SPRINGER
REFERENCE

Robert A. Meyers
Editor-in-Chief

VOLUME 1
A–Cl

Encyclopedia of Complexity and Systems Science

Springer

Robert A. Meyers (Ed.)

Encyclopedia of Complexity and Systems Science

With 4300 Figures and 420 Tables

Springer

Robert A. Meyers, Ph. D.
Editor-in-Chief
RAMTECH LIMITED
122 Escalle Lane
Larkspur, CA 94939
USA
robert.meyers@ramtechlimited.org

Library of Congress Control Number: 2008933604

ISBN: 978-0-387-30440-3

This publication is available also as:
Print publication under ISBN: 978-0-387-75888-6 and
Print and electronic bundle under ISBN: 978-0-387-69572-3

© 2009 SpringerScience+Buisiness Media, LLC.

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC., 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

springer.com

Printed on acid free paper SPIN: 11560258 2109letex – 5 4 3 2 1 0

Preface

The Encyclopedia of Complexity and System Science is an authoritative single source for understanding and applying the basic tenets of complexity and systems theory as well as the tools and measures for analyzing complex systems in science, engineering and many areas of social, financial and business interactions. It is written for an audience of advanced university undergraduate and graduate students, professors, and professionals in a wide range of fields who must manage complexity on scales ranging from the atomic and molecular to the societal and global. Each article was selected and peer reviewed by one of our 36 Section Editors with advice and consultation provided by our 15 Board Members and Editor-in-Chief. This level of coordination assures that the reader can have a level of confidence in the relevance and accuracy of the information far exceeding that generally found on the World Wide Web. Accessablilty is also a priority and for this reason each article includes a glossary of important terms and a concise definition of the subject.

Complex systems are systems that comprise many interacting parts with the ability to generate a new quality of collective behavior through self-organization, e. g. the spontaneous formation of temporal, spatial or functional structures. They are therefore adaptive as they evolve and may contain self-driving feedback loops. Thus, complex systems are much more than a sum of their parts. Complex systems are often characterized as having extreme sensitivity to initial conditions as well as emergent behavior that are not readily predictable or even completely deterministic. The conclusion is that a reductionist (bottom-up) approach is often an incomplete description of a phenomenon. This recognition, that the collective behavior of the whole system cannot be simply inferred from the understanding of the behavior of the individual components, has led to many new concepts and sophisticated mathematical and modeling tools for application to many scientific, engineering, and societal issues that can be adequately described only in terms of complexity and complex systems.

Examples of Grand Scientific Challenges which can be approached through complexity and systems science include: the structure, history and future of the universe; the biological basis of consciousness; the true complexity of the genetic makeup and molecular functioning of humans (genetics and epigenetics) and other life forms; human longevity limits; unification of the laws of physics; the dynamics and extent of climate change and the effects of climate change; extending the boundaries of and understanding the theoretical limits of computing; sustainability of life on the earth; workings of the interior of the earth; predictability, dynamics and extent of earthquakes, tsunamis, and other natural disasters; dynamics of turbulent flows and the motion of granular materials; the structure of atoms as expressed in the Standard Model and the formulation of the Standard Model and gravity into a Unified Theory; the structure of water; control of global infectious diseases and also evolution and quantification of (ultimately) human cooperative behavior in politics, economics, business systems and social interactions. In fact, most of these issues have identified nonlinearities and are beginning to be addressed with nonlinear techniques – e. g. human longevity limits; the Standard Model; climate change, earthquake prediction, workings of the earth's interior, natural disaster prediction, etc.

The complex systems mathematical and modeling tools and scientific and engineering applications that comprise the Encyclopedia of Complexity and Systems Science are as follows:

- Agent Based Modeling and Simulation
- Applications of Physics and Mathematics to Social Science
- Cellular Automata, Mathematical Basis of
- Chaos and Complexity in Astrophysics
- Climate Modeling, Global Warming and Weather Prediction
- Complex Networks and Graph Theory

- Complexity and Nonlinearity in Autonomous Robotics
- Complexity in Computational Chemistry
- Complexity in Earthquakes, Tsunamis, and Volcanoes, and Forecasting and Early Warning of their Hazards
- Computational and Theoretical Nanoscience
- Control and Dynamical Systems
- Data Mining and Knowledge Discovery
- Ecological Complexity
- Ergodic Theory
- Finance and Econometrics
- Fractals and Multifractals
- Game Theory
- Granular Computing
- Intelligent Systems
- Nonlinear Ordinary Differential Equations and Dynamical Systems
- Nonlinear Partial Differential Equations
- Percolation
- Perturbation Theory
- Probability and Statistics in Complex Systems
- Quantum Information Science
- Social Network Analysis
- Soft Computing
- Solitons
- Statistical and Nonlinear Physics
- Synergetics
- System Dynamics
- Systems Biology
- Traffic Management, Complex Dynamics of
- Unconventional Computing
- Wavelets

The 15 members of our Editorial Advisory Board include: Ahmed Zewail of Caltech, Nobel Prize in chemistry, who provided guidance in chemistry and physical biology applications of complexity and systems science; Thomas Schelling, Nobel Prize in Economics, provided guidance in economics and political science applications; Mario J. Molina, Nobel Prize in Chemistry, whose current interests include: atmospheric chemical processes and science-policy issues related to urban and regional air pollution and to global change; Manfred Eigen, Nobel Prize in Chemistry, who is expert in molecular self-organization and evolutionary biotechnology; Benoit B. Mandelbrot of Yale University and Battelle Pacific Northwest Laboratories, Wolf Prize for Physics, who provided guidance in physical, mathematical or social phenomena that are characterized by abundant data but wild variability; Richard E. Stearns, 1993 Turing Award, whose current interests include computational complexity, automata theory, analysis of algorithms and game theory; Pierre-Louis Lions, 1994 Fields Medal, whose interests are nonlinear partial differential equations and applications; Leroy Hood of the Institute for Systems Biology, Lasker Award, whose specialty is systems approaches to biology and medicine; Steven Wolfram, creator of Mathematica®; and Lotfi Zadeh, Honda and Okawa Prizes and IEEE Medal of Honor, whose current interests are fuzzy logic and soft computing. Sadly, Paul Lauterbur, Nobel Prize in Medicine or Physiology, who provided guidance in evolution and origin of life and overall relationship of determinism to complexity passed away during the project. A complete listing of our Board is presented immediately following this Preface.

Our 36 Section Editors, who supervised the 35 Sections, represent some of the best and brightest in their respective fields. It is notable that some are rather young, and yet are highly accomplished, as seems appropriate for the very modern scientific approach inherent in complexity and systems science. The Section Editors have selected both the articles (described below) and also nominated our authors and either provided peer review or supervised outside peer reviewers. A complete listing of the Section Editors is presented immediately following this Preface.

The efforts of the 41 members of our Board and Section Editor team have resulted in near 300 articles on the mathematical and modeling basis of complexity, e. g. fractals and multifractals, cellular automata, game theory, quantum information processing, unconventional computing, ergodic theory, percolation, non-linear ordinary differential equations, non-linear partial differential equations, perturbation theory, probability and statistics, solitons, wavelets, agent based modeling and simulation, complex networks and graph theory, data mining and knowledge discovery, granular computing and soft computing; as well as near 300 articles describing progress using these tools in physics, econometrics, ecosystems, climate prediction, nanoelectronics, complex networks, quantum computing, chemistry, astrophysics, geophysics, systems biology, bioinformatics, medicine, system dynamics, engineering, control and dynamical systems, traffic management, and robotics as well as social, economics and political sciences.

This Encyclopedia in total comprises 592 peer-reviewed, in-depth, 15–20 page articles presented alphabetically as prepared by more than 700 authors. The result is a carefully organized, accessible, and easily searchable 11,000-page reference work that places complexity and systems science at the center of modern scientific, engineering, and societal problems and potential solutions. A complete alphabetized listing of articles, extensive cross-references, glossaries of terms, and a detailed index are all included. We believe that there is no other treatment of this field with near the depth and authority of that prepared by our team of prize winning Board members, Section Editors and authors. This new publication will extend the influence of complexity and systems science to a much wider audience than has been possible up until now.

Acknowledgments

I wish to thank, David Packer, Executive Editor, who proposed the project to me and then provided invaluable counsel in performance of many elements of the publication process.

I wish to express our profound gratitude to Kerstin Kindler, Reference Publisher and Julia Koerting, Editorial Assistant for their outstanding editorial efforts in producing this Encyclopedia. Their interaction with our team of over 700 authors, 36 section editors and 15 board members was highly professional, courteous, pleasant and yet firm, which was integral in obtaining high quality and timely delivery of our manuscripts. Their mastery of formal as well as idiomatic English facilitated many of these interactions.

Robert A. Meyers
Editor in Chief
Larkspur, California
March 2009

Topical Table of Contents

Agent Based Modeling and Simulation, Section Editor: Filippo Castiglione

Agent Based Computational Economics
Agent Based Modeling and Artificial Life
Agent Based Modeling and Computer Languages
Agent Based Modeling and Simulation, Introduction to
Agent Based Modeling, Large Scale Simulations
Agent Based Modeling, Mathematical Formalism for
Agent-Based Modeling and Simulation
Cellular Automaton Modeling of Tumor Invasion
Computer Graphics and Games, Agent Based Modeling in
Embodied and Situated Agents, Adaptive Behavior in
Interaction Based Computing in Physics
Logic and Geometry of Agents in Agent-Based Modeling
Social Phenomena Simulation
Swarm Intelligence

Autonomous Robotics, Complexity and Nonlinearity in, Section Editor: Warren Dixon

Adaptive Visual Servo Control
Cognitive Robotics
Complexity and Non-Linearity in Autonomous Robotics, Introduction to
Continuum Robots
Distributed Controls of Multiple Robotic Systems, An Optimization Approach
Distributed Robotic Teams: A Framework for Simulated and Real-World Modeling
Foraging Robots
Human Robot Interaction
Image Based State Estimation
Modular Self-Reconfigurable Robots
Motion Prediction for Continued Autonomy
Multiple Mobile Robot Teams, Path Planning and Motion Coordination in
Neuro-fuzzy Control of Autonomous Robotics
Self-replicating Robotic Systems
Software Architectures for Autonomy

Cellular Automata, Mathematical Basis of, Section Editor: Andrew Adamatzky

Additive Cellular Automata
Algorithmic Complexity and Cellular Automata
Cellular Automata and Groups
Cellular Automata and Language Theory
Cellular Automata as Models of Parallel Computation
Cellular Automata in Hyperbolic Spaces

Cellular Automata Modeling of Physical Systems
Cellular Automata on Triangular, Pentagonal and Hexagonal Tessellations
Cellular Automata with Memory
Cellular Automata, Classification of
Cellular Automata, Emergent Phenomena in
Cellular Automata, Universality of
Chaotic Behavior of Cellular Automata
Dynamics of Cellular Automata in Non-compact Spaces
Ergodic Theory of Cellular Automata
Evolving Cellular Automata
Firing Squad Synchronization Problem in Cellular Automata
Gliders in Cellular Automata
Growth Phenomena in Cellular Automata
Identification of Cellular Automata
Mathematical Basis of Cellular Automata, Introduction to
Phase Transitions in Cellular Automata
Quantum Cellular Automata
Reversible Cellular Automata
Self-organised Criticality and Cellular Automata
Self-Replication and Cellular Automata
Structurally Dynamic Cellular Automata
Tiling Problem and Undecidability in Cellular Automata
Topological Dynamics of Cellular Automata

Chaos and Complexity in Astrophysics, Section Editor: Steve N. Shore

Acceleration Mechanisms
Astronomical Time Series, Complexity in
Astrophysics, Chaos and Complexity in
Astrophysics: Dynamical Systems
Chaos and Complexity in Astrophysics, Introduction to
Cosmic Gravitational Background, Stochastic
Cosmic Strings
Exobiology (theoretical), Complexity in
Exobiology and Complexity
Orbital Dynamics, Chaos in
Self-Organization in Magnetohydrodynamic Turbulence
Space Plasmas, Dynamical Complexity in
Stellar Dynamics, N-body Methods for
Topological Magnetohydrodynamics and Astrophysics

Climate Modeling, Global Warming and Weather Prediction, Section Editor: Hartmut Grassl

Abrupt Climate Change Modeling
Climate Change and Agriculture
Climate Change and Human Health
Climate Change, Economic Costs of
Climate Modeling, Global Warming and Weather Prediction, Introduction to
Cryosphere Models
Regional Climate Models: Linking Global Climate Change to Local Impacts
Single Column Modeling of Atmospheric Boundary Layers
and the Complex Interactions with the Land Surface

Complex Networks and Graph Theory, Section Editor: Geoffrey Canright

Community Structure in Graphs
Complex Gene Regulatory Networks – From Structure to Biological Observables: Cell Fate Determination
Complex Networks and Graph Theory
Complex Networks, Visualization of
Food Webs
Growth Models for Networks
Human Sexual Networks
Internet Topology
Link Analysis and Web Search
Motifs in Graphs
Non-negative Matrices and Digraphs
Random Graphs, A Whirlwind Tour of
Synchronization Phenomena on Networks
World Wide Web, Graph Structure

Complexity in Computational Chemistry, Section Editor: Danail Bonchev

Biochemistry, Chaotic Dynamics, Noise, and Fractal Space in
Biological Complexity and Biochemical Information
Biological Development and Evolution, Complexity and Self-Organization in
Cellular Automata Modeling of Complex Biochemical Systems
Composites, Multifunctional
Computational Chemistry, Introduction to Complexity in
Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis
DNA-templated Self-assembly of Protein Arrays and Highly Conductive Nanowires
Drug Design with Artificial Intelligence Methods
Drug Design with Artificial Neural Networks
Drug Design with Machine Learning
Drug Design, Molecular Descriptors in
Information Theoretic Complexity Measures
Molecular Evolution, Networks in
Nanoscale Atomic Clusters, Complexity of
Polymers, Nonlinearity in
QSAR Modeling and QSAR Based Virtual Screening, Complexity and Challenges of Modern
Quantum Similarity and Quantum Quantitative Structure-Properties Relationships (QQSPR)
Self-assembled Materials
Topological Complexity of Molecules

Complexity in Earthquakes, Tsunamis, and Volcanoes, and Forecast, Section Editor: William H. K. Lee

Brittle Tectonics: A Non-linear Dynamical System
Complexity in Earthquakes, Tsunamis, and Volcanoes, and Forecast, Introduction to
Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of
Earthquake Clusters over Multi-dimensional Space, Visualization of
Earthquake Damage: Detection and Early Warning in Man-Made Structures
Earthquake Early Warning System in Southern Italy
Earthquake Engineering, Non-linear Problems in
Earthquake Forecasting and Verification
Earthquake Location, Direct, Global-Search Methods
Earthquake Magnitude
Earthquake Monitoring and Early Warning Systems

Earthquake Networks, Complex
Earthquake Nucleation Process
Earthquake Occurrence and Mechanisms, Stochastic Models for
Earthquake Scaling Laws
Earthquake Source Parameters, Rapid Estimates for Tsunami Warning
Earthquake Source: Asymmetry and Rotation Effects
Earthquakes, Dynamic Triggering of
Earthquakes, Electromagnetic Signals of
Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in
Geo-Complexity and Earthquake Prediction
GPS: Applications in Crustal Deformation Monitoring
Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures
Infrasound from Earthquakes, Tsunamis and Volcanoes
Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation
Seismic Wave Propagation in Media with Complex Geometries, Simulation of
Seismic Waves in Heterogeneous Earth, Scattering of
Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space
Seismicity, Statistical Physics Approaches to
Slug Flow: Modeling in a Conduit and Associated Elastic Radiation
Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy
Tomography, Seismic
Tsunami Earthquakes
Tsunami Forecasting and Warning
Tsunami Inundation, Modeling of
Tsunamis, Inverse Problem of
Volcanic Eruptions, Explosive: Experimental Insights
Volcanic Eruptions: Cyclicity During Lava Dome Growth
Volcanic Eruptions: Stochastic Models of Occurrence Patterns
Volcanic Hazards and Early Warning
Volcano Seismic Signals, Source Quantification of
Volcanoes, Non-linear Processes in
Wedge Mechanics: Relation With Subduction Zone Earthquakes and Tsunamis

Computational and Theoretical Nanoscience, Section Editor: Yong Suk Joe

Carbon Nanotubes, Thermo-mechanical and Transport Properties of
Charge Based Solid-State Flying Qubits
Computational and Theoretical Nanoscience, Introduction to
Field Computation in Natural and Artificial Intelligence
Geometric Phase and Related Phenomena in Quantum Nanosystems
Multimillion Atom Simulations with Nemo3D
Nanoscale Processes, Modeling Coupled and Transport Phenomena in Nanotechnology
Quantum Dot Spin Transistors, Self-consistent Simulation of
Quantum Dots: Fano Resonances in Aharonov–Bohm Ring
Quantum Impurity Physics in Coupled Quantum Dots
Quantum Phenomena in Semiconductor Nanostructures
Quantum Simulations of Ballistic Nanowire Field Effect Transistors
Resonances in Electronic Transport Through Quantum Wires and Rings
Semiclassical Spin Transport in Spin-Orbit Coupled Systems
Spin Dependent Exchange and Correlation in Two-Dimensional Electron Layers
Spin Dynamics in Disordered Solids
Spin-polarized Quantum Transport in Mesoscopic Conductors: Computational Concepts and Physical Phenomena

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation
Tunneling Through Quantum Dots with Discrete Symmetries
Viral Protein Nano-Actuators, Computational Studies of Bio-nanomachines

Data Mining and Knowledge Discovery, Section Editor: Peter Kokol

Data and Dimensionality Reduction in Data Analysis and System Modeling
Data-Mining and Knowledge Discovery, Introduction to
Data-Mining and Knowledge Discovery, Neural Networks in
Data-Mining and Knowledge Discovery: Case Based Reasoning, Nearest Neighbor and Rough Sets
Decision Trees
Discovery Systems
Genetic and Evolutionary Algorithms and Programming: General Introduction and Application to Game Playing
Knowledge Discovery: Clustering
Machine Learning, Ensemble Methods in
Manipulating Data and Dimension Reduction Methods: Feature Selection

Ecological Complexity, Section Editor: Bai-Lian Li

Ecological Complexity
Ecological Topology and Networks
Entropy Maximization and Species Abundance
Human-Environment Interactions, Complex Systems Approaches for Dynamic Sustainable Development

EiC Selections, Section Editor: Robert A. Meyers

Catastrophe Theory
Coordination Dynamics
Infinite Dimensional Controllability
Philosophy of Science, Mathematical Models in
Self-organizing Systems

Ergodic Theory, Section Editor: Bryna Kra

Chaos and Ergodic Theory
Entropy in Ergodic Theory
Ergodic Theorems
Ergodic Theory on Homogeneous Spaces and Metric Number Theory
Ergodic Theory, Introduction to
Ergodic Theory: Basic Examples and Constructions
Ergodic Theory: Fractal Geometry
Ergodic Theory: Interactions with Combinatorics and Number Theory
Ergodic Theory: Non-singular Transformations
Ergodic Theory: Recurrence
Ergodic Theory: Rigidity
Ergodicity and Mixing Properties
Isomorphism Theory in Ergodic Theory
Joinings in Ergodic Theory
Measure Preserving Systems
Pressure and Equilibrium States in Ergodic Theory
Smooth Ergodic Theory
Spectral Theory of Dynamical Systems
Symbolic Dynamics
Topological Dynamics

Finance and Econometrics, Section Editor: Bruce Mizrach

Bayesian Methods in Non-linear Time Series
Corporate and Municipal Bond Market Microstructure in the U.S.
Econometrics: Models of Regime Changes
Econometrics: Nonlinear Cointegration
Econometrics: Panel Data Methods
Econophysics, Observational
Finance and Econometrics, Introduction to
Finance, Agent Based Modeling in
Financial Economics, Fat-Tailed Distributions
Financial Economics, Non-linear Time Series in
Financial Economics, Return Predictability and Market Efficiency
Financial Economics, The Cross-Section of Stock Returns and the Fama-French Three Factor Model
Financial Economics, Time Variation in the Market Return
Financial Forecasting, Non-linear Time Series in
Financial Forecasting, Sensitive Dependence
GARCH Modeling
Macroeconomics, Nonlinear Time Series in
Market Microstructure
Market Microstructure, Foreign Exchange
Microeconometrics
Nonparametric Tests for Independence
Stochastic Volatility
Treasury Market, Microstrucuture of the U.S.

Fractals and Multifractals, Section Editor: Daniel ben-Avraham and Shlomo Havlin

Anomalous Diffusion on Fractal Networks
Dynamics on Fractals
Fractal and Multifractal Scaling of Electrical Conduction in Random Resistor Networks
Fractal and Multifractal Time Series
Fractal and Transfractal Scale-Free Networks
Fractal Geometry, A Brief Introduction to
Fractal Growth Processes
Fractal Structures in Condensed Matter Physics
Fractals and Economics
Fractals and Multifractals, Introduction to
Fractals and Percolation
Fractals and Wavelets: What can we Learn on Transcription and Replication
from Wavelet-Based Multifractal Analysis of DNA Sequences?
Fractals in Biology
Fractals in Geology and Geophysics
Fractals in the Quantum Theory of Spacetime
Fractals Meet Chaos
Phase Transitions on Fractals and Networks
Reaction Kinetics in Fractals

Game Theory, Section Editor: Marilda Sotomayor

Bayesian Games: Games with Incomplete Information
Cooperative Games
Cooperative Games (Von Neumann–Morgenstern Stable Sets)

Correlated Equilibria and Communication in Games
Cost Sharing
Differential Games
Dynamic Games with an Application to Climate Change Models
Evolutionary Game Theory
Fair Division
Game Theory and Strategic Complexity
Game Theory, Introduction to
Implementation Theory
Inspection Games
Learning in Games
Market Games and Clubs
Mechanism Design
Networks and Stability
Principal-Agent Models
Repeated Games with Complete Information
Repeated Games with Incomplete Information
Reputation Effects
Signaling Games
Static Games
Stochastic Games
Two-Sided Matching Models
Voting
Voting Procedures, Complexity of
Zero-sum Two Person Games

Granular Computing, Section Editor: Tsau Y. Lin

Cooperative Multi-Hierarchical Query Answering Systems
Dependency and Granularity in Data Mining
Fuzzy Logic
Fuzzy Probability Theory
Fuzzy System Models Evolution from Fuzzy Rulebases to Fuzzy Functions
Genetic-Fuzzy Data Mining Techniques
Granular Model for Data Mining
Granular Computing and Data Mining for Ordered Data: The Dominance-Based Rough Set Approach
Granular Computing and Modeling of the Uncertainty in Quantum Mechanics
Granular Computing System Vulnerabilities: Exploring the Dark Side of Social Networking Communities
Granular Computing, Information Models for
Granular Computing, Introduction to
Granular Computing, Philosophical Foundation for
Granular Computing, Principles and Perspectives of
Granular Computing: Practices, Theories and Future Directions
Granular Neural Network
Granulation of Knowledge: Similarity Based Approach in Information and Decision Systems
Multi-Granular Computing and Quotient Structure
Non-standard Analysis, An Invitation to
Rough and Rough-Fuzzy Sets in Design of Information Systems
Rough Set Data Analysis
Rule Induction, Missing Attribute Values and Discretization
Social Networks and Granular Computing

Intelligent Systems, Section Editor: James A. Hendler

Artificial Intelligence in Modeling and Simulation
Intelligent Control
Intelligent Systems, Introduction to
Learning and Planning (Intelligent Systems)
Mobile Agents
Semantic Web

Non-Linear Ordinary Differential Equations and Dynamical Systems, Section Editor: Ferdinand Verhulst

Center Manifolds
Dynamics of Hamiltonian Systems
Dynamics of Parametric Excitation
Existence and Uniqueness of Solutions of Initial Value Problems
Hyperbolic Dynamical Systems
Lyapunov–Schmidt Method for Dynamical Systems
Non-linear Ordinary Differential Equations and Dynamical Systems, Introduction to
Numerical Bifurcation Analysis
Periodic Orbits of Hamiltonian Systems
Periodic Solutions of Non-autonomous Ordinary Differential Equations
Relaxation Oscillations
Stability Theory of Ordinary Differential Equations

Non-Linear Partial Differential Equations, Section Editor: Italo Capuzzo Dolcetta

Biological Fluid Dynamics, Non-linear Partial Differential Equations
Control of Nonlinear Partial Differential Equations
Dispersion Phenomena in Partial Differential Equations
Hamilton-Jacobi Equations and weak KAM Theory
Hyperbolic Conservation Laws
Navier-Stokes Equations: A Mathematical Analysis
Non-linear Partial Differential Equations, Introduction to
Non-linear Partial Differential Equations, Viscosity Solution Method in
Non-linear Stochastic Partial Differential Equations
Scaling Limits of Large Systems of Nonlinear Partial Differential Equations
Vehicular Traffic: A Review of Continuum Mathematical Models

Percolation, Section Editor: Muhammad Sahimi

Bootstrap Percolation
Conduction and Diffusion in Percolating Systems
Continuum Percolation
Correlated Percolation
Elastic Percolation Networks
Invasion Percolation
Networks, Flexibility and Mobility in
Percolation and Polymer Morphology and Rheology
Percolation in Complex Networks
Percolation in Porous Media
Percolation Lattices, Efficient Simulation of Large
Percolation Phase Transition
Percolation Thresholds, Exact
Percolation, and Faults and Fractures in Rock

Percolation, Introduction to
Scaling Properties, Fractals, and the Renormalization Group Approach to Percolation

Perturbation Theory, Section Editor: Giuseppe Gaeta

Diagrammatic Methods in Classical Perturbation Theory
Hamiltonian Perturbation Theory (and Transition to Chaos)
Kolmogorov-Arnold-Moser (KAM) Theory
N-body Problem and Choreographies
Nekhoroshev Theory
Non-linear Dynamics, Symmetry and Perturbation Theory in
Normal Forms in Perturbation Theory
Perturbation Analysis of Parametric Resonance
Perturbation of Equilibria in the Mathematical Theory of Evolution
Perturbation of Systems with Nilpotent Real Part
Perturbation Theory
Perturbation Theory and Molecular Dynamics
Perturbation Theory for Non-smooth Systems
Perturbation Theory for PDEs
Perturbation Theory in Celestial Mechanics
Perturbation Theory in Quantum Mechanics
Perturbation Theory, Introduction to
Perturbation Theory, Semiclassical
Perturbative Expansions, Convergence of
Quantum Bifurcations

Probability and Statistics in Complex Systems, Section Editor: Henrik Jeldtoft Jensen

Bayesian Statistics
Branching Processes
Complexity in Systems Level Biology and Genetics: Statistical Perspectives
Correlations in Complex Systems
Entropy
Extreme Value Statistics
Field Theoretic Methods
Fluctuations, Importance of: Complexity in the View of Stochastic Processes
Hierarchical Dynamics
Levy Statistics and Anomalous Transport: Levy Flights and Subdiffusion
Probability and Statistics in Complex Systems, Introduction to
Probability Densities in Complex Systems, Measuring
Probability Distributions in Complex Systems
Random Matrix Theory
Random Walks in Random Environment
Record Statistics and Dynamics
Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems
Stochastic Processes

Quantum Information Science, Section Editor: Joseph F. Traub

Quantum Algorithms
Quantum Algorithms and Complexity for Continuous Problems
Quantum Computational Complexity
Quantum Computing Using Optics

Quantum Computing with Trapped Ions
Quantum Cryptography
Quantum Error Correction and Fault Tolerant Quantum Computing
Quantum Information Processing
Quantum Information Science, Introduction to

Social Network Analysis, Section Editor: John Scott

Network Analysis, Longitudinal Methods of
Positional Analysis and Blockmodelling
Social Network Analysis, Estimation and Sampling in
Social Network Analysis, Graph Theoretical Approaches to
Social Network Analysis, Large-Scale
Social Network Analysis, Overwiew of
Social Network Analysis, Two-Mode Concepts in
Social Network Visualization, Methods of
Social Networks, Algebraic Models for
Social Networks, Diffusion Processes in
Social Networks, Exponential Random Graph (p*) Models for

Social Science, Physics and Mathematics Applications in, Section Editor: Andrzej Nowak

Agent Based Modeling and Neoclassical Economics: A Critical Perspective
Agent Based Models in Economics and Complexity
Applications of Physics and Mathematics to Social Science, Introduction to
Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies
Consciousness and Complexity
Development, Complex Dynamic Systems of
Development, Evolution, and the Emergence of Novel Behavior
Dynamics and Evaluation: The Warm Glow of Processing Fluency
Dynamics of Language
Evolution of Culture, Memetics
Extreme Events in Socio-economic and Political Complex Systems, Predictability of
Human Behavior, Dynamics of
Intermittency and Localization
Investment Decision Making in Finance, Models of
Marketing: Complexity Modeling, Theory and Applications in
Minority Games
Moral Dynamics
Opinions Dynamics and Sociophysics
Physics and Mathematics Applications in Social Science
Rational, Goal-Oriented Agents
Social Cognitive Complexity
Social Coordination, from the Perspective of Coordination Dynamics
Social Organizations with Complexity Theory: A Dramatically Different Lens for the Knowledge Economy
Social Processes, Physical Models of
Social Processes, Simulation Models of
Social Psychology, Applications of Complexity to
Traffic and Crowd Dynamics: The Physics of the City

Soft Computing, Section Editor: Janusz Kacprzyk

Aggregation Operators and Soft Computing
Evolving Fuzzy Systems

Fuzzy Logic, Type-2 and Uncertainty
Fuzzy Optimization
Fuzzy Sets Theory, Foundations of
Hybrid Soft Computing Models for Systems Modeling and Control
Neuro-fuzzy Systems
Possibility Theory
Rough Sets in Decision Making
Rough Sets: Foundations and Perspectives
Soft Computing, Introduction to
Statistics with Imprecise Data

Solitons, Section Editor: Mohamed A. Helal

Adomian Decomposition Method Applied to Non-linear Evolution Equations in Soliton Theory
Inverse Scattering Transform and the Theory of Solitons
Korteweg–de Vries Equation (KdV), Different Analytical Methods for Solving the
Korteweg–de Vries Equation (KdV) and Modified Korteweg–de Vries Equations (mKdV), Semi-analytical Methods for Solving the
Korteweg–de Vries Equation (KdV), Some Numerical Methods for Solving the
Korteweg–de Vries Equation (KdV) History, Exact N-Soliton Solutions and Further Properties
Non-linear Internal Waves
Partial Differential Equations that Lead to Solitons
Shallow Water Waves and Solitary Waves
Soliton Perturbation
Solitons and Compactons
Solitons Interactions
Solitons, Introduction to
Solitons, Tsunamis and Oceanographical Applications of
Solitons: Historical and Physical Introduction
Water Waves and the Korteweg–de Vries Equation

Statistical and Nonlinear Physics, Section Editor: M. Cristina Marchetti

Anisotropic Networks, Elastomers and Gels
Cell Biology: Networks, Regulation and Pathways
Chaotic Dynamics in Nonequilibrium Statistical Mechanics
Collective Transport and Depinning
Complex Systems and Emergent Phenomena
Cytoskeleton and Cell Motility
Disordered Elastic Media
Econophysics, Statistical Mechanics Approach to
Fluctuation Theorems, Brownian Motors and Thermodynamics of Small Systems
Glasses and Aging, A Statistical Mechanics Perspective on
Granular Flows
Jamming of Granular Matter
Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of
Microfluidics
Monte Carlo Simulations in Statistical Physics
Networks: Structure and Dynamics
Neuronal Dynamics
Noise and Stability in Modelocked Soliton Lasers
Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence
Optimization Problems and Algorithms from Computer Science

Polymer Physics
Protein Mechanics at the Single-Molecule Level
Quantum Chaos
Statistical and Non-linear Physics, Introduction to
Ultracold Atomic Gases: Novel States of Matter

Synergetics, Section Editor: Hermann Haken

Brain Pacemaker
Fluid Dynamics, Pattern Formation
Fluid Dynamics, Turbulence
Intentionality: A Naturalization Proposal on the Basis of Complex Dynamical Systems
Linear and Non-linear Fokker–Planck Equations
Movement Coordination
Patterns and Interfaces in Dissipative Dynamics
Self-Organization and Clinical Psychology
Self-Organization and the City
Synergetics, Introduction to
Synergetics: Basic Concepts

System Dynamics, Section Editor: Brian Dangerfield

Business Policy and Strategy, System Dynamics Applications to
Delay and Disruption in Complex Projects
Diffusion of Innovations, System Dynamics Analysis of the
Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation
Group Model Building
Health Care in the United Kingdom and Europe, System Dynamics Applications to
Health Care in the United States, System Dynamics Applications to
Public Policy, System Dynamics Applications to
Scenario-Driven Planning with System Dynamics
System Dynamics and Its Contribution to Economics and Economic Modeling
System Dynamics and Organizational Learning
System Dynamics in the Evolution of the Systems Approach
System Dynamics Modeling: Validation for Quality Assurance
System Dynamics Models of Environment, Energy and Climate Change
System Dynamics Models, Optimization of
System Dynamics Philosophical Background and Underpinnings
System Dynamics, Analytical Methods for Structural Dominance Analysis in
System Dynamics, Introduction to
System Dynamics, The Basic Elements of

Systems and Control Theory, Section Editor: Matthias Kawski

Chronological Calculus in Systems and Control Theory
Discrete Control Systems
Finite Dimensional Controllability
Hybrid Control Systems
Learning, System Identification, and Complexity
Maximum Principle in Optimal Control
Mechanical Systems: Symmetries and Reduction
Nonsmooth Analysis in Systems and Control Theory
Observability (Deterministic Systems) and Realization Theory

Robotic Networks, Distributed Algorithms for
Stability and Feedback Stabilization
Stochastic Noises, Observation, Identification and Realization with
System Regulation and Design, Geometric and Algebraic Methods in
Systems and Control, Introduction to

Systems Biology, Section Editor: Timothy P. Galitski

Biological Data Integration and Model Building
Biological Models of Molecular Network Dynamics
Biomolecular Network Structure and Function
Boolean Modeling of Biological Networks
Ecological Systems
Functional Genomics for Characterization of Genome Sequences
Genome Organization
Metabolic Systems Biology
Stochastic Models of Biological Processes
Systems Biology of Human Immunity and Disease
Systems Biology, Introduction to
Systems Genetics and Complex Traits

Traffic Management, Complex Dynamics of, Section Editor: Boris Kerner

Air Traffic Control, Complex Dynamics of
Complex Dynamics of Traffic Management, Introduction to
Evacuation as a Communication and Social Phenomenon
Evacuation Dynamics: Empirical Results, Modeling and Applications
Freeway Traffic Management and Control
Pedestrian, Crowd and Evacuation Dynamics
Traffic Breakdown, Probabilistic Theory of
Traffic Congestion, Modeling Approaches to
Traffic Congestion, Spatiotemporal Features of
Traffic Networks, Optimization and Control of Urban
Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment
Traffic Prediction of Congested Patterns
Travel Behaviour and Demand Analysis and Prediction

Unconventional Computing, Section Editor: Andrew Adamatzky

Amorphous Computing
Analog Computation
Artificial Chemistry
Bacterial Computing
Cellular Computing
Computing in Geometrical Constrained Excitable Chemical Systems
Computing with Solitons
DNA Computing
Evolution in Materio
Immunecomputing
Mechanical Computing: The Computational Complexity of Physical Devices
Membrane Computing
Molecular Automata
Nanocomputers

Optical Computing
Quantum Computing
Reaction-Diffusion Computing
Reversible Computing
Thermodynamics of Computation
Unconventional Computing, Introduction to
Unconventional Computing, Novel Hardware for

Wavelets, Section Editor: Edward Aboufadel

Bivariate (Two-dimensional) Wavelets
Comparison of Discrete and Continuous Wavelet Transforms
Curvelets and Ridgelets
Multivariate Splines and Their Applictions
Multiwavelets
Numerical Issues When Using Wavelets
Popular Wavelet Families and Filters and Their Use
Statistical Applications of Wavelets
Wavelets and PDE Techniques in Image Processing, A Quick Tour of
Wavelets and the Lifting Scheme
Wavelets, Introduction to

About the Editor-in-Chief

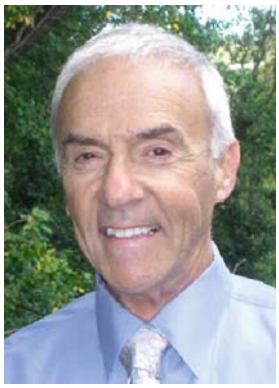

Robert A. Meyers

President: RAMTECH Limited
Manager, Chemical Process Technology, TRW Inc.
Post-doctoral Fellow: California Institute of Technology
Ph. D. Chemistry, University of California at Los Angeles
B. A., Chemistry, California State University, San Diego

Biography

Dr. Meyers has worked with more than 25 Nobel laureates during his career.

Research

Dr. Meyers was Manager of Chemical Technology at TRW (now Northrop Grumman) in Redondo Beach, CA and is now President of RAMTECH Limited. He is co-inventor of the Gravimelt process for desulfurization and demineralization of coal for air pollution and water pollution control. Dr. Meyers is the inventor of and was project manager for the DOE-sponsored Magnetohydrodynamics Seed Regeneration Project which has resulted in the construction and successful operation of a pilot plant for production of potassium formate, a chemical utilized for plasma electricity generation and air pollution control. Dr. Meyers managed the pilot-scale DoE project for determining the hydrodynamics of synthetic fuels. He is a co-inventor of several thermo-oxidative stable polymers which have achieved commercial success as the GE PEI, Upjohn Polyimides and Rhone-Polenc bismaleimide resins. He has also managed projects for photochemistry, chemical lasers, flue gas scrubbing, oil shale analysis and refining, petroleum analysis and refining, global change measurement from space satellites, analysis and mitigation (carbon dioxide and ozone), hydrometallurgical refining, soil and hazardous waste remediation, novel polymers synthesis, modeling of the economics of space transportation systems, space rigidizable structures and chemiluminescence-based devices.

He is a senior member of the American Institute of Chemical Engineers, member of the American Physical Society, member of the American Chemical Society and serves on the UCLA Chemistry Department Advisory Board. He was a member of the joint USA-Russia working group on air pollution control and the EPA-sponsored Waste Reduction Institute for Scientists and Engineers.

Dr. Meyers has more than 20 patents and 50 technical papers. He has published in primary literature journals including Science and the Journal of the American Chemical Society, and is listed in Who's Who in America and Who's Who in the World. Dr. Meyers' scientific achievements have been reviewed in feature articles in the popular press in publications such as The New York Times Science Supplement and The Wall Street Journal as well as more specialized publications such as Chemical Engineering and Coal Age. A public service film was produced by the Environmental Protection Agency of Dr. Meyers' chemical desulfurization invention for air pollution control.

Scientific Books

Dr. Meyers is the author or Editor-in-Chief of 12 technical books one of which won the Association of American Publishers Award as the best book in technology and engineering.

Encyclopedias

Dr. Meyers conceived and has served as Editor-in-Chief of the Academic Press (now Elsevier) Encyclopedia of Physical Science and Technology. This is an 18-volume publication of 780 twenty-page articles written to an audience of university students and practicing professionals. This encyclopedia, first published in 1987, was very successful, and because of this, was revised and reissued in 1992 as a second edition. The Third Edition was published in 2001 and is now on-line. Dr. Meyers has completed two editions of the Encyclopedia of Molecular Cell Biology and Molecular Medicine for Wiley VCH publishers (1995 and 2004). These cover molecular and cellular level genetics, biochemistry, pharmacology, diseases and structure determination as well as cell biology. His eight-volume Encyclopedia of Environmental Analysis and Remediation was published in 1998 by John Wiley & Sons and his 15-volume Encyclopedia of Analytical Chemistry was published in 2000, also by John Wiley & Sons, all of which are available on-line.

Editorial Board Members

Professor Manfred Eigen *
of the MPI for Biophysical Chemistry
1967 Nobel Prize in chemistry
Current interests include: molecular self-organization and evolutionary biotechnology.
* Informal Advisor

Paul C. Lauterbur† (1929–2007)
2003 Nobel Prize in Medicine or Physiology
for magnetic resonance imaging
Current interests include: molecular imprints, origin of life, and complexity versus determinism.

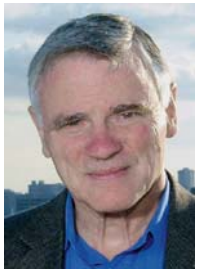

Leroy Hood
Institute for Systems Biology
1987 Lasker Award
2002 Kyoto Prize and
2003 Lemelson-MIT Prize
Current interests include: systems approaches to biology and medicine.

Pierre-Louis Lions
1994 Fields Medal
Current interests include: nonlinear partial differential equations and applications

Professor Benoit B. Mandelbrot
Sterling Professor Emeritus of Mathematical Sciences at Yale University
1993 Wolf Prize for Physics and the
2003 Japan Prize for Science and Technology
Current interests include: seeking a measure of order in physical, mathematical or social phenomena that are characterized by abundant data but wild variability.

Richard E. Stearns
1993 Turing Award for foundations of computational complexity
Current interests include: computational complexity, automata theory, analysis of algorithms, and game theory.

Mario J. Molina
1995 Nobel Prize in Chemistry for atmospheric chemistry, particularly the formation and decomposition of ozone
Current interests include: atmospheric chemical processes, and science-policy issues related to urban and regional air pollution and to global change.

Stephen Wolfram
Founder and CEO, Wolfram Research
Creator, Mathematica®
Author, *A New Kind of Science*

Thomas C. Schelling
Distinguished University Professor, University of Maryland, Emeritus
2005 Nobel Prize in Economics for understanding of conflict and cooperation through game-theory analysis
Current interests include: applications of complexity in economics and political science.

Lotfi A. Zadeh
Professor in the Graduate School,
Computer Science Division
Department of Electrical Engineering
and Computer Sciences
University of California, Berkeley

Ahmed Zewail
1999 Nobel Prize in Chemistry for his pioneering developments in the field of femtoscience
Current interests include: physical biology and complexity

Jerrold E. Marsden
Professor of Control & Dynamical Systems
California Institute of Technology

Joseph P. S. Kung
Professor
Department of Mathematics
University of North Texas

John Scott
Professor
Department of Sociology
University of Essex

William H. K. Lee
Scientist Emeritus
US Geological Survey
Menlo Park, CA 94025, USA

Steve N. Shore
Professor of Astrophysics
University of Pisa and Indiana University

Section Editors

Agent Based Modeling and Simulation

Filippo Castiglione
Research Scientist
Institute for Computing Applications (IAC) "M. Picone"
National Research Council (CNR), Italy

Cellular Automata, Mathematical Basis of

Andrew Adamatzky
Professor
Faculty of Computing, Engineering
and Mathematical Science
University of the West of England

Autonomous Robotics, Complexity and Nonlinearity in

Warren Dixon
Professor
Department of Mechanical and Aerospace
Engineering Department, University of Florida

Chaos and Complexity in Astrophysics

Steve N. Shore
Professor of Astrophysics
University of Pisa and Indiana University

Climate Modeling, Global Warming and Weather Prediction

HARTMUT GRASSL
Professor emeritus, Hamburg University
Former Director of the Max Planck Institute of Meteorology, Hamburg
Former Director World Climate Research Program 1994–1999

Complex Networks and Graph Theory

GEOFFREY CANRIGHT
Senior Research Scientist
Telenor Research and Innovation
Fornebu, Norway

Complexity in Computational Chemistry

DANAIL BONCHEV
Professor of Mathematics
Senior Fellow and Director of Research, Networks and Pathways Center for the Study of Biological Complexity
Virginia Commonwealth University

Complexity in Earthquakes, Tsunamis and Volcanoes and Forecast

WILLIAM H. K. LEE
Scientist Emeritus, US Geological Survey, Menlo Park

Computational and Theoretical Nanoscience

YONG SUK JOE
Professor and Director of Center for Computational Nanoscience
Department of Physics and Astronomy
Ball State University

Data Mining and Knowledge Discovery

PETER KOKOL
Professor
Department of Computer Science
University of Maribor, Slovenia

Ecological Complexity

Bai-Lian (Larry) Li
Professor of Ecology
University of California, Riverside

Ergodic Theory

Bryna Kra
Professor
Department of Mathematics
Northwestern University

Finance and Econometrics

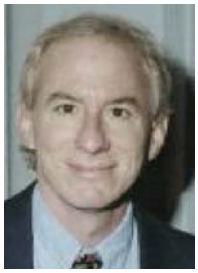

Bruce Mizrach
Professor
Department of Economics
Rutgers University

Fractals and Multifractals

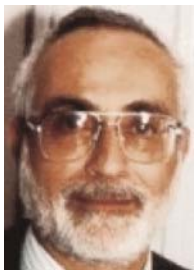

Shlomo Havlin
Professor
Department of Physics
Bar Ilan University

and

Daniel ben-Avraham
Professor
Department of Physics
Clarkson University

Game Theory

Marilda Sotomayor
Professor
Department of Economics
University of São Paulo, Brazil
Department of Economics
Brown University, Providence

Granular Computing

Tsau Y. Lin
Professor
Computer Science Department
San Jose State University

Intelligent Systems

James A. Hendler
Senior Constellation Professor of the Tetherless World
Research Constellation
Rensselaer Polytechnic Institute

Non-linear Ordinary Differential Equations and Dynamical Systems

Ferdinand Verhulst
Professor
Mathematisch Institut
University of Utrecht

Non-linear Partial Differential Equations

Italo Capuzzo Dolcetta
Professor
Dipartimento di Matematica “Guido Castelnuovo”
Università Roma La Sapienza

Percolation

Muhammad Sahimi
Professor of Chemical Engineering
and Materials Science
University of Southern California

Perturbation Theory

Giuseppe Gaeta
Professor in Mathematical Physics
Dipartimento di Matematica
Università di Milano, Italy

Probability and Statistics in Complex Systems

Henrik Jeldtoft Jensen
Professor of Mathematical Physics
Department of Mathematics and Institute for Mathematical Sciences
Imperial College London

Quantum Information Science

Joseph F. Traub
Edwin Howard Armstrong Professor of Computer Science
Computer Science Department
Columbia University

Social Network Analysis

John Scott
Professor of Sociology
School of Social Science and Law
University of Plymouth

Social Science, Physics and Mathematics Applications in

Andrzej Nowak
Director of the Center for Complex Systems
University of Warsaw
Assistant Professor, Psychology Department
Florida Atlantic University

Soft Computing

Janusz Kacprzyk
Deputy Director for Scientific Affairs, Professor
Systems Research Institute
Polish Academy of Sciences

Solitons

Mohamed A. Helal
Professor
Department of Mathematics
Faculty of Science
University of Cairo

Statistical and Nonlinear Physics

M. Cristina Marchetti
William R. Kenan, Jr. Professor of Physics
Physics Department
Syracuse University

Synergetics

Hermann Haken
Professor
Center for Synergetics
Universität Stuttgart

System Dynamics

Brian Dangerfield
Professor of Systems Modelling & Executive Editor
System Dynamics Review Centre for OR & Applied
Statistics Salford Business School
Faculty of Business, Law & the Built Environment
University of Salford

Systems and Control Theory

Matthias Kawski
Professor, Department of Mathematics and Statistics
Arizona State University

Systems Biology

Timothy P. Galitski
Associate Professor
Institute for Systems Biology
Seattle, Washington

Traffic Management, Complex Dynamics of

Boris Kerner
Head of “Traffic”
DaimlerChrysler AG

Unconventional Computing

Andrew Adamatzky
Professor
Faculty of Computing, Engineering
and Mathematical Science
University of the West of England

Wavelets

Edward Aboufadel
Professor of Mathematics
Grand Valley State University

Contributors

Abelson, Hal
Massachusetts Institute of Technology
Cambridge
USA

Abe, Sumiyoshi
Mie University
Tsu
Japan
Institut Supérieur des Matériaux et Mécaniques
Le Mans
France

Abou-Dina, Moustafa S.
Cairo University
Giza
Egypt

Aboufadel, Edward
Grand Valley State University
Allendale
USA

Abramsky, Samson
Oxford University Computing Laboratory
Oxford
UK

Achúcarro, Ana
University of Leiden
Leiden
The Netherlands
University of the Basque Country UPV-EHU
Leioa
Spain

Ackermann, Fran
University of Strathclyde
Glasgow
UK

Adamatzky, Andrew
University of the West of England
Bristol
UK

Adami, Christoph
State University of New York
Claremont
USA

Adamic, Lada A.
University of Michigan
Ann Arbor
USA

Adeoye, Mosobalaje O.
Obafemi Awolowo University
Ile-Ife
Nigeria

Adler, Pierre M.
UPMC-Sisyphe
Paris
France

Ahmed, Shaikh
Purdue University
West Lafayette
USA
Southern Illinois University
Carbondale
USA

Akin, Ethan
The City College
New York City
USA

Aktosun, Tuncay
University of Texas at Arlington
Arlington
USA

Alabau-Boussouira, Fatiha
Université de Metz
Metz
France

Alava, Mikko J.
Espoo University of Technology
Espoo
Finland

Albano, Ezequiel V.
Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA) CCT La Plata
La Plata
Argentina

Alonso-Sanz, Ramón
Universidad Politécnica de Madrid
Madrid
Spain

Alouges, François
Université Paris-Sud
Orsay cedex
France

Amos, Martyn
Manchester Metropolitan University
Manchester
UK

Andersen, David F.
University at Albany
Albany
USA

Andersen, Jørgen Vitting
Université de Nice-Sophia Antipolis
Valbonne
France

Andersen, Torben G.
Northwestern University
Evanston
USA
NBER
Cambridge
USA
CREATES
Aarhus
Denmark

Andrews, Steven S.
Lawrence Berkeley National Laboratory
Berkeley
USA

Ángeles Gil, María
University of Oviedo
Oviedo
Spain

Angelov, Plamen
Lancaster University
Lancaster
UK

Aon, Miguel Antonio
The Johns Hopkins University
Baltimore
USA

Araújo, Vitor
CMUP
Porto
Portugal
IM-UFRJ
Rio de Janeiro
Brazil

Arkin, Adam P.
Lawrence Berkeley National Laboratory
Berkeley
USA
University of California
Berkeley
USA

Arneodo, Alain
ENS-Lyon CNRS
Lyon Cedex
France

Asai, Tetsuya
Hokkaido University
Sapporo
Japan

Audit, Benjamin
ENS-Lyon CNRS
Lyon Cedex
France

Avenhaus, Rudolf
Armed Forces University Munich
Neubiberg
Germany

Avishai, Yshai
Ben-Gurion University of the Negev
Beer-Sheva
IsraelRTRA-Triangle de la Physique, LPS (Orsay) and CEA-SPHT (Saclay)
Gif sur Yvette
France

Bae, Hansang
Purdue University
West Lafayette
USA

Bagnoli, Franco
University of Florence
Florence
Italy

Balaban, Alexandru T.
Texas A&M University
Galveston
USA

Balberg, Isaac
The Hebrew University
Jerusalem
Israel

Bambusi, Dario
Università degli Studi di Milano
Milano
Italia

Bandini, Stefania
University of Milan-Bicocca
Milan
Italy

Banzhaf, Wolfgang
Memorial University of Newfoundland
St. John's
Canada

Barmin, Alexei A.
Moscow State University
Moscow
Russia

Barry, Roger G.
University of Colorado
Boulder
USA

Bascompte, Jordi
Estación Biológica de Doñana, CSIC
Seville
Spain

Batagelj, Vladimir
University of Ljubljana
Ljubljana
Slovenia

Batty, Michael
University College London
London
UK

Bays, Carter
University of South Carolina
Columbia
USA

Bazzani, Armando
Physics Department and INFN Sezione di Bologna
Bologna
Italy

Beal, Jacob
Massachusetts Institute of Technology
Cambridge
USA

Beaubouef, Theresa
Southeastern Louisiana University
Hammond
USA

Bebbington, Mark S.
Massey University
Palmerston North
New Zealand

Beer, Michael
National University of Singapore
Kent Ridge
Singapore

Behn, Mark
Woods Hole Oceanographic Institution
Woods Hole
USA

Behringer, Bob
Duke University
Durham
USA

Bellemans, T.
Hasselt University
Diepenbeek
Belgium

ben-Avraham, Daniel
Clarkson University
Potsdam
USA

Benedetto, D.
Dipartimento di Matematica, Università di Roma 'La Sapienza'
Roma
Italy

Beni, Gerardo
University of California Riverside
Riverside
USA

Ben-Zion, Yehuda
University of Southern California
Los Angeles
USA

Benzoni, Luca
Federal Reserve Bank of Chicago
Chicago
USA

Berger, Mitchell A.
University of Exeter
Devon
UK
UCL
London
UK

Bergström, Lennart
Stockholm University
Stockholm
Sweden

Berkhin, Pavel
eBay Inc.
San Jose
USA

Berman, Abraham
Technion – Israel Institute of Technology
Haifa
Israel

Berthier, Ludovic
Université Montpellier II and CNRS
Montpellier
France

Bertoni, Andrea
CNR-INFM National Research Center on NanoStructures and BioSystems at Surfaces (S3)
Modena
Italy

Bestehorn, Michael
Brandenburg University of Technology
Cottbus
Germany

Bialek, William
Princeton University
Princeton
USA

Binder, Kurt
Johannes Gutenberg Universität
Mainz
Germany

Biroli, Giulio
IPhT, CNRS, MPPU, URA2306, Saclay
Gif-sur-Yvette
France

Bjelland, Johannes
Telenor R&I
Fornebu
Norway

Bjorklund, David F.
Florida Atlantic University
Boca Raton
USA

Boccara, Nino
University of Illinois
Chicago
USA
CE Saclay
Gif-sur-Yvette
France

Bodin, Örjan
Stockholm University
Stockholm
Sweden

Bonča, Janez
Jožef Stefan Institute
Ljubljana
Slovenia
University of Ljubljana
Ljubljana
Slovenia

Bonchev, Danail G.
Virginia Commonwealth University
Richmond
USA

Borgatti, Stephen P.
University of Kentucky
Lexington
USA

Bormann, Peter
GeoForschungsZentrum Potsdam
Potsdam
Germany

Boykin, Timothy B.
The University of Alabama
Huntsville
USA

Bracci, Luciano
Università di Pisa
Pisa
Italy
Sezione di Pisa
Pisa
Italy

Brameier, Markus
University of Aarhus
Århus
Denmark

Brams, Steven J.
New York University
New York
USA

Brandenburg, Axel
AlbaNova University Center
Stockholm
Sweden

Breier, Georg
Technische Universität Dresden
Dresden
Germany

Bressan, Alberto
Penn State University
University Park
USA

Brodie of Brodie, Edward-Benedict
ENS-Lyon CNRS
Lyon Cedex
France

Broer, Henk W.
University of Groningen
Groningen
The Netherlands

Brooks, Benjamin A.
University of Hawaii
Honolulu
USA

Browne, Will
The University of Reading
Reading
UK

Brunel, Nicolas
Université Paris Descartes
Paris
France
CNRS
Paris
France

Buldyrev, Sergey V.
Yeshiva University
New York
USA

Bullo, Francesco
University of California
Santa Barbara
USA

Bunde, Armin
Justus-Liebig-Universität
Giessen
Germany

Buzzi, Jérôme
C.N.R.S. and Université Paris-Sud
Orsay
France

Byrne, Raymond H.
Sandia National Laboratories
Albuquerque
USA

Camps, Octavia
Northeastern University
Boston
USA

Cannarsa, Piermarco
Università di Roma “Tor Vergata”
Rome
Italy

Canright, Geoffrey
Telenor R&I
Fornebu
Norway

Canty, Morton J.
Forschungszentrum Jülich
Jülich
Germany

Carbó-Dorca, Ramon
Universitat de Girona
Girona
Spain

Carbone, Vincenzo
Università della Calabria
Arcavacata di Rende
Italy

Carrión-Vázquez, Mariano
CSIC & CIBERNED (Network on Degenerative Deseases)
Madrid
Spain

Carter, Gregory W.
Institute for Systems Biology
Seattle
USA

Casati, Giulio
Universita dell'Insubria
Como
Italy

Castellano, Claudio
"Sapienza" Università di Roma
Roma
Italy

Castiglione, Filippo
Institute for Computing Applications (IAC) – National Research Council (CNR)
Rome
Italy

Castillo, Oscar
Tijuana Institute of Technology
Tijuana
Mexico

Castro Neto, Antonio H.
Boston University
Boston
USA

Caulfield, H. John
Fisk University
Nashville
USA

Caverlee, James
Texas A&M University
College Station
USA

Ceccherini-Silberstein, Tullio
Università del Sannio
Benevento
Italy

Celletti, Alessandra
Università di Roma Tor Vergata
Roma
Italy

Cenek, Martin
Portland State University
Portland
USA

Cercone, Nick
York University
Toronto
Canada

Cervelle, Julien
Université Paris-Est
Marne la Vallée
France

Chaisson, Eric J.
Tufts University
Massachusetts
USA

Chakraborty, Bulbul
Brandeis University
Waltham
USA

Champneys, Alan
University of Bristol
Bristol
UK

Chang, Kow-Lung
National Taiwan University
Taipeh
Taiwan

Chang, Tom
Massachusetts Institute of Technology
Cambridge
USA

Chan, Tony F.
University of California
Los Angeles
USA

Chatterjee, Kalyan
The Pennsylvania State University
University Park
USA

Chazottes, Jean-René
CNRS/École Polytechnique
Palaiseau
France

Chechkin, Aleksei V.
Institute for Theoretical Physics NSC KIPT
Kharkov
Ukraine

Chen, Chun-Hao
National Cheng–Kung University
Tainan
Taiwan

Chen, Guanrong
City University of Hong Kong
Hong Kong
China

Chen, Zhengxin
University of Nebraska at Omaha
Omaha
USA

Chicone, Carmen
University of Missouri-Columbia
Columbia
USA

Chielens, Klaas
Vrije Universiteit Brussel
Brussels
Belgium

Chierchia, Luigi
Università "Roma Tre"
Roma
Italy

Chirikjian, Gregory S.
Johns Hopkins University
Baltimore
USA

Chopard, Bastien
University of Geneva
Geneva
Switzerland

Chouet, Bernard
US Geological Survey
Menlo Park
USA

Chung, Fan
University of California
San Diego
USA

Cieplak, Marek
Polish Academy of Sciences
Warsaw
Poland

Clarke, Francis
Institut universitaire de France et Université de Lyon
Lyon
France

Clark, Steve
Purdue University
West Lafayette
USA

Cohen, Reuven
Bar-Ilan University
Ramat-Gan
Israel

Coniglio, Antonio
Università di Napoli "Federico II", Complesso
Universitario di Monte Sant'Angelo
Naples
Italy

Conte, Rosaria
CNR
Rome
Italy

Contopoulos, George
Research Center for Astronomy
Athens
Greece

Convertito, Vincenzo
Istituto Nazionale di Geofisica e Vulcanologia
(RISSC-Lab)
Napoli
Italy

Coornaert, Michel
Université Louis Pasteur et CNRS
Strasbourg
France

Corchón, Luis C.
Universidad Carlos III
Madrid
Spain

Cortassa, Sonia
The Johns Hopkins University
Baltimore
USA

Cortés, Jorge
University of California
San Diego
USA

Costa, Antonio
University of Bristol
Bristol
UK
Istituto Nazionale di Geofisica e Vulcanologia
Naples
Italy

Cpałka, Krzysztof
Częstochowa University of Technology
Częstochowa
Poland
Academy of Humanities and Economics
Lodz
Poland

Crane III, Carl D.
University of Florida
Gainesville
USA

Creutz, Michael
Brookhaven National Laboratory
Upton
USA

Crilly, Tony
Middlesex University
London
UK

Culcer, Dimitrie
Argonne National Laboratory
Argonne
USA
Northern Illinois University
De Kalb
USA

Curtis, Andrew
The University of Edinburgh
Edinburgh
UK

Dahmen, Karin A.
University of Illinois at Urbana-Champaign
Urbana
USA

D'Ancona, Piero
Unversità di Roma "La Sapienza"
Roma
Italy

Dangerfield, Brian
University of Salford
Salford
UK

Danilenko, Alexandre I.
Ukrainian National Academy of Sciences
Kharkov
Ukraine

Da Prato, Giuseppe
Scuola Normale Superiore
Pisa
Italy

Dardzinska, Agnieszka
Białystok Technical University
Białystok
Poland

d'Aubenton-Carafa, Yves
CNRS
Gif-sur-Yvette
France

D'Auria, Luca
Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Napoli
Naples
Italy

DAVIDSSON, PAUL
Blekinge Institute of Technology
Ronneby
Sweden

DAWSON, KENNETH A.
University College Dublin
Dublin
Ireland

DEBNATH, LOKENATH
University of Texas – Pan American
Edinburg
USA

DE GREGORIO, PAOLO
Cornell University
Ithaca
USA

DE LA RUE, THIERRY
CNRS – Université de Rouen
Saint Étienne du Rouvray
France

D'EMILIO, EMILIO
Università di Pisa
Pisa
Italy

DEL JUNCO, ANDRÉS
University of Toronto
Toronto
Canada

DEMURJIAN, STEVEN A.
The University of Connecticut
Storrs
USA

DENNUNZIO, ALBERTO
Università degli Studi di Milano-Bicocca
Milan
Italy

DE NOOY, WOUTER
University of Amsterdam
Amsterdam
The Netherlands

DERCOLE, FABIO
Politecnico di Milano
Milano
Italy

DERKS, GIANNE
University of Surrey
Guildford
UK

DE SCHUTTER, B.
TU Delft
Delft
The Netherlands

DE SILVA, CLARENCE W.
University of British Columbia
Vancouver
Canada

DESIMONE, ANTONIO
SISSA-International School for Advanced Studies
Trieste
Italy

DEUTSCH, ANDREAS
Technische Universität Dresden
Dresden
Germany

DEUTSCHBAUER, ADAM M.
Lawrence Berkeley National Laboratory
Berkeley
USA

DHARMA-WARDANA, M.W. CHANDRE
National Research Council of Canada
Ottawa
Canada

DHILLON, INDERJIT S.
University of Texas
Austin
USA

DIKS, CEES
University of Amsterdam
Amsterdam
The Netherlands

DING, BAOQUAN
Lawrence Berkeley National Lab
Berkeley
USA

DINH, TUAN
Lawrence Berkeley National Laboratory
Berkeley
USA

DITTRICH, PETER
Friedrich Schiller University Jena
Jena
Germany

DIXON, WARREN E.
University of Florida
Gainesville
USA

DOLCETTA, ITALO CAPUZZO
Sapienza Universita' di Roma
Rome
Italy

DOMOTOR, ZOLTAN
University of Pennsylvania
Philadelphia
USA

DOREIAN, PATRICK
University of Pittsburgh
Pittsburgh
USA

DORFMAN, J. ROBERT
University of Maryland
College Park
USA

DOROGOVTSEV, SERGEY N.
Universidade de Aveiro
Aveiro
Portugal
A. F. Ioffe Physico-Technical Institute
St. Petersburg
Russia

DRAPER, DAVID
University of California
Santa Cruz
USA

DRENNER, ANDREW
University of Minnesota
Minneapolis
USA

DUBEY, ATUL
Rutgers, The State University of New Jersey
Piscataway
USA

DUBOIS, DIDIER
Universite Paul Sabatier
Toulouse Cedex
France

DUDLEY, AIMÉE M.
Institute for Systems Biology
Seattle
USA

DUFTY, JAMES W.
University of Florida
Gainesville
USA

DUNNE, JENNIFER A.
Santa Fe Institute
Santa Fe
USA
Pacific Ecoinformatics and Computational Ecology Lab
Berkeley
USA

DURAND-LOSE, JÉRÔME
Université d'Orléans
Orléans
France

DUTTA, PRAJIT K.
Columbia University
New York
USA

DUXBURY, PHILLIP M.
Michigan State University
East Lansing
USA

DŽEROSKI, SAŠO
Jožef Stefan Institute
Ljubljana
Slovenia

DZHEPAROV, FRIDRIKH S.
Institute for Theoretical and Experimental Physics
Moscow
Russia

DZWINEL, WITOLD
AGH University of Sci. and Technol.
Kraków
Poland

EDDY, JAMES A.
University of Illinois
Urbana-Champaign
USA

EDELMAN, GERALD M.
The Neurosciences Institute
San Diego
USA

EDEN, COLIN
University of Strathclyde
Glasgow
UK

ELIA, LUCA
Istituto Nazionale di Geofisica e Vulcanologia
(RISSC-Lab)
Napoli
Italy

ENGØ-MONSEN, KENTH
Telenor R&I
Fornebu
Norway

ERGÜN, GÜLER
University of Bath
Bath
UK

ESCANCIANO, JUAN-CARLOS
Indiana University
Bloomington
USA

ESCRIBANO, ALVARO
Universidad Carlos III de Madrid
Madrid
Spain

FADILI, JALAL
Ecole Nationale Supèrieure d'Ingènieurs de Caen
Caen Cedex
France

FALOUTSOS, MICHALIS
University of California
Riverside
USA

FEDDEMA, JOHN T.
Sandia National Laboratories
Albuquerque
USA

FEIL-SEIFER, DAVID
University of Southern California
Los Angeles
USA

FIERRO, ANNALISA
Università di Napoli "Federico II"
Naples
Italy

FOGEDBY, HANS C.
University of Aarhus
Aarhus
Denmark
Niels Bohr Institute
Copenhagen
Denmark

FORD, ANDREW
Washington State University, Pullman
Washington
USA

FORGACS, GABOR
University of Missouri
Columbia
USA

FORGES, FRANÇOISE
Université Paris-Dauphine
Paris
France

FORMENTI, ENRICO
Université de Nice Sophia Antipolis
Sophia Antipolis
France

FORTUNATO, SANTO
ISI Foundation
Torino
Italy

FOSTER, JAMES H.
University of Hawaii
Honolulu
USA

FRANK, OVE
Stockholm University
Stockholm
Sweden

FRANK, TILL D.
University of Connecticut
Storrs
USA

FRANTZIKINAKIS, NIKOS
University of Memphis
Memphis
USA

FREEMAN, LINTON C.
University of California
Irvine
USA

FRENKEL, ANATOLY I.
Yeshiva University
New York
USA

FRIEDRICH, RUDOLF
University of Münster
Münster
Germany

FUCHS, ARMIN
Florida Atlantic University
Boca Raton
USA

GABRIELOV, ANDREI
Purdue University
West Lafayette
USA

GAETA, GIUSEPPE
Università di Milano
Milan
Italy

GALDI, GIOVANNI P.
University of Pittsburgh
Pittsburgh
USA

GALLAVOTTI, GIOVANNI
Università di Roma I "La Sapienza"
Roma
Italy

GALLEGATI, MAURO
Università Politecnica delle Marche
Ancona
Italy

GALLEGOS, ANA
Universitat de Girona
Girona
Spain

GALLOS, LAZAROS K.
City College of New York
New York
USA

GANS, NICHOLAS
University of Florida
Gainesville
USA

GARCES, MILTON
HIGP, SOEST, University of Hawaii, Manoa
Kailua–Kona
USA

GARDINER, AMY K.
Florida Atlantic University
Boca Raton
USA

GARTNER, NATHAN H.
University of Massachusetts
Lowell
USA

GASPARINI, PAOLO
Università di Napoli "Federico II" (RISSC-Lab)
Napoli
Italy

GAUTHIER, JEAN-PAUL ANDRÉ
University of Toulon
Toulon
France

GENTILE, GUIDO
Università di Roma Tre
Roma
Italy

GEORGANTZAS, NICHOLAS C.
Fordham University Business Schools
New York
USA

GHOSAL, SANDIP
Northwestern University
Evanston
USA

GIAMARCHI, THIERRY
University of Geneva
Geneva
Switzerland

GIORGINI, BRUNO
Physics Department and INFN Sezione di Bologna
Bologna
Italy

GLASS, OLIVIER
Université Pierre et Marie Curie
Paris
France

GLUSMAN, GUSTAVO
Institute for Systems Biology
Seattle
USA

GOEBEL, RAFAL
Loyola University
Chicago
USA

GOLDENBERG, JACOB
Hebrew University
Jerusalem
Israel

GONZÁLEZ-RIVERA, GLORIA
University of California
Riverside
USA

GORECKA, JOANNA NATALIA
Polish Academy of Science
Warsaw
Poland

GORECKI, JERZY
Polish Academy of Science
Warsaw
Poland
Cardinal Stefan Wyszynski University
Warsaw
Poland

GOSSNER, OLIVIER
Northwestern University
Paris
France

GOUDIE, DOUGLAS
James Cook University
Townsville
Australia

GOULIAS, KONSTADINOS G.
University of California Santa Barbara
Santa Barbara
USA

GRAMCHEV, TODOR
Università di Cagliari
Cagliari
Italy

GRASMAN, JOHAN
Wageningen University and Research Centre
Wageningen
The Netherlands

GRASSL, HARTMUT
Max Planck Institute for Meteorology
Hamburg
Germany

GRASSL, MARKUS
Austrian Academy of Sciences
Innsbruck
Austria

GRAVNER, JANKO
University of California
Davis
USA

GRECO, SALVATORE
University of Catania
Catania
Italy

GREEN, KEITH E.
Clemson University
Clemson
USA

GROESSER, STEFAN
University of St. Gallen
St. Gallen
Switzerland

GRZYMALA-BUSSE, JERZY W.
University of Kansas
Lawrence
USA
Polish Academy of Sciences
Warsaw
Poland

HAAS, MARKUS
University of Munich
Munich
Germany

HADJIDEMETRIOU, JOHN D.
University of Thessaloniki
Thessaloniki
Greece

HAFNER, CHRISTIAN M.
Université catholique de Louvain
Louvain-la-Neuve
Belgium

HAINZL, SEBASTIAN
GFZ German Research Centre for Geosciences
Potsdam
Germany

HAKEN, HERMANN
Universität Stuttgart
Stuttgart
Germany

HAKIM, VINCENT
CNRS
Paris
France

HALEY, BENJAMIN
Purdue University
West Lafayette
USA

HALL, WENDY
University of Southampton
Southampton
UK

HAN, BIN
University of Alberta
Edmonton
Canada

HÄNGGI, PETER
University of Augsburg
Augsburg
Germany

HAN, JIANCHAO
California State University
Dominguez Hills, Carson
USA

HANSON, JAMES E.
IBM T.J. Watson Research Center
Yorktown Heights
USA

HANSSMANN, HEINZ
Universiteit Utrecht
Utrecht
The Netherlands

HARDING, SIMON
Memorial University
St. John's
Canada

HATZIKIROU, HARALAMBOS
Technische Universität Dresden
Dresden
Germany

HAUPTMAN, AMI
Ben-Gurion University
Beer-Sheva
Israel

HAUPTMANN, CHRISTIAN
Research Center Jülich
Jülich
Germany

HAVLIN, SHLOMO
Bar-Ilan University
Ramat-Gan
Israel

HEDIN, ERIC R.
Ball State University
Muncie
USA

HEGSELMANN, RAINER
Bayreuth University
Bayreuth
Germany

HEGYI, A.
TU Delft
Delft
The Netherlands

HE, JIANGHENG
Geological Survey of Canada
Sidney
Canada

HE, JI-HUAN
Donghua University
Shanghai
China

HE, YIHUA
University of California
Riverside
USA

HELAL, MOHAMED A.
Cairo University
Giza
Egypt

HELBING, DIRK
ETH Zurich
Zurich
Switzerland
Collegium Budapest
Budapest
Hungary

HENDLER, JAMES
Rensselaer Polytechnic Institute
Troy
USA

HEREMAN, WILLY
Colorado School of Mines
Golden
USA

HEYLIGHEN, FRANCIS
Vrije Universiteit Brussel
Brussels
Belgium

HILL, DAVID P.
Volcano Hazards Program
Menlo Park
USA

HIRANO, SHOJI
Shimane University, School of Medicine
Enya-cho Izumo City, Shimane
Japan

HIRSCH, GARY
Independent Consultant
Wayland
USA

HIRSHORN, BARRY
NOAA/NWS/Pacific Tsunami Warning Center
Ewa Beach
USA

HOFFMAN, CHRISTOPHER
University of Washington
Seattle
USA

HO, KAI-MING
Iowa State University
Ames
USA

HOLLIDAY, JAMES R.
University of California
Davis
USA

HOLSCHNEIDER, MATTHIAS
University of Potsdam
Potsdam
Germany

HOLTSLAG, ALBERT A. M.
Wageningen University
Wageningen
The Netherlands

HOMER, JACK
Independent Consultant
Voorhees
USA

HONG, TZUNG-PEI
National University of Kaohsiung
Kaohsiung
Taiwan

HOWICK, SUSAN
University of Strathclyde
Glasgow
UK

HRYNIEWICZ, OLGIERD
Systems Research Institute
Warsaw
Poland

HUANG, SUI
Department of Biological Sciences, University of Calgary
Calgary
Canada

HUBER, DAVID E.
University of California
San Diego
USA

HUDRY, OLIVIER
École Nationale Supérieure des Télécommunications
Paris
France

HUGHES, BARRY D.
University of Melbourne
Melbourne
Australia

HU, GUOQIANG
University of Florida
Gainesville
USA

HUVET, MAXIME
CNRS
Gif-sur-Yvette
France

HU, YAN
University of Victoria
Victoria
Canada

IANNACCONE, GIOVANNI
Istituto Nazionale di Geofisica e Vulcanologia
(RISSC-Lab)
Napoli
Italy

ICHIHARA, MIE
University of Tokyo
Tokyo
Japan

IERVOLINO, IUNIO
Università di Napoli "Federico II"
Napoli
Italy

IGEL, HEINER
Ludwig-Maximilians-University
Munich
Germany

IIO, YOSHIHISA
Kyoto University
Kyoto
Japan

ILACHINSKI, ANDREW
Center for Naval Analyses
Alexandria
USA

INC, MUSTAFA
Fırat University
Elazığ
Turkey

ISIDORI, ALBERTO
University of Rome
La Sapienza
Italy

IVANCIUC, OVIDIU
University of Texas, Medical Branch
Galveston
USA

JACOB, DANIELA
Max-Planck-Institute for Meteorology
Hamburg
Germany

JAMES, MICHAEL R.
Lancaster University
Lancaster
UK

JAMSHIDI, NEEMA
University of California
San Diego, La Jolla
USA

JANEŽIČ, DUŠANKA
National Institute of Chemistry
Ljubljana
Slovenia

JANSSEN, MICHAEL
University of Minnesota
Minneapolis
USA

JARRAH, ABDUL S.
Virginia Polytechnic Institute and State University
Virginia
USA

JENSEN, ARNE
Aalborg University
Aalborg East
Denmark

JENSEN, HENRIK JELDTOFT
Imperial College London
London
UK

JOE, YONG S.
Ball State University
Muncie
USA

JOHANSSON, ANDERS
ETH Zurich
Zurich
Switzerland

JOHN, ROBERT I.
De Montfort University
Leicester
UK

JOHNSON, DUANE D.
University of Illinois at Urbana-Champaign
Urbana
USA

JORGENSEN, PALLE E. T.
The University of Iowa
Iowa City
USA

JURDJEVIC, VELIMIR
University of Toronto
Toronto
Canada

KACPRZYK, JANUSZ
Polish Academy of Sciences
Warsaw
Poland

KADLEC, BEN
University of Colorado
Boulder
USA

KAMIGAICHI, OSAMU
Japan Meteorological Agency
Tokyo
Japan

KAMOGAWA, MASASHI
Tokyo Gakugei University
Koganei-shi
Japan

KAMPMANN, CHRISTIAN ERIK
Copenhagen Business School
Copenhagen
Denmark

KAMSTRA, MARK J.
York University
Toronto
Canada

KANAMORI, H.
Caltech
Pasadena
USA

KANTELHARDT, JAN W.
Martin-Luther-University Halle-Wittenberg
Halle
Germany

KARI, JARKKO
University of Turku
Turku
Finland

KARLSSON, ANETTE M.
University of Delaware
Newark
USA

KÄSER, MARTIN
Ludwig-Maximilians-University
Munich
Germany

KAUFFMAN, STUART A.
Department of Biological Sciences, University of Calgary
Calgary
Canada

KAWAMURA, KAZUHIKO
Vanderbilt University
Nashville
USA

KAWSKI, MATTHIAS
Department of Mathematics, Arizona State University
Tempe
USA

KAYA, DOĞAN
Firat University
Elazig
Turkey

KEILIS-BOROK, VLADIMIR
University of California
Los Angeles
USA
Russian Academy of Sciences
Moscow
Russia

KEINERT, FRITZ
Iowa State University
Ames
USA

KELLER, GERHARD
Universität Erlangen-Nürnberg
Erlangen
Germany

KELSO, JAMES A. S.
Florida Atlantic University
Boca Raton
USA

KENDON, VIV
University of Leeds
Leeds
UK

KENNEL, RICK
Purdue University
West Lafayette
USA

KERNER, BORIS S.
Daimler AG
Sindelfingen
Germany

KHARCHE, NEERAV
Purdue University
West Lafayette
USA

KIER, LEMONT B.
Virginia Commonwealth University
Richmond
USA

KIKOIN, KONSTANTIN
Tel-Aviv University
Tel-Aviv
Israel

KING, JONATHAN L. F.
University of Florida
Gainesville
USA

KING, PETER
Imperial College London
London
UK

KLAFTER, JOSEPH
Tel Aviv University
Tel Aviv
Israel
University of Freiburg
Freiburg
Germany

KLEINBOCK, DMITRY
Brandeis University
Waltham
USA

KLENOV, SERGEY L.
Moscow Institute of Physics and Technology
Dolgoprudny
Russia

KLIMECK, GERHARD
Purdue University
West Lafayette
USA
California Institute of Technology
Pasadena
USA

KLINGSCH, WOLFRAM
University of Wuppertal
Wuppertal
Germany

KLÜPFEL, HUBERT
TraffGo HT GmbH
Duisburg
Germany

KLÜVER, JÜRGEN
Duisburg-Essen University
Essen
Germany

KNACKSTEDT, MARK
Australian National University
Canberra
Australia

KOIJEN, RALPH S. J.
Tilburg University
Tilburg
The Netherlands

KOIKE, SHIGEAKI
Saitama University
Saitama
Japan

KOKALJ, ANTON
Jožef Stefan Institute
Ljubljana
Slovenia

KOKOL, PETER
University of Maribor
Maribor
Slovenia

KOLNER, BRIAN H.
University of California
Davis
USA

Korenok, Oleg
VCU School of Business
Richmond
USA

Korkusinski, Marek
National Research Council of Canada
Ottawa
Canada

Koster, Maurice
University of Amsterdam
Amsterdam
Netherlands

Kra, Bryna
Northwestern University
Evanston
USA

Kramer, Lisa A.
University of Toronto
Toronto
Canada

Krawitz, Peter
Ludwig Maximilians Universität
München
Germany

Kretz, Tobias
PTV Planung Transport Verkehr AG
Karlsruhe
Germany

Kumagai, Hiroyuki
National Research Institute for Earth Science and Disaster Prevention
Tsukuba
Japan
IAVCEI/IASPEI Joint Commission on Volcano Seismology
Tsukuba
Japan

Kůrka, Petr
Université de Nice Sophia Antipolis
Nice
France
Academy of Sciences and Charles University
Prague
Czechia

Kuter, Ugur
University of Maryland
College Park
USA

Kutrib, Martin
Universität Giessen
Giessen
Germany

Laaksonen, Aatto
Stockholm University
Stockholm
Sweden

La Cour–Harbo, Anders
Aalborg University
Aalborg East
Denmark

Lai, Ming-Jun
The University of Georgia
Athens
USA

Lakshmanan, M.
Bharathidasan University
Tiruchirapalli
India

Lancieri, Maria
Istituto Nazionale di Geofisica e Vulcanologia (RISSC-Lab)
Napoli
Italy

Lane, Stephen J.
Lancaster University
Lancaster
UK

Lange, Wolfgang
University of Sussex
Brighton
UK

Laubenbacher, Reinhard
Virginia Polytechnic Institute and State University
Virginia
USA

Lauritsen, Kent Bækgaard
Danish Meteorological Institute
Copenhagen
Denmark

Lavi, Ron
The Technion – Israel Institute of Technology
Haifa
Israel

Lawlor, Aonghus
University College Dublin
Dublin
Ireland

Leburton, Jean-Pierre
University of Illinois
Urbana
USA

Lee, Gun-Do
Seoul National University
Seoul
Korea

Lee, Sunhee
Purdue University
West Lafayette
USA

Lee, Tae-Hwy
University of California
Riverside
USA

Lee, Taeyoung
University of Michigan
Ann Arbor
USA

Lee, William H. K.
US Geological Survey
Menlo Park
USA

Lefebvre, Aline
Université Paris-Sud
Orsay cedex
France

Lemańczyk, Mariusz
Nicolaus Copernicus University
Toruń
Poland

Leok, Melvin
Purdue University
West Lafayette
USA

Le Pichon, Alexis
CEA/DASE/LD
Bruyères-le-Châtel
France

Lepreti, Fabio
Università della Calabria
Arcavacata di Rende
Italy

Levy, Moshe
The Hebrew University
Jerusalem
Israel

Lewis, Frank L.
University of Texas at Arlington
Fort Worth
USA

Lewis, Nathan E.
University of California
San Diego, La Jolla
USA

Liau, Churn-Jung
Academia Sinica
Taipei
Taiwan

Lichtman, Allan
American University
Washington D.C.
USA

Liljeros, Fredrik
Stockholm University
Stockholm
Sweden

Lin, Tsau Young
San Jose State University
San Jose
USA

Liu, Huan
Arizona State University
Tempe
USA

Liu, Yan
Arizona State University
Tempe
USA

Lloyd, Seth
MIT
Cambridge
USA

Lodwick, Weldon A.
University of Colorado Denver
Denver
USA

Lohmann, Gerrit
Alfred Wegener Institute for Polar and Marine Research
Bremerhaven
Germany

Lohman, Rowena
Cornell University
Ithaca
USA

Lo, Hoi-Kwong
University of Toronto
Toronto
Canada

Lomax, Anthony
ALomax Scientific
Mouans-Sartoux
France

Lyneis, James M.
Worcester Polytechnic Institute
Worcester
USA

Lynett, Patrick J.
Texas A&M University
College Station
USA

Maani, Kambiz
The University of Queensland
Brisbane
Australia

Macal, Charles M.
Center for Complex Adaptive Agent Systems Simulation
(CAS^2)
Argonne
USA

Macdonald, Joanne
Columbia University
New York
USA

MacDonald, Roderick
University at Albany
Albany
USA

Macho-Stadler, Inés
Universitat Autònoma de Barcelona
Barcelona
Spain

MacLennan, Bruce J.
University of Tennessee
Knoxville
USA

Mac Namee, Brian
Dublin Institute of Technology
Dublin
Ireland

Madariaga, Raul
Laboratoire de Géologie
Paris
France

Maier, Frank H.
International University in Germany
Bruchsal
Germany

Mailath, George J.
University of Pennsylvania
Philadelphia
USA

Mai, P. Martin
Institute of Geophysics, ETH
Zürich
Switzerland

Makino, Junichiro
National Astronomical Observatory of Japan
Tokyo
Japan

Makse, Hernán A.
City College of New York
New York
USA

Mange, Daniel
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Lausanne
Switzerland

Manzan, Sebastiano
Baruch College CUNY
New York
USA

Manzoni, Sara
University of Milan-Bicocca
Milan
Italy

Marchetti, M. Cristina
Syracuse University
Syracuse
USA

Marconi, Lorenzo
University of Bologna
Bologna
Italy

Marcus, Brian
University of British Columbia
Vancouver
Canada

Margenstern, Maurice
Université Paul Verlaine
Metz
France

Marin, François
Laboratoire Ondes et Milieux Complexes, Fre CNRS 3102
Le Havre Cedex
France

Marion, Russ
Clemson University
Clemson
USA

Markellos, Raphael N.
Loughborough University
Loughborough
UK
Athens University of Economics and Business
Athens
Greece

Marsden, Jerrold E.
California Institute of Technology
Pasadena
USA

Martínez, Sonia
University of California
San Diego
USA

Martini, Marcello
Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Napoli
Naples
Italy

Martino, Claudio
Università di Napoli "Federico II" (RISSC-Lab)
Napoli
Italy

Martins, Carlos J. A. P.
Universidade do Porto
Centro de Física do Porto
Porto
Portugal
University of Cambridge
Cambridge
UK

Masihi, Mohsen
Sharif University of Technology
Tehran
Iran

Matarazzo, Benedetto
University of Catania
Catania
Italy

Matarić, Maja J.
University of Southern California
Los Angeles
USA

Mathey, Ludwig
Harvard University
Cambridge
USA

Maurer, Brian A.
Michigan State University
East Lansing
USA

Mawhin, Jean
Université Catholique de Louvain
Maryland
USA

McClamroch, Harris
University of Michigan
Ann Arbor
USA

McCutcheon, Randall
University of Memphis
Memphis
USA

McGuire, Jeffrey J.
Woods Hole Oceanographic Institution
Woods Hole
USA

McKane, Alan J.
University of Manchester
Manchester
UK

McLeish, T.C.B.
University of Leeds
Leeds
UK

McLennan, Michael
Purdue University
West Lafayette
USA

Meijer, Hil
University of Twente
Enschede
The Netherlands

Melin, Patricia
Tijuana Institute of Technology
Tijuana
Mexico

Melnik, Oleg
Moscow State University
Moscow
Russia
University of Bristol
Bristol
UK

Melnikov, Dmitriy V.
University of Illinois
Urbana
USA

Melnik, Roderick
Wilfrid Laurier University
Waterloo
Canada
University of Waterloo
Waterloo
Canada

Melrose, Don B.
University of Sydney
Sydney
Australia

Menasalvas, Ernestina
Facultad de Informatica
Madrid
Spain

Mendel, Jerry M.
University of Southern California
Los Angeles
USA

Metzler, Ralf
Technical University of Munich
Garching
Germany

Meyers, Robert A.
Ramtech Limited
Larkspur
USA

Michelini, Alberto
Istituto Nazionale di Geofisica e Vulcanologia
Roma
Italy

Milburn, Gerard J.
The University of Queensland
Brisbane
Australia

Miličević, Ante
The Institute of Medical Research and Occupational Health
Zagreb
Croatia

Miller, Christian S.
Board of Governors of the Federal Reserve System
Washington DC
USA

Miller, Julian F.
University of York
Heslington
UK

Milling, Peter M.
Mannheim University
Mannheim
Germany

MILLS, TERENCE C.
Loughborough University
Loughborough
UK

MITCHELL, MELANIE
Portland State University
Portland
USA

MIZRACH, BRUCE
Rutgers University
New Brunswick
USA

MORITA, KENICHI
Hiroshima University
Higashi-Hiroshima
Japan

MORLEY, JAMES
Washington University
St. Louis
USA

MORTVEIT, HENNING S.
Virginia Polytechnic Institute and State University
Virginia
USA

MOSCA, MICHELE
University of Waterloo
Waterloo
Canada
St. Jerome's University
Waterloo
Canada
Perimeter Institute for Theoretical Physics
Waterloo
Canada

MOSES, MATTHEW
Johns Hopkins University
Baltimore
USA

MOSS, SCOTT
Manchester Metropolitan University Business School
Manchester
UK

MOURZENKO, VALERI V.
CNRS-LCD
Chasseneuil du Poitou
France

MURRAY-MORALEDA, JESSICA
US Geological Survey
Menlo Park
USA

MUTO, SHIGEO
Institute of Technology
Tokyo
Japan

MUZY, ALEXANDRE
Università di Corsica
Corte
France

NACHBAR, JOHN
Washington University
St. Louis
USA

NAGAO, TOSHIYASU
Tokai University
Shizuoka
Japan

NAKAYAMA, TSUNEYOSHI
Toyota Physical and Chemical Research Institute
Nagakute
Japan

NAUGHTON, THOMAS J.
National University of Ireland
Maynooth County Kildare
Ireland
University of Oulu, RFMedia Laboratory
Ylivieska
Finland

NAUMOV, MAXIM
Purdue University
West Lafayette
USA

NEELY, CHRISTOPHER J.
Federal Reserve Bank of St. Louis
St. Louis
USA

NEWMAN, LENORE LAURI
Royal Roads University
Victoria
Canada

NEWMAN, STUART A.
New York Medical College
Valhalla
USA

NICODEMI, MARIO
University of Warwick
Coventry
UK

NICOLAY, SAMUEL
Université de Liège
Liège
Belgium

NICOL, MATTHEW
University of Houston
Houston
USA

NIEDERMAN, LAURENT
Université Paris
Paris
France
IMCCE
Paris
France

NIGRO, GIUSY
Università della Calabria
Arcavacata di Rende
Italy

NIKOLIĆ, SONJA
The Rugjer Bošković Institute
Zagreb
Croatia

NILSSON JACOBI, MARTIN
Chalmers University of Technology
Gothenburg
Sweden

NISHIMURA, TAKESHI
Tohoku University
Sendai
Japan

NIȚICĂ, VIOREL
West Chester University
West Chester
USA
Institute of Mathematics
Bucharest
Romania

NITTA, KOH-HEI
Graduate School of Natural Science and Technology,
Kanazawa University
Kanazawa
Japan

NOLFI, STEFANO
National Research Council (CNR)
Rome
Italy

NORTH, MICHAEL J.
Center for Complex Adaptive Agent Systems Simulation
(CAS^2)
Argonne
USA

NOTTALE, LAURENT
Paris Observatory and Paris Diderot University
Paris
France

NOWAK, ANDRZEJ
University of Warsaw
Warsaw
Poland

NOWICKI, ROBERT
Częstochowa University of Technology
Częstochowa
Poland

NUZZO, RALPH G.
University of Illinois at Urbana-Champaign
Urbana
USA

OBERHAUSER, ANDRÉS F.
University of Texas Medical Branch
Galveston
USA

O'HARA, KIERON
University of Southampton
Southampton
UK

OLAYA, CAMILO
Universidad de Los Andes
Bogotá
Colombia

OLDEMAN, BART
Concordia University
Montreal
Canada

OLHEDE, SOFIA
University College London
London
UK

Oliva, Rogelio
Texas A&M University
College Station
USA

Orbach, Raymond L.
University of California
Riverside
USA

Orlov, Michael
Ben-Gurion University
Beer-Sheva
Israel

Osipenko, George
State Polytechnic University
St. Petersburg
Russia

Osler, Carol
Brandeis University
Waltham
USA

Oullier, Olivier
Aix-Marseille Université
Marseille
France
Florida Atlantic University
Boca Raton
USA

Özak, Ömer
Brown University
Providence
USA

Page Jr., Frank H.
Indiana University
Bloomington
USA
Universite Paris 1
Pantheon–Sorbonne
France

Palsson, Bernhard Ø.
University of California
San Diego, La Jolla
USA

Panati, Gianluca
Università di Roma "La Sapienza"
Roma
Italy

Panov, Panče
Jožef Stefan Institute
Ljubljana
Slovenia

Papageorgiou, Anargyros
Columbia University
New York
USA

Papanikolopoulos, Nikolaos
University of Minnesota
Minneapolis
USA

Parker, Lynne E.
University of Tennessee
Knoxville
USA

Park, Michael
University of Pennsylvania
Philadelphia
USA

Parry, Hazel R.
Central Science Laboratory
York
UK

Paterson, Lincoln
CSIRO Petroleum
Clayton
Australia

Pattison, Philippa
University of Melbourne
Parkville
Australia

Păun, Gheorghe
Institute of Mathematics of the Romanian Academy
București
Romania

Pedrycz, Witold
University of Alberta
Edmonton
Canada
Polish Academy of Sciences
Warsaw
Poland

Peinke, Joachim
Carl-von-Ossietzky University Oldenburg
Oldenburg
Germany

PENTA, ANTONIO
University of Pennsylvania
Philadelphia
USA

PEPER, FERDINAND
National Institute of Information and Communications Technology
Kobe
Japan

PÉREZ-CASTRILLO, DAVID
Universitat Autònoma de Barcelona
Barcelona
Spain

PERLITZ, VOLKER
Universitätsklinikum der RWTH Aachen
Aachen
Germany

PETERSEN, KARL
University of North Carolina
Chapel Hill
USA

PETERS, JAMES F.
University of Manitoba
Winnipeg
Canada

PETKOVA, RALITSA
Texas A&M University
College Station
USA

PETRY, FREDERICK
Stennis Space Center
Mississippi
USA

PICASSO, LUIGI E.
Università di Pisa
Pisa
Italy
Sezione di Pisa
Pisa
Italy

PICCI, GIORGIO
University of Padua
Padua
Italy

PICCOLI, BENEDETTO
Consiglio Nazionale delle Ricerche
Rome
Italy

PIGER, JEREMY
University of Oregon
Eugene
USA

PIGORSCH, CHRISTIAN
University of Bonn
Bonn
Germany

PISMEN, L.M.
Technion – Israel Institute of Technology
Haifa
Israel

PIVATO, MARCUS
Trent University
Peterborough
Canada

PIWOWAR, MICHAEL S.
Securities Litigation and Consulting Group, Inc.
Fairfax
USA

PODGORELEC, VILI
University of Maribor
Maribor
Slovenia

POGACH, JONATHAN
University of Pennsylvania
Philadelphia
USA

POKROPIŃSKA, AGATA
Jan Dlugosz University
Częstochowa
Poland

POLET, JASCHA
California State Polytechnic University
Pomona
USA

POLKOWSKI, LECH
Polish-Japanese Institute of Information Technology
Warsaw
Poland

Popovych, Oleksandr V.
Research Center Jülich
Jülich
Germany

Port, Robert
Indiana University
Bloomington
USA

Portugali, Juval
Tel Aviv University
Tel Aviv
Israel

Povalej, Petra
University of Maribor
Maribor
Slovenia

Prade, Henri
Universite Paul Sabatier
Toulouse Cedex
France

Prejean, Stephanie G.
Alaska Science Center
Anchorage
USA

Price, Nathan D.
University of Illinois
Urbana-Champaign
USA

Prosen, Tomaž
Univerza v Ljubljani
Ljubljana
Slovenia

Pruessner, Gunnar
Imperial College London
London
UK

Pu, Calton
Georgia Institute of Technology
Atlanta
USA

Pujol, Jose
The University of Memphis
Memphis
USA

Pulvirenti, M.
Dipartimento di Matematica, Università di Roma 'La Sapienza'
Roma
Italy

Qian, Lei
Fisk University
Nashville
USA

Quas, Anthony
University of Victoria
Victoria
Canada

Quincampoix, Marc
Université de Bretagne Occidentale
Brest
France

Radzicki, Michael J.
Worcester Polytechnic Institute
Worcester
USA

Rafii-Tabar, H.
Shahid Beheshti University of Medical Sciences
Tehran
Iran
Institute for Research in Fundamental Sciences (IPM)
Tehran
Iran

Raghavan, T.E.S.
University of Illinois
Chicago
USA

Rahman, Rajib
Purdue University
West Lafayette
USA

Rakha, Hesham
Virginia Polytechnic Institute and State University
Blacksburg
USA

Rambaldi, Sandro
Physics Department and INFN Sezione di Bologna
Bologna
Italy

Rand, Darren
Massachusetts Institute of Technology
Lexington
USA

Ras, Zbigniew W.
University of North Carolina
Charlotte
USA
Polish Academy of Sciences
Warsaw
Poland

Ratiu, Tudor S.
École Polytechnique Fédérale de Lausanne
Lausanne
Switzerland

Ravaioli, Umberto
University of Illinois at Urbana-Champaign
Urbana
USA

Ravi, S.S.
University at Albany – State University of New York
New York
USA

Redner, Sidney
Boston University
Boston
USA

Regan, Erzsébet Ravasz
Beth Israel Deaconess Medical Center
Boston
USA

Regev, Oded
Technion–Israel – Institute of Technology
Haifa
Israel
Columbia University
New York
USA

Rehborn, Hubert
Daimler AG
Sindelfingen
Germany

Reif, John H.
Duke University
Durham
USA

Renault, Jérôme
Université Paris Dauphine
Paris
France

Ren, Yu-Jie
Shantou University
Shantou
People's Republic of China
Dalian Polytechnic University
Dalian
People's Republic of China
Beijing Institute of Applied Physics and Computational Mathematics
Beijing
People's Republic of China

Reza Rahimi Tabar, M.
Sharif University of Technology
Theran
Iran

Richardson, George P.
University at Albany, State University of New York
Albany
USA

Rich, Eliot
University at Albany
Albany
USA

Richiardi, Matteo G.
Università Politecnica delle Marche
Ancona
Italy
Collegio Carlo Alberto – LABORatorio R. Revelli
Moncalieri
Italy

Richter, Klaus
Universität Regensburg
Regensburg
Germany

Rieger, Heiko
Universität des Saarlandes
Saarbrücken
Germany

Rinker, Sherri
Arizona State University
Tempe
USA

Risler, Thomas
Laboratoire Physicochimie Curie (CNRS-UMR 168)
Paris
France
Université Pierre et Marie Curie Paris VI
Paris
France

Ritort, Felix
Universitat de Barcelona
Barcelona
Spain

Roach, Jared C.
Seattle Childrens Hospital
Seattle
USA

Robinett, Rush D.
Sandia National Laboratories
Albuquerque
USA

Robins, Garry
University of Melbourne
Melbourne
Australia

Roehner, Bertrand M.
University of Paris 7
Paris
France

Rogsch, Christian
University of Wuppertal
Wuppertal
Germany

Rosenzweig, Cynthia
Columbia University
New York
USA

Rosier, Lionel
Institut Elie Cartan
Vandoeuvre-lès-Nancy
France

Roth, Frederick P.
Harvard Medical School
Boston
USA
Dana-Farber Cancer Institute
Boston
USA

Rötteler, Martin
NEC Laboratories America, Inc.
Princeton
USA

Rouwette, Etiënne A. J. A.
Radboud University
Nijmegen
The Netherlands

Rozenfeld, Hernán D.
City College of New York
New York
USA

Rundle, John B.
University of California
Davis
USA

Rutkowski, Leszek
Częstochowa University of Technology
Częstochowa
Poland

Ryu, Hoon
Purdue University
West Lafayette
USA

Sabourian, Hamid
University of Cambridge
Cambridge
UK

Sacchetti, Andrea
Universitá di Modena e Reggio Emilia
Modena
Italy

Saeed, Khalid
Worcester Polytechnic Institute
Worcester
USA

Sahimi, Muhammad
University of Southern California
Los Angeles
USA

Saied, Faisal
Purdue University
West Lafayette
USA

SÁNCHEZ, ANGEL
Universidad Carlos III de Madrid
Madrid
Spain
Universidad de Zaragoza
Zaragoza
Spain

SANDER, LEONARD M.
The University of Michigan
Ann Arbor
USA

SANDHOLM, WILLIAM H.
University of Wisconsin
Madison
USA

SANDRONI, ALVARO
University of Pennsylvania
Philadelphia
USA

SANFELICE, RICARDO G.
University of Arizona
Tucson
USA

SANNS, WERNER
University of Applied Sciences
Darmstadt
Germany

SASTRA, JIMMY
University of Pennsylvania
Philadelphia
USA

SATAKE, KENJI
University of Tokyo
Tokyo
Japan

SATANIN, ARKADY M.
Russian Academy of Sciences
Nizhny Novgorod
Russia

SATO, HARUO
Tohoku University
Sendai-shi, Miyagi-ken
Japan

SATRIANO, CLAUDIO
Università di Napoli "Federico II" (RISSC-Lab)
Napoli
Italy

SAUL, JOACHIM
GeoForschungsZentrum Potsdam
Potsdam
Germany

SAURO, HERBERT M.
University of Washington
Seattle
USA

SBANO, LUCA
University of Warwick
Warwick
UK

SCARGLE, JEFFREY D.
NASA Ames Research Center
Moffett Field
USA

SCHADSCHNEIDER, ANDREAS
Universität zu Köln
Köln
Germany
Interdisziplinäres Zentrum für Komplexe Systeme
Bonn
Germany

SCHEID, MATTHIAS
Universität Regensburg
Regensburg
Germany

SCHERER, RAFAŁ
Częstochowa University of Technology
Częstochowa
Poland

SCHIEPEK, GÜNTER
Paracelsus Medical University
Salzburg
Austria

SCHMELING, JÖRG
Lund University
Lund
Sweden

SCHOENWALD, DAVID A.
Sandia National Laboratories
Albuquerque
USA

SCHOLZ, CHRISTOPHER H.
Columbia University
New York
USA

SCHWANINGER, MARKUS
University of St. Gallen
St. Gallen
Switzerland

SCOTT, JOHN
University of Plymouth
Plymouth
UK

SELMAN, DENIZ
University of Pennsylvania
Philadelphia
USA

SERRANO, ROBERTO
Brown University
Providence
USA
IMDEA-Social Sciences
Madrid
Spain

SETH, ANIL K.
University of Sussex
Brighton
UK

SEYBOLD, PAUL G.
Wright State University
Dayton
USA

SEYFRIED, ARMIN
Research Centre Jülich
Jülich
Germany

SHAKED-MONDERER, NAOMI
Emek Yezreel College
Emek Yezreel
Israel

SHAPIRA, DANIEL
Ben–Gurion University
Beer Sheva
Israel

SHEN, JIANHONG
Barclays Capital
New York
USA

SHETH, KAPIL
NASA Ames Research Center
Moffet Field
USA

SHIN, MINCHEOL
Information and Communications University
Yuseong
Korea

SHINTANI, MOTOTSUGU
Vanderbilt University
Nashville
USA

SHIPLEY, BILL
Université de Sherbrooke
Sherbrooke
Canada

SHMULEVICH, ILYA
Department of Bioengineering University of Washington
Seattle
USA

SHORE, STEVEN N.
Università di Pisa
Sezione di Pisa
Pisa
Italy

SIBANI, PAOLO
SDU
Odense
Denmark

SICONOLFI, ANTONIO
"La Sapienza" Università di Roma
Roma
Italy

SIGANOS, GEORGOS
University of California
Riverside
USA

SILVA, CESAR E.
Williams College
Williamstown
USA

SIPPER, MOSHE
Ben-Gurion University
Beer-Sheva
Israel

SKOWRON, ANDRZEJ
Warsaw University
Warsaw
Poland

Slanina, František
Academy of Sciences of the Czech Republic
Prague
Czech Republic
Center for Theoretical Study
Prague
Czech Republic

Słowiński, Roman
Poznan University of Technology
Poznan
Poland
Polish Academy of Sciences
Warsaw
Poland

Smit, Arian F. A.
Institute for Systems Biology
Seattle
USA

Snijders, Tom A. B.
University of Oxford
Oxford
UK

Sobel, Joel
University of California
San Diego
USA

Sokolov, Igor M.
Humboldt-Universität zu Berlin
Berlin
Germany

Solan, Eilon
Tel Aviv University
Tel Aviv
Israel

Solé, Ricard V.
Santa Fe Institute
Santa Fe
USA

Solomon, Sorin
Hebrew University
Jerusalem
Israel

Solomon, Tom H.
Bucknell University
Lewisburg
USA

Soloviev, Alexandre
Russian Academy of Sciences
Moscow
Russia
The Abdus Salam International Center for Theoretical Physics
Trieste
Italy

Song, Chaoming
City College of New York
New York
USA

Song, Myung-Sin
Southern Illinois University
Edwardsville
USA

Sontag, Eduardo D.
Rutgers University
New Brunswick
USA

Soomere, Tarmo
Tallinn University of Technology
Tallinn
Estonia

Sornette, Didier
Technology and Economics
ETH Zurich
Switzerland

Sotomayor, Marilda
University of São Paulo/SP
São Paulo
Brazil
Brown University
Providence
USA

Sparks, R. Stephen J.
University of Bristol
Bristol
UK

Sridhar, Banavar
NASA Ames Research Center
Moffet Field
USA

Stamatiadis, Chronis
University of Massachusetts
Lowell
USA

STARCK, JEAN-LUC
CEA/Saclay
Gif sur Yvette
France

STAUFFER, ANDRÉ
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Lausanne
Switzerland

STAUFFER, DIETRICH
Cologne University
Köln
Germany

STEENEVELD, GERT-JAN
Wageningen University
Wageningen
The Netherlands

STEFANOVIC, DARKO
University of New Mexico
Albuquerque
USA

STEIGLITZ, KEN
Princeton University
Princeton
USA

STEINMETZ, LARS M.
European Molecular Biology Laboratory
Heidelberg
Germany

STEPANIUK, JAROSŁAW
Białystok University of Technology
Białystok
Poland

STEPHENS, DAVID A.
McGill University
Montreal
Canada

STIGLIC, GREGOR
University of Maribor
Maribor
Slovenia

STINGU, PETRU EMANUEL
University of Texas at Arlington
Fort Worth
USA

STOJANOVIC, MILAN
Columbia University
New York
USA

STRAWIŃSKA, URSZULA
Warsaw School for Social Psychology
Warsaw
Poland

STRELNIKER, YAKOV M.
Bar–Ilan University
Ramat–Gan
Israel

STUPAZZINI, MARCO
Politecnico di Milano
Milano
Italy

SURI, NIRANJAN
Institute for Human and Machine Cognition
Pensacola
USA

SUSSMAN, GERALD JAY
Massachusetts Institute of Technology
Cambridge
USA

SUTHAKORN, JACKRIT
Mahidol University
Salaya
Thailand

SUTNER, KLAUS
Carnegie Mellon University
Pittsburgh
USA

SUZUKI, NORIKAZU
Nihon University
Chiba
Japan

SZNAIER, MARIO
Northeastern University
Boston
USA

TAKAYASU, HIDEKI
Sony Computer Science Laboratories Inc
Tokyo
Japan

TAKAYASU, MISAKO
Tokyo Institute of Technology
Tokyo
Japan

TAKEI, YASUKO
University of Tokyo
Tokyo
Japan

TANG, LEI-HAN
Hong Kong Baptist University
Kowloon Tong, Hong Kong SAR
China

TASS, PETER A.
Research Center Jülich
Jülich
Germany
University of Cologne
Cologne
Germany

TÄUBER, UWE CLAUS
Virginia Polytechnic Institute and State University
Blacksburg
USA

TAWFIK, ALY
Virginia Polytechnic Institute and State University
Blacksburg
USA

TEEL, ANDREW R.
University of California
Santa Barbara
USA

TEISSEYRE, ROMAN
Polish Academy of Sciences
Warsaw
Poland

TEIXEIRA, MARCO ANTÔNIO
Universidade Estadual de Campinas
Campinas
Brazil

TEMPESTI, GIANLUCA
University of York
York
UK

TERENTJEV, EUGENE M.
University of Cambridge
Cambridge
UK

TERRACINI, SUSANNA
Università di Milano Bicocca
Milano
Italia

TEUSCHER, CHRISTOF
Los Alamos National Laboratory
Los Alamos
USA

THERMES, CLAUDE
CNRS
Gif-sur-Yvette
France

THIELE, INES
University of California
San Diego
USA

THORPE, MICHAEL F.
Arizona State University
Tempe
USA

THOVERT, JEAN-FRANÇOIS
CNRS-LCD
Chasseneuil du Poitou
France

TILLING, ROBERT I.
US Geological Survey
Menlo Park
USA

TIMMIS, JON
University of York
York
UK

TINCANI, MICHELA
University of Pennsylvania
Philadelphia
USA

TKAČIK, GAŠPER
Princeton University
Princeton
USA

TODOROVSKA, MARIA I.
University of Southern California
Los Angeles
USA

TOL, RICHARD S. J.
Economic and Social Research Institute
Dublin
Ireland
Vrije Universiteit
Amsterdam
The Netherlands
Carnegie Mellon University
Pittsburgh
USA

TOMALA, TRISTAN
HEC Paris
Paris
France

TOMASSONE, M. SILVINA
Rutgers, The State University of New Jersey
Piscataway
USA

TORRA, VICENÇ
Institut d'Investigació en Intelligència Artificial – CSIC
Bellaterra
Spain

TOSIN, ANDREA
Consiglio Nazionale delle Ricerche
Rome
Italy

TOUCHON, MARIE
CNRS
Paris
France
Université Pierre et Marie Curie
Paris
France

TOUCHTON, ROBERT A.
Honeywell International
Phoenix
USA

TRAUB, JOSEPH F.
Columbia University
New York
USA

TRIFUNAC, MIHAILO D.
University of Southern California
Los Angeles
USA

TRINAJSTIĆ, NENAD
The Rugjer Bošković Institute
Zagreb
Croatia

TRIVEDI, PRAVIN K.
Indiana University
Bloomington
USA

TROITZSCH, KLAUS G.
Universität Koblenz-Landau
Koblenz
Germany

TROPSHA, ALEXANDER
University of North Carolina at Chapel Hill
Chapel Hill
USA

TSAI, SHAN-WEN
University of California
Riverside
USA

TSALLIS, CONSTANTINO
Centro Brasileiro de Pesquisas Físicas
Rio de Janeiro
Brazil
Santa Fe Institute
Santa Fe
USA

TSCHACHER, WOLFGANG
University of Bern
Bern
Switzerland

TSENG, VINCENT S.
National Cheng–Kung University
Tainan
Taiwan

TSUMOTO, SHUSAKU
Faculty of Medicine, Shimane University
Shimane
Japan

TURCOTTE, DONALD L.
University of California
Davis
USA

TÜRKŞEN, I. BURHAN
TOBB-ETÜ, (Economics and Technology University of the Union of Turkish Chambers and Commodity Exchanges)
Ankara
Republic of Turkey

UMEO, HIROSHI
University of Osaka
Osaka
Japan

UNGARELLI, CARLO
CNR-Institute of Geosciences and Earth Resources
Pisa
Italy

UNTIEDT, ELIZABETH A.
University of Colorado Denver
Denver
USA

USMAN, MUHAMMAD
Purdue University
West Lafayette
USA

UYEDA, SEIYA
Tokai University
Shizuoka
Japan

VALENTE, THOMAS W.
University of Southern California
Alhambra
USA

VALLACHER, ROBIN R.
Florida Atlantic University
Boca Raton
USA

VALVERDE, SERGI
Parc de Recerca Biomedica de Barcelona
Barcelona
Spain

VANDERBAUWHEDE, ANDRÉ
Ghent University
Gent
Belgium

VAN GEERT, PAUL
The Heymans Institute
Groningen
The Netherlands

VAN NIEUWERBURGH, STIJN
New York University
New York
USA

VARGIAMIDIS, VASSILIOS
Aristotle University
Thessaloniki
Greece

VEGA, CLARA
Board of Governors of the Federal Reserve System
Washington DC
USA

VELTRI, PIERLUIGI
Università della Calabria
Arcavacata di Rende
Italy

VENNIX, JAC A. M.
Radboud University
Nijmegen
The Netherlands

VERE-JONES, DAVID
Statistical Research Associates and Victoria University
Wellington
New Zealand

VERHAGEN, HARKO
Stockholm University and Royal Institute of Technology
Stockholm
Sweden

VERHULST, FERDINAND
University of Utrecht
Utrecht
The Netherlands

VERLIC, MATEJA
University of Maribor
Maribor
Slovenia

VIANA, MARCELO
IMPA
Rio de Janeiro
Brazil

VIDYASAGAR, M.
Software Units Layout
Hyderabad
India

VITEK, JAN
Purdue University
West Lafayette
USA

VIZZARI, GIUSEPPE
University of Milan-Bicocca
Milan
Italy

VOLIJ, OSCAR
Ben-Gurion University
Beer-Sheva
Israel

VOORHEES, BURTON
Athabasca University
Athabasca
Canada

WAGNER, ANDREAS
University of Zurich
Zurich
Switzerland
The Santa Fe Institute
New Mexico
USA

WAKO, JUN
Gakushuin University
Tokyo
Japan

WALCHER, SEBASTIAN
RWTH Aachen
Aachen
Germany

WALKER, IAN D.
Clemson University
Clemson
USA

WANG, BING-HONG
University of Science and Technology of China
Hefei Anhui
China
Shanghai Academy of System Science
Shanghai
China

WANG, CAI-ZHUANG
Iowa State University
Ames
USA

WANG, KELIN
Geological Survey of Canada
Sidney
Canada
University of Victoria
Victoria
Canada

WARD, TOM
University of East Anglia
Norwich
UK

WASILEWSKA, ANITA
Stony Brook University
Stony Brook
USA

WATROUS, JOHN
University of Waterloo
Waterloo
Canada

WAZWAZ, ABDUL-MAJID
Saint Xavier University
Chicago
USA

WEBB, STEVE
Georgia Institute of Technology
Atlanta
USA

WEBER, EMANUEL
Istituto Nazionale di Geofisica e Vulcanologia
(RISSC-Lab)
Napoli
Italy

WEINSTEIN, STUART
NOAA/NWS/Pacific Tsunami Warning Center
Ewa Beach
USA

WERNER, MAXIMILIAN J.
Institute of Geophysics
ETH Zurich
Switzerland

WHITE, ANDREW G.
The University of Queensland
Brisbane
Australia

WHITE, DOUGLAS R.
University of California
Irvine
USA
Santa Fe Institute
Santa Fe
USA

WHITE, PAUL
University of Pennsylvania
Philadelphia
USA

WIERMAN, JOHN C.
Johns Hopkins University
Baltimore
USA

WIESNER, KAROLINE
University of Bristol
Bristol
UK

WILKINSON, AMIE
Northwestern University
Evanston
USA

WILLIAMS, TERRY
Southampton University
Southampton
UK

WIMMER, MICHAEL
Universität Regensburg
Regensburg
Germany

WINFIELD, ALAN FT
University of the West of England
Bristol
UK

WINKIELMAN, PIOTR
University of California
San Diego
USA

WOLSTENHOLME, ERIC
South Bank University
London
UK
Symmetric SD
Brighton
UK

WOODERS, MYRNA
Vanderbilt University
Nashville
USA
University of Warwick
Coventry
UK

WOODS, DAMIEN
University College Cork
Cork
Ireland
University if Seville
Seville
Spain

WOOLDRIDGE, JEFFREY M.
Michigan State University
East Lansing
USA

WORSCH, THOMAS
Universität Karlsruhe
Karlsruhe
Germany

WU, YIH-MIN
National Taiwan University
Taipei
Taiwan

XIONG, SHI-JIE
Nanjing University
Nanjing
China

YAARI, GUR
Institute for Scientific Interchange
Turin
Italy
Hebrew University
Jerusalem
Israel

YAKOVENKO, VICTOR M.
University of Maryland
College Park
USA

YANG, JUDITH C.
University of Pittsburgh
Pittsburgh
USA

YANG, WEI-ZHE
National Taiwan University
Taipei
Taiwan

YAN, HAO
Arizona State University
Tempe
USA

YEUNG, CHI HO
The Hong Kong University of Science and Technology
Hong Kong
China
Université de Fribourg
Pérolles, Fribourg
Switzerland
University of Electronic Science and Technology of China (UESTC)
Chengdu
China

YILMAZ, LEVENT
Auburn University
Alabama
USA

YIM, MARK
University of Pennsylvania
Philadelphia
USA

YUEN, DAVID A.
University of Minnesota
Minneapolis
USA

YULMETYEV, RENAT M.
Tatar State University of Pedagogical and Humanities Sciences
Kazan
Russia

ZADEH, LOTFI A.
University of California
Berkeley
USA

ZAMIR, SHMUEL
Hebrew University
Jerusalem
Israel

ZANG, YI
Zhejiang Normal University
Jinhua
China

ZEIGLER, BERNARD
University of Arizona
Tucson
USA

ZEITOUNI, OFER
University of Minnesota
Minneapolis
USA

ŽENKO, BERNARD
Jožef Stefan Institute
Ljubljana
Slovenia

ZHANG, BO
Tsinghua University
Beijing
China

ZHANG, LAN V.
Harvard Medical School
Boston
USA

ZHANG, LING
Anhui University, Hefei
Anhui
China

ZHANG, YAN-QING
Georgia State University
Atlanta
USA

ZHANG, YI-CHENG
The Hong Kong University of Science and Technology
Hong Kong
China
Université de Fribourg
Pérolles, Fribourg
Switzerland
University of Electronic Science and Technology of China (UESTC)
Chengdu
China

ZHAO, MING
University of Science and Technology of China
Hefei Anhui
China

ZHAO, YI
University of Toronto
Toronto
Canada

ZHAO, ZHENG
Arizona State University
Tempe
USA

ZHILINSKIÍ, BORIS
Université du Littoral
Dunkerque
France

ZHOU, HAO-MIN
Georgia Institute of Technology
Atlanta
USA

ZHOU, TAO
University of Science and Technology of China
Hefei Anhui
China

ZHU, SHUN-DONG
Zhejiang Lishui University
Lishui
China

ZIFF, ROBERT M.
University of Michigan
Ann Arbor
USA

ŽITKO, ROK
Jožef Stefan Institute
Ljubljana
Slovenia

ZÖLLER, GERT
University of Potsdam
Potsdam
Germany

ZOLLO, ALDO
Università di Napoli "Federico II" (RISSC-Lab)
Napoli
Italy

ZORMAN, MILAN
University of Maribor
Maribor
Slovenia

Peer Reviewers

Filippo Castiglione, Thomas Lux, Marco Pedicini, Dietrich Stauffer, Warren Dixon, Andrew Adamatzky, Steve N. Shore, Hartmut Grassl, Yakir Berchenko, Geoffrey Canright, Niloy Ganguly, Gerhard Weikum, Danail Bonchev, Sumiyoshi Abe, Richard Allen, J. P. Ampuero, Mark Bebbington, Margaret Boettcher, Yehuda Ben-Zion, Susan Bilek, Peter Bormann, Michele Caputo, H. D. Ceniceros, Dave Chadwell, Bernard Chouet, Ton Correig, Luca D'Auria, Renata Dmowska, Douglas P. Drob, Donna Eberhart-Phillips, John R. Evans, Delphine Fitzenz, Eric L. Geist, A. J. Hale, Matt Haney, Jeanne Hardebeck, Ruth A. Harris, Dave P. Hill, James Holliday, Heiner Igel, Erol Kalkan, Hiroo Kanamori, Vladimir Keilis-Borok, Annabel Kelly, Fred Klein, Michael Korn, Rod Lakes, Serge Lallemand, Steve Lane, John Langbein, Markus Lazar, William H. K. Lee, Anthony Lomax, Cinna Lomnitz, Anthony Lowry, Vladimir Lyakhovsky, Martin Mai, Warner Marzocchi, Art McGarr, Steve McNutt, Jim Moore, Patrick Muffler, Jessica Murray, Masaru Nakano, Takeshi Nishimura, Edo Nyland, Emile Okal, Paolo Papale, John Power, German Prieto, Jose Pujol, David Rhoades, Luis Rivera, Russell Robinson, Malcolm Sambridge, Charles Sammis, Kenji Satake, William U Savage, Rick Paik Schoenberg, Rick Sibson, Mark Simons, Roel Snieder, Didier Sornette, Chris Stephens, Ta-Liang Teng, Mihailo Trifunac, David Vere-Jones, Kelin Wang, Ru-Shan Wu, Harry Yeh, Zbigniew Zembaty, Gert Zöller, Yong Suk Joe, Peter Kokol, Bai-Lian Li, Robert A. Meyers, JonAaronson, Ethan Coven, Thierry de la Rue, Andres del Junco, Dimitry Dolgopyat, Nikos Frantzikinakis, Katrin Gelfert, Eli Glasner, Aimee Johnson, Bryna Kra, Mariusz Lemanczyk, Doug Lind, Randall McCutcheon, Karl Petersen, Anthony Quas, Omri Sarig, Akshay Venkatesh, Tom Ward, Barak Weiss, Bruce Mizrach, Daniel ben-Avraham, Shlomo Havlin, Okada Akira, Nizar Allouch, Bezalel Beleg, Pierre Bernhard, Ulle Endriss, Marta Faias, Thomas Ferguson, Olivier Gossner, Aviad Heifetz, Johannes Horner, Marc Kilgour, Jerome Lang, Jihong Lee, Ehud Lehrer, Xiao Luo, Hervé Moulin, John Nachbar, Mikio Nakayama, Yuichi Noguchi, Slawomir Plaskacz, Roy Radner, Dinah Rosenberg, Roberto Serrano, Marilda Sotomayor, Vincent Vannetelbosch, Vera Hernandez Marcos, Reinhilde Veugelers, Peter Vida, Shmuel Zamir, Tsau Y. Lin, James A. Hendler, Ferdinand Verhulst, Italo Capuzzo Dolcetta, Muhammad Sahimi, Giuseppe Gaeta, Henrik Jeldtoft Jensen, Joseph F. Traub, John Scott, Andrzej Nowak, Janusz Kacprzyk, Mohamed A. Helal, Dogan Kaya, Hermann Haken, Matthias Kawski, Gabor Balazsi, Hamid Bolouri, Ed Dougherry, Tim Galitski, Simon Levin, Adrian Ozinsky, Nathan Price, Prahlad Ram, Jeff Ranish, Lee Rowen, Zhaolei Zhang, BobCavana, Brian Dangerfield, Susan Howick, Jim Lyneis, Geoff McDonnell, Mohammad Mojtahedzadeh, John Morecroft, Erling Moxnes, Dave Packer, Kathy Taylor, Kim Warren, David Wheat, Aparna Baskaran, Mark Bowick, Mongwea Jeng, Ulrich Kuhl, M. Cristina Marchetti, Olivier Martin, Jennifer Schwarz, Xiangjun Xing, Hermann Haken, Armando Bazzani, Moshe E. Ben-Akiva, Michael J. Demetsky, Song Gao, Boris Kerner, Sergey Klenov, N. Harris McClamroch, Hubert Rehborn, Andreas Schadschneider, Martin Treiber, Andrew Adamatzky, Edward Aboufadel

How to Access the Articles

Each Section Editor prepared an introductory article describing their field, the articles comprising the section and the interrelationship of the articles. Thus, our readership can utilize each section introduction as the reference point to the individual articles or can look up individual articles by alphabetical entry. The introductory articles occur alphabetically, under the section name, among the nearly 600 alphabetically listed entries.

A

Abrupt Climate Change Modeling

Gerrit Lohmann
Alfred Wegener Institute for Polar and Marine Research, Bremerhaven, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
A Mathematical Definition
Earth System Modeling and Analysis
Example: Glacial-Interglacial Transitions
Example: Cenozoic Climate Cooling
Examples: Transient Growth
Future Directions
Bibliography

Glossary

Atmosphere The atmosphere is involved in many processes of abrupt climate change, providing a strong non-linearity in the climate system and propagating the influence of any climate forcing from one part of the globe to another. Atmospheric temperature, composition, humidity, cloudiness, and wind determine the Earth's energy fluxes. Wind affects the ocean's surface circulation and upwelling patterns. Atmospheric moisture transport determines the freshwater balance for the oceans, overall water circulation, and the dynamics of glaciers.

Oceans Because water has enormous heat capacity, oceans typically store 10–100 times more heat than equivalent land surfaces. The oceans exert a profound influence on climate through their ability to transport heat from one location to another. Changes in ocean circulation have been implicated in abrupt climate change of the past. Deglacial meltwater has freshened the North Atlantic and reduced the ability of the water to sink, inducing long-term coolings.

Land surface The reflective capacity of the land can change greatly, with snow or ice sheets reflecting up to 90% of the sunlight while dense forests absorb more than 90%. Changes in surface characteristics can also affect solar heating, cloud formation, rainfall, and surface-water flow to the oceans, thus feeding back strongly on climate.

Cryosphere The portion of the Earth covered with ice and snow, the cryosphere, greatly affects temperature. When sea ice forms, it increases the planetary reflective capacity, thereby enhancing cooling. Sea ice also insulates the atmosphere from the relatively warm ocean, allowing winter air temperatures to steeply decline and reduce the supply of moisture to the atmosphere. Glaciers and snow cover on land can also provide abrupt-change mechanisms. The water frozen in a glacier can melt if warmed sufficiently, leading to possibly rapid discharge, with consequent effects on sea level and ocean circulation. Meanwhile, snow-covered lands of all types maintain cold conditions because of their high reflectivity and because surface temperatures cannot rise above freezing until the snow completely melts.

External factors Phenomena external to the climate system can also be agents of abrupt climate change. For example, the orbital parameters of the Earth vary over time, affecting the latitudinal distribution of solar energy. Furthermore, fluctuations in solar output, prompted by sunspot activity or the effects of solar wind, as well as volcanoes may cause climate fluctuations.

Climate time scales The climate system is a composite system consisting of five major interactive components: the atmosphere, the hydrosphere, including the oceans, the cryosphere, the lithosphere, and the biosphere. All subsystems are open and non-isolated, as the atmosphere, hydrosphere, cryosphere and biosphere act as cascading systems linked by complex

feedback processes. Climate refers to the average conditions in the Earth system that generally occur over periods of time, usually several decades or longer. This time scale is longer than the typical response time of the atmosphere. Parts of the other components of the Earth system (ice, ocean, continents) have much slower response times (decadal to millennial).

Climate variables and forcing State variables are temperature, rainfall, wind, ocean currents, and many other variables in the Earth system. In our notation, the variables are described by a finite set of real variables in a vector $x(t) \in \mathbb{R}^n$. The climate system is subject to two main external forcings $F(x, t)$ that condition its behavior, solar radiation and the action of gravity. Since $F(x, t)$ has usually a spatial dependence, F is also a vector $\in \mathbb{R}^n$. Solar radiation must be regarded as the primary forcing mechanism, as it provides almost all the energy that drives the climate system. The whole climate system can be regarded as continuously evolving, as solar radiation changes on diurnal, seasonal and longer time scales, with parts of the system leading or lagging in time. Therefore, the subsystems of the climate system are not always in equilibrium with each other. Indeed, the climate system is a dissipative, highly non-linear system, with many instabilities.

Climate models are based on balances of energy, momentum, and mass, as well as radiation laws. There are several model categories, full circulation models, low-order models, and models of intermediate complexity. Climate models simulate the interactions of the atmosphere, oceans, land surface, and ice. They are used for a variety of purposes from study of the dynamics of the weather and climate system, past climate to projections of future climate.

Global climate models or General circulation models (GCMs) The balances of energy, momentum, and mass are formulated in the framework of fluid dynamics on the rotating Earth. GCMs discretize the equations for fluid motion and energy transfer and integrate these forward in time. They also contain parametrization for processes – such as convection – that occur on scales too small to be resolved directly. The dimension of the state vector is in the order of $n \sim 10^5 - 10^8$ depending on the resolution and complexity of the model.

Model categories In addition to complex numerical climate models, it can be of great utility to reduce the system to low-order, box, and conceptual models. This complementary approach has been successfully applied to a number of questions regarding feedback mechanisms and the basic dynamical behavior, e. g. [48,84]. In some cases, e. g. the stochastic climate model of Hasselmann [32], such models can provide a null hypothesis for the complex system. The transition from highly complex dynamical equations to a low-order description of climate is an important topic of research. In his book "Dynamical Paleoclimatology", Saltzman [77] formulated a dynamical system approach in order to differentiate between fast-response and slow-response variables. As an alternative to this method, one can try to derive phenomenologically based concepts of climate variability, e. g. [21,43]. In between the comprehensive models and conceptual models, a wide class of "models of intermediate complexity" were defined [12].

Earth-system models of intermediate complexity (EMICs) Depending on the nature of questions asked and the pertinent time scales, different types of models are used. There are, on the one extreme, conceptual models, and, on the other extreme, comprehensive models (GCMs) operating at a high spatial and temporal resolution. Models of intermediate complexity bridge the gap [12]. These models are successful in describing the Earth system dynamics including a large number of Earth system components. This approach is especially useful when considering long time scales where the complex models are computationally too expensive, e. g. [47]. Improvements in the development of coupled models of intermediate complexity have led to a situation where modeling a glacial cycle, even with prognostic atmospheric CO_2 is becoming possible.

Climate simulation A climate simulation is the output of a computer program that attempts to simulate the climate evolution under appropriate boundary conditions. Simulations have become a useful part of climate science to gain insight into the sensitivity of the system.

Climate variability pattern Climate variability is defined as changes in integral properties of the climate system. True understanding of climate dynamics and prediction of future changes will come only with an understanding of the Earth system as a whole, and over past and present climate. Such understanding requires identification of the patterns of climate variability and their relationships to known forcing. Examples for climate variability patterns are the North Atlantic Oscillation (NAO) or the El Niño-Southern Oscillation (ENSO).

Abrupt climate change One can define abrupt climate change in the time and frequency domain. (a) Time domain: Abrupt climate change refers to a large shift in climate that persists for years or longer, such

as marked changes in average temperature, or altered patterns of storms, floods, or droughts, over a widespread area that takes place so rapidly that the natural system has difficulty adapting to it. In the context of past abrupt climate change, "rapidly" typically means on the order of a decade. (b) Frequency domain: An abrupt change means that the characteristic periodicity changes. Also the phase relation between certain climate variables may change in a relatively short time. For both types of changes examples will be provided.

Regime shifts are defined as rapid transitions from one state to another. In the marine environment, regimes may last for several decades, and shifts often appear to be associated with changes in the climate system. If the shifts occur regularly, they are often referred to as an oscillation (e. g., Atlantic Multi-decadal Oscillation, Pacific Decadal Oscillation). Similarly, one can define a regime shift in the frequency domain.

Anthropogenic climate change Beginning with the industrial revolution in the 1850s and accelerating ever since, the human consumption of fossil fuels has elevated CO_2 levels from a concentration of ~ 280 ppm to more than 380 ppm today. These increases are projected to reach more than 560 ppm before the end of the 21st century. As an example, a concomitant shift of ocean circulation would have serious consequences for both agriculture and fishing.

Multiple equilibria Fossil evidence and computer models demonstrate that the Earth's complex and dynamic climate system has more than one mode of operation. Each mode produces different climate patterns. The evidence of models and data analysis shows that the Earth's climate system has sensitive thresholds. Pushed past a threshold, the system can jump from one stable operating mode to a completely different one.

Long-term climate statistics Starting with a given initial state, the solutions $x(t)$ of the equations that govern the dynamics of a non-linear system, such as the atmosphere, result in a set of long-term statistics. If all initial states ultimately lead to the same set of statistical properties, the system is ergodic or transitive. If, instead, there are two or more different sets of statistical properties, where some initial states lead to one set, while the other initial states lead to another, the system is called intransitive (one may call the different states regimes). If there are different sets of statistics that a system may assume in its evolution from different initial states through a long, but finite, period of time, the system is called almost intransitive [50,51,53]. In the transitive case, the equilibrium climate statistics are both stable and unique. Long-term climate statistics will give a good description of the climate. In the almost intransitive case, the system in the course of its evolution will show finite periods during which distinctly different climatic regimes prevail. The almost intransitive case arises because of internal feedbacks, or instabilities involving the different components of the climatic system. The climatic record can show rapid step-like shifts in climate variability that occur over decades or less, including climatic extremes (e. g. drought) that persist for decades.

Feedbacks A perturbation in a system with a negative feedback mechanism will be reduced whereas in a system with positive feedback mechanisms, the perturbation will grow. Quite often, the system dynamics can be reduced to a low-order description. Then, the growth or decay of perturbations can be classified by the systems' eigenvalues or the pseudospectrum. Consider the stochastic dynamical system

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = f(x) + g(x)\xi + F(x,t)\,, \tag{1}$$

where ξ is a stochastic process. The functions f, g describe the climate dynamics, in this case without explicit time dependence. The external forcing $F(x,t)$ is generally time-, variable-, and space-dependent. In his theoretical approach, Hasselmann [32] formulated a linear stochastic climate model

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = Ax + \sigma\xi + F(t)\,, \tag{2}$$

with system matrix $A \in \mathbb{R}^{n\times n}$, constant noise term σ, and stochastic process ξ. Interestingly, many features of the climate system can be well described by (2), which is analogous to the Ornstein–Uhlenbeck process in statistical physics [89]. In the climate system, linear and non-linear feedbacks are essential for abrupt climate changes.

Paleoclimate Abrupt climate change is evident in model results and in instrumental records of the climate system. Much interest in the subject is motivated by the evidence in archives of extreme changes. Proxy records of paleoclimate are central to the subject of abrupt climate change. Available paleoclimate records provide information on many environmental variables, such as temperature, moisture, wind, currents, and isotopic compositions.

Thermohaline circulation stems from the Greek words "thermos" (heat) and "halos" (salt). The ocean is driven to a large extent by surface heat and freshwater

fluxes. As the ocean is non-linear, it cannot be strictly separated from the wind-driven circulation. The expressions thermohaline circulation (THC) and meridional overturning circulation (MOC) in the ocean are quite often used as synonyms although the latter includes all effects (wind, thermal, haline forcing) and describes the ocean transport in meridional direction. Another related expression is the ocean conveyor belt. This metaphor is motivated by the fact that the North Atlantic is the source of the deep limb of a global ocean circulation system [10]. If North Atlantic surface waters did not sink, the global ocean circulation would cease, currents would weaken or be redirected. The resulting reorganization would reconfigure climate patterns, especially in the Atlantic Ocean. One fundamental aspect of this circulation is the balance of two processes: cooling of the deep ocean at high latitudes, and heating of deeper levels from the surface through vertical mixing.

Definition of the Subject

The occurrence of abrupt change of climate at various time scales has attracted a great deal of interest for its theoretical and practical significance [2,3,9]. To some extent, a definition of what constitutes an abrupt climatic change depends on the sampling interval of the data being examined [28]. For the instrumental period covering approximately the last 100 years of annually or seasonally sampled data, an abrupt change in a particular climate variable will be taken to mean a statistically highly significant difference between adjacent 10-year sample means. In the paleoclimate context (i. e. on long time scales), an abrupt climate change can be in the order of decades to thousands of years. Since the climate dynamics can be often projected onto a limited number of modes or patterns of climate variability (e. g., [21,22]), the definition of abrupt climate change is also related to spatio-temporal patterns.

The concept of abrupt change of climate is therefore applied for different time scales. For example, changes in climatic regimes were described associated with surface temperature, precipitation, atmospheric circulation in North America during the 1920s and 1960s [19,75]. Sometimes, the term “climate jump” is used instead of “abrupt climate change”, e. g. [92]. Flohn [25] expanded the concept of abrupt climate change to include both singular events and catastrophes such as the extreme El Niño of 1982/1983, as well as discontinuities in paleoclimate indices taken from ice cores and other proxy data. In the instrumental record covering the last 150 years, there is a well-documented abrupt shift of sea surface temperature and atmospheric circulation features in the Northern Hemisphere in the mid-1970s, e. g. [22,67,88]. Some of the best-known and best-studied widespread abrupt climate changes started and ended during the last deglaciation, most pronounced at high latitudes.

In his classic studies of chaotic systems, Lorenz has proposed a deterministic theory of climate change with his concept of the "almost-intransitivity" of the highly nonlinear climate systems. In this set of equations, there exists the possibility of multiple stable solutions to the governing equations, even in the absence of any variations in external forcing [51]. More complex models, e. g. [11,20] also demonstrated this possibility. On the other hand, variations in external forcing, such as the changes of incoming solar radiation, volcanic activity, deglacial meltwater, and increases of greenhouse gas concentration have also been proposed to account for abrupt changes in addition to climate intransitivity [9,25,38,41,49]. A particular climate change is linked to the widespread continental glaciation of Antarctica during the Cenozoic (65 Ma to present) at about 34 Ma, e. g. [93]. It should be noted that many facets of regional climate change are abrupt changes although the global means are rather smoothly changing.

Besides abrupt climate change as described in the time domain, we can find abrupt shifts in the frequency domain. A prominent example for an abrupt climate change in the frequency domain is the mid-Pleistocene transition or revolution (MPR), which is the last major “event” in a secular trend towards more intensive global glaciation that characterizes the last few tens of millions of years. The MPR is the term used to describe the transition between 41 ky (ky $= 10^3$ years) and 100 ky glacial-interglacial cycles which occurred about one million years ago (see a recent review in [61]). Evidence of this is provided by high-resolution oxygen isotope data from deep sea cores, e. g. [45,83].

Another example is the possibility of greenhouse gas-driven warming leading to a change in El Niño events. Modeling studies indicate that a strong enhancement of El Niño conditions in the future is not inconceivable [85]. Such a shift would have enormous consequences for both the biosphere and humans. The apparent phase shifts during the 1970s seems unique over this time period, and may thus represent a real climate shift although the available time series is probably too short to unequivocally prove that the shift is significant [90]. The inability to resolve questions of this kind from short instrumental time series provides one of the strongest arguments for extending the instrumental record of climate variability with well-dated, temporally finely resolved and rigorously calibrated proxy data.

Introduction

One view of climate change is that the Earth's climate system has changed gradually in response to both natural and human-induced processes. Researchers became intrigued by abrupt climate change when they discovered striking evidence of large, abrupt, and widespread changes preserved in paleoclimatic archives, the history of Earth's climate recorded in tree rings, ice cores, sediments, and other sources. For example, tree rings show the frequency of droughts, sediments reveal the number and type of organisms present, and gas bubbles trapped in ice cores indicate past atmospheric conditions.

The Earth's climate system is characterized by change on all time and space scales, and some of the changes are abrupt even relative to the short time scales of relevance to human societies. Paleoclimatic records show that abrupt climate changes have affected much or all of the Earth repeatedly over the last ice-age cycle as well as earlier – and these changes sometimes have occurred in periods as short as a few years, as documented in Greenland ice cores. Perturbations at northern high latitudes were spectacularly large: some had temperature increases of up to 10–20°C and a local doubling of precipitation within decades.

In the frequency domain, abrupt climate shifts are due to changes in the dominant oscillations (as in the case of the MPR), or due to a shift in the phase between different climate signals. As an example, the phase between the Indian Monsoon and ENSO exhibits significant shifts for the past 100 years [59].

The period of regular instrumental records of global climate is relatively short (100–200 years). Even so, this record shows many climatic fluctuations, some abrupt or sudden, as well as slow drifts in climate. Climatic changes become apparent on many temporal and spatial scales. Most abrupt climate changes are regional in their spatial extent. However, regional changes can have remote impacts due to atmospheric and oceanic teleconnections. Some of these shifts may be termed abrupt or sudden in that they represent relatively rapid changes in otherwise comparatively stable conditions, but they can also be found superimposed on other much slower climatic changes.

The definition of "abrupt" or "rapid" climate changes is therefore necessarily subjective, since it depends in large measure on the sample interval used in a particular study and on the pattern of longer-term variation within which the sudden shift is embedded. It is therefore useful to avoid a too general approach, but instead to focus on different types of rapid transitions as they are detected and modeled for different time periods. Although distinctions between types are somewhat arbitrary, together they cover a wide range of shifts in dominant climate mode on time scales ranging from the Cenozoic (the last 65 millions of years) to the recent and future climate.

A Mathematical Definition

Time Domain

Abrupt climate change is characterized by a transition of the climate system into a different state (of temperature, rainfall, and other variables) on a time scale that is faster than variations in the neighborhood (in time). Abrupt climate change could be related to a forcing or internally generated. Consider $x(t) \in \mathbb{R}^n$ as a multi-dimensional climate state variable (temperature, rainfall, and other variables). We define an abrupt climate shift of degree ϵ and amplitude B, if

$$\frac{\mathrm{d}}{\mathrm{d}t} x_i(t) \quad \text{can be approximated by a function} \quad \frac{B}{\pi} \frac{\epsilon}{x_i^2 + \epsilon^2} \tag{3}$$

for one $i \in \{1, \ldots, n\}$ in a time interval $[t_1, t_2]$. The case $\epsilon \to 0$ is called instantaneous climate shift, i. e. $x_i(t)$ can be approximated by the Heaviside step function. The degree of approximation can be specified by a proper norm.

An alternative way of defining an abrupt climate shift is through the identification of probable breaks in a time series (e. g., the surface temperature series). The formulation of a two-phase regression (TPR) test, e. g. [55,79], describing a series $x(t)$ is given by

$$x(t) = \mu_1 + \alpha_1 t + \epsilon(t) \quad \text{for } t \le c \tag{4}$$

$$x(t) = \mu_2 + \alpha_2 t + \epsilon(t) \quad \text{for } t > c\,. \tag{5}$$

Under the null hypothesis of no changepoint, the two phases of the regression should be statistically equivalent and both the difference in means $\mu_{1,2}$, and the difference in slopes, $\alpha_{1,2}$, should be close to zero for each possible changepoint c.

In a stochastic framework one may use an appropriate stochastic differential equation (Langevin equation)

$$\frac{\mathrm{d}}{\mathrm{d}t} x(t) = f(x) + g(x)\xi\,, \tag{6}$$

where ξ is a stationary stochastic process and the functions $f, g\colon \mathbb{R}^n \to \mathbb{R}^n$ describe the climate dynamics. Abrupt climate change can be defined as a transition of a short period of time $[t_1, t_2]$, where the probability of an event is larger than a threshold. The properties of the random force

are described through its distribution and its correlation properties at different times. In the Ornstein–Uhlenbeck process ξ is assumed to have a Gaussian distribution of zero average,

$$\langle \xi(t) \rangle = 0 \tag{7}$$

and to be δ-correlated in time,

$$\langle \xi(t)\xi(t+\tau) \rangle = \delta(\tau) \,. \tag{8}$$

The brackets indicate an average over realizations of the random force. For a Gaussian process only the average and second moment need to be specified since all higher moments can be expressed in terms of the first two. Note that the dependence of the correlation function on the time difference τ assumes that ξ is a stationary process.

The probability density $p(x,t)$ for the variable $x(t)$ in (6) obeys the Fokker–Planck equation

$$\partial_t p = -\frac{\partial}{\partial x}\left[f(x)p\right] + \frac{\partial}{\partial x}\left[g(x)\frac{\partial}{\partial x}\left\{g(x)p\right\}\right] \,. \tag{9}$$

Its stationary probability density of (6) is given by

$$p_{st}(x) = \aleph \exp\left(-2\int_{x_0}^{x} \frac{f(y) - g(y)g'(y)}{g(y)^2}\,\mathrm{d}y\right) \,, \tag{10}$$

where $\aleph$ is a normalization constant. $g'(y)$ stands for the derivative of g with respect to its argument. The extrema x_m of the steady state density obey the equation

$$f(x_m) - g(x_m)g'(x_m) = 0 \tag{11}$$

for $g(x_m) \neq 0$. Here is the crux of the noise-induced transition phenomenon: one notes that this equation is not the same as the equation $f(x_m) = 0$ that determines the steady states of the system in the absence of multiplicative noise. As a result, the most probable states of the noisy system need not coincide with the deterministic stationary states. More importantly, new solutions may appear or existing solutions may be destabilized by the noise. These are the changes in the asymptotic behavior of the system caused by the presence of the noise, e. g. [84].

Climate Variability and Climate Change

The temporal evolution of climate can be expressed in terms of two basic modes: the forced variations which are the response of the climate system to changes in the external forcing $F(x,t)$ (mostly called climate change), and the free variations owing to internal instabilities and feedbacks leading to non-linear interactions among the various components of the climate system [68] (mostly called climate variability). The external causes $F(x,t)$, operate mostly by causing variations in the amount of solar radiation received by or absorbed by the Earth, and comprise variations in both astronomical (e. g. orbital parameters) and terrestrial forcings (e. g. atmospheric composition, aerosol loading). For example, the diurnal and seasonal variations in climate are related to external astronomical forcings operating via solar radiation, while ice ages are related to changes in Earth orbital parameters. Volcanic eruptions are one example of a terrestrial forcing which may introduce abrupt modifications over a period of 2 or 3 years. The more rapid the forcing, the more likely it is that it will cause an abrupt change. The resulting evolution may be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = f(x) + g(x)\xi + F(x,t) \,. \tag{12}$$

$F(x,t)$ is independent of x if the forcing does not depend on climate (external forcing).

The internal free variations within the climate system are associated with both positive and negative feedback interactions between the atmosphere, oceans, cryosphere and biosphere. These feedbacks lead to instabilities or oscillations of the system on all time scales, and can either operate independently or reinforce external forcings. Investigations of the properties of systems which are far from equilibrium show that they have a number of unusual properties. In particular, as the distance from equilibrium increases, they can develop complex oscillations with both chaotic and periodic characteristics. They also may show bifurcation points where the system may switch between various regimes. Under non-equilibrium conditions, local events have repercussions throughout the whole system. These long-range correlations are at first small, but increase with distance from equilibrium, and may become essential at bifurcation points.

When applying (12), different concepts of climate change are in the literature. Quite often, the dynamics is governed by the following stochastic differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = -\frac{\mathrm{d}}{\mathrm{d}x}U(x) + \sigma\xi + F(t) \tag{13}$$

with potential

$$U(x) = a_4x^4 + a_3x^3 + a_2x^2 + a_1x \,. \tag{14}$$

If the potential is quadratic and $F(t) = 0$, the Orstein–Uhlenbeck process is retained. In contrast, a bistable nonlinear system with two minima in $U(x)$ has been assumed in which shifts between the two distinctly different states are triggered randomly by stochastic forcing, e. g. [7]. In

such a system, climate variability and change in the potential can interact due to stochastic resonance [1,7]. Stochastic resonance occurs when the signal-to-noise ratio of a non-linear device is maximized for a moderate value of noise intensity σ. It often occurs in bistable and excitable systems with sub-threshold inputs. For lower noise intensities, the signal does not cause the device to cross threshold, so little signal is passed through it. For large noise intensities, the output is dominated by the noise, also leading to a low signal-to-noise ratio. For moderate intensities, the noise allows the signal to reach threshold, but the noise intensity is not so large as to swamp it.

Strictly speaking, stochastic resonance occurs in bistable systems, when a small periodic force $F(t)$ (which is external) is applied together with a large wide-band stochastic force $\sigma\xi$ (which is internal). The system response is driven by the combination of the two forces that compete/cooperate to make the system switch between the two stable states. The degree of order is related to the amount of periodic function that it shows in the system response. When the periodic force is chosen small enough in order to not make the system response switch, the presence of a non-negligible noise is required for it to happen. When the noise is small very few switches occur, mainly at random with no significant periodicity in the system response. When the noise is very strong a large number of switches occur for each period of the periodic force and the system response does not show remarkable periodicity. Quite surprisingly, between these two conditions, there exists an optimal value of the noise that cooperatively concurs with the periodic forcing in order to make almost exactly one switch per period (a maximum in the signal-to-noise ratio).

Furthermore, non-linear oscillators have been proposed where the timing of the deterministic external forcing is crucial for generating oscillations [51,77,78]. Some aspects of non-equilibrium systems can be found in the climatic system. On the climatological scale, it exhibits abrupt jumps in the long-term rate of temperature change, which are often associated with changes in circulation patterns.

Frequency Domain

In the frequency domain, there are different ways to describe abrupt climate change. A stationary process exhibits an autocovariance function of the form

$$\mathrm{Cov}(\tau) = \langle (x(t+\tau) - \langle x \rangle)(x(t) - \langle x \rangle) \rangle \tag{15}$$

where $\langle\rangle$ denotes the expectation or the statistical mean. Normalized to the variance (i. e. the autocovariance function at $\tau = 0$) one gets the autocorrelation function $C(\tau)$:

$$C(\tau) = \mathrm{Cov}(\tau)/\,\mathrm{Cov}(0)\,. \tag{16}$$

Many stochastic processes in nature exhibit short-range correlations, which decay exponentially:

$$C(\tau) \sim \exp(-\tau/\tau_0)\,, \quad \text{for} \;\; \tau \to \infty\,. \tag{17}$$

These processes exhibit a typical time scale τ_0.

As the frequency domain counterpart of the autocovariance function of a stationary process, one can define the spectrum as

$$S(\omega) = \widehat{\mathrm{Cov}(\tau)}\,, \tag{18}$$

where the hat denotes the Fourier transformation. However, geophysical processes are furthermore often non-stationary. In this regard, the optimal method is continuous wavelet analysis as it intrinsically adjusts the time resolution to the analyzed scale, e. g. [16,59].

Wavelet Spectra A major question concerns the significance testing of wavelet spectra. Torrence and Compo [86] formulated pointwise significance tests against reasonable background spectra. However, Maraun and Kurths [58] pointed out a serious deficiency of pointwise significance testing: Given a realization of white noise, large patches of spurious significance are detected, making it – without further insight – impossible to judge which features of an estimated wavelet spectrum differ from background noise and which are just artifacts of multiple testing. Under these conditions, a reliable corroboration of a given hypothesis is impossible. This demonstrates the necessity to study the significance testing of continuous wavelet spectra in terms of sensitivity and specificity. Given the set of all patches with pointwise significant values, areawise significant patches are defined as the subset of additionally areawise significant wavelet spectral coefficients given as the union of all critical areas that completely lie inside the patches of pointwise significant values. Whereas the specificity of the areawise test appears to be – almost independently of the signal-to-noise ratio – close to one, that of the pointwise test decreases for high background noise, as more and more spurious patches appear [58].

Eigenvalues and Pseudospectrum Another spectral method characterizing the abruptness of climate change is related to the resonance of the linear system (1). As we will see later in the context of atmosphere and ocean instabilities, an eigenvalue analysis is inappropriate in describing the dynamics of the system (12). Inspection of many geo-

physical systems shows that most of the systems fail the normality condition

$$A\,A^{\dagger} = A^{\dagger}\,A\,, \tag{19}$$

where † denotes the adjoint-complex operator. If a matrix is far from normal, its eigenvalues (the spectrum) have little to do with its temporal evolution [71,87]. More about the dynamics can be learned by examining the pseudospectrum of A in the complex plane. The ϵ-pseudospectrum of operator A is defined by two equivalent formulations:

$$\begin{aligned}\Lambda_\epsilon(A) &= \{z \in \mathbb{C} : ||(zI - A)^{-1}|| \geq \epsilon^{-1}\} \\ &= \{z \in \mathbb{C} : [\text{ smallest singular value of } \\ &\qquad (zI - A)] \leq \epsilon\}\,.\end{aligned} \tag{20}$$

This set of values z in the complex plane are defined by contour lines of the resolvent $(zI - A)^{-1}$. The resolvent determines the system's response to a forcing as supplied by external forcing $F(x, t)$, stochastic forcing $g(x)\xi$, or initial/boundary conditions. The pseudospectrum reflects the robustness of the spectrum and provides information about instability and resonance. One theorem is derived from Laplace transformation stating that transient growth is related to how far the ϵ-pseudospectrum extends into the right half plane:

$$||\exp(A\,t)|| \geq \frac{1}{\epsilon} \sup_{z \in \Lambda_\epsilon(A)} \mathrm{Real}(z)\,. \tag{21}$$

In terms of climate theory, the pseudospectrum indicates resonant amplification. Maximal amplification is at the poles of $(zI - A)^{-1}$, characterized by the eigenfrequencies. In a system satisfying (19), the system's response is characterized solely by the proximity to the eigenfrequencies. In the non-normal case, the pseudospectrum shows large resonant amplification for frequencies which are not eigenfrequencies. This transient growth mechanism is important for both initial value and forced problems.

Earth System Modeling and Analysis

Hierarchy of Models

Modeling is necessary to produce a useful understanding of abrupt climate processes. Model analyses help to focus research on possible causes of abrupt climate change, such as human activities; on key areas where climatic thresholds might be crossed; and on fundamental uncertainties in climate-system dynamics. Improved understanding of abrupt climatic changes that occurred in the past and that are possible in the future can be gained through climate models. A comprehensive modeling strategy designed to address abrupt climate change includes vigorous use of a hierarchy of models, from theory and conceptual models through models of intermediate complexity, to high-resolution models of components of the climate system, to fully coupled earth-system models. The simpler models are well-suited for use in developing new hypotheses for abrupt climate change. Model-data comparisons are needed to assess the quality of model predictions. It is important to note that the multiple long integrations of enhanced, fully coupled Earth system models required for this research are not possible with the computer resources available today, and thus, these resources are currently being enhanced.

Feedback

One particularly convincing example showing that the feedbacks in the climate system are important is the drying of the Sahara about 5000 years before present which is triggered by variations in the Earth's orbit around the sun. Numerous modeling studies, e. g. [31], suggest that the abruptness of the onset and termination of the early to mid-Holocene humid period across much of Africa north of the equator depends on the presence of non-linear feedbacks associated with both ocean circulation and changes in surface hydrology and vegetation, e. g. [18]. Without including these feedbacks alongside gradual insolation forcing, it is impossible for existing models to come even close to simulating the rapidity or the magnitude of climatic change associated with the extension of wetlands and plant cover in the Sahara/Sahel region prior to the onset of desiccation around 5000 years before present.

Climate Archives and Modeling

Systematic measurements of climate using modern instruments have produced records covering the last 150 years. In order to reconstruct past variations in the climate system further back in time, scientists use natural archives of climatic and environmental changes, such as ice cores, tree rings, ocean and lake sediments, corals, and historical evidence. Scientists call these records proxies because, although they are not usually direct measures of temperature or other climatic variables, they are affected by temperature, and using modern calibrations, the changes in the proxy preserved in the fossil record can be interpreted in terms of past climate.

Ice core data, coral data, ring width of a tree, or information from marine sediments are examples of a proxy for temperature, or in some cases rainfall, because the thick-

ness of the ring can be statistically related to temperature and/or rainfall in the past. The most valuable proxies are those that can be scaled to climate variables, and those where the uncertainty in the proxy can be measured. Proxies that cannot be quantified in terms of climate or environment are less useful in studying abrupt climate change because the magnitude of change cannot be determined. Quite often, the interpretation of proxy data is already a model of climate change since it involves constraints (dating, representativeness etc.). Uncertainties in the proxies, and uncertainties in the dating, are the main reasons that abrupt climate change is one of the more difficult topics in the field of paleoclimatology.

Example: Glacial-Interglacial Transitions

Astronomical Theory of Ice Ages

Over the past half million years, marine, polar ice core and terrestrial records all highlight the sudden and dramatic nature of glacial terminations, the shifts in global climate that occurred as the world passed from dominantly glacial to interglacial conditions, e. g. [23,69]. These climate transitions, although probably of relatively minor relevance to the prediction of potential future rapid climate change, do provide the most compelling evidence available in the historical record for the role of greenhouse gas, oceanic and biospheric feedbacks as non-linear amplifiers in the climate system. It is such evidence of the dramatic effect of non-linear feedbacks that shows relatively minor changes in climatic forcing may lead to abrupt climate response.

A salient feature of glacial-interglacial climate change is furthermore its asymmetry (Fig. 1). Warmings are rapid, usually followed by slower descent into colder climate. Given the symmetry of orbital forcings $F(t)$, the cause of rapid warming at glacial "terminations" must lie in a climate feedback [37,65]. Clearly, the asymmetric feedback is due to the albedo (reflectance) of ice and snow changing from high values under glacial climates to low values under warm climates. The albedo feedback helps explain the rapidity of deglaciations and their beginnings in spring and summer. Increased absorption of sunlight caused by lower albedo provides the energy for rapid ice melt. The build-up of snow and ice takes much longer than melting.

Many simplified climate models consist of only a few coupled ordinary differential equations controlled by carefully selected parameters. It is generally acknowledged that the "best" models will be those that contain a minimum of adjustable parameters [77] and are robust with respect to changes in those parameters. Rial [72] formulated a logistic-delayed and energy balance model to understand the saw-tooth shape in the paleoclimate record: A fast warming-slow cooling is described by

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = R\left(1 - \frac{x(t-\tau)}{K(t)}\right)x(t-\tau) \tag{22}$$

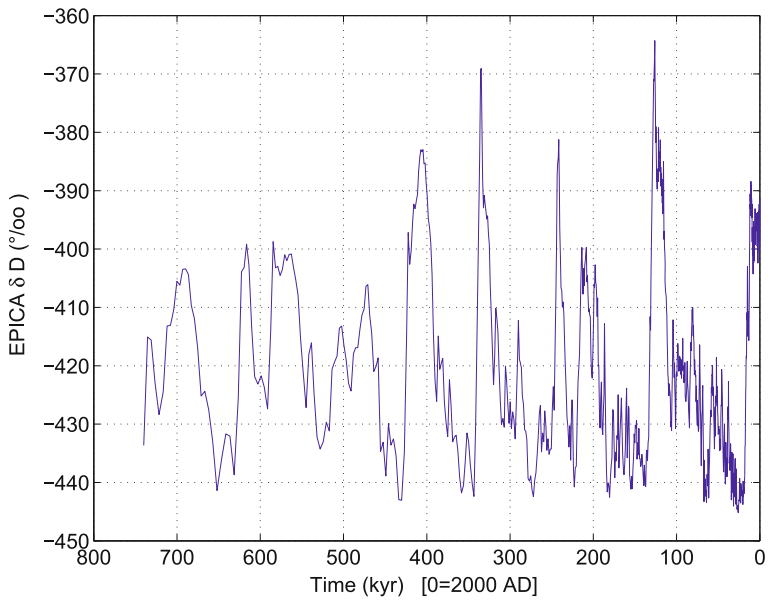

Abrupt Climate Change Modeling, Figure 1
Oxygen isotope record from a southern hemisphere ice core [23] showing the glacial-interglacial changes. Note the asymmetry: the state is longer in the cold (glacials) phases than in the warm phases (interglacials)

$$C\frac{\mathrm{d}}{\mathrm{d}t}T(t) = Q\,(1-\alpha(x)) - (A + BT) \tag{23}$$

with $x(t)$ for the normalized ice extent, τ time delay, $K(t) = 1 + e(t)T(t)$ carrying capacity, $1/R$ response time of the ice sheet, $T(t)$ global mean temperature, $\alpha(x)$ planetary albedo, external parameter $e(t)$, and $R\tau$ bifurcation parameter. A, B, C, Q are constants for the energy balance of the climate. The equation is calibrated so that for $x(t) = 1$ the albedo $\alpha(x) = 0.3$ and $T(t) = 15°\mathrm{C}$. With (23), saw-toothed waveforms and frequency modulation can be understood [72]. The delayed equation yields damped oscillations of $x(t)$ about the carrying capacity for small τ. If τ becomes long compared to the natural response time of the system, the oscillations will become strong, and will grow in amplitude, period and duration. As in the logistic equation for growth, here the product $R\tau$ is a bifurcation parameter, which when crossing the threshold value $\pi/2$ makes the solutions undergo a Hopf bifurcation and settle to a stable limit cycle with fundamental period $\sim 4\tau$ [73].

The astronomical theory of ice ages – also called Milankovitch theory [62] – gained the status of a paradigm for explaining the multi-millennial variability. A key element of this theory is that summer insolation at high latitudes of the northern hemisphere determines glacial-interglacial transitions connected with the waxing and waning of large continental ice sheets, e.g. [33,37], the dominant signal in the climate record for the last million years. Climate conditions of glacials and interglacials are very different. During the Last Glacial Maximum, about 20,000 years before present, surface temperature in the north Atlantic realm was 10–20°C lower than today [13]. A recent study of Huybers and Wunsch [36] has shown that the most simple system for the phase of ice volume $x(t)$ is given by

$$x(t+1) = x(t) + \sigma\xi \tag{24}$$

with ξ a Gaussian white noise process, but with mean $\mu = 1$, and $\sigma = 2$. ξ represents the unpredictable background weather and climate variability spanning all time scales out to the glacial/interglacial. This highly simplified model posits 1-ky steps in ice volume $x(t)$. The non-zero mean biases the Earth toward glaciation. Once $x(t)$ reaches a threshold, a termination is triggered, and ice-volume is linearly reset to zero over 10 ky. The following threshold condition for a termination makes it more likely for a termination of ice volume to occur when obliquity $\Theta(t)$ is large:

$$x(t) \geq T_0 - a\Theta(t)\,. \tag{25}$$

$\Theta(t)$ has a frequency of about 41 ky, and is furthermore normalized to zero mean with unit variance. The other parameters are: amplitude $a = 15$, $T_0 = 105$. Furthermore, the initial ice volume at 700 ky before present is set to $x(t = -700) = 30$. Equation (24) resembles an order-one autoregressive process, similar to (2), plus the threshold

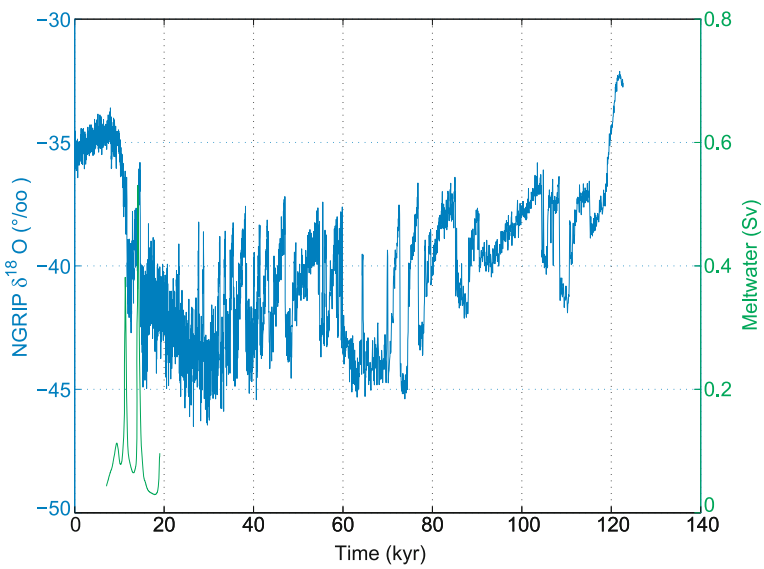

Abrupt Climate Change Modeling, Figure 2
Oxygen isotope record from a Greenland ice core record [64] using an updated time scale for this record [23]. *Green*: Sea-level derived rate of deglacial meltwater discharge [24] which is strong after deglacial warming

condition (25). Models like (24), (25) are not theories of climate change, but rather attempts at efficient kinematic descriptions of the data, and different mechanisms can be consistent with the limited observational records. In the next section, the process of deglaciation is modeled in a three-dimensional model including the spatial dimension.

Deglaciation

The question is what causes the abrupt warming at the onset of the Boelling as seen in the Greenland ice cores (Fig. 2). There is a clear antiphasing seen in the deglaciation interval between 20 and 10 ky ago: During the first half of this period, Antarctica steadily warmed, but little change occurred in Greenland. Then, at the time when Greenland's climate underwent an abrupt warming, the warming in Antarctica stopped. Knorr and Lohmann [42], also summarizing numerous modeling studies for deglaciation, describe how global warming (which may be induced by greenhouse gases and feedbacks) can induce a rapid intensification of the ocean circulation (Fig. 3). During the Boelling/Alleroed, a sudden increase of the northward heat transport draws more heat from the south, and leads to a strong warming in the north. This "heat piracy" from the South Atlantic has been formulated by Crowley [15]. A logical consequence of this heat piracy is the Antarctic Cold Reversal (ACR) during the Northern Hemisphere warm Boelling/Alleroed. This particular example shows that an abrupt climate change of the ocean circulation (with large climate impacts in the North Atlantic) is related to a smooth global warming. To understand the dynamical behavior of the system, the concept of hysteresis is applied, using the global warming after the last ice ages as the control parameter [42]. The system exhibits multiple steady states (Fig. 4): a weak glacial ocean circulation and a stronger circulation (which is comparable in strength to the modern mode of operation). Deglacial warming induces a transition from a weak glacial THC state to a stronger THC state, characterizing the abrupt warming during the deglaciation.

Millennial Climate Variability

Within glacial periods, and especially well documented during the last one, spanning from around 110 to 11.6 ky ago, there are dramatic climate oscillations, including high-latitude temperature changes approaching the same magnitude as the glacial cycle itself, recorded in archives from the polar ice caps, high to middle latitude marine sediments, lake sediments and continental loess sections. These oscillations are usually referred to as the Dansgaard–Oeschger Cycle and occur mostly on 1 to 2 ky time scales, e. g. [6], although regional records of these transitions can show much more rapid change. The termination of the Younger Dryas cold event, for example, is manifested in ice core records from central Greenland as a near doubling of snow accumulation rate and a temperature shift of around 10°C occurring within a decade with world-wide teleconnections. One hypothesis for explaining these climatic transitions is that the ocean thermohaline circulation flips between different modes, with warm intervals reflecting periods of strong deep water formation in the northern North Atlantic and vice versa [29]. As an alternative approach, one can estimate the underlying dynamics (13), (14) directly from data [43]. The method is based on the unscented Kalman filter, a non-linear extension of the conventional Kalman filter. This technique allows one to consistently estimate parameters in deterministic and stochastic non-linear models. The optimiza-

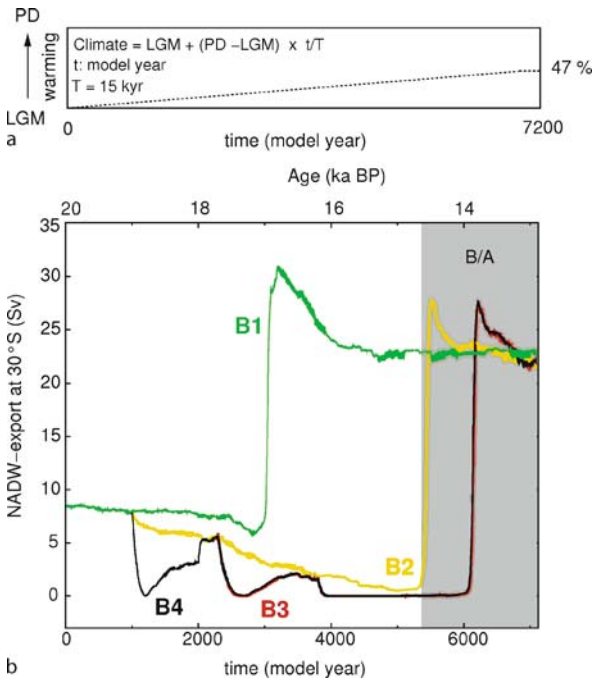

Abrupt Climate Change Modeling, Figure 3
Forcing and model response of the ocean overturning rate. a The background climate conditions are linearly interpolated between glacial (LGM), and modern (PD), conditions. Gradual warming is stopped after 7000 model years, which is related to ∼ 47% of the total warming. b Circulation strength (export at 30°S) versus time. The *green curve* B1 represents the experiment without any deglacial freshwater release to the North Atlantic. Experiments B2 (*yellow curve*), B3 (*red curve*), and B4 (*black curve*), exhibit different successions of deglacial meltwater pulse scenarios to the North Atlantic [42]

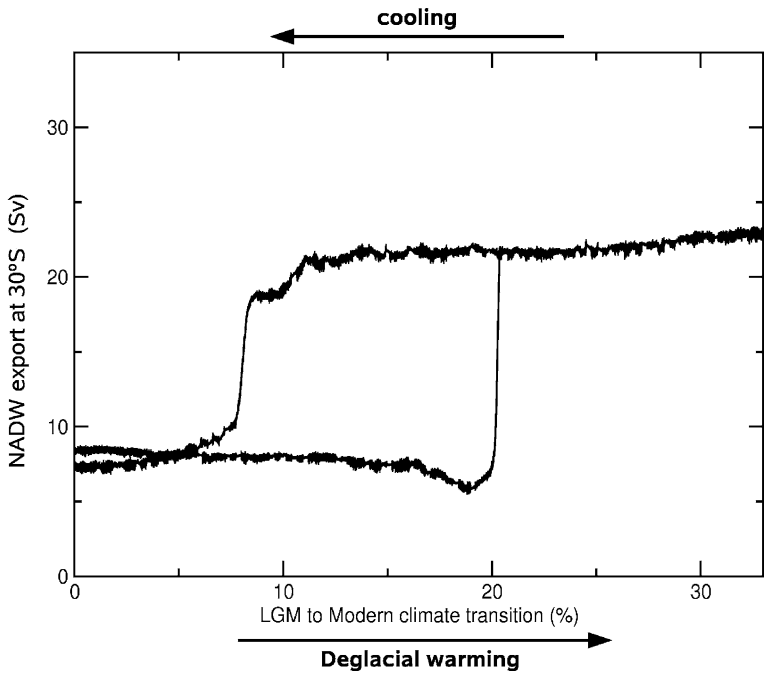

Abrupt Climate Change Modeling, Figure 4
Hysteresis loop of the ocean overturning strength (*black curve*) with respect to slowly varying climate background conditions. The transition values are given in % of a full glacial-interglacial transition [42]

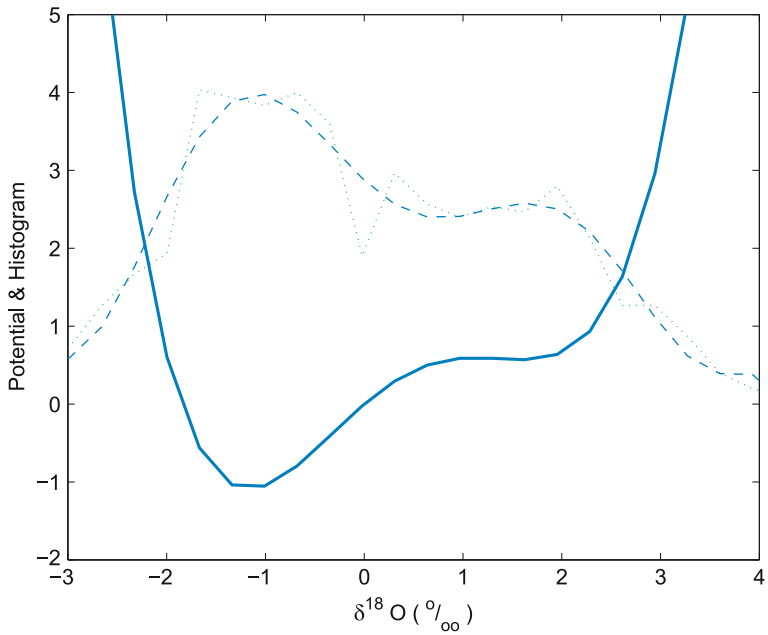

Abrupt Climate Change Modeling, Figure 5
Potential derived from the data (*solid*) together with probability densities of the model (*dashed*) and the data (*dotted*)

tion yields for the coefficients $a_4 = 0.13 \pm 0.01$, $a_3 = -0.27 \pm 0.02$, $a_2 = -0.36 \pm 0.08$, and $a_1 = 1.09 \pm 0.23$. The dynamical noise level of the system σ is estimated to be 2.4. The potential is highly asymmetric and degenerate (that is, close to a bifurcation): there is one stable cold stadial state and one indifferently stable warm interstadial state (Fig. 5). This seems to be related to the fact that the warm intervals are relatively short-lasting.

Coming back to the ice cores and a potential linkage of the hemispheres, Stocker and Johnson [81] proposed a conceptual model linking the isotopic records from Antarctica and Greenland. The basis is an energy balance with temperatures in the North and South Atlantic Ocean, as well as a "southern heat reservoir". It is assumed that the change in heat storage of a "southern heat reservoir" T_S is given by the temperature difference between the reservoir T_S and the Southern Ocean temperature T, with a characteristic time scale τ:

$$\frac{\mathrm{d}}{\mathrm{d}t} T_S(t) = \frac{1}{\tau} \left[T - T_S \right] \ . \tag{26}$$

T_N denotes the time-dependent temperature anomaly of the North Atlantic. The Southern Ocean temperature T is assumed to be $-T_N$ according to the bipolar seesaw (North Atlantic cold $\leftrightarrow$ South Atlantic warm). Using Laplace transform, one can solve for T_S

$$T_S = -\frac{1}{\tau} \int_0^t T_N(t-t') \exp(-t'/\tau) \mathrm{d}t' + T_S(0) \exp(-t/\tau) . \tag{27}$$

The reservoir temperature is therefore a convolution of the northern temperature using the time scale τ ranging from 100 to 4000 years. Equation (27) demonstrates that T_S and T_N will have entirely different time characteristics. Abrupt changes in the north appear damped and integrated in time in the southern reservoir. A sudden reduction in the thermohaline circulation causes a cooling in the North Atlantic and a warming in the South, a situation similar to the Younger Dryas period [80], see also Fig. 2.

Example: Cenozoic Climate Cooling

Antarctic Glaciation

During the Cenozoic (65 million years ago (Ma) to present), there was the widespread glaciation of the Antarctic continent at about 34 Ma, e. g. [93]. Antarctic glaciation is the first part of a climate change from relatively warm and certainly ice-free conditions to massive ice sheets in both, the southern and northern hemispheres [44]. Opening of circum-Antarctic seaways is one of the factors that have been ascribed as a cause for Antarctic climate change so far [40,93]. Besides gateway openings, the atmospheric carbon dioxide concentration is another important factor affecting the evolution of the Cenozoic climate [17,93]. As a third component in the long-term evolution of Antarctic glaciation, land topography is able to insert certain thresholds for abrupt ice sheet build-up. Whereas tectonics, land topography, and long-term Cenozoic CO_2-decrease act as preconditioning for Antarctic land ice formation, the cyclicities of the Earth's orbital configuration are superimposed on shorter time scales and may have served as the ultimate trigger and pacemaker for ice-sheet growth at the Eocene-Oligocene boundary around 34 Ma [14].

DeConto and Pollard [17] varied Southern Ocean heat transport to mimic gateway opening instead of an explicit simulation of ocean dynamics. They found a predominating role of pCO_2 in the onset of glaciation instead of a dominating tectonic role for "thermal isolation".

Mid-Pleistocene Revolution

Glaciation in the Northern Hemisphere lagged behind, with the earliest recorded glaciation anywhere in the Northern Hemisphere occurring between 10 and 6 Ma and continuing through to the major increases in global ice volume around 2–3 Ma [60]. A recent compilation of 57 globally distributed records [45] is shown in Fig. 6. Let us focus now on the mid-Pleistocene transition or revolution (MPR), describing the transition from 41 ky to 100 ky glacial-interglacial cycles.

Milankovitch [62] initially suggested that the critical factor was total summer insolation at about 65°N, because for an ice sheet to grow some additional ice must survive each successive summer. In contrast, the Southern Hemisphere is limited in its response because the expansion of ice sheets is curtailed by the Southern Ocean around Antarctica. The conventional view of glaciation is thus that low summer insolation in the temperate North Hemisphere allows ice to survive summer and thus start to build up on the northern continents. If so, how then do we account for the MPR? Despite the pronounced change in Earth system response evidenced in paleoclimatic records, the frequency and amplitude characteristics of the orbital parameters which force long-term global climate change, e. g., eccentricity ($\sim$ 100 ky), obliquity ($\sim$ 41 ky) and precession ($\sim$ 21 and $\sim$ 19 ky), do not vary during the MPR [8]. This suggests that the cause of change in response at the MPR is internal rather than external to the global climate system.

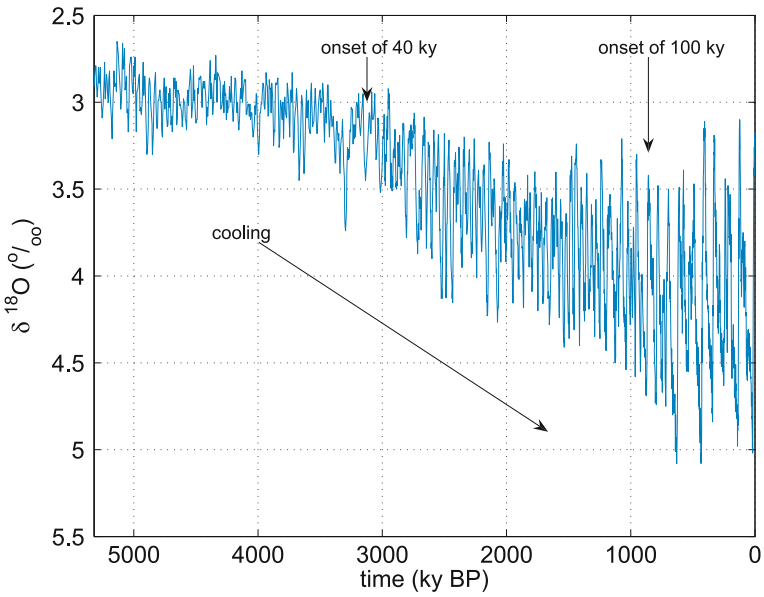

Abrupt Climate Change Modeling, Figure 6
A compilation of 57 globally distributed records by Lisiecki and Raymo [45]: The $\delta^{18}O$ record reflects mainly the climate variables temperature and ice volume

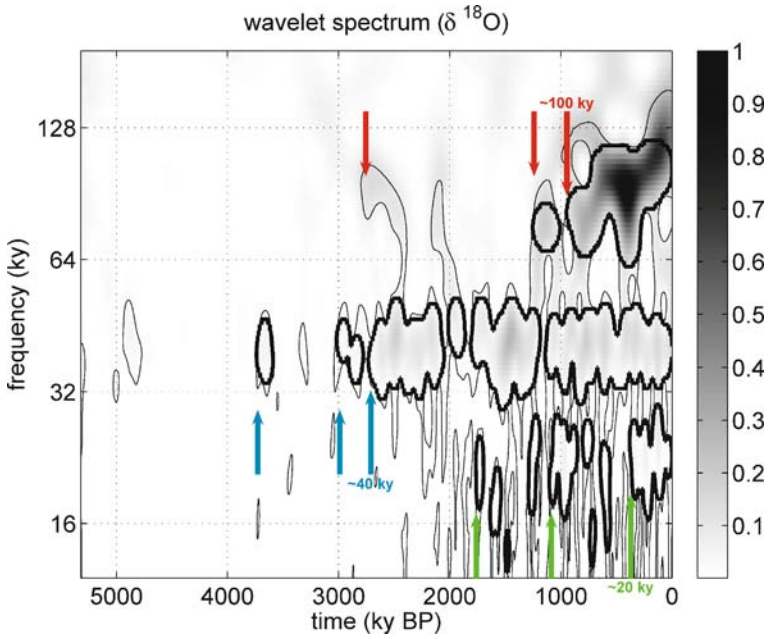

Abrupt Climate Change Modeling, Figure 7
Lisiecki and Raymo [45]: The corresponding wavelet sample spectrum calculated using Morlet wavelet with $\omega_0 = 6$. *Thin* and *thick lines* surround pointwise and areawise significant patches, respectively

The result of a wavelet spectral analysis (Fig. 7) suggests several abrupt climate changes in the frequency domain (shown as schematic arrows in the figure). These abrupt climate shifts represent major reorganizations in the climate system. Some of them are possibly linked to the development of Northern Hemisphere ice volume. The MPR marked a prolongation to and intensification of the $\sim$ 100 ky glacial-interglacial climate. Not only does

the periodicity of glacial-interglacial cycles increase going through the MPR, but there is also an increase in the amplitude of global ice volume variations.

It is likely that the MPR is a transition to a more intense and prolonged glacial state, and associated subsequent rapid deglaciation becomes possible. The first occurrence of continental-scale ice sheets, especially on Greenland, is recorded as ice-rafted detritus released from drifting icebergs into sediments of the mid- and high-latitude ocean. After a transient precursor event at 3.2 Ma, signals of large-scale glaciations suddenly started in the subpolar North Atlantic in two steps, at 2.9 and 2.7 Ma, e. g. [5].

The ice volume increase may in part be attributed to the prolonging of glacial periods and thus of ice accumulation. The amplitude of ice volume variation is also accentuated by the extreme warmth of many interglacial periods. Thus, a colder climate with larger ice sheets should have the possibility of a greater sudden warming [45]. The MPR therefore marks a dramatic sharpening of the contrast between warm and cold periods. Note however, that the amount of energy at the 40 ka period is hardly changed in the time after 1 Ma, and notably, one sees the addition of energy at longer periods, without any significant reduction in obliquity-band energy. After about 1 Ma, large glacial-interglacial changes begin to occur on an approximately 100 ka time scale (but not periodically) superimposed upon the variability which continues largely unchanged [91]. Why did 100 ka glacial-interglacials also become possible in addition to the ice volume variability? Lowering of global CO_2 below some critical threshold, or changes in continental configuration, or atmospheric circulation patterns, or all together, are among the conceivable possibilities, e. g. [70].

Examples: Transient Growth

The former examples show the power of the combination of models, data analysis, and interpretation for abrupt climate change. In the next two examples, it is shown how important the transient growth mechanism is for abrupt climate change.

Conceptual Model of the Ocean Circulation

In this section, a category of the non-linear models following the simple thermohaline model of Stommel [82] is analyzed. The common assumption of these box models is that the oceanic overturning rate Φ can be expressed by the meridional density difference:

$$\Phi = -c(\alpha\Delta T - \beta\Delta S)\,, \tag{28}$$

where α and β are the thermal and haline expansion coefficients, c is a tunable parameter, and Δ denotes the meridional difference operator applied to the variables temperature T and salinity S, respectively. Stommel [82] considered a two-box ocean model where the boxes are connected by an overflow at the top and a capillary tube at the bottom, such that the capillary flow is directed from the high-density vessel to the low-density vessel following (28).

The equations for temperature T and salinity S are the heat and salt budgets using an upstream scheme for the advective transport and fluxes with the atmosphere:

$$\frac{\mathrm{d}}{\mathrm{d}t}T = -\frac{\Phi}{V}\Delta T - \frac{F_{\mathrm{oa}}}{\rho_0 c_p h} \tag{29}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}S = -\frac{\Phi}{V}\Delta S - \frac{S_0}{h}(P-E)\,, \tag{30}$$

where V is the volume of the box with depth h, and $(P-E)$ denotes the freshwater flux (precipitation minus evaporation plus runoff). F_{oa} is the heat flux at the ocean-atmosphere interface, S_0 is a reference salinity, and $\rho_0 c_p$ denotes the heat capacity of the ocean.

Denoting furthermore $x \in \mathbb{R}^2$ for the anomalies of $(\Delta T, \Delta S)$, Lohmann and Schneider [48] have shown the evolution equation is of the following structure:

$$\frac{\mathrm{d}}{\mathrm{d}t}x = Ax + \langle b|x\rangle\, x\,. \tag{31}$$

The brackets $\langle|\rangle$ denote the Euclidean scalar product. This evolution Equation (31) can be transferred to a

$$x(t) = \frac{1}{\gamma(t)}\exp(At)x_0\,, \tag{32}$$

with a scaling function $\gamma(t, x_0)$. The models of Stommel [82], and many others are of this type, and their dynamics are therefore exactly known.

It is worth knowing that (29), (30) is equivalent to the multi-dimensional Malthus–Verhulst model (also known as the logistic equation), which was originally proposed to describe the evolution of a biological population. Let x denote the number (or density) of individuals of a certain population. This number will change due to growth, death, and competition. In the simplest version, birth and death rates are assumed proportional to n, but accounting for limited resources and competition it is modified by $(1-x)$:

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = a(1-x)x\,. \tag{33}$$

In climate, the logistic equation is important for Lorenz's [52] error growth model: where $x(t)$ is the algebraic forecast error at time t and a is the linear growth rate.

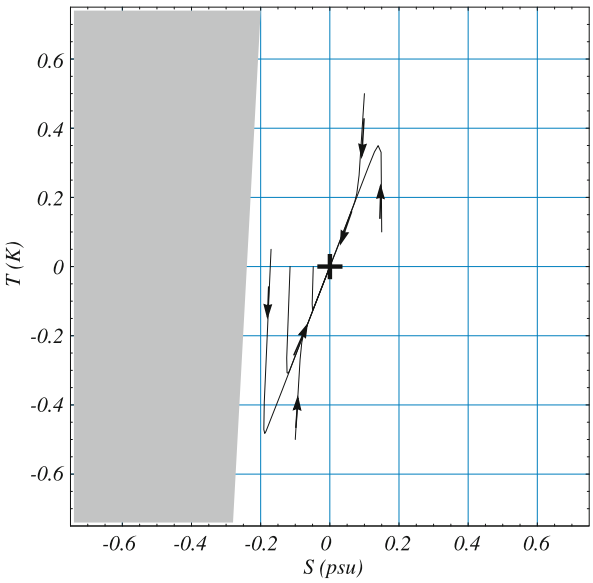

Abrupt Climate Change Modeling, Figure 8
The basin of attraction (*white area*) and the dynamics in the thermohaline phase space. With initial conditions outside the gray area, the trajectories converge asymptotically to the origin corresponding to the thermally driven solution of the THC. Due to the fast thermal response during the first decade of relaxation, the distance of the trajectories from zero can increase temporarily

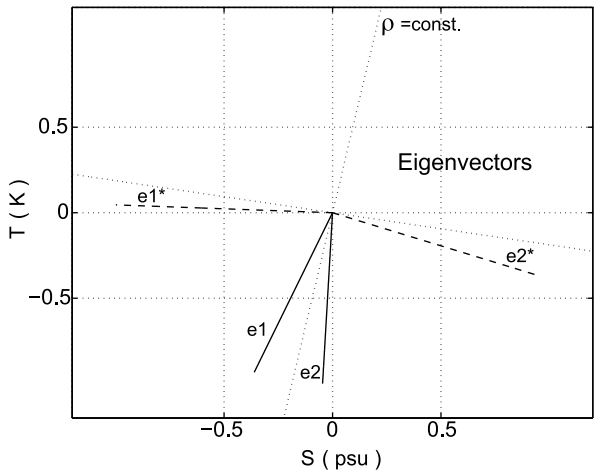

Abrupt Climate Change Modeling, Figure 9
Eigenvectors e_1, e_2, and adjoint eigenvectors e_1^*, e_2^* of the tangent linear operator $A^\dagger$. The *dotted lines* show the line of constant density and the perpendicular

It is useful to analyze the dynamics in the phase space spanned by temperature and salinity anomalies and investigate the model sensitivity under anomalous high latitude forcing, induced as an initial perturbation. The lines in Fig. 8 are phase space trajectories after perturbations of different magnitude have been injected into the North Atlantic. We notice that for most trajectories, the distances from zero $(0, 0)$ increase temporarily, where the maximal distance from zero is after a decade. After about 10 years the trajectories in Fig. 8 point into a "mixed temperature/salinity direction", denoted further as e_1.

Figure 8 implies that the adjustment of the THC involves two phases: A fast thermal response and a slower response on the e_1-direction. The vector e_1 is identical with the most unstable mode in the system. Because the scaling function $\gamma(t)$ acts upon both temperature and salinity (32), the evolution of the non-linear model can be well characterized by the eigenvectors of the matrix A, which is discussed in the following.

In our system, the operator A of the box model is found to be non-normal, and the eigenvectors are not orthogonal. One eigenvalue (e_2) is closely related to temperature anomalies, whereas the other (e_1) is a "mixed temperature/salinity eigenvector" (Fig. 9). The eigenvectors of the adjoint matrix $A^\dagger$ are denoted by e_1^* and e_2^*, respectively. For the non-normal matrix A, the eigenvectors of A and $A^\dagger$ do not coincide, but fulfill the "biorthogonality condition":

$$e_1^* \perp e_2 \quad \text{and} \quad e_2^* \perp e_1 \,. \tag{34}$$

Both eigenvalues $\lambda_{1,2}$ are real and negative. Because of $\lambda_2 < \lambda_1$, the first term dominates for long time scales and the second for short time scales. Using the biorthogonality condition, we get furthermore the coefficients

$$c_i = \frac{\langle e_i^* | x_0 \rangle}{\langle e_i^* | e_i \rangle} \qquad \text{for} \quad i = 1, 2 \,. \tag{35}$$

A perturbation is called "optimal", if the initial error vector has minimal projection onto the subspace with fastest decaying perturbations, or equivalently if the coefficient c_1 is maximal. This is according to (35) equivalent to x_0 pointing into the direction of e_1^*. This unit vector e_1^* is called the "biorthogonal" [66] to the most unstable eigenvector e_1 which we want to excite. In order to make a geometrical picture for the mathematical considerations, assume that the tail of the vector x_0 is placed on the e_1-line and its tip on the e_2-line. This vector is stretched maximally because the tail decays to zero quickly, whereas the tip is hardly unchanged due to the larger eigenvalue λ_1. The most unstable mode e_1 and its biorthogonal e_1^* differ greatly from each other, and the perturbation that optimally excites the mode bears little resemblance to the mode itself.

It is remarkable that the optimal initial perturbation vector e_1^* does not coincide with a perturbation in sea surface density at high latitudes, which would reside on the dotted line perpendicular to $\rho = \text{const}$ in Fig. 9. Even when using a space spanned by $(\alpha T, \beta S)$ instead of (T, S), to take into account the different values for the thermal and haline expansion coefficients, vector e_1^* is much more dominated by the scaled salinity anomalies than the temperature anomalies of the high latitudinal box.

Numerical simulations by Manabe and Stouffer [57] showed, for the North Atlantic, that between two and four times the preindustrial CO_2 concentration, a threshold value is passed and the thermohaline circulation ceases completely. One other example of early Holocene rapid climate change is the "8200-yr BP" cooling event recorded in the North Atlantic region possibly induced by freshwater. One possible explanation for this dramatic regional cooling is a shutdown in the formation of deep water in the northern North Atlantic due to freshwater input caused by catastrophic drainage of Laurentide lakes [4,46]. The theoretic considerations and these numerical experiments suggest that formation of deep water in the North Atlantic is highly sensitive to the freshwater forcing.

Resonance in an Atmospheric Circulation Model

An atmospheric general circulation model PUMA [26] is applied to the problem. The model is based on the multi-level spectral model described by Hoskins and Simmons [35]. For our experiments we chose five vertical levels and a T21 horizontal resolution. PUMA belongs to the class of models of intermediate complexity [12]; it has been used to understand principle feedbacks [56], and dynamics on long time scales [76]. For simplicity, the equations are scaled here such that they are dimensionless. The model is linearized about a zonally symmetric mean state providing for a realistic storm track at mid-latitudes [27]. In a simplified version of the model and calculating the linear model A with $n = 214$, one can derive the pseudospectrum. Figure 10 indicates resonances besides the poles (the eigenvalues) indicated by crosses. The $\mathrm{Im}(z)$-axis shows the frequencies, the $\mathrm{Re}(z)$-axis the damping/amplification of the modes. Important modes for the climate system are those with $-0.5 < \mathrm{Im}(z) < 0.5$ representing planetary Rossby waves. The basic feature is that transient growth of initially small perturbations can occur even if all the eigenmodes decay exponentially. Mathematically, an arbitrary matrix A can be decomposed as a sum

$$A = D + N \tag{36}$$

where A is diagonalizable, and N is nilpotent (there exists an integer $q \in \mathbb{N}$ with $N^q = 0$), and D commutes with N (i. e. $DN = NA$). This fact follows from the Jordan–Chevalley decomposition theorem. This means we can compute the exponential of (A t) by reducing to the cases:

$$\exp(At) = \exp((D + N)t) = \exp(Dt)\exp(Nt) \tag{37}$$

where the exponential of Nt can be computed directly from the series expansion, as the series terminates after a finite number of terms. Basically, the number $q \in \mathbb{N}$ is related to the transient growth of the system ($q = 1$ means no transient growth).

The resonant structures are due to the mode interaction: It is not possible to change one variable without the others, because they are not orthogonal. Interestingly, one can also compute the $A^\dagger$ model, showing the optimal perturbation of a mode e_i through its biorthogonal vector (35).

The analysis indicates that non-normality of the system is a fundamental feature of the atmospheric dynamics. This has consequences for the error growth dynamics, and instability of the system, e. g. [48,66]. Similar features are obtained in shear flow systems [71,87] and other hydrodynamic applications. This transient growth mechanism is important for both initial value and forced problems of the climate system.

Future Directions

Until now, details of abrupt climate change are not well known to accurately predict it. With better information, the society could take more confident action to reduce the potential impact of abrupt changes on agriculture, water resources, and the built environment, among other impacts. A better understanding of sea-ice and glacier stability, land-surface processes, and atmospheric and oceanic circulation patterns is needed. Moreover, to effectively use any additional knowledge of these and other physical processes behind abrupt climate change, more sophisticated ways of assessing their interactions must be developed, including:

Better models. At present, the models used to assess climate and its impacts cannot simulate the size, speed, and extent of past abrupt changes, let alone predict future abrupt changes. Efforts are needed to improve how the mechanisms driving abrupt climate change are represented in these models and to more rigorously test models against the climate record.

More theory. There are concepts to find the underlying dynamical system, to derive a theory from a high-order to low-order description similar to what is done in

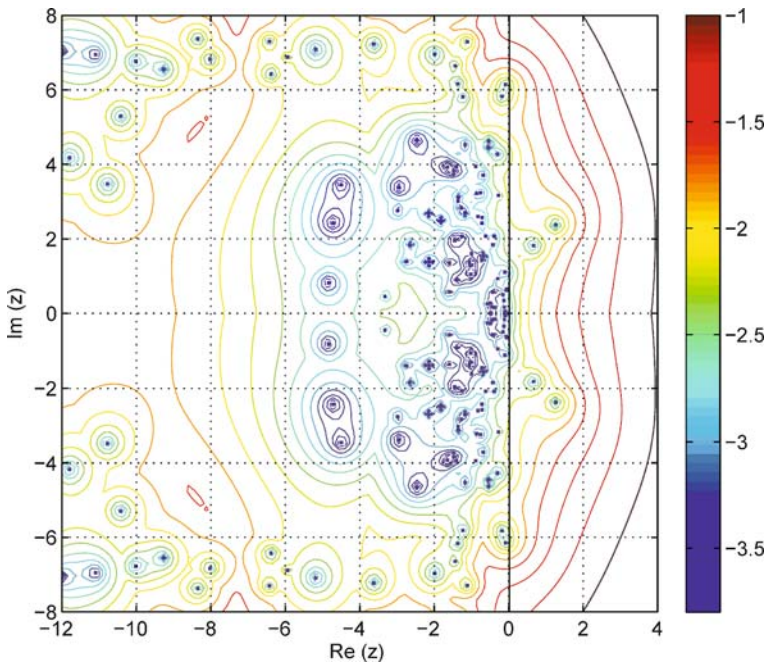

Abrupt Climate Change Modeling, Figure 10

Contours of $\log_{10}(1/\epsilon)$. The figure displays resonant structures of the linearized atmospheric circulation model. The modes extend to the right half plane and are connected through resonant structures, indicating the transient growth mechanism inherent in atmospheric dynamics

statistical physics (Mori–Zwanzig approach [63,94], Master equation), or in stochastic differential equations. A systematic reduction of the complex system into fewer degrees of freedom shall bring a deeper level of understanding about the underlying physics. A systematic approach was suggested by Saltzman [77]. Spectral and pseudo-spectral concepts have not been used too much in climate theory. There is a variety of phenomenological stochastic models in which non-linearity and fluctuations coexist, and in which this coexistence leads to interesting phenomena that would not arise without the complex interplay.

Paleoclimatic data. More climate information from the distant past would go a long way toward strengthening our understanding of abrupt climate changes and models of past climate. In particular, an enhanced effort is needed to expand the geographic coverage, temporal resolution, and variety of paleoclimatic data. Although the present climate has no direct analogon to the past [54], the dynamical interpretation of data will improve the understanding of thresholds and non-linearities in the Earth system.

Appropriate statistical tools. Because most statistical calculations at present are based on the assumption that climate is not changing but is stationary, they have limited value for non-stationary (changing) climates and for climate-related variables that are often highly skewed by rapid changes over time such as for abrupt-change regimes. Available statistical tools themselves need to be adapted or replaced with new approaches altogether to better reflect the properties of abrupt climate change.

Synthesis. Physical, ecological, and human systems are complex, non-linear, dynamic and imperfectly understood. Present climate change is producing conditions outside the range of recent historical experience and observation, and it is unclear how the systems will interact with and react to further climate changes. Hence, it is crucial to be able to better understand and recognize abrupt climate changes quickly. This capability will involve improved monitoring of parameters that describe climatic, ecological, and economic systems. Some of the desired data are not uniquely associated with abrupt climate change and, indeed, have broad applications. Other data take on particular importance because they concern properties or regions implicated in postulated mechanisms of abrupt climate change. Research to increase our understanding of abrupt climate change should be designed specifically within the context of the various mechanisms thought to be involved. Focus is required to provide data for process studies from key regions where triggers of abrupt climate change are likely to occur, and to obtain reliable time series of climate indicators that play crucial roles in the postu-

lated mechanisms. Observations could enable early warning of the onset of abrupt climate change. New observational techniques and data-model comparisons will also be required.

Bibliography

Primary Literature

1. Alley RB, Anandakrishnan S, Jung P (2001) Stochastic resonance in the North Atlantic. Paleoceanogr 16:190–198
2. Alley RB, Marotzke J, Nordhaus W, Overpeck J, Peteet D, Pielke R Jr, Pierrehumbert R, Rhines P, Stocker T, Talley L, Wallace JM (2002) Abrupt Climate Change: Inevitable Surprises. US National Academy of Sciences, National Research Council Committee on Abrupt Climate Change, National Academy Press, Washington
3. Alverson K, Oldfield F (2000) Abrupt Climate Change. In: Joint Newsletter of the Past Global Changes Project (PAGES) and the Climate Variability and Predictability Project (CLIVAR), vol 8, no 1. Bern, pp 7–10
4. Barber DC, Dyke A, Hillaire-Marcel C, Jennings AE, Andrews JT, Kerwin MW, Bilodeau G, McNeely R, Southon J, Morehead MD, Gagnonk JM (1999) Forcing of the cold event of 8,200 years ago by catastrophic drainage of Laurentide lakes. Nature 400:344–348
5. Bartoli G, Sarnthein M, Weinelt M, Erlenkeuser H, Garbe-Schönberg D, Lea DW (2005) Final closure of Panama and the onset of northern hemisphere glaciation. Earth Planet Sci Lett 237:33–44
6. Bender M, Malaize B, Orchardo J, Sowers T, Jouzel J (1999) Mechanisms of Global Climate Change. Clark P et al (eds) AGU 112:149–164
7. Benzi R, Parisi G, Sutera A, Vulpiani A (1982) Stochastic resonance in climatic change. Tellus 34:10
8. Berger A, Loutre MF (1991) Insolation values for the climate of the last 10 million years. Quat Sci Rev 10:297–317
9. Berger WH, Labeyrie LD (1987) Abrupt Climatic Change, Evidence and Implications. NATO ASI Series, Series C, Mathematical and Physical Sciences, vol 216. D Reidel, Dordrecht, pp 425
10. Broecker WS et al (1985) Does the Ocean-atmosphere System Have More than One Stable Mode of Operation? Nature 315:21–26
11. Bryan F (1986) High Latitude Salinity Effects and Inter-hemispheric Thermohaline Circulations. Nature 323:301–304
12. Claussen M, Mysak LA, Weaver AJ, Crucifix M, Fichefet T, Loutre M-F, Weber SL, Alcamo J, Alexeev VA, Berger A, Calov R, Ganopolski A, Goosse H, Lohmann G, Lunkeit F, Mokhov II, Petoukhov V, Stone P, Wang Z (2002) Earth System Models of Intermediate Complexity: Closing the Gap in the Spectrum of Climate System Models. Clim Dyn 18:579–586
13. CLIMAP project members (1976) The surface of the ice age Earth. Science 191:1131–1137
14. Coxall HK, Wilson PA, Pälike H, Lear CH, Backman J (2005) Rapid stepwise onset of Antarctic glaciation and deeper calcite compensation in the Pacific Ocean. Nature 433:53–57. doi:10.1038/nature03135
15. Crowley TJ (1992) North Atlantic deep water cools the southern hemisphere. Paleoceanogr 7:489–497
16. Daubechies I (1992) Ten Lectures on Wavelets. Society for Industrial and Applied Mathematics (SIAM). CBMS-NSF Regional Conference Series in Applied Mathematics, vol 61, Philadelphia
17. DeConto RM, Pollard D (2003) Rapid Cenozoic glaciation of Antarctica induced by declining atmospheric CO2. Nature 421:245–249. doi:10.1038/nature01290
18. DeMenocal et al (2000) Abrupt onset and termination of the African Humid Period: Rapid climate response to gradual insolation forcing. Quat Sci Rev 19:347–361
19. Diaz HF, Quayle RG (1980) The climate of the United States since 1895: spatial and temporal changes. Mon Wea Rev 108:149–226
20. Dijkstra HA, Te Raa L, Weijer W (2004) A systematic approach to determine thresholds of the ocean's thermohaline circulation. Tellus 56A(4):362
21. Dima M, Lohmann G (2002) Fundamental and derived modes of climate variability. Application to biennial and interannual timescale. Tellus 56A:229–249
22. Dima M, Lohmann G (2007) A hemispheric mechanism for the Atlantic Multidecadal Oscillation. J Clim 20:2706–2719
23. EPICA Community Members (2006) One-to-one coupling of glacial climate variability in Greenland and Antarctica. Nature 444:195–198. doi:10.1038/nature05301
24. Fairbanks RG (1989) A 17,000-year glacio-eustatic sea level record: influence of glacial melting rates on the Younger Dryas event and deep-ocean circulation. Nature 342:637–642
25. Flohn H (1986) Singular events and catastrophes now and in climatic history. Naturwissenschaften 73:136–149
26. Fraedrich K, Kirk E, Lunkeit F (1998) Portable University Model of the Atmosphere. DKRZ Report 16, Hamburg
27. Frisius T, Lunkeit F, Fraedrich K, James IN (1998) Storm-track organization and variability in a simplified atmospheric global circulation model. Q J R Meteorol Soc 124:1019–1043
28. Fu C, Diaz HF, Dong D, Fletcher JO (1999) Changes in atmospheric circulation over northern hemisphere oceans associated with the rapid warming of the 1920s. Int J Climatol 19(6):581–606
29. Ganopolski A, Rahmstorf S (2001) Rapid changes of glacial climate simulated in a coupled climate model. Nature 409: 153–158
30. Ganopolski A, Rahmstorf S (2002) Abrupt glacial climate changes due to stochastic resonance. Phys Rev Let 88(3):038501
31. Ganopolski A, Kubatzki C, Claussen M, Brovkin V, Petoukhov V (1998) The influence of vegetation-atmosphere-ocean interaction on climate during the mid-Holocene. Science 280:1916
32. Hasselmann K (1976) Stochastic climate models, Part 1, Theory. Tellus 28:289–485
33. Hays JD, Imbrie J, Shackleton NJ (1976) Variations in the Earth's Orbit: Pacemaker of the Ice Ages. Science 194:1121–1132
34. Henderson GM, Slowey NC (2000) Evidence from U-Th dating against Northern Hemisphere forcing of the penultimate deglaciation. Nature 404:61–66
35. Hoskins BJ, Simmons AJ (1975) A multi-layer spectral model and the semi-implicit method. Q J R Meteorol Soc 101: 1231–1250
36. Huybers P, Wunsch C (2005) Obliquity pacing of the late Pleistocene glacial terminations. Nature 434:491–494. doi:10.1038/nature03401
37. Imbrie J, Imbrie JZ (1980) Modeling the climatic response to orbital variations. Science 207:943–953

38. IPCC (2007) Summary for Policymakers. In: Climate change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge and New York
39. Iwashima T, Yamamoto R (1986) Time-space spectral model of low order barotropic system with periodic forcing. J Meterol Soc Jpn 64:183–196
40. Kennett JP, Houtz RE, Andrews PB, Edwards AE, Gostin VA, Hajos M, Hampton M, Jenkins DG, Margolis SV, Ovenshine AT, Perch-Nielsen K (1974) Development of the circum-Antarctic current. Science 186:144–147
41. Knorr G, Lohmann G (2003) Southern Ocean Origin for the resumption of Atlantic thermohaline circulation during deglaciation. Nature 424:532–536
42. Knorr G, Lohmann G (2007) Rapid transitions in the Atlantic thermohaline circulation triggered by global warming and meltwater during the last deglaciation. Geochem Geophys Geosyst 8(12), Q12006:1–22. doi:10.1029/2007GC001604
43. Kwasniok F, Lohmann G (2008) Underlying Dynamics of Glacial Millennial-Scale Climate Transitions Derived from Ice-Core Data. Phys Rev E (accepted)
44. Lawver LA, Gahagan LM (2003) Evolution of Cenozoic seaways in the circum-Antarctic region. Palaeogeography, Palaeoclimatology, Palaeoecology 198:11–37. doi:10.1016/S0031-0182(03)00392-4
45. Lisiecki LE, Raymo ME (2005) A Pliocene-Pleistocene stack of 57 globally distributed benthic O-18 records. Paleoceanography 20:PA1003. doi:10.1029/2004PA001071
46. Lohmann G (2003) Atmospheric and oceanic freshwater transport during weak Atlantic overturning circulation. Tellus 55A: 438–449
47. Lohmann G, Gerdes R (1998) Sea ice effects on the Sensitivity of the Thermohaline Circulation in simplified atmosphere-ocean-sea ice models. J Climate 11:2789–2803
48. Lohmann G, Schneider J (1999) Dynamics and predictability of Stommel's box model: A phase space perspective with implications for decadal climate variability. Tellus 51A:326–336
49. Lohmann G, Schulz M (2000) Reconciling Boelling warmth with peak deglacial meltwater discharge. Paleoceanography 15:537–540
50. Lorenz EN (1963) Deterministic nonperiodic flow. J Atmos Sci 20:130–141
51. Lorenz EN (1976) Nondeterministic theories of climatic change. Quat Res 6:495–506
52. Lorenz EN (1982) Atmospheric predictability experiments with a large numerical model. Tellus 34:505–513
53. Lorenz EN (1990) Can chaos and intransitivity lead to interannual variability? Tellus 42A:378–389
54. Lorenz S, Lohmann G (2004) Acceleration technique for Milankovitch type forcing in a coupled atmosphere-ocean circulation model: method and application for the Holocene. Climate Dyn 23(7–8):727–743. doi:10.1007/s00382-004-0469-y
55. Lund R, Reeves J (2002) Detection of undocumented changepoints: A revision of the two-phase regression model. J Climate 15:2547–2554
56. Lunkeit F, Bauer SE, Fraedrich K (1998) Storm tracks in a warmer climate: Sensitivity studies with a simplified global circulation model. Clim Dyn 14:813–826
57. Manabe S, Stouffer RJ (1993) Century-scale effects of increased atmospheric CO_2 on the ocean-atmosphere system. Nature 364:215–218
58. Maraun D, Kurths J (2004) Cross wavelet analysis. Significance testing and pitfalls. Nonlin Proc Geoph 11:505–514
59. Maraun D, Kurths J (2005) Epochs of phase coherence between El Niño/Southern Oscillation and Indian monsoon. Geophys Res Lett 32:L15709. doi:10.1029/2005GL023225
60. Maslin MA, Li XS, Loutre MF, Berger A (1998) The contribution of orbital forcing to the progressive intensification of Northern Hemisphere Glaciation. Quat Sci Rev 17:411–426
61. Maslin MA, Ridgewell A (2005) Mid-Pleistocene Revolution and the eccentricity myth. Special Publication of the Geological Society of London 247:19–34
62. Milankovitch M (1941) Kanon der Erdbestrahlung. Royal Serb Acad Spec Publ, Belgrad, 132, Sect. Math Nat Sci 33:484
63. Mori H (1965) A Continued-Fraction Representation of the Time-Correlation Functions Prog Theor Phys 33:423–455. doi:10.1143/PTP.34.399
64. North Greenland Ice Core Project members (2004) High-resolution record of Northern Hemisphere climate extending into the last interglacial period. Nature 431:147–151
65. Paillard D (1998) The timing of Pleistocene glaciations from a simple multiple-state climate model. Nature 391:378–381
66. Palmer TN (1996) Predictability of the atmosphere and oceans: From days to decades. In: Anderson DTA, Willebrand J (eds) Large-scale transport processes in oceans and atmosphere. NATO ASI Series 44. Springer, Berlin, pp 83–155
67. Parker DE, Jones PD, Folland CK, Bevan A (1994) Interdecadal changes of surface temperature since the late nineteenth century. J Geophys Res 99:14,373-14,399
68. Peixoto JP, Oort AH (1992) Physics of Climate. American Institute of Physics, New York, p 520
69. Petit JR, Jouzel J, Raynaud D, Barkov NI, Barnola JM, Basile I, Bender M, Chappellaz J, Davis M, Delaygue G, Delmotte M, Kotlyakov VM, Legrand M, Lipenkov VY, Lorius C, Pepin L, Ritz C, Saltzman E, Stievenard M (1999) Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. Nature 399:429–436
70. Raymo M, Ganley K, Carter S, Oppo DW, McManus J (1998) Millennial-scale climate instability during the early Pleistocene epoch. Nature 392:699–701
71. Reddy SC, Schmidt P, Henningson D (1993) Pseudospectra of the Orr-Sommerfeld operator. SIAM J Appl Math 53:15–47
72. Rial JA (1999) Pacemaking the Ice Ages by Frequency Modulation of Earth's Orbital Eccentricity. Science 285:564–568
73. Rial JA (2004) Abrupt Climate Change: Chaos and Order at Orbital and Millennial Scales. Glob Plan Change 41:95–109
74. Ridgwell AJ, Watson AJ, Raymo ME (1999) Is the spectral signature of the 100 Kyr glacial cycle consistent with a Milankovitch origin? Paleoceanography 14:437–440
75. Rogers JC (1985) Atmospheric circulation changes associated with the warming over the northern North Atlantic in the 1920s. J Climate Appl Meteorol 24:1303–1310
76. Romanova V, Lohmann G, Grosfeld K, Butzin M (2006) The relative role of oceanic heat transport and orography on glacial climate. Quat Sci Rev 25:832–845. doi:10.1016/j.quascirev.2005.07.007
77. Saltzman (2002) Dynamical Paleoclimatology. Generalized Theory of Global Climate Change. In: International Geophysics Series, vol 80. Harcourt-Academic Press (Elsevier Science), San Diego, p 354

78. Schulz M, Paul A, Timmermann A (2004) Glacial-Interglacial Contrast in Climate Variability at Centennial-to-Millennial Timescales: Observations and Conceptual Model. Quat Sci Rev 23:2219
79. Seidel DJ, Lanzante JR (2004) An assessment of three alternatives to linear trends for characterizing global atmospheric temperature changes. J Geophy Res 109:D14108. doi:10.1029/2003JD004414
80. Stocker TF (1998) The seesaw effect. Science 282:61–62
81. Stocker TF, Johnsen SJ (2003) A minimum thermodynamic model for the bipolar seesaw. Paleoceanography 18(4):1087
82. Stommel H (1961) Thermohaline Convection with Two Stable Regimes of Flow. Tellus 13:224–230
83. Tiedemann R, Sarnthein M, Shackleton NJ (1994) Astronomic time scale for the Pliocene Atlantic δ^{18}O and dust flux records of Ocean Drilling Program site 659. Paleoceanography 9:19–638
84. Timmermann A, Lohmann G (2000) Noise-Induced Transitions in a simplified model of the thermohaline circulation. J Phys Oceanogr 30(8):1891–1900
85. Timmermann A, Oberhuber J, Bracher A, Esch M, Latif M, Roeckner E (1999) Increased El Niño frequency in a climate model forced by future greenhouse warming. Nature 398:694–696
86. Torrence C, Compo G (1998) A practical guide to wavelet analysis. Bull Amer Meteor Soc 79:61–78
87. Trefethen LN, Trefethen AE, Reddy SC, Driscoll TA (1993) Hydrodynamic stability without eigenvalues. Science 261:578–584
88. Trenberth KE (1990) Recent observed interdecadal climate changes in the Northern Hemisphere. Bull Am Meteorol Soc 71:988–993
89. Uhlenbeck GE, Ornstein LS (1930) On the theory of Brownian Motion. Phys Rev 36:823–841
90. Wunsch C (1999) The interpretation of short climate records, with comments on the North Atlantic and Southern Oscillation. Bull Amer Meteor Soc 80:245–255
91. Wunsch C (2004) Quantitative estimate of the Milankovitch-forced contribution to observed Quaternary climate change. Quat Sci Rev 23(9–10):1001–1012
92. Yamamoto R, Iwashima T, Sanga NK (1985) Climatic jump: a hypothesis in climate diagnosis. J Meteorol Soc Jpn 63:1157–1160
93. Zachos J, Pagani M, Sloan L, Thomas E, Billups K (2001) Trends, Rhythms, and Aberrations in Global Climate 65 Ma to Present. Science 292(5517):686–693
94. Zwanzig R (1980) Thermodynamic modeling of systems far from equilibrium. In: Garrido L (ed) Lecture Notes in Physics 132, in Systems Far From Equilibrium. Springer, Berlin

Books and Reviews

Dijkstra HA (2005) Nonlinear Physical Oceanography, 2nd revised and extended edition. Springer, New York, pp 537
Hansen J, Sato M, Kharecha P (2007) Climate change and trace gases. Phil Trans R Soc A 365:1925–1954. doi:10.1098/rsta.2007.2052
Lockwood JG (2001) Abrupt and sudden climate transitions and fluctuations: a review. Int J Climat 21:1153–1179
Rial JA, Pielke RA Sr, Beniston M, Claussen M, Canadell J, Cox P, Held H, N deNoblet-Ducudre, Prinn R, Reynolds J, Salas JD (2004) Nonlinearities, Feedbacks and Critical Thresholds Within the Earth's Climate System. Clim Chang 65:11–38
Ruddiman WF (2001) Earth's Climate. Past and Future. WH Freeman, New York, p 465
Stocker TF (1999) Abrupt climate changes from the past to the future-a review. Int J Earth Sci 88:365–374

Acceleration Mechanisms

Don B. Melrose
School of Physics, University of Sydney, Sydney, Australia

Article Outline

Glossary
Definition of the Subject
Introduction
Stochastic Acceleration
Diffusive Shock Acceleration
DSA at Multiple Shocks
Applications of DSA
Resonant Scattering
Acceleration by Parallel Electric Fields
Other Acceleration Mechanisms
Future Directions
Appendix A: Quasilinear Equations
Bibliography

Glossary

Acceleration mechanism Any process, or sequence of processes, that results in a small fraction of the particles in a plasma each gaining an energy greatly in excess of the thermal energy.

Heliosphere The region around the Sun populated by plasma escaping from the Sun in the solar wind, and extending to well beyond the orbits of the planets.

Magnetosphere The region around a planet or star populated by plasma on magnetic field lines tied to the planet or star.

Resonance The gyroresonance condition (6) corresponds to the wave frequency being an integral, s, multiple of the particle gyrofrequency in the rest frame of the particle gyrocenter; $s = 0$ is also called the Cerenkov condition.

Shock A fast magnetoacoustic shock is a discontinuity in plasma flow that propagates into the upstream plasma faster than the Alfvén speed, compressing both the plasma and the magnetic field.

Solar corona The hot ($> 10^6$ K) outer atmosphere of the Sun.

Solar flare An explosion in the solar corona that leads to copious acceleration of fast particles.

Definition of the Subject

A typical astrophysical or space plasma consists of thermal plasma components, a magnetic field, various nonthermal distributions of fast particles and turbulence involving various kinds of waves. An acceleration mechanism is the process, or sequence of processes, by which the fast particles gain their energy. There is no single, universal acceleration mechanism, and it is now recognized that the data require a rich variety of different acceleration mechanisms operating under different conditions in different space and astrophysical plasmas. Three generic mechanisms are emphasized in this review: stochastic acceleration, diffusive shock acceleration and acceleration by parallel electric fields.

Introduction

The major observational discoveries in this field were almost all unexpected and initially quite baffling. The first notable discovery was that of cosmic rays (CRs), by Victor Hess in 1912. Hess measured the rate of ion production inside a hermetically sealed container in a balloon flight to an altitude of 5300 m, and concluded that "The results of my observation are best explained by the assumption that a radiation of very great penetrating power enters our atmosphere from above". By the late 1920s CRs were identified as charged particles, and not photons, and due to their isotropy it was inferred that they come from the Galaxy. Perhaps the earliest idea on the origin of Galactic CRs (GCRs) has stood the test of time: in 1934 Baade and Zwicky [8] proposed that GCRs are accelerated in supernova explosions. Supernovae are the only adequate source of the energy needed to replace GCRs, whose lifetime in the Galaxy is about 10^7 yr [34]. The observed momentum distribution of GCRs is a power law, $f(p) \propto p^{-b}$, with more recent observations implying $b = 4.7$ below a knee in the distribution at $p \approx 10^{15}$ eV/c with the spectrum steepening at higher momenta; Axford [6] referred to these as GCR1 and GCR2, respectively.

The next major discoveries followed the development of radio astronomy in the 1940s. This led to the identification of a wide class of sources radiating by a nonthermal mechanism, identified in the late 1940s as synchrotron radiation. The presence of highly relativistic electrons in synchrotron sources raises the question of how they are accelerated. Subsequent development of X-ray and gamma-ray astronomy placed severe constraints on the acceleration of the relativistic electrons in synchrotron sources, due to the short life times of electrons radiating at high frequencies. Solar radio astronomy led to the identification of several different types of radio bursts associated with solar flares, including type III bursts due to ≈ 10 keV electrons escaping from the Sun, type II bursts associated with shocks that result from the flare and type I and type IV emission that involve radio continua. All these phenomena require specific acceleration mechanisms to explain the presence of the radio-emitting electrons.

Space physics began in the late 1950s with the discovery of the Earth's Van Allen radiation belts, raising the question of how these energetic electrons and ions trapped in the Earth's magnetosphere are accelerated. Further space exploration led to the identification of energetic particle distributions in essentially all space plasmas: other planetary magnetospheres, the solar wind, and more specifically associated with bow shocks and interplanetary shocks. Observations from spacecraft provide in situ data on particles and fields, in the Earth's magnetosphere, in other planetary magnetospheres, and throughout the heliosphere. In the case of the magnetosphere, Earth-based data on precipitation of particles, notably in the auroral zones, and on radio waves originating from the magnetosphere complement the spacecraft data on particles and waves in situ.

Current-day ideas on acceleration mechanisms are dominated by three generic mechanisms, plus a rich variety of specific mechanisms of various kinds. Two of the generic mechanisms had forerunners in two papers by Fermi [30,31]. In the first of these [30], Fermi suggested that the acceleration of GCRs is due to reflections off moving interstellar clouds. CRs gain energy in head-on collisions with clouds, and lose energy in overtaking collisions; a net average gain results from head-on collisions being more probable than overtaking collisions. This model is recognized as the first example of the generic mechanism of stochastic acceleration. In the second paper [31], Fermi proposed acceleration between 'closing jaws' in which all reflections are head-on and hence energy-increasing. This idea is an essential ingredient in the astrophysically most important of the generic mechanisms: diffusive shock acceleration (DSA). The third generic mechanism is acceleration by a parallel electric field, $E_\parallel$; in particular, there is compelling evidence for auroral electrons resulting from such acceleration. Both stochastic acceleration and DSA require that the particles be scattered efficiently in order to be accelerated, and an important step in the development of the theory is that very efficient pitch-angle scattering of fast particles can result from gyroresonant inter-

actions with low-frequency waves. The following remarks relate to the three generic mechanisms and to resonance acceleration.

Stochastic Acceleration

The original version of Fermi mechanism [30], involving reflection off magnetized interstellar clouds, is too slow to account for GCRs, basically because such collisions are so infrequent. The acceleration mechanism is more efficient for more frequent, smaller energy changes than for less frequent, larger energy changes. This led to the recognition that a more efficient Fermi-type acceleration can result from MHD turbulence [23,50,82,83,95]. One form of the idea is that the compression and rarefaction of B, associated with the MHD turbulence, conserves the magnetic moment, $p_\perp^2/B$, and when combined with a scattering process that maintains the isotropy of the particles, this implies a 'betatron' acceleration, or 'magnetic pumping' [15,88]. A closer analogy with Fermi acceleration is due to the small fraction of the particles that reflect off the compressions in B, which propagate at the Alfvén speed, v_A. Besides the betatron-type acceleration and the reflection off moving compressions, there is also a transit acceleration [90] due to particles diffusing through the compressions and rarefactions. All these energy changes contribute to stochastic acceleration by MHD turbulence [57].

The mathematical description of Fermi-type acceleration is in terms of isotropic diffusion in momentum space. In the early literature stochastic acceleration was treated using a Fokker–Planck approach that includes a term that describes a systematic acceleration and a term that describes a diffusion in energy. In was shown by Tverskoi [98] that for Fermi's original mechanism this equation reduces to an isotropic diffusion in momentum space. The same equation is derived for acceleration by isotropic turbulence [40,57,93]. However, MHD turbulence is never isotropic: it consists of a spectrum of MHD waves, in both the (non-compressive) Alfvén mode and the (compressive) fast magnetoacoustic mode. It is now assumed that the isotropy is due to efficient resonant scattering, with no implication that the turbulence itself is isotropic.

Diffusive Shock Acceleration

Fermi [31] suggested that the acceleration can be efficient if all reflections are head-on. A simple example is when two magnetized clouds approach each other, and a particle bounces back and forth reflecting off the approaching clouds. A related example is for a shock propagating into a closed magnetic loop, where a particle trapped in the loop bounces back and forth being reflected head-on by the approaching shock [69,106]. However, such examples are constrained by the fact that acceleration ceases when the approaching structures meet. In 1977–8 several different authors [7,10,17,56] recognized that a single shock can accelerate particles efficiently, provided that the particles are scattered on either side of the shock, as illustrated in Fig. 1. This mechanism is called diffusive shock acceleration (DSA).

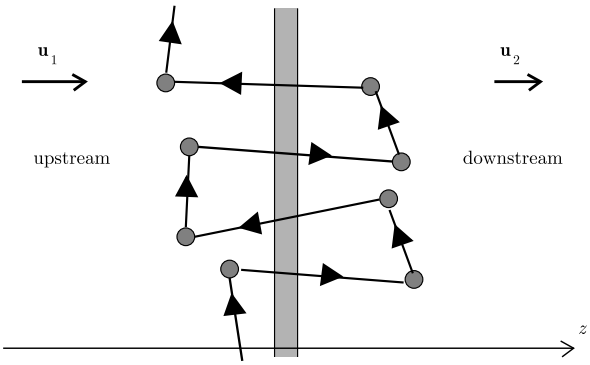

Acceleration Mechanisms, Figure 1
Diffusive shock acceleration is illustrated: the *shaded vertical region* is the shock, the *circular blobs* denote idealized scattering centers, and the *solid line with arrows* denotes the path of an idealized fast particle

To understand why DSA occurs, first note that the fluid velocity changes discontinuously across the shock. Let the two sides of the shock be labeled 1 and 2. Consider a particle on side 1 about to cross the shock and enter side 2. The scattering centers on side 2 are convected along with the plasma flow (velocities u_1, u_2), and the particle sees them approaching it head on at $|u_1 - u_2|$. Once the particle crosses the shock and is scattered it gains energy due to the oncoming motion of the scatterers. After being scattered a number of times on side 2, the particle can diffuse back to the shock and cross back onto side 1. On doing so, it sees the scattering centers on side 1 approaching it head on at $|u_1 - u_2|$. Again it gains energy on being scattered. DSA requires efficient scattering, and this can be achieved only by resonant scattering.

Resonant Scattering

Both stochastic acceleration and DSA require that the fast particles be scattered such that their distribution remains almost isotropic. The recognition that fast particles are scattered very efficiently developed in two other context before its significance to acceleration mechanism become fully appreciated. These developments essentially defined a new field of "plasma astrophysics" that emerged in the mid 1960s.

Efficient scattering is needed to explain precipitation of particles trapped in the magnetosphere. Under quiescent conditions, ions and electrons precipitate steadily (albeit in localized bursts) from the radiation belts, implying that there is a continuous resupply of fast particles. The precipitation is interpreted in terms of pitch-angle scattering of trapped particles into the loss cone. Pitch-angle scattering of ions [24,105] involves ion cyclotron waves, and the scattering of electrons involves whistlers [27,52]. Efficient scattering is also needed to explain the data on GCRs. Both their estimated residence time, $\sim 10^7$ yr, in the Galactic disc and their small observed anisotropy imply that GCRs diffuse slowly through the Galaxy. The net flow speed is of order the Alfvén speed, $v_A \sim 10^{-4}c$. The efficient scattering is attributed to resonant scattering by Alfvén (and magnetoacoustic waves) in the interstellar medium.

Resonant scattering, also called pitch-angle scattering, occurs when the gyroresonance condition is satisfied for a fast particle interacting with a wave whose phase speed is much less than the speed of the particle. To the particle, the wave appears as a spatially periodic oscillation in the magnetic field. The resonance condition corresponds to the wavelength of the magnetic fluctuation being equal to the distance that the particle propagates along the field line in a gyroperiod. As a consequence, the Lorentz force, $q\mathbf{v} \times \delta\mathbf{B}$, due to the magnetic fluctuation, $\delta\mathbf{B}$, in the wave is in the same direction at the same phase each gyroperiod. This causes a systematic change in the momentum of the particle, without any change in energy, corresponding to a change in the pitch angle, $\alpha = \arctan(p_\parallel/p_\perp)$. The sense of the change remains the same until phase coherence is lost. Such interactions cause random changes in α, as a result of which the particles diffuse in α. The resonance condition for a given wave (given $\omega, k_\parallel$) defines a resonance curve in velocity ($v_\parallel$–$v_\perp$) space or in momentum ($p_\parallel$–$p_\perp$) space; this curve is a conic section in general and it reduces to a circle in the limit of slow phase speed. Pitch-angle diffusion corresponds to particles diffusing around the circumference of this circle.

The resonant interaction results in the waves either growing or damping, depending on whether energy is transferred from the particles to the waves, or vice versa. For any given distribution of particles, all the particles that can resonate with a given wave lie on the resonance curve. If the gradient in the distribution function along this curve is an increasing function of ε, more particles lose energy to the wave than gain energy from the wave, and the wave grows. For any anisotropic distribution, there is a choice of the relevant signs (harmonic number $s = \pm 1$, $k_\parallel/|k_\parallel|$ and $v_\parallel/|v_\parallel|$) that implies growth of those waves that carry away the excess momentum. The back reaction on the distribution of particles, called quasilinear relaxation, reduces the anisotropy that causes the wave growth [74]. This implies that the resonant waves needed to scatter the particles can be generated by the anisotropic particles themselves.

In DSA the resonant waves can be generated by a more subtle anisotropy. Upstream of the shock the density of the fast particles decreases with distance from the shock, and the spatial gradient implies a form of anisotropy that can cause the resonant waves to grow. Analogous to the scattering of streaming CRs, the growth of the resonant waves decreases rapidly with increasing particle energy, and some other source of resonant waves is required for higher energy particles. Scattering downstream of the shock is less problematic, with several possible sources of resonant waves, including waves generated in the upstream region and swept back across the shock.

Acceleration by Parallel Electric Fields

There is overwhelming evidence that acceleration by $E_\parallel$ occurs in the Earth's auroral zones. Acceleration by $E_\parallel$ plays a central role in populating pulsar magnetospheres with electron-positron pairs, and is a possible mechanism for the "bulk energization" mechanism for energetic electrons in solar flares.

Acceleration by $E_\parallel$ is necessarily associated with a parallel current, $J_\parallel$: the relevant source term from Maxwell's equations for the energy transfer is $J_\parallel E_\parallel$. Historically, the existence of $J_\parallel$ in the auroral zones is perhaps the oldest idea relevant to the theory of acceleration, having been predicted by Kristian Birkeland in 1903. Understanding acceleration by $E_\parallel$ remains problematical. One conceptual difficulty is that astrophysical plasmas are typically described using ideal magnetohydrodynamics (MHD), in which one has $E_\parallel = 0$ by hypothesis: theories involving $E_\parallel$ are necessarily outside the scope of MHD. A related difficulty is that $E_\parallel$ accelerates particles of oppositely signed charge in opposite directions, setting up a current flow that is opposed by inductive effects. Both inductive effects and the charge separation resulting from the current flow tend to suppress the $E_\parallel$ and hence the acceleration.

A generic model has emerged for the acceleration by $E_\parallel$ of auroral electrons [103]. The model requires a mechanical driver, which sets up an EMF, and a closed circuit around which the current flows. The EMF is made available by magnetic reconnection, which acts like a switch setting up the current flow around the circuit. The energy released is transported away through a Poynting flux, propagating at v_A. This energy is transferred, at least in part, to energetic particles, in an acceleration region, which acts as a load in the circuit. The available po-

tential localizes across the acceleration region, producing the $E_\parallel$.

Suggested astrophysical applications of circuit models tend to be viewed with suspicion, especially because the physics involved is specifically excluded in the assumptions of ideal MHD. Some non-ideal-MHD effect is needed to explain why $E_\parallel \neq 0$ develops. One such mechanism is resistivity, which is included in non-ideal or resistive MHD, but the $E_\parallel$ allowed by classical resistivity is too weak to play any significant role in the fully ionized plasmas of interest. Currently favored possibilities include double layers and inertial Alfvén waves.

Despite its apparent simplicity, acceleration by $E_\parallel$ is the least understood of the three generic mechanisms.

Stochastic Acceleration

Second-order Fermi acceleration [30] is now identified as an archetypical form of stochastic acceleration, and the term Fermi acceleration is sometimes used in a generic sense to describe stochastic acceleration.

Diffusion in Momentum Space

The isotropic diffusive equation in momentum space that describes stochastic acceleration is [57,98,99]

$$\frac{\partial\langle f\rangle(p)}{\partial t} = \frac{1}{p^2}\frac{\partial}{\partial p}\left[p^2 D_{pp}(p)\frac{\partial}{\partial p}\right]\langle f\rangle(p)\,, \tag{1}$$

where $\langle f\rangle(p)$ is the particle distribution function, $f(p)$, averaged over pitch angle. The diffusion coefficient is

$$D_{pp}(p) = \zeta_A \frac{cp^2}{4v}\left(1-\frac{v_A^2}{v^2}\right)^2\,, \tag{2}$$

where ζ_A is the acceleration rate. The final factor in (2) does not appear in simpler treatments of Fermi-type acceleration; this factor may be ignored only in the limit of particle speeds much greater than the Alfvén speed, $v \gg v_A$. The meaning of ζ_A is most easily understood by estimating the mean rate of acceleration [72,97]

$$\left\langle\frac{d\varepsilon}{dt}\right\rangle = \frac{1}{p^2}\frac{\partial}{\partial p}\left[vp^2 D_{pp}(p)\right]\,, \tag{3}$$

where $\varepsilon = (m^2c^4 + p_\parallel^2 c^2 + p_\perp^2 c^2)^{1/2}$ is the energy. In the limit $v \gg v_A$ (3) gives

$$\left\langle\frac{d\varepsilon}{dt}\right\rangle \approx \zeta_A pc\,, \tag{4}$$

which reduces for highly relativistic particles to the familiar form $\langle d\varepsilon/dt\rangle \approx \zeta_A\varepsilon$. The acceleration rate may be estimated as

$$\zeta_A = \frac{\pi}{4}\bar{\omega}\left(\frac{\delta B}{B}\right)^2\,, \tag{5}$$

where δB is the magnetic amplitude in the waves and $\bar{\omega}$ is their mean frequency. A remarkable feature of (5) is that, although effective scattering is an essential ingredient in the theory, the acceleration rate is independent of the details of the scattering. The parameters that appear in (5) refer only to the MHD turbulence that is causing the acceleration.

Treatment in Terms of Resonant Interactions

Further insight into Fermi-type acceleration was given by Achterberg [1] who showed that it is equivalent to resonant damping of magnetoacoustic waves in the presence of efficient pitch-angle scattering. The general gyroresonance condition is

$$\omega - s\Omega - k_\parallel v_\parallel = 0\,, \tag{6}$$

where ω is the wave frequency, $k_\parallel$ is the component of its wave vector parallel to $\mathbf{B}$, $s = 0, \pm 1\ldots.$ is the harmonic number, $\Omega = |q|B/\gamma m$ is the relativistic gyrofrequency of the particle with charge q, mass m, Lorentz factor γ and velocity $v_\parallel$ parallel to $\mathbf{B}$. The effect of gyroresonant interactions on the particles and the waves is described by the quasilinear equations, which are written down in the Appendix. The particular resonant interaction of interest here is at the Cerenkov resonance, $s = 0$. Resonance at $s = 0$ is possible for waves that have a component of their electric field along $\mathbf{B}$, and this is the case for magnetoacoustic waves, but not for Alfvén waves (in ideal MHD). The resonant interaction alone would cause the particles to diffuse in $p_\parallel$, with $p_\perp$ remaining constant. This corresponds to the special case

$$\frac{df(\mathbf{p})}{dt} = \frac{\partial}{\partial p_\parallel}\left[D_{\parallel\parallel}(\mathbf{p})\frac{\partial f(\mathbf{p})}{\partial p_\parallel}\right] \tag{7}$$

of the general diffusion equation (33). In the absence of pitch-angle scattering, the distribution of fast particles becomes increasingly anisotropic, favoring $p_\parallel$ over $p_\perp$, leading to a suppression of the acceleration. In the presence of efficient pitch-angle scattering, one averages (7) over pitch-angle, leading to the isotropic diffusion equation (3), with $D_{pp}(p) = \frac{1}{2}\int d\mu\mu^2 D_{\parallel\parallel}(\mathbf{p})$, where $\mu = p_\parallel/p$ is the cosine of the pitch angle. The resulting expression for $D_{pp}(p)$, given by (2) with (5), confirms that this resonant interaction is equivalent to Fermi acceleration.

Diffusive Shock Acceleration

Diffusive shock acceleration (DSA) is now recognized as the most important acceleration mechanism in astrophysical plasmas: its identification was a major achievement in

the field. Prior to the identification of DSA, it was difficult to explain how acceleration can be as efficient as it appears to be, and why power-law energy spectra are so common. DSA is so efficient that one needs to look for self-regulation processes that limit its efficiency, and DSA implies power-law spectra of the form observed without any additional assumptions.

The treatment of DSA given below is a nonrelativistic, single-particle theory. The assumption that collective effects of the fast particles can be neglected is not necessarily valid: DSA is so efficient that the fast particles can become an important component of the upstream plasma. Once the pressure associated with the fast particles becomes comparable with the thermal pressure, the structure of the shock is modified by this pressure. This can result in the stresses being transferred from the downstream to the upstream plasma primarily through the fast particles, with no discontinuity in the density of the thermal gas [67]. Such nonlinear effects provide a constraint on DSA.

Diffusive Treatment of DSA

Consider a distribution of particles $f(p,z)$ that is averaged over pitch angle and is a function of distance z from a shock in a frame in which the shock is at rest. It is assumed that scattering causes the particles to diffuse in the z direction with diffusion coefficient $\kappa(z)$. The particles are also assumed to be streaming with the streaming speed u. The diffusion is described by

$$\frac{\partial f(p,z)}{\partial t} + u\frac{\partial f(p,z)}{\partial z} + \dot{p}\frac{\partial f(p,z)}{\partial p} = \frac{\partial}{\partial z}\left(\kappa(z)\frac{\partial f(p,z)}{\partial z}\right) + Q(p,z) - f_{\mathrm{esc}}(p)\,, \quad (8)$$

where $Q(p,z)$ is a source term, where the sink term $f_{\mathrm{esc}}(p)$ takes account of escape of particles downstream from the shock, and where the term involving $\dot{p} = -(1/3)p\partial u/\partial z$ determines the energy changes. It is assumed that the speed u changes abruptly across the shock:

$$u = \begin{cases} u_1 & \text{for} \quad z<0 \quad \text{(upstream)}\,, \\ u_2 & \text{for} \quad z>0 \quad \text{(downstream)}\,, \end{cases} \qquad \frac{\partial u}{\partial z} = (u_1-u_2)\delta(z)\,. \quad (9)$$

A stationary solution of (8) exists when both the source and the sink term are neglected, such that the equation reduces to $u\,\partial f/\partial z = (\partial/\partial z)(\kappa\,\partial f/\partial z)$; a general solution is

$$f(p,z) = A + B\exp\left[u\int \mathrm{d}z\,\frac{1}{\kappa(z)}\right]\,, \quad (10)$$

with u constant on either side of the shock, and where A, B are constants. In the upstream region, $z<0$, one has $u/\kappa(z)\neq 0$, and the solution (10) diverges at $z\to-\infty$ unless one has $B=0$ for $z<0$. Writing $f_\pm(p) = \lim_{z\to\pm\infty} f(p,z)$, one has

$$f(p,z) = \begin{cases} f_-(p) + [f_+(p)-f_-(p)]\exp\left[u_1\int_0^\infty \mathrm{d}z\,\frac{1}{\kappa(z)}\right] & \text{for } z<0\,, \\ f_+(p) \qquad \text{for } z>0\,. \end{cases} \quad (11)$$

On matching the discontinuity in the derivative of this solution with the discontinuity due to the acceleration term, one finds

$$f_+(p) = bp^{-b}\int_0^p \mathrm{d}p'\,p'^{(b-1)}f_-(p')\,, \quad b = \frac{3u_1}{u_2-u_1}\,, \quad (12)$$

which determines the downstream solution in terms of the upstream solution. That is, if the shock propagates into a region where the distribution is $f_-(p)$, then after the shock has passed the distribution is $f_+(p)$.

For monoenergetic injection, $f_-(p)\propto\delta(p-p_0)$ say, (12) implies $f_+(p)\propto p^{-b}$. The power law index, b, is determined by the ratio u_1/u_2, which is determined by the strength of the shock. In terms of the Mach number $M = u_1/v_{\mathrm{A1}}$, where v_{A1} is the Alfvén speed upstream of the shock, and the ratio Γ of specific heats, one finds

$$b = \frac{3r}{r-1} = \frac{3(\Gamma+1)M^2}{2(M^2-1)}\,, \quad r = \frac{(\Gamma+1)M^2}{2+(\Gamma-1)M^2}\,, \quad (13)$$

where r is the compression factor across the shock. For $\Gamma = 5/3$ and $M^2 \gg 1$, the maximum strength of the shock is $r=4$ and the minimum value of the spectral index is $b=4$.

Alternative Treatment of DSA

An alternative treatment of DSA is as follows. Consider a cycle in which a particle with momentum p crosses the shock from downstream to upstream and back again. Let $\Delta(p)$ be the change in p in a cycle, and $P(p)$ be the probability per cycle of the particle escaping far downstream. Simple theory implies

$$\Delta(p) = \frac{4(r-1)}{3r}\,\frac{u_1}{v}\,p\,, \quad P(p) = \frac{4u_1}{vr}\,. \quad (14)$$

After one cycle, $p\to p' = p+\Delta(p)$ implies that the number of particles per unit momentum changes according to $4\pi p^2 f(p)\mathrm{d}p \to 4\pi p'^2 f(p')\mathrm{d}p' = [1-$

$P(p)]4\pi p^2 f(p)\mathrm{d}p$. After integration, one imposes the requirement that the downstream flux of particles balances the upstream flux, and a result equivalent to (12) is derived.

This model allows one to estimate the acceleration time, t_a, in terms of the cycle time t_c. One has $t_c = (\lambda_1/u_1 + \lambda_2/u_2) = (\lambda_1 + r\lambda_2)/u_1$, where $\lambda_{1,2}$ are the scattering mean free paths in the upstream and downstream regions, respectively. For isotropic scattering, the scattering mean free path, λ, is related to the spatial diffusion coefficient, κ, by $\kappa = \lambda v/3$, where v is the speed of the particle. The mean free path may be written as $\lambda = \eta r_g$, where r_g is the gyroradius, and it is usually assumed that η is a constant of order unity, called Bohm diffusion. The rate of momentum gain is given by

$$\frac{\mathrm{d}p}{\mathrm{d}t} = \frac{p}{t_a} - \left(\frac{\mathrm{d}p}{\mathrm{d}t}\right)_{\mathrm{loss}}, \quad t_a = \frac{u_1^2}{c\bar{\lambda}}, \quad \bar{\lambda} = \frac{3r(\lambda_1 + r\lambda_2)}{4(r-1)}, \tag{15}$$

where t_a is an acceleration time, $\bar{\lambda}$ is a mean scattering free path, and where a loss term is included.

Injection and Preacceleration

DSA requires efficient scattering, and this is attributed to resonant scattering, Sect. "Resonant Scattering". For ions, scattering by Alfvén and magnetoacoustic waves requires $v \gg v_A$, and for nonrelativistic electrons, scattering by whistler waves requires $v \gg 43 v_A$. These conditions are typically not satisfied for thermal particles, and one needs to postulate an injection or preacceleration mechanism to create a seed population of ions or electrons above the respective thresholds for DSA to operate on them.

One possible type of injection mechanism involves collisions populating a depleted thermal tail. The thresholds ($v \approx v_A$ for ions or $v \gg 43 v_A$ for electrons) are somewhere in the tail of the thermal distribution of particles, and the acceleration removes these particles, leaving a depleted thermal tail. Collisions between thermal particles cause a net flux of particles in momentum space into the high-energy tail, tending to restore a thermal distribution. The rate at which nonthermal particles are created by this process depends on the collision frequency and on the details of the acceleration mechanism [38]. However, this process encounters two serious difficulties. First, it is too slow to account for the required injection. Second, for ions it is extremely sensitive to the charge and mass, implying that the relative isotopic abundances of accelerated ions should differ by very large factors from the relative isotopic abundances of the background plasma, and this is simply not the case in general. Other preacceleration mechanisms involve either electric fields associated with the shock itself, or waves associated with the shock. These are discussed briefly below in connection with shock drift acceleration.

DSA at Relativistic Shocks

In the extension of DSA to highly relativistic shocks, the assumption of near isotropy of the pitch angle distribution is not valid, and the simple analytic theory does not generalize in a straightforward way. Numerical calculations show that the resulting spectrum is a power law, with $b = 4.2$–4.3 [4,41].

DSA at Multiple Shocks

In many of the applications of DSA, it is unlikely that a single shock is responsible for the acceleration [2]. Extension of the theory to multiple shocks shows that DSA is related to Fermi-type acceleration: there are energy gains associated with DSA at each shock, and energy losses due to decompression between shocks. It is straightforward to include synchrotron losses in such a model, and the competition between acceleration and synchrotron losses can possibly explain flat or slightly inverted synchrotron spectra in some sources.

Multiple Shock Model

Consider the following model for DSA at multiple shocks.

(a) All shocks are identical with compression ratio, $r = b/(b-3)$, and all have the same injection spectrum, $\phi_0(p)$, assumed to be monoenergetic, $\phi_0(p) \propto \delta(p - p_0)$.

(b) The distribution downstream of the first shock, $f_1(p)$, results from acceleration of the injection spectrum, $\phi_0(p)$, at the shock, cf. (12):

$$f_1(p) = bp^{-b}\int_0^p \mathrm{d}q\, q^{b-1}\phi_0(q), \quad b = \frac{3r}{r-1}. \tag{16}$$

(c) Between each shock a decompression occurs. Assuming that scattering keeps the particles isotropic, the distribution function after decompression is $f_1'(p) = f_1(p/R)$, with $R = r^{-1/3}$. Hence the injection spectrum into the second shock is

$$f_1'(p) = b(p/R)^{-b}\int_0^{p/R} \mathrm{d}q\, q^{b-1}\phi_0(q). \tag{17}$$

(d) The injection spectrum at any subsequent shock consists of the sum of $\phi_0(p)$ and the decompressed spectrum resulting from acceleration at the previous shock.

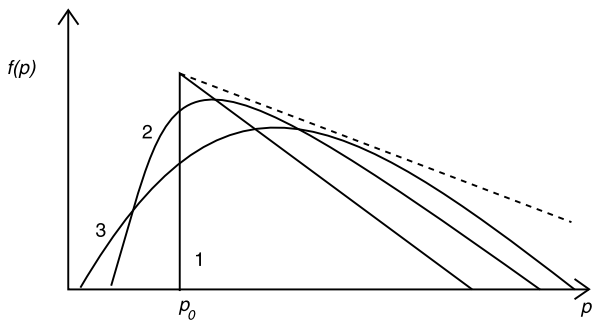

Acceleration Mechanisms, Figure 2
The distribution function $f_i'(p, p_0)$ for particles accelerated after $i = 1$–3 shock. The *dashed line* indicates the asymptotic solution, $f_\infty(p)$ given by (18)

The model implies that downstream of the nth shock (after decompression) the distribution is

$$f_n'(p) = \sum_{i=1}^{n} f_i'(p, p_0)$$
$$f_i'(p, p_0) = \frac{Kb^i}{p_0}\left(\frac{p}{R^i p_0}\right)^{-b} \frac{(\ln\, p/R^i p_0)^{i-1}}{(i-1)!}\,. \tag{18}$$

After an arbitrarily large number of shocks, the spectrum approaches $f_\infty(p) \propto p^{-3}$ at $p > p_0$, as illustrated in Fig. 2. This spectrum is harder, by one power of p, than the spectrum $\propto p^{-4}$ from a single strong shock. Moreover, the spectrum $f_\infty(p) \propto p^{-3}$ is approached irrespective of the strength of the shocks, although the stronger the shocks the faster it is approached. An interpretation is that the combination of DSA and decompression leads to a Fermi-like acceleration mechanism: the asymptotic solution for Fermi-type acceleration for constant injection at $p = p_0$ is a power law with $b = 3$ for $p > p_0$. Such a distribution corresponds to a flat synchrotron spectrum: the intensity of synchrotron radiation is a power law in frequency, ν, with $I(\nu) \propto \nu^{-\alpha}, \alpha = (b-3)/2$.

Inclusion of Synchrotron Losses

Synchrotron losses cause the momentum of the radiating electron to decrease at the rate $\mathrm{d}p/\mathrm{d}t = -C_1 B^2 p^2$, with C_1 a constant. Synchrotron losses limit DSA: there is a momentum, $p = p_c$, above which the synchrotron loss rate exceeds the rate of acceleration, and acceleration to $p > p_c$ is not possible. The average acceleration rate over one cycle due to DSA is $\mathrm{d}p/\mathrm{d}t = \Delta(p)/t_c = C_2 p$, and with the inclusion of synchrotron losses this is replaced by $\mathrm{d}p/\mathrm{d}t = C_2 p - C_1 B^2 p^2 = C_2 p(1 - p/p_c)$ with $p_c = C_2/C_1 B^2$. It is straightforward to repeat the calculation of DSA at multiple shocks including the synchrotron losses [75]. In Fig. 3 the logarithm of the distribution function is plotted as a function of $\log(p/p_0)$. The synchrotron cutoff momentum is chosen to be $p_c = 10^3 p_0$, and all the shocks have $r = 3.8$. The distribution below the synchrotron cutoff, $p < p_c$, due to the cumulative effect of injection at every shock, becomes harder than $b = 3$, such that the slope has a peak (with $b \approx 2$) just below p_c. Such a distribution corresponds to a weakly inverted spectrum with a peak just below a sharp cutoff due to synchrotron losses.

Applications of DSA

There in an extensive literature on applications of DSA, and only a few remarks on some of the applications are made here. The applications are separated into astrophysical, solar and space-physics, and the emphasis in each of these in qualitatively different. In astrophysical applications the emphasis is on the acceleration of highly relativistic particles, particularly GCRs and synchrotron-emitting electrons. In solar applications one needs to explain detailed data on the properties of the solar energetic particles (SEPs) observed in space, and on the radio, X-ray and gamma-ray signatures of them in the solar atmosphere. In space-physics, the objective is to model the detailed distributions of particles and fields measured in situ. The comments here indicate these different emphases.

Acceleration in Young Supernova Remnants

Recent observational and theoretical developments have been combined in the currently favored suggestion that GCRs are accelerated at the shocks where young supernova remnants (SNRs) interact with the interstellar medium. The observational development includes X-ray, gamma ray and TeV photon data from young supernovae, including the remnants of historical supernovae $\lesssim 10^3$ yr old, that show edge-brightening, which is interpreted in terms of synchrotron emission from relativistic electrons accelerated at an outer shock [43,101,104]. The suggestion is that GCRs are also accelerated at these shocks. The theoretical development is the suggestion [11,12] that the CRs can cause the magnetic field to be amplified by a very large factor in association with these shocks.

Since the original suggestion by Baade and Zwicky, it has been widely believed that the acceleration of GCRs is due to shocks generated by supernovae, and DSA provides an effective mechanism by which this occurs. Hillas [43] gave four arguments in favor of the location being the shocks in young SNRs:

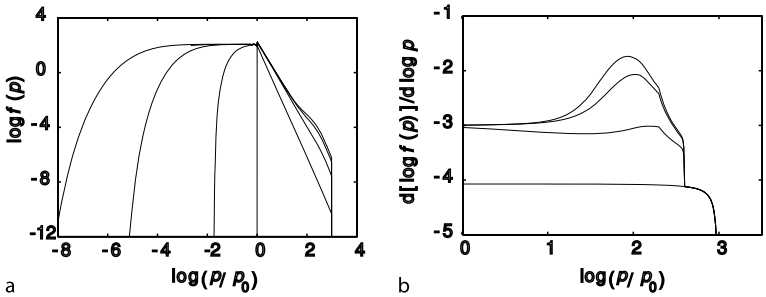

Acceleration Mechanisms, Figure 3
The cumulative effect of DSA plus synchrotron losses after many shocks with injection at each shock; a the distribution, b the slope of the distribution for $p > p_0$: $r = 3.8$, $p_c/p_0 = 10^3$, $N = 1, 10, 30, 50$

(a) DSA produces a power-law spectrum of about the right slope.
(b) The model is consistent with the long-standing arguments that the only adequate energy source for GCRs is supernovae.
(c) The composition of GCR1 is consistent with ordinary interstellar matter being injected into DSA, and it is at least plausible that such composition-independent injection occurs at such shocks.
(d) When coupled with self-generation of magnetic field, DSA can explain the energy at which the knee between GCR1 and GCR2 occurs [44], cf. however [81].
Although the details are not complete, these arguments provide a plausible overview as to how GCR1 are accelerated.

The amplification of the magnetic field is attributed to the nonlinear development of a nonresonant plasma instability driven by the CRs [11,12,86]. A strong, self-generated magnetic field allows DSA to accelerate particles to higher energy than otherwise, due to the maximum energy being proportional to B.

Acceleration in Astrophysical Jets

Astrophysical jets are associated with accretion discs, around proto-stars and various classes of compact objects, including the supermassive black holes in active galactic nuclei (AGN). Radio jets associated with AGN can extend to enormous distances from the AGN; these jets require that the synchrotron-emitting electrons be accelerated in situ [16]. The appearance (characterized by various knots) of the jet in M87 [70] in radio and optical is remarkably similar, whereas the much shorter synchrotron loss time of the optically-emitting electrons leads one to expect the optical emission to be much more sharply localized about acceleration sites. A possible model for the acceleration involves DSA at multiple shocks, which develop naturally in the flow, provided the acceleration is sufficiently fast to overcome the losses of the optically-emitting electrons. The required scattering of the optically-emitting electrons can plausibly be attributed to a Kolmogorov spectrum of turbulence [76].

A subclass of radio sources associated with AGN have flat or inverted spectra, corresponding to power laws $I(\nu) \propto \nu^{\alpha}$ with $\alpha \approx 0$ or $\alpha > 0$ over at least a decade in frequency. Synchrotron self-absorption was proposed for such spectra [51]. A specific model can account for flat spectra [68], but it requires such special conditions that it was referred to as a "cosmic conspiracy" [22]. A possible alternative explanation, for spectra with $\alpha < 1/3$, is in terms of an acceleration mechanism that produces a power-law electron spectrum with $b = 3 - 2\alpha$. DSA at a single nonrelativistic shock produces a spectrum with $b > 4$ corresponding to $\alpha < -0.5$. A relativistic shock can produce a somewhat harder spectrum but cannot account for flat spectra. DSA at multiple shocks tends towards $b = 3$, $\alpha = 0$, and provides a possible alternative explanation for flat spectra. The pile-up effect [87] occurs naturally when synchorotron losses are included in DSA at multiple shocks [75], and can account for inverted spectra with $\alpha \leq 1/3$.

Solar Energetic Particles

It has long been recognized that acceleration of fast particles occurs in the solar corona in connection with solar flares. There is a power-law distribution of flares in energy or in area: the energy released in a flare is approximately proportional to the area that brightens in Hα. All flares produce fast nonrelativistic (10–20 keV) electrons, observed through their radio (type III bursts) or

X-ray (bright points) emission. Large flares also produce energetic ions and relativistic electrons. In the early literature [108], it was assumed that the acceleration occurs in two phases: the first (impulsive) phase involving nonrelativistic electrons (and perhaps ions), attributed to some form of "bulk energization", and the second phase involves a slower acceleration of solar CRs, attributed to the shocks [59], such as those that produce type II radio bursts. However, this simple picture is not consistent with other data: gamma-ray data show that the relativistic electrons and energetic ions appear immediately after the onset of the flare, without the predicted delay. Spacecraft data on solar energetic particles (SEPs) reveal further phenomena and correlations that are not consistent with the two-phase model. Various different acceleration mechanisms have been proposed for SEPs [77].

A new paradigm has emerged for the interpretation of SEPs observed in space: events are separated into flare-associated SEPs and CME-associated SEPs [62]. In a CME (coronal mass ejection) a previously magnetically bound mass of corona plasma becomes detached, accelerates away from the Sun, and drives a shock ahead of it. The detachment of a CME involves some form of magnetic reconnection in the corona. Earlier ideas on a tight correlation between flares and CMEs have not been confirmed by more recent data. Although there is some correlation, it is not one-to-one, and flares and CMEs are now regarded as distinct phenomena. The acceleration of CME-associated SEPs is plausibly due to DSA, but how the flare-associated SEPs are accelerated remains poorly understood.

There are anomalies in the ionic composition in flare-associated SEPs, the most notable of which concerns ^{3}He. The ratio of ^{3}He to ^{4}He in the solar photosphere is 5×10^{-4}, but the ratio in flare-associated SEPs is highly variable, and can be greatly enhanced, even exceeding unity in exceptional cases. A plausible explanation for ^{3}He-rich events is preferential preacceleration feeding into DSA. A specific preacceleration mechanism that selects ^{3}He involves the generation of ion cyclotron waves between the proton and ^{4}He cyclotron resonances, with these waves damping by accelerating ^{3}He [32]. The formulation of a self-consistent model for the acceleration of flare-associated SEPs is complicated by the paucity of direct signatures of processes in the solar corona, and the difficulty in reconciling the data on various emissions from energetic particles precipitating into the photosphere with the SEP data.

Bulk energization of electrons is not understood. A large fraction of the energy released in a flare goes into bulk energization, requiring acceleration of virtually all the electrons in the flaring region of the corona. One model that involves shock acceleration is illustrated in Fig. 4: magnetic reconnection leads to an outflow of plasma from the reconnection site, with the upflow associated with a CME and the downflow running into an underlying closed magnetic loop. In such a model, the electron heating is identified with shock acceleration where the downflow is stopped by the underlying loop. There are also various non-shock-associated mechanisms, involving $E_{\|}$, for bulk energization.

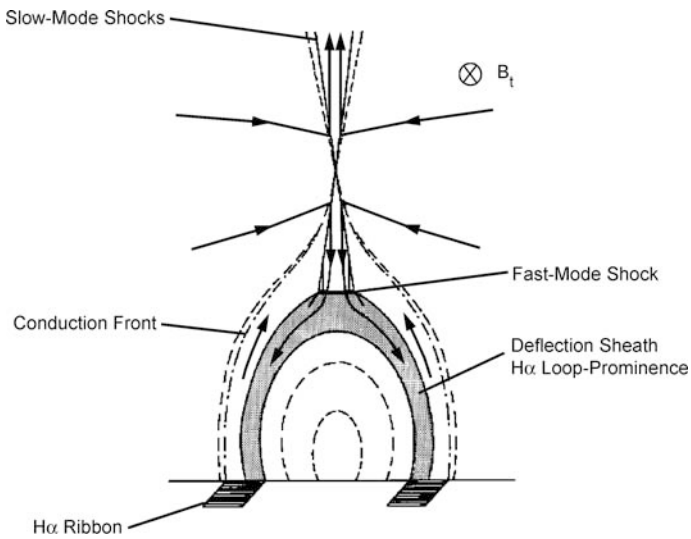

Acceleration Mechanisms, Figure 4
A cartoon [33] illustrating one suggestion for how magnetic reconnection leads to acceleration of particles in a solar flare

Resonant Scattering

The theory of resonant scattering was developed initially in two different contexts: the diffusion of GCRs through the ISM and the formation and stability of the Earth's Van Allen belts.

Streaming of GCRs

A simple model for CRs streaming along the field lines at v_{CR} is

$$f(p,\alpha) = f_0(p)\left(1 + \frac{v_{\mathrm{CR}}}{v}\cos\alpha\right),$$
$$f_0(p) = K_{\mathrm{CR}}\left(\frac{p}{p_0}\right)^{-b}, \tag{19}$$

where the non-streaming part of the distribution function, $f_0(p)$, is measured to be a power law above some minimum p_0, with K_{CR} a normalization constant.

The anisotropic CRs can generate the resonant waves that scatter them. Growth of Alfvén (A) and magnetoacoustic (M) waves may be described in terms of a negative absorption coefficient. On evaluating the absorption coefficient using the general expression (32) and inserting numerical values for CRs one finds

$$\gamma_{\mathrm{A},M}(\omega,\theta) = -2.7\times 10^{-7}\left(\frac{n_{\mathrm{e}}}{1\,\mathrm{cm}^{-3}}\right)^{-1/2}\left(\frac{p}{p_0}\right)^{-1.6}\left(\frac{\cos\theta}{|\cos\theta|}\frac{v_{\mathrm{CR}}}{c} - \frac{b}{3}\frac{v_{\mathrm{A}}}{c}\right), \tag{20}$$

where $|\cos\theta|$ is approximated by unity. It follows from (20) that the waves grow in the forward streaming direction ($\cos\theta > 0$ for $v_{\mathrm{CR}} > 0$) for $v_{\mathrm{CR}} > (b/3)v_{\mathrm{A}}$. The scattering of the CRs by the waves generated by the streaming particles themselves causes v_{CR} to reduce. The rate of reduction may be evaluated using the quasilinear equation written down in the Appendix. One finds

$$\frac{\mathrm{d}v_{\mathrm{CR}}}{\mathrm{d}t} = -\nu_s\left(v_{\mathrm{CR}} - \frac{\zeta-1}{2}\frac{|\cos\theta|}{\cos\theta}\frac{b}{3}v_{\mathrm{A}}\right), \tag{21}$$

with $\zeta = (\mathcal{F}_{\mathrm{A}}^{+} - \mathcal{F}_{\mathrm{A}}^{-})/(\mathcal{F}_{\mathrm{A}}^{+} + \mathcal{F}_{\mathrm{A}}^{-})$, where $\mathcal{F}_{\mathrm{A}}^{\pm}$ are the fluxes of Alfvén waves in the forward (+) and backward (−) streaming directions, and with the scattering frequency given by

$$\nu_s = 3\int_{-1}^{+1}\mathrm{d}\cos\alpha\,\sin^2\alpha\,D_{\alpha\alpha}, \tag{22}$$

with $D_{\alpha\alpha}$ determined by (34).

The growth rate (20) is fast enough to account for growth of the waves that resonate with lower-energy GCRs, and the scattering time is fast enough to reduce the streaming to close to $(b/3)v_{\mathrm{A}}$. However, the growth time increases rapidly with p, with an approximate one-to-one relation between the resonant wave number k and the momentum of the resonant particles $kp = |q|B$. For $p \gg 10^{13}\,\mathrm{eV}/c$ the growth time becomes ineffective, and self-generation of the resonant waves cannot account for efficient scattering.

Scattering of Higher Energy CRs

For higher-energy GCRs, effective scattering requires an external source of MHD waves. There is a turbulent spectrum in the ISM, and the observations [5] show it to be a Kolmogorov-like spectrum

$$W(k) \propto k^{-5/3}, \tag{23}$$

where $W(k)$ is the energy in the waves per unit wavenumber k. Resonant scattering by a Kolmogorov spectrum implies a scattering frequency that decreases only slowly, $\nu_s \propto p^{-1/3}$, with increasing p. This scattering must dominate scattering by self-generated waves at higher p. It is likely that the distribution of turbulence is far from homogeneous, and that the most effective scattering occurs in localized regions of enhanced turbulence. How effective scattering of higher energy particles occurs is a poorly understood aspect of acceleration mechanisms.

Nonresonant Versus Resonant Growth

Plasma instabilities can occur in two forms: resonant and nonresonant. A nonresonant instability involves an intrinsically growing wave, described by a complex solution of a real dispersion equation. A recent suggestion that has stimulated considerable interests involves an instability [11,12] that is a nonresonant counterpart of (20) [73]. The potential importance of this instability is twofold. First, it can allow the lower-energy particles to generate a wide spectrum of waves including those needed to scatter higher-energy particles. Second, its nonlinear development can lead to amplification of the magnetic field, which ultimately causes the instability to saturate [11,12,86]. A similar idea has been proposed in connection with generation of magnetic fields through a Weibel instability associated with the relativistic shocks that generate gamma-ray bursts [107]. Such nonresonant growth is an important ingredient in the currently favored acceleration mechanism for GCR1 [43].

Formation of the Van Allen Belts

A simple model for the formation of the radiation belts is based on adiabatic invariants, which are associated with quasi-periodic motions. The first adiabatic invariant is associated with the gyration about a field line, and it implies that $M \propto p_\perp^2/B$ is an invariant. After averaging over the gyrations, a fast particle bounces back and forth along a field line between two reflection points. The adiabatic invariant associated with this bounce motion, denoted J, is the integral of $p_\parallel$ along the field lines between the reflection points. After averaging over bounce motion, the remaining motion is a periodic drift around the Earth. There is a third adiabatic invariant, Φ, which corresponds to the integral for the vector potential around the Earth. For a dipole-like magnetic field it is convenient to label a field line by the radial distance, L, to its midpoint. Then one has $M \propto p_\perp^2 L^3$, $J \propto p_\parallel L$, $\Phi \propto 1/L$. Violation of the third adiabatic invariant, due to perturbations in the Earth magnetic field with a period equal to that of the drift motion, causes the particles to diffuse in L at constant M and J. The steady state density that would result from such diffusion is proportional to $p_\perp^2 p_\parallel \propto 1/L^4$. Particles that enter the magnetosphere from the solar wind tend to diffuse inwards to build up this density. As the particles diffuse to smaller L, their momentum increases according to $p \propto L^{-3/4}$. The inward diffusion ceases at the point where pitch-angle scattering (which violates conservation of M) becomes effective and causes the particles to be scattered into the loss cone and to precipitate into the Earth's neutral atmosphere. Due to their loss-cone anisotropy, illustrated schematically in Fig. 5, the anisotropic particles can generate the resonant waves needed to scatter them, with the growth rate increasing rapidly with decreasing L. In this way, resonant scattering provides the sink (at smaller L) that allows the Van Allen belts to remain in a steady state when the magnetosphere is quiescent. The absorption coefficient for the waves tends to be maintained near zero, referred to as marginal stability, such that sporadic localized bursts of wave growth lead to sporadic localized bursts of particle precipitation. Such bursts of wave growth can be triggered by artificially generated VLF signals [42].

Acceleration by Parallel Electric Fields

Acceleration by an electric field parallel to the magnetic field, $E_\parallel$, is the simplest conceivable acceleration mechanism, but also the least understood. The argument for acceleration by $E_\parallel$ is compelling in connection with aurorae, is very strong in connection with pulsars, and is strong for bulk energization in solar flares.

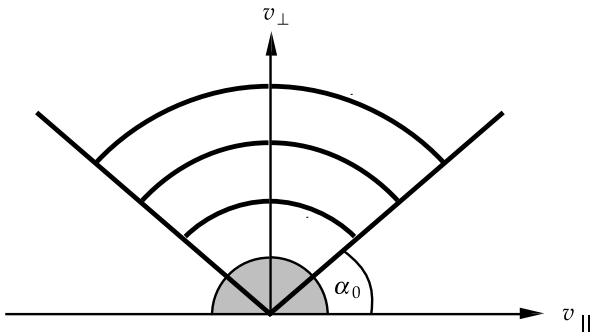

Acceleration Mechanisms, Figure 5
An idealized loss-cone distribution is illustrated; α_0 is the loss cone angle, the *shaded region* is filled by isotropic thermal particles, and the *circular arcs* denote contours of constant *f*

Available Potential

It is sometimes helpful to note that all relevant acceleration mechanisms can be attributed to an inductive electric field, and to identify the specific inductive field. In particular, the maximum available potential, or EMF (electromotive force), determines the maximum energy to which particles can be accelerated. There are two general useful semi-quantitative estimates of the available potential: one is the EMF associated with changing magnetic flux through a circuit; the other involves the power radiated by a rotating magnet in vacuo.

EMF Due to a Relative Motions The EMF in a simple circuit, due to changing magnetic flux through the circuit, is $\Phi = -\mathrm{d}(BA)/\mathrm{d}t$, where A is the area of the circuit. Although Φ can be due to a changing magnetic field, $\Phi = A\mathrm{d}B/\mathrm{d}t$, in most cases of interest it is due to relative motions, $\Phi = B\mathrm{d}A/\mathrm{d}t = BLv$, where L and v need to be identified in a specific model. A $\Phi = BLv$ appears along a magnetic flux tube line that connects two MHD regions in relative motion to each other, with v the relative velocity perpendicular to the flux tube, with a magnetic field B in one of the regions, with L the perpendicular dimension of the flux tube orthogonal to the relative velocity in this region. This simple argument needs to be complemented with an identification of how the current closes, and how this potential is distributed around the closed circuit. This model provides a plausible estimate for the energy of auroral electrons and for electrons accelerated in Jupiter's magnetosphere. In the case of auroral particles, the velocity, v, is between the Earth and the solar wind, and the acceleration occurs along the auroral field lines that connect these two regions. Parameters B and L across the magnetotail and v for the solar wind, give Φ of several kV,

consistent with the observed auroral electrons of a several keV. In an early model for the Io–Jupiter interaction, the moon Io was assumed to drag the Jovian magnetic field lines threading it through the corotating magnetosphere, leading to the estimate $\Phi = B_{\rm Io} L_{\rm Io} v_{\rm Io}$, with $B_{\rm Io}$ is the Jovian magnetic field at the orbit of Io, $L_{\rm Io}$ is the diameter of Io and $v_{\rm Io}$ is the relative velocity of Io to the corotating magnetic field lines [37]. This gives Φ of a few MV, consistent with a copious supply of few-MeV electrons associated with Io's motion through the Jovian magnetosphere.

Another application of this estimate of Φ provides a limit on the maximum energy to which particles can be acceleration due to DSA. The relative flow on opposite sides of the shock implies an electric field in the plane of the shock front: one may identify this electric field as Bv, with v the relative flow speed perpendicular to **B**. Then $\Phi = BvL$ is the potential available across the lateral extent, L, of the shock. The maximum energy to which a particle can be accelerated through DSA is limited to $< eBvL$ per unit charge.

A point of historical interest is that the earliest recognized suggestion for acceleration of CRs involved an inductive field due to $\mathrm{d}B/\mathrm{d}t$: in 1933 Swann [94] proposed that the acceleration occurs in starspots due to the magnetic field increasing from zero to $B = 0.2\,\mathrm{T}$ in a time $\tau = 10^6\,\mathrm{s}$ in a region of size $L = 3 \times 10^8\,\mathrm{m}$, giving $\Phi = BL^2/\tau = 2 \times 10^{10}\,\mathrm{V}$. Although this model is not considered realistic, it is of interest that a similar value for Φ results from $\Phi = BLv$ with the same B, L and with L/τ replaced by $v = 3\,\mathrm{km\,s^{-1}}$; these numbers are close to what one would estimate for the potential difference between field lines on opposite sides of a rotating sunspot. Such large potential drops are available in the solar atmosphere, but the implications of this are poorly understood.

EMF Due to a Rotating Compact Object In a simple model for the electrodynamics of a rotating magnetized compact object, such as a neutron star or a black hole, the slowing down is estimated assuming that the rotational energy loss is due to magnetic dipole radiation. The power radiated may be written as $I\Phi$, where the effective displacement current, I, and the available potential, Φ, are related by $\Phi = Z_0 I$, with $Z_0 = \mu_0 c = 377\,\Omega$ the impedance of the vacuum. Equating Φ^2/Z_0 to the rate of rotational energy loss provides as estimate of Φ. Although the slowing down for pulsars is not actually due to magnetic dipole radiation, estimates based on this model underpin the basic interpretation of pulsars, including estimates of the age, the surface magnetic field and the available potential. The Φ may be interpreted as between the stellar surface and the pulsar wind region outside the light cylinder, leading to an $E_\parallel$ along the polar-cap field lines that connect these regions.

Acceleration of Auroral Electrons

Magnetic reconnection in the Earth's magnetotail, in a so-called magnetospheric substorm, results in energetic electrons precipitating into the upper atmosphere, producing the visible aurora. A build up of the magnetic stresses prior to a substorm results from reconnection, at the sunward edge of the magnetosphere, between terrestrial field lines and with interplanetary field lines, with the solar wind transferring magnetic flux from the front of the magnetosphere to the magnetotail. There is a build up of the cross-tail current associated with the build up of the magnetic flux. A generic model for a substorm [103] involves relaxation of these stresses through reconnection, which creates closed field lines, and allows transfer of magnetic flux to the front of the magnetosphere as these reconnected field rotate. The reconnection also causes the cross-tail current to be partly redirected around a circuit along the reconnecting field lines, so that the current closes by flowing along field lines to the ionosphere, across the field lines due to the Pedersen conductivity there, and returning to the magnetotail along a nearby field lines. The magnetic energy released is goes partly into an Alfvénic flux towards the Earth, and this energy flux is partially converted into energetic electrons in an acceleration region. The acceleration region is far removed from the reconnection site, and is at a height $> 10^3$ km, well above the location of the visible aurora, as indicated schematically Fig. 6. The acceleration is due to an $E_\parallel$ in the acceleration region, but the mechanism that sets up the $E_\parallel$ is still not clear. Three ideas are discussed briefly here.

A parallel potential difference develops between the ionosphere, $\Phi_{\rm ion}$, and the magnetosphere, $\Phi_{\rm mag}$, along a current-carrying field line. The current implies a flow of electrons relative to ions. There is a limited supply of magnetospheric electrons to carry the current, due to most electrons mirroring before they reach the ionosphere. The potential difference develops to allow the current to flow. Let $R_{\rm m}$ be the mirror ratio, along a current-carrying field line. For $1 \ll e\Phi_\parallel/T_{\rm e} \ll R_{\rm m}$, the so-called Knight relation [54,63,102] gives

$$J_\parallel \approx K(\Phi_{\rm ion} - \Phi_{\rm mag})\,, \quad K = \frac{\mathrm{e}^2 n_{\rm e}}{(2\pi)^{1/2} m_{\rm e} V_{\rm e}}\,, \qquad (24)$$

where $T_{\rm e} = m_{\rm e} V_{\rm e}^2$ is the electron temperature. The Knight relation (24) is useful for a number of purposes, but it is not particularly useful in considering how the potential localizes to create the $E_\parallel$ in the acceleration region.

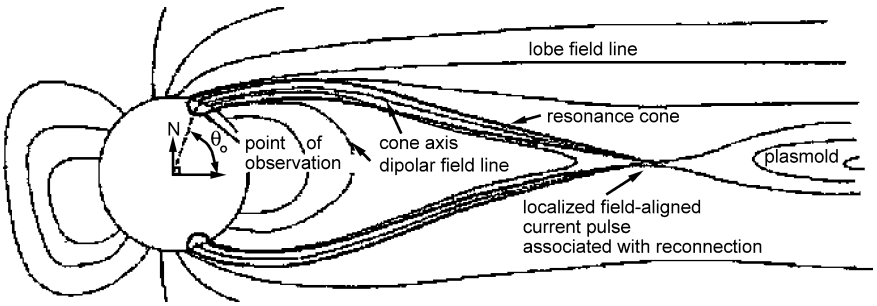

Acceleration Mechanisms, Figure 6
A cartoon [13] showing the regions in the Earth's magnetosphere associated with the acceleration of auroral electrons. The generation region is where magnetic reconnection occurs, launching Alfvén waves, which accelerate electrons in an acceleration region where the EMF from the generator is localized along field lines

In situ observations of $E_\parallel$ in the acceleration region have been interpreted in a variety of ways. Localized structures sweeping across a spacecraft were initially called electrostatic shocks [78] and also referred to as double layers (DLs) [19,28]. The theory of DLs is well established for laboratory plasmas, and models for DLs have been adapted to space and astrophysical plasma applications [18,85]. DLs are classified as weak or strong, depending on whether the potential difference is of order or much greater than T_e/e, respectively. Some models for the formation and structure of DLs involve two other ideas that have been discussed widely in the literature: anomalous resistivity and electron phase space holes [80]. However, the details of how localized DLs accelerate particles and produce the observed fluxes of energetic electrons is still not clear.

Another way in which an $E_\parallel$ can arise is through the Alfvén waves that transport the energy released in the magnetotail [92,103,109]. The condition $E_\parallel = 0$ is strictly valid for an Alfvén wave only in ideal MHD, and $E_\parallel \neq 0$ can arise due to electron inertia, due to finite thermal gyroradii and due to (anomalous) resistivity. Low-frequency Alfvén waves with large $k_\perp$ become either inertial Alfvén waves (IAW) or kinetic Alfvén waves, depending on whether inertia or thermal effects dominate. The dispersion relation and the ratio of the parallel to the perpendicular electric fields in an IAW depend on the skin depth, $\lambda_e = c/\omega_p$:

$$\frac{\omega}{|k_\parallel|} = \frac{v_A}{(1 + k_\perp^2 \lambda_e^2)^{1/2}}\,, \qquad \frac{E_\parallel}{E_\perp} = \frac{k_\parallel k_\perp \lambda_e^2}{1 + k_\perp^2 \lambda_e^2}\,. \tag{25}$$

The $E_\parallel$ associated with IAWs is one suggested mechanism for auroral acceleration: the skin depth is several kilometers, of the order, 10 km, of observed changes in $J_\parallel$ across field lines [103,109]. Acceleration by IAWs is a relatively old idea [36], favored for electrons accelerated by Alfvén waves generated by Io's motion through the Jovian magnetosphere [35]. An important detail is that this $E_\parallel$, being a wave field, reverses sign, accelerating electrons in opposite directions, every half wave period.

Another relevant idea is that of an Alfvénic resonator. Inhomogeneities along field lines cause v_A to change, and this leads to reflection of Alfvén waves at the top of the ionosphere. Suppose there is also reflection at some higher region. This combination leads to localized trapping of wave energy in the Alfvén resonator between the reflection points [29,64]. A self-consistent acceleration model requires that the upper reflection site be where the $E_\parallel$ develops and the electrons are accelerated, perhaps resulting in a localized current instability and associated anomalous resistivity that cause the Alfvén waves to be reflected.

An extreme example of auroral acceleration occurs in inverted-V electrons. The upward current can essentially depopulate a localized auroral flux tube of all thermal electrons, leading to a plasma cavity in which the inverted-V electrons are observed, Fig. 7. The "inverted-V" refers to the dynamic spectrum observed by a spacecraft as it crosses the cavity: lowest energies near the edges and maximum energy near the center. Inverted-V electrons correlate with auroral kilometric radiation [39], which is due to electron cyclotron maser emission.

One can conclude that despite significant progress, the details of the acceleration of auroral electrons still need further investigation [20].

Acceleration in Pulsars Gaps

Pulsars are rapidly rotating neutron stars with superstrong magnetic fields. There are three broad classes: ordinary, recycled (or millisecond), and magnetars; only ordinary pulsars are considered here. These have periods $P \approx 0.1$–1 s, surface magnetic fields $B \approx 10^7$–10^9 T and are slow-

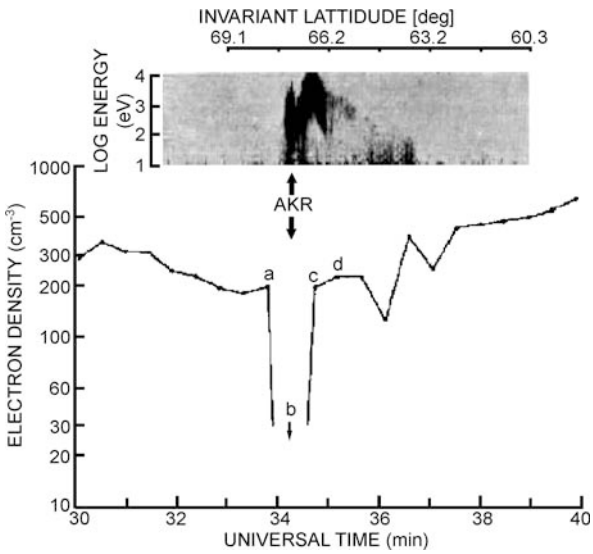

Acceleration Mechanisms, Figure 7
A density depletion in an auroral region of the Earth; inverted-V electrons carry a large upward current in such regions [14]

ing down at a rate $\dot{P} \sim 10^{-15}$. The age, $P/2\dot{P}$, and the surface magnetic field, $B \propto (P\dot{P})^{1/2}$, are estimated assuming the rotating magnetic dipole model in which the rotational energy loss is due to magnetic dipole radiation. Equating the energy loss due to the observed slowing down to Φ^2/Z_0 gives an estimate for the potential available along field lines in the polar-cap region. For the Crab pulsar the rotational energy loss is approximately 10^{31} W, implying $\Phi \approx 6 \times 10^{16}$ V; the available potentials in most other pulsars are somewhat smaller than this. In the most widely favored model, the electric field is assumed to cause a charge-limited flow from the stellar surface, providing the "Goldreich–Julian" charge density required by the corotation electric field. However, the Goldreich–Julian charge density cannot be set up globally in this way, and requires additional source of charge within the magnetosphere. The available potential is assumed to develop across localized regions called "gaps" where acceleration of "primary" particles, due to $E_\parallel \neq 0$, results in the emission of gamma rays which trigger a pair cascade that produces the necessary secondary (electron–positron) charges. This additional source of charge is assumed to screen $E_\parallel$ outside the gaps. However, there is no fully consistent model and no consensus on the details of the structure and location of the gaps. The present author believes that existing models, based on a stationary distribution of plasma in a corotating frame, are violently unstable to temporal perturbations, resulting in large-amplitude oscillations in $E_\parallel$ [61].

Although the details of acceleration by $E_\parallel$ are quite different in the pulsar case from the auroral case, the underlying difficulties in understanding such acceleration are similar. A static model for acceleration by $E_\parallel$ encounters seemingly insurmountable difficulties, and an oscillating model involves unfamiliar physics.

Acceleration of Nonrelativistic Solar Electrons

A large fraction of the energy released in a solar flare goes into very hot electrons: all the electrons in a relatively large volume are accelerated in bulk to 10–20 keV. Several models for bulk heating involving acceleration by $E_\parallel$ have been proposed, but all have unsatisfactory features. One idea is to invoke a runaway process [45]. Consider the equation of motion of an electron in an electric field subject to Coulomb collisions with a collision frequency $\nu_e(v) = \nu_e(V_e/v)^3$, where V_e is the thermal speed of electrons:

$$\frac{d\mathbf{v}}{dt} = -\frac{e\mathbf{E}}{m} - \nu_e \mathbf{v}\left(\frac{V_e}{v}\right)^3 . \tag{26}$$

It follows that electrons with speed

$$\frac{v}{V_e} > \left(\frac{E_D}{E}\right)^{1/2}, \quad E_D = \frac{mV_e\nu_e}{e}, \tag{27}$$

are freely accelerated; E_D is called the Dreicer field. This so-called electron runaway sets up a charge separation that should quickly screen $E_\parallel$ and suppress the acceleration. It also causes $J_\parallel$ to change, and this is opposed by inductive effects. Runaway on its own is not a plausible acceleration mechanism.

There are inconsistencies in estimates of the available potential, Φ. Assuming that bulk energization is due to acceleration by an $E_\parallel$ implies $\Phi \approx 10$–20 kV, that is of order 10^4 V. Two arguments suggest a much larger potential. On the one hand, there is observational evidence that large flares tend to occur in flux loops carrying large currents, $I \gtrsim 10^{12}$ A. If one identifies I with the observed current, then to produce the power $I\Phi$ of order 10^{22} W released in a large flare requires Φ of order 10^{10} V. On the other hand, as indicated above, the potential available due to photospheric motions is estimated above to be 10^9–10^{10} V. This potential cannot be ignored when two magnetic loops come in contact and reconnect: the potential difference between the reconnecting field lines is expected to be of order 10^9–10^{10} V.

There are several problem that have yet to be clearly resolved. One problem is how the required $E_\parallel \neq 0$ develops. The suggestions as to how this problem is resolved in the

auroral acceleration region need to be explored for the solar application. (An exception is inertial Alfvén waves: the skin depth being far too small to be relevant in the solar corona.) The suggestion that double layers might play an important role in solar flares has a long history [85], as do the suggestions involving anomalous resistivity [26] and an Alfvénic resonator [47,48]. However, none of the existing models is satisfactory. Another problem is how to reconcile the estimates $\Phi \approx 10^4$ V and 10^9–10^{10} V. It appears that the available 10^9–10^{10} V must localize into $> 10^5$ regions with $E_\parallel \neq 0$ and $\Phi \approx 10^4$ V, where the acceleration occurs, such that the accelerated electrons that escape from one region do not pass through any other. A third problem is that, like the auroral and pulsar cases, any model for acceleration by a static $E_\parallel$ encounters serious difficulties. Although a large-scale oscillating $E_\parallel$ might plausibly overcome these difficulties, there is no model for such oscillations.

Other Acceleration Mechanisms

The three most important acceleration mechanisms are DSA, stochastic acceleration and acceleration by parallel electric fields. Various other mechanisms are relevant in specific applications. Examples mentioned above are the acceleration of energetic particles in the Earth's Van Allen belts and various preacceleration mechanisms for DSA. Some comments are made here on three other mechanisms: shock drift acceleration, gyroresonant acceleration and acceleration in current sheets.

Shock Drift Acceleration (SDA)

When a particle encounters a shock it is either transmitted across the shock front or reflected from it, and in either case it tends to gain energy [96]. Shock drift acceleration (SDA) is attributed to the grad-B drift when the particle encounters the abrupt change in B: the scalar product of drift velocity and acceleration by the convective electric field, $-q\mathbf{u} \times \mathbf{B}$, is positive, implying energy gain. The energy gained by a particle depends on how far it drifts along the front. A simple model for DSA is based on $p_\perp^2/B$ being conserved when the particle encounters the shock. This conservation law is somewhat surprising (it cannot be justified on the usual basis that $p_\perp^2/B$ is an adiabatic invariant), and it applies only approximately and only after averaging over gyrophase [25,84].

In treating SDA it is convenient to make a Lorentz transformation to the de Hoffmann–Teller frame. The frame most widely used in discussing a shock is the shock-normal frame, in which the shock front is stationary and the upstream flow velocity is normal to it. In this frame

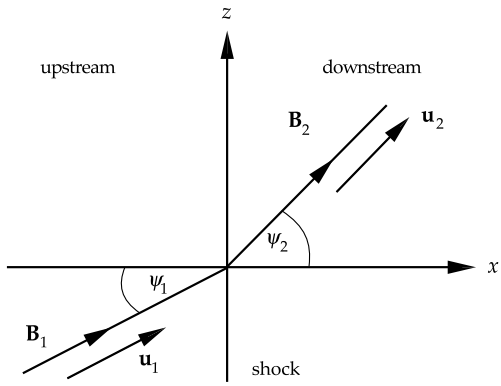

Acceleration Mechanisms, Figure 8
In the de Hoffmann–Teller frame the fluid velocities are parallel to the magnetic fields on either side of the shock

there is a convective electric field, $-\mathbf{u} \times \mathbf{B}$, on both sides of the shock, and the point of intersection of $\mathbf{B}$ with the shock moves in the plane of the shock. In the de Hoffmann–Teller frame this point of intersection is stationary, $\mathbf{u}$ and $\mathbf{B}$ are parallel to each other on both sides of the shock, and there is no electric field, Fig. 8. The de Hoffmann–Teller frame exists only for "subluminal" shocks for which the velocity of intersection is less than c; in the "superluminal" case, where the velocity of intersection is greater than c, there exists a frame in which $\mathbf{B}$ is perpendicular to the shock normal. In the de Hoffmann–Teller frame the energy of a particle is unchanged when it is transmitted across the shock, or is reflected from the shock. A simple theory for SDA involves the energy being conserved in the de Hoffmann–Teller frame on transmission or reflection, with the implications of this being found by Lorentz transforming to the shock-normal frame.

Let quantities in the de Hoffmann–Teller and shock-normal frames be denoted with and without primes, respectively. Let ψ_1 be the angle between $\mathbf{u}_1$ and $\mathbf{B}_1$. Provided that the relative velocity between the two frames is nonrelativistic, the angle ψ_1 is the same in the two frames, and the relative velocity is $u_0 = u_1 \tan\psi_1$. In the nonrelativistic case, the pitch angles in the two frames are related by $v' \cos\alpha' = v\cos\alpha + u_0$, $v' \sin\alpha' = v \sin\alpha$. Reflection is possible only for $\sin^2\alpha_1' \geq B_1/B_2$, and the maximum change in energy occurs for reflected particles at the threshold, $\alpha_1' = \alpha_c$, $\sin^2\alpha_c = B_1/B_2$. The maximum ratio of the energy after reflection (ref) to before (inc) is for a particle with $v_1 \cos\alpha_1 = -u_0(1-\cos\alpha_c)$, that is propagating away from the shock, and is overtaken by the shock. The maximum ratio is

$$\left(\frac{E_{\mathrm{ref}}}{E_{\mathrm{inc}}}\right)_{\max} = \frac{1+(1-B_1/B_2)^{1/2}}{1-(1-B_1/B_2)^{1/2}}\,, \tag{28}$$

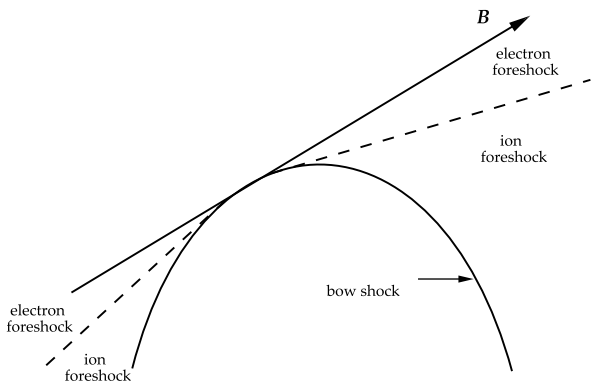

Acceleration Mechanisms, Figure 9
The foreshock may be separated into an electron foreshock, nearer to the field line that is tangent to the bow shock, and an ion foreshock, nearer to the bow shock; the *dashed line* indicates the ill-defined boundary between these two regions, which actually merge continuously into each other

which is also the case in a relativistic treatment [53]. For a strong shock, $B_2 \to 4B_1$, the ratio (28) is $(2+\sqrt{3})/2 - \sqrt{3} = 13.93$. However, the energy ratio decreases rapidly away from the point in the momentum space of the incident particles where this maximum occurs [9].

SDA accounts well for relevant observations of particles associated with shocks in the heliosphere, in particular, the Earth's bow shock and other planetary bow shocks. SDA produces a foreshock region ahead of any (convex) curved shock. The foreshock is the region ahead of the shock and behind the field line that is tangent to the shock, as illustrated in Fig. 9. Fast particles reflected from the shock propagate along the field lines as they are swept towards the shock by the upstream flow. Electrons have higher speeds than ions, so that they propagate further before encountering the shock, producing an electron foreshock region that is much larger than the ion foreshock region. SDA should similarly populate a foreshock region ahead of any curved shock.

Type II solar radio bursts are directly associated with shocks in the solar corona, and extending into the solar wind [79]. Acceleration by SDA is a favored mechanism for the acceleration of electrons that produce type II radio bursts [46,55,60]. SDA is also a potential candidate for preacceleration for DSA. Although both suggestions seem plausible, additional ideas are required to account for the details of the observations.

A modification of DSA, called shock surfing acceleration [100], involves the particles being trapped in the shock front for much longer than the simple theory of SDA implies. In other modifications, the assumption that the shock is a simple discontinuity is relaxed. Various additional possible acceleration mechanisms arise: due to a potential difference along field lines [49]; various instabilities that can lead to large-amplitude waves that can trap and preaccelerate particles ("surfatron" acceleration [66]); and nonstationarity that involves the shock continually breaking up and reforming [89].

Resonant Acceleration

MHD turbulence consists of a mixture of the Alfvén and fast modes. Fermi-type stochastic acceleration involves only the fast mode, and may be interpreted in terms of resonant damping at harmonic number $s = 0$. Resonances at $s \neq 0$ are possible and lead to dissipation of turbulence in both modes, with the energy going into the particles. For highly relativistic particles, resonances at high harmonics lead to a diffusion in momentum space, somewhat analogous to Fermi-type acceleration, but with a diffusion coefficient that depends on a higher power of p [58]. Although this acceleration must occur, for say GCRs due to turbulence in the interstellar medium, there is no strong case for it playing an important role in specific applications.

Acceleration During Magnetic Reconnection

Magnetic energy release through magnetic reconnection is known to be associated with powerful events, notably magnetospheric substorms and solar flares, that can lead to copious acceleration of fast particles. There have been various suggestions that reconnection can lead directly to acceleration of fast particles. For example, for magnetic reconnection in a current sheet, around the surface where **B** changes sign, a test particle that enters the sheet typically emerges with a higher energy [91]. Although such energy gains occur, they do not appear to constitute an important acceleration mechanism. Both theory and observation suggest that the important acceleration associated with reconnection is indirect and spatially separated from the reconnection, and is not a direct consequence of the reconnection. The magnetic energy released in reconnection goes primarily into an Alfvénic flux and into mass motions accelerated by the magnetic tension in the reconnected field lines.

A different application of heating in current sheets has been proposed in connection with pulsar winds [21,65]. Beyond the light cylinder there is a wind containing wound up magnetic field lines in a nearly azimuthal direction, with the sign of **B** reversing periodically with radial distance. The wind accelerates particles at a termination shock, where it encounters the surrounding synchrotron nebula. An outstanding problem is that the energy flux is

Poynting dominated near the light cylinder, but must be kinetic-energy dominated before it reaches the termination shock. The suggestion is that magnetic energy dissipation in the current sheets separating the regions of opposite **B** provide the necessary conversion of magnetic energy into particle energy [21,65]. An alternative suggestion is that the number of current carriers becomes inadequate leading to the development of an accelerating field [71].

Future Directions

Three generic acceleration mechanisms are widely recognized, and there are also numerous other specific mechanisms that are important in specific applications. The pioneering works of Fermi [30,31] anticipated two of the three generic mechanisms: stochastic acceleration and diffusive shock acceleration (DSA), respectively. The third generic mechanism, acceleration by a parallel electric field, $E_\parallel$, was suggested even earlier [94], but remains the least adequately understood of the three.

The most important acceleration mechanism in astrophysics is DSA. It is the accepted mechanism for the acceleration of GCRs and for relativistic electrons in most synchrotron sources. The favored location for the acceleration of GCRs is at the shocks in young supernova remnants [43]. Despite the basic theory being well established there are essential details where our current understanding is incomplete and where further progress is to be expected. The well-established theory is a test-particle model for DSA at a single nonrelativistic shock. Several modifications to the theory are known to be important: the generalizations to relativistic shocks, and to multiple shocks, and the dynamical effects of the accelerated particles on the shock structure. There are several aspects of the required efficient scattering in DSA that are inadequately understood. One is the injection problem: resonant scattering is essential and requires that nonrelativistic ions and electrons have speeds $v > v_A$ and $v > 43v_A$, respectively. Preacceleration to above these thresholds is required before DSA can operate. For ions, the elemental abundances of GCRs suggests an injection mechanism that is insensitive to ionic species, whereas for flare-associated solar energetic particles there are extreme elemental anomalies (notably ^{3}He). For electrons the injection problem is more severe, and less understood. Second is the long-standing question of the ratio of relativistic electrons to ions [34]. In situ data on shocks in the heliosphere provide little evidence for DSA producing relativistic electrons. It is speculated that the shocks in synchrotron sources have much higher Mach numbers than those in the heliosphere, and that such shocks are more efficient in accelerating electrons. A third aspect relates to resonant scattering of the highest energy particles. Unlike lower-energy particles, the resonant waves needed to scatter higher-energy particles cannot be generated by the particles themselves, and other sources of the required waves are speculative. A suggestion that at least alleviates this difficulty is that the lower-energy particles cause waves to grow through a nonresonant instability, which may lead to amplification of the magnetic field by a very large factor [11]: an increase in B increases the maximum energy to which DSA can operate. Fourth, spatial diffusion across field lines is postulated to occur in DSA at perpendicular shocks, but is incompletely understood [3].

Despite its apparent simplicity, acceleration by $E_\parallel$ is the least understood of the three generic mechanisms. An important qualitative point is that $E_\parallel \neq 0$ is inconsistent with MHD, and acceleration by $E_\parallel$ necessarily involves concepts that are relatively unfamiliar in the context of MHD. The best understood application is to acceleration of auroral electrons associated with magnetospheric substorms, but even in this case, how the $E_\parallel$ develops remains uncertain. Existing models for acceleration by $E_\parallel$ in 'gaps' in pulsars and in bulk energization of electrons in solar flares have serious flaws. Some of the present difficulties are alleviated by assuming that the relevant $E_\parallel$ is oscillating: specific models include inertial Alfvén waves and an Alfvén resonator in the magnetosphere, and large-amplitude oscillations in pulsar magnetospheres. Further progress in understanding acceleration by $E_\parallel$ is likely to come from more detailed in situ data on acceleration of auroral electrons.

Compared with the rapid evolution in ideas in other subfields of astrophysics, the ideas involved in understanding particle acceleration have developed only gradually over many decades. Further progress is certain to occur, but is likely to involve gradual acceptance and modification of the existing ideas outlined above, rather than radical new insights.

Appendix A: Quasilinear Equations

The quasilinear equations are written down here using a semiclassical formalism. The waves are regarded as a collection of wave quanta, with energy $\hbar\omega$ and momentum $\hbar\mathbf{k}$. Waves in an arbitrary mode, labeled M, have dispersion relation $\omega = \omega_M(\mathbf{k})$, polarization vector $\mathbf{e}_M(\mathbf{k})$ and ratio of electric to total energy, $R_M(\mathbf{k})$. Their distribution is described by their occupation number $N_M(\mathbf{k})$. A resonant interaction occurs when the gyroresonance condition is satisfied:

$$\omega - s\Omega - k_\parallel v_\parallel = 0\,, \tag{A1}$$

where s is an integer, $\Omega = |q|B/m\gamma$ is the relativistic gyrofrequency, and $k_\parallel$, $v_\parallel$ are the components of $\mathbf{k}$, $\mathbf{v}$ parallel to $\mathbf{B}$. Resonances at $s > 0$ are said to be via the normal Doppler effect, and those at $s < 0$ are said to be via the anomalous Doppler effect. Resonances at $s \leq 0$ are possible only for waves with refractive index greater than unity. The effect of a wave-particle interaction is described by the probability of spontaneous emission, $w_M(\mathbf{p}, \mathbf{k}, s)$, which is the probability per unit time that the particle emit a wave quantum in the wave mode M in the elemental range $d^3\mathbf{k}/(2\pi)^3$. For a particle of charge q and mass m, the probability of spontaneous emission is given by

$$\begin{aligned} w_M(\mathbf{p}, \mathbf{k}, s) &= \frac{2\pi q^2 R_M(\mathbf{k})}{\varepsilon_0 \hbar \omega_M(\mathbf{k})} \left|\mathbf{e}^*_M(\mathbf{k}) \cdot \mathbf{V}(\mathbf{k}, \mathbf{p}; s)\right|^2 \\ &\quad \cdot \delta\big(\omega_M(\mathbf{k}) - s\Omega - k_\parallel v_\parallel\big)\,, \\ \mathbf{V}(\mathbf{k}, \mathbf{p}; s) &= \Big(v_\perp \frac{s}{k_\perp R} J_s(k_\perp R)\,, \\ &\quad - i\eta v_\perp J'_s(k_\perp R)\,, v_\parallel J_s(k_\perp R)\Big)\,, \\ \Omega &= \frac{\Omega_0}{\gamma}\,, \quad \Omega_0 = \frac{|q|B}{m}\,, \quad R = \frac{v_\perp}{\Omega} = \frac{p_\perp}{|q|B}\,. \end{aligned} \tag{A2}$$

An advantage of the semiclassical formalism is that the Einstein coefficients imply that the probability of stimulated emission and true absorption are given by multiplying this probability by $N_M(\mathbf{k})$. This allows the effect on the occupation numbers of the waves and the particles to be written down using a simple bookkeeping argument. For the waves one finds

$$\frac{\mathrm{d}N_M(\mathbf{k})}{\mathrm{d}t} = \left(\frac{\mathrm{d}N_M(\mathbf{k})}{\mathrm{d}t}\right)_{\text{spont}} - \gamma_M(\mathbf{k}) N_M(\mathbf{k})\,, \tag{A3}$$

where the two terms on the right hand side describe spontaneous emission and absorption, respectively, with the absorption coefficient given by

$$\begin{aligned} \gamma_M(\mathbf{k}) &= -\sum_{s=-\infty}^{\infty} \int \mathrm{d}^3\mathbf{p}\, w_M(\mathbf{p}, \mathbf{k}, s) \hat{D}_s f(\mathbf{p})\,, \\ \hat{D}_s &= \hbar \left(\frac{s\Omega}{v_\perp}\frac{\partial}{\partial p_\perp} + k_\parallel \frac{\partial}{\partial p_\parallel}\right) \\ &= \frac{\hbar\omega}{v}\left(\frac{\partial}{\partial p} + \frac{\cos\alpha - n_M \beta \cos\theta}{p \sin\alpha}\frac{\partial}{\partial \alpha}\right)\,. \end{aligned} \tag{A4}$$

The two forms correspond to, respectively, cylindrical and polar coordinates in momentum space.

The evolution of the distribution of particles due to the resonant interaction includes a term, neglected here, that describes the effect of spontaneous emission, and a quasi-linear diffusion equation that describes the effect of the induced processes. In cylindrical and spherical polar coordinates, this equation is

$$\begin{aligned} \frac{\mathrm{d}f(\mathbf{p})}{\mathrm{d}t} &= \frac{1}{p_\perp}\frac{\partial}{\partial p_\perp} \\ &\quad \cdot \left\{p_\perp \left[D_{\perp\perp}(\mathbf{p})\frac{\partial}{\partial p_\perp} + D_{\perp\parallel}(\mathbf{p})\frac{\partial}{\partial p_\parallel}\right] f(\mathbf{p})\right\} \\ &\quad + \frac{\partial}{\partial p_\parallel}\left\{\left[D_{\parallel\perp}(\mathbf{p})\frac{\partial}{\partial p_\perp} + D_{\parallel\parallel}(\mathbf{p})\frac{\partial}{\partial p_\parallel}\right] f(\mathbf{p})\right\} \\ &= \frac{1}{\sin\alpha}\frac{\partial}{\partial\alpha} \\ &\quad \cdot \left\{\sin\alpha \left[D_{\alpha\alpha}(\mathbf{p})\frac{\partial}{\partial\alpha} + D_{\alpha p}(\mathbf{p})\frac{\partial}{\partial p}\right] f(\mathbf{p})\right\} \\ &\quad + \frac{1}{p^2}\frac{\partial}{\partial p}\left\{p^2\left[D_{p\alpha}(\mathbf{p})\frac{\partial}{\partial\alpha} + D_{pp}(\mathbf{p})\frac{\partial}{\partial p}\right] f(\mathbf{p})\right\}\,, \end{aligned} \tag{A5}$$

respectively, with the diffusion coefficients in either set of coordinates given by

$$\begin{aligned} D_{QQ'}(\mathbf{p}) &= \sum_{s=-\infty}^{\infty} \int \frac{\mathrm{d}^3\mathbf{k}}{(2\pi)^3} w_M(\mathbf{p}, \mathbf{k}, s)\, \Delta Q\, \Delta Q'\, N_M(\mathbf{k}), \\ \Delta Q &= \hat{D}_s Q\,, \\ \Delta p_\perp &= \frac{s\Omega}{v_\perp}\,, \quad \Delta p_\parallel = \hbar k_\parallel\,, \\ \Delta\alpha &= \frac{\hbar(\omega\cos\alpha - k_\parallel v)}{p v \sin\alpha}\,, \quad \Delta p = \frac{\hbar\omega}{v}\,. \end{aligned} \tag{A6}$$

Bibliography

Primary Literature

1. Achterberg A (1981) On the nature of small amplitude Fermi acceleration. Astron Astrophys 97:259–264
2. Achterberg A (1990) Particle acceleration by an ensemble of shocks. Astron Astrophys 231:251–258
3. Achterberg A, Ball L (1994) Particle acceleration at superluminal quasi-perpendicular shocks. Appl SN1978K SN1987A. Astron Astrophys 285:687–704
4. Achterberg A, Gallant YA, Kirk JG, Guthmann AW (2001) Particle acceleration by ultrarelativistic shocks: theory and simulations. Mon Not Roy Astron Soc 328:393–408
5. Armstrong JW, Rickett BJ, Spangler SR (1995) Electron density power spectra in the local interstellar medium. J Atrophys 443:209–221
6. Axford WI (1994) The origins of high-energy cosmic rays. J Astrophys Suppl 90:937–944
7. Axford WI, Leer E, Skadron G (1977) Acceleration of cosmic rays by shock waves. 15th Internat Cosmic Ray Conf Papers (Plovdiv) 11:132–137
8. Baade W, Zwicky F (1934) Cosmic rays from super-novae. Proc Natl Acad Sci 20:259

9. Ball L, Melrose DB (2001) Shock drift acceleration of electrons. Publ Astron Soc Aust 18:361–373
10. Bell AR (1978) The acceleration of cosmic rays in shock fronts 1&2. Mon Not Roy Astron Soc 182:147–156; 182:443–455
11. Bell AR (2004) Turbulent amplification of magnetic field and diffusive shock acceleration of cosmic rays. Mon Not Roy Astron Soc 353:550–558
12. Bell AR, Lucek SG (2001) Cosmic ray acceleration to very high energy trhough the non-linear amplification by cosmic rays of the seed magnetic field. Mon Not Roy Astron Soc 321:433–438
13. Bellan PM (1996) New model for ULF Pc5 pulsations: Alfvén cones. Geophys Res Lett 23:1717–1720
14. Benson RF, Calvert W, Klumpar DM (1980) Simultaneous wave and particle observations in the auroral kilometric radiation source region. Geophys Res Lett 7:959–962
15. Berger JM, Newcomb WA, Dawson JM, Frieman EA, Kulsrud RM, Lenard A (1958) Heating of a confined plasma by oscillating electromagnetic fields. Phys Fluids 1:301–307
16. Bicknell GV, Melrose DB (1982) In situ acceleration in extragalactic radio jets. J Astrophys 262:511–528
17. Blandford RD, Ostriker FR (1978) Particle acceleration by astrophysical shocks. J Astrophys 221:L29–L32
18. Block LP (1978) A double layer review. Astrophys Space Sci 55:59–83
19. Boström R, Gustafsson G, Holback B, Holmgren G, Koskinen H (1988) Characteristics of solitary waves and weak double layers in the magnetospheric plasma. Phys Rev Lett 61:82–85
20. Chaston CC, Bonnell JW, Carlson CW, Berthomier M, Peticolas LM, Roth I, McFadden JP, Ergun RE, Strangeway RJ (2002) Electron acceleration in the ionospheric Alfven resonator. J Geophys Res 107:SMP 41–1–16
21. Coroniti FV (1990) Magnetically striped relativistic magnetohydrodynamic winds. The Crab Nebula revisited. J Astrophys 349:538–545
22. Cotton WD, Wittels JJ, Shapiro II, Marcaide J, Owen FN, Spangler SR, Rius A, Angulo C, Clark TA, Knight CA (1980) The very flat radio spectrum of 0735+178: A cosmic conspiracy? J Astrophys 238:L123–L128
23. Davis L Jr (1956) Modified Fermi mechanism for the acceleration of cosmic rays. Phys Rev 101:351–358
24. Dragt AJ (1961) Effect of hydromagnetic waves of the lifetime of Van Allen radiation protons. J Geophys Res 66:1641–1649
25. Drury LO'C (1983) An introduction to the theory of diffusive shock acceleration of energetic particles in tenuous plasma. Rep Prog Phys 46:973–1027
26. Duijveman A, Hoyng P, Ionson JA (1981) Fast plasma heating by anomalous and inertial resistivity effects in the solar atmosphere. J Astrophys 245:721–735
27. Dungey JW (1963) Loss of Van Allen electrons due to whistlers. Planet Space Sci 11:591–595
28. Ergun RE, Andersson L, Main D, Su Y-J, Newman DL, Goldman MV, Carlson CW, Hull AJ, McFadden JP, Mozer FS (2004) Auroral particle acceleration by strong double layers: the upward current region. J Geophys Res 109:A12220
29. Ergun RE, Su Y-J, Andersson L, Bagenal F, Delemere PA, Lysak RL, Strangeway RJ (2006) S bursts and the Jupiter ionospheric Alfvn resonator. J Geophys Res 111:A06212
30. Fermi E (1949) On the origin of cosmic radiation. Phys Rev 75:1169–1174
31. Fermi E (1954) Galactic magnetic field and the origin of cosmic radiation. J Astrophys 119:1–6
32. Fisk LA (1978) ^{3}He-rich flares: a possible explanation. J Astrophys 224:1048–1055
33. Forbes TG, Malherbe JM (1986) A shock condensation mechanism for loop prominences. J Astrophys 302:L67–L70
34. Ginzburg VL, Syrovatskii SI (1964) The origin of cosmic rays. Pergamon Press, Oxford
35. Goertz CK (1980) Io's interaction with the plasma torus. J Geophys Res 85:2949–2956
36. Goertz CK, Boswell RW (1979) Magnetosphere-ionosphere coupling. J Geophys Res 84:7239–7246
37. Goldreich P, Lynden-Bell D (1969) Io, a Jovian unipolar inductor. J Astrophys 56:59–78
38. Gurevich AV (1960) On the amount of accelerated particles in an ionized gas under various acceleration mechanisms. Sov Phys JETP 11:1150–1157
39. Gurnett DA (1974) The earth as a radio source: terrestrial kilometric radiation. J Geophys Res 79:4227–4238
40. Hall DE, Sturrock PA (1967) Diffusion, scattering, and acceleration of particles by stochastic electromagnetic fields. Plasma Phys 10:2620–2628
41. Heavens AF, Drury LO'C (1989) Relativistic shocks and particle acceleration. Mon Not Roy Astron Soc 235:997–1009
42. Helliwell RA (1967) A theory of discrete VLF emissions from the magnetosphere. J Geophys Res 72:4773–4790
43. Hillas AM (2006) Cosmic rays: recent progress and some current questions. In: Klöckner H-R, Jarvis M, Rawlings S (eds) Cosmology, galaxy formation and astroparticle physics on the pathway to the SKA. astro-ph/0607109 (to be published)
44. Hillas AM (2006) The cosmic-ray knee and ensuing spectrum seen as a consequence of Bell's SNR shock acceleration process. J Phys Conf Ser 47:168–177
45. Holman GD (1985) Acceleration of runaway electrons and joule heating in solar flares. J Astrophys 293:584–594
46. Holman GD, Pesses ME (1983) Type II radio emission and the shock drift acceleration of electrons. J Astrophys 267:837–843
47. Ionson JA (1982) Resonant electrodynamic heating of stellar coronal loops: an LRC circuit analog. J Astrophys 254:318–334
48. Ionson JA (1985) Coronal heating by resonant (AC) and nonresonant (DC) mechanisms. Astron Astrophys 146:199–203
49. Jones FC, Ellison DC (1987) Noncoplanar magnetic fields, shock potentials, and ion deflection. J Geophys Res 92: 11205–11207
50. Kaplan SA (1956) The theory of the acceleration of charged particles by isotropic gas magnetic turbulent fields. Sov Phys JETP 2:203–210
51. Kellermann KI, Pauliny-Toth IIK (1969) The spectra of opaque radio sources. J Astrophys 155:L71–L78
52. Kennel CF, Petschek HE (1966) Limit on stably trapped particle fluxes. J Geophys Res 71:1–28
53. Kirk JG (1994) Particle Acceleration. In: Benz AO, Courvoisier TJ-L (eds) Saas-Fee Advanced Course 24, Plasma Astrophysics. Springer, New York, pp 225–314
54. Knight S (1973) Parallel electric fields. Planet Space Sci 21:741–750
55. Knock SA, Cairns IH, Robinson PA, Kuncic Z (2001) Theory of type II radio emission from the foreshock of an interplanetary shock. J Geophys Res 106:25041–25052
56. Krimsky GF (1977) A regular mechanism for the acceleration

of charged particles on the front of a shock wave. Sov Phys Doklady 234:1306–1308
57. Kulsrud R, Ferrari A (1971) The relativistic quasilinear theory of particle acceleration by hydromagnetic turbulence. Astrophys Space Sci 12:302–318
58. Lacombe C (1977) Acceleration of particles and plasma heating by turbulent Alfvén waves in a radiogalaxy. Astron Astrophys 54:1–16
59. Lee MA, Ryan JM (1986) Time-dependent shock acceleration of energetic solar particles. J Astrophys 303:829–842
60. Leroy MM, Mangeney A (1984) A theory of energization of solar wind electrons by the Earth's bow shock. Ann Geophys 2:449–455
61. Levinson A, Melrose D, Judge A, Luo Q (2005) Large-amplitude, pair-creating oscilllations in pulsar magnetospheres. J Astrophys 631:456–465
62. Lin RP (2006) Particle acceleration by the Sun: electrons, hard X-rays/gamma-rays. Space Sci Rev 124:233–248
63. Lyons LR (1980) Generation of large-scale regions of auroral currents, electric potentials, and precipitation by the divergence of convection electric fields. J Geophys Res 85:17–24
64. Lysak RL, Song Y (2005) Nonlocal interactions between electrons and Alfvn waves on auroral field lines. J Geophys Res 110:A10S06
65. Lyubarsky YE, Kirk JG (2001) Reconnection in a striped pulsar wind. J Astrophys 547:437–448
66. McClements KG, Dieckmann ME, Ynnerman A, Chapman SC, Dendy RO (2001) Surfatron and stochastic acceleration of electrons at supernova remant shocks. Pys Rev Lett 87:255002
67. Malkov MA, Drury LO'C (2001) Nonlinear theory of diffusive acceleration of particles by shock waves. Rep Prog Phys 64:429–481
68. Marscher AP (1977) Structure of radio sources with remarkably flat spectra: PKS 0735+178. Astron J 82:781–784
69. McLean DJ, Sheridan KV, Stewart RT, Wild JP (1971) Regular pulses from the Sun and a possible clue to the origin of solar cosmic rays. Nature 234:140–142
70. Meisenheimer K, Röser H-J, Schlötelburg M (1996) Astron Astrophys 307:61–79
71. Melatos A, Melrose DB (1996) Energy transport in a rotation-modulated pulsar wind. Mon Not Roy Astron Soc 279:1168–1190
72. Melrose DB (1968) The emission and absorption of waves by charged particles in magnetized plasmas. Astrophys Space Sci 2:171–235
73. Melrose DB (2005) Nonresonant Alfvén waves driven by cosmic rays. In: Li G, Zank GP, Russell CT (eds) The physics of collisionless shocks. AIP Conference Proceedings #781, American Institute of Physics, pp 135–140
74. Melrose DB, Wentzel DG (1970) The interaction of cosmic-ray electrons with cosmic-ray protons. J Astrophys 161:457–476
75. Melrose D, Couch A (1997) Effect of synchrotron losses on multiple diffusive shock acceleration. Autralian J Phys 14:251–257
76. Micono M, Zurlo N, Massaglia S, Ferrari A, Melrose DB (1999) Diffusive shock acceleration in extragalactic jets. Astron Astrophys 349:323–333
77. Miller JA, Cargill PJ, Emslie AG, Holman GD, Dennis BR, LaRosa TN, Winglee RM, Benka SG, Tsuneta S (1997) Critical issues for understanding particle acceleration in impulsive solar flares. J Geophys Res 102:4631–14660
78. Mozer FS, Cattell CA, Hudson MK, Lysak RL, Temerin M, Torbert RB (1980) Satellite measurements and theories of low altitude auroral particle acceleration. Space Sci Rev 27: 155–213
79. Nelson GJ, Melrose DB (1985) Type II bursts. In: McLean DJ, Labrum NR (eds) Solar Radiophysics, chapter 13. Cambridge University Press, pp 333–359
80. Newman DL, Goldman MV, Ergun RE, Mangeney A (2001) Formation of double layers and electron holes in a current-driven space plasma. Phys Rev Lett 87:255001
81. Parizot E, Marcowith A, Ballet J, Gallant YA (2006) Observational constraints on energetic particle diffusion in young supernovae remnants: amplified magnetic field and maximum energy. Astron Astrophys 453:387–395
82. Parker EN (1957) Acceleration of cosmic rays in solar flares. Phys Rev 107:830–836
83. Parker EN, Tidman DA (1958) Suprathermal particles. Phys Rev 111:1206–1211
84. Pesses ME (1981) On the conservation of the first adiabatic invariant in perpendicular shocks. J Geophys Res 86: 150–152
85. Raadu MA (1989) The physics of double layers and their role in astrophysics. Phys Rep 178:25–97
86. Reville B, Kirk JG, Duffy P (2006) A current-driven instability in parallel, relativistic shocks. Plasma Phys Cont Fus 48:1741–1747
87. Schlickeiser R (1984) An explanation of abrupt cutoffs in the optical-infrared spectra of non-thermal sources. A new pile-up mechanism for relativistic electron spectra. Astron Astrophys 136:227–236
88. Schlüter A (1957) Der Gyro-Relaxations-Effekt. Z Nat 12a:822–825
89. Schmitz H, Chapman SC, Dendy RO (2002) Electron preacceleration mechanisms in the foot region of high Alfvénic Mach number shocks. J Astrophys 579:327–336
90. Shen CS (1965) Transit acceleration of charged particles in an inhomogeneous electromagnetic field. J Astrophys 141: 1091–1104
91. Speiser TW (1965) Particle trajectories in model current sheets, 1. Analytic solutions. J Geophys Res 70:1717–1788
92. Stasiewicz K, et al. (2000) Small scale Alfvénic structures in the aurora. Space Sci Rev 92:423–533
93. Sturrock PA (1966) Stochastic acceleration. Phys Rev 141:186–191
94. Swann WFG (1933) A mechanism of acquirement of cosmic-ray energies by electrons. Phys Rev 43:217–220
95. Thompson WB (1955) On the acceleration of cosmic-ray particles by magneto-hydrodynamic waves. Proc Roy Soc Lond 233:402–406
96. Toptygin IN (1980) Acceleration of particles by shocks in a cosmic plasma. Space Sci Rev 26:157–213
97. Tsytovich VN (1966) Statistical acceleration of particles in a turbulent plasma. Sov Phys Uspeki 9:370–404
98. Tverskoĭ BA (1967) Contribution to the theory of Fermi statistical acceleration. Sov Phys JETP 25:317–325
99. Tverskoĭ BA (1968) Theory of turbulent acceleration of charged particles in a plasma. Sov Phys JETP 26:821–828
100. Ucer D, Shapiro VD (2001) Unlimited relativistic shock surfing acceleration. Phys Rev Lett 87:075001

101. van der Swaluw E, Achterberg A (2004) Non-thermal X-ray emission from young supernova remnants. Astron Astrophys 421:1021–1030
102. Vedin J, Rönnmark K (2004) A linear auroral current-voltage relation in fluid theory. Ann Geophys 22:1719–1728
103. Vogt J (2002) Alfvén wave coupling in the auroral current circuit. Surveys Geophys 23:335–377
104. Völk HJ (2006) Shell-type supernova remnants. astro-ph/0603502
105. Wentzel DG (1961) Hydromagnetic waves and the trapped radiation Part 1. Breakdown of the adiabatic invariance. J Geophys Res 66:359–369
106. Wentzel DG (1964) Motion across magnetic discontinuities and Fermi acceleration of charged particles. J Astrophys 140:1013–1024
107. Wiersma J, Achterberg A (2004) Magnetic field generation in relativistic shocks. An early end of the exponential Weibel instability in electron-proton plasmas. Astron Astrophys 428:365–371
108. Wild JP, Smerd SF, Weiss AA (1963) Solar bursts. Ann Rev Astron Astrophys 1:291–366
109. Wygant JR, Keiling A, Cattell CA, Lysak RL, Temerin M, Mozer FS, Kletzing CA, Scudder JD, Streltsov V, Lotko W, Russell CT (2002) Evidence for kinetic Alfén waves and parallel electron energization at 4–6 RE altitudes in the plasma sheet boundary layer. J Geophys Res 107:A900113

Books and Reviews

Alfvén H, Fälthammar C-G (1963) Cosmical Electrodynamics. Oxford University Press, Oxford
Aschwanden MJ (2004) Physics of the solar corona. Springer, Berlin
Benz AO (1993) Plasma astrophysics: kinetic processes in solar and stellar coronae. Kluwer Academic Publishers, Dordrecht
Berezinskii VS, Bulanov SV, Dogiel VA, Ginzburg VL, Ptuskin VS (1990) Astrophysics of cosmic rays. North Holland, Amsterdam
Dorman LI (2006) Cosmic ray interactions, propagation, and acceleration in space plasmas. Springer, New York
Drury LO'C (1983) An introduction to the theory of diffusive shock acceleration of energetic particles in tenuous plasmas. Rep Prog Phys 46:973–1027
McLean DJ, Labrum NR (eds) (1986) Solar Radiophysics. Cambridge University Press, Cambridge
Melrose DB (1980) Plasma Astrophysics, vol I & II. Gordon, New York
Melrose DB (1986) Instabilities in Space and Laboratory Plasmas. Cambridge University Press
Malkov MA, Drury LO'C (2001) Nonlinear theory of diffusive acceleration of particles by shock waves. Rep Prog Phys 64:429–481
Michel FC (1991) Physics of neutron star magnetospheres. The University of Chicago Press, Chicago
Priest ER, Forbes T (2000) Magnetic reconnection–MHD theory and applications. Cambridge University Press, Cambridge
Schlickeiser R (2002) Cosmic ray astrophysics. Springer, Berlin
Stone RG, Tsurutani BT (eds) (1985) Collisionless shocks in the heliosphere: a tutorial review. Geophysicial Monograph 34. American Geophysical Union, Washington DC
Sturrock PA (1980) Solar flares. Colorado Associated University Press, Boulder
Svestka Z (1976) Solar Flares. D Reidel, Dordrecht
Vogt J (2002) Alfvén wave coupling in the auroral current circuit. Surv Geophys 23:335–377

Adaptive Visual Servo Control

Guoqiang Hu[1], Nicholas Gans[2], Warren E. Dixon[2]
[1] Department of Mechanical and Nuclear Engineering, Kansas State University, Manhattan, USA
[2] Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Visual Servo Control Methods
Compensating for Projection Normalization
Visual Servo Control via an Uncalibrated Camera
Future Directions
Bibliography

Glossary

Camera-in-hand configuration The camera-in-hand configuration refers to the case when the camera is attached to a moving robotic system (e. g., held by the robot end-effector).

Camera-to-hand configuration The camera-to-hand configuration refers to the case when the camera is stationary and observing moving targets (e. g., a fixed camera observing a moving robot end-effector).

Euclidean reconstruction Euclidean reconstruction is the act of reconstructing the Euclidean coordinates of feature points based on the two-dimensional image information obtained from a visual sensor.

Extrinsic calibration parameters The extrinsic calibration parameters are defined as the relative position and orientation of the camera reference frame to the world reference frame (e. g., the frame affixed to the robot base). The parameters are represented by a rotation matrix and a translation vector.

Feature point Different computer vision algorithms have been developed to search images for distinguishing features in an image (e. g., lines, points, corners, textures). Given a three dimensional Euclidean point, its projection to the two dimensional image plane is called a feature point. Feature points are selected based on

contrast with surrounding pixels and the ability for the point to be tracked from frame to frame. Sharp corners (e. g., the windows of a building), or center of a homogenous region in the image (e. g., the center of car headlights) are typical examples of feature points.

Homography The geometric concept of homography is a one-to-one and on-to transformation or mapping between two sets of points. In computer vision, homography refers to the mapping between points in two Euclidean-planes (Euclidean homography), or to the mapping between points in two images (projective homography).

Image Jacobian The image-Jacobian is also called the interaction matrix, feature sensitivity matrix, or feature Jacobian. It is a transformation matrix that relates the change of the pixel coordinates of the feature points on the image to the velocity of the camera.

Intrinsic calibration parameters The intrinsic calibration parameters map the coordinates from the normalized Euclidean-space to the image-space. For the pinhole camera, this invertible mapping includes the image center, camera scale factors, and camera magnification factor.

Pose The position and orientation of an object is referred to as the pose. The pose has a translation component that is an element of $\mathbb{R}^3$ (i. e., Euclidean-space), and the rotation component is an element of the special orthogonal group $SO(3)$, though local mappings of rotation that exist in $\mathbb{R}^3$ (Euler angles, angle/axis) or $\mathbb{R}^4$ (unit quaternions). Pose of an object has specific meaning when describing the relative position and orientation of one object to another or the position and orientation of an object at different time instances. Pose is typically used to describe the position and orientation of one reference frame with respect to another frame.

Projection normalization The Euclidean coordinates of a point are projected onto a two-dimensional image via the projection normalization. For a given camera, a linear projective matrix can be used to map points in the Euclidean-space to the corresponding pixel coordinates on the image-plane. This relationship is not bijective because there exist infinite number of Euclidean points that have the same normalized coordinates, and the depth information is lost during the normalization.

Unit quaternion The quaternion is a four dimensional vector which is composed of real numbers (i. e., $q \in \mathbb{R}^4$). The unit quaternion is a quaternion subject to a nonlinear constraint that the norm is equal to 1. The unit quaternion can be used as a globally nonsingular representation of a rotation matrix.

Visual servo control system Control systems that use information acquired from an imaging source (e. g., a digital camera) in the feedback control loop are defined as visual servo control systems. In addition to providing feedback relating the local position and orientation (i. e., pose) of the camera with respect to some target, an image sensor can also be used to relate local sensor information to an inertial reference frame for global control tasks. Visual servoing requires multidisciplinary expertise to integrate a vision system with the controller for tasks including: selecting the proper imaging hardware; extracting and processing images at rates amenable to closed-loop control; image analysis and feature point extraction/tracking; recovering/estimating necessary state information from an image; and operating within the limits of the sensor range (i. e., field-of-view).

Lyapunov function A Lyapunov function is a continuously differentiable positive definite function with a negative definite or semi-definite time-derivative.

Definition of the Subject

Control systems that use information acquired from an imaging source in the feedback loop are defined as visual servo control systems. Visual servo control has developed into a large subset of robotics literature because of the enabling capabilities it can provide for autonomy. The use of an image sensor for feedback is motivated by autonomous task execution in an unstructured or unknown environments. In addition to providing feedback relating the local position and orientation (i. e., pose) of the camera with respect to some target, an image sensor can also be used to relate local sensor information to an inertial reference frame for global control tasks.

However, the use of image-based feedback adds complexity and new challenges for the control system design. Beyond expertise in design of mechanical control systems, visual servoing requires multidisciplinary expertise to integrate a vision system with the controller for tasks including: selecting the proper imaging hardware; extracting and processing images at rates amenable to closed-loop control; image analysis and feature point extraction/tracking; recovering/estimating necessary state information from an image; and operating within the limits of the sensor range (i. e., field-of-view). While each of the aforementioned tasks are active topics of research interest, the scope of this chapter is focused on issues associated with using reconstructed and estimated state information from an image to develop a stable closed-loop error system. That is, the topics in this chapter are based on the assumption that images

can be acquired, analyzed, and the resulting data can be provided to the controller without restricting the control rates.

Introduction

The different visual servo control methods can be divided into three main categories including: image-based, position-based, and approaches that use of blend of image and position-based approaches. Image-based visual servo control (e. g., [1,2,3,4,5]) consists of a feedback signal that is composed of pure image-space information (i. e., the control objective is defined in terms of an image pixel error). This approach is considered to be more robust to camera calibration and robot kinematic errors and is more likely to keep the relevant image features in the field-of-view than position-based methods because the feedback is directly obtained from the image without the need to transfer the image-space measurement to another space. A drawback of image-based visual servo control is that since the controller is implemented in the robot joint space, an image-Jacobian is required to relate the derivative of the image-space measurements to the camera's linear and angular velocities. However, the image-Jacobian typically contains singularities and local minima [6], and the controller stability analysis is difficult to obtain in the presence of calibration uncertainty [7]. Another drawback of image-based methods is that since the controller is based on image-feedback, the robot could be commanded to some configuration that is not physically possible. This issue is described as Chaumette's conundrum (see [6,8]).

Position-based visual servo control (e. g., [1,5,9,10,11]) uses reconstructed Euclidean information in the feedback loop. For this approach, the image-Jacobian singularity and local minima problems are avoided, and physically realizable trajectories are generated. However, the approach is susceptible to inaccuracies in the task-space reconstruction if the transformation is corrupted (e. g., uncertain camera calibration). Also, since the controller does not directly use the image features in the feedback, the commanded robot trajectory may cause the feature points to leave the field-of-view. A review of these two approaches is provided in [5,12,13].

The third class of visual servo controllers use some image-space information combined with some reconstructed information as a means to combine the advantages of these two approaches while avoiding their disadvantages (e. g., [8,14,15,16,17,18,19,20,21,22,23]). One particular approach was coined 2.5D visual servo control in [15,16,17,18] because this class of controllers exploits some two dimensional image feedback and reconstructed three-dimensional feedback. This class of controllers is also called homography-based visual servo control in [19,20,21,22] because of the underlying reliance of the construction and decomposition of a homography.

A key issue that is pervasive to all visual servo control methods is that the image-space is a two-dimensional projection of the three-dimensional Euclidean-space. To compensate for the lack of depth information from the two-dimensional image data, some control methods exploit additional sensors (e. g., laser and sound ranging technologies) along with sensor fusion methods or the use of additional cameras in a stereo configuration that triangulate on corresponding images. However, the practical drawbacks of incorporating additional sensors include: increased cost, increased complexity, decreased reliability, and increased processing burden. Geometric knowledge of the target or scene can be used to estimate depth, but this information is not available for many unstructured tasks.

Motivated by these practical constraints, direct approximation and estimation techniques (cf. [4,5]) and dynamic adaptive compensation (cf. [19,20,21,22,24,25,26,27,28]) have been developed. For example, an adaptive kinematic controller has been developed in [24] to achieve a uniformly ultimately bounded regulation control objective provided conditions on the translation velocity and bounds on the uncertain depth parameters are satisfied. In [24], a three dimensional depth estimation procedure has been developed that exploits a prediction error provided a positive definite condition on the interaction matrix is satisfied. In [20,26,27], homography-based visual servo controllers have been developed to asymptotically regulate a manipulator end-effector and a mobile robot, respectively, by developing an adaptive update law that actively compensates for an unknown depth parameter. In [19,21], adaptive homography-based visual servo controllers have been developed to achieve asymptotic tracking control of a manipulator end-effector and a mobile robot, respectively. In [22,28], adaptive homography-based visual servo controllers via a quaternion formulation have been developed to achieve asymptotic regulation and tracking control of a manipulator end-effector, respectively.

In addition to normalizing the three-dimensional coordinates, the projection from the Euclidean-space to the image-space also involves a transformation matrix that may contain uncertainties. Specifically, a camera model (e. g., the pinhole model) is often required to relate pixel coordinates from an image to the (normalized) Euclidean coordinates. The camera model is typically assumed to be exactly known (i. e., the intrinsic calibration parameters are assumed to be known); however, de-

spite the availability of several popular calibration methods (cf. [29,30,31,32,33,34,35]), camera calibration can be time-consuming, requires some level of expertise, and has inherent inaccuracies. If the calibration parameters are not exactly known, performance degradation and potential unpredictable response from the visual servo controller may occur.

Motivated by the desire to incorporate robustness to camera calibration, different control approaches that do not depend on exact camera calibration have been proposed (cf. [18,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51]). Efforts such as [36,37,38,39,40] have investigated the development of methods to estimate the image and robot manipulator Jacobians. These methods are composed of some form of recursive Jacobian estimation law and a control law. Specifically, a visual servo controller is developed in [36] based on a weighted recursive least-squares update law to estimate the image Jacobian. In [37], a Broyden–Jacobian estimator is applied and a nonlinear least-square optimization method is used for the visual servo control development. In [38], the authors used a nullspace-biased Newton-step visual servo strategy with a Broyden Jacobian estimation for online singularity detection and avoidance in an uncalibrated visual servo control problem. In [39,40] a recursive least-squares algorithm is implemented for Jacobian estimation, and a dynamic Gauss–Newton method is used to minimize the squared error in the image plane.

In [41,42,43,44,45,46,47], robust and adaptive visual servo controllers have been developed. The development of traditional adaptive control methods to compensate for the uncertainty in the transformation matrix is inhibited because of the time-varying uncertainty injected in the transformation from the normalization of the Euclidean coordinates. As a result, initial adaptive control results were limited to scenarios where the optic axis of the camera was assumed to be perpendicular with the plane formed by the feature points (i. e., the time-varying uncertainty is reduced to a constant uncertainty). Other adaptive methods were developed that could compensate for the uncertain calibration parameters assuming an additional sensor (e. g., ultrasonic sensors, laser-based sensors, additional cameras) could be used to measure the depth information.

More recent approaches exploit geometric relationships between multiple spatiotemporal views of an object to transform the time-varying uncertainty into known time-varying terms multiplied by an unknown constant [18,48,49,50,51,52]. In [48], an on-line calibration algorithm was developed for position-based visual servoing. In [49], an adaptive controller was developed for image-based dynamic control of a robot manipulator. One problem with methods based on the image-Jacobian is that the estimated image-Jacobian may contain singularities. The development in [49] exploits an additional potential force function to drive the estimated parameters away from the values that result in a singular Jacobian matrix. In [52], an adaptive homography-based controller was proposed to address problems of uncertainty in the intrinsic camera calibration parameters and lack of depth measurements. In Subsect. "Adaptive Homography–Based Visual Servo Control Approach", a combined high-gain control approach and an adaptive control strategy are proposed to regulate the robot end-effector to a desired pose asymptotically with considering this time-varying scaling factor.

Robust control approaches based on static best-guess estimation of the calibration matrix have been developed to solve the uncalibrated visual servo regulation problem (cf. [18,44,50,51]). Specifically, under a set of assumptions on the rotation and calibration matrix, a kinematic controller was developed in [44] that utilizes a constant, best-guess estimate of the calibration parameters to achieve local set-point regulation for the six degree-of-freedom visual servo control problem. Homography-based visual servoing methods using best-guess estimation are used in [18,50,51] to achieve asymptotic or exponential regulation with respect to both camera and hand-eye calibration errors for the six degree-of-freedom problem.

This chapter is organized to provide the basic underlying principles of both historic and recent visual servo control methods with a progression to open research issues due to the loss of information through the image projection and the disturbances due to uncertainty in the camera calibration. Section "Visual Servo Control Methods" provides an overview of the basic methodologies. This section includes topics such as: geometric relationships and Euclidean reconstruction using the homography in Subsect. "Geometric Relationships"; and image-based visual servo control, position-based visual servo control, and hybrid visual servo control approaches in Subsects. "Image-based Visual Servo Control"–"Homography–Based Visual Servo Control". In Sect. "Compensating for Projection Normalization", an adaptive technique that actively compensates for unknown depth parameters in visual servo control is presented. Specifically, this adaptive compensation approach has been combined with the homography-based visual servo control scheme and image-based visual servo control scheme to generate asymptotic results in Subsects. "Adaptive Depth Compensation in Homography–Based Visual Servo Control" and "Adaptive Depth Compensation in Image–Based Visual Servo Control", respectively. In Sect. "Visual Servo Con-

trol via an Uncalibrated Camera", several visual control approaches that do not require exact camera calibration knowledge are presented. For example, in Subsect. "Jacobian Estimation Approach", a visual servo controller that is based on a Jacobian matrix estimator is described. In Subsect. "Robust Control Approach", a robust controller based on static best-guess estimation of the calibration matrix is developed to solve the regulation problem. In Subsect. "Adaptive Image–Based Visual Servo Control Approach", an adaptive controller for image-based dynamic regulation control of a robot manipulator is described based on a depth-independent interaction matrix. In Subsect. "Adaptive Homography–Based Visual Servo Control Approach", a combined high-gain control approach and an adaptive control strategy are used to regulate the robot end-effector to a desired pose based on the homography technique.

Visual Servo Control Methods

The development in this chapter is focused on the development of visual servo controllers for robotic systems. To develop the controllers, the subsequent development is based on the assumption that images can be acquired, analyzed, and the resulting data can be provided to the controller without restricting the control rates. The sensor data from image-based feedback are feature points. Feature points are pixels in an image that can be identified and tracked between images so that the motion of the camera/robot can be discerned from the image. Image processing techniques can be used to select coplanar and non-collinear feature points within an image. For simplicity, the development in this chapter is based on the assumption that four stationary coplanar and non-collinear feature points [53] denoted by $O_i \forall i = 1, 2, 3, 4$ can be determined from a feature point tracking algorithm (e. g., Kanade–Lucas–Tomasi (KLT) algorithm discussed in [54,55]).

The plane defined by the four feature points is denoted by π as depicted in Fig. 1. The assumption that feature points lie on a plane is not restrictive and does not reduce the generality of the results. For example, if four coplanar target points are not available then the subsequent development can also exploit methods such as the virtual parallax algorithm (e. g., see [16,56]) to create virtual planes from non-planar feature points.

Geometric Relationships

The camera-in-hand configuration is depicted in Fig. 1. In Fig. 1, $\mathcal{F}$ denotes a coordinate frame that is considered to be affixed to the single current camera viewing the object, and a stationary coordinate frame $\mathcal{F}^*$ denotes a constant (a priori determined) desired camera position and orientation that is defined as a desired image. The Euclidean coordinates of the feature points O_i expressed in the frames $\mathcal{F}$ and $\mathcal{F}^*$ are denoted by $\bar{m}_i(t) \in \mathbb{R}^3$ and $\bar{m}_i^* \in \mathbb{R}^3$, respectively, as

$$\bar{m}_i \triangleq \begin{bmatrix} x_i(t) & y_i(t) & z_i(t) \end{bmatrix}^{\mathrm{T}}, \quad \bar{m}_i^* \triangleq \begin{bmatrix} x_i^* & y_i^* & z_i^* \end{bmatrix}^{\mathrm{T}}, \tag{1}$$

where $x_i(t), y_i(t), z_i(t) \in \mathbb{R}$ denote the Euclidean coordinates of the ith feature point, and $x_i^*, y_i^*, z_i^* \in \mathbb{R}$ denote the Euclidean coordinates of the corresponding feature points in the desired/reference image. From standard Euclidean geometry, relationships between $\bar{m}_i(t)$ and $\bar{m}_i^*$ can be determined as

$$\bar{m}_i = x_f + R\bar{m}_i^*, \tag{2}$$

where $R(t) \in SO(3)$ denotes the orientation of $\mathcal{F}^*$ with respect to $\mathcal{F}$, and $x_f(t) \in \mathbb{R}^3$ denotes the translation vector from $\mathcal{F}$ to $\mathcal{F}^*$ expressed in the coordinate frame $\mathcal{F}$. The normalized Euclidean coordinate vectors, denoted by $m_i(t) \in \mathbb{R}^3$ and $m_i^* \in \mathbb{R}^3$, are defined as

$$m_i = \begin{bmatrix} m_{xi} & m_{yi} & 1 \end{bmatrix}^{\mathrm{T}} \triangleq \begin{bmatrix} \frac{x_i}{z_i} & \frac{y_i}{z_i} & 1 \end{bmatrix}^{\mathrm{T}}, \tag{3}$$

$$m_i^* = \begin{bmatrix} m_{xi}^* & m_{yi}^* & 1 \end{bmatrix}^{\mathrm{T}} \triangleq \begin{bmatrix} \frac{x_i^*}{z_i^*} & \frac{y_i^*}{z_i^*} & 1 \end{bmatrix}^{\mathrm{T}}. \tag{4}$$

Based on the Euclidean geometry in Eq. (2), the relationships between $m_i(t)$ and m_i^* can be determined as [53]

$$m_i = \underbrace{\frac{z_i^*}{z_i}}_{\alpha_i} \underbrace{\left(R + \frac{x_f}{d^*} n^{*\mathrm{T}}\right)}_{H} m_i^*, \tag{5}$$

where $\alpha_i(t) \in \mathbb{R}$ is a scaling term, and $H(t) \in \mathbb{R}^{3\times 3}$ denotes the Euclidean homography. The Euclidean homography is composed of: a scaled translation vector which is equal to the translation vector $x_f(t)$ divided by the distance $d^* \in \mathbb{R}$ from the origin of $\mathcal{F}^*$ to the plane π; the rotation between $\mathcal{F}$ and $\mathcal{F}^*$ denoted by $R(t)$; and $n^* \in \mathbb{R}^3$ denoting a constant unit normal to the plane π.

Each feature point O_i on π also has a pixel coordinate $p_i(t) \in \mathbb{R}^3$ and $p_i^* \in \mathbb{R}^3$ expressed in the image coordinate frame for the current image and the desired image denoted by

$$p_i \triangleq \begin{bmatrix} u_i & v_i & 1 \end{bmatrix}^{\mathrm{T}}, \quad p_i^* \triangleq \begin{bmatrix} u_i^* & v_i^* & 1 \end{bmatrix}^{\mathrm{T}}, \tag{6}$$

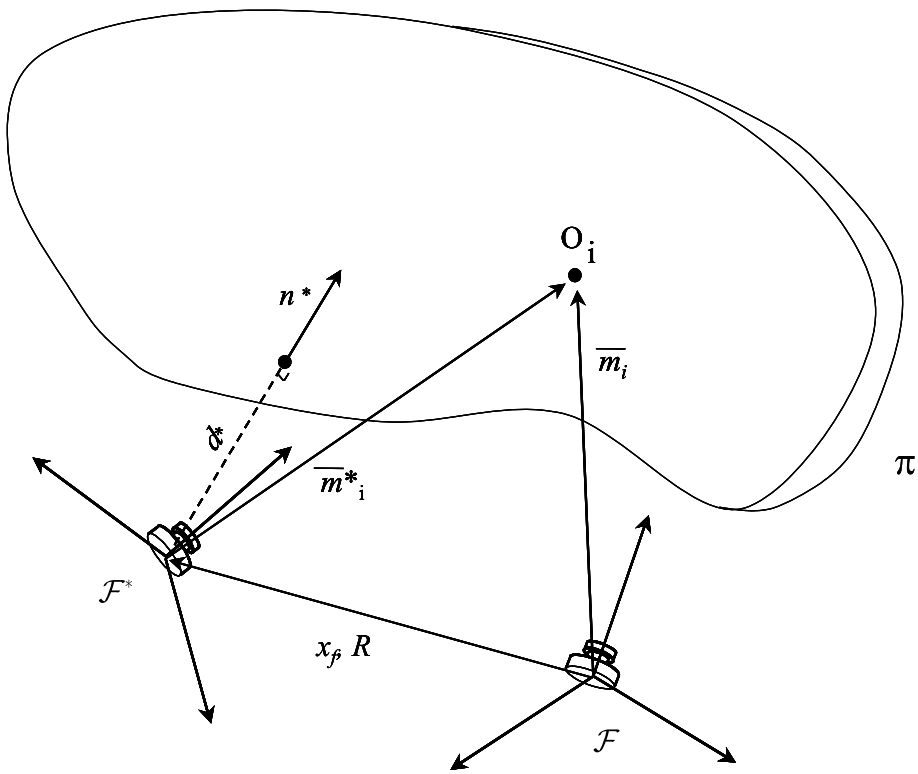

Adaptive Visual Servo Control, Figure 1
Coordinate frame relationships between a camera viewing a planar patch at different spatiotemporal instances

where $u_i(t), v_i(t), u_i^*, v_i^* \in \mathbb{R}$. The pixel coordinates $p_i(t)$ and p_i^* are related to the normalized task-space coordinates $m_i(t)$ and m_i^* by the following global invertible transformation (i. e., the pinhole camera model)

$$p_i = Am_i \,, \quad p_i^* = Am_i^* \,, \tag{7}$$

where $A \in \mathbb{R}^{3\times 3}$ is a constant, upper triangular, and invertible intrinsic camera calibration matrix that is explicitly defined as [57]

$$A \triangleq \begin{bmatrix} \alpha & -\alpha \cot\varphi & u_0 \\ 0 & \frac{\beta}{\sin\varphi} & v_0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & 1 \end{bmatrix} , \tag{8}$$

where $a_{11}, a_{12}, a_{13}, a_{22}, a_{23} \in \mathbb{R}$. In Eq. (8), $u_0, v_0 \in \mathbb{R}$ denote the pixel coordinates of the principal point (i. e., the image center that is defined as the frame buffer coordinates of the intersection of the optical axis with the image plane), $\alpha, \beta \in \mathbb{R}$ represent the product of the camera scaling factors and the focal length, and $\varphi \in \mathbb{R}$ is the skew angle between the camera axes. Based on the physical meaning, the diagonal calibration elements are positive (i. e., $a_{11}, a_{22} > 0$). Based on Eq. (7), the Euclidean relationship in Eq. (5) can be expressed in terms of the image coordinates as

$$p_i = \alpha_i \underbrace{\left(AHA^{-1}\right)}_{G} p_i^* \,, \tag{9}$$

where $G(t) \in \mathbb{R}^{3\times 3}$ denotes the projective homography.

Based on the feature point pairs $\left(p_i^*, p_i(t)\right)$ with $i = 1, 2, 3, 4$, the projective homography can be determined up to a scalar multiple (i. e., $\frac{G(t)}{g_{33}(t)}$ where $g_{33}(t) \in \mathbb{R}$ denotes the bottom right element of $G(t)$) [19]. Various methods can then be applied (e. g., see [58,59]) to decompose the Euclidean homography to obtain the rotation matrix $R(t)$, scaled translation $\frac{x_f(t)}{d^*}$, normal vector n^*, and the depth ratio $\alpha_i(t)$.

Image-Based Visual Servo Control

As described previously, image-based visual servo control uses pure image-space information (i. e., pixel coordinates) in the control development. The image error for the feature point O_i between the current and desired poses

is defined as

$$e_{ii}(t) = \begin{bmatrix} u_i(t) - u_i^* \\ v_i(t) - v_{vi}^* \end{bmatrix} \in \mathbb{R}^2 .$$

The derivative of $e_{ii}(t)$ can be written as (e. g., see [3,4,5])

$$\dot{e}_{ii} = L_{ii} \begin{bmatrix} v_c \\ \omega_c \end{bmatrix} , \tag{10}$$

where $L_{ii}(m_i, z_i) \in \mathbb{R}^{2n\times 6}$ is defined as

$$L_{ii} = \begin{bmatrix} a_{11} & a_{12} \\ 0 & a_{22} \end{bmatrix} \times \begin{bmatrix} -\frac{1}{z_i} & 0 & \frac{m_{xi}}{z_i} & m_{xi}m_{yi} & -(1+m_{xi}^2) & m_{yi} \\ 0 & -\frac{1}{z_i} & \frac{m_{yi}}{z_i} & 1+m_{yi}^2 & -m_{xi}m_{yi} & -m_{xi} \end{bmatrix} ,$$

and $v_c(t), \omega_c(t) \in \mathbb{R}^3$ are linear and angular velocities of the camera, respectively, expressed in the camera coordinate frame $\mathcal{F}$. If the image error is defined based on the normalized pixel coordinates as

$$e_{ii}(t) = \begin{bmatrix} m_{xi}(t) - m_{xi}^* \\ m_{yi}(t) - m_{yi}^* \end{bmatrix} \in \mathbb{R}^2$$

then the derivative of $e_{ii}(t)$ can be written as (e. g., see [13])

$$\dot{e}_{ii} = \begin{bmatrix} -\frac{1}{z_i} & 0 & \frac{m_{xi}}{z_i} & m_{xi}m_{yi} & -(1+m_{xi}^2) & m_{yi} \\ 0 & -\frac{1}{z_i} & \frac{m_{yi}}{z_i} & 1+m_{yi}^2 & -m_{xi}m_{yi} & -m_{xi} \end{bmatrix} \times \begin{bmatrix} v_c \\ \omega_c \end{bmatrix} . \tag{11}$$

The image error for n ($n \geq 3$) feature points is defined as

$$e_i(t) = \begin{bmatrix} e_{i1}^T(t) & e_{i2}^T(t) & \cdots & e_{in}^T(t) \end{bmatrix}^T \in \mathbb{R}^{2n} .$$

Based on Eq. (10), the open-loop error system for $e_i(t)$ can be written as

$$\dot{e}_i = L_i \begin{bmatrix} v_c \\ \omega_c \end{bmatrix} , \tag{12}$$

where the matrix

$$L_i(t) = \begin{bmatrix} L_{i1}^T(t) & L_{i2}^T(t) & \cdots & L_{in}^T(t) \end{bmatrix}^T \in \mathbb{R}^{2n\times 6}$$

is called the image Jacobian, interaction matrix, feature sensitivity matrix, or feature Jacobian (cf. [2,3,4,5]). When the image Jacobian matrix is nonsingular at the desired pose (i. e., $e_i(t) = 0$), and the depth information $z_i(t)$ is available, a locally exponentially result (i. e., $e_i(t) \to 0$ exponentially) can be obtained by using the controller

$$\begin{bmatrix} v_c \\ \omega_c \end{bmatrix} = -k_i L_i^+ e_i , \tag{13}$$

where $k_i \in \mathbb{R}$ is a positive control gain, and $L_i^+(t)$ is the pseudo inverse of $L_i(t)$. See [13] for details regarding the stability analysis.

In the controller Eq. (13), $z_i(t)$ is unknown. Direct approximation and estimation techniques (cf. [4,5]) have been developed to estimate $z_i(t)$. For example, an approximate estimate of $L_i(t)$ using constant z_i^* at the desired pose has been used in [4]. Besides the direct approximation and estimation approaches, dynamic adaptive compensation approaches (cf. [19,20,21,22,24,25,26,27,28]) have also been developed recently to compensate for this unknown depth information, which will be represented in Sect. "Compensating for Projection Normalization".

Position–Based Visual Servo Control

The feedback loop in position-based visual servo control (e. g., [1,5,9,10,11]) uses Euclidean pose information that is reconstructed from the image information. Various methods can be applied (cf. [5,60,61]) to compute the pose estimate based on the image information and a priori geometric knowledge of the environment. The resulting outcome of these methods are the translation vector $x_f(t)$ and rotation matrix $R(t)$ between the current and desired poses of the robot end-effector. The rotation matrix can be further decomposed as a rotation axis $u(t) \in \mathbb{R}^3$ and a rotation angle $\theta(t) \in \mathbb{R}$.

The pose error $e_p(t) \in \mathbb{R}^6$ can be defined as

$$e_p(t) = \begin{bmatrix} x_f \\ u\theta \end{bmatrix} ,$$

where the corresponding open-loop error system can then be determined as

$$\dot{e}_p = L_p \begin{bmatrix} v_c \\ \omega_c \end{bmatrix} , \tag{14}$$

where $L_p(t) \in \mathbb{R}^{6\times 6}$ is defined as

$$L_p = \begin{bmatrix} R & 0 \\ 0 & L_\omega \end{bmatrix} . \tag{15}$$

In Eq. (15), $L_\omega(u(t), \theta(t)) \in \mathbb{R}^{3\times 3}$ is defined as [15]

$$L_\omega = I_3 - \frac{\theta}{2}[u]_\times + \left(1 - \frac{\operatorname{sinc}(\theta)}{\operatorname{sinc}^2\left(\frac{\theta}{2}\right)}\right)[u]_\times^2 . \tag{16}$$

Based on the open-loop error system in Eq. (14), an example position-based visual servo controller can be designed as

$$\begin{bmatrix} v_c \\ \omega_c \end{bmatrix} = -k_p L_p^{-1} e_p \tag{17}$$

provided that $L_p(t)$ is nonsingular, where $k_p \in \mathbb{R}$ is a positive control gain. The controller in Eq. (17) yields an asymptotically stable result in the sense that $e_p(t)$ asymptotically converges to zero. See [13] for details regarding the stability analysis.

A key issue with position-based visual servo control is the need to reconstruct the Euclidean information to find the pose estimate from the image. Given detailed geometric knowledge of the scene (such as a CAD model) approaches such as [5,60,61] can deliver pose estimates. When geometric knowledge is not available, homography-based Euclidean reconstruction is also a good option. Given a set of feature points as stated in Subsect. "Geometric Relationships" and based on Eqs. (5) and (7), the homography $H(t)$ between the current pose and desired pose can be obtained if the camera calibration matrix A is known. Various methods can then be applied (e. g., see [58,59]) to decompose the homography to obtain the rotation matrices $R(t)$ and the scaled translation vector $\frac{x_f(t)}{d^*}$. If $\frac{x_f(t)}{d^*}$ is used in the control development, then either d^* should be estimated directly or adaptive control techniques would be required to compensate for/identify the unknown scalar.

Homography–Based Visual Servo Control

The 2.5D visual servo control approach developed in [15,16,17,18] decomposes the six-DOF motion into translation and rotation components and uses separate translation and rotation controllers to achieve the asymptotic regulation. This class of controllers is also called homography-based visual servo control in [19,20,21,22] because of the underlying reliance of the construction and decomposition of a homography.

The translation error signal $e_v(t) \in \mathbb{R}^3$ can be chosen as a combination of image information (or normalized Euclidean information) and reconstructed Euclidean information as

$$e_v = \left[m_{xi} - m_{xi}^* \quad m_{yi} - m_{yi}^* \quad \ln\left(\frac{z_i}{z_i^*}\right) \right]^{\mathrm{T}} , \tag{18}$$

where the current and desired Euclidean coordinates are related to the chosen feature point O_i. The rotation error signal $e_\omega(t) \in \mathbb{R}^3$ is typically chosen in terms of the reconstructed rotation parameterization as

$$e_\omega = u\theta \, .$$

The corresponding translation and rotation error system can be written as

$$\dot{e}_v = L_v v_c + L_{v\omega} \omega_c \tag{19}$$

$$\dot{e}_\omega = L_\omega \omega_c \, , \tag{20}$$

where $L_\omega(t)$ was defined in Eq. (16), and $L_v(t), L_{v\omega}(t) \in \mathbb{R}^{3\times 3}$ are defined as

$$\begin{aligned} L_v &= -\frac{\alpha_i}{z_i^*} \begin{bmatrix} 1 & 0 & -m_{xi} \\ 0 & 1 & -m_{yi} \\ 0 & 0 & 1 \end{bmatrix} , \\ L_{v\omega} &= \begin{bmatrix} m_{xi} m_{yi} & -1 - m_{xi}^2 & m_{yi} \\ 1 + m_{yi}^2 & -m_{xi} m_{yi} & -m_{xi} \\ -m_{yi} & m_{xi} & 0 \end{bmatrix} . \end{aligned} \tag{21}$$

The rotation and translation controllers can be designed as

$$\begin{aligned} \omega_c &= -k L_\omega^{-1} e_\omega = -k e_\omega \\ v_c &= -L_v^{-1} (k e_v - k L_{v\omega} e_\omega) \end{aligned} \tag{22}$$

to yield an asymptotically stable result.

There are some other control approaches that combine image-space and reconstructed Euclidean information in the error system such as [8,23]. The partitioned visual servo control approach proposed in [8] decouples the z-axis motions (including both the translation and rotation components) from the other degrees of freedom and derives separate controllers for these z-axis motions. The switched visual servo control approach proposed in [23] partitions the control along the time axis, instead of the usual partitioning along spatial degrees of freedom. This system uses both position-based and image-based controllers, and a high-level decision maker invokes the appropriate controller based on the values of a set of Lyapunov functions.

Simulation Comparisons

The controllers given in Eqs. (13), (17), and (22) are implemented based on the same initial conditions. The errors asymptotically converge for all three controllers. The different image-space and Euclidean trajectories in the simulation are shown in Fig. 2.

Subplots a–c of Fig. 2 are two-dimensional image-space trajectories of the four feature points generated by the controllers Eqs. (13), (17), and (22), respectively. The trajectories start from the initial positions and asymptotically converge to the desired positions. A comparison of the results indicate that the image-based visual servo control and homography-based visual servo control have better performance in the image-space (see subplots a and c).

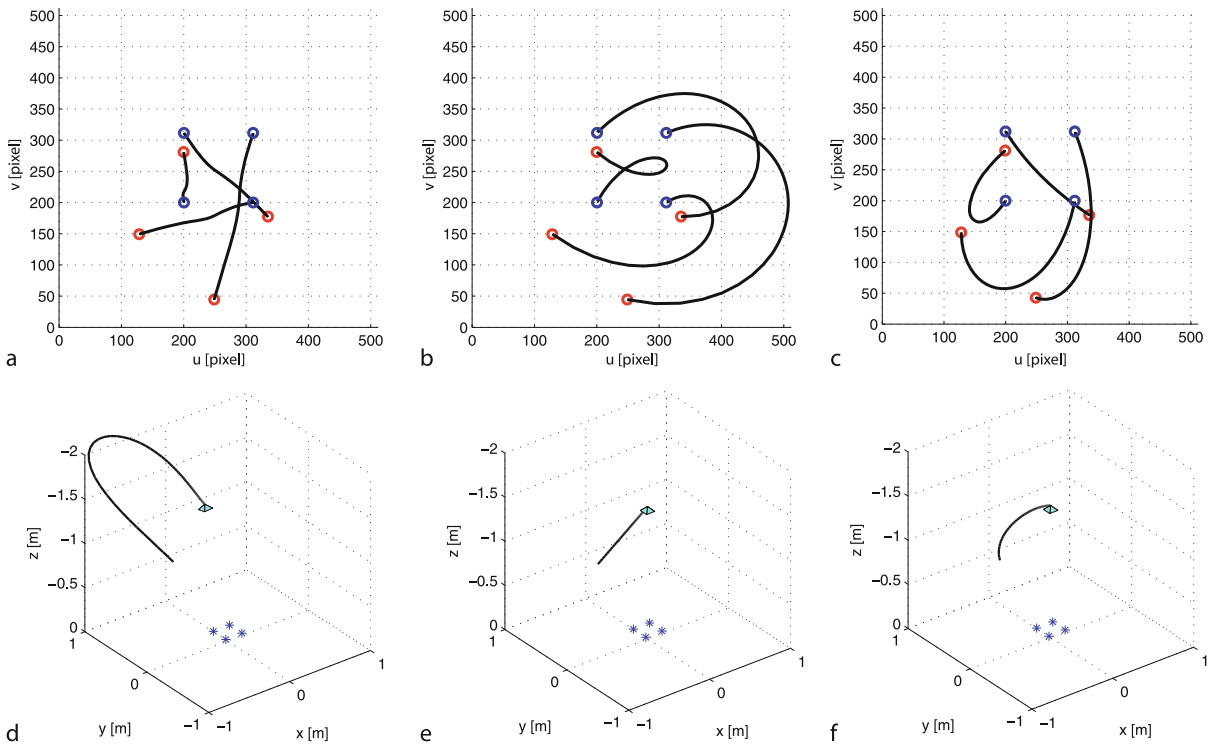

Adaptive Visual Servo Control, Figure 2
Simulation results for image-based, position-based, and combined visual servo control approaches

Also, since the position-based visual servo controller does not use image-features in the feedback, the feature points may leave the camera's field-of-view (see subplot b).

Subplots d–f of Fig. 2 are the three-dimensional Euclidean trajectories of the camera generated by the controllers Eqs. (13), (17), and (22), respectively. Since the position-based visual servo control uses the Euclidean information directly in the control loop, it has the best performance in the Euclidean trajectory (subplot e), followed by the homography-based controller (subplot f). The image-based visual servo controller in subplot d generates the longest Euclidean trajectory. Since image-based visual servo controller doesn't guarantee the Euclidean performance, it may generate a physically unimplementable trajectory.

Compensating for Projection Normalization

A key issue that impacts visual servo control is that the image-space is a two-dimensional projection of the three-dimensional Euclidean-space. To compensate for the lack of depth information from the two-dimensional image data, some researchers have focused on the use of alternate sensors (e. g., laser and sound ranging technologies). Other researchers have explored the use of a camera-based vision system in conjunction with other sensors along with sensor fusion methods or the use of additional cameras in a stereo configuration that triangulate on corresponding images. However, the practical drawbacks of incorporating additional sensors include: increased cost, increased complexity, decreased reliability, and increased processing burden. Recently, some researchers have developed adaptive control methods that can actively adapt for the unknown depth parameter (cf. [19,20,21,22,24,25,26,27,28]) along with methods based on direct estimation and approximation (cf. [4,5]).

Adaptive Depth Compensation in Homography–Based Visual Servo Control

As in [19,20,21,22,26,27,28], homography-based visual servo controllers have been developed to achieve asymptotic regulation or tracking control of a manipulator end-effector or a mobile robot, by developing an adaptive update law that actively compensates for an unknown depth parameter. The open-loop error system and control development are discussed in this section.

Open–Loop Error System The translation error signal $e_v(t)$ is defined as in Subsect "Homography–Based Visual Servo Control". However, the rotation error signal is defined as a unit quaternion $q(t) \triangleq \begin{bmatrix} q_0(t) & q_v^{\mathrm{T}}(t) \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^4$ where $q_0(t) \in \mathbb{R}$ and $q_v(t) \in \mathbb{R}^3$ (e. g., see [22,28]). The corresponding translation and rotation error systems are determined as

$$z_i^* \dot{e}_v = -\alpha_i \bar{L}_v v_c + z_i^* L_{v\omega} \omega_c \tag{23}$$

and

$$\begin{bmatrix} \dot{q}_0 \\ \dot{q}_v \end{bmatrix} = \frac{1}{2} \begin{bmatrix} -q_v^{\mathrm{T}} \\ q_0 I_3 + q_v^{\times} \end{bmatrix} \omega_c \,, \tag{24}$$

where the notation $q_v^{\times}(t)$ denotes the skew-symmetric form of the vector $q_v(t)$ (e. g., see [22,28]), and $\bar{L}_v(t) \in \mathbb{R}^{3\times 3}$ is defined as

$$\bar{L}_v = \begin{bmatrix} 1 & 0 & -m_{xi} \\ 0 & 1 & -m_{yi} \\ 0 & 0 & 1 \end{bmatrix} . \tag{25}$$

The structure of Eq. (25) indicates that $\bar{L}_v(t)$ is measurable and can be used in the controller development directly.

Controller and Adaptive Law Design The rotation and translation controllers can be developed as (e. g., see [15,19,22,28])

$$\omega_c = -K_\omega q_v \,, \tag{26}$$

$$v_c = \frac{1}{\alpha_i} \bar{L}_v^{-1} \left(K_v e_v + \hat{z}_i^* L_{v\omega} \omega_c \right) \,, \tag{27}$$

where $K_\omega, K_v \in \mathbb{R}^{3\times 3}$ denote two diagonal matrices of positive constant control gains. In Eq. (27), the parameter estimate $\hat{z}_i^*(t) \in \mathbb{R}$ is developed for the unknown constant z_i^*, and is defined as (e. g., see [19,22,28])

$$\dot{\hat{z}}_i^* = \gamma_i e_v^{\mathrm{T}} L_{v\omega} \omega_c \,, \tag{28}$$

where $\gamma_i \in \mathbb{R}$ denotes a positive constant adaptation gain. Based on Eqs. (23)–(28), the closed-loop error system is obtained as

$$\dot{q}_0 = \frac{1}{2} q_v^{\mathrm{T}} K_\omega q_v \,, \tag{29}$$

$$\dot{q}_v = -\frac{1}{2} \left(q_0 I_3 + q_v^{\times} \right) K_\omega q_v \,, \tag{30}$$

$$z_i^* \dot{e}_v = -K_v e_v + \tilde{z}_i L_{v\omega} \omega_c \,, \tag{31}$$

where $\tilde{z}_i(t) \in \mathbb{R}$ denotes the following parameter estimation error:

$$\tilde{z}_i = z_i^* - \hat{z}_i^* \,. \tag{32}$$

Closed–Loop Stability Analysis The controller given in Eqs. (26) and (27) along with the adaptive update law in Eq. (28) ensures asymptotic translation and rotation regulation in the sense that

$$\|q_v(t)\| \to 0 \quad \text{and} \quad \|e_v(t)\| \to 0 \quad \text{as} \quad t \to \infty \,. \tag{33}$$

To prove the stability result, a Lyapunov function candidate $V(t) \in \mathbb{R}$ can be defined as

$$V = \frac{z_i^*}{2} e_v^{\mathrm{T}} e_v + q_v^{\mathrm{T}} q_v + (1 - q_0)^2 + \frac{1}{2\gamma_i} \tilde{z}_i^2 \,. \tag{34}$$

The time-derivative of $V(t)$ can be determined as

$$\begin{aligned} \dot{V} = {} & e_v^{\mathrm{T}} \left(-K_v e_v + \tilde{z}_i L_{v\omega} \omega_c \right) - q_v^{\mathrm{T}} \left(q_0 I_3 + q_v^{\times} \right) K_\omega q_v \\ & - (1 - q_0) q_v^{\mathrm{T}} K_\omega q_v - \tilde{z}_i e_v^{\mathrm{T}} L_{v\omega} \omega_c \,, \end{aligned} \tag{35}$$

where Eqs. (29)-(32) were utilized. The following negative semi-definite expression is obtained after simplifying Eq. (35):

$$\dot{V} = -e_v^{\mathrm{T}} K_v e_v - q_v^{\mathrm{T}} K_\omega q_v \,. \tag{36}$$

Barbalat's Lemma [62] can then be used to prove the result given in Eq. (33).

Based on the controller in Eq. (26) and (27) and the adaptive law in Eq. (28), the resulting asymptotic translation and rotation errors are plotted in Fig. 3 and Fig. 4, respectively. The image-space trajectory is shown in Fig. 5, and also in Fig. 6 in a three-dimensional format, where the vertical axis is time. The parameter estimate for z_i^* is shown in Fig. 7.

Adaptive Depth Compensation in Image–Based Visual Servo Control

This section provides an example of how an image-based visual servo controller can be developed that adaptively compensates for the uncertain depth. In this section, only three feature points are required to be tracked in the image to develop the error system (i. e., to obtain an image-Jacobian with at least equal rows than columns), but at least four feature points are required to solve for the unknown depth ratio $\alpha_i(t)$.

Open–Loop Error System The image error $e_i(t)$ defined in Subsect. "Image-based Visual Servo Control" can be

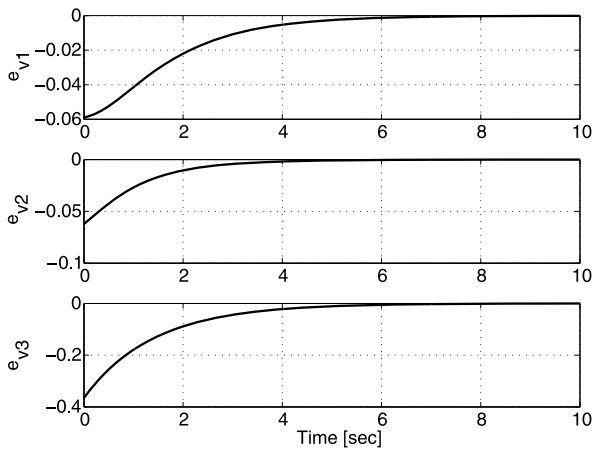

Adaptive Visual Servo Control, Figure 3
Unitless translation error $e_v(t)$

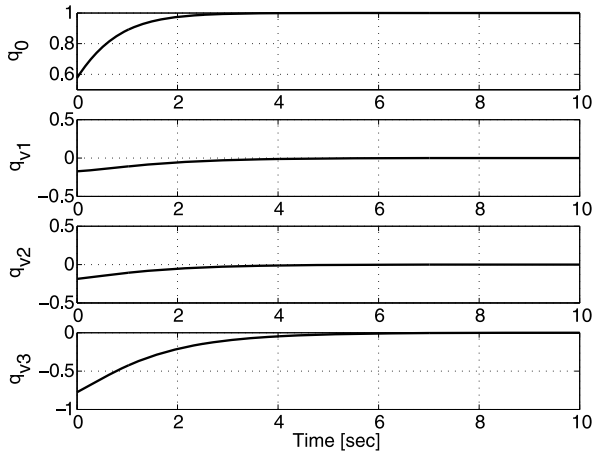

Adaptive Visual Servo Control, Figure 4
Quaternion rotation error $q(t)$

rewritten as

$$e_i(t) = \begin{bmatrix} m_{x1}(t) - m_{x1}^* \\ m_{y1}(t) - m_{y1}^* \\ m_{x2}(t) - m_{x2}^* \\ m_{y2}(t) - m_{y2}^* \\ m_{x3}(t) - m_{x3}^* \\ m_{y3}(t) - m_{y3}^* \end{bmatrix} \triangleq \begin{bmatrix} e_{x1} \\ e_{y1} \\ e_{x2} \\ e_{y2} \\ e_{x3} \\ e_{y3} \end{bmatrix}, \tag{37}$$

when three feature points are used for notation simplicity. The definition of image error $e_i(t)$ is expandable to more than three feature points. The time derivative of $e_i(t)$ is given by

$$\dot{e}_i(t) = L_i \begin{bmatrix} v_c \\ \omega_c \end{bmatrix}, \tag{38}$$

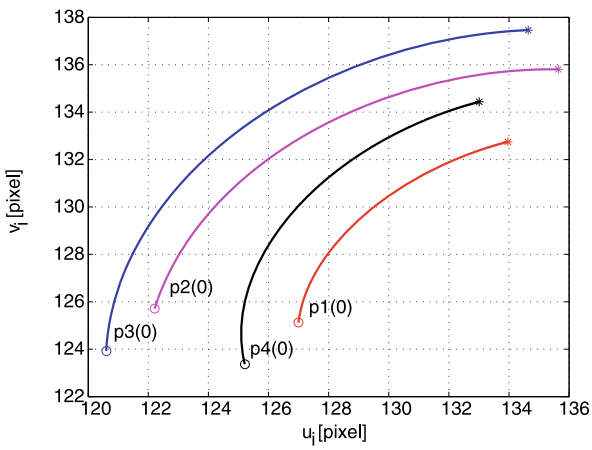

Adaptive Visual Servo Control, Figure 5
Image-space error in pixels between $p(t)$ and p^*. In the figure, "O" denotes the initial positions of the 4 feature points in the image, and "*" denotes the corresponding final positions of the feature points

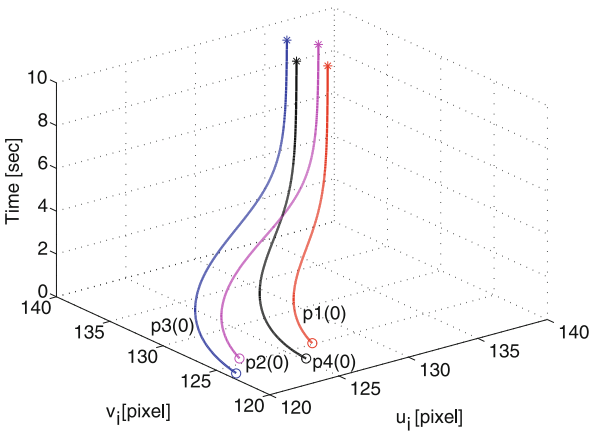

Adaptive Visual Servo Control, Figure 6
Image-space error in pixels between $p(t)$ and p^* shown in a 3D graph. In the figure, "O" denotes the initial positions of the 4 feature points in the image, and "*" denotes the corresponding final positions of the feature points

where the image Jacobian $L_i(t)$ is given by

$$L_i = \begin{bmatrix} \frac{\alpha_1}{z_1^*} & 0 & \frac{\alpha_1 m_{x1}}{z_1^*} & -m_{x1}m_{y1} & 1+m_{x1}^2 & -m_{y1} \\ 0 & \frac{\alpha_1}{z_1^*} & \frac{\alpha_1 m_{y1}}{z_1^*} & -1-m_{y1}^2 & m_{x1}m_{y1} & m_{x1} \\ \frac{\alpha_2}{z_2^*} & 0 & \frac{\alpha_2 m_{x2}}{z_2^*} & -m_{x2}m_{y2} & 1+m_{x2}^2 & -m_{y2} \\ 0 & \frac{\alpha_2}{z_2^*} & \frac{\alpha_2 m_{y2}}{z_2^*} & -1-m_{y2}^2 & m_{x2}m_{y2} & m_{x2} \\ \frac{\alpha_3}{z_3^*} & 0 & \frac{\alpha_3 m_{x3}}{z_3^*} & -m_{x3}m_{y3} & 1+m_{x3}^2 & -m_{y3} \\ 0 & \frac{\alpha_3}{z_3^*} & \frac{\alpha_3 m_{y3}}{z_3^*} & -1-m_{y3}^2 & m_{x3}m_{y3} & m_{x3} \end{bmatrix}, \tag{39}$$

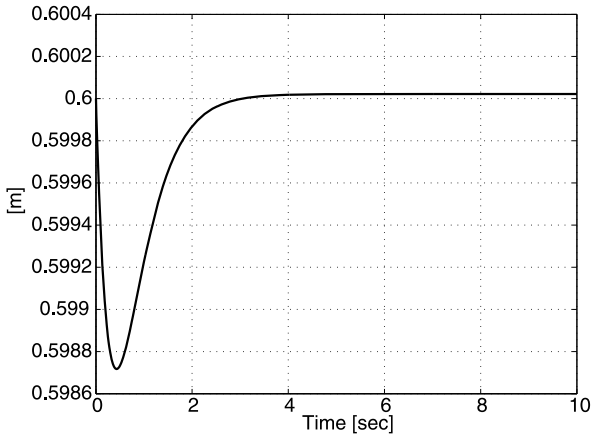

Adaptive Visual Servo Control, Figure 7
Adaptive on-line estimate of z_i^*

where the depth ratio $\alpha_i(t)$, $i = 1, 2, 3$ is defined in (5). Let $\phi = \begin{bmatrix}\phi_1 & \phi_2 & \phi_3\end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^3$ denote the unknown constant parameter vector as

$$\phi = \begin{bmatrix}\frac{1}{z_1^*} & \frac{1}{z_2^*} & \frac{1}{z_3^*}\end{bmatrix}^{\mathrm{T}} . \tag{40}$$

Even though the error system (38) is not linearly parameterizable in ϕ, an adaptive law can still be developed based on the following property in Paragraph "Controller and Adaptive Law Design".

Controller and Adaptive Law Design Based on an open-loop error system as in Eq. (38), the controller is developed as

$$\begin{bmatrix} v_c \\ \omega_c \end{bmatrix} = -k_i \hat{L}_i^{\mathrm{T}} e_i , \tag{41}$$

where $k_i \in \mathbb{R}$ is the control gain,

$$\hat{\phi}(t) = \begin{bmatrix}\hat{\phi}_1(t) \hat{\phi}_2(t) \hat{\phi}_3(t)\end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^3$$

is the adaptive estimation of the unknown constant parameter vector defined in Eq. (40), and $\hat{L}_i(t) \in \mathbb{R}^{6\times 6}$ is an estimation of the image Jacobian matrix as

$$\hat{L}_i = \begin{bmatrix} \alpha_1\hat{\phi}_1 & 0 & \alpha_1 m_{x1}\hat{\phi}_1 & -m_{x1}m_{y1} & 1+m_{x1}^2 & -m_{y1} \\ 0 & \alpha_1\hat{\phi}_1 & \alpha_1 m_{y1}\hat{\phi}_1 & -1-m_{y1}^2 & m_{x1}m_{y1} & m_{x1} \\ \alpha_2\hat{\phi}_2 & 0 & \alpha_2 m_{x2}\hat{\phi}_2 & -m_{x2}m_{y2} & 1+m_{x2}^2 & -m_{y2} \\ 0 & \alpha_2\hat{\phi}_2 & \alpha_2 m_{y2}\hat{\phi}_2 & -1-m_{y2}^2 & m_{x2}m_{y2} & m_{x2} \\ \alpha_3\hat{\phi}_3 & 0 & \alpha_3 m_{x3}\hat{\phi}_3 & -m_{x3}m_{y3} & 1+m_{x3}^2 & -m_{y3} \\ 0 & \alpha_3\hat{\phi}_3 & \alpha_3 m_{y3}\hat{\phi}_3 & -1-m_{y3}^2 & m_{x3}m_{y3} & m_{x3} \end{bmatrix} . \tag{42}$$

The adaptive law for $\hat{\phi}(t)$ in Eq. (42) is designed as

$$\dot{\hat{\phi}} = -\Gamma e_M \hat{L}_{iM} \hat{L}_i^{\mathrm{T}} e_i , \tag{43}$$

where $\Gamma \in \mathbb{R}^{3\times 3}$ is a positive diagonal matrix, $e_M(t) \in \mathbb{R}^{3\times 6}$ and $\hat{L}_{iM}(t) \in \mathbb{R}^{6\times 6}$ are defined as

$$e_M = \begin{bmatrix} e_{x1} & e_{y1} & 0 & 0 & 0 & 0 \\ 0 & 0 & e_{x2} & e_{y2} & 0 & 0 \\ 0 & 0 & 0 & 0 & e_{x3} & e_{y3} \end{bmatrix}$$

$$\hat{L}_{iM} = \begin{bmatrix} \alpha_1 & 0 & m_{x1}\alpha_1 & 0 & 0 & 0 \\ 0 & \alpha_1 & m_{y1}\alpha_1 & 0 & 0 & 0 \\ \alpha_2 & 0 & m_{x2}\alpha_2 & 0 & 0 & 0 \\ 0 & \alpha_2 & m_{y2}\alpha_2 & 0 & 0 & 0 \\ \alpha_3 & 0 & m_{x3}\alpha_3 & 0 & 0 & 0 \\ 0 & \alpha_3 & m_{y3}\alpha_3 & 0 & 0 & 0 \end{bmatrix} .$$

The adaptive law is motivated by the following two properties:

$$L_i - \hat{L}_i = \tilde{L}_i = \Phi_M \hat{L}_{iM} , \tag{44}$$

$$e_i^{\mathrm{T}} \Phi_M \hat{L}_{iM} = \tilde{\phi}^{\mathrm{T}} e_M \hat{L}_{iM} . \tag{45}$$

In Eqs. (44) and (45), the adaptive estimation error $\tilde{\phi}(t) \in \mathbb{R}^3$ and $\Phi_M(t) \in \mathbb{R}^{6\times 6}$ are defined as

$$\begin{aligned} \tilde{\phi} &= \begin{bmatrix}\tilde{\phi}_1 & \tilde{\phi}_2 & \tilde{\phi}_3\end{bmatrix}^{\mathrm{T}} \\ &= \begin{bmatrix}\phi_1 - \hat{\phi}_1 & \phi_2 - \hat{\phi}_2 & \phi_3 - \hat{\phi}_3\end{bmatrix}^{\mathrm{T}} , \\ \Phi_M &= \mathrm{diag}\left\{\tilde{\phi}_1, \tilde{\phi}_1, \tilde{\phi}_2, \tilde{\phi}_2, \tilde{\phi}_3, \tilde{\phi}_3\right\} . \end{aligned}$$

Under the controller Eq. (41) and the adaptive law Eq. (43), the closed-loop error system can be written as

$$\begin{aligned} \dot{e}_i &= -k_i L_i \hat{L}_i^{\mathrm{T}} e_i = -k_i \hat{L}_i \hat{L}_i^{\mathrm{T}} e_i - k_i \tilde{L}_i \hat{L}_i^{\mathrm{T}} e_i \\ &= -k_i \hat{L}_i \hat{L}_i^{\mathrm{T}} e_i - k_i \Phi_M \hat{L}_{iM} \hat{L}_i^{\mathrm{T}} e_i . \end{aligned} \tag{46}$$

Closed–Loop Stability Analysis The controller given in Eq. (41) and the adaptive update law in Eq. (43) ensures asymptotic regulation in the sense that

$$\|e_i(t)\| \to 0 \quad \text{as} \quad t \to \infty . \tag{47}$$

To prove the stability result, a Lyapunov function candidate $V(t) \in \mathbb{R}$ can be defined as

$$V = \frac{1}{2} e_i^{\mathrm{T}} e_i + \frac{1}{2} k_i \tilde{\phi} \Gamma^{-1} \tilde{\phi} .$$

The time-derivative of $V(t)$ can be determined as

$$\begin{aligned} \dot{V} &= e_i^{\mathrm{T}} \dot{e}_i - k_i \tilde{\phi} \Gamma^{-1} \dot{\hat{\phi}} \\ &= e_i^{\mathrm{T}} \left(-k_i \hat{L}_i \hat{L}_i^{\mathrm{T}} e_i - k_i \Phi_M \hat{L}_{iM} \hat{L}_i^{\mathrm{T}} e_i\right) \\ &\quad - k_i \tilde{\phi} \Gamma^{-1} \left(\Gamma e_M \hat{L}_{iM} e_i\right) \\ &= -k_i e_i^{\mathrm{T}} \hat{L}_i \hat{L}_i^{\mathrm{T}} e_i - k_i e_i^{\mathrm{T}} \Phi_M \hat{L}_{iM} \hat{L}_i^{\mathrm{T}} e_i + k_i \tilde{\phi} e_M \hat{L}_{iM} \hat{L}_i^{\mathrm{T}} e_i \end{aligned}$$

Based on Eqs. (44) and (45)

$$e_i^{\mathrm{T}} \Phi_M \hat{L}_{iM} \hat{L}_i^{\mathrm{T}} = \tilde{\phi}^{\mathrm{T}} e_M \hat{L}_{iM} \hat{L}_i^{\mathrm{T}} ,$$

and $\dot{V}(t)$ can be further simplified as

$$\begin{aligned}\dot{V} &= -k_i e_i^{\mathrm{T}} \hat{L}_i \hat{L}_i^{\mathrm{T}} e_i - k_i \tilde{\phi} e_M \hat{L}_{iM} \hat{L}_i^{\mathrm{T}} e_i + k_i \tilde{\phi} e_M \hat{L}_{iM} \hat{L}_i^{\mathrm{T}} e_i \\ &= -k_i e_i^{\mathrm{T}} \hat{L}_i \hat{L}_i^{\mathrm{T}} e_i .\end{aligned}$$

Barbalat's Lemma [62] can then be used to conclude the result given in Eq. (47).

Visual Servo Control via an Uncalibrated Camera

Motivated by the desire to incorporate robustness to camera calibration, different control approaches (e. g., Jacobian estimation-based least-square optimization, robust control, and adaptive control) that do not depend on exact camera calibration have been proposed (cf. [18,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,63]).

Jacobian Estimation Approach

The uncalibrated visual servo controllers that are based on a Jacobian matrix estimator have been developed in [36,37,38,39,40]. The camera calibration matrix and kinematic model is not required with the Jacobian matrix estimation. This type of visual servo control scheme is composed of a recursive Jacobian estimation law and a control law. Specifically, in [36], a weighted recursive least-square update law is used for Jacobian estimation, and a visual servo controller is developed via the Lyapunov approach. In [37], a Broyden Jacobian estimator is applied for Jacobian estimation and a nonlinear least-square optimization method is used for the visual servo controller development. In [39,40] a recursive least-squares algorithm is implemented for Jacobian estimation, and a dynamic Gauss–Newton method is used to minimize the squared error in the image plane to get the controller.

Let the robot end-effector features $y(\theta) \in \mathbb{R}^m$ be a function of the robot joint angle vector $\theta(t) \in \mathbb{R}^n$, and let the target features $y^*(t) \in \mathbb{R}^m$ be a function of time. The tracking error in the image-plane can then be defined as

$$f(\theta, t) = y(\theta) - y^*(t) \in \mathbb{R}^m .$$

The trajectory $\theta(t)$ that causes the end-effector to follow the target is established by minimizing the squared image error

$$F(\theta, t) = \frac{1}{2} f^{\mathrm{T}}(\theta, t) f(\theta, t) .$$

By using the dynamic Gauss–Newton method, the joint angle vector update law (controller) can be designed as [40]

$$\theta_{k+1} = \theta_k - \left(\hat{J}_k^{\mathrm{T}} \hat{J}_k\right)^{-1} \hat{J}_k^{\mathrm{T}} \left(f_k + \frac{\partial f_k}{\partial t} h_t \right) , \tag{48}$$

where h_t is the increment of t. In Eq. (48), the estimated Jacobian $\hat{J}(t)$, can be determined by multiple recursive least-square Jacobian estimation algorithms (e. g., Broyden method [37], dynamic Broyden method [40], or an exponentially weighted recursive least-square algorithm [40]).

Robust Control Approach

Robust control approaches based on static best-guess estimation of the calibration matrix have been developed to solve the uncalibrated visual servo regulation problem (cf. [18,44,50,51]). Specifically, a kinematic controller was developed in [44] that utilizes a constant, best-guess estimate of the calibration parameters to achieve local setpoint regulation for the six-DOF problem; although, several conditions on the rotation and calibration matrix are required. Model-free homography-based visual servoing methods using best-guess estimation are used in [18,50] to achieve exponential or asymptotic regulation with respect to both camera and hand-eye calibration errors for the six-DOF problem. A quaternion-based robust controller is developed in [51] that achieved a regulation control objective.

State Estimation Let $\hat{A} \in \mathbb{R}^{3\times3}$ be a constant best-guess estimate of the camera calibration matrix A. The normalized coordinate estimates, denoted by $\hat{m}_i(t)$, $\hat{m}_i^* \in \mathbb{R}^3$, can be expressed as

$$\hat{m}_i = \hat{A}^{-1} p_i = \tilde{A} m_i , \tag{49}$$

$$\hat{m}_i^* = \hat{A}^{-1} p_i^* = \tilde{A} m_i^* , \tag{50}$$

where the upper triangular calibration error matrix $\tilde{A} \in \mathbb{R}^{3\times3}$ is defined as

$$\tilde{A} = \hat{A}^{-1} A . \tag{51}$$

Since $m_i(t)$ and m_i^* can not be exactly determined, the estimates in Eqs. (49) and (50) can be substituted into Eq. (5) to obtain the following relationship:

$$\hat{m}_i = \alpha_i \hat{H} \hat{m}_i^* , \tag{52}$$

where $\hat{H}(t) \in \mathbb{R}^{3\times3}$ denotes the estimated Euclidean homography [18] defined as

$$\hat{H} = \tilde{A} H \tilde{A}^{-1} . \tag{53}$$

Since $\hat{m}_i(t)$ and $\hat{m}_i^*$ can be determined from Eqs. (49) and (50), a set of twelve linear equations can be developed from the four image point pairs, and Eq. (52) can be used to solve for $\hat{H}(t)$. If additional information (e. g., at least four vanishing points) is provided (see [64] for a description of how to determine vanishing points in an image), various techniques (e. g., see [58,59]) can be used to decompose $\hat{H}(t)$ to obtain the estimated rotation and translation components as

$$\hat{H} = \tilde{A}R\tilde{A}^{-1} + \tilde{A}x_h n^{*T}\tilde{A}^{-1} = \hat{R} + \hat{x}_h \hat{n}^{*T} , \tag{54}$$

where $\hat{R}(t) \in \mathbb{R}^{3\times 3}$ and $\hat{x}_h(t), \hat{n}^* \in \mathbb{R}^3$ denote the estimate of $R(t)$, $x_h(t)$ and n^*, respectively, as [18]

$$\hat{R} = \tilde{A}R\tilde{A}^{-1} , \tag{55}$$

$$\hat{x}_h = \sigma \tilde{A} x_h , \tag{56}$$

$$\hat{n}^* = \frac{1}{\sigma}\tilde{A}^{-\mathrm{T}} n^* , \tag{57}$$

where $\sigma \in \mathbb{R}$ denotes the positive constant

$$\sigma = \left\| \tilde{A}^{-\mathrm{T}} n^* \right\| . \tag{58}$$

Based on Eq. (55), the rotation matrix estimate $\hat{R}(t)$ is similar to $R(t)$. By exploiting the properties of similar matrices (i. e., similar matrices have the same trace and eigenvalues), the following estimates can be determined as [18]

$$\hat{\theta} = \theta , \quad \hat{u} = \mu \tilde{A} u ,$$

where $\hat{\theta}(t) \in \mathbb{R}$ and $\hat{u}(t) \in \mathbb{R}^3$ denote the estimates of $\theta(t)$ and $u(t)$, respectively, and $\mu \in \mathbb{R}$ is defined as

$$\mu = \frac{1}{\left\| \tilde{A} u \right\|} .$$

Alternatively, if the rotation matrix is decomposed as a unit quaternion. The quaternion estimate

$$\hat{q}(t) \triangleq \left[\hat{q}_0(t) \quad \hat{q}_v^{\mathrm{T}}(t) \right]^{\mathrm{T}} \in \mathbb{R}^4$$

can be related to the actual quaternion $q(t)$ as [51]

$$\hat{q}_0 = q_0 , \quad \hat{q}_v = \pm \mu \tilde{A} q_v , \tag{59}$$

where

$$\mu = \frac{\| q_v \|}{\left\| \tilde{A} q_v \right\|} . \tag{60}$$

Since the sign ambiguity in Eq. (59) does not affect the control development and stability analysis [51], only the positive sign in Eq. (59) needs to be considered in the following control development and stability analysis.

Control Development and Stability Analysis The rotation open-loop error system can be developed by taking the time derivative of $q(t)$ as

$$\begin{bmatrix} \dot{q}_0 \\ \dot{q}_v \end{bmatrix} = \frac{1}{2} \begin{bmatrix} -q_v^{\mathrm{T}} \\ q_0 I_3 + q_v^{\times} \end{bmatrix} \omega_c . \tag{61}$$

Based on the open-loop error system in (61) and the subsequent stability analysis, the angular velocity controller is designed as

$$\omega_c = -k_\omega \hat{q}_v , \tag{62}$$

where $k_\omega \in \mathbb{R}$ denotes a positive control gain. Substituting Eq. (62) into Eq. (61), the rotation closed-loop error system can be developed as

$$\dot{q}_0 = \frac{1}{2} k_\omega q_v^{\mathrm{T}} \hat{q}_v \tag{63}$$

$$\dot{q}_v = -\frac{1}{2} k_\omega \left(q_0 I_3 + q_v^{\times} \right) \hat{q}_v . \tag{64}$$

The open-loop translation error system can be determined as [50]

$$\dot{e} = -\frac{1}{z_i^*} v_c - \omega_c^{\times} e + \left[m_i^* \right]^{\times} \omega_c , \tag{65}$$

where the translation error $e(t) \in \mathbb{R}^3$ is defined as

$$e = \frac{z_i}{z_i^*} m_i - m_i^* .$$

The translation controller is designed as (see also [18,50])

$$v_c = k_v \hat{e} , \tag{66}$$

where $k_v \in \mathbb{R}$ denotes a positive control gain, and $\hat{e}(t) \in \mathbb{R}^3$ is defined as

$$\hat{e} = \frac{z_i}{z_i^*} \hat{m}_i - \hat{m}_i^* . \tag{67}$$

In Eq. (67), $\hat{m}_i(t)$ and $\hat{m}_i^*$ can be computed from Eq. (49) and (50), respectively, and the ratio $(z_i(t))/z_i^*$ can be computed from the decomposition of the estimated Euclidean homography in Eq. (52).

After substituting Eqs. (62) and (66) into Eq. (65), the resulting closed-loop translation error system can be determined as

$$\dot{e} = \left(-k_v \frac{1}{z_i^*} \tilde{A} + \left[k_\omega \hat{q}_v \right]^{\times} \right) e - k_\omega \left[m_i^* \right]^{\times} \hat{q}_v . \tag{68}$$

The controller given in Eqs. (62) and (66) ensures asymptotic regulation in the sense that

$$\| q_v(t) \| \to 0 , \quad \| e(t) \| \to 0 \quad \text{as} \quad t \to \infty \tag{69}$$

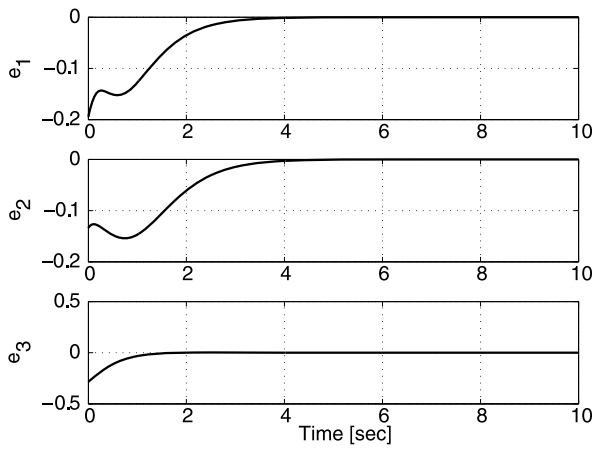

Adaptive Visual Servo Control, Figure 8
Unitless translation error between $m_1(t)$ and m_1^*

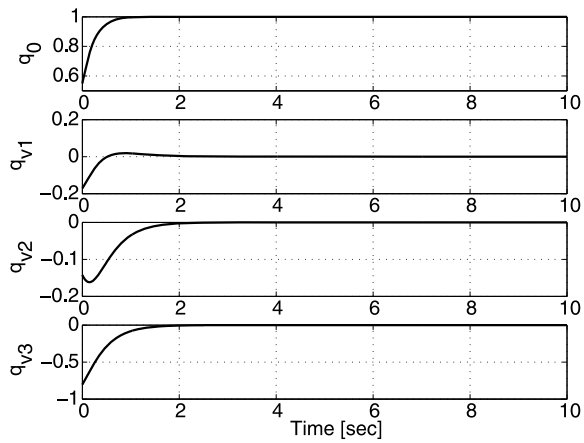

Adaptive Visual Servo Control, Figure 9
Quaternion rotation error $q(t)$

provided k_v is selected sufficiently large, and the following inequalities are satisfied (cf. [18,50,51])

$$\lambda_{\min}\left\{\frac{1}{2}\left(\tilde{A}+\tilde{A}^{\mathrm{T}}\right)\right\} \geq \lambda_0\,, \quad \lambda_{\max}\left\{\frac{1}{2}\left(\tilde{A}+\tilde{A}^{\mathrm{T}}\right)\right\} \leq \lambda_1\,,$$

where $\lambda_0, \lambda_1 \in \mathbb{R}$ are positive constants, and $\lambda_{\min}\{\cdot\}$ and $\lambda_{\max}\{\cdot\}$ denote the minimal and maximal eigenvalues of $1/2(\tilde{A}+\tilde{A}^{\mathrm{T}})$, respectively.

To prove the stability result, a Lyapunov function candidate $V(t) \in \mathbb{R}$ can be defined as

$$V = q_v^{\mathrm{T}} q_v + (1-q_0)^2 + e^{\mathrm{T}} e\,. \tag{70}$$

The time derivative of Eq. (70) can be proven to be negative semi-definite as [51]

$$\dot{V} \leq -\frac{k_\omega \lambda_0 \mu}{K_{v1}}\|q_v\|^2 - \frac{\lambda_0}{z_i^*}\|e\|^2\,, \tag{71}$$

where K_{v1} is a positive large enough number. Using the signal chasing argument in [51], Barbalat's Lemma [62] can then be used to prove the result given in Eq. (69). Based on the controller in Eqs. (62) and (66), the resulting asymptotic translation and rotation errors are plotted in Fig. 8 and Fig. 9, respectively.

Adaptive Image–Based Visual Servo Control Approach

In [49], an adaptive controller for image-based dynamic regulation control of a robot manipulator is proposed using a fixed camera whose intrinsic and extrinsic parameters are not known. A novel depth-independent interaction matrix (cf. [49,65,66]) is proposed therein to map the visual signals onto the joints of the robot manipulator. Due to the depth-independent interaction matrix, the closed-loop dynamics can be linearly parameterized in the constant unknown camera parameters. An adaptive law can then be developed to estimate these parameters online.

Open–Loop Error System Denote the homogeneous coordinate of the feature point expressed in the robot base frame as

$$\bar{x}(t) = \begin{bmatrix} x(t) & y(t) & z(t) & 1 \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^4\,,$$

and denote its coordinate in the camera frame as

$${}^c\bar{x}(t) = \begin{bmatrix} {}^c x(t) & {}^c y(t) & {}^c z(t) \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^3\,.$$

The current and desired positions of the feature point on the image plane are respectively defined as

$$p(t) = \begin{bmatrix} u(t) & v(t) & 1 \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^3 \quad \text{and}$$
$$p_d = \begin{bmatrix} u_d & v_d & 1 \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^3\,.$$

The image error between the current and desired positions of the feature point on the image plane, denoted as $e_i(t) \in \mathbb{R}^3$, is defined as

$$e_i = p - p_d = \begin{bmatrix} u & v & 1 \end{bmatrix}^{\mathrm{T}} - \begin{bmatrix} u_d & v_d & 1 \end{bmatrix}^{\mathrm{T}}\,. \tag{72}$$

Under the camera perspective projection model, $p(t)$ can be related to $\bar{x}(t)$ as

$$p = \frac{1}{{}^c z} M \bar{x}(t)\,, \tag{73}$$

where $M \in \mathbb{R}^{3\times 4}$ is a constant perspective projection matrix that depends only on the intrinsic and extrinsic calibration parameters. Based on Eq. (73), the depth of the

feature point in the camera frame is given by

$$^{c}z(t) = m_3^{\mathrm{T}} \bar{x}(t) \,, \tag{74}$$

where $m_3 \in \mathbb{R}^{1\times 4}$ is the third row of M. Based on Eqs. (72)–(74), the derivative of image error $\dot{e}_i(t)$ is related to $\dot{\bar{x}}(t)$ as

$$\dot{e}_i = \dot{p} = \frac{1}{^{c}z}\left(M - pm_3^{\mathrm{T}}\right)\dot{\bar{x}} \,. \tag{75}$$

The robot kinematics are given by

$$\dot{\bar{x}} = J(q)\dot{q} \,, \tag{76}$$

where $q(t) \in \mathbb{R}^n$ denotes the joint angle of the manipulator, and $J(q) \in \mathbb{R}^{4\times n}$ denotes the manipulator Jacobian. Based on Eqs. (75) and (76),

$$\dot{e}_i = \frac{1}{^{c}z}\left(M - pm_3^{\mathrm{T}}\right) J(q)\dot{q} \,. \tag{77}$$

The robot dynamics are given by

$$H(q)\ddot{q} + \left(\frac{1}{2}\dot{H}(q) + C(q,\dot{q})\right)\dot{q} + g(q) = \tau \,. \tag{78}$$

In Eq. (78), $H(q) \in \mathbb{R}^{n\times n}$ is the positive-definite and symmetric inertia matrix, $C(q,\dot{q}) \in \mathbb{R}^{n\times n}$ is a skew-symmetric matrix, $g(q) \in \mathbb{R}^n$ denotes the gravitational force, and $\tau(t) \in \mathbb{R}^n$ is the joint input of the robot manipulator.

Controller and Adaptive Law Design As stated in Property 2 of [49], the perspective projection matrix M can only be determined up to a scale, so it is reasonable to fix one component of M and treat the others as unknown constant parameters. Without loss of generality, let m_{34} be fixed, and define the unknown parameter vector as

$$\phi = [m_{11}\ m_{12}\ m_{13}\ m_{14}\ m_{21}\ m_{22}\ m_{23}\ m_{24}\ m_{31}\ m_{32}\ m_{33}]^{\mathrm{T}} \in \mathbb{R}^{11} \,.$$

To design the adaptive law, the estimated projection error of the feature point is defined as

$$\begin{aligned} e(t_j,t) &= p(t_j)\hat{m}_3^{\mathrm{T}}(t)\bar{x}(t_j) - \hat{M}(t)\bar{x}(t_j) \\ &= W\left(\bar{x}(t_j), p(t_j)\right)\tilde{\phi} \end{aligned} \tag{79}$$

at time instant t_j. In Eq. (79), $W\left(\bar{x}(t_j), p(t_j)\right) \in \mathbb{R}^{3\times 11}$ is a constant regression matrix and the estimation error $\tilde{\phi}(t) \in \mathbb{R}^{11}$ is defined as $\tilde{\phi}(t) = \phi - \hat{\phi}(t)$ with $\hat{\phi}(t) \in \mathbb{R}^{11}$ equal to the estimation of ϕ. The controller can then be designed as [49]

$$\tau = g(q) - K_1\dot{q} - J^{\mathrm{T}}(q)\left(\hat{M}^{\mathrm{T}} - \hat{m}_3 p^{\mathrm{T}} + \frac{1}{2}\hat{m}_3 e_i\right)Be_i \,, \tag{80}$$

where $K_1 \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{3\times 3}$ are positive-definite gain matrices. The adaptive law is designed as

$$\begin{aligned} \dot{\hat{\phi}} = -\Gamma^{-1}\Big(&Y^{\mathrm{T}}(q,p)\dot{q} \\ &+ \Sigma_{j=1}^{m} W^{\mathrm{T}}\left(\bar{x}(t_j), p(t_j)\right) K_3 e(t_j,t)\Big) \,, \end{aligned} \tag{81}$$

where $\Gamma \in \mathbb{R}^{11\times 11}$ and $K_3 \in \mathbb{R}^{3\times 3}$ are positive-definite gain matrices, and the regression matrix $Y(q,p)$ is determined as

$$\begin{aligned} -J^{\mathrm{T}}(q)\Big[&\left(\hat{M}^{\mathrm{T}} - \hat{m}_3 p^{\mathrm{T}} - M^{\mathrm{T}} + m_3 p^{\mathrm{T}}\right) \\ &+ \frac{1}{2}\left(\hat{m}_3 - m_3\right) e_i^{\mathrm{T}}\Big]Be_i = Y(q,p)\tilde{\phi} \,. \end{aligned}$$

As stated in [49], if the five positions (corresponding to five time instant t_j, $j = 1, 2, \ldots, 5$) of the feature point for the parameter adaptation are so selected that it is impossible to find three collinear projections on the image plane, under the control of the proposed controller Eq. (80) and the adaptive algorithm Eq. (81) for parameter estimation, then the image error of the feature point is convergent to zero, i. e., $\|e_i(t)\| \to 0$ as $t \to \infty$ (see [49] for the closed-loop stability analysis).

Adaptive Homography–Based Visual Servo Control Approach

Problem Statement and Assumptions In this section, a combined high-gain control approach and an adaptive control strategy are used to regulate the robot end-effector to a desired pose asymptotically. A homography-based visual servo control approach is used to address the six-DOF regulation problem. A high-gain controller is developed to asymptotically stabilize the rotation error system by exploiting the upper triangular form of the rotation error system and the fact that the diagonal elements of the camera calibration matrix are positive. The translation error system can be linearly parameterized in terms of some unknown constant parameters which are determined by the unknown depth information and the camera calibration parameters. An adaptive controller is developed to stabilize the translation error by compensating for the unknown depth information and intrinsic camera calibration parameters. A Lyapunov-based analysis is used to examine the stability of the developed controller.

The following development is based on two mild assumptions.

Assumption 1: The bounds of a_{11} and a_{22} (defined in Subsect. "Geometric Relationships") are assumed to be known as

$$\underline{\zeta}_{a_{11}} < a_{11} < \overline{\zeta}_{a_{11}} \,, \quad \underline{\zeta}_{a_{22}} < a_{22} < \overline{\zeta}_{a_{22}} \,. \tag{82}$$

The absolute values of a_{12}, a_{13}, and a_{23} are upper bounded as

$$|a_{12}| < \overline{\zeta}_{a_{12}}\,, \quad |a_{13}| < \overline{\zeta}_{a_{13}}\,, \quad |a_{23}| < \overline{\zeta}_{a_{23}}\,. \tag{83}$$

In (82) and (83), $\underline{\zeta}_{a_{11}}, \overline{\zeta}_{a_{11}}, \underline{\zeta}_{a_{22}}, \overline{\zeta}_{a_{22}}, \overline{\zeta}_{a_{12}}, \overline{\zeta}_{a_{13}}$ and $\overline{\zeta}_{a_{23}}$ are positive constants.

Assumption 2: The reference plane is within the camera field-of-view and not at infinity. That is, there exist positive constants $\underline{\zeta}_{z_i}$ and $\overline{\zeta}_{z_i}$ such that

$$\underline{\zeta}_{z_i} < z_i(t) < \overline{\zeta}_{z_i}. \tag{84}$$

Euclidean Reconstruction Using Vanishing Points Based on Eqs. (5)–(7), the homography relationship based on measurable pixel coordinates is

$$p_i = \alpha_i A H A^{-1} p_i^*\,. \tag{85}$$

Since A is unknown, standard homography computation and decomposition algorithms can't be applied to extract the rotation and translation from the homography. If some additional information can be applied (e. g., four vanishing points), the rotation matrix can be obtained. For the vanishing points, $d^* = \infty$, so that

$$H = R + \frac{x_f}{d^*} n^{*T} = R. \tag{86}$$

Based on Eq. (86), the relationship in Eq. (85) can be expressed as

$$p_i = \alpha_i \bar{R} p_i^*\,, \tag{87}$$

where $\bar{R}(t) \in \mathbb{R}^{3\times 3}$ is defined as

$$\bar{R} = A R A^{-1}\,. \tag{88}$$

For the four vanishing points, twelve linear equations can be obtained based on Eq. (87). After normalizing $\bar{R}(t)$ by one nonzero element (e. g., $\bar{R}_{33}(t) \in \mathbb{R}$ which is assumed to be the third row third column element of $\bar{R}(t)$ without loss of generality) twelve equations can be used to solve for twelve unknowns. The twelve unknowns are given by the eight unknown elements of the normalized $\bar{R}(t)$, denoted by $\bar{R}_n(t) \in \mathbb{R}^{3\times 3}$ defined as

$$\bar{R}_n \triangleq \frac{\bar{R}}{\bar{R}_{33}}\,, \tag{89}$$

and the four unknowns are given by $\bar{R}_{33}(t)\alpha_i(t)$. From the definition of $\bar{R}_n(t)$ in Eq. (89), the fact that

$$\det(\bar{R}) = \det(A)\det(R)\det(A^{-1}) = 1 \tag{90}$$

can be used to conclude that

$$\bar{R}_{33}^3 \det(\bar{R}_n) = 1\,, \tag{91}$$

and hence,

$$\bar{R} = \frac{\bar{R}_n}{\sqrt[3]{\det(\bar{R}_n)}}\,. \tag{92}$$

After $\bar{R}(t)$ is obtained, the original four feature points on the reference plane can be used to determine the depth ratio $\alpha_i(t)$.

Open–Loop Error System

Rotation Error System If the rotation matrix $R(t)$ introduced in (5) were known, then the corresponding unit quaternion $q(t) \triangleq \left[q_0(t) \quad q_v^{\mathrm{T}}(t)\right]^{\mathrm{T}}$ can be calculated using the numerically robust method presented in [28] and [67] based on the corresponding relationships

$$R(q) = \left(q_0^2 - q_v^{\mathrm{T}} q_v\right) I_3 + 2 q_v q_v^{\mathrm{T}} - 2 q_0 q_v^{\times}\,, \tag{93}$$

where I_3 is the 3×3 identity matrix, and the notation $q_v^{\times}(t)$ denotes the skew-symmetric form of the vector $q_v(t)$. Given $R(t)$, the quaternion $q(t)$ can also be written as

$$q_0 = \frac{1}{2}\sqrt{1 + \mathrm{tr}(R)}\,, \quad q_v = \frac{1}{2} u \sqrt{3 - \mathrm{tr}(R)}\,, \tag{94}$$

where $u(t) \in \mathbb{R}^3$ is a unit eigenvector of $R(t)$ with respect to the eigenvalue 1. The open-loop rotation error system for $q(t)$ can be obtained as [68]

$$\begin{bmatrix} \dot{q}_0 \\ \dot{q}_v \end{bmatrix} = \frac{1}{2}\begin{bmatrix} -q_v^{\mathrm{T}} \\ q_0 I_3 + q_v^{\times} \end{bmatrix} \omega_{\mathrm{c}}\,, \tag{95}$$

where $\omega_{\mathrm{c}}(t) \in \mathbb{R}^3$ defines the angular velocity of the camera expressed in $\mathcal{F}$.

The quaternion $q(t)$ given in Eq. (93)–(95) is not measurable since $R(t)$ is unknown. However, since $\bar{R}(t)$ can be determined as described in Eq. (92), the same algorithm as shown in Eq. (94) can be used to determine a corresponding measurable quaternion $\left(\bar{q}_0(t), \bar{q}_v^{\mathrm{T}}(t)\right)^{\mathrm{T}}$ as

$$\bar{q}_0 = \frac{1}{2}\sqrt{1 + tr(\bar{R})}\,, \quad \bar{q}_v = \frac{1}{2}\bar{u}\sqrt{3 - \mathrm{tr}(\bar{R})}\,, \tag{96}$$

where $\bar{u}(t) \in \mathbb{R}^3$ is a unit eigenvector of $\bar{R}(t)$ with respect to the eigenvalue 1. Based on Eq. (88), $\mathrm{tr}(\bar{R}) = \mathrm{tr}(ARA^{-1}) = \mathrm{tr}(R)$, where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. Since $R(t)$ and $\bar{R}(t)$ are similar matrices, the relationship between $\left(q_0(t), q_v^{\mathrm{T}}(t)\right)^{\mathrm{T}}$ and $\left(\bar{q}_0(t), \bar{q}_v^{\mathrm{T}}(t)\right)^{\mathrm{T}}$ can be determined as

$$\bar{q}_0 = q_0\,, \quad \bar{q}_v = \frac{\|q_v\|}{\|Aq_v\|} A q_v \triangleq \gamma A q_v\,, \tag{97}$$

where $\gamma(t) \in \mathbb{R}$ is a positive, unknown, time-varying scalar.

The square of $\gamma(t)$ is

$$\gamma^2 = \frac{q_v^{\mathrm{T}} q_v}{(Aq_v)^{\mathrm{T}} A q_v} = \frac{q_v^{\mathrm{T}} q_v}{q_v^{\mathrm{T}} A^{\mathrm{T}} A q_v} . \tag{98}$$

Since A is of full rank, the symmetric matrix $A^{\mathrm{T}}A$ is positive definite. Hence, the Rayleigh–Ritz theorem can be used to conclude that

$$\lambda_{\min}(A^{\mathrm{T}}A) q_v^{\mathrm{T}} q_v \le q_v^{\mathrm{T}} A^{\mathrm{T}} A q_v \le \lambda_{\max}(A^{\mathrm{T}}A) q_v^{\mathrm{T}} q_v . \tag{99}$$

From Eqs. (98) and (99), it can be concluded that

$$\begin{gathered} \frac{1}{\lambda_{\max}(A^{\mathrm{T}}A)} \le \gamma^2 = \frac{q_v^{\mathrm{T}} q_v}{(Aq_v)^{\mathrm{T}} A q_v} \le \frac{1}{\lambda_{\min}(A^{\mathrm{T}}A)} . \\ \sqrt{\frac{1}{\lambda_{\max}(A^{\mathrm{T}}A)}} \le \gamma \le \sqrt{\frac{1}{\lambda_{\min}(A^{\mathrm{T}}A)}} . \end{gathered} \tag{100}$$

Based on Eq. (100), there exist positive bounding constants $\underline{\zeta}_\gamma, \overline{\zeta}_\gamma \in \mathbb{R}$ that satisfy the following inequalities:

$$\underline{\zeta}_\gamma < \gamma(t) < \overline{\zeta}_\gamma . \tag{101}$$

The inverse of the relationship between $\bar{q}_v(t)$ and $q_v(t)$ in Eq. (97) can also be developed as

$$\begin{aligned} q_v &= \frac{1}{\gamma} A^{-1} \bar{q}_v \\ &= \frac{1}{\gamma} \begin{bmatrix} \frac{1}{a_{11}} \bar{q}_{v1} - \frac{a_{12}}{a_{11} a_{22}} \bar{q}_{v2} - \left(\frac{a_{13}}{a_{11}} - \frac{a_{12} a_{23}}{a_{11} a_{22}} \right) \bar{q}_{v3} \\ \frac{1}{a_{22}} \bar{q}_{v2} - \frac{a_{23}}{a_{22}} \bar{q}_{v3} \\ \bar{q}_{v3} \end{bmatrix} . \end{aligned} \tag{102}$$

Translation Error System The translation error, denoted by $e(t) \in \mathbb{R}^3$, is defined as

$$e(t) = p_e(t) - p_e^* , \tag{103}$$

where $p_e(t), p_e^* \in \mathbb{R}^3$ are defined as

$$p_e = \begin{bmatrix} u_i & v_i & -\ln(\alpha_i) \end{bmatrix}^{\mathrm{T}} , \quad p_e^* = \begin{bmatrix} u_i^* & v_i^* & 0 \end{bmatrix}^{\mathrm{T}} , \tag{104}$$

where $i \in \{1, \cdots, 4\}$. The translation error $e(t)$ is measurable since the first two elements are image coordinates, and $\alpha_i(t)$ is obtained from the homography decomposition. The open-loop translation error system can be obtained by taking the time derivative of $e(t)$ and multiplying the resulting expression by z_i^* as

$$z_i^* \dot{e} = -\alpha_i A_e v_c + z_i^* A_e \left(A^{-1} p_i \right)^{\times} \omega_c , \tag{105}$$

where $v_c(t) \in \mathbb{R}^3$ defines the linear velocity of the camera expressed in $\mathcal{F}$, and $A_e(t) \in \mathbb{R}^{3\times 3}$ is defined as

$$A_e = \begin{bmatrix} a_{11} & a_{12} & a_{13} - u_i \\ 0 & a_{22} & a_{23} - v_i \\ 0 & 0 & 1 \end{bmatrix} .$$

Rotation Controller Development and Stability Analysis Based on the relationship in Eq. (97), the open-loop error system in Eq. (95), and the subsequent stability analysis, the rotation controller is designed as

$$\begin{aligned} \omega_{c1} &= -k_{\omega 1} \bar{q}_{v1} = -(k_{\omega 11} + 2) \bar{q}_{v1} \\ \omega_{c2} &= -k_{\omega 2} \bar{q}_{v2} = -(k_{\omega 21} + k_{\omega 22} + 1) \bar{q}_{v2} \\ \omega_{c3} &= -k_{\omega 3} \bar{q}_{v3} = -(k_{\omega 31} + k_{\omega 32} + k_{\omega 33}) \bar{q}_{v3} , \end{aligned} \tag{106}$$

where $k_{\omega i} \in \mathbb{R}, i = 1, 2, 3$ and $k_{\omega ij} \in \mathbb{R}, i, j = 1, 2, 3, j < i$, are positive constants. The expressed form of the controller in Eq. (106) is motivated by the use of completing the squares in the subsequent stability analysis. In Eq. (106), the damping control gains $k_{\omega 21}, k_{\omega 31}, k_{\omega 32}$ are selected according to the following sufficient conditions to facilitate the subsequent stability analysis

$$\begin{aligned} k_{\omega 21} &> \frac{1}{4} k_{\omega 1}^2 \frac{\overline{\zeta}_{a_{12}}^2}{\underline{\zeta}_{a_{11}} \underline{\zeta}_{a_{22}}} , \\ k_{\omega 31} &> \frac{1}{4} k_{\omega 1}^2 \frac{1}{\underline{\zeta}_{a_{11}}} \left(\frac{\overline{\zeta}_{a_{12}} \overline{\zeta}_{a_{23}}}{\underline{\zeta}_{a_{22}}} + \overline{\zeta}_{a_{13}} \right)^2 , \\ k_{\omega 32} &> \frac{1}{4} k_{\omega 2}^2 \frac{\overline{\zeta}_{a_{23}}^2}{\underline{\zeta}_{a_{22}}} , \end{aligned} \tag{107}$$

where $\underline{\zeta}_{a_{11}}, \overline{\zeta}_{a_{11}}, \underline{\zeta}_{a_{22}}, \overline{\zeta}_{a_{22}}, \overline{\zeta}_{a_{12}}, \overline{\zeta}_{a_{13}}$ and $\overline{\zeta}_{a_{23}}$ are defined in Eqs. (82) and (83), and $k_{\omega 11}, k_{\omega 22}, k_{\omega 33}$ are feedback gains that can be selected to adjust the performance of the rotation control system.

Provided the sufficient gain conditions given in Eq. (107) are satisfied, the controller in Eq. (106) ensures asymptotic regulation of the rotation error in the sense that

$$\|q_v(t)\| \to 0 , \quad \text{as} \quad t \to \infty . \tag{108}$$

To prove the stability result, a Lyapunov function candidate $V_1(t) \in \mathbb{R}$ can be defined as

$$V_1 \triangleq q_v^{\mathrm{T}} q_v + (1 - q_0)^2 . \tag{109}$$

Based on the open-loop error system in Eq. (95), the time-derivative of $V_1(t)$ can be determined as

$$\begin{aligned} \dot{V}_1 &= 2 q_v^{\mathrm{T}} \dot{q}_v - 2(1 - q_0) \dot{q}_0 \\ &= q_v^{\mathrm{T}} \omega_c = q_{v1} \omega_{c1} + q_{v2} \omega_{c2} + q_{v3} \omega_{c3} . \end{aligned} \tag{110}$$

After substituting Eq. (102) for $q_v(t)$, substituting Eq. (106) for $\omega_c(t)$, and completing the squares, the expression in Eq. (110) can be written as

$$\dot{V}_1 < -\frac{1}{\bar{\zeta}_\gamma}\left(k_{\omega 11}\frac{1}{a_{11}}\tilde{q}_{v1}^2 + k_{\omega 22}\frac{1}{a_{22}}\tilde{q}_{v2}^2 + k_{\omega 33}\tilde{q}_{v3}^2\right) . \tag{111}$$

Barbalat's lemma [62] can now be used to conclude that $\|q_v(t)\| \to 0$ as $t \to \infty$.

Translation Control Development and Stability Analysis After some algebraic manipulation, the translation error system in Eq. (105) can be rewritten as

$$\begin{aligned} \frac{z_i^*}{a_{11}}\dot{e}_1 &= -\alpha_i v_{c1} + Y_1(\alpha_i, u_i, v_i, \omega_c, v_{c2}, v_{c3})\phi_1 \\ \frac{z_i^*}{a_{22}}\dot{e}_2 &= -\alpha_i v_{c2} + Y_2(\alpha_i, u_i, v_i, \omega_c, v_{c3})\phi_2 \\ z_i^*\dot{e}_3 &= -\alpha_i v_{c3} + Y_3(\alpha_i, u_i, v_i, \omega_c)\phi_3 , \end{aligned} \tag{112}$$

where $\phi_1 \in \mathbb{R}^{n_1}$, $\phi_2 \in \mathbb{R}^{n_2}$, and $\phi_3 \in \mathbb{R}^{n_3}$ are vectors of constant unknown parameters, and $Y_1(\cdot) \in \mathbb{R}^{1\times n_1}$, $Y_2(\cdot) \in \mathbb{R}^{1\times n_2}$, and $Y_3(\cdot) \in \mathbb{R}^{1\times n_3}$ are known regressor vectors.

The control strategy used in [52] is to design $v_{c3}(t)$ to stabilize $e_3(t)$, and then design $v_{c2}(t)$ to stabilize $e_2(t)$ given $v_{c3}(t)$, and then design $v_{c1}(t)$ to stabilize $e_1(t)$ given $v_{c3}(t)$ and $v_{c2}(t)$. Following this design strategy, the translation controller $v_c(t)$ is designed as [52]

$$\begin{aligned} v_{c3} &= \frac{1}{\alpha_i}(k_{v3}e_3 + Y_3(\alpha_i, u_i, v_i, \omega_c)\hat{\phi}_3) \\ v_{c2} &= \frac{1}{\alpha_i}(k_{v2}e_2 + Y_2(\alpha_i, u_i, v_i, \omega_c, v_{c3})\hat{\phi}_2) \\ v_{c1} &= \frac{1}{\alpha_i}(k_{v1}e_1 + Y_1(\alpha_i, u_i, v_i, \omega_c, v_{c2}, v_{c3})\hat{\phi}_1) , \end{aligned} \tag{113}$$

where the depth ratio $\alpha_i(t) > 0 \forall t$. In (113), $\hat{\phi}_1(t) \in \mathbb{R}^{n_1}, \hat{\phi}_2(t) \in \mathbb{R}^{n_2}, \hat{\phi}_3(t) \in \mathbb{R}^{n_3}$ denote adaptive estimates that are designed according to the following adaptive update laws to cancel the respective terms in the subsequent stability analysis

$$\dot{\hat{\phi}}_1 = \Gamma_1 Y_1^{\mathrm{T}} e_1 , \quad \dot{\hat{\phi}}_2 = \Gamma_2 Y_2^{\mathrm{T}} e_2 , \quad \dot{\hat{\phi}}_3 = \Gamma_3 Y_3^{\mathrm{T}} e_3 , \tag{114}$$

where $\Gamma_1 \in \mathbb{R}^{n_1\times n_1}, \Gamma_2 \in \mathbb{R}^{n_2\times n_2}, \Gamma_3 \in \mathbb{R}^{n_3\times n_3}$ are diagonal matrices of positive constant adaptation gains. Based on Eqs. (112) and (113), the closed-loop translation error system is

$$\begin{aligned} \frac{z_i^*}{a_{11}}\dot{e}_1 &= -k_{v1}e_1 + Y_1(\alpha_i, u_i, v_i, \omega_c, v_{c2}, v_{c3})\tilde{\phi}_1 \\ \frac{z_i^*}{a_{22}}\dot{e}_2 &= -k_{v2}e_2 + Y_2(\alpha_i, u_i, v_i, \omega_c, v_{c3})\tilde{\phi}_2 \\ z_i^*\dot{e}_3 &= -k_{v3}e_3 + Y_3(\alpha_i, u_i, v_i, \omega_c)\tilde{\phi}_3 , \end{aligned} \tag{115}$$

where $\tilde{\phi}_1(t) \in \mathbb{R}^{n_1}, \tilde{\phi}_2(t) \in \mathbb{R}^{n_2}, \tilde{\phi}_3(t) \in \mathbb{R}^{n_3}$ denote the intrinsic calibration parameter mismatch defined as

$$\tilde{\phi}_1(t) = \phi_1 - \hat{\phi}_1(t) , \quad \tilde{\phi}_2(t) = \phi_2 - \hat{\phi}_2(t) , \quad \tilde{\phi}_3(t) = \phi_3 - \hat{\phi}_3(t) .$$

The controller given in Eq. (113) along with the adaptive update law in Eq. (114) ensures asymptotic regulation of the translation error system in the sense that

$$\|e(t)\| \to 0 \quad \text{as} \quad t \to \infty .$$

To prove the stability result, a Lyapunov function candidate $V_2(t) \in \mathbb{R}$ can be defined as

$$\begin{aligned} V_2 = &\frac{1}{2}\frac{z_i^*}{a_{11}}e_1^2 + \frac{1}{2}\frac{z_i^*}{a_{22}}e_2^2 + \frac{1}{2}z_i^* e_3^2 + \frac{1}{2}\tilde{\phi}_1^{\mathrm{T}}\Gamma_1^{-1}\tilde{\phi}_1 \\ &+ \frac{1}{2}\tilde{\phi}_2^{\mathrm{T}}\Gamma_2^{-1}\tilde{\phi}_2 + \frac{1}{2}\tilde{\phi}_3^{\mathrm{T}}\Gamma_3^{-1}\tilde{\phi}_3 . \end{aligned} \tag{116}$$

After taking the time derivative of Eq. (116) and substituting for the closed-loop error system developed in Eq. (115), the following simplified expression can be obtained:

$$\dot{V}_2 = -k_{v1}e_1^2 - k_{v2}e_2^2 - k_{v3}e_3^2 . \tag{117}$$

Barbalat's lemma [62] can now be used to show that $e_1(t), e_2(t), e_3(t) \to 0$ as $t \to \infty$.

The controller given in Eqs. (106) and (113) along with the adaptive update law in Eq. (114) ensures asymptotic translation and rotation regulation in the sense that

$$\|q_v(t)\| \to 0 \quad \text{and} \quad \|e(t)\| \to 0 \quad \text{as} \quad t \to \infty ,$$

provided the control gains satisfy the sufficient conditions given in Eq. (107). Based on the controller in Eqs. (106) and (113) and the adaptive law in Eq. (114), the resulting asymptotic translation and rotation errors are plotted in Figs. 10 and 11, respectively. The image-space trajectory is shown in Fig. 12, and it is also shown in Fig. 13 in a three dimensional format.

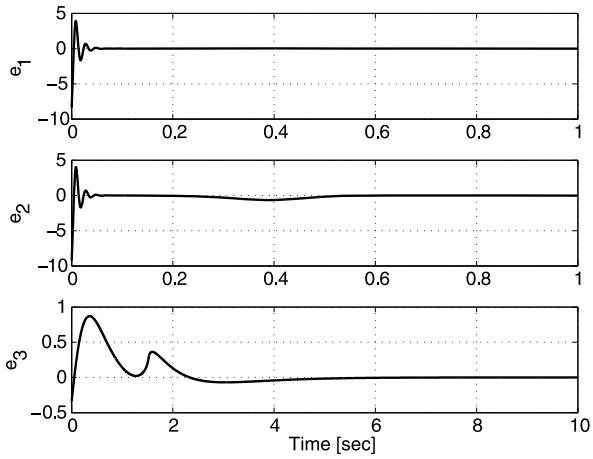

Adaptive Visual Servo Control, Figure 10
Unitless translation error $e(t)$

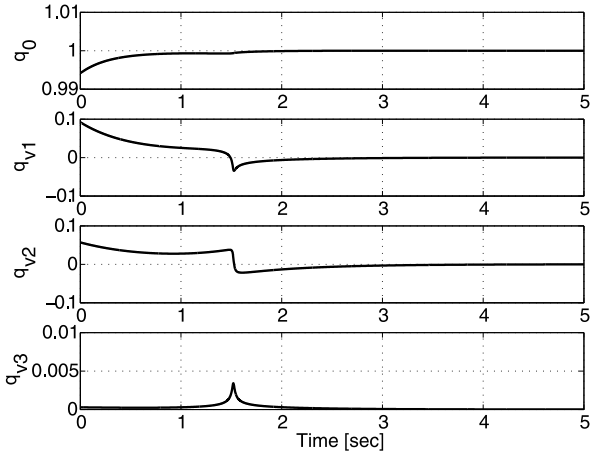

Adaptive Visual Servo Control, Figure 11
Quaternion rotation error $q(t)$

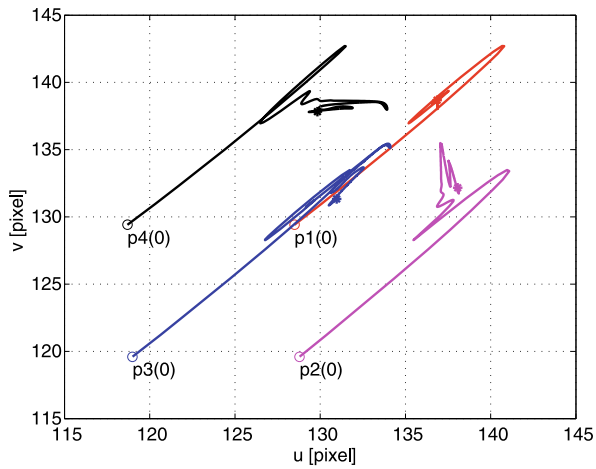

Adaptive Visual Servo Control, Figure 12
Image-space error in pixels between $p(t)$ and p^*. In the figure, "O" denotes the initial positions of the 4 feature points in the image, and "*" denotes the corresponding final positions of the feature points

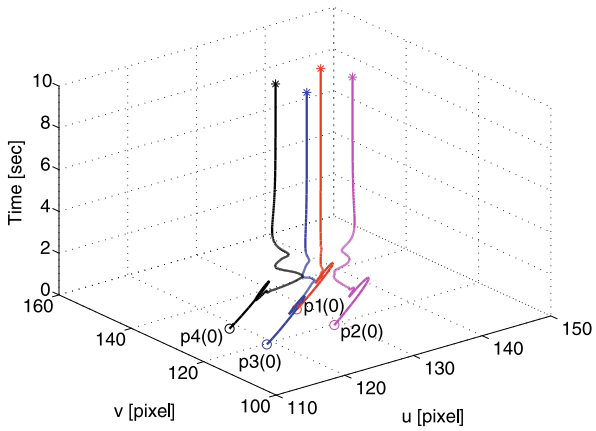

Adaptive Visual Servo Control, Figure 13
Image-space error in pixels between $p(t)$ and p^* shown in a 3D graph. In the figure, "O" denotes the initial positions of the 4 feature points in the image, and "*" denotes the corresponding final positions of the feature points

Future Directions

Image-based visual servo control is problematic because it may yield impossible Euclidean trajectories for the robotic system, and the image-Jacobian can contain singularities. Position-based visual servo control also has deficits in that the feedback signal does not exploit the image information directly, so the resulting image-trajectory may take the feature points out of the field-of-view. Homography-based visual servo control combines the positive aspects of the former methods; however, this approach requires the homography construction and decomposition. The homography construction is sensitive to image noise and there exists a sign ambiguity in the homography decomposition. Open problems that are areas of future research include new methods to combine image-based and reconstructed information in the feedback loop without the homography decomposition, perhaps using methods that directly servo on the homography feedback without requiring the homography decomposition.

The success of visual servo control for traditional robotic systems provides an impetus for other autonomous systems such as the navigation and control of unmanned air vehicles (UAV). For new applications, new hurdles may arise. For example, the inclusion of the dynamics in traditional robotics applications is a straight-

forward task through the use of integrator backstepping. However, when a UAV is maneuvering (including roll, pitch, yaw motions) to a desired pose, the feature points may be required to leave the field-of-view because of the unique flight dynamics. Other examples include systems that are hyper-redundant, such as robotic elephant trunks, where advanced path-planning must be incorporated with the controller to account for obstacle avoidance and maintaining a line-of-sight to a target.

Bibliography

1. Weiss LE, Sanderson AC, Neuman CP (1987) Dynamic sensor-based control of robots with visual feedback. IEEE J Robot Autom RA-3:404–417
2. Feddema J, Mitchell O (1989) Vision-guided servoing with feature-based trajectory generation. IEEE Trans Robot Autom 5:691–700
3. Hashimoto K, Kimoto T, Ebine T, Kimura H (1991) Manipulator control with image-based visual servo. Proc IEEE Int Conf Robot Autom, pp 2267–2272
4. Espiau B, Chaumette F, Rives P (1992) A new approach to visual servoing in robotics. IEEE Trans Robot Autom 8:313–326
5. Hutchinson S, Hager G, Corke P (1996) A tutorial on visual servo control. IEEE Trans Robot Autom 12:651–670
6. Chaumette F (1998) Potential problems of stability and convergence in image-based and position-based visual servoing. In: Kriegman D, Hager G, Morse A (eds), The confluence of vision and control LNCIS series, vol 237. Springer, Berlin, pp 66–78
7. Espiau B (1993) Effect of camera calibration errors on visual servoing in robotics. 3rd Int Symp. Exp Robot, pp 182–192
8. Corke P, Hutchinson S (2001) A new partitioned approach to image-based visual servo control. IEEE Trans Robot Autom 17:507–515
9. Wilson WJ, Hulls CW, Bell GS (1996) Relative end-effector control using cartesian position based visual servoing. IEEE Trans Robot Autom 12:684–696
10. Martinet P, Gallice J, Khadraoui D (1996) Vision based control law using 3D visual features. Proc World Autom Cong, pp 497–502
11. Daucher N, Dhome M, Lapresté J, Rives G (1997) Speed command of a robotic system by monocular pose estimate. Proc Int Conf Intell Robot Syst, pp 55–62
12. Hashimoto K (2003) A review on vision-based control of robot manipulators. Adv Robot 17:969–991
13. Chaumette F, Hutchinson S (2006) Visual servo control part i: Basic approaches. IEEE Robot Autom Mag 13:82–90
14. Deguchi K (1998) Optimal motion control for image-based visual servoing by decoupling translation and rotation. Proc Int Conf Intell Robot Syst, pp 705–711
15. Malis E, Chaumette F, Bodet S (1999) 2 1/2D visual servoing. IEEE Trans Robot Autom 15:238–250
16. Malis E, Chaumette F (2000) 2 1/2D visual servoing with respect to unknown objects through a new estimation scheme of camera displacement. Int J Comput Vis 37:79–97
17. Chaumette F, Malis E (2000) 2 1/2D visual servoing: a possible solution to improve image-based and position-based visual servoings. Proc IEEE Int Conf Robot Autom, pp 630–635
18. Malis E, Chaumette F (2002) Theoretical improvements in the stability analysis of a new class of model-free visual servoing methods. IEEE Trans Robot Autom 18:176–186
19. Chen J, Dawson DM, Dixon WE, Behal A (2005) Adaptive homography-based visual servo tracking for a fixed camera configuration with a camera-in-hand extension. IEEE Trans Control Syst Technol 13:814–825
20. Fang Y, Dixon WE, Dawson DM, Chawda P (2005) Homography-based visual servoing of wheeled mobile robots. IEEE Trans Syst Man Cybern – Part B: Cybern 35:1041–1050
21. Chen J, Dixon WE, Dawson DM, McIntyre M (2006) Homography-based visual servo tracking control of a wheeled mobile robot. IEEE Trans Robot 22:406–415
22. Hu G, Gupta S, Fitz-coy N, Dixon WE (2006) Lyapunov-based visual servo tracking control via a quaternion formulation. Proc IEEE Conf Decision Contr, pp 3861–3866
23. Gans N, Hutchinson S (2007) Stable visual servoing through hybrid switched-system control. IEEE Trans Robot 23:530–540
24. Conticelli F, Allotta B (2001) Nonlinear controllability and stability analysis of adaptive image-based systems. IEEE Trans Robot Autom 17:208–214
25. Conticelli F, Allota B (2001) Discrete-time robot visual feedback in 3-D positioning tasks with depth adaptation. IEEE/ASME Trans Mechatron 6:356–363
26. Fang Y, Behal A, Dixon WE, Dawson DM (2002) Adaptive 2.5D visual servoing of kinematically redundant robot manipulators. Proc IEEE Conf Decision Contr, pp 2860–2865
27. Fang Y, Dawson DM, Dixon WE, de Queiroz MS (2002) 2.5D visual servoing of wheeled mobile robots. Proc IEEE Conf Decision Contr, pp 2866–2871
28. Hu G, Dixon WE, Gupta S, Fitz-coy N (2006) A quaternion formulation for homography-based visual servo control. Proc IEEE Int Conf Robot Autom, pp 2391–2396
29. Tsai RY (1987) A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf tv cameras and lenses. IEEE J Robot Autom 3:323–344
30. Tsai RY (1989) Synopsis of recent progress on camera calibration for 3D machine vision. MIT Press, Cambridge
31. Robert L (1996) Camera calibration without feature extraction. Comput Vis Image Underst 63:314–325
32. Heikkila J, Silven O (1997) A four-step camera calibration procedure with implicit image correction. Proc IEEE Conf Comput Vis Pattern Recogn, pp 1106–1112
33. Clarke TA, Fryer JG (1998) The development of camera calibration methods and models. Photogrammetric Rec 16:51–66
34. Sturm PF, Maybank SJ (1999) On plane-based camera calibration: A general algorithm, singularities, applications. Proc IEEE Conf Comput Vis Pattern Recogn, pp 432–437
35. Zhang Z (1999) Flexible camera calibration by viewing a plane from unknown orientations. Proc IEEE Int Conf Comput Vis, pp 666–673
36. Hosoda K, Asada M (1994) Versatile visual servoing without knowledge of true jacobian. Proc IEEE/RSJ Int Conf Intell Robot Syst, pp 186–193
37. Jagersand M, Fuentes O, Nelson R (1997) Experimental evaluation of uncalibrated visual servoing for precision manipulation. Proc Int Conf Robot Autom, pp 2874–2880
38. Shahamiri M, Jagersand M (2005) Uncalibrated visual servoing using a biased newton method for on-line singularity detection and avoidance. Proc IEEE/RSJ Int Conf Intell Robot Syst, pp 3953–3958

39. Piepmeier JA, Lipkin H (2003) Uncalibrated eye-in-hand visual servoing. Int J Robot Res 22:805–819
40. Piepmeier JA, McMurray GV, Lipkin H (2004) Uncalibrated dynamic visual servoing. IEEE Trans Robot Autom 24:143–147
41. Kelly R (1996) Robust asymptotically stable visual servoing of planar manipulator. IEEE Trans Robot Autom 12:759–766
42. Bishop B, Spong MW (1997) Adaptive calibration and control of 2D monocular visual servo system. Proc IFAC Symp Robot Contr, pp 525–530
43. Hsu L, Aquino PLS (1999) Adaptive visual tracking with uncertain manipulator dynamics and uncalibrated camera. Proc IEEE Conf Decision Contr, pp 1248–1253
44. Taylor CJ, Ostrowski JP (2000) Robust vision-based pose control. Proc IEEE Int Conf Robot Autom, pp 2734–2740
45. Zergeroglu E, Dawson DM, de Queiroz M, Behal A (2001) Vision-based nonlinear tracking controllers in the presence of parametric uncertainty. IEEE/ASME Trans Mechatron 6:322–337
46. Dixon WE, Dawson DM, Zergeroglu E, Behal A (2001) Adaptive tracking control of a wheeled mobile robot via an uncalibrated camera system. IEEE Trans Syst Man Cybern – Part B: Cybern 31:341–352
47. Astolfi A, Hsu L, Netto M, Ortega R (2002) Two solutions to the adaptive visual servoing problem. IEEE Trans Robot Autom 18:387–392
48. Ruf A, Tonko M, Horaud R, Nagel H-H (1997) Visual tracking of an end-effector by adaptive kinematic prediction. Proc IEEE/RSJ Int Conf Intell Robot Syst, pp 893–898
49. Liu Y, Wang H, Wang C, Lam K (2006) Uncalibrated visual servoing of robots using a depth-independent interaction matrix. IEEE Trans Robot 22:804–817
50. Fang Y, Dixon WE, Dawson DM, Chen J (2003) An exponential class of model-free visual servoing controllers in the presence of uncertain camera calibration. Proc IEEE Conf Decision Contr, pp 5390–5395
51. Hu G, Gans N, Dixon WE (2007) Quaternion-based visual servo control in the presence of camera calibration error. Proc IEEE Multi-Conf Syst Contr, pp 1492–1497
52. Chen J, Behal A, Dawson DM, Dixon WE (2003) Adaptive visual servoing in the presence of intrinsic calibration uncertainty. Proc IEEE Conf Decision Contr, pp 5396–5401
53. Faugeras O (1993) Three-dimensional computer vision: A geometric viewpoint. MIT Press, Cambridge
54. Shi J, Tomasi C (1994) Good features to track. Proc IEEE Conf Comput Vis Pattern Recogn, pp 593–600
55. Tomasi C, Kanade T (1991) Detection and tracking of point features. Technical report. Carnegie Mellon University, Pittsburgh
56. Boufama B, Mohr R (1995) Epipole and fundamental matrix estimation using virtual parallax. Proc IEEE Int Conf Comput Vis, pp (1030)–1036
57. Hartley R, Zisserman A (2000) Multiple view geometry in computer vision. Cambridge University Press, New York
58. Faugeras O, Lustman F (1988) Motion and structure from motion in a piecewise planar environment. Int J Pattern Recogn Artif Intell 2:485–508
59. Zhang Z, Hanson AR (1995) Scaled euclidean 3D reconstruction based on externally uncalibrated cameras. IEEE Symp Comput Vis, pp 37–42
60. DeMenthon D, Davis L (1995) Model-based object pose in 25 lines of code. Int J Comput Vis 15:123–141
61. Quan L, Lan Z-D (1999) Linear *n*-point camera pose determination. IEEE Trans Pattern Anal Machine Intell 21:774–780
62. Slotine JJ, Li W (1991) Applied nonlinear control. Prentice Hall, Inc., Englewood Cliff
63. Zachi A, Hsu L, Ortega R, Lizarralde F (2006) Dynamic control of uncertain manipulators through immersion and invariance adaptive visual servoing. Int J Robot Res 25:1149–1159
64. Almansa A, Desolneux A, Vamech S (2003) Vanishing point detection without any a priori information. IEEE Trans Pattern Analysis and Machine Intelligence 25:502–507
65. Liu Y, Wang H, Lam K (2005) Dynamic visual servoing of robots in uncalibrated environments. Proc IEEE Int Conf Robot Autom, pp 3142–3148
66. Liu Y, Wang H, Zhou D (2006) Dynamic tracking of manipulators using visual feedback from an uncalibrated fixed camera. Proc IEEE Int Conf Robot Autom, pp 4124–4129
67. Shuster M (1993) A survey of attitude representations. J Astron Sci 41:439–518
68. Dixon WE, Behal A, Dawson DM, Nagarkatti S (2003) Nonlinear control of engineering systems: A Lyapunov–based approach. Birkhäuser, Boston

Additive Cellular Automata

Burton Voorhees
Center for Science, Athabasca University, Athabasca, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
Notation and Formal Definitions
Additive Cellular Automata in One Dimension
d-Dimensional Rules
Future Directions
Bibliography

Glossary

Cellular automata Cellular automata are dynamical systems that are discrete in space, time, and value. A state of a cellular automaton is a spatial array of discrete cells, each containing a value chosen from a finite alphabet. The state space for a cellular automaton is the set of all such configurations.

Alphabet of a cellular automaton The alphabet of a cellular automaton is the set of symbols or values that can appear in each cell. The alphabet contains a distinguished symbol called the null or quiescent symbol, usually indicated by 0, which satisfies the condition of an additive identity: $0 + x = x$.

Cellular automata rule The rule, or update rule of a cellular automaton describes how any given state is transformed into its successor state. The update rule of a cellular automaton is described by a rule table, which defines a local neighborhood mapping, or equivalently as a global update mapping.

Additive cellular automata An additive cellular automaton is a cellular automaton whose update rule satisfies the condition that its action on the sum of two states is equal to the sum of its actions on the two states separately.

Linear cellular automata A linear cellular automaton is a cellular automaton whose update rule satisfies the condition that its action on the sum of two states separately equals action on the sum of the two states plus its action on the state in which all cells contain the quiescent symbol. Note that some researchers reverse the definitions of additivity and linearity.

Neighborhood The neighborhood of a given cell is the set of cells that contribute to the update of value in that cell under the specified update rule.

Rule table The rule table of a cellular automaton is a listing of all neighborhoods together with the symbol that each neighborhood maps to under the local update rule.

Local maps of a cellular automaton The local mapping for a cellular automaton is a map from the set of all neighborhoods of a cell to the automaton alphabet.

State transition diagram The state transition diagram (STD) of a cellular automaton is a directed graph with each vertex labeled by a possible state and an edge directed from a vertex x to a vertex y if and only if the state labeling vertex x maps to the state labeling vertex y under application of the automaton update rule.

Transient states A transient state of a cellular automaton is a state that can at most appear only once in the evolution of the automaton rule.

Cyclic states A cyclic state of a cellular automaton is a state lying on a cycle of the automaton update rule, hence it is periodically revisited in the evolution of the rule.

Basins of attraction The basins of attraction of a cellular automaton are the equivalences classes of cyclic states together with their associated transient states, with two states being equivalent if they lie on the same cycle of the update rule.

Predecessor state A state x is the predecessor of a state y if and only if x maps to y under application of the cellular automaton update rule. More specifically, a state x is an nth order predecessor of a state y if it maps to y under n applications of the update rule.

Garden-of-Eden A Garden-of-Eden state is a state that has no predecessor. It can be present only as an initial condition.

Surjectivity A mapping is surjective (or onto) if every state has a predecessor.

Injectivity A mapping is injective (one-to-one) if every state in its domain maps to a unique state in its range. That is, if states x and y both map to a state z then $x = y$.

Reversibility A mapping $\mathcal{X}$ is reversible if and only if a second mapping $\mathcal{X}^{-1}$ exists such that if $\mathcal{X}(x) = y$ then $\mathcal{X}^{-1}(y) = x$. For finite state spaces reversibility and injectivity are identical.

Definition of the Subject

Cellular automata are discrete dynamical systems in which an extended array of symbols from a finite alphabet is iteratively updated according to a specified local rule. Originally developed by John von Neumann [1,2] in 1948, following suggestions from Stanislaw Ulam, for the purpose of showing that self-replicating automata could be constructed. Von Neumann's construction followed a complicated set of reproduction rules but later work showed that self-reproducing automata could be constructed with only simple update rules, e. g. [3]. More generally, cellular automata are of interest because they show that highly complex patterns can arise from the application of very simple update rules. While conceptually simple, they provide a robust modeling class for application in a variety of disciplines, e. g. [4], as well as fertile grounds for theoretical research. Additive cellular automata are the simplest class of cellular automata. They have been extensively studied from both theoretical and practical perspectives.

Introduction

A wide variety of cellular automata applications, in a number of differing disciplines, has appeared in the past fifty years, see, e. g. [5,6,7,8]. Among other things, cellular automata have been used to model growth and aggregation processes [9,10,11,12]; discrete reaction-diffusion systems [13,14,15,16,17]; spin exchange systems [18,19]; biological pattern formation [20,21]; disease processes and transmission [22,23,24,25,26]; DNA sequences, and gene interactions [27,28]; spiral galaxies [29]; social interaction networks [30]; and forest fires [31,32]. They have been used for language and pattern recognition [33,34,35,36,37,38,39]; image processing [40,41]; as parallel computers [42,43,44,45,46,47]; parallel multipliers [48]; sorters [49]; and prime number sieves [50].

In recent years, cellular automata have become important for VLSI logic circuit design [51]. Circuit designers need "simple, regular, modular, and cascadable logic circuit structure to realize a complex function" and cellular automata, which show a significant advantage over linear feedback shift registers, the traditional circuit building block, satisfy this need (see [52] for an extensive survey). Cellular automata, in particular additive cellular automata, are of value for producing high-quality pseudorandom sequences [53,54,55,56,57]; for pseudoexhaustive and deterministic pattern generation [58,59,60,61,62,63]; for signature analysis [64,65,66]; error correcting codes [67,68]; pseudoassociative memory [69]; and cryptography [70].

In this discussion, attention focuses on the subclass of additive cellular automata. These are the simplest cellular automata, characterized by the property that the action of the update rule on the sum of two states is equal to the sum of the rule acting on each state separately. Hybrid additive rules (i. e., with different cells evolving according to different additive rules) have proved particularly useful for generation of pseudorandom and pseudoexhaustive sequences, signature analysis, and other circuit design applications, e. g. [52,71,72].

The remainder of this article is organized as follows: Sect. "Notation and Formal Definitions" introduces definitions and notational conventions. In Sect. "Additive Cellular Automata in One Dimension", consideration is restricted to one-dimensional rules. The influence of boundary conditions on the evolution of one-dimensional rules, conditions for rule additivity, generation of fractal space-time outputs, equivalent forms of rule representation, injectivity and reversibility, transient lengths, and cycle periods are discussed using several approaches. Taking $\mathcal{X}$ as the global operator for an additive cellular automata, a method for analytic solution of equations of the form $\mathcal{X}(\mu) = \beta$ is described. Section "*d*-Dimensional Rules" describes work on d-dimensional rules defined on tori. The discrete baker transformation is defined and used to generalize one-dimensional results on transient lengths, cycle periods, and similarity of state transition diagrams. Extensive references to the literature are provided throughout and a set of general references is provided at the end of the bibliography.

Notation and Formal Definitions

Let $S(\mathcal{L}) = \{s_i\}$ be the set of lattice sites of a d-dimensional lattice $\mathcal{L}$ with n_r equal to the number of lattice sites on dimension r. Denote by $\mathcal{A}$ a finite symbols set with $|\mathcal{A}| = p$ (usually prime). An $\mathcal{A}$-configuration on $\mathcal{L}$ is a surjective map $v : \mathcal{A} \mapsto S(\mathcal{L})$ that assigns a symbol from $\mathcal{A}$ to each site in $S(\mathcal{L})$. In this way, every $\mathcal{A}$-configuration defines a size $n_1 \times \cdots \times n_d$, d-dimensional matrix μ of symbols drawn from $\mathcal{A}$. Denote the set of all $\mathcal{A}$-configurations on $\mathcal{L}$ by $\mathcal{E}(\mathcal{A}, \mathcal{L})$.

Each $s_i \in S(\mathcal{L})$ is labeled by an integer vector $\vec{i} = (i_1, \ldots, i_d)$ where i_r is the number of sites along the rth dimension separating s_i from the assigned origin in $\mathcal{L}$. The shift operator on the rth dimension of $\mathcal{L}$ is the map $\sigma_r : \mathcal{L} \mapsto \mathcal{L}$ defined by

$$\sigma_r(s_i) = s_j \,, \quad \vec{j} = (i_1, \ldots, i_r - 1, \ldots, i_d) \,. \tag{1}$$

Equivalently, the shift maps the value at site $\vec{i}$ to the value at site $\vec{j}$.

Let $\mu(s_i; t) = \mu(i_1, \ldots, i_d; t) \in \mathcal{A}$ be the entry of μ corresponding to site s_i at iteration t for any discrete dynamical system having $\mathcal{E}(\mathcal{A}, \mathcal{L})$ as state space. Given a finite set of integer d-tuples $\mathcal{N} = \{(k_1, \ldots, k_d)\}$, define the $\mathcal{N}$-neighborhood of a site $s_i \in S(\mathcal{L})$ as

$$N(s_i) = \left\{ s_j \,\middle|\, \vec{j} = \vec{i} + \vec{k}, \vec{k} \in \mathcal{N} \right\} \tag{2}$$

A neighborhood configuration is a surjective map $\underline{v} : \mathcal{A} \mapsto N(s_0)$. Denote the set of all neighborhood configurations by $\mathcal{E}_{\mathcal{N}}(\mathcal{A})$.

The rule table for a cellular automata acting on the state space $\mathcal{E}(\mathcal{A}, \mathcal{L})$ with standard neighborhood $N(s_0)$ is defined by a map $x\colon \mathcal{E}_{\mathcal{N}}(\mathcal{A}) \mapsto \mathcal{A}$ (note that this map need not be surjective or injective). The value of x for a given neighborhood configuration is called the (value of the) rule component of that configuration. The map $x : \mathcal{E}_{\mathcal{N}}(\mathcal{A}) \mapsto \mathcal{A}$ induces a global map $\mathcal{X} : \mathcal{E}(\mathcal{A}, \mathcal{L}) \mapsto \mathcal{E}(\mathcal{A}, \mathcal{L})$ as follows: For any given element $\mu(t) \in \mathcal{E}(\mathcal{A}, \mathcal{L})$, the set $C(s_i) = \left\{ \mu(s_j; t) \,\middle|\, s_j \in N(s_i) \right\}$ is a neighborhood configuration for the site s_i, hence the map $\mu(s_i; t) \mapsto x(C(s_i))$ for all s_i produces a new symbol $\mu(s_i; t + 1)$. The site s_i is called the mapping site. When taken over all mapping sites, this produces a matrix $\mu(t + 1)$ that is the representation of $\mathcal{X}(\mu(t))$. A cellular automaton is indicated by reference to its rule table or to the global map defined by this rule table.

A cellular automaton with global map $\mathcal{X}$ is additive if and only if, for all pairs of states μ and β,

$$\mathcal{X}(\mu + \beta) = \mathcal{X}(\mu) + \mathcal{X}(\beta) \tag{3}$$

Addition of states is carried out site-wise mod(p) on the matrix representations of μ and β; for example, for a one-dimensional six-site lattice with $p = 3$ the sum of 120112 and 021212 is 111021.

The definition for additivity given in [52] differs slightly from this standard definition. There, a binary valued cellular automaton is called "linear" if its local rule

only involves the XOR operation and "additive" if it involves XOR and/or XNOR. A rule involving XNOR can be written as the binary complement of a rule involving only XOR. In terms of the global operator of the rule, this means that it has the form $\mathbf{1} + \mathcal{X}$ where $\mathcal{X}$ satisfies Eq. (3) and $\mathbf{1}$ represents the rule that maps every site to 1. Thus, $(\mathbf{1} + \mathcal{X})(\mu + \beta)$ equals $1 \ldots 1 + \mathcal{X}(\mu + \beta)$ while

$$\begin{aligned}(\mathbf{1} + \mathcal{X})(\mu) &+ (\mathbf{1} + \mathcal{X})(\beta) \\ &= 1 \ldots 1 + 1 \ldots 1 + \mathcal{X}(\mu) + \mathcal{X}(\beta) \\ &= \mathcal{X}(\mu) + \mathcal{X}(\beta) \bmod(2) \,.\end{aligned}$$

In what follows, an additive rule is defined strictly as one obeying Eq. (3), corresponding to rules that are "linear" in [52].

Much of the formal study of cellular automata has focused on the properties and forms of representation of the map $\mathcal{X} : \mathcal{E}(\mathcal{A}, \mathcal{L}) \mapsto \mathcal{E}(\mathcal{A}, \mathcal{L})$. The structure of the state transition diagram (STD($\mathcal{X}$)) of this map is of particular interest.

Example 1 (Continuous Transformations of the Shift Dynamical System) Let $\mathcal{L}$ be isomorphic to the set of integers $\mathcal{Z}$. Then $\mathcal{E}(\mathcal{A}, \mathcal{Z})$ is the set of infinite sequences with entries from $\mathcal{A}$. With the product topology induced by the discrete topology on $\mathcal{A}$ and σ as the left shift map, the system $(\mathcal{E}(\mathcal{A}, \mathcal{Z}), \sigma)$ is the shift dynamical system on $\mathcal{A}$. The set of cellular automata maps $\mathcal{X} : \mathcal{E}(\mathcal{A}, \mathcal{Z}) \mapsto \mathcal{E}(\mathcal{A}, \mathcal{Z})$ constitutes the class of continuous shift-commuting transformations of $(\mathcal{E}(\mathcal{A}, \mathcal{Z}), \sigma)$, a fundamental result of Hedlund [73].

Example 2 (Elementary Cellular Automata) Let $\mathcal{L}$ be isomorphic to $\mathcal{L}$ with $\mathcal{A} = \{0, 1\}$ and $\mathcal{N} = \{-1, 0, 1\}$. The neighborhood of site s_i is $\{s_{i-1}, s_i, s_{i+1}\}$ and $\mathcal{E}_{\mathcal{N}}(\mathcal{A}) = \{000, 001, 010, 011, 100, 101, 110, 111\}$. In this one-dimensional case, the rule table can be written as $x_i = x(i_0 i_1 i_2)$ where $i_0 i_1 i_2$ is the binary form of the index i. Listing this gives the standard form for the rule table of an elementary cellular automata.

000	001	010	011	100	101	110	111
x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7

The standard labeling scheme for elementary cellular automata was introduced by Wolfram [74], who observed that the rule table for elementary rules defines the binary number $\sum_{1=0}^{7} x_i 2^i$ and used this number to label the corresponding rule.

Example 3 (The Game of Life) This simple 2-dimensional cellular automata was invented by John Conway to illustrate a self-reproducing system. It was first presented in 1970 by Martin Gardner [75,76]. The game takes place on a square lattice, either infinite or toridal. The neighborhood of a cell consists of the eight cells surrounding it. The alphabet is {0,1}: a 1 in a cell indicates that cell is alive, a 0 indicates that it is dead. The update rules are: (a) If a cell contains a 0, it remains 0 unless exactly three of its neighbors contain a 1; (b) If a cell contains a 1 then it remains a 1 if and only if two or three of its neighbors are 1. This cellular automata produces a number of interesting patterns including a variety of fixed points (still life); oscillators (period 2); and moving patterns (gliders, spaceships); as well as more exotic patterns such as glider guns which generate a stream of glider patterns.

Additive Cellular Automata in One Dimension

Much of the work on cellular automata has focused on rules in one dimension ($d = 1$). This section reviews some of this work.

Boundary Conditions and Additivity

In the case of one-dimensional cellular automata, the lattice $\mathcal{L}$ can be isomorphic to the integers; to the non-negative integers; to the finite set $\{0, \ldots, n-1\} \in \mathcal{Z}$; or to the integers modulo an integer n. In the first case, there are no boundary conditions; in the remaining three cases, different boundary conditions apply. If $\mathcal{L}$ is isomorphic to $\mathcal{Z}_n$, the integers mod(n), the boundary conditions are periodic and the lattice is circular (it is a p-adic necklace). This is called a cylindrical cellular automata [77] because evolution of the rule can be represented as taking place on a cylinder. If the lattice is isomorphic to $\{0, \ldots, n-1\}$, null, or Dirchlet boundary conditions are set [78,79,80]. That is, the symbol assigned to all sites in $\mathcal{L}$ outside of this set is the null symbol. When the lattice is isomorphic to the non-negative integers $\mathcal{Z}^+$, null boundary conditions are set at the left boundary. In these latter two cases, the neighborhood structure assumed may influence the need for null conditions.

Example 4 (Elementary Rule 90) Let δ represent the global map for the elementary cellular automata rule 90, with rule table

000	001	010	011	100	101	110	111
0	1	0	1	1	0	1	0

For a binary string μ in $\mathcal{Z}$ or $\mathcal{Z}_n$ the action of rule 90 is defined by $[\delta(\mu)]_i = \mu_{i-1} + \mu_{i+1} \bmod(2)$, where all indices are taken mod(n) in the case of $\mathcal{Z}_n$. In the remaining

cases,

$$[\delta(\mu)]_i = \begin{cases} \mu_1 & i = 0 \\ \mu_{i-1} + \mu_{i+1} & 0 < i < n-1 \\ \mu_{n-2} & i = n-1 \end{cases}$$

null conditions (4)

$$[\delta(\mu)]_i = \begin{cases} \mu_1 & i = 0 \\ \mu_{i-1} + \mu_{i+1} & 0 < i \end{cases}$$

half-infinite conditions

Note that $\mathcal{L}$ and $\mathcal{Z}_n$ are representations of the intervals $[-1, 1]$ and $[0, 1]$ respectively. Cellular automata rules are not quite functions on these intervals, however, since they are generally double valued on rational points having distinct representations as binary strings [81]. For example, both $0\underline{1}$ and $1\underline{0}$ in $\mathcal{Z}^+$, where underlining indicates infinite repetition, are numerically 1/2 but $\delta(0\underline{1}) = 11\underline{0} = 3/4$ while $\delta(1\underline{0}) = 01\underline{0} = 1/4$.

The state space $\mathcal{E}(\{0, 1\}, \mathcal{L})$ for binary valued one-dimensional cellular automata is just the set of binary sequences over the specified one-dimensional lattice. For the cases of $\mathcal{Z}$ and $\mathcal{Z}_n$ all such rules commute with the shift operator σ. When null boundary conditions are involved, however, commutativity fails at the boundary sites. For example, let $\mathcal{X}$ be the global operator for an elementary cellular automata operating on strings $\mu = \mu_0 \ldots \mu_{n-1}$ of length n with null boundary conditions. Noting that $-1 = 1$ mod(2), the commutator $[\mathcal{X}, \sigma]$ has components

$$\begin{aligned} \big[[\mathcal{X}, \sigma](\mu)\big]_i &= \big[\mathcal{X}\sigma(\mu) + \sigma\mathcal{X}(\mu)\big]_i \\ &= \begin{cases} \mathcal{X}(0\mu_1\mu_2) + X(\mu_0\mu_1\mu_2) \bmod(2) & i = 0 \\ 0 & 0 < i < n-1 \\ \mathcal{X}(\mu_{n-1}00) & i = n-1 \end{cases} \end{aligned} \tag{5}$$

The relation $[\mathcal{X}, \sigma] = \underline{0}$ can be preserved for rules defined on $\mathcal{Z}^+$ by altering the neighborhood structure. The case $\mathcal{N} = \{-1, 0, 1\}$ is called nearest neighbor because the value introduced at site i at the next iteration of a rule is determined by the values at site i and its immediately neighboring sites $i-1$ and $i+1$. Taking $\mathcal{N} = \{0, 1, 2\}$ yields left justified neighborhoods. This eliminates the need for null boundary conditions at the left boundary in $\mathcal{Z}^+$.

Changes in the neighborhood structure of this sort are equivalent to changes in the mapping site. It is important to recognize that such changes can significantly alter the topological structure of the state transition diagram (STD) for a rule. If $\mathcal{X}$ represents the global map for a rule with nearest neighbor neighborhoods, then the global map for the same rule with left justified neighborhoods is $\sigma\mathcal{X}$ and the presence of the shift operator can change cycle periods. For example, the maximum cycle period for nearest neighbor rule 90 acting on $\mathcal{Z}_6$ is 2 while the maximum cycle period for this same rule with left justified neighborhoods is 3.

$100010 \mapsto 010100 \mapsto 100010$ nearest neighbor case

$100010 \mapsto 101000 \mapsto 001010 \mapsto 100010$ left justified case

The additivity condition of Eq. (3) has an expression in terms of rule table components. For $\mathcal{A} = \{0, 1\}$ this is given by:

Theorem 1 ([81])
A k-site rule $\mathcal{X} : \mathcal{E}(\{0, 1\}, \mathcal{L}) \mapsto \mathcal{E}(\{0, 1\}, \mathcal{L})$ with rule components x_i is additive if and only if for $i = i_0 \ldots i_{k-1}$ and $j = j_0 \ldots j_{k-1}$, with $(i + j)_r = i_r + j_r$ mod(2), it is true that $x_i + x_j = x_{i+j}$ mod(2).

Corollary 1 ([81])
A k-site rule $\mathcal{X} : \mathcal{E}(\{0, 1\}, \mathcal{L}) \mapsto \mathcal{E}(\{0, 1\}, \mathcal{L})$ is additive if and only if for all $i = i_0 \ldots i_{k-1}$

$$x_i = \sum_{r=0}^{k-1} i_{k-r-1} x_{2^r} \ mod(2) \,. \tag{6}$$

By Eq. (6), the k rule components x_{2^r}, $0 \le r \le k-1$, determine the set of k-site additive rules. Hence only 2^k of the possible 2^{2^k} k-site rules are additive, including the **0** rule that maps all sites to 0.

Additive Cellular Automata and Fractals

There is a direct connection between the space-time output patterns of additive cellular automata and self-similar fractal patterns [82,83,84,85,86,87,88]. The simplest examples are elementary rules 102 and 90. When acting on a doubly infinite sequence with the initial state $\underline{0}1\underline{0}$, iteration of these rules yields the space-time output indicated in Fig. 1. In the case of rule 60, this output is the mod(2) Pascal triangle while for rule 90 it consists of every other row of this triangle [89].

The pattern generated by rule 60 (or, equivalently, by rule 102) rescales to yield the fractal known as the Sirpinski gasket [90,91]. Direct connections between cellular automata outputs and the fractal generation schemes of matrix substitution systems and hierarchical iterated function systems are shown in [92,93,94,95,96,97,98,99]. In [100,101] the dimension spectrum associated to the

Rule 60 Space-Time Output (32 Iterations)	Rule 90 Space-Time Output (16 Iterations)
1	1
11	101
101	10001
1111	110011
10001	11111111
110011	1100000011
1010101	111100001111
11111111	11001100110011
100000001	1111111111111111
1100000011	110000000000000011
10100000101	11110000000000001111
111100001111	1100110000000000110011
1000100010001	111111110000000011111111
11001100110011	11000000110000000110000001
101010101010101	1111000011110000111100001111
1111111111111111	110011001100110011001100110011
10000000000000001	11111111111111111111111111111111
110000000000000011	
1010000000000000101	
11110000000000001111	
100010000000000010001	
1100110000000000110011	
10101010000000001010101	
111111110000000011111111	
1000000010000000100000001	
11000000110000001100000011	
101000001010000010100000101	
1111000011110000111100001111	
10001000100010001000100010001	
110011001100110011001100110011	
1010101010101010101010101010101	
11111111111111111111111111111111	

Additive Cellular Automata, Figure 1
Space-time output of rules 60 and 90 from initial state with a single 1

space-time output of additive cellular automata is shown to be equal to the singularity spectrum of an associated multifractal.

Forms of Representation

Unless otherwise noted, the lattice in this section will be $\mathcal{Z}_n$ with periodic boundary conditions. Several forms of representation for additive rules with periodic boundary conditions appear in the literature. Rules have been represented as dipolynomials over finite fields [102]; as recursion relations [103,104]; as circulant matrices [105,106]; and as polynomials in roots of unity or in powers of the left shift [81].

The global operator for a k-site additive rule with neighborhood structure $\mathcal{N} = \{-r, \ldots, k-r-1\}$ can be written as

$$\mathcal{X} = \sigma^{-r} \sum_{s=0}^{k-1} a_s \sigma^s \quad a_s \in \mathcal{A} \,. \tag{7}$$

In the dipolynomial representation, a state $\mu = \mu_0 \ldots \mu_{n-1}$ defines a dipolynomial $\mu(x)$ while the rule $\mathcal{X}$ is represented by the dipolynomial equivalent of Eq. (7):

$$\begin{aligned} \mu(x) &= \sum_{s=0}^{n-1} \mu_s x^s \,, \quad a_s \in \mathcal{A} \\ \mathcal{X} &= x^{-r} \sum_{s=0}^{k-1} a_s x^s \,, \quad a_s \in \mathcal{A} \end{aligned} \tag{8}$$

The action of $\mathcal{X}$ on a state μ is obtained by multiplication of the corresponding dipolynomials: $\mathcal{X}(x)\mu(x)$, with all products reduced mod(n) [102].

The rule $\mathcal{X}$ is also representable in terms of powers of the shift operator, with the expression

$$\mathcal{X} = \sum_{s=0}^{k-1} a_s \sigma^s \quad a_s \in \mathcal{A} \tag{9}$$

where the coefficients in this equation differ from those in Eq. (7). For example, rule 90 acting on $\mathcal{Z}_n$ is represented by $\sigma^{-1} + \sigma$ in the form of Eq. (7), and by $\sigma + \sigma^5$ in the form of Eq. (9) because $-1 = 5$ mod(6).

When a rule $\mathcal{X}$ acting on $\mathcal{Z}_n$ is represented as in Eq. (9) the string $(a_0, \ldots, a_{n-1})$ is directly connected to the representation of $\mathcal{X}$ as a circulant matrix.

A right circulant matrix is a matrix in which each successive row is obtained from the row immediatly above by shifting that row one unit to the right, with the final row entry shifted to the front. Thus,

$$\mathrm{circ}(a_0, a_1, a_2, \ldots, a_{n-2}, a_{n-1}) = \begin{pmatrix} a_0 & a_1 & a_2 & \cdots & a_{n-2} & a_{n-1} \\ a_{n-1} & a_0 & a_1 & \cdots & a_{n-3} & a_{n-2} \\ & & & \vdots & & \\ a_1 & a_2 & a_3 & \cdots & a_{n-1} & a_0 \end{pmatrix}$$

If $\mu \in \mathcal{E}(\mathcal{A}, \mathcal{Z}_n)$ is written as a column vector, the operation of $\mathcal{X}$ on μ is given as multiplication by the right circulant matrix $\mathrm{circ}(a_0, \ldots, a_{n-1})$ with all terms reduced $\mathrm{mod}(p)$. The value of this representation is that properties of circulant matrices are well known and this provides significant information about the cellular automata rule. This approach has a natural extension to the case of null boundary conditions [107,108], although the matrix involved is no longer a complete circulant. For example, the nearest neighbor form of rule 90 acting on $\mathcal{Z}_6$ with periodic boundary conditions has matrix representation

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{10a}$$

while the matrix representation for null boundary conditions is

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{10b}$$

The circulant representation of the left shift σ on $\mathcal{Z}_n$ is

$$\sigma = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ & & & & \vdots & & & \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix} \tag{10c}$$

The following lemmas [109] show the direct connection to Eq. (9) and to Hedlund's condition:

Lemma 1 ([109]) *An $n \times n$ matrix $\mathbf{A}$ is circulant if and only if $[\mathbf{A}, \sigma] = \mathbf{0}$.*

Lemma 2 ([109]) *An $n \times n$ matrix $\mathbf{A}$ is circulant if and only if $\mathbf{A} = P_A(\sigma)$ for some polynomial P_A of degree less than or equal to n. Further, if $\mathbf{A} = circ(a_0, \ldots, a_{n-1})$ then*

$$P_A(\sigma) = \sum_{s=0}^{n-1} a_s \sigma^s \tag{11}$$

Other properties of circulant matrices provide links to the representation of rules over $\mathcal{Z}_n$ by polynomials in the nth roots of unity. Let $\omega = e^{2\pi i/n}$ be the first complex nth root of unity.

Definition 1 The Fourier matrix of order n is the matrix

$$F_n = \frac{1}{\sqrt{n}} \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 & 1 \\ 1 & \omega^{n-1} & \omega^{n-2} & \ldots & \omega^2 & \omega \\ 1 & \omega^{n-2} & \omega^{n-4} & \ldots & \omega^4 & \omega^2 \\ & & & \vdots & & \\ 1 & \omega^2 & \omega^4 & \ldots & \omega^{n-4} & \omega^{n-2} \\ 1 & \omega & \omega^2 & \ldots & \omega^{n-2} & \omega^{n-1} \end{pmatrix} \tag{12}$$

Denote the Hermitian conjugate (transpose, complex conjugate) of this by F_n^*.

Lemma 3 ([109])

a) *The Fourier matrix is unitary: $F_n F_n^* = F_n^* F_n = I$.*

b) *The eigenvalues of F_n are ± 1 and $\pm i$ with multiplicities depending on n.*

c) *The characteristic polynomials of F_n^* are*

$$\begin{array}{ll} (\lambda-1)^2(\lambda-i)(\lambda+1)\left(\lambda^4-1\right)^{\frac{n}{4}-1} & n = 0 \ mod(4) \\ (\lambda-1)\left(\lambda^4-1\right)^{(n-1)/4} & n = 1 \ mod(4) \\ \left(\lambda^2-1\right)\left(\lambda^4-1\right)^{(n-2)/4} & n = 2 \ mod(4) \\ (\lambda-i)\left(\lambda^2-1\right)\left(\lambda^4-1\right)^{(n-3)/4} & n = 3 \ mod(4) \end{array} \tag{13}$$

Every $n \times n$ circulant matrix $\mathbf{A}$ is diagonalized by F_n. Further, if P_A is the polynomial defined by Eq. (11) then

$$F_n \mathbf{A} F_n^* = \Lambda(\mathbf{A}) = \mathrm{diag}\left[P_A(1), P_A(\omega), \ldots, P_A(\omega^{n-1})\right] \tag{14}$$

hence the rth eigenvalue of $\mathbf{A}$ is $P_A(\omega^r)$. Define the $n \times n$ matrix $\pi_r = \mathrm{diag}(0, \ldots, 0, 1, 0, \ldots, 0)$ with the 1 in the rth position and set $\Pi_r = F_n^* \pi_r F_n$. The matrices Π_r are Hermitian and satisfy the conditions

$$\Pi_r \Pi_s = \begin{cases} 0 & r \neq s \\ \Pi_s & r = s \end{cases}$$

$$\sum_{r=0}^{n-1} \Pi_r = I$$

Thus, they are orthogonal, idempotent, and form a resolution of unity. Hence, they are a complete set of projection matrices.

Lemma 4 ([109]) *Let* $\mathbf{A} = circ(a_0, \ldots, a_{n-1})$. *Then*

$$\mathbf{A} = \sum_{r=0}^{n-1} P_A(\omega^r)\Pi_r \tag{15}$$

Representation of an additive rule in terms of complex polynomials yields an interesting result on injectivity. A rule $\mathcal{X} : \mathcal{E}(\mathcal{A}, \mathcal{Z}) \mapsto \mathcal{E}(\mathcal{A}, \mathcal{Z})$ is *surjective* if every configuration has a predecessor. If the predecessor of a configuration having a predecessor is unique, the rule is *injective*. A rule that is both surjective and injective is bijective. For cellular automata, injectivity is equivalent to reversibility. If a rule $\mathcal{X}$ is reversible, there is an inverse rule; that is, a reversible rule $\mathcal{X}^{-1}$ such that if $\xi \in \mathcal{E}(\mathcal{A}, \mathcal{Z})$ and $\mathcal{X}(\eta) = \xi$, then $\mathcal{X}^{-1}(\xi) = \eta$ [110,111,112,113,114,115]. The question of whether or not a cellular automata rule is surjective or injective is decidable only in dimension one [112]. Injective additive rules are also called group rules [52]. Those with maximum period cycles provide an effective means of generating pseudorandom sequences. For a binary valued rule operating on strings of length n and all divisors of n, the maximal possible cycle period is $2^n - 1$ (i. e., every state but the 0 state is on the cycle). If a rule operating on strings of length n has periodic boundary conditions, the shift operator produces cycles of length n. As a result, no additive rule acting on strings with periodic boundary conditions can produce cycles of maximal period. If null boundary conditions are used, however, maximal period cycles can appear [79].

For a given cellular automata rule, the set of states having no predecessor is called the Garden-of-Eden (GoE). All additive rules acting on $\mathcal{E}(\mathcal{A}, \mathcal{Z})$ are surjective but this is not true of rules acting on $\mathcal{E}(\mathcal{A}, \mathcal{Z}_n)$. For additive rules acting on strings of length n with periodic boundary conditions there will be configurations having no predecessors. Nevertheless, these Garden-of-Eden states are not intrinsic to the rule since they do have predecessors when the state space $\mathcal{E}(\mathcal{A}, \mathcal{Z}_n)$ is embedded in $\mathcal{E}(\mathcal{A}, \mathcal{Z})$, as the following example shows.

Example 5 Let $\mathcal{X} : \mathcal{E}(\{0, 1\}, \mathcal{Z}) \mapsto E(\{0, 1\}, \mathcal{Z})$ be defined by the rule table

000	001	010	011	100	101	110	111
0	0	1	1	1	1	0	0

For $\mathcal{X} : \mathcal{E}(\{0, 1\}, \mathcal{Z}_3) \mapsto \mathcal{E}(\{0, 1\}, \mathcal{Z}_3)$ the states $\{001, 010, 100, 111\}$ have no predecessors. It is easy to show, however, that these states do have predecessors in $\mathcal{Z}_6$ when $\mathcal{Z}_3$ and $\mathcal{Z}_6$ are embedded in $\mathcal{Z}$:

$$\{\underline{000111}, \underline{111000}\} \mapsto \underline{001};$$
$$\{\underline{001110}, \underline{110001}\} \mapsto \underline{010};$$
$$\{\underline{011100}, \underline{100011}\} \mapsto \underline{100};$$
$$\{\underline{01}, \underline{10}\} \mapsto \underline{1}$$

While all additive rules are surjective in this sense, not all additive rules are injective. The condition for injectivity is easily seen using the representation in terms of nth roots of unity. From Eq. (14) a rule will be injective on $\mathcal{E}(\mathcal{A}, \mathcal{Z}_n)$ if and only if its diagonalized matrix representation is invertible, hence there can be no zero eigenvalues.

Lemma 5 ([81,116]) *Let* $\mathcal{X} : \mathcal{E}(\mathcal{A}, \mathcal{Z}_n) \mapsto \mathcal{E}(\mathcal{A}, \mathcal{Z}_n)$ *be an additive rule represented by* $\mathbf{A} = circ(a_0, \ldots, a_{n-1})$. $\mathcal{X}$ *is injective on* $\mathcal{Z}_n \subset \mathcal{Z}$ *if and only if* $P_A(\omega^s) \neq 0 \ mod(p)$ *for* $0 \leq s \leq n-1$.

If the rule in Lemma 5 is to be injective on $\mathcal{Z}$ then it must be so on $\mathcal{Z}_n$ for all n. Thus the complex polynomial $P_A(z)$ must be irreducible with respect to all nth roots of unity, for all n. This leads to the next theorem:

Theorem 2 ([81,116]) *Let* $\mathcal{X} : \mathcal{E}(\mathcal{A}, \mathcal{Z}) \mapsto \mathcal{E}(\mathcal{A}, \mathcal{Z})$ *be an additive rule represented as in Eq. (9).* $\mathcal{X}$ *will be injective if and only if*

$$\lim_{\varepsilon \to 0} \frac{1}{2\pi i} \oint_{c(\varepsilon)} \frac{P'_A(z)}{P_A(z)} \, dz = 0 \tag{16}$$

where $P'_A(z)$ *is the derivative of* $P_A(z)$ *and* $c(\varepsilon)$ *is a contour consisting of circles of radius* $1 + \varepsilon$ *and* $1 - \varepsilon$.

This follows since the integral in Eq. (16) counts the number of zeros minus the number of poles of $P_A(z)$ contained within the contour. Since there are no poles, in the limit this counts the number of zeros on the unit circle.

Corollary 2 ([81]) *Let* $\mathcal{X} : \mathcal{E}(\{0, 1\}, \mathcal{Z}_n) \mapsto \mathcal{E}(\{0, 1\}, \mathcal{Z}_n)$ *be defined by* $X = \sum_{s=0}^{k-1} \sigma^s$.

1. *If k is even,* $\mathcal{X}$ *is not injective.*
2. *If k is odd,* $\mathcal{X}$ *will be injective for all n such that* $n \neq mk$ *for any m.*

Table 1 lists all rules that are injective on $\mathcal{E}(\{0, 1\}, \mathcal{Z}_n)$ for at least some n, with $k \leq 5$.

Transient Lengths and Cycle Periods

For any cellular automata acting on a finite state space, every state eventually maps to a fixed point or cycle. If

Additive Cellular Automata, Table 1
Injective Rules for k ≤ 5

$\mathcal{X}$	k	injectivity condition
σ^4	5	none
σ^3	4	none
σ^2	3	none
σ	2	none
I	1	none
$\sigma^2+\sigma^3+\sigma^4$	5	$n \neq 3m$
$\sigma+\sigma^3+\sigma^4$	5	none
$\sigma+\sigma^2+\sigma^4$	5	none
$I+\sigma^3+\sigma^4$	5	none
$I+\sigma^2+\sigma^4$	5	$n \neq 3m$
$I+\sigma+\sigma^4$	5	none
$\sigma+\sigma^2+\sigma^3$	4	$n \neq 3m$
$I+\sigma^2+\sigma^3$	4	none
$I+\sigma+\sigma^3$	4	none
$I+\sigma+\sigma^2$	3	$n \neq 3m$
$I+\sigma+\sigma^2+\sigma^3+\sigma^4$	5	$n \neq 5m$

a rule is injective, it is reversible and every state is a fixed point, or is on a cycle. If not injective, there will be states without predecessors, Garden-of-Eden states. As indicate, however, if a rule is additive its Garden-of-Eden states are spurious in the sense that they do have predecessors if the state space is enlarged.

The following theorem lists several significant properties of cellular automata rules acting on $\mathcal{E}(\mathcal{A},\mathcal{Z})$ or $\mathcal{E}(\mathcal{A},\mathcal{Z}^+)$ with left justified neighborhoods.

Theorem 3 ([81]) *Let $\mathcal{X}$ be a k-site cellular automata rule acting on $\mathcal{E}(\mathcal{A},\mathcal{Z})$ or on $\mathcal{E}(\mathcal{A},\mathcal{Z}^+)$ with left justified neighborhoods. Then the following statements are equivalent: (a) $\mathcal{X}$ is surjective, (b) $\mathcal{X}$ has an empty Garden-of-Eden, (c) Every finite sequence $\mu_0 \ldots \mu_{n-1}$ has exactly p^{k-1} pre-images and every state μ has at most p^{k-1} predecessors, (d) $\mathcal{X}$ maps eventually periodic states to eventually periodic states and non-periodic states to non-periodic states, (e) as a map of the interval [0,1] $\mathcal{X}$ maps rationals to rationals and irrationals to irrationals.*

If $\mathcal{X}:\mathcal{E}(\mathcal{A},\mathcal{Z}_n)\mapsto\mathcal{E}(\mathcal{A},\mathcal{Z}_n)$ is a k-site rule with $|\mathcal{A}|=p$ and either periodic or null boundary conditions, the state transition diagram, STD($\mathcal{X}$) is a graph with p^n vertices labeled by the set of p-adic numbers $\{i_0,\ldots,i_{n-1}|0\le i_r\le p-1\}$. An edge is directed from the vertex $i_0,\ldots,i_{n-1}$ to the vertex $j_0,\ldots,j_{n-1}$ if and only if $\mathcal{X}(i_0,\ldots,i_{n-1})=j_0,\ldots,j_{n-1}$. Each state μ maps to a unique state $\mathcal{X}(\mu)$ so STD($\mathcal{X}$) consists of a set of trees rooted on fixed points or cycles. States at the top of trees are Garden-of-Eden states.

If $h(\mathcal{X},n)$ is the maximum tree height, states at heights $\underline{h}\le h(\mathcal{X},n)$ cannot appear after $h(\mathcal{X},n)-\underline{h}+1$ iterations and after $h(\mathcal{X},n)$ iterations only fixed points and states on cycles remain. Thus, iteration of a non-injective rule on $\mathcal{E}(\mathcal{A},\mathcal{Z}_n)$ decreases the number of available states with a corresponding reduction in entropy. On the other hand, non-injective additive rules acting on $\mathcal{E}(\mathcal{A},\mathcal{Z}^+)$ do not reduce entropy [117] even though the do so on $\mathcal{E}(\mathcal{A},\mathcal{Z}_n)$ for all n. The explanation for this apparent paradox is that the Garden-of-Eden states that appear in $\mathcal{E}(\mathcal{A},\mathcal{Z}_n)$ are artifacts of the finite length of states in this space. When embedded in $\mathcal{Z}^+$, states in $\mathcal{Z}_n$ correspond to periodic configurations, hence to rational numbers in [0,1] and the set of all rationals has measure 0 in the reals.

Parameters of interest for characterizing state transition diagrams of rules acting on $\mathcal{E}(\mathcal{A},\mathcal{Z}_n)$ are the maximum tree height $h(\mathcal{X},n)$ and the cycle periods $c_s(\mathcal{X},n)$.

Theorem 4 ([102])

1. *Trees rooted at all vertices on cycles or at fixed points of the STD for additive cellular automata are isomorphic to the tree rooted at the fixed point $\underline{0}$.*
2. *The periods of all cycles of an additive rule acting on $\mathcal{Z}_n$ are divisors of the period for cycles obtained by starting from an initial state containing only a single 1.*
3. *Let $c(m)$ be the maximum cycle period for an additive cellular automaton acting on $\mathcal{Z}_m$ and take $n=2m$. Then $c(n)$ divides $2c(m)$.*
4. *Let $n=2^k m$, m odd. The maximum cycle period $c(n)$ for an additive rule acting on $\mathcal{Z}_n$ satisfies $c(n)|2^{ord(n,m)}-2^m$ where $ord(n,m)=min\{r|\,2^r=2^m\ mod(m)\}$*

In most cases, the maximum cycle period equals $2^{\mathrm{ord}(n,m)}-2^m$ or, if the rule is symmetric, $2^{\mathrm{sord}(n,m)}-2^m$. In [104,118], Jen shows that when this is not the case, it is a number theoretic consequence of an anomalous shift that reduces the maximum cycle period. As indicated, the choice of mapping site can influence cycle periods and the effect of anomalous shifts is analogous. This is an immediate result of the next theorem.

Theorem 5 ([77,118]) *Let $\mathcal{X}:\mathcal{E}(\mathcal{A},\mathcal{Z}_n)\mapsto\mathcal{E}(\mathcal{A},\mathcal{Z}_n)$ be the global mapping for a cellular automaton rule. A state μ is on a cycle of $\mathcal{X}$ if and only if there exist integers r and s such that $\mathcal{X}^r(\mu)=\sigma^{-s}(\mu)$.*

In some cases $r=c(n)$ and $s=n$ so the theorem is not as strong as it might first appear. If $\mathcal{X}^r(\mu)=\sigma^{-s}(\mu)$ for all states on cycles of maximum period, however, then $c(n)=rt$ where $t=\min\{j\,|\,js=0\ \mathrm{mod}(n)\}$. A change of mapping site is equivalent to multiplication by a power

of the shift σ. If μ is on a cycle of maximum period then $(\sigma^k \mathcal{X})^r(\mu) = \sigma^{k-s}(\mu)$ and the cycle period is rq with $q = \min\{j \mid j(k-s) = 0 \bmod(n)\}$. In general, there is no requirement that $q = t$.

Comprehensive results on cycle periods and maximum transient lengths in the case of null or periodic boundary conditions were first obtained by Elspasa [119] in work characterizing the cycle sets in the state transition diagrams of linear machines. Since then a number of researchers have independently derived similar results [120,121,122,123,124,125,126]. Let $\mathbf{A} = \mathrm{circ}(a_0, \ldots, a_{n-1})$ and let μ be a state in $\mathcal{E}(\mathcal{A}, \mathcal{Z}_n)$. The *minimal annihilating polynomial* of μ is the monic polynomial $P_\mu(z)$ such that $P_\mu(\mathbf{A})\mu = \mathbf{0} \bmod(p)$. This polynomial exists since $\mathbf{A}$ always satisfies its characteristic equation. Let $P_\mu(z) = z^k \underline{P}_\mu(z)$ with $\underline{P}_\mu(0) \neq 0$. The order of $P_\mu(z)$, $\mathrm{ord}(P_\mu(z))$, is defined as the smallest natural number c such that $\underline{P}_\mu(z) | (z^c - 1)$. The following theorem is given in [121,122]. Alternate versions appear in [123,124,126].

Theorem 6 *Let $\mu \in \mathcal{E}(\mathcal{A}, \mathcal{Z}_n)$ with minimal annihilating polynomial $P_\mu(z) = z^k \underline{P}_\mu(z)$. Then $\mathbf{A}^k \mu$ belongs to a cycle with period $c = ord(P_\mu(z))$.*

Since the minimal annihilating polynomial always divides the minimal polynomial, which, for additive cellular automata represented by circulant matrices, or by the corresponding null boundary condition matrix, is the same as the characteristic polynomial of that matrix, all cycle periods and maximum transient lengths can be found from the characteristic polynomial. Hence, the maximum cycle period is the order of the characteristic polynomial since there always exists a state whose minimal annihilating polynomial is the minimal polynomial.

The questions of reachability, and conditions for states to be on cycles is also addressed using the formulation in terms of roots of unity. If an additive rule acting on $\mathcal{E}(\mathcal{A}, \mathcal{Z}_n)$ is represented by a circulant matrix $\mathbf{A}$ and a state μ is represented by $\mu = \sum_{s=0}^{n-1} \mu_s \omega^s$, write

$$P_A(\omega) = \prod_{j=1}^{r} \rho_j(\omega) \prod_{k=1}^{s} \Omega_k(\omega) = \rho(\omega)\Omega(\omega) \tag{17}$$

where $\rho(\omega)$ is a product of the irreducible factors of $P_A(\omega)$ representing injective rules, and $\Omega(\omega)$ is a product of the irreducible factors representing non-injective rules. Diagonalization of $\mathbf{A}$ yields $\Lambda(\mathbf{A}) = \Lambda(\rho)\Lambda(\Omega)$ and $\Lambda^{-1}(\rho)$ exists since $\rho(\omega)$ represents injective (hence reversible) rules. Let ν denote the nullity of $\Lambda(\Omega)$.

Theorem 7 ([81,127]) *A state $\mu(\omega)$ is reachable by an additive cellular automata rule with circulant matrix representation $\mathbf{A}$ if and only if $\Lambda(\Omega)|\mu(\omega)$. The fraction of reachable configurations is $2^{-\nu}$. Further, if $d(\mu)$ is the in degree of the STD vertex labeled by μ then*

$$d(\mu) = \begin{cases} 0 & \mu \text{ is a Garden-of-Eden state} \\ 2^\nu & \text{otherwise} \end{cases} \tag{18}$$

Example 6a (Rule 90 acting on $\mathcal{E}(\{0,1\}, \mathcal{Z}_n)$ with periodic boundary conditions) The circulant matrix for nearest neighbor rule 90 acting on $\mathcal{Z}_n$ with periodic boundary conditions is $\mathbf{A}(\delta) = \mathrm{circ}(0, 1, 0, \ldots, 0, 1)$. The characteristic polynomial of this matrix satisfies the recurrence relation $Q_{n+1}(x) = xQ_n(x) + Q_{n-1}(x) \bmod(2)$. For $n = 2^s m$ with m odd, the characteristic polynomial of $\mathbf{A}(\delta)$ has the form

$$\begin{aligned} Q_n(x) &= x^{2^s} r_m^{2^{s+1}}(x) \\ r_{2k+1}(x) &= x r_{2k-1}(x) + r_{2k-3}(x) \end{aligned} \tag{19}$$

and the minimal annihilating polynomial is [108]

$$\begin{aligned} x r_m(x) \quad & s = 0 \\ x^{2^{s-1}} r_m^{2^s}(x) \quad & s > 0 \end{aligned} \tag{20}$$

where $r_m(0) \neq 0$. For $n = 6$ ($s = 1$, $m = 3$), the circulant matrix representing rule 90 is given by Eq. (10a). The characteristic polynomial of this matrix, with coefficients reduced mod(2), is $x^2(x^4 + 1) = x^2(x + 1)^4$. Thus, the minimal polynomial is $x(x + 1)^2 = x(x^2 + 1)$ and $r_3(x) = (x + 1)$. The minimum integer c such that $(x^2 + 1)$ divides $(x^c + 1)$ is just $c = 2$, showing that the maximum cycle period is 2.

Example 6b (Rule 90 Represented by Roots of Unity, Acting on $\mathcal{Z}_6$) The Theorem 5 condition $\mathcal{X}^r(\mu) = \sigma^{-s}(\mu)$ for a state μ to be on a cycle can be written as $\Lambda^r(\mathcal{X})F_n(\mu) = \Lambda^{n-s}(\sigma)F_n(\mu)$. This gives the conditions for μ to be on a cycle as a set of linear equations [81]. For $n = 6$, elementary rule 90 has the form $\delta = \sigma + \sigma^5$, or in terms of the sixth root of unity, $\delta(\omega) = \omega + \omega^5$. Using $\omega^3 = -1$ with all sums taken mod(2)

$$\begin{aligned} \Lambda(\delta) &= \mathrm{diag}\left(0, \omega + \omega^5, \omega^2 + \omega^4, 0, \omega^2 + \omega^4, \omega + \omega^5\right) \\ \Lambda^2(\delta) &= \mathrm{diag}\left(0, \omega^2 + \omega^4, \omega^2 + \omega^4, 0, \omega^2 + \omega^4, \omega^2 + \omega^4\right) \\ \Lambda^3(\delta) &= \mathrm{diag}\left(0, \omega + \omega^5, \omega^2 + \omega^4, 0, \omega^2 + \omega^4, \omega + \omega^5\right) \end{aligned} \tag{21}$$

hence $\Lambda^3(\delta) = \Lambda(\delta)$ or equivalently, $\Lambda(\delta)(\Lambda^2(\delta) + I) = 0 \bmod(2)$. Thus, the maximum tree height is one and the maximum cycle period is two. In addition, $\nu = 2$ so that 1/4 of the total of 64 states will be on cycles. Further,

for $n = 6$ (observing that $-1 = 1 \bmod(2)$)

$$F_6(\mu) = \frac{1}{\sqrt{6}} \begin{pmatrix} \mu_0 + \mu_1 + \mu_2 + \mu_3 + \mu_4 + \mu_5 \\ \mu_0 + \mu_1\omega^5 + \mu_2\omega^4 + \mu_3\omega^3 + \mu_4\omega^2 + \mu_5\omega \\ \mu_0 + \mu_1\omega^4 + \mu_2\omega^2 + \mu_3 + \mu_4\omega^4 + \mu_5\omega^2 \\ \mu_0 + \mu_1\omega^3 + \mu_2 + \mu_3\omega^3 + \mu_4 + \mu_5\omega^3 \\ \mu_0 + \mu_1\omega^2 + \mu_2\omega^4 + \mu_3 + \mu_4\omega^2 + \mu_5\omega^4 \\ \mu_0 + \mu_1\omega + \mu_2\omega^2 + \mu_3\omega^3 + \mu_4\omega^4 + \mu_5\omega^5 \end{pmatrix} \tag{22}$$

Since $\omega^2 + \omega^4 = \omega^3 = -1 = 1 \bmod(2)$, $\Lambda^2(\delta) = \mathrm{diag}(0, 1, 1, 0, 1, 1)$ and the condition $\Lambda^2(\delta)F_6(\mu) = F_6(\mu)$ requires that $[F_6(\mu)]_0 = [F_6(\mu)]_3 = 0$ which reduces to $\mu_0+\mu_2+\mu_4 = \mu_1+\mu_3+\mu_5 = 0$ or equivalently, $\mu_4 = \mu_0+\mu_2$ and $\mu_5 = \mu_1+\mu_3$. Hence a state μ will be on a cycle if and only if it has the form $\mu = (\mu_0, \mu_1, \mu_2, \mu_3, \mu_0 + \mu_2, \mu_1 + \mu_3)$.

Computing Predecessor States

A problem of general interest for cellular automata is computation of predecessor states. For a rule $\mathcal{X} : \mathcal{E}(\mathcal{A}, \mathcal{L}) \mapsto \mathcal{E}(\mathcal{A}, \mathcal{L})$ with a state β given this requires solution of the equation $\mathcal{X}(\mu) = \beta$. It is always possible to construct solutions for this equation, or to show that none exist by a method of backward reconstruction based on the rule table.

Example 7 (Rule 60 Acting on $\mathcal{Z}_4$ With Periodic Boundary Conditions) Rule 60 is a 2-site rule, defined by $(00, 11) \mapsto 0, (01, 10) \mapsto 1$. Given the state 0110 the predecessors of this state can be computed as follows:

1. The initial 0 in 0110 can arise from either 00 or 11.
2. Starting with a 00, the next symbol in 0110 is a 1 and this can arise from a 01 or a 10, but this must also connect to the original 00 so only 01 is allowed, giving 001. Starting from a 11, on the other hand, the same reasoning requires 110.
3. The third symbol in 0110 is also a 1. To be consistent with 001 requires that 10 be selected, and to be consistent with 110 requires that 01 be selected, thus giving the two partially constructed possibilities as 0010 and 1101.
4. Finally, the fourth symbol must be a 0. This requires that the predecessor string conclude with either 00 or 11. Since the strings are in $\mathcal{Z}_4$ with periodic boundary conditions, the final symbol in the predecessor string must also be the first symbol in that string. Thus, both 0010 and 1101 are seen to be predecessors of 0110.

Other ways of computing predecessor states for finite strings is through the construction of a rule matrix [81] or the use of de Bruijn diagrams [81,113]. Backward reconstruction, the rule matrix, and use of a de Bruijn diagram are valid methods for computing predecessor states for all one-dimensional rules. For additive rules, however, there is an analytic means for computing predecessor states, starting from left justified neighborhoods defined on $\mathcal{E}(\mathcal{A}, \mathcal{Z}_n)$ or $\mathcal{E}(\mathcal{A}, \mathcal{Z}^+)$ [81,128]. This can be illustrated for rules defined on $\mathcal{E}(\{0, 1\}, \mathcal{Z}^+)$. This method also works for rules defined on $\mathcal{E}(\{0, 1\}, \mathcal{Z}_n)$ if it is embedded in $\mathcal{E}(\{0, 1\}, \mathcal{Z}^+)$ as the subset of half-infinite periodic sequences with periods that divide n. Define operators $\mathcal{B} : \mathcal{E}(\{0, 1\}, \mathcal{Z}^+) \mapsto \mathcal{E}(\{0, 1\}, \mathcal{Z}^+)$ and $\sigma^{-1} : \mathcal{E}(\{0, 1\}, \mathcal{Z}^+) \mapsto \mathcal{E}(\{0, 1\}, \mathcal{Z}^+)$ by

$$\begin{aligned} \left[B(\mu)\right]_s &= \sum_{i=0}^{s} \mu_i \bmod(2) \\ \left[\sigma^{-1}(\mu)\right]_s &= \begin{cases} 0 & s = 0 \\ \mu_{s-1} & s > 0 \end{cases} \end{aligned} \tag{23}$$

Theorem 8 ([81,128]) *Let $D = I + \sigma$ be the global operator for elementary rule 60 acting on $\mathcal{E}(\{0, 1\}, \mathcal{Z}^+)$ and define the state $\alpha^{(s)}$ in $\mathcal{Z}^+$ by*

$$\left[\alpha^{(s)}\right]_i = \begin{cases} 0 & i \neq s \\ 1 & i = s \end{cases}$$

1. *The general solution of $D(\mu) = \beta$ is*

$$\mu = a_0\mathcal{B}(\alpha^{(0)}) + \mathcal{B}\sigma^{-1}(\beta)\,.$$

2. *The general solution of $D^k(\mu) = \beta$ is*

$$\mu = \sum_{s=0}^{k-1} a_s B^{s+1}(\alpha^{(s)}) + B^k\sigma^{-k}(\beta)\,,$$

where the coefficients a_s provide initial conditions.

The general technique for computing predecessors can be illustrated with the case of rule 150, expressed in left justified form as $\mathcal{X} = I+\sigma+\sigma^2$. To solve $(I+\sigma+\sigma^2)(\mu) = \beta$ define four sequences:

$$\begin{aligned} \mu_i^{(0)} &= \mu_{2i}, \quad \mu_i^{(1)} = \mu_{2i+1} \\ \beta_i^{(0)} &= \beta_{2i}, \quad \beta_i^{(1)} = \beta_{2i+1} \quad i = 0, 1, 2, \ldots \end{aligned} \tag{24}$$

The equation $(I + \sigma + \sigma^2)(\mu) = \beta$ reduces to the pair of coupled equations

$$\begin{aligned} \mu_i^{(0)} + \mu_{i+1}^{(0)} &= \mu_i^{(1)} + \beta_i^{(0)} \\ \mu_i^{(1)} + \mu_{i+1}^{(1)} &= \mu_{i+1}^{(0)} + \beta_i^{(0)} \end{aligned} \tag{25}$$

and these can be written as

$$D\left(\mu^{(0)}\right) = \mu^{(1)} + \beta^{(0)}$$
$$D\left(\mu^{(1)}\right) = \sigma\left(\mu^{(0)}\right) + \beta^{(1)} \tag{26}$$

From Theorem 8, Eq. (26) can be formally solved to obtain

$$\mu^{(0)} = a_0 B\left(\alpha_{(0)}\right) + B\sigma^{-1}\left(\mu^{(1)} + \beta^{(0)}\right)$$
$$\mu^{(1)} = a_1 B\left(\alpha_{(0)}\right) + B\sigma^{-1}\left(\sigma\mu^{(0)} + \beta^{(1)}\right) \tag{27}$$

Substituting the second equation of (27) into the first, making use of the identity $\sigma^{-2}\sigma(\mu^{(0)}) = \sigma^{-1}(\mu^{(0)} + \mu_0^{(0)}\alpha^{(0)})$, rearranging terms and absorbing the term $\mu_0^{(0)}\alpha^{(0)}$ into the constant parameter, results in the equation

$$\left(I + B^2\sigma^{-1}\right)\left(\mu^{(0)}\right) = B\left(a_0\alpha^{(0)} + a_1 B\left(\alpha^{(1)}\right)\right) + B\sigma^{-1}\left(\beta^{(0)}\right) + B^2\sigma^{-2}\left(\beta^{(1)}\right) \tag{28}$$

Theorem 9 ([81,128]) *The operator $(I + B^2\sigma^{-1})$ is invertible with $(I + B^2\sigma^{-1})^{-1} = I + C_{(2,1)}$ with*

$$\left[C_{(2,1)}(\xi)\right]_i = \begin{cases} 0 & i = 0 \\ \sum_{s=0}^{\left\lceil \frac{(i-1)}{3} \right\rceil} \left(\xi_{i-3s-2} + \xi_{i-3s-1}\right) & i > 0 \end{cases} \tag{29}$$

where all sums are taken mod(2), $\xi_r = 0$ for $r < 0$, and $\lceil x \rceil$ indicates the greatest integer less than or equal to x.

The solution for $\mu^{(0)}$ is substituted into the second equation of (27) yielding a solution for $\mu^{(1)}$. These are recombined to get the general solution for μ. This technique of reducing a single equation to a set of coupled equations involving simpler additive rules works in general although the form for partitioning of sequences is specific to the particular case. Computation of predecessors involves inversion of operators of the form $I + B^r\sigma^{-s}$. The general form for the inverse of this operator is $I + C_{(r,s)}$ where $C_{(r,s)}$ is the lower triangular matrix that is the solution of the equation

$$\sum_{m=j}^{j+r} \binom{r+1}{m-j+1} [C_{(r,s)}]_{im} + [C_{(r,s)}]_{i,j+s} = \delta_{i,j+s} \tag{30}$$

d-Dimensional Rules

Both [102,105] discuss the extension from one-dimensional to d-dimensional rules defined on tori. In [102] this discussion uses a formalism of multinomials defined over finite fields. In [105], the one-dimensional analysis based on circulant matrices is generalized. The matrix formulism of state transitions is retained by defining a d-fold "circulant of circulants," which is not, of itself, necessarily a circulant. Computation of the the non-zero eigenvalues of this matrix yields results on transient lengths and cycle periods.

More recently, an extensive analysis of additive rules defined on multi-dimensional tori has appeared [129]. A d-dimensional integer vector $\vec{n} = (n_1, \ldots, n_d)$ defines a discrete toridal lattice $\mathcal{L}(\vec{n})$. Every d-dimensional matrix of size $\vec{n}$ with entries in $\mathcal{A}$, $|\mathcal{A}| = p$ (prime), defines an additive rule acting on $\mathcal{E}(\mathcal{A}, \mathcal{L}(\vec{n}))$ as follows: Let $\mathcal{T}$ and $\mu(t)$ be elements of $\mathcal{E}(\mathcal{A}, \mathcal{L}(\vec{n}))$ with $\mathcal{X}$ the rule defined by $\mathcal{T}$ and $\mu(t)$ a state at time t. The state transition defined by $\mathcal{X}$ is $\mu(t+1) = \mathcal{X}(\mu(t))$ and this is given by

$$\left[\mu(t+1)\right]_{i_1\ldots i_d} = \sum_{k_1,\ldots,k_d} \left[C(\mathcal{T})\right]_{i_1\ldots i_d}^{k_1\ldots k_d} \left[\mu(t)\right]_{k_1\ldots k_d}$$
$$\left[C(\mathcal{T})\right]_{i_1\ldots i_d}^{k_1\ldots k_d} = \mathcal{T}_{j_1\ldots j_d} \quad j_s = k_s - i_s \bmod(n_s) \tag{31}$$

The matrix $C(\mathcal{T})$ is the d-dimensional generalization of a circulant matrix with $\mathcal{T}$ as the equivalent of its first row. For example, if $d = 1$ and $p = 2$ with $\mathcal{T} = (0, 1, 0, 0, 0, 1)$ this defines the additive rule $\sigma + \sigma^5$ (rule 90) and the matrix $C(\mathcal{T})$ is given in Eq. (10a).

Let S and $\mathcal{T}$ be elements of $\mathcal{E}(\mathcal{A}, \mathcal{L}(\vec{n}))$ and define the binary operation $\psi : \mathcal{E}(\mathcal{A}, \mathcal{L}(\vec{n})) \times \mathcal{E}(\mathcal{A}, \mathcal{L}(\vec{n})) \mapsto \mathcal{E}(\mathcal{A}, \mathcal{L}(\vec{n}))$ by

$$\left[\psi(S, \mathcal{T})\right]_{i_1\ldots i_d} = \sum_{\substack{k_1,\ldots,k_d \\ 0 \le k_s < n_s}} S_{k_1\ldots k_d} \mathcal{T}_{i_1-k_1\ldots i_d-k_d} \tag{32}$$

with all sums taken mod(p).

Lemma 6 ([129]) *Let S and $\mathcal{T}$ be as above, with generalized circulant matrices $C(S)$, $C(\mathcal{T})$. The product $C(S)C(\mathcal{T})$ defined by*

$$\left[C(S)C(\mathcal{T})\right]_{i_1\ldots i_d}^{j_1\ldots j_d} = \sum_{k_1\ldots k_d} \left[C(S)\right]_{i_1\ldots i_d}^{k_1\ldots k_d} \left[C(\mathcal{T})\right]_{k_1\ldots k_d}^{j_1\ldots j_d} \; mod(p) \tag{33}$$

is also a generalized circulant and $C(S)C(\mathcal{T}) = C(\psi(S, \mathcal{T}))$.

An important tool for analysis of additive rules on multi-dimensional tori is the discrete baker transformation. This

is a discrete version of the baker transformation (Bernoulli shift) for continuous one dimensional dynamical systems [130]. The discrete baker transformation is the operation $B_p : \mathcal{E}(\mathcal{A}, \mathcal{L}(\vec{n})) \mapsto \mathcal{E}(\mathcal{A}, \mathcal{L}(\vec{n}))$ defined by

$$\left[B_p * \mathcal{T}\right]_{i_1 \ldots i_d} = \sum_{i_s = p k_s \bmod(n_s)} \mathcal{T}_{k_1 \ldots k_d} \bmod(p) \quad (34)$$

with the empty sum set to 0. In the one dimensional case with $p = 2$ and $\mathcal{T} = (a_0, \ldots, a_{n-1})$ this becomes

$$[B_2 * \mathcal{T}]_i = \sum_{k:\, i = 2k \bmod(n)} a_k \bmod(2) \quad (35)$$

Example 8 If $d = 1$, $p = 2$ and $n = 7$ then $B_2(a_0, a_1, a_2, a_3, a_4, a_5, a_6) = (a_0, a_4, a_1, a_5, a_2, a_6, a_3)$. On the other hand, if $n = 6$ then $B_2(a_0, a_1, a_2, a_3, a_4, a_5) = (a_0 + a_3, 0, a_1 + a_4, 0, a_2 + a_5, 0)$.

In one dimension, if $\mathcal{X}$ is the rule defined by $\mathcal{T} = (a_0, \ldots, a_{n-1})$ then $B_p \mathcal{X} = \mathcal{X}^2$.

Example 9 If $d = 2$ and $p = 3$ with $n_1 = 5$, $n_2 = 6$ and

$$\mathcal{T} = \begin{bmatrix} a_{00} & a_{01} & a_{02} & a_{03} & a_{04} \\ a_{10} & a_{11} & a_{12} & a_{13} & a_{14} \\ a_{20} & a_{21} & a_{22} & a_{23} & a_{24} \\ a_{30} & a_{31} & a_{32} & a_{33} & a_{34} \\ a_{40} & a_{41} & a_{42} & a_{43} & a_{44} \\ a_{50} & a_{51} & a_{52} & a_{53} & a_{54} \end{bmatrix}$$

then

$$B_2 * \mathcal{T} = \begin{bmatrix} a_{00} + a_{20} + a_{40} & a_{02} + a_{22} + a_{42} & a_{04} + a_{24} + a_{44} & a_{01} + a_{21} + a_{41} & a_{03} + a_{23} + a_{43} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ a_{30} + a_{10} + a_{50} & a_{32} + a_{12} + a_{52} & a_{34} + a_{14} + a_{54} & a_{31} + a_{11} + a_{51} & a_{33} + a_{13} + a_{53} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

with all sums mod(3).

The discrete baker transformation exponentially speeds up rule evolution.

Theorem 10 ([129])
If p is prime then $[C(\mathcal{T})]^{p^r} = C\left(B_p^r * \mathcal{T}\right)$.

Thus, if the matrix $\mathcal{T}$ defines a rule $\mathcal{X} : \mathcal{E}(\mathcal{A}, \mathcal{L}(\vec{n})) \mapsto \mathcal{E}(\mathcal{A}, \mathcal{L}(\vec{n}))$ then $\mathcal{X}^{p^r}$ is obtained from the matrix $B_p^r * \mathcal{T}$.

The operator B_p is a permutation if every sum in Eq. (34) contains at most one non-zero summand. In Example 7, B_2 is a permutation for $n = 7$ but is not for $n = 6$.

Lemma 7 ([129]) *Set $\iota = max\{ex_p n_s | 1 \le s \le d\}$. That is, ι is the largest integer k such that for some s, p^k divides n_s. Then for all $r > \iota$ and any $\mathcal{T} \in \mathcal{E}(\mathcal{A}, \mathcal{L}(\vec{n}))$, B_p is a permutation on $B_p^r * \mathcal{T}$.*

Theorem 11 ([129]) *Let q be prime and let $ord_m q$ be the order of q in m when this is defined and 1 otherwise. Write $n_s = p^{k_s} m_s$ and set $c = lcm\left(ord_{m_1} p, \ldots, ord_{m_d} p\right)$. Then, for any rule $\mathcal{X}$ defined by a matrix $\mathcal{T} \in \mathcal{E}(\mathcal{A}, \mathcal{L}(\vec{n}))$ the following are true:*

1. $B_p^{\iota + c} * \mathcal{T} = B_p^{\iota} * \mathcal{T}$
2. *The period of any cycle in $STD(\mathcal{X})$ divides $p^{\iota}(p^c - 1)$*
3. *The heignt of trees in $STD(\mathcal{X})$ does not exceed p^{ι}.*

The baker transformation is a linear transformation on the space of d-dimensional matrices with entries from $\mathcal{A}$. Since each element of this space defines an additive cellular automata rule, the vertices of the state transition diagram for the baker transformation can be labeled by these rules, and this is exhaustive.

Definitions:

1. An oriented graph $G = (V, E)$ is a set V of vertices together with an edge set $E \subseteq V \times V$. If $(v, w) \in E$ then there is an edge directed from vertex v to vertex w.
2. An oriented graph $G_1 = (V_1, E_1)$ reduces to an oriented graph $G_2 = (V_2, E_2)$ modulo p ($G_1 <_p G_2$) if (a) $V_1 = V_2 = V$, and (b) $(v, w) \in E$ if and only if there is an oriented chain of length p from v to w in G_2.

Example 10 Figure 2 shows the reduction modulo 3 of a connected graph with 12 vertices to a graph with three connected components.

The baker diagram for the space $\mathcal{E}(\mathcal{A}, \mathcal{L}(\vec{n}))$ has an edge directed from rule $\mathcal{X}$ to rule $\mathcal{Y}$ if and only if $\mathcal{Y} = B_p \mathcal{X}$.

Lemma 8 ([129]) *If $(\mathcal{X}, \mathcal{Y})$ is an edge in $STD(B_p)$ then $STD(\mathcal{Y}) <_p STD(\mathcal{X})$.*

Since $\mathcal{E}(\mathcal{A}, \mathcal{L}(\vec{n}))$ is finite, iteration of B_p on $\mathcal{E}(\mathcal{A}, \mathcal{L}(\vec{n}))$ must eventually reach a fixed point or cycle. If $\mathcal{T} \in \mathcal{E}(\mathcal{A}, \mathcal{L}(\vec{n}))$ set $\tau(\mathcal{T})$ equal to the sum mod(p) of all entries in $\mathcal{T}$. If $\mathcal{X}$ is the rule defined by $\mathcal{T}$ take $\eta(\mathcal{X})$ as the maximum height of trees in STD($\mathcal{X}$), $\pi(\mathcal{X})$ as the number of vertices in STD($\mathcal{X}$) lying on cycles, and $\theta_s(\mathcal{X})$ as the number of cycles of length s contained in STD($\mathcal{X}$).

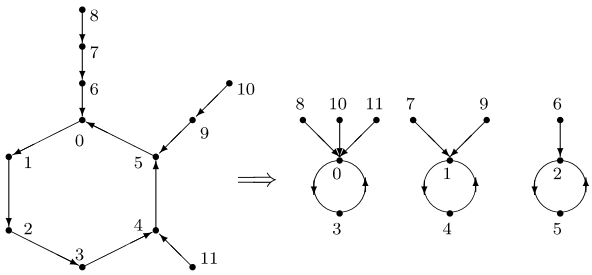

Additive Cellular Automata, Figure 2
Example of Graph Reduction Modulo 3 for a Graph with 12 Vertices (from [129])

Theorem 12 ([129])

1. *If* $STD(\mathcal{X}) <_p STD(\mathcal{Y})$ *then*
   a) $\eta(\mathcal{X}) \leq \left\lceil \frac{\eta(\mathcal{Y})}{p} \right\rceil$
   b) *If* j *is the largest number such that* $j\theta_{pj}(\mathcal{Y}) > 0$ *then* $\theta_{pj}(\mathcal{X}) = 0$
   c) $\pi(\mathcal{X}) = \pi(\mathcal{Y})$.
2. *Let* $\mathcal{X}$, $\mathcal{Y}$ *be two rules belonging to a cycle* R *of* $STD(B_p)$ *with period* l_R *and let* $\{c_1, \ldots, c_k\}$ *be the set of all cycle periods in* $STD(\mathcal{X})$. *Then*
   a) $STD(\mathcal{X})$ *and* $STD(\mathcal{Y})$ *are isomorphic as graphs*
   b) $\eta(\mathcal{X}) \leq 1$ *and* $p \not| c_s$ *for all* s
   c) *For all* s, $c_s | (p^{l_R} - 1)$, $l_R = lcm(ord_{c_1} p, \ldots, ord_{c_k} p)$, $l_R | c$
   d) *In addition, if* $\mathcal{X}$ *is defined by the matrix* $\mathcal{T}$ *and* $\tau(\mathcal{T}) = 0$ *then the number of connected components of* $STD(\mathcal{X})$ *is divisible by* p.

Future Directions

Much work remains on both the theoretical analysis and the applications of cellular automata. While much of this work will utilize non-additive automata, there are still many open questions on additive cellular automata as well. Several theoretical questions revolve around the issue of surjectivity. While all additive cellular automata are surjective (bracketing the spurious Garden-of-Eden states that exist for rules defined on finite state spaces) and for general rules surjectivity in dimensions higher than one is undecidable, it is desirable to have a simple general criteria for surjectivity in one dimension. It would also be useful to know if any surjective cellular automaton (additive or non-additive) is capable of universal computation. If there are surjective rules that are universal computers this would connect to work in mathematical logic on the computational limits of formal systems [133]. This connection arises because surjective rules only exhibit Garden-of-Eden states when acting on finite state spaces. The appearance of predecessor states when the state space is enlarged seems analogous to the increase in computational power of a formal system when it is enlarged by the addition of a new axiom.

Section "*d*-Dimensional Rules" introduced the discrete baker transformation of additive cellular automata rules. This transformation is a linear operator on the space of additive rules and further research into its properties can contribute to a deeper understanding of the structure of additive rules. In addition, there appears to be a connection between certain universal numbers that can be defined from this transformation and the well-known Mersenne and Fermat numbers. Elucidation of this connection would provide a significant link to number-theoretic aspects of cellular automata.

When applications are considered, [52] indicates many avenues of continued development for additive cellular automata. In addition, many of the references in the general bibliography point to directions of current cellular automata research in a number of areas. As indicated in the introduction, applications in physics (crystal growth, hydrodynamics, reaction-diffusion systems, astronomy), medicine and biology (pattern formation, genetic interaction networks, disease modeling, ecosystem modeling), pattern recognition and image processing, and computation (random number generation, language and pattern recognition, test pattern generation for VLSI chips, signature analysis, error correcting codes, cryptography) are only a small portion of the cellular automata applications that continue to be studied.

Bibliography

This article provides a brief survey of some of the significant theoretical results on additive cellular automata, together with references to applications of both cellular automata in general and additive cellular automata in particular. For historical information, references [1,2,3,5,7] are recommended. References [4,6,8,9] provide a general background in the use of cellular automata in modeling, as well as a number of examples. Specific exemplary cases of applications are found in [10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72]. [52], which deals extensively with the use of additive cellular automata in computing applications and VLSI chip design, is of particular value. [82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100,101] provide a good background in

the relation between cellular automata and fractal patterns. [81,102,103,104,105,106,107,108,109,110,111,112,113,114,115,116,117,118,119,120,121,122,123,124,125,126,127,128,129] deal with the theoretical analysis of additive cellular automata. A good survey of work in cellular automata up to the mid-1990s is [131]. An important computational survey of cellular automata dynamics is given in [132].

Primary Literature

1. von Neumann J (1963) The general and logical theory of automata. In: Taub A (ed) J. von Neumann Collected Works, vol 5. Pergamon, NY, pp 288–328
2. von Neumann J, Burk AW (ed) (1966) Theory of Self-Reproducing Automata. University of Illinois Press, Urbana
3. Arbib M (1966) Simple self-reproducing universal automata. Inf Control 9:177–189
4. Toffoli T, Margolis N (1987) Cellular Automata Machines: A New Environment for Modeling. MIT Press, Cambridge
5. Codd EF (1968) Cellular Automata. Academic Press, NY
6. Duff MJB, Preston K, Jr. (1984) Modern Cellular Automata: Theory and Applications. Plenum, NY
7. Sarkar P (2000) A brief history of cellular automata. ACM Comput Syst 32(1):80–107
8. Chopard B, Droz M (1998) Cellular Automata Modelling of Physical Systems. Cambridge University Press, Cambridge
9. Lindenmayer A, Rozenberg G (1976) Automata, Languages, Development. North Holland, Amsterdam
10. Mackay AL (1976) Crystal Symmetry. Phys Bull 27:495–497
11. Langer JS (1980) Instabilities and pattern formation in crystal growth. Rev Mod Phys 52:1
12. Lin F, Goldenfeld N (1990) Generic features of late-stage crystal growth. Phys Rev A 42:895–903
13. Greenberg JM, Hastings SP (1978) Spatial patterns for discrete models of diffusion in excitable media. SIAM J Appl Math 34(3):515–523
14. Greenberg JM, Hassard BD, Hastings SP (1978) Pattern formation and periodic structures in systems modeled by reaction-diffusion equations. Bull Am Math Soc 84:1296–1327
15. Madore BF, Freedman WL (1983) Computer simulations of the Belousov-Zhabotinsky reaction. Science 222:615–616
16. Adamatzky A, Costello BDL, Asai T (2005) Reaction-Diffusion Computers. Elsevier, London
17. Oono Y, Kohmoto M (1985) A discrete model for chemical turbulance. Phys Rev Lett 55:2927–2931
18. Falk H (1986) Comments on a simple cellular automata in spin representation. Physica D 20:447–449
19. Canning A, Droz M (1991) A comparison of spin exchange and cellular automata models for diffusion-controlled reactions. In: Gutowitz HA (ed) Cellular Automata: Theory and Experiment. MIT Press, Cambridge, pp 285–292
20. Vitanni P (1973) Sexually reproducing cellular automata. Math Biosci 18:23–54
21. Young D (1984) A local activator-inhibitor model of vertebrate skin patterns. Math Biosci 72:51–58
22. Dutching W, Vogelsaenger T (1985) Recent progress in modeling and simulation of three dimensional tumor growth and treatment. Biosystems 18(1):79–104
23. Moreira J, Deutsch A (2002) Cellular automata models of tumor development: A critical review. Adv Complex Syst 5(2&3):247–269
24. Sieburg HB, McCutchan JA, Clay OK, Cabalerro L, Ostlund JJ (1991) Simulation on HIV infection in artificial immune systems. In: Gutowitz HA (ed) Cellular Automata: Theory and Experiment. MIT Press, Cambridge, pp 208–227
25. Santos RMZD, Continho S (2001) Dynamics of HIAV approach: A cellular automata approach. Phys Rev Lett 87(16):102–104
26. Beauchemin C, Samuel J, Tuszynski J (2005) A simple cellular automaton model for influenza A viral infection. J Theor Biol 232(2):223–234
27. Burks C, Farmer D (1984) Towards modeling DNA sequences as automata. Physica D 10:157–167
28. Moore JH, Hahn LW (2002) Cellular automata and genetic algorithms for parallel problem solving in human genetics. In: Merelo JJ, Panagiotis A, Beyer H-G (eds) Lecture Notes in Computer Science. Springer, Berlin, pp 821–830
29. Gerola H, Seiden P (1978) Stochastic star formation and spiral structure of galaxies. Astrophys J 223:129–135
30. Flache A, Hegselmann R (1998) Understanding complex social dynamics: A plead for cellular automata based modeling. J Artif Soc Social Simul 1(3):1
31. Chen K, Bak P, Jensen MH (1990) A deterministic critical forest-fire model. Phys Lett A 149:207–210
32. Drossel B, Schwabl F (1992) Self-organized critical forest-fire model. Phys Rev Lett 69:1629–1632
33. Smith III AR (1972) Real-time language recognition by one-dimensional cellular automata. J Comput Syst Sci 6:233–253
34. Sommerhalder R, van Westrhenen SC (1983) Parallel language recognition in constant time by cellular automata. Acta Inform 19:397–407
35. Ibarra OH, Palis MA, Kim SM (1985) Fast parallel language recognition by cellular automata. Theor Comput Sci 41: 231–246
36. Morita K, Ueno S (1994) Parallel generation and parsing of array languages using reversible cellular automata. Int J Pattern Recognit Artif Intell 8:543–561
37. Jen E (1986) Invariant strings and pattern recognizing properties of 1D CA. J Stat Phys 43:243–265
38. Raghavan R (1993) Cellular automata in pattern recognition. Inf Sci 70:145–177
39. Chattopadhyay S, Adhikari S, Sengupta S, Pal M (2000) Highly regular, modular, and cascadable design of cellular automata-based pattern classifier. IEEE Trans VLSI Syst 8(6):724–735
40. Rosenfeld A (1979) Picture Languages. Academic Press, NY
41. Sternberg SR (1980) Language and architecture for parallel image processing. In: Gelesma ES, Kanal LN (eds) Pattern Recognition in Practice. North-Holland, Amsterdam, p 35
42. Hopcroft JE, Ullman JD (1972) Introduction to Automata Theory, Language, and Computation. Addison-Wesley, Reading
43. Cole SN (1969) Real time computation by n-dimensional iterative arrays of finite state machines. IEEE Trans Comput C-18:349
44. Benjamin SC, Johnson NF (1997) A possible nanometer-scale computing device based on an additive cellular automata. Appl Phys Lett 70(17):2321–2323
45. Carter F (1984) The molecular device computer: point of departure for large scale cellular automata. Physica D 10: 175–194

46. Hillis WD (1984) The connection machine: A computer architecture based on cellular automata. Physica D 10:213–228
47. Manning FB (1977) An approach to highly integrated computer-maintained cellular arrays. IEEE Trans Comput C-26:536
48. Atrubin AJ (1965) A one-dimensional real time iterative multiplier. IEEE Trans Comput EC-14:394
49. Nishio H (1981) Real time sorting of binary numbers by a 1-dimensional cellular automata. Kyoto University Technical Report
50. Fischer PC (1965) Generation of primes by a one-dimensional real time iterative array. J Assoc Comput Machin 12:388
51. Pries W, Thanailakis A, Card HC (1986) Group properties of cellular automata and VLSI applications. IEEE Trans Comput C-35:1013–1024
52. Chaudhuri PP, Chowdhury DR, Nandi S, Chattopadhyay S (1997) Additive Cellular Automata: Theory and Applications, vol 1. IEEE Computer Society Press, Los Alamitos
53. Bardell PH, McAnney WH (1986) Pseudo-random arrays for built-in tests. IEEE Trans Comput C-35(7):653–658
54. Hortensius PD, McLeod RD, Card HC (1989) Parallel random number generation for VLSI systems using cellular automata. IEEE Trans Comput 38(10):1466–1473
55. Tsalides P, York TA, Thanailakis A (1991) Pseudorandom number generation for VLSI systems using cellular automata. IEEE Proc E: Comput Digit Technol 138:241–249
56. Matsumoto M (1998) Simple cellular automata as pseudorandom m-sequence generators for built-in self-test. ACM Trans Modeling Comput Simul (TOMACS) 8(1):31–42
57. Tomassini M, Sipper M, Perrenoud M (2000) On the generation of high-quality random numbers by two-dimensional cellular automata. IEEE Trans Comput 49(10):1146–1151
58. Das AK, Chaudhuri PP (1989) An efficient on-chip deterministic test pattern generation scheme. Euromicro J, Microprocess, Microprogramm 26:195–204
59. Serra M (1990) Algebraic analysis and algorithms for linear cellular automata over GF(2) and their applications to digital circuit testing. Congressus Numerantium 75:127–139
60. Das AK, Chaudhuri PP (1993) Vector space theoretical analysis of additive cellular automata and its applications for pseudo-exhaustive test pattern generation. IEEE Trans Comput 42:340–352
61. Tziones P, Tsalides P, Thanailakis A (1994) A new cellular automaton-based nearest neighbor pattern classifier and its VLSI implementation. IEEE Trans VLSI Implement 2(3): 343–353
62. Mrugalski G, Rajski J, Tyszer J (2000) Cellular automata-based test pattern generators with phase shifters. IEEE Trans Comput-Aided Des 19(8):878–893
63. Sikdar BK, Ganguly N, Chaudhuri PP (2002) Design of hierarchical cellular automata for on-chip test pattern generator. IEEE Trans Comput Assist Des 21(12):1530–1539
64. Hortensius PD, McLeod RD, Card HC (1990) Cellular automata based signature analysis for built-in self-test. IEEE Trans Comput C-39:1273–1283
65. Serra M, Slater T, Muzio JC, Miller DM (1990) Analysis of one dimensional cellular automata and their aliasing probabilities. IEEE Trans Comput-Aided Des 9:767–778
66. Dasgupta P, Chattopadhyay S, Sengupta I (2001) Theory and application of nongroup cellular automata for message authentification. J Syst Architecture 47(7):383–404
67. Chowdhury DR, Basu S, Gupta IS, Chaudhuri PP (1994) Design of CAECC cellular automata based error correcting code. IEEE Trans Comput 43:759–764
68. Chowdhury DR, Gupta IS, Chaudhuri PP (1995) Cellular automata based byte error correcting code. IEEE Trans Comput 44:371–382
69. Chowdhury DR, Gupta IS, Chaudhuri PP (1995) A low-cost high-capacity associative memory design using cellular automata. IEEE Trans Comput 44:1260–1264
70. Nandi S, Kar BK, Chaudhuri PP (1994) Theory and application of cellular automata in cryptography. IEEE Trans Comput 43(12):1346–1357
71. Cattell K, Muzio JC (1996) Synthesis of one-dimensional linear hybrid cellular automata. IEEE Trans Comput-Aided Des 15:325–335
72. Cattell K, Zhang S, Serra M, Zmuzio JC (1999) 2-by-n hybrid cellular automata with regular configuration: Theory and applications. IEEE Trans Comput 48(3):285–295
73. Hedlund GA (1969) Endomorphisms and automorphisms of the shift dynamical system. Math Syst Theory 4:320–375
74. Wolfram S (1983) Statistical mechanics of cellular automata. Rev Mod Phys 55:601–644
75. Gardner M (1970) The fantastic combinations of John Conway's new solitaire game 'life'. Sci Am 223:120–123
76. Gardner M (1971) On cellular automata self-reproduction, the Garden of Eden and the game of 'life'. Sci Am 224:112–117
77. Jen E (1988) Cylindrical cellular automata. Commun Math Phys 118:569–590
78. Tadaki S, Matsufuji S (1993) Periodicity in one-dimensional finite linear cellular automata. Prog Theor Phys 89(2):325–331
79. Nandi S, Pal Chaudhuri P (1996) Analysis of periodic and intermediate boundary 90/150 cellular automata. IEEE Trans Comput 45(1):1–12
80. Chin W, Cortzen B, Goldman J (2001) Linear cellular automata with boundary conditions. Linear Algebra Appl 322:193–206
81. Voorhees BH (1996) Computational Analysis of One Dimensional Cellular Automata. World Scientific, Singapore
82. Willson S (1984) Cellular automata can generate fractals. Discret Appl Math 8:91–99
83. Willson S (1984) Growth rates and fractional dimension in cellular automata. Physica D 10:69–74
84. Peitgen H-O, Richter PH (1986) The Beauty of Fractals: Images of Complex Dynamical Systems. Springer, NY
85. Willson S (1987) The equality of fractional dimension for certain cellular automata. Physica D 24:179–189
86. Willson S (1987) Computing fractional dimension of additive cellular automata. Physica D 24:190–206
87. Culik II K, Dube S (1989) Fractals and recurrent behavior of cellular automata. Complex Syst 3:253–267
88. Willson S (1992) Calculating growth rates and moments for additive cellular automata. Discret Appl Math 35:47–65
89. Voorhees B (1988) Cellular automata, Pascal's triangle, and generation of order. Physica D 31:135–140
90. von Haeseler F, Peitgen H-O, Skordev G (1992) Pascal's triangle, dynamical systems and attractors. Ergod Theory Dyn Syst 12:479–486
91. Allouche JP, von Haeseler F, Peitgen H-O, Skordev G (1996) Linear cellular automata, finite automata and Pascal's triangle. Discret Appl Math 66:1–22
92. von Haeseler F, Peitgen H-O, Skordev G (1992) Linear cellular automata, substitutions, hierarchical iterated function sys-

tems. In: Encarnacao JL, Peitgen H-O, Sakas G, Englert G (eds) Fractal Geometry and Computer Graphics. Springer, NY
93. von Haeseler F, Peitgen H-O, Skordev G (1993) Cellular automata, matrix substitution, and fractals. Ann Math Artif Intell 8(3,4):345–362
94. von Haeseler F, Peitgen H-O, Skordev G (1995) Global analysis of self-similar features of cellular automata: selected examples. Physica D 86:64–80
95. Barbé A, von Haeseler F, Peitgen H-O, Skordev G (1995) Course-graining invariant patterns of one-dimensional two-state linear cellular automata. Int J Bifurc Chaos 5:1611–1631
96. von Haeseler F, Peitgen H-O, Skordev G (2001) Self-similar structures of rescaled evolution sets of cellular automata I. Int J Bifurc Chaos 11(4):913–926
97. Nagler J, Claussen JC (2005) $1/f^{\alpha}$ spectra in elementary cellular automata and fractal signals. Phys Rev E 71:067103
98. von Haeseler F, Peitgen H-O, Skordev G (2001) Self-similar structures of rescaled evolution sets of cellular automata II. Int J Bifurc Chaos 11(4):927–941
99. Barbé A, von Haeseler F, Peitgen H-O, Skordev G (2003) Rescaled evolution sets of linear cellular automata on a cylinder. Int J Bifurc Chaos 13(4):815–842
100. Takahashi S (1990) Cellular automata and multifractals: Dimension spectra of linear cellular automata. Physica D 45: 36–48
101. Takahashi S (1992) Self-similarity of linear cellular automata. J Comput Syst Sci 44(1):114–140
102. Martin O, Odlyzko A, Wolfram S (1984) Algebraic properties of cellular automata. Commun Math Phys 93:219–258
103. Jen E (1986) Global properties of cellular automata. J Stat Phys 43(1/2):219–242
104. Jen E (1988) Linear cellular automata and recurring sequences in finite fields. Commun Math Phys 119:13–28
105. Guan P, He Y (1986) Exact results for deterministic cellular automata with additive rules. J Stat Phys 43(3/4):463–478
106. Das AK, Sanyal A, Chaudhuri PP (1992) On characterization of cellular automata with matrix algebra. Inf Sci 61:251–277
107. Tadaki S (1994) Orbits in one-dimensional finite linear cellular automata. Phys Rev E 49(2):1168–1173
108. Voorhees B (2008) Representations of rule 90 and related rules for periodic, null, and half-infinite boundary conditions. J Cell Autom 3(1):1–25
109. Davis PJ (1979) Circulant Matrices. Wiley-Interscience, NY
110. Culik II K (1987) On invertible cellular automata. Complex Syst 1:1035–1044
111. Toffoli T, Margolis N (1990) Invertible cellular automata: A review. Physica D 45:229–253
112. Kari J (1990) Reversibility of 2D cellular is undecidable. Physica D 45:379–385
113. Sutner K (1991) De Bruijn graphs and linear cellular automata. Complex Syst 5(1):19–30
114. Morita K (1994) Reversible cellular automata. J Inf Process Soc Japan 35:315–321
115. Moraal H (2000) Graph-theoretical characterization of invertible cellular automata. Physica D 141:1–18
116. Voorhees B (1994) A note on injectivity of additive cellular automata. Complex Syst 8(3):151–159
117. Lind DA (1984) Applications of ergodic theory and sofic systems to cellular automata. Physica D 10:36–44
118. Jen E (1989) Limit cycles in one-dimensional cellular automata. In: Stein DL (ed) Lectures in the Sciences of Complexity. Addison Wesley, Reading
119. Elspas B (1959) The theory of autonomous linear sequential networks. TRE Trans Circuits CT-6:45–60
120. Stevens JG, Rosensweig RE, Cerkanowicz AE (1993) Transient and cyclic behavior of cellular automata with null boundary conditions. J Stat Phys 73(1,2):159–174
121. Stevens JG (1999) On the construction of state diagrams for cellular automata with additive rules. Inf Sci 115:43–59
122. Thomas DM, Stevens JG, Letteiri S (2006) Characteristic and minimal polynomials of linear cellular automata. Rocky Mountain J Math 36(3):1077–1092
123. Sutner K (1988) On σ-automata. Complex Syst 2(1):1–28
124. Sutner K (1988) Additive automata on graphs. Complex Syst 2:649–661
125. Sutner K (2000) Sigma-automata and Chebyschev polynomials. Theor Comput Sci 230:49–73
126. Sutner K (2001) Decomposition of additive CA. Complex Syst 13(2):245–270
127. Lidl R, Pilz G (1984) Applied Abstract Algebra. Springer, NY
128. Voorhees B (1993) Predecessors of cellular automata states: I. Additive automata. Physica D 68:283–292
129. Bulitko V, Voorhees B, Bulitko V (2006) Discrete baker transformation for linear cellular automata analysis. J Cell Autom 1:40–70
130. Arnold V, Aviz A (1968) Ergodic Problems of Classical Mechanics. Benjamin, NY
131. Illichinsky A (2001) Cellular Automata: A Discrete Universe. World Scientific, Singapore
132. Wuensche A, Lesser M (1992) The Global Dynamics of Cellular Automata. Addison Wesley, Reading
133. Chaitin G (2006) Meta Math! Vintage, NY

Books and Reviews

Wolfram S (1994) Cellular Automata and Complexity. Addison Wesley, Reading

Wolfram S (2002) A New Kind of Science. Wolfram Media, Champaign

Legendi T (1987) Parallel Processing by Cellular Automata and Arrays. Kluwer, Dordrecht

Rietman E (1989) Exploring the Geometry of Nature: Computer Modeling of Chaos, Fractals, Cellular Automata, and Neural Networks. McGraw-Hill, NY

Manneville P, Boccara N, Vishniac GY, Bidaux R (1990) Cellular Automata and Modeling of Complex Physical Systems. Springer, NY

Goles E, Martinez S (eds) (1992) Statistical Physics, Automata Networks and Dynamical Systems. Kluwer, Dordrecht

Boccara N, Goles E, Martenez S (1993) Cellular Automata and Cooperative Systems. Kluwer, Dordrecht

Goles E, Martinez S (1994) Cellular Automata, Dynamical Systems and Neural Networks. Kluwer, Dordrecht

Adamatzky A (1994) Identification of Cellular Automata. Taylor & Francis, London

Perdang JM, Lejeune A (eds) (1994) Cellular Automata: Prospects in Astronomy and Astrophysics. World Scientific, Singapore

Goles E, Martenez S (1996) Dynamics of Complex Interacting Systems. Kluwer, Dordrecht

Sipper M (1997) Evolution of Parallel Cellular Machines: The Cellular Programming Approach. Springer, NY

Delorme M, Mazoyer J (eds) (1998) Cellular Automata: A Parallel Model. Kluwer, Dordrecht
Goles E, Martinez S (eds) (1999) Cellular Automata and Complex Systems. Kluwer, Dordrecht
Crutchfield JP, Hanson JE (1999) Computational Mechanics of Cellular Processes. University of California Press, Berkeley
Wolf-Gladrow DA (2000) Lattice-Gas Cellular Automata and Lattice Boltzmann Models. Springer, NY
Lafe O (2000) Cellular Automata Transformations: Theory and Applications in Multimedia Compression, Encryption and Modeling. Kluwer, Dordrecht
Yang T (2001) Cellular Image Processing. Nova, NY
Griffeath D, Moore C (eds) (2003) New Constructions in Cellular Automata. Oxford University Press, NY
Deutsch A, Dormann S (2004) Cellular Automata Modeling of Biological Pattern Formation: Characterization, Applications, and Analysis. Springer, Berlin
Amos M (2004) Cellular Computing. Oxford University Press, NY
Rothman DH, Zaleski S (2004) Lattice-Gas Cellular Automata: Simple Models of Complex Hydrodynamics. Cambridge University Press, Cambridge
Batty M (2005) Cities and Complexity: Understanding Cities with Cellular Automata, Agent-Based Models, and Fractals. MIT Press, Cambridge
Kier LB, Seybold PG, Cheng C-K (2005) Cellular Automata Modeling of Chemical Systems: A Testbook and Laboratory Manual. Springer, NY
Schiff JL (2007) Cellular Automata: A Discrete View. Wiley-Interscience, NY
Gros C (2007) Complex and Adaptive Dynamical Systems: A Primer. Springer, NY
Bandini S, Moroni L (eds) (1996) ACRI '96: Proceedings of the 2nd International Conference on Cellular Automata for Research and Industry. Springer, NY
Bandini S, Serra R, Liverani FS (eds) (1998) Cellular Automata: Research Towards Industry, ACRI '98: Proceedings of the 3rd International Conference on Cellular Automata for Research and Industry. Springer, NY
Bandini S, Worsch T (eds) (2000) Theory and Practical Issues on Cellular Automata: Proceedings of the 4th International Conference on Cellular Automata for Research and Industry. Springer, NY
Tomassini M, Chopard B (eds) (2002) Cellular Automata: Proceedings of the 5th International Conference on Cellular Automata for Research and Industry. Springer, NY
Sloot PMA, Chopard B, Hoekstra AG (eds) (2004) Cellular Automata: Proceedings of the 6th International Conference on Cellular Automata for Research and Industry. Springer, NY
El Yacoubi S, Chopard B, Bandini S (eds) (2006) Cellular Automata: Proceedings of the 7th International Conference on Cellular Automata for Research and Industry. Springer, NY

Websites

http://cell-auto.com/links/ Gives many links to other sites on cellular automata
http://www.theory.org/complexity/cdpt/html/node4.html Provides reviews of theoretical aspects of cellular automata
http://www.ddlab.com An excellent site; it provides access to the Discrete Dynamics Lab program, a valuable asset in work on cellular automata and random Boolean networks
http://cellular.ci.ulsa.mx Provides access to a number of worthwhile unpublished papers and a number of useful references

Adomian Decomposition Method Applied to Non-linear Evolution Equations in Soliton Theory

Abdul-Majid Wazwaz
Department of Mathematics, Saint Xavier University, Chicago, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Adomian Decomposition Method and Adomian Polynomials
Modified Decomposition Method and Noise Terms Phenomenon
Solitons, Peakons, Kinks, and Compactons
Solitons of the KdV Equation
Kinks of the Burgers Equation
Peakons of the Camassa–Holm Equation
Compactons of the K(n,n) Equation
Future Directions
Bibliography

Glossary

Solitons Solitons appear as a result of a balance between a weakly nonlinear convection and a linear dispersion. The solitons are localized highly stable waves that retains its identity: shape and speed, upon interaction, and resemble particle like behavior. In the case of a collision, solitons undergo a phase shift.

Types of solitons *Solitary waves*, which are localized traveling waves, are asymptotically zero at large distances and appear in many structures such as solitons, kinks, peakons, cuspons, and compactons, among others. *Solitons* appear as a bell-shaped sech profile. *Kink waves* rise or descend from one asymptotic state to another. *Peakons* are peaked solitary-wave solutions. *Cuspons* exhibit cusps at their crests. In the peakon structure, the traveling-wave solutions are smooth except for a peak at a corner of its crest. Peakons are the points at which spatial derivative changes sign so that peakons have a finite jump in the first derivative of the solution $u(x, t)$. Unlike peakons, where the derivatives

at the peak differ only by a sign, the derivatives at the jump of a cuspon diverge.

Compactons are solitons with compact spatial support such that each compacton is a soliton confined to a finite core or a soliton without exponential tails. Compactons are generated due to the delicate interaction between the effect of the genuine nonlinear convection and the genuinely nonlinear dispersion.

Adomian method The Adomian decomposition method approaches linear and nonlinear, and homogeneous and inhomogeneous differential and integral equations in a unified way. The method provides the solution in a rapidly convergent series with terms that are elegantly determined in a recursive manner. The method can be used to obtain closed-form solutions, if such solutions exist. A truncated number of the obtained series can be used for numerical purposes. The method was modified to accelerate the computational process. The noise terms phenomenon, which may appear for inhomogeneous cases, can give the exact solution in two iterations only.

Definition of the Subject

Nonlinear phenomena play a significant role in many branches of applied sciences such as applied mathematics, physics, biology, chemistry, astronomy, plasma, and fluid dynamics. Nonlinear dispersive equations that govern these phenomena have the genuine soliton property. Solitons are pulses that propagate without any change of its identity, i. e., shape and speed, during their travel through a nonlinear dispersive medium [1,5,34]. Solitons resemble properties of a particle, hence the suffix *on* is used [19,20].

Solitons exist in many scientific branches, such as optical fiber photonics, fiber lasers, plasmas, molecular systems, laser pulses propagating in solids, liquid crystals, nonlinear optics, cosmology, and condensed-matter physics. Based on its importance in many fields, a huge amount of research work has been conducted during the last four decades to make more progress in understanding the soliton phenomenon. A variety of very powerful algorithms has been used to achieve this goal. The Adomian decomposition method introduced in [2,6,21,22,23,24,25,26], which will be used in this work, is one of the reliable methods that has been used recently.

Introduction

The aim of this work is to apply the Adomian decomposition method to derive specific types of soliton solutions. Solitons were discovered *experimentally* by John Scott Russell in 1844. Korteweg and de Vries in 1895 investigated *analytically* the soliton concept [10], where they derived the pioneer equation of solitons, well known as the KdV equation, that models the height of the surface of shallow water in the presence of solitary waves. Moreover, Zabusky and Kruskal [35] investigated this phenomenon *analytically* in 1965. Since 1965, a huge number of research works have been conducted on nonlinear dispersive and dissipative equations. The aim of these works has been to study thoroughly the characteristics of soliton solutions and to study various types of solitons that appear as a result of these equations. Several reliable methods were used in the literature to handle nonlinear dispersive equations. Hirota's bilinear method [8,9] has been used for single- and multiple-soliton solutions. The inverse scattering method [1] has been widely used. For single-soliton solutions, several methods, such as the tanh method [13,14,15], the pseudospectral method, and the truncated Painlevé expansion, are used.

However, in this work, the decomposition method, introduced by Adomian in 1984, will be applied to derive the desired types of soliton solutions. The method approaches all problems in a unified way and in a straightforward manner. The method computes the solution in a fast convergent series with components that are elegantly determined. Unlike other methods, the initial or boundary conditions are necessary for the use of the Adomian method.

Adomian Decomposition Method and Adomian Polynomials

The Adomian decomposition method, developed by George Adomian in 1984, has been receiving much attention in applied mathematics in general and in the area of initial value and boundary value problems in particular. Moreover, it is also used in the area of series solutions for numerical purposes. The method is powerful and effective and can be used in linear or nonlinear, ordinary or partial differential equations, and integral equations. The decomposition method demonstrates fast convergence of the solution and provides numerical approximation with a high level of accuracy. The method handles applied problems directly and in a straightforward manner without using linearization, perturbation, or any other restrictive assumption that might change the physical behavior of the physical model under investigation.

The method is effectively addressed and thoroughly used by many researchers in the literature [1,21,22,23,24,25,26]. It is important to indicate that well-known methods, such as Backlund transformation, inverse scattering method, Hirota's bilinear formalism, the tanh method, and

many others, can handle problems without using initial or boundary value conditions. For the Adomian decomposition method, the initial or boundary conditions are necessary to conduct the determination of the components recursively. However, some of the standard methods require huge calculations, whereas the Adomian method minimizes the volume of computational work.

The Adomian decomposition method consists in decomposing the unknown function $u(x,t)$ of any equation into a sum of an infinite number of components given by the decomposition series

$$u(x,t) = \sum_{n=0}^{\infty} u_n(x,t)\,, \tag{1}$$

where the components $u_n(x,t), n \geq 0$ are to be determined in a recursive manner. The decomposition method concerns itself with the determination of the components $u_0(x,t), u_1(x,t), u_2(x,t), \ldots$ individually. The determination of these components can be obtained through a recursive relation that usually involves evaluation of simple integrals.

We now give a clear overview of Adomian decomposition method. Consider the linear differential equation written in an operator form by

$$Lu + Ru = f\,, \tag{2}$$

where L is the lower-order derivative, which is assumed to be invertible, R is a linear differential operator of order greater than L, and f is a source term. We next apply the inverse operator L^{-1} to both sides of Eq. (2) and use the given initial or boundary condition to get

$$u(x,t) = g - L^{-1}(Ru)\,, \tag{3}$$

where the function g represents the terms arising from integrating the source term f and from using the given conditions; all are assumed to be prescribed. Substituting the infinite series of components

$$u(x,t) = \sum_{n=0}^{\infty} u_n(x,t) \tag{4}$$

into both sides of (3) yields

$$\sum_{n=0}^{\infty} u_n = g - L^{-1}\left(R\left(\sum_{n=0}^{\infty} u_n\right)\right)\,. \tag{5}$$

The decomposition method suggests that the zeroth component u_0 is usually defined by all terms not included under the inverse operator L^{-1}, which arise from the initial data and from integrating the inhomogeneous term. This in turn gives the formal recursive relation

$$\begin{aligned} u_0(x,t) &= g\,, \\ u_{k+1}(x,t) &= -L^{-1}\left(R(u_k)\right)\,, \quad k \geq 0\,, \end{aligned} \tag{6}$$

or, equivalently,

$$\begin{aligned} u_0(x,t) &= g\,, \\ u_1(x,t) &= -L^{-1}\left(R(u_0)\right)\,, \\ u_2(x,t) &= -L^{-1}\left(R(u_1)\right)\,, \\ u_3(x,t) &= -L^{-1}\left(R(u_2)\right)\,, \\ &\vdots\,. \end{aligned} \tag{7}$$

The differential equation under consideration is now reduced to integrals that can be easily evaluated. Having determined these components $u_0(x,t), u_1(x,t), u_2(x,t), \ldots$, we then substitute the obtained components into (4) to obtain the solution in a series form. The determined series may converge very rapidly to a closed-form solution, if an exact solution exists. For concrete problems, where a closed-form solution is not obtainable, a truncated number of terms is usually used for numerical purposes. Few terms of the truncated series give an approximation with a high degree of accuracy. The convergence concept of the decomposition series is investigated thoroughly in the literature.

Several significant studies were conducted to compare the performance of the Adomian method with other methods, such as Picard's method, Taylor series method, finite differences method, perturbation techniques, and others. The conclusions emphasized the fact that the Adomian method has many advantages and requires less computational work compared to existing techniques.

The Adomian method and many others applied this method to many deterministic and stochastic problems. However, the Adomian method, like some other methods, suffers if the zeroth component $u_0(x,t) = 0$ and makes the integrand of the right side in (7) $u_1(x,t) = 0$. If the right side integrand in (7) does not vanish, if it is of a form such as e^{u_0} or $\ln(\alpha + u_0), \alpha > 0$, then the method works effectively.

As stated before, the Adomian method decomposes the unknown function $u(x,t)$ into an infinite number of components. However, for nonlinear functions of $u(x,t)$ such as $u^2, u^3, \ln(1+u), \cos u, e^u, uu_x$, etc., a special representation for nonlinear terms was developed by Adomian and others. Adomian introduced a formal algorithm to establish the proper representation for all forms of nonlinear functions of $u(x,t)$. This representation of nonlin-

ear terms is necessary to apply the Adomian method properly. Several alternative algorithms have been introduced in the literature by researchers to calculate Adomian polynomials. However, the Adomian algorithm remains the most commonly applied because it is simple and practical; therefore, it will be used in this work.

Adomian assumed that the nonlinear term $F(u)$ can be expressed by an infinite series of the so-called Adomian polynomials A_n given in the form

$$F(u) = \sum_{n=0}^{\infty} A_n(u_0, u_1, u_2, \ldots, u_n) , \tag{8}$$

where A_n can be evaluated for all forms of nonlinearity.

Adomian polynomials A_n for the nonlinear term $F(u)$ can be evaluated using the following expression:

$$A_n = \frac{1}{n!} \frac{\mathrm{d}^n}{\mathrm{d}\lambda^n} \left[F\left(\sum_{i=0}^{n} \lambda^i u_i \right) \right]_{\lambda 0} , n = 0, 1, 2, \cdots . \tag{9}$$

The general formula (9) can be simplified as follows. Assuming that the nonlinear function is $F(u)$, using (9), Adomian polynomials are given by

$$\begin{aligned}
A_0 &= F(u_0) , \\
A_1 &= u_1 F'(u_0) , \\
A_2 &= u_2 F'(u_0) + \frac{1}{2!} u_1^2 F''(u_0) , \\
A_3 &= u_3 F'(u_0) + u_1 u_2 F''(u_0) + \frac{1}{3!} u_1^3 F'''(u_0) , \\
A_4 &= u_4 F'(u_0) + \left(\frac{1}{2!} u_2^2 + u_1 u_3 \right) F''(u_0) \\
&\quad + \frac{1}{2!} u_1^2 u_2 F'''(u_0) + \frac{1}{4!} u_1^4 F^{(iv)}(u_0) , \\
A_5 &= u_5 F'(u_0) + (u_2 u_3 + u_1 u_4) F''(u_0) \\
&\quad + \left(\frac{1}{2!} u_1 u_2^2 + \frac{1}{2!} u_1^2 u_3 \right) F'''(u_0) \\
&\quad + \frac{1}{3!} u_1^3 u_2 F^{(iv)}(u_0) + \frac{1}{5!} u_1^5 F^{(v)}(u_0) .
\end{aligned} \tag{10}$$

Other polynomials can be generated in a similar manner. It is clear that A_0 depends only on u_0, A_1 depends only on u_0 and u_1, A_2 depends only on u_0, u_1, and u_2, and so on.

For $F(u) = u^2$, we find

$$\begin{aligned}
A_0 &= F(u_0) = u_0^2 , \\
A_1 &= u_1 F'(u_0) = 2 u_0 u_1 , \\
A_2 &= u_2 F'(u_0) + \frac{1}{2!} u_1^2 F''(u_0) = 2 u_0 u_2 + u_1^2 , \\
A_3 &= u_3 F'(u_0) + u_1 u_2 F''(u_0) + \frac{1}{3!} u_1^3 F'''(u_0) \\
&= 2 u_0 u_3 + 2 u_1 u_2 .
\end{aligned} \tag{11}$$

Modified Decomposition Method and Noise Terms Phenomenon

A reliable modification of the Adomian decomposition method [22] was developed by Wazwaz in 1999. The modification will further accelerate the convergence of the series solution. As presented earlier, the standard decomposition method admits the use of the recursive relation

$$\begin{aligned}
u_0 &= g , \\
u_{k+1} &= -L^{-1}(R u_k) , \quad k \geq 0 .
\end{aligned} \tag{12}$$

The modified decomposition method introduces a slight variation to the recursive relation (12) that will lead to the determination of the components of u in a faster and easier way. For specific cases, the function g can be set as the sum of two partial functions, g_1 and g_2. In other words, we can set

$$g = g_1 + g_2 . \tag{13}$$

This assumption gives a slight qualitative change in the formation of the recursive relation (12). To reduce the size of calculations, we identify the zeroth component u_0 by one part of g, that is, g_1 or g_2. The other part of g can be added to the component u_1 among other terms. In other words, the modified recursive relation can be defined as

$$\begin{aligned}
u_0 &= g_1 , \\
u_1 &= g_2 - L^{-1}(R u_0) , \\
u_{k+1} &= -L^{-1}(R u_k) , \quad k \geq 1 .
\end{aligned} \tag{14}$$

The change occurred in the formation of the first two components u_0 and u_1 only [22,24,25]. Although this variation in the formation of u_0 and u_1 is slight, it plays a major role in accelerating the convergence of the solution and in minimizing the size of calculations. It is interesting to point out that by selecting the parts g_1 and g_2 properly, the exact solution $u(x, t)$ may be obtained by using very few iterations, and sometimes by evaluating only two components. Moreover, if g consists of one term only, the standard decomposition method should be employed in this case.

Another useful feature of the Adomian method is the noise terms phenomenon. The noise terms may appear for inhomogeneous problems only. This phenomenon was addressed by Adomian in 1994. In 1997, Wazwaz investigated the necessary conditions for the appearance of noise terms in the decomposition series. **Noise terms** are defined as identical terms with opposite signs [21] that arise in the components u_0 and u_1 particularly, and in other components as well. By canceling the noise terms between u_0 and u_1, even though u_1 contains other terms, the remaining

noncanceled terms of u_0 may give the exact solution of the equation. Therefore, it is necessary to verify that the noncanceled terms of u_0 satisfy the equation. The noise terms, if they exist in the components u_0 and u_1, will provide the solution in a closed form with only two successive iterations.

It was formally proved by Wazwaz in 1997 that a necessary condition for the appearance of the noise terms is required. The conclusion made is that the zeroth component u_0 must contain the exact solution u, among other terms [21]. Moreover, it was shown that the nonhomogeneity condition does not always guarantee the appearance of the noise terms.

Solitons, Peakons, Kinks, and Compactons

There are many types of *solitary waves*. *Solitons*, which are localized traveling waves, are asymptotically zero at large distances [1,5,7,21,22,23,24,25,26,27,28,29,30,31,32,33]. Solitons appear as a bell-shaped sech profile. Soliton solution $u(\xi)$ results as a balance between nonlinearity and dispersion, where $u(\xi), u'(\xi), u''(\xi), \cdots \to 0$ as $\xi \to \pm\infty$, where $\xi = x - ct$, and c is the speed of the wave prorogation. The soliton solution either decays exponentially as in the KdV equation, or it converges to a constant at infinity such as the kinks of the sine-Gordon equation. This means that the soliton solutions appear as sech^{α} or $\arctan(e^{\alpha(x-ct)})$. Moreover, one soliton interacts with other solitons preserving its permanent form.

Another type of solitary wave is the *kink wave*, which rises or descends from one asymptotic state to another [18]. The Burgers equation and the sine-Gordon equation are examples of nonlinear wave equations that exhibit kink solutions. It is to be noted that the kink $u(\xi)$ converges to $\pm\alpha$, where α is a constant. However, $u'(\xi), u''(\xi), \cdots \to 0$ as $\xi \to \pm\infty$.

Peakons are peaked solitary-wave solutions and are another type of solitary-wave solution. In this structure, the traveling wave solutions are smooth except for a peak at a corner of its crest. Peakons are the points at which a spatial derivative changes signs so that peakons have a finite jump in the first derivative of the solution $u(x, t)$ [3,4,11,12,32]. This means that the first derivative of $u(x, t)$ has identical values with opposite signs around the peak.

A significant type of soliton is the *compacton*, which is a soliton with compact spatial support such that each compacton is a soliton confined to a finite core or a soliton without exponential tails [16,17]. Compactons were formally derived by Rosenau and Hyman in 1993 where a special kind of the KdV equation was used to derive this new discovery. Unlike a soliton that narrows as the amplitude increases, a compacton's width is independent of the amplitude [16,17,27,28,29,30]. Classical solitons are analytic solutions, whereas compactons are nonanalytic solutions [16,17]. As will be shown by a graph below, compactons are solitons that are free of exponential wings or tails. Compactons arise as a result of the delicate interaction between genuine nonlinear terms and the genuinely nonlinear dispersion, as is the case with the K(n,n) equation.

Solitons of the KdV Equation

In 1895, Korteweg, together with his Ph.D student de Vriesf, derived *analytically* a nonlinear partial differential equation, well known now by the KdV equation given in its simplest form by

$$u_t + 6uu_x + u_{xxx} = 0, u(x, 0) = \frac{2c^2 e^{cx}}{(1 + e^{cx})^2}. \tag{15}$$

The KdV equation is the simplest equation embodying both nonlinearity and dispersion [5,34]. This equation has served as a model for the development of solitary-wave theory. The KdV equation is a completely integrable bi-Hamiltonian equation. The KdV equation is used to model the height of the surface of shallow water in the presence of solitary waves. The nonlinearity represented by uu_x tends to localize the wave, while the linear dispersion represented by u_{xxx} spreads it out. The balance between the weak nonlinearity and the linear dispersion gives solitons that consist of single humped waves. The equilibrium between the nonlinearity uu_x and the dispersion u_{xxx} of the KdV equation is stable.

In 1965, Zabusky and Kruskal [35] investigated *numerically* the nonlinear interaction of a large solitary wave overtaking a smaller one and discovered that solitary waves underwent nonlinear interactions following the KdV equation. Further, the waves emerged from this interaction retaining their original shape and amplitude, and therefore conserved energy and mass. The only effect of the interaction was a phase shift.

To apply the Adomian decomposition method, we first write the KdV Eq. (15) in an operator form:

$$L_t u = -u_{xxx} - 6uu_x \,, \tag{16}$$

where the differential operator L_t is

$$L_t = \frac{\partial}{\partial t} \,, \tag{17}$$

assuming that the integral operator L_t^{-1} exists and may be regarded as a onefold definite integral defined by

$$L_t^{-1}(.) = \int_0^t (.)\mathrm{d}t \,. \tag{18}$$

This means that

$$L_t^{-1}\, L_t\, u(x,t) = u(x,t) - u(x,0) \,. \tag{19}$$

Applying L_t^{-1} to both sides of (16) and using the initial condition we find

$$u(x,t) = \frac{2c^2 \mathrm{e}^{cx}}{(1+\mathrm{e}^{cx})^2} + L_t^{-1}(-u_{xxx} - 6uu_x) \,. \tag{20}$$

Notice that the right-hand side contains a linear term u_{xxx} and a nonlinear term uu_x. Accordingly, Adomian polynomials for this nonlinear term are given by

$$\begin{aligned}
A_0 &= F(u_0) = u_0 u_{0_x} \,, \\
A_1 &= \tfrac{1}{2} L_x(2u_0u_1) = u_{0_x}u_1 + u_0 u_{1_x} \,, \\
A_2 &= \tfrac{1}{2} L_x(2u_0u_2 + u_1^2) = u_{0_x}u_2 + u_{1_x}u_1 + u_{2_x}u_0 \,, \\
A_3 &= \tfrac{1}{2} L_x(2u_0u_3 + 2u_1u_2) \\
&= u_{0_x}u_3 + u_{1_x}u_2 + u_{2_x}u_1 + u_{3_x}u_0 \,.
\end{aligned} \tag{21}$$

Recall that the Adomian decomposition method suggests that the linear function u may be represented by a decomposition series

$$u = \sum_{n=0}^{\infty} u_n \,, \tag{22}$$

whereas nonlinear term $F(u)$ can be expressed by an infinite series of the so-called Adomian polynomials A_n given in the form

$$F(u) = \sum_{n=0}^{\infty} A_n(u_0, u_1, u_2, \ldots, u_n) \,. \tag{23}$$

Using the decomposition identification for the linear and nonlinear terms in (20) yields

$$\sum_{n=0}^{\infty} u_n(x,t) = \frac{2c^2\mathrm{e}^{cx}}{(1+\mathrm{e}^{cx})^2} - L_t^{-1}\left(\sum_{n=0}^{\infty} u_n(x,t)\right)_{xxx} - 6L_t^{-1}\left(\sum_{n=0}^{\infty} A_n\right) \,. \tag{24}$$

The Adomian method allows for the use of the recursive relation

$$\begin{aligned}
u_0(x,t) &= \frac{2c^2\mathrm{e}^{cx}}{(1+\mathrm{e}^{cx})^2} \,, \\
u_{k+1}(x,t) &= -L_t^{-1}\left(u_{k_{xxx}}\right) - 6L_t^{-1}(A_k) \,, \quad k \geq 0 \,.
\end{aligned} \tag{25}$$

The components u_n, $n \geq 0$ can be elegantly calculated by

$$\begin{aligned}
u_0(x,t) &= \frac{2c^2\mathrm{e}^{cx}}{(1+\mathrm{e}^{cx})^2} \,, \\
u_1(x,t) &= -L_t^{-1}\left(u_{0_{xxx}}\right) - 6L_t^{-1}(A_0) \\
&= \frac{2c^5\mathrm{e}^{cx}(\mathrm{e}^{cx}-1)}{(1+\mathrm{e}^{cx})^3}\, t \,, \\
u_2(x,t) &= -L_t^{-1}\left(u_{1_{xxx}}\right) - 6L_t^{-1}(A_1) \\
&= \frac{c^8\mathrm{e}^{cx}(\mathrm{e}^{2cx}-4\mathrm{e}^{cx}+1)}{(1+\mathrm{e}^{cx})^4}\, t^2 \,, \\
u_3(x,t) &= -L_t^{-1}\left(u_{2_{xxx}}\right) - 6L_t^{-1}(A_2) \\
&= \frac{c^{11}\mathrm{e}^{cx}(\mathrm{e}^{3cx}-11\mathrm{e}^{2cx}+11\mathrm{e}^{cx}-1)}{3(1+\mathrm{e}^{cx})^5}\, t^3 \,, \\
u_4(x,t) &= -L_t^{-1}\left(u_{3_{xxx}}\right) - 6L_t^{-1}(A_3) \\
&= \frac{c^{14}\mathrm{e}^{cx}(\mathrm{e}^{4cx}-26\mathrm{e}^{3cx}+66\mathrm{e}^{2cx}-26\mathrm{e}^{cx}+1)}{12(1+\mathrm{e}^{cx})^6}\, t^4 \,,
\end{aligned} \tag{26}$$

and so on. The series solution is thus given by

$$\begin{aligned}
u(x,t) = {} & \frac{2c^2\mathrm{e}^{cx}}{(1+\mathrm{e}^{cx})^2} \\
& \times \left(1 + \frac{c^3(\mathrm{e}^{cx}-1)}{(1+\mathrm{e}^{cx})}\, t + \frac{c^6(\mathrm{e}^{2cx}-4\mathrm{e}^{cx}+1)}{2(1+\mathrm{e}^{cx})^2}\, t^2 \right. \\
& \left. + \frac{c^9(\mathrm{e}^{3cx}-11\mathrm{e}^{2cx}+11\mathrm{e}^{cx}-1)}{6(1+\mathrm{e}^{cx})^3}\, t^3 + \cdots \right) \,,
\end{aligned} \tag{27}$$

and in a closed form by

$$u(x,t) = \frac{2c^2\mathrm{e}^{c(x-c^2t)}}{(1+\mathrm{e}^{c(x-c^2t)})^2} \,. \tag{28}$$

The last equation emphasizes the fact that the dispersion relation is $\omega = c^3$. Moreover, the exact solution (28) can be rewritten as

$$u(x,t) = \frac{c^2}{2}\,\mathrm{sech}^2\left[\frac{c}{2}\left(x - c^2 t\right)\right] \,. \tag{29}$$

This in turn gives the bell-shaped soliton solution of the KdV equation. A typical graph of a bell-shaped soliton is given in Fig. 1.

The graph shows that solitons are characterized by exponential wings or tails. The graph also confirms that solitons become asymptotically zero at large distances.

Kinks of the Burgers Equation

Burgers (1895–1981) introduced one of the fundamental model equations in fluid mechanics [18] that demonstrates

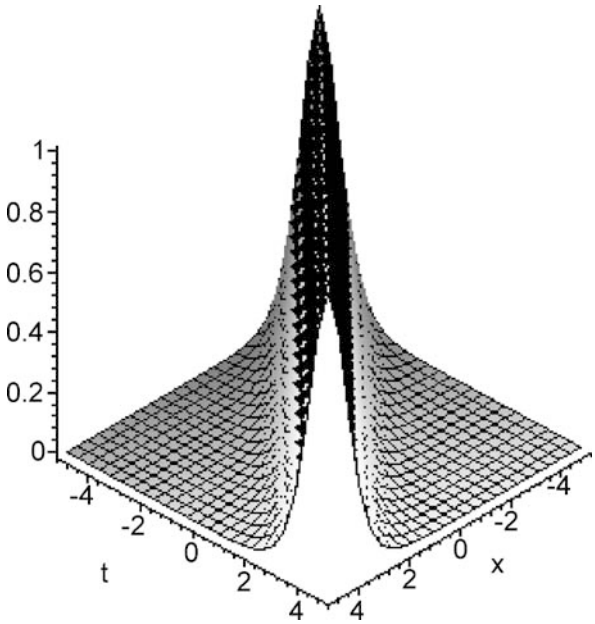

Adomian Decomposition Method Applied to Non-linear Evolution Equations in Soliton Theory, Figure 1
Figure 1 shows the soliton graph $u(x,t) = \mathrm{sech}^2(x - ct), c = 1, -5 \le x, t \le 5$

coupling between nonlinear advection uu_x and linear diffusion u_{xx}. The Burgers equation appears in gas dynamics and traffic flow. Burgers introduced this equation to capture some of the features of turbulent fluid in a channel caused by the interaction of the opposite effects of convection and diffusion. The standard form of the Burgers equation is given by

$$\begin{aligned} u_t + uu_x &= \nu u_{xx} \,, \\ u(x,0) &= \frac{2c}{(1+\mathrm{e}^{\frac{cx}{\nu}})} \,, \quad t > 0 \,, \quad \nu \neq 0 \,, \end{aligned} \tag{30}$$

where $u(x,t)$ is the velocity and ν is a constant that defines the kinematic viscosity. If the viscosity $\nu = 0$, then the equation is called an *inviscid* Burgers equation. The inviscid Burgers equation will not be examined in this work.

It is the goal of this work to apply the Adomian decomposition method to the Burgers equation; therefore we write (30) in an operator form

$$L_t u = \nu u_{xx} - uu_x \,, \tag{31}$$

where the differential operator L_t is

$$L_t = \frac{\partial}{\partial t} \,, \tag{32}$$

and as a result

$$L_t^{-1}(.) = \int_0^t (.)\mathrm{d}t \,. \tag{33}$$

This means that

$$L_t^{-1} L_t\, u(x,t) = u(x,t) - u(x,0) \,. \tag{34}$$

Applying L_t^{-1} to both sides of (31) and using the initial condition we find

$$u(x,t) = \frac{2c}{(1+\mathrm{e}^{\frac{cx}{\nu}})} + L_t^{-1}(\nu u_{xx} - uu_x) \,. \tag{35}$$

Notice that the right-hand side contains a linear term and a nonlinear term uu_x. The Adomian polynomials for this nonlinear term uu_x are the same as in the KdV equation.

Using the decomposition identification for the linear and nonlinear terms in (35) yields

$$\begin{aligned} \sum_{n=0}^{\infty} u_n(x,t) = \frac{2c}{(1+\mathrm{e}^{\frac{cx}{\nu}})} &+ L_t^{-1}\left(\nu \sum_{n=0}^{\infty} u_n(x,t)\right)_{xx} \\ &- L_t^{-1}\left(\sum_{n=0}^{\infty} A_n\right) \,. \end{aligned} \tag{36}$$

The Adomian method allows for the use of the recursive relation

$$\begin{aligned} u_0(x,t) &= \frac{2c}{(1+\mathrm{e}^{\frac{cx}{n}})} \,, \\ u_{k+1}(x,t) &= L_t^{-1}\left(\nu u_{k_{xxx}}\right) - L_t^{-1}(A_k) \,, \quad k \ge 0 \,. \end{aligned} \tag{37}$$

The components $u_n, n \ge 0$ can be elegantly calculated by

$$\begin{aligned} u_0(x,t) &= \frac{2c}{(1+\mathrm{e}^{\frac{cx}{\nu}})} \,, \\ u_1(x,t) &= L_t^{-1}\left(\nu u_{0_{xxx}}\right) - L_t^{-1}(A_0) \\ &= \frac{2c^3 \mathrm{e}^{\frac{cx}{\nu}}}{\nu(1+\mathrm{e}^{\frac{cx}{\nu}})^2}\, t \,, \\ u_2(x,t) &= L_t^{-1}\left(\nu u_{1_{xxx}}\right) - L_t^{-1}(A_1) \\ &= \frac{c^5 \mathrm{e}^{\frac{cx}{\nu}}(\mathrm{e}^{\frac{cx}{\nu}} - 1)}{\nu^2(1+\mathrm{e}^{\frac{cx}{\nu}})^3}\, t^2 \,, \\ u_3(x,t) &= L_t^{-1}\left(\nu u_{2_{xxx}}\right) - L_t^{-1}(A_2) \\ &= \frac{c^7 \mathrm{e}^{\frac{cx}{\nu}}(\mathrm{e}^{\frac{2cx}{\nu}} - 4\mathrm{e}^{\frac{cx}{\nu}} + 1)}{3\nu^3(1+\mathrm{e}^{\frac{cx}{\nu}})^4}\, t^3 \,, \\ u_4(x,t) &= L_t^{-1}\left(\nu u_{3_{xxx}}\right) - L_t^{-1}(A_3) \\ &= \frac{c^9 \mathrm{e}^{\frac{cx}{\nu}}(\mathrm{e}^{\frac{3cx}{\nu}} - 11\mathrm{e}^{\frac{2cx}{\nu}} + 11\mathrm{e}^{\frac{cx}{\nu}} - 1)}{12\nu^4(1+\mathrm{e}^{\frac{cx}{\nu}})^5}\, t^4 \,, \end{aligned} \tag{38}$$

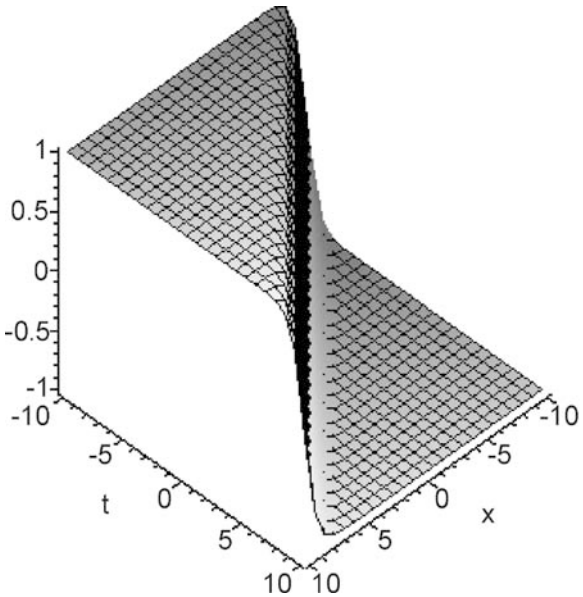

Adomian Decomposition Method Applied to Non-linear Evolution Equations in Soliton Theory, Figure 2
Figure 2 shows the kink graph $u(x,t) = \tanh(x - ct)$, $c = 1$, $-10 \le x,\ t \le 10$

and so on. The series solution is thus given by

$$u(x,t) = \frac{2ce^{\frac{cx}{\nu}}}{(1+e^{\frac{cx}{\nu}})} \times \left(e^{-\frac{cx}{\nu}} + \frac{c^2}{\nu(1+e^{\frac{cx}{\nu}})}\, t + \frac{c^4(e^{\frac{cx}{\nu}}-1)}{2\nu^2(1+e^{\frac{cx}{\nu}})^2}\, t^2 + \frac{c^6(e^{\frac{2cx}{\nu}} - 4e^{\frac{cx}{\nu}} + 1)}{6\nu^3(1+e^{\frac{cx}{\nu}})^3}\, t^3 + \cdots \right), \quad (39)$$

and in a closed form by

$$u(x,t) = \frac{2c}{\left(1 + e^{\frac{c}{\nu}(x-ct)}\right)}, \quad \nu \neq 0, \quad (40)$$

or equivalently

$$u(x,t) = c\left(1 - \tanh\left[\frac{c}{2\nu}(x-ct)\right]\right). \quad (41)$$

Figure 2 shows a kink graph.

The graph shows that the kink converges to ± 1 as $\xi \to \pm\infty$.

Peakons of the Camassa–Holm Equation

Camassa and Holm [4] derived in 1993 a completely integrable wave equation

$$u_t + 2ku_x - u_{xxt} + 3uu_x = 2u_x u_{xx} + uu_{xxx} \quad (42)$$

by retaining two terms that are usually neglected in the small-amplitude shallow-water limit [3]. The constant k is related to the critical shallow-water wave speed. This equation models the unidirectional propagation of water waves in shallow water. The CH Eq. (42) differs from the well-known regularized long wave (RLW) equation only through the two terms on the right-hand side of (42). Moreover, this equation has an integrable bi-Hamiltonian structure and arises in the context of differential geometry, where it can be seen as a reexpression for geodesic flow on an infinite-dimensional Lie group. The CH equation admits a second-order isospectral problem and allows for peaked solitary-wave solutions, called peakons.

In this work, we will apply the Adomian method to the CH equation

$$u_t - u_{xxt} + 3uu_x = 2u_x u_{xx} + uu_{xxx}, \quad u(x,0) = ce^{-|x|}, \quad (43)$$

or equivalently

$$u_t = u_{xxt} - 3uu_x + 2u_x u_{xx} + uu_{xxx}, \quad u(x,0) = ce^{-|x|}, \quad (44)$$

where we set $k = 0$.

Proceeding as before, the CH Eq. (44) in an operator form is as follows:

$$L_t u = u_{xxt} - 3uu_x + 2u_x u_{xx} + uu_{xxx}, \quad u(x,0) = ce^{-|x|}, \quad (45)$$

where the differential operator L_t is as defined above. It is important to point out that the CH equation includes three nonlinear terms; therefore, we derive the following three sets of Adomian polynomials for the terms uu_x, $u_x u_{xx}$, and uu_{xxx}:

$$\begin{aligned} A_0 &= u_0 u_{0_x}, \\ A_1 &= u_{0_x} u_1 + u_0 u_{1_x}, \\ A_2 &= u_{0_x} u_2 + u_{1_x} u_1 + u_{2_x} u_0, \end{aligned} \quad (46)$$

$$\begin{aligned} B_0 &= u_{0_x} u_{0_{xx}}, \\ B_1 &= u_{0_{xx}} u_{1_x} + u_{0_x} u_{1_{xx}}, \\ B_2 &= u_{0_{xx}} u_{2_x} + u_{1_x} u_{1_{xx}} + u_{0_x} u_{2_{xx}}, \end{aligned} \quad (47)$$

and

$$\begin{aligned} C_0 &= u_0 u_{0_{xxx}}, \\ C_1 &= u_{0_{xxx}} u_1 + u_0 u_{1_{xxx}}, \\ C_2 &= u_{0_{xxx}} u_2 + u_1 u_{1_{xxx}} + u_0 u_{2_{xxx}}. \end{aligned} \quad (48)$$

Applying the inverse operator L_t^{-1} to both sides of (45) and using the initial condition we find

$$u(x,t) = ce^{-|x|} + L_t^{-1}\left(u_{xxt} - 3uu_x + 2u_x u_{xx} + uu_{xxx}\right). \tag{49}$$

Case 1. For $x > 0$, the initial condition will be $u(x,0) = e^{-x}$. Using the decomposition series (1) and Adomian polynomials in (49) yields

$$\sum_{n=0}^{\infty} u_n(x,t) = ce^{-x} + L_t^{-1}\left(\sum_{n=0}^{\infty} u_n(x,t)\right)_{xxt} + L_t^{-1}\left(-3\sum_{n=0}^{\infty} A_n + 2\sum_{n=0}^{\infty} B_n + \sum_{n=0}^{\infty} C_n\right). \tag{50}$$

The Adomian scheme allows the use of the recursive relation

$$\begin{aligned} u_0(x,t) &= ce^{-x}\,, \\ u_{k+1}(x,t) &= L_t^{-1}\left(u_k(x,t)\right)_{xxt} \\ &\quad + L_t^{-1}\left(-3A_k + 2B_k + C_k\right), \quad k \geq 0\,. \end{aligned} \tag{51}$$

The components u_n, $n \geq 0$ can be elegantly computed by

$$\begin{aligned} u_0(x,t) &= ce^{-x}\,, \\ u_1(x,t) &= L_t^{-1}\left(u_0(x,t)\right)_{xxt} + L_t^{-1}\left(-3A_0 + 2B_0 + C_0\right) \\ &= c^2 e^{-x}\, t\,, \\ u_2(x,t) &= L_t^{-1}\left(u_1(x,t)\right)_{xxt} + L_t^{-1}\left(-3A_1 + 2B_1 + C_1\right) \\ &= \frac{1}{2!}c^3 e^{-x}\, t^2\,, \\ u_3(x,t) &= L_t^{-1}\left(u_2(x,t)\right)_{xxt} + L_t^{-1}\left(-3A_2 + 2B_2 + C_2\right) \\ &= \frac{1}{3!}c^4 e^{-x}\, t^3\,, \\ u_4(x,t) &= L_t^{-1}\left(u_3(x,t)\right)_{xxt} + L_t^{-1}\left(-3A_3 + 2B_3 + C_3\right) \\ &= \frac{1}{4!}c^4 e^{-x}\, t^4\,, \end{aligned} \tag{52}$$

and so on. The series solution for $x > 0$ is given by

$$u(x,t) = ce^{-x} \times \left(1 + ct + \frac{1}{2!}(ct)^2 + \frac{1}{3!}(ct)^3 + \frac{1}{4!}(ct)^4 + \cdots\right), \tag{53}$$

and in a closed form by

$$u(x,t) = ce^{-(x-ct)}\,. \tag{54}$$

Case 2. For $x < 0$, the initial condition will be $u(x,0) = e^{x}$. Proceeding as before, we obtain

$$\begin{aligned} u_0(x,t) &= ce^{x}\,, \\ u_1(x,t) &= L_t^{-1}\left(u_0(x,t)\right)_{xxt} + L_t^{-1}\left(-3A_0 + 2B_0 + C_0\right) \\ &= -c^2 e^{x}\, t\,, \\ u_2(x,t) &= L_t^{-1}\left(u_1(x,t)\right)_{xxt} + L_t^{-1}\left(-3A_1 + 2B_1 + C_1\right) \\ &= \frac{1}{2!}c^3 e^{x}\, t^2\,, \\ u_3(x,t) &= L_t^{-1}\left(u_2(x,t)\right)_{xxt} + L_t^{-1}\left(-3A_2 + 2B_2 + C_2\right) \\ &= -\frac{1}{3!}c^4 e^{x}\, t^3\,, \\ u_4(x,t) &= L_t^{-1}\left(u_3(x,t)\right)_{xxt} + L_t^{-1}\left(-3A_3 + 2B_3 + C_3\right) \\ &= \frac{1}{4!}c^4 e^{x}\, t^4\,, \end{aligned} \tag{55}$$

and so on. The series solution for $x < 0$ is given by

$$u(x,t) = ce^{x} \times \left(1 - ct + \frac{1}{2!}(ct)^2 - \frac{1}{3!}(ct)^3 + \frac{1}{4!}(ct)^4 + \cdots\right), \tag{56}$$

and in a closed form by

$$u(x,t) = ce^{(x-ct)}\,. \tag{57}$$

Combining the results for both cases gives the peakon solution

$$u(x,t) = ce^{-|x-ct|}\,. \tag{58}$$

Figure 3 shows a peakon graph.

The graph shows that a peakon with a peak at a corner is generated with equal first derivatives at both sides of the peak, but with opposite signs.

Compactons of the K(n,n) Equation

The K(n,n) equation [16,17] was introduced by Rosenau and Hyman in 1993. This equation was investigated experimentally and analytically. The K(m,n) equation is a genuinely nonlinear dispersive equation, a special type of the KdV equation, named K(m,n), of the form

$$u_t + (u^m)_x + (u^n)_{xxx} = 0\,, \quad n > 1\,. \tag{59}$$

Compactons, which are solitons with compact support or strict localization of solitary waves, have been investigated thoroughly in the literature. The delicate interaction between the effect of the genuine nonlinear convection $(u^n)_x$ and the genuinely nonlinear dispersion of $(u^n)_{xxx}$ generates solitary waves with exact compact support that are called *compactons*. It was also discovered that solitary

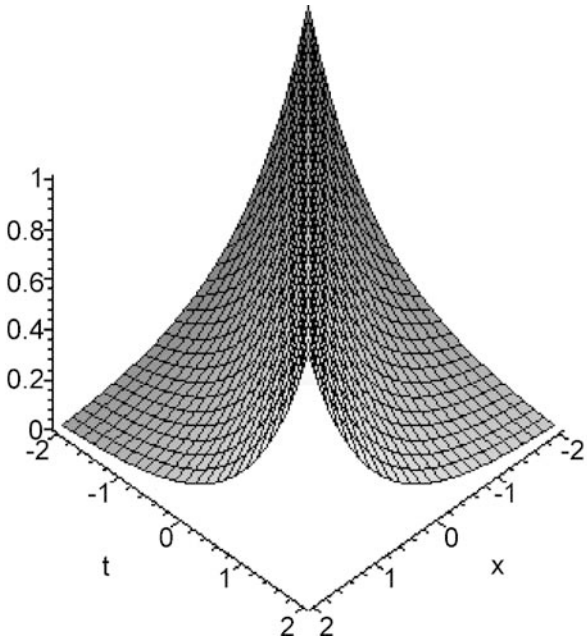

Adomian Decomposition Method Applied to Non-linear Evolution Equations in Soliton Theory, Figure 3
Figure 3 shows the peakon graph $u(x,t) = e^{-|(x-ct)|}$, $c = 1$, $-2 \le x,\ t \le 2$

waves may compactify under the influence of nonlinear dispersion, which is capable of causing deep qualitative changes in the nature of genuinely nonlinear phenomena [16,17]. Unlike solitons that narrow as the amplitude increases, a compacton's width is independent of the amplitude. Compactons such as drops do not possess infinite wings; hence they interact among themselves only across short distances.

Compactons are nonanalytic solutions, whereas classical solitons are analytic solutions. The points of nonanalyticity at the edge of a compacton correspond to points of genuine nonlinearity for the differential equation and introduce singularities in the associated dynamical system for the traveling waves [16,17,27,28,29,30]. Compactons were proved to collide elastically and vanish outside a finite core region. This discovery was studied thoroughly by many researchers who were involved with identical nonlinear dispersive equations. It is to be noted that solutions were obtained only for cases where $m = n$. However, for $m \neq n$, solutions have not yet been determined.

Without loss of generality, we will examine the special K(3,3) initial value problem

$$u_t + (u^3)_x + (u^3)_{xxx} = 0\,, \quad u(x,0) = \frac{\sqrt{6c}}{2}\cos\left(\frac{1}{3}x\right). \tag{60}$$

We first write the K(3,3) Eq. (60) in an operator form:

$$L_t u = -(u^3)_x - (u^3)_{xxx}\,, \quad u(x,0) = \frac{\sqrt{6c}}{2}\cos\left(\frac{1}{3}x\right), \tag{61}$$

where the differential operator L_t is as defined above. The differential operator L_t is defined by

$$L_t = \frac{\partial}{\partial t}\,. \tag{62}$$

Applying L_t^{-1} to both sides of (61) and using the initial condition we find

$$u(x,t) = \frac{\sqrt{6c}}{2}\cos\left(\frac{1}{3}x\right) - L_t^{-1}\left((u^3)_x + (u^3)_{xxx}\right). \tag{63}$$

Notice that the right-hand side contains two nonlinear terms $(u^3)_x$ and $(u^3)_{xxx}$. Accordingly, the Adomian polynomials for these terms are given by

$$\begin{aligned} A_0 &= (u_0^3)_x\,, \\ A_1 &= (3u_0^2 u_1)_x\,, \\ A_2 &= (3u_0^2 u_2 + 3u_0 u_1^2)_x\,, \\ A_3 &= (3u_0^2 u_3 + 6u_0 u_1 u_2 + u_1^3)_x\,, \end{aligned} \tag{64}$$

and

$$\begin{aligned} B_0 &= (u_0^3)_{xxx}\,, \\ B_1 &= (3u_0^2 u_1)_{xxx}\,, \\ B_2 &= (3u_0^2 u_2 + 3u_0 u_1^2)_{xxx}\,, \\ B_3 &= (3u_0^2 u_3 + 6u_0 u_1 u_2 + u_1^3)_{xxx}\,, \end{aligned} \tag{65}$$

respectively.

Proceeding as before, Eq. (63) becomes

$$\sum_{n=0}^{\infty} u_n(x,t) = \frac{\sqrt{6c}}{2}\cos\left(\frac{1}{3}x\right) - L_t^{-1}\left(\sum_{n=0}^{\infty} A_n + B_n\right). \tag{66}$$

This gives the recursive relation

$$\begin{aligned} u_0(x,t) &= \frac{\sqrt{6c}}{2}\cos\left(\frac{1}{3}x\right)\,, \\ u_{k+1}(x,t) &= -L_t^{-1}\left(L_t^{-1}(A_k + B_k)\right)\,, \quad k \ge 0\,. \end{aligned} \tag{67}$$

The components u_n, $n \geq 0$ can be recursively determined as

$$\begin{aligned}
u_0(x,t) &= \frac{\sqrt{6c}}{2}\cos\left(\frac{1}{3}x\right), \\
u_1(x,t) &= \frac{c\sqrt{6c}}{6}\sin\left(\frac{1}{3}x\right)t, \\
u_2(x,t) &= -\frac{c^2\sqrt{6c}}{36}\cos\left(\frac{1}{3}x\right)t^2, \\
u_3(x,t) &= -\frac{c^3\sqrt{6c}}{324}\sin\left(\frac{1}{3}x\right)t^3, \\
u_4(x,t) &= \frac{c^4\sqrt{6c}}{3888}\cos\left(\frac{1}{3}x\right)t^4, \\
u_5(x,t) &= \frac{c^5\sqrt{6c}}{58320}\sin\left(\frac{1}{3}x\right)t^5,
\end{aligned} \tag{68}$$

and so on. The series solution is thus given by

$$\begin{aligned}
&u(x,t) \\
&= \frac{\sqrt{6c}}{2}\cos\left(\frac{1}{3}x\right)\left(1-\frac{1}{2!}\left(\frac{ct}{3}\right)^2+\frac{1}{4!}\left(\frac{ct}{3}\right)^4+\cdots\right) \\
&\quad+\frac{\sqrt{6c}}{2}\sin\left(\frac{1}{3}x\right)\left(\frac{ct}{3}-\frac{1}{3!}\left(\frac{ct}{3}\right)^3+\frac{1}{5!}\left(\frac{ct}{5}\right)^5+\cdots\right).
\end{aligned} \tag{69}$$

This is equivalent to

$$\begin{aligned}
&u(x,t) \\
&= \frac{\sqrt{6c}}{2}\left(\cos\left(\frac{1}{3}x\right)\cos\left(\frac{1}{3}ct\right)+\sin\left(\frac{1}{3}x\right)\sin\left(\frac{1}{3}ct\right)\right),
\end{aligned} \tag{70}$$

or equivalently

$$u(x,t) = \begin{cases} \left\{\frac{\sqrt{6c}}{2}\cos\left[\frac{1}{3}(x-ct)\right]\right\}, & |\,x-ct\,| \leq \frac{3\pi}{2}, \\ 0 & \text{otherwise}. \end{cases} \tag{71}$$

This in turn gives the compacton solution of the K(3,3) equation. Figure 4 shows a compacton confined to a finite core without exponential wings.

The graph shows a compacton: a soliton free of exponential wings.

It is interesting to point out that the generalized solution of the K(n,n) equation, $n > 1$, is given by the compactons

$$\begin{aligned}
&u(x,t) \\
&= \begin{cases} \left\{\frac{2cn}{(n+1)}\cos^2\left[\frac{n-1}{2n}(x-ct)\right]\right\}^{\frac{1}{n-1}}, & |\,x-ct\,| \leq \frac{\pi}{2\mu}, \\ 0 & \text{otherwise}, \end{cases}
\end{aligned} \tag{72}$$

and

$$\begin{aligned}
&u_3(x,t) \\
&= \begin{cases} \left\{\frac{2cn}{(n+1)}\sin^2\left[\frac{n-1}{2n}(x-ct)\right]\right\}^{\frac{1}{n-1}}, & |\,x-ct\,| \leq \frac{\pi}{\mu}, \\ 0 & \text{otherwise}. \end{cases}
\end{aligned} \tag{73}$$

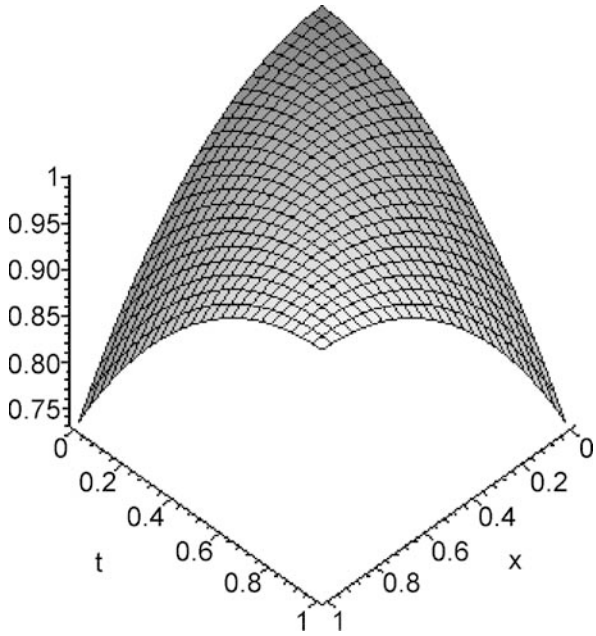

Adomian Decomposition Method Applied to Non-linear Evolution Equations in Soliton Theory, Figure 4

Figure 4 shows the compacton graph $u(x,t) = \cos^{\frac{1}{2}}(x-ct)$, $c = 1$, $0 \leq x$, $t \leq 1$

Future Directions

The most significant advantage of the Adomian method is that it attacks any problem without any need for a transformation formula or any restrictive assumption that may change the physical behavior of the solution. For singular problems, it was possible to overcome the singularity phenomenon and to attain practically a series solution in a standard way. Moreover, for problems on unbounded domains, combining the obtained series solution

by the Adomian method with Padé approximants provides a promising tool to handle boundary value problems. The Padé approximants, which often show superior performance over series approximations, provide a promising tool for use in applied fields.

As stated above, unlike other methods such as Hirota and the inverse scattering method, where solutions can be obtained without using prescribed conditions, the Adomian method requires such conditions. Moreover, these conditions must be of a form that does not provide zero for the integrand of the first component $u_1(x, t)$. Such a case should be addressed to enhance the performance of the method. Another aspect that should be addressed is Adomian polynomials. The existing techniques require tedious work to evaluate it. Most importantly, the Adomian method guarantees only one solution for nonlinear problems. This is an issue that should be investigated to improve the performance of the Adomian method compared to other methods.

On the other hand, it is important to further examine the N-soliton solutions by simplified forms, such as the form given by Hereman and Nuseir in [7]. The bilinear form of Hirota is not that easy to use and is not always attainable for nonlinear models.

Bibliography

Primary Literature

1. Ablowitz MJ, Clarkson PA (1991) Solitons, nonlinear evolution equations and inverse scattering. Cambridge University Press, Cambridge
2. Adomian G (1984) A new approach to nonlinear partial differential equations. J Math Anal Appl 102:420–434
3. Boyd PJ (1997) Peakons and cashoidal waves: travelling wave solutions of the Camassa–Holm equation. Appl Math Comput 81(2–3):173–187
4. Camassa R, Holm D (1993) An integrable shallow water equation with peaked solitons. Phys Rev Lett 71(11):1661–1664
5. Drazin PG, Johnson RS (1996) Solitons: an introduction. Cambridge University Press, Cambridge
6. Helal MA, Mehanna MS (2006) A comparison between two different methods for solving KdV-Burgers equation. Chaos Solitons Fractals 28:320–326
7. Hereman W, Nuseir A (1997) Symbolic methods to construct exact solutions of nonlinear partial differential equations. Math Comp Simul 43:13–27
8. Hirota R (1971) Exact solutions of the Korteweg–de Vries equation for multiple collisions of solitons. Phys Rev Lett 27(18):1192–1194
9. Hirota R (1972) Exact solutions of the modified Korteweg–de Vries equation for multiple collisions of solitons. J Phys Soc Jpn 33(5):1456–1458
10. Korteweg DJ, de Vries G (1895) On the change of form of long waves advancing in a rectangular canal and on a new type of long stationary waves. Philos Mag 5th Ser 36:422–443
11. Lenells J (2005) Travelling wave solutions of the Camassa–Holm equation. J Differ Equ 217:393–430
12. Liu Z, Wang R, Jing Z (2004) Peaked wave solutions of Camassa–Holm equation. Chaos Solitons Fractals 19:77–92
13. Malfliet W (1992) Solitary wave solutions of nonlinear wave equations. Am J Phys 60(7):650–654
14. Malfliet W, Hereman W (1996) The tanh method: I. Exact solutions of nonlinear evolution and wave equations. Phys Scr 54:563–568
15. Malfliet W, Hereman W (1996) The tanh method: II. Perturbation technique for conservative systems. Phys Scr 54:569–575
16. Rosenau P (1998) On a class of nonlinear dispersive-dissipative interactions. Phys D 230(5/6):535–546
17. Rosenau P, Hyman J (1993) Compactons: solitons with finite wavelengths. Phys Rev Lett 70(5):564–567
18. Veksler A, Zarmi Y (2005) Wave interactions and the analysis of the perturbed Burgers equation. Phys D 211:57–73
19. Wadati M (1972) The exact solution of the modified Korteweg–de Vries equation. J Phys Soc Jpn 32:1681–1687
20. Wadati M (2001) Introduction to solitons. Pramana J Phys 57(5/6):841–847
21. Wazwaz AM (1997) Necessary conditions for the appearance of noise terms in decomposition solution series. Appl Math Comput 81:265–274
22. Wazwaz AM (1998) A reliable modification of Adomian's decomposition method. Appl Math Comput 92:1–7
23. Wazwaz AM (1999) A comparison between the Adomian decomposition method and the Taylor series method in the series solutions. Appl Math Comput 102:77–86
24. Wazwaz AM (2000) A new algorithm for calculating Adomian polynomials for nonlinear operators. Appl Math Comput 111(1):33–51
25. Wazwaz AM (2001) Exact specific solutions with solitary patterns for the nonlinear dispersive K(m,n) equations. Chaos, Solitons and Fractals 13(1):161–170
26. Wazwaz AM (2002) General solutions with solitary patterns for the defocusing branch of the nonlinear dispersive K(n,n) equations in higher dimensional spaces. Appl Math Comput 133(2/3):229–244
27. Wazwaz AM (2003) An analytic study of compactons structures in a class of nonlinear dispersive equations. Math Comput Simul 63(1):35–44
28. Wazwaz AM (2003) Compactons in a class of nonlinear dispersive equations. Math Comput Model l37(3/4):333–341
29. Wazwaz AM (2004) The tanh method for travelling wave solutions of nonlinear equations. Appl Math Comput 154(3):713–723
30. Wazwaz AM (2005) Compact and noncompact structures for variants of the KdV equation. Int J Appl Math 18(2):213–221
31. Wazwaz AM (2006) New kinds of solitons and periodic solutions to the generalized KdV equation. Numer Methods Partial Differ Equ 23(2):247–255
32. Wazwaz AM (2006) Peakons, kinks, compactons and solitary patterns solutions for a family of Camassa–Holm equations by using new hyperbolic schemes. Appl Math Comput 182(1):412–424
33. Wazwaz AM (2007) The extended tanh method for new solitons solutions for many forms of the fifth-order KdV equations. Appl Math Comput 184(2):1002–1014
34. Whitham GB (1999) Linear and nonlinear waves. Wiley, New York

35. Zabusky NJ, Kruskal MD (1965) Interaction of solitons in a collisionless plasma and the recurrence of initial states. Phys Rev Lett 15:240–243

Books and Reviews

Adomian G (1994) Solving Frontier Problems of Physics: The Decomposition Method. Kluwer, Boston
Burgers JM (1974) The nonlinear diffusion equation. Reidel, Dordrecht
Conte R, Magri F, Musette M, Satsuma J, Winternitz P (2003) Lecture Notes in Physics: Direct and Inverse methods in Nonlinear Evolution Equations. Springer, Berlin
Hirota R (2004) The Direct Method in Soliton Theory. Cambridge University Press, Cambridge
Johnson RS (1997) A Modern Introduction to the Mathematical Theory of Water Waves. Cambridge University Press, Cambridge
Kosmann-Schwarzbach Y, Grammaticos B, Tamizhmani KM (2004) Lecture Notes in Physics: Integrability of Nonlinear Systems. Springer, Berlin
Wazwaz AM (1997) A First Course in Integral Equations. World Scientific, Singapore
Wazwaz AM (2002) Partial Differential Equations: Methods and Applications. Balkema, The Netherlands

Agent Based Computational Economics

Moshe Levy
The Hebrew University, Jerusalem, Israel

Article Outline

Glossary
Definition of the Subject
Introduction
Some of the Pioneering Studies
Illustration with the LLS Model
Summary and Future Directions
Bibliography

Glossary

Agent-based simulation A simulation of a system of multiple interacting agents (sometimes also known as "microscopic simulation"). The "micro" rules governing the actions of the agents are known, and so are their rules of interaction. Starting with some initial conditions, the dynamics of the system are investigated by simulating the state of the system through discrete time steps. This approach can be employed to study general properties of the system, which are not sensitive to the initial conditions, or the dynamics of a specific system with fairly well-known initial conditions, e. g. the impact of the baby boomers' retirement on the US stock market.

Bounded-rationality Most economic models describe agents as being *fully* rational – given the information at their disposal they act in the optimal way which maximizes their objective (or utility) function. This optimization may be technically very complicated, requiring economic, mathematical and statistical sophistication. In contrast, *bounded* rational agents are limited in their ability to optimize. This limitation may be due to limited computational power, errors, or various psychological biases which have been experimentally documented.

Market anomalies Empirically documented phenomena that are difficult to explain within the standard rational representative agent economic framework. Some of these phenomena are the over-reaction and under-reaction of prices to news, the auto-correlation of stock returns, various calendar and day-of-the-week effects, and the excess volatility of stock returns.

Representative agent A standard modeling technique in economics, by which an entire class of agents (e. g. investors) are modeled by a single "representative" agent. If agents are completely homogeneous, it is obvious that the representative agent method is perfectly legitimate. However, when agents are heterogeneous, the representative agent approach can lead to a multitude of problems (see [16]).

Definition of the Subject

Mainstream economic models typically make the assumption that an entire group of agents, e. g. "investors", can be modeled with a single "rational representative agent". While this assumption has proven extremely useful in advancing the science of economics by yielding analytically tractable models, it is clear that the assumption is not realistic: people are different one from the other in their tastes, beliefs, and sophistication, and as many psychological studies have shown, they often deviate from rationality in systematic ways.

Agent Based Computational Economics is a framework allowing economics to expand beyond the realm of the "rational representative agent". By modeling and simulating the behavior of each agent and the interaction among agents, agent based simulation allows us to investigate the dynamics of complex economic systems with many heterogeneous and not necessarily fully rational agents.

The agent based simulation approach allows economists to investigate systems that can not be studied with

the conventional methods. Thus, the following key questions can be addressed: How do heterogeneity and systematic deviations from rationality affect markets? Can these elements explain empirically observed phenomena which are considered "anomalies" in the standard economics literature? How robust are the results obtained with the analytical models? By addressing these questions the agent based simulation approach complements the traditional analytical analysis, and is gradually becoming a standard tool in economic analysis.

Introduction

For solving the dynamics of two bodies (e. g. stars) with some initial locations and velocities and some law of attraction (e. g. gravitation) there is a well-known analytical solution. However, for a similar system with three bodies there is no known analytical solution. Of course, this does not mean that physicists can't investigate and predict the behavior of such systems. Knowing the state of the system (i. e. the location, velocity, and acceleration of each body) at time t, allows us to calculate the state of the system an instant later, at time $t + \Delta t$. Thus, starting with the initial conditions we can predict the dynamics of the system by simply simulating the "behavior" of each element in the system over time.

This powerful and fruitful approach, sometimes called "Microscopic Simulation", has been adopted by many other branches of science. Its application in economics is best known as "Agent Based Simulation" or "Agent Based Computation". The advantages of this approach are clear – they allow the researcher to go where no analytical models can go. Yet, despite of the advantages, perhaps surprisingly, the agent based approach was not adopted very quickly by economists. Perhaps the main reason for this is that a particular simulation only describes the dynamics of a system with a particular set of parameters and initial conditions. With other parameters and initial conditions the dynamics may be different. So economists may ask: what is the value of conducting simulations if we get very different results with different parameter values? While in physics the parameters (like the gravitational constant) may be known with great accuracy, in economics the parameters (like the risk-aversion coefficient, or for that matter the entire decision-making rule) are typically estimated with substantial error. This is a strong point. Indeed, we would argue that the "art" of agent based simulations is the ability to understand the general dynamics of the system and to draw general conclusions from a finite number of simulations. Of course, one simulation is sufficient as a counter-example to show that a certain result does not hold, but many more simulations are required in order to convince of an alternative general regularity.

This manuscript is intended as an introduction to agent-based computational economics. An introduction to this field has two goals: (i) to explain and to demonstrate the agent-based methodology in economics, stressing the advantages and disadvantages of this approach relative to the alternative purely analytical methodology, and (ii) to review studies published in this area. The emphasis in this paper will be on the first goal. While Sect. "Some of the Pioneering Studies" does provide a brief review of some of the cornerstone studies in this area, more comprehensive reviews can be found in [19,24,32,39,40], on which part of Sect. "Some of the Pioneering Studies" is based. A comprehensive review of the many papers employing agent based computational models in economics will go far beyond the scope of this article. To achieve goal (i) above, in Sect. "Illustration with the LLS Model" we will focus on one particular model of the stock market in some detail. Section "Summary and Future Directions" concludes with some thoughts about the future of the field.

Some of the Pioneering Studies

Schelling's Segregation Model

Schelling's [36] classical segregation model is one of the earliest models of population dynamics. Schelling's model is not intended as a realistic tool for studying the actual dynamics of specific communities as it ignores economic, real-estate and cultural factors. Rather, the aim of this very simplified model is to explain the emergence of macroscopic single-race neighborhoods even when individuals are not racists. More precisely, Schelling found that the collective effect of neighborhood racial segregation results even from individual behavior that presents only a very mild preference for same-color neighbors. For instance, even the minimal requirement by each individual of having (at least) one neighbor belonging to ones' own race leads to the segregation effect.

The agent based simulation starts with a square mesh, or lattice, (representing a town) which is composed of cells (representing houses). On these cells reside agents which are either "blue" or "green" (the different races). The crucial parameter is the minimal percentage of same-color neighbors that each agent requires. Each agent, in his turn, examines the color of all his neighbors. If the percentage of neighbors belonging to his own group is above the "minimal percentage", the agent does nothing. If the percentage of neighbors of his own color is less then the minimal percentage, the agent moves to the closest unoccupied cell. The agent then examines the color of the neighbors of the

new location and acts accordingly (moves if the number of neighbors of his own color is below the minimal percentage and stays there otherwise). This goes on until the agent is finally located at a cite in which the minimal percentage condition holds. After a while, however, it might happen that following the moves of the other agents, the minimal percentage condition ceases to be fulfilled and then the agent starts moving again until he finds an appropriate cell. As mentioned above, the main result is that even for very mild individual preferences for same-color neighbors, after some time the entire system displays a very high level of segregation.

A more modern, developed and sophisticated reincarnation of these ideas is the Sugarscape environment described by Epstein and Axtell [6]. The model considers a population of moving, feeding, pairing, procreating, trading, warring agents and displays various qualitative collective events which their populations incur. By employing agent based simulation one can study the macroscopic results induced by the agents' individual behavior.

The Kim and Markowitz Portfolio Insurers Model

Harry Markowitz is very well known for being one of the founders of modern portfolio theory, a contribution for which he has received the Nobel Prize in economics. It is less well known, however, that Markowitz is also one of the pioneers in employing agent based simulations in economics.

During the October 1987 crash markets all over the globe plummeted by more than 20% within a few days. The surprising fact about this crash is that it appeared to be spontaneous – it was not triggered by any obvious event. Following the 1987 crash researchers started to look for endogenous market features, rather than external forces, as sources of price variation. The Kim-Markowitz [15] model explains the 1987 crash as resulting from investors' "Constant Proportion Portfolio Insurance" (CPPI) policy. Kim and Markowitz proposed that market instabilities arise as a consequence of the individual insurers' efforts to cut their losses by selling once the stock prices are going down.

The Kim Markowitz agent based model involves two groups of individual investors: rebalancers and insurers (CPPI investors). The rebalancers are aiming to keep a constant composition of their portfolio, while the insurers make the appropriate operations to insure that their eventual losses will not exceed a certain fraction of the investment per time period.

The rebalancers act to keep a portfolio structure with (for instance) half of their wealth in cash and half in stocks. If the stock price rises, then the stocks weight in the portfolio will increase and the rebalancers will sell shares until the shares again constitute 50% of the portfolio. If the stock price decreases, then the value of the shares in the portfolio decreases, and the rebalancers will buy shares until the stock again constitutes 50% of the portfolio. Thus, the rebalancers have a stabilizing influence on the market by selling when the market rises and buying when the market falls.

A typical CPPI investor has as his/her main objective not to lose more than (for instance) 25% of his initial wealth during a quarter, which consists of 65 trading days. Thus, he aims to insure that at each cycle 75% of the initial wealth is out of reasonable risk. To this effect, he assumes that the current value of the stock will not fall in one day by more than a factor of 2. The result is that he always keeps in stock twice the difference between the present wealth and 75% of the initial wealth (which he had at the beginning of the 65 days investing period). This determines the amount the CPPI agent is bidding or offering at each stage. Obviously, after a price fall, the amount he wants to keep in stocks will fall and the CPPI investor will sell and further destabilize the market. After an increase in the prices (and personal wealth) the amount the CPPI agent wants to keep in shares will increase: he will buy, and may support a price bubble.

The simulations reveal that even a relatively small fraction of CPPI investors (i. e. less than 50%) is enough to destabilize the market, and crashes and booms are observed. Hence, the claim of Kim and Markowitz that the CPPI policy may be responsible for the 1987 crash is supported by the agent based simulations. Various variants of this model were studied intensively by Egenter, Lux and Stauffer [5] who find that the price time evolution becomes unrealistically periodic for a large number of investors (the periodicity seems related with the fixed 65 days quarter and is significantly diminished if the 65 day period begins on a different date for each investor).

The Arthur, Holland, Lebaron, Palmer and Tayler Stock Market Model

Palmer, Arthur, Holland, Lebaron and Tayler [30] and Arthur, Holland, Lebaron, Palmer and Tayler [3] (AHLPT) construct an agent based simulation model that is focused on the concept of co-evolution. Each investor adapts his/her investment strategy such as to maximally exploit the market dynamics generated by the investment strategies of all others investors. This leads to an ever-evolving market, driven endogenously by the ever-changing strategies of the investors.

The main objective of AHLPT is to prove that market fluctuations may be induced by this endogenous co-evolution, rather than by exogenous events. Moreover, AHLPT study the various regimes of the system: the regime in which rational fundamentalist strategies are dominating vs. the regime in which investors start developing strategies based on technical trading. In the technical trading regime, if some of the investors follow fundamentalist strategies, they may be punished rather than rewarded by the market. AHLPT also study the relation between the various strategies (fundamentals vs. technical) and the volatility properties of the market (clustering, excess volatility, volume-volatility correlations, etc.).

In the first paper quoted above, the authors simulated a single stock and further limited the bid/offer decision to a ternary choice of: i) bid to buy one share, ii) offer to sell one share, or: iii) do nothing. Each agent had a collection of rules which described how he should behave (i, ii or iii) in various market conditions. If the current market conditions were not covered by any of the rules, the default was to do nothing. If more than one rule applied in a certain market condition, the rule to act upon was chosen probabilistically according to the "strengths" of the applicable rules. The "strength" of each rule was determined according to the rule's past performance: rules that "worked" became "stronger". Thus, if a certain rule performed well, it became more likely to be used again.

The price is updated proportionally to the relative excess of offers over demands. In [3], the rules were used to predict future prices. The price prediction was then transformed into a buy/sell order through the use of a Constant Absolute Risk Aversion (CARA) utility function. The use of CARA utility leads to demands which do not depend on the investor's wealth.

The heart of the AHLPT dynamics are the trading rules. In particular, the authors differentiate between "fundamental" rules and "technical" rules, and study their relative strength in various market regimes. For instance, a "fundamental" rule may require a market conditions of the type:

$$\text{dividend/current price} > 0.04$$

in order to be applied. A "technical" rule may be triggered if the market fulfills a condition of the type:

$$\text{current price} > \text{10-period moving-average of past prices.}$$

The rules undergo genetic dynamics: the weakest rules are substituted periodically by copies of the strongest rules and all the rules undergo random mutations (or even versions of "sexual" crossovers: new rules are formed by combining parts from 2 different rules). The genetic dynamics of the trading rules represent investors' learning: new rules represent new trading strategies. Investors examine new strategies, and adopt those which tend to work best. The main results of this model are:

For a Few Agents, a Small Number of Rules, and Small Dividend Changes

- The price converges towards an equilibrium price which is close to the fundamental value.
- Trading volume is low.
- There are no bubbles, crashes or anomalies.
- Agents follow homogeneous simple fundamentalist rules.

For a Large Number of Agents and a Large Number of Rules

- There is no convergence to an equilibrium price, and the dynamics are complex.
- The price displays occasional large deviations from the fundamental value (bubbles and crashes).
- Some of these deviations are triggered by the emergence of collectively self-fulfilling agent price-prediction rules.
- The agents become heterogeneous (adopt very different rules).
- Trading volumes fluctuate (large volumes correspond to bubbles and crashes).
- The rules evolve over time to more and more complex patterns, organized in hierarchies (rules, exceptions to rules, exceptions to exceptions, and so on ...).
- The successful rules are time dependent: a rule which is successful at a given time may perform poorly if reintroduced after many cycles of market co-evolution.

The Lux and Lux and Marchesi Model

Lux [27] and Lux and Marchesi [28] propose a model to endogenously explain the heavy tail distribution of returns and the clustering of volatility. Both of these phenomena emerge in the Lux model as soon as one assumes that in addition to the fundamentalists there are also chartists in the model. Lux and Marchesi [28] further divide the chartists into optimists (buyers) and pessimists (sellers). The market fluctuations are driven and amplified by the fluctuations in the various populations: chartists converting into fundamentalists, pessimists into optimists, etc.

In the Lux and Marchesi model the stock's fundamental value is exogenously determined. The fluctuations of the fundamental value are inputted exogenously as a white

noise process in the logarithm of the value. The market price is determined by investors' demands and by the market clearance condition.

Lux and Marchesi consider three types of traders:

- **Fundamentalists** observe the fundamental value of the stock. They anticipate that the price will eventually converge to the fundamental value, and their demand for shares is proportional to the difference between the market price and the fundamental value.
- **Chartists** look more at the present trends in the market price rather than at fundamental economic values; the chartists are divided into
- **Optimists** (they buy a fixed amount of shares per unit time)
- **Pessimists** (they sell shares).

Transitions between these three groups (optimists, pessimists, fundamentalists) happen with probabilities depending on the market dynamics and on the present numbers of traders in each of the three classes:

- **The transition probabilities of chartists** depend on the majority opinion (through an "opinion index" measuring the relative number of optimists minus the relative number of pessimists) and on the actual price trend (the current time derivative of the current market price), which determines the relative profit of the various strategies.
- **The fundamentalists decide to turn into chartists** if the profits of the later become significantly larger than their own, and vice versa (the detailed formulae used by Lux and Marchesi are inspired from the exponential transition probabilities governing statistical mechanics physical systems).

  The main results of the model are:
  - No long-term deviations between the current market price and the fundamental price are observed.
  - The deviations from the fundamental price, which do occur, are unsystematic.
  - In spite of the fact that the variations of the fundamental price are normally distributed, the variations of the market price (the market returns) are not. In particular the returns exhibit a frequency of extreme events which is higher than expected for a normal distribution. The authors emphasize the amplification role of the market that transforms the input normal distribution of the fundamental value variations into a leptokurtotic (heavy tailed) distribution of price variation, which is encountered in the actual financial data.
  - clustering of volatility.

The authors explain the volatility clustering (and as a consequence, the leptokurticity) by the following mechanism. In periods of high volatility, the fundamental information is not very useful to insure profits, and a large fraction of the agents become chartists. The opposite is true in quiet periods when the actual price is very close to the fundamental value. The two regimes are separated by a threshold in the number of chartist agents. Once this threshold is approached (from below) large fluctuations take place which further increase the number of chartists. This destabilization is eventually dampened by the energetic intervention of the fundamentalists when the price deviates too much from the fundamental value. The authors compare this temporal instability with the on-off intermittence encountered in certain physical systems. According to Egenter et al. [5], the fraction of chartists in the Lux Marchesi model goes to zero as the total number of traders goes to infinity, when the rest of the parameters are kept constant.

Illustration with the LLS Model

The purpose of this section is to give a more detailed "hands on" example of the agent based approach, and to discuss some of the practical dilemmas arising when implementing this approach, by focusing on one specific model. We will focus on the so called LLS Model of the stock market (for more detail, and various versions of the model, see [11,17,22,23,24,25]. This section is based on the presentation of the LLS Model in Chap. 7 of [24]).

Background

Real life investors differ in their investment behavior from the investment behavior of the idealized representative rational investor assumed in most economic and financial models. Investors differ one from the other in their preferences, their investment horizon, the information at their disposal, and their interpretation of this information. No financial economist seriously doubts these observations. However, modeling the empirically and experimentally documented investor behavior and the heterogeneity of investors is very difficult and in most cases practically impossible to do within an analytic framework. For instance, the empirical and experimental evidence suggests that most investors are characterized by Constant Relative Risk Aversion (CRRA), which implies a power (myopic) utility function (see Eq. (2) below). However, for a general distribution of returns it is impossible to obtain an analytic solution for the portfolio optimization problem of investors with these preferences. Extrapolation of future returns from past returns, biased probability weighting, and partial deviations from rationality are also all ex-

perimentally documented but difficult to incorporate in an analytical setting. One is then usually forced to make the assumptions of rationality and homogeneity (at least in some dimension) and to make unrealistic assumptions regarding investors' preferences, in order to obtain a model with a tractable solution. The hope in these circumstances is that the model will capture the essence of the system under investigation, and will serve as a useful benchmark, even though some of the underlying assumptions are admittedly false.

Most homogeneous rational agent models lead to the following predictions: no trading volume, zero autocorrelation of returns, and price volatility which is equal to or lower than the volatility of the "fundamental value" of the stock (defined as the present value of all future dividends, see [37]). However, the empirical evidence is very different:

- Trading volume can be extremely heavy [1,14].
- Stock returns exhibit short-run momentum (positive autocorrelation) and long-run mean reversion (negative autocorrelation) [7,13,21,31].
- Stock returns are excessively volatile relative to the dividends [37].

As most standard rational-representative-agent models cannot explain these empirical findings, these phenomena are known as "anomalies" or "puzzles". Can these "anomalies" be due to elements of investors' behavior which are unmodeled in the standard rational-representative-agent models, such as the experimentally documented deviations of investors' behavior from rationality and/or the heterogeneity of investors? The agent based simulation approach offers us a tool to investigate this question. The strength of the agent based simulation approach is that since it is not restricted to the scope of analytical methods, one is able to investigate virtually any imaginable investor behavior and market structure. Thus, one can study models which incorporate the experimental findings regarding the behavior of investors, and evaluate the effects of various behavioral elements on market dynamics and asset pricing.

The LLS model incorporates some of the main empirical findings regarding investor behavior, and we employ this model in order to study the effect of each element of investor behavior on asset pricing and market dynamics. We start out with a benchmark model in which all of the investors are rational, informed and identical, and then, one by one, we add elements of heterogeneity and deviations from rationality to the model in order to study their effects on the market dynamics.

In the benchmark model all investors are Rational, Informed and Identical (RII investors). This is, in effect, a "representative agent" model. The RII investors are informed about the dividend process, and they rationally act to maximize their expected utility. The RII investors make investment decisions based on the present value of future cash flows. They are essentially fundamentalists who evaluate the stock's fundamental value and try to find bargains in the market. The benchmark model in which all investors are RII yields results which are typical of most rational-representative-agent models: in this model prices follow a random walk, there is no excess volatility of the prices relative to the volatility of the dividend process, and since all agents are identical, there is no trading volume.

After describing the properties of the benchmark model, we investigate the effects of introducing various elements of investor behavior which are found in laboratory experiments but are absent in most standard models. We do so by adding to the model a minority of investors who do not operate like the RII investors. These investors are Efficient Market Believers (EMB from now on). The EMBs are investors who believe that the price of the stock reflects all of the currently available information about the stock. As a consequence, they do not try to time the market or to buy bargain stocks. Rather, their investment decision is reduced to the optimal diversification problem. For this portfolio optimization, the *ex-ante* return distribution is required. However, since the *ex-ante* distribution is unknown, the EMB investors use the *ex-post* distribution in order to estimate the *ex-ante* distribution. It has been documented that in fact, many investors form their expectations regarding the future return distribution based on the distribution of past returns.

There are various ways to incorporate the investment decisions of the EMBs. This stems from the fact that there are different ways to estimate the *ex-ante* distribution from the *ex-post* distribution. How far back should one look at the historical returns? Should more emphasis be given to more recent returns? Should some "outlier" observations be filtered out? etc. Of course, there are no clear answers to these questions, and different investors may have different ways of forming their estimation of the *ex-ante* return distribution (even though they are looking at the same series of historical returns). Moreover, some investors may use the objective *ex-post* probabilities when constructing their estimation of the *ex-ante* distribution, whereas others may use biased subjective probability weights. In order to build the analysis step-by-step we start by analyzing the case in which the EMB population is homogeneous, and then introduce various forms of heterogeneity into this population.

An important issue in market modeling is that of the degree of investors' rationality. Most models in economics and finance assume that people are fully rational. This assumption usually manifests itself as the maximization of an expected utility function by the individual. However, numerous experimental studies have shown that people deviate from rational decision-making [41,42,43,44,45]. Some studies model deviations from the behavior of the rational agent by introducing a sub-group of liquidity, or "noise", traders. These are traders that buy and sell stocks for reasons that are not directly related to the future payoffs of the financial asset - their motivation to trade arises from outside of the market (for example, a "noise trader's" daughter unexpectedly announces her plans to marry, and the trader sells stocks because of this unexpected need for cash). The exogenous reasons for trading are assumed random, and thus lead to random or "noise" trading (see [10]). The LLS model takes a different approach to the modeling of noise trading. Rather than dividing investors into the extreme categories of "fully rational" and "noise traders", the LLS model assumes that most investors try to act as rationally as they can, but are influenced by a multitude of factors causing them to deviate to some extent from the behavior that would have been optimal from their point of view. Namely, all investors are characterized by a utility function and act to maximize their expected utility; however, some investors may deviate to some extent from the optimal choice which maximizes their expected utility. These deviations from the optimal choice may be due to irrationality, inefficiency, liquidity constraints, or a combination of all of the above.

In the framework of the LLS model we examine the effects of the EMBs' deviations from rationality and their heterogeneity, relative to the benchmark model in which investors are informed, rational and homogeneous. We find that the behavioral elements which are empirically documented, namely, extrapolation from past returns, deviation from rationality, and heterogeneity among investors, lead to all of the following empirically documented "puzzles":

- Excess volatility
- Short-term momentum
- Longer-term return mean-reversion
- Heavy trading volume
- Positive correlation between volume and contemporaneous absolute returns
- Positive correlation between volume and lagged absolute returns

The fact that all these anomalies or "puzzles", which are hard to explain with standard rational-representative-agent models, are generated naturally by a simple model which incorporates the experimental findings regarding investor behavior and the heterogeneity of investors, leads one to suspect that these behavioral elements and the diversity of investors are a crucial part of the workings of the market, and as such they cannot be "assumed away". As the experimentally documented bounded-rational behavior and heterogeneity are in many cases impossible to analyze analytically, agent based simulation presents a very promising tool for investigating market models incorporating these elements.

The LLS Model

The stock market consists of two investment alternatives: a stock (or index of stocks) and a bond. The bond is assumed to be a riskless asset, and the stock is a risky asset. The stock serves as a proxy for the market portfolio (e. g., the Standard & Poors 500 index). The extension from one risky asset to many risky assets is possible; however, one stock (the index) is sufficient for our present analysis because we restrict ourselves to global market phenomena and do not wish to deal with asset allocation across several risky assets. Investors are allowed to revise their portfolio at given time points, i. e. we discuss a discrete time model.

The bond is assumed to be a riskless investment yielding a constant return at the end of each time period. The bond is in infinite supply and investors can buy from it as much as they wish at a given rate of r_f. The stock is in finite supply. There are N outstanding shares of the stock. The return on the stock is composed of two elements:

a) Capital Gain: If an investor holds a stock, any rise (fall) in the price of the stock contributes to an increase (decrease) in the investor's wealth.

b) Dividends: The company earns income and distributes dividends at the end of each time period. We denote the dividend per share paid at time t by D_t. We assume that the dividend is a stochastic variable following a multiplicative random walk, i. e., $\tilde{D}_t = D_{t-1}(1+\tilde{z})$, where $\tilde{z}$ is a random variable with some probability density function $f(z)$ in the range $[z_1, z_2]$. (In order to allow for a dividend cut as well as a dividend increase we typically choose: $z_1 < 0, z_2 > 0$).

The total return on the stock in period t, which we denote by R_t is given by:

$$\tilde{R}_t = \frac{\tilde{P}_t + \tilde{D}_t}{P_{t-1}} , \tag{1}$$

where $\tilde{P}_t$ is the stock price at time t.

All investors in the model are characterized by a von Neuman-Morgenstern utility function. We assume that all

investors have a power utility function of the form:

$$U(W) = \frac{W^{1-\alpha}}{1-\alpha}\,, \tag{2}$$

where α is the risk aversion parameter. This form of utility function implies Constant Relative Risk Aversion (CRRA). We employ the power utility function (Eq. (2)) because the empirical evidence suggests that relative risk aversion is approximately constant (see, for example [8,9,18,20]), and the power utility function is the unique utility function which satisfies the CRRA condition. Another implication of CRRA is that the optimal investment choice is independent of the investment horizon [33,34]. In other words, regardless of investors' actual investment horizon, they choose their optimal portfolio as though they are investing for a single period. The myopia property of the power utility function simplifies our analysis, as it allows us to assume that investors maximize their one-period-ahead expected utility.

We model two different types of investors: Rational, Informed, Identical (RII) investors, and Efficient Market Believers (EMB). These two investor types are described below.

Rational Informed Identical (RII) Investors RII investors evaluate the "fundamental value" of the stock as the discounted stream of all future dividends, and thus can also be thought of as "fundamentalists". They believe that the stock price may deviate from the fundamental value in the short run, but if it does, it will eventually converge to the fundamental value. The RII investors act according to the assumption of asymptotic convergence: if the stock price is low relative to the fundamental value they buy in anticipation that the underpricing will be corrected, and vice versa. We make the simplifying assumption that the RII investors believe that the convergence of the price to the fundamental value will occur in the next period, however, our results hold for the more general case where the convergence is assumed to occur some T periods ahead, with $T > 1$.

In order to estimate next period's return distribution, the RII investors need to estimate the distribution of next period's price, $\tilde{P}_{t+1}$, and of next period's dividend, $\tilde{D}_{t+1}$. Since they know the dividend process, the RII investors know that $\tilde{D}_{t+1} = D_t(1+\tilde{z})$ where $\tilde{z}$ is distributed according to $f(z)$ in the range $[z_1, z_2]$. The RII investors employ Gordon's dividend stream model in order to calculate the fundamental value of the stock:

$$P^f_{t+1} = \frac{E_{t+1}[\tilde{D}_{t+2}]}{k-g}\,, \tag{3}$$

where the superscript f stands for the *fundamental* value, $E_{t+1}[\tilde{D}_{t+2}]$ is the dividend corresponding to time $t+2$ as expected at time $t+1$, k is the discount factor or the expected rate of return demanded by the market for the stock, and g is the *expected* growth rate of the dividend, i. e., $g = E(\tilde{z}) = \int_{z_1}^{z_2} f(z)z\mathrm{d}z$.

The RII investors believe that the stock price may temporarily deviate from the fundamental value; however, they also believe that the price will eventually converge to the fundamental value. For simplification we assume that the RII investors believe that the convergence to the fundamental value will take place next period. Thus, the RII investors estimate P_{t+1} as:

$$P_{t+1} = P^f_{t+1}\,.$$

The expectation at time $t+1$ of $\tilde{D}_{t+2}$ depends on the realized dividend observed at $t+1$:

$$E_{t+1}[\tilde{D}_{t+2}] = D_{t+1}(1+g)\,.$$

Thus, the RII investors believe that the price at $t+1$ will be given by:

$$P_{t+1} = P^f_{t+1} = \frac{D_{t+1}(1+g)}{k-g}\,.$$

At time t, D_t is known, but D_{t+1} is not; therefore P^f_{t+1} is also not known with certainty at time t. However, given D_t, the RII investors know the distribution of $\tilde{D}_{t+1}$:

$$\tilde{D}_{t+1} = D_t(1+\tilde{z}),$$

where $\tilde{z}$ is distributed according to the known $f(z)$. The realization of $\tilde{D}_{t+1}$ determines P^f_{t+1}. Thus, at time t, RII investors believe that P_{t+1} is a random variable given by:

$$\tilde{P}_{t+1} = \tilde{P}^f_{t+1} = \frac{D_t(1+\tilde{z})(1+g)}{k-g}.$$

Notice that the RII investors face uncertainty regarding next period's price. In our model we assume that the RII investors are certain about the dividend growth rate g, the discount factor k, and the fact that the price will converge to the fundamental value next period. In this framework the only source of uncertainty regarding next period's price stems from the uncertainty regarding next period's dividend realization. More generally, the RII investors' uncertainty can result from uncertainty regarding any one of the above factors, or a combination of several of these factors. Any mix of these uncertainties is possible to investigate in the agent based simulation framework, but very hard, if not impossible, to incorporate in an analytic

framework. As a consequence of the uncertainty regarding next period's price and of their risk aversion, the RII investors do not buy an infinite number of shares even if they perceive the stock as underpriced. Rather, they estimate the stock's next period's return distribution, and find the optimal mix of the stock and the bond which maximizes their expected utility. The RII investors estimate next period's return on the stock as:

$$\tilde{R}_{t+1} = \frac{\tilde{P}_{t+1} + \tilde{D}_{t+1}}{P_t} = \frac{\frac{D_t(1+\tilde{z})(1+g)}{k-g} + D_t(1+\tilde{z})}{P_t}, \quad (4)$$

where $\tilde{z}$, the next year growth in the dividend, is the source of uncertainty. The demands of the RII investors for the stock depend on the price of the stock. For any *hypothetical* price P_h investors calculate the proportion of their wealth x they should invest in the stock in order to maximize their expected utility. The RII investor i believes that if she invests a proportion x of her wealth in the stock at time t, then at time $t+1$ her wealth will be:

$$\tilde{W}^i_{t+1} = W^i_h[(1-x)(1+r_f) + x\tilde{R}_{t+1}], \quad (5)$$

where $\tilde{R}_{t+1}$ is the return on the stock, as given by Eq. (1), and W^i_h is the wealth of investor i at time t given that the stock price at time t is P_h.

If the price in period t is the hypothetical price P_h, the $t+1$ expected utility of investor i is the following function of her investment proportion in the stock, x:

$$EU(\tilde{W}^i_{t+1}) = EU\left(W^i_h\left[(1-x)(1+r_f) + x\tilde{R}_{t+1}\right]\right). \quad (6)$$

Substituting $\tilde{R}_{t+1}$ from Eq. (4), using the power utility function (Eq. (2)), and substituting the hypothetical price P_h for P_t, the expected utility becomes the following function of x:

$$EU(\tilde{W}^i_{t+1}) = \frac{(W^i_h)^{1-\alpha}}{1-\alpha}\int_{z_1}^{z_2}\left[(1-x)(1+r_f) + x\left(\frac{\frac{D_t(1+z)(1+g)}{k-g} + D_t(1+z)}{P_h}\right)\right]^{1-\alpha} f(z)\mathrm{d}z\,, \quad (7)$$

where the integration is over all possible values of z. In the agent based simulation framework, this expression for the expected utility, and the optimal investment proportion x, can be solved numerically for any general choice of distribution $f(z)$. For the sake of simplicity we restrict the present analysis to the case where $\tilde{z}$ is distributed uniformly in the range $[z_1, z_2]$. This simplification leads to the following expression for the expected utility:

$$\begin{aligned} EU(\tilde{W}^i_{t+1}) &= \frac{(W^i_h)^{1-\alpha}}{(1-\alpha)(2-\alpha)}\frac{1}{(z_2-z_1)}\left(\frac{k-g}{k+1}\right)\frac{P_h}{xD_t} \\ &\left\{\left[(1-x)(1+r_f) + \frac{x}{P_h}\left(\frac{k+1}{k-g}\right)D_t(1+z_2)\right]^{(2-\alpha)}\right. \\ &\left. -\left[(1-x)(1+r_f) + \frac{x}{P_h}\left(\frac{k+1}{k-g}\right)D_t(1+z_1)\right]^{(2-\alpha)}\right\} \end{aligned} \quad (8)$$

For any hypothetical price P_h, each investor (numerically) finds the optimal proportion x_h which maximizes his/her expected utility given by Eq. (8). Notice that the optimal proportion, x_h, is independent of the wealth, W^i_h. Thus, if all RII investors have the same degree of risk aversion, α, they will have the same optimal investment proportion in the stock, regardless of their wealth. The number of shares demanded by investor i at the hypothetical price P_h is given by:

$$N^i_h(P_h) = \frac{x^i_h(P_h)W^i_h(P_h)}{P_h}. \quad (9)$$

Efficient Market Believers (EMB) The second type of investors in the LLS model are EMBs. The EMBs believe in market efficiency - they believe that the stock price accurately reflects the stock's fundamental value. Thus, they do not try to time the market or to look for "bargain" stocks. Rather, their investment decision is reduced to the optimal diversification between the stock and the bond. This diversification decision requires the *ex-ante* return distribution for the stock, but as the *ex-ante* distribution is not available, the EMBs assume that the process generating the returns is fairly stable, and they employ the *ex-post* distribution of stock returns in order to estimate the *ex-ante* return distribution.

Different EMB investors may disagree on the optimal number of *ex-post* return observations that should be employed in order to estimate the *ex-ante* return distribution. There is a trade-off between using more observations for better statistical inference, and using a smaller number of only more recent observations, which are probably more representative of the *ex-ante* distribution. As in reality, there is no "recipe" for the optimal number of observations to use. EMB investor i believes that the m^i most recent returns on the stock are the best estimate of the *ex-ante* distribution. Investors create an estimation of the *ex-ante* return distribution by assigning an equal probability

to each of the m^i most recent return observations:

$$Prob^i(\tilde{R}_{t+1} = R_{t-j}) = \frac{1}{m^i} \quad \text{for } j = 1, \ldots, m^i \tag{10}$$

The expected utility of EMB investor i is given by:

$$EU(W^i_{t+1}) = \frac{(W^i_h)^{1-\alpha}}{(1-\alpha)} \frac{1}{m^i} \sum_{j=1}^{m^i} \left[(1-x)(1+r_f) + xR_{t-j}\right]^{1-\alpha}, \tag{11}$$

where the summation is over the set of m^i most recent *ex-post* returns, x is the proportion of wealth invested in the stock, and as before W^i_h is the wealth of investor i at time t given that the stock price at time t is P_h. Notice that W^i_h does not change the optimal diversification policy, i. e., x. Given a set of m^i past returns, the optimal portfolio for the EMB investor i is an investment of a proportion x^{*i} in the stock and (1- x^{*i}) in the bond, where x^{*i} is the proportion which maximizes the above expected utility (Eq. (11)) for investor i. Notice that x^{*i} generally cannot be solved for analytically. However, in the agent based simulation framework this does not constitute a problem, as one can find x^{*i} numerically.

Deviations from Rationality Investors who are efficient market believers, and are rational, choose the investment proportion x^* which maximizes their expected utility. However, many empirical studies have shown that the behavior of investors is driven not only by rational expected utility maximization but by a multitude of other factors (see, for example, [34,41,42,43,44]). Deviations from the optimal rational investment proportion can be due to the cost of resources which are required for the portfolio optimization: time, access to information, computational power, etc., or due to exogenous events (for example, an investor plans to revise his portfolio, but gets distracted because his car breaks down). We assume that the different factors causing the investor to deviate from the optimal investment proportion x^* are random and uncorrelated with each other. By the central limit theorem, the aggregate effect of a large number of random uncorrelated influences is a normally distributed random influence, or "noise". Hence, we model the effect of all the factors causing the investor to deviate from his optimal portfolio by adding a normally distributed random variable to the optimal investment proportion. To be more specific, we assume:

$$x^i = x^{*i} + \tilde{\varepsilon}^i, \tag{12}$$

where $\tilde{\varepsilon}^i$ is a random variable drawn from a truncated normal distribution with mean zero and standard deviation σ. Notice that noise is investor-specific, thus, $\tilde{\varepsilon}^i$ is drawn separately and independently for each investor.

The noise can be added to the decision-making of the RII investors, the EMB investors, or to both. The results are not much different with these various approaches. Since the RII investors are taken as the benchmark of rationality, in this chapter we add the noise only to the decision-making of the EMB investors.

Market Clearance The number of shares demanded by each investor is a monotonically decreasing function of the hypothetical price P_h (see [24]). As the total number of outstanding shares is N, the price of the stock at time t is given by the market clearance condition: P_t is the unique price at which the total demand for shares is equal to the total supply, N:

$$\sum_i N^i_h(P_t) = \sum_i \frac{x_h(P_t) W^i_h(P_t)}{P_t} = N\,, \tag{13}$$

where the summation is over all the investors in the market, RII investors as well as EMB investors.

Agent Based Simulation The market dynamics begin with a set of initial conditions which consist of an initial stock price P_0, an initial dividend D_0, the wealth and number of shares held by each investor at time $t = 0$, and an initial "history" of stock returns. As will become evident, the general results do not depend on the initial conditions. At the first period ($t = 1$), interest is paid on the bond, and the time 1 dividend $\tilde{D}_1 = D_0(1 + \tilde{z})$ is realized and paid out. Then investors submit their demand orders, $N^i_h(P_h)$, and the market clearing price P_1 is determined. After the clearing price is set, the new wealth and number of shares held by each investor are calculated. This completes one time period. This process is repeated over and over, as the market dynamics develop.

We would like to stress that even the simplified benchmark model, with only RII investors, is impossible to solve analytically. The reason for this is that the optimal investment proportion, $x_h(P_h)$, cannot be calculated analytically. This problem is very general and it is encountered with almost any choice of utility function and distribution of returns. One important exception is the case of a negative exponential utility function and normally distributed returns. Indeed, many models make these two assumptions for the sake of tractability. The problem with the assumption of negative exponential utility is that it implies Constant Absolute Risk Aversion (CARA), which is

very unrealistic, as it implies that investors choose to invest the same dollar amount in a risky prospect *independent of their wealth*. This is not only in sharp contradiction to the empirical evidence, but also excludes the investigation of the two-way interaction between wealth and price dynamics, which is crucial to the understanding of the market.

Thus, one contribution of the agent based simulation approach is that it allows investigation of models with realistic assumptions regarding investors' preferences. However, the main contribution of this method is that it permits us to investigate models which are much more complex (and realistic) than the benchmark model, in which all investors are RII. With the agent based simulation approach one can study models incorporating the empirically and experimentally documented investors' behavior, and the heterogeneity of investors.

Results of the LLS Model

We begin by describing the benchmark case where all investors are rational and identical. Then we introduce to the market EMB investors and investigate their affects on the market dynamics.

Benchmark Case: Fully Rational and Identical Agents In this benchmark model all investors are RII: rational, informed and identical. Thus, it is not surprising that the benchmark model generates market dynamics which are typical of homogeneous rational agent models:

No Volume All investors in the model are identical; they therefore always agree on the optimal proportion to invest in the stock. As a consequence, all the investors always achieve the same return on their portfolio. This means that at any time period the ratio between the wealth of any two investors is equal to the ratio of their initial wealths, i. e.:

$$\frac{W_t^i}{W_t^j} = \frac{W_0^i}{W_0^j} \,. \tag{14}$$

As the wealth of investors is always in the same proportion, and as they always invest the same fraction of their wealth in the stock, the number of shares held by different investors is also always in the same proportion:

$$\frac{N_t^i}{N_t^j} = \frac{\frac{x_t W_t^i}{P_t}}{\frac{x_t W_t^j}{P_t}} = \frac{W_t^i}{W_t^j} = \frac{W_0^i}{W_0^j} \,. \tag{15}$$

Since the total supply of shares is constant, this implies that each investor always holds the same number of shares, and there is *no trading volume* (the number of shares held may vary from one investor to the other as a consequence of different initial endowments).

Log-Prices Follow a Random Walk In the benchmark model all investors believe that next period's price will converge to the fundamental value given by the discounted dividend model (Eq. (3)). Therefore, the actual stock price is always close to the fundamental value. The fluctuations in the stock price are driven by fluctuations in the fundamental value, which in turn are driven by the fluctuating dividend realizations. As the dividend fluctuations are (by assumption) uncorrelated over time, one would expect that the price fluctuations will also be uncorrelated. To verify this intuitive result, we examine the return autocorrelations in simulations of the benchmark model.

Let us turn to the simulation of the model. We first describe the parameters and initial conditions used in the simulation, and then report the results. We simulate the benchmark model with the following parameters:

- Number of investors = 1000
- Risk aversion parameter α= 1.5. This value roughly conforms with the estimate of the risk aversion parameter found empirically and experimentally.
- Number of shares = 10,000.
- We take the time period to be a quarter, and accordingly we choose:
- Riskless interest rate $r_f = 0.01$.
- Required rate of return on stock k = 0.04.
- Maximal one-period dividend decrease z_1 = -0.07.
- Maximal one-period dividend growth z_2 = 0.10.
- $\tilde{z}$ is uniformly distributed between these values. Thus, the average dividend growth rate is $g = (z_1 + z_2)/2 = 0.015$.

Initial Conditions: Each investor is endowed at time $t = 0$ with a total wealth of \$1000, which is composed of 10 shares worth an initial price of \$50 per share, and \$500 in cash. The initial quarterly dividend is set at \$0.5 (for an annual dividend yield of about 4%). As will soon become evident, the dynamics are not sensitive to the particular choice of initial conditions.

Figure 1 shows the price dynamics in a typical simulation with these parameters (simulations with the same parameters differ one from the other because of the different random dividend realizations). Notice that the vertical axis in this figure is logarithmic. Thus, the roughly constant slope implies an approximately exponential price growth, or an approximately constant *average* return.

The prices in this simulation seem to fluctuate randomly around the trend. However, Fig. 1 shows only one simulation. In order to have a more rigorous analysis we

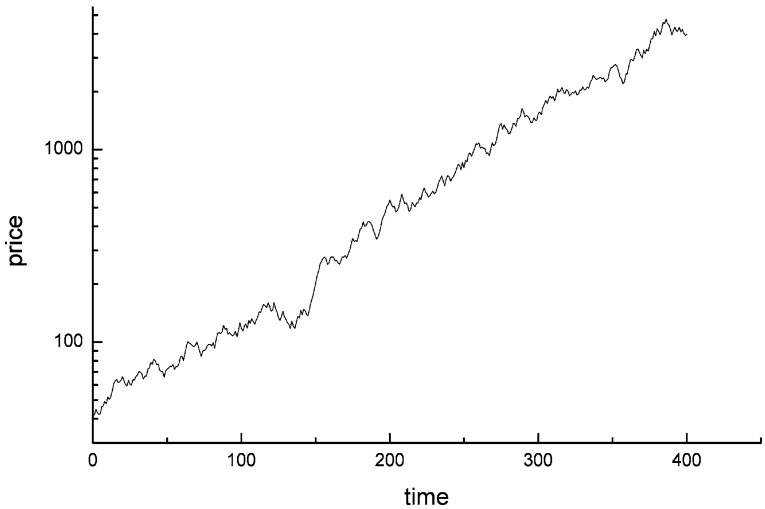

Agent Based Computational Economics, Figure 1
Price Dynamics in the Benchmark Model

perform many independent simulations, and employ statistical tools. Namely, for each simulation we calculate the autocorrelation of returns. We perform a univariate regression of the return in time t on the return on time $t-j$:

$$R_t = \alpha_j + \beta_j R_{t-j} + \varepsilon \,,$$

where R_t is the return in period t, and j is the lag. The autocorrelation of returns for lag j is defined as:

$$\rho_j = \frac{cov(R_t, R_{t-j})}{\hat{\sigma}^2(R)} \,,$$

and it is estimated by $\hat{\beta}$. We calculate the autocorrelation for different lags, $j = 1, \ldots .40$. Figure 2 shows the average autocorrelation as a function of the lag, calculated over 100 independent simulations. It is evident both from the figure that the returns are uncorrelated in the benchmark model, conforming with the random-walk hypothesis.

No Excess Volatility Since the RII investors believe that the stock price will converge to the fundamental value next period, in the benchmark model prices are always close to the fundamental value given by the discounted dividend stream. Thus, we do not expect prices to be more volatile than the value of the discounted dividend stream. For a formal test of excess volatility we follow the technique in [37]. For each time period we calculate the actual price P_t, and the fundamental value of discounted dividend stream, P_t^f, as in Eq. (3). Since prices follow an upward trend, in order to have a meaningful measure of the volatility, we must detrend these price series. Following Shiller, we run the regression:

$$\ln P_t = bt + c + \varepsilon_t \,, \tag{16}$$

in order to find the average exponential price growth rate (where b and c are constants). Then, we define the detrended price as: $p_t = P_t/e^{\hat{b}t}$. Similarly, we define the detrended value of the discounted dividend stream p_t^f, and compare $\sigma(p_t)$ with $\sigma(p_t^f)$. For 100 1000-period simulations we find an average $\sigma(p_t)$ of 22.4, and an average $\sigma(p_t^f)$ of 22.9. As expected, the actual price and the fundamental value have almost the same volatility.

To summarize the results obtained for the benchmark model, we find that when all investors are assumed to be rational, informed and identical, we obtain results which are typical of rational-representative-agent models: no volume, no return autocorrelations, and no excess volatility. We next turn to examine the effect of introducing into the market EMB investors, which model empirically and experimentally documented elements of investors' behavior.

The Introduction of a Small Minority of EMB Investors

In this section we will show that the introduction of a small minority of heterogeneous EMB investors generates many of the empirically observed market "anomalies" which are absent in the benchmark model, and indeed, in most other rational-representative-agent models. We take this as strong evidence that the "non-rational" elements of investor behavior which are documented in experimental studies, and the heterogeneity of investors, both of which

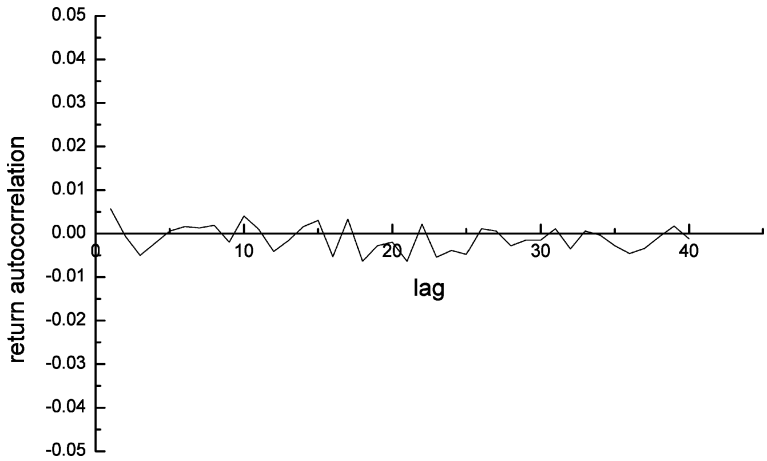

Agent Based Computational Economics, Figure 2
Return Autocorrelation in Benchmark Model

are incorporated in the LLS model, are crucial to understanding the dynamics of the market.

In presenting the results of the LLS model with EMB investors we take an incremental approach. We begin by describing the results of a model with a small sub-population of *homogeneous* EMB believers. This model produces the above mentioned market "anomalies"; however, it produces unrealistic cyclic market dynamics. Thus, this model is presented both for analyzing the source of the "anomalies" in a simplified setting, and as a reference point with which to compare the dynamics of the model with a *heterogeneous* EMB believer population.

We investigate the effects of investors' heterogeneity by first analyzing the case in which there are two types of EMBs. The two types differ in the method they use to estimate the *ex-ante* return distribution. Namely, the first type looks at the set of the last m_1 *ex-post* returns, whereas the second type looks at the set of the last m_2 *ex-post* returns. It turns out that the dynamics in this case are much more complicated than a simple "average" between the case where all EMB investors have m_1 and the case where all EMB investors have m_2. Rather, there is a complex non-linear interaction between the two EMB sub-populations. This implies that the heterogeneity of investors is a very important element determining the market dynamics, an element which is completely absent in representative-agent models.

Finally, we present the case where there is an entire spectrum of EMB investors differing in the number of *ex-post* observations they take into account when estimating the *ex-ante* distribution. This general case generates very realistic-looking market dynamics with all of the above mentioned market anomalies.

Homogeneous Sub-Population of EMBs When a very small sub-population of EMB investors is introduced to the benchmark LLS model, the market dynamics change dramatically. Figure 3 depicts a typical price path in a simulation of a market with 95% RII investors and 5% EMB investors. The EMB investors have $m = 10$ (i. e., they estimate the *ex-ante* return distribution by observing the set of the last 10 *ex-post* returns). σ, the standard deviation of the random noise affecting the EMBs' decision making is taken as 0.2. All investors, RII and EMB alike, have the same risk aversion parameter $\alpha = 1.5$ (as before). In the first 150 trading periods the price dynamics look very similar to the typical dynamics of the benchmark model. However, after the first 150 or so periods the price dynamics change. From this point onwards the market is characterized by periodic booms and crashes. Of course, Fig. 3 describes only one simulation. However, as will become evident shortly, different simulations with the same parameters may differ in detail, but the pattern is general: at some stage (not necessarily after 150 periods) the EMB investors induce cyclic price behavior. It is quite astonishing that such a small minority of only 5% of the investors can have such a dramatic impact on the market.

In order to understand the periodic booms and crashes let us focus on the behavior of the EMB investors. After every trade, the EMB investors revise their estimation of the *ex-ante* return distribution, because the set of *ex-post* returns they employ to estimate the *ex-ante* distribution changes. Namely, investors add the latest return generated by the stock to this set and delete the oldest return from this set. As a result of this update in the estimation of the *ex-ante* distribution, the optimal investment proportion x^* changes, and EMB investors revise their portfolios at

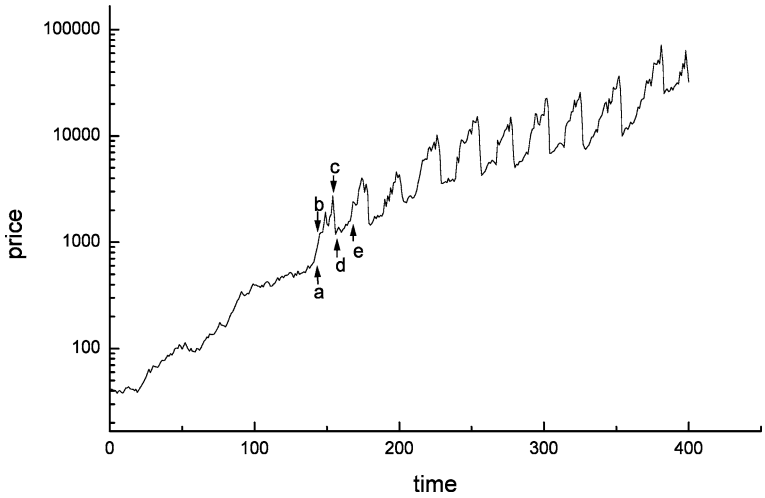

Agent Based Computational Economics, Figure 3
5% of Investors are Efficient Market Believers, 95% Rational Informed Investors

next period's trade. During the first 150 or so periods, the informed investors control the dynamics and the returns fluctuate randomly (as in the benchmark model). As a consequence, the investment proportion of the EMB investors also fluctuates irregularly. Thus, during the first 150 periods the EMB investors do not effect the dynamics much. However, at point **a** the dynamics change qualitatively (see Fig. 3). At this point, a relatively high dividend is realized, and as a consequence, a relatively high return is generated. This high return leads the EMB investors to increase their investment proportion in the stock at the next trading period. This increased demand of the EMB investors is large enough to effect next period's price, and thus a second high return is generated. Now the EMB investors look at a set of *ex-post* returns with two high returns, and they increase their investment proportion even further. Thus, a positive feedback loop is created.

Notice that as the price goes up, the informed investors realize that the stock is overvalued relative to the fundamental value P^f and they decrease their holdings in the stock. However, this effect does not stop the price increase and break the feedback loop because the EMB investors continue to buy shares aggressively. The positive feedback loop pushes the stock price further and further up to point **b**, at which the EMBs are invested 100% in the stock. At point **b** the positive feedback loop "runs out of gas". However, the stock price remains at the high level because the EMB investors remain fully invested in the stock (the set of past m=10 returns includes at this stage the very high returns generated during the "boom" – segment **a**–**b** in Fig. 3).

When the price is at the high level (segment **b**–**c**), the dividend yield is low, and as a consequence, the returns are generally low. As time goes by and we move from point **b** towards point **c**, the set of $m = 10$ last returns gets filled with low returns. Despite this fact, the extremely high returns generated in the boom are also still in this set, and they are high enough to keep the EMB investors fully invested. However, 10 periods after the boom, these extremely high returns are pushed out of the set of relevant *ex-post* returns. When this occurs, at point **c**, the EMB investors face a set of low returns, and they cut their investment proportion in the stock sharply. This causes a dramatic crash (segment **c**–**d**). Once the stock price goes back down to the "fundamental" value, the informed investors come back into the picture. They buy back the stock and stop the crash.

The EMB investors stay away from the stock as long as the *ex-post* return set includes the terrible return of the crash. At this stage the informed investors regain control of the dynamics and the stock price remains close to its fundamental value. 10 periods after the crash the extremely negative return of the crash is excluded from the *ex-post* return set, and the EMB investors start increasing their investment proportion in the stock (point **e**). This drives the stock price up, and a new boom-crash cycle is initiated. This cycle repeats itself over and over almost periodically.

Figure 3 depicts the price dynamics of a single simulation. One may therefore wonder how general the results discussed above are. Figure 4 shows two more simulations with the same parameters but different dividend realiza-

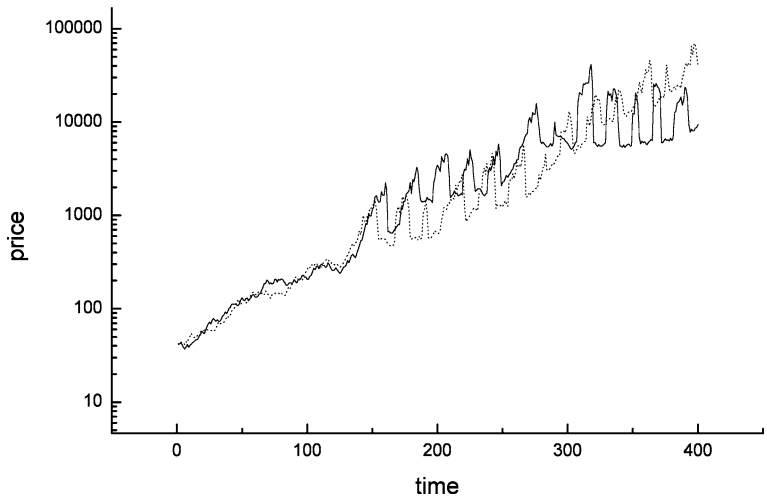

Agent Based Computational Economics, Figure 4
Two More Simulations – same Parameters as Fig. 3, Different Divident Realizations

tions. It is evident from this figure that although the simulations vary in detail (because of the different dividend realizations), the overall price pattern with periodic boom-crash cycles is robust.

Although these dynamics are very unrealistic in terms of the periodicity, and therefore the predictability of the price, they do shed light on the mechanism generating many of the empirically observed market phenomena. In the next section, when we relax the assumption that the EMB population is homogeneous with respect to m, the price is no longer cyclic or predictable, yet the mechanisms generating the market phenomena are the same as in this homogeneous EMB population case. The homogeneous EMB population case generates the following market phenomena:

Heavy Trading Volume As explained above, shares change hands continuously between the RII investors and the EMB investors. When a "boom" starts the RII investors observe higher ex-post returns and become more optimistic, while the RII investor view the stock as becoming overpriced and become more pessimistic. Thus, at this stage the EMBs buy most of the shares from the RIIs. When the stock crashes, the opposite is true: the EMBs are very pessimistic, but the RII investors buy the stock once it falls back to the fundamental value. Thus, there is substantial trading volume in this market. The average trading volume in a typical simulation is about 1000 shares per period, which are 10% of the total outstanding shares.

Autocorrelation of Returns The cyclic behavior of the price yields a very definite return autocorrelation pattern. The autocorrelation pattern is depicted graphically in Fig. 5. The autocorrelation pattern is directly linked to the length of the price cycle, which in turn are determined by m. Since the moving window of *ex-post* returns used to estimate the *ex-ante* distribution is $m = 10$ periods long, the price cycles are typically a little longer than 20 periods long: a cycle consists of the positive feedback loop (segment **a–b** in Fig. 3) which is about 2–3 periods long, the upper plateau (segment **b–c** in Fig. 3) which is about 10 periods long, the crash that occurs during one or two periods, and the lower plateau (segment **d–e** in Fig. 3) which is again about 10 periods long, for a total of about 23–25 periods. Thus, we expect positive autocorrelation for lags of about 23–25 periods, because this is the lag between one point and the corresponding point in the next (or previous) cycle. We also expect negative autocorrelation for lags of about 10–12 periods, because this is the lag between a boom and the following (or previous) crash, and vice versa. This is precisely the pattern we observe in Fig. 5.

Excess Volatility The EMB investors induce large deviations of the price from the fundamental value. Thus, price fluctuations are caused not only by dividend fluctuations (as the standard theory suggests) but also by the endogenous market dynamics driven by the EMB investors. This "extra" source of fluctuations causes the price to be more volatile than the fundamental value P^f.

Indeed, for 100 1000-period independent simulations with 5% EMB investors we find an average $\sigma(p_t)$ of 46.4, and an average $\sigma(p_t^f)$ of 30.6; i. e., we have excess volatility of about 50%.

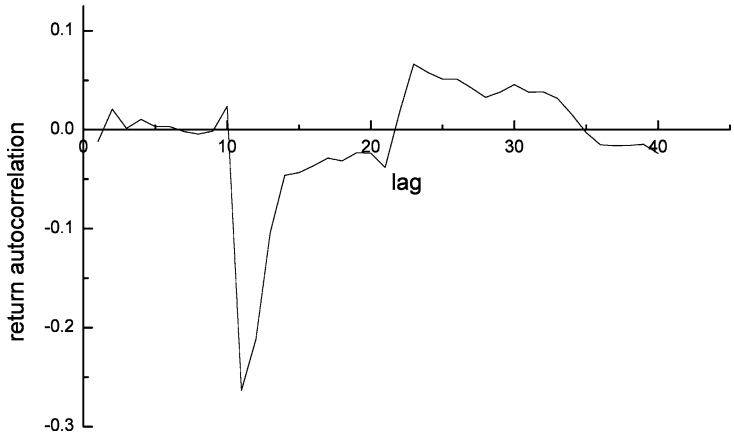

Agent Based Computational Economics, Figure 5
Return Autocorrelation 5%, Efficient Market Believers, $m = 10$

As a first step in analyzing the effects of heterogeneity of the EMB population, in the next section we examine the case of two types of EMB investors. We later analyze a model in which there is a full spectrum of EMB investors.

Two Types of EMBs One justification for using a representative agent in economic modeling is that although investors are heterogeneous in reality, one can model their collective behavior with one representative or "average" investor. In this section we show that this is generally not true. Many aspects of the dynamics result from the non-linear interaction between different investor types. To illustrate this point, in this section we analyze a very simple case in which there are only two types of EMB investors: one with $m = 5$ and the other with $m = 15$. Each of these two types consists of 2% of the investor population, and the remaining 96% are informed investors. The representative agent logic may tempt us to think that the resulting market dynamics would be similar to that of one "average" investor, i. e. an investor with $m = 10$. Figure 6 shows that this is clearly not the case. Rather than seeing periodic cycles of about 23–25 periods (which correspond to the average m of 10, as in Fig. 3), we see an irregular pattern. As before, the dynamics are first dictated by the informed investors. Then, at point **a**, the EMB investors with $m = 15$ induce cycles which are about 30 periods long. At point **b** there is a transition to shorter cycles induced by the $m = 5$ population, and at point **c** there is another transition back to longer cycles. What is going on?

These complex dynamics result from the non-linear interaction between the different sub-populations. The transitions from one price pattern to another can be partly understood by looking at the wealth of each sub-population. Figure 7 shows the proportion of the total wealth held by each of the two EMB populations (the remaining proportion is held by the informed investors). As seen in Fig. 7, the cycles which start at point **a** are dictated by the $m = 15$ rather than the $m = 5$ population, because at this stage the $m = 15$ population controls more of the wealth than the $m = 5$ population. However, after 3 cycles (at point **b**) the picture is reversed. At this point the $m = 5$ population is more powerful than the $m = 15$ population, and there is a transition to shorter boom-crash cycles. At point **c** the wealth of the two sub-populations is again almost equal, and there is another transition to longer cycles. Thus, the complex price dynamics can be partly understood from the wealth dynamics. But how are the wealth dynamics determined? Why does the $m = 5$ population become wealthier at point **b**, and why does it lose most of this advantage at point **c**? It is obvious that the wealth dynamics are influenced by the price dynamics, thus there is a complicated two-way interaction between the two. Although this interaction is generally very complex, some principle ideas about the mutual influence between the wealth and price patterns can be formulated. For example, a population that becomes dominant and dictates the price dynamics, typically starts under-performing, because it affects the price with its actions. This means pushing the price up when buying, and therefore buying high, and pushing the price down when selling. However, a more detailed analysis must consider the specific investment strategy employed by each population. For a more comprehensive analysis of the interaction between heterogeneous EMB populations see [25].

The two EMB population model generates the same market phenomena as did the homogeneous population

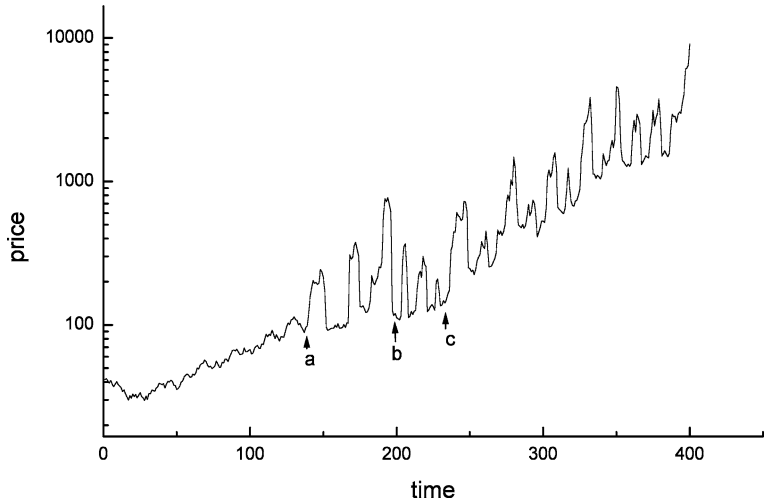

Agent Based Computational Economics, Figure 6
2% EMB $m = 5$, 2% EMB $m = 15$, 96% RII

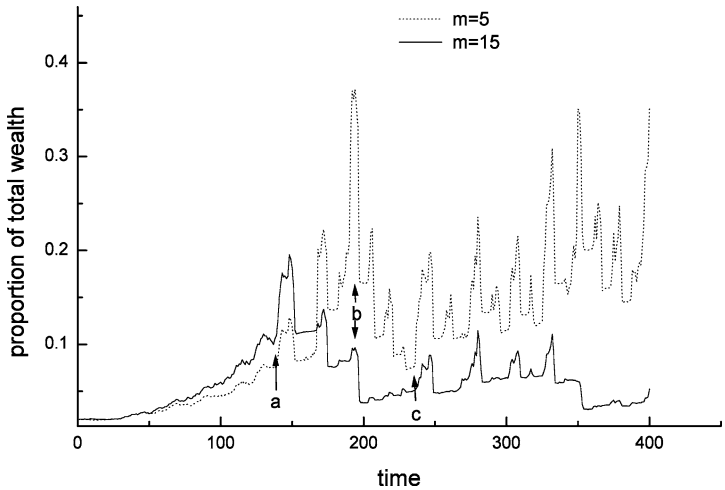

Agent Based Computational Economics, Figure 7
Proportion of the total wealth held by the two EMB populations

case: heavy trading volume, return autocorrelations, and excess volatility. Although the price pattern is much less regular in the two-EMB-population case, there still seems to be a great deal of predictability about the prices. Moreover, the booms and crashes generated by this model are unrealistically dramatic and frequent. In the next section we analyze a model with a continuous spectrum of EMB investors. We show that this fuller heterogeneity of investors leads to very realistic price and volume patterns.

Full Spectrum of EMB Investors Up to this point we have analyzed markets with at most three different subpopulations (one RII population and two EMB populations). The market dynamics we found displayed the empirically observed market anomalies, but they were unrealistic in the magnitude, frequency, and semi-predictability of booms and crashes. In reality, we would expect not only two or three investor types, but rather an entire spectrum of investors. In this section we consider a model with a full spectrum of different EMB investors. It turns out that *more is different*. When there is an entire range of investors, the price dynamics become realistic: booms and crashes are not periodic or predictable, and they are also less frequent and dramatic. At the same time, we still obtain all of the market anomalies described before.

In this model each investor has a different number of *ex-post* observations which he utilizes to estimate the *ex-ante* distribution. Namely, investor i looks at the set of the m^i most recent returns on the stock, and we assume that m^i is distributed in the population according to a truncated normal distribution with average $\bar{m}$ and standard deviation σ_m (as $m \leq 0$ is meaningless, the distribution is truncated at $m = 0$).

Figure 8 shows the price pattern of a typical simulation of this model. In this simulation 90% of the investors are RII, and the remaining 10% are heterogeneous EMB investors with $\bar{m} = 40$, and $\sigma_m = 10$. The price pattern seems very realistic with "smoother" and more irregular cycles. Crashes are dramatic, but infrequent and unpredictable.

The heterogeneous EMB population model generates the following empirically observed market phenomena:

Return Autocorrelation: Momentum and Mean-Reversion In the heterogeneous EMB population model trends are generated by the same positive feedback mechanism that generated cycles in the homogeneous case: high (low) returns tend to make the EMB investors more (less) aggressive, this generates more high (low) returns, etc. The difference between the two cases is that in the heterogeneous case there is a very complicated interaction between all the different investor sub-populations and as a result there are no distinct regular cycles, but rather, smoother and more irregular trends. There is no single cycle length – the dynamics are a combination of many different cycles. This makes the autocorrelation pattern also smoother and more continuous. The return autocorrelations in the heterogeneous model are shown in Fig. 9. This autocorrelation pattern conforms with the empirical findings. In the short-run (lags 1–4) the autocorrelation is positive – this is the empirically documented phenomena known as momentum: in the short-run, high returns tend to be followed by more high returns, and low returns tend to be followed by more low returns. In the longer-run (lags 5–13) the autocorrelation is negative, which is known as mean-reversion. For even longer lags the autocorrelation eventually tends to zero. The short-run momentum, longer-run mean-reversion, and eventual diminishing autocorrelation creates the general "U-shape" which is found in empirical studies [7,13,31] and which is seen in Fig. 9.

Excess Volatility The price level is generally determined by the fundamental value of the stock. However, as in the homogeneous EMB population case, the EMB investors occasionally induce temporary departures of the price away from the fundamental value. These temporary departures from the fundamental value make the price more volatile than the fundamental value. Following Shiller's methodology we define the detrended price, p, and fundamental value, p^f. Averaging over 100 independent simulations we find $\sigma(p) = 27.1$ and $\sigma(p^f) = 19.2$, which is an excess volatility of 41% .

Heavy Volume As investors in our model have different information (the informed investors know the dividend process, while the EMB investors do not), and different ways of interpreting the information (EMB investors with different memory spans have different estimations regarding the *ex-ante* return distribution), there is a high level of trading volume in this model. The average trading volume in this model is about 1700 shares per period (17% of the total outstanding shares). As explained below, the volume is positively correlated with contemporaneous and lagged absolute returns.

Volume is Positively Correlated with Contemporaneous and Lagged Absolute Returns Investors revise their portfolios as a result of changes in their beliefs regarding the future return distribution. The changes in the beliefs can be due to a change in the current price, to a new dividend realization (in the case of the informed investors), or to a new observation of an *ex-post* return (in the case of the EMB investors). If all investors change their beliefs in the same direction (for example, if everybody becomes more optimistic), the stock price can change substantially with almost no volume – everybody would like to increase the proportion of the stock in his portfolio, this will push the price up, but a very small number of shares will change hands. This scenario would lead to zero or perhaps even negative correlation between the magnitude of the price change (or return) and the volume. However, the typical scenario in the LLS model is different. Typically, when a positive feedback trend is induced by the EMB investors, the opinions of the informed investors and the EMB investors change in opposite directions. The EMB investors see a trend of rising prices as a positive indication about the *ex-ante* return distribution, while the informed investors believe that the higher the price level is above the fundamental value, the more overpriced the stock is, and the harder it will eventually fall. The exact opposite holds for a trend of falling prices. Thus, price trends are typically interpreted differently by the two investor types, and therefore induce heavy trading volume. The more pronounced the trend, the more likely it is to lead to heavy volume, and at the same time, to large price changes which are due to the positive feedback trading on behalf of the EMB investors.

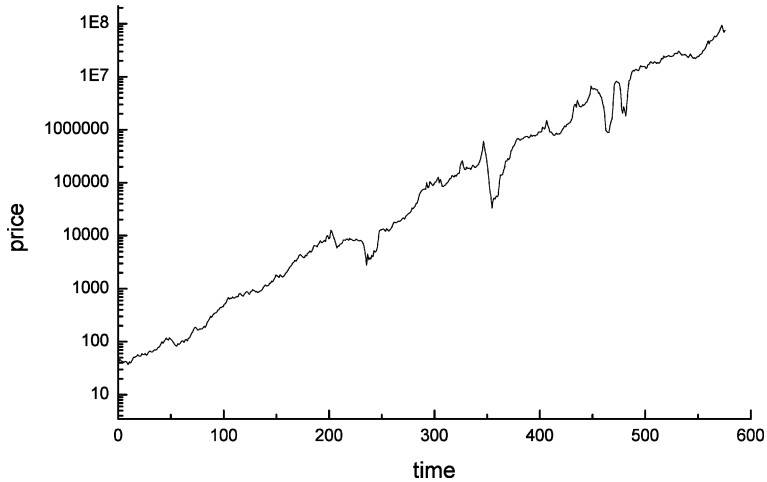

Agent Based Computational Economics, Figure 8
Spectrum of Heterogeneous EMB Investors (10% EMB Investors, 90% RII Investors)

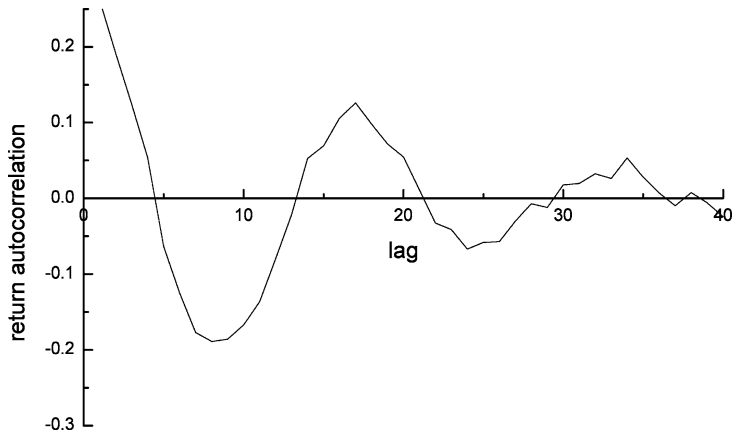

Agent Based Computational Economics, Figure 9
Return Autocorrelation – Heterogeneous EMB Population

This explains not only the positive correlation between volume and contemporaneous absolute rates of return, but also the positive correlation between volume and lagged absolute rates of return. The reason is that the behavior of the EMB investors induces short-term positive return autocorrelation, or momentum (see above). That is, a large absolute return this period is associated not only with high volume this period, but also with a large absolute return next period, and therefore with high volume next period. In other words, when there is a substantial price increase (decrease), EMB investors become more (less) aggressive and the opposite happens to the informed traders. As we have seen before, when a positive feedback loop is started, the EMB investors are more dominant in determining the price, and therefore another large price increase (decrease) is expected next period. This large price change is likely to be associated with heavy trading volume as the opinions of the two populations diverge. Furthermore, this large increase (decrease) is expected to make the EMB investors even more optimistic (pessimistic) leading to another large price increase (decrease) and heavy volume next period.

In order to verify this relationship quantitatively, we regress volume on contemporaneous and lagged absolute rates of return for 100 independent simulations. We run the regressions:

$$\begin{aligned} V_t &= \alpha + \beta_C \, |R_t - 1| + \varepsilon_t \,, \quad \text{and} \\ V_t &= \alpha + \beta_L \, |R_{t-1} - 1| + \varepsilon_t \,, \end{aligned} \tag{17}$$

where V_t is the volume at time t and R_t is the total return on the stock at time t, and the subscripts C and L stand for contemporaneous and lagged. We find an average value of 870 for $\hat{\beta}_C$ with an average t-value of 5.0 and an average value of 886 for $\hat{\beta}_L$ with an average t-value of 5.1.

Discussion of the LLS Results The LLS model is an Agent Based Simulation model of the stock market which incorporates some of the fundamental experimental findings regarding the behavior of investors. The main non-standard assumption of the model is that there is a small minority of investors in the market who are uninformed about the dividend process and who believe in market efficiency. The investment decision of these investors is reduced to the optimal diversification between the stock and the bond.

The LLS model generates many of the empirically documented market phenomena which are hard to explain in the analytical rational-representative-agent framework. These phenomena are:

- Short term momentum;
- Longer term mean reversion;
- Excess volatility;
- Heavy trading volume;
- Positive correlation between volume and contemporaneous absolute returns;
- Positive correlation between volume and lagged absolute returns;
- Endogenous market crashes.

The fact that so many "puzzles" are explained with a simple model built on a small number of empirically documented behavioral elements leads us to suspect that these behavioral elements are very important in understanding the workings of the market. This is especially true in light of the observations that a very small minority of the non-standard bounded-rational investors can have a dramatic influence on the market, and that these investors are not wiped out by the majority of rational investors.

Summary and Future Directions

Standard economic models typically describe a world of homogeneous rational agents. This approach is the foundation of most of our present day knowledge in economic theory. With the Agent Based Simulation approach we can investigate a much more complex and "messy" world with different agent types, who employ different strategies to try to survive and prosper in a market with structural uncertainty. Agents can learn over time, from their own experience and from their observation about the performance of other agents. They co-evolve over time and as they do so, the market dynamics change continuously. This is a world view closer to biology, than it is to the "clean" realm of physical laws which classical economics has aspired to.

The Agent Based approach should not and can not replace the standard analytical economic approach. Rather, these two methodologies support and complement each other: When an analytical model is developed, it should become standard practice to examine the robustness of the model's results with agent based simulations. Similarly, when results emerge from agent based simulation, one should try to understand their origin and their generality, not only by running many simulations, but also by trying to capture the essence of the results in a simplified analytical setting (if possible).

Although the first steps in economic agent based simulations were made decades ago, economics has been slow and cautious to adopt this new methodology. Only in recent years has this field begun to bloom. It is my belief and hope that the agent based approach will prove as fruitful in economics as it has been in so many other branches of science.

Bibliography

Primary Literature

1. Admati A, Pfleiderer P (1988) A theory of intraday patterns: Volume and price variability. Rev Financ Stud 1:3–40
2. Arthur WB (1994) Inductive reasoning and bounded rationality (The El Farol problem). Am Econ Rev 84:406–411
3. Arthur WB, Holland JH, Lebaron B, Palmer RG, Tayler P (1997) Asset pricing under endogenous expectations in an artificial stock market. In: Arthur WB, Durlauf S, Lane D (eds) The economy as an evolving complex system II. Addison-Wesley, Redwood City
4. Brock WA, Hommes CA (1998) Heterogeneous beliefs and routes to chaos in a simple asset pricing model. J Econ Dyn Control 22:1235–1274
5. Egenter E, Lux T, Stauffer D (1999) Finite size effects in Monte Carlo Simulations of two stock market models. Physica A 268:250–256
6. Epstein JM, Axtell RL (1996) Complex adaptive systems. In: Growing artificial societies: Social science from the bottom up. MIT Press, Washington DC
7. Fama E, French K (1988) Permanent and temporary components of stock prices. J Political Econ 96:246–273
8. Friend I, Blume ME (1975) The demand for risky assets. Am Econ Rev 65:900–922
9. Gordon J, Paradis GE, Rorke CH (1972) Experimental evidence on alternative portfolio decision rules. Am Econ Rev 62(1):107–118
10. Grossman S, Stiglitz J (1980) On the impossibility of informationally efficient markets. Am Econ Rev 70:393–408
11. Hellthaler T (1995) The influence of investor number on a microscopic market. Int J Mod Phys C 6:845–852

12. Hommes CH (2002) Modeling the stylized facts in finance through simple nonlinear adaptive systems. PNAS 99:7221–7228
13. Jegadeesh N, Titman S (1993) Returns to buying winners and selling losers: Implications for stock market efficiency. J Finance 48:65–91
14. Karpoff J (1987) The relationship between price changes and trading volume: A survey. J Financ Quantitative Anal 22:109–126
15. Kim GW, Markowitz HM (1989) Investment rules, margin, and market volatility. J Portf Manag 16:45–52
16. Kirman AP (1992) Whom or what does the representative agent represent? J Econ Perspectiv 6:117–136
17. Kohl R (1997) The influence of the number of different stocks on the Levy, Levy Solomon model. Int J Mod Phys C 8:1309–1316
18. Kroll Y, Levy H, Rapoport A (1988) Experimental tests of the separation theorem and the capital asset pricing model. Am Econ Rev 78:500–519
19. LeBaron B (2000) Agent-based computational finance: Suggested readings and early research. J Econ Dyn Control 24:679–702
20. Levy H (1994) Absolute and relative risk aversion: An experimental study. J Risk Uncertain 8:289–307
21. Levy H, Lim KC (1998) The economic significance of the cross-sectional autoregressive model: Further analysis. Rev Quant Finance Acc 11:37–51
22. Levy M, Levy H (1996) The danger of assuming homogeneous expectations. Financ Analyst J 52:65–70
23. Levy M, Levy H, Solomon S (1994) A microscopic model of the stock market: Cycles, booms, and crashes. Econs Lett 45:103–111
24. Levy M, Levy H, Solomon S (2000) Microscopic simulation of financial markets. Academic Press, San Diego
25. Levy M, Persky N, Solomon, S (1996) The complex dyn of a simple stock market model. Int J High Speed Comput 8:93–113
26. Lux T (1995) Herd behaviour, bubbles and crashes. Econ J 105:881
27. Lux T (1998) The socio-economic dynamics of speculative bubbles: Interacting agents, chaos, and the fat tails of returns distributions. J Econ Behav Organ 33:143–165
28. Lux T, Marchesi M (1999) Volatility clustering in financial markets: A micro-simulation of interacting agents. Nature 397:498
29. Orcutt GH, Caldwell SB, Wertheimer R (1976) Policy exploration through microanalytic simulation. The Urban Institute, Washington DC
30. Palmer RG, Arthur WB, Holland JH, LeBaron B, Tayler P (1994) Artificial economic life: A simple model of a stock market. Physica D 75:264–274
31. Poterba JM, Summers LH (1988) Mean reversion in stock returns: Evidence and implications. J Financial Econs 22:27–59
32. Samanidou E, Zschischang E, Stauffer D, Lux T (2007) Agent-based models of financial markets. Rep Prog Phys 70:409–450
33. Samuelson PA (1989) The judgement of economic science on rational portfolio management: Timing and long horizon effects. J Portfolio Manag 16:4–12
34. Samuelson PA (1994) The long term case for equities and how it can be oversold. J Portf Management 21:15–24
35. Sargent T (1993) Bounded rationality and macroeconomics. Oxford University Press, Oxford
36. Schelling TC (1978) Micro motives and macro behavior. Norton & Company, New York
37. Shiller RJ (1981) Do stock prices move too much to be justified by subsequent changes in dividends? Am Econ Rev 71: 421–436
38. Stauffer D, de Oliveira PMC, Bernardes AT (1999) Monte Carlo Simulation of volatility correlation in microscopic market model. Int J Theor Appl Finance 2:83–94
39. Tesfatsion L (2002) Agent-based computational economics: Growing economies from the bottom up. Artif Life 8:55–82
40. Tesfatsion L (2001) Special issue on agent-based computational economics. J Econ Dyn Control 25:281–293
41. Thaler R (ed) (1993) Advances in behavioral finance. Russel Sage Foundation, New York
42. Thaler R (1994) Quasi rational economics. Russel Sage Foundation, New York
43. Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. Science 211:453–480
44. Tversky A, Kahneman D (1986) Rational choice and the framing of decision. J Bus 59(4):251–278
45. Tversky A, Kahneman D (1992) Advances in prospect theory: Cumulative representation of uncertainty. J Risk Uncertain 5:297–323

Books and Reviews

Anderson PW, Arrow J, Pines D (eds) (1988) The economy as an evolving complex system. Addison-Wesley, Redwood City

Axelrod R (1997) The complexity of cooperation: Agent-based models of conflict and cooperation. Princeton University Press, Princeton

Moss de Oliveira S, de Oliveira H, Stauffer D (1999) Evolution, money, war and computers. BG Teubner, Stuttgart-Leipzig

Solomon S (1995) The microscopic representation of complex macroscopic phenomena. In: Stauffer D (ed) Annual Rev Comput Phys II. World Scientific, Singapore

Agent Based Modeling and Artificial Life

Charles M. Macal
Center for Complex Adaptive Agent Systems Simulation (CAS2), Decision and Information Sciences Division, Argonne National Laboratory, Argonne, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Artificial Life
Alife in Agent-Based Modeling
Future Directions
Bibliography

Glossary

Adaptation The process by which organisms (agents) change their behavior or by which populations of agents change their collective behaviors with respect to their environment.

Agent-based modeling (ABM) A modeling and simulation approach applied to a complex system or complex adaptive system, in which the model is comprised of a large number of interacting elements (agents).

Ant colony optimization (ACO) A heuristic optimization technique motivated by collective decision processes followed by ants in foraging for food.

Artificial chemistry Chemistry based on the information content and transformation possibilities of molecules.

Artificial life (ALife) A field that investigates life's essential qualities primarily from an information content perspective.

Artificial neural network (ANN) A heuristic optimization and simulated learning technique motivated by the neuronal structure of the brain.

Autocatalytic set A closed set of chemical reactions that is self-sustaining.

Autonomous The characteristic of being capable of making independent decisions over a range of situations.

Avida An advanced artificial life computer program developed by Adami [1] and others that models populations of artificial organisms and the essential features of life such as interaction and replication.

Biologically-inspired computational algorithm Any kind of algorithm that is based on biological metaphors or analogies.

Cellular automata (CA) A mathematical construct and technique that models a system in discrete time and discrete space in which the state of a cell depends on transition rules and the states of neighboring cells.

Coevolution A process by which many entities adapt and evolve their behaviors as a result of mutually effective interactions.

Complex system A system comprised of a large number of strongly interacting components (agents).

Complex adaptive system (CAS) A system comprised of a large number of strongly interacting components (agents) that adapt at the individual (agent) level or collectively at the population level.

Decentralized control A feature of a system in which the control mechanisms are distributed over multiple parts of the system.

Digital organism An entity that is represented by its essential information-theoretic elements (genomes) and implemented as a computational algorithm or model.

Downward causation The process by which a higher-order emergent structure takes on its own emergent behaviors and these behaviors exert influence on the constituent agents of the emergent structure.

Dynamic network analysis Network modeling and analysis in which the structure of the network, i. e., nodes (agents) and links (agent interactions) is endogenous to the model.

Echo An artificial life computer program developed by Holland [36] that models populations of complex adaptive systems and the essential features of adaptation in nature.

Emergence The process by which order is produced in nature.

Entropy A measure of order, related to the information needed to specify a system and its state.

Evolution (artificial) The process by which a set of instructions is transmitted and changes over successive generations.

Evolutionary game A repeated game in which agents adapt their strategies in recognition of the fact that they will face their opponents in the future.

Evolution strategies A heuristic optimization technique motivated by the genetic operation of selection and mutation.

Evolutionary algorithm Any algorithm motivated by the genetic operations including selection, mutation, crossover, etc.

Evolutionary computing A field of computing based on the use of evolutionary algorithms.

Finite state machine A mathematical model consisting of entities with a finite (usually small) number of possible discrete states.

Game of life, life A cellular automaton developed by Conway [8] that illustrates a maximally complex system based on simple rules.

Generative social science Social science investigation with the goal of understanding how social processes emerge out of social interaction.

Genetic algorithm (GA) A specific kind of evolutionary algorithm motivated by the genetic operations of selection, mutation, and crossover.

Genetic programming (GP) A specific kind of evolutionary algorithm that manipulates symbols according to prescribed rules, motivated by the genetic operations of selection, mutation, and crossover.

Genotype A set of instructions for a developmental process that creates a complex structure, as in a genotype for transmitting genetic information and seeding a developmental process leading to a phenotype.

Hypercycle A closed set of functional relations that is

self-sustaining, as in an autocatalytic chemical reaction network.

Individual-based model An approach originating in ecology to model populations of agents that emphasizes the need to represent diversity among individuals.

Langton's ant An example of a very simple computational program that computes patterns of arbitrary complexity after an initial series of simple structures.

Langton's loop An example of a very simple computational program that computes replicas of its structures according to simple rules applied locally as in cellular automata.

Learning classifier system (LCS) A specific algorithmic framework for implementing an adaptive system by varying the weights applied to behavioral rules specified for individual agents.

Lindenmeyer system (L-system) A formal grammar, which is a set of rules for rewriting strings of symbols.

Machine learning A field of inquiry consisting of algorithms for recognizing patterns in data (e. g., data mining) through various computerized learning techniques.

Mind–body problem A field of inquiry that addresses how human consciousness arises out of material processes, and whether consciousness is the result of a logical-deductive or algorithmic process.

Meme A term coined by Dawkins [17] to refer to the minimal encoding of cultural information, similar to the genome's role in transmitting genetic information.

Particle swarm optimization An optimization technique similar to ant colony optimization, based on independent particles (agents) that search a landscape to optimize a single objective or goal.

Phenotype The result of an instance of a genotype interacting with its environment through a developmental process.

Reaction-diffusion system A system that includes mechanisms for both attraction and transformation (e. g., of agents) as well as repulsion and diffusion.

Recursively generated object An object that is generated by the repeated application of simple rules.

Self-organization A process by which structure and organization arise from within the endogenous instructions and processes inherent in an entity.

Self-replication The process by which an agent (e. g., organism, machine, etc.) creates a copy of itself that contains instructions for both the agent's operation and its replication.

Social agent-based modeling Agent-based modeling applied to social systems, generally applied to people and human populations, but also animals.

Social network analysis (SNA) A collection of techniques and approaches for analyzing networks of social relationships.

Stigmergy The practice of agents using the environment as a tool for communication with other agents to supplement direct agent-to-agent communication.

Swarm An early agent-based modeling toolkit designed to model artificial life applications.

Swarm intelligence Collective intelligence based on the actions of a set of interacting agents behaving according to a set of prescribed simple rules.

Symbolic processing A computational technique that consists of processing symbols rather than strictly numerical data.

Sugarscape An abstract agent-based model of artificial societies developed by Epstein and Axtell [26] to investigate the emergence of social processes.

Tierra An early artificial life computer program developed by Ray [59] that models populations of artificial organisms and the essential features of life such as interaction and replication.

Universal Turing machine (UTM) An abstract representation of the capabilities of any computable system.

Update rule A rule or transformation directive for changing or updating the state of an entity or agent, as for example updating the state of an agent in an agent-based model or updating the state of an L-system.

Definition of the Subject

Agent-based modeling began as the computational arm of artificial life some 20 years ago. Artificial life is concerned with the emergence of order in nature. How do systems self-organize themselves and spontaneously achieve a higher-ordered state? Agent-based modeling then, is concerned with exploring and understanding the processes that lead to the emergence of order through computational means. The essential features of artificial life models are translated into computational algorithms through agent-based modeling. With its historical roots in artificial life, agent-based modeling has become a distinctive form of modeling and simulation. Agent-based modeling is a bottom-up approach to modeling complex systems by explicitly representing the behaviors of large numbers of agents and the processes by which they interact. These essential features are all that is needed to produce at least rudimentary forms of emergent behavior at the systems level. To understand the current state of agent-based mod-

eling and where the field aspires to be in the future, it is necessary to understand the origins of agent-based modeling in artificial life.

Introduction

The field of Artificial Life, or "ALife," is intimately connected to Agent-Based Modeling, or "ABM." Although one can easily enumerate some of life's distinctive properties, such as reproduction, respiration, adaptation, emergence, etc., a precise definition of life remains elusive.

Artificial Life had its inception as a coherent and sustainable field of investigation at a workshop in the late 1980s [43]. This workshop drew together specialists from diverse fields who had been working on related problems in different guises, using different vocabularies suited to their fields.

At about the same time, the introduction of the personal computer suddenly made computing accessible, convenient, inexpensive and compelling as an experimental tool. The future seemed to have almost unlimited possibilities for the development of ALife computer programs to explore life and its possibilities. Thus several ALife software programs emerged that sought to encapsulate the essential elements of life through incorporation of ALife-related algorithms into easily usable software packages that could be widely distributed. Computational programs for modeling populations of digital organisms, such as Tierra, Avida, and Echo, were developed along with more general purpose agent-based simulators such as Swarm.

Yet, the purpose of ALife was never restricted to understanding or re-creating life as it exists today. According to Langton:

> Artificial systems which exhibit lifelike behaviors are worthy of investigation on their own right, whether or not we think that the processes that they mimic have played a role in the development or mechanics of life as we know it to be. Such systems can help us expand our understanding of life as it could be. (p. xvi in [43])

The field of ALife addresses life-like properties of systems at an abstract level by focusing on the information content of such systems independent of the medium in which they exist, whether it be biological, chemical, physical or *in silico*. This means that computation, modeling, and simulation play a central role in ALife investigations.

The relationship between ALife and ABM is complex. A case can be made that the emergence of ALife as a field was essential to the creation of agent-based modeling. Computational tools were both required and became possible in the 1980s for developing sophisticated models of digital organisms and general purpose artificial life simulators. Likewise, a case can be made that the possibility for creating agent-based models was essential to making ALife a promising and productive endeavor. ABM made it possible to understand the logical outcomes and implications of ALife models and life-like processes. Traditional analytical means, although valuable in establishing baseline information, were limited in their capabilities to include essential features of ALife. Many threads of ALife are still intertwined with developments in ABM and vice verse. Agent-based models demonstrate the emergence of life-like features using ALife frameworks; ALife algorithms are widely used in agent-based models to represent agent behaviors. These threads are explored in this article. In ALife terminology, one could say that ALife and ABM have co-evolved to their present states. In all likelihood they will continue to do so.

This chapter covers in a necessarily brief and, perhaps superficial, but broad way, these relationships between ABM and ALife and extrapolates to future possibilities. This chapter is organized as follows. Section "Artificial Life" introduces Artificial Life, its essential elements and its relationship to computing and agent-based modeling. Section "Alife in Agent-Based Modeling" describes several examples of ABM applications spanning many scales. Section "Future Directions" concludes with future directions for ABM and ALife. A bibliography is included for further reading.

Artificial Life

Artificial Life was initially motivated by the need to model biological systems and brought with it the need for computation. The field of ALife has always been multi-disciplinary and continues to encompass a broad research agenda covering a variety of topics from a number of disciplines, including:

- Essential elements of life and artificial life,
- Origins of life and self-organization,
- Evolutionary dynamics,
- Replication and development processes,
- Learning and evolution,
- Emergence,
- Computation of living systems,
- Simulation systems for studying ALife, and
- Many others.

Each of these topics has threads leading into agent-based modeling.

The Essence of ALife

The essence of Artificial Life is summed up by Langton (p. xxii in [43]) with a list of essential characteristics:

- Lifelike behavior on the part of man-made systems
- Semi-autonomous entities whose local interactions with one another are governed by a set of simple rules
- Populations, rather than individuals
- Simple rather than complex specifications
- Local rather than global control
- Bottom-up rather than top-down modeling
- Emergent rather than pre-specified behaviors.

Langton observes that complex high-level dynamics and structures often emerge (in living and artificial systems), developing over time out of the local interactions among low-level primitives. Agent-based modeling has grown up around the need to model the essentials of ALife.

Self-Replication and Cellular Automata Artificial Life traces its beginnings to the work of John von Neumann in the 1940s and investigations into the theoretical possibilities for developing a *self-replicating* machine [64]. Such a self-replicating machine not only carries instructions for its operations, but also for its replication. The issue concerned how to replicate such a machine that contained the instructions for its operation along with the instructions for its replication. Did a machine to replicate such a machine need to contain both the instructions for the machine's operation and replication, as well as instructions for replicating the instructions on how to replicate the original machine? (see Fig. 1). Von Neumann used the abstract mathematical construct of *cellular automata*, originally conceived in discussions with Stanislaw Ulam, to prove that such a machine could be designed, at least in theory. Von Neumann was never able to build such a machine due to the lack of sophisticated computers that existed at the time.

Cellular automata (CA) have been central to the development of computing Artificial Life models. Virtually all of the early agent-based models that required agents to be spatially located were in the form of von Neumann's original cellular automata. A cellular automata is a *finite-state machine* in which time and space are treated as discrete rather than continuous, as would be the case, for example in differential equation models. A typical CA is a two-dimensional grid or lattice consisting of cells. Each cell assumes one of a finite number of states at any time. A cell's neighborhood is the set of cells surrounding a cell, typically, a five-cell neighborhood (von Neumann neighborhood) or a nine-cell neighborhood (Moore neighborhood), as in Fig. 2.

A set of simple state transition rules determines the value of each cell based on the cell's state and the states of neighboring cells. Every cell is updated at each time according to the transition rules. Each cell is identical in terms of its update rules. Cells differ only in their initial states. A CA is deterministic in the sense that the same state for a cell and its set of neighbors always results in the same updated state for the cell. Typically, CAs are set up with periodic boundary conditions, meaning that the set of cells on one edge of the grid boundary are the neighbor cells to the cells on the opposite edge of the grid boundary. The space of the CA grid forms a surface on a toroid, or donut-shape, so there is no boundary per se. It is straightforward to extend the notion of cellular automata to two, three, or more dimensions.

Von Neumann solved the self-replication problem by developing a cellular automaton in which each cell had 29 possible states and five neighbors (including the updated cell itself). In the von Neumann neighborhood, neighbor cells are in the north, south, east and west directions from the updated cell.

The Game Of Life Conway's Game of Life, or Life, developed in the 1970s, is an important example of a CA [8,31,57]. The simplest way to illustrate some of the basic ideas of agent-based modeling is through a CA. The Game of Life is a two-state, nine-neighbor cellular automata with three rules that determine the state (either On, i. e., shaded, or Off, i. e., white) of each cell:

1. A cell will be On in the next generation if exactly three of its eight neighboring cells are currently On.
2. A cell will retain its current state if exactly two of its neighbors are On.
3. A cell will be Off otherwise.

Initially, a small set of On cells is randomly distributed over the grid. The three rules are then applied repeatedly to all cells in the grid.

After several updates of all cells on the grid, distinctive patterns emerge, and in some cases these patterns can sustain themselves indefinitely throughout the simulation (Fig. 3). The state of each cell is based only on the current state of the cell and the cells touching it in its immediate neighborhood. The nine-neighbor per neighborhood assumption built into Life determines the scope of the locally available information for each cell to update its state.

Conway showed that, at least in theory, the structures and patterns that can result during a Life computation are complex enough to be the basis for a fully functional

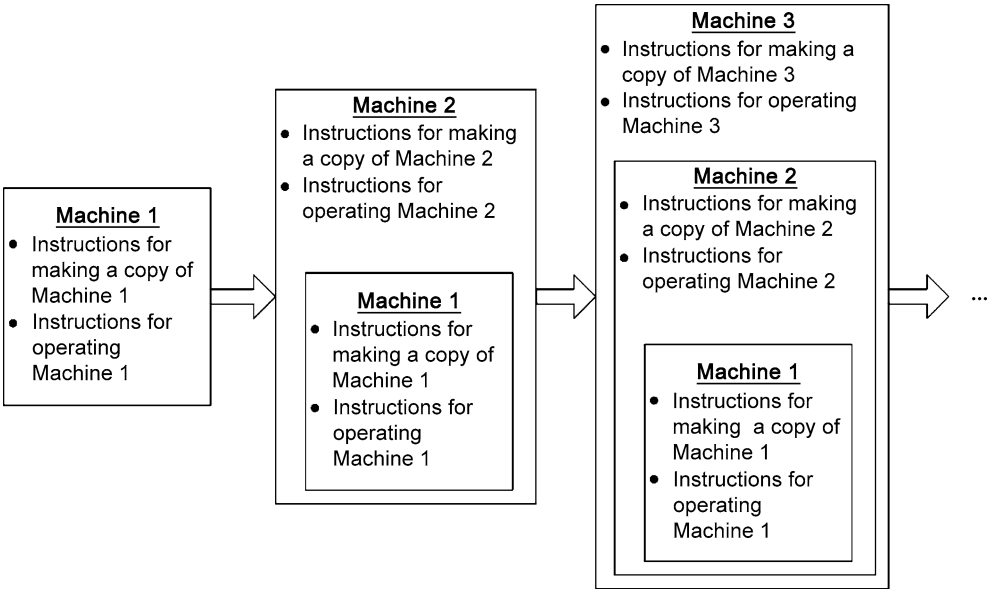

Agent Based Modeling and Artificial Life, Figure 1
Von Neumann's Self-Replication Problem

Agent Based Modeling and Artificial Life, Figure 2
Cellular Automata Neighborhoods

computer that is complex enough to spontaneously generate self-replicating structures (see the section below on universal computation). Two observations are important about the Life rules:

- As simple as the state transition rules are, by using only local information, structures of arbitrarily high complexity can emerge in a CA.
- The specific patterns that emerge are extremely sensitive to the specific rules used. For example, changing Rule 1 above to "A cell will be On in the next generation if exactly four of its eight neighboring cells are currently On" results in the development of completely different patterns.
- The Game of Life provides insights into the role of information in fundamental life processes.

Cellular Automata Classes Wolfram investigated the possibilities for complexity in cellular automata across the

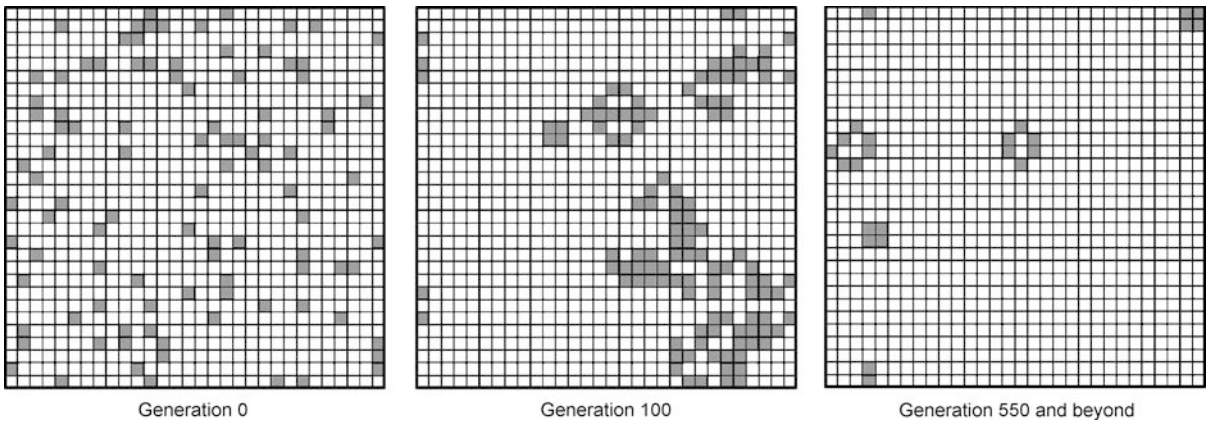

Agent Based Modeling and Artificial Life, Figure 3
Game of Life

full range of transition rules and initial states, using one-dimensional cellular automata [70]. He categorized four distinct classes for the resulting patterns produced by a CA as it is solved repeatedly over time. These are:

- Class I: homogeneous state,
- Class II: simple stable or periodic structure,
- Class III: chaotic (non-repeating) pattern, and
- Class IV: complex patterns of localized structures.

The most interesting of these is Class IV cellular automata, in which very complex patterns of non-repeating localized structures emerge that are often long-lived. Wolfram showed that these Class IV structures were also complex enough to support universal computation [72]. Langton [45] coined the term "life at the edge of chaos" to describe the idea that Class IV systems are situated in a thin region between Class II and Class III systems. Agent-based models often yield Class I, II, and III behaviors.

Other experiments with CAs investigated the simplest representations that could replicate themselves and produce emergent structures. *Langton's Loop* is a self-replicating two-dimensional cellular automaton, much simpler than von Neumann's [41]. Although not complex enough to be a universal computer, Langton's Loop was the simplest known structure that could reproduce itself. *Langton's Ant* is a two-dimensional CA with a simple set of rules, but complicated emergent behavior. Following a simple set of rules for moving from cell to cell, a simulated ant displays unexpectedly complex behavior. After an initial period of chaotic movements in the vicinity of its initial location, the ant begins to build a recurrent pattern of regular structures that repeats itself indefinitely [42]. Langton's Ant has behaviors complex enough to be a universal computer.

Genotype/Phenotype Distinction Biologists distinguish between the *genotype* and the *phenotype* as hallmarks of biological systems. The genotype is the template – the set of instructions, the specification, the blueprint – for an organism. DNA is the genotype for living organisms, for example. A DNA strand contains the complete instructions for the replication and development of the organism. The phenotype is the organism – the machine, the product, the result – that develops from the instructions in the genotype (Fig. 4).

Morphogenesis is the developmental process by which the phenotype develops in accord with the genotype, through interactions with and resources obtained from its environment. In a famous paper, Turing [66] modeled the dynamics of morphogenesis, and more generally, the problem of how patterns self-organize spontaneously in nature. Turing used differential equations to model a simple set of reaction-diffusion chemical reactions. Turing demonstrated that only a few assumptions were necessary to bring about the emergence of wave patterns and gradients of chemical concentration, suggestive of morphological patterns that commonly occur in nature. *Reaction-diffusion systems* are characterized by the simultaneous processes of attraction and repulsion, and are the basis for the agent behavioral rules (attraction and repulsion) in many social agent-based models.

More recently, Bonabeau extended Turing's treatment of morphogenesis to a theory of pattern formation based on agent-based modeling. Bonabeau [12] states the reason for relying on ABM: "because pattern-forming sys-

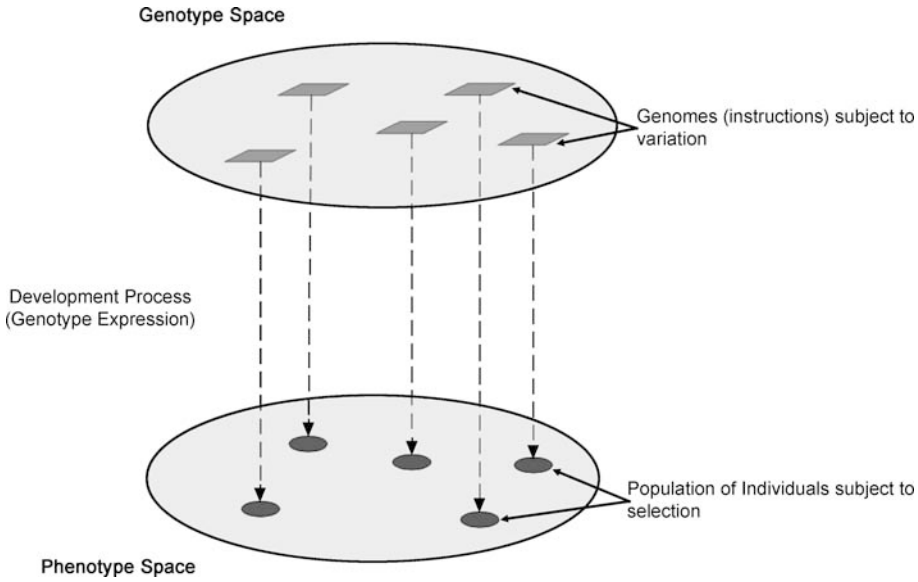

Agent Based Modeling and Artificial Life, Figure 4
Genotype and Phenotype Relations

tems based on agents are (relatively) more easily amenable to experimental observations."

Information Processes One approach to building systems from a genotype specification is based on the methodology of *recursively generated objects*. Such recursive systems are compact in their specification, and their repeated application can result in complex structures, as demonstrated by cellular automata.

Recursive systems are logic systems in which strings of symbols are recursively rewritten based a minimum set of instructions. Recursive systems, or term replacement systems, as they have been called, can result in complex structures. Examples of recursive systems include cellular automata, as described above, and *Lindenmayer systems*, called *L-systems* [46]. An L-system consists of a formal grammar, which is a set of rules for rewriting strings of symbols. L-systems have been used extensively for modeling living systems, for example, plant growth and development, producing highly realistic renderings of plants, with intricate morphologies and branching structures.

Wolfram [71] used symbolic recursion as a basis for developing *Mathematica*, the computational mathematics system based on *symbolic processing* and term replacement. Unlike numeric programming languages, a symbolic programming language allows a variable to be a basic object and does not require a variable to be assigned a value before it is used in a program.

Any agent-based model is essentially a recursive system. Time is simulated by the repeated application of the agent updating rules. The genotype is the set of rules for the agent behaviors. The phenotype is set of the patterns and structures that emerge from the computation. As in cellular automata and recursive systems, extremely complex structures emerge in agent-based models that are often unpredictable from examination of the agent rules.

Emergence One of the primary motivations for the field of ALife is to understand *emergent* processes, that is, the processes by which life emerges from its constituent elements. Langton writes: "The 'key' concept in ALife, is emergent behavior." (p. 2 in [44]). Complex systems exhibit patterns of emergence that are not predictable from inspection of the individual elements. Emergence is described as unexpected, unpredictable or otherwise surprising. That is, the modeled system exhibits behaviors that are not explicitly built into the model. Unpredictability is due to the non-linear effects that result from the interactions of entities having simple behaviors. Emergence by these definitions is something of a subjective process.

In biological systems, emergence is a central issue whether it be the emergence of the phenotype from the genotype, the emergence of protein complexes from genomic information networks [39], or the emergence of consciousness from networks of millions of brain cells.

One of the motivations for agent-based modeling is to explore the emergent behaviors exhibited by the simulated system. In general, agent-based models often exhibit patterns and relationships that emerge from agents interactions. An example is the observed formation of groups of

agents that collectively act in coherent and coordinated patterns. Complex adaptive systems, widely investigated by Holland in his agent-based model Echo [37], are often structured in hierarchies of emergent structures. Emergent structures can collectively form higher-order structures, using the lower-level structures as building blocks. An emergent structure itself can take on new emergent behaviors. These structures in turn affect the agents from which the structure has emerged in a process called *downward causation* [32]. For example, in the real world people organize and identify with groups, institutions, nations, etc. They create norms, laws, and protocols that in turn act on the individuals comprising the group.

From the perspective of agent-based modeling, emergence has some interesting challenges for modeling:

- How does one operationally define emergence with respect to agent-based modeling?
- How does one automatically identify and measure the emergence of entities in a model?
- How do agents that comprise an emergent entity perceived by an observer recognize that they are part of that entity?

Artificial Chemistry *Artificial chemistry* is a subfield of ALife. One of the original goals of Artificial chemistry was to understand how life could originate from pre-biotic chemical processes. Artificial chemistry studies self-organization in chemical reaction networks by simulating chemical reactions between artificial molecules. Artificial chemistry specifies well-understood chemical reactions and other information such as reaction rates, relative molecular concentrations, probabilities of reaction, etc. These form a network of possibilities. The artificial substances and the networks of chemical reactions that emerge from the possibilities are studied through computation. Reactions are specified as recursive algebras and activated as term replacement systems [30].

Hypercycles The emergence of *autocatalytic sets*, or *hypercycles*, has been a prime focus of artificial chemistry [22]. A hypercycle is a self-contained system of molecules and a self-replicating, and thereby self-sustaining, cyclic linkage of chemical reactions. Hypercycles evolve through a process by which self-replicating entities compete for selection.

The hypercycle model illustrates how an ALife process can be adopted to the agent-based modeling domain. Inspired by the hypercycle model, Padgett [54] developed an agent-based model of the co-evolution of economic production and economic firms, focusing on skills. Padgett used the model to establish three principles of social organization that provide foundations for the evolution of technological complexity:

- Structured topology (how interaction networks form)
- Altruistic learning (how cooperation and exchange emerges), and
- Stigmergy (how agent communication is facilitated by using the environment as a means of information exchange among agents).

Digital Organisms The widespread availability of personal computers spurred the development of ALife programs used to study evolutionary processes *in silico*. Tierra was the first system devised in which computer programs were successfully able to evolve and adapt [59]. Avida extended Tierra to account for the spatial distribution of organisms and other features [52,68]. Echo is a simulation framework for implementing models to investigate mechanisms that regulate diversity and information-processing in complex adaptive systems (CAS), systems comprised of many interacting adaptive agents [36,37]. In implementations of Echo, populations evolve interaction networks, resembling species communities in ecological systems, which regulate the flow of resources.

Systems such as Tierra, Avida, and Echo simulate populations of digital organisms, based on the genotype/phenotype schema. They employ computational algorithms to mutate and evolve populations of organisms living in a simulated computer environment. Organisms are represented as strings of symbols, or agent attributes, in computer memory. The environment provides them with resources (computation time) they need to survive, compete, and reproduce. Digital organisms interact in various ways and develop strategies to ensure survival in resource-limited environments.

Digital organisms are extended to agent-based modeling by implementing individual-based models of food webs in a system called DOVE [69]. Agent-based models allow a more complete representation of agent behaviors and their evolutionary adaptability at both the individual and population levels.

ALife and Computing

Creating life-like forms through computation is central to Artificial Life. Is it possible to create life through computation? The capabilities and limitations of computation constrain the types of artificial life that can be created. The history of ALife has close ties with important events in the history of computation.

Alan Turing [65] investigated the limitations of computation by developing an abstract and idealized computer, called a *Universal Turing Machine* (UTM). A UTM has an infinite tape (memory) and is therefore an idealization of any actual computer that may be realized. A UTM is capable of computing anything that is computable, that is, anything that can be derived via a logical, deductive series of statements. Are the algorithms used in today's computers, and in ALife calculations and agent-based models in particular, as powerful as universal computers?

Any system that can effectively simulate a small set of logical operations (such as AND and NOT) can effectively produce any possible computation. Simple rule systems in cellular automata were shown to be equivalent to universal computers [67,72], and in principal able to compute anything that is computable – *perhaps, even life!*

Some have argued that life, in particular human consciousness, is not the result of a logical-deductive or algorithmic process and therefore not computable by a Universal Turing Machine. This problem is more generally referred to as the *mind-body problem* [48]. Dreyfus [20] argues against the assumption often made in the field of artificial intelligence that human minds function like general purpose symbol manipulation machines. Penrose [56] argues that the rational processes of the human mind transcend formal logic systems. In a somewhat different view, biological naturalism contends [63] that human behavior might be able to be simulated, but human consciousness is outside the bounds of computation.

Such philosophical debates are as relevant to agent-based modeling as they are to artificial intelligence, for they are the basis of answering the question of what kind of systems and processes agent-based models will ultimately be able, or unable, to simulate.

Artificial Life Algorithms

ALife use several *biologically-inspired computational algorithms* [53]. Bioinspired algorithms include those based on Darwinian evolution, such as evolutionary algorithms, those based on neural structures, such as neural networks, and those based on decentralized decision making behaviors observed in nature. These algorithms are commonly used to model adaptation and learning in agent-based modeling or to optimize the behaviors of whole systems.

Evolutionary Computing *Evolutionary computing* includes a family of related algorithms and programming solution techniques inspired by evolutionary processes, especially the genetic processes of DNA replication and cell division [21]. These techniques are known as *evolutionary algorithms* and include the following [7]:

- Genetic algorithms [34,35,36,38,51]
- Evolution strategies [60]
- Learning classifier systems [38]
- Genetic programming [40]
- Evolutionary programming [28]

Genetic algorithms (GA) model the dynamic processes by which populations of individuals evolve to improved levels of fitness for their particular environment over repeated generations. GAs illustrate how evolutionary algorithms process a population and apply the genetic operations of mutation and crossover (see Fig. 5). Each behavior is represented as a chromosome consisting of a series of symbols, for example, as a series of 0s and 1s. The encoding process establishing correspondence between behaviors and their chromosomal representations is part of the modeling process.

The general steps in a genetic algorithm are as follows:

1. *Initialization:* Generate an initial population of individuals. The individuals are unique and include specific encoding of attributes in chromosomes that represents the characteristics of the individuals.
2. *Evaluation:* Calculate the fitness of all individuals according to a specified fitness function.
3. *Checking:* If any of the individuals has achieved an acceptable level of fitness, stop, the problem is solved. Otherwise, continue with selection.
4. *Selection:* Select the best pair of individuals in the population for reproduction according to their high fitness levels.
5. *Crossover:* Combine the chromosomes for the two best individuals through a crossover operation and produce a pair of offspring.
6. *Mutation:* Randomly mutate the chromosomes for the offspring.
7. *Replacement:* Replace the least fit individuals in the population with the offspring.
8. Continue at Step 2

Steps 5 and 6 above, the operations of crossover and mutation, comprise the set of genetic operators inspired by nature. This series of steps for a GA comprise a basic framework rather than a specific implementation. Actual GA implementations include numerous variations and alternative implementations in several of the GA steps.

Evolution strategies (ES) are similar to genetic algorithms but rely on mutation as its primary genetic operator.

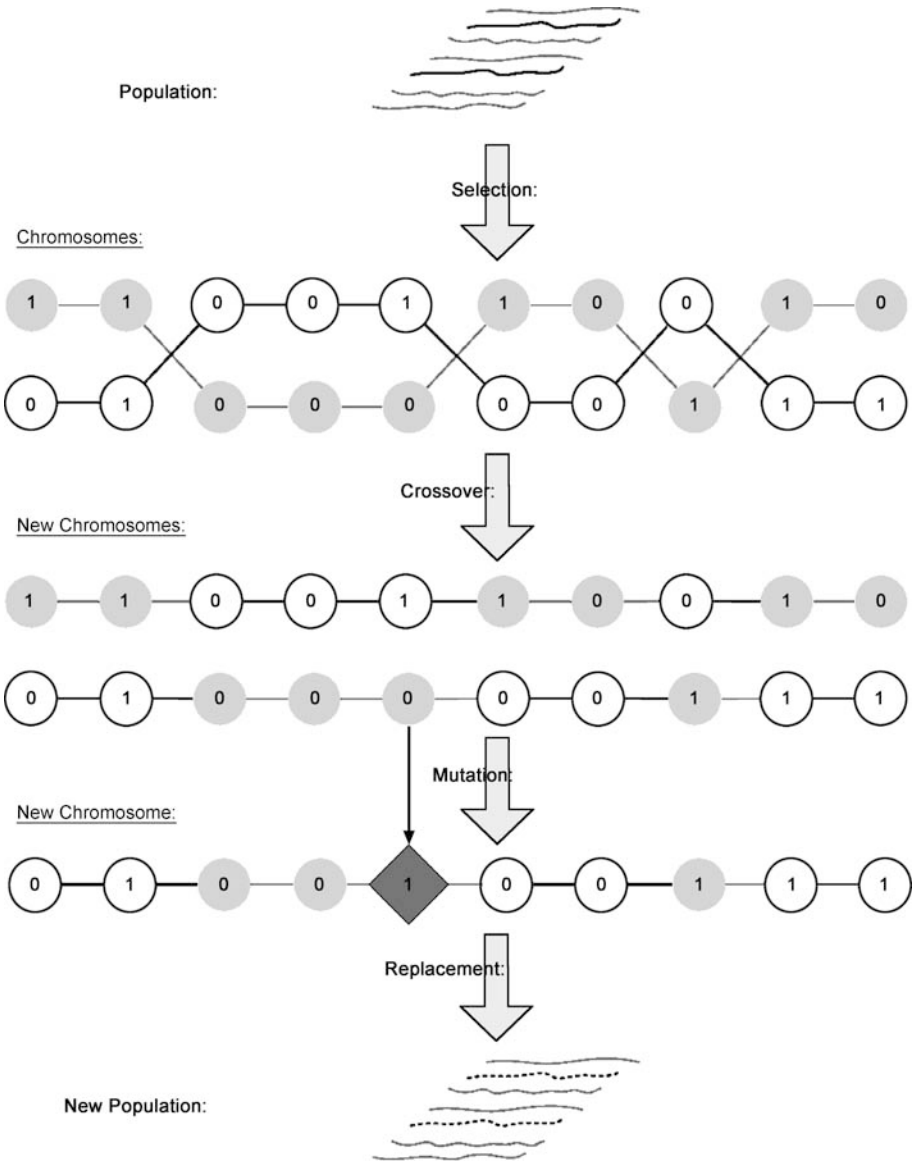

Agent Based Modeling and Artificial Life, Figure 5
Genetic Algorithm

Learning classifier systems (LCS) build on genetic algorithms and adaptively assign relative weights to sensor-action sets that result in the most positive outcomes relative to a goal.

Genetic programming (GP) has similar features to genetic algorithms, but instead of using 0s and 1s or other symbols for comprising chromosomes, GPs combine logical operations and directives in a tree structure. In effect, chromosomes in GPs represent whole computer programs that perform a variety of functions with varying degrees of success and efficiencies. GP chromosomes are evaluated against fitness or performance measures and recombined. Better performing chromosomes are maintained and expand their representation in the population. For example, an application of a GP is to evolve a better-performing rule set that represents an agent's behavior.

Evolutionary programming (EP) is a similar technique to genetic programming, but relies on mutation as its primary genetic operator.

Biologically Inspired Computing *Artificial neural networks* (ANN) are another type of commonly used bio-

logically inspired algorithm [50]. An artificial neural network uses mathematical models based on the structures observed in neural systems. An artificial neuron contains a stimulus-response model of neuron activation based on thresholds of stimulation. In modeling terms, neural networks are equivalent to nonlinear, statistical data modeling techniques. Artificial neural networks can be used to model complex relationships between inputs and outputs and to find patterns in data that are dynamically changing. An ANN is adaptive in that changes in its structure are based on external or internal information that flows through the network. The adaptive capability makes ANN an important technique in agent-based models.

Swarm intelligence refers to problem solving techniques, usually applied to solving optimization problems that are based on decentralized problem solving strategies that have been observed in nature. These include:

- Ant colony optimization [19].
- Particle swarm optimization [16].

Swarm intelligence algorithms simulate the movement and interactions of large numbers of ants or particles over a search space. In terms of agent-based modeling, the ants or particles are the agents, and the search space is the environment. Agents have position and state as attributes. In the case of particle swarm optimization, agents also have velocity.

Ant colony optimization (ACO) mimics techniques that ants use to forage and find food efficiently [13,24]. The general idea of ant colony optimization algorithms is as follows:

1. In a typical ant colony, ants search randomly until one of them finds food.
2. Then they return to their colony and lay down a chemical pheromone trail along the way.
3. When other ants find such a pheromone trail, they are more likely to follow the trail rather than to continue to search randomly.
4. As other ants find the same food source, they return to the nest, reinforcing the original pheromone trail as they return.
5. As more and more ants find the food source, the ants eventually lay down a strong pheromone trail to the point that virtually all the ants are directed to the food source.
6. As the food source is depleted, fewer ants are able to find the food and fewer ants lay down a reinforcing pheromone trail; the pheromone naturally evaporates and eventually no ants proceed to the food source, as the ants shift their attention to searching for new food sources.

In an ant colony optimization computational model, the optimization problem is represented as a graph, with nodes representing places and links representing possible paths. An ant colony algorithm mimics ant behavior with simulated ants moving from node to node in the graph, laying down pheromone trails, etc. The process by which ants communicate indirectly by using the environment as an intermediary is known as *stigmergy* [13], and is commonly used in agent-based modeling.

Particle swarm optimization (PSO) is another decentralized problem solving technique in which a swarm of particles is simulated as it moves over a search space in search of a global optimum. A particle stores its best position found so far in its memory and is aware of the best positions obtained by its neighboring particles. The velocity of each particle adapts over time based on the locations of the best global and local solutions obtained so far, incorporating a degree of stochastic variation in the updating of the particle positions at each iteration.

Artificial Life Algorithms and Agent-Based Modeling Biologically-inspired algorithms are often used with agent-based models. For example, an agent's behavior and its capacity to learn from experience or to adapt to changing conditions can be modeled abstractly through the use of genetic algorithms or neural networks. In the case of a GA, a chromosome effectively represents a single agent action (output) given a specific condition or environmental stimulus (input). Behaviors that are acted on and enable the agent to respond better to environmental challenges are reinforced and acquire a greater share of the chromosome pool. Behaviors that fail to improve the organism's fitness diminish in their representation in the population.

Evolutionary programming can be used to directly evolve programs that represent agent behaviors. For example, Manson [49] develops a bounded rationality model using evolutionary programming to solve an agent multicriteria optimization problem.

Artificial neural networks have also been applied to modeling adaptive agent behaviors, in which an agent derives a statistical relationship between the environmental conditions it faces, its history, and its actions, based on feedback on the success or failures of its actions and the actions of others. For example, an agent may need to develop a strategy for bidding in a market, based on the success of its own and other's previous bids and outcomes.

Finally, swarm intelligence approaches are agent-based in their basic structure, as described above. They can

also be used for system optimization through the selection of appropriate parameters for agent behaviors.

ALife Summary

Based on the previous discussion, the essential features of an ALife program can be summarized as follows:

- *Population:* A population of organisms or individuals is considered. The population may be diversified, and individuals may vary their characteristics, behaviors, and accumulated resources, in both time and space.
- *Interaction:* Interaction requires sensing of the immediate locale, or neighborhood, on the part of an individual. An organism can simply become "aware" of other organisms in its vicinity or it may have a richer set of interactions with them. The individual also interacts with its (non-agent) environment in its immediate locale. This requirement introduces spatial aspects into the problem, as organisms must negotiate the search for resources through time and space.
- *Sustainment and renewal:* Sustainment and renewal requires the acquisition of resources. An organism needs to sense, find, ingest, and metabolize resources, or nourishment as an energy source for processing into other forms of nutrients. Resources may be provided by the environment, i. e., outside of the agents themselves, or by other agents. The need for sustainment leads to competition for resources among organisms. Competition could also be a precursor to cooperation and more complex emergent social structures if this proves to be a more effective strategy for survival.
- *Self-reproduction and replacement:* Organisms reproduce by following instructions at least partially embedded within themselves and interacting with the environment and other agents. Passing on traits to the next generation implies a requirement for trait transmission. Trait transmission requires encoding an organism's traits in a reduced form, that is, a form that contains less than the total information representing the entire organism. It also requires a process for transforming the organism's traits into a viable set of possible new traits for a new organism. Mutation and crossover operators enter into such a process. Organisms also leave the population and are replaced by other organisms, possibly with different traits. The organisms can be transformed through changes in their attributes and behaviors, as in, for example, learning or aging. The populations of organisms can be transformed through the introduction of new organisms and replacement, as in evolutionary adaptation.

As we will see in the section that follows, many of the essential aspects of ALife have been incorporated into the development of agent-based models.

Alife in Agent-Based Modeling

This section briefly touches on the ways in which ALife has motivated agent-based modeling. The form of agent-based models, in terms of their structure and appearance, is directly based on early models from the field of ALife. Several application disciplines in agent-based modeling have been spawned and infused by ALife concepts. Two are covered here. These are how agent-based modeling is applied to social and biological systems.

Agent-Based Modeling Topologies

Agent-based modeling owes much to artificial life in both form and substance. Modeling a population of heterogeneous agents with a diverse set of characteristics is a hallmark of agent-based modeling. The agent perspective is unique among simulation approaches, unlike the process perspective or the state-variable approach taken by other simulation approaches.

As we have seen, agents interact with a small set of neighbor agents in a local area. Agent neighborhoods are defined by how agents are connected, the agent interaction topology. Cellular automata represent agent neighborhoods by using a grid in which the agents exist in the cells, one agent per cell, or as the nodes of the lattice of the grid. The cells immediately surrounding an agent comprise the agent's neighborhood and the agents that reside in the neighborhood cells comprise the neighbors. Many agent-based models have been based on this cellular automata spatial representation. The transition from a cellular automata, such as the game of Life, to an agent-based model is accomplished by allowing agents to be distinct from the cells on which they reside and allowing the agents to move from cell to cell across the grid. Agents move according to the dictates of their behaviors, interacting with other agents that happen to be in their local neighborhoods along the way.

Agent interaction topologies have been extended beyond cellular automata to include networks, either predefined and static as in the case of autocatalytic chemical networks, or endogenous and dynamic, according to the results of agent interactions that occur in the model. Networks allow an agent's neighborhood to be defined more generally and flexibly, and in the case of social agents, more accurately describe social agents' interaction patterns. In addition to cellular automata grids and networks, agent interaction topologies have also been ex-

tended across a variety of domains. In summary, agent interaction topologies include:

- Cellular automata grids (agents are cells or are within in cells) or lattices (agents are grid points),
- Networks, in which agents of vertices and agent relationships are edges,
- Continuous space, in one, two or three dimensions;
- Aspatial random interactions, in which pairs of agents are randomly selected; and
- Geographical Information Systems (GIS), in which agents move over geographically-defined patches, relaxing the one-agent per cell restriction.

Social Agent-Based Modeling

Early social agent-based models were based on ALife's cellular automata approach. In applications of agent-based modeling to social processes, agents represent people or groups of people, and agent relationships represent processes of social interaction [33].

Social Agents Sakoda [61] formulated one of the first social agent-based models, the Checkerboard Model, which had some of the key features of a cellular automaton. Following a similar approach, Schelling developed a model of housing segregation in which agents represent homeowners and neighbors, and agent interactions represent agents' perceptions of their neighbors [62]. Schelling studied housing segregation patterns and posed the question of whether it is it possible to get highly segregated settlement patterns even if most individuals are, in fact, "colorblind?" The Schelling model demonstrated that segregated housing areas can develop spontaneously in the sense that system-level patterns can emerge that are not necessarily implied by or consistent with the objectives of the individual agents (Fig. 6). In the model, agents operated according to a fixed set of rules and were not adaptive.

Identifying the social interaction mechanisms for how cooperative behavior emerges among individuals and groups has been addressed using agent-based modeling and evolutionary game theory. Evolutionary game theory accounts for how the repeated interactions of players in a game-theoretic framework affect the development and evolution of the players' strategies. Axelrod showed, using a cellular automata approach, in which agents on the grid employed a variety of different strategies, that a simple Tit-For-Tat strategy of reciprocal behavior toward individuals is enough to establish sustainable cooperative behavior [4,5]. In addition, Axelrod investigated strategies that were self-sustaining and robust in that they reduced the possibility of invasion by agents having other strategies.

Agent Based Modeling and Artificial Life, Figure 6
Schelling Housing Segregation Model

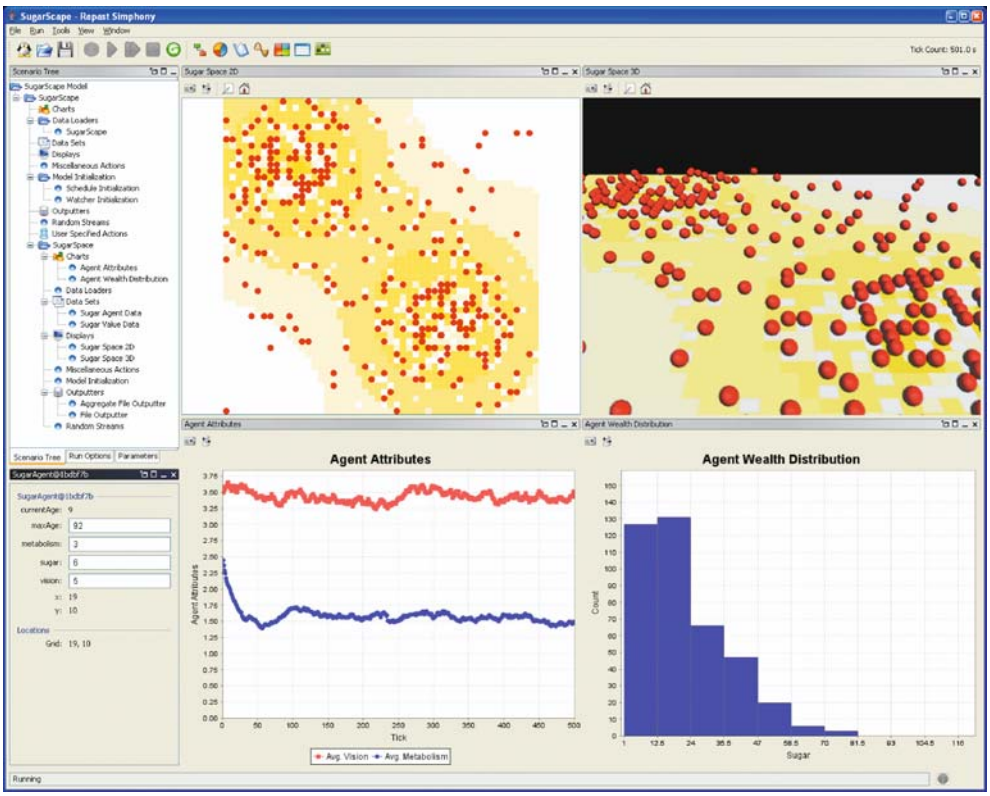

Agent Based Modeling and Artificial Life, Figure 7
Sugarscape Artificial Society Simulation in the Repast Agent-Based Modeling Toolkit

Epstein and Axtell introduced the notion of an external environment that agents interact with in addition to other agents. In their groundbreaking Sugarscape model of artificial societies, agents interacted with their environment depending on their location in the grid [26]. This allowed agents to access environmental variables, extract resources, etc., based on location. In numerous computational experiments, Sugarscape agents emerged with a variety of characteristics and behaviors, highly suggestive of a realistic, although rudimentary and abstract, society (Fig. 7). They observed emergent processes that they interpreted as death, disease, trade, wealth, sex and reproduction, culture, conflict and war, as well as externalities such as pollution. As agents interacted with their neighbors as they moved around the grid, the interactions resulted in a contact network, that is, a network consisting of nodes and links. The nodes are agents and the links indicate the agents that have been neighbors at some point in the course of their movements over the grid. Contact networks were the basis for studying contagion and epidemics in the Sugarscape model. Understanding the agent rules that govern how networks are structured and grow, how quickly information is communicated through networks, and the kinds of relationships that networks embody are important aspects of modeling agents.

Culture and Generative Social Science Dawkins, who has written extensively on aspects of Darwinian evolution, coined the term *meme* as the smallest element of culture that is transmissible between individuals, similar to the notion of the gene as being the primary unit of transmitting genetic information [17]. Several social agent-based models are based on a meme representation of culture as shared or collective agent attributes.

In the broadest terms, social agent-based simulation is concerned with social interaction and social processes. Emergence enters into social simulation through *generative social science* whose goal is to model social processes as emergent processes and their emergence as the result of social interactions. Epstein has argued that social processes are not fully understood unless one is able to theorize how they work at a deep level and have social processes emerge as part of a computational model [25]. More recent work has treated culture as a fluid and dynamic process subject

to interpretation of individual agents, more complex than the genotype/phenotype framework would suggest.

ALife and Biology

ALife research has motivated many agent-based computational models of biological systems, and at all scales, ranging from the cellular level, or even the subcellular molecular level, as the basic unit of agency, to complex organisms embedded in larger structures such as food webs or complex ecosystems.

From Cellular Automata to Cells Cellular automata are a natural application to modeling cellular systems [2,27]. One approach uses the cellular automata grid and cells to model structures of stationary cells comprising a tissue matrix. Each cell is a tissue agent. Mobile cells consisting of pathogens and antibodies are also modeled as agents. Mobile agents diffuse through tissue and interact with tissue and other co-located mobile cells. This approach is the basis for agent-based models of the immune system. Celada and Seiden [15] used bit-strings to model the cell receptors in a cellular automaton model of the immune system called IMMSIM. This approach was extended to a more general agent-based model in and implemented to maximize the number of cells that could be modeled in the CIMMSIM and ParImm systems [9]. The Basic Immune Simulator uses a general agent-based framework (the Repast agent-based modeling toolkit) to model the interactions between the cells of the innate and adaptive immune system [29]. These approaches for modeling the immune system have inspired several agent-based models of intrusion detection for computer networks (see for example, [6]), and have found use in modeling the development and spread of cancer [58].

At the more macroscopic level, agent-based epidemic models have been developed using network topologies. These models include people and some representation of pathogens as individual agents for natural [11] and potentially man-made [14] epidemics.

Modeling bacteria and their associated behaviors in their natural environments is another direction of agent-based modeling. Expanding beyond the basic cellular automata structure into continuous space and network topologies, Emonet, et al. [23] developed AgentCell, a multi-scale agent-based model of *E. Coli* bacteria motility (Fig. 8). In this multi-scale agent-based simulation, molecules within a cell are modeled as individual agents. The molecular reactions comprising the signal transduction network for chemotaxis are modeled using an embedded stochastic simulator, StochSim [46]. This multi-scale approach allows the motile (macroscopic) behavior of colonies of bacteria to be modeled as a direct result of the modeled micro-level processes of protein production within the cells, which are based on individual molecular interactions.

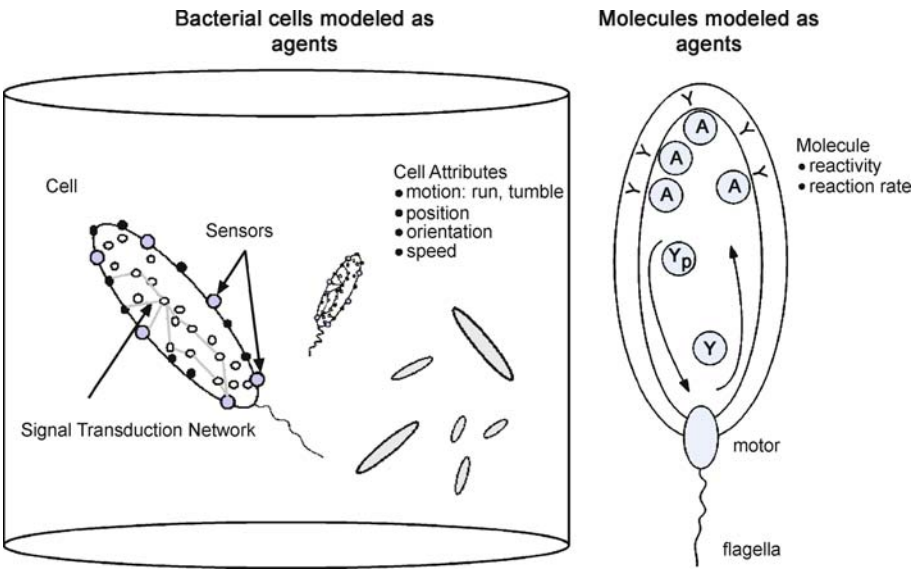

Agent Based Modeling and Artificial Life, Figure 8
AgentCell Multi-Scale Agent-Based Model of Bacterial Chemotaxis

Artificial Ecologies Early models of ecosystems used approaches adapted from physical modeling, especially models of idealized gases based on statistical mechanics. More recently, individual-based models have been developed to represent the full range of individual diversity by explicitly modeling individual attributes or behaviors, and aggregating across individuals for an entire population [18]. Agent-based approaches model a diverse set of agents and their interactions based on their relationships, incorporating adaptive behaviors as appropriate. For example, food webs represent the complex, hierarchical network of agent relationships in local ecosystems [55]. Agents are individuals or species representatives. Adaptation and learning for agents in such food webs can be modeled to explore diversity, relative population sizes, and resiliency to environmental insult.

Adaptation and Learning in Agent-Based Models Biologists consider adaptation to be an essential part of the process of evolutionary change. Adaptation occurs at two levels: the individual level and the population level. In parallel with these notions, agents in an ABM adapt by changing their individual behaviors or by changing their proportional representation in the population. Agents adapt their behaviors at the individual level through learning from experience in their modeled environment.

With respect to agent-based modeling, theories of learning by individual agents or collectives of agents, as well as algorithms for how to model learning, become important. Machine learning is a field consisting of algorithms for recognizing patterns in data (such as data mining) through techniques such as supervised learning, unsupervised learning and reinforcement learning [3,10]. Genetic algorithms [34] and related techniques such as learning classifier systems [38] are commonly used to represent agent learning in agent-based models. In ABM applications, agents learn through interactions with the simulated environment in which they are embedded as the simulation precedes through time, and agents modify their behaviors accordingly.

Agents may also adapt collectively at the population level. Those agents having behavioral rules better suited to their environments survive and thrive, and those agents not so well suited are gradually eliminated from the population.

Future Directions

Agent-based modeling continues to be inspired by ALife – in the fundamental questions it is trying to answer, in the algorithms that it employs to model agent behaviors and solve agent-based models, and in the computational architectures that are employed to implement agent-based models. The future of the fields of both ALife and ABM will continue to be intertwined in essential ways in the coming years.

Computational advances will continue at an ever-increasing pace, opening new vistas for computational possibilities in terms of expanding the scale of models that are possible. Computational advances will take several forms, including advances in computer hardware including new chip designs, multi-core processors, and advanced integrated hardware architectures. Software that take advantage of these designs and in particular computational algorithms and modeling techniques and approaches will continue to provide opportunities for advancing the scale of applications and allow more features to be included in agent-based models as well as ALife applications. These will be opportunities for advancing applications of ABM to ALife in both the realms of scientific research and in policy analysis.

Real-world optimization problems routinely solved by business and industry will continue to be solved by ALife-inspired algorithms. The use of ALife-inspired agent-based algorithms for solving optimization problems will become more widespread because of their natural implementation and ability to handle ill-defined problems.

Emergence is a key theme of ALife. ABM offers the capability to model the emergence of order in a variety of complex and complex adaptive systems. Inspired by ALife, identifying the fundamental mechanisms responsible for higher order emergence and exploring these with agent-based modeling will be an important and promising research area.

Advancing social sciences beyond the genotype/phenotype framework to address the generative nature of social systems in their full complexity is a requirement for advancing computational social models. Recent work has treated culture as a fluid and dynamic process subject to interpretation of individual agents, more complex in many ways than that provided by the genotype/phenotype framework.

Agent-based modeling will continue to be the avenue for exploring new constructs in ALife. If true artificial life is ever developed *in silico*, it will most likely be done using the methods and tools of agent-based modeling.

Bibliography

Primary Literature

1. Adami C (1998) Introduction to Artificial Life. TELOS, Santa Clara

2. Alber MS, Kiskowski MA, Glazier JA, Jiang Y (2003) On Cellular Automaton Approaches to Modeling Biological Cells. In: Rosenthal J, Gilliam DS (eds) Mathematical Systems Theory in Biology, Communication, and Finance, IMA Volume. Springer, New York, pp 1–39
3. Alpaydın E (2004) Introduction to Machine Learning. MIT Press, Cambridge
4. Axelrod R (1984) The Evolution of Cooperation. Basic Books, New York
5. Axelrod R (1997) The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration. Princeton University Press, Princeton
6. Azzedine B, Renato BM, Kathia RLJ, Joao Bosco MS, Mirela SMAN (2007) An Agent Based and Biological Inspired Real-Time Intrusion Detection and Security Model for Computer Network Operations. Comp Commun 30(13):2649–2660
7. Back T (1996) Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms. Oxford University Press, New York
8. Berlekamp ER, Conway JH, Guy RK (2003) Winning Ways for Your Mathematical Plays, 2nd edn. AK Peters, Natick
9. Bernaschi M, Castiglione F (2001) Design and Implementation of an Immune System Simulator, Computers in Biology and Medicine 31(5):303–331
10. Bishop CM (2007) Pattern Recognition and Machine Learning. Springer, New York
11. Bobashev GV, Goedecke DM, Yu F, Epstein JM (2007) A Hybrid Epidemic Model: Combining the Advantages of Agent-Based and Equation-Based Approaches. In: Henderson SG, Biller B, Hsieh M-H, Shortle J, Tew JD, Barton RR (eds) Proc. 2007 Winter Simulation Conference, Washington, pp 1532–1537
12. Bonabeau E (1997) From Classical Models of Morphogenesis to Agent-Based Models of Pattern Formation. Artif Life 3:191–211
13. Bonabeau E, Dorigo M, Theraulaz G (1999) Swarm Intelligence: From Natural to Artificial Systems. Oxford University Press, New York
14. Carley KM, Fridsma DB, Casman E, Yahja A, Altman N, Chen LC, Kaminsky B, Nave D (2006) Biowar: Scalable Agent-Based Model of Bioattacks. IEEE Trans Syst Man Cybern Part A: Syst Hum 36(2):252–265
15. Celada F, Seiden PE (1992) A Computer Model of Cellular Interactions in the Immune System. Immunol Today 13(2):56–62
16. Clerc M (2006) Particle Swarm Optimization. ISTE Publishing Company, London
17. Dawkins R (1989) The Selfish Gene, 2nd edn. Oxford University Press, Oxford
18. DeAngelis DL, Gross LJ (eds) (1992) Individual-Based Models and Approaches in Ecology: Populations, Communities and Ecosystems. Proceedings of a Symposium/Workshop, Knoxville, 16–19 May 1990. Chapman & Hall, New York. ISBN 0-412-03171-X
19. Dorigo M, Stützle T (2004) Ant Colony Optimization. MIT Press, Cambridge
20. Dreyfus HL (1979) What Computers Can't Do: The Limits of Artificial Intelligence. Harper & Row, New York
21. Eiben AE, Smith JE (2007) Introduction to Evolutionary Computing, 2nd edn. Springer, New York
22. Eigen M, Schuster P (1979) The Hypercycle: A Principle of Natural Self-Organization. Springer, Berlin
23. Emonet T, Macal CM, North MJ, Wickersham CE, Cluzel P (2005) AgentCell: A Digital Single-Cell Assay for Bacterial Chemotaxis. Bioinformatics 21(11):2714–2721
24. Engelbrecht AP (2006) Fundamentals of Computational Swarm Intelligence. Wiley, Hoboken
25. Epstein JM (2007) Generative Social Science: Studies in Agent-Based Computational Modeling. Princeton University Press, Princeton
26. Epstein JM, Axtell R (1996) Growing Artificial Societies: Social Science from the Bottom Up. MIT Press, Cambridge
27. Ermentrout GB, Edelstein-Keshet L (1993) Cellular Automata Approaches to Biological Modeling. J Theor Biol 160(1):97–133
28. Fogel LJ, Owens AJ, Walsh MJ (1966) Artificial Intelligence through Simulated Evolution. Wiley, Hoboken
29. Folcik VA, An GC, Orosz CG (2007) The Basic Immune Simulator: An Agent-Based Model to Study the Interactions between Innate and Adaptive Immunity. Theor Biol Med Model 4(39):1–18. http://www.tbiomed.com/content/4/1/39
30. Fontana W (1992) Algorithmic Chemistry. In: Langton CG, Taylor C, Farmer JD, Rasmussen S (eds) Artificial Life II: Proceedings of the Workshop on Artificial Life, Santa Fe, February 1990, Santa Fe Institute Studies in the Sciences of the Complexity, vol X. Addison-Wesley, Reading, pp 159–209
31. Gardner M (1970) The Fantastic Combinations of John Conway's New Solitaire Game Life. Sci Am 223:120–123
32. Gilbert N (2002) Varieties of Emergence. In: Macal C, Sallach D (eds) Proceedings of the Agent 2002 Conference on Social Agents: Ecology, Exchange and Evolution, Chicago, 11–12 Oct 2002, pp 1–11. Available on CD and at www.agent2007.anl.gov
33. Gilbert N, Troitzsch KG (1999) Simulation for the Social Scientist. Open University Press, Buckingham
34. Goldberg DE (1989) Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, Reading
35. Goldberg DE (1994) Genetic and Evolutionary Algorithms Come of Age. Commun ACM 37(3):113–119
36. Holland JH (1975) Adaptation in Natural and Artificial Systems. University of Michigan, Ann Arbor
37. Holland J (1995) Hidden Order: How Adaptation Builds Complexity. Addison-Wesley, Reading
38. Holland JH, Booker LB, Colombetti M, Dorigo M, Goldberg DE, Forrest S, Riolo RL, Smith RE, Lanzi PL, Stolzmann W, Wilson SW (2000) What Is a Learning Classifier System? In: Lanzi PL, Stolzmann W, Wilson SW (eds) Learning Classifier Systems, from Foundations to Applications. Springer, London, pp 3–32
39. Kauffman SA (1993) The Origins of Order: Self-Organization and Selection in Evolution. Oxford University Press, Oxford
40. Koza JR (1992) Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge. 840 pp
41. Langton CG (1984) Self-Reproduction in Cellular Automata. Physica D 10:135–144
42. Langton CG (1986) Studying Artificial Life with Cellular Automata. Physica D 22:120–149
43. Langton CG (1989) Preface. In: Langton CG (ed) Artificial Life: Proceedings of an Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems. Los Alamos, Sept 1987, Addison-Wesley, Reading, pp xv-xxvi
44. Langton CG (1989) Artificial Life. In: Langton CG (ed) Artificial Life: The Proceedings of an Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems. Los Alamos, Sept 1987, Santa Fe Institute Studies in the Sciences of Complexity, vol VI. Addison-Wesley, Reading, pp 1–47

45. Langton CG (1992) Life at the Edge of Chaos. In: Langton CG, Taylor C, Farmer JD, Rasmussen S (eds) Artificial Life II: Proceedings of the Workshop on Artificial Life. Santa Fe, Feb 1990, Santa Fe Institute Studies in the Sciences of the Complexity, vol X. Addison-Wesley, Reading, pp 41–91
46. Le Novere N, Shimizu TS (2001) Stochsim:Modelling of Stochastic Biomolecular Processes. Bioinformatics 17(6):575–576
47. Lindenmeyer A (1968) Mathematical Models for Cellular Interaction in Development. J Theor Biol 18:280–315
48. Lucas JR (1961) Minds, Machines and Godel. Philosophy 36(137):112–127
49. Manson SM (2006) Bounded Rationality in Agent-Based Models: Experiments with Evolutionary Programs. Intern J Geogr Inf Sci 20(9):991–1012
50. Mehrotra K, Mohan CK, Ranka S (1996) Elements of Artificial Neural Networks. MIT Press, Cambridge
51. Mitchell M, Forrest S (1994) Genetic Algorithms and Artificial Life. Artif Life 1(3):267–289
52. Ofria C, Wilke CO (2004) Avida: A Software Platform for Research in Computational Evolutionary Biology. Artif Life 10(2):191–229
53. Olariu S, Zomaya AY (eds) (2006) Handbook of Bioinspired Algorithms and Applications. Chapman, Boca Raton, pp 679
54. Padgett JF, Lee D, Collier N (2003) Economic Production as Chemistry. Ind Corp Chang 12(4):843–877
55. Peacor SD, Riolo RL, Pascual M (2006) Plasticity and Species Coexistence: Modeling Food Webs as Complex Adaptive Systems. In: Pascual M, Dunne JA (eds) Ecological Networks: Linking Structure to Dynamics in Food Webs. Oxford University Press, New York, pp 245–270
56. Penrose R (1989) The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics. Oxford University Press, Oxford
57. Poundstone W (1985) The Recursive Universe. Contemporary Books, Chicago. 252 pp
58. Preziosi L (ed) (2003) Cancer Modelling and Simulation. Chapman, Boca Raton
59. Ray TS (1991) An Approach to the Synthesis of Life (Tierra Simulator). In: Langton CG, Taylor C, Farmer JD, Rasmussen S (eds) Artificial Life Ii: Proceedings of the Workshop on Artificial Life. Wesley, Redwood City, pp 371–408
60. Rechenberg I (1973) Evolutionsstrategie: Optimierung Technischer Systeme Nach Prinzipien Der Biologischen Evolution. Frommann-Holzboog, Stuttgart
61. Sakoda JM (1971) The Checkerboard Model of Social Interaction. J Math Soc 1:119–132
62. Schelling TC (1971) Dynamic Models of Segregation. J Math Soc 1:143–186
63. Searle JR (1990) Is the Brain a Digital Computer? Presidential Address to the American Philosophical Association
64. Taub AH (ed) (1961) John Von Neumann: Collected Works. vol V: Design of Computers, Theory of Automata and Numerical Analysis (Delivered at the Hixon Symposium, Pasadena, Sept 1948). Pergamon Press, Oxford
65. Turing AM (1938) On Computable Numbers with an Application to the Entscheidungsproblem. Process Lond Math Soc 2(42):230–265
66. Turing AM (1952) The Chemical Basis of Morphogenesis. Philos Trans Royal Soc B 237:37–72
67. von Neumann J (1966) Theory of Self-Reproducing Automata, edited by Burks AW. University of Illinois Press, Champaign
68. Wilke CO, Adami C (2002) The Biology of Digital Organisms. Trends Ecol Evol 17(11):528–532
69. Wilke CO, Chow SS (2006) Exploring the Evolution of Ecosystems with Digital Organisms. In: Pascual M, Dunne JA (eds) Ecological Networks: Linking Structure to Dynamics in Food Webs. Oxford University Press, New York, pp 271–286
70. Wolfram S (1984) Universality and Complexity in Cellular Automata. Physica D:1–35
71. Wolfram S (1999) The Mathematica Book, 4th edn. Wolfram Media/Cambridge University Press, Champaign
72. Wolfram S (2002) A New Kind of Science. Wolfram Media, Champaign

Books and Reviews

Artificial Life (journal) web page (2008) http://www.mitpressjournals.org/loi/artl. Accessed 8 March 2008

Banks ER (1971) Information Processing and Transmission in Cellular Automata. Ph. D. dissertation, Massachusetts Institute of Technology

Batty M (2007) Cities and Complexity: Understanding Cities with Cellular Automata, Agent-Based Models, and Fractals. MIT Press, Cambridge

Bedau MA (2002) The Scientific and Philosophical Scope of Artificial Life, Leonardo 35:395–400

Bedau MA (2003) Artificial Life: Organization, Adaptation and Complexity from the Bottom Up. TRENDS Cognit Sci 7(11):505–512

Copeland BJ (2004) The Essential Turing. Oxford University Press, Oxford. 613 pp

Ganguly N, Sikdar BK, Deutsch A, Canright G, Chaudhuri PP (2008) A Survey on Cellular Automata. www.cs.unibo.it/bison/publications/CAsurvey.pdf

Griffeath D, Moore C (eds) (2003) New Constructions in Cellular Automata. Santa Fe Institute Studies in the Sciences of Complexity Proceedings. Oxford University Press, New York, 360 pp

Gutowitz H (ed) (1991) Cellular Automata: Theory and Experiment. Special Issue of Physica D 499 pp

Hraber T, Jones PT, Forrest S (1997) The Ecology of Echo. Artif Life 3:165–190

International Society for Artificial Life web page (2008) www.alife.org. Accessed 8 March 2008

Jacob C (2001) Illustrating Evolutionary Computation with Mathematica. Academic Press, San Diego. 578 pp

Michael CF, Fred WG, Jay A (2005) Simulation Optimization: A Review, New Developments, and Applications. In: Proceedings of the 37th conference on Winter simulation, Orlando

Miller JH, Page SE (2007) Complex Adaptive Systems: An Introduction to Computational Models of Social Life. Princeton University Press, Princeton

North MJ, Macal CM (2007) Managing Business Complexity: Discovering Strategic Solutions with Agent-Based Modeling and Simulation. Oxford University Press, New York

Pascual M, Dunne JA (eds) (2006) Ecological Networks: Linking Structure to Dynamics in Food Webs. Santa Fe Institute Studies in the Sciences of the Complexity. Oxford University Press, New York

Simon H (2001) The Sciences of the Artificial. MIT Press, Cambridge

Sims K (1991) Artificial Evolution for Computer Graphics. ACM SIGGRAPH '91, Las Vegas, July 1991, pp 319–328

Sims K (1994) Evolving 3D Morphology and Behavior by Competition. Artificial Life IV, pp 28–39
Terzopoulos D (1999) Artificial Life for Computer Graphics. Commun ACM 42(8):33–42
Toffoli T, Margolus N (1987) Cellular Automata Machines: A New Environment for Modeling. MIT Press, Cambridge, 200 pp
Tu X, Terzopoulos D, Artificial Fishes: Physics, Locomotion, Perception, Behavior. In: Proceedings of SIGGRAPH'94, 24–29 July 1994, Orlando, pp 43–50
Weisbuch G (1991) Complex Systems Dynamics: An Introduction to Automata Networks, translated from French by Ryckebusch S. Addison-Wesley, Redwood City
Wiener N (1948) Cybernetics, or Control and Communication in the Animal and the Machine. Wiley, New York
Wooldridge M (2000) Reasoning About Rational Agents. MIT Press, Cambridge

Agent Based Modeling and Computer Languages

Michael J. North, Charles M. Macal
Argonne National Laboratory, Decision and Information Sciences Division, Center for Complex Adaptive Agent Systems Simulation (CAS^2), Argonne, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Types of Computer Languages
Requirements of Computer Languages for Agent-Based Modeling
Example Computer Languages Useful for Agent-Based Modeling
Future Directions
Bibliography

Glossary

Agent An agent is a self-directed component in an agent-based model.

Agent-based model An agent-based model is a simulation made up of a set of agents and an agent interaction environment.

Annotations Annotations are a Java feature for including metadata in compiled code.

Attributes Attributes are a C# feature for including metadata in compiled code.

Aspects Aspects are a way to implement dispersed but recurrent tasks in one location.

Bytecode Bytecode is compiled Java binary code.

Common intermediate language Common Intermediate Language (CIL) is compiled binary code for the Microsoft .NET Framework. CIL was originally called Microsoft Intermediate Language (MSIL).

Computational algebra systems Computational algebra systems (CAS) are computational mathematics systems that calculate using symbolic expressions.

Computational mathematics systems Computational Mathematics Systems (CMS) are software programs that allow users to apply powerful mathematical algorithms to solve problems through a convenient and interactive user interface. CMS typically supply a wide range of built-in functions and algorithms.

Class A class is the object-oriented inheritable binding of procedures and data that provides the basis for creating objects.

Computer language A computer language is a method of noting directives for computers. Computer programming languages, or more simply programming languages, are an important category of computer languages.

Computer programming language Please see glossary entry for "Programming language".

C# C# (Archer [1]) is an object-oriented programming language that was developed and is maintained by Microsoft. C# is one of many languages that can be used to generate Microsoft .NET Framework code. This code is run using a 'virtual machine' that potentially gives it a consistent execution environment on different computer platforms.

C++ C++ is a widely used object-oriented programming language that was created by Bjarne Stroustrup (Stroustrup [39]) at AT&T. C++ is widely noted for both its object-oriented structure and its ability to be easily compiled into native machine code.

Design pattern Design patterns form a "common vocabulary" describing tried and true solutions for commonly faced software design problems (Coplien [6]).

Domain-specific language Domain-specific languages (DSL's) are computer languages that are highly customized to support a well defined application area or 'domain'. DSL's commonly include a substantial number of keywords that are nouns and verbs in the area of application as well as overall structures and execution patterns that correspond closely with the application area.

Declarative language According to Watson [47] a "declarative language (or non-procedural language) involves the specification of a set of rules defining the solution to the problem; it is then up to the computer to determine how to reach a solution consistent with the given rules".

Dynamic method invocation Dynamic method invocation, combined with reflection, is a Java and C# approach to higher-order programming.

Encapsulation Encapsulation is the containment of details inside a module.

Field A field is a piece of object-oriented data.

Functional language According to Watson [47] "in functional languages (sometimes called applicative languages) the fundamental operation is function application".

Function pointers Function pointers are part of C++'s approach to higher-order programming. Runtime Type Identification is another component of this approach.

Generics Generics are a Java and C# feature for generalizing classes.

Goto statement A goto statement is an unconditional jump in a code execution flow.

Headless A headless program executes without the use of a video monitor. This is generally done to rapidly execute models while logging results to files or databases.

Higher-order programming According to Reynolds [35], higher-order programming involves the use of "procedures or labels … as data" such that they "can be used as arguments to procedures, as results of functions, or as values of assignable variables".

Imperative language According to Watson [47] in imperative languages "there is a fundamental underlying dependence on the assignment operation and on variables implemented as computer memory locations, whose contents can be read and altered".

Inheritance Inheritance is the ability of an object-oriented class to assume the methods and data of another class called the parent class.

Java Java (Foxwell [12]) is a widely used object-oriented programming language that was developed and is maintained by Sun Microsystems. Java is known for its widespread cross-platform availability on many different types of hardware and operating systems. This capability comes from Java's use of a 'virtual machine' that allows code to have a consistent execution environment on many different computer platforms.

Logic programming language According to Watson [47] "in a logic programming language, the programmer needs only to supply the problem specification in some formal form, as it is the responsibility of the language system to infer a method of solution".

Macro language Macro languages are simple domain-specific languages that are used to write script for tools such as spreadsheets.

Mathematica *Mathematica* is a commercial software program for computational mathematics. Information on *Mathematica* can be found at http://www.wolfram.com/.

Method A method is an object-oriented procedure.

Methodological individualism A reductionist approach to social science originally developed by Max Weber that focuses on the interaction of well defined and separate individuals (Heath [20]). Alternatives theories usually focus on more holistic views of interaction (Heath [20]).

MATLAB The MATrix Laboratory (MATLAB) is a commercial software program for computational mathematics. Information on MATLAB can be found at http://www.mathworks.com/.

Mobile agents Mobile agents are light-weight software proxies that roam the world-wide web and perform various functions programmed by their owners such as gathering information from web sites.

Module According to Stevens et al. [38], "the term module is used to refer to a set of one or more contiguous program statements having a name by which other parts of the system can invoke it and preferably having its own distinct set of variable names".

NetLogo NetLogo (Wilensky [48]) is an agent-based modeling and simulation platform that uses a domain-specific language to define models. NetLogo models are built using a metaphor of turtles as agents and patches as environmental components (Wilensky [48]). NetLogo itself is Java-based. NetLogo is free for use in education and research. More information on NetLogo and downloads can be found at http://ccl.northwestern.edu/netlogo/.

Non-procedural language Please see glossary entry for "Declarative language".

Object An object is the instantiation of a class to produce executable instances.

Object-oriented language Object-oriented languages are structured languages that have special features for binding data with procedures; inheritance; encapsulation; and polymorphism. Careful abstraction that omits unimportant details is an important design principle associated with the use of object-oriented languages.

Objective-C Objective-C is an object-oriented language that extends the C language.

Observer The observer is a NetLogo agent that has a view of an entire model. There is exactly one observer in every NetLogo model.

Patch A patch is a NetLogo agent with a fixed location on a master grid.

Polymorphism Polymorphism is the ability of an object-oriented class to respond to multiple related messages, often method calls with the same name but different parameters.

Procedural language According to Watson [47] "procedural languages … are those in which the action of the program is defined by a series of operations defined by the programmer".

Programming language A programming language is a computer language that allows any computable activity to be expressed.

Record A record is an independently addressable collection of data items.

Reflection Reflection, combined with dynamic method invocation, is a Java and C# approach to higher-order programming.

Repast The Recursive Porous Agent Simulation Toolkit (Repast) is a free and open source family of agent-based modeling and simulation platforms (ROAD [44]). Information on Repast and free downloads can be found at http://repast.sourceforge.net/.

Repast simphony Repast Simphony (Repast S) is the newest member of the Repast family of free and open source agent-based modeling and simulation platforms (North et al. [32], North et al. [30]). The Java-based Repast S system includes advanced features for specifying, executing, and analyzing agent-based simulations.

Repast simphony score Repast Simphony Score is an XML metadata file format that describes the components (e. g., agents and spaces) allowed in a Repast Simphony simulation.

Runtime type identification Runtime Type Identification (RTTI) is part of C++'s approach to higher-order programming. Function pointers are another component of this approach.

Structured language Structured languages are languages that divide programs into separate modules each of which has one controlled entry point, a limited number of exit points, and no internal jumps (Dijkstra [10]).

Swarm Swarm (Swarm Development Group [40]) is a free and open source agent-based modeling and simulation platform maintained by the Swarm Development Group. The core Swarm system uses Objective-C. A Java-based "Java Swarm" wrapper for the Objective-C core is also available. Information on Swarm and free downloads can be found at http://www.swarm.org/.

Templates Templates are a C++ feature for generalizing classes.

Turtle A turtle is a mobile NetLogo agent.

Virtual machine A virtual machine is a software environment that allows user code to have a consistent execution environment on many different computer platforms.

Unstructured language Unstructured languages are languages that rely on step-by-step solutions such that the solutions can contain arbitrary jumps between steps.

Definition of the Subject

Agent-based modeling is a bottom-up approach to representing and investigating complex systems. Agent-based models can be implemented either computationally (e. g., through computer simulation) or non-computationally (e. g., with participatory simulation). The close match between the capabilities of computational platforms and the requirements of agent-based modeling make these platforms a natural choice for many agent-based models. Of course, realizing the potential benefits of this natural match necessitates the use of computer languages to express the designs of agent-based models. A wide range of computer programming languages can play this role including both domain-specific and general purpose languages. The domain-specific languages include business-oriented languages (e. g., spreadsheet programming tools); science and engineering languages (e. g., Mathematica); and dedicated agent-based modeling languages (e. g., NetLogo). The general purpose languages can be used directly (e. g., Java programming) or within agent-based modeling toolkits (e. g., Repast). The choice that is most appropriate for each modeling project depends on both the requirements of that project and the resources available to implement it.

Introduction

The term agent-based modeling (ABM) refers to the computational modeling of a system as comprised of a number of independent, interacting entities, which are referred to as 'agents'. Generally, an agent-based system is made up of agents that interact, adapt, and sustain themselves while interacting with other agents and adapting to a changing environment. The fundamental feature of an agent is its autonomy, the capability of the agent to act independently without the need for direction from external sources. Agents have behaviors that make them active rather than passive entities. Agent behaviors allow agents to take in information from their environment, which in-

cludes their interactions with other agents, process the information and make some decision about their next action, and take the action. Jennings [21] provides a rigorous computer science view of agency emphasizing the essential characteristic of autonomous behavior.

Beyond the essential characteristic of autonomy, there is no universal agreement on the precise definition of the term "agent", as used in agent-based modeling. Some consider any type of independent component, whether it be a software model or a software model of an extant individual, to be an agent [4]. An independent component's behaviors can be modeled as consisting of anything from simple reactive decision rules to multi-dimensional behavior complexes based on adaptive artificial intelligence (AI) techniques.

Other authors insist that a component's behavior must be adaptive in order for the entity to be considered an agent. The agent label is reserved for components that can adapt to their environment, by learning from the successes and failures with their interactions with other agents, and change their behaviors in response. Casti [5] argues that agents should contain both base-level rules for behavior as well as a higher-level set of "rules to change the rules". The base-level rules provide responses to the environment while the "rules to change the rules" provide adaptation [5].

From a practical modeling standpoint, agent characteristics can be summarized as follows:

- Agents are identifiable as self-contained individuals. An agent has a set of characteristics and rules governing its behaviors.
- Agents are autonomous and self-directed. An agent can function independently in its environment and in its interactions with other agents, at least over a limited range of situations that are of interest.
- An agent is situated, living in an environment with which it interacts along with other agents. Agents have the ability to recognize and distinguish the traits of other agents. Agents also have protocols for interaction with other agents, such as for communication, and the capability to respond to the environment.
- An agent may be goal-directed, having targets to achieve with respect to its behaviors. This allows an agent to compare the outcome of its behavior to its goals. An agent's goals need not be comprehensive or well-defined. For example, an agent does not necessarily have formally stated objectives it is trying to maximize.
- An agent might have the ability to learn and adapt its behaviors based on its experiences. An agent might have rules that modify its behavior over time. Generally, learning and adaptation at the agent level requires some form of memory to be built into the agents behaviors.

Often, in an agent-based model, the population of agents varies over time, as agents are born and die. Another form of adaptation can occur at the agent population level. Agents that are fit are better able to sustain themselves and possibly reproduce as time in the simulation progresses, while agents that have characteristics less suited to their continued survival are excluded from the population.

Another basic assumption of agent-based modeling is that agents have access only to local information. Agents obtain information about the rest of the world only through their interactions with the limited number of agents around them at any one time, and from their interactions with a local patch of the environment in which they are situated.

These aspects of how agent-based modeling treats agents highlight the fact that the full range of agent diversity can be incorporated into an agent-based model. Agents are diverse and heterogeneous as well as dynamic in their attributes and behavioral rules. There is no need to make agents homogeneous through aggregating agents into groups or by identifying the 'average' agent as representative of the entire population. Behavioral rules vary in their sophistication, how much information is considered in the agent decisions (i. e., cognitive 'load'), the agent's internal models of the external world including the possible reactions or behaviors of other agents, and the extent of memory of past events the agent retains and uses in its decisions. Agents can also vary by the resources they have manage to accumulate during the simulation, which may be due to some advantage that results from specific attributes. The only limit on the number of agents in an agent-based model is imposed by the computational resources required to run the model.

As a point of clarification, agent-based modeling is also known by other names. ABS (agent-based systems), IBM (individual-based modeling), and MAS (multi-agent systems) are widely-used acronyms, but 'ABM' will be used throughout this discussion. The term 'agent' has connotations other than how it is used in ABM. For example, ABM agents are different from the typical agents found in mobile agent systems. 'Mobile agents' are light-weight software proxies that roam the world-wide web and perform various functioned programmed by their owners such as gathering information from web sites. To this extent, mobile agents are autonomous and share this characteristic with agents in ABM.

Types of Computer Languages

A 'computer language' is a method of noting directives for computers. 'Computer programming languages,' or 'programming languages,' are an important category of computer languages. A programming language is a computer language that allows any computable activity to be expressed. This article focuses on computer programming languages rather than the more general computer languages since virtually all agent-based modeling systems require the power of programming languages. This article sometimes uses the simpler term 'computer languages' when referring to computer programming languages. According to Watson [47]:

> Programming languages are used to describe algorithms, that is, sequences of steps that lead to the solution of problems … A programming language can be considered to be a 'notation' that can be used to specify algorithms with precision.

Watson [47] goes on to say that "programming languages can be roughly divided into four groups: imperative languages, functional languages, logic programming languages, and others". Watson [47] states that in imperative languages "there is a fundamental underlying dependence on the assignment operation and on variables implemented as computer memory locations, whose contents can be read and altered". However, "in functional languages (sometimes called applicative languages) the fundamental operation is function application" [47]. Watson cites LISP as an example. Watson [47] continues by noting that "in a logic programming language, the programmer needs only to supply the problem specification in some formal form, as it is the responsibility of the language system to infer a method of solution".

A useful feature of most functional languages, many logic programming languages, and some imperative languages is higher-order programming. According to Reynolds [35]:

> In analogy with mathematical logic, we will say that a programming language is higher-order if procedures or labels can occur as data, i. e., if these entities can be used as arguments to procedures, as results of functions, or as values of assignable variables. A language that is not higher-order will be called first-order.

Watson [47] offers that "another way of grouping programming languages is to classify them as procedural or declarative languages". Elaborating, Watson [47] states that:

> Procedural languages … are those in which the action of the program is defined by a series of operations defined by the programmer. To solve a problem, the programmer has to specify a series of steps (or statements) which are executed in sequence.

On the other hand, Watson [47] notes that:

> Programming in a declarative language (or non-procedural language) involves the specification of a set of rules defining the solution to the problem; it is then up to the computer to determine how to reach a solution consistent with the given rules … The language Prolog falls into this category, although it retains some procedural aspects. Another widespread non-procedural system is the spreadsheet program.

Imperative and functional languages are usually procedural while logic programming languages are generally declarative. This distinction is important since it implies that most imperative and functional languages require users to define how each operation is to be completed while logic programming languages only require users to define what is to be achieved. However, when faced with multiple possible solutions with different execution speeds and memory requirements, imperative and functional languages offer the potential for users to explicitly choose more efficient implementations over less efficient ones. Logic programming languages generally need to infer which solution is best from the problem description and may or may not choose the most efficient implementation. Naturally, this potential strength of imperative and functional languages may also be cast as a weakness. With imperative and functional language users need to correctly choose a good implementation among any competing candidates that may be available.

Similarly to Watson [47], Van Roy and Haridi [45] define several common computational models namely those that are object-oriented, those that are logic-based, and those that are functional. Object-oriented languages are procedural languages that bind procedures (i. e., 'encapsulated methods') to their corresponding data (i. e., 'fields') in nested hierarchies (i. e., 'inheritance' graphs) such that the resulting 'classes' can be instantiated to produce executable instances (i. e., 'objects') that respond to multiple related messages (i. e., 'polymorphism'). Logic-based languages correspond to Watson's [47] logic programming languages. Similarly, Van Roy and Haridi [45] functional languages correspond to those of Watson [47].

Two additional types of languages can be added to Van Roy and Haridi's [45] list of three. These are unstruc-

tured and structured languages [10]. Both unstructured and structured languages are procedural languages.

Unstructured languages are languages that rely on step-by-step solutions such that the solutions can contain arbitrary jumps between steps [10]. BASIC, COBOL, FORTRAN, and C are examples of unstructured languages. The arbitrary jumps are often implemented using 'goto' statements. Unstructured languages where famously criticized by Edsger Dijkstra in his classic paper "Go To Statement Considered Harmful" [10]. This and related criticism lead to the introduction of structured languages.

Structured languages are languages that divide programs into separate modules each of which has one controlled entry point, a limited number of exit points, and no internal jumps [10]. Following Stevens et al. [38] "the term module is used to refer to a set of one or more contiguous program statements having a name by which other parts of the system can invoke it and preferably having its own distinct set of variable names". Structured language modules, often called procedures, are generally intended to be small. As such, large numbers of them are usually required to solve complex problems. Standard Pascal is an example of structured, but not object-oriented, language. As stated earlier, C is technically an unstructured language (i. e., it allows jumps within procedures and 'long jumps' between procedures), but it is used so often in a structured way that many people think of it as a structured language.

The quality of modularization in structured language code is often considered to be a function of coupling and cohesion [38]. Coupling is the tie between modules such that the proper functioning of one module depends on the functioning of another module. Cohesion is the ties within a module such that proper functioning of one line of code in a module depends on the functioning of another one line of code in the same module. The goal for modules is maximizing cohesion while minimizing coupling.

Object-oriented languages are a subset of structured languages. Object-oriented methods and classes are structured programming modules that have special features for binding data, inheritance, and polymorphism. The previously introduced concepts of coupling and cohesion apply to classes, objects, methods, and fields the same way that they apply to generic structured language modules. Objective-C, C++, C#, and Java are all examples of object-oriented languages. As with C, the languages Objective-C, C++, and C# offer goto statements but they have object-oriented features and are generally used in a structured way. Java is an interesting case in that the word 'goto' is reserved as a keyword in the language specification, but it is not intended to be implemented.

It is possible to develop agent-based models using any of the programming languages discussed above namely, unstructured languages, structured languages, object-oriented languages, logic-based languages, and functional languages. Specific examples are provided later in this article. However, certain features of programming languages are particularly well suited for supporting the requirements of agent-based modeling and simulation.

Requirements of Computer Languages for Agent-Based Modeling

The requirements of computer languages for agent-based modeling and simulation include the following:

- There is a need to create well defined modules that correspond to agents. These modules should bind together agent state data and agent behaviors into integrated independently addressable constructs. Ideally these modules will be flexible enough to change structure over time and to optionally allow fuzzy boundaries to implement models that go beyond methodological individualism [20].
- There is a need to create well defined containers that correspond to agent environments. Ideally these containers will be recursively nestable or will otherwise support sophisticated definitions of containment.
- There is a need to create well defined spatial relationships within agent environments. These relationships should include notions of abstract space (e. g., lattices), physical space (e. g., maps), and connectedness (e. g., networks).
- There is a need to easily setup model configurations such as the number of agents; the relationships between agents; the environmental details; and the results to be collected.
- There is a need to conveniently collect and analyze model results.

Each of the kinds of programming languages namely, unstructured languages, structured languages, object-oriented languages, logic-based languages, and functional languages can address these requirements.

Unstructured languages generally support procedure definitions which can be used to implement agent behaviors. They also sometimes support the collection of diverse data into independently addressable constructs in the form of data structures often called 'records'. However, they generally lack support for binding procedures to individual data items or records of data items. This lack of support for creating integrated constructs also typically limits the

language-level support for agent containers. Native support for implementing spatial environments is similarly limited by the inability to directly bind procedures to data.

As discussed in the previous section, unstructured languages offer statements to implement execution jumps. The use of jumps within and between procedures tends to reduce module cohesion and increase module coupling compared to structured code. The result is reduced code maintainability and extensibility compared to structured solutions. This is a substantial disadvantage of unstructured languages.

In contrast, many have argued that, at least theoretically, unstructured languages can achieve the highest execution speed and lowest memory usage of the language options since nearly everything is left to the application programmers. In practice, programmers implementing agent-based models in unstructured languages usually need to write their own tools to form agents by correlating data with the corresponding procedures. Ironically, these tools are often similar in design, implementation, and performance to some of the structured and object-oriented features discussed later.

Unstructured languages generally do not provide special support for application data configuration, program output collection, or program results analysis. As such, these tasks usually need to be manually implemented by model developers.

In terms of agent-based modeling, structured languages are similar to unstructured languages in that they do not provide tools to directly integrate data and procedures into independently addressable constructs. Therefore, structured language support for agents, agent environments, and agent spatial relationships is similar to that provided by unstructured languages. However, the lack of jump statements in structured languages tends to increase program maintainability and extensibility compared to unstructured languages. This generally gives structured languages a substantial advantage over unstructured languages for implementing agent-based models.

Object-oriented languages build on the maintainability and extensibility advantages of structured languages by adding the ability to bind data to procedures. This binding in the form of classes provides a natural way to implement agents. In fact, object-oriented languages have their roots in Ole-Johan Dahl and Kristen Nygaard's Simula simulation language [7,8,45]! According to Dahl and Nygaard [7]:

> SIMULA (SIMULation LAnguage) is a language designed to facilitate formal description of the layout and rules of operation of systems with discrete events (changes of state). The language is a true extension of ALGOL 60 [2], i. e., it contains ALGOL 60 as a subset. As a programming language, apart from simulation, SIMULA has extensive list processing facilities and introduces an extended coroutine concept in a high-level language.

Dahl and Nygaard go on to state the importance of specific languages for simulation [7] as follows:

> Simulation is now a widely used tool for analysis of a variety of phenomena: nerve networks, communication systems, traffic flow, production systems, administrative systems, social systems, etc. Because of the necessary list processing, complex data structures and program sequencing demands, simulation programs are comparatively difficult to write in machine language or in ALGOL or FORTRAN. This alone calls for the introduction of simulation languages.
>
> However, still more important is the need for a set of basic concepts in terms of which it is possible to approach, understand and describe all the apparently very different phenomena listed above. A simulation language should be built around such a set of basic concepts and allow a formal description which may generate a computer program. The language should point out similarities and differences between systems and force the research worker to consider all relevant aspects of the systems. System descriptions should be easy to read and print and hence useful for communication.

Again, according to Dahl and Nygaard [8]:

> SIMULA I (1962–65) and Simula 67 (1967) are the two first object-oriented languages. Simula 67 introduced most of the key concepts of object-oriented programming: both objects and classes, subclasses (usually referred to as inheritance) and virtual procedures, combined with safe referencing and mechanisms for bringing into a program collections of program structures described under a common class heading (prefixed blocks).

The Simula languages were developed at the Norwegian Computing Center, Oslo, Norway by Ole-Johan Dahl and Kristen Nygaard. Nygaard's work in Operational Research in the 1950s and early 1960s created the need for precise tools for the description and simulation of com-

plex man-machine systems. In 1961 the idea emerged for developing a language that both could be used for system description (for people) and for system prescription (as a computer program through a compiler). Such a language had to contain an algorithmic language, and Dahl's knowledge of compilers became essential … When the inheritance mechanism was invented in 1967, Simula 67 was developed as a general programming language that also could be specialized for many domains, including system simulation.

Generally, object-oriented classes are used to define agent templates and instantiated objects are used to implement specific agents. Agent environment templates and spatial relationships patterns are also typically implemented using classes. Recursive environment nesting as well as abstract spaces, physical spaces, and connectedness can all be represented in relatively straightforward ways. Instantiated objects are used to implement specific agent environments and spatial relationships in individual models. Within these models, model configurations are also commonly implemented as objects instantiated from one or more classes. However, as with unstructured and structured languages, object-oriented languages generally do not provide special support for application data configuration, program output collection, or program results analysis. As such, these tasks usually need to be manually implemented by model developers. Regardless of this, the ability to bind data and procedures provides such a straightforward method for implementing agents that most agent-based models are written using object-oriented languages.

It should be noted that traditional object-oriented languages do not provide a means to modify class and object structures once a program begins to execute. Newer 'dynamic' object-oriented languages such as Groovy [22] offer this capability. This potentially allows agents to gain and lose data items and methods during the execution of a model based on the flow of events in a simulation. This in turn offers the possibility of implementing modules with fuzzy boundaries that are flexible enough to change structure over time.

As discussed in the previous section, logic-based languages offer an alternative to the progression formed by unstructured, structured, and object-oriented languages. Logic-based languages can provide a form of direct support for binding data (e. g., asserted propositions) with actions (e. g., logical predicates), sometimes including the use of higher-order programming. In principle, each agent can be implemented as a complex predicate with multiple nested sub-terms. The sub-terms, which may contain unresolved variables, can then be activated and resolved as needed during model execution. Agent templates which are analogous to object-oriented classes can be implemented using the same approach but with a larger number of unresolved variables. Agent environments and the resulting relationships between agents can be formed in a similar way. Since each of these constructs can be modified at any time, the resulting system can change structure over time and may even allow fuzzy boundaries. In practice this approach is rarely, if ever, used. As with the previously discussed approaches, logic-based languages usually do not provide special support for application data configuration, program output collection, or program results analysis so these usually need to be manually developed.

Functional languages offer yet another alternative to the previously discussed languages. Like logic-based and object-oriented languages, functional languages often provide a form of direct support for binding data with behaviors. This support often leverages the fact that most functional languages support higher-order programming. As a result, the data is usually in the form of nested lists of values and functions while the behaviors themselves are implemented in the form of functions. Agent templates (i. e., 'classes'), agent environments, and agent relationships can be implemented similarly. Each of the lists can be dynamically changed during a simulation run so the model structure can evolve and can potentially have fuzzy boundaries. Unlike the other languages discussed so far, a major class of functional languages, namely those designed for computational mathematics usually include sophisticated support for program output collection and results analysis. An example is *Mathematica* (Wolfram [49]). If the application data is configured in mathematically regular ways then these systems may also provide support for application data setup.

Example Computer Languages Useful for Agent-Based Modeling

Domain-Specific Languages

Domain-specific languages (DSL's) are computer languages that are highly customized to support a well defined application area or 'domain'. DSL's commonly include a substantial number of keywords that are nouns and verbs in the area of application as well as overall structures and execution patterns that correspond closely with the application area. DSL's are intended to allow users to write in a language that is closely aligned with their area of expertise.

DSL's often gain their focus by losing generality. For many DSL's there are activities that can be programmed in most computer languages that cannot be programmed

in the given DSL. This is consciously done to simplify the DSL's design and make it easier to learn and use. If a DSL is properly designed then the loss of generality is often inconsequential for most uses since the excluded activities are chosen to be outside the normal range of application. However, even the best designed DSL's can occasionally be restrictive when the bounds of the language are encountered. Some DSL's provide special extension points that allow their users to program in a more general language such as C or Java when the limits of the DSL are reached. This feature is extremely useful, but requires more sophistication on the part of the user in that they need to know and simultaneously use both the DSL and the general language.

DSL's have the potential to implement specific features to support 'design patterns' within a given domain. Design patterns form a "common vocabulary" describing tried and true solutions for commonly faced software design problems (Coplien [6]). Software design patterns were popularized by Gamma et al. [13]. North and Macal [29] describe three design patterns for agent-based modeling itself.

In principle, DSL's can be unstructured, structured, object-oriented, logic-based, or functional. In practice, DSL's are often structured languages or object-oriented languages and occasionally are functional languages. Commonly used ABM DSL's include business-oriented languages (e. g., spreadsheet programming tools); science and engineering languages (e. g., *Mathematica*); and dedicated agent-based modeling languages (e. g., NetLogo).

Business Languages Some of the most widely used business computer languages are those available in spreadsheet packages. Spreadsheets are usually programmed using a 'macro language'. As discussed further in North and Macal [29], any modern spreadsheet program can be used to do basic agent-based modeling. The most common convention is to associate each row of a primary spreadsheet worksheet with an agent and use consecutive columns to store agent properties. Secondary worksheets are then used to represent the agent environment and to provide temporary storage for intermediate calculations. A simple loop is usually used to scan down the list of agents and to allow each one to execute in turn. The beginning and end of the scanning loop are generally used for special setup activities before and special cleanup activities after each round. An example agent spreadsheet from North and Macal [29] is shown in Fig. 1and Fig. 2. Agent spreadsheets have both strengths and weaknesses compared to the other ABM tools. Agent spreadsheets tend to be easy to build but they also tend to have limited capabilities. This balance makes spreadsheets ideal for agent-based model exploration, scoping, and prototyping. Simple agent models can be implemented on the desktop using environments outside of spreadsheets as well.

Science and Engineering Languages Science and engineering languages embodied in commercial products such as *Mathematica*, MATLAB, Maple, and others can be used as a basis for developing agent-based models. Such systems usually have a large user base, are readily available on desktop platforms, and are widely integrated into academic training programs. They can be used as rapid prototype development tools or as components of large-scale modeling systems. Science and engineering languages have been applied to agent-based modeling. Their advantages include a fully integrated development environment, their interpreted (as opposed to compiled) nature provides immediate feedback to users during the development process, and a packaged user interface. Integrated tools provide support for data import and graphical display. Macal [25] describes the use of *Mathematica* and MATLAB in agent-based simulation and Macal and Howe [26] detail investigations into linking *Mathematica* and MATLAB to the Repast ABM toolkit to make use of Repast's simulation scheduling algorithms. In the following sections we focus on MATLAB and *Mathematica* as representative examples of science and engineering languages.

MATLAB and *Mathematica* are both examples of Computational Mathematics Systems (CMS). CMS allow users to apply powerful mathematical algorithms to solve problems through a convenient and interactive user interface. CMS typically supply a wide range of built-in functions and algorithms. MATLAB, *Mathematica*, and Maple are examples of commercially available CMS whose origins go back to the late 1980s. CMS are structured in two main parts: (1) the user interface that allows dynamic user interaction, and (2) the underlying computational engine, or kernel, that performs the computations according to the user's instructions. Unlike conventional programming languages, CMS are interpreted instead of compiled, so there is immediate feedback to the user, but some performance penalty is paid. The underlying computational engine is written in the C programming language for these systems, but C coding is unseen by the user. The most recent releases of CMS are fully integrated systems, combining capabilities for data input and export, graphical display, and the capability to link to external programs written in conventional languages such as C or Java using inter-process communication protocols. The powerful features of CMS, their convenience of use, the need to

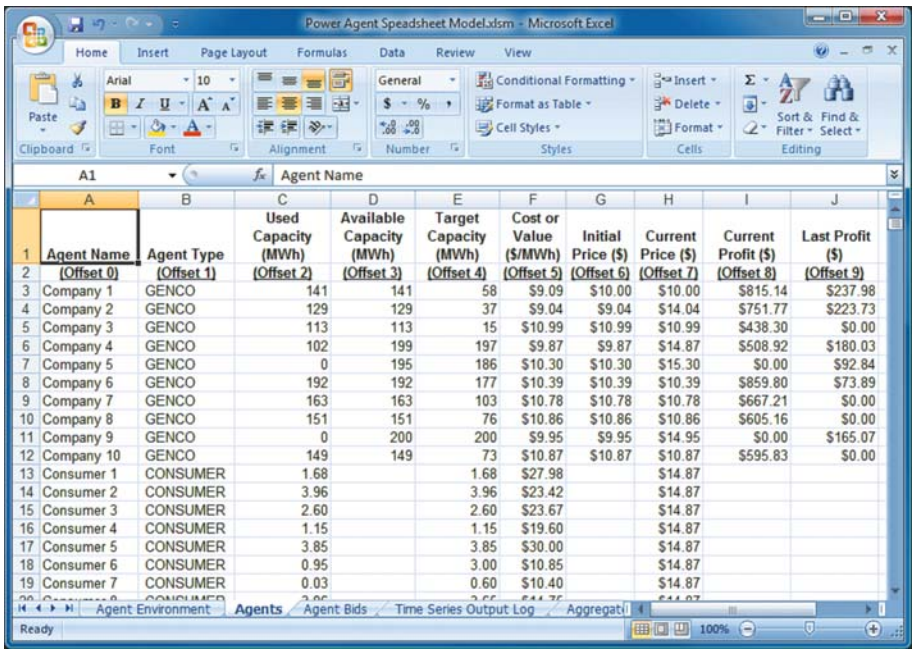

	Agent Name (Offset 0)	Agent Type (Offset 1)	Used Capacity (MWh) (Offset 2)	Available Capacity (MWh) (Offset 3)	Target Capacity (MWh) (Offset 4)	Cost or Value ($/MWh) (Offset 5)	Initial Price ($) (Offset 6)	Current Price ($) (Offset 7)	Current Profit ($) (Offset 8)	Last Profit ($) (Offset 9)
3	Company 1	GENCO	141	141	58	$9.09	$10.00	$10.00	$815.14	$237.98
4	Company 2	GENCO	129	129	37	$9.04	$9.04	$14.04	$751.77	$223.73
5	Company 3	GENCO	113	113	15	$10.99	$10.99	$10.99	$438.30	$0.00
6	Company 4	GENCO	102	199	197	$9.87	$9.87	$14.87	$508.92	$180.03
7	Company 5	GENCO	0	195	186	$10.30	$10.30	$15.30	$0.00	$92.84
8	Company 6	GENCO	192	192	177	$10.39	$10.39	$10.39	$859.80	$73.89
9	Company 7	GENCO	163	163	103	$10.78	$10.78	$10.78	$667.21	$0.00
10	Company 8	GENCO	151	151	76	$10.86	$10.86	$10.86	$605.16	$0.00
11	Company 9	GENCO	0	200	200	$9.95	$9.95	$14.95	$0.00	$165.07
12	Company 10	GENCO	149	149	73	$10.87	$10.87	$10.87	$595.83	$0.00
13	Consumer 1	CONSUMER	1.68		1.68	$27.98		$14.87		
14	Consumer 2	CONSUMER	3.96		3.96	$23.42		$14.87		
15	Consumer 3	CONSUMER	2.60		2.60	$23.67		$14.87		
16	Consumer 4	CONSUMER	1.15		1.15	$19.60		$14.87		
17	Consumer 5	CONSUMER	3.85		3.85	$30.00		$14.87		
18	Consumer 6	CONSUMER	0.95		3.00	$10.85		$14.87		
19	Consumer 7	CONSUMER	0.03		0.60	$10.40		$14.87		

Agent Based Modeling and Computer Languages, Figure 1
An example agent spreadsheet [29]

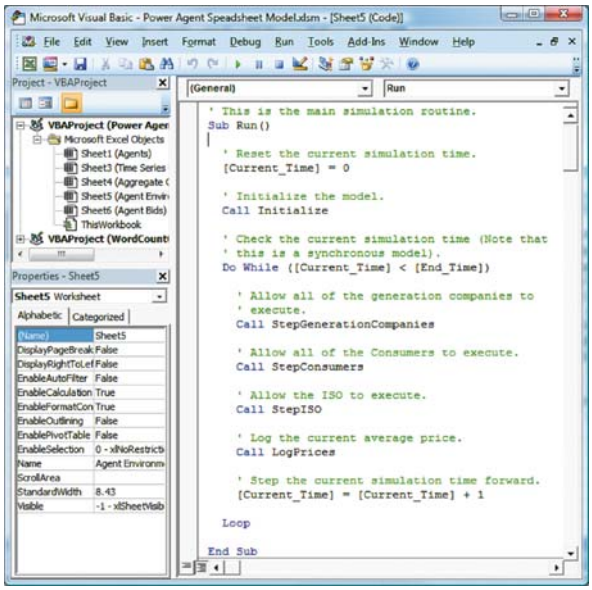

```
' This is the main simulation routine.
Sub Run()

  ' Reset the current simulation time.
  [Current_Time] = 0

  ' Initialize the model.
  Call Initialize

  ' Check the current simulation time (Note that
  ' this is a synchronous model).
  Do While ([Current_Time] < [End_Time])

    ' Allow all of the generation companies to
    ' execute.
    Call StepGenerationCompanies

    ' Allow all of the Consumers to execute.
    Call StepConsumers

    ' Allow the ISO to execute.
    Call StepISO

    ' Log the current average price.
    Call LogPrices

    ' Step the current simulation time forward.
    [Current_Time] = [Current_Time] + 1

  Loop

End Sub
```

Agent Based Modeling and Computer Languages, Figure 2
An example agent spreadsheet code [29]

learn only a limited number of instructions on the part of the user, and the immediate feedback provided to users are features of CMS that make them good candidates for developing agent-based simulations.

A further distinction can be made among CMS. A subset of CMS are what is called Computational Algebra Systems (CAS). CAS are computational mathematics systems that calculate using symbolic expressions. CAS owe their origins to the LISP programming language, which was the earliest functional programming language [24]. Macsyma (www.scientek.com/macsyma) and Scheme [37] (www.swiss.ai.mit.edu/projects/scheme) are often mentioned as important implementations leading to present day CAS. Typical uses of CAS are equation solving, symbolic integration and differentiation, exact calculations in linear algebra, simplification of mathematical expressions, and variable precision arithmetic. Computational mathematics systems consist of numeric processing systems or symbolic processing systems, or possibly a combination of both. Especially when algebraic and numeric capabilities are combined into a multi-paradigm programming environment, new modeling possibilities open up for developing sophisticated agent-based simulations with minimal coding.

Mathematica *Mathematica* is a commercially available numeric processing system with enormous integrated numerical processing capability (http://www.wolfram.com). Beyond numeric processing, *Mathematica* is a fully functional programming language. Unlike MATLAB, *Mathematica* is a symbolic processing system that uses term replacement as its primary operation. Symbolic processing

means that variables can be used before they have values assigned to them; in contrast a numeric processing language requires that every variable have a value assigned to it before it is used in the program. In this respect, although *Mathematica* and MATLAB may appear similar and share many capabilities, *Mathematica* is fundamentally much different than MATLAB, with a much different style of programming and ultimately with a different set of capabilities applicable to agent-based modeling.

Mathematica's symbolic processing capabilities allow one to program in multiple programming styles, either as alternatives or in combination, such as functional programming, logic programming, procedural programming, and even object-oriented programming styles. Like MATLAB, *Mathematica* is also an interpreted language, with the kernel of *Mathematica* running in the background in C. In terms of data types, everything is an expression in *Mathematica*. An expression is a data type with a head and a list of arguments in which even the head of the expression is part of the expression's arguments.

The *Mathematica* user interface consists of a what is referred to as a notebook (Fig. 3). A *Mathematica* notebook is a fully integratable development environment and a complete publication environment. The *Mathematica* Application Programming Interface (API) allows programs written in C, FORTRAN, or Java to interact with *Mathematica*. The API has facilities for dynamically calling routines from *Mathematica* as well as calling *Mathematica* as a computational engine.

Figure 3 shows *Mathematica* desktop notebook environment. A *Mathematica* notebook is displayed in its own window. Within a notebook, each item is contained in a cell. The notebook cell structure has underlying coding that is accessible to the user.

In *Mathematica*, a network representation consists of combining lists of lists, or more generally expressions of expressions, to various depths. For example, in *Mathematica*, an agent can be represented explicitly as an expression that includes a head named agent, a sequence of agent attributes, and a list of the agent's neighbors. Agent data and methods are linked together by the use of what are called up values.

Example references for agent-based simulation using *Mathematica* include Gaylord and Davis [15], Gaylord and Nishidate [16], and Gaylord and Wellin [17]. Gaylord and D'Andria [14] describe applications in social agent-based modeling.

MATLAB The MATrix LABoratory (MATLAB) is a numeric processing system with enormous integrated numerical processing capability (http://www.mathworks.com). It uses a scripting-language approach to programming. MATLAB is a high-level matrix/array language with control flow, functions, data structures, input/output, and object-oriented programming features. The user interface consists of the MATLAB Desktop, which is a fully integrated and mature development environment. There is an application programming interface (API). The MATLAB API allows programs written in C, Fortran, or Java to interact with MATLAB. There are facilities for calling routines from MATLAB (dynamic linking) as well as routines for calling MATLAB as a computational engine, as well as for reading and writing specialized MATLAB files.

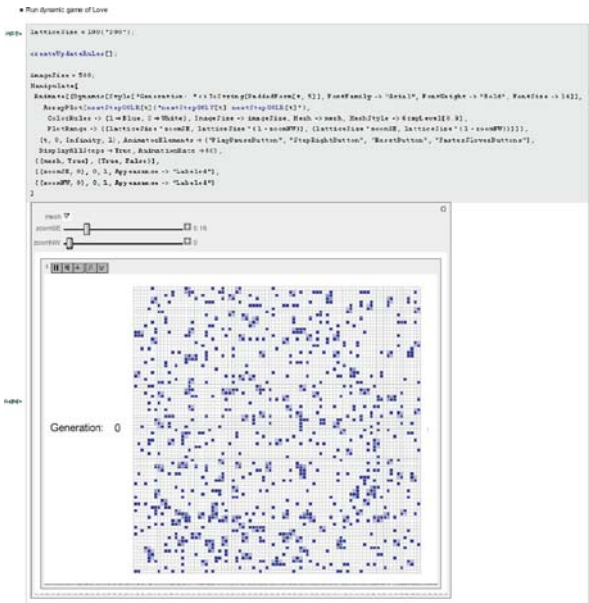

Agent Based Modeling and Computer Languages, Figure 3
Example Mathematica cellular automata model

Figure 4 shows the MATLAB Desktop environment illustrating the Game of Life, which is a standard MATLAB demonstration. The desktop consist of four standard windows: a command window, which contains a command line, the primary way of interacting with MATLAB, the workspace, which indicates the values of all the variables currently existing in the session, a command history window that tracks the entered command, and the current directory window. Other windows allow text editing of programs and graphical output display.

When it comes to agent-based simulation, as in most types of coding, the most important indicator of the power of a language for modeling is the extent of and the sophistication of the allowed data types and data structures. As Sedgewick [36] observes:

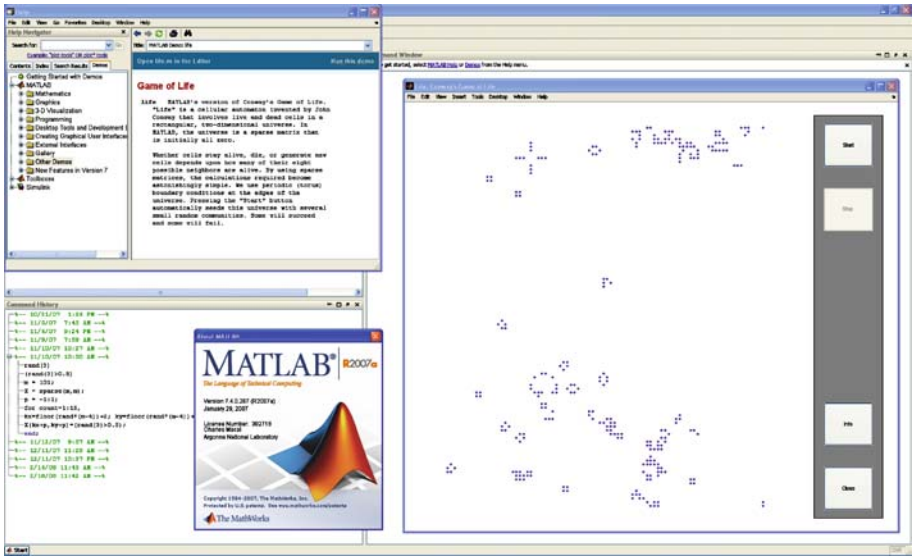

Agent Based Modeling and Computer Languages, Figure 4
Example MATLAB cellular automata model

> For many applications, the choice of the proper data structure is really the only major decision involved in the implementation; once the choice has been made only very simple algorithms are needed [36].

The flexibility of data types plays an important role in developing large-scale, extensible models for agent-based simulation. In MATLAB the primary data type is the double array, which is essentially a two-dimensional numeric matrix. Other data types include logical arrays, cell arrays, structures, and character arrays.

For agent-based simulations that define agent relationships based on networks, connectivity of the links defines the scope of agent interaction and locally available information. Extensions to modeling social networks require the use of more complex data structures than the matrix structure commonly used for grid representations. Extensions from grid topologies to network topologies are straightforward in MATLAB and similarly in *Mathematica*. In MATLAB, a network representation consists of combining cell arrays or structures in various ways.

The MATLAB desktop environment showing the Game of Life demonstration appears in Fig. 4. The Game of Life is a cellular automaton invented by mathematician John Conway that involves live and dead cells in cellular automata grid. In MATLAB, the agent environment is a sparse matrix that is initially set to all zeros. Whether cells stay alive, die, or generate new cells depends upon how many of their eight possible neighbors are alive. By using sparse matrices, the calculations required become very simple. Pressing the "Start" button automatically seeds this universe with several small random communities and initiates a series of cell updates. After a short period of simulation, the initial random distribution of live (i. e., highlighted) cells develops into sets of sustainable patterns that endure for generations.

Several agent-based models using MATLAB have been published in addition to the Game of Life. These include a model of political institutions in modern Italy [3], a model of pair interactions and attitudes [34], a bargaining model to simulate negotiations between water users [43], and a model of sentiment and social mitosis based on Heider's Balance Theory [18,46]. The latter model uses Euler, a MATLAB-like language. Thorngate argues for the use of MATLAB as an important tool to teach simulation programming techniques [42].

Dedicated Agent-Based Modeling Languages Dedicated agent-based modeling languages are DSL's that are designed to specifically support agent-based modeling. Several such languages currently exist. These languages are functionally differentiated by the underlying assumptions their designers made about the structures of agent-based models. The designers of some of these languages assume quite a lot about the situations being modeled and use this information to provide users with pre-completed or template components. The designers of other languages make comparatively fewer assumptions and encourage users to

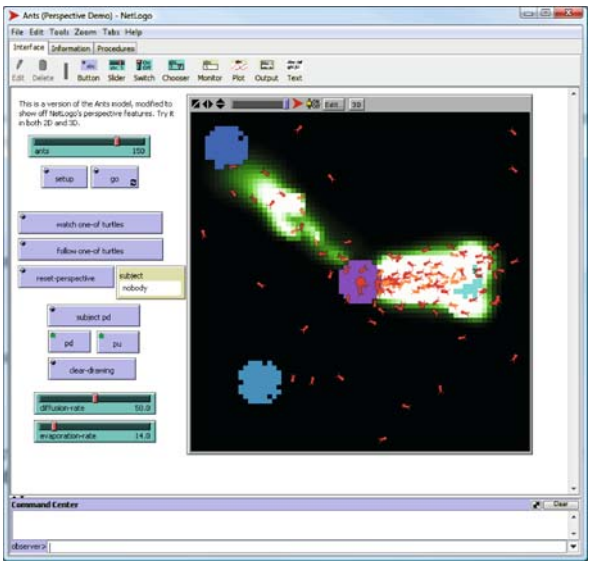

Agent Based Modeling and Computer Languages, Figure 5
Example NetLogo ant colony model [48]

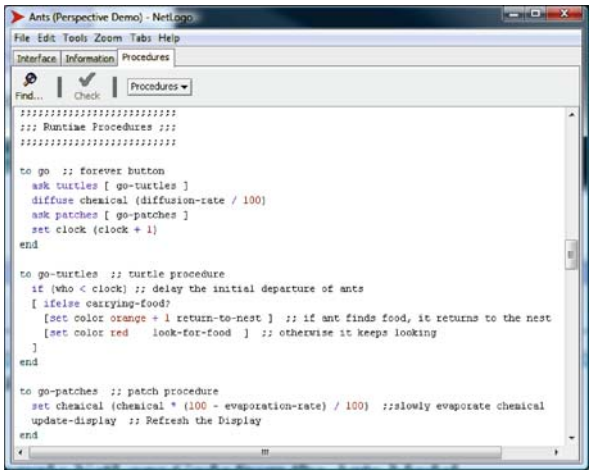

Agent Based Modeling and Computer Languages, Figure 6
Example NetLogo code from the ant colony model [48]

implement a wider range of models. However, more work is often needed to build models in these systems. This article will discuss two selected examples, namely NetLogo and the visual interface for Repast Simphony.

NetLogo NetLogo is an education-focused ABM environment (Wilensky [48]). The NetLogo language uses a modified version of the Logo programming language (Harvey [19]). NetLogo itself is Java-based and is free for use in education and research. More information on NetLogo and downloads can be found at http://ccl.northwestern.edu/netlogo/.

NetLogo is designed to provide a basic computational laboratory for teaching complex adaptive systems concepts. NetLogo was originally developed to support teaching, but it can be used to develop a wider range of applications. NetLogo provides a graphical environment to create programs that control graphic 'turtles' that reside in a world of 'patches' that is monitored by an 'observer'. NetLogo's DSL is limited to its turtle and patch paradigm. However, NetLogo models can be extended using Java to provide for more general programming capabilities. An example NetLogo model of an ant colony [48] (center) feeding on three food sources (upper left corner, lower left corner, and middle right) is shown in Fig. 5. Example code [48] from this model is shown in Fig. 6.

Repast Simphony Visual Interface The Recursive Porous Agent Simulation Toolkit (Repast) is a free and open source family of agent-based modeling and simulation platforms (ROAD [44]). Information on Repast and free downloads can be found at http://repast.sourceforge.net/. Repast Simphony (Repast S) is the newest member of the Repast family [30,32]. The Java-based Repast S system includes advanced features for specifying, executing, and analyzing agent-based simulations. Repast Simphony offers several methods for specifying agents and agent environments including visual specification, specification with the dynamic object-oriented Groovy language [22], and specification with Java. In principle, Repast S's visual DSL can be used for any kind of programming, but models beyond a certain level of complexity are better implemented in Groovy or Java. As discussed later, Groovy and Java are general purpose languages. All of Repast S's languages can be fluidly combined in a single model. An example Repast S flocking model is shown in Fig. 7 [33]. The visual specification approach uses a tree to define model contents (Fig. 8 middle panel with "model.score" header) and a flowchart to define agent behaviors (Fig. 8 right side panel with "SearchExampleAgent.agent" header). In all cases, the user has a choice of a visual rich point-and-click interface or a 'headless' batch interface to execute models.

General Languages Unlike DSL's, general languages are designed to take on any programming challenge. However, in order to meet this challenge they are usually more complex than DSL's. This tends to make them more difficult to learn and use. Lahtinen et al. [23] documents some of the challenges users face in learning general purpose programming languages. Despite these issues, general purpose programming languages are essential for al-

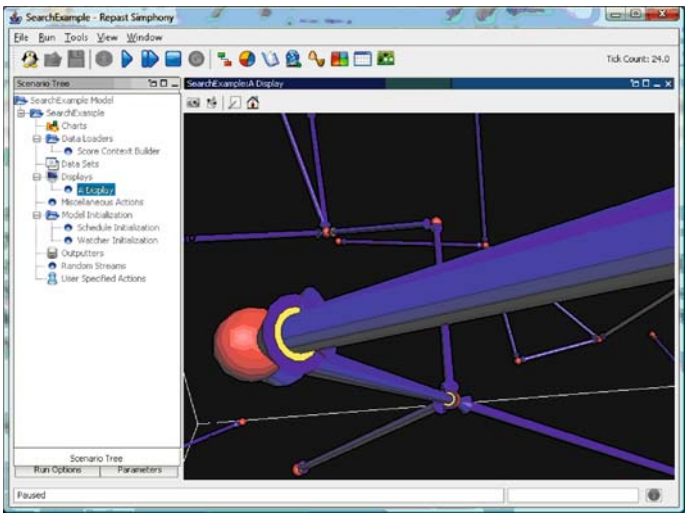

Agent Based Modeling and Computer Languages, Figure 7
Example Repast Simphony flocking model [33]

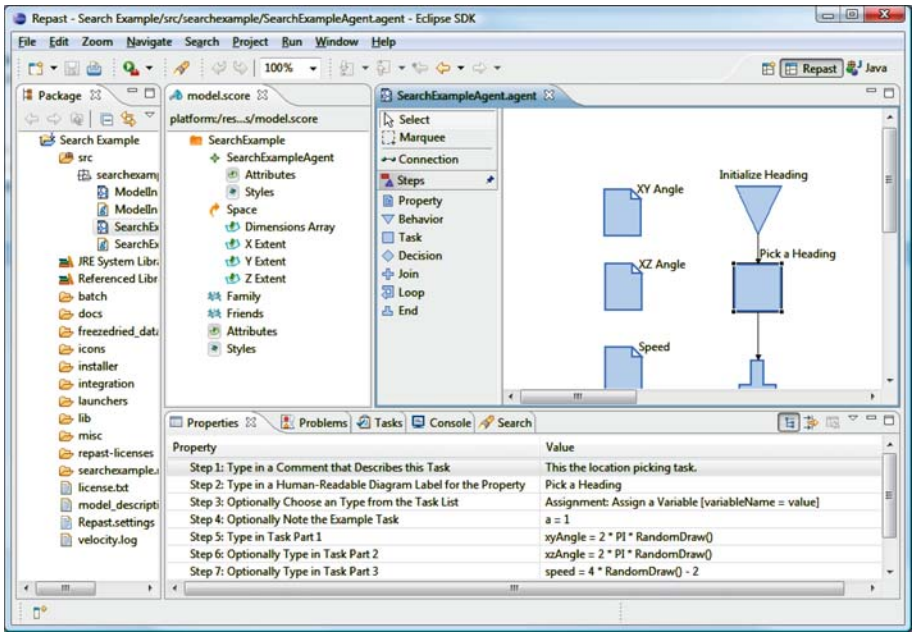

Agent Based Modeling and Computer Languages, Figure 8
Example Repast Simphony visual behavior from the flocking model [33]

lowing users to access the full capabilities of modern computers. Naturally, there are a huge number of general purpose programming languages. This article considers these options from two perspectives. First, general language toolkits are discussed. These toolkits provide libraries of functions to be used in a general purpose host language. Second, the use of three raw general purpose languages, namely Java, C#, and C++, is discussed.

General Language Toolkits As previously stated, general language toolkits are libraries that are intended be used in a general purpose host language. These toolkits usually provide model developers with software for functions such as simulation time scheduling, results visualization, results logging, and model execution as well as domain-specific tools [31]. Users of raw general purpose languages have to write all of the needed features by themselves by hand.

A wide range of general language toolkits currently exist. This article will discuss two selected examples, namely Swarm and the Groovy and Java interfaces for Repast Simphony.

Swarm Swarm [28] is a free and open source agent-based modeling library. Swarm seeks to create a shared simulation platform for agent modeling and to facilitate the development of a wide range of models. Users build simulations by incorporating Swarm library components into their own programs. Information on Swarm and free downloads can be found at http://www.swarm.org/ from Marcus Daniels [9]:

> Swarm is a set of libraries that facilitate implementation of agent-based models. Swarm's inspiration comes from the field of Artificial Life. Artificial Life is an approach to studying biological systems that attempts to infer mechanism from biological phenomena, using the elaboration, refinement, and generalization of these mechanisms to identify unifying dynamical properties of biological systems … To help fill this need, Chris Langton initiated the Swarm project in 1994 at the Santa Fe Institute. The first version was available by 1996, and since then it has evolved to serve not only researchers in biology, but also anthropology, computer science, defense, ecology, economics, geography, industry, and political science.

The Swarm simulation system has two fundamental components. The core component runs general-purpose simulation code written in Objective-C, Tcl/Tk, and Java. This component handles most of the behind the scenes details. The external wrapper components run user-specific simulation code written in either Objective-C or Java. These components handle most of the center stage work. An example Swarm supply chain model is shown in Fig. 9.

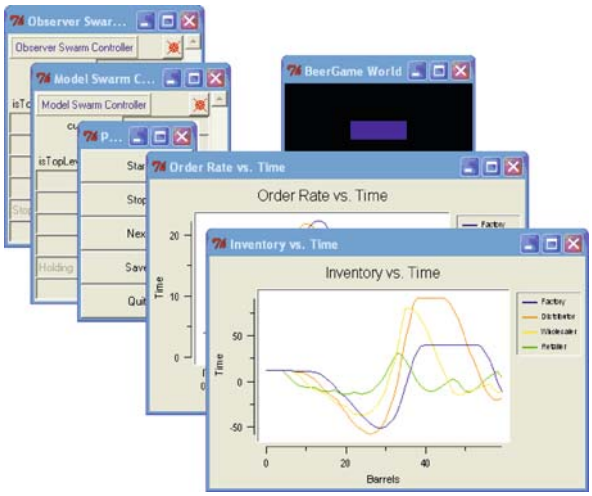

Agent Based Modeling and Computer Languages, Figure 9
Example Swarm supply chain model

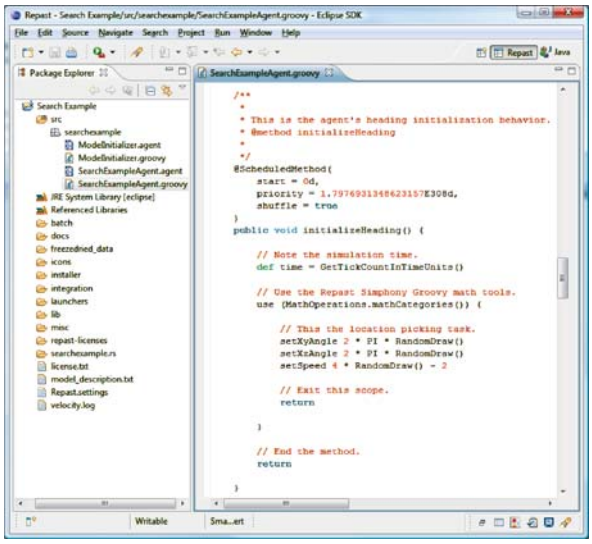

Agent Based Modeling and Computer Languages, Figure 10
Example Repast Simphony Groovy code from the flocking model in Fig. 7 [33]

Repast Simphony Java and Groovy As previously discussed, Repast is a free and open source family of agent-based modeling and simulation platforms (ROAD [44]). Information on Repast and free downloads can be found at http://repast.sourceforge.net/. Repast S is the newest member of the Repast family [30,32]. The Java-based Repast S system includes advanced features for specifying, executing, and analyzing agent-based simulations. An example Repast S flocking model is shown in Fig. 7 [33].

Repast Simphony offers several intermixable methods for specifying agents and agent environments including visual specification, specification with the dynamic object-oriented Groovy language [22], and specification with Java. The Groovy approach uses the dynamic object-oriented Groovy language as shown in Fig. 10. The Java approach for an example predator-prey model is shown in Fig. 11.

Java Java [12] is a widely used object-oriented programming language that was developed and is maintained by Sun Microsystems. Java is known for its widespread 'cross-platform' availability on many different types of hard-

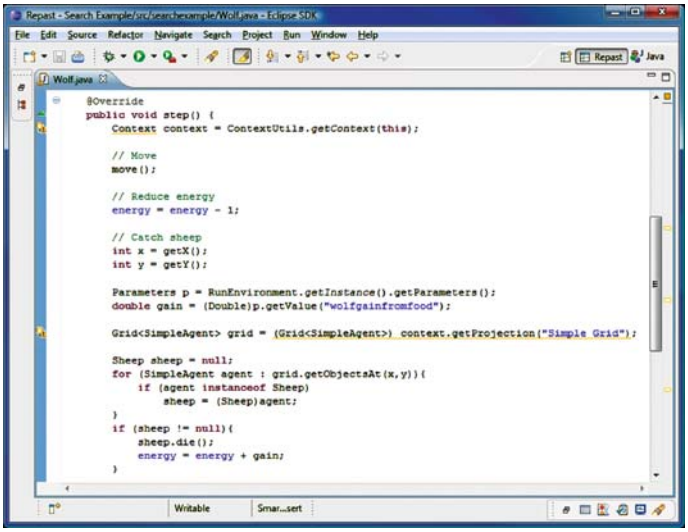

Agent Based Modeling and Computer Languages, Figure 11
Example Repast Simphony Java code for a predator-prey model [41]

ware and operating systems. This capability comes from Java's use of a 'virtual machine' that allows binary code or 'bytecode' to have a consistent execution environment on many different computer platforms. A large number of tools are available for Java program development including the powerful Eclipse development environment [11] and many supporting libraries. Java uses reflection and dynamic method invocation to implement a variant of higher-order programming. Reflection is used for runtime class structure examination while dynamic method invocation is used to call newly referenced methods at runtime. Java's object-orientation, cross platform availability, reflection, and dynamic method invocation along with newer features such as annotations for including metadata in compiled code, generics for generalizing class, and aspects to implement dispersed but recurrent tasks make it a good choice for agent-based model development.

C# C# [1] is an object-oriented programming language that was developed and is maintained by Microsoft. C# is one of many languages that can be used to generate Microsoft .NET Framework code or Common Intermediate Language (CIL). Like Java bytecode, CIL is run using a 'virtual machine' that potentially gives it a consistent execution environment on different computer platforms. A growing number of tools are emerging to support C# development. C#, and the Microsoft .NET Framework more generally, are in principle cross platform, but in practice they are mainly executed under Microsoft Windows.

The Microsoft .NET Framework provides for the compilation into CIL of many different languages such as C#, Managed C++, and Managed Visual Basic to name just a few. Once these languages are compiled to CIL, the resulting modules are fully interoperable. This allows users to conveniently develop integrated software using a mixture of different languages. Like Java, C# supports reflection and dynamic method invocation for higher-order programming. C#'s object-orientation, multi-lingual integration, generics, attributes for including metadata in compiled code, aspects, reflection, and dynamic method invocation make it well suited for agent-based model development, particularly on the Microsoft Windows platform.

C++ C++ is a widely used object-oriented programming language that was created by Bjarne Stroustrup (Stroustrup [39]) at AT&T. C++ is widely noted for both its object-oriented structure and its ability to be easily compiled into native machine code. C++ gives users substantial access to the underlying computer but also requires substantial programming skills.

Most C++ compilers are actually more properly considered C/C++ compilers since they can compile non-object-oriented C code as well as object-oriented C++ code. This allows sophisticated users to the opportunity highly optimize selected areas of model code. However, this also opens the possibility of introducing difficult to resolve errors and hard to maintain code. It is also more difficult to port C++ code from one computer architecture to an-

other than it is for virtual machine-based languages such as Java.

C++ can use a combination of Runtime Type Identification (RTTI) and function pointers to implement higher-order programming. Similar to the Java approach, C++ RTTI can be used for runtime class structure examination while function pointers can be used to call newly referenced methods at runtime. C++'s object-orientation, RTTI, function pointers, and low-level machine access make it a reasonable choice for the development of extremely large or complicated agent-based models.

Future Directions

Future developments in computer languages could have enormous implications for the development of agent-based modeling. Some of the challenges of agent-based modeling for the future include (1) scaling up models to handle large numbers of agents running on distributed heterogeneous processors across the grid, (2) handling the large amounts of data generated by agent models and making sense out of it, and (3) developing user-friendly interfaces and modular components in a collaborative environment that can be used by domain experts with little or no knowledge of standard computer coding techniques. Visual and natural language development environments that can be used by non-programmers are continuing to advance but remain to be proven at reducing the programming burden. There are a variety of next steps for the development of computer languages for agent-based modeling including the further development of DSL's; increasing visual modeling capabilities; and the development of languages and language features that better support pattern-based development. DSL's are likely to become increasing available as agent-based modeling grows into a wider range of domains. More agent-based modeling systems are developing visual interfaces for specifying model structures and agent behaviors. Many of these visual environments are themselves DSL's. The continued success of agent-based modeling will likely yield an increasing number of design patterns. Supporting and even automating implementations of these patterns may form a natural source for new language features. Many of these new features are likely to be implemented within DSL's.

Bibliography

1. Archer T (2001) Inside C#. Microsoft Press, Redmond
2. Backus J, Bauer F, Green J, Katz C, McCarthy J, Naur P, Perlis A, Rutishauser H, Samuelson K, Vauquois B, Wegstein J, van Wijngaarden A, Woodger M (1963) Revised Report on the Algorithmic Language ALGOL 60. In: Naur P (ed) Communications of the ACM, vol 6. ACM, New York, pp 1–17
3. Bhavnani R (2003) Adaptive agents, political institutions and civic traditions in modern italy. J Artif Soc Soc Simul 6(4). Available at http://jasss.soc.surrey.ac.uk/6/4/1.html
4. Bonabeau E (2001) Agent-based modeling: methods and techniques for simulating human systems. Proc Natl Acad Sci 99(3):7280–7287
5. Casti J (1997) Would-be worlds: how simulation is changing the world of science. Wiley, New York
6. Coplien J (2001) Software patterns home page. Available as http://hillside.net/patterns/
7. Dahl O-J, Nygaard K (1966) SIMULA – an ALGOL-based simulation language. Commun ACM 9:671–678
8. Dahl O-J, Nygaard K (2001) How object-oriented programming started. Available at http://heim.ifi.uio.no/~kristen/FORSKNINGSDOK_MAPPE/F_OO_start.html
9. Daniels M (1999) Integrating simulation technologies with swarm. In: Proc of the agent 1999 workshop on agent simulation: applications, models, and tools. Argonne National Laboratory, Argonne
10. Dijkstra E (1968) Go to statement considered harmful. Commun ACM 11(3):147–148
11. Eclipse (2008) Eclipse home page. Available at http://www.eclipse.org/
12. Foxwell H (1999) Java 2 software development kit. Linux J
13. Gamma E, Helm R, Johnson R, Vlissides J (1995) Design patterns: elements of reusable object-oriented software. Addison-Wesley, Wokingham
14. Gaylord R, D'Andria L (1998) Simulating society: a mathematica toolkit for modeling socioeconomic behavior. Springer/TELOS, New York
15. Gaylord R, Davis J (1999) Modeling nonspatial social interactions. Math Edu Res 8(2):1–4
16. Gaylord R, Nishidate K (1994) Modeling nature: cellular automata simulations with Mathematica. Springer, New York
17. Gaylord R, Wellin P (1995) Computer simulations with Mathematica: explorations in complex physical and biological systems. Springer/TELOS, New York
18. Guetzkow H, Kotler P, Schultz R (eds) (1972) Simulation in social and administrative science. Prentice Hall, Englewood Cliffs
19. Harvey B (1997) Computer science logo style. MIT Press. Boston
20. Heath J (2005) Methodological individualism. In: Zalta E (ed) Stanford encyclopedia of philosophy. Stanford University, Stanford, Aviable at http://plato.stanford.edu
21. Jennings N (2000) On agent-based software engineering. Artif Intell 117:277–296
22. Koenig D, Glover A, King P, Laforge G, Skeet J (2007) Groovy in action. Manning Publications, Greenwhich
23. Lahtinen E, Ala-Mutka K, Jarvinen H-M (2005) A study of the difficulties of novice programmers. In: Proc of the 10th annual SIGCSE conference on innovation and technology in computer science education. Caparica, Portugal. ACM
24. McCarthy J (1960) Recursive functions of symbolic expressions and their computation by machine I. J ACM 3:184–195
25. Macal C (2004) Agent-based modeling and social simulation with Mathematica and MATLAB. In: Macal C, Sallach D, North M (eds) Proc of the agent 2004 conference on social dynamics: interaction, reflexivity and emergence. Argonne National Laboratory, Argonne
26. Macal C, Howe T (2005) Linking repast to computational mathematics systems: Mathematica and MATLAB. In: Macal C, Sallach D, North M (eds) Proc of the agent 2005 conference

on generative social processes, models, and mechanisms. Argonne National Laboratory, Argonne
27. Macal C, North M (2007) Tutorial on Agent-based Modeling and Simulation: Desktop ABMS. In: Henderson SG, Biller B, Hsieh MH, Shortle J, Tew JD, Barton RR (eds) Proceedings of the 2007 Winter Simulation Conference, December 9-12, 2007, pp 95–106, http://www.informs-sim.org/wsc07papers/011.pdf
28. Minar N, Burkhart R, Langton C, Askenazi M (1996) The swarm simulation system: a toolkit for building multi-agent simulations. Available at http://alumni.media.mit.edu/~nelson/research/swarm/
29. North M, Macal C (2007) Managing business complexity: discovering strategic solutions with agent-based modeling and simulation. Oxford, New York
30. North M, Howe T, Collier N, Vos J (2005) Repast simphony runtime system. In: Macal C, North M, Sallach D (eds) Proc of the agent 2005 conference on generative social processes, models, and mechanisms. Argonne National Laboratory, Argonne
31. North M, Collier N, Vos R (2006) Experiences creating three implementations of the repast agent modeling toolkit. ACM Trans Model Comput Simul 16(1):1–25. ACM. New York
32. North M, Tatara E, Collier N, Ozik J (2007) Visual agent-based model development with repast simphony. In: Macal C, North M, Sallach D (eds) Proc of the agent 2007 conference on complex interaction and social emergence. Argonne National Laboratory, Argonne
33. North M, Howe T, Collier N, Tatara E, Ozik J, Macal C (2008) Search and emergence in agent-based models. In: Agent-based societies: Societal and cultural interactions. IGI Global Publishing, New York
34. Pearson D, Boudarel M-R (2001) Pair interactions: real and perceived attitudes. J Artif Soc Soc Simul 4(4). Available at http://www.soc.surrey.ac.uk/JASSS/4/4/4.html
35. Reynolds J (1998) Definitional Interpreters for Higher-Order Programming Languages. In: Higher-Order and Symbolic Computation. Kluwer, Boston, pp 363–397
36. Sedgewick R (1988) Algorithms, 2nd edn. Addison-Wesley, Reading, pp 657
37. Springer G, Freeman D (1989) Scheme and the art of programming. McGraw-Hill, New York
38. Stevens W, Meyers G, Constantine L (1974) Structured design. IBM Syst J 2
39. Stroustrup B (2008) Bjarne Stroustrup's FAQ. Available at http://www.research.att.com/~bs/bs_faq.html#invention
40. Swarm Development Group (2008) SDG home page. Available at http://www.swarm.org/
41. Tatara E, North M, Howe T, Collier N, Vos J (2006) An Introduction to Repast Simphony Modeling using a simple Predator-Prey Example. In: Proceedings of the Agent 2006 Conference on Social Agents: Results and Prospects. Argonne National Laboratory, Argonne
42. Thorngate W (2000) Teaching social simulation with MATLAB. J Artif Soc Soc Simul 3(1). Available at http://www.soc.surrey.ac.uk/JASSS/3/1/forum/1.html
43. Thoyer S,Morardet S, Rio P, Simon L, Goodhue R, Rausser G (2001) A bargaining model to simulate negotiations between water users. J Artif Soc Soc Simul 4(2). Available at http://www.soc.surrey.ac.uk/JASSS/4/2/6.html
44. ROAD (2008) Repast home page. Available at http://repast.sourceforge.net/
45. van Roy P, Haridi S (2004) Concepts, techniques, and models of computer programming. MIT Press, Cambridge
46. Wang Z, Thorngate W (2003) Sentiment and social mitosis: implications of Heider's balance theory. J Artif Soc Soc Simul 6(3). Available at http://jasss.soc.surrey.ac.uk/6/3/2.html
47. Watson D (1989) High-level languages and their compilers. Addison-Wesley, Wokingham
48. Wilensky U (1999) NetLogo. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL. http://ccl.northwestern.edu/netlogo/
49. Wolfram Research (2008) Mathematica home page. Available at http://www.wolfram.com/

Agent Based Modeling, Large Scale Simulations

Hazel R. Parry
Central Science Laboratory, York, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Large Scale Agent Based Models: Guidelines for Development
Parallel Computing
Example
Future Directions
Acknowledgments
Bibliography

Glossary

Agent A popular definition of an agent, particularly in AI research, is that of Wooldridge [45], pp. 29: "an agent is a computer system that is situated in some environment, and that is capable of autonomous action in this environment in order to meet its design objectives". In particular, it is the autonomy, flexibility, inter-agent communication, reactivity and proactiveness of the agents that distinguishes the paradigm and gives power to agent-based models and multi-agent simulation [15,21]. Multi-agent systems (MAS) comprise of numerous agents, which are given rules by which they act and interact with one another to achieve a set of goals.

Block mapping A method of partitioning an array of elements between nodes of a distributed system, where the array elements are partitioned as evenly as possible into blocks of consecutive elements and assigned to processors. The size of the blocks approximates to

the number of array elements divided by the number of processors.

Complexity Complexity and complex systems pertain to ideas of randomness and irregularity in a system, where individual-scale interactions may result in either very complex or surprisingly simple patterns of behavior at the larger scale. Complex agent-based systems are therefore usually made up of agents interacting in a non-linear fashion. The agents are capable of generating emergent behavioral patterns, of deciding between rules and of relying upon data across a variety of scales. The concept allows for studies of interaction between hierarchical levels rather than fixed levels of analysis.

Cyclic mapping A method of partitioning an array of elements between nodes of a distributed system, where the array elements are partitioned by cycling through each node and assigning individual elements of the array to each node in turn.

Grid Computer 'Grids' are comprised of a large number of disparate computers (often desktop PCs) that are treated as a virtual cluster when linked to one another via a distributed communication infrastructure (such as the internet or an intranet). Grids facilitate sharing of computing, application, data and storage resources. Grid computing crosses geographic and institutional boundaries, lacks central control, and is dynamic as nodes are added or removed in an uncoordinated manner. BOINC computing is a form of distributed computing is where idle time on CPUs may be used to process information (http://boinc.berkeley.edu/).

Ising-type model Ising-type models have been primarily used in the physical sciences. They simulate behavior in which individual elements (e. g., atoms, animals, social behavior, etc.) modify their behavior so as to conform to the behavior of other individuals in their vicinity. Conway's Game of Life is a Ising-type model, where cells are in one of two states: dead or alive. In biology, the technique is used to model neural networks and flocking birds, for example.

Message passing (MP) Message passing (MP) is the principle way by which parallel clusters of machines are programmed. It is a widely-used, powerful and general method of enabling distribution and creating efficient programs [30]. Key advantages of using MP architectures are an ability to scale to many processors, flexibility, 'future-proofing' of programs and portability [29].

Message passing interface (MPI) A computing standard that is used for programming parallel systems. It is implemented as a library of code that may be used to enable message passing in a parallel computing system. Such libraries have largely been developed in C and Fortran, but are also used with other languages such as Java (MPIJava http://www.hpjava.org). It enables developers of parallel software to write parallel programs that are both portable and efficient.

Multiple instruction multiple data (MIMD) Parallelization where different algorithms are applied to different data items on different processors.

Parallel computer architecture A parallel computer architecture consists of a number of identical units that contain CPUs (Central Processing Units) which function as ordinary serial computers. These units, called nodes, are connected to one another (Fig. 1). They may transfer information and data between one another (e. g. via MPI) and simultaneously perform calculations on different data.

Single instruction multiple data (SIMD) SIMD techniques exploit data level parallelism: when a large mass of data of a uniform type needs the same instruction performed on it. An example is a vector or array processor. An application that may take advantage of SIMD is one where the same value is being added (or subtracted) to a large number of data points.

Vector computer/vector processor Vector computers contain a CPU designed to run mathematical operations on multiple data elements simultaneously (rather than sequentially). This form of processing is essentially a SIMD approach. The Cray Y-MP and the Convex C3880 are two examples of vector processors used for supercomputing in the 1980s and 1990s. Today, most recent commodity CPU designs include some vector processing instructions.

Definition of the Subject

'Large scale' simulations in the context of agent-based modelling are not only simulations that are large in terms of the size of the simulation (number of agents simulated), but they are also complex. Complexity is inherent in agent-based models, as they are usually composed of dynamic, heterogeneous, interacting agents. Large scale agent-based models have also be referred to as 'Massively Multi-agent Systems (MMAS)' [18]. MMAS is defined as " 'beyond resource limitation": the number of agents exceeds local computer resources, or the situations are too complex to design/program given human cognitive resource limits' [18], Preface. Therefore, for agent-based modelling 'large scale' is not simply a size problem, it is also a problem of managing complexity to ensure scalability of the agent model.

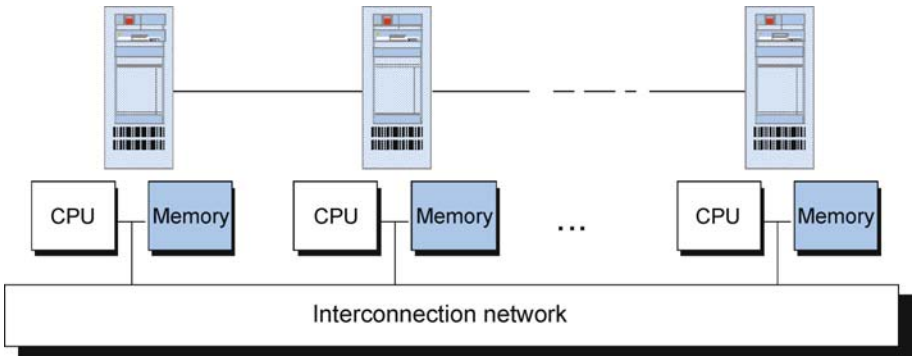

Agent Based Modeling, Large Scale Simulations, Figure 1
A network with interconnected separate memory and processors (after [30], pp. 19)

Multi-agent simulation models increase in scale as the modeller requires many agents to investigate whole system behavior, or the modeller wishes to fully examine the response of a single agent in a realistic context. Two key problems may be introduced by increasing the scale of a multi-agent system: (1) Computational resources limit the simulation time and/or data storage capacity and (2) Agent model analysis may become more difficult. Difficulty in analyzing the model may be due to the model system having a large number of complex components or to memory for model output storage being restricted by computer resources.

Introduction

Many systems that are now simulated using agent-based models are systems where agent numbers are large and potentially complex. These large scale models are constructed under a number of diverse scientific disciplines, with differing histories of agent simulation and methodologies emerging with which to deal with large scale simulations: for example molecular physics, social science (e. g. crowd simulation, city growth), telecommunications, ecology and military research. The primary methodology to emerge is parallel computing, where an agent model is distributed across a number of CPUs to increase the memory and processing power available to the simulation. However, there are a range of potential methods with which to simulate large numbers of agents. Some suggestions are listed in Table 1.

The simplest solution to enable larger scale agent simulation is usually to improve the computer hardware that is used and run the model on a server or invest in a more powerful PC. However, this option may be too costly or may not provide enough scaling so other options may then be considered. Another simple solutions may be to reduce the number of agents or to revert to a simpler modelling approach such as a population model, but both of these solutions would significantly alter the model and the philosophy behind the model, probably not addressing the research question that the model was initially constructed for. There are a number of unique advantages and insights that may be gained from an agent-based approach. Agent simulations are often constructed to enable the analysis of emergent spatial patterns and individual life histories. The realism of an agent-based approach may be lost by using a simpler modelling technique.

The structure of agent simulations (often with asynchronous updating and heterogeneous data types) would mean that running a simulation on a vector computer may make little difference to the simulation performance. This is because an agent model typically has few elements that could take advantage of SIMD: rarely the same value will be added (or subtracted) to a large number of data points. Vector processors are less successful when a program does not have a regular structure, and they don't scale to arbitrarily large problems (the upper limit on the speed of a vector program will be some multiple of the speed of the CPU [30]).

Another relatively simple option is to implement an aggregation of the individual agents into 'super-agents', such as the 'super-individual' approach of Scheffer et al. [35]. The basic concept of this approach is shown in Fig. 2. These 'super-agents' are formed from individual agents that share the same characteristics, such as age and sex. However, it may not be possible to group agents in a simulation in this way and, importantly, this method has been proven ineffective in a spatial context [31,32].

The most challenging solution, to reprogram the model in parallel, is a popular solution due to the shortcomings of the other approaches outlined above. A parallel solution may also have some monetary cost and require

Agent Based Modeling, Large Scale Simulations, Table 1
Potential solutions to implement when faced with a large number of agents to model

Solution	Pro	Con
Reduce the number of agents in order for model to run	No reprogramming of model	Unrealistic population. Alters model behavior
Revert to a population based modelling approach	Could potentially handle any number of individuals	Lose insights from agent approach. Unsuitable for research questions. Construction of entirely new model (non-agent-based)
Invest in an extremely powerful computer	No reprogramming of model	High cost
Run the model on a vector computer	Potentially more efficient as more calculations may be performed in a given time	This approach only works more efficiently with SIMD, probably unsuitable for agent-based models
Super-individuals [35]	Relatively simple solution, keeping model formulation the same	Reprogramming of model. Inappropriate in a spatial context [31,32]
Invest in a powerful computer network and reprogram the model in parallel	Makes available high levels of memory and processing power	High cost. Advanced computing skills required for restructuring of model

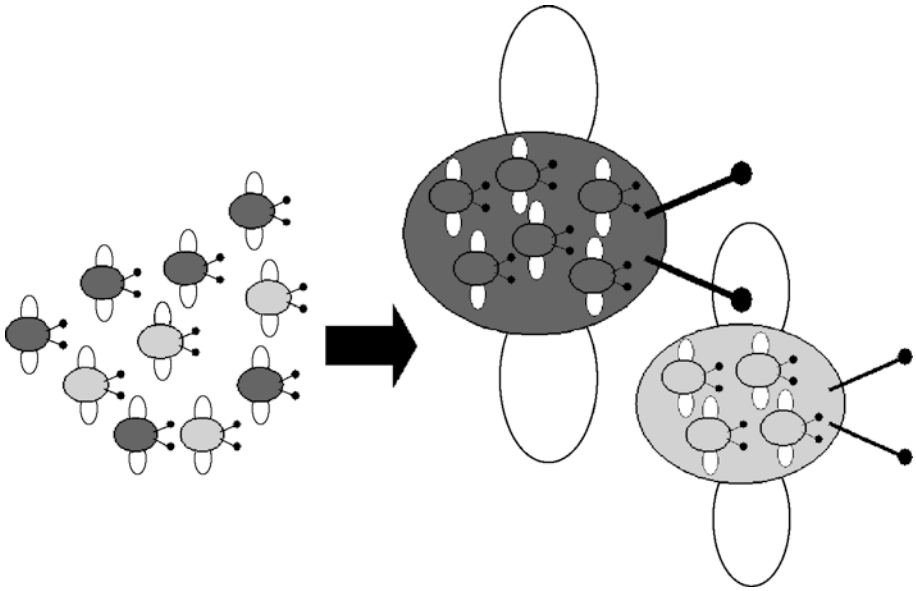

Agent Based Modeling, Large Scale Simulations, Figure 2
'Super-agents': Grouping of individuals into single objects that represent the collective

advanced computing skills to implement, but it can potentially greatly increase the scale of the agent simulation. Extremely large scale object-oriented simulations that simulate individual particles on massively parallel computer systems have been successfully developed in the physical sciences of fluid dynamics, meteorology and materials science. In the early 1990s, work in the field of molecular-dynamics (MD) simulations proved parallel platforms to be highly successful in enabling large-scale MD simulation of up to 131 million particles [26]. Today the same code has been tested and used to simulate up to 320 billion atoms on the BlueGene/L architecture containing 131,072 IBM PowerPC440 processors [22]. These simulations include calculations based upon the short-range interaction between the individual atoms, thus in some ways approximate to agent simulations although in other ways molecular-dynamic simulations lack the complexity of most agent-based models (see Table 2).

There are significant decisions to be made when considering the application of a computing solution such as

Agent Based Modeling, Large Scale Simulations, Table 2
Key elements of a 'bottom-up' simulation that may affect the way in which it may scale. Agent simulations tend to be complex (to the right of the table), though may have some elements that are less complex, such as local or fixed interactions

Element	Least complex →	Most complex
Spatial structure	Aspatial or Lattice of cells (1d, 2d or 3d +)	Continuous space
Internal state	Simple representation (boolean true or false)	Complex representation (many states from an enumerable set) or fuzzy variable values
Agent heterogeneity	No	Yes
Interactions	Local and fixed (within a neighborhood)	Multiple different ranges and stochastic
Synchrony of model updates	Synchronous update	Not synchronous: asynchrony due to state-transition rules

parallel programming to solve the problem of large numbers of agents. In addition to the issue of reprogramming the model to run on a parallel computer architecture, it is also necessary to consider the additional complexity of agents (as opposed to atoms), so that programming models and tools facilitate the deployment, management and control of agents in the distributed simulation [10]. For example, distributed execution resources and timelines must be managed, full encapsulation of agents must be enforced, and tight control over message-based multi-agent interactions is necessary [10]. Agent models can vary in complexity, but most tend to be complex especially in the key model elements of spatial structure and agent heterogeneity. Table 2 gives an indication of the relative complexity of model elements found in models that focus on individual interactions (which encompasses both multi-agent models and less complex, 'Ising'-type models).

The following sections detail guidelines for the development of a large scale agent-based model, highlighting in particular the challenges faced in writing large scale, high performance Agent based Modelling (ABM) simulations and giving a suggested development protocol. Following this, an example is given of the parallelization of a simple agent-based model, showing some of the advantages but also the pitfalls of this most popular solution. Key challenges, including difficulties that may arise in the analysis of agent-based models at a large scale, are highlighted. Alternative solutions are then discussed and some conclusions are drawn on the way in which large scale agent-based simulation may develop in coming years.

Large Scale Agent Based Models: Guidelines for Development

Key Considerations

There is no such thing as a standard agent-based model, or even a coherent methodology for agent simulation development (although recent literature in a number of fields sets out some design protocols, e. g. Gilbert [11] and Grimm et al. [12]). Thus, there can be no standard method to develop a large scale agent-based model. However, there are certain things to consider when planning to scale up a model. Some key questions to ask about the model are as follows:

1. What program design do you already have and what is the limitation of this design?
   (a) What is it the memory footprint for any existing implementation?
   (b) What are your current run times?
2. What are your scaling requirements?
   (a) How much do you need to scale now?
   (b) How far do you need to scale eventually?
   (c) How soon do you need to do it?
3. How simple is your model and how is it structured?
4. What are your agent complexities?
5. What are your output requirements?

The first question is to identify the limitations in the program design that you are using and to focus on the primary 'bottlenecks' in the model. These limitations will either be due to memory or speed (or perhaps both). Therefore it will be necessary to identify the memory footprint for your existing model, and analyze run times, identifying where the most time is taken or memory used by the simulation. It is primarily processor power that controls the speed of the simulation. Runtime will also increase massively once Random Access Memory (RAM) is used up, as most operating systems will resort to virtual memory (i. e. hard drive space), thus a slower mechanism with mechanical parts rather than solid-state technology engages. At this stage, it may be such that simple adjustments to the code may improve the scalability of the model. However, if the code is efficient, other solutions will then need to be sought.

The second question is how much scaling is actually necessary for the model. It may be such that a simple or interim solution (e. g. upgrading computer hardware) may be acceptable whilst only moderate scaling is required,

but longer term requirements should also be considered – a hardware upgrade may be a quick fix but if the model may eventually be used for much larger simulations it is necessary to plan for the largest scaling that will potentially be required.

The third question, relating to model simplicity and structure, is key to deciding a methodology that can be used to scale a model up. A number of factors will affect whether a model will be easy to distribute in parallel, for example. These include whether the model iterates at each time step or is event driven, whether it is aspatial or spatial and the level/type of agent interaction (both with one another and with the environment). More detail on the implications of these factors is given in Sect. "Parallel Computing".

Agent complexity, in addition to model structure, may limit the options available for scaling up a model. For example, a possible scaling solution may be to group individual agents together as 'super-individuals' [35]. However, if agents are too complex it may not be possible to determine a simple grouping system (such as by age), as agent behavior may be influenced heavily by numerous other state variables.

Output requirements are also important to consider. These may already be limiting the model, in terms of memory for data storage. Even if they are not currently limiting the model in this way, once the model is scaled up output data storage needs may be an issue, for example, if the histories of individual agents need to be stored. In addition, the way that output data is handled by the model may be altered if the model structure is altered (e. g. if agents are grouped together output will be at an aggregate level). Thus, an important consideration is to ensure that output data is comparable to the original model and that it is feasible to output once the model structure is altered.

A Protocol

In relation to the key considerations highlighted above, a simple protocol for developing a large scale agent-based simulation can be defined as follows:

1. Optimize existing code.
2. Clearly identify scaling requirements (both for now and in the future).
3. Consider simple solutions first (e. g. a hardware upgrade).
4. Consider more challenging solutions.
5. Evaluate the suitability of the chosen scaling solution on a simplified version of the model before implementing on the full model.

The main scaling solution to implement (e. g. from Table 1) is defined by the requirements of the model. Implementation of more challenging solutions should be done in stages, where perhaps a simplified version of the model is implemented on a larger scale. Agent simulation development should originate with a local, flexible 'prototype' and then as the model development progresses and stabilizes larger scale implementations can be experimented with [10]. This is necessary for a parallel implementation of a model, for example, as a simplified model enables and assessment of whether it is likely to provide the desired improvements in model efficiency. This is particularly the case for improvements in model speed, as this depends on improved processing performance that is not easily calculated in advance.

Parallel Computing

Increasing the capacity of an individual computer in terms of memory and processing power has limited ability to perform large scale agent simulations, particularly due to the time the machine would take to run the model using a single processor. However, by using multiple processors and a mix of distributed and shared memory working simultaneously, the scale of the problem for each individual computer is much reduced. Subsequently, simulations can run in a fraction of the time that would be taken to perform the same complex, memory intensive, operations. This is the essence of parallel computing. 'Parallel computing' encompasses a wide range of computer architectures, from a HPC (High performance computing) Linux box, to dedicated multi-processor/multi-core systems (such as a Beowulf cluster), super clusters, local computer clusters or Grids and public computing facilities (e. g. Grid computers, such as the White Rose Grid, UK http://www.wrgrid.org.uk/). The common factor is that these systems consist of a number of interconnected 'nodes' (processing units), that may perform simultaneous calculations on different data. These calculations may be the same or different, depending whether a 'Single Instruction Multiple Data' (SIMD) or 'Multiple Instruction Multiple data' (MIMD) approach is implemented.

In terms of MAS, parallel computing has been used to develop large scale agent simulations in a number of disciplines. These range from ecology, e. g. [1,17,27,40,41,42,43] and biology, e. g. [5,7] to social science, e. g. [38] and computer science, e. g. [34], including artificial intelligence and robotics, e. g. [3,4].

Several key challenges arise when implementing an agent model in parallel, which may affect the increase in performance achieved. These include load balancing be-

tween nodes, synchronizing events to ensure causality, monitoring of the distributed simulation state, managing communication between nodes and dynamic resource allocation [39]. Good load balancing and inter-node communication with event synchronisation are central to the development of an efficient parallel simulation, and are further discussed below.

Load Balancing

In order to ensure the most efficient use of memory and processing resources in a parallel computing system the data load must be balanced between processors and the work load equally distributed. If this is not the case then one computer may be idle as others are working, resulting in time delays and inefficient use of the system's capacity. There are a number of ways in which data can be 'mapped' to different nodes and the most appropriate depends on the model structure. Further details and examples are given in Pacheco [30], including 'block mapping' and 'cyclic mapping'. An example of 'block mapping' load balancing is given below, in Sect. "Example".

In many simulations the computational demands on the nodes may alter over time, as the intensity of the agents' or environment's processing requirements varies on each node over time. In this case, dynamic load balancing techniques can be adopted to further improve the parallel model performance. For example, Jang [19] and Jang and Agha [20], use a form of dynamic load balancing with object migration they term "Adaptive Actor Architecture". Each agent platform monitors the workload of its computer node and the communication patterns of agents executing on it in order to redistribute agents according to their communication localities as agent platforms become overloaded. However, this approach does introduce additional processing overheads, so is only worth implementing for large scale agent simulations where some agents communicate with one another more intensely than other agents (communication locality is important) or communication patterns are continuously changing so static agent allocation is not efficient.

Communication Between Nodes

It is important to minimize inter-node communication when constructing a parallel agent simulation, as this may slow the simulation down significantly if the programmer is not careful [37,38]. The structure of the model itself largely determines the way in which data should be split and information transferred between nodes to maximize efficiency. Agent simulations generally by definition act spatially within an environment. Thus, an important first consideration is whether to split the agents or the environment between nodes. The decision as to whether to split the agents between processors or elements of the environment such as grid cells largely depends upon the complexity of the environment, the mobility of the agents, and the number of interactions between the agents. If the environment is relatively simple (thus information on the whole environment may be stored on all nodes), it is probably most efficient to distribute the agents. This is particularly the case if the agents are highly mobile, as a key problem when dividing the environment between processors is the transfer of agents or information between processors. However, if there are complex, spatially defined interactions between agents, splitting agents between nodes may be problematic, as agents may be interacting with other agents that are spatially local in the context of the whole simulation but are residing on different processors. Therefore conversely, if the agents are not very mobile but have complex, local interactions and/or the agents reside in a complex environment, it is probably best to split the environment between nodes [25]. Further efficiency may be achieved by clustering the agents which communicate heavily with each other [37].

In models where there is high mobility and high interaction it is often possible, especially for ecological models, to find a statistical commonality that can be used as a replacement for more detailed interaction. For example, as will be shown in our example, if the number of local agent interactions is the only important aspect of the interactions, a density map of the agents, transferred to a central node, aggregated and redistributed, might allow agents to be divided between nodes without the issue of having to do detailed inter-agent communication between nodes a large number of times.

The way in which the simulation iterates may influence the approach taken when parallelizing the model. The model may update synchronously at a given time step or asynchronously (usually because the system is event-driven). In addition, agents may update asynchronously but the nodes may be synchronized at each time step or key model stage. Asynchronous updating may be a problem if there is communication between nodes, as some nodes may have to wait for others to finish processes before communication takes place and further processing is possible, resulting in blocking (see below). Communication between nodes then becomes highly complex [44]. It is important that messages communicating between agents are received in the correct order, however a common problem in distributed simulations is ensuring that this is so as other factors, such as latency in message transmission across the network, may affect communication [44].

A number of time management mechanisms exist that may be implemented to manage message passing in order to ensure effective node to node communication, e. g. [8].

Blocking and Deadlocking

Deadlock occurs when two or more processes are waiting for communication from one of the other processes. When programming a parallel simulation it is important to avoid deadlock to ensure the simulation completes. The simplest example is when two processors are programmed to receive from the other processor before that processor has sent. This may be simply resolved by changing the order that tasks are executed, or to use 'non-blocking' message passing. Where blocking is used, processing on nodes waits until a message is transmitted. However, when 'non-blocking' is used, processing continues even if the message hasn't been transmitted yet. The use of a non-blocking MPI may reduce computing times, and work can be performed while communication is in progress.

Example

To demonstrate some of the benefits and pitfalls of parallel programming for a large scale agent-based model, a simple example is given here. This summarizes a simplified agent-based model of aphid population dynamics in agricultural landscapes of the UK, which was parallelized to cope with millions of agents, as described in detail in Parry [31], Parry and Evans [32] and Parry et al. [33].

A key problem with the original, non-parallel, aphid simulation was that it was hindered by memory requirements, which were far larger than could be accommodated at any individual processing element. This is a common computing problem [6]. The data storage required for each aphid object in a landscape scale simulation quickly exceeded the storage capacity of a PC with up to 2097 MB of RAM. The combined or 'virtual shared' memory of several computers was used to cope with the amount of data needed, using a Single Instruction Multiple-Data approach (SIMD).

Message-passing techniques were used to transfer information between processors, to distribute the agents in the simulation across a Beowulf cluster (a 30-node distributed memory parallel computer). A Message-passing Interface (MPI) for Java was used, MPIJava (http://www.hpjava.org). 'MPIJava wraps around the open-source' open-source native MPI 'LAM' (http://www.lam-mpi.org/). Further details on the methods used to incorporate the MPI into the model are given in Parry [31] and Parry et al. [33].

Effective parallelization minimizes the passing of information between nodes, as it is processor intensive. In the example model, only the environment object and information on the number of agents to create on each node are passed from a single control node to each of the other nodes in the cluster, and only density information is returned to the control node for redistribution and display. The control node manages the progress of the model, acts as a central communication point for the model and handles any code that may not be distributed to all nodes (such as libraries from an agent toolkit or a GUI). Structuring a model without a control node is possible, or the control node may also be used to process data, depending on the requirements of the simulation. Transfer of density values, rather than agents, significantly reduced the computational overheads for message passing between the nodes. The model was simple enough that specific inter-agent communication between nodes was not necessary.

Even distribution of data between nodes was achieved by splitting immigrant agents evenly across the system, with each node containing information on the environment and local densities passed from the control node. The number of immigrants to be added to each node was calculated by a form of 'block mapping', pp. 35 in [30], which partitioned the number of immigrants into blocks which were then assigned to each node. So, if there were three nodes ($n = 3$) and thirteen immigrants ($i = 13$), the immigrants mapped to each node would be as follows:

$$\begin{aligned} i_0, i_1, i_2, i_3 &\rightarrow n_1 \\ i_4, i_5, i_6, i_7 &\rightarrow n_2 \\ i_8, i_9, i_{10}, i_{11}, i_{12} &\rightarrow n_3 \ . \end{aligned}$$

As thirteen does not divide evenly by three, the thirteenth agent is added to the final node.

Benefits

Simulation runtime and memory availability was greatly improved by implementing the simple aphid model in parallel across a large number of nodes. The greatest improvement in simulation runtime and memory availability was seen when the simulation was run across the maximum number of nodes (25) (Figs. 3 and 4). The largest improvement in speed given by the parallel model in comparison to the non-parallel model is when more than 500,000 agents are run across twenty-five nodes, although the parallel model is slower by comparison for lower numbers. This means that additional processing power is required in the parallel simulation compared to the original model, such that only when very large numbers of agents are run does it become more efficient.

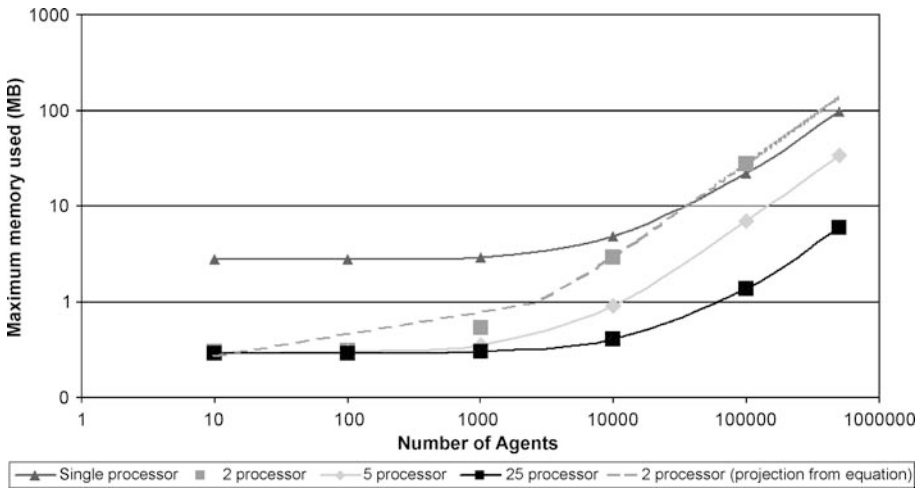

Agent Based Modeling, Large Scale Simulations, Figure 3
Plot of the mean maximum memory used (per node) against number of agents for the model: comparison between simulations using 2, 5 and 25 nodes and the non-parallel model (single processor)

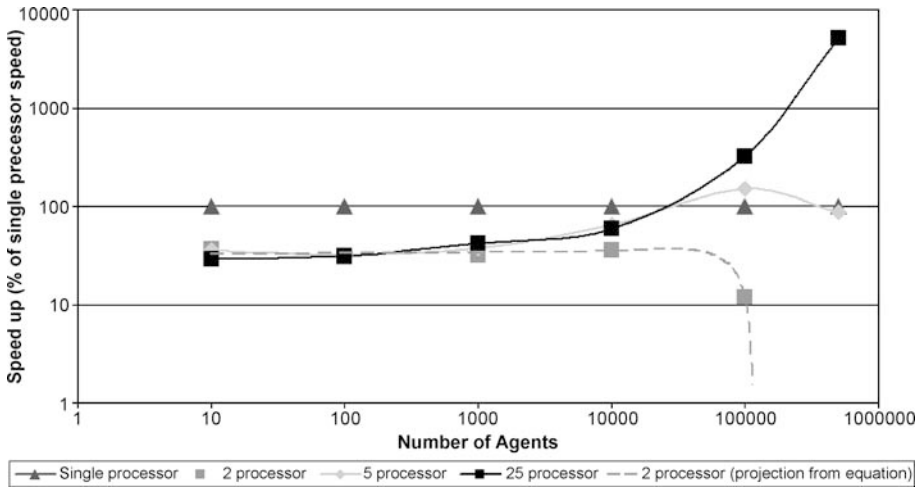

Agent Based Modeling, Large Scale Simulations, Figure 4
Plot of the percentage speed up (per node) from the non-parallel model against number of agents for the model: comparison between simulations using 2, 5 and 25 nodes and the non-parallel model (single processor)

Pitfalls

Although there are clear benefits of distributing the example simulation across a large number of nodes, the results highlight that the parallel approach is not always more efficient than the original single processor implementation of a model. In the example, the two node simulation used more memory on the worker node than the non-parallel model when the simulation had 100,000 agents or above. This is due to additional memory requirements introduced by message passing and extra calculations required in the parallel implementation (which are less significant when more nodes are used as these requirements remain relatively constant).

The results also highlight that adding more processors does not necessarily increase the model speed. The example model shows that for simulations run on two nodes (one control node, one worker node) the simulation takes longer to run in parallel compared to the non-parallel model. Message passing time delay and the modified structure of the code are responsible. The greatest improvement in speed is when more than 500,000 agents are run across twenty-five nodes, however when lower numbers of nodes are used the relationship between the num-

ber of nodes and speed is complex: for 100,000 agents five nodes are faster than the non-parallel model, but for 500,000 the non-parallel model is faster. Overall, these results suggest that when memory is sufficient on a single processor, it is unlikely to ever be efficient to parallelize the code, as when the number of individuals was low the parallel simulation took longer and was less efficient than the non-parallel model run on a single node.

This demonstrates that in order to effectively parallelize an agent model, the balance between the advantage of increasing the memory availability and the cost of communication between nodes must be assessed. By following an iterative development process as suggested in Sect. "A Protocol", the threshold below which parallelization is not efficient and whether this option is suitable for the model should become apparent in the early stages of model development. Here, the simplified study confirmed the value of further development to build a parallel version of the full model.

Future Directions

Complexity and Model Analysis

In addition to the processing and data handling issues faced by large scale agent-based simulations, as agent simulations increase in scale and complexity, model analysis may become more difficult and the system may become intractable. There is no clearly defined way of dealing with increased difficulties of model analysis introduced by the greater complexity of large scale agent-based models, in the same way that there is no clear way to deal with the complexity inherent in most agent-based systems. However, some guidelines to model analysis have recently been developed for agent simulations, e. g. [12], and some suggestions are put forward on ways in which agent simulation complexity may be described, for example by the use of a model 'overview', 'design concepts' and 'details' (ODD) protocol for agent model description. In particular, the protocol requires that design concepts are linked to general concepts identified in the field of Complex Adaptive Systems [13], including emergence and stochasticity.

To address issues of complexity in a large scale agent simulation, one possibility is to include an additional layer of structuring, 'organizational design', into the system. This is where the peer group of an agent, it's roles and responsibilities are assigned in the model and made explicit, pp. 121 in [16]. In general agent simulations already have an implicit organization structure; Horling and Lesser [16] argue that explicit organizational design highlights hidden inefficiencies in the model and allows the model to take full advantages of the resources available.

Grid Computing

For large scale agent-based models, many researchers have reached the limitations of ordinary PCs. However, there are 'critical tensions' [10] between agent simulations built on ordinary PCs and heterogeneous, distributed parallel programming approaches. The architecture of a distributed system is very different to that of an ordinary PC, thus to transfer a simulation to a computer cluster additional system properties must be taken into account, including management of the distribution of the simulation and concurrency [10]. This is particularly apparent when parallelization is attempted on heterogeneous, non-dedicated systems such as a public Grid. The Grid may offer a number of advantages for large scale agent-based simulation, such as collaboration between modellers, access to resources and geographically distributed datasets [47]. However, in such systems, issues of infrastructure reliability, functional completeness and the state of documentation for some kinds of environments exist [10]. In order to use such a system the 'fundamental' issue of partial failures must be addressed (e. g. with a dynamic agent replication strategy [14]).

Dissemination of Techniques

For parallel computing as a solution to large scale agent-based simulation, there is an interesting and useful future challenge to develop user friendly, high performance, versatile hardware architectures and software systems. Many developers of agent simulations are not computer scientists by training, and still rely upon numerous agent toolkits for simulation development (e. g. Swarm and Repast). Automatic distribution of agents to whatever resources are available would be a great tool for many agent software developers. Therefore, perhaps the greatest challenge would be to develop a system that would allow for paralellisation to be performed in an agent simulation automatically, where agents may be written in a high-level language and could be automatically partitioned to nodes in a network. One example of an attempt to achieve this is Graphcode (http://parallel.hpc.unsw.edu.au/rks/graphcode/). Based upon MPI, it maps agent-based models onto parallel computers, where agents are written based upon their graph topography to minimize communication overhead. Another example is HLA_GRID_Repast [47], 'a system for executing large scale distributed simulations of agent-based systems over the Grid', for users of the popular Repast agent toolkit. HLA_GRID_Repast is a middleware layer which enables the execution of a federation of multiple interacting instances of Repast models across a grid network with a High

Level Architecture (HLA). This is a 'centralized coordination approach' to distributing an agent simulation across a network [39]. Examples of algorithms designed to enable dynamic distribution of agent simulations are given in Scheutz and Schermerhorn [36].

Although parallel computing is often the most effective way of handling large scale agent-based simulations, there are still some significant obstacles to the use of parallel computing for MAS. As shown with the simple example given here, this may not always be the most effective solution depending upon the increase in scale needed and the model complexity. Other possible methods were suggested in Sect. "Introduction", but these may also be unsuitable. Another option could be to deconstruct the model and simplify only certain elements of the model using either parallel computing or one of the other solutions suggested in Sect. "Introduction". Such a 'hybrid' approach is demonstrated by Zhang and Lui [46], who combine equation-based approaches and multi-agent simulation with a Cellular Automata to simulate the complex interactions in the process of human immune response to HIV. The result is a model where equations are used to represent within-site processes of HIV infection, and agent-based simulation is used to represent the diffusion of the virus between sites.

It is therefore important to consider primarily the various ways in which the model may be altered, hybridized or simplified yet still address the core research questions, before investing money in hardware or investing time in the development of complex computational solutions. Making the transition from a serial application to a parallel version is a process that requires a fair degree of formalism and program restructuring, so is not to be entered into lightly without exploring the other options and the needs of the simulation first.

Overall, it is clear that disparate work is being done in a number of disciplines to facilitate large scale agent-based simulation, and knowledge is developing rapidly. Some of this work is innovative and highly advanced, yet inaccessible to researchers in other disciplines who may unaware of key developments outside of their field. This chapter synthesizes and evaluates large scale agent simulation to date, providing a reference for a wide range of agent simulation developers.

Acknowledgments

Many thanks to Andrew Evans (Multi-Agent Systems and Simulation Research Group, University of Leeds, UK) and Phil Northing (Central Science Laboratory, UK) for their advice on this chapter.

Bibliography

Primary Literature

1. Abbott CA, Berry MW, Comiskey EJ, Gross LJ, Luh H-K (1997) Parallel Individual-Based Modeling of Everglades Deer Ecology. IEEE Comput Sci Eng 4(4):60–78
2. Anderson J (2000) A Generic Distributed Simulation System for Intelligent Agent Design and Evaluation. In: Proceedings of the AI, Simulation & Planning In High Autonomy Systems, Arizona
3. Bokma A, Slade A, Kerridge S, Johnson K (1994) Engineering large-scale agent- based systems with consensus. Robot Comput-Integr Manuf 11(2):81–91
4. Bouzid M, Chevrier V, Vialle S, Charpillet F (2001) Parallel simulation of a stochastic agent/environment interaction model. Integr Comput-Aided Eng 8(3):189–203
5. Castiglione F, Bernaschi M, Succi S (1997) Simulating the Immune Response on a Distributed Parallel Computer. Int J Mod Phys C 8(3):527–545
6. Chalmers A, Tidmus J (1996) Practical Parallel Processing: An Introduction to Problem Solving in Parallel. International Thomson Computer Press, London
7. Da-Jun T, Tang F, Lee TA, Sarda D, Krishnan A, Goryachev A (2004) Parallel computing platform for the agent-based modeling of multicellular biological systems. In: Parallel and Distributed Computing: Applications and Technologies. Lecture Notes in Computer Science, vol 3320, pp 5–8
8. Fujimoto RM (1998) Time management in the high level architecture. Simul 71:388–400
9. Gasser L, Kakugawa K (2002) MACE3J: Fast flexible distributed simulation of large, large-grain multi-agent systems. In: Proceedings of AAMAS
10. Gasser L, Kakugawa K, Chee B, Esteva M (2005) Smooth scaling ahead: progressive MAS simulation from single PCs to Grids. In: Davidsson P, Logan B, Takadama K (eds) Multi-agent and multi-agent-based simulation. Joint Workshop MABS 2004 New York, NY, USA, July 19, 2004. Springer, Berlin
11. Gilbert N (2007) Agent-based Models (Quantitative applications in the social siences). SAGE, London
12. Grimm V, Berger U, Bastiansen F, Eliassen S, Ginot V, Giske J, Goss-Custard J, Grand T, Heinz S, Huse G, Huth A, Jepsen JU, Jorgensen C, Mooij WM, Muller B, Pe'er G, Piou C, Railsback SF, Robbins AM, Robbins MM, Rossmanith E, Ruger N, Strand E, Souissi S, Stillman RA, Vabo R, Visser U, DeAngelis DL (2006) A standard protocol for describing individual-based and agent-based models. Ecol Model 198(1–2):115–126
13. Grimm V, Railsback SF (2005) Individual-based Modeling and Ecology. Princeton Series in Theoretical and Computational Biology. Princeton University Press, Princeton, 480 pp
14. Guessoum Z, Briot J-P, Faci N (2005) Towards fault-tolerant massively multiagent system. In: Ishida T, Gasser L, Nakashima H (eds) Massively Multi-Agent Systems I: First International Workshop MMAS 2004, Kyoto, Japan, December 2004. Springer, Berlin
15. Heppenstall AJ (2004) Application of Hybrid Intelligent Agents to Modelling a Dynamic, Locally Interacting Retail Market. Ph D thesis, University of Leeds, UK
16. Horling B, Lesser V (2005) Quantitative organizational models for large-scale agent systems. In: Ishida T, Gasser L, Nakashima H (eds) Massively Multi-Agent Systems I: First Inter-

national Workshop MMAS 2004, Kyoto, Japan, December 2004. Springer, Berlin
17. Immanuel A, Berry MW, Gross LJ, Palmer M, Wang D (2005) A parallel implementation of ALFISH: simulating hydrological compartmentalization effects on fish dynamics in the Florida Everglades. Simul Model Pract Theory 13:55–76
18. Ishida T, Gasser L, Nakashima H (eds) (2005) Massively Multi-Agent Systems I. First International Workshop, MMAS 2004, Kyoto, Japan. Springer, Berlin
19. Jang MW (2006) Agent framework services to reduce agent communication overhead in large-scale agent-based simulations. Simul Model Pract Theory 14(6):679–694
20. Jang MW, Agha G (2005) Adaptive agent allocation for massively multi-agent applications. In: Ishida T, Gasser L, Nakashima H (eds) Massively Multi-Agent Systems I: First International Workshop MMAS 2004, Kyoto, Japan, December 2004. Springer, Berlin
21. Jennings NR (2000) On agent-based software engineering. Artif Intell 117:277–296
22. Kadau K, Germann TC, Lomdahl PS (2006) Molecular dynamics comes of age: 320 billion atom simulation on BlueGene/L. Int J Mod Phys C 17(12):1755
23. Lees M, Logan B, Oguara T, Theodoropoulos G (2003) Simulating Agent-Based Systems with HLA: The case of SIM_AGENT – Part II. In: Proceedings of the 2003 European Simulation Interoperability Workshop
24. Lees M, Logan B, Theodoropoulos G (2002) Simulating Agent-Based Systems with HLA: The case of SIM_AGENT. In: Proceedings of the 2002 European Simulation Interoperability Workshop, pp 285–293
25. Logan B, Theodoropolous G (2001) The Distributed Simulation of Multi-Agent Systems. Proc IEEE 89(2):174–185
26. Lomdahl PS, Beazley DM, Tamayo P, Gronbechjensen N (1993) Multimillion particle molecular-dynamics on the CM-5. Int J Mod Phys C: Phys Comput 4(6):1075–1084
27. Lorek H, Sonnenschein M (1995) Using parallel computers to simulate individual-oriented models in ecology: a case study. In: Proceedings: ESM '95 European Simulation Multiconference, Prague, June 1995
28. Luke S, Cioffi-Revilla C, Panait L, Sullivan K (2004) MASON: A New Multi-Agent Simulation Toolkit. In: Proceedings of the 2004 SwarmFest Workshop
29. Openshaw S, Turton I (2000) High performance computing and the art of parallel programming: an introduction for geographers, social scientists, engineers. Routledge, London
30. Pacheco PS (1997) Parallel Programming with MPI. Morgan Kauffman Publishers, San Francisco
31. Parry HR (2006) Effects of Land Management upon Species Population Dynamics: A Spatially Explicit, Individual-based Model. Ph D thesis, University of Leeds
32. Parry HR, Evans AJ (in press) A comparative analysis of parallel processing and super-individual methods for improving the computational performance of a large individual-based model. Ecological Modelling
33. Parry HR, Evans AJ, Heppenstall AJ (2006) Millions of Agents: Parallel Simulations with the Repast Agent-Based Toolkit. In: Trappl R (ed) Cybernetics and Systems 2006, Proceedings of the 18th European Meeting on Cybernetics and Systems Research
34. Popov K, Vlassov V, Rafea M, Holmgren F, Brand P, Haridi S (2003) Parallel agent-based simulation on a cluster of workstations. In: EURO-PAR 2003 Parallel Processing, vol 2790, pp 470–480
35. Scheffer M, Baveco JM, DeAngelis DL, Rose KA, van Nes EH (1995) Super-Individuals: a simple solution for modelling large populations on an individual basis. Ecol Model 80:161–170
36. Scheutz M, Schermerhorn P (2006) Adaptive Algorithms for the Dynamic Distribution and Parallel Execution of Agent-Based Models. J Parallel Distrib Comput 66(8):1037–1051
37. Takahashi T, Mizuta H (2006) Efficient agent-based simulation framework for multi-node supercomputers. In: Perrone LF, Wieland FP, Liu J, Lawson BG, Nicol DM, Fujimoto RM (eds) Proceedings of the 2006 Winter Simulation Conference
38. Takeuchi I (2005) A massively multi-agent simulation system for disaster mitigation. In: Ishida T, Gasser L, Nakashima H (eds) Massively Multi-Agent Systems I: First International Workshop MMAS 2004, Kyoto, Japan, December 2004. Springer, Berlin
39. Timm IJ, Pawlaszczyk D (2005) Large Scale Multiagent Simulation on the Grid. In: Veit D, Schnizler B, Eymann T (eds) Proceedings of the Workshop on Agent-based grid Economics (AGE 2005) at the IEEE International Symposium on Cluster Computing and the Grid (CCGRID). Cardiff University, Cardiff
40. Wang D, Berry MW, Carr EA, Gross LJ (2006a) A parallel fish landscape model for ecosystem modeling. Simul 82(7):451–465
41. Wang D, Berry MW, Gross LJ (2006b) On parallelization of a spatially-explicit structured ecological model for integrated ecosystem simulation. Int J High Perform Comput Appl 20(4):571–581
42. Wang D, Carr E, Gross LJ, Berry MW (2005a) Toward ecosystem modeling on computing grids. Comput Sci Eng 7:44–52
43. Wang D, Gross L, Carr E, Berry M (2004) Design and implementation of a Parallel Fish Model for South Florida. In: Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04)
44. Wang F, Turner SJ, Wang L (2005b) Agent Communication in Distributed Simulations. In: Davidsson P, Logan B, Takadama K (eds) Multi-agent and multi-agent-based simulation. Joint Workshop MABS 2004 New York, NY, USA, July 19, 2004. Springer, Berlin
45. Wooldridge M (1999) Intelligent agents. In: Weiss G (ed) Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence. MIT Press, Cambridge, pp 27–78
46. Zhang S, Lui J (2005) A massively multi-agent system for discovering HIV-immune interaction dynamics. In: Ishida T, Gasser L, Nakashima H (eds) Massively Multi-Agent Systems I: First International Workshop MMAS 2004, Kyoto, Japan, December 2004. Springer, Berlin
47. Zhang Y, Thedoropoulos G, Minson R, Turner SJ, Cai W, Xie Y, Logan B (2005) Grid-aware Large Scale Distributed Simulation of Agent-based Systems. In: 2004 European Simulation Interoperability Workshop (EuroSIW 2005), 05E-SIW-047, Toulouse, France

Books and Reviews

Agent libraries and toolkits with distributed or parallel features:

- Distributed GenSim: Supports distributed parallel execution [2].
- Ecolab: http://ecolab.sourceforge.net/. EcoLab models may also use the Graphcode library to implement a distributed network of agents over an MPI-based computer cluster.

- Graphcode system http://parallel.hpc.unsw.edu.au/rks/graphcode/.
- MACE3J: http://www.isrl.uiuc.edu/amag/mace/ an experimental agent platform supporting deployment of agent simulations across a variety of system architectures [9,10].
- MASON http://cs.gmu.edu/~eclab/projects/mason/. MASON was 'not intended to include parallelization of a single simulation across multiple networked processors' [28]. However, it does provide two kinds of simple parallelization:
  1. Any given step in the simulation can be broken into parallel sub-steps each performed simultaneously.
  2. A simulation step can run asynchronously in the background independent of the simulation.
- Repast http://repast.sourceforge.net and HLA_GRID_Repast [47]. The Repast toolkit has in-built capabilities for performing batch simulation runs.
- SimAgent: http://www.cs.bham.ac.uk/research/projects/poplog/packages/simagent.html. Two developments support distributed versions of SimAgent:
  1. The use of HLA to distribute SimAgent [23,24]
  2. The SWAGES package: http://www.nd.edu/~airolab/software/index.html. This allows SimAgent to be distributed over different computers and interfaced with other packages.

Message Passing Interfaces (MPI):
- Background and tutorials http://www-unix.mcs.anl.gov/mpi/
- MPICH2 http://www-unix.mcs.anl.gov/mpi/mpich/
- MPI forum http://www.mpi-forum.org/
- MPIJava http://www.hpjava.org
- OpenMP http://www.openmp.org
- OpenMPI http://www.open-mpi.org/

Parallel computing and distributed agent simulation websites:
- Further references and websites http://www.cs.rit.edu/~ncs/parallel.html.
- Introduction to parallel programming http://www.mhpcc.edu/training/workshop/parallel_intro/MAIN.html
- http://www.agents.cs.nott.ac.uk/research/simulation/simulators/ (implementations of HLA_GRID_Repast and distributed SimAgent).
- Globus Grid computing resources http://www.globus.org/.
- Beowulf computer clusters http://www.beowulf.org/

Agent Based Modeling, Mathematical Formalism for

Reinhard Laubenbacher[1], Abdul S. Jarrah[1], Henning S. Mortveit[1], S.S. Ravi[2]
[1] Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Virginia, USA
[2] Department of Computer Science, University at Albany – State University of New York, New York, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Existing Mathematical Frameworks
Finite Dynamical Systems
Finite Dynamical Systems as Theoretical and Computational Tools
Mathematical Results on Finite Dynamical Systems
Future Directions
Bibliography

Glossary

Agent-based simulation An agent-based simulation of a complex system is a computer model that consists of a collection of agents/variables that can take on a typically finite collection of states. The state of an agent at a given point in time is determined through a collection of rules that describe the agent's interaction with other agents. These rules may be deterministic or stochastic. The agent's state depends on the agent's previous state and the state of a collection of other agents with whom it interacts.

Mathematical framework A mathematical framework for agent-based simulation consists of a collection of mathematical objects that are considered mathematical abstractions of agent-based simulations. This collection of objects should be general enough to capture the key features of most simulations, yet specific enough to allow the development of a mathematical theory with meaningful results and algorithms.

Finite dynamical system A finite dynamical system is a time-discrete dynamical system on a finite state set. That is, it is a mapping from a Cartesian product of finitely many copies of a finite set to itself. This finite set is often considered to be a field. The dynamics is generated by iteration of the mapping.

Definition of the Subject

Agent-based simulations are generative or computational approaches used for analyzing "complex systems." What is a "system?" Examples of systems include a collection of molecules in a container, the population in an urban area, and the brokers in a stock market. The entities or *agents* in these three systems would be molecules, individuals and stock brokers, respectively. The agents in such systems interact in the sense that molecules collide, individuals come into contact with other individuals and bro-

kers trade shares. Such systems, often called *multiagent systems*, are not necessarily complex. The label "complex" is typically attached to a system if the number of agents is large, if the agent interactions are involved, or if there is a large degree of heterogeneity in agent character or their interactions.

This is of course not an attempt to define a complex system. Currently there is no generally agreed upon definition of complex systems. It is not the goal of this article to provide such a definition – for our purposes it will be sufficient to think of a complex system as a collection of agents interacting in some manner that involves one or more of the complexity components just mentioned, that is, with a large number of agents, heterogeneity in agent character and interactions and possibly stochastic aspects to all these parts. The global properties of complex systems, such as their global dynamics, emerge from the totality of local interactions between individual agents over time. While these local interactions are well understood in many cases, little is known about the emerging global behavior arising through interaction. Thus, it is typically difficult to construct global mathematical models such as systems of ordinary or partial differential equations, whose properties one could then analyze. Agent-based simulations are one way to create computational models of complex systems that take their place.

An *agent-based simulation*, sometimes also called an *individual-based* or *interaction-based simulation* (which we prefer), of a complex system is in essence a computer program that realizes some (possibly approximate) model of the system to be studied, incorporating the agents and their rules of interaction. The simulation might be deterministic (i. e., the evolution of agent-states is governed by deterministic rules) or stochastic. The typical way in which such simulations are used is to initialize the computer program with a particular assignment of agent states and to run it for some time. The output is a temporal sequence of states for all agents, which is then used to draw conclusions about the complex system one is trying to understand. In other words, the computer program is the model of the complex system, and by running the program repeatedly, one expects to obtain an understanding of the characteristics of the complex system.

There are two main drawbacks to this approach. First, it is difficult to validate the model. Simulations for most systems involve quite complex software constructs that pose challenges to code validation. Second, there are essentially no rigorous tools available for an analysis of model properties and dynamics. There is also no widely applicable formalism for the comparison of models. For instance, if one agent-based simulation is a simplification of another, then one would like to be able to relate their dynamics in a rigorous fashion. We are currently lacking a mathematically rich formal framework that models agent-based simulations. This framework should have at its core a class of mathematical objects to which one can map agent-based simulations. The objects should have a sufficiently general mathematical structure to capture key features of agent-based simulations and, at the same time, should be rich enough to allow the derivation of substantial mathematical results. This chapter presents one such framework, namely the class of *time-discrete dynamical systems over finite state sets*.

The building blocks of these systems consist of a collection of variables (mapping to agents), a graph that captures the dependence relations of agents on other agents, a local update function for each agent that encapsulates the rules by which the state of each agent evolves over time, and an update discipline for the variables (e. g. parallel or sequential). We will show that this class of mathematical objects is appropriate for the representation of agent-based simulations and, therefore, complex systems, and is rich enough to pose and answer relevant mathematical questions. This class is sufficiently rich to be of mathematical interest in its own right and much work remains to be done in studying it. We also remark that many other frameworks such as probabilistic Boolean networks [80] fit inside the framework described here.

Introduction

Computer simulations have become an integral part of today's research and analysis methodologies. The ever-increasing demands arising from the complexity and sheer size of the phenomena studied constantly push computational boundaries, challenge existing computational methodologies, and motivate the development of new theories to improve our understanding of the potential and limitations of computer simulation. Interaction-based simulations are being used to simulate a variety of biological systems such as ecosystems and the immune system, social systems such as urban populations and markets, and infrastructure systems such as communication networks and power grids.

To model or describe a given system, one typically has several choices in the construction and design of agent-based models and representations. When agents are chosen to be simple, the simulation may not capture the behavior of the real system. On the other hand, the use of highly sophisticated agents can quickly lead to complex behavior and dynamics. Also, use of sophisticated agents may lead to a system that *scales poorly*. That is, a linear

increase in the number of agents in the system may require a non-linear (e. g., quadratic, cubic, or exponential) increase in the computational resources needed for the simulation.

Two common methods, namely *discrete event simulation* and *time-stepped simulation*, are often used to implement agent-based models [1,45,67]. In the discrete event simulation method, each event that occurs in the system is assigned a time of occurrence. The collection of events is kept in increasing order of their occurrence times. (Note that an event occurring at a certain time may give rise to events which occur later.) When all the events that occur at a particular time instant have been carried out, the simulation clock is advanced to the next time instant in the order. Thus, the differences between successive values of the simulation clock may not be uniform. Discrete event simulation is typically used in contexts such as queuing systems [58]. In the time-stepped method of simulation, the simulation clock is always advanced by the same amount. For each value of the simulation clock, the states of the system components are computed using equations that model the system. This method of simulation is commonly used for studying, e. g., fluid flows or chemical reactions. The choice of model (discrete event versus time-stepped) is typically guided by an analysis of the computational speed they can offer, which in turn depends on the nature of the system being modeled, see, e. g., [37].

Tool-kits for general purpose agent-based simulations include Swarm [29,57] and Repast [68]. Such tool-kits allow one to specify more complex agents and interactions than would be possible using, e. g., ordinary differential equations models. In general, it is difficult to develop a software package that is capable of supporting the simulation of a wide variety of physical, biological, and social systems.

Standard or classical approaches to modeling are often based on continuous techniques and frameworks such as ordinary differential equations (ODEs) and partial differential equations (PDEs). For example, there are PDE based models for studying traffic flow [38,47,85]. These can accurately model the emergence of traffic jams for simple road/intersection configurations through, for example, the formation of shocks. However, these models fail to scale to the size and the specifications required to accurately represent large urban areas. Even if they hypothetically were to scale to the required size, the answers they provide (e. g. car density on a road as a function of position and time) cannot answer questions pertaining to specific travelers or cars. Questions of this type can be naturally described and answered through agent-based models. An example of such a system is TRANSIMS (see Subsect. "TRANSIMS (Transportation, Analysis, Simulation System)"), where an agent-based simulation scheme is implemented through a cellular automaton model. Another well-known example of the change in modeling paradigms from continuous to discrete is given by lattice gas automata [32] in the context of fluid dynamics. Stochastic elements are inherent in many systems, and this usually is reflected in the resulting models used to describe them. A stochastic framework is a natural approach in the modeling of, for example, noise over a channel in a simulation of telecommunication networks [6]. In an economic market or a game theoretic setting with competing players, a player may sometimes decide to provide incorrect information. The state of such a player may therefore be viewed and modeled by a random variable. A player may make certain probabilistic assumptions about other players' environment. In biological systems, certain features and properties may only be known up to the level of probability distributions. It is only natural to incorporate this stochasticity into models of such systems.

Since applications of stochastic discrete models are common, it is desirable to obtain a better understanding of these simulations both from an application point of view (reliability, validation) and from a mathematical point of view. However, an important disadvantage of agent-based models is that there are few mathematical tools available at this time for the analysis of their dynamics.

Examples of Agent-Based Simulations

In order to provide the reader with some concrete examples that can also be used later on to illustrate theoretical concepts we describe here three examples of agent-based descriptions of complex systems, ranging from traffic networks to the immune system and voting schemes.

TRANSIMS (Transportation, Analysis, Simulation System) TRANSIMS is a large-scale computer simulation of traffic on a road network [64,66,76]. The simulation works at the resolution level of individual travelers, and has been used to study large US metropolitan areas such as Portland, OR, Washington D.C. and Dallas/Fort Worth. A TRANSIMS-based analysis of an urban area requires (i) a population, (ii) a location-based activity plan for each individual for the duration of the analysis period and (iii) a network representation of all transportation pathways of the given area. The data required for (i) and (ii) are generated based on, e. g., extensive surveys and other information sources. The network representation is typically very close to a complete description of the real transportation network of the given urban area.

TRANSIMS consists of two main modules: the *router* and the cellular automaton based *micro-simulator*. The router maps each activity plan for each individual (obtained typically from activity surveys) into a travel route. The micro-simulator executes the travel routes and sends each individual through the transportation network so that its activity plan is carried out. This is done in such a way that all constraints imposed on individuals from traffic driving rules, road signals, fellow travelers, and public transportation schedules are respected. The time scale is typically 1 s.

The micro-simulator is the part of TRANSIMS responsible for the detailed traffic dynamics. Its implementation is based on *cellular automata* which are described in more detail in Subsect. "Cellular Automata". Here, for simplicity, we focus on the situation where each individual travels by car. The *road network representation* is in terms of *links* (e. g. road segments) and *nodes* (e. g. intersections). The network description is turned into a cell-network description by discretizing each lane of every link into cells. A cell corresponds to a 7.5 m lane segment, and can have up to four neighbor cells (front, back, left and right).

The *vehicle dynamics* is specified as follows. Vehicles travel with discrete velocities 0, 1, 2, 3, 4 or 5 which are constant between time steps. Each update time step brings the simulation one time unit forward. If the time unit is one second then the maximum speed of $v_{\max} = 5$ cells per time unit corresponds to an actual speed of $5 \times 7.5\,\mathrm{m/s} = 37.5\,\mathrm{m/s}$ which is 135 km/h or approximately 83.9 mph.

Ignoring intersection dynamics, the micro-simulator executes three functions for each vehicle in every update: (a) lane-changing, (b) acceleration and (c) movement. These functions can be implemented through four cellular automata, one each for lane change decision and execution, one for acceleration and one for movement. For instance, the acceleration automaton works as follows. A vehicle in TRANSIMS can increase its speed by at most 1 cell per second, but if the road ahead is blocked, the vehicle can come to a complete stop in the same time. The function that is applied to each cell that has a car in it uses the gap ahead and the maximal speed to determine if the car will increase or decrease its velocity. Additionally, a car may have its velocity decreased one unit as determined by a certain deceleration probability. The random deceleration is an important element of producing realistic traffic flow. A major advantage of this representation is that it leads to very light-weight agents, a feature that is critical for achieving efficient scaling.

CImmSim Next we discuss an interaction-based simulation that models certain aspects of the human immune system. Comprised of a large number of interacting cells whose motion is constrained by the body's anatomy, the immune system lends itself very well to agent-based simulation. In particular, these models can take into account three-dimensional anatomical variations as well as small-scale variability in cell distributions. For instance, while the number of T-cells in the human body is astronomical, the number of antigen-specific T-cells, for a specific antigen, can be quite small, thereby creating many spatial inhomogeneities. Also, little is known about the global structure of the system to be modeled.

The first discrete model to incorporate a useful level of complexity was *ImmSim* [22,23], developed by Seiden and Celada as a stochastic cellular automaton. The system includes B-cells, T-cells, antigen presenting cells (APCs), antibodies, antigens, and antibody-antigen complexes. Receptors on cells are represented by bit strings, and antibodies use bit strings to represent their epitopes and peptides. Specificity and affinity are defined by using bit string similarity. The bit string approach was initially introduced in [31]. The model is implemented on a regular two-dimensional grid, which can be thought of as a slice of a lymph node, for instance. It has been used to study various phenomena, including the optimal number of human leukocyte antigens in human beings [22], the autoimmunity and T-lymphocyte selection in the thymus [60], antibody selection and hyper-mutation [24], and the dependence of the selection and maturation of the immune response on the antigen-to-receptor's affinity [15]. The computational limitations of the Seiden-Celada model have been overcome by a modified model, *CimmSim* [20], implemented on a parallel architecture. Its complexity is several orders of magnitude larger than that of its predecessor. It has been used to model hypersensitivity to chemotherapy [19] and the selection of escape mutants from immune recognition during HIV infection [14]. In [21] the *CimmSim* framework was applied to the study of mechanisms that contribute to the persistence of infection with the Epstein–Barr virus.

A Voting Game The following example describes a hypothetical voting scheme. The voting system is constructed from a collection of voters. For simplicity, it is assumed that only two candidates, represented by 0 and 1, contest in the election. There are N voters represented by the set $\{v_1, v_2, \ldots, v_N\}$. Each voter has a candidate preference or a *state*. We denote the state of voter v_i by x_i. Moreover, each voter knows the preferences or states of some of his or her friends (fellow voters). This friendship relation is captured by the *dependency graph* which we describe later in Subsect. "Definitions, Background, and Examples". In-

formally, the dependency graph has as vertices the voters with an edge between each pair of voters that are friends.

Starting from an initial configuration of preferences, the voters cast their votes in some order. The candidate that receives the most votes is the winner. A number of rules can be formulated to decide how each voter chooses a candidate. We will provide examples of such rules later, and as will be seen, the outcome of the election is governed by the order in which the votes are cast as well as the structure of the dependency graph.

Existing Mathematical Frameworks

The field of agent-based simulation currently places heavy emphasis on implementation and computation rather than on the derivation of formal results. Computation is no doubt a very useful way to discover potentially interesting behavior and phenomena. However, unless the simulation has been set up very carefully, its outcome does not formally validate or guarantee the observed phenomenon. It could simply be caused by an artifact of the system model, an implementation error, or some other uncertainty.

A first step in a theory of agent-based simulation is the introduction of a formal framework that on the one hand is precise and computationally powerful, and, on the other hand, is natural in the sense that it can be used to effectively describe large classes of both deterministic and stochastic systems. Apart from providing a common basis and a language for describing the model using a sound formalism, such a framework has many advantages. At a first level, it helps to clarify the key structure in a system. Domain specific knowledge is crucial to deriving good models of complex systems, but domain specificity is often confounded by domain conventions and terminology that eventually obfuscate the real structure.

A formal, context independent framework also makes it easier to take advantage of existing general theory and algorithms. Having a model formulated in such a framework also makes it easier to establish results. Additionally, expressing the model using a general framework is more likely to produce results that are widely applicable. This type of framework also supports implementation and validation. Modeling languages like UML [16] serve a similar purpose, but tend to focus solely on software implementation and validation issues, and very little on mathematical or computational analysis.

Cellular Automata

In this section we discuss several existing frameworks for describing agent-based simulations. Cellular automata (CA) were introduced by Ulam and von Neumann [84] as biologically motivated models of computation. Early research addressed questions about the computational power of these devices. Since then their properties have been studied in the context of dynamical systems [40], language theory [52], and ergodic theory [51] to mention just a few areas. Cellular automata were popularized by Conway [35] (Game of Life) and by Wolfram [55,86,88]. Cellular automata (both deterministic and stochastic) have been used as models for phenomena ranging from lattice gases [32] and flows in porous media [77] to traffic analysis [33,63,65].

A cellular automaton is typically defined over a regular grid. An example is a two-dimensional grid such as $\mathbf{Z}^2$. Each grid point (i, j) is referred to as a *site* or *node*. Each site has a state $x_{i,j}(t)$ which is often taken to be binary. Here t denotes the time step. Furthermore, there is a notion of a *neighborhood* for each site. The neighborhood N of a site is the collection of sites that can influence the future state of the given site. Based on its current state $x_{i,j}(t)$ and the current states of the sites in its neighborhood N, a function $f_{i,j}$ is used to compute the next state $x_{i,j}(t+1)$ of the site (i, j). Specifically, we have

$$x_{i,j}(t+1) = f_{i,j}(\bar{x}_{i,j}(t)) \,, \tag{1}$$

where $\bar{x}_{i,j}(t)$ denotes the tuple consisting of all the states $x_{i',j'}(t)$ with $(i', j') \in N$. The tuple consisting of the states of all the sites is the CA *configuration* and is denoted $x(t) = (x_{i,j}(t))_{i,j}$. Equation (1) is used to map the configuration $x(t)$ to $x(t+1)$. The cellular automaton map or dynamical system, is the map Φ that sends $x(t)$ to $x(t+1)$.

A central part of CA research is to understand how configurations evolve under iteration of the map Φ and what types of dynamical behavior can be generated. A general introduction to CA can be found in [43].

Hopfield Networks

Hopfield networks were proposed as a simple model of associative memories [42]. A discrete Hopfield neural network consists of an undirected graph $Y(V, E)$. At any time t, each node $v_i \in V$ has a state $x_i(t) \in \{+1, -1\}$. Further, each node $v_i \in V$ has an associated *threshold* $\tau_i \in \mathbf{R}$. Each edge $\{v_i, v_j\} \in E$ has an associated weight $w_{i,j} \in \mathbf{R}$. For each node v_i, the neighborhood N_i of v_i includes v_i and the set of nodes that are adjacent to v_i in Y. Formally,

$$N_i = \{v_i\} \cup \{v_j \in V : \{v_i, v_j\} \in E\} \,.$$

States of nodes are updated as follows. At time t, node v_i computes the function f_i defined by

$$f_i(t) = \operatorname{sgn}\left(-\tau_i + \sum_{v_j \in N_i} w_{i,j}\, x_j(t)\right) ,$$

where sgn is the map from $\mathbf{R}$ to $\{+1, -1\}$, defined by

$$\operatorname{sgn}(x) = \begin{cases} 1\,, & \text{if } x \geq 0 \quad \text{and} \\ -1\,, & \text{otherwise}\,. \end{cases}$$

Now, the state of v_i at time $t+1$ is

$$x_i(t+1) = f_i(t)\,.$$

Many references on Hopfield networks (see for example [42,78]) assume that the underlying undirected graph is complete; that is, there is an edge between every pair of nodes. In the definition presented above, the graph need not be complete. However, this does not cause any difficulties since the missing edges can be assigned weight 0. As a consequence, such edges will not play any role in determining the dynamics of the system. Both synchronous and asynchronous update models of Hopfield neural networks have been considered in the literature. For theoretical results concerning Hopfield networks see [69,70] and the references cited therein. Reference [78] presents a number of applications of neural networks. In [54], a Hopfield model is used to study polarization in dynamic networks.

Communicating Finite State Machines

The model of communicating finite state machines (CFSM) was proposed to analyze protocols used in computer networks. In some of the literature, this model is also referred to as "concurrent transition systems" [36].

In the CFSM model, each agent is a process executing on some node of a distributed computing system. Although there are minor differences among the various CFSM models proposed in the literature [17,36], the basic set-up models each process as a finite state machine (FSM). Thus, each agent is in a certain state at any time instant t. For each pair of agents, there is a bidirectional channel through which they can communicate. The state of an agent at time $t+1$ is a function of the current state and the input (if any) on one or more of the channels. When an agent (FSM) undergoes a transition from one state to another, it may also choose to send a message to another agent or receive a message from an agent. In general, such systems can be synchronous or asynchronous. As can be seen, CFSMs are a natural formalism for studying protocols used in computer networks. The CFSM model has been used extensively to prove properties (e. g. deadlock freedom, fairness) of a number of protocols used in practice (see [17,36] and the references cited therein).

Other frameworks include interacting particle systems [50], and Petri nets [59]. There is a vast literature on both, but space limitations prevent a discussion here.

Finite Dynamical Systems

Another, quite general, modeling framework that has been proposed is that of *finite dynamical systems*, both synchronous and asynchronous. Here the proposed mathematical object representing an agent-based simulation is a time-discrete dynamical system on a finite state set. The description of the systems is modeled after the key components of an agent-based simulation, namely agents, the dependency graph, local update functions, and an update order. This makes a mapping to agent-based simulations natural. In the remainder of this chapter we will show that finite dynamical systems satisfy our criteria for a good mathematical framework in that they are general enough to serve as a broad computing tool and mathematically rich enough to allow the derivation of formal results.

Definitions, Background, and Examples

Let $x_1, \ldots, x_n$ be a collection of variables, which take values in a finite set X. (As will be seen, the variables represent the entities in the system being modeled and the elements of X represent their states.) Each variable x_i has associated to it a "local update function" $f_i\colon X^n \longrightarrow X$, where "local" refers to the fact that f_i takes inputs from the variables in the "neighborhood" of x_i, in a sense to be made precise below. By abuse of notation we also let f_i denote the function $X^n \longrightarrow X^n$ which changes the ith coordinate and leaves the other coordinates unchanged. This allows for the sequential composition of the local update functions. These functions assemble to a dynamical system

$$\Phi = (f_1, \ldots, f_n)\colon X^n \longrightarrow X^n\,,$$

with the dynamics generated by iteration of Φ. As an example, if $X = \{0, 1\}$ with the standard Boolean operators AND and OR, then Φ is a Boolean network.

The assembly of Φ from the local functions f_i can be done in one of several ways. One can update each of the variables simultaneously, that is,

$$\Phi(x_1, \ldots, x_n) = (f_1(x_1, \ldots, x_n), \ldots, f_n(x_1, \ldots, x_n))\,.$$

In this case one obtains a *parallel dynamical system*.

Alternatively, one can choose to update the states of the variables according to some fixed update order, for example, a permutation $(\pi_1, \pi_2, \ldots, \pi_n)$ of the set $\{1, \ldots, n\}$. More generally, one could use a *word* on the set

$\{1, \ldots, n\}$, that is, $\pi = (\pi_1, \ldots, \pi_t)$ where t is the length of the word. The function composition

$$\Phi_\pi = f_{\pi_t} \circ f_{\pi_{t-1}} \circ \cdots \circ f_{\pi_1} \,, \tag{2}$$

is called a *sequential dynamical system (SDS)* and, as before, the dynamics of Φ_π is generated by iteration. The case when π is a permutation on $\{1, \ldots, n\}$ has been studied extensively [2,3,5,9]. It is clear that using a different permutation or word σ may result in a different dynamical system Φ_σ. Using a word rather than a permutation allows one to capture the case where some vertices have states that are updated more frequently than others.

Remark 1 Notice that for a fixed π, the function Φ_π is a parallel dynamical system: once the update order π is chosen and the local update functions are composed according to π, that is, the function Φ_π has been computed, then $\Phi_\pi(x_1, \ldots, x_n) = g(x_1, \ldots, x_n)$ where g is a parallel update dynamical system. However, the maps g_i are not local functions.

The dynamics of Φ is usually represented as a directed graph on the vertex set X^n, called the *phase space* of Φ. There is a directed edge from $\mathbf{v} \in X^n$ to $\mathbf{w} \in X^n$ if and only if $\Phi(\mathbf{v}) = \mathbf{w}$. A second graph that is usually associated with a finite dynamical system is its *dependency graph* $Y(V, E)$. In general, this is a directed graph, and its vertex set is $V = \{1, \ldots, n\}$. There is a directed edge from i to j if and only if x_i appears in the function f_j. In many situations, the interaction relationship between pairs of variables is symmetric; that is, variable x_i appears in f_j if and only if x_j appears in f_i. In such cases, the dependency graph can be thought of as an undirected graph. We recall that the dependency graphs mentioned in the context of the voting game (Subsect. "A Voting Game") and Hopfield networks (Subsect. "Hopfield Networks") are undirected graphs. The dependency graph plays an important role in the study of finite dynamical systems and is sometimes listed explicitly as part of the specification of Φ.

Example 2 Let $X = \{0, 1\}$ (the Boolean case). Suppose we have four variables and the local Boolean update functions are

$$\begin{aligned} f_1 &= x_1 + x_2 + x_3 + x_4, \\ f_2 &= x_1 + x_2, \\ f_3 &= x_1 + x_3, \\ f_4 &= x_1 + x_4 \,, \end{aligned}$$

where "+" represents sum modulo 2. The dynamics of the function $\Phi = (f_1, \ldots, f_4)\colon X^4 \longrightarrow X^4$ is the directed graph in Fig. 1a while the dependency graph is in Fig. 1b.

Example 3 Consider the local functions in the Example 2 above and let $\pi = (2, 1, 3, 4)$. Then

$$\Phi_\pi = f_4 \circ f_3 \circ f_1 \circ f_2 \colon X^4 \longrightarrow X^4 \,.$$

The phase space of Φ_π is the directed graph in Fig. 2a, while the phase space of Φ_γ, where $\gamma = id$ is in Fig. 2b.

Notice that the phase space of any function is a directed graph in which every vertex has out-degree one; this is a characteristic property of deterministic functions.

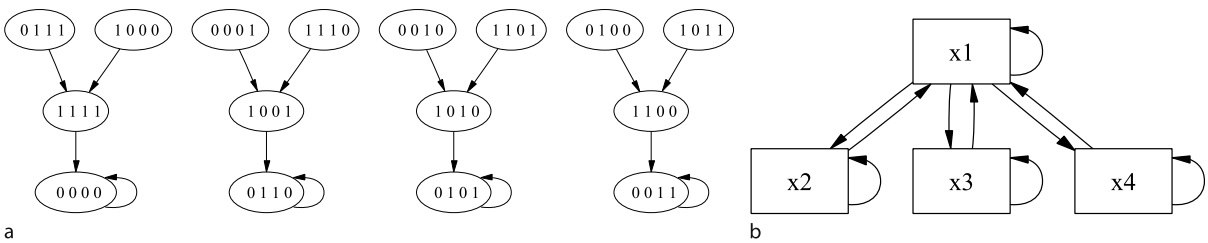

Agent Based Modeling, Mathematical Formalism for, Figure 1
The phase space of the parallel system Φ (a) and dependency graph of the Boolean functions from Example 2 (b)

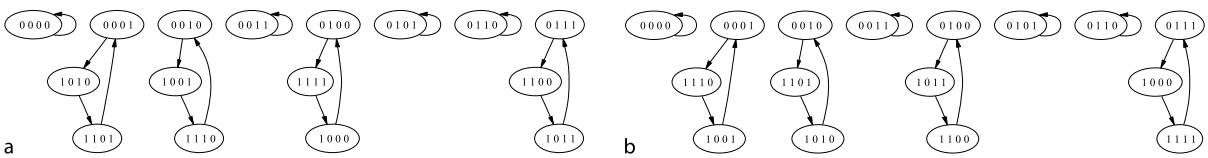

Agent Based Modeling, Mathematical Formalism for, Figure 2
The phase spaces from Example 3: Φ_π (a) and Φ_{id} (b)

Making use of Boolean arithmetic is a powerful tool in studying Boolean networks, which is not available in general. In order to have available an enhanced set of tools it is often natural to make an additional assumption regarding X, namely that it is a finite number system, a *finite field* [49]. This amounts to the assumption that there are "addition" and "multiplication" operations defined on X that satisfy the same rules as ordinary addition and multiplication of numbers. Examples include $\mathbf{Z}_p$, the integers modulo a prime p. This assumption can be thought of as the discrete analog of imposing a coordinate system on an affine space.

When X is a finite field, it is easy to show that for any local function g, there exists a polynomial h such that $g(x_1,\ldots,x_n) = h(x_1,\ldots,x_n)$ for all $(x_1,\ldots,x_n) \in X^n$. To be precise, suppose X is a finite field with q elements. Then

$$g(x_1,\ldots,x_n) = \sum_{(c_1,\ldots,c_n)\in X^n} g(c_1,\ldots,c_n) \prod_{i=1}^{n} (1-(x_i-c_i)^{q-1}) \,. \quad (3)$$

This observation has many useful consequences, since polynomial functions have been studied extensively and many analytical tools are available.

Notice that cellular automata and Boolean networks, both parallel and sequential, are special classes of polynomial dynamical systems. In fact, it is straightforward to see that

$$x \wedge y = x \cdot y \quad x \vee y = x + y + xy \quad \text{and} \quad \neg x = x + 1 \,. \quad (4)$$

Therefore, the modeling framework of finite dynamical systems includes that of cellular automata, discussed earlier. Also, since a Hopfield network is a function $X^n \longrightarrow X^n$, which can be represented through its local constituent functions, it follows that Hopfield networks also are special cases of finite dynamical systems.

Stochastic Finite Dynamical Systems

The deterministic framework of finite dynamical systems can be made stochastic in several different ways, making one or more of the system's defining data stochastic. For example, one could use one or both of the following criteria.

- Assume that each variable has a nonempty set of local functions assigned to it, together with a probability distribution on this set, and each time a variable is updated, one of these local functions is chosen at random to update its state. We call such systems *probabilistic finite dynamical systems* (PFDS), a generalization of probabilistic Boolean networks [81].
- Fix a subset of permutations $T \subseteq S_n$ together with a probability distribution. When it is time for the system to update its state, a permutation $\pi \in T$ is chosen at random and the agents are updated sequentially using π. We call such systems *stochastic finite dynamical systems* (SFDS).

Remark 4 By Remark, each system Φ_π is a parallel system. Hence a SFDS is nothing but a set of parallel dynamical systems $\{\Phi_\pi : \pi \in T\}$, together with a probability distribution. When it is time for the system to update its state, a system Φ_π is chosen at random and used for the next iteration.

To describe the phase space of a stochastic finite dynamical system, a general method is as follows. Let Ω be a finite collection of systems $\Phi_1,\ldots,\Phi_t$, where $\Phi_i : X^n \longrightarrow X^n$ for all i, and consider the probabilities $p_1,\ldots,p_t$ which sum to 1. We obtain the stochastic phase space

$$\Gamma_\Omega = p_1\Gamma_1 + p_2\Gamma_2 + \cdots + p_t\Gamma_t \,, \quad (5)$$

where Γ_i is the phase space of Φ_i. The associated probability space is $\mathcal{F} = (\Omega, 2^\Omega, \mu)$, where the probability measure μ is induced by the probabilities p_i. It is clear that the stochastic phase space can be viewed as a *Markov chain* over the state space X^n. The adjacency matrix of Γ_Ω directly encodes the Markov transition matrix. This is of course not new, and has been done in, e. g., [28,81,83]. But we emphasize the point that even though SFDS give rise to Markov chains *our study of SFDS is greatly facilitated by the rich additional structure* available in these systems. To understand the effect of structural components such as the topology of the dependency graph or the stochastic nature of the update, it is important to study them not as Markov chains but as SFDS.

Example 5 Consider Φ_π and Φ_γ from Example 3 and let Γ_π and Γ_γ be their phases spaces as shown in Fig. 2. Let $p_1 = p_2 = 1/2$. The phase space $(1/2)\Gamma_\pi + (1/2)\Gamma_\gamma$ of the stochastic sequential dynamical system obtained from Φ_π and Φ_γ (with equal probabilities) is presented in Fig. 3.

Agent-Based Simulations as Finite Dynamical Systems

In the following we describe the generic structure of the systems typically modeled and studied through agent-based simulations. The central notion is naturally that of an *agent*.

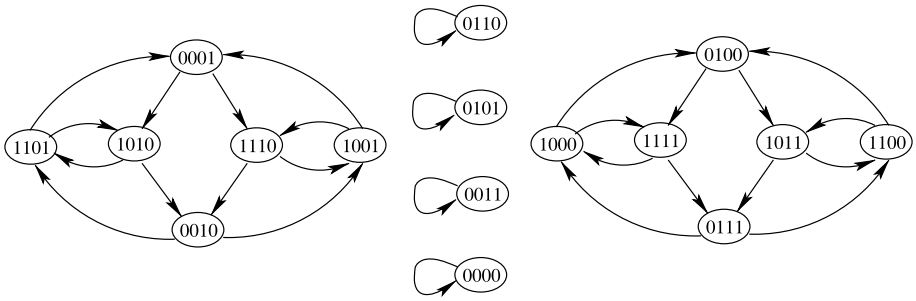

Agent Based Modeling, Mathematical Formalism for, Figure 3
The stochastic phase space for Example 5 induced by the two deterministic phase spaces of Φ_π and Φ_γ from Fig. 2. For simplicity the weights of the edges have been omitted

Each agent carries a state that may encode its preferences, internal configuration, perception of its environment, and so on. In the case of TRANSIMS, for instance, the agents are the cells making up the road network. The cell state contains information about whether or not the cell is occupied by a vehicle as well as the velocity of the vehicle. One may assume that each cell takes on states from the same set of possible states, which may be chosen to support the structure of a finite field.

The agents interact with each other, but typically an agent only interacts with a small subset of agents, its *neighbors*. Through such an interaction an agent may decide to change its state based on the states (or other aspects) of the agents with which it interacts. We will refer to the process where an agent modifies its state through interaction as an *agent update*. The precise way in which an agent modifies its state is governed by the nature of the particular agent. In TRANSIMS the neighbors of a cell are the adjacent road network cells. From this adjacency relation one obtains a dependency graph of the agents. The local update function for a given agent can be obtained from the rules governing traffic flow between cells.

The updates of all the agents may be scheduled in different ways. Classical approaches include synchronous, asynchronous and event-driven schemes. The choice will depend on system properties or particular considerations about the simulation implementation.

In the case of *CImmSim*, the situation is somewhat more complicated. Here the agents are also the spatial units of the system, each representing a small volume of lymph tissue. The total volume is represented as a 2-dimensional CA, in which every agent has 4 neighbors, so that the dependency graph is a regular 2-dimensional grid. The state of each agent is a collection of counts for the various immune cells and pathogens that are present in this particular agent (volume). Movement between cells is implemented as diffusion. Immune cells can interact with each other and with pathogens while they reside in the same volume. Thus, the local update function for a given cell of the simulation is made up of the two components of movement between cells and interactions within a cell. For instance, a B cell could interact with the Epstein–Barr virus in a given volume and perform a transition from uninfected to infected by the next time step. Interactions as well as movement are stochastic, resulting in a stochastic finite dynamical system. The update order is parallel.

Example 6 (The voting game (Subsect. "A Voting Game")) The following scenario is constructed to illustrate how implementation choices for the system components have a clear and direct bearing on the dynamics and simulation outcomes.

Let the voter dependency graph be the star graph on 5 vertices with center vertex a and surrounding vertices b, c, d and e. Furthermore, assume that everybody votes opportunistically using the majority rule: the vote cast by an individual is equal to the preference of the majority of his/her friends with the person's own preference included. For simplicity, assume candidate 1 is preferred in the case of a tie.

If the initial preference is $x_a = 1$ and $x_b = x_c = x_d = x_e = 0$ then if voter a goes first he/she will vote for candidate 0 since that is the choice of the majority of the neighbor voters. However, if b and c go first they only know a's preference. Voter b therefore casts his/her vote for candidate 1 as does c. Note that this is a tie situation with an equal number of preferences for candidate 1 (a) and for candidate 2 (b). If voter a goes next, then the situation has changed: the preference of b and c has already changed to 1. Consequently, voter a picks candidate 1. At the end of the day candidate 1 is the election winner, and the choice of update order has tipped the election!

This example is of course constructed to illustrate our point. However, in real systems it can be much more difficult to detect the presence of such sensitivities and their implications. A solid mathematical framework can be very helpful in detecting such effects.

Finite Dynamical Systems as Theoretical and Computational Tools

If finite dynamical systems are to be useful as a modeling paradigm for agent-based simulations it is necessary that they can serve as a fairly universal model of computation. We discuss here how such dynamical systems can mimic Turing Machines (TMs), a standard universal model for computation. For a more thorough exposition, we refer the reader to the series of papers by Barrett et al. [7,8,10,11,12]. To make the discussion reasonably self-contained, we provide a brief and informal discussion of the TM model. Additional information on TMs can be found in any standard text on the theory of computation (e. g. [82]).

A Computational View of Finite Dynamical Systems: Definitions

In order to understand the relationship of finite dynamical systems to TMs, it is important to view such systems from a computational stand point. Recall that a finite dynamical system $\Phi : X^n \longrightarrow X^n$, where X is a finite set, has an underlying dependency graph $Y(V, E)$. From a computational point of view, the nodes of the dependency graph (the agents in the simulation) are thought of as devices that compute appropriate functions. For simplicity, we will assume in this section that the dependency graph is undirected, that is, all dependency relations are symmetric. At any time, the state value of each node $v_i \in V$ is from the specified domain X. The inputs to f_i are the current state of v_i and the states of the neighbors of v_i as specified by Y. The output of f_i, which is also a member of X, becomes the state of v_i at the next time instant. The discussion in this section will focus on sequentially updated systems (SDS), but all results discussed apply to parallel systems as well.

Each step of the computation carried out by an SDS can be thought as consisting of n "mini steps;" in each mini step, the value of the local transition function at a node is computed and the state of the node is changed to the computed value. Given an SDS Φ, a *configuration* C of Φ is a vector $(c_1, c_2, \ldots, c_n) \in X^n$. It can be seen that each computational step of an SDS causes a transition from one configuration to another.

Configuration Reachability Problem for SDSs

Based on the computational view, a number of different problems can be defined for SDSs (see for example, [4,10,12]). To illustrate how SDSs can model TM computations, we will focus on one such problem, namely the *configuration reachability* (CR) problem: Given an SDS Φ, an initial configuration C and another configuration C', will Φ, starting from C, ever reach configuration C'? The problem can also be expressed in terms of the phase space of Φ. Since configurations such as C and C' are represented by nodes in the phase space, the CR problem boils down to the question of whether there is a directed path in the phase space from C to C'. This abstract problem can be mapped to several problems in the simulation of multi-agent systems. Consider for example the TRANSIMS context. Here, the initial configuration C may represent the state of the system early in the day (when the traffic is very light) and C' may represent an "undesirable" state of the system (such as heavy traffic congestion). Similarly, in the context of modeling an infectious disease, C may represent the initial onset of the disease (when only a small number of people are infected) and C' may represent a situation where a large percentage of the population is infected. The purpose of studying computational problems such as CR is to determine whether one can efficiently predict the occurrence of certain events in the system from a description of the system. If computational analysis shows that the system can indeed reach undesirable configurations as it evolves, then one can try to identify steps needed to deal with such situations.

Turing Machines: A Brief Overview

A Turing machine (TM) is a simple and commonly used model for general purpose computational devices. Since our goal is to point out how SDSs can also serve as computational devices, we will present an informal overview of the TM model. Readers interested in a more formal description may consult [82].

A TM consists of a set Q of states, a one-way infinite input tape and a read/write head that can read and modify symbols on the input tape. The input tape is divided into cells and each cell contains a symbol from a special finite alphabet. An input consisting of n symbols is written on the leftmost n cells of the input tape. (The other cells are assumed to contain a special symbol called *blank*.) One of the states in Q, denoted by q_s, is the designated *start* state. Q also includes two other special states, denoted by q_a (the *accepting* state) and q_r (the *rejecting* state). At any time, the machine is in one of the states in Q. The *transition function* for the TM specifies for each combination of the current

state and the current symbol under the head, a new state, a new symbol for the current cell (which is under the head) and a movement (i. e., left or right by one cell) for the head. The machine starts in state q_s with the head on the first cell of the input tape. Each step of the machine is carried out in accordance with the transition function. If the machine ever reaches either the accepting or the rejecting state, it halts with the corresponding decision; otherwise, the machine runs forever.

A *configuration* of a TM consists of its current state, the current tape contents and the position of the head. Note that the transition function of a TM specifies how a new configuration is obtained from the current configuration.

The above description is for the basic TM model (also called *single tape* TM model). For convenience in describing some computations, several variants of the above basic model have been proposed. For example, in a *multi-tape* TM, there are one or more *work tapes* in addition to the input tape. The work tapes can be used to store intermediate results. Each work tape has its own read/write head and the definitions of configuration and transition function can be suitably modified to accommodate work tapes. While such an enhancement to the basic TM model makes it easier to carry out certain computations, it does not add to the machine's computational power. In other words, any computation that can be carried out using the enhanced model can also be carried out using the basic model.

As in the case of dynamical systems, one can define a *configuration reachability* (CR) problem for TMs: Given a TM M, an initial configuration $\mathcal{I}_M$ and another configuration C_M, will the TM starting from $\mathcal{I}_M$ ever reach C_M? We refer to the CR problem in the context of TMs as CR-TM. In fact, it is this problem for TMs that captures the essence of what can be effectively computed. In particular, by choosing the state component of C_M to be one of the halting states (q_a or q_r), the problem of determining whether a function is computable is transformed into an appropriate CR-TM problem. By imposing appropriate restrictions on the resources used by a TM (e. g. the number of steps, the number of cells on the work tapes), one obtains different versions of the CR-TM problem which characterize different computational complexity classes [82].

How SDSs Can Mimic Turing Machines

The above discussion points out an important similarity between SDSs and TMs. Under both of these models, each computational step causes a transition from one configuration to another. It is this similarity that allows one to construct a discrete dynamical system Φ that can simulate a TM. Typically, each step of a TM is simulated by a short sequence of successive iterations Φ. As part of the construction, one also identifies a suitable mapping between the configurations of the TM being simulated and those of the dynamical system. This is done in such a way that the answer to CR-TM problem is "yes" if and only if the answer to the CR problem for the dynamical system is also "yes."

To illustrate the basic ideas, we will informally sketch a construction from [10]. For simplicity, this construction produces an SDS Φ that simulates a restricted version of TMs; the restriction being that for any input containing n symbols, the number of work tape cells that the machine may use is bounded by a linear function of n. Such a TM is called a *linear bounded automaton* (LBA) [82]. Let M denote the given LBA and let n denote the length of the input to M. The domain X for the SDS Φ is chosen to be a finite set based on the allowed symbols in the input to the TM. The dependency graph is chosen to be a simple path on n nodes, where each node serves as a representative for a cell on the input tape. The initial and final configurations C and C' for Φ are constructed from the corresponding configurations of M. The local transition function for each node of the SDS is constructed from the given transition function for M in such a way that each step of M corresponds to exactly one step of Φ. Thus, there is a simple bijection between the sequence of configurations that M goes through during its computation and the sequence of states that Φ goes through as it evolves. Using this bijection, it is shown in [10] that the answer to the CR-TM problem is "yes" if and only if Φ reaches C' starting from C. Reference [10] also presents a number of sophisticated constructions where the resulting dynamical system is very simple; for example, it is shown that an LBA can be simulated by an SDS in which X is the Boolean field, the dependency graph is d-regular for some constant d and the local transition functions at all the nodes are *identical*. Such results point out that one does not need complicated dynamical systems to model TM computations.

References [7,8,10] present constructions that show how more general models of TMs can also be simulated by appropriate SDSs. As one would expect, these constructions lead to SDSs with more complex local transition functions.

Barrett et al. [11] have also considered *stochastic* SDS (SSDS), where the local transition function at each node is stochastic. For each combination of inputs, a stochastic local transition function specifies a probability distribution over the domain of state values. It is shown in [11] that SSDSs can effectively simulate computations carried

out by probabilistic TMs (i. e., TMs whose transition functions are stochastic).

TRANSIMS Related Questions

Section "TRANSIMS (Transportation, Analysis, Simulation System)" gave an overview of some aspects of the TRANSIMS model. The micro-simulator module is specified as a functional composition of four cellular automata of the form $\Delta_4 \circ \Delta_3 \circ \Delta_2 \circ \Delta_1$. (We only described Δ_3 which corresponds to velocity updates.) Such a formulation has several advantages. First, it has created an abstraction of the essence of the system in a precise mathematical way without burying it in contextual domain details. An immediate advantage of this specification is that it makes the whole implementation process more straightforward and transparent. While the local update functions for the cells are typically quite simple, for any realistic study of an urban area the problem size would typically require a sequential implementation, raising a number of issues that are best addressed within a mathematical framework like the one considered here.

Mathematical Results on Finite Dynamical Systems

In this section we outline a collection of mathematical results about finite dynamical systems that is representative of the available knowledge. The majority of these results are about deterministic systems, as the theory of stochastic systems of this type is still in its infancy. We will first consider synchronously updated systems.

Throughout this section, we make the assumption that the state set X carries the algebraic structure of a finite field. Accordingly, we use the notation k instead of X. It is a standard result that in this case the number q of elements in k has the form $q = p^t$ for some prime p and $t \geq 1$. The reader may keep the simplest case $k = \{0, 1\}$ in mind, in which case we are effectively considering Boolean networks.

Recall Eq. (3). That is, any function $g\colon k^n \longrightarrow k$ can be represented by a multivariate polynomial with coefficients in k. If we require that the exponent of each variable be less than q, then this representation is unique. In particular Eq. (4) implies that every Boolean function can be represented uniquely as a polynomial function.

Parallel Update Systems

Certain classes of finite dynamical systems have been studied extensively, in particular cellular automata and Boolean networks where the state set is $\{0, 1\}$. Many of these studies have been experimental in nature, however. Some more general mathematical results about cellular automata can be found in the papers of Wolfram and collaborators [87]. The results there focus primarily on one-dimensional Boolean cellular automata with a particular fixed initial state. Here we collect a sampling of more general results.

We first consider linear and affine systems

$$\Phi = (f_1, \ldots, f_n)\colon k^n \longrightarrow k^n .$$

That is, we consider systems for which the coordinate functions f_i are linear, resp. affine, polynomials. (In the Boolean case this includes functions constructed from XOR (sum modulo 2) and negation.) When each f_i is a linear polynomial of the form $f_i(x_1, \ldots, x_n) = a_{i1}x_1 + \cdots + a_{in}x_n$, the map Φ is nothing but a *linear transformation* on k^n over k, and, by using the standard basis, Φ has the matrix representation

$$\Phi\left(\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}\right) = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix},$$

where $a_{ij} \in k$ for all i, j.

Linear finite dynamical systems were first studied by Elspas [30]. His motivation came from studying feedback shift register networks and their applications to radar and communication systems and automatic error correction circuits. For linear systems over finite fields of prime cardinality, that is, q is a prime number, Elspas showed that the exact number and length of each limit cycle can be determined from the *elementary divisors* of the matrix A. Recently, Hernandez [41] re-discovered Elspas' results and generalized them to arbitrary finite fields. Furthermore, he showed that the structure of the *tree of transients* at each node of each limit cycle is the same, and can be completely determined from the nilpotent elementary divisors of the form x^a. For *affine* Boolean networks (that is, finite dynamical systems over the Boolean field with two elements, whose local functions are linear polynomials which might have constant terms), a method to analyze their cycle length has been developed in [56]. After embedding the matrix of the transition function, which is of dimension $n \times (n + 1)$, into a square matrix of dimension $n + 1$, the problem is then reduced to the linear case. A fast algorithm based on [41] has been implemented in [44], using the symbolic computation package *Macaulay2*.

It is not surprising that the phase space structure of Φ should depend on invariants of the matrix $A = (a_{ij})$. The rational canonical form of A is a block-diagonal matrix and one can recover the structure of the phase space of A

from that of the blocks in the rational form of A. Each block represents either an invertible or a nilpotent linear transformation. Consider an invertible block B. If $\mu(x)$ is the minimal polynomial of B, then there exists s such that $\mu(x)$ divides $x^s - 1$. Hence $B^s - I = 0$ which implies that $B^s\mathbf{v} = \mathbf{v}$. That is, every state vector $\mathbf{v}$ in the phase space of B is in a cycle whose length is a divisor of s.

Definition 7 For any polynomial $\lambda(x)$ in $k[x]$, the *order* of $\lambda(x)$ is the least integer s such that $\lambda(x)$ divides $x^s - 1$.

The cycle structure of the phase space of Φ can be completely determined from the orders of the irreducible factors of the minimal polynomial of Φ. The computation of these orders involves in particular the factorization of numbers of the form $q^r - 1$, which makes the computation of the order of a polynomial potentially quite costly. The nilpotent blocks in the decomposition of A determine the tree structure at the nodes of the limit cycles. It turns out that all trees at all periodic nodes are identical. This generalizes a result in [55] for additive cellular automata over the field with two elements.

While the fact that the structure of the phase space of a linear system can be determined from the invariants associated with its matrix may not be unexpected, it is a beautiful example of how the right mathematical viewpoint provides powerful tools to completely solve the problem of relating the structure of the local functions to the resulting (or emerging) dynamics. Linear and affine systems have been studied extensively in several different contexts and from several different points of view, in particular the case of cellular automata. For instance, additive cellular automata over more general rings as state sets have been studied, e. g., in [25]. Further results on additive CAs can also be found there. One important focus in [25] is on the problem of finding CAs with limit cycles of maximal length for the purpose of constructing pseudo random number generators.

Unfortunately, the situation is more complicated for nonlinear systems. For the special class of Boolean synchronous systems whose local update functions consist of monomials, there is a polynomial time algorithm that determines whether a given monomial system has only fixed points as periodic points [26]. This question was motivated by applications to the modeling of biochemical networks. The criterion is given in terms of invariants of the dependency graph Y. For a strongly connected directed graph Y (i. e., there is a directed path between any pair of vertices), its *loop number* is the greatest common divisor of all directed loops at a particular vertex. (This number is independent of the vertex chosen.)

Theorem 8 ([26]) *A Boolean monomial system has only fixed points as periodic points if and only if the loop number of every strongly connected component of its dependency graph is equal to 1.*

In [27] it is shown that the problem for general finite fields can be reduced to that of a Boolean system and a linear system over rings of the form $\mathbf{Z}/p^r\mathbf{Z}$, p prime. Boolean monomial systems have been studied before in the cellular automaton context [13].

Sequential Update Systems

The update order in a sequential dynamical system has been studied using combinatorial and algebraic techniques. A natural question to ask here is how the system depends on the update schedule. In [3,5,61,72] this was answered on several levels for the special case where the update schedule is a permutation. We describe these results in some detail. Results about the more general case of update orders described by words on the indices of the local update functions can be found in [34].

Given local update functions $f_i\colon k^n \longrightarrow k$ and permutation update orders σ, π, a natural question is when the two resulting SDS Φ_σ and Φ_π are identical and, more generally, how many different systems one obtains by varying the update order over all permutations. Both questions can be answered in great generality. The answer involves invariants of two graphs, namely the acyclic orientations of the dependency graph Y of the local update functions and the *update graph* of Y. The update graph $U(Y)$ of Y is the graph whose vertex set consists of all permutations of the vertex set of Y [72]. There is an (undirected) edge between two permutations $\sigma = (\sigma_1, \ldots, \sigma_n)$ and $\tau = (\tau_1, \ldots, \tau_n)$ if they differ by a transposition of two adjacent entries σ_i and σ_{i+1} such that there is no edge in Y between σ_i and σ_{i+1}.

The update graph encodes the fact that one can commute two local update functions f_i and f_j without affecting the end result Φ if i and j are not connected by an edge in Y. That is, $\cdots f_i \circ f_j \cdots = \cdots f_j \circ f_i \cdots$ if and only if i and j are not connected by an edge in Y.

All permutations belonging to the same connected component in $U(Y)$ give identical *SDS* maps. The number of (connected) components in $U(Y)$ is therefore an upper bound for the number of functionally inequivalent *SDS* that can be generated by just changing the update order. It is convenient to introduce an equivalence relation $\sim_Y$ on S_Y by $\pi \sim_Y \sigma$ if π and σ belong to the same connected component in the graph $U(Y)$. It is then clear that if $\pi \sim_Y \sigma$ then corresponding sequential dynamical systems are identical as maps. This can also be characterized

in terms of acyclic orientations of the graph Y: each component in the update graph induces a unique acyclic orientation of the graph Y. Moreover, we have the following result:

Proposition 9 ([72]) *There is a bijection*

$$F_Y : S_Y/\sim_Y \longrightarrow \mathrm{Acyc}(Y) ,$$

where $S_Y/\sim_Y$ *denotes the set of equivalence classes of* $\sim_Y$ *and* $\mathrm{Acyc}(Y)$ *denotes the set of acyclic orientations of* Y.

This upper bound on the number of functionally different systems has been shown in [72] to be sharp for Boolean systems, in the sense that for a given Y one construct this number of different systems, using appropriate combinations of NOR functions.

For two permutations σ and τ it is easy to determine if they give identical *SDS* maps: one can just compare their induced acyclic orientations. The number of acyclic orientations of the graph Y tells how many functionally different *SDS* maps one can obtain for a fixed graph and fixed vertex functions. The work of Cartier and Foata [18] on partially commutative monoids studies a similar question, but their work is not concerned with finite dynamical systems.

Note that permutation update orders have been studied sporadically in the context of cellular automata on circle graphs [71] but not in a systematic way, typically using the order $(1, 2, \ldots, n)$ or the even-odd/odd-even orders. As a side note, we remark that this work also confirms our findings that switching from a parallel update order to a sequential order turns the complex behavior found in Wolfram's "class III and IV" automata into much more regular or mundane dynamics, see e. g. [79].

The work on functional equivalence was extended to dynamical equivalence (topological conjugation) in [5,61]. The automorphism group of the graph Y can be made to act on the components of the update graph $U(Y)$ and therefore also on the acyclic orientations of Y. All permutations contained in components of an orbit under $\mathrm{Aut}(Y)$ give rise to dynamically equivalent sequential dynamical systems, that is, to isomorphic phase spaces. However, here one needs some more technical assumptions, i. e., the local functions must be symmetric and induced, see [9]. This of course also leads to a bound for the number of dynamically inequivalent systems that can be generated by varying the update order alone. Again, this was first done for permutation update orders. The theory was extended to words over the vertex set of Y in [34,75].

The structure of the graph Y influences the dynamics of the system. As an example, graph invariants such as the independent sets of Y turn out to be in a bijective correspondence with the invariant set of sequential systems over the Boolean field $k = \{0, 1\}$ that have $\mathrm{nor}_t : k^t \to k$ given by $\mathrm{nor}_t(x_1, \ldots, x_t) = (1 + x_1) \cdots (1 + x_t)$ as local functions [73]. This can be extended to other classes such as those with order independent invariant sets as in [39]. We have already seen how the automorphisms of a graph give rise to equivalence [61]. Also, if the graph Y has nontrivial covering maps we can derive simplified or reduced (in an appropriate sense) versions of the original SDS over the image graphs of Y, see e. g. [62,74].

Parallel and sequential dynamical systems differ when it comes to invertibility. Whereas it is generally computationally intractable to determine if a CA over $\mathbf{Z}^d$ is invertible for $d \geq 2$ [46], it is straightforward to determine this for a sequential dynamical system [61]. For example, it turns out that the only invertible Boolean sequential dynamical systems with symmetric local functions are the ones where the local functions are either the parity function or the logical complement of the parity function [5].

Some classes of sequential dynamical systems such as the ones induced by the nor-function have desirable stability properties [39]. These systems have minimal invariant sets (i. e. periodic states) that do not depend on the update order. Additionally, these invariant sets are stable with respect to configuration perturbations.

If a state $\mathbf{c}$ is perturbed to a state $\mathbf{c}'$ that is not periodic this state will evolve to a periodic state $\mathbf{c}''$ in one step; that is, the system will quickly return to the invariant set. However, the states $\mathbf{c}$ and $\mathbf{c}''$ may not necessarily be in the same periodic orbit.

The Category of Sequential Dynamical Systems

As a general principle, in order to study a given class of mathematical objects it is useful to study transformations between them. In order to provide a good basis for a mathematical analysis the objects and transformations together should form a *category*, that is, the class of transformations between two objects should satisfy certain reasonable properties (see, e. g., [53]). Several proposed definitions of a transformation of SDS have been published, notably in [48] and [74]. One possible interpretation of a transformation of SDS from the point of view of agent-based simulation is that the transformation represents the approximation of one simulation by another or the embedding/projection of one simulation into/onto another. These concepts have obvious utility when considering different simulations of the same complex system.

One can take different points of view in defining a transformation of SDS. One approach is to require that

a transformation is compatible with the defining structural elements of an SDS, that is, with the dependency graph, the local update functions, and the update schedule. If this is done properly, then one should expect to be able to prove that the resulting transformation induces a transformation at the level of phase spaces. That is, transformations between SDS should preserve the local and global dynamic behavior. This implies that transformations between SDS lead to transformations between the associated global update functions.

Since the point of view of SDS is that global dynamics emerges from system properties that are defined locally, the notion of SDS transformation should focus on the local structure. This is the point of view taken in [48]. The definition given there is rather technical and the details are beyond the scope of this article. The basic idea is as follows. Let $\Phi_\pi = f_{\pi(n)} \circ \cdots \circ f_{\pi(1)}$ and $\Phi_\sigma = g_{\sigma(m)} \circ \cdots \circ g_{\sigma(1)}$ with the dependency graphs Y_π and Y_γ, respectively. A transformation $F\colon \Phi_\pi \longrightarrow \Phi_\sigma$ is determined by

- a graph mapping $\varphi\colon Y_\pi \longrightarrow Y_\gamma$ (reverse direction);
- a family of maps $k_{\phi(v)} \longrightarrow k_v$ with $v \in Y_\pi$;
- an order preserving map $\sigma \longmapsto \pi$ of update schedules.

These maps are required to satisfy the property that they "locally" assemble to a coherent transformation. Using this definition of transformation, it is shown (Theorem 2.6 in [48]) that the class of SDS forms a category. One of the requirements, for instance, is that the composition of two transformations is again a transformation. Furthermore, it is shown (Theorem 3.2 in [48]) that a transformation of SDS induces a map of directed graphs on the phase spaces of the two systems. That is, a transformation of the local structural elements of SDS induces a transformation of global dynamics. One of the results proven in [48] is that every SDS can be decomposed uniquely into a direct product (in the categorical sense) of indecomposable SDS.

Another possible point of view is that a transformation

$$F\colon (\Phi_\pi : k^n \to k^n) \longrightarrow (\Phi_\gamma : k^m \to k^m)$$

is a function $F\colon k^n \longrightarrow k^m$ such that $F \circ \Phi_\pi = \Phi_\gamma \circ F$, without requiring specific structural properties. This is the approach in [74]. This definition also results in a category, and a collection of mathematical results. Whatever definition chosen, much work remains to be done in studying these categories and their properties.

Future Directions

Agent-based computer simulation is an important method for modeling many complex systems, whose global dynamics emerges from the interaction of many local entities. Sometimes this is the only feasible approach, especially when available information is not enough to construct global dynamic models. The size of many realistic systems, however, leads to computer models that are themselves highly complex, even if they are constructed from simple software entities. As a result it becomes challenging to carry out verification, validation, and analysis of the models, since these consist in essence of complex computer programs. This chapter argues that the appropriate approach is to provide a formal mathematical foundation by introducing a class of mathematical objects to which one can map agent-based simulations. These objects should capture the key features of an agent-based simulation and should be mathematically rich enough to allow the derivation of general results and techniques. The mathematical setting of dynamical systems is a natural choice for this purpose.

The class of finite dynamical systems over a state set X which carries the structure of a finite field satisfies all these criteria. Parallel, sequential, and stochastic versions of these are rich enough to serve as the mathematical basis for models of a broad range of complex systems. While finite dynamical systems have been studied extensively from an experimental point of view, their mathematical theory should be considered to be in its infancy, providing a fruitful area of research at the interface of mathematics, computer science, and complex systems theory.

Bibliography

Primary Literature

1. Bagrodia RL (1998) Parallel languages for discrete-event simulation models. IEEE Comput Sci Eng 5(2):27–38
2. Barrett CL, Reidys CM (1999) Elements of a theory of simulation I: Sequential CA over random graphs. Appl Math Comput 98:241–259
3. Barrett CL, Mortveit HS, Reidys CM (2000) Elements of a theory of simulation II: Sequential dynamical systems. Appl Math Comput 107(2–3):121–136
4. Barrett CL, Hunt III HB, Marathe MV, Ravi SS, Rosenkrantz DJ, Stearns RE, Tosic P (2001) Garden of eden and fixed point configurations in sequential dynamical systems. In: Proc International Conference on Combinatorics, Computation and Geometry DM-CCG'2001. pp 95–110
5. Barrett CL, Mortveit HS, Reidys CM (2001) Elements of a theory of simulation III: Equivalence of SDS. Appl Math Comput 122:325–340
6. Barrett CL, Marathe MV, Smith JP, Ravi SS (2002) A mobility and traffic generation framework for modeling and simulating ad hoc communication networks. In: SAC'02 Madrid, ACM, pp 122–126
7. Barrett CL, Hunt III HB, Marathe MV, Ravi SS, Rosenkrantz DJ, Stearns RE (2003) On some special classes of sequential dynamical systems. Ann Comb 7(4):381–408
8. Barrett CL, Hunt III HB, Marathe MV, Ravi SS, Rosenkrantz DJ,

Stearns RE (2003) Reachability problems for sequential dynamical systems with threshold functions. Theor Comput Sci 295(1–3):41–64
9. Barrett CL, Mortveit HS, Reidys CM (2003) Elements of a theory of computer simulation. IV. sequential dynamical systems: fixed points, invertibility and equivalence. Appl Math Comput 134(1):153–171
10. Barrett CL, Hunt III HB, Marathe MV, Ravi SS, Rosenkrantz DJ, Stearns RE (2006) Complexity of reachability problems for finite discrete sequential dynamical systems. J Comput Syst Sci 72:1317–1345
11. Barrett CL, Hunt III HB, Marathe MV, Ravi SS, Rosenkrantz DJ, Stearns RE, Thakur M (2007) Computational aspects of analyzing social network dynamics. In: Proc International Joint Conference on Artificial Intelligence IJCAI 2007. pp 2268–2273
12. Barrett CL, Hunt III HB, Marathe MV, Ravi SS, Rosenkrantz DJ, Stearns RE, Thakur M (2007) Predecessor existence problems for finite discrete dynamical systems. Theor Comput Sci 1-2: 3–37
13. Bartlett R, Garzon M (1993) Monomial cellular automata. Complex Syst 7(5):367–388
14. Bernaschi M, Castiglione F (2002) Selection of escape mutants from immune recognition during hiv infection. Immunol Cell Biol 80:307–313
15. Bernaschi M, Succi S, Castiglione F (2000) Large-scale cellular automata simulations of the immune system response. Phys Rev E 61:1851–1854
16. Booch G, Rumbaugh J, Jacobson I (2005) Unified Modeling Language User Guide, 2nd edn. Addison-Wesley, Reading
17. Brand D, Zafiropulo P (1983) On communicating finite-state machines. J ACM 30:323–342
18. Cartier P, Foata D (1969) Problemes combinatoires de commutation et reárrangements. Lecture Notes in Mathematics, vol 85. Springer, Berlin
19. Castiglione F, Agur Z (2003) Analyzing hypersensitivity to chemotherapy in a cellular automata model of the immune system. In: Preziosi L (ed) Cancer modeling and simulation. Chapman and Hall/CRC, London
20. Castiglione F, Bernaschi M, Succi S (1997) Simulating the Immune Response on a Distributed Parallel Computer. Int J Mod Phys C 8:527–545; 10.1142/S0129183197000424
21. Castiglione F, Duca K, Jarrah A, Laubenbacher R, Hochberg D, Thorley-Lawson D (2007) Simulating Epstein–Barr virus infection with C-ImmSim. Bioinformatics 23(11):1371–1377
22. Celada F, Seiden P (1992) A computer model of cellular interactions in the immune syste. Immunol Today 13(2):56–62
23. Celada F, Seiden P (1992) A model for simulating cognate recognition and response in the immune system. J Theor Biol 158:235–270
24. Celada F, Seiden P (1996) Affinity maturation and hypermutation in a simulation of the humoral immune response. Eur J Immunol 26(6):1350–1358
25. Chaudhuri PP (1997) Additive Cellular Automata. Theory and Applications, vol 1. IEEE Comput Soc Press
26. Colón-Reyes O, Laubenbacher R, Pareigis B (2004) Boolean monomial dynamical systems. Ann Comb 8:425–439
27. Colón-Reyes O, Jarrah A, Laubenbacher R, Sturmfels B (2006) Monomial dynamical systems over finite fields. Complex Syst 16(4):333–342
28. Dawson D (1974) Synchronous and asynchronous reversible Markov systems. Canad Math Bull 17(5):633–649
29. Ebeling W, Schweitzer F (2001) Swarms of particle agents with harmonic interactions. Theor Biosci 120–3/4:207–224
30. Elspas B (1959) The theory of autonomous linear sequential networks. IRE Trans Circuit Theor, pp 45–60
31. Farmer J, Packard N, Perelson A (1986) The immune system, adaptation, and machine learning. Phys D 2(1–3):187–204
32. Frish U, Hasslacher B, Pomeau Y (1986) Lattice-gas automata for the Navier–Stokes equations. Phys Rev Lett 56:1505–1508
33. Fukś H (2004) Probabilistic cellular automata with conserved quantities. Nonlinearity 17:159–173
34. Garcia LD, Jarrah AS, Laubenbacher R (2006) Sequential dynamical systems over words. Appl Math Comput 174(1): 500–510
35. Gardner M (1970) The fantastic combinations of John Conway's new solitaire game "life". Sci Am 223:120–123
36. Gouda M, Chang C (1986) Proving liveness for networks of communicating finite-state machines. ACM Trans Program Lang Syst 8:154–182
37. Guo Y, Gong W, Towsley D (2000) Time-stepped hybrid simulation (tshs) for large scale networks. In: INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proc. IEEE, vol 2. pp 441–450
38. Gupta A, Katiyar V (2005) Analyses of shock waves and jams in traffic flow. J Phys A 38:4069–4083
39. Hansson AÅ, Mortveit HS, Reidys CM (2005) On asynchronous cellular automata. Adv Complex Syst 8(4):521–538
40. Hedlund G (1969) Endomorphisms and automorphisms of the shift dynamical system. Math Syst Theory 3:320–375
41. Hernández-Toledo A (2005) Linear finite dynamical systems. Commun Algebra 33(9):2977–2989
42. Hopfield J (1982) Neural networks and physical systems with emergent collective computational properties. Proc National Academy of Sciences of the USA 79:2554–2588
43. Ilachinsky A (2001) Cellular Automata: A Discrete Universe. World Scientific, Singapore
44. Jarrah A, Laubenbacher R, Stillman M, Vera-Licona P (2007) An efficient algorithm for the phase space structure of linear dynamical systems over finite fields. (submitted)
45. Jefferson DR (1985) Virtual time. ACM Trans Program Lang Syst 7(3):404–425
46. Kari J (2005) Theory of cellular automata: A survey. Theor Comput Sci 334:3–33
47. Keyfitz BL (2004) Hold that light! modeling of traffic flow by differential equations. Stud Math Libr 26:127–153
48. Laubenbacher R, Pareigis B (2003) Decomposition and simulation of sequential dynamical systems. Adv Appl Math 30: 655–678
49. Lidl R, Niederreiter H (1997) Finite Fields. Cambridge University Press, Cambridge
50. Liggett TM (2005) Interacting particle systems. Classics in Mathematics. Springer, Berlin. Reprint of the 1985 original
51. Lind DA (1984) Applications of ergodic theory and sofic systems to cellular automata. Physica D 10D:36–44
52. Lindgren K, Moore C, Nordahl M (1998) Complexity of two-dimensional patterns. J Statistical Phys 91(5–6):909–951
53. Mac Lane S (1998) Category Theory for the Working Mathematician, 2nd edn. No 5. in GTM. Springer, New York
54. Macy MW, Kitts JA, Flache A (2003) Polarization in Dynamic Networks: A Hopfield Model of Emergent Structure. In: Dynamic social network modeling and analysis. The National Academies Press, Washington D.C., pp 162–173

55. Martin O, Odlyzko A, Wolfram S (1984) Algebraic properties of cellular automata. Commun Math Phys 93:219–258
56. Milligan D, Wilson M (1993) The behavior of affine boolean sequential networks. Connect Sci 5(2):153–167
57. Minar N, Burkhart R, Langton C, Manor A (1996) The swarm simulation system: A toolkit for building multi-agent simulations. Santa Fe Institute preprint series. http://www.santafe.edu/research/publications/wpabstract/199606042 Accessed 11 Aug 2008
58. Misra J (1986) Distributed discrete-event simulation. ACM Comput Surv 18(1):39–65
59. Moncion T, Hutzler G, Amar P (2006) Verification of biochemical agent-based models using petri nets. In: Robert T (ed) International Symposium on Agent Based Modeling and Simulation, ABModSim'06, Austrian Society for Cybernetics Studies, pp 695–700; http://www.ibisc.univ-evry.fr/pub/basilic/OUT/Publications/2006/MHA06
60. Morpurgo D, Serentha R, Seiden P, Celada F (1995) Modelling thymic functions in a cellular automaton. Int Immunol 7: 505–516
61. Mortveit HS, Reidys CM (2001) Discrete, sequential dynamical systems. Discret Math 226:281–295
62. Mortveit HS, Reidys CM (2004) Reduction of discrete dynamical systems over graphs. Adv Complex Syst 7(1):1–20
63. Nagel K, Schreckenberg M (1992) A cellular automaton model for freeway traffic. J Phys I 2:2221–2229
64. Nagel K, Wagner P (2006) Traffic Flow: Approaches to Modelling and Control. Wiley
65. Nagel K, Schreckenberg M, Schadschneider A, Ito N (1995) Discrete stochastic models for traffic flow. Phys Rev E 51: 2939–2949
66. Nagel K, Rickert M, Barrett CL (1997) Large-scale traffic simulation. Lecture notes in computer science, vol 1215. Springer, Berlin, pp 380–402
67. Nance RE (1993) A history of discrete event simulation programming languages. ACM SIGPLAN Notices 28:149–175
68. North MJ, Collier NT, Vos JR (2006) Experiences creating three implementations of the repast agent modeling toolkit. ACM Trans Modeling Comput Simulation 16:1–25
69. Orponen P (1994) Computational complexity of neural networks: A survey. Nordic J Comput 1:94–110
70. Orponen P (1996) The computational power of discrete hopfield networks with hidden units. Neural Comput 8:403–415
71. Park JK, Steiglitz K, Thruston WP (1986) Soliton-like behavior in automata. Physica D 19D:423–432
72. Reidys C (1998) Acyclic Orientations of Random Graphs. Adv Appl Math 21:181–192
73. Reidys CM (2001) On acyclic orientations and sequential dynamical systems. Adv Appl Math 27:790–804
74. Reidys CM (2005) On Certain Morphisms of Sequential Dynamical Systems. Discret Math 296(2–3):245–257
75. Reidys CM (2006) Sequential dynamical systems over words. Ann Combinatorics 10(4):481–498
76. Rickert M, Nagel K, Schreckenberg M, Latour A (1996) Two lane traffic simulations using cellular automata. Physica A 231: 534–550
77. Rothman DH (1988) Cellular-automaton fluids: A model for flow in porous media. Geophysics 53:509–518
78. Russell S, Norwig P (2003) Artificial Intelligence: A Modern Approach. Prentice-Hall, Upper Saddle River
79. Schönfisch B, de Roos A (1999) Synchronous and asynchronous updating in cellular automata. BioSystems 51: 123–143
80. Shmulevich I, Dougherty ER, Kim S, Zhang W (2002) Probabilistic boolean networks: A rule-based uncertainty model for gene regulatory networks. Bioinformatics 18(2):261–274
81. Shmulevich I, Dougherty ER, Zhang W (2002) From boolean to probabilistic boolean networks as models of genetic regulatory networks. Proc IEEE 90(11):1778–1792
82. Sipser M (1997) Introduction to the Theory of Computation. PWS Publishing Co, Boston
83. Vasershtein L (1969) Markov processes over denumerable products of spaces describing large system of automata. Probl Peredachi Informatsii 5(3):64–72
84. von Neumann J, Burks AW (ed) (1966) Theory of Self-Reproducing Automata. University of Illinois Press, Champaign
85. Whitham G (1999) Linear and Nonlinear Waves, reprint edition edn. Pure and Applied Mathematics: A Wiley-Interscience Series of Texts, Monographs and Tracts, Wiley-Interscience, New York
86. Wolfram S (1983) Statistical mechanics of cellular automata. Rev Mod Phys 55:601–644
87. Wolfram S (1986) Theory and Applications of Cellular Automata, Advanced Series on Complex Systems, vol 1. World Scientific Publishing Company, Singapore
88. Wolfram S (2002) A New Kind of Science. Wolfram Media, Champaign

Books and Reviews

Hopcroft JE, Motwani R, Ullman JD (2000) Automata Theory, Languages and Computation. Addison Wesley, Reading
Kozen DC (1997) Automata and Computability. Springer, New York
Wooldridge M (2002) Introduction to Multiagent Systems. Wiley, Chichester

Agent Based Modeling and Neoclassical Economics: A Critical Perspective*

Scott Moss
Centre for Policy Modeling, Manchester Metropolitan University Business School, Manchester, UK

Article Outline

Introduction
Economic Modeling Approaches
Methodological Issues
Conditions for Complexity
Complexity and the Role of Evidence
Future Directions
Bibliography

*The remarks about neoclassical economics are drawn from my inaugural lecture [28]

Introduction

Agent Based Modeling and Neoclassical Economics based modeling naturally generates complexity whereas neoclassical economics is incompatible in principle with complexity. The reasons that preclude complexity in neoclassical economic models also ensure that neoclassical economics cannot describe any society ever observed or that could ever be observed.

The meaning of complexity has been developed, mainly by physicists, to cover unpredictable, episodic volatility and also particular network topologies. In both cases there are nodes representing the components of a system and there are links among the components that can represent interactions amongst those components. Unpredictable, episodic volatility can result from particular forms of behavior by components and the interactions amongst those components. I am not aware of any investigations into relationships between that type of complexity and network topology.

The point to be made here is that the core assumptions *and* the methodology of conventional neoclassical economics preclude the emergence of episodic volatility and render social network topology inconsequential. When elaborated with heterogeneous agents, network topologies might have some effects on the outputs from computational neoclassical economic models – but, again, I am not aware of any systematic investigations into this possibility.

Economic Modeling Approaches

All definitions of complexity take for granted that there will be some specifications of individual components, that in general each component will interact with some other components and there will be some macro level phenomena that could not be described or understood except on the basis of the components and their interactions. The purpose of this section is to categorize the ways in which economists, agent based modelers and complexity scientists approach this micro-macro issue.

There are several strands in the economics and social sciences literatures for building macro analyzes explicitly on micro foundations. The main strands are computable general equilibrium (CGE), agent based computational economics (ACE), agent based social simulation (ABSS) and complexity science (CS) including econophysics and sociophysics. These strands are not all mutually exclusive although there are some conflicting elements among several of them.

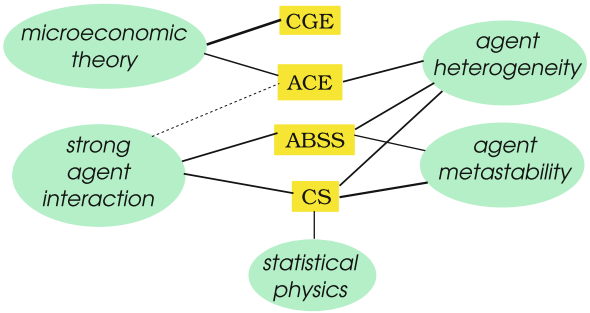

Agent Based Modeling and Neoclassical Economics: A Critical Perspective, Figure 1
Constraints on model designs

Computable General Equilibrium

CGE is the most theoretically constrained of the four strands under consideration. As with general equilibrium theory, it is predicated on the assumptions that households maximize utility and firms maximize profits and that markets clear. The computational load associated with explicit representation of every household and firm leads to the adoption of representative agents intended to capture the behavior of a group such as all households or firms in a particular industrial sector. Some CGE models represent technology with input-output tables; others with marginalist production functions.

Agent Based Computational Economics

An objection to the representative agent device is raised in the ACE literature where the effects of agent heterogeneity are explored. In these models, households can differ in their utility functions (or at least the parameters of those functions) or agents can adopt different game theoretic strategies and firms can employ different production functions. The theoretical core of ACE is not essentially different from that of CGE, both relying on conventional economic theory.

Agent Based Social Simulation

Models reported in the ABSS literature are by and large not driven by traditional theoretical concerns. There is a very wide range in the degree of empirical content: many models being developed to explore "stylized facts", others driven by qualitative case studies. The latter are often validated against both qualitative micro level data provided by stakeholders and against macro level statistical data.

Complexity Science

Because neoclassical economic theory excludes social embeddedness, the social complexity research that could be

relevant to a consideration of neoclassical economics must be concerned with unpredictable, episodic turbulence. The CS literature on financial markets was seminal and remains well known. The interest in financial markets goes back to Mandelbrot [25] who used financial market data both because it exhibits "outliers" in the relative change series and because of the fineness of the time grain of the price and volume data. A seminal article by Palmer et al. [35] reported a simulation model in which individuals were represented by an early form of software agent and which produced time series marked by the occasional episodes of volatility of the sort observed in real financial market data. However, similar unpredictable episodes of turbulence and volatility have emerged in models of the early post-Soviet Russian economy [31], domestic water consumption [8,15] and models of transactions in intermediated markets [29]. Fine grain data exhibiting the same patterns of volatility were found *subsequent to the original publication* of each model.

Relationships Among the Four Approaches

The common thread between CGE and ACE is their common reliance on longstanding economic concepts of utility, continuous production functions, profit maximization, the use of game theory et sic hoc omnes. The common thread between ABSS and complexity science is the importance of social interaction and specifications of individual behavior that are either more ad hoc or based on detailed qualitative evidence for specific cases.

Complex, as distinct from rational, agents' behavior is "sticky": it takes non-trivial events or social pressures to make them change. This is the social meaning of metastability. They are also socially embedded in the sense that they interact densely with other agents and are influenced by some of those other agents, generally speaking others who are most like themselves and who they have reason to like and respect [10]. The social difference between influence and imitation is the social equivalent of the physical difference between dissipative and non-dissipative processes. Of course, such influence is meaningless unless the agents differ in some way – they must be heterogeneous.

Methodological Issues

Neoclassical economic theory has no empirically based micro foundation. It has agents of two types: households that maximize utility and firms that maximize profits. Time and expectations are allowed to influence these maximization processes by substituting "expected utility" or "expected profits" for the realized magnitudes. In such models, agents (households or firms) act as if they know with certainty a population distribution of possible outcomes from their actions. In the terminology introduced by Knight [23], risk pertains when the agent knows the *frequency* distribution of past outcomes that, as in actuarial contexts, are expected with confidence to pertain in the future. When no such frequency distribution is known, then uncertainty prevails. In the sense of Knight (though the terminology gets muddled in the economics literature), the assumption that agents maximize expected utility or expected profits is tenable in conditions of risk but not in conditions of uncertainty. Moreover, it has long been known (with Nobel prizes awarded to Allais [4] and to Daniel Kahneman of Kahneman and Tversky [21] for the demonstrations) that individuals do not act as if they were maximizers of utility or expected utility. Nor is there any evidence that enterprises actually maximize profits. Many economists acknowledge that rationality is bounded and that we lack the cognitive capacity to absorb the required amount of information and then to process that information in order to identify some optimal decision or action. This has given rise to a variety of schools of economic thought such as evolutionary economics (Nelson and Winter [32] is the seminal work here) and Keynesian economics [22] being amongst the best known.

There is evidently a recognizable (and often recognized) divide between the behavioral assumptions of neoclassical economics on the one hand and, on the other hand, common observation, experimental observation (cf. [5]) and a host of business histories (the work of Chandler [12,13] and Penrose [37] being surely the most influential). The evidence shows that the assumptions of neoclassical economics are inaccurate descriptions of the behavior the theories and models are purported to represent. There are two classes of defense for these descriptively inaccurate assumptions. On is the *as-if* defense and the other is the *for-simplicity* defense. These are considered in turn.

The as-if defense was enunciated in several forms by Samuelson [39], Friedman [17] and Alchian [3] in the years around 1950. The details of the differences between Samuelson and Friedman are not germane here. Both argued that their purpose was to model aggregate economic entities such as markets or national economies and descriptively inaccurate assumptions at micro level are permissible provided that the models are descriptively accurate at macro level. Alchian's contribution was to propose a mechanism. He asserted that, at least in the case of firms, those that were more profitably would be more likely to survive than firms that were less profitable. Consequently, over time, more and more firms would approach more closely to the maximum of profits available to them so that, even if they did not actually seek to maximize profits, the

surviving population of firms would be those that implicitly did actually maximize profits.

The as-if defense is in practice an equilibrium argument. Neoclassical economic models are solved for the simultaneous maximization of utility and profits by all agents – or, at least it is proved that such a solution exists. In such a configuration, no agent has any incentive to change its behavior so the equilibrium presumed to be stable in the small (that is, once reached it is maintained). There is no proof that any general equilibrium model with an arbitrary number of agents is stable in the large (that is, that any feasible solution is an attractor of the system as a whole). The Alchian form of the as-if defense does not take into account any effects of agent interaction or influence of any agent by any other agent. In empirical – that is to say, econometric – testing of neoclassical models, extreme events and observations are dismissed as outliers and removed from the data set being used for the testing or else their effect is encapsulated by specially constructed dummy variables.

The for-simplicity defense rests on the presumption that simpler models are always to be preferred to more complicated models and the achievement of simplicity justifies making assumptions about behavior and environment that are not justified by evidence. The author has for many years now justified this claim by choosing any arbitrary leading economics journal and searching the most recent issue for an assumption made "for simplicity". On every occasion, the assumption made "for simplicity" has turned out to be an assumption that changed the nature of an empirical problem being addressed so that it conformed to the requirements (such as convexity or absence of externalities) of the mathematical technique being applied to the analysis. Seven of the eleven papers in, at the time of writing, the most recent (November 2007) issue of the *Quarterly Journal of Economics* appealed to the value of simplicity or, in one case, tractability to justify assumptions or specifications that were not justified empirically. The direct quotations are:

> Tractability obviously dictated the use of a simple summary statistic of the distribution of legislator ideologies (see p. 1418 in [14]).
>
> The fact, established below, that deriving the properties of the seats-votes relationship requires consideration only of the properties of the univariate distribution of ζ as opposed to those of the bivariate distribution of σ and μ considerably simplifies the analysis (see p. 1480 in [9]).
>
> We make three key assumptions to simplify the analysis. First, we assume that all jobs last indefinitely once found (i. e., there is no subsequent job destruction). Second, anticipating our empirical findings, we assume that wages are exogenously fixed, eliminating reservation-wage choices. Third, we assume that utility is separable in consumption and search effort (see p. 1516 in [11]).
>
> A more conventional timing assumption in search models without savings is that search in period t leads to a job that begins in period $t+1$. Assuming that search in period t leads to a job in period t itself simplifies the analytic expressions ... (see p. 1517 in [11]).
>
> For simplicity, we'll assume that $\beta_{\text{SAT}}\left(X_{i,s}\right)$, $\beta_{\text{TEST}}\left(X_{i,s}\right)$ and $\beta_{\text{OTH}}\left(X_{i,s}\right)$ are linear in $X_{i,s}$ and can thus be written .... We will further assume that the random utility component is independent and identically distributed (i. i. d.) from a type 1 extreme value distribution (see p. 1616 in [19]).
>
> For simplicity, all households represent two-earner married couples of the same age (see p. 1683 in [33]).
>
> For simplicity, the model assumes that the highest 35 years of earnings correspond to the ages between 25 and 59 (see p. 1685 in [33]).
>
> We follow Auerbach and Kotlikoff (1987) by measuring efficiency gains from social security privatization using an LSRA that compensates households that would otherwise lose from reform. To be clear, the LSRA is not being proposed as an actual government institution. Instead, it is simply a hypothetical mechanism that allows us to measure the standard Hicksian efficiency gains in general equilibrium associated with privatization (see p. 1687 in [33]).
>
> Assume for simplicity that these batch sizes are fixed for each product class .... Given these fixed batch sizes for the two classes of product, the firm maximizes profits by deciding how many production runs ... [to] undertake ... (see pp. 1731–1732 in [7]).
>
> We adopt a number of simplifying assumptions to focus on the main implications of this framework. First, we assume that the relationship between the firm and each manager is short-term. Second, when $x_{i,k} = z_{i,k}$, the manager obtains a private benefit. We assume that managers are credit-constrained and cannot compensate principals for these private benefits and that these private benefits are sufficiently large so that it is not profitable for the principal to utilize incentive contracts to induce managers

> to take the right action. These assumptions imply that delegation will lead to the implementation of the action that is preferred by the manager . . . (see p. 1769 in [1]).

All but the first of these quotations are from theoretical papers and the "simplifications" enable the authors to produce equilibrium solutions to their models. No one has ever knowingly observed an equilibrium and in a world where not everything is convex due to (for example) economies of large scale production and where computational capacities limit cognitive abilities, in principle no equilibrium ever will be observed. Indeed, Radner [38] showed that a *necessary* condition for general equilibrium to exist is that all agents have unlimited computational capacities if trading takes place at a sequence of dates. In the core general equilibrium model, all transactions are agreed at a single moment for all time. The "simplifications" required to produce equilibrium models cannot therefore be justified on the basis of relevance to empirical observation. They also ensure that the models cannot capture complexity.

Conditions for Complexity

The social and behavioral conditions for complexity manifest as unpredictable, episodic volatility appears to be the following:

- Individuals behave in routine ways unless some nontrivial event or events or social pressure from other individuals induce them to change their behavior.
- Individuals interact with other individuals.
- Individuals influence but do not generally imitate one another.
- Interactions amongst individuals and individual behavior are not dominated by events that do not arise from that interaction and behavior.

These conditions were first noticed as a general phenomenon in physical models and articulated by Jensen [20] as metastability, dense interaction, dissipation, and coldness of the system, respectively. The phenomenon of unpredictable, clustered volatility in social models had previously been noticed as had its similarity to self organized criticality as described by statistical physicists starting with Bak et al. [6].

Complexity and Social Volatility

Volatility in social statistical time series and power law distributed cross sectional data have long been observed by statisticians and social scientists. Vilfredo Pareto [36] discovered that the personal distribution of income is power law distributed, a finding which has been replicated widely across countries and time. The same phenomenon is now well known to characterize word use [43] city sizes [44], firm sizes (including market shares) [41], distributions of links between internet sites [2] and a host of other cross sectional distributions. Where firm sizes and market shares are concerned, there have been strands in the industrial economics literature reporting models yielding that result. However, the observed results have not been explained by models in which households maximize utility and firms maximize profits. As Simon and Bonini [41] point out, some variant to Gibrat's Law (or the law of proportional effect), which states that the growth rate of individuals (say firms) is not correlated with individual size, will generate one highly skewed distribution or another and the particular distribution can be refined by an appropriate choice of the representation of the law.

These desired results also emerged from a series of models based on plausible or empirically based specifications of individual behavior and social interaction in agent based social simulation models. In capturing stakeholders' perceptions of the behavior and social interactions of relevant classes of individuals and also in relying on well validated propositions from social psychology and cognitive science, models were implemented that produced the sort of skewed distributions that we observe in practice. Episodic volatility followed from the metastability and social embeddedness of agents. In the nature of this process, most changes are relatively small in magnitude but a few changes are very large. This results in fat-tailed distributions of relative changes in variable values at macro level and also in cross sectional data as described by Gibrat's Law. In practice, the large relative changes tend to be bunched together in unpredictable episodes of volatility.

While these results arise naturally in evidence-driven ABSS models, they are not easily reconciled with neoclassical economic theory. As Krugman [24] (quoted by Eeckhout [16]) had it, "We have to say that the rank-size rule is a major embarrassment for economic theory: one of the strongest statistical relationships we know, lacking any clear basis in theory".

Complexity and Social Network Topology

Social network topologies can obviously have no meaning in a model comprised by representative agents. In ACE, ABSS and CS models, there will always be a network of social interaction. However, the nature of the interaction can be very different across the different approaches.

In ACE models, it is common to find social interaction taking the form of games – typically the Prisoners' Dilemma. A nice example of such a model is Tesfatsion's labor market model using McFadzean and Tesfatsions's [26,42] Trading Network Game. In Tesfatsions's labour market model, there is a fixed number of workers and a fixed number of employers identical in their total offers of labour and of employment, respectively. Each worker (resp. employer) ascribes a value of utility to an employment arrangement with any employer (resp. worker). The utility starts out at a some exogenous value and is then increased or reduced depending on the experience at each trading date. The experience is the combination of cooperation and defection by each party to the employment relation at each time step. The social network in this model

> … is represented in the form of a directed graph in which the vertices $V(E)$ of the graph represent the work suppliers and employers, the edges of the graph (directed arrows) represent work offers directed from work suppliers to employers, and the edge weight on any edge denotes the number of accepted work offers (contracts) between the work supplier and employer connected by the edge (see p. 431 in [42]).

The topology of this network depends on the outcomes of sequences of prisoners' dilemma games determining the utilities of workers and employers to one another. Every worker can see every employer and conversely so that the directed links between agents are limited by the number of work contracts into which each agent can engage. After some arbitrary number of time steps, the strategies of the agents are represented as genes and a genetic algorithm is applied so that, over a whole simulation, the elements of the most successful defect/cooperate strategies become more dominant. Since these strategies determine the outcomes of the prisoners' dilemma games, the social network continues to evolve with utility enhancing strategies becoming more dominant.

In a recent (at the time of writing) issue of The Journal of Economic Dynamics and Control, Page and Tassier [34] modeled the development of chain stores across markets. A firm was defined by its product. Each product was assigned an "intrinsic quality" represented by an integer drawn at random from a distribution $\theta\left(q\right)$, and a set of I "hedonic attributes" represented by I positive integers in a range from 0 to some arbitrary, user selected number A. Consumers are represented by utility functions that are positively related to "quality" and negatively related to the difference between some desired set of hedonic attributes and the hedonic attributes of the product. There are a number (set by the model user) of discrete markets. Page and Tassier then run a variety of simulations that allow for firms to replicate themselves across markets or, through lack of demand, to leave markets.

These two models seem to be representative of a wide class of ACE models. In the first place, agents are defined by utility functions or game theoretic strategies so that the behavior of any individual agent is either fixed or responds smoothly to infinitesimal changes in prices, incomes or whatever other arguments might populate its utility function. In either event, agents cannot be metastable and follow behavioral routines until (but only until) some significant stimulus causes them to change their behavioral responses. In the second place, agents' preferences and responses are not influenced by the preferences or actions of any other agents like themselves. That is, their behavior as determined by their utility functions or game theoretic strategies will respond to market signals or the actions of the other agent in their game but not to communications with or observations of any other agents. These agents are not, in the words of Granovetter [18], socially embedded especially since it is rare in a neoclassical model for there to be more than two players in a game and unheard-of for there to be more than three (cf. [27]).

Whilst we cannot state with authority that the conditions of metastability, social influence and the consistency principle are necessary for complexity to emerge at macro level from micro level behavior, these conditions have characterized the social simulation models that have produced the episodic and unpredictable volatility associated with complexity. The absence of social embeddedness in the neoclassical ACE models must also explain their lack of any representation of social (as distinct from merely economic) networks.

Complexity and the Role of Evidence

An interesting and fairly typical feature of papers reporting neoclassical models – both theoretical and computational with agents – is that they motivate the modeling exercise by appeal to some empirical, macro level economic phenomenon and then ignore evidence about the micro level behavior that might bring about such phenomena. This practice can be seen in both of the ACE examples described in Sect. "Conditions for Complexity".

Tesfatsion [42] motivates her model on more theoretical grounds than do Page and Tassier [34]. She wrote:

> Understanding the relationship between market structure, market behavior, and market power in

> markets with multiple agents engaged in repeated strategic interactions has been a major focus of analytical, empirical, and human-subject experimental researchers in industrial organization since the early 1970s. To date, however, definitive conclusions have been difficult to obtain.

She goes on to cite "a unified theoretical treatment of oligopoly decision-making", an article on empirical findings with respect to market power that looks only at industry level statistical measures, and some work with experimental subjects. No references are made, either by Tesfatsion or those she cites, to any case studies of the "repeated strategic interactions" in which the "multiple agents" engage.

Page and Tassier give more historical detail. Their motivation turns on:

> Chain stores and franchises dominate the American economic landscape. A drive through any moderately sized city reveals remarkable conformity in restaurants, stores, and service centers. Anyone who so desires can eat at Applebee's, shop at Wal-Mart, and grab a Starbuck's latte grande while getting her car brakes done at Midas (see p. 3428 in [34]).
>
> For example, in many markets, Lowe's and Home Depot capture a significant portion of the home improvement market. These big box stores drove many small independent hardware stores and lumber yards out of business. The residual demand resides in niches that can be filled by hardware chains specializing in upscale home furnishings like Restoration Hardware. … Often, when Lowe's enters a market, it creates a niche for Restoration Hardware as well. And, as both Lowe's and Restoration Hardware enter more and more markets, they in turn create additional common niches that can be filled by even more chains. Thus, chains beget chains (see p. 3429 in [34]).

So this article claims a clear and direct historical basis. And yet

> … To capture the increasing correlation in niches formally, we introduce two new concepts, the niche landscape and the differential niche landscape. The former plots the quality required to enter a market at a given set of hedonic attributes. The latter plots the differences in two niche landscapes. In the presence of chains, differential niche landscapes become flat, i. e. the niche landscapes become correlated across markets (see p. 3429 in [34]).

The representation of the actors in this framework has been discussed above. At no stage is the agent design discussed in relation to any empirical accounts of the behavior and motivations of the managers of Wal-Mart, Starbucks, Lowes, Restoration Hardware or any other enterprise or any consumer.

This is, of course, the way of neoclassical economics and it has extended to ACE research as well. What is perhaps more unsettling is that it has also extended to the bastions of complexity science – the econophysicists.

There is a long literature now on complexity and financial markets and also on complexity an the formation of opinions – opinion dynamics. There are at least two reasons for the popularity amongst physicists of financial market modeling. First, there are long series of very fine grain data. Second, the data exhibits the unpredictable, episodic volatility associated with complexity. The popularity of opinion dynamics cannot be based on the quality of the data – even at macro level – because that quality is much more coarse grain and inconsistent over time than financial market data. Nonetheless, the two literatures are marked by the heavy presence and influence of physicists and by the lack of connection between their agent designs and any available evidence about the behavior of traders in financial markets or voters or others acting on or expressing their opinions.

A good example from the opinion dynamics literature – chosen at random from *The European Physical Journal B* – is by Schweitzer and Hoyst [40], "Modeling collective opinion formation by means of active Brownian particles". The motivation for their article is

> The formation of public opinion is among the challenging problems in social science, because it reveals a complex dynamics, which may depend on different internal and external influences. We mention the influence of political leaders, the biasing effect of mass media, as well as individual features, such as persuasion or support for other opinions.

We immediately have complexity and social relevance to motivate an article on social dynamics in a physics journal. However, there is no empirical justification for modeling individuals who form opinions as active Brownian particles. The apparent complexity in the outcomes of social processes of opinion formation can be produced by the non-linear feedbacks of fields of active Brownian particles. Whether individuals actually behave in this way is not addressed by Schweitzer and Hoyst or, as far as I know, by any contributor to the opinion dynamics literature.

Much the same can be said of the econophysics literature on financial markets. The clustered volatility associated with complexity is readily produced by physical models with characteristics of metastability, dissipation and dense patterns of interaction. What the econophysicists fail to address is the question of whether their particular formulations – and active Brownian particle is just one of many examples – are *descriptively accurate* representations of the individual actors whose behavior they are seeking to analyze.

In this regard, the econophysicists are not better scientists than neoclassical economists. It can be said in favor of neoclassical (including ACE) economists that they are at least following in a long tradition when they ignore the relationship between what people actually do and how agents are modeled. In the long history of the physical sciences, however, observation and evidence at micro and macro level and all levels in between has dominated theory (cf. [30]). There are some at least in the ABSS research community who would prefer our colleagues with backgrounds in the physical scientists to follow their own methodological tradition in this regard and not that of the neoclassical economists.

Future Directions

Complexity science is not a niche research interest in the social sciences. Societies are complex and all social science should be complexity science. However, any social science that excludes social interaction and inertia or routine necessarily suppresses complexity. As noted here, the adoption of utility theory and representative agents by neoclassical economists (and other social scientists influenced by them) amounts to the exclusion of behavioral inertia and social interaction, respectively. To drop both utility and representative agents and to build analyzes bottom up from a sound basis in evidence would produce a better – very likely, a good – body of economic analysis. But the transition from present convention would be enormous – a transition that experience shows to be beyond the capacity of current and previous generations of mainstream economists. Not only would they have to abandon theories that drive and constrain their research but also their whole epistemological and wider methodological stance. They would have to accept that prediction and forecasting cannot be core methodological objectives and that theories are built by abstracting from detailed evidence based social simulation models the designs and outputs from which have been validated by stakeholders in a range of contexts. This would be a future direction guided by good science.

Bibliography

1. Acemoglu D, Aghion P, Lelarge C, Van Reenen J, Zilibotti F (2007) Technology, information, and the decentralization of the firm. Q J Econ 122(4):1759–1799. http://www.mitpressjournals.org/doi/abs/10.1162/qjec.2007.122.4.1759
2. Adamic LA, Huberman BA (1999) Power-law distribution of the World Wide Web. Science 287(5461):2115
3. Alchian AA (1950) Uncertainty, evolution and economic theory. J Political Econ 58(2):211–221
4. Allais M (1953) Le comportement de l'homme rationnel devant le risque: Critiques des postulats et axiomes de l'ecole americaine. Econometrica 21(4):503–546
5. Anderson JR (1993) Rules of the mind. Lawrence Erlbaum Associates, Hillsdale
6. Bak P, Tang C, Weisenfeld K (1987) Self organized criticality: An explanation of 1/f noise. Phys Rev Lett 59(4):381–384
7. Bartel A, Ichniowski C, Shaw K (2007) How does information technology affect productivity? Plant-level comparisons of product innovation, process improvement, and worker skills. Q J Econ 122(4):1721–1758. http://www.mitpressjournals.org/doi/abs/10.1162/qjec.2007.122.4.1721
8. Barthelemy O (2006) Untangling scenario components with agent based modeling: An example of social simulations of water demand forecasts. Ph D thesis, Manchester Metropolitan University
9. Besley T, Preston I (2007) Electoral bias and policy choice: Theory and evidence. Q J Econ 122(4):1473–1510. http://www.mitpressjournals.org/doi/abs/10.1162/qjec.2007.122.4.1473
10. Brown R (1965) Social psychology. The Free Press, New York
11. Card D, Chetty R, Weber A (2007) Cash-on-hand and competing models of intertemporal behavior: New evidence from the labor market. Q J Econ 122(4):1511–1560. http://www.mitpressjournals.org/doi/abs/10.1162/qjec.2007.122.4.1511
12. Chandler AD (1962) Strategy and structure: Chapters in the history of the american industrial enterprise. MIT Press, Cambridge
13. Chandler AD (1977) The visible hand: The managerial revolution in american business. Harvard University Press
14. Coate S, Brian K (2007) Socially optimal districting: A theoretical and empirical exploration. Q J Econ 122(4):1409–1471
15. Downing TE, Moss S, Pahl Wostl C (2000) Understanding climate policy using participatory agent based social simulation. In: Moss S, Davidsson P (eds) Multi agent based social simulation. Lecture Notes in Artificial Intelligence, vol 1979. Springer, Berlin, pp 198–213
16. Eeckhout J (2004) Gibrat's law for (all) cities. Am Econ Rev 94(5):1429–1451
17. Friedman M (1953) The methodology of positive economics. In: Essays on Positive Economics. University of Chicago Press, Chicago
18. Granovetter M (1985) Economic action and social structure: The problem of embeddedness. Am J Sociol 91(3):481–510
19. Jacob BA, Lefgren L (2007) What do parents value in education? An empirical investigation of parents' revealed preferences for teachers. Q J Econ 122(4):1603–1637. http://www.mitpressjournals.org/doi/abs/10.1162/qjec.2007.122.4.1603
20. Jensen H (1998) Self-organized criticality: Emergent complex behavior in physical and biological systems. Cambridge University Press, Cambridge

21. Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. Econometrica 47(2):263–292
22. Keynes JM (1935) The general theory of employment, interest and money. Macmillan, London
23. Knight FH (1921) Risk, uncertainty and profit. Houghton-Mifflin
24. Krugman P (1995) Development, geography and economic theory. MIT Press, Cambridge
25. Mandelbrot B (1963) The variation of certain speculative prices. J Bus 36(4):394–419
26. McFadzean D, Tesfatsion L (1999) A c++ platform for the evolution of trade networks. Comput Econ 14:109–134
27. Moss S (2001) Game theory: Limitations and an alternative. J Artif Soc and Soc Simul 4(2)
28. Moss S (1999) Relevance, realism and rigour: A third way for social and economic research. Technical Report 99-56, Centre for Policy Modeling, Manchester Metropolitan University
29. Moss S (2002) Policy analysis from first principles. Proc US Nat Acad Sci 99(Suppl. 3):7267–7274
30. Moss S, Edmonds B (2005) Towards good social science. J Artif Soc and Soc Simul 8(4). ISSN 1460-7425. http://jasss.soc.surrey.ac.uk/8/4/13.html
31. Moss S, Kuznetsova O (1996) Modeling the process of market emergence. In: Owsinski JW, Nahorski Z (eds) Modeling and analyzing economies in transition. MODEST, Warsaw, pp 125–138
32. Nelson RR, Winter SG (1982) An evolutionary theory of economic change. Harvard University Press, Cambridge
33. Nishiyama S, Smetters K (2007) Does social security privatization produce efficiency gains? Q J Econ 122(4):1677–1719. http://www.mitpressjournals.org/doi/abs/10.1162/qjec.2007.122.4.1677
34. Page SE, Tassier T (2007) Why chains beget chains: An ecological model of firm entry and exit and the evolution of market similarity. J Econ Dyn Control 31(10):3427–3458
35. Palmer R, Arthur WB, Holland JH, LeBaron B, Taylor P (1993) Artificial economic life: A simple model for a stock market. Physica D 75:264–274
36. Pareto V (1896–1897) Cours d'économie politique professé à l'Université de Lausanne. Rouge, Lausanne
37. Penrose ET (1959) The theory of the growth of the firm. Wiley, New York
38. Radner R (1968) Competitive equilibrium under uncertainty. Econometrica 36(1):31–58
39. Samuelson PA (1949) Foundations of economic analysis. Harvard University Press
40. Schweitzer F, Hoyst JA (2000) Modeling collective opinion formation by means of active Brownian particles. Europ Phys J B – Condens Matter Complex Syst 15(4):723–732
41. Simon HA, Bonini CP (1958) The size distribution of business firms. Am Econ Rev 48(4):607–617. ISSN 0002-8282 http://links.jstor.org/sici?sici=0002-8282%28195809%2948%3A4%3C607%3ATS%DOBF%3E2.0.CO%3B2-3
42. Tesfatsion L (2001) Structure, behavior, and market power in an evolutionary labor market with adaptive search. J Econ Dyn Control 25(3–4):419–457
43. Zipf GK (1935) The psycho-biology of language. Houghton Mifflin, Boston
44. Zipf GK (1949) Human behavior and the principle of least effort. Addison-Wesley, Cambridge

Agent Based Modeling and Simulation

Stefania Bandini, Sara Manzoni, Giuseppe Vizzari
Complex Systems and Artificial Intelligence Research Center, University of Milan-Bicocca, Milan, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Agent-Based Models for Simulation
Platforms for Agent-Based Simulation
Future Directions
Bibliography

Glossary

Agent The definition of the term agent is controversial even inside the restricted community of computer scientists dealing with research on agent models and technologies [25]. A weak definition, that could be suited to describe the extremely heterogeneous approaches in the agent-based simulation context, is "an autonomous entity, having the ability to decide the actions to be carried out in the environment and interactions to be established with other agents, according to its perceptions and internal state".

Agent architecture The term agent *architecture* [53] refers to the internal structure that is responsible of effectively selecting the actions to be carried out, according to the perceptions and internal state of an agent. Different architectures have been proposed in order to obtain specific agent behaviors and they are generally classified into *deliberative* and *reactive* (respectively, *hysteretic* and *tropistic*, according to the classification reported in [29]).

Autonomy The term autonomy has different meanings, for it represents (in addition to the control of an agent over its own internal state) different aspects of the possibility of an agent to decide about its own actions. For instance, it may represent the possibility of an agent to decide (i) about the timing of an action, (ii) whether or not to fulfill a request, (iii) to act without the need of an external trigger event (also called pro-activeness or proactivity) or even (iv) basing on its personal experience instead of hard-wired knowledge [53]. It must be noted that different agent models do not generally embody all the above notions of autonomy.

Interaction "An interaction occurs when two or more agents are brought into a dynamic relationship through a set of reciprocal actions" [22].

Environment "The environment is a first-class abstraction that provides the surrounding conditions for agents to exist and that mediates both the interaction among agents and the access to resources" [66].

Platform for agent-based simulation a software framework specifically aimed at supporting the realization of agent-based simulation systems; this kind of framework often provides abstractions and mechanisms for the definition of agents and their environments, to support their interaction, but also additional functionalities like the management of the simulation (e. g. set-up, configuration, turn management), its visualization, monitoring and the acquisition of data about the simulated dynamics.

Definition of the Subject

Agent-Based Modeling and Simulation – an approach to the modeling and simulation of a system in which the overall behavior is determined by the local action and interaction of a set of agents situated in an environment. Every agent chooses the action to be carried out on the basis of its own behavioral specification, internal state and perception of the environment. The environment, besides enabling perceptions, can regulate agents' interactions and constraint their actions.

Introduction

Computer simulation represents a way to exploit a computational model to evaluate designs and plans without actually bringing them into existence in the real world (e. g. architectural designs, road networks and traffic lights), but also to evaluate theories and models of complex systems (e. g. biological or social systems) by envisioning the effect of the modeling choices, with the aim of gaining insight of their functioning. The use of these "synthetic environments" is sometimes necessary because the simulated system cannot actually be observed, since it is actually being designed or also for ethical or practical reasons. A general schema (based on several elaborations, such as those described in [19,31]) describing the role of simulation as a predictive or explanatory instrument is shown in Fig. 1.

Several situations are characterized by the presence of autonomous entities whose actions and interactions determine (in a non-trivial way) the evolution of the overall system. Agent-based models are particularly suited to represent these situations, and to support the study and analysis of topics like decentralized decision making, local-global interactions, self-organization, emergence, effects of heterogeneity in the simulated system. The interest in this relatively recent approach to modeling and simulation is demonstrated by the number of scientific events focused in this topic (see, to make some examples rooted in the computer science context, the Multi Agent Based Simulation workshop series [17,35,45,55,56,57], the IMA workshop on agent-based modeling[1] and the Agent-Based Modeling and Simulation symposium [7]). Agent-based models and multi-agent systems (MAS) have been adopted to simulate complex systems in very different contexts, ranging from social and economical simulation (see, e. g., [18]) to logistics optimization (see, e. g., [64]), from biological systems (see, e. g., [8]) to traffic (see, e. g., [4,12,60]) and crowd simulation (see, e. g., [11]).

This heterogeneity in the application domains also reflects the fact that, especially in this context of agent focused research, influences come from most different re-

[1] http://www.ima.umn.edu/complex/fall/agent.html

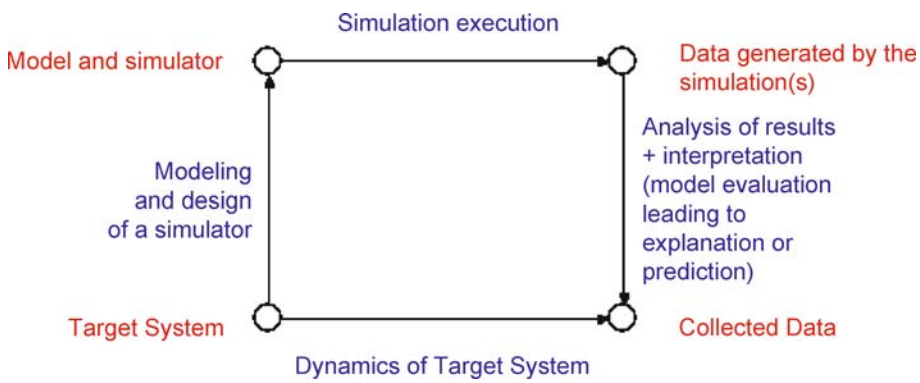

Agent Based Modeling and Simulation, Figure 1
A general schema describing the usage of simulation as a predictive or explanatory instrument

search areas. Several *traffic and crowd* agent models, to make a relevant example, are deeply influenced by *physics*, and the related models provide agents that are modeled as particles subject to forces generated by the environment as well as by other agents (i. e. *active walker* models, such as [37]). Other approaches to crowd modeling and simulation build on experiences with *Cellular Automata* (CA) approaches (see, e. g., [54]) but provide a more clear separation between the environment and the entities that inhabit, act and interact in it (see, e. g., [6,38]). This line of research leads to the definition of models for situated MASs, a type of model that was also defined and successfully applied in the context of (reactive) robotics and control systems [61,62]. Models and simulators defined and developed in the context of social sciences [31] and economy [50] are instead based on different theories (often non-classical ones) of *human behavior* in order to gain further insight on it and help building and validating new theories.

The common standpoint of all the above-mentioned approaches and of many other ones that describe themselves as agent-based is the fact that the analytical unit of the system is represented by the individual agent, acting and interacting with other entities in a shared environment: the overall system dynamic is not defined in terms of a global function, but rather the result of individuals' actions and interactions. On the other hand, it must also be noted that in most of the introduced application domains, the environment plays a prominent role because:

- It deeply influences the behaviors of the simulated entities, in terms of perceptions and allowed actions for the agents.
- The aim of the simulation is to observe some aggregate level behavior (e. g. the density of a certain type of agent in an area of the environment, the average length of a given path for mobile agents, the generation of clusters of agents), that can actually only be observed in the environment.

Besides these common elements, the above introduced approaches often dramatically differ in the way agents are described, both in terms of properties and behavior. A similar consideration can be done for their environment.

Considering the above introduced considerations, the aim of this article is not to present a specific technical contribution but rather to introduce an abstract reference model that can be applied to analyze, describe and discuss different models, concrete simulation experiences, platforms that, from different points of view, legitimately claim to adopt an agent-based approach. The reference model is illustrated in Fig. 2.

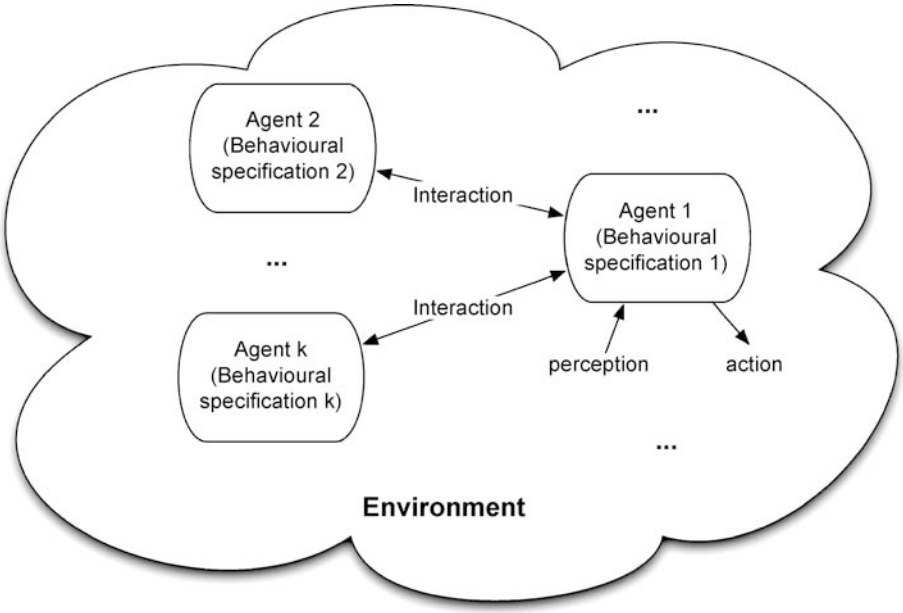

Agent Based Modeling and Simulation, Figure 2
An abstract reference model to analyze, describe and discuss different models, concrete simulation experiences, platforms legitimately claiming to adopt an agent-based approach

In particular, the following are the main elements of this reference model:

- *Agents* encompassing a possibly heterogeneous behavioral specification;
- Their *environment*, supplying agents their perceptions and enabling their actions;
- Mechanisms of *interaction* among agents.

Agent interaction, in fact, can be considered as a specific type of action having a central role in agent-based models. In the following Section, these elements will be analyzed, with reference to the literature on agent-based models and technologies but also considering the concrete experiences and results obtained by researches in other areas, from biology, to urban planning, to social sciences, in an attempt to enhance the mutual awareness of the points of contact of the respective efforts. Section "Platforms for Agent-Based Simulation" briefly discusses the available platforms supporting the rapid prototyping or development of agent-based simulations. A discussion on the future directions for this multi-disciplinary research area will end the article.

Agent-Based Models for Simulation

A model is an abstract and simplified representation of a given part of the world, either existing or planned (a target system, in the terminology adopted for Fig. 1). Models are commonly defined in order to study and explain observed phenomena or to forecast future phenomena. Agent-based models for simulation, as previously mentioned, are characterized by the presence of agents performing some kind of behavior in a shared environment. The notion of agent, however, is controversial even inside the restricted community of computer scientists dealing with research on agent models and technologies [25]. The most commonly adopted definition of agent [68] specifies a set of properties that must characterize an entity to effectively call it an agent, and in particular *autonomy* (the possibility to operate without intervention by humans, and a certain degree of control over its own state), *social ability* (the possibility to interact employing some kind of agent communication language, a notion that will be analyzed more in details in Subsect. "Agent Interaction"), *reactivity* (the possibility to perceive an environment in which it is situated and respond to perceived changes) and *pro-activeness* (the possibility to take the initiative, starting some activity according to internal goals rather than as a reaction to an external stimulus). This definition, considered by the authors a *weak* definition of agency, is generally already too restrictive to describe as agents most of the entities populating agent-based models for simulation in different fields. Even if the distance between the context of research on *intelligent* agents and agent-based simulation, that is often more focused on the resulting behavior of the local action and interaction of relatively simple agents, cannot be neglected, the aim of this section is to present some relevant results of research on agent models and technologies in computer science and put them in relation with current research on agent-based simulation in other research areas.

As previously suggested, agent-based models (ABMs) can be considered models of complex systems and the ABM approach considers that simple and complex phenomena can be the result of interactions between autonomous and independent entities (i. e. agents) which operate within communities in accordance with different modes of interaction. Thus, agents and ABMs should not be considered simply as a technology [43,69] but also a modeling approach that can be exploited to represent some system properties that are not simply described as properties or functionalities of single system components but sometimes *emerge from collective behaviors* [22]. The study of such emerging behaviors is a typical activity in complex systems modeling [10] and agent-based models are growingly employed for the study of complex systems (see, e. g., [3,36,65] satellite workshops of the 2007 edition of the European Conference on Complex Systems).

Agent Behavior Specification

This section will first of all discuss the possible notions of *actions* available to an agent, then it will discuss the way an action actually chooses the actions to be carried out, introducing the notion of *agent architecture*.

Agent Actions – Actions are the elements at the basis of agent behavior. Agent actions can cause modifications in their environment or in other agents that constitutes the ABM. Different modeling solutions can be provided in order to describe agent actions: as *transformation of a global state*, as *response to influences*, as *computing processes*, as *local modification*, as *physical displacement*, and as *command* (more details about the reported methods to represent agent actions can be found in [22]).

- *Functional transformation of states* is based on concepts of states and state transformation and constitutes the most classical approach of artificial intelligence to action representation (mainly used in the planning and multi-agent planning contexts). According to this approach, agent actions are defined as operators whose effect is a change in the state of the world [24,29,30].

- Modeling *actions as local modification* provides an approach opposite to the one based on transformation of a global state. Agents perceive their local environment (the only part of the environment that they can access) and according to their perceptions, they modify their internal state. Automata networks [32] and cellular automata [67] are examples of this type of models. They are dynamical systems whose behavior is defined in terms of local relationships and local transformations; in turn, these changes influence the overall system state. Within the ABMs context cellular automata are often exploited to represent the dynamic behavior of agent environment [58] or to simulate population dynamics [21] in artificial life [1,41].
- Modeling *actions as response to influences* [23] extends the previous approach introducing elements to consider the effects of agent interactions and the simultaneous execution of actions in an ABM. Agent actions are conditioned and represent a reaction to other agent actions or to environment modifications.
- Agents can also be considered *as computational processes* (in the vein of the actor model [2]). A computing system, and thus a ABM, can be considered as a set of activities (i. e. processes) that are executed sequentially or in parallel. This modeling approach in particular focuses on single computing entities, their behaviors and interactions. The most studied and well known methods to represent processes are finite state automata, Petri nets and their variants [46].
- *Approaches derived from physics* can also be found in order to represent agent actions. In these cases, actions mainly concern movements and their applications are in the robotics contexts (i. e. reactive and situated agents) [42]. One of the most used notion derived from physics is the one of *field* (e. g. gravitational, electrical, magnetic). Agents are attracted or repelled by given objects or environment areas that emit fields to indicate their position.
- According to cybernetics and control system theory, actions can be represented *as system commands* that regulate and control agent behavior. In this way, actions are complex tasks that the agent executes in order to fulfill given goals and that take into account the environment reactions and correctives to previous actions.

Agent Architecture – The term *architecture* [53] refers to the model of agent internal structure that is responsible of effectively selecting the actions to be carried out, according to the perceptions and internal state of an agent. Different architectures have been proposed in order to obtain specific agent behaviors and they are generally classified into *deliberative* and *reactive* (respectively, *hysteretic* and *tropistic* according to the classification reported in [29]).

Reactive agents are elementary (and often memoryless) agents with a defined position in the environment. Reactive agents perform their actions as a consequence of the perception of stimuli coming either from other agents or from the environment; generally, the behavioral specification of this kind of agent is a set of condition-action rules, with the addition of a selection strategy for choosing an action to be carried out whenever more rules could be activated. In this case, the motivation for an action derives from a triggering event detected in the environment; these agents cannot be pro-active.

Deliberative or *cognitive* agents, instead, are characterized by a more complex action selection mechanism, and their behavior is based on so called mental states, on facts representing agent knowledge about the environment and, possibly, also on memories of past experiences. Deliberative agents, for every possible sequence of perceptions, try to select a sequence of actions, allowing them to achieve a given goal. Deliberative models, usually defined within the planning context, provide a symbolic and explicit representation of the world within agents and their decisions are based on logic reasoning and symbol manipulation. The BDI model (belief, desire, intention [51,52]) is perhaps the most widespread model for deliberative agents. The internal state of agents is composed of three "data structures" concerning agent beliefs, desires and intentions. Beliefs represent agent information about its surrounding world, desires are the agent goals, while intentions represent the desire an agent has effectively selected, that it has to some extend committed.

Hybrid architectures can also be defined by combining the previous ones. Agents can have a layered architecture, where deliberative layers are based on a symbolic representation of the surrounding world, generate plans and take decisions, while reactive layers perform actions as effect of the perception of external stimuli. Both vertical and horizontal architectures have been proposed in order to structure layers [13]. In horizontal architecture no priorities are associated to layers and the results of the different layers must be combined to obtain agent's behavior. When layers are instead arranged in a vertical structure, reactive layers have higher priority over deliberative ones, that are activated only when no reactive behavior is triggered by the perception of an external stimulus.

A MAS can be composed of cognitive agents (generally a relatively low number of deliberative agents), each one possessing its own knowledge-model determining its behavior and its interactions with other agents and the environment. By contrast, there could be MAS made only by

reactive agents. This type of system is based on the idea that it is not necessary for a single agent to be individually intelligent for the system to demonstrate complex (intelligent) behaviors. Systems of reactive agents are usually more robust and fault tolerant than other agent-based systems (e. g. an agent may be lost without any catastrophic effect for the system). Other benefits of reactive MAS include flexibility and adaptability in contrast to the inflexibility that sometimes characterizes systems of deliberative agents [14]. Finally, a system might also present an *heterogeneous* composition of reactive and deliberative agents.

Environment

Weyns et al. in [66] provide a definition of the notion of environment for MASs (and thus of an environment for an ABM), and also discuss the core responsibilities that can be ascribed to it. In particular, in the specific context of simulation the environment is typically responsible for the following:

- Reflecting/reifying/managing the structure of the physical/social arrangement of the overall system;
- Embedding, supporting regulated access to objects and parts of the system that are not modeled as agents;
- Supporting agent perception and situated action (it must be noted that agent interaction should be considered a particular kind of action);
- Maintain internal dynamics (e. g. spontaneous growth of resources, dissipation signals emitted by agents);
- Define/enforce rules.

In order to exemplify this schema, we will now consider agent-based models and simulators that are based on a physical approach; the latter generally consider agents as particles subject to and generating forces. In this case, the environment comprises laws regulating these influences and relevant elements of the simulated system that are not agents (e. g. point of reference that generate attraction/repulsion forces). It is the environment that determines the overall dynamics, combining the effects that influence each agent and applying them generally in discrete time steps. In this cycle, it captures all the above introduced responsibilities, and the role of agents is minimal (according to some definitions they should not be called agents at all), and running a simulation is essentially reduced to computing iteratively a set equations (see, e. g., [4,37]). In situated ABM approaches agents have a higher degree of autonomy and control over their actions, since they evaluate their perceptions and choose their actions according to their behavioral specification. The environment retains a very relevant role, since it provides agents with their perceptions that are generated according to the current structure of the system and to the arrangement of agents situated in it. Socioeconomic models and simulations provide various approaches to the representation of the simulated system, but are generally similar to situated ABMs.

It is now necessary to make a clarification on how the notion of environment in the context of MAS-based simulation can be turned into a software architecture. Klügl et al. [40] argue that the notion of environment in multi-agent simulation is actually made up of two conceptually different elements: the simulated environment and the simulation environment. The former is a part of the computational model that represents the reality or the abstraction that is the object of the simulation activity. The simulation environment, on the other hand, is a software infrastructure for executing the simulation. In this framework, to make an explicit decoupling between these levels is a prerequisite for good engineering practice. It must be noted that also a different work [33], non-specifically developed in the context of agent-based simulation, provided a model for the deployment environment, that is the specific part of the software infrastructure that manages the interactions among agents.

Another recent work is focused on clarifying the notion of ABM environment and describes a three layered model for situated ABM environments [63]. This work argues that environmental abstractions (as well as those related to agents) crosscut all the system levels, from application specific ones, to the execution platform, to the physical infrastructure. There are thus application specific aspects of agents' environment that must be supported by the software infrastructure supporting the execution of the ABM, and in particular the ABM framework (MAS framework in the figure). Figure 3 compares the two above-described schemas.

The fact that the environment actually crosscuts all system levels in a deployment model represents a problem making difficult the separation between simulated environment and simulation infrastructure. In fact, the modeling choices can have a deep influence on the design of the underlying ABM framework and, vice versa, design choices on the simulation infrastructure make it suitable for some ABM and environment models but not usable for other ones. As a result, general ABM framework supporting simulation actually exist, but they cannot offer a specific form of support to the modeler, although they can offer basic mechanisms and abstractions.

SeSAm, for instance, offers a general simulation infrastructure but relies on plugins [40]. Those plugins, for example, could be used to define and manage the spatial

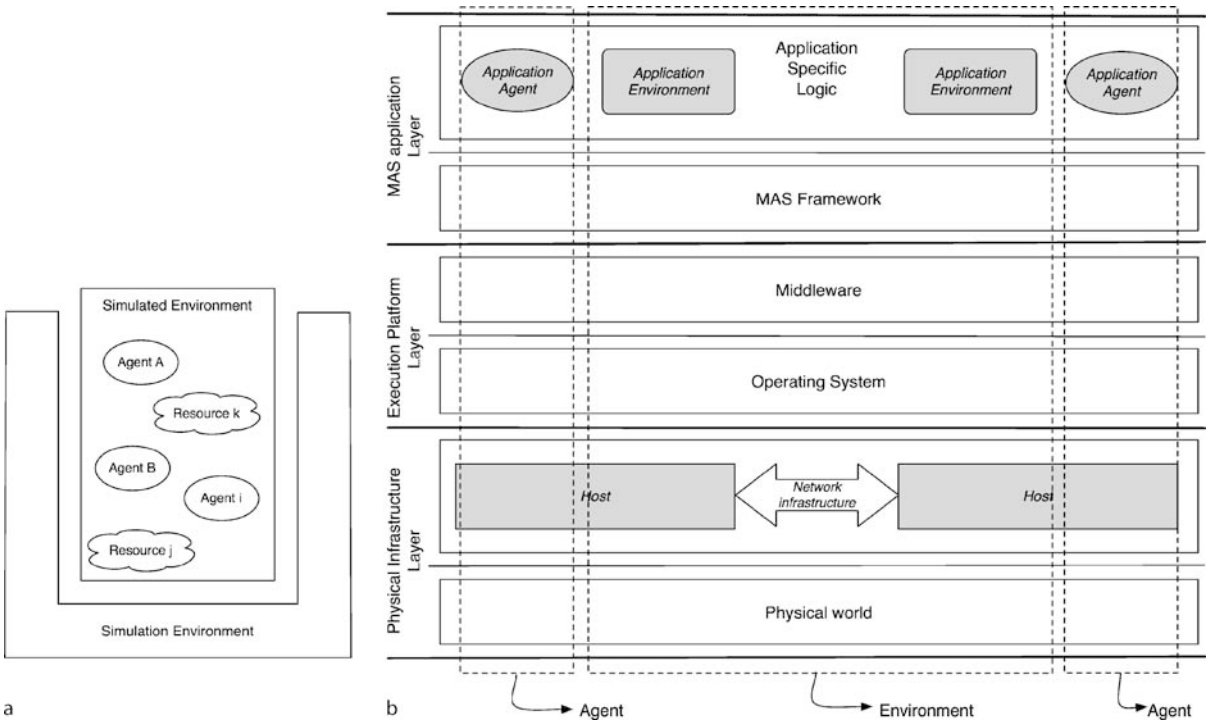

Agent Based Modeling and Simulation, Figure 3
A schema introduced in [40] to show differences and relationships between simulated and simulation environment (a), and a three layer deployment model for situated MAS introduced in [63] highlighting the crosscutting abstractions *agent* and *environment* (b)

features of the simulated environment, including the associated basic functions supporting agent movement and perception in that kind of environment. With reference to Fig. 3b, such a plugin would be associated to the application environment module, in the ABM application layer. However, these aspects represent just some of the features of the simulated environment, that can actually comprise rules and laws that extend their influence over the agents and the outcomes of their attempts to act in the environment.

Agent Interaction

Interaction is a key aspect in ABM. There is a plethora of definitions for the concept of agent and most of them emphasize the fact that this kind of entity should be able to interact with their environment and with other entities in order to solve problems or simply reach their goals according to coordination, cooperation or competition schemes. The essence of a ABM is the fact that the global system dynamics emerges from the local behaviors and interactions among its composing parts. Strictly speaking, for some kind of ABM the global dynamics is just the sum of local behaviors and interactions, so we cannot always speak of emergent behavior when we talk about a ABM. However the assumptions that underlie the design of an interaction model (or the choice of an existing one for the design and implementation of a specific application) are so important that they have a deep impact on the definition of agent themselves (e. g. an interpreter of a specific language, a perceiver of signals). Therefore it is almost an obvious consequence that interaction mechanisms have a huge impact on the modeling, design and development of applications based on a specific kind of ABM, which, in turn, is based on a particular interaction model. It is thus not a surprise that a significative part of the research that was carried out in the agent area was focused on this aspect.

This section presents a conceptual taxonomy of currently known/available agent interaction models, trying to define advantages and issues related to them, both from a conceptual and a technical point of view.

There are many possible dimensions and aspects of agent interaction models that can be chosen and adopted in order to define a possible taxonomy. The first aspect that is here considered to classify agent interaction models is related to the fact that agents communicate di-

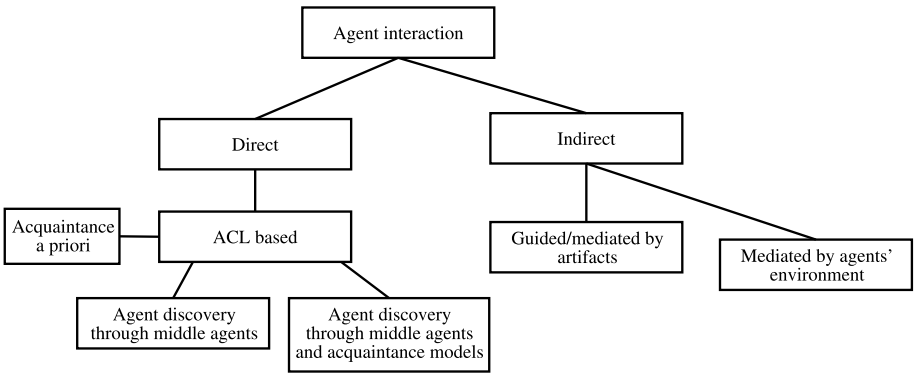

Agent Based Modeling and Simulation, Figure 4
The proposed taxonomy of agent interaction models

rectly (for instance exchanging messages), and thus the model does not include an abstraction of the actual communication channel, or there are some media interposed among the communication partners which is explicitly included in the interaction model. While the former approach, with specific reference to Agent Communication Language (ACL)-based models, is the most widely adopted in the agent area it has its drawbacks, and most of them are related to the issue of agents acquaintance. The way ACL-based agent interaction models deal with this issue is the subject of another dimension of the taxonomy, providing direct a priori acquaintance among agents, the adoption of middle agents for information discovery and the development of more complex acquaintance models to tackle issues related to the representation and maintenance of acquaintance information but also to robustness and scalability of the agent infrastructure. However there are other agent interaction models providing an indirect communication among them. Some of these approaches provide the creation and exploitation of artifacts that represent a medium for agents' interaction. Other indirect interaction models are more focused on modeling agent environment as the place where agent interactions take place, thus influencing interactions and agent behavior.

Direct Interaction Models The first and most widely adopted kind of agent interaction model provide a direct information exchange between communication partners. This approach ignores the details related to the communication channel that allows the interaction, and does not include it as an element of the abstract interaction model. Generally the related mechanisms provide a point-to-point message-passing protocol regulating the exchange of messages among agents. There are various aspects of the communicative act that must be modeled (ranging from low-level technical considerations on message format to conceptual issues related to the formal semantics of messages and conversations), but generally this approach provides the definition of suitable languages to cover these aspects. While this approach is generally well-understood and can be implemented in a very effective way (especially as it is substantially based on the vast experience of computer networks protocols), in the agent context, it requires specific architectural and conceptual solutions to tackle issues related to the agent acquaintance/discovery and ontological issues.

Intuitively an Agent Communication Language (ACL) provides agents with a means of exchanging information and knowledge. This vague definition inherently includes the point of view on the conception of the term agent, which assumes that an agent is an intelligent autonomous entity whose features include some sort of *social ability* [68]. According to some approaches this kind of feature is the one that ultimately defines the essence of agency [28]. Leaving aside the discussion on the definition and conception of agency, this section will focus on what the expression "social ability" effectively means. To do so we will briefly compare these ACL share with those approaches that allow the exchange of information among distributed components (e. g. in legacy systems[2]) some basic issues: in particular, the definition of a communication channel allowing the reliable exchange of messages over a computer network (i. e. the lower level aspects of the communication). What distinguishes ACLs from such systems are the objects of discourse and their semantic complexity; in particular there are two aspects which distributed computing protocols and architectures do not have to deal with:

[2]With this expression we mean pieces of software which are not designed to interact with agents and agent based systems.

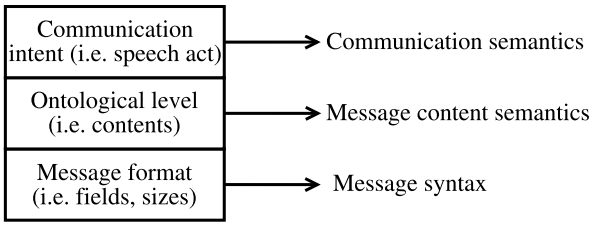

Agent Based Modeling and Simulation, Figure 5
Layers and concerns of an agent communication language

- *Autonomy* of interacting components: modern systems' components (even though they can be quite complex and can be considered as self-sufficient with reference to supplying a specific service) have a lower degree of autonomy than the one that is generally associated to agents.
- The information conveyed in messages does not generally require a comprehensive *ontological approach*, as structures and categories can be considered to be shared by system components.

Regarding the autonomy, while traditional software components offer services and generally perform the required actions as a reaction to the external requests, agents may decide not to carry out a task that was required by some other system entity. Moreover generally agents are considered temporally continuous and proactive, while this is not generally true for common software components.

For what concerns the second point, generally components have specific interfaces which assume an agreement on a set of shared data structures. The semantics of the related information, and the semantics of messages/method invocation/service requests, is generally given on some kind of (more or less formally specified) modeling language, but is tightly related to component implementation. For agent interaction a more explicit and comprehensive view on domain concepts must be specified. In order to be able to effectively exchange knowledge, agents must share an *ontology* (see, e. g., [34]), that is a representation of a set of categories of objects, concepts, entities, properties and relations among them. In other words, the same concept, object or entity must have a uniform meaning and set of properties across the whole system.

Indirect Interaction Models From a strictly technical point of view, agent communication is generally indirect even in direct agent interaction models. In fact most of these approaches adopt some kind of communication infrastructure supplying a reliable end-to-end message passing mechanism. Nonetheless, the adoption of a *conceptually* direct agent interaction model brings specific issues that were previously introduced. The remaining of this section will focus on models providing the presence of an intermediate entity mediating (allowing and regulating) agent interaction. This communication abstraction is not merely a low-level implementation detail, but a first-class concept of the model.

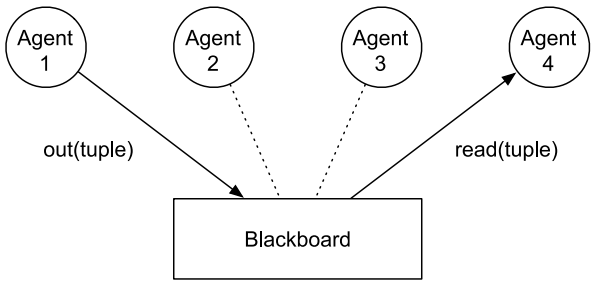

Agent Based Modeling and Simulation, Figure 6
A conceptual diagram for a typical blackboard architecture, including two sample primitives of the Linda coordination model, that is, the output of a tuple into the blackboard (the *out* operation) and the non-destructive input of an agent from the blackboard (the *read* operation)

Agent interaction models which provide indirect mechanisms of communication will be classified into *artifact mediated* and *spatially grounded* models. The distinction is based on the inspiration and metaphor on which these models are rooted. The former provide the design and implementation of an artifact which emulates concrete objects of agents' environment whose goal is the communication of autonomous entities. Spatially grounded agent interaction models bring the metaphor of modeling agent environment to the extreme, recognizing that there are situations in which spatial features and information represent a key factor and cannot be neglected in analyzing and modeling a system.

Both of these approaches provide interaction mechanisms that are deeply different from point-to-point message exchange among entities. In fact, the media which enable the interaction intrinsically represent a *context* influencing agent communication.

In the real world a number of physical agents interact sharing resources, by having a competitive access to them (e. g. cars in streets and crossroads), but also collaborating in order to perform tasks which could not be carried out by single entities alone, due to insufficient competencies or abilities (e. g. people that carry a very heavy burden together). Very often, in order to regulate the interactions related to these resources, we build concrete artifacts, such as traffic lights on the streets, or neatly placed

handles on large heavy boxes. Exploiting this metaphor, some approaches to agent interaction tend to model and implement abstractions allowing the cooperation of entities through a shared resource, whose access is regulated according to precisely defined rules. *Blackboard-based architectures* are the first examples of this kind of models. A blackboard is a shared data repository that enables cooperating software modules to communicate in an indirect and anonymous way [20]. In particular the concept of *tuple space*, first introduced in Linda [26], represents a pervasive modification to the basic blackboard model.

Linda [26] coordination language probably represents the most relevant blackboard-based model. It is based on the concept of tuple space, that is an associative blackboard allowing agents to share and retrieve data (i. e. tuples) through some data-matching mechanism (such as pattern-matching or unification) integrated within the blackboard. Linda also defines a very simple language defining mechanisms for accessing the tuple space.

The rationale of this approach is to keep separated computation and coordination contexts as much as possible [27], by providing specific abstractions for agent interaction. With respect to direct interaction models, part of the burden of coordination is in fact moved from the agent to the infrastructure. The evolution of this approach has basically followed two directions: the extension of the coordination language and infrastructure in order to increase its expressiveness or usability, and the modeling and implementation of distributed tuple spaces [16,48,49].

While the previously described indirect approaches define artifacts for agent interaction taking inspiration from actual concrete object of the real world, other approaches bring the metaphor of agent environment to the extreme by taking into account its spatial features.

In these approaches agents are situated in an environment whose spatial features are represented possibly in an explicit way and have an influence on their perception, interaction and thus on their behavior. The concept of perception, which is really abstract and metaphoric in direct interaction models and has little to do with the physical world (agents essentially perceive their state of mind, which includes the effect of received messages, like new facts in their knowledge base), here is related to a more direct modeling of what is often referred to as "local point of view". In fact these approaches provide the implementation of an infrastructure for agent communication which allows them to perceive the state of the environment in their position (and possibly in nearby locations). They can also cause local modifications to the state of the environment, generally through the emission of signals, emulating some kind of physical phenomenon

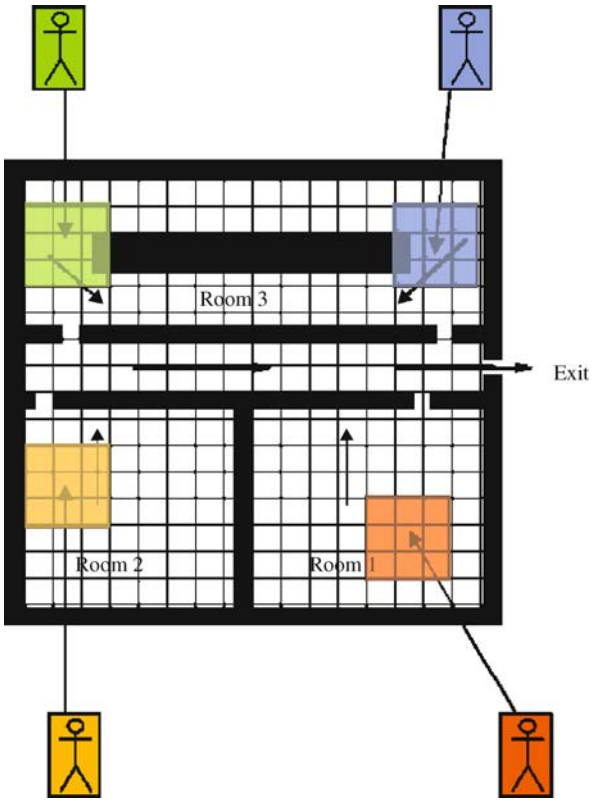

Agent Based Modeling and Simulation, Figure 7
A sample schema exemplifying an environment mediated form of interaction in which the spatial structure of the environment has a central role in determining agents' perceptions and their possibility to interact

(e. g. pheromones [15], or fields [5,44]) or also by simply observing the actions of other agents and reacting to this perception, in a "behavioral implicit communication" schema [59].

In all these cases, however, the structuring function of the environment is central, since it actually defines what can be perceived by an agent in its current position and how it can actually modify the environment, to which extent its actions can be noted by other agents and thus interact with them.

Platforms for Agent-Based Simulation

Considering the pervasive diffusion and adoption of agent-based approaches to modeling and simulation, it is not surprising the fact that there is a growing interest in software frameworks specifically aimed at supporting the realization of agent-based simulation systems.

This kind of framework often provides abstractions and mechanisms for the definition of agents and their en-

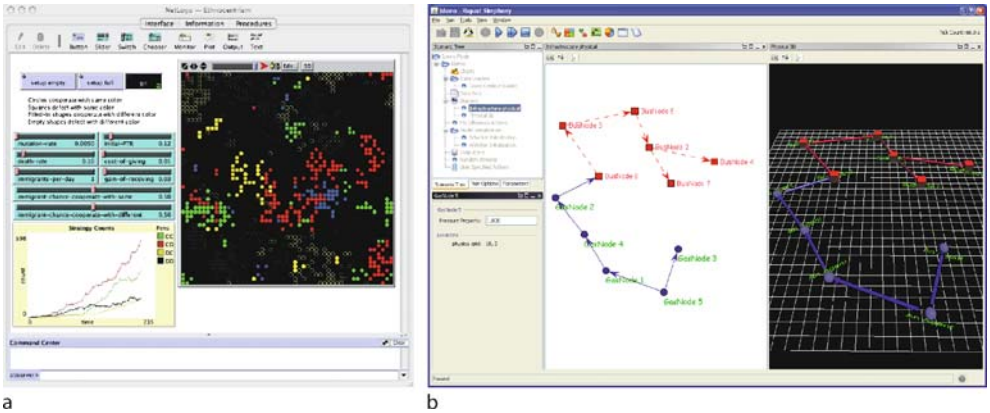

Agent Based Modeling and Simulation, Figure 8
A screenshot of a Netlogo simulation applet, (a), and a Repast simulation model, (b)

vironments, to support their interaction, but also additional functionalities like the management of the simulation (e. g. set-up, configuration, turn management), its visualization, monitoring and the acquisition of data about the simulated dynamics. It is not the aim of this article to provide a detailed review of the current state of the art in this sector, but rather sketch some classes of instruments that have been used to support the realization of agent-based simulations and provide a set of references to relevant examples of platforms facilitating the development of agent-based simulations.

A first category of these platforms instruments provides *general purpose frameworks* in which agents mainly represent *passive abstractions*, sort of data structures that are manipulated by an overall simulation process. A relevant example of such tools is **NetLogo**[3], a dialect of the Logo language specifically aimed at modeling phenomena characterized by a decentralized, interconnected nature. NetLogo does not even adopt the term agent to denote individuals, but it rather calls them *turtles*; a typical simulation consists in a cycle choosing and performing an action for every turtle, considering its current situation and state. It must be noted that, considering some of the previously mentioned definitions of autonomous agent a turtle should not be considered an agent, due to the almost absent autonomy of these entities. The choice of a very simple programming language that does not require a background on informatics, the possibility to deploy in a very simple way simulations as Java applets, and the availability of simple yet effective visualization tools, made NetLogo extremely popular.

A second category of platforms provides frameworks that are developed with a similar rationale, providing for very similar support tools, but these instruments are based on *general purpose programming languages* (generally object oriented). **Repast**[4] [47] represents a successful representative of this category, being a widely employed agent-based simulation platform based on the Java language. The object-oriented nature of the underlying programming language supports the definition of computational elements that make these agents more autonomous, closer to the common definitions of agents, supporting the encapsulation of state (and state manipulation mechanisms), actions and action choice mechanism in agent's class. The choice of adopting a general purpose programming language, on one hand, makes the adoption of these instruments harder for modelers without a background in informatics but, on the other, it simplifies the integration with external and existing libraries. Repast, in its current version, can be easily connected to instruments for statistical analysis, data visualization, reporting and also geographic information systems.

While the above-mentioned functionalities are surely important in simplifying the development of an effective simulator, and even if in principle it is possible to adapt frameworks belonging to the previously described categories, it must be noted that their neutrality with respect to the specific adopted agent model leads to a necessary preliminary phase of adaptation of the platform to the specific features of the model that is being implemented. If the latter defines specific abstractions and mechanisms for agents, their decision-making activities, their environ-

[3] http://ccl.northwestern.edu/netlogo/

[4] http://repast.sourceforge.net/

ment and the way they interact, then the modeler must in general develop proper computational supports to be able to fruitfully employ the platform. These platforms, in fact, are not endowed with specific supports to the realization of agent deliberation mechanisms or infrastructures for interaction models, either direct or indirect (even if it must be noted that all the above platforms generally provide some form of support to agent environment definition, such as grid-like or graph structures).

A third category of platforms represent an attempt to provide a *higher level linguistic support* trying to reduce the distance between agent-based models and their implementations. The latest version of Repast, for instance, is characterized by the presence of a high level interface for "point-and-click" definition of agent's behaviors, that is based on a set of primitives for the specification of agent's actions. **SimSesam**[5] [39] defines a set of *primitive functions* as basic elements for describing agents' behaviors, and it also provides visual tools supporting model implementation. At the extreme border of this category, we can mention efforts that are specifically aimed at supporting the development of simulations based on a precise agent model, approach and sometimes even for a specific area of application, such as the one described in [9,64].

Future Directions

Agent-based modeling and simulation is a relatively young yet already widely diffused approach to the analysis, modeling and simulation of complex systems. The heterogeneity of the approaches, modeling styles and applications that legitimately claim to be "agent-based", as well as the fact that different disciplines are involved in the related research efforts, they are all factors that hindered the definition of a generally recognized view of the field. A higher level framework of this kind of activity would be desirable in order to relate different efforts by means of a shared schema. Moreover it could represent the first step in effectively facing some of the epistemological issues related to this approach to the modeling and analysis of complex systems. The future directions in this broad research area are thus naturally aimed at obtaining vertical analytical results on specific application domains, but they must also include efforts aimed at "building bridges" between the single disciplinary results in the attempt to reach a more general and shared understanding of how these bottom-up modeling approaches can be effectively employed to study, explain and (maybe) predict the overall behavior of complex systems.

[5] http://www.simsesam.de/

Bibliography

1. Adami C (1998) Introduction to Artificial Life. Springer, New York
2. Agha G (1986) Actors: A Model of Concurrent Computation in Distributed Systems. MIT press, Cambridge
3. Alfi V, Galla T, Marsili M, Pietronero L (eds) (2007) Interacting Agents, Complexity and Inter-Disciplinary Applications (IACIA)
4. Balmer M, Nagel K (2006) Shape morphing of intersection layouts using curb side oriented driver simulation. In: van Leeuwen JP, Timmermans HJ (eds) Innovations in Design & Decision Support Systems in Architecture and Urban Planning. Springer, Dordrecht, pp 167–183
5. Bandini S, Manzoni S, Simone C (2002) Heterogeneous agents situated in heterogeneous spaces. Appl Artific Intell 16:831–852
6. Bandini S, Manzoni S, Vizzari G (2004) Situated cellular agents: a model to simulate crowding dynamics. IEICE Transactions on Information and Systems: Special Issues on Cellular Automata E87-D, pp 669–676
7. Bandini S, Petta P, Vizzari G (eds) (2006) International Symposium on Agent Based Modeling and Simulation (ABModSim 2006), vol Cybernetics and Systems. Austrian Society for Cybernetic Studies (2006) 18th European Meeting on Cybernetics and Systems Research (EMCSR 2006)
8. Bandini S, Celada F, Manzoni S, Puzone R, Vizzari G (2006) Modelling the immune system with situated agents. In: Apolloni B, Marinaro M, Nicosia G, Tagliaferri R (eds) Proceedings of WIRN/NAIS 2005. Lecture Notes in Computer Science, vol 3931. Springer, Berlin, pp 231–243
9. Bandini S, Federici ML, Vizzari G (2007) Situated cellular agents approach to crowd modeling and simulation. Cybernet Syst 38:729–753
10. Bar-Yam Y (1997) Dynamics of Complex Systems. Addison-Wesley, Reading
11. Batty M (2001) Agent based pedestrian modeling. Env Plan B: Plan Des 28:321–326
12. Bazzan ALC, Wahle J, Klügl F (1999) Agents in traffic modelling – from reactive to social behaviour. In: Burgard W, Christaller T, Cremers AB (eds) KI-99: Advances in Artificial Intelligence, 23rd Annual German Conference on Artificial Intelligence, Bonn, Germany, 13–15 September 1999. Lecture Notes in Computer Science, vol 1701. Springer, Berlin, pp 303–306
13. Brooks RA (1986) A robust layered control system for a mobile robot. IEEE J Robot Autom 2:14–23
14. Brooks RA (1990) Elephants don't play chess. Robot Autonom Syst 6:3–15
15. Brueckner S (2000) An analytic approach to pheromone-based coordination. In: ICMAS IEEE Comp Soc, pp 369–370
16. Cabri G, Leonardi L, Zambonelli F (2000) MARS: a programmable coordination architecture for mobile agents. IEEE Inter Comp 4:26–35
17. Davidsson P, Logan B, Takadama K (eds) (2005) Multi-Agent and Multi-Agent-Based Simulation, Joint Workshop MABS (2004), New York, 19 July 2004, Revised Selected Papers. In: Davidsson P, Logan B, Takadama K (eds) MABS. Lecture Notes in Computer Science, vol 3415. Springer, Berlin
18. Dosi G, Fagiolo G, Roventini A (2006) An evolutionary model of endogenous business cycles. Comput Econ 27:3–34
19. Edmonds B (2001) The use of models – making MABS more informative. In: Multi-Agent-Based Simulation, Second Interna-

tional Workshop MABS (2000), Boston MA, USA, July (2000), Revised and Additional Papers. Lecture Notes in Computer Science, vol 1979. Springer, Berlin, pp 15–32
20. Englemore RS, Morgan T (eds) (1988) Blackboard Systems. Addison-Wesley, Reading
21. Epstein JM, Axtell R (1996) Growing Artificial Societies. MIT Press, Boston
22. Ferber J (1999) Multi–Agent Systems. Addison-Wesley, London
23. Ferber J, Muller J (1996) Influences and reaction: A model of situated multiagent systems. In: Proceedings of the 2nd International Conference on Multiagent Systems
24. Fikes RE, Nilsson NJ (1971) STRIPS: a new approach to the application of theorem proving to problem solving. Artif Intell 2:189–208
25. Franklin S, Graesser A (1997) Is it an agent, or just a program?: A taxonomy for autonomous agents. In: Müller JP, Wooldridge M, Jennings NR (eds) Intelligent Agents III, Agent Theories, Architectures, and Languages ECAI '96 Workshop (ATAL), Budapest, 12–13 August 1996. Lecture Notes in Computer Science, vol 1193. Springer, Berlin, pp 21–36
26. Gelernter D (1985) Generative communication in Linda. ACM Trans Program Lang Syst 7:80–112
27. Gelernter D, Carriero N (1992) Coordination languages and their significance. Commun ACM 35:97–107
28. Genesereth MR, Ketchpel SP (1994) Software agents. Commun ACM 37(7):48–53
29. Genesereth MR, Nilsson N (1987) Logical Foundations of Artificial Intelligence. Morgan Kaufmann, San Mateo
30. Georgeff M (1984) A theory of action in multi–agent planning. In: Proceedings of the AAAI84, pp 121–125
31. Gilbert N, Troitzsch KG (2005) Simulation for the Social Scientist 2nd edn. Open University Press, Maidenhead
32. Goles E, Martinez S (1990) Neural and automata networks, dynamical behavior and applications. Kluwer, Norwell
33. Gouaich A, Michel F, Guiraud Y (2005) MIC: A deployment environment for autonomous agents. In: Environments for Multi-Agent Systems, First International Workshop (E4MAS 2004). Lecture Notes in Computer Science, vol 3374. Springer, Berlin, pp 109–126
34. Gruber TR (1995) Toward principles for the design of ontologies used for knowledge sharing. Int J Hum Comp Stud 43:907–928
35. Hales D, Edmonds B, Norling E, Rouchier J (eds) (2003) Multi-Agent-Based Simulation III, 4th International Workshop MABS (2003), Melbourne, 14 July 2003. Revised Papers. In: Hales D, Edmonds B, Norling E, Rouchier J (eds) MABS. Lecture Notes in Computer Science, vol 2927. Springer, Berlin
36. Hassas S, Serugendo GDM, Phan D (eds) (2007) Multi-Agents for modelling Complex Systems (MA4CS). http://bat710.univ-lyon1.fr/~farmetta/MA4CS07
37. Helbing D, Schweitzer F, Keltsch J, Molnár P (1997) Active walker model for the formation of human and animal trail systems. Phys Rev E 56:2527–2539
38. Henein CM, White T (2005) Agent-based modelling of forces in crowds. In: Davidsson P, Logan B, Takadama K (eds) Multi-Agent and Multi-Agent-Based Simulation, Joint Workshop MABS (2004), New York 19 July 2004. Revised Selected Papers. Lecture Notes in Computer Science, vol 3415. Springer, Berlin, pp 173–184
39. Klügl F, Herrler R, Oechslein C (2003) From simulated to real environments: How to use sesam for software development. In: Schillo M, Klusch M, Müller JP, Tianfield H (eds) MATES. Lecture Notes in Computer Science, vol 2831. Springer, Berlin, pp 13–24
40. Klügl F, Fehler M, Herrler R (2005) About the role of the environment in multi-agent simulations. In: Weyns D, Parunak HVD, Michel F (eds) Environments for Multi-Agent Systems, First International Workshop E4MAS (2004), New York 19 July 2004. Revised Selected Papers. vol 3374, pp 127–149
41. Langton C (1995) Artificial life: An overview. MIT Press, Cambridge
42. Latombe JC (1991) Robot Motion Planning. Kluwer, Boston
43. Luck M, McBurney P, Sheory O, Willmott S (eds) (2005) Agent Technology: Computing as Interaction. University of Southampton, Southampton
44. Mamei M, Zambonelli F, Leonardi L (2002) Co-fields: towards a unifying approach to the engineering of swarm intelligent systems. In: Engineering Societies in the Agents World III: Third International Workshop (ESAW 2002). Lecture Notes in Artificial Intelligence, vol 2577. Springer, Berlin, pp 68–81
45. Moss S, Davidsson P (eds) (2001) Multi-Agent-Based Simulation, Second International Workshop MABS (2000), Boston, July, (2000), Revised and Additional Papers. Lecture Notes in Computer Science, vol 1979. Springer, Berlin
46. Murata T (1989) Petri nets: properties, analysis and applications. Proc IEEE 77:541–580
47. North MJ, Collier NT, Vos JR (2006) Experiences creating three implementations of the repast agent modeling toolkit. ACM Trans Model Comp Sim 16:1–25
48. Omicini A, Zambonelli F (1999) Coordination for Internet application development. Autono Agents Multi-Agent Syst 2:251–269 Special Issue: Coordination Mechanisms for Web Agents
49. Picco GP, Murphy AL, Roman GC (1999) Lime: Linda meets mobility. In: Proceedings of the 21st International Conference on Software Engineering (ICSE 99) ACM Press, Los Angeles, pp 368–377
50. Pyka A, Fagiolo G (2007) Agent-Based Modelling: A Methodology for Neo-Schumpeterian Economics. In: Hanusch H, Pyka A (eds) Elgar Companion to Neo-Schumpeterian Economics. Edward Elgar Publishing, pp 467–487
51. Rao A, Georgeff M (1991) Modeling rational agents within a BDI-architecture. In: Proc Knowledge Representation and Reasoning (KR&R 1991)
52. Rao A, Georgeff M (1995) BDI agents: from theory to practice. In: Proceedings of the International Conference on Multi-Agent Systems
53. Russel S, Norvig P (1995) Artificial Intelligence: A Modern Approach. Prentice Hall, Upper Saddle River
54. Schadschneider A, Kirchner A, Nishinari K (2002) CA approach to collective phenomena in pedestrian dynamics. In: Bandini S, Chopard B, Tomassini M (eds) Cellular Automata, 5th International Conference on Cellular Automata for Research and Industry ACRI 2002. Lecture Notes in Computer Science, vol 2493. Springer, Berlin, pp 239–248
55. Sichman JS, Antunes L (eds) (2006) Multi-Agent-Based Simulation VI, International Workshop MABS (2005), Utrecht, The Netherlands, 25 July 2005, Revised and Invited Papers. In: Sichman JS, Antunes L (eds) MABS. Lecture Notes in Computer Science, vol 3891. Springer, Berlin
56. Sichman JS, Conte R, Gilbert N (eds) (1998) Multi-Agent Systems and Agent-Based Simulation, First International Work-

shop MABS '98, Paris, France, 4–6 July 1998. Proceedings. In: Sichman JS, Conte R, Gilbert N (eds): MABS. Lecture Notes in Computer Science, vol 1534. Springer, Berlin
57. Sichman JS, Bousquet F, Davidsson P (eds) (2003) Multi-Agent-Based Simulation, Third International Workshop MABS (2002), Bologna, Italy, 15–16 July 2002, Revised Papers. In: Sichman JS, Bousquet F, Davidsson P (eds) MABS. Lecture Notes in Computer Science, vol 2581. Springer, Berlin
58. Torrens P (2002) Cellular automata and multi-agent systems as planning support tools. In: Geertman S, Stillwell J (eds) Planning Support Systems in Practice. Springer, London, pp 205–222
59. Tummolini L, Castelfranchi C, Ricci A, Viroli M, Omicini A (2004) "Exhibitionists" and "voyeurs" do it better: A shared environment approach for flexible coordination with tacit messages. In: Weyns D, Parunak HVD, Michel F (eds) 1st International Workshop on Environments for MultiAgent Systems (E4MAS 2004), pp 97–111
60. Wahle J, Schreckenberg M (2001) A multi-agent system for online simulations based on real-world traffic data. In: Annual Hawaii International Conference on System Sciences (HICSS-34), IEEE Computer Society
61. Weyns D, Holvoet T (2006) From reactive robots to situated multi-agent systems: a historical perspective on the role of environment in multi-agent systems. In: Dikenelli O, Gleizes MP, Ricci A (eds) Engineering Societies in the Agents World VI, 6th International Workshop ESAW (2005). Lecture Notes in Computer Science, vol 3963. Springer, Berlin, pp 63–88
62. Weyns D, Schelfthout K, Holvoet T, Lefever T (2005) Decentralized control of E'GV transportation systems. In: AAMAS Industrial Applications. ACM Press, Utrecht, pp 67–74
63. Weyns D, Vizzari G, Holvoet T (2006) Environments for situated multi-agent systems: beyond infrastructure. In: Weyns D, Parunak HVD, Michel F (eds) Environments for Multi-Agent Systems II, Second International Workshop E4MAS (2005), Utrecht, 25 July 2005. Selected Revised and Invited Papers. Lecture Notes in Computer Science, vol 3830. Springer, Berlin, pp 1–17
64. Weyns D, Boucké N, Holvoet T (2006) Gradient field-based task assignment in an AGV transportation system. In: AAMAS '06: Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems. ACM Press, Hakodate, pp 842–849
65. Weyns D, Brueckner SA, Demazeau Y (eds) (2008) Engineering Environment-Mediated Multi-Agent Systems: International Workshop, EEMMAS 2007, Dresden, Germany, October 2007. Selected Revised and Invited Papers. Lecture Notes in Computer Science, vol 5049. Springer, Berlin
66. Weyns D, Omicini A, Odell J (2007) Environment as a first class abstraction in multiagent systems. Auton Agents Multi-Agent Syst 14:5–30
67. Wolfram S (1986) Theory and Applications of Cellular Automata. World Press, Singapore
68. Wooldridge MJ, Jennings NR (1995) Intelligent agents: theory and practice. Knowl Eng Rev 10:115–152
69. Zambonelli F, Parunak HVD (2003) Signs of a revolution in computer science and software engineering. In: Petta P, Tolksdorf R, Zambonelli F (eds) Engineering Societies in the Agents World III, Third International Workshop, ESAW 2002, Madrid, Spain, September 2002, Revised Papers. Lecture Notes in Computer Science, vol 2577. Springer, Berlin, pp 13–28

Agent Based Modeling and Simulation, Introduction to

Filippo Castiglione
Institute for Computing Applications (IAC) – National Research Council (CNR), Rome, Italy

Agent-based modeling (ABM) is a relatively new computational modeling paradigm that is markedly useful in studying complex systems composed of a large number of interacting entities with many degrees of freedom. Other names for ABM are individual-based modeling (IBM) or multi-agent systems (MAS). Physicists often use the term micro-simulation or interaction-based computing.

The basic idea of ABM is to construct the computational counterpart of a conceptual model of a system under study on the basis of discrete entities (agents) with some properties and behavioral rules, and then to simulate them in a computer to mimic the real phenomena.

The definition of agent is somewhat controversial as witnessed by the fact that the models found in the literature adopt an extremely heterogeneous rationale. The agent is an autonomous entity having its own internal state reflecting its perception of the environment and interacting with other entities according to more or less sophisticated rules. In practice, the term agent is used to indicate entities ranging all the way from simple pieces of software to "conscious" entities with learning capabilities. For example, there are "helper" agents for web retrieval, robotic agents to explore inhospitable environments, buyer/seller agents in an economy, and so on. Roughly speaking, an entity is an "agent" if it has some degree of autonomy, that is, if it is distinguishable from its environment by some kind of spatial, temporal, or functional attribute: an agent must be identifiable. Moreover, it is usually required that an agent must have some autonomy of action and that it must be able to engage in tasks in an environment without direct external control.

From simple agents, which interact locally with simple rules of behavior, merely responding befittingly to environmental cues, and not necessarily striving for an overall goal, we observe a synergy which leads to a higher-level whole with much more intricate behavior than the component agents (*holism*, meaning all, entire, total). Agents can be identified on the basis of a set of properties that must characterize an entity, and in particular, *autonomy* (the capability of operating without intervention by humans, and a certain degree of control over its own state); *social ability* (the capability of interacting by employing some kind of agent communication language); *reactivity* (the ability

to perceive an environment in which it is situated and respond to perceived changes); and *pro-activeness* (the ability to take the initiative, starting some activity according to internal goals rather than as a reaction to an external stimulus). Moreover, it is also conceptually important to definite what the agent "environment" in an ABM is.

In general, given the relative immaturity of this modeling paradigm and the broad spectrum of disciplines in which it is applied, a clear cut and widely accepted definition of high level concepts of agents, environment, interactions and so on, is still lacking. Therefore a real ABM ontology is needed to address the epistemological issues related to the agent-based paradigm of modeling of complex systems in order to attempt to reach a more general comprehension of emergent properties which, though ascribed to the definition of a specific application domain, are also universal (see ► Agent Based Modeling and Simulation).

Historically, the first simple conceptual form of agent-based models was developed in the late 1940s, and it took the advent of the computer to show its modeling power. This is the Von Neumann machine, a theoretical machine capable of reproduction. The device von Neumann proposed would follow precisely detailed instructions to produce an identical copy of itself. The concept was then improved by Stanislaw Ulam. He suggested that the machine be built on paper, as a collection of cells on a grid. This idea inspired von Neumann to create the first of the models later termed cellular automata (CA). John Conway then constructed the well-known "Game of Life". Unlike the von Neumann's machine, Conway's *Game of Life* operated by simple rules in a virtual world in the form of a two-dimensional checkerboard.

Conway's Game of Life has become a paradigmatic example of models concerned with the emergence of order in nature. How do systems self-organize themselves and spontaneously achieve a higher-ordered state? These and other questions have been deeply adressed in the first workshop on Artificial Life held in the late 1980s in Santa Fe. This workshop shaped the ALife field of research. Agent-based modeling is historically connected to ALife because it has become a distinctive form of modeling and simulation in this field. In fact, the essential features of ALife models are translated into computational algorithms throug agent-based modeling (see ► Agent Based Modeling and Artificial Life).

Agent-based models can be seen as the natural extension of the CA-like models, which have been very successful in the past decades in shedding light on various physical phenomena. One important characteristic of ABMs which distinguishes them from cellular automata, is the potential asynchrony of the interactions among agents and between agents and their environments. In ABMs, agents typically do not simultaneously perform actions at constant time-steps, as in CAs or boolean networks. Rather, their actions follow discrete-event cues or a sequential schedule of interactions. The discrete-event setup allows for the cohabitation of agents with different environmental experiences. Also ABMs are not necessarily grid-based nor do agents "tile" the environment.

Physics investigation is based on building models of reality. It is a common experience that, even using simple "building blocks", one usually obtains systems whose behavior is quite complex. This is the reason why CA-like, and therefore agent-based models, have been used extensively among physicists to investigate experimentally (that is, on a computer) the essential ingredients of a complex phenomenon. Rather than being derived from some fundamental law of physics, these essential ingredients constitute artificial worlds. Therefore, there exists a pathway from Newton's laws to CA and ABM simulations in classical physics that has not yet expressed all its potential (see ► Interaction Based Computing in Physics).

CA-like models also proved very successful in theoretical biology to describe the aggregation of cells or microorganisms in normal or pathological conditions (see ► Cellular Automaton Modeling of Tumor Invasion).

Returning to the concept of agent in the ABM paradigm, an agent may represent a particle, a financial trader, a cell in a living organism, a predator in a complex ecosystem, a power plant, an atom belonging to a certain material, a buyer in a closed economy, customers in a market model, forest trees, cars in large traffic vehicle system, etc. Once the level of description of the system under study has been defined, the identification of such entities is quite straightforward. For example, if one looks at the world economy, then the correct choice of agents are nations, rather than individual companies. On the other hand, if one is interested in looking at the dynamics of a stock, then the entities determining the price evolution are the buyers and sellers.

This example points to a field where ABM provides a very interesting and valuable instrument of research. Indeed, mainstream economic models typically make the assumption that an entire group of agents, for example, "investors", can be modeled with a single "rational representative agent". While this assumption has proven extremely useful in advancing the science of economics by yielding analytically tractable models, it is clear that the assumption is not realistic: people differ in their tastes, beliefs, and sophistication, and as many psychological studies have shown, they often deviate from rationality in

systematic ways. Agent-based computational economics (ACE) is a framework allowing economics to expand beyond the realm of the "rational representative agent". By modeling and simulating the behavior of each agent and interactions among agents, agent-based simulation allows us to investigate the dynamics of complex economic systems with many heterogeneous (and not necessarily fully rational) agents. Agent-based computational economics complements the traditional analytical approach and is gradually becoming a standard tool in economic analysis (see ► Agent Based Computational Economics).

Because the paradigm of agent-based modeling and simulation can handle richness of detail in the agent's description and behavior, this methodology is very appealing for the study and simulation of *social systems*, where the behavior and the heterogeneity of the interacting components are not safely reducible to some stylized or simple mechanism. Social phenomena simulation in the area of agent-based modeling and simulation, concerns the emulation of the individual behavior of a group of social entities, typically including their cognition, actions and interaction. This field of research aims at "growing" artificial society following a bottom-up approach.

Historically, the birth of the agent-based model as a model for social systems can be primarily attributed to a computer scientist, Craig Reynolds. He tried to model the reality of lively biological agents, known as artificial life, a term coined by Christopher Langton. In 1996 Joshua M. Epstein and Robert Axtell developed the first large scale agent model, the Sugarscape, to simulate and explore the role of social phenomenon such as seasonal migrations, pollution, sexual reproduction, combat, transmission of disease and even culture (see ► Social Phenomena Simulation).

In the field of artificial intelligence, the collective behavior of agents that without central control, collectively carry out tasks normally requiring some form of "intelligence", constitutes the central concept in the field of *swarm intelligence*. The term "swarm intelligence" first appeared in 1989. As the use of the term swarm intelligence has increased, its meaning has broadened to a point in which it is often understood to encompass almost any type of collective behavior. Technologically, the importance of "swarms" is mainly based on potential advantages over centralized systems. The potential advantages are: economy (the swarm units are simple, hence, in principle, mass producible, *modularizable*, interchangeable, and disposable; reliability (due to the redundancy of the components; destruction/death of some units has negligible effect on the accomplishment of the task, as the swarm adapts to the loss of few units); ability to perform tasks beyond those of centralized systems, for example, escaping enemy detection. From this initial perspective on potential advantages, the actual application of swarm intelligence has extended to many areas, and inspired potential future applications in defense and space technologies, (for example, control of groups of unmanned vehicles in land, water, or air), flexible manufacturing systems, and advanced computer technologies (bio-computing), medical technologies, and telecommunications (see ► Swarm Intelligence).

Similarly, robotics has adopted the ABM paradigm to study, by means of simulation, the crucial features of adaptation and cooperation in the pursuit of a global goal. Adaptive behavior concerns the study of how organisms develop their behavioral and cognitive skills through a synthetic methodology, consisting of designing artificial agents which are able to adapt to their environment autonomously. These studies are important both from a modeling point of view (that is, for better understanding intelligence and adaptation in natural beings) and from an engineering point of view (that is, for developing artifacts displaying effective behavioral and cognitive skills) (see ► Embodied and Situated Agents, Adaptive Behavior in).

What makes ABM a novel and interesting paradigm of modeling is the idea that agents are individually represented and "monitored" in the computer's memory. One can, at any time during the simulation, ask a question such as "what is the age distribution of the agents?", or "how many stocks have accumulated buyers following that specific strategy?", or "what is the average velocity of the particles?". "Large scale" simulations in the context of agent-based modeling are not only simulations that are large in terms of size (number of agents simulated), but are also complex. Complexity is inherent in agent-based models, as they are usually composed of dynamic, heterogeneous, interacting agents. Large-scale agent-based models have also been referred to as "Massively Multi-agent Systems (MMAS)". MMAS is defined as "beyond resource limitation": the number of agents exceeds local computer resources, or the situations are too complex to design and program given human cognitive resource limits. Therefore, for agent-based modeling, "large scale" is not simply a size problem, it is also a problem of managing complexity to ensure scalability of the agent model. Agent-based models increase in scale as the modeler requires many agents to investigate whole system behavior, or as the modeler wishes to fully examine the response of a single agent in a realistic context. There are two key problems that have to be tackled as the scale of a multi-agent system increases: computational resources limit the simulation time and/or data storage capacity; and agent model analysis may be-

come more difficult. Difficulty in analyzing the model may be due to the model system having a large number of complex components or due to memory for model output storage being restricted by computer resources (see ► Agent Based Modeling, Large Scale Simulations).

For implementation of agent-based models, both domain-specific and general-purpose languages are routinely used. Domain-specific languages include business-oriented languages (for example, spreadsheet programming tools); science and engineering languages (such as Mathematica); and dedicated agent-based modeling languages (for example, NetLogo). General-purpose languages can be used directly (as in the case of Java programming) or within agent-based modeling toolkits (for example, Repast). The choice that is most appropriate for a given modeling project depends on both the requirements of that project and the resources available to implement it (see ► Agent Based Modeling and Computer Languages).

Interestingly, ABM is not being used exclusively in science. In fact, the entertainment industry has promoted its own interest in the ABM technology. As graphics technology has improved in recent years, more and more importance has been placed on the behavior of virtual characters in applications set in virtual worlds in areas such as games, movies and simulations. The behavior of virtual characters should be believable in order to create the illusion that these virtual worlds are populated with living characters. This has led to the application of agent-based modeling to the control of these virtual characters. There are a number of advantages of using agent-based modeling techniques which include the fact that they remove the requirement for hand controlling all agents in a virtual environment, and allow agents in games to respond to unexpected actions by players (see ► Computer Graphics and Games, Agent Based Modeling in).

Since it is difficult to formally analyze complex multi-agent systems, they are mainly studied through computer simulations. While computer simulations can be very useful, results obtained through simulations do not formally validate the observed behavior. It is widely recognized that building a sound and widely applicable theory for ABM systems will require an inter-disciplinary approach and the development of new mathematical and computational concepts. In other words, there is a compelling need for a mathematical framework which one can use to represent ABM systems and formally establish their properties. Fortunately, some known mathematical frameworks already exist that can be used to formally describe multi-agent systems, for example, that of finite *dynamical systems* (both deterministic and stochastic). A sampling of the results from this field of mathematics shows that they can be used to carry out rigorous studies of the properties of multi-agent systems and, in general, that they can also serve as universal models for computation. Moreover, special cases of dynamical systems (sequential dynamical systems) can be structured in accordance with the theory of categories and therefore provide the basis for a formal theory to describe ABM behavior (see ► Agent Based Modeling, Mathematical Formalism for).

On the same line of thought, agents and interaction can be studied from the perspective of logic and computer science. In particular, ideas about *logical dynamics*, *games semantics* and *geometry of interaction*, which have been developed over the past two decades, lead towards a structural theory of agents and interaction. This provides a basis for powerful logical methods such as compositionality, types and high-order calculi, which have proved so fruitful in computer science, to be applied in the domain of ABM and simulation (see ► Logic and Geometry of Agents in Agent-Based Modeling).

The appeal of ABM methodology in science increases manifestly with advances in computational power of modern computers. However, it is important to bear in mind that increasing the complexity of a model does not necessarily bring more understanding of the fundamental laws governing the overall dynamics. Actually, beyond a certain level of model complexity, the model loses its ability to explain or predict reality, thus reducing model building to a mere surrogate of reality where things may happen with a surprising adherence to reality, but we are unable to explain *why* this happens. Therefore, model construction must proceed incrementally, step by step, possibly validating the model at each stage of development before adding more details. ABM technology is very powerful but, if badly used, could reduce science to a mere exercise consisting of mimicking reality.

Agent Based Models in Economics and Complexity

Mauro Gallegati[1], Matteo G. Richiardi[1,2]
[1] Università Politecnica delle Marche, Ancona, Italy
[2] Collegio Carlo Alberto – LABORatorio R. Revelli, Moncalieri, Italy

Article Outline

Glossary
Definition of the Subject
Introduction

Some Limits of the Mainstream Approach
The Economics of Complexity
Additional Features of Agent-Based Models
An Ante Litteram Agent-Based Model: Thomas Schelling's Segregation Model
The Development of Agent-Based Modeling
A Recursive System Representation of Agent-Based Models
Analysis of Model Behavior
Validation and Estimation
The Role of Economic Policy
Future Directions
Bibliography

Glossary

Abduction also called *inference to the best explanation*, abduction is a method of reasoning in which one looks for the hypothesis that would best explain the relevant evidence.

Agents entities of a model that (i) are perceived as a unit from the outside, (ii) have the ability to *act*, and possibly to *react* to external stimuli and *interact* with the environment and other agents.

Agent-based computational economics (ACE) is the computational study of economic processes modeled as dynamic systems of interacting agent.

Agent-based models (ABM) are models where (i) there is a multitude of objects that interact with each other and with the environment; (ii) the objects are autonomous, i. e. there is no central, or top-down control over their behavior; and (iii) the outcome of their interaction is numerically computed.

Complexity there are more than 45 existing definitions of complexity (Seth Lloyd, as reported on p. 303 in [97]). However, they can be grouped in just two broad classes: a *computational* view and a *descriptive* view. Computational (or algorithmic) complexity is a measure of the amount of information necessary to compute a system; descriptive complexity refers to the amount of information necessary to describe a system. We refer to this second view, and define complex systems as systems characterized by emergent properties (see *emergence*).

Deduction the logical derivation of conclusions from given premises.

Economics is the science about the intended and unintended consequences of individual actions, in an environment characterized by scarce resources that both requires and forces to interaction.

Emergence the spontaneous formation of self-organized structures at different layers of a hierarchical system configuration.

Evolution in biology, is a change in the inherited traits of a population from one generation to the next. In social sciences it is intended as an endogenous change over time in the behavior of the population, originated by competitive pressure and/or learning.

Heterogeneity non-degenerate distribution of characteristics in a population of agents.

Induction the intuition of general patterns from the observation of statistical regularities.

Interaction a situation when the actions or the supposed actions of one agent may affect those of other agents within a reference group.

Out-of-equilibrium a situation when the behavior of a system, in terms of individual strategies or aggregate outcomes, is not stable.

Definition of the Subject

A crucial aspect of the complexity approach is how interacting elements produce aggregate patterns that those elements in turn react to. This leads to the emergence of aggregate properties and structures that cannot be guessed by looking only at individual behavior.

It has been argued [144] that complexity is ubiquitous in economic problems (although this is rarely acknowledged in economic modeling), since (i) the economy is inherently characterized by the interaction of individuals, and (ii) these individuals have cognitive abilities, e. g. they form expectations on aggregate outcomes and base their behavior upon them: "Imagine how hard physics would be if electrons could think", is how the Nobel prize winner Murray Gell–Mann, a physicist, has put it (as reported by Page [131]).

Explicitly considering how heterogeneous elements dynamically develop their behavior through interaction is a hard task analytically, the equilibrium analysis of mainstream (neoclassical) economics being a shortcut that in many cases is at risk of throwing the baby out with the bath water, so to speak. On the other hand, numerical computation of the dynamics of the process started to be a feasible alternative only when computer power became widely accessible. The computational study of heterogeneous interaction agents is called agent-based modeling (ABM). Interestingly, among its first applications a prominent role was given to economic models [4], although it was quickly found of value in other disciplines too (from sociology to ecology, from biology to medicine). The goal of this chapter is to motivate the use of the complexity approach

and agent-based modeling in economics, by discussing the weaknesses of the traditional paradigm of mainstream economics, and then explain what ABM is and which research and policy questions it can help to analyze.

Introduction

Economics is in troubled waters. Although there exists a mainstream approach, its internal coherence and ability to explain the empirical evidence are increasingly questioned. The causes of the present state of affairs go back to the middle of the eighteenth century, when some of the Western economies were transformed by the technological progress which lead to the industrial revolution. This was one century after the Newtonian revolution in physics: from the small apple to the enormous planets, all objects seemed to obey the simple *natural* law of gravitation. It was therefore natural for a new figure of social scientist, the economist, to borrow the method (*mathematics*) of the most successful hard science, physics, allowing for the mutation of *political economy* into *economics*. It was (and still is) the mechanical physics of the seventeenth century, which ruled economics. In the final chapter of his General Theory, Keynes wrote of politicians as slaves of late economists: in their turn, they are slaves of late physicists of the seventeenth century (see also [125]).

From then on, economics lived its own evolution based on the classical physics assumptions (reductionism, determinism and mechanicism). Quite remarkably, the approach of statistical physics, which deeply affected physical science at the turn of the nineteenth century by emphasizing the difference between micro and macro, was adopted by Keynes around the mid 1930s. However, after decades of extraordinary success it was rejected by the neoclassical school around the mid 1970s, which framed the discipline into the old approach and ignored, by definition, any interdependencies among agents and difference between individual and aggregate behavior (being agents, electrons, nations or planets).

The ideas of natural laws and equilibrium have been transplanted into economics *sic et simpliciter*. As a consequence of the adoption of the classical mechanics paradigm, the difference between micro and macro was analyzed under a reductionist approach. In such a setting, aggregation is simply the process of summing up market outcomes of individual entities to obtain economy-wide totals. This means that there is no difference between micro and macro: the dynamics of the whole is nothing but a summation of the dynamics of its components (in term of physics, the motion of a planet can be described by the dynamics of the atoms composing it). This approach does not take into consideration that there might be two-way interdependencies between the agents and the aggregate properties of the system: interacting elements produce aggregate patterns that those elements in turn react to. What macroeconomists typically fail to realize is that the correct procedure of aggregation is not a sum: this is when emergence enters the drama. With the term emergence we mean the arising of complex structures from simple individual rules [147,153,171]. Empirical evidence, as well as experimental tests, shows that aggregation generates regularities, i. e. simple individual rules, when aggregated, produce statistical regularities or well-shaped aggregate functions: regularities emerge from individual chaos [106]. The concept of equilibrium is quite a dramatic example. In many economic models equilibrium is described as a state in which (individual and aggregate) demand equals supply. The notion of *statistical equilibrium*, in which the aggregate equilibrium is compatible with individual disequilibrium, is outside the box of tools of the mainstream economist. The same is true for the notion of *evolutionary equilibrium* (at an aggregate level) developed in biology. The equilibrium of a system no longer requires that every single element be in equilibrium by itself, but rather that the statistical distributions describing aggregate phenomena be stable, i. e. in "[...] a state of macroscopic equilibrium maintained by a large number of transitions in opposite directions" (p. 356 in [64]).

According to this view, an individual organism is in equilibrium *only when it is dead*. A consequence of the idea that macroscopic phenomena can emerge is that reductionism is wrong.

Ironically, since it can be argued, as we will do in the section below, that economics strongly needs this methodological twist [144], ABM has received lees attention in economics than in other sciences ([110]; but [82] is a counter-example). The aim of this chapter is not to provide a review of applications of the complexity theory to economics (the interested reader is referred to [15,26,60,124,140,142]), but rather to describe the development of the *Agent-Based Modeling* (ABM) approach to complexity.

The chapter is structured as follows: after reviewing some limits of mainstream economics (Sect. "Additional Features of Agent-Based Models"), Sects. "The Economics of Complexity" and "Additional Features of Agent-Based Models" describe how the complexity perspective differs from the traditional one, and how many problems of the mainstream approach can be overcome by ABM. As an example, we present a prototypical example of ABM, based on the work of Thomas Schelling on the dynamics of segregation. After dedicating some sections to, respectively,

a skeleton history of ABM, a recursive system representation of these models, a discussion on how ABM can be interpreted, estimated and validated, we finally discus how the complexity approach can be used to guide policy intervention and analysis. A final section discusses the achievements of the ABM agenda.

Some Limits of the Mainstream Approach

The research program launched by the neoclassical school states that macroeconomics should be explicitly grounded on microfoundations. This is how Robert Lucas put it: "The most interesting recent developments in macroeconomic theory seem to me describable as the reincorporation of aggregative problems [...] within the general framework of 'microeconomic' theory. If these developments succeed, the term 'macroeconomic' will be simply disappear from use and the modifier 'micro' will become superfluous. We will simply speak, as did Smith, Marshall and Walras, of *economic theory*" (pp. 107–108 in [115]). According to the mainstream, this implies that economic phenomena at a macroscopic level should be explained as a summation of the activities undertaken by individual decision makers. This procedure of microfoundation is very different from that now used in physics. The latter starts from the micro-dynamics of the single particle, as expressed by the Liouville equation and, through the master equation, ends up with the macroscopic equations. In the aggregation process, the dynamics of the agents lose their degree of freedom and behave coherently in the aggregate. In mainstream economics, while the procedure is formally the same (from micro to macro), it is assumed that the dynamics of the agents are those of the aggregate. The reduction of the degree of freedom, which is characteristic of the aggregation problem in physics, is therefore ruled out: a rational agent with complete information chooses to implement the individually optimal behavior, without additional constraints. There are three main pillars of this approach: (i) the precepts of the rational choice-theoretic tradition; (ii) the equilibrium concept of the Walrasian analysis; and (iii) the reductionist approach of classical physics. In the following, we will show that assumptions (i)–(ii), which constitute the necessary conditions for reducing macro to micro, are logically flawed (and empirically unfounded), while rejection of (iii) opens the road to complexity.

Mainstream economics is axiomatic and based on unrealistic (or unverifiable) assumptions. According to the supporters of this view, such an abstraction is necessary since the real world is complicated: rather than compromising the epistemic worth of economics, such assumptions are essential for economic knowledge. However, this argument does not invalidate the criticism of unrealistic assumptions [136]. While it requires internal coherence, so that theorems can be logically deduced from a set of assumptions, it abstracts from external coherence between theoretical statements and empirical evidence. Of course, this implies an important epistemological detachment from falsifiable sciences like physics. In setting the methodological stage for the dynamic stochastic general equilibrium (DSGE) macroeconomic theory, Lucas and Sargent declared:

> "An economy following a multivariate stochastic process is now routinely described as being in equilibrium, by which is meant nothing more that at each point in time (a) markets clears and (b) agents act in their own self-interest. This development, which stemmed mainly from the work of Arrow [...] and Debreu [...], implies that simply to look at any economic time series and conclude that it is a disequilibrium phenomenon is a meaningless observation. [...] The key elements of these models are that agents are rational, reacting to policy changes in a way which is in their best interests privately, and that the impulses which trigger business fluctuations are mainly unanticipated shocks." (p. 7 in [116]).

The self-regulating order of Adam Smith [153] is transformed into a competitive general equilibrium (GE) in the form elaborated in the 1870s by Walras, that is a configuration of (fully flexible) prices and plans of action such that, at those prices, all agents can carry out their chosen plans and, consequently, markets clear. In a continuous effort of generalization and analytical sophistication, modern (neoclassical) economists interested in building microfoundations for macroeconomics soon recurred to the refinement proposed in the 1950s by Arrow and Debreu [14], who showed that also individual intertemporal (on an infinite horizon) optimization yields a GE, as soon as the economy is equipped with perfect price foresight for each future state of nature and a complete set of Arrow-securities markets [11], all open at time zero and closed simultaneously. Whenever these conditions hold true, the GE is an allocation that maximizes a properly defined social welfare function, or the equilibrium is Pareto-efficient (*First Welfare Theorem*).

The literature has pointed out several logical inconsistencies of the mainstream approach. Davis [44] identifies three impossibility results, which determine the breakdown of the mainstream, i. e. neoclassical, economics:

(i) Arrow's 1951 theorem showing that neoclassical theory is unable to explain social choice [10]; (ii) the Cambridge capital debate pointing out that mainstream is contradictory with respect to the concept of aggregate capital [40]; and (iii) the Sonnenschein–Mantel–Debreu results showing that the standard comparative static reasoning is inapplicable in general equilibrium models. In particular, a few points are worth remembering here.

1. The GE is neither unique nor locally stable under general conditions. This negative result, which refers to the work of Sonnenschein [155], Debreu [46] and Mantel [119], can be summarized along the following lines. Let the aggregate excess demand function $F(p)$ – obtained from aggregating among individual excess demands $f(p)$ – be a mapping from the price simplex Π to the commodity space P^N. A GE is defined as a price vector p such that $F(p*) = 0$. It turns out that the only conditions that $F(\cdot)$ inherits from $f(\cdot)$ are continuity, homogeneity of degree zero and the Walras' law (i. e., the total value of excess demand is zero). These assure the existence, but neither the uniqueness nor the local stability of $p*$, unless preferences generating individual demand functions are restricted to very implausible cases.
2. The existence of a GE is proved *via* the Brower's fix point theorem, i. e. by finding a continuous function $g(\cdot)\colon \Pi \to \Pi$ so that any fixed point for $g(\cdot)$ is also an equilibrium price vector $F(p*) = 0$. Suppose that we are interested in finding an algorithm which, starting from an arbitrary price vector p, chooses price sequences to check for $p*$ and halts when it finds it. In other terms, to find the GE price vector $F(p*) = 0$ means that halting configurations are decidable. As this violates the undecidability of the halting problem for Turing machines, from a recursion theoretic viewpoint the GE solution is incomputable [138,167]. Notice that the same problem applies, in spite of its name, to the class of computable GE models [169].
3. By construction, in a GE all transactions are undertaken at the same equilibrium price vector. Economic theory has worked out two mechanisms capable of reaching this outcome. First, one can assume that buyers and sellers adjust, costless, their optimal supplies and demands to prices called out by a (explicit or implicit) fictitious auctioneer, who continues to do his job until he finds a price vector which clears all markets. Only then transactions take place (Walras' assumption). Alternatively, buyers and sellers sign provisional contracts and are allowed to freely (i. e., without any cost) recontract until a price vector is found which makes individual plans fully compatible. Once again, transactions occur only after the equilibrium price vector has been established (Edgeworth's assumption). Regardless of the mechanism one adopts, the GE model is one in which the formation of prices precedes the process of exchange, instead of being the result of it, through a *tatonnement* process occurring in a meta-time. Real markets work the other way round and operates in real time, so that the GE model cannot be considered a scientific explanation of real economic phenomena [9].
4. It has been widely recognized since Debreu [45], that integrating money in the theory of value represented by the GE model is at best problematic. No economic agent can individually decide to monetize alone; monetary trade should be the equilibrium outcome of market interactions among optimizing agents. The use of money – that is, a common medium of exchange and a store of value – implies that one party to a transaction gives up something valuable (for instance, his endowment or production) for something inherently useless (a fiduciary token for which he has no immediate use) in the hope of advantageously re-trading it in the future. Given that in a GE model actual transactions take place only after a price vector coordinating all trading plans has been freely found, money can be consistently introduced into the picture only if the logical keystone of the absence of transaction costs is abandoned. By the same token, since credit makes sense only if agents can sign contracts in which one side promises future delivery of goods or services to the other side, in equilibrium markets for debt are meaningless, and bankruptcy can be safely ignored. Finally, as the very notion of a GE implies that all transactions occur only when individual plans are mutually compatible, and this has to be true also in the labor market, the empirically observed phenomenon of involuntary unemployment and the microfoundation program put forth by Lucas and Sargent are logically inconsistent.
5. The very absence of money and credit is a consequence of the fact that in GE there is no time. The only role assigned to time in a GE model is, in fact, that of dating commodities. Products, technologies and preferences are exogenously given and fixed from the outset. The convenient implication of banning out-of-equilibrium transactions is simply that of getting rid of any disturbing influence of intermediary modifications of endowments – and therefore of individual excess demands – on the final equilibrium outcome. The introduction of non-Walrasian elements into the GE microfoundations program – such as fixed or sticky prices, imperfect competition and incomplete markets leading to temporary

equilibrium models – yields interesting Keynesian features such as the breaking of the Say's law and scope for a monetary theory of production, a rationale for financial institutions and a more persuasive treatment of informational frictions. As argued in Vriend [165], however, all these approaches preserve a Walrasian perspective in that models are invariably closed by a GE solution concept which, implicitly or (more often) not, implies the existence of a fictitious auctioneer who processes information, calculates equilibrium prices and quantities, and regulates transactions. As a result, if the Walrasian auctioneer is removed the decentralized economy becomes dynamically incomplete, as we are not left with any mechanism determining how quantities and prices are set and how exchanges occur.

The flaws of the solution adopted by mainstream macroeconomists to overcome the problems of uniqueness and stability of equilibrium on the one hand, and of analytical-tractability on the other one – i. e. the usage of a representative agent (RA) whose choices summarize those of the whole population of agents – are so pervasive that we discuss them hereafter.

6. Although the RA framework has a long history, it is standard to build the microfoundation procedure on it only after Lucas' critique paper [114]. Mainstream models are characterized by an explicitly stated optimization problem of the RA, while the derived individual demand or supply curves are used to obtain the aggregate demand or supply curves. Even when the models allow for heterogeneity, interaction is generally absent (the so-called *weak interaction hypothesis* [139]). The use of RA models should allow one to avoid the Lucas critique, to provide microfoundations to macroeconomics, and, *ça va sans dire*, to build Walrasian general equilibrium models. Since models with many heterogeneous interacting agents are complicated and no closed form solution is often available (aggregation of heterogenous interacting agents is analyzed in [5,6,7,53,78]), economists assume the existence of an RA: a simplification that makes it easier to solve for the competitive equilibrium allocation, since direct interaction is ruled out by definitions. Unfortunately, as Hildenbrand and Kirman [95] noted:

   > "There are no assumptions on isolated individuals, which will give us the properties of aggregate behavior. We are reduced to making assumptions at the aggregate level, which cannot be justified, by the usual individualistic assumptions. This problem is usually avoided in the macroeconomic literature by assuming that the economy behaves like an individual. Such an assumption cannot be justified in the context of the standard model".

The equilibria of general equilibrium models with a RA are characterized by a complete absence of trade and exchange, which is a counterfactual idea. Kirman [99], Gallegati [76] and Caballero [36] show that RA models ignore valid aggregation concerns, by neglecting interaction and emergence, hence committing fallacy of composition (what in philosophy is called fallacy of division, i. e. to attribute properties to a different level than where the property is observed: game theory offers a good case in point with the concept of Nash equilibrium, by assuming that social regularities come from the agent level equilibrium). Those authors provide examples in which the RA does not represent the individuals in the economy so that the reduction of a group of heterogeneous agents to an RA is not just an analytical convenience, but it is both unjustified and leads to conclusions which are usually misleading and often wrong ([99]; see also [98]). A further result, which is a proof of the logical fallacy in bridging the micro to the macro is the *impossibility theorem* of Arrow: it shows that an ensemble of people, which has to collectively take a decision, cannot show the same rationality of an individual [123]. Moreover, the standard econometric tools are based upon the assumption of an RA. If the economic system is populated by heterogeneous (not necessarily interacting) agents, then the problem of the microfoundation of macroeconometrics becomes a central topic, since some issues (e. g., *co-integration*, *Granger-causality*, *impulse-response function of structural VAR*) lose their significance [69].

All in all, we might say that the failure of the RA framework, points out the *vacuum* of the mainstream microfoundation literature, which ignores interactions: no box of tools is available to connect the micro and the macro levels, beside the RA whose existence is at odds with the empirical evidence [30,158] and the equilibrium theory as well [99].

The Economics of Complexity

According to the mainstream approach there is no direct interaction among economic units (for a pioneeristic and neglected contribution see [68]; see also [101]). In the most extreme case, any individual strategy is excluded (*principle of excluded strategy*, according to Schumpeter [149]) and agents are homogeneous. Small departures from the

perfect information hypothesis are incoherent with the Arrow–Debreu general equilibrium model, as shown by Grossman and Stiglitz [88], since they open the chance of having direct links among agents [156]. In particular, if prices convey information about the quality there cannot be an equilibrium price as determined by the demand-supply schedule, since demand curves depend on the probability distribution of the supply (p. 98 in [87]).

What characterizes a complex system is the notion of emergence, that is the spontaneous formation of self-organized structures at different layers of a hierarchical system configuration [43]. Rather, mainstream economics conceptualizes economic systems as consisting of several identical and isolated components, each one being a copy of a RA. The aggregate solution can thus be obtained by means of a simple summation of the choices made by each optimizing agent. The RA device, of course, is a way of avoiding the problem of aggregation by eliminating heterogeneity. But heterogeneity is still there. If the macroeconomist takes it seriously, he/she has to derive aggregate quantities and their relationships from the analysis of the micro-behavior of different agents. This is exactly the key point of the *aggregation problem*: starting from the *micro-equations* describing/representing the (optimal) choices of the economic units, what can we say about the *macro-equations*? Do they have the same functional form of the micro-equations (the *analogy principle*)? If not, how is the macro-theory derived?

The complexity approach to economics discards the GE approach to the microfoundation program, as well as its RA shorthand version. Instead of asking to deductively prove the existence of an equilibrium price vector $p*$ such that $F(p*) = 0$, it aims at explicitly constructing it by means of an algorithm or a rule. From an epistemological perspective, this implies a shift from the realm of classical to that of constructive theorizing [168]. Clearly, the act of computationally constructing a coordinated state – instead of imposing it via the Walrasian auctioneer – for a decentralized economic system requires complete description of goal-directed economic agents and their interaction structure.

Agent-based modeling represents an effective implementation of this research agenda ([60,124], see also [24, 67,81,175]). ABM is a methodology that allows one to construct, based on simple rules of behavior and interaction, models with heterogeneous agents, where the resulting aggregate dynamics and empirical regularities are not known a priori and are not deducible from individual behavior. It is characterized by three main tenets: (i) there is a multitude of objects that interact with each other and with the environment; (ii) the objects are autonomous, i. e. there is no central, or top-down control over their behavior; and (iii) the outcome of their interaction is numerically computed. Since the objects are autonomous, they are called agents ([3,4]; see also the repository of ACE-related material maintained by Leigh Tesfatsion at http://www.econ.iastate.edu/tesfatsi/ace.htm): "Agent-based Computational Economics is the computational study of economic processes modeled as dynamic systems of interacting agent" [161].

Agents can be anything from cells to biological entities, from individuals to social groups like families or firms. Agents can be composed by other agents: the only requirement being that they are perceived as a unit from the outside, and that they do something, i. e. they have the ability to *act*, and possibly to *react* to external stimuli and *interact* with the environment and other agents. The environment, which may include physical entities (infrastructures, geographical locations, etc.) and institutions (markets, regulatory systems, etc.), can also be modeled in terms of agents (e. g. a central bank, the order book of a stock exchange, etc.), whenever the conditions outlined above are met. When not, it should be thought of simply as a set of variables (say, temperature or business confidence).

The methodological issues are the real *litmus paper* of the competing approaches. According to one of the most quoted economic papers, Friedman [71], the ultimate goal of a positive science is to develop hypotheses that yield valid and meaningful *predictions* about actual phenomena. Not a word on predictions at the *meso*-level or on the realism of the hypotheses. Even the Occam rule is systematically ignored: e. g. to get a downward sloping aggregate demand curve, mainstream economics has to assume indifference curves which are: (i) defined only in the positive quadrant of commodity-bundle quantities; (ii) negatively sloped; (iii) complete; (iv) transitive, and (v) strictly convex, while ABM has to assume only the existence of reservation prices. Moreover, to properly aggregate from microbehavior, i. e. to get a well shaped aggregate demand from the individual ones, it has to be assumed that the propensity to consume out of income has to be homogeneous for all the agents (*homothetic* Engel curves) and that distribution is independent from relative prices. This methodology resembles the scientific procedure of the aruspexes, who predicted the future by reading the animals' bowels. The ABM methodology is bottom-up and focuses on the interaction between many heterogenous interacting agents, which might produce a statistical equilibrium, rather than a *natural* one as the mainstream approach assumes. The bottom-up approach models individual behavior according to simple behavioral rules; agents are allowed to have *local interaction* and to change

the *individual rule* (*through adaptation*) as well as the *interaction nodes*. By aggregating, some *statistical regularity* emerges, which cannot be inferred from individual behavior *(self emerging regularities)*: this *emergent behavior* feeds back to the individual level *(downward causation)* thus establishing a macrofoundation of micro. As a consequence, each and every proposition may be falsified at *micro*, *meso* and *macro* levels. This approach opposes the axiomatic theory of economics, where the optimization procedure is the standard for a scientific, i. e. not ad-hoc, modeling procedure.

The agent-based methodology can also be viewed as a way to reconcile the two opposing philosophical perspectives of *methodological individualism* and *holism*. Having agents as the unit of analysis, ABM is deeply rooted in methodological individualism, a philosophical method aimed at explaining and understanding broad society-wide developments as the aggregation of decisions by individuals [13,172]. Methodological individualism suggests – in its most extreme (and erroneous) version – that a system can be understood by analyzing *separately* its constituents, the reductionist approach that the whole is nothing but the sum of its parts [51,127]. However, the ability to reduce everything to simple fundamental objects and laws does not imply the ability to start from those objects and laws and reconstruct the universe. In other terms, reductionism does not imply constructionism [2].

The Austrian school of economics championed the use of methodological individualism in economics in the twentieth century, of which Friederich von Hayek has been one of the main exponents. The legacy of Hayek to ABM and the complex system approach has been recognized [166]. Methodological individualism is also considered an essential part of modern neoclassical economics, with its analysis of collective action in terms of rational, utility-maximizing individuals: should the microfoundations in terms of individual rational behavior be abandoned, the Lucas Critique [114] would kick in. However, it is hard to recognize the imprinting of methodological individualism in the RA paradigm, which claims that the whole society can be analyzed in terms of the behavior of a single, representative, individual and forgets to apply to it the Lucas critique. On the other hand, focusing on aggregate phenomena arising from the bottom up [61] from the interaction of many different agents, ABM also adopts a holistic approach when it claims that these phenomena cannot be studied without looking at the entire context in which they are embedded. Indeed, holism is the idea that all the properties of a given system cannot be determined or explained by the sum of its component parts alone. Instead, the system as a whole determines in an important way that the parts behave. The general principle of holism was concisely summarized by Aristotle in his *Metaphysics*: "The whole is more than the sum of its parts", a *manifesto* of the complexity approach. However, ABM (and more in general complexity theory) should not be confused with general systems theory, an holistic approach developed in the 1950s and 1960s that in its most radical form argued that everything affects everything else: according to systems theory, phenomena that appear to have simple causes, such as unemployment, actually have a variety of complex causes – complex in the sense that the causes are interrelated, nonlinear, and difficult to determine [133]. Conversely, the complexity approach looks for *simple* rules that underpin complexity, an agenda that has been entirely transferred to ABM.

Also, ABM can be thought of as a bridge between methodological individualism and methodological holism. In agent-based models aggregate outcomes (the whole, e. g. the unemployment rate) are computed as the sum of individual characteristics (its parts, e. g. individual employment status). However, aggregate behavior can often be recognized as distinct from the behavior of the comprising agents, leading to the discovery of emergent properties. In this sense, the whole is more than – and different from – the sum of its parts. It might even be the case that the whole appears to act as if it followed a distinct logic, with its own goals and means, as in the example of a cartel of firms that act in order to influence the market price of a good. From the outside, the whole appears no different from a new agent type (e. g. a family, a firm). A new entity is born; the computational experiment has been successful in growing artificial societies from the bottom up [61].

This *bottom-up* approach to complexity consists in deducing the macroscopic objects (*macros*) and their phenomenological complex *ad-hoc* laws in terms of a multitude of elementary microscopic objects (*micros*) interacting by simple fundamental laws [154], and ABM provides a technique that allows one to systematically follow the birth of these complex macroscopic phenomenology. The *macros* at a specific scale can become the *micros* at the next scale.

Depending on the scope of the analysis, it is generally convenient to stop at some scale in the way down to reconstruct aggregate, top-level dynamics from the bottom up. When applied to economics, only a few levels (e. g. a *micro*, a *meso* and a *macro* level) are in general sufficient to provide a thorough understanding of the system. Defining the elementary units of analysis amounts to fixing the limits for the reductionist approach, which is not aprioristically discarded but rather integrated in the analysis. These units

are in fact characterized by an inner structure that does not depend on the environment in which they are embedded. They can thus be analyzed separately.

The need for the ABM approach at any given scale is often linked to the existence of some underlying autocatalytic process at a lower level. *Autocatalytic processes* are dynamic processes with positive feedbacks, where the growth of some quantity is to some extent self-perpetuating, as in the case when it is proportional to its initial value. The importance of positive feedbacks has been recognized in the literature on increasing returns, in particular with respect to the possibility of multiple equilibria [151], since the time of Marshall. However, the traditional analysis is static, and does not address how an equilibrium out of several might be selected. Looking at the problem from a dynamic stochastic process perspective, selection is explained in terms of one set of small historical events magnified by increasing returns.

Moreover, the existence of an autocatalytic process implies that looking at the average, or most probable, behavior of the constituent units is non representative of the dynamics of the system: autocatalyticity insures that the behavior of the entire system is dominated by the elements with the highest auto-catalytic growth rate rather than by the typical or average element [154]. In presence of autocatalytic processes, even a small amount of individual heterogeneity invalidates any description of the behavior of the system in terms of its average element: the real world is controlled as much by the *tails* of distributions as by means or averages. We need to free ourselves from *average* thinking [3].

The fact that autocatalytic dynamics are scale invariant (i. e. after a transformation that multiplies all the variables by a common factor) is a key to understanding the emergence of scale invariant distributions of these variables (e. g. power laws), at an aggregate level. The relevance of scale free distributions in economics (e. g. of firm size, wealth, income, etc.) is now extensively recognized (Brock, 1999), and has been the subject of through investigation in the econophysics literature [120].

Additional Features of Agent-Based Models

We have so far introduced the three fundamental characteristics of ABM: there are agents that play the role of actors, there is no script or *Deus ex-machina* and the story is played live, i. e. it is computed. Following Epstein [58,59,60], we can further characterize the methodology, by enumerating a number of features that, although not necessary to define an agent-based model, are often present. These are:

Heterogeneity

While in analytical models there is a big advantage in reducing the ways in which individuals differ, the computational burden of ABM does not change at all if different values of the parameters (e. g. preferences, endowments, location, social contacts, abilities etc.) are specified for different individuals. Normally, a distribution for each relevant parameter is chosen, and this simply implies that a few parameters (those governing the distribution) are added to the model.

Explicit Space

This can be seen as specification of the previous point: individuals often differ in the physical place where they are located, and/or in the neighbors with whom they can or have to interact (which define the network structure of the model, see below).

Local Interaction

Again, this can be seen as a specification of the network structure connecting the agents. Analytical models often assume either global interaction (as in Walrasian markets), or very simple local interaction. ABM allow for much richer specifications. No direct interaction (only through prices) is allowed in the GE, while direct interaction (*local and stochastic*, usually [101]) is the rule for the complexity approach: figures 1a-c give a graphical representation of Walrasian, random and scale-free interaction respectively. Note that the empirical evidence supports the third case: hubs and power laws are the rule in the real world [38,52].

Actually, some neoclassical economists asked for an analysis of how social relations affect the allocation of resources (e. g., [12,107,134]). They went almost completely unheard, however, until the upsurge in the early 1990s of a brand new body of work aimed at understanding and modeling the social context of economic decisions, usually labeled *new social economics* or *social interaction economics* [56]. Models of social interactions (Manski [118] offers an operational classification of the channels through which the actions of one agent may affect those of other agents within a reference group) are generally able to produce several properties, such as *multiple equilibria* [34]; *non-ergodicity* and *phase transition* [54]; *equilibrium stratification* in social and/or spatial dimension [27,83]; the existence of a *social multiplier* of behaviors [84]. The key idea consists in recognizing that the social relationships in which individual economic agents are embedded can have a large impact on economic decisions. In fact, the social context impacts on individual economic decisions

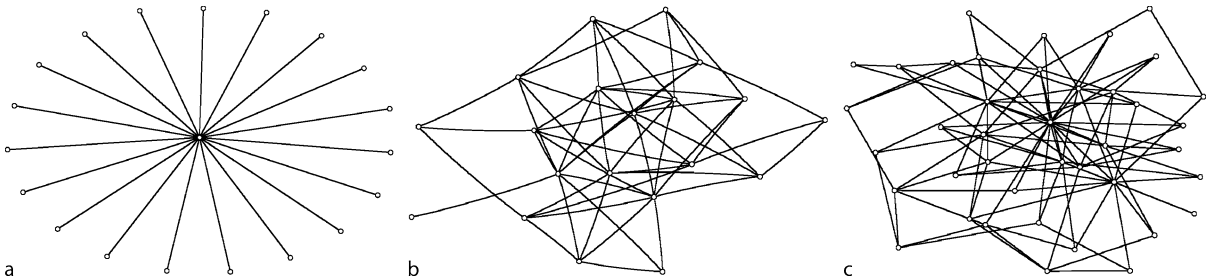

Agent Based Models in Economics and Complexity, Figure 1
a a Walrasian GE representation; b a random graph; c a scale free graph in which several hubs can be identified

through several mechanisms. First, social norms, cultural processes and economic institutions may influence motivations, values, and tastes and, ultimately, make preferences endogenous [31]. Second, even if we admit that individuals are endowed with exogenously given preferences, the pervasiveness of information asymmetries in real-world economies implies that economic agents voluntarily share values, notions of acceptable behavior and socially based enforcement mechanisms in order to reduce uncertainty and favor coordination [50]. Third, the welfare of individuals may depend on some social characteristics like honor, popularity, stigma or status [41]. Finally, interactions not mediated by enforceable contracts may occur because of pure technological externalities in network industries [152] or indirect effects transmitted through prices (pecuniary externalities) in non-competitive markets [28], which may lead to coordination failures due to strategic complementarities [42].

Bounded Rationality

Interestingly, while in analytical models it is generally easier to implement some form of optimal behavior rather than solving models where individuals follow "reasonable" rules of thumb, or learn either by looking at what happened to others or what happened to them in the past, for ABM the opposite is true. However, it can be argued that *real* individuals also face the same difficulties in determining and following the optimal behavior, and are characterized by some sort of bounded rationality: "There are two components of this: bounded information and bounded computing power. Agents have neither global information nor infinite computational capacity. Although they are typically purposive, they are not global optimizers; they use simple rules based on local information" (p. 1588 in [59]).

The requirement on full rationality is indeed very strong, since it requires an infinite computational capacity (the ability of processing tons of data in a infinitesimal amount of time) and all the information. Moreover, according to the mainstream approach, information is complete and free for all the agents. Note that one of the assumptions in the Walrasian approach is that each agent has only private information: this is equivalent to say that *strategic behavior* about information collection and dissemination is ruled out and the collection of the whole set of the information is left to the market *via* the auctioneer (or a benevolent dictator [25]). Indeed, one could read the rational expectation "revolution" as the tentative to decentralize the price setting procedure by defenestrating the auctioneer. Limited information is taken into account, but the constraints have to affect every agent in the same way (the so-called Lucas' islands hypothesis) and the Greenwald–Stiglitz theorem [86] states that in this case the equilibrium is not even Pareto-constrained. If information is asymmetric or private, agents have to be heterogeneous and direct interaction has to be considered: this destroys the mainstream model and generates coordination failures.

On the contrary, agent-based models are build upon the hypothesis that agents have limited information. Once again, the ABM approach is much more parsimonious, since it only requires that the agents do not commit systematic errors. Moreover, given the limited information setting, the economic environment might change affecting, and being affected by, agents' behavior: individuals learn through experience and by interacting with other agents.

Non-equilibrium Dynamics

As we will explain in more details below, ABM are recursive models, in which the state of the system at time $t + 1$ is computed starting from the state at time t. Hence, they allow the investigation of what happens all along the route, not only at the start and at the end of the journey. This point is, we believe, the most important. Brian Arthur (p. 1552 in [16]) offers an effective statement of

its relevance for economic theory: "Standard neoclassical economics asks what agents' actions, strategies, or expectations are in equilibrium with (consistent with) the outcome or pattern these behaviors aggregatively create. Agent-based computational economics enables us to ask a wider question: how agents' actions, strategies or expectations might react to – might endogenously change with – the pattern they create. [...] This out-of-equilibrium approach is not a minor adjunct to standard economic theory; it is economics done in a more general way. [...] The static equilibrium approach suffers two characteristic indeterminacies: it cannot easily resolve among multiple equilibria; nor can it easily model individuals' choices of expectations. Both problems are ones of formation (of an equilibrium and of an 'ecology' of expectations, respectively), and when analyzed in formation – that is, out of equilibrium – these anomalies disappear".

As we have seen, continuous market clearing is *assumed* by the mainstream. It is a necessary condition to obtain "efficiency and optimality" and it is quite curious to read of a theory assuming the *explenandum*. In such a way, every out of equilibrium dynamics or path dependency is ruled out and initial conditions do not matter. The GE model assumes that transactions happen only after the vector of the equilibrium prices has been reached: instead of being the result of the exchange, it foresees it *par tatonnement* in a logical, fictitious time. Because the real markets operate in real, historical, time and the exchange process determines prices, the GE model is not able to describe any real economy [9]. Clower [39] suggested (resemblance Edgeworth, [57]) that exchange might happen out of equilibrium (*at false prices*). In such a case, agents will be quantity-rationed in their supply of-demand for: because of it, the intertemporal maximization problem has to be quantity-constraints (the so-called *Clower constraint*) and if the economy would reach equilibrium, it will be non-optimal and inefficient.

The requirement on rationality is also very strong, since it requires an infinite computational capacity (the ability of processing tons of data in a infinitesimal amount of time) and all the information. In fact, if information is limited, the outcome of a rational choice may be non-optimal. Once again, all the ABM approach is much more parsimonious, since it requires that the agents do not commit systematic errors. Moreover, given the limited information setting, the economic environment might change affecting, and being affected by, agents' behavior: learning and adaptive behavior are therefore contemplated.

Finally, according to Beinhocker [26], the approaches differ also as regard *dynamics* (*Complex Systems* are open, dynamic, non-linear systems, far from equilibrium; *Mainstream* economics are closed, static, linear systems in equilibrium) and *evolution* (*Complex Systems* have an evolutionary process of differentiation, selection and amplification which provides the system with novelty and is responsible for its growth in order and complexity, while *Mainstream* has no mechanism for endogenously creating novelty, or growth in order and complexity.

An Ante Litteram Agent-Based Model: Thomas Schelling's Segregation Model

One of the early and most well known examples of an agent-based model is the segregation model proposed by Thomas Schelling [145,146], who in 2005 received the Nobel prize for his studies in game theory (surveys of more recent applications of ABM to economics can be found in [159,160,161,163]). To correctly assess the importance of the model, it must be evaluated against the social and historical background of the time. Up to the end of the 1960s racial segregation was institutionalized in the United States. Racial laws required that public schools, public places and public transportation, like trains and buses, had separate facilities for whites and blacks. Residential segregation was also prescribed in some States, although it is now widely recognized that it mainly came about through organized, mostly private efforts to ghettoize blacks in the early twentieth century – particularly the years between the world wars [63,126]. But if the social attitude was the strongest force in producing residential segregation, the Civil Rights movement of the 1960s greatly contributed to a change of climate, with the white population exhibiting increasing levels of tolerance. Eventually, the movement gained such strength to achieve its main objective, the abolition of the racial laws: this was sealed in the Civil Rights Act of 1968 which, among many other things, outlawed a wide range of discriminatory conduct in housing markets. Hence, both the general public attitude and the law changed dramatically during the 1960s. As a consequence, many observers predicted a rapid decline in housing segregation. The decline, however, was almost imperceptible. The question then was why this happened. Schelling's segregation model brought an answer, suggesting that small differences in tolerance level or initial location could trigger high level of segregation even without formal (i. e. legal) constraints, and even for decent levels of overall tolerance. In the model, whites and blacks are (randomly) located over a grid, each individual occupying one cell. As a consequence, each individual has at most eight neighbors (Moore neighborhood), located on adjacent cells. Preferences over residential patterns are represented as the maximum quota of racially different

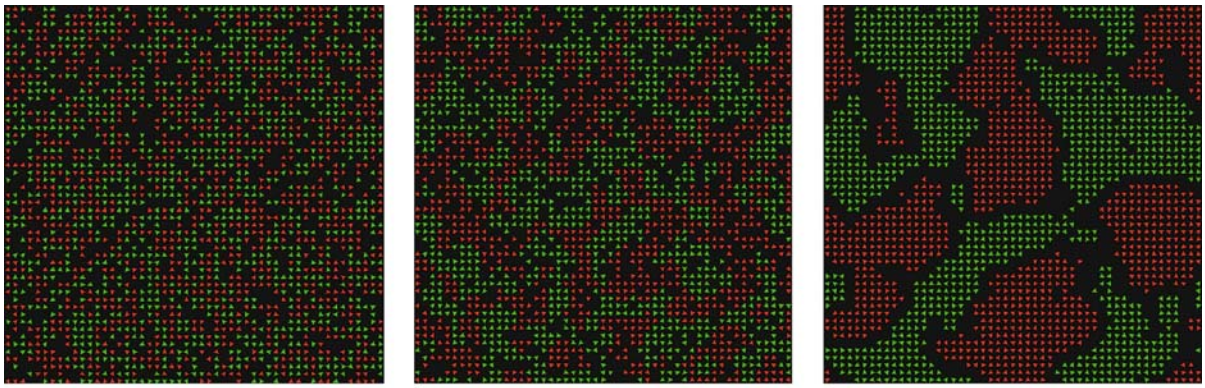

Agent Based Models in Economics and Complexity, Figure 2
NETLOGO implementation of Schelling's segregation model. a Initial (random) pattern. The average share of racially similar neighbors is roughly 50%. With a tolerance level of 70% (40%), less than 20% (more than 80%) of the individuals are not satisfied. b Final pattern. The average share of racially similar neighbors is 72.1%. Everyone is satisfied. c Final pattern. The average share of racially similar neighbors is 99.7%. Everyone is satisfied

neighbors that an individual tolerates. For simplicity, we can assume that preferences are identical: a unique number defines the level of tolerance in the population. For example, if the tolerance level is 50% and an individual has only five neighbors, he would be satisfied if no more than two of his neighbors are racially different. If an individual is not satisfied by his current location, he tries to move to a different location where he is satisfied.

The mechanism that generates segregation is the following. Since individuals are initially located randomly on the grid, by chance there will be someone who is not satisfied. His decision to move creates two externalities: one in the location of origin and the other in the location of destination. For example, suppose a white individual decides to move because there are too many black people around. As he leaves, the ethnic composition of his neighborhood is affected (there is one white less). This increases the possibility that another white individual, who was previously satisfied, becomes eager to move. A similar situation occurs in the area of destination. The arrival of a white individual affects the ethnic composition of the neighborhood, possibly causing some black individual to become unsatisfied. Thus, a small non-homogeneity in the initial residential pattern triggers a chain effect that eventually leads to high levels of segregation. This mechanism is reinforced when preferences are not homogeneous in the population.

Figure 2, which shows the NETLOGO implementation of the Schelling model, exemplifies [173]. The left panel depicts the initial residential pattern, for a population of 2000 individuals, evenly divided between green and red, living on a 51×51 cells torus (hence the population density is 76.9%). Two values for the tolerance threshold are tested: in the first configuration, tolerance is extremely high (70%), while in the second it is significantly lower (30%), although at a level that would still be considered decent by many commentators. The initial residential pattern (obviously) shows no levels of segregation: every individual has on average 50% of neighbors of a different race. However, after just a few periods the equilibrium configurations of the middle (for a tolerance level of 70%) and right (for tolerance level of 30%) panels are obtained. The level of segregation is high: more than three quarters of neighbors are on average of the same racial group, even in case (b), when individuals are actually happy to live in a neighborhood dominated by a different racial group! Moreover, most people live in perfectly homogeneous clusters, with different ethnic clusters being often physically separated from each other by a no man's land. Only the relative mix brought by confining clusters keeps down the measure of overall segregation. Should the overall composition of the population be biased in favor of one ethnic group, we would clearly recognize the formation of ghettoes.

Note that the formation of racially homogeneous ethnic clusters and ghettoes is an emergent property of the system, which could hardly be deduced by looking at individual behavior alone, without considering the effects of interaction. Moreover, the clusters themselves could be considered as the elementary unit of analysis at a different, more aggregate level, and their behavior, e. g. whether they shrink, expand, merge or vanish, studied with respect to some exogenous changes in the environment. Not only a *property*, i. e. a statistical regularity, has emerged, but also a whole new *entity* can be recognized. However, this

new entity is nothing else but a subjective interpretation by some external observer of an emergent property of the system.

The Development of Agent-Based Modeling

The early example of the segregation model notwithstanding, the development of agent-based computational economics is closely linked with the work conducted at the Santa Fe Institute for the study of complexity, a private, non-profit, independent research and education center founded in 1984 in Santa Fe, New Mexico. The purpose of the institute has been, since its foundation, to foster multidisciplinary collaboration in pursuit of understanding the common themes that arise in natural, artificial, and social systems. This unified view is the dominant theme of what has been called the *new science of complexity*.

The outcomes of this research program are well depicted in three books, all bearing the title *The economy as an evolving complex system* [4,17,29]. The following quotation, from the preface of the 1997 volume, summarizes very accurately the approach:

> "In September 1987 twenty people came together at the Santa Fe Institute to talk about 'the economy as a evolving, complex system'. Ten were theoretical economists, invited by Kenneth J. Arrow, and ten were physicists, biologists and computer scientists, invited by Philip W. Anderson. The meeting was motivated by the hope that new ideas bubbling in the natural sciences, loosely tied together under the rubric of 'the sciences of complexity', might stimulate new ways of thinking about economic problems. For ten days, economists and natural scientists took turns talking about their respective worlds and methodologies. While physicists grappled with general equilibrium analysis and non-cooperative game theory, economists tried to make sense of spin glass models, Boolean networks, and genetic algorithms. The meeting left two legacies. The first was the 1988 volume of essays; the other was the founding, in 1988, of the Economics Program at the Santa Fe Institute, the Institute's first resident research program. The Program's mission was to encourage the understanding of economic phenomena from a complexity perspective, which involved the development of theory as well as tools for modeling and for empirical analysis. [...] But just what is the complexity perspective in economics? That is not an easy question to answer. [...] Looking back over the developments in the past decade, and of the papers produced by the program, we believe that a coherent perspective – sometimes called the 'Santa Fe approach' – has emerged within economics."

The work carried out at the Santa Fe Institute greatly contributed to popularize the complexity approach to economics, although a similar line of research was initiated in Europe by chemists and physicists concerned with emergent structures and disequilibrium dynamics (more precisely, in Brussels by the group of the Nobel prize winner physical chemist Ilya Prigogine ([128]) and in Stuttgart by the group of the theoretical physicist Hermann Haken [91], as discussed in length by Rosser [141]).

Two main reasons can help explaining why the Santa Fe approach gained some visibility outside the restricted group of people interested in the complexity theory (perhaps contributing in this way to mount what Horgan [96,97], called an intellectual fad). Together, they offered an appealing suggestion of both what to do and how to do it. The first reason was the ability to present the complexity paradigm as a unitary perspective. This unitary vision stressed in particular the existence of feedbacks between *functionalities* and *objectives*: individual objectives determine to some extent the use and modification of existing functionalities, but functionalities direct to some extent the choice of individual objectives. It is this analytical focus that proved to be valuable in disciplines as diverse as the social sciences, the biological sciences and even architecture [70]. The second reason has to do with the creation of a specific simulation platform that allowed relatively inexperienced researchers to build their own toy models that, thanks to the enormous and sustained increase in commonly available computing power, could run quickly even on small PCs. This simulation platform was called SWARM [18], and consisted in a series of libraries that implemented many of the functionalities and technicalities needed to build an agent-based simulation, e. g. the schedule of the events, the passing of time and graphical widgets to monitor the simulation. In addition to offering a practical tool to write agent-based simulations, the SWARM approach proposed a protocol in simulation design, which the SWARM libraries exemplified.

The principles at the basis of the SWARM protocol are:

(i) The use of an *object-oriented programming* language (SWARM was first written in OBJECTIVE C, and later translated into JAVA), with different objects (and object types) being a natural counterpart for different agents (and agent types);
(ii) A *separate implementation* of the model and the tools used for monitoring and conducting experiments on the model (the so-called observer);

(iii) An architecture that allows nesting models one into another, in order to build a *hierarchy* of swarms – a swarm being a group of objects and a schedule of actions that the objects execute. One swarm can thus contain lower-level swarms whose schedules are integrated into the higher-level schedule.

A number of different simulation platforms that adhered to the SWARM protocol for simulation design have been proposed since, the most widespread being REPAST ([129]; see also [135]). However, other alternative approaches to writing agent-based models exist. Some rely on general-purpose mathematical software, like MATHEMATICA, MATLAB or MATCAD. Others, exemplified by the STARLOGO/NETLOGO experience [137], are based on the idea of an agent-based specific language.

Finally, despite the fact that ABM are most often computer models, and that the methodology could not develop in the absence of cheap and easy-to-handle personal computers, it is beneficial to remember that one of the most well-known agent-based models, the segregation model we have already described, abstracted altogether from the use of computers. As Schelling recalls, he had the original idea while seated on plane, and investigated it with paper and pencil. When he arrived home, he explained the rules of the game to his son and got him to move zincs and coppers from the child's own collection on a checkerboard, looking for the results: The dynamics were sufficiently intriguing to keep my twelve-year-old engaged p. 1643 in [148].

A Recursive System Representation of Agent-Based Models

Although the complexity theory is, above all, a mathematical concept, a rather common misunderstanding about agent-based simulations is that they are not as sound as mathematical models. In an often-quoted article, Thomas Ostrom [130] argued that *computer simulation* is a third symbol system in its own right, aside *verbal description* and *mathematics*: simulation is no mathematics at all (see [79]). An intermediate level of abstraction, according to this view, characterizes computer simulations: they are more abstract than verbal descriptions, but less abstract than pure mathematics. Ostrom (p. 384 in [130]) also argued that any theory that can be expressed in either of the first two symbol systems can also be expressed in the third symbol system. This implies that there might be verbal theories, which cannot be adequately expressed in the second symbol system of mathematics, but can be in the third [79].

This view has become increasingly popular among social simulators themselves, apparently because it offers a shield to the perplexity of the mathematicians, while hinting at a sort of superiority of computer simulations. Our opinion is that both statements are simply and plainly wrong. Agent-based modeling – and more in general simulation – *is* mathematics, as we argue in this paragraph. Moreover, the conjecture that any theory can be expressed via simulation is easily contradicted: think for instance of simulating Hegel's philosophical system.

Actually, agent-based simulations are nothing else but recursive systems [59,110], where the variables s that describe at time t the state of each individual unit are determined, possibly in a stochastic way, as a function of the past states s and some parameters a:

$$s_{i,t} = f_i(s_{i,t-1},\, s_{-i,t-1};\, a_i,\, a_{-i}\,;\, t) \tag{1}$$

The individual state variables could include the memory of past values, as in the case when an unemployed person is characterized not only by the fact that he is unemployed, but also by when he last had a job. The function f_i and the parameters a_i determine individual behavior. They can possibly change over time, either in a random way or depending on some lagged variable or on higher-order parameters (as in the *Environment-Rule-Agent* framework of Gilbert and Terna [80]); when this is the case, their expression can simply be substituted for in Eq. (1). Equation (1) allows the recursive computation of the system: at any moment in time the state of each unit can be expressed as a (possibly stochastic) function of the initial values X_0 only, where X_0 includes the initial states and parameters of all the individual units:

$$s_{i,t} = g_i(X_0;\, t) \tag{2}$$

The aggregate state of the system is simply defined as

$$S_t = \sum_i s_{i,t} \tag{3}$$

Equilibrium in this system is described as a situation where the aggregate state S, or some other aggregate statistics Y computed on the individual states or the individual parameters are stationary.

Notice that this formalization describes both traditional dynamic micro models and agent-based simulations. In principle an agent-based model, not differently from traditional dynamic micro models, *could* be solved analytically. The problem is that the expressions involved quickly become unbearable, as (i) the level of heterogeneity, as measured by the distribution of the parame-

ters a_i and functional forms f_i, increases; (ii) the amount of interaction, as measured by the dependency of $s_{i,t}$ on $s_{-i,t-1}$, increases; (iii) the functional forms f become more complicated, e. g. with the introduction of *if-else* conditions, etc.

Hence, the resort to numerical simulation. Traditional analytical models on the other hand must take great care that the system can be solved analytically, i. e. by symbolic manipulation. Hence the use of simple functions as the omnipresent Cobb–Douglas, the assumption of homogeneous units (that can then be replaced by a RA), the choice of simple interaction processes, often mediated by a centralized coordination mechanism. However, analytical tractability alone is a poor justification of any modeling choice. As the Nobel laureate Harry Markowitz wrote, "if we restrict ourselves to models which can be solved analytically, we will be modeling for our mutual entertainment, not to maximize explanatory or predictive power" (as reported in [112]). Restricting to analytically solvable *modls* – as they are called in the not sufficiently well-known paper by Axel Leijonhufvud [108] – looks dangerously close to the tale of the man who was searching for his keys under the light of a street lamp at night and, once asked if he had lost them there, he answered "No, but this is where the light is".

Analysis of Model Behavior

Being able to reach a close solution means that it is possible to connect inputs and outputs of the model, at any point in time, in a clear way: the *input-output transformation function*, or *reduced form*, implied by the *structural form* in which the model is expressed, is analytically obtained (e. g. the equilibrium expression of some aggregate variable of interest, as a function of the model parameters). Hence, theorems can be proved and laws expressed.

On the contrary, in a simulation model the reduced form remains unknown, and only *inductive* evidence about the input/output transformation implied by the model can be collected. Performing multiple runs of the simulation with different parameters does this. In other words, simulations suffer from the problem of stating general propositions about the dynamics of the model starting only from point observations. Since scientific explanations are generally defined as the derivation of general laws, which are able to replicate the phenomena of interests [93,94], simulations appear to be less scientific than analytical models. As Axelrod [19] points out, "like deduction, [a simulation] starts with a set of explicit assumptions. But unlike deduction, it does not prove theorems. Instead, a simulation generates data that can be analyzed inductively". Induction comes at the moment of explaining the behavior of the model. It should be noted that although induction is used to obtain knowledge about the behavior of a given simulation model, the use of a simulation model to obtain knowledge about the behavior of the real world refers to the logical process of abduction [109,117]. Abduction [66,132], also called *inference to the best explanation*, is a method of reasoning in which one looks for the hypothesis that would best explain the relevant evidence, as in the case when the observation that the grass is wet allows one to suppose that it rained.

Being constrained to unveil the underlying input-output transformation function by repetitively sampling the parameter space, simulations cannot prove necessity, i. e. they cannot provide in the traditional sense necessary conditions for any behavior to hold. This is because nothing excludes a priori that the system will behave in a radically different way as soon as the value of some parameter is changed, while it is generally not possible to sample all values of the parameter space. In other words, the artificial data may not be representative of all outcomes the model can produce. While analytical results are conditional on the specific hypothesis made about the *model* only, simulation results are conditional both on the specific hypothesis of the model and the specific values of the *parameters* used in the simulation runs: each run of such a model yields is a sufficiency theorem, [yet] a single run does not provide any information on the robustness of such theorems [20].

The sampling problem becomes increasingly harder as the number of the parameters increase. This has been referred to as *the curse of dimensionality* [143]. To evaluate its implications, two arguments should be considered. The first one is theoretical: if the impossibility to gain a full knowledge of the system applies to the *artificial* world defined by the simulation model, it also applies to the *real* world. The real data generating process being itself unknown, stylized facts (against which all models are in general evaluated) could in principle turn wrong, at some point in time. From an epistemological point of view, our belief that the sun will rise tomorrow remains a probabilistic assessment. The second, and more decisive, consideration is empirical: we should not worry too much about the behavior of a model for particular evil combinations of the parameters, as long as these combinations remain extremely rare (one relevant exception is when rare events are the focus of the investigation, e. g. in risk management, see [150]). If the design of the experiments is sufficiently accurate, the problem of how imprecise is the estimated input-output transformation function becomes marginal:

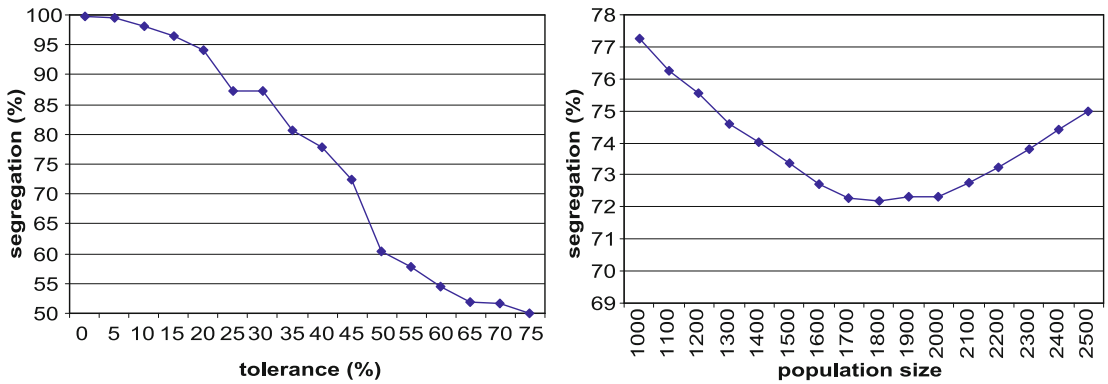

Agent Based Models in Economics and Complexity, Figure 3
Sensitivity analysis for the Schelling's segregation model. Segregation is measured as the share of racially similar neighbors. The reference parameter configuration is population size = 2000, tolerance level = 40%

While the curse of dimensionality places a practical upper bound on the size of the parameter space that can be checked for robustness, it is also the case that vast performance increases in computer hardware are rapidly converting what was once perhaps a fatal difficulty into a manageable one [20].

In conclusion, extensive experimentation is the only way to get a full understanding of the simulation behavior. Sampling of the parameter space can be done either systematically, i. e. by grid exploration, or randomly. Following Leombruni et al. [111], we can further distinguish between two levels at which sampling can be done: a *global* level and a *local* level. Local sampling is conducted around some specific parameter configurations of interest, by letting each parameter vary and keeping all the others unchanged. This is known as *sensitivity analysis*, and is the equivalent to the study of the partial derivatives of the input-output transformation function in an analytical model.

As an example, Fig. 3 reports a plot of the equilibrium level of segregation in the Schelling model, for decreasing values of tolerance (left panel) and increasing population density (right panel). Tolerance level is sampled in the range $[0, .7]$ by increasing steps of .05, while population size is sampled in the range $[1000, 2000]$ by increasing steps of 100. To get rid of random effects (in the initial residential pattern and in the choice of a different location of unsatisfied individuals), 100 runs are performed for every value of the parameter being changed, and average outcomes are reported. This gives an idea of the local effects of the two parameters around the central parameter configuration where the population size is equal to 2000 and the tolerance level is equal to 70%.

For what concerns the effect of tolerance on segregation (left panel), it should be noted that the somewhat irregular shape of the relationship is a consequence not of the sample size but of the small neighborhood individuals take into consideration (a maximum of eight adjacent cells, as we have seen), and the discretization it brings. As the effect of population size on segregation (right panel) is concerned, it may seem at a first glance counter-intuitive that segregation initially diminishes, as the population density increases. This is due to the fact that clusters can separate more if there are more free locations. Of course, nothing excludes the possibility that these marginal effects are completely different around a different parameter configuration. To check whether this is the case, it is necessary either to repeat the sensitivity analysis around other configurations, or to adopt a multivariate perspective.

Allowing all parameters to change performs *global sampling*, thus removing the reference to any particular configuration. To interpret the results of such a global analysis, a relationship between inputs and outputs in the *artificial* data can be estimated, e. g.:

$$Y = m(X_0) \,. \tag{4}$$

Where Y is the statistics of interest (say, the Gini coefficient of wealth), computed in equilibrium, i. e. when it has reached stationarity, and X_0 contains the initial conditions and the structural parameters of the model: $X_0 = \{s_0, a\}$. If the (not necessary unique) steady state is independent of the initial conditions, Eq. 4 simplifies to:

$$Y = m(A) \,. \tag{5}$$

Where A contains only the parameters of the simulation. The choice of the functional form m to be estimated, which

Agent Based Models in Economics and Complexity, Table 1

Regression results for Schelling's segregation model. Instead of repeating the experiment *n* times for each parameter configuration, in order to average out the random effects of the model, we preferred to test a number of different parameter configurations *n* times higher. Thus, population size is explored in the range [1000, 2000] by increasing steps of 10, and tolerance level is explored in the range [0, 7] by increasing steps of .05

Source	SS	df	MS
Model	666719.502	6	111119.917
Residual	12033.9282	2108	5.70869461
Total	678753.43	2114	321.075416

Number of obs = 2115
F(6, 2108) = 19465.03
Prob > F = 0.0000
R-squared = 0.9823
Adj R-squared = 0.9822
Root MSE = 2.3893

Segregation	Coef.	Std. Err.	t	$P > \|t\|$	[95% Conf. Interval]	
tolerance	3.379668	.0819347	41.25	0.000	3.218987	3.54035
tolerance_2	− .0655574	.0013175	− 49.76	0.000	− .0681411	− .0629737
tolerance_3	.0003292	6.73e− 06	48.94	0.000	.000316	.0003424
density	− 23.83033	3.274691	− 7.28	0.000	− 30.25229	− 17.40837
density_2	20.05102	2.372174	8.45	0.000	15.39897	24.70306
interaction	− .1745321	.0153685	− 11.36	0.000	− .2046712	− .144393
_cons	57.31189	1.957341	29.28	0.000	53.47336	61.15041

is sometimes referred to as *metamodel* [103] is to a certain extent arbitrary, and should be guided by the usual criteria for model specification for the analysis of real data.

As an example, we performed a multivariate analysis on the artificial data coming out of the Schelling's segregation model, by letting both the population size and the tolerance threshold to vary. Overall, 2115 parameter configurations are tested. After some data mining, our preferred specification is an OLS regression of the segregation level on a third order polynomial of the *tolerance* threshold, a second order polynomial of population *density*, plus an *interaction* term given by the product of tolerance and density. The interaction term, that turns out to be highly significant, implies that the local analysis of Fig. 3 has no general validity.

The regression outcome is reported in Table 1.

Such a model allows predicting the resulting segregation level for any value of the parameters. Of course, as the complexity of the model increases (e. g. leading to multiple equilibria) finding an appropriate meta-model becomes increasingly arduous.

Finally, let's remark that the curse of dimensionality strongly suggests that the flexibility in model specification characterizing agent-based models is to be used with care, never neglecting the KISS (*Keep it simple, Stupid*) principle. Schelling's segregation model is in this respect an example of simplicity, since it has but a few parameters: this is not incoherent with the complexity approach, since it stresses how simple behavioral rules can generate very complex dynamics.

Validation and Estimation

The previous section has dealt with the problem of interpreting the behavior of an agent-based model, and we have seen that this can be done by appropriately generating and analyzing *artificial* data. We now turn to the relationship between artificial and *real* data, that is (i) the problem of choosing the parameter values in order to have the behavior of the model being as close as possible to the real data, and (ii) the decision whether a model is good enough, which often entails a judgment on "how close" as close as possible is. The first issue is referred to as the problem of *calibration* or *estimation* of the model, while the second one is known as *validation.*

Note that all models have to be understood. Thus, for agent-based models analysis of the artificial data is always an issue. However, not all models have to be estimated or validated. Some models are built with a theoretical focus (e. g. Akerlof's market for lemons), and thus comparison with the real data is not an issue – although it could be argued that some sort of evaluation is still needed, although of a different kind.

Estimation

Although the terms calibration and estimation are sometimes given slightly different meanings (e. g. [105]), we agree with Hansen and Heckman (p. 91 in [92]) that "the distinction drawn between calibrating and estimating the parameters of a model is artificial at best. Moreover, the justification for what is called *calibration* is vague and con-

fusing. In a profession that is already too segmented, the construction of such artificial distinctions is counterproductive."

Our understanding is that, too often, calibration simply refers to a sort of rough estimation, e. g. by means of visual comparison of the artificial and real data. However, not all parameters ought to be estimated by means of formal statistical methods. Some of them have very natural real counterparts and their value is known (e. g. the interest rate): the simulation is run with empirical data. Unknown parameters have on the other had to be properly estimated.

In analytical models the reduced form coefficients, e. g. the coefficients linking output variables to inputs, can be estimated in the real data. If the model is identified, there is a one-to-one relationship between the structural and the reduced form coefficients. Thus, estimates for the structural coefficients can be recovered. In a simulation model this can't be done. However, we could compare the outcome of the simulation with the real data, and change the structural coefficient values until the distance between the simulation output and the real data is minimized. This is called *indirect inference* [85], and is also applied to analytical models e. g. when it is not possible to write down the likelihood. There are many ways to compare real and artificial data. For instance, simple statistics can be computed both in real and in artificial data, and then aggregated in a unique measure of distance. Clearly, these statistics have to be computed just once in the real data (which does not change), and once every iteration until convergence in the artificial data, which depends on the value of the structural parameters. The change in the value of the parameters of each iteration is determined according to some optimization algorithm, with the aim to minimize the distance.

In the *method of simulated moments* different order of moments are used, and then weighted to take into account their uncertainty (while the uncertainty regarding the simulated moments can be reduced by increasing the number of simulation runs, the uncertainty in the estimation of the real, population moment on the basis of real sample data cannot be avoided). The intuition behind this is to allow parameters estimated with a higher degree of uncertainty to count less, in the final measure of distance between the real and artificial data [174]. Having different weights (or no weights at all) impinges on the efficiency of the estimates, not on their consistency. If the number of moments is equal to the number of structural parameters to be estimated, the model is just-identified and the minimized distance, for the estimated values of the parameters, is 0. If the number of moments is higher than the number of parameters the model is over-identified and the minimized distance is greater than 0. If it is lower it is under-identified. Another strategy is to estimate an *auxiliary model* both in the real and in the artificial data, and then compare the two sets of estimates obtained. The regression coefficients have the same role as the moments in the method of simulated moments: they are just a way of summarizing the data. Hence, if the number of coefficients in the auxiliary model is the same as the number of structural parameters to be estimated the model is just-identified and the minimized distance is 0. The specification of the auxiliary model is not too important. It can be proved that misspecification (e. g. omission of a relevant variable in the relationship to be estimated) only affects efficiency, while the estimates of the structural parameters remain consistent. A natural choice is of course the meta-model of Eq. 4.

Validation

A different issue is determining "how good" a model is. Of course, an answer to this question cannot be unique, but must be made in respect to some evaluation criterion. This in turn depends on the objectives of the analysis [62,102,111,164]. The need for evaluation of the model is no different in agent-based models and in traditional analytical models. However, like all simulations agent-based models require an additional layer of evaluation: the validity of the simulator (the program that simulates) relative to the model (*program validity*).

Assuming this is satisfied and the program has no bugs, Marks [121] formalizes the assessment of the *model validity* as follows: the model is said to be *useful* if it can exhibit at least some of the observed historical behaviors, *accurate* if it exhibits only behaviors that are compatible with those observed historically, and *complete* if it exhibits all the historically observed behaviors. In particular, letting $\mathbf{R}$ be the real world output, and $\mathbf{M}$ be the model output, four cases are possible:

a. No intersection between $\mathbf{R}$ and $\mathbf{M}$ ($\mathbf{R} \cap \mathbf{M} = \emptyset$): the model is *useless*;
b. $\mathbf{M}$ is a subset of $\mathbf{R}$ ($\mathbf{M} \subset \mathbf{R}$): the model is accurate, but *incomplete*;
c. $\mathbf{R}$ is a subset of $\mathbf{M}$ ($\mathbf{M} \supset \mathbf{R}$): the model is complete, but *inaccurate* (or *redundant*, since the model might tell something about what could yet happen in the world);
d. $\mathbf{M}$ is equivalent to $\mathbf{R}$ ($\mathbf{M} \Leftrightarrow \mathbf{R}$): the model is *complete* and *accurate*.

Of course, the selection of the relevant historical behaviors is crucial, and amounts to defining the criteria against which the model is to be evaluated. Moreover, the recogni-

tion itself of historical behavior passes through a process of analysis and simplification that leads to the identification of *stylized facts*, which are generally defined in stochastic terms. Thus, a model is eventually evaluated according to the extent to which it is able to statistically replicate the selected stylized facts.

Finally, let's note that the behavior of the model might change significantly for different values of the parameters. Hence, the process of validation always regards both the *structure* of the model and the *values* of the parameters. This explains why and how validation and estimation are connected: as we have already noted, estimation is an attempt to make the behavior of the model as close as possible to real behavior; validation is a judgment on how far the two behaviors (still) are. A model where the parameters have not been properly estimated and are e. g. simple guesses can of course be validated. However, by definition its performance can only increase should the values of the parameters be replaced with their estimates.

The Role of Economic Policy

Before *economics* was *political economy*. According to the classical economists, the economic science has to be used to control the real economies and steer them towards desirable outcomes. If one considers the economic system as an analogue of the physical one, it is quite obvious to look for *natural* economic policy prescriptions (*one policy fits all*). This is the approach of mainstream (neoclassical) economists. There is a widespread opinion, well summarized by Brock and Colander [33], that, with respect to the economic policy analysis of the mainstream, (i) complexity does not add anything new to the box of tools. This point needs substantial corrections (see also the reflections by Durlauf [3]). The complexity approach showed us that the age of certainty ended with the non-equilibrium revolution, exemplified by the works of Prigogine. Considering the economy as an evolving (adaptive) system we have to admit that our understanding of it is limited (there is no room for *Laplace' demon* in complexity). Individual behavioral rules evolve according to their past performance: this provides a mechanism for an endogenous change of the environment. As a consequence the rational expectation hypothesis loses significance. However, agents are still rational in that they do what they can in order not to commit systematic errors [113]. In this setting there is still room for policy intervention outside the mainstream myth of a neutral and optimal policy. Because emergent facts are transient phenomena, policy recommendations are less certain, and they should be institution and historically oriented [65,170]. In particular, it has been emphasized that complex systems can either be extremely fragile and turbulent (a slight modification in some minor detail brings macroscopic changes), or relatively robust and stable: in such a context, policy prescriptions ought to be case sensitive.

In a heterogenous interacting agents environment, there is also room for an extension of the Lucas critique. It is well known that, according to it, because the underlying parameters are not policy-invariant any policy advice derived from large-scale econometric models that lack microfoundations would be misleading. The Lucas Critique implies that in order to predict the effect of a policy experiment, the so-called *deep parameters* (*preferences, technology* and *resource constraints*) that govern individual behavior have to be modeled. Only in this case it is possible to predict the behaviors of individuals, conditional on the change in policy, and aggregate them to calculate the macroeconomic outcome. But here is the trick: aggregation is a sum only if interaction is ignored. If non-price interactions (or other non-linearities) are important, then the interaction between agents may produce very different outcomes. Mainstream models focus on analytical solvable solutions: to get them, they have to simplify the assumptions e. g. using the RA approach or a Gaussian representation of heterogeneity. At the end, the main objective of these models is to fit the theory, not the empirics: how to explain, e. g., the scale-free network of the real economy represented in Fig. 1c by using the non interacting network of the mainstream model of Fig. 1a? At a minimum, one should recognize that the mainstream approach is a very primitive framework and, as a consequence, the economic policy recommendations derivable from it are very far from being adequate prescriptions for the real world.

Real economies are composed by millions of interacting agents, whose distribution is far from being stochastic or normal. As an example, Fig. 4 reports the distribution of the firms' trade-credit relations in the electronic-equipment sector in Japan in 2003 (see [47]). It is quite evident that there exist several hubs, i. e. firms with many connections: the distribution of the degree of connectivity is *scale free*, i. e. there are a lot of firms with one or two links, and very a few firms with a lot of connections. Let us assume the Central Authority has to prevent a *financial collapse* of the system, or the spreading of a financial crisis (the so-called *domino effect*, see e. g. [104] and [157]). Rather than looking at the average risk of bankruptcy (in power law distributions the mean may even not exist, i. e. there is an empirical mean, but it is not stable), and to infer it is a measure of the stability of the system, by means of a network analysis the economy can be analyzed in terms of differ-

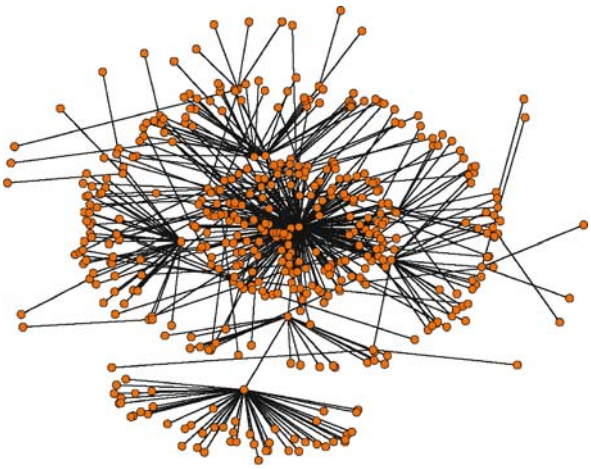

Agent Based Models in Economics and Complexity, Figure 4
Network of firms (electrical machinery and other machines sector, Japan). Source: De Masi et al. [47]

ent interacting sub-systems, and local intervention can be recommended to prevent failures and their spread.

Instead of a helicopter drop of liquidity, one can make targeted interventions to a given agent or sector of activity: Fujiwara, [72], show how to calculate the probability of going bankrupt by *solo*, i. e. because of idiosyncratic elements, or *domino* effect, i. e. because of the failure or other agents with which there exist credit or commercial links.

One of the traditional fields of applications of economic policy is *redistribution*. It should be clear that a sound policy analysis requires a framework built without the RA straight jacket. A redistributive economic policy has to take into account that individuals are different: not only they behave differently, e. g. with respect to saving propensities, but they also have different fortunes: the so-called *St.Thomas* (13:12) effect (*to anyone who has, more will be given and he will grow rich; from anyone who has not, even what he has will be taken away*), which is the road to Paradise for Catholics, and to the power-law distribution of income and wealth for the econophysicists.

Gaffeo et al. [75], show that there is a robust link between firms' size distribution, their growth rate and GDP growth. This link determines the distributions of the amplitude frequency, size of recessions and expansion *etc.* Aggregate firms' size distribution can be well approximated by a *power law* [21,74], while sector distribution is still right skewed, but without scale-free characteristics [22]. Firms' growth rates are far from being *normal*: in the central part of the distribution they are tent shaped with very fat tails. Moreover, empirical evidence shows that exit is an inverse function of firms' age and size and proportional to financial fragility. In order to reduce the volatility of fluctuations, policy makers should act on the firms' size distribution, allowing for a growth of their capitalization, their financial solidity and wealth redistribution [48,49]. Since these emerging facts are policy sensitive, if the aggregate parameters change the shape of the curve will shift as well.

Differently from Keynesian economic policy, which theorizes aggregate economic policy tools, and mainstream neoclassical economics, which prescribes individual incentives because of the Lucas critique but ignores interaction which is a major but still neglected part of that critique, the ABM approach proposes a bottom up analysis. What generally comes out is not a one-size-fits-all policy since it depends on the general as well as the idiosyncratic economic conditions; moreover, it generally has to be conducted at different levels (from micro to meso to macro). In short, ABM can offer new answers to old unresolved questions, although it is still in a far too premature stage to offer definitive tools.

Future Directions

We have shown that mainstream approach to economics uses a methodology [71], which is so weak in its assumptions as to have been repeatedly ridiculed by the epistemologists [37], and dates back to the classical mechanical approach, according to which reductionism is possible. We have also seen that adopting the reductionist approach in economics is to say that agents *do not interact directly*: this is a very implausible assumption (billions of Robinson Crusoes who never meet Friday) and cannot explain the *emerging* characteristics of our societies, as witnessed by the empirical evidence. The reductionist approach of the mainstream is also theoretically incoherent, since it can be given no sound microfoundations [8,100].

In the fourth edition of his *Principles*, Marshall wrote, "The Mecca of the economist is biology". What he meant to say was that, because economics deals with learning agents, evolution and change are the *granum salis* of our economic world. A theory built upon the issue of allocations of given quantities is not well equipped for the analysis of change. This allocation can be optimal only if there are no externalities (increasing returns, non-price interactions etc.) and information is complete, as in the case of the *invisible hand* parabola. In the history of science, there is a passage from a view emphasizing centralized intelligent design to a view emphasizing *self organized criticality* [27], according to which a system with many heterogenous interacting agents reaches a statistical aggregate equilibrium, characterized by the appearance of some (often scale free) stable distributions. These distributions are no longer op-

timal or efficient according to some welfare criterion: they are simply the natural outcome of individual interaction.

Because of the above-mentioned internal and external inconsistencies of the mainstream approach, a growing strand of economists is now following a different methodology based upon the analysis of systems with many heterogenous interacting agents. Their interaction leads to empirical regularities, which emerge from the system as a whole and cannot be identified by looking at any single agent in isolation: these emerging properties are, according to us, the main distinguishing feature of a complex system. The focus on interaction allows the scientist to abandon the heroic and unrealistic RA framework, in favor of the ABM approach, the *science of complexity* popularized by the SFI. Where did the Santa Fe approach go? Did it really bring a revolution in social science, as some of its initial proponents ambitiously believed? Almost twenty years and two "The economy as an evolving complex system" volumes later, Blume and Durlauf summarized this intellectual Odyssey as follows:

> "On some levels, there has been great success. Much of the original motivation for the Economics Program revolved around the belief that economic research could benefit from an injection of new mathematical models and new substantive perspectives on human behavior. [...] At the same time, [...] some of the early aspirations were not met" (Chaps. 1–2 in [29]).

It is probably premature to try to give definitive answers. For sure, ABM and the complexity approach are a very tough line of research whose empirical results are very promising (see e. g., Chaps. 2–3 in [77]). Modeling an agent-based economy however remains in itself a complex and complicated adventure.

Bibliography

1. Allen PM, Engelen G, Sanglier M (1986) Towards a general dynamic model of the spatial evolution of urban systems. In: Hutchinson B, Batty M (eds) Advances in urban systems modelling. North-Holland, Amsterdam, pp 199–220
2. Anderson PW (1972) More is different. Science 177:4047
3. Anderson PW (1997) Some thoughts about distribution in economics. In: Arthur WB, Durlaf SN, Lane D (eds) The economy as an evolving complex system II. Addison-Wesley, Reading
4. Anderson PW, Arrow K, Pines D (eds) (1988) The economy as an evolving complex system. Addison-Wesley, Redwood
5. Aoki M (1996) New approaches to macroeconomic modelling: evolutionary stochastic dynamics, multiple equilibria, and externalities as field effects. Cambridge University Press, Cambridge
6. Aoki M (2002) Modeling aggregate behaviour and fluctuations in economics. Cambridge University Press, Cambridge
7. Aoki M, Yoshikawa H (2006) Reconstructing macroeconomics. Cambridge University Press, Cambridge
8. Aoki M, Yoshikawa H (2007) Non-self-averaging in economic models. Economics Discussion Papers No. 2007-49, Kiel Institute for the World Economy
9. Arrow KJ (1959) Towards a theory of price adjustment. In: Abramovits M (ed) Allocation of economic resources. Stanford University Press, Stanford
10. Arrow KJ (1963) Social choice and individual values, 2nd edn. Yale University Press, New Haven
11. Arrow KJ (1964) The role of securities in the optimal allocation of risk-bearing. Rev Econ Stud 31:91–96
12. Arrow KJ (1971) A utilitarian approach to the concept of equality in public expenditures. Q J Econ 85(3):409–415
13. Arrow KJ (1994) Methodological individualism and social knowledge. Am Econ Rev 84:1–9
14. Arrow KJ, Debreu G (1954) Existence of an equilibrium for a competitive economy. Econometrica 22:265–290
15. Arthur WB (2000) Complexity and the economy. In: Colander D (ed) The complexity vision and the teaching of economics. Edward Elgar, Northampton
16. Arthur WB (2006) Out-of-equilibrium economics and agent-based modeling. In: Tesfatsion L, Judd KL (eds) Handbook of computational economics, vol 2: Agent-Based Computational Economics, ch 32. North-Holland, Amsterdam, pp 1551–1564
17. Arthur WB, Durlauf S, Lane D (eds) (1997) The economy as an evolving complex system II. Addison-Wesley, Reading
18. Askenazi M, Burkhart R, Langton C, Minar N (1996) The swarm simulation system: A toolkit for building multi-agent simulations. Santa Fe Institute, Working Paper, no. 96-06-042
19. Axelrod R (1997) Advancing the art of simulation in the social sciences. In: Conte R, Hegselmann R, Terna P (eds) Simulating social phenomena. Springer, Berlin, pp 21–40
20. Axtell RL (2000) Why agents? On the varied motivations for agent computing in the social sciences. Proceedings of the Workshop on Agent Simulation: Applications, Models and Tools. Argonne National Laboratory, Chicago
21. Axtell RL (2001) Zipf distribution of US firm sizes. Science 293:1818–1820
22. Axtell RL, Gallegati M, Palestrini A (2006) Common components in firms' growth and the scaling puzzle. Available at SSRN: http://ssrn.com/abstract=1016420
23. Bak P (1997) How nature works. The science of self-organized criticality. Oxford University Press, Oxford
24. Batten DF (2000) Discovering artificial economics. Westview Press, Boulder
25. Barone E (1908) Il ministro della produzione nello stato collettivista. G Econ 267–293, 391–414
26. Beinhocker ED (2006) The origin of wealth: Evolution, complexity, and the radical remaking of economics. Harvard Business School Press, Cambridge
27. Bénabou R (1996) Heterogeneity, stratification and growth: Macroeconomic implications of community structure and school finance. Am Econ Rev 86:584–609
28. Blanchard OJ, Kiyotaki N (1987) Monopolistic competition and the effects of aggregate demand. Am Econ Rew 77:647–666
29. Blume L, Durlauf S (eds) (2006) The economy as an evolving

complex system, III. Current perspectives and future directions. Oxford University Press, Oxford
30. Blundell R, Stoker TM (2005) Heterogeneity and aggregation. J Econ Lit 43:347–391
31. Bowles S (1998) Endogenous preferences: The cultural consequences of markets and other economic institutions. J Econ Lit 36:75–111
32. Brock WA (1999) Scaling in economics: a reader's guide. Ind Corp Change 8(3):409–446
33. Brock WA, Colander D (2000) Complexity and policy. In: Colander D (ed) The complexity vision and the teaching of economics. Edward Elgar, Northampton
34. Brock WA, Durlauf SN (2001) Discrete choice with social interactions. Rev Econ Stud 68:235–260
35. Brock WA, Durlauf SN (2005) Social interactions and macroeconomics. UW-Madison, SSRI Working Papers n.5
36. Caballero RJ (1992) A Fallacy of composition. Am Econ Rev 82:1279–1292
37. Calafati AG (2007) Milton Friedman's epistemology UPM working paper n.270
38. Caldarelli G (2006) Scale-free networks. Complex webs in nature and technology. Oxford University Press, Oxford
39. Clower RW (1965) The keynesian counterrevolution: A theoretical appraisal. In: Hahn F, Brechling F (eds) The theory of interst rates. Macmillan, London
40. Cohen A, Harcourt G (2003) What ever happened to the Cambridge capital theory controversies. J Econ Perspect 17:199–214
41. Cole HL, Mailath GJ, Postlewaite A (1992) Social norms, savings behaviour, and growth. J Political Econ 100(6):1092–1125
42. Cooper RW (1999) Coordination games: Complementarities and macroeconomics. Cambridge University Press, Cambridge
43. Crutchfield J (1994) Is anything ever new? Considering emergence. In: Cowan G, Pines D, Meltzer D (eds) Complexity: Metaphors, models, and reality. Addison-Wesley, Reading, pp 515–537
44. Davis JB (2006) The turn in economics: Neoclassical dominance to mainstream pluralism? J Inst Econ 2(1):1–20
45. Debreu G (1959) The theory of value. Wiley, New York
46. Debreu G (1974) Excess demand functions. J Math Econ 1:15–23
47. De Masi G, Fujiwara Y, Gallegati M, Greenwald B, Stiglitz JE (2008) Empirical evidences of credit networks in Japan. mimeo
48. Delli Gatti D, Di Guilmi C, Gaffeo E, Gallegati M, Giulioni G, Palestrini A (2004) Business cycle fluctuations and firms' size distribution dynamics. Adv Complex Syst 7(2):1–18
49. Delli Gatti D, Di Guilmi C, Gaffeo E, Gallegati M, Giulioni G, Palestrini A (2005) A new approach to business fluctuations: Heterogeneous interacting agents, scaling laws and financial fragility. J Econ Behav Organ 56(4):489–512
50. Denzau AT, North DC (1994) Shared mental models: Ideologies and institutions. Kyklos 47(1):3–31
51. Descartes R (1637) Discours de la méthode pour bien conduire sa raison, et chercher la verité dans les sciences, tr. Discourse on Method and Meditations. The Liberal Arts Press, 1960, New York
52. Dorogovtsev SN, Mendes JFF (2003) Evolution of networks from biological nets to the internet and the WWW. Oxford University Press, Oxford
53. Di Guilmi C, Gallegati M, Landini S (2007) Economic dynamics with financial fragility and mean-field interaction: a model. arXiv:0709.2083
54. Durlauf SN (1993) Nonergodic economic growth. Rev Econ Stud 60:349–366
55. Durlauf SN (1997) What should policymakers know about economic complexity? Wash Q 21(1):157–165
56. Durlauf SN, Young HP (2001) Social dynamics. The MIT Press, Cambridge
57. Edgeworth FY (1925) The pure theory of monopoly In: Papers relating to political economy. McMillan, London
58. Epstein JM (1999) Agent-based computational models and generative social science. Complexity 4:41–60
59. Epstein JM (2006) Remarks on the foundations of agent-based generative social science. In: Tesfatsion L, Judd KL (eds) Handbook of computational economics. Agent-based computational economics, vol 2, ch 34. North-Holland, Amsterdam, pp 1585–1604
60. Epstein JM (2006) Generative social science: Studies in agent-based computational modeling. Princeton University Press, New York
61. Epstein JM, Axtell RL (1996) Growing artificial societies: Social science from the bottom up. The MIT Press, Cambridge
62. Fagiolo G, Moneta A, Windrum P (2007) A critical guide to empirical validation of agent-based models in economics: Methodologies, procedures, and open problems. Comput Econ 30:195–226
63. Farley R (1986) The residential segregation of blacks from whites: Trends, causes, and consequences. In: US Commission on Civil Rights, Issues in housing discrimination. US Commission on Civil Rights
64. Feller W (1957) An introduction to probability. Theory and its applications. Wiley, New York
65. Finch J, Orillard M (eds) (2005) Complexity and the economy: Implications for economy policy. Edward Elgar, Cheltenham
66. Flach PA, Kakas AC (eds) (2000) Abduction and induction. Essays on their relation and integration. Kluwer, Dordrecht
67. Flake GW (1998) The computational beauty of nature. The MIT Press, Cambridge
68. Foellmer H (1974) Random economies with many interacting agents. J Math Econ 1:51–62
69. Forni M, Lippi M (1997) Aggregation and the micro-foundations of microeconomics. Oxford University Press, Oxford
70. Frazer J (1995) An evolutionary architecture. Architectural Association Publications, London
71. Friedman M (1953) Essays in positive economics. University of Chicago Press, Chicago
72. Fujiwara Y (2006) Proceedings of the 9th Joint Conference on Information Sciences (JCIS), Advances in Intelligent Systems Research Series. Available at http://www.atlantis-press.com/publications/aisr/jcis-06/index_jcis
73. Gabaix X (2008) Power laws in Economics and Finance, 11 Sep 2008. Available at SSRN: http://ssrn.com/abstract=1257822
74. Gaffeo E, Gallegati M, Palestrini A (2003) On the size distribution of firms, additional evidence from the G7 countries. Phys A 324:117–123
75. Gaffeo E, Russo A, Catalano M, Gallegati M, Napoletano M (2007) Industrial dynamics, fiscal policy and R&D: Evi-

dence from a computational experiment. J Econ Behav Organ 64:426–447
76. Gallegati M (1993) Composition effects and economic fluctuations. Econ Lett 44(1–2):123–126
77. Gallegati M, Delli Gatti D, Gaffeo E, Giulioni G, Palestrini A (2008) Emergent macroeconomics. Springer, Berlin
78. Gallegati M, Palestrini A, Delli Gatti D, Scalas E (2006) Aggregation of heterogeneous interacting agents: The variant representative agent framework. J Econ Interact Coord 1(1):5–19
79. Gilbert N (ed) (1999) Computer simulation in the social sciences, vol 42. Sage, Thousand Oaks
80. Gilbert N, Terna P (2000) How to build and use agent-based models in social science. Mind Soc 1:57–72
81. Gilbert N, Troitzsch K (2005) Simulation for the social scientist. Open University Press, Buckingham
82. Gintis H (2007) The dynamics of general equilibrium. Econ J 117:1280–1309
83. Glaeser E, Sacerdote B, Scheinkman J (1996) Crime and social interactions. Q J Econ 111:507–548
84. Glaeser J, Dixit J, Green DP (2002) Studying hate crime with the internet: What makes racists advocate racial violence? J Soc Issues 58(122):177–194
85. Gourieroux C, Monfort A (1997) Simulation-based econometric methods. Oxford University Press, Oxford
86. Greenwald B, Stiglitz JE (1986) Externalities in economies with imperfect information and incomplete markets. Q J Econ 101(2):229–264
87. Grossman SJ, Stiglitz JE (1976) Information and competitive price systems. Am Econ Rev 66:246–253
88. Grossman SJ, Stiglitz JE (1980) On the impossibility of informationally efficient markets. Am Econ Rev 70(3):393–408
89. Guesnerie R (1993) Successes and failures in coordinating expectations. Eur Econ Rev 37:243–268
90. Hahn F (1982) Money and inflation. Blackwell, Oxford
91. Haken H (1983) Synergetics. Nonequilibrium phase transitions and social measurement, 3rd edn. Springer, Berlin
92. Hansen L, Heckman J (1996) The empirical foundations of calibration. J Econ Perspect 10:87–104
93. Hempel CV (1965) Aspects of scientific explanation. Free Press, London
94. Hempel CV, Oppenheim P (1948) Studies in the logic of explanation. Philos Sci 15:135–175
95. Hildenbrand W, Kirman AP (1988) Equilibrium analysis: Variations on the themes by edgeworth and walras. North-Holland, Amsterdam
96. Horgan J (1995) From complexity to perplexity. Sci Am 272:104
97. Horgan J (1997) The end of science: Facing the limits of knowledge in the twilight of the scientific age. Broadway Books, New York
98. Jerison M (1984) Aggregation and pairwise aggregation of demand when the distribution of income is fixed. J Econ Theory 33(1):1–31
99. Kirman AP (1992) Whom or what does the representative individual represent. J Econ Perspect 6:117–136
100. Kirman AP (1996) Microfoundations – built on sand? A review of Maarten Janssen's microfoundations: A Critical Inquiry. J Econ Methodol 3(2):322–333
101. Kirman AP (2000) Interaction and markets. In: Gallegati M, Kirman AP (eds) Beyond the representative agent. Edward Elgar, Cheltenham
102. Kleijnen JPC (1998) Experimental design for sensitivity analysis, optimization, and validation of simulation models. In: Banks J (ed) Handbook of simulation. Wiley, New York, pp 173–223
103. Kleijnen JPC, Sargent RG (2000) A methodology for the fitting and validation of metamodels in simulation. Eur J Oper Res 120(1):14–29
104. Krugman P (1998) Bubble, boom, crash: theoretical notes on Asia's crisis. mimeo
105. Kydland FE, Prescott EC (1996) The computational experiment: An econometric tool. J Econ Perspect 10:69–85
106. Lavoie D (1989) Economic chaos or spontaneous order? Implications for political economy of the new view of science. Cato J 8:613–635
107. Leibenstein H (1950) Bandwagon, snob, and veblen effects in the theory of consumers' demand. Q J Econ 64:183–207
108. Leijonhufvud A (1973) Life among the econ. Econ Inq 11:327–337
109. Leombruni R (2002) The methodological status of agent-based simulations, LABORatorio Revelli. Working Paper No. 19
110. Leombruni R, Richiardi MG (2005) Why are economists sceptical about agent-based simulations? Phys A 355:103–109
111. Leombruni R, Richiardi MG, Saam NJ, Sonnessa M (2005) A common protocol for agent-based social simulation. J Artif Soc Simul 9:1
112. Levy M, Levy H, Solomon S (2000) Microscopic simulation of financial markets. In: From Investor Behavior to Market Phenomena. Academica Press, New York
113. Lewontin C, Levins R (2008) Biology under the influence: Dialectical essays on the coevolution of nature and society. Monthly Review Press, US
114. Lucas RE (1976) Econometric policy evaluation: A critique. Carnegie-Rochester Conference Series, vol 1, pp 19–46
115. Lucas RE (1987) Models of business cycles. Blackwell, New York
116. Lucas RE, Sargent T (1979) After keynesian macroeconomics. Fed Reserv Bank Minneap Q Rev 3(2):270–294
117. Magnani L, Belli E (2006) Agent-based abduction: Being rational through fallacies. In: Magnani L (ed) Model-based reasoning in science and engineering, Cognitive Science, Epistemology, Logic. College Publications, London, pp 415–439
118. Manski CF (2000) Economic analysis of social interactions. J Econ Perspect 14:115–136
119. Mantel R (1974) On the characterization of aggregate excess demand. J Econ Theory 7:348–353
120. Mantegna RN, Stanley HE (2000) An introduction to econophysics. Cambridge University Press, Cambridge
121. Marks RE (2007) Validating Simulation Models: A general framework and four applied examples. Comput Econ 30:265–290
122. May RM (1976) Simple mathematical models with very complicated dynamics. Nature 261:459–467
123. Mas-Colell A, Whinston MD, Green J (1995) Microeconomic theory. Oxford University Press, Oxford
124. Miller JH, Page SE (2006) Complex adaptive systems: An introduction to computational models of social life. Princeton University Press, New York
125. Mirowski P (1989) More heat than light. Cambridge University Press, Cambridge
126. Muth RF (1986) The causes of housing segregation. US Com-

mission on Civil Rights, Issues in Housing Discrimination. US Commission on Civil Rights
127. Nagel E (1961) The structure of science. Routledge and Paul Kegan, London
128. Nicolis G, Prigogine I (1989) Exploring complexity: An introduction. WH Freeman, New York
129. North MJ, Howe TR, Collier NT, Vos JR (2005) Repast simphony runtime system. In: Macal CM, North MJ, Sallach D (eds) Proceedings of the agent 2005 Conference on Generative Social Processes, Models, and Mechanisms, 13–15 Oct 2005
130. Ostrom T (1988) Computer simulation: the third symbol system. J Exp Soc Psycholog 24:381–392
131. Page S (1999) Computational models from A to Z. Complexity 5:35–41
132. Peirce CS (1955) Abduction and induction. In: J Buchler (ed) Philosophical writings of peirce. Dover, New York
133. Phelan S (2001) What is complexity science, really? Emergence 3:120–136
134. Pollack R (1975) Interdependent preferences. Am Econ Rev 66:309–320
135. Railsback SF, Lytinen SL, Jackson SK (2006) Agent-based simulation platforms: Review and development recommendations. Simulation 82:609–623
136. Rappaport S (1996) Abstraction and unrealistic assumptions in economics. J Econ Methodol 3(2):215–36
137. Resnick M (1994) Turtles, termites and traffic jams: Explorations in massively parallel microworlds. MIT, Cambidge
138. Richter MK, Wong K (1999) Non-computability of competitive equilibrium. Econ Theory 14:1–28
139. Rioss Rull V (1995) Models with heterogeneous agents. In: Cooley TF (ed) Frontiers of business cycle research. Princeton University Press, New York
140. Rosser JB (1999) On the complexities of complex economic dynamics. J Econ Perspect 13:169–192
141. Rosser JB (2000) Integrating the complexity vision into the teaching of mathematical economics. In: Colander D (ed) The complexity vision and the teaching of economics. Edward Elgar, Cheltenham, pp 209–230
142. Rosser JB (2003) Complexity in economics. Edward Elgar, Cheltenham
143. Rust J (1997) Using randomization to break the curse of dimensionality. Econometrica 65:487–516
144. Saari DG (1995) Mathematical complexity of simple economics. Notices Am Math Soc 42:222–230
145. Schelling TC (1969) Models of segregation. Am Econ Rev 59:488–493
146. Schelling TC (1971) Dynamic models of segregration. J Math Sociol 1:143–186
147. Schelling TC (1978) Micromotives and macrobehaviour. W.W. Norton, New York
148. Schelling TC (2006) Some fun, thirty-five years ago. In: Tesfatsion L, Judd KL (eds) Handbook of computational economics. Agent-based computational economics, vol 2, ch 37. North-Holland, Amsterdam, pp 1639–1644
149. Schumpeter JA (1960) History of economic analysis. Oxford University Press, Oxford
150. Segre-Tossani L, Smith LM (2003) Advanced modeling, visualization, and data mining techniques for a new risk landscape. Casualty Actuarial Society, Arlington, pp 83–97
151. Semmler W (2005) Introduction (multiple equilibria). J Econ Behav Organ 57:381–389
152. Shy O (2001) The economics of network industries. Cambridge University Press, Cambridge
153. Smith A (1776/1937) The wealth of nations. Random House, New York
154. Solomon S (2007) Complexity roadmap. Institute for Scientific Interchange, Torino
155. Sonnenschein H (1972) Market excess demand functions. Econometrica 40:549–563
156. Stiglitz JE (1992) Methodological issues and the new keynesian economics. In: Vercelli A, Dimitri N (eds) Macroeconomics: A survey of research strategies. Oxford University Press, Oxford, pp 38–86
157. Stiglitz JE (2002) Globalization and its discontents. Northon, New York
158. Stoker T (1995) Empirical approaches to the problem of aggregation over individuals. J Econ Lit 31:1827–1874
159. Tesfatsion L (ed) (2001) Special issue on agent-based computational economics. J Econ Dyn Control 25
160. Tesfatsion L (ed) (2001) Special issue on agent-based computational economics. Comput Econ 18
161. Tesfatsion L (2001) Agent-based computational economics: A brief guide to the literature. In: Michie J (ed) Reader's guide to the social sciences. Fitzroy-Dearborn, London
162. Tesfatsion L (2002) Agent-based computational economics: Growing economies from the bottom up. Artif Life 8:55–82
163. Tesfatsion L (2006) Agent-based computational economics: A constructive approach to economic theory. In: Tesfatsion L, Judd KL (eds) Handbook of computational economics. Agent-based computational economics, vol 2, ch 16. North-Holland, Amsterdam, pp 831–880
164. Troitzsch KG (2004) Validating simulation models. In: Horton G (ed) Proceedings of the 18th european simulation multiconference. Networked simulations and simulation networks. SCS Publishing House, Erlangen, pp 265–270
165. Vriend NJ (1994) A new perspective on decentralized trade. Econ Appl 46(4):5–22
166. Vriend NJ (2002) Was Hayek an ace? South Econ J 68:811–840
167. Velupillai KV (2000) Computable economics. Oxford University Press, Oxford
168. Velupillai KV (2002) Effectivity and constructivity in economic theory. J Econ Behav Organ 49:307–325
169. Velupillai KV (2005) The foundations of computable general equilibrium theory. In: Department of Economics Working Papers No 13. University of Trento
170. Velupillai KV (2007) The impossibility of an effective theory of policy in a complex economy. In: Salzano M, Colander D (eds) Complexity hints for economic policy. Springer, Milan
171. von Hayek FA (1948) Individualism and economic order. University of Chicago Press, Chicago
172. von Mises L (1949) Human action: A treatise on economics. Yale University Press, Yale
173. Wilensky U (1998) NetLogo segregation model. Center for connected learning and computer-based modeling. Northwestern University, Evanston. http://ccl.northwestern.edu/netlogo/models/Segregation
174. Winker P, Gilli M, Jeleskovic V (2007) An objective function for simulation based inference on exchange rate data. J Econ Interact Coord 2:125–145
175. Wooldridge M (2001) An introduction to multiagent systems. Wiley, New York

Aggregation Operators and Soft Computing

VICENÇ TORRA
Institut d'Investigació en Intelligència Artificial – CSIC, Bellaterra, Spain

Article Outline

Glossary
Definition of the Subject
Introduction
Applications
Aggregation, Integration and Fusion
Aggregation Operators and Their Construction
Some Aggregation Operators
Future Directions
Bibliography

Glossary

Information integration The whole process of obtaining some information from different sources and then using this information to achieve a concrete task.

Information fusion A general term that defines the whole area that studies techniques and methods to combine information for achieving a particular goal.

Aggregation operators Particular operators that are actually used for combining the information.

Definition of the Subject

Aggregation operators are the particular functions used for combining information in systems where several information sources have to be taken into consideration for achieving a particular goal.

Formally, aggregation operators are the particular techniques of information fusion which can be mathematically expressed in a relatively simple way. They are part of the more general process of information integration. That is the process that goes from the acquisition of data to the accomplishment of the final task.

As current systems (from search engines to vision and robotics systems) need to consider information from different sources, aggregation operators are of major importance. Such operators permit one to reduce the quantity of information and improve its quality.

Introduction

Information fusion is a broad field that studies computational methods and techniques for combining data and information. At present, different methods have been developed. Differences are on the type of data available and on the assumptions data and information sources (data suppliers) satisfy. For example, methods exist for combining numerical data (e. g. readings from sensors) as well as for combining structured information (e. g. partial plans obtained from AI planners). Besides, the complexity of the methods ranges from simple functions (as the arithmetic mean or the voting technique) that combine a few data to more complex ones (as procedures for dealing with satellite images) that combine a huge burden of data.

Soft computing is an area in computer science that encompasses a few techniques with the goal of solving complex problems when there are no analytic or complete solutions. Two typical problems are the control of a variable (when standard methods, as PID controllers, are not applicable) and the optimization of a function (when standard methods, as quadratic programming, cannot be applied because the objective function to be optimized is highly non linear). Among the techniques in soft computing, we can underline genetic algorithms and evolutionary computation, neural networks, and fuzzy sets and fuzzy systems [21].

Some of the techniques included in soft computing have been used for combining and fusing data. This is the case of neural networks. It is well known that neural networks (when applied in a supervised manner) can be used to build a system that reproduces up to some extent the behavior of a set of examples. So, in the case of aggregation and data fusion, they can be used to build models when we have a data set consisting on some examples (pairs of input data, expected outcome) that are to be reproduced by the system. Aggregation operators, which can also be seen from the soft computing perspective, follow a different principle. They are defined mathematically and are characterized in terms of their properties.

Aggregation operators are often studied in the framework of fuzzy sets and fuzzy systems, one of the components of soft computing. Such operators, when operating on the $[0, 1]$ interval are closely related to with t-norms and t-conorms, the basic functions in fuzzy logic to model conjunction and disjunction. Formally, t-norms and t-conorms are used in fuzzy logic to combine the membership degrees either in a conjunctive or disjunctive way. In this setting aggregation operators are also used to combine information but additional flexibility is permitted in the combination. Weighted minimum, and the fuzzy integrals are some examples of aggregation operators that are typically studied in this framework.

In this work, we present an overview of the area of information fusion and we focus on the aggregation oper-

ators. We review a few of them with some of their properties. We focus on those that have a tight link with soft computing, and especially on the ones that belong to the field of fuzzy sets and fuzzy logic.

Applications

Aggregation operators are used in a large variety of contexts and applications [8]. We underline a few of them below. For example, there are applications in economy (to define all kind of indices as e. g. the retail price index), biology (to combine DNA sequences and classifications, either expressed as dendrograms – trees – or partitions [7,12]), in education (to rate students, universities and educational systems [8]), in bibliometrics (to evaluate the quality of the researchers and rate them by means of citation indices as e. g. the number of citations or the h index [16,25]), and in computer science.

In computer science, aggregation operators are used for e. g. evaluating hardware and software, and for fusing preferences. In the first case, different criteria or performance measures are combined to define an index. For example, performance measures for evaluating runtime systems are defined as the aggregation of the execution times of different benchmarks. Some means (e. g. geometric means) are used for such aggregation. In the second case, different sets of preferences are fused into a new aggregated one. For example, meta-search engines proceed in this way aggregating the results of different search engines. Note that the result of each search engine is a preference on the links and the final ranking is another preference that is built as an aggregation of the initial preferences.

In addition, aggregation operators and information fusion are also used in artificial intelligence, robotics and vision. Knowledge integration, knowledge revision or belief merging for knowledge based systems can also be seen as a case of information fusion at large. In this case, a knowledge base, which is defined using e. g. description logic in terms of propositions, is checked to detect incoherences and, eventually, to correct them. In vision, fusion is used e. g. for improving the quality of images. Fusion is also used in robotics for a similar purpose. In particular, it is used for improving the quality of the information obtained from the sensors [18].

The different uses of information fusion in artificial intelligence and robotics can be classified into two main categories. On the one hand, we have the uses related to making decisions and, on the other hand, the uses related with the obtainment of a better description of the application domain (so that the system has a better understanding).

Decision making typically corresponds to the process of selecting the best alternative. Difficulties in decision making are found when each alternative is evaluated using different (conflicting) criteria. This case corresponds to the multi criteria decision making framework [13,20]. In such framework, depending on the criterion used in the selection, different alternatives are selected. That is, different Pareto optimal solutions can be selected according to the criteria. One approach to solve this situation is to use aggregation operators to combine the degrees of satisfaction for each criteria into a new aggregated degree of satisfaction (for a new *aggregated criterion*). Then selection uses this aggregated criterion to select the best rated alternative. Another approach consists of considering different preferences (one for each criterion) and then such preferences are aggregated into a new aggregated preference. Then, selection is based on the aggregated preference.

A different case, also related to decision making, is when the alternative to be selected (e. g. the solution of a problem) is constructed from the other ones. For example, several partial solutions are retrieved or obtained by a set of subsystems (maybe because such solutions have been applied before with success into similar problems, as in case-based reasoning). Then, if none of them is completely appropriate in the new situation, the new solution is built from such partial solutions. Some applications of case-based reasoning can be seen from this perspective. This is also the case in plan merging, where partial plans are combined to build the final plan. Another case corresponds to the ensemble methods in machine learning. Such methods assign an outcome (e. g., a class in a classification problem) for a given data taking into account the outcomes of different subsystems (e. g., different classifiers based on different assumptions constructed from the data in the learning set). Then, a final outcome is decided on the basis of the outcomes of the subsystems (e. g., using a voting strategy).

The second use of information fusion and aggregation operators is for improving the understanding of the application domain. This tries to solve some of the problems related to data of low quality, e. g., data that lacks accuracy due to errors caused by the information source or due to the errors introduced in the transmission. Another difficulty is that the data from a single source might describe only part of the application's domain, and thus, several sources are needed to have a more comprehensive description of the domain, e. g. either several video cameras or a sequence of images from the same video camera are needed to have a full description of a room or a building. Knowledge integration and knowledge revision can be seen from this perspective as the knowledge in a system

needs to be updated and refined when the system is under execution.

The different uses of aggregation operators and the large number of applications have stimulated the definition, study and characterization of aggregation operators on different types of data. Due to this, there are nowadays operators for a large number of different domains. For example, there are operators to combine data in nominal scales (no ordering or structure exists between the elements to be aggregated), ordinal scales (there is ordering between pairs of categories but no other operators are allowed), numerical data (ratio and measure scales), partitions (input data correspond to a partition of the elements of a given set), fuzzy partitions (as in the previous case, but fuzzy partitions are considered), dendrograms (e. g., tree-like structures for classifying species), DNA sequences (i. e., sequences of A, T, C, and G), images (in remote sensing applications), and logical propositions (in knowledge integration).

Aggregation, Integration and Fusion

In the context of information fusion there are three terms that are in common use. The terms are information integration, information fusion, and aggregation operator. Although these terms are often used as synonymous their meaning is not exactly equivalent. We define them below.

Information integration This corresponds to the whole process of obtaining some information from different sources and then using this information to achieve a concrete task. It is also possible that instead of information from several sources, information is extracted from the same source but at different times.

Information fusion This is a general term that defines the whole area that studies techniques and methods to combine information for achieving a particular goal. It is also understood as the actual process of combining the information (a particular algorithm that uses the data to obtain a single datum).

Aggregation operators They are some of the particular operators that are actually used for combining the information. They can be understood as particular information fusion techniques. In fact, such operators are the ones that can be formalized straightforwardly in a mathematical way. The most well known aggregation operators are the means. Other operators as t-norms and t-conorms, already pointed out above, are also considered by some as aggregation operators. In general, we consider operators that combine N values in a given domain D obtaining a new single datum in the same domain D. For example, in the case of the arithmetic mean, we combine N values on the real line and obtain a new datum that is also a value in the real line.

Taking all this into account, we would say that a robot that combines sensory information includes some information integration processes in its internal system. When sensory data is combined with a mean, or a similar function, we would say that it uses a fusion method that is an aggregation operator. In contrast, if the robot combines the data using a neural network or another black-box approach, we would say that it uses an information fusion method but that no aggregation operator can be distinguished.

As said above, some people distinguish between aggregation operators and means. In this case, a mean is a particular aggregation operator that satisfies unanimity and monotonicity. Other operators that combine N values but that do neither satisfy unanimity nor monotonicity are not means but aggregation operators. Besides, there are aggregation operators that are not required to satisfy monotonicity because they are not defined on ordered domains. This is the case of operators on nominal scales. The plurality rule or voting is one of such operators.

In this work we consider aggregation in a narrow sense. Therefore, an aggregation operator is a function $\mathbb{C}$, from consensus, such that combines N values in a domain D and returns another value in the same domain, i. e., $\mathbb{C}\colon D^N \to D$. Such function satisfies unanimity (i. e., $\mathbb{C}(a,\ldots,a) = a$ for all a in D) and, if applicable, monotonicity (i. e., $\mathbb{C}(a_1,\ldots,a_N) \le \mathbb{C}(b_1,\ldots,b_N)$ when $a_i < b_i$. i. e., this latter condition is valid, as stated before, only when there is an order on the elements in D, and, otherwise, the condition is removed.

Operators on ordered domains defined in this way satisfy internality. That is, the aggregation of values $a_1,\ldots,a_N$ is a value between the minimum and the maximum of these values:

$$\min_i a_i \le \mathbb{C}(a_1,\ldots,a_N) \le \max_i a_i \,.$$

Aggregation Operators and Their Construction

The definition of an aggregation operator has to take into account several aspects. On the one hand, it should consider the type of data to be combined (this aspect has been mentioned before). That is, whether data is numerical, categorical, etc. On the second hand, it should consider the level of abstraction. That is, whether aggregation has to be done at a lower level or at a higher one. Due to the data flow, the data in a system typically moves from low levels to higher ones. Then, the fusion of two data can be done

at different abstraction levels. Note that data at different levels have different representations, e. g. information at a lower level might be numerical (e. g. the values of a measurement), while at a higher level it might be symbolical (e. g. propositions in logic, or belief measures). The level of abstraction and the type of data that are appropriate for fusion depends on a third element, the type of information. That is, information from different sources might be redundant or complementary. Redundant information is usually aggregated at a low level (e. g. the values of measuring the same object using two similar sensors), while complementary information is usually aggregated at a higher level (e. g., the fusion of images to have a better representation of 3D objects cannot be done at a pixel–signal–level).

Once the appropriate environment for a function has been settled, it is the time for the definition of the function. Two main cases can be distinguished, (i) definition from examples and (ii) definitions from properties.

(i) The use of examples for defining the operators follows machine learning theory. Two main cases can be distinguished. The first one is when we assume that the model of aggregation is known. e. g., we know that we will use a weighted mean for the aggregation, and then the problem is to find the appropriate parameters for such weighted mean. The second one is when the model is not known and only a loose model – black box model – is assumed. Then the examples are used to settle such model. The use of neural networks is an example of this kind of application.

(ii) In the case of definition from properties, the system designer establishes the properties that the aggregation operator needs to satisfy. Then the operator is derived from such properties.

Functional equations [1,2] can be used for this purpose. Functional equations are equations that have functions as their unknowns. For example, $\phi(x+y) = \phi(x) + \phi(y)$ is an example of a functional equation (one of Cauchy's equations) with an unknown function ϕ. In this equation, the goal is to find the functions ϕ that satisfy such equation. In this case, when ϕ are continuous, solutions are functions $\phi(x)$ of the form $\phi(x) = \alpha x$.

In the case of aggregation, we might be interested in, for instance, functions $\phi(x, y)$ that aggregate two numerical values x and y, and that satisfy unanimity (i. e., $\phi(x, x) = x$), invariance to transformations of the form $\gamma(x) = x + t$ (i.e, $\phi(x_1 + t, x_2 + t) = \phi(x_1, x_2) + t$), and invariance to transformations of the form $\gamma(x) = rx$ – positive homogeneity (i. e., $\phi(rx_1, rx_2) = r\phi(x_1, x_2)$). Such functions are completely characterized and all of them are of the form $\phi(x_1, x_2) = (1-k)x_1 + kx_2$.

Naturally, the definition from properties might cause the definition of an overconstrained problem. That is, we may include too restrictive constraints and define a problem with no possible solution. This is the case of the problem of aggregation of preferences [5] that lead to Arrow's impossibility theorem [4]. That is, there is a set of natural properties that no function for the aggregation of preferences can satisfy.

The definition of the aggregated value as the one that is located at a minimum distance from the objects being aggregated is another approach for defining aggregation operators. In this case, we need to define a distance between pairs of objects on the domain of application. Let D be the domain of application, let d be a distance function for pairs of objects in D (that is, $d\colon D \times D \to \mathbb{R}$). Then, the aggregation function $\mathbb{C}\colon D^N \to D$ applied to $a_1, \ldots, a_N$ is formally defined as

$$\mathbb{C}(a_1, \ldots, a_N) = \arg\min_c \left\{ \sum_{i=1}^{N} d(c, a_i) \right\} .$$

Such definition is often known as the median rule. In the numerical setting, i. e., if the domain D is a subset of the real set, when $d(a, b) = (a-b)^2$, the aggregation operator results to be the arithmetic mean. When $d(a, b) = |a-b|$, the aggregation operator results into the median. Formally, the median is the only solution when N is of the form $2k+1$ or when N is of the form $2k$ but the two central values are equal. Otherwise, (when N is of the form $2k$ and the two central values are different), the median is one of the solutions of the equation above but other solutions might be also valid. For example, let us assume that $N = 4$ and that $a_1 \le a_2 < a_3 \le a_4$, then another solution is,

$$\mathbb{C}(a_1, \ldots, a_N) = \frac{(a_4 a_3 - a_2 a_1)}{(a_4 + a_3) - (a_2 + a_1)} .$$

When $d(a, b) = |a-b|^\infty$ for $k \to \infty$, we have that the midrange of the elements $a_1, \ldots, a_N$ is a solution of the equation above. That is,

$$\mathbb{C}(a_1, \ldots, a_N) = \frac{(\min_i a_i - \max_i a_i)}{2} .$$

Some Aggregation Operators

There exist a large number of aggregation operators. We will discuss in this section a few of them, classified into different families. We begin with the quasi-arithmetic mean, a family of operators that encompasses some of the most well-known aggregation operators, the arithmetic mean, the geometric mean, and the harmonic mean.

Quasi-Arithmetic Means

A quasi-arithmetic mean is an aggregation operator $\mathbb{C}$ of the form

$$\mathrm{QAM}(a_1, \ldots, a_N) = \varphi^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\varphi(a_i)\right) .$$

with φ an invertible function.

We list below a few examples of quasi-arithmetic means.

- Arithmetic mean:

$$\mathrm{AM}(a_1, \ldots, a_N) = \frac{1}{N}\sum_{i=1}^{N} a_i .$$

- Geometric mean:

$$\mathrm{GM}(a_1, \ldots, a_N) = \sqrt[N]{\frac{1}{N}\sum_{i=1}^{N} a_i} .$$

- Harmonic mean:

$$\mathrm{HM}(a_1, \ldots, a_N) = \frac{N}{\sum_{i=1}^{N}\frac{1}{a_i}} .$$

- Power mean (Hölder mean, or root-mean-power, RMP):

$$\mathrm{RMP}(a_1, \ldots, a_N) = \sqrt[\alpha]{\frac{1}{N}\sum_{i=1}^{N} a_i^{\alpha}} .$$

- Quadratic mean (root-mean-square):

$$\mathrm{RMS}(a_1, \ldots, a_N) = \sqrt[2]{\frac{1}{N}\sum_{i=1}^{N} a_i^{2}} .$$

These means are obtained from the quasi-arithmetic mean using the following definition for the function φ: $\varphi(x) = x$ (in the arithmetic mean), $\varphi(x) = \log x$ (in the geometric mean), $\varphi(x) = 1/x$ (in the harmonic mean), $\varphi(x) = x^{\alpha}$ (in the power mean), and $\varphi(x) = x^2$ (in the quadratic mean).

Several properties have been proved for these functions. In particular, some inequalities have been proven [10,15]. For example, it is known that the harmonic mean (HM) always return a value that is smaller than the one of the geometric mean (GM) and that the value of the arithmetic mean is smaller than the one of the arithmetic mean (AM). That is, the following inequality holds for all x, y,

$$\mathrm{HM}(x, y) \le \mathrm{GM}(x, y) \le \mathrm{AM}(x, y) .$$

In the case of the power means, the following inequality holds,

$$\mathrm{RMP}_r(x, y) \le \mathrm{RMP}_s(x, y) \quad \text{for all } \ r < s .$$

That is, the power means is monotonic with respect to its parameter.

Also related with the power means, we have that when $\alpha \to 0$, $\mathrm{RMP}_{\alpha} = \mathrm{GM}$, that when $\alpha \to \infty$, $\mathrm{RMP}_{\alpha} = \max$ and that when $\alpha \to -\infty$, $\mathrm{RMP}_{\alpha} = \min$.

From the point of view of the applications, it is of relevance whether the outcome of an aggregation operator is, in general, large or small. That is, whether the aggregation operator A is larger or smaller than the aggregation operator B. For example, whether it holds or not that

$$A(x, y) \le B(x, y) \quad \text{for all } \ x, y.$$

Due to the importance in real practice of such ordering, some indices have been developed for analyzing the outcomes of the operators. The average value, the orness and the andness are examples of such indices.

The orness computes the similarity of the operator with the maximum. Recall that in fuzzy sets, the maximum is used to model the disjunction, i. e., or. In contrast to that, the andness computes the similarity of the operator with the minimum. Recall that in fuzzy sets, the minimum is used to model the conjunction, i. e., and. For these measures, it can be proven that (i) the operator maximum has a maximum orness (an orness equal to one) and a minimum andness (an andness equal to zero), (ii) the orness of the arithmetic mean is larger than the orness of the geometric mean, and this latter orness is larger than the one of the harmonic mean. This property follows from the following inequality that holds for all x and y,

$$\mathrm{HM}(x, y) \le \mathrm{GM}(x, y) \le \mathrm{AM}(x, y) .$$

The interpretations for the andness and the orness ease the selection of the appropriate aggregation function. This is so because the orness can be interpreted as a degree of compensation. That is, when different criteria are considered, a high value of orness means that a few good criteria can compensate a large number of bad ones. In contrast to that, a low value of orness means that all criteria have to be satisfied in order to have a good final score. Taking this into account, and using the example above, when only low compensation is permitted, it is advisable to use the geometric mean or the harmonic mean. In contrast, a larger compensation requires the user to consider the arithmetic mean.

A similar approach can be applied for the power mean. The larger the parameter r in the function (i. e., RMP_r) the larger the compensation.

Weighted Means and Quasi-weighted Means

The weighted mean is the most well-known aggregation operator with weights, and the quasi-weighted mean is one of its generalizations. In such operator, weights are represented by means of a weighting vector. That is, a vector $w = (w_1, \ldots, w_N)$ that satisfies

$$w_i \geq 0 \quad \text{and} \quad \Sigma w_i = 1 \,.$$

- Weighted mean: given a weighting vector $w = (w_1, \ldots, w_N)$, the weighted mean is defined as follows

  $$\mathrm{WM}(a_1, \ldots, a_N) = \sum_{i=1}^{N} w_i a_i \,.$$

  Naturally, when $w_i = 1/N$ for all $i = 1, \ldots, N$, the weighted mean corresponds to the arithmetic mean.
  In this definition, the weights can be understood as their reliability. The larger the importance, the larger the weights.
- Quasi-weighted mean: given a weighting vector $w = (w_1, \ldots, w_N)$, the quasi-weighted mean (or quasi-linear mean) is defined as follows

  $$\mathrm{QWM}(a_1, \ldots, a_N) = \varphi^{-1}\left(w_i \sum_{i=1}^{N} \varphi(a_i)\right) ,$$

  with φ an invertible function, and $w = (w_1, \ldots, w_N)$ a weighting vector.
  The quasi-weighted mean generalizes the quasi-arithmetic mean given above introducing weights w_i for each of the source. Using appropriate functions φ, we obtain the weighted mean (with $\varphi(x) = x$), the weighted geometric mean (with $\varphi(x) = \log x$), the weighted harmonic mean (with $\varphi(x) = 1/x$), the weighted power mean (with $\varphi(x) = x^{\alpha}$), and the weighted quadratic mean (with $\varphi(x) = x^2$).

Losonczi's Means

This family of means, defined in [6], is related to the quasi-weighted means. While in the former the weights only depend on the information source (there is one weight for each a_i but such weight does not depend on the value a_i), this is not the case in the Losonczi's means. In this operator, weights depend on the values being aggregated.

Formally, a Losonczi's mean is an aggregation operator $\mathbb{C}$ of the form,

$$\mathbb{C}(a_1, \ldots, a_N) = \varphi^{-1}\left(\frac{\sum_{i=1}^{N} \pi_i(a_i)\varphi(a_i)}{\sum_{i=1}^{N} \pi_i(a_i)}\right) ,$$

with φ an invertible function.

Naturally, when $\pi_i(a)$ is constant for all a, we have that the function reduces to a quasi-weighted means. When $\pi_i(a) = a^q$ and $\varphi(a) = a^{p-q}$ with $q = p - 1$, the Losonczi's means reduces to the counter-harmonic mean, also known by Lehmer mean. That is,

$$\mathbb{C}(a_1, \ldots, a_N) = \frac{\sum_{i=1}^{N} a_i^p}{\sum_{i=1}^{N} a_i^{p-1}} \,.$$

When $\pi_i(a) = w_i a^q$ and $\varphi(a) = a^{p-q}$ with $q = p - 1$, the Losonczi's means reduces to the weighted Lehmer mean. That is,

$$\mathbb{C}(a_1, \ldots, a_N) = \frac{\sum_{i=1}^{N} w_i a_i^p}{\sum_{i=1}^{N} w_i a_i^{p-1}} \,.$$

Linear Combination of Order Statistics and OWA Operators

Given $a_1, \ldots, a_N$ the ith order statistics [3] is the element in $a_1, \ldots, a_N$ that occupies the ith position when such elements are ordered in increasing order. Formally, let i be an index $i \in \{1, \ldots, N\}$. Then, the ith order statistics OS_i of dimension N is defined as follows:

$$\mathrm{OS}_i(a_1, \ldots, a_N) = a_{s(i)} \,.$$

Where $\{s(1), \ldots, s(N)\}$ is a permutation of $\{1, \ldots, N\}$ such that $a_{s(j)} \leq a_{s(j+1)}$ for all $j \in \{1, \ldots, N-1\}$.

When N is of the form $2k + 1$, $\mathrm{OS}_{(N+1)/2} = \mathrm{OS}_{k+1}$ is the median; when N is of the form $2k$, the median corresponds to the mean value of $\mathrm{OS}_{N/2}$ and $\mathrm{OS}_{N/2+1}$. That is,

$$\mathrm{median}(a_1, \ldots, a_N) = (\mathrm{OS}_{N/2} + \mathrm{OS}_{N/2+1})/2 \,.$$

The OWA (ordered weighted averaging) operator is a linear combination of order statistics (an L-estimator). Its definition is as follows.

- OWA operator: given a weighting vector $w = (w_1, \ldots, w_N)$ (i. e., a vector such that $w_i \geq 0$ and $\Sigma_i w_i = 1$), the OWA operator is defined as follows

  $$\mathrm{OWA}(a_1, \ldots, a_N) = \sum_{i=1}^{N} w_i a_{\sigma(i)} ,$$

  where $\{\sigma(1), \ldots, \sigma(N)\}$ is a permutation of $\{1, \ldots, N\}$ such that $a_{\sigma(i-1)} \geq a_{\sigma(i)}$ for all $i \in \{2, \ldots, N\}$. Equivalently,

  $$\mathrm{OWA}(a_1, \ldots, a_N) = \sum_{i=1}^{N} w_i \cdot \mathrm{OS}_i(a_1, \ldots, a_N) \,.$$

Naturally, when $w_i = 1/N$ for all $i = 1, \ldots, N$, the OWA operator corresponds to the arithmetic mean.
The ordering of the elements, permit the user to express the importance of some of the values being aggregated (importance of the values with respect to their position). For example, the OWA permits us to model situations where importance is given to low values (e. g. in a robot low values are more important than larger ones to avoid collisions). So, informally, we have that the w_i for $i = N$ and $i = N - 1$ should be large.

- Geometric OWA operator: given a weighting vector $w = (w_1, \ldots, w_N)$, the geometric OWA (GOWA) operator is defined as follows

$$\mathrm{GOWA}(a_1, \ldots, a_N) = \prod_{i=1}^{N} a_{\sigma(i)}^{w_i} ,$$

where $\{\sigma(1), \ldots, \sigma(N)\}$ is a permutation of $\{1, \ldots, N\}$ such that $a_{\sigma(i-1)} \geq a_{\sigma(i)}$ for all $i \in \{2, \ldots, N\}$. That is, the GOWA operator is a geometric mean of the order statistics.

- Generalized OWA operator: given a weighting vector $w = (w_1, \ldots, w_N)$, the generalized OWA operator is defined as follows

$$\text{Generalized OWA}(a_1, \ldots, a_N) = \left(\sum_{i=1}^{N} w_i a_{\sigma(i)}^{\alpha} \right)^{1/\alpha} ,$$

where $\{\sigma(1), \ldots, \sigma(N)\}$ is a permutation of $\{1, \ldots, N\}$ such that $a_{\sigma(i-1)} \geq a_{\sigma(i)}$ for all $i \in \{2, \ldots, N\}$. That is, the generalized OWA operator is a generalized mean (root-mean-powers) of the order statistics.

- Quasi-OWA operator: given a weighting vector $w = (w_1, \ldots, w_N)$, the quasi-OWA (QOWA) operator is defined as follows

$$\mathrm{QOWA}(a_1, \ldots, a_N) = \varphi^{-1} \left(\sum_{i=1}^{N} w_i \varphi\big(a_{\sigma(i)}\big) \right) .$$

where $\{\sigma(1), \ldots, \sigma(N)\}$ is a permutation of $\{1, \ldots, N\}$ such that $a_{\sigma(i-1)} \geq a_{\sigma(i)}$ for all $i \in \{2, \ldots, N\}$. That is, the quasi-OWA operator is a quasi-weighted mean of the order statistics.

$$\mathrm{OWA}(a_1, \ldots, a_N) = \prod_{i=1}^{N} a_{\sigma(i)}^{w_i} ,$$

where $\{\sigma(1), \ldots, \sigma(N)\}$ is a permutation of $\{1, \ldots, N\}$ such that $a_{\sigma(i-1)} \geq a_{\sigma(i)}$ for all $i \in \{2, \ldots, N\}$.

- BADD-OWA operator: given a weighting vector $w = (w_1, \ldots, w_N)$, the BADD-OWA (basic defuzzification distribution OWA) corresponds to the counter-harmonic mean (see above). That is,

$$\mathbb{C}(a_1, \ldots, a_N) = \frac{\sum_{i=1}^{N} a_i^p}{\sum_{i=1}^{N} a_i^{p-1}} .$$

- Induced OWA operator: given a weighting vector $w = (w_1, \ldots, w_N)$ and a priority vector $b = (b_1, \ldots, b_N)$, the induced OWA (IOWA) operator is defined as follows

$$\mathrm{IOWA}_b(a_1, \ldots, a_N) = \sum_{i=1}^{N} w_i a_{\sigma(i)} ,$$

where $\{\sigma(1), \ldots, \sigma(N)\}$ is a permutation of $\{1, \ldots, N\}$ such that $b_{\sigma(i-1)} \geq b_{\sigma(i)}$ for all $i \in \{2, \ldots, N\}$. Naturally, the OWA corresponds to the IOWA when a is used as the priority vector.

The WOWA Operator

Let p and w be two weighting vectors of dimension N. Then, the WOWA [23] operator (weighted ordered weighted averaging operator) of dimension N is defined as

$$\mathrm{WOWA}_{p,w}(a_1, \ldots, a_N) = \sum_{i=1}^{N} \omega_i a_{\sigma(i)} ,$$

where σ is defined as above (i.e, $a_{\sigma(i)}$ is the ith largest element in the collection $a_1, \ldots, a_N$), and the weight ω_i is defined as

$$\omega_i = w^* \left(\sum_{j \leq i} p_{\sigma(j)} \right) - w^* \left(\sum_{j < i} p_{\sigma(j)} \right) ,$$

with w^* being a nondecreasing function that interpolates the points

$$\left\{ \left(i/N, \sum_{j \leq i} w_j \right) \right\}_{i=1,\ldots,N} \cup \{(0, 0)\} .$$

The function w^* is required to be a straight line when the points can be interpolated in this way.

The WOWA operator generalizes the OWA and the weighted mean. When $p = (1/N, \ldots, 1/N)$, the WOWA reduces to an OWA operator and when $w = (1/N, \ldots, 1/N)$, the WOWA reduces to a weighted mean. That is,

$$\mathrm{WOWA}_{p,w}(a_1, \ldots, a_N) = \mathrm{OWA}_w(a_1, \ldots, a_N), \quad \text{when } p = (1/N, \ldots, 1/N),$$

and

$$\mathrm{WOWA}_{p,w}(a_1,\ldots,a_N) = \mathrm{WM}_p(a_1,\ldots,a_N), \quad \text{when } w = (1/N,\ldots,1/N).$$

As the OWA and the weighted mean generalize the arithmetic mean, when the weights are all equal to $1/N$, the same happens with the WOWA. i. e.,

$$\mathrm{WOWA}_{p,w}(a_1,\ldots,a_N) = \mathrm{AM}(a_1,\ldots,a_N), \quad \text{when } p = (1/N,\ldots,1/N) \text{ and } w = (1/N,\ldots,1/N).$$

As stated here, the WOWA operator reduces to the OWA and the weighted mean for appropriate weighting vectors. This property permits the user to interpret the two weighting vectors p and w used in the WOWA in terms of the corresponding interpretation in the OWA and the weighted mean. That is, p corresponds to the reliability of the information sources and w permits to express the importance of the values (with respect to their ordering). Thus, the WOWA permits the user to express which are the most reliable information sources (e. g., sensors) and which are the values that have to be considered more relevant (e. g. larger importance to small values in sensors to avoid collisions, or larger importance to large values if a high degree of compensation is permitted in multicriteria decision making).

Fuzzy Integrals

Fuzzy integrals can be used in aggregation [14,25]. Formally, they integrate a function with respect to a fuzzy measure. From the aggregation point of view, the function corresponds to the data to be aggregated and the fuzzy measure corresponds to some background knowledge (following the jargon in artificial intelligence) about the reliability or importance of the information sources. In this sense, the fuzzy measure can be seen as a generalization of the weights as they are used for the same purpose. The difference is that while the weights can only be used for a single information source, the measures are applied to sets of them.

In order to use the fuzzy integrals in aggregation we need first some formalization. First, we need to define the set of information sources, let it be $X = \{x_1,\ldots,x_N\}$. That is, X is the set of sensors in a robot or the set of criteria in a multicriteria decision making problem. Then, we need to express the datum (a number) supplied by source x_i. This is represented by a function f from X to $\mathbb{R}$. That is, $f(x_i)$ is the data supplied by sensor x_i or the evaluation of the criterion x_i. In terms of the notation used above in this work, $f(x_i) = a_i$. Then, the aggregation of the data supplied by the information sources X can be computed as the integral of the function f with respect to a fuzzy measure on X. The measure is a function defined on the subsets of X. So, for any set of information sources, we have its *weight*.

Formally, a fuzzy measure is a set function on X into $[0,1]$. That is, a function from $\wp(X)$ into $[0,1]$ or, in other words, a function that assigns a value in $[0,1]$ to each subset A of X. Furthermore, fuzzy measures are monotonic with respect to set inclusion. Their definition is as follows.

A fuzzy measure μ is a set function $\mu\colon \wp(X) \to [0,1]$ that satisfies:

- $\mu(\emptyset) = 0$,
- $\mu(X) = 1$,
- $A \subseteq B$ implies $\mu(A) \le \mu(B)$.

When the measure of A is understood as its importance, the last condition in this definition states that the more sources we have, the larger their importance. Boundary conditions (the first two conditions) imply that the minimum importance (the one of the empty set) is zero and that the maximum importance (the one of the full set) is one. The boundary condition $\mu(X) = 1$ is arbitrary; nevertheless, from the point of view of the aggregation, it can be justified as being similar to the one existing for the weights in the weighted mean, and having a similar rationale. That is, the weights w_i in a weighting vector add to one: $\Sigma_i w_i = 1$. Note that such restriction is appropriate in the case of the weighted mean because it implies that the weighted mean satisfies unanimity ($\mathrm{WM}(a,\ldots,a) = a$). Note that any other arbitrary bound K (i. e., $\Sigma_i w_i = K$) would just result into a weighted mean such that unanimity does not hold but that $\mathrm{WM}(a,\ldots,a) = Ka$.

Due to the usefulness of the fuzzy measures [14], several families have been defined and studied in the literature. Some of them are the following: Sugeno lambda measures, decomposable fuzzy measures, hierarchically decomposable fuzzy measures, k-order additive fuzzy measures.

As said, fuzzy measures are used in combination with fuzzy integrals. Several fuzzy integrals have been defined in the literature. We review a few of them below.

- Choquet integral [11]: The Choquet integral of a function $f\colon X \to \mathbb{R}^+$ with respect to a fuzzy measure μ is defined by:

$$\mathrm{CI}_\mu(f) = (C)\int f\,\mathrm{d}\mu = \sum_{i=1}^{N}\left[f\left(x_{s(i)}\right) - f\left(x_{s(i-1)}\right)\right]\mu\left(A_{s(i)}\right).$$

where $f(x_{s(i)})$ indicates that the indices have been permuted so that $0 \le f(x_{s(1)}) \le \cdots \le f(x_{s(N)}) \le 1$, where $f(x_{s(0)}) = 0$ and where $A_{s(i)} = \{x_{s(i)}, \ldots, x_{s(N)}\}$.

The Choquet integral generalizes the arithmetic mean, the weighted mean, the OWA operator and the WOWA operator.

(a) It generalizes the arithmetic mean when we have that the measure is proportional to the cardinality of the set, i. e., when $\mu(A) = |A|/|X|$, the Choquet integral of f corresponds to the arithmetic mean of $a_i = f(x_i)$.

(b) Given a weighting vector $w = (w_1, \ldots, w_N)$, if we define a fuzzy measure as follows

$$\mu(A) = \sum_{x_i \in A} w_i \ ,$$

we have that the Choquet integral of f with respect to μ corresponds to the weighted mean of a_i with respect to w.

(c) Given a weighting vector $w = (w_1, \ldots, w_N)$, if we define a fuzzy measure μ as follows

$$\mu(A) = \sum_{i=1}^{|A|} w_i \ ,$$

we have that the Choquet integral of f with respect to μ corresponds to the OWA of a_i with respect to w.

(d) Given weighting vectors w and p, we can define a fuzzy measure in terms of the interpolating function w^*. Once such function is built, we define the fuzzy measure μ as follows:

$$\mu(A) = w^* \left(\sum_{i=1}^{|A|} w_i \right) \ ,$$

we have that the Choquet integral of f with respect to such μ corresponds to the WOWA of the a_i with respect to weighting vectors w and p.

- Sugeno integral [22]: The Sugeno integral of a function $f\colon X \to \mathbb{R}^+$ with respect to a fuzzy measure μ is defined by:

$$\mathrm{SI}_\mu(f) = (S) \int f \, \mathrm{d}\mu = \max_{i=1}^{N} \left(\min\left(f(x_{s(i)}), \mu(A_{s(i)}) \right) \right) .$$

Here, s and $A_{s(i)}$ are defined as above.

Other fuzzy integrals exist. Among them, we have the twofold integral that generalizes both Choquet and Sugeno integrals. The generalization uses two fuzzy measures μ_C and μ_S that correspond, respectively, to the measures used in the Choquet and Sugeno integrals. The measure of the Choquet integral has a probabilistic flavor as the Choquet integral generalizes the weighted mean, that can be seen as the expectation of the a_i with respect to the probability distribution $w = (w_1, \ldots, w_N)$. Note that the weighting vector in the weighted mean is compatible with a probability distribution because weights are positive and add to one. Instead, the Sugeno integral has a fuzzy flavor as it generalizes the weighted min. and the weighted max., which are used in fuzzy systems. Such operators are defined in terms of weighting vectors that do not add to one but have a maximum of one. So, they are like possibility distributions.

In short, the twofold integral integrates a function with respect to two fuzzy measures μ_C and μ_S. We define such integral below.

- Twofold integral [17,23]: The twofold integral of a function $f\colon X \to \mathbb{R}^+$ with respect to a fuzzy measure μ is defined by:

$$\mathrm{TI}_{\mu_\mathrm{S},\mu_\mathrm{C}}(f) = \sum_{i=1}^{N} \left(\left(\max_{j=1}^{i} \left(\min\left(f(x_{s(i)}), \mu_\mathrm{S}(A_{s(i)}) \right) \right) \right) \cdot \left(\mu_\mathrm{C}(A_{s(i)}) - \mu_\mathrm{C}(A_{s(i+1)}) \right) \right) .$$

Here, s and $A_{s(i)}$ are defined as above. Naturally, $A_{s(n+1)} = \emptyset$.

When $\mu_\mathrm{C} = \mu^*$, the twofold integral reduces to the Sugeno integral and when $\mu_\mathrm{S} = \mu^*$, the twofold integral reduces to the Choquet integral. Here, μ^* represents ignorance and is defined by $\mu(\emptyset) = 0$ and $\mu^*(A) = 1$ when $A \neq \emptyset$. In addition, when $\mu_\mathrm{C} = \mu_\mathrm{S} = \mu^*$, we have that the twofold integral reduces to the maximum.

Future Directions

Aggregation operators have been studied for a long time. For example, the inequality

$$\mathrm{HM}(x, y) \le \mathrm{GM}(x, y) \le \mathrm{AM}(x, y) \ ,$$

that involves the harmonic mean (HM), the geometric mean (GM), and the arithmetic mean (AM) was already known by Pappus of Alexandria (fl. c. 300–c. 350). It is given in his book Synagoge (book III). Since then, a large number of aggregation operators have been proposed fostered by the new applications and the theoretical studies.

From the practical point of view, current research is oriented towards the embedding of such operators into real applications. This has motivated the study and the de-

velopment of methods and techniques for parameter determination. e. g., the development of methods that permits a user to find the appropriate parameters in a given application. For example, methods have been defined to determine the weights in the weighted mean, and to determine the fuzzy measure in a fuzzy integral. This is an active line of research.

Another related topic is the selection of the appropriate function. Now, the methods for parameter determination mainly presume that the appropriate operator is known beforehand and that the only elements to be determined are the parameters. Tools and methods are needed to help the users to select which is the appropriate function.

Finally, there is the need for developing architectures for information integration. That is, architectures that encompass all the processes related with fusion: from the input data (acquisition of the data) to the final decision or output.

All these aspects, oriented towards the real application of aggregation operators are combined with new research on the theoretical side. Such research includes the definition of new functions (needed so that systems can work on new types of data), the characterization of the functions (needed so that we can know which are the properties of the functions and so that a user can then select the appropriate one based on his/her problem requirements) and the study of composite models using aggregation operators (i. e., hierarchical models that combine several aggregation operators).

Bibliography

Primary Literature

1. Aczél J (1966) Lectures on functional equations and their applications. Academic, New York, London
2. Aczél J (1987) A short course on functional equations. Reidel, Dordrecht
3. Arnold BC, Balakrishnan N, Nagaraja HN (1992) A first course in order statistics. Wiley, New York
4. Arrow KJ (1951) Social choice and individual values, 2nd edn. Wiley, New York
5. Arrow KJ, Sen AK, Suzumura K (eds) (2002) Handbook of social choice and welfare. Elsevier, Amsterdam
6. Bajraktarević M (1958) Sur une équation fonctionnelle aux valeurs moyennes. Glasnik Mat Fiz I Astr 13(4):243–248
7. Barthelemy JP, McMorris FR (1986) The median procedure for n-trees. J Classif 3:329–334
8. Bouyssou D, Marchant T, Pirlot M, Perny P, Tsoukiàs A, Vincke P (2000) Evaluation and decision models: A critical perspective. Kluwer's International Series. Kluwer, Dordrecht
9. Bullen PS (2003) Handbook of means and their inequalities. Kluwer, Dordrecht
10. Bullen PS, Mitrinović DS, Vasić PM (1988) Means and their Inequalities. Reidel, Dordrecht
11. Choquet G (1953) Theory of capacities. Ann Inst Fourier 5:131–295
12. Fishburn PC, Rubinstein A (1986) Aggregation of equivalence relations. J Classif 3:61–65
13. Fodor J, Roubens M (1994) Fuzzy preference modelling and multicriteria decision support. Kluwer, Dordrecht
14. Grabisch M, Murofushi T, Sugeno M (2000) Fuzzy measures and integrals: Theory and applications. Physica, Heidelberg
15. Hardy GH, Littlewood JE, Pólya G (1934) Inequalities, 2nd edn. Cambridge University Press, Cambridge
16. Hirsch JE (2005) An index to quantify an individual's scientific research output. Proc Natl Acad Sci 102(45):16569–16572
17. Narukawa Y, Torra V (2004) Twofold integral and Multi-step Choquet integral. Kybernetika 40(1):39–50
18. Mitchell HB (2007) Multi-sensor data fusion. An introduction. Springer, Heidelberg
19. Pappus (1982) La collection mathématique. Librairie Scientifique et Technique Albert Blanchard, Paris
20. Roy B (1996) Multicriteria methodology for decision aiding. Kluwer, Dordrecht
21. Ruspini EH, Bonissone PP, Pedrycz W (1998) Handbook of fuzzy computation. IOP, London
22. Sugeno M (1974) Theory of fuzzy integrals and its applications. Ph D dissertation, Tokyo Institute of Technology, Japan
23. Torra V (1997) The weighted OWA operator. Int J Intell Syst 12:153–166
24. Torra V (2003) La integral doble o *twofold* integral: Una generalització de les integrals de Choquet i Sugeno. Butlletí de l'Associació Catalana d'Intel·ligència Artificial 29:13–19. Preliminary version in English: Twofold integral: A generalization of Choquet and Sugeno integral. IIIA Technical Report TR-2003-08
25. Torra V, Narukawa Y (2007) Modeling decisions: Information fusion and aggregation operators. Springer, Heidelberg
26. Torra V, Narukawa Y (2008) The h-index and the number of citations: Two fuzzy integrals. IEEE Trans Fuzzy Syst 16(3): 795–797

Books and Reviews

[26] gives a general description of the field of aggregation operators, it defines the main operators and discusses a few practical topics about their applications (e. g. parameter determination). (Calvo et al, 2002) is an edited book that contains state-of-the-art chapter on different topics related with aggregation and fusion. A few properties on the aggregation operators (mainly related with inequalities) can be found in the books by Bullen [9] and Bullen, Mitrinović and Vasić [10], and the excellent book by Hardy, Littlewood and Pólya [34]. [14] is an edited volume on fuzzy measures and fuzzy integrals.

[18] is an introduction to multisensor data fusion, a topic very much related with aggregation operators

Alsina C, Frank MJ, Schweizer B (2006) Associative functions: Triangular norms and copulas. World Scientific, Singapore

Calvo T, Mayor G, Mesiar R (2002) Aggregation operators. Physica, Heidelberg

Pap E (2002) Handbook of measure theory, vols I, II. North-Holland, Amsterdam

Yager RR (1988) On ordered weighted averaging aggregation operators in multi-criteria decision making. IEEE Trans Syst Man Cybern 18:183–190

Air Traffic Control, Complex Dynamics of

Banavar Sridhar, Kapil Sheth
NASA Ames Research Center, Moffet Field, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Complex Network Analysis
Current US Air Traffic Network
Future Air Traffic Scenarios
Concluding Remarks
Bibliography

Glossary

Air traffic flow Air Traffic Flow represents the dynamic distribution of air traffic over a region of space. Air traffic is undergoing major changes both in developed and developing countries. The demand for air traffic depends on population growth and other economic factors. Air traffic in the United States is expected to grow to 2 or 3 times the current levels of traffic in the next few decades. An understanding of the characteristics of the current and future flows is essential to design a good traffic flow management strategy.

Traffic flow management Safety limits the number of aircraft arriving at an airport or in a region of the airspace. Airspace Capacity, the maximum number of aircraft in a region, depends on the technology to keep aircraft separated by a safe distance and weather conditions. Airspace capacity decreases in the presence of severe weather and aircraft may have to be rerouted or delayed on the ground to maintain safety. The imbalance between airspace capacity and traffic flow demand leads to delays. Traffic flow management tries to maintain efficiency of the flows while not exceeding capacity limits.

Complex networks A network connects components of a system. The connections and the number of components vary with the function of the network. It is extremely difficult to analyze and visualize the behavior of networks when the number of components in the system becomes large. Recently, there has been a major advance in our understanding of the behavior of networks with large number of components. Several theories have been advanced about the evolution of large biological and engineering networks by authors in diverse disciplines like physics, mathematics, biology and computer science.

Scale-free networks Several large biological and engineering networks exhibit a scale-free property in the sense that the probabilistic distribution of their nodes as a function of connections decreases slower than an exponential. These networks are characterized by the fact that a small number of components have a disproportionate influence on the performance of the network. Scale-free networks are tolerant to random failure of components; but are vulnerable to selective attack on components.

Definition of the Subject

Civil Aviation is a vital sector of the US economy. In 2004, air transportation and related industries generated an output of $1.37 trillion employing 12.3 million people in the US [5]. The National Airspace System (NAS) refers to all the hardware, software and people involved in managing air traffic in the United States. More than 50 000 commercial flights operated in the US airspace alone on a typical day at the present time. Demand for air transportation has seen a six-fold increase in the past 30 years and estimates call for a strong average growth rate of 4.7% during the next 20 years [1]. This increase in demand will put a further strain on the airports and airspace resulting in delays. Air Transport Association, a group representing airlines, estimated the cost of delays to airlines in 2005 at 5.9 billion dollars. To address the changes required in the air transportation system, the US has created a multi-agency Joint Planning and Development Office to lead the required transformations. Similar activities are being pursued in Europe and Asia [14].

Introduction

Air traffic in the United States continues to grow at a steady pace except for a dip immediately after the tragic events of September 11, 2001. There are different growth scenarios associated both with the magnitude and the composition of the future air traffic. The Terminal Area Forecast (TAF), prepared every year by the Federal Aviation Administration (FAA), projects the growth of traffic in the United States [4]. Both Boeing and Airbus publish market outlooks for air travel annually. Although predicting the future growth of traffic is difficult, there are two sig-

nificant trends: (1) heavily congested major airports continue to see an increase in traffic, and (2) the emergence of regional jets and other smaller aircraft with fewer passengers operating directly between non-major airports. The interaction between air traffic demand and the ability of the system to provide the necessary airport and airspace resources can be modeled as a network. The size of the resulting network varies depending on the choice of its nodes. It would be useful to understand the properties of this network to guide future design and development. Many questions, such as the growth of delay with increasing traffic demand and impact of the en route weather on future air traffic, require a systematic understanding of the properties of the air traffic network.

Recently, there has been a major advance in the understanding of the behavior of networks with large number of components. Several theories have been advanced about the evolution of large biological and engineering networks by authors in diverse disciplines like physics, mathematics, biology and computer science [21]. Several networks exhibit a scale-free property in the sense that the probabilistic distribution of their nodes as a function of connections decreases slower than an exponential. These networks are characterized by the fact that a small number of components have a disproportionate influence on the performance of the network. Scale-free networks are tolerant to random failure of components, but are vulnerable to selective attack on components [2,17].

This paper examines two network representations for the current air traffic system. A network defined with the 40 major airports as nodes and with standard flight routes as links has a characteristic scale: all nodes have 60 or more links and no node has more than 460 links. Another network defined with current aircraft routing structure exhibits an exponentially truncated scale-free behavior. Its degree ranges from 2 connections to 2900 connections, and 217 nodes have more than 250 connections. Furthermore, those high-degree nodes are homogeneously distributed in the airspace. A consequence of this scale-free behavior is that the random loss of a single node has little impact, but the loss of multiple high-degree nodes (such as occurs during major storms in busy airspace) can adversely impact the system. Two future scenarios of air traffic growth are used to predict the growth of air traffic in the United States. It is shown that a three-times growth in the overall traffic may result in a ten-times impact on the density of traffic in certain parts of the United States.

The paper is organized as follows. The Sect. "Complex Network Analysis" provides a brief overview and terminology of the complex networks useful in the analysis of air traffic. This is followed by an application of the complex network methodology to the current air traffic system. The Sect. "Future Air Traffic Scenarios considers different scenarios for the evolution of air traffic in the United States during the next 25 years and looks for changes in the behavior of the network. An analysis of the future air traffic scenarios shows that a three-times growth in overall traffic can result in a tenfold impact on the density of highly connected nodes in certain parts of the United States. Thus, in addition to bottlenecks at major airports, the risk of airspace congestion calls for route restructuring and the introduction of new procedures and automation to increase airspace capacity. Concluding remarks and possible research in this area are discussed in the last section.

Complex Network Analysis

Complex systems have many agents or components interacting with each other, and their collective behavior is not a simple combination of the individual behavior. The pattern of interaction between agents can be studied as a network of connections between agents. Networks of many types are pervasive in modern society, and scientists from many fields are trying to broaden their understanding of the behavior of the networks. A network is made up of basic components called either vertices or nodes. Each node is connected to other nodes in the system. The line connecting two nodes is referred to as a link or an edge. An edge is directed if it runs only in one direction and undirected if it runs in both directions. Degree is the number of edges connected to a node. Figure 1 shows a simple network with nodes A, B, C and D and undirected edges connecting them. The degrees of nodes A, B, C and D are 3, 2, 2 and 1, respectively.

The structure of the network strongly influences the functions performed by the network. It is possible to analyze networks of small sizes (less than 100) by drawing the picture of the network and analyzing its properties. The number of nodes in a complex engineering system, such as the worldwide web [10], can easily be as large as a few million nodes. The behavior of Random Graphs, graphs

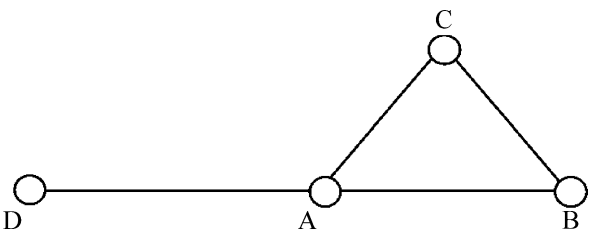

Air Traffic Control, Complex Dynamics of, Figure 1
Nodes and edges in a network

where nodes are connected to other nodes randomly, as the number of nodes becomes large has been studied extensively by mathematicians [13]. The behavior of Regular Networks, idealized systems where each component is identical, has been studied by several authors [17]. Networks describing real systems are neither random nor regular. The ability to model real systems and capture the behavior of key variables is extremely difficult. Recent developments in complex networks examine the statistical properties of large networks and help answer questions about the dynamics and stability of such networks. However, studies on the effects of structure on system behavior are still in their early stages.

A major area of interest is the role played by the distribution of the degree of nodes in a network. Let p_k be the fraction of nodes in the network that has degree k. The degree distribution for the network can be computed by generating a histogram of p_k. The degree distribution for large random graphs, where each edge is present or absent with equal probability, can be modeled as a Poisson distribution. The distribution tends to peak for a small value of k and decays exponentially for large values of k. When using real data, to reduce the effect of noise in small datasets, the distribution of degree is expressed in terms of the complementary cumulative distribution function (ccdf), P_k, the probability of the number of nodes in the graph with degree greater than k. P_k is computed using the expression

$$P_k = \sum_{i=k}^{\infty} p_i \,.$$

Figure 2 shows the ccdf of the Poisson distribution, $e^{-\lambda}\lambda^k/k!$, for $\lambda = 4$. Figure 3 shows the ccdf for the Poisson distribution on a logarithmic scale.

It has been observed that the degree distributions for some real-world networks, such as the Internet and biological networks, are highly skewed with tails several times longer than the mean. These real networks have a small number of nodes, or hubs, with a high level of connectivity. Hubs play an important role in influencing the properties of the network. Real networks exhibiting a small number of hubs are referred to as "scale-free" networks [6]. The distributions of degree for a number of networks, e. g., Internet, World Wide Web, collaboration network of mathematicians, etc., show a power law in their tails: $P_k \approx k^{-\alpha}$ for some constant value of α. Figure 4 shows the ccdf for $\alpha = 2$. Power law distributions (Fig. 4) appear linear in a logarithmic scale (Fig. 5).

The appearance of hubs in scale-free networks is explained in terms of two behaviors both in people and networks, namely, real systems-growth and preferential at-

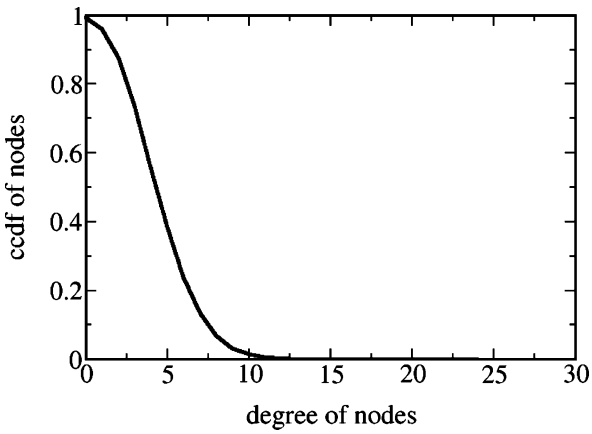

Air Traffic Control, Complex Dynamics of, Figure 2
Poisson distribution

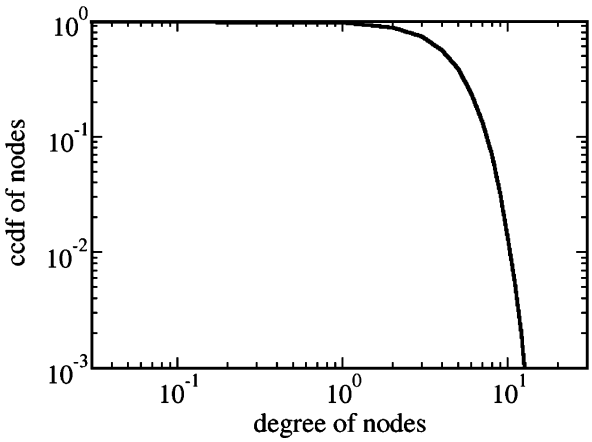

Air Traffic Control, Complex Dynamics of, Figure 3
Poisson distribution in logarithmic scale

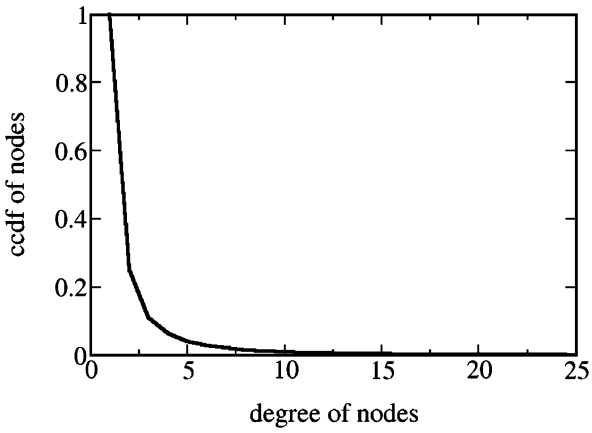

Air Traffic Control, Complex Dynamics of, Figure 4
Power law distribution

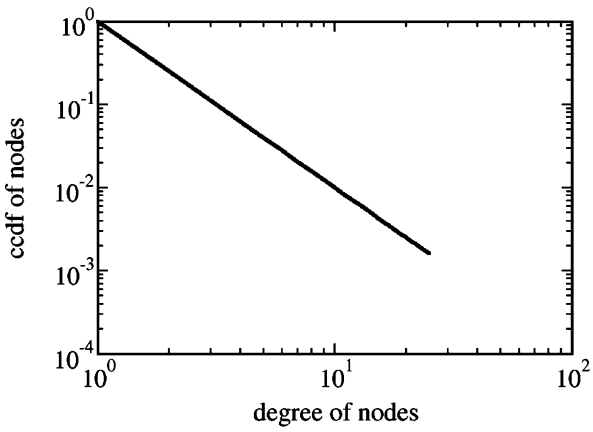

Air Traffic Control, Complex Dynamics of, Figure 5
Power law distribution in logarithmic scale

tachment. There is a tendency for new growth to gravitate towards desirable locations. Generally, all the desirable locations already have some existing nodes. The preferential attachment mechanism leads to the creation of more powerful nodes or hubs. The various hubs compete to attract new nodes and absorb them, resulting in an increase of the number of links in the initial hubs.

Scale-free networks with power law distributions exhibit two important properties: (a) remarkable resistance to random failure of nodes and (b) extreme vulnerability to targeted attacks. The functionality and performance of the network depends on the existence of the edges between pairs of nodes. The distance between the remaining vertices gets longer as nodes are removed from a network and the removal of a certain number of nodes may result in a collapse of the entire system. The tolerance of networks to failures or removal of nodes varies with their level of resilience. The nodes from a network could be removed randomly or in a coordinated approach. Several studies have shown that scale-free networks are generally robust to random removal of nodes. However, they are less tolerant to the selective removal of hubs.

The next section will examine the air traffic in the United States to see if it exhibits the properties of a scale-free network and draw analogies from the attack vulnerabilities of other complex networks. The methodology can also be used to understand the properties of the Next Generation Air Transportation System as it evolves over the next few decades [18].

Current US Air Traffic Network

The National Airspace System (NAS) refers to the collection of hardware, software and people, including runways, radars, networks, FAA, airlines, etc., involved in Air Traffic Management (ATM) in the US. It has been pointed out that the NAS should be modeled as a complex adaptive system [12]. The NAS can be looked at as a network at several different levels [11]. Some examples of networks involving components of the NAS are the route network of an airline connecting different airports, a network of sectors (geographical region used to monitor safe separation between aircraft) interconnected by geographical proximity, network of airline crew located in different cities and a surface network of ramps, runways and departure gates at an airport. The network components and links vary with the planning interval and the modeling problem. The analysis of current traffic uses data from 2006 and constructs two different types of networks for a typical day of operations in the NAS. The analysis is based on actual traffic for a day in July 2006. The traffic data are processed using Future Air traffic management Concept Evaluation Tool (FACET) simulation software [19] for conducting the network flow analysis.

The major distinctions regarding the two air traffic networks are described in Table 1. The nodes in the Airport Network (AN) correspond to the 40 major airports in the US. Today, a set of predefined alternative routes are used for flying between particular city pairs. Each node in network AN is connected to another node through one or more routes. The Airport and Airspace Network (AAN) includes all airports in the US and routes connecting these airports. A route between two airports is defined by a series of geographical positions or fixes in the airspace. Each fix together with the airports represent a node in the AAN network.

Computation of the Degree of a Node and the Distribution Function

The computation of the degree of a node and the distribution function is illustrated by an example. The FACET simulation software generates the departure airport, arrival airport and the fixes through which the aircraft travels en route based on air traffic data. Figure 6 shows the intersection of different routes connecting airports through common fixes. In the figure, aircraft traveling from Seattle (SEA) to Atlanta (ATL), aircraft traveling from Oakland (OAK) to Newark (EWR), and aircraft traveling from Los Angeles (LAX) to Boston (BOS) pass through the fix Garden City, Kansas (GCK). The degree of GCK is equal to the total number of aircraft traveling along these three routes during the day. The degree of node GCK is three if a single aircraft travels between each of these three pairs

Air Traffic Control, Complex Dynamics of, Table 1
Different types of Air Traffic Management networks

Network	Nodes	Number of Nodes	Edges
Airport Network (AN)	40 major airports	40	Routes between 40 major airports
Airport and Airspace Network (AAN)	All airports and fixes along routes connecting the airports	8170	Routes between all airports

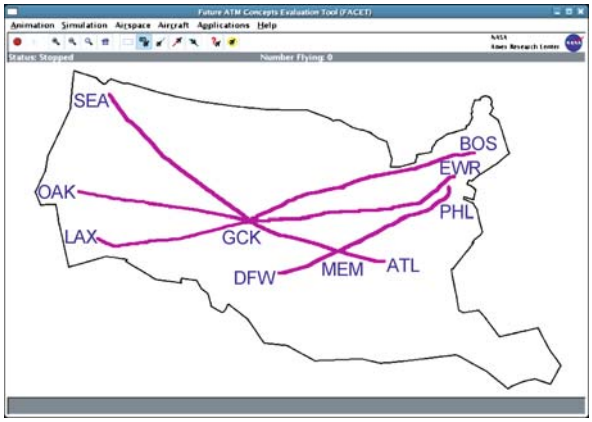

Air Traffic Control, Complex Dynamics of, Figure 6
Example of intersection of routes at common fixes

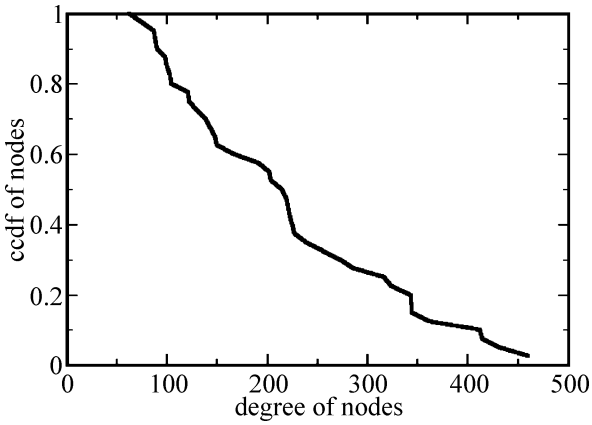

Air Traffic Control, Complex Dynamics of, Figure 7
Cumulative distribution function for nodes (40 major airports) in network AN

of airports. Similarly, aircraft traveling from Dallas (DFW) to Philadelphia (PHL) and Seattle to Atlanta pass through the fix Memphis, TN (MEM). The degree of fix MEM is two if a single aircraft travels between each of these city pairs. The distribution and the ccdf can be computed easily given the degree associated with all the airports and en route fixes and are described in the next section.

Behavior of the Degree of Nodes in Air Traffic Management Network

The ccdf for the nodes in network AN is shown in Fig. 7. The figure shows the probability that an airport in the network has more than a certain number of routes originating or terminating at the airport. The number of connections at all airports exceeds 60 and Chicago O'Hare airport has the maximum number (460) of connections. The distribution of nodes in the airport network does not show scale-free behavior. The airport connection distribution in the U.S is similar to the distribution of the connections between the networks of world airports [3].

Figure 8 shows the ccdf of the nodes in network AAN. The ccdf of the nodes in network AAN shows a rapid decay of nodes with the degree of the node. However, it has 217 nodes with more than 250 links, indicative of high volumes

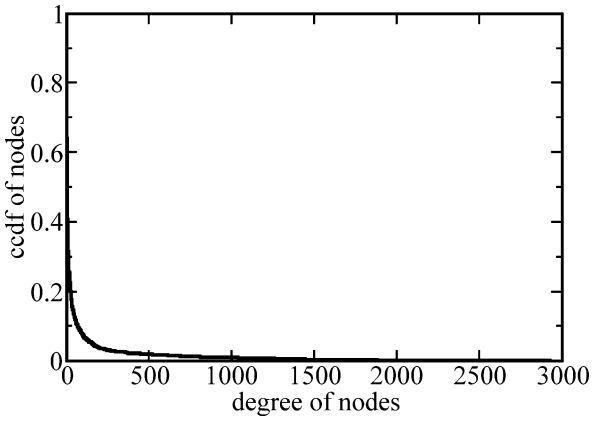

Air Traffic Control, Complex Dynamics of, Figure 8
Distribution of nodes in AAN with current traffic

of traffic through these nodes. Figure 9 shows the logarithmic behavior of the ccdf of the nodes in network AAN.

The nodes of network AAN initially follow the power law curve similar to several scale-free networks [17]. However, towards the end of the tail, the distribution shows a deviation from the power law behavior indicating a limit on the number of nodes that have a high number of connections. The limit on the growth of the large hub nodes, in the distribution of the ATM network, is due to the con-

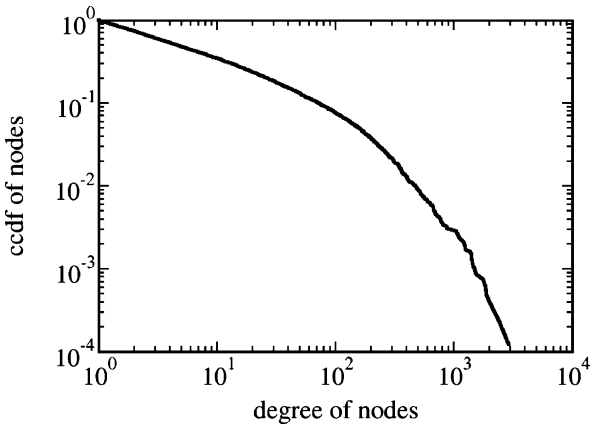

Air Traffic Control, Complex Dynamics of, Figure 9
Logarithmic behavior of nodes in network AAN

Air Traffic Control, Complex Dynamics of, Figure 10
Geographical distribution of 250G nodes in network AAN with current traffic and severe weather polygons

straints imposed on traffic demand by the location of the cities, economic development and government policies.

The existence of hub and secondary airports characterizes current air traffic operations [8]. The network analysis provides an additional ability to study traffic behavior in the en route airspace. Figure 10 shows geographical distribution of the nodes with degree higher than 250 in network AAN. These nodes will be referred to as 250G nodes. 250G nodes represent 2.7% of the total nodes in the network. Table 2 shows the number of 250G nodes in different traffic control regions, referred to as Centers. The Centers, Chicago (ZAU), Boston (ZBW), Atlanta (ZTL), Kansas City (ZKC), Washington, DC (ZDC), Indianapolis (ZID), Jacksonville (ZJX), Los Angeles (ZLA), Cleveland (ZOB) and New York (ZNY), each have more than ten of the 250G nodes. The Centers in the western part of the United States have fewer 250G nodes indicative of the lower traffic density in these Centers. It is informative to express the distribution of the 250G nodes in units of number of nodes per 10,000 square nautical miles (10 Ksqnm). Using this measure of nodal traffic density, ZNY has the highest nodal density with six 250G nodes per 10 Ksqnm. Table 3 shows the nodal traffic density for the 20 Centers in the continental US.

The behavior of the degree of the nodes in an ATM network should not be surprising, since the ATM network has evolved to serve population densities in the US. Network analysis helps to visualize and quantify the characteristics of the network.

Resilience of Air Traffic Management Networks

The tolerance of complex networks to random and targeted failures depends on their network structure. As observed earlier, a scale-free network is tolerant to random failure, since the hubs are few and the chance of a hub being selected randomly is low. However, the same network may be prone to targeted attacks on a small percentage of vital nodes. In air traffic management networks, weather can be regarded as an agent of attacks on the system.

Convective weather is a major source of uncertainty in ATM networks. One effect of severe weather is to make airspace unavailable for the flow of air traffic. The removal of airspace may result in the failure of hubs and increase the average path length in the network. The impact of weather on ATM system performance appears as delay [20]. The tolerance of the ATM network to weather depends on the geographical distribution of the weather and the coverage of nodes with high degree.

Future Air Traffic Scenarios

National and international projections of traffic growth indicate a tripling of passengers by 2025 [7]. There may be increased traffic due to the growing presence of on-demand air taxis and unmanned air vehicles. It is estimated that 5000 micro jets may be operational by 2010 and 13 500 by 2022 [22]. The TAF provides forecasts for airports in the NAS. The forecast for the major airports in the United States receives more detailed modeling that takes into account local economic conditions and airline costs. TAF is the basis for most aviation demand forecasts. The TAF data published in March 2006 provides traffic growth rates for the period 2004–2020. The AvDemand model [16] provides three times (3X) current traffic scenarios starting with TAF and assuming slightly different conditions for traffic growth beyond 2025. The Transportation System Analysis Model (TSAM) [23] predicts

future traffic demand based on demographics and population at the county level. It provides a complimentary alternative to the FAA forecast. The results in this study are based on data generated by AvDemand.

Two future scenarios are considered. These two scenarios and the current traffic provide a baseline to compare the impact of future traffic changes on the ATM system. The initial values of aircraft routes and schedules used in the extrapolation are based on traffic data for a day in May 2002. Assuming traffic growth from TAF and the nominal traffic schedule, Fratar algorithm [23] can be used to create future daily total number of flights between each origin and destination pair of the current schedule. The traffic growth scenarios using TAF airport growth out to 2025 predict a future NAS-wide traffic growth by a factor of 1.49. To achieve 3-times current day traffic (3X) scenarios assumed in future planning [9], consider two different scenarios for post-2025 traffic growth. The compound extrapolation approach grows traffic until it reaches 3X by assuming the TAF airport growth rates for traffic beyond 2025. As an alternative, the homogeneous extrapolation approach assumes the same growth rate at all airports until the traffic level reaches 3X. These two scenarios will be referred to as 3X Compound (3XC) and 3X Homogeneous (3XH) scenarios in the rest of the chapter. The scenarios are processed to compute the changes in the network properties of future ATM systems.

Behavior of Future ATM Networks

The two scenarios described earlier can be used to study the characteristics of the 3X traffic under the two assumptions. The properties of future ATM networks can be derived similar to the properties of the current ATM networks. The computations can be used to compare the geographical distribution of the hubs compared to the distribution today. Figures 11 and 12 show the distribution of the nodes in the AAN network for the 3XC traffic scenario. The results are similar for the 3XH traffic scenario.

The future ATM networks, under both scenarios, show exponentially truncated scale-free behavior similar to the current ATM network. The distribution of the hubs will have an impact on the growth of delay in future ATM networks subjected to severe weather. The vulnerability of the network to reduction in capacity caused by certain weather patterns will be significant and disruptive to the operation of the system. The impact of the weather on future ATM networks will be described subsequently in the paper.

The geographical distribution of the 250G nodes under 3XC is shown in Fig. 13. Table 2 shows the same distribu-

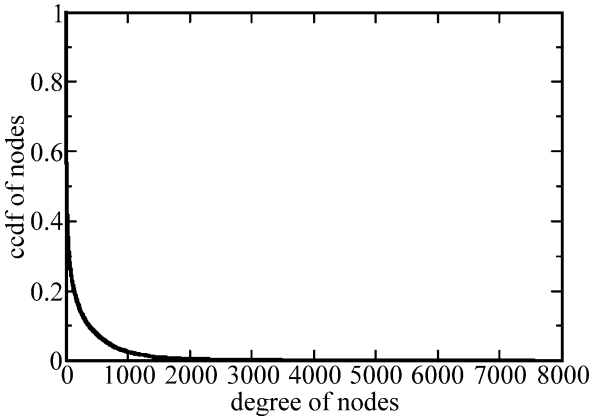

Air Traffic Control, Complex Dynamics of, Figure 11
Distribution of nodes in AAN using 3X Compound Scenario

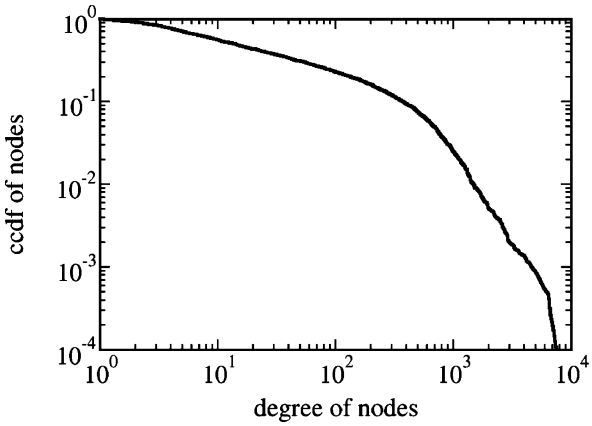

Air Traffic Control, Complex Dynamics of, Figure 12
Logarithmic behavior of nodes in network AAN under 3X Compound traffic

Air Traffic Control, Complex Dynamics of, Figure 13
Geographical distribution of 250G nodes in network N5 with 3X Compound traffic and severe weather polygons

Air Traffic Control, Complex Dynamics of, Table 2
Distribution of 250G nodes by Centers under current and future traffic scenarios

Center	Current	3XH	3XC
ZAB	7	70	78
ZAU	15	67	70
ZBW	16	77	87
ZDC	22	134	143
ZDV	8	51	68
ZFW	3	59	68
ZHU	8	60	72
ZID	15	71	71
ZJX	14	71	81
ZKC	11	58	50
ZLA	16	102	119
ZLC	5	29	40
ZMA	4	44	54
ZME	10	38	42
ZMP	6	46	53
ZNY	14	80	88
ZOA	10	57	67
ZOB	13	90	89
ZSE	5	35	41
ZTL	15	59	62
Total	217	1298	1443

Air Traffic Control, Complex Dynamics of, Table 3
Nodal density of 250G nodes by Centers under current and future traffic scenarios

Center	Area	ρ1X	ρ3XH	ρ3XC
ZAB	18.04	0.39	3.88	4.32
ZAU	7.60	1.97	8.82	9.21
ZBW	11.62	1.38	6.63	7.49
ZDC	12.69	1.73	10.56	11.27
ZDV	20.18	0.40	2.53	3.37
ZFW	12.32	0.24	4.79	5.52
ZHU	27.64	0.29	2.17	2.61
ZID	7.07	2.12	10.04	10.04
ZJX	14.55	0.96	4.88	5.57
ZKC	13.37	0.82	4.34	3.74
ZLA	13.55	1.18	7.53	8.78
ZLC	31.57	0.16	0.92	1.27
ZMA	28.66	0.14	1.53	1.88
ZME	10.75	0.93	3.53	3.91
ZMP	28.82	0.21	1.60	1.84
ZNY	2.43	5.76	32.92	36.21
ZOA	13.54	0.74	4.21	4.95
ZOB	6.74	1.93	13.36	13.21
ZSE	19.64	0.25	1.78	2.09
ZTL	9.18	1.63	6.43	6.75
Total	309.96	0.70	4.19	4.66

tion by Centers for current traffic and 3X scenarios. An examination of Table 2 is helpful in drawing distinctions between current day traffic and 3X demand traffic. The 3X traffic demand creates close to a six-fold increase in the 250G traffic nodes. Earlier, it was noted that ZNY has the highest 250G traffic nodal density. Under the 3X demand, as can be seen from Table 3, ZNY nodal density of 2006 is equaled or exceeded by a majority of the 20 Centers. Even more alarmingly, the nodal density of 250G nodes in ZNY is 6 times the current value, and Cleveland and Washington Centers have twice the nodal density of ZNY in 2006. The 3X traffic also gives rise to nodes with even more connections. The hubs, 500G, 750G and 1000G, are defined similarly to the 250G hub. The number of 250G, 500G, 750G and 1000G hubs under different traffic conditions are shown in Table 4.

Air Traffic Control, Complex Dynamics of, Table 4
Total number of nodes for different traffic scenarios

Types of nodes	Current	3XH	3XC
250G	217	1312	1468
500G	72	620	806
750G	33	304	452
1000G	22	165	262

Impact of Weather on ATM Networks

Another way to view the growth of traffic is to compare how similar weather patterns may affect current traffic and 3X traffic. The geographical distribution of the 250G nodes shown in Fig. 13 is based on traffic during July 2, 2006. This was a calm weather day with a total NAS aggregate delay of 11 997 minutes [15]. The traffic on July 2, 2006 will be assumed as traffic unaffected by weather in this discussion [20].

Whenever there is severe weather in the NAS, airspace capacity is reduced and traffic is rerouted or held on the ground causing delays in the system. The Collaborative Convective Forecast Product (CCFP) is a model of severe weather activity and the areas marked blue in Fig. 10 shows the CCFP for July 13, 2006. On July 13, 2006 the NAS experienced a significant total delay of 219 350 minutes with ZNY, ZDC and ZTL Centers contributing a delay of 116 654, 49 400 and 23 822 minutes respectively. Next, as shown in Fig. 13, the same severe weather is overlaid in blue polygons on the geographical distribution of the 250G

Air Traffic Control, Complex Dynamics of, Table 5
Total number of nodes affected by severe weather

Types of Nodes	1X	3XH	3XC
250G	34	191	207
500G	9	96	118
750G	4	49	67
1000G	3	23	41

nodes under 3XC demand. Table 5 shows hub nodes affected by the weather today and in the future.

A greater number of hubs, between six to ten times, are affected by the same weather pattern under future traffic scenarios than today. If one considers the non-linear growth of delay, in regions such as ZNY, ZDC and ZTL where the demand on the airspace is close to capacity, the increased density of high traffic nodes in these regions will result in much larger delays compared to 2006.

Concluding Remarks

Air traffic in the United States can be modeled as a network to understand the impact of predicted growth in demand on the performance of the system. It is demonstrated that the air traffic network with current en route flight plan intersections as nodes and with the flight plans as links shows scale-free properties typical of several large engineering and biological networks. A consequence of this property is the non-linear growth of traffic in certain regions of the United States. A preliminary analysis indicates that a three-times growth in the overall traffic may result in a ten-times impact on the density of traffic in certain parts of the United States. The air traffic system currently experiences significant delay during periods of severe weather activity. The impact of weather of same severity will be magnified several times, especially in the northeastern parts of the United States, leading to lower system performance. Research must be conducted to determine whether this risk can be mitigated through restructuring routes or by introducing new operational concepts, such as automation assistance to controllers to increase airspace capacity. The network analysis described in the chapter can be used to guide the development of various traffic flow management concepts to increase the efficiency of air traffic systems.

Bibliography

1. Airbus SAS (2002) Airbus Global Market Forecast 2001–2002. Reference CB 390.0008/02
2. Albert R, Jeong H, Barabasi AL (2000) Error and attack tolerance of complex networks. Nature 406:378–382
3. Amaral LAN, Scala A, Barthelemy M, Stanley HE (2000) Classes of Small-world Networks. Proc Natl Acad Sci USA 97:1149–1152
4. Anonymous (2005) Terminal Area Forecast Summary, Fiscal Years 2004–2020. Report No. FAA-APO-05-1, US Department of Transportation, Federal Aviation Administration
5. Anonymous (2006) Commercial Aviation and the American Economy. The Campbell-Hill Aviation Group Inc, Washington DC
6. Barabasi AL, Bonabeau E (2003) Scale-Free Networks. Sci Am 288:60–69
7. Boeing (2007) Current Market Outlook. http://www.boeing.com/commercial/cmo Accessed June 26, 2007
8. Bonnefoy PA, Hansman RJ (2004) Emergence and Impact of Secondary Airports in the United States. 4th AIAA Aviation Technology, Integration and Operations Conference, Chicago, IL, September 2004
9. Borener S, Carr G, Ballard D, Hasan S (2006) Can NGATS Meet the Demands of the Future? J Air Traffic Control 48(1):34–38
10. Broder A, et al (2000) Graph structure in the web. Comput Netw 33:309–320
11. DeLaurentis D, Han E, Kotegawa T (2006) Establishment of a Network-based Simulation of Future Air Transportation Concepts. 6th AIAA Aviation Technology, Integration and Operations Conference, Wichita, KS, September 2006
12. Donohue G (2003) Air Transportation is a complex Adaptive System: Not an Aircraft Design. AIAA/ICAS International Air and Space Symposium and Exposition, Dayton, Ohio, July 2003
13. Erdos P, Renyi A (1960) On the Evolution of Random Graphs. Publ Math Inst, Hung Acad Sci 5:17–61
14. Eurocontrol (2007) An Assessment of Air Traffic Management in Europe during the Calender Year 2006: Performance Review Report. Eurocontrol, Brussels, Belgium
15. Federal Aviation Administration (2004) Order 7210.55C: Operational Data Reporting Requirements. US Department of Transportation
16. Huang AS, Schleicher D, Hunter G (2006) Future Flight demand Generation Tool. 6th AIAA Aviation Technology, Integration and Operations Conference, Wichita, KS, September 2006
17. Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45:167–256
18. Pearce RA (2006) The next generation air transportation system: Transformation starts now. J Air Traffic Control 48(1): 7–10
19. Sridhar B, Sheth K, Smith P, Leber W (2005) Migration of FACET from Simulation Environment to Dispatcher Decision Support System. Proceedings of Digital Avionics Systems Conference, Washington, DC, November 2005
20. Sridhar B, Swei SM (2006) Relationship between Weather, Traffic and Delay Based on Empirical Methods. 6th AIAA Aviation Technology, Integration and Operations Conference, Wichita, KS, September 2006
21. Strogatz SH (2001) Exploring complex networks. Nature 410:268–276
22. Teal Group (2003) World Unmanned Aerial Vehicle Systems. Market Profile and Forecast, Fairfax
23. Viken J, et al (2006) NAS Demand Predictions, Transportation Systems Analysis Model (TSAM) Compared with Other Forecasts. 6th AIAA Aviation Technology, Integration and Operations Conference, Wichita, KS, September 2006

Algorithmic Complexity and Cellular Automata

Julien Cervelle[1], Enrico Formenti[2]
[1] Laboratoire d'Informatique de l'Institut Gaspard–Monge, Université Paris-Est, Marne la Vallée, France
[2] Laboratoire I3S, Université de Nice-Sophia Antipolis, Sophia Antipolis, France

Article Outline

Glossary
Definition of the Subject
Introduction
Kolmogorov Complexity
Dynamical Systems and Symbolic Factors
Cellular Automata
Algorithmic Complexity as a Demonstration Tool
Measuring CA Structural Complexity
Measuring the Complexity of CA Dynamics
Future Directions
Acknowledgments
Bibliography

Glossary

Algorithmic complexity of object x shortest program for outputting a description for x (w.r.t. a universal representation system)

Equicontinuity all points are equicontinuity points (in compact settings)

Equicontinuity point a point for which the orbits of nearby points remain close

Expansivity from two distinct points orbits eventually separate

Incompressible word a word for which the shortest program outputting it has "almost" the same length as the word itself

Injectivity the next state function is injective

Kolmogorov complexity see "algorithmic complexity"

Rich configuration a configuration that contains all possible finite patterns over a given alphabet

Sensitivity to initial conditions for any point x there exist arbitrary close points whose orbits eventually separate from the orbit of x

Surjectivity the next state function is surjective

Transitivity there always exist points that eventually move from any arbitrary neighborhood to any other

Definition of the Subject

In the last 10 years the field of complex systems has enjoyed astonishing development with meaningful applications in most scientific domains.

Cellular automata (CA) are a very used model for complex systems characterized by a multitude of small identical agents capable of building a complex behavior on the basis of local interactions.

The huge variety of CA dynamical behaviors has been popularized since the early 1980s by the work of S. Wolfram (see [18] for an exhaustive review).

Later, researchers started a systematic study of CA. Most of the results are summarized in the chapters of this book (see ► Chaotic Behavior of Cellular Automata, ► Dynamics of Cellular Automata in Non-compact Spaces, ► Topological Dynamics of Cellular Automata and ► Ergodic Theory of Cellular Automata for instance).

However, many questions seem to be intractable and vanquisch researchers' efforts (some such efforts are reported by ► Chaotic Behavior of Cellular Automata, ► Dynamics of Cellular Automata in Non-compact Spaces, ► Topological Dynamics of Cellular Automata and [14]). We strongly believe that Kolmogorov complexity can be a valuable tool in the quest for answers in this domain.

Introduction

Kolmogorov (or algorithmic) complexity was introduced to precisely formalize the notion of "randomness." Former definitions of randomness were based on probabilistic concepts: for instance, a sequence of 0 and 1 is *random* if it is obtained by repeated coin tosses.

The drawback of such a definition is that it does not formally specify what is and what is not a random sequence since all sequences (of a fixed length) have the same probability to be obtained through unbiased coin tosses. However, if one considers the following two sequences:

0000000000000000000000000000

1001001001011101011010111 0101

one would say that the first one is not random while the second one is.

The idea of Kolmogorov complexity comes from this simple reasoning. What makes the first sequence simple is that it can briefly be described by the sentence "twenty-nine zeros," while the second cannot be described in a shorter way than by describing it literally: "the sequence one zero zero one zero zero one zero zero one

zero one one one zero one zero one one zero one zero one one one zero one zero one." If we use regular expressions instead of English as a description language, the first sequence is 0^{29} and the second is itself.

A first idea to formalize this process is to say that the complexity of a sequence is the length of its compressed form for a given lossless compressor. The issue of such a definition is that there is no universal compression tool as compression usually depends on the type of data (sound, picture, text). However, Kolmogorov and Solomonoff proved that there exists a universally optimal compressor that compresses sequences better up to some additive constant w.r.t. any other compressor (Theorem 1).

Kolmogorov complexity is not computable, but its combinatorial properties make it a useful demonstration tool. The major technique in Kolmogorov complexity is the *incompressibility method*, and it is briefly reviewed in Sect. "Algorithmic Complexity as a Demonstration Tool". For more on the subject see [13].

On the one hand, Kolmogorov complexity allows one to give a formal context to study randomness in discrete dynamical systems; on the other hand, it is a helpful tool for decreasing the combinatorial complexity of problems. Both of these aspects of Kolmogorov complexity will figure prominently in this chapter.

Kolmogorov Complexity

In this section we give the formal definition of Kolmogorov complexity and its main properties. For simplicity sake we restrict ourselves to the alphabet {0,1} although Kolmogorov complexity can be defined on more general alphabets and even on integers [13].

Let $\{0,1\}^*$ be the set of words on $\{0,1\}$, $|w|$ the length of word w and for $0 \leqslant i < |w|$, w_i the ith letter of w (indexed from 0), and ε the empty word.

A *partial computable function* is a computable function that is defined on a subset of $\{0,1\}^*$ (being undefined on an input means that the program that computes the function does not halt). Denote by D_φ the set of inputs on which φ is defined. The *composition* of partial computable functions f and g is the partial computable function $f \circ g$ defined on the set $\{x \in D_g,\ g(x) \in D_f\}$ by $f \circ g(x) = f(g(x))$. If f is injective, the *inverse* of f is the partial computable function f^{-1} defined by

$$f^{-1}(w) = \begin{cases} x \text{ the unique word such that } f(x) = w \text{ if it exists,} \\ \textit{undefined} \text{ otherwise .} \end{cases}$$

If f is computable, then f^{-1} is computed by the algorithm, which tries all possible x and halts when it finds one whose image is w. If no such x exists, then the program does not halt.

A *representation system* is a partial computable function φ from $(\{0,1\}^*)^2$ to $\{0,1\}^*$. It plays the role of a decompressor program used to build words from their compressed form. There are no requirements, and, in particular, it is not mandatory either that each word have a compressed form (i. e., the image set of φ need not be $\{0,1\}^*$) or that a word have only one compressed form (i. e., φ need not be injective).

Definition 1 (Kolmogorov complexity given a representation system) Let φ be a representation system. The *Kolmogorov complexity given φ of a word w knowing v*, denoted by $K_\varphi(w|v)$, is defined by the length of the shortest word y such that $\varphi(y,v) = w$ or $+\infty$ if such a word does not exist:

$$K_\varphi(w|v) = \min\{|y|,\ y \in \{0,1\}^* \text{ s.t. } \varphi(y,v) = w\} \quad \text{with the convention } \min \emptyset = \infty\,.$$

If y is such that $\varphi(y,v) = w$, we say that y is a *program* for w knowing v. If y is the shortest word such that $\varphi(y,v) = w$, we say that y is the *shortest program* for w knowing v. The *Kolmogorov complexity given φ of a word w*, denoted $K_\varphi(w)$, is defined by

$$K_\varphi(w) = K_\varphi(w|\varepsilon)\,.$$

If y is such that $\varphi(y,\varepsilon) = w$, we say that y is a *program* for w. If y is the shortest word such that $\varphi(y,\varepsilon) = w$, we say that y is the *shortest program* for w.

In the previous definition, the name *program* can be a little confusing since φ seems to be the program and y seems to be its input. This name comes from the additively optimal universal representation system introduced in Theorem 1, where the input is a program for the universal Turing machine.

As stated in the introduction, this definition lacks robustness since it depends on a particular (de)compressor. However, the following results prove that we can find a best compressor in order to define a robust notion of complexity.

Definition 2 (Additively optimal universal representation system) A representation system φ_0 is *additively optimal universal* if for any representation system φ there exists a constant c such that for all w and v in $\{0,1\}^*$

$$K_\varphi(w|v) + c \geqslant K_{\varphi_0}(w|v)\,. \tag{1}$$

This definition means that a universal representation system compresses better than any other compression system up to some additive constant.

Remark 1 The constant in Inequality (1) is mandatory since a representation system cannot be better than any other one because of very specialized compressors: for any word w, the representation system δ_w by $\delta_w(y, v) = w$ for all words y and v is one of the best for word w (i. e., $K_{\delta_w}(w|v) = 0$).

Proposition 1 *Let φ_0 and φ_1 be two additively optimal universal representation systems. Then, there is a constant c such that for all words w and v*

$$|K_{\varphi_0}(w|v) - K_{\varphi_1}(w|v)| < c\,.$$

Proof From the definition of additively optimal universal representation system, there are constants c_0 and c_1 such that for all words w and v and for all representation systems φ

$$K_\varphi(w|v) + c_0 \geqslant K_{\varphi_0}(w|v) \text{ and } K_\varphi(w|v) + c_1 \geqslant K_{\varphi_1}(w|v)\,.$$

The thesis is obtained by choosing $c = \max\{c_0, c_1\}$. □

In order to define completely Kolmogorov complexity, one needs an additively optimal universal representation system. The key idea for conceiving such a system is to embed the program for decompression in the compressed form.

In order to join the decompressing program and the compressed word in one string, we need a special combining function. We denote by $w^{\times 2}$ a word u of length $2|w|$ such that for all $0 \leqslant i < |w|$, $u_{2i} = u_{2i+1} = w_i$, which is word w, where all letters are repeated twice. Define

$$p \odot w = p^{\times 2}01w\,.$$

For instance, if $p = 001$ and $w = 1001011$, then $p \odot w = 000011\,01\,1001011$.

This way of combining words has two interesting properties. First, from $p \odot w$ a program can compute p and w since, while all bits are repeated twice, we read p, and after skipping the 01, we read w. Second, it holds that

$$|p \odot w| = 2|p| + 2 + |w|\,.$$

This last equation is fundamental to proving the additively optimality property.

Remark 2 The word $p^{\times 2}01$ is called a self-delimiting encoding for p since it encodes p in a unique way and the end of the code is computable. Hence, if one concatenates this code for p and another word, then one can compute back p and the second word. Using the same reasoning, from $p^{\times 2}01q^{\times 2}01w$ one can compute each of the three words p, q, and w.

Theorem 1 *There exists an additively optimal universal representation system.*

Proof Let φ_0 be the function defined from $(\{0,1\}^*)^2$ to $\{0,1\}^*$ by

$$\varphi_0 : \begin{cases} (\{0,1\}^*)^2 \to \{0,1\}^*\,, \\ (p \odot w, v) \mapsto T_p(w, v)\,, \end{cases}$$

where T_p is the function from $(\{0,1\}^*)^2$ to $\{0,1\}^*$ computed by the pth Turing machine (two tapes for the two inputs). From computability theory we know that this function is computable since there exists a universal Turing machine.

Let φ be a representation system. Since φ is computable, it is computed by a Turing machine whose number is p. For all words w and v, let y be a shortest program for w knowing v with representation system φ. Then

$$\varphi_0(p \odot y, v) = T_p(y, v) = \varphi(y, v) = w\,,$$

and so $p \odot y$ is a program (perhaps not the shortest) for w knowing v with representation system φ_0. Since

$$|p \odot y| = 2|p| + 2 + |y| = 2|p| + 2 + K_\varphi(w|v)\,,$$

it holds that, for all words w,

$$K_{\varphi_0}(w|v) \leqslant K_\varphi(w|v) + 2|p| + 2\,.$$

As p depends only on φ, we have that for all representation systems φ, there is a constant $c = 2|p| + 2$ such that

$$K_{\varphi_0}(w|v) < K_\varphi(w|v) + c\,.$$

We conclude that φ_0 is an additively optimal representation system. □

Now we are able to define Kolmogorov complexity.

Definition 3 Let φ_0 be an additively optimal representation system. The *Kolmogorov complexity of a word w knowing v* is defined by

$$K(w|v) = K_{\varphi_0}(w|v)\,,$$

and the *Kolmogorov complexity of a word w* is defined by

$$K(w) = K_{\varphi_0}(w) = K(w|\varepsilon)\,.$$

Remark 3 Kolmogorov complexity is defined up to an additive constant from Proposition 1. This means that any meaningful result must include in its statement something like "there is a constant c such that."

Properties

Basic Relations Kolmogorov complexity allows one to express relations that seem natural between complexities of words. For instance, the complexity of ww is close to the complexity of w. The complexity of uv is less than the sum of the complexities of u and v (in fact up to the logarithm of the size of u or v). They will be used in the sequel.

We give here a few examples of how to prove such results.

Theorem 2 *There exists a constant c such that for all words w and v,*

$$K(w|v) < |w| + c\,.$$

Proof Let π be the representation system defined by $\pi(w, x) = w$. As φ_0 is additively optimal, there is a constant c such that

$$K(w|v) < K_\pi(w|v) + c\,.$$

Since w is the only word such that $\pi(w) = w$, one has $K_\pi(w|v) = |w|$. We conclude that

$$K(w|v) < K_\pi(w|v) + c = |w| + c\,.$$

□

This result is interesting since it gives a computable upper bound for Kolmogorov complexity. Moreover, in Subsect. "Incompressible Words" we will see that this bound is tight.

Theorem 3 *Let f and g be partial computable functions. There exists a constant c such that for all words $w \in D_f$ and all $v \in D_g$*

$$K(f(w)|v) < K(w|g(v)) + c\,.$$

Proof As f and g are computable, function h, defined by $h(x, y) = f(\varphi_0(x, g(y)))$, is a representation system. Hence, by additive optimality of φ_0, there is a constant c such that for all w and v

$$K(f(w)|v) < K_h(f(w)|v) + c\,. \tag{2}$$

Let x be a shortest program for w knowing $g(v)$ with representation system φ_0. Since $\varphi_0(x, g(v)) = w$, $h(x, v) = f(\varphi_0(x, g(v)) = f(w)$, and so x is a program for $f(w)$ given v with representation system h. We conclude that

$$K_h(f(w)|v) \leqslant |y| = K(w|g(v))\,,$$

and hence, from Eq. (2),

$$K(f(w)|v) < K_h(f(w)|v) + c \leqslant K(w|g(v)) + c\,.$$

□

This proposition formalizes the intuition that if a word w can be computed from x, then w is simpler than x, of course up to some additive constant. This means that w contains less information than x. Of course, if f is injective, as we can go back from $f(w)$ to w, then w and $f(w)$ have the same amount of information. This is stated by the next corollary.

Corollary 1 *Let f and g be injective partial computable functions. There exists a constant c such that for all words w and v,*

$$|K(f(w)|v) - K(w|g(v))| < c\,.$$

Proof From the previous theorem, as f, g, f^{-1}, and g^{-1} are computable, there are constants c_1 and c_2 such that for all w and v

$$K(f(w)|v) < K(w|g(v)) + c_1 \text{ and}$$
$$K(f^{-1}(w)|v) < K(w|g^{-1}(v)) + c_2\,.$$

Hence, $K(w|g(v)) < K(f(w)|v) + c_2$. Choosing $c = \max(c_1, c_2)$ completes the proof. □

Using this corollary, one finds that the complexity of the representation by binary strings of a mathematical object depends only on the mathematical object and not on its representation, provided that an algorithm can compute one representation from another. Hence, for any mathematical object o, we denote by notation abuse $K(o)$ the complexity of a representation of o with alphabet $\{0, 1\}$. In particular, we use this for integers, finite sets, graphs, etc.

However, for objects that cannot be represented by finite binary strings (for instance, real numbers, infinite strings), we cannot define a Kolmogorov complexity. In this case, a per-case definition has to be given since for infinite objects several reasonable definitions of complexity can be found. For instance, for infinite strings one can use the limit superior or the limit inferior of the conditional complexity of the prefix given its length. One can use the limit superior of the complexity of the prefix divided by its length, etc.

Now we state a result about the complexity of a pair compared to the complexity of each member.

Theorem 4 *There is constant c such that, for any words x, y, and v,*

$$K(\langle x, y\rangle|v) < K(x|v) + K(y|\langle v, x\rangle) + 2\log_2(K(x|v)) + c\,,$$

where $\langle .,. \rangle$ *is any bijective computable function between* $\mathbb{N}^2$ *and* $\mathbb{N}$.

Proof Let $\widehat{x,y}$ be defined by $\widehat{x,y} = \ell(x)^{\times 2}01xy$, where $\ell(x)$ is the length of x written in binary. Note that x and y can both be reconstructed from $\widehat{x,y}$ and that

$$|\widehat{x,y}| = 2 + 2\lceil \log_2 |x| \rceil + |x| + |y| \,.$$

Let φ be the representation system defined by

$$\varphi(\widehat{x,y}, v) = \langle \varphi_0(x,v), \varphi_0(y, \langle v, \varphi_0(x,v)\rangle)\rangle \,.$$

As φ_0 is additively optimal, there is a constant c such that for all w and v

$$K(w|v) < K_\varphi(w|v) + c \,.$$

Let z be a shortest program for x knowing v, and let z' be a shortest program for y knowing $\langle v, x\rangle$. As $\widehat{z,z'}$ is a program for $\langle x, y\rangle$ knowing v with representation system φ, one has

$$\begin{aligned} &K(\langle x,y\rangle|v) \\ &< K_\varphi(w|v) + c \leqslant |\widehat{z,z'}| + c \\ &\leqslant K(x|v) + K(y|\langle v,x\rangle) + 2\log_2(K(x|v)) + 3 + c \,, \end{aligned}$$

which completes the proof. □

Of course, this inequality is not an equality since if $x = y$, the complexity of $K(\langle x,y\rangle)$ is the same as $K(x)$ up to some constant. When the equality holds, it is said that x and y have no mutual information.

Note that if we can represent the combination of programs z and z' in a shorter way, the upper bound is decreased by the same amount. For instance, if we choose $\widehat{x,y} = \ell(\ell(x))^{\times 2}01\ell(x)xy$, it becomes

$$2\log_2\log_2|x| + \log_2|x| + |x| + |y| \,.$$

If we know that a program never contains a special word u, then with $\widehat{x,y} = xuy$ it becomes $|x| + |y|$.

Examples The properties seen so far allow one to prove most of the relations between complexities of words.

For instance, choosing $f\colon w \mapsto ww$ and g being the identity in Corollary 1, we obtain that there is a constant c such that $|K(ww|v) - K(w|v)| < c$.

In the same way, letting f be the identity and g be defined by $g(x) = \varepsilon$ in Theorem 3, we get that there is a constant c such that $K(w|v) < K(w) + c$.

By Theorem 4, Corollary 1, and choosing f as $\langle x,y\rangle \mapsto xy$, and g as the identity, we have that there is a constant c such that

$$K(xy) < K(x) + K(y) + 2\min\log_2(K(x)), \log_2(K(y))) + c.$$

Incompressible Words Kolmogorov complexity gives a computer-science notion of randomness. By Theorem 2, we have an upper bound on Kolmogorov complexity: the length of the word. When this upper bound is reached, we say that the word is random. However, Kolmogorov complexity is difficult to use since it is not computable (Theorem 2.3.2 in [13]). The good news is that it is approximable from above.

The following definition formalizes the notion of incompressible word.

Definition 4 Let $c \in \mathbb{R}^+$. Word w is *c-incompressible* if

$$K(w) \geqslant |w| - c \,.$$

The next result proves the existence of incompressible words.

Proposition 2 *For any* $c \in \mathbb{R}^+$, *there are at least* $(2^{n+1} - 2^{n-c})$ *c-incompressible words w such that* $|w| \leqslant n$.

Proof Each program produces only one word; therefore there cannot be more words whose complexity is below n than the number of programs of size n. Hence, one finds

$$|\{w,\ K(w) < n\}| \leqslant 2^n - 1 \,.$$

This implies that

$$\begin{aligned} &|\{w,\ K(w) < |w| - c \text{ and } |w| \leqslant n\}| \\ &\qquad\qquad \leqslant |\{w,\ K(w) < n\}| < 2^{n-c} \,, \end{aligned}$$

and, since

$$\begin{aligned} &\{w,\ K(w) \geqslant |w| - c \text{ and } |w| \leqslant n\} \\ &\subset \{w,\ |w| \leqslant n\} \backslash \{w,\ K(w) < |w| - c \text{ and } |w| \leqslant n\} \,, \end{aligned}$$

it holds that

$$|\{w,\ K(w) \geqslant |w| - c \text{ and } |w| \leqslant n\}| > 2^{n+1} - 2^{n-c} \,.$$

□

From this proposition one deduces that half of the words whose size is less than or equal to n reach the upper bound of their Kolmogorov complexity. This incompressibility method relies on the fact that most words are incompressible.

Martin–Löf Tests

Martin–Löf tests give another equivalent definition of random sequences. The idea is simple: a sequence is random if it does not pass any computable test of singularity (i. e.,

a test that selects words from a "negligible" recursively enumerable set). As for Kolmogorov complexity, this notion needs a proper definition and implies a universal test. However, in this review we prefer to express tests in terms of Kolmogorov complexity. (We refer the reader to [13] for more on Martin–Löf tests.)

Let us start with an example. Consider the test "to have the same number of 0s and 1s." Define the set

$$E = \{w,\ w \text{ has the same number of 0s and 1s}\}$$

and order it in military order (length first, then lexicographic order), $E = \{e_0, e_1, \dots\}$. Consider the computable function $f\colon x \mapsto e_x$ (which is in fact computable for any decidable set E). Note that there are $\binom{2n}{n}$ words in E of length $2n$. Thus, if e_x has length $2n$, x is less than $\binom{2n}{n}$, whose logarithm is $2n - \log_2 \sqrt{n} + O(1)$. Then, using Theorem 3 with function f, one finds that

$$K(e_x) < K(x) < \log_2 x < |e_x| - \log_2 \sqrt{\frac{|e_x|}{2}}\ ,$$

up to an additive constant. We conclude that all members e_x of E are not c-incompressible whenever they are long enough.

This notion of test corresponds to "belonging to a small set" in Kolmogorov complexity terms. The next proposition formalizes these ideas.

Proposition 3 *Let E be a recursively enumerable set such that $|E \cap \{0,1\}^n| = o(2^n)$. Then, for all constants c, there is an integer M such that all words of length greater than M of E are not c-incompressible.*

Proof In this proof, we represent integers as binary strings. Let $\widehat{x,y}$ be defined, for integers x and y, by

$$\widehat{x,y} = |x|^{\times 2} 01 x y\ .$$

Let f be the computable function

$$f\colon \begin{cases} \{0,1\}^* & \to \quad \{0,1\}^*\ , \\ \widehat{x,y} & \mapsto \quad \text{the } y\text{th word of length } x \\ & \qquad \text{in the enumeration of } E\ . \end{cases}$$

From Theorems 2 and 3, there is a constant d such that for all words u of E

$$K(e) < |f^{-1}(e)| + d\ .$$

Note $u_n = |E \cap \{0,1\}^n|$. From the definition of function f one has

$$|f^{-1}(e)| \leqslant 3 + 2\log_2(|e|) + \log_2(u_{|e|})\ .$$

As $u_n = o(2^n)$, $\log_2(u_{|e|}) - |e|$ tends to $-\infty$, so there is an integer M such that when $|e| \geqslant M$, $K(e) < |f^{-1}(e)| + d < |e| - c$. This proves that no members of E whose length is greater than M are c-incompressible. □

Dynamical Systems and Symbolic Factors

A (discrete-time) dynamical system is a structure $\langle X, f\rangle$ where X is the set of *states* of the system and f is a function from X to itself that, given a state, tells which is the next state of the system. In the literature, the state space is usually assumed to be a compact metric space and f is a continuous function.

The study of the asymptotic properties of a dynamical system is, in general, difficult. Therefore, it can be interesting to associate the studied system with a simpler one and deduce the properties of the original system from the simple one. Indeed, under the compactness assumption, one can associate each system $\langle X, f\rangle$ with its *symbolic factor* as follows.

Consider a finite open covering $\{\beta_0, \beta_1, \dots, \beta_k\}$ of X. Label each set β_i with a symbol α_i. For any orbit of initial condition $x \in X$, we build an infinite word w_x on $\{\alpha_0, \alpha_1, \dots, \alpha_k\}^*$ such that for all $i \in \mathbb{N}$, $w_x(i) = \alpha_j$ if $f^i(x) \in \beta_j$ for some $j \in \{0, 1, \dots, k\}$. If $f^i(x)$ belongs to $\beta_j \cap \beta_h$ (for $j \neq h$), then arbitrarily choose either α_j or α_h. Denote by S_x the set of infinite words associated with the initial condition $x \in X$ and set $S = \bigcup_{x\in X} S_x$. The system $\langle S, \sigma\rangle$ is the symbolic system associated with $\langle X, f\rangle$; σ is the *shift map* defined as $\forall x \in S \forall i \in \mathbb{N},\ \sigma(x)_i = x_{i+1}$.

When $\{\beta_0, \beta_1, \dots, \beta_k\}$ is a clopen partition, then $\langle S, \sigma\rangle$ is a factor of $\langle X, f\rangle$ (see [12] for instance).

Dynamical systems theory has made great strides in recent decades, and a huge quantity of new systems have appeared. As a consequence, scientists have tried to classify them according to different criteria. From interactions among physicists was born the idea that in order to simulate a certain physical phenomenon, one should use a dynamical system whose complexity is not higher than that of the phenomenon under study. The problem is the meaning of the word "complexity"; in general it has a different meaning for each scientist. From a computer-science point of view, if we look at the state space of factor systems, they can be seen as sets of bi-infinite words on a fixed alphabet. Hence, each factor $\langle S, \sigma\rangle$ can be associated with a language as follows. For each pair of words $u\ w$ write $u \prec v$ if u is a factor of w. Given a finite word u and a bi-infinite word v, with an abuse of notation we write $u \prec v$ if u occurs as a factor in v. The language $L(v)$ associated with a bi-infinite word $v \in A^*$ is defined as

$$L(v) = \{u \in A^*, u \prec v\}\ .$$

Finally, the language $L(S)$ associated with the symbolic factor $\langle S, \sigma\rangle$ is given by

$$L(S) = \{u \in A^*, u \prec v\}\ .$$

The idea is that the complexity of the system $\langle X, f\rangle$ is proportional to the language complexity of its symbolic factor $\langle S, \sigma\rangle$ (see, for example, ► Topological Dynamics of Cellular Automata).

In [4,5], Brudno proposes to evaluate the complexity of symbolic factors using Kolmogorov complexity. Indeed, the complexity of the orbit of initial condition $x \in X$ according to finite open covering $\{\beta_0, \beta_1, \ldots, \beta_k\}$ is defined as

$$K(x, f, \{\beta_0, \beta_1, \ldots, \beta_k\}) = \limsup_{n\to\infty} \min_{w\in S_x} \frac{K(w_{0:n})}{n} .$$

Finally, the complexity of the orbit x is given by

$$K(x, f) = \sup K(x, f, \beta) ,$$

where the supremum is taken w.r.t. all possible finite open coverings β of X. Brudno has proven the following result.

Theorem 5 ([4]) *Consider a dynamical system $\langle X, f\rangle$ and an ergodic measure μ. For μ-almost all $x \in X$, $K(x, f) = H_\mu(f)$, where $H_\mu(f)$ is the measure entropy of $\langle X, f\rangle$ (for more on the measure entropy see, for example, ► Ergodic Theory of Cellular Automata).*

Cellular Automata

Cellular automata (CA) are (discrete-time) dynamical systems that have received increased attention in the last few decades as formal models of complex systems based on local interaction rules. This is mainly due to their great variety of distinct dynamical behavior and to the possibility of easily performing large-scale computer simulations. In this section we quickly review the definitions and useful results, which are necessary in the sequel.

Consider the set of configurations C, which consists of all functions from $\mathbb{Z}^D$ into A. The space C is usually equipped with the Cantor metric d_C defined as

$$\forall a, b \in C, \quad d_C(a, b) = 2^{-n} , \quad \text{with } n = \min_{\vec{v}\in\mathbb{Z}^D} \{\|\vec{v}\|_\infty : a(\vec{v}) \neq b(\vec{v})\} , \tag{3}$$

where $\|\vec{v}\|_\infty$ denotes the maximum of the absolute value of the components of $\vec{v}$. The topology induced by d_C coincides with the product topology induced by the discrete topology on A. With this topology, C is a compact, perfect, and totally disconnected space.

Let $N = \{\vec{u}_1, \ldots, \vec{u}_s\}$ be an ordered set of vectors of $\mathbb{Z}^D$ and $f\colon A^s \mapsto A$ be a function.

Definition 5 (CA) The D-dimensional CA based on the local rule δ and the neighborhood frame N is the pair $\langle C, f\rangle$, where $f\colon C \mapsto C$ is the global transition rule defined as follows:

$$\forall c \in C, \quad \forall \vec{v} \in \mathbb{Z}^D , \quad f(c)(\vec{v}) = \delta\left(c(\vec{v} + \vec{u}_1), \ldots, c(\vec{v} + \vec{u}_s)\right) . \tag{4}$$

Note that the mapping f is (uniformly) continuous with respect to d_C. Hence, the pair $\langle C, f\rangle$ is a proper (discrete-time) dynamical system.

In [7], a new metric on the phase space is introduced to better match the intuitive idea of chaotic behavior with its mathematical formulation. More results in this research direction can be found in [1,2,3,15]. This volume dedicates a whole chapter to the subject (► Dynamics of Cellular Automata in Non-compact Spaces). Here we simply recall the definition of the Besicovitch distance since it will be used in Subsect. "Example 2".

Consider the Hamming distance between two words $u\,v$ on the same alphabet A, $\#(u, v) = |\{i \in \mathbb{N} \mid u_i \neq v_i\}|$. This distance can be easily extended to work on words that are factors of bi-infinite words as follows:

$$\#_{h,k}(u, v) = |\{i \in [h, k] \mid u_i \neq v_i\}| ,$$

where $h, k \in \mathbb{Z}$ and $h < k$. Finally, the *Besicovitch pseudodistance* is defined for any pair of bi-infinite words $u\,v$ as

$$d_{\mathrm{B}}(u, v) = \limsup_{n\to\infty} \frac{\#_{-n,n}(u, v)}{2n + 1} .$$

The pseudodistance d_{B} can be turned into a distance when taking its restriction to $A^{\mathbb{Z}}|_{\doteq}$, where $\doteq$ is the relation of "being at null d_{B} distance." Roughly speaking, d_{B} measures the upper density of differences between two bi-infinite words.

The *space-time diagram* Γ is a graphical representation of a CA orbit. Formally, for a D-dimensional CA f with state set A, Γ is a function from $A^{\mathbb{Z}^D} \times \mathbb{N} \times \mathbb{Z}$ to A defined as $\Gamma(x, i) = f^i(x)_j$, for all $x \in A^{\mathbb{Z}^D}$, $i \in \mathbb{N}$ and $j \in \mathbb{Z}$.

The *limit set* Ω_f contains the long-term behavior of a CA on a given set U and is defined as follows:

$$\Omega_f(U) = \bigcap_{n\in\mathbb{N}} f^n(U) .$$

Unfortunately, any nontrivial property on the CA limit set is undecidable [11]. Other interesting information on a CA's long-term behavior is given by the *orbit limit set* Δ defined as follows:

$$\Delta_f(U) = \bigcup_{u\in U} \mathcal{O}'_f(u) ,$$

where H' is the set of adherence points of H. Note that in general the limit set and the orbit limit sets give different information. For example, consider a *rich* configuration c i. e., a configuration containing all possible finite patterns (here we adopt the terminology of [6]) and the shift map σ. Then, $\Delta_\sigma(\{c\}) = A^{\mathbb{Z}}$, while $\Omega_\sigma(\{c\})$ is countable.

Algorithmic Complexity as a Demonstration Tool

The incompressibility method is a valuable formal tool to decrease the combinatorial complexity of problems. It is essentially based on the following ideas (see also Chap. 6 in [13]):

- Incompressible words cannot be produced by short programs. Hence, if one has an incompressible (infinite) word, it cannot be algorithmically obtained.
- Most words are incompressible. Hence, a word taken at random can usually be considered to be incompressible without loss of generality.
- If one "proves" a recursive property on incompressible words, then, by Proposition 3, we have a contradiction.

Application to Cellular Automata

A portion of a space-time diagram of a CA can be computed given its local rule and the initial condition. The part of the diagram that depends only upon it has at most the complexity of this portion (Fig. 1). This fact often implies great computational dependencies between the initial part and the final part of the portion of the diagram: if the final part has high complexity, then the initial part must be at least as complex. Using this basic idea, proofs are structured as follows: assume that the final part has high complexity; use the hypothesis to prove that the initial part is not complex; then we have a contradiction. This technique provides a faster and clearer way to prove results that could be obtained by technical combinatorial arguments.

The first example illustrates this fact by rewriting a combinatorial proof in terms of Kolmogorov complexity. The second one is a result that was directly written in terms of Kolmogorov complexity.

Example 1

Consider the following result about languages recognizable by CA.

Proposition 4 ([17]) *The language $L = \{uvu, u, v \in \{0,1\}^*, |u| > 1\}$ is not recognizable by a real-time one-way CA.*

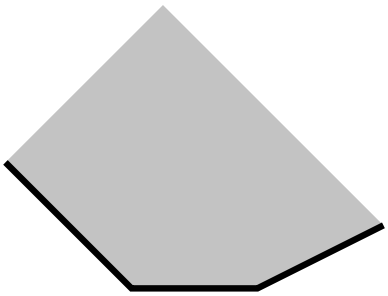

Algorithmic Complexity and Cellular Automata, Figure 1
States in the *gray zone* can be computed for the states of the *black line*: K(gray zone) $\leqslant K$(black line) up to an additive constant

Before giving the proof in terms of Kolmogorov complexity, we must recall some concepts. A language L is accepted in real-time by a CA if is accepted after $|x|$ transitions. An input word is accepted by a CA if after accepting state. A CA is *one-way* if the neighborhood is $\{0, 1\}$ (i. e., the central cell and the one on its right).

Proof One-way CA have some interesting properties. First, from a one-way CA recognizing L, a computer can build a one-way CA recognizing $L_\Sigma = \{uvu,\ u, v \in \Sigma^*, |u| > 1\}$, where Σ is any finite alphabet. The idea is to code the ith letter of Σ by $1u_0 1u_1 \cdots 1u_k 00$, where $u_0 u_1 \cdots u_k$ is i written in binary, and to apply the former CA.

Second, for any one-way CA of global rule f we have that

$$f^{|w|}(xwy) = f^{|w|}(xw) f^{|w|}(wy)$$

for all words x, y, and w.

Now, assume that a one-way CA recognizes L in real time. Then, there is an algorithm F that, for any integer n, computes a local rule for a one-way CA that recognizes $L_{\{0,\ldots,n-1\}}$ in real time.

Fix an integer n and choose a 0-incompressible word w of length $n(n-1)$. Let A be the subset of Z of pairs $(x,\ y)$ for $x, y \in \{0, \ldots, n-1\}$ and let $x \neq y$ be defined by

$$A = \{(x, y), \text{the } nx + y\text{th bit of } w \text{ is } 1\}\ .$$

Since A can be computed from w and vice versa, one finds that $K(A) = K(w) = |w| = n(n-1)$ up to an additive constant that does not depend on n.

Order the set $A = \{(x_0, y_0), (x_1, y_1), \ldots, \}$ and build a new word u as follows:

$$u_0 = y_0 x_0\ ,$$
$$u_{i+1} = u_i y_{i+1} x_{i+1} u_i\ ,$$
$$u = u_{|A|-1}\ ,$$

From Lemma 1 in [17], one finds that for all $x, y \in \{0, \ldots, n-1\}$, $(x, y) \in A \Leftrightarrow xuy \in L$. Let δ be the local rule produced by $F(n)$, Q the set of final states, and f the associated global rule. Since

$$f^{|u|}(xuy) = f^{|u|}(xu)f^{|u|}(uy),$$

one has that

$$(x, y) \in A \Leftrightarrow xuy \in L \Leftrightarrow \delta(f^{|u|}(xu), f^{|u|}(uy)) \in Q\,.$$

Hence, from the knowledge of each $f^{|u|}(xu)$ and $f^{|u|}(ux)$ for $x \in \{0, \ldots, n-1\}$, a list of $2n$ integers of $\{0, \ldots, n-1\}$, one can compute A. We conclude that

$$K(A) \leqslant 2n \log_2(n)$$

up to an additive constant that does not depend on n. This contradicts $K(A) = n(n-1)$ for large enough n. □

Example 2

Notation Given $x \in A^{\mathbb{Z}^D}$, we denote by $x_{\to n}$ the word $w \in S^{(2n+1)^D}$ obtained by taking all the states x_i for $i \in (2n+1)^D$ in the martial order.

The next example shows a proof that directly uses the incompressibility method.

Theorem 6 *In the Besicovitch topological space there is no transitive CA.*

Recall that a CA f is *transitive* if for all nonempty sets A B there exists $n \in \mathbb{N}$ such that $f^n(A) \cap B \neq \emptyset$.

Proof By contradiction, assume that there exists a transitive CA f of radius r with $C = |S|$ states. Let x and y be two configurations such that

$$\forall n \in \mathbb{N},\ K(x_{\to n}|y_{\to n}) \geq \frac{n}{2}\,.$$

A simple counting argument proves that such configurations x and y always exist. Since f is transitive, there are two configurations x' and y' such that for all $n \in \mathbb{N}$

$$\begin{aligned} \Delta(x_{\to n}, x'_{\to n}) &\leq 4\varepsilon n\,, \\ \Delta(y_{\to n}, y'_{\to n}) &\leq 4\delta n\,, \end{aligned} \tag{5}$$

and an integer u (which only depends on ε and δ) such that

$$f^u(y') = x'\,, \tag{6}$$

where $\varepsilon = \delta = (4e^{10\log_2 C})^{-1}$.

In what follows only n varies, while C, u, x, y, x', y', δ, and ε are fixed and independent of n.

By Eq. (6), one may compute the word $x'_{\to n}$ from the following items:

- $y'_{\to n}, f, u$ and n;
- The two words of y' of length ur that surround $y'_{\to n}$ and that are missing to compute $x'_{\to n}$ with Eq. (6).

We obtain that

$$K(x'_{\to n}|y'_{\to n}) \leq 2ur + K(u) + K(n) + K(f) + O(1) \leq o(n) \tag{7}$$

(the notations O and o are defined with respect to n). Note that n is fixed and hence $K(n)$ is a constant bounded by $\log_2 n$. Similarly, r and S are fixed, and hence $K(f)$ is constant and bounded by $C^{2r+1}\log_2 C + O(1)$.

Let us evaluate $K(y'_{\to n}|y_{\to n})$. Let $a_1, a_2, a_3, \ldots, a_k$ be the positive positions that $y_{\to n}$ and $y'_{\to n}$ differ at, sorted in increasing order. Let $b_1 = a_1$ and $b_i = a_i - a_{i-1}$, for $2 \leq i \leq k$. By Eq. (5) we know that $k \leq 4\delta n$. Note that $\sum_{i=1}^{k} b_i = a_k \leq n$.

Symmetrically let $a'_1, a'_2, a'_3, \ldots, a'_{k'}$ be the absolute values of the strictly negative positions that $y_{\to n}$ and $y'_{\to n}$ differ at, sorted in increasing order. Let $b'_1 = a'_1$ and $b'_i = a'_i - a'_{i-1}$, where $2 \leq i \leq k'$. Equation (5) states that $k' \leq 4\delta n$.

Since the logarithm is a concave function, one has

$$\sum \frac{\ln b_i}{k} \leq \ln \frac{\sum b_i}{k} \leq \ln \frac{n}{k}\,,$$

and hence

$$\sum \ln b_i \leq k \ln \frac{n}{k}\,, \tag{8}$$

which also holds for b'_i and k'.

Knowledge of the b_i, b'_i, and $k + k'$ states of the cells of $y'_{\to n}$, where $y_{\to n}$ differs from $y'_{\to n}$, is enough to compute $y'_{\to n}$ from $y_{\to n}$. Hence,

$$\begin{aligned} K(y'_{\to n}|y_{\to n}) \leq &\sum \ln(b_i) \\ &+ \sum \ln(b'_i) + (k + k')\log_2 C + O(1)\,. \end{aligned}$$

Equation (8) states that

$$K(y'_{\to n}|y_{\to n}) \leq k \ln \frac{n}{k} + k' \ln \frac{n}{k'} + (k + k')\log_2 C + O(1).$$

The function $k \mapsto k \ln \frac{n}{k}$ is increasing on $[0, \frac{n}{e}]$. As $k \leq 4\delta n \leq \frac{n}{e^{10\log_2 C}}$, we have that

$$k \ln \frac{n}{k} \leq 4\delta n \ln \frac{n}{4\delta n} \leq \frac{n}{e^{10\log_2 C}} \ln e^{10\log_2 C} \leq \frac{10n}{e^{10\log_2 C}}$$

and that

$$(k + k')\log_2 C \leq \frac{2n\log_2 C}{e^{10\log_2 C}}\,.$$

Replacing a, b, and k by a', b', and k', the same sequence of inequalities leads to a similar result. One deduces that

$$K(y'_{\to n}|y_{\to n}) \leq \frac{(2\log_2 C + 20)n}{e^{10\log_2 C}} + O(1)\,. \tag{9}$$

Similarly, Eq. (9) is also true with $K(x_{\to n}|x'_{\to n})$.

The triangular mequality for Kolmogorov complexity $K(a|b) \leq K(a|c) + K(b|c)$ (consequence of theorems 3 and 4) gives:

$$K(x_{\to n}|y_{\to n}) \leq K(x_{\to n}|x'_{\to n}) + K(x'_{\to n}|y'_{\to n}) + K(y'_{\to n}|y_{\to n}) + O(1)\,.$$

Equations (9) and (7) allow one to conclude that

$$K(x_{\to n}|y_{\to n}) \leq \frac{(2\log_2 C + 20)n}{e^{10\log_2 C}} + o(n)\,.$$

The hypothesis on x and y was $K(x_{\to n}|y_{\to n}) \geq \frac{n}{2}$. This implies that

$$\frac{n}{2} \leq \frac{(2C + 20)n}{e^{10\log_2 C}} + o(n)\,.$$

The last inequality is false for big enough n. □

Measuring CA Structural Complexity

Another use of Kolmogorov complexity in the study of CA is to understand what maximum complexity they can produce extracting examples of CA that show high complexity characteristics. The question is to define what is the meaning of "show high complexity characteristics" and, more precisely, what characteristic to consider. This section is devoted to structural characteristics of CA, that is to say, complexity that can be observed through static particularities.

The Case of Tilings

In this section, we give the original example that was given for tilings, often considered as a static version of CA.

In [10], Durand et al. construct a tile set whose tilings have maximum complexity. This paper contains two main results. The first one is an upper bound for the complexity of tilings, and the second one is an example of tiling that reaches this bound. First we recall the definitions about tilings of a plane by Wang tiles.

Definition 6 (Tilings with Wang tiles) Let C be a finite set of *colors*. A *Wang tile* is a quadruplet (n, s, e, w) of four colors from C corresponding to a square tile whose top color is n, left color is w, right color is e, and bottom color is s. A Wang tile cannot be rotated but can be copied to infinity.

Given a set of Wang tiles T, we say that a plane can be tiled by T if one can place tiles from T on the square grid $\mathbb{Z}^2$ such that the adjacent border of neighboring tiles has the same color. A set of tiles T that can tile the plane is called a *palette*.

The notion of *local constraint* gives a point of view closer to CA than tilings. Roughly speaking, it gives the local constraints that a tiling of the plane using 0 and 1 must satisfy. Note that this notion can be defined on any alphabet, but we can equivalently code any letter with 0 and 1.

Definition 7 (Tilings by local constraints) Let r be a positive integer called *radius*. Let C be a set of square patterns of size $2r + 1$ made of 0 and 1 (formally a function from $[\![-r, r]\!]$ to $\{0, 1\}$). The set is said to tile the plane if there is a way to put zeros and ones on the 2-D grid (formally a function from $\mathbb{Z}^2$ to $\{0, 1\}$) whose patterns of size $2r + 1$ are all in C. The possible layouts of zeros and ones on the (2-D) grid are called the *tilings acceptable for C*.

Seminal papers on this subject used Wang tiles. We translate these results in terms of local constraints in order to more smoothly apply them to CA.

Theorem 7 proves that, among tilings acceptable by a local constraint, there is always one that is not too complex. Note that this is a good notion since the use of a constraint of radius 1, $\{\boxed{0}, \boxed{1}\}$, which allows for all possible patterns, provides for acceptable high-complexity tilings since all tilings are acceptable.

Theorem 7 ([10]) *Let C be a local constraint. There is tiling τ acceptable for C such that the Kolmogorov complexity of the central pattern of size n is $O(n)$ (recall that the maximal complexity for a square pattern $n \times n$ is $O(n^2)$).*

Proof The idea is simple: if one knows the bits present in a border of width r of a pattern of size n, there are finitely many possibilities to fill the interior, so the first one in any computable order (for instance, lexicographically when putting all horizontal lines one after the other) has at most the complexity of the border since an algorithm can enumerate all possible fillings and take the first one in the chosen order.

Then, if one knows the bits in borders of width r of all central square patterns of size 2^n for all positive integers n (Fig. 2), one can recursively compute for each gap the first possible filling for a given computable order. The tiling obtained in this way (this actually defines a tiling since all cells are eventually assigned a value) has the required complexity: in order to compute the central pattern of size n,

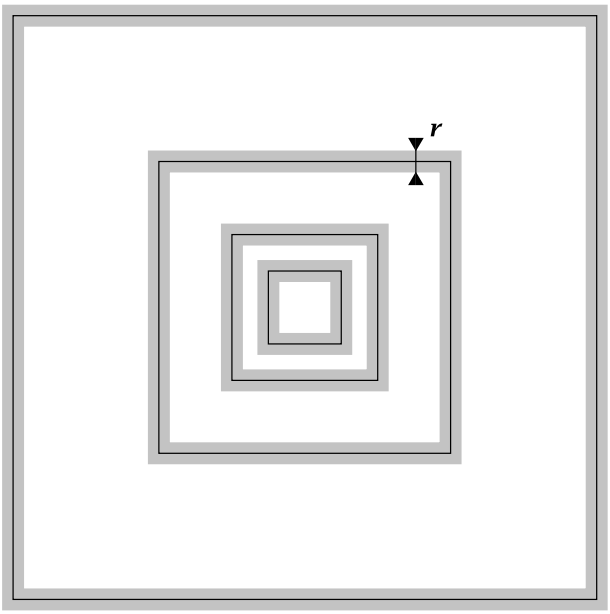

Algorithmic Complexity and Cellular Automata, Figure 2
Nested squares of size 2^k and border width r

the algorithm simply needs all the borders of size 2^k for $k \leqslant 2n$, which is of length at most $O(n)$. □

The next result proves that this bound is almost reachable.

Theorem 8 *Let r be any computable monotone and unbounded function. There exists a local constraint C_r such that for all tilings τ acceptable for C_r, the complexity of the central square pattern of τ is $O(n/r(n))$.*

The original statement does not have the O part. However, note that the simulation of Wang tiles by a local constraint uses a square of size $\ell = \lceil \sqrt{\log k} \rceil$ to simulate a single tile from a set of k Wang tiles; a square pattern of size n in the tiling corresponds to a square pattern of size $\frac{n}{\ell}$ in the original tiling.

Function r can grow very slowly (for instance, the inverse of the Ackermann function) provided it grows monotonously to infinity and is computable.

Proof The proof is rather technical, and we only give a sketch. The basic idea consists in taking a tiling constructed by Robinson [16] in order to prove that it is undecidable to test if a local constraint can tile the plane. This tiling is self-similar and is represented in Fig. 3. As it contains increasingly larger squares that occur periodically (note that the whole tiling is not periodic but only quasiperiodic), one can perform more and more computation steps within these squares (the periodicity is required to be sure that squares are present).

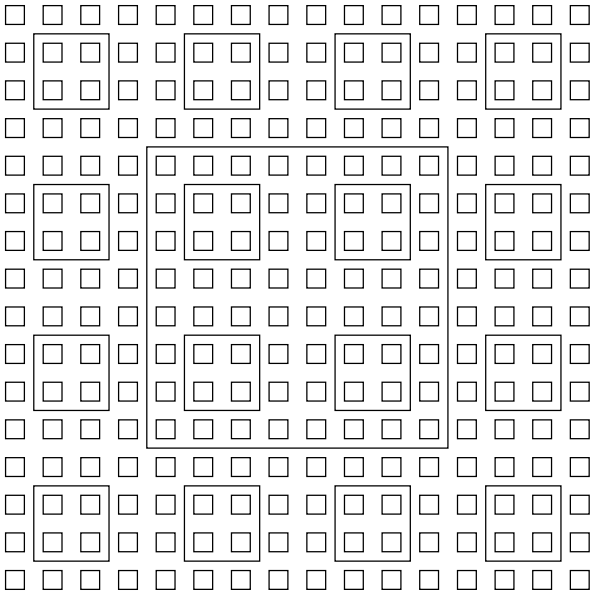

Algorithmic Complexity and Cellular Automata, Figure 3
Robinson's tiling

Using this tiling, Robinson can build a local constraint that simulates any Turing machine. Note that in the present case, the situation is more tricky than it seems since some technical features must be assured, like the fact that a square must deal with the smaller squares inside it or the constraint to add to make sure that smaller squares

have the same input as the bigger one. Using the constraint to forbid the occurrence of any final state, one gets that the compatible tilings will only simulate computations for inputs for which it does not halt.

To finish the proof, Durand et al. build a local constraint C that simulates a special Turing machine that halts on inputs whose complexity is small. Such a Turing machine exists since, though Kolmogorov complexity is not computable, testing all programs from ε to 1^n allows a Turing machine to compute all words whose complexity is below n and halt if it finds one. Then all tilings compatible with C contain in each square an input on which the Turing machine does not halt yet is of high complexity.

Function r occurs in the technical arguments since computable zones do not grow as fast as the length of squares. □

The Case of Cellular Automata

As we have seen so far, one of the results of [10] is that a tile set always produces tilings whose central square patterns of size n have a Kolmogorov complexity of $O(n)$ and not n^2 (which is the maximal complexity).

In the case of CA, something similar holds for space-time diagrams. Indeed, if one knows the initial row. One can compute the triangular part of the space-time diagram which depends on it (see Fig. 1). Then, as with tilings, the complexity of an $n \times n$-square is the same as the complexity of the first line, i. e., $O(n)$. However, unlike tilings, in CA, there is no restriction as to the initial configuration, hence CA have simple space-time diagrams.

Thus in this case Kolmogorov complexity is not of great help. One idea to improve the results could be to study how the complexity of configurations evolves during the application of the global rule. This aspect is particularly interesting with respect to dynamical properties. This is the subject of the next section.

Consider a CA F with radius r and local rule δ. Then, its orbit limit set cannot be empty. Indeed, let a be a state of f. Consider the configuration ${}^{\omega}a^{\omega}$. Let $s_a = \delta(a, \ldots, a)$. Then, $f({}^{\omega}a^{\omega}) = {}^{\omega}s_a^{\omega}$. Consider now a graph whose vertices are the states of the CA and the edges are (a, s_a). Since each vertex has an outgoing edge (actually exactly one), it must contain a cycle $a_0 \to a_1 \to \cdots \to a_k \to a_0$. Then each of the configurations ${}^{\omega}a_i^{\omega}$ for $0 \leqslant i \leqslant k$ is in the limit set since $f^k({}^{\omega}a_i^{\omega}) = {}^{\omega}a_i^{\omega}$.

This simple fact proves that any orbit limit set (and any limit set) of a CA must contain at least a monochromatic configuration whose complexity is low (for any reasonable definition). However, one can build a CA whose orbit limit set contains only complex configurations, except for the mandatory monochromatic one using the local-constraints technique discussed in the previous section.

Proposition 5 *There exists a CA whose orbit limit set contains only complex configurations.*

Proof Let r be any computable monotone and unbounded function and C_r the associated local constraint. Let A be the alphabet that C_r is defined on. Consider the 2-D CA f_r on the alphabet $A \cup \{\#\}$ (we assume $\{\#\} \neq A$). Let r be the radius of f_r (the same radius as C_r). Finally, the local rule δ of f_r is defined as follows:

$$\delta(P) = \begin{cases} P_0 & \text{if } P \in C_r\,, \\ \# & \text{otherwise}\,. \end{cases}$$

Using this local rule, one can verify the following fact. If the configuration c is not acceptable for C_r, then $(\mathcal{O}(c))' = \{{}^{\omega}\#^{\omega}\}$; $(\mathcal{O}(c))' = \{c\}$ otherwise. Indeed, if c is acceptable for C_r, then $f(c) = c$; otherwise there is a position i such that $c(i)$ is not a valid pattern for C_r. Then $f(c)(i) = \#$. By simple induction, this means that all cells that are at a distance less than kr from position i become $\#$ after k steps. Hence, for all $n > 0$, after $k \geqslant \frac{n+2|i|}{r}$ steps, the Cantor distance between $f^k(c)$ and ${}^{\omega}\#^{\omega}$ is less than 2^{-n}, i. e., $\mathcal{O}(c)$ tends to ${}^{\omega}\#^{\omega}$. □

Measuring the Complexity of CA Dynamics

The results of the previous section have limited range since they tell something about the quasicomplexity but nothing about the plain complexity of the limit set. To enhance our study we need to introduce some general concepts, namely, randomness spaces.

Randomness Spaces

Roughly speaking, a randomness space is a structure made of a topological space and a measure that helps in defining which points of the space are random. More formally we can give the following.

Definition 8 ([6]) A *randomness space* is a structure $\langle X, B, \mu\rangle$ where X is a topological space, $B\colon \mathbb{N} \to 2^X$ a total numbering of a subbase of X, and μ a Borel measure.

Given a numbering B for a subbase for an open set of X, one can produce a numbering B' for a base as follows:

$$B'(i) = \bigcap_{j \in D_{i+1}} B(j)\,,$$

where $D\colon \mathbb{N} \to \{E \mid E \subseteq \mathbb{N} \text{ and } E \text{ is finite}\}$ is the bijection defined by $D^{-1}(E) = \sum_{i \in E} 2^i$. B' is called the *base derived* from the subbase B. Given two sequences of open sets (V_n)

and (U_n) of X, we say that (V_n) is *U-computable* if there exists a recursively enumerable set $H \in \mathbb{N}$ such that

$$\forall n \in \mathbb{N}, V_n = \bigcup_{i \in \mathbb{N}, \langle n,i \rangle \in H} U_i ,$$

where $\langle i, j \rangle = (i + j)(i + j + 1) + j$ is the classical bijection between $\mathbb{N}^2$ and $\mathbb{N}$. Note that this bijection can be extended to $\mathbb{N}^D$ (for $D > 1$) as follows: $\langle x_1, x_2, \ldots, x_k \rangle = \langle x_1, \langle x_2, \ldots, x_k \rangle \rangle$.

Definition 9 ([6]) Given a randomness space $\langle X, B, \mu \rangle$, a *randomness test* on X is a B'-computable sequence (U_n) of open sets such that $\forall n \in \mathbb{N},\ \mu(U_n) \leq 2^{-n}$.

Given a randomness test (U_n), a point $x \in X$ is said to *pass the test* (U_n) if $x \in \cap_{n \in \mathbb{N}} U_n$. In other words, tests select points belonging to sets of null measure. The computability of the tests assures the computability of the selected null measure sets.

Definition 10 ([6]) Given a randomness space $\langle X, B, \mu \rangle$, a point $x \in X$ is *nonrandom* if x passes some randomness test. The point $x \in X$ is *random*k if it is not nonrandom.

Finally, note that for any $D \geq 1$, $\langle A^{\mathbb{Z}^D}, B, \mu \rangle$ is a randomness space when setting B as

$$B(j + D \cdot \langle i_1, \ldots, i_D \rangle) = \left\{ c \in A^{\mathbb{Z}^D},\ c_{i_1,\ldots,i_D} = a_j \right\} ,$$

where $A = \{a_1, \ldots, a_j, \ldots, a_D\}$ and μ is the classical product measure built from the uniform Bernoulli measure over A.

Theorem 9 ([6]) *Consider a D-dimensional CA f. Then, the following statements are equivalent:*

1. *f is surjective;*
2. *$\forall c \in A^{\mathbb{Z}^D}$, if c is rich (i. e., c contains all possible finite patterns), then $f(c)$ is rich;*
3. *$\forall c \in A^{\mathbb{Z}^D}$, if c is random, then $f(c)$ is random.*

Theorem 10 ([6]) *Consider a D-dimensional CA f. Then, $\forall c \in A^{\mathbb{Z}^D}$, if c is not rich, then $f(c)$ is not rich.*

Theorem 11 ([6]) *Consider a 1-D CA f. Then, $\forall c \in A^{\mathbb{Z}^D}$, if c is nonrandom, then $f(c)$ is nonrandom.*

Note that the result in Theorem 11 is proved only for 1-D CA; its generalization to higher dimensions is still an open problem.

Open problem 1 *Do D-dimensional CA for $D > 1$ preserve nonrandomness?*

From Theorem 9 we know that the property of preserving randomness (resp. richness) is related to surjectivity. Hence, randomness (resp. richness) preserving is decidable in one dimension and undecidable in higher dimensions. The opposite relations are still open.

Open problem 2 *Is nonrichness (resp. nonrandomness) a decidable property?*

Algorithmic Distance

In this section we review an approach to the study of CA dynamics from the point of view of algorithmic complexity that is completely different than the one reported in the previous section. For more details on the algorithmic distance see ► Chaotic Behavior of Cellular Automata.

In this new approach we are going to define a new distance using Kolmogorov complexity in such a way that two points x and y are near if it is "easy" to transform x into y or vice versa using a computer program. In this way, if a CA turns out to be sensitive to initial conditions, for example, then it means that it is able to create new information. Indeed, we will see that this is not the case.

Definition 11 The *algorithmic distance* between $x \in A^{\mathbb{Z}^D}$ and $y \in A^{\mathbb{Z}^D}$ is defined as follows:

$$d_{\mathrm{a}}(x, y) = \limsup_{n \to \infty} \frac{K(x \to n | y \to n) + K(y \to n | x \to n)}{2(2n + 1)^D} .$$

It is not difficult to see that d_{a} is only a pseudodistance since there are many points that are at null distance (those that differ only on a finite number of cells, for example). Consider the relation $\cong$ of "being at null d_{a} distance, i. e., $\forall x, y \in A^{\mathbb{Z}^D}$, $x \cong y$ if and only if $d_{\mathrm{a}}(x, y) = 0$. Then, $\langle A^{\mathbb{Z}^D}/_{\cong}, d_{\mathrm{a}} \rangle$ is a metric space.

Note that the definition of d_{a} does not depend on the chosen additively optimal universal description mode since the additive constant disappears when dividing by $2(2n + 1)^D$ and taking the superior limit. Moreover, by Theorem 2, the distance is bounded by 1.

The following results summarize the main properties of this new metric space.

Theorem 12 (**► Chaotic Behavior of Cellular Automata**) *The metric space $\langle A^{\mathbb{Z}^D}/_{\cong}, d_{\mathrm{a}} \rangle$ is perfect, pathwise connected, infinite dimensional, nonseparable, and noncompact.*

Theorem 12 says that the new topological space has enough interesting properties to make worthwhile the study of CA dynamics on it.

The first interesting result obtained by this approach concerns surjective CA. Recall that surjectivity plays a central role in the study of chaotic behavior since it is a nec-

essary condition for many other properties used to define deterministic chaos like expansivity, transitivity, ergodicity, and so on (see ▶ Topological Dynamics of Cellular Automata and [8] for more on this subject). The following result proves that in the new topology $A^{\mathbb{Z}^D}/_{\cong}$ the situation is completely different.

Proposition 6 (▶ **Chaotic Behavior of Cellular Automata**) *If f is a surjective CA, then $d_A(x, f(x)) = 0$ for any $x \in A^{\mathbb{Z}^D}/_{\cong}$. In other words, every CA behaves like the identity in $A^{\mathbb{Z}^D}/_{\cong}$.*

Proof In order to compute $f(x)_{\to n}$ from $x_{\to n}$, one need only know the index of f in the set of CA with radius r, state set S, and dimension d; therefore

$$K(f(x)_{\to n}|x_{\to n}) \leqslant 2dr(2n+2r+1)^{d-1}\log_2|S| + K(f) + 2\log_2 K(f) + c\,,$$

and similarly

$$K(x_{\to n}|f(x)_{\to n}) \leqslant gdr(2n+1)^{d-1}\log_2|S| + K(f) + 2\log_2 K(f) + c\,.$$

Dividing by $2(2n+1)^d$ and taking the superior limit one finds that $d_a(x, f(x)) = 0$. □

The result of Proposition 6 means that surjective CA can neither create new information nor destroy it. Hence, they have a high degree of stability from an algorithmic point of view. This contrasts with what happens in the Cantor topology. We conclude that the classical notion of deterministic chaos is orthogonal to "algorithmic chaos," at least in what concerns the class of CA.

Proposition 7 (▶ **Chaotic Behavior of Cellular Automata**) *Consider a CA f that is neither surjective nor constant. Then, there exist two configurations $x, y \in A^{\mathbb{Z}^D}/_{\cong}$ such that $d_a(x, y) = 0$ but $d_a(f(x), f(y)) \neq 0$.*

In other words, Proposition 7 says that nonsurjective, nonconstant CA are not compatible with $\cong$ and hence not continuous. This means that, for any pair of configurations $x\,y$, this kind of CA either destroys completely the information content of $x\,y$ or preserves it in one, say x, and destroys it in y. However, the following result says that some weak form of continuity still persists (see ▶ Topological Dynamics of Cellular Automata and [8] for the definitions of equicontinuity point and sensibility to initial conditions).

Proposition 8 (▶ **Chaotic Behavior of Cellular Automata**) *Consider a CA f and let $\underline{a}$ be the configuration made with all cells in state a. Then, $\underline{a}$ is both a fixed point and an equicontinuity point for f.*

Even if CA were noncontinuous on $A^{\mathbb{Z}^D}/_{\cong}$, one could still wonder what happens with respect to the usual properties used to define deterministic chaos. For instance, by Proposition 8, it is clear that no CA is sensitive to initial conditions. The following question is still open.

Open problem 3 *Is $\underline{a}$ the only equicontinuity point for CA on $A^{\mathbb{Z}^D}/_{\cong}$?*

Future Directions

In this paper we have illustrated how algorithmic complexity can help in the study of CA dynamics. We essentially used it as a powerful tool to decrease the combinatorial complexity of problems. These kinds of applications are only at their beginnings and much more are expected in the future. For example, in view of the results of Subsect. "Example 1", we wonder if Kolmogorov complexity can help in proving the famous conjecture that languages recognizable in real time by CA are a strict subclass of linear-time recognizable languages (see [9,14]).

Another completely different development would consist in finding how and if Theorem 11 extends to higher dimensions. How this property can be restated in the context of the algorithmic distance is also of great interest.

Finally, how to extend the results obtained for CA to other dynamical systems is a research direction that must be explored. We are rather confident that this can shed new light on the complexity behavior of such systems.

Acknowledgments

This work has been partially supported by the ANR Blanc Project "Sycomore".

Bibliography

Primary Literature

1. Blanchard F, Formenti E, Kurka P (1999) Cellular automata in the Cantor, Besicovitch and Weyl topological spaces. Complex Syst 11:107–123
2. Blanchard F, Cervelle J, Formenti E (2003) Periodicity and transitivity for cellular automata in Besicovitch topologies. In: Rovan B, Vojtas P (eds) (MFCS'2003), vol 2747. Springer, Bratislava, pp 228–238
3. Blanchard F, Cervelle J, Formenti E (2005) Some results about chaotic behavior of cellular automata. Theor Comput Sci 349(3):318–336
4. Brudno AA (1978) The complexity of the trajectories of a dynamical system. Russ Math Surv 33(1):197–198
5. Brudno AA (1983) Entropy and the complexity of the trajectories of a dynamical system. Trans Moscow Math Soc 44:127
6. Calude CS, Hertling P, Jürgensen H, Weihrauch K (2001) Randomness on full shift spaces. Chaos, Solitons Fractals 12(3):491–503

7. Cattaneo G, Formenti E, Margara L, Mazoyer J (1997) A shift-invariant metric on S^Z inducing a non-trivial topology. In: Privara I, Rusika P (eds) (MFCS'97), vol 1295. Springer, Bratislava, pp 179–188
8. Cervelle J, Durand B, Formenti E (2001) Algorithmic information theory and cellular automata dynamics. In: Mathematical Foundations of Computer Science (MFCS'01). Lectures Notes in Computer Science, vol 2136. Springer, Berlin pp 248–259
9. Delorme M, Mazoyer J (1999) Cellular automata as languages recognizers. In: Cellular automata: A parallel model. Kluwer, Dordrecht
10. Durand B, Levin L, Shen A (2001) Complex tilings. In: STOC '01: Proceedings of the 33rd annual ACM symposium on theory of computing, pp 732–739
11. Kari J (1994) Rice's theorem for the limit set of cellular automata. Theor Comput Sci 127(2):229–254
12. Kůrka P (1997) Languages, equicontinuity and attractors in cellular automata. Ergod Theory Dyn Syst 17:417–433
13. Li M, Vitányi P (1997) An introduction to Kolmogorov complexity and its applications, 2nd edn. Springer, Berlin
14. M Delorme EF, Mazoyer J (2000) Open problems. Research Report LIP 2000-25, Ecole Normale Supérieure de Lyon
15. Pivato M (2005) Cellular automata vs. quasisturmian systems. Ergod Theory Dyn Syst 25(5):1583–1632
16. Robinson RM (1971) Undecidability and nonperiodicity for tilings of the plane. Invent Math 12(3):177–209
17. Terrier V (1996) Language not recognizable in real time by one-way cellular automata. Theor Comput Sci 156:283–287
18. Wolfram S (2002) A new kind of science. Wolfram Media, Champaign. http://www.wolframscience.com/

Books and Reviews

Batterman RW, White HS (1996) Chaos and algorithmic complexity. Fund Phys 26(3):307–336
Bennet CH, Gács P, Li M, Vitányi P, Zurek W (1998) Information distance. EEE Trans Inf Theory 44(4):1407–1423
Calude CS (2002) Information and randomness. Texts in theoretical computer science, 2nd edn. Springer, Berlin
Cover TM, Thomas JA (2006) Elements of information theory, 2nd edn. Wiley, New York
White HS (1993) Algorithmic complexity of points in dynamical systems. Ergod Theory Dyn Syst 13:807–830

Amorphous Computing

Hal Abelson, Jacob Beal, Gerald Jay Sussman
Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, USA

Article Outline

Glossary
Definition of the Subject
Introduction
The Amorphous Computing Model
Programming Amorphous Systems
Amorphous Computing Paradigms
Primitives for Amorphous Computing
Means of Combination and Abstraction
Supporting Infrastructure and Services
Lessons for Engineering
Future Directions
Bibliography

Glossary

Amorphous computer A collection of computational particles dispersed irregularly on a surface or throughout a volume, where individual particles have no *a priori* knowledge of their positions or orientations.

Computational particle A (possibly faulty) individual device for an amorphous computer. Each particle has modest computing power and a modest amount of memory. The particles are not synchronized, although they are all capable of operating at similar speeds, since they are fabricated by the same process. All particles are programmed identically, although each particle has means for storing local state and for generating random numbers.

Field A function assigning a value to every particle in an amorphous computer.

Gradient A basic amorphous computing primitive that estimates the distance from each particle to the nearest particle designated as a source of the gradient.

Definition of the Subject

The goal of amorphous computing is to identify organizational principles and create programming technologies for obtaining intentional, pre-specified behavior from the cooperation of myriad unreliable parts that are arranged in unknown, irregular, and time-varying ways. The heightened relevance of amorphous computing today stems from the emergence of new technologies that could serve as substrates for information processing systems of immense power at unprecedentedly low cost, if only we could master the challenge of programming them.

Introduction

Even as the foundations of computer science were being laid, researchers could hardly help noticing the contrast between the robustness of natural organisms and the fragility of the new computing devices. As John von Neumann remarked in 1948 [53]:

> With our artificial automata we are moving much more in the dark than nature appears to be with

its organisms. We are, and apparently, at least at present, have to be much more 'scared' by the occurrence of an isolated error and by the malfunction which must be behind it. Our behavior is clearly that of overcaution, generated by ignorance.

Amorphous computing emerged as a field in the mid-1990s, from the convergence of three factors:

- Inspiration from the cellular automata models for fundamental physics [13,34].
- Hope that understanding the robustness of biological development could both help overcome the brittleness typical of computer systems and also illuminate the mechanisms of developmental biology.
- The prospect of nearly free computers in vast quantities.

Microfabrication

One technology that has come to fruition over the past decade is micro-mechanical electronic component manufacture, which integrates logic circuits, micro-sensors, actuators, and communication on a single chip. Aggregates of these can be manufactured extremely inexpensively, provided that not all the chips need work correctly, and that there is no need to arrange the chips into precise geometrical configurations or to establish precise interconnections among them. A decade ago, researchers envisioned *smart dust* elements small enough to be borne on air currents to form clouds of communicating sensor particles [26].

Airborne sensor clouds are still a dream, but networks of millimeter-scale particles are now commercially available for environmental monitoring applications [40]. With low enough manufacturing costs, we could mix such particles into bulk materials to form coatings like "smart paint" that can sense data and communicate its actions to the outside world. A smart paint coating on a wall could sense vibrations, monitor the premises for intruders, or cancel noise. Bridges or buildings coated with smart paint could report on traffic and wind loads and monitor structural integrity. If the particles have actuators, then the paint could even heal small cracks by shifting the material around. Making the particles mobile opens up entire new classes of applications that are beginning to be explored by research in *swarm robotics* [35] and *modular robotics* [49].

Cellular Engineering

The second disruptive technology that motivates the study of amorphous computing is microbiology. Biological organisms have served as motivating metaphors for computing since the days of calculating engines, but it is only over the past decade that we have begun to see how biology could literally be a substrate for computing, through the possibility of constructing digital-logic circuits within individual living cells. In one technology logic signals are represented not by electrical voltages and currents, but by concentrations of DNA binding proteins, and logic elements are realized as binding sites where proteins interact through promotion and repression. As a simple example, if A and B are proteins whose concentrations represent logic levels, then an "inverter" can be implemented in DNA as a genetic unit through which A serves as a repressor that blocks the production of B [29,54,55,61]. Since cells can reproduce themselves and obtain energy from their environment, the resulting information processing units could be manufactured in bulk at a very low cost.

There is beginning to take shape a technology of cellular engineering that can tailor-make programmable cells to function as sensors or delivery vehicles for pharmaceuticals, or even as chemical factories for the assembly of nanoscale structures. Researchers in this emerging field of *synthetic biology* are starting to assemble registries of standard logical components implemented in DNA that can be inserted into *E. coli* [48]. The components have been engineered to permit standard means of combining them, so that biological logic designers can assemble circuits in a mix-and-match way, similar to how electrical logic designers create circuits from standard TTL parts. There's even an *International Genetically Engineered Machine Competition* where student teams from universities around the world compete to create novel biological devices from parts in the registry [24].

Either of these technologies—microfabricated particles or engineered cells—provides a path to cheaply fabricate aggregates of massive numbers of computing elements. But harnessing these for computing is quite a different matter, because the aggregates are unstructured. Digital computers have always been constructed to behave as precise arrangements of reliable parts, and almost all techniques for organizing computations depend upon this precision and reliability. Amorphous computing seeks to discover new programming methods that do not require precise control over the interaction or arrangement of the individual computing elements and to instantiate these techniques in new programming languages.

The Amorphous Computing Model

Amorphous computing models the salient features of an unstructured aggregate through the notion of an *amor-*

phous computer, a collection of *computational particles* dispersed irregularly on a surface or throughout a volume, where individual particles have no *a priori* knowledge of their positions or orientations. The particles are possibly faulty, may contain sensors and effect actions, and in some applications might be mobile. Each particle has modest computing power and a modest amount of memory. The particles are not synchronized, although they are all capable of operating at similar speeds, since they are fabricated by the same process. All particles are programmed identically, although each particle has means for storing local state and for generating random numbers. There may also be several distinguished particles that have been initialized to particular states.

Each particle can communicate with a few nearby neighbors. In an electronic amorphous computer the particles might communicate via short-distance radio, whereas bioengineered cells might communicate by chemical signals. Although the details of the communication model can vary, the maximum distance over which two particles can communicate effectively is assumed to be small compared with the size of the entire amorphous computer. Communication is assumed to be unreliable and a sender has no assurance that a message has been received (Higher-level protocols with message acknowledgement can be built on such unreliable channels).

We assume that the number of particles may be very large (on the order of 10^6 to 10^{12}). Algorithms appropriate to run on an amorphous computer should be relatively independent of the number of particles: the performance should degrade gracefully as the number of particles decreases. Thus, the entire amorphous computer can be regarded as a massively parallel computing system, and previous investigations into massively parallel computing, such as research in cellular automata, are one source of ideas for dealing with amorphous computers. However, amorphous computing differs from investigations into cellular automata, because amorphous mechanisms must be independent of the detailed configuration, reliability, and synchronization of the particles.

Programming Amorphous Systems

A central theme in amorphous computing is the search for programming paradigms that work within the amorphous model. Here, biology has been a rich source of metaphors for inspiring new programming techniques. In embryonic development, even though the precise arrangements and numbers of the individual cells are highly variable, the genetic "programs" coded in DNA nevertheless produce well-defined intricate shapes and precise forms. Amorphous computers should be able to achieve similar results.

One technique for programming an amorphous computer uses diffusion. One particle (chosen by some symmetry-breaking process) broadcasts a message. This message is received by each of its neighbors, which propagate it to their neighbors, and so on, to create a wave that spreads throughout the system. The message contains a count, and each particle stores the received count and increments it before re-broadcasting. Once a particle has stored its count, it stops re-broadcasting and ignores future count messages. This *count-up wave* gives each particle a rough measure of its distance from the original source. One can also produce regions of controlled size, by having the count message relayed only if the count is below a designated bound.

Two such count-up waves can be combined to identify a chain of particles between two given particles *A* and *B*. Particle *A* begins by generating a count-up wave as above. This time, however, each intermediate particle, when it receives its count, performs a handshake to identify its "predecessor"—the particle from which it received the count (and whose own count will therefore be one less). When the wave of count messages reaches *B*, *B* sends a "successor" message, informing its predecessor that it should become part of the chain and should send a message to *its* predecessor, and so on, all the way back to *A*. Note that this method works, even though the particles are irregularly distributed, provided there is a path from *A* to *B*.

The motivating metaphor for these two programs is chemical gradient diffusion, which is a foundational mechanism in biological development [60]. In nature, biological mechanisms not only generate elaborate forms, they can also maintain forms and repair them. We can modify the above amorphous line-drawing program so that it produces self-repairing line: first, particles keep re-broadcasting their count and successor messages. Second, the status of a particle as having a count or being in the chain decays over time unless it is refreshed by new messages. That is, a particle that stops hearing successor messages intended for it will eventually revert to not being in the chain and will stop broadcasting its own successor messages. A particle that stops hearing its count being broadcast will start acting as if it never had a count, pick up a new count from the messages it hears, and start broadcasting the count messages with the new count. Clement and Nagpal [12] demonstrated that this mechanism can be used to generate self-repairing lines and other patterns, and even re-route lines and patterns around "dead" regions where particles have stopped functioning.

The relationship with biology flows in the other direction as well: the amorphous algorithm for repair is a model which is not obviously inconsistent with the facts of angiogenesis in the repair of wounds. Although the existence of the algorithm has no bearing on the facts of the matter, it may stimulate systems-level thinking about models in biological research. For example, Patel et al. use amorphous ideas to analyze the growth of epithelial cells [43].

Amorphous Computing Paradigms

Amorphous computing is still in its infancy. Most of linguistic investigations based on the amorphous computing model have been carried out in simulation. Nevertheless, this work has yielded a rich variety of programming paradigms that demonstrate that one can in fact achieve robustness in face of the unreliability of individual particles and the absence of precise organization among them.

Marker Propagation for Amorphous Particles

Weiss's Microbial Colony Language [55] is a marker propagation language for programming the particles in an amorphous computer. The program to be executed, which is the same for each particle, is constructed as a set of rules. The state of each particle includes a set of binary markers, and rules are enabled by boolean combinations of the markers. The rules, which have the form (*trigger*, *condition*, *action*) are triggered by the receipt of labelled messages from neighboring particles. A rule may test conditions, set or clear various markers, and it broadcast further messages to its neighbors. Each message carries a count that determines how far it will diffuse, and each marker has a lifetime that determines how long its value lasts. Supporting these language's rules is a runtime system that automatically propagates messages and manages the lifetimes of markers, so that the programmer need not deal with these operations explicitly.

Weiss's system is powerful, but the level of abstraction is very low. This is because it was motivated by cellular engineering—as something that can be directly implemented by genetic regulatory networks. The language is therefore more useful as a tool set in which to implement higher-level languages such as GPL (see below), serving as a demonstration that in principle, these higher-level languages can be implemented by genetic regulatory networks as well.

Figure 1 shows an example simulation programmed in this language, that organizes an initially undifferentiated column of particles into a structure with band of two alternating colors: a caricature of somites in developing vertebrae.

The Growing Point Language

Coore's *Growing Point Language* (GPL) [15] demonstrates that an amorphous computer can be configured by a program that is common to all the computing elements to generate highly complex patterns, such as the pattern representing the interconnection structure of an arbitrary electrical circuit as shown in Fig. 2.

GPL is inspired by a botanical metaphor based on growing points and tropisms. A growing point is a locus of activity in an amorphous computer. A growing point propagates through the computer by transferring its activity from one computing element to a neighbor. As a growing point passes through the computer it effects the differentiation of the behaviors of the particles it visits. Particles secrete "chemical" signals whose count-up waves define gradients, and these attract or repel grow-

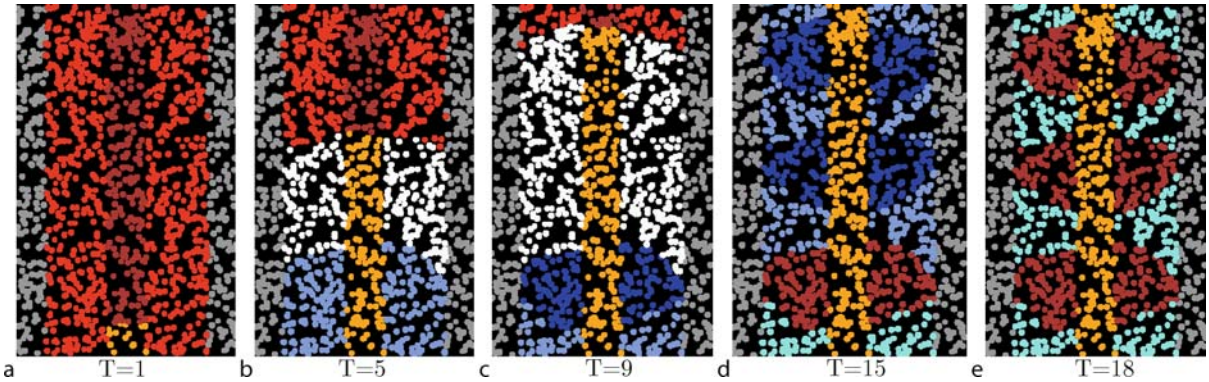

Amorphous Computing, Figure 1
A Microbial Colony Language program organizes a tube into a structure similar to that of somites in the developing vertebrate (from [55]).

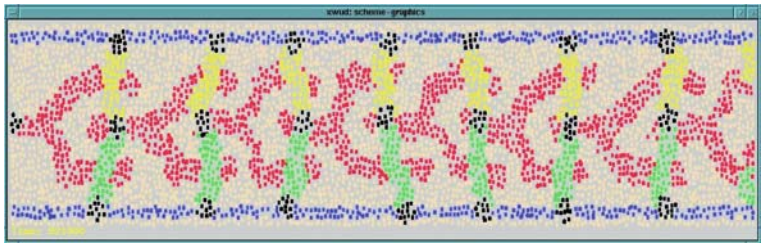

Amorphous Computing, Figure 2
A pattern generated by GPL whose shape mimics a chain of CMOS inverters (from [15])

ing points as directed by programmer-specific "tropisms". Coore demonstrated that these mechanisms are sufficient to permit amorphous computers to generate any arbitrary prespecified graph structure pattern, up to topology. Unlike real biology, however, once a pattern has been constructed, there is no clear mechanism to maintain it in the face of changes to the material. Also, from a programming linguistic point of view, there is no clear way to compose shapes by composing growing points. More recently, Gayle and Coore have shown how GPL may be extended to produce arbitrarily large patterns such as arbitrary text strings[21]. D'Hondt and D'Hondt have explored the use of GPL for geometrical constructions and its relations with computational geometry [18,19].

Origami-Based Self-Assembly

Nagpal [38] developed a prototype model for controlling programmable materials. She showed how to organize a program to direct an amorphous sheet of deformable particles to cooperate to construct a large family of globally-specified predetermined shapes. Her method, which is inspired by the folding of epithelial tissue, allows a programmer to specify a sequence of folds, where the set of available folds is sufficient to create any origami shape (as shown by Huzita's axioms for origami [23]). Figure 3 shows a sheet of amorphous particles, where particles can cooperate to create creases and folds, assembling itself into the well-known origami "cup" structure.

Nagpal showed how this language of folds can be compiled into a a low-level program that can be distributed to all of the particles of the amorphous sheet, similar to Coore's GPL or Weiss's MCL. With a few differences of initial state (for example, particles at the edges of the sheet know that they are edge particles) the particles run their copies of the program, interact with their neighbors, and fold up to make the predetermined shape. This technique is quite robust. Nagpal studied the range of shapes that can be constructed using her method, and on their sensitivity to errors of communication, random cell death, and density of the cells.

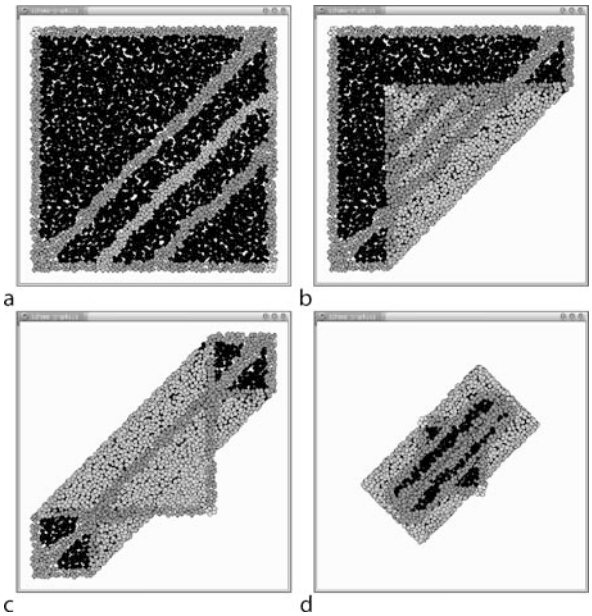

Amorphous Computing, Figure 3
Folding an envelope structure (from [38]). A pattern of lines is constructed according to origami axioms. Elements then coordinate to fold the sheet using an actuation model based on epithelial cell morphogenesis. In the figure, black indicates the front side of the sheet, grey indicates the back side, and the various colored bands show the folds and creases that are generated by the amorphous process. The small white spots show gaps in the sheet cause by "dead" or missing cells—the process works despite these

As a programming framework, the origami language has more structure than the growing point language, because the origami methods allow composition of shape constructions. On the other hand, once a shape has been constructed, there is no clear mechanism to maintain existing patterns in the face of changes to the material.

Dynamic Recruitment

In Butera [10]'s "paintable computing", processes dynamically recruit computational particles from an amorphous computer to implement their goals. As one of his examples Butera uses dynamic recruitment to implement a robust storage system for streaming audio and images (Fig. 4). Fragments of the image and audio stream circulate freely through the amorphous computer and are marshaled to a port when needed. The audio fragments also sort themselves into a playable stream as they migrate to the port.

To enhance robustness, there are multiple copies of the lower resolution image fragments and fewer copies of the higher resolution image fragments. Thus, the image is hard to destroy; with lost fragments the image is degraded but not destroyed.

Clement and Nagpal [12] also use dynamic recruitment in the development of active gradients, as described below.

Growth and Regeneration

Kondacs [30] showed how to synthesize arbitrary two-dimensional shapes by growing them. These computing units, or "cells", are identically-programmed and decentralized, with the ability to engage in only limited, local communication. Growth is implemented by allowing cells to multiply. Each of his cells may create a child cell and place it randomly within a ring around the mother cell. Cells may also die, a fact which Kondacs puts to use for temporary scaffolding when building complex shapes. If a structure requires construction of a narrow neck between two other structures it can be built precisely by laying down a thick connection and later trimming it to size.

Attributes of this system include scalability, robustness, and the ability for self-repair. Just as a starfish can regenerate its entire body from part of a limb, his system can self-repair in the event of agent death: his sphere-network representation allows the structure to be grown starting from any sphere, and every cell contains all necessary information for reproducing the missing structure.

Abstraction to Continuous Space and Time

The amorphous model postulates computing particles distributed throughout a space. If the particles are dense, one can imagine the particles as actually filling the space, and create programming abstractions that view the space itself as the object being programmed, rather than the collection of particles. Beal and Bachrach [1,7] pursued this approach by creating a language, Proto, where programmers specify the behavior of an amorphous computer as though it were a continuous material filling the space it occupies. Proto programs manipulate fields of values spanning the entire space. Programming primitives are designed to make it simple to compile global operations to operations at each point of the continuum. These operations are approximated by having each device represent a nearby chunk of space. Programs are specified in space and time units that are independent of the distribution of particles and of the particulars of communication and execution on those particles (Fig. 5). Programs are composed functionally, and many of the details of communication and composition are made implicit by Proto's runtime system, allowing complex programs to be expressed simply. Proto has been applied to applications in sensor networks like target tracking and threat avoidance, to swarm robotics and to modular robotics, e. g., generating a planar wave for coordinated actuation.

Newton's language Regiment [41,42] also takes a continuous view of space and time. Regiment is organized in terms of stream operations, where each stream represents

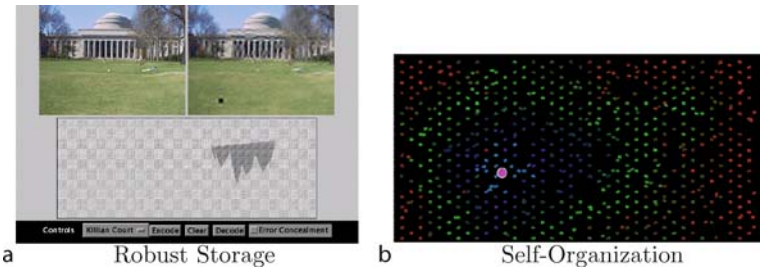

Amorphous Computing, Figure 4

Butera dynamically controls the flow of information through an amorphous computer. In a image fragments spread through the computer so that a degraded copy can be recovered from any segment; the original image is on the left, the blurry copy on the right has been recovered from the small region shown below. In b audio fragments sort themselves into a playable stream as they migrate to an output port; cooler colors are earlier times (from [10]).

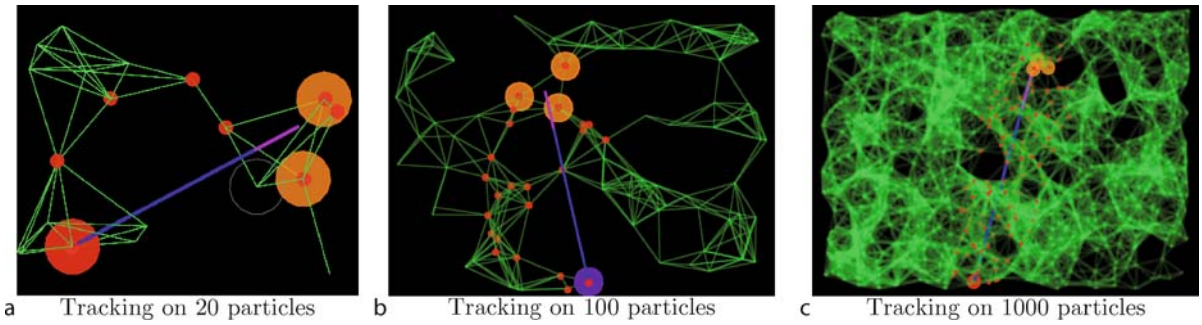

Amorphous Computing, Figure 5
A tracking program written in Proto sends the location of a target region (*orange*) to a listener (*red*) along a channel (*small red dots*) in the network (indicated by *green lines*). The continuous space and time abstraction allows the same program to run at different resolutions

a time-varying quantity over a part of space, for example, the average value of the temperature over a disc of a given radius centered at a designated point. Regiment, also a functional language, is designed to gather streams of data from regions of the amorphous computer and accumulate them at a single point. This assumption allows Regiment to provide region-wide summary functions that are difficult to implement in Proto.

Primitives for Amorphous Computing

The previous section illustrated some paradigms that have been developed for programming amorphous systems, each paradigm building on some organizing metaphor. But eventually, meeting the challenge of amorphous systems will require a more comprehensive linguistic framework. We can approach the task of creating such a framework following the perspective in [59], which views languages in terms of primitives, means of combination, and means of abstraction.

The fact that amorphous computers consist of vast numbers of unreliable and unsynchronized particles, arranged in space in ways that are locally unknown, constrains the primitive mechanisms available for organizing cooperation among the particles. While amorphous computers are naturally massively parallel, the kind of computation that they are most suited for is parallelism that does not depend on explicit synchronization and the use of atomic operations to control concurrent access to resources. However, there are large classes of useful behaviors that can be implemented without these tools. Primitive mechanisms that are appropriate for specifying behavior on amorphous computers include gossip, random choice, fields, and gradients.

Gossip

Gossip, also known as epidemic communication [17,20], is a simple communication mechanism. The goal of a gossip computation is to obtain an agreement about the value of some parameter. Each particle broadcasts its opinion of the parameter to its neighbors, and computation is performed by each particle combining the values that it receives from its neighbors, without consideration of the identification of the source. If the computation changes a particle's opinion of the value, it rebroadcasts its new opinion. The process concludes when the are no further broadcasts.

For example, an aggregate can agree upon the minimum of the values held by all the particles as follows. Each particle broadcasts its value. Each recipient compares its current value with the value that it receives. If the received value is smaller than its current value, it changes its current value to that minimum and rebroadcasts the new value.

The advantage of gossip is that it flows in all directions and is very difficult to disrupt. The disadvantage is that the lack of source information makes it difficult to revise a decision.

Random Choice

Random choice is used to break symmetry, allowing the particles to differentiate their behavior. The simplest use of random choice is to establish local identity of particles: each particle chooses a random number to identify itself to its neighbors. If the number of possible choices is large enough, then it is unlikely that any nearby particles will choose the same number, and this number can thus be used as an identifier for the particle to its neighbors. Random choice can be combined with gossip to elect

leaders, either for the entire system or for local regions. (If collisions in choice can be detected, then the number of choices need not be much higher than then number of neighbors. Also, using gossip to elect leaders makes sense only when we expect a leader to be long-lived, due to the difficulty of changing the decision to designate a replacement leader.)

To elect a single leader for the entire system, every particle chooses a value, then gossips to find the minimum. The particle with the minimum value becomes the leader. To elect regional leaders, we instead use gossip to carry the identity of the first leader a particle has heard of. Each particle uses random choice as a "coin flip" to decide when to declare itself a leader; if the flip comes up heads enough times before the particle hears of another leader, the particle declares itself a leader and broadcasts that fact to its neighbors. The entire system is thus broken up into contiguous domains of particles who first heard some particular particle declare itself a leader.

One challenge in using random choice on an amorphous computer is to ensure that the particulars of particle distribution do not have an unexpected effect on the outcome. For example, if we wish to control the expected size of the domain that each regional leader presides over, then the probability of becoming a leader must depend on the density of the particles.

Fields

Every component of the state of the computational particles in an amorphous computer may be thought of as a field over the discrete space occupied by those particles. If the density of particles is large enough this field of values may be thought of as an approximation of a field on the continuous space.

We can make amorphous models that approximate the solutions of the classical partial-differential equations of physics, given appropriate boundary conditions. The amorphous methods can be shown to be consistent, convergent and stable.

For example, the algorithm for solving the Laplace equation with Dirichlet conditions is analogous to the way it would be solved on a lattice. Each particle must repeatedly update the value of the solution to be the average of the solutions posted by its neighbors, but the boundary points must not change their values. This algorithm will eventually converge, although very slowly, independent of the order of the updates and the details of the local connectedness of the network. There are optimizations, such as over-relaxation, that are just as applicable in the amorphous context as on a regular grid.

Katzenelson [27] has shown similar results for the diffusion equation, complete with analytic estimates of the errors that arise from the discrete and irregularly connected network. In the diffusion equation there is a conserved quantity, the amount of material diffusing. Rauch [47] has shown how this can work with the wave equation, illustrating that systems that conserve energy and momentum can also be effectively modeled with an amorphous computer. The simulation of the wave equation does require that the communicating particles know their relative positions, but it is not hard to establish local coordinate systems.

Gradients

An important primitive in amorphous computing is the gradient, which estimates the distance from each particle to the nearest particle designated as a source. The gradient is inspired by the chemical-gradient diffusion process that is crucial to biological development. Amorphous computing builds on this idea, but does not necessarily compute the distance using diffusion because simulating diffusion can be expensive.

The common alternative is a linear-time mechanism that depends on active computation and relaying of information rather than passive diffusion. Calculation of a gradient starts with each source particle setting its distance estimate to zero, and every other particle setting its distance estimate to infinity. The sources then broadcast their estimates to their neighbors. When a particle receives a message from its neighbor, it compares its current distance estimate to the distance through its neighbor. If the distance through its neighbor is less, it chooses that to be its estimate, and broadcasts its new estimate onwards.

Although the basic form of the gradient is simple, there are several ways in which gradients can be varied to better match the context in which they are used. These choices may be made largely independently, giving a wide variety of options when designing a system. Variations which have been explored include:

Active Gradients An active gradient [12,14,16] monitors its validity in the face of changing sources and device failure, and maintains correct distance values. For example, if the supporting sources disappear, the gradient is deallocated. A gradient may also carry version information, allowing its source to change more smoothly.

Active gradients can provide self-repairing coordinate systems, as a foundation for robust construction of patterns. Figure 6 shows the "count-down wave" line described above in the introduction to this article. The line's

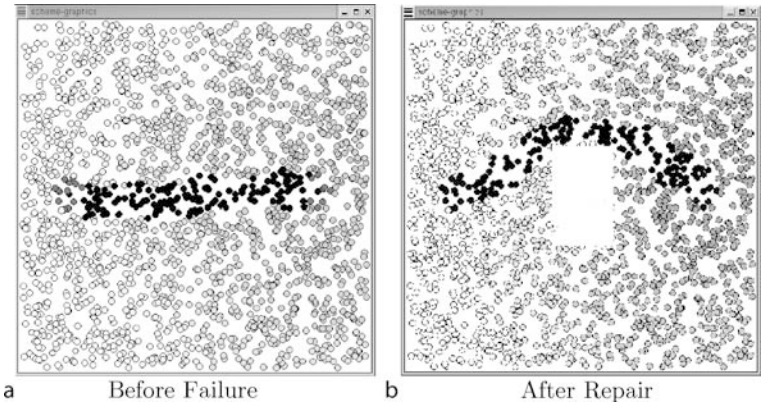

Amorphous Computing, Figure 6
A line being maintained by active gradients, from [12]. A line (*black*) is constructed between two anchor regions (*dark grey*) based on the active gradient emitted by the right anchor region (*light grays*). The line is able to rapidly repair itself following failures because the gradient actively maintains itself

implementation in terms of active gradients provides for self-repair when the underlying amorphous computer is damaged.

Polarity A gradient may be set to count down from a positive value at the source, rather than to count up from zero. This bounds the distance that the gradient can span, which can help limit resource usage, but may limit the scalability of programs.

Adaptivity As described above, a gradient relaxes once to a distance estimate. If communication is expensive and precision unimportant, the gradient can take the first value that arrives and ignore all subsequent values. If we want the gradient to adapt to changes in the distribution of particles or sources, then the particles need to broadcast at regular intervals. We can then have estimates that converge smoothly to precise estimates by adding a restoring force which acts opposite to the relaxation, allowing the gradient to rise when unconstrained by its neighbors. If, on the other hand, we value adaptation speed over smoothness, then each particle can recalculate its distance estimate from scratch with each new batch of values.

Carrier Normally, the distance value calculated by the gradient is the signal we are interested in. A gradient may instead be used to carry an arbitrary signal outward from the source. In this case, the value at each particle is the most recently arrived value from the nearest source.

Distance Measure A gradient's distance measure is, of course, dependent on how much knowledge we have about the relative positions of neighbors. It is sometimes advantageous to discard good information and use only hop-count values, since it is easier to make an adaptive gradient using hop-count values. Non-linear distance measures are also possible, such as a count-down gradient that decays exponentially from the source. Finally, the value of a gradient may depend on more sources than the nearest (this is the case for a chemical gradient), though this may be very expensive to calculate.

Coordinates and Clusters

Computational particles may be built with restrictions about what can be known about local geometry. A particle may know that it can reliably communicate with a few neighbors. If we assume that these neighbors are all within a disc of some approximate communication radius then distances to others may be estimated by minimum hop count [28]. However, it is possible that more elaborate particles can estimate distances to near neighbors. For example, the Cricket localization system [45] uses the fact that sound travels more slowly than radio, so the distance is estimated by the difference in time of arrival between simultaneously transmitted signals. McLurkin's swarmbots [35] use the ISIS communication system that gives bearing and range information. However, it is possible for a sufficiently dense amorphous computer to produce local coordinate systems for its particles with even the crudest method of determining distances. We can make an atlas of overlapping coordinate systems, using random symmetry breaking to make new starting baselines [3]. These coordinate systems can be combined and made consistent to form a manifold, even if the amorphous computer is not flat or simply connected.

One way to establish coordinates is to choose two initial particles that are a known distance apart. Each one serves as the source of a gradient. A pair of rectangular axes can be determined by the shortest path between them and by a bisector constructed where the two gradients are equal. These may be refined by averaging and calibrated using the known distance between the selected particles. After the axes are established, they may source new gradients that can be combined to make coordinates for the region near these axes. The coordinate system can be further refined using further averaging. Other natural coordinate constructions are bipolar elliptical. This kind of construction was pioneered by Coore [14] and Nagpal [37]. Katzenelson [27] did early work to determine the kind of accuracy that can be expected from such a construction.

Spatial clustering can be accomplished with any of a wide variety of algorithms, such as the clubs algorithm [16], LOCI [36], or persistent node partitioning [4]. Clusters can themselves be clustered, forming a hierarchical clustering of logarithmic height.

Means of Combination and Abstraction

A programming framework for amorphous systems requires more than primitive mechanisms. We also need suitable means of combination, so that programmers can combine behaviors to produce more complex behaviors, and means of abstraction so that the compound behaviors can be named and manipulated as units. Here are a few means of combination that have been investigated with amorphous computing.

Spatial and Temporal Sequencing

Several behaviors can be strung together in a sequence. The challenge in controlling such a sequence is to determine when one phase has completed and it is safe to move on to the next. Trigger rules can be used to detect completion locally.

In Coore's Growing Point Language [15], all of the sequencing decisions are made locally, with different growing points progressing independently. There is no difficulty of synchronization in this approach because the only time when two growing points need to agree is when they have become spatially coincident. When growing points merge, the independent processes are automatically synchronized.

Nagpal's origami language [38] has long-range operations that cannot overlap in time unless they are in non-interacting regions of the space. The implementation uses barrier synchronization to sequence the operations: when completion is detected locally, a signal is propagated throughout a marked region of the sheet, and the next operation begins after a waiting time determined by the diameter of the sheet.

With adaptive gradients, we can use the presence of an inducing signal to run an operation. When the induction signal disappears, the operation ceases and the particles begin the next operation. This allows sequencing to be triggered by the last detection of completion rather than by the first.

Pipelining

If a behavior is self-stabilizing (meaning that it converges to a correct state from any arbitrary state) then we can use it in a sequence without knowing when the previous phase completes. The evolving output of the previous phase serves as the input of this next phase, and once the preceding behavior has converged, the self-stabilizing behavior will converge as well.

If the previous phase evolves smoothly towards its final state, then by the time it has converged, the next phase may have almost converged as well, working from its partial results. For example, the coordinate system mechanism described above can be pipelined; the final coordinates are being formed even as the farther particles learn that they are not on one of the two axes.

Restriction to Spatial Regions

Because the particles of an amorphous computer are distributed in space it is natural to assign particular behaviors to specific spatial regions. In Beal and Bachrach's work, restriction of a process to a region is a primitive[7]. As another example, when Nagpal's system folds an origami construction, regions on different faces may differentiate so that they fold in different patterns. These folds may, if the physics permits, be performed simultaneously. It may be necessary to sequence later construction that depends on the completion of the substructures.

Regions of space can be named using coordinates, clustering, or implicitly through calculations on fields. Indeed, one could implement solid modelling on an amorphous computer. Once a region is identified, a particle can test whether it is a member of that region when deciding whether to run a behavior. It is also necessary to specify how a particle should change its behavior if its membership in a region may vary with time.

Modularity and Abstraction

Standard means of abstraction may be applied in an amorphous computing context, such as naming procedures,

data structures, and processes. The question for amorphous computing is what collection of entities is useful to name.

Because geometry is essential in an amorphous computing context, it becomes appropriate to describe computational processes in terms of geometric entities. Thus there are new opportunities for combining geometric structures and naming the combinations. For example, it is appropriate to compute with, combine, and name regions of space, intervals of time, and fields defined on them [7,42]. It may also be useful to describe the propagation of information through the amorphous computer in terms of the light cone of an event [2].

Not all traditional abstractions extend nicely to an amorphous computing context because of the challenges of scale and the fallibility of parts and interconnect. For example, atomic transactions may be excessively expensive in an amorphous computing context. And yet, some of the goals that a programmer might use atomic transactions to accomplish, such as the approximate enforcement of conservation laws, can be obtained using techniques that are compatible with an amorphous environment, as shown by Rauch [47].

Supporting Infrastructure and Services

Amorphous computing languages, with their primitives, means of combination, and means of abstraction, rest on supporting services. One example, described above, is the automatic message propagation and decay in Weiss's Microbial Colony Language [55]. MCL programs do not need to deal with this explicitly because it is incorporated into the operating system of the MCL machine. Experience with amorphous computing is beginning to identify other key services that amorphous machines must supply.

Particle Identity

Particles must be able to choose identifiers for communicating with their neighbors. More generally, there are many operations in an amorphous computation where the particles may need to choose numbers, with the property that individual particles choose different numbers.

If we are willing to pay the cost, it is possible to build unique identifiers into the particles, as is done with current macroscopic computers. We need only locally unique identifiers, however, so we can obtain them using pseudorandom-number generators. On the surface of it, this may seem problematic, since the particles in the amorphous are assumed to be manufactured identically, with identical programs. There are, however, ways to obtain individualized random numbers. For example, the particles are not synchronized, and they are not really physically identical, so they will run at slightly different rates. This difference is enough to allow pseudorandom-number generators to get locally out of synch and produce different sequences of numbers. Amorphous computing particles that have sensors may also get seeds for their pseudorandom-number generators from sensor noise.

Local Geometry and Gradients

Particles must maintain connections with their neighbors, tracking who they can reliably communicate with, and whatever local geometry information is available. Because particles may fail or move, this information needs to be maintained actively. The geometry information may include distance and bearing to each neighbor, as well as the time it takes to communicate with each neighbor. But many implementations will not be able to give significant distance or bearing information. Since all of this information may be obsolete or inaccurately measured, the particles must also maintain information on how reliable each piece of information is.

An amorphous computer must know the dimension of the space it occupies. This will generally be a constant—either the computer covers a surface or fills a volume. In rare cases, however, the effective dimension of a computer may change: for example, paint is three-dimensional in a bucket and two-dimensional once applied. Combining this information with how the number of accessible correspondents changes with distance, an amorphous process can derive curvature and local density information.

An amorphous computer should also support gradient propagation as part of the infrastructure: a programmer should not have to explicitly deal with the propagation of gradients (or other broadcast communications) in each particle. A process may explicitly initiate a gradient, or explicitly react to one that it is interested in, but the propagation of the gradient through a particle should be automatically maintained by the infrastructure.

Implementing Communication

Communication between neighbors can occur through any number of mechanisms, each with its own set of properties: amorphous computing systems have been built that communicate through directional infrared [35], RF broadcast [22], and low-speed serial cables [9,46]. Simulated systems have also included other mechanisms such as signals superimposed on the power connections [11] and chemical diffusion [55].

Communication between particles can be made implicit with a neighborhood shared memory. In this arrangement, each particle designates some of its internal state to be shared with its neighbors. The particles regularly communicate, giving each particle a best-effort view of the exposed portions of the states of its neighbors. The contents of the exposed state may be specified explicitly [10,35,58] or implicitly [7]. The shared memory allows the system to be tuned by trading off communication rate against the quality of the synchronization, and decreasing transmission rates when the exposed state is not changing.

Lessons for Engineering

As von Neumann remarked half a century ago, biological systems are strikingly robust when compared with our artificial systems. Even today software is fragile. Computer science is currently built on a foundation that largely assumes the existence of a perfect infrastructure. Integrated circuits are fabricated in clean-room environments, tested deterministically, and discarded if even a single defect is uncovered. Entire software systems fail with single-line errors. In contrast, biological systems rely on local computation, local communication, and local state, yet they exhibit tremendous resilience.

Although this contrast is most striking in computer science, amorphous computing can provide lessons throughout engineering. Amorphous computing concentrates on making systems flexible and adaptable at the expense of efficiency. Amorphous computing requires an engineer to work under extreme conditions. The engineer must arrange the cooperation of vast numbers of identical computational particles to accomplish prespecified goals, but may not depend upon the numbers. We may not depend on any prespecified interconnect of the particles. We may not depend on synchronization of the particles. We may not depend on the stability of the communications system. We may not depend on the long-term survival of any individual particles. The combination of these obstacles forces us to abandon many of the comforts that are available in more typical engineering domains.

By restricting ourselves in this way we obtain some robustness and flexibility, at the cost of potentially inefficient use of resources, because the algorithms that are appropriate are ones that do not take advantage of these assumptions. Algorithms that work well in an amorphous context depend on the average behavior of participating particles. For example, in Nagpal's origami system a fold that is specified will be satisfactory if it is approximately in the right place and if most of the particles on the specified fold line agree that they are part of the fold line: dissenters will be overruled by the majority. In Proto a programmer can address only regions of space, assumed to be populated by many particles. The programmer may not address individual particles, so failures of individual particles are unlikely to make major perturbations to the behavior of the system.

An amorphous computation can be quite immune to details of the macroscopic geometry as well as to the interconnectedness of the particles. Since amorphous computations make their own local coordinate systems, they are relatively independent of coordinate distortions. In an amorphous computation we accept a wide range of outcomes that arise from variations of the local geometry. Tolerance of local variation can lead to surprising flexibility: the mechanisms which allow Nagpal's origami language to tolerate local distortions allow programs to distort globally as well, and Nagpal shows how such variations can account for the variations in the head shapes of related species of *Drosophila* [38]. In Coore's language one specifies the topology of the pattern to be constructed, but only limited information about the geometry. The topology will be obtained, regardless of the local geometry, so long as there is sufficient density of particles to support the topology. Amorphous computations based on a continuous model of space (as in Proto) are naturally scale independent.

Since an amorphous computer is composed of unsynchronized particles, a program may not depend upon a priori timing of events. The sequencing of phases of a process must be determined by either explicit termination signals or with times measured dynamically. So amorphous computations are time-scale independent by construction.

A program for an amorphous computer may not depend on the reliability of the particles or the communication paths. As a consequence it is necessary to construct the program so as to dynamically compensate for failures. One way to do this is to specify the result as the satisfaction of a set of constraints, and to build the program as a homeostatic mechanism that continually drives the system toward satisfaction of those constraints. For example, an active gradient continually maintains each particle's estimate of the distance to the source of the gradient. This can be used to establish and maintain connections in the face of failures of particles or relocation of the source. If a system is specified in this way, repair after injury is a continuation of the development process: an injury causes some constraints to become unsatisfied, and the development process builds new structure to heal the injury.

By restricting the assumptions that a programmer can rely upon we increase the flexibility and reliability of the

programs that are constructed. However, it is not yet clear how this limits the range of possible applications of amorphous computing.

Future Directions

Computer hardware is almost free, and in the future it will continue to decrease in price and size. Sensors and actuators are improving as well. Future systems will have vast numbers of computing mechansisms with integrated sensors and actuators, to a degree that outstrips our current approaches to system design. When the numbers become large enough, the appropriate programming technology will be amorphous computing. This transition has already begun to appear in several fields:

Sensor networks The success of sensor network research has encouraged the planning and deployment of ever-larger numbers of devices. The ad-hoc, time-varying nature of sensor networks has encouraged amorphous approaches, such as communication through directed diffusion [25] and Newton's Regiment language [42].

Robotics Multi-agent robotics is much like sensor networks, except that the devices are mobile and have actuators. *Swarm robotics* considers independently mobile robots working together as a team like ants or bees, while *modular robotics* consider robots that physically attach to one another in order to make shapes or perform actions, working together like the cells of an organism. Gradients are being used to create "flocking" behaviors in swarm robotics [35,44]. In modular robotics, Stoy uses gradients to create shapes [51] while De Rosa et al. form shapes through stochastic growth and decay [49].

Pervasive computing Pervasive computing seeks to exploit the rapid proliferation of wireless computing devices throughout our everyday environment. Mamei and Zambonelli's TOTA system [32] is an amorphous computing implementation supporting a model of programming using fields and gradients [33]. Servat and Drogoul have suggested combining amorphous computing and reactive agent-based systems to produce something they call "pervasive intelligence" [50].

Multicore processors As it becomes more difficult to increase processor speed, chip manufacturers are looking for performance gains through increasing the number of processing cores per chip. Butera's work [10] looks toward a future in which there are thousands of cores per chip and it is no longer reasonable to assume they are all working or have them communicate all-to-all.

While much of amorphous computing research is inspired by biological observations, it is also likely that insights and lessons learned from programming amorphous computers will help elucidate some biological problems [43]. Some of this will be stimulated by the emerging engineering of biological systems. Current work in synthetic biology [24,48,61] is centered on controlling the molecular biology of cells. Soon synthetic biologists will begin to engineer biofilms and perhaps direct the construction of multicellular organs, where amorphous computing will become an essential technological tool.

Bibliography

Primary Literature

1. Bachrach J, Beal J (2006) Programming a sensor network as an amorphous medium. In: DCOSS 2006 Posters, June 2006
2. Bachrach J, Beal J, Fujiwara T (2007) Continuous space-time semantics allow adaptive program execution. In: IEEE International Conference on Self-Adaptive and Self-Organizing Systems, 2007
3. Bachrach J, Nagpal R, Salib M, Shrobe H (2003) Experimental results and theoretical analysis of a self-organizing global coordinate system for ad hoc sensor networks. Telecommun Syst J, Special Issue on Wireless System Networks 26(2–4):213–233
4. Beal J (2003) A robust amorphous hierarchy from persistent nodes. In: Commun Syst Netw
5. Beal J (2004) Programming an amorphous computational medium. In: Unconventional Programming Paradigms International Workshop, September 2004
6. Beal J (2005) Amorphous medium language. In: Large-Scale Multi-Agent Systems Workshop (LSMAS). Held in Conjunction with AAMAS-05
7. Beal J, Bachrach J (2006) Infrastructure for engineered emergence on sensor/actuator networks. In: IEEE Intelligent Systems, 2006
8. Beal J, Sussman G (2005) Biologically-inspired robust spatial programming. Technical Report AI Memo 2005-001, MIT, January 2005
9. Beebee W M68hc11 gunk api book. http://www.swiss.ai.mit.edu/projects/amorphous/HC11/api.html. Accessed 31 May 2007
10. Butera W (2002) Programming a Paintable Computer. Ph D thesis, MIT
11. Campbell J, Pillai P, Goldstein SC (2005) The robot is the tether: Active, adaptive power routing for modular robots with unary inter-robot connectors. In: IROS 2005
12. Clement L, Nagpal R (2003) Self-assembly and self-repairing topologies. In: Workshop on Adaptability in Multi-Agent Systems, RoboCup Australian Open, January 2003
13. Codd EF (1968) Cellular Automata. Academic Press, New York
14. Coore D (1998) Establishing a coordinate system on an amorphous computer. In: MIT Student Workshop on High Performance Computing, 1998
15. Coore D (1999) Botanical Computing: A Developmental Approach to Generating Interconnect Topologies on an Amorphous Computer. Ph D thesis, MIT

16. Coore D, Nagpal R, Weiss R (1997) Paradigms for structure in an amorphous computer. Technical Report AI Memo 1614, MIT
17. Demers A, Greene D, Hauser C, Irish W, Larson J, Shenker S, Stuygis H, Swinehart D, Terry D (1987) Epidemic algorithms for replicated database maintenance. In: 7th ACM Symposium on Operating Systems Principles, 1987
18. D'Hondt E, D'Hondt T (2001) Amorphous geometry. In: ECAL 2001
19. D'Hondt E, D'Hondt T (2001) Experiments in amorphous geometry. In: 2001 International Conference on Artificial Intelligence
20. Ganesan D, Krishnamachari B, Woo A, Culler D, Estrin D, Wicker S (2002) An empirical study of epidemic algorithms in large scale multihop wireless networks. Technical Report IRB-TR-02-003, Intel Research Berkeley
21. Gayle O, Coore D (2006) Self-organizing text in an amorphous environment. In: ICCS 2006
22. Hill J, Szewcyk R, Woo A, Culler D, Hollar S, Pister K (2000) System architecture directions for networked sensors. In: ASPLOS, November 2000
23. Huzita H, Scimemi B (1989) The algebra of paper-folding. In: First International Meeting of Origami Science and Technology, 1989
24. igem 2006: international genetically engineered machine competition (2006) http://www.igem2006.com. Accessed 31 May 2007
25. Intanagonwiwat C, Govindan R, Estrin D (2000) Directed diffusion: a scalable and robust communication paradigm for sensor networks. In: Mobile Computing and Networking, pp 56–67
26. Kahn JM, Katz RH, Pister KSJ (1999) Mobile networking for smart dust. In: ACM/IEEE Int. Conf. on Mobile Computing and Networking (MobiCom 99), August 1999
27. Katzenelson J (1999) Notes on amorphous computing. (Unpublished Draft)
28. Kleinrock L, Sylvester J (1978) Optimum transmission radii for packet radio networks or why six is a magic number. In: IEEE Natl Telecommun. Conf, December 1978, pp 4.3.1–4.3.5
29. Knight TF, Sussman GJ (1998) Cellular gate technology. In: First International Conference on Unconventional Models of Computation (UMC98)
30. Kondacs A (2003) Biologically-inspired self-assembly of 2d shapes, using global-to-local compilation. In: International Joint Conference on Artificial Intelligence (IJCAI)
31. Mamei M, Zambonelli F (2003) Spray computers: Frontiers of self-organization for pervasive computing. In: WOA 2003
32. Mamei M, Zambonelli F (2004) Spatial computing: the tota approach. In: WOA 2004, pp 126–142
33. Mamei M, Zambonelli F (2005) Physical deployment of digital pheromones through rfid technology. In: AAMAS 2005, pp 1353–1354
34. Margolus N (1988) Physics and Computation. Ph D thesis, MIT
35. McLurkin J (2004) Stupid robot tricks: A behavior-based distributed algorithm library for programming swarms of robots. Master's thesis, MIT
36. Mittal V, Demirbas M, Arora A (2003) Loci: Local clustering service for large scale wireless sensor networks. Technical Report OSU-CISRC-2/03-TR07, Ohio State University
37. Nagpal R (1999) Organizing a global coordinate system from local information on an amorphous computer. Technical Report AI Memo 1666, MIT
38. Nagpal R (2001) Programmable Self-Assembly: Constructing Global Shape using Biologically-inspired Local Interactions and Origami Mathematics. Ph D thesis, MIT
39. Nagpal R, Mamei M (2004) Engineering amorphous computing systems. In: Bergenti F, Gleizes MP, Zambonelli F (eds) Methodologies and Software Engineering for Agent Systems, The Agent-Oriented Software Engineering Handbook. Kluwer, New York, pp 303–320
40. Dust Networks. http://www.dust-inc.com. Accessed 31 May 2007
41. Newton R, Morrisett G, Welsh M (2007) The regiment macroprogramming system. In: International Conference on Information Processing in Sensor Networks (IPSN'07)
42. Newton R, Welsh M (2004) Region streams: Functional macroprogramming for sensor networks. In: First International Workshop on Data Management for Sensor Networks (DMSN), August 2004
43. Patel A, Nagpal R, Gibson M, Perrimon N (2006) The emergence of geometric order in proliferating metazoan epithelia. Nature 442:1038–1041
44. Payton D, Daily M, Estowski R, Howard M, Lee C (2001) Pheromone robotics. Autonomous Robotics 11:319–324
45. Priyantha N, Chakraborty A, Balakrishnan H (2000) The cricket location-support system. In: ACM International Conference on Mobile Computing and Networking (ACM MOBICOM), August 2000
46. Raffle H, Parkes A, Ishii H (2004) Topobo: A constructive assembly system with kinetic memory. CHI
47. Rauch E (1999) Discrete, amorphous physical models. Master's thesis, MIT
48. Registry of standard biological parts. http://parts.mit.edu. Accessed 31 May 2007
49. De Rosa M, Goldstein SC, Lee P, Campbell J, Pillai P (2006) Scalable shape sculpting via hole motion: Motion planning in lattice-constrained module robots. In: Proceedings of the 2006 IEEE International Conference on Robotics and Automation (ICRA '06), May 2006
50. Servat D, Drogoul A (2002) Combining amorphous computing and reactive agent-based systems: a paradigm for pervasive intelligence? In: AAMAS 2002
51. Stoy K (2003) Emergent Control of Self-Reconfigurable Robots. Ph D thesis, University of Southern Denmark
52. Sutherland A (2003) Towards rseam: Resilient serial execution on amorphous machines. Master's thesis, MIT
53. von Neumann J (1951) The general and logical theory of automata. In: Jeffress L (ed) Cerebral Mechanisms for Behavior. Wiley, New York, p 16
54. Weiss R, Knight T (2000) Engineered communications for microbial robotics. In: Sixth International Meeting on DNA Based Computers (DNA6)
55. Weiss R (2001) Cellular Computation and Communications using Engineered Genetic Regular Networks. Ph D thesis, MIT
56. Welsh M, Mainland G (2004) Programming sensor networks using abstract regions. In: Proceedings of the First USENIX/ACM Symposium on Networked Systems Design and Implementation (NSDI '04), March 2004
57. Werfel J, Bar-Yam Y, Nagpal R (2005) Building patterned structures with robot swarms. In: IJCAI
58. Whitehouse K, Sharp C, Brewer E, Culler D (2004) Hood: a neighborhood abstraction for sensor networks. In: Proceedings of the 2nd international conference on Mobile systems, applications, and services.

59. Abelson H, Sussman GJ, Sussman J (1996) Structure and Interpretation of Computer Programs, 2nd edn. MIT Press, Cambridge
60. Ashe HL, Briscoe J (2006) The interpretation of morphogen gradients. Development 133:385–94
61. Weiss R, Basu S, Hooshangi S, Kalmbach A, Karig D, Mehreja R, Netravali I (2003) Genetic circuit building blocks for cellular computation, communications, and signal processing. Natural Computing 2(1):47–84

Books and Reviews

Abelson H, Allen D, Coore D, Hanson C, Homsy G, Knight T, Nagpal R, Rauch E, Sussman G, Weiss R (1999) Amorphous computing. Technical Report AIM-1665, MIT

Analog Computation

Bruce J. MacLennan
Department of Electrical Engineering & Computer Science, University of Tennessee,
Knoxville, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Fundamentals of Analog Computing
Analog Computation in Nature
General-Purpose Analog Computation
Analog Computation and the Turing Limit
Analog Thinking
Future Directions
Bibliography

Glossary

Accuracy The closeness of a computation to the corresponding primary system.

BSS The theory of computation over the real numbers defined by Blum, Shub, and Smale.

Church–Turing (CT) computation The model of computation based on the Turing machine and other equivalent abstract computing machines; commonly accepted as defining the limits of digital computation.

EAC Extended analog computer defined by Rubel.

GPAC General-purpose analog computer.

Nomograph A device for the graphical solution of equations by means of a family of curves and a straightedge.

ODE Ordinary differential equation.

PDE Partial differential equation.

Potentiometer A variable resistance, adjustable by the computer operator, used in electronic analog computing as an attenuator for setting constants and parameters in a computation.

Precision The quality of an analog representation or computation, which depends on both resolution and stability.

Primary system The system being simulated, modeled, analyzed, or controlled by an analog computer, also called the *target system*.

Scaling The adjustment, by constant multiplication, of variables in the primary system (including time) so that the corresponding variables in the analog systems are in an appropriate range.

TM Turing machine.

Definition of the Subject

Although analog computation was eclipsed by digital computation in the second half of the twentieth century, it is returning as an important alternative computing technology. Indeed, as explained in this article, theoretical results imply that analog computation can escape from the limitations of digital computation. Furthermore, analog computation has emerged as an important theoretical framework for discussing computation in the brain and other natural systems.

Analog computation gets its name from an *analogy*, or systematic relationship, between the physical processes in the computer and those in the system it is intended to model or simulate (the *primary system*). For example, the electrical quantities voltage, current, and conductance might be used as analogs of the fluid pressure, flow rate, and pipe diameter. More specifically, in traditional analog computation, physical quantities in the computation obey the same mathematical laws as physical quantities in the primary system. Thus the computational quantities are proportional to the modeled quantities. This is in contrast to *digital computation*, in which quantities are represented by strings of symbols (e. g., binary digits) that have no direct physical relationship to the modeled quantities. According to the *Oxford English Dictionary* (2nd ed., s.vv. analogue, digital), these usages emerged in the 1940s.

However, in a fundamental sense all computing is based on an analogy, that is, on a systematic relationship between the states and processes in the computer and those in the primary system. In a digital computer, the relationship is more abstract and complex than simple proportionality, but even so simple an analog computer as a slide rule goes beyond strict proportion (i. e., distance on

the rule is proportional to the logarithm of the number). In both analog and digital computation—indeed in all computation—the relevant abstract mathematical structure of the problem is realized in the physical states and processes of the computer, but the realization may be more or less direct [40,41,46].

Therefore, despite the etymologies of the terms "analog" and "digital", in modern usage the principal distinction between digital and analog computation is that the former operates on discrete representations in discrete steps, while the later operated on continuous representations by means of continuous processes (e. g., MacLennan [46], Siegelmann [p. 147 in 78], Small [p. 30 in 82], Weyrick [p. 3 in 89]).

That is, the primary distinction resides in the *topologies* of the states and processes, and it would be more accurate to refer to *discrete* and *continuous computation* [p. 39 in 25]. (Consider so-called analog and digital clocks. The principal difference resides in the continuity or discreteness of the representation of time; the motion of the two (or three) hands of an "analog" clock do not mimic the motion of the rotating earth or the position of the sun relative to it.)

Introduction

History

Pre-electronic Analog Computation Just like digital calculation, analog computation was originally performed by hand. Thus we find several analog computational procedures in the "constructions" of Euclidean geometry (Euclid, fl. 300 BCE), which derive from techniques used in ancient surveying and architecture. For example, Problem II.51 is "to divide a given straight line into two parts, so that the rectangle contained by the whole and one of the parts shall be equal to the square of the other part". Also, Problem VI.13 is "to find a mean proportional between two given straight lines", and VI.30 is "to cut a given straight line in extreme and mean ratio". These procedures do not make use of measurements in terms of any fixed unit or of digital calculation; the lengths and other continuous quantities are manipulated directly (via compass and straightedge). On the other hand, the techniques involve discrete, precise operational steps, and so they can be considered *algorithms*, but over continuous magnitudes rather than discrete numbers.

It is interesting to note that the ancient Greeks distinguished continuous *magnitudes* (Grk., *megethoi*), which have physical dimensions (e. g., length, area, rate), from discrete *numbers* (Grk., *arithmoi*), which do not [49]. Euclid axiomatizes them separately (magnitudes in Book V, numbers in Book VII), and a mathematical system comprising both discrete and continuous quantities was not achieved until the nineteenth century in the work of Weierstrass and Dedekind.

The earliest known mechanical analog computer is the "Antikythera mechanism", which was found in 1900 in a shipwreck under the sea near the Greek island of Antikythera (between Kythera and Crete). It dates to the second century BCE and appears to be intended for astronomical calculations. The device is sophisticated (at least 70 gears) and well engineered, suggesting that it was not the first of its type, and therefore that other analog computing devices may have been used in the ancient Mediterranean world [22]. Indeed, according to Cicero (*Rep*. 22) and other authors, Archimedes (c. 287–c. 212 BCE) and other ancient scientists also built analog computers, such as armillary spheres, for astronomical simulation and computation. Other antique mechanical analog computers include the astrolabe, which is used for the determination of longitude and a variety of other astronomical purposes, and the *torquetum*, which converts astronomical measurements between equatorial, ecliptic, and horizontal coordinates.

A class of special-purpose analog computer, which is simple in conception but may be used for a wide range of purposes, is the *nomograph* (also, *nomogram*, *alignment chart*). In its most common form, it permits the solution of quite arbitrary equations in three real variables, $f(u, v, w) = 0$. The nomograph is a chart or graph with scales for each of the variables; typically these scales are curved and have non-uniform numerical markings. Given values for any two of the variables, a straightedge is laid across their positions on their scales, and the value of the third variable is read off where the straightedge crosses the third scale. Nomographs were used to solve many problems in engineering and applied mathematics. They improve intuitive understanding by allowing the relationships among the variables to be visualized, and facilitate exploring their variation by moving the straightedge. Lipka (1918) is an example of a course in graphical and mechanical methods of analog computation, including nomographs and slide rules.

Until the introduction of portable electronic calculators in the early 1970s, the *slide rule* was the most familiar analog computing device. Slide rules use logarithms for multiplication and division, and they were invented in the early seventeenth century shortly after John Napier's description of logarithms.

The mid-nineteenth century saw the development of the *field analogy method* by G. Kirchhoff (1824–1887) and others [33]. In this approach an electrical field in an elec-

trolytic tank or conductive paper was used to solve two-dimensional boundary problems for temperature distributions and magnetic fields [p. 34 in 82]. It is an early example of *analog field computation* (see ► Field Computation in Natural and Artificial Intelligence).

In the nineteenth century a number of mechanical analog computers were developed for integration and differentiation (e. g., Litka 1918, pp. 246–256; Clymer [15]). For example, the *planimeter* measures the area under a curve or within a closed boundary. While the operator moves a pointer along the curve, a rotating wheel accumulates the area. Similarly, the *integraph* is able to draw the integral of a given function as its shape is traced. Other mechanical devices can draw the derivative of a curve or compute a tangent line at a given point.

In the late nineteenth century William Thomson, Lord Kelvin, constructed several analog computers, including a "tide predictor" and a "harmonic analyzer", which computed the Fourier coefficients of a tidal curve [85,86]. In 1876 he described how the mechanical integrators invented by his brother could be connected together in a feedback loop in order to solve second and higher order differential equations (Small [pp. 34–35, 42 in 82], Thomson [84]). He was unable to construct this *differential analyzer*, which had to await the invention of the torque amplifier in 1927.

The torque amplifier and other technical advancements permitted Vannevar Bush at MIT to construct the first practical differential analyzer in 1930 [pp. 42–45 in 82]. It had six integrators and could also do addition, subtraction, multiplication, and division. Input data were entered in the form of continuous curves, and the machine automatically plotted the output curves continuously as the equations were integrated. Similar differential analyzers were constructed at other laboratories in the US and the UK.

Setting up a problem on the MIT differential analyzer took a long time; gears and rods had to be arranged to define the required dependencies among the variables. Bush later designed a much more sophisticated machine, the Rockefeller Differential Analyzer, which became operational in 1947. With 18 integrators (out of a planned 30), it provided programmatic control of machine setup, and permitted several jobs to be run simultaneously. Mechanical differential analyzers were rapidly supplanted by electronic analog computers in the mid-1950s, and most were disassembled in the 1960s (Bowles [10], Owens [61], Small [pp. 50–45 in 82]).

During World War II, and even later wars, an important application of optical and mechanical analog computation was in "gun directors" and "bomb sights", which performed ballistic computations to accurately target artillery and dropped ordnance.

Electronic Analog Computation in the 20th Century It is commonly supposed that electronic analog computers were superior to mechanical analog computers, and they were in many respects, including speed, cost, ease of construction, size, and portability [pp. 54–56 in 82]. On the other hand, mechanical integrators produced higher precision results (0.1% vs. 1% for early electronic devices) and had greater mathematical flexibility (they were able to integrate with respect to any variable, not just time). However, many important applications did not require high precision and focused on dynamic systems for which time integration was sufficient.

Analog computers (non-electronic as well as electronic) can be divided into *active-element* and *passive-element* computers; the former involve some kind of amplification, the latter do not [pp. 2-1–4 in 87]. Passive-element computers included the *network analyzers*, which were developed in the 1920s to analyze electric power distribution networks, and which continued in use through the 1950s [pp. 35–40 in 82]. They were also applied to problems in thermodynamics, aircraft design, and mechanical engineering. In these systems networks or grids of resistive elements or reactive elements (i. e., involving capacitance and inductance as well as resistance) were used to model the spatial distribution of physical quantities such as voltage, current, and power (in electric distribution networks), electrical potential in space, stress in solid materials, temperature (in heat diffusion problems), pressure, fluid flow rate, and wave amplitude [p. 2-2 in 87]. That is, network analyzers dealt with partial differential equations (PDEs), whereas active-element computers, such as the differential analyzer and its electronic successors, were restricted to ordinary differential equations (ODEs) in which time was the independent variable. Large network analyzers are early examples of *analog field computers* (see ► Field Computation in Natural and Artificial Intelligence).

Electronic analog computers became feasible after the invention of the *DC operational amplifier* ("op amp") c. 1940 [pp. 64, 67–72 in 82]. Already in the 1930s scientists at Bell Telephone Laboratories (BTL) had developed the DC-coupled feedback-stabilized amplifier, which is the basis of the op amp. In 1940, as the USA prepared to enter World War II, DL Parkinson at BTL had a dream in which he saw DC amplifiers being used to control an anti-aircraft gun. As a consequence, with his colleagues CA Lovell and BT Weber, he wrote a series of papers on "electrical mathematics", which described electrical circuits to

"operationalize" addition, subtraction, integration, differentiation, etc. The project to produce an electronic gun-director led to the development and refinement of DC op amps suitable for analog computation.

The war-time work at BTL was focused primarily on control applications of analog devices, such as the gun-director. Other researchers, such as E. Lakatos at BTL, were more interested in applying them to general-purpose analog computation for science and engineering, which resulted in the design of the General Purpose Analog Computer (GPAC), also called "Gypsy", completed in 1949 [pp. 69–71 in 82]. Building on the BTL op amp design, fundamental work on electronic analog computation was conducted at Columbia University in the 1940s. In particular, this research showed how analog computation could be applied to the simulation of dynamic systems and to the solution of nonlinear equations.

Commercial general-purpose analog computers (GPACs) emerged in the late 1940s and early 1950s [pp. 72–73 in 82]. Typically they provided several dozen integrators, but several GPACs could be connected together to solve larger problems. Later, large-scale GPACs might have up to 500 amplifiers and compute with 0.01%–0.1% precision [p. 2-33 in 87].

Besides integrators, typical GPACs provided adders, subtracters, multipliers, fixed function generators (e. g., logarithms, exponentials, trigonometric functions), and variable function generators (for user-defined functions) [Chaps. 1.3, 2.4 in 87]. A GPAC was programmed by connecting these components together, often by means of a patch panel. In addition, parameters could be entered by adjusting potentiometers (attenuators), and arbitrary functions could be entered in the form of graphs [pp. 1-72–81, 2-154–156 in 87]. Output devices plotted data continuously or displayed it numerically [pp. 3-1–30 in 87].

The most basic way of using a GPAC was in *single-shot mode* [pp. 168–170 in 89]. First, parameters and initial values were entered into the potentiometers. Next, putting a master switch in "reset" mode controlled relays to apply the initial values to the integrators. Turning the switch to "operate" or "compute" mode allowed the computation to take place (i. e., the integrators to integrate). Finally, placing the switch in "hold" mode stopped the computation and stabilized the values, allowing them to be read from the computer (e. g., on voltmeters). Although single-shot operation was also called "slow operation" (in comparison to "repetitive operation", discussed next), it was in practice quite fast. Because all of the devices computed in parallel and at electronic speeds, analog computers usually solved problems in real-time but often much faster (Truitt and Rogers [pp. 1-30–32 in 87], Small [p. 72 in 82]).

One common application of GPACs was to explore the effect of one or more parameters on the behavior of a system. To facilitate this exploration of the parameter space, some GPACs provided a *repetitive operation mode*, which worked as follows (Weyrick [p. 170 in 89], Small [p. 72 in 82]). An electronic clock switched the computer between reset and compute modes at an adjustable rate (e. g., 10–1000 cycles per second) [p. 280, n. 1 in 2]. In effect the simulation was rerun at the clock rate, but if any parameters were adjusted, the simulation results would vary along with them. Therefore, within a few seconds, an entire family of related simulations could be run. More importantly, the operator could acquire an intuitive understanding of the system's dependence on its parameters.

The Eclipse of Analog Computing A common view is that electronic analog computers were a primitive predecessor of the digital computer, and that their use was just a historical episode, or even a digression, in the inevitable triumph of digital technology. It is supposed that the current digital hegemony is a simple matter of technological superiority. However, the history is much more complicated, and involves a number of social, economic, historical, pedagogical, and also technical factors, which are outside the scope of this article (see Small [81] and Small [82], especially Chap. 8, for more information). In any case, beginning after World War II and continuing for twenty-five years, there was lively debate about the relative merits of analog and digital computation.

Speed was an oft-cited advantage of analog computers [Chap. 8 in 82]. While early digital computers were much faster than mechanical differential analyzers, they were slower (often by several orders of magnitude) than electronic analog computers. Furthermore, although digital computers could perform individual arithmetic operations rapidly, complete problems were solved sequentially, one operation at a time, whereas analog computers operated in parallel. Thus it was argued that increasingly large problems required more time to solve on a digital computer, whereas on an analog computer they might require more hardware but not more time. Even as digital computing speed was improved, analog computing retained its advantage for several decades, but this advantage eroded steadily.

Another important issue was the comparative *precision* of digital and analog computation [Chap. 8 in 82]. Analog computers typically computed with three or four digits of precision, and it was very expensive to do much better, due to the difficulty of manufacturing the parts and other factors. In contrast, digital computers could perform

arithmetic operations with many digits of precision, and the hardware cost was approximately proportional to the number of digits. Against this, analog computing advocates argued that many problems did not require such high precision, because the measurements were known to only a few significant figures and the mathematical models were approximations. Further, they distinguished between precision and *accuracy*, which refers to the conformity of the computation to physical reality, and they argued that digital computation was often less accurate than analog, due to numerical limitations (e. g., truncation, cumulative error in numerical integration). Nevertheless, some important applications, such as the calculation of missile trajectories, required greater precision, and for these, digital computation had the advantage. Indeed, to some extent precision was viewed as inherently desirable, even in applications where it was unimportant, and it was easily mistaken for accuracy. (See Sect. "Precision" for more on precision and accuracy.)

There was even a social factor involved, in that the written programs, precision, and exactness of digital computation were associated with mathematics and science, but the hands-on operation, parameter variation, and approximate solutions of analog computation were associated with engineers, and so analog computing inherited "the lower status of engineering *vis-à-vis* science" [p. 251 in 82]. Thus the status of digital computing was further enhanced as engineering became more mathematical and scientific after World War II [pp. 247–251 in 82].

Already by the mid-1950s the competition between analog and digital had evolved into the idea that they were complementary technologies. This resulted in the development of a variety of *hybrid* analog/digital computing systems [pp. 251–253, 263–266 in 82]. In some cases this involved using a digital computer to control an analog computer by using digital logic to connect the analog computing elements, set parameters, and gather data. This improved the accessibility and usability of analog computers, but had the disadvantage of distancing the user from the physical analog system. The intercontinental ballistic missile program in the USA stimulated the further development of hybrid computers in the late 1950s and 1960s [81]. These applications required the speed of analog computation to simulate the closed-loop control systems and the precision of digital computation for accurate computation of trajectories. However, by the early 1970s hybrids were being displaced by all digital systems. Certainly part of the reason was the steady improvement in digital technology, driven by a vibrant digital computer industry, but contemporaries also pointed to an inaccurate perception that analog computing was obsolete and to a lack of education about the advantages and techniques of analog computing.

Another argument made in favor of digital computers was that they were general-purpose, since they could be used in business data processing and other application domains, whereas analog computers were essentially special-purpose, since they were limited to scientific computation [pp. 248–250 in 82]. Against this it was argued that *all* computing is essentially computing by analogy, and therefore analog computation was general-purpose because the class of analog computers included digital computers! (See also Sect. "Definition of the Subject" on computing by analogy.) Be that as it may, analog computation, as normally understood, is restricted to continuous variables, and so it was not immediately applicable to discrete data, such as that manipulated in business computing and other nonscientific applications. Therefore business (and eventually consumer) applications motivated the computer industry's investment in digital computer technology at the expense of analog technology.

Although it is commonly believed that analog computers quickly disappeared after digital computers became available, this is inaccurate, for both general-purpose and special-purpose analog computers have continued to be used in specialized applications to the present time. For example, a general-purpose *electrical* (vs. electronic) analog computer, the Anacom, was still in use in 1991. This is not technological atavism, for "there is no doubt considerable truth in the fact that Anacom continued to be used because it effectively met a need in a historically neglected but nevertheless important computer application area" [3]. As mentioned, the reasons for the eclipse of analog computing were not simply the technological superiority of digital computation; the conditions were much more complex. Therefore a change in conditions has necessitated a reevaluation of analog technology.

Analog VLSI In the mid-1980s, Carver Mead, who already had made important contributions to digital VLSI technology, began to advocate for the development of analog VLSI [51,52]. His motivation was that "the nervous system of even a very simple animal contains computing paradigms that are orders of magnitude more effective than are those found in systems made by humans" and that they "can be realized in our most commonly available technology—silicon integrated circuits" [pp. xi in 52]. However, he argued, since these natural computation systems are analog and highly non-linear, progress would require understanding neural information processing in animals and applying it in a new analog VLSI technology.

Because analog computation is closer to the physical laws by which all computation is realized (which are continuous), analog circuits often use fewer devices than corresponding digital circuits. For example, a four-quadrant adder (capable of adding two signed numbers) can be fabricated from four transistors [pp. 87–88 in 52], and a four-quadrant multiplier from nine to seventeen, depending on the required range of operation [pp. 90–96 in 52]. Intuitions derived from digital logic about what is simple or complex to compute are often misleading when applied to analog computation. For example, two transistors are sufficient to compute the logarithm or exponential, five for the hyperbolic tangent (which is very useful in neural computation), and three for the square root [pp. 70–71, 97–99 in 52]. Thus analog VLSI is an attractive approach to "post-Moore's Law computing" (see Sect. "Future Directions" below). Mead and his colleagues demonstrated a number of analog VLSI devices inspired by the nervous system, including a "silicon retina" and an "electronic cochlea" [Chaps. 15–16 in 52], research that has lead to a renaissance of interest in electronic analog computing.

Non-Electronic Analog Computation As will be explained in the body of this article, analog computation suggests many opportunities for future computing technologies. Many physical phenomena are potential media for analog computation provided they have useful mathematical structure (i. e., the mathematical laws describing them are mathematical functions useful for general- or special-purpose computation), and they are sufficiently controllable for practical use.

Article Roadmap

The remainder of this article will begin by summarizing the fundamentals of analog computing, starting with the continuous state space and the various processes by which analog computation can be organized in time. Next it will discuss analog computation in nature, which provides models and inspiration for many contemporary uses of analog computation, such as neural networks. Then we consider general-purpose analog computing, both from a theoretical perspective and in terms of practical general-purpose analog computers. This leads to a discussion of the theoretical power of analog computation and in particular to the issue of whether analog computing is in some sense more powerful than digital computing. We briefly consider the cognitive aspects of analog computing, and whether it leads to a different approach to computation than does digital computing. Finally, we conclude with some observations on the role of analog computation in "post-Moore's Law computing".

Fundamentals of Analog Computing

Continuous State Space

As discussed in Sect. "Introduction", the fundamental characteristic that distinguishes analog from digital computation is that the state space is continuous in analog computation and discrete in digital computation. Therefore it might be more accurate to call analog and digital computation *continuous* and *discrete computation*, respectively. Furthermore, since the earliest days there have been *hybrid computers* that combine continuous and discrete state spaces and processes. Thus, there are several respects in which the state space may be continuous.

In the simplest case the state space comprises a finite (generally modest) number of variables, each holding a continuous quantity (e. g., voltage, current, charge). In a traditional GPAC they correspond to the variables in the ODEs defining the computational process, each typically having some independent meaning in the analysis of the problem. Mathematically, the variables are taken to contain bounded real numbers, although complex-valued variables are also possible (e. g., in AC electronic analog computers). In a practical sense, however, their precision is limited by noise, stability, device tolerance, and other factors (discussed below, Sect. "Characteristics of Analog Computation").

In typical analog neural networks the state space is larger in dimension but more structured than in the former case. The artificial neurons are organized into one or more layers, each composed of a (possibly large) number of artificial neurons. Commonly each layer of neurons is densely connected to the next layer. In general the layers each have some meaning in the problem domain, but the individual neurons constituting them do not (and so, in mathematical descriptions, the neurons are typically numbered rather than named).

The individual artificial neurons usually perform a simple computation such as this:

$$y = \sigma(s)\,, \quad \text{where } s = b + \sum_{i=1}^{n} w_i x_i\,,$$

and where y is the activity of the neuron, $x_1, \ldots, x_n$ are the activities of the neurons that provide its inputs, b is a bias term, and $w_1, \ldots, w_n$ are the *weights* or strengths of the connections. Often the *activation function* σ is a real-valued *sigmoid* ("S-shaped") function, such as the *logistic*

sigmoid,

$$\sigma(s) = \frac{1}{1 + e^{-s}} ,$$

in which case the neuron activity y is a real number, but some applications use a discontinuous *threshold function*, such as the *Heaviside function*,

$$U(s) = \begin{cases} +1 & \text{if } s \geq 0 \\ 0 & \text{if } s < 0 \end{cases}$$

in which case the activity is a discrete quantity. The *saturated-linear* or *piecewise-linear* sigmoid is also used occasionally:

$$\sigma(s) = \begin{cases} +1 & \text{if } s > 1 \\ s & \text{if } 0 \leq s \leq 1 \\ 0 & \text{if } s < 0 . \end{cases}$$

Regardless of whether the activation function is continuous or discrete, the bias b and connection weights $w_1, \ldots, w_n$ are real numbers, as is the "net input" $s = \sum_i w_i x_i$ to the activation function. Analog computation may be used to evaluate the linear combination s and the activation function $\sigma(s)$, if it is real-valued. The biases and weights are normally determined by a learning algorithm (e. g., back-propagation), which is also a good candidate for analog implementation.

In summary, the continuous state space of a neural network includes the bias values and net inputs of the neurons and the interconnection strengths between the neurons. It also includes the activity values of the neurons, if the activation function is a real-valued sigmoid function, as is often the case. Often large groups ("layers") of neurons (and the connections between these groups) have some intuitive meaning in the problem domain, but typically the individual neuron activities, bias values, and interconnection weights do not.

If we extrapolate the number of neurons in a layer to the continuum limit, we get a *field*, which may be defined as a continuous distribution of continuous quantity (see ► Field Computation in Natural and Artificial Intelligence). Treating a group of artificial or biological neurons as a continuous mass is a reasonable mathematical approximation if their number is sufficiently large and if their spatial arrangement is significant (as it generally is in the brain). Fields are especially useful in modeling *cortical maps*, in which information is represented by the pattern of activity over a region of neural cortex.

In field computation the state space in continuous in two ways: it is continuous in variation but also in space. Therefore, field computation is especially applicable to solving PDEs and to processing spatially extended information such as visual images. Some early analog computing devices were capable of field computation [pp. 1-14–17, 2-2–16 in 87]. For example, as previously mentioned (Sect. "Introduction"), large resistor and capacitor networks could be used for solving PDEs such as diffusion problems. In these cases a discrete ensemble of resistors and capacitors was used to approximate a continuous field, while in other cases the computing medium was spatially continuous. The latter made use of conductive sheets (for two-dimensional fields) or electrolytic tanks (for two- or three-dimensional fields). When they were applied to steady-state spatial problems, these analog computers were called *field plotters* or *potential analyzers*.

The ability to fabricate very large arrays of analog computing devices, combined with the need to exploit massive parallelism in realtime computation and control applications, creates new opportunities for field computation [37,38,43]. There is also renewed interest in using physical fields in analog computation. For example, Rubel [73] defined an abstract *extended analog computer* (EAC), which augments Shannon's [77] general purpose analog computer with (unspecified) facilities for field computation, such as PDE solvers (see Sects. Shannon's Analysis – Rubel's Extended Analog Computer below). JW. Mills has explored the practical application of these ideas in his *artificial neural field networks* and VLSI EACs, which use the diffusion of electrons in bulk silicon or conductive gels and plastics for 2D and 3D field computation [53,54].

Computational Process

We have considered the continuous state space, which is the basis for analog computing, but there are a variety of ways in which analog computers can operate on the state. In particular, the state can change continuously in time or be updated at distinct instants (as in digital computation).

Continuous Time Since the laws of physics on which analog computing is based are differential equations, many analog computations proceed in continuous real time. Also, as we have seen, an important application of analog computers in the late 19th and early 20th centuries was the integration of ODEs in which time is the independent variable. A common technique in analog simulation of physical systems is *time scaling*, in which the differential equations are altered systematically so the simulation proceeds either more slowly or more quickly than the primary system (see Sect. "Characteristics of Analog Computation" for more on time scaling). On the other hand, because analog computations are close to the physical processes that

realize them, analog computing is rapid, which makes it very suitable for real-time control applications.

In principle, any mathematically describable physical process operating on time-varying physical quantities can be used for analog computation. In practice, however, analog computers typically provide familiar operations that scientists and engineers use in differential equations [70,87]. These include basic arithmetic operations, such as algebraic sum and difference ($u(t) = v(t) \pm w(t)$), constant multiplication or scaling ($u(t) = cv(t)$), variable multiplication and division ($u(t) = v(t)w(t)$, $u(t) = v(t)/w(t)$), and inversion ($u(t) = -v(t)$). Transcendental functions may be provided, such as the exponential ($u(t) = \exp v(t)$), logarithm ($u(t) = \ln v(t)$), trigonometric functions ($u(t) = \sin v(t)$, etc.), and *resolvers* for converting between polar and rectangular coordinates. Most important, of course, is definite integration ($u(t) = v_0 + \int_0^t v(\tau)\,\mathrm{d}\tau$), but differentiation may also be provided ($u(t) = \dot{v}(t)$). Generally, however, direct differentiation is avoided, since noise tends to have a higher frequency than the signal, and therefore differentiation amplifies noise; typically problems are reformulated to avoid direct differentiation [pp. 26–27 in 89]. As previously mentioned, many GPACs include *(arbitrary) function generators*, which allow the use of functions defined only by a graph and for which no mathematical definition might be available; in this way empirically defined functions can be used [pp. 32–42 in 70]. Thus, given a graph $(x, f(x))$, or a sufficient set of samples, $(x_k, f(x_k))$, the function generator approximates $u(t) = f(v(t))$. Rather less common are generators for arbitrary functions of two variables, $u(t) = f(v(t), w(t))$, in which the function may be defined by a surface, $(x, y, f(x, y))$, or by sufficient samples from it.

Although analog computing is primarily continuous, there are situations in which discontinuous behavior is required. Therefore some analog computers provide *comparators*, which produce a discontinuous result depending on the relative value of two input values. For example,

$$u = \begin{cases} k & \text{if } v \geq w\,, \\ 0 & \text{if } v < w\,. \end{cases}$$

Typically, this would be implemented as a Heaviside (unit step) function applied to the difference of the inputs, $w = kU(v - w)$. In addition to allowing the definition of discontinuous functions, comparators provide a primitive decision making ability, and may be used, for example to terminate a computation (switching the computer from "operate" to "hold" mode).

Other operations that have proved useful in analog computation are time delays and noise generators [Chap. 7 in 31]. The function of a *time delay* is simply to retard the signal by an adjustable delay $T > 0$: $u(t + T) = v(t)$. One common application is to model delays in the primary system (e. g., human response time).

Typically a *noise generator* produces time-invariant Gaussian-distributed noise with zero mean and a flat power spectrum (over a band compatible with the analog computing process). The standard deviation can be adjusted by scaling, the mean can be shifted by addition, and the spectrum altered by filtering, as required by the application. Historically noise generators were used to model noise and other random effects in the primary system, to determine, for example, its sensitivity to effects such as turbulence. However, noise can make a positive contribution in some analog computing algorithms (e. g., for symmetry breaking and in simulated annealing, weight perturbation learning, and stochastic resonance).

As already mentioned, some analog computing devices for the direct solution of PDEs have been developed. In general a PDE solver depends on an analogous physical process, that is, on a process obeying the same class of PDEs that it is intended to solve. For example, in Mill's EAC, diffusion of electrons in conductive sheets or solids is used to solve diffusion equations [53,54]. Historically, PDEs were solved on electronic GPACs by discretizing all but one of the independent variables, thus replacing the differential equations by difference equations [70], pp. 173–193. That is, computation over a field was approximated by computation over a finite real array.

Reaction-diffusion computation is an important example of continuous-time analog computing ► Reaction-Diffusion Computing. The state is represented by a set of time-varying chemical concentration fields, $c_1, \ldots, c_n$. These fields are distributed across a one-, two-, or three-dimensional space Ω, so that, for $\mathbf{x} \in \Omega$, $c_k(\mathbf{x}, t)$ represents the concentration of chemical k at location $\mathbf{x}$ and time t. Computation proceeds in continuous time according to reaction-diffusion equations, which have the form:

$$\partial \mathbf{c}/\partial t = \mathrm{D}\nabla^2 \mathbf{c} + \mathbf{F}(\mathbf{c})\,,$$

where $\mathbf{c} = (c_1, \ldots, c_n)^{\mathrm{T}}$ is the vector of concentrations, $\mathrm{D} = \mathrm{diag}(d_1, \ldots, d_n)$ is a diagonal matrix of positive diffusion rates, and $\mathbf{F}$ is nonlinear vector function that describes how the chemical reactions affect the concentrations.

Some neural net models operate in continuous time and thus are examples of continuous-time analog computation. For example, Grossberg [26,27,28] defines the activity of a neuron by differential equations such as this:

$$\dot{x}_i = -a_i x_i + \sum_{j=1}^{n} b_{ij} w_{ij}^{(+)} f_j(x_j) - \sum_{j=1}^{n} c_{ij} w_{ij}^{(-)} g_j(x_j) + I_i \,.$$

This describes the continuous change in the activity of neuron i resulting from passive decay (first term), positive feedback from other neurons (second term), negative feedback (third term), and input (last term). The f_j and g_j are nonlinear activation functions, and the $w_{ij}^{(+)}$ and $w_{ij}^{(-)}$ are adaptable excitatory and inhibitory connection strengths, respectively.

The continuous Hopfield network is another example of continuous-time analog computation [30]. The output y_i of a neuron is a nonlinear function of its internal state x_i, $y_i = \sigma(x_i)$, where the hyperbolic tangent is usually used as the activation function, $\sigma(x) = \tanh x$, because its range is $[-1, 1]$. The internal state is defined by a differential equation,

$$\tau_i \dot{x}_i = -a_i x_i + b_i + \sum_{j=1}^{n} w_{ij} y_j \,,$$

where τ_i is a time constant, a_i is the decay rate, b_i is the bias, and w_{ij} is the connection weight to neuron i from neuron j. In a Hopfield network every neuron is symmetrically connected to every other ($w_{ij} = w_{ji}$) but not to itself ($w_{ii} = 0$).

Of course analog VLSI implementations of neural networks also operate in continuous time (e. g., [20,52].

Concurrent with the resurgence of interest in analog computation have been innovative reconceptualizations of continuous-time computation. For example, Brockett [12] has shown that dynamical systems can perform a number of problems normally considered to be intrinsically sequential. In particular, a certain system of ODEs (a *nonperiodic finite Toda lattice*) can sort a list of numbers by continuous-time analog computation. The system is started with the vector $\mathbf{x}$ equal to the values to be sorted and a vector $\mathbf{y}$ initialized to small nonzero values; the $\mathbf{y}$ vector converges to a sorted permutation of $\mathbf{x}$.

Sequential Time computation refers to computation in which discrete computational operations take place in succession but at no definite interval [88]. Ordinary digital computer programs take place in sequential time, for the operations occur one after another, but the individual operations are not required to have any specific duration, so long as they take finite time.

One of the oldest examples of sequential analog computation is provided by the compass-and-straightedge constructions of traditional Euclidean geometry (Sect. "Introduction"). These computations proceed by a sequence of discrete operations, but the individual operations involve continuous representations (e. g., compass settings, straightedge positions) and operate on a continuous state (the figure under construction). Slide rule calculation might seem to be an example of sequential analog computation, but if we look at it, we see that although the operations are performed by an analog device, the intermediate results are recorded digitally (and so this part of the state space is discrete). Thus it is a kind of hybrid computation.

The familiar digital computer automates sequential digital computations that once were performed manually by human "computers". Sequential analog computation can be similarly automated. That is, just as the control unit of an ordinary digital computer sequences digital computations, so a digital control unit can sequence analog computations. In addition to the analog computation devices (adders, multipliers, etc.), such a computer must provide variables and registers capable of holding continuous quantities between the sequential steps of the computation (see also Sect. "Discrete Time" below).

The primitive operations of sequential-time analog computation are typically similar to those in continuous-time computation (e. g., addition, multiplication, transcendental functions), but integration and differentiation with respect to sequential time do not make sense. However, continuous-time integration within a single step, and space-domain integration, as in PDE solvers or field computation devices, are compatible with sequential analog computation.

In general, any model of digital computation can be converted to a similar model of sequential analog computation by changing the discrete state space to a continuum, and making appropriate changes to the rest of the model. For example, we can make an analog Turing machine by allowing it to write a bounded real number (rather than a symbol from a finite alphabet) onto a tape cell. The Turing machine's finite control can be altered to test for tape markings in some specified range.

Similarly, in a series of publications Blum, Shub, and Smale developed a theory of computation over the reals, which is an abstract model of sequential-time analog computation [6,7]. In this "BSS model" programs are represented as flowcharts, but they are able to operate on real-valued variables. Using this model they were able to prove a number of theorems about the complexity of sequential analog algorithms.

The BSS model, and some other sequential analog computation models, assume that it is possible to make exact comparisons between real numbers (analogous to exact comparisons between integers or discrete symbols in digital computation) and to use the result of the comparison

to control the path of execution. Comparisons of this kind are problematic because they imply infinite precision in the comparator (which may be defensible in a mathematical model but is impossible in physical analog devices), and because they make the execution path a discontinuous function of the state (whereas analog computation is usually continuous). Indeed, it has been argued that this is not "true" analog computation [p. 148 in 78].

Many artificial neural network models are examples of sequential-time analog computation. In a simple feed-forward neural network, an input vector is processed by the layers in order, as in a pipeline. That is, the output of layer n becomes the input of layer $n + 1$. Since the model does not make any assumptions about the amount of time it takes a vector to be processed by each layer and to propagate to the next, execution takes place in sequential time. Most *recurrent* neural networks, which have feedback, also operate in sequential time, since the activities of all the neurons are updated synchronously (that is, signals propagate through the layers, or back to earlier layers, in lock-step).

Many artificial neural-net learning algorithms are also sequential-time analog computations. For example, the back-propagation algorithm updates a network's weights, moving sequentially backward through the layers.

In summary, the correctness of sequential time computation (analog or digital) depends on the *order* of operations, not on their *duration*, and similarly the efficiency of sequential computations is evaluated in terms of the *number* of operations, not on their *total duration*.

Discrete Time analog computation has similarities to both continuous-time and sequential analog computation. Like the latter, it proceeds by a sequence of discrete (analog) computation steps; like the former, these steps occur at a constant rate in real time (e. g., some "frame rate"). If the real-time rate is sufficient for the application, then discrete-time computation can approximate continuous-time computation (including integration and differentiation).

Some electronic GPACs implemented discrete-time analog computation by a modification of repetitive operation mode, called *iterative analog computation* [Chap. 9 in 2]. Recall (Sect. "Electronic Analog Computation in the 20th Century") that in repetitive operation mode a clock rapidly switched the computer between reset and compute modes, thus repeating the same analog computation, but with different parameters (set by the operator). However, each repetition was independent of the others. Iterative operation was different in that analog values computed by one iteration could be used as initial values in the next. This was accomplished by means of an analog memory circuit (based on an op amp) that sampled an analog value at the end of one compute cycle (effectively during hold mode) and used it to initialize an integrator during the following reset cycle. (A modified version of the memory circuit could be used to retain a value over several iterations.) Iterative computation was used for problems such as determining, by iterative search or refinement, the initial conditions that would lead to a desired state at a future time. Since the analog computations were iterated at a fixed clock rate, iterative operation is an example of discrete-time analog computation. However, the clock rate is not directly relevant in some applications (such as the iterative solution of boundary value problems), in which case iterative operation is better characterized as sequential analog computation.

The principal contemporary examples of discrete-time analog computing are in neural network applications to time-series analysis and (discrete-time) control. In each of these cases the input to the neural net is a sequence of discrete-time samples, which propagate through the net and generate discrete-time output signals. Many of these neural nets are recurrent, that is, values from later layers are fed back into earlier layers, which allows the net to remember information from one sample to the next.

Analog Computer Programs

The concept of a *program* is central to digital computing, both practically, for it is the means for programming general-purpose digital computers, and theoretically, for it defines the limits of what can be computed by a universal machine, such as a universal Turing machine. Therefore it is important to discuss means for describing or specifying analog computations.

Traditionally, analog computers were used to solve ODEs (and sometimes PDEs), and so in one sense a mathematical differential equation is one way to represent an analog computation. However, since the equations were usually not suitable for direct solution on an analog computer, the process of *programming* involved the translation of the equations into a schematic diagram showing how the analog computing devices (integrators etc.) should be connected to solve the problem. These diagrams are the closest analogies to digital computer programs and may be compared to flowcharts, which were once popular in digital computer programming. It is worth noting, however, that flowcharts (and ordinary computer programs) represent sequences among operations, whereas analog computing diagrams represent functional relationships among variables, and therefore a kind of parallel data flow.

Differential equations and schematic diagrams are suitable for continuous-time computation, but for sequential analog computation something more akin to a conventional digital program can be used. Thus, as previously discussed (Sect. "Sequential Time"), the BSS system uses flowcharts to describe sequential computations over the reals. Similarly, C. Moore [55] defines recursive functions over the reals by means of a notation similar to a programming language.

In principle any sort of analog computation might involve constants that are arbitrary real numbers, which therefore might not be expressible in finite form (e. g., as a finite string of digits). Although this is of theoretical interest (see Sect. "Real-valued Inputs, Outputs, and Constants" below), from a practical standpoint these constants could be set with about at most four digits of precision [p. 11 in 70]. Indeed, automatic potentiometer-setting devices were constructed that read a series of decimal numerals from punched paper tape and used them to set the potentiometers for the constants [pp. 3-58–60 in 87]. Nevertheless it is worth observing that analog computers do allow continuous inputs that need not be expressed in digital notation, for example, when the parameters of a simulation are continuously varied by the operator. In principle, therefore, an analog program can incorporate constants that are represented by a real-valued physical quantity (e. g., an angle or a distance), which need not be expressed digitally. Further, as we have seen (Sect. "Electronic Analog Computation in the 20th Century"), some electronic analog computers could compute a function by means of an arbitrarily drawn curve, that is, not represented by an equation or a finite set of digitized points. Therefore, in the context of analog computing it is natural to expand the concept of a program beyond discrete symbols to include continuous representations (scalar magnitudes, vectors, curves, shapes, surfaces, etc.).

Typically such continuous representations would be used as adjuncts to conventional discrete representations of the analog computational process, such as equations or diagrams. However, in some cases the most natural static representation of the process is itself continuous, in which case it is more like a "guiding image" than a textual prescription [42]. A simple example is a potential surface, which defines a continuum of trajectories from initial states (possible inputs) to fixed-point attractors (the results of the computations). Such a "program" may define a deterministic computation (e. g., if the computation proceeds by gradient descent), or it may constrain a nondeterministic computation (e. g., if the computation may proceed by any potential-decreasing trajectory). Thus analog computation suggests a broadened notion of programs and programming.

Characteristics of Analog Computation

Precision Analog computation is evaluated in terms of both accuracy and precision, but the two must be distinguished carefully [pp. 25–28 in 2,pp. 12–13 in 89,pp. 257–261 in 82]. *Accuracy* refers primarily to the relationship between a simulation and the primary system it is simulating or, more generally, to the relationship between the results of a computation and the mathematically correct result. Accuracy is a result of many factors, including the mathematical model chosen, the way it is set up on a computer, and the precision of the analog computing devices. *Precision*, therefore, is a narrower notion, which refers to the quality of a representation or computing device. In analog computing, precision depends on *resolution* (fineness of operation) and *stability* (absence of drift), and may be measured as a fraction of the represented value. Thus a precision of 0.01% means that the representation will stay within 0.01% of the represented value for a reasonable period of time. For purposes of comparing analog devices, the precision is usually expressed as a fraction of *full-scale variation* (i. e., the difference between the maximum and minimum representable values).

It is apparent that the precision of analog computing devices depends on many factors. One is the choice of physical process and the way it is utilized in the device. For example a linear mathematical operation can be realized by using a linear region of a nonlinear physical process, but the realization will be approximate and have some inherent imprecision. Also, associated, unavoidable physical effects (e. g., loading, and leakage and other losses) may prevent precise implementation of an intended mathematical function. Further, there are fundamental physical limitations to resolution (e. g., quantum effects, diffraction). Noise is inevitable, both intrinsic (e. g., thermal noise) and extrinsic (e. g., ambient radiation). Changes in ambient physical conditions, such as temperature, can affect the physical processes and decrease precision. At slower time scales, materials and components age and their physical characteristics change. In addition, there are always technical and economic limits to the control of components, materials, and processes in analog device fabrication.

The precision of analog and digital computing devices depend on very different factors. The precision of a (binary) digital device depends on the number of bits, which influences the amount of hardware, but not its quality. For example, a 64-bit adder is about twice the size of a 32-bit adder, but can made out of the same components. At worst, the size of a digital device might increase with the

square of the number of bits of precision. This is because binary digital devices only need to represent two states, and therefore they can operate in saturation. The fabrication standards sufficient for the first bit of precision are also sufficient for the 64th bit. Analog devices, in contrast, need to be able to represent a continuum of states precisely. Therefore, the fabrication of high-precision analog devices is much more expensive than low-precision devices, since the quality of components, materials, and processes must be much more carefully controlled. Doubling the precision of an analog device may be expensive, whereas the cost of each additional bit of digital precision is incremental; that is, the cost is proportional to the logarithm of the precision expressed as a fraction of full range.

The forgoing considerations might seem to be a convincing argument for the superiority of digital to analog technology, and indeed they were an important factor in the competition between analog and digital computers in the middle of the twentieth century [pp. 257–261 in 82]. However, as was argued at that time, many computer applications do not require high precision. Indeed, in many engineering applications, the input data are known to only a few digits, and the equations may be approximate or derived from experiments. In these cases the very high precision of digital computation is unnecessary and may in fact be misleading (e. g., if one displays all 14 digits of a result that is accurate to only three). Furthermore, many applications in image processing and control do not require high precision. More recently, research in artificial neural networks (ANNs) has shown that low-precision analog computation is sufficient for almost all ANN applications. Indeed, neural information processing in the brain seems to operate with very low precision (perhaps less than 10% [p. 378 in 50]), for which it compensates with massive parallelism. For example, by *coarse coding* a population of low-precision devices can represent information with relatively high precision [pp. 91–96 in 74,75].

Scaling An important aspect of analog computing is *scaling*, which is used to adjust a problem to an analog computer. First is *time scaling*, which adjusts a problem to the characteristic time scale at which a computer operates, which is a consequence of its design and the physical processes by which it is realized [pp. 37–44 in 62,pp. 262–263 in 70,pp. 241–243 in 89]. For example, we might want a simulation to proceed on a very different time scale from the primary system. Thus a weather or economic simulation should proceed faster than real time in order to get useful predictions. Conversely, we might want to slow down a simulation of protein folding so that we can observe the stages in the process. Also, for accurate results it is necessary to avoid exceeding the maximum response rate of the analog devices, which might dictate a slower simulation speed. On the other hand, too slow a computation might be inaccurate as a consequence of instability (e. g., drift and leakage in the integrators).

Time scaling affects only time-dependent operations such as integration. For example, suppose t, time in the primary system or "problem time", is related to τ, time in the computer, by $\tau = \beta t$. Therefore, an integration $u(t) = \int_0^t v(t')\mathrm{d}t'$ in the primary system is replaced by the integration $u(\tau) = \beta^{-1} \int_0^\tau v(\tau')\mathrm{d}\tau'$ on the computer. Thus time scaling may be accomplished simply by decreasing the input gain to the integrator by a factor of β.

Fundamental to analog computation is the representation of a continuous quantity in the primary system by a continuous quantity in the computer. For example, a displacement x in meters might be represented by a potential V in volts. The two are related by an *amplitude* or *magnitude scale factor*, $V = \alpha x$, (with units volts/meter), chosen to meet two criteria [pp. 103–106 in 2,Chap. 4 in 62,pp. 127–128 in 70,pp. 233–240 in 89]. On the one hand, α must be sufficiently small so that the range of the problem variable is accommodated within the range of values supported by the computing device. Exceeding the device's intended operating range may lead to inaccurate results (e. g., forcing a linear device into nonlinear behavior). On the other hand, the scale factor should not be too small, or relevant variation in the problem variable will be less than the resolution of the device, also leading to inaccuracy. (Recall that precision is specified as a fraction of full-range variation.)

In addition to the explicit variables of the primary system, there are implicit variables, such as the time derivatives of the explicit variables, and scale factors must be chosen for them too. For example, in addition to displacement x, a problem might include velocity $\dot{x}$ and acceleration $\ddot{x}$. Therefore, scale factors α, α', and α'' must be chosen so that αx, $\alpha' \dot{x}$, and $\alpha'' \ddot{x}$ have an appropriate range of variation (neither too large nor too small).

Once a scale factor has been chosen, the primary system equations are adjusted to obtain the analog computing equations. For example, if we have scaled $u = \alpha x$ and $v = \alpha' \dot{x}$, then the integration $x(t) = \int_0^t \dot{x}(t')\mathrm{d}t'$ would be computed by scaled equation:

$$u(t) = \frac{\alpha}{\alpha'} \int_0^t v(t')\mathrm{d}t'.$$

This is accomplished by simply setting the input gain of the integrator to α/α'.

In practice, time scaling and magnitude scaling are not independent [p. 262 in 70]. For example, if the derivatives

of a variable can be large, then the variable can change rapidly, and so it may be necessary to slow down the computation to avoid exceeding the high-frequency response of the computer. Conversely, small derivatives might require the computation to be run faster to avoid integrator leakage etc. Appropriate scale factors are determined by considering both the physics and the mathematics of the problem [pp. 40–44 in 62]. That is, first, the physics of the primary system may limit the ranges of the variables and their derivatives. Second, analysis of the mathematical equations describing the system can give additional information on the ranges of the variables. For example, in some cases the natural frequency of a system can be estimated from the coefficients of the differential equations; the maximum of the nth derivative is then estimated as the n power of this frequency [p. 42 in 62,pp. 238–240 in 89]. In any case, it is not necessary to have accurate values for the ranges; rough estimates giving orders of magnitude are adequate.

It is tempting to think of magnitude scaling as a problem unique to analog computing, but before the invention of floating-point numbers it was also necessary in digital computer programming. In any case it is an essential aspect of analog computing, in which physical processes are more directly used for computation than they are in digital computing. Although the necessity of scaling has been a source of criticism, advocates for analog computing have argued that it is a blessing in disguise, because it leads to improved understanding of the primary system, which was often the goal of the computation in the first place [5,Chap. 8 in 82]. Practitioners of analog computing are more likely to have an intuitive understanding of both the primary system and its mathematical description (see Sect. "Analog Thinking").

Analog Computation in Nature

Computational processes—that is to say, information processing and control—occur in many living systems, most obviously in nervous systems, but also in the self-organized behavior of groups of organisms. In most cases natural computation is analog, either because it makes use of continuous natural processes, or because it makes use of discrete but stochastic processes. Several examples will be considered briefly.

Neural Computation

In the past neurons were thought of binary computing devices, something like digital logic gates. This was a consequence of the "all or nothing" response of a neuron, which refers to the fact that it does or does not generate an *action potential* (voltage spike) depending, respectively, on whether its total input exceeds a threshold or not (more accurately, it generates an action potential if the membrane depolarization at the axon hillock exceeds the threshold and the neuron is not in its refractory period). Certainly some neurons (e. g., so-called "command neurons") do act something like logic gates. However, most neurons are analyzed better as analog devices, because the *rate* of impulse generation represents significant information. In particular, an *amplitude code*, the membrane potential near the axon hillock (which is a summation of the electrical influences on the neuron), is translated into a *rate code* for more reliable long-distance transmission along the axons. Nevertheless, the code is low precision (about one digit), since information theory shows that it takes at least N milliseconds (and probably more like $5N$ ms) to discriminate N values [39]. The rate code is translated back to an amplitude code by the synapses, since successive impulses release neurotransmitter from the axon terminal, which diffuses across the synaptic cleft to receptors. Thus a synapse acts as a leaky integrator to time-average the impulses.

As previously discussed (Sect. "Continuous State Space"), many artificial neural net models have real-valued neural activities, which correspond to rate-encoded axonal signals of biological neurons. On the other hand, these models typically treat the input connections as simple real-valued weights, which ignores the analog signal processing that takes place in the dendritic trees of biological neurons. The dendritic trees of many neurons are complex structures, which often have thousand of synaptic inputs. The binding of neurotransmitters to receptors causes minute voltage fluctuations, which propagate along the membrane, and ultimately cause voltage fluctuations at the axon hillock, which influence the impulse rate. Since the dendrites have both resistance and capacitance, to a first approximation the signal propagation is described by the "cable equations", which describe passive signal propagation in cables of specified diameter, capacitance, and resistance [Chap. 1 in 1]. Therefore, to a first approximation, a neuron's dendritic net operates as an adaptive linear analog filter with thousands of inputs, and so it is capable of quite complex signal processing. More accurately, however, it must be treated as a *nonlinear* analog filter, since voltage-gated ion channels introduce nonlinear effects. The extent of analog signal processing in dendritic trees is still poorly understood.

In most cases, then, neural information processing is treated best as low-precision analog computation. Although individual neurons have quite broadly tuned responses, accuracy in perception and sensorimotor control is achieved through coarse coding, as already discussed

(Sect. "Characteristics of Analog Computation"). Further, one widely used neural representation is the *cortical map*, in which neurons are systematically arranged in accord with one or more dimensions of their stimulus space, so that stimuli are represented by patterns of activity over the map. (Examples are *tonotopic maps*, in which pitch is mapped to cortical location, and *retinotopic maps*, in which cortical location represents retinal location.) Since neural density in the cortex is at least 146 000 neurons per square millimeter [p. 51 in 14], even relatively small cortical maps can be treated as fields and information processing in them as analog field computation (see ► Field Computation in Natural and Artificial Intelligence). Overall, the brain demonstrates what can be accomplished by massively parallel analog computation, even if the individual devices are comparatively slow and of low precision.

Adaptive Self-Organization in Social Insects

Another example of analog computation in nature is provided by the self-organizing behavior of social insects, microorganisms, and other populations [13]. Often such organisms respond to concentrations, or gradients in the concentrations, of chemicals produced by other members of the population. These chemicals may be deposited and diffuse through the environment. In other cases, insects and other organisms communicate by contact, but may maintain estimates of the relative proportions of different kinds of contacts. Because the quantities are effectively continuous, all these are examples of analog control and computation.

Self-organizing populations provide many informative examples of the use of natural processes for analog information processing and control. For example, diffusion of pheromones is a common means of self-organization in insect colonies, facilitating the creation of paths to resources, the construction of nests, and many other functions [13]. Real diffusion (as opposed to sequential simulations of it) executes, in effect, a massively parallel search of paths from the chemical's source to its recipients and allows the identification of near-optimal paths. Furthermore, if the chemical degrades, as is generally the case, then the system will be adaptive, in effect continually searching out the shortest paths, so long as source continues to function [13]. Simulated diffusion has been applied to robot path planning [32,69].

Genetic Circuits

Another example of natural analog computing is provided by the *genetic regulatory networks* that control the behavior of cells, in multicellular organisms as well as single-celled ones [16]. These networks are defined by the mutually interdependent regulatory genes, promoters, and repressors that control the internal and external behavior of a cell. The interdependencies are mediated by proteins, the synthesis of which is governed by genes, and which in turn regulate the synthesis of other gene products (or themselves). Since it is the quantities of these substances that is relevant, many of the regulatory motifs can be described in computational terms as adders, subtracters, integrators, etc. Thus the genetic regulatory network implements an analog control system for the cell [68].

It might be argued that the number of intracellular molecules of a particular protein is a (relatively small) discrete number, and therefore that it is inaccurate to treat it as a continuous quantity. However, the molecular processes in the cell are stochastic, and so the relevant quantity is the *probability* that a regulatory protein will bind to a regulatory site. Further, the processes take place in continuous real time, and so the rates are generally the significant quantities. Finally, although in some cases gene activity is either on or off (more accurately: very low), in other cases it varies continuously between these extremes [pp. 388–390 in 29].

Embryological development combines the analog control of individual cells with the sort of self-organization of populations seen in social insects and other colonial organisms. Locomotion of the cells and the expression of specific genes is controlled by chemical signals, among other mechanisms [16,17]. Thus PDEs have proved useful in explaining some aspects of development; for example *reaction-diffusion equations* have been used to describe the formation of hair-coat patterns and related phenomena [13,48,57]; see ► Reaction-Diffusion Computing. Therefore the developmental process is governed by naturally occurring analog computation.

Is Everything a Computer?

It might seem that any continuous physical process could be viewed as analog computation, which would make the term almost meaningless. As the question has been put, is it meaningful (or useful) to say that the solar system is *computing* Kepler's laws? In fact, it is possible and worthwhile to make a distinction between computation and other physical processes that happen to be described by mathematical laws [40,41,44,46].

If we recall the original meaning of analog computation (Sect. "Definition of the Subject"), we see that the computational system is used to solve some mathematical problem with respect to a primary system. What makes this possible is that the computational system and the pri-

mary system have the same, or systematically related, abstract (mathematical) structures. Thus the computational system can inform us about the primary system, or be used to control it, etc. Although from a practical standpoint some analogs are better than others, in principle any physical system can be used that obeys the same equations as the primary system.

Based on these considerations we may define computation as a physical process the purpose of which is the abstract manipulation of abstract objects (i. e., information processing); this definition applies to analog, digital, and hybrid computation [40,41,44,46]. Therefore, to determine if a natural system is computational we need to look to its purpose or function within the context of the living system of which it is a part. One test of whether its function is the abstract manipulation of abstract objects is to ask whether it could still fulfill its function if realized by different physical processes, a property called *multiple realizability*. (Similarly, in artificial systems, a simulation of the economy might be realized equally accurately by a hydraulic analog computer or an electronic analog computer [5].) By this standard, the majority of the nervous system is purely computational; in principle it could be replaced by electronic devices obeying the same differential equations. In the other cases we have considered (self-organization of living populations, genetic circuits) there are instances of both pure computation and computation mixed with other functions (for example, where the specific substances used have other – e. g. metabolic – roles in the living system).

General-Purpose Analog Computation

The Importance of General-Purpose Computers

Although special-purpose analog and digital computers have been developed, and continue to be developed, for many purposes, the importance of general-purpose computers, which can be adapted easily for a wide variety of purposes, has been recognized since at least the nineteenth century. Babbage's plans for a general-purpose digital computer, his *analytical engine* (1835), are well known, but a general-purpose differential analyzer was advocated by Kelvin [84]. Practical general-purpose analog and digital computers were first developed at about the same time: from the early 1930s through the war years. General-purpose computers of both kinds permit the prototyping of special-purpose computers and, more importantly, permit the flexible reuse of computer hardware for different or evolving purposes.

The concept of a general-purpose computer is useful also for determining the limits of a computing paradigm. If one can design—theoretically or practically—a *universal computer*, that is, a general-purpose computer capable of simulating any computer in a relevant class, then anything uncomputable by the universal computer will also be uncomputable by any computer in that class. This is, of course, the approach used to show that certain functions are uncomputable by any Turing machine because they are uncomputable by a universal Turing machine. For the same reason, the concept of general-purpose analog computers, and in particular of *universal analog computers* are theoretically important for establishing limits to analog computation.

General-Purpose Electronic Analog Computers

Before taking up these theoretical issues, it is worth recalling that a typical electronic GPAC would include linear elements, such as adders, subtracters, constant multipliers, integrators, and differentiators; nonlinear elements, such as variable multipliers and function generators; other computational elements, such as comparators, noise generators, and delay elements (Sect. "Electronic Analog Computation in the 20th Century"). These are, of course, in addition to input/output devices, which would not affect its computational abilities.

Shannon's Analysis

Claude Shannon did an important analysis of the computational capabilities of the differential analyzer, which applies to many GPACs [76,77]. He considered an abstract differential analyzer equipped with an unlimited number of integrators, adders, constant multipliers, and function generators (for functions with only a finite number of finite discontinuities), with at most one source of drive (which limits possible interconnections between units). This was based on prior work that had shown that almost all the generally used elementary functions could be generated with addition and integration. We will summarize informally a few of Shannon's results; for details, please consult the original paper.

First Shannon offers proofs that, by setting up the correct ODEs, a GPAC with the mentioned facilities can generate any function if and only if is not hypertranscendental (Theorem II); thus the GPAC can generate any function that is algebraic transcendental (a very large class), but not, for example, Euler's gamma function and Riemann's zeta function. He also shows that the GPAC can generate functions derived from generable functions, such as the integrals, derivatives, inverses, and compositions of generable functions (Thms. III, IV). These results can be generalized to functions of any number of variables, and to their com-

positions, partial derivatives, and inverses with respect to any one variable (Thms. VI, VII, IX, X).

Next Shannon shows that a function of any number of variables that is continuous over a closed region of space can be approximated arbitrarily closely over that region with a finite number of adders and integrators (Thms. V, VIII).

Shannon then turns from the generation of functions to the solution of ODEs and shows that the GPAC can solve any system of ODEs defined in terms of non-hypertranscendental functions (Thm. XI).

Finally, Shannon addresses a question that might seem of limited interest, but turns out to be relevant to the computational power of analog computers (see Sect. "Analog Computation and the Turing Limit" below). To understand it we must recall that he was investigating the differential analyzer—a mechanical analog computer—but similar issues arise in other analog computing technologies. The question is whether it is possible to perform an arbitrary constant multiplication, $u = kv$, by means of gear ratios. He show that if we have just two gear ratios a and $b(a, b \neq 0, 1)$, such that b is not a rational power of a, then by combinations of these gears we can approximate k arbitrarily closely (Thm. XII). That is, to approximate multiplication by arbitrary real numbers, it is sufficient to be able to multiply by a, b, and their inverses, provided a and b are not related by a rational power.

Shannon mentions an alternative method of constant multiplication, which uses integration, $kv = \int_0^v k\mathrm{d}v$, but this requires setting the integrand to the constant function k. Therefore, multiplying by an arbitrary real number requires the ability to input an arbitrary real as the integrand. The issue of real-valued inputs and outputs to analog computers is relevant both to their theoretical power and to practical matters of their application (see Sect. "Real-valued Inputs, Output, and Constants").

Shannon's proofs, which were incomplete, were eventually refined by M. Pour-El [63] and finally corrected by L. Lipshitz and L.A. Rubel [35]. Rubel [72] proved that Shannon's GPAC cannot solve the Dirichlet problem for Laplace's equation on the disk; indeed, it is limited to initial-value problems for algebraic ODEs. Specifically, *the Shannon–Pour-El Thesis* is that the outputs of the GPAC are exactly the solutions of the *algebraic differential equations*, that is, equations of the form

$$P[x, y(x), y'(x), y''(x), \ldots, y^{(n)}(x)] = 0 ,$$

where P is a polynomial that is not identically vanishing in any of its variables (these are the *differentially algebraic* functions) [71]. (For details please consult the cited papers.) The limitations of Shannon's GPAC motivated Rubel's definition of the Extended Analog Computer.

Rubel's Extended Analog Computer

The combination of Rubel's [71] conviction that the brain is an analog computer together with the limitations of Shannon's GPAC led him to propose the *Extended Analog Computer* (EAC) [73].

Like Shannon's GPAC (and the Turing machine), the EAC is a conceptual computer intended to facilitate theoretical investigation of the limits of a class of computers. The EAC extends the GPAC in a number of respects. For example, whereas the GPAC solves equations defined over a single variable (time), the EAC can generate functions over any finite number of real variables. Further, whereas the GPAC is restricted to initial-value problems for ODEs, the EAC solves both initial- and boundary-value problems for a variety of PDEs.

The EAC is structured into a series of levels, each more powerful than the ones below it, from which it accepts inputs. The inputs to the lowest level are a finite number of real variables ("settings"). At this level it operates on real polynomials, from which it is able to generate the differentially algebraic functions. The computing on each level is accomplished by conceptual analog devices, which include constant real-number generators, adders, multipliers, differentiators, "substituters" (for function composition), devices for analytic continuation, and inverters, which solve systems of equations defined over functions generated by the lower levels. Most characteristic of the EAC is the "boundary-value-problem box", which solves systems of PDEs and ODEs subject to boundary conditions and other constraints. The PDEs are defined in terms of functions generated by the lower levels. Such PDE solvers may seem implausible, and so it is important to recall field-computing devices for this purpose were implemented in some practical analog computers (see Sect. "History") and more recently in Mills' EAC [54]. As Rubel observed, PDE solvers could be implemented by physical processes that obey the same PDEs (heat equation, wave equation, etc.). (See also Sect. "Future Directions" below.)

Finally, the EAC is required to be "extremely well-posed", which means that each level is relatively insensitive to perturbations in its inputs; thus "all the outputs depend in a strongly deterministic and stable way on the initial settings of the machine" [73].

Rubel [73] proves that the EAC can compute everything that the GPAC can compute, but also such functions as the gamma and zeta, and that it can solve the Dirichlet problem for Laplace's equation on the disk, all of which

are beyond the GPAC's capabilities. Further, whereas the GPAC can compute differentially algebraic functions of time, the EAC can compute differentially algebraic functions of any finite number of real variables. In fact, Rubel did not find any real-analytic (C^∞) function that is *not* computable on the EAC, but he observes that if the EAC can indeed generate every real-analytic function, it would be too broad to be useful as a model of analog computation.

Analog Computation and the Turing Limit

Introduction

The Church–Turing Thesis asserts that anything that is effectively computable is computable by a Turing machine, but the Turing machine (and equivalent models, such as the lambda calculus) are models of discrete computation, and so it is natural to wonder how analog computing compares in power, and in particular whether it can compute beyond the "Turing limit". Superficial answers are easy to obtain, but the issue is subtle because it depends upon choices among definitions, none of which is obviously correct, it involves the foundations of mathematics and its philosophy, and it raises epistemological issues about the role of models in scientific theories. Nevertheless this is an active research area, but many of the results are apparently inconsistent due to the differing assumptions on which they are based. Therefore this section will be limited to a mention of a few of the interesting results, but without attempting a comprehensive, systematic, or detailed survey; Siegelmann [78] can serve as an introduction to the literature.

A Sampling of Theoretical Results

Continuous-Time Models P. Orponen's [59] 1997 survey of continuous-time computation theory is a good introduction to the literature as of that time; here we give a sample of these and more recent results.

There are several results showing that—under various assumptions—analog computers have at least the power of Turing machines (TMs). For example, M.S. Branicky [11] showed that a TM could be simulated by ODEs, but he used non-differentiable functions; O. Bournez et al. [8] provide an alternative construction using only analytic functions. They also prove that the GPAC computability coincides with (Turing-)computable analysis, which is surprising, since the gamma function is Turing-computable but, as we have seen, the GPAC cannot generate it. The paradox is resolved by a distinction between *generating* a function and *computing* it, with the latter, broader notion permitting convergent computation of the function (that is, as $t \to \infty$). However, the computational power of general ODEs has not been determined in general [p. 149 in 78]. MB Pour-El and I Richards exhibit a Turing-computable ODE that does not have a Turing-computable solution [64,66]. M. Stannett [83] also defined a continuous-time analog computer that could solve the halting problem.

C. Moore [55] defines a class of continuous-time recursive functions over the reals, which includes a zero-finding operator μ. Functions can be classified into a hierarchy depending on the number of uses of μ, with the lowest level (no μs) corresponding approximately to Shannon's GPAC. Higher levels can compute non-Turing-computable functions, such as the decision procedure for the halting problem, but he questions whether this result is relevant in the physical world, which is constrained by "noise, quantum effects, finite accuracy, and limited resources". O. Bournez and M. Cosnard [9] have extended these results and shown that many dynamical systems have super-Turing power.

S. Omohundro [58] showed that a system of ten coupled nonlinear PDEs could simulate an arbitrary cellular automaton (see ► Mathematical Basis of Cellular Automata, Introduction to), which implies that PDEs have at least Turing power. Further, D. Wolpert and B.J. MacLennan [90,91] showed that any TM can be simulated by a field computer with linear dynamics, but the construction uses Dirac delta functions. Pour-El and Richards exhibit a wave equation in three-dimensional space with Turing-computable initial conditions, but for which the unique solution is Turing-uncomputable [65,66].

Sequential-Time Models We will mention a few of the results that have been obtained concerning the power of sequential-time analog computation.

Although the BSS model has been investigated extensively, its power has not been completely determined [6,7]. It is known to depend on whether just rational numbers or arbitrary real numbers are allowed in its programs [p. 148 in 78].

A *coupled map lattice* (CML) is a cellular automaton with real-valued states ► Mathematical Basis of Cellular Automata, Introduction to; it is a sequential-time analog computer, which can be considered a discrete-space approximation to a simple sequential-time field computer. P. Orponen and M. Matamala [60] showed that a finite CML can simulate a universal Turing machine. However, since a CML can simulate a BSS program or a recurrent neural network (see Sect. "Recurrent Neural Networks" below), it actually has super-Turing power [p. 149 in 78].

Recurrent neural networks are some of the most important examples of sequential analog computers, and so the following section is devoted to them.

Recurrent Neural Networks With the renewed interest in neural networks in the mid-1980s, may investigators wondered if recurrent neural nets have super-Turing power. M. Garzon and S. Franklin showed that a sequential-time net with a countable infinity of neurons could exceed Turing power [21,23,24]. Indeed, Siegelmann and E.D. Sontag [80] showed that finite neural nets with real-valued weights have super-Turing power, but W. Maass and Sontag [36] showed that recurrent nets with Gaussian or similar noise had *sub*-Turing power, illustrating again the dependence on these results on assumptions about what is a reasonable mathematical idealization of analog computing.

For recent results on recurrent neural networks, we will restrict our attention of the work of Siegelmann [78], who addresses the computational power of these networks in terms of the classes of languages they can recognize. Without loss of generality the languages are restricted to sets of binary strings. A string to be tested is fed to the network one bit at a time, along with an input that indicates when the end of the input string has been reached. The network is said to *decide* whether the string is in the language if it correctly indicates whether it is in the set or not, after some finite number of sequential steps since input began.

Siegelmann shows that, if exponential time is allowed for recognition, finite recurrent neural networks with real-valued weights (and saturated-linear activation functions) can compute *all* languages, and thus they are more powerful than Turing machines. Similarly, stochastic networks with rational weights also have super-Turing power, although less power than the deterministic nets with real weights. (Specifically, they compute $P/POLY$ and BPP/log^* respectively; see Siegelmann [Chaps. 4, 9 in 78] for details.) She further argues that these neural networks serve as a "standard model" of (sequential) analog computation (comparable to Turing machines in Church-Turing computation), and therefore that the limits and capabilities of these nets apply to sequential analog computation generally.

Siegelmann [p. 156 in 78] observes that the super-Turing power of recurrent neural networks is a consequence of their use of non-rational real-valued weights. In effect, a real number can contain an infinite number of bits of information. This raises the question of how the non-rational weights of a network can ever be set, since it is not possible to define a physical quantity with infinite precision. However, although non-rational weights may not be able to be set from outside the network, they can be computed within the network by learning algorithms, which are analog computations. Thus, Siegelmann suggests, the fundamental distinction may be between *static computational models*, such as the Turing machine and its equivalents, and *dynamically evolving computational models*, which can tune continuously variable parameters and thereby achieve super-Turing power.

Dissipative Models Beyond the issue of the power of analog computing relative to the Turing limit, there are also questions of its relative efficiency. For example, could analog computing solve NP-hard problems in polynomial or even linear time? In traditional computational complexity theory, efficiency issues are addressed in terms of the asymptotic number of computation steps to compute a function as the size of the function's input increases. One way to address corresponding issues in an analog context is by treating an analog computation as a *dissipative system*, which in this context means a system that decreases some quantity (analogous to energy) so that the system state converges to an *point attractor*. From this perspective, the initial state of the system incorporates the input to the computation, and the attractor represents its output. Therefore, HT Sieglemann, S Fishman, and A Ben-Hur have developed a complexity theory for dissipative systems, in both sequential and continuous time, which addresses the rate of convergence in terms of the underlying rates of the system [4,79]. The relation between dissipative complexity classes (e. g., P_d, NP_d) and corresponding classical complexity classes (P, NP) remains unclear [p. 151 in 78].

Real-Valued Inputs, Outputs, and Constants

A common argument, with relevance to the theoretical power of analog computation, is that an input to an analog computer must be determined by setting a dial to a number or by typing a number into digital-to-analog conversion device, and therefore that the input will be a rational number. The same argument applies to any internal constants in the analog computation. Similarly, it is argued, any output from an analog computer must be measured, and the accuracy of measurement is limited, so that the result will be a rational number. Therefore, it is claimed, real numbers are irrelevant to analog computing, since any practical analog computer computes a function from the rationals to the rationals, and can therefore be simulated by a Turing machine. (See related arguments by Martin Davis [18,19].)

There are a number of interrelated issues here, which may be considered briefly. First, the argument is couched in terms of the input or output of *digital representations*, and the numbers so represented are necessarily rational (more generally, computable). This seems natural enough when we think of an analog computer as a calculating device, and in fact many historical analog computers were used in this way and had digital inputs and outputs (since this is our most reliable way of recording and reproducing quantities).

However, in many analog *control systems*, the inputs and outputs are continuous physical quantities that vary continuously in time (also a continuous physical quantity); that is, according to current physical theory, these quantities are real numbers, which vary according to differential equations. It is worth recalling that physical quantities are neither rational nor irrational; they can be so classified only in comparison with each other or with respect to a unit, that is, only if they are measured and digitally represented. Furthermore, physical quantities are neither computable nor uncomputable (in a Church-Turing sense); these terms apply only to discrete representations of these quantities (i. e., to numerals or other digital representations).

Therefore, in accord with ordinary mathematical descriptions of physical processes, analog computations can can be treated as having arbitrary real numbers (in some range) as inputs, outputs, or internal states; like other continuous processes, continuous-time analog computations pass through all the reals in some range, including non-Turing-computable reals. Paradoxically, however, these same physical processes can be simulated on digital computers.

The Issue of Simulation by Turing Machines and Digital Computers

Theoretical results about the computational power, relative to Turing machines, of neural networks and other analog models of computation raise difficult issues, some of which are epistemological rather than strictly technical. On the one hand, we have a series of theoretical results proving the super-Turing power of analog computation models of various kinds. On the other hand, we have the obvious fact that neural nets are routinely simulated on ordinary digital computers, which have at most the power of Turing machines. Furthermore, it is reasonable to suppose that any physical process that might be used to realize analog computation—and certainly the known processes—could be simulated on a digital computer, as is done routinely in computational science. This would seem to be incontrovertible proof that analog computation is no more powerful than Turing machines. The crux of the paradox lies, of course, in the non-Turing-computable reals. These numbers are a familiar, accepted, and necessary part of standard mathematics, in which physical theory is formulated, but from the standpoint of Church-Turing (CT) computation they do not exist. This suggests that the the paradox is not a contradiction, but reflects a divergence between the goals and assumptions of the two models of computation.

The Problem of Models of Computation

These issues may be put in context by recalling that the Church-Turing (CT) model of computation is in fact a *model*, and therefore that it has the limitations of all models. A model is a cognitive tool that improves our ability to understand some class of phenomena by preserving relevant characteristics of the phenomena while altering other, irrelevant (or less relevant) characteristics. For example, a *scale model* alters the size (taken to be irrelevant) while preserving shape and other characteristics. Often a model achieves its purposes by making *simplifying* or *idealizing assumptions*, which facilitate analysis or simulation of the system. For example, we may use a linear mathematical model of a physical process that is only approximately linear. For a model to be effective it must preserve characteristics and make simplifying assumptions that are appropriate to the domain of questions it is intended to answer, its *frame of relevance* [46]. If a model is applied to problems outside of its frame of relevance, then it may give answers that are misleading or incorrect, because they depend more on the simplifying assumptions than on the phenomena being modeled. Therefore we must be especially cautious applying a model outside of its frame of relevance, or even at the limits of its frame, where the simplifying assumptions become progressively less appropriate. The problem is aggravated by the fact that often the frame of relevance is not explicit defined, but resides in a tacit background of practices and skills within some discipline.

Therefore, to determine the applicability of the CT model of computation to analog computing, we must consider the frame of relevance of the CT model. This is easiest if we recall the domain of issues and questions it was originally developed to address: issues of effective calculability and derivability in formalized mathematics. This frame of relevance determines many of the assumptions of the CT model, for example, that information is represented by finite discrete structures of symbols from a finite alphabet, that information processing proceeds by the application of definite formal rules at discrete instants of time, and

that a computational or derivational process must be completed in a finite number of these steps.[1] Many of these assumptions are incompatible with analog computing and with the frames of relevance of many models of analog computation.

Relevant Issues for Analog Computation

Analog computation is often used for control. Historically, analog computers were used in control systems and to simulate control systems, but contemporary analog VLSI is also frequently applied in control. Natural analog computation also frequently serves a control function, for example, sensorimotor control by the nervous system, genetic regulation in cells, and self-organized cooperation in insect colonies. Therefore, control systems provide one frame of relevance for models of analog computation.

In this frame of relevance real-time response is a critical issue, which models of analog computation, therefore, ought to be able to address. Thus it is necessary to be able to relate the speed and frequency response of analog computation to the rates of the physical processes by which the computation is realized. Traditional methods of algorithm analysis, which are based on sequential time and asymptotic behavior, are inadequate in this frame of relevance. On the one hand, the constants (time scale factors), which reflect the underlying rate of computation are absolutely critical (but ignored in asymptotic analysis); on the other hand, in control applications the asymptotic behavior of algorithm is generally irrelevant, since the inputs are typically fixed in size or of a limited range of sizes.

The CT model of computation is oriented around the idea that the purpose of a computation is to evaluate a mathematical function. Therefore the basic criterion of adequacy for a computation is *correctness*, that is, that given a precise representation of an input to the function, it will produce (after finitely many steps) a precise representation of the corresponding output of the function. In the context of natural computation and control, however, other criteria may be equally or even more relevant. For example, *robustness* is important: how well does the system respond in the presence of noise, uncertainty, imprecision, and error, which are unavoidable in real natural and artificial control systems, and how well does it respond to defects and damage, which arise in many natural and artificial contexts. Since the real world is unpredictable, *flexibility* is also important: how well does an artificial system respond to inputs for which it was not designed, and how well does a natural system behave in situations outside the range of those to which it is evolutionarily adapted. Therefore, *adaptability* (through learning and other means) is another important issue in this frame of relevance.[2]

Transcending Turing Computability

Thus we see that many applications of analog computation raise different questions from those addressed by the CT model of computation; the most useful models of analog computing will have a different frame of relevance. In order to address traditional questions such as whether analog computers can compute "beyond the Turing limit", or whether they can solve NP-hard problems in polynomial time, it is necessary to construct models of analog computation within the CT frame of relevance. Unfortunately, constructing such models requires making commitments about many issues (such as the representation of reals and the discretization of time), that may affect the answers to these questions, but are fundamentally unimportant in the frame of relevance of the most useful applications of the concept of analog computation. Therefore, being overly focused on traditional problems in the theory of computation (which was formulated for a different frame of relevance) may distract us from formulating models of analog computation that can address important issues in its own frame of relevance.

Analog Thinking

It will be worthwhile to say a few words about the *cognitive implications* of analog computing, which are a largely forgotten aspect of analog vs. digital debates of the late 20th century. For example, it was argued that analog computing provides a deeper intuitive understanding of a system than the alternatives do [5,Chap. 8 in 82]. On the one hand, analog computers afforded a means of understanding analytically intractable systems by means of "dynamic models". By setting up an analog simulation, it was possible to vary the parameters and explore interactively the behavior of a dynamical system that could not be analyzed mathematically. Digital simulations, in contrast, were orders of magnitude slower and did not permit this kind of interactive investigation. (Performance has improved sufficiently in contemporary digital computers so that in many cases digital simulations can be used as dynamic models, sometimes with an interface that mimics an analog computer; see [5].)

Analog computing is also relevant to the cognitive distinction between *knowing how* (*procedural knowledge*)

[1] See MacLennan [45,46] for a more detailed discussion of the frame of relevance of the CT model.

[2] See MacLennan [45,46] for a more detailed discussion of the frames of relevance of natural computation and control.

and *knowing that* (*factual knowledge*) [Chap. 8 in 82]. The latter ("know-that") is more characteristic of scientific culture, which strives for generality and exactness, often by designing experiments that allow phenomena to be studied in isolation, whereas the former ("know-how") is more characteristic of engineering culture; at least it was so through the first half of the twentieth century, before the development of "engineering science" and the widespread use of analytic techniques in engineering education and practice. Engineers were faced with analytically intractable systems, with inexact measurements, and with empirical relationships (characteristic curves, etc.), all of which made analog computers attractive for solving engineering problems. Furthermore, because analog computing made use of physical phenomena that were mathematically analogous to those in the primary system, the engineer's intuition and understanding of one system could be transferred to the other. Some commentators have mourned the loss of hands-on intuitive understanding attendant on the increasingly scientific orientation of engineering education and the disappearance of analog computer [5,34,61,67].

I will mention one last cognitive issue relevant to the differences between analog and digital computing. As already discussed Sect. "Characteristics of Analog Computation", it is generally agreed that it is less expensive to achieve high precision with digital technology than with analog technology. Of course, high precision may not be important, for example when the available data are inexact or in natural computation. Further, some advocates of analog computing argue that high precision digital result are often misleading [p. 261 in 82]. Precision does not imply accuracy, and the fact that an answer is displayed with 10 digits does not guarantee that it is accurate to 10 digits; in particular, engineering data may be known to only a few significant figures, and the accuracy of digital calculation may be limited by numerical problems. Therefore, on the one hand, users of digital computers might fall into the trap of trusting their apparently exact results, but users of modest-precision analog computers were more inclined to healthy skepticism about their computations. Or so it was claimed.

Future Directions

Certainly there are many purposes that are best served by digital technology; indeed there is a tendency nowadays to think that everything is done better digitally. Therefore it will be worthwhile to consider whether analog computation should have a role in future computing technologies. I will argue that the approaching end of *Moore's Law* [56], which has predicted exponential growth in digital logic densities, will encourage the development of new analog computing technologies.

Two avenues present themselves as ways toward greater computing power: faster individual computing elements and greater densities of computing elements. Greater density increases power by facilitating parallel computing, and by enabling greater computing power to be put into smaller packages. Other things being equal, the fewer the layers of implementation between the computational operations and the physical processes that realize them, that is to say, the more directly the physical processes implement the computations, the more quickly they will be able to proceed. Since most physical processes are continuous (defined by differential equations), analog computation is generally faster than digital. For example, we may compare analog addition, implemented directly by the additive combination of physical quantities, with the sequential process of digital addition. Similarly, other things being equal, the fewer physical devices required to implement a computational element, the greater will be the density of these elements. Therefore, in general, the closer the computational process is to the physical processes that realize it, the fewer devices will be required, and so the continuity of physical law suggests that analog computation has the potential for greater density than digital. For example, four transistors can realize analog addition, whereas many more are required for digital addition. Both considerations argue for an increasing role of analog computation in post-Moore's Law computing.

From this broad perspective, there are many physical phenomena that are potentially usable for future analog computing technologies. We seek phenomena that can be described by well-known and useful mathematical functions (e. g., addition, multiplication, exponential, logarithm, convolution). These descriptions do not need to be exact for the phenomena to be useful in many applications, for which limited range and precision are adequate. Furthermore, in some applications speed is not an important criterion; for example, in some control applications, small size, low power, robustness, etc. may be more important than speed, so long as the computer responds quickly enough to accomplish the control task. Of course there are many other considerations in determining whether given physical phenomena can be used for practical analog computation in a given application [47]. These include stability, controllability, manufacturability, and the ease of interfacing with input and output transducers and other devices. Nevertheless, in the post-Moore's Law world, we will have to be willing to consider all physical phenomena as potential computing technologies, and in many cases we

will find that analog computing is the most effective way to utilize them.

Natural computation provides many examples of effective analog computation realized by relatively slow, low-precision operations, often through massive parallelism. Therefore, post-Moore's Law computing has much to learn from the natural world.

Bibliography

Primary Literature

1. Anderson JA (1995) An Introduction to Neural Networks. MIT Press, Cambridge
2. Ashley JR (1963) Introduction to Analog Computing. Wiley, New York
3. Aspray W (1993) Edwin L. Harder and the Anacom: Analog computing at Westinghouse. IEEE Ann Hist of Comput 15(2):35–52
4. Ben-Hur A, Siegelmann HT, Fishman S (2002) A theory of complexity for continuous time systems. J Complex 18:51–86
5. Bissell CC (2004) A great disappearing act: The electronic analogue computer. In: IEEE Conference on the History of Electronics, Bletchley, June 2004. pp 28–30
6. Blum L, Cucker F, Shub M, Smale S (1998) Complexity and Real Computation. Springer, Berlin
7. Blum L, Shub M, Smale S (1988) On a theory of computation and complexity over the real numbers: NP completeness, recursive functions and universal machines. Bulletin Am Math Soc 21:1–46
8. Bournez O, Campagnolo ML, Graça DS, Hainry E. The General Purpose Analog Computer and computable analysis are two equivalent paradigms of analog computation. In: Theory and Applications of Models of Computation (TAMC 2006). Lectures Notes in Computer Science, vol 3959. Springer, Berlin, pp 631–643
9. Bournez O, Cosnard M (1996) On the computational power of dynamical systems and hybrid systems. Theor Comput Sci 168(2):417–59
10. Bowles MD (1996) US technological enthusiasm and British technological skepticism in the age of the analog brain. Ann Hist Comput 18(4):5–15
11. Branicky MS (1994) Analog computation with continuous ODEs. In: Proceedings IEEE Workshop on Physics and Computation, Dallas, pp 265–274
12. Brockett RW (1988) Dynamical systems that sort lists, diagonalize matrices and solve linear programming problems. In: Proceedings 27th IEEE Conference Decision and Control, Austin, December 1988, pp 799–803
13. Camazine S, Deneubourg J-L, Franks NR, Sneyd G, Theraulaz J, Bonabeau E (2001) Self-organization in Biological Systems. Princeton Univ. Press, New York
14. Changeux J-P (1985) Neuronal Man: The Biology of Mind (trans: Garey LL). Oxford University Press, Oxford
15. Clymer AB (1993) The mechanical analog computers of Hannibal Ford and William Newell. IEEE Ann Hist Comput 15(2): 19–34
16. Davidson EH (2006) The Regulatory Genome: Gene Regulatory Networks in Development and Evolution. Academic Press, Amsterdam
17. Davies JA (2005) Mechanisms of Morphogensis. Elsevier, Amsterdam
18. Davis M (2004) The myth of hypercomputation. In: Teuscher C (ed) Alan Turing: Life and Legacy of a Great Thinker. Springer, Berlin, pp 195–212
19. Davis M (2006) Why there is no such discipline as hypercomputation. Appl Math Comput 178:4–7
20. Fakhraie SM, Smith KC (1997) VLSI-Compatible Implementation for Artificial Neural Networks. Kluwer, Boston
21. Franklin S, Garzon M (1990) Neural computability. In: Omidvar OM (ed) Progress in Neural Networks, vol 1. Ablex, Norwood, pp 127–145
22. Freeth T, Bitsakis Y, Moussas X, Seiradakis JH, Tselikas A, Mangou H, Zafeiropoulou M, Hadland R, Bate D, Ramsey A, Allen M, Crawley A, Hockley P, Malzbender T, Gelb D, Ambrisco W, Edmunds MG (2006) Decoding the ancient Greek astronomical calculator known as the Antikythera mechanism. Nature 444:587–591
23. Garzon M, Franklin S (1989) Neural computability ii (extended abstract). In: Proceedings, IJCNN International Joint Conference on Neural Networks, vol 1. Institute of Electrical and Electronic Engineers, New York, pp 631–637
24. Garzon M, Franklin S (1990) Computation on graphs. In: Omidvar OM (ed) Progress in Neural Networks, vol 2. Ablex, Norwood
25. Goldstine HH (1972) The Computer from Pascal to von Neumann. Princeton, Princeton
26. Grossberg S (1967) Nonlinear difference-differential equations in prediction and learning theory. Proc Nat Acad Sci USA 58(4):1329–1334
27. Grossberg S (1973) Contour enhancement, short term memory, and constancies in reverberating neural networks. Stud Appl Math LII:213–257
28. Grossberg S (1976) Adaptive pattern classification and universal recoding: I. parallel development and coding of neural feature detectors. Biol Cybern 23:121–134
29. Hartl DL (1994) Genetics, 3rd edn. Jones & Bartlett, Boston
30. Hopfield JJ (1984) Neurons with graded response have collective computational properties like those of two-state neurons. Proc Nat Acad Sci USA 81:3088–92
31. Howe RM (1961) Design Fundamentals of Analog Computer Components. Van Nostrand, Princeton
32. Khatib O (1986) Real-time obstacle avoidance for manipulators and mobile robots. Int J Robot Res 5:90–99
33. Kirchhoff G (1845) Ueber den Durchgang eines elektrischen Stromes durch eine Ebene, insbesondere durch eine kreisförmige. Ann Phys Chem 140(64)(4):497–514
34. Lang GF (2000) Analog was *not* a computer trademark! Why would anyone write about analog computers in year 2000? Sound Vib 34(8):16–24
35. Lipshitz L, Rubel LA (1987) A differentially algebraic replacment theorem. Proc Am Math Soc 99(2):367–72
36. Maass W, Sontag ED (1999) Analog neural nets with Gaussian or other common noise distributions cannot recognize arbitrary regular languages. Neural Comput 11(3):771–782
37. MacLennan BJ (1987) Technology-independent design of neurocomputers: The universal field computer. In: Caudill M, Butler C (eds) Proceedings of the IEEE First International Conference on Neural Networks, vol 3, IEEE Press, pp 39–49

38. MacLennan BJ (1990) Field computation: A theoretical framework for massively parallel analog computation, parts I–IV. Technical Report CS-90-100, Department of Computer Science, University of Tennessee, Knoxville. Available from http://www.cs.utk.edu/~mclennan. Accessed 20 May 2008
39. MacLennan BJ (1991) Gabor representations of spatiotemporal visual images. Technical Report CS-91-144, Department of Computer Science, University of Tennessee, Knoxville. Available from http://www.cs.utk.edu/~mclennan. Accessed 20 May 2008
40. MacLennan BJ (1994) Continuous computation and the emergence of the discrete. In: Pribram KH (ed) Origins: Brain & Self-Organization, Lawrence Erlbaum, Hillsdale, pp 121–151.
41. MacLennan BJ (1994) "Words lie in our way". Minds Mach 4(4):421–437
42. MacLennan BJ (1995) Continuous formal systems: A unifying model in language and cognition. In: Proceedings of the IEEE Workshop on Architectures for Semiotic Modeling and Situation Analysis in Large Complex Systems, Monterey, August 1995. pp 161–172. Also available from http://www.cs.utk.edu/~mclennan. Accessed 20 May 2008
43. MacLennan BJ (1999) Field computation in natural and artificial intelligence. Inf Sci 119:73–89
44. MacLennan BJ (2001) Can differential equations compute? Technical Report UT-CS-01-459, Department of Computer Science, University of Tennessee, Knoxville. Available from http://www.cs.utk.edu/~mclennan. Accessed 20 May 2008
45. MacLennan BJ (2003) Transcending Turing computability. Minds Mach 13:3–22
46. MacLennan BJ (2004) Natural computation and non-Turing models of computation. Theor Comput Sci 317:115–145
47. MacLennan BJ (in press) Super-Turing or non-Turing? Extending the concept of computation. Int J Unconv Comput, in press
48. Maini PK, Othmer HG (eds) (2001) Mathematical Models for Biological Pattern Formation. Springer, New York
49. Maziarz EA, Greenwood T (1968) Greek Mathematical Philosophy. Frederick Ungar, New York
50. McClelland JL, Rumelhart DE, the PDP Research Group. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol 2: Psychological and Biological Models. MIT Press, Cambridge
51. Mead C (1987) Silicon models of neural computation. In: Caudill M, Butler C (eds) Proceedings, IEEE First International Conference on Neural Networks, vol I. IEEE Press, Piscataway, pp 91–106
52. Mead C (1989) Analog VLSI and Neural Systems. Addison-Wesley, Reading
53. Mills JW (1996) The continuous retina: Image processing with a single-sensor artificial neural field network. In: Proceedings IEEE Conference on Neural Networks. IEEE Press, Piscataway
54. Mills JW, Himebaugh B, Kopecky B, Parker M, Shue C, Weilemann C (2006) "Empty space" computes: The evolution of an unconventional supercomputer. In: Proceedings of the 3rd Conference on Computing Frontiers, New York, May 2006. ACM Press, pp 115–126
55. Moore C (1996) Recursion theory on the reals and continuous-time computation. Theor Comput Sci 162:23–44
56. Moore GE (1965) Cramming more components onto integrated circuits. Electronics 38(8):114–117
57. Murray JD (1977) Lectures on Nonlinear Differential-Equation Models in Biology. Oxford, Oxford
58. Omohundro S (1984) Modeling cellular automata with partial differential equations. Physica D 10:128–34, 1984.
59. Orponen P (1997) A survey of continous-time computation theory. In: Advances in Algorithms, Languages, and Complexity, Kluwer, Dordrecht, pp 209–224
60. Orponen P, Matamala M (1996) Universal computation by finite two-dimensional coupled map lattices. In: Proceedings, Physics and Computation 1996, New England Complex Systems Institute, Cambridge, pp 243–7
61. Owens L (1986) Vannevar Bush and the differential analyzer: The text and context of an early computer. Technol Culture 27(1):63–95
62. Peterson GR (1967) Basic Analog Computation. Macmillan, New York
63. Pour-El MB (1974) Abstract computability and its relation to the general purpose analog computer (some connections between logic, differential equations and analog computers). Trans Am Math Soc 199:1–29
64. Pour-El MB, Richards I (1979) A computable ordinary differential equation which possesses no computable solution. Ann Math Log 17:61–90
65. Pour-EL MB, Richards I (1981) The wave equation with computable initial data such that its unique solution is not computable. Adv Math 39:215–239
66. Pour-El MB, Richards I (1982) Noncomputability in models of physical phenomena. Int J Theor Phys, 21:553–555
67. Puchta S (1996) On the role of mathematics and mathematical knowledge in the invention of Vannevar Bush's early analog computers. IEEE Ann Hist Comput 18(4):49–59
68. Reiner JM (1968) The Organism as an Adaptive Control System. Prentice-Hall, Englewood Cliffs
69. Rimon E, Koditschek DE (1989) The construction of analytic diffeomorphisms for exact robot navigation on star worlds. In: Proceedings of the 1989 IEEE International Conference on Robotics and Automation, Scottsdale AZ. IEEE Press, New York, pp 21–26
70. Rogers AE, Connolly TW (1960) Analog Computation in Engineering Design. McGraw-Hill, New York
71. Rubel LA (1985) The brain as an analog computer. J Theor Neurobiol 4:73–81
72. Rubel LA (1988) Some mathematical limitations of the general-purpose analog computer. Adv Appl Math 9:22–34
73. Rubel LA (1993) The extended analog computer. Adv Appl Math 14:39–50
74. Rumelhart DE, McClelland JL, the PDP Research Group (1986) Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol 1: Foundations. MIT Press, Cambridge
75. Sanger TD (1996) Probability density estimation for the interpretation of neural population codes. J Neurophysiol 76:2790–2793
76. Shannon CE (1941) Mathematical theory of the differential analyzer. J Math Phys Mass Inst Technol 20:337–354
77. Shannon CE (1993) Mathematical theory of the differential analyzer. In: Sloane NJA, Wyner AD (eds) Claude Elwood Shannon: Collected Papers. IEEE Press, New York, pp 496–513
78. Siegelmann HT (1999) Neural Networks and Analog Computation: Beyond the Turing Limit. Birkhäuser, Boston
79. Siegelmann HT, Ben-Hur A, Fishman S (1999) Computational complexity for continuous time dynamics. Phys Rev Lett 83(7):1463–6

80. Siegelmann HT, Sontag ED (1994) Analog computation via neural networks. Theor Comput Sci 131:331–360
81. Small JS (1993) General-purpose electronic analog computing. IEEE Ann Hist Comput 15(2):8–18
82. Small JS (2001) The Analogue Alternative: The electronic analogue computer in Britain and the USA, 1930–1975. Routledge, London & New York
83. Stannett M (19901) X-machines and the halting problem: Building a super-Turing machine. Form Asp Comput 2:331–341
84. Thomson W (Lord Kelvin) (1876) Mechanical integration of the general linear differential equation of any order with variable coefficients. Proc Royal Soc 24:271–275
85. Thomson W (Lord Kelvin) (1878) Harmonic analyzer. Proc Royal Soc 27:371–373
86. Thomson W (Lord Kelvin) (1938). The tides. In: The Harvard Classics, vol 30: Scientific Papers. Collier, New York, pp 274–307
87. Truitt TD, Rogers AE (1960) Basics of Analog Computers. John F. Rider, New York
88. van Gelder T (1997) Dynamics and cognition. In: Haugeland J (ed) Mind Design II: Philosophy, Psychology and Artificial Intelligence. MIT Press, Cambridge MA, revised & enlarged edition, Chap 16, pp 421–450
89. Weyrick RC (1969) Fundamentals of Analog Computers. Prentice-Hall, Englewood Cliffs
90. Wolpert DH (1991) A computationally universal field computer which is purely linear. Technical Report LA-UR-91-2937. Los Alamos National Laboratory, Loa Alamos
91. Wolpert DH, MacLennan BJ (1993) A computationally universal field computer that is purely linear. Technical Report CS-93-206. Dept. of Computer Science, University of Tennessee, Knoxville

Books and Reviews

Fifer S (1961) Analog computation: Theory, techniques and applications, 4 vols. McGraw-Hill, New York
Bissell CC (2004) A great disappearing act: The electronic analogue computer. In: IEEE conference on the history of electronics, 28–30 Bletchley, June 2004
Lipka J (1918) Graphical and mechanical computation. Wiley, New York
Mead C (1989) Analog VLSI and neural systems. Addison-Wesley, Reading
Siegelmann HT (1999) Neural networks and analog computation: Beyond the Turing limit. Birkhäuser, Boston
Small JS (2001) The analogue alternative: The electronic analogue computer in Britain and the USA, 1930–1975. Routledge, London & New York
Small JS (1993) General-purpose electronic analog computing: 1945–1965. IEEE Ann Hist Comput 15(2):8–18

Anisotropic Networks, Elastomers and Gels

Eugene M. Terentjev
Cavendish Laboratory, University of Cambridge, Cambridge, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Nematic Elastomers: Reversible Shape Change
Nematic Elastomers: Soft Elasticity and Dynamics
Swelling and Anisotropic Gels
Cholesteric Elastomers: Photonics
New Frontiers
Bibliography

Glossary

Elastomer Nominally equivalent to "rubber", this defines a weakly crosslinked network of polymer chains which retain high thermal mobility of the strands between crosslinks. In this case the entropic effects dominate the response of such networks. Entropic rubber elasticity arises when such a network percolates the whole macroscopic system and the crosslinks constrain it such that it remembers its equilibrium shape, and respond elastically to deformations. In this sense, an *elastomer* is contrasted to a polymer *glass*: The latter can refer to a network so densely crosslinked that the individual chain segments have no significant mobility – or a polymer system below its structural glass transition. The elastic modulus of elastomers is usually in the range 10–1000 kPa, the estimate arising from the thermal energy kT per network strand, while the glass modulus is usually 0.1–10 GPa, as in most solids.

Gel This word is used in many different contexts, in all cases referring to a soft object that is capable of retaining its shape in ambient conditions against, e. g., gravity. The latter condition distinguishes a gel from a "sol", or a nominal liquid. The softness is a relative concept; usually it refers to the modulus at or below the "human" scale, around 1–100 kPa, but there are many examples of even softer gels. In this article, a *gel* is contrasted to an *elastomer*, referring to a crosslinked network of polymer chains which is swollen by a good solvent. In this case the effective modulus becomes very low, nevertheless the system remains elastic as long as the integrity of percolating network remains intact.

Quenched constraints The term *quenched* (as opposed to *annealed*) refers to objects or systems that are prevented from exploring their full phase or conformational space during thermal motion. The classical example of quenched system is a structural glass, where the mobility of elements is very limited. In elastomers and gels, network crosslinks are the quenched objects, constraining the polymer chains connecting to them.

The effect of randomly quenched constraints is profound in many physical systems, as the local thermodynamic equilibrium has to establish among the mobile (annealed) elements while observing the constraints, the random local realization of which is determined by external factors, such as the preparation history.

Liquid crystal This refers to a group of anisotropic phases with incomplete translational order (which would represent a crystalline lattice). Classical examples are: The *nematic* phase – a fluid with no translational order at all, but with an orientational order of anisotropic molecules, and *smectic* or *lamellar* phases, which in addition to orientational order also have a 1-dimensional density modulation (a stack of parallel layers). Liquid crystallinity is a state of spontaneous, equilibrium anisotropy (breaking of orientational symmetry), in contrast to anisotropy induced by external factors such as electric/magnetic field or mechanical deformation.

Shape memory Strictly, any elastic material "remembers" its equilibrium shape and returns to it after deformation. The term *shape-memory* was introduced to distinguish materials that could be made to preserve their deformed state, until a trigger (e. g. heating above a transition temperature) induces the return to the original equilibrium shape. In shape-memory alloys, this is due to the martensitic transition, in shape-memory polymers the deformed state can be fixed by a glass transition or partial crystallization. Liquid crystal elastomers and gels have a reversible, or *equilibrium* shape-memory in that the shape of the body is determined by the current state of order at any given temperature.

Definition of the Subject

Anisotropic (liquid crystalline) elastomers and gels bring together, as nowhere else, three important ideas: *Orientational order* in amorphous soft materials, *responsive molecular shape* and *quenched topological constraints*. Acting together, they create many new physical phenomena that are briefly reviewed in this article. Classical liquid crystals are typically fluids of relatively stiff rod molecules with long range orientational order. Long polymer chains, with incorporated rigid anisotropic units can also form orientationally ordered liquid crystalline phases. By contrast with free rigid rods, these flexible chains change their average molecular shape, from isotropic spherical to ellipsoidal, when their component rods align. Linking the polymer chains together into a network fixes their relative topology, and the polymer melt (or solution) becomes an elastic solid – an elastomer (or gel). Radically new properties arise from the ability to change average molecular shape of anisotropic polymer chains in the solid state.

In ordinary solids the deformations are created by relative movement of the same atoms (or molecules) that form the bonded lattice. Hence, when the deformation is small, the lattice symmetry is preserved and one obtains an ordinary elastic response (often anisotropic). There is a classical elastic response found in glasses as well (either isotropic or anisotropic), where in place of the crystalline lattice recording the preferred position of atoms, they are confined by constraints of local cages. Either way, the elements of the body "know" their positions and the system responds with an increase of elastic energy when these elements are displaced. Large deformations destroy the lattice (or cage) integrity and simply break the material.

In contrast, in elastomers and gels the macroscopic elastic response arises from the entropy change of chains on relative movement of their crosslinked end-points, which are relatively far apart. Monomers remain highly mobile and thus liquid-like. What happens to chain segments (monomer moieties) on a smaller length scale is a relatively independent matter and such weakly crosslinked network behaves as a liquid on length scales smaller than the end-point separation of strands. In particular, the liquid crystalline order can be established within these segments. The magnitude of this order can be altered and its director can rotate, in principle, independently of deformation of the crosslinking points. Such an internal degree of freedom within, and coupled to the elastic body provides additional local forces and torques, intricately connected in the overall macroscopic response of the body.

In the simplest and most straightforward case of nematic elastomers and gels, the uniaxial ellipsoidal average anisotropy of chain constituting the network leads to two key physical properties that make these materials so unique. If one can change the level of nematic (orientational) order, which is not difficult to achieve by e. g. changing temperature, changes at the molecular level induce a corresponding mechanical strains: A block of rubber can contract or elongate by a large amount on heating or cooling, respectively. This process is perfectly reversible. This leads to a group of possible applications in artificial muscles and actuators, which can be driven by any stimulus that affects the local nematic order: temperature, solvent intake, irradiation, etc.

It is also possible to rotate the nematic director axis and the rubber matrix independently, although in contrast to ordinary liquid crystals it costs energy to uni-

formly rotate the director within the matrix. This penalty leads to suppression of orientational fluctuations and high optical clarity of birefringent elastomers and gels. Local rotations also yield a subtle and spectacular elastic phenomenon which is called 'soft elasticity'. Contrary to intuition, there is an infinite number of non-trivial mechanical deformations that can accommodate the rotation of anisotropic distribution of chains without its distortion. As a result, the entropy of the chains does not change, in spite of macroscopic deformations, and the material can be distorted without any significant energy cost! A special combination of shears and elongations/compressions is required, but it turns out not very difficult to achieve. Elastic softness, or attempts by the material to achieve it, pervade much of the elasticity of nematic elastomers and gels. For instance, another unique and spectacular property of these systems is anomalously high damping, where the mechanical energy is dissipated on these soft rotational deformation modes.

Cholesteric liquid crystals have a helical director texture. When crosslinked into elastomers or gels, this periodic helical anisotropy is made to couple to the mechanical state of the whole body. Their optical and mechanical responses to imposed stress are exceedingly rich as a result. Cholesteric elastomers are brightly colored due to selective reflection, and change color as they are stretched – their photonic band structure can be powerfully manipulated by applied deformations. Such periodic (distributed feedback) photonic systems can emit laser radiation, the color of which can be continuously shifted by mechanical means across the whole visible spectrum. Further, the effect of topological imprinting can select and extract molecules of specific handedness from a mixed solvent. Cholesteric gels can act as a mechanical separator of chirality.

Smectic or lamellar elastomers and gels have plane-like, layered modulation of density in one direction (SmA), or additionally a tilt of the director away from the layer normal (SmC). Many other more complex smectic phases exist and could also be made into elastomers. Here the layers are often constrained by crosslinking density modulation not to move relative to the rubber matrix. Deformations along the layer normal can be resisted by a modulus up to 100 times greater than the normal rubber (shear) modulus of the matrix. Thus smectic elastomers are rubbery in the two dimensions of their layer planes, but respond as hard conventional solids in their third dimension. Such extreme mechanical anisotropy promises interesting applications.

The director tilt associated with the transition from SmA to SmC induces distortion in the polymer chain shape distribution. Since chain shape is coupled to mechanical shape for an elastomer with quite low symmetry, one expects a large variety of independent soft-elasticity modes. The tilted SmC systems also exist in chiral forms which must on symmetry grounds be ferroelectric, with spontaneous polarization pointing along the vector product $[\hat{\boldsymbol{k}} \times \boldsymbol{n}]$, see Fig. 1d. Ferroelectric elastomer is very special: Mechanically it is soft, about 10^4 times lower in modulus than conventional solid ferro- and piezoelectrics. Distortions give polarization changes comparable to those

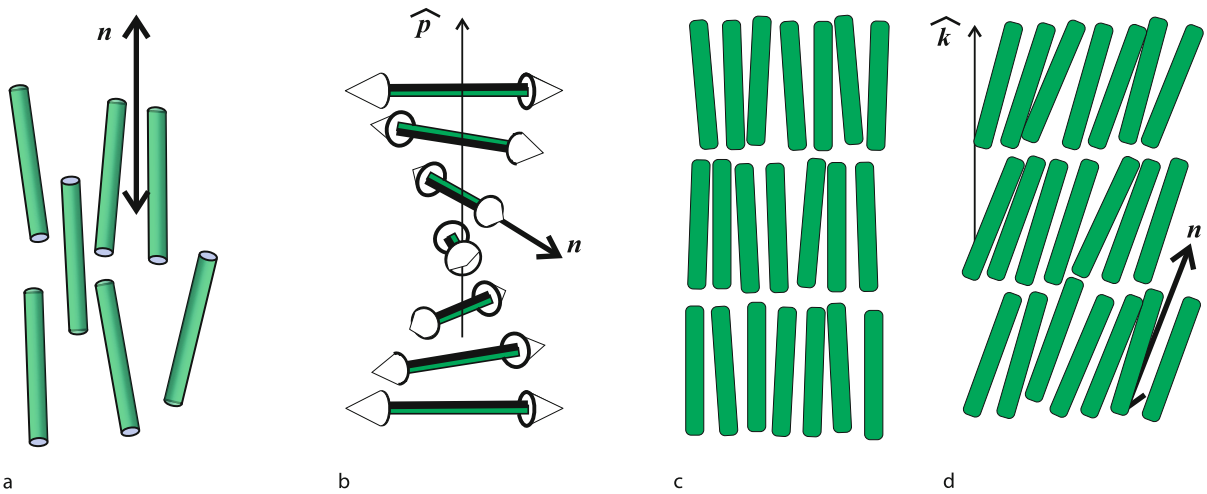

Anisotropic Networks, Elastomers and Gels, Figure 1

Arrangement of anisotropic (*rod-like*) molecular moieties in common liquid crystal phases: a nematic, b cholesteric, c smectic-A and d smectic-C. The average orientation of long axes is the director $\boldsymbol{n}$. The director, as well as the cholesteric helix axis $\hat{p}$ and the smectic layer normal $\hat{k}$, are all directionless "quadrupolar" vectors, e. g. $\boldsymbol{n} = -\boldsymbol{n}$

in ordinary ferroelectrics, but the response to an applied stress must necessarily much larger than in conventional materials due to the very low mechanical impedance.

It is debatable, whether an anisotropic glass could be called a liquid crystal, although nominally a glass can be made to possess orientational but not a translational order. The lack of molecular mobility makes the question of thermodynamic equilibrium very ambiguous.

Introduction

Within the simplest affine deformation approach one regards the change in each chain end-to-end vector $\boldsymbol{R}$ as $\boldsymbol{R}' = \boldsymbol{F} \cdot \boldsymbol{R}$, when a deformation of the overall polymer network is characterized by a deformation gradient tensor F_{ij}. In continuum elasticity this tensor is often called $\boldsymbol{F} = \nabla \boldsymbol{f}$, a gradient of the geometrical mapping $\boldsymbol{f}$ from the reference to the current state of deformation. Assuming the chain connecting the two crosslinks is long enough, the Gaussian approximation for the number of its conformations $Z(\boldsymbol{R})$ gives for the free energy (per chain): $W_{\text{ch}} = -kT \ln Z(\boldsymbol{R}') \simeq (kT/a^2N)[\boldsymbol{F}^T \cdot \boldsymbol{F}]_{ij} R_i R_j$, where a is the step length of the chain random walk and N the number of such steps. In order to find the total elastic work function of all chains affinely deforming in the network, one needs to add the contributions $W_{\text{ch}}(R)$ with the statistical weight to find a chain with a given initial end-to-end distance $\boldsymbol{R}$ in the system. This procedure, called the quenched averaging, produces the average $\langle R_i R_j \rangle \simeq \frac{1}{3} a^2 N \delta_{ij}$ in W_{ch}. The resulting rubber-elastic free energy (per unit volume) is $W_{\text{el}} = \frac{1}{2} n_c kT(\boldsymbol{F}^T : \boldsymbol{F})$, with n_c a number of chains per unit volume of the network. This is a remarkably robust expression, with many seemingly relevant effects, such as the fluctuation of crosslinking points, only contributing a small quantitative change in the prefactor. It is, however, incomplete since the simplified Gaussian statistics of chains does not take into account their physical volume and thus does not address the compressibility issue. The proper way of dealing with it is by adding an additional independent penalty for the volume change: $\frac{1}{2} K(\det \boldsymbol{F} - 1)^2$ in elastomers, with K the typical bulk modulus of the order 10 GPa in organic materials. This should be modified into $\frac{1}{2} K(\det \boldsymbol{F} - \Phi)^2$ for gels swollen by a volume fraction Φ of solvent.

If one assumes that the rubber elastic modulus is low, then a very reasonable approximation is to simply impose the volume-conservation constraint on deformation tensor: $\det \boldsymbol{F} = 1$. The value of the rubber modulus is found on expanding the deformation gradient tensor in small strains, say for a small extension, $F_{zz} = 1 + \varepsilon$, and obtaining $W_{\text{el}} \simeq \frac{3}{2} n_c kT \varepsilon^2$. This means the extension (Young) modulus $E = 3 n_c kT$; the analogous construction for a small simple shear will give the shear modulus $G = n_c kT$, exactly a third of the Young modulus as required in an incompressible medium. This shear modulus G, having its origin in the entropic effect of reduction of conformational freedom on polymer chain deformation, is usually so much smaller than the bulk modulus (determined by the enthalpy of compressing the dense polymer liquid) that the assumption of rubber or gel deforming at constant volume is justified. This constraint leads to the familiar rubber-elastic expression $W_{\text{el}} = \frac{1}{2} n_c kT(\lambda^2 + 2/\lambda)$ where one has assumed that the imposed extension $F_{zz} = \lambda$ is accompanied by the symmetric contraction in both transverse directions, $F_{xx} = F_{yy} = 1/\sqrt{\lambda}$ due to the incompressibility.

When the chains forming the rubbery network are liquid crystalline, their end-to-end distance distribution becomes anisotropic. The case of smectic/lamellar ordering is much more complicated. For a simple uniaxial nematic one obtains $\langle R_\parallel R_\parallel \rangle = \frac{1}{3} \ell_\parallel L$ and $\langle R_\perp R_\perp \rangle = \frac{1}{3} \ell_\perp L$, with $L = aN$ the chain contour length and $\ell_\parallel / \ell_\perp$ the ratio of average chain step lengths along and perpendicular to the nematic director. In the isotropic phase one recovers $\ell_\parallel = \ell_\perp = a$. The uniaxial anisotropy of polymer chains has a principal axis along the nematic director $\boldsymbol{n}$, with a prolate ($\ell_\parallel / \ell_\perp > 1$) or oblate ($\ell_\parallel / \ell_\perp < 1$) ellipsoidal conformation of polymer backbone. The ability of this principal axis to rotate independently under the influence of network strains makes the rubber elastic response non-symmetric, so that the elastic work function is [6]

$$W_{\text{el}} = \frac{1}{2} G \mathrm{Tr} \left(\boldsymbol{F}^T \cdot \boldsymbol{\ell}^{-1} \cdot \boldsymbol{F} \cdot \boldsymbol{\ell}_0 \right) + \frac{1}{2} K (\det \boldsymbol{F} - 1)^2 , \quad (1)$$

with $\boldsymbol{\ell}$ the uniaxial matrices of chain step-lengths before ($\boldsymbol{\ell}_0$) and after the deformation to the current state: $\ell_{ij} = \ell_\perp \delta_{ij} + [\ell_\parallel - \ell_\perp] n_i n_j$. Note that the deformation gradient tensor $\boldsymbol{F}$ does no longer enter the elastic energy in the symmetric combination $\boldsymbol{F}^T \cdot \boldsymbol{F}$, but is "sandwiched" between the matrices $\boldsymbol{\ell}$ with different principal axes. This means that antisymmetric components of strain will now have a non-trivial physical effect, in contrast to isotropic rubbers and, more crucially, to elastic solids with uniaxial anisotropy. There, the anisotropy axis is immobile and the response is anisotropic but symmetric in stress and strain. The uniqueness of nematic rubbers stems from the competing microscopic interactions and the difference in characteristic length scales: The uniaxial anisotropy is established on a small (monomer) scale of nematic coherence length, while the strains are defined (and the elastic response is arising) on a much greater length scale of polymer chain end-to-end distance, see Fig. 2, [67,68].

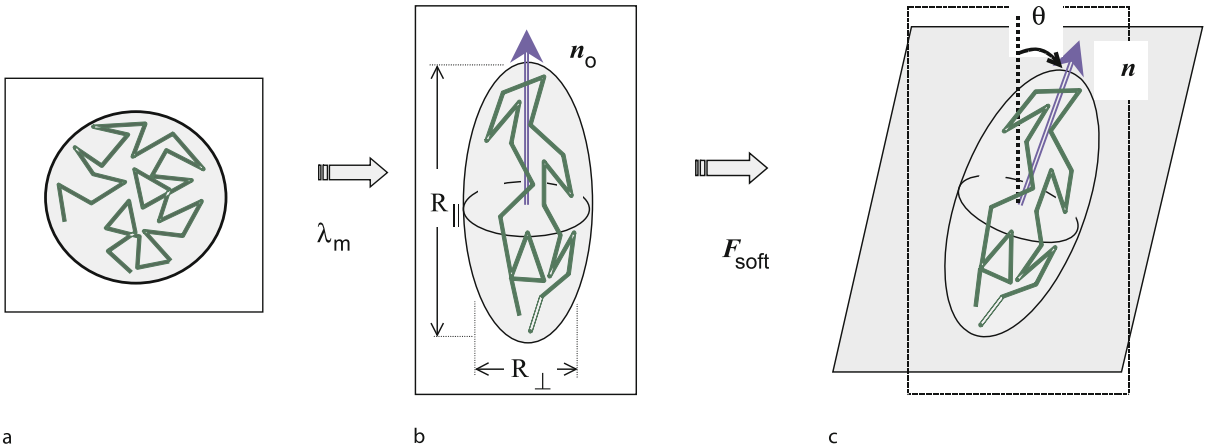

Anisotropic Networks, Elastomers and Gels, Figure 2
Relation between the equilibrium chain shape and deformations in nematic elastomers. When the network of initially isotropic chains, each o average forming a spherical shape of gyration a, is brought into the nematic phase, a corresponding spontaneous deformation of the sample occurs in proportion to the backbone anisotropy, $\lambda_m = (\ell_\parallel/\ell_\perp)^{-1/3}$, b. An example of soft deformation is given in c, when rotating anisotropic chains can affinely accommodate all strains (a combination of compression along the initial director, extension across it and a shear in the plane of director rotation), not causing any entropic rubber-elastic response

Nematic Elastomers: Reversible Shape Change

The "Trace formula" (1) has proven very successful in describing many physical properties of nematic and smectic elastomers and gels. One of the most important consequences of coupling the nematic order to the shape of an elastic body (and perhaps the most relevant for applications) is the effect of spontaneous uniaxial extension/contraction. It is a very simple phenomenon, pictorially illustrated in transformation between states (a) and (b) of Fig. 2. Mathematically, it is straightforward to obtain from (1) for the fixed director orientation $n_z = 1$ and $\det[\boldsymbol{F}] = 1$, that

$$W_{\mathrm{el}} = \frac{1}{2}G\left(\lambda^2\frac{\ell_\parallel^{(0)}}{\ell_\parallel} + \frac{2}{\lambda}\frac{\ell_\perp^{(0)}}{\ell_\perp}\right) , \tag{2}$$

minimizing which one obtains the equilibrium uniaxial extension λ along the nematic director. In the case when the reference state $\boldsymbol{\ell}_0$ is isotropic, $\ell_\parallel^{(0)} = \ell_\perp^{(0)} = a$, this spontaneous extension along the director takes the especially simple form [1,69]:

$$\lambda_{\mathrm{m}} = (\ell_\parallel/\ell_\perp)^{1/3} . \tag{3}$$

The particular way $\ell_\parallel$ and $\ell_\perp$ depend on the value of nematic order parameter Q depends on the molecular structure of the system. For weak chain anisotropy it is certainly linear: $\ell_\parallel/\ell_\perp = 1 + \alpha Q$, however, there are situations (e. g. in highly anisotropic main-chain nematic polymers) when the ratio $\ell_\parallel/\ell_\perp \propto \exp[1/(1-Q)]$ takes very high values. The key relation (3) lies at the heart of nematic elastomers performing as thermal actuators or artificial muscles, Fig. 3, [34,54,60]. The analogy is further enhanced by the fact that the typical stress exerted on such actuation is in the range of 10–100 kPa, the same as in human muscle [54]. A large amount of work has been recently reporting on photo-actuation of nematic elastomers containing e. g. azobenzene moieties, Fig. 4, or on the effect of adding/removing solvents [22,26,37] – in all cases the effect is due to the changing of underlying nematic order parameter Q, affecting the magnitude of the ratio $(\ell_\parallel/\ell_\perp)$ and thus the length of elastomer sample or the exerted force if the length is constrained [19]. Finally, creating an inhomogeneous internal deformation results in the bending actuation response, Fig. 5, [9,53,71].

However useful, the concise expression (1) does arise from a highly idealized molecular model based on the simple Gaussian statistics of polymer chains between crosslinks and, as such, may seem oversimplified. It turns out that the symmetry features, correctly captured in the Trace formula, are more important than various corrections and generalizations of the model. In particular, a much more complex theory of nematic rubber elasticity taking into account chain entanglements (which have to play a very important role in crosslinked networks, where the constraint release by diffusion is prohibited), devel-

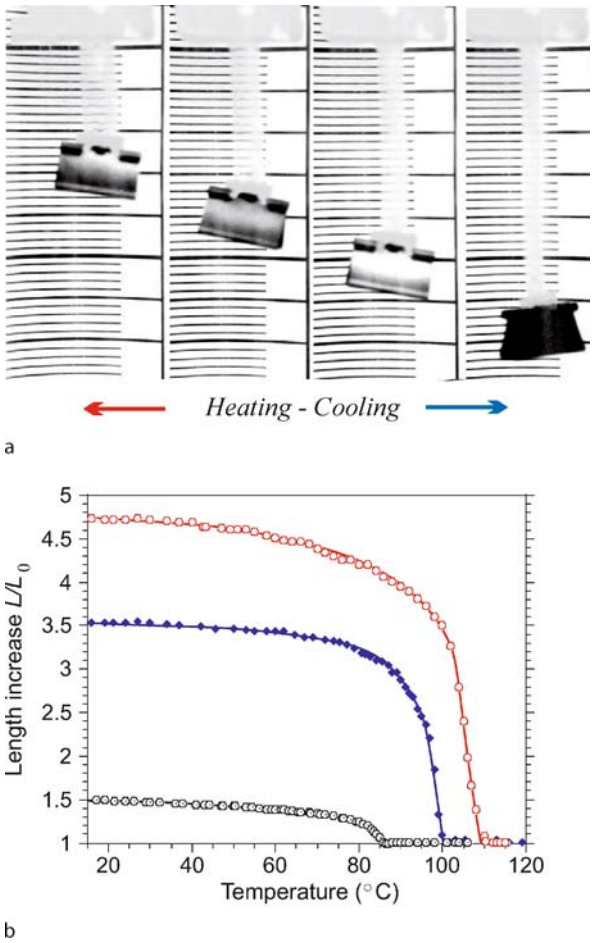

Anisotropic Networks, Elastomers and Gels, Figure 3
Thermal actuation of nematic elastomers. Images a show a strip of rubber with its director aligned along its length, reversibly changing its length on heating and cooling. The same underlying mechanism of changing the orientational order parameter $Q(T)$ translates itself into the mechanical effect differently, depending on the molecular structure of the polymer. Plot b shows the relative change in sample length (the strain λ_m) for different materials, illustrating the range of motion and the different point of transition into the isotropic phase, where $Q = 0$ and $L = L_0$

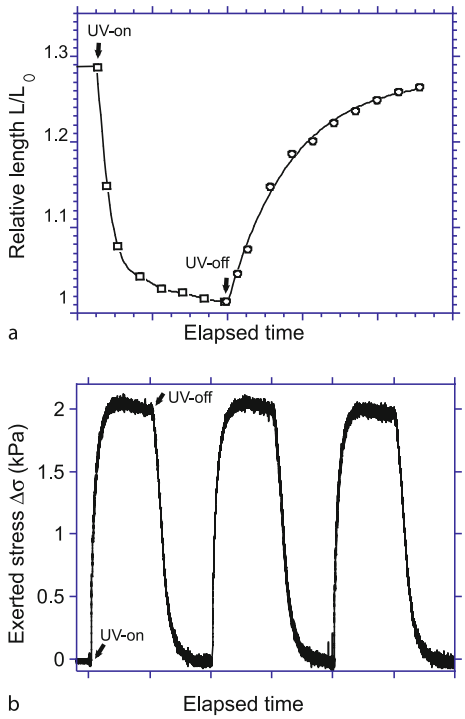

Anisotropic Networks, Elastomers and Gels, Figure 4
Photo-induced actuation of nematic elastomers. Plot a shows the contraction of a strip of azobenzene-containing nematic rubber on UV-irradiation, returning back to its equilibrium value when the light is turned off. Plot b shows the exerted stress under illumination, when the sample is mechanically held. The time-scales of photo-actuation can vary greatly depending on the molecular structure of chromophores

oped in [33], shows the same key symmetries:

$$W_{\mathrm{el}} = \frac{2}{3}G\frac{2M+1}{3M+1}\mathrm{Tr}\left(\boldsymbol{F}^T\cdot\boldsymbol{\ell}^{-1}\cdot\boldsymbol{F}\cdot\boldsymbol{\ell}_0\right) + \frac{3}{2}G(M-1)\frac{2M+1}{3M+1}\left(\overline{\left|\boldsymbol{\ell}^{-1/2}\cdot\boldsymbol{F}\cdot\boldsymbol{\ell}_0^{1/2}\right|}\right)^2 + G(M-1)\overline{\ln\left|\boldsymbol{\ell}^{-1/2}\cdot\boldsymbol{F}\cdot\boldsymbol{\ell}_0^{1/2}\right|}\,, \tag{4}$$

where M is a number of entanglements on a typical chain of length N ($M = 1$ in the ideal Trace formula), and the notation $\overline{|\ldots|}$ denotes an orientational average of the matrix … applied to an arbitrary tube segment. The important feature of the complicated expression (4) is the "sandwiching" of the deformation gradient tensor $\boldsymbol{F}$ between the two orientational tensors defined in the current and the reference states, respectively. The continuum model of fully non-linear elasticity is also quite complicated in its applications, but the fundamental requirements that any elastic free energy expression must be (i) invariant with respect to body rotations of the reference state and the current state, separately, and (ii) reduce to the classical expression $\mathrm{Tr}(\boldsymbol{F}^T\cdot\boldsymbol{F})$ when the underlying chains become isotropic, requires it to have the form

$$W_{\mathrm{el}} \propto \mathrm{Tr}\left(\boldsymbol{F}^T\cdot[\boldsymbol{\ell}]^{-m}\cdot\boldsymbol{F}\cdot[\boldsymbol{\ell}_0]^m\right)\,, \tag{5}$$

where the power is only allowed to be $m = 1, 2, 3$. Once again, one recognizes the key symmetry of separating the two powers of deformation gradient tensor by the orien-

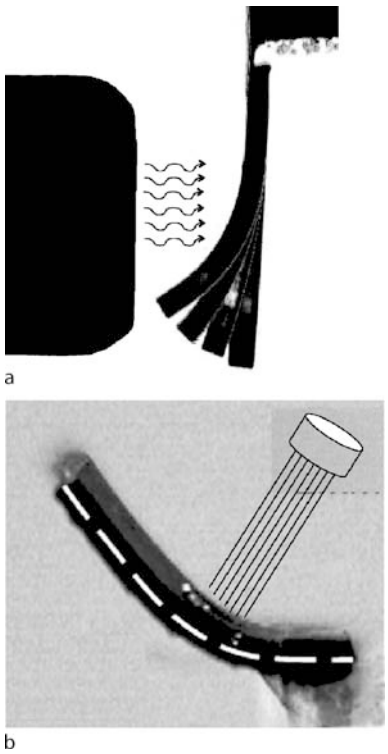

Anisotropic Networks, Elastomers and Gels, Figure 5
Bending of nematic elastomer cantilevers. Image collage a shows the effect of heat (IR radiation) applied from one side of the rubber strip, resulting in the near side contracting more than the far side. Image b shows the similar effect on photo-elastomer illuminated from one side

tation tensors defined in the current and reference frames, respectively.

Nematic Elastomers: Soft Elasticity and Dynamics

Two essential consequences of the coupling between the elastic modes of polymer network, described by the general strain tensor $\boldsymbol{F}$, and the rotational modes of the nematic director, which are the principal axes of chain anisotropy tensor $\boldsymbol{\ell}$, are the reduction of the effective elastic response and the penalty on the director fluctuations. The first effect has received the name "soft elasticity" and is the result of the special symmetry of the coupling between the orientational modes and deformation gradients. It is easy to see [44] that there is a whole continuous class of deformations described by the form

$$\boldsymbol{F}_{\text{soft}} = \boldsymbol{\ell}_{\theta}^{1/2} \cdot \boldsymbol{V} \cdot \boldsymbol{\ell}_{0}^{-1/2}, \tag{6}$$

with $\boldsymbol{V}$ an arbitrary unitary matrix, which leave the elastic free energy W_{el} at its constant minimum value (for an incompressible system). Remarkably, this remains true whether one examines the Trace formula (1), or any other expression above, (4) or (5), as long as they respect the correct symmetries of the two independent degrees of freedom [24,50]. Figure 2c illustrates one example of such soft deformation.

This phenomenon is unique to anisotropic elastomers and gels, which have an internal microstructure (liquid crystalline order) that is capable of continuously changing its tensor structure. It can also be seen in smectic elastomers, although the layer constraints make it more complicated mathematically. Possible the closest elastic system is the famous shape-memory alloy [5], where the martensite transition creates a crystalline unit cell with several (usually just two) equal-energy states. The ability of the lattice to switch between these states in response to certain deformation modes gives the "soft deformation" modes to these alloys, albeit only a few discrete ones. In elastomers and gels the "soft deformation" reflects the ability of anisotropic polymer chains to rotate their long axis to accommodate some imposed elastic deformations without changing their shape.

If one instead chooses to focus on the director modes in a nematic elastomer with a fixed (constrained) shape, the coupling between the internal mobile order parameter and the network elasticity provide a large energy penalty for uniform director rotations $\delta\boldsymbol{n}$ (with respect to the elastically constrained network). This penalty, which appears as a large mass term in the expression for mean-square director fluctuation, results in the suppression of such fluctuations and the related scattering of light from a nematic elastomer [48]. In contrast to optically turbid ordinary liquid nematics, where light is scattered by long-wavelength director fluctuations, the aligned monodomain anisotropic elastomer or gel is totally transparent. However, when the elastic deformations in the network are not constrained and are free to relax, there are certain combinations of polarization and wave vectors of director fluctuations (corresponding to the soft deformation modes), for which the "effective mass" vanishes and the fluctuation spectrum should appear as in ordinary liquid nematics [44].

It is natural to expect that if a sample of monodomain, uniformly aligned nematic elastomer (which usually implies that it has been crosslinked in the aligned nematic phase [34,35]) is stretched along the axis perpendicular to the nematic director $\boldsymbol{n}_0$, the director will switch and point along the axis of uniaxial extension. The early theory (ignoring the effect of soft elasticity) [6] has predicted, and the experiment on polyacrylate LCE [43] confirmed

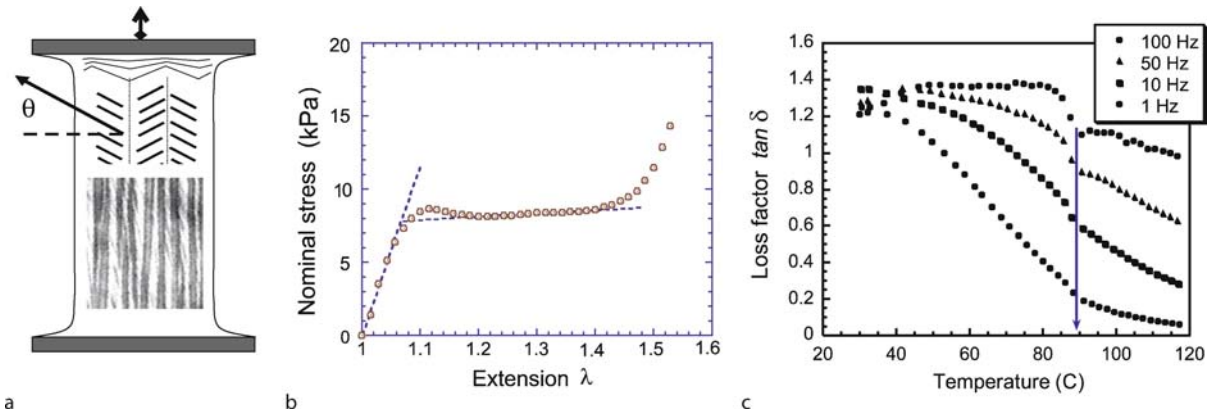

Anisotropic Networks, Elastomers and Gels, Figure 6

a The scheme of stripe domains in nematic elastomers (*inset* is their image from a polarizing microscope). b The stress-strain curve during the stripe domain regime: After a threshold strain, the soft plateau extends over the whole region where the internal director rotation takes place ($\theta \in 0 \div 90°$). c The effect of anomalous vibration damping, demonstrated by the consistently high loss factor across the whole range of nematic phase; the arrow points at the onset of the isotropic phase where the high damping disappears

that this switching may occur in an abrupt discontinuous fashion when the natural long dimension of anisotropic polymer chains can fit into the new shape of the sample, much extended in the perpendicular direction. However the same experiment performed on a polysiloxane LCE [31] has shown an unexpected stripe domain pattern. Further investigation has proven that the nematic director rotates continuously from $\boldsymbol{n}_0$ towards the new perpendicular axis, over a substantial range of deformations, but the direction of this rotation alternates in semi-regular stripes of several microns width oriented along the stretching direction, Fig. 6a. Later the same textures have been observed by other groups and on different materials, including polyacrylates [32,73], although there also have been reports confirming the single director switching mode [46].

Theoretical description of stripe domains [66] has been very successful and has led to several long-reaching consequences. First of all, this phenomenon has become a conclusive proof of the soft elasticity principles. As such, the nematic elastomers have been recognized as part of a greater family of elastic systems with microstructure (shape memory alloys being the most famous member of this family to date), all exhibiting similar phenomena. Finally, the need to understand the threshold value of extension λ_c has led to deeper understanding of internal microscopic constraints in LCE and resulted in introduction of the concept of semi-softness: A class of deformations that has the "soft" symmetry, as in Eq. (6), penalized by a small but physically distinct elastic energy due to such random constraints.

The main result of theoretical model [66] gives the director angle variation with strain,

$$\theta(\lambda) = \pm \arcsin\left[\frac{\ell_\parallel}{\ell_\parallel - \ell_\perp}\left(1 - \frac{\lambda_c^2}{\lambda^2}\right)\right]^{1/2} \tag{7}$$

with only one free parameter, the threshold strain λ_c. In fact, if one uses the pure Trace formula (1), the threshold is not present at all, $\lambda_c = 1$. The backbone chain anisotropy $\ell_\parallel/\ell_\perp$, which enters the theory, is an independent experimentally accessible quantity related, e. g. to the spontaneous shape change of LCE on heating it into the isotropic phase, $\lambda_m \approx (\ell_\parallel/\ell_\perp)^{1/3}$ in Fig. 2. This allowed the data for director angle to be collapsed onto the same master curve, spanning the whole range of non-linear deformations.

The physical reason for the stretched LCE to break into stripe domains with the opposite director rotation $\pm\theta(\lambda)$ becomes clear when one recalls the idea of soft elasticity [44,67]. The polymer chains forming the network are anisotropic, in most cases having the average shape of uniaxial prolate ellipsoid, see Fig. 2. If the assumption of affine deformation is made, the strain applied to the whole sample is locally applied to all network strands. The origin of (entropic) rubber elasticity is the corresponding change of shape of the chains, away from their equilibrium shape frozen at network formation, which results in the reduction in entropy and rise in the elastic free energy. However, the nematic (e. g. prolate anisotropic) chains may find another way of accommodating the deformation: If the sample is stretched perpendicular to the director $\boldsymbol{n}$ (the

long axis of chain gyration volume), the chains may rotate their *undeformed* ellipsoidal shapes – thus providing an extension, but necessarily in combination with simple shear – and keep their entropy constant and elastic free energy zero! This, of course, is unique to nematic elastomers: Isotropic chains (with spherical shape) have to deform to accommodate any deformation. It is also important that with deformations costing no elastic energy, there is no need for the material to change its preferred nematic order parameter Q, so the ratio $\ell_\parallel/\ell_\perp$ remains constant for the duration of soft deformations (this is not the case when, e. g., one stretched the elastomer along the director axis). The physical explanation of stripe domains is now clear: the stretched nematic network attempts to follow the soft deformation route to reduce its elastic energy, but this requires a global shear deformation which is prohibited by rigid clamps on two opposite ends of the sample, Fig. 6a. The resolution of this macroscopic constraint is to break into small parallel stripes, each with a completely soft deformation (and a corresponding shear) but with the sign of director rotation (and thus – the shear) alternating between the stripes. Then there is no global shear and the system can lower its elastic energy in the bulk, although it now has to pay the penalty for domain walls and for the non-uniform region of deformation near the clamps. The balance of gains and losses determines the domain size d.

The argument above seems to provide a reason for the threshold strain λ_c, which is necessary to overcome the barrier for creating domain walls between the "soft" stripes. However, it turns out that the numbers do not match. The energy of domain walls must be determined by the gradient Frank elasticity of the nematic (with the Frank elastic constant independently measured, $K \sim 10^{-11}$ N) and thus should be very small, since the characteristic length scale of nematic rubbers is $\xi = \sqrt{K/G} \sim 10^{-9}$ m. Hence the threshold provided by domain walls alone would be vanishingly small whereas most experiments have reported $\lambda_c \sim 1.1$ or more. This mystery has led to a whole new concept of what is now called semi-softness of LCE. The idea is that, due to several different microscopic mechanisms [65], a small addition to the classical nematic rubber-elastic free energy is breaking the symmetry required for the soft deformations:

$$W_{\mathrm{el}} \approx \frac{1}{2} G \Big[\mathrm{Tr}\big(\boldsymbol{F}^T \cdot \boldsymbol{\ell}_\theta^{-1} \cdot \boldsymbol{F} \cdot \boldsymbol{\ell}_0\big) + \alpha\,(\delta \boldsymbol{n} + \boldsymbol{n} \cdot \boldsymbol{F} \times \delta \boldsymbol{n})^2 \Big] \quad (8)$$

(usually $\alpha \ll 1$). The soft-elastic pathways are still representing the low-energy deformations, but the small penalty $\sim \alpha G$ provides the threshold for stripe domain formation ($\lambda_c \sim 1 + \alpha$) and also makes the slope of the stress-strain soft-elastic plateau small but nonzero, Fig. 6b. Compare this with the ordinary extension modulus of rubber before and after the region of stripe domains. Regardless of small complications of semi-soft corrections, the main advance in identifying the whole class of special low-energy soft deformations in LCE and proving their existence by direct experiment is worth noting.

The presence of the underlying nematic order, with its director capable of independently moving with respect to the matrix [16], leads to another remarkable feature of anisotropic elastomers and gels: the anomalous dissipation of mechanical energy, see Fig. 6c. This effect is represented in very high values of the so-called loss factor $\tan\delta = G''/G'$, representing the ratio of the imaginary and real parts of the linear modulus of response to the oscillating shear. $\tan\delta > 1$ in a fully elastic material is a very unusual situation, no doubt to be explored in applications in vibration and impact damping [15]. Another aspect of high internal dissipation due to the rotating director modes is the very high amplitude and very long times of stress relaxation of stress in all liquid crystalline elastomers and gels [14].

Swelling and Anisotropic Gels

Polymer gels are usually isotropic and swell and shrink equally in all directions. Volume phase transition phenomena of polymer gels have been extensively investigated and it is known that an infinitesimal change in an environmental intensive variable such as solvent composition, pH and temperature yields a discontinuous volume change for some polymer gels [4,27,55]. These volume transitions are driven by the balance between the repulsive and attractive forces acting on the network chains such as van der Waals, hydrophobic, ionic, hydrogen bonding [49].

If gels of anisotropic molecular structure were prepared, anisotropic mechanical behavior will be expected. Additional volume transition can be triggered by nematic ordering of anisotropic molecules inside the gel [62]. This effect adds a new driving molecular force for volume transition phenomena. Temperature-sensitive gels exhibiting volume transition at a certain temperature have attracted much attention of scientist and technologists because of their applications to drug delivery systems and sensors, originally concentrating on N-isopropylacrylamide-based isotropic systems. Spontaneously anisotropic gels undergoing a sharp and large volume change accompanied by nematic-isotropic transition is a new type of such temperature-sensitive gels, which will extend the potential applications.

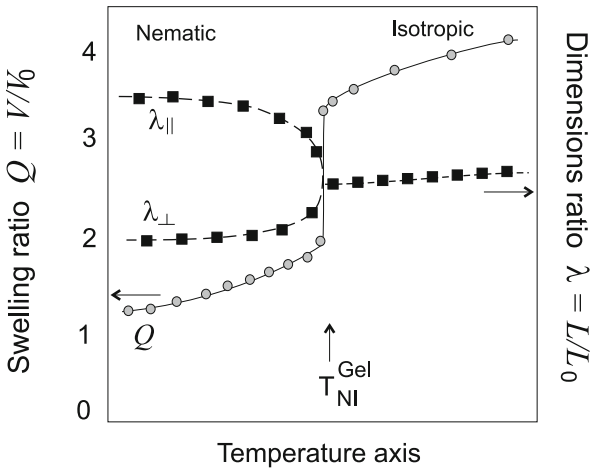

a

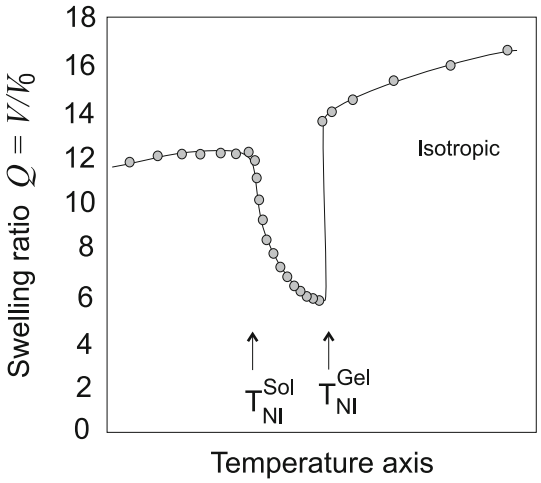

b

Anisotropic Networks, Elastomers and Gels, Figure 7
Volume transitions and anisotropic swelling of nematic gels. Plot a shows the generic effect of equilibrium swelling by isotropic solvent, at different temperatures (with T_{NI}^{Gel} marking the point of nematic-isotropic transition). The anisotropy of swelling is indicated by the growing difference between the expanding length along the nematic director (a factor $\lambda_{\parallel}$) and perpendicular to the director ($\lambda_{\perp}$), with $Q = \lambda_{\parallel}\lambda_{\perp}^2$. Plot b indicates the transitions when the network is swollen by the liquid crystalline solvent, which has its own transition temperature T_{NI}^{Sol}

Monodomain nematic networks in the dry state show a large spontaneous elongation parallel to the director axis upon nematic ordering, as discussed in earlier sections. This shape change in dry anisotropic elastomers occurs without appreciable volume change. Large anisotropic volume changes occur in such networks since the addition of isotropic solvent also has the effect of reducing the nematic order, but also remove the incompressibility constraint [29,61,72]. Upon cooling, the swollen isotropic gel continuously changes into the shrunken monodomain nematic gel around the transition point T_{NI}^{Gel}, see Fig. 7a. Importantly, the volume decrease upon nematic ordering is accompanied by the shape change of a gel with a considerable degree of anisotropy, as indicated on the generic plot. The driving force behind the abrupt and pronounced shrinking of the nematic gel is the thermodynamic force (effective chemical potential) that tries to expel the "wrong" solvent and allow the network to lower its free energy of the nematic phase. Liquid crystalline networks swollen in a liquid crystalline solvent is an even richer system, expected to act as a soft actuator driven under electric field utilizing the strong dielectric anisotropy of mesogenic units controllable by external field [30]. Figure 7b shows a generic plot of the swelling ratio $Q = V/V_0$. This plot indicates that the large shrinking occurs in the window of temperatures when the solvent is still isotropic. However, the solvent is taken into the gel once again when it becomes nematic itself, since this now promotes the order in the network and lowers the joint free energy [41,62].

In stark contrast with ordinary liquid crystals, dry anisotropic elastomers do not show any significant response to an applied electric or magnetic field. The reason is that the local elastic resistance (measured by the elastic energy density, essentially the modulus $G \sim 10^{4-6}$ Pa) is much greater than the energy density associated with local dielectric or diamagnetic anisotropy of the material. Essentially, the fields are too weak to rotate the director pinned to the elastic matrix. However, in highly swollen gels the elastic modulus is much lower and one expects (and indeed finds) an electro-optical and electro-mechanical response [10,59,63]. An additional interesting possibility is when a gel sample is mechanically unconstrained and therefore capable of finding a soft-deformation route in response to a dielectric torque applied to its local director. In this case the only resistance to dielectric torque would come from the weak semi-softness. Figure 8 shows what happens when a monodomain nematic gel, freely floating in a solvent, is subjected to an electric field perpendicular to its director [63]. The director rotates, and the shape of the gel changes accordingly, see Fig. 2c, contracting in the parallel direction but remaining constant in the perpendicular direction ($\lambda_{\perp} = 1$) as required by the soft-deformation geometry.

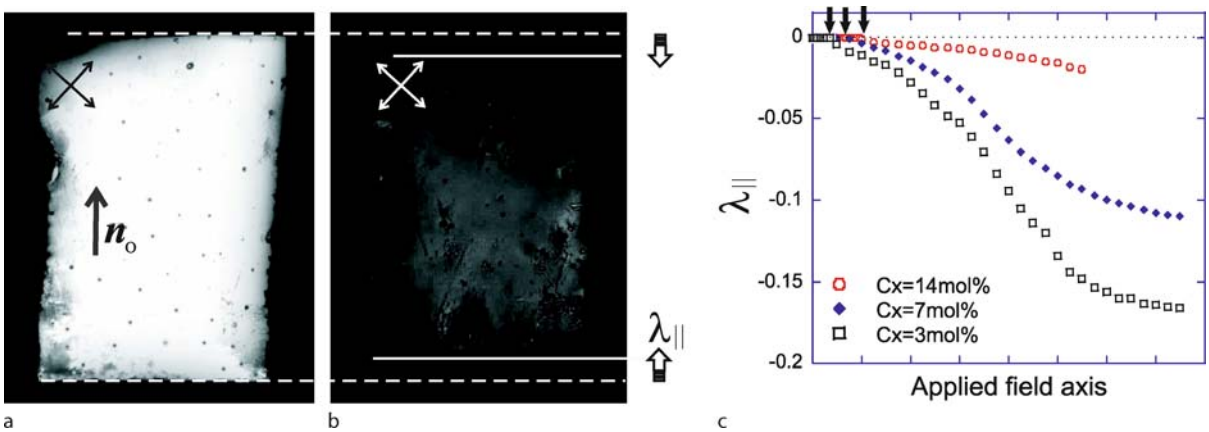

Anisotropic Networks, Elastomers and Gels, Figure 8

Response of an unconstrained nematic gel to an electric field. Image a shows the gel with the director n_0 aligned in the plane, viewed from above between crossed polars. Electric field is applied perpendicular to the image plane and the director rotates out of the plane, so that image b shows extinction between crossed polars. Note that the sample width remains constant. The associated contraction of the gel along the original director, $\lambda_{||}$ is plotted against applied field for several crosslinking density values. The *arrows* on the plot indicate the semi-soft threshold for the director rotation onset, shifting with crosslinking density. (The data by K. Urayama)

Cholesteric Elastomers: Photonics

When chiral molecular moieties (or dopant) are is added to nematic polymers, they form cholesteric phases. Crosslinked networks made from such polymers accordingly have a spontaneously twisted director distribution. Their director distribution being periodic with characteristic length scale in the optical range, they display their inhomogeneous modulation in brilliant colors of selective reflection. Being macroscopically non-centrosymmetric, such elastic materials possess correspondingly rare physical properties, such as the piezoelectricity.

As with nematic LCE, in order to access many unique properties of cholesteric elastomers one needs to form well-aligned monodomain samples, which would remain ordered due to crosslinking. As with nematic systems, the only way to achieve this is to introduce the network crosslinks when the desired phase is established by either external fields or boundary conditions – otherwise a polydomain elastomer would result. As with nematics, there are certain practical uses of cholesteric polydomain state – for instance, the applications of stereo-selective swelling (see below) may work even better in a system with micron-size cholesteric domains [17].

There are two established ways to create monodomain cholesteric elastomer films with the director in the film plane and the helical axis uniformly aligned perpendicular to the film. Each has its advantages, but also unavoidable drawbacks. Traditionally, a technique of photopolymerization (and network crosslinking when di-functional molecules are used, such as di-acrylates) was used in cells where the desired cholesteric alignment was achieved and maintained by proper boundary conditions. Many elaborate and exciting applications have been developed using this approach, e. g. [8,25]. However, it is well-known that the influence of surface anchoring only extends a maximum of few microns into the bulk of a liquid crystal (unless a main-chain mesogenic polymer, with dramatically increased planar anchoring strength is employed [57]). Therefore, only very thin cholesteric LCE films can be prepared by this surface-alignment and photopolymerization technique.

A different approach to preparing monodomain cholesteric elastomers has been developed by Kim and Finkelmann [28], utilizing the two-step crosslinking principle originally used in nematic LCE. One prepares a sample of cholesteric polymer gel, crosslinked only partially, and the deposits it on a substrate to de-swell. Since the area in contact with the substrate remains constrained, the sample cannot change its two lateral dimensions in the plane to accommodate the loss of volume; only the thickness can change. These constraints in effect apply a large uniaxial compression, equivalent to a biaxial extension in the flat plane, which serves to align the director in the plane and the helical axis perpendicular to it. After this is achieved, the full crosslinking of the network is initiated and the well-aligned free-standing cholesteric elastomer results. This is a delicate technique, which requires a careful balance of crosslinking/deswelling rates, solvent content and temperature of the phases. In princi-

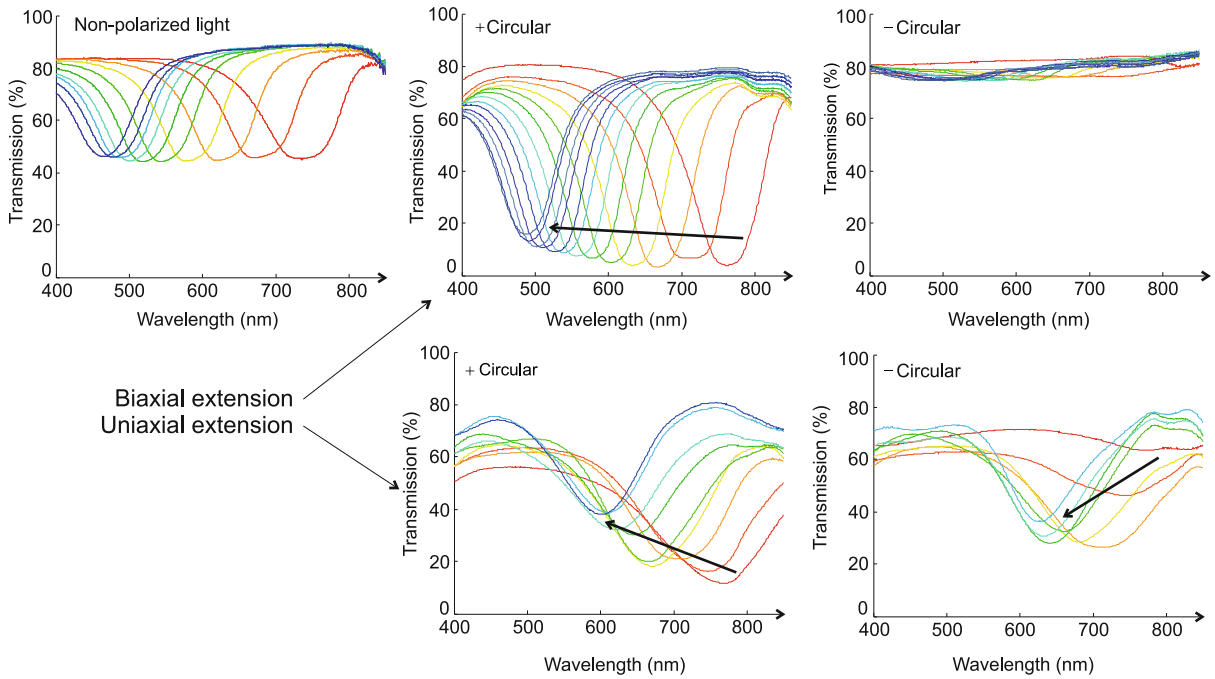

Anisotropic Networks, Elastomers and Gels, Figure 9

Selective reflection of light by cholesteric elastomers shifts under deformation. The *top row of plots* shows the results for a biaxial extension in the plane, leading to uniaxial compression along the helix from 1 to 0.65, separating the transmission of right- and left-handed circular polarized light. The *bottom row of plots* shows the analogous results for a uniaxial extension in the plane, from 1 to 1.6. The key difference to note is the emergence of the matching selective reflection for the opposite chirality of light in the uniaxial extension case [13]

ple it has no significant restrictions on the resulting elastomer area or thickness. However, there is a limitation of a different nature: one cannot achieve too long cholesteric pitch, e. g. in the visible red or infra-red range. The reason is the required high strength of chiral twisting at the stage of biaxial extension on deswelling: With no or weak twisting strength the material is close to the ordinary nematic and although the symmetric biaxial extension does confine the director to its plane, it does nothing to counter the quenched disorder of the establishing crosslinks and the resulting propensity to form a 2-dimensional (planar) polydomain state. Only at sufficiently high chiral twisting power the cholesteric propensity overcomes this planar disorder – and the resulting cholesteric pitch will end up relatively short. Another unavoidable drawback of this 2-step crosslinking technique is the disordering effect of the portion of crosslinks established at the first stage of gel formation (one has to have them to sustain elastic stress and prevent flow on deswelling), so the optical quality of samples prepared in this way is always worse than that in thin films photo-polymerized in a perfect surface-aligned cholesteric state.

Cholesteric elastomers and gels respond to imposed mechanical deformations in different ways. Bending of flexible free-standing cholesteric LCE has been used to great effect [40]. The more mechanically straightforward idea is to compress the elastomer along the helical axis, which is easiest to achieve by imposing a symmetric biaxial extension in the director plane [21]. This leads to the affine contraction of the helical pitch and the corresponding dramatic shift of the selective reflection band, see Fig. 9. If a more traditional uniaxial stretching of elastomer fils is applied, with the initial helical pitch perpendicular to the direction of stretching, the modified director distribution is usually no longer simply helical although the characteristic length scales also affinely scale with sample dimensions. On deformation the director texture remains periodic along the former helical pitch, but becomes a non-chiral step-wise modulation [12,70], leading to new photonic bandgaps and selective reflection in both right- and left-handed circular polarized light, Fig. 9. Laser emission has been predicted and observed in cholesteric systems when light was excited in their bulk, e. g. by a photo-dye stimulated by a pumping laser emit-

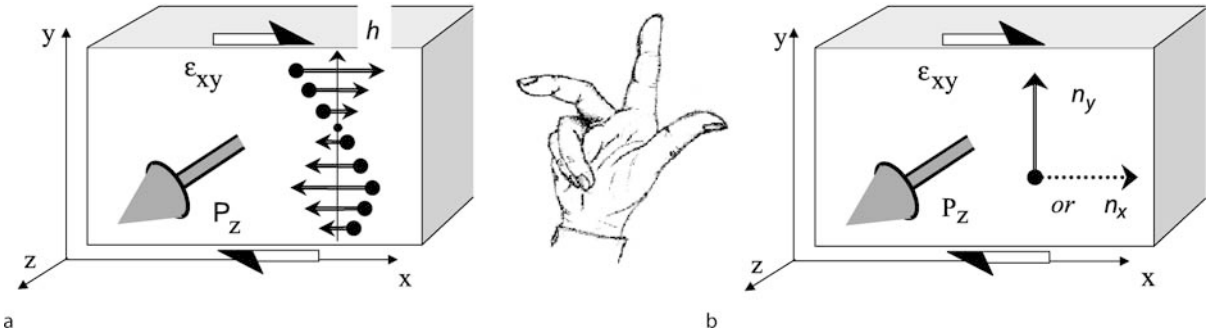

Anisotropic Networks, Elastomers and Gels, Figure 10

Piezoelectric effects in chiral elastomers. a Polarization induced by the shear applied to helically twisted textures (flexoelectric effect); b Polarization induced by shearing a uniformly aligned chiral nematic, with the director either along *x*- or *y*-axes (the true piezoelectric effect $\mathbf{P} = \gamma[\mathbf{n} \times (\boldsymbol{\varepsilon} \cdot \mathbf{n})]$). The sketch illustrates the chiral geometry that produces a polarization vector along *z* due to the shear and anisotropy axis lying in the *x*-*y* plane

ting near the bandgap edge. The ability to continuously change the position of bandgap by mechanically deforming cholesteric elastomers has led to an attractive application opportunity of tunable lasers [21,40,47].

Another possibility is topological imprinting of helical director distribution. Being anchored to the rubber matrix due to crosslinking, the director texture can remain macroscopically chiral even when all chiral dopants are removed from the material [18,39]. Internally stored mechanical twist can cause cholesteric elastomers to interact selectively with solvents according to its imprinted handedness. Separation of racemic mixtures by stereo-selective swelling is an interesting application possibility.

The lack of centro-symmetric invariance in chiral elastic materials leads to interesting and rare physical properties [56], in particular, piezoelectric effect and non-linear optical properties. Such characteristics of elastomers with a chiral smectic C* order (the ferroelectric liquid crystalline elastomers, FLCE) have been studied with some intensity for several years now. After the permanently monodomain (fixed by crosslinking) free-standing samples of FLCE were prepared [20,23], several useful experiments targeting various electro-mechanical and electro-optical properties, in particular – the piezoelectricity and the non-linear optical response, have been reported in recent years [7,36]. Clearly, the prospect of important applications will continue driving this work. In this short review we shall concentrate on the analogous effect in chiral nematic LCE which do not possess a spontaneous polarization.

With much higher symmetry of nematic and cholesteric elastomers (the point group D_∞ in a chiral material, in contrast to a very low symmetry: C_2 plus the translational effect of layers in ferroelectric smectic C*), there is a real possibility to identify microscopic origins of piezoelectricity in amorphous polymers or indeed elastomers, if one aims to have an equilibrium effect in a stress-resistant material. The piezoelectric effect in a uniform chiral nematic LCE has been described phenomenologically [56] and by a fully non-linear microscopic theory similar in its spirit to the basic Trace formula for nematic rubber elasticity [58]. All experimental research so far has concentrated on the helically twisted cholesteric elastomers [11,42,64]. However, the cholesteric texture under the required shear deformation [45], Fig. 10a, will produce highly non-uniform distortions giving rise to the well-understood flexoelectric effect and masking the possible chiral piezoelectricity.

The uniform linear piezoelectricity, i. e. the polarization induced by a uniform strain, Fig. 10b, (with the small deformation $\varepsilon = \lambda - 1$), is unknown in completely amorphous polymers and rubbers. Even the famous PVDF polymer-based piezoelectric has its response due to inhomogeneity of crystalline regions affected by macroscopic deformation. The molecular theory [58] has examined the effect of chirality in molecular structure of chain monomers and the bias in their short-axis alignment when the chains are stretched at an angle to the average nematic director $\boldsymbol{n}$. If the monomers possess a transverse dipole moment, this bias leads to macroscopic polarization:

$$\mathbf{P} \simeq -\frac{1}{2}(n_c \Delta)\underline{\underline{\epsilon}} : \left(\boldsymbol{\lambda}^T \cdot \boldsymbol{\ell}_\theta^{-1} \cdot \boldsymbol{\lambda} \cdot \boldsymbol{\ell}_0\right) . \tag{9}$$

Here, as in the underlying rubber elasticity, n_c is the number of network strands per unit volume (the crosslinking density) and the parameter Δ is the measure of monomer chirality with the transverse dipole moment d_t, see [58] for detail. Compare this with the original Trace formula

of Eq. (1) to note the characteristic "sandwiching" of strain between the two different orientation tensors. This expression involves the full deformation gradient tensor $\boldsymbol{\lambda}$ and, therefore can describe large deformations of a chiral nematic rubber. When shear deformations are small, the linear approximation of (9) gives, for a symmetric shear,

$$\mathbf{P} \simeq \gamma[\mathbf{n} \times (\boldsymbol{\varepsilon} \cdot \mathbf{n})] ,$$

with the linear coefficient $\gamma = \partial P/\partial \varepsilon \approx -\frac{1}{2} n_c \Delta(\ell_\parallel^2 - \ell_\perp^2)/\ell_\parallel \ell_\perp$ clearly proportional to the degree of local nematic order through the chain anisotropy $\ell_\parallel - \ell_\perp$. Piezoelectricity in amorphous rubbers is not only interesting from the point of view of fundamentals of chiral polymer random walks and symmetry breaking. On the practical side, due to the rubber modulus much lower than in ordinary solid piezoelectrics (typically $G \sim 10^5$ Pa), the relevant coefficient $d = \partial P/\partial \sigma = \gamma/G$ is much higher than the corresponding response to stress in, for instance, quartz or PVDF. The corresponding low mechanical impedance should make the piezoelectric rubber attractive for many energy transducing applications.

New Frontiers

When positional ordering, in the form of layering, is added to the orientational order of nematics, we have the smectic phase, Fig. 1c,d. When made into networks, they remain locally liquid-like, can suffer large extensions and have mobile internal degrees of freedom – much like with translationally homogeneous nematic elastomers and gels. This freedom, which gave spontaneous distortions and soft elasticity in nematics, is restricted in smectics by the layers to which the matrix is often strongly coupled. The mechanics of smectic elastomers [38,51] is decidedly more complex than that of nematic elastomers. In many cases the effect of layers means that the material has the nonlinear mechanics of a 2D rubber in the layer plane, while perform as a 1-dimensional crystalline solid in the third dimension [2] when the layer translation with respect to the rubbery matrix is penalized by a periodic potential provided by the lamellar order. Alternatively, concentrating on the layer structure, one finds the behavior radically altered from the liquid crystal case. The layers interact via the effective background provided by the elastomer network and, as a result, no longer display the celebrated Landau–Peierls loss of long range order in one- or two- dimensions. The layer buckling instabilities occur when deformations along the layer normal are applied. An even more rich system is the smectic-C elastomer and its variants. This phase has the director tilted in the layer planes and a much lower symmetry, allowing for a great variety of soft deformation modes [3,52]. The chiral variants of smectic-C networks are ferroelectric solids: Soft, greatly extensible and non-crystalline and thus of great technological significance.

This article examines some of the recent and relevant findings about the new class of materials – liquid crystalline elastomers and gels. Nematic rubbers have already proved themselves an unusual and exciting system, with a number of unique optical and mechanical properties, indeed a real example of the Cosserat media with couple-stress elasticity. Polymer networks with cholesteric order are an equally provocative system promising new physical properties.

The basic molecular theory of nematic rubber appears to be exceptionally simple in its foundation and does not involve any model parameters apart from the backbone polymer chain anisotropy $\ell_\parallel/\ell_\perp$, which can be independently measured. This is a great advantage over many other soft condensed matter systems requiring complicated, sometimes ambiguous theoretical approaches. Of course, in many real situations and materials one finds a need to look deeper into the microscopic properties; an example of this is the "semi-softness" of nematic networks. Most of these systems are characterized by non-uniform deformations: Even in the simplest experimental set-up a large portion of the sample near the clamps is subjected to non-uniform strains and, therefore, responds with a non-uniform director field.

Looking into the future, many challenging and fundamental problems in this field are still outstanding. The field of dynamics and relaxation in rubbery networks, although not young by any means, is still not offering an unambiguous physical picture of stress relaxation. Adding the liquid crystalline order, we find an additional (director) field undergoing its own relaxation process and coupled to that of an elastic network. In the particular case of polydomain (i. e. randomly disordered in equilibrium) elastomers, we can identify a hierarchical sequence of physical processes in the underlying network (above its T_g) and the superimposed glass-like nematic order. This leads to a particularly slow relaxation, but much remains to be done to understand the physics of such complex systems.

The general problem of dynamic mechanical properties, rheology and relaxation in Cosserat-like incompressible solids, also characterized by the effect of soft elasticity, brings to mind a number of possible applications. An example would be the selective attenuation of certain acoustic waves, with polarization and propagation direction satisfying the condition for softness, essentially leading to an acoustic filtering system. Another example of application of soft elasticity, also related to the problem of relaxation,

is the damping of shear vibrations in an engineering component when its surface is covered by a layer of nematic or smectic rubber, particularly aligned to allow the director rotation and softness.

Other very important area of applications is based on polarizational properties of materials with chirality. Most non-linear optical applications (which have a great technological potential) deal with elastomers in the ferroelectric smectic C^* phase. The low symmetry (in particular, chirality) and large spontaneous polarization of C^* smectics have a strong effect on the underlying elastic network, and vice versa. This also returns back to the general problem of mechanical properties of smectic rubbers and gels. In conclusion, after the initial period of exploration and material synthesis, liquid crystalline elastomers, in all their variety, now offer themselves as an exciting ground for both fundamental research and for technology. Smectic and lamellar liquid-crystalline elastomers and gels are a much more recent area of study, with much more theoretical and experimental work required to underpin their dramatically anisotropic and non-linear mechanical properties combining a two-dimensional rubber-elastic response and a solid-like properties in the third direction.

Bibliography

1. Abramchuk SS, Khokhlov AR (1987) Molecular theory of high elasticity of the polymer networks with orientational ordering of links. Doklady Akad Nauk SSSR 297:385
2. Adams JM, Warner M (2005) Elasticity of smectic-A elastomers. Phys Rev E 71:021708
3. Adams JM, Warner M (2005) Soft elasticity in smectic elastomers. Phys Rev E 72:011703
4. Annaka M, Tanaka T (1992) Multiple phases of polymer gels. Nature 355:430
5. Bhattacharya K (2003) Microstructure of martensite. Oxford University Press, Oxford
6. Bladon P, Terentjev EM, Warner M (1994) Deformation-induced orientational transitions in liquid crystal elastomers. J Phys II 4:75
7. Brehmer M, Zentel R, Giesselmann F, Germer R, Zugemaier P (1996) Coupling of liquid crystalline and polymer network properties in LC-elastomers. Liq Cryst 21:589
8. Broer DJ, Lub J, Mol GN (1995) Wide-band reflective polarizers from cholesteric polymer networks with a pitch gradient. Nature 378:467
9. Camacho-Lopez M, Finkelmann H, Palffy-Muhoray P, Shelley M (2004) Fast liquid-crystal elastomer swims into the dark. Nat Mater 3:307
10. Chang C-C, Chien L-C, Meyer RB (1997) Electro-optical study of nematic elastomer gels. Phys Rev E 56:595
11. Chang C-C, Chien L-C, Meyer RB (1997) Piezoelectric effects in cholesteric elastomer gels. Phys Rev E 55:534
12. Cicuta P, Tajbakhsh AR, Terentjev EM (2002) Evolution of photonic structure on deformation of cholesteric elastomers. Phys Rev E 65:051704
13. Cicuta P, Tajbakhsh AR, Terentjev EM (2004) Photonic bandgaps and optical activity in cholesteric elastomers. Phys Rev E 70:011703
14. Clarke SM, Terentjev EM (1998) Slow stress relaxation in randomly disordered nematic elastomers and gels. Phys Rev Lett 81:4436
15. Clarke SM, Tajbakhsh AR, Terentjev EM, Remillat C, Tomlinson GR, House JR (2001) Soft elasticity and mechanical damping in liquid crystalline elastomers. J Appl Phys 89:6530
16. Clarke SM, Tajbakhsh AR, Terentjev EM, Warner M (2001) Anomalous viscoelastic response of nematic elastomers. Phys Rev Lett 86:4044
17. Courty S, Tajbakhsh AR, Terentjev EM (2003) Phase chirality and stereo-selective swelling of cholesteric elastomers. Eur Phys J E 12:617
18. Courty S, Tajbakhsh AR, Terentjev EM (2003) Stereo-selective swelling of imprinted cholesterics networks. Phys Rev Lett 91:085503
19. Cviklinski J, Tajbakhsh AR, Terentjev EM (2002) UV-isomerisation in nematic elastomers as a route to photo-mechanical transducer. Eur Phys J E 9:427
20. Finkelmann H, Benne I, Semmler K (1995) Smectic liquid single-crystal elastomers. Macromol Symp 96:169
21. Finkelmann H, Kim ST, Munoz A, Palffy-Muhoray P, Taheri B (2001) Tunable mirrorless lasing in cholesteric liquid crystalline elastomers. Adv Mater 13:1069
22. Finkelmann H, Nishikawa E, Pereira GG, Warner M (2001) A new opto-mechanical effect in solids. Phys Rev Lett 87:015501
23. Gebhard E, Zentel R (1998) Freestanding ferroelectric elastomer films. Macromol Rapid Comm 19:341
24. Golubović L and Lubensky TC (1989) Nonlinear elasticity of amorphous solids. Phys Rev Lett 63:1082
25. Hikmet RAM, Kemperman H (1998) Electrically switchable mirrors and optical components made from liquid-crystal gels. Nature 392:476
26. Hogan PM, Tajbakhsh AR, Terentjev EM (2002) UV-manipulation of order and macroscopic shape in nematic elastomers. Phys Rev E 65:041720
27. Ilmain F, Tanaka T, Kokufuta E (1991) Volume transition in a gel driven by hydrogen-bonding. Nature 349:400
28. Kim ST, Finkelmann H (2001) Cholesteric liquid single-crystal elastomers (LSCE) obtained by the anisotropic deswelling method. Macromol Rapid Comm 22:429
29. Kishi R, Shishido M, Tazuke S (1990) Liquid-crystalline polymer gels: Anisotropic swelling of poly(gamma-benzyl L-glutamate) gel crosslinked under a magnetic field. Macromolecules 23:3868
30. Kishi R, Suzuki Y, Ichijo H, Hirasa H (1997) Electrical deformation of thermotropic liquid-crystalline polymer gels. Mol Cryst Liq Cryst 294:411
31. Kundler I, Finkelmann H (1995) Strain-induced director reorientation in nematic liquid single-crystal elastomers. Macromol Rapid Comm 16:679
32. Kundler I, Finkelmann H (1998) Director reorientation via stripe-domains in nematic elastomers. Macromol Chem Phys 199:677
33. Kutter S, Terentjev EM (2001) Tube model for the elasticity of entangled nematic rubbers. Eur Phys J E 6:221
34. Küpfer J, Finkelmann H (1991) Nematic liquid single-crystal elastomers. Macromol Rapid Comm 12:717

35. Legge CH, Davis FJ, Mitchell GR (1991) Memory effects in liquid-crystal elastomers. J Phys II 1:1253
36. Lehmann W, Gattinger P, Keck M, Kremer F, Stein P, Eckert T, Finkelmann H (1998) The inverse electromechanical effect in mechanically oriented SmC*-elastomers. Ferroelectrics 208:373
37. Li MH, Keller P, Li B, Wang XG, Brunet M (2003) Light-driven side-on nematic elastomer actuators. Adv Mater 15:569
38. Lubensky TC, Terentjev EM, Warner M (1994) Layer-network coupling in smectic elastomers. J Phys II 4:1457
39. Mao Y, Warner M (2000) Theory of chiral imprinting. Phys Rev Lett 84:5335
40. Matsui T, Ozaki R, Funamoto K, Ozaki M, Yoshino K (2002) Flexible mirrorless laser based on a free-standing film of photopolymerized cholesteric liquid crystal. Appl Phys Lett 81:3741
41. Matsuyama A, Kato T (2002) Nematic ordering-induced volume phase transitions of liquid crystalline gels. J Chem Phys 116:8175
42. Meier W, Finkelmann H (1990) Piezoelectricity of cholesteric elastomers. Macromol Chem Rapid Comm 11:599
43. Mitchell GR, Davis FJ, Guo W (1993) Strain-induced transitions in liquid-crystal elastomers. Phys Rev Lett 71:2947
44. Olmsted PD (1994) Rotational invariance and goldstone modes in nematic elastomers and gels. J Phys II 4:2215
45. Pelcovits RA, Meyer RB (1995) Piezoelectricity of cholesteric elastomers. J Phys II 5:877
46. Roberts PMS, Mitchell GR, Davis FJ (1997) A single director switching mode for monodomain liquid crystal elastomers. J Phys II 7:1337
47. Schmidtke J, Stille W, Finkelmann H (2003) Defect mode emission of a dye doped cholesteric polymer network. Phys Rev Lett 90:083902
48. Schönstein M, Stille W, Strobl G (2001) Effect of the network on the director fluctuations in a nematic side-group elastomer analysed by static and dynamic light scattering. Eur Phys J E 5:511
49. Shibayama M, Tanaka T (1993) Volume phase-transition and related phenomena of polymer gels. Adv Polym Sci 109:1
50. Stenull O, Lubensky TC (2004) Anomalous elasticity of nematic and critically soft elastomers. Phys Rev E 69:021807
51. Stenull O, Lubensky TC (2005) Phase transitions and soft elasticity of smectic elastomers. Phys Rev Lett 94:018304
52. Stenull O, Lubensky TC (2006) Soft elasticity in biaxial smectic and smectic-C elastomers. Phys Rev E 74:051709
53. Tabiryan N, Serak S, Dai X-M, Bunning T (2005) Polymer film with optically controlled form and actuation. Opt Express 13:7442
54. Tajbakhsh AR, Terentjev EM (2001) Spontaneous thermal expansion of nematic elastomers. Eur Phys J E 6:181
55. Tanaka T (1978) Collapse of gels and critical endpoint. Phys Rev Lett 40:820
56. Terentjev EM (1993) Phenomenological theory of non-uniform nematic elastomers: Free energy of deformations and electric field effects. Europhys Lett 23:27
57. Terentjev EM (1995) Density functional model of anchoring energy at a liquid crystalline polymer-solid interface. J Phys II 5:159
58. Terentjev EM, Warner M (1999) Piezoelectricity of chiral nematic elastomers. Eur Phys J B 8:595
59. Terentjev EM, Warner M, Bladon P (1994) Orientation of liquid crystal elastomers and gels by an electric field. J Phys II 4:667
60. Thomsen DL, Keller P, Naciri J, Pink R, Jeon H, Shenoy D, Ratna BR (2001) Liquid crystal elastomers with mechanical properties of a muscle. Macromolecules 34:5868
61. Urayama K, Arai YO, Takigawa T (2005) Volume phase transition of monodomain nematic polymer networks in isotropic solvents. Macromolecules 38:3469
62. Urayama K, Okuno Y, Nakao T, Kohjiya S (2003) Volume transition of nematic gels in nematogenic solvents. J Chem Phys 118:2903
63. Urayama K, Kondo H, Arai YO, Takigawa T (2005) Electrically driven deformations of nematic gels. Phys Rev E 71:051713
64. Vallerien SU, Kremer F, Fischer EW, Kapitza H, Zentel R, Poths H (1990) Experimental proof of piezoelectricity in cholesteric and chiral smectic C* phases of LC-elastomers. Macromol Chem Rapid Comm 11:593
65. Verwey GC, Warner M (1997) Compositional fluctuations and semisoftness in nematic elastomers. Macromolecules 30:4189
66. Verwey GC, Warner M, Terentjev EM (1996) Elastic instability and stripe domains in liquid crystalline elastomers. J Phys II 6:1273
67. Warner M, Bladon P, Terentjev EM (1994) 'Soft Elasticity' – Deformations without resistance in liquid crystal elastomers. J Phys II 4:93
68. Warner M, Terentjev EM (2007) Liquid crystal elastomers, 2nd edn. Oxford University Press, Oxford
69. Warner M, Gelling KP, Vilgis TA (1988) Theory of nematic networks. J Chem Phys 88:4008
70. Warner M, Terentjev EM, Meyer RB, Mao Y (2000) Untwisting of a cholesteric elastomer by a mechanical field. Phys Rev Lett 85:2320
71. Yu Y, Nakano M, Ikeda T (2003) Directed bending of a polymer film by light – miniaturizing a simple photomechanical system. Nature 425:145
72. Yusuf Y, Ono Y, Sumisaki Y, Cladis PE, Brand HR, Finkelmann H, Kai S (2004) Swelling dynamics of liquid crystal elastomers swollen with low molecular weight liquid crystals. Phys Rev E 69:021710
73. Zubarev ER, Talroze RV, Yuranova TI, Vasilets VN, Plate NA (1996) Influence of crosslinking conditions on the phase behavior of a polyacrylate-based liquid-crystalline elastomer. Macromol Rapid Comm 17:43

Anomalous Diffusion on Fractal Networks

Igor M. Sokolov
Institute of Physics, Humboldt-Universität zu Berlin, Berlin, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Random Walks and Normal Diffusion
Anomalous Diffusion
Anomalous Diffusion on Fractal Structures
Percolation Clusters

Scaling of PDF and Diffusion Equations on Fractal Lattices
Further Directions
Bibliography

Glossary

Anomalous diffusion An essentially diffusive process in which the mean squared displacement grows, however, not as $\langle \mathbf{R}^2 \rangle \propto t$ like in normal diffusion, but as $\langle \mathbf{R}^2 \rangle \propto t^\alpha$ with $\alpha \neq 0$, either asymptotically faster than in normal diffusion ($\alpha > 1$, superdiffusion) or asymptotically slower ($\alpha < 1$, subdiffusion).

Comb model A planar network consisting of a backbone (spine) and teeth (dangling ends). A popular simple model showing anomalous diffusion.

Fractional diffusion equation A diffusion equation for a non-Markovian diffusion process, typically with a memory kernel corresponding to a fractional derivative. A mathematical instrument which adequately describes many processes of anomalous diffusion, for example the continuous-time random walks (CTRW).

Walk dimension The fractal dimension of the trajectory of a random walker on a network. The walk dimension is defined through the mean time T or the mean number of steps n which a walker needs to leave for the first time the ball of radius R around the starting point of the walk: $T \propto R^{d_w}$.

Spectral dimension A property of a fractal structure which substitutes the Euclidean dimension in the expression for the probability to be at the origin at time t, $P(\mathbf{0}, t) \propto t^{-d_s/2}$. Defines the behavior of the network Laplace operator.

Alexander–Orbach conjecture A conjecture that the spectral dimension of the incipient infinite cluster in percolation is equal to 4/3 independently on the Euclidean dimension. The invariance of spectral dimension is only approximate and holds within 2% accuracy even in $d = 2$. The value of 3/4 is achieved in $d > 6$ and on trees.

Compact visitation A property of a random walk to visit practically all sites within the domain of the size of the mean squared displacement. The visitation is compact if the walk dimension d_w exceeds the fractal dimension of the substrate d_f.

Definition of the Subject

Many situations in physics, chemistry, biology or in computer science can be described within models related to random walks, in which a "particle" (walker) jumps from one node of a network to another following the network's links. In many cases the network is embedded into Euclidean space which allows for discussing the displacement of the walker from its initial site. The displacement's behavior is important e. g. for description of charge transport in disordered semiconductors, or for understanding the contaminants' behavior in underground water. In other cases the discussion can be reduced to the properties which can be defined for a whatever network, like return probabilities to the starting site or the number of different nodes visited, the ones of importance for chemical kinetics or for search processes on the corresponding networks. In the present contribution we only discuss networks embedded in Euclidean space.

One of the basic properties of normal diffusion is the fact that the mean squared displacement of a walker grows proportionally to time, $\mathbf{R}^2 \propto t$, and asymptotic deviations from this law are termed anomalous diffusion. Microscopic models leading to normal diffusion rely on the spatial homogeneity of the network, either exact or statistical. If the system is disordered and its local properties vary from one its part to another, the overall behavior at very large scales, larger than a whatever inhomogeneity, can still be diffusive, a property called homogenization.

The ordered systems or systems with absent long-range order but highly homogeneous on larger scales do not exhaust the whole variety of relevant physical systems. In many cases the system can not be considered as homogeneous on a whatever scale of interest. Moreover, in some of these cases the system shows some extent of scale invariance (dilatation symmetry) typical for fractals. Examples are polymer molecules and networks, networks of pores in some porous media, networks of cracks in rocks, and so on. Physically all these systems correspond to a case of very strong disorder, do not homogenize even at largest relevant scales, and show fractal properties. The diffusion in such systems is anomalous, exhibiting large deviations from the linear dependence of the mean squared displacement on time. Many other properties, like return probabilities, mean number of different visited nodes etc. also behave in a way different from the one they behave in normal diffusion. Understanding these properties is of primary importance for theoretical description of transport in such systems, chemical reactions in them, and in devising efficient search algorithms or in understanding ones used in natural search, as taking place in gene expression or incorporated into the animals' foraging strategies.

Introduction

The theory of diffusion processes was initiated by A. Fick who was interested in the nutrients' transport in a living

body and put down the phenomenological diffusion equation (1855). The microscopic, probabilistic discussion of diffusion started with the works of L. Bachelier on the theory of financial speculation (1900), of A. Einstein on the motion of colloidal particles in a quiescent fluid (the Brownian motion) (1905), and with a short letter of K. Pearson to the readers of Nature in the same year motivated by his work on animal motion. These works put forward the description of diffusion processes within the framework of random walk models, the ones describing the motion as a sequence of independent bounded steps in space requiring some bounded time to be completed. These models lead to larger scales and longer times to a typical diffusion behavior as described by Fick's laws. The most prominent property of this so-called normal diffusion is the fact that the mean squared displacement of the diffusing particle from its initial position at times much longer than the typical time of one step grows as $\langle x^2(t)\rangle \propto t$. Another prominent property of normal diffusion is the fact that the distribution of the particle's displacements tends to a Gaussian (of width $\langle x^2(t)\rangle^{1/2}$), as a consequence of the independence of different steps.

The situations described by normal diffusion are abundant, but looking closer into many other processes (which still can be described within a random walk picture, however, with only more or less independent steps, or with steps showing a broad distribution of their lengths or times) revealed that these processes, still resembling or related to normal diffusion, show a vastly different behavior of the mean squared displacement, e. g. $\langle x^2(t)\rangle \propto t^\alpha$ with $\alpha \neq 1$. In this case one speaks about anomalous diffusion [11]. The cases with $\alpha < 1$ correspond to *subdiffusion*, the ones with $\alpha > 1$ correspond to *superdiffusion*. The subdiffusion is typical for the motion of charge carriers in disordered semiconductors, for contaminant transport by underground water or for proteins' motion in cell membranes. The models of subdiffusion are often connected either with the continuous-time random walks or with diffusion in fractal networks (the topic of the present article) or with combinations thereof. The superdiffusive behavior, associated with divergent mean square displacement per complete step, is encountered e. g. in transport in some Hamiltonian models and in maps, in animal motion, in transport of particles by flows or in diffusion in porous media in Knudsen regime.

The diffusion on fractal networks (typically subdiffusion) was a topic of extensive investigation in 1980s–1990s, so that the most results discussed here can be considered as well-established. Therefore most of the references in this contribution are given to the review articles, and only few of them to original works (mostly to the pioneering publications or to newer works which didn't find their place in reviews).

Random Walks and Normal Diffusion

Let us first turn to normal diffusion and concentrate on the mean squared displacement of the random walker. Our model will correspond to a random walk on a regular lattice or in free space, where the steps of the random walker will be considered independent, identically distributed random variables (step lengths $\mathbf{s}_i$) following a distribution with a given symmetric probability density function (PDF) $p(\mathbf{s})$, so that $\langle \mathbf{s}\rangle = 0$. Let a be the root mean square displacement in one step, $\langle \mathbf{s}_i^2\rangle = a^2$. The squared displacement of the walker after n steps is thus

$$\langle r_n^2\rangle = \left\langle \left(\sum_{i=1}^{n} \mathbf{s}_i\right)^2 \right\rangle = \sum_{i=1}^{n} \langle \mathbf{s}_i^2\rangle + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} \langle \mathbf{s}_i \mathbf{s}_j\rangle \,.$$

The first sum is simply na^2, and the second one vanishes if different steps are uncorrelated and have zero mean. Therefore $\langle r_n^2\rangle = a^2 n$: The mean squared displacement in a walk is proportional to the number of steps.

Provided the mean time τ necessary to complete a step exists, this expression can be translated into the temporal dependence of the displacement $\langle r^2(t)\rangle = (a^2/\tau)t$, since the mean number of steps done during the time t is $n = t/\tau$. This is the time-dependence of the mean squared displacement of a random walker characteristic for normal diffusion. The prefactor a^2/τ in the $\langle r^2(t)\rangle \propto t$ dependence is connected with the usual diffusion coefficient K, defined through $\langle r^2(t)\rangle = 2dKt$, where d is the dimension of the Euclidean space. Here we note that in the theory of random walks one often distinguishes between the situations termed as genuine "walks" (or velocity models) where the particle moves, say, with a constant velocity and changes its direction at the end of the step, and "random flight" models, when steps are considered to be instantaneous ("jumps") and the time cost of a step is attributed to the waiting time at a site before the jump is made. The situations considered below correspond to the flight picture.

The are three important postulates guaranteeing that random walks correspond to a normal diffusion process:

- The existence of the mean squared displacement per step, $\langle \mathbf{s}^2\rangle < \infty$
- The existence of the mean time per step (interpreted as a mean waiting time on a site) $\tau < \infty$ and
- The uncorrelated nature of the walk $\langle \mathbf{s}_i \mathbf{s}_j\rangle = \langle \mathbf{s}^2\rangle \delta_{ij}$.

These are exactly the three postulates made by Einstein in his work on Brownian motion. The last postulate can be

weakened: The persistent (i. e. correlated) random walks still lead to normal diffusion provided $\sum_{j=i}^{\infty}\langle \mathbf{s}_i \mathbf{s}_j\rangle < \infty$.

In the case when the steps are independent, the central limit theorem immediately states that the limiting distribution of the displacements will be Gaussian, of the form

$$P(\mathbf{r}, t) = \frac{1}{(4\pi K t)^{d/2}} \exp\left(-\frac{\mathrm{d}r^2}{4Kt}\right) .$$

This distribution scales as a whole: rescaling the distance by a factor of λ, $x \to \lambda x$ and of the time by the factor of λ^2, $t \to \lambda^2 t$, does not change the form of this distribution. The moments of the displacement scale according to $\langle \mathbf{r}^k\rangle = \int P(r, t)\mathbf{r}^k d^d\mathbf{r}^k = \mathrm{const}(k)t^{k/2}$. Another important property of the the PDF is the fact that the return probability, i. e. the probability to be at time t at the origin of the motion, $P(\mathbf{0}, t)$, scales as

$$P(\mathbf{0}, t) \propto t^{-d/2} , \tag{1}$$

i. e. shows the behavior depending on the substrate's dimension.

The PDF $P(\mathbf{r}, t)$ is the solution of the Fick's diffusion equation

$$\frac{\partial}{\partial t} P(\mathbf{r}, t) = k\Delta P(\mathbf{r}, t) , \tag{2}$$

where Δ is a Laplace operator. Equation 2 is a parabolic partial differential equation, so that its form is invariant under the scale transformation $t \to \lambda^2 t, r \to \lambda r$ discussed above. This invariance of the equation has very strong implications. If follows, for example, that one can invert the scaling relation for the mean squared displacement

$$R^2 \propto t$$

and interpret this inverse one

$$T \propto r^2$$

as e. g. a scaling rule governing the dependence of a characteristic passage time from some initial site 0 to the sites at a distance r from this one. Such scaling holds e. g. for the mean time spent by a walk in a prescribed region of space (i. e. for the mean first passage time to its boundary). For an isotropic system Eq. (2) can be rewritten as an equation in the radial coordinate only:

$$\frac{\partial}{\partial t} P(r, t) = K \frac{1}{r^{d-1}} \frac{\partial}{\partial r}\left[r^{d-1}\frac{\partial}{\partial r} P(r, t)\right] . \tag{3}$$

Looking at the trajectory of a random walk at scales much larger than the step's length one infers that it is self-similar in statistical sense, i. e. that its whatever portion (of the size considerably larger then the step's length) looks like the whole trajectory, i. e. it is a statistical, random fractal. The fractal (mass) dimension of the trajectory of a random walk can be easily found. Let us interpret the number of steps as the "mass" M of the object and $\langle \mathbf{r}_n^2\rangle^{1/2}$ as its "size" L. The mass dimension D is then defined as $M \propto L^D$, it corresponds to the scaling of the mass of a solid D-dimensional object with its typical size. The scaling of the mean squared displacement of a simple random walk with its number of steps suggests that the corresponding dimension $D = d_\mathrm{w}$ for a walk is $d_\mathrm{w} = 2$. The large-scale properties of the random walks in Euclidean space are universal and do not depend on exact form of the PDF of waiting times and of jumps' lengths (provided the first possesses at least one and the second one at least two moments).

This fact can be used for an interpretation of several known properties of simple random walks. Thus, random walks on a one-dimensional lattice ($d = 1$) are recurrent, i. e. each site earlier visited by the walk is revisited repeatedly afterwards, while for $d \geq 3$ they are nonrecurrent. This property can be rationalized through arguing that an intrinsically two-dimensional object cannot be "squeezed" into a one-dimensional space without self-intersections, and that it cannot fill fully a space of three or more dimensions. The same considerations can be used to rationalize the behavior of the mean number of different sites visited by a walker during n steps $\langle S_n\rangle$. $\langle S_n\rangle$ typically goes as a power law in n. Thus, on a one-dimensional lattice $\langle S_n\rangle \propto n^{1/2}$, in lattices in three and more dimensions $\langle S_n\rangle \propto n$, and in a marginal two-dimensional situation $\langle S_n\rangle \propto n/\log n$. To understand the behavior in $d = 1$ and in $d = 3$, it is enough to note that in a recurrent one-dimensional walk all sites within the span of the walk are visited by a walker at least once, so that the mean number of different sites visited by a one-dimensional walk ($d = 1$) grows as a span of the walk, which itself is proportional to a mean squared displacement, $\langle S_n\rangle \propto L = \langle \mathbf{r}_n^2\rangle^{1/2}$, i. e. $\langle S_n\rangle \propto n^{1/2} = n^{1/d_\mathrm{w}}$. In three dimensions ($d = 3$), on the contrary, the mean number of different visited sites grows as the number of steps, $\langle S_n\rangle \propto n$, since different sites are on the average visited only finite number of times. The property of random walks on one-dimensional and two-dimensional lattices to visit practically all the sites within the domain of the size of their mean square displacement is called the *compact visitation* property.

The importance of the number of different sites visited (and its moments like $\langle S_n\rangle$ or $\langle S(t)\rangle$) gets clear when one considers chemical reactions or search problems on the corresponding networks. The situation when each site of

a network can with the probability p be a trap for a walker, so that the walker is removed whenever it visits such a marked site (reaction scheme $A + B \to B$, where the symbol A denotes the walker and symbol B the "immortal" and immobile trap) corresponds to a trapping problem of reaction kinetics. The opposite situation, corresponding to the evaluation of the survival probability of an immobile A-particle at one of the network's nodes in presence of mobile B-walkers (the same $A + B \to B$ reaction scheme, however now with immobile A and mobile B) corresponds to the target problem, or scavenging. In the first case the survival probability $\Phi(N)$ for the A-particle after N steps goes for p small as $\Phi(t) \propto \langle \exp(-pS_N)\rangle$, while for the second case it behaves as $\Phi(t) \propto \exp(-p\langle S_N\rangle)$. The simple $A + B \to B$ reaction problems can be also interpreted as the ones of the distribution of the times necessary to find one of multiple hidden "traps" by one searching agent, or as the one of finding one special marked site by a swarm of independent searching agents.

Anomalous Diffusion

The Einstein's postulates leading to normal diffusion are to no extent the laws of Nature, and do not have to hold for a whatever random walk process. Before discussing general properties of such anomalous diffusion, we start from a few simple examples [10].

Simple Geometric Models Leading to Subdiffusion

Let us first discuss two simple models of geometrical structures showing anomalous diffusion. The discussion of these models allows one to gain some intuition about the emergence and properties of anomalous diffusion in complex geometries, as opposed to Euclidean lattices, and discuss mathematical notions appearing in description of such systems. In all cases the structure on which the walk takes place will be considered as a lattice with the unit spacing between the sites, and the walk corresponds to jumps from a site to one of its neighbors. Moreover, we consider the random walk to evolve in discrete time; the time t and the number of steps n are simply proportional to each other.

Diffusion on a Comb Structure The simplest example of how the subdiffusive motion can be created in a more or less complex (however yet non-fractal) geometrical structure is delivered by the diffusion on a comb. We consider the diffusion of a walker along the backbone of a comb with infinite "teeth" (dangling ends), see Fig. 1. The walker's displacement in x-direction is only possible when it is on the backbone; entering the side branch switches

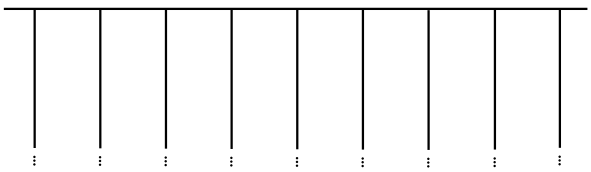

Anomalous Diffusion on Fractal Networks, Figure 1
A comb structure: Trapping in the dangling ends leads to the anomalous diffusion along the comb's backbone. This kind of anomalous diffusion is modeled by CTRW with power-law waiting time distribution and described by a fractional diffusion equation

off the diffusion along the backbone. Therefore the motion on the backbone of the comb is interrupted by waiting times. The distribution of these waiting times is given by the power law $\psi(t) \propto t^{-3/2}$ corresponding to the first return probability of a one-dimensional random walk in a side branch to its origin. This situation – simple random walk with broad distribution of waiting times on sites – corresponds to a continuous-time random walk model. In this model, for the case of the power-law waiting time distributions of the form $\psi(t) \propto t^{-1-\alpha}$ with $\alpha < 1$, the mean number of steps $\langle n(t)\rangle$ (in our case: the mean number of steps over the backbone) as a function of time goes as $\langle n(t)\rangle \propto t^{\alpha}$. Since the mean squared displacement along the backbone is proportional to this number of steps, it is also governed by $\langle x^2(t)\rangle \propto t^{\alpha}$, i. e. in our case we have to do with strongly subdiffusive behavior $\langle x^2(t)\rangle \propto t^{1/2}$.

The PDF of the particle's displacements in CTRW with the waiting time distribution $\psi(t)$ is governed by a non-Markovian generalization of the diffusion equation

$$\frac{\partial}{\partial t}P(x,t) = a^2 \frac{\partial}{\partial t}\Delta \int_0^{\infty} M(t-t')P(x,t') \,,$$

with the memory kernel $M(t)$ which has a physical meaning of the time-dependent density of steps and is given by the inverse Laplace transform of $\tilde{M}(u) = \tilde{\psi}(u)/[1-\tilde{\psi}(u)]$ (with $\tilde{\psi}(u)$ being the Laplace-transform of $\psi(t)$). For the case of a power-law waiting-time probability density of the type $\psi(t) \propto t^{-1-\alpha}$ the structure of the right-hand side of the equation corresponds to the *fractional Riemann–Liouvolle derivative* ${}_0D_t^{\beta}$, an integro-differential operator defined as

$${}_0D_t^{\beta} f(t) = \frac{1}{\Gamma(1-\beta)} \frac{\mathrm{d}}{\mathrm{d}t} \int_0^t \mathrm{d}t' \frac{f(t')}{(t-t')^{\beta}} \,, \tag{4}$$

(here only for $0 < \beta < 1$), so that in the case of the diffusion on a backbone of the comb we have

$$\frac{\partial}{\partial t}P(x,t) = K_0 D_t^{1-\alpha} \Delta P(x,t) \,, \tag{5}$$

(here with $\alpha = 1/2$). An exhaustive discussion of such equations is provided in [13,14]. Similar fractional diffusion equation was used in [12] for the description of the diffusion in a percolation system (as measured by NMR). However one has to be cautious when using CTRW and corresponding fractional equations for the description of diffusion on fractal networks, since they may capture some properties of such diffusion and fully disregard other ones. This question will be discussed to some extent in Sect. "Anomalous Diffusion on Fractal Structures".

Equation (5) can be rewritten in a different but equivalent form

$$ {}_0^*D_t^\alpha \frac{\partial}{\partial t} P(x,t) = K \Delta P(x,t) \,, $$

with the operator ${}_0^*D_t^\alpha$, a Caputo derivative, conjugated to the Riemann–Liouville one, Eq. (4), and different from Eq. (4) by interchanged order of differentiation and integration (temporal derivative inside the integral). The derivation of both forms of the fractional equations and their interpretation is discussed in [21]. For the sake of transparency, the corresponding integrodifferential operators ${}_0D_t^\beta$ and ${}_0^*D_t^\beta$ are simply denoted as a fractional partial derivative $\partial^\beta/\partial t^\beta$ and interpreted as a Riemann–Liouville or as a Caputo derivative depending whether they stand on the left or on the right-hand side of a generalized diffusion equation.

Random Walk on a Random Walk and on a Self-Avoiding Walk Let us now turn to another source of anomalous diffusion, namely the nondecaying correlations between the subsequent steps, introduced by the structure of the possible ways on a substrate lattice.

The case where the anomalies of diffusion are due to the tortuosity of the path between the two sites of a fractal "network" is illustrated by the examples of random walks on polymer chains, being rather simple, topologically one-dimensional objects. Considering lattice models for the chains we encounter two typical situations, the simpler one, corresponding to the random walk (RW) chains and a more complex one, corresponding to self-avoiding ones. A conformation of a RW-chain of N monomers corresponds to a random walk on a lattice, self-intersections are allowed. In reality, RW chains are a reasonable model for polymer chains in melts or in θ-solutions, where the repulsion between the monomers of the same chain is compensated by the repulsion from the monomers of other chains or from the solvent molecules, so that when not real intersections, then at least very close contacts between the monomers are possible. The other case corresponds to chains in good solvents, where the effects of steric repulsion dominate. The end-to-end distance in a random walk chain grows as $R \propto l^{1/2}$ (with l being its contour length). In a self-avoiding-walk (SAW) chain $R \propto l^\nu$, with ν being a so-called Flory exponent (e. g. $\nu \approx 3/5$ in $3d$). The corresponding chains can be considered as topologically one-dimensional fractal objects with fractal dimensions $d_\mathrm{f} = 2$ and $d_\mathrm{f} = 1/\nu \approx 5/3$ respectively.

We now concentrate on an walker (excitation, enzyme, transcription factor) performing its diffusive motion on a chain, serving as a "rail track" for diffusion. This chain itself is embedded into the Euclidean space. The contour length of the chain corresponds to the *chemical coordinate* along the path, in each step of the walk at a chain the chemical coordinate changes by ± 1. Let us first consider the situation when a particle can only travel along this chemical path, i. e. cannot change its track at a point of a self-intersection of the walk or at a close contact between the two different parts of the chain. This walk starts with its first step at $\mathbf{r} = 0$. Let K be the diffusion coefficient of the walker along the chain. The typical displacement of a walker along the chain after time t will then be $l \propto t^{1/2} = t^{1/d_\mathrm{w}}$. For a RW-chain this displacement along the chain is translated into the displacement in Euclidean space according to $R = \langle \mathbf{R}^2 \rangle^{1/2} \propto l^{1/2}$, so that the typical displacement in Euclidean space after time t goes as $R \propto t^{1/4}$, a strongly subdiffusive behavior known as one of the regimes predicted by a repetition theory of polymer dynamics. The dimension of the corresponding random walk is, accordingly, $d_\mathrm{w} = 4$. The same discussion for the SAW chains leads to $d_\mathrm{w} = 1/2\nu$.

After time t the displacement in the chemical space is given by a Gaussian distribution

$$ p_\mathrm{c}(l,t) = \frac{1}{\sqrt{4\pi K t}} \exp\left(-\frac{l^2}{4Kt}\right) \,. $$

The change in the contour length l can be interpreted as the number of steps along the random-walk "rail", so that the Euclidean displacement for the given l is distributed according to

$$ p_\mathrm{E}(\mathbf{r},l) = \left(\frac{d}{2\pi a^2 |l|}\right) \exp\left(-\frac{d r^2}{2a^2|l|}\right) \,. \tag{6} $$

We can, thus, obtain the PDF of the Euclidean displacement:

$$ P(\mathbf{r},l) = \int_0^\infty p_\mathrm{E}(\mathbf{r},l) p_\mathrm{c}(l,t) \mathrm{d}l \,. \tag{7} $$

The same discussion can be pursued for a self-avoiding

walk for which

$$p_{\mathrm{E}}(\mathbf{r}, l) \simeq \left(\frac{r}{l^{\nu}}\right)^{\frac{\gamma-1}{\nu}} \exp\left[-\mathrm{const}\left(\frac{r}{l^{\nu}}\right)^{1/(1-\nu)}\right], \quad (8)$$

with γ being the exponent describing the dependence of the number of realizations of the corresponding chains on their lengths. Equation (6) is a particular case of Eq. (8) with $\gamma = 1$ and $\nu = 1/2$. Evaluating the integral using the Laplace method, we get

$$P(\mathbf{r}, t) \simeq \exp\left[-\mathrm{const}\left(\frac{r}{t^{1/d_{\mathrm{w}}}}\right)^{(1-1/d_{\mathrm{w}})^{-1}}\right]$$

up to the preexponential. We note that in this case the reason for the anomalous diffusion is the tortuosity of the chemical path. The topologically one-dimensional structure of the lattice allowed us for discussing the problem in considerable detail.

The situation changes when we introduce the possibility for a walker to jump not only to the nearest-neighbor in chemical space by also to the sites which are far away along the chemical sequence but close to each other in Euclidean space (say one lattice site of the underlying lattice far). In this case the structure of the chain in the Euclidean space leads to the possibility to jump very long in chemical sequence of the chain, with the jump length distribution going as $p(l) \propto l^{-3/2}$ for an RW and $p(l) \propto l^{-2.2}$ for a SAW in $d = 3$ (the probability of a close return of a SAW to its original path). The corresponding distributions lack the second moment (and even the first moment for RW chains), and therefore one might assume that also the diffusion in a chemical space of a chain will be anomalous. It indeed shows considerable peculiarities. If the chain's structure is frozen (i. e. the conformational dynamics of a chain is slow compared to the diffusion on the chain), the situation in both cases corresponds to "paradoxical diffusion" [22]: Although the PDF of displacements in the chemical space lacks the second (and higher) moments, the width of the PDF (described e. g. as its interquartile distance) grows, according to $W \propto t^{1/2}$. In the Euclidean space we have here to do with the diffusion on a static fractal structure, i. e. with a network of fractal dimension d_{f} coinciding with the fractal dimension of the chain. Allowing the chain conformation changing in time (i. e. considering the opposite limiting case when the conformation changes are fast enough compared to the diffusion along the chain) leads to another, superdiffusive, behavior, namely to Lévy flights along the chemical sequence. The difference between the static, fixed networks and the transient ones has to be borne in mind when interpreting physical results.

Anomalous Diffusion on Fractal Structures

Simple Fractal Lattices: the Walk's Dimension

As already mentioned, the diffusion on fractal structures is often anomalous, $\langle x^2(t)\rangle \propto t^{\alpha}$ with $\alpha \neq 1$. Parallel to the situation with random walks in Euclidean space, the trajectory of a walk is typically self similar (on the scales exceeding the typical step length), so that the exponent α can be connected with the fractal dimension of the walk's trajectory d_{w}. Assuming the mean time cost per step to be finite, i. e. $t \propto n$, one readily infers that $\alpha = 2/d_{\mathrm{w}}$. The fractal properties of the trajectory can be rather easily checked, and the value of d_{w} can be rather easily obtained for simple regular fractals, the prominent example being a Sierpinski gasket, discussed below. Analytically, the value of d_{w} is often obtained via scaling of the first passage time to a boundary of the region (taken to be a sphere of radius r), i. e. using not the relation $R \propto t^{1/d_{\mathrm{w}}}$ but the inverse one, $T \propto r^{d_{\mathrm{w}}}$. The fact that both values of d_{w} coincide witnesses strongly in favor of the self-similar nature of diffusion on fractal structures.

Diffusion on a Sierpinski Gasket Serpinski gasket (essentially a Sierpinski lattice), one of the simplest regular fractal network is a structure obtained by iterating a generator shown in Fig. 2. Its simple iterative structure and the ease of its numerical implementation made it a popular toy model exemplifying the properties of fractal networks. The properties of diffusion on this network were investigated in detail, both numerically and analytically. Since the structure does not possess dangling ends, the cause of the diffusion anomalies is connected with tortuosity of the typical paths available for diffusion, however, differently from the previous examples, these paths are numerous and non-equivalent.

The numerical method allowing for obtaining exact results on the properties of diffusion is based on exact enumeration techniques, see e. g. [10]. The idea here is to calculate the displacement probability distribution based on the exact number of ways $W_{i,n}$ the particle, starting at a given site 0 can arrive to a given site i at step n. For a given network the number $W_{i,n}$ is simply a sum of the numbers of ways $W_{j,n-1}$ leading from site 0 to the nearest neighbors j of the site i, $W_{j,n-1}$. Starting from $W_{0,0} = 1$ and $W_{i,0} = 0$ for $i \neq 0$ one simply updates the array of $W_{i,n}$ according to the rule $W_{i,n} = \sum_{j=nn(i)} W_{j,n-1}$, where $nn(i)$ denotes the nearest neighbors of the node i in the network. The probability $P_{i,n}$ is proportional to the number of these equally probably ways, and is obtained from $W_{i,n}$ by normalization. This is essentially a very old method used to solve numerically a diffusion equation

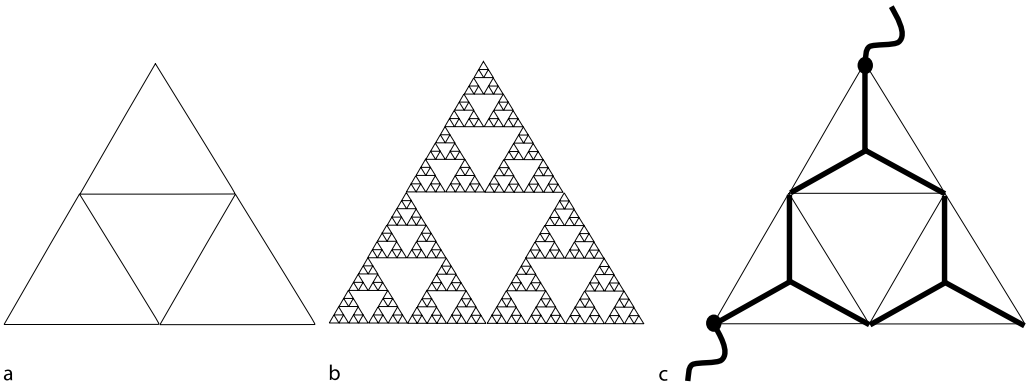

Anomalous Diffusion on Fractal Networks, Figure 2
A Sierpinski gasket: a its generator b a structure after 4th iteration. The lattice is rescaled by the factor of 4 to fit into the picture. c A constriction obtained after triangle-star transformation used in the calculation of the walk's dimension, see Subsect. "The Spectral Dimension"

in the times preceding the computer era. For regular lattices this gives us the exact values of the probabilities. For statistical fractals (like percolation clusters) an additional averaging over the realizations of the structure is necessary, so that other methods of calculation of probabilities and exponents might be superior. The value of d_w can then be obtained by plotting $\ln n$ vs. $\ln\langle \mathbf{r}_n^2\rangle^{1/2}$: the corresponding points fall on a straight line with slope d_w. The dimension of the walks on a Sierpinski gasket obtained by direct enumeration coincides with its theoretical value $d_w = \ln 5/\ln 2 = 2.322\ldots$, see Subsect. "The Spectral Dimension".

Loopless Fractals (Trees) Another situation, the one similar the the case of a comb, is encountered in loopless fractals (fractal trees), as exemplified by a Vicsek's construction shown in Fig. 3, or by diffusion-limited aggregates (DLA). For the structure depicted in Fig. 3, the walk's dimension is $d_w = \log 15/\log 3 = 2.465\ldots$ [3]. Parallel to the case of the comb, the main mechanism leading to subdiffusion is trapping in dangling ends, which, different from the case of the comb, now themselves have a fractal structure. Parallel to the case of the comb, the time of travel from one site of the structure to another one is strongly affected by trapping. However, trapping inside the dangling end does not lead to a halt of the particle, but only to confining its motion to some scale, so that the overall equations governing the PDF might be different in the form from the fractional diffusion ones.

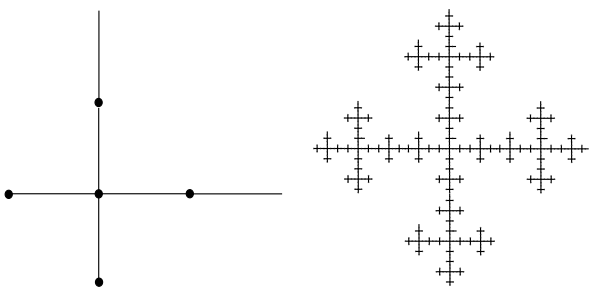

Anomalous Diffusion on Fractal Networks, Figure 3
A Vicsek fractal: the generator (*left*) and its 4th iteration. This is a loopless structure; the diffusion anomalies here are mainly caused by trapping in dangling ends

Diffusion and Conductivity

The random walks of particles on a network can be described by a master equation (which is perfectly valid for the exponential waiting time distribution on a site and asymptotically valid in case of all other waiting time distributions with finite mean waiting time τ): Let $\mathbf{p}(t)$ be the vector with elements $p_i(t)$ being the probabilities to find a particle at node i at time t. The master equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{p} = \mathbf{W}\mathbf{p} \tag{9}$$

gives then the temporal changes in this probability. A similar equation with the temporal derivative changed to $\mathrm{d}/\mathrm{d}n$ describes the n-dependence for the probabilities $p_{i,n}$ in a random walk as a function of the number of steps, provided n is large enough to be considered as a continuous variable. The matrix $\mathbf{W}$ describes the transition probabilities between the nodes of the lattice or network. The nondiagonal elements of the corresponding matrix are w_{ij}, the transition probabilities from site i to site j per unit time

or in one step. The diagonal elements are the sums of all non-diagonal elements in the corresponding column taken with the opposite sign: $w_{ii} = -\sum_j w_{ji}$, which represents the probability conservation law. The situation of unbiased random walks corresponds to a symmetric matrix $\mathbf{W}$: $w_{ij} = w_{ji}$. Considering homogeneous networks and putting all nonzero w_{ij} to unity, one sees that the difference operator represented by each line of the matrix is a symmetric difference approximation to a Laplacian.

The diffusion problem is intimately connected with the problem of conductivity of a network, as described by Kirchhoff's laws. Indeed, let us consider a stationary situation in which the particles enter the network at some particular site A at constant rate, say, one per unit time, and leave it at some site B (or a given set set of sites B). Let us assume that after some time a stationary distribution of the particles over the lattice establishes, and the particles' concentration on the sites will be described by the vector proportional to the vector of stationary probabilities satisfying

$$\mathbf{Wp} = 0 \,. \tag{10}$$

Calculating the probabilities $\mathbf{p}$ corresponds then formally to calculating the voltages on the nodes of the resistor network of the same geometry under given overall current using the Kirchhoff's laws. The conductivities of resistors connecting nodes i and j have to be taken proportional to the corresponding transition probabilities w_{ij}. The condition given by Eq. (10) corresponds then to the second Kirchhoff's law representing the particle conservation (the fact that the sum of all currents to/from the node i is zero), and the fact that the corresponding probability current is proportional to the probability difference is replaced by the Ohm's law. The first Kirchhoff's law follows from the uniqueness of the solution. Therefore calculation of the dimension of a walk can be done by discussing the scaling of conductivity with the size of the fractal object. This typically follows a power law.

As an example let us consider a Sierpinski lattice and calculate the resistance between the terminals A and B (depicted by thick wires outside of the triangle in Fig. 2c) of a fractal of the next generation, assuming that the conductivity between the corresponding nodes of the lattice of the previous generation is $R = 1$. Using the triangle-star transformation known in the theory of electric circuits, i. e. passing to the structure shown in Fig. 2c by thick lines inside the triangle, with the conductivity of each bond $r = 1/2$ (corresponding to the same value of the conductivity between the terminals), we get the resistance of a renormalized structure $R' = 5/3$. Thus, the dependence of R on the spacial scale L of the object is $R \propto L^{\zeta}$ with $\zeta = \log(5/3)/\log 2$. The scaling of the conductivity G is correspondingly $G \propto L^{-\zeta}$.

Using the flow-over-population approach, known in calculation of the mean first passage time in a system, we get $I \propto N/\langle t\rangle$, where I is the probability current through the system (the number of particles entering A per unit time), N is the overall stationary number of particles within the system and $\langle t\rangle$ is the mean time a particle spends inside the system, i. e. the mean first passage time from A to B. The mean number of particles inside the system is proportional to a typical concentration (say to the probability p_{A} to find a particle at site A for the given current I) and to the number of sites. The first one, for a given current, scales as the system's resistivity, $p_{\mathrm{A}} \propto R \propto L^{\zeta}$, and the second one, clearly, as $L^{d_{\mathrm{f}}}$ where L is the system's size. On the other hand, the mean first passage time scales according to $\langle t\rangle \propto L^{d_{\mathrm{w}}}$ (this time corresponds to the typical number of steps during which the walk transverses an Euclidean distance L), so that

$$d_{\mathrm{w}} = d_{\mathrm{f}} + \zeta \,. \tag{11}$$

Considering the d-dimensional generalizations of Sierpinski gaskets and using analogous considerations we get $\zeta = \log[(d+3)/(d+1)]/\log 2$. Combining this with the fractal dimension $d_{\mathrm{f}} = \log(d+1)/\log 2$ of a gasket gives us for the dimension of the walks $d_{\mathrm{w}} = \log(d+3)/\log 2$.

The relation between the scaling exponent of the conductivity and the dimension of a random walk on a fractal system can be used in the opposite way, since d_{w} can easily be obtained numerically for a whatever structure. On the other hand, the solutions of the Kirchhoff's equations on complex structures (i. e. the solution of a large system of algebraic equations), which is typically achieved using relaxation algorithms, is numerically much more involved.

We note that the expression for ζ can be rewritten as $\zeta = d_{\mathrm{f}}(2/d_{\mathrm{s}} - 1)$, where d_{s} is the spectral dimension of the network, which will be introduced in Subsect. "The Spectral Dimension". The relation between the walk dimension and this new quality therefore reads

$$d_{\mathrm{w}} = \frac{2d_{\mathrm{f}}}{d_{\mathrm{s}}} \,. \tag{12}$$

A different discussion of the same problem based on the Einstein's relation between diffusion coefficient and conductivity, and on crossover arguments can be found in [10,20].

The Spectral Dimension

In mathematical literature spectral dimension is defined through the return probability of the random walk, i. e. the

probability to be at the origin after n steps or at time t. We consider a node on a network and take it to be an origin of a simple random walk. We moreover consider the probability $P(\mathbf{0}, t)$ to be at this node at time t (a return probability). The spectral dimension defines then the asymptotic behavior of this probability for n large:

$$P(\mathbf{0}, t) \propto t^{-d_s/2}\,, \quad \text{for } t \to \infty\,. \tag{13}$$

The spectral dimension d_s is therefore exactly a quantity substituting the Euclidean dimension d is Eq. (1). In the statistical case one should consider an average of $p(t)$ over the origin of the random walk and over the ensemble of the corresponding graphs.

For fixed graphs the spectral dimension and the fractal (Hausdorff) dimension are related by

$$\frac{2d_f}{1+d_f} \le d_s \le d_f\,, \tag{14}$$

provided both exist [5]. The same relation is also shown true for some random geometries, see e. g. [6,7] and references therein. The bounds are optimal, i. e. both equalities can be realized in some structures.

The connection between the walk dimension d_w and the spectral dimension d_s of a network can be easily rationalized by the following consideration. Let the random walk have a dimension d_w, so that after time t the position of the walker can be considered as more or less homogeneously spread within the spatial region of the linear size $R \propto n^{1/d_w}$ or $R \propto t^{1/d_w}$, with the overall number of nodes $N \propto R^{d_f} \propto t^{d_f/d_w}$. The probability to be at one particular node, namely at $\mathbf{0}$, goes then as $1/N$, i. e. as t^{-d_f/d_w}. Comparing this with the definition of the spectral dimension, Eq. (13) we get exactly Eq. (12). The lower bound in the inequality Eq. (14) follows from Eq. (13) and from the the observation that the value of ζ in Eq. (11) is always smaller or equal to one (the value of $\zeta = 1$ would correspond to a one-dimensional wire without shunts). The upper bound can be obtained using Eq. (12) and noting that the walks' dimension never gets less than 2, its value for the regular network: the random walk on a whatever structure is more compact than one on a line.

We note that the relation Eq. (12) relies on the assumption that the spread of the walker within the r-domain is homogeneous, and needs reinterpretation for strongly anisotropic structures like combs, where the spread in teeth and along the spine are very different (see e. g. [2]). For the planar comb with infinite teeth $d_f = 2$, and d_s as calculated through the return time is $d_s = 3/2$, while the walk's dimension (mostly determined by the motion in teeth) is $d_w = 2$, like for random walks in Euclidean space. The inequality Eq. (14) is, on the contrary, universal.

Let us discuss another meaning of the spectral dimension, the one due to which it got its name. This one has to do with the description of random walks within the master equation scheme. From the spectral (Laplace) representation of the solution of the master equation, Eq. (9), we can easily find the probability $P(\mathbf{0}, t)$ that the walker starting at site $\mathbf{0}$ at $t = 0$ is found at the same site at time t. This one reads

$$P(\mathbf{0}, t) = \sum_{i=1}^{\infty} a_i \exp(-\epsilon_i t)\,,$$

where ϵ_i is the ith eigenvalue of the matrix $\mathbf{W}$ and a_i is the amplitude of its ith eigenvector at site $\mathbf{0}$. Considering the lattice as infinite, we can pass from discrete eigenvalue decomposition to a continuum

$$P(\mathbf{0}, t) = \int_0^{\infty} \mathcal{N}(\epsilon) a(\epsilon) \exp(-\epsilon t) \mathrm{d}\epsilon\,.$$

For long times, the behavior of $P(\mathbf{0}, t)$ is dominated by the behavior of $\mathcal{N}(\epsilon)$ for small values of ϵ. Here $\mathcal{N}(\epsilon)$ is the density of states of a system described by the matrix $\mathbf{W}$. The exact forms of such densities are well known for many Euclidean lattices, since the problem is equivalent to the calculating of spectrum in tight-binding approximation used in the solid state physics. For all Euclidean lattices $\mathcal{N}(\epsilon) \propto \epsilon^{d/2-1}$ for $\epsilon \to 0$, which gives us the forms of famous van Hove singularities of the spectrum. Assuming $a(\epsilon)$ to be nonsingular at $\epsilon \to 0$, we get $P(\mathbf{0}, t) \propto t^{-d/2}$. For a fractal structure, the value of d changed for the one of the spectral dimension d_s. This corresponds to the density of states $\mathcal{N}(\epsilon) \propto \epsilon^{d_s/2-1}$ which describes the properties of spectrum of a fractal analog of a Laplace operator. The dimension d_s is also often called fracton dimension of the structure, since the corresponding eigenvectors of the matrix (corresponding to eigenstates in tight-binding model) are called fractons, see e. g. [16].

In the examples of random walks on quenched polymer chains the dimension d_w was twice the fractal dimension of the underlying structures, so that the spectral dimension of the corresponding structures (without intersections) was exactly 1 just like their topological dimension.

The spectral dimension of the network governs also the behavior of the mean number of different sites visited by the random walk. A random walk of n steps having a property of compact visitation typically visits all sites within the radius of the order of its typical displacement $R_n \propto n^{1/d_w}$. The number of these sites is $S_n \propto R_n^{d_f}$ where d_f is the fractal dimension of the network, so that $S_n \propto n^{d_s/2}$ (provided $d_s \le 2$, i. e. provided the random

walk is recurrent and shows compact visitation). The spectral dimension of a structure plays important role in many other applications [16]. This and many other properties of complex networks related to the spectrum of **W** can often be easier obtained by considering random walk on the lattice than by obtaining the spectrum through direct diagonalization.

Percolation Clusters

Percolation clusters close to criticality are one of the most important examples of fractal networks. The properties of these clusters and the corresponding fractal and spectral dimensions are intimately connected to the critical indices in percolation theory.

Thus, simple crossover arguments show that the fractal dimension of an incipient infinite percolation cluster in $d \leq 6$ dimensions is $d_f = d - \beta/\nu$, where β is the critical exponent of the density of the infinite cluster, $P_\infty \propto (p - p_c)^\beta$ and ν is the critical exponent of the correlation length, $\xi \propto |p - p_c|^\nu$, and stagnates at $d_f = 4$ for $d > 6$, see e. g. [23].

On the other hand, the critical exponent t, describing the behavior of the resistance of a percolation system close to percolation threshold, $\rho \propto (p - p_c)^{-t}$, is connected with the exponent ζ via the following crossover argument. The resistance of the system is the one of the infinite cluster. For $p > p_c$ this cluster can be considered as fractal at scales $L < \xi$ and as homogeneous at larger scales. Our value for the fractal dimension of the cluster follows exactly from this argument by matching the density of the cluster $P(L) \propto L^{(d_f - d)}$ at $L = \xi \propto (p - p_c)^{-\nu}$ to its density $P_\infty \propto (p - p_c)^\beta$ at larger scales. The infinite cluster for $p > p_c$ is then a dense regular assembly of subunits of size ξ which on their turn are fractal. If the typical resistivity of each fractal subunit is R_ξ then the resistivity of the overall assembly goes as $R \propto R_\xi (L/\xi)^{2-d}$. This is a well-known relation showing that the resistivity of a wire grows proportional to its length, the resistivities of similar flat figures are the same etc. This relation holds in the homogeneous regime. On the other hand, in the fractal regime $R_L \propto L^\zeta$, so that overall $R \propto \xi^{\zeta + d - 2} \propto (p - p_c)^{-\nu(\zeta + d - 2)}$ giving the value of the critical exponent $t = \nu(\zeta + d - 2)$. The value of ζ can in its turn be expressed through the values of spectral and fractal dimensions of a percolation cluster.

The spectral dimension of the percolation cluster is very close to 4/3 in any dimension larger then one. This surprising finding lead to the conjecture by Alexander and Orbach that this value of $d_s = 4/3$ might be exact [1] (it is exact for percolation systems in $d \geq 6$ and for trees). Much effort was put into proving or disproving the conjecture, see the discussion in [10]. The latest, very accurate simulations of two-dimensional percolation by Grassberger show that the conjecture does not hold in $d = 2$, and the prediction $d_s = 4/3$ is off by around 2% [9]. In any case it can be considered as a very useful mnemonic rule.

Anomalous diffusion on percolation clusters corresponds theoretically to the most involved case, since it combines all mechanisms generating diffusion anomalies. The infinite cluster of the percolation system consists of a backbone, its main current-carrying structure, and the dangling ends (smaller clusters on all scales attached to the backbone through only one point). The anomalous diffusion on the infinite cluster is thus partly caused by trapping in this dangling ends. If one considers a shortest (chemical) way between the two nodes on the cluster, this way is tortuous and has fractal dimension larger than one.

The Role of Finite Clusters

The exponent t of the conductivity of a percolation system is essentially the characteristic of an infinite cluster only. Depending on the physical problem at hand, one can consider the situations when only the infinite cluster plays the role (like in the experiments of [12], where only the infinite cluster is filled by fluid pressed through the boundary of the system), and situations when the excitations can be found in infinite as well as in the finite clusters, as it is the case for optical excitation in mixed molecular crystals, a situation discussed in [24].

Let us concentrate on the case $p = p_c$. The structure of large but finite clusters at lower scales is indistinguishable from those of the infinite cluster. Thus, the mean squared displacement of a walker on a cluster of size L (one with $N \propto L^{d_f}$ sites) grows as $R^2(t) \propto t^{2/d_w}$ until is stagnates at the value of R of the order of the cluster's size, $R^2 \propto N^{2/d_f}$. At a given time t we can subdivide all clusters into two classes: those whose size is small compared to t^{1/d_w} (i. e. the ones with $N < t^{d_s/2}$), on which the mean square displacement stagnates, and those of the larger size, on which it still grows:

$$R^2(t) \propto \begin{cases} t^{2/d_w}, & t^{1/d_w} < N^{1/f_f} \\ N^{2/d_f}, & \text{otherwise.} \end{cases}$$

The characteristic size $N_{\text{cross}}(t)$ corresponds to the crossover between these two regimes for a given time t. The probability that a particle starts at a cluster of size N is proportional to the number of its sites and goes as $P(N) \propto Np(N) \propto N^{1-\tau}$ with $\tau = (2d\nu - \beta)/(d\nu - \beta)$, see e. g. [23]. Here $p(N)$ is the probability to find a cluster

of N sites among all clusters. The overall mean squared displacement is then given by averaging over the corresponding cluster sizes:

$$\langle r^2(t)\rangle \propto \sum_{N=1}^{N_{\text{cross}}(t)} N^{2/d_f} N^{1-\tau} + \sum_{N_{\text{cross}}(t)}^{\infty} t^{2/d_w} N^{1-\tau} .$$

Introducing the expression for $N_{\text{cross}}(t)$, we get

$$\langle r^2(t)\rangle \propto t^{2/d_w+(2-\tau)d_f/d_w} = t^{(1/d_w)(2d_f-\beta/\nu)} .$$

A similar result holds for the number of distinct sites visited, where one has to perform the analogous averaging with

$$S(t) \propto \begin{cases} t^{d_s/2} , & t^{1/d_w} < N^{1/f_f} \\ N , & \text{otherwise} \end{cases}$$

giving $S(t) \propto t^{(d_s/2)(1-\beta\nu/d_f)}$. This gives us the effective spectral dimension of the system $\tilde{d}_s = (d_s/2)(1-\beta\nu/d_f)$.

Scaling of PDF and Diffusion Equations on Fractal Lattices

If the probability density of the displacement of a particle on a fractal scales, the overall scaling form of this displacement can be obtained rather easily: Indeed, the typical displacement R during the time t goes as $r \propto t^{1/d_w}$, so that the corresponding PDF has to scale as

$$P(r,t) = \frac{r^{d_f-d}}{t^{d_s/2}} f\left(\frac{r}{t^{1/d_w}}\right) . \tag{15}$$

The prefactor takes into account the normalization of the probability density on a fractal structure (denominator) and also the scaling of the density of the fractal structure in the Euclidean space (enumerator). The simple scaling assumes that all the moments of the distance scale in the same way, so that $\langle \mathbf{r}^k\rangle = \int P(r,t)\mathbf{r}^k d^d\mathbf{r}^k = \text{const}(k)t^{k/d_w}$. The overall scaling behavior of the PDF was confirmed numerically for many fractal lattices like Sierpinski gaskets and percolation clusters, and is corroborated in many other cases by the coincidence of the values of d_w obtained via calculation of the mean squared displacement and of the mean first passage time (or resistivity of the network). The existence of the scaling form leads to important consequences e. g. for quantification of anomalous diffusion in biological tissues by characterizing the diffusion-time dependence of the magnetic resonance signal [17], which, for example, allows for differentiation between the healthy and the tumor tissues. However, even in the cases when the scaling form Eq. (15) is believed to hold, the exact form of the scaling function $f(x)$ is hard to get; the question is not yet resolved even for simple regular fractals.

For applications it is often interesting to have a kind of a phenomenological equation roughly describing the behavior of the PDF. Such an approach will be analogous to putting down Richardson's equation for the turbulent diffusion. Essentially, the problem here is to find a correct way of averaging or coarse-graining the microscopic master equation, Eq. (9), to get its valid continuous limit on larger scales. The regular procedure to do this is unknown; therefore, several phenomenological approaches to the problem were formulated.

We are looking for an equation for the PDF of the walker's displacement from its initial position and assume the system to be isotropic on the average. We look for an analog of the classical Fick's equation, $\frac{\partial}{\partial t}P(r,t) = \nabla K(r)\nabla P(r,t)$ which in spherical coordinates and for $K = K(r)$ takes the form of Eq. (3):

$$\frac{\partial}{\partial t}P(r,t) = \frac{1}{r^{d-1}}\frac{\partial}{\partial r}\left[r^{d-1}K(r)\frac{\partial}{\partial r}P(r,t)\right] .$$

On fractal lattices one changes from the Euclidean dimension of space to the fractal dimension of the lattice d_f, and takes into account that the effective diffusion coefficient $K(r)$ decays with distance as $K(r) \simeq Kr^{-\theta}$ with $\theta = d_w - 2$ to capture the slowing down of anomalous diffusion on a fractal compared to the Euclidean lattice situation. The corresponding equation put forward by O'Shaughnessy and Procaccia [18] reads:

$$\frac{\partial}{\partial t}P(r,t) = K\frac{1}{r^{d_f-1}}\frac{\partial}{\partial r}\left[r^{d_f-d_w+1}\frac{\partial}{\partial r}P(r,t)\right] . \tag{16}$$

This equation was widely used in description of anomalous diffusion on fractal lattices and percolation clusters, but can be considered only as a rough phenomenological approximation. Its solution

$$P(r,t) = A\frac{r^{d_f-d}}{t^{d_s/2}}\exp\left(-B\frac{r^{d_w}}{t}\right) ,$$

(with $B = 1/Kd_w^2$ and A being the corresponding normalization constant) corresponds exactly to the type of scaling given by Eq. (15), with the scaling function $f(x) = \exp(x^{d_w})$. This behavior of the scaling function disagrees e. g. with the results for random walks on polymer chains, for which case we had $f(x) = \exp(x^{\mu})$ with $\mu = (1-1/d_w)^{-1}$ for large enough x.

In literature, several other proposals (based on plausible assumptions but to no extent following uniquely from

the models considered) were made, taking into account possible non-Markovian nature of the motion. These are the fractional equations of the type

$$\frac{\partial^{1/d_w}}{\partial t^{1/d_w}} P(r,t) = -K_1 \frac{1}{r^{(d_s-1)/2}} \frac{\partial}{\partial r}\left[r^{(d_s-1)/2} P(r,t)\right] , \quad (17)$$

resembling "half" of the diffusion equation and containing a fractional derivative [8], as well as the "full" fractional diffusion equation

$$\frac{\partial^{2/d_w}}{\partial t^{2/d_w}} P(r,t) = K_2 \frac{1}{r^{d_s-1}} \frac{\partial}{\partial r}\left[r^{d_s-1} \frac{\partial}{\partial r} P(r,t)\right] , \quad (18)$$

as proposed in [15]. All these equation are invariant under the scale transformation

$$t \to \lambda t , \quad r \to \lambda^{1/d_w} r ,$$

and lead to PDFs showing correct overall scaling properties, see [19]. None of them reproduces correctly the PDF of displacements on a fractal (i. e. the scaling function $f(x)$) in the whole range of distances. Ref. [19] comparing the corresponding solutions with the results of simulation of anomalous diffusion an a Sierpinski gasket shows that the O'Shaughnessy–Procaccia equation, Eq. (16), performs the best for the central part of the distribution (small displacements), where Eq. (18) overestimates the PDF and Eq. (17) shows an unphysical divergence. On the other hand, the results of Eqs. (17) and (18) reproduce equally well the PDF's tail for large displacements, while Eq. (16) leads to a PDF decaying considerably faster than the numerical one. This fact witnesses favor of strong non-Markovian effects in the fractal diffusion, however, the physical nature of this non-Markovian behavior observed here in a fractal network without dangling ends, is not as clear as it is in a comb model and its more complex analogs. Therefore, the question of the correct equation describing the diffusion in fractal system (or different correct equations for the corresponding different classes of fractal systems) is still open. We note also that in some cases a simple fractional diffusion equation (with the corresponding power of the derivative leading to the correct scaling exponent d_w) gives a reasonable approximation to experimental data, as the ones of the model experiment of [12].

Further Directions

The physical understanding of anomalous diffusion due to random walks on fractal substrates may be considered as rather deep and full although it does not always lead to simple pictures. For example, although the spectral dimension for a whatever graph can be rather easily calculated, it is not quite clear what properties of the graph are responsible for its particular value. Moreover, there are large differences with respect to the degree of rigor with which different statements are proved. Mathematicians have recognized the problem, and the field of diffusion on fractal networks has become a fruitful field of research in probability theory. One of the examples of recent mathematical development is the deep understanding of spectral properties of fractal lattices and a proof of inequalities like Eq. (14).

A question which is still fully open is the one of the detailed description of the diffusion within a kind of (generalized) diffusion equation. It seems clear that there is more than one type of such equations depending on the concrete fractal network described. However even the classification of possible types of such equations is still missing.

All our discussion (except for the discussion of the role of finite clusters in percolation) was pertinent to infinite structures. The recent work [4] has showed that finite fractal networks are interesting on their own and opens a new possible direction of investigations.

Bibliography

Primary Literature

1. Alexander S, Orbach RD (1982) Density of states on fractals–fractons. J Phys Lett 43:L625–L631
2. Bertacci D (2006) Asymptotic behavior of the simple random walk on the 2-dimensional comb. Electron J Probab 45:1184–1203
3. Christou A, Stinchcombe RB (1986) Anomalous diffusion on regular and random models for diffusion-limited aggregation. J Phys A Math Gen 19:2625–2636
4. Condamin S, Bénichou O, Tejedor V, Voituriez R, Klafter J (2007) First-passage times in complex scale-invariant media. Nature 450:77–80
5. Coulhon T (2000) Random Walks and Geometry on Infinite Graphs. In: Ambrosio L, Cassano FS (eds) Lecture Notes on Analysis on Metric Spaces. Trento, CIMR, (1999) Scuola Normale Superiore di Pisa
6. Durhuus B, Jonsson T, Wheather JF (2006) Random walks on combs. J Phys A Math Gen 39:1009–1037
7. Durhuus B, Jonsson T, Wheather JF (2007) The spectral dimension of generic trees. J Stat Phys 128:1237–1260
8. Giona M, Roman HE (1992) Fractional diffusion equation on fractals – one-dimensional case and asymptotic-behavior. J Phys A Math Gen 25:2093–2105; Roman HE, Giona M, Fractional diffusion equation on fractals – 3-dimensional case and scattering function, ibid., 2107–2117
9. Grassberger P (1999) Conductivity exponent and backbone dimension in 2-*d* percolation. Physica A 262:251–263
10. Havlin S, Ben-Avraham D (2002) Diffusion in disordered media. Adv Phys 51:187–292
11. Klafter J, Sokolov IM (2005) Anomalous diffusion spreads its wings. Phys World 18:29–32

12. Klemm A, Metzler R, Kimmich R (2002) Diffusion on random-site percolation clusters: Theory and NMR microscopy experiments with model objects. Phys Rev E 65:021112
13. Metzler R, Klafter J (2000) The random walk's guide to anomalous diffusion: a fractional dynamics approach. Phys Rep 339:1–77
14. Metzler R, Klafter J (2004) The restaurant at the end of the random walk: recent developments in the description of anomalous transport by fractional dynamics. J Phys A Math Gen 37:R161–R208
15. Metzler R, Glöckle WG, Nonnenmacher TF (1994) Fractional model equation for anomalous diffusion. Physica A 211:13–24
16. Nakayama T, Yakubo K, Orbach RL (1994) Dynamical properties of fractal networks: Scaling, numerical simulations, and physical realizations. Rev Mod Phys 66:381–443
17. Özarslan E, Basser PJ, Shepherd TM, Thelwall PE, Vemuri BC, Blackband SJ (2006) Observation of anomalous diffusion in excised tissue by characterizing the diffusion-time dependence of the MR signal. J Magn Res 183:315–323
18. O'Shaughnessy B, Procaccia I (1985) Analytical solutions for diffusion on fractal objects. Phys Rev Lett 54:455–458
19. Schulzky C, Essex C, Davidson M, Franz A, Hoffmann KH (2000) The similarity group and anomalous diffusion equations. J Phys A Math Gen 33:5501–5511
20. Sokolov IM (1986) Dimensions and other geometrical critical exponents in the percolation theory. Usp Fizicheskikh Nauk 150:221–255 (1986) translated in: Sov. Phys. Usp. 29:924
21. Sokolov IM, Klafter J (2005) From diffusion to anomalous diffusion: A century after Einstein's Brownian motion. Chaos 15:026103
22. Sokolov IM, Mai J, Blumen A (1997) Paradoxical diffusion in chemical space for nearest-neighbor walks over polymer chains. Phys Rev Lett 79:857–860
23. Stauffer D (1979) Scaling theory of percolation clusters. Phys Rep 54:1–74
24. Webman I (1984) Diffusion and trapping of excitations on fractals. Phys Rev Lett 52:220–223

Additional Reading

The present article gave a brief overview of what is known about the diffusion on fractal networks, however this overview is far from covering all the facets of the problem. Thus, we only discussed unbiased diffusion (the effects of bias may be drastic due to e. g. stronger trapping in the dangling ends), and considered only the situations in which the waiting time at all nodes was the same (we did not discuss e. g. the continuous-time random walks on fractal networks), as well as left out of attention many particular systems and applications. Several review articles can be recommended as a further reading, some of them already mentioned in the text. One of the best-known sources is [10] being a reprint of the text from the "sturm und drang" period of investigation of fractal geometries. A lot of useful information on random walks models in general an on walks on fractals is contained in the review by Haus and Kehr from approximately the same time. General discussion of the anomalous diffusion is contained in the work by Bouchaud and Georges. The classical review of the percolation theory is given in the book of Stauffer and Aharony. Some additional information on anomalous diffusion in percolation systems can be found in the review by Isichenko. A classical source on random walks in disordered systems is the book by Hughes.

Haus JW, Kehr K (1987) Diffusion in regular and disordered lattices. Phys Rep 150:263–406
Bouchaud JP, Georges A (1990) Anomalous diffusion in disordered media – statistical mechanisms, models and physical applications. Phys Rep 195:127–293
Stauffer D, Aharony A (2003) Introduction to Percolation Theory. Taylor & Fransis, London
Isichenko MB (1992) Percolation, statistical topography, and transport in random-media. Rev Mod Phys 64:961–1043
Hughes BD (1995) Random Walks and random Environments. Oxford University Press, New York

Applications of Physics and Mathematics to Social Science, Introduction to

Andrzej Nowak[1], Urszula Strawińska[2]
[1] Institute for Social Studies, University of Warsaw, Warsaw, Poland
[2] Warsaw School for Social Psychology, Warsaw, Poland

The human mind is often regarded as the most complex structure in the universe. If anything could be argued to be of higher complexity it is a collection of interacting minds: social groups and societies. Kluver argues that social cognitive complexity (see ► Social Cognitive Complexity) stems from the fact that such systems consist of multiple (at least two – social and cognitive) levels that are constantly at flux due to the multiple influences both between and within levels.

The complexity of the subject matter of the social sciences has made adequate description using the traditional models of mathematics and physics difficult. For a long time it has been argued that tools of mathematics and physics are not adequate for social processes and therefore science had to proceed along two independent fronts that seemed irreconcilable (Snow, 1993). One approach, epitomized by such disciplines as sociology and anthropology, was characterized by deep insights into the nature of psychological social phenomena. These insights reflected socially constructed metrics and meanings that emphasized the contextual aspects of human experience. In the other scientific culture, epitomized in the physical sciences by precise measurement, qualitative laws and formal models defined the paradigm within which the invariant and general features of phenomena are understood.

The bridge between these two approaches consisted mainly of statistical models of psychological and social phenomena and models based on the assumption that rational choice underlies human behavior. The discoveries in the late 70s and 80s of the physical and mathematical sciences have produced a profound change in the social

sciences by providing concepts and tools that integrate the deep understanding inherent in social sciences with the precision of mathematics and physics. This perspective has provided solutions for many apparent paradoxes of social sciences, such as how altruism can emerge in self-centered individuals. It has led to surprising insights, such as the realization that the avoidance by individuals of being in the minority leads to full segregation at the level of social groups.

The core of the complexity approach into social domain is the application of mathematics and physics. These applications, described and explained by Solomon and Stauffer (see ► Applications of Physics and Mathematics to Social Science, Introduction to), are often based on the Ising model, where a binary state of an element corresponds to individual opinions (e. g. yes-no opinions), choices (e. g. cooperate or compete), or behaviors. The binary representations of individuals are arranged into structures such as a grid structure of cellular automata or networks to represent the structure of interactions of social groups and societies. The characteristics of every individual, such as opinions, depend on "neighbors": adjacent cells in a lattice, or linked nodes in a network. The tools of statistical physics, especially the mean field approach, may be used to analyze dynamics of such systems. The dynamics of such a system is often characterized by dramatic changes, phase transitions in response to small and continuous changes in the factors that influence the system (such as percolation). Physical models of social processes, described by Slanina (see ► Social Processes, Physical Models of) concentrate on the link between the micro (individual) and the macro (societies, social groups, organizations) levels of description of social reality. This approach can explain how cooperation emerges among self-centered individuals, how social classes are formed, and explains properties of behaviors stemming from bounded confidence models.

In models of opinion dynamics explained by Stauffer (see ► Opinion Dynamics and Sociophysics), individuals either move in the social space, usually seeking others with similar opinions (Schelling's model), or change their opinions in a way that depends on the opinions of the interacting partners. In a Volter model, for example, everyone takes an opinion of a random neighbor, in the majority model one adopts the opinion of the majority.

The primary tool for analyzing dynamics of complexity models of social processes are computer simulations. Simulation models described by Troitzsch (see ► Social Processes, Simulation Models of) allow one to discover dynamical properties of models that, due to their complexity and nonlinearity of interactions, may be unsolvable using analytical tools. Most computer simulations utilize agent-based models. In such models individuals are represented as agents who interact with other agents following the rules specified by the model. Models taken from physics assume extremely simple agents who are often represented by a single binary variable such as an opinion in a position in a social space. In such models the rules assumed on the level of a single agent are usually trivial; however interaction between agents leads to the emergence of complex spatial and temporal patterns on the social level. The rules specifying the behavior of a single agent may sometimes become complex. As Conte points out, an approach originating from computer science also assumes much richer representation of individuals (see ► Rational, Goal-Oriented Agents).

Complexity models of social processes are very different from traditional mathematical models. Traditional mathematical and physical models describe the temporal evolution of the model in terms of equations involving the global characteristics of systems. Yaari, Stauffer and Solomon discuss intermittency and localization, central properties of the dynamics of complexity-based models (see ► Intermittency and Localization). Most classical equations describing social and demographic processes, such as the logistic equation of growth, acquire unexpected and fascinating properties when they define local rather than global dynamics. These models produce a variety of behaviors that are observed in many real-life systems, for example in economics, sociology, biology, and ecology. According to Kelso and Oullier coordination is one of the pivotal aspects of emergent social processes (see ► Social Coordination, from the Perspective of Coordination Dynamics). Individuals not only act alone but they coordinate in the achievement of their goals. Understanding how this coordination occurs and what forms it may take allows us to gain insight into how individuals are combined into dyads and higher level social structures.

White suggests that human behavior may be analyzed at different levels (see ► Human Behavior, Dynamics of). Other people provide the critical context for actions of individuals. Moreover, the actions of individuals change the structure of society, for example by creating connections between cooperating individuals. Cohesion provides an important characteristic for group processes. Networks organize individuals into social groups and societies where the cohesion is an important characteristic of the resultant aggregates. Whereas the individual level of description may involve rules, the dynamics of aggregates may be described by equations.

Processes occurring in organizations, social groups and societies to high degree depend on structure of in-

teractions among individuals. Reed–Tsochas explains that the structure of human interactions and human relations may be described as a complex network where individuals correspond to nodes and links to social context or social relations. Most networks existing in nature as well as in societies have several common properties such as scale free distribution: there is a small number of nodes with a very high number of connections (degree) and a large number of nodes with a small number of connections. Networks tend to have a small world structure. In societies it is possible to go from one node to most of the others by traversing no more than six links. Autoaffiliate character means that similar nodes tend to be connected. There is a high chance of being connected for nodes that are similar than for nodes that are dissimilar. The network formalism offers the very tool to analyze structure of complex relations among individuals, groups and organizations in a society. These network properties have important consequences for the dynamics of processes occurring in social groups and organizations.

Complexity models are applied to different levels and within levels to different domains of social processes. Complexity models based in the neurophysiology of the human brain have provided the first precise model of the functioning of consciousness. As described by Edelman and Seth, momentary progressive coordination of activity of selective neuronal groups underlies the phenomena of consciousness (see ► Consciousness and Complexity). This precise complexity-based model suggests that consciousness is not a property of a specific brain region but an emergent, collective phenomenon. This model explains a variety of experimental results from psychology. Although affective responses are produced by specific brain regions, they may be linked to the dynamical properties of patterns of activity of the cognitive system. In particular, emotions, as described by Winkielman and Huber in ► Dynamics and Evaluation: The Warm Glow of Processing Fluency, depend on the dynamics of perception. Positive affect is elicited by the fluent processing of stimuli resulting from high coherence in the processing mechanisms, whereas negative emotions result from low fluency.

As Port points out, the dynamics of language (see ► Dynamics of Language) can be discussed either on the level of an individual or a society. At the individual level language can be understood as the skills of individuals processing speech in real time, at the social level as a common pattern of the speech of speakers in a community.

Rational rules governing the interaction of individual agents can produce moral dynamics in which the norms of a society evolve among individuals who are linked by social interdependence. In his article, Hegselman explains how rational choices may underlie the emergence of norms and the evolution of solidarity and trust (see ► Moral Dynamics).

Applications of complexity to social psychology involve dynamical models of various individual and group-level processes described by Vallacher (see ► Social Psychology, Applications of Complexity to). The attractor is the key concept for finding stability in the constant dynamics of all psychological processes. Processes oscillate around stable points which function as attractors for their dynamics. Dynamical minimalism describes the core of the approach of complexity – how to build very simple models of very complex psychological and social processes. At the abstract level many aspects of individual interactions may be described in terms of social influence.

The intimate link between complexity and the dynamics of psychological processes is most clearly seen in development. The concept of development may be captured in terms of the increasing complexity of the human system. Van Geert explains one of the main contributions of the complexity approach to development, namely the realization that mental and behavioral phenomena result, not from individual factors, but from the joint impact of many factors acting at different time scales (see ► Development, Complex Dynamic Systems of). The question of time scales goes beyond individuals to generations when one considers similarities between development and evolution. Evolutionary principles as applied to human systems, described by Bjorklund and Gardiner, define the subject matter of memetics (see ► Development, Evolution, and the Emergence of Novel Behavior). Human opinions, norms and beliefs decide on the fitness of a given individual to the social environment. Successful individuals, because of the higher fitness, have a higher chance of transmitting this property to others. Through such processes human modeling occurs. From the evolutionary point of view, the characteristics of individuals (such as opinions, norms, values) may be treated as genes and the process of transmission of these properties can be captured in terms of reproduction, which depends on fitness. As explained by Heylighen and Chielens, evolution of culture can simply be modeled by genetic algorithms resembling evolutionary processes and the approach taken by memetics (see ► Evolution of Culture, Memetics).

The complexity approach provides principles to describe the evolution of cities in time, and explain patterns of urban dynamics (see ► Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies by Batty). Comprehensive models of cities show the interrelation between various processes in

the city. Bazzani, Giorgini, and Rambaldi present traffic and crowd dynamics that provided focus for studies concerning the dynamics of urban processes (see ► Traffic and Crowd Dynamics: The Physics of the City).

Interactions within the organization of multiple people, factors, and ideas may be understood in terms of knowledge economy. As described by Marion in ► Social Organizations with Complexity Theory: A Dramatically Different Lens for the Knowledge Economy, creativity, adaptability, and learning may be described in terms of knowledge transfers and transformations within an organization.

The spread of product adoption may be described by the diffusion model. Barriers in communication between early adopters and followers may result in a dual market phenomenon, which may be analyzed using the tools of complexity modeling. Goldenberg presents models in which cellular automata describe the geographical spread of a product (see ► Marketing: Complexity Modeling, Theory and Applications in). The interdisciplinary approach allows one to develop a universal framework for modeling market penetration dynamics.

The most precise models of complexity are being developed to model economic processes. Gallegati defines agent-based models as the core of complexity modeling of economic processes (see ► Agent Based Models in Economics and Complexity). These models are usually characterized by heterogeneity, explicit space, and local interactions. Agents' behaviors are usually governed by the principles of bounded rationality in that they have limited information and therefore limited processing capacity.

Minority game, as described by Zheng (see ► Minority Games), provides a formal model of how the financial market functions. In this model, individuals in a group must make a binary choice. Those who are in the minority and choose the less popular option win. The interesting feature of the game is that any winning strategy becomes a losing strategy as soon as it is adopted by the majority of the group. It forces constant evolution of the rules governing the market, where each rule is invalidated as soon as the majority of the agents discovers it.

Agent-based modeling challenges many assumptions of neoclassical economics. Moss suggests in his critical analysis (see ► Agent Based Modeling and Neoclassical Economics: A Critical Perspective) that the neoclassical economy describes an idealized society and does not apply to any real social system in which economic decisions are made by heterogeneous interacting agents.

Andersen describes models of investment decision making in finance that are constructed by assuming different decision rules of agents trying to maximize their profit (see ► Investment Decision Making in Finance, Models of). Researchers try to understand patterns of change in financial markets in relation to simple rules of agents trying to maximize their outcomes. These models allow one to study typical patterns of market dynamics as the link between decision rules of agents and the resultant dynamics of the market indexes.

The dynamics of social and political processes are highly nonlinear and often abrupt. The state of societies is shaped to a substantial degree by rare extreme events rather than by the mass of average processes. Keilis–Borok, Soloviev, and Lichtman claim that predictability of extreme events in socio-economics and politics (see ► Extreme Events in Socio-economic and Political Complex Systems, Predictability of) is thus an important consideration for the prediction of any social, political, and economic process. New developments concerning the predictability of extreme events improve the possibility of predicting the course of dynamics of various social processes such as unemployment, recession, surges in homicides, etc.

In contrast to traditional mathematical and physical models, the approach of complexity allows us to combine deep understanding of psychological and social phenomena with the precision of the formal sciences. It provides a platform for the integration of natural and social sciences. It describes phenomena occurring at higher levels of reality as emergent effects of the interaction of lower level elements. It allows, therefore, for the integration of theories describing various levels of psychological and social reality, from the level of consciousness to high social levels such as economies, cities, and societies. In addition to its heuristic value this approach has a strong integrative potential also between different disciplines of behavioral and social sciences.

Everyone agrees that psychological and social processes display the greatest complexity of all the phenomena research can encounter. The surprising discovery of the complexity approach is that extremely complex processes may be produced by very simple rules and interactions. From this perspective simplicity is a coin flip of complexity rather than its opposition. The approach of complexity can therefore provide for simple, precise explanations of extremely complex psychological and social processes. The title of this section could be recast as Finding simplicity in complex psychological and social processes.

Acknowledgment

Preparation of this article was supported by a CO3 grant from the European NEST PATHFINDER initiative.

Artificial Chemistry

Peter Dittrich
Department of Mathematics and Computer Science, Friedrich Schiller University Jena, Jena, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Basic Building Blocks of an Artificial Chemistry
Structure-to-Function Mapping
Space
Theory
Evolution
Information Processing
Future Directions
Bibliography

Glossary

Molecular species A molecular species is an abstract class denoting an ensemble of identical molecules. Equivalently the terms "species", "compound", or just "molecule" are used; in some specific context also the terms "substrate" or "metabolite".

Molecule A molecule is a concrete instance of a molecular species. Molecules are those entities of an artificial chemistry that react. Note that sometimes the term "molecule" is used equivalently to molecular species.

Reaction network A set of molecular species together with a set of reaction rules. Formally, a reaction network is equivalent to a Petri network. A reaction network describes the stoichiometric structure of a reaction system.

Order of a reaction The order of a reaction is the sum of the exponents of concentrations in the kinetic rate law (if given). Note that an order can be fractional. If only the stoichiometric coefficients of the reaction rule are given (Eq. (1)), the order is the sum of the coefficients of the left-hand side species. When assuming mass-action kinetics, both definitions are equivalent.

Autocatalytic set A (self-maintaining) set where each molecule is produced catalytically by molecules from that set. Note that an autocatalytic may produce molecules not present in that set.

Closure A set of molecules A is closed, if no combination of molecules from A can react to form a molecule outside A. Note that the term "closure" has been also used to denote the catalytical closure of an autocatalytic set.

Self-maintaining A set of molecules is called self-maintaining, if it is able to maintain all its constituents. In a purely autocatalytic system under flow condition, this means that every molecule can be catalytically produced by at least one reaction among molecule from the set.

(Chemical) Organization A closed and self-maintaining set of molecules.

Multiset A multiset is like a set but where elements can appear more than once; that is, each element has a multiplicity, e. g., $\{a, a, b, c, c, c\}$.

Definition of the Subject

Artificial chemistries are chemical-like systems or abstract models of chemical processes. They are studied in order to illuminate and understand fundamental principles of chemical systems as well as to exploit the chemical metaphor as a design principle for information processing systems in fields like chemical computing or nonlinear optimization.

An artificial chemistry (AC) is usually a formal (and, more seldom, a physical) system that consists of objects called molecules, which interact according to rules called reactions. Compared to conventional chemical models, artificial chemistries are more abstract in the sense that there is usually not a one-to-one mapping between the molecules (and reactions) of the artificial chemistry to real molecules (and reactions). An artificial chemistry aims at capturing the logic of chemistry rather than trying to explain a particular chemical system. More formally, an artificial chemistry can be defined by a set of molecular species $\mathcal{M}$, a set of reaction rules $\mathcal{R}$, and specifications of the dynamics, e. g., kinetic laws, update algorithm, and geometry of the reaction vessel.

The scientific motivation for AC research is to build abstract models in order to understand chemical-like systems in all kind of domains ranging from physics, chemistry, biology, computer science, and sociology. Abstract chemical models to study fundamental principles, such as spatial pattern formation and self-replication, can be traced back to the beginning of modern computer science in the 40s [91,97]. Since then a growing number of approaches have tackled questions concerning the origin of complex forms, the origin of life itself [19,81], or cellular diversity [48]. In the same way a variety of approaches for constructing ACs have appeared, ranging from simple ordinary differential equations [20] to complex individual-based algebraic transformation systems [26,36,51].

The engineering motivation for AC research aims at employing the chemical metaphor as a programming and computing paradigm. Approaches can be distinguished according to whether chemistry serves as a paradigm to program or construct conventional *in-silico* information processing systems [4], Or whether real molecules are used for information processing like in molecular computing [13] or DNA computing [2] ► Cellular Computing, ► DNA Computing.

Introduction

The term artificial chemistry appeared around 1990 in the field of artificial life [72]. However, models that fell under the definition of artificial chemistry appeared also decades before, such as von Neumann's universal self-replicating automata [97]. In the 50s, a famous abstract chemical system was introduced by Turing [91] in order to show how spatial diffusion can destabilize a chemical system leading to spatial pattern formation. Turing's artificial chemistries consist of only a handful of chemical species interacting according to a couple of reaction rules, each carefully designed. The dynamics is simulated by a partial differential equation.

Turing's model possesses the structure of a conventional reaction kinetics chemical model (Subsect. "The Chemical Differential Equation"). However it does not aim at modeling a particular reaction process, but aims at exploring how, in principle, spatial patterns can be generated by simple chemical processes. Turing's artificial chemistries allow one to study whether a particular mechanism (e. g., destabilization through diffusion) can explain a particular phenomena (e. g., symmetry breaking and pattern formation).

The example illustrates that studying artificial chemistries is more a synthetic approach. That is, understanding should be gained through the synthesis of an artificial system and its observation under various conditions [45,55]. The underlying assumption is that understanding a phenomena and how this phenomena can be (re-)created (without copying it) is closely related [96].

Today's artificial chemistries are much more complex than Turing's model. They consist of a huge, sometimes infinite, amount of different molecular species and their reaction rules are defined not manually one by one, but implicitly (Sect. "Structure-to-Function Mapping") and can evolve (Sect. "Evolution"). Questions that are tackled are the structure and organization of large chemical networks, the origin of life, prebiotic chemical evolution, and the emergence of chemical information and its processing.

Introductory Example

It is relatively easy to create a complex, infinitely sized reaction network. Assume that the set of possible molecules $\mathcal{M}$ are strings over an alphabet. Next, we have to define a reaction function describing what happens when two molecules $a_1, a_2 \in \mathcal{M}$ collide. A general approach is to map a_1 to a machine, operator, or function $f = \text{fold}(a_1)$ with $f : \mathcal{M} \to \mathcal{M} \cup \{\text{elastic}\}$. The mapping fold can be defined in various ways, such as by interpreting a_1 as a computer program, a Turing machine, or a lambda expression (Sect. "Structure-to-Function Mapping"). Next, we apply the machine f derived from a_1 to the second molecule. If $f(a_2) = \text{elastic}$ the molecules collided elastically and nothing happens. This could be the case if the machine f does not halt within a predefined amount of time. Otherwise the molecules react and a new molecule $a_3 = f(a_2)$ is produced. How this new molecule changes the reaction vessel depends on the algorithm simulating the dynamics of the vessel. Algorithm 1 is a–simple, widely applied algorithm that simulates second-order catalytic reactions under flow conditions: Because the population size is constant and thus limited, there is competition and only those molecules that are somehow created will sustain.

First, the algorithm chooses two molecules randomly, which simulates a **collision**. Note that a realistic measure of time takes the amount of collision and not the amount of reaction events into account. Then, we determine stochastically whether the selected molecules react. If the molecules react, a randomly selected molecule from the population is replaced by the new product, which assures that the population size stays constant. The replaced molecules constitute the outflow. The inflow is implicitly assumed and is generated by the catalytic production of molecules. From a chemical point of view, the algorithm assumes that a product is built from an energy-rich substrate, which does not appear in the algorithm and which is assumed to be available at constant concentration.

The following section (Sect. "Basic Building Blocks of an Artificial Chemistry") describes the basic building blocks of an artificial chemistry in more detail. In Sect. "Structure-to-Function Mapping" we will explore various techniques to define reaction rules. The role of space is briefly discussed in Sect. "Space". Then the important theoretical concepts of an autocatalytic set and a chemical organization are explained (Sect. "Theory"). Sect. "Evolution" shows how artificial chemistries can be used to study (chemical) evolution. This article does not discuss information processing in detail, because there are a series of other specialized articles on that topic, which are summarized in Sect. "Information Processing".

```
INPUT: P : population (array of k molecules)

randomPosition() := randomInt(0, k-1);
reaction(r1, r2) := (fold(r1))(r2);

while not terminate do
  reactant1 := P[randomPosition()];
  reactant2 := P[randomPosition()];
  product := reaction(reactant1, reactant2);
  if product == not elastic
     P[randomPosition()] := product;
  fi
  t := t + 1/k  ;; increment simulated time
od
```

Artificial Chemistry, Algorithm 1

Basic Building Blocks of an Artificial Chemistry

Molecules

The first step in defining an AC requires that we define the set of all molecular species that can appear in the AC. The easiest way to specify this set is to enumerate explicitly all molecules as symbols. For example: $\mathcal{M} = \{H_2, O_2, H_2O, H_2O_2\}$ or, equivalently, $\mathcal{M} = \{a, b, c, d\}$. A symbol is just a name without any structure. Most conventional (bio-)chemical reaction system models are defined this way. Also many artificial chemistries where networks are designed "by hand" [20], are created randomly [85], or are evolved by changing the reaction rules [41,42,43] use only symbolic molecules without any structure.

Alternatively, molecules can posses a structure. In that case, the set of molecules is implicitly defined. For example: $\mathcal{M} = \{$all polymers that can be made from the two monomers a and b$\}$. In this case, the set of all possible molecules can even become infinite, or at least quite large. A vast variety of such rule-based molecule definitions can be found in different approaches. For example, structured molecules may be abstract character sequences [3,23,48,64], sequences of instructions [1,39], lambda-expressions [26], combinators [83], binary strings [5,16,90], numbers [8], machine code [72], graphs [78], swarms [79], or proofs [28].

We can call a molecule's representation its **structure**, in contrast to its **function**, which is given by the reaction rules $\mathcal{R}$. The description of the valid molecules and their structure is usually the first step when an AC is defined. This is analogous to a part of chemistry that describes what kind of atom configurations form stable molecules and how these molecules appear.

Reaction Rules

The set of reaction rules $\mathcal{R}$ describes how molecules from $\mathcal{M} = \{a_1, a_2, \ldots\}$ interact. A rule $\rho \in \mathcal{R}$ can be written according to the chemical notation of reaction rules in the form

$$l_{a_1,\rho} a_1 + l_{a_2,\rho} a_2 + \cdots \rightarrow r_{a_1,\rho} a_1 + r_{a_2,\rho} a_2 + \cdots . \quad (1)$$

The stoichiometric coefficients $l_{a,\rho}$ and $r_{a,\rho}$ describe the amount of molecular species $a \in \mathcal{M}$ in reaction $\rho \in \mathcal{R}$ on the left-hand and right-hand side, respectively. Together, the stoichiometric coefficients define the **stoichiometric matrix**

$$\mathbf{S} = (s_{a,\rho}) = (r_{a,\rho} - l_{a,\rho}) . \quad (2)$$

An entry $s_{a,\rho}$ of the stoichiometric matrix denotes the net amount of molecules of type a produced in reaction ρ.

A reaction rule determines a multiset of molecules (the left-hand side) that can react and subsequently be replaced by the molecules on the right-hand side. Note that the sign "+" is not an operator here, but should only separate the components on either side. The set of all molecules $\mathcal{M}$ and a set of reaction rules $\mathcal{R}$ define the **reaction network** $\langle \mathcal{M}, \mathcal{R} \rangle$ of the AC. The reaction network is equivalent to a Petri net [71]. It can be represented by two matrices, $(l_{a,\rho})$ and $(r_{a,\rho})$, made up of the stoichiometric coefficients. Equivalently we can represent the reaction network as a hyper-graph or a directed bipartite graph with two node types denoting reaction rules and molecular species, respectively, and edge labels for the stoichiometry.

A rule is applicable only if certain conditions are fulfilled. The major condition is that all of the left-hand side components must be available. This condition can

be broadened easily to include other parameters such as neighborhood, rate constants, probability for a reaction, influence of modifier species, or energy consumption. In such a case, a reaction rule would also contain additional information or further parameters. Whether or not these additional predicates are taken into consideration depends on the objectives of the artificial chemistry. If it is meant to simulate real chemistry as accurate as possible, it is necessary to integrate these parameters into the simulation system. If the goal is to build an abstract model, many of these parameters can be omitted.

Like for the set of molecules we can define the set of reaction rules explicitly by enumerating all rules symbolically [20], or we can define it implicitly by referring to the structure of the molecules. Implicit definitions of reaction rules use, for example, string matching/string concatenation [3,60,64], lambda-calculus [26,27], Turing machines [40,90], finite state machines [98] or machine code language [1,16,78,87], proof theory [27], matrix multiplication [5], swarm dynamics [79], or simple arithmetic operations [8]. Note that in some cases the reactions can emerge as the result of the interaction of many atomic particles, whose dynamics is specified by force fields or rules [79]. Section "Structure-to-Function Mapping" will discuss implicitly defined reactions in more details.

Dynamics

The third element of an artificial chemistry is its specification of how the reaction rules give rise to a dynamic process of molecules reacting in some kind of reaction vessel. It is assumed that the reaction vessel contains multiple copies of molecules, which can be seen as instances of the molecular species $\mathcal{M}$.

This section summarizes how the dynamics of a reaction vessel (which usually contains a huge number of molecules) can be modeled and simulated. The approaches can be characterized roughly by whether each molecule is treated explicitly or whether all molecules of one type are represented by a number denoting their frequency or concentration.

The Chemical Differential Equation A reaction network $\langle \mathcal{M}, \mathcal{R} \rangle$ specifies the structure of a reaction system, but does not contain any notion of time. A common way to specify the dynamics of the reaction system is by using a system of ordinary differential equations of the following form:

$$\dot{\mathbf{x}}(t) = \mathbf{S}\mathbf{v}(\mathbf{x}(t)) \,, \tag{3}$$

where $\mathbf{x} = (x_1, \ldots, x_m)^T \in \mathbb{R}^m$ is a concentration vector depending on time t, $\mathbf{S}$ the stoichiometric matrix derived from $\mathcal{R}$ (Eq. (2)), and $\mathbf{v} = (v_1, \ldots, v_r)^T \in \mathbb{R}^r$ a flux vector depending on the current concentration vector. A flux $v_\rho \geq 0$ describes the velocity or turnover rate of reaction $\rho \in \mathcal{R}$. The actual value of v_ρ depends usually on the concentration of the species participating in the reaction ρ (i. e., $\mathrm{LHS}(\rho) \equiv \{a \in \mathcal{M} \colon l_{a,\rho} > 0\}$) and sometimes on additional modifier species, whose concentration is not influenced by that reaction. There are two important assumptions that are due to the nature of reaction systems. These assumptions relate the kinetic function $\mathbf{v}$ to the reaction rules $\mathcal{R}$:

Assumption 1 A reaction ρ can only take place if all species of its left-hand side LHS(()ρ) are present. This implies that for all molecules $a \in \mathcal{M}$ and reactions $\rho \in \mathcal{R}$ with $a \in \mathrm{LHS}(\rho)$, if $x_a = 0$ then $v_\rho = 0$. The flux v_ρ must be zero, if the concentration x_a of a molecule appearing on the left-hand side of this reaction is zero. This assumption meets the obvious fact that a molecule has to be present to react.

Assumption 2 If all species $\mathrm{LHS}(\rho)$ of a reaction $\rho \in \mathcal{R}$ are present in the reactor (e. g. for all $a \in \mathrm{LHS}(\rho), x_a > 0$) the flux of that reaction is positive, (i. e., $v_\rho > 0$). In other words, the flux v_ρ must be positive, if all molecules required for that reaction are present, even in small quantities. Note that this assumption implies that a modifier species necessary for that reaction must appear as a species on the left- (and right-) hand side of ρ.

In short, taking Assumption 1 and 2 together, we demand for a chemical differential equation:

$$v_\rho > 0 \Leftrightarrow \forall a \in \mathrm{LHS}(\rho), x_a > 0 \,. \tag{4}$$

Note that Assumption 1 is absolutely necessary. Although Assumption 2 is very reasonable, there might be situations where it is not true. Assume, for example, a reaction of the form $a + a \rightarrow b$. If there is only one single molecule a left in the reaction vessel, the reaction cannot take place. Note however, if we want to model this effect, an ODE is the wrong approach and we should choose a method like those described in the following sections.

There are a large number of kinetic laws fulfilling these assumptions (Eq. (4)), including all laws that are usually applied in practice. The most fundamental of such kinetic laws is **mass-action kinetics**, which is just the product of the concentrations of the interacting molecules (cf. [32]):

$$v_\rho(\mathbf{x}) = \prod_{a \in \mathcal{M}} x_a^{l_{a,\rho}} \,. \tag{5}$$

The sum of the exponents in the kinetic law $\sum_{a \in \mathcal{M}} l_{a,\rho}$ is called the **order** of the reaction. Most more compli-

cated laws like Michaelis–Menten kinetics are derived from mass-action kinetics. In this sense, mass-action kinetics is the most general approach, allowing one to emulate all more complicated laws derived from it, but at the expense of a larger amount of molecular species.

Explicitly Simulated Collisions The chemical differential equation is a time and state continuous approximation of a discrete system where molecules collide stochastically and react. In the introduction we saw how this can be simulated by a rather simple algorithm, which is widely used in AC research [5,16,26]. The algorithm presented in the introduction simulates a second-order catalytic flow system with constant population size. It is relatively easy to extend this algorithm to include reaction rates and arbitrary orders.

First, the algorithm chooses a subset from P, which simulates a collision among molecules. Note that the resulting subset could be empty, which can be used to simulate an inflow (i. e., reactions of the form $\rightarrow A$, where the left-hand side is empty). By defining the probability of obtaining a subset of size n, we can control the rate of reactions of order n. If the molecules react, the reactands are removed from the population and the products are added. Note that this algorithm can be interpreted as a stochastic rewriting system operating on a multiset of molecules [70,88].

For large population size k, the dynamics created by the algorithm tends to the continuous dynamics of the chemical differential equation assuming mass-action kinetics. For the special case of second-order catalytic flow, as described by the algorithm in the introduction, the dynamics can be described by the catalytic network equation [85]

$$\begin{aligned} \dot{x}_k &= \sum_{i,j\in\mathcal{M}} \alpha^k_{i,j} x_i x_j - x_k \Phi(\mathbf{x}) \,, \quad \text{with} \\ \Phi(\mathbf{x}) &= \sum_{i,j,k\in\mathcal{M}} \alpha^k_{i,j} x_i x_j \,, \end{aligned} \tag{6}$$

which is a generalization of the replicator equation [35]. If the reaction function is deterministic, we can set the kinetic constants to

$$\alpha^k_{i,j} = \begin{cases} 1 & \text{reaction}(i,j) = k \,, \\ 0 & \text{otherwise} \,. \end{cases} \tag{7}$$

This dynamic is of particular interest, because competition is generated due to a limited population size. Note that the limited population size is equivalent to an unlimited population size with a limited substrate, e. g., [3].

Discrete-Event Simulation If the copy number of a particular molecular species becomes high or if the reaction rates differ strongly, it is more efficient to use discrete-event simulation techniques. The most famous technique is Gillespie's algorithm [32]. Roughly, in each step, the algorithm generates two random numbers depending on the current population state to determine the next reaction to occur as well as the time interval Δt when it will occur. Then the simulated time is advanced $t := t + \Delta t$, and the molecule count of the population updated according to the reaction that occurred.

The Gillespie algorithm generates a statistically correct trajectory of the chemical master equation. The chemical master equation is the most general way to formulate the stochastic dynamics of a chemical system. The chemical master equation is a first-order differential equation describing the time-evolution of the probability of a reaction system to occupy each one of its possible discrete set of states. In a well-stirred (non spatial) AC a state is equivalent to a multiset of molecules.

Structure-to-Function Mapping

A fundamental "logic" of real chemical systems is that molecules posses a structure and that this structure determines how the molecule interacts with other molecules. Thus, there is a mapping from the structure of a molecule to its function, which determines the reaction rules.

In real chemistry the mapping from structure to function is given by natural or physical laws. In artificial chemistries the structure-to-function mapping is usually given by some kind of algorithm. In most cases, an artificial chemistry using a structure-to-function mapping does not aim at modeling real molecular dynamics in detail. Rather the aim is to capture the fact that there is a structure, a function, and a relation between them.

Having an AC with a structure-to-function mapping at hand, we can study strongly constructive dynamical systems [27,28]. These are systems where, through the interaction of its components, novel molecular species appear.

Example: Lambda-Calculus (*AlChemy*)

This section demonstrates a structure-to-function that employs a concept from computer science: the λ-calculus. The λ-calculus has been used by Fontana [26] and Fontana and Buss [27] to define a constructive artificial chemistry.

In the so-called *AlChemy*, a molecule is a normalized λ-expression. A λ-expression is a word over an alphabet $A = \{\lambda, ., (,)\} \cup V$ where $V = \{x_1, x_2, \ldots\}$ is an infinite set of available variable names. The set of λ expres-

```
while not terminate() do
  reactands := choseASubsetFrom(P);
  if randomReal(0, 1) < reactionProbability(reactands);
     products = reaction(reactands);
     P := remove(P, reactands);
     P := insert(P, products);
  fi
  t := t + 1/ sizeOf(P)  ;; increment simulated time
od
```

Artificial Chemistry, Algorithm 2

sions Γ is defined for $x \in V, s_1 \in \Gamma, s_2 \in \Gamma$ by

$$\begin{aligned} x \in \Gamma &\quad \text{variable name}\,, \\ \lambda x.s_2 \in \Gamma &\quad \text{abstraction}\,, \\ (s_2)s_1 \in \Gamma &\quad \text{application}\,. \end{aligned}$$

An abstraction $\lambda x.s_2$ can be interpreted as a function definition, where x is the parameter in the "body" s_2. The expression $(s_2)s_1$ can be interpreted as the application of s_2 on s_1, which is formalized by the rewriting rule

$$(\lambda x.s_2)s_1 = s_2[x \longleftarrow s_1]\,, \tag{8}$$

where $s_2[x \longleftarrow s_1]$ denotes the term which is generated by replacing every unbounded occurrence of x in s_2 by s_1. A variable x is bounded if it appears in in a form like $\ldots(\lambda x.\ldots.x\ldots)\ldots$. It is also not allowed to apply the rewriting rule if a variable becomes bounded. Example: Let $s_1 = \lambda x_1.(x_1)\lambda x_2.x_2$ and $s_2 = \lambda x_3.x_3$ then we can derive:

$$\begin{aligned} (s_1)s_2 &\Rightarrow (\lambda x_1.(x_1)\lambda x_2.x_2)\lambda x_3.x_3 \\ &\Rightarrow (\lambda x_3.x_3)\lambda x_2.x_2 \Rightarrow \lambda x_2.x_2\,. \end{aligned} \tag{9}$$

The simplest way to define a set of second-order catalytic reactions $\mathcal{R}$ by applying a molecule s_1 to another molecule s_2:

$$s_1 + s_2 \rightarrow s_1 + s_2 + \mathit{normalForm}((s_1)s_2)\,. \tag{10}$$

The procedure *normalForm* reduces its argument term to normal form, which is in practice bounded by a maximum of available time and memory; if these resources are exceeded before termination, the collision is considered to be elastic. It should be noted that the λ-calculus allows an elegant generalization of the collision rule by defining it by λ-expression $\Phi \in \Gamma\colon s_1 + s_2 \rightarrow s_1 + s_2 + \mathit{normalForm}(((\Phi)s_1)s_2)$.

In order to simulate the artificial chemistry, Fontana and Buss [27] used an algorithm like the algorithm described in Subsect. "Explicitly Simulated Collisions", which simulates second-order catalytic mass-action kinetics in a well-stirred population under flow condition.

AlChemy possesses a couple of important general properties: Molecules come in two forms: as passive data possessing a structure and as active machines operating on those structures. This generates a "strange loop" like in typogenetics [36,94], which allows molecules to refer to themselves and to operate on themselves. Structures operate on structures, and by doing so change the way structures operate on structures, and so on; creating a "strange", self-referential dynamic loop. Obviously, there are always some laws that cannot be changed, which are here the rules of the lambda-calculus, defining the structure-to-function mapping. Thus, we can interpret these fixed and predefined laws of an artificial chemistry as the natural or physical laws.

Arithmetic and Logic Operations

One of the most easy ways to define reaction rules implicitly is to apply arithmetic operations taken from mathematics. Even the simplest rules can generate interesting behaviors. Assume, for example, that the molecules are natural numbers: $\mathcal{M} = \{0, 1, 2 \ldots, n-1\}$ and the reaction rules are defined by division: reaction$(a_1, a_2) = a_1/a_2$ if a_1 is a multiple of a_2, otherwise the molecules do not react. For the dynamics we assume a finite, discrete population and an algorithm like the one described in Subsect. "Explicitly Simulated Collisions". With increasing population size the resulting reaction system displays a phase transition, at which the population is able to produce prime numbers with high probability (see [8] for details). In a typical simulation, the initially high diversity of a random population is reduced leaving a set of non-reacting prime numbers.

A quite different behavior can be obtained by simply replacing the division by addition: reaction$(a_1, a_2) = a_1 + a_2 \mod n$ with $n = |\mathcal{M}|$ the number of molecular species. Initializing the well-stirred tank reactor only with

the molecular species 1, the diversity rapidly increases towards its maximum and the reactor behaves apparently totally chaotic after a very short transient phase (for a sufficiently large population size). However, there are still regularities, which depend on the prime factors of n (cf. [15]).

Matrix-Multiplication Chemistry More complex reaction operators can be defined that operate on vectors or strings instead of scalar numbers as molecules. The matrix multiplication chemistry introduce by Banzhaf [5,6,7] uses binary strings as molecules. The reaction between two binary strings is performed by folding one of them into a matrix which then operates on the other string by multiplication.

Example of a reaction for 4-bit molecules Assume a reaction $s_1 + s_2 \Longrightarrow s_3$. The general approach is:

1. Fold s_1 to matrix M. Example: $s_1 = (s_1^1, s_1^2, s_1^3, s_1^4)$

$$M = \begin{pmatrix} s_1^1 & s_1^2 \\ s_1^3 & s_1^4 \end{pmatrix} . \tag{11}$$

2. Multiply M with subsequences of s_2. Example: Let $s_2 = \left(s_2^1, s_2^2, s_2^3, s_2^4\right)$ be divided into two subsequences $s_2^{12} = \left(s_2^1, s_2^2\right)$ and $s_2^{34} = (s_2^3, s_2^4)$. Then we can multiply M with the subsequences:

$$s_3^{12} = M \odot s_2^{12} , \quad s_3^{34} = M \odot s_2^{34} . \tag{12}$$

3. Compose s_3 by concatenating the products. Example: $s_3 = s_3^{12} \oplus s_3^{34}$.

There are various ways of defining the vector matrix product $\odot$. It was mainly used with the following **threshold multiplication**. Given a bit vector $x = (x_1, \ldots, x_n)$ and a bit matrix $M = (M_{ij})$ then the term $y = M \odot x$ is defined by:

$$y_j = \begin{cases} 0 & \text{if } \sum_{i=1}^n x_i M_{i,j} \leq \Phi . \\ 1 & \text{otherwise} . \end{cases} \tag{13}$$

The threshold multiplication is similar to the common matrix-vector product, except that the resulting vector is mapped to a binary vector by using the threshold Φ.

Simulating the dynamics by an ODE or explicit molecular collisions (Subsect. "Explicitly Simulated Collisions"), we can observe that such a system would develop into a steady state where some string species support each other in production and thus become a stable autocatalytic cycle, whereas others would disappear due to the competition in the reactor.

The system has a couple of interesting properties: Despite the relatively simple definition of the basic reaction mechanism, the resulting complexity of the reaction system and its dynamics is surprisingly high. Like in typogenetics and Fontana's lambda-chemistry, molecules appear in two forms; as passive data (binary strings) and as active operators (binary matrix). The folding of a binary string to a matrix is the central operation of the structure-to-function mapping. The matrix is multiplied with substrings of the operand and thus some kind of locality is preserved, which mimics local operation of macromolecules (e. g. ribosomes or restriction enzymes) on polymers (e. g. RNA or DNA). Locality is also conserved in the folding, because bits that are close in the string are also close in the matrix.

Autocatalytic Polymer Chemistries

In order to study the emergence and evolution of autocatalytic sets [19,46,75] Bagley, Farmer, Fontana, Kauffman and others [3,23,47,48,59] have used artificial chemistries where the molecules are character sequences (e. g., $\mathcal{M} = \{a, b, aa, ab, ba, bb, aaa, aab, \ldots\}$) and the reactions are concatenation and cleavage, for example:

$$aa + babb \leftrightharpoons aababb \quad \text{(slow)} . \tag{14}$$

Additionally, each molecule can act as a catalyst enhancing the rate of a concatenation reaction.

$$aa + babb + bbb \leftrightharpoons bbb + aababb \quad \text{(fast)} . \tag{15}$$

Figure 1 shows an example of an autocatalytic network that appeared. There are two time scales. Reactions that are not catalyzed are assumed to be very slow and not depicted, whereas catalyzed reactions, inflow, and decay are fast.

Typical experiments simulate a well-stirred flow reactor using meta-dynamical ODE frame work [3]; that is, the ODE model is dynamically changed when new molecular species appear or present species vanish. Note that the structure of the molecules does not fully determine the reaction rules. Which molecule catalyzes which reaction (the dotted arrows in Fig. 1) is assigned explicitly randomly. An interesting aspect is that the catalytic or autocatalytic polymer sets (or reaction networks) evolve without having a genome [47,48]. The simulation studies show how small, spontaneous fluctuations can be amplified by an autocatalytic network, possibly leading to a modification of the entire network [3]. Recent studies by Fernando and Rowe [24] include further aspects like energy conservation and compartmentalization.

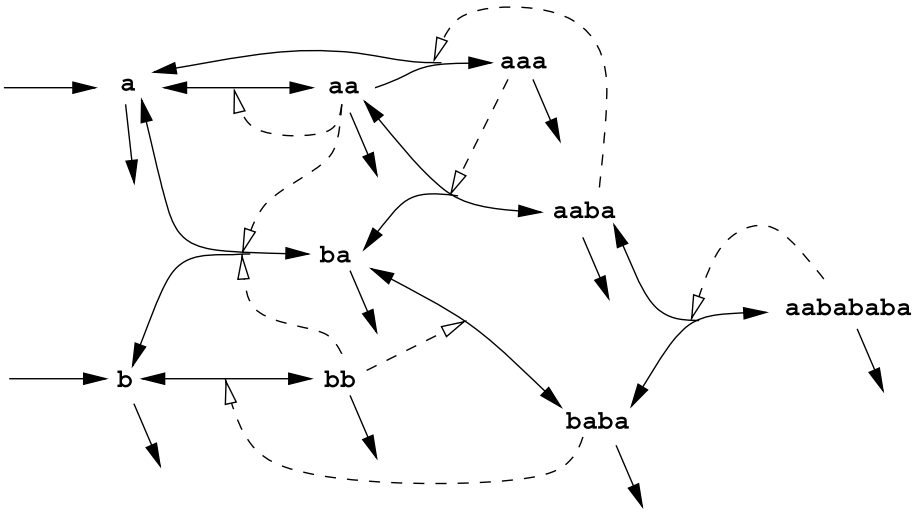

Artificial Chemistry, Figure 1
Example of an autocatalytic polymer network. The *dotted lines* represent catalytic links, e. g., $aa + ba + aaa \rightarrow aaba + aaa$. All molecules are subject to a dilution flow. In this example, all polymerization reactions are reversible and there is a continuous inflow of the monomers *a* and *b*. Note that the network contains further auto-catalytic networks, for example, {*a*, *b*, *aa*, *ba*} and {*a*, *aa*}, of which some are closed and thus organizations, for example, {*a*, *b*, *aa*, *ba*}. The set {*a*, *aa*} is not closed, because *b* and *ba* can be produced

Bagley and Farmer [3] showed that autocatalytic sets can be silhouetted against a noisy background of spontaneously reacting molecules under moderate (that is, neither too low nor too high) flow.

Artificial Chemistries Inspired by Turing Machines

The concept of an abstract automaton or Turing machine provides a base for a variety of structure-function mappings. In these approaches molecules are usually represented as a sequence of bits, characters, or instructions [36,51,63,86]. A sequence of bits specifies the behavior of a machine by coding for its state transition function. Thus, like in the matrix-multiplication chemistry and the lambda-chemistry, a molecule appears in two forms, namely as passive data (e. g., a binary string) and as an active machine. Also here we can call the mapping from a binary string into its machine **folding**, which might be indeterministic and may depend on other molecules (e. g., [40]).

In the early 1970s Laing [51] argued for abstract, non-analogous models in order to develop a general theory for living systems. For developing such general theory, so-called **artificial organisms** would be required. Laing suggested a series of artificial organisms [51,52,53] that should allow one to study general properties of life and thus would allow one to derive a theory which is not restricted to the instance of life we observe on earth. The artificial organisms consist of different compartments, e. g., a "brain" plus "body" parts. These compartments contain binary strings as molecules. Strings are translated to a sequence of instructions forming a three-dimensional shape (cf. [78,86,94]). In order to perform a reaction, two molecules are attached such that they touch at one position (Fig. 2). One of the molecules is executed like a Turing machine, manipulating the passive data molecule as its tape. Laing proved that his artificial organisms are able to perform universal computation. He also demonstrated different forms of self-reproduction, self-description and self-inspection using his molecular machines [53].

Typogenetics is a similar approach, which was introduced in 1979 by Hofstadter in order to illustrate the "formal logic of life" [36]. Later, *typogenetics* was simulated and investigated in more detail [67,94,95]. The molecules of *typogenetics* are character sequences (called strands) over the alphabet `A, C, G, T`. The reaction rules are "typographic" manipulations based on a set of predefined basic operations like cutting, insertion, or deletion of characters. A sequence of such operations forms a unit (called an enzyme) which may operate on a character sequence like a Turing machine on its tape. A character string can be "translated" to an enzyme (i. e., a sequence of operations) by mapping two characters to an operation according to a predefined "genetic code".

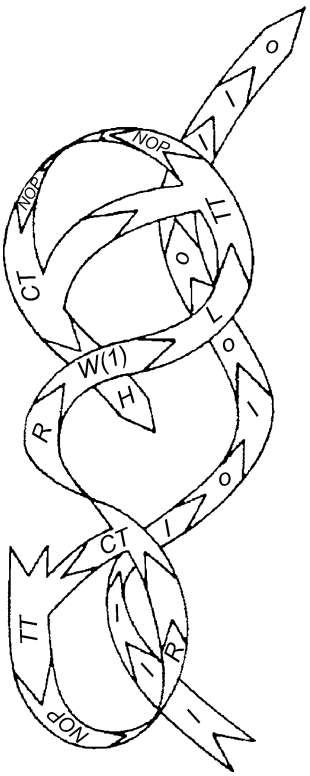

Artificial Chemistry, Figure 2
Illustration of Laing's molecular machines. A program molecule is associated with a data molecule. Figure from [52]

Another artificial chemistry whose reaction mechanism is inspired by the Turing machine was suggested by McCaskill [63] in order to study the self-organization of complex chemical systems consisting of catalytic polymers. Variants of this AC were realized later in a specially designed parallel reconfigurable computer based on FPGAs – Field Programmable Gate Arrays [12,18,89]. In this approach, molecules are binary strings of fixed length (e. g., 20 [63]) or of variable length [12]. As in previous approaches, a string codes for an automaton able to manipulate other strings. And again, pattern matching is used to check whether two molecules can react and to obtain the "binding site"; i. e., the location where the active molecule (machine) manipulates the passive molecule (tape). The general reaction scheme can be written as:

$$s_{\mathrm{M}} + s_{\mathrm{T}} \rightarrow s_{\mathrm{M}} + s'_{\mathrm{T}} \,. \tag{16}$$

In experimental studies, self-replicating strings appeared frequently. In coevolution with parasites, an evolutionary arms race started among these species and the self-replicating string diversified to an extent that the parasites could not coadapt and went extinct. In a spatial environment (e. g., a 2-D lattice), sets of cooperating polymers evolved, interacting in a hypercyclic fashion. The authors also observed a *chemoton*-like [30] cooperative behavior, with spatially isolated, membrane-bounded evolutionary stable molecular organizations.

Machines with Fixed Tape Size In order to perform large-scale systematic simulation studies like the investigation of noise [40] or intrinsic evolution [16], it makes sense to limit the study to molecules of fixed tractable length. Ikegami and Hashimoto [40] developed an abstract artificial chemistry with two types of molecular species: 7 bit long tapes, which are mapped to 16 bit long machines. Tapes and machines form two separated populations, simulated in parallel. Reactions take place between a tape and a machine according to the following reaction scheme:

$$s_{\mathrm{M}} + s_{\mathrm{T}} \longrightarrow s_{\mathrm{M}} + s_{\mathrm{T}} + s'_{\mathrm{M}} + s'_{\mathrm{T}} \,. \tag{17}$$

A machine s_M can react with a tape s_T, if its head matches a substring of the tape s_T and its tail matches a different substring of the tape s_T. The machine operates only between these two substrings (called reading frame) which results in a tape s'_{T}. The tape s'_{T} is then "translated" (folded) into a machine s'_{M}.

Ikegami and Hashimoto [40] showed that under the influence of low noise, simple autocatalytic loops are formed. When the noise level is increased, the reaction network is destabilized by parasites, but after a relatively long transient phase (about 1000 generations) a very stable, dense reaction network appears, called **core network** [40]. A core network maintains its relatively high diversity even if the noise is deactivated. The active mutation rate is high. When the noise level is very high, only small, degenerated core networks emerge with a low diversity and very low (even no) active mutation. The core networks which emerged under the influence of external noise are very stable so that their is no further development after they have appeared.

Assembler Automata

An **Assembler automaton** is like a parallel von Neumann machine. It consists of a core memory and a set of processing units running in parallel. Inspired by *Core Wars* [14], assembler automata have been used to create certain artificial life systems, such as *Coreworld* [72,73], *Tierra* [74], Avida [1], and Amoeba [69]. Although these machines have been classified as artificial chemistries [72], it is in general difficult to identify molecules or reactions. Furthermore, the assembler automaton *Tierra* has explicitly

been designed to imitate the Cambrian explosion and *not* a chemical process. Nevertheless, in some cases we can interpret the execution of an assembler automaton as a chemical process, which is especially possible in a clear way in experiments with Avida. Here, a molecule is a single assembler program, which is protected by a memory management system. The system is initialized with manually written programs that are able to self-replicate. Therefore, in a basic version of Avida only unimolecular first-order reactions occur, which are of replicator type. The reaction scheme can be written as $a \rightarrow a + \text{mutation}(a)$. The function mutation represents the possible errors that can occur while the program is self-replicating.

Latticei Molecular Systems

In this section systems are discussed which consist of a regular lattice, where each lattice site can hold a part (e. g. atom) of a molecule. Between parts, bonds can be formed so that a molecule covers many lattice sites. This is different to systems where a lattice site holds a complete molecule, like in Avida. The important difference is that in lattice molecular systems, the space of the molecular structure is identical to the space in which the molecules are floating around. In systems where a molecule covers just one lattice site, the molecular structure is described in a different space independently from the space in which the molecule is located.

Lattice molecular systems have been intensively used to model polymers [92], protein folding [82], and RNA structures. Besides approaches which intend to model real molecular dynamics as accurately as possible, there are also approaches which try to build abstract models. These models should give insight into statistical properties of polymers like their energy landscape and folding processes [76,77], but should not give insight into questions concerning origin and evolution self-maintaining organizations or molecular information processing. For these questions more abstract systems are studied as described in the following.

Varela, Maturana, and Uribe introduced in [93] a lattice molecular system to illustrate their concept of autopoiesis (cf. [65,99]). The system consists of a 2-D square lattice. Each lattice site can be occupied by one of the following atoms: Substrate *S*, catalyst *K*, and monomer *L*. Atoms may form bonds and thus form molecular structures on the lattice. If molecules come close to each other they may react according to the following reaction rules:

$K + 2S \rightarrow K + L$ (1) Composition: Formation of a monomer.

$\ldots$-L-L $+ L \rightarrow \ldots$-L-L-L (2) Concatenation: L + L → L-L Formation of a bond between a monomer and another monomer with no more than one bond. This reaction is inhibited by a double-bounded monomer [65].

$L \rightarrow 2S$ (3) Disintegration: Decomposition of a monomer

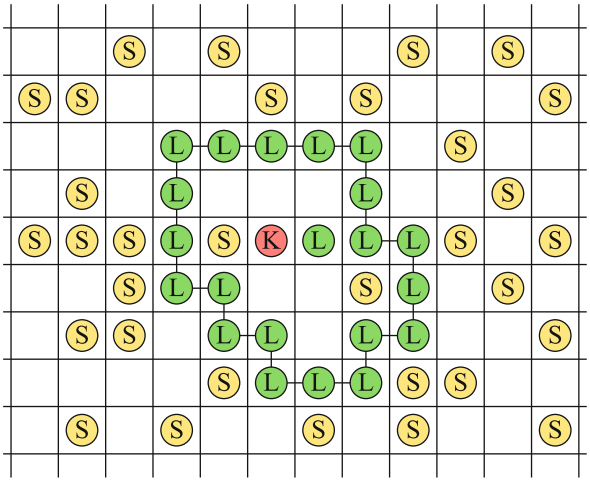

Artificial Chemistry, Figure 3
Illustration of a lattice molecular automaton that contains an autopoietic entity. Its membrane is formed by a chain of L monomers and encloses a catalyst *K*. Only substrate S may diffuse through the membrane. Substrate inside is catalyzed by *K* to form free monomers. If the membrane is damaged by disintegration of a monomer *L* it can be quickly repaired by a free monomer. See [65,93] for details

Figure 3 illustrates an autopoietic entity that may arise. Note that it is quite difficult to find the right dynamical conditions under which such autopoietic structures are stable [65,68].

Cellular Automata

Cellular automata ► Mathematical Basis of Cellular Automata, Introduction to can be used as a medium to simulate chemical-like processes in various ways. An obvious approach is to use the cellular automata to model space where each cell can hold a molecule (like in Avida) or atom (like in lattice molecular automata). However there are approaches where it is not clear at the onset what a molecule or reaction is. The model specification does not contain any notion of a molecule or reaction, so that an observer has to identify them. Molecules can be equated

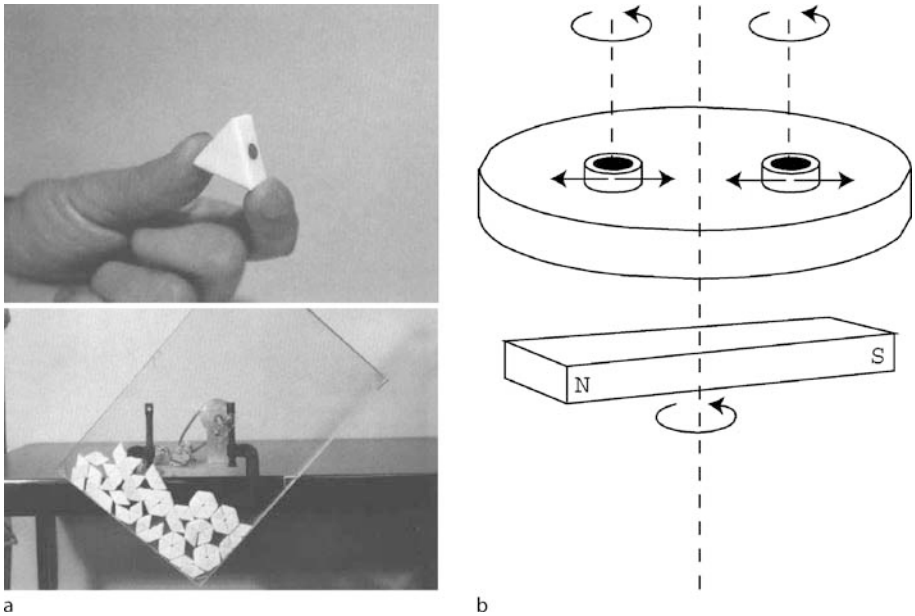

Artificial Chemistry, Figure 4
Example of mechanical artificial chemistries. a Self-assembling magnetic tiles by Hosokawa et al. [38]. b Rotating magnetic discs by Grzybowski et al. [33]. Figures taken (and modified) from [33,38]

with self-propagating patterns like gliders, self-reproducing loops [54,80], or with the moving boundary between two homogeneous domains [37]. Note that in the latter case, particles become visible as space-time structures, which are defined by boundaries and not by connected set of cells with specific states. An interaction between these boundaries is then interpreted as a reaction.

Mechanical Artificial Chemistries

There are also physical systems that can be regarded as artificial chemistries. Hosokawa et al. [38] presented a mechanical self-assembly system consisting of triangular shapes that form bonds by permanent magnets. Interpreting attachment and detachment as chemical reactions, Hosokawa et al. [38] derived a chemical differential equation (Subsect. "The Chemical Differential Equation") modeling the kinetics of the structures appearing in the system. Another approach uses millimeter-sized magnetic disks at a liquid-air interface, subject to a magnetic field produced by a rotating permanent magnet [33]. These magnetic discs exhibit various types of structures, which might even interact resembling chemical reactions. The important difference to the first approach is that the rotating magnetic discs form dissipative structures, which require continuous energy supply.

Semi-Realistic Molecular Structures and Graph Rewriting

Recent development in artificial chemistry aims at more realistic molecular structures and reactions, like *oo chemistry* by Bersini [10] (see also [57]) or *toy chemistry* [9]. These approaches apply graph rewriting, which is a powerful and quite universal approach for defining transformations of graph-like structures [50] In toy chemistry, molecules are represented as labeled graphs; i. e., by their structural formulas; their basic properties are derived by a simplified version of the extended Hückel MO theory that operates directly on the graphs; chemical reaction mechanisms are implemented as graph rewriting rules acting on the structural formulas; reactivities and selectivities are modeled by a variant of the frontier molecular orbital theory based on the extended Hückel scheme. Figure 5 shows an example of a graph-rewriting rule for unimolecular reactions.

Space

Many approaches discussed above assume that the reaction vessel is well-stirred. However, especially in living systems, space plays an important role. Cells, for example, are not a bag of enzymes, but spatially highly structured.

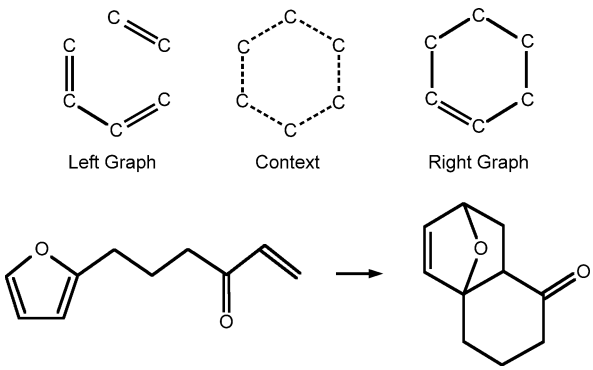

Artificial Chemistry, Figure 5
An example of a graph-rewriting rule (*top*) and its application to the synthesis of a bridge ring system (*bottom*). Figure from [9]

Moreover, it is assumed that space has played an important role in the origin of life.

Techniques for Modeling Space

In chemistry, space is usually modeled by assuming an Euclidean space (partial differential equation or particle-based) or by compartments, which we can find also in many artificial chemistries [29,70]. However, some artificial chemistries use more abstract, "computer friendly" spaces, such as a core-memory like in *Tierra* [74] or a planar triangular graph [83]. There are systems like MGS [31] that allow to specify various topological structures easily. Approaches like P-systems and membrane computing [70] allow one even to change the spatial structure dynamically ► Membrane Computing.

Phenomena in Space

In general, space delays the extinction of species by fitter species, which is for example, used in Avida to obtain more complex evolutionary patterns [1]. Space leads usually to higher diversity. Moreover, systems that are instable in a well-stirred reaction vessel can become stable in a spatial situation. An example is the hypercycle [20], which stabilizes against parasites when space is introduced [11]. Conversely, space can destabilize an equilibrium, leading to symmetry breaking, which has been suggested as an important mechanism underlying morphogenesis [91]. Space can support the co-existence of chemical species, which would not co-exist in a well-stirred system [49]. Space is also necessary for the formation of autopoietic structures [93] and the formation of units that undergo Darwinian evolution [24].

Theory

There is a large body of theory from domains like chemical reaction system theory [21], Petri net theory [71], and rewriting system theory (e. g., P-systems [70]), which applies also to artificial chemistries. In this section, however, we shall investigate in more detail those theoretical concepts which have emerged in artificial chemistry research.

When working with complex artificial chemistries, we are usually more interested in the qualitative than quantitative nature of the dynamics. That is, we study how the set of molecular species present in the reaction vessel changes over time, rather than studying a more detailed trajectory in concentration space.

The most prominent qualitative concept is the autocatalytic set, which has been proposed as an important element in the origin of life [19,46,75]. An **autocatalytic set** is a set of molecular species where each species is produced by at least one catalytic reaction within the set [41,47] (Fig. 1). This property has also been called self-maintaining [27] or (catalytic) closure [48]. Formally: A set of species $A \subseteq \mathcal{M}$ is called an autocatalytic set (or sometimes, catalytically closed set), if for all species[1] $a \in A$ there is a catalyst $a' \in A$, and a reaction $\rho \in \mathcal{R}$ such that a' is catalyst in ρ (i. e., $a' \in \mathrm{LHS}(\rho)$ and $a' \in \mathrm{RHS}(\rho)$) and ρ can take place in A (i. e., $\mathrm{LHS}(\rho) \subseteq A$).

Example 1 (autocatalytic sets) $\mathcal{R} = \{a \to 2a, a \to a + b, a \to, b \longrightarrow\}$. Molecule a catalyzes its own production and the production of b. Both molecules are subject to a dilution flow (or, equivalently, spontaneous decay), which is usually assumed in ACs studying autocatalytic sets. In this example there are three autocatalytic sets: two non-trivial autocatalytic sets $\{a\}$ and $\{a, b\}$, and the empty set $\{\}$, which is from a mathematical point of view also an autocatalytic set.

The term "autocatalytic set" makes sense only in ACs where catalysis is possible, it is not useful when applied to arbitrary reaction networks. For this reason Dittrich and Speroni di Fenizio [17] introduced a general notion of self-maintenance, which includes the autocatalytic set as a special case: Formally, given a reaction network $\langle \mathcal{M}, \mathcal{R} \rangle$ with $m = |\mathcal{M}|$ molecules and $r = |\mathcal{R}|$ reactions, and let $\mathbf{S} = (s_{a,\rho})$ be the $(m \times r)$ stoichiometric matrix implied by the reaction rules $\mathcal{R}$, where $s_{a,\rho}$ denotes the number

[1] For general reaction systems, this definition has to be refined. When A contains species that are part of the inflow, like a and b in Fig. 1, but which are not produced in a catalytic way, we might want them to be part of an autocatalytic set. Assume, for example, the set $A = \{a, b, aa, ba\}$ from Fig. 1, where aa catalyzes the production of aa and ba, while using up "substrate" a and b, which are not catalytically produced.

of molecules of type a produced in reaction ρ. A set of molecules $C \subseteq \mathcal{M}$ is called **self-maintaining**, if there exists a flux vector $\mathbf{v} \in \mathbb{R}^r$ such that the following three conditions apply: (1) for all reactions ρ that can take place in C (i. e., $\mathrm{LHS}(\rho) \subseteq C$) the flux $v_\rho > 0$; (2) for all remaining reactions ρ (i. e., $\mathrm{LHS}(\rho) \not\subseteq C$), the flux $v_\rho = 0$; and (3) for all molecules $a \in C$, the production rate $(\mathbf{Sv})_a \geq 0$. v_ρ denotes the element of $\mathbf{v}$ describing the flux (i. e. rate) of reaction ρ. $(\mathbf{Sv})_a$ is the production rate of molecule a given flux vector $\mathbf{v}$. In Example 1 there are three self-maintaining (autocatalytic) sets.

Interestingly, there are self-maintaining sets that cannot make up a stationary state. In our Example 1 only two of the three self-maintaining sets are species combinations that can make up a stationary state (i. e., a state $\mathbf{x}^0$ for which $\mathbf{0} = \mathbf{Sv}(\mathbf{x}^0)$ holds). The self-maintaining set $\{a\}$ cannot make up a stationary state because a generates b through the reaction $a \rightarrow a + b$. Thus, there is no stationary state (of Eq. (3) with Assumptions 1 and 2) in which the concentration of a is positive while the concentration of b is zero. In order to filter out those less interesting self-maintaining sets, Fontana and Buss [27] introduced a concept taken from mathematics called closure: Formally, a set of species $A \subseteq \mathcal{M}$ is **closed**, if for all reactions ρ with $\mathrm{LHS}(\rho) \subseteq A$ (the reactions that can take place in A), the products are also contained in A, i. e., $\mathrm{RHS}(\rho) \subseteq A$).

Closure and self-maintenance lead to the important concept of a chemical organization [17,27]: Given an arbitrary reaction network $\langle \mathcal{M}, \mathcal{R} \rangle$, a set of molecular species that is closed and self-maintaining is called an **organization**. The importance of an organization is illustrated by a theorem roughly saying that given a fixed point of the chemical ODE (Eq. (3)), then the species with positive concentrations form an organization [17]. In other words, we have only to check those species combinations for stationary states that are organizations. The set of all organizations can be visualized nicely by a Hasse-diagram, which sketches the hierarchical (organizational) structure of the reaction network (Fig. 6). The dynamics of the artificial chemistry can then be explained within this Hasse-diagram as a movement from organization to organization [61,84].

Note that in systems under flow condition, the (finite) set of all organizations forms an algebraic lattice [17]. That is, given two organizations, there is always a unique organization union and organization intersection.

Although the concept of **autopoiesis** [93] has been described informally in quite some detail, a stringent formal definition is lacking. In a way, we can interpret the formal concepts above as an attempt to approach necessary properties of an autopoietic system step by step by precise formal means. Obviously, being (contained in) at least one organization is a necessary condition for a chemical autopoietic system. But it is not sufficient. Missing are the notion of robustness and a spatial concept that formalizes a system's ability to maintain (and self-create) its own identity, e. g., through maintaining a membrane (Fig. 3).

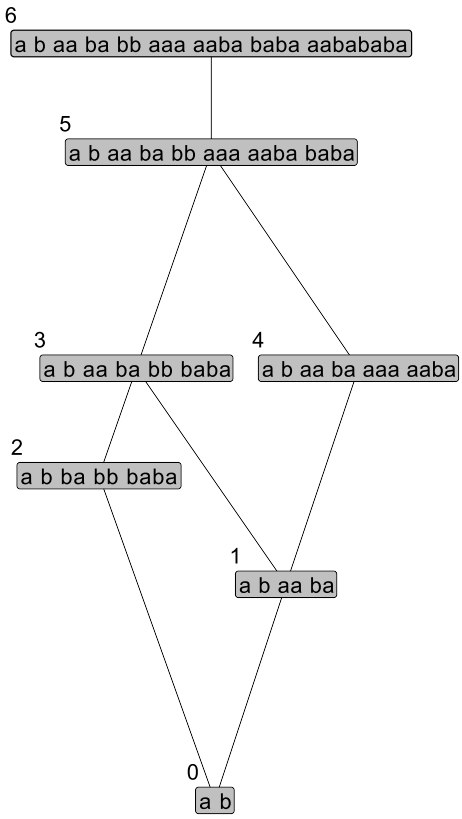

Artificial Chemistry, Figure 6
Lattice of organizations of the autocatalytic network shown in Fig. 1. An organization is a closed and self-maintaining set of species. Two organizations are connected by a line, if one is contained in the other and there is no organization in between. The vertical position of an organization is determined by the number of species it contains

Evolution

With artificial chemistries we can study chemical evolution. Chemical evolution (also called "pre-biotic evolution") describes the first step in the development of life, such as the formation of complex organic molecules from simpler (in-)organic compounds [62]. Because there are no pre-biotic fossils, the study of chemical evolution has to rely on experiments [56,66] and theoretic (simulation) models [19]. Artificial chemistries aim at capturing the

constructive nature of these chemical systems and try to reproduce their evolution in computer simulations. The approaches can be distinguished by whether the evolution is driven by external operators like mutation, or whether variation and selection appears intrinsically through the chemical dynamics itself.

Extrinsic Evolution

In extrinsic approaches, an external variation operator changes the reaction network by adding, removing, or manipulating a reaction, which may also lead to the addition or removal of chemical species. In this approach, a molecule does not need to posses a structure. The following example by Jain and Krishna [41] shows that this approach allows one to create an evolving system quite elegantly:

Let us assume that the reaction network consists of m species $\mathcal{M} = \{1, \ldots, m\}$. There are first-order catalytic reaction rules of the form $(i \to i + j) \in \mathcal{R}$ and a general dilution flow $(a \to) \in \mathcal{R}$ for all species $a \in \mathcal{M}$. The reaction network is completely described by a directed graph represented by the adjacency matrix $C = (c_{i,j})$, $i, j \in \mathcal{M}$, $c_{i,j} \in \{0, 1\}$, with $c_{i,j} = 1$ if molecule j catalyzes the production of i, and $c_{i,j} = 0$ otherwise. In order to avoid that self-replicators dominate the system, Jain and Krishna assume $c_{i,i} = 0$ for all molecules $i \in \mathcal{M}$.

At the beginning, the reaction network is randomly initialized, that is, (for $i \neq j) c_{i,j} = 1$ with probability p and $c_{i,j} = 0$, otherwise. In order to simulate the dynamics, we assume a population of molecules represented by the concentration vector $\mathbf{x} = (x_1, \ldots, x_m)$, where x_i represents the current concentration of species i. The whole system is simulated in the following way:

Step 1: Simulate the chemical differential equation

$$\dot{x}_i = \sum_{i \in \mathcal{M}} c_{j,i} - x_i \sum_{k,j \in \mathcal{M}} c_{k,j} x_j \,, \tag{18}$$

until a steady state is reached. Note that this steady state is usually independent of the initial concentrations, cf. [85].

Step 2: Select the "mutating" species i, which is the species with smallest concentration in the steady state.

Step 3: Update the reaction network by replacing the mutating species by a new species, which is created randomly in the same way as the initial species. That is, the entries of the ith row and ith column of the adjacency matrix C are replaced by randomly chosen entries with the same probability p as during initialization.

Step 4: Go to Step 1.

Note that there are two explicitly simulated time scales: a slow, evolutionary time scale at which the reaction network evolves (Step 2 and 3), and a fast time scale at which the molecules catalytically react (Step 1).

In this model we can observe how an autocatalytic set inevitably appears after a period of disorder. After its arrival the largest autocatalytic set increases rapidly its connectivity until it spans the whole network. Subsequently, the connectivity converges to a steady state determined by the rate of external perturbations. The resulting highly non-random network is not fully stable, so that we can also study the causes for crashes and recoveries. For example, Jain and Krishna [44] identified that in the absence of large external perturbation, the appearance of a new *viable* species is a major cause of large extinction and recoveries. Furthermore, crashes can be caused by extinction of a "keystone species". Note that for these observations, the new species created in Step 3 do not need to inherit any information from other species.

Intrinsic Evolution

Evolutionary phenomena can also be caused by the intrinsic dynamics of the (artificial) chemistry. In this case, external operators like those mutating the reaction network are not required. As opposed to the previous approach, we can define the whole reaction network at the onset and let only the composition of molecules present in the reaction vessel evolve. The reaction rules are (usually) de-

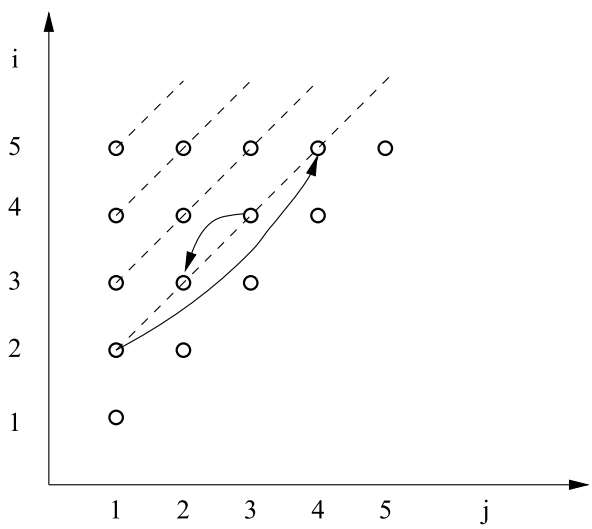

Artificial Chemistry, Figure 7
Illustration of syntactically and semantically closed organizations. Each organization consists of an infinite number of molecular species (connected by a *dotted line*). A *circle* represents a molecule having the structure: $\lambda x_1.\lambda x_2.\ldots.27$

fined by a structure-to-function mapping similar to those described in Sect. "Structure-to-Function Mapping". The advantage of this approach is that we need not define an external operator changing the reaction network's topology. Furthermore, the evolution can be more realistic because, for example, molecular species that have vanished can reenter at a later time, which does not happen in an approach like the one described previously. Also, when we have a structure-to-function mapping, we can study how the structure of the molecules is related to the emergence of chemical organizations.

There are various approaches using, for example, Turing machines [63], lambda-calculus [26,27], abstract automata [16], or combinators [83] for the structure-to-function mapping.

In all of those approaches we can observe an important phenomenon: While the AC evolves autocatalytic sets or more general, chemical organizations become visible. Like in the system by Jain and Krishna this effect is indicated also by an increase of the connectivity of the species within the population.

When an organization becomes visible, the system has focused on a sub-space of the whole set of possible molecules. Those molecules, belonging to the emerged organization, posses usually relatively high concentrations, because they are generated by many reactions. The closure of this set is usually smaller than the universe $\mathcal{M}$. Depending on the setting the emerged organization can consist of a single self-replicator, a small set of mutually producing molecules, or a large set of different molecules. Note that in the latter case the organization can be so large (and even infinite) that not all its molecules are present in the population (see Fig. 7 for an example). Nevertheless the population can carry the organization, if the system can keep a generating set of molecules within the population.

An emerged organization can be quite stable but can also change either by external perturbations like randomly inserted molecules or by internal fluctuations caused by reactions among molecules that are not part of the emerged organization, but which have remained in small quantities in the population, cf. for an example [61].

Dittrich and Banzhaf [16] have shown that it is even possible to obtain evolutionary behavior without any externally generated variation, that is, even without any inflow of random molecules. And without any explicitly defined fitness function or selection process. Selection emerges as a result of the dilution flow and the limited population size.

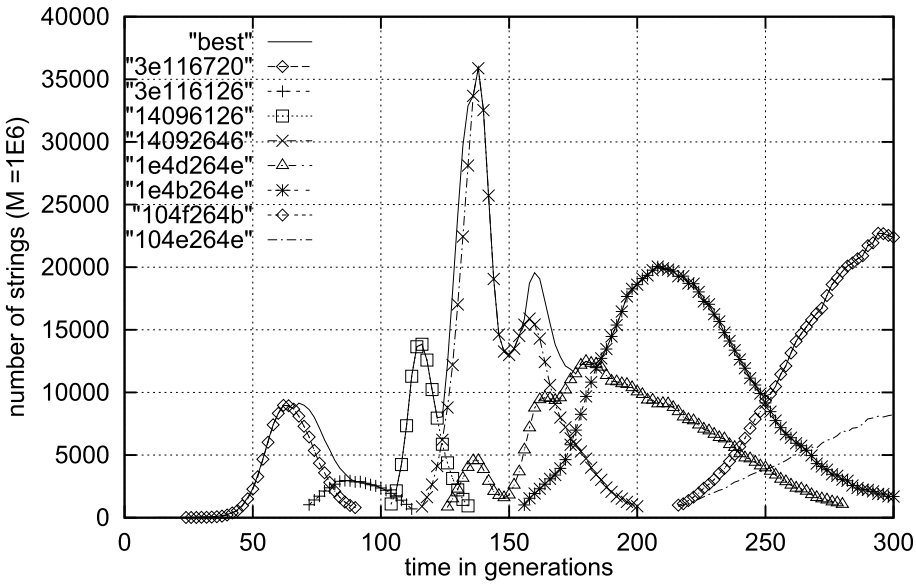

Artificial Chemistry, Figure 8

Example of a self-evolving artificial chemistry. The figure shows the concentration of some selected species of a reaction vessel containing approximately 10^4 different species. In this example, molecules are binary strings of length 32 bit. Two binary strings react by mapping one of them to a finite state machine operating on the second binary string. There is only a general, non-selective dilution flow. No other operators like mutation, variation, or selection are applied. Note that the structure of a new species tends to be similar to the structure of the species it has been created from. The dynamics simulated by the algorithm of the introduction. Population size $k = 10^6$ molecules. Figure from [16]

Syntactic and Semantic Closure

In many "constructive" implicitly defined artificial chemistries we can observe species appearing that share syntactical and functional similarities that are invariant under the reaction operation. Fontana and Buss [27] called this phenomenon syntactic and semantic closure.

Syntactic closure refers to the observation that the molecules within an emerged organization $O \subset \mathcal{M}$ are structurally similar. That is, they share certain structural features. This allows one to describe the set of molecules O by a formal language or grammar in a compact way. If we would instead pick a subset A from $\mathcal{M}$ randomly, the expected length of A's description is on the order of its size. Assume, for example, if we pick one million strings of length 100 randomly from the set of all strings of length 100, then we would need about 100 million characters to describe the resulting set. Interestingly, the organizations that appear in implicitly defined ACs can be described much more compactly by a grammar.

Syntactic and semantic closure should be illustrated with an example taken from [27], where $O \subset \mathcal{M}$ is even infinite in size. In particular experiments with the lambda-chemistry, molecules appeared that posses the following structure:

$$A_{i,j} \equiv \lambda x_1.\lambda x_2.\ldots.\lambda x_i.x_j \quad \text{with} \quad j \leq i\,, \tag{19}$$

for example $\lambda x_1.\lambda x_2.\lambda x_3.x_2$. Syntactical closure means that we can specify such structural rules specifying the molecules of O and that these structural rules are invariant under the reaction mechanism. If molecules with the structure given by Eq. (19) react, their product can also be described by Eq. (19).

Semantic closure means that we can describe the reactions taking place within O by referring to the molecule's grammatical structure (e. g. Eq. (19)). In our example, all reactions within O can be described by the following laws (illustrated by Fig. 7):

$$\begin{aligned} \forall i, j > 1, k, l\colon A_{i,j} + A_{k,l} &\Longrightarrow A_{i-1,j-1}\,, \\ \forall i > 1, j = 1, k, l\colon A_{i,j} + A_{k,l} &\Longrightarrow A_{k+i-1,l+i-1}\,. \end{aligned} \tag{20}$$

For example $\lambda x_1.\lambda x_2.\lambda x_3.\lambda x_4.x_3 + \lambda x_1.\lambda x_2.x_2 \Longrightarrow \lambda x_1.\lambda x_2.\lambda x_3.x_2$ and $\lambda x_1.\lambda x_2.x_1 + \lambda x_1.\lambda x_2.\lambda x_3.\lambda x_4.x_3 \Longrightarrow \lambda x_1.\lambda x_2.\lambda x_3.\lambda x_4.\lambda x_5.x_4$, respectively (Fig. 7). As a result of the semantic closure we need not to refer to the underlying reaction mechanism (e. g., the lambda-calculus). Instead, we can explain the reactions within O on a more abstract level by referring to the grammatical structure (e. g., Eq. (19)).

Note that syntactical and semantical closure appears also in real chemical system and is exploited by chemistry to organize chemical explanations. It might even be stated that without this phenomenon, chemistry as we know it would not be possible.

Information Processing

The chemical metaphor gives rise to a variety of computing concepts, which are explained in detail in further articles of this encyclopedia. In approaches like amorphous computing ► Amorphous Computing chemical mechanisms are used as a sub-system for communication between a huge number of simple, spatially distributed processing units. In membrane computing ► Membrane Computing, rewriting operations are not only used to react molecules, but also to model active transport between compartments and to change the spatial compartment structure itself.

Beside these theoretical and in-silico artificial chemistries there are approaches that aim at using real molecules: Spatial patterns like waves and solitons in excitable media can be exploited to compute, see ► Reaction-Diffusion Computing and ► Computing in Geometrical Constrained Excitable Chemical Systems. Whereas other approaches like conformational computing and DNA computing ► DNA Computing do not rely on spatially structured reaction vessels. Other closely related topics are molecular automata ► Molecular Automata and bacterial computing ► Bacterial Computing.

Future Directions

Currently, we can observe a convergence of artificial chemistries and chemical models towards more realistic artificial chemistries [9]. On the one hand, models of systems biology are inspired by techniques from artificial chemistries, e. g., in the domain of rule-based modeling [22,34]. On the other hand, artificial chemistries become more realistic, for example, by adopting more realistic molecular structures or using techniques from computational chemistry to calculate energies [9,57]. Furthermore, the relation between artificial chemistries and real biological systems is made more and more explicit [25, 45,58].

In the future, novel theories and techniques have to be developed to handle complex, implicitly defined reaction systems. This development will especially be driven by the needs of system biology, when implicitly defined models (i. e., rule-based models) become more frequent.

Another open challenge lies in the creation of realistic (chemical) open-ended evolutionary systems. For exam-

ple, an artificial chemistry with "true" open-ended evolution or the ability to show a satisfying long transient phase, which would be comparable to the natural process where novelties continuously appear, has not been presented yet. The reason for this might be lacking computing power, insufficient man-power to implement what we know, or missing knowledge concerning the fundamental mechanism of (chemical) evolution. Artificial chemistries could provide a powerful platform to study whether the mechanisms that we believe explain (chemical) evolution are indeed candidates.

As sketched above, there is a broad range of currently explored practical application domains of artificial chemistries in technical artifacts, such as ambient computing, amorphous computing, organic computing, or smart materials. And eventually, artificial chemistry might come back to the realm of real chemistry and inspire the design of novel computing reaction systems like in the field of molecular computing, molecular communication, bacterial computing, or synthetic biology.

Bibliography

Primary Literature

1. Adami C, Brown CT (1994) Evolutionary learning in the 2D artificial life system avida. In: Brooks RA, Maes P (eds) Prof artificial life IV. MIT Press, Cambridge, pp 377–381. ISBN 0-262-52190-3
2. Adleman LM (1994) Molecular computation of solutions to combinatorical problems. Science 266:1021
3. Bagley RJ, Farmer JD (1992) Spontaneous emergence of a metabolism. In: Langton CG, Taylor C, Farmer JD, Rasmussen S (eds) Artificial life II. Addison-Wesley, Redwood City, pp 93–140. ISBN 0-201-52570-4
4. Banâtre J-P, Métayer DL (1986) A new computational model and its discipline of programming. Technical Report RR-0566. INRIA, Rennes
5. Banzhaf W (1993) Self-replicating sequences of binary numbers – foundations I and II: General and strings of length n = 4. Biol Cybern 69:269–281
6. Banzhaf W (1994) Self-replicating sequences of binary numbers: The build-up of complexity. Complex Syst 8:215–225
7. Banzhaf W (1995) Self-organizing algorithms derived from RNA interactions. In: Banzhaf W, Eeckman FH (eds) Evolution and Biocomputing. LNCS, vol 899. Springer, Berlin, pp 69–103
8. Banzhaf W, Dittrich P, Rauhe H (1996) Emergent computation by catalytic reactions. Nanotechnology 7(1):307–314
9. Benkö G, Flamm C, Stadler PF (2003) A graph-based toy model of chemistry. J Chem Inf Comput Sci 43(4):1085–1093. doi:10.1021/ci0200570
10. Bersini H (2000) Reaction mechanisms in the oo chemistry. In: Bedau MA, McCaskill JS, Packard NH, Rasmussen S (eds) Artificial life VII. MIT Press, Cambridge, pp 39–48
11. Boerlijst MC, Hogeweg P (1991) Spiral wave structure in prebiotic evolution: Hypercycles stable against parasites. Physica D 48(1):17–28
12. Breyer J, Ackermann J, McCaskill J (1999) Evolving reaction-diffusion ecosystems with self-assembling structure in thin films. Artif Life 4(1):25–40
13. Conrad M (1992) Molecular computing: The key-lock paradigm. Computer 25:11–22
14. Dewdney AK (1984) In the game called core war hostile programs engage in a battle of bits. Sci Amer 250:14–22
15. Dittrich P (2001) On artificial chemistries. Ph D thesis, University of Dortmund
16. Dittrich P, Banzhaf W (1998) Self-evolution in a constructive binary string system. Artif Life 4(2):203–220
17. Dittrich P, Speroni di Fenizio P (2007) Chemical organization theory. Bull Math Biol 69(4):1199–1231. doi:10.1007/s11538-006-9130-8
18. Ehricht R, Ellinger T, McCascill JS (1997) Cooperative amplification of templates by cross-hybridization (CATCH). Eur J Biochem 243(1/2):358–364
19. Eigen M (1971) Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften 58(10):465–523
20. Eigen M, Schuster P (1977) The hypercycle: A principle of natural self-organisation, part A. Naturwissenschaften 64(11):541–565
21. Érdi P, Tóth J (1989) Mathematical models of chemical reactions: Theory and applications of deterministic and stochastic models. Pinceton University Press, Princeton
22. Faeder JR, Blinov ML, Goldstein B, Hlavacek WS (2005) Rule-based modeling of biochemical networks. Complexity. doi:10.1002/cplx.20074
23. Farmer JD, Kauffman SA, Packard NH (1986) Autocatalytic replication of polymers. Physica D 22:50–67
24. Fernando C, Rowe J (2007) Natural selection in chemical evolution. J Theor Biol 247(1):152–167. doi:10.1016/j.jtbi.2007.01.028
25. Fernando C, von Kiedrowski G, Szathmáry E (2007) A stochastic model of nonenzymatic nucleic acid replication: Elongators sequester replicators. J Mol Evol 64(5):572–585. doi:10.1007/s00239-006-0218-4
26. Fontana W (1992) Algorithmic chemistry. In: Langton CG, Taylor C, Farmer JD, Rasmussen S (eds) Artificial life II. Addison-Wesley, Redwood City, pp 159–210
27. Fontana W, Buss LW (1994) 'The arrival of the fittest': Toward a theory of biological organization. Bull Math Biol 56:1–64
28. Fontana W, Buss LW (1996) The barrier of objects: From dynamical systems to bounded organization. In: Casti J, Karlqvist A (eds) Boundaries and barriers. Addison-Wesley, Redwood City, pp 56–116
29. Furusawa C, Kaneko K (1998) Emergence of multicellular organisms with dynamic differentiation and spatial pattern. Artif Life 4:79–93
30. Gánti T (1975) Organization of chemical reactions into dividing and metabolizing units: The chemotons. Biosystems 7(1):15–21
31. Giavitto J-L, Michel O (2001) MGS: A rule-based programming language for complex objects and collections. Electron Note Theor Comput Sci 59(4):286–304
32. Gillespie DT (1976) General method for numerically simulating stochastic time evolution of coupled chemical-reaction. J Comput Phys 22(4):403–434
33. Grzybowski BA, Stone HA, Whitesides GM (2000) Dynamic self-

assembly of magnetized, millimetre-sized objects rotating at a liquid-air interface. Nature 405(6790):1033–1036
34. Hlavacek W, Faeder J, Blinov M, Posner R, Hucka M, Fontana W (2006) Rules for modeling signal-transduction systems. Sci STKE 2006:re6
35. Hofbauer J, Sigmund K (1988) Dynamical systems and the theory of evolution. University Press, Cambridge
36. Hofstadter DR (1979) Gödel, Escher, Bach: An eternal golden braid. Basic Books Inc, New York. ISBN 0-465-02685-0
37. Hordijk W, Crutchfield JP, Mitchell M (1996) Embedded-particle computation in evolved cellular automata. In: Toffoli T, Biafore M, Leäo J (eds) PhysComp96. New England Complex Systems Institute, Cambridge, pp 153–8
38. Hosokawa K, Shimoyama I, Miura H (1994) Dynamics of self-assembling systems: Analogy with chemical kinetics. Artif Life 1(4):413–427
39. Hutton TJ (2002) Evolvable self-replicating molecules in an artificial chemistry. Artif Life 8(4):341–356
40. Ikegami T, Hashimoto T (1995) Active mutation in self-reproducing networks of machines and tapes. Artif Life 2(3):305–318
41. Jain S, Krishna S (1998) Autocatalytic sets and the growth of complexity in an evolutionary model. Phys Rev Lett 81(25):5684–5687
42. Jain S, Krishna S (1999) Emergence and growth of complex networks in adaptive systems. Comput Phys Commun 122:116–121
43. Jain S, Krishna S (2001) A model for the emergence of cooperation, interdependence, and structure in evolving networks. Proc Natl Acad Sci USA 98(2):543–547
44. Jain S, Krishna S (2002) Large extinctions in an evolutionary model: The role of innovation and keystone species. Proc Natl Acad Sci USA 99(4):2055–2060. doi:10.1073/pnas.032618499
45. Kaneko K (2007) Life: An introduction to complex systems biology. Springer, Berlin
46. Kauffman SA (1971) Cellular homeostasis, epigenesis and replication in randomly aggregated macromolecular systems. J Cybern 1:71–96
47. Kauffman SA (1986) Autocatalytic sets of proteins. J Theor Biol 119:1–24
48. Kauffman SA (1993) The origins of order: Self-organization and selection in evolution. Oxford University Press, New York
49. Kirner T, Ackermann J, Ehricht R, McCaskill JS (1999) Complex patterns predicted in an in vitro experimental model system for the evolution of molecular cooperation. Biophys Chem 79(3):163–86
50. Kniemeyer O, Buck-Sorlin GH, Kurth W (2004) A graph grammar approach to artificial life. Artif Life 10(4):413–431. doi:10.1162/1064546041766451
51. Laing R (1972) Artificial organisms and autonomous cell rules. J Cybern 2(1):38–49
52. Laing R (1975) Some alternative reproductive strategies in artificial molecular machines. J Theor Biol 54:63–84
53. Laing R (1977) Automaton models of reproduction by self-inspection. J Theor Biol 66:437–56
54. Langton CG (1984) Self-reproduction in cellular automata. Physica D 10D(1–2):135–44
55. Langton CG (1989) Artificial life. In: Langton CG (ed) Proc of artificial life. Addison-Wesley, Redwood City, pp 1–48
56. Lazcano A, Bada JL (2003) The 1953 Stanley L. Miller experiment: Fifty years of prebiotic organic chemistry. Orig Life Evol Biosph 33(3):235–42
57. Lenaerts T, Bersini H (2009) A synthon approach to artificial chemistry. Artif Life 9 (in press)
58. Lenski RE, Ofria C, Collier TC, Adami C (1999) Genome complexity, robustness and genetic interactions in digital organisms. Nature 400(6745):661–4
59. Lohn JD, Colombano S, Scargle J, Stassinopoulos D, Haith GL (1998) Evolution of catalytic reaction sets using genetic algorithms. In: Proc IEEE International Conference on Evolutionary Computation. IEEE, New York, pp 487–492
60. Lugowski MW (1989) Computational metabolism: Towards biological geometries for computing. In: Langton CG (ed) Artificial Life. Addison-Wesley, Redwood City, pp 341–368. ISBN 0-201-09346-4
61. Matsumaru N, Speroni di Fenizio P, Centler F, Dittrich P (2006) On the evolution of chemical organizations. In: Artmann S, Dittrich P (eds) Proc of the 7th german workshop of artificial life. IOS Press, Amsterdam, pp 135–146
62. Maynard Smith J, Szathmáry E (1995) The major transitions in evolution. Oxford University Press, New York
63. McCaskill JS (1988) Polymer chemistry on tape: A computational model for emergent genetics. Internal report. MPI for Biophysical Chemistry, Göttingen
64. McCaskill JS, Chorongiewski H, Mekelburg D, Tangen U, Gemm U (1994) Configurable computer hardware to simulate long-time self-organization of biopolymers. Ber Bunsenges Phys Chem 98(9):1114–1114
65. McMullin B, Varela FJ (1997) Rediscovering computational autopoiesis. In: Husbands P, Harvey I (eds) Fourth european conference on artificial life. MIT Press, Cambridge, pp 38–47
66. Miller SL (1953) A production of amino acids under possible primitive earth conditions. Science 117(3046):528–9
67. Morris HC (1989) Typogenetics: A logic for artificial life. In: Langton CG (ed) Artif life. Addison-Wesley, Redwood City, pp 341–368
68. Ono N, Ikegami T (2000) Self-maintenance and self-reproduction in an abstract cell model. J Theor Biol 206(2):243–253
69. Pargellis AN (1996) The spontaneous generation of digital "life". Physica D 91(1–2):86–96
70. Păun G (2000) Computing with membranes. J Comput Syst Sci 61(1):108–143
71. Petri CA (1962) Kommunikation mit Automaten. Ph D thesis, University of Bonn
72. Rasmussen S, Knudsen C, Feldberg R, Hindsholm M (1990) The coreworld: Emergence and evolution of cooperative structures in a computational chemistry. Physica D 42:111–134
73. Rasmussen S, Knudsen C, Feldberg R (1992) Dynamics of programmable matter. In: Langton CG, Taylor C, Farmer JD, Rasmussen S (eds) Artificial life II. Addison-Wesley, Redwood City, pp 211–291. ISBN 0-201-52570-4
74. Ray TS (1992) An approach to the synthesis of life. In: Langton CG, Taylor C, Farmer JD, Rasmussen S (eds) Artificial life II. Addison-Wesley, Redwood City, pp 371–408
75. Rössler OE (1971) A system theoretic model for biogenesis. Z Naturforsch B 26(8):741–746
76. Sali A, Shakhnovich E, Karplus M (1994) How does a protein fold? Nature 369(6477):248–251
77. Sali A, Shakhnovich E, Karplus M (1994) Kinetics of protein folding: A lattice model study of the requirements for folding to the native state. J Mol Biol 235(5):1614–1636
78. Salzberg C (2007) A graph-based reflexive artificial chemistry. Biosystems 87(1):1–12

79. Sayama H (2009) Swarm chemistry. Artif Life. (in press)
80. Sayama H (1998) Introduction of structural dissolution into Langton's self-reproducing loop. In: Adami C, Belew R, Kitano H, Taylor C (eds) Artificial life VI. MIT Press, Cambridge, pp 114–122
81. Segre D, Ben-Eli D, Lancet D (2000) Compositional genomes: Prebiotic information transfer in mutually catalytic noncovalent assemblies. Proc Natl Acad Sci USA 97(8):4112–4117
82. Socci ND, Onuchic JN (1995) Folding kinetics of proteinlike heteropolymers. J Chem Phys 101(2):1519–1528
83. Speroni di Fenizio P (2000) A less abstract artficial chemistry. In: Bedau MA, McCaskill JS, Packard NH, Rasmussen S (eds) Artificial life VII. MIT Press, Cambridge, pp 49–53
84. Speroni Di Fenizio P, Dittrich P (2002) Artificial chemistry's global dynamics. Movement in the lattice of organisation. J Three Dimens Images 16(4):160–163. ISSN 1342-2189
85. Stadler PF, Fontana W, Miller JH (1993) Random catalytic reaction networks. Physica D 63:378–392
86. Suzuki H (2007) Mathematical folding of node chains in a molecular network. Biosystems 87(2–3):125–135. doi:10.1016/j.biosystems.2006.09.005
87. Suzuki K, Ikegami T (2006) Spatial-pattern-induced evolution of a self-replicating loop network. Artif Life 12(4):461–485. doi:10.1162/artl.2006.12.4.461
88. Suzuki Y, Tanaka H (1997) Symbolic chemical system based on abstract rewriting and its behavior pattern. Artif Life Robotics 1:211–219
89. Tangen U, Schulte L, McCaskill JS (1997) A parallel hardware evolvable computer polyp. In: Pocek KL, Arnold J (eds) IEEE symposium on FPGAs for custopm computing machines. IEEE Computer Society, Los Alamitos
90. Thürk M (1993) Ein Modell zur Selbstorganisation von Automatenalgorithmen zum Studium molekularer Evolution. Ph D thesis, Universität Jena
91. Turing AM (1952) The chemical basis of morphogenesis. Phil Trans R Soc London B 237:37–72
92. Vanderzande C (1998) Lattice models of polymers. Cambridge University Press, Cambridge
93. Varela FJ, Maturana HR, Uribe R (1974) Autopoiesis: The organization of living systems. BioSystems 5(4):187–196
94. Varetto L (1993) Typogenetics: An artificial genetic system. J Theor Biol 160(2):185–205
95. Varetto L (1998) Studying artificial life with a molecular automaton. J Theor Biol 193(2):257–285
96. Vico G (1710) De antiquissima Italorum sapientia ex linguae originibus eruenda librir tres. Neapel
97. von Neumann J, Burks A (ed) (1966) The theory of self-reproducing automata. University of Illinois Press, Urbana
98. Zauner K-P, Conrad M (1996) Simulating the interplay of structure, kinetics, and dynamics in complex biochemical networks. In: Hofestädt R, Lengauer T, Löffler M, Schomburg D (eds) Computer science and biology GCB'96. University of Leipzig, Leipzig, pp 336–338
99. Zeleny M (1977) Self-organization of living systems: A formal model of autopoiesis. Int J General Sci 4:13–28

Books and Reviews

Adami C (1998) Introduction to artificial life. Springer, New York
Dittrich P, Ziegler J, Banzhaf W (2001) Artificial chemistries – a review. Artif Life 7(3):225–275
Hofbauer J, Sigmund K (1998) Evolutionary games and population dynamics. Cambridge University Press, Cambridge

Artificial Intelligence in Modeling and Simulation

Bernard Zeigler[1], Alexandre Muzy[2], Levent Yilmaz[3]
[1] Arizona Center for Integrative Modeling and Simulation, University of Arizona, Tucson, USA
[2] CNRS, Università di Corsica, Corte, France
[3] Auburn University, Alabama, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Review of System Theory and Framework for Modeling and Simulation
Fundamental Problems in M&S
AI-Related Software Background
AI Methods in Fundamental Problems of M&S
Automation of M&S
SES/Model Base Architecture for an Automated Modeler/Simulationist
Intelligent Agents in Simulation
Future Directions
Bibliography

Glossary

Behavior *The observable manifestation of an interaction with a system.*

DEVS Discrete Event System Specification formalism describes models developed for simulation; applications include simulation based testing of collaborative services.

Endomorphic agents *Agents that contain models of themselves and/or of other endomorphic Agents.*

Levels of interoperability Levels at which systems can interoperate such as syntactic, semantic and pragmatic. The higher the level, the more effective is information exchange among participants.

Levels of system specification Levels at which dynamic input/output systems can be described, known, or specified ranging from behavioral to structural.

Metadata Data that describes other data; a hierarchical concept in which metadata are a descriptive abstraction above the data it describes.

Model-based automation Automation of system development and deployment that employs models or system specifications, such as DEVS, to derive artifacts.

Modeling and simulation ontology The SES is interpreted as an ontology for the domain of hierarchical, modular simulation models specified with the DEVS formalism.

Net-centric environment Network Centered, typically Internet-centered or web-centered information exchange medium.

Ontology Language that describes a state of the world from a particular conceptual view and usually pertains to a particular application domain.

Pragmatic frame A means of characterizing the consumer's use of the information sent by a producer; formalized using the concept of processing network model.

Pragmatics Pragmatics is based on Speech Act Theory and focuses on elucidating the intent of the semantics constrained by a given context. Metadata tags to support pragmatics include Authority, Urgency/Consequences, Relationship, Tense and Completeness.

Predicate logic An expressive form of declarative language that can describe ontologies using symbols for individuals, operations, variables, functions with governing axioms and constraints.

Schema An advanced form of XML document definition, extends the DTD concept.

Semantics Semantics determines the content of messages in which information is packaged. The meaning of a message is the eventual outcome of the processing that it supports.

Sensor Device that can sense or detect some aspect of the world or some change in such an aspect.

System specification Formalism for describing or specifying a system. There are levels of system specification ranging from behavior to structure.

Service-oriented architecture Web service architecture in which services are designed to be (1) accessed without knowledge of their internals through well-defined interfaces and (2) readily discoverable and composable.

Structure The internal mechanism that produces the behavior of a system.

System entity structure Ontological basis for modeling and simulation. Its pruned entity structures can describe both static data sets and dynamic simulation models.

Syntax Prescribes the form of messages in which information is packaged.

UML Unified Modeling Language is a software development language and environment that can be used for ontology development and has tools that map UML specifications into XML.

XML eXtensible Markup Language provides a syntax for document structures containing tagged information where tag definitions set up the basis for semantic interpretation.

Definition of the Subject

This article discusses the role of Artificial Intelligence (AI) in Modeling and Simulation (M&S). AI is the field of computer science that attempts to construct computer systems that emulate human problem solving behavior with the goal of understanding human intelligence. M&S is a multidisciplinary field of systems engineering, software engineering, and computer science that seeks to develop robust methodologies for constructing computerized models with the goal of providing tools that can assist humans in all activities of the M&S enterprise. Although each of these disciplines has its core community there have been numerous intersections and cross-fertilizations between the two fields. From the perspective of this article, we view M&S as presenting some fundamental and very difficult problems whose solution may benefit from the concepts and techniques of AI.

Introduction

To state the M&S problems that may benefit from AI we first briefly review a system-theory based framework for M&S that provides a language and concepts to facilitate definitive problem statement. We then introduce some key problem areas: verification and validation, reuse and composability, and distributed simulation and systems of systems interoperability. After some further review of software and AI-related background, we go on to outline some areas of AI that have direct applicability to the just given problems in M&S. In order to provide a unifying theme for the problem and solutions, we then raise the question of whether all of M&S can be automated into an integrated autonomous artificial modeler/simulationist. We then proceed to explore an approach to developing such an intelligent agent and present a concrete means by which such an agent could engage in M&S. We close with consideration of an advanced feature that such an agent must have if it is to fully emulate human capability—the ability, to a limited, but significant extent, to construct and employ models of its own "mind" as well of the "minds" of other agents.

Review of System Theory and Framework for Modeling and Simulation

Hierarchy of System Specifications

Systems theory [1] deals with a hierarchy of system specifications which defines levels at which a system may be known or specified. Table 1 shows this Hierarchy of System Specifications (in simplified form, see [2] for full exposition).

- At level 0 we deal with the input and output interface of a system.
- At level 1 we deal with purely observational recordings of the behavior of a system. This is an I/O relation which consists of a set of pairs of input behaviors and associated output behaviors.
- At level 2 we have knowledge of the initial state when the input is applied. This allows partitioning the input/output pairs of level 1 into non-overlapping subsets, each subset associated with a different starting state.
- At level 3 the system is described by state space and state transition functions. The transition function describes the state-to-state transitions caused by the inputs and the outputs generated thereupon.
- At level 4 a system is specified by a set of components and a coupling structure. The components are systems on their own with their own state set and state transition functions. A coupling structure defines how those interact. A property of a coupled system that is called "closure under coupling" guarantees that a coupled system at level 3 itself specifies a system. This property allows hierarchical construction of systems, i. e., that coupled systems can be used as components in larger coupled systems.

As we shall see in a moment, the system specification hierarchy provides a mathematical underpinning to define a framework for modeling and simulation. Each of the entities (e. g., real world, model, simulation, and experimental frame) will be described as a system known or specified at some level of specification. The essence of modeling and simulation lies in establishing relations between pairs of system descriptions. These relations pertain to the validity of a system description at one level of specification relative to another system description at a different (higher, lower, or equal) level of specification.

On the basis of the arrangement of system levels as shown in Table 1, we distinguish between vertical and horizontal relations. A vertical relation is called an association mapping and takes a system at one level of specification and generates its counterpart at another level of specification. The downward motion in the structure-to-behavior direction, formally represents the process by which the behavior of a model is generated. This is relevant in simulation and testing when the model generates the behavior which then can be compared with the desired behavior.

The opposite upward mapping relates a system description at a lower level with one at a higher level of specification. While the downward association of specifications is straightforward, the upward association is much less so. This is because in the upward direction information is introduced while in the downward direction information is reduced. Many structures exhibit the same behavior and recovering a unique structure from a given behavior is not possible. The upward direction, however, is fundamental in the design process where a structure (system at level 3) has to be found that is capable of generating the desired behavior (system at level 1).

Framework for Modeling and Simulation

The *Framework for M&S* as described in [2] establishes *entities* and their *relationships* that are central to the M&S enterprise (see Fig. 1. The entities of the Framework are: *source system, model, simulator*, and *experimental frame*; they are related by the *modeling* and the *simulation* relationships. Each entity is formally characterized as a sys-

Artificial Intelligence in Modeling and Simulation, Table 1
Hierarchy of system specifications

Level	Name	What we specify at this level
4	Coupled systems	System built up by several component systems that are coupled together
3	I/O System	System with state and state transitions to generate the behavior
2	I/O Function	Collection of input/output pairs constituting the allowed behavior partitioned according to the initial state the system is in when the input is applied
1	I/O Behavior	Collection of input/output pairs constituting the allowed behavior of the system from an external Black Box view
0	I/O Frame	Input and output variables and ports together with allowed values

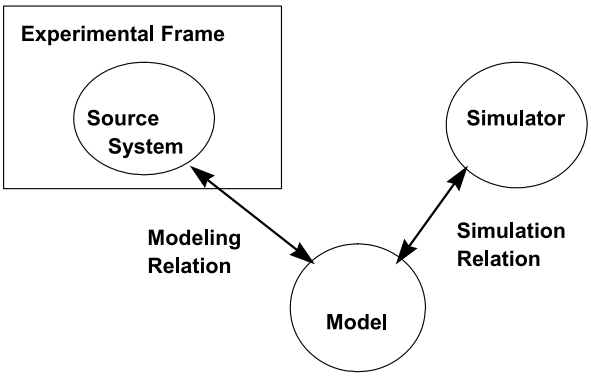

Artificial Intelligence in Modeling and Simulation, Figure 1
Framework entities and relationships

tem at an appropriate level of specification of a generic dynamic system.

Source System

The source system is the real or virtual environment that we are interested in modeling. It is viewed as a source of observable data, in the form of time-indexed trajectories of variables. The data that has been gathered from observing or otherwise experimenting with a system is called the system behavior database. These data are viewed or acquired through experimental frames of interest to the model development and user. As we shall see, in the case of model validation, these data are the basis for comparison with data generated by a model. Thus, these data must be sufficient to enable reliable comparison as well as being accepted by both the model developer and the test agency as the basis for comparison. Data sources for this purpose might be measurement taken in prior experiments, mathematical representation of the measured data, or expert knowledge of the system behavior by accepted subject matter experts.

Experimental Frame

An experimental frame is a specification of the conditions under which the system is observed or experimented with [3]. An experimental frame is the operational formulation of the objectives that motivate a M&S project. A frame is realized as a system that interacts with the system of interest to obtain the data of interest under specified conditions.

An experimental frame specification consists of four major subsections:

Input stimuli Specification of the class of admissible input time-dependent stimuli. This is the class from which individual samples will be drawn and injected into the model or system under test for particular experiments.

Control Specification of the conditions under which the model or system will be initialized, continued under examination, and terminated.

Metrics Specification of the data summarization functions and the measures to be employed to provide quantitative or qualitative measures of the input/output behavior of the model. Examples of such metrics are performance indices, goodness of fit criteria, and error accuracy bound.

Analysis Specification of means by which the results of data collection in the frame will be analyzed to arrive at final conclusions. The data collected in a frame consists of pairs of input/output time functions.

When an experimental frame is realized as a system to interact with the model or system under test the specifications become components of the driving system. For example, a generator of output time functions implements the class of input stimuli.

An experimental frame is the operational formulation of the objectives that motivate a modeling and simulation project. Many experimental frames can be formulated for the same system (both source system and model) and the same experimental frame may apply to many systems. Why would we want to define many frames for the same system? Or apply the same frame to many systems? For the same reason that we might have different objectives in modeling the same system, or have the same objective in modeling different systems. There are two equally valid views of an experimental frame. One, views a frame as a definition of the type of data elements that will go into the database. The second views a frame as a system that interacts with the system of interest to obtain the data of interest under specified conditions. In this view, the frame is characterized by its implementation as a measurement system or observer. In this implementation, a frame typically has three types of components (as shown in Fig. 2 and Fig. 3): a *generator*, that generates input segments to the system; *acceptor* that monitors an experiment to see the desired experimental conditions are met; and *transducer* that observes and analyzes the system output segments.

Figure 2b illustrates a simple, but ubiquitous, pattern for experimental frames that measure typical job processing performance metrics, such as round trip time and throughput. Illustrated in the web context, a generator produces service request messages at a given rate. The time that has elapsed between sending of a request and its return from a server is the round trip time. A transducer notes the departures and arrivals of requests allow-

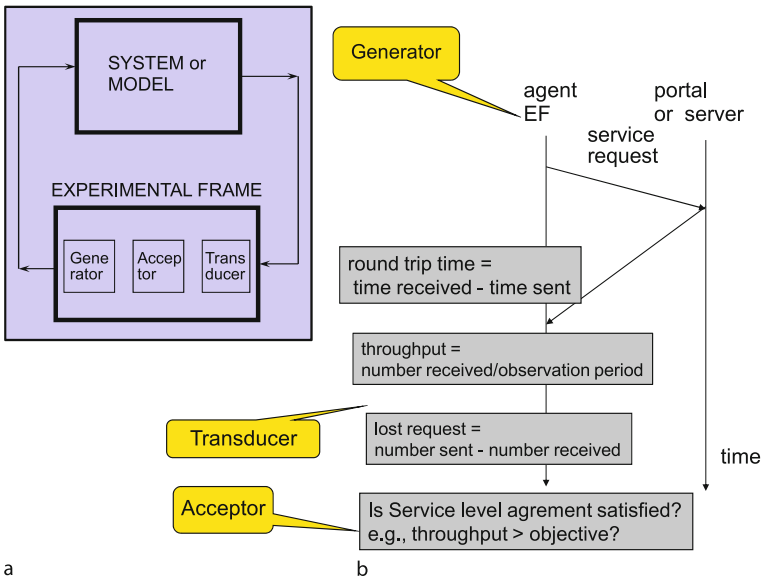

Artificial Intelligence in Modeling and Simulation, Figure 2
Experimental frame and components

ing it to compute the average round trip time and other related statistics, as well as the throughput and unsatisfied (or lost) requests. An acceptor notes whether performance achieves the developer's objectives, for example, whether the throughput exceeds the desired level and/or whether say 99% of the round trip times are below a given threshold.

Objectives for modeling relate to the role of the model in systems design, management or control. Experimental frames translate the objectives into more precise experimentation conditions for the source system or its models. We can distinguish between objectives concerning those for verification and validation of (a) models and (b) systems. In the case of models, experimental frames translate the objectives into more precise experimentation conditions for the source system and/or its models. A model under test is expected to be valid for the source system in each such frame. Having stated the objectives, there is presumably a best level of resolution to answer these questions. The more demanding the questions, the greater the resolution likely to be needed to answer them. Thus, the choice of appropriate levels of abstraction also hinges on the objectives and their experimental frame counterparts.

In the case of objectives for verification and validation of systems, we need to be given, or be able to formulate, the requirements for the behavior of the system at the IO behavior level. The experimental frame then is formulated to translate these requirements into a set of possible experiments to test whether the system actually performs its

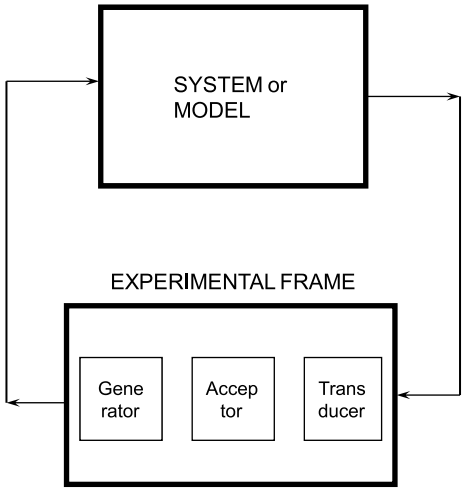

Artificial Intelligence in Modeling and Simulation, Figure 3
Experimental frame and its components

required behavior. In addition we can formulate measures of the effectiveness (MOE) of a system in accomplishing its goals. We call such measures, *outcome* measures. In order to compute such measures, the system must expose relevant variables, we'll call these *output* variables, whose values can be observed during execution runs of the system.

Model

A model is a system specification, such as a set of instructions, rules, equations, or constraints for generating in-

put/output behavior. Models may be expressed in a variety of formalisms that may be understood as means for specifying subclasses of dynamic systems. The Discrete Event System Specification (DEVS) formalism delineates the subclass of discrete event systems and it can also represent the systems specified within traditional formalisms such as differential (continuous) and difference (discrete time) equations [4]. In DEVS, as in systems theory, a model can be *atomic*, i. e., not further decomposed, or *coupled*, in which case it consists of components that are coupled or interconnected together.

Simulator

A simulator is any computation system (such as a single processor, or a processor network, or more abstractly an algorithm), capable of executing a model to generate its behavior.

The more general purpose a simulator is the greater the extent to which it can be configured to execute a variety of model types. In order of increasing capability, simulators can be:

- Dedicated to a particular model or small class of similar models
- Capable of accepting all (practical) models from a wide class, such as an application domain (e. g., communication systems)
- Restricted to models expressed in a particular modeling formalism, such as continuous differential equation models
- Capable of accepting multi-formalism models (having components from several formalism classes, such as continuous and discrete event).

A simulator can take many forms such as on a single computer or multiple computers executing on a network.

Fundamental Problems in M&S

We have now reviewed a system-theory-based framework for M&S that provides a language and concepts in which to formulate key problems in M&S. Next on our agenda is to discuss problem areas including: verification and validation, reuse and composability, and distributed simulation and systems of systems interoperability. These are challenging, and heretofore, unsolved problems at the core of the M&S enterprise.

Validation and Verification

The basic concepts of verification and validation (V&V) have been described in different settings, levels of details, and points of views and are still evolving. These concepts have been studied by a variety of scientific and engineering disciplines and various flavors of validation and verification concepts and techniques have emerged from a modeling and simulation perspective. Within the modeling and simulation community, a variety of methodologies for V&V have been suggested in the literature [5,6,7]. A categorization of 77 verification, validation and testing techniques along with 15 principles has been offered to guide the application of these techniques [8]. However, these methods vary extensively – e. g., alpha testing, induction, cause and effect graphing, inference, predicate calculus, proof of correctness, and user interface testing and are only loosely related to one another. Therefore, such a categorization can only serve as an informal guideline for the development of a process for V&V of models and systems.

Validation and verification concepts are themselves founded on more primitive concepts such as system specifications and homomorphism as discussed in the framework of M&S [2]. In this framework, the entities *system, experimental frame, model, simulator* take on real importance only when properly related to each other. For example, we build a model of a particular system for some objective and only some models, and not others, are suitable. Thus, it is critical to the success of a simulation modeling effort that certain relationships hold. Two of the most important are *validity* and *simulator correctness*.

The basic modeling relation, *validity*, refers to the relation between a model, a source system and an *experimental frame*. The most basic concept, *replicative validity*, is affirmed if, for all the experiments possible within the experimental frame, the behavior of the model and system agree within acceptable tolerance. The term *accuracy* is often used in place of validity. Another term, *fidelity*, is often used for a combination of both validity and detail. Thus, a high fidelity model may refer to a model that is both highly detailed and valid (in some understood experimental frame). However, when used this way, the assumption seems to be that high detail alone is needed for high fidelity, as if validity is a necessary consequence of high detail. In fact, it is possible to have a very detailed model that is nevertheless very much in error, simply because some of the highly resolved components function in a different manner than their real system counterparts.

The basic approach to model validation is comparison of the behavior generated by a model and the source system it represents within a given experimental frame. The basis for comparison serves as the reference Fig. 4 against which the accuracy of the model is measured.

The basic *simulation* relation, simulator correctness, is a relation between a *simulator* and a *model*. A *simulator*

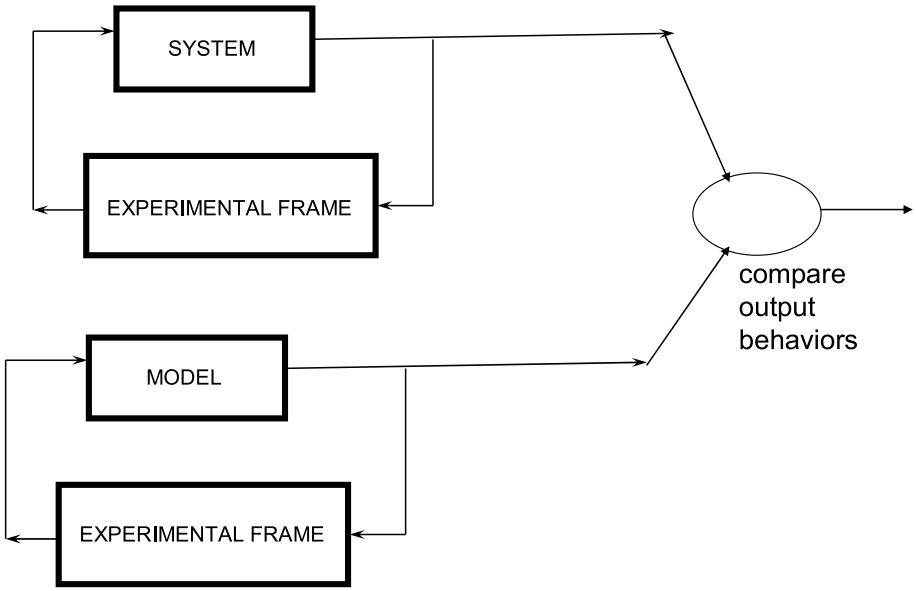

Artificial Intelligence in Modeling and Simulation, Figure 4
Basic approach to model validation

correctly simulates a model if it is guaranteed to faithfully generate the model's output behavior given its initial state and its input trajectory. In practice, as suggested above, simulators are constructed to execute not just one model but also a family of possible models. For example, a network simulator provides both a simulator and a class of network models it can simulate. In such cases, we must establish that a simulator will correctly execute the particular class of models it claims to support. Conceptually, the approach to testing for such execution, illustrated in Fig. 5, is to perform a number of test cases in which the same model is provided to the simulator under test and to a "gold standard simulator" which is known to correctly simulate the model. Of course such test case models must lie within the class supported by the simulat or under test as well as presented in the form that it expects to receive them. Comparison of the output behaviors in the same manner as with model validation is then employed to check the agreement between the two simulators.

If the specifications of both the simulator and the model are available in separated form where each can be accessed independently, it may be possible to prove correctness mathematically.

The case of system validation is illustrated in Fig. 6. Here the system is considered as a hardware and/or software implementation to be validated against requirements for its input/output behavior. The goal is to develop test models that can stimulate the implemented system with inputs and can observe its outputs to compare them with those required by the behavior requirements. Also shown is a dotted path in which a reference model is constructed that is capable of simulation execution. Construction of such a reference model is more difficult to develop than the test models since it requires not only knowing in advance what output to test for, but to actually to generate such an output. Although such a reference model is not required, it may be desirable in situations in which the extra cost of development is justified by the additional range of tests that might be possible and the consequential increased coverage this may provide.

Model Reuse and Composability

Model reuse and composability are two sides of the same coin—it is patently desirable to reuse models, the fruits of earlier or others work. However, typically such models will become components in a larger composite model and must be able to interact meaningfully with them. While software development disciplines are successfully applying a component-based approach to build software systems, the additional systems dynamics involved in simulation models has resisted straight forward reuse and composition approaches. A model is only reusable to the extent that its original dynamic systems assumptions are consistent with the constraints of the new simulation application. Consequently, without contextual information to guide selection and refactoring, a model may not be reused to advantage within a new experimental frame. Davis and

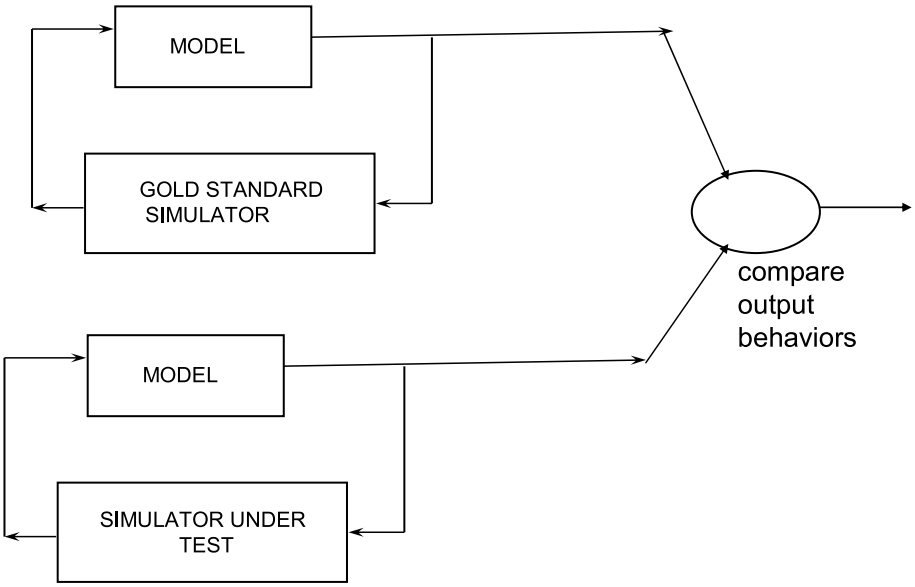

Artificial Intelligence in Modeling and Simulation, Figure 5
Basic approach to simulator verification

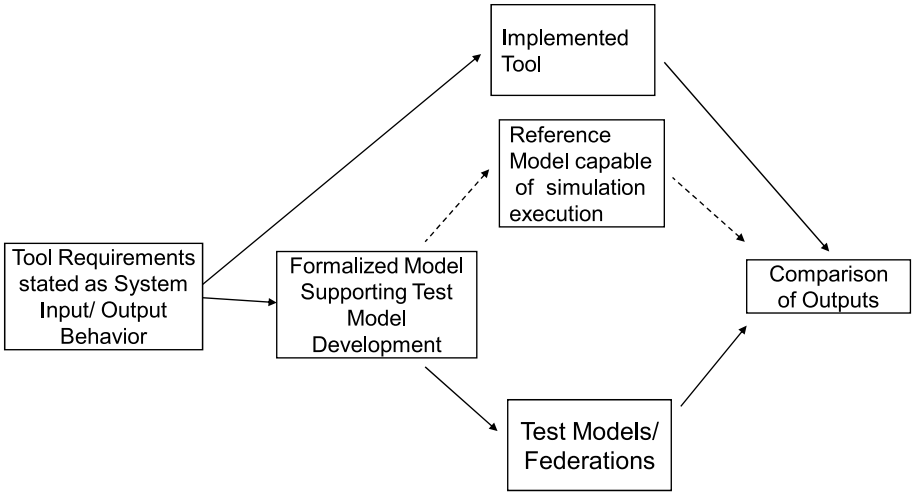

Artificial Intelligence in Modeling and Simulation, Figure 6
Basic approach to system validation

Anderson [9] argue that to foster such reuse, model representation methods should distinguish, and separately specify, the model, simulator, and the experimental frame. However, Yilmaz and Oren [10] pointed out that more contextual information is needed beyond the information provided by the set of experimental frames to which a model is applicable [11], namely, the characterization of the context in which the model was constructed. These authors extended the basic model-simulator-experimental frame perspective to emphasize the role of context in reuse. They make a sharp distinction between the objective context within which a simulation model is originally defined and the intentional context in which the model is being qualified for reuse. They extend the system theoretic levels of specification discussed earlier to define certain behavioral model dependency relations needed to formalize conceptual, realization, and experimental aspects of context.

As the scope of simulation applications grows, it is increasingly the case that more than one modeling paradigm is needed to adequately express the dynamics of the different components. For systems composed of models with dynamics that are intrinsically heterogeneous, it is crucial to use multiple modeling formalisms to describe them. However, combining different model types poses a variety of challenges [9,12,13]. Sarjoughian [14], introduced an approach to multi-formalism modeling that employs an interfacing mechanism called a Knowledge Interchange Broker to compose model components expressed in diverse formalisms. The KIB supports translation from the semantics of one formalism into that of a second to ensure coordinated and correct execution simulation algorithms of distinct modeling formalisms.

Distributed Simulation and System of Systems Interoperability

The problems of model reuse and composability manifest themselves strongly in the context of distributed simulation where the objective is to enable existing geographically dispersed simulators to meaningfully interact, or federate, together. We briefly review experience with interoperability in the distributed simulation context and a linguistically based approach to the System of Systems (SoS) interoperability problem [15]. Sage and Cuppan [16] drew the parallel between viewing the construction of SoS as a federation of systems and the federation that is supported by the High Level Architecture (HLA), an IEEE standard fostered by the DoD to enable composition of simulations [17,18].

HLA is a network middleware layer that supports message exchanges among simulations, called federates, in a neutral format. However, experience with HLA has been disappointing and forced acknowledging the difference between technical interoperability and substantive interoperability [19]. The first only enables heterogeneous simulations to exchange data but does not guarantee the second, which is the desired outcome of exchanging meaningful data, namely, that coherent interaction among federates takes place. Tolk and Muguirra [20] introduced the Levels of Conceptual Interoperability Model (LCIM) which identified seven levels of interoperability among participating systems. These levels can be viewed as a refinement of the *operational* interoperability type which is one of three defined by [15]. The operational type concerns linkages between systems in their interactions with one another, the environment, and with users. The other types apply to the context in which systems are constructed and acquired. They are *constructive*—relating to linkages between organizations responsible for system construction and *programmatic*—linkages between program offices to manage system acquisition.

AI-Related Software Background

To proceed to the discussion of the role of AI in addressing key problems in M&S, we need to provide some further software and AI-related background. We offer a brief historical account of object-orientation and agent-based systems as a springboard to discuss the upcoming concepts of object frameworks, ontologies and endomorphic agents.

Object-Orientation and Agent-Based Systems

Many of the software technology advances of the last 30 years have been initiated from the field of M&S. Objects, as code modules with both structure and behavior, were first introduced in the SIMULA simulation language [21]. Objects blossomed in various directions and became institutionalized in the widely adopted programming language C++ and later in the infrastructure for the web in the form of Java [22] and its variants. The freedom from straight-line procedural programming that object-orientation championed was taken up in AI in two directions: various forms of knowledge representation and of autonomy. Rule-based systems aggregate modular if-then logic elements—the rules—that can be activated in some form of causal sequence (inference chains) by an execution engine [23]. In their passive state, rules represent static discrete pieces of inferential logic, called declarative knowledge. However, when activated, a rule influences the state of the computation and the activation of subsequent rules, providing the system a dynamic or procedural, knowledge characteristic as well. Frame-based systems further expanded knowledge representation flexibility and inferencing capability by supporting slots and constraints on their values—the frames—as well as their taxonomies based on generalization/specialization relationships [24]. Convergence with object-orientation became apparent in that frames could be identified as objects and their taxonomic organization could be identified with classes within object-style organizations that are based on sub-class hierarchies.

On the other hand, the modular nature of objects together with their behavior and interaction with other objects, led to the concept of agents which embodied increased autonomy and self-determination [25]. Agents represent individual threads of computation and are typically deployable in distributed form over computer networks where they interact with local environments and communicate/coordinate with each other. A wide variety

of agent types exists in large part determined by the variety and sophistication of their processing capacities—ranging from agents that simply gather information on packet traffic in a network to logic-based entities with elaborate reasoning capacity and authority to make decisions (employing knowledge representations just mentioned.) The step from agents to their aggregates is natural, thus leading to the concept of multi-agent systems or societies of agents, especially in the realm of modeling and simulation [26].

To explore the role of AI in M&S at the present, we project the just-given historical background to the concurrent concepts of object frameworks, ontologies and endomorphic agents. The Unified Modeling Language (UML) is gaining a strong foothold as the defacto standard for object-based software development. Starting as a diagrammatical means of software system representation, it has evolved to a formally specified language in which the fundamental properties of objects are abstracted and organized [27,28]. Ontologies are models of the world relating to specific aspects or applications that are typically represented in frame-based languages and form the knowledge components for logical agents on the Semantic Web [29]. A convergence is underway that re-enforces the commonality of the object-based origins of AI and software engineering. UML is being extended to incorporate ontology representations so that software systems in general will have more explicit models of their domains of operation. As we shall soon see, endomorphic agents refer to agents that include abstractions of their own structure and behavior within their ontologies of the world [30].

The M&S Framework Within Unified Modeling Language (UML)

With object-orientation as unified by UML and some background in agent-based systems, we are in a position to discuss the computational realization of the M&S framework discussed earlier. The computational framework is based on the Discrete Event System Specification (DEVS) formalism and implemented in various object-oriented environments. Using UML we can represent the framework as a set of classes and relations as illustrated in Figs. 7 and 8. Various software implementations of DEVS support different subsets of the classes and relations. In particular, we mention a recent implementation of DEVS within a Service-Oriented Architecture (SOA) environment called DEVS/SOA [31,32]. This implementation exploits some of the benefits afforded by the web environment mentioned earlier and provides a context for consideration of the primary target of our discussion, comprehensive automation of the M&S enterprise.

We use one of the UML constructs, the use case diagram, to depict the various capabilities that would be involved in automating all or parts of the M&S enterprise. Use cases are represented by ovals that connect to at least one actor (stick figure) and to other use cases through "includes" relations, shown as dotted arrows. For example, a sensor (actor) collects data (use case) which includes storage of data (use case). A memory actor stores and retrieves models which include storage and retrieval (respectively) of data. Constructing models includes retriev-

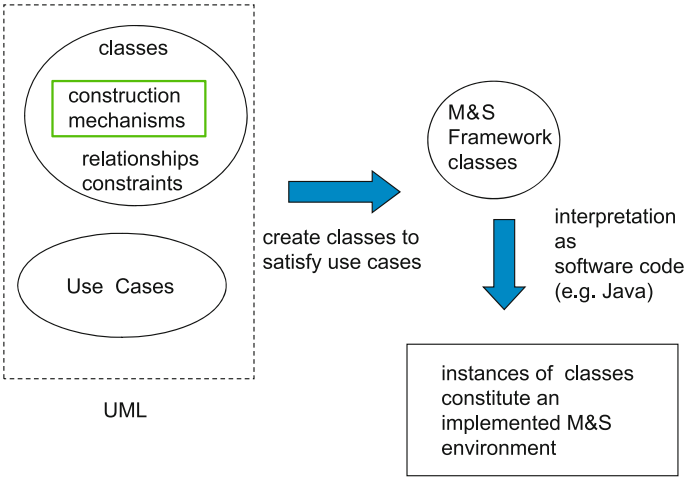

Artificial Intelligence in Modeling and Simulation, Figure 7
M&S framework formulated within UML

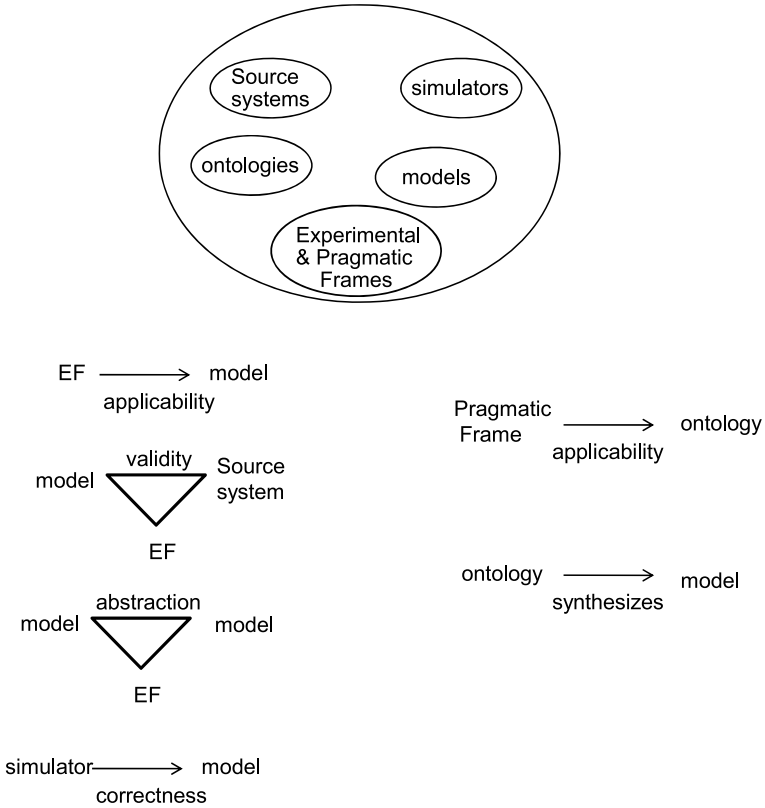

Artificial Intelligence in Modeling and Simulation, Figure 8
M&S framework classes and relations in a UML representation

ing stored data within an experimental frame. Validating models includes retrieving models from memory as components and simulating the composite model to generate data within an experimental frame. The emulator-simulator actor does the simulating to execute the model so that its generated behavior can be matched against the stored data in the experimental frame. The objectives of the human modeler drive the model evaluator and hence the choice of experimental frames to consider as well as models to validate. Models can be used in (at least) two time frames [33]. In the long term, they support planning of actions to be taken in the future. In the short term, models support execution and control in real-time of previously planned actions.

AI Methods in Fundamental Problems of M&S

The enterprise of modeling and simulation is characterized by activities such as model, simulator and experimental frame creation, construction, reuse, composition, verification and validation. We have seen that valid model construction requires significant expertise in all the components of the M&S enterprise, e. g., modeling formalisms, simulation methods, and domain understanding and knowledge. Needless to say, few people can bring all such elements to the table, and this situation creates a significant bottleneck to progress in such projects. Among the contributing factors are lack of trained personnel that must be brought in, expense of such high capability experts, and the time needed to construct models to the resolution required for most objectives. This section introduces some AI-related technologies that can ameliorate this situation: Service-Oriented Architecture and Semantic Web, ontologies, constrained natural language capabilities, and genetic algorithms. Subsequently we will consider these as components in unified, comprehensive, and autonomous automation of M&S.

Service-Oriented Architecture and Semantic Web

On the World Wide Web, a Service-Oriented Architecture is a market place of open and discoverable web-services

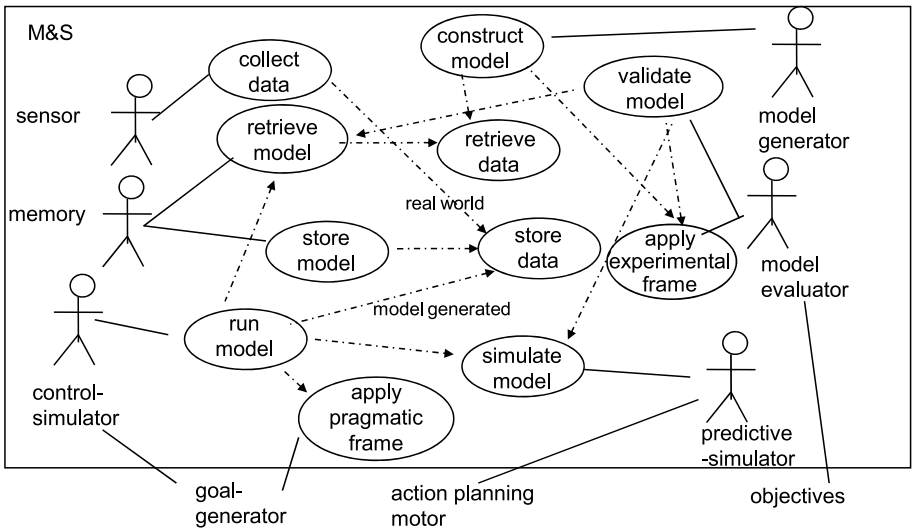

Artificial Intelligence in Modeling and Simulation, Figure 9
UML use case formulation of the overall M&S enterprise

incorporating, as they mature, Semantic Web technologies [34]. The eXtensible Markup Language (XML) is the standard format for encoding data sets and there are standards for sending and receiving XML [35]. Unfortunately, the problem just starts at this level. There are myriad ways, or Schemata, to encode data into XML and a good number of such Schemata have already been developed. More often than not, they are different in detail when applied to the same domains. What explains this incompatibility?

In a Service-Oriented Architecture, the producer sends messages containing XML documents generated in accordance with a schema. The consumer receives and interprets these messages using the same schema in which they were sent. Such a message encodes a world state description (or changes in it) that is a member of a set delineated by an ontology. The ontology takes into account the pragmatic frame, i. e., a description of how the information will be used in downstream processing. In a SOA environment, data dissemination may be dominated by "user pull of data," incremental transmission, discovery using metadata, and automated retrieval of data to meet user pragmatic frame specifications. This is the SOA concept of data-centered, interface-driven, loose coupling between producers and consumers. The SOA concept requires the development of platform-independent, community-accepted, standards that allow raw data to be syntactically packaged into XML and accompanied by metadata that describes the semantic and pragmatic information needed to effectively process the data into increasingly higher-value products downstream.

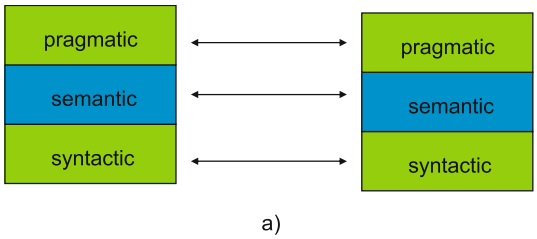

Artificial Intelligence in Modeling and Simulation, Figure 10
Interoperability levels in distributed simulation

Ontologies

Semantic Web researchers typically seek to develop intelligent agents that can draw logical inferences from diverse, possibly contradictory, ontologies such as a web search might discover. Semantic Web research has led to a focus on ontologies [34]. These are logical languages that provide a common vocabulary of terms and axiomatic relations among them for a subject area. In contrast, the newly emerging area of ontology integration assumes that human understanding and collaboration will not be replaced by intelligent agents. Therefore, the goal is to create concepts and tools to help people develop practical solutions to incompatibility problems that impede "effective" exchange of data and ways of testing that such solutions have been correctly implemented.

As illustrated in Fig. 10, interoperability of systems can be considered at three linguistically-inspired levels: *syntactic, semantic*, and *pragmatic*. The levels are summarized in Table 2. More detail is provided in [36].

Artificial Intelligence in Modeling and Simulation, Table 2
Linguistic levels

Linguistic Level	A collaboration of systems or services interoperates at this level if:	Examples
Pragmatic – how information in messages is used	The receiver reacts to the message in a manner that the sender intends	An order from a commander is obeyed by the troops in the field as the commander intended. A necessary condition is that the information arrives in a timely manner and that its meaning has been preserved (semantic interoperability)
Semantic – shared understanding of meaning of messages	The receiver assigns the same meaning as the sender did to the message.	An order from a commander to multi -national participants in a coalition operation is understood in a common manner despite translation into different languages. A necessary condition is that the information can be unequivocally extracted from the data (syntactic interoperability)
Syntactic – common rules governing composition and transmitting of messages	The consumer is able to receive and parse the sender's message	A common network protocol (e. g. IPv4) is employed ensuring that all nodes on the network can send and receive data bit arrays adhering to a prescribed format.

Constrained Natural Language

Model development can be substantially aided by enabling users to specify modeling constructs using some form of constrained natural language [37]. The goal is to overcome modeling complexity by letting users with limited or nonexistent formal modeling or programming background convey essential information using natural language, a form of expression that is natural and intuitive. Practicality demands constraining the actual expressions that can be used so that the linguistic processing is tractable and the input can be interpreted unambiguously. Some techniques allow the user to narrow down essential components for model construction. Their goal is to reduce ambiguity between the user's requirements and essential model construction components. A natural language interface allows model specification in terms of a verb phrase consisting of a verb, noun, and modifier, for example "build car quickly." Conceptual realization of a model from a verb phrase ties in closely with Checkland's [38] insight that an appropriate verb should be used to express the root definition, or core purpose, of a system. The main barrier between many people and existing modeling software is their lack of computer literacy and this provides an incentive to develop natural language interfaces as a means of bridging this gap. Natural language expression could create modelers out of people who think semantically, but do not have the requisite computer skills to express these ideas. A semantic representation frees the user to explore the system on the familiar grounds of natural language and opens the way for brain storming, innovation and testing of models before they leave the drawing board.

Genetic Algorithms

The genetic algorithm is a subset of evolutionary algorithms that model biological processes to search in highly complex spaces. A genetic algorithm (GA) allows a population composed of many individuals to evolve under specified selection rules to a state that maximizes the "fitness." The theory was developed by John Holland [39] and popularized by Goldberg who was able to solve a difficult problem involving the control of gas pipeline transmission for his dissertation [40]. Numerous applications of GAs have since been chronicled [41,42]. Recently, GAs have been applied to cutting-edge problems in automated construction of simulation models, as discussed below [43].

Automation of M&S

We are now ready to suggest a unifying theme for the problems in M&S and possible AI-based solutions, by raising the question of whether all of M&S can be automated into an integrated autonomous artificial modeler/simulationist. First, we provide some background needed to explore an approach to developing such an intelligent agent based on the System Entity Structure/Model Base framework, a hybrid methodology that combines elements of AI and M&S.

System Entity Structure

The System Entity Structure (SES) concepts were first presented in [44]. They were subsequently extended and implemented in a knowledge-based design environment [45]. Application to model base management originated with [46]. Subsequent formalizations and imple-

mentations were developed in [47,48,49,50,51]. Applications to various domains are given in [52].

A System Entity Structure is a knowledge representation formalism which focuses on certain elements and relationships that relate to M&S. Entities represent things that exist in the real world or sometimes in an imagined world. Aspects represent ways of decomposing things into more fine grained ones. Multi-aspects are aspects for which the components are all of one kind. Specializations represent categories or families of specific forms that a thing can assume provides the means to represent a family of models as a labeled tree. Two of its key features are support for decomposition and specialization. The former allows decomposing a large system into smaller systems. The latter supports representation of alternative choices. Specialization enables representing a generic model (e. g., a computer display model) as one of its specialized variations (e. g., a flat panel display or a CRT display). On the basis of SES axiomatic specifications, a family of models (design-space) can be represented and further automatically pruned to generate a simulation model. Such models can be systematically studied and experimented with based on alternative design choices. An important, salient feature of SES is its ability to represent models not only in terms of their decomposition and specialization, but also their aspects. The SES represents alternative decompositions via aspects. The system entity structure (SES) formalism provides an operational language for specifying such hierarchical structures. An SES is a structural knowledge representation scheme that systematically organizes a family of possible structures of a system. Such a family characterizes decomposition, coupling, and taxonomic relationships among entities. An *entity* represents a real world object. The decomposition of an entity concerns how it may be broken down into sub-entities. In addition, coupling specifications tell how sub-entities may be coupled together to reconstitute the entity and be associated with an aspect. The *taxonomic* relationship concerns admissible variants of an entity. The SES/Model-Base framework [52] is a powerful means to support the plan-generate-evaluate paradigm in systems design. Within the framework, entity structures organize models in a model base. Thus, modeling activity within the framework consists of three subactivities: specification of model composition structure, specification of model behavior, and synthesis of a simulation model.

The SES is governed by an axiomatic framework in which entities alternate with the other items. For example, a thing is made up of parts; therefore, its entity representation has a corresponding aspect which, in turn, has entities representing the parts. A System Entity Structure specifies a family of hierarchical, modular simulation models, each of which corresponds to a complete pruning of the SES. Thus, the SES formalism can be viewed as an ontology with the set of all simulation models as its domain of discourse. The mapping from SES to the Systems formalism, particularly to the DEVS formalism, is discussed in [36]. We note that simulation models include both static and dynamic elements in any application domain, hence represent an advanced form of ontology framework.

SES/Model Base Architecture for an Automated Modeler/Simulationist

In this section, we raise the challenge of creating a fully automated modeler/simulationist that can autonomously carry out all the separate functions identified in the M&S framework as well as the high level management of these functions that is currently under exclusively human control. Recall the use case diagrams in Fig. 9 that depict the various capabilities that would need to be involved in realizing a completely automated modeler/simulationist. To link up with the primary modules of mind, we assign model construction to the belief generator—interpreting beliefs as models [54]. Motivations outside the M&S component drive the belief evaluator and hence the choice of experimental frames to consider as well as models to validate. External desire generators stimulate the imaginer/envisioner to run models to make predictions within pragmatic frames and assist in action planning.

The use case diagram of Fig. 9, is itself a model of how modeling and simulation activities may be carried out within human minds. We need not be committed to particular details at this early stage, but will assume that such a model can be refined to provide a useful representation of human mental activity from the perspective of M&S. This provides the basis for examining how such an artificially intelligent modeler/simulationist might work and considering the requirements for comprehensive automation of M&S.

The SES/MB methodology, introduced earlier, provides a basis for formulating a conceptual architecture for automating all of the M&S activities depicted earlier in Fig. 9 into an integrated system. As illustrated in Fig. 11, the dichotomy into real-time use of models for acting in the real world and the longer-term development of models that can be employed in such real-time activities is manifested in the distinction between passive and active models. Passive models are stored in a repository that can be likened to long-term memory. Such models under go the life-cycle mentioned earlier in which they are validated and employed within experimental frames of interest for

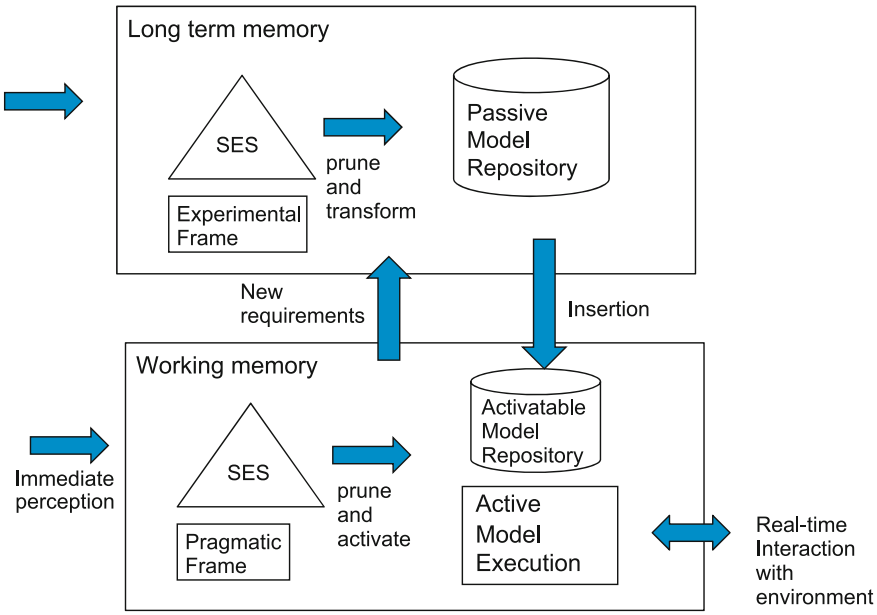

Artificial Intelligence in Modeling and Simulation, Figure 11
SES/Model base architecture for automated M&S

long-term forecasting or decision-making. However, in addition to this quite standard concept of operation, there is an input pathway from the short term, or working, memory in which models are executed in real-time. Such execution, in real-world environments, often results in deficiencies, which provide impetus and requirements for instigating new model construction. Whereas long-term model development and application objectives are characterized by experimental frames, the short-term execution objectives are characterized by pragmatic frames. As discussed in [36], a pragmatic frame provides a means of characterizing the use for which an executable model is being sought. Such models must be simple enough to execute within, usually, stringent real-time deadlines.

The M&S Framework Within Mind Architecture

An influential formulation of recent work relating to mind and brain [53] views mind as the behavior of the brain, where mind is characterized by a massively modular architecture. This means that mind is, composed of a large number of modules, each responsible for different functions, and each largely independent and sparsely connected with others. Evolution is assumed to favor such differentiation and specialization since under suitably weakly interactive environments they are less redundant and more efficient in consuming space and time resources. Indeed, this formulation is reminiscent of the class of problems characterized by Systems of Systems (SoS) in which the attempt is made to integrate existing systems, originally built to perform specific functions, into a more comprehensive and multifunctional system. As discussed in [15], the components of each system can be viewed as communicating with each other within a common ontology, or model of the world that is tuned to the smooth functioning of the organization. However, such ontologies may well be mismatched to support integration at the systems level. Despite working on the results of a long history of pre-human evolution, the fact that consciousness seems to provide a unified and undifferentiated picture of mind, suggests that human evolution has to a large extent solved the SoS integration problem.

The Activity Paradigm for Automated M&S

At least for the initial part of its life, a modeling agent needs to work on a "first order" assumption about its environment, namely, that it can exploit only semantics-free properties [39]. Regularities, such as periodic behaviors, and stimulus-response associations, are one source of such semantics-free properties. In this section, we will focus on a fundamental property of systems, such as the brain's neural network, that have a large number of components. This distribution of activity of such a system over space and time provides a rich and semantics-free substrate from which models can be generated. Proposing structures and

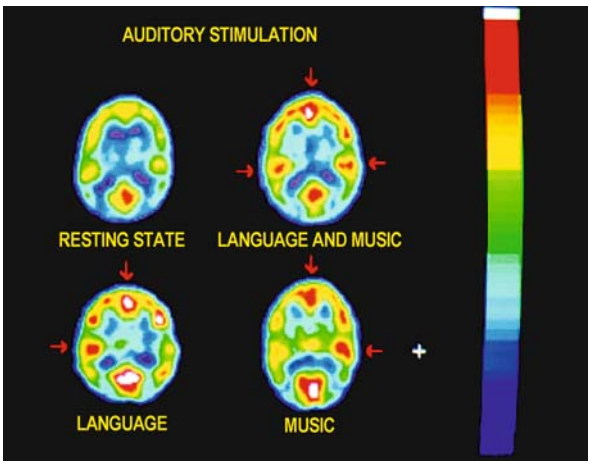

Artificial Intelligence in Modeling and Simulation, Figure 12
Brain module activities

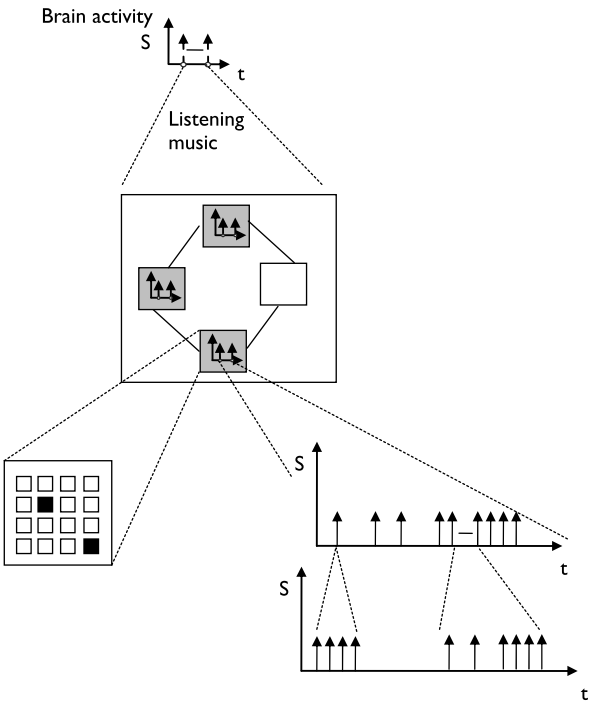

Artificial Intelligence in Modeling and Simulation, Figure 13
Hierarchy of components, activity and discrete-event abstractions

algorithms to track and replicate this activity should support automated modeling and simulation of patterns of reality. The goal of the activity paradigm is to extract mechanisms from natural phenomena and behaviors to automate and guide the M&S process.

The brain offers a quintessential illustration of activity and its potential use to construct models. Figure 12 describes brain electrical activities [54]. Positron emission tomography (PET) is used to record electrical activity inside the brain. The PET method scans show what happens inside the brain when resting and when stimulated by words and music. The red areas indicate high brain activities. Language and music produce responses in opposite sides of the brain (showing the sub-system specializations). There are many levels of activity (ranging from low to high.)

There is a strong link between modularity and the applicability of activity measurement as a useful concept. Indeed, modules represent loci for activity—a distribution of activity would not be discernable over a network were there no modules that could be observed to be in different states of activity. As just illustrated, neuroscientists are exploiting this activity paradigm to associate brain areas with functions and to gain insight into areas that are active or inactive over different, but related, functions, such as language and music processing. We can generalize this approach as a paradigm for an automated modeling agent.

Component, Activity and Discrete-Event Abstractions

Figure 13 depicts a brain description through components and activities. First, the modeler considers one brain activity (e. g., listening music.) This first level activity corresponds to simulation components at a lower level. At this level, activity of the components can be considered to be only on (grey boxes) or off (white boxes.) At lower levels, structure and behaviors can be decomposed. The structure of the higher component level is detailed at lower levels (e. g., to the neuronal network). The behavior is also detailed through activity and discrete-event abstraction [55]. At the finest levels, activity of components can be detailed.

Using the pattern detection and quantization methods continuously changing variables can also be treated within the activity paradigm. As illustrated in Fig. 14, small slopes and small peaks can signal low activity whereas high slopes and peaks can signal high activity levels. To provide for scale, discrete event abstraction can be achieved using quantization [56]. To determine the activity level, a *quantum* or measure of significant change has to be chosen. The quantum size acts as a filter on the continuous flow. For example, one can notice that in the figure, using the displayed quantum, the smallest peaks will not be significant. Thus, different levels of resolution can be achieved by employing different quantum sizes. A genetic algorithm can be used to find the optimum such level of resolution given for a given modeling objective [43].

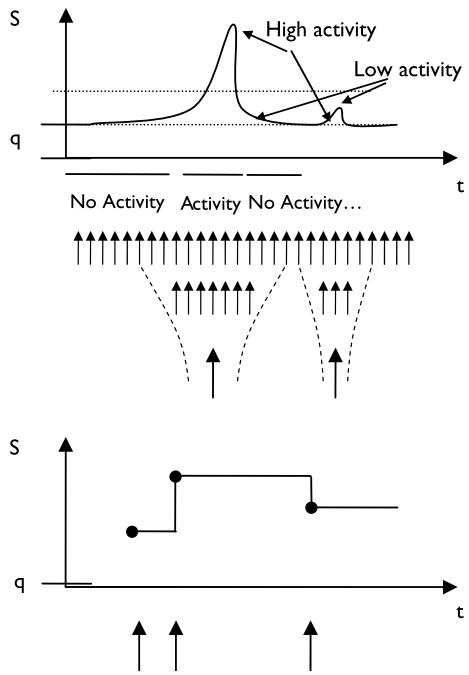

Artificial Intelligence in Modeling and Simulation, Figure 14
Activity sensitivity and discrete-events

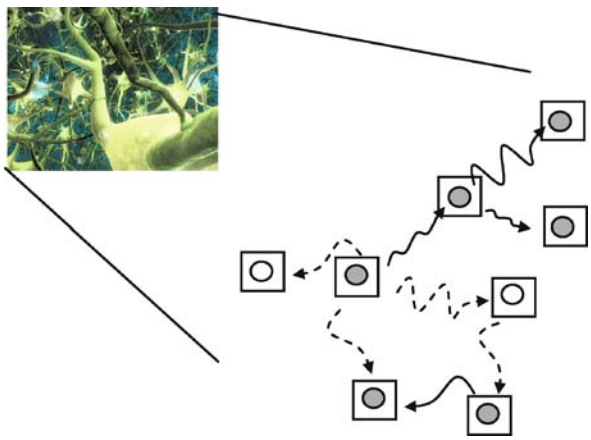

Artificial Intelligence in Modeling and Simulation, Figure 15
Activity paths in neurons

Activity Tracking

Within the activity paradigm, M&S consists of capturing activity paths through component processing and transformations. To determine the basic structure of the whole system, an automated modeler has to answer questions of the form: *where and how is activity produced, received, and transmitted*? Figure 15 represents a component-based view of activity flow in a neuronal network. Activity paths through components are represented by full arrows. Activity is represented by full circles. Components are represented by squares. The modeler must generate such a graph based on observed data—but how does it obtain such data? One approach is characteristic of current use of PET scans by neuroscientists. This approach exploits a relationship between activity and energy—activity requires consumption of energy, therefore, observing areas of high energy consumption signals areas of high activity. Notice that this correlation requires that energy consumption be localizable to modules in the same way that activity is so localized. So for example, current computer architectures that provide a single power source to all components do not lend themselves to such observation. How activity is passed on from component to component can be related to the modularity styles (none, weak, strong) of the components. Concepts relating such modularity styles to activity transfer need to be developed to support an activity tracking methodology that goes beyond reliance on energy correlation.

Activity Model Validation

Recall that having generated a model of an observed system (whether through activity tracking or by other means), the next step is validation. In order to perform such validation, the modeler needs an approach to generating activity profiles from simulation experiments on the model and to comparing these profiles with those observed in the real system. Muzy and Nutaro [57] have developed algorithms that exploit activity tracking to achieve efficient simulation of DEVS models. These algorithms can be adopted to provide an activity tracking pattern applicable to a given simulation model to extract its activity profiles for comparison with those of the modeled system.

A forthcoming monograph will develop the activity paradigm for M&S in greater detail [58].

Intelligent Agents in Simulation

Recent trends in technology as well as the use of simulation in exploring complex artificial and natural information processes [62,63] have made it clear that simulation model fidelity and complexity will continue to increase dramatically in the coming decades. The dynamic and distributed nature of simulation applications, the significance of exploratory analysis of complex phenomena [64], and the need for modeling the micro-level interactions, collaboration, and cooperation among real-world entities is bringing a shift in the way systems are being conceptualized. Using intelligent agents in simulation models is based on the idea that it is possible to represent the behavior of active entities in the world in terms of the interactions of

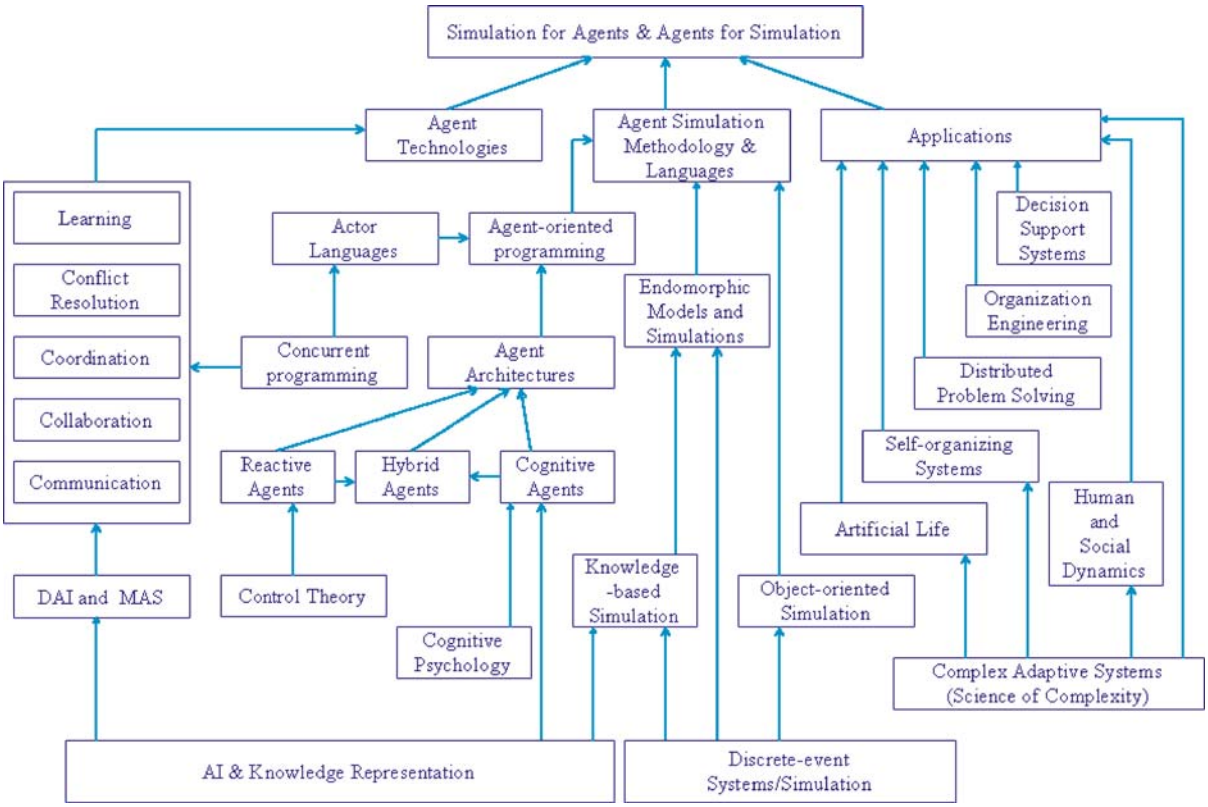

Artificial Intelligence in Modeling and Simulation, Figure 16
Evolution of the use of intelligent agents in simulation

an assembly of agents with their own operational autonomy.

The early pervading view on the use of agents in simulation stems from the developments in Distributed Artificial Intelligence (DAI), as well as advances in agent architectures and agent-oriented programming. The DAI perspective to modeling systems in terms of entities that are capable of solving problems by means of reasoning through symbol manipulation resulted in various technologies that constitute the basic elements of agent systems. The early work on design of agent simulators within the DAI community focused on answering the question of how goals and intentions of agents emerge and how they lead to execution of actions that change the state of their environment. The agent-directed approach to simulating agent systems lies at the intersection of several disciplines: DAI, Control Theory, Complex Adaptive Systems (CAS), and Discrete-event Systems/Simulation. As shown in Fig. 16, these core disciplines gave direction to technology, languages, and possible applications, which then influenced the evolution of the synergy between simulation and agent systems.

Distributed Artificial Intelligence and Simulation

While progress in agent simulators and interpreters resulted in various agent architectures and their computational engines, the ability to coordinate agent ensembles was recognized early as a key challenge [65]. The MACE system [66] is considered as one of the major milestones in DAI. Specifically, the proposed DAI system integrated concepts from concurrent programming (e. g., actor formalism [67]) and knowledge representation to symbolically reason about skills and beliefs pertaining to modeling the environment. Task allocation and coordination were considered as fundamental challenges in early DAI systems. The contract net protocol developed by [68] provided the basis for modeling collaboration in simulation of distributed problem solving.

Agent Simulation Architectures

One of the first agent-oriented simulation languages, AGENT-0 [69], provided a framework that enabled the representation of beliefs and intentions of agents. Unlike object-oriented simulation languages such as SIM-

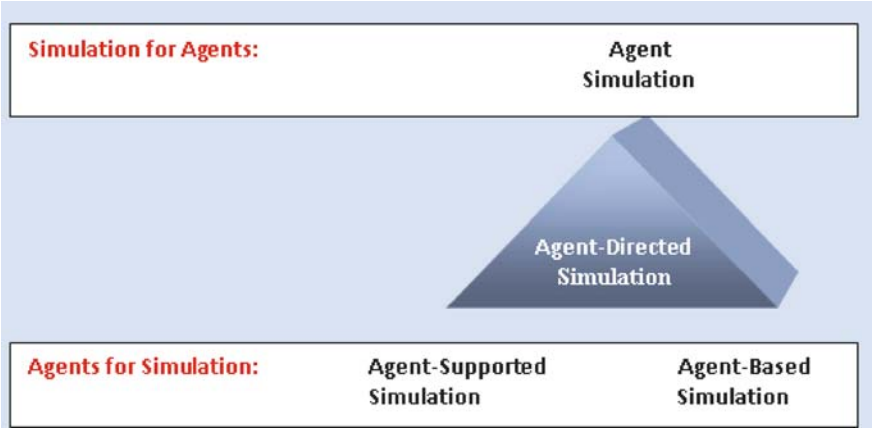

Artificial Intelligence in Modeling and Simulation, Figure 17
Agent-directed simulation framework

ULA 67 [70], the first object-oriented language for specifying discrete-event systems, AGENT-O and McCarthy's Elephant2000 language incorporated speech act theory to provide flexible communication mechanisms for agents. DAI and cognitive psychology influenced the development of cognitive agents such as those found in AGENT-0, e. g., the Belief-Desires-Intentions (BDI) framework [71]. However, procedural reasoning and control theory provided a basis for the design and implementation of reactive agents. Classical control theory enables the specification of a mathematical model that describes the interaction of a control system and its environment. The analogy between an agent and control system facilitated the formalization of agent interactions in terms of a formal specification of dynamic systems. The shortcomings of reactive agents (i. e., lack of mechanisms of goal-directed behavior) and cognitive agents (i. e., issues pertaining to computational tractability in deliberative reasoning) led to the development of hybrid architectures such as the RAP system [72].

Agents are often viewed as design metaphors in the development of models for simulation and gaming. Yet, this narrow view limits the potential of agents in improving various other dimensions of simulation. To this end, Fig. 17 presents a unified paradigm of Agent-Directed Simulation that consists of two categories as follows: (1) Simulation for Agents (agent simulation), i. e., simulation of systems that can be modeled by agents (in engineering, human and social dynamics, military applications etc.) and (2) Agents for Simulation that can be grouped under two groups: agent-supported simulation and agent-based simulation.

Agent Simulation

Agent simulation involves the use of agents as design metaphors in developing simulation models. Agent simulation involves the use of simulation conceptual frameworks (e. g., discrete-event, activity scanning) to simulate the behavioral dynamics of agent systems and incorporate autonomous agents that function in parallel to achieve their goals and objectives. Agents possess high-level interaction mechanisms independent of the problem being solved. Communication protocols and mechanisms for interaction via task allocation, coordination of actions, and conflict resolution at varying levels of sophistication are primary elements of agent simulations. Simulating agent systems requires understanding the basic principles, organizational mechanisms, and technologies underlying such systems.

Agent-Based Simulation

Agent-based simulation is the use of agent technology to monitor and generate model behavior. This is similar to the use of AI techniques for the generation of model behavior (e. g., qualitative simulation and knowledge-based simulation). Development of novel and advanced simulation methodologies such as multisimulation suggests the use of intelligent agents as simulator coordinators, where run-time decisions for model staging and updating takes place to facilitate dynamic composability. The perception feature of agents makes them pertinent for monitoring tasks. Also, agent-based simulation is useful for having complex experiments and deliberative knowledge processing such as planning, deciding, and reasoning. Agents are

also critical enablers to improve composability and interoperability of simulation models [73].

Agent-Supported Simulation

Agent-supported simulation deals with the use of agents as a support facility to enable computer assistance by enhancing cognitive capabilities in problem specification and solving. Hence, agent-supported simulation involves the use of intelligent agents to improve simulation and gaming infrastructures or environments. Agent-supported simulation is used for the following purposes:

- to provide computer assistance for front-end and/or back-end interface functions;
- to process elements of a simulation study symbolically (for example, for consistency checks and built-in reliability); and
- to provide cognitive abilities to the elements of a simulation study, such as learning or understanding abilities.

For instance, in simulations with defense applications, agents are often used as support facilities to

- see the battlefield,
- fuse, integrate and de-conflict the information presented by the decision-maker,
- generate alarms based on the recognition of specific patterns,
- Filter, sort, track, and prioritize the disseminated information, and
- generate contingency plans and courses of actions.

A significant requirement for the design and simulation of agent systems is the distributed knowledge that represents the mental model that characterizes each agent's beliefs about the environment, itself, and other agents. Endomorphic agent concepts provide a framework for addressing the difficult conceptual issues that arise in this domain.

Endomorphic Agents

We now consider an advanced feature that an autonomous, integrated and comprehensive modeler/simulationist agent must have if it is fully emulate human capability. This is the ability, to a limited, but significant extent, to construct and employ models of its own mind as well of the minds of other agents. We use the term "mind" in the sense just discussed.

The concept of an endomorphic agent is illustrated in Fig. 18 and 19 in a sequence of related diagrams. The diagram labeled with an oval with embedded number 1 is that of Fig. 9 with the modifications mentioned earlier to match up with human motivation and desire generation modules. In diagram 2, the label "mind" refers to the set of M&S capabilities depicted in Fig. 9. As in [30], an agent, human or technological, is considered to be composed of a mind and a body. Here, "body" represents the external manifestation of the agent, which is observable by other agents. Whereas, in contrast, mind is hidden from view and must be a construct, or model, of other agents. In other words, to use the language of evolutionary psychology, agents must develop a "theory of mind" about other agents from observation of their external behavior. An endomorphic agent is represented in diagram 3 with a mental model of the body and mind of the agent in diagram 2. This second agent is shown more explicitly in diagram 4, with a mental representation of the first agent's body and mind. Diagram 5 depicts the recursive aspect of endomorphism, where the (original) agent of diagram 2 has developed a model of the second agent's body and mind. But the latter model contains the just-mentioned model of the first agent's body and mind. This leads to a potentially infinite regress in which—apparently—each agent can have a representation of the other agent, and by reflection, of himself, that increases in depth of nesting without end. Hofstadter [59] represents a similar concept in the diagram on page 144 of his book, in which the comic character Sluggo is "dreaming of himself dreaming of himself dreaming of himself, without end." He then uses the label on the Morton Salt box on page 145 to show how not all self reference involves infinite recursion. On the label, the girl carrying the salt box obscures its label with her arm, thereby shutting down the regress. Thus, the salt box has a representation of itself on its label but this representation is only partial.

Reference [30] related the termination in self-reference to the agent's objectives and requirements in constructing models of himself, other agents, and the environment. Briefly, the agent need only to go as deep as needed to get a reliable model of the other agents. The agent can stop at level 1 with a representation of the other's bodies. However, this might not allow predicting another's movements, particularly if the latter has a mind in control of these movements. This would force the first agent to include at least a crude model of the other agent's mind. In a competitive situation, having such a model might give the first agent an advantage and this might lead the second agent to likewise develop a predictive model of the first agent. With the second agent now seeming to become less predictable, the first agent might develop a model of the second agent's mind that restores lost predictive power. This would likely have to include a reflected representation of himself, although the impending regression could be

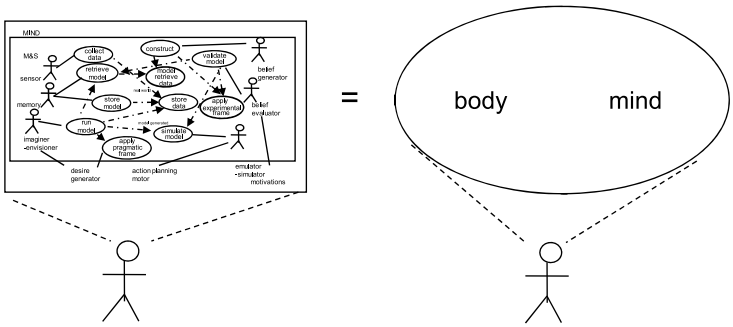

Artificial Intelligence in Modeling and Simulation, Figure 18
M&S within mind

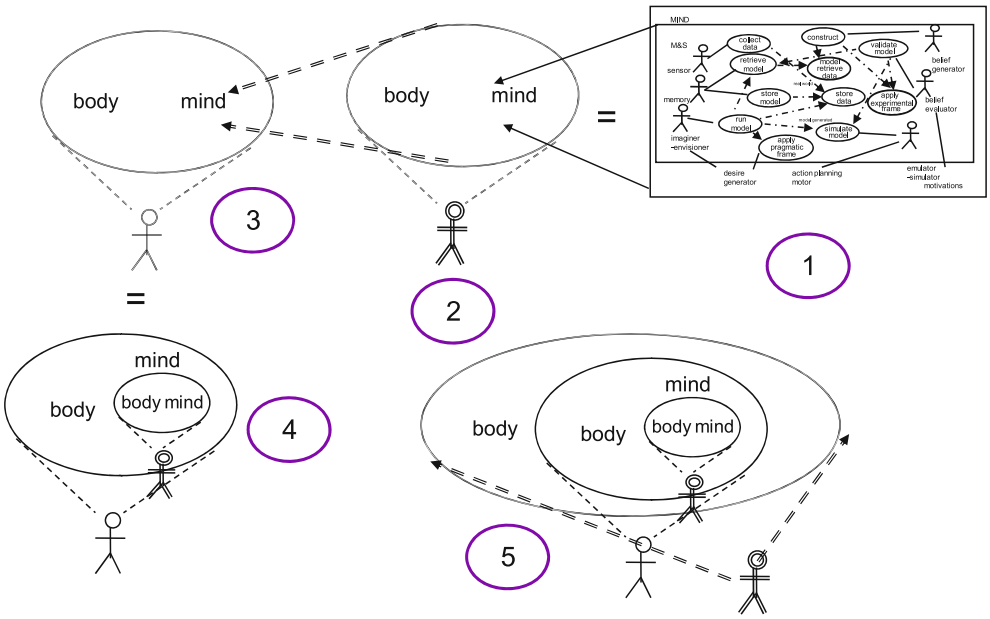

Artificial Intelligence in Modeling and Simulation, Figure 19
Emergence of endomorphic agents

halted if this representation did not, itself, contain a model of the other agent. Thus, the depth to which competitive endomorphic agents have models of themselves and others might be the product of a co-evolutionary "mental arms race" in which an improvement in one side triggers a contingent improvement in the other—the improvement being an incremental refinement of the internal models by successively adding more levels of nesting.

Minsky [60] conjectured that termination of the potentially infinite regress in agent's models of each other within a society of mind might be constrained by shear limitations on the ability to martial the resources required to support the necessary computation. We can go further by assuming that agents have differing mental capacities to support such computational nesting. Therefore, an agent with greater capacity might be able to "out think" one of lesser capability. This is illustrated by the following real-life story drawn from a recent newspaper account of a critical play in a baseball game.

Interacting Models of Others in Competitive Sport

The following account is illustrated in Fig. 20.

A Ninth Inning to Forget
Cordero Can't Close, Then Base-Running Gaffe Ends Nats' Rally

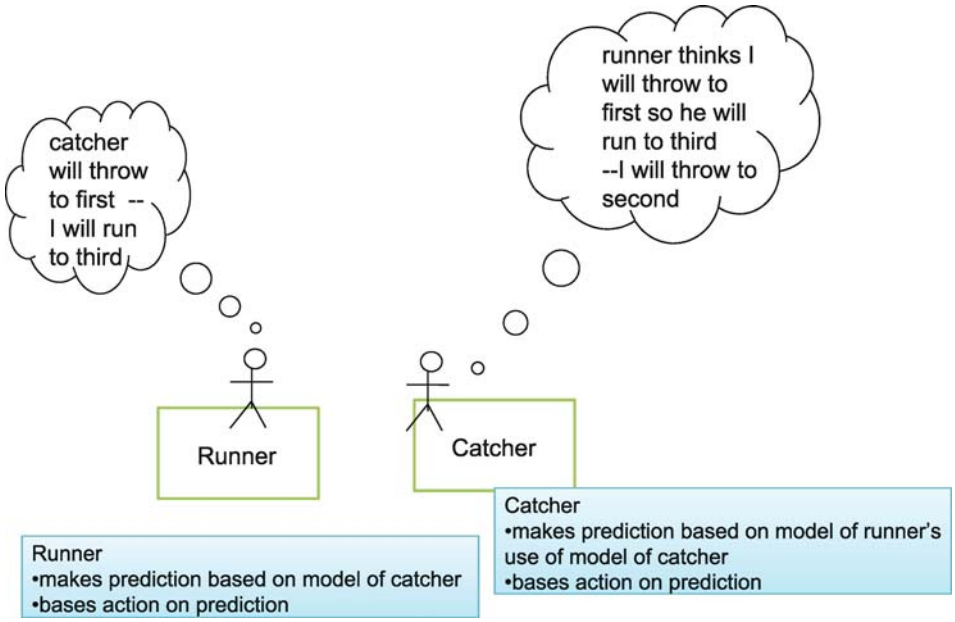

Artificial Intelligence in Modeling and Simulation, Figure 20
Interacting models of others in baseball

Steve Yanda – Washington Post Staff Writer
Jun 24, 2007

Copyright The Washington Post Company Jun 24, 2007Indians 4, Nationals 3

Nook Logan played out the ending of last night's game in his head as he stood on second base in the bottom of the ninth inning. The bases were loaded with one out and the Washington Nationals trailed by one run. Even if Felipe Lopez, the batter at the plate, grounded the ball, say, right back to Cleveland Indians closer Joe Borowski, the pitcher merely would throw home. Awaiting the toss would be catcher Kelly Shoppach, who would tag the plate and attempt to nail Lopez at first. By the time Shoppach's throw reached first baseman Victor Martinez, Logan figured he would be gliding across the plate with the tying run. Lopez did ground to Borowski, and the closer did fire the ball home. However, Shoppach elected to throw to third instead of first, catching Logan drifting too far off the bag for the final out in the Nationals' 4–3 loss at RFK Stadium. "I thought [Shoppach] was going to throw to first," Logan said. And if the catcher had, would Logan have scored all the way from second? "Easy."

We'll analyze this account to show how it throws light on the advantage rendered by having an endomorphic capability to process to a nesting depth exceeding that of an opponent.

The situation starts the bottom of the ninth inning with the Washington Nationals at bats having the bases loaded with one out and trailing by one run. The runner on second base, Nook Logan plays out the ending of the game in his head. This can be interpreted in terms of endomorphic models as follows. Logan makes a prediction using his models of the opposing pitcher and catcher, namely that the pitcher would throw home and the catcher would tag the plate and attempt to nail the batter at first. Logan then makes a prediction using a model of himself, namely, that he would be able to reach home plate while the pitcher's thrown ball was traveling to first base.

In actual play, the catcher threw the ball to third and caught Logan out. This is evidence that the catcher was able to play out the simulation to a greater depth then was Logan. The catcher's model of the situation agreed with that of Logan as it related to the other actors. The difference was that the catcher used a model of Logan that predicted that the latter (Logan) would predict the he (the catcher) would throw to first. Having this prediction, the catcher decided instead to throw the ball to the second baseman which resulted in putting Logan out. We note that the catcher's model was based on his model of Logan's model so it was at one level greater in depth of nesting than the latter. To have succeeded, Logan would have had to be

able to support one more level, namely, to have a model of the catcher that would predict that the catcher would use the model of himself (Logan) to out-think him and then make the counter move, not to start towards third base.

The enigma of such endomorphic agents provides extreme challenges to further research in AI and M&S. The formal and computational framework that the M&S framework discussed here provides may be of particular advantage to cognitive psychologists and philosophers interested in an active area of investigation in which the terms "theory of mind," "simulation," and "mind reading" are employed without much in the way of definition [61].

Future Directions

M&S presents some fundamental and very difficult problems whose solution may benefit from the concepts and techniques of AI. We have discussed some key problem areas including verification and validation, reuse and composability, and distributed simulation and systems of systems interoperability. We have also considered some areas of AI that have direct applicability to problems in M&S, such as Service-Oriented Architecture and Semantic Web, ontologies, constrained natural language, and genetic algorithms. In order to provide a unifying theme for the problem and solutions, we raised the question of whether all of M&S can be automated into an integrated autonomous artificial modeler/simulationist. We explored an approach to developing such an intelligent agent based on the System Entity Structure/Model Base framework, a hybrid methodology that combines elements of AI and M&S. We proposed a concrete methodology by which such an agent could engage in M&S based on activity tracking. There are numerous challenges to AI that implementing such a methodology in automated form presents. We closed with consideration of endomorphic modeling capability, an advanced feature that such an agent must have if it is to fully emulate human M&S capability. Since this capacity implies an infinite regress in which models contain models of themselves without end, it can only be had to a limited degree. However, it may offer critical insights into competitive co-evolutionary human or higher order primate behavior to launch more intensive research into model nesting depth. This is the degree to which an endomorphic agent can marshal mental resources needed to construct and employ models of its own "mind" as well of the 'minds" of other agents. The enigma of such endomorphic agents provides extreme challenges to further research in AI and M&S, as well as related disciplines such as cognitive science and philosophy.

Bibliography

Primary Literature

1. Wymore AW (1993) Model-based systems engineering: An introduction to the mathematical theory of discrete systems and to the tricotyledon theory of system design. CRC, Boca Raton
2. Zeigler BP, Kim TG, Praehofer H (2000) Theory of modeling and simulation. Academic Press, New York
3. Ören TI, Zeigler BP (1979) Concepts for advanced simulation methodologies. Simulation 32(3):69–82
4. http://en.wikipedia.org/wiki/DEVS Accessed Aug 2008
5. Knepell PL, Aragno DC (1993) Simulation validation: a confidence assessment methodology. IEEE Computer Society Press, Los Alamitos
6. Law AM, Kelton WD (1999) Simulation modeling and analysis, 3rd edn. McGraw-Hill, Columbus
7. Sargent RG (1994) Verification and validation of simulation models. In: Winter simulation conference. pp 77–84
8. Balci O (1998) Verification, validation, and testing. In: Winter simulation conference.
9. Davis KP, Anderson AR (2003) Improving the composability of department of defense models and simulations, RAND technical report. http://www.rand.org/pubs/monographs/MG101/. Accessed Nov 2007; J Def Model Simul Appl Methodol Technol 1(1):5–17
10. Ylmaz L, Oren TI (2004) A conceptual model for reusable simulations within a model-simulator-context framework. Conference on conceptual modeling and simulation. Conceptual Models Conference, Italy, 28–31 October, pp 28–31
11. Traore M, Muxy A (2004) Capturing the dual relationship between simulation models and their context. Simulation practice and theory. Elsevier
12. Page E, Opper J (1999) Observations on the complexity of composable simulation. In: Proceedings of winter simulation conference, Orlando, pp 553–560
13. Kasputis S, Ng H (2000) Composable simulations. In: Proceedings of winter simulation conference, Orlando, pp 1577–1584
14. Sarjoughain HS (2006) Model composability. In: Perrone LF, Wieland FP, Liu J, Lawson BG, Nicol DM, Fujimoto RM (eds) Proceedings of the winter simulation conference, pp 104–158
15. DiMario MJ (2006) System of systems interoperability types and characteristics in joint command and control. In: Proceedings of the 2006 IEEE/SMC international conference on system of systems engineering, Los Angeles, April 2006
16. Sage AP, Cuppan CD (2001) On the systems engineering and management of systems of systems and federation of systems. Information knowledge systems management, vol 2, pp 325–345
17. Dahmann JS, Kuhl F, Weatherly R (1998) Standards for simulation: as simple as possible but not simpler the high level architecture for simulation. Simulation 71(6):378
18. Sarjoughian HS, Zeigler BP (2000) DEVS and HLA: Complimentary paradigms for M&S? Trans SCS 4(17):187–197
19. Yilmaz L (2004) On the need for contextualized introspective simulation models to improve reuse and composability of defense simulations. J Def Model Simul 1(3):135–145
20. Tolk A, Muguira JA (2003) The levels of conceptual interoperability model (LCIM). In: Proceedings fall simulation interoperability workshop, http://www.sisostds.org Accessed Aug 2008
21. http://en.wikipedia.org/wiki/Simula Accessed Aug 2008

22. http://en.wikipedia.org/wiki/Java Accessed Aug 2008
23. http://en.wikipedia.org/wiki/Expert_system Accessed Aug 2008
24. http://en.wikipedia.org/wiki/Frame_language Accessed Aug 2008
25. http://en.wikipedia.org/wiki/Agent_based_model Accessed Aug 2008
26. http://www.swarm.org/wiki/Main_Page Accessed Aug 2008
27. Unified Modeling Language (UML) http://www.omg.org/technology/documents/formal/uml.htm
28. Object Modeling Group (OMG) http://www.omg.org
29. http://en.wikipedia.org/wiki/Semantic_web Accessed Aug 2008
30. Zeigler BP (1990) Object Oriented Simulation with Hierarchical, Modular Models: Intelligent Agents and Endomorphic Systems. Academic Press, Orlando
31. http://en.wikipedia.org/wiki/Service_oriented_architecture Accessed Aug 2008
32. Mittal S, Mak E, Nutaro JJ (2006) DEVS-Based dynamic modeling & simulation reconfiguration using enhanced DoDAF design process. Special issue on DoDAF. J Def Model Simul, Dec (3)4:239–267
33. Ziegler BP (1988) Simulation methodology/model manipulation. In: Encyclopedia of systems and controls. Pergamon Press, England
34. Alexiev V, Breu M, de Bruijn J, Fensel D, Lara R, Lausen H (2005) Information integration with ontologies. Wiley, New York
35. Kim L (2003) Official XMLSPY handbook. Wiley, Indianapolis
36. Zeigler BP, Hammonds P (2007) Modeling & simulation-based data engineering: introducing pragmatics into ontologies for net-centric information exchange. Academic Press, New York
37. Simard RJ, Zeigler BP, Couretas JN (1994) Verb phrase model specification via system entity structures. AI and Planning in high autonomy systems, 1994. Distributed interactive simulation environments. Proceedings of the Fifth Annual Conference, 7–9 Dec 1994, pp 192–1989
38. Checkland P (1999) Soft systems methodology in action. Wiley, London
39. Holland JH (1992) Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence. MIT Press, Cambridge
40. Goldberg DE (1989) Genetic algorithms in search, optimization and machine learning. Addison-Wesley Professional, Princeton
41. Davis L (1987) Genetic algorithms and simulated annealing. Morgan Kaufmann, San Francisco
42. Zbigniew M (1996) Genetic algorithms + data structures = evolution programs. Springer, Heidelberg
43. Cheon S (2007) Experimental frame structuring for automated model construction: application to simulated weather generation. Doct Diss, Dept of ECE, University of Arizona, Tucson
44. Zeigler BP (1984) Multifaceted modelling and discrete event simulation. Academic Press, London
45. Rozenblit JW, Hu J, Zeigler BP, Kim TG (1990) Knowledge-based design and simulation environment (KBDSE): foundational concepts and implementation. J Oper Res Soc 41(6):475–489
46. Kim TG, Lee C, Zeigler BP, Christensen ER (1990) System entity structuring and model base management. IEEE Trans Syst Man Cyber 20(5):1013–1024
47. Zeigler BP, Zhang G (1989) The system entity structure: knowledge representation for simulation modeling and design. In: Widman LA, Loparo KA, Nielsen N (eds) Artificial intelligence, simulation and modeling. Wiley, New York, pp 47–73
48. Luh C, Zeigler BP (1991) Model base management for multifaceted systems. ACM Trans Model Comp Sim 1(3):195–218
49. Couretas J (1998) System entity structure alternatives enumeration environment (SEAS). Doctoral Dissertation Dept of ECE, University of Arizona
50. Hyu C Park, Tag G Kim (1998) A relational algebraic framework for VHDL models management. Trans SCS 15(2):43–55
51. Chi SD, Lee J, Kim Y (1997) Using the SES/MB framework to analyze traffic flow. Trans SCS 14(4):211–221
52. Cho TH, Zeigler BP, Rozenblit JW (1996) A knowledge based simulation environment for hierarchical flexible manufacturing. IEEE Trans Syst Man Cyber- Part A: Syst Hum 26(1):81–91
53. Carruthers P (2006) Massively modular mind architecture the architecture of the mind. Oxford University Press, USA, pp 480
54. Wolpert L (2004) Six impossible things before breakfast: The evolutionary origin of belief, W.W. Norton London
55. Zeigler BP (2005) Discrete event abstraction: an emerging paradigm for modeling complex adaptive systems perspectives on adaptation, In: Booker L (ed) Natural and artificial systems, essays in honor of John Holland. Oxford University Press, Oxford
56. Nutaro J, Zeigler BP (2007) On the stability and performance of discrete event methods for simulating continuous systems. J Comput Phys 227(1):797–819
57. Muzy A, Nutaro JJ (2005) Algorithms for efficient implementation of the DEVS & DSDEVS abstract simulators. In: 1st Open International Conference on Modeling and Simulation (OICMS). Clermont-Ferrand, France, pp 273–279
58. Muzy A The activity paradigm for modeling and simulation of complex systems. (in process)
59. Hofstadter D (2007) I am a strange loop. Basic Books
60. Minsky M (1988) Society of mind. Simon & Schuster, Goldman
61. Alvin I (2006) Goldman simulating minds: the philosophy, psychology, and neuroscience of mindreading. Oxford University Press, USA
62. Denning PJ (2007) Computing is a natural science. Commun ACM 50(7):13–18
63. Luck M, McBurney P, Preist C (2003) Agent technology: enabling next generation computing a roadmap for agent based computing. Agentlink, Liverpool
64. Miller JH, Page SE (2007) Complex adaptive systems: an introduction to computational models of social life. Princeton University Press, Princeton
65. Ferber J (1999) Multi-Agent systems: an introduction to distributed artificial intelligence. Addison-Wesley, Princeton
66. Gasser L, Braganza C, Herman N (1987) Mace: a extensible testbed for distributed AI research. Distributed artificial intelligence – research notes in artificial intelligence, pp 119–152
67. Agha G, Hewitt C (1985) Concurrent programming using actors: exploiting large-scale parallelism. In: Proceedings of the foundations of software technology and theoretical computer science, Fifth Conference, pp 19–41
68. Smith RG (1980) The contract net protocol: high-level communication and control in a distributed problem solver. IEEE Trans Comput 29(12):1104–1113
69. Shoham Y (1993) Agent-oriented programming. Artif Intell 60(1):51–92
70. Dahl OJ, Nygaard K (1967) SIMULA67 Common base definiton. Norweigan Computing Center, Norway

71. Rao AS, George MP (1995) BDI-agents: from theory to practice. In: Proceedings of the first intl. conference on multiagent systems, San Francisco
72. Firby RJ (1992) Building symbolic primitives with continuous control routines. In: Procedings of the First Int Conf on AI Planning Systems. College Park, MD pp 62–29
73. Yilmaz L, Pasputelli S (2005) Toward a meta-level framework for agent-supported interoperation of defense simulations. J Def Model Simul 2(3):161–175

Books and Reviews

Alexiev V, Breu M, de Bruijn J, Fensel D, Lara R, Lausen H (2005) Information integration with ontologies. Wiley, New York
Alvin I (2006) Goldman Simulating Minds: The philosophy, psychology, and neuroscience of mindreading. Oxford University Press, USA
Carruthers P (2006) Massively modular mind architecture The architecture of the mind. http://www.amazon.com/exec/obidos/search-handle-url/102-1253221-6360121
John H (1992) Holland adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. The MIT Press Cambridge
Zeigler BP (1990) Object oriented simulation with hierarchical, modular models: intelligent agents and endomorphic systems. Academic Press, Orlando
Zeigler BP, Hammonds P (2007) Modeling & simulation-based data engineering: introducing pragmatics into ontologies for net-centric information exchange. Academic Press, New York
Zeigler BP, Kim TG, Praehofer H (2000) Theory of modeling and simulation. Academic Press, New York

Astronomical Time Series, Complexity in

Jeffrey D. Scargle
Space Science Division, NASA Ames Research Center, Moffett Field, USA

Article Outline

Glossary
Prolog: The Story of Goldilocks and the Three Time Series
Definition of the Subject
Introduction
Complex Time Series in Astronomy
Chaotic Dynamics in Celestial Mechanics
Chaotic Time Series Analysis in Astronomy
Future Directions
Bibliography

Glossary

Dynamical system A mathematical or physical system of which the evolution forward in time is determined by certain laws of motion – deterministic, random, or a combination.

Deterministic system A dynamical system the future of which, except for noise, can be exactly predicted based on past data.

Chaotic system A dynamical system the future of which can be predicted for short times, but due to the exponential growth of noise ("sensitive dependence on initial conditions") long term prediction is not possible.

Random system A dynamical system the future of which cannot be predicted, based on past data.

Linear system Consider a dynamical system described in part by two variables (representing forces, displacements, or the like), one of which is viewed as *input* and the other as *output*. The part of the system so described is said to be *linear* if the output is a linear function of the input. Examples: (a) a spring is linear in a small force applied to the spring, if the displacement is small; (b) the Fourier transform (output) of a time series (input) is linear. Example (b) is connected with the myth that a linear transform is not useful for "linear data." (To describe given time series data as linear is an abuse of the term. The concept of linearity cannot apply to data, but rather to a model or model class describing the data. In general given data can be equally well described by linear and nonlinear models.)

Phase portrait A picture of the evolution of a dynamical system, in a geometrical space related to the phase space of the system. This picture consists of a sampling of points tracing out the system trajectory over time.

Stationary system If the statistical descriptors of a dynamical system do not change with time, the system is said to be stationary. In practice, with finite data streams, the mathematical definitions of the various kinds of stationarity cannot be implemented, and one is forced to consider local or approximate stationarity.

Prolog: The Story of Goldilocks and the Three Time Series

Once upon a time there was a little girl named Goldilocks. She went for a walk in the forest. Pretty soon she came upon an observatory. She knocked and, when no one answered, she walked right in.

On the server in the computer room, there were three data sets. Goldilocks was hungry to find magical, simple dynamics underlying an apparently complicated time series. She tested the time series from the first data set.

"This time series is too noisy!" she exclaimed.

So, she tested the time series from the second data set.

"This time series is too short," she said

So, she analyzed the last time series.

"Ahhh, this time series is just right," she said happily and she analyzed it. "Wow, look, I found a 'chaotic attractor' and a small, non-integer embedding dimension, so this must be a chaotic astronomical system!" Goldilocks began to speculate about the underlying physics, but there were many detailed models possible. Goldilocks fell asleep. When she awoke later, she realized that this had all been a fairy-tale dream.

THE END

Definition of the Subject

The subject of this article is *time series analysis as a tool for analyzing complex astrophysical systems*. A time series is a set of sequential measurements where information about the measurement times is contained in the data record. It is implicit that the goal is to study the variation of the measured quantity, ruling out measurements repeated simply to more accurately determine a value, say be averaging.

A number of developments are combining to markedly increase the importance of this topic in modern astrophysics. The first is the growing realization that many astronomical objects are dynamically changing in complicated ways, yielding intensity changes that have an unpredictable, stochastic nature. In short, many of the objects previously thought to be evolving only slowly and serenely, and on very long time scales, are in actuality complex systems – if not chaotic in a rigorous sense, at least behaving stochastically. This development has obviously been fueled by the recent rapid progress in complex systems theory. Less obvious is the role of improved observational techniques. It is now much easier to distinguish randomness in time series data due to observational errors from intrinsic – regular or stochastic – variability in the source itself. And in this Age of Digital Astronomy analysis of large sets of high-quality time series is more and more crucial to progress.

Introduction

Let's begin with a brief description of *dynamical systems*, in both the mathematical and physical sense. The former is an algorithm put forth as a model of the latter. *Dynamics* refers to the fact that the algorithm advances an initial state in time, modeling physical forces that drive the temporal behavior of an astrophysical system.

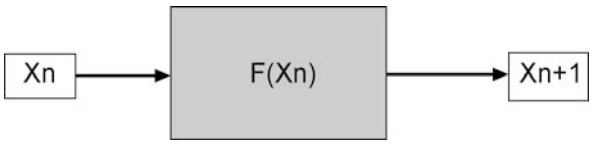

Astronomical Time Series, Complexity in, Figure 1
Black box representation of a discrete, one-dimensional dynamical system

Now consider some simple dynamical systems that exhibit complex behavior of the sort that has been called *chaos*. Figure 1 depicts the essence of a simple, one dimensional dynamical system, namely a procedure for generating the value of the variable X at the next future time $n+1$ (the output), given the value of X at the current time n (the input). The procedure starts at time zero, with initial condition X_0 as input, and is repeated indefinitely, generating the time series $\{X_0, X_1, X_2, X_3, \ldots\}$. Algebraically, the above picture is represented as simply

$$X_{n+1} = f(X_n)\,, \tag{1}$$

where f is the function representing the dynamics of the process. The following figure plots an example of such a dynamical system for a specific function f. The values of X look rather random, although the behavior for small values of X are a hint that some non-randomness is lurking here.

A commonly used test of whiteness is to compute the autocorrelation function and/or the power spectrum of X. Here are the results for the same process as above:

Note that the (symmetric) auto-correlation function is essentially a delta function, significantly different from zero only at zero lag. The power spectrum, of course, is irregular, but it is consistent with being constant with frequency. So our process appears to be white noise. Indeed, the autocorrelation and power spectrum of this process are a delta function and constant, respectively. The process X is indeed white noise!

But you have just been led down a garden path, because the interpretation of this as being a random process is wrong. Equation (1) precisely defines deterministic dynamics. The next future value is a definite, single-valued function of the past value. So our white process can hardly be truly random noise.

It is time to reveal X, as well as the fact that there is a stronger form of randomness than described by the term *uncorrelated*. The above figures all refer to the famous *lo-*

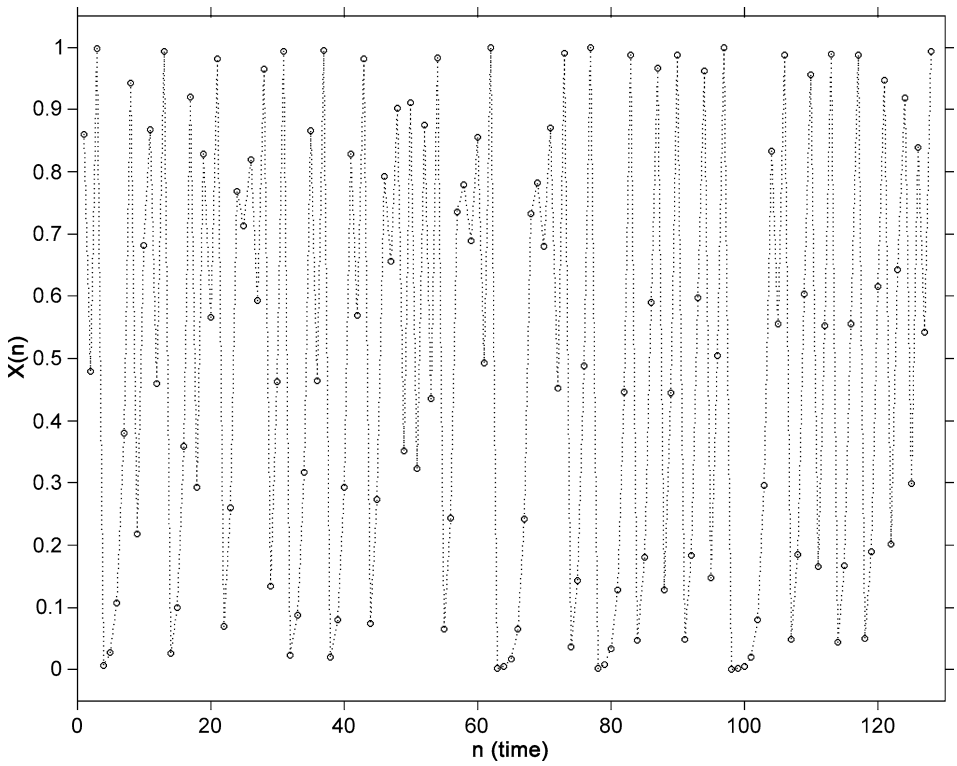

Astronomical Time Series, Complexity in, Figure 2
Time series generated by a simple nonlinear dynamical system. There are $N = 128$ time samples plotted

gistic map, namely

$$X_{n+1} = f(X_n) = \lambda X_n(1 - X_n)\,; \tag{2}$$

in what follows $\lambda = 4$, a special value that makes X maximally chaotic. The interested reader should consult one of the many books discussing the extraordinarily complex behavior of this nonlinear system as a function of λ, including proof that it is a "white" process, i. e. has a constant power spectrum and a delta-function autocorrelation, for $\lambda = 4$.

The point here is that the simple first-order diagnostics – autocorrelations and power spectra – as useful as they are, are not helpful at identifying the nature of our process. Let's turn to basics. The essence of a random process is in its joint probability distributions, not in its correlations. The joint distribution of X_n and X_m is defined as the probability that the two variables take arbitrary values, thusly:

$$P(a, b) = \text{Prob}(X_n = a \;\text{ and }\; X_m = b)\,, \tag{3}$$

where the notation here emphasizes that the arguments of P are dummy variables. The more common, if slightly misleading, notation is $P(X_n, X_m)$. One can estimate this quantity with 2D histograms – i. e., simply count pairs of values in 2D bins. Figure 4 is such an estimate of $P(X_n, X_{n+1})$. If the process X were truly random, all of the variables X_n would be statistically independent of each other. Independence is defined by the relation

$$\begin{aligned}\text{Prob}(X_n = a \;\text{ and }\; X_m = b)\\ = \text{Prob}(X_n = a)\;\text{Prob}(X_m = b)\,.\end{aligned} \tag{4}$$

In words, two variables are independent if their joint probability distribution is the product of their individual distributions. It should be clear from the figure that the logistic process is very far from independent; on the contrary, the value of X_{n+1} is determined completely by the value of X_n. Unfortunately, the concepts of uncorrelatedness and independence are frequently confused. In addition, scientists often derive results using the correlation function when they could derive much more powerful results using properties of joint probability distributions.

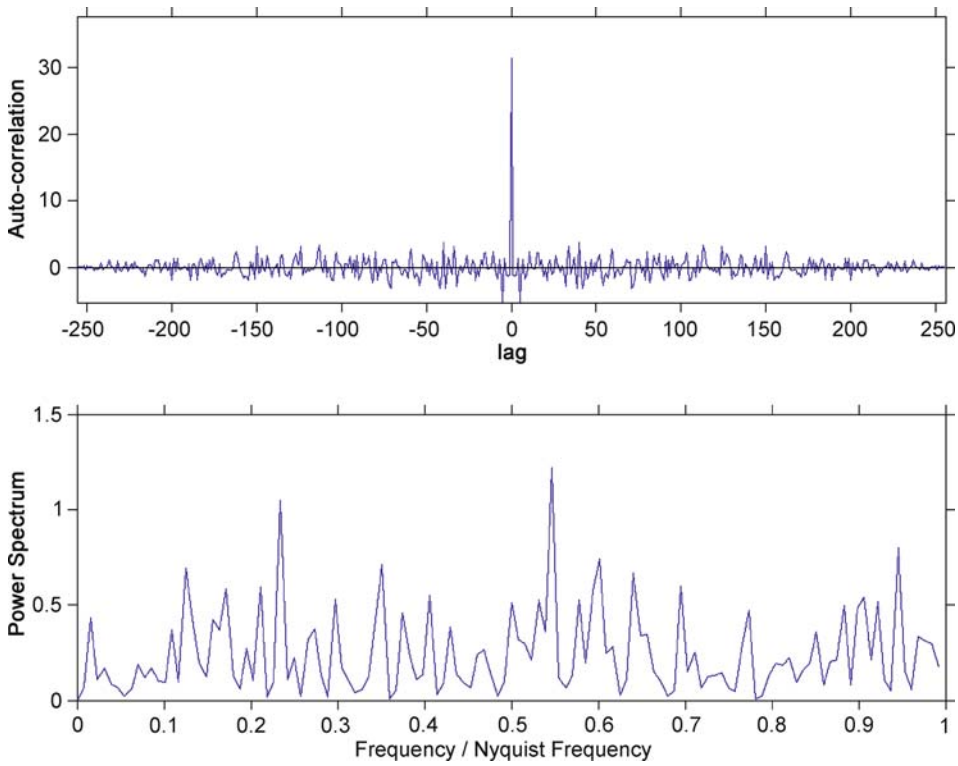

Astronomical Time Series, Complexity in, Figure 3
Autocorrelation and power spectrum from the same process as in Fig. 2 (but with twice as many time points, $N = 256$)

Complex Time Series in Astronomy

The first time series data in recorded history were probably astronomical in origin. It was important for the ancients to keep track of the motions of celestial objects, if only for predicting the arrival of each year's planting season. This agricultural service involved noting the passage of time and the location of the sun relative to other objects in the sky. Well-known examples of this connection between astronomical observations over time and practical aspects of agriculture punctuate the history of the ancient Egyptian the Mayan culture. There are presumably many other examples that have not survived the ravages of time – or the ravages of "all the wrong explorers and discoverers – thieves planting flags, murderers carrying crosses whom we have been raised to honor" (Peter S. Beagle, preface to J. R. R. Tolkien's "The Hobbit").

Perhaps the most famous historical incident of relevance is Gauss's successful determination in 1801 of the orbit of Ceres from a mere 41 days of data, over which the asteroid moved only 9 degrees on the sky. Gauss had invented the method of least squares some years earlier, and it was relatively easy (only 100 hours of computation time!) to apply this method to the problem of determining an orbit of an object moving around the sun, from a time series of its positions on the sky.

> "… for it is now clearly shown that the orbit of a heavenly body may be determined quite nearly from good observations embracing only a few days; and this without any hypothetical assumption."
>
> Carl Friedrich Gauss

From the dawn of recorded astronomical observations, as perhaps implemented in prehistoric sites such as Stonehenge, to the Space Age, astronomical time series have been central to scientific exploration. The underlying goal is to be able to understand the dynamical forces driving the underlying physical processes and evidenced by variability in the observed time series. Starting with Poincaré in the later part of the 19th century, whose work was largely in the context of planetary motions, astronomers faced the peculiar possibility that simple, deterministic astrophysical systems could have complex, apparently random behavior. For example, celestial mechanicians realized that some of the planetary motions they had been dealing with

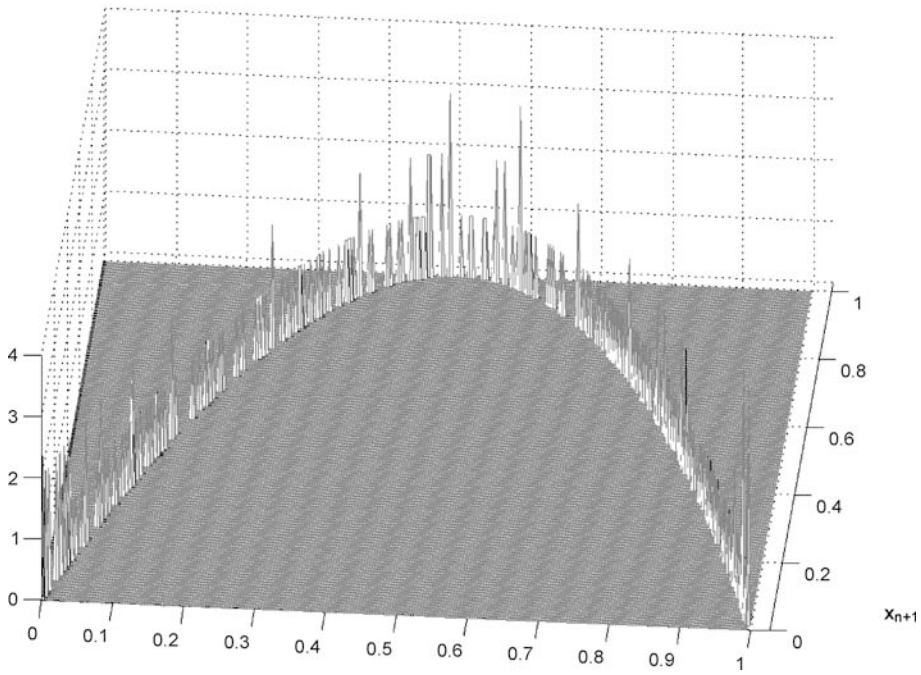

Astronomical Time Series, Complexity in, Figure 4
Histogram estimate of the joint probability distribution of X_n and X_{n+1} using data from the logistic process described in the previous figures

over the ages are in reality examples of chaos – the evolution of physical systems in time in a deterministic yet disordered way, often summarized as "simple systems with complex behavior."

How can a simple and perfectly deterministic physical system produce complex and unpredictable behavior? The answer lies in a kind of instability – sensitivity to initial conditions. A nonlinear, chaotic system has this property in the sense that evolutionary paths starting from two arbitrarily close initial conditions separate exponentially with time. This property leads to an amplification of the inevitable initial uncertainties that forecloses prediction of the future. In the hope that complex, seemingly random behaviors observed in some systems could have simple explanations, there was in previous decades a spasm of activity in astrophysics, primarily in two complementary areas: (a) theoretical computations of the nonlinear dynamical models of astrophysical systems, and (b) analysis of time series and other data from astrophysical systems, in an effort to uncover the simplicity underlying the apparent complexity. The first of these is treated in the article Chaos and Complexity in Astrophysics, by Oded Regev, elsewhere in this Encyclopedia and rather briefly in the next section, while the second – being the current focus – will be treated in more detail in the rest of this article.

Chaotic Dynamics in Celestial Mechanics

The period around 1980–1990 saw intensive development in dynamical systems theory, particularly in celestial mechanics – i. e. the theoretical study of the dynamics and kinematics of astronomical bodies, in our solar system, in our galaxy, and elsewhere in the Universe. Lecar, Franklin, Holman and Murray [40] review the topic of chaotic dynamics in the solar system.

As in the history of earlier developments in dynamics (Galileo, Newton, Einstein, to name a few of the great minds illuminating this grand passage in human knowledge), astronomy played an important role in the origin of complex systems theory. Jack Lissauer [41] provides a comprehensive review of nonlinear dynamics in the solar system, covering the extensive work of Wisdom, Laskar, Duncan, and many other workers in this arena. The thread here is that in many important situations the equations of motion defining the dynamical system are known exactly, and the driving forces are known with great precision, so that highly accurate numerical computations of the dynamical evolution of the systems are achievable. These computations, and accompanying analytic studies, have detailed a number of key examples of sensitivity to initial conditions and quasi-random wandering of state space

that are the hallmarks of deterministic chaos. One of the earliest of these success stories was the seminal Ph. D. thesis of Jack Wisdom [76], which elucidated the structure of gaps known in the spatial distribution of asteroids by showing that the missing regions correspond to chaotic orbital behavior.

For the most part, these computations relate to the very long timescales of many astronomical phenomena. One exception – the astronomical system showing chaotic behavior on the shortest time scale – is the fascinating story of the rotation of Hyperion, an irregularly shaped satellite of the planet Saturn. This satellite is unusual in several ways: it shape is very irregular, as can be seen in the dramatic false-color image in Fig. 5, its orbit is eccentric (eccentricity $e = 0.1$), and it is relatively close to its mother planet (the semi-major axis of its orbit is less than 25 times the radius of Saturn). These features of Hyperion and its orbit conspire to produce strong and regularly varying tidal forces on the satellite. Specifically, the differential equation describing the time evolution of the angle θ between the long axis of the satellite and a reference direction is

$$\mathrm{d}^2\theta/\mathrm{d}t^2 + (\omega_0^2/2r^3)\sin 2(\theta - f) = 0\,, \tag{5}$$

where f is an angle (the true anomaly) specifying the position of the satellite in its orbit, r is the radial distance of the satellite from the planet, and ω_0 is a function of the principle moments of inertia of the satellite [42]. This nonlinear equation has very interesting properties, and several authors (e. g. Wisdom, Peale and Mignard [77]) have shown that this system undergoes chaotic motion. The surface of section, or state space comprised by θ and $\mathrm{d}\theta/\mathrm{d}t$, consists of large regular regions, within which are chaotic islands roughly associated with the spin orbit resonances 1:2, 1:1, 3:2, 2:1, 5:2, etc. This configuration has marked similarities with the three-body system, although it is of course much simpler.

Detailed numerical computations of the Hyperion-Saturn system predicted chaotic rotation of the satellite. That is to say, the satellite's rotation axis should tumble chaotically and not remain even roughly pointed in a fixed direction. In a real sense, the magnitude and direction of rotation should not even have well defined values in the short term; further, the meaning of the long-term averages would be problematic. This peculiar state of motion, unprecedented among the other satellites of the Solar System, was consistent with a number of close-up images of Hyperion obtained by the Voyager spacecraft, and more recently by the Cassini spacecraft (Fig. 5). In addition, it was noted that a consequence of the irregular rotation is that the brightness of Hyperion also be chaotic, because the shape and reflectivity of the surface is irregular. Klavetter [36] made time series measurements of Hyperion's brightness to study this question. The observations were difficult, due to the faintness of the satellite and the background of light scattered in the telescope by the very bright planet Saturn. In addition, he attempted to elucidate the state of chaotic motion by finding parameter values for a model of the rotational dynamics, by fitting the model to the time series data [36].

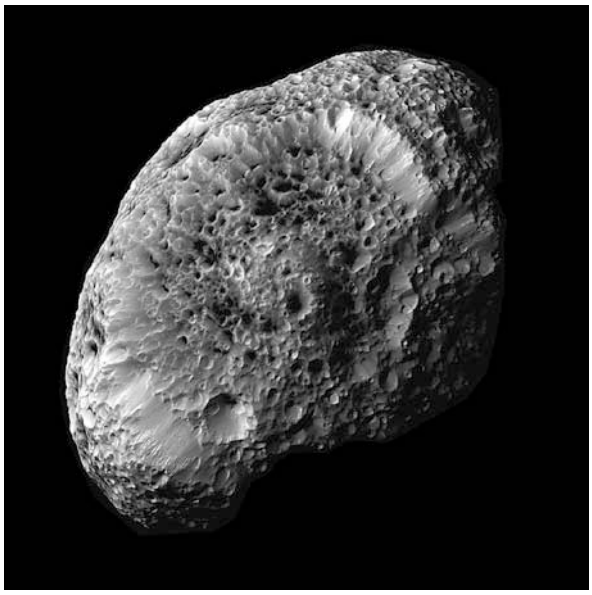

Astronomical Time Series, Complexity in, Figure 5
This image of Hyperion shows strange details across the strange, tumbling moon's surface. Differences in color could indicate differences in the surface composition. This image was taken by the NASA Cassini spacecraft during a close flyby on Sept. 26, 2005

The chaotic rotation of Hyperion is the best case of the connection between dynamical systems theory and time series analysis, and the only case where the time scales are short enough for detailed observational studies. The reader is referred to the excellent article, Chaos and Complexity in Astrophysics, by Oded Regev, in this Encyclopedia, for more details on the relevant theoretical background of both this and other complex astronomical systems.

Chaotic Time Series Analysis in Astronomy

We now turn to our main topic, the extraction of information about complex astronomical systems by analyzing observational time series data. Astronomical time series have much in common with sequential data in other sciences, but also have features of special concern. Astronomy is fundamentally an observational science. Direct ex-

periments are for the most part not feasible, and in some cases one cannot even obtain repeated observations. Consider the distribution of galaxies, cosmic background radiation, etc. on the sky. Short of sending an astronomical observatory to another part of the Universe, there is only one view of the sky. As a result, it can be very difficult to judge whether certain peculiar structures are accidental (e. g. statistical fluctuations connected with observational errors) or physically significant. Indeed, concepts such as *randomness*, *determinism*, and *probability* need to be interpreted in somewhat special ways in cosmology and the study of large-scale structure of the Universe. (For an example, see the discussion of the term *cosmic variance*, e. g., on Wikipedia.)

More germane to this discussion, many observational time series cannot be obtained at evenly spaced time intervals, as is assumed in the application of most standard analysis methods. Gaps or other sampling irregularities can interfere with the ability to characterize complex variations. This is especially true of low signal-to-noise observations with random sample times – for then, the complexities of observational errors, sampling, and true system behavior are confusingly tangled together. Even worse, the sampling can be partially random, partially semi-regular, dependent on previous values or estimated signal components, or even dependent on the real-time perceived importance of the observations to the public or to a funding agency.

Another consideration, in many but by no means all cases, is simply the quantity and quality of the data. If each data point requires dedicating a large amount observing time with an expensive telescope, long time series will not be common. Too, the object of interest may be faint and affected by instrumental and cosmic backgrounds. The analyst may thus have to be content with short, noisy time series. Early in the history of this subject most of the data analysis methods were developed in the context of laboratory or computer studies, for which large amounts of data were readily available and could be repeated (with different initial conditions or parameter values) at will. As will be seen, diagnosing complex or chaotic behavior fundamentally requires access to data on the system's return to states very close to each other, but at wide range of times. Hence the difficulties described here are particularly limiting. Accordingly many standard methods were ineffective on most astronomical time series data.

Now consider the setting in which most time series analysis of complex systems is carried out, the *embedding* procedure, with which it is possible to derive dynamical information from observations of a subset of the variables needed to specify the full dynamical behavior of the system, or even from variables that are not even dynamical in nature. This remarkable mathematical result means that observations of subsidiary quantities that are not directly involved in the dynamics can provide clues to underlying chaotic dynamical processes. For example, observations of brightness variations in stars can yield understanding of the dynamics of the stellar interior. An unfortunate feature of this mathematics is that the lack of direct connection between observable and the dynamics leaves room for data analytic results of questionable physical significance. The interested reader should supplement the intuitive, nonmathematical discussion here, with reference to one of the many excellent books on this topic, such as Sprott [67].

A primary question facing the analyst is whether the noisy data at hand were generated by a (simple) deterministic, chaotic, or random process – or some combination. The deterministic nature of chaos means that knowledge of the dynamical laws of the system and the current state, as long as they are both known accurately enough, allow accurate prediction of future states. Of course sensitive dependence on initial data prevents this kind of prediction from extending very far into the future. In contrast, if the physical process is random, there is an intrinsic uncertainty of future values even a short time into the future – perhaps due to quantum mechanical effects or the presence of complex, unmeasured hidden variables. One of the methodologies of time series analysis is to use features relating to these kinds of predictability (or lack thereof) to diagnose the nature of the underlying physical system. For practical purposes, then, short-term predictability is the key: chaotic processes may be predicted accurately over short periods of time, and random ones cannot.

It is important to note that even random noise can be predicted to some extent. A trivial example is that even white noise can be predicted by guessing that the next value will be the mean value, either as determined externally or from past observations. More germane is the fact that nonwhite noise can be predicted to an even greater degree. Correlations present in such a process can yield a prediction that is even better than the mean value. This partial predictability needs to be distinguished from the more accurate, but short term, predictability of a chaotic process.

At this point the reader will appreciate the difficulties of detecting or characterizing the various kinds of complex behavior from time series analysis. The remainder of this section will describe various properties of time series data that are useful in diagnosing chaotic dynamics and other complex behavior. Some are merely consistent with chaotic dynamics and do not really prove that chaos is present. Others are properties that of necessity must be present if the underlying process is chaotic.

An obvious first question is: To what extent the visual appearance of a plot of the time series indicate the presence of chaos? Rather easily one can visually distinguish stochastic processes from linearly deterministic ones, such as periodic oscillations or polynomial trends. (Here the term *stochastic* refers to the quality of disorder that is common to random and chaotic processes.) But the distinction between chaos and randomness is more subtle. As is evident in Fig. 1, plots of chaotic and random data with the same degree of correlation (e. g. white noise against the output of the logistic map for $\lambda = 4$) are nearly indistinguishable. The reader is encouraged to make such plots using Eq. (2) for white chaos and a random number generator for white noise.

The eye seems good at distinguishing order from disorder. Why do we have so much difficulty visualizing the difference between chaos and randomness? The problem lies not with the abilities of the human eye but largely with the way the data are presented. The distinction between various kinds of disorder, or stochastic behavior, reside in a different space than the standard time series plot, as will soon be evident.

Part of the problem with visual recognition is that the dependence of successive values of the observable X is not revealed very well by plotting them, next to each other on a time axis. This standard time series plot places too much emphasis on the time at which the measurements were made. What matters are the relationships between observations at successive times, not the absolute times of the observations. This suggests plotting X_n against X_{n+1}, as opposed to X_n against n. Compare Figs. 2 and 4.

This simple idea has been around in many guises for a long time, and probably cannot be credited to any one person. A charming introduction to the resulting plot, sometimes called a *return map* was given by Shaw [66] – with the slight twist that he plotted successive time intervals between discrete events (the falling of a water drop from a drippy faucet) against each other, rather than the values of an observable. This simple idea has grown into the most powerful tool of chaotic time series analysis. To see why this is so, we make what seems like a diversion – but we'll return soon.

This diversion is *state space*. It is generally useful to study a physical system in this abstract space, the coordinates of which are a complete set of independent dynamical variables. The use of this space lies in the fact that any configuration of the system corresponds to a unique point in this space, and vice versa. The system's complete evolution similarly corresponds to a trajectory or *orbit* in state space. This surprisingly useful geometrical view is old (cf. Poincare, whose studies of the three-body problem of celestial mechanics, in the late 19th century, are considered by some to be the origin of the idea of chaos, and certainly of the geometric view of dynamics) and well described in the modern literature of dynamical systems, e. g. Wiggins [75].

Unfortunately, it is rare that an experimentalist can measure all the physical coordinates necessary to specify the state of the system. Practical considerations often permit measurement of only a few state variables or, even worse, only peripheral quantities that depend indirectly on the state. An example of this relation is the case of the chaotic tumbling of Saturn's satellite Hyperion [36], discussed above, where the overall brightness of the satellite must serve as a surrogate for the set of variables needed to specify the rotational state (rotation direction and magnitude, plus orbital parameters). The passive nature of astronomical observations exacerbates this disconnect between observables and dynamically relevant variables. A chemist or physicist can insert a probe and measure things directly, whereas the astronomer must accept those photons, particles, or perturbations of space-time that Nature chooses to send our way.

Remarkably, extensive measurements of just one variable can reveal the structure of the orbits in the full multivariate state space. Packard, Crutchfield, Farmer, and Shaw [51] and Takens [70] show that under certain conditions on the dynamical system there is a multidimensional *embedding space*, the coordinates of which can be derived from a single observed variable. The time series data generate trajectories in this space that are simply related to the system's state-space orbits. This relation is in the form of a smooth map, from the state space to the embedding space, that preserves the topology of the evolutionary paths. Thus the essential features of the unobservable state-space orbits can be understood by studying the accessible trajectories in the embedding space. In this analysis, one does not necessarily need to know the dimension of the state space – or even the identity of its coordinates.

Of the many choices of embedding space coordinates the most common is the observed variable evaluated at a set of lagged times:

$$X = (X_n, X_{n+k}, X_{n+2k}, X_{n+3k}, \ldots, X_{n+(M-1)k}) \,. \quad (6)$$

The lag k and the embedding dimension M are positive integers. In principle the lags need not be equal, but in practice they are almost always chosen to be so. As time goes by, points defined by Eq. (6) fall on embedding-space trajectories topologically equivalent to the system's state-space orbits. The trajectories can be traced out using time series data, albeit crudely if the noise level is large. When the system satisfies the requisite mathematical conditions

(and M is high enough) one says that one has constructed a *suitable embedding space*.

We thus arrive at a most important tool for the analysis of chaos, the *phase portrait*, also called the state-space portrait. The phase portrait plays a key role in the search for order in chaos. It can be constructed directly from the time series data. In simplest form it is just a plot of X_{n+1} against X_n. In other cases the process must be described in a space of higher dimensions, and we add coordinates $X_{n+2}, X_{n+3}, \ldots$ Also, the lag k need not be one unit. A caution: the example described by Harding, Shinbrot and Cordes [32] shows that one can be fooled by the appearance of the phase portrait. They exhibit purely random, non-chaotic models for their pulsar data which produce phase portraits that show the interesting structure that is sometimes thought to uniquely characterize chaos.

In practice, with real data, there are further complications. Observational errors always produce noise that somewhat blurs the phase portrait. Yet with a sufficiently large signal-to-noise ratio this plot will reveal the nature of the dynamics in spite of noise. One does not know a priori the value of the lag. In the theorems justifying the embedding procedure – namely infinite, noise-free data streams – the value of k does not matter. In practice k does matter, and one must figure out a good value from the data. One does not know a priori the value of the dimension M of the embedding space. In theory M must be larger than $1 + 2d$ (d is the dimension of the physical state space), but in practice considerably smaller values provide suitable embeddings. However, it may well be that the correct dimension for viewing the underlying process is higher than 3, in which case there are problems displaying and visualizing the data.

The *attractor* of a physical system is, simply put, the part of state space toward which the system tends to evolve. The attractor ranges from simple closed curves in the case of periodic motion to highly complex, high-dimensional fractal sets for chaotic motion [2].

Another view of the embedding procedure is that it is the experimentalist's view of the attractor of the system. That is, the smooth transformation that maps state space into the embedding space maps the physical attractor into a set contained in the embedding space. The experimentalist cannot deal with the attractor itself, and must learn about it by studying its image in the embedding space. An important point about the phase portrait is sometimes overlooked: **the time sequence with which the sample points traverse the embedding space is important**. In simple terms, if the points smoothly trace out a figure, no matter how complex it is, the process is indicated to be regular. On the other hand, if the points jump randomly, from one part of the embedding space to another, and back again, the process is indicated to be chaotic. While this behavior can be depicted with an animated display on a computer screen, and captured in a video file, it is much harder to display on a piece of paper or in a publication.

The dimension of the attractor (which clearly cannot be larger than the embedding dimension) is an important quantity. The relation between the attractor dimension, the embedding dimension, the dimension of the embedding-space image of the attractor, and a host of other dimension quantities that have been introduced is a complex subject. See [23,70] for the pioneering work, and textbooks such as [67] for up-to-date discussions.

In turn, estimation of the dimension of the attractor from the time series data is a complex topic. There are many different kinds of dimensions that can be defined, most of which can be estimated from data. There are many different techniques for estimating a given dimension. And a single number does not always characterize fully the dimension of an attractor. Generalized dimensions and a function which represents the distribution of dimensions have been introduced (e. g., Mandelbrot [43,44], Theiler [72]).

The literature on this subject is too large to review here. An overview [72] contains 143 references. The classic paper is by the Santa Cruz group [28]; see also [23] for a very nice physical introduction. The book [45] contains many articles on dimension and dimension estimation. The statistics of dimension estimates (e. g. Holzfuss and Meyer-Kress [34], Theiler [72], Smith 1990, Ramsey and Yuan [54]) is of interest to experimentalists, for they need to assess the reliability of their results. Unfortunately the uncertainties due to other effects [14], some of which we will soon meet, may often overwhelm purely statistical uncertainties.

The technique most commonly used by experimentalists provides an estimate of the correlation dimension [30,31]. This method implements the notion that the rate at which the volume contained in a hyper sphere embedded in M dimensional space grows with radius as R^M.

Unfortunately there are severe problems with the use of dimension estimates to diagnose chaotic dynamics. It is fair to say that the technique is essentially useless for detecting the presence of chaos in most astronomical applications. To start with, large amounts of data are necessary to determine dimensions with any reliability. Next, even modest noise can affect the results [49]. Smoothing of the data, due to either experimental effects (e. g., finite-time-constant instrumentation) or physical effects (e. g., radiative transfer effects in the case of astronom-

ical objects of variable luminosity) distorts the dimension [6,7,15,47,52,59,60,61,62].

Furthermore, there are more fundamental difficulties with extracting useful dimensional information from time series. Contrary to assumptions frequently made by early workers in this field, determination of a low dimension, fractional or otherwise, does not prove the existence of a chaotic process underlying the time series. In a seminal paper, Osborne and Provenzale [50] showed that purely non-chaotic colored noise has a low dimension as determined with the Grassberger–Procaccia and similar algorithms. They showed that the dimension is a simple function of the steepness (quantified by a power-law index) of the power spectrum. Accordingly, essentially any dimension can be produced by a purely random (no chaos need apply!) process, with the appropriate power spectrum. But see [53] for some methods to address the problem of distinguishing chaos from randomness, and [74] for a broad overview of the problems encountered with astronomical time series.

A second fundamental issue was raised by Ruelle and Eckmann [58], who showed that there is a simple upper limit to the value that can be obtained using the Grassberger–Procaccia or any other approach to dimension estimation where one regresses counts of points in boxes (hyper cubes) in the embedding space against the box volume. By noting the limiting cases of one point per box and all N points in the box, one sees that the maximum logarithmic slope (which is the dimension estimate) of the regression curve is $\log(N)$, approximately. This is an order of magnitude estimate, and it hardly matters whether the base of the logarithm is e or 10.

Hence if the true dimension is any value greater than the logarithm of the number of data points, one will estimate a dimension $D \sim \log(N)$. Unfortunately the astronomical literature is littered with dimension estimates in the range $D \sim 2$–3 or so, based on 100 to 1000 or so data points, accompanied by breathless claims of detection of chaos. See also [57] for other comments aimed at dampening some of the enthusiasm in this arena.

In summary, for a number of reasons, estimates of the "fractal dimension" of time series data is nearly useless as a litmus test for nonlinear deterministic dynamics, or chaos.

Period doubling and *bifurcations* are sometimes associated with chaotic dynamics. The idea is as follows: suppose a dynamical system has periodic behavior, and the value of the period changes as some parameter of the system changes. An example is the logistic process of Eq. (2), for some values of λ less than 4. It is well known that this exceedingly simple system has exceedingly complex behavior, including critical parameter values when one period becomes two. Although some variable stars and other periodic astronomical systems exhibit this kind of behavior, as a function of time and possibly related to underlying changes in physical parameters, these phenomena are not very common in astronomy, and therefore will not be discussed here. In any case, much of the theoretical research in this area focuses on control – that is, where the parameter or parameters can be adjusted more or less arbitrarily. Once again the passive, observational nature of astronomy comes in to play. Nevertheless, there is at least one feature of time series data that has been taken as indicative of possible period doubling, namely systematic amplitude changes of an otherwise periodic process. For example, in the light curve of a variable star, "the alternation of deep and shallow minima" is some indication of period doubling [13]. Clearly such a conclusion is speculative, and would need to be supported by further observations, e. g., of the light curve as conditions in the stellar interior or atmosphere undergo secular changes.

It was remarked earlier that sensitivity to initial conditions (SIC) is a key feature of chaos. There is a measure of the exponential divergence of trajectories, namely Lyapunov exponents. There is a huge literature on this concept (e. g. early papers such as, Eckmann, Kamphorst, Ruelle and Ciliberto [21], Wolf, Swift, Swinney and Vastano [78], and Sprott [67]) that measure this sensitivity. To assess divergence of nearby trajectories from time series, it is necessary to have data describing more than one trajectory, or a trajectory that returns rather close to earlier points of itself. Measuring Lyapunov exponents, either averaged over the trajectory or as a function of position in the embedding space is obviously a very difficult problem. As the Goldilocks tale at the beginning of this article hinted, the short, noisy time series available to many fields of observational astronomy are simply not sufficient to determine the exponential divergences contained in the Lyapunov exponent. There are no cases in astronomy where a reliable Lyapunov exponent has been determined from time series data.

Related analytic topics are prediction of time series and determination of the equations of motion of the observed system, to assess possible nonlinearity, randomness, nonstationarity, *etc.* Prediction of a process from observations of its past is as important in the analysis of chaotic time series as it is for random ones. Accordingly a large literature has developed (e. g. Abarbanel, Brown and Kadtke [1], Crutchfield [16], Crutchfield and McNamara [18], Crutchfield and Packard [19], Crutchfield and Young [20], Farmer and Sidorowich [24]). Various authors have pursued the related topic of *noise reduc-*

tion [25,38,69]. Relevant scaling properties are discussed by Atmanspacher, Scheingraber and Voges [4]. Hempelmann and Kurths [33] discuss analysis of the luminosity of cataclysmic variables (SS Cygni in particular) using symbolic dynamics, an approach that has considerable promise in this area [20]. But again, data limitations have prevented much success in the application of these methods in astronomy.

Let's end with a brief discussion of variability in brightness of astronomical objects, in various wavelength ranges. Regev [55] and Buchler and Regev [13] reviewed theoretical work on possible chaotic dynamics of stellar pulsations and analysis of variable star data. Numerical work that suggests chaotic behavior in stellar models includes [3,11,12,39,48]. Gillet [29] on the other hand does not confirm the period-doubling bifurcation scenario suggested by Kovács and Buchler [39] and intermittency and period-doubling bifurcations suggested in the work of Aikawa [4]. Serre, Buchler, and Goupil [65] study forecasting of variable star light curves using nonlinear methods. See also Kollath [37].

A final study demonstrates a data-driven approach. Scargle, Steiman-Cameron, Young, Donoho, Crutchfield and Imamura [64] started with a beautiful time series of X-ray observations of the bright x-ray source in Scorpius (Sco X-1) from the EXOSAT satellite. The observations covered nearly 10 hours, with a time resolution of 2 ms – thus covering 7 decades of time-scales. The only significant source of noise in these observations was the fluctuations associated with photon counting statistics, often called Poisson noise. Above such noise background this source evidences random variations of large amplitude and over a wide range of time scales.

In carrying out simple exploratory time series plots, it became evident that the general appearance of this variability is independent of time scale. This *fractal* or *scaling behavior* is common astrophysical systems (see Mandelbrot [43,44], and Flandrin [26] for the mathematics). This led to consideration of nonlinear physical models with outputs having such behavior. The classical dripping faucet system, the first archetypical chaotic system [66], had been generalized from a point system to a spatially extended one [17], the *dripping handrail* – so-called because the physical metaphor involved condensation of dew on a horizontal rail. Substituting accreting hot plasma for water, the inner edge of the accretion disk for the handrail, and a density driven instability for the drip mechanism, we hypothesized that the stochastic behavior in Sco X-1 can be described in terms of this kind of nonlinear dynamical system (see Young and Scargle [79] for a more detailed exposition of the model).

We studied the x-ray luminosity time series using the wavelet spectrum (called the *scalegram*), an analog of the ordinary Fourier power spectrum better suited to displaying power-law power spectra ("$1/f$ noise") associated with self-scaling in time series data. The basic idea was to compare scalegrams from the data against ones computed from the output of numerical simulations of the dripping handrail model. It was natural that the model produced power-law wavelet spectra, but astonishingly it also produced a local peak in the scalegram – a spectral feature indicating what astronomers call a quasi-periodic oscillation, or QPO. Astonishing because we were not even trying to fit this feature, and because the conventional wisdom was that the $1/f$ noise and the QPO are produced by completely different phenomena. Here a single, incredibly simple, physical model yielded both spectral features, fitting the scalegram estimated from the time series data quite well. Note that wavelets [22,73] are useful tools for diagnosing stochastic (random or chaotic) time series. Since the corresponding transforms are linear, this is a counterexample to the myth that linear tools are not useful for "nonlinear data."

A similar model to the dripping handrail was proposed by Mineshige and co-workers (Mineshige, Takeuchi, and Nishimori [46] and earlier references therein), based on another non-linear dynamical paradigm called *self-organized criticality*. The physical metaphor for their work was the instabilities when sand is gradually dropped onto a flat surface, ultimately generating avalanches with interesting self-scaling properties, similar to those of some accretion systems [8]. Hence this is sometime called the *sand pile model*. Accretion systems in astronomy are complex and still not well understood. While there is not general acceptance of nonlinear dynamics as explaining the scaling and QPO features in the power spectra, this approach is competitive with the others under study.

There reader may wish to consult several general collections of papers dealing with chaotic phenomena in astrophysics [9,10,56].

Future Directions

Admitedly this article has a somewhat gloomy tone, since it has emphasized the difficulties of applying many of the standard data analysis methods to astronomical time series, including problems with the data as well as limitations on existing analysis methods. But there will be great future progress in this topic, due to both increased data quantity and quality and the development of more sophisticated data analysis technology. As we enter the Age of Digital Astronomy, the improved detectors, telescopes, and space

platforms are already producing large-scale surveys of the sky with good time coverage and resolution, coupled with photometric accuracy. These platforms will yield cornucopias of data which – coupled with advanced machine learning and data mining techniques, adapted and applied to the study of nonlinear dynamical systems from time series data – will greatly improve our knowledge of the astonishingly violent, irregular processes occurring throughout the Universe. [As a specific aside, it may be noted that "projection pursuit" [27,35] provides ways to deal with multidimensional data of any kind. While most workers have used it only to search for clustering in multidimensional data, projection pursuit may prove to be useful for state space structures.] In addition, multivariate visualization techniques, especially those aimed at elucidating the time evolution of states of a complex system, will be crucial to future progress.

Bibliography

Primary Literature

1. Abarbanel H, Brown R, Kadtke J (1989) Prediction and system identification in chaotic nonlinear systems: Time series with broadband spectra. Phys Lett A 138:401–408
2. Abraham R, Shaw C (1992) Dynamics, the Geometry of Behavior. Addison Wesley, Reading
3. Aikawa T (1987) The Pomeau–Manneville Intermittent Transition to Chaos in Hydrodynamic Pulsation Models. Astrophys Space Sci 139:218–293
4. Aikawa T (1990) Intermittent chaos in a subharmonic bifurcation sequence of stellar pulsation models. Astrophys Space Sci 164:295
5. Atmanspacher H, Scheingraber H, Voges W (1988) Global scaling properties of a chaotic attractor reconstructed from experimental data. Phys Rev A 37:1314
6. Badii R, Broggi G, Derighetti B, Ravani M, Cilberto S, Politi A, Rubio MA (1988) Dimension Increase in Filtered Chaotic Signals. Phys Rev Lett 60:979–982
7. Badii R, Politi A (1986) On the Fractal Dimension of Filtered Chaotic Signals. In: Mayer-Kress G (ed) Dimensions and Entropies in Chaotic Systems, Quantification of Complex Behavior. Springer Series in Synergetics, vol 32. Springer, New York
8. Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality: An explanation of 1/f noise. Phys Rev Lett 59:381–384
9. Buchler J, Eichhorn H (1987) Chaotic Phenomena in Astrophysics. Proceedings of the Second Florida Workshop in Nonlinear Astronomy. Annals New York Academy of Sciences, vol 497. New York Academy of Science, New York
10. Buchler J, Perdang P, Spiegel E (1985) Chaos in Astrophysics. Reidel, Dordrecht
11. Buchler JR, Goupil MJ, Kovács G (1987) Tangent Bifurcations and Intermittency in the Pulsations of Population II Cepheid Models. Phys Lett A 126:177–180
12. Buchler JR, Kovács G (1987) Period-Doubling Bifurcations and Chaos in W Virginis Models. Ap J Lett 320:L57–62
13. Buchler JR, Regev O (1990) Chaos in Stellar Variability. In: Krasner S (ed) The Ubiquity of Chaos. American Association for the Advancement of Science, Washington DC, pp 218–222
14. Caswell WE, York JA (1986) Invisible Errors in Dimension Calculation: geometric and systematic effects. In: Mayer-Kress G (ed) Dimensions and Entropies in Chaotic Systems, Quantification of Complex Behavior. Springer Series in Synergetics, vol 32. Springer, New York
15. Chennaoui A, Pawelzik K, Liebert W, Schuster H, Pfister G (1988) Attractor reconstruction from filtered chaotic time series. Phys Rev A 41:4151–4159
16. Crutchfield J (1989) Inferring the dynamic, quantifying physical Complexity. In: Abraham N et al (eds) Measures of Complexity and Chaos. Plenum Press, New York
17. Crutchfield J, Kaneko K (1988) Are Attractors Relevant to Fluid Turbulence? Phys Rev Let 60:2715–2718
18. Crutchfield J, McNamara B (1987) Equations of motion from a data series, Complex Syst 1:417–452
19. Crutchfield J, Packard NH (1983) Symbolic Dynamics of Noisy Chaos. Physica D 7:201–223
20. Crutchfield J, Young K (1989) Inferring Statistical Complexity. Phys Rev Lett 63:105–108
21. Eckmann J-P, Kamphorst SO, Ruelle D, Ciliberto S (1986) Liapunov Exponents from Time Series. Phys Rev A 34:4971–4979
22. Fang L-Z, Thews RL (1998) Wavelets in Physics. World Scientific, Singapore. A physics-oriented treatment of wavelets, with several astronomical applications, mostly for spatial or spectral data
23. Farmer JD, Ott E, Yorke JA (1983) The dimension of chaotic attractors. Physica D 7:153–180
24. Farmer JD, Sidorowich J (1987) Predicting chaotic time series. Phys Rev Lett 59:845–848
25. Farmer JD, Sidorowich J (1988) Exploiting chaos to predict the future and reduce noise. In: Lee Y (ed) Evolution, Learning and Cognition. World Scientific Pub Co, Singapore
26. Flandrin P (1999) Time-Frequency/Time-Scale Analysis. Academic Press, San Diego . Volume 10 in an excellent series, Wavelet Analysis and Its Applications. The approach adopted by Flandrin is well suited to astronomical and physical applications
27. Friedman JH, Tukey JW (1974) A Projection Pursuit Algorithm for Exploratory Data Analysis. IEEE Trans Comput 23:881–890
28. Froehling H, Crutchfield J, Farmer JD, Packard NH, Shaw R (1981) On Determining the Dimension of Chaotic Flows. Physica D 3:605–617. C Herculis. Astron Astrophys 259:215–226
29. Gillet D (1992) On the origin of the alternating deep and shallow light minima in RV Tauri stars: R Scuti and A. Astron Astrophys 259:215
30. Grassberger P, Procaccia I (1983) Characterization of strange attractors. Phys Rev Lett 50:346–349
31. Grassberger P, Procaccia I (1983) Estimation of the Kolmogorov entropy from a chaotic signal. Phys Rev A 28:2591–2593
32. Harding A, Shinbrot T, Cordes J (1990) A chaotic attractor in timing noise from the VELA pulsar? Astrophys J 353:588–596
33. Hempelmann A, Kurths J (1990) Dynamics of the Outburst Series of SS Cygni. Astron Astrophys 232:356–366
34. Holzfuss J, Mayer-Kress G (1986) An Approach to Error-Estimation in the Application of Dimension Algorithms. In: Mayer-Kress G (ed) Dimensions and Entropies in Chaotic Systems, Quantification of Complex Behavior. Springer Series in Synergetics, vol 32. Springer, New York

35. Jones MC, Sibson R (1987) What is Projection Pursuit? J Roy Statis Soc A 150:1–36
36. Klavetter JJ (1989) Rotation of Hyperion. I – Observations. Astron J 97:570. II – Dynamics. Astron J 98:1855
37. Kollath Z (1993) On the observed complexity of chaotic stellar pulsation. Astrophys Space Sci 210:141–143
38. Kostelich E, Yorke J (1988) Noise reduction in dynamical systems. Phys Rev A 38:1649–1652
39. Kovács G and Buchler J R (1988) Regular and Irregular Pulsations in Population II Cepheids. Ap J 334:971–994
40. Lecar M, Franklin F, Holman M, Murray N (2001) Chaos in the Solar System. Ann Rev Astron Astrophys 39:581–631
41. Lissauer JJ (1999) Chaotic motion in the Solar System. Rev Mod Phys 71:835–845
42. Lissauer JJ, Murray CD (2007) Solar System Dynamics: Regular and Chaotic Motion. Encyclopedia of the Solar System, Academic Press, San Diego
43. Mandelbrot B (1989) Multifractal Measures, Especially for the Geophysicist. Pure Appl Geophys 131:5–42
44. Mandelbrot B (1990) Negative Fractal Dimensions and Multifractals. Physica A 163:306–315
45. Mayer-Kress G (ed) (1986) Dimensions and Entropies in Chaotic Systems, Quantification of Complex Behavior. Springer Series in Synergetics, vol 32. Springer, New York
46. Mineshige S, Takeuchi M, Nishimori H (1994) Is a Black Hole Accretion Disk in a Self-Organized Critical State. ApJ 435:L12
47. Mitschke F, Moeller M, Lange W (1988), Measuring Filtered Chaotic Signals. Phys Rev A 37:4518–4521
48. Moskalik P, Buchler JR (1990) Resonances and Period Doubling in the Pulsations of Stellar Models. Ap J 355:590–601
49. Norris JP, Matilsky TA (1989) Is Hercules X-1 a Strange Attractor? Ap J 346:912–918
50. Osborne AR, Provenzale A (1989) Finite correlation dimension for stochastic systems with power-law spectra. Physica D 35:357–381
51. Packard NH, Crutchfield JP, Farmer JD, Shaw RS (1980) Geometry from a Time Series. Phys Rev Lett 45:712–716
52. Paoli P, Politi A, Broggi G, Ravani M, Badii R (1988) Phase Transitions in Filtered Chaotic Signals. Phys Rev Lett 62:2429–2432
53. Provenzale A, Smith LA, Vio R, Murante G (1992) Distinguishing between low-dimensional dynamics and randomness in measured time series. Physica D 58:31
54. Ramsey JB, Yuan H-J (1990) The Statistical Properties of Dimension Calculations Using Small Data Sets. Nonlinearlity 3:155–176
55. Regev O (1990) Complexity from Thermal Instability. In: Krasner S (ed) The Ubiquity of Chaos. American Association for the Advancement of Science, Washington DC, pp 223–232. Related work in press, MNRAS
56. Regev O (2006) Chaos and complexity in astrophysics, Cambridge University Press, Cambridge
57. Ruelle D (1990) Deterministic Chaos: The Science and the Fiction (the Claude Bernard Lecture for 1989). Proc Royal Soc London Ser A Math Phys Sci 427(1873):241–248
58. Ruelle D, Eckmann J-P (1992) Fundamental limitations for estimating dimensions and Lyapunov exponents in dynamical systems. Physica B 56:185–187
59. Scargle JD (1989) An introduction to chaotic and random time series analysis. Int J Imag Syst Tech 1:243–253
60. Scargle JD (1989) Random and Chaotic Time Series Analysis: Minimum Phase-Volume Deconvolution. In: Lam L, Morris H (eds) Nonlinear Structures in Physical Systems. Proceedings of the Second Woodward Conference. Springer, New York, pp 131–134
61. Scargle JD (1990) Astronomical Time Series Analysis: Modeling of Chaotic and Random Processes. In: Jaschek C, Murtagh F (eds) Errors, Bias and Uncertainties in Astronomy. Cambridge U Press, Cambridge, pp 1–23
62. Scargle JD (1990) Studies in astronomical time series analysis. IV: Modeling chaotic and random processes with linear filters. Ap J 343:469–482
63. Scargle JD (1992) Predictive deconvolution of chaotic and random processes. In: Brillinger D, Parzen E, Rosenblatt M (eds) New Directions in Time Series Analysis. Springer, New York
64. Scargle JD, Steiman-Cameron T, Young K, Donoho D, Crutchfield J, Imamura J (1993) The quasi-periodic oscillations and very low frequency noise of Scorpius X-1 as transient chaos – A dripping handrail? Ap J Lett 411(part 2):L91–94
65. Serre T, Buchler JR, Goupil M-J (1991) Predicting White Dwarf Light Curves. In: Vauclair G, Sion E (eds) Proceedings of the 7th European Workshop. NATO Advanced Science Institutes (ASI) Series C, vol 336. Kluwer, Dordrecht, p 175
66. Shaw R (1984) The Dripping Faucet as a Model Chaotic System. Aerial Press, Santa Cruz
67. Sprott JC (2003) Chaos and Time-Series Analysis, Oxford University Press, Oxford. This is a comprehensive and excellent treatment, with some astronomical examples
68. Steiman-Cameron T, Young K, Scargle J, Crutchfield J, Imamura J, Wolff M, Wood K (1994) Dripping handrails and the quasi-periodic oscillations of the AM Herculis. Ap J 435:775–783
69. Sugihara G, May R (1990) Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. Nature 344:734
70. Takens F (1981) Detecting strange attractors in turbulence. In: Rand DA, Young L-S (eds) Dynamical Systems and Turbulence, Lecture Notes in Mathematics, vol 898. Springer, pp 366–381
71. Theiler J (1986) Spurious Dimension from Correlation Algorithms Applied to Limited Time-series data. Phys Rev A 34:2427–2432
72. Theiler J (1990) Statistical Precision of Dimension Estimators. Phys Rev A 41:3038–3051
73. van den Berg JC (1998) Wavelets in Physics, Cambridge University Press, Cambridge. A physics-oriented treatment of wavelets, with a chapter by A Bijaoui on astrophysical applications, some relating to complex systems analysis
74. Vio R, Cristiani S, Lessi O, Provenzale A (1992) Time Series Analysis in Astronomy: An Application to Quasar Variability Studies. Ap J 391:518–530
75. Wiggins S (2003) Introduction to Applied Nonlinear Dynamical Systems and Chaos, 2nd edn. Springer, New York
76. Wisdom J (1981) The origin of the Kirkwood gaps: A mapping for asteroidal motion near the 3/1 commensurability. The resonance overlap criterion and the onset of stochastic behavior in the restricted three-body problem. Ph D Thesis, California Inst of Technology
77. Wisdom J, Peale SJ, Mignard F (1984) Icarus 58:137
78. Wolf A, Swift JB, Swinney HL, Vastano JA (1985) Determining Lyapunov exponents from a time series, Physica D 16: 285–317
79. Young K, Scargle J (1996) The Dripping Handrail Model: Transient Chaos in Accretion Systems. Ap J 468:617

Books and Reviews

Heck A, Perdang JM (eds) (1991) Applying Fractals in Astronomy. Springer, New York. A collection of 10 papers

Jaschek C, Murtagh F (eds) (1990) Errors, Bias and Uncertainties in Astronomy. Cambridge U Press, Cambridge

Lowen SB, Teich MC (2005) Fractal-Based Point Processes, Wiley-Interscience, Hoboken. An excellent treatment of the unusual subject of fractal characteristics of event data generated by complex processes

Maoz D, Sternverg A, Leibowitz EM (1997) Astronomical Time Series, Kluwer, Dordrecht. Proceedings of the Florence and George Wise Observatory 25th Anniversary Symposium

Ruelle D (1989) Chaotic Evolution and Strange Attractors: The Statistical Analysis of Time Series for Deterministic Nonlinear Systems. Cambridge U Press, Cambridge

Statistical Challenges in Modern Astronomy, a series of books derived from the excellent Penn State conference series, with many papers on time series or related topics. http://astrostatistics.psu.edu/scma4/index.html

Subba Rao T, Priestly MB, Lessi O (eds) (1997) Applications of Time Series Analysis in Astronomy and Meteorology. Chapman & Hall, London. Conference proceedings; sample data were available to the participants

Astrophysics, Chaos and Complexity in

Oded Regev[1,2]

[1] Department of Physics, Technion–Israel – Institute of Technology, Haifa, Israel

[2] Department of Astronomy, Columbia University, New York, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Hamiltonian Chaos in Planetary, Stellar and Galactic Dynamics
Chaotic Time Variability of Astronomical Sources
Spatio-Temporal Patterns and Complexity in Extended Systems
Future Directions
Bibliography

Glossary

Accretion disk A flat gaseous structure into which matter endowed with angular momentum settles when it is being gravitationally attracted to a relatively compact object. Dissipative processes, which are needed to allow for the extraction of angular momentum from the orbiting gas so as to allow for it to be accreted by the star, heat the disk and make it observable by the outgoing radiation.

Attractor A geometrical object in the state space of a dynamical system to which the system's trajectories tend for large times (or iteration numbers). Attractors of dissipative systems have a dimension that is lower than that of the state space and when this dimension is fractal, the attractor is called *strange*.

Bifurcation A qualitative change in the essential behavior of a mathematical system (e. g., algebraic equations, iterated maps, differential equations), as a parameter is varied. Usually involves an abrupt appearance and/or disappearance of solutions.

Close binary stars Two stars in a Keplerian orbit around each other, having a small enough separation, so that they may affect each other during their evolution (for example by mass transfer). When the pair contains, in addition to a normal unevolved star, a compact object (in *cataclysmic binaries* it is a white dwarf and in *X-ray binaries* a neutron star or black hole) the system may be observationally prominent. Mass transfer from the normal star via an accretion disk onto the compact object may result in violent phenomena, such as nova explosions or X-ray bursts.

Dynamical system (DS) A set of rules by application of which the state of a physical (or some other, well defined) system can be found, if an initial state is known. Iterated maps (IM) in which the evolution proceeds in discrete steps, i. e., one defined by appropriate iterative formula, and differential equations, both ordinary (ODE) and partial (PDE) are typical DS.

DS flow All *trajectories* (or *orbits*) of a DS in its *state space*, which is spanned by the relevant state variables. The flow is fully determined by the DS and may serve as its geometrical representation.

Hubble time Estimate for the age of the universe obtained from taking the inverse of the *Hubble constant*, the constant ratio between the recession velocity of galaxies and the distance to them (Hubble law). Current estimates give the Hubble time as ~ 14 Gyr.

Interstellar medium (ISM) Diffuse matter of low density, filling the space between stars in a galaxy. The average number density of the ISM is typically $n \sim 1\,\mathrm{cm}^{-3}$ but its distribution is highly non-uniform, with the densest clouds reaching $n \sim 10^6\,\mathrm{cm}^{-3}$.

Fractal A set whose (suitably defined) geometrical dimension is non-integral. Typically, the set appears self-similar on all scales. A number of geometrical objects associated with chaos (e. g. strange attractors) are fractals.

Integrable DS A DS whose exact solution may be calculated analytically in terms of elementary or special functions and possibly quadratures. In integrable Hamiltonian DS the motion proceeds on a set of invariant tori in phase space.

Integral of motion A function of a DS's state variables which remains constant during the evolution. The existence of integrals of motion allows for a dimensional reduction of the DS. A sufficient number of independent integrals of motion guarantees integrability.

KAM theorem An important theorem by Kolmogorov, Arnold and Moser (KAM), elucidating the loss of integrability and transition to chaos in Hamiltonian DS by the breakdown of invariant tori.

Liapunov exponents A quantitative measure of the local divergence of nearby trajectories in a DS. If the largest such exponent, Λ, is positive, trajectories diverge locally exponentially, giving rise to chaotic behavior. In that case a measure for the time that chaos is manifested, the *Liapunov time*, is $\propto 1/\Lambda$.

Poincaré section A geometrical construction with the help of which the behavior of a multi-dimensional DS can be examined by means of a two-dimensional IM. The IM is obtained by finding the system's trajectory successive crossings of a given surface cutting the state space of the DS (*surface of section*). Essential features of the dynamics can be captured if the surface is placed in an appropriate position.

Pulsating stars Stars that, due to an instability, exhibit radial or non-radial pulsation. Most have periodic variations, with periods ranging from a few days to years, but some are irregular. The classical Cepheids, which are the most common regular pulsators, were instrumental in determining cosmological distances by using their known period-luminosity relation.

Shilnikov scenario One of the known mathematical structures that guarantee chaos in dissipative DS. Let an unstable fixed point exist in the DS state space and have a fast unstable direction and a stable subspace of trajectories that slowly spiral in. If a homoclinic (connecting the point to itself) orbit exists then there are chaotic orbits around it.

Definition of the Subject

Astronomy is the science that deals with the origin, evolution, composition, distance to, and motion of all bodies and scattered matter in the universe. It includes astrophysics, which is usually considered to be the theoretical part of astronomy, and as such focuses on the physical properties and structure of cosmic bodies, scattered matter and the universe as a whole. Astrophysics exploits the knowledge acquired in physics and employs the latter's methods, in an effort to model astronomical systems and understand the processes taking place in them.

In recent decades a new approach to nonlinear dynamical systems (DS) has been introduced and applied to a variety of disciplines in which DS are used as mathematical models. The theory of chaos and complexity, as this approach is often called, evolved from the study of diverse DS which behave unpredictably and exhibit complex characteristics, despite their seeming simplicity and deterministic nature. The complex behavior is attributed to the property of *sensitivity to initial conditions* (SIC), whereby despite their completely deterministic behavior, two identical systems in initial states differing only by a minute amount, relatively rapidly develop in very different ways. The theory has been most fruitful in physics, but it is important to note that its first paradigms were actually deeply related to astrophysical systems.

A large majority of astrophysical systems are theoretically modeled by nonlinear DS. The application of the ideas and methods of chaos and complexity theory to astrophysics seems thus to be natural. Indeed, these methods have already been exploited in the study of some systems on a vast variety of scales – from planetary satellites through pulsating stars and up to the large scale structure of the universe. The main importance of these approaches is in their ability to provide new analytical insights into the intricate physical processes taking place in cosmic matter, shaping the various astronomical objects and causing the prominent observable phenomena that occur in them.

Introduction

Newton's laws of mechanics and gravitation had their most prominent application in the significant work of Laplace, *Celestial Mechanics*. It appeared in five volumes between 1798 and 1827 and summarized his mathematical development and application of Newton's work. Laplace offered a complete mechanical interpretation of the Solar System – planets and their satellites, including the effects of perturbations and tidal interactions. This work had immediately been adopted as an unequivocal manifestation of Nature's determinism, that is, the possibility to precisely determine the future of any DS – a set of ordinary differential equations (ODE) in this case – if only appropriate initial conditions are known. Laplace himself was so confident in his results that, according to one story, when Napoleon asked him what is the role of the Creator in his theory, he replied that "this hypothesis" was redundant.

Remarkably, the same system in celestial mechanics (and actually its simplest form – containing just three bodies) which had been considered to be the primary paradigm of determinism gave rise, almost a century after Laplace's work, to profound analytical difficulties. Another great French mathematician, Henri Poincaré, made the fundamental discovery that the gravitational n-body system is non-integrable, that is, its general solution can not be expressed analytically, already for $n > 2$. This work initiated the development of the theory of chaos in Hamiltonian systems, which culminated, about 50 years ago, in the Kolmogorov–Arnold–Moser (KAM) theorem. Section "Hamiltonian Chaos in Planetary, Stellar and Galactic Dynamics" of this article reviews the application of this theory to planetary, stellar and galactic dynamics.

In unrelated investigations of dissipative DS (Hamiltonian systems are conservative) in the 1960s, aperiodic behavior was observed in numerical studies of simplistic models of thermal convection in geophysical and astrophysical contexts by Lorentz [40] and Moore and Spiegel [44]. It was conjectured that this behavior is due to SIC, with which these systems were endowed. By the same time, the American mathematician Stephen Smale discovered a new class of "strange" attractors on which the dynamics is chaotic and which naturally arise in DS – sets of ODE or iterated maps (IM) – endowed with SIC. In the early 1980s, it was realized, following the work of Mitchell Feigenbaum and others on bifurcations in quadratic IM, that such DS possess a universal behavior as they approach chaos. Some applications of these developments to astrophysical systems are discussed in Sect. "Chaotic Time Variability of Astronomical Sources".

A variety of important physical (and astrophysical) systems, notably fluids, are modeled by partial differential equations (PDE). This class of DS provides the *spatio-temporal* information, necessary for the description and understanding of the relevant phenomena. These DS, especially when they are nonlinear (as is often the case), are significantly more complicated than ODE and IM, but appropriate reduction techniques can sometimes help in applying to them the knowledge gained for ODE and IM. These and a variety of other techniques constitute what is called today *pattern theory*. Spatio-temporal patterns have been identified and categorized and they help to understand, using analytical and semi-analytical (usually perturbative) approaches, some essential properties of the solutions. It is of interest to properly quantify, describe and understand the processes that create spatio-temporal complexity, like the shapes of interstellar clouds, patterns in thermal convection, turbulent flows and other astrophysically relevant topics. Section "Spatio-Temporal Patterns and Complexity in Extended Systems" is devoted to a review of these issues.

The quest for the universal and the generic, from which an understanding of complicated and seemingly erratic and heterogenous natural phenomena can emerge, is central to the applications of the theory of chaos, patterns and complexity in DS to astrophysics. It may provide a powerful investigative tool supplementing fully fledged numerical simulations, which have traditionally been employed in the study of nonlinear DS in astrophysics.

Hamiltonian Chaos in Planetary, Stellar and Galactic Dynamics

Planetary systems, star clusters of various richness, galaxies, galaxy clusters and super-clusters (clusters of clusters) can be approached most directly by considering the gravitational n-body problem (GNBP). In this approach all other interactions, save the gravitational one, are neglected, and the bodies are considered to be point masses. This model is viable, at least approximately, when the distances between the bodies are much larger than their typical size and any scattered matter between the bodies has only a very small effect on the dynamics.

The GNBP is one of the paradigms of classical dynamics, whose development originated in Newton's laws and through the implementation of techniques of mathematical analysis (mainly by Lagrange in his work *Analytical Mechanics*, published in 1788) finally acquired a powerful abstract formulation in the form of Hamiltonian canonical formalism. Hamiltonian systems, that is, those that obey the Hamilton equations are endowed with important conservation properties linked to symmetries of the Hamiltonian, a function that completely describes the system. This function is defined on the configuration space which consists of the *phase space* (spanned by the generalized coordinates and momenta) and the time coordinate. A Hamiltonian system conserves phase volume and sometimes also other quantities, notably the total energy (when the Hamiltonian does not depend explicitly on time, as is the case in the GNBP).

The Hamilton equations consist of two first order ODE for each degree of freedom, thus the GNBP in a three-dimensional physical space yields, in general, $6n$ equations and so for large n it is quite formidable. The $n = 2$ case is reducible to an equivalent one body problem, known as the Kepler problem. The complete solution of this problem was first given by Johann Bernoulli in 1710, quite long before the Lagrange–Hamilton formalism was introduced. The gravitational two-body problem has been successfully applied to various astrophysical systems, e. g. the

motion of planets and their satellites and the dynamics of binary stars. The quest for a similar reduction for systems with $n > 2$, was immediately undertaken by several great mathematicians of the time. The side benefit of these studies, conducted for almost two centuries, was a significant progress in mathematics (mainly in the theory of ODE), but definite answers were found in only some particular, rather limited cases. Attempts to treat even the simplest problem of this kind, (the restricted, planar, three-body, i. e, where one of the bodies is so light that its gravity has no effect on the dynamics and all the orbits are restricted to a plane) ended in a failure. All general perturbative approaches invariably led to diverging terms because of the appearance of "small divisors" in the perturbation series.

The first real breakthrough came only in the 1880s, when Henri Poincaré worked on the GNBP, set for a prize by King Oscar II of Sweden. Poincaré did not provide a solution of the problem, but he managed to understand why it is so hard to solve. By ingenious geometrical arguments, he showed that the orbits in the restricted three-body problem are too complicated to be described by any explicit formula. In more technical terms, Poincaré showed that the restricted three-body problem, and therefore the general GNBP, is *non-integrable*. He did so by introducing a novel idea, now called a Poincaré section, with the help of which he was able to visualize the essentials of the dynamics by means of a two-dimensional area preserving IM. Figure 1 shows schematically a typical mathematical structure, a *homoclinic tangle* in this case, that is behind chaotic behavior in systems like the restricted three-body problem (and, incidentally, also the forced pendulum). The existence of a hyperbolic fixed point in the appropriate equivalent IM and the transversal intersection of its stable and unstable manifolds gives rise to the complex behavior. Remarkably, Poincaré was able to visualize such a structure without the aid of computer graphics.

The non-integrability of the GNBP naturally prompted the question of the Solar System stability. No definite answer to this problem could however be reasonably expected on the basis of analytical work alone. Before electronic computers became available, this work had largely been based on perturbation or mean field methods. Despite the difficulties, these efforts yielded new and deep insights on chaotic behavior in Hamiltonian systems after intensive work of close to 50 years. First and foremost among these is the KAM theorem, which elucidated the mathematical process by which an integrable Hamiltonian system transits to chaotic behavior by losing its integrability, when a suitably defined control parameter (e. g. the relative size of the non-integrable perturbation to an integrable Hamiltonian) is gradually increased from

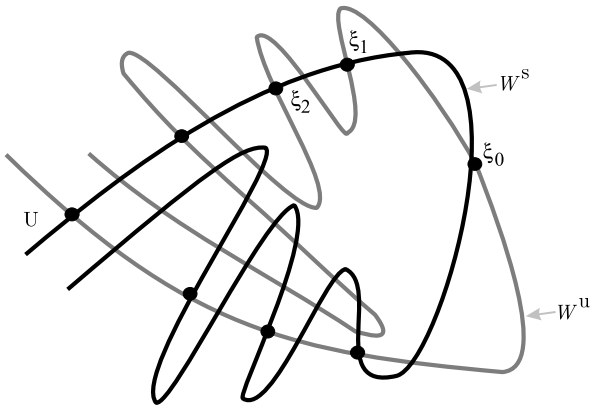

Astrophysics, Chaos and Complexity in, Figure 1
Schematic illustration of the homoclinic tangle of the unstable saddle fixed point *U*. W^s and W^u are, respectively, the stable and unstable invariant sets (manifolds) of the saddle point. A transversal intersection of these manifolds (ξ_0) necessarily leads to an infinite number of subsequent intersections, resulting in chaotic behavior of the DS

zero. In particular, the crucial role of resonant tori in this process has been recognized, the transition starting with the increasing distortion of these tori. The resonant tori become ultimately corrugated on all scales, acquiring a fractal shape and allowing orbits to break-up from them and diffuse in the regions between the surviving non-resonant tori. Over 20 years after the final formulation of the KAM theorem, Boris Chirikov suggested a diagnostic criterion for the onset of chaos in Hamiltonian systems. He studied numerically the *standard map*

$$\begin{aligned} I_{j+1} &= I_j + K \sin \Theta_j \\ \Theta_{j+1} &= \Theta_j + I_j \,, \end{aligned} \tag{1}$$

a particular area preserving two-dimensional IM, where I and Θ are typically action-angle variables of a Hamiltonian system and K is a constant. He showed that fully-spread chaos arises when $K > 1$ and found that this happens because *resonances overlap*. This is schematically illustrated in Fig. 2.

The basic defining property of deterministic chaos is SIC (due to the divergence of initially arbitrarily close phase space trajectories of the DS) and it is quantified by the positivity of the largest *Liapunov exponent*, which guarantees chaotic behavior in all DS. In Hamiltonian systems, additional subdivision of different degrees of chaotic motion is available. As the Hamiltonian DS is driven farther into the chaotic regime the motion becomes more strongly irregular, in the sense that the DS trajectories in phase space explore progressively larger volumes. All

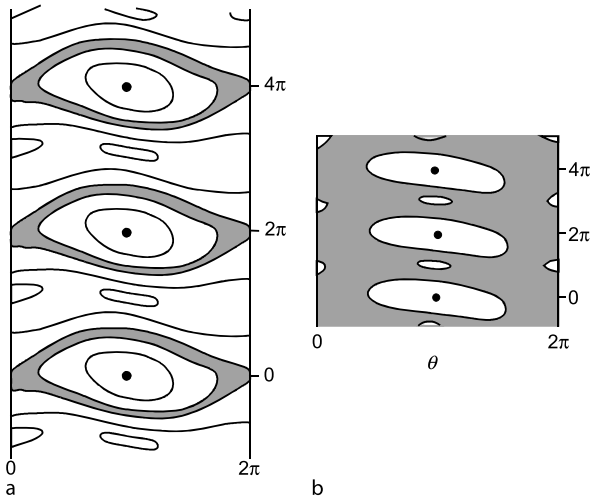

Astrophysics, Chaos and Complexity in, Figure 2
Non-resonant KAM tori and chaotic regions (*shaded*) in the standard map. Schematic drawing for and an almost integrable case ($K \lesssim 1$) (a) and the fully chaotic case ($K > 1$) (b). The broken up resonant tori, displayed in a which is expanded so as to show more details, overlap and merge, giving the situation shown in b – widespread chaos

Hamiltonian DS are *recurrent*, i. e., their trajectories return infinitely many times arbitrarily close to the initial point (according to the Poincaré recurrence theorem this is true for almost all orbits). A DS is called *ergodic* when long-time averages of a variable is equivalent to its phase-space average. Hamiltonian system are ergodic on non-resonant tori. If any initial phase-space volume eventually spreads over the whole space, then the chaotic system is said to have the property of *mixing*. Deeper into the chaotic regime a Hamiltonian system becomes a *K-system*. This happens when trajectories in a connected neighborhood diverge exponentially on average and finally, when every trajectory has a positive Liapunov exponent, there is global instability and the system is called a *C-system*.

The first Hamiltonian system, directly relevant to astrophysics, which was explicitly shown to exhibit deterministic chaos, was that studied by Hénon and Heiles [29]. They numerically investigated the orbits of a point mass (a star) in a model potential (approximating that of an axially symmetric galaxy). Casting the gravitational influence of all the other masses in a system into a mean field and computing the orbits of a point mass in it, is significantly easier than a full n-body integration and has always been one of the primary techniques of galactic dynamics. The question addressed in this study involved the possible existence of a third isolating integral, which would guarantee integrability in this case. The fact that the motion in the azimuthal (with respect to the symmetry axis) direction can be separated out and that total energy and angular momentum are two independent integrals of motion leaves out a three-dimensional phase volume that is accessible to the system. A third isolating (i. e., independent of the other two) integral would thus guarantee motion on a two-dimensional phase surface only, that is, integrability. Using a model axi-symmetric potential, whose non-dimensional form

$$V(x, y) = \frac{1}{2}\left(x^2 + y^2\right) + \left(x^2 y - \frac{y^3}{3}\right) , \qquad (2)$$

is clearly seen to contain a simple nonlinear part in addition to the harmonic one, Hénon and Heiles integrated the Hamilton equations of motion for several values of E, the non-dimensional energy parameter. By examining the Poincaré section in each case, they found that the system is integrable and the trajectory proceeds on tori (Fig. 3) as long as E is small enough. For a certain, larger value of E some of the tori are destroyed, with a chaotic "sea" appearing in between the surviving ones, but finally above some other, still higher, value of E, almost all tori are destroyed and chaos becomes widespread (Fig. 4). These results largely follow the scenario set forth by the KAM theorem. They also settled the issue: A third integral of motion of galactic dynamics does *not* exist in general. The implications of these findings have since then been thoroughly investigated (see also below), primarily by George Con-

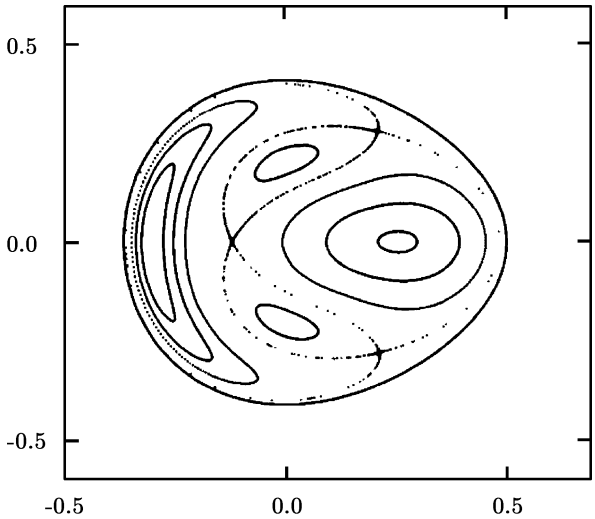

Astrophysics, Chaos and Complexity in, Figure 3
Surface of section in the x–y plane for the Hénon–Heiles system with a relatively low value of the energy parameter $E = 1/12$. The system is integrable and the motion proceeds on invariant tori

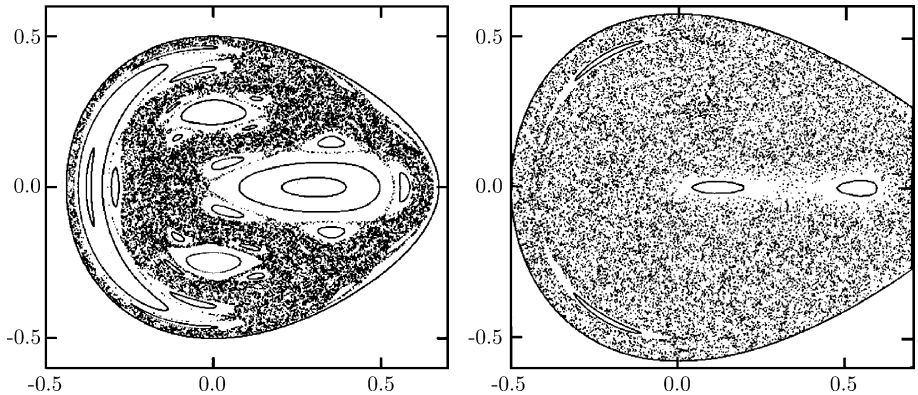

Astrophysics, Chaos and Complexity in, Figure 4
Same as Fig. 3, but for values $E = 1/8$ (*left panel*) and $E = 1/16$. Chaotic orbits that exist between the surviving tori eventually spread and fill almost all of the available space, with only limited islands of regular motions surviving

topoulos and his collaborators. Moreover, the pioneering work of Hénon and Heiles could perhaps not directly predict the properties of the stellar orbits and their distribution in some particular galaxy. It has, however, remained to this day a fundamental contribution to the theory of Hamiltonian DS.

The Hénon and Heiles system is actually a one-body problem (in a smooth external potential) and owes its chaotic behavior to the nonlinearity brought about by the anharmonic terms in the potential. Perhaps the simplest and most intriguing astrophysical realizations of the GNBP involve the motion of bodies in the Solar System. There exists today ample theoretical evidence, based largely on series of gradually improving numerical studies, that chaos permeates the Solar System and can be found essentially everywhere one looks. The numerical calculations have recently been supplemented by analytical and semi-analytical work.

The earliest definite finding (theoretical *and* observational) of chaotic motion among the permanent bodies in the Solar System was in the rotational dynamics of the small and rather inconspicuous Saturn's moon, Hyperion. In 1981 the Voyager 2 mission to Saturn discovered that Hyperion's orbit, residing well outside the Saturnian rings, is quite eccentric and the moon itself is unusually flat and its spin behavior is quite strange. It was found, for example, that Hyperion is spinning along neither its longest axis nor the shortest one; thus its spinning motion must be unstable. On the theoretical side, even though this problem is not a realization of the GNBP with $n > 2$, the inclusion of this moon's rotation around its axis may still allow for complicated dynamics in the system. The problem was analyzed by Wisdom, Peale and Mignard [63], who employed a known nonlinear equation for the time evolution of the angle $\theta(t)$ between the moon's principal axis, having the smallest moment of inertia, and a fixed direction in space. The equation,

$$\frac{1}{n^2}\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} + \frac{\eta^2}{2}\left(\frac{a}{r}\right)^3 \sin\left[2(\theta-\phi)\right] = 0\,, \tag{3}$$

includes the orbital parameters of the moon $a(t), \phi(t)$ and the orbital mean motion frequency $n \equiv \sqrt{GM_S/a^3}$, assumed to be given, and also the "asphericity" parameter of the moon's figure, $\eta \equiv \sqrt{3(I_B - I_A)/I_C}$, with $I_C \geq I_B > I_A$ being the principal moments of inertia of the moon.

Equation 3 can be perceived as describing a nonlinear oscillator subject to periodic parametric forcing, a non-integrable Hamiltonian system, and thus it could not be analytically solved, save for special cases that allow for perturbative treatment. Wisdom et al. integrated the equation numerically and examined the appropriate Poincaré section for various solutions. Re-scaling the time variable (formally $t \to nt$, so that the orbital period becomes 2π), one can easily see that the equation of motion results from a time-dependent Hamiltonian

$$H(x,p,t) = \tfrac{1}{2}p^2 + \tfrac{1}{2}\eta^2\alpha^2(t)\cos\{[2x-\phi(t)]\}\,, \tag{4}$$

where $x \equiv \theta$, $p \equiv \dot{\theta}$ and $\alpha(t) \equiv 2^{-1/2}[a/r(t)]^{3/2}$. Now, choosing orbital parameters appropriate for Hyperion's orbit Wisdom et al. found that even for the relatively small value of $\eta = 0.2$ (Hyperion actually has $\eta = 0.89$) narrow disconnected chaotic layers appear in the $p - x$ plane of section and they reside close to the resonances $p = 1/2, 1, 3/2$. For $\eta = 0.89$ there appears an extended chaotic zone. These characteristics of the surface of the section are reminiscent of the transition to chaos in the generic case depicted in Fig. 2, where the resonance over-

lap criterion is illustrated. The direction of Hyperion's axis of rotation is, thus, bound to vary chaotically and calculations show that significant changes can occur during just a few orbital times.

Chaotic tumbling of the kind found for Hyperion cannot be expected for planets orbiting the Sun, because for them $I_A \approx I_B$. Complex behavior in the rotational dynamics may arise, however, from the perturbations of the orbit caused by the presence of other planets. Laskar and Robutel [35] explored this possibility for all the inner Solar System planets and found that the obliquity of Mars is currently varying chaotically between 0° and $\sim 60°$ on a time-scale of a few million years, and it has probably always been varying in this manner. The rotational dynamics of Mercury and Venus do not appear to be in the chaotic regime, but it is reasonable that these planets may have experienced, in their past, slow chaotic changes in the obliquity of the rotational axis. The Earth's case is quite intriguing because this analysis indicates that it should be now in the state of chaotic variation of its obliquity between 0° and $\sim 80°$ with a time-scale of a few million years. Laskar, Joutel and Robutel [36] repeated the calculations for the Earth, including also the effects of the Moon, and obtained the comforting result that the Moon's torque appears to stabilize the Earth's rotational state, with the obliquity varying in a regular manner and by no more than $\pm 1.3°$ about the mean value of 23.3°.

Just as the periodic orbital motion in the two-body problem may induce complex rotational dynamics in one of the bodies, tidal interaction due to the bodies' finite size may affect the orbit in a manner that is far from being simple. Mardling [41] investigated this possibility in the context of binary stars, formulating the problem by means of a Hamiltonian system including stellar, tidally induced, oscillations. She found that for some specific initial orbital parameters the radial separation and eccentricity may experience chaotic variations, giving rise to the possibility that an initially unbound orbit may change so as to give rise to a stable bound system (a binary). In subsequent studies the effect of such tidal captures on an evolution of a start cluster was investigated. In addition, Mardling noticed and pointed out the formal similarity of the tidal-capture problem to the three-body problem.

Since the early work of Daniel Kirkwood, it has been known that the space distribution of asteroids (also known as *minor planets*) is not uniform and, in particular, displays prominent gaps, which, as already Kirkwood himself had recognized, appear when the their period would be equal to a simple fraction of Jupiter's period (see Fig. 5). Early studies, based on the restricted three-body problem (an asteroid moving in the field of the Sun and Jupiter, orbiting around each other in a Keplerian circle), had employed perturbative averaging techniques to the motion of the large bodies and failed to predict the Kirkwood gaps. Wisdom [62] noticed that these studies could not account for possible spreading of asteroid orbits through the entire chaotic regions accessible to them in phase space, because the existing KAM tori prevent it. This is a purely dimensional effect and such spreading is possible and may proceed (albeit on a very long time-scale) even if the system is close to integrability and resonance overlap does not occur, but the dimension of the accessible phase space has to be higher by at least two than that of the tori. To allow for the possibility of *Arnold* diffusion (so named after its discoverer), Wisdom suggested abandoning the previous approach (giving rise to only four-dimensional phase space) and use, instead, the three-dimensional restricted *elliptic* problem. He proposed studying an equivalent discrete map (because full numerical integration over a long enough time was not feasible then), whose structure near one particular commensurability (the 3:1 ratio with Jupiter's motion) is similar to that of the full Hamiltonian flow. Test particles, initially placed near the 3:1 commensurability in low eccentricity orbits, were found to typically hover around the initial orbit for almost a million years, and then suffered a sudden large increase of eccentricity, which would put them in a Mars-crossing orbit. Thus an asteroid could not survive for a long time and be observed in an orbit close to the 3:1 resonance. It would ultimately be removed from it by an encounter with Mars. Subsequent full numerical integration confirmed this idea and provided an explanation for the formation of this prominent Kirkwood gap and, in principle, the other gaps as well. Another important finding of these studies was that even though the Liapunov time for this system is relatively short ($\sim 10^3$ years), the chaotic increase of eccentricity occurred only after a significantly longer time. This is typical for systems close to integrability, when chaotic trajectories can explore the available phase space only very slowly, in this sense, by means of Arnold diffusion. This situation is sometimes referred to as *stable chaos* (see also below).

The outer Solar System is abundant in minor bodies as well – asteroids and comets. Comets are thought to originate from icy planetesimals that formed at the outskirts of the primitive Solar nebula. Some of them found stable orbits in the Kuiper belt, a relatively narrow and flat ring-like region starting just outside the orbit of Neptune and ending ~ 50 AU from the Sun. Other comets were probably scattered by encounters with nearby stars into a vast spherical shell-like reservoir, ranging between approximately 30,000 AU and 50,000 AU from the Sun – the

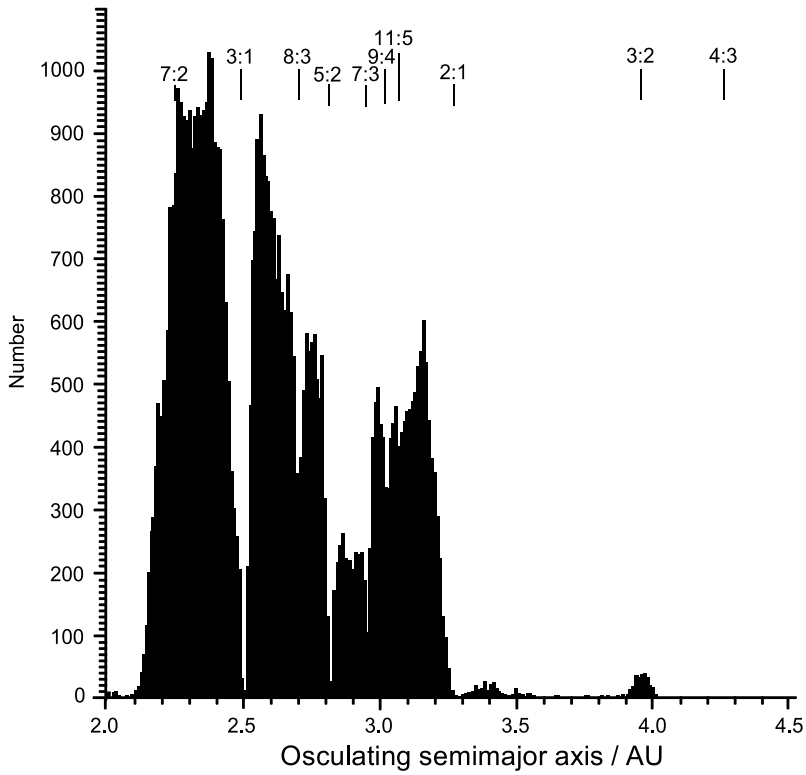

Astrophysics, Chaos and Complexity in, Figure 5
The number of asteroids as a function of the size (in AU). The location of the Kirkwood gaps, marked by the corresponding resonances with Jupiter motion, is clearly apparent. Reproduced, with permission, from [37], © Annual Reviews

Oort cloud. This topic is still quite controversial nowadays, but extensive numerical calculations suggest that, in any case, chaos seems to be playing a major role in cometary dynamics (cf. [19]). Most comets coming from the Oort cloud (long-period comets) are thought to have chaotic orbits, the origin of chaos being attributed in this case to repeated close encounters with massive planets. Sagdeev and Zaslavsky [52] showed that encounters between Jupiter and a comet on an almost parabolic orbit can account for a large chaotic zone, extending into the Oort cloud. More specifically, attempts to numerically calculate the orbit of the famous Halley comet, close to the time of its latest visit in 1985, failed to agree with each other and with observations. This problem may have been caused by the fact that Halley's orbit is in fact chaotic. Chirikov and Vecheslavov [15]) supported this suggestion and estimated that the timescale for practical unpredictability in this case could be as short as ~ 30 years. Regarding the short-period comets, numerical studies have indicated that there are practically no stable orbits among the outer planets. Such orbits have, however, been found above ~ 40 AU from the Sun and it is thus reasonable that planetesimals indeed could survive for very long times in the Kuiper belt. Close to the stable regions there exist also regions of instability, from which comets could be injected into the inner Solar System by means of Arnold diffusion and ultimately (but probably only temporarily) be captured into resonant orbits, by interaction with other planets, and appear as a short-period comets.

The most difficult dynamical problem in the Solar System results from the inclusion of all (or, at least, a few massive) planets and the Sun in a full GNBP. Since the late 1980s a number of numerical integrations, employing several different approaches, have been performed on this problem. The results of these studies were generally in rather remarkable agreement, although different techniques had been used to allow for long integration times. Positive Liapunov exponents, indicating chaos, have been found for the orbits of most planets and the Liapunov times t_Λ have been estimated, as follows. Sussman and Wisdom [57] found $10 \lesssim t_\Lambda \lesssim 20$ Myr for Pluto, $t_\Lambda \sim 4$ Myr for the inner planets and $3 \lesssim t_\Lambda \lesssim 30$ Myr for

the outer ones. Murray and Holman [46] refined the result for the Jovian planets, and determined it more accurately, at $t_\Lambda \sim 5\,\mathrm{Myr}$. Laskar [33] found $t_\Lambda \sim 4\,\mathrm{Myr}$ for the inner planets and, in addition, pointed out that despite this value, the effects of chaos on the orbits of these planets remain small for at least 5 Gyr (with the exception of Mercury, whose orbit was found to be able to change drastically its eccentricity within $\sim 3.5\,\mathrm{Gye}$). Similar conclusions have also been reached for the outer planets. It thus appears that the disparity between the Liapunov time and the time after which a dramatic change in the orbit may actually appear is similar for the planetary orbits to that of the asteroid orbits and the term "stable chaos" is appropriate to planetary orbits as well.

Stellar dynamics in clusters and galaxies is, in principle, a realization of the GNBP with large n. Direct numerical simulations of such problems are still, even nowadays, often requiring prohibitive computational resources. Historically, the approach to such problems had largely been based on statistical methods, inspired by the impressive success of statistical mechanics in handling systems composed of an enormous number of microscopic particles. Typical relaxation times, relevant to dynamical astronomy are, however, too long for stars to have settled into the invoked (and observed) statistical equilibria in clusters and galaxies. The understanding of the underlying mechanism of a *violent* relaxation process, that was proposed to solve this problem, calls for a better understanding of the n-body dynamics. To circumvent the full GNBP, studies of orbits in a given mean field potential were conducted, but the question of self-consistency (of the possibly evolving background potential) posed a nontrivial challenge to this approach.

Already the pioneering study of Hénon and Heiles [29] showed that chaotic orbits are expected even in simplistic steady potentials. In a number of more recent works it has been found that chaotic orbits in realistic galactic potentials tend to diffuse over the entire phase space available to them, in times shorter than Hubble time. For example, Hasan, Pfenniger and Norman [27] showed this for barred galaxies and Merritt and Fridman [42] for those having triaxial, strongly cusped mass distributions. In addition, Merritt and Valluri [43] found that random time-dependent perturbations in the mean potential are likely to enhance the rate of diffusion of the orbits. Habib, Kandrup and Mahon [26] investigated these type of effects in more detail by explicitly including weak friction and noise in the calculation of orbits in non-integrable two-dimensional potentials. Habib et al. found that even in two-dimensional, but strongly chaotic, potentials, the statistical properties of orbit ensembles (and indeed the structure of phase space itself) were altered dramatically (on a timescale much shorter than t_R) by the presence of friction and noise.

All these studies have essentially been modifications of the mean field dynamics, but it is reasonable to conclude from them and some similar calculations that the full GNBP can be expected to contain a significant amount of irregularities and exhibit evolution on a short (compared to t_R) timescale. Thus, quite a strong evolution of the phase space distribution is expected to occur already in time-independent (but non-integrable) potentials. The mechanism driving this evolution was identified to be associated with chaotic diffusion or *mixing*. The numerical experiments mentioned above suggest that mixing behavior indeed does occur in some realistic potentials. Moreover, it is significantly enhanced by cusps and irregularities (which have been modeled by noise and friction) in the potential.

What do these findings imply for galactic dynamics? It is quite clear that this question cannot be answered within the realm of the mean field approach. If the stochastic orbits are causing a slow (but not too slow) evolution of the halo, chaos probably plays a role in galactic evolution. Moreover, it is possible that the underlying potential is not only evolving in time, but also cannot be assumed to be smooth. One has to treat thus the full GNBP, which is clearly the fully self-consistent approach to the problem of cluster or galactic dynamics. While it is clear that the system is chaotic, it is still not clear into which category of chaotic Hamiltonian system it should be classified. Already Gurzadyan and Savvidy [25] have shown that the GNBP does not satisfy the requirements for it to be classified as a C-system. Subsequently, El-Zant [20] pointed out that conventional methods used for quantifying phase-space mixing are not adequate for the GNBP because of singularities (e. g. collisions), and proposed to use local geometric properties rather than global ones. He also addressed, in a more recent study, the influence of potential "softening", usually employed in n-body simulations to prevent singularities. The role of chaos in cluster and galactic dynamics and in dynamical astronomy is still not fully understood and some of the issues remain controversial. The subject is discussed in detail in the recent book by Contopoulos [16].

Chaotic Time Variability of Astronomical Sources

A variety of astronomical sources emit radiation whose intensity is intrinsically a time variable. This means that the variability is the object's inherent property and does not result from geometrical (e. g. due to the object's motion) or

other external (e. g. because of the Earth's atmosphere or other obscuring matter) effects. Such variability may provide a valuable addition to the usual fixed intensity and spectral information on the source.

Time-dependent signals that are measured in lab experiments and astronomical observations are usually recorded as digital *time series* data. Such data are also provided when a DS is numerically solved by a computer program. Time series analysis is, in general, a non-trivial process. In particular, this is so if the signal is contaminated by observational noise and the amount of data is limited, as is the case in a typical astronomical observation. Traditional techniques, like Fourier analysis, work well when the signal is periodic or multi-periodic, but they fail if it is irregular. Irregular variability may result from a random process, but it can also be due to deterministic chaos.

The questions then arise how to distinguish between a deterministically chaotic signal and a random one, and what is the relevant information that can be extracted from the former. Theoretical and practical issues related to this problem have indeed been addressed in parallel to the developments in chaos theory and have yielded significant understanding and a number of effective algorithms (see Abarbanel et al. [1]). The central idea, based on the pioneering work of David Ruelle, is based on a reconstruction of the essential dynamics of the DS which produces the signal, by embedding it in a multidimensional (pseudo) state space. The dimension of the set on which this dynamics takes place can then be found, using an algorithm of the kind that was originally proposed by Grassberger and Procaccia [24]. Low-dimensional fractal dimension is indicative of deterministic chaos.

A number of attempts have been made to discover chaos in aperiodic time series arising from photometric observation of several types of astronomical objects. While quite reliable results have been achieved for some pulsating stars, the situation has been less satisfactory for accreting systems (X-ray binaries and active galactic nuclei). Norris and Matilsky [47] pointed out some of the pitfalls of a straightforward application of the Grassberger–Procaccia method on the reconstructed state space. They and Scargle [53], in a more general work, have reached the conclusion that an adequate predictive analysis (like a model or mapping) is most probably indispensable in the studies of irregular astronomical series which suffer from an insufficient amount of data and a relatively poor signal-to-noise ratio. Further general caveats and in particular those relevant to studies of active galactic nuclei variability were given by Vio et al. [60].

The most convincing evidence of low-dimensional dynamics and chaos in an astronomical signal was given by Buchler et al. [11]. They proposed and used an appropriate discrete global mapping in the reconstructed pseudo state-space, which allowed them to use enough data points to faithfully determine the dimension of the attractor present in the equivalent dissipative DS, arising from the properly reduced data set. The aperiodic pulsating star R Scuti, whose light-curve had been followed for over 30 years, mainly by amateur observers (see Fig. 6), was analyzed first, giving a positive Liapunov exponent and a low fractal dimension ($D = 3.1$) of the attractor. In a subsequent work Buchler, Kolláth and Cadmus [12] presented analyzes of the light curves of several additional large-amplitude, irregularly pulsating stars (astronomically classified as *semi-regular*). They eliminated multi-periodicity and stochasticity as possible causes of the irregular behavior, leaving low-dimensional chaos as the only viable alternative. They then used their reconstruction technique in an attempt to extract quantitative information from the light curves and to uncover common physical features in this class of irregular variable stars. Similarly to the case of R Sct discussed above, it was found that the dynamics of at least three stars (out of the four examined) takes place in a four-dimensional dynamical state space, suggesting that two vibrational modes are involved in the pulsation. The physical nature of the irregular light curves was identified as resulting from a resonant mechanism with the first overtone mode, having a frequency close to twice that of the fundamental one. If the fundamental mode is self-excited while the resonant one is by itself stable, chaotic behavior can be expected on the basis of what is known as the Shilnikov scenario.

Convincing detection of chaotic behavior in signals from variable accreting systems has not, as yet, been definitely found. Contamination by observational, and possible also inherent, stochastic noise and an insufficient amount of data constitute the main challenges in this case.

On the theoretical side, some simplistic and idiosyncratic models, as well as quite sophisticated methodical approaches, have been proposed to account for chaotic behavior in time-variable astrophysical systems. Buchler and Regev [10] devised a one-zone pulsator stellar model, in which ionization and radiative-transfer were incorporated. Such simplistic models had been quite effective before in elucidating the mechanisms responsible for pulsational instability in stars. The model was of an extended ionization zone atop an inert stellar core, the oscillation arising from possible dynamical instability due to partial ionization. At the same time the zone was driven thermally by the radiation diffusing through the envelope, from the core outward. When the dynamical time was significantly different from the thermal time, the model exhibited pe-

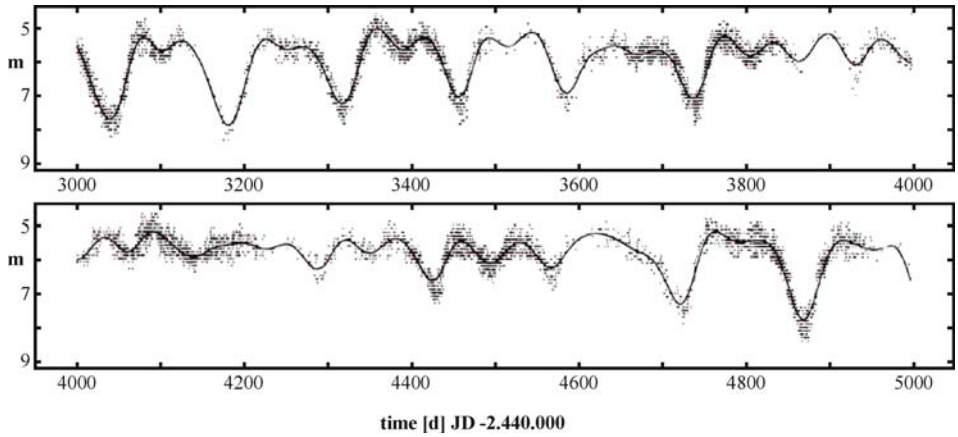

Astrophysics, Chaos and Complexity in, Figure 6
Typical observed light-curve segments of the irregularly pulsating star R Scuti. The raw data points are indicated by the *dots* and the smoothed filtered signal by the *continuous line*. Reproduced, with permission, from [11]

riodic oscillations. However if these two time scales were set to be comparable, Buchler and Regev found chaotic behavior. The mathematical structure of this model was quite similar to the simplistic model of thermal convection, in which Moore and Spiegel [44] had found aperiodic variability, proposing also (together with Baker, whose one-zone model was instrumental in understanding pulsational instability in general) that a similar model can be used to get irregular stellar pulsation. In Fig. 7 a projection of the system's trajectory in its three-dimensional state-space on the R–$\dot{R}$ (outer radius–outer velocity) plane is shown for four cases. In cases (b) and (c) there is chaotic behavior. The layered structure in those cases is reminiscent of chaotic attractors in typical DS. Using some standard criteria, Buchler and Regev have shown that this variability has indeed the properties of deterministic chaos. Other models of this kind have also discovered chaotic behavior, resulting from other mechanisms – e. g. a nonlinear non-adiabatic oscillator, but without dynamical instability and relaxation oscillations; a dynamically driven (by the inner core) Hamiltonian oscillator, etc. It is difficult to decide which of the above mechanisms, if any, is relevant to a realistic stellar model or indeed a pulsating star. Detailed numerical modeling of pulsating stars have, however, also shown transition to chaotic behavior via a sequence of period-doubling bifurcations, when a parameter (in this case the effective temperature) was varied. It has been suggested by Buchler and his collaborators that the transition to chaos is associated, in this case, with the excitation of vibrational overtones, brought about by resonances.

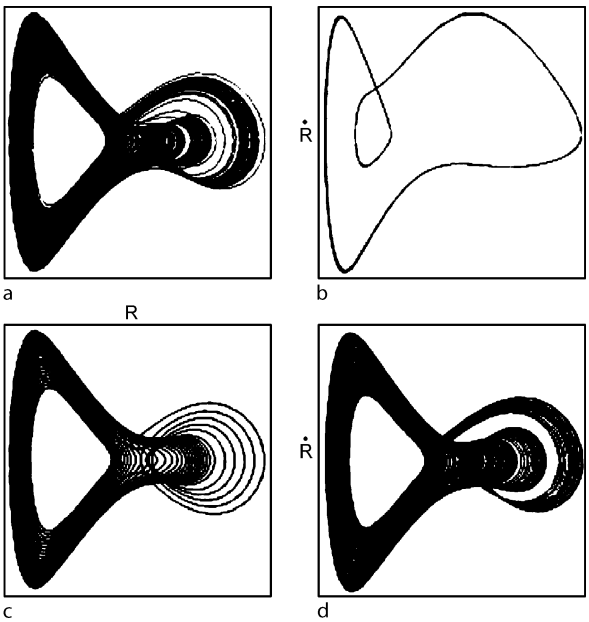

Astrophysics, Chaos and Complexity in, Figure 7
Projections of the attractor on the $R - \dot{R}$ plane in the model of Buchler and Regev [10]. In cases a and d the oscillations are periodic while in b and c there is chaos. Reproduced, with permission, from [10]

A systematic approach to stellar pulsation, based on DS theory has also been put forward. It employs the idea of dimensional reduction (using either perturbative analysis or a projection technique). The use of such approaches is quite widespread, in particular when the original DS is

formulated in terms of rather complicated PDE, as is the case in stellar pulsation theory. The idea is to capture the essentials of the dynamics in the form of a simpler mathematical system like a model PDE or a set of ODE – *amplitude equations*, or even IM. This is usually feasible close to the instability threshold, where the existence of a relatively low-dimensional *center manifold* to which the dynamics of the dissipative system is attracted, is guaranteed. Spiegel [56] and Buchler [8] discussed the possible techniques which can be applicable to stellar pulsation and Goupil and Buchler [23] summarized their work on amplitude equations, by giving their formal form for a rather general (non-radial, non-adiabatic) case. By studying the relevant amplitude equations in the various cases, it is possible to identify the nature of the possible mode couplings that are expected to occur in the nonlinear regime. In particular, the excitation of multi-mode pulsation and the effect of resonances can be investigated and then applied to phenomenological findings on pulsating stars (cf. Buchler [9]). For example, the amplitude equations, derived in this manner for a case in which two, almost resonant (having frequencies (Ω_1 and $\Omega_2 \approx 2\Omega_1$) radial modes are taken into account are

$$\begin{aligned} \frac{\mathrm{d}A_1}{\mathrm{d}t} &= \kappa_1 A_1 + i\frac{\Omega\Delta}{2}A_1 + P_1 A_1^* A_2 \\ \frac{\mathrm{d}A_2}{\mathrm{d}t} &= \kappa_2 A_2 - i\Omega\Delta A_2 + P_2 A_2^2 \,, \end{aligned} \tag{5}$$

where A_i are the complex amplitudes of the modes, κ_i their linear growth-rates, $\Omega \equiv (2\Omega_1 + \Omega_2)/2$, $\Delta \equiv (2\Omega_1 + \Omega_2)/(2\Omega)$ and P_i are dependent on the specific model and can be derived from its structure. These equations capture the essential *slow* dynamics for the weakly nonlinear case.

Amplitude equations derived by perturbative techniques based on the separation of time-scales (as above) are not adequate for the case in which chaotic pulsation occurs with the dynamical and thermal timescales being comparable. However, dimensional reduction is still possible in this case, but more sophisticated approaches, e. g., the one suggested by Spiegel [56], are called for.

In the case of variable accreting sources several simplistic models, based on DS that can give rise to complex time-variability, were also proposed. Celnikier [13] pointed out that the essential features of the, so called, rapid X-ray burster, can be reproduced by a deterministic simple nonlinear equation, of the kind used in population dynamics studies. The IM employed was similar to the one used by Feingenbaum in his pioneering study, which demonstrated the universal period-doubling bifurcation route to chaos. The rapid burster belongs to a class of variable X-ray sources, thought to originate in a close binary system in which a normal star is transferring mass onto its companion neutron star, giving rise to thermonuclear flashes on the latter. A simplistic two-zone model that captures some main characteristics of regular X-ray bursters has been proposed by Regev and Livio [51] and such models are still popular nowadays (e. g. [17]). For the rapid burster Livio and Regev [39] extended the model so as to include a third degree of freedom, thus providing the possibility of positive and negative feedback via mass accretion, and obtained a chaotic series of bursts, whose characteristics were very similar to the rapid burster.

Accreting close binary systems are known to exhibit irregularities in their light-curves in a number of spectral windows. Among them are quasi-periodic oscillations and short-time flickering. Young and Scargle [66] proposed a model, in this context, based on a coupled map lattice – an IM in which the value of a variable is dependent on the values of this variable in the previous iteration step, not only at the same location (lattice point), but also on this value at adjacent spatial locations. For some range of the parameters, quasi-periodic oscillations similar to the ones observed in several accreting binaries, were found. Young and Scargle found that persistent positive value of the largest Liapunov exponent cannot be guaranteed in this case, however chaotic behavior can be guaranteed during long transients. Such behavior was given the name *transient chaos*. Another type of a DS, a *cellular automaton* (essentially a coupled map lattice, but also with the dynamical variable allowed to take discrete values only) was proposed by Yonehara, Mineshige and Welsh [65] to model an accretion flow through a disk in close binary systems. The DS they invoked was in general found to evolve towards a characteristic state, whose qualitative behavior – mass accumulation during an extended time interval followed by rapid avalanches – resembled other well-studied DS. In most cases the avalanches are localized and quite small and terminate rapidly, but occasionally a local avalanche may trigger some very prominent events. This generic behavior has been found to occur in a variety of open dissipative systems and the state in which it occurs was called by Bak, Tang and Weisenfeld [4] *self-organized criticality*. Such states generally do not have positive Liapunov exponents and are thus not truly chaotic, but they clearly exhibit scaling behavior. The most famous of these systems is the sand pile model, reminding some perhaps also of the behavior of stock markets. In Fig. 8 the flickering light-curves of a *cataclysmic variable* in two different wavelengths are shown along with the power spectral density (PSD) as calculated by Yonehara et al. for one particular choice of the parameters.

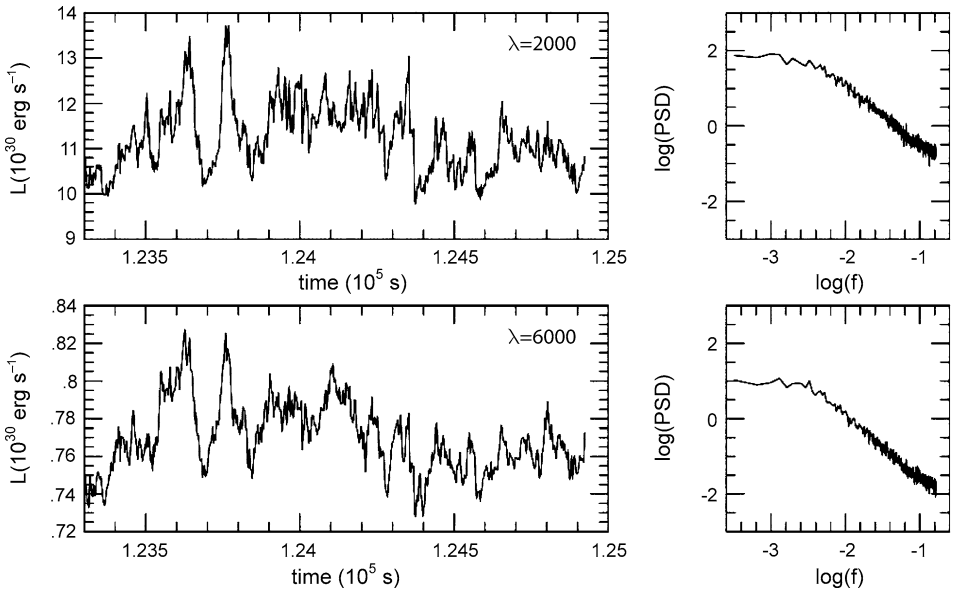

Astrophysics, Chaos and Complexity in, Figure 8
Light-curves and power spectral densities produced by the cellular automaton model of Yonehara et al. [65] for the flickering of Cataclysmic Variables. The *upper panels* correspond to 2000 Å emission, while the *lower ones* to emission at 6000 Å. Reproduced, with permission, from [65]

Spatio-Temporal Patterns and Complexity in Extended Systems

Extended astrophysical systems, that is, those that can be spatially resolved in observations often display very complex structures. The distribution of scattered matter in galaxies, the interstellar medium (ISM), is far from being uniform. Observations in a variety of wavelengths indicate that clouds of different densities are common and exhibit very complicated spatial forms (see Fig. 9 for one example). Observable phenomena in more compact objects than interstellar clouds, e. g. accretion disks of different types and stars (which are usually not spatially resolvable in direct observation), often require consideration of spatial processes as well. The spatial distribution of the stars themselves (seen as point sources) appears in the observations to be highly non-uniform. Hierarchical clustering, from star clusters of varying richness through galaxies and galactic clusters and up to the largest super-clusters of galaxies, is apparent on a variety of spatial scales, up to that of the entire visible universe. It is of interest to quantify the observational data and identify the physical processed responsible for shaping the various structures.

The time evolution of spatially extended astrophysical objects is naturally modeled by PDE. These DS are often very complex and include the equations of (magneto)fluid dynamics and also of thermal and radiative processes. Attempts to actually solve such nonlinear PDE (usually referred to as *microscopic equations*) can only be made by using numerical simulation. Despite the recent significant progress in computer power and visualization methods, such numerical calculations are often still inadequate, in particular because the Reynolds numbers of typical astrophysical flows are so high, such that a faithful resolution of the whole dynamical range is impossible. Modern approaches to spatio-temporal complexity utilize mathematical techniques to reduce the usually formidable mi-

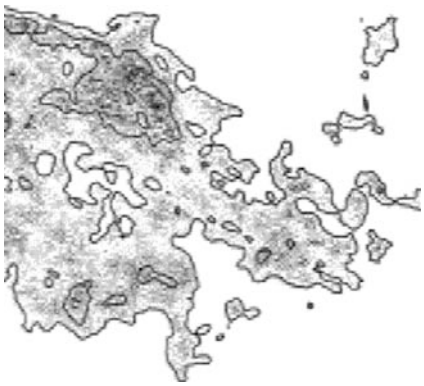

Astrophysics, Chaos and Complexity in, Figure 9
Contours of infra-red radiation intensity in some clouds observed by IRAS (detail). Adapted, by permission, from [6]

croscopic PDE to significantly simpler and analytically (or semi-analytically) tractable DS (simple generic *model* PDE, ODE and IM). These approaches constitute what is now called *theory of patterns and complexity* and have so far been successfully employed in the theoretical study of laboratory and atmospheric fluid dynamical phenomena, chemical reactor systems, as well as in other diverse applications (see Cross and Hohenberg [18] for a review).

Physical processes that are instrumental in shaping the ISM complexity undoubtedly include fluid turbulence, which unfortunately still defies adequate understanding. Dynamical system approaches to turbulence have yielded some success, but only for rather limited simple issues, like transition (e. g. Bohr et al. [7]). In an effort to identify simpler physical mechanisms inducing ISM complexity, Elphick, Regev and Spiegel [21] began to examine the effects that thermal bi-stability (alone) may have on an extended system like the ISM. They used a one-dimensional simplified version for the relevant non-dimensional microscopic equation,

$$\partial_t Z = F(Z; p) + Z^{\beta} \partial_x^2 Z \,, \tag{6}$$

where the dependent variable $Z(x, t)$ is a positive power of the temperature, β is determined from the temperature dependence of the heat-diffusion coefficient and $F(Z; p)$ is the relevant cooling function with the pressure p considered to be a fixed parameter. Approximating $F(Z; p)$ by a suitable polynomial, they were able to reduce the problem to a set of ODE for the motion of fronts (between the two stable phases). In pattern theory this kind of approach is called *defect* dynamics. Elphick et al. found that an initially random distribution of front/anti-front pairs (which can be regarded as "clouds") develops by front – anti-front annihilation and gives rise to an inverse cascade of larger and larger clouds, on the way to thermal phase separation. This process is similar to what is known in condensed matter physics as spinodal decomposition or roughening. When the simplest spatial forcing (as it should exist in the ISM, due to the non-uniform distribution of the heating sources) was introduced, possible complex steady states were found. Such states can be described by the following pattern map

$$\begin{aligned} I_{j+1} &= I_j - K \sin \Theta_j \\ \Theta_{j+1} &= \Theta_j + L \log I_j \,, \end{aligned} \tag{7}$$

where Θ and I are simple functions of the front position and separation (cloud size), respectively. The constants K and L are dependent on the system specifics. This IM bears resemblance to the standard map (Eq. 1) and similarly to the latter is chaotic. This means that thermal instability (and some forcing) alone can produce spatial complexity.

In subsequent studies Aharonson, Regev and Shaviv [3] used a multi-dimensional extension of Eq. 6 and solved it numerically in two dimensions, starting from a random distribution. It is well known in pattern theory that front curvature is driving evolution of structures in systems governed by reaction-diffusion equations, like Eq. 6, and indeed the system advanced relatively fast towards phase-separation. However, the inclusion of simple and quite slight spatio-temporal forcing atop the cooling function resulted in a final locked steady pattern with complex cloud boundaries (see the left panels of Fig. 10). Shaviv and Regev [54] included also fluid motion to account for mass conservation through the moving fronts, but this did not change the result qualitatively (the right panels of Fig. 10). Calculation of the cloud boundaries' fractal dimension in the final state of the above numerical simulations gave a result that was similar to the one found in observations by Bazel and Désert [6]; Falgarone, Philips and Walker [22] and others. These studies analyzed observational data in a number of wavelengths and reported fractal dimensions $1.2 \lesssim D \lesssim 1.5$ for the various cloud boundaries. These observational claims have, however, remained quite controversial and several researchers have claimed that fractal structures can not be actually faithfully inferred from such observations. Still, the fact that thermal instability and diffusive spatial coupling alone can give rise to complicated cloud forms may be important in ongoing efforts to understand ISM complexity.

Ideas from pattern theory can also be used for astrophysical systems governed by microscopic equations that are considerably more complicated than those describing just a thermally bi-stable diffusive medium. Among the first paradigms of deterministic chaos were the dissipative ODE systems of Lorentz, Moore and Spiegel, derived by some drastic simplifications of the equations of fluid dynamics in order to model thermal convection. More judicious reduction procedures have been developed since then and successfully applied to different spatio-temporal systems. In addition to the mathematical techniques that allow one to extract the dynamics of coherent structures (defects) in spatio-temporal systems, when such structures do indeed exist, there are a number of other methods for the reduction of the original PDE system. An approach of this kind, which works well for a system close to its instability threshold, was recently applied, by Umurhan, Menou and Regev [58], to the magneto-rotational instability which is thought to be of primary importance in accretion disks. Employing a perturbative asymptotic analysis, valid close to the instability threshold

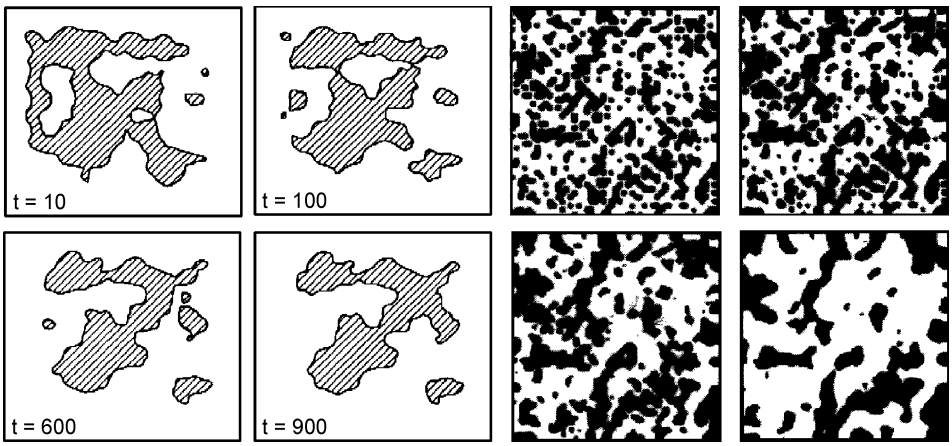

Astrophysics, Chaos and Complexity in, Figure 10
Evolution of cloud patterns in a reaction-diffusion system of the kind (6) but in 2D. The *left panels* show the purely thermal case and the *right one* display the effect of the inclusion of fluid motion. The *rightmost low picture* shows the ultimate state that persists due to forcing. Reproduced, with permission from [3] and [54]

and for a weakly nonlinear case, Umurhan et al. reduced the full microscopic magneto-hydrodynamical equations to a well-known, generic PDE for the spatio-temporal amplitude of a disturbance – the real Ginzburg–Landau equation. Using then-known solutions of this equation, one could estimate various important flow properties as the instability develops and nonlinearly saturates.

The dynamo problem (i. e., the creation of magnetic fields in flows of ionized fluids) is of obvious astrophysical interest and it too can be approached, in addition to numerical calculations, by methods of nonlinear DS theory. The work of Balmforth et al. [5], in which the advection of vector fields by chaotic flows was investigated, is an example of such an approach. Childress and Gilbert [14] summarize a number of ideas along these lines, that can be fruitful for future research on the dynamo problem.

The nature of the distribution of stars in the universe, with its hierarchial clustering, has raised some important questions. Virtually all the viable cosmological models are derived from Einstein's general theory of relativity and are based on the *cosmological principle*, which asserts that the universe is homogeneous and isotropic. However, direct observations of *luminous* matter have always revealed a substantial degree of clumpiness. Straightforward examination of the pictures reveals galaxy groups, clusters, super-clusters, voids, filaments and walls (see Fig. 11). Hierarchial clustering is typical of fractal sets and thus the question of whether the large scale distribution of matter in the universe is fractal or homogeneous has naturally been raised. Wu, Lahav and Rees [64] have recently compiled observational surveys of the galaxy distribution and concluded that relevant statistical measures, which enable the determination of dimension, point to the fact that the luminous matter distribution's dimension is scale dependent. On relatively small scales, less than about 50 h^{-1} Mpc (h is the Hubble constant in units of 100 km/s/Mpc), the distribution is indeed undoubtedly fractal. It is the smoothness of the galaxy distribution (or lack thereof) on the largest scales that is still the key point in the debate between the proponents of a "fractal universe" and the more traditional cosmologists.

The cosmic microwave background (CMB) radiation, for which years of data (since the launch of the COBE satellite) are now available, can provide a measure of the spatial fluctuations in the universe on the grandest scales – of the order of $\sim 1000\, h^{-1}$ Mpc. Fairly recent results indicate that the dimension of the mass distribution at these scales is not different from the value of three by more than a few tens of 10^{-6}, indicating that the universe is quite smooth on the largest scale. Various attempts to evaluate the mass distribution in the universe are based on the observation of different entities, from luminous matter (galaxies) and up to the CMB (the universe at a very early epoch, when galaxies were still absent). Some very recent optical surveys aim at an impressively deep coverage (up to $\sim$ 300–400 h^{-1} Mpc) but the grandest scales can be probed, at the present time, only by other means. Modern observations in radio, microwave, infra-red and X-ray bands also enabled a deeper look, well beyond the existing and upcoming deep optical surveys. Some recent radio surveys and X-ray background (XRB) data can probably probe scales of up to $\sim 500\, h^{-1}$ Mpc and thus bridge the

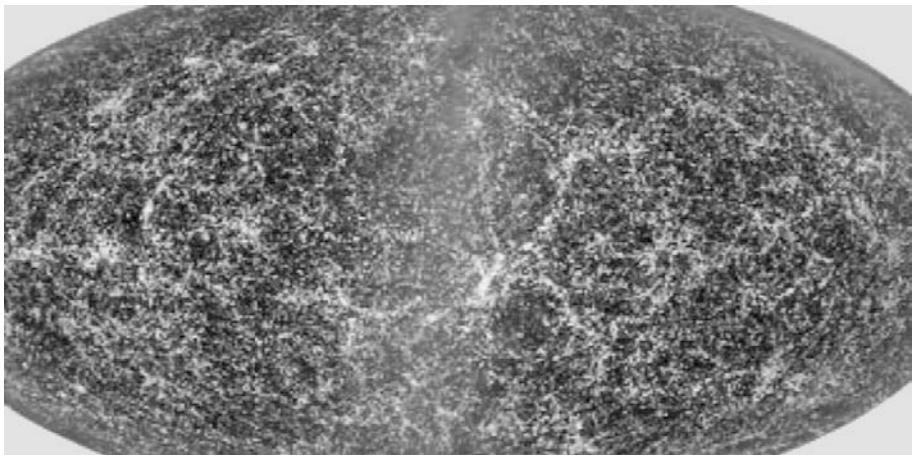

Astrophysics, Chaos and Complexity in, Figure 11
An image of close to one million of the brightest galaxies detected by the Two Micron All Sky Survey -2MASS (detail). The non-uniform distribution with clusters, voids and filaments, is clearly visible. The blurred strip is due to the Galactic plane [59].

gap between the optical surveys and the largest CMB scale. Other probing tools, like the Lyman α forest, may also be valuable in this respect.

The question of how (and at what scale) the transition from a fractal distribution to a smooth one occur, is still quite open. There seems even to be no agreement on the question of what the relevant statistical measures are to find the scale this transition. The recent suggestion by Pietronero, Gabrielli and Sylos Labini [48] to define it as the crossover scale between power-law behavior and exponential decay of the correlation function appears to be the most meaningful.

Theoretical efforts, aimed at finding and understanding the physical mechanism behind the above results on the spatial cosmic distributions, have also been made quite recently. A model for nonlinear gravitational clustering was proposed by Provenzale et al. [49], who introduced a simple, dynamically motivated family of multi-fractals (sets whose fractal dimension changes with scale), which are well suited for a bottom-up scenario of gravitational evolution. Murante et al. [45] suggested that a galaxy distribution whose dimension varies with scale may naturally result from the presence of a number of very strong local density maxima. They randomly superimposed a few tens of singularities and let them computationally generate point sets, whose dimension could be estimated as a function of scale. A result resembling real galaxy data, on small enough scales, was obtained for sufficiently strong singularities. Provenzale, Spiegel and Thieberger [50] suggested that additional statistical measures of the galaxy distribution may be useful for a meaningful characterization and perhaps deeper understanding of the scaling behavior at intermediate scales and found that lacunarity, a well-defined mathematical property found in multi-fractals and giving rise to oscillations around the mean value of their generalized dimensions, may have important astrophysical consequences.

Future Directions

Astronomy counted among the main disciplines in which the first studies on chaos and complexity arose, in both Hamiltonian and dissipative nonlinear DS. The most significant mathematical developments in the theory took place in the 20th century, mainly in its second half, and the efforts to extend it have continued until the present time. Detailed applications to particular astrophysical systems have, however, always been marred by the difficulty of obtaining a sufficient amount of high-quality data. Carefully controlled experiments, which are possible in the laboratory, are impossible when cosmic phenomena are merely passively observed.

The ongoing fast development of modern means for astronomical data acquisition – sophisticated telescopes, both earthbound and on dedicated satellites, increasingly powerful computers and new data reduction techniques – have recently opened new possibilities. On the theoretical side, intensive efforts are also being made to understand the behavior of the extremely complicated DS as those encountered in the study of astrophysical systems. Thus the prospects for fruitful and significant future developments appear to be very good.

Among the possible future directions, research on the following topics seems to be the most promising:

- Near the end of the 20th century, it was observationally confirmed that stars other than the Sun, some of them in binary systems, have orbiting objects around them that appear to be planets. In the last decade over 200 of such exo-planets have been discovered and there is growing evidence that chaos plays a role in both their formation as well as in the evolution of exo-planetary systems. Holman, Touma and Tremaine [30] studied such newly discovered exo-planets and concluded that their results imply that planetary orbits in binary stellar systems commonly experience periods of high eccentricity and dynamical chaos, and that such planets may occasionally collide with the primary star. Lissauer [38] pointed out the possible role of chaos in the formation of planets, in the Solar Systems and exo-planetary systems and Kiseleva-Eggleton et al. [32] found that within a few known exo-planetary systems at least one has chaotic orbits. The issue of planet migration, which acquired importance after the discovery of exo-planets, and the possible role of chaos in this process, has been addressed by Adams and Laughlin [2]. It appears, thus, that the inclusion of chaotic processes is indispensable in the important quest to understand the formation of planetary systems and their evolution. Such studies are expected to shed new light on the probability of finding exo-planets within "habitable" zones and to have implications for possible existence and survival of life in the universe.
- The most direct approach to the study of star cluster and galaxy dynamics is the one employing the GNBP with a large value of n. The direct problem can only be approached numerically and because chaos is bound to be present in this patently non-integrable Hamiltonian DS, it is important to assess its possible role in the simulations. This issue has not been fully resolved yet, but there are some recent findings which may elucidate outstanding questions and future directions in this context. For example, Kandrup and Sideris [31] addressed the difference between the behavior found in full n-body simulations and that of orbits in a smooth potential and Hayes [28] compared the results of n-body calculations with softened potential to the exact ones. Much remains to be done in this line of research, because the theoretical efforts towards understanding the formation of structure in the universe, on a large range of scales, heavily rely on n-body simulations.
  Recent work on galactic dynamics, employing self-consistent models, indicates that it is important to carefully examine the orbits, distinguishing between the regular and chaotic ones, and then try to understand what the role is that the chaotic orbits play in shaping the galaxy. The work of Voglis, Stavropoulos and Kalapotharakos [61], for example, yielded a surprising new result – that long living spiral arms are excited on the disk, composed almost completely of chaotic orbits. Other studies of this kind, led mainly by George Contopoulos and his collaborators and students, have already provided additional important results in this context. It is reasonable to expect that new understanding on the role of chaos in galactic dynamics as well as on ways to handle its influence on n-body simulations will emerge as a result of future studies. Such efforts may also contribute to new results in the theory of Hamiltonian DS of this sort.
- Astrophysical fluid dynamics is a broad subject, in which input from DS and pattern theory may provide significant progress in the future. Most existing studies of fluid systems in astrophysics have been based, in addition to linear stability analyzes, on very simplified models or brute-force numerical simulations. New knowledge and understanding of nonlinear processes that numerical simulations alone can not provide, can be expected, as it has been achieved in other fluid dynamical applications, if nonlinear DS and pattern theory is applied to astrophysical fluids as well. Ed Spiegel and his collaborators have advocated approaches of this kind for some time (e. g. Spiegel [55]) and interesting results have already been obtained. Selected examples are given in the book edited by Zahn and Zinn-Justin [67] and more recent such contributions have already found their way to the astrophysical and general fluid-dynamical literature. They mark a promising new direction in the theoretical studies of important, but not yet fully understood, astrophysical (magneto)fluid-dynamical processes as accretion, convection, explosions, jets and so forth.

Bibliography

Primary Literature

1. Abarbanel HDI, Brown R, Sidorovich JJ, Tsimring SS (1993) The analysis of observed chaotic data in physical systems. Rev Mod Phys 65:1331–1392
2. Adams FC, Laughlin G (2003) Migration and dynamical relaxation in crowded systems of giant planets. Icarus 163:290–306
3. Aharonson V, Regev O, Shaviv N (1994) Pattern evolution in thermally bistable media. Astrophys J 426:621–628
4. Bak P, Tang C, Weisenfeld K (1987) Self-organized criticality – An explanation of 1/f noise. Phys Rev Lett 59:381–384
5. Balmforth NJ, Cvitanović P, Ierley GR, Spiegel EA, Vattay G (1993) Ann NY Acad Sci 706:148–154
6. Bazell D, Désert FX (1988) Fractal structure of interstellar cirrus. Astrophys J 333:312–319

7. Bohr T, Jensen MH, Paladin G, Vulpiani M (1998) Dynamical system approach to turbulence. Cambridge University Press, Cambridge
8. Buchler JR (1993) A dynamical systems approach to nonlinear stellar pulsations. Astrophys Space Sci 210:9–31
9. Buchler JR (1998) Nonlinear pulsations. ASP Conference Series 135:220–230
10. Buchler JR, Regev O (1982) Oscillations of an extended ionization region in a star. Astrophys J 263:312–319
11. Buchler JR, Serre T, Kolláth Z, Mattei J (1996) Nonlinear analysis of the light curve of the variable star R Scuti. Astrophys J 462:489–501
12. Buchler JR, Kolláth Z, Cadmus RR Jr (2004) Evidence for low-dimensional chaos in semiregular variable stars. Astrophys J 613:532–547
13. Celnikier LM (1977) A simple mathematical simulacrum of the X-ray burster phenomenon. Astron Astrophys 60:421–423
14. Childress S, Gilbert AD (1995) Stretch, twist and fold: The fast dynamo. Springer, New York
15. Chirikov RV, Vecheslavov VV (1989) Chaotic dynamics of Comet Halley. Astron Astrophys 221:146–154
16. Contopoulos G (2002) Order and chaos in dynamical astronomy. Springer, New York
17. Cooper RL, Narayan R (2006) A two-zone model for type I X-ray bursts on accreting neutron stars. Astrophys J 652:584–596
18. Cross MC, Hohenberg PC (1993) Pattern formation outside equilibrium. Rev Mod Phys 65:851–1112
19. Duncan MJ, Quinn T (1993) The long-term dynamical evolution of the Solar System. Ann Rev Astron Astrophys 31:265–295
20. El-Zant AA (1997) On the stability of motion of N-body systems: A geometric approach. Astron Astrophys 326:113–129
21. Elphick C, Regev O, Spiegel EA (1991) Complexity from thermal instability. Mon Not Roy Astr Soc 250:617–628
22. Falgarone E, Phillips TG, Walker CK (1991) The edges of molecular clouds – Fractal boundaries and density structure. Astrophys J 378:186–201
23. Goupil M-J, Buchler JR (1994) Amplitude equations for nonadiabatic nonradial pulsators. Astron Astrophys 291:481–499
24. Grassberger P, Procaccia I (1983) Measuring the strangeness of strange attractors. Physica 9D:189–208
25. Gurzadyan VG, Savvidy GK (1986) Collective relaxation of stellar systems. Astron Astrophys 160:203–210
26. Habib S, Kandrup HE, Mahon ME (1997) Chaos and noise in galactic potentials. Astrophys J 480:155–166
27. Hasan H, Pfeniger C, Norman C (1993) Galactic bars with central mass concentrations – Three-dimensional dynamics. Astrophys J 409:91–109
28. Hayes WB (2003) Shadowing high-dimensional Hamiltonian systems: The gravitational N-body problem. Phys Rev Lett 90(1–4):054104
29. Hénon M, Heiles C (1964) The applicability of the third integral of motion: Some numerical experiments. Astronomical J 69:73–79
30. Holman M, Touma J, Tremaine S (1995) Chaotic variations in the eccentricity of the planet orbiting 16 Cygni B. Nature 386:254–256
31. Kandrup HE, Sideris IV (2003) Smooth potential chaos and *N*-body simulations. Astrophys J 585:244–249
32. Kiseleva-Eggleton L, Bois L, Rambaux N, Dvorak R (2002) Global dynamics and stability limits for planetary systems around HD 12661, HD 38529, HD 37124, and HD 160691. Astrophys J 578:L145–L148
33. Laskar J (1994) Large-scale chaos in the Solar System. Astron Astrophys 64:115–162
34. Laskar J (1996) Large scale chaos and marginal stability in the Solar System. Celestial Mech Dynamical Astron Astrophys 287:L9–L12
35. Laskar J, Robutel P (1993) The chaotic obliquity of planets. Nature 361:608–611
36. Laskar J, Joutel F, Robutel P (1993) Stabilization of the earth's obliquity by the moon. Nature 361:615–617
37. Lecar M, Franklin FA, Holman MJ, Murray NJ (2001) Chaos in the Solar System. Ann Rev Astron Astrophys 39:581–631
38. Lissauer JJ (1999) Chaotic motion in the Solar System. Rev Mod Phys 71:835–845
39. Livio M, Regev O (1985) Xray burst sources: A model for chaotic behavior. Astron Astrophys J 148:133–137
40. Lorentz EN (1963) Deterministic nonperiodic flow. J Atmos Sci 20:130–141
41. Mardling RA (1995) The role of chaos in the circularization of tidal capture binaries – I. The chaos boundary. II. Long-time evolution. Astrophys J 450:722–731, 732–747
42. Merritt D, Fridman T (1996) Triaxial galaxies with cusps. Astrophys J 460:136–162
43. Merritt D, Valluri M (1996) Chaos and mixing in triaxial stellar systems. Astrophys J 471:82–105
44. Moore DW, Spiegel EA (1966) A thermally excited non-linear oscillator. Astrophys J 143:871–887
45. Murante G, Provenzale A, Spiegel EA, Thieberger R (1997) Density singularities and cosmic structures. Mon Not Roy Astr Soc 291:585–592
46. Murray N, Holman M (1999) The origin of chaos in the outer Solar System. Science 283:1877–1881
47. Norris N, Matilsky T (1989) Is Hercules X-1 a strange attractor? Astrophys J 346:912–918
48. Pietronero L, Gabrielli A, Sylos Labini F (2002) Statistical physics for cosmic structures. Physica A 306:395–401
49. Provenzale A, Galeotti P, Murante G, Villone B (1992) A simple multifractal model of nonlinear gravitational clustering. Astrophys J 401:455–460
50. Provenzale A, Spiegel EA, Thieberger R (1997) Cosmic lacunarity. Chaos 7:82–88
51. Regev O, Livio M (1984) Thermal cycles from a two-zone accreting model: X-ray bursts and shell flashes. Astron Astrophys 134:123–128
52. Sagdeev RZ, Zaslavsky GM (1987) Stochasticity in the Kepler problem and a model of possible dynamics of comets in the Oort cloud. Nuovo Cimento B 97:119–130
53. Scargle JD (1990) Studies in astronomical time series analysis IV – Modelling chaotic and random processes with linear filters. Astrophys J 359:469–482
54. Shaviv NJ, Regev O (1994) Interface dynamics and domain growth in thermally bistable fluids. Phys Rev E 50:2048–2056
55. Spiegel EA (1980) Fluid dynamical form of the linear and nonlinear Schrödinger equations. Physica D 1:236–240
56. Spiegel EA (1993) Patterns of aperiodic pulsation. Astrophys Space Sci 210:33–49
57. Sussman GJ, Wisdom J (1992) Chaotic evolution of the Solar System. Science 257:56–62
58. Umurhan M, Menou K, Regev O (2007) Weakly nonlinear anal-

ysis of the magnetorotational instability in a model channel flow. Phys Rev Lett 98(1–4):034501
59. University of Massachusetts and Infrared Processing and Analysis Center/California Institute of Technology (2003) Joint project: Two Micron All Sky Survey. http://www.ipac.caltech.edu/2mass/releases/allsky. Accessed 2008
60. Vio R, Cristiani S, Lessi O, Provenzale A (1992) Time series analysis in astronomy – An application to quasar variability studies. Astrophys J 391:518–530
61. Voglis N, Stavropoulos I, Kalapotharakos C (2006) Chaotic motion and spiral structure in self-consistent models of rotating galaxies. Mon Not Roy Astr Soc 372:901–922
62. Wisdom J (1982) The origin of the Kirkwood gaps – A mapping for asteroidal motion near the 3/1 commensurability. Astron J 87:577–593
63. Wisdom J, Peale SJ, Mignard F (1984) The chaotic rotation of Hyperion. Icarus 58:137–152
64. Wu KKS, Lahav O, Rees MJ (1999) The large-scale smoothness of the universe. Nature 397:225–230
65. Yonehara A, Mineshige S, Welsh WF (1997) Cellular-automaton model for flickering of cataclysmic variables. Astrophys J 486:388–396
66. Young K, Scargle JD (1996) The dripping handrail model: Transient chaos in accretion systems. Astrophys J 468:617–632
67. Zahn J-P, Zinn-Justin J (eds) (1993) Astrophysical fluid dynamics (Les-Houches – Session XLVII). North-Holland, Amsterdam

Books and Reviews

Barnsley M (1998) Fractals everywhere. Academic, San Diego
Heggie D, Hut P (2003) The gravitational million-body problem: A multidisciplinary approach to star cluster dynamics. Cambridge University Press, Cambridge
Manneville P (2004) Instability, chaos and turbulence. Imperial College Press, London
Pismen LM (2006) Patterns and interfaces in dissipative dynamics. Springer, Berlin
Regev O (2006) Chaos and complexity in astrophysics. Cambridge University Press, Cambridge
Shore SN (1992) An introduction to astrophysical fluid dynamics. Academic, San Diego
Sussman GJ, Wisdom J (2001) Structure and interpretation of classical mechanics. MIT Press, Cambridge

Astrophysics: Dynamical Systems

George Contopoulos
Academy of Athens, Research Center for Astronomy, Athens, Greece

Article Outline

Definition of the Subject
Introduction
Classification
Integrable Systems
Formal Integrals. KAM and Nekhoroshev Theory
Periodic Orbits
Transition from Order to Chaos
Dynamical Spectra
Application: Order and Chaos in Galaxies
Future Directions
Bibliography

Definition of the Subject

Dynamical systems is a broad subject, where research has been very active in recent years. It deals with systems that change in time according to particular laws. Examples of such systems appear in astronomy and astrophysics, in cosmology, in various branches of physics and chemistry, and in all kinds of other applications, like meteorology, geodynamics, electronics and biology. In the present article we deal with the mathematical theory of dynamical systems from the point of view of dynamical astronomy. This theory is generic in the sense that its mathematics can be applied to diverse problems of physics and related sciences, ranging from elementary particles to cosmology.

Introduction

The theory of dynamical systems has close relations with astronomy and astrophysics, in particular with dynamical astronomy. Dynamical astronomy followed two quite different traditions until the middle of the 20th century, namely celestial mechanics and statistical mechanics.

Celestial mechanics was the prototype of order. The use of perturbation theories in the solar system was quite effective for predicting the motion of the planets, satellites and comets. Despite the fact that the perturbation series were in most cases extremely long, their effectiveness was never doubted. One of the most painstaking developments in this field was the "Theory of the Motion of the Moon" of Delaunay [29] and his successors. Today the theory of the Moon has been extended to an unprecedented accuracy by computer algebra, giving the position of the Moon with an accuracy of a few centimeters [8,30].

Celestial mechanics had proved its effectiveness already by the 19th century with the discovery of the planet Neptune by Leverrier and Adams. Its subsequent development reached its culmination in the work of Poincaré: "Méthodes Nouvelles de la Mécanique Céleste" [66]. But at the same time it reached its limits. Poincaré proved that most of the series of celestial mechanics are divergent. Therefore their accuracy is limited. On the other hand, it was made clear that the most important problem of celestial mechanics, the N-body problem, cannot be solved analytically. The work of Poincaré contains the basic ideas of most of the recent theories of dynamical systems, in particular what we call today the "Theory of Chaos".

A completely different approach of the theory of dynamical systems was based on statistical mechanics. This approach dealt with the N-body problem, but from a quite different point of view. Instead of a dominant central body (the Sun) and many small bodies (planets), it dealt with N-bodies of equal masses (or of masses of the same order) with large N.

Statistical mechanics, developed by Boltzmann, Gibbs and others, dominated physics in general. In this approach the notion of individual orbits of particles became not only secondary, but even irrelevant. One might consider different types of orbits, but their statistical properties should be the same.

Nevertheless, a basic problem was present in the foundations of statistical mechanics. If the N-body systems are deterministic, how is it possible to derive the random properties of the ensembles of statistical mechanics? This question led to the development of a new theoretical approach that is called "ergodic" theory. If almost all orbits are ergodic, then the random behavior of the system can be proved, and not simply assumed. As an example, consider a room of gas of constant density. The constancy of the density may be considered as a probabilistic effect. But if the orbits of the molecules are ergodic, all particles stay on the average equal intervals of time in equal volumes inside this room. Thus, the constancy of density is due to the properties of the orbits of all particles.

In dynamical astronomy there was a complete dichotomy between celestial mechanics and stellar dynamics. Celestial mechanics continued its work with perturbation series, and no chaos at all, while the new branch of stellar dynamics (and galactic dynamics in particular) was completely influenced by statistical considerations, as in statistical mechanics. In particular, one dealt with distribution functions of stellar velocities that were well described by a velocity ellipsoid. But it was realized that the time required for the establishment of a statistical equilibrium was much longer than the age of the universe. Thus the stars had not enough time to settle into a statistical equilibrium. These problems led to a reconsideration of the foundations of galactic dynamics, namely to the study of the individual orbits of the stars. Such studies started only in the sixties. The first time when celestial mechanicians and galactic astronomers met at a common meeting was during the 1964 Thessaloniki Symposium on the "Theory of Orbits in the Solar System and in Stellar Systems" [16]. At that time a completely new tool was available in dynamical astronomy, namely the use of computers.

The first computer experiment with a N-body system was the so-called Fermi–Pasta–Ulam paradox, in 1955. These authors considered particles along a line attracted by linear forces, plus a small nonlinear force. The energy was initially given to a few modes. The energy of each mode cannot change if the forces are only linear. But because of the non linearities, the energies of all the modes of the system were expected to change continually, in a way consistent with a statistical mechanics prediction. Instead of that, the computer results have shown that the energy changes were almost periodic and no tendency to a final statistical situation ever appeared. This was a paradox, a great surprise for people grown up with statistical mechanics.

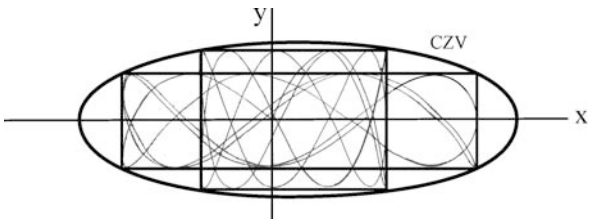

Astrophysics: Dynamical Systems, Figure 1
The first published orbits (1958) in the meridian plane of an axisymmetric galaxy inside the curve of zero velocity (CZV = equipotential)

A little later (1956), a computer was used by Per-Olof Lindblad to calculate orbits of stars in a plane galaxy, in order to find the formation of spiral arms [51].

At about the same time, we calculated [13] two orbits of stars in three dimensions. Based on the prevalent assumptions of statistical mechanics, we expected these orbits to be ergodic and fill all the space inside the energy surface. Instead, we found that the orbits did not fill all the available space, but filled only curvilinear parallelograms, like deformed Lissajous figures (Fig. 1). Later it was shown that such orbits can be explained, qualitatively and quantitatively, by a formal third integral of motion [14].

The theory of the third integral goes back to the work of Birkhoff [5] and Whittaker [77,78]. In particular Whittaker found the so-called adelphic integral, i. e. an integral similar to the energy in simple dynamical systems. In the case of two harmonic oscillators this integral is reduced to the energy of one oscillator only. In more general cases higher order terms have to be added to find a constant of motion. However, in resonant cases the form of the third integral may be quite different. For example, in the case of two equal frequencies it is a generalization of the angular momentum.

The series giving the third integral are in general divergent [70]. But even so in numerical applications the third integral is better conserved if it is truncated at higher and higher orders, up to a certain maximum order. The usefulness of these series was emphasized by Moser [61].

The most important result of these studies was that, in general, dynamical systems are neither integrable nor ergodic. This was a surprise, because it was generally assumed that generic dynamical systems are either integrable or ergodic [48]. This change of paradigm was emphasized by Lynden Bell [54].

The existence of ordered domains in generic dynamical systems is the basic content of the famous Kolmogorov–Arnold–Moser (KAM) theorem. Kolmogorov [47] announced a theorem proving the existence of invariant tori in dynamical systems. Such tori contain quasi-periodic motions with frequencies whose ratios are far from all rationals (see Sect. "Formal Integrals. KAM and Nekhoroshev Theory").

The details of the proof were given by Arnold [2,3] in the analytical case, and independently by Moser [59,60] in sufficiently differentiable cases (with 333 derivatives!). More recently the existence of such tori was proven for systems that are differentiable only a few times. Furthermore, invariant tori appear generically near stable periodic orbits. Such orbits appear even in systems with arbitrarily large perturbations. Thus complete lack of ordered regions containing invariant tori is rather exceptional.

Classification

There are two main types of dynamical systems:

(1) Maps (or mappings) of the form $\vec{x}_{n+1} = \vec{f}(\vec{x}_n)$, where $\vec{x}_n$ is a vector of N-dimensions and $\vec{f}$ is a set of N-functions.
(2) Systems of differential equations of the form $\dot{\vec{x}} = \vec{f}(\vec{x}, t)$, where $\vec{x}$ is a N-dimensional vector, and the dot denotes derivative with respect to a continuous time t.

Another separation of dynamical systems, both of maps and differential equations, is in conservative and dissipative systems. Conservative systems preserve the volume in phase space, while in dissipative systems the volume decreases on the average. If we reverse the time direction we have systems with increasing volume.

A large class of systems of differential equations are the Hamiltonian systems, with conjugate variables $\vec{x}\,(x_1, x_2, \ldots, x_N)$ and $\vec{y}\,(y_1, y_2, \ldots, y_N)$ that satisfy equations of the form:

$$\dot{\vec{x}} = \frac{\partial H}{\partial \vec{y}}\,, \quad \dot{\vec{y}} = -\frac{\partial H}{\partial \vec{x}} \tag{1}$$

where H is the Hamiltonian function $H = H(\vec{x}, \vec{y}, t)$ and N is the number of degrees of freedom. The equations of motion (1) are called canonical, or Hamiltonian equations. The space of the variables $\vec{x}$ and $\vec{y}$ is called phase space. The phase space of a map and of a system of differential equations is N-dimensional, while the phase space of a Hamiltonian is $2N$-dimensional. A change of variables $(\vec{x}, \vec{y}) \to (\vec{x}', \vec{y}')$ is called canonical if the equations of motion in the new variables are also canonical.

The most important separation of dynamical systems is between integrable systems and chaotic systems.

The most simple integrable systems are the solvable systems, i. e. systems that can be solved explicitly, to give the variables as functions of time. This definition is too restricted, if by functions we mean known functions of time. A more general definition is in terms of single-valued functions of time [79], even if such functions can be given only numerically.

A more general definition of integrable systems is systems that have N independent analytic integrals of motion in involution $I_i(\vec{x}, t) = I_i(\vec{x}_0, t_0)$, $i = 1, 2, \ldots, N$. (By involution we mean that any two such integrals have a zero Poisson bracket (see Eq. 6)).

As regards chaotic systems, their definition and classification is a difficult task. The basic property of chaos is sensitive dependence on initial conditions. Namely two orbits of a compact system starting very close to each other deviate considerably later on (exponentially fast in time).

A particular class of chaotic systems are the ergodic systems, in which the orbits go everywhere on a surface of constant energy. Namely, the orbits pass through the *neighborhood* of every point of the energy surface. In an ergodic system the time average of a function is equal to its phase average, according to Birkhoff's theorem [6]. An ergodic system is called "mixing" if two nearby particles on different (ergodic) orbits can go very far from each other.

A further distinction of mixing systems is related to the speed of deviation of nearby orbits. If the deviation is exponential in time the system is called "Kolmogorov" or K-system. The deviation $\vec{\xi}$ of nearby orbits that are initially (at time t_0) at an infinitesimal distance $\vec{\xi}_0$ is measured by the Lyapunov characteristic number (LCN) (or simply the Lyapunov exponent).

For almost all deviations $\vec{\xi}_0$ from the initial condition $\vec{x}_0$ of an orbit the Lyapunov characteristic number is the same, equal to

$$\mathrm{LCN} = \lim_{t\to\infty} \frac{\ln |\vec{\xi}/\vec{\xi}_0|}{t} \tag{2}$$

In a system of N degrees of freedom one can define N Lyapunov characteristic numbers. If not specified explicitly otherwise the term Lyapunov characteristic number means the maximal LCN. A system is said to have sensitive dependence on the initial conditions if its LCN is positive.

A simple system with positive LCN is given by $\xi = \xi_0 \exp(qt)$. In this system $\ln(\xi/\xi_0) = qt$, therefore LCN $= q$ (constant). A system in which the deviation ξ increases linearly in time (or as a power of time) is not Kolmogorov although it is mixing. For example, if $\xi = \xi_0 + \alpha t$, we have $\ln(\xi/\xi_0) \approx \ln t$, and the limit LCN $= \lim_{t\to\infty}(\ln t/t)$ is zero.

If in a Kolmogorov system the LCN for all orbits is between certain positive limits, i. e. if $C_1 \leq \text{LCN} \leq C_2$ the system is called "Anosov", or C-system.

A C-system has infinite periodic orbits, but none of them is stable. Furthermore the Anosov systems are hyperbolic, i. e. the stable and unstable manifolds of each unstable periodic orbit intersect transversally. The Anosov systems are *structurally stable*, i. e. a small perturbation of an Anosov system leads to another Anosov system.

The usual classification [50] of dynamical systems is the following: Integrable systems, ergodic systems, mixing systems, Kolmogorov systems, Anosov systems.

This classification represents the view that if a system is not integrable it is at least ergodic. The first edition of the book of Landau and Lifshitz [48] separated all dynamical systems into two classes, integrable and ergodic. (However after the 3rd edition this part was omitted). This classification misses the most important property of chaos, the fact that in general chaos co-exists with order in the same system.

In fact, in most systems that have been studied numerically up to now, one finds both chaotic and ordered orbits. Such systems are called systems with divided phase space. Chaotic orbits have a positive Lyapunov characteristic number, while ordered orbits have LCN $= 0$. Both chaotic and ordered orbits cover parts of the phase space. This is best seen in numerical studies of systems of two degrees of freedom. The ordered and chaotic domains are intricately mixed. However, there are regions where order is predominant, and other regions where chaos is predominant.

Systems of two degrees of freedom that are close to integrable have only small regions of chaos and most orbits are ordered. On the other hand, systems that are far from integrable may seem completely chaotic, but after a closer look we usually find small islands of stability in them.

Completely integrable systems are exceptional. On the other hand, truly ergodic systems are also exceptional. For example, some piecewise linear maps, that are given modulo a constant, or systems with abrupt reflections, like the *stadium* have been proven to be ergodic. But if we add generic nonlinear terms in these maps the structure of phase space changes and islands of stability appear.

The present day classification of dynamical systems is

	Ordered	Chaotic		Random
Compact	Integrable	(General Case) Systems with Divided Phase Space	(Limiting Cases) Ergodic Mixing Kolmogorov Anosov	
Non-compact	Integrable with escapes	Nonintegrable with escapes	–	

A particular class are the noncompact systems (i. e. systems with escapes). In such systems some initial conditions lead to escapes, while others do not. These systems are either integrable, or chaotic. In the latter case chaos appears in the form of chaotic scattering. But properly speaking chaos refers only to compact systems. Finally, the random systems are limiting cases of Anosov systems, because they have infinitely large Lyapunov characteristic numbers.

Integrable Systems

The simplest integrable systems can be solved explicitly. Autonomous Hamiltonian equations of the form (1) with one degree of freedom ($N = 1$) are always solvable.

A very useful set of variables are the action-angle variables $(\vec{J}, \vec{\theta})$. In particular the change of variables

$$x_i = \sqrt{2J_i}\,\sin\theta_i\,, \quad y_i = \sqrt{2J_i}\,\cos\theta_i\,, \quad (i = 1, 2 \ldots N) \tag{3}$$

is canonical and the new variables $(\vec{J}, \vec{\theta})$ satisfy canonical equations of the form

$$\dot{\theta} = \frac{\partial H}{\partial J}\,, \quad \dot{J} = -\frac{\partial H}{\partial \theta} \tag{4}$$

In the particular case that H is a function of the actions only $H = H(\vec{J})$ the actions $\vec{J}(J_1, J_2, \ldots J_N)$ are integrals of motion, because $\dot{J}_i = 0$, and consequently the derivatives of H with respect to J_i are constant, i. e. $\partial H/\partial \vec{J} = \vec{\omega}(\vec{J})$ where $\vec{\omega}$ represents the frequencies $(\omega_1, \omega_2, \ldots, \omega_N)$. Then the angles $(\theta_1, \theta_2, \ldots, \theta_N)$ are linear functions of the time $\theta_i = \omega_i(t - t_{i0})$. In such a case the motion takes place on a N-dimensional torus.

A function $I(\vec{x}, \vec{y}, t)$ in a Hamiltonian system is called an integral of motion if it remains constant along any orbit, i. e. its total derivative is zero. Then

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \frac{\partial I}{\partial \vec{x}}\frac{\mathrm{d}\vec{x}}{\mathrm{d}t} + \frac{\partial I}{\partial \vec{y}}\frac{\mathrm{d}\vec{y}}{\mathrm{d}t} + \frac{\partial I}{\partial t} = [I, H] + \frac{\partial I}{\partial t} = 0 \tag{5}$$

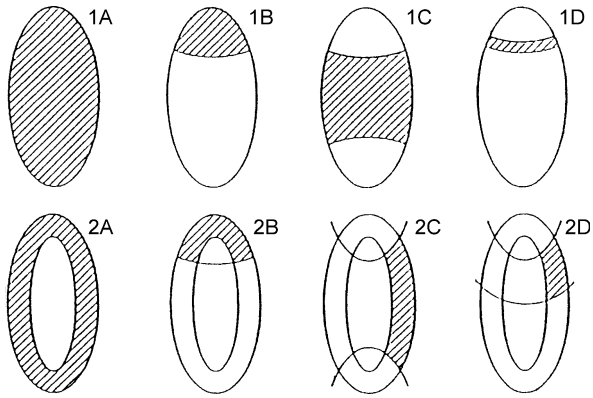

Astrophysics: Dynamical Systems, Figure 2
Regions filled by the various types of orbits in a Stäckel potential

where $\vec{x}$, $\vec{y}$ are additive and

$$[I, H] \equiv \frac{\partial I}{\partial \vec{x}} \frac{\partial H}{\partial \vec{y}} - \frac{\partial I}{\partial \vec{y}} \frac{\partial H}{\partial \vec{x}} \tag{6}$$

is called the *Poisson bracket* between the functions I and H.

An important class of integrable systems are the Stäckel potentials [33,71,76]. They are given in elliptical coordinates (λ, μ) by potentials of the form

$$V = -\frac{[F_1(\lambda) - F_2(\mu)]}{\lambda - \mu} \tag{7}$$

where λ and μ are the roots of the second degree equation in τ

$$\frac{x^2}{\tau - a^2} + \frac{y^2}{\tau - b^2} = 1 \tag{8}$$

with Cartesian coordinates x, y and $a^2 \geq b^2$. The Eq. (8) represents confocal ellipses and hyperbolas. From every point (x, y) passes an ellipse (defined by λ) and a hyperbola (defined by μ). Both have the same foci at $y = \overline{a} = \pm(a^2 - b^2)^{1/2}$. The orbits in the plane (x, y) fill regions between such ellipses and hyperbolas (Fig. 2).

The Stäckel potentials are in general non-rotating. Up to now only one rotating Stäckel potential has been found, besides the trivial case of a homogeneous ellipsoid, namely the potential

$$V = -\frac{k_1}{[(x^2 + y^2)^2 + 2\overline{a}^2(x^2 - y^2) + \overline{a}^4]^{1/2}} + \frac{1}{2}\Omega_s^2 r^2 \tag{9}$$

The orbits in the rotating case are in general tubes around both foci, or tubes around one focus only [24] (Fig. 3).

Stäckel potentials in three degrees of freedom were considered by many authors like [31,45,53]. There are 32 possible forms of 3-D Stäckel orbits [21].

The Stäckel potentials have been used extensively in recent years in constructing self-consistent models of elliptical galaxies [7,32,46,72]. Such models represent well real galaxies.

Formal Integrals. KAM and Nekhoroshev Theory

If a system is close to an integrable one, we may use perturbation methods to find approximate integrals that are valid for long intervals of time, and sometimes for all times. Usually such integrals are in the form of formal power series. Formal integrals for polynomial Hamiltonians were first derived by Whittaker [77,78].

Another form of such integrals was derived by Birkhoff [5] and by Cherry [9]. Formal integrals appropriate for galactic dynamics were derived by Contopoulos [14,15]. In the case of an axisymmetric galaxy an integral of this form is called a *third integral* (the first integral

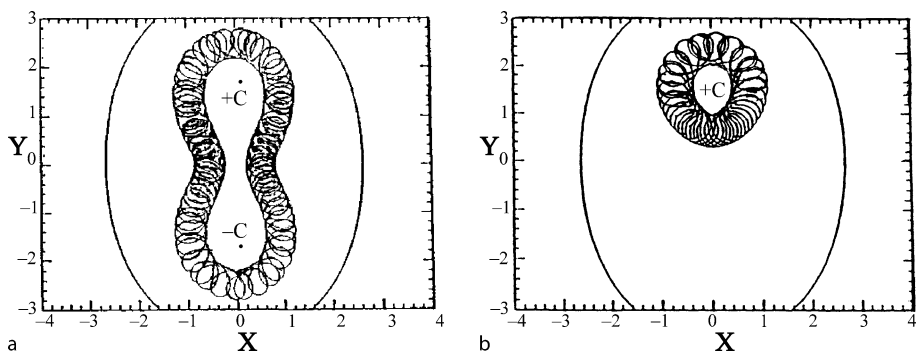

Astrophysics: Dynamical Systems, Figure 3
Orbits in a rotating Stäckel potential

being the energy and the second the z component of the angular momentum).

Birkhoff [5] and Cherry ([9] use Hamiltonians of the form)

$$H = i\omega_1\xi_1\eta_1 + i\omega_2\xi_2\eta_2 + \cdots + i\omega_N\xi_N\eta_N + H_3 + H_4 + \ldots \tag{10}$$

where H_k is of degree k in $(\xi_\lambda, \eta_\lambda)$, and $\xi_\lambda, \eta_\lambda$ are complex variables ($\lambda = 1, 2, \ldots N$).

By successive canonical changes of variables the Hamiltonian takes the *normal form* $\overline{H} = \overline{H}_2 + \overline{H}_4 + \ldots$ where $\overline{H}_k$ is a function of the products $\Omega_\lambda = (\overline{\xi}_\lambda \overline{\eta}_\lambda)$ only (in the nonresonant case), thus it is an even function of the variables. Then the quantities Ω_λ are integrals of motion, and the equations of motion can be solved explicitly.

The third integral in a galaxy refers to a Hamiltonian of the form [14]

$$H = \frac{1}{2}\sum_{\lambda=1}^{2}\left(y_\lambda^2 + \omega_\lambda^2 x_\lambda^2\right) + H_3 + H_4 + \ldots \tag{11}$$

where H_k is a function of degree k in x_λ, y_λ ($\lambda = 1, 2$). This Hamiltonian represents a galactic potential on a plane of symmetry. If $r = r_0$ represents a circular orbit on this plane, we have $x_1 = r - r_0$, $x_2 = z$ and $y_1 = \dot{x}_1$, $y_2 = \dot{x}_2$. Then the term H_3 of the Hamiltonian takes the form

$$H_3 = -\epsilon x_1 x_2^2 - \frac{\epsilon'}{3}x_1^3 \tag{12}$$

An integral of motion has its total derivative equal to zero. The successive terms of the third integral Φ can be found from the equation

$$\frac{\mathrm{d}\Phi}{\mathrm{d}t} = [\Phi, H] = 0 \tag{13}$$

(where $\Phi = \Phi_2 + \Phi_3 + \ldots$, and $[\Phi, H]$ is the Poisson bracket) by equating terms of equal degree.

The Hamiltonian H_2 has two independent integrals of motion $\Phi_2 = \frac{1}{2}\left[y_1^2 + \omega_1^2 x_1^2\right]$ and $\Phi_2' = H_2 - \Phi_2$. Then, we find

$$\Phi_3 = -\frac{1}{(\omega_1^2 - 4\omega_2^2)}\cdot\left[\left(\omega_1^2 - 2\omega_2^2\right)x_1x_2^2 - 2x_1y_2^2 + 2y_1x_2y_2\right] \tag{14}$$

etc. The expression (14) and the higher order terms Φ_k of the third integral contain denominators of the form $(m\omega_1 - n\omega_2)$ that become zero if $\omega_1/\omega_2 = n/m$ (rational). Then Φ contains secular terms. For example, the form (14) of Φ_3 cannot be valid if $\omega_1 - 2\omega_2 = 0$. It is not valid also if $\omega_1 - 2\omega_2$ is small (case of a "small divisor") because then Φ_3 is large, while the series expansion of Φ implies that the successive terms Φ_2, Φ_3, etc. become smaller as the order increases.

In such resonance cases we can construct a third integral by a different method. If ω_1/ω_2 is rational there are further isolating integrals of motion of the lowest order Hamiltonian H_2 beyond the partial energies Φ_2, Φ_2'. For example, if $\omega_1/\omega_2 = n/m$ then H_2 has also the integrals

$$\begin{aligned} \begin{matrix} S_0 \\ C_0 \end{matrix} &= (2\Phi_2)^{m/2}(2\Phi'_2)^{n/2}\begin{matrix}\sin\\ \cos\end{matrix}\left[n\omega_2 T - m\omega_1(T - T_0)\right] \\ &= \text{constant} \end{aligned} \tag{15}$$

which can be written as polynomials of degree $m + n$ in the variables x_λ, y_λ [15]. The integrals S_0, C_0 can be used for constructing series of the form $S = S_0 + S_1 + S_2 + \ldots$, $C = C_0 + C_1 + C_2 + \ldots$. These series also contain secular terms. However, we can prove that there are combinations of the three integrals Φ_2, Φ_2' and S_0 (or C_0) that do not contain secular terms of any order, thus they are formal integrals of motion.

The series of the third integral are in general not convergent. However, they may give very good applications, if truncated at an appropriate level. A simple example of such a series was provided by Poincaré ([66], Vol. II). The series

$$f_1 = \sum_{n=0}^{\infty}\frac{n!}{(1000)^n} \tag{16}$$

is divergent (i. e. f_1 is infinite), but if it is truncated at any large order n smaller than 1000, it gives approximately the same value, and it seems to converge, as n increases, to a finite value. On the other hand the series

$$f_2 = \sum_{n=0}^{\infty}\frac{(1000)^n}{n!} \tag{17}$$

is convergent, but in order to give an approximate value of f_2 one has to take more than 1000 terms.

There is an important theorem establishing the existence of N-dimensional invariant tori in systems of N degrees of freedom. This is the celebrated KAM theorem, announced by Kolmogorov [47]. The details of the proof were given by Arnold [2,3] and independently by Moser [59,60].

The KAM theorem states that in an autonomous Hamiltonian system, close to an integrable system expressed in action-angle variables, there are N-dimensional

invariant tori, i. e. tori containing quasi-periodic motions with frequencies ω_i, satisfying a diophantine condition (after the ancient Greek mathematician Diophantos) $|\vec{\omega}k| > \gamma/|k|^{N+1}$, where $\vec{\omega}k = \omega_1 k_1 + \omega_2 k_2 + \cdots + \omega_N k_N$ with $k_1, k_2, \ldots, k_N$ integers and $|k| = |k_1| + |k_2| + \cdots + |k_N| \neq 0$, while γ is a small quantity depending on the perturbation. The set of such tori has a measure different from zero if γ is small.

In the case of two degrees of freedom the ratio ω_1/ω_2 is far from all resonances, if it satisfies a diophantine condition $|\omega_1/\omega_2 - n/m| > \epsilon/m^3$. If we exclude the sets of all irrationals that are near every rational n/m, on both its sides, the set of the excluded irrationals is smaller than

$$\sum_{m=1}^{\infty} \sum_{n=1}^{m-1} \frac{2\epsilon}{m^3} < 2\epsilon \sum_{m=1}^{\infty} \frac{1}{m^2} = 2C\epsilon \tag{18}$$

where $C = \pi^2/6 \approx 1.64$. Therefore the set of diophantine numbers ω_1/ω_2 in the interval $(0, 1)$ has a measure larger than $1 - C\epsilon$ and if ϵ is small this is not only positive, but it can even be close to 1.

The formal integrals Φ are convergent only in integrable cases. In other cases Φ are asymptotic series that may not represent a particular functions. The question then arises for how long the formal integrals $\Phi = \Phi_2 + \Phi_3 + \ldots$ are valid within a given approximation. The answer to this question is provided by the theory of Nekhoroshev [62]. This theory establishes that the formal integrals are valid over exponentially long times. Namely, if the deviation from an integrable system is of order ϵ the Nekhoroshev time is of order

$$t = t_* \exp\left(\frac{M}{\epsilon^m}\right) \tag{19}$$

where t_*, M and m are positive constants. The formal integrals must be truncated at an optimal order, such that the variation of the truncated integrals is minimum. An analytical estimate of the optimal order of truncation was provided recently by Efthymiopoulos et al. [36].

Let us consider the orbits in the phase space of an autonomous Hamiltonian system of two degrees of freedom $H \equiv H(x, y, p_x, p_y) = h$. As the energy of the system, h, is an integral of motion, one variable, say p_y, can be derived from H if the other three variables are known. All orbits with fixed energy equal to h, lie on the above 3-D surface, in the 4-D phase space. Thus, we may consider the orbits in the 3 dimensional phase space (x, y, p_x).

If every orbit is intersected by a particular surface $S(p_x, p_y, x, y) = \text{const}$ within every fixed interval of time T, then this surface is called a Poincaré surface of section. The intersections of the points of S are called consequents; they define a map on the surface of section S that is called a Poincaré map. In particular, a periodic orbit of multiplicity three is represented by three points (e. g. O_1, O_2, O_3 in Fig. 4).

Periodic Orbits

The most important orbits in a dynamical system are the periodic orbits. If the periodic orbits are stable they trap around them a set of nonperiodic orbits with the same average topology. If they are unstable they repel the orbits in their neighborhood. In integrable systems the main unstable orbits separate the resonant orbits from the nonresonant orbits. In nonintegrable systems the unstable periodic orbits introduce chaos in the system. Thus, the study of a dynamical system usually starts with an exploration of its periodic orbits.

In the case of an area preserving map $x_1 = f(x_0, y_0, \epsilon)$, $y_1 = g(x_0, y_0, \epsilon)$, where ϵ is a parameter, a periodic orbit is an invariant point $x_1 = x_0$, $y_1 = y_0$. Its stability is found by linearizing the equations around this point. We have $\Delta x_1 = a\Delta x_0 + b\Delta y_0$, $\Delta y_1 = c\Delta x_0 + d\Delta y_0$, where $a = \partial f/\partial x_0$, $b = \partial f/\partial y_0$, $c = \partial g/\partial x_0$, $d = \partial g/\partial y_0$. The conservation of areas implies $ad - bc = 1$.

In a Hamiltonian system of two degrees of freedom the periodic orbits and the deviations $\vec{\xi} = (\Delta x, \Delta y)$ are defined in the 4-D phase space $(x, y, \dot{x}, \dot{y})$. Along particular directions in the 4-D phase space the original deviation $\vec{\xi}_0 = (\Delta x_0, \Delta y_0)$ becomes $\vec{\xi}_1 = (\Delta x_1, \Delta y_1)$ after one iteration, where $\vec{\xi}_1$ is equal to $\vec{\xi}_0$ multiplied by a factor λ. Such vectors $\vec{\xi}_0$ are called *eigenvectors* (with arbitrary length) and the factors λ are called *eigenvalues*. The eigenvector forms an angle Φ with the x-axis, given by $\tan\Phi = (\lambda - a)/b = c/(\lambda - d)$, that has two opposite solutions Φ and $\Phi + 180°$ for each real eigenvalue λ.

The two eigenvalues λ_1, λ_2 are found from the equation

$$\begin{vmatrix} a - \lambda & b \\ c & d - \lambda \end{vmatrix} = \lambda^2 - (a + d)\lambda + 1 = 0 \tag{20}$$

The roots λ_1, λ_2 of this equation are inverse. If $|a + d| > 2$ the roots λ are real ($|\lambda_1| > 1$ and $|\lambda_2| < 1$). Then, the orbit is unstable and the eigenvectors $\vec{\xi}_{01}, \vec{\xi}_{02}$ form the angles Φ with the x-axis. On the other hand if $|a + d| < 2$ the roots are complex conjugate with $|\lambda_i| = 1$ and the orbit is stable.

Ordered orbits lie on invariant tori that intersect a surface of section along invariant curves. Orbits that do not lie on closed invariant curves and do not escape to infinity are chaotic. On a surface of section such orbits are represented by the irregular distribution of their consequents.

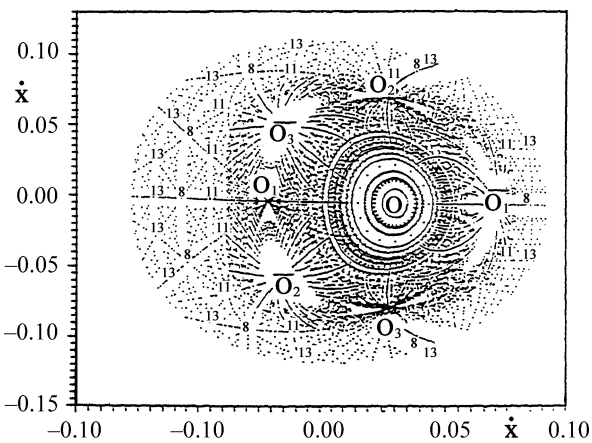

Astrophysics: Dynamical Systems, Figure 4
Distribution of periodic orbits of various multiplicities in the potential $V = \frac{1}{2}(\omega_1^2 x_1^2 + \omega_2^2 x_2^2) - \epsilon x_1 x_2^2$ with $\omega_1^2 = 1.6$, $\omega_2^2 = 0.9$, $\epsilon = 0.3$

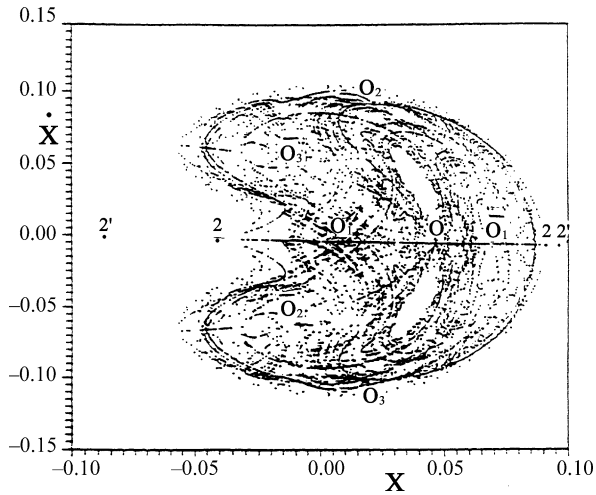

Astrophysics: Dynamical Systems, Figure 5
As in Fig. 4 for $\epsilon = 4.5$

Such chaotic orbits appear near every unstable periodic orbit of a nonintegrable system. The domains filled by the chaotic orbits surround the corresponding islands of stability of the same chaotic zone that contains the unstable periodic orbits. For example, the chaotic domains of the Hamiltonian (11) with $H_3 = -\epsilon x_1 x_2^2$ are near the unstable points O_1, O_2, O_3 of type $n/m = 2/3$ in Fig. 4. These chaotic domains surround also the islands around the stable periodic orbit $(\overline{O}_1 \overline{O}_2 \overline{O}_3)$.

For relatively small perturbations the various resonances are well separated by invariant curves that close around the central periodic orbit O, therefore chaos is limited. But when the perturbation increases the various chaotic domains increase and join each other, destroying the separating invariant curves, and producing a large connected chaotic domain (Fig. 5) (for the same Hamiltonian but with $\epsilon = 4.5$). This is a manifestation of a "resonance overlap", or "resonance interaction".

The asymptotic curves of a hyperbolic point of an integrable system join into one separatrix, unless these curves extend to infinity. The separatrix may be considered as an orbit (homoclinic orbit), or as an invariant curve containing the images of the initial points on this curve. An orbit starting very close to the origin along the unstable branch U will move far from O, but eventually it will return close to O along the stable branch S.

However, in a nonintegrable system there are no closed separatrices, and the unstable and stable asymptotic curves intersect at infinite points called homoclinic points.

All initial points on an asymptotic curve generate *asymptotic orbits*, i. e. orbits approaching the periodic orbit O either in the forward or in the backward direction of time. The homoclinic points define *doubly asymptotic orbits* because they approach the orbit O both for $t \to -\infty$ and for $t \to \infty$.

The intersecting asymptotic curves of an unstable periodic orbit form the so-called *homoclinic tangle*. If we start an orbit in this region, its consequents fill the region in a practically random way, forming the chaotic domain that we see near every unstable periodic orbit of a nonintegrable system.

Transition from Order to Chaos

Dissipative Systems. The Logistic Map

A simple one dimensional dissipative system, where we can see the transition from order to chaos, is the logistic map

$$x_{i+1} = f(x_i) = 4\lambda x_i(1 - x_i) \tag{21}$$

This is the prototype of a quadratic map, and it is studied extensively in the pioneering article of May [56] and in books on chaos (see e. g. [1,12]). Thus, we will only describe here its main properties. We consider only values of λ between 0 and 1 in order to have x always between 0 and 1. The logistic map has two simple periodic orbits on the intersection of the parabola (21) with the diagonal $x_{i+1} = x_i$, namely the points $x = 0$ and $x = x_0 = 1 - \frac{1}{4\lambda}$.

A periodic orbit x_0 is stable if the derivative $f' = \mathrm{d}f/\mathrm{d}x_0$ is absolutely smaller than 1. This happens if $1/4 < \lambda < \Lambda_1 = 3/4$.

It is easily found that when the orbit $x_0 = 1 - (1/4\lambda)$ of period 1 becomes unstable, then a period -2 family of

periodic orbits (x_a, x_b) bifurcates, which is stable if $\Lambda_1 < \lambda < \Lambda_2 = (1+\sqrt{6})/4$.

There is an infinite set of period doubling bifurcations (Fig. 6a). The intervals between successive bifurcations decrease almost geometrically at every period doubling, i. e.

$$\lim_{n\to\infty} \frac{\Lambda_n - \Lambda_{n-1}}{\Lambda_{n+1} - \Lambda_n} = \delta = 4.669201609\ldots \tag{22}$$

This relation is asymptotically exact (i. e. for $n \to \infty$) [28,37]. The number δ is universal, i. e. it is the same in generic dissipative systems.

The curves $F^m(\lambda)$, with $m = 2^n$, starting at the nth period doubling bifurcation, are similar to each other, decreasing by a factor $\alpha = 2.50\ldots$ in size at each successive period doubling bifurcation.

As the ratios (22) decrease almost geometrically, the value Λ_n converges to the limiting value $\Lambda_\infty = 0.893$.

In 2-D maps the areas are reduced at every iteration in the dissipative case and conserved in the conservative case. An example is provided by the Hénon map [42]

$$x_{i+1} = 1 - Kx_i^2 + y_i, y_{i+1} = bx_i \pmod 1 \tag{23}$$

(the original map was not given modulo 1). The Jacobian of the transformation is $J = b$. If $0 < b < 1$ the system is dissipative and if $b = 1$ it is conservative.

In the dissipative case there are *attractors*, i. e. manifolds to which tend many orbits as $t \to \infty$ (or $i \to \infty$ in maps). These attractors may be points, curves (limit cycles), or strange attractors [39,43,52], which are composed of infinite lines.

Other mechanisms of transition to chaos in dissipative systems are reviewed by Contopoulos (Section 2.6.3 in [21]).

Conservative Systems

The first numerical study of the transition to chaos in a conservative system was provided by Hénon and Heiles [44]. The transition to chaos follows a number of scenarios.

Infinite Period Doubling Bifurcations This scenario is similar to the corresponding scenario of the dissipative case. In fact a large degree of chaos is introduced after an infinite number of period doubling bifurcations along the characteristics of the periodic orbits ($x = x(\lambda)$). However, there are three differences in the conservative case.

(1) Infinite period doubling bifurcations appear generically only in systems of two degrees of freedom.
(2) The bifurcation ratio is different ($\delta \cong 8.72$ in conservative systems versus $\delta \cong 4.67$ in dissipative systems). Also the scaling factors are different.
(3) The introduction of chaos follows a different pattern. While in the dissipative case chaos appears only *after* infinite bifurcations (Fig. 6a), in the conservative systems some chaos appears around every unstable orbit. As the perturbation increases the chaotic domains increase in size and they merge to form a large connected domain (Fig. 6b).

In some cases the appearance of infinite bifurcations may be followed by their corresponding disappearance, as the perturbation increases further. Then the infinite unstable families terminate in the opposite way forming infinite bubbles.

Infinite Bifurcations from the same Family While in the infinite bifurcations scenario we have pitchfork bifurcations from the successive bifurcating families (Fig. 6), in the present case the original family of periodic orbits becomes successively stable and unstable, an infinite number of times. For example, this happens in the case of the Hamiltonian

$$H = \frac{1}{2}(\dot{x}^2 + \dot{y}^2 + x^2 + y^2) + xy^2 = h \tag{24}$$

[26]. The stable bifurcating families undergo further period doubling bifurcations. The successive stable and unstable intervals along the original family have a bifurcation ratio $\delta \cong 9.22$. But this bifurcation ratio is not universal. It depends on the particular dynamical system considered. In the case (24) it was proven analytically by Heggie [41] that this ratio is $\delta = \exp(\pi/\sqrt{2})$ but in other systems it is different. In all cases near the limiting point there is an infinity of unstable families that produce a large degree of chaos.

Infinite Gaps In a dynamical system representing a rotating galaxy, the central periodic family, which consists of circular orbits in the unperturbed case, has two basic frequencies. These are the rotational frequency in the rotating frame $(\Omega - \Omega_s)$ (where Ω is the angular velocity along a circular orbit and Ω_s the angular velocity of the system, e. g. of the spiral arms), and the epicyclic frequency κ of radial oscillations from the circular orbit.

The ratio of the two frequencies (rotation number) goes through infinite rational values $\kappa/(\Omega - \Omega_s) = n/m$ (resonances) as the energy increases.

The most important resonances are those with $m = 1$. At every even resonance ($n = 2n_1$, $m = 1$), we have a bifurcation of a resonant family of periodic orbits in the

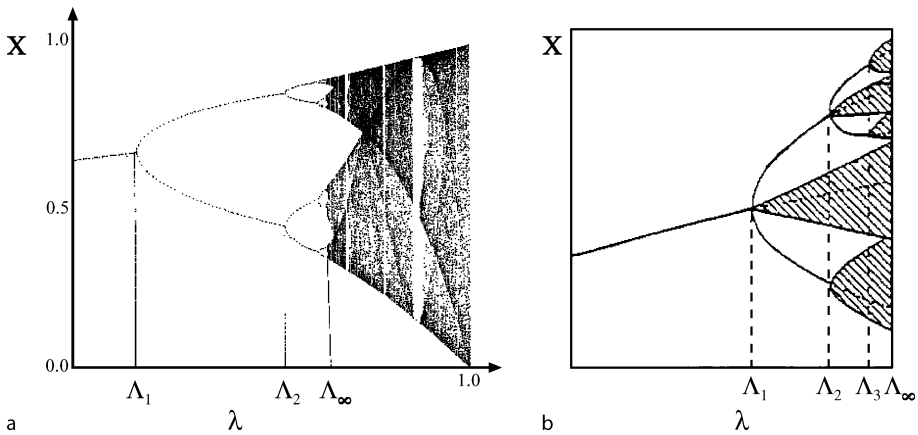

Astrophysics: Dynamical Systems, Figure 6
Bifurcations and chaos: a in a dissipative system (the logistic map) and b in a conservative system

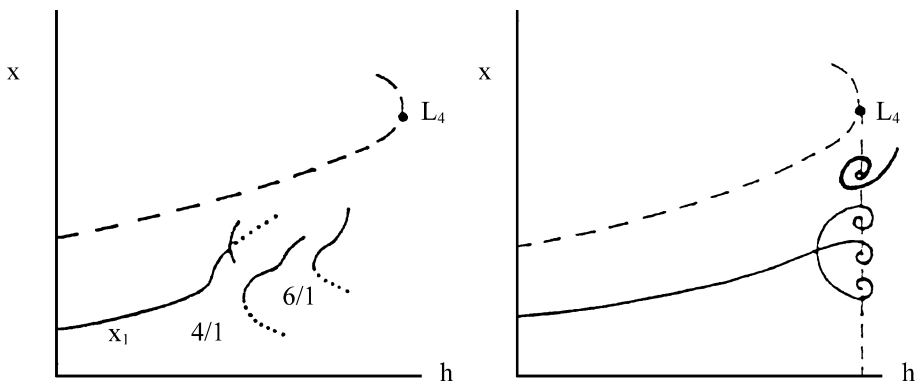

Astrophysics: Dynamical Systems, Figure 7
a Gaps appear along the characteristic of the family x_1 at all even resonances $2n_1/m$. b Spiral characteristics near corotation (L_4)

unperturbed case, which becomes a gap in the perturbed case (when a spiral, or a bar, is added to the axisymmetric background, Fig. 7a). The gap is formed by separating the main family into two parts and joining the first part with a branch of the resonant family, and the second part containing the other branch of the resonant family.

As we approach corotation (where $\Omega = \Omega_s$) the rotation number tends to infinity. Therefore, an infinity of gaps are formed, followed by the appearance of an infinity of unstable periodic orbits, and this leads to a large degree of chaos [19].

Infinite Spirals There is an infinite number of families of periodic orbits whose characteristics form spirals near corotation (Fig. 7b). These families contain an infinite number of unstable periodic orbits that interact with each other, producing chaos [21,65].

Resonance Overlap

In a nonintegrable Hamiltonian we have several types of resonant islands and their size can be found by means of the third integral, calculated for each resonance separately.

We consider a Hamiltonian of the form (11) with ω_1/ω_2 near a resonance n/m. The maximum size D of an island of type n/m is of $O(\epsilon^{(m+n-4)/2})$ [17]. The same is true for the area covered by the islands of the resonance n/m.

If the islands of various nearby resonances are well separated, then there is no resonance overlap and no large degree of chaos. However, as the perturbation increases the various islands increase in size and the theoretical islands overlap. As this overlapping is not possible in reality, what really happens is that the region between the islands becomes chaotic.

There are two ways to estimate the critical perturbation of the resonance overlap.

(1) We calculate the areas of the various islands. When this quantity becomes equal to the total area of phase space on the surface of section, we start to have large degree of chaos [17].
(2) We may find the critical value for the production of large chaos between two main resonances of the system, say $n/m = 4/3$ and $n/m = 3/2$. The positions of the triple and double periodic orbits are given by the corresponding forms of the third integral as functions of the perturbation ε (characteristics). The theoretical characteristics intersect for a certain value of ε. On the other hand, the real characteristics cannot intersect. Instead, between the two characteristics, a chaotic region is formed, due to the resonance overlap.

The resonance overlap criterion for large chaos was considered first by Rosenbluth et al. [67] and Contopoulos [17]. It was later described in detail by Walker and Ford [75], Chirikov et al. [11], Chirikov [10], and many others.

Arnold Diffusion

Nearly integrable autonomous Hamiltonian systems of N degrees of freedom have a $2N-1$ dimensional phase space of constant energy, and a large set of N-dimensional invariant surfaces (invariant tori), according to the KAM theorem. If $N = 2$ there are $N = 2$ dimensional surfaces separating the $2N-1 = 3$ dimensional phase space, and diffusion through these surfaces is impossible. But if $N \geq 3$ the N-dimensional surfaces do not separate the $2N-1$ dimensional phase space, and diffusion is possible all over the phase space. This is called "Arnold diffusion" [3]. Such a diffusion is possible even if the perturbation is infinitesimal.

In order to understand better this phenomenon consider a system of three degrees of freedom, that has 3-D tori in a 5-D phase space of constant energy. If we reduce the number of both dimensions by two we have 1-D tori (lines) in a 3-D space. In an integrable case there is a line passing through every point of the space and all motions are regular. However, in a nonintegrable case there are gaps between the lines and these gaps communicate with each other. Therefore, if a chaotic orbit starts in such a gap it may go everywhere in the 3-D space.

Extensive numerical studies (e. g. [49]) in a 4-D map have shown that in the same dynamical system there are both ordered domains where diffusion is very slow (Arnold diffusion), and chaotic domains, where diffusion is dominated by resonance overlap.

As an example we consider two coupled standard maps

$$\begin{aligned} x_1' &= x_1 + y_1' \,, \\ y_1' &= y_1 + \frac{K}{2\pi}\sin 2\pi x_1 - \frac{\beta}{\pi}\sin 2\pi(x_2 - x_1) \\ & \qquad\qquad (\mathrm{mod}\ 1) \\ x_2' &= x_2 + y_2' \,, \\ y_2' &= y_2 + \frac{K}{2\pi}\sin 2\pi x_2 - \frac{\beta}{\pi}\sin 2\pi(x_1 - x_2) \end{aligned} \tag{25}$$

Here K is the nonlinearity and β the coupling constant [25].

Taking the same initial conditions, we calculate the time of diffusion from an ordered region to the large chaotic sea. If the coupling β is larger than a critical value $\beta = \beta_c \cong 0.305$, the diffusion time T increases exponentially with decreasing β. If however, $\beta < \beta_c$, the time T increases superexponential as β decreases. We can identify the exponential increase case as due to resonance overlap diffusion and the superexponential increase as due Arnold diffusion.

Cantori

When the perturbation increases and a particular torus is destroyed, it becomes a cantorus, i. e. a Cantor set of points that is nowhere dense [4,64]. The cantorus has a countable infinity of gaps, but it contains a noncountable infinity of orbits. The most surprising property of the cantori is that their measure is zero and their fractal dimension is also zero.

In the case of the standard map, cantori with golden rotation number $rot = (\sqrt{5}-1)/2$ (expressed in continuous fraction representation as $[1, 1, 1, \ldots]$) exist for all values of K larger than a critical value $K_{cr} = 0.972$ ([55]).

The cantorus is formed when all the periodic orbits corresponding to the various truncations of the golden number become unstable. In the limit $K = K_{cr}$ the invariant curve (torus) is a fractal with self-similar structure on all scales [40,69].

What happens in general around an island of stability is the following. Before the destruction of the last KAM curve surrounding the island, there is a chaotic layer, just inside the last KAM curve, while outside the last KAM curve there is a large chaotic sea. When the last KAM curve becomes a cantorus the orbits from the inner chaotic layer can escape to the large chaotic sea, but only after a long time. Thus, the region just inside a cantorus with small holes is differentiated by the larger density of the points inside it (Fig. 8). This is the phenomenon of *stickiness*.

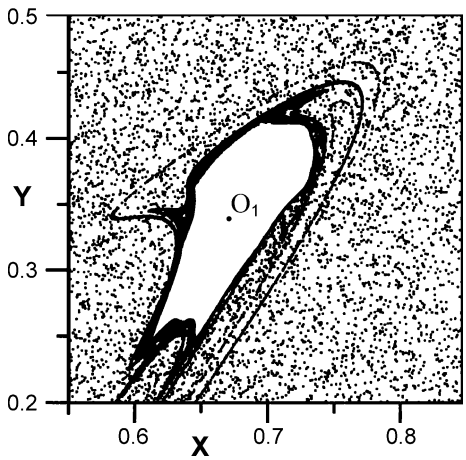

Astrophysics: Dynamical Systems, Figure 8
Stickiness in the standard map for $K = 5$. The sticky region (*dark*) surrounds an island of stability, and is surrounded by a large chaotic sea

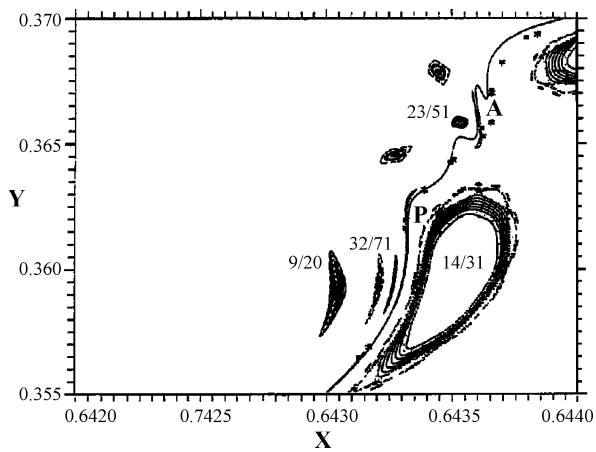

Astrophysics: Dynamical Systems, Figure 9
An asymptotic curve of the periodic orbit P in the standard map with $K = 5$ makes several oscillations inside the cantorus (*stars*) and then escapes. We mark some islands inside the cantorus (on the *right*) and outside it (on the *left*)

Stickiness

This phenomenon was found numerically by Contopoulos [18]. Stickiness has been observed in practically all nonintegrable dynamical systems that contain islands [57]. In Fig. 8 the sticky zone (the dark region surrounding the island O_1) is limited on its inner side by the outer boundary of the island O_1 and on its outer side by a set of cantori, but mainly by the cantorus that has the smallest gaps. An orbit starting in the sticky zone requires a long time to escape outside. The details of the escape can be found if we calculate the asymptotic curves of unstable periodic orbits just inside the main cantorus.

As an example consider an unstable asymptotic curve from the periodic orbit of period 215 in the case of the standard map for $K = 5$ (Fig. 9; [35]). If we start with a small interval along the asymptotic curve, we find an image of this curve that crosses the cantorus after two iterations along the same asymptotic curve. In Fig. 9 we see that the asymptotic curve starts from P by going first to the left and downwards and then it moves to the right and upwards, crossing the cantorus in one of its largest gaps (marked A in Fig. 9). But then the asymptotic curve makes several oscillations back and forth, entering again inside the cantorus several times, before going to a large distance in the chaotic sea. Then it moves around for a long time before coming back and getting trapped again in the sticky zone close to the island.

Another type of stickiness refers to the unstable asymptotic curves that extend far into the chaotic sea (dark lines in Fig. 8). These asymptotic curves attract nearby orbits and thus thick dark lines are formed [23].

Dynamical Spectra

Despite the importance of the Lyapunov characteristic numbers in distinguishing between order and chaos, their practical application is limited by the very long calculations that are usually required for their evaluation. In fact, if we calculate the finite time LCN

$$\chi = \frac{\ln |\xi/\xi_0|}{t} \tag{26}$$

the variation of χ is irregular for relatively small t and only for large t the value of χ of a chaotic orbit stabilizes and tends to a constant limiting value (curve (2) of Fig. 10), which is the Lyapunov characteristic number

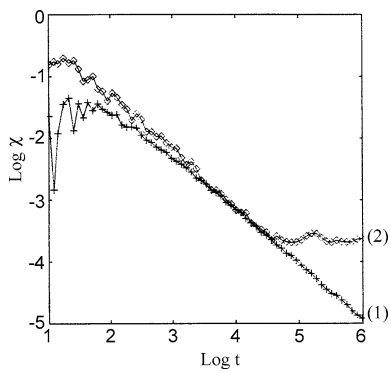

Astrophysics: Dynamical Systems, Figure 10
The variation of $\chi(t)$ of (1) an ordered orbit, and (2) a chaotic orbit

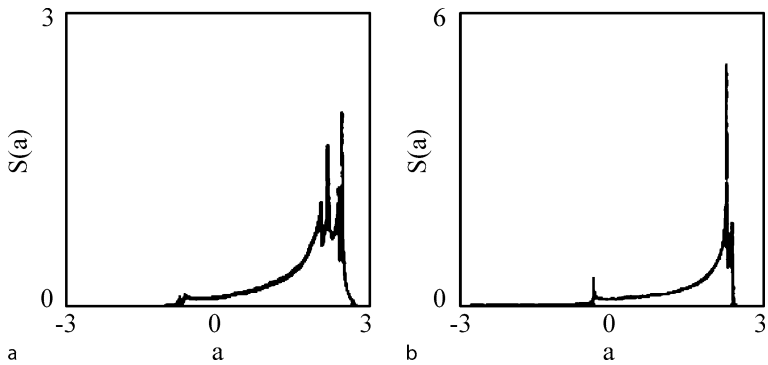

Astrophysics: Dynamical Systems, Figure 11
The spectra of stretching numbers: a of the standard map and b of the Hénon map

LCN $= \lim_{t\to\infty} \chi$. If, on the other hand, χ varies approximately as $1/t$, the LCN is zero (ordered orbit (1) of Fig. 10).

Of special interest is to find a finite time LCN after the shortest possible time. In the case of a map the shortest time is one iteration ($t = 1$). Thus one finds the quantities

$$a_i = \ln |\xi_{i+1}/\xi_i| \tag{27}$$

which are called Lyapunov indicators by Froeschlé et al. [38], or stretching numbers by Voglis and Contopoulos [73].

The distribution of successive values of the stretching numbers a_i along an orbit is called the "spectrum of stretching numbers". Namely the spectrum gives the proportion $\mathrm{d}N/N$ of the values of a_i in a given interval $(a, a + \mathrm{d}a)$ divided by $\mathrm{d}a$, i. e.

$$S(a) = \mathrm{d}N/(N\mathrm{d}a) \tag{28}$$

where $\mathrm{d}a$ is a small quantity and N a large number of iterations.

The main properties of the spectra of stretching numbers are [73]:

(1) The spectrum is invariant along the same orbit.
(2) The spectrum does not depend on the initial deviation $\vec{\xi}_0$ from the same initial point in the case of two dimensional maps.
(3) The spectrum does not depend on the initial conditions of orbits in the same connected chaotic domain.
(4) In the case of ordered orbits in two dimensional maps the spectrum does not depend on the initial conditions on the same invariant curve.

The Lyapunov characteristic number is equal to LCN $= \int S(a)\, a\, \mathrm{d}a$. The spectrum gives much more information about a system than LCN. E. g. the standard map

$$x' = x + y', \quad y' = y + \frac{K}{2\pi} \sin 2\pi x \pmod 1 \tag{29}$$

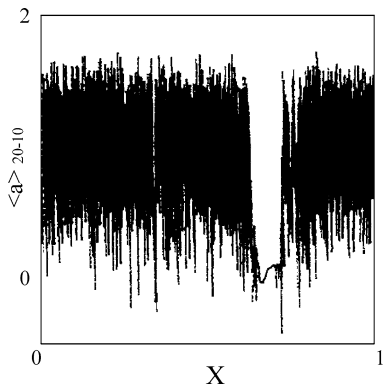

Astrophysics: Dynamical Systems, Figure 12
The average value of the stretching number, $\langle\alpha\rangle$, for 10 iterations (beyond the first 10 transients) as a function of *x* for constant *y*

for $K = 10$ and the Hénon map (23) for $K' = 7.407$ and $b = 1$ give two apparently random distributions of points with the same LCN $= 1.62$. However, their spectra are quite different (Fig. 11).

The use of the stretching numbers allows a fast distinction between ordered and chaotic orbits. E.g. while the calculation of the LCN requires $10^5 - 10^6$ iterations, in Fig. 12 we can separate the chaotic from the ordered orbits after only 20 iterations. In Fig. 12 the noise is large, nevertheless the chaotic orbits have roughly the same value of LCN $= \langle a\rangle > 0$, while the ordered orbits have $\langle a\rangle$ very close to zero. Further details about this topic can be found in the book of Contopoulos [21].

Application: Order and Chaos in Galaxies

Galactic Orbits

Galaxies are composed of stars, gas (including dust), and dark matter. The stars and the dark matter produce the

Astrophysics: Dynamical Systems, Figure 13
The basic types of Galaxies a M59/NGC 4621 Elliptical [E], b M51/NGC 5194 normal spiral [Sb] and c NGC1300 barred spiral [SBb]

main part of the galactic potential and force, while the contribution of the gas in the potential and force is small.

In the Hubble classification of galaxies, the elliptical galaxies (E) and the early type spiral galaxies (Sa, Sb) and barred galaxies (SBa, SBb) have less gas than the late-type galaxies (Sc, SBc) and the irregular galaxies (I) (Fig. 13).

Galaxies that have a well developed spiral structure (usually two symmetric spiral arms) are called "grand design", while galaxies with irregular and multiple fragments of spirals are called "flocculent".

The study of stellar (or mainly stellar) galaxies is based on a systematic exploration of their orbits.

The exploration of orbits in galaxies is very important, because the orbits are needed in constructing self-consistent models of galaxies.

Self-consistency is a new type of problem that does not appear, e. g. in accelerators, or in the solar system. It requires the construction of appropriate sets of orbits of stars, such that their superposition gives a response density that matches the imposed density of the model. Such a construction is done in many cases by taking a grid of initial conditions and calculating the orbits numerically in a computer and then populating these orbits with stars. But it is much more helpful and illuminating if one studies the types of orbits that form the building blocks of the galaxies.

Thus, the first step in understanding the structure and dynamics of a galaxy is to calculate its orbits, periodic, quasi-periodic and chaotic.

The most important orbits in a galactic system are the periodic orbits. The stable orbits trap around them sets of quasi-periodic orbits, that give the main features of the galaxy. On the other hand the unstable orbits separate the various types of ordered orbits, and also characterize the chaotic domains in a galaxy.

The simplest galactic orbits are the circular periodic orbits on the plane of symmetry of an axisymmetric galaxy. The circular orbits in such a galaxy are in general stable.

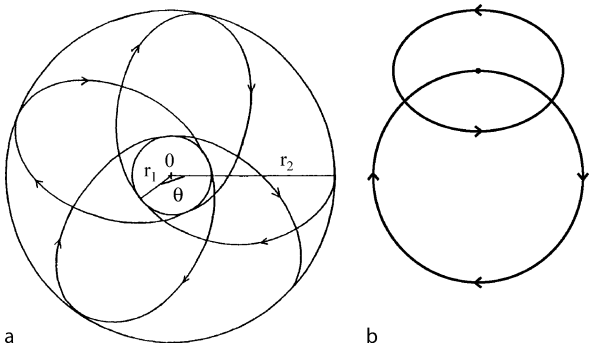

Astrophysics: Dynamical Systems, Figure 14
a A rosette orbit in an axisymmetric galaxy. b The corresponding epicyclic orbit

Orbits close to the stable circular periodic orbits are called epicyclic orbits. Such orbits fill circular rings in the plane of symmetry of the galaxy and are called "rosette" orbits (Fig. 14a). In a frame of reference rotating with the angular velocity (frequency) Ω of the center of the epicycle (epicenter, or guiding center) these orbits are closed and they are approximately ellipses around the center of the epicycle (Fig. 14b).

If V_0 is the axisymmetric potential, we have $\Omega^2 = V_0'/r$ and $\kappa^2 = V_0'' + 3V_0'/r$. The frequency κ along the epicycle is called "epicyclic frequency".

In most cases (spiral or barred galaxies) the frame of reference is not inertial, but rotates with angular velocity Ω_s. This is called "pattern velocity". In this rotating frame the form of the galaxy is stationary. Then the two basic frequencies of a moving star are $(\Omega - \Omega_s)$ and κ. If the ratio $\kappa/(\Omega - \Omega_s) = (n/m)$ is rational, then we have a resonant periodic orbit in the rotating system. Two most important resonances are the Lindblad resonances $\kappa/(\Omega - \Omega_s) = \pm 2/1$ (+ inner Lindblad resonance, ILR, − outer Lindblad resonance, OLR). A third most important resonance is corotation (or particle resonance), where $\Omega = \Omega_s$. The angular velocity Ω of an axisymmet-

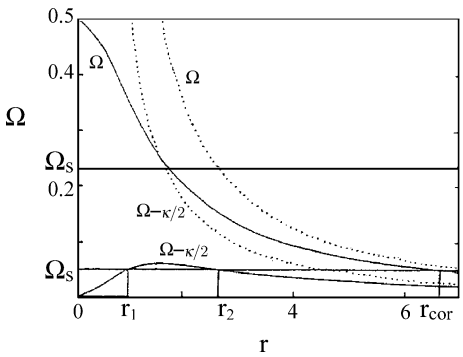

Astrophysics: Dynamical Systems, Figure 15
The curves Ω and $\Omega - \kappa/2$ in two cases: (1) with only one ILR (*dotted*) and (2) two ILR's (*solid*), for two different values of the pattern velocity Ω_s

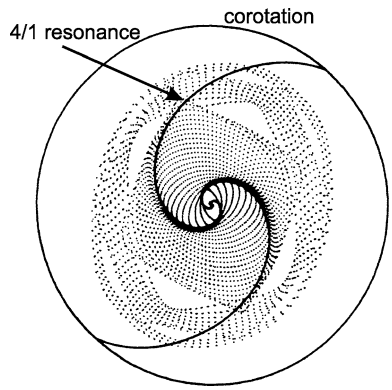

Astrophysics: Dynamical Systems, Figure 16
Periodic orbits in a spiral galaxy support the spiral up to the 4/1 resonance, but not beyond it

ric family is a function of the distance r. Ω is a monotonically decreasing function of r (Fig. 15).

Corotation and the inner and outer Lindblad resonances appear at the intersections of the line $\Omega = \Omega_s$ (Fig. 15) with the curves Ω, $\Omega - \kappa/2$ and $\Omega + \kappa/2$. There is only one distance corresponding to corotation and one distance corresponding to the outer Lindblad resonance OLR. However, depending on the model used, we may have one inner Lindblad resonance, or two ILRs, or no ILR at all.

The study of nonperiodic orbits on the plane of symmetry of a galaxy is realized by means of a surface of section.

The invariant curves of the ordered nonperiodic orbits surround either the point x_1 or the point x_4, that represent direct or retrograde periodic orbits that are reduced to circular orbits in the circular case.

The periodic orbits of type x_1 in the case of a spiral are like ellipses close to the inner Lindblad resonance (Fig. 16), but become like squares near the 4/1 resonance. These orbits and the nonperiodic orbits around them support the spiral up to the 4/1 resonance. However, beyond the 4/1 resonance these orbits are out of phase and do not support the spiral (Fig. 16). Thus, self-consistent spirals should terminate near the 4/1 resonance [20]. On the other hand, in strong bars we may have spirals outside corotation (see Sect. "Chaotic Orbits"). More details about orbits in galaxies can be found in [63] and [21].

Integrals of Motion in Spiral and Barred Galaxies

The Hamiltonian on the plane of symmetry of a galaxy (spiral or barred) rotating with angular velocity Ω_s is

$$H = \frac{\dot{r}^2}{2} + \frac{J_0^2}{2r^2} + V_0(r) - \Omega_s J_0 + V_1 \qquad (30)$$

where $\dot{r}$ is the radial velocity, J_0 the angular momentum, $V_0(r)$ the axisymmetric potential and V_1 the spiral (or barred) perturbation. The value of H is conserved (Jacobi constant). The Hamiltonian in action-angle variables is of the form $H = H_0 + V_1$, where

$$H_0 = \omega_1 I_1 + \omega_2 I_2 + a I_1^2 + 2b I_1 I_2 + c I_2^2 + \ldots \qquad (31)$$

and

$$\begin{aligned} V_1 &= \mathrm{Re}\,\{A(r)\,\exp[i(\Phi(r) - 2\theta)]\} \\ &= \mathrm{Re} \sum_{mn} V_{mn}(I_1, I_2) \exp[i(m\theta_1 - n\theta_2)] \end{aligned} \qquad (32)$$

The expansion in action-angle variables is found by expanding A, Φ and θ in terms of two angles, θ_1 (epicyclic angle) and θ_2 (azimuth of the epicenter measured from a certain direction), and $r - r_c = (2I_1/\omega_1)^{1/2}$, where r_c is the radius of a circular orbit that has the same Jacobi constant as the real orbit.

The main terms of V_1 are of order A and contain the trigonometric terms $\cos(\theta_1 - 2\theta_2)$, $\cos(\theta_1 + 2\theta_2)$ and $\cos 2\theta_2$. Away from the main resonances of the galaxy (inner and outer Lindblad resonances and corotation) it is possible to eliminate these terms by means of a canonical transformation that gives V_1 as a function of new action variables I_1^*, I_2^*. Thus, in this case the Hamiltonian H is a function of I_1^*, I_2^* only, i. e. I_1^* and I_2^* are integrals of motion.

The usual density wave theory of spiral structure deals with the forms of the integrals I_1^*, I_2^*, truncated after the first order terms.

However, the transformation from I_i to I_i^* contains denominators of the forms $\omega_1 - 2\omega_2$, $\omega_1 + 2\omega_2$ or ω_2, which

tend to zero close to the inner and outer Lindblad resonances and corotation respectively. For example, near the inner Lindblad resonance we cannot eliminate the term with $\cos(\theta_1 - 2\theta_2)$, because the transformation would have the denominator $\omega_1 - 2\omega_2$ which is close to zero. In this case we eliminate only the terms with $\cos(\theta_1 + 2\theta_2)$ and $\cos 2\theta_2$. Instead of eliminating $\cos(\theta_1 - 2\theta_2)$, we write this term as $\cos\psi_1$, introducing now resonant action-angle variables $J_1 = I_1^*$, $J_2 = I_2^* + 2I_1^*$, $\psi_1 = \theta_1^* - 2\theta_2^*$, $\psi_2 = \theta_2^*$. Then H is expressed in terms of J_1, J_2 and ψ_1, but does not contain ψ_2. Therefore, the conjugate action J_2 is an integral of motion. Thus, we have again two integrals of motion, namely J_2 and the Jacobi constant H.

The resonant form of H explains the forms of the orbits near the inner Lindblad resonance. In particular, near this resonance we have not one but two periodic orbits roughly perpendicular to each other.

The resonant theory is an extension of the linear density wave theory and it is applicable near the ILR. In a similar way we find the forms of the integrals near the outer Lindblad resonance and near corotation.

The three resonant forms of the Hamiltonian tend to the nonresonant form away from all resonances. In fact, if the amplitude of the spiral, or the bar, is small, the resonant effects are restricted in small regions around each resonance. However, if the amplitude of a spiral, or a bar, is large, we have an overlapping of resonances and the second integral is no more applicable.

Chaotic Orbits

Chaos in galaxies is always present near corotation, because this region contains an infinite number of higher order resonances that interact with each other.

Besides corotation, chaos may appear near the center of the galaxy when there is a large mass concentration (e. g. a central black hole). Furthermore large scale chaos appears when the amplitude of the spiral, or the bar, is large. Appreciable chaos has been found in N-body simulations of galaxies [27,58,68].

Chaos in barred galaxies near corotation appears around the unstable Lagrangian points L_1 and L_2. Around these points there are short period unstable periodic orbits for values of the Jacobi constant larger than the value $H(L_1)$ appropriate for L_1. Every unstable periodic orbit, PL_1, PL_2 of this family has a stable (S etc) and an unstable (U etc) manifold attached to it. Orbits near L_1 or L_2 follow the unstable manifolds of the corresponding orbits PL_1 and PL_2 and form trailing spiral arms along U and U' (Fig. 17), which are clearly seen in strong bars [22,74].

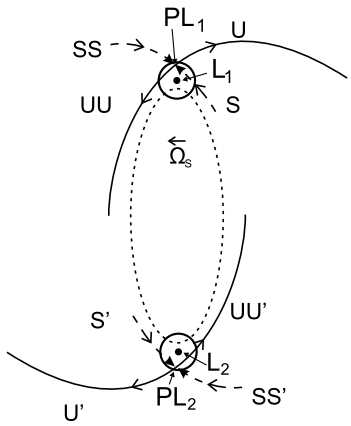

Astrophysics: Dynamical Systems, Figure 17
Orbits starting close to the unstable orbits PL_1, PL_2 move towards PL_1, PL_2 along the stable asymptotic curves S, SS and S', SS' and away from PL_1, PL_2 along the unstable asymptotic curves U, UU and U', UU'

Thus, in strong bars chaos is important in explaining the outer spiral arms.

Future Directions

The subject of dynamical systems in astrophysics is progressing along three main directions: (1) Exploration of further basic phenomena, especially in systems of more than two degrees of freedom, (2) Rigorous mathematical theorems, and (3) Applications on various problems of astrophysics.

(1) Much exploratory work exists now on systems of two degrees of freedom, but relatively little work has been done with systems of three or more degrees of freedom. In particular it is important to find the applicability of Arnold diffusion in various cases. A better separation of ordered and chaotic domains in phase-space is also necessary. Extensions of this work to find ordered and chaotic domains in quantum mechanical systems is also important. ([34] and references therein). Finally N-body problems with large N are also of great interest, especially as regards the validity and applications of statistical mechanics. In particular problems connected with the evolution of dynamical systems, are quite important.

(2) Many problems that have been explored numerically up to now require rigorous mathematical proofs. For example, the use of formal integrals of motion, the application of basic theorems, like the KAM and the Nekhorosev theorems and the applicability of many statistical methods require a better mathematical foundation.

(3) Presently much is being done with astrophysical applications of the theory of dynamical systems. For example, galactic dynamics has experienced an important expansion in recent years, with the exploration of the role of chaos in galaxies, especially in the formation and evolution of the outer spiral arms, and in the evolution of the central black holes. There are many other astrophysical problems where order and chaos play an important role, like the structure and evolution of stars, stellar variability, solar and stellar activity, the role of magnetic fields in stars and galaxies, the properties of extrasolar planetary systems, and the evolution of the whole Universe.

As regards dynamical astronomy a list of 72 problems of current interest that require further serious analytical and numerical work (possible theses topics) is provided at the end of the book of Contopoulos [21], together with many references.

Bibliography

1. Argyris J, Faust G, Haase M (1994) An Exploration of Chaos. North Holland, Amsterdam
2. Arnold VI (1961) Sov Math Dokl 2:245; (1963) Russ Math Surv 18(5):9
3. Arnold VI (1964) Sov Math Dokl 5:581
4. Aubry S (1978) In: Bishop AR, Schneider T (eds) Solitons and Condensed Matter Physics. Springer, New York, p 264
5. Birkhoff GD (1927) Dynamical Systems. Am Math Soc, Providence
6. Birkhoff GD (1931) Proc Nat Acad Sci 17:656
7. Bishop JL (1987) Astrophys J 322:618
8. Chapront-Touzé M, Henrard J (1980) Astron Astrophys 86:221
9. Cherry TM (1924) Mon Not R Astr Soc 84:729
10. Chirikov BV (1979) Phys Rep 52:263
11. Chirikov BV, Keil E, Sessler AM (1971) J Stat Phys 3:307
12. Collet P, Eckmann J-P (1980) Iterated Maps on the Interval as Dynamical Systems. Birkhäuser, Boston
13. Contopoulos G (1958) Stockholm Ann 20(5)
14. Contopoulos G (1960) Z Astrophys 49:273
15. Contopoulos G (1963) Astron J 68:763
16. Contopoulos G (ed) (1966) The Theory of Orbits in the Solar System and in Stellar Systems. IAU Symp 25. Academic Press, London
17. Contopoulos G (1967) In: Hénon M, Nahon F (eds) Les Nouvelles Méthodes de la Dynamique Stellaire. Bull Astron 2(3):223
18. Contopoulos G (1971) Astron J 76:147
19. Contopoulos G (1983) Lett Nuovo Cim 37:149
20. Contopoulos G (1985) Comments Astrophys 11:1
21. Contopoulos G (2002) Order and Chaos in Dynamical Astronomy. Springer, New York; Reprinted 2004
22. Contopoulos G (2008) In: Contopoulos G, Patsis P (eds) Chaos in Astronomy. Springer, New York, p 3
23. Contopoulos G, Harsoula M (2008) Int J Bif Chaos, in press
24. Contopoulos G, Vandervoort P (1992) Astrophys J 389:118
25. Contopoulos G, Voglis N (1996) Cel Mech Dyn Astron 64:1
26. Contopoulos G, Zikides M (1980) Astron Astrophys 90:198
27. Contopoulos G, Efthymiopoulos C, Voglis N (2000) Cel Mech Dyn Astron 78:243
28. Coullet P, Tresser C (1978) J Phys 39:C5–25
29. Delaunay C (1867) Theorie du Mouvement de la Lune. Paris
30. Deprit A, Henrard J, Rom A (1971) Astron J 76:273
31. de Zeeuw T (1985) Mon Not R Astr Soc 216:273
32. de Zeeuw T, Hunter C, Schwarzschild M (1987) Astrophys J 317:607
33. Eddington AS (1915) Mon Not R Astr Soc 76:37
34. Efthymiopoulos C, Contopoulos G (2006) J Phys A 39:1819
35. Efthymiopoulos C, Contopoulos G, Voglis N, Dvorak R (1997) J Phys A 30:8167
36. Efthymiopoulos C, Giorgilli A, Contopoulos G (2004) J Phys A 37:1831
37. Feigenbaum M (1978) J Stat Phys 19:25
38. Froeschlé C, et al. (1993) Cel Mech Dyn Astron 56:307
39. Grassberger P, Procaccia I (1983) Phys D 9:189
40. Greene JM et al. (1981) Phys D 3:468
41. Heggie DC (1983) Cel Mech 29:207
42. Hénon M (1969) Quart J Appl Math 27:291
43. Hénon M (1976) Commun Math Phys 50:69
44. Hénon M, Heiles C (1964) Astron J 69:73
45. Hunter C (1988) In: Buchler et al (eds) Integrability in Dynamical Systems. N Y Acad Sci Ann 536:25
46. Hunter C (1990) In: Buchler et al (eds) Galactic Models. N Y Acad Ann 596:187
47. Kolmogorov AN (1954) Dokl Akad Nauk SSSR 98:527
48. Landau LD, Lifshitz EM (1960) Mechanics, 1st edn; 3rd edn (1976) Pergamon Press, Oxford
49. Laskar J (1993) Cel Mech Dyn Astron 56:191
50. Lichtenberg AJ, Lieberman MA (1992) Regular and Chaotic Dynamics, 2nd edn. Springer, New York
51. Lindblad PO (1960) Stockholm Ann 21:3; 21:4
52. Lorentz EN (1963) J Atmos Sci 20:130
53. Lynden-Bell D (1962) Mon Not R Astr Soc 124:1
54. Lynden-Bell D (1998) In: Buchler et al (eds) Nonlinear Dynamics and Chaos in Astrophysics. NY Acad Sci Ann 867:3
55. Mather JN (1982) Erg Theory Dyn Syst 2:397; (1982) Topology 21:457
56. May RM (1976) Nature 261:459
57. Menjuk CR (1985) Phys Rev A 31:3282
58. Merritt D, Valluri M (1999) Astron J 118:1177
59. Moser J (1962) Nachr Acad Wiss Göttingen II. Math Phys Kl:1
60. Moser J (1967) Math Ann 169:136
61. Moser J (1968) Mem Amer Math Soc 81:1
62. Nekhoroshev NN (1977) Russ Math Surv 32(6):1
63. Ollongren A (1965) In: Blaauw A et al (eds) Stars and Stellar Systems V Galactic Structure. Univ of Chicago Press, Chicago, p 501
64. Percival IC (1979) In: Month M, Herrera JC (eds) Nonlinear Dynamics and the Beam-Beam Interaction. Am Inst Phys, New York, p 302
65. Pinotsis A (1988) In: Roy AE (ed) Long Term Behaviour of Natural and Artificial N-Body Systems. Kluwer, Dordrecht, p 465
66. Poincaré H (1892) Les Méthodes Nouvelles de la Mécanique Céleste Gauthier Villars. Paris I (1892); Paris II (1893); Paris III (1899); Dover edn (1957)
67. Rosenbluth MN, Sagdeev RA, Taylor JB, Zaslavsky GM (1966) Nucl Fusion 6:217
68. Schwarzschild M (1993) Astrophys J 409:563

69. Shenker SJ, Kadanoff LP (1982) J Stat Phys 27:631
70. Siegel CL (1956) Vorlesungen über Himmelsmechanik. Springer, Berlin
71. Stäckel P (1890) Math Ann 35:91
72. Statler TS (1987) Astrophys J 321:113
73. Voglis N, Contopoulos G (1994) J Phys A 27:4899
74. Voglis N, Tsoutsis P, Efthymiopoulos C (2006) Mon Not R Astr Soc 373:280
75. Walker GH, Ford J (1969) Phys Rev 188:416
76. Weinacht J (1924) Math Ann 91:279
77. Whittaker ET (1916) Proc Roy Soc Edinburgh 37:95
78. Whittaker ET (1937) A Treatise on the Analytical Dynamics of Particles and Rigid Bodies, 4th edn. Cambridge Univ Press, Cambridge
79. Wintner A (1947) Analytical Foundations of Celestial Mechanics. Princeton Univ Press, Princeton

B

Bacterial Computing

Martyn Amos
Department of Computing and Mathematics,
Manchester Metropolitan University, Manchester, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Motivation for Bacterial Computing
The Logic of Life
Rewiring Genetic Circuitry
Successful Implementations
Future Directions
Bibliography

Glossary

DNA Deoxyribonucleic acid. Molecule that encodes the genetic information of cellular organisms.

Operon Set of functionally related genes with a common promoter ("on switch").

Plasmid Small circular DNA molecule used to transfer genes from one organism to another.

RNA Ribonucleic acid. Molecule similar to DNA, which helps in the conversion of genetic information to proteins.

Transcription Conversion of a genetic sequence into RNA.

Translation Conversion of an RNA sequence into an amino acid sequence (and, ultimately, a protein).

Definition of the Subject

Bacterial computing is a conceptual subset of *synthetic biology*, which is itself an emerging scientific discipline largely concerned with the *engineering* of biological systems. The goals of synthetic biology may be loosely partioned into four sets: (1) To better understand the fundamental operation of the biological system being engineered, (2) To extend synthetic chemistry, and create improved systems for the synthesis of molecules, (3) To investigate the "optimization" of existing biological systems for human purposes, (4) To develop and apply rational *engineering* principles to the design and construction of biological systems. It is on these last two goals that we focus in the current article.

The main benefits that may accrue from these studies are both theoretical and practical; the construction and study of synthetic biosystems could improve our quantitative understanding of the fundamental underlying processes, as well as suggesting plausible applications in fields as diverse as pharmaceutical synthesis and delivery, biosensing, tissue engineering, bionanotechnology, biomaterials, energy production and environmental remediation.

Introduction

Complex natural processes may often be described in terms of networks of computational components, such as Boolean logic gates or artificial neurons. The interaction of biological molecules and the flow of information controlling the development and behavior of organisms is particularly amenable to this approach, and these models are well-established in the biological community. However, only relatively recently have papers appeared proposing the use of such systems to perform useful, *human-defined* tasks. For example, rather than merely using the network analogy as a convenient technique for clarifying our understanding of complex systems, it may now be possible to harness the power of such systems for the purposes of computation.

Despite the relatively recent emergence of biological computing as a distinct research area, the link between biology and computer science is not a new one. Of course, for years biologists have used computers to store and analyze experimental data. Indeed, it is widely accepted that the huge advances of the Human Genome Project (as well

as other genome projects) were only made possible by the powerful computational tools available. Bioinformatics has emerged as "the science of the 21st century", requiring the contributions of truly interdisciplinary scientists who are equally at home at the lab bench or writing software at the computer.

However, the seeds of the relationship between biology and computer science were sown over fifty years ago, when the latter discipline did not even exist. When, in the 17th century, the French mathematician and philosopher René Descartes declared to Queen Christina of Sweden that animals could be considered a class of machines, she challenged him to demonstrate how a clock could reproduce. Three centuries later in 1951, with the publication of *"The General and Logical Theory of Automata"* [38] John von Neumann showed how a machine could indeed construct a copy of itself. Von Neumann believed that the behavior of *natural* organisms, although orders of magnitude more complex, was similar to that of the most intricate machines of the day. He believed that life was based on *logic*. We now begin to look at how this view of life may be used, not simply as a useful analogy, but as the practical foundation of a whole new engineering discipline.

Motivation for Bacterial Computing

Here we consider the main motivations behind recent work on bacterial computing (and, more broadly, synthetic biology). Before recombinant DNA technology made it possible to construct new genetic sequences, biologists were restricted to crudely "knocking out" individual genes from an organism's genome, and then assessing the damage caused (or otherwise). Such knock-outs gradually allowed them to piece together fragments of causality, but the process was very time-consuming and error-prone.

Since the dawn of genetic engineering – with the ability to synthesize novel gene segments – biologists have been in a position to make much more finely-tuned modifications to their organism of choice, this generating much more refined data. Other advances in biochemistry have also contributed, allowing scientists to – for example – investigate new types of genetic systems with, for example, *twelve* bases, rather than the traditional four [14]. Such creations have yielded valuable insights into the mechanics of mutation, adaptation and evolution. Researchers in synthetic biology are now extending their work beyond the synthesis of single genes, and are now introducing whole new gene *complexes* into organisms.

The objectives behind this work are both theoretical and practical. As Benner and Seymour argue [5], " … a synthetic goal forces scientists to cross uncharted ground to encounter and solve problems that are not easily encountered through [top-down] analysis. This drives the emergence of new paradigms ["world views"] in ways that analysis cannot easily do." Drew Endy agrees. "Worst-case scenario, it's a complementary approach to traditional discovery science" [18]. "Best-case scenario, we get systems that are simpler and easier to understand … " Or, to put it bluntly, "Let's build new biological systems – systems that are easier to understand because we made them that way" [31]. As well as shedding new light on the underlying biology, these novel systems may well have significant practical utility. Such new creations, according to Endy's "personal wish list" might include "generating biological machines that could clean up toxic waste, detect chemical weapons, perform simple computations, stalk cancer cells, lay down electronic circuits, synthesize complex compounds and even produce hydrogen from sunlight" [18]. In the next Section we begin to consider how this might be achieved, by first describing the underlying logic of genetic circuitry.

The Logic of Life

Twenty years after von Neumann's seminal paper, Francois Jacob and Jacques Monod identified *specific* natural processes that could be viewed as behaving according to logical principles:

> "The logic of biological regulatory systems abides not by Hegelian laws but, like the workings of computers, by the propositional algebra of George Boole." [29]

This conclusion was drawn from earlier work of Jacob and Monod [30]. In addition, Jacob and Monod described the "lactose system" [20], which is one of the archetypal examples of a Boolean biosystem. We describe this system shortly, but first give a brief introduction to the operation of genes in general terms.

DNA as the Carrier of Genetic Information

The *central dogma* of molecular biology [9] is that DNA produces RNA, which in turn produces proteins. The basic "building blocks" of genetic information are known as *genes*. Each gene codes for one specific *protein* and may be turned on (*expressed*) or off (*repressed*) when required.

Transcription and Translation

We now describe the processes that determine the structure of a protein, and hence its function. Note that in what follows we assume the processes described occur in bacte-

ria, rather than in higher organisms such as humans. For a full description of the structure of the DNA molecule, see the chapter on *DNA computing*. In order for a DNA sequence to be converted into a protein molecule, it must be read (*transcribed*) and the transcript converted (*translated*) into a protein. Transcription of a gene produces a *messenger RNA* (mRNA) copy, which can then be translated into a protein.

Transcription proceeds as follows. The mRNA copy is synthesized by an enzyme known as *RNA polymerase*. In order to do this, the RNA polymerase must be able to recognize the specific region to be transcribed. This specificity requirement facilitates the regulation of genetic expression, thus preventing the production of unwanted proteins. Transcription begins at specific sites within the DNA sequence, known as *promoters*. These promoters may be thought of as "markers", or "signs", in that they are not transcribed into RNA. The regions that *are* transcribed into RNA (and eventually translated into protein) are referred to as *structural* genes. The RNA polymerase recognizes the promoter, and transcription begins. In order for the RNA polymerase to begin transcription, the double helix must be opened so that the sequence of bases may be read. This opening involves the breaking of the hydrogen bonds between bases. The RNA polymerase then moves along the DNA *template* strand in the $3 \rightarrow 5'$ direction. As it does so, the polymerase creates an *antiparallel* mRNA chain (that is, the mRNA strand is the equivalent of the Watson-Crick complement of the template). However, there is one significant difference, in that RNA contains uracil instead of thymine. Thus, in mRNA terms, "*U* binds with *A*."

The RNA polymerase moves along the DNA, the DNA re-coiling into its double-helix structure behind it, until it reaches the end of the region to be transcribed. The end of this region is marked by a *terminator* which, like the promoter, is not transcribed.

Genetic Regulation

Each step of the conversion, from stored information (DNA), through mRNA (messenger), to protein synthesis (effector), is itself catalyzed by other effector molecules. These may be enzymes or other factors that are required for a process to continue (for example, sugars). Consequently, a loop is formed, where products of one gene are required to produce further gene products, and may even influence that gene's own expression. This process was first described by Jacob and Monod in 1961 [20], a discovery that earned them a share of the 1965 Nobel Prize in Physiology or Medicine.

Genes are composed of a number of distinct regions, which control and encode the desired product. These regions are generally of the form promoter–gene–terminator. Transcription may be regulated by effector molecules known as *inducers* and *repressors*, which interact with the promoter and increase or decrease the level of transcription. This allows effective control over the expression of proteins, avoiding the production of unnecessary compounds. It is important to note at this stage that, in reality, genetic regulation does not conform to the digital "on-off" model that is popularly portrayed; rather, it is continuous or analog in nature.

The *Lac* Operon

One of the most well-studied genetic systems is the *lac operon*. An *operon* is a set of functionally related genes with a common promoter. An example of this is the *lac* operon, which contains three structural genes that allow *E. coli* to utilize the sugar lactose.

When *E.coli* is grown on the sugar glucose, the product of the (separate, and unrelated to the *lac* operon) *lacI* gene *represses* the transcription of the *lacZYA* operon (i. e., the operon is turned off). However, if lactose is supplied *together with* glucose, a lactose by-product is produced which interacts with the repressor molecule, preventing it from repressing the *lacZYA* operon. This de-repression does not itself initiate transcription, since it would be inefficient to utilize lactose if the more common sugar glucose were still available. The operon is *positively* regulated (i. e., "encouraged") by a different molecule, whose level increases as the amount of available glucose decreases. Therefore, if lactose were present as the sole carbon source, the *lacI* repression would be relaxed and the high "encouraging" levels would activate transcription, leading to the synthesis of the *lacZYA* gene products. Thus, the promoter is under the control of two sugars, and the *lacZYA* operon is only transcribed when lactose is *present* and glucose is *absent*.

In essence, Jacob and Monod showed how a gene may be thought of (in very abstract terms) as a binary switch, and how the state of that switch might be affected by the presence or absence of certain molecules. Monod's point, made in his classic book *Chance and Necessity* and quoted above, was that the workings of biological systems operate not by Hegel's philosophical or metaphysical logic of understanding, but according to the formal, mathematically-grounded logical system of George Boole.

What Jacob and Monod found was that the transcription of a gene may be regulated by molecules known as *inducers* and *repressors*, which either increase or decrease the

"volume" of a gene (corresponding to its level of transcription, which isn't always as clear cut and binary as Monod's quote might suggest). These molecules interact with the promoter region of a gene, allowing the gene's level to be finely "tuned". The *lac* genes are so-called because, in the *E. coli* bacterium, they combine to produce a variety of proteins that allow the cell to metabolise the sugar lactose (which is most commonly found in milk, hence the derivation from the Latin, *lact*, meaning milk).

For reasons of efficiency, these proteins should only be produced (i. e., the genes be turned on) when lactose is present in the cell's environment. Making these proteins when lactose is *absent* would be a waste of the cell's resources, after all. However, a different sugar – glucose – will *always* be preferable to lactose, if the cell can get it, since glucose is an "easier" form of sugar to metabolise. So, the input to and output from the *lac* operon may be expressed as a truth table, with G and L standing for glucose and lactose (1 if present, 0 if absent), and O standing for the output of the operon (1 if on, 0 if off):

G	L	O
0	0	0
0	1	1
1	0	0
1	1	0

The Boolean function that the *lac* operon therefore *physically* computes is (L AND (NOT G)), since it only outputs 1 if L=1 (lactose present) *and* G=0 (glucose is absent). By showing how one gene could affect the expression of another – just like a transistor feeds into the input of another and affects its state – Jacob and Monod laid the foundations for a new way of thinking about genes; not simply in terms of protein blueprints, but of *circuits* of interacting parts, or dynamic networks of interconnected switches and logic gates. This view of the genome is now well-established [22,23,27], but in the next Section we show how it might be used to guide the *engineering* of biological systems.

Rewiring Genetic Circuitry

A key difference between the wiring of a computer chip and the circuitry of the cell is that "electronics engineers know exactly how resistors and capacitors are wired to each other because they installed the wiring. But biologists often don't have a complete picture. They may not know which of thousands of genes and proteins are interacting at a given moment, making it hard to predict how circuits will behave inside cells" [12]. This makes the task of reengineering vastly more complex.

Rather than trying assimilate the huge amounts of data currently being generated by the various genome projects, synthetic biologists are taking a novel route – simplify and build. "They create models of genetic circuits, build the circuits, see if they work, and adjust them if they don't – learning about biology in the process. 'I view it as a reductionist approach to systems biology,' says biomedical engineer Jim Collins of Boston University" [12].

The field of *systems biology* has emerged in recent years as an alternative to the traditional reductionist way of doing science. Rather than simply focussing on a single level of description (such as individual proteins), researchers are now seeking to *integrate* information from many different layers of complexity. By studing how different biological components *interact*, rather than simply looking at their structure, systems biologists are attempting to build models of systems from the bottom up. A model is simply an abstract description of how a system operates – for example, a set of equations that describe how a disease spreads throughout a population. The point of a model is to capture the *essence* of a system's operation, and it should therefore be as simple as possible. Crucially, a model should also be capable of making *predictions*, which can then be tested against reality using real data (for example, if infected people are placed in quarantine for a week, how does this affect the progress of the disease in question, or what happens if I feed this signal into the chip?). The results obtained from these tests may then feed back in to further refinements of the model, in a continuous cycle of improvement.

When a model suggests a plausible structure for a synthetic genetic circuit, the next stage is to engineer it into the chosen organism, such as a bacterium. Once the structure of the DNA molecule was elucidated and the processes of transcription and translation were understood, molecular biologists were frustrated by the lack of suitable experimental techniques that would facilitate more detailed examination of the genetic material. However, in the early 1970s, several techniques were developed that allowed previously impossible experiments to be carried out (see [8,33]). These techniques quickly led to the first ever successful cloning experiments [19,25]. Cloning is generally defined as " … the production of multiple identical copies of a single gene, cell, virus, or organism." [35]. This is achieved as follows: a specific sequence (corresponding, perhaps, to a novel gene) is inserted in a circular DNA molecule, known as a *plasmid*, or *vector*, producing a *recombinant DNA molecule*. The vector acts as a *vehicle*, transporting the sequence into a *host* cell (usually a bacterium, such as *E.coli*). Cloning single genes is well-established, but is often done on an ad hoc basis. If biological

computing is to succeed, it requires some degree of *standardization*, in the same way that computer manufacturers build different computers, but using a standard library of components. "Biobricks are the first example of standard biological parts," explains Drew Endy [21]. "You will be able to use biobricks to program systems that do whatever biological systems do." He continues. "That way, if in the future, someone asks me to make an organism that, say, counts to 3,000 and then turns left, I can grab the parts I need off the shelf, hook them together and predict how they will perform" [15].

Each biobrick is a simple component, such as an AND gate, or an inverter (NOT). Put them together one after the other, and you have a NAND (NOT-AND) gate, which is all that is needed to build any Boolean circuit (an arbitrary circuit can be translated to an equivalent circuit that uses only NAND gates. It will be much bigger than the original, but it will compute the same function. Such considerations were important in the early stages of integrated circuits, when building *different* logic gates was difficult and expensive). Just as transistors can be used together to build logic gates, and these gates then combined into circuits, there exists a hierarchy of complexity with biobricks. At the bottom are the "parts", which generally correspond to coding regions for proteins. Then, one level up, we have "devices", which are built from parts – the oscillator of Elowitz and Leibler, for example, could be constructed from three inverter devices chained together, since all an inverter does is "flip" its signal from 1 to 0 (or vice versa). This circuit would be an example of the biobricks at the top of the conceptual tree – "systems", which are collections of parts to do a significant task (like oscillating or counting).

Tom Knight at MIT made the first 6 biobricks, each held in a plasmid ready for use. As we have stated, plasmids can be used to insert novel DNA sequences in the genomes of bacteria, which act as the "testbed" for the biobrick circuits. "Just pour the contents of one of these vials into a standard reagent solution, and the DNA will transform itself into a functional component of the bacteria," he explains [6]. Drew Endy was instrumental in developing this work further, one invaluable resource being the Registry of Standard Biological Parts [32], the definitive catalogue of new biological component. At the start of 2006, it contained 224 "basic" parts and 459 "composite" parts, with 270 parts "under construction". Biobricks are still at a relatively early stage, but "Eventually we'll be able to design and build *in silico* and go out and have things synthesized," says Jay Keasling, head of Lawrence Berkeley National Laboratory's new synthetic biology department [12].

Successful Implementations

Although important foundational work had been performed by Arkin and Ross as early as 1994 [2], the year 2000 was a particularly significant one for synthetic biology. In January two foundational papers appeared back-to-back in the same issue of *Nature*. "Networks of interacting biomolecules carry out many essential functions in living cells, but the 'design principles' underlying the functioning of such intracellular networks remain poorly understood, despite intensive efforts including quantitative analysis of relatively simple systems. Here we present a complementary approach to this problem: the design and construction of a synthetic network to implement a particular function" [11]. That was the introduction to a paper that Drew Endy would call "the high-water mark of a synthetic genetic circuit that does something" [12]. In the first of the two articles, Michael Elowitz and Stanislau Leibler (then both at Princeton) showed how to build microscopic "Christmas tree lights" using bacteria.

Synthetic Oscillator

In physics, an *oscillator* is a system that produces a regular, periodic "output". Familiar examples include a pendulum, a vibrating string, or a lighthouse. Linking several oscillators together in some way gives rise to *synchrony* – for example, heart cells repeatedly firing in unison, or millions of fireflies blinking on and off, seemingly as one [36].

Leibler actually had *two* articles published in the same high-impact issue of *Nature*. The other was a short communication, co-authored with Naama Barkai – also at Princeton, but in the department of Physics [3]. In their paper, titled "Circadian clocks limited by noise", Leibler and Barkai showed how a simple model of biochemical networks could oscillate reliably, even in the presence of noise. They argued that such oscillations (which might, for example, control the internal circadian clock that tells us when to wake up and when to be tired) are based on networks of *genetic regulation*. They built a simulation of a simple regulatory network using a *Monte Carlo* algorithm. They found that, however they perturbed the system, it still oscillated reliably, although, at the time, their results existed only *in silico*. The other paper by Leibler was much more applied, in the sense that they had constructed a biological circuit [11]. Elowitz and Leibler had succeeded in constructing an artificial genetic oscillator in cells, using a synthetic network of repressors. They called this construction the *repressilator*.

Rather than investigating *existing* oscillator networks, Elowitz and Leibler decided to build one entirely from first principles. They chose three repressor-promoter pairs that

had already been sequenced and characterized, and first built a mathematical model in software. By running the sets of equations, they identified from their simulation results certain molecular characteristics of the components that gave rise to so-called *limit cycle* oscillations; those that are robust to perturbations. This information from the model's results lead Elowitz and Leibler to select strong promoter molecules and repressor molecules that would rapidly decay. In order to implement the oscillation, they chose three genes, each of which affected one of the others by repressing it, or turning it off. For the sake of illustration, we call the genes A, B and C. The product of gene A turns off (represses) gene B. The absence of B (which represses C) allows C to turn on. C is chosen such that it turns gene A *off* again, and the three genes loop continuously in a "daisy chain" effect, turning on and off in a repetitive cycle. However, some form of *reporting* is necessary in order to confirm that the oscillation is occurring as planned.

Green fluorescent protein (GFP) is a molecule found occurring naturally in the jellyfish *Aequorea victoria*. Biologists find it invaluable because it has one interesting property – when placed under ultraviolet light, it *glows*. Biologists quickly sequenced the gene responsible for producing this protein, as they realized that it could have many applications as a *reporter*. By inserting the gene into an organism, you have a ready-made "status light" – when placed into bacteria, they glow brightly if the gene is turned on, and look normal if it's turned off. We can think of it in terms of a Boolean circuit – if the circuit outputs the value 1, the GFP promoter is produced to turn on the light. If the value is 0, the promoter isn't produced, the GFP gene isn't expressed, and the light stays off.

Elowitz and Leibler set up their gene network so that the GFP gene would be expressed whenever gene C was turned *off* – when it was turned on, the GFP would gradually decay and fade away. They synthesized the appropriate DNA sequences and inserted them into a *plasmid*, eventually yielding a population of bacteria that blinked on and off in a repetitive cycle, like miniature lighthouses. Moreover, and perhaps most significantly, the period between flashes was longer than the time taken for the cells to divide, showing that the state of the system had been passed on during reproduction.

Synthetic Toggle Switch

Rather than model an existing circuit and then altering it, Elowitz and Leibler had taken a "bottom up" approach to learning about how gene circuits operate. The other notable paper to appear in that issue was written by Timothy Gardner, Charles Cantor and Jim Collins, all of Boston University in the US. In 2000, Gardner, Collins and Cantor observed that genetic switching (such as that observed in the lambda phage [34]) had not yet been "demonstrated in networks of non-specialized regulatory components" [13]. That is to say, at that point nobody had been able to construct a switch out of genes that hadn't already been "designed" by evolution to perform that specific task. The team had a similar philosophy to that of Elowitz and Leibler, in that their main motivation was being able to test theories about the fundamental behaviour of gene regulatory networks. "Owing to the difficulty of testing their predictions," they explained, "these theories have not, in general, been verified experimentally. Here we have integrated theory and experiment by constructing and testing a synthetic, bistable [two-state] gene circuit based on the predictions of a simple mathematical model."

"We were looking for the equivalent of a light switch to flip processes on or off in the cell," explained Gardner [10]. "Then I realized a way to do this with genes instead of with electric circuits." The team chose two genes that were mutually inhibitory – that is, each produced a molecule that would turn the other off. One important thing to bear in mind is that the system didn't have a single input. Although the team acknowledged that bistability might be possible – in theory – using only a single promoter that regulated *itself*, they anticipated possible problems with robustness and experimental tunability if they used that approach. Instead, they decided to use a system whereby each "side" of the switch could be "pressed" by a different stimulus – the addition of a chemical on one, and a change in temperature on the other. Things were set up so that if the system was in the state induced by the chemical, it would stay in that state until the temperature was changed, and would only change *back* again if the chemical was reintroduced. Importantly, these stimuli did not have to be applied continuously – a "short sharp" burst was enough to cause the switch to flip over. As with the other experiment, Gardner and his colleagues used GFP as the system state reporter, so that the cells glowed in one state, and looked "normal" in the other.

In line with the bottom-up approach, they first created a mathematical model of the system and made some predictions about how it would behave inside a cell. Within a year, Gardner had spliced the appropriate genes into the bacteria, and he was able to flip them –at will – from one state the the other. As McAdams and Arkin observed, synthetic "one way" switches had been created in the mid-1980s, but "this is perhaps the first engineered design exploiting bistability to produce a switch with capability of *reversibly* switching between two … stable states" [28].

The potential applications of such a bacterial switch were clear. As they state in the conclusion to their article, "As a practical device, the toggle switch … may find applications in gene therapy and biotechnology." They also borrowed from the language of computer programming, using an analogy between their construction and the short "applets" written in the Java language, which now allow us to download and run programs in our web browser. "Finally, as a cellular memory unit, the toggle forms the basis for 'genetic applets' – self-contained, programmable, synthetic gene circuits for the control of cell function".

Engineered Communication

Towards the end of his life, Alan Turing did some foundational work on pattern formation in nature, in an attempt to explain how zebras get their striped coats or leopards their spots. The study of *morphogenesis* (from the Greek, *morphe* – shape, and *genesis* – creation. "Amorphous" therefore means "without shape or structure") is concerned with how cells split to assume new roles and communicate with another to form very precise shapes, such as tissues and organs. Turing postulated that the diffusion of chemical signals both within and between cells is the main driving force behind such complex pattern formation [37].

Although Turing's work was mainly concerned with the processes occurring amongst cells inside a developing embryo, it is clear that chemical signalling also goes on between bacteria. Ron Weiss of Princeton University was particularly interested in *Vibrio fischeri*, a bacterium that has a symbiotic relationship with a variety of aquatic creatures, including the Hawaiian squid. This relationship is due mainly to the fact that the bacteria exhibit *bioluminescence* – they generate a chemical known as a *Luciferase* (coded by the *Lux* gene), a version of which is also found in fireflies, and which causes them to glow when gathered together in numbers. Cells within the primitive light organs of the squid draw in bacteria from the seawater and encourage them to grow. Crucially, once enough bacteria are packed into the light organ they produce a signal to tell the squid cells to stop attracting their colleagues, and only then do they begin to glow. The cells get a safe environment in which to grow, protected from competition, and the squid has a light source by which to navigate and catch prey. The mechanism by which the *Vibrio* "know" when to start glowing is known as *quorum sensing*, since there have to be sufficient "members" present for luminiscence to occur.

The bacteria secrete an *autoinducer* molecule, known as VAI (*Vibrio* Auto Inducer), which diffuses through the cell wall. The *Lux* gene (which generates the glowing chemical) needs to be activated (turned on) by a particular protein – which attracts the attention of the polymerase – but the protein can only do this with help from the VAI. Its particular 3D structure is such that it can't fit tightly onto the gene unless it's been slightly bent out of shape. Where there's enough VAI present, it locks onto the protein and alters its conformation, so that it can turn on the gene. Thus, the concentration of VAI is absolutely crucial; once a critical threshold has been passed, the bacteria "know" that there are enough of them present, and they begin to glow.

Weiss realized that this quorum-based cell-to-cell communication mechanism could provide a powerful framework for the construction of bacterial devices – imagine, for example, a tube of solution containing engineered bacteria that can be added to a sample of seawater, causing it to glow only if the concentration of a particular pollutant exceeds a certain threshold. Crucially, as we will see shortly, it also allows the possibility of generating precise "complex patterned" development.

Weiss set up two colonies of *E. coli*, one containing "sender", and the other "receivers". The idea was that the senders would generate a chemical signal made up of VAI, which could diffuse across a gap and then be picked up by the receivers. Once a strong enough signal was being communicated, the receivers would glow using GFP to say that it had been picked up. Weiss cloned the appropriate gene sequences (corresponding to a type of biobrick) into his bacteria, placed colonies of receiver cells on a plate, and the receivers started to glow in acknowledgment.

Synthetic Circuit Evolution

In late 2002, Weiss and his colleagues published another paper, this time describing how rigorous engineering principles may be brought to bear on the problem of designing and building entirely new genetic circuitry. The motivation was clear – "biological circuit engineers will have to confront their inability to predict the precise behavior of even the most simple synthetic networks, a serious shortcoming and challenge for the design and construction of more sophisticated genetic circuitry in the future" [39].

Together with colleagues Yohei Yokobayashi and Frances Arnold, Weiss proposed a two stage strategy: first, design a circuit from the bottom up, as Elowitz and others had before, and clone it into bacteria. Such circuits are highly unlikely to work first time, "because the behavior of biological components inside living cells is highly context-dependent, the actual circuit performance will likely differ from the design predictions, often resulting in a poorly performing or nonfunctional circuit." Rather than simply

abandoning their design, Weiss and his team decided to then *tune* the circuit inside the cell itself, by applying the principles of evolution. By inducing mutations in the DNA that they had just introduced, they were able to slightly modify the behaviour of the circuit that it represented. Of course, many of these changes would be catastrophic, giving even worse performance than before, but, occasionally, they observed a minor improvement. In that case, they kept the "winning" bacteria, and subjected it to another round of mutation, in a repeated cycle. In a microcosmic version of Darwinian evolution, mutation followed by selection of the fittest took an initially unpromising pool of broken circuits and transformed them into winners. "Ron is utilizing the power of evolution to design networks in ways so that they perform exactly the way you want them to," observed Jim Collins [16]. In a commentary article in the same issue of the journal, Jeff Hasty called this approach "design then mutate" [17]. The team showed how a circuit made up of three genetic gates could be fine-tuned *in vivo* to give the correct performance, and they concluded that "the approach we have outlined should serve as a robust and widely applicable route to obtaining circuits, as well as new genetic devices, that function inside living cells."

Pattern Formation

The next topic studied by Weiss and his team was the problem of *space* – specifically, how to get a population of bacteria to cover a surface with a specific density. This facility could be useful when designing bacterial biosensors – devices that detect chemicals in the environment and produce a response. By controlling the density of the microbial components, it might be possible to tune the sensitivity of the overall device. More importantly, the ability for cells to control their own density would provide a useful "self-destruct" mechanism were these genetically-modified bugs ever to be released into the environment for "real world" applications.

In "Programmed population control by cell-cell communication and regulated killing" [40], Weiss and his team built on their previous results to demonstrate the ability to keep the density of an *E. coli* population artificially low – that is, below the "natural" density that could be supported by the available nutrients. They designed a genetic circuit that caused the bacteria to generate a different *Vibrio* signalling molecule, only this time, instead of making the cells glow, a sufficient concentration would flip a switch inside the cell, turning on a *killer gene*, encoding a protein that was toxic in sufficient quantities. The system behaved exactly as predicted by their mathematical model. The culture grew at an exponential rate (that is, doubling every time step) for seven hours, before hitting the defined density threshold. At that point the population dropped sharply, as countless cells expired, until the population settled at a steady density significantly (ten times) lower than an unmodified "control" colony. The team concluded that "The population-control circuit lays the foundations for using cell-cell communication to programme interactions among bacterial colonies, allowing the concept of communication -regulated growth and death to be extended to engineering synthetic ecosystems" [40].

The next stage was to programme cells to form *specific* spatial patterns in the dish. As we have already mentioned briefly, pattern formation is one characteristic of multicellular systems. This is generally achieved using some form of chemical signalling, combined with a differential response – that is, different cells, although genetically identical, may "read" the environmental signals and react in different ways, depending on their internal state. For example, one cell might be "hungry" and choose to move towards a food source, while an identical cell might choose to remain in the same spot, since it has adequate levels of energy.

The team used a variant of the sender-receiver model, only this time adding a "distance detection" component to the receiver circuit. The senders were placed in the center of the dish, and the receivers distributed uniformly across the surface. The receivers constructed so that they could measure the *strength* of the signal being "beamed" from the senders, a signal which decayed over distance (a little like a radio station gradually breaking up as you move out of the reception area). The cells were engineered so that only those that were either "near" to the senders or "far" from the senders would generate a response (those in the middle region were instructed to keep quiet). These cells are genetically identical, and are uniformly distributed over the surface – the differential response comes in the way that they assess the strength of the signal, and make a decision on whether or not to respond. The power of the system was increased further by making the near cells glow green, and those far away glow red (using a different fluorescent protein).

When the team set the system running, they observed the formation of a "dartboard" pattern, with the "bullseye" being the colony of senders (instructed to glow cyan, or light blue), which was surrounded by a green ring, which in turn was surrounded by a red ring. By placing three sender colonies in a triangle, they were also able to obtain a green heart-shaped pattern, formed by the intersection of three green circles, as well as other patterns, determined solely by the initial configuration of senders [4].

Bacterial Camera

Rather than generating light, a different team decided to use bacteria to *detect* light – in the process, building the world's first microbial camera. By engineering a dense bed of *E. coli*, a team of students led by Chris Voight at Berkeley developed light-sensitive "film" capable of storing images at a resolution of 100 megapixels per square inch. *E. coli* are not normally sensitive to light, so the group took genes coding for *photoreceptors* from blue-green algae, and spliced them into their bugs [24]. When light was shone on the cells, it turned on a genetic switch that cause a chemical inside them to permanently darken, thus generating a black "pixel". By projecting an image onto a plate of bacteria, the team were able to obtain several monochrome images, including the *Nature* logo and the face of team member Andrew Ellington. Nobel Laureate Sir Harry Kroto, discoverer of "buckballs", called the team's camera an "extremely exciting advance" [26], going on to say that "I have always thought that the first major nanotechnology advances would involve some sort of chemical modification of biology."

Future Directions

Weiss and his team team suggest that "the integration of such systems into higher-level organisms and with different cell functions will have practical applications in three-dimensional tissue engineering, biosensing, and biomaterial fabrication." [4] One possible use for such a system might lie in the detection of bio-weapons – spread a culture of bacteria over a surface, and, with the appropriate control circuit, they will be able to accurately pinpoint the location of any pathogens. Programmed cells could eventually replace artificial tissues, or even organs – current attempts to build such constructions in the laboratory rely on cells arranging themselves around an artificial scaffold. Controlled cellular structure formation could do away with the need for such support – "The way we're doing tissue engineering, right now, … is very unnatural," argues Weiss. "Clearly cells make scaffolds themselves. If we're able to program them to do that, we might be able to embed them in the site of injury and have them figure out for themselves what the pattern should be" [7]. In addition to building structures, others are considering engineering cells to act as miniature drug delivery systems – fighting disease or infection from the *inside*. Adam Arkin and Chris Voight are currently investigating the use of modified *E. coli* to battle against cancer tumours, while Jay Keasling and co-workers at Berkeley are looking at engineering circuits into the same bacteria to persuade them to generate a potent antimalarial drug that is normally found in small amounts in wormwood plants.

Clearly, bacterial computing/synthetic biology is still at a relatively early stage in its development, although the field is growing at a tremendous pace. It could be argued, with some justification, that the dominant science of the new millennium may well prove to be at the intersection of biology and computing. As biologist Roger Brent argues, "I think that synthetic biology … will be as important to the 21st century as [the] ability to manipulate bits was to the 20th" [1].

Bibliography

Primary Literature

1. Anon (2004) Roger Brent and the alpha project. ACM Ubiquity, 5(3)
2. Arkin A, Ross J (1994) Computational functions in biochemical reaction networks. Biophysical J 67:560–578
3. Barkai N, Leibler S (2000) Circadian clocks limited by noise. Nature 403:267–268
4. Basu S, Gerchman Y, Collins CH, Arnold FH, Weiss R (2005) A synthetic multicellular system for programmed pattern formation. Nature 434:1130–1134
5. Benner SA, Sismour M (2005) Synthetic biology. Nature Rev Genet 6:533–543
6. Brown C (2004) BioBricks to help reverse-engineer life. EE Times, June 11
7. Brown S (2005) Command performances. San Diego Union-Tribune, December 14
8. Brown TA (1990) Gene cloning: an introduction, 2nd edn. Chapman and Hall, London
9. Crick F (1970) Central dogma of molecular biology. Nature 227:561–563
10. Eisenberg A (2000) Unlike viruses, bacteria find a welcome in the world of computing. New York Times, June 1
11. Elowitz M, Leibler S (2000) A synthetic oscillatory network of transcriptional regulators. Nature 403:335–338
12. Ferber D (2004) Synthetic biology: microbes made to order. Science 303(5655):158–161
13. Gardner T, Cantor R, Collins J (2000) Construction of a genetic toggle switch in Escherichia coli. Nature 403:339–342
14. Geyer CR, Battersby TR, Benner SA (2003) Nucleobase pairing in expanded Watson-Crick-like genetic information systems. Structure 11:1485–1498
15. Gibbs WW (2004) Synthetic life. Scientific Am April 26
16. Gravitz L (2004) 10 emerging technologies that will change your world. MIT Technol Rev February
17. Hasty J (2002) Design then mutate. Proc Natl Acad Sci 99(26):16516–16518
18. Hopkin K (2004) Life: the next generation. The Scientist 18(19):56
19. Jackson DA, Symons RH, Berg P (1972) Biochemical method for inserting new genetic information into DNA of simian virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli. Proc Natl Acad Sci 69:2904–2909

20. Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3:318–356
21. Jha A (2005) From the cells up. The Guardian, March 10
22. Kauffman S (1993) Gene regulation networks: a theory for their global structure and behaviors. Current topics in developmental biology 6:145–182
23. Kauffman SA (1993) The origins of order: Self-organization and selection in evolution. Oxford University Press, New York
24. Levskaya A, Chevalier AA, Tabor JJ, Simpson ZB, Lavery LA, Levy M, Davidson EA, Scouras A, Ellington AD, Marcotte EM, Voight CA (2005) Engineering Escherichia coli to see light. Nature 438:441–442
25. Lobban PE, Sutton CA (1973) Enzymatic end-to-end joining of DNA molecules. J Mol Biol 78(3):453–471
26. Marks P (2005) For ultrasharp pictures, use a living camera. New Scientist, November 26, p 28
27. McAdams HH, Shapiro L (1995) Circuit simulation of genetic networks. Science 269(5224):650–656
28. McAdams HH, Arkin A (2000) Genetic regulatory circuits: Advances toward a genetic circuit engineering discipline. Current Biol 10:318–320
29. Monod J (1970) Chance and Necessity. Penguin, London
30. Monod J, Changeux JP, Jacob F (1963) Allosteric proteins and cellular control systems. J Mol Biol 6:306–329
31. Morton O (2005) Life, Reinvented. Wired 13(1), January
32. Registry of Standard Biological Parts. http://parts.mit.edu/
33. Old R, Primrose S (1994) Principles of Gene Manipulation, an Introduction to Genetic Engineering, 5th edn. Blackwell, Boston
34. Ptashne M (2004) A Genetic Switch, 3rd edn. Phage Lambda Revisited. Cold Spring Harbor Laboratory Press, Woodbury
35. Roberts L, Murrell C (eds) (1998) An introduction to genetic engineering. Department of Biological Sciences, University of Warwick
36. Strogatz S (2003) Sync: The Emerging Science of Spontaneous Order. Penguin, London
37. Turing AM (1952) The chemical basis of morphogenesis. Phil Trans Roy Soc B 237:37–72
38. von Neumann J (1941) The general and logical theory of automata. In: Cerebral Mechanisms in Behavior. Wiley, New York
39. Yokobayashi Y, Weiss R, Arnold FH (2002) Directed evolution of a genetic circuit. Proc Natl Acad Sci 99(26):16587–16591
40. You L, Cox III RS, Weiss R, Arnold FH (2004) Programmed population control by cell-cell communication and regulated killing. Nature 428:868–871

Books and Reviews

Alon U (2006) An Introduction to Systems Biology: Design Principles of Biological Circuits. Chapman and Hall/CRC
Amos M (ed) (2004) Cellular Computing. Series in Systems Biology Oxford University Press
Amos M (2006) Genesis Machines: The New Science of Biocomputing. Atlantic Books, London
Benner SA (2003) Synthetic biology: Act natural. Nature 421:118
Endy D (2005) Foundations for engineering biology. Nature 436:449–453
Kobayashi H, Kaern M, Araki M, Chung K, Gardner TS, Cantor CR, Collins JJ (2004) Programmable cells: interfacing natural and engineered gene networks. Proc Natl Acad Sci 101(22):8414–8419
Sayler GS, Simpson ML, Cox CD (2004) Emerging foundations: nano-engineering and bio-microelectronics for environmental biotechnology. Curr Opin Microbiol 7:267–273

Bayesian Games: Games with Incomplete Information

Shmuel Zamir
Center for the Study of Rationality, Hebrew University, Jerusalem, Israel

Article Outline

Glossary
Definition of the Subject
Introduction
Harsanyi's Model: The Notion of *Type*
Aumann's Model
Harsanyi's Model and Hierarchies of Beliefs
The Universal Belief Space
Belief Subspaces
Consistent Beliefs and Common Priors
Bayesian Games and Bayesian Equilibrium
Bayesian Equilibrium and Correlated Equilibrium
Concluding Remarks and Future Directions
Acknowledgments
Bibliography

Glossary

Bayesian game An interactive decision situation involving several decision makers (players) in which each player has beliefs about (i. e. assigns probability distribution to) the payoff relevant parameters and the beliefs of the other players.

State of nature Payoff relevant data of the game such as payoff functions, value of a random variable, etc. It is convenient to think of a state of nature as a full description of a 'game-form' (actions and payoff functions).

Type Also known as *state of mind*, is a full description of player's beliefs (about the state of nature), beliefs about beliefs of the other players, beliefs about the beliefs about his beliefs, etc. ad infinitum.

State of the world A specification of the state of nature (payoff relevant parameters) and the players' types (belief of all levels). That is, a state of the world is a state of nature and a list of the states of mind of all players.

Common prior and consistent beliefs The beliefs of players in a game with incomplete information are said to be consistent if they are derived from the same

probability distribution (the common prior) by conditioning on each player's private information. In other words, if the beliefs are consistent, the only source of differences in beliefs is difference in information.

Bayesian equilibrium A Nash equilibrium of a Bayesian game: A list of behavior and beliefs such that each player is doing his best to maximize his payoff, according to his beliefs about the behavior of the other players.

Correlated equilibrium A Nash equilibrium in an extension of the game in which there is a chance move, and each player has only partial information about its outcome.

Definition of the Subject

Bayesian games (also known as *Games with Incomplete Information*) are models of interactive decision situations in which the decision makers (players) have only partial information about the data of the game and about the other players. Clearly this is typically the situation we are facing and hence the importance of the subject: The basic underlying assumption of classical game theory according to which the data of the game is *common knowledge* (CK) among the players, is too strong and often implausible in real situations. The importance of Bayesian games is in providing the tools and methodology to relax this implausible assumption, to enable modeling of the overwhelming majority of real-life situations in which players have only partial information about the payoff relevant data. As a result of the interactive nature of the situation, this methodology turns out to be rather deep and sophisticated, both conceptually and mathematically: Adopting the classical Bayesian approach of statistics, we encounter the need to deal with an *infinite hierarchy of beliefs*: what does each player believe that the other player believes about what he believes… is the actual payoff associated with a certain outcome? It is not surprising that this methodological difficulty was a major obstacle in the development of the theory, and this article is largely devoted to explaining and resolving this methodological difficulty.

Introduction

A game is a mathematical model for an interactive decision situation involving several decision makers (players) whose decisions affect each other. A basic, often implicit, assumption is that the data of the game, which we call the *state of nature*, are *common knowledge* (CK) among the players. In particular the actions available to the players and the payoff functions are CK. This is a rather strong assumption that says that every player knows all actions and payoff functions of all players, every player knows that all other players know all actions and payoff functions, every player knows that every player knows that every player knows… etc. ad infinitum. *Bayesian games* (also known as *games with incomplete information*), which is the subject of this article, are models of interactive decision situations in which each player has only partial information about the payoff relevant parameters of the given situation.

Adopting the Bayesian approach, we assume that a player who has only partial knowledge about the state of nature has some *beliefs*, namely prior distribution, about the parameters which he does not know or he is uncertain about. However, unlike in a statistical problem which involves a single decision maker, this is not enough in an interactive situation: As the decisions of other players are relevant, so are their beliefs, since they affect their decisions. Thus a player must have beliefs about the beliefs of other players. For the same reason, a player needs beliefs about the beliefs of other players about his beliefs and so on. This interactive reasoning about beliefs leads unavoidably to *infinite hierarchies of beliefs* which looks rather intractable. The natural emergence of hierarchies of beliefs is illustrated in the following example:

Example 1 Two players, P1 and P2, play a 2×2 game whose payoffs depend on an unknown state of nature $s \in \{1, 2\}$. Player P1's actions are $\{T, B\}$, player P2's actions are $\{L, R\}$ and the payoffs are given in the following matrices:

		P2: L	P2: R
P1	T	$0, 1$	$1, 0$
	B	$1, 0$	$0, 1$

a Payoffs when $s = 1$

		P2: L	P2: R
P1	T	$1, 0$	$0, 1$
	B	$0, 1$	$1, 0$

b Payoffs when $s = 2$

Assume that the belief (prior) of P1 about the event $\{s = 1\}$ is p and the belief of P2 about the same event is q. The best action of P1 depends both on his prior and on the action of P2, and similarly for the best action of P2. This is given in the following tables:

	P2's action L	R
$p < 0.5$	T	B
$p > 0.5$	B	T

a Best reply of P1

P1's action		$q < 0.5$	$q > 0.5$
	T	R	L
	B	L	R

b Best reply of P2

Now, since the optimal action of P1 depends not only on his belief p but also on the, unknown to him, action of P2, which depends on his belief q, player P1 must therefore have beliefs about q. These are his *second-level beliefs*, namely beliefs about beliefs. But then, since this is relevant and unknown to P2, he must have beliefs about that which will be *third-level beliefs* of P2, and so on. The whole infinite hierarchies of beliefs of the two players pop out naturally in the analysis of this simple two-person game of incomplete information.

The objective of this article is to model this kind of situation. Most of the effort will be devoted to the modeling of the mutual beliefs structure and only then we add the underlying game which, together with the beliefs structure, defines a *Bayesian game* for which we define the notion of *Bayesian equilibrium*.

Harsanyi's Model: The Notion of *Type*

As suggested by our introductory example, the straightforward way to describe the mutual beliefs structure in a situation of incomplete information is to specify explicitly the whole hierarchies of beliefs of the players, that is, the beliefs of each player about the unknown parameters of the game, each player's beliefs about the other players' beliefs about these parameters, each player's beliefs about the other players' beliefs about his beliefs about the parameters, and so on ad infinitum. This may be called the *explicit* approach and is in fact feasible and was explored and developed at a later stage of the theory (see [18,5,6,7]). We will come back to it when we discuss the universal belief space. However, for obvious reasons, the explicit approach is mathematically rather cumbersome and hardly manageable. Indeed this was a major obstacle to the development of the theory of games with incomplete information at its early stages. The breakthrough was provided by John Harsanyi [11] in a seminal work that earned him the Nobel Prize some thirty years later. While Harsanyi actually formulated the problem verbally, in an explicit way, he suggested a solution that 'avoided' the difficulty of having to deal with infinite hierarchies of beliefs, by providing a much more workable *implicit*, encapsulated model which we present now.

The key notion in Harsanyi's model is that of *type*. Each player can be of several types where a type is to be thought of as a full description of the player's beliefs about the *state of nature* (the data of the game), beliefs about the beliefs of other players about the state of nature and about his own beliefs, etc. One may think of a player's type as his *state of mind*; a specific configuration of his brain that contains an answer to any question regarding beliefs about the state of nature and about the types of the other players. Note that this implies self-reference (of a type to itself through the types of other players) which is unavoidable in an interactive decision situation. A Harsanyi game of incomplete information consists of the following ingredients (to simplify notations, assume all sets to be finite):

- I – Player's set.
- S – The set of states of nature.
- T_i – The type set of player $i \in I$.
  Let $T = \times_{i \in I} T_i$ – denote the *type set*, that is, the set type profiles.
- $Y \subset S \times T$ – a set of *states of the world*.
- $p \in \Delta(Y)$ – probability distribution on Y, called the *common prior*.

(For a set A, we denote the set of probability distributions on A by $\Delta(A)$.)

Remark A state of the world ω thus consists of a state of nature and a list of the types of the players. We denote it as

$$\omega = (s(\omega); t_1(\omega), \ldots, t_n(\omega)) .$$

We think of the state of nature as a full description of the game which we call a *game-form*. So, if it is a game in strategic form, we write the state of nature at state of the world ω as:

$$s(\omega) = (I, (A_i(\omega))_{i \in I}, (u_i(\cdot; \omega))_{i \in I}) .$$

The payoff functions u_i depend only on the state of nature and not on the types. That is, for all $i \in I$:

$$s(\omega) = s(\omega') \Rightarrow u_i(\cdot; \omega) = u_i(\cdot; \omega') .$$

The game with incomplete information is played as follows:

(1) A chance move chooses $\omega = (s(\omega); t_1(\omega), \ldots, t_n(\omega)) \in Y$ using the probability distribution p.
(2) Each player is told his chosen type $t_i(\omega)$ (but not the chosen state of nature $s(\omega)$ and not the other players' types $t_{-i}(\omega) = (t_j(\omega))_{j \neq i}$).
(3) The players choose simultaneously an action: player i chooses $a_i \in A_i(\omega)$ and receives a payoff $u_i(a; \omega)$ where $a = (a_1, \ldots, a_n)$ is the vector of chosen actions and ω is the state of the world chosen by the chance move.

Remark The set $A_i(\omega)$ of actions available to player i in state of the world ω must be known to him. Since his only information is his type $t_i(\omega)$, we must impose that $A_i(\omega)$ is T_i-measurable, i. e.,

$$t_i(\omega) = t_i(\omega') \Rightarrow A_i(\omega) = A_i(\omega') .$$

Note that if $s(\omega)$ was commonly known among the players, it would be a regular game in strategic form. We use the term 'game-form' to indicate that the players have only partial information about $s(\omega)$. The players do not know which $s(\omega)$ is being played. In other words, in the extensive form game of Harsanyi, the game-forms $(s(\omega))_{\omega \in Y}$ are not subgames since they are interconnected by information sets: Player i does not know which $s(\omega)$ is being played since he does not know ω; he knows only his own type $t_i(\omega)$.

An important application of Harsanyi's model is made in *auction theory*, as an auction is a clear situation of incomplete information. For example, in a closed private-value auction of a single indivisible object, the type of a player is his private-value for the object, which is typically known to him and not to other players. We come back to this in the section entitled "Examples of Bayesian Equilibria".

Aumann's Model

A frequently used model of incomplete information was given by Aumann [2].

Definition 2 An Aumann model of incomplete information is $(I, Y, (\pi_i)_{i \in I}, P)$ where:

- I is the players' set.
- Y is a (finite) set whose elements are called *states of the world*.
- For $i \in I$, π_i is a partition of Y.
- P is a probability distribution on Y, also called the *common prior*.

In this model a state of the world $\omega \in Y$ is chosen according to the probability distribution P, and each player i is informed of $\pi_i(\omega)$, the element of his partition that contains the chosen state of the world ω. This is the informational structure which becomes a game with incomplete information if we add a mapping $s\colon Y \to S$. The state of nature $s(\omega)$ is the game-form corresponding to the state of the world ω (with the requirement that the action sets $A_i(\omega)$ are π_i-measurable).

It is readily seen that Aumann's model is a Harsanyi model in which the type set T_i of player i is the set of his partition elements, i. e., $T_i = \{\pi_i(\omega) | \omega \in Y\}$, and the common prior on Y is P. Conversely, any Harsanyi model is an Aumann model in which the partitions are those defined by the types, i. e., $\pi_i(\omega) = \{\omega' \in Y | t_i(\omega') = t_i(\omega)\}$.

Harsanyi's Model and Hierarchies of Beliefs

As our starting point in modeling incomplete information situations was the appearance of hierarchies of beliefs, one may ask how is the Harsanyi (or Aumann) model related to hierarchies of beliefs and how does it capture this unavoidable feature of incomplete information situations? The main observation towards answering this question is the following:

Proposition 3 *Any state of the world in Aumann's model or any type profile $t \in T$ in Harsanyi's model defines (uniquely) a hierarchy of mutual beliefs among the players.*

Let us illustrate the idea of the proof by the following example:

Example Consider a Harsanyi model with two players, I and II, each of which can be of two types: $T_I = \{I_1, I_2\}$, $T_{II} = \{II_1, II_2\}$ and thus: $T = \{(I_1, II_1), (I_1, II_2), (I_2, II_1), (I_2, II_2)\}$. The probability p on types is given by:

	II_1	II_2
I_1	$\frac{1}{4}$	$\frac{1}{4}$
I_2	$\frac{1}{3}$	$\frac{1}{6}$

Denote the corresponding states of nature by $a = s(I_1 II_1)$, $b = s(I_1 II_2)$, $c = s(I_2 II_1)$ and $d = s(I_2 II_2)$. These are the *states of nature* about which there is incomplete information.

The game in extensive form:

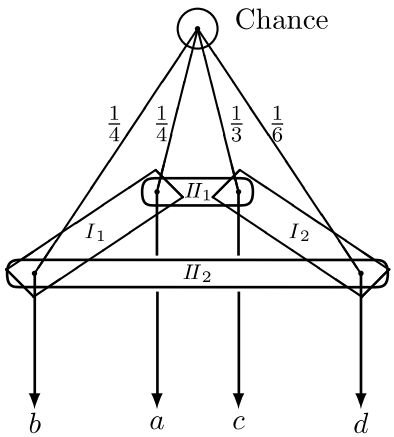

Assume that the state of nature is a. What are the belief hierarchies of the players?

	II_1	II_2
I_1	a: $\frac{1}{4}$	b: $\frac{1}{4}$
I_2	c: $\frac{1}{3}$	d: $\frac{1}{6}$

First-level beliefs are obtained by each player from p, by conditioning on his type:

- I_1: With probability $\frac{1}{2}$ the state is a and with probability $\frac{1}{2}$ the state is b.
- I_2: With probability $\frac{2}{3}$ the state is c and with probability $\frac{1}{3}$ the state is d.
- II_1: With probability $\frac{3}{7}$ the state is a and with probability $\frac{4}{7}$ the state is c.
- II_2: With probability $\frac{3}{5}$ the state is b and with probability $\frac{2}{5}$ the state is d.

Second-level beliefs (using short-hand notation for the above beliefs: $\left(\frac{1}{2}a \overset{*}{+} \frac{1}{2}b\right)$, etc.):

- I_1: With probability $\frac{1}{2}$, player II believes $\left(\frac{3}{7}a \overset{*}{+} \frac{4}{7}c\right)$, and with probability $\frac{1}{2}$, player II believes $\left(\frac{3}{5}b \overset{*}{+} \frac{2}{5}d\right)$.
- I_2: With probability $\frac{2}{3}$, player II believes $\left(\frac{3}{7}a \overset{*}{+} \frac{4}{7}c\right)$, and with probability $\frac{1}{3}$, player II believes $\left(\frac{3}{5}b \overset{*}{+} \frac{2}{5}d\right)$.
- II_1: With probability $\frac{3}{7}$, player I believes $\left(\frac{1}{2}a \overset{*}{+} \frac{1}{2}b\right)$, and with probability $\frac{4}{7}$, player I believes $\left(\frac{2}{3}c \overset{*}{+} \frac{1}{3}d\right)$.
- II_2: With probability $\frac{3}{5}$, player I believes $\left(\frac{1}{2}a \overset{*}{+} \frac{1}{2}b\right)$, and with probability $\frac{2}{5}$, player I believes $\left(\frac{2}{3}c \overset{*}{+} \frac{1}{3}d\right)$.

Third-level beliefs:

- I_1: With probability $\frac{1}{2}$, player II believes that: "With probability $\frac{3}{7}$, player I believes $\left(\frac{1}{2}a \overset{*}{+} \frac{1}{2}b\right)$ and with probability $\frac{4}{7}$, player I believes $\left(\frac{2}{3}c \overset{*}{+} \frac{1}{3}d\right)$".
  And with probability $\frac{1}{2}$, player II believes that: "With probability $\frac{3}{5}$, player I believes $\left(\frac{1}{2}a \overset{*}{+} \frac{1}{2}b\right)$ and with probability $\frac{2}{5}$, player I believes $\left(\frac{2}{3}c \overset{*}{+} \frac{1}{3}d\right)$".

and so on and so on. The idea is very simple and powerful; since each player of a given type has a probability distribution (beliefs) both about the types of the other players and about the set S of states of nature, the hierarchies of beliefs are constructed inductively: If the kth level beliefs (about S) are defined for each type, then the beliefs about types generates the $(k+1)$th level of beliefs.

Thus the compact model of Harsanyi *does* capture the whole hierarchies of beliefs and it is rather tractable. The natural question is whether this model can be used for *all* hierarchies of beliefs. In other words, given any hierarchy of mutual beliefs of a set of players I about a set S of states of nature, can it be represented by a Harsanyi game? This was answered by Mertens and Zamir [18], who constructed the universal belief space; that is, given a set S of states of nature and a finite set I of players, they looked for the space Ω of *all possible* hierarchies of mutual beliefs about S among the players in I. This construction is outlined in the next section.

The Universal Belief Space

Given a finite set of players $I = \{1, \ldots, n\}$ and a set S of states of nature, which are assumed to be compact, we first identify the mathematical spaces in which lie the hierarchies of beliefs. Recall that $\Delta(A)$ denotes the set of probability distributions on A and define inductively the sequence of spaces $(X_k)_{k=1}^{\infty}$ by

$$X_1 = \Delta(S) \tag{1}$$

$$X_{k+1} = X_k \times \Delta(S \times X_k^{n-1}), \quad \text{for } k = 1, 2, \ldots . \tag{2}$$

Any probability distribution on S can be a first-level belief and is thus in X_1. A second-level belief is a joint probability distribution on S and the first-level beliefs of the other $(n-1)$ players. This is an element in $\Delta(S \times X_1^{n-1})$ and therefore a two-level hierarchy is an element of the product space $X_1 \times \Delta(S \times X_1^{n-1})$, and so on for any level.

Note that at each level belief is a joint probability distribution on S and the previous level beliefs, allowing for correlation between the two. In dealing with these probability spaces we need to have some mathematical structure. More specifically, we make use of the *weak** topology:

Definition 4 A sequence $(F_n)_{n=1}^{\infty}$ of probability measures (on Ω) converges in the *weak** topology to the probability F if and only if $\lim_{n\to\infty} \int_\Omega g(\omega)dF_n = \int_\Omega g(\omega)dF$ for all bounded and continuous functions $g\colon \Omega \to \mathbb{R}$.

It follows from the compactness of S that all spaces defined by (1)–(2) are compact in the *weak** topology. However, for $k > 1$, not every element of X_k represents a coherent hierarchy of beliefs of level k. For example, if $(\mu_1, \mu_2) \in X_2$ where $\mu_1 \in \Delta(S) = X_1$ and $\mu_2 \in \Delta(S \times X_1^{n-1})$, then for this to describe meaningful beliefs of a player, the marginal distribution of μ_2 on S must coincide with μ_1. More generally, any event A in the space of k-level beliefs has to have the same (marginal) probability in any higher-level beliefs. Furthermore, not only are each player's beliefs coherent, but he also considers only coherent beliefs of the other players (only those that are in support of his beliefs). Expressing formally this coherency condition yields a selection $T_k \subseteq X_k$ such that $T_1 = X_1 = \Delta(S)$. It is proved that the projection of T_{k+1} on X_k is T_k (that is, any coherent k-level hierarchy can be extended to a coherent $k+1$-level hierarchy) and that all the sets T_k are compact. Therefore, the *projective limit*, $T = \lim_{\infty \leftarrow k} T_k$, is well defined and nonempty.[1]

Definition 5 The *universal type space* T is the projective limit of the spaces $(T_k)_{k=1}^{\infty}$.

That is, T is the set of *all coherent infinite hierarchies of beliefs* regarding S, of a player in I. It does not depend on i since by construction it contains *all possible* hierarchies of beliefs regarding S, and it is therefore the same for all players. It is determined only by S and the number of players n.

Proposition 6 *The universal type space T is compact and satisfies*

$$T \approx \Delta(S \times T^{n-1}) \,. \tag{3}$$

The $\approx$ sign in (3) is to be read as an *isomorphism* and Proposition 6 says that a type of player can be identified with a joint probability distribution on the state of nature and the types of the other players. The implicit equation (3) reflects the self-reference and circularity of the notion of type: The type of a player is his beliefs about the state of nature and about all the beliefs of the other players, in particular, their beliefs about his own beliefs.

Definition 7 The *universal belief space* (UBS) is the space Ω defined by:

$$\Omega = S \times T^n \,. \tag{4}$$

An element of Ω is called a *state of the world*.

Thus a state of the world is $\omega = (s(\omega); t_1(\omega), t_2(\omega), \ldots, t_n(\omega))$ with $s(\omega) \in S$ and $t_i(\omega) \in T$ for all i in I. This is the specification of the states of nature and the types of all players. The universal belief space Ω is what we looked for: the set of *all* incomplete information and mutual belief configurations of a set of n players regarding the state of nature. In particular, as we will see later, all Harsanyi and Aumann models are embedded in Ω, but it includes also belief configurations that cannot be modeled as Harsanyi games. As we noted before, the UBS is determined only by the set of states of nature S and the set of players I, so it should be denoted as $\Omega(S, I)$. For the sake of simplicity we shall omit the arguments and write Ω, unless we wish to emphasize the underlying sets S and I.

The execution of the construction of the UBS according to the outline above involves some non-trivial mathematics, as can be seen in Mertens and Zamir [18]. The reason is that even with a finite number of states of nature, the space of first-level beliefs is a continuum, the second level is the space of probability distributions on a continuum and the third level is the space of probability distributions on the space of probability distributions on a continuum. This requires some structure for these spaces: For a (Borel) measurable event E let $B_i^p(E)$ be the event "player i of type t_i believes that the probability of E is at least p", that is,

$$B_i^p(E) = \{\omega \in \Omega | t_i(E) \geq p\}$$

Since this is the object of beliefs of players other than i (beliefs of $j \neq i$ about the beliefs of i), this set must also be measurable. Mertens and Zamir used the *weak** topology which is the minimal topology with which the event $B_i^p(E)$ is (Borel) measurable for any (Borel) measurable event E. In this topology, if A is a compact set then $\Delta(A)$, the space of all probability distributions on A, is also compact. However, the hierarchic construction can also be made with stronger topologies on $\Delta(A)$ (see [9,12,17]). Heifetz and Samet [14] worked out the construction of the universal belief space without topology, using only a measurable structure (which is implied by the assumption that the beliefs of the players are measurable). All these explicit constructions of the belief space are within what is

[1] The projective limit (also known as the *inverse limit*) of the sequence $(T_k)_{k=1}^{\infty}$ is the space T of all sequences $(\mu_1, \mu_2, \ldots) \in \times_{k=1}^{\infty} T_k$ which satisfy: For any $k \in \mathbb{N}$, there is a probability distribution $\nu_k \in \Delta(S \times T_k^{n-1})$ such that $\mu_{k+1} = (\mu_k, \nu_k)$.

called the *semantic* approach. Aumann [6] provided another construction of a belief system using the *syntactic* approach based on *sentences and logical formulas* specifying explicitly what each player believes about the state of nature, about the beliefs of the other players about the state of nature and so on. For a detailed construction see Aumann [6], Heifetz and Mongin [13], and Meier [16]. For a comparison of the syntactic and semantic approaches see Aumann and Heifetz [7].

Belief Subspaces

In constructing the universal belief space we implicitly assumed that each player knows his own type since we specified only his beliefs about the state of nature and about the beliefs of the *other* players. In view of that, and since by (3) a type of player i is a probability distribution on $S \times T^{I\setminus\{i\}}$, we can view a type t_i also as a probability distribution on $\Omega = S \times T^I$ in which the marginal distribution on T_i is a degenerate delta function at t_i; that is, if $\omega = (s(\omega); t_1(\omega), t_2(\omega), \ldots, t_n(\omega))$, then for all i in I,

$$t_i(\omega) \in \Delta(\Omega) \quad \text{and} \quad t_i(\omega)[t_i = t_i(\omega)] = 1 \,. \tag{5}$$

In particular it follows that if $\mathrm{Supp}(t_i)$ denotes the support of t_i, then

$$\omega' \in \mathrm{Supp}(t_i(\omega)) \Rightarrow t_i(\omega') = t_i(\omega) \,. \tag{6}$$

Let $P_i(\omega) = \mathrm{Supp}(t_i(\omega)) \subseteq \Omega$. This defines a *possibility correspondence*; at state of the world ω, player i does not consider as possible any point not in $P_i(\omega)$. By (6),

$$P_i(\omega) \cap P_i(\omega') \neq \phi \Rightarrow P_i(\omega) = P_i(\omega') \,.$$

However, unlike in Aumann's model, P_i does not define a partition of Ω since it is possible that $\omega \notin P_i(\omega)$, and hence the union $\cup_{\omega\in\Omega} P_i(\omega)$ may be strictly smaller than Ω (see Example 7). If $\omega \in P_i(\omega) \subseteq Y$ holds for all ω in some subspace $Y \subset \Omega$, then $(P_i(\omega))_{\omega\in Y}$ is a partition of Y.

As we said, the universal belief space includes all possible beliefs and mutual belief structures over the state of nature. However, in a specific situation of incomplete information, it may well be that only part of Ω is relevant for describing the situation. If the state of the world is ω then clearly all states of the world in $\cup_{i\in I} P_i(\omega)$ are relevant, but this is not all, because if $\omega' \in P_i(\omega)$ then all states in $P_j(\omega')$, for $j \neq i$, are also relevant in the considerations of player i. This observation motivates the following definition:

Definition 8 A *belief subspace* (*BL*-subspace) is a closed subset Y of Ω which satisfies:

$$P_i(\omega) \subseteq Y \quad \forall i \in I \quad \text{and} \quad \forall \omega \in Y \,. \tag{7}$$

A belief subspace is *minimal* if it has no proper subset which is also a belief subspace. Given $\omega \in \Omega$, the belief subspace at ω, denoted by $Y(\omega)$, is the minimal subspace containing ω.

Since Ω is a *BL*-subspace, $Y(\omega)$ is well defined for all $\omega \in \Omega$. A *BL*-subspace is a closed subset of Ω which is also *closed under beliefs* of the players. In any $\omega \in Y$, it contains all states of the world which are relevant to the situation: If $\omega' \notin Y$, then no player believes that ω' is possible, no player believes that any other player believes that ω' is possible, no player believes that any player believes that any player believes..., etc.

Remark 9 The subspace $Y(\omega)$ is meant to be the minimal subspace which is belief-closed by all players *at the state* ω. Thus, a natural definition would be: $\tilde{Y}(\omega)$ is the minimal *BL*-subspace containing $P_i(\omega)$ for all i in I. However, if for every player the state ω is not in $P_i(\omega)$ then $\omega \notin \tilde{Y}(\omega)$. Yet, even if it is not in the belief closure of the players, the real state ω is still relevant (at least for the analyst) because it determines the true state of nature; that is, it determines the true payoffs of the game. This is the reason for adding the true state of the world ω, even though "it may not be in the mind of the players".

It follows from (5), (6) and (7) that a *BL*-subspace Y has the following structure:

Proposition 10 *A closed subset Y of the universal belief space Ω is a BL-subspace if and only if it satisfies the following conditions:*

1. *For any $\omega = (s(\omega); t_1(\omega), t_2(\omega), \ldots, t_n(\omega)) \in Y$, and for all i, the type $t_i(\omega)$ is a probability distribution on Y.*
2. *For any ω and ω' in Y,*

$$\omega' \in \mathrm{Supp}(t_i(\omega)) \Rightarrow t_i(\omega') = t_i(\omega) \,.$$

In fact condition 1 follows directly from Definition 8 while condition 2 follows from the general property of the UBS expressed in (6).

Given a *BL*-subspace Y in $\Omega(S, I)$ we denote by T_i *the type set of player i*,

$$T_i = \{t_i(\omega) | \omega \in Y\} \,,$$

and note that unlike in the UBS, in a specific model Y, the type sets are typically *not the same* for all i, and the analogue of (4) is

$$Y \subseteq S \times T_1 \times \cdots \times T_n \,.$$

A *BL*-subspace is a model of incomplete information about the state of nature. As we saw in Harsanyi's model,

in any model of incomplete information about a fixed set S of states of nature, involving the same set of players I, a state of the world ω defines (encapsulates) an infinite hierarchy of mutual beliefs of the players I on S. By the universality of the belief space $\Omega(S, I)$, there is $\omega' \in \Omega(S, I)$ with the same hierarchy of beliefs as that of ω. The mapping of each ω to its corresponding ω' in $\Omega(S, I)$ is called a *belief morphism*, as it preserves the belief structure. Mertens and Zamir [18] proved that the space $\Omega(S, I)$ is universal in the sense that *any* model Y of incomplete information of the set of players I about the state of nature $s \in S$ can be embedded in $\Omega(S, I)$ via belief morphism $\varphi\colon Y \to \Omega(S, I)$ so that $\varphi(Y)$ is a belief subspace in $\Omega(S, I)$. In the following examples we give the BL-subspaces representing some known models.

Examples of Belief Subspaces

Example 1 (A game with complete information) If the state of nature is $s_0 \in S$ then in the universal belief space $\Omega(S, I)$, the game is described by a BL-subspace Y consisting of a single state of the world:

$$Y = \{\omega\} \quad \text{where} \quad \omega = (s_0; [1\omega], \ldots, [1\omega]) \,.$$

Here $[1\omega]$ is the only possible probability distribution on Y, namely, the trivial distribution supported by ω. In particular, the state of nature s_0 (i. e., the data of the game) is commonly known.

Example 2 (Commonly known uncertainty about the state of nature) Assume that the players' set is $I = \{1, \ldots, n\}$ and there are k states of nature representing, say, k possible n-dimensional payoff matrices $G_1, \ldots, G_k$. At the beginning of the game, the payoff matrix is chosen by a chance move according to the probability vector $p = (p_1, \ldots, p_k)$ which is commonly known by the players but no player receives any information about the outcome of the chance move. The set of states of nature is $S = \{G_1, \ldots, G_k\}$. The situation described above is embedded in the UBS, $\Omega(S, I)$, as the following BL-subspace Y consisting of k states of the world (denoting $p \in \Delta(Y)$ by $[p_1\omega_1, \ldots, p_k\omega_k]$):

- $Y = \{\omega_1, \ldots, \omega_k\}$
- $\omega_1 = (G_1; [p_1\omega_1, \ldots, p_k\omega_k], \ldots, [p_1\omega_1, \ldots, p_k\omega_k])$
- $\omega_2 = (G_2; [p_1\omega_1, \ldots, p_k\omega_k], \ldots, [p_1\omega_1, \ldots, p_k\omega_k])$
- $\ldots\ldots$
- $\omega_k = (G_k; [p_1\omega_1, \ldots, p_k\omega_k], \ldots, [p_1\omega_1, \ldots, p_k\omega_k])$.

There is a single type, $[p_1\omega_1, \ldots, p_k\omega_k]$, which is the same for all players. It should be emphasized that the type is a distribution on Y (and not just on the states of nature), which implies that the beliefs $[p_1G_1, \ldots, p_kG_k]$ on the state of nature are commonly known by the players.

Example 3 (Two players with incomplete information on one side) There are two players, $I = \{I, II\}$, and two possible payoff matrices, $S = \{G_1, G_2\}$. The payoff matrix is chosen at random with $P(s = G_1) = p$, known to both players. The outcome of this chance move is known only to player I. Aumann and Maschler have studied such situations in which the chosen matrix is played repeatedly and the issue is how the informed player strategically uses his information (see Aumann and Maschler [8] and its references). This situation is presented in the UBS by the following BL-subspace:

- $Y = \{\omega_1, \omega_2\}$
- $\omega_1 = (G_1; [1\omega_1], [p\omega_1, (1-p)\omega_2])$
- $\omega_2 = (G_2; [1\omega_2], [p\omega_1, (1-p)\omega_2])$.

Player I has two possible types: $I_1 = [1\omega_1]$ when he is informed of G_1, and $I_2 = [1\omega_2]$ when he is informed of G_2. Player II has only one type, $II = [p\omega_1, (1-p)\omega_2]$. We describe this situation in the following *extensive form-like* figure in which the oval forms describe the types of the players in the various vertices.

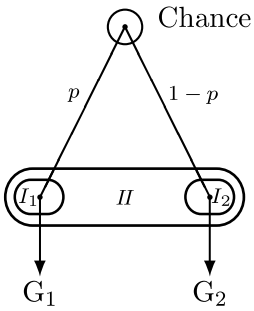

Example 4 (Incomplete information about the other players' information) In the next example, taken from Sorin and Zamir [23], one of two players always knows the state of nature but he may be uncertain whether the other player knows it. There are two players, $I = \{I, II\}$, and two possible payoff matrices, $S = \{G_1, G_2\}$. It is commonly known to both players that the payoff matrix is chosen at random by a toss of a fair coin: $P(s = G_1) = 1/2$. The outcome of this chance move is told to player I. In addition, if (and only if) the matrix G_1 was chosen, another fair coin toss determines whether to inform player II which payoff matrix was chosen. In any case player I is *not told* the result of the second coin toss. This situation is described by the following belief space with three states of the world:

- $Y = \{\omega_1, \omega_2, \omega_3\}$
- $\omega_1 = (G_1; [\frac{1}{2}\omega_1, \frac{1}{2}\omega_2], [1\omega_1])$

- $\omega_2 = (G_1; [\frac{1}{2}\omega_1, \frac{1}{2}\omega_2], [\frac{1}{3}\omega_2, \frac{2}{3}\omega_3])$
- $\omega_3 = (G_2; [1\omega_3], [\frac{1}{2}\omega_2, \frac{1}{2}\omega_3])$

Each player has two types and the type sets are:

$$T_I = \{I_1, I_2\} = \left\{ \left[\frac{1}{2}\omega_1, \frac{1}{2}\omega_2\right], [1\omega_3] \right\}$$
$$T_{II} = \{II_1, II_2\} = \left\{ [1\omega_1], \left[\frac{1}{3}\omega_2, \frac{2}{3}\omega_3\right] \right\} .$$

Note that in all our examples of belief subspaces, condition (6) is satisfied; the support of a player's type contains only states of the world in which he has that type. The game with incomplete information is described in the following figure:

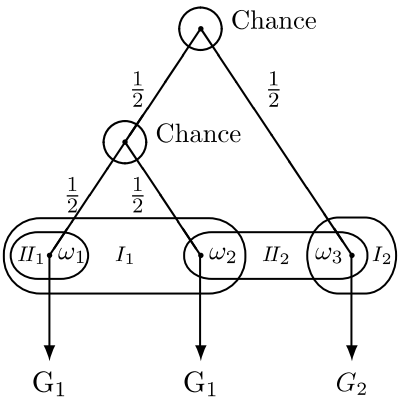

Example 5 (Incomplete information on two sides: A Harsanyi game) In this example, the set of players is again $I = \{I, II\}$ and the set of states of nature is $S = \{s_{11}, s_{12}, s_{21}, s_{22}\}$. In the universal belief space $\Omega(S, I)$ consider the following *BL*-subspace consisting of four states of the world.

- $Y = \{\omega_{11}, \omega_{12}, \omega_{21}, \omega_{22}\}$
- $\omega_{11} = \left(s_{11}; [\frac{3}{7}\omega_{11}, \frac{4}{7}\omega_{12}], [\frac{3}{5}\omega_{11}, \frac{2}{5}\omega_{21}]\right)$
- $\omega_{12} = \left(s_{12}; [\frac{3}{7}\omega_{11}, \frac{4}{7}\omega_{12}], [\frac{4}{5}\omega_{12}, \frac{1}{5}\omega_{22}]\right)$
- $\omega_{21} = \left(s_{21}; [\frac{2}{3}\omega_{21}, \frac{1}{3}\omega_{22}], [\frac{3}{5}\omega_{11}, \frac{2}{5}\omega_{21}]\right)$
- $\omega_{22} = \left(s_{22}; [\frac{2}{3}\omega_{21}, \frac{1}{3}\omega_{22}], [\frac{4}{5}\omega_{12}, \frac{1}{5}\omega_{22}]\right)$

Again, each player has two types and the type sets are:

$$T_I = \{I_1, I_2\} = \left\{ \left[\frac{3}{7}\omega_{11}, \frac{4}{7}\omega_{12}\right], \left[\frac{2}{3}\omega_{21}, \frac{1}{3}\omega_{22}\right] \right\}$$
$$T_{II} = \{II_1, II_2\} = \left\{ \left[\frac{3}{5}\omega_{11}, \frac{2}{5}\omega_{21}\right], \left[\frac{4}{5}\omega_{12}, \frac{1}{5}\omega_{22}\right] \right\} .$$

The type of a player determines his beliefs about the type of the other player. For example, player I of type I_1 assigns probability 3/7 to the state of the world ω_{11} in which player II is of type II_1, and probability 4/7 to the state of the world ω_{12} in which player II is of type II_2. Therefore, the beliefs of type I_1 about the types of player II are $P(II_1) = 3/7$, $P(II_2) = 4/7$. The mutual beliefs about each other's type are given in the following tables:

	II_1	II_2
I_1	3/7	4/7
I_2	2/3	1/3

Beliefs of player I

	II_1	II_2
I_1	3/5	4/5
I_2	2/5	1/5

Beliefs of player II

These are precisely the beliefs of Bayesian players if the pair of types (t_I, t_{II}) in $T = T_I \times T_{II}$ is chosen according to the prior probability distribution p below, and each player is then informed of his own type:

	II_1	II_2
I_1	0.3	0.4
I_2	0.2	0.1

Prior distribution p on T

In other words, this *BL*-subspace is a Harsanyi game with type sets T_I, T_{II} and the prior probability distribution p on the types. Actually, as there is one-to-one mapping between the type set T and the set S of states of nature, the situation is generated by a chance move choosing the state of nature $s_{ij} \in S$ according to the distribution p (that is, $P(s_{ij}) = P(I_i, II_j)$ for i and j in $\{1, 2\}$) and then player I is informed of i and player II is informed of j. As a matter of fact, *all* the *BL*-subspaces in the previous examples can also be written as Harsanyi games, mostly in a trivial way.

Example 6 (Inconsistent beliefs) In the same universal belief space, $\Omega(S, I)$ of the previous example, consider now another *BL*-subspace $\tilde{Y}$ which differs from Y only by changing the type II_1 of player II from $[\frac{3}{5}\omega_{11}, \frac{2}{5}\omega_{21}]$ to $[\frac{1}{2}\omega_{11}, \frac{1}{2}\omega_{21}]$, that is,

- $\tilde{Y} = \{\omega_{11}, \omega_{12}, \omega_{21}, \omega_{22}\}$
- $\omega_{11} = \left(s_{11}; [\frac{3}{7}\omega_{11}, \frac{4}{7}\omega_{12}], [\frac{1}{2}\omega_{11}, \frac{1}{2}\omega_{21}]\right)$
- $\omega_{12} = \left(s_{12}; [\frac{3}{7}\omega_{11}, \frac{4}{7}\omega_{12}], [\frac{4}{5}\omega_{12}, \frac{1}{5}\omega_{22}]\right)$
- $\omega_{21} = \left(s_{21}; [\frac{2}{3}\omega_{21}, \frac{1}{3}\omega_{22}], [\frac{1}{2}\omega_{11}, \frac{1}{2}\omega_{21}]\right)$
- $\omega_{22} = \left(s_{22}; [\frac{2}{3}\omega_{21}, \frac{1}{3}\omega_{22}], [\frac{4}{5}\omega_{12}, \frac{1}{5}\omega_{22}]\right)$

with type sets:

$$T_I = \{I_1, I_2\} = \left\{ \left[\frac{3}{7}\omega_{11}, \frac{4}{7}\omega_{12}\right], \left[\frac{2}{3}\omega_{21}, \frac{1}{3}\omega_{22}\right] \right\}$$
$$T_{II} = \{II_1, II_2\} = \left\{ \left[\frac{1}{2}\omega_{11}, \frac{1}{2}\omega_{21}\right], \left[\frac{4}{5}\omega_{12}, \frac{1}{5}\omega_{22}\right] \right\} .$$

Now, the mutual beliefs about each other's type are:

	II_1	II_2
I_1	3/7	4/7
I_2	2/3	1/3

Beliefs of player I

	II_1	II_2
I_1	1/2	4/5
I_2	1/2	1/5

Beliefs of player II

Unlike in the previous example, these beliefs *cannot be derived from a prior distribution p*. According to Harsanyi, these are *inconsistent beliefs*. A *BL*-subspace with inconsistent beliefs *cannot be described* as a Harsanyi or Aumann model; it cannot be described as a game in extensive form.

Example 7 ("Highly inconsistent" beliefs) In the previous example, even though the beliefs of the players were inconsistent in all states of the world, the true state was considered possible by all players (for example in the state ω_{12} player I assigns to this state probability 4/7 and player II assigns to it probability 4/5). As was emphasized before, the UBS contains *all* belief configurations, including *highly inconsistent* or *wrong* beliefs, as the following example shows. The belief subspace of the two players I and II concerning the state of nature which can be s_1 or s_2 is given by:

- $Y = \{\omega_1, \omega_2\}$
- $\omega_1 = \left(s_1; [\frac{1}{2}\omega_1, \frac{1}{2}\omega_2], [1\omega_2]\right)$
- $\omega_2 = \left(s_2; [\frac{1}{2}\omega_1, \frac{1}{2}\omega_2], [1\omega_2]\right)$.

In the state of the world ω_1, the state of nature is s_1, player I assigns equal probabilities to s_1 and s_2, but player II assigns probability 1 to s_2. In other words, he does not consider as possible the true state of the world (and also the true state of nature): $\omega_1 \notin P_I(\omega_1)$ and consequently $\cup_{\omega \in Y} P_I(\omega) = \{\omega_2\}$ which is strictly smaller than Y. By the definition of belief subspace and condition (6), this also implies that $\cup_{\omega \in \Omega} P_I(\omega)$ is strictly smaller than Ω (as it does not contain ω_1).

Consistent Beliefs and Common Priors

A *BL*-subspace Y is a semantic belief system presenting, via the notion of types, the hierarchies of belief of a set of players having incomplete information about the state of nature. A state of the world captures the situation at what is called the *interim stage*: Each player knows his own type and has beliefs about the state of nature and the types of the other players. The question "what is the *real state of the world* ω?" is not addressed. In a *BL*-subspace, there is no chance move with explicit probability distribution that chooses the state of the world, while such a probability distribution is part of a Harsanyi or an Aumann model. Yet, in the belief space Y of Example 5 in the previous section, such a *prior distribution p emerged endogenously* from the structure of Y. More specifically, if the state $\omega \in Y$ is chosen by a chance move according to the probability distribution p and each player i is told his type $t_i(\omega)$, then his beliefs are *precisely* those described by $t_i(\omega)$. This is a property of the *BL*-subspace that we call *consistency* (which does not hold, for instance, for the *BL*-subspace $\tilde{Y}$ in Example 6) and that we define now: Let $Y \subseteq \Omega$ be a *BL*-subspace.

Definition 11

(i) A probability distribution $p \in \Delta(Y)$ is said to be *consistent* if for any player $i \in I$,

$$p = \int_Y t_i(\omega) \mathrm{d}p \,. \tag{8}$$

(ii) A *BL*-subspace Y is said to be consistent if there is a consistent probability distribution p with $\mathrm{Supp}(p) = Y$. A consistent *BL*-subspace will be called a *C*-subspace. A state of the world $\omega \in \Omega$ is said to be consistent if it is a point in a *C*-subspace.

The interpretation of (8) is that the probability distribution p is "the average" of the types $t_i(\omega)$ of player i (which are also probability distributions on Y), when the average is taken on Y according to p. This definition is not transparent; it is not clear how it captures the consistency property we have just explained, in terms of a chance move choosing $\omega \in Y$ according to p. However, it turns out to be equivalent.

For $\omega \in Y$ denote $\pi_i(\omega) = \{\omega' \in Y | t_i(\omega') = t_i(\omega)\}$; then we have:

Proposition 12 *A probability distribution $p \in \Delta(Y)$ is consistent if and only if*

$$t_i(\omega)(A) = p(A|\pi_i(\omega)) \tag{9}$$

holds for all $i \in I$ and for any measurable set $A \subseteq Y$.

In particular, a Harsanyi or an Aumann model is represented by a consistent *BL*-subspace since, by construction, the beliefs are derived from a common prior distribution which is part of the data of the model. The role of the prior distribution p in these models is actually not that of an additional parameter of the model but rather that of an additional assumption on the belief system, namely, the *consistency assumption*. In fact, if a minimal belief subspace is consistent, then the common prior p is uniquely determined by the beliefs, as we saw in Example 5; *there is no need to specify p as additional data of the system*.

Proposition 13 *If $\omega \in \Omega$ is a consistent state of the world, and if $Y(\omega)$ is the smallest consistent BL-subspace containing ω, then the consistent probability distribution p on $Y(\omega)$ is uniquely determined.*

(The formulation of this proposition requires some technical qualification if $Y(\omega)$ is a continuum.)

The consistency (or the existence of a common prior), is quite a strong assumption. It assumes that differences in beliefs (i. e., in probability assessments) are due *only* to differences in information; players having precisely *the same information* will have precisely *the same beliefs*. It is no surprise that this assumption has strong consequences, the most known of which is due to Aumann [2]: Players with consistent beliefs *cannot agree to disagree*. That is, if at some state of the world it is commonly known that one player assigns probability q_1 to an event E and another player assigns probability q_2 to the same event, then it must be the case that $q_1 = q_2$. Variants of this result appear under the title of *"No trade theorems"* (see, e. g., [19]): Rational players with consistent beliefs cannot believe that they both can gain from a trade or a bet between them.

The plausibility and the justification of the common prior assumption was extensively discussed in the literature (see, e. g., [4,10,11]). It is sometimes referred to in the literature as the *Harsanyi doctrine*. Here we only make the observation that within the set of *BL*-subspaces in Ω, the set of consistent *BL*-subspaces is a set of measure zero. To see the idea of the proof, consider the following example:

Example 8 (Generalization of Examples 5 and 6) Consider a *BL*-subspace as in Examples 5 and 6 but with type sets:

$$\begin{aligned} T_I &= \{I_1, I_2\} \\ &= \{[\alpha_1\omega_{11}, (1-\alpha_1)\omega_{12}], [\alpha_2\omega_{21}, (1-\alpha_2)\omega_{22}]\} \\ T_{II} &= \{II_1, II_2\} \\ &= \{[\beta_1\omega_{11}, (1-\beta_1)\omega_{21}], [\beta_2\omega_{12}, (1-\beta_2)\omega_{22}]\} . \end{aligned}$$

For any $(\alpha_1, \alpha_2, \beta_1, \beta_2) \in [0,1]^4$ this is a *BL*-subspace. The mutual beliefs about each other's type are:

	II_1	II_2
I_1	α_1	$1-\alpha_1$
I_2	α_2	$1-\alpha_2$

Beliefs of player I

	II_1	II_2
I_1	β_1	β_2
I_2	$1-\beta_1$	$1-\beta_2$

Beliefs of player II

If the subspace is consistent, these beliefs are obtained as conditional distributions from some prior probability distribution p on $T = T_I \times T_{II}$, say, by p of the following matrix:

	II_1	II_2
I_1	p_{11}	p_{12}
I_2	p_{21}	p_{22}

Prior distribution p on T

This implies (assuming $p_{ij} \neq 0$ for all i and j),

$$\frac{p_{11}}{p_{12}} = \frac{\alpha_1}{1-\alpha_1}\,; \quad \frac{p_{21}}{p_{22}} = \frac{\alpha_2}{1-\alpha_2}$$
$$\text{and hence} \qquad \frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{\alpha_1}{1-\alpha_1}\frac{1-\alpha_2}{\alpha_2} .$$

Similarly,

$$\frac{p_{11}}{p_{21}} = \frac{\beta_1}{1-\beta_1}\,; \quad \frac{p_{12}}{p_{22}} = \frac{\beta_2}{1-\beta_2}$$
$$\text{and hence} \qquad \frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{\beta_1}{1-\beta_1}\frac{1-\beta_2}{\beta_2} .$$

It follows that the types must satisfy:

$$\frac{\alpha_1}{1-\alpha_1}\frac{1-\alpha_2}{\alpha_2} = \frac{\beta_1}{1-\beta_1}\frac{1-\beta_2}{\beta_2} , \tag{10}$$

which is *generally* not the case. More precisely, the set of $(\alpha_1, \alpha_2, \beta_1, \beta_2) \in [0,1]^4$ satisfying the condition (10) is a set of measure zero; it is a three-dimensional set in the four-dimensional set $[0,1]^4$. Nyarko [21] proved that even the ratio of the dimensions of the set of consistent *BL*-subspaces to the dimension of the set of *BL*-subspaces goes to zero as the latter goes to infinity. Summing up, *most BL*-subspaces are inconsistent and thus do not satisfy the common prior condition.

Bayesian Games and Bayesian Equilibrium

As we said, a game with incomplete information played by Bayesian players, often called a *Bayesian game*, is a game in which the players have incomplete information about the data of the game. Being a Bayesian, each player has beliefs (probability distribution) about any relevant data he does not know, including the beliefs of the other players. So far, we have developed the belief structure of such a situation which is a *BL*-subspace Y in the universal belief space $\Omega(S, I)$. Now we add the action sets and the payoff functions. These are actually part of the description of the state of nature: The mapping $s\colon \Omega \to S$ assigns to each state of the world ω the *game-form* $s(\omega)$ played at this state. To emphasize this interpretation of $s(\omega)$ as a game-form, we denote it also as Γ_ω:

$$\Gamma_\omega = (I, A_i(t_i(\omega))_{i\in I}, (u_i(\omega))_{i\in I}) ,$$

where $A_i(t_i(\omega))$ is the actions set (pure strategies) of player i at ω and $u_i(\omega)\colon A(\omega) \to \mathbb{R}$ is his payoff function and $A(\omega) = \times_{i\in I} A_i(t_i(\omega))$ is the set of action profiles at state ω. Note that while the actions of a player depend only on his type, his payoff depends on the actions and types of all the players. For a vector of actions $a \in A(\omega)$, we write $u_i(\omega; a)$ for $u_i(\omega)(a)$. Given a *BL*-subspace $Y \subseteq \Omega(S, I)$ we define *the Bayesian game on* Y as follows:

Definition 14 The Bayesian game on Y is a vector payoff game in which:

- $I = \{1, \ldots, n\}$ – the players' set.
- Σ_i – the strategy set of player i, is the set of mappings

  $$\sigma_i\colon Y \longrightarrow A_i \quad \text{which are } T_i\text{–measurable}\,.$$

  In particular:

  $$t_i(\omega_1) = t_i(\omega_2) \Longrightarrow \sigma_i(\omega_1) = \sigma_i(\omega_2)\,.$$

  Let $\Sigma = \times_{i\in I}\Sigma_i$.
- The payoff function u_i for player i is a *vector-valued function* $u_i = \left(u_{t_i}\right)_{t_i\in T_i}$, where u_{t_i} (the payoff function of player i of type t_i) is a mapping

  $$u_{t_i}\colon \Sigma \longrightarrow \mathbb{R}$$

  defined by

  $$u_{t_i}(\sigma) = \int_Y u_i(\omega; \sigma(\omega))\mathrm{d}t_i(\omega)\,. \tag{11}$$

Note that u_{t_i} is T_i–measurable, as it should be. When Y is a finite *BL*-subspace, the above-defined Bayesian game is an n-person "game" in which the payoff for player i is a vector with a payoff for each one of his types (therefore, a vector of dimension $|T_i|$). It becomes a *regular game-form* for a given state of the world ω since then the payoff to player i is $u_{t_i(\omega)}$. However, these game-forms are not regular games since they are *interconnected*; the players do not know which of these "games" they are playing (since they do not know the state of the world ω). Thus, just like a Harsanyi game, a Bayesian game on a *BL*-subspace Y consists of a family of *connected* game-forms, one for each $\omega \in Y$. However, unlike a Harsanyi game, a Bayesian game has no chance move that chooses the state of the world (or the vector of types). A way to transform a Bayesian game into a regular game was suggested by R. Selten and was named by Harsanyi as the *Selten game* G^{**} (see p. 496 in [11]). This is a game with $|T_1| \cdot |T_2| \cdot \ldots \cdot |T_n|$ players (one for each type) in which each player $t_i \in T_i$ chooses a strategy and then *selects* his $(n-1)$ partners, one from each T_j; $j \neq i$, according to his beliefs t_i.

Bayesian Equilibrium

Although a Bayesian game is not a regular game, the Nash equilibrium concept based on the notion of *best reply* can be adapted to yield the solution concept of *Bayesian equilibrium* (also called *Nash–Bayes equilibrium*).

Definition 15 A vector of strategies $\sigma = (\sigma_1, \ldots, \sigma_n)$, in a Bayesian game, is called a *Bayesian equilibrium* if for all i in I and for all t_i in T_i,

$$u_{t_i}(\sigma) \geq u_{t_i}(\sigma_{-i}; \tilde{\sigma}_i)\,, \quad \forall \tilde{\sigma}_i \in \Sigma_i\,, \tag{12}$$

where, as usual, $\sigma_{-i} = (\sigma_j)_{j\neq i}$ denotes the vector of strategies of players other than i.

Thus, a Bayesian equilibrium specifies a behavior for each player which is a *best reply* to what he believes is the behavior of the other players, that is, a best reply to the strategies of the other players given his type. In a game with complete information, which corresponds to a *BL*-subspace with one state of the world ($Y = \{\omega\}$), as there is only one type of each player, and the beliefs are all probability one on a singleton, the Bayesian equilibrium is just the well-known Nash equilibrium.

Remark 16 It is readily seen that when Y is finite, any Bayesian equilibrium is a Nash equilibrium of the Selten game G^{**} in which each type is a player who selects the types of his partners according to his beliefs. Similarly, we can transform the Bayesian game into an ordinary game in strategic form by defining the payoff function to player i to be $\tilde{u}_i = \sum_{t_i\in T_i} \gamma_{t_i} u_{t_i}$ where γ_{t_i} are strictly positive. Again, independently of the values of the constants γ_{t_i}, any Bayesian equilibrium is a Nash equilibrium of this game and vice versa. In particular, if we choose the constants so that $\sum_{t_i\in T_i} \gamma_{t_i} = 1$, we obtain the game suggested by Aumann and Maschler in 1967 (see p. 95 in [8]) and again, the set of Nash equilibria of this game is precisely the set of Bayesian equilibria.

The Harsanyi Game Revisited

As we observed in Example 5, the belief structure of a consistent *BL*-subspace is the same as in a Harsanyi game *after the chance move choosing the types*. That is, the embedding of the Harsanyi game as a *BL*-subspace in the universal belief space is only at *the interim stage*, *after* the moment that each player gets to know his type. The Harsanyi game on the other hand is at the *ex ante stage*, *before* a player knows his type. Then, what is the relation between the Nash equilibrium in the Harsanyi game at the ex ante stage and the equilibrium at the interim stage, namely, the Bayesian equilibrium of the corresponding *BL*-subspace?

This is an important question concerning the embedding of the Harsanyi game in the UBS since, as we said before, the chance move choosing the types *does not appear explicitly in the UBS*. The answer to this question was given by Harsanyi (1967–8) (assuming that each type t_i has a positive probability):

Theorem 17 (Harsanyi) *The set of Nash equilibria of a Harsanyi game is identical to the set of Bayesian equilibria of the equivalent BL-subspace in the UBS.*

In other words, this theorem states that any equilibrium in the ex ante stage is also an equilibrium at the interim stage and vice versa.

In modeling situations of incomplete information, the interim stage is the natural one; if a player *knows his beliefs (type)*, then why should he analyze the situation, as Harsanyi suggests, from the ex ante point of view as if his type was not known to him and he could equally well be of another type? Theorem 17 provides a technical answer to this question: The equilibria are the *same* in both games and the equilibrium strategy of the ex ante game specifies for each type precisely his equilibrium strategy at the interim stage. In that respect, for a player who knows his type, the Harsanyi model is just an *auxiliary game* to compute his equilibrium behavior. Of course the deeper answer to the question above comes from the interactive nature of the situation: Even though player i knows he is of type t_i, he knows that his partners do not know that and that they may consider the possibility that he is of type $\tilde{t}_i$, and since this affects their behavior, the behavior of type $\tilde{t}_i$ is also relevant to player i who knows he is of type t_i. Finally, Theorem 17 makes the Bayesian equilibrium the natural extension of the Nash equilibrium concept to games with incomplete information for consistent or inconsistent beliefs, when the Harsanyi ordinary game model is unavailable.

Examples of Bayesian Equilibria In Example 6, there are two players of two types each, and with *inconsistent* mutual beliefs given by

	II_1	II_2
I_1	3/7	4/7
I_2	2/3	1/3

Beliefs of player I

	II_1	II_2
I_1	1/2	4/5
I_2	1/2	1/5

Beliefs of player II

Assume that the payoff matrices for the four type's of profiles are:

		II	
		L	R
I	T	2, 0	0, 1
	B	0, 0	1, 0

G_{11}: Payoffs when $t = (I_1, II_1)$

		II	
		L	R
I	T	0, 0	0, 0
	B	1, 1	1, 0

G_{12}: Payoffs when $t = (I_1, II_2)$

		II	
		L	R
I	T	0, 0	0, 0
	B	1, 1	0, 0

G_{21}: Payoffs when $t = (I_2, II_1)$

		II	
		L	R
I	T	0, 0	2, 1
	B	0, 0	0, 2

G_{22}: Payoffs when $t = (I_2, II_2)$

As the beliefs are inconsistent they cannot be presented by a Harsanyi game. Yet, we can compute the Bayesian equilibrium of this Bayesian game. Let (x, y) be the strategy of player I, which is:

- Play the mixed strategy $[x(T), (1-x)(B)]$ when you are of type I_1.
- Play the mixed strategy $[y(T), (1-y)(B)]$ when you are of type I_2.

and let (z, t) be the strategy of player II, which is:

- Play the mixed strategy $[z(L), (1-z)(R)]$ when you are of type II_1.
- Play the mixed strategy $[t(L), (1-t)(R)]$ when you are of type II_2.

For $0 < x, y, z, t < 1$, each player of each type must be indifferent between his two pure actions; that yields the values in equilibrium:

$$x = \frac{3}{5}, \quad y = \frac{2}{5}, \quad z = \frac{7}{9}, \quad t = \frac{2}{9}.$$

There is no "expected payoff" since this is a Bayesian game and not a game; the expected payoffs depend on the *actual* state of the world, i. e., the actual types of the players and the actual payoff matrix. For example, the state of the world is $\omega_{11} = (G_{11}; I_1, II_1)$; the expected payoffs are:

$$\pi(\omega_{11}) = \left(\frac{3}{5}, \frac{2}{5}\right) G_{11} \begin{pmatrix} 7/9 \\ 2/9 \end{pmatrix} = \left(\frac{46}{45}, \frac{6}{45}\right).$$

Similarly:

$$\pi(\omega_{12}) = \left(\frac{3}{5}, \frac{2}{5}\right) G_{12} \begin{pmatrix} 2/9 \\ 7/9 \end{pmatrix} = \left(\frac{18}{45}, \frac{4}{45}\right)$$

$$\pi(\omega_{21}) = \left(\frac{2}{5}, \frac{3}{5}\right) G_{21} \begin{pmatrix} 7/9 \\ 2/9 \end{pmatrix} = \left(\frac{21}{45}, \frac{21}{45}\right)$$

$$\pi(\omega_{22}) = \left(\frac{2}{5}, \frac{3}{5}\right) G_{22} \begin{pmatrix} 2/9 \\ 7/9 \end{pmatrix} = \left(\frac{28}{45}, \frac{70}{45}\right).$$

However, these are the *objective* payoffs as viewed by the analyst; they are viewed differently by the players. For

player i of type t_i the relevant payoff is his subjective payoff $u_{t_i}(\sigma)$ defined in (11). For example, at state ω_{11} (or ω_{12}) player I believes that with probability 3/7 the state is ω_{11} in which case his payoff is 46/45 and with probability 4/7 the state is ω_{12} in which case his payoff is 18/45. Therefore his *subjective* expected payoff at state ω_{11} is $3/7 \times 46/45 + 4/7 \times 18/45 = 2/3$. Similar computations show that in states ω_{21} or ω_{22} player I "expects" a payoff of 7/15 while player II "expects" 3/10 at states ω_{11} or ω_{21} and 86/225 in states ω_{12} or ω_{22}.

Bayesian equilibrium is widely used in Auction Theory, which constitutes an important and successful application of the theory of games with incomplete information. The simplest example is that of two buyers bidding in a first-price auction for an indivisible object. If each buyer i has a private value v_i for the object (which is independent of the private value v_j of the other buyer), and if he further believes that v_j is random with uniform probability distribution on $[0, 1]$, then this is a Bayesian game in which the type of a player is his private valuation; that is, the type sets are $T_1 = T_2 = [0, 1]$, which is a continuum. This is a consistent Bayesian game (that is, a Harsanyi game) since the beliefs are derived from the uniform probability distribution on $T_1 \times T_2 = [0, 1]^2$. A Bayesian equilibrium of this game is that in which each player bids half of his private value: $b_i(v_i) = v_i/2$ (see, e. g., Chap. III in [25]). Although auction theory was developed far beyond this simple example, almost all the models studied so far are Bayesian games with consistent beliefs, that is, Harsanyi games. The main reason of course is that consistent Bayesian games are more manageable since they can be described in terms of an equivalent ordinary game in strategic form. However, inconsistent beliefs are rather plausible and exist in the market place in general and even more so in auction situations. An example of that is the case of collusion of bidders: When a bidding ring is formed, it may well be the case that some of the bidders outside the ring are unaware of its existence and behave under the belief that all bidders are competitive. The members of the ring may or may not know whether the other bidders know about the ring, or they may be uncertain about it. This rather plausible mutual belief situation is typically inconsistent and has to be treated as an inconsistent Bayesian game for which a Bayesian equilibrium is to be found.

Bayesian Equilibrium and Correlated Equilibrium

Correlated equilibrium was introduced in Aumann (1974) as the Nash equilibrium of a game extended by adding to it random events about which the players have partial information. Basically, starting from an ordinary game, Aumann added a probability space and information structure and obtained a game with incomplete information, the equilibrium of which he called a *correlated equilibrium of the original game*. The fact that the Nash equilibrium of a game with incomplete information is the Bayesian equilibrium suggests that the concept of correlated equilibrium is closely related to that of Bayesian equilibrium. In fact Aumann noticed that and discussed it in a second paper entitled "Correlated equilibrium as an expression of Bayesian rationality" [3]. In this section, we review briefly, by way of an example, the concept of correlated equilibrium, and state formally its relation to the concept of Bayesian equilibrium.

Example 18 Consider a two-person game with actions $\{T, B\}$ for player **1** and $\{L, R\}$ for player **2** with corresponding payoffs given in the following matrix:

		2	
		L	R
1	T	6,6	2,7
	B	7,2	0,0

G: Payoffs of the basic game

This game has three Nash equilibria: (T, R) with payoff $(2, 7)$, (B, L) with payoff $(7, 2)$ and the mixed equilibrium $([\frac{2}{3}(T), \frac{1}{3}(B)], [\frac{2}{3}(L), \frac{1}{3}(R)])$ with payoff $(4\frac{2}{3}, 4\frac{2}{3})$. Suppose that we add to the game a chance move that chooses an element in $\{T, B\} \times \{L, R\}$ according to the following probability distribution μ:

	L	R
T	1/3	1/3
B	1/3	0,0

μ: Probability distribution on $\{T,B\} \times \{L,R\}$

Let us now extend the game G to a game with incomplete information G^* in which a chance move chooses an element in $\{T, B\} \times \{L, R\}$ according to the probability distribution above. Then player **1** is informed of the first (left) component of the chosen element and player **2** is informed of the second (right) component. Then each player chooses an action in G and the payoff is made. If we interpret the partial information as a suggestion of which action to choose, then it is readily verified that following the suggestion is a Nash equilibrium of the extended game yielding a payoff $(5, 5)$. This was called by Aumann a *correlated equilibrium of the original game* G. In our terminology, the extended game G^* is a Bayesian game and its Nash equilibrium is its Bayesian equilibrium. Thus what

we have here is that a correlated equilibrium of a game is just the Bayesian equilibrium of its extension to a game with incomplete information. We now make this a general formal statement. For simplicity, we use the Aumann model of a game with incomplete information.

Let $G = (I, (A_i)_{i\in I}, (u_i)_{i\in I})$ be a game in strategic form where I is the set of players, A_i is the set of actions (pure strategies) of player i and u_i is his payoff function.

Definition 19 Given a game in strategic form G, an incomplete information extension (the I-extension) of the game G is the game G^* given by

$$G^* = (I, (A_i)_{i\in I}, (u_i)_{i\in I}, (Y, p)), (\pi_i)_{i\in I}) ,$$

where (Y, p) is a finite probability space and π_i is a partition of Y (the information partition of player i).

This is an Aumann model of incomplete information and, as we noted before, it is also a Harsanyi type-based model in which the type of player i at state $\omega \in Y$ is $t_i(\omega) = \pi_i(\omega)$, and a strategy of player i is a mapping from his type set to his mixed actions: $\sigma_i : T_i \to \Delta(A_i)$.

We identify a correlated equilibrium in the game G by the probability distribution μ on the vectors of actions $A = A_1\times, \ldots, \times A_n$. Thus $\mu \in \Delta(A)$ is a correlated equilibrium of the game G if when $a \in A$ is chosen according to μ and each player i is suggested to play a_i, his best reply is in fact to play the action a_i.

Given a game with incomplete information G^* as in definition 19, any vector of strategies of the players $\sigma = (\sigma_1, \ldots, \sigma_n)$ induces a probability distribution on the vectors of actions $a \in A$. We denote this as $\mu_\sigma \in \Delta(A)$.

We can now state the relation between correlated and Bayesian equilibria:

Theorem 20 *Let σ be a Bayesian equilibrium in the game of incomplete information $G^* = (I, (A_i)_{i\in I}, (u_i)_{i\in I}, (Y, p)), (\pi_i)_{i\in I})$; then the induced probability distribution μ_σ is a correlated equilibrium of the basic game $G = (I, (A_i)_{i\in I}, (u_i)_{i\in I})$.*

The other direction is:

Theorem 21 *Let μ be a correlated equilibrium of the game $G = (I, (A_i)_{i\in I}, (u_i)_{i\in I})$; then G has an extension to a game with incomplete information $G^* = (I, (A_i)_{i\in I}, (u_i)_{i\in I}, (Y, p)), (\pi_i)_{i\in I})$ with a Bayesian equilibrium σ for which $\mu_\sigma = \mu$.*

Concluding Remarks and Future Directions

The Consistency Assumption

To the heated discussion of the merits and justification of the consistency assumption in economic and game-theoretical models, we would like to add a couple of remarks. In our opinion, the appropriate way of modeling an incomplete information situation is at the *interim stage*, that is, when a player *knows his own beliefs (type)*. The Harsanyi *ex ante* model is just an auxiliary construction for the analysis. Actually this was also the view of Harsanyi, who justified his model by proving that it provides the same equilibria as the interim stage situation it generates (Theorem 17). The *Harsanyi doctrine* says roughly that our models *"should be consistent"* and if we get an inconsistent model it must be the case that it not be a "correct" model of the situation at hand. This becomes less convincing if we agree that the interim stage is what we are interested in: Not only are *most* mutual beliefs inconsistent, as we saw in the section entitled "Consistent Beliefs and Common Priors" above, but it is hard to argue convincingly that the model in Example 5 describes an *adequate* mutual belief situation while the model in Example 6 does not; the only difference between the two is that in one model, a certain type's beliefs are $[\frac{3}{5}\omega_{11}, \frac{2}{5}\omega_{21}]$ while in the other model his beliefs are $[\frac{1}{2}\omega_{11}, \frac{1}{2}\omega_{21}]$.

Another related point is the fact that if players' beliefs are the data of the situation (in the interim stage), then these are typically imprecise and rather hard to measure. Therefore any meaningful result of our analysis should be robust to small changes in the beliefs. This cannot be achieved within the consistent belief systems which are a thin set of measure zero in the universal belief space.

Knowledge and Beliefs

Our interest in this article was mostly in the notion of *beliefs* of players and less in the notion of *knowledge*. These are two related but different notions. Knowledge is defined through a knowledge operator satisfying some axioms. Beliefs are defined by means of probability distributions. Aumann's model, discussed in the section entitled "Aumann's Model" above, has both elements: The knowledge was generated by the partitions of the players while the beliefs were generated by the probability P on the space Y (and the partitions). Being interested in the subjective beliefs of the player we could understand "at state of the world $\omega \in \Omega$ player i knows the event $E \subseteq \Omega$" to mean "at state of the world $\omega \in \Omega$ player i assigns to the event $E \subseteq \Omega$ probability 1". However, in the universal belief space, "belief with probability 1" does not satisfy a central axiom of the knowledge operator. Namely, if at $\omega \in \Omega$ player i knows the event $E \subseteq \Omega$, then $\omega \in E$. That is, if a player *knows* an event, then this event in fact happened. In the universal belief space where all coherent beliefs are possible, in a state $\omega \in \Omega$ a player may assign probabil-

ity 1 to the event $\{\omega'\}$ where $\omega' \neq \omega$. In fact, if in a BL-subspace Y the condition $\omega \in P_i(\omega)$ is satisfied for all i and all $\omega \in Y$, then belief with probability 1 is a knowledge operator on Y. This in fact was the case in Aumann's and in Harsanyi's models where, by construction, the support of the beliefs of a player in the state ω always included ω. For a detailed discussion of the relationship between knowledge and beliefs in the universal belief space see Vassilakis and Zamir [24].

Future Directions

We have not said much about the existence of Bayesian equilibrium, mainly because it has not been studied enough and there are no general results, especially in the non-consistent case. We can readily see that a Bayesian game on a finite BL-subspace in which each state of nature $s(\omega)$ is a finite game-form has a Bayesian equilibrium in mixed strategies. This can be proved, for example, by transforming the Bayesian game into an ordinary finite game (see Remark 16) and applying the Nash theorem for finite games. For games with incomplete information with a continuum of strategies and payoff functions not necessarily continuous, there are no general existence results. Even in consistent auction models, existence was proved for specific models separately (see [20,15,22]). Establishing general existence results for large families of Bayesian games is clearly an important future direction of research. Since, as we argued before, most games are Bayesian games, the existence of a Bayesian equilibrium should, and could, reach at least the level of generality available for the existence of a Nash equilibrium.

Acknowledgments

I am grateful to two anonymous reviewers for their helpful comments.

Bibliography

1. Aumann R (1974) Subjectivity and Correlation in Randomized Strategies. J Math Econ 1:67–96
2. Aumann R (1976) Agreeing to disagree. Ann Stat 4:1236–1239
3. Aumann R (1987) Correlated equilibrium as an expression of Bayesian rationality. Econometrica 55:1–18
4. Aumann R (1998) Common priors: A reply to Gul. Econometrica 66:929–938
5. Aumann R (1999) Interactive epistemology I: Knowledge. Intern J Game Theory 28:263–300
6. Aumann R (1999) Interactive epistemology II: Probability. Intern J Game Theory 28:301–314
7. Aumann R, Heifetz A (2002) Incomplete Information. In: Aumann R, Hart S (eds) Handbook of Game Theory, vol 3. Elsevier, pp 1666–1686
8. Aumann R, Maschler M (1995) Repeated Games with Incomplete Information. MIT Press, Cambridge
9. Brandenburger A, Dekel E (1993) Hierarchies of beliefs and common knowledge. J Econ Theory 59:189–198
10. Gul F (1998) A comment on Aumann's Bayesian view. Econometrica 66:923–927
11. Harsanyi J (1967–8) Games with incomplete information played by 'Bayesian' players, parts I–III. Manag Sci 8:159–182, 320–334, 486–502
12. Heifetz A (1993) The Bayesian formulation of incomplete information, the non-compact case. Intern J Game Theory 21:329–338
13. Heifetz A, Mongin P (2001) Probability logic for type spaces. Games Econ Behav 35:31–53
14. Heifetz A, Samet D (1998) Topology-free topology of beliefs. J Econ Theory 82:324–341
15. Maskin E, Riley J (2000) Asymmetric auctions. Rev Econ Stud 67:413–438
16. Meier M (2001) An infinitary probability logic for type spaces. CORE Discussion Paper 2001/61
17. Mertens J-F, Sorin S, Zamir S (1994) Repeated Games, Part A: Background Material. CORE Discussion Paper No. 9420
18. Mertens J-F, Zamir S (1985) Foundation of Bayesian analysis for games with incomplete information. Intern J Game Theory 14:1–29
19. Milgrom PR, Stokey N (1982) Information, trade and common knowledge. J Eco Theory 26:17–27
20. Milgrom PR, Weber RJ (1982) A Theory of Auctions and Competitive Bidding. Econometrica 50:1089–1122
21. Nyarko Y (1991) Most games violate the Harsanyi doctrine. C.V. Starr working paper #91–39, NYU
22. Reny P, Zamir S (2004) On the existence of pure strategy monotone equilibria in asymmetric first price auctions. Econometrica 72:1105–1125
23. Sorin S, Zamir S (1985) A 2-person game with lack of information on $1\frac{1}{2}$ sides. Math Oper Res 10:17–23
24. Vassilakis S, Zamir S (1993) Common beliefs and common knowledge. J Math Econ 22:495–505
25. Wolfstetter E (1999) Topics in Microeconomics. Cambridge University Press, Cambridge

Bayesian Methods in Non-linear Time Series

Oleg Korenok
Department of Economics, VCU School of Business, Richmond, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Threshold Autoregressive Model
Smooth Transition Autoregressive Model
Markov-Switching Model

Future Directions
Acknowledgments
Bibliography

Glossary

Autoregressive model describes a stochastic process as a weighted average of its previous values and a stochastic error term.

Threshold autoregressive model is an autoregressive model in which parameters change depending on the time index or the previous values of the process.

Markov-switching autoregressive model is an autoregressive model in which parameters change over time depending on an unobserved Markov chain.

Prior distribution summarizes the information about the parameters of interest after observing the data.

Posterior distribution summarizes the information about the parameters of interest after observing the data.

Definition of the Subject

Economic fluctuations display definite nonlinear features. Recessions, wars, financial panics, and varying government policies change the dynamics of almost all macroeconomic and financial time series. In the time series literature, such events are modeled by modifying the standard linear autoregressive (abbreviated, AR) model

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t \,,$$

where y_t is a covariance stationary process, ϵ_t is an independent and identically distributed noise process, $\epsilon_t \sim i.i.d.N(0, \sigma^2)$, and the parameters c, ϕ_i, and σ^2 are fixed over time. In particular, the literature assumes that y_t follows two or more regimes. The three most commonly used nonlinear models differ in their description of the transition between regimes. In the threshold autoregressive (abbreviated, TAR) model, regime changes abruptly; in the smooth threshold autoregressive (abbreviated, STAR) model, regime changes slowly. Nevertheless, in both models the regime change depends on the time index or lagged values of y_t. In the Markov-switching autoregressive (abbreviated, MAR) model, however, the regime change depends on the past values of an unobserved random variable, the state of the Markov chain, and possibly the lagged values of y_t.

Arguably, the best-known example of the nonlinear time series model is the model of cyclical fluctuations of the US economy. It was first introduced and estimated by Hamilton [45] for quarterly US real Gross National Product over the 1952(II)–1984(IV) period. The model has two discrete regimes. The first regime is associated with a positive 1.2% growth rate and the second regime is associated with a negative −0.4% growth rate. Against his original motivation to find decade-long changes in growth rate trends for the US economy, Hamilton finds that negative growth regimes occur at the business cycle frequency. Positive growth regimes last, on average, 10 quarters, and negative growth regimes last, on average, 4 quarters. Moreover, he finds that the estimated regimes coincide closely with the official National Bureau of Economic Research (abbreviated, NBER) recession dates.

Figure 1 illustrates Hamilton's results for the extended 1952(II)–2006(IV) sample. Panel (a) shows the quarterly growth rate of the US real Gross Domestic Product, currently the more common measure of output; panel (b) plots the estimated probability that the US economy is in a negative growth regime. The shaded regions represent recessionary periods as determined informally and with some delay by the NBER: It took nine months for the NBER's Business Cycle Dating Committee to determine the latest peak of the US economy, which occurred in March 2001 but was officially announced in November 2001. Even though the NBER dates were not used in the model, the periods with high probability of a negative growth rate coincide almost perfectly with the NBER dates.

In addition to the formal recession dating methodology, Hamilton [45] presents clear statistical evidence for the proposition that the US business cycle is asymmetric: Behavior of output during normal times, when labor, capital, and technology determine long-run economic growth, is distinct from behavior during recessions, when all these factors are underutilized.

Introduction

Hamilton's paper triggered an explosion of interest in nonlinear time series. The purpose of this paper is to give a survey of the main developments from the Bayesian perspective. The Bayesian framework treats model parameters as random variables and interprets probability as a degree of belief about particular realizations of a random variable conditional on available information. Given the observed sample, the inference updates prior beliefs, formulated before observing the sample, into posterior beliefs using Bayes' theorem

$$p(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{f(y)} \,,$$

where y is the sample observations $y = (y_1, \ldots, y_T)$, θ is the vector of parameters $\theta = (c, \phi_1, \ldots, \phi_p, \sigma^2)$, $\pi(\theta)$ is

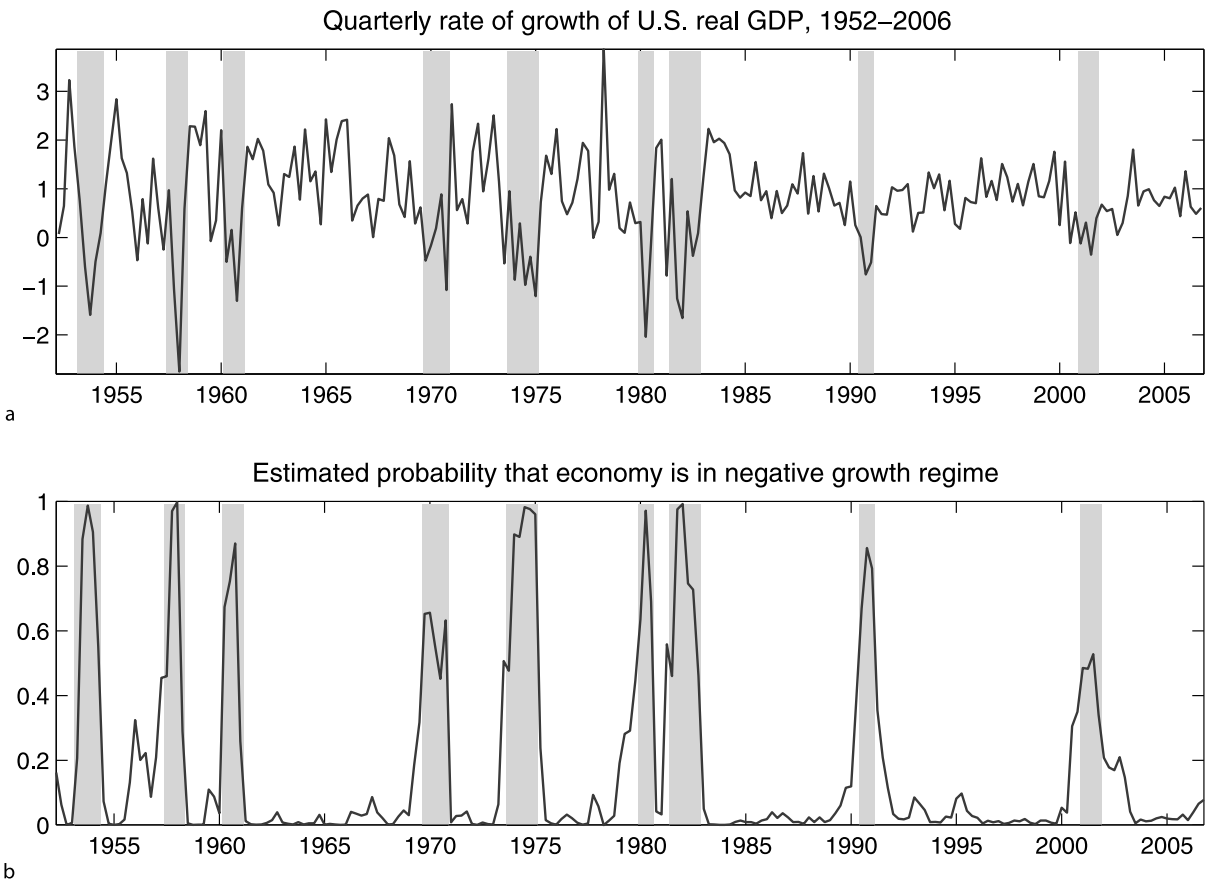

Bayesian Methods in Non-linear Time Series, Figure 1
Output growth and recession probabilities

the prior distribution that describes beliefs prior to observing the data, $f(y|\theta)$ is the distribution of the sample conditional on the parameters, $f(y)$ is the marginal distribution of the sample, and $p(\theta|y)$ is the posterior distribution that describes the beliefs after observing the sample. Zellner [100], Bauwens, Lubrano, and Richard [7], Koop [58], Lancaster [61], and Geweke [39] cover Bayesian econometrics extensively and provide excellent introductions to relevant computational techniques.

We review the three most commonly used nonlinear models in three separate sections. We start each section by describing a baseline model and discussing possible extensions and applications (Matlab implementation of baseline models is available at http://www.people.vcu.edu/~okorenok/share/mlab.zip). Then we review the choice of prior, inference, tests against the linear hypothesis, and conclude with models selection. A short discussion of recent progress in incorporating regime changes into theoretical macroeconomic models concludes our survey.

Our survey builds on reviews of the TAR and STAR models in Tong [95], Granger and Terasvirta [41], Terasvirta [90], Bauwens, Lubrano, and Richard [7], Lubrano [63], Potter [74], Franses and van Dijk [34], van Dijk, Terasvirta, and Franses [98], and on reviews of the MAR models in Hamilton [46], Potter [74], and Kim and Nelson [51].

We limit our survey of nonlinear models only to the TAR, STAR, and MAR models. For a reader interested in a wider range of time series models from a Bayesian prospective, we recommend Steel's [84] survey: He overviews linear, as well as nonlinear, and parametric, as well as nonparametric, models.

Threshold Autoregressive Model

A threshold regression was introduced by Quandt [75] and was extended to the threshold autoregressive model by Tong [92,93] and Tong and Lim [94]. Tong [95] had a great impact on popularizing TAR models.

We limit our baseline model to a single switching variable z_t. The choice of the switching variable depends on the purpose of the investigation. For the analysis of structural breaks at an unknown point in time, Perron and Vogelsang [70], as well as DeJong [24], among many others, use the time index ($z_t = t$). For the purpose of prediction, Geweke and Terui [37], Chen and Lee [15], and others, use a lagged value of the time series ($z_t = y_{t-d}$), the self-exciting threshold autoregressive (abbreviated, SETAR) model.

In our discussion, the number of lags in the model p and a delay d is fixed. We also limit the baseline model to the homoscedastic case so that the variance of ϵ_t is constant in both regimes.

Introducing a more general notation, $x_t' = (1, y_{t-1}, \ldots, y_{t-p})$, $\beta' = (c, \phi_1, \ldots, \phi_p)$, the two-regime TAR model becomes

$$\begin{aligned} y_t &= x_t'\beta_1 + \epsilon_t \quad \text{if} \quad z_t < \tau \quad \text{(first regime)}, \\ y_t &= x_t'\beta_2 + \epsilon_t \quad \text{if} \quad z_t \geq \tau \quad \text{(second regime)}, \end{aligned}$$

or more succinctly

$$y_t = [1 - I_{[\tau,\infty)}(z_t)]x_t'\beta_1 + I_{[\tau,\infty)}(z_t)x_t'\beta_2 + \epsilon_t\,, \quad (1)$$

where $I_A(x)$ is an indicator function that is equal to one if $x \in A$, in particular $I_{[\tau,\infty)}(z_t) = 1$ if $z_t \in [\tau, \infty)$. The indicator function introduces the abrupt transition between regimes. It is convenient to rewrite the model in a more compact form

$$y_t = x_t'(\tau)\beta + \epsilon_t\,, \quad (2)$$

where $x_t'(\tau) = (x_t', I_{[\tau,\infty)}(z_t)x_t')$ and $\beta' = (\beta_1', \delta')$ with $\delta = \beta_2 - \beta_1$.

If the number of observations in regime i is less than or equal to the number of parameters, we cannot estimate parameters, or the model is not identified. In the Bayesian inference, we resolve the identification problem by restricting the region of possible parameter values to the one where the number of observations per regime is greater than the number of regressors.

The baseline model can be extended in several ways. First, we can allow the variance of the error term to differ in each regime. In this case, we rescale the data and introduce an additional parameter $\phi = \sigma_2^2/\sigma_1^2$, as in Lubrano [63]. Second, we can allow the number of lags to differ in each regime. Then p equals to $\max\{p_1, p_2\}$.

A more substantial change is required if we want to increase the number of regimes r. We can either use a single transition variable

$$y_t = x_t\beta_i(t) + \sigma_i(t)\epsilon_t,$$

where $i(t) = 1$ if $z_t < \tau_1$, $i(t) = 2$ if $\tau_1 \leq z_t < \tau_2$, $\ldots$, $i(t) = r$ if $\tau_{r-1} \leq z_t$; or we can use a combination of two (or more) transition variables as in Astatkie, Watts, and Watt [5], where first stage transition is nested in the second stage transition

$$\begin{aligned} y_t = {} & [(1 - I_{[\tau_1,\infty)}(z_{1t}))x_t'\beta_1 + I_{[\tau_1,\infty)}(z_{1t})x_t'\beta_2] \\ & \cdot [1 - I_{[\tau_2,\infty)}(z_{2t})] \\ & + [(1 - I_{[\tau_1,\infty)}(z_{1t}))x_t'\beta_3 + I_{[\tau_1,\infty)}(z_{1t})x_t'\beta_4] \\ & \cdot I_{[\tau_2,\infty)}(z_{2t}) + \epsilon_t\,, \end{aligned}$$

nested TAR model.

Also, we can treat either the choice of number of lags, the delay, or the number of regimes as an inference problem. Then p, d, and r are added to the vector of the model parameters, as in Geweke and Terui [37] and Koop and Potter [57].

Finally, the univariate TAR model can be extended to describe a vector of time series as in Tsay [96]. The n dimensional two-regime TAR model can be specified in a manner similar to Eq. (1) as

$$\begin{aligned} Y_t = {} & [1 - I_{[\tau,\infty)}(z_t)](C_1 + \Phi_{11}Y_{t-1} + \cdots + \Phi_{1p}Y_{t-p}) \\ & + I_{[\tau,\infty)}(z_t)(C_2 + \Phi_{21}Y_{t-1} + \cdots + \Phi_{2p}Y_{t-p}) + \epsilon_t\,, \end{aligned}$$

where $Y_t = (y_{1t}, \ldots, y_{nt})'$ is a $(n \times 1)$ vector, C_1 is a $(n \times 1)$ vector, Φ_{ji}, $j = 1, 2$, $i = 1, \ldots, p$ are $(n \times n)$ matrices, and $\epsilon_t = (\epsilon_{1t}, \ldots, \epsilon_{nt})$ is a vector of error terms with mean zero and positive definite covariance matrix Σ.

The TAR model has a wide range of applications. Tiao and Tsay [91], Potter [73], Pesaran and Potter [71], Rothman [78], and Koop and Potter [54] demonstrate both statistically significant and economically important nonlinearities in the US business cycle. Pfann, Schotman, and Tschernig [72] find strong evidence of high volatility and low volatility regimes in the behavior of US short-term interest rates. Dwyer, Locke, and Yu [26], Martens, Kofman, and Vorst [66], and Forbes, Kalb, and Kofman [33] describe the relationship between spot and futures prices of the S&P 500 index and model financial arbitrage in these markets as a threshold process. Obstfeld and Taylor [68] study the law of one price and purchasing power parity convergences and find strong evidence of two regimes. They demonstrate fast, months rather than years, convergence when price differences are higher than transaction costs, and slow or no convergence otherwise.

To simplify the exposition, our discussion of inference for all models will be conditional on the initial observations in the sample. We assume that $y_{1-p}, \ldots, y_0$ are observable. Two alternative treatments are possible.

One can treat the initial observations as unobserved random variables and include the marginal density of initial observations into the likelihood. Alternatively, in the Bayesian analysis, one can treat the initial observations as any other parameter and augment the parameter space, θ, with $y_{1-p}, \ldots, y_0$.

Prior

The first step in Bayesian inference is to formalize prior beliefs about the model's parameters by choosing functional forms and parameters of prior distributions.

The prior density for τ depends on our choice of z_t. First, we can limit the prior support by the minimum and the maximum of z_t. Second, if $z_t = t$ the threshold is a date, and so the prior density is naturally discrete. If, however, $z_t = y_{t-d}$, the threshold τ is continuous and so is the prior density.

For a model to be identified, we restrict the support of the prior density to the region where the number of observations per regime is greater than the number of regressors. We assign an equal weight to the entire support to get the 'non-informative' prior for τ that is proportional to a constant

$$\pi(\tau) \propto I_{[z_{(k_1)}, z_{(T-k_2)}]}(\tau) \,, \tag{3}$$

where k_1 and k_2 are the number of regressors in the first and second regimes, and the subscript (t) indicates the order in the sample, $z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(T)}$. For example, $z_{(1)} = 1$ and $z_{(T)} = T$ if z_t is a time index since the ordering is natural. For an alternative prior distribution of τ see Ferreira [31].

We assume that the prior density for β and σ^2 is independent of the prior density for τ. Also, because, conditional on τ, the model (2) is linear, we use the natural conjugate prior for β and σ^2

$$\begin{aligned} \pi(\beta|\sigma^2) &= N(\beta|\beta_0, \sigma^2 M_0^{-1}), \\ \pi(\sigma^2) &= IG_2(\sigma^2|\nu_0, s_0), \end{aligned}$$

where $IG_2(.)$ denotes the density of the Inverted Gamma-2 distribution. The functional form of the Inverted Gamma-2 density is given by

$$IG_2(\sigma^2|\nu, s) = \Gamma\left(\frac{\nu}{2}\right)^{-1} \left(\frac{s}{2}\right)^{\frac{\nu}{2}} (\sigma^2)^{-\frac{1}{2}(\nu+2)} \exp\left(-\frac{s}{2\sigma^2}\right).$$

The natural conjugate prior allows us to use analytical integration that considerably simplifies the inference.

Estimation

The next step of the Bayesian analysis is to combine sample information with our prior beliefs to form the posterior beliefs. Given prior distributions, we update prior distributions with the sample likelihood into posterior distributions using Bayes' theorem. The posterior distribution can be further summarized for each parameter with its marginal expectation and variance.

Using the assumption of Normal errors, the likelihood function of the model (2) is

$$f(\beta, \sigma^2, \tau|y) \propto \sigma^{-T} \exp\left\{-\frac{1}{2\sigma^2} \sum (y_t - x_t'(\tau)\beta)^2\right\} . \tag{4}$$

The posterior density is a product of the prior and the likelihood

$$p(\beta, \sigma^2, \tau|y) = \pi(\beta|\sigma^2)\pi(\sigma^2)\pi(\tau) f(\beta, \sigma^2, \tau|y) . \tag{5}$$

Conditional on the threshold parameter, model (2) is linear. Applying the results from the standard natural conjugate analysis in the linear regression model (for details see Zellner [100]), the posteriors density of β, conditional on threshold and the data, can be obtained by integrating the posterior with respect to σ^2

$$\begin{aligned} p(\beta|\tau, y) &= \int p(\beta, \sigma^2|\tau, y)\, \mathrm{d}\sigma^2 \\ &= t(\beta|\beta(\tau), s(\tau), M(\tau), \nu) \,, \end{aligned} \tag{6}$$

where $t(.)$ denotes the density of the multivariate Student t-distribution with

$$\begin{aligned} M(\tau) &= M_0 + \sum x_t(\tau)' x_t(\tau), \\ \beta(\tau) &= M(\tau)^{-1} \left(\sum x_t(\tau) y_t + M_0 \beta_0\right), \\ s(\tau) &= s_0 + \beta_0' M_0 \beta_0 + \sum y_t^2 - \beta'(\tau) M(\tau) \beta(\tau), \\ \nu &= \nu_0 + T \,. \end{aligned}$$

Further, by integrating Eq. (6) with respect to β, we obtain the marginal posterior density for τ, which is proportional to the inverse of the integrating constant of $t(\beta|\beta(\tau), s(\tau), M(\tau), \nu)$ times the threshold prior density

$$p(\tau|y) \propto s(\tau)^{-\nu/2} |M(\tau)|^{-1/2} \pi(\tau) \,. \tag{7}$$

Though analytical integration of this function is not available, the fact that it is a univariate function defined on bounded support greatly simplifies the numerical integration.

By integrating numerically the posterior for β conditional on the threshold and the data, we find marginal posterior density for β

$$p(\beta|y) = \int p(\beta|\tau, y)p(\tau|y)\,\mathrm{d}\tau.$$

Finally, using analytical results for the expectation of the conditional density β, we can find the marginal moments of β by integrating only over τ

$$\begin{aligned} E(\beta|y) &= \int E(\beta|\tau, y)p(\tau|y)\,\mathrm{d}\tau, \\ \mathrm{Var}(\beta|y) &= \int \mathrm{Var}(\beta|\tau, y)p(\tau|y)\,\mathrm{d}\tau \\ &\quad + \int (E(\beta|\tau, y) - E(\beta|y))(E(\beta|\tau, y) \\ &\quad - E(\beta|y))' p(\tau|y)\,\mathrm{d}\tau. \end{aligned}$$

Similarly, applying the results from the standard natural conjugate analysis, we obtain the posterior density of σ^2 conditional on the threshold and the data. Then we integrate out τ numerically to get the marginal posterior density for σ^2

$$p(\sigma^2|y) = \int IG_2(\sigma^2|\nu, s(\tau))p(\tau|y)\,\mathrm{d}\tau\ ,$$

and the marginal moments $E(\sigma^2|y)$ and $\mathrm{Var}(\sigma^2|y)$.

Testing for Linearity and Model Selection

After estimating the TAR model, we might ask whether our data are best characterized by two regimes or a single regime? Model (2) becomes linear when both regimes have identical regression coefficients, so that the difference $\beta_1 - \beta_2 = \delta$ is zero. There are two methods to the null hypothesis test $H_0 : \delta = 0$. The first approach is the Bayesian equivalent of the F-test. Taking into account that β conditional on τ has a Student t-distribution and that the linear transformation of a Student random vector is also a Student, the quadratic transformation of δ

$$\xi(\delta|\tau, y) = (\delta - \delta(\tau))' M_{22.1}(\tau)(\delta - \delta(\tau))\frac{T-k}{k_2 s(\tau)} \qquad (8)$$

has a Fisher distribution, where

$$M_{22.1}(\tau) = M_{22}(\tau) - M_{21}(\tau)M_{11}^{-1}(\tau)M_{12}\ ,$$

and $\delta(\tau)$ is our estimate. $M(\tau)$ is partitioned by dividing β into β_1 and δ. The posterior 'p-value' of the Bayesian F-test gives the unconditional probability that $\xi(\delta|y)$ exceeds $\xi(\delta = 0|y)$. It can be computed numerically as

$$\begin{aligned} &\Pr(\xi(\delta) > \xi(\delta = 0)|y) \\ &\quad = \int F(\xi(\delta = 0|y), k_2, T-k) \cdot p(\tau|y)\,\mathrm{d}\tau\ , \qquad (9) \end{aligned}$$

where $F(\xi(\delta = 0|y), k_2, T-k)$ is the Fisher distribution function with k_2 and $T-k$ degrees of freedom. The null hypothesis is accepted if, for example, $\Pr(\xi(\delta) > \xi(\delta = 0)|y)$ is larger than 5%.

The second approach, the posterior odds, is more general, and can also be used to select the number of lags p, the delay parameter d, or the number of regimes r. Koop and Potter [55,56] advocate and illustrate this approach in the context of the TAR model. To choose between two competing models, m_1 with $\theta_1 = (\beta_1, \delta, \tau, \sigma^2)$ and m_2 with $\theta_2 = (\beta_1, 0, \tau, \sigma^2)$, we calculate the posterior odds ratio

$$po_{12} = \frac{f(y|m_1)\pi(m_1)}{f(y|m_2)\pi(m_2)}\ ,$$

where $\pi(m_i)$ is the prior probability for the model i, and $f(y|m_i)$ is the marginal likelihood or marginal density of the sample. Since $f(y|m_i)$ is a normalizing constant of the posterior density, it can be calculated as

$$f(y|m_i) = \int f(y|\theta_i, m_i)\pi(\theta_i|m_i)\,\mathrm{d}\theta_i\ .$$

With a 'non-informative' prior that assigns equal weight to each model, the posterior odds reduces to the ratio of marginal likelihoods, or the Bayes factor. Again, applying the standard natural conjugate analysis of the linear regression model, the marginal likelihood for model i is

$$\begin{aligned} f(y|m_i) = \int &\frac{\Gamma\left(\frac{\nu(\tau_i|m_i)}{2}\right) s_0^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)\pi^{\frac{T}{2}}} s(\tau_i|m_i)^{-\frac{\nu(\tau_i|m_i)}{2}} \\ &\cdot \left(\frac{|M_0|}{|M(\tau_i|m_i)|}\right)^{\frac{1}{2}} \pi(\tau_i|m_i)\,\mathrm{d}\tau\ , \qquad (10) \end{aligned}$$

which can be calculated numerically. The model with the highest marginal likelihood is preferred.

Smooth Transition Autoregressive Model

In some applications, imposing an abrupt transition between regimes might be undesirable. For example, if the initial estimate of output is slightly below the threshold, even a small upward revision will result in a substantial change of the forecast in the TAR model. Bacon and

Watts [6], in a regression model context, and Chan and Tong [14], in the TAR model context, propose to make the transition between regimes smooth. Terasvirta [89] develops a modeling cycle for the STAR model that includes specification, estimation, and evaluation stages as in the Box and Jenkins [9] modeling cycle for the linear time series model.

In the STAR model, a smooth transition is imposed by replacing the indicator function in Eq. (1) by the cumulative distribution function

$$y_t = [1-F(\gamma(z_t-\tau))]x_t'\beta_1+F(\gamma(z_t-\tau))x_t'\beta_2+\epsilon_t\,. \quad (1a)$$

Terasvirta [89] uses the logistic function

$$F(\gamma(z_t-\tau)) = \frac{1}{1+\exp(-\gamma(z_t-\tau))}\,,$$

where $\gamma \in [0,\infty)$ determines the degree of smoothness. As γ increases, smoothness decreases. In the limit, as γ approaches infinity, $F(.)$ becomes an indicator function, with $F(\gamma(z_t-\tau)) \sim 1$ when $z_t \geq \tau$. We can rewrite Eq. (1a) as

$$y_t = x_t'(\gamma,\tau)\beta + \epsilon_t\,, \quad (2a)$$

where $x_t'(\gamma,\tau) = (x_t', F(\gamma(z_t-\tau))x_t')$.

Note that the identification problem discussed for the TAR model does not occur in the STAR model. We cannot have fewer observations than regressors because we no longer classify observations into regimes. The new parameter γ, however, introduces a new identification problem. If $\gamma = 0$, the logistic function equals $\frac{1}{2}$ for any value of τ, so τ is not identified. Also $x_t'(\gamma,\tau)$ is perfectly collinear unless the two regimes have no common regressors. Perfect collinearity implies that δ is also not identified. As in the TAR model, we choose such prior densities that resolve the identification problem.

The baseline model can be extended in several directions. Generally, the transition function $F(.)$ is not limited to the logistic function. Any continuous, monotonically increasing function $F(.)$ with $F(-\infty) = 0$ and $F(\infty) = 1$ can be used. For example, the popular alternative to the logistic function is the exponential function

$$F(\gamma(z_t-\tau)) = 1-\exp(-\gamma(z_t-\tau)^2)\,.$$

In the regression model context, Bacon and Watts [6] show that results are not sensitive to the choice of $F(.)$. As in the TAR model, we can increase the number of regimes either with a single transition variable

$$\begin{aligned} y_t = {} & x_t'\beta_1 + F(\gamma_1(z_t-\tau_1))x_t'(\beta_2-\beta_1)+\ldots \\ & + F(\gamma_r(z_t-\tau_r))x_t'(\beta_r-\beta_{r-1})+\epsilon_t\,, \end{aligned}$$

or with a combination of transition variables

$$\begin{aligned} y_t = {} & [(1-F(\gamma_1(z_{1t}-\tau_1)))x_t'\beta_1 + F(\gamma_1(z_{1t}-\tau_1))x_t'\beta_2] \\ & \cdot [(1-F(\gamma_2(z_{2t}-\tau_2)))] \\ & + [(1-F(\gamma_1(z_{1t}-\tau_1)))x_t'\beta_3 \\ & + F(\gamma_1(z_{1t}-\tau_1))x_t'\beta_4]\cdot[F(\gamma_2(z_{2t}-\tau_2))] + \epsilon_t\,. \end{aligned}$$

See van Dijk and Franses [97] for a discussion of the multiple regime STAR model.

Also, we can treat the choice of number of lags p, delay d, or number of regimes r as an inference problem, adding p, d, and r to the vector of parameters in the model. In addition, we can allow the variance of the error term to change between regimes, or more generally, use an autoregressive conditional heteroscedasticity form as in Lundbergh and Terasvirta [64], or a stochastic volatility form as in Korenok and Radchenko [59].

Finally, similar to the TAR model, the univariate STAR model can be extended to model a vector of time series as in Granger and Swanson [42]. The n dimensional two-regime STAR model can be specified as

$$\begin{aligned} Y_t = {} & [1-F(\gamma(z_t-\tau))](C_1+\Phi_{11}Y_{t-1}+\cdots+\Phi_{1p}Y_{t-p}) \\ & + F(\gamma(z_t-\tau))(C_2+\Phi_{21}Y_{t-1}+\cdots+\Phi_{2p}Y_{t-p}) \\ & + \epsilon_t\,, \end{aligned}$$

where we use the same notation as in the multivariate TAR model.

Applications of the STAR model include models of the business cycles, real exchange rates, stock and futures prices, interest rates, and monetary policy. Terasvirta and Anderson [88] and van Dijk and Franses [97] demonstrate nonlinearities in the US business cycles. Skalin and Terasvirta [82] find similar nonlinearities in Swedish business cycles. Michael, Nobay, and Peel [67], Sarantis [80], and Taylor, Peel, and Sarno [87] show that the real exchange rate nonlinearly depends on the size of the deviation from purchasing power parity; Lundbergh and Terasvirta [65] and Korenok and Radchenko [59] use the STAR model to fit the behavior of exchange rates inside a target zone. Taylor, van Dijk, Franses, and Lucas [86] describe the nonlinear relationship between spot and futures prices of the FTSE100 index. Anderson [1] uses the STAR model to study yield movements in the US Treasury Bill Market. Finally, Rothman, van Dijk, and Franses [79] find evidence of a nonlinear relationship between money and output; Weise [99] demonstrates that monetary policy has a stronger effect on output during recessions.

Prior

As in the TAR model, the natural conjugate priors for β and σ^2 facilitate analytical integration. Bauwens, Lubrano, and Richard [7] impose the identification at $\gamma = 0$ by modifying the prior density of β

$$\pi(\beta|\sigma^2, \gamma) = N(\beta|0, \sigma^2 M_0^{-1}(\gamma)) ,$$

where, assuming prior independence between β_1 and δ, M_0 is defined as

$$M_0(\gamma) = \begin{pmatrix} M_{0,11} & 0 \\ 0 & M_{0,22}/\exp(\gamma) \end{pmatrix} .$$

As γ gets closer to zero, the prior variance falls, increasing precision around $\delta = 0$. The choice of $\delta = 0$ is consistent with the linear hypothesis, which can be formulated as either $\delta = 0$ or $\gamma = 0$. When γ is positive, prior precision about $\delta = 0$ decreases as variance rises, so more weight is given to the information in the sample. We keep the natural conjugate prior of σ^2 without modifications.

We do not modify the prior for the threshold parameter τ. When γ is large, the smooth transition function is close to the step transition function. Thus, we prefer to limit the prior to the region where the number of observations per regime is greater than the number of regressors to avoid the TAR identification problem.

The prior for the smoothness parameter, γ, cannot be 'non-informative' or flat. As $\gamma \to \infty$ the smooth transition function becomes a step transition with a strictly positive likelihood. This means that the marginal likelihood function of γ is not integrable. To avoid the integration problem, Bauwens, Lubrano, and Richard [7] use the truncated Cauchy density

$$\pi(\gamma) \propto (1 + \gamma^2)^{-1} I_{[0,\infty)}(\gamma) .$$

Estimation

Inference in the STAR model follows the TAR methodology, taking into account the additional parameter γ, and the new definitions of $M_0(\gamma)$ and $x_t(\tau, \gamma)$.

In particular, the likelihood function of model (2a) is

$$f(\beta, \sigma^2, \tau, \gamma|y) \propto \sigma^{-T} \exp\left\{-\frac{1}{2\sigma^2}\sum(y_t - x_t'(\tau,\gamma)\beta)^2\right\} , \quad (4a)$$

the posterior density is

$$p(\beta, \sigma^2, \tau, \gamma|y) = \pi(\beta|\sigma^2)\pi(\sigma^2)\pi(\tau)\pi(\gamma) f(\beta, \sigma^2, \tau, \gamma|y) , \quad (5a)$$

and the joint posterior density of τ and γ is proportional to the inverse of the integrating constant of the Student t-density $t(\beta|\beta(\tau,\gamma), s(\tau,\gamma), M(\tau,\gamma), \nu)$ times the prior densities for c and γ

$$p(\tau, \gamma|y) \propto |s(\tau,\gamma)|^{-(T-k)/2}|M(\tau,\gamma)|^{-1/2} \pi(\tau)\pi(\gamma) , \quad (7a)$$

where

$$\begin{aligned} M(\tau,\gamma) &= M_0(\gamma) + \sum x_t(\tau,\gamma)' x_t(\tau,\gamma) , \\ \beta(\tau,\gamma) &= M(\tau,\gamma)^{-1}\left(\sum x_t(\tau,\gamma) y_t + M_0(\gamma)\beta_0\right) , \\ s(\tau,\gamma) &= s_0 + \beta_0' M_0(\gamma)\beta_0 \\ &\quad + \sum y_t^2 - \beta'(\tau,\gamma)M(\tau,\gamma)\beta(\tau,\gamma) , \\ \nu &= \nu_0 + T . \end{aligned}$$

This function is bivariate and can be integrated numerically with respect to τ and γ. Then, as in the TAR model, we use numerical integration to obtain marginal densities and moments for β and σ^2.

Compared to the TAR model, β_1 and β_2 cannot be interpreted as regression coefficients in regime 1 and regime 2. Smooth transition implies that the effect of change in x_t on y_t is a weighted average of two regimes with weights changing from one observation to the other.

Testing for Linearity and Model Selection

The STAR model becomes linear when either $\delta = 0$ or $\gamma = 0$. The test for $H_0 : \delta = 0$ is equivalent to the test in the TAR model. The quadratic transformation of δ

$$\xi(\delta|\tau,\gamma,y) = (\delta - \delta(\tau,\gamma))' M_{22.1}(\tau,\gamma)(\delta - \delta(\tau,\gamma)) \frac{T-k}{k_2 s(\tau,\gamma)} , \quad (8a)$$

where

$$M_{22.1}(\tau,\gamma) = M_{22}(\tau,\gamma) - M_{21}(\tau,\gamma)M_{11}^{-1}(\tau,\gamma)M_{12}(\tau,\gamma),$$

has a Fisher distribution. We can find the posterior 'p-value' of the Bayesian F-test numerically as

$$\Pr(\xi(\delta) > \xi(\delta = 0)|y) = \iint F(\xi(\delta = 0|y), k_2, T-k) p(\tau,\gamma|y)\, \mathrm{d}\tau\, \mathrm{d}\gamma . \quad (9a)$$

The null hypothesis is accepted, for example, if $\Pr(\xi(\delta) > \xi(\delta = 0)|y)$ is larger than 5%.

The test for $H_0 : \gamma = 0$ can be conducted using the 95% highest posterior density interval (abbreviated, HPDI), defined as the smallest interval with 95% probability of γ to be in the interval

$$\max_h \mathrm{PDI}(h) = \left\{ \gamma | \int p(\tau, \gamma)\pi(\tau)\, \mathrm{d}\tau \geq h \right\}, \quad \text{s.t. } \Pr(\mathrm{PDI}(h)) \geq 0.95.$$

The null hypothesis is accepted, for example, if $\gamma = 0$ is inside the 95% HPDI.

As in the TAR model, linearity tests and model selection can be conducted using posterior odds. In the STAR model, the marginal likelihood for model i is given by

$$f(y|m_i) = \iint \frac{\Gamma\left(\frac{\nu(\tau_i,\gamma_i|m_i)}{2}\right) s_0^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)\pi^{\frac{T}{2}}} s(\tau_i, \gamma_i|m_i)^{-\frac{\nu(\tau_i,\gamma_i|m_i)}{2}} \cdot \left(\frac{|M_0|}{|M(\tau_i, \gamma_i|m_i)|}\right)^{\frac{1}{2}} \pi(\tau_i|m_i)\pi(\gamma_i|m_i)\, \mathrm{d}\tau_i\, \mathrm{d}\gamma_i, \tag{10a}$$

which can be calculated numerically. The model with the highest marginal likelihood is preferred.

Markov-Switching Model

Unlike the threshold models, where the regime transition depends on a time index or on lagged values of y_t, the Markov-switching autoregressive model relies on a random variable, s_t. A Markov-switching regression was introduced in econometrics by Goldfeld and Quandt [40] and was extended to the Markov-switching autoregressive model by Hamilton [45].

As in the threshold models, we limit our baseline MAR model to two regimes that differ only in mean. The variance of the error term is constant. The number of lags p is determined by the model choice. The two-regime MAR model becomes

$$\begin{aligned} (y_t - \mu_{s_t}) &= \sum_{i=1}^{p} \phi_i (y_{t-i} - \mu_{s_{t-i}}) + \epsilon_t, \\ \mu_{s_t} &= \mu_0 \quad \text{if} \quad s_t = 0 \quad \text{(first regime)}, \\ \mu_{s_t} &= \mu_0 + \mu_1 \quad \text{if} \quad s_t = 1 \quad \text{(second regime)}, \end{aligned} \tag{11}$$

where $\mu_{s_t} = \mu_0 + s_t\mu_1$. An unobserved discreet random variable s_t takes only integer values of 0 or 1. The transition probability $\Pr(s_t = j|s_{t-1} = i) = p_{ij}$ that state i will be followed by state j depends only on s_{t-1}, the first order Markov-switching process, with transition probability matrix

$$P = \begin{pmatrix} p_{11} & p_{21} \\ p_{12} & p_{22} \end{pmatrix}.$$

Since we have only two possible regimes and $p_{i1} + p_{i2} = 1$, we estimate only two free parameters, the probabilities of remaining in the same regime p_{11} and p_{22}. We also assume that, conditional on previous history of states $s = (s_1, \ldots, s_T)'$, the transition probabilities are independent of other parameters and the data.

In general, we do not have a clear association between regimes and the state indicator. This introduces an identification problem when we change regime identifiers, 0 and 1, and accordingly change $\mu_0^* = \mu_0 + \mu_1$ and $\mu_1^* = -\mu_1$. For example, if $s_t = 0$ during recessions, then the long run average during recessions is μ_0 and the long-run average during expansions is $\mu_0 + \mu_1$. On the other hand, if $s_t = 0$ during expansions, then the long-run average during expansions is $\mu_0^* = \mu_0 + \mu_1$ and the long-run average during recessions is $\mu_0^* - \mu_1$ or $\mu_1^* = -\mu_1$.

The second identification problem occurs in the MAR model when $\mu_1 = 0$; the model becomes linear. In this case, the conditional mean $E(y_t|s_t = 0) = E(y_t|s_t = 1) = \mu_0$ is independent of the state realizations, s, and transition probability matrix, P. Neither s nor P are identified.

The baseline model can be extended in several directions. The Markov-switching component can be modified by increasing the number of regimes as in Calvet and Fisher [11] and Sims and Zha [81] or by increasing the order of the Markov-switching process so that s_t depends on $s_{t-1}, \ldots, s_{t-r}$. Both changes can be incorporated by increasing the number of states in the baseline model, as in Hamilton [46].

Diebold, Lee, and Weinbach [22], Filardo [32], and Peria [69] relax the assumption of time invariant Markov-switching by making the transition probabilities depend on lagged values of y_t. In most applications, however, relatively few transitions between regimes makes it difficult to estimate the transition probabilities and restricts model choice to two or three regimes with time-invariant probabilities.

The error term can be modified by introducing regime-switching for the variance of the error term as in Hamilton and Susmel [47], and Cai [10]; by relaxing the assumption of Gaussian density for the error term as in Dueker [25]; or by specifying a general Markov-switching moving average structure for the error term as in Billio, Monfort, and Robert [8].

Finally, the univariate Markov-switching model can be extended to a multivariate model. Diebold and Rude-

busch [23] propose a model where a number of time series are driven by a common unobserved Markov-switching variable, the dynamic factor model. The dynamic factor model captures the fact that many economic series show similar changes in dynamic behavior during recessions. Krolzig [60] provides a detailed exposition of how the baseline model can be extended to the Markov-switching vector autoregressive model.

The applications of the MAR model include models of business cycles, interest rates, financial crises, portfolio diversification, options pricing, and changes in government policy. Hamilton [45], Filardo [32], Diebold and Rudebusch [23], Kim and Nelson [51], Kim and Piger [53], and Hamilton [48] find statistically significant evidence that expansionary and contractionary phases of the US business cycle are distinct. Hamilton [44], Cai [10], Garcia and Perron [35], Gray [43], Dueker [25], Smith [83], Hamilton [48], and Dai, Singleton, and Yang [18] describe dramatic changes in interest rate volatility associated with the OPEC oil shocks, the changes in the Federal Reserve operating procedures in 1979–1982, and the stock market crash of October 1987. Ang and Bekaert [3] show a similar increase in volatility in Germany during the reunification period. Jeanne and Masson [49] use the MAR model to describe the crisis of the European Monetary System in 1992–1993; Cerra and Saxena [13] find permanent losses in output after the Asian crisis. Ang and Bekaert [2] report that the correlation between international equity returns is higher during bear markets relative to bull markets. Radchenko [76] shows that gasoline prices respond faster to a permanent oil price change compared to a transitory change. Finally, Sims and Zha [81] document abrupt changes of shocks to US monetary policy, and Davig and Leeper [20] document the regime changes in fiscal policy.

Prior

As in the threshold models, the natural conjugate priors facilitate considerably the integration of the posterior density. Conditional on s_t, μ_0, and μ_1, the MAR model is linear

$$y_t(s_t) = x_t'(s_t)\tilde{\phi} + \epsilon_t \,, \tag{12}$$

where

$$y_t(s_t) = y_t - \mu_{s_t} \,,$$
$$x_t'(s_t) = (y_{t-1} - \mu_{s_{t-1}}, \ldots, y_{t-p} - \mu_{s_{t-p}}) \,,$$

and $\tilde{\phi} = (\phi_1, \ldots, \phi_p)'$. For the regression coefficient $\tilde{\phi}$ and the variance of the error term σ^2, the natural conjugate prior is given by

$$\pi(\tilde{\phi}|\sigma^2) = N(\tilde{\phi}|\tilde{\phi}_0, \sigma^2 M_{0,\phi}^{-1}) I_A(\tilde{\phi}),$$
$$\pi(\sigma^2) = IG_2(\sigma^2|\nu_0, s_0) \,,$$

where A is a region where the roots of polynomial $1 - \phi_1 L - \cdots - \phi_p L^p = 0$ lie outside the complex unit circle. This restriction imposes stationarity on $y_t(s_t)$.

Conditional on s_t and $\tilde{\phi}$, the MAR model is also linear

$$y_t(\tilde{\phi}) = x_t'(\tilde{\phi})\tilde{\mu} + \epsilon_t \,, \tag{13}$$

where

$$y_t(\tilde{\phi}) = y_t - \sum_{i=1}^{p} \phi_i y_{t-p} \,,$$
$$x_t'(\tilde{\phi}) = \left(1, s_t - \sum_{i=1}^{p} \phi_i s_{t-p}\right) \,,$$

and $\tilde{\mu} = (\mu_0, \mu_1)'$. The natural conjugate prior for $\tilde{\mu}$ is

$$\pi(\tilde{\mu}) = N(\tilde{\mu}|\tilde{\mu}_0, M_{0,\mu}^{-1}) I_{(0,\infty)}(\mu_1) \,,$$

where the indicator function imposes an identification constraint. In particular, we constrain the mean of the second regime to be greater than the mean of the first regime and in this way fix the order of regimes. We also impose $\mu_1 \neq 0$.

Kim and Nelson [51] show that the natural conjugate prior for the vector of transition probabilities $\tilde{p} = (p_{11}, p_{22})'$ is

$$\pi(\tilde{p}) = B(p_{11}|\alpha_1, \beta_1) B(p_{22}|\alpha_2, \beta_2) \,,$$

where $B(.)$ denotes the density of Beta distribution defined on the interval $[0, 1]$.

Estimation

In the Bayesian approach, we add realizations of the vector of states to the model parameters: $\theta = (\mu_0, \mu_1, \phi_1, \ldots, \phi_p, \sigma, p_{11}, p_{22}, s_1, \ldots, s_T)'$. Analytical or numerical integration of the posterior density $p(\theta|y)$, where θ is $p + 5 + T \times 1$, may be difficult.

Albert and Chib [4] developed inference methodology that overcomes the curse of dimensionality using Gibbs-sampling, a Markov chain Monte Carlo simulation method of integration. The technique was further refined by Kim and Nelson [50]. Monte Carlo integration takes random draws from the posterior density and, by averaging them, produces estimates of moments. In particular, Gibbs-sampling allows us to generate many draws

$\theta^{(g)}, g = 1, \ldots, G$, from joint density of $p(\theta|y)$ using only conditional densities $p(\theta_i|\theta_{i\neq j}, y)$ either for all i or for blocks of parameters. The joint and marginal distribution of $\theta^{(g)}$ converge at an exponential rate to the joint and marginal distribution of θ under fairly weak conditions. Casella and George [12], Gelfand and Smith [36], and Geweke [38] provide the details.

To implement the Gibbs-sampling simulation, we have to describe the conditional posterior distributions for all parameters or parameter blocks. It is convenient to separate parameters into five blocks: the state vector s, the transition probabilities $\tilde{p}$, the regression coefficients $\tilde{\phi}$ in the conditional linear model (12), the regression coefficients $\tilde{\mu}$ in the conditional linear model (13), and the variance of the error term σ^2.

The state vector s is a first-order Markov process, which implies that given s_{t+1} all information, for example $s_{t+2}, \ldots, s_T$ and $y_{t+1}, \ldots, y_T$, is irrelevant in describing s_t. Then the posterior density of s conditional on other parameters becomes

$$p(s|\tilde{p}, \tilde{\phi}, \tilde{\mu}, \sigma^2, y) = p(s_T|\tilde{p}, \tilde{\phi}, \tilde{\mu}, \sigma^2, y) \prod_{t=1}^{T-1} p(s_t|s_{t+1}, \tilde{p}, \tilde{\phi}, \tilde{\mu}, \sigma^2, y^t) , \tag{14}$$

where $y^t = (y_1, \ldots, y_t)'$. The functional form of the posterior density suggests that we can generate draw of the state vector recursively. First we generate the last element s_T. Then, conditional on s_T, we generate s_{T-1}. More generally, conditional on s_{t+1}, we generate s_t for $t = T-1, T-2, \ldots, 1$.

To generate the state vector, Kim and Nelson [50] use the output from Hamilton's [45] filter. To facilitate exposition, we suppress the conditioning on parameters and consider first a model without lags.

Hamilton's filter starts from the observation that, before observing the data, the probability of finding the state in regime j, $\Pr(s_0 = j|y^0)$, equals the unconditional probability, $\Pr(s_t = j)$, which is proportional to the eigenvector of P associated with unitary eigenvalue.

Using transition probabilities and the probability of observing regime j conditional on observations obtained through date t, $\Pr(s_t = j|y^t)$, we predict the next period regime

$$\Pr(s_{t+1} = j|y^t) = \Pr(s_t = 0|y^t)p_{0j} + \Pr(s_t = 1|y^t)p_{1j} . \tag{15}$$

Once y_{t+1} is observed, we update the prediction using Bayes rule

$$\Pr(s_{t+1} = j|y^{t+1}) = \Pr(s_{t+1} = j|y_{t+1}, y^t) = \frac{f(y_{t+1}|s_{t+1} = j, y^t)\Pr(s_{t+1} = j|y^t)}{f(y_{t+1}|y^t)} , \tag{16}$$

where the numerator is the joint probability of observing y_{t+1} and $s_{t+1} = j$, which is a product of the probability of observing y_{t+1} given that state s_{t+1} is in regime j (for example $f(y_{t+1}|s_{t+1} = 0, y^t) = N(\mu_0, \sigma^2)$) and our prediction from Eq. (15). The denominator is the unconditional density of observing y_{t+1}, which is a sum of the numerator over all possible regimes

$$f(y_{t+1}|y^t) = \sum_j f(y_{t+1}|s_{t+1} = j, y^t)\Pr(s_{t+1} = j|y^t) . \tag{17}$$

Starting from $\Pr(s_0 = j|y^0)$, the filter iterates through Eqs. (15)–(17) until we calculate $\Pr(s_t = j|y^t)$ for every t and j. As a by-product of the filter we obtain the likelihood function

$$f(\tilde{\phi}, \tilde{\mu}, \tilde{p}, \sigma^2, s|y) = \prod_t f(y_{t+1}|y^t) . \tag{18}$$

For the AR(1) model, the filter should be adjusted. Given $\Pr(s_t = j|y^t)$, we forecast the next period regime and the previous period regime jointly, taking one summand in Eq. (15) at a time

$$\Pr(s_{t+1} = j, s_t = i|y^t) = p_{ij}\Pr(s_t = i|y^t) , \tag{15a}$$

for $j = 0, 1$ and $i = 0, 1$. After y_{t+1} is observed, we update our prediction to

$$\Pr(s_{t+1} = j, s_t = i|y^{t+1}) = \frac{f(y_{t+1}|s_{t+1} = j, s_t = i, y^t)\ \Pr(s_{t+1} = j, s_t = i|y^t)}{f(y_{t+1}|y^t)} , \tag{16a}$$

where $f(y_{t+1}|s_{t+1} = j, s_t = i, y^t)$ is the density of observing $y_t + 1$ given that state $s_t + 1$ is in regime j and state s_t is in regime i (for example $f(y_{t+1}|s_{t+1} = 0, s_t = 0, y^t) = N(\mu_0 + \phi_1(y_t - \mu_0), \sigma^2)$)

$$f(y_{t+1}|y^t) = \sum_j \sum_i f(y_{t+1}|s_{t+1} = j, s_t = i, y^t) \cdot \Pr(s_{t+1} = j, s_t = i|y^t) . \tag{17a}$$

Summing (16a) over i,

$$\Pr(s_{t+1} = j|y^{t+1}) = \sum_i \Pr(s_{t+1} = j, s_t = i|y^{t+1}), \tag{19}$$

finishes the iteration. Iterating through Eqs. (15a)–(17a) and (19) we get $\Pr(s_t = j|y^t)$ for every t and j. The extension to a more general AR(p) model is similar.

The output of Hamilton's filter gives only the first term in the product (14), which is sufficient to generate s_T. To generate the other states s_t conditional on y^t and $s_t + 1$, $t = T-1, T-2, \ldots, 1$, we again use Bayes rule

$$\Pr(s_t = j|s_{t+1} = i, y^t) = \frac{p_{ji} \Pr(s_t = j|y^t)}{\sum_j p_{ji} \Pr(s_t = j|y^t)}, \quad (20)$$

where $\Pr(s_t = j|y^t)$ is the output from Hamilton's filter. Since s_t is a discrete random variable taking on values 0 and 1, we can generate it by drawing random numbers from uniform distribution between 0 and 1, and comparing them to $\Pr(s_t = 1|s_{t+1} = i, y^t)$.

Conditional on other parameters in the model, the likelihood function of transition probabilities reduces to a simple count n_{ij} of transitions from state i to state j

$$f(\tilde{p}|\tilde{\mu}, \tilde{\phi}, \sigma_2, s, y) = p_{11}^{n_{11}}(1-p_{11})^{n_{12}} p_{22}^{n_{22}}(1-p_{22})^{n_{21}},$$

which is the product of the independent beta distributions. The posterior distribution for the transition probabilities conditional on the other parameters is a product of independent beta distributions

$$\begin{aligned} &p(\tilde{p}|\tilde{\phi}, \tilde{\mu}, \sigma^2, s, y) \\ &\quad = B(\alpha_1 + n_{11}, \beta_1 + n_{12}) \cdot B(\alpha_2 + n_{22}, \beta_2 + n_{21}). \end{aligned}$$

To derive posterior distributions for $\tilde{\phi}$, $\tilde{\mu}$, and σ^2 conditional on other parameters, we use standard results for a linear model with the natural conjugate priors. The natural conjugate priors are reviewed, for example, by Geweke [39], Koop [58], or Lancaster [61]. In particular, the conditional distribution of the regression coefficients is Normal

$$\begin{aligned} &p(\tilde{\phi}|\tilde{p}, \tilde{\mu}, \sigma^2, s, y) \\ &\quad = N\left(\Sigma_\phi \left(\sigma^{-2} M_{0,\phi}\tilde{\phi}_0 + \sigma^{-2}\sum x_t(s)' y_t(s)\right), \Sigma_\phi\right) \\ &\qquad \cdot I_A(\tilde{\phi}), \\ &p(\tilde{\mu}|\tilde{p}, \tilde{\phi}, \sigma^2, s, y) \\ &\quad = N\left(\Sigma_\mu \left(M_{0,\mu}\tilde{\mu}_0 + \sigma^{-2}\sum x_t(\tilde{\phi})' y_t(\tilde{\phi})\right), \Sigma_\mu\right) \\ &\qquad \cdot I_{(0,\infty)}(\mu_1), \end{aligned}$$

where

$$\begin{aligned} \Sigma_\phi &= \left(\sigma^{-2} M_{0,\phi} + \sigma^{-2}\sum x_t(s)' x_t(s)\right)^{-1}, \\ \Sigma_\mu &= \left(M_{0,\mu} + \sigma^{-2}\sum x_t(\tilde{\phi})' x_t(\tilde{\phi})\right)^{-1}. \end{aligned}$$

The conditional distribution for the variance of error term is Inverted Gamma-2

$$\begin{aligned} &p(\sigma^2|\tilde{p}, \tilde{\phi}, \tilde{\mu}, s, y) \\ &\quad = IG_2\left(s_0 + \sum (y_t(s_t) - x_t'(s_t)\tilde{\phi})^2, \nu_0 + T\right). \end{aligned}$$

Testing for Linearity and Model Selection

Given our prior, the linear model is not nested in the MAR model. To test against a linear model, we use the Bayes factor. We also use the Bayes factor to select the number of regimes and the number of lags.

The Bayes factor is a ratio of marginal likelihoods of the alternative models. To find the marginal likelihood, we need to integrate the product of the likelihood function and the prior density with respect to all parameters. Chib [16] shows that the marginal likelihood can be computed from the output of the Gibbs sampler requiring only that the integrating constants of the conditional posterior distributions be known. This requirement is satisfied for the natural conjugate priors.

From the Bayes's theorem it follows that the identity

$$f(y) = \frac{f(y|\theta)\pi(\theta)}{p(\theta|y)},$$

holds for any θ. The complete functional form of the numerator is given by the product of the likelihood (18) and the prior densities. Chib suggests evaluating the denominator, the posterior density, at the posterior mode θ^*. Then the posterior density at the posterior mode can be written as

$$\begin{aligned} p(\theta^*|y) &= p(\tilde{\mu}^*|y)\, p(\tilde{\phi}^*|\tilde{\mu}^*, y) \\ &\quad \cdot p(\tilde{\sigma}^{2*}|\tilde{\mu}^*, \tilde{\phi}^*, y)\, p(\tilde{p}^*|y, \mu^*, \tilde{\phi}^*, \sigma^{2*}). \end{aligned}$$

The first term

$$\begin{aligned} &p(\tilde{\mu}^*|y) = \\ &\int p(\tilde{\mu}^*|\tilde{\phi}, \sigma^2, \tilde{p}, s, y)\, p(\tilde{\phi}, \sigma^2, \tilde{p}, s|y)\, d\tilde{\phi}\, d\sigma^2\, d\tilde{p}\, ds, \end{aligned}$$

can be estimated by averaging over the full conditional density

$$\hat{p}(\tilde{\mu}^*|y) = G^{-1}\sum_{g=1}^{G} p\left(\tilde{\mu}^*|\tilde{\phi}^{(g)}, \sigma^{2(g)}, \tilde{p}^{(g)}, s^{(g)}, y\right).$$

This estimate converges at an exponential rate to the true marginal distribution of $\tilde{\mu}$.

In the second term,

$$p\left(\tilde{\phi}|\tilde{\mu}^*, y\right) = \int p\left(\tilde{\phi}^*|\tilde{\mu}^*, \sigma^2, \tilde{p}, s, y\right) p\left(\sigma^2, \tilde{p}, s|\tilde{\mu}^*, y\right) \, \mathrm{d}\sigma^2 \, \mathrm{d}\tilde{p} \, \mathrm{d}s \, ,$$

the complete conditional density of $\tilde{\phi}$ cannot be averaged directly because the Gibbs sampler does not provide draws conditional on $\tilde{\mu}^*$. We generate necessary draws by additional G iterations of the original Gibbs sampler, but instead of generating $\tilde{\mu}$ we set it equal to $\tilde{\mu}^*$. Then the estimate of the second term

$$\hat{p}\left(\tilde{\phi}^*|\tilde{\mu}^*, y\right) = G^{-1} \sum_{g=G+1}^{2G} p\left(\tilde{\phi}^*|\tilde{\mu}^*, \sigma^{2(g)}, \tilde{p}^{(g)}, s^{(g)}, y\right) \, ,$$

converges at an exponential rate to the true $p\left(\tilde{\phi}|\tilde{\mu}^*, y\right)$. Similarly, by generating additional draws from the Gibbs sampler we compute $\hat{p}\left(\tilde{\sigma}^{2*}|\tilde{\mu}^*, \tilde{\phi}^*, y\right)$ and $\hat{p}(\tilde{p}^*|y, \mu^*, \tilde{\phi}^*, \sigma^{2*})$.

Substituting our estimate of posterior density into marginal likelihood results in

$$\begin{aligned}\ln f\left(y\right) = {} & \ln f\left(y|\theta^*\right) + \ln \pi\left(\theta^*\right) - \ln \hat{p}\left(\tilde{\mu}^*|y\right) \\ & - \ln \hat{p}\left(\tilde{\phi}^*|\tilde{\mu}^*, y\right) - \ln \hat{p}\left(\tilde{\sigma}^{2*}|\tilde{\mu}^*, \tilde{\phi}^*, y\right) \\ & - \ln \hat{p}\left(\tilde{p}^*|y, \mu^*, \tilde{\phi}^*, \sigma^{2*}\right) \, .\end{aligned}$$

The model with the highest marginal likelihood is preferred.

Future Directions

Given the large volume of evidence collected in the nonlinear time series, incorporating regime-switching policies and disturbances into general equilibrium models may lead to a better understanding of monetary and fiscal policies.

Over the years, the time series literature has collected substantial statistical evidence that output, unemployment, and interest rates in the US exhibit different behavior in recessions and expansions. Contrary to the real business cycle models in which short-run and long-run fluctuations have the same origin, the statistical evidence suggests that the forces that cause output to rise may be quite different from those that cause it to fall.

Also, many studies provide evidence that monetary and fiscal policies have changed substantially throughout US history. Taylor [85], Clarida, Gali, and Gertler [17], Romer and Romer [77], and Lubik and Schorfheide [62] show that, since the mid-1980s, the Fed reacted more forcefully to inflation. Favero and Monacelli [30] and Davig and Leeper [20] demonstrate that US fiscal policy has fluctuated frequently responding to wars, recessions, and more generally to the level of debt. Sims and Zha [81], after extensive comparison of 17 regime-switching structural VAR models, report that their best-fitting model requires nine regimes to incorporate the large shocks, for example, generated by the OPEC oil embargo or the Vietnam War. They conclude that, "It is time to abandon the idea that policy change is best modelled as a once-and-for-all, nonstochastic regime switch" (p. 56).

The research by Davig and Leeper [19,20,21] and Farmer, Waggoner, and Zha [27,28,29] show considerable promise in introducing nonlinear regime-switching components into dynamic stochastic general equilibrium models. For example, Davig and Leeper [20] estimate regime-switching rules for monetary policy and tax policy and incorporate them into the otherwise standard new-Keynesian model. Unlike expansionary fiscal policy in the fixed-regime model, fiscal expansion in the regime-switching model increases inflation and output.

Acknowledgments

The author is grateful to Bruce Mizrach, the Finance and Econometrics Section editor, as well as to David Harless, Carol S. Lehr, Ming Lo, Stan Radchenko, Philip Rothman, and Tara Sinclair for many helpful comments.

Bibliography

1. Anderson HM (1997) Transaction costs and nonlinear adjustment towards equilibrium in the US treasury bill market. Oxf Bull Econ Stat 59:465–484
2. Ang A, Bekaert G (2002) International asset allocation with regime shifts. Rev Financ Stud 15:1137–1187
3. Ang A, Bekaert G (2002) Regime switches in interest rates. J Bus Econ Stat 20:163–197
4. Albert JH, Chib S (1993) Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. J Bus Econ Stat 11(1):1–15
5. Astatkie T, Watts DG, Watt WE (1997) Nested threshold autoregressive NeTAR models. Int J Forecast 13:105–116
6. Bacon DW, Watts DG (1971) Estimating the transition between intersecting straight lines. Biometrica 62:525–534
7. Bauwens L, Lubrano M, Richard JF (1999) Bayesian Inference in Dynamic Econometric Models. Oxford University Press, New York
8. Billio M, Monfort A, Robert CP (1999) Bayesian estimation of switching ARMA models. J Econ 93:229–255
9. Box GEP, Jenkins GM (1970) Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco
10. Cai J (1994) A Markov model of switching-regime ARCH. J Bus Econ Stat 12:309–316
11. Calvet L, Fisher A (2004) Regime switching and the estimation of multifractal processes. J Financ Econ 2:49–83
12. Casella G, George EI (1992) Explaining the Gibbs sampler. Am Stat 46:167–174

13. Cerra V, Saxena SC (2005) Did output recover from the Asian crisis? IMF Staff Papers 52:1–23
14. Chan KS, Tong H (1986) On estimating thresholds in autoregressive models. J Time Ser Anal 7:178–190
15. Chen CWS, Lee JC (1995) Bayesian inference of threshold autoregressive models. J Time Ser Anal 16:483–492
16. Chib S (1995) Marginal likelihood from the Gibbs output. J Am Stat Assoc 90:1313–1321
17. Clarida R, Gali J, Gertler M (2000) Monetary policy rules and macroeconomic stability: evidence and some theory. Q J Econ 115:147–180
18. Dai Q Singleton KJ, Yang W (2007) Regime shifts in a dynamic term structure model of US treasury bonds. Rev Financ Stud 20(5):1669–1706
19. Davig T, Leeper E (2005) Generalizing the Taylor principle. National Bureau of Economic Research, Working Paper No 11874
20. Davig T, Leeper E (2006) Fluctuating macro policies and the fiscal theory. In: Acemoglu D, Rogoff, Woodford M (eds) NBER Macroeconomic Annual. MIT Press, Cambridge
21. Davig T, Leeper E (2007) Generalizing the Taylor principle. Am Econ Rev 97(3):607–635
22. Diebold FX, Lee JH, Weinbach GC (1994) Regime switching with time-varying transition probabilities. In: Hargreaves C (ed) Nonstationary Time Series Analysis and Cointegration. Oxford University Press, Oxford
23. Diebold FX, Rudebusch GD (1996) Measuring business cycles: a modern perspective. Rev Econ Stat 78:67–77
24. DeJong DN (1996) A Bayesian Search for Structural Breaks in US GNP. In: Fomby TB (ed) Advances in Econometrics: Bayesian Methods Applied to Time Series Data, vol 11, part B. JAI Press, Greenwich, Connecticut, pp 109–146
25. Dueker M (1997) Markov switching in GARCH processes and mean-reverting stock-market volatility. J Bus Econ Stat 15:26–34
26. Dwyer GP, Locke P, Yu W (1996) Index arbitrage and nonlinear dynamics between the S&P 500 futures and cash. Rev Financ Stud 9:301–332
27. Farmer RE, Waggoner DF, Zha T (2006) Indeterminacy in a forward looking regime switching model. NBER Working Paper No 12540
28. Farmer RE, Waggoner DF, Zha T (2006) Minimal state variable solutions to Markov-switching rational expectations models. (unpublished manuscript)
29. Farmer RE, Waggoner DF, Zha T (2007) Understanding the New-Keynesian model when monetary policy switches regimes. (unpublished manuscript)
30. Favero CA, Monacelli T (2005) Fiscal policy rules and regime (in)stability: evidence from the US Manuscript, IGIER
31. Ferreira PE (1975) A Bayesian analysis of a switching regression model: known number of regimes. J Am Stat Assoc 70:370–374
32. Filardo AJ (1994) Business cycle phases and their transitional dynamics. J Bus Econ Stat 12:299–308
33. Forbes CS, Kalb GRJ, Kofman P (1999) Bayesian arbitrage threshold analysis. J Bus Econ Stat 17:364–372
34. Franses PH, van Dijk D (2000) Nonlinear Time Series Models in Empirical Finance. Cambridge University Press, Cambridge
35. Garcia R, Perron P (1996) An analysis of real interest under regime shift. Rev Econ Stat 78:111–125
36. Gelfand AE, Smith AFM (1990) Sampling-based approaches to calculating marginal densities. J Am Stat Assoc 85(410): 398–409
37. Geweke J, Terui N (1993) Bayesian threshold autoregressive models for nonlinear time series. J Time Ser Anal 14: 441–454
38. Geweke J (1999) Using simulation methods for Bayesian econometric models: inference, development and communication. Econ Rev 18:1–127
39. Geweke J (2005) Contemporary Bayesian Econometrics and Statistics. Wiley, Hoboken
40. Goldfeld SM, Quandt RE (1973) A Markov model for switching regressions. J Econ 1:3–16
41. Granger CWJ, Terasvirta T (1993) Modeling Nonlinear Economic Relationships. Oxford University Press, Oxford
42. Granger CWJ, Swanson NR (1996) Future developments in the study of cointegrated variables. Oxf Bull Econ Stat 58: 537–553
43. Gray SF (1996) Modeling the conditional distribution of interest rates as a regime-switching process. J Financ Econ 42:27–62
44. Hamilton JD (1988) Rational-expectations econometric analysis of changes in regime: an investigation of the term structure of interest rates. J Econ Dyn Control 12:385–423
45. Hamilton JD (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57(2):357–384
46. Hamilton JD (1994) Time Series Analysis. Princeton University Press, Princeton
47. Hamilton JD, Susmel R (1994) Autoregressive conditional heteroskedasticity and changes in regime. J Econ 64:207–333
48. Hamilton JD (2005) Whats real about the business cycle? Fed Reserve Bank St Louis Rev 87(4):435–452
49. Jeanne O, Masson P (2000) Currency crises, sunspots, and Markov- switching regimes. J Int Econ 50:327–350
50. Kim CJ, Nelson CR (1998) Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime-switching. Rev Econ Stat 80(2):188–201
51. Kim CJ, Nelson CR (1999) Friedman's plucking model of business fluctuations: tests and estimates of permanent and transitory components. J Money Credit Bank 31:317–334
52. Kim CJ, Nelson CR (1999) State-Space Models with Regime Switching: Classical and Gibbs-sampling Approaches with Applications. MIT Press, Cambridge
53. Kim CJ, Piger J (2002) Common stochastic trends, common cycles, and asymmetry in economic fluctuations. J Monet Econ 49:1189–1211
54. Koop G, Potter SM (1999) Dynamic asymmetries in US unemployment. J Bus Econ Stat 17:198–312
55. Koop G, Potter SM (1999) Bayes factors and nonlinearity: evidence from economic time series. J Econ 88:251–281
56. Koop G, Potter SM (2000) Nonlinearity, structural breaks or outliers in economic time series? In: Barnett B, Johansen S (eds) Nonlinear Econometric Modeling in Time Series. Cambridge University Press, Cambridge
57. Koop G, Potter SM (2003) Bayesian analysis of endogenous delay threshold models. J Bus Econ Stat 21(1):93–103
58. Koop G (2003) Bayesian Econometrics. Wiley, Chichester
59. Korenok O, Radchenko R (2005) The smooth transition autoregressive target zone model with the Gaussian stochastic

volatility and TGARCH error terms with applications. VCU Economics Department, No 0505
60. Krolzig HM (1997) Markov-Switching Vector Autoregressions: Modeling, Statistical Inference, and Application to Business Cycle Analysis. Springer, Berlin
61. Lancaster T (2004) An Introduction to Modern Bayesian Econometrics. Blackwell Publishing, Malden
62. Lubik TA, Schorfheide F (2004) Testing for indeterminacy: an application to US monetary policy. Am Econ Rev 94:190–217
63. Lubrano M (1999) Bayesian Analysis of Nonlinear Time Series Models with a threshold. In: Barnett WA, Hendry DF, Hylleberg S, Terasvirta T, Tjostheim D, Wurts A (eds) Nonlinear Econometric Modeling. Cambridge University Press, Cambridge
64. Lundbergh S, Terasvirta T (1998) Modelling economic high-frequency time series with STAR-GARCH models. Working Paper Series in Economics and Finance No. 291, Stockholm School of Economics
65. Lundbergh S, Terasvirta T (2006) A time series model for an exchange rate in a target zone with applications. J Econ 131:579–609
66. Martens M, Kofman P, Vorst ACF (1998) A threshold error correction for intraday futures and index returns. J Appl Econ 13:245–263
67. Michael P, Nobay AR, Peel DA (1997) Transaction costs and nonlinear adjustment in real exchange rates: an empirical investigation. J Political Econ 105:862–879
68. Obstfeld M, Taylor AM (1997) Nonlinear aspects of goods-market arbitrage and adjustment: Heckscher's commodity points revisited. J Japan Int Econ 11:441–479
69. Peria MSM (2002) A regime-switching approach to the study of speculative attacks: a focus on EMS crises. In: Hamilton JD, Raj B (eds) Advances in Markov-Switching Models. Physica-Verlag, Heidelberg
70. Perron P, Vogelsang TJ (1992) Nonstationarity and level shifts with an application to purchasing power parity. J Bus Econ Stat 10:301–320
71. Pesaran MH, Potter S (1997) A floor and ceiling model of US output. J Econ Dyn Control 21:661–695
72. Pfann GA, Schotman PC, Tschernig R (1996) Nonlinear interest rate dynamics and the implications for the term structure. J Econ 74:149–176
73. Poter SM (1995) A nonlinear approach to US GNP. J Appl Econ 10:109–125
74. Potter SM (1999) Nonlinear time series modelling: an introduction. J Econ Surv 13:505–528
75. Quandt RE (1958) The estimation of the parameters of a linear regression system obeying two separate regimes. J Am Stat Assoc 53:873–880
76. Radchenko S (2005) Lags in the response of gasoline prices to changes in crude oil prices: the role of short-term and long-term shocks. Energy Econ 27:573–602
77. Romer CD, Romer DH (2002) A rehabilitation of monetary policy in the 1950s. Am Econ Rev 92:121–127
78. Rothman P (1998) Forecasting asymmetric unemployment rates. Rev Econ Stat 80:164–168
79. Rothman P, van Dijk D, Franses PH (2001) A multivariate STAR analysis of the relationship between money and output. Macroecon Dyn 5:506–532
80. Sarantis N (1999) Modeling non-linearities in real effective exchange rates. J Int Money Finance 18:27–45
81. Sims C, Zha T (2006) Were there switches in US monetary policy? Am Econ Rev 96:54–81
82. Skalin J, Terasvirta T (1999) Another look at swedish business cycles. J Appl Econ 14:359–378
83. Smith DR (2002) Markov-switching and stochastic volatility diffusion models of short-term interest rates. J Bus Econ Stat 20:183–197
84. Steel M (2008) Bayesian Time Series Analysis. In: Durlauf S, Blume L (eds) The New Palgrave Dictionary of Economics, 2nd ed. Palgrave Macmillan, London
85. Taylor JB (1999) An historical analysis of monetary policy rules. In: Taylor JB (ed) Monetary Policy Rules, pp 319–341
86. Taylor N, van Dijk D, Franses PH, Lucas A (2000) SETS, arbitrage activity, and stock price dynamics. J Bank Finance 24: 1289–1306
87. Taylor MP, Peel DA, Sarno L (2001) Nonlinear mean-reversion in exchange rate rates: towards a solution to the purchasing power parity puzzles. Int Econ Rev 42:1015–1042
88. Terasvirta T, Anderson H (1992) Characterising nonlinearities in business cycles using smooth transition autoregressive models. J Appl Econom 7S:119–136
89. Terasvirta T (1994) Specification, estimation and evaluation of smooth transition autoregressive models. J Am Stat Assoc 89:208–219
90. Terasvirta T (1998) Modelling economic relationships with smooth transition regressions. In Handbook of Appl Economic Statistics. Marcel Dekker, New York, pp 507–552
91. Tiao CG, Tsay RS (1994) Some advances in non linear and adaptive modelling in time series. J Forecast 13:109–131
92. Tong H (1978) On a threshold model. In: Chan CH (ed) Pattern Recognition and Signal Processing. Sijthoff and Noordhoff, Amsterdam
93. Tong H (1983) Threshold Models in Non-Linear Time Series Analysis. Lecture Notes in Statistics, no 21. Springer, Heidelberg
94. Tong H, Lim KS (1980) Threshold autoregression, limit cycles and cyclical data. J Royal Stat Soc B 42:245–292
95. Tong H (1990) Nonlinear Time Series: A Dynamical System Approach. Oxford University Press, Oxford
96. Tsay RS (1998) Testing and modeling multivariate threshold models. J Am Stat Assoc 93:1188–1202
97. van Dijk D, Franses PH (1999) Modeling multiple regimes in the business cycle. Macroecon Dyn 3:311–340
98. van Dijk D, Terasvirta T, Franses PH (2002) Smooth transition autoregressive models – a survey of recent developments. Econ Rev 21:1–47
99. Weise CL (1999) The Asymmetric effects of monetary policy. J Money Credit Bank 31:85–108
100. Zellner A (1971) An Introduction to Bayesian Inference in Econometrics. Wiley, New York

Bayesian Statistics

DAVID DRAPER
Department of Applied Mathematics and Statistics, Baskin School of Engineering, University of California, Santa Cruz, USA

Article Outline

Glossary
Definition of the Subject and Introduction
The Bayesian Statistical Paradigm
Three Examples
Comparison with the Frequentist Statistical Paradigm
Future Directions
Bibliography

Glossary

Bayes' theorem; prior, likelihood and posterior distributions Given (a) θ, something of interest which is unknown to the person making an uncertainty assessment, conveniently referred to as You, (b) y, an information source which is relevant to decreasing Your uncertainty about θ, (c) a desire to learn about θ from y in a way that is both internally and externally logically consistent, and (d) $\mathcal{B}$, Your background assumptions and judgments about how the world works, as these assumptions and judgments relate to learning about θ from y, it can be shown that You are compelled in this situation to reason within the standard rules of probability as the basis of Your *inferences* about θ, *predictions* of future data y^*, and *decisions* in the face of uncertainty (see below for contrasts between inference, prediction and decision-making), and to quantify Your uncertainty about any unknown quantities through conditional probability distributions. When inferences about θ are the goal, Bayes' Theorem provides a means of combining all relevant information internal and external to y:

$$p(\theta|y,\mathcal{B}) = c\, p(\theta|\mathcal{B})\, l(\theta|y,\mathcal{B})\,. \tag{1}$$

Here, for example in the case in which θ is a real-valued vector of length k, (a) $p(\theta|\mathcal{B})$ is Your *prior distribution* about θ given $\mathcal{B}$ (in the form of a probability density function), which quantifies all relevant information available to You about θ external to y, (b) c is a positive normalizing constant, chosen to make the density on the left side of the equation integrate to 1, (c) $l(\theta|y,\mathcal{B})$ is Your *likelihood distribution* for θ given y and $\mathcal{B}$, which is defined to be a density-normalized multiple of Your *sampling distribution* $p(\cdot|\theta,\mathcal{B})$ for future data values y^* given θ and $\mathcal{B}$, but re-interpreted as a function of θ for fixed y, and (d) $p(\theta|y,\mathcal{B})$ is Your *posterior distribution* about θ given y and $\mathcal{B}$, which summarizes Your current total information about θ and solves the basic inference problem.

Bayesian parametric and non-parametric modeling (1) Following de Finetti [23], a Bayesian statistical model is a joint predictive distribution $p(y_1,\ldots,y_n)$ for observable quantities y_i that have not yet been observed, and about which You are therefore uncertain. When the y_i are real-valued, often You will not regard them as probabilistically *independent* (informally, the y_i are independent if information about any of them does not help You to predict the others); but it may be possible to identify a *parameter vector* $\theta = (\theta_1,\ldots,\theta_k)$ such that You would judge the y_i *conditionally independent* given θ, and would therefore be willing to model them via the relation

$$p(y_1,\ldots,y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta)\,. \tag{2}$$

When combined with a prior distribution $p(\theta)$ on θ that is appropriate to the context, this is *Bayesian parametric modeling*, in which $p(y_i|\theta)$ will often have a standard distributional form (such as binomial, Poisson or Gaussian). (2) When a (finite) parameter vector that induces conditional independence cannot be found, if You judge your uncertainty about the real-valued y_i *exchangeable* (see below), then a representation theorem of de Finetti [21] states informally that all internally logically consistent predictive distributions $p(y_1,\ldots,y_n)$ can be expressed in a way that is equivalent to the *hierarchical model* (see below)

$$\begin{aligned} (F|\mathcal{B}) &\sim p(F|\mathcal{B}) \\ (y_i|F,\mathcal{B}) &\overset{\text{IID}}{\sim} F\,, \end{aligned} \tag{3}$$

where (a) F is the cumulative distribution function (CDF) of the underlying process $(y_1, y_2, \ldots)$ from which You are willing to regard $p(y_1,\ldots,y_n)$ as (in effect) like a random sample and (b) $p(F|\mathcal{B})$ is Your prior distribution on the space $\mathcal{F}$ of all CDFs on the real line. This (placing probability distributions on infinite-dimensional spaces such as $\mathcal{F}$) is *Bayesian non-parametric modeling*, in which priors involving *Dirichlet processes* and/or *Pólya trees* (see Sect. "Inference: Parametric and Non-Parametric Modeling of Count Data") are often used.

Exchangeability A sequence $y = (y_1,\ldots,y_n)$ of random variables (for $n \geq 1$) is *(finitely) exchangeable* if the joint probability distribution $p(y_1,\ldots,y_n)$ of the elements of y is invariant under permutation of the indices $(1,\ldots,n)$, and a countably infinite sequence $(y_1, y_2, \ldots)$ is *(infinitely) exchangeable* if every finite subsequence is finitely exchangeable.

Hierarchical modeling Often Your uncertainty about something unknown to You can be seen to have a *nested* or *hierarchical* character. One class of examples arises in *cluster sampling* in fields such as education and medicine, in which students (level 1) are nested within classrooms (level 2) and patients (level 1) within hospitals (level 2); cluster sampling involves random samples (and therefore uncertainty) at two or more levels in such a data hierarchy (examples of this type of hierarchical modeling are given in Sect. "Strengths and Weaknesses of the Two Approaches"). Another, quite different, class of examples of Bayesian hierarchical modeling is exemplified by equation (3) above, in which is was helpful to decompose Your overall predictive uncertainty about $(y_1, \ldots, y_n)$ into (a) uncertainty about F and then (b) uncertainty about the y_i given F (examples of this type of hierarchical modeling appear in Sect. "Inference and Prediction: Binary Outcomes with No Covariates" and "Inference: Parametric and Non-Parametric Modeling of Count Data").

Inference, prediction and decision-making; samples and populations Given a data source y, *inference* involves drawing probabilistic conclusions about the underlying process that gave rise to y, *prediction* involves summarizing uncertainty about future observable data values y^*, and decision-making involves looking for optimal behavioral choices in the face of uncertainty (about either the underlying process, or the future, or both). In some cases inference takes the form of reasoning backwards from a *sample* of data values to a *population*: a (larger) universe of possible data values from which You judge that the sample has been drawn in a manner that is *representative* (i. e., so that the sampled and unsampled values in the population are (likely to be) similar in relevant ways).

Mixture modeling Given y, unknown to You, and $\mathcal{B}$, Your background assumptions and judgments relevant to y, You have a choice: You can either model (Your uncertainty about) y directly, through the probability distribution $p(y|\mathcal{B})$, or (if that is not feasible) You can identify a quantity x upon which You judge y to depend and model y hierarchically, in two stages: first by modeling x, through the probability distribution $p(x|\mathcal{B})$, and then by modeling y given x, through the probability distribution $p(y|x, \mathcal{B})$:

$$p(y|\mathcal{B}) = \int_{\mathcal{X}} p(y|x, \mathcal{B})\, p(x|\mathcal{B})\, \mathrm{d}x \,, \tag{4}$$

where $\mathcal{X}$ is the space of possible values of x over which Your uncertainty is expressed. This is *mixture modeling*, a special case of hierarchical modeling (see above). In hierarchical notation (4) can be re-expressed as

$$y = \left\{ \begin{array}{c} x \\ (y|x) \end{array} \right\} . \tag{5}$$

Examples of mixture modeling in this article include (a) equation (3) above, with F playing the role of x; (b) the basic equation governing Bayesian prediction, discussed in Sect. "The Bayesian Statistical Paradigm"; (c) Bayesian model averaging (Sect. "The Bayesian Statistical Paradigm"); (d) de Finetti's representation theorem for binary outcomes (Sect. "Inference and Prediction: Binary Outcomes with No Covariates"); (e) random-effects parametric and non-parametric modeling of count data (Sect. "Inference: Parametric and Non-Parametric Modeling of Count Data"); and (f) integrated likelihoods in Bayes factors (Sect. "Decision-Making: Variable Selection in Generalized Linear Models; Bayesian Model Selection").

Probability – frequentist and Bayesian In the *frequentist* probability paradigm, attention is restricted to phenomena that are inherently repeatable under (essentially) identical conditions; then, for an event A of interest, $P_f(A)$ is the limiting relative frequency with which A occurs in the (hypothetical) repetitions, as the number of repetitions $n \to \infty$. By contrast, Your Bayesian probability $P_B(A|\mathcal{B})$ is the numerical weight of evidence, given Your background information $\mathcal{B}$ relevant to A, in favor of a true-false proposition A whose truth status is uncertain to You, obeying a series of reasonable axioms to ensure that Your Bayesian probabilities are internally logically consistent.

Utility To ensure internal logical consistency, optimal decision-making proceeds by (a) specifying a *utility* function $U(a, \theta_0)$ quantifying the numerical value associated with taking action a if the unknown is really θ_0 and (b) *maximizing expected utility*, where the expectation is taken over uncertainty in θ as quantified by the posterior distribution $p(\theta|y, \mathcal{B})$.

Definition of the Subject and Introduction

Statistics may be defined as the study of uncertainty: how to measure it, and how to make choices in the face of it. Uncertainty is quantified via *probability*, of which there are two leading paradigms, *frequentist* (discussed in Sect. "Comparison with the Frequentist Statistical Paradigm") and *Bayesian*. In the Bayesian approach to probability the primitive constructs are true-false *propositions* A whose truth status is uncertain, and the probability of A is the numerical weight of evidence in favor of A,

constrained to obey a set of axioms to ensure that Bayesian probabilities are *coherent* (internally logically consistent).

The discipline of statistics may be divided broadly into four activities: *description* (graphical and numerical summaries of a data set y, without attempting to reason outward from it; this activity is almost entirely non-probabilistic and will not be discussed further here), *inference* (drawing probabilistic conclusions about the underlying process that gave rise to y), *prediction* (summarizing uncertainty about future observable data values y^*), and *decision-making* (looking for optimal behavioral choices in the face of uncertainty). Bayesian statistics is an approach to inference, prediction and decision-making that is based on the Bayesian probability paradigm, in which uncertainty about an unknown θ (this is the inference problem) is quantified by means of a conditional probability distribution $p(\theta|y, \mathcal{B})$; here y is all available relevant data and $\mathcal{B}$ summarizes the background assumptions and judgments of the person making the uncertainty assessment. Prediction of a future y^* is similarly based on the conditional probability distribution $p(y^*|y, \mathcal{B})$, and optimal decision-making proceeds by (a) specifying a *utility* function $U(a, \theta_0)$ quantifying the numerical reward associated with taking action a if the unknown is really θ_0 and (b) *maximizing expected utility*, where the expectation is taken over uncertainty in θ as quantified by $p(\theta|y, \mathcal{B})$.

The Bayesian Statistical Paradigm

Statistics is the branch of mathematical and scientific inquiry devoted to the study of uncertainty: its consequences, and how to behave sensibly in its presence. The subject draws heavily on *probability*, a discipline which predates it by about 100 years: basic probability theory can be traced [48] to work of Pascal, Fermat and Huygens in the 1650s, and the beginnings of statistics [34,109] are evident in work of Bayes published in the 1760s.

The Bayesian statistical paradigm consists of three basic ingredients:

- θ, something of interest which is unknown (or only partially known) to the person making the uncertainty assessment, conveniently referred to, in a convention proposed by Good (1950), as You. Often θ is a *parameter* vector of real numbers (of finite length k, say) or a matrix, but it can literally be almost anything: for example, a function (three leading examples are a cumulative distribution function (CDF), a density, or a regression surface), a phylogenetic tree, an image of a region on the surface of Mars at a particular moment in time, ...
- y, an information source which is relevant to decreasing Your uncertainty about θ. Often y is a vector of real numbers (of finite length n, say), but it can also literally be almost anything: for instance, a time series, a movie, the text in a book, ...
- A desire to learn about θ from y in a way that is both *coherent* (internally consistent: in other words, free of internal logical contradictions; Bernardo and Smith [11] give a precise definition of coherence) and *well-calibrated* (externally consistent: for example, capable of making accurate predictions of future data y^*).

It turns out [23,53] that You are compelled in this situation to reason within the standard rules of probability (see below) as the basis of Your inferences about θ, predictions of future data y^*, and decisions in the face of uncertainty, and to quantify Your uncertainty about any unknown quantities through conditional probability distributions, as in the following three basic equations of Bayesian statistics:

$$\begin{aligned} p(\theta|y, \mathcal{B}) &= c\, p(\theta|\mathcal{B})\, l(\theta|y, \mathcal{B}) \\ p(y^*|y, \mathcal{B}) &= \int_\Theta p(y^*|\theta, \mathcal{B})\, p(\theta|y, \mathcal{B})\, \mathrm{d}\theta \\ a^* &= \operatorname*{argmax}_{a\in\mathcal{A}} E_{(\theta|y,\mathcal{B})}\left[U(a, \theta)\right] . \end{aligned} \tag{6}$$

(The basic rules of probability [71] are: for any true-false propositions A and B and any background assumptions and judgments $\mathcal{B}$, (*convexity*) $0 \le P(A|\mathcal{B}) \le 1$, with equality at 1 iff A is known to be true under $\mathcal{B}$; (*multiplication*) $P(A \text{ and } B|\mathcal{B}) = P(A|\mathcal{B})\, P(B|A, \mathcal{B}) = P(B|\mathcal{B})\, P(A|B, \mathcal{B})$; and (*addition*) $P(A \text{ or } B|\mathcal{B}) = P(A|\mathcal{B}) + P(B|\mathcal{B}) - P(A \text{ and } B|\mathcal{B})$.)

The meaning of the equations in (6) is as follows.

- $\mathcal{B}$ stands for Your background (often not fully stated) assumptions and judgments about how the world works, as these assumptions and judgments relate to learning about θ from y. $\mathcal{B}$ is often omitted from the basic equations (sometimes with unfortunate consequences), yielding the simpler-looking forms

$$\begin{aligned} p(\theta|y) &= c\, p(\theta)\, l(\theta|y) \\ p(y^*|y) &= \int_\Theta p(y^*|\theta)\, p(\theta|y)\, \mathrm{d}\theta \\ a^* &= \operatorname*{argmax}_{a\in\mathcal{A}} E_{(\theta|y)}\left[U(a, \theta)\right] . \end{aligned} \tag{7}$$

- $p(\theta|\mathcal{B})$ is Your *prior information* about θ given $\mathcal{B}$, in the form of a probability density function (PDF) or probability mass function (PMF) if θ lives continuously or discretely on $\mathbb{R}^k$ (this is generically referred to as

Your *prior distribution*), and $p(\theta|y,\mathcal{B})$ is Your *posterior distribution* about θ given y and $\mathcal{B}$, which summarizes Your current total information about θ and solves the basic inference problem. These are actually not very good names for $p(\theta|\mathcal{B})$ and $p(\theta|y,\mathcal{B})$, because (for example) $p(\theta|\mathcal{B})$ really stands for all (relevant) information about θ (given $\mathcal{B}$) external to y, whether that information was obtained before (or after) y arrives, but (a) they do emphasize the sequential nature of learning and (b) through long usage it would be difficult for more accurate names to be adopted.

- c (here and throughout) is a generic positive normalizing constant, inserted into the first equation in (6) to make the left-hand side integrate (or sum) to 1 (as any coherent distribution must).
- $p(y^*|\theta,\mathcal{B})$ is Your *sampling distribution* for future data values y^* given θ and $\mathcal{B}$ (and presumably You would use the same sampling distribution $p(y|\theta,\mathcal{B})$ for (past) data values y, mentally turning the clock back to a point before the data arrives and thinking about what values of y You might see). This assumes that You are willing to regard Your data as like random draws from a *population* of possible data values (an heroic assumption in some cases, for instance with observational rather than randomized data; this same assumption arises in the frequentist statistical paradigm, discussed below in Sect. "Comparison with the Frequentist Statistical Paradigm").
- $l(\theta|y,\mathcal{B})$ is Your *likelihood function* for θ given y and $\mathcal{B}$, which is defined to be any positive constant multiple of the sampling distribution $p(y|\theta,\mathcal{B})$ but re-interpreted as a function of θ for fixed y:

$$l(\theta|y,\mathcal{B}) = c\, p(y|\theta,\mathcal{B})\,. \tag{8}$$

The likelihood function is also central to one of the main approaches to frequentist statistical inference, developed by Fisher [37]; the two approaches are contrasted in Sect."Comparison with the Frequentist Statistical Paradigm".

All of the symbols in the first equation in (6) have now been defined, and this equation can be recognized as *Bayes' Theorem*, named after Bayes [5] because a special case of it appears prominently in work of his that was published posthumously. It describes how to pass coherently from information about θ external to y (quantified in the prior distribution $p(\theta|\mathcal{B})$) to information both internal and external to y (quantified in the posterior distribution $p(\theta|y,\mathcal{B})$), via the likelihood function $l(\theta|y,\mathcal{B})$: You multiply the prior and likelihood pointwise in θ and normalize so that the posterior distribution $p(\theta|y,\mathcal{B})$ integrates (or sums) to 1.

- According to the second equation in (6), $p(y^*|y,\mathcal{B})$, Your (posterior) *predictive distribution* for future data y^* given (past) data y and $\mathcal{B}$, which solves the basic prediction problem, must be a weighted average of Your sampling distribution $p(y^*|\theta,\mathcal{B})$ weighted by Your current best information $p(\theta|y,\mathcal{B})$ about θ given y and $\mathcal{B}$; in this integral Θ is the space of possible values of θ over which Your uncertainty is expressed. (The second equation in (6) contains a simplifying assumption that should be mentioned: in full generality the first term $p(y^*|\theta,\mathcal{B})$ inside the integral would be $p(y^*|y,\theta,\mathcal{B})$, but it is almost always the case that the information in y is redundant in the presence of complete knowledge of θ, in which case $p(y^*|y,\theta,\mathcal{B}) = p(y^*|\theta,\mathcal{B})$; this state of affairs could be described by saying that *the past and future are conditionally independent given the truth.* A simple example of this phenomenon is provided by coin-tossing: if You are watching a Bernoulli(θ) process unfold (see Sect. "Inference and Prediction: Binary Outcomes with No Covariates") whose success probability θ is unknown to You, the information that 8 of the first 10 tosses have been heads is definitely useful to You in predicting the 11th toss, but if instead You somehow knew that θ was 0.7, the outcome of the first 10 tosses would be irrelevant to You in predicting any future tosses.)
- Finally, in the context of making a choice in the face of uncertainty, $\mathcal{A}$ is Your set of possible actions, $U(a,\theta_0)$ is the numerical value (*utility*) You attach to taking action a if the unknown is really θ_0 (specified, without loss of generality, so that large utility values are preferred by You), and the third equation in (6) says that to make the choice coherently You should find the action a^* that *maximizes expected utility* (MEU); here the expectation

$$E_{(\theta|y,\mathcal{B})}\left[U(a,\theta)\right] = \int_{\Theta} U(a,\theta)\, p(\theta|y,\mathcal{B})\, \mathrm{d}\theta \tag{9}$$

is taken over uncertainty in θ as quantified by the posterior distribution $p(\theta|y,\mathcal{B})$.

This summarizes the entire Bayesian statistical paradigm, which is driven by the three equations in (6). Examples of its use include clinical trial design [56] and analysis [105]; spatio-temporal modeling, with environmental applications [101]; forecasting and dynamic linear models [115]; non-parametric estimation of receiver operating characteristic curves, with applications in medicine and agriculture [49]; finite selection models, with health policy applications [79]; Bayesian CART model search, with applications in breast cancer research [16]; construc-

tion of radiocarbon calibration curves, with archaeological applications [15]; factor regression models, with applications to gene expression data [114]; mixture modeling for high-density genotyping arrays, with bioinformatic applications [100]; the EM algorithm for Bayesian fitting of latent process models [76]; state-space modeling, with applications in particle-filtering [92]; causal inference [42,99]; hierarchical modeling of DNA sequences, with genetic and medical applications [77]; hierarchical Poisson regression modeling, with applications in health care evaluation [17]; multiscale modeling, with engineering and financial applications [33]; expected posterior prior distributions for model selection [91]; nested Dirichlet processes, with applications in the health sciences [96]; Bayesian methods in the study of sustainable fisheries [74,82]; hierarchical non-parametric meta-analysis, with medical and educational applications [81]; and structural equation modeling of multilevel data, with applications to health policy [19].

Challenges to the paradigm include the following:

- **Q:** How do You specify the sampling distribution/likelihood function that quantifies the information about the unknown θ internal to Your data set y? **A:** (1) The solution to this problem, which is common to all approaches to statistical inference, involves imagining future data y^* from the same process that has yielded or will yield Your data set y; often the variability You expect in future data values can be quantified (at least approximately) through a standard parametric family of distributions (such as the Bernoulli/binomial for binary data, the Poisson for count data, and the Gaussian for real-valued outcomes) and the parameter vector of this family becomes the unknown θ of interest. (2) Uncertainty in the likelihood function is referred to as *model uncertainty* [67]; a leading approach to quantifying this source of uncertainty is *Bayesian model averaging* [18,25,52], in which uncertainty about the models M in an ensemble $\mathcal{M}$ of models (specifying $\mathcal{M}$ is part of $\mathcal{B}$) is assessed and propagated for a quantity, such as a future data value y^*, that is common to all models via the expression

$$p(y^*|y,\mathcal{B}) = \int_{\mathcal{M}} p(y^*|y,M,\mathcal{B})\, p(M|y,\mathcal{B})\, \mathrm{d}M\,. \quad (10)$$

  In other words, to make coherent predictions in the presence of model uncertainty You should form a weighted average of the conditional predictive distributions $p(y^*|y,M,\mathcal{B})$, weighted by the posterior model probabilities $p(M|y,\mathcal{B})$. Other potentially useful approaches to model uncertainty include *Bayesian non-parametric modeling*, which is examined in Sect. "Inference: Parametric and Non-Parametric Modeling of Count Data", and methods based on *cross-validation* [110], in which (in Bayesian language) part of the data is used to specify the prior distribution on $\mathcal{M}$ (which is an input to calculating the posterior model probabilities) and the rest of the data is employed to update that prior.
- **Q:** How do You quantify information about the unknown θ external to Your data set y in the prior probability distribution $p(\theta|\mathcal{B})$? **A:** (1) There is an extensive literature on *elicitation* of prior (and other) probabilities; notable references include O'Hagan et al. [85] and the citations given there. (2) If θ is a parameter vector and the likelihood function is a member of the *exponential family* [11], the prior distribution can be chosen in such a way that the prior and posterior distributions for θ have the same mathematical form (such a prior is said to be *conjugate* to the given likelihood); this may greatly simplify the computations, and often prior information can (at least approximately) be quantified by choosing a member of the conjugate family (see Sect."Inference and Prediction: Binary Outcomes with No Covariates" for an example of both of these phenomena).

  In situations where it is not precisely clear how to quantify the available information external to y, two sets of tools are available:
  - *Sensitivity analysis* [30], also known as *pre-posterior analysis* [4]: Before the data have begun to arrive, You can (a) generate data similar to what You expect You will see, (b) choose a plausible prior specification and update it to the posterior on the quantities of greatest interest, (c) repeat (b) across a variety of plausible alternatives, and (d) see if there is substantial stability in conclusions across the variations in prior specification. If so, fine; if not, this approach can be combined with hierarchical modeling [68]: You can collect all of the plausible priors and add a layer hierarchically to the prior specification, with the new layer indexing variation across the prior alternatives.
  - *Bayesian robustness* [8,95]: If, for example, the context of the problem implies that You only wish to specify that the prior distribution belongs to an infinite-dimensional class (such as, for priors on (0, 1), the class of monotone non-increasing functions) with (for instance) bounds on the first two moments, You can in turn quantify bounds on summaries of the resulting posterior distribution, which may be narrow enough to demonstrate that Your

uncertainty in specifying the prior does not lead to differences that are large in practical terms.

Often context suggests specification of a prior that has relatively little information content in relation to the likelihood information; for reasons that are made clear in Sect. "Inference and Prediction: Binary Outcomes with No Covariates", such priors are referred to as relatively *diffuse* or *flat* (the term *non-informative* is sometimes also used, but this seems worth avoiding, because any prior specification takes a particular position regarding the amount of relevant information external to the data). See Bernardo [10] and Kass and Wasserman [60] for a variety of formal methods for generating diffuse prior distributions.

- **Q:** How do You quantify Your utility function $U(a, \theta)$ for optimal decision-making? **A:** There is a rather less extensive statistical literature on elicitation of utility than probability; notable references include Fishburn [35,36], Schervish et al. [103], and the citations in Bernardo ans Smith [11]. There is a parallel (and somewhat richer) economics literature on utility elicitation; see, for instance, Abdellaoui [1] and Blavatskyy [12]. Sect. "Decision-Making: Variable Selection in Generalized Linear Models; Bayesian Model Selection" provides a decision-theoretic example.
- Suppose that $\theta = (\theta_1, \ldots, \theta_k)$ is a parameter vector of length k. Then (a) computing the normalizing constant in Bayes' Theorem

$$c = \left(\int \cdots \int p(y|\theta_1, \ldots, \theta_k, \mathcal{B}) \cdot p(\theta_1, \ldots, \theta_k|\mathcal{B})\, \mathrm{d}\theta_1 \cdots \mathrm{d}\theta_k \right)^{-1} \quad (11)$$

involves evaluating a k-dimensional integral; (b) the predictive distribution in the second equation in (6) involves another k-dimensional integral; and (c) the posterior $p(\theta_1, \ldots, \theta_k|y, \mathcal{B})$ is a k-dimensional probability distribution, which for $k > 2$ can be difficult to visualize, so that attention often focuses on the *marginal* posterior distributions

$$p(\theta_j|y, \mathcal{B}) = \int \cdots \int p(\theta_1, \ldots, \theta_k|y, \mathcal{B})\, \mathrm{d}\theta_{-j} \quad (12)$$

for $j = 1, \ldots k$, where θ_{-j} is the θ vector with component j omitted; each of these marginal distributions involves a $(k-1)$-dimensional integral. If k is large these integrals can be difficult or impossible to evaluate exactly, and a general method for computing accurate approximations to them proved elusive from the time of Bayes in the eighteenth century until recently (in the late eighteenth century Laplace [63]) developed an analytical method, which today bears his name, for approximating integrals that arise in Bayesian work [11], but his method is not as general as the computationally-intensive techniques in widespread current use). Around 1990 there was a fundamental shift in Bayesian computation, with the belated discovery by the statistics profession of a class of techniques – *Markov chain Monte Carlo* (MCMC) methods [41,44] – for approximating high-dimensional Bayesian integrals in a computationally-intensive manner, which had been published in the chemical physics literature in the 1950s [78]; these methods came into focus for the Bayesian community at a moment when desktop computers had finally become fast enough to make use of such techniques.

MCMC methods approximate integrals associated with the posterior distribution $p(\theta|y, \mathcal{B})$ by (a) creating a Markov chain whose equilibrium distribution is the desired posterior and (b) sampling from this chain from an initial $\theta_{(0)}$ (i) until equilibrium has been reached (all draws up to this point are typically discarded) and (ii) for a sufficiently long period thereafter to achieve the desired approximation accuracy. With the advent and refinement of MCMC methods since 1990, the Bayesian integration problem has been solved for a wide variety of models, with more ambitious sampling schemes made possible year after year with increased computing speeds: for instance, in problems in which the dimension of the parameter space is not fixed in advance (an example is *regression change-point* problems [104], where the outcome y is assumed to depend linearly (apart from stochastic noise) on the predictor(s) x but with an unknown number of changes of slope and intercept and unknown locations for those changes), ordinary MCMC techniques will not work; in such problems methods such as *reversible-jump MCMC* [47,94] and *Markov birth-death processes* [108], which create Markov chains that permit trans-dimensional jumps, are required.

The main drawback of MCMC methods is that they do not necessarily scale well as n (the number of data observations) increases; one alternative, popular in the machine learning community, is *variational* methods [55], which convert the integration problem into an optimization problem by (a) approximating the posterior distribution of interest by a family of distributions yielding a closed-form approximation to the integral and (b) finding the member of the family that maximizes the accuracy of the approximation.

- Bayesian decision theory [6], based on maximizing expected utility, is unambiguous in its normative rec-

ommendation for how a single agent (You) should make a choice in the face of uncertainty, and it has had widespread success in fields such as economics (e. g., [2,50]) and medicine (e. g., [87,116]). It is well known, however [3,112]), that Bayesian decision theory (or indeed any other formal approach that seeks an optimal behavioral choice) can be problematic when used normatively for group decision-making, because of conflicts in preferences among members of the group. This is an important unsolved problem.

Three Examples

Inference and Prediction: Binary Outcomes with No Covariates

Consider the problem of measuring the quality of care at a particular hospital H. One way to do this is to examine the outcomes of that care, such as mortality, after adjusting for the burden of illness brought by the patients to H on admission. As an even simpler version of this problem, consider just the n binary mortality observables $y = (y_1, \dots, y_n)$ (with mortality measured within 30 days of admission, say; 1 = died, 0 = lived) that You will see from all of the patients at H with a particular admission diagnosis (heart attack, say) during some prespecified future time window. You acknowledge Your uncertainty about which elements in the sequence will be 0s and which 1s, and You wish to quantify this uncertainty using the Bayesian paradigm. As de Finetti [20] noted, in this situation Your fundamental imperative is to construct a predictive distribution $p(y_1, \dots, y_n|\mathcal{B})$ that expresses Your uncertainty about the future observables, rather than – as is perhaps more common – to reach immediately for a standard family of parametric models for the y_i (in other words, to posit the existence of a vector $\theta = (\theta_1, \dots, \theta_k)$ of parameters and to model the observables by appeal to a family $p(y_i|\theta, \mathcal{B})$ of probability distributions indexed by θ).

Even though the y_i are binary, with all but the smallest values of n it still seems a formidable task to elicit from Yourself an n-dimensional predictive distribution $p(y_1, \dots, y_n|\mathcal{B})$. De Finetti [20] showed, however, that the task is easier than it seems. In the absence of any further information about the patients, You notice that Your uncertainty about them is *exchangeable*: if someone (without telling You) were to rearrange the order in which their mortality outcomes become known to You, Your predictive distribution would not change. This still seems to leave $p(y_1, \dots, y_n|\mathcal{B})$ substantially unspecified (where $\mathcal{B}$ now includes the judgment of exchangeability of the y_i), but de Finetti [20] proved a remarkable theorem which shows (in effect) that all exchangeable predictive distributions for a vector of binary observables are representable as *mixtures* of Bernoulli sampling distributions: if You're willing to regard $(y_1, \dots, y_n)$ as the first n terms in an *infinitely exchangeable* binary sequence $(y_1, y_2, \dots)$ (which just means that every finite subsequence is exchangeable), then to achieve coherence Your predictive distribution must be expressible as

$$p(y_1, \dots, y_n|\mathcal{B}) = \int_0^1 \theta^{s_n}(1-\theta)^{n-s_n}\, p(\theta|\mathcal{B})\, \mathrm{d}\theta\,, \quad (13)$$

where $s_n = \sum_{i=1}^n y_i$. Here the quantity θ on the right side of (13) is more than just an integration variable: the equation says that in Your predictive modeling of the binary y_i You may as well proceed *as if*

- There is a quantity called θ, interpretable both as the marginal death probability $p(y_i = 1|\theta, \mathcal{B})$ for each patient and as the long-run mortality rate in the infinite sequence $(y_1, y_2, \dots)$ (which serves, in effect, as a population of values to which conclusions from the data can be generalized);
- Conditional on θ and $\mathcal{B}$, the y_i are independent identically distributed (IID) Bernoulli (θ); and
- θ can be viewed as a realization of a random variable with density $p(\theta|\mathcal{B})$.

In other words, exchangeability of Your uncertainty about a binary process is functionally equivalent to assuming the simple Bayesian *hierarchical* model [27]

$$\begin{aligned} (\theta|\mathcal{B}) &\sim p(\theta|\mathcal{B}) \\ (y_i|\theta, \mathcal{B}) &\overset{\text{IID}}{\sim} \text{Bernoulli}(\theta)\,, \end{aligned} \quad (14)$$

and $p(\theta|\mathcal{B})$ is recognizable as Your prior distribution for θ, the underlying death rate for heart attack patients similar to those You expect will arrive at hospital H during the relevant time window.

Consider now the problem of quantitatively specifying prior information about θ. From (13) and (14) the likelihood function is

$$l(\theta|y, \mathcal{B}) = c\,\theta^{s_n}(1-\theta)^{n-s_n}\,, \quad (15)$$

which (when interpreted in the Bayesian manner as a density in θ) is recognizable as a member of the Beta family of probability distributions: for $\alpha, \beta > 0$ and $0 < \theta < 1$,

$$\theta \sim \text{Beta}(\alpha, \beta) \text{ iff } p(\theta) = c\,\theta^{\alpha-1}(1-\theta)^{\beta-1}\,. \quad (16)$$

Moreover, this family has the property that the product of two Beta densities is another Beta density, so by Bayes'

Theorem if the prior $p(\theta|\mathcal{B})$ is chosen to be Beta(α, β) for some (as-yet unspecified) $\alpha > 0$ and $\beta > 0$, then the posterior will be Beta$(\alpha + s_n, \beta + n - s_n)$: this is conjugacy (Sect. "The Bayesian Statistical Paradigm") of the Beta family for the Bernoulli/binomial likelihood. In this case the conjugacy leads to a simple interpretation of α and β: the prior acts like a data set with α 1s and β 0s, in the sense that if person 1 does a Bayesian analysis with a Beta(α, β) prior and sample data $y = (y_1, \ldots, y_n)$ and person 2 instead merges the corresponding "prior data set" with y and does a maximum-likelihood analysis (Sect. "Comparison with the Frequentist Statistical Paradigm") on the resulting merged data, the two people will get the same answers. This also shows that the *prior sample size* n_0 in the Beta–Bernoulli/binomial model is $(\alpha + \beta)$. Given that the mean of a Beta(α, β) distribution is $\alpha / (\alpha + \beta)$, calculation reveals that the posterior mean $(\alpha + s_n) / (\alpha + \beta + n)$ of θ is a weighted average of the prior mean and the data mean $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$, with prior and data weights n_0 and n, respectively:

$$\frac{\alpha + s_n}{\alpha + \beta + n} = \frac{n_0 \left(\frac{\alpha}{\alpha+\beta}\right) + n\overline{y}}{n_0 + n} . \tag{17}$$

These facts shed intuitive light on how Bayes' Theorem combines information internal and external to a given data source: thinking of prior information as equivalent to a data set is a valuable intuition, even in non-conjugate settings.

The choice of α and β naturally depends on the available information external to y. Consider for illustration two such specifications:

- Analyst 1 does a web search and finds that the 30-day mortality rate for heart attack (given average quality of care and average patient sickness at admission) in her country is 15%. The information she has about hospital H is that its care and patient sickness are not likely to be wildly different from the country averages but that a mortality deviation from the mean, if present, would be more likely to occur on the high side than the low. Having lived in the community served by H for some time and having not heard anything either outstanding or deplorable about the hospital, she would be surprised to find that the underlying heart attack death rate at H was less than (say) 5% or greater than (say) 30%. One way to quantify this information is to set the prior mean to 15% and to place (say) 95% of the prior mass between 5% and 30%.
- Analyst 2 has little information external to y and thus wishes to specify a relatively diffuse prior distribution that does not dramatically favor any part of the unit interval.

Numerical integration reveals that $(\alpha, \beta) = (4.5, 25.5)$, with a prior sample size of 30.0, satisfies Analyst 1's constraints. Analyst 2's diffuse prior evidently corresponds to a rather small prior sample size; a variety of positive values of α and β near 0 are possible, all of which will lead to a relatively flat prior.

Suppose for illustration that the time period in question is about four years in length and H is a medium-size US hospital; then there will be about $n = 385$ heart attack patients in the data set y. Suppose further that the observed mortality rate at H comes out $\overline{y} = s_n / n = 69/385 \doteq 18\%$. Figure 1 summarizes the prior-to-posterior updating with this data set and the two priors for Analysts 1 (left panel) and 2 (right panel), with $\alpha = \beta = 1$ (the Uniform distribution) for Analyst 2. Even though the two priors are rather different – Analyst 1's prior is skewed, with a prior mean of 0.15 and $n_0 \doteq 30$; Analyst 2's prior is flat, with a prior mean of 0.5 and $n_0 = 2$ – it is evident that the posterior distributions are nearly the same in both cases; this is because the data sample size $n = 385$ is so much larger than either of the prior sample sizes, so that the likelihood information dominates. With both priors the likelihood and posterior distributions are nearly the same, another consequence of $n_0 \ll n$. For Analyst 1 the posterior mean, standard deviation, and 95% central posterior interval for θ are $(0.177, 0.00241, 0.142, 0.215)$, and the corresponding numerical results for Analyst 2 are $(0.181, 0.00258, 0.144, 0.221)$; again it is clear that the two sets of results are almost identical. With a large sample size, careful elicitation – like that undertaken by Analyst 1 – will often yield results similar to those with a diffuse prior.

The posterior predictive distribution $p(y_{n+1}|y_1, \ldots y_n, \mathcal{B})$ for the next observation, having observed the first n, is also straightforward to calculate in closed form with the conjugate prior in this model. It is clear that $p(y_{n+1}|y, \mathcal{B})$ has to be a Bernoulli(θ^*) distribution for some θ^*, and intuition says that θ^* should just be the mean $\alpha^* / (\alpha^* + \beta^*)$ of the posterior distribution for θ given y, in which $\alpha^* = \alpha + s_n$ and $\beta^* = \beta + n - s_n$ are the parameters of the Beta posterior. To check this, making use of the fact that the normalizing constant in the Beta(α, β) family is $\Gamma(\alpha + \beta) / \Gamma(\alpha)\, \Gamma(\beta)$, the second equation in (6) gives

$$\begin{aligned} &p(y_{n+1}|y_1, \ldots y_n, \mathcal{B}) \\ &\quad = \int_0^1 \theta^{y_{n+1}} (1-\theta)^{1-y_{n+1}} \frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^*)\, \Gamma(\beta^*)} \theta^{\alpha^*-1} \\ &\qquad \cdot (1-\theta)^{\beta^*-1}\, \mathrm{d}\theta \end{aligned}$$

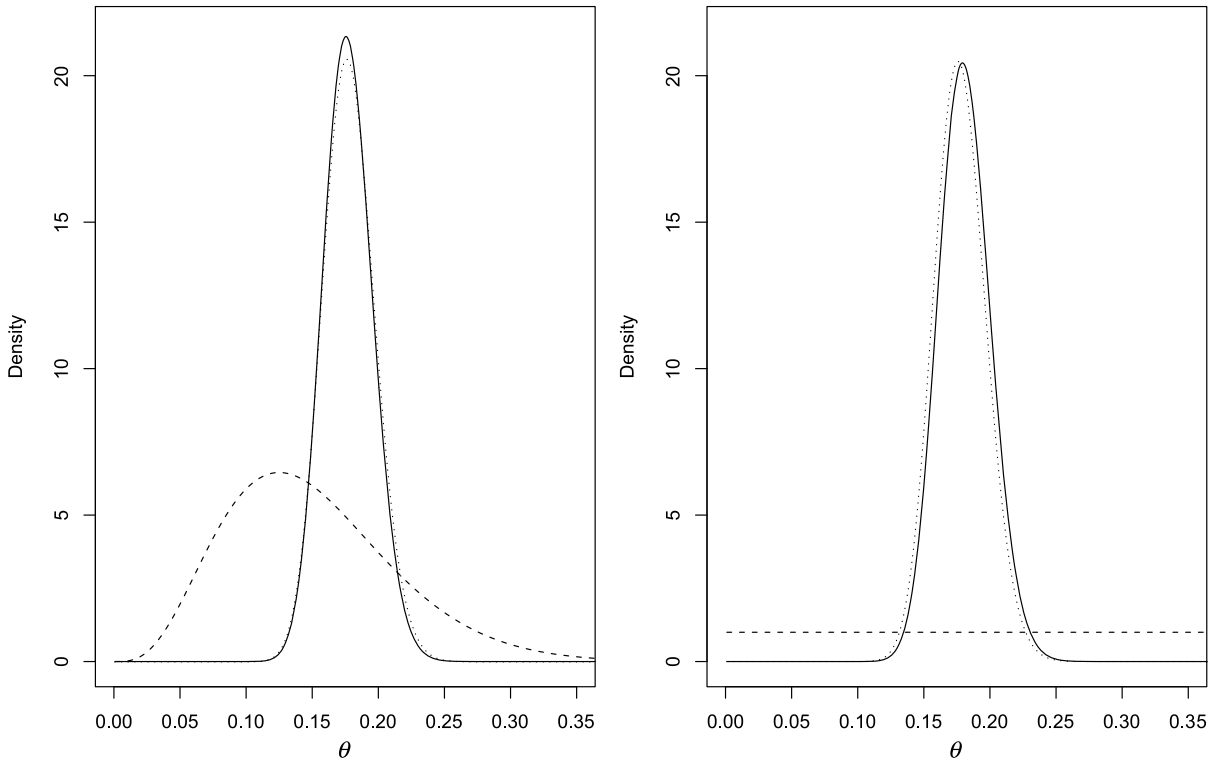

Bayesian Statistics, Figure 1
Prior-to-posterior updating with two prior specifications in the mortality data set (in both panels, prior: *long dotted lines*; likelihood: *short dotted lines*; posterior: *solid lines*). The left and right panels give the updating with the priors for Analysts 1 and 2, respectively

$$
\begin{aligned}
&= \frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^*)\,\Gamma(\beta^*)} \int_0^1 \theta^{\alpha^* + y_{n+1} - 1} \\
&\quad \cdot (1-\theta)^{(\beta^* - y_{n+1} + 1) - 1}\, \mathrm{d}\theta \\
&= \left[\frac{\Gamma(\alpha^* + y_{n+1})}{\Gamma(\alpha^*)}\right]\left[\frac{\Gamma(\beta^* - y_{n+1} + 1)}{\Gamma(\beta^*)}\right] \\
&\quad \cdot \left[\frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^* + \beta^* + 1)}\right]; \qquad (18)
\end{aligned}
$$

this, combined with the fact that $\Gamma(x+1)/\Gamma(x) = x$ for any real x, yields, for example in the case $y_{n+1} = 1$,

$$
\begin{aligned}
p(y_{n+1} = 1 | y, \mathcal{B}) &= \left[\frac{\Gamma(\alpha^* + 1)}{\Gamma(\alpha^*)}\right]\left[\frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^* + \beta^* + 1)}\right] \\
&= \frac{\alpha^*}{\alpha^* + \beta^*}, \qquad (19)
\end{aligned}
$$

confirming intuition.

Inference: Parametric and Non-Parametric Modeling of Count Data

Most elderly people in the Western world say they would prefer to spend the end of their lives at home, but many instead finish their lives in an institution (a nursing home or hospital). How can elderly people living in their communities be offered health and social services that would help to prevent institutionalization? Hendriksen et al. [51] conducted an experiment in the 1980s in Denmark to test the effectiveness of *in-home geriatric assessment* (IHGA), a form of preventive medicine in which each person's medical and social needs are assessed and acted upon individually. A total of $n = 572$ elderly people living in non-institutional settings in a number of villages were randomized, $n_C = 287$ to a control group, who received standard health care, and $n_T = 285$ to a treatment group, who received standard care plus IHGA. The number of hospitalizations during the two-year life of the study was an outcome of particular interest.

The data are presented and summarized in Table 1. Evidently IHGA lowered the mean hospitalization rate per two years (for the elderly Danish people in the study, at least) by $(0.944 - 0.768) \doteq 0.176$, which is about an 18% reduction from the control level, a clinically large difference. The question then becomes, in Bayesian inferential language: what is the posterior distribution for the treatment effect in the entire population $\mathcal{P}$ of patients judged exchangeable with those in the study?

Bayesian Statistics, Table 1
Distribution of number of hospitalizations in the IHGA study

Group	Number of Hospitalizations									Sample	
	0	1	2	3	4	5	6	7	n	Mean	Variance
Control	138	77	46	12	8	4	0	2	287	0.944	1.54
Treat-ment	147	83	37	13	3	1	1	0	285	0.768	1.02

Continuing to refer to the relevant analyst as You, with a binary outcome variable and no covariates in Sect. "Inference and Prediction: Binary Outcomes with No Covariates" the model arose naturally from a judgment of exchangeability of Your uncertainty about all n outcomes, but such a judgment of *unconditional* exchangeability would not be appropriate initially here; to make such a judgment would be to assert that the treatment and control interventions have the same effect on hospitalization, and it was the point of the study to see if this is true. Here, at least initially, it would be more scientifically appropriate to assert exchangeability separately and in parallel within the two experimental groups, a judgment de Finetti [22] called *partial exchangeability* and which has more recently been referred to as *conditional exchangeability* [28,72] given the treatment/control status covariate.

Considering for the moment just the control group outcome values C_i, $i = 1, \ldots, n_C$, and seeking as in Sect. "Inference and Prediction: Binary Outcomes with No Covariates" to model them via a predictive distribution $p(C_1, \ldots, C_{n_C}|\mathcal{B})$, de Finetti's previous representation theorem is not available because the outcomes are real-valued rather than binary, but he proved [21] another theorem for this situation as well: if You're willing to regard $(C_1, \ldots, C_{n_C})$ as the first n_C terms in an infinitely exchangeable sequence $(C_1, C_2, \ldots)$ of values on $\mathbb{R}$ (which plays the role of the population $\mathcal{P}$, under the control condition, in this problem), then to achieve coherence Your predictive distribution must be expressible as

$$p(C_1, \ldots, C_{n_C}|\mathcal{B}) = \int_{\mathcal{F}} \prod_{i=1}^{n_C} F(C_i)\, \mathrm{d}G(F|\mathcal{B})\,; \tag{20}$$

here (a) F has an interpretation as $F(t) = \lim_{n_C \to \infty} F_{n_C}(t)$, where F_{n_C} is the empirical CDF based on $(C_1, \ldots, C_{n_C})$; (b) $G(F|\mathcal{B}) = \lim_{n_C \to \infty} p(F_{n_C}|\mathcal{B})$, where $p(\cdot|\mathcal{B})$ is Your joint probability distribution on $(C_1, C_2, \ldots)$; and (c) $\mathcal{F}$ is the space of all possible CDFs on $\mathbb{R}$. Equation (20) says informally that exchangeability of Your uncertainty about an observable process unfolding on the real line is functionally equivalent to assuming the Bayesian hierarchical model

$$\begin{aligned} (F|\mathcal{B}) &\sim p(F|\mathcal{B}) \\ (y_i|F, \mathcal{B}) &\overset{\text{IID}}{\sim} F\,, \end{aligned} \tag{21}$$

where $p(F|\mathcal{B})$ is a prior distribution on $\mathcal{F}$. Placing distributions on functions, such as CDFs and regression surfaces, is the topic addressed by the field of *Bayesian nonparametric* (BNP) modeling [24,80], an area of statistics that has recently moved completely into the realm of day-to-day implementation and relevance through advances in MCMC computational methods. Two rich families of prior distributions on CDFs about which a wealth of practical experience has recently accumulated include (mixtures of) *Dirichlet processes* [32] and *Pólya trees* [66].

Parametric modeling is of course also possible with the IHGA data: as noted by Krnjajić et al. [62], who explore both parametric and BNP models for data of this kind, Poisson modeling is a natural choice, since the outcome consists of counts of relatively rare events. The first Poisson model to which one would generally turn is a *fixed-effects* model, in which $(C_i|\lambda_C)$ are IID Poisson(λ_C) $(i = 1, \ldots, n_C = 287)$ and $(T_j|\lambda_T)$ are IID Poisson(λ_T) $(j = 1, \ldots, n_T = 285)$, with a diffuse prior on (λ_C, λ_T) if little is known, external to the data set, about the underlying hospitalization rates in the control and treatment groups. However, the last two columns of Table 1 reveal that the sample variance is noticeably larger than the sample mean in both groups, indicating substantial Poisson over-dispersion. For a second, improved, parametric model this suggests a *random-effects* Poisson model of the form

$$\begin{aligned} (C_i|\lambda_{iC}) &\overset{\text{indep}}{\sim} \text{Poisson}(\lambda_{iC}) \\ \left[\log(\lambda_{iC})|\beta_{0C}, \sigma_C^2\right] &\overset{\text{IID}}{\sim} N(\beta_{0C}, \sigma_C^2)\,, \end{aligned} \tag{22}$$

and similarly for the treatment group, with diffuse priors for $(\beta_{0C}, \sigma_C^2, \beta_{0T}, \sigma_T^2)$. As Krnjajić et al. [62] note, from a medical point of view this model is more plausible than the fixed-effects formulation: each patient in the control group has his/her own *latent* (unobserved) underlying rate of hospitalization λ_{iC}, which may well differ from the underlying rates of the other control patients because of unmeasured differences in factors such as health status at the beginning of the experiment (and similarly for the treatment group).

Model (22), when complemented by its analogue in the treatment group, specifies a Lognormal mixture of Poisson distributions for each group and is straightforward to fit by MCMC, but the Gaussian assumption for the mixing distribution is conventional, not motivated by the underlying science of the problem, and if the distribution of the

latent variables is not Gaussian – for example, if it is multimodal or skewed – model (22) may well lead to incorrect inferences. Krnjajić et al. [62] therefore also examine several BNP models that are centered on the random-effects Poisson model but which permit learning about the true underlying distribution of the latent variables instead of assuming it is Gaussian. One of their models, when applied (for example) to the control group, was

$$\begin{aligned} (C_i|\lambda_{iC}) &\overset{\text{indep}}{\sim} \text{Poisson}(\lambda_{iC}) \\ \left[\log(\lambda_{iC})|G\right] &\overset{\text{IID}}{\sim} G \\ (G|\alpha,\mu,\sigma^2) &\sim \text{DP}[\alpha\, N(\mu,\sigma^2)]\,. \end{aligned} \tag{23}$$

Here $\text{DP}[\alpha\, N(\mu,\sigma^2)]$ refers to a Dirichlet process prior distribution, on the CDF G of the latent variables, which is centered at the $N(\mu,\sigma^2)$ model with *precision parameter* α. Model (23) is an *expansion* of the random-effects Poisson model (22) in that the latter is a special case of the former (obtained by letting $\alpha \to \infty$). *Model expansion* is a common Bayesian analytic tool which helps to assess and propagate model uncertainty: if You are uncertain about a particular modeling detail, instead of fitting a model that assumes this detail is correct with probability 1, embed it in a richer model class of which it is a special case, and let the data tell You about its plausibility.

With the IHGA data, models (22) and (23) turned out to arrive at similar inferential conclusions – in both cases point estimates of the ratio of the treatment mean to the control mean were about 0.82 with a posterior standard deviation of about 0.09, and a posterior probability that the (population) mean ratio was less than 1 of about 0.95, so that evidence is strong that IHGA lowers mean hospitalizations not just in the sample but in the collection $\mathcal{P}$ of elderly people to whom it is appropriate to generalize. But the two modeling approaches need not yield similar results: if the latent variable distribution is far from Gaussian, model (22) will not be able to adjust to this violation of one of its basic assumptions. Krnjajić et al. [62] performed a simulation study in which data sets with 300 observations were generated from various Gaussian and non-Gaussian latent variable distributions and a variety of parametric and BNP models were fit to the resulting count data; Fig. 2 summarizes the prior and posterior predictive distributions from models (22; top panel) and (23; bottom panel) with a bimodal latent variable distribution. The parametric Gaussian random-effects model cannot fit the bimodality on the data scale, but the BNP model – even though centered on the Gaussian as the random-effects distribution – adapts smoothly to the underlying bimodal reality.

Decision-Making: Variable Selection in Generalized Linear Models; Bayesian Model Selection

Variable selection (choosing the "best" subset of predictors) in generalized linear models is an old problem, dating back at least to the 1960s, and many methods [113] have been proposed to try to solve it; but virtually all of them ignore an aspect of the problem that can be important: the cost of data collection of the predictors. An example, studied by Fouskakis and Draper [39], which is an elaboration of the problem examined in Sect. "Inference and Prediction: Binary Outcomes with No Covariates", arises in the field of quality of health care measurement, where patient sickness at admission is often assessed by using logistic regression of an outcome, such as mortality within 30 days of admission, on a fairly large number of sickness indicators (on the order of 100) to construct a sickness scale, employing standard variable selection methods (for instance, backward selection from a model with all predictors) to find an "optimal" subset of 10–20 indicators that predict mortality well. The problem with such *benefit-only* methods is that they ignore the considerable differences among the sickness indicators in the *cost* of data collection; this issue is crucial when admission sickness is used to drive programs (now implemented or under consideration in several countries, including the US and UK) that attempt to identify substandard hospitals by comparing observed and expected mortality rates (given admission sickness), because such quality of care investigations are typically conducted under cost constraints. When both data-collection cost and accuracy of prediction of 30-day mortality are considered, a large variable-selection problem arises in which the only variables that make it into the final scale should be those that achieve a cost-benefit tradeoff.

Variable selection is an example of the broader process of *model selection*, in which questions such as "Is model M_1 better than M_2?" and "Is M_1 good enough?" arise. These inquiries cannot be addressed, however, without first answering a new set of questions: good enough (better than) for what purpose? Specifying this purpose [26,57,61,70] identifies model selection as a decision problem that should be approached by constructing a contextually relevant utility function and maximizing expected utility. Fouskakis and Draper [39] create a utility function, for variable selection in their severity of illness problem, with two components that are combined additively: a data-collection component (in monetary units, such as US$), which is simply the negative of the total amount of money required to collect data on a given set of patients with a given subset of the sickness indicators; and a predictive-accuracy component, in which a method

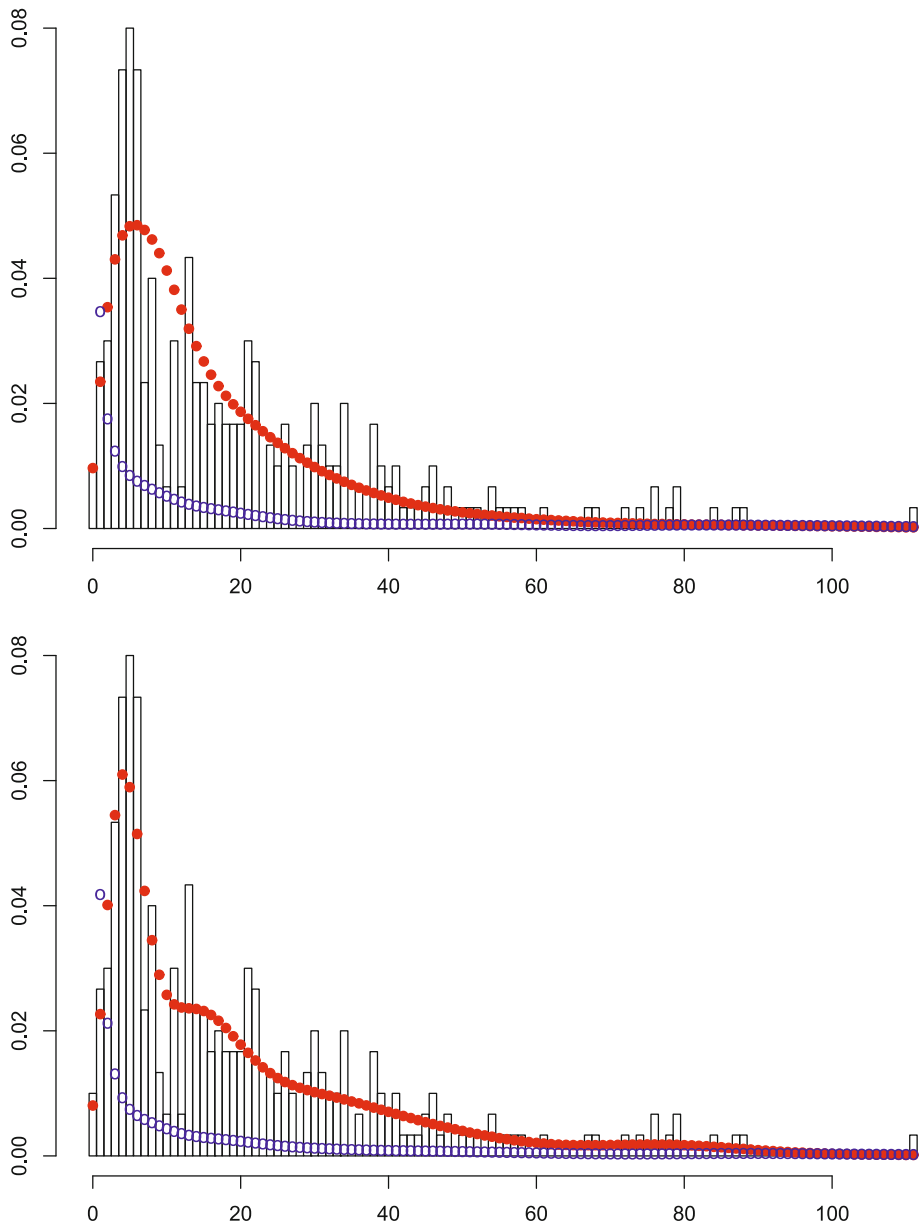

Bayesian Statistics, Figure 2

Prior (*open circles*) and posterior (*solid circles*) predictive distributions under models (22) and (23) (*top* and *bottom* panels, respectively) based on a data set generated from a bimodal latent variable distribution. In each panel, the histogram plots the simulated counts

is devised to convert increased predictive accuracy into decreased monetary cost by thinking about the consequences of labeling a hospital with bad quality of care "good" and vice versa. One aspect of their work, with a data set (from a RAND study: [58]) involving $p = 83$ sickness indicators gathered on a representative sample of $n = 2{,}532$ elderly American patients hospitalized in the period 1980–86 with pneumonia, focused only on the $p = 14$ variables in the original RAND sickness scale; this was chosen because $2^{14} = 16\,384$ was a small enough number of possible models to do brute-force enumeration of the estimated expected utility (EEU) of all the models. Figure 3 is a parallel boxplot of the EEUs of all 16 384 variable subsets, with the boxplots sorted by the number of variables in each model. The model with no predictors does poorly, with an EEU of about US$−14.5, but from a cost-benefit point of

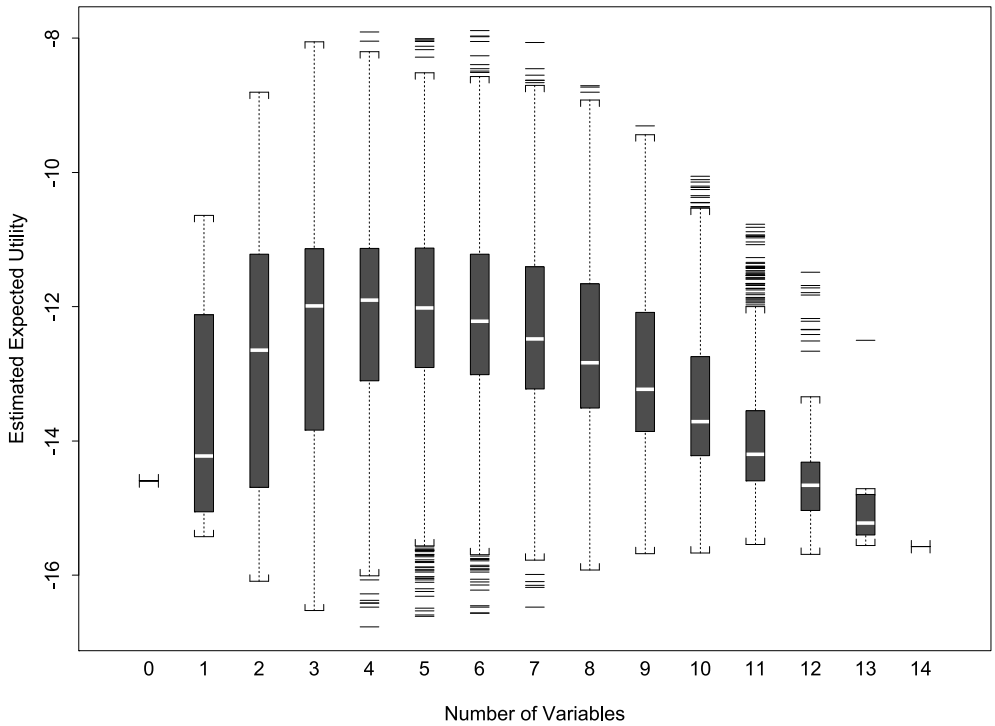

Bayesian Statistics, Figure 3
Estimated expected utility of all 16 384 variable subsets in the quality of care study based on RAND data

view the RAND sickness scale with all 14 variables is even worse (US$−15.7), because it includes expensive variables that do not add much to the predictive power in relation to cheaper variables that predict almost as well. The best subsets have 4–6 variables and would save about US$8 per patient when compared with the entire 14–variable scale; this would amount to significant savings if the observed-versus-expected assessment method were applied widely.

Returning to the general problem of Bayesian model selection, two cases can be distinguished: situations in which the precise purpose to which the model will be put can be specified (as in the variable-selection problem above), and settings in which at least some of the end uses to which the modeling will be put are not yet known. In this second situation it is still helpful to reason in a decision-theoretic way: the hallmark of a good (bad) model is that it makes good (bad) predictions, so a utility function based on predictive accuracy can be a good general-purpose choice. With (a) a single sample of data y, (b) a future data value y^*, and (c) two models M_j $(j = 1, 2)$ for illustration, what is needed is a *scoring rule* that measures the discrepancy between y^* and its predictive distribution $p(y^*|y, M_j, \mathcal{B})$ under model M_j. It turns out [46,86] that the optimal (*impartial, symmetric, proper*) scoring rules are linear functions of $\log p(y^*|y, M_j, \mathcal{B})$, which has a simple intuitive motivation: if the predictive distribution is Gaussian, for example, then values of y^* close to the center (in other words, those for which the prediction has been good) will receive a greater reward than those in the tails. An example [65], in this one-sample setting, of a model selection criterion (a) based on prediction, (b) motivated by utility considerations and (c) with good model discrimination properties [29] is the *full-sample log score*

$$LS_{FS}(M_j|y, \mathcal{B}) = \frac{1}{n}\sum_{i=1}^{n} \log p^*(y_i|y, M_j, \mathcal{B}) \,, \qquad (24)$$

which is related to the *conditional predictive ordinate* criterion [90]. Other Bayesian model selection criteria in current use include the following:

- *Bayes factors* [60]: Bayes' Theorem, written in odds form for discriminating between models M_1 and M_2, says that

$$\frac{p(M_1|y, \mathcal{B})}{p(M_2|y, \mathcal{B})} = \left[\frac{p(M_1|\mathcal{B})}{p(M_2|\mathcal{B})}\right] \cdot \left[\frac{p(y|M_1, \mathcal{B})}{p(y|M_2, \mathcal{B})}\right] ; \qquad (25)$$

here the prior odds in favor of M_1,

$$\frac{p(M_1|\mathcal{B})}{p(M_2|\mathcal{B})}\,,$$

are multiplied by the *Bayes factor*

$$\frac{p(y|M_1,\mathcal{B})}{p(y|M_2,\mathcal{B})}$$

to produce the posterior odds

$$\frac{p(M_1|y,\mathcal{B})}{p(M_2|y,\mathcal{B})}\,.$$

According to the logic of this criterion, models with high posterior probability are to be preferred, and if all the models under consideration are equally plausible a priori this reduces to preferring models with larger Bayes factors in their favor. One problem with this approach is that – in parametric models in which model M_j has parameter vector θ_j defined on parameter space Θ_j – the *integrated likelihoods* $p(y|M_j,\mathcal{B})$ appearing in the Bayes factor can be expressed as

$$\begin{aligned} p(y|M_j,\mathcal{B}) &= \int_{\Theta_j} p(y|\theta_j,M_j,\mathcal{B})\,p(\theta_j|M_j,\mathcal{B})\,\mathrm{d}\theta_j \\ &= E_{(\theta_j|M_j,\mathcal{B})}\left[p(y|\theta_j,M_j,\mathcal{B})\right]\,. \end{aligned} \tag{26}$$

In other words, the numerator and denominator ingredients in the Bayes factor are each expressible as expectations of likelihood functions with respect to the *prior* distributions on the model parameters, and if context suggests that these priors should be specified diffusely the resulting Bayes factor can be unstable as a function of precisely how the diffuseness is specified. Various attempts have been made to remedy this instability of Bayes factors (for example, {partial, intrinsic, fractional} Bayes factors, well calibrated priors, conventional priors, intrinsic priors, expected posterior priors, …; [9]); all of these methods appear to require an appeal to ad-hockery which is absent from the log score approach.

- *Deviance Information Criterion* (DIC): Given a parametric model $p(y|\theta_j,M_j,\mathcal{B})$, Spiegelhalter et al. [106] define the *deviance information criterion* (*DIC*) (by analogy with other information criteria) to be a tradeoff between (a) an estimate of the model lack of fit, as measured by the *deviance* $D(\overline{\theta}_j)$ (where $\overline{\theta}_j$ is the posterior mean of θ_j under M_j; for the purpose of DIC, the deviance of a model [75] is minus twice the logarithm of the likelihood for that model), and (b) a penalty for model complexity equal to twice the effective number of parameters p_{Dj} of the model:

$$DIC(M_j|y,\mathcal{B}) = D(\overline{\theta}_j) + 2\,\hat{p}_{Dj}\,. \tag{27}$$

When p_{Dj} is difficult to read directly from the model (for example, in complex hierarchical models, especially those with random effects), Spiegelhalter et al. motivate the following estimate, which is easy to compute from standard MCMC output:

$$\hat{p}_{Dj} = \overline{D(\theta_j)} - D(\overline{\theta}_j)\,; \tag{28}$$

in other words, $\hat{p}_{Dj}$ is the difference between the posterior mean of the deviance and the deviance evaluated at the posterior mean of the parameters.
DIC is available as an option in several MCMC packages, including `WinBUGS` [107] and `MLwiN` [93]. One difficulty with DIC is that the MCMC estimate of p_{Dj} can be poor if the marginal posteriors for one or more parameters (using the parameterization that defines the deviance) are far from Gaussian; reparameterization (onto parameter scales where the posteriors are approximately Normal) helps but can still lead to mediocre estimates of p_{Dj}.

Other notable recent references on the subject of Bayesian variable selection include Brown et al. [13], who examine multivariate regression in the context of compositional data, and George and Foster [43], who use *empirical Bayes* methods in the Gaussian linear model.

Comparison with the Frequentist Statistical Paradigm

Strengths and Weaknesses of the Two Approaches

Frequentist statistics, which has concentrated mainly on inference, proceeds by (i) thinking of the values in a data set y as like a random sample from a *population* $\mathcal{P}$ (a set to which it is hoped that conclusions based on the data can validly be generalized), (ii) specifying a summary θ of interest in $\mathcal{P}$ (such as the population mean of the outcome variable), (iii) identifying a function $\hat{\theta}$ of y that can serve as a reasonable estimate of θ, (iv) imagining repeating the random sampling from $\mathcal{P}$ to get other data sets y and therefore other values of $\hat{\theta}$, and (v) using the random behavior of $\hat{\theta}$ across these repetitions to make inferential probability statements involving θ. A leading implementation of the frequentist paradigm [37] is based on using the value $\hat{\theta}_{\mathrm{MLE}}$ that maximizes the likelihood function as the estimate of θ and obtaining a measure of uncertainty

for $\hat{\theta}_{\text{MLE}}$ from the curvature of the logarithm of the likelihood function at its maximum; this is *maximum likelihood inference*.

Each of the frequentist and Bayesian approaches to statistics has strengths and weaknesses.

- The frequentist paradigm has the advantage that repeated-sampling calculations are often more tractable than manipulations with conditional probability distributions, and it has the clear strength that it focuses attention on the scientifically important issue of *calibration*: in settings where the true data-generating process is known (e. g., in simulations of random sampling from a known population $\mathcal{P}$), how often does a particular method of statistical inference recover known truth? The frequentist approach has the disadvantage that it only applies to inherently repeatable phenomena, and therefore cannot be used to quantify uncertainty about many true-false propositions of real-world interest (for example, if You are a doctor to whom a new patient (male, say) has just come, strictly speaking You cannot talk about the frequentist probability that *this* patient is HIV positive; he either is or he is not, and his arriving at Your office is not the outcome of any repeatable process that is straightforward to identify). In practice the frequentist approach also has the weaknesses that (a) model uncertainty is more difficult to assess and propagate in this paradigm, (b) predictive uncertainty assessments are not always straightforward to create from the frequentist point of view (the *bootstrap* [31] is one possible solution) and (c) inferential calibration may not be easy to achieve when the sample size n is small.
  An example of several of these drawbacks arises in the construction of *confidence intervals* [83], in which repeated-sampling statements such as

$$P_f(\hat{\theta}_{\text{low}} < \theta < \hat{\theta}_{\text{high}}) = 0.95 \tag{29}$$

  (where P_f quantifies the frequentist variability in $\hat{\theta}_{\text{low}}$ and $\hat{\theta}_{\text{high}}$ across repeated samples from $\mathcal{P}$) are interpreted in the frequentist paradigm as suggesting that the unknown θ lies between $\hat{\theta}_{\text{low}}$ and $\hat{\theta}_{\text{high}}$ with 95% "confidence." Two difficulties with this are that (a) equation (29) looks like a probability statement about θ but is not, because in the frequentist approach θ is a fixed unknown constant that cannot be described probabilistically, and (b) with small sample sizes *nominal* 95% confidence intervals based on maximum likelihood estimation can have actual *coverage* (the percentage of time in repeated sampling that the interval includes the true θ) substantially less than 95%.
- The Bayesian approach has the following clear advantages: (a) It applies (at least in principle) to uncertainty about anything, whether associated with a repeatable process or not; (b) inference is unambiguously based on the first equation in (6), without the need to face questions such as what constitutes a "reasonable" estimate of θ (step (iii) in the frequentist inferential paradigm above); (c) prediction is straightforwardly and unambiguously based on the second equation in (6); and (d) in the problem of decision analysis a celebrated theorem of Wald [111] says informally that all good decisions can be interpreted as having been arrived at by maximizing expected utility as in the third equation of (6), so the Bayesian approach appears to be the way forward in decision problems rather broadly (but note the final challenge at the end of Sect. "The Bayesian Statistical Paradigm"). The principal disadvantage of the Bayesian approach is that coherence (internal logical consistency) by itself does not guarantee good calibration: You are free in the Bayesian paradigm to insert strong prior information in the modeling process (without violating coherence), and – if this information is seen after the fact to have been out of step with the world – Your inferences, predictions and/or decisions may also be off-target (of course, the same is true in the both the frequentist and Bayesian paradigms with regard to Your modeling of the likelihood information).

Two examples of frequentist inferences having poor calibration properties in small samples were given by Browne and Draper [14]. Their first example again concerns the measurement of quality of care, which is often studied with *cluster samples*: a random sample of J hospitals (indexed by j) and a random sample of N total patients (indexed by i) nested in the chosen hospitals is taken, and quality of care for the chosen patients and various hospital- and patient-level predictors are measured. With y_{ij} as the quality of care score for patient i in hospital j, a first step would often be to fit a *variance-components model* with random effects at both the hospital and patient levels, to assess the relative magnitudes of within- and between-hospital variability in quality of care:

$$\begin{gathered} y_{ij} = \beta_0 + u_j + e_{ij},\ \ i = 1, \ldots, n_j,\ \ j = 1, \ldots, J; \\ \sum_{j=1}^{J} n_j = N,\ \ (u_j|\sigma_u^2) \overset{\text{IID}}{\sim} N(0, \sigma_u^2),\ \ (e_{ij}|\sigma_e^2) \overset{\text{IID}}{\sim} N(0, \sigma_e^2). \end{gathered} \tag{30}$$

Browne and Draper [14] used a simulation study to show that, with a variety of maximum-likelihood-based meth-

ods for creating confidence intervals for σ_u^2, the actual coverage of nominal 95% intervals ranged from 72–94% across realistic sample sizes and true parameter values in the fields of education and medicine, versus 89–94% for Bayesian methods based on diffuse priors.

Their second example involved a re-analysis of a Guatemalan National Survey of Maternal and Child Health [89,97], with three-level data (births nested within mothers within communities), working with the random-effects logistic regression model

$$(y_{ijk} \mid p_{ijk}) \overset{\text{indep}}{\sim} \text{Bernoulli}(p_{ijk}) \quad \text{with}$$
$$\text{logit}(p_{ijk}) = \beta_0 + \beta_1 x_{1ijk} + \beta_2 x_{2jk} + \beta_3 x_{3k} + u_{jk} + v_k, \tag{31}$$

where y_{ijk} is a binary indicator of modern prenatal care or not and where $u_{jk} \sim N(0, \sigma_u^2)$ and $v_k \sim N(0, \sigma_v^2)$ were random effects at the mother and community levels (respectively). Simulating data sets with 2 449 births by 1 558 women living in 161 communities (as in the Rodríguez and Goldman study [97]), Browne and Draper [14] showed that things can be even worse for likelihood-based methods in this model, with actual coverages (at nominal 95%) as low as 0–2% for intervals for σ_u^2 and σ_v^2, whereas Bayesian methods with diffuse priors again produced actual coverages from 89–96%. The technical problem is that the marginal likelihood functions for random-effects variances are often heavily skewed, with maxima at or near 0 even when the true variance is positive; Bayesian methods, which integrate over the likelihood function rather than maximizing it, can have (much) better small-sample calibration performance as a result.

Some Historical Perspective

The earliest published formal example of an attempt to do statistical inference – to reason backwards from effects to causes – seems to have been Bayes [5], who defined conditional probability for the first time and noted that the result we now call Bayes' Theorem was a trivial consequence of the definition. From the 1760s til the 1920s, all (or almost all) statistical inference was Bayesian, using the paradigm that Fisher and others referred to as *inverse probability*; prominent Bayesians of this period included Gauss [40], Laplace [64] and Pearson [88]. This Bayesian consensus changed with the publication of Fisher [37], which laid out a user-friendly program for maximum-likelihood estimation and inference in a wide variety of problems. Fisher railed against Bayesian inference; his principal objection was that in settings where little was known about a parameter (vector) θ external to the data, a number of prior distributions could be put forward to quantify this relative ignorance. He believed passionately in the late Victorian–Edwardian goal of scientific *objectivity*, and it bothered him greatly that two analysts with somewhat different diffuse priors might obtain somewhat different posteriors. (There is a Bayesian account of objectivity: a probability is objective if many different people more or less agree on its value. An example would be the probability of drawing a red ball from an urn known to contain 20 red and 80 white balls, if a sincere attempt is made to thoroughly mix the balls without looking at them and to draw the ball in a way that does not tend to favor one ball over another.)

There are two problems with Fisher's argument, which he never addressed:

1. He would be perfectly correct to raise this objection to Bayesian analysis if investigators were often forced to do inference based solely on prior information with no data, but in practice with even modest sample sizes the posterior is relatively insensitive to the precise manner in which diffuseness is specified in the prior, because the likelihood information in such situations is relatively so much stronger than the prior information; Sect. "Inference and Prediction: Binary Outcomes with No Covariates" provides an example of this phenomenon.
2. If Fisher had looked at the entire process of inference with an engineering eye to sensitivity and stability, he would have been forced to admit that uncertainty in how to specify the likelihood function has inferential consequences that are often an order of magnitude larger than those arising from uncertainty in how to specify the prior. It is an inescapable fact that subjectivity, through assumptions and judgments (such as the form of the likelihood function), is an integral part of any statistical analysis in problems of realistic complexity.

In spite of these unrebutted flaws in Fisher's objections to Bayesian inference, two schools of frequentist inference – one based on Fisher's maximum-likelihood estimation and *significance tests* [38], the other based on the confidence intervals and *hypothesis tests* of Neyman [83] and Neyman and Pearson [84] – came to dominate statistical practice from the 1920s at least through the 1980s. One major reason for this was practical: the Bayesian paradigm is based on integrating over the posterior distribution, and accurate approximations to high-dimensional integrals were not available during the period in question. Fisher's technology, based on differentiation (to find the maximum and curvature of the logarithm of the likelihood function) rather than integration, was a much more

tractable approach for its time. Jeffreys [54], working in the field of astronomy, and Savage [102] and Lindley [69], building on de Finetti's results, advocated forcefully for the adoption of Bayesian methods, but prior to the advent of MCMC techniques (in the late 1980s) Bayesians were often in the position of saying that they knew the best way to solve statistical problems but the computations were beyond them. MCMC has removed this practical objection to the Bayesian paradigm for a wide class of problems.

The increased availability of affordable computers with decent CPU throughput in the 1980s also helped to overcome one objection raised in Sect. "Strengths and Weaknesses of the Two Approaches" against likelihood methods, that they can produce poorly-calibrated inferences with small samples, through the introduction of the bootstrap by Efron [31] in 1979. At this writing (a) both the frequentist and Bayesian paradigms are in vigorous inferential use, with the proportion of Bayesian articles in leading journals continuing an increase that began in the 1980s; (b) Bayesian MCMC analyses are often employed to produce meaningful predictive conclusions, with the use of the bootstrap increasing for frequentist predictive calibration; and (c) the Bayesian paradigm dominates decision analysis.

A Bayesian-Frequentist Fusion

During the 20th century the debate over which paradigm to use was often framed in such a way that it seemed it was necessary to choose one approach and defend it against attacks from people who had chosen the other, but there is nothing that forces an analyst to choose a single paradigm. Since both approaches have strengths and weaknesses, it seems worthwhile instead to seek a *fusion* of the two that makes best use of the strengths. Because (a) the Bayesian paradigm appears to be the most flexible way so far developed for quantifying all sources of uncertainty and (b) its main weakness is that coherence does not guarantee good calibration, a number of statisticians, including Rubin [98], Draper [26], and Little [73], have suggested a fusion in which inferences, predictions and decisions are formulated using Bayesian methods and then evaluated for their calibration properties using frequentist methods, for example by using Bayesian models to create 95% predictive intervals for observables not used in the modeling process and seeing if approximately 95% of these intervals include the actual observed values. Analysts more accustomed to the purely frequentist (likelihood) paradigm who prefer not to explicitly make use of prior distributions may still find it useful to reason in a Bayesian way, by integrating over the parameter uncertainty in their likelihood functions rather than maximizing over it, in order to enjoy the superior calibration properties that integration has been demonstrated to provide.

Future Directions

Since the mid- to late-1980s the Bayesian statistical paradigm has made significant advances in many fields of inquiry, including agriculture, archaeology, astronomy, bioinformatics, biology, economics, education, environmetrics, finance, health policy, and medicine (see Sect. "The Bayesian Statistical Paradigm" for recent citations of work in many of these disciplines). Three areas of methodological and theoretical research appear particularly promising for extending the useful scope of Bayesian work, as follows:

- *Elicitation of prior distributions and utility functions:* It is arguable that too much use is made in Bayesian analysis of diffuse prior distributions, because (a) accurate elicitation of non-diffuse priors is hard work and (b) lingering traces still remain of a desire to at least appear to achieve the unattainable Victorian–Edwardian goal of objectivity, the (false) argument being that the use of diffuse priors somehow equates to an absence of subjectivity (see, e. g., the papers by Berger [7] and Goldstein [45] and the ensuing discussion for a vigorous debate on this issue). It is also arguable that too much emphasis was placed in the 20th century on inference at the expense of decision-making, with inferential tools such as the Neyman–Pearson hypothesis testing machinery (Sect. "Some Historical Perspective") used incorrectly to make decisions for which they are not optimal; the main reason for this, as noted in Sect. "Strengths and Weaknesses of the Two Approaches" and "Some Historical Perspective", is that (a) the frequentist paradigm was dominant from the 1920s through the 1980s and (b) the high ground in decision theory is dominated by the Bayesian approach. Relevant citations of excellent recent work on elicitation of prior distributions and utility functions were given in Sect. "The Bayesian Statistical Paradigm"; it is natural to expect that there will be a greater emphasis on decision theory and non-diffuse prior modeling in the future, and elicitation in those fields of Bayesian methodology is an important area of continuing research.
- *Group decision-making:* As noted in Sect. "The Bayesian Statistical Paradigm", maximizing expected utility is an effective method for decision-making by a single agent, but when two or more agents are involved in the decision process this approach cannot be guaranteed to

yield a satisfying solution: there may be conflicts in the agents' preferences, particularly if their relationship is at least partly adversarial. With three or more possible actions, *transitivity* of preference – if You prefer action a_1 to a_2 and a_2 to a_3, then You should prefer a_1 to a_3 – is a criterion that any reasonable decision-making process should obey; informally, a well-known theorem by Arrow [3] states that even if all of the agents' utility functions obey transitivity, there is no way to combine their utility functions into a single decision-making process that is guaranteed to respect transitivity. However, Arrow's theorem is temporally static, in the sense that the agents do not share their utility functions with each other and iterate after doing so, and it assumes that all agents have the same set $\mathcal{A}$ of feasible actions. If agents A_1 and A_2 have action spaces $\mathcal{A}_1$ and $\mathcal{A}_2$ that are not identical and they share the details of their utility specification with each other, it is possible that A_1 may realize that one of the actions in $\mathcal{A}_2$ that (s)he had not considered is better than any of the actions in $\mathcal{A}_1$ or vice versa; thus a temporally dynamic solution to the problem posed by Arrow's theorem may be possible, even if A_1 and A_2 are partially adversarial. This is another important area for new research.

- *Bayesian computation:* Since the late 1980s, simulation-based computation based on Markov chain Monte Carlo (MCMC) methods has made useful Bayesian analyses possible in an increasingly broad range of application areas, and (as noted in Sect. "The Bayesian Statistical Paradigm") increases in computing speed and sophistication of MCMC algorithms have enhanced this trend significantly. However, if a regression-style data set is visualized as a matrix with n rows (one for each subject of inquiry) and k columns (one for each variable measured on the subjects), MCMC methods do not necessarily scale well in either n or k, with the result that they can be too slow to be of practical use with large data sets (e.g, at current desktop computing speeds, with n and/or k on the order of 10^5 or greater). Improving the scaling of MCMC methods, or finding a new approach to Bayesian computation that scales better, is thus a third important area for continuing study.

Bibliography

1. Abdellaoui M (2000) Parameter-free elicitation of utility and probability weighting functions. Manag Sci 46:1497–1512
2. Aleskerov F, Bouyssou D, Monjardet B (2007) Utility Maximization, Choice and Preference, 2nd edn. Springer, New York
3. Arrow KJ (1963) Social Choice and Individual Values, 2nd edn. Yale University Press, New Haven CT
4. Barlow RE, Wu AS (1981) Preposterior analysis of Bayes estimators of mean life. Biometrika 68:403–410
5. Bayes T (1764) An essay towards solving a problem in the doctrine of chances. Philos Trans Royal Soc Lond 53:370–418
6. Berger JO (1985) Statistical Decision Theory and Bayesian Analysis. Springer, New York
7. Berger JO (2006) The case for objective Bayesian analysis (with discussion). Bayesian Anal 1:385–472
8. Berger JO, Betro B, Moreno E, Pericchi LR, Ruggeri F, Salinetti G, Wasserman L (eds) (1995) Bayesian Robustness. Institute of Mathematical Statistics Lecture Notes-Monograph Series, vol 29. IMS, Hayward CA
9. Berger JO, Pericchi LR (2001) Objective Bayesian methods for model selection: introduction and comparison. In: Lahiri P (ed) Model Selection. Monograph Series, vol 38. Institute of Mathematical Statistics Lecture Notes Series, Beachwood, pp 135–207
10. Bernardo JM (1979) Reference posterior distributions for Bayesian inference (with discussion). J Royal Stat Soc, Series B 41:113–147
11. Bernardo JM, Smith AFM (1994) Bayesian Theory. Wiley, New York
12. Blavatskyy P (2006) Error propagation in the elicitation of utility and probability weighting functions. Theory Decis 60:315–334
13. Brown PJ, Vannucci M, Fearn T (1998) Multivariate Bayesian variable selection and prediction. J Royal Stat Soc, Series B 60:627–641
14. Browne WJ, Draper D (2006) A comparison of Bayesian and likelihood methods for fitting multilevel models (with discussion). Bayesian Anal 1:473–550
15. Buck C, Blackwell P (2008) Bayesian construction of radiocarbon calibration curves (with discussion). In: Case Studies in Bayesian Statistics, vol 9. Springer, New York
16. Chipman H, George EI, McCulloch RE (1998) Bayesian CART model search (with discussion). J Am Stat Assoc 93:935–960
17. Christiansen CL, Morris CN (1997) Hierarchical Poisson regression modeling. J Am Stat Assoc 92:618–632
18. Clyde M, George EI (2004) Model uncertainty. Stat Sci 19:81–94
19. Das S, Chen MH, Kim S, Warren N (2008) A Bayesian structural equations model for multilevel data with missing responses and missing covariates. Bayesian Anal 3:197–224
20. de Finetti B (1930) Funzione caratteristica di un fenomeno aleatorio. Mem R Accad Lincei 4:86–133
21. de Finetti B (1937). La prévision: ses lois logiques, ses sources subjectives. Ann Inst H Poincaré 7:1–68 (reprinted in translation as de Finetti B (1980) Foresight: its logical laws, its subjective sources. In: Kyburg HE, Smokler HE (eds) Studies in Subjective Probability. Dover, New York, pp 93–158)
22. de Finetti B (1938/1980). Sur la condition d'équivalence partielle. Actual Sci Ind 739 (reprinted in translation as de Finetti B (1980) On the condition of partial exchangeability. In: Jeffrey R (ed) Studies in Inductive Logic and Probability. University of California Press, Berkeley, pp 193–206)
23. de Finetti B (1970) Teoria delle Probabilità, vol 1 and 2. Eunaudi, Torino (reprinted in translation as de Finetti B (1974–75) Theory of probability, vol 1 and 2. Wiley, Chichester)
24. Dey D, Müller P, Sinha D (eds) (1998) Practical Nonparametric and Semiparametric Bayesian Statistics. Springer, New York

25. Draper D (1995) Assessment and propagation of model uncertainty (with discussion). J Royal Stat Soc, Series B 57:45–97
26. Draper D (1999) Model uncertainty yes, discrete model averaging maybe. Comment on: Hoeting JA, Madigan D, Raftery AE, Volinsky CT (eds) Bayesian model averaging: a tutorial. Stat Sci 14:405–409
27. Draper D (2007) Bayesian multilevel analysis and MCMC. In: de Leeuw J, Meijer E (eds) Handbook of Multilevel Analysis. Springer, New York, pp 31–94
28. Draper D, Hodges J, Mallows C, Pregibon D (1993) Exchangeability and data analysis (with discussion). J Royal Stat Soc, Series A 156:9–37
29. Draper D, Krnjajić M (2008) Bayesian model specification. Submitted
30. Duran BS, Booker JM (1988) A Bayes sensitivity analysis when using the Beta distribution as a prior. IEEE Trans Reliab 37:239–247
31. Efron B (1979) Bootstrap methods. Ann Stat 7:1–26
32. Ferguson T (1973) A Bayesian analysis of some nonparametric problems. Ann Stat 1:209–230
33. Ferreira MAR, Lee HKH (2007) Multiscale Modeling. Springer, New York
34. Fienberg SE (2006) When did Bayesian inference become "Bayesian"? Bayesian Anal 1:1–40
35. Fishburn PC (1970) Utility Theory for Decision Making. Wiley, New York
36. Fishburn PC (1981) Subjective expected utility: a review of normative theories. Theory Decis 13:139–199
37. Fisher RA (1922) On the mathematical foundations of theoretical statistics. Philos Trans Royal Soc Lond, Series A 222:309–368
38. Fisher RA (1925) Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh
39. Fouskakis D, Draper D (2008) Comparing stochastic optimization methods for variable selection in binary outcome prediction, with application to health policy. J Am Stat Assoc, forthcoming
40. Gauss CF (1809) Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium, vol 2. Perthes and Besser, Hamburg
41. Gelfand AE, Smith AFM (1990) Sampling-based approaches to calculating marginal densities. J Am Stat Assoc 85:398–409
42. Gelman A, Meng X-L (2004) Applied Bayesian Modeling and Causal Inference From Incomplete-Data Perspectives. Wiley, New York
43. George EI, Foster DP (2000) Calibration and empirical Bayes variable selection. Biometrika 87:731–747
44. Gilks WR, Richardson S, Spiegelhalter DJ (eds) (1996) Markov Chain Monte Carlo in Practice. Chapman, New York
45. Goldstein M (2006) Subjective Bayesian analysis: principles and practice (with discussion). Bayesian Anal 1:385–472
46. Good IJ (1950) Probability and the Weighing of Evidence. Charles Griffin, London
47. Green P (1995) Reversible jump Markov chain Monte carlo computation and Bayesian model determination. Biometrika 82:711–713
48. Hacking I (1984) The Emergence of Probability. University Press, Cambridge
49. Hanson TE, Kottas A, Branscum AJ (2008) Modelling stochastic order in the analysis of receiver operating characteristic data: Bayesian non-parametric approaches. J Royal Stat Soc, Series C (Applied Statistics) 57:207–226
50. Hellwig K, Speckbacher G, Weniges P (2000) Utility maximization under capital growth constraints. J Math Econ 33:1–12
51. Hendriksen C, Lund E, Stromgard E (1984) Consequences of assessment and intervention among elderly people: a three year randomized controlled trial. Br Med J 289:1522–1524
52. Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: a tutorial. Stat Sci 14:382–417
53. Jaynes ET (2003) Probability Theory: The Logic of Science. Cambridge University Press, Cambridge
54. Jeffreys H (1931) Scientific Inference. Cambridge University Press, Cambridge
55. Jordan MI, Ghahramani Z, Jaakkola TS, Saul L (1999) An introduction to variational methods for graphical models. Mach Learn 37:183–233
56. Kadane JB (ed) (1996) Bayesian Methods and Ethics in a Clinical Trial Design. Wiley, New York
57. Kadane JB, Dickey JM (1980) Bayesian decision theory and the simplification of models. In: Kmenta J, Ramsey J (eds) Evaluation of Econometric Models. Academic Press, New York
58. Kahn K, Rubenstein L, Draper D, Kosecoff J, Rogers W, Keeler E, Brook R (1990) The effects of the DRG-based Prospective Payment System on quality of care for hospitalized Medicare patients: An introduction to the series (with discussion). J Am Med Assoc 264:1953–1955
59. Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90:773–795
60. Kass RE, Wasserman L (1996) The selection of prior distributions by formal rules. J Am Stat Assoc 91:1343–1370
61. Key J, Pericchi LR, Smith AFM (1999) Bayesian model choice: what and why? (with discussion). In: Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds) Bayesian Statistics 6. Clarendon Press, Oxford, pp 343–370
62. Krnjajić M, Kottas A, Draper D (2008) Parametric and nonparametric Bayesian model specification: a case study involving models for count data. Comput Stat Data Anal 52: 2110–2128
63. Laplace PS (1774) Mémoire sur la probabilité des causes par les évenements. Mém Acad Sci Paris 6:621–656
64. Laplace PS (1812) Théorie Analytique des Probabilités. Courcier, Paris
65. Laud PW, Ibrahim JG (1995) Predictive model selection. J Royal Stat Soc, Series B 57:247–262
66. Lavine M (1992) Some aspects of Pólya tree distributions for statistical modelling. Ann Stat 20:1222–1235
67. Leamer EE (1978) Specification searches: Ad hoc inference with non-experimental data. Wiley, New York
68. Leonard T, Hsu JSJ (1999) Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers. Cambridge University Press, Cambridge
69. Lindley DV (1965) Introduction to Probability and Statistics. Cambridge University Press, Cambridge
70. Lindley DV (1968) The choice of variables in multiple regression (with discussion). J Royal Stat Soc, Series B 30:31–66
71. Lindley DV (2006) Understanding Uncertainty. Wiley, New York
72. Lindley DV, Novick MR (1981) The role of exchangeability in inference. Ann Stat 9:45–58
73. Little RJA (2006) Calibrated Bayes: A Bayes/frequentist roadmap. Am Stat 60:213–223

74. Mangel M, Munch SB (2003) Opportunities for Bayesian analysis in the search for sustainable fisheries. ISBA Bulletin 10:3–5
75. McCullagh P, Nelder JA (1989) Generalized Linear Models, 2nd edn. Chapman, New York
76. Meng XL, van Dyk DA (1997) The EM Algorithm: an old folk song sung to a fast new tune (with discussion). J Royal Stat Soc, Series B 59:511–567
77. Merl D, Prado R (2007) Detecting selection in DNA sequences: Bayesian modelling and inference (with discussion). In: Bernardo JM, Bayarri MJ, Berger JO, Dawid AP, Heckerman D, Smith AFM, West M (eds) Bayesian Statistics 8. University Press, Oxford, pp 1–22
78. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equations of state calculations by fast computing machine. J Chem Phys 21:1087–1091
79. Morris CN, Hill J (2000) The Health Insurance Experiment: design using the Finite Selection Model. In: Morton SC, Rolph JE (eds) Public Policy and Statistics: Case Studies from RAND. Springer, New York, pp 29–53
80. Müller P, Quintana F (2004) Nonparametric Bayesian data analysis. Stat Sci 19:95–110
81. Müller P, Quintana F, Rosner G (2004) Hierarchical meta-analysis over related non-parametric Bayesian models. J Royal Stat Soc, Series B 66:735–749
82. Munch SB, Kottas A, Mangel M (2005) Bayesian nonparametric analysis of stock-recruitment relationships. Can J Fish Aquat Sci 62:1808–1821
83. Neyman J (1937) Outline of a theory of statistical estimation based on the classical theory of probability. Philos Trans Royal Soc Lond A 236:333–380
84. Neyman J, Pearson ES (1928) On the use and interpretation of certain test criteria for purposes of statistical inference. Biometrika 20:175–240
85. O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T (2006) Uncertain Judgements. Eliciting Experts' Probabilities. Wiley, New York
86. O'Hagan A, Forster J (2004) Bayesian Inference, 2nd edn. In: Kendall's Advanced Theory of Statistics, vol 2B. Arnold, London
87. Parmigiani G (2002) Modeling in medical decision-making: A Bayesian approach. Wiley, New York
88. Pearson KP (1895) Mathematical contributions to the theory of evolution, II. Skew variation in homogeneous material. Proc Royal Soc Lond 57:257–260
89. Pebley AR, Goldman N (1992) Family, community, ethnic identity, and the use of formal health care services in Guatemala. Working Paper 92-12. Office of Population Research, Princeton
90. Pettit LI (1990) The conditional predictive ordinate for the Normal distribution. J Royal Stat Soc, Series B 52:175–184
91. Pérez JM, Berger JO (2002) Expected posterior prior distributions for model selection. Biometrika 89:491–512
92. Polson NG, Stroud JR, Müller P (2008) Practical filtering with sequential parameter learning. J Royal Stat Soc, Series B 70:413–428
93. Rashbash J, Steele F, Browne WJ, Prosser B (2005) A User's Guide to `MLwiN`, Version 2.0. Centre for Multilevel Modelling, University of Bristol, Bristol UK; available at www.cmm.bristol.ac.uk Accessed 15 Aug 2008
94. Richardson S, Green PJ (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). J Royal Stat Soc, Series B 59:731–792
95. Rios Insua D, Ruggeri F (eds) (2000) Robust Bayesian Analysis. Springer, New York
96. Rodríguez A, Dunston DB, Gelfand AE (2008) The nested Dirichlet process. J Am Stat Assoc, 103, forthcoming
97. Rodríguez G, Goldman N (1995) An assessment of estimation procedures for multilevel models with binary responses. J Royal Stat Soc, Series A 158:73–89
98. Rubin DB (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann Stat 12:1151–1172
99. Rubin DB (2005) Bayesian inference for causal effects. In: Rao CR, Dey DK (eds) Handbook of Statistics: Bayesian Thinking, Modeling and Computation, vol 25. Elsevier, Amsterdam, pp 1–16
100. Sabatti C, Lange K (2008) Bayesian Gaussian mixture models for high-density genotyping arrays. J Am Stat Assoc 103:89–100
101. Sansó B, Forest CE, Zantedeschi D (2008) Inferring climate system properties using a computer model (with discussion). Bayesian Anal 3:1–62
102. Savage LJ (1954) The Foundations of Statistics. Wiley, New York
103. Schervish MJ, Seidenfeld T, Kadane JB (1990) State-dependent utilities. J Am Stat Assoc 85:840–847
104. Seidou O, Asselin JJ, Ouarda TMBJ (2007) Bayesian multivariate linear regression with application to change point models in hydrometeorological variables. In: Water Resources Research 43, W08401, doi:10.1029/2005WR004835.
105. Spiegelhalter DJ, Abrams KR, Myles JP (2004) Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Wiley, New York
106. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002) Bayesian measures of model complexity and fit (with discussion). J Royal Stat Soc, Series B 64:583–640
107. Spiegelhalter DJ, Thomas A, Best NG (1999) `WinBUGS` Version 1.2 User Manual. MRC Biostatistics Unit, Cambridge
108. Stephens M (2000) Bayesian analysis of mixture models with an unknown number of components – an alternative to reversible-jump methods. Ann Stat 28:40–74
109. Stigler SM (1986) The History of Statistics: The Measurement of Uncertainty Before 1900. Harvard University Press, Cambridge
110. Stone M (1974) Cross-validation choice and assessment of statistical predictions (with discussion). J Royal Stat Soc, Series B 36:111–147
111. Wald A (1950) Statistical Decision Functions. Wiley, New York
112. Weerahandi S, Zidek JV (1981) Multi-Bayesian statistical decision theory. J Royal Stat Soc, Series A 144:85–93
113. Weisberg S (2005) Applied Linear Regression, 3rd edn. Wiley, New York
114. West M (2003) Bayesian factor regression models in the "large p, small n paradigm." Bayesian Statistics 7:723–732
115. West M, Harrison PJ (1997) Bayesian Forecasting and Dynamic Models. Springer, New York
116. Whitehead J (2006) Using Bayesian decision theory in dose-escalation studies. In: Chevret S (ed) Statistical Methods for Dose-Finding Experiments. Wiley, New York, pp 149–171

Biochemistry, Chaotic Dynamics, Noise, and Fractal Space in

Miguel Antonio Aon, Sonia Cortassa
Institute of Molecular Cardiobiology, The Johns Hopkins University, Baltimore, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Cellular Self-Organization and Nonlinear Dynamics
Fractals as Geometric, Statistic, and Dynamic Objects
Fractals in Biochemistry and Physiology
Chaos in Biochemistry and Physiology
Chaos and Colored Noise
Noise in Biochemical and Cellular Systems
Noise and Network Behavior
Future Directions
Bibliography

Glossary

Scaling a quantity scales if it changes according to a power law whose exponent is insensitive to the details of the processes involved, a feature usually referred to as universality.

Scale free refers to the geometry or dynamic of structures or processes with no single characteristic scale (space) or frequency (temporal).

Critical applies to states or phenomena characterized by their extreme sensitivity to small changes or perturbations.

Fractal is an object of geometric, statistical, or dynamical nature that obeys power laws of the form $M(L) \propto L^D$, with D as the non-integer fractal dimension. Mandelbrot introduced "fractal geometry" and defined a fractal "as a shape made of parts similar to the whole in some way".

Chaotic is a dynamic, non-periodic, behavior that describes 'turbulent' motions in time that exhibiting sensitive dependence on initial conditions.

Deterministic dynamic behavior specified by difference or differential equations.

Noise refer to random time series originating from stochastic fluctuations of events of different nature (e. g. molecular, electric, energetic as in the biological realm) that in a certain context may represent a signal.

Mitochondria sub-cellular organelles considered to be the main energy-producers of the cell as well as the source of important signaling functions, among them apoptosis or cell death.

Network the collective organization of an ensemble of objects, or groups of them, in space (structural and topological) and time (dynamics).

Definition of the Subject

The description of fractals, noise, and chaotic dynamics, is deeply associated with the concepts of *scaling* and *critical* phenomena. The recent discovery of the occurrence of scaling, criticality, and colored noise behavior in mitochondrial networks – at a fundamental level of cellular organization and dynamics as it relates to energetics, life and death – further emphasizes the role of fractals and chaos in biochemistry.

The concept of fractals introduced by Mandelbrot was initially applied to the geometric description of irregular objects but quickly spread to the field of dynamics where it joined chaos. Initially described by Lorenz in the context of meteorology, chaos extended his influence from fluid dynamics to biology and biochemistry.

Analysis of time series of biological variables with techniques inspired by fractal and chaos theories are providing a more clear understanding of the relationship between scaling in space and time exhibited by cells and organs in health and disease. These studies have also contributed to differentiate between chaos and noise, while revealing the increasing role of the power law description in networks' topology and dynamics.

The example of the cardiac mitochondrial network clearly states that the major elements of a theory of organized complexity based on the integration of Systems Biology and the Complex Systems Approach are at hand.

Introduction

In the undergoing shift between the *analytical* and *integrative* periods of biology, the importance of the concept of complex systems and systemic approaches has become dominant [1]. Cells, organisms and ecosystems are complex because they consist of a large number of mutually interacting parts. By exchanging energy, matter or information with their environment complex systems exhibit emergence, that is they self-organize their internal structure and their dynamics with novel and sometimes surprising macroscopic properties [2,3,4,5]. The appearance of emergent properties – that do not result from the existence of a central controller – is the single most distin-

guishing feature of complex systems, and self-organization the main mechanism of generating complexity [2,5,6,7,8].

The *Complex Systems Approach* involves analyzing inter-connections and relationships rather than component parts. It is nowadays pervasive, playing an increasing role in most of the scientific disciplines, including biology (cellular biochemistry, molecular biology, cellular and developmental biology). Its main foundation is the concept of self-organization based on non-equilibrium thermodynamics of nonlinear open systems, i. e. those that exchange energy and matter like cells, organisms, and ecosystems. Three modern developments, fractals, chaos, and networks' topology and dynamics, have further extended the complex systems approach.

The advent of high throughput technologies, and the feasibility to mathematically model complicated systems, have made possible the quantitative integration of information to gain insights into the nature of health and disease, and their control. This resumes the approach of *Systems Biology* as understood at present. It is timely to integrate Systems Biology to the group of important theoretical concepts as well as experimental and mathematical techniques loosely grouped here as the Complex Systems Approach (Fig. 1). This may prove fruitful and insightful for biochemistry and physiology.

Originally, the concept of fractals arose to explain the geometry of irregular objects that is the great majority either from the macroscopic or microscopic world that surrounds us [9,10]. The discovery of chaos in dynamics [11]

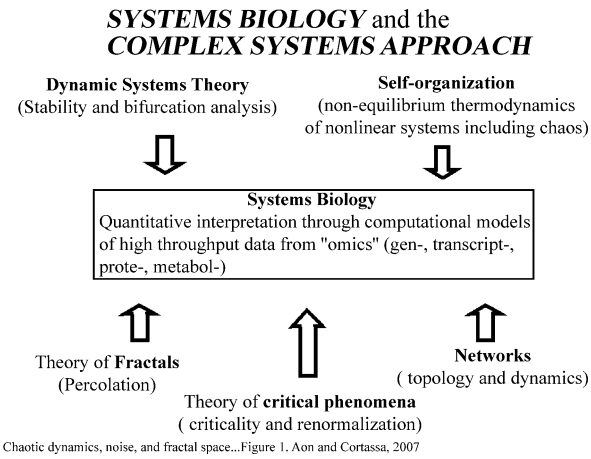

Biochemistry, Chaotic Dynamics, Noise, and Fractal Space in, Figure 1

This diagram shows the main theoretical concepts from the Complex Systems Approach that may contribute to a more integrated and realistic view of Systems Biology as presently understood

(see [12] for a historical review) and criticality in phase transitions [5,13] allowed us to clearly see that *scaling* is a common, foundational, concept of chaos and criticality as it is with fractals. Thus, scaling appears in the description of fractals, in the theory of critical phenomena and in chaotic dynamics [14].

Scaling appeals to the interaction between the multiple levels of organization exhibited by cells and organisms, thus linking the spatial and temporal aspects of their organization [15,16,17]. Scaling is a feature displayed by an unexpectedly large selection of biological systems. Scaling means that a quantity depends through a power law on its argument. Thus, if a variable value changes according to a *power law* when the parameter it depends on is growing linearly, we say it *scales*, and the corresponding exponent is called a *critical exponent*. The concept of scaling has been extended to the fractal description and characterization of the dynamic performance of systems in general (e. g. biological, economical, geophysical) exhibiting self-organized (chaotic) or (colored) noise behavior (see below). A great insight there has been to show that in time series from, e. g. market prices or mitochondrial membrane potential, the short-term fluctuations are intrinsically related to the long-term trends, appealing to the notion of networks.

Three main directions have characterized the study of cellular networks [18]: (i) architectural (structural morphology); (ii) topological (connectivity properties); and (iii) dynamical (time-dependent behavior). Network organization is at the basis of the collective and large-scale spreading consequences exhibited by complex phenomena [19,20]. Networks exhibit scale-free topologies, i. e. most of the nodes in a network will have only a few links and these will be held together by a small number of nodes exhibiting high connectivity [21]. This is a manifestation of the scaling and inverse power law features exhibited by network organization in the topological and dynamical sense. As a result, inverse power laws characterize the normal, physiological, as well as the pathophysiological behavior of the cardiac mitochondrial network [18,22,23].

Biologists have the notion of scaling from long ago, especially associated with the relationship between the size of organisms and their metabolic rates, or the so-called allometric relationship [2,24]. The concept of organisms' design in terms of fractal geometry as an underlying explanation of the link between size and metabolism has reinvigorated this research field (see [25] and Refs. therein). Why is the concept of scaling so pervading and fundamental for biology and biochemistry? Because it shows the intrinsic interdependence among the different levels of organization exhibited by living systems, which is expressed by the correlation among the different scales of space and

time involved [2,15,16,25,26,27]. Geometric and dynamic fractals capture the essence of this crucial feature through scaling and self-similarity [26].

Scaling also expresses the essential organizational traits of different systems independent of their nature. Some relevant examples are the complex set of dynamic motions or chaos detected at the onset of turbulence, as well as critical phenomena. Chaos and critical behavior display *universality* since universal numbers [28,29] or a few critical exponents [5,10] characterize them. In this context, *universality* means the insensitivity of the exponent of power laws to the details of the processes considered. Critical exponents also characterize the transition of the cardiac mitochondrial network from physiological to pathophysiological behavior [30], as another recent example described later. If one day we may presume of having a theory of complexity the concept of scaling will certainly be in its foundations.

A central insight derived from the Complex Systems Approach has been to show that networks of social or economic nature, or complex systems – such as human-engineered machines or biological ones tinkered by evolution – when stressed, can collapse, crash, or rupture. These events that happen under extreme conditions will trigger fundamental biological and societal changes and as such qualify as crises ranging from disease and epidemics to large natural catastrophes (e. g. earthquakes, hurricanes) [5,17] or stock markets crash. The cascade of failures shown to scale from the subcellular (mitochondrial network) to the organ (heart) resulting in reperfusion-related arrhythmias after ischemic injury, and eventually the death of the organism, is another thrilling example [31]. From our innermost biochemistry and physiology to the large scale of natural and human phenomena, the same laws are at play.

Cellular Self-Organization and Nonlinear Dynamics

A cell is the autonomous unit of life. Its autonomy is based on two fundamental concepts: "operational closure" [32] and self-organization [4]. Operational closure refers to the capacity of living systems for self-making of their organization (autopoietic networks) through self-organization by continuous exchange of matter and energy [32]. According to the emerging view of living systems, cells consist of self-making networks of mass, energy and information continually flowing, but the overall system maintains a stable form autonomously through self-organization, i. e. dynamic organization [1,2,7,18]. Chaos along with multiple stationary states and oscillations belong to the range of dynamic self-organized behaviors exhibited by cells.

Thermodynamically, cells and organisms behave as open systems continuously exchanging matter and energy. Under non-equilibrium steady state conditions, free energy is continuously dissipated thus the entropy production from irreversible processes inside the system, d_iS, always increases whereas the flow of entropy exchange with the surroundings, d_eS, may be negative. It is in the non-equilibrium state that self-organized structures can emerge spontaneously implying a decrease in entropy maintained by energy and matter dissipation and as such called *dissipative structures* [4]. Although living organisms continually increase their entropy (positive d_iS) and tend to approach the dangerous state of maximum entropy (or death), they stay away from it by continually drawing from its environment negative entropy (d_eS) or, more rigorously, free energy to keep its organization [33,34].

In many cases, the stationary state represents a delicate balance among the units of the system but avalanche-like series of changes may take place when it is close to a critical state. Such large events express more than anything else the underlying "forces" usually hidden under almost perfect balance providing a potential for a better understanding of complex systems [5,35]. Under those conditions, the system becomes very sensitive to perturbations. Many biological systems can be considered both, self-organized and critical.

Fractals as Geometric, Statistic, and Dynamic Objects

Fractals are of geometrical, statistical and dynamical nature [14]. Geometrically, fractals can be looked at as structures exhibiting *scaling* in *space* since their mass as a function of size, or their density as a function of distance, behave as a power law [9,35]. Additionally, the scaling is self-similar meaning that the object remains invariant at several length scales. The fractal dimension, D_f, characterizes the invariant scaling of objects exhibiting this type of geometry.

Mandelbrot (see [36] for a review), and West (1990) among others, expanded the concept of fractals from geometric objects to dynamic processes exhibiting fractal statistics. Fractal dynamics correspond to processes that are scale free, i. e. they do not exhibit a fundamental scale, and are represented by inverse power laws.

The double nature of fractals as geometric (spatial) objects and dynamic processes has allowed the use of common techniques for their characterization and quantification. For instance, lacunarity used as a measure of the non-uniformity (heterogeneity) of structures, is quantified through mass related distributions. These distribu-

tions can be characterized by the coefficient of variation or relative dispersion, RD (= standard deviation, SD/mean) that is a strong function of scale [9,26,37]. Relative dispersion as applied to the temporal domain suggests that events in different time scales are tied together through fractal statistics whose span will be as large as the scaling holds for at least two orders of magnitude. For self-similar time series, Relative Dispersional Analysis (RDA) shows that RD is constant with scale (i. e., the object looks the same at all scales) [17,26]. Since RDA implies the binning of the data set at successively larger time scales, for a system exhibiting completely random fluctuations in time, its RD drops off more rapidly as compared with a system that shows long-term memory. As a result, an inverse power law with a slope corresponding to a D_f close to 1.0 suggests high correlation among the components of a system exhibiting dynamics processes with long-term memory (see next section below).

In the temporal domain, the existence of fractals has also been investigated using Power Spectral Analysis (PSA). As in geometrically fractal objects, the dynamics of processes with long-term memory correspond to self-similar scaling in frequency. Thus, correlations in time series are also revealed using PSA; the power spectrum of such data usually follows an inverse power law proportional to $1/f^{\beta}$. This power law, more generally known as $1/f$ noise, describes colored noise depending on the value of β or the spectral exponent (e. g. $\beta = 1, 2$ or 3, for pink, brown, or black noise, respectively). The spectral exponent equals zero for white noise; thus, the spectrum of white noise is independent of frequency. If we integrate a white noise over time, we get "brown" noise that has a power spectrum that is proportional to f^{-2} over an extended frequency.

Fractals in Biochemistry and Physiology

In cellular biochemistry, the fractal approach has been directed to understanding the organization and behavior of (macro)molecules in cells [16,38,39,40] whereas in physiology it has been used as a theoretical and analytical tool of signals [26,41,42]. The former field is still in its infancy whereas the latter has already shown a relevant progress.

The recent finding of fractal organization and scaling during transitions in mitochondrial physiology [30] is significant because it links cellular biochemistry with organ (patho)physiology [31,43,44] (Fig. 2). An early proposal [2,45] suggested that in the cellular cytoplasm fractal-like kinetics could arise as a result of cytoskeletal proteins' fractal organization and anomalous diffusion [46]. Recently, this proposal has been further substantiated [16,40,47] as primary steps towards understanding the dynamics of biochemical reactions in the intracellular environment [48,49,50]. The approach to fractal kinetics in cells differs between these authors. All of them depart from Kopelman (1988) whose main observation was that in chemical reactions occurring in heterogeneous media the rate constants become time-dependent scaling with time as t^{-h}, the h parameter reflecting the characteristics of the medium, e. g. obstacle density. The dependence of the rate constant upon h has been modeled according to a fractal Zipf–Mandelbrot distribution [40], assuming Michaelis–Menten kinetics in 2D lattices using Monte Carlo simulations with time-dependent rate constant [47], or in terms of the dependence of the parameter h on the spectral dimension, D_s, for Michaelis–Menten or sigmoidal kinetics [16]. The main findings show that: (i) spatial segregation of substrate and products increase with the degree of obstruction in the medium making stronger the deviation of the rate constants at longer times and, consequently, the fractal kinetic description as compared with the classical approach [40,47]; (ii) at short times the reaction rate becomes much larger, at low substrate levels, in fractal than in Euclidean space; this behavior being explained by the time-dependence of the Michaelis–Menten constant (K_M) or an increase in cooperativity and reaction amplification in allosteric kinetics [16]. The latter behavior is of utmost importance for cells when challenged by sudden changes in environmental conditions since the quick relaxation of molecular mechanisms will allow fast and precise adaptation.

Heart physiology is one of the fields that benefited the most from applying fractal ideas. Pioneering efforts to apply fractal concepts in cardiac physiology uncovered that healthiness, as judged from electrocardiographic (EKG) data, display a broadband frequency spectrum of the QT interval [51]. The finding of an inverse power-law spectrum in the QRS complex arising from the EKG [51] suggested that heart dynamics exhibits no fundamental time scale (period) [41,52,53] unlike in cardiac disease where there is "loss of spectral reserve". The latter was suggested by spectrum simplification and the appearance of characteristic frequencies [14,26,54]. The loss of "spectral reserve" in cardiac interbeat interval spectra [54] has been determined in a number of disease processes including development of heart failure or postinfarction arrhythmias [55,56,57].

Inverse power law behavior is a manifestation of self-similar, fractal, behavior in the dynamics of a process. Fractal time series are associated with power spectra in which the higher the frequency component, the lower its power [14]. On the basis of the fractal organization of the

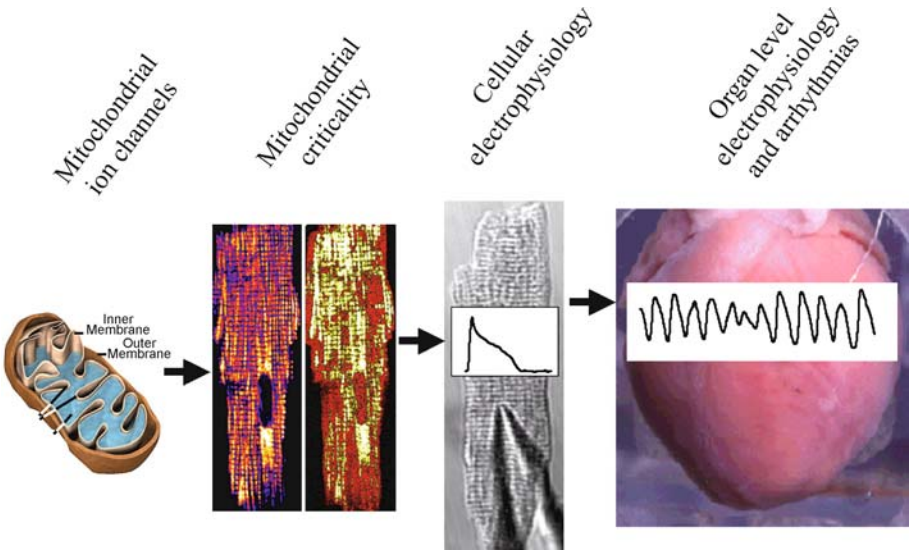

Biochemistry, Chaotic Dynamics, Noise, and Fractal Space in, Figure 2
Scaling mitochondrial dynamics to arrhythmias in the heart. Opening of inner mitochondrial membrane ion channels results in loss of membrane potential ($\Delta\Psi_m$) (*left panel*) that under criticality (a mitochondrial cluster with threshold levels of reactive oxygen species, ROS) can scale to the mitochondrial network through neighbor-neighbor interaction (*second panel*). Whole-cell electrical and mechanical dysfunction occurs due to the coupling of metabolism with energy-sensitive ion channels in the sarcolemma and the Ca^{2+}-handling subsystem (*third panel*). Dispersion of action potential repolarization contributes to impaired propagation in the myocardial syncytium, resulting in reentrant cardiac arrhythmias in the postischemic heart (*fourth panel*)

His–Purkinje conduction system, an interpretation of the inverse power law spectrum of the QRS-complex has been offered [14,51]. It was proposed that the effect of the finely branching His–Purkinje conduction network might be to subtly "decorrelate" the individual pulses that superpose to form the QRS-complex. A fractal decorrelation cascade of N voltage-time pulses results whose time trace exhibits a spectrum obeying an inverse power-law behavior. What is the mechanism of such inverse power law spectra? The inverse power law spectrum can be viewed as the resultant of many processes interacting over a myriad of interdependent scales [14]. Multiple control processes including neurohumoral regulation and local electrochemical factors regulate fluctuations in heart rate. Competition between sympathetic and parasympathetic stimulation on pacemaker cells of the sinoatrial node represent the autonomic component of the heart rate variability. A minor role of the sympathetic nervous system on the $1/f$ spectrum following administration of propanolol, a β-adrenergic blocker, has been reported [58]. The contribution of non-autonomic factors such as mitochondrial energetics to the long-range correlations in the human heartbeat time series has not been investigated.

Inverse power laws also characterize the physiological and pathophysiological behaviors of the cardiac mitochondrial network [22] (Fig. 3a). At the center stage of cell energetics, mitochondrial oxidative phosphorylation is a prime example of a biological system regulated so as to rapidly and robustly match changes in energy supply to meet cellular demand. A profound implication of the power law behavior exhibited by the mitochondrial network is that beyond a critical level, the loss of a few mitochondria may cause the entire network to fail [30] because of their multiplicative interdependency [14] (Fig. 4). In heart, the loss of mitochondrial function through collapse of $\Delta\Psi_m$ following metabolic stress directly affects the sarcolemmal K_{ATP} channel and alters the cellular action potential [AP] [59,60]. The rapid uncoupling of oxidative phosphorylation during depolarization of $\Delta\Psi_m$ is closely linked to the activation of sarcolemmal K_{ATP} currents, consequently shortening the cellular AP [60] and rendering the myocyte electrically inexcitable during the nadir of $\Delta\Psi_m$ oscillation [59]. This mechanism has been recently shown to contribute to destabilize the AP repolarization during reperfusion after ischemic injury in the whole heart, leading to arrhythmias [31] (Fig. 2). This confirms the prediction derived from the inherent organization of mitochondria as a network, i. e. that failures can scale to higher levels of organization [43,44].

In cardiac cells, the collective dynamics of the mitochondrial network was characterized from time series using RDA and PSA analysis (see previous section). The

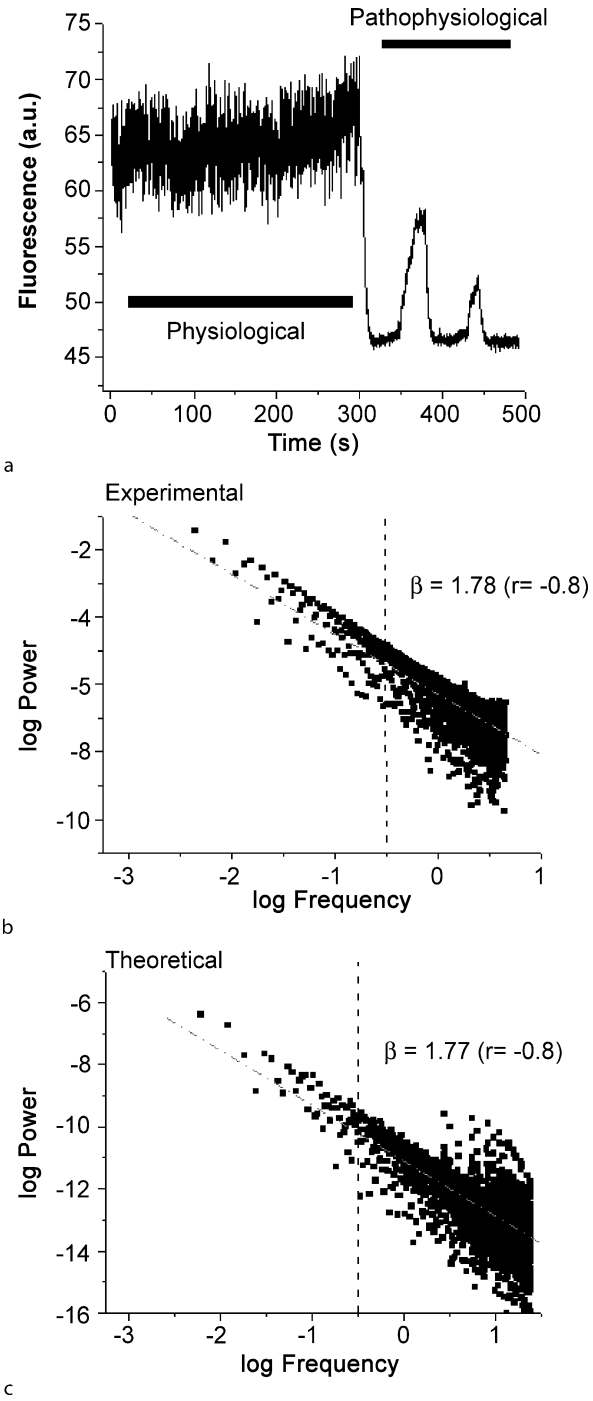

◀ **Biochemistry, Chaotic Dynamics, Noise, and Fractal Space in, Figure 3**

Physiological and pathophysiological behaviors of the mitochondrial network in heart cells. a Freshly isolated ventricular cardiomyocytes were loaded with 100 nM of the cationic potentiometric fluorescent dye TMRM, and imaged by two photon microscopy (150 ms time resolution). The results obtained from a stack of 3,720 images are shown. Before the mitochondrial network reaches criticality, the $\Delta\Psi_m$ (as measured by TMRM fluorescence) oscillates at high frequencies and small amplitudes. After criticality, the network dynamics evolve into "pathophysiological" behavior characterized by low frequency high-amplitude oscillations. b Statistical analysis of experimentally obtained $\Delta\Psi_m$ time series of the mitochondrial network of living cardiomyocytes loaded with TMRM. Power Spectral Analysis (PSA) of $\Delta\Psi_m$ time series after Fast Fourier Transform reveals a broad spectrum of oscillation in normally polarized mitochondria with a spectral exponent, $\beta = 1.78$ (a random process gives a $\beta = 0$ meaning that there is no relationship between the amplitude and the frequency in a random signal). The observed relationship between the amplitude squared (power) and frequency obeys a homogeneous power law ($1/f^{\beta}$; with f, frequency, and β, the spectral exponent), and is statistically self-similar which means that there is no dominant frequency. c After simulation of oscillations with our computational model of a mitochondrial oscillator, we selected five oscillatory periods in the high frequency domain (between 70 and 300 ms) and one from the low frequency (1 min period) domain and attributed each one of them proportionally to a network composed by 500 mitochondria [22]. A matrix containing a total of 500 columns (mitochondria) and 6,000 rows was constructed. The time steps represented by the rows correspond to a fixed integration step of 20 ms for the numerical integration of the system of ordinary differential equations. Shown are the results after applying PSA to the average value of each row of the matrix at e.g. time 1, T_1, that represents the experimental average value of fluorescence intensity of the $\Delta\Psi_m$ probe (corresponding to mV) obtained every 150 ms from 500 mitochondria (on average as in *panel* a) from each image of our stack. In *panels* b and c, the vertical dashed line at 0.3 Hz delimits the high frequency region that contains 93% of the data points (to the right of the line corresponds to the high frequency low-amplitude domain of the spectrum). Notice the close parallel between the experimental (b) and theoretical data (c)

collective dynamics of the mitochondrial membrane potential, $\Delta\Psi_m$, belongs to a statistically fractal, self-similar process characterized by a large number of frequencies in multiple timescales, rather than an inherent "characteristic" frequency [22]. The dynamic behavior of a spatially heterogeneous system can be interpreted from the correlation between nearest neighbors in terms of covariance through the spatial correlation coefficient, r_1 [62]:

$$r_1 = 2^{3-2D} - 1 \,. \tag{1}$$

The RD of aggregates of nearest neighbors obeys a fractal relationship [26]. Equation (1) can also be applied to fractal random time series, in addition to spatially inhomogeneous data sets, where D (or D_f) is the fractal dimension [17]. The derivation of the single-time-lag correlation coefficient, r_1, of a time series [17] is based on similar arguments as those used for Eq. (1). With perfect correla-

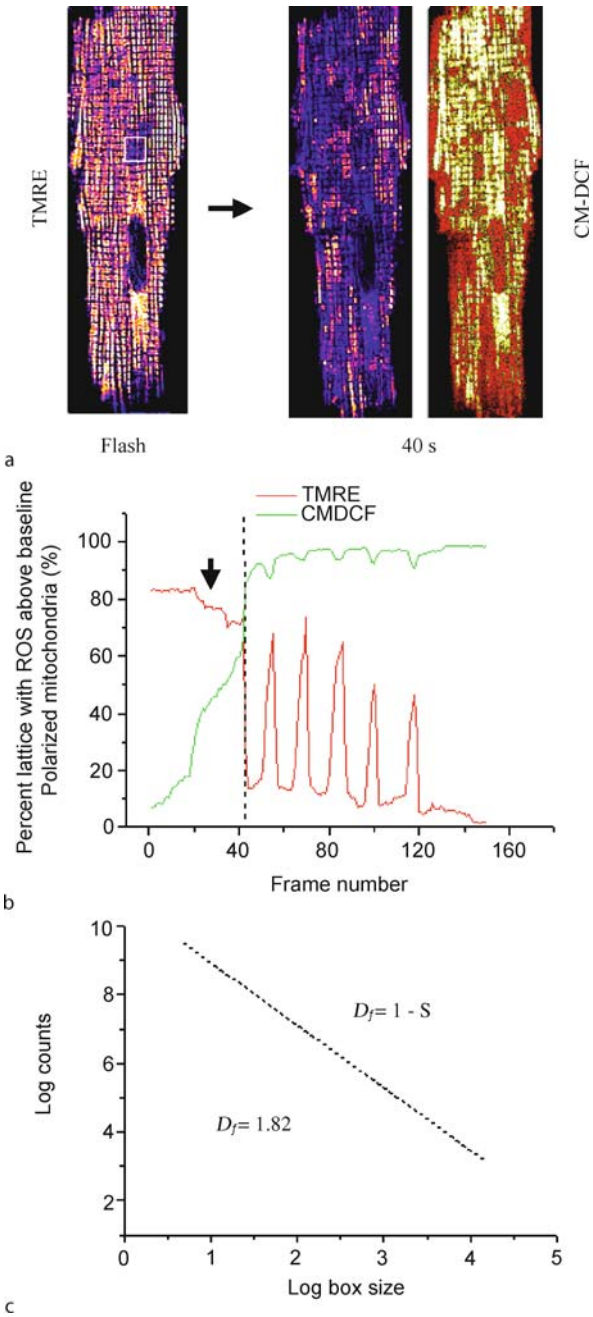

◀ **Biochemistry, Chaotic Dynamics, Noise, and Fractal Space in, Figure 4**

Mitochondrial criticality and fractal organization at the percolation threshold. a The mitochondrial lattice of cardiomyocytes, as imaged by two-photon laser scanning microscopy, was labeled with the membrane potential marker TMRE (*left* and *middle panels*; TMRE labeled cell, intensity scaled from blue to white;) and a reactive oxygen species (ROS)-sensitive fluorescent probe (*right panel*; CM-H_2DCFDA signal, scaled from red [reduced] to yellow [oxidized]). b The time course of $\Delta\Psi_m$ depolarization (the laser flash is indicated by the arrow in this panel and corresponds to the white square on the *left image of panel* a). After the laser flash a mitochondrial spanning cluster develops (within 40 s on average) as mitochondrial ROS accumulates. Criticality is reached when about 60% of the mitochondria have CM-DCF fluorescence levels ∼ 20% above baseline (*right image on panel* a), and after which sustained oscillations in $\Delta\Psi_m$ happen as depicted. This coincides with the expected theoretical percolation threshold according to standard 2D percolation theory [16,30]. c The fractal dimension, D_f, of the mitochondrial cluster, as revealed by the ROS probe (*right image on panel* a), was also consistent with percolation theory

tion, $r_1 = 1$, the mitochondrial behavior is correlated everywhere [26]. Processes without memory such as white or brown noise show completely random behavior which is characterized by an exponential law with slopes corresponding to $D_f = 1.5$ [17,26] whereas pink noise exhibits D_f close to 1.0 [22].

In the physiological regime, the power spectrum of mitochondria corresponds to a network of coupled oscillators obeying an inverse power law proportional to $1/f^{\beta}$ ($\beta = 1.7$) (Fig. 3b). This is indicative of a broad range of frequencies and amplitudes spanning at least three orders of magnitude in time (from ms to a few min). The broad frequency distribution, and the long-term temporal correlations exhibited by the mitochondrial network, allows a change in one time scale to be felt across the frequency range. This is reminiscent of heart rate regulation in a normal subject, suggesting the intriguing possibility that a change in mitochondrial network dynamics could entrain pathological consequences for the cell and the organism.

The inverse power law behavior exhibited by the mitochondrial network dynamics obeys a precise mechanism. Amplitude and oscillatory frequency exhibit an inverse relationship that, mechanistically, reflects the balance between the rate of ROS production and scavenging [22]. This relationship demonstrates two essential features of the oscillatory dynamics displayed by mitochondria: (i) the link between period and amplitude, (i. e., the longer the period, the higher the amplitude), and (ii) at periods > 300 ms, the transition to a single dominant frequency with large $\Delta\Psi_m$ depolarization that define the limit of the physiological domain.

Chaos in Biochemistry and Physiology

Homeodynamics rather than homeostasis characterize the dynamic motion of biological systems [7]. In a biological system, homeodynamics can be better visualized in the *state space*, which is n-dimensional and represents the set

of all possible states of the system with n as the number of variables. In the case of the *phase space* (i. e. a two-dimensional state space) the system's motion is given by the trajectories (whose ensemble is the phase portrait) of the system like a fluid flowing around itself.

An *attractor* is a restricted region of the phase space to which all the trajectories settle or approach over time. *Strange attractors* characterize chaotic dynamics when represented in phase space. It is well known that strange attractors can be quantitatively described as fractals, its dimension being a non-integer that is less than the dimensionality, n, of the phase space (see below). In contrast, time series of colored noise will have an integer dimension equal to the dimensionality of the phase space.

A great number of unstable limit cycles (with different periods) are embedded in a strange attractor that may be selected (at least theoretically) to a stable periodic output by small perturbations of their parameters [63,64]. Mechanistically, chaos can be controlled by a combination of periodic perturbations in a feedback control circuit with delay [65,66]. A controlled chaotic attractor can, potentially, be a multi-oscillator that can be tuned into different time domains. Evidence has been provided that in the heart, multifrequency oscillations in rhythm and amplitude create a highly unstable prefibrillatory condition that presages fibrillation [67,68].

In recent years, an ultradian clock that functions as an intracellular timekeeper coordinating biochemical activity has been proposed [3,69]. This ultradian (circahoralian) clock with ~ 40 min period has been described in the yeast *Saccharomyces cerevisiae* growing aerobically in continuous culture [70,71,72]. Respiratory oscillations are (among many) a main physiological manifestation of the ~ 40 min temperature-compensated periodicity (see [73] for a review). Genome-wide expression (transcriptome) during the time frame period provided by the ultradian clock revealed the existence of two blocks of redox superclusters [74]. A genome-wide activity of ca. 5300 transcripts expressed in two phases represented by two blocks of ~ 600 and ~ 4700 maximally expressed genes during active respiration (oxidative) and low respiration (reductive), respectively [27] (Fig. 5). Within the 40 min time frame, a 10–15 min period is conducive to DNA synthesis and replication, a time window that opens during the reductive phase of the clock cycle [74], suggesting that this might be an evolutionary strategy to avoid oxidative damage [27,75]. These gene expression clusters extend beyond central metabolism and mitochondrial energetics to other highly conserved cellular processes such as peroxisome function and repair and turnover of protein, RNA and DNA synthesis [27,75]. By separating into different

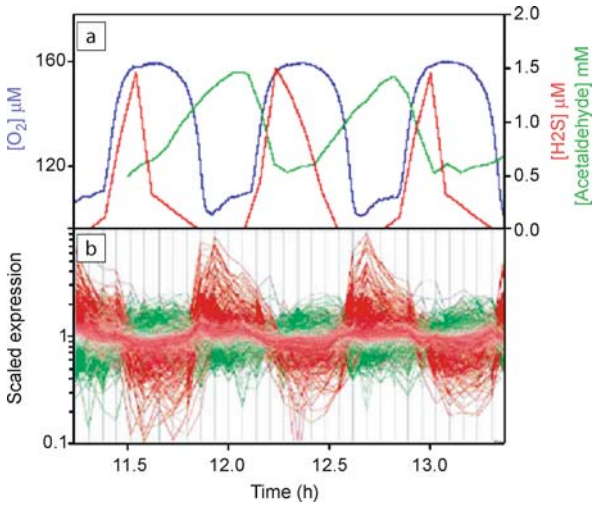

Biochemistry, Chaotic Dynamics, Noise, and Fractal Space in, Figure 5

The transcriptome of *Saccharomyces cerevisiae* (yeast) in continuous culture: activity with respect to the respiratory oscillations. a Dissolved O_2 in the culture, acetaldehyde and H_2S. b Transcripts in phase with the oxidative phase of the culture (*red*), and those in phase with the reductive phase (*green*). The scale for the transcripts was obtained by dividing the intensity of expression of that gene at that point by the median intensity for the 32 points obtained over the duration of the experiment. The genome-wide transcriptional activity during respiratory oscillations has shown that 5329 transcripts are expressed and, of these, two blocks of redox superclusters could be assigned: (i) 650 that give maximal expression during the phase of active respiration (the oxidative phase of the cycle), and (ii) 4679 that are maximally observed during the "reductive" phase of the cycle (Reproduced from [27])

temporal windows contemporary incompatible biochemical reactions, the time keeping function of the ultradian clock uncovers a powerful biological rationale. Because of the sensitivity to reactive oxygen species and oxidative stress exhibited by biochemical processes such as DNA synthesis, their occurrence manifests at a time when oxygen uptake is minimal, and antioxidant species (e. g. reduced glutathione) are highest [76].

Recent contributions show that mitochondria oscillate as coupled oscillators either during respiratory oscillations in yeast [77] or in cardiac mitochondria [22]. It is now clear that a yeast population functions like a multicellular system, a network of coupled oscillators, through chemical communication by metabolites like acetaldehyde [78] or hydrogen sulfide, H_2S [79]. The very recent experimental finding of a chaotic attractor of metabolic origin in continuous cultures of yeast could, potentially, give support to the idea that "controlled chaos" may be at the origin of ultradian (even circadian) rhythms [73,80]. Several periods

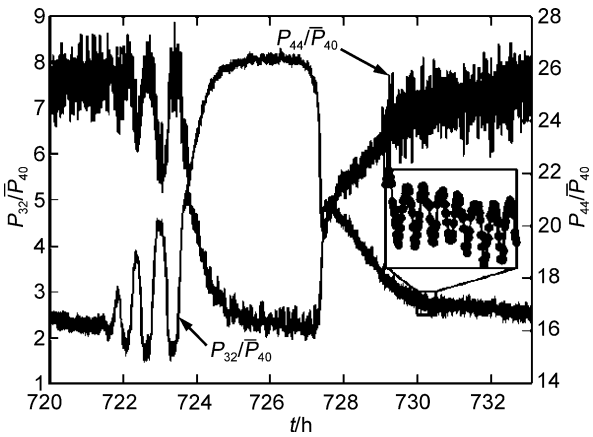

Biochemistry, Chaotic Dynamics, Noise, and Fractal Space in, Figure 6

Self-organized multioscillatory behavior of yeast respiration in continuous culture. One oscillatory period of the cell cycle-dependent, circahoralian, and fast oscillations exhibited by yeast in chemostat cultures corresponding to the relative membrane-inlet mass spectrometry signals of the m/z 32 and 44 components (i. e. relative O_2 and CO_2 signals, respectively) vs. time. The *inset* shows the individual data points for $m/z > 1/4$ 32 (O_2) for a 30-min span (Reproduced from [80])

of 13 h, 40 min and 4 min are embedded in the strange attractor of yeast metabolism and likely more may be found with greater time resolution in data collection (Fig. 6).

Chaos and Colored Noise

A number of investigators have shown that while it is true that a low-dimensional chaotic attractor will generate a fractal time series, so too will processes other than white noise exhibiting power-law spectra like colored noise [81]; see also [14]. These colored noises are clearly not necessarily related to any deterministic dynamical process. However, it has been demonstrated that colored noise may emerge from deterministic (oscillatory) dynamics in a mitochondrial network (Fig. 3c and see below) [22].

Random noises exhibit a well defined, monotonically decreasing function of the correlation dimension, ν, and the spectral exponent α, $\nu(\alpha)$ (see Fig. 3b and 3c where $\beta = \alpha$). The correlation integral $C(r)$, that for a chaotic or strange attractor is based on the correlation method proposed by Grassberger and Procaccia (1983) (see also [14]), has the power law form:

$$C(r) \sim r^{\nu} \tag{2}$$

where ν, the "correlation dimension", is closely related to the fractal dimension, D, and the "information dimension" of the attractor [82]. The D of a stochastic trajectory generated by a colored noise process can be related to the slope of the inverse power law spectrum, α, through the following equation:

$$D = 2/(\alpha - 1)\,. \tag{3}$$

Osborne and Provenzale (1988) found that a finite value for the correlation dimension ν may be found even for non-deterministic, random signals. For $\alpha = 1$ (corresponds to a white noise spectrum) they found that scaling is present, and by repeating this analysis for a number of values of α they found a quantitative relation between ν and α. Additionally, systems dominated by some kind of stochastic process are thought to provide a non-saturating value of ν.

These results have implications for experimental studies of deterministic chaos. Dynamic processes characterized by a finite and small D correspond to low dimensional strange attractors whereas a stochastic or random time series is believed to completely fill the available 3D space [26].

Noise in Biochemical and Cellular Systems

The intrinsic autonomous dynamics of biochemical processes may generate noise through stochastic fluctuations of molecular events involved in cellular physiology and biochemistry [83]. These dynamics generate time series that are characterized as colored noise when exhibiting an inverse power law spectrum. For this class of processes there is a one-to-one correspondence between the power-law dependence of the spectrum and fractal behavior, i. e. the power law index is related to the fractal dimension [17,81]. Moreover, these fractal processes are scale free, i. e. they do not exhibit a characteristic time scale, and, unlike white or brown noise, they exhibit long-term memory. The latter is a property of the autonomous system's dynamics as opposed to the so-called "extrinsic noise" or short-term random fluctuations produced by the interaction of the system with the environment [17]. Although in principle, colored noises are not related to any deterministic dynamic process as in the case of "deterministic chaos" (see above), noise-like behavior may arise from deterministic oscillatory dynamics.

Despite the stochastic function of regulatory and signaling circuits within cells, most cellular events show a precise spatio-temporal regulation. What is the role of noise, then? Several cellular processes are either noise-driven or are influenced by noise such as ion channel gating [38] and gene expression among others [83]. However, intracellular signaling networks have the ability to filter noise that in the case of periodic signals may function as *frequency filters*. A filter that passes only low frequencies is

a *low-pass* filter whereas one that passes only high frequencies is a *high-pass* filter [84]. In terms of signal processing, cascades and relays such as two-component systems and the mitogen-activated protein kinase pathway function as low-pass filters as they transduce low-frequency signals whereas high frequencies are attenuated [83]. A chemical low-pass filter can at low frequencies amplify a signal [84]. Considering nonlinear biochemical reactions, band-pass filtering is also possible. This is based on *integral feedback*, a form of negative feedback that uses an internal memory state to amplify intermediate frequencies. Bacterial chemotaxis uses integral feedback that by measuring temporal changes in chemical concentrations rather than steady state values, results in biased motion towards attractants [83].

In a cascade-type signaling pathway there seems to be a trade off between noise attenuation (arising from the input signal) and inherent noise at each step of the pathway (by molecular fluctuations). As a consequence, from the point of view of noise attenuation, there seems to be an optimal cascade length [85]. The biological ubiquitous negative feedback is the most common noise-attenuating device that in terms of signal processing functions as a low-pass filter. This has been experimentally demonstrated using the constitutive expression of GFP that is highly variable in *E. coli*. The addition of a negative feedback using the tetracycline repressor significantly reduced the measured variability [86]. Other noise-attenuating mechanisms are gene dosage, parallel cascades and regulatory checkpoints that either increase the likelihood of gene expression, establish the consensus from multiple signals, or ensure that each step in a pathway is successfully completed before proceeding to the next step.

There are mechanisms that amplify or exploit noise as well. Among these we count mechanisms that give rise to population heterogeneity or mechanisms that use noise to attenuate noise. Autocatalysis or positive feedback amplify the effects of noise giving rise to heterogeneity in a cell population as has been demonstrated in yeast using the tetracycline transactivator and a GFP marker [61]. Noise may actually enhance a signal when certain nonlinear effects are present as in stochastic resonance [87]. Concerning the latter, several biological examples have been reported [83].

Biochemical switches may be differentially sensitive to noise. Switches are based on ultrasensitive or bistable mechanisms depending upon their dynamic behavior based on a series of continuous or discontinuous steady values as a function of a control parameter [48]. Bistable mechanisms that may function in futile cycles of phosphorylation-dephosphorylation cascades, exhibit hysteresis or "memory" effects since depending on the direction of change in the control parameter (i. e. increasing or decreasing) two steady state values are possible for the same value of the control parameter [2]. Ultrasensitivity is a type of signal amplification by which the percentage change in the response of a biological system increases several-fold in a narrow variation range of the stimulus [48,49,88,89]. One main advantage of ultrasensitivity is to allow increases in flux by several-fold after fluctuations in substrate or effector concentration slightly change over the background. Bistable mechanisms are more robust to noisy signals as compared with ultrasensitive ones that in turn may exploit noise such as in slightly changing illumination conditions in cyanobacteria. These relatively small fluctuations over the background allow the triggering of the mechanism of light harvesting and polysaccharide storage even under poor stimulus conditions and quickly deactivating it hindering ATP wastage [50].

Several modeling approaches are available to address the problem of how cells control, exploit and tolerate noise. Among them are the Langevin equation and the continuous or discrete descriptions by the Fokker–Planck or master equations, respectively [83].

Noise and Network Behavior

Complex networks involve multiple feedback loops and the coexistence of mechanisms of noise attenuation, amplification and exploitation. A network keeping a reliable functionality is robust if tolerating a wide range of parametric variation and inherent noise due to stochastic molecular fluctuations. A complex network such as the chemotaxis pathway in *E. coli* is robust since it keeps its functionality despite a wide range of variation in enzymatic activities and protein concentrations [19,90,91]. The intrinsic stochasticity of biochemical reactions can be bypassed then bestowing noise resistance when perturbations affect variables that are not directly responsible for the dynamics or when molecular species are present in low numbers for shorts periods of time, as shown in a simple model of circadian rhythms [92].

Timing exerted by oscillatory mechanisms are found throughout the biological world and their periods span a wide range from milliseconds, as in the action potential of neurons and the heart, to the slow evolutionary changes that require thousands of generations [7,93]. Rhythms with temperature compensation are biological clocks. In virtually all light-sensitive organisms from cyanobacteria to humans, biological clocks adapt cyclic physiology to geophysical time with time-keeping properties in the circadian (24 h), ultradian (< 24 h) and infradian (> 24 h)

domains [94]. Within this framework, it becomes clear why it is important to understand the synchronization of a population of coupled oscillators as an important problem for the dynamics of physiology in living systems (see Sect. "Chaos in Biochemistry and Physiology"). Through frequency and amplitude modulation, oscillatory dynamics may function as a temporal-encoding signaling mechanism. Recent experimental findings are in agreement with the hypothesis that mitochondrial oscillatory behavior may function as a frequency- and/or amplitude-encoded signaling mechanism under physiological conditions and over a wide range of time scales [22,95]. This indicates that each mitochondrion affects the temporal behavior of its neighbors such that the collective ensemble of mitochondria in a cardiac cell acts as a network. The colored noise dynamics exhibited by mitochondria could be simulated from deterministic (oscillatory) dynamics by mixing a certain number of frequencies (Fig. 3c). This colored noise does not necessarily belong to chaotic dynamics.

Future Directions

Is it possible to have a theory of organized complexity in biology? The apparent problem in formulating such a theory is that *contingency* and *specificity* are attributes of biological systems. These two features emphasize the importance of detailed observations in apparent contradiction with the theoretical aim of condensing specific system details into a small number of mathematical equations, similar to the fundamental laws of physics. What makes the situation different for biology as compared to physics is that the details are important because the outcome (sometimes, life or death) is important. Thus, we are aimed at finding universal behavior taking into account most of the biologically relevant details [35].

The existence of self-organization, scaling, criticality, percolation, fractals in the cardiac mitochondrial network show that there exists a clear crossroad between the universality of physical concepts and a defined biological system such as the heart [22,30]. The cardiac mitochondrial network exhibits the transition from physiological to pathophysiological behavior with all the characteristics of systems at critical state, i. e. a mitochondrial percolation cluster attains criticality at a certain threshold level of ROS. The transition that happens at the percolation threshold is self-organized and occurs with all the traits of universality, i. e. similar critical exponents as predicted by percolation theory, and the mitochondrial cluster exhibiting fractal organization. Moreover, the time series of mitochondrial membrane potential (a crucial cellular variable) exhibit fractal organization and long-term memory. Inverse power laws statistically describe this behavior, and, in this case, the mechanisms underlying its origin are known.

Another major breakthrough comes from respiratory oscillations in yeast and their relationship to an ultradian clock. The genome-wide activity of yeast during respiratory oscillations [27,74] has now been linked to a chaotic attractor embedding multiple oscillatory periods [80].

The multiple time scales exhibited by the heart and yeast systems suggest that intracellular network dynamic organization manifests itself as scale-free in the form of dynamic fractals. Scale-free architecture has been shown in network connectivity but not in dynamics. The scale free behavior exhibited by mitochondrial network dynamics would allow simultaneous modulation of intracellular timekeeping in several time scales [18]. Two main features of mitochondria as metabolic "hubs" and as producers of ROS as signaling molecules with scale-free dynamics are in agreement with: (i) the cardiac mitochondrial network having both properties of constancy and flexibility, i. e. providing a steady supply of ATP to fuel contraction, and adapting the rate of energy production to meet changing metabolic demand [96,97]; and (ii) achieving the right ROS balance both in amount and dynamics, compatible with intracellular signaling.

We witness exciting times since several fundamental discoveries done in unrelated areas of research start to reveal that scaling and critical states are fundamental in normal physiology and pathology at the (sub)cellular, tissue and organ levels. Under steady conditions, biochemical and physiological functions exhibit a delicate balance that under stress unleashes as crises. When stressed, systems of disparate nature (e. g. markets, shuttles, cells) become extremely sensitive to small changes or perturbations exhibiting an exceptional large susceptibility to external factors and strong correlation between different parts of the system. Overall, the collective, repetitive actions of interactive nonlinear influences on many scales lead to large-scale correlations, determining the astonishing autonomy and interdependence of the living.

Bibliography

Primary Literature

1. Aon MA, Cortassa S (2006) Metabolic dynamics in cells viewed as multilayered, distributed, mass-energy-information networks. In: Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. Wiley Interscience, New York
2. Aon MA, Cortassa S (1997) Dynamic biological organization. Fundamentals as applied to cellular systems. Chapman & Hall, London
3. Lloyd D (1992) Intracellular time keeping: epigenetic oscillations reveal the functions of an ultradian clock. In: Lloyd D,

Rossi EL (eds) Ultradian rhythms in life processes. Springer, London, pp 5–22
4. Nicolis G, Prigogine I (1977) Self-organization in nonequilibrium systems: from dissipative structures to order through fluctuations. Wiley, New York
5. Sornette D (2000) Critical Phenomena in Natural Sciences. Chaos, fractals, selforganization and disorder: Concepts and tools. Springer, Berlin
6. Boccara N (2004) Modeling Complex Systems. Springer, New York
7. Lloyd D, Aon MA, Cortassa S (2001) Why homeodynamics, not homeostasis? Sci World J 1:133–145
8. Mainzer K (2005) Symmetry and Complexity. The spirit and beauty of nonlinear science. World Scientific, Singapore
9. Mandelbrot BB (1983) The fractal geometry of nature. WH Freeman, New York
10. Schroeder M (1991) Fractals, chaos, power laws. Minutes from an infinite paradise. WH Freeman, New York
11. Lorenz EN (1963) Deterministic nonperiodic flow. J Atmos Sci 20:130–141
12. Gleick J (1988) Chaos: making a new science. Penguin Books, New York
13. Wilson KG (1979) Problems in physics with many scales of length. Sci Am 241:158–179
14. West BJ (1990) Fractal Physiology and Chaos in Medicine. World Scientific, Singapore
15. Aon MA, Cortassa S (1993) An allometric interpretation of the spatio-temporal organization of molecular and cellular processes. Mol Cell Biochem 120:1–13
16. Aon MA, O'Rourke B, Cortassa S (2004) The fractal architecture of cytoplasmic organization: scaling, kinetics and emergence in metabolic networks. Mol Cell Biochem:256–257:169–184
17. West BJ (1999) Physiology, promiscuity and prophecy at The Millennium: A tale of tails. World Scientific, Singapore
18. Aon MA, Cortassa S, O'Rourke B (2007) On the network properties of mitochondria. In: Saks V (ed) Molecular system bioenergetics: energy for life. Wiley, Weinheim, pp 111–135
19. Alon U (2007) An introduction to systems biology. Design principles of biological circuits. Chapman & Hall/CRC, London
20. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5:101–113
21. Wagner A, Fell DA (2001) The small world inside large metabolic networks. Proc Biol Sci 268:1803–1810
22. Aon MA, Cortassa S, O'Rourke B (2006) The fundamental organization of cardiac mitochondria as a network of coupled oscillators. Biophys J 91:4317–4327
23. Aon MA, Cortassa S, O'Rourke B (2008) Mitochondrial oscillations in physiology and pathophysiology. Adv Exp Med Biol 641:98–117
24. Enquist BJ, West GB, Brown JH (2000) Quarter-power allometric scaling in vascular plants: Functional basis and ecological consequences. Oxford University Press, New York
25. Brown JH, West GB, Enquist BJ (2000) Scaling in biology. Oxford University Press, New York
26. Bassingthwaighte JB, Liebovitch LS, West BJ (1994) Fractal Physiology. Oxford University Press for the American Physiological Society, New York
27. Lloyd D, Murray DB (2005) Ultradian metronome: timekeeper for orchestration of cellular coherence. Trends Biochem Sci 30:373–377
28. Cvitanovic P (1989) Universality in chaos. Adam Hilger, Bristol
29. Feigenbaum MJ (1980) Universal behavior in nonlinear systems. Los Alamos Sci 1:4–27
30. Aon MA, Cortassa S, O'Rourke B (2004) Percolation and criticality in a mitochondrial network. Proc Natl Acad Sci USA 101:4447–4452
31. Akar FG, Aon MA, Tomaselli GF, O'Rourke B (2005) The mitochondrial origin of postischemic arrhythmias. J Clin Invest 115:3527–3535
32. Varela F, Maturana H, Uribe R (1974) Autopoiesis: the organization of living sytems, its characterization and a model. Biosystems 5:187–196
33. Schneider ED, Sagan D (2005) Into the cool. Energy flow, thermodynamics and life. The University of Chicago Press, Chicago
34. Schrodinger E (1944) What is life? The physical aspects of the living cell. Cambridge University Press, Cambridge
35. Vicsek T (2001) Fluctuations and Scaling in Biology. Oxford University Press, New York
36. Mandelbrot BB, Hudson RL (2004) The (mis)behavior of markest. A fractal view of risk, ruin, and reward. Basic Books, New York
37. Smith TGJ, Lange GD (1996) Fractal studies of neuronal and glial cellular morphology. CRC Press, Boca Raton
38. Liebovitch LS, Todorov AT (1996) Using fractals and nonlinear dynamics to determine the physical properties of ion channel proteins. Crit Rev Neurobiol 10:169–187
39. Savageau MA (1995) Michaelis-Menten mechanism reconsidered: implications of fractal kinetics. J Theor Biol 176:115–124
40. Schnell S, Turner TE (2004) Reaction kinetics in intracellular environments with macromolecular crowding: simulations and rate laws. Prog Biophys Mol Biol 85:235–260
41. Goldberger AL, Amaral LA, Hausdorff JM, Ivanov P, Peng CK, Stanley HE (2002) Fractal dynamics in physiology: alterations with disease and aging. Proc Natl Acad Sci USA 99 Suppl 1:2466–2472
42. Ivanov PC, Amaral LA, Goldberger AL, Havlin S, Rosenblum MG, Struzik ZR, Stanley HE (1999) Multifractality in human heartbeat dynamics. Nature 399:461–465
43. Aon MA, Cortassa S, Akar FG, O'Rourke B (2006) Mitochondrial criticality: a new concept at the turning point of life or death. Biochim Biophys Acta 1762:232–240
44. O'Rourke B, Cortassa S, Aon MA (2005) Mitochondrial ion channels: gatekeepers of life and death. Physiol (Bethesda) 20:303–315
45. Rabouille C, Cortassa S, Aon MA (1992) Fractal organisation in biological macromolecular lattices. J Biomol Struct Dyn 9:1013–1024
46. Kopelman R (1988) Fractal reaction kinetics. Science 241:1620–1626
47. Berry H (2002) Monte carlo simulations of enzyme reactions in two dimensions: fractal kinetics and spatial segregation. Biophys J 83:1891–1901
48. Aon MA, Gomez-Casati DF, Iglesias AA, Cortassa S (2001) Ultrasensitivity in (supra)molecularly organized and crowded environments. Cell Biol Int 25:1091–1099
49. Gomez-Casati DF, Aon MA, Iglesias AA (1999) Ultrasensitive glycogen synthesis in Cyanobacteria. FEBS Lett 446:117–121
50. Gomez-Casati DF, Cortassa S, Aon MA, Iglesias AA (2003) Ultrasensitive behavior in the synthesis of storage polysaccharides in cyanobacteria. Planta 216:969–975
51. Goldberger AL, West BJ (1987) Applications of nonlinear dynamics to clinical cardiology. Ann N Y Acad Sci 504:195–213

52. Ivanov PC, Nunes Amaral LA, Goldberger AL, Havlin S, Rosenblum MG, Stanley HE, Struzik ZR (2001) From 1/f noise to multifractal cascades in heartbeat dynamics. Chaos 11:641–652
53. Peng CK, Mietus J, Hausdorff JM, Havlin S, Stanley HE, Goldberger AL (1993) Long-range anticorrelations and non-gaussian behavior of the heartbeat. Phys Rev Lett 70:1343–1346
54. Goldberger AL, Bhargava V, West BJ, Mandell AJ (1985) On a mechanism of cardiac electrical stability. The fractal hypothesis. Biophys J 48:525–528
55. Casolo G, Balli E, Taddei T, Amuhasi J, Gori C (1989) Decreased spontaneous heart rate variability in congestive heart failure. Am J Cardiol 64:1162–1167
56. Ewing DJ (1991) Heart rate variability: an important new risk factor in patients following myocardial infarction. Clin Cardiol 14:683–685
57. Skinner JE, Pratt CM, Vybiral T (1993) A reduction in the correlation dimension of heartbeat intervals precedes imminent ventricular fibrillation in human subjects. Am Heart J 125:731–743
58. Yamamoto Y, Hughson RL (1994) On the fractal nature of heart rate variability in humans: effects of data length and beta-adrenergic blockade. Am J Physiol 266:R40–49
59. Aon MA, Cortassa S, Marban E, O'Rourke B (2003) Synchronized whole cell oscillations in mitochondrial metabolism triggered by a local release of reactive oxygen species in cardiac myocytes. J Biol Chem 278:44735–44744
60. O'Rourke B, Ramza BM, Marban E (1994) Oscillations of membrane current and excitability driven by metabolic oscillations in heart cells. Science 265:962–966
61. Becskei A, Seraphin B, Serrano L (2001) Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. Embo J 20:2528–2535
62. van Beek JHGM, Roger SA, Bassingthwaighte JB (1989) Regional myocardial flow heterogeneity explained with fractal networks. Am J Physiol 257:H1670–H1680
63. Garfinkel AM, Spano L, Ditto WL, Weiss JN (1992) Controlling cardiac chaos. Science 257:1230–1235
64. Shinbrot T, Ditto W, Grebogi C, Ott E, Spano M, Yorke JA (1992) Using the sensitive dependence of chaos (the "butterfly effect") to direct trajectories in an experimental chaotic system. Phys Rev Lett 68:2863–2866
65. Meucci R, Labate A, Ciofini C (1998) Experimental techniques for controlling chaos in lasers. Int J Bifurcat Chaos 8:1759–1768
66. Moss F (1994) Neurobiology: Chaos under control. Nature 370:596–597
67. Garfinkel A, Chen PS, Walter DO, Karagueuzian HS, Kogan B, Evans SJ, Karpoukhin M, Hwang C, Uchida T, Gotoh M, Nwasokwa O, Sager P, Weiss JN (1997) Quasiperiodicity and chaos in cardiac fibrillation. J Clin Invest 99:305–314
68. Verrier RL, Nearing BD, Lovett EG (1997) Complex oscillatory heart rhythm: a dance macabre. J Clin Invest 99:156–157
69. Lloyd AL, Lloyd D (1993) Hypothesis: the central oscillator of the circadian clock is a controlled chaotic attractor. Biosystems 29:77–85
70. Keulers M, Suzuki T, Satroutdinov AD, Kuriyama H (1996) Autonomous metabolic oscillation in continuous culture of Saccharomyces cerevisiae grown on ethanol. FEMS Microbiol Lett 142:253–258
71. Lloyd D, Salgado LE, Turner MP, Suller MT, Murray D (2002) Cycles of mitochondrial energization driven by the ultradian clock in a continuous culture of Saccharomyces cerevisiae. Microbiology 148:3715–3724
72. Satroutdinov AD, Kuriyama H, Kobayashi H (1992) Oscillatory metabolism of Saccharomyces cerevisiae in continuous culture. FEMS Microbiol Lett 77:261–267
73. Lloyd D (2008) Respiratory oscillations in yeast. In: Maroto M, Monk N (eds) Cellular oscillatory mechanisms. Landes Bioscience (in press)
74. Klevecz RR, Bolen J, Forrest G, Murray DB (2004) A genomewide oscillation in transcription gates DNA replication and cell cycle. Proc Natl Acad Sci USA 101:1200–1205
75. Lloyd D, Murray DB (2006) The temporal architecture of eukaryotic growth. FEBS Lett 580:2830–2835
76. Tu BP, Kudlicki A, Rowicka M, McKnight SL (2005) Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. Science 310:1152–1158
77. Aon MA, Cortassa S, Lemar KM, Hayes AJ, Lloyd D (2007) Single and cell population respiratory oscillations in yeast: a 2-photon scanning laser microscopy study. FEBS Lett 581:8–14
78. Murray DB, Klevecz RR, Lloyd D (2003) Generation and maintenance of synchrony in Saccharomyces cerevisiae continuous culture. Exp Cell Res 287:10–15
79. Lloyd D (2006) Hydrogen sulfide: clandestine microbial messenger? Trends Microbiol 14:456–462
80. Roussel MR, Lloyd D (2007) Observation of a chaotic multioscillatory metabolic attractor by real-time monitoring of a yeast continuous culture. Febs J 274:1011–1018
81. Osborne AR, Provenzale A (1989) Finite correlation dimension for stochastic systems with power-law spectra. Physica D 35:357–381
82. Grassberger P, Procaccia I (1983) Measuring the strangeness of strange attractors. Physica D 9:189–208
83. Rao CV, Wolf DM, Arkin AP (2002) Control, exploitation and tolerance of intracellular noise. Nature 420:231–237
84. Arkin AP (2000) Signal processing by biochemical reaction networks. Cambridge University Press, Cambridge
85. Thattai M, van Oudenaarden A (2002) Attenuation of noise in ultrasensitive signaling cascades. Biophys J 82:2943–2950
86. Becskei A, Serrano L (2000) Engineering stability in gene networks by autoregulation. Nature 405:590–593
87. Kosko B (2006) Noise. Viking, New York
88. Goldbeter A, Koshland DE Jr (1982) Sensitivity amplification in biochemical systems. Q Rev Biophys 15:555–591
89. Koshland DE Jr, Goldbeter A, Stock JB (1982) Amplification and adaptation in regulatory and sensory systems. Science 217:220–225
90. Alon U, Surette MG, Barkai N, Leibler S (1999) Robustness in bacterial chemotaxis. Nature 397:168–171
91. Barkai N, Leibler S (1997) Robustness in simple biochemical networks. Nature 387:913–917
92. Vilar JM, Kueh HY, Barkai N, Leibler S (2002) Mechanisms of noise-resistance in genetic oscillators. Proc Natl Acad Sci USA 99:5988–5992
93. Lloyd D (1998) Circadian and ultradian clock-controlled rhythms in unicellular microorganisms. Adv Microb Physiol 39:291–338
94. Schibler U, Naef F (2005) Cellular oscillators: rhythmic gene expression and metabolism. Curr Opin Cell Biol 17:223–229
95. Cortassa S, Aon MA, Winslow RL, O'Rourke B (2004) A mitochondrial oscillator dependent on reactive oxygen species. Biophys J 87:2060–2073

96. Cortassa SC, Aon MA, O'Rourke B, Jacques R, Tseng HJ, Marban E, Winslow RL (2006) A computational model integrating electrophysiology, contraction and mitochondrial bioenergetics in the ventricular myocyte. Biophys J 91:1564–1589
97. Saks V, Dzeja P, Schlattner U, Vendelin M, Terzic A, Wallimann T (2006) Cardiac system bioenergetics: metabolic basis of the Frank-Starling law. J Physiol 571:253–273

Books and Reviews

Bak P (1996) How nature works: the science of self-organized criticality. Copernicus, New York
Barabasi AL (2003) Linked. Plume, New York
Capra F (1996) The web of life. Anchor books Doubleday, New York
Dewey GT (1997) Fractals in molecular biophysics. Oxford University Press, New York
Cortassa S, Aon MA, Iglesias AA, Lloyd D (2002) An Introduction to Metabolic and Cellular Engineering. World Scientific, Singapore
Feder J (1988) Fractals. Plenum Press, New York
Glass L, Mackey MC (1988) From clocks to chaos. The rhythms of life. Princeton University Press, Princeton
Haken H (1978) Synergetics. Springer, Heidelberg
Jantsch E (1989) The self-organizing universe. Scientific and human implications of the emerging paradigm of evolution. Pergamon Press, Oxford
Kauffman SA (1993) Origins of order: Self-organization and selection in evolution. Oxford University Press, New York
Lane N (2005) Power, sex, suicide. Mitochondria and the meaning of life. Oxford University Press, Oxford
Noble D (2006) The music of life. Oxford University Press, New York
Sornette D (2003) Why stock markets crash. Critical events in complex financial systems. Princeton University Press, New Jersey
Varela FJ (1989) Autonomie et connaissance. Seuil, Paris
West BJ, Deering B (1995) The Lure of Modern Science. Fractal Thinking, vol 3. World Scientific, Singapore
Yates EF (1987) Self-organizing systems. The emergence of order. Plenum Press, New York

Biological Complexity and Biochemical Information

Christoph Adami
Keck Graduate Institute of Applied Life Sciences, State University of New York, Claremont, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Measures of Biological Complexity
Biochemical Information
Network Complexity
Future Directions
Acknowledgments
Bibliography

Glossary

C-value The haploid genome size of an organism, measured either in picograms (pg) or base pairs (bp).

Degree distribution The probability distribution $P(d)$ to find a node with d edges in a network.

Entropic profile A graph of the per-site entropy along the sites of a biomolecular sequence, such as a DNA, RNA, or protein sequence.

Epistasis Generally, an interaction between genes, where the fitness effect of the modification of one gene influences the fitness effect of the modification of another gene. More specifically, an interaction between *mutations* that can be either positive (reinforcing or *synergistic*), or negative (mitigating or *antagonistic*).

Erdös–Rényi network A random graph with a binomial degree distribution.

Fitness A numerical measure predicting the long-term success of a lineage.

Jensen–Shannon divergence In probability and statistics, a measure for the similarity of probability distributions, given by the symmetrized relative entropy of the distributions.

Module In network theory, a group of nodes that is closely associated in connections or function, but only weakly associated to other such groups.

Motif In network theory, a subgraph of small size.

Network diameter For networks, the average geodesic distance between nodes, defined as $D = 1/m \sum_{i=1}^{n} \sum_{j=1}^{n} d(i, j)$, where m is the number of edes of the graph, n is the number of nodes, and $d(i, j)$ is the shortest path distance between nodes i and j.

Phylogenetic depth A measure of the genetic distance between a genome and its ancestor on the same line of descent, given by the number of genetically different genomes on the line between the genomes plus one.

Random variable In probability and statistics, a mathematical object with discrete or continuous states that the object takes on with probabilities drawn from a probability distribution associated to the random variable.

Source entropy The entropy of a sequence generated by a process that generates symbols with a given probability distribution.

Wright–Fisher process In population genetics, a stochastic process that describes how genes are transmitted from one generation to the next.

Turing machine In mathematics, an abstract automaton that manipulates symbols on a tape directed by a finite set of rules.

Watson-Crick pairing In biochemistry, the pairing be-

tween nucleotides adenine and thymine (A-T), and guanine and cytosine (G-C).

Zipf's law A relationship between the frequency f and the rank k of words in a text, of the form $f(k) \sim k^s$, where s is the exponent of the distribution.

Definition of the Subject

Biological complexity refers to a measure of the intricateness, or complication, of a biological organism that is directly related to that organism's ability to successfully function in a complex environment. Because organismal complexity is difficult to define, several different measures of complexity are often used as proxies for biological complexity, such as structural, functional, or sequence complexity. While the complexity of single proteins can be estimated using tools from information theory, a whole organism's biological complexity is reflected in its set of expressed proteins and its interactions, whereas the complexity of an ecosystem is summarized by the network of interacting species and their interaction with the environment.

Introduction

Mankind's need to classify the world around him is perhaps nowhere more apparent than in our zeal to attach to each and every living organism a tag that reveals its relationship to ourselves. The idea that all forms of life (and even inanimate matter) belong to a "Great Chain of Being" goes back to medieval times [1] and usually depicts rocks at the bottom of the chain, with plants next, followed by fish, birds, mammals, humans, angels, and ultimately god (see Fig. 1).

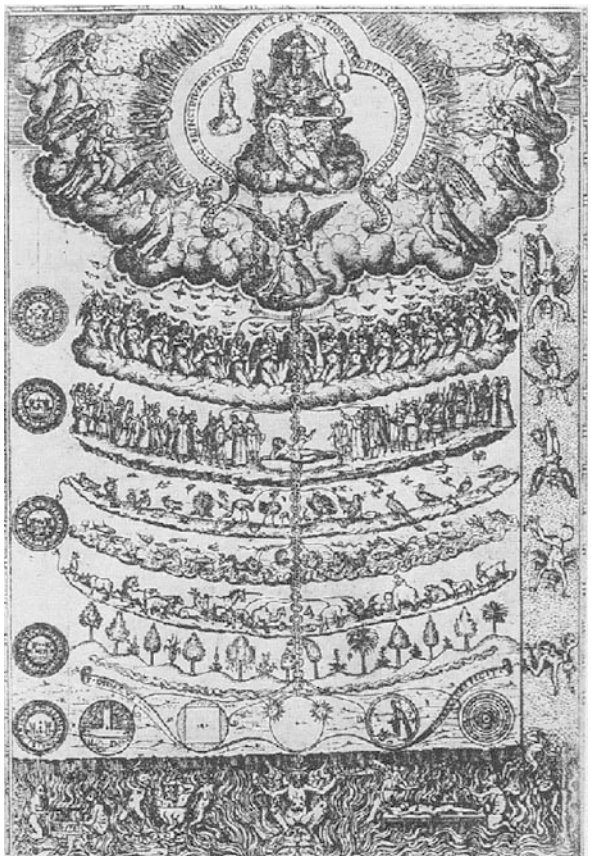

Biological Complexity and Biochemical Information, Figure 1
A depiction of the Great Chain of Being by Fra Diego de Valdes, in *Rhetorica Christiana, from 1579*

In more modern times, biologists and lay people alike have, sometimes unconsciously, given in to the same need by classifying organisms according to their perceived complexity. So, for example, viruses are often perceived as the least complex organisms, followed by bacteria, unicellular eukaryotes, fungi, plants, invertebrates, vertebrates, mammals, and, ultimately, ourselves. Such a hierarchical view of biological complexity with *Homo sapiens* at its top has repeatedly been decried (see, e. g., [2] and recently [3]). However, in the absence of a quantitative measure of biological complexity, neither a position in favor nor against a hierarchical ordering of species can authoritatively be maintained. Indeed, even though it has become less and less accepted to view *Homo* as the "crown jewel" of evolution, there is no reason a priori why this cannot in fact be true.

Many different measures for biological complexity have been introduced in the past literature, some of them with obvious biological roots (such as number of cells, number of different tissue types, genome length, etc.), some of them inspired by the physics of dynamical systems. Neither of these measures has managed to convince a majority that it represents the best proxy for a biological complexity measure, and most have very obvious shortcomings. As we learn more about the genome and proteome of different organisms, it seems as if we are well on our way to accumulate a sufficient amount of detailed knowledge concerning the function of organisms that a universal measure of organism complexity is at last imaginable. For example, the set of all expressed proteins and their interactions (including their regulation and post-translational modification) could conceivably enter into an equation whose result is a number: the biological (functional) complexity of an organism that allows for comparisons across the tree of life.

Such a fantasy, however, is almost surely misguided unless we can enumerate along with the list of proteins and their interactions the possible environments within which

the organism can make a living. After all, the complexity of an organism must depend crucially on the environment within which it functions. Consider as an example a 150 pound rock and an average human, transplanted to the surface of the moon. Functionally, they are very comparable there. Likewise, a human is much less complex than even an algal bloom, if forced under water for an extended period of time. Given that we are unable to characterize even a single environment for most bacteria (environmental microbiologists estimate that fewer than 2% of all bacteria can be cultured in the laboratory) it is highly unlikely that the full functional complexity of most organisms can be ascertained.

Still, for organisms that make a living in comparable environments, a classification in terms of their biological complexity given their bioinformatic pedigree is not unreasonable, and the concepts and issues discussed below are steps toward such a goal. In the following section, I review standard types of complexity measures, how they compare to each other and how they fail to capture the essence of biological complexity. I then discuss a type of sequence complexity that can be shown to reduce to the information encoded into the biochemistry of DNA and proteins, and discuss its advantage over other measures, as well as its practical problems. I discuss complexity measures for biological and biochemical networks in the following section, and close with some thoughts on future directions.

Measures of Biological Complexity

The different measures put forward in the literature to quantify an organism's complexity can be grouped into three main categories: structural, functional, and sequence complexities. A fourth category—network complexity—is very recent, and I will consider that last.

Structural Complexity

Perhaps the most obvious way to define complexity is with reference to the structural complication of an organism. Clearly, structural complexity is not limited to biological organisms alone, and a good measure of this kind might allow us to compare biological organisms to human feats of engineering. Two problems are immediately apparent for any measure of structural complexity. First, to generate a scalar (that is, a single number) that will rank all possible structures appears to be a formidable task. Second, there is no guarantee that structural complexity is always a good predictor for an organism's success in the biosphere. On the one hand, something that looks like a complicated contraption could, in principle, be an evolutionary artifact: a non-adaptive feature that is either a necessity or a consequence of another feature (called a "spandrel" by Gould and Lewontin [4]). On the other hand, a complicated device could conceivably be functional only in a very different environment, perhaps one in which the organism can no longer survive. In other words, complex structure is not necessarily predictive of complex function, although we expect this to be true in the majority of cases.

The second difficulty, however, pales compared to the first. Commonly, a measure of structural complexity attempts to count the number of parts and their connection. Several such measures are reviewed by McShea [5]. A typical example for a structural complexity measure is the number of *different* cell types within an organism [6], perhaps normalized by the total number of cells in order to counter the bias of organism size. Bell and Mooers analyzed such a measure [7], and found that it robustly classifies animals as more complex than plants, and plants more complex than algae. However, the measure remains a very crude estimate of complexity, unable to shed light on finer gradations of the tree of life.

Hierarchical measures of complexity represent a different type of structural complexity. Generally speaking, a measure of hierarchical complexity seeks to quantify the number of levels needed to build a biological system, for example as the minimum amount of hierarchical structuring needed to build an understanding of the system [8]. The problem with a hierarchical scale of complexity for biological systems is that there are only four clear hierarchies: the prokaryote cell, the eukaryotic cell viewed as a symbiotic assembly of prokaryotic cells, the multicellular organism, and colonial individuals or integrated societies [9]. However, it is possible to introduce a higher-resolution scale by decomposing each hierarchy into levels and sublevels, for example by differentiating a monomorphic aggregate of elements of the lower hierarchy from a differentiated aggregate, and an aggregate of nested differentiated elements [9]. Even though a hierarchical measure of complexity necessarily represents a fairly coarse scale, it is one of only few measures of structural complexity that shows an unambiguous increase in complexity throughout the fossil record.

Functional Complexity

Ideally, any measure of biological complexity should be functional, that is, reflecting how the organism functions in a complex world. Because of the obvious difficulty in relating either form to function, or sequence to function, function-based measures are currently even less well-formed than structural ones. Because function

is understood to be a product of natural selection, it is often implicitly assumed that a measure of functional complexity should increase in evolution. However, arguments advanced earlier clearly destroy this notion: it is certainly possible that following a drastic change in the environment the functional complexity of an organism decreases, for example because a preferred metabolic substrate is eliminated from the environment, abolishing the metabolic pathway—and thus the function associated with it—for the unlucky species. A functional complexity measure for single molecules that is based on information theory has recently been introduced by Szostak [10]. I will discuss this measure in more detail in Sect. "Biochemical Information and Functional Complexity".

Attempts to count the number of different functions that an organism can perform are (and perhaps will remain) hopeless. This problem is directly related to our inability to characterize the necessary, or even sufficient, elements of an organism's environment. McShea proposed to catalogue the different behaviors of an organism as proxy for its different functions [11], as behaviors are more amenable to observation than functions. In turn, he suggested that the number of *behavioral parts* of an organism could be used as a proxy for the number of behaviors, if it is true that parts usually play a role only in one function or behavior. He gives the example of the Swiss Army knife as a device more complex than a screwdriver, on account of the different parts that represent different functions. McShea's analysis of functional parts leads him to consider networks of interacting parts that display varying degrees of modularity. I will return to such network-based measures further below.

Sequence Complexity

Given that all forms of life on Earth contain a genetic code that is responsible for generating their form and function, we might naively assume that the amount of haploid DNA (measured either in picograms (pg) as was done before the advent of whole genome sequencing or in millions of base pairs (mbp) as is more common today) would reflect—even if only roughly—the complexity of the organism. This hope was quashed relatively early on: Britten and Davidson showed conclusively [12] that no correlation between genome size and perceived complexity exists. This disconnect has been termed the "C-value paradox" [13] (reviewed in [14]) and is perhaps best exemplified by the giant free living amoeba *Amoeba dubia*, whose total DNA content was estimated at 700 pg, which would correspond to about 675,000 mpb if it was all haploid. This would correspond to about 200 times the C-value of humans. However, the haploidy of the *A. dubia* genome is now in doubt [15]. The variation in genome size and the absence of a correlation to a complexity scale such as that given by the classical chain of being is depicted in Fig. 2.

Because sequence length cannot be used, the focus for measures of sequence complexity instead has been on mathematical measures. The literature on mathematical sequence complexity, or more specifically, the complexity of symbolic strings, is far richer than that concerning functional or structural and morphological complexity. Among sequence complexities there are many different types, such as Kolmogorov, compositional, or information-theoretic ones. I shall give brief expositions of examples of each of these types here, without attempting to be even nearly exhaustive. For a good review of sequence complexities used in physics, see Ref. [16].

The most well-known sequence complexity is that introduced by the Russian mathematician Andrey Kolmogorov. He proposed to assign to each symbolic sequence a scalar that represents the *regularity* of the sequence [17]. So, for example, a string consisting of the repetition of a symbol or pattern is classified as regular, whereas a string with no discernable pattern would be irregular, and therefore complex. Note however that this algorithm does not classify a string as complex just because no pattern is readily identifiable. The Kolmogorov measure assumes that *all possible* computer programs (of a universal Turing machine) are tested so as to find the shortest one that produces the sequence in question. Mathematically, the Kolmogorov complexity of a string s is given by the length of the shortest program p (denoted as $|p|$) that produces s when executed on a universal Turing machine T:

$$K(s) = \min\{|p| : s = C_{\mathrm{T}}(p)\} \ , \tag{1}$$

where $C_{\mathrm{T}}(p)$ denotes the result of running program p on Turing machine T. So, for example, the binary equivalent of the irrational number π is random prima facie, however a concise algorithm (a short program p_π) exists to produce it, leading to a fairly low complexity for s_π. There is a vast amount of mathematical and information-theoretical literature concerning the Kolmogorov complexity (see, e. g., [18]), but when applied to biological sequences, this measure has two obvious flaws.

First, the procedure to generate the Kolmogorov complexity is uncomputable because the search for the smallest program may never end (the computation may not halt). Second, truly random strings, that is, those that cannot be generated by any computation, are assigned maximal complexity. But in physics (and in particular in biology!) truly random sequences are meaningless, and there-

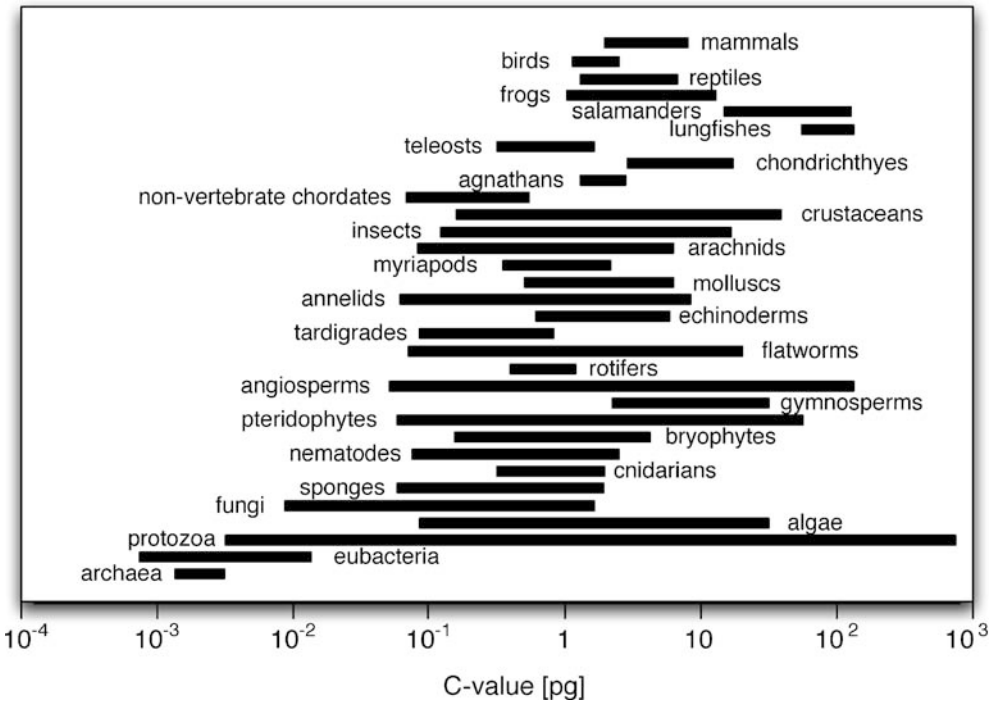

Biological Complexity and Biochemical Information, Figure 2
Ranges in haploid genome size measured in picograms (the C-value) in different organisms [14]

fore should not be assigned a large complexity. In fact, the Kolmogorov complexity is even logically inconsistent. Remember that the procedure to calculate the Kolmogorov complexity is one where an automaton returns the smallest program that produces the sequence in question. For a random sequence, the procedure is to return the sequence itself (and thus the complexity estimate of a random sequence is the length of that sequence, the largest possible result). But, logically, a random sequence can never be the result of a computation, because a computation is a deterministic process. These flaws can be partly fixed by focusing on conditional and mutual Kolmogorov complexities, as I outline further below. In summary, while the Kolmogorov complexity is a mathematically sound measure of *mathematical* sequence regularity, it is not a measure of the complexity of a sequence that describes a physical object.

The key concept to account for objects in the physical—rather than mathematical—world is the *conditional Kolmogorov complexity* (first introduced in [17]) of a string s given another string t, defined as the length of the smallest program $|p|$ that produces s using that program and an external string t:

$$K(s|t) = \min\{|p| : s = C_{\mathrm{T}}(p|t)\} \ , \qquad (2)$$

with the notation $(p|t)$ pronounced as "p given t". Now, t can be a sequence from the physical world, or describing an object in the physical world. The program p is small if the sequence s can be obtained from t using a simple computation. Using this construction, the conditional complexity of a random sequence r can be defined rigorously, because the shortest program to produce a random sequence r involves both the sequence r as input to the machine, and the vanishingly small (in the limit of infinite strings) program p="print". In other words, the conditional Kolmogorov complexity of random strings *vanishes* in the limit of long sequences, rather than being maximal.

To define a physical complexity of a symbolic string (physical, because it refers to a physical world, rather than to abstract mathematics), we can imagine a sequence e that represents everything that can be measured in that world. We now ask of our Turing machine to compute string s given everything that is knowable about the physical world, that is, given e. In that case, the conditional complexity $K(s|e)$ represents everything that *cannot* be obtained from knowing the physical world. In other words, it represents the "remaining randomness": the unmeasurable. Naturally, the physical complexity of the string is then just the length of the sequence *minus* the remaining

randomness [19]:

$$C_P(s) = K_0(s) - K(s|e) \,, \tag{3}$$

where $K_0(s)$, the Kolmogorov complexity in the absence of both a "world" string e and the rules of mathematics, is just the length of s: $K_0(s) = |s|$. This notation is chosen in anticipation of results from information theory introduced in the following section. In particular, it allows us to see that $C_P(s)$ is simply a *mutual Kolmogorov complexity* [17], between string s and the world e. Put in another way, the physical complexity of a sequence is that part of the sequence that can be obtained from the world string e using a concise (and therefore short in the limit of very long sequences) computation using program p. Only *those* sequences can be obtained by computation from the world e that *mean* something in world e, or refer to something there. Note that by its construction, the physical complexity represents a special case of the "effective complexity" measure introduced earlier by Gell-Mann and Lloyd [20].

As an example, consider a world of machines, whose blueprints are stored in such a way that they can be represented as sequences of symbols. In this world, e represents all the blueprints of all the possible machines that exist there. A sequence s is complex in this world if a part—or all—of the sequence can be obtained by manipulating, or translating, the world tape e. (It is conceivable that s contains part of a blueprint from e in encrypted form, in which case the program p must try to compare e to encrypted forms of s.) Of course, it would be an accident bordering on the unreasonable to find that a string that is complex in e is also complex mathematically. Instead, from a mathematical point of view, such a sequence most likely would be classified as random, or rather, the search for a shortest program would not halt. Similarly, it is extremely unlikely that sequence s would be classified as complex in a world in which e represents, say, all the literature produced on Earth (unless there are a few books on the blueprints of certain machines!). Thus, the complexity of a sequence s, by this construction, is never absolute (like in mathematics), but always conditional with respect to the world within which the sequence is to be *interpreted*.

This is precisely what we need in order to quantify the complexity of biological sequences, because it is immediately clear that a biological sequence only means something in a very specific environment, given very specific rules of chemistry. So, according to this argument, the sequence describing a particular organism is complex only with respect to the environment within which that organism "makes its living", that is, its niche. Take the organism out of its niche, and it is unlikely to function as well as in its native niche; some of its structural complexity may turn into a useless appendage or, worse, a liability. In the following section, we will see under what circumstances this physical complexity can be understood in terms of information theory.

Biochemical Information

The inception of information theory [21] created the widely accepted expectation that biology would ultimately become a subdiscipline of cybernetics, information theory, and control theory [22]. That this expectation has not come to pass lies partly in a gross underestimate of the complexity of molecular and cellular biology by engineers and physicists, and partly in a fundamentally misguided application of information theory to genomes and molecular sequences in general. Most of the early applications of information theory to biomolecular sequences focused on an estimation of the entropy of sequences [23], rather than the information content. For example, a significant amount of work was expended on estimating the compressibility of DNA and protein sequences by studying long range correlations [24,25,26], all the while mistaking the average per-site entropy for information content [27]. These studies concluded, for the most part, that coding DNA sequences and proteins are *essentially random*, or uncompressible (non-coding DNA was found to be less random, due to repeats). From the point of view of coding theory, such a finding should not have been surprising, as Shannon's coding theorem [28] implies that the length of a message conveyed by a symbolic sequence is limited by the per-site (source) entropy [29]. In other words, we expect evolution to try to maximize the source entropy. However, source entropy does not equal information content. Let us briefly review the basics of information theory as applied to molecular sequences [30].

Entropy

For a random variable X that can take on states $x_1, \ldots, x_D$ with probabilities $p(x_i)$, the entropy of X is given by

$$H(X) = -\sum_{i}^{D} p(x_i) \log p(x_i) \,, \tag{4}$$

where the logarithm is taken to a convenient base, and determines the units of the entropy. If the base is chosen to be 2, for example, the entropy is given in bits. Often, it is convenient to take the number of possible states D as the basis. In that case, the entropy is bounded from above by 1, and a sequence of N such random variables (a polymer) has a maximal entropy of N.

Suppose we are given a DNA sequence

$$\texttt{AGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGA}\\ \texttt{TCCCGCATAGCTCCACCA} \tag{5}$$

We can either take the stance that this is one of 4^{56} possible DNA sequences of length $N = 56$, or we can imagine that this is the record of 56 consecutive independent trials of a single DNA random variable that can take on the states $x = \texttt{A}, \texttt{T}, \texttt{G}, \texttt{C}$ only. If we take the latter view, then we can try to estimate the *source entropy* by using the results of the 56 measurements to calculate the *plug-in probabilities*:

$$\begin{aligned} p_{\texttt{A}} &\approx \frac{10}{56} \approx 0.18\,, \quad p_{\texttt{T}} \approx \frac{11}{56} \approx 0.20\,, \\ p_{\texttt{G}} &\approx \frac{16}{56} \approx 0.29\,, \quad p_{\texttt{C}} \approx \frac{19}{56} \approx 0.34\,. \end{aligned} \tag{6}$$

These probabilities are quite uncertain due to the lack of statistics, and only give us limited knowledge about the nature of the gene. The deviation of the calculated probabilities from the unbiased prediction $p_i = 1/4$ (an independent identically distributed, or i.i.d, random variable) is not significant. Indeed, using the four probabilities above, we obtain as an estimate of the source entropy

$$H(X) = -\sum_{i=\texttt{A},\texttt{C},\texttt{G},\texttt{T}} p(x_i)\log_4 p(x_i) \approx 0.978\,, \tag{7}$$

which is almost maximal and confirms that, whatever message is encoded in the sequence, it is encoded in a nearly maximally compressed manner. For longer sequences, an analysis of the usage probabilities can reveal important information such as a bias in nucleotide usage, which will reduce the source entropy.

A typical application of sequence entropy is the calculation of "n-gram" entropies for blocks of sequences of length n:

$$H_n = -\sum_{i}^{4^n} P_i^{(n)} \log P_i^{(n)}\,, \tag{8}$$

where $P_i^{(n)}$ represents the probability to find the ith of the 4^n blocks of length n in the sequence. H_n then measures the average uncertainty within an average n-gram. We can write down the difference [31]

$$h_n = H_{n+1} - H_n\,, \tag{9}$$

which is sometimes called the amount of information necessary to predict the next symbol on a sequence given the n previous symbols. For example, for the sequence Eq. (5), $H_2 = 1.877$ (based on 55 2-grams) so that $h_1 = H_2 - H_1 = 0.899$, using the per-site entropy Eq. (7) for H_1. Note, however, that the difference Eq. (9) is, strictly speaking, not a measure of information but rather the conditional entropy (see below) to find a symbol A given the sequence S_n of length n: $h_n = H(A|S_n)$.

Using the entropy per letter $h = \lim_{n\to\infty} h_n$, we can calculate Grassberger's *effective measure complexity* [31]

$$EMC = \sum_{n=0}^{\infty}(h_n - h)\,. \tag{10}$$

This measure sums up the "memory" effects within the string, that is, it sums up the correlations at all scales. This measure vanishes if all sites are independent (because then $H_n = nh$ and $h_n = h$) and is maximal if there are strong long-range correlations within a sequence. The utility of this measure to capture the complexity of a gene is much in doubt, however, because the ability to predict the next symbol on a sequence, as discussed before, is unlikely to shed light on the function and utility of the gene if the information contained in the sequence is *encoded*.

The *compositional complexity* is another usage of n-gram entropies to capture sequence complexity, but this measure attempts to find the optimal segmentation of the sequence into m partitions based on a significance criterion. The more heterogeneous a sequence is, the higher the compositional complexity C_{comp} of a sequence s with length L [32]

$$C_{\text{comp}} = \max_{\pi} J(s_m)\,, \tag{11}$$

which is the maximum over all possible partitions π of the Jensen–Shannon divergence

$$J(s_m) = H(s) - \sum_{i=1}^{m} \frac{l_i}{L} H(s_i)\,, \tag{12}$$

where $H(s)$ is again the source entropy, and $H(s_i)$ is the entropy of the ith segment of length l_i. As with most of the sequence measures discussed up to now, this measure addresses coding-style more than function, and is unlikely to capture the essence of functional information.

Conditional Entropy

Before we can talk about information, we have to introduce the concept of *conditional probabilities*, that is, the probability $p(x_i|y_j)$ that a random variable X takes on one of its states x_i given that another variable Y is in one of its

states y_j. Then, the conditional entropy of X given that Y is in state y_j can be written as

$$H(X|Y = y_j) = -\sum_{i}^{D} p(x_i|y_j) \log p(x_i|y_j) \,. \tag{13}$$

This concept allows us to discover the *relative state* of two different random variables, or in this case two different nucleotide positions. Suppose we are given 32 more sequences like the one in Eq. (5), and suppose we are told that these sequences represent the *same information*, that is, same gene, but from different samples. After aligning these 33 sequences, the result looks like Table 1.

The extra symbols '–' and '.' in this alignment need further explanation. The symbol '–' denotes a "missing" symbol, that is, a deletion in that sequence with respect to the other sequences. The symbol '.', on the other hand, signals that the identity of the nucleotide at that position is not known. Here, we will treat each of these symbols as if there was no data for the random variable at that position, reducing the statistics at that site. Clearly, we now are viewing the set of sequences in Table 1 as representing independent samples of a joint random variable

$$Z = X_1 X_2 \cdots X_{56} \tag{14}$$

Using the conditional entropy, we can now study *correlations* between nucleotide positions. For example, we can study position 8, or more precisely, the random variable X_8. We can gather a probability estimate for this position by gathering statistics *vertically*, rather than horizontally as before. Here, we find

$$\begin{aligned} p_{\mathrm{A}}(8) &= 5/33 \,, \quad p_{\mathrm{C}}(8) = 17/33 \,, \\ p_{\mathrm{G}}(8) &= 5/33 \,, \quad p_{\mathrm{T}}(8) = 6/33 \,. \end{aligned} \tag{15}$$

which, even though 33 sequences does not represent a large sample, are significantly different from the random assumption $p_i = 1/4$. Even more important deviations from the i.i.d case can be found for X_{33} and X_{36}: only one particular nucleotide can be found at these positions, all others are absent. The reasons for this *abnormal* distribution is immediately clear when we reveal the origin of these sequences: they are the last 56 nucleotides of the tRNA gene of the *E. coli* bacterium [33], and the function of this molecule requires positions 33 and 36 to have that particular nucleotide. A mutation leading to an alternative nucleotide at these positions implies the death of the organism carrying the mutant, which explains why we do not find it in a sample of living organisms. We can now check how uneven the probabilities per-site are, by plotting the per-site entropy (normalized to lie between zero and one by choosing base $D = 4$ for the logarithm) against site-number in Fig. 3.

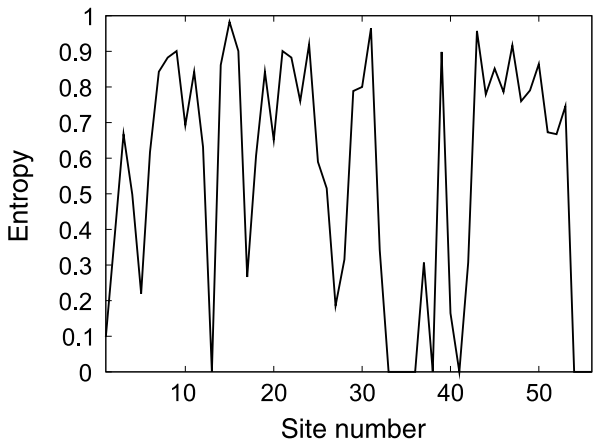

Biological Complexity and Biochemical Information, Figure 3
Per-site entropy of the last 56 positions of *E. coli* tRNA from the alignment in Table 1

We notice that there are sites that are almost random, and some sites that are absolutely conserved. While this pattern, conserved by evolution over billions of years, clearly reflects the function of the organism, it does not tell us whether the states of any two sites are correlated.

The relative state of two random variables will tell us whether knowing the state of one variable will reveal to us information about another variable. For example, we can ask whether knowing, say, X_8, allows us to say something about other sites that we would not already know in the absence of that knowledge. Let us see whether the probabilities characterizing X_{21}, say, depend on knowing X_8. Collecting frequencies for X_{21} gives approximately

$$\begin{aligned} p_{21}(\mathrm{A}) &= 0.24 \,, \; p_{21}(\mathrm{C}) = 0.46 \,, \\ p_{21}(\mathrm{G}) &= 0.21 \,, \; p_{21}(\mathrm{T}) = 0.09 \,, \end{aligned} \tag{16}$$

while $p(X_{21}|X_8) =$

$$\begin{pmatrix} p(\mathrm{A}|\mathrm{A}) & p(\mathrm{A}|\mathrm{C}) & p(\mathrm{A}|\mathrm{G}) & p(\mathrm{A}|\mathrm{T}) \\ p(\mathrm{C}|\mathrm{A}) & p(\mathrm{C}|\mathrm{C}) & p(\mathrm{C}|\mathrm{G}) & p(\mathrm{C}|\mathrm{T}) \\ p(\mathrm{G}|\mathrm{A}) & p(\mathrm{G}|\mathrm{C}) & p(\mathrm{G}|\mathrm{G}) & p(\mathrm{G}|\mathrm{T}) \\ p(\mathrm{T}|\mathrm{A}) & p(\mathrm{T}|\mathrm{C}) & p(\mathrm{T}|\mathrm{G}) & p(\mathrm{T}|\mathrm{T}) \end{pmatrix} = \begin{pmatrix} 0.2 & 0.235 & 0 & 0.5 \\ 0 & 0.706 & 0.2 & 0.333 \\ 0.8 & 0 & 0.4 & 0.167 \\ 0 & 0.059 & 0.4 & 0 \end{pmatrix} . \tag{17}$$

Biological Complexity and Biochemical Information, Table 1
Alignment of the last 56 nucleotides of sequences of tRNA genes of *E. coli*

```
AGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
AGAGCGCTTGCATGGCATGCAAGAGGTCAGCGGTTCGATCCCGCTTAGCTCCACCA
TATGTAGCGGATTGCAAATCCGTCTA-GTCCGGTTCGACTCCGGAACGCGCCTCCA
AGAATACCTGCCTGTCACGCAGGGGGTCGCGGGTTCGAGTCCCGTCCGTTCCGCCA
AGGACACCGCCCTTTCACGGCGGTAA-CAGGGGTTCGAATCCCCTAGGGGACGCCA
AGAGCAGGGGATTGAAAATCCCCGTGTCCTTGGTTCGATTCCGAGTCCGGGCACCA
ATTACCTCAGCCTTCCAAGCTGATGA-TGCGGGTTCGATTCCCGCTGCCCGCTCCA
AGAGCACGACCTTGCCAAGGTCGGGGTCGCGAGTTCGAGTCTCGTTTCCCGCTCCA
AGAACGAGAGCTTCCCAAGCTCTATA-CGAGGGTTCGATTCCCTTCGCCCGCTCCA
AGAGCCCTGGATTGTGATTCCAGTTGTCGTGGGTTCGAATCCCATTAGCCACCCCA
AGAGCGCACCCCTGATAAGGGTGAGGTCGGTGGTTCAAGTCCACTCAGGCCTACCA
AGAGCAGGCGACTCATAATCGCTTGGTCGCTGGTTCAAGTCCAGCAGGGGCCACCA
AGAGCAGTTGACTTTTAATCAATTGGTCGCAGGTTCGAATCCTGCACGACCCACCA
AGAGCACATCACTCATAATGATGGGGTCACAGGTTCGAATCCCGTCGTAGCCACCA
AGAACGGCGGACTGTTAATCCGTATGTCACTGGTTCGAGTCCAGTCAGAGGAGCCA
AGCGCAACTGGTTTGGGACCAGTGGGTCGGAGGTTCGAATCCTCTCTCGCCGACCA
AGCGCACTTCGTTCGGGACGAAGGGGTCGGAGGTTCGAATCCTCTATCACCGACCA
AGCGCACCGTCATGGGGTGTCGGGGGTCGGAGGTTCAAATCCTCTCGTGCCGACCA
AAGGCACCGGTTTTTGATACCGGCATTCCCTGGTTCGAATCCAGGTACCCCAGCCA
AAGGCACCGGATTCTGATTCCGGCATTCCGAGGTTCGAATCCTCGTACCCCAGCCA
AGAGCGCTGCCCTCCGGAGGCAGAGGTCTCAGGTTCGAATCCTGTCGGGCGCGCCA
AGAGCAACGACCTTCTAAGTCGTGGGCCGCAGGTTCGAATCCTGCAGGGCGCGCCA
AGAGCAACGACCTTCTAAGTCGTGGGCCGCAGGTTCGAATCCTGCAGGGCGCGCCA
AGAGTACTCGGCTACGAACCGAGCGGTCGGAGGTTCGAATCCTCCCGGATGCACCA
ATAACGAGCCCCTCCTAAGGGCTAAT-TGCAGGTTCGATTCCTGCAGGGGACACCA
AGAGCGCACCCTTGGTAGGGGTGGGGTCCCCAGTTCGACTCTGGGTATCAGCACCA
AGAGCGCACCCTTGGTAAGGGTGAGGTCGGCAGTTCGAATCTGCCTATCAGCACCA
AGAGCAACTGACTTGTAATCAGTAGGTCACCAGTTCGATTCCGGTA.TCGGCACCA
AGAGCAGCGCATTCGTAATGCGAAGGTCGTAGGTTCGACTCCTATTATCGGCACCA
AGAGCGCACCCTTGGTAAGGGTGAGGTCCCCAGTTCGACTCTGGGTATCAGCACCA
AGAGCACCGGTCTCCAAAACCGGGTGTTGGGAGTTCGAGTCTCTCCGCCCCTGCCA
AGCTCGTCGGGCTCATAACCCGAAGATCGTCGGTTCAAATCCGGCCCCCGCAACCA
AGCTCGTCGGGCTCATAACCCGAAGGTCGTCGGTTCAAATCCGGCCCCCGCAACCA
```

Knowing X_8 would help us in predicting the state of X_{21} if any of the numbers in any of the columns of Eq. (17) is significantly larger than 0.25, which is the probability to guess X_{21} right by random chance. In fact, however, only a very few of the probabilities are even above 0.5. Clearly, the knowledge of position 8 does not help much in predicting position 21. Given $X_8 = \mathtt{G}$, for example, the probabilities to correctly predict A, C, G or T at position 21 are given by the third column of Eq. (17), and none of those probabilities exceed 0.5. The only case where knowing X_8 helps us significantly in predicting X_{21} is if $X_8 = \mathtt{A}$, since then $X_{21} = \mathtt{G}$ with probability 0.8. But let us look instead at another pair of sites, this time X_8 and X_{22}. We can obtain the *unconditional* probabilities for X_{22} from the alignment (as before, we are rounding the probabilities)

$$\begin{aligned} p_{22}(\mathtt{A}) &= 0.18\,, \quad p_{22}(\mathtt{C}) = 0.15\,, \\ p_{22}(\mathtt{G}) &= 0.52\,, \quad p_{22}(\mathtt{T}) = 0.15\,, \end{aligned} \tag{18}$$

but the conditional probability matrix is very different from Eq. (17):

$$p(X_{22}|X_8) = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \end{array}. \tag{19}$$

This matrix immediately tells us that, based on the alignment in Table 1, if we see a T at position 8, then we can be certain ($p = 1$) to find A at position 22, while if we see a G at position 8, we are sure to encounter C at 22, and so forth. Obviously, these are the associations implied by Watson-Crick pairing, so this conditional probability matrix suggests to us that position 8 and 22 are in fact *paired* within the molecule: this is how biochemistry stores information in molecules.

Calculating the conditional entropy $H(X_{22}|X_8 = \mathtt{T})$ makes the correlation apparent in another way. As the entropy of a random variable measures the amount of uncertainty that we have about it, we immediately see that

knowing X_8 tells us everything about the state of X_{22} that there is to know: for every one of X_8's possible states, $H(X_{22}|X_8 = \mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}) = 0$. We can now introduce the *average conditional entropy*:

$$H(X|Y) = \sum_{j=1}^{D} p(y_j) H(X|Y = y_j)\,, \tag{20}$$

which gives us the average value of the entropy of X given any of the states of Y. For the pair (X_{22}, X_8) this gives of course $H(X_{22}|X_8) = 0$, while for the pair (X_{21}, X_8) we find

$$H(X_{21}|X_8) = 0.58\,. \tag{21}$$

We saw earlier that knowing site 8 does not significantly affect our probability to predict site 21, but still the conditional entropy Eq. (21) is significantly smaller than the unconditional one, $H(X_{21}) = 0.9$.

Information

From the example we just discussed, the definition of biochemical information is clear: it is just the reduction in entropy (or uncertainty) of one variable using the knowledge of the state of another. So, if Y is again a variable whose entropy we seek to decrease using our knowledge of X, then the information X conveys about Y is just

$$I(X:Y) = H(Y) - H(Y|X)\,. \tag{22}$$

This definition of information (also sometimes called "shared" or "mutual" entropy) is symmetric: $I(X:Y) = I(Y:X)$, and we can therefore equivalently write

$$I(X:Y) = H(X) - H(X|Y)\,. \tag{23}$$

Thus, "what X knows about Y, Y also knows about X". This is particularly clear for the example of biochemical information through Watson-Crick pairing that we looked at above. We can calculate the information position 8 has about position 22 as

$$I(X_{22}:X_8) = H(X_{22}) - H(X_{22}|X_8) = H(X_{22}) \tag{24}$$

because $H(X_{22}|X_8) = 0$ as we found above. At the same time,

$$I(X_8:X_{22}) = H(X_8) - H(X_8|X_{22}) = H(X_8) \tag{25}$$

because $H(X_8|X_{22})$ also vanishes. Then, $H(X_8)$ and $H(X_{22})$ have to be equal. We can repeat this analysis for the other pair we looked at:

$$I(X_8:X_{21}) = H(X_{21}) - H(X_{21}|X_8) = 0.9 - 0.58 = 0.42\,. \tag{26}$$

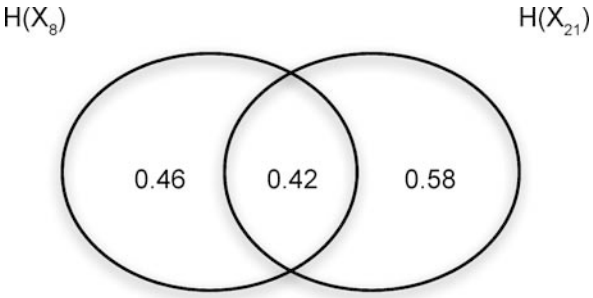

Biological Complexity and Biochemical Information, Figure 4
Venn diagram of entropies between sites 8 and 21 of the sequence alignment (1)

A simple diagram (Fig. 4) helps to see how entropies are distributed among shared and conditional entropies.

As we discussed earlier, if two sites share most of their entropy, we can conclude that they bind to each other in a Watson-Crick pair. Because this binding is responsible for the structure of the molecule, information theory can help us to determine the molecule's *secondary structure*, that is, how the molecule is arranged as a chain in two dimensions [34].

Molecular Complexity

While it is important to understand how biochemistry encodes information in chemical bonds, calculating information-theoretical correlations between nucleotides (or between residues in proteins [35,36,37,38]) does not reveal how much information a molecule stores about the environment within which it functions. We can achieve this by imagining that the physical and chemical world has a *description* in terms of a symbolic sequence. Then, given this particular sequence (and thus, given the identity of the world within which we study the entropy of biomolecular sequences), the set of information-rich sequences can be determined, and the information content calculated. The mathematical form of the measure of a molecule's information content will be shown to be closely related to the physical complexity measure discussed earlier in Eq. (3).

We start by considering a molecule that is guaranteed *not* to encode information about its environment: a random molecule. Let us imagine a polypeptide P of L residues, written as a joint random variable $P = P_1 P_2 \cdots P_L$. If P is truly random, then each of its 20^L states are equally likely, and the entropy of P is

$$H(P) = \log 20^L\,. \tag{27}$$

If we choose to take 20 as the base of the logarithm, we see that the entropy of the random L-mer is $H(P) = L$, that is, one unit of entropy for each random monomer. Of course, functional proteins are nowhere near random. Instead, the probabilities to find a particular residue x at position n, $p_n(x)$, are strongly constrained by evolution. The opposite extreme of a protein would be the case where one and only one particular residue is allowed at each position in a protein that has a specific function. Then, only one state of the polymer is consistent with that function, and $H(P) = 0$. In general, the entropy of P is given by

$$H(P) = -\sum_{i}^{20^L} p_i \log p_i \,, \tag{28}$$

where p_i is the probability to find any of the possible 20^L states of P in an infinitely large ensemble of polymers of length L. In practice, however, ensembles large enough to estimate these p_i cannot exist (note that 20^{100} is approximately 10^{130}), so we need to find approximations to calculate $H(P)$ if we would like to measure a polymer's entropy. A common approximation is to neglect interactions between the different sites n, so that the entropy of P can be written in terms of a sum over the entropies of each monomer:

$$H(P) \approx \sum_{n=1}^{L} H(P_n) \,. \tag{29}$$

In practice this is not a bad approximation [39], but we should keep in mind that interactions between residues, known as *epistasis*, are extremely important in evolution and are often studied in detail [40,41,42]. Typically, while many pairs of residues interact epistatically, some do so positively and some negatively, so that on average the approximation Eq. (29) often holds.

If we can consider each residue separately, we can now focus on the probabilities $p_n(x)$ introduced above. Typically, we expect those sites that are very important for the function of the protein to be strongly conserved (all the $p_n(x)$ are either one or zero at position n), leading to a vanishing entropy at that site, while those that are less important are less constrained, leading to a larger entropy at those sites. We can now define the amount of information a sequence P stores about its environment as the mutual entropy (information) between the random variable describing protein P and a random variable E describing all possible environments (we imagine, as before, that all possible environments can be described in terms of sequences, and listed)

$$I(P:E) = H(P) - H(P|E) \,, \tag{30}$$

which is the same formula as Eq. (46). The first term on the right hand side of Eq. (30) is the *unconditional* entropy of the polymer P. An entropy that is not conditional on any environment is an entropy that is unspecified. This is the same as the entropy of a random protein, and thus $H(P) = H_{\max}(P) = L$. The second term in that equation is the average conditional entropy of the protein, averaged over all the possible environments described by the random variable E. Our world, of course, is just one particular such environment $E = e$, and therefore the amount of information stored in a polymer P about environment $E = e$ is given by

$$I(P:e) = L - H(P|e) \,, \tag{31}$$

where

$$H(P|e) = \sum_{i=1}^{L} H(P_i|e) = -\sum_{n=1}^{L}\sum_{x=1}^{20} p_n(x) \log_{20} p_n(x) \,, \tag{32}$$

and the sum over x goes over the possible 20 amino acids at each site.

The probabilities $p_n(x)$ can be obtained by a sequence alignment of structurally identical proteins of different species, such as those listed in the *Pfam* database [43]. As an example, we can align the sequences of the homeobox proteins of the rodent family *Rodentia*. Aligning 703 of the sequences of 57 residues in the database (as of July 2006) allows us to estimate the $p_n(x)$ necessary for calculating the rodent homeodomain information content. Note that for this analysis to be significant, we have to ensure that the sequences of all the aligned proteins code for a protein with the *same* functionality. Indeed, any time two different residues are allowed at a particular position, we must be able to imply that these are neutral substitutions in the protein. Adaptive changes that influence protein function should not appear in the alignment. The average sequence identity of the set of 703 proteins is about 39%, which gives confidence that the set is composed of sequences coding for proteins that are at least structurally identical. (Sequences with more than 30% identity have more than a 90% chance of coding for structurally identical proteins [44].)

When estimating entropies from finite ensembles, care must be taken to correct the estimates for a bias that arises when the ensemble is small. The method for correcting this bias is well-known [45,46], and applied to all data shown here. We start by calculating the per-site entropy of all 57 aligned sites, giving the *entropy profile* of the sequence shown in Fig. 5. This view reveals a curious alterna-

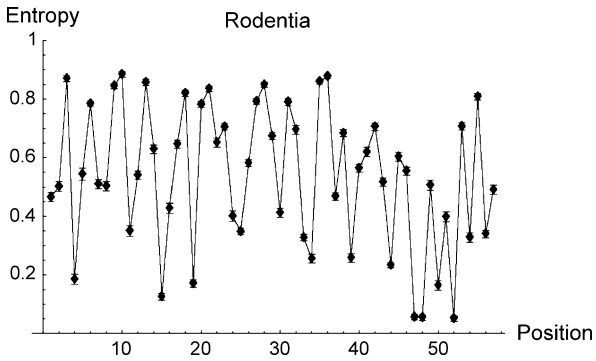

Biological Complexity and Biochemical Information, Figure 5
Entropic profile of the 57 amino acid rodent homeodomain, obtained from 703 sequences in Pfam (St. Louis mirror, accessed July 20, 2006). Per-site entropy between zero and one by taking logs to base 20

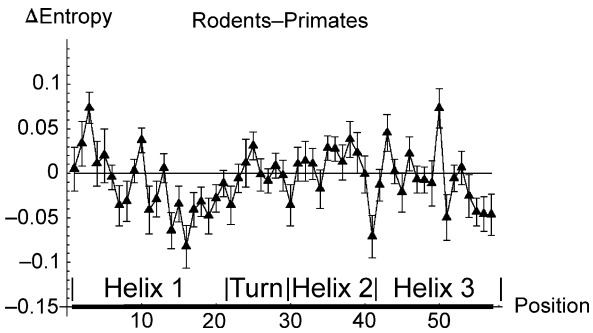

Biological Complexity and Biochemical Information, Figure 6
Difference between entropic profile of the homeobox protein of rodents and primates (the latter from 504 sequences in Pfam St. Louis, accessed July 20, 2006)

tion between high and low entropy sites in the homeobox protein.

The information content is obtained by summing up the per-site entropies and subtracting this value from the length of the sequence, as implied by Eq. (31). If we call the unit of information obtained by using logarithms to the base of the size of the alphabet the "mer" (such that an L-mer has a maximum information content of L mers) then the information content of rodent homeobox proteins is

$$I(P_{\text{rodents}}) = 26.83 \pm 0.13 \text{ mers} , \tag{33}$$

where the quoted error reflects the statistical nature of the probability estimate, quantified in [45,46]. Note that for amino acids, 1 mer equals about 4.32 bits, while for nucleic acids 1 mer equals 2 bits.

From the point of view of evolution, we may ask whether this information content has changed as species evolved. For example, we might ask whether animals that, according to our intuitive understanding of the "great chain", are considered "higher" than rodents show a different information content for this protein. The homeobox protein sequences of many other families of organisms are available in *Pfam* to examine this. As an example, the alignment of 504 sequences of primate homeobox proteins results in an entropic profile remarkably similar to that of Fig. 5. The information content for these proteins can be calculated to be

$$I(P_{\text{primates}}) = 26.71 \pm 0.14 \text{ mers} , \tag{34}$$

in other words, identical to that of the rodents within statistical error. But just because the total information content is the same does not imply that information is coded in a similar manner. We can study this by subtracting one entropic profile form the other, to see if the entropy of some sites has increased, and some others decreased in the evolution from rodents to primates. We can see this difference plot in Fig. 6, which suggest that some recoding did indeed take place, in particular in the first helix of the protein, but the significance of this result is not strong.

Thus, information theory can track some aspects of evolutionary changes between proteins, even if the total amount of information is unchanged.

Biochemical Information and Functional Complexity

It is tempting to speculate that the information content of a biomolecule is related to the functional complexity of an organism. After all, possessing information about an ensemble enables the *prediction* of the possible states of an ensemble with accuracy better than random, something that is highly valuable for a biological organism whose ensemble is an uncertain environment. Let us recall the measure of sequence complexity ("physical complexity") introduced in the previous section:

$$C_P(s) = K_0(s) - K(s|e) , \tag{35}$$

for any sequence s, given an environment sequence e. If we take the average of this quantity over an infinite ensemble of sequences s_i drawn from an ensemble S, we obtain

$$\langle C_P(s)\rangle_S = \sum_{s_i} p(s_i)\,(K_0(s_i) - K(s_i|e)) , \quad s_i \in S . \tag{36}$$

It is easy to prove the following inequality between average Kolmogorov complexities and the Shannon entropy [47]:

$$\sum_{s_i} p(s_i)K(s_i) \geq \sum_{s_i} p(s_i) \log \frac{1}{p(s_i)} = H(S) . \tag{37}$$

The inequality in Eq. (37) reflects the possibility that the complexities $K(s_i)$, which can be viewed as compressed *encodings* of the sequences s_i, do not necessarily form a *perfect code* that saturates the Kraft inequality [48]. However, because the $K(s_i)$ do represent the *smallest* program encoding s_i, it is reasonable to assume that the average Kolmogorov complexity is given by the Shannon entropy of the ensemble up to an additive constant, which represents the length of a program that tells one computer how to simulate another (see, e. g., [48]):

$$\langle K(s_i)\rangle_S \approx H(S) + c\,. \tag{38}$$

In that case (and assuming that the overall constant cancels from the difference), the average physical complexity becomes

$$\langle C_P(s)\rangle_S \approx H(S) - H(S|e)\,, \tag{39}$$

where $H(S)$ is the unconditional Shannon entropy of the ensemble of sequences, and $H(S|e)$ is the conditional entropy. If the ensemble S consists of sequences of fixed length L, then the unconditional entropy is $H(S) = L$, and $H(S|e)$ is the conditional entropy of the sequences as in Eq. (31). (Note that, technically, the Kolmogorov complexity for fixed length sequences $K(s|L)$ is related to the arbitrary length complexity $K(s)$ via $K(s) \leq K(s|L) + 2\log L + c$, where c is again the "simulation constant".) To summarize, the average physical complexity is (assuming perfect coding) equal to the Shannon information that the ensemble has about the environment, that is, a sequence's information content.

This interpretation of complexity is particularly satisfying from an evolutionary point of view. The value of information lies in the ability of the observer who is in possession of it to make *predictions* about the system that the information is *about*. Organisms, armed with the functionality bestowed upon them by their genetic code, do precisely that to survive. An organism's metabolism is a chemical machine making predictions about the availability and concentrations of the surrounding chemicals. A cell's surface proteins make predictions about the type of cells it might interact with, and so on. Viewed in this way, informational complexity should be a near perfect proxy for functional complexity, because information must be used for function: if it is not so used, a sequence represents entropy, not information. An investigation of the informational complexity of evolving computer programs (an instance of "digital life" [49,50]) has shown that the complexity increases in evolution [51] and correlates well with the functional complexity of the programs [52]. A good example for how the equivalence of function and information is achieved in biochemistry is the evolution of functionality in ribozymes by in-vitro evolution, achieved by Jack Szostak's group at Massachussets General Hospital.

This group evolved short GTP-binding RNAs (aptamers) in vitro, and found eleven distinct structures with different binding affinities [53]. By measuring the information content as outlined here (within each pool of sequences that evolved for each structure, there were sufficient mutants that could be aligned in order to determine the substitution probabilities $p_n(x)$), it was possible to show that increased functional activity went hand-in-hand with increased information content, so much so that the group was able to derive a simple law that predicts, within this GTP-binding set of ribozymes, that a ten-fold higher binding affinity is achieved by about 10 bits of extra information. In other words, the informational complexity is linearly proportional to the functional complexity. Even more, the structural complexity, as measured by the number of different stems (ladders) within the secondary structure of the enzyme, also seemed to increase with functional activity.

Based on this type of observation, Szostak has proposed a new measure of functional complexity [10,54] that is based both on function and on information. For a particular function x, let E_x represent the degree of that function achieved by a system. Then the *functional information* is defined as [54]

$$I(E_x) = -\log(F(E_x))\,, \tag{40}$$

where $F(E_x)$ is the fraction of all possible configurations of the system that possess a degree larger or equal to E_x. For sequences, the function could represent a binding affinity, or the number of ATPs produced by a pathway within which the enzyme is the bottleneck factor, or any other real-valued attribute that characterizes the performance of the sequence. This measure introduces a clear link between information and function, but fundamentally turns out to be a coarse-grained version of the information content Eq. (31), as can be seen as follows.

Suppose we are interested in measuring the information content of a sequence s that performs function x to the degree E_x. We can obtain the functional information of s by creating all possible mutants of s, and measuring the fraction $F(E_x)$ of sequences that have the same function as s, given by $\nu(s)/N$, where $\nu(s)$ is the number of neutral mutants of s within S, and N is the total number of possible sequences. Thus,

$$I(E_x) = \log N - \log \nu(s)\,. \tag{41}$$

The conditional probability to find a sequence s_i given environment e in an evolving populations of sequences of the type s is given by $p(s_i|e)$. If e specifies the function x at level E_x for s_i, then $p(s_i|e) = 1$ if s_i performs the function at the required level, and zero otherwise (coarse-graining of the entropy). There are $\nu(s)$ such sequences in the ensemble, and thus

$$H(S|e) = -\sum_i p(s_i|e) \log p(s_i|e)$$
$$= -\sum_{\nu(s)} \frac{1}{\nu(s)} \log \frac{1}{\nu(s)} = \log \nu(s) \,. \qquad (42)$$

As $\log N = \log D^L = L$, Eq. (41) recovers Eq. (31), that is, functional information is a coarse-grained version of the Shannon information content of a sequence.

Network Complexity

If we are given the informational complexity of every single protein in an organism, is the sum of these complexities equal to the complexity of the cell, or the organism? Certainly not, because much of the complexity of an organism lies in how the proteins interact, and in particular in the complicated temporal sequence of events dictated by the regulation of the expression of proteins as a function of environmental influences.

It is well-known that functional biological networks have properties that distinguish them from a randomly connected set of nodes (a random graph, or Erdös–Rényi network). In fact, many biological networks have degree distributions that are approximately scale-free (Fig. 7a), in stark contrast to the binomial distribution of the random graph (Fig. 7b). Also, the diameter of biological networks, defined as the average of all internode distances $d(i, j)$, is small ("small-world-network", see [55]) and depends only weakly on the number of nodes in the network, again unlike what we find for random graphs.

Assessing the complexity of a network usually entails measuring the structural complexity of the network, as compared to a random graph. Often, the *modularity* of a network is used as a proxy for the structural or the functional complexity, even though the concept of a module for networks is not universally defined [56,57]. Usually, a module is defined as a discrete entity whose function is separable from those of other modules, but in biology modules can have significant overlap. Within protein-protein interaction networks, putative modules can be obtained by clustering, so that modules are sets of proteins that are strongly interconnected, but only weakly connected to other such sets. In this section I discuss several ways to understand network complexity, starting with an estimate of the information content of a functional network by assessing the information content of the genome that gave rise to it, using the methods discussed in Sect. "Molecular Complexity". I then review an information-theoretic approach to network modularity, followed by an exposition of a method to assess network structure and information by measuring subnetwork (subgraph) abundances.

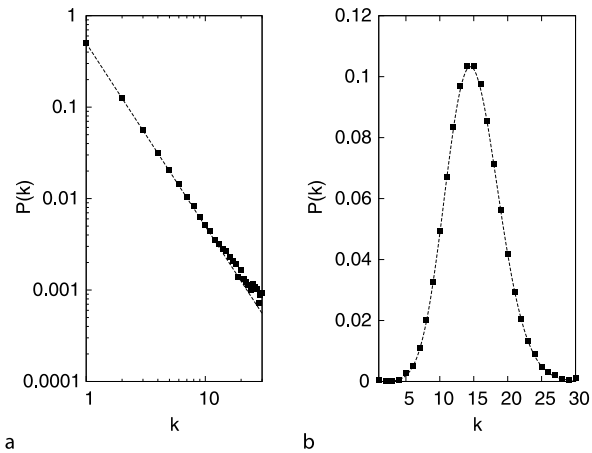

Biological Complexity and Biochemical Information, Figure 7
a Degree distribution $P(k)$ to find a node with k edges in a scale free network (squares) with degree distribution $P(k) \sim 1/k^2$ (*dashed line*), b degree distribution for a random graph with 1,000 nodes and a connectivity $p = 0.015$ (*squares*) and a Poisson distribution with mean $\langle k \rangle = 15$ (*line*)

Evolution of Information in Networks

The difficulty in assessing network complexity from network topology is clear to anyone who has studied the multitudes of networks arising in engineering and biology. Biological networks usually have thousands of nodes and several thousand edges, and often appear to be unstructured. For example, the network summarizing the connectivity of neurons in the brain of the nematode *C. elegans* shows very little modularity or structure at first sight, but is markedly different from a random network [58] (see Fig. 8).

Because the functionality of a network is not necessarily reflected in its topological structure, the best hope for assessing the complexity of a network is to measure the complexity of the set of rules used to construct it. In biology, this set of rules is encoded in the genome, so a first-order estimate of the complexity of a network should be given by the complexity of the genome that produced it. Of course, this is difficult for all the reasons given in the

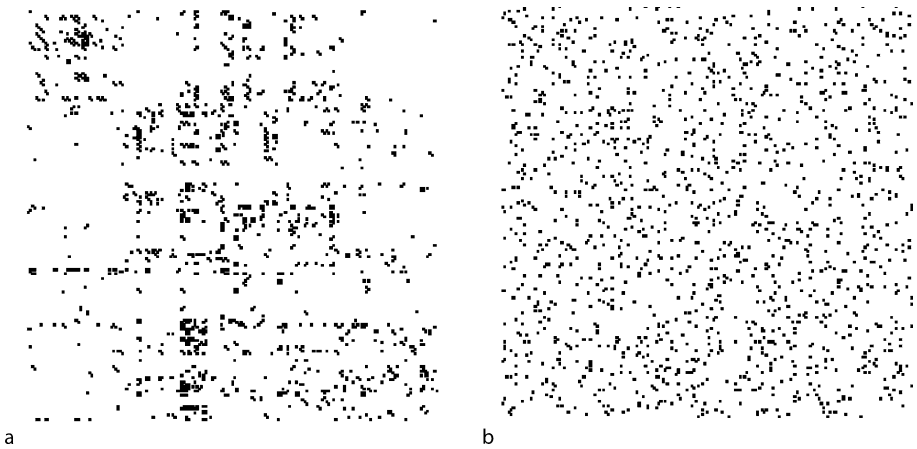

Biological Complexity and Biochemical Information, Figure 8
Adjacency matrix of 179 of the 302-neuron neural network of a *C. elegans* brain (*left*), and a random network of the same size and connectivity (*right*)

previous section, but even more difficult in this case because a network of proteins, for example, is specified not just by the open reading frames coding for the proteins, but also all the untranslated regulatory regions as well as the transcription factors affecting them.

We can test the evolution of network complexity in computational models where a genome represents the functionality of a cellular network, as was done recently in Ref. [59]. In this work, an artificial chemistry and genetics was encoded in a simple linear (circular) code based on the monomers `0,1,2,3`, where enzymatic proteins with variable specificity act on 53 precursor molecules to form up to 555 metabolites. The metabolic reactions involving transport and enzymatic proteins are obtained by a translation of the genetic code into reactions, and implementing chemostat physics and reaction kinetics. Evolution proceeded from a simple ancestral genome with only 3 genes to large and complex metabolic networks of thousands of nodes and several thousand edges, in a completely asexual Wright–Fisher process acting on two chromosomes.

In order to be considered fit, an artificial cell has to import precursor molecules that are available outside the cell walls and convert them into metabolites. The fitness of an organism was determined by calculating the produced biomass of metabolites (see [59]). Evolution was carried out in three different environments that differ in their predictability. In the simplest environment, the location of precursor sources and their abundance is constant during evolution (the "static" environment), while in the quasistatic environment one randomly selected precursor source location is changed per update. In the dynamic environment, the source location of all precursors is changed randomly, and 25% of all precursors are made unavailable, giving rise to a highly unpredictable environment.

The information content of the genomes was measured as outlined above, that is

$$I = L - H(s) \,, \tag{43}$$

where L is the total length of the sequence and $H(s)$ is the sum of per-site entropies

$$H(s) = \sum_{x=1}^{L} H(x) \tag{44}$$

and the per-site entropy $H(x)$ is obtained by summing over the substitution probabilities p_i at that site:

$$H(x) = -\sum_{i=0}^{3} p_i \log_4 p_i \,. \tag{45}$$

Because we only have 4 possible symbols per site, taking the logarithm to base 4 again ensures that the per-site entropy lies between 0 and 1.

As the evolutionary mechanism allows for insertions and deletion of entire genes or genetic regions along with point mutations, genomes can change length during evolution. As a consequence, an alignment of genomes in a population to ascertain substitution probabilities is problematic. Instead, an approach can be used that determines the substitution probabilities p_i from the fitness effect of the substitution on organism fitness, along with an application of population genetics theory. If a substitution of allele i has fitness effect w_i, then the probability to find this allele in an equilibrated population evolving at mutation

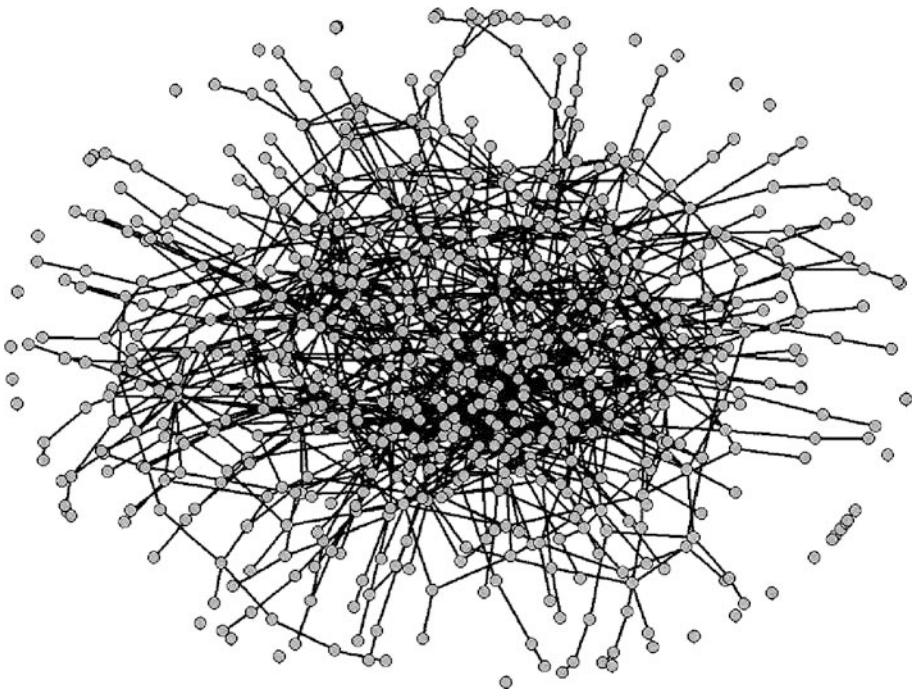

Biological Complexity and Biochemical Information, Figure 9
Evolved metabolic network with 969 nodes and 1,698 edges, rendered with PAJEK [60]

rate μ is given by [61]

$$p_i = \frac{p_i w_i}{\bar{w}}(1-\mu) + \frac{\mu}{4}\sum_{j=0}^{3}\frac{p_j w_j}{\bar{w}}\,, \tag{46}$$

where $\bar{w} = \sum_{i=0}^{3} p_i w_i$ is the mean fitness of the 4 possible alleles at that position.

Figure 10 shows the evolution of network fitness, size and complexity (measured in units of monomer entropy or "mer", where one mer equals 2 bits) for the three different environments discussed above), as a function of the phylogenetic depth of the organism. The phylogenetic depth of an organism is its position on the line of descent of the particular evolutionary run: The organism with the highest fitness at the end of the run is used to reconstruct the line of descent by following its direct ancestry and increasing the depth counter whenever an organism's genome differs from that of its direct parent. When arriving at the initial organism, this counter is set to zero, so that the phylogenetic depth counter increases up until the last organism on the line. Because on average we find about one new organism on the line of descent per generation in these runs, the phylogenetic depth is a good proxy for evolutionary time, even though in principle many generations could pass without an advance on the line of descent.

Because the fitness of an organism is multiplicative in the biomass (discovering how to produce a new metabolite multiplies the previous fitness by a number greater than one), the log of the fitness grows about linearly for all three environments (Fig. 10a). The fitness grows fastest for the static environment that is the easiest to predict (dotted line), while it takes longer for complexity to emerge in the dynamic environment (solid line). The same trend is reflected in the growth of the number of nodes and edges in these environments (Fig. 10b). Finally, the information content as calculated by Eq. (43) using the substitution probabilities Eq. (46) follows the same trend: the informational complexity grows the fastest for the static and quasi-static environments, and lags behind for evolution in a dynamic environment. The reason for the slower growth of complexity for networks evolving in dynamic environments is clear: because the availability of precursors necessary for the production of complex metabolites cannot be relied upon in such environments, the cells end up manufacturing the precursor molecules within the cells (rather than importing them from the outside). This machinery is complex in itself, but takes time to evolve. Ultimately, we expect networks evolving in dynamic environments to be more complex than those evolving in static environments because of the added flexibility of producing precursor molecules within the cells. However, such net-

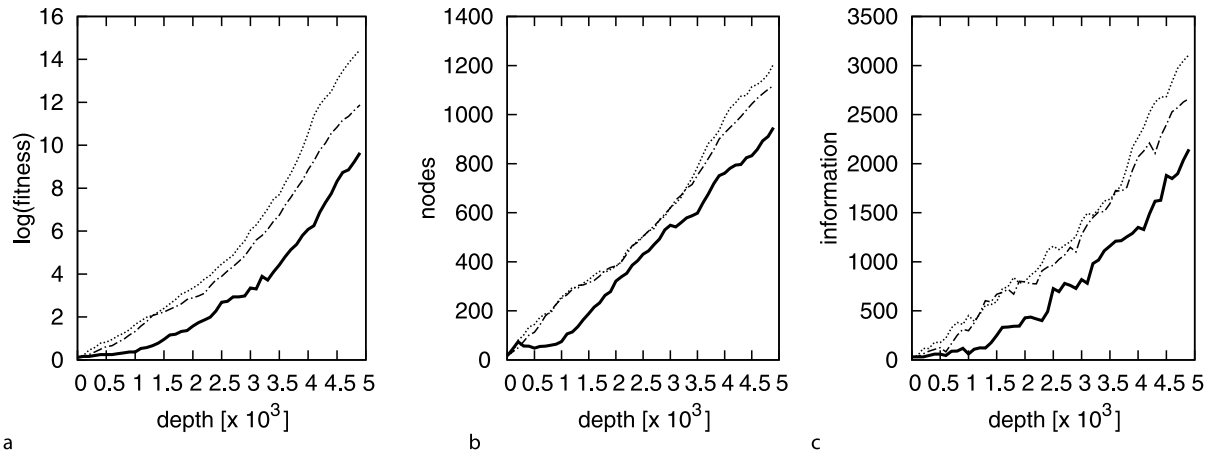

Biological Complexity and Biochemical Information, Figure 10
Evolution of complexity in artificial metabolic networks. a log fitness for networks evolving in a static (*dotted line*), quasistatic (dash-dotted), and dynamic environment (*solid line*). b Evolution of the number of nodes (number of edges follows a similar trend). c Evolution of informational complexity, lines as in a

works lag behind slightly during the time this complexity is generated.

Note that the informational complexity of the networks used to seed these evolutionary experiments is rather low: the network with three genes (two for importing precursors and one for metabolizing those) is specified with a genome of informational complexity of just 36 mers (72 bits), even though the starting genome has 1,000 "nucleotide" positions in each of the two chromosomes. The non-coding part of these initial genomes thus does not contribute to the informational complexity, because changing any of these positions to any other allele cannot change the fitness of the organism (we do not take beneficial mutations into account in the fitness tests). For these non-coding nucleotides, the p_i calculated by Eq. (46) all are exactly equal ($p_i = 1/4$), guaranteeing that they do not contribute to I, as is easily checked. However, the informational complexity grows rather quickly once more and more metabolic reactions are discovered and optimized, at a pace of about 0.5 mers (1 bit) per depth step (roughly one bit per generation).

Modules from Information Theory

The path towards an understanding of the functional organization of a network in the absence of genomic information usually involves the *decomplexification* of the network, either by clustering nodes that are related in function [62], removing those nodes that are immaterial to (or redundant in) function, or analyzing the subgraph decomposition [63] as discussed further below. A particularly insightful method to decompose networks into modules—both overlapping and non-overlapping—uses information theory to estimate how much information about the original network is present in the abstract—that is, decomplexified— version, while maximizing a variable that measures the *relevance* of the abstraction. This method, sometimes called the "information-bottleneck" approach [64] was applied to biological and engineering networks by Ziv et al. [65].

The *Network Information Bottleneck* (NIB) approach attempts to replace a complex network by a simpler one while still retaining the essential aspects of the network. For example, a highly connected star topology could be replaced by a single node that represents the modular function of the star, as in Fig. 11.

The main idea of the method is that while there are many different ways in which one can collapse a topology, the optimal mapping is one where the new topology retains as much information as possible about the original

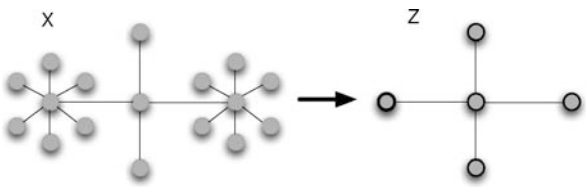

Biological Complexity and Biochemical Information, Figure 11
Collapse of a topology described by the nodes *X* to a more succinct one described by *Z* in which clusters are replaced by a cluster assignment variable *Z*

one, while maintaining as much *relevance* of the description as possible. Say, for example, that a random variable X stands for nodes $x \in X$ in a network that occur with probability $p(x)$, and total number of states $|X| = N$, where N is the size of the network. A *model* of this network can then be made by a random variable Z with *fewer* states $|Z| < N$, and where a cluster assignment is given by a set of probabilities $p(z|x)$: the probability that z is assigned to a particular cluster given the input node x. Ideally, we would like to maximize the mutual entropy (information) between the random variables X and Z, but we shall do this with a constraint given by a relevance variable mentioned earlier. This relevance variable will distinguish different ways in which clusters are assigned. In this application of the NIB to networks, the relevance variable involves diffusion on the network: those nodes that are close to each other are preferentially visited by a diffusion algorithm, and are more likely to be clustered together. In this algorithm, the relevance is represented by a random variable Y that is defined such that a diffusive process determines the probability to arrive at node y. The relation to the network variable X is given by the joint probability $p(x, y) = p(y|x)p(x)$, the probability to arrive at node y via a diffusive process given the process started at node x, times the probability that we started at node x. The latter probability is always assumed to be uniform, that is, $p(x) = 1/N$.

The NIB algorithm allows optimal solutions where several different nodes z could be assigned to any given input node x. This flexibility allows for the possibility of *overlapping* clusters or modules (soft clustering), but in the following we will follow only the algorithm for hard clustering, so that for any choice of x and z, $p(z|x)$ is either one or zero.

One version of the algorithm (agglomerative clustering) begins with a random variable Z that has exactly one fewer nodes than X, and attempts to find the optimal pair of x nodes to join to produce a model of the network with one fewer nodes. For each possible cluster assignment $p(z|x)$ we can execute a diffusion process to determine the matrix $p(y|z)$, that is, the probability to arrive at node y given a node z as starting point. We choose to merge those nodes that maximize $I(Y : Z)$, and the algorithm repeats with a set Z smaller by one node until all nodes have been merged. At each step, we can calculate the normalized variables $0 < I(Z : X)/H(X) < 1$ and $0 < I(Z : Y)/I(X : Y) < 1$ and plot them against each other, giving rise to the *information curve* [65], as in Fig. 12.

A completely random network gives rise to the diagonal in Fig. 12, and represents the least modular network. We can define a modularity score, the *network modularity*,

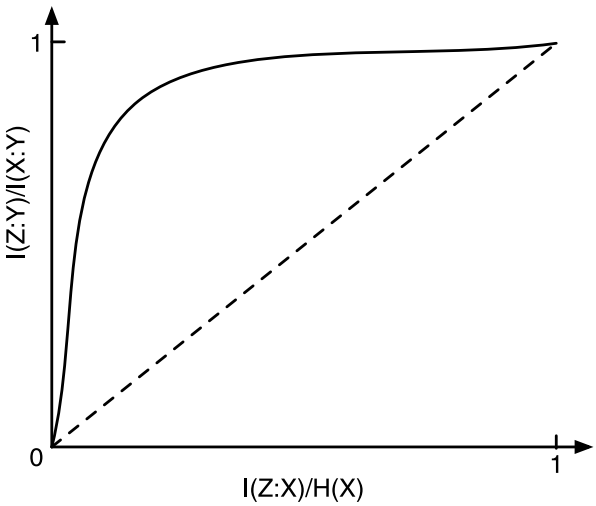

Biological Complexity and Biochemical Information, Figure 12
The information curve for a modular network (*solid line*), obtained by starting with a model network Z of the the same size as X (maximal information $I(Z : X)/H(X) = 1$ and $I(Z : Y)/I(X : Y) = 1$, *upper right corner*), and merging nodes while maximizing $I(Z : Y)$. This process generates the information curve from the *upper right corner* all the way down to the *lower left corner*, where $|Z| = 1$ and the mutual entropy vanishes. The dashed line represents the information curve for a random network. The modularity score is given by the area under the information curve

as the area under the information curve. Perfectly modular networks then have a maximal modularity score of 1, whereas random networks have a score of 1/2. Several different networks have been analyzed using this modularity measure in Ref. [65], such as the network of coauthors for papers presented at a meeting of the American Physical Society. This network has 5,604 nodes and 19,761 edges, and yielded a modularity core of 0.9775. The regulatory network of the bacterium *E. coli* (328 nodes and 456 edges), also analyzed by these authors, yielded a score of 0.9709. Thus, both of these networks are highly modular according to this measure.

Interestingly, the network of connections of the *C. elegans* brain (see Fig. 8) has a network modularity score 0.9027, whereas a randomized version retaining the same number of nodes and edges scores 0.4984 on average, as predicted for a random network. The evolved metabolic networks discussed in Sect. "Evolution of Information in Networks" also score high on this scale. For the largest connected component of a 453 node network, we find a modularity score of 0.8486 for the 5,000*th* organism on the line of descent (about one organism per generation). While a score for small networks is fairly meaningless, the score increases slightly as the networks become more complex.

Information in Motifs

We have seen that networks that are functional are built from modules, or at least can be understood in terms of strongly connected sets of nodes that are only weakly connected to other such clusters. This clustering—carried out using information theory in the previous section—can also be performed on the basis of topology alone. For example, every network can be analyzed in terms of its *subgraph composition* [63], that is, the frequency with which particular subgraphs or motifs appear within the entire network. The degree with which certain motifs are overutilized—and some others underutilized—compared to a uniform distribution or one obtained from a random network, reflects the *local structure* of the network and can be used to classify networks from very different realms into similar categories [66].

Subgraph abundances can also be used to study the modular composition of networks, in analogy to the modular composition of sentences in written language. Meaning can be conveyed in text only because the utilization frequency of letters in words, and words in sentences, is different from uniform. For example, the letters e,t,a,i,o, and n appear in decreasing frequency in an average text written in English, while the rank-abundance distribution of the words follows a scale-free distribution (Zipf's Law [67,68]). If we assume that a random sequence of letters contains no information, then the deviation from the uniform distribution could be used to distinguish and perhaps classify functional (that is, meaningful) text from gibberish. In the same vein, it is possible that functional networks differ significantly from random networks in the subgraph utilization, and we can study this difference by estimating the *subgraph information content* as follows.

Suppose we compute the probability to find any of the two possible motifs that can be made of three nodes (see Fig. 13). For simplicity, we are considering here only undirected graphs, and do not allow self-interactions, that is, nodes that link to themselves. We can then compare these empirical probabilities to the probabilities with which these subgraphs appear in a random network.

A priori, we might think that any of the two motifs of size $n = 3$ should appear with equal probability in a random network, giving rise to a motif entropy that is maximal:

$$H_3(\tfrac{1}{2}, \tfrac{1}{2}) = -\tfrac{1}{2}\log_2 \tfrac{1}{2} - \tfrac{1}{2}\log \tfrac{1}{2} = 1\,. \tag{47}$$

Here, we defined the size-n motif entropy

$$H_n(p_1, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log_m p_i\,, \tag{48}$$

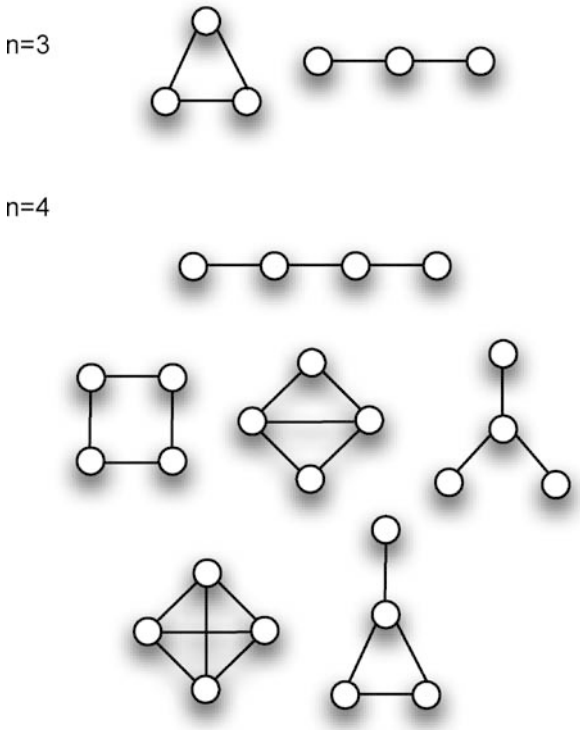

Biological Complexity and Biochemical Information, Figure 13
Undirected motifs of size $n = 3$ and $n = 4$, without self-interaction

where m is the number of possible connected motifs of size n, and the p_i are the probabilities to find the ith motif in the network. (Because the base of the logarithm is also m, this entropy is normalized to lie between zero and one.) The information stored within $n = 3$-motifs would then be (the superscript (u) refers to the uniform baseline distribution)

$$I_3^{(\mathrm{u})} = H_3(\tfrac{1}{2}, \tfrac{1}{2}) - H_3(p_1, p_2) = 1 - H_3(p_1, p_2)\,, \tag{49}$$

while the information stored in n-motifs is naturally

$$\begin{aligned} I_n^{(\mathrm{u})} &= H_n\left(\frac{1}{m}, \ldots, \frac{1}{m}\right) - H_n(p_1, \ldots, p_m) \\ &= 1 - H_n(p_1, \ldots, p_m)\,. \end{aligned} \tag{50}$$

However, even random networks do not have a uniform distribution of motifs, and we can instead consider the information stored in motifs compared to a random network as baseline, as (for $n = 3$)

$$I_3^{(\mathrm{r})} = H_3(p_1^{(\mathrm{r})}, p_2^{(\mathrm{r})}) - H_3(p_1, p_2)\,. \tag{51}$$

where the $p_i^{(\mathrm{r})}$ refer to the probability of finding motif i in a random network.

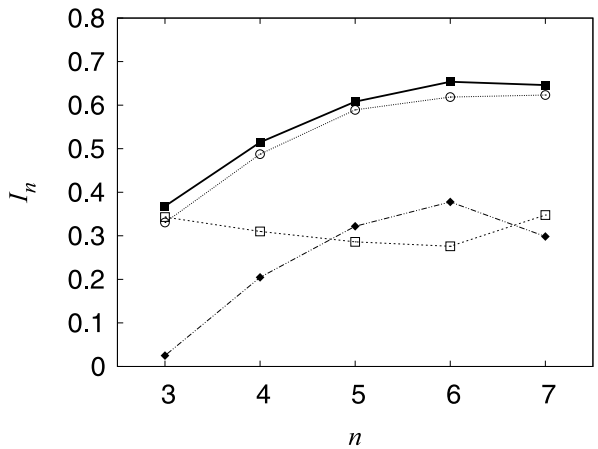

Biological Complexity and Biochemical Information, Figure 14
Motif information I_n for motifs of size n for evolved networks with a uniform distribution as a baseline (*solid line, filled squares*), an Erdös–Rényi network of the same size and number of edges and connectivity $p \approx 0.005$ (*dashed line, open squares*), and a randomized network of the same size with a scale-free edge distribution (*dotted line, open circles*). The motif information using the probability distribution of a random network as the baseline Eq. ($I_n^{(r)}$) is the difference between the solid and dashed lines (*dash-dotted line, filled diamonds*)

In Fig. 14, we see the information $I_n^{(u)}$ stored in motifs of size n for $n = 3 - 7$ (filled squares, solid line), for a single, evolved, functional, simulated metabolic network discussed in Sect. "Evolution of Information in Networks", of 598 nodes. The network information increases as more complex motifs (larger n) are used for encoding, but appears to stabilize. This behavior mirrors the statistics of n-gram entropies in English text, as noted early on by Shannon [68]. Note that because the shortest path between any two nodes is on average of the order of 4-5 in these networks [59], motifs of size 7 or larger are not well-sampled.

We can study whether the network information is dictated by functionality, edge distribution, or both, by constructing analogous networks that have the functionality removed by randomizing connections but keeping the scale-free edge distribution, and by randomizing the network but destroying also the edge distribution. If, for example, we randomize the connections in our functional evolved network while keeping the scale-free edge distribution, we find that the network information is only slightly lowered (open circles and dotted line in Fig. 14). On the other hand, if we randomize the network in such a way that the degree distribution is that of a random graph (but still keeping the same number of nodes and edges), the dependence of the network information as a function of the subgraph size is markedly different (open squares, dashed line in Fig. 14). This suggests that the network information is significantly dictated by the biological scale-free distribution of edges per nodes, and only weakly by the actual function of the network.

Because random graphs do not utilize subgraphs with equal probability (as witnessed by the non-zero network information in Erdös–Rényi networks), it is more appropriate to use the random network probabilities as the baseline to calculate the network information. In this case, we see the network information $I_n^{(r)}$ increase from small values at $n = 3$ up to $n = 6$ (dash-dotted line, filled diamonds in Fig. 14). The decrease noted for $n = 7$ is due to incomplete sampling of size $n = 7$ networks for this small graph (598 nodes), and is not significant.

In summary, network information as measured by subgraph "n-gram" entropies behaves similar to the dependence of n-gram entropies in written language, and can be used to distinguish functional networks from random ones. However, the network information appears to be controlled mostly by the form of the edge distribution. Insight into the modular structure of networks will likely depend on understanding how the subgraphs of networks are assembled into modules, in analogy to how letters are assembled into words in written text.

Future Directions

Several billions of years of evolution have shaped our biosphere to become the complicated, interdependent, hierarchical complex system we witness today. Among the parts and members of this system, there are certainly differences in complexity—some obvious to any observer—some less so. Structural complexity, the most intuitive of all measures, is notoriously difficult to define because there is no universal system to rank all possible physical structures. We have seen that automata theory and information theory allow us to quantify the complexity of a sequence in terms of its information content about the environment within which it has evolved, and that this information is a good proxy for the functional complexity of the organism precisely because the information is used by the organism to function in a complex environment.

But the measure is limited because so far it is only practical for short stretches of DNA or single proteins. Thus, a quantitative measure for the complexity of a whole genome using information theory will only be possible when a large number of complete genomes of closely related species is available. Another shortcoming of the informational complexity is that it refers to a particular niche only, and furthermore cannot quantify the complexity of genes that are adapted to varying environments. So, for example, so far we cannot use the informational complex-

ity to estimate the complexity of an ecosystem, nor of an organism that spends part of its lifecycle in one environment, and another part in a completely different one (simple forms of life such as arboviruses are notorious for such a cycle). However, a natural extension of the informational complexity exists that may cover multiple environments both in time and space. Recall that the informational complexity of an ensemble of sequences S refers to a single environment description $E = e$:

$$I(S : e) = L - H(S|e) \,, \tag{52}$$

where $H(S|e)$ is the ensemble entropy of the sequences. We can generalize this expression by promoting the environment to a true random variable E that can take on states e_i with probability $p(e_i)$. This formalism can describe environments that are composed of different (spatially separated or overlapping) niches e_i, as well as environments that take on the different states e_i periodically or even randomly, in time. The informational complexity then becomes

$$\begin{aligned} I(S : e) \to \sum_e p(e) I(S : e) \\ = I(S : E) = L - H(S|E) \,. \end{aligned} \tag{53}$$

Here, $H(S|E)$ is the average conditional entropy of the sequence ensemble S given the environment E. Whether this construction will turn out to be useful for characterizing the complexity of ecosystems or variable environments remains to be seen, as the practical obstacles are only amplified by having to measure the informational complexity in multiple environments. But at the very least this construction addresses one of the fundamental problems in assessing functional complexity that we encountered in the introduction, namely that for organisms that have adapted to be functional in a variety of environments, we can find genes that appear to show no phenotype upon knockout in the laboratory. Such genes, however, may very well show a phenotype in a particular environment that the organism encounters in the wild, and this functional capacity of an organism needs to be taken into account when assessing functional complexity. If the random variable E accounts for the multitude of environments with the correct probabilities $p(e)$, then the full functional complexity of the organism may be characterized using Eq. (53). But to apply this measure, more efforts need to be expended towards understanding the modes in which an organism functions in its native environment(s) (as opposed to the unnatural laboratory conditions that are the norm today). If such an effort is made, then as we can expect an exponential increase in sequence data in the coming years, the prospects for a general understanding of biological complexity in terms of sequence complexity are good.

Acknowledgments

I am grateful to Arend Hintze for the collaborative work in Sect. "Network Complexity", as well as numerous discussions. I am also indebted to Matthew Rupp for the analysis shown in Figs. 5 and 6. This work was supported in part by the National Science Foundations Frontiers in Integrative Biological Research grant FIBR-0527023, a Templeton Foundation research grant, and DARPA's FunBio initiative.

Bibliography

1. Lovejoy AO (1936) The great chain of being: A study of the history of the idea. Harvard University Press, Cambridge
2. Gould S (1996) Full house: The spread of excellence from Plato to Darwin. Harmony Books, New York
3. Nee S (2005) The great chain of being. Nature 435:429
4. Gould S, Lewontin R (1979) The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. Proc R Soc London B 205:581–598
5. McShea DW (1996) Metazoan complexity and evolution: Is there a trend? Evolution 50:477–492
6. Valentine J, Collins A, Meyer C (1994) Morphological complexity increase in metazoans. Paleobiology, 20:131–142
7. Bell G, Mooers A (1997) Size and complexity among multicellular organisms. Biol J Linnean Soc 60:345–363
8. Nehaniv CL, Rhodes JL (2000) The evolution and understanding of hierarchical complexity in biology from an algebraic perspective. Artif Life 6:45–67
9. McShea D (2001) The hierarchical structure of organisms: A scale and documentation of a trend in the maximum. Paleobiology 27:405–423
10. Szostak JW (2003) Functional information: Molecular messages. Nature 423:689
11. McShea DW (2000) Functional complexity in organisms: Parts as proxies. Biol Philosoph 15:641–668
12. Britten RJ, Davidson EH (1971) Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. Q Rev Biol 46:111–138
13. Cavalier-Smith T (1985) Eukaryotic gene numbers, non-coding DNA and genome size. In: Cavalier-Smith T (ed) The evolution of genome size. Wiley, New York, pp. 69–103
14. Gregory TR (2004) Macroevolution, hierarchy theory, and the c-value enigma. Paleobiology 30:179–202
15. Gregory TR (2005) Genome size evolution in animals. In: Gregory TR (ed) The evolution of the genome. Elsevier, San Diego, pp. 3–87
16. Badii R, Politi A (1997) Complexity: Hierarchical structures and scaling in physics, Cambridge Nonlinear Science Series, vol. 6. Cambridge University Press, Cambridge (UK)
17. Kolmogorov A (1965) Three approaches to the quantitative definition of information. Probl Inf Transm 1:4
18. Li M, Vitanyi P (1997) An introduction to Kolmogorov complexity and its applications. Springer, New York
19. Adami C, Cerf NJ (2000) Physical complexity of symbolic sequences. Physica D 137:62–69

20. Gell-Mann M, Lloyd S (1996) Information measures, effective complexity, and total information. Complexity 2:44–52
21. Shannon C, Weaver W (1949) The mathematical theory of communication. University of Illinois Press, Urbana
22. Quastler H (ed) (1953) Information theory in biology. University of Illinois Press, Urbana
23. Gatlin L (1972) Information theory and the living system. Columbia University Press, New York
24. Mantegna RN, Buldyrev SV, Goldberger AL, Havlin S, Peng CK, et al (1994) Linguistic features of noncoding DNA sequences. Phys Rev Lett, 73:3169–3172
25. Schmitt AO, Herzel H (1997) Estimating the entropy of DNA sequences. J Theor Biol 188:369–377
26. Weiss O, Jimenez-Montaño MA, Herzel H (2000) Information content of protein sequences. J theor Biol, 206:379–386
27. Herzel H, Ebeling W, Schmitt AO (1994) Entropy of biosequences: The role of repeats. Phys Rev E 50:5061–5071
28. Shannon C (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423, 623–656
29. MacKay DJC (2002) Information theory, inference and learning algorithms. Cambridge University Press, Cambridge
30. Adami C (2004) Information theory in molecular biology. Phys Life Rev 1:3–22
31. Grassberger P (1986) Toward a quantitative theory of self-generated complexity. Int J Theor Phys 25:907–938
32. Bernaola-Galvan P, Roman-Roldan R, Oliver J (1996) Compositional segmentation and long-range fractal correlations in DNA sequences. Phys Rev E 53:5181–5189
33. Sprinzl M, Horn C, Brown M, Ioudovitch A, Steinberg S (1998) Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res, 26:148–153
34. Eddy SR, Durbin R (1994) RNA sequence analysis using covariance models. Nucl Acids Res, 22:2079–2088
35. Korber BT, Farber RM, Wolpert DH, Lapedes AS (1993) Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc Natl Acad Sci USA 90:7176–7180
36. Clarke ND (1995) Covariation of residues in the homeodomain sequence family. Protein Sci, 4:2269–2278
37. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW (2000) Correlations among amino acid sites in bhlh protein domains: an information theoretic analysis. Mol Biol Evol 17:164–178
38. Wang LY (2005) Covariation analysis of local amino acid sequences in recurrent protein local structures. J Bioinform Comput Biol 3:1391–1409
39. Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK (eds) Molecular systematic, 2nd edn, Sinauer, Sunderland, pp. 407–514
40. Wolf JB, Brodie III ED, Wade MJ (eds) (2000) Epistasis and the evolutionary process. Oxford University Press, Oxford
41. Bridgham JT, Carroll SM, Thornton JW (2006) Evolution of hormone-receptor complexity by molecular exploitation. Science 312:97–101
42. Cowperthwaite MC, Bull JJ, Ancel Meyers L (2006) From bad to good: Fitness reversals and the ascent of deleterious mutations. PLoS Comput Biol 2:e141
43. Finn RD et al (2006) Pfam: Clans, web tools and services. Nucleic Acids Res 34:D247–D251
44. Brenner SE, Chothia C, Hubbard TJP (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci USA, 95:6073–6078
45. Miller GA, Madow WG (1954) On the maximum likelihood estimate of the Shannon-Wiener measure of information. Technical Report 54–75, Air Force Cambridge Research Center, Bedford
46. Basharin GP (1959) On a statistical estimate for the entropy of a sequence of independent random variables. Theory Probab Appl 4:333–337
47. Zurek WH (1990) Algorithmic information content, Church-Turing thesis, physical entropy, and Maxwell's demon. In: Zurek WH (ed) Complexity, entropy, and the physics of information. SFI Studies in the Sciences of Complexity, vol. 8 Addison-Wesley. Redwood City pp. 73–89
48. Cover TM, Thomas JA (1991) Elements of Information Theory. John Wiley, New York
49. Adami C (1998) Introduction to Artificial Life. Springer, New York
50. Adami C (2006) Digital genetics: Unravelling the genetic basis of evolution. Nat Rev Genet 7:109–118
51. Adami C, Ofria C, Collier T (1999) Evolution of biological complexity. Proc Natl Acad Sci USA 97:4463–4468
52. Ofria C, Huang W, Torng E (2008) On the gradual evolution of complexity and the sudden emergence of complex features. Artif Life 14, to appear
53. Carothers JM, Oestreich SC, Davis JH, Szostak JW (2004) Informational complexity and functional activity of RNA structures. J Amer Chem Soc, 126:5130–5137
54. Hazen RM, Griffin PL, Carothers JM, Szostak JW (2007) Functional information and the emergence of biocomplexity. Proc Natl Acad Sci USA 104:8574–8581
55. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5:101–113
56. Schlosser G, Wagner GP (eds) (2004) Modularity in development and evolution. University of Chicago Press, Chicago, IL
57. Callebaut W, Rasskin-Gutman D (eds) (2005) Modularity: Understanding the development and evolution of natural complex systems. MIT Press, Cambridge, Mass
58. Reigl M, Alon U, Chklovskii DB (2004) Search for computational modules in the C. elegans brain. BMC Biol 2:25
59. Hintze A, Adami C (2008) Evolution of complex modular biological networks. PLoS Comput Biol 4:e23
60. Batagelj V, Mrvar A (2003) Pajek: Analysis and visulaization of large networks. In: M Jünger PM (ed) Graph Drawing Software. Springer, Berlin, pp. 77–103
61. Huang W, Ofria C, Torng E (2004) Measuring biological complexity in digital organisms. In: Pollack J, Bedau MA, Husbands P, Ikegami T, Watson R (eds) Proceedings of Artificial Life IX, MIT Press, Cambridge, pp. 315–321
62. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442
63. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D et al. (2002) Network motifs: simple building blocks of complex networks. Science 298:824–827
64. Tishby N, Pereira F, Bialek W (1999) The information bottleneck method. In: Hajek B, Sreenivas RS (eds) Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing, University of Illinois Press, Champaign, IL, pp. 368–377
65. Ziv E, Middendorf M, Wiggins CH (2005) Information-theoretic approach to network modularity. Phys Rev E 71:046117

66. Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S et al. (2004) Superfamilies of evolved and designed networks. Science, 303:1538–1542
67. Zipf GK (1935) The psycho-biology of languages. Houghton-Mifflin, Boston
68. Shannon CE (1951) Prediction and entropy of printed English. Bell System Tech J 30:50–64

Biological Data Integration and Model Building

James A. Eddy[1], Nathan D. Price[1,2,3]
[1] Department of Bioengineering, University of Illinois, Urbana-Champaign, USA
[2] Department of Chemical and Biomolecular Engineering, University of Illinois, Urbana-Champaign, USA
[3] Institute for Genomic Biology, University of Illinois, Urbana-Champaign, USA

Article Outline

Glossary
Definition of the Subject
Introduction
The Challenge of Heterogeneous Data Types in Biology
Integration Through Interaction Network Representations
Model Building as Informative Data Integration
Data Integration Using Influence Network Models
Data Integration Using Biochemical Network Models
Comparison of Biochemical and Statistical Network Models for Data Integration
Future Directions
Bibliography

Glossary

Constraint-based analysis A modeling framework based on excluding infeasible network states via environmental, physicochemical, and regulatory constraints to improve predictions of achievable cellular states and behavior.

Data space Multidimensional space containing all possible states of a system; this space can be reduced using defined constraints.

Interaction network A graph where the nodes represent biomolecules (e. g. genes) and the edges represent defined interactions between the nodes, whether they be direct physical interactions (e. g. protein–protein binding, protein–DNA binding) or functional relationships (e. g. synthetic lethality).

Biochemical reaction network Collection of metabolic, signaling, or regulatory chemical reactions described in stoichiometric detail.

Statistical inferrence network A network model designed from statistical inference from large-scale biological data sets to be quantitatively predictive for novel perturbations and/or environmental conditions.

Genome The complete DNA nucleotide sequence in all chromosomes of an organism.

Transcriptome The complete set of RNA transcripts produced from an organism's genome under a particular set of conditions.

Proteome The complete set of expressed proteins produced by the genome.

Metabolome The complete set of small molecules which are the intermediates and products of an organism's metabolism.

Boolean network A set of N discrete-valued variables, $\sigma_1, \sigma_2, \ldots, \sigma_N$ where $\sigma_n \in \{0, 1\}$. To each node a set of k_n nodes, $\sigma_{n_1}, \sigma_{n_2}, \ldots, \sigma_{n_{k_n}}$ is assigned, which controls the value of σ_n through the equation $\sigma_n(t+1) = f_n(\sigma_{n_1}(t), \ldots, \sigma_{n_{k_n}}(t))$. In the case of Boolean networks, the functions f_n can be chosen from the ensemble of all possible Boolean functions.

Definition of the Subject

Data integration and model building have become essential activities in biological research as technological advancements continue to empower the measurement of biological data of increasing diversity and scale. High-throughput technologies provide a wealth of global data sets (e. g. genomics, transcriptomics, proteomics, metabolomics), and the challenge becomes how to integrate this data to maximize the amount of useful biological information that can be extracted. Integrating biological data is important and challenging because of the nature of biology. Biological systems have evolved over the course of billions of years, and in that time biological mechanisms have become very diverse, with molecular machines of intricate detail. Thus, while there are certainly great general scientific principles to be distilled – such as the foundational evolutionary theory – much of biology is found in the details of these evolved systems. This emphasis on the details of systems and the history by which they came into being (i. e. evolution) are distinct features of biology as a science, and influence the need for large-scale data integration. Also, biological systems are responsive to vary-

ing environments, with potential system states influenced by the combinatorics of all possible molecular and environmental perturbations. Thus, data space in this realm is extraordinarily large. There is no shortage of possibilities to explore, and vast amounts of data will be needed to fully reengineer biological systems and understand their workings at a (near) complete level of detail or at a level where accurate prediction can be achieved.

Another reason data integration is essential, is that biology arises through the complex interworking of many components. Thus, to understand biology we must integrate biological data in a way that will allow us, not only to access all our acquired data efficiently, but even more importantly allow us to study, predict, modify, and even engineer the emergent properties of biological systems, where the whole is more than the sum of the parts. As Linus Pauling stated, "Life is a relationship among molecules and not a property of any molecule." Thus data integration and model building to study the emergent properties of biomolecules operating in concert is a pursuit fundamental to the study of biology.

Introduction

With the emergence of full genome sequencing capabilities in the mid-nineties - marked by the first sequencing of a free-living organism, *Haemophilus influenzae* [77] – the stage was set to greatly expand the scope of data integration through computational models in biology. While modeling in biology certainly predates this era (for review, see e. g. [86]), the availability of known genome sequences has enabled the enumeration of the components of an entire cell to become a potentially definable goal. While it is still true that no cell's component 'parts list' is fully defined, the existence of genome sequences allows a much more complete view, and, importantly, an indication of the amount of genomic and gene product information that is yet to be explored. In this way, genome sequencing has catalyzed the reconstruction of genome-scale models, with an eye towards the eventual creation of whole-cell simulators and beyond.

The creation of models of the integrated functions of genes and proteins in cells is of fundamental and immediate importance to the emerging field of systems biology, and represents a powerful means for integrating biological data to aid in understanding emergent properties of biological systems. Some of the most successful attempts at cell-scale modeling to date have been based on reconstructing networks that represent hundreds to thousands of experimentally-determined biochemical interactions, while others have been very successful at inferring statistical networks from large amounts of high-throughput data. These biological networks (e. g. metabolic, regulatory, or signaling) can be analyzed, and predictions about cellular behavior under a variety of conditions and perturbations can be made and tested. Many types of models have been built and applied to study cellular behavior. In this review we focus on two broad types: biochemical network models and statistical inference models. Through iterative model prediction, experimentation, and network refinement, the molecular circuitry and functions of biological networks can be elucidated. The construction of genome-scale models that integrate data about the myriad components that produce cellular behavior is a fundamental goal of systems biology research today.

The Challenge of Heterogeneous Data Types in Biology

Data-space is highly varied in biology for a number of reasons deriving from the nature of biological systems and the heterogeneity of methods that are used to probe different aspects of these systems. To begin thinking about the heterogeneity of biological data, we may first consider the numerous classes of biological molecules in the cell, such as the genome, messenger RNAs, proteins, protein complexes, metabolites and so forth. The global measurement of each of these types of data leads to the study of genomics, transcriptomics, proteomics, metabolomics and so forth. Beyond these biomolecule classifications, heterogeneity in biological data is increased further because of the many forms in which each of the biomolecules in these global data can exist. These modifications give rise to difficulties in characterizing precisely what constitutes a gene or a protein for analysis purposes. For example, a protein can take on many different forms, it may be phosphorylated or not, sites on the protein may be glycosylated or not, it may have different shapes into which it folds depending on its microenvironment, and so forth. In one sense then, one can imagine that integration databases require data on a protein (or a gene) to consist of a packet of information defining both the range of chemical entities defined under the umbrella of the protein (or gene) name, as well as a description of the states in which it is known to be able to exist. Also, the environmental context can greatly affect the values measured for (mostly) all of these data types. Thus, there seem to be confounding factors, uncertainties, and combinatorics at virtually every level of attempting to describe biology – a daunting challenge for data integration moving forward.

Another key factor in defining the heterogeneity of biological data is that the different chemistries of the vari-

ous molecules means that our ability to measure them is often based on very different technologies, with very different capacities for accuracy and global coverage. Also, the history of the development of the technology can also play a role, in that the typical pattern for successful emerging technologies is to increase their capacities logarithmically over time, as has been seen for genomic and transcriptomic data as well as, famously, for computing power (for review of this phenomenon in many fields, see [45]). Thus, factors such as the rate at which these technologies are improving as well as the technical challenges involved in their measurement means that for different data types, we are often at very different stages of our ability to measure them accurately and globally. For example, because of the complementary nature of RNA molecules, we can use template chemistry, as well as amplification of signal using PCR, to enable global measurements. Because of this chemistry, and a history of development, these technologies are now quite mature – i. e. they are available commercially and their accuracy and global coverage are benefiting from the economies of scale and a competitive economic environment. Thus, researchers in the field can currently generate quite accurate data for transcriptomes compared to what is currently possible for proteomics or metabolomics. Genome sequencing has reached such an advanced state that it is becoming possible to sequence individual copies of genomes. Indeed, this progress can be seen in the notable sequencing of the individual genomes of both James Watson and J. Craig Venter [48]. Emerging technologies will enable sequencing capabilities not just from single individuals, but even single cells.

That data space in biology is varied not only because of the nature of biological systems, but also because of variations in our ability to measure different components of biological systems, is a critical issue because what we are really interested in studying is variation in biology. In a sense, the interaction between the biological system and the measurement technology leads to a type of biological 'uncertainty principle' whereby the process of measurement typically involves significant perturbation to the system. Thus, it is important to try to distinguish between variation in biological data that occurs because of measurement technologies and that which occurs because of underlying biology. This task is a fundamental challenge to biological data integration. With any given measurement technology, repeated measurements give rise to estimates of the 'technical variance', while biological replicates are used to study differences due to subtle changes affecting the biology of the organisms. It is thus a more challenging endeavor to develop methods that account for the different accuracies of measuring heterogeneous data while attempting to integrate genomics, proteomics, transcriptomics and their interactions into cohesive networks. One example of a methods arising to meet this challenge is the *Pointillist* software package, which enables the reconstruction of interaction networks using p-value-based methods for integrating various types of biological data [33,34].

Another factor to consider is that, even within each of these different data types, many hierarchies of information are contained. For example, the core information of the genome is typically represented as a linear sequence of base pairs (A,T,C,G), since this sequence does contain a tremendous amount of biological information. Even just within the linear sequence of genomic information are different levels of information. For example, many analyzes focus on quantizing the genome into genes and focusing on the properties of the resulting gene products. However, the position in the linear sequence of the genome effects genes – their functioning is not independent of their position. Of course, all of these molecules exist in three dimensional space, and thus their three dimensional structure and the ways in which those three dimensional structures interact with each other is at the heart of biology. However, most network techniques in use today do not account for three dimensional structure. While there is much work in computational biology focused on three dimensional structures in, for example, the problem of protein folding based on genomic sequence, these endeavors are typically limited in the number of proteins that they can consider. Thus, three-dimensional interaction modeling is not feasible today for the simulation of cells or large networks, leaving such detailed whole cell simulators out of reach, based both on availability of data and on computational power. Simulations of large-scale networks using the three dimensional properties of molecules, including their chemical interactions as well as their biomechanical properties, remains a very large challenge yet to be met in systems biology.

Finally, another critical aspect to biological data integration is that most biological data types do not translate easily from one form to the next. For example, although transcribed genes yield messenger RNAs which are then converted into proteins, the correlation between the measured transcriptome and the measured proteome is typically weak. This fact means that the inferences that can accurately be made about protein concentrations from transcriptome measurements are very limited (although done quite often in practice). This lack of correlation can be understood, given that there is a time delay in the translation of protein from mRNA, that proteins have different degradation rates, and that protein synthesis and degradation are regulated processes. One could in principle model the

relationship between gene expression and protein concentration accurately if one had access to sufficient data (e. g. protein degradation rates, chemical conversion pathways, tRNA binding rates, etc.), but generating the data at the level of detail required for such approaches is currently prohibitive at the genome-scale. Thus, researchers in the field seek to identify shorter and less data-intensive paths to gain the ability to correlate and predict across such heterogeneous data sets.

Integration Through Interaction Network Representations

Technological advancements in the post-genome era have enabled high-throughput screens for protein-protein, protein-DNA, and genetic interactions. The enormous amount of data resulting from these experimental studies has encouraged the development of methods to functionally characterize and visualize the interactions between components using interaction networks. Interaction networks are graphs where the nodes represent biomolecules and the edges represent defined relationships between the nodes. Methods have been developed to integrate interaction network data with complementary high-throughput data, such as from gene expression microarrays, providing insight into the relation of network interactions to cellular states under different conditions. Interaction networks can serve as databases storing a large amount of information about the components in a system and the interactions between them. Interaction network maps and databases generally serve to integrate biological data in a graph, as DNA, transcripts, proteins, and metabolites are organized into networks of direct and indirect interactions. These interaction networks are also useful for visually integrating information from genomics, proteomics, and metabolomics studies. They can be very broadly applied, and are thus used for functional relationships as well as specific binding events. For example, one can define a synthetic lethality network [82], where each node is a gene and an edge indicates that the knockout of the two genes is lethal to the cell whereas each individual knockout is not. In addition to serving as databases and means to allow for visualization of a large amount of data, these interaction networks can be studied to reveal the global topological properties of these networks.

Analysis of large and complex systems biology datasets requires automated principled methods based on robust statistical procedures. One important method that has been developed recently is an integrated framework, named *Pointillist* [33,34], for combining diverse datasets and inferential algorithms, which are incorporated into *Cytoscape* [73] for visualization and simulation. *Cytoscape* is a powerful tool for visualizing and integrating interaction network data with gene expression and other types of state data [38,73]. In addition to its strong visualization capabilities, *Cytoscape* has the potential to be expanded using plug-ins to facilitate a number of analyzes on the topological features of the network, as well as on how high-throughput data maps onto the network. For example, *Cytoscape* implements methods of searching interaction networks for connected sets of differentially expressed genes. These methods also examine which conditions significantly affect gene expression in each set (also referred to as active subnetworks). A statistical scoring system is used to quantify the gene expression change in each active subnetwork. The integration methodology of *Pointillist* is able to handle data of different types and sizes, yielding a combined p-value from multiple data sources for the existence of each link in an interaction network. A novel aspect of this methodology is that it does not require a 'gold standard' set of data to be used for training nor does it make assumptions about the underlying statistical distributions in the form of parametric models. This involves designing an efficient deterministic optimization algorithm to minimize the numbers of misses and false positives via an iterative element by element procedure. The methodology is general-purpose so that it can be applied to integrate data from any existing and future technologies. Such approaches are critical to the improved reconstruction of high-quality interaction networks from high-throughput data sources.

Protein interaction networks have also been studied from an evolutionary standpoint, comparing networks of different species to identify conserved components and to yield insight into the functional importance of selected proteins, complexes, and pathways. *PathBLAST* is a network alignment and search tool used to compare protein interaction networks across species [41], providing a framework for addressing some of the challenges in studying interaction networks. For example, common signals present in both networks are reinforced while independent noise from false-positive reactions is reduced. Additionally, these comparisons can systematically catalog conserved network regions arising from shared homologous components. Similar to using *BLAST* to align protein sequences, *PathBLAST* is used to rapidly align protein interaction networks by searching for high-scoring pathway alignments. Each alignment score is a product of independent probabilities, where the probability of each protein pair is based on the *BLAST E*-value of sequence alignment and the protein interaction probability is based on false-positive rates associated with interactions in the target net-

work. Score distributions cannot currently be calculated within acceptable time frames, but it is expected that as more organisms become available this approach will become increasingly powerful. Other methods have focused on using the existing protein interaction network in one organism to predict the network in a different organism.

Interaction networks are an informative and powerful way to visualize large amounts of biological data. They also serve as a basis from which to study the topological properties of biological networks [25,29,36,38,70,78]. However, they fall short in terms of explaining or predicting quantitatively the emergent dynamic behaviors of these component interactions. Thus, while interaction networks are important inputs for building models, they are not in themselves quantitative systems models. For example, they are generally insufficient to predict how a perturbation to a particular node (e. g. gene) in the network will affect other nodes, as will the statistical and biochemical network models that will be the focus of the rest of this article. While interaction networks by themselves are not predictive of states and functions, they do provide insight into evolutionary relationships as well as provide a basis for further experimental investigation. Interaction networks are highly useful for differentiating direct and indirect effects that are found from statistical inference, and also provide essential input information for reconstructing statistical inference networks of any large scale because they help to constrain the number of combinations that need to be evaluated (as will be discussed more below).

Model Building as Informative Data Integration

The challenge of biological data integration begins with high-quality relational databases to house this information in an accessible manner [13,85]. However, it does not end there. While extremely useful, building these large databases does not by itself provide the ability to quantitatively integrate data to predict the emergent properties of interacting systems of components. Such highly-informative data integration, however, is possible, through the use of quantitative computational models.

As discussed above, high throughput technologies have initiated a phase transition in biology, making possible the study of large scale biological systems as a whole. Biology has become an information rich science, where the limiting step in biological discovery is now often not data generation but data analysis. Many in the scientific community have responded to this transformation by suggesting that the fundamental way in which much biology is done must change in order to best harness the large amount of biological data becoming available [1,31,37,43,44,]. Harnessing this information is critical to furthering our ultimate goal of improving human health and understanding the mechanisms underlying life. In response to these opportunities, three priorities have arisen in recent years from discussions at the National Institutes of Health: (1) the integration of heterogeneous biological data sets; (2) the integration of experiment and computations; and (3) the development of methods for systems analysis in biology. All of these high priority objectives for biological data integration can be met through the creation of quantitative models of biological systems – many of which can be expanded to the genome-scale.

Quantitative models provide for the integration of heterogeneous data sets in mathematically precise terms. Within the context of a cell-scale model, various data types interact in precisely defined ways such that predictions of cellular behavior can be made. These precise interactions are represented through defined mathematical relationships that mirror the underlying physical mechanisms to varying degrees of detail. Thus, quantitative models provide a most compact and powerful representation of diverse data sets.

Quantitative models themselves also serve as detailed hypotheses of how cells function, and thus constructing these models leads naturally to an integral connection between computational and experimental programs. When a quantitative model is constructed, the relationships between various aspects of a cell are related through mathematics in such a way as to reflect a physical process. Failure modes of these models point towards gaps in understanding the basic physical processes that underlie how cellular processes work. Thus, predictions from quantitative models, when found to be incorrect, lead to directed lines of questioning from which understanding of the cell's workings can be increased in a systematic fashion. Constructing and revising these models based on data forms a powerful iterative model building approach that can drive robust experimental programs.

Finally, the quest to create reliable cell-scale models provides a fertile ground for the development of systems analysis methods in biology. Developing the capacity to correctly model each new facet of cellular function provides its own set of challenges and requires the development of methods for its incorporation into the existing theoretical framework of a developing whole cell model. In this way, many modeling approaches are developed based upon the specific questions being asked and the problem being studied. Historically, mathematics have been developed primarily from the necessity to solve physics problems. Now, as biology becomes more and more amenable to mathematical analysis, it is likely that new mathematics

and new computational tools will be developed in order to meet the specific needs of modeling biology [10]. One example of this is the constraint-based modeling philosophy that will be discussed below, which has arisen to model biological systems because of the lack of complete information about the systems being studied and the inherent adaptability of living systems [55,67]. Creating increasingly accurate and detailed whole-cell models promises to be a rewarding effort that will result in the development of novel modeling methods.

The creation of cell-scale models is thus a fundamental and immediate goal of systems biology today and a powerful means for integrating large-scale data sets into an overarching cohesive framework. Such models are of great importance because they allow for the quantitative statement of a hypothesis of how a cell functions. Some disagreement may exist about what exactly constitutes a whole-cell model, but ideally one would like to be able to: 1) represent the multiple ways cells can accomplish the same task (redundancy); 2) model their dynamic and steady state behavior; 3) elucidate differences in pathological and normal states; 4) lead directly towards treatments for disease (where applicable); 5) provide mechanistic description (not just statistical outcome); 6) be specific for a personal genotype to phenotype relationship; 7) remain valid under a wide range of environmental conditions; 8) scale well such that models of relevant size are computable in reasonable length of time; 9) predict cellular behavior resulting from changes in gene content and the environment (preferably should not require fitting of parameters for each condition); 10) represent cells with models that are simpler than the cell itself. Additional and/or modified criteria could no doubt be stated. Suffice it to say, however, that models that satisfy all of these criteria do not yet exist, although models that meet some of these criteria are becoming available for various cellular processes.

As described above, quantitative models integrate heterogeneous data sets, combine experiment and computation, and drive the development of methods of systems analysis in biology. Thus, as is widely recognized, the pursuit of the goal of creating quantititive models is fundamental to the development of biology as a whole. Reliable cell-scale simulations hold the long-term promise of significantly increasing our ability to rationally design and modify biological systems. This power is fundamental to such applications as treating disease, developing new drugs, and using biosystems to improve our environment and create useful products.

For the purposes of this article, we will discuss two general modeling approaches, based on network models of two different kinds, namely, biochemical reaction networks, and statistical inferrence networks (Fig. 1). These two types of quantitative models represent fundamentally different approaches to network reconstruction and typically use quite different mathematical techniques. However, they share the same fundamental goal: namely, the quantitative prediction of the behavior of a biological system under a wide range of conditions and perturbations. Each of these modeling approaches will be discussed in detail in the following sections.

Data Integration Using Influence Network Models

One approach to modeling biological networks that holds tremendous potential for advancing knowledge of biology is the use of statistical associations in high-throughput data types to infer predictive network models. These statistical inference networks differ from interaction networks in that the connections between different network components is accompanied by mathematical functions precisely relating how perturbations to one (or a set of) component(s) effects the others. In essence, statistical inferrence networks are predictive and computable, while interaction networks as a rule are not. Interaction networks can, however, be an important – even essential – information source for the reconstruction of predictive statistical inference networks. The growing amounts of high-quality microarray gene expression data, coupled with input from gene-protein interaction networks, are now making it possible to make robust statistical inferences about the structure and dynamics of biomolecular control systems, such as transcriptional regulatory networks. Many approaches attempt to infer relationships between gene expression measurements using deterministic or stochastic formalisms. The fundamental idea behind these approaches is that models that faithfully capture such relationships have predictive capacity as regards system behavior and can be used to gain insight about system-wide properties, such as steady-state behavior or responses to perturbations or specific stimuli. There are a number of ways in which such relationships can be represented, both in the discrete and continuous domains.

One popular modeling approach that captures the nonlinear multivariate relationships exhibited by biological control circuits, such as gene regulatory networks, is the class of Boolean networks, which owes its inception to the work of Stuart Kauffman in the late 1960s [40]. In the Boolean network model, the variables (e. g., genes or proteins) are binary-valued, meaning that their states can be either on or off, and the relationships between the variables are captured by Boolean functions. Each (target) gene is assigned a Boolean rule that determines its value as a func-

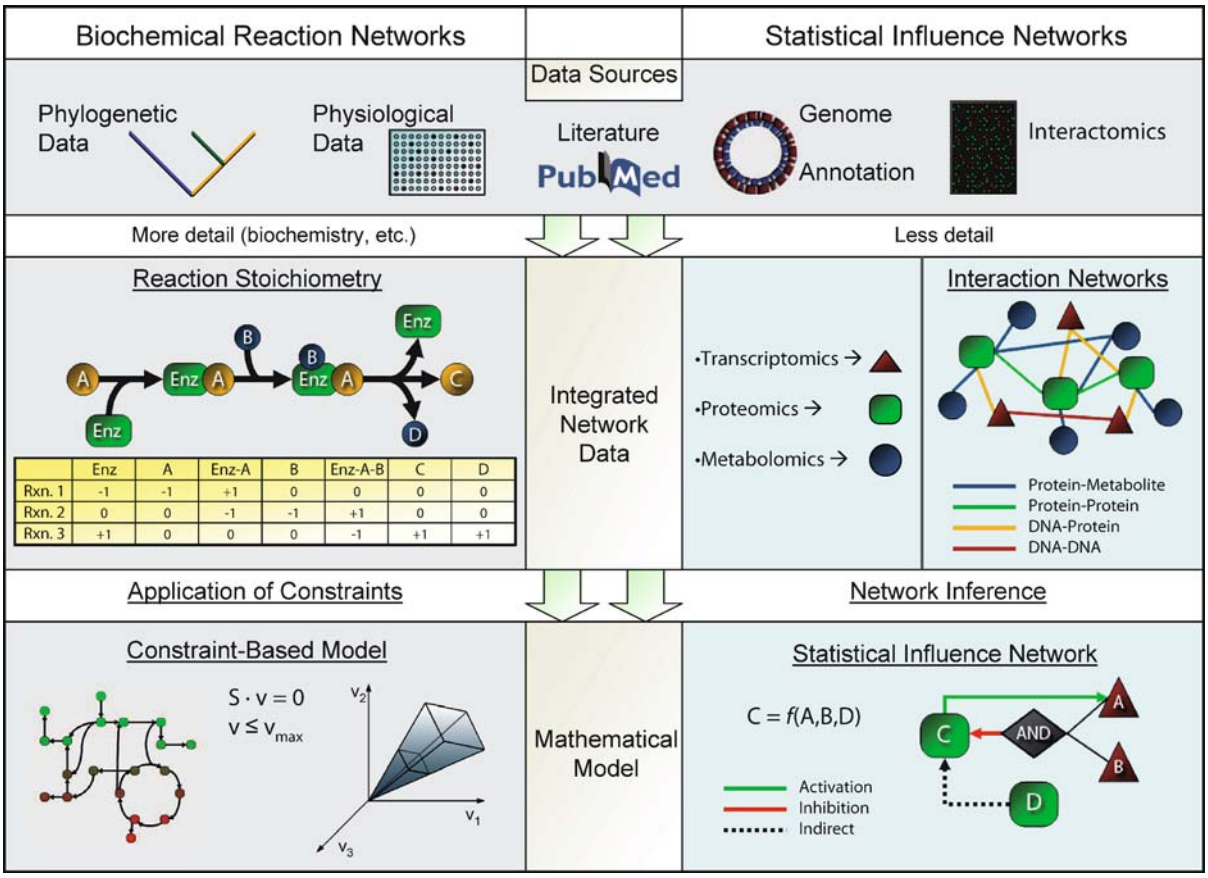

Biological Data Integration and Model Building, Figure 1
Data integration and model building through biochemical reaction networks and statistical influence networks

tion of the values of a set of other (predictor) genes, possibly including the target gene itself. System dynamics are generated by updating the Boolean functions, either synchronously or asynchronously, causing the system to transition from state to state in accordance with its Boolean update rules, where a state is a binary representation of the activities of all of the variables in the system (i. e., a binary vector representing the genes that are on or off at any given time).

Boolean network models have been constructed and analyzed for a number of developmental and physiological processes. For example, Albert et al. constructed a Boolean network model for a subset of genes of the fruitfly *Drosophila melanogaster*, which describes different stable gene expression patterns in the segmentation process of the developing embryo [1]. The steady-state behavior of this model was in excellent agreement with experimentally observed expression patterns under wild type and several gene mutation conditions. This study highlighted the importance of the network topology in determining biologically correct asymptotic states of the system. Indeed, when the segment polarity gene control network was modeled with more detailed kinetic models, such as systems of nonlinear differential equations, exceptional robustness to changes in the kinetic parameters was observed [83].

Boolean networks have also been used to model the yeast and mammalian cell cycle [20,49]. Li et al. demonstrated that the cell cycle sequence of protein states, which is a globally attracting trajectory of the dynamics, is extremely robust with respect to small perturbations to the network. The Boolean network formalism was also recently used to model systems-level regulation of the host immune response, which resulted in experimentally validated predictions regarding cytokine regulation and the effects of perturbations [79]. Boolean rules can be learned from gene expression data using methods from computational learning theory [47] and statistical signal processing [74,75,76].

A limitation of the Boolean network approach is its inherent determinism. Because of the inherent stochasticity

of gene expression and the uncertainty associated with the measurement process due to experimental noise and possible interacting latent variables (e. g. protein concentrations or activation states that are not measured), the inference of a single deterministic function may result in poor predictive accuracy, particularly in the context of small sample sizes (e. g., number of microarrays) relative to the number of genes.

One approach to "absorb" this uncertainty is to infer a number of simple functions (having few variables), each of which performs relatively well, and probabilistically synthesize them into a stochastic model, called a probabilistic Boolean network (PBN) [75]. The contribution of each function is proportional to its determinative potential as captured by statistical measures such as the coefficient of determination, which are estimated from the data [76]. The dynamical behavior of PBNs can be studied using the theory of Markov chains, which allows the determination of steady-state behavior as well as systematic intervention and control strategies designed to alter system behavior in a specified manner [77]. The PBN formalism has been used to construct networks in the context of several cancer studies, including glioma [26], melanoma [54], and leukemia [50]. PBNs, which are stochastic rule-based models, bear a close relationship to dynamic Bayesian networks [46] – a popular model class for representing the dynamics of gene expression.

Bayesian networks are graphical models that have been used to represent conditional dependencies and independencies among the variables corresponding to gene expression measurements [23]. One limitation of Bayesian networks for modeling genetic networks is that these models must be in the form of directed acyclic graphs and, as such, are not able to represent feedback control mechanisms. Dynamic Bayesian networks, on the other hand, are Bayesian networks that are capable of representing temporal processes [42,87] that may include such feedback loops. Since not all causal relationships can be inferred from correlation data, meaning that there can be different directed graphs that explain the data equally well, intervention experiments where genes are manipulated by overexpression or deletion have been proposed to learn networks [62]. The Bayesian network formalism has also been used to infer signaling networks from multicolor flow cytometry data [71].

There exist a number of other approaches for inferring large-scale molecular regulatory networks from high-throughput data sets. One example is a method, called the Inferelator, that selects the most likely regulators of a given gene using a nonlinear model that can incorporate combinatorial nonlinear influences of a regulator on target gene expression, coupled with a sparse regression approach to avoid overfitting [69]. In order to constrain the network inference, the Inferelator performs a preprocessing step of biclustering using the cMonkey algorithm [69], which results in a reduction of dimensionality and places the inferred interactions into experiment-specific contexts. The authors used this approach to construct a model of transcriptional regulation in *Halobacterium* that relates 80 transcription factors to 500 predicted gene targets.

Another method that predicts functional associations among genes by extracting statistical dependencies between gene expression measurements is the ARACNe algorithm [52]. This information-theoretic method uses a pairwise mutual information criterion across gene expression profiles to determine significant interactions. A key step in the method is the use of the so-called data processing inequality, which is intended to eliminate indirect relationships in which two genes are co-regulated through one or more intermediaries. Thus, the relationships in the final reconstructed network are more likely to represent the direct regulatory interactions. The ARACNe algorithm was applied to 336 genome-wide expression profiles of human B cells, resulting in the identification of MYC as a major regulatory hub along with newly identified and validated MYC targets [4].

A method related to the ARACNe algorithm, called the context likelihood of relatedness (CLR), also uses the mutual information measure but applies an adaptive background correction step to eliminate false correlations and indirect influences [18]. CLR was applied to a compendium of 445 *E. coli* microarray experiments collected under various conditions and compared to other inference algorithms on the same data set. The CLR algorithm had superior performance as compared to the other algorithms, which included Bayesian networks and ARACNe, when tested against experimentally determined interactions curated in the RegulonDB database. It also identified many novel interactions, a number of which were verified with chromatin immunoprecipitation [18].

Data Integration Using Biochemical Network Models

When the necessary biochemical detail is known, we can reconstruct biochemical reaction networks. These networks represent the underlying chemistry of the system, and thus at a minimum represent stoichiometric relationships between inter-converted biomolecules. Biochemical reaction networks are thus directly based on chemistry, rather than a reflection of statistical associations, as are statistical inferrence networks. These stoichiomet-

ric reconstructions have been most commonly applied to small-molecule interconversions in metabolic networks, but this formalism can easily incorporate biological transformations of all types, including for metabolic, signaling, protein–protein interactions, and gene regulatory networks [55]. The stoichiometry of biochemical reaction networks can now be reconstructed at the genome-scale, and at smaller scale with sufficient detail to generate kinetic models. These biochemical reaction networks represent many years of accumulated experimental data and can be interrogated *in silico* to determine their functional states. Genome-scale models based on biochemical networks provide a comprehensive, yet concise, description of cellular functions.

For metabolism, the reconstruction of the biochemical reaction network is a well-established procedure [14,22,27,29,66,68,81], while methods for the reconstruction of the associated regulatory [11,24] and signaling networks [56,57,58] with stoichiometric detail are developing. In any case, the reconstruction process is typically labor-intensive, involving massive amounts of data collection, biochemical characterization, and concurrent validation. A metabolic network reconstruction begins by collecting the list of enzymatic genes for the organism of interest. Using online databases and published literature, the predicted products and functions of each gene are recorded. It is often advantageous to model an organism whose annotated genome is publicly available. *BLAST* searches and other bioinformatics approaches seeking specifically to uncover functions the cell must have the capacity to carry out, however, can be used to identify the functions of previously uncharacterized genes. This is a simple example of how the process of iterative network model building guides discovery of new biological hypotheses. Along with the gene and peptide properties for each enzyme in the network, the stoichiometric description for each reaction is also recorded, with care taken to ensure mass and charge balance. Each enzyme may be the product of one or multiple genes, whereas a given enzyme may catalyze one or multiple reactions. The association between genes, proteins, and reactions (GPRs) are described using Boolean language. Examples of "OR" relationships are 1) isozymes, where either one protein or the other is required for the reaction to occur; or 2) transcript variants, where at least one in a set of alternatively spliced transcripts is needed to produce the enzyme corresponding to a particular reaction. In contrast, an "AND" relationship might describe a multimeric protein, where multiple genes are required for production, or a complex of proteins, all necessary to catalyze a given reaction. Finally, for organisms with multiple cellular compartments, the cellular localization is determined for all enzymes and reactions. Metabolic networks can be reconstructed in a pathway-by-pathway manner, allowing a team of researchers to work in parallel to complete the model more quickly. Computational methods can be used to interrogate the properties of each pathway, or subnetwork, to evaluate the completeness of each of the submodels given the proper 'exchange' molecules that will be trafficked back and forth between subsystems.

Signaling network reconstructions follow the general process of metabolic networks, but are based on a slightly different set of data. These networks can again be reconstructed in a pathway-by-pathway manner, and in some cases, only one pathway may be the focus of study. While stand-alone signaling network models typically do not rely on GPR associations, the Boolean rules described above become necessary when integrating signaling and regulatory networks. Published literature is most often used, with signaling databases more scarce than metabolic, to identify the relevant signaling interactions (e. g. phosphorylation, dephosphorlyation, receptor binding, poly-ubiquitination, etc.). These interactions are then represented as mass-balanced biochemical reactions. Reconstructing both metabolic and signaling pathways in the form of biochemical reactions allows for each type of network to be represented as a stoichiometric matrix, serving as an integrated database to synthesize large amounts of data about signaling networks. In the stoichiometric matrix (S), each row represents a molecular compound and each column represents a reaction. For metabolism, these networks are often focused on just the metabolites, where the existence of a protein that catalyzes this reaction is used to allow that reaction to be present in the network. It is also possible (and truer to the realities in the system) to represent the proteins themselves as compounds in the network, which enables the integration of proteomics, metabolomics, and flux data (Fig. 1). For signaling, as well as regulatory networks, the inclusion of the proteins as compounds is essential. Transcriptional regulatory networks are generally constructed based on experimental gene expression data, high-throughput protein-DNA interaction data from, for example, CHIP–chip experiments, as well as compilations of data from specific transcription factor binding experiments. Regulatory reactions are represented as Boolean switches, where active or inactive genes correspond to both metabolic and signaling activity. Regulatory reconstructions are most useful when coupled with either metabolic, signaling networks, or both – thus fully capturing cellular behavior under different states. The collective process of reconstructing biochemical reaction networks has been referred to as the "two-dimensional annotation" of genomes [55].

Genome-scale models can be generated using a constraint-based approach, as has been reviewed in detail elsewhere [67]. This modeling process involves a three-step procedure. First, the biochemical network is reconstructed, as described above. Second, the physico-chemical and environmental constraints under which the reconstructed network operates are represented mathematically. Physico-chemical constraints are based upon properties such as enzyme capacity, reaction stoichiometry and thermodynamics associated with reaction directionality and biochemical loops.

Constraints are represented in two fundamental forms: balances and bounds. Balance constraints dictate the conservation of quantities such as mass, energy, and momentum, as well as electroneutrality, solvent capacity, and osmotic pressure. The conservation of mass, for example, is commonly used to describe fluxes through biochemical networks at steady state (i. e. flux balance analysis). At steady-state, there is no accumulation or loss of metabolites within the network. Therefore, the rate of production of each metabolite must equal its rate of consumption. The balance of fluxes is typically represented mathematically as $S \bullet v = 0$, where v is a vector of fluxes through the metabolic network and S is the stoichiometric matrix containing all the reactions in the reconstructed network (Fig. 1). Bounds limit numerical ranges of individual variables and parameters such as fluxes, concentrations, and kinetic constants. For example, upper and lower limits can be applied to individual fluxes ($v_{min} \leq v \leq v_{max}$).

Environmental constraints include nutrient availability, pH, temperature, osmolarity, and the availability of electron acceptors. These constraints are important for the quantitative analysis of biochemical network models. For example, defined media and well-documented environmental conditions are used to integrate data from various laboratories into quantitative models that are accurately descriptive and predictive. Environmental constraints can also be represented with bounds, defining the range of metabolite fluxes into and out of the model system to agree with a desired medium or condition. For example, to simulate cellular growth on a glucose-free medium, the upper and lower bounds on glucose flux into the cell would be set equal to zero ($v_{glucose} = 0$).

Regulatory constraints are enforced via the Boolean rules describing GPR associations. In the case of an isozyme or multimeric protein (both "AND" relationships), the absence of one gene will prevent the production of its protein product, effectively turning off the corresponding reaction in the system. Gene knockout or deletion simulations can therefore provide further insight to the relation between sets of genes and network states. Regulatory constraints also provide a link between environmental conditions and the reconstructed network; deletion tests can also be used to determine gene essentiality under various conditions. Finally, regulatory constraints are essential for network reconstructions aiming to integrate signaling and metabolic pathways.

The statement of constraints leads to the definition of a solution space that contains all non-excluded network states, describing the possible functions of the reconstructed network or all the allowable phenotypes. The third step is the determination of the possible solutions in this space that correspond to physiologically meaningful states (Fig. 2). Numerous *in silico* techniques have been developed to determine optimal network states (flux balance analysis, extreme pathways analysis, etc.) under different biological objectives (biomass, ATP production) [68]. This three-step constraint-based modeling procedure has been successfully utilized to study phenotypes in various model [3,16,21,35] and infectious [59,63] microorganisms. Recently, this approach has passed a significant milestone, namely the reconstruction of the human metabolic network [14]. Constraint-based analysis has been applied to selected human systems in the past [65,80], but with this reconstruction in place, this approach is poised to have a much more significant impact on modeling human systems.

Stoichiometric reaction networks can be expanded into dynamic models using standard differential equation models and the addition of information in the form of kinetic rate constants for network reactions. These systems, unlike those generally used in the constraint-based approach, do not make the steady-state assumption for all compounds but rather can simulate detailed dynamic behavior. Thus, they can model a greater degree of complexity for a given network and simultaneously account for both concentrations of compounds and fluxes through reactions. The disadvantage of these systems compared with constraint-based analysis is that they require many more parameters and are thus more data intensive to create and/or can be more prone to overfitting when large numbers of parameters are unknown. These models are not typically created at the whole cell scale, with a notable exception being cell-scale models of red blood cell metabolism [16,53].

In recent years, these types of dynamic models have been used very successfully to provide mechanistic models of, for example, key signaling pathways involved in critical physiological and pathophysiological processes. For example, a differential equation model of NF-kB signaling [30] was updated and used to elucidate the role of tumor necrosis factor in controlling the sustained phase of lipopolysac-

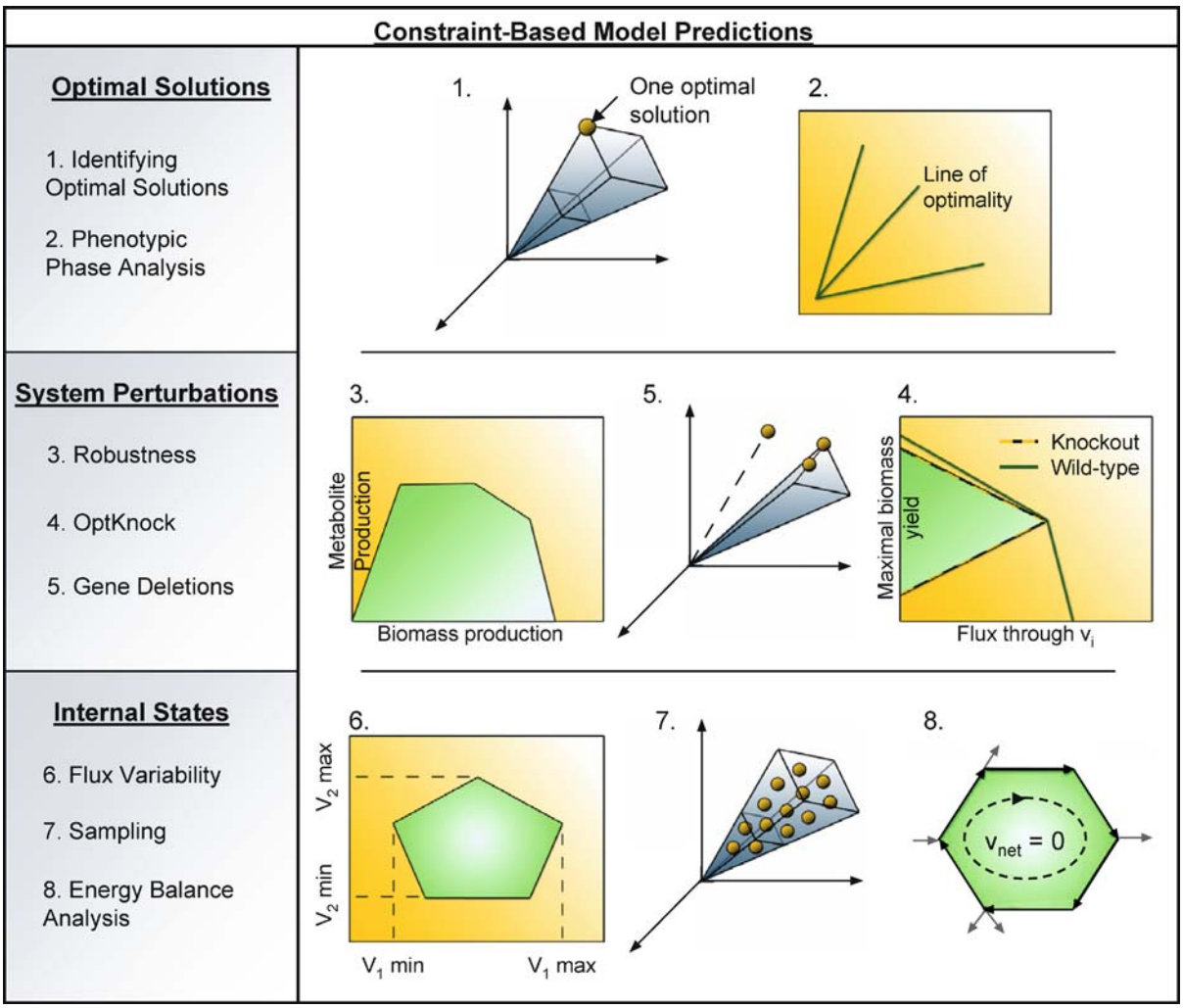

Biological Data Integration and Model Building, Figure 2

Representative computational predictions that can be made using constraint-based modeling of biochemical reaction networks. Example computational predictions include: 1) identifying optimal solutions, such as predicting an optimal growth rate [16], 2) defining phenotypic phase planes that can be used to interpret and predict the outcome of adaptive evolution [35], 3) robustness analyzes evaluating the effect of changing fluxes on optimal growth [17], 4) OptKnock and related methods for predicting the effect of genetic modifications on the relationship between cell growth and the production of a desired product [8,32,60,61], 5) the lethality of gene deletions [19,29], 6) the analysis of the range of allowable values for each flux in a biochemical network under a set of specified constraints [51,67], 7) monte-carlo methods for characterizing the space of potential network steady states [3,65,80], and 8) the introduction of thermodynamic constraints to more realistically constrain the range of internal biochemical fluxes [5,6].

charide-induced IKK activation [12,84], demonstrating the usefulness of these models for deciphering functions of biological networks. Another example was the use of a model to study the integrated dynamic behavior of a network of epidermal growth factor receptor family members [28]. Christopher et al. have constructed a dynamic simulator of gene expression and signaling networks in a human cancer cell [9]. In the long run, such dynamic models set the stage for personalized medicine by offering the promise of rational drug design and control of the outcome of induced molecular perturbations.

Comparison of Biochemical and Statistical Network Models for Data Integration

There are fundamental differences between the biochemical and statistical classes of network modeling described herein. One clear difference is the manner in which these

underlying networks are reconstructed. For biochemical networks, reconstruction is typically a work-intensive process requiring significant biochemical characterization with little network inference done (other than inferences of single gene function for catalyzing a reaction based on sequence homology). Thus, the ability to rapidly generate networks for organisms that are relatively uncharacterized from high-throughput data is an inherent advantage of the inferred statistical networks. One advantage of the biochemical network models is that, once reconstructed, the networks are not as subject to change (other than addition) since many of the links are based directly on biochemical evidence. Inferred networks, on the other hand, can undergo substantial changes in light of additional data. Another common difference, although not fundamental, is that constraint-based biochemical network models have mostly been used to model flux, whereas inference networks have mostly been used to predict substance amounts (e. g. mRNA expression). One way this can be thought of is that the biochemical network models currently link more closely to functional phenotype (i. e. fluxes) [72], while the inferred networks relate more directly to available high-throughput data (i. e. transcriptomes). The kinetic biochemical network models, of course, have the capacity to account for both flux and abundance, but suffer from the limitation that they are by far the most data intensive to reconstruct. Another key advantage of biochemical reaction networks, stemming from their basis in chemistry, is that physico-chemical laws apply, such as mass-energy balance, while such laws are not generally applicable to the inferred networks. Of course, the advantage of the inferred networks is that, since they do not need to be mechanistic or require biochemical detail, they can be applied very broadly to systems that are not yet well characterized and can link very disparate data types as long as underlying correlations exist. In summary, both modeling types are essential to contemporary computational systems biology and provide advantages over each other in different settings.

Future Directions

The future for biological data integration through model building is bright. High throughput experimental technologies keep improving, yielding exponentially increasing sources of data. Computing power continues to increase exponentially as well, meaning that both the necessary computing power and data streams are expected to keep the need for data integration through modeling at a high and increasing premium in biological research. However, very significant challenges clearly remain in integrating biological data going forward. Even fundamental concepts such as what exactly constitutes a gene or a protein remain somewhat ambiguous, and so the definition of data sets and the rigorous data storage itself are challenges. New and improved mathematical and computational approaches will also need to be developed for accurate biological modeling, adaptive over time along with the emergence of new measurement technologies. One interesting challenge for model building going forward is whether hybrid models that take advantage of the strengths of both biochemical and statistical inference modeling approaches can be constructed to move us further towards the goal of predictive whole-cell models and beyond. Early attempts have been done to link Boolean regulatory networks with constraint-based flux models [11], but the extent to which these approaches can be married to provide significant advances in our ability to model biological networks remains an open question.

Bibliography

1. Albert R, Othmer HG (2003) The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in Drosophila melanogaster. J Theor Biol 223(1):1–18
2. Alm E, Arkin AP (2003) Biological networks. Curr Opin Struct Biol 13(2):193–202
3. Almaas E, Kovacs B et al (2004) Global organization of metabolic fluxes in the bacterium Escherichia coli. Nature 427(6977):839–43
4. Basso K, Margolin AA et al (2005) Reverse engineering of regulatory networks in human B cells. Nat Genet 37(4):382–90
5. Beard DA, Babson E et al (2004) Thermodynamic constraints for biochemical networks. J Theor Biol 228(3):327–33
6. Beard DA, Liang SD et al (2002) Energy balance for analysis of complex metabolic networks. Biophys J 83(1):79–86
7. Bonneau R, Reiss DJ et al (2006) The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. Genome Biol 7(5):R36
8. Burgard AP, Pharkya P et al (2003) Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol Bioeng 84(6):647–57
9. Christopher R, Dhiman A et al (2004) Data-driven computer simulation of human cancer cell. Ann NY Acad Sci 1020:132–53
10. Cohen JE (2004) Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. PLoS Biol 2(12):e439
11. Covert MW, Knight EM et al (2004) Integrating high-throughput and computational data elucidates bacterial networks. Nature 429(6987):92–6
12. Covert MW, Leung TH et al (2005) Achieving stability of lipopolysaccharide-induced NF-kappaB activation. Science 309(5742):1854–1857
13. Deshpande N, Addess KJ et al (2005) The RCSB Protein Data Bank: a redesigned query system and relational database

based on the mmCIF schema. Nucleic Acids Res (Database issue) 33:D233–7
14. Duarte NC, Becker SA et al (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci USA 104(6):1777–82
15. Duarte NC, Herrgard MJ et al (2004) Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. Genome Res 14(7):1298–309
16. Edwards JS, Ibarra RU et al (2001) In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nat Biotechnol 19(2):125–30
17. Edwards JS, Palsson BO (2000) Robustness analysis of the Escherichia coli metabolic network. Biotechnol Prog 16(6):927–39
18. Faith JJ, Hayete B et al (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol 5(1):e8
19. Famili I, Forster J et al (2003) Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. Proc Natl Acad Sci USA 100(23):13134–13139
20. Faure A, Naldi A et al (2006) Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle. Bioinformatics 22(14):e124–e131
21. Forster J, Famili I et al (2003) Large-scale evaluation of in silico gene deletions in Saccharomyces cerevisiae. Omics 7(2):193–202
22. Francke C, Siezen RJ et al (2005) Reconstructing the metabolic network of a bacterium from its genome. Trends Microbiol 13(11):550–558
23. Friedman N (2004) Inferring cellular networks using probabilistic graphical models. Science 303(5659):799–805
24. Gianchandani EP, Papin JA et al (2006) Matrix formalism to describe functional states of transcriptional regulatory systems. PLoS Comput Biol 2(8):e101
25. Han JD, Bertin N et al (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430(6995):88–93
26. Hashimoto RF, Kim S et al (2004) Growing genetic regulatory networks from seed genes. Bioinformatics 20(8):1241–7
27. Heinemann M, Kummel A et al (2005) In silico genome-scale reconstruction and validation of the Staphylococcus aureus metabolic network. Biotechnol Bioeng 92(7):850–864
28. Hendriks BS, Wiley HS et al (2003) HER2-mediated effects on EGFR endosomal sorting: analysis of biophysical mechanisms. Biophys J 85(4):2732–2745
29. Herrgard MJ, Palsson BO (2005) Untangling the web of functional and physical interactions in yeast. J Biol 4(2):5
30. Hoffmann A, Levchenko A et al (2002) The IkappaB-NF-kappaB signaling module: temporal control and selective gene activation. Science 298(5596):1241–1245
31. Hood L, Heath JR et al (2004) Systems biology and new technologies enable predictive and preventative medicine. Science 306(5696):640–643
32. Hua Q, Joyce AR et al (2006) Metabolic analysis of adaptive evolution for in silico-designed lactate-producing strains. Biotechnol Bioeng 95(5):992–1002
33. Hwang D, Rust AG et al (2005) A data integration methodology for systems biology. Proc Natl Acad Sci USA 102(48):17296–17301
34. Hwang D, Smith JJ et al (2005) A data integration methodology for systems biology: experimental verification. Proc Natl Acad Sci USA 102(48):17302–17307
35. Ibarra RU, Edwards JS et al (2002) Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature 420(6912):186–189
36. Ideker T (2004) A systems approach to discovering signaling and regulatory pathways–or, how to digest large interaction networks into relevant pieces. Adv Exp Med Biol 547:21–30
37. Ideker T, Galitski T et al (2001) A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet 2:343–372
38. Ideker T, Ozier O et al (2002) Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics (Suppl) 18(1):S233–S2340
39. Jamshidi N, Edwards JS et al (2001) Dynamic simulation of the human red blood cell metabolic network. Bioinformatics 17(3):286–287
40. Kauffman SA (1993) The origins of order : self organization and selection in evolution. Oxford University Press, New York
41. Kelley BP, Yuan B et al (2004) PathBLAST: a tool for alignment of protein interaction networks. Nucleic Acids Res (Web Server issue) 32:W83–W88
42. Kim SY, Imoto S et al (2003) Inferring gene networks from time series microarray data using dynamic Bayesian networks. Brief Bioinform 4(3):228–235
43. Kirschner MW (2005) The meaning of systems biology. Cell 121(4):503–504
44. Kitano H (2002) Computational systems biology. Nature 420(6912):206–210
45. Kurzweil R (2005) The singularity is near: when humans transcend biology. Penguin, London
46. Lahdesmaki H, Hautaniemi S et al (2006) Relationships between probabilistic Boolean networks and dynamic Bayesian networks as models of gene regulatory networks. Signal Processing 86(4):814–834
47. Lahdesmaki H, Shmulevich I et al (2003) On Learning Gene Regulatory Networks Under the Boolean Network Model. Machine Learning 52(1–2):147–167
48. Levy S, Sutton G et al (2007) The Diploid Genome Sequence of an Individual Human. PLoS Biol 5(10):e254
49. Li F, Long T et al (2004) The yeast cell-cycle network is robustly designed. Proc Natl Acad Sci USA 101(14):4781–4786
50. Li H, Zhan M (2006) Systematic intervention of transcription for identifying network response to disease and cellular phenotypes. Bioinformatics 22(1):96–102
51. Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab Eng 5(4):264–276
52. Margolin AA, Wang K et al (2006) Reverse engineering cellular networks. Nat Protoc 1(2):662–671
53. Mulquiney PJ, Kuchel PW (2003) Modelling metabolism with Mathematica, detailed examples including erythrocyte metabolism. CRC Press, Boca Raton
54. Pal R, Datta A et al (2005) Intervention in context-sensitive probabilistic Boolean networks. Bioinformatics 21(7):1211–1218
55. Palsson B (2004) Two-dimensional annotation of genomes. Nat Biotechnol 22(10):1218–1219
56. Papin JA, Hunter T et al (2005) Reconstruction of cellular sig-

nalling networks and analysis of their properties. Nat Rev Mol Cell Biol 6(2):99–111
57. Papin JA, Palsson BO (2004) The JAK-STAT signaling network in the human B-cell: an extreme signaling pathway analysis. Biophys J 87(1):37–46
58. Papin JA, Palsson BO (2004) Topological analysis of mass-balanced signaling networks: a framework to obtain network properties including crosstalk. J Theor Biol 227(2):283–297
59. Papin JA, Price ND et al (2002) The genome-scale metabolic extreme pathway structure in Haemophilus influenzae shows significant network redundancy. J Theor Biol 215(1):67–82
60. Pharkya P, Burgard AP et al (2003) Exploring the overproduction of amino acids using the bilevel optimization framework OptKnock. Biotechnol Bioeng 84(7):887–899
61. Pharkya P, Burgard AP et al (2004) OptStrain: a computational framework for redesign of microbial production systems. Genome Res 14(11):2367–76
62. Pournara I, Wernisch L (2004) Reconstruction of gene networks using Bayesian learning and manipulation experiments. Bioinformatics 20(17):2934–2942
63. Price ND, Papin JA et al (2002) Determination of redundancy and systems properties of the metabolic network of Helicobacter pylori using genome-scale extreme pathway analysis. Genome Res 12(5):760–769
64. Price ND, Reed JL et al (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2(11):886–897
65. Price ND, Schellenberger J et al (2004) Uniform sampling of steady-state flux spaces: means to design experiments and to interpret enzymopathies. Biophys J 87(4):2172–2186
66. Reed JL, Palsson BO (2003) Thirteen years of building constraint-based in silico models of Escherichia coli. J Bacteriol 185(9):2692–2699
67. Reed JL, Palsson BO (2004) Genome-scale in silico models of E. coli have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. Genome Res 14(9):1797–1805
68. Reed JL, Vo TD et al (2003) An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). Genome Biol 4(9):R54
69. Reiss DJ, Baliga NS et al (2006) Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. BMC Bioinformatics 7:280
70. Rual JF, Venkatesan K et al (2005) Towards a proteome-scale map of the human protein-protein interaction network. Nature 437(7062):1173–1178
71. Sachs K, Perez O et al (2005) Causal protein-signaling networks derived from multiparameter single-cell data. Science 308(5721):523–529
72. Sauer U (2004) High-throughput phenomics: experimental methods for mapping fluxomes. Curr Opin Biotechnol 15(1):58–63
73. Shannon P, Markiel A et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–504
74. Shmulevich I, Dougherty ER et al (2002) Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics 18(2):261–274
75. Shmulevich I, Dougherty ER et al (2002) From Boolean to probabilistic Boolean networks as models of genetic regulatory networks. Proceedings of the IEEE 90(11):1778–1792
76. Shmulevich I, Dougherty ER et al (2002) Gene perturbation and intervention in probabilistic Boolean networks. Bioinformatics 18(10):1319–1331
77. Smith HO, Tomb JF et al (1995) Frequency and distribution of DNA uptake signal sequences in the Haemophilus influenzae Rd genome. Science 269(5223):538–540
78. Stelzl U, Worm U et al (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122(6):957–968
79. Thakar J, Pillione M et al (2007) Modelling Systems-Level Regulation of Host Immune Responses. PloS Comput Biol 3(6):e109
80. Thiele I, Price ND et al (2005) Candidate metabolic network states in human mitochondria. Impact of diabetes, ischemia, and diet. J Biol Chem 280(12):11683–11695
81. Thiele I, Vo TD et al (2005) Expanded metabolic reconstruction of Helicobacter pylori (iIT341 GSM/GPR): an in silico genome-scale characterization of single- and double-deletion mutants. J Bacteriol 187(16):5818–5830
82. Tong AH, Lesage G et al (2004) Global mapping of the yeast genetic interaction network. Science 303(5659):808–813
83. von Dassow G, Meir E et al (2000) The segment polarity network is a robust developmental module. Nature 406(6792):188–192
84. Werner SL, Barken D et al (2005) Stimulus specificity of gene expression programs determined by temporal control of IKK activity. Science 309(5742):1857–1861
85. Westbrook J, Feng Z et al (2002) The Protein Data Bank: unifying the archive. Nucleic Acids Res 30(1):245–248
86. Westerhoff HV, Palsson BO (2004) The evolution of molecular biology into systems biology. Nat Biotechnol 22(10):1249–1252
87. Zou M, Conzen SD (2005) A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics 21(1):71–79

Biological Development and Evolution, Complexity and Self-organization in

Stuart A. Newman[1], Gabor Forgacs[2]
[1] New York Medical College, Valhalla, USA
[2] University of Missouri, Columbia, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Dynamic Multistability: Cell Differentiation
Differential Adhesion: Gastrulation
Biochemical Oscillations: Segmentation
Reaction-Diffusion Mechanisms: Body Axis Formation
Evolution of Developmental Mechanisms
Future Directions
Bibliography

Glossary

Differential adhesion The capacity of cells to adhere to each other in a cell type-dependent manner. The strength of adhesion between two cells of type *A* typically differs from that between cells of type *B*. This may be due to differences either in the number or type of cell adhesion molecules.

Differential gene expression The main regulatory basis of cell differentiation. Cells of different type in a given organism carry the same set of genes (the genome) but vary in which of these genes are active, that is, expressed.

Biochemical oscillation The variation in the concentration of a given molecule in principle either in space or time, although typically the term is reserved for the latter.

Reaction-diffusion mechanism A conceptual framework for describing spatiotemporal pattern formation in a system of several interacting and diffusing chemical species.

Canalization An evolved property of developmental pathways that permits the robust generation of a phenotype in the face of perturbations. The perturbations can be those internal to the organism, in the form of gene mutation or developmental noise, or external to it, in the form of environmental variability.

Definition of the Subject

Much work over the past half-century in developmental and evolutionary biology has focused on a subset of an organism's components, its genes. The hierarchical regulatory relationships among genes have been a major emphasis in studies in development, while the variation of genes has played a corresponding role in evolutionary studies. In the past decade, however, investigators have increasingly considered the part played by physical and dynamical properties of cells and tissues, and their molecular components, in producing biological characteristics over the course of ontogeny and phylogeny. Living cells and tissues encompass numerous interpenetrating multicomponent systems in which dynamical interactions among intracellular gene products, metabolites, ions, etc., and interactions between cells, directly via their adhesive surfaces, or indirectly via secreted extracellular matrix (ECM) molecules or diffusible signaling molecules ("morphogens"), generate topologically complex, three-dimensional self-perpetuating entities consisting of up to several hundred different cell types in elaborate arrangements. In this description, the systems properties that give rise to organisms and their substructures, however, are nowhere specified in any "genetic program".

While these "generic" (i. e., common to living and nonliving systems) physical and chemical-dynamic processes contribute to the production of form and pattern in all modern-day organisms, it is likely that early in the evolution of multicellular organisms such processes were even more efficacious in determining organismal forms. In particular, morphologies originally generated by physics and chemical dynamics could have provided templates for genetic evolution that stabilized and reinforced (rather than innovated) multicellular body plans and organ forms. The hierarchical genetic control of development seen in modern organisms can thus be considered an outcome of this evolutionary interplay between genetic change and generic physicochemical processes.

Introduction

The field of developmental biology has as its major concern *embryogenesis*: the generation of fully-formed organisms from a fertilized egg, the *zygote*. Other issues in this field, organ *regeneration* and tissue *repair* in organisms that have already passed through the embryonic stages, have in common with embryogenesis three interrelated phenomena: *cell differentiation*, the production of distinct cell types, *cell pattern formation*, the generation of specific spatial arrangements of cells of different types, and *morphogenesis*, the molding and shaping of tissues [25]. The cells involved in these developmental processes and outcomes generally have the same genetic information encoded in their DNA, the *genome* of the organism, so that the different cell behaviors are largely associated with *differential gene expression*.

Because of the ubiquity and importance of differential gene expression during development, and the fact that each type of organism has its own unique genome, a highly gene-centered view of development prevailed for several decades after the discovery of DNA's capacity to encode information. In particular, development was held to be the unfolding of a "genetic program" specific to each kind of organism and organ, based on differential gene expression. This view became standard despite the fact that no convincing models had ever been presented for how genes or their products (proteins and RNA molecules) could alone build three dimensional shapes and forms, or even generate populations of cells that utilized the common pool of genetic information in different ways. By "alone" is meant without the help of physics and chemical dynamics, the scientific disciplines traditionally invoked to ex-

plain changes in shape, form, and chemical composition in nonliving material systems.

It has long been recognized, however, that biological systems, in addition to being repositories of genes, are also physicochemical systems, and that phenomena first identified in the nonliving world can also provide models for biological processes. Indeed, at the level of structure and function of biomolecules and macromolecular assemblages such as membranes and molecular motors, biology has historically drawn on ideas from chemistry and physics. More recently there has been intensified cooperation between biological and physical scientists at the level of complex biological systems, including developmental systems. The rise of systems developmental, and evolutionary-developmental, biology required the generation by physical and information scientists of theoretical tools and computational power sufficient to model systems of great complexity. But also, and equally important, has been the growing recognition by modern biologists that organisms are more than programmed expressions of their genes, and that the behavior of systems of many interacting components is neither obvious nor predictable on the basis of the behavior of their parts.

Since biology is increasingly studied at the systems level, there has also been new attention to the origination of these complex systems. In many instances, it is reasonable to assume that complexity and integration of subsystems in living organisms has evolved in the context of forms and functions that originally emerged (in evolutionary history) by simpler physicochemical means. Thus the elaborate system of balanced, antagonistic signaling interactions that keep cell metabolism homeostatic, and embryogenesis on-track, can be seen as the result of accretion, by natural selection, of stabilizing mechanisms for simpler physicochemical generative processes that would otherwise be less reliable.

In what follows we will give a fairly wide-ranging, though hardly comprehensive, survey of the physicochemical processes organizing, and presumed to have promoted the emergence of the complex systems underlying the multicellular development of animals.

Dynamic Multistability: Cell Differentiation

The early embryos of multicellular organisms are referred to as *blastulae*. These are typically hollow clusters of several dozen to several hundred cells that arise through the cleavage of the zygote (cell divisions without volume increase of the embryo). While the zygote is "totipotent" – it has the potential to give rise to any of the more than 200 specialized cell types (e. g., bone, cartilage, the various kinds of muscle, blood, and nerve cells) of the mature human body – its first few divisions generate cells that are "pluripotent" – capable of giving rise to only a limited range of cell types. These cells, in turn, diversify into ones with progressively limited potency, ultimately generating all the (generally unipotent) specialized cells of the body [25].

The transition from wider to narrower developmental potency is referred to as *determination*. This stage of cell specialization generally occurs with no overt change in the appearance of cells. Instead, subtle modifications, only discernable at the molecular level, set the altered cells on new and restricted developmental pathways. A later stage of cell specialization, referred to as *differentiation*, results in cells with vastly different appearances and functional modifications – electrically excitable neurons with extended processes up to a meter long, bone and cartilage cells surrounded by solid matrices, red blood cells capable of soaking up and disbursing oxygen, and so forth.

Since each cell of the organism contains an identical set of genes (except for the egg and sperm and their immediate precursors, and some cells of the immune system), a fundamental question of development is how the same genetic instructions can produce different types of cells. This question pertains to both determination and differentiation. Multicellular organisms solve the problem of specialization by activating a type-specific subset of genes in each cell type.

The *biochemical state* of a cell can be defined as the list of all the different types of molecules contained within it, along with their concentrations. The dynamical state of a cell, like that of any dynamical system, resides in a multidimensional space, the "state space", with dimensionality equal to the number of system variables (e. g., chemical components) [96]. During the cell division cycle (i. e., the sequence of changes that produces two cells from one), also referred to as the cell cycle, the biochemical state changes periodically with time. (This, of course, assumes that cells are not undergoing differentiation). If two cells have the same complement of molecules at corresponding stages of the cell cycle, then, they can be considered to be of the same differentiated state. The cell's biochemical state also has a spatial aspect – the concentration of a given molecule might not be uniform throughout the cell. We will discuss an example of this related to *Drosophila* early development in Sect. "Evolution of Development Mechanisms". The state of differentiation of the cell (its *type*) can be identified with the collection of proteins it is capable of producing.

Of the estimated 25,000 human genes [76], a large proportion constitutes the "housekeeping genes," involved in

functions common to all or most cells types. In contrast, a relatively small number of genes – possibly fewer than a thousand – specify the type of determined or differentiated cells; these genes are *developmentally regulated* (i. e., turned on and off in an embryonic stage- and embryonic position-dependent fashion) during embryogenesis. Cell type diversity is based to a great extent on sharp transitions in the dynamical state of embryonic cells, particularly with respect to their developmentally-regulated genes.

When a cell divides it inherits not just a set of genes and a particular mixture of molecular components, but also a dynamical system at a particular *dynamical state*. Dynamical states of many-component systems can be transient, stable, unstable, oscillatory, or chaotic [96]. The cell division cycle in the early stages of frog embryogenesis, for example, is thought to be controlled by a *limit cycle* oscillator [5]. A limit cycle is a continuum of dynamical states that define a stable orbit in the state space surrounding an unstable *node*, a node being a stationary (i. e., time-independent) point, or steady state, of the dynamical system [102]. The fact that cells can inherit dynamical states was demonstrated experimentally by Elowitz and Leibler [20]. These investigators used genetic engineering techniques to provide the bacterium *Escherichia coli* with a set of feedback circuits involving transcriptional repressor proteins such that a biochemical oscillator not previously found in this organism was produced. Individual cells displayed a chemical oscillation with a period longer than the cell cycle. This implied that the dynamical state of the artificial oscillator was inherited across cell generations (if it had not, no periodicity distinct from that of the cell cycle would have been observed). Because the biochemical oscillation was not tied to the cell cycle oscillation, newly divided cells in successive generations found themselves at different phases of the engineered oscillation.

The ability of cells to pass on dynamical states (and not just "informational" macromolecules such as DNA) to their progeny has important implications for developmental regulation, since continuity and stability of a cell's biochemical identity is key to the performance of its role in a fully developed organism. Inheritance which does not depend directly on genes is called "epigenetic inheritance" [68] and the biological or biochemical states inherited in this fashion are called "epigenetic states". Although epigenetic states can be determined by reversible chemical modifications of DNA [41], they can also represent alternative steady states of a cell's network of active or expressed genes [46]. The ability of cells to undergo transitions among a limited number of discrete, stable epigenetic states and to propagate such decisions from one cell generation to the next is essential to the capacity of the embryo to generate diverse cell types.

All dividing cells exhibit oscillatory dynamical behavior in the subspace of the full state space whose coordinates are defined by cell cycle-related molecules. In contrast, cells exhibit alternative stable steady states in the subspace defined by molecules related to states of cell determination and differentiation. During development, a dividing cell might transmit its particular system state to each of its daughter cells, but it is also possible that some internal or external event accompanying cell division could push one or both daughter cells out of the "basin of attraction" in which the precursor cell resided and into an alternative state. (The basin of attraction of a stable node is the region of state space surrounding the node, in which all system trajectories, present in this region, terminate at that node [96].

It follows that not every molecular species needs to be considered simultaneously in modeling a cell's transitions between alternative biochemical states. Changes in the concentrations of the small molecules involved in the cell's housekeeping functions such as energy metabolism and amino acid, nucleotide, and lipid synthesis, occur much more rapidly than changes in the pools of macromolecules such as RNAs and proteins. The latter class of molecules is indicative of the cell's gene expression profile and can therefore be considered against a metabolic background defined by the average concentrations of molecules of the former class. Most of a cell's active genes are kept in the "on" state during the cell's lifetime, since like the metabolites they are also mainly involved in housekeeping functions. The pools of these "constitutively active" gene products can often be considered constant, with their concentrations entering into the dynamic description of the developing embryo as fixed parameters rather than variables. (See Goodwin [27] for an early discussion of separation of time scales in cell activities).

As mentioned above, it is primarily the regulated genes that are important to consider in analyzing determination and differentiation. And of these regulated genes, the most important ones for understanding cell type specification are those whose products control the activity of other genes.

Genes are regulated by a set of proteins called *transcription factors*. Like all proteins, these factors are themselves gene products, in this case having the specific function of turning genes on and off. They do this by binding to specific sequences of DNA usually "upstream" of the gene's transcription start site, called the "promoter". Transitions between cell types during development are frequently controlled by the relative levels of transcription

factors. Because the control of most developmentally-regulated genes is a consequence of the synthesis of the factors that regulate their transcription, cell-type transitions can be driven by changes in the relative levels of a fairly small number of transcription factors [14]. We can thus gain insight into the dynamical basis of cell type switching (i. e., determination and differentiation) by focusing on molecular circuits, or networks, consisting solely of transcription factors and the genes that specify them. Networks in which the components mutually regulate one another's expression are termed "autoregulatory". Although a variety of autoregulatory transcription factor circuits appear during development, it is clear that such circuits will have different properties depending on their particular "wiring diagrams", that is the interaction patterns among their components.

Transcription factors can be classified either as *activators*, which bind to a site on a gene's promoter, and enhance the rate of that gene's transcription over its basal rate, or *repressors*, which decrease the rate of a gene's transcription when bound to a site on its promoter. Repression can be competitive or noncompetitive. In the first case, the repressor will interfere with activator binding and can only depress the gene's transcription rate to the basal level. In the second case, the repressor acts independently of any activator and can therefore potentially depress the transcription rate below basal levels. The basal rate of transcription depends on constitutive transcription factors, housekeeping proteins which are distinct from those in the autoregulatory circuits that we consider below.

Networks of autoregulatory transcription factors constitute dynamical systems. Like such systems encountered in physics and mathematics they can productively be modeled as systems of ordinary differential equations (ODE). A classic approach, exemplified by the work of Keller [47], and more recently by Cinquin and Demongeot [10], is to set up systems of ODEs representing experimentally motivated topologies and kinetics of transcription factor interactions. The solutions of these systems of ODEs are then examined with respect to stationary states, limit cycles, bifurcations, etc., under variations in initial conditions, external inputs and kinetic parameters. When considered a part of a multicellular aggregate, this single-cell dynamical system can switch between its intrinsic alternative states depending on signals it experiences from its microenvironment – other such cells, the ECM, or in certain cases the external world.

In his analysis of developmental switching, Keller [47] used simulation methods to investigate the behavior of several autoregulatory transcription factor networks with a range of wiring diagrams (Fig. 1). Each network was represented by a set of n coupled ordinary differential equations – one for the concentration of each factor in the network – and the steady-state behaviors of the systems were explored. The questions asked were: how many stationary states exist; are they stable or unstable?

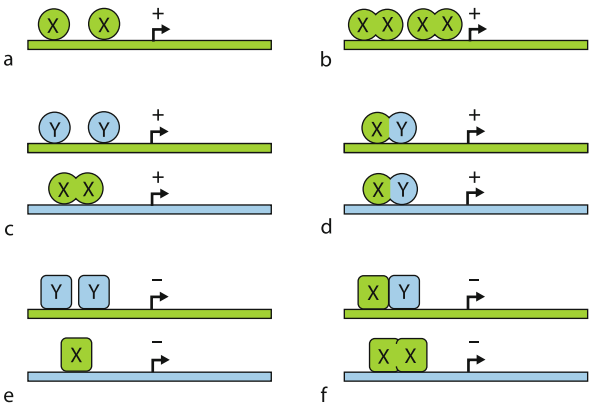

Biological Development and Evolution, Complexity and Self-organization in, Figure 1
Six model genetic circuits discussed by Keller [47]. a Autoactivation by monomer *X*; b autoactivation by dimer X_2; c mutual activation by monomer *Y* and dimer X_2; d autoactivation by heterodimer *XY*; e mutual repression by monomers *X* and *Y*; f mutual repression by dimer X_2 and heterodimer *XY*. Activated and repressed states of gene expression are represented respectively by + and –. The various transcription factors (circles for activators and squares for repressors) bind to the promoters of the genes, the colors of which correspond to the transcription factors they specify. Based on Keller [47]; see original paper for details. Figure modified and reprinted, with permission, from [21]

One such network, designated as the "Mutual repression by dimer and heterodimer" (MRDH) (Fig. 1f), comprises the genes encoding the transcriptional repressor, X, and the gene encoding the protein, Y, and thus represents a two component network. The stationary values of the two components are governed by the following equations:

$$\frac{\mathrm{d}[X]}{\mathrm{d}t} = S_{X_B} - \left\{d_X[X] + 2d_{X_2}K_{X_2}[X]^2 + d_{XY}K_{XY}[X][Y]\right\} = 0 \quad (1)$$

$$\frac{\mathrm{d}[Y]}{\mathrm{d}t} = \frac{1+\rho K_X K_{X_2}[X]^2}{1+K_X K_{X_2}[X]^2} S_{Y_B} - \{d_Y[Y] + d_{XY}K_{XY}[X][Y]\} = 0\,, \quad (2)$$

where $[\ldots]$ represents concentrations of the components, the S's synthesis rates, and the d's decay constants for the indicated monomeric ($[X], [Y]$) and dimeric ($[X]^2$) species. (See Keller [47] for details).

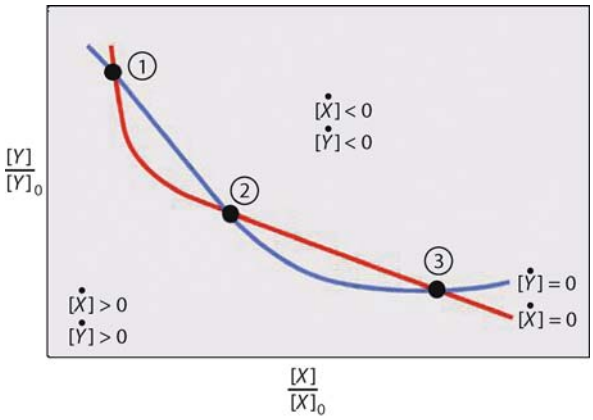

Biological Development and Evolution, Complexity and Self-organization in, Figure 2

The solutions of the steady-state Eqs. (1) and (2), given in terms of the solutions $[X_0]$ and $[Y_0]$. Here $[X_0]$ is defined as the steady state cellular level of monomer X produced in the presence of steady state cellular level $[Y_0]$ of monomer Y by the rate of transcription S_{X_0}. By definition (see Eq. (1)), $S_{X_0} = d_X[X_0] + 2d_{X_2}K_{X_2}[X_0]^2 + d_{XY}K_{XY}[X_0][Y_0]$. Since along the *blue* and *red lines*, respectively, $d[X]/dt \equiv [\dot{X}] = 0$ and $d[Y]/dt \equiv [\dot{Y}] = 0$, the intersections of these curves correspond to the steady state solutions of the system of equations, Eqs. (1) and (2). Steady states 1 and 3 are stable, while steady state 2 is unstable. Based on Keller [47]; see original paper for details. Figure modified and reprinted, with permission, from [21]

It was found that if in the absence of the repressor X the rate of synthesis of protein Y is high, in its presence, the system described by Eqs. (1) and (2) exhibits three steady states, as shown in Fig. 2. Steady states 1 and 3 are stable, thus could be considered as defining two distinct cell types, while steady state 2 is unstable. In an example using realistic kinetic constants, the steady-state values of $[X]$ and $[Y]$ at the two stable steady states differ substantially from one another, showing that the dynamical properties of these autoregulatory networks of transcription factors can provide the basis for generating stable alternative cell fates during early development.

The biological validity of Keller's model depends on whether switching between these alternative states is predicted to occur under realistic conditions. By this criterion the model works: the microenvironment of a cell, containing an autoregulatory network of transcription factors, could readily induce changes in the rate of synthesis of one or more of the factors via signal transduction pathways that originate outside the cell ("outside-in" signaling [24]). Moreover, the microenvironment can also affect the activity of transcription factors in an autoregulatory network by indirectly interfering with their localization in the cell's nucleus, where transcription takes place [59]. In addition, cell division may perturb the cellular levels of autoregulatory transcription factors, particularly if they or their mRNAs are unequally partitioned between the daughter cells. Any jump in the concentration of one or more factors in the autoregulatory system can bring it into a new basin of attraction and thereby lead to a new stable cell state.

The Keller model shows that the existence of multiple steady states in an embryonic cell's state space makes it possible, in principle, for more than one cell type to arise among its descendents. However this capability does not, by itself, provide the conditions under which such a potentially divergent cell population would actually be produced and persist as long as it is required.

Keller's is a single cell analysis. Another, less conventional, approach, acknowledges from the start the multicellular nature of developing organisms. Here the dynamical attractors sought are ones that pertain to the multicellular system as a whole, which consists of multiple copies of initially identical model cells in interaction with one another [43,44].

The need for an inherently multicellular approach to the cell type switching problem is indicated by experimental observations that suggest that cell differentiation depends on properties of cell aggregates rather than simply those of individual cells. For example, during muscle differentiation in the early frog embryo, the muscle precursor cells must be in contact with one another throughout gastrulation (the set of rearrangements that establish the body's main tissue layers referred to as the "germ layers", the ectoderm, mesoderm and endoderm) in order to develop into terminally differentiated muscle [29].

The need for cells to act in groups in order to acquire new identities during development has been termed the "community effect" [29]. This phenomenon is a developmental manifestation of the general property of cells and other dynamical systems of assuming one or another of their possible internal states in a fashion that is dependent on inputs from their external environment. In the case noted above the external environment consists of other cells of the same genotype.

Kaneko, Yomo and co-workers [43,44] have described a previously unknown chemical-dynamic process, termed "isologous diversification", by which replicate copies of the same dynamical system (i. e., cells of the same initial type) can undergo stable differentiation simply by virtue of exchanging chemical substances with one another. This differs from Keller's model described above in that the final state achieved exists only in the phase space of the collective "multicellular" system. Whereas the distinct local states of each cell within the collectivity are mutually re-

inforcing, these local states are not necessarily attractors of the dynamical system representing the individual cell, as they are in Keller's model. The Kaneko–Yomo system thus provides a model for the community effect.

The following simple version of the model, based on Kaneko and Yomo [44], was initially presented in Forgacs and Newman [21]. Improvements and generalizations of the model presented in subsequent publications [43] do not change its qualitative features.

Consider a system of originally identical cells with intra- and inter-cell dynamics, which incorporate cell growth, cell division and cell death (for the general scheme of the model, see Fig. 3a). The dynamical variables are the concentrations of molecular species ("chemicals") inside and outside the cells. The criterion by which differentiated cells are distinguished is the average of the intracellular concentrations of these chemicals (over the cell cycle). Cells are assumed to interact with each other through their effect on the intercellular concentrations of the chemicals A and B (only two are considered in this simplified model). Chemicals enter the cells by either active transport or diffusion. It is further assumed that a source chemical S is supplied by a flow from an external tank to the chamber containing the cells.

Kaneko and Yomo [44] consider cell division to be equivalent to the accumulation of a threshold quantity of DNA. To avoid infinite growth in cell number, a condition for cell death has to also be imposed. It is assumed that a cell will die if the amount of chemicals A and B in its interior is below the "starvation" threshold.

The system described can be expressed as the following set of equations (see also Fig. 3b):

$$\frac{\mathrm{d}x_i^S}{\mathrm{d}t} = -E^S x_i^S + \mathrm{Transp}_i^S + \mathrm{Diff}_i^S \tag{3}$$

$$\frac{\mathrm{d}x_i^A}{\mathrm{d}t} = (e_B x_i^B)x_i^B - (e_A x_i^B)x_i^A + E^S x_i^S + \mathrm{Transp}_i^A + \mathrm{Diff}_i^A \tag{4}$$

$$\frac{\mathrm{d}x_i^B}{\mathrm{d}t} = (e_A x_i^B)x_i^A - (e_B x_i^B)x_i^B - k x_i^B + \mathrm{Transp}_i^B + \mathrm{Diff}_i^B \tag{5}$$

$$\frac{\mathrm{d}X^S}{\mathrm{d}t} = f\left(\overline{X^S} - X^S\right) - \sum_{i=1}^{N}\left(\mathrm{Transp}_i^S + \mathrm{Diff}_i^S\right), \tag{6}$$

Here the x_i are concentrations of the various chemicals in the ith cell, X^S the external concentration of the source chemical. E^S is the (constant) concentration in each cell of an enzyme that converts S into A, which in turn is catalyzed by the enzyme (with concentration $E_i^A = e_A x_i^B$) to produce B. B is catalyzed by an enzyme (with concentration $E_i^B = e_B x_i^B$) to produce A and it also produces DNA (at a rate k). *Transp* and *Diff* are respectively active transport and diffusion functions, which depend on both the intracellular (x_i) and extracellular (X_i) concentrations. (See Forgacs and Newman [21] for details).

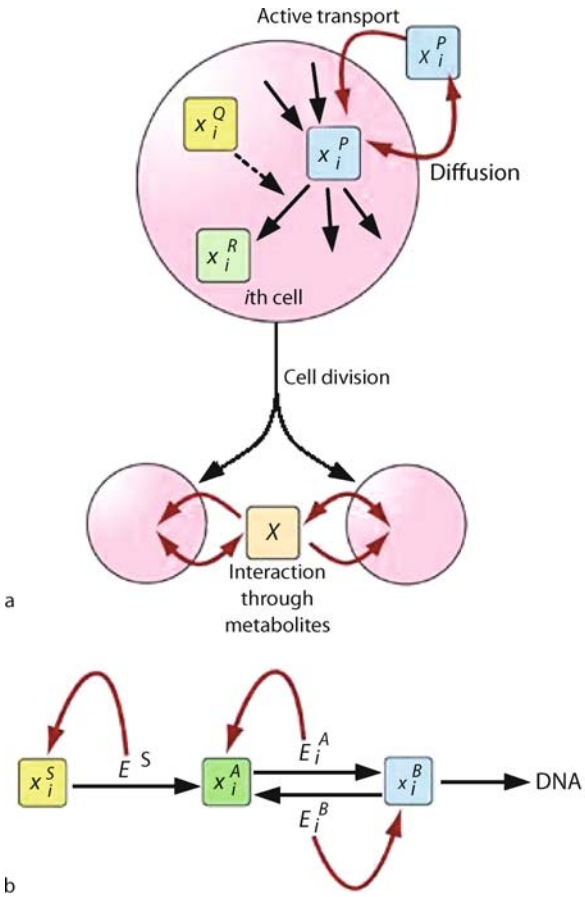

Biological Development and Evolution, Complexity and Self-organization in, Figure 3

a General scheme of the model of Kaneko and Yomo. The concentrations of each chemical may differ in the various cells (cell *i* is shown in the *center*) but some (*P* in the example) may have a fixed value in the extracellular environment. External chemicals can enter the cells by passive diffusion or active transport. The concentration of any chemical inside a given cell depends on chemical reactions in which other chemicals are precursors or products (*solid arrows*) or cofactors (*dashed arrow*). Once cell division has taken place by the synthesis of DNA (one of the "chemicals") exceeding a threshold, cells communicate with one another by exchange of transportable chemicals. b Schematic representation of the intracellular dynamics of a simple version of the Kaneko–Yomo model. *Red arrows* symbolize catalysis. The variables $x_i^A(t)$, $x_i^B(t)$, $x_i^S(t)$ and E_i^A, E_i^B, E^S denote respectively the concentrations of chemicals *A*, *B* and *S* and their enzymes in the *i*th cell, as described in the text. Reprinted, with permission, from [21]

Simulations based on the above model and its generalizations, using a larger number of chemicals [43], led to the following general features, which are likely to pertain to real, interacting cells as well:

1. As the model cells replicate (by division) and interact with one another, eventually multiple biochemical states corresponding to distinct cell types appear. The different types are related to each other by a hierarchical structure in which one cell type stands at the apex, cell types derived from it stand at subnodes, and so on. Such pathways of generation of cell type, which are seen in real embryonic systems, are referred to as developmental lineages.
2. The hierarchical structure appears gradually. Up to a certain number of cells (which depends on the model parameters), all cells have the same biochemical state (Fig. 4a). When the total number of cells rises above a certain threshold value, the state with identical cells is no longer stable. Small differences between cells first introduced by random fluctuations in chemical concentrations start to be amplified. For example, synchrony of biochemical oscillations in different cells of the cluster may break down (Fig. 4b). Ultimately, the population splits into a few groups ("dynamical clusters"), with the phase of the oscillator in each group being offset from that in other groups, like groups of identical clocks in different time zones.
3. When the ratio of the number of cells in the distinct clusters falls within some range (depending on model parameters), the differences in intracellular biochemical dynamics are mutually stabilized by cell-cell interactions.
4. With further increase of cell number, the average concentrations of the chemicals over the cell cycle become different (Fig. 4c). That is to say, groups of cells come to differ not only in the phases of the same biochemical oscillations, but also in their average chemical composition integrated over the entire lifetimes of the cells. After the formation of cell types, the chemical compositions of each group are inherited by their daughter cells (Fig. 4d).

In contrast to the Keller model in which different cell types represent a choice among basins of attraction for a multi-attractor system, with external influences having the potential to bias such preset alternatives, in the Kaneko–Yomo model interactions between cells can give rise to stable intracellular states, which would not exist without such interactions. Isologous diversification thus provides a plausible model for the community effect [29]. It is reasonable to expect that both intrinsic multistability of a dy-

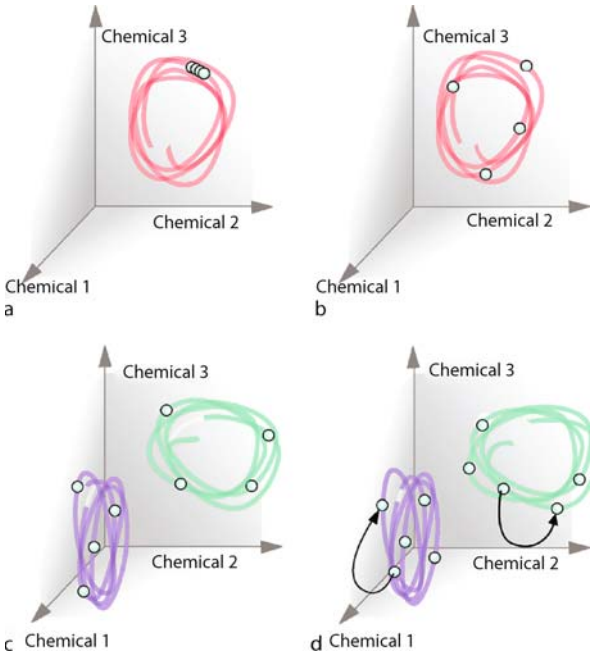

Biological Development and Evolution, Complexity and Self-organization in, Figure 4

Schematic representation of the differentiation scenario in the isologous diversification model of Kaneko and Yomo. When there are *N* cells and *C* chemicals ($C = 3$ in the figure), the state space of the multicellular system is $N \times C$ dimensional. A point in this space corresponds to the instantaneous value of all the chemicals and in each cell the orbits represent the evolution of these concentrations in the multicellular system. As long as the biochemical states of the replicating cells are identical, points along the orbit could characterize the synchronous states of the cells. This is illustrated in panel a, where the four circles, representing cells with the same phase and magnitude of their chemicals, overlap. With further replication, cells with differing biochemical states appear. First, chemicals in different cells differ only in their phases, thus the circles in panel b still fall on the same orbit, albeit are well separated in space. With further increase in cell number differentiation takes place: not only the phases but also the magnitudes (i. e., the averages over the cell cycle) of the chemicals in different cells will differ. The two orbits in panel c represent two distinct cell types, each different from the original cell type shown in panels a and b. Panel d illustrates the "breeding true" of the differentiated cells. After the formation of distinct cell types, the chemical compositions of each group are inherited by their daughter cells. That is, chemical compositions of cells are recursive over subsequent divisions as the result of stabilizing interactions. Cell division is represented here by an arrow from a progenitor cell to its progeny. Adapted, with changes, from Kaneko [42]. Reprinted, with permission, from [21]

namical system of the sort analyzed by Keller, and interaction-dependent multistability, as described by Kaneko, Yomo, and coworkers, based as they are on generic properties of complex dynamical systems, are utilized in initiating developmental decisions in various contexts in different organisms.

Differential Adhesion: Gastrulation

As cells differentiate they become biochemically and structurally specialized and capable of forming multicellular structures with characteristic shapes, such as those of the earliest developmental stages, *blastulae* and multilayered *gastrulae*. Later during development tubular vessels, ducts and crypts emerge. The appearance and function of these specialized structures reflect, among other things, differences in the ability of cells to adhere to each other and the distinct mechanisms by which they do so. The distinct adhesion molecules expressed by differentiating cells thus mobilize physical forces to produce stereotypical multicellular structures.

The organization of tissues is determined not only by the chemical nature of the cells' adhesion molecules but also the distribution of these molecules on the cell surface. *Epithelioid* cells express cell adhesion molecules (CAMs), such as cadherins, uniformly over their surfaces and tend to form solid cell masses. *Epithelial* cells, which form two dimensional sheets must have CAMs predominantly along their lateral surfaces, whereas substrate adhesion molecules (SAMs), such as integrins, must populate their basal surfaces, along which interaction with the planar ECM, the basal lamina, is carried out. Such distribution of CAMs and SAMs renders epithelial cells *polarized*. Finally, *mesenchymal* cells express SAMs over their surfaces by which they interact with interstitial ECM.

Adhesion enters into the phenomena of embryonic development in several distinct ways. The first, and most straightforward way, is in simply holding tissues together. Whereas mature tissues contain the definitive, relatively long-lived forms of CAM-containing junctions, during early development the CAM-containing junctions are present in apparently immature forms [17], consistent with the provisional arrangement of cells and their capacity to rearrange during these stages.

The other roles for adhesion during development are based on its modulation – the phenomenon of *differential adhesion*. The regulated spatiotemporal modulation of adhesion is an important driving force for major morphogenetic transitions during embryogenesis. The simplest form of this is the detachment of cell populations from existing tissues. This is usually followed by their relocation.

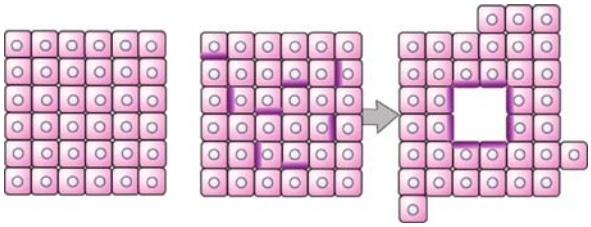

Biological Development and Evolution, Complexity and Self-organization in, Figure 5
Schematic illustration of lumen formation by polarized cells expressing cell adhesion molecules only on restricted parts of their surface. The shading in the eight cells in the *middle* panel represents the lack of adhesion molecules on corresponding regions of the cells. As a consequence, if minimization of configurational energy is the driving force in cell rearrangement, a lumen, shown on the *right* is bound to appear. (Based on Newman [62]) Reprinted, with permission, from [21]

But modulation of adhesive strength without complete detachment also has morphogenetic consequences, whether it occurs locally, on the scale of the single cell surface, or more globally, on the scale of whole cells within a common tissue.

Polar expression of CAMs can lead rather directly to morphogenetic change, as illustrated in Fig. 5. In the process of differentiation some of the originally non-polarized cells may lose their CAMs along part of their surface, thus preventing them from adhering to one another at those sites [98]. If the cells move around randomly, maximizing their contacts with one another in accordance with this constraint, a hollow region (a lumen) will naturally arise.

One of the most dramatic morphogenetic processes exhibited by embryonic cells is sorting, the ability of cells of distinct types to segregate into distinct tissues which do not intermix at their common boundary. Steinberg postulated that cells of different origin adhere to each other with different strengths and, in analogy with immiscible liquids such as oil and water, undergo a process of phase separation in which the final configuration corresponds to the minimum of interfacial and surface free energies [93]. (Here the different tissues play a role analogous to liquids and the constituent cells mimic the molecules of the liquid). This "differential adhesion hypothesis" (DAH) was expressed in quantitative terms by Phillips [77] based on geometric analysis of surface and interfacial tensions of immiscible droplets of liquids. According to the DAH, the final "phase-separated" state of two adjacent tissues is an equilibrium configuration not dependent on the pathway by which it was achieved. That is, it will be the same whether arrived at by fusion of two intact fragments of tissue or by the sorting out of their respective cells from a binary mixture (Fig. 6). Another implication of the DAH is

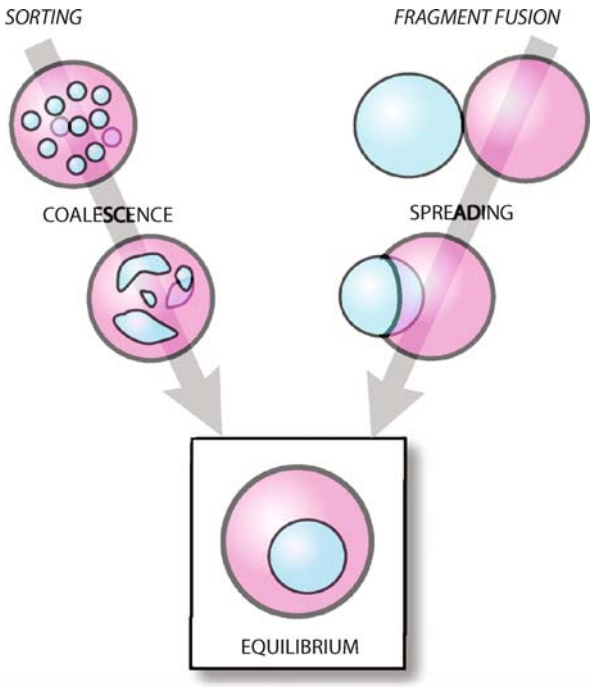

Biological Development and Evolution, Complexity and Self-organization in, Figure 6
Different paths by which two immiscible liquids or cell aggregates (composed of cells with differing adhesive properties) may arrive at the same equilibrium state. The path on the *left* shows sorting or nucleation, which proceeds through the gradual coalescence of groups of cells. The path on the *right* corresponds to engulfment, which occurs through spreading. Reprinted, with permission, from [21]

that tissue engulfment relationships should form a hierarchy: if tissue A engulfs tissue B and B engulfs C in separate experiments, it follows that A will engulf C if that experiment is performed. Finally, the DAH predicts that the values of tissue surface tensions should fall in a quantitative order that corresponds to the engulfment hierarchy. Each of these predictions of the DAH has been amply confirmed experimentally [19,21,22,93].

By the time the blastula has formed, the embryo already contains, or begins to generate, a number of differentiated cell types. Insofar as these cell types have or acquire distinct adhesive properties, compartmentalization or other forms of regional segregation start taking place. This regionalization, accompanied by the collective movement of the resulting cell masses – that is common to all forms of gastrulation – gives rise to embryos consisting of two or three germ layers along with some subsidiary populations of cells.

To understand how cells can rearrange by differential adhesion it is useful to describe the notion of the "work of adhesion". Consider a unit interfacial area between two materials A and B immersed in a medium, denoted by M (which in particular could be vacuum or, in the case of tissues, the extracellular medium or tissue culture medium). We define the work of adhesion w_{AB} as the energy input required to separate A and B across the unit area in medium M. We can imagine such a unit area to be formed in the following way. First, we separate a rectangular column of A and B to produce two free unit surfaces of each substance (Fig. 7). This requires w_{AA} and w_{BB} amounts of work, respectively. These quantities are called the works of cohesion. (Note that the magnitudes of the work of adhesion and cohesion depend on the medium). We then combine these pieces to end up with two unit interfacial areas between A and B as shown in Fig. 7. Thus the total work, Δ_{AB}, needed to produce a unit interfacial area between A and B is given by

$$\Delta_{AB} = \tfrac{1}{2}(w_{AA} + w_{BB}) - w_{AB} \tag{7}$$

The quantity Δ_{AB} is called the interfacial energy and Eq. (7) is known as the Dupré equation [39]. If the interface is formed by two immiscible liquids then Eq. (7) can readily be expressed in terms of liquid surface and interfacial tensions. By its definition the surface tension is the energy required to increase the surface of the liquid by one unit of area. Since the works w_{AA} and w_{BB} create two units of area, we obtain

$$\sigma_{AB} = \sigma_{AM} + \sigma_{BM} - w_{AB}\ , \tag{8}$$

where σ_{AM}, σ_{BM} and σ_{AB} are respectively the surface tensions of liquids A and B and their mutual interfacial tension [39]. (Whereas for solids Δ_{AB} depends on the amount of interfacial area between A and B already formed, for liquids, σ_{AB} does not). Note that immiscibility of A and B implies $\sigma_{AB} > 0$. If, on the contrary, $\sigma_{AB} \leq 0$, it is energetically more beneficial for the molecules of liquid A to be surrounded by molecules of liquid B, and vice versa; that is, A and B are miscible.

If we now invoke the liquid-like behavior of tissues, we can apply Eq. (7) or Eq. (8) to obtain the conditions for sorting in terms of the w's or σ's. We imagine the cells of tissues A and B to be initially randomly intermixed and surrounded by tissue culture medium and allow them to sort, as discussed above. Let us assume that tissue A is the more cohesive one. This implies that w_{AA} is the largest of the three quantities on the right hand side of Eq. (7). In the energetically most favorable configuration at the end of the sorting process cells of tissue A form a sphere, the configuration in which they have minimal contact with their environment and maximal contact with each other. Then, depending on the relative magnitudes of the w's,

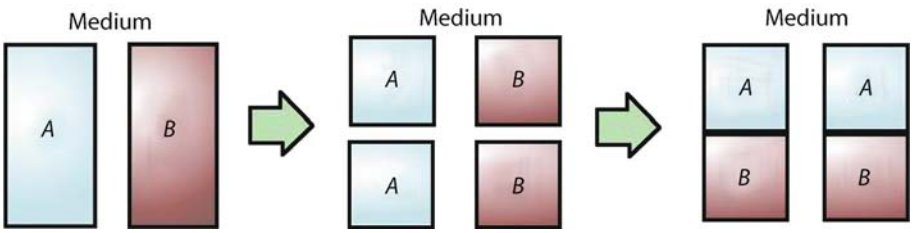

Biological Development and Evolution, Complexity and Self-organization in, Figure 7
Schematic illustration of creation of an interface between two materials *A* and *B*, immersed in a common medium. In the two-step process shown, first free interfaces of materials *A* or *B* are produced (*middle* panel), which requires the separation of their corresponding subunits (molecules, cells) from one another, and thus involves the works of cohesion, w_{AA} and w_{BB}. In the second step the free interfaces are combined to form the *AB* interface (by rearranging the *A* and *B* blocks; *right* panel). The separation of *A* and *B* requires the work of adhesion, w_{AB}. If the cross-sectional area of columns *A* and *B* is of unit magnitude, then the operation shown results in two units of interfacial area between *A* and *B*. Reprinted, with permission, from [21]

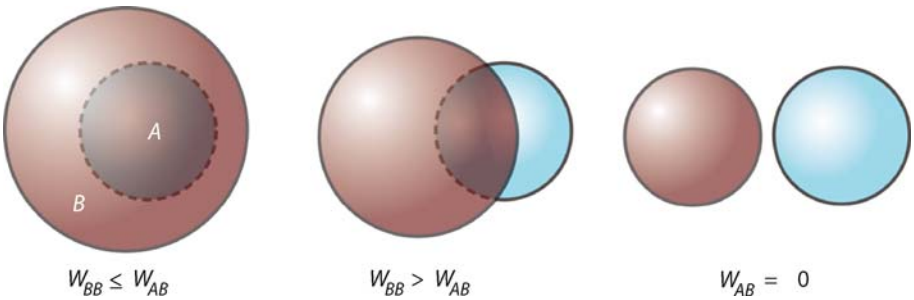

Biological Development and Evolution, Complexity and Self-organization in, Figure 8
Geometric configurations of immiscible liquids *A* (represented by the interior smaller sphere in the *left* panel) and *B* (represented by the larger sphere in the *left* panel, and the corresponding relations between the works of cohesion and adhesion. It is assumed that *A* is more cohesive than *B*: $w_{AA} > w_{BB}$ and $\sigma_{AM} > \sigma_{BM}$ (*M* denotes the surrounding medium). The *left* and *middle* panels correspond respectively to complete and partial envelopment. Figure modified and reprinted, with permission, from [21]

the sphere of tissue *B* may completely or partially envelop the sphere of tissue *A* or the two spheres may separate (Fig. 8) [21,93].

Biochemical Oscillations: Segmentation

A wide variety of animal types, ranging across groups as diverse as insects, annelids (e. g., earthworms), and vertebrates, undergo *segmentation* early in development, whereby the embryo, or a major portion of it, becomes subdivided into a series of tissue modules [25]. These modules typically appear similar to each other when initially formed; later they may follow distinct developmental fates and the original segmental organization may be all but obscured in the adult form. Somite formation (or somitogenesis) is a segmentation process in vertebrate embryos in which the tissue to either side of the central axis of the embryo (where the backbone will eventually form) becomes organized into parallel blocks of tissue.

Somitogenesis takes place in a sequential fashion. The first somite begins forming as a distinct cluster of cells in the anterior region (towards the head) of the embryo's body. Each new somite forms just posterior (towards the tail) to the previous one, budding off from the anterior portion of the unsegmented presomitic mesoderm (PSM) (Fig. 9). Eventually, 50 (chick), 65 (mouse), or as many as 500 (certain snakes) of these segments will form.

In the late nineteeth century the biologist William Bateson speculated that the formation of repetitive blocks of tissue, such as the somites of vertebrates or the segments of earthworms might be produced by an oscillatory process inherent to developing tissues [2,65]. More recently, Pourquié and coworkers made the significant observation that the gene *c-hairy1*, which specifies a transcription factor, is expressed in the PSM of avian embryos in cyclic waves whose temporal periodicity corresponds to the formation time of one somite [72,78] (Fig. 9). The *c-hairy1* mRNA and protein product are expressed in a temporally-periodic fashion in individual cells, but since the phase of the oscillator is different at different points along the embryo's axis, the areas of maximal expression sweep along the axis in a periodic fashion.

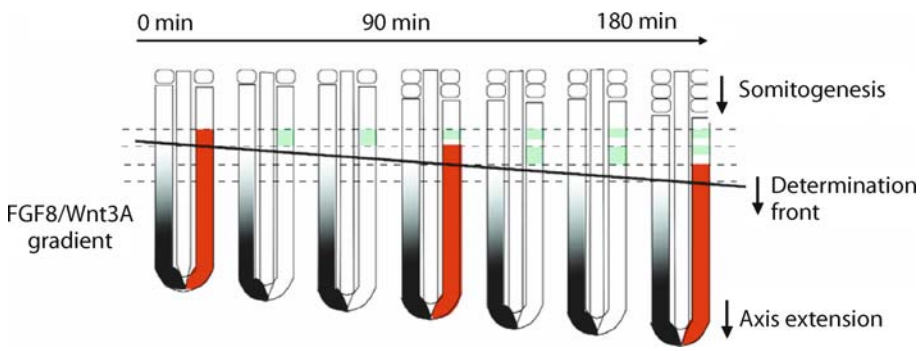

Biological Development and Evolution, Complexity and Self-organization in, Figure 9
Model proposed by Pourquié for segment formation in vertebrates, based on mouse and chick data. A gradient of *FGF8* (see text), shown in *black*, regresses posteriorly during somitogenesis. The anterior boundary of the gradient defines the determination front, which corresponds to the position of the wavefront (*thick black line*). (A coordinately-expressed gradient of *Wnt3A* plays a similar role; Aulehla et al. [2]). Oscillatory (i. e., waxing and waning with developmental stage) expression of *c-hairy1* and related genes is shown in *red*. Expression of genes of the Mesp family, which encode transcription factors involved in somite boundary formation, is shown in *green*. (Reprinted, with permission, from Pourquié [78]; see original paper for details)

Experimental evidence suggests that somite boundaries form when cells which have left a posterior growth zone move sufficiently far away from a source of a diffusible protein known as fibroblast growth factor 8 (FGF8) in the tailbud at the posterior end of the embryo [18]. The FGF gradient thus acts as a "gate" that, when its low end coincides with a particular phase of the segmentation clock, results in formation of a boundary [18,78]. The general features of this mechanism (called the "clock and wavefront" model) were predicted on the basis of dynamical principles two decades before there was any direct evidence for a somitic oscillator [13].

During somitogenesis in the zebrafish a pair of transcription factors known as *her1* and *her7*, which are related to chicken *hairy1* (see above), oscillate in the PSM in a similar fashion, as does the cell surface signaling ligand delta C. Lewis [52] has suggested that *her1* and *her7* constitute an autoregulatory transcription factor gene circuit of the sort treated by Keller [47] (see Sect. "Dynamic Multistability: Cell Differentiation", above), and are the core components of the somitic oscillator in zebrafish. He also hypothesizes that deltaC, whose signaling function is realized by activating the Notch receptors on adjacent cells, is a downstream effector of this oscillation. The two *her* genes negatively regulate their own expression [33,71] and are positively regulated by signaling via Notch [71]. Certain additional experimental results [34] led Lewis to the conclusion that signaling by the Notch pathway, usually considered to act in the determination of cell fate [1] in this case acts to keep cells in the segment-generating growth zone in synchrony [52]. Such synchrony has been experimentally confirmed in chicken embryos [79,94].

Lewis [52] and Monk [58] have independently provided a simple mechanism for the oscillatory expression of the *her1* and *her7* genes, which we briefly summarize here. The model is based on the assumption, mentioned above, that there exists a feedback loop in which the Her1 and Her7 proteins directly bind to the regulatory DNA of their own genes to inhibit transcription. Also incorporated into the model is the recognition that there is always a delay between the initiation of transcription and the initiation of translation, T_m (since it takes time for the mRNA molecule to translocate into the cytoplasm), as well as between the initiation of translation and the emergence of a complete functional protein molecule, T_p (see Fig. 10a).

These ingredients are put into mathematical language in the following way. For a given autoregulatory gene, let $m(t)$ be the number of mRNA molecules in a cell at time t and let $p(t)$ be the number of the corresponding protein molecules. The rate of change of m and p, are then assumed to obey the following equations

$$\frac{\mathrm{d}p(t)}{\mathrm{d}t} = am(t - T_p) - bp(t) \tag{9}$$

$$\frac{\mathrm{d}m(t)}{\mathrm{d}t} = f\left[p(t - T_m)\right] - cm(t)\,. \tag{10}$$

Here the constants b and c are the decay rates of the protein and its mRNA, respectively, a is the rate of production of new protein molecules and $f(p)$ is the rate of production of new mRNA molecules. The function $f(p)$ is assumed to be a decreasing function of the amount of protein (for its form see Lewis [52]. The results of simulations turned out to be quite insensitive to the specific form of $f(p)$).

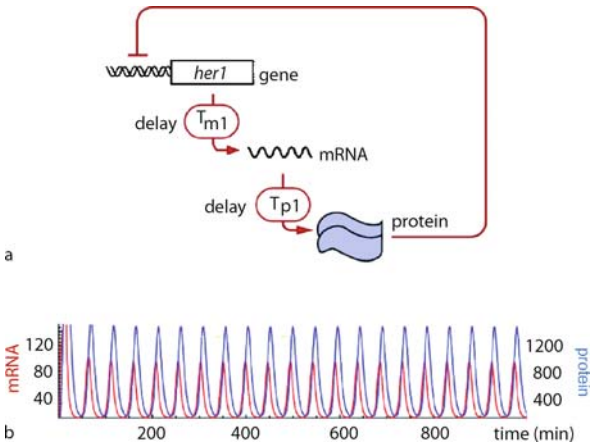

Biological Development and Evolution, Complexity and Self-organization in, Figure 10

Cell autonomous gene expression oscillator for zebrafish somitogenesis. a Molecular control circuitry for a single gene, *her1*, whose protein product acts as a homodimer to inhibit *her1* expression. b Computed behavior for the system in a (defined by Eqs. (9) and (10)) in terms of the number of mRNA molecules per cell in *red* and protein molecules in *blue*. Parameter values were chosen appropriate for the *her1* homodimer oscillator based on experimental results. (Reprinted from Lewis [52] with permission from Elsevier; see original paper for details)

The above delay differential equations were numerically solved for *her1* and *her7* (for which Lewis was able to estimate the values of all the model parameters in Eqs. (9) and (10) using experimental results). The solutions indeed exhibit sustained oscillations in the concentration of Her1 and Her7, with the predicted periods close to the observed ones (Fig. 10b). In a subsequent study Lewis and his coworkers have experimentally confirmed the main features of this oscillatory model for segmentation [26].

Reaction-Diffusion Mechanisms: Body Axis Formation

The nonuniform distribution of any chemical substance, whatever the mechanism of its formation, can clearly provide spatial information to cells. For at least a century embryologists have considered models for pattern formation and its regulation that employ diffusion gradients [25]. Only in the last decade, however, has convincing evidence been produced that this mechanism is utilized in early development. The prime evidence comes from studies of mesoderm *induction*, a key event preceding gastrulation in the frog *Xenopus*. Nieuwkoop [69] originally showed that mesoderm (the middle of the three germ layers in the three-layered gastrula – the one that gives rise to muscle, skeletal tissue, and blood) only appeared when tissue from the upper half of an early embryo ("animal cap") was juxtaposed with tissue from the lower half of the embryo ("vegetal pole"). By themselves, animal cap and vegetal pole cells, respectively, only produce ectoderm, which gives rise to skin and nervous tissue, and endoderm, which gives rise to the intestinal lining. Later it was found that several released, soluble, factors of the TGF-β protein superfamily and the FGF protein family could substitute for the inducing vegetal pole cells [28]. Both TGF-β [55] and FGFs [9] can diffuse over several cell diameters.

While none of this proves beyond question that simple diffusion of such released signal molecules (called "morphogens") between and among cells, rather than some other, cell-dependent mechanism, actually establishes the gradients in question, work by Lander and co-workers, among others [49], has shown that diffusion-like processes, mediated by secreted molecules interacting with cell surfaces and ECMs, are likely to be involved in much developmental patterning. Analysis of such effects can best be performed by considering generalized reaction-diffusion systems, which we now describe (see [21] for additional details).

The rate of change in the concentrations of n interacting molecular species (c_i, $i = 1, 2, \ldots, n$) is determined by their reaction kinetics and can be expressed in terms of ordinary differential equations

$$\frac{\mathrm{d}c_i}{\mathrm{d}t} = F_i(c_1, c_2, \ldots, c_n) \,. \tag{11}$$

The explicit form of the functions F_i in Eq. (11) depends on the details of the reactions. Spatial inhomogeneities also cause time variations in the concentrations even in the absence of chemical reactions. If these inhomogeneities are governed by diffusion, then in one spatial dimension,

$$\frac{\partial c_i}{\partial t} = D_i \frac{\partial^2 c_i}{\partial x^2} \,. \tag{12}$$

Here D_i is the diffusion coefficient of the ith species. In general, both diffusion and reactions contribute to the change in concentration and the time dependence of the c_i's is governed by reaction-diffusion equations

$$\frac{\partial c_i}{\partial t} = D_i \frac{\partial^2 c_i}{\partial x^2} + F_i(c_1, c_2, \ldots, c_n) \,. \tag{13}$$

Reaction-diffusion systems exhibit characteristic parameter-dependent bifurcations (the "Turing instability"), which are thought to serve as the basis for pattern formation in several embryonic systems, including butterfly wing spots, stripes on fish skin, distribution of feathers on the skin of birds, the skeleton of the vertebrate limb, and the primary axis of the developing vertebrate embryo [21].

Gastrulation in the frog embryo is initiated by the formation of an indentation, the "blastopore", through which the surrounding cells invaginate, or tuck into, to the hollow blastula. Spemann and Mangold [91] discovered that the anterior blastopore lip constitutes an *organizer*: a population of cells that directs the movement of other cells. The action of the Spemann–Mangold organizer ultimately leads to the formation of the notochord, the rod of connective tissue that first defines the anteroposterior body axis, and to either side of which the somites later appear (see Sect. "Biochemical Oscillations: Segmentation", above). These investigators also found that an embryo with an organizer from another embryo at the same stage transplanted at some distance from its own organizer would form two axes, and conjoined twins would result. Other classes of vertebrates have similarly acting organizers.

A property of this tissue is that if it is removed, adjacent cells differentiate into organizer cells and take up its role. This indicates that one of the functions of the organizer is to suppress nearby cells with similar potential from exercising it. This makes the body axis a partly self-organizing system. The formation of the body axis in vertebrates also exhibits another unusual feature: while it takes place in an apparently symmetrical fashion, with the left and right sides of the embryo seemingly equivalent to one another, at some point the symmetry is broken. Genes such as *nodal* and *lefty* start being expressed differently on the two sides of the embryo [90], and the whole body eventually assumes a partly asymmetric morphology, particularly with respect to internal organs, such as the heart.

Turing [99] first demonstrated that reaction-diffusion systems like that represented in Eq. (12) will, with appropriate choice of parameters and boundary conditions, generate self-organizing patterns, with a particular propensity to exhibit symmetry breaking across more than one axis. Using this class of models, Meinhardt [56] has presented an analysis of axis formation in vertebrates and the breaking of symmetry around these axes.

The first goal a model of axis formation has to accomplish is to generate an organizer de novo. For this high local concentrations and graded distributions of signaling molecules are needed. This can be accomplished by the coupling of a self-enhancing feedback loop acting over a short range with a competing inhibitory reaction acting over a longer range. The simplest system that can produce such a molecular pattern in the $x - y$ plane consists of a positively autoregulatory activator (with concentration $A(x, y; t)$) and an inhibitor (with concentration $I(x, y; t)$). The activator controls the production of the inhibitor, which in turn limits the production of the activator. This process can be described by the following reaction-diffusion system [56]

$$\frac{\partial A}{\partial t} = D_A \left(\frac{\partial^2 A}{\partial x^2} + \frac{\partial^2 A}{\partial y^2} \right) + s \frac{A^2 + I_A}{I\left(1 + s_A A^2\right)} - k_A A \quad (14)$$

$$\frac{\partial I}{\partial t} = D_I \left(\frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2} \right) + sA^2 - k_I I + I_I \quad (15)$$

The A^2 terms specify that the feedback of the activator on its own production and that of the inhibitor in both cases is non-linear. The factor $s > 0$ describes positive autoregulation, the capability of a factor to induce positive feedback on its own synthesis. This may occur by purely chemical means ("autocatalysis"), which is the mechanism assumed by Turing [99] when he first considered systems of this type. More generally, in living tissues, positive autoregulation occurs if a cell's exposure to a factor it has secreted causes it to make more of the same factor [100]. The inhibitor slows down the production of the activator (i. e., the $1/I$ factor in the second term in Eq. (14)). Both activator and inhibitor diffuse (i. e., spread) and decay with respective diffusion (D_A, D_I) and rate constants (k_A, k_I). The small baseline inhibitor concentrations, I_A and I_I can initiate activator self-enhancement or suppress its onset, respectively, at low values of A. The factor s_A, when present, leads to saturation of positive autoregulation. Once the positive autoregulatory reaction is under way, it leads to a stable, self-regulating pattern in which the activator is in dynamic equilibrium with the surrounding cloud of the inhibitor.

The various organizers and subsequent inductions leading to symmetry breaking, axis formation and the appearance of the three germ layers in amphibians during gastrulation, can all, in principle, be modeled by the reaction-diffusion system in Eqs. (14) and (15), or by the coupling of several such systems. The biological relevance of such reaction-diffusion models depends on whether there exist molecules that can be identified as activator-inhibitor pairs. Meinhardt's model starts with a default state, which consists of ectoderm. Patch-like activation generates the first "hot spot", the *vegetal pole organizer*, which induces endoderm formation (simulation in Fig. 11a). A candidate for the diffusible activator in the corresponding self-enhancing loop for endoderm specification is the TGF-β-like factor Derriere, which activates the VegT transcription factor [97].

VegT expression remains localized to the vegetal pole, but not because of lack of competence of the surrounding cells to produce VegT [11]. These findings provide circumstantial evidence for the existence of the inhibitor required by the reaction-diffusion model. Subsequently, a second feedback loop forms a second hot spot in the

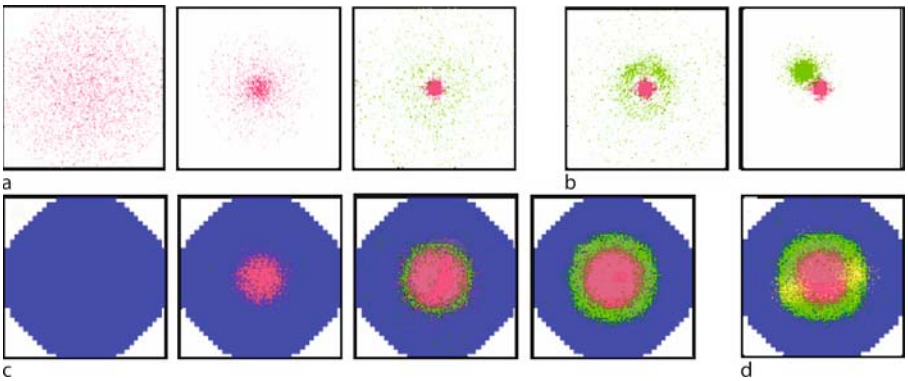

Biological Development and Evolution, Complexity and Self-organization in, Figure 11
Pattern formation in the reaction-diffusion model of Meinhardt. a Induction of the vegetal pole organizer. *Left*: The interaction of an autocatalytic activator, the TGF-β-like factor Derriere (*red*), with a long-ranging inhibitor (whose production it controls, and which in turn, limits the activator's production), creates an unstable state in an initially near-uniform distribution of the substances (inhibitor not shown). *Middle* and *right*: Small local elevation in the activator concentration above steady-state levels triggers a cascade of events governed by Eqs. (14)–(15): further increase of the activator due to autocatalysis, spread of the concomitantly produced surplus of inhibitor into the surrounding, where it suppresses activator production (*middle*), and the establishment of a new stable state, in which the activator maximum (hot spot) is in a dynamic equilibrium with the surrounding cloud of inhibitor (*right*). The *right* panel also shows the near-uniform distribution of the activator (*green*) in the second reaction-diffusion system discussed in b. b Induction of the Nieuwkoop center. Once the first hot spot has formed, it activates a second self-enhancing feedback loop. The production of the activator (*green*) in this reaction is inhibited by the vegetal pole organizer itself. As a consequence, the Nieuwkoop center is displaced from the pole. c Zonal separation of ectoderm, endoderm and mesoderm. The competition of several positive feedback loops assures that in one cell only one of these loops is active. As the results indicate, reaction-diffusion systems can produce not only spot-like organizers, but also zones. Endoderm (*red*) forms as shown from a default ectodermal state (*blue*). The mesodermal zone (*green*; which forms by the involvement of the *FGF*/*brachyury* feedback loop) develops a characteristic width by an additional self-inhibitory influence in Eqs. (14) and (15). d Induction of the Spemann–Mangold organizer. The activation of an organizing region (yellow) within the still symmetric mesodermal zone would imply a competition over relatively long distances. In such a case, Eqs. (14) and (15) lead to the occurrence of two organizing regions, a biologically unacceptable result. The strong asymmetry shown in b prevents this and suggests a reason for the existence of the Nieuwkoop center. (Reprinted from Meinhardt [56] with permission from the University of the Basque Country Press; see original paper for details)

vicinity of the first, in the endoderm. This is identified with the "Nieuwkoop center", a second organizing region, which appears in a specific quadrant of the blastula. A candidate for the second self-enhancing loop is FGF together with Brachyury [87]. Interestingly, the inhibitor for this loop is hypothesized to be the first loop itself (i. e., the vegetal pole organizer), which acts as local repressor for the second. As a result of this local inhibitory effect, the Nieuwkoop center is displaced from the pole (simulation in Fig. 11b). With the formation of the Nieuwkoop center the spherical symmetry of the embryo is broken. In Meinhardt's model this symmetry breaking "propagates" and thus forms the basis of further symmetry breakings, in particular the left-right asymmetry.

By secreting several diffusible factors, the Nieuwkoop center induces the formation of the Spemann–Mangold organizer [31] (If the second feedback loop, responsible for the Nieuwkoop center is not included in the model, two Spemann–Mangold organizers appear symmetrically with respect to the animal-vegetal axis and no symmetry breaking occurs). The organizer triggers gastrulation, in the course of which the germ layers acquire their relative positions and the notochord forms. This long thin structure marks the midline of the embryo, which itself inherits organizer function and eventually establishes the primary (AP) embryonic axis. A simulation of midline formation, based on Meinhardt's model is shown in Fig. 12.

Evolution of Developmental Mechanisms

Many key gene products that participate in and regulate multicellular development emerged over several billions of years of evolution in a world containing only single-celled organisms. Less than a billion years ago multicellular organisms appeared, and the gene products that had evolved in the earlier period were now raw material to be acted upon by physical and chemical-dynamic mechanisms on a more macroscopic scale [62]. The mechanisms described in the previous sections – chemical multistability, chemical oscillation, changes in cell adhesive differentials (both between cell types and across the surface of individual cells), and reaction-diffusion-based symmetry

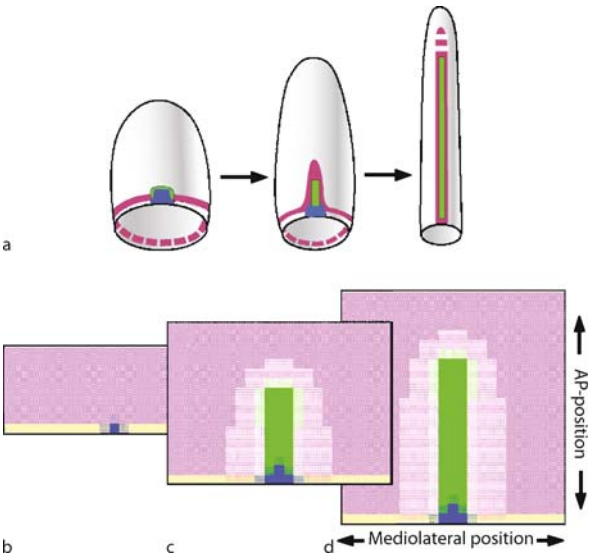

Biological Development and Evolution, Complexity and Self-organization in, Figure 12
Formation of the midline and enfolding of the anteroposterior (AP) axis according to the reaction-diffusion model of Meinhardt. a Schematics of the hypothesized processes involved in axis formation. Cells close to the blastopore (*red* annulus) move towards the Spemann–Mangold organizer (*blue*). The spot-like organizer redirects the cells in its vicinity: initially they are drawn to the organizer, but then lose their attraction to it, so that they leave as a unified stripe-like band. Stem cells in the organizer region may contribute to this band to form the most central element of the midline (*green*). Such highly coordinated cellular motion and strong positional specification along the enfolding AP axis requires the simultaneous action of several feedback loops. b–d Simplified simulation of the scenario described in a. A reaction-diffusion system tuned to make stripes (*green*; compare with zone formation in Fig. 11) is triggered by the organizer (*blue*). The organizer itself is the result of a self-enhancing system, activated in a spot-like manner. (In b–d the blastopore is shown in yellow.) Repulsion between the spot system (the organizer) and the stripe system (the notochord) causes the elongation of the latter. Saturation in self-enhancement (due to the factor s_A in Eq. (14)) ensures that the stripe system does not disintegrate into individual patches and thus establishes the midline. This, in turn, acts as a sink for a ubiquitously produced substance (*pink*), which could be the product of *BMP-4*. The local concentration of this substance (*shades of pink*) is a measure of the distance from the midline. This simulation is simplified in that the actual movement of cells toward the organizer is not considered; instead, the midline elongates by the addition of lines of new cells next to the blastopore. (a redrawn, and b–d reprinted, from Meinhardt [56] with permission from the University of the Basque Country Press; see original paper for details)

breaking, would have caused these ancient multicellular aggregates to take on a wide range of biological forms. Any set of physicochemical activities that generated a new form in a reliable fashion within a genetically uniform population of cell aggregates would have constituted a primitive developmental mechanism.

But any developmental process that depends solely on physical mechanisms of morphogenesis would have been variable in its outcome – subject to inconstant external parameters such as temperature and the chemical microenvironment [67]. Modern-day embryos develop in a more rigidly programmed fashion: the operation of cell type- and pattern-generating physical and chemical-dynamic mechanisms is constrained and focused by hierarchical systems of coordinated gene activities. These developmental-genetic programs are the result of eons of molecular evolution that occurred mainly *after* the origination of multicellularity.

The relation between physical mechanisms of morphogenesis and hierarchical developmental programs can be appreciated by considering a major puzzle in the field of evolutionary developmental biology ("EvoDevo"): the disparity between the mechanisms of segment formation in "short germ-band" insects such as beetles and "long germ-band" insects such as fruit flies. As we will see, one solution to this conundrum touches on all categories of multicomponent dynamical behaviors discussed in earlier sections.

Similarly to somitogenesis in vertebrates (see Sect. "Biochemical Oscillations: Segmentation"), in short germ-band insects [75], as well as in other arthropods, such as the horseshoe crab [40], segmental primordia are added in sequence from a zone of cell proliferation ("growth zone") (Fig. 13). In contrast, in long germ-band insects, such as the fruit fly *Drosophila*, a series of chemical stripes (i. e., parallel bands of high concentration of a molecule) forms in the embryo, which at this stage is a *syncytium*, a large cell with single cytoplasmic compartment containing about 6000 nuclei arranged in a single layer on the inner surface of the plasma membrane [50]. These stripes are actually alternating, evenly-spaced bands of transcription factors of the "pair-rule" class. The pair-rule genes include *even-skipped*, *fushi tarazu*, and *hairy*, which is the insect homolog of the *c-hairy1* gene expressed in a periodic fashion during vertebrate somitogenesis (see Sect. "Biochemical Oscillations: Segmentation"). When cellularization (the enclosure of each nucleus and nearby cytoplasm in their own complete plasma membrane) takes place shortly thereafter, the cells of the resulting blastoderm will have periodically-distributed identities, determined by the particular mix of transcription factors they have incorporated. The different cell states are later transformed into states of differential adhesivity [36], and morphological segments form as a consequence.

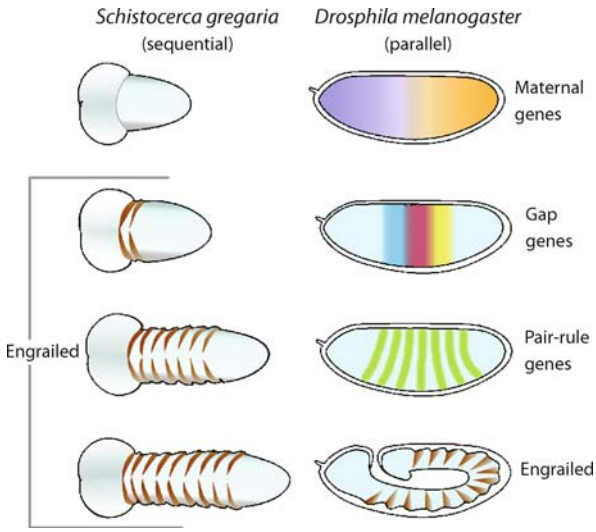

Biological Development and Evolution, Complexity and Self-organization in, Figure 13

Comparison of sequential and parallel segmentation modes in short and long germ-band insects. Left: In short germ-band insects (the embryo of the grasshopper, *Schistocerca gregaria*, is shown schematically) one or groups of a few segments appear in succession. Brown stripes indicate expression of a segment polarity gene such as *engrailed*. With further development additional segments appear sequentially from a zone of proliferation that remains posterior to the most recently-added segment. Right: In long germ-band insects (the embryo of the fruit-fly, *Drosophila melanogaster*, is shown) gradients of maternal gene products (e. g., *bicoid* and *nanos*) are present in the egg before cellularization of the blastoderm (see text). As development proceeds, the maternal gene products induce the expression of gap genes (e. g., *hunchback*, *Krüppel*), the products of which, in turn, induce the expression of pair-rule genes in a total of 14 stripes (e. g., *eve*, *fushi tarazu*, *hairy*). Each pair-rule gene is expressed in a subset of the stripes: *eve*, for example, is expressed in seven alternating stripes, shown in *green*. The pair-rule gene products provide a prepattern for the alternating bands of nuclear expression of segment polarity genes. Once cellularization has occurred (*bottom panel*), these are distributed similarly to that in short germ-band embryos. (*Schistocerca* series after Patel [73]; *Drosophila* series after Ingham [36]. Reprinted, with permission, from [21]

The formation of overt segments in both short (e. g., grasshoppers [75], Patel [74] and beetles [6]) and long germ-band insects [45] requires the prior expression of a stripe of the product of the *engrailed* (*en*) gene, a transcription factor, in the cells of the posterior border of each of the presumptive segments. In *Drosophila*, the positions of the engrailed stripes are largely determined by the activity of the pair-rule genes *even-skipped* (*eve*) and *fuhsi tarazu* (*ftz*), which, as mentioned above, exhibit alternating, complementary seven stripe patterns prior to the formation of the blastoderm [23,35].

On theoretical [4] and experimental [51] grounds it has long been recognized that the kinetic properties that give rise to a chemical oscillation (Sect. "Biochemical Oscillations: Segmentation"), can, when one or more of the components is diffusible, also give rise to standing or traveling spatial periodicities of chemical concentration (Sect. "Reaction-Diffusion Mechanisms: Body Axis Formation"). (See Fig. 14 for a simple dynamical system that exhibits temporal oscillation or standing waves, depending on whether or not diffusion is permitted). This connection between oscillatory and reaction-diffusion mechanisms can potentially unify the different segmentation mechanisms found in short and long germ-band insects. This would be quite straightforward if the *Drosophila* embryo were patterned by a reaction-diffusion system, which can readily give rise to a series of chemical standing waves ("stripes").

While the stripe patterns of the pair-rule genes in *Drosophila* indeed have the distinct appearance of being produced by a reaction-diffusion system (Fig. 13), at least one of the stripes is generated instead by a complex set of interactions among transcription factors in the syncytial embryo. The formation of *eve* stripe number 2 requires the existence of sequences in the *eve* promotor that switch on the *eve* gene in response to a set of spatially-distributed morphogens that under normal circumstances have the requisite values only at the stripe 2 position (Fig. 15) [88,89]. In particular, these promoter sequences respond to specific combinations of products of the "gap" genes (e. g., *giant*, *knirps*, the embryonically-produced version of *hunchback*). These proteins are transcription factors that are expressed in a spatially-nonuniform fashion and act as activators and competitive repressors of the pair-rule gene promoters (also see discussion of the Keller model in Sect. "Dynamic Multistability: Cell Differentiation"). The patterned expression of the gap genes, in turn, is controlled by the responses of their own promoters to particular combinations of products of "maternal" genes (e. g., *bicoid*, *staufen*), which are distributed as gradients along the embryo at even earlier stages (Fig. 13). As the category name suggests, the maternal gene products are deposited in the egg during oogeneis.

While the expression of *engrailed* along the posterior margin of each developing segment is a constant theme during development of arthropods, the expression patterns of pair-rule genes is less well-conserved over evolution [15]. The accepted view is that the short germ-band "sequential" mode is the more ancient way of making segments, and that the long germ-band "simultaneous" mode seen in *Drosophila*, which employs pair-rule stripes, is more recently evolved [53].

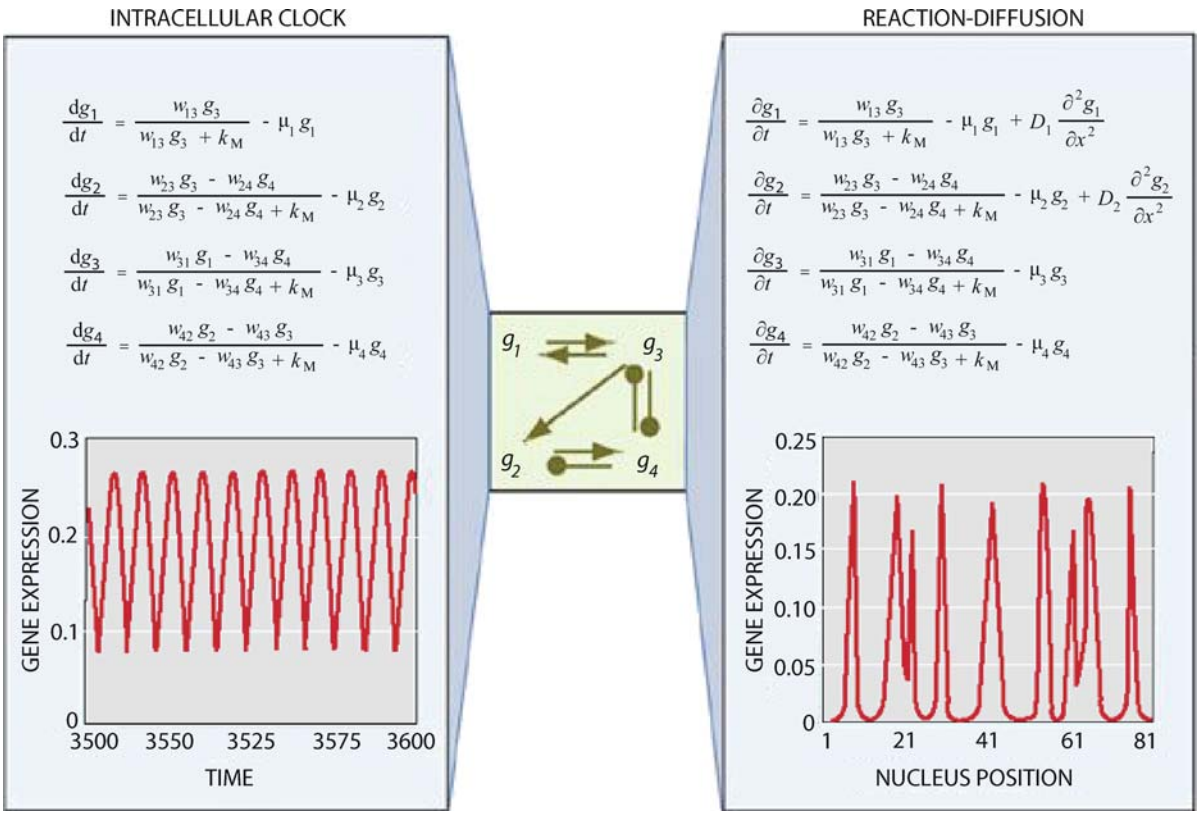

Biological Development and Evolution, Complexity and Self-organization in, Figure 14

Example of a network that can produce (for the same parameter values) sequential stripes when acting as an intracellular biochemical clock in a one-dimensional cellularized blastoderm with a posterior proliferative zone, and simultaneously forming stripes when acting in a one-dimensional diffusion-permissive syncytium. The network is shown in the central box. Arrows indicate positive regulation; lines terminating in circles, negative regulation. In the upper boxes, the equations governing each of the two behaviors are shown. The four genes involved in the central network diagram, as well as their levels of expression, are denoted by g_1, g_2, g_3, and g_4. In the reaction-diffusion case, g_1 and g_2 can diffuse between nuclei (note that the two sets of equations differ only in the presence of a diffusion term for the products of genes 1 and 2). The lower boxes indicate the levels of expression of gene 2 for the two systems. For the intracellular clock the *x*-axis represents time, and the indicated time variation of the gene's expression level refers to a single nucleus. In the reaction-diffusion system this axis represents space, the distance along a chain of 81 nuclei. In the pattern shown on the *right*, the initial condition consisted of all gene product levels set to zero except gene 1 in the central of 81 nuclei, which was assigned a small value (the exact quantity did not affect the pattern). All parameters (except for D_1 and D_2) had the same value when generating the two patterns (for details see Salazar-Ciudad et al. [85], from which the figure was redrawn). Figure modified and reprinted, with permission, from [21]

No intracellular oscillations have thus far been identified during segmentation of invertebrates, unlike the case in vertebrates such as the mouse, chicken, and zebrafish. However, the sequential appearance of gene expression stripes from the posterior proliferative zone of short germ-band insects and other arthropods such as spiders, has led to the suggestion that these patterns in fact arise from a segmentation clock like that found to control vertebrate somitogenesis [95].

Based on the experimental findings, theoretical considerations and evolutionary inferences described above, it has been hypothesized that the ancestor of *Drosophila* generated its segments by a reaction-diffusion system [60,85], built upon the presumed chemical oscillator underlying short germ-band segmentation. Modern-day *Drosophila* contains (retains, according to the hypothesis) the ingredients for this type of mechanism. Specifically, in the syncytial embryo several of the pair-rule proteins (e. g., eve, ftz) diffuse over short distances among the cell nuclei that synthesize their mRNAs, and positively regulate their own synthesis [30,38].

The hypothesized evolutionary scenario can be summarized as follows: the appearance of the syncytial mode of embryogenesis converted the chemical oscillation-de-

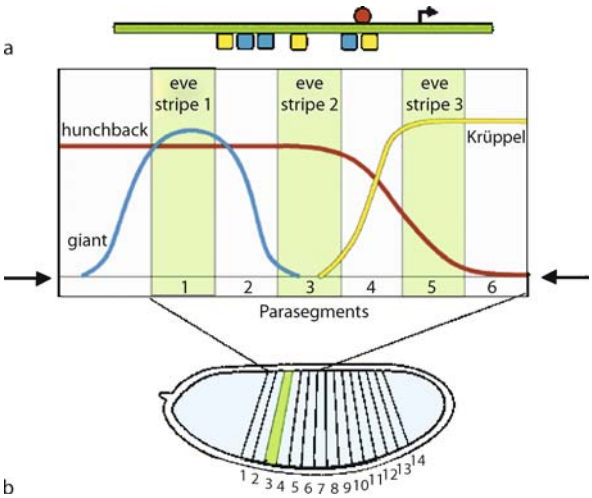

Biological Development and Evolution, Complexity and Self-organization in, Figure 15

a Schematic representation of a portion of the *even-skipped* gene, including the promoter and transcription start site. Contained within the promoter is a subregion ("stripe 2 enhancer") responsible for the expression of the second eve stripe. This sequence contains binding sites for gap gene–class transcription factors that positively (hunchback, *red*) or negatively (giant, *blue*; Krüppel, *orange*) regulate *eve* expression. b Illustration of *Drosophila* syncytial blastula, showing distribution of giant, hunchback and Krüppel proteins in region of first six prospective parasegments. At the position of prospective parasegment 3 (the posterior half of the last head segment plus the anterior half of the first thoracic – upper body – segment) the levels of the activator hunchback is high and those of the inhibitors giant and Krüppel are low. This induces eve stripe 2. To either side of this position giant and Krüppel are high, restricting expression of eve to a narrow band. (a based on Small et al. [88]; b after Wolpert [104], 2002. Reprinted, with permission, from [21])

pendent temporal mechanism found in the more ancient short germ-band insects into the spatial standing wave mechanism seen in the more recently evolved long germ-band forms. The pair-rule stripes formed by the proposed reaction-diffusion mechanism would have been equivalent to one another. That is, they would have been generated by a single mechanism acting in a spatially-periodic fashion, not by stripe-specific molecular machinery (see above). Thus despite suggesting an underlying physical connection between modern short germ-band segmentation and segmentation in the presumed anscestor of long germ-band forms, this hypothesis introduces a new puzzle of its own: Why does modern-day *Drosophila* not use a reaction-diffusion mechanism to produce its segments?

Genes are always undergoing random mutation, but morphological change does not always track genetic change. Particularly interesting are those cases in which the outward form of a body plan or organ does not change, but its genetic "underpinning" does. One example of this is seen in the role of the transcription factor sloppy-paired (slp) in the beetle *Tribolium* and *Drosophila*. In relation to, but in contrast with the paired (prd) protein, which activates *engrailed* and *wingless* in the same odd-even parasegmental register in both species, slp has an opposite parasegmental register in *Tribolium* and *Drosophila* [8].

This situation can result from a mechanism that is initially "plastic", i. e., having variable morphological outcomes despite utilizing the same genes (as physically based mechanisms would typically be), followed by a particular kind of natural selection, termed "canalizing selection" by Waddington [101] (see also [86]). Canalizing selection will preserve those randomly acquired genetic alterations that happen to enhance the reliability of a developmental process. Development would thereby become more complex at the molecular level, but correspondingly more resistant ("robust") to external perturbations or internal noise that could disrupt non-reinforced physical mechanisms of determination.

If the striped expression of pair-rule genes in the ancestor of modern *Drosophila* was generated by a reaction-diffusion mechanism, this inherently variable developmental system would have been a prime candidate for canalizing evolution. The elaborate systems of multiple promoter elements responsive to pre-existing, nonuniformly distributed molecular cues (e. g., maternal and gap gene products), seen in *Drosophila* is therefore not inconsistent with this pattern having originated as a reaction-diffusion process.

In light of the discussion in the previous paragraphs, a tentative answer to why modern-day *Drosophila* does not use a reaction-diffusion mechanism to produce its segments is that such pattern-forming systems are inherently unstable to environmental and random genetic changes, and would therefore, under pressure of natural selection, have been replaced, or at least reinforced, by more hierarchically-organized genetic control systems. We do not know, at present, whether the short germ-band mode of segmentation is similarly hierarchically reinforced.

This evolutionary hypothesis has been examined computationally in a simple physical model by Salazar-Ciudad and coworkers [84]. The model consists of a fixed number of nuclei arranged in a row within a syncytium. Each nucleus has the same genome (i. e., the same set of genes), and the same *epigenetic system* (i. e., the same activating and inhibitory relationships among these genes). The genes in these networks specify receptors or transcription factors that act within the cells that produce them, or paracrine factors that diffuse between cells (Fig. 16). These genes in-

teract with each other according to a set of simple rules that embody unidirectional interactions in which an upstream gene activates a downstream one, as well as reciprocal interactions, in which genes feed back (via their products) on each other's activities. This formalism was based on a similar one devised by Reinitz and Sharp [80], who considered the specific problem of segmentation in the *Drosophila* embryo.

One may ask whether a system of this sort, with particular values of the gene-gene coupling and diffusion constants, can form a spatial pattern of differentiated cells. Salazar-Ciudad and coworkers performed simulations on systems containing 25 nuclei, and a pattern was considered to arise if after some time different nuclei stably expressed one or more of the genes at different levels. The system was isolated from external influences ("zero-flux boundary conditions" were used, that is the boundaries of the domain were impermeable to diffusible morphogens), and initial conditions were set such that at $t = 0$ the levels of all gene products had zero value except for that of an arbitrarily chosen gene, which had a non-zero value in the nucleus at the middle position.

Isolated single nuclei, or isolated patches of contiguous nuclei expressing a given gene, are the one-dimensional analogue of isolated stripes of gene expression in a two-dimensional sheet of nuclei, such as those in the *Drosophila* embryo prior to cellularization [84]. Regardless of whether a particular system initially gives rise to a pattern, if within the model it evolves by random changes in the values of the gene-gene coupling constants it may acquire or lose pattern-forming ability over subsequent "generations". Salazar-Ciudad and coworkers therefore used such systems to perform a large number of computational "evolutionary experiments". Here we will describe a few of these that are relevant to the questions raised by the evolution of segmentation in insects, discussed above.

It had earlier been determined that the core mechanisms responsible for all stable patterns fell into two non-overlapping topological categories [83]. These were referred to as "emergent" and "hierarchical". Mechanisms in which reciprocal positive and negative feedback interactions give rise to the pattern, are emergent. In contrast, those that form patterns by virtue of the unidirectional influence of one gene on the next, in an ordered succession ("feed-forward" loops [54]), as in the "maternal gene induces gap gene induces pair-rule gene" scheme described above for early steps of *Drosophila* segmentation, are hierarchical. Emergent systems are equivalent to excitable dynamical systems, like the transcription factor networks discussed in Sect. "Dynamic Multistability: Cell Differentiation" and the reaction-diffusion systems discussed in Sect. "Reaction-Diffusion Mechanisms: Body Axis Formation".

In their computational studies of the evolution of developmental mechanisms the investigators found that randomly-chosen emergent networks were much more likely than randomly-chosen hierarchical networks to generate complex patterns, i. e., patterns with three or more (one-dimensional) stripes. This was taken to suggest that the evolutionary origination of complex forms, such as segmented body plans, would have been more readily achieved in a world of organisms in which dynamical (e. g., multistable, reaction-diffusion, oscillatory) mechanisms were also active [84] and not just mechanisms based on genetic hierarchies.

Following up on these observations, Salazar-Ciudad and co-workers performed a set of computational studies on the evolution of developmental mechanisms after a pattern had been originated. First they identified emergent and hierarchical networks that produced a particular pattern – e. g., three "stripes" [84]. (Note that patterns themselves are neither emergent nor hierarchical – these terms apply to the mechanisms that generate them). They next asked if given networks would "breed true" phenotypically, despite changes to their underlying circuitry. That is, would their genetically-altered "progeny" exhibit the same pattern as the unaltered version? Genetic alterations in these model systems consisted of point mutations (i. e., changes in the value of a gene-gene coupling constants), duplications, recombinations (i. e., interchange of coupling constant values between pairs of genes), and the acquisition of new interactions (i. e., a coupling constant that was initially equal to zero was randomly assigned a small positive or negative value).

It was found that hierarchical networks were much less likely to diverge from the original pattern (after undergoing simulated evolution as described) than emergent networks [84]. That is to say, a given pattern would be more robust (and thus evolutionarily more stable) under genetic mutation if it were generated by a hierarchical, than an emergent network. Occasionally it was observed that networks that started out as emergent were converted into hierarchical networks with the same number of stripes. The results of Salazar-Ciudad and coworkers on how network topology influences evolutionary stability imply that these "converted" networks would produce the original pattern in the face of further genetic evolution. Recall that this is precisely the scenario that was hypothesized above to have occurred during the evolution of *Drosophila* segmentation [84].

Subject to the caveats to what is obviously a highly schematic analysis, the possible implications of these com-

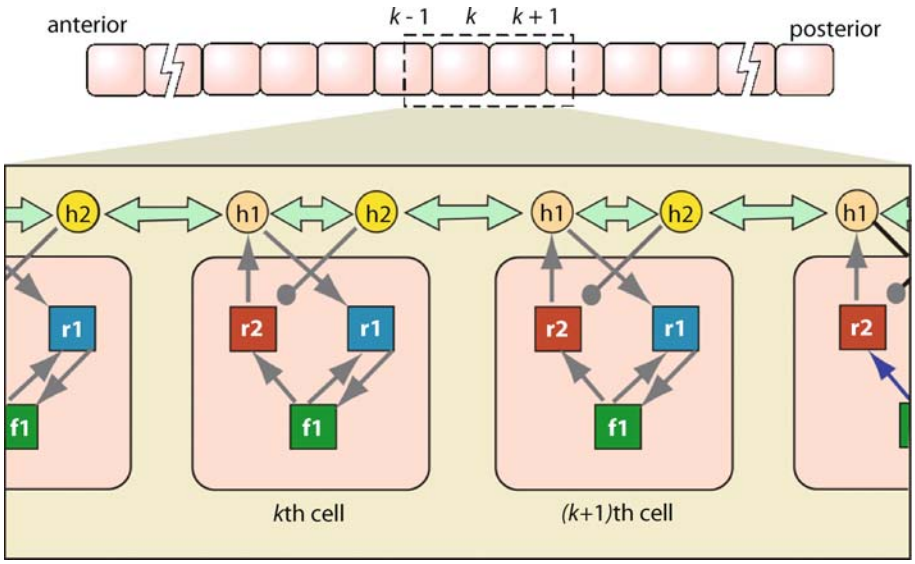

Biological Development and Evolution, Complexity and Self-organization in, Figure 16
Schematic of the gene-gene interactions in the model of Salazar-Ciudad et al. [84]. A line of cells is represented at the *top*. *Below*, types of genes are illustrated. Genes whose products act solely on or within the cells that produce them (receptors: *r*, transcription factors: *f*) are represented by squares; diffusible factors (paracrine factors or hormones: h) that pass between cells and enable genes to affect one another's activities, are represented by circles. Activating and inhibitory interactions are denoted by *small arrows* and *lines terminating in circles*, respectively. *Double-headed green arrows* denote diffusion. (After Salazar-Ciudad et al. [84]) Reprinted, with permission, from [21]

putational experiments for the evolution of segmentation in long germ-band insects are the following: (i) if the ancestral embryo indeed generated its seven-stripe pair-rule protein patterns by a reaction-diffusion mechanism, and (ii) if this pattern was sufficiently well-adapted so as to provide a premium on breeding true, then (iii) genetic changes that preserved the pattern but converted the underlying network from an emergent one to a hierarchic one (as seen in present-day *Drosophila*) would have been favored.

As this discussion of the short germ-band/long germ-band puzzle suggests, developmental mechanisms can change over the course of evolution while their morphological outcomes remain constant. While forms (few tissue layers, segments, tubes, and so forth) remain simple and reminiscent of structures that can be generated by generic physical processes, the associated gene-expression networks can become increasingly more complex. This is because the latter tend to evolve so as to bring about a developmental result in a fashion protected against perturbation from external and internal variability, including, in the latter case, mutation and noise.

The evolutionary changes that will lead to increasingly canalized or robust outcomes can include genetic redundancies [70,103] such as duplication of developmental control genes [32] and multiplication of their regulatory elements [88,89], as well as the utilization of chaperone proteins as "phenotypic capacitors" [81]. The effect of this mode of evolution is to convert organisms into more stable types that are less morphologically plastic than ones molded by relatively unconstrained physical mechanisms (Fig. 17).

Future Directions

While all living systems are part of the physical world and thus must obey its physical laws, it is not self-evident that all biological problems can be adequately addressed by the concepts of physics as we know them today. On a fundamental level this is because whereas the laws of physics have not changed during the lifetime of the Earth, living systems have. Evolution is the hallmark of life, but has little relevance for physics. As living systems evolved, some lineages transformed from single cells into multicellular organisms with a myriad of different interacting components organized on multiple levels. Such systems are considerably more complex than even the most elaborate materials studied by physicists. Thus, on the practical level, the tools of present day physics do not make it possible to study a fully developed adult organism with all its parts or even the subcellular economy of intracellular biochemical networks.

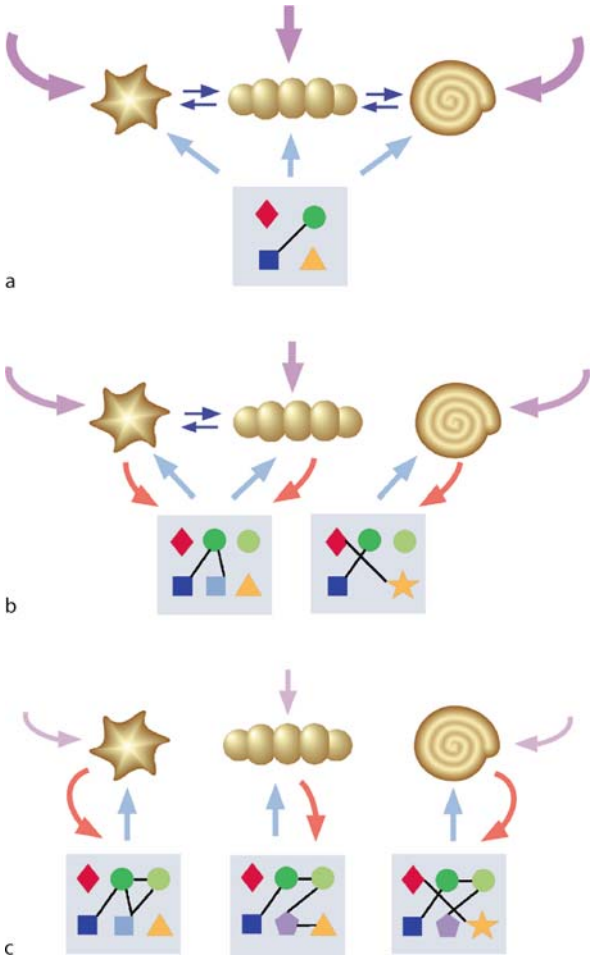

Biological Development and Evolution, Complexity and Self-organization in, Figure 17
Schematic representation of evolutionary partitioning of a morphologically plastic ancestral organism into distinct morphotypes associated with unique genotypes. a A hypothetical primitive metazoan is shown with a schematic representation of its genome in the box below it. Specific genes are shown as colored geometric objects; interactions between them by lines. Determinants of the organism's form include the materials provided by expression of its genes (*light blue arrows*) and the external environment, including physical causes (*purple arrows*) acting on its inherent physical properties. At this stage of evolution the organism is highly plastic, exhibiting several condition-dependent forms that are mutually interconvertible (*dark blue arrows*). b Descendents of organism in a after some stabilizing evolution. Gene duplication, mutation, etc. have led to non-interbreeding populations that are biased toward subsets of the original morphological phenotypes. Determinants of form are still gene products and the physical environment, but the effect of the latter has become attenuated (*smaller, fainter purple arrows*) as development has become more programmatic. There is also causal influence of the form on the genotype (*orange arrows*), exerted over evolutionary time, as ecological establishment of forms filters out those variant genotypes that are not compatible with the established form. Some morphotypes remain interconvertible at this stage of evolution. c Modern organisms descended from those in b. Further stabilizing evolution has now led to each morphotype being uniquely associated with its own genotype. Physical causation is even more attenuated. Note that in this idealized example the forms have remained unchanged while the genes and mechanisms for generating the forms have undergone extensive evolution. (From Newman et al. [67]; with permission from the University of the Basque Country Press)

The developing embryo, however, considered at the appropriate level of resolution, is not beyond physical comprehension of its main episodes and transitions. Each of the processes discussed in the previous sections, multistability of dynamical systems, differential adhesion and interfacial tension, chemical oscillation and symmetry-breaking by reaction-diffusion interaction, have their counterparts in the world of nonliving chemically and mechanically excitable [57] "soft matter" [16]. These descriptions of the behavior of matter at an intermediate scale (i. e., between molecules and bulk materials), represent elegant simplifications of highly complex many-body interactions. They are based on models developed by physicists using sophisticated mathematical techniques and systems-level approaches (nonlinear dynamics, and scaling and renormalization group theory, to name a few), which, for the most part, were unavailable before the mid-20th century.

Embryos are approachable by these physical principles because they are, to a major extent, chemically and mechanically excitable soft matter. Developing tissues consist of cells containing complex biochemical networks which are capable of switching between a limited number of stationary or oscillatory states. Their cells are mobile with respect to one another, giving them liquid-like properties, and they secrete diffusible signal molecules, permitting them to become spatially heterogeneous by straightforward physical means.

As we have seen, the forms generated by self-organizing physical processes of soft matter will become increasingly regulated by genetic hierarchies over the course of evolution. But the major structural motifs of the animal body and its organs (hollow, multilayered cell masses, segments, tubes) had their origin more than 500 million years ago [12,61,68]. During this period "before programs" [67], certain molecules of the "developmental-genetic toolkit" [7] – cadherins, Notch and its ligands, secreted molecules such as Wnt, BMP, hedgehog, FGF – though having evolved in the world of single cells, took on, when present in a multicellular context, the role of mediating intermediate-scale physical effects [66]. Once this new set (for organisms) of physical processes was thus mobi-

lized, morphological evolution, unconstrained in this early period by highly integrated programs, could proceed very rapidly. The otherwise puzzling high degree of conservation, in disparate animal forms, of the toolkit genes for morphological development is to be expected under these circumstances [64].

Although multicellular forms are relatively simple, and the physical processes that likely first brought them about comprehensible, the molecular mechanisms of development in modern-day organisms are enormously complex. The challenge for both developmental and evolutionary studies in the future, then, will be to experimentally disentangle and conceptually reintegrate the roles physics and developmental genetics in modern organisms and their ancestors.

Bibliography

Primary Literature

1. Artavanis-Tsakonas S, Rand MD, Lake RJ (1999) Notch signaling: cell fate control and signal integration in development. Science 284:770–6
2. Aulehla A, Wehrle C, Brand-Saberi B, Kemler R, Gossler A, Kanzler B, Herrmann BG (2003) Wnt3a plays a major role in the segmentation clock controlling somitogenesis. Dev Cell 4:395–406
3. Bateson W (1894) Materials for the study of variation. Macmillan, London
4. Boissonade J, Dulos E, DeKepper P (1994) Turing patterns: From myth to reality. In: Kapral R, Showalter K (eds) Chemical waves and patterns. Boston, Kluwer
5. Borisuk MT, Tyson JJ (1998) Bifurcation analysis of a model of mitotic control in frog eggs. J Theor Biol 195:69–85
6. Brown SJ, Patel NH, Denell RE (1994) Embryonic expression of the single Tribolium engrailed homolog. Dev Genet 15:7–18
7. Carroll SB, Grenier JK, Weatherbee SD (2001) From DNA to diversity: molecular genetics and the evolution of animal design. Blackwell Science, Malden
8. Choe CP, Brown SJ (2007) Evolutionary flexibility of pair-rule patterning revealed by functional analysis of secondary pair-rule genes, paired and sloppy-paired in the short-germ insect, Tribolium castaneum. Dev Biol 302:281–94
9. Christen B, Slack J (1999) Spatial response to fibroblast growth factor signalling in Xenopus embryos. Development 126:119–125
10. Cinquin O, Demongeot J (2005) High-dimensional switches and the modelling of cellular differentiation. J Theor Biol 233:391–411
11. Clements D, Friday RV, Woodland HR (1999) Mode of action of VegT in mesoderm and endoderm formation. Development 126:4903–11
12. Conway Morris S (2006) Darwin's dilemma: The realities of the cambrian 'explosion'. Philos Trans Royal Soc Lond B Biol Sci 361:1069–83
13. Cooke J, Zeeman EC (1976) A clock and wavefront model for control of the number of repeated structures during animal morphogenesis. J Theor Biol 58:455–76
14. Davidson EH (2006) The regulatory genome: gene regulatory networks in development and evolution. Elsevier Academic Press, Amsterdam
15. Dawes R, Dawson I, Falciani F et al (1994) Dax, a locust Hox gene related to fushi-tarazu but showing no pair-rule expression. Development 120:1561–72
16. de Gennes PG (1992) Soft matter. Science 256:495–497
17. DeMarais AA, Moon RT (1992) The armadillo homologs beta-catenin and plakoglobin are differentially expressed during early development of Xenopus laevis. Dev Biol 153:337–46
18. Dubrulle J, McGrew MJ, Pourquié O (2001) FGF signaling controls somite boundary position and regulates segmentation clock control of spatiotemporal Hox gene activation. Cell 106:219–32
19. Duguay D, Foty RA, Steinberg MS (2003) Cadherin-mediated cell adhesion and tissue segregation: qualitative and quantitative determinants. Dev Biol 253:309–23
20. Elowitz MB, Leibler S (2000) A synthetic oscillatory network of transcriptional regulators. Nature 403:335–8
21. Forgacs G, Newman SA (2005) Biological physics of the developing embryo. Cambridge University Press, Cambridge
22. Foty RA, Pfleger CM, Forgacs G, Steinberg MS (1996) Surface tensions of embryonic tissues predict their mutual envelopment behavior. Development 122:1611–1620
23. Frasch M, Levine M (1987) Complementary patterns of even-skipped and fushi tarazu expression involve their differential regulation by a common set of segmentation genes in Drosophila. Genes Dev 1:981–95
24. Giancotti FG, Ruoslahti E (1999) Integrin signaling. Science 285:1028–32
25. Gilbert SF (2006) Developmental biology, 8th edn. Sinauer Associates, Sunderland
26. Giudicelli F, Ozbudak EM, Wright GJ, Lewis J (2007) Setting the tempo in development: an investigation of the zebrafish somite clock mechanism. PLoS Biol 5:e150
27. Goodwin BC (1963) Temporal organization in cells; a dynamic theory of cellular control processes. Academic Press, London
28. Green J (2002) Morphogen gradients, positional information, and Xenopus: Interplay of theory and experiment. Dev Dyn 225:392–408
29. Gurdon JB (1988) A community effect in animal development. Nature 336:772–4
30. Harding K, Hoey T, Warrior R et al (1989) Autoregulatory and gap gene response elements of the even-skipped promoter of Drosophila. EMBO J 8:1205–12
31. Harland R, Gerhart J (1997) Formation and function of Spemann's organizer. Annu Rev Cell Dev Biol 13:611–67
32. Holland PW (1999) Gene duplication: past, present and future. Semin Cell Dev Biol 10:541–7
33. Holley SA, Geisler R, Nusslein-Volhard C (2000) Control of her1 expression during zebrafish somitogenesis by a Delta-dependent oscillator and an independent wave-front activity. Genes Dev 14:1678–1690
34. Holley SA, Julich D, Rauch GJ et al (2002) her1 and the notch pathway function within the oscillator mechanism that regulates zebrafish somitogenesis. Development 129:1175–83
35. Howard K, Ingham P (1986) Regulatory interactions between the segmentation genes fushi tarazu, hairy, and engrailed in the Drosophila blastoderm. Cell 44:949–57
36. Ingham PW (1988) The molecular genetics of embryonic pattern formation in Drosophila. Nature. 335:25–34

37. Irvine KD, Wieschaus E (1994) Cell intercalation during Drosophila germband extension and its regulation by pair-rule segmentation genes. Development 120:827–41
38. Ish-Horowicz D, Pinchin SM, Ingham PW et al (1989) Autocatalytic ftz activation and instability induced by ectopic ftz expression. Cell 57:223–232
39. Israelachvili JN (1991) Intermolecular and surface forces. Academic Press, London
40. Itow T (1986) Inhibitors of DNA synthesis change the differentiation of body segments and increase the segment number in horseshoe crab embryos. Roux's Arch Dev Biol 195:323–333
41. Jablonka E, Lamb MJ (1995) Epigenetic inheritance and evolution. Oxford University Press, Oxford
42. Kaneko K (2003) Organization through intra-inter dynamics. In: Müller GB, Newman SA (eds) Origination of Organismal Form: Beyond the Gene in Developmental and Evolutionary Biology. MIT Press, Cambridge, pp 195–220
43. Kaneko K (2006) Life: an introduction to complex systems biology. Springer, New York
44. Kaneko K, Yomo T (1994) Cell division, differentiation and dynamic clustering. Physica D 75:89–102
45. Karr TL, Weir MP, Ali Z et al (1989) Patterns of engrailed protein in early Drosophila embryos. Development 105:605–612
46. Kauffman SA (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor Biol 22:437–67
47. Keller AD (1995) Model genetic circuits encoding autoregulatory transcription factors. J Theor Biol 172:169–85
48. Kerszberg M, Wolpert L (1998) Mechanisms for positional signaling by morphogen transport: a theoretical study. J Theor Biol 191:103–14
49. Lander AD (2007) Morpheus unbound: reimagining the morphogen gradient. Cell 128:245–56
50. Lawrence PA (1992) The making of a fly: the genetics of animal design. Blackwell Scientific Publications, Oxford
51. Lengyel I, Epstein IR (1992) A chemical approach to designing Turing patterns in reaction-diffusion systems. Proc Natl Acad Sci USA 89:3977–9
52. Lewis J (2003) Autoinhibition with transcriptional delay: a simple mechanism for the zebrafish somitogenesis oscillator. Curr Biol 13:1398–408
53. Liu PZ, Kaufman TC (2005) Short and long germ segmentation: unanswered questions in the evolution of a developmental mode. Evol Dev 7:629–46
54. Mangan S, Zaslaver A, Alon U (2003) The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. J Mol Biol 334:197–204
55. McDowell N, Gurdon JB, Grainger DJ (2001) Formation of a functional morphogen gradient by a passive process in tissue from the early Xenopus embryo. Int J Dev Biol 45 (1 Spec No):199–207
56. Meinhardt H (2001) Organizer and axes formation as a self-organizing process. Int J Dev Biol 45(1 Spec No):177–88
57. Mikhailov AS (1990) Foundations of synergetics I. Springer, Berlin
58. Monk NA (2003) Oscillatory expression of Hes1, p53, and NF-kappaB driven by transcriptional time delays. Curr Biol 13:1409–13
59. Morisco C, Seta K, Hardt SE et al (2001) Glycogen synthase kinase 3beta regulates GATA4 in cardiac myocytes. J Biol Chem 276:28586–97
60. Newman SA (1993) Is segmentation generic? BioEssays 15:277–283
61. Newman SA (1994) Generic physical mechanisms of tissue morphogenesis: A common basis for development and evolution. J Evol Bio 7:467–488
62. Newman SA (1998) Epithelial morphogenesis: a physico-evolutionary interpretation. In: Chuong C-M (ed) Molecular Basis of Epithelial Appendage Morphogenesis. Landes, Austin, pp 341–358
63. Newman SA (2003) From physics to development: the evolution of morphogenetic mechanisms. In: Müller GB, Newman SA (eds) Origination of Organismal Form: Beyond the Gene in Developmental and Evolutionary Biology. MIT Press, Cambridge
64. Newman SA (2006) The developmental-genetic toolkit and the molecular homology-analogy paradox. Biol Theory 1:12–16
65. Newman SA (2007) William Bateson's physicalist ideas. In: Laubichler M, Maienschein J (eds) From Embryology to Evo-Devo: a History of Evolutionary Development. Cambridge, MA, MIT Press, pp 83–107
66. Newman SA, Bhat R (2008) Dynamical patterning modules: physico-genetic determinants of morphological development and evolution. Phys Biol 5:15008
67. Newman SA, Forgacs G, Müller GB (2006) Before programs: The physical origination of multicellular forms. Int J Dev Biol 50:289–99
68. Newman SA, Müller GB (2000) Epigenetic mechanisms of character origination. J Exp Zool 288:304–17
69. Nieuwkoop PD (1969) The formation of mesoderm in Urodelean amphibians. I. Induction by the endoderm. Wilhelm Roux' Arch Entw Mech Org 162:341–373
70. Nowak MA, Boerlijst MC, Cooke J, Smith JM (1997) Evolution of genetic redundancy. Nature 388:167–71
71. Oates AC, Ho RK (2002) Hairy/E(spl)-related (Her) genes are central components of the segmentation oscillator and display redundancy with the Delta/Notch signaling pathway in the formation of anterior segmental boundaries in the zebrafish. Development 129:2929–46
72. Palmeirim I, Henrique D, Ish-Horowicz D et al (1997) Avian hairy gene expression identifies a molecular clock linked to vertebrate segmentation and somitogenesis. Cell 91:639–48
73. Patel NH (1994) Developmental evolution: insights from studies of insect segmentation. Science 266:581–590
74. Patel NH, Ball EE, Goodman CS (1992) Changing role of even-skipped during the evolution of insect pattern formation. Nature 357:339–342
75. Patel NH, Kornberg TB, Goodman CS (1989) Expression of engrailed during segmentation in grasshopper and crayfish. Development 107:201–212
76. Pennisi E (2003) A low number wins the GeneSweep pool. Science 300:1484
77. Phillips HM (1969) Equilibrium measurements of embryonic cell adhesiveness: Physical formulation and testing of the differential adhesion hypothesis. Ph D thesis. Johns Hopkins University
78. Pourquié O (2003) The segmentation clock: converting embryonic time into spatial pattern. Science 301:328–30
79. Primmett DR, Norris WE, Carlson GJ et al (1989) Periodic segmental anomalies induced by heat shock in the chick embryo are associated with the cell cycle. Development 105:119–30

80. Reinitz J, Mjolsness E, Sharp DH (1995) Model for cooperative control of positional information in Drosophila by bicoid and maternal hunchback. J Exp Zool 271:47–56
81. Rutherford SL, Lindquist S (1998) Hsp90 as a capacitor for morphological evolution. Nature 396:336–42
82. Sakuma R, Ohnishi Yi Y, Meno C et al (2002) Inhibition of Nodal signalling by Lefty mediated through interaction with common receptors and efficient diffusion. Genes Cells 7:401–12
83. Salazar-Ciudad I, Garcia-Fernandez J, Sole RV (2000) Gene networks capable of pattern formation: from induction to reaction-diffusion. J Theor Biol 205:587–603
84. Salazar-Ciudad I, Newman SA, Solé R (2001) Phenotypic and dynamical transitions in model genetic networks. I. Emergence of patterns and genotype-phenotype relationships. Evol Dev 3:84–94
85. Salazar-Ciudad I, Solé R, Newman SA (2001) Phenotypic and dynamical transitions in model genetic networks. II. Application to the evolution of segmentation mechanisms. Evol Dev 3:95–103
86. Schmalhausen II (1949) Factors of evolution. Blakiston, Philadelphia
87. Schulte-Merker S, Smith JC (1995) Mesoderm formation in response to Brachyury requires FGF signalling. Curr Biol 5:62–7
88. Small S, Blair A, Levine M (1992) Regulation of even-skipped stripe 2 in the Drosophila embryo. EMBO J 11:4047–4057
89. Small S, Kraut R, Hoey T et al (1991) Transcriptional regulation of a pair-rule stripe in Drosophila. Genes Dev 5:827–39
90. Solnica-Krezel L (2003) Vertebrate development: taming the nodal waves. Curr Biol 13:R7–9
91. Spemann H, Mangold H (1924) Über Induktion von Embryonalanlagen durch Implantation artfremder Organisatoren. Wilhelm Roux' Arch Entw Mech Org 100:599–638
92. St Johnston D, Nusslein-Volhard C (1992) The origin of pattern and polarity in the Drosophila embryo. Cell 68:201–19
93. Steinberg MS (1963) Reconstruction of tissues by dissociated cells. Some morphogenetic tissue movements and the sorting out of embryonic cells may have a common explanation. Science 141:401–8
94. Stern CD, Bellairs R (1984) Mitotic activity during somite segmentation in the early chick embryo. Anat Embryol (Berl) 169:97–102
95. Stollewerk A, Schoppmeier M, Damen WG (2003) Involvement of Notch and Delta genes in spider segmentation. Nature 423:863–5
96. Strogatz SH (1994) Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering. Perseus Pub, Cambridge
97. Sun B, Bush S, Collins-Racie L et al (1999) derriere: a TGF-beta family member required for posterior development in Xenopus. Development 126:1467–1482
98. Tsarfaty I, Resau JH, Rulong S, Keydar I, Faletto DL, Vande Woude GF (1992) The met proto-oncogene receptor and lumen formation. Science 257:1258–61
99. Turing AM (1952) The chemical basis of morphogenesis. Phil Trans Royal Soc Lond B 237:37–72
100. Van Obberghen-Schilling E, Roche NS, Flanders KC et al (1988) Transforming growth factor beta-1 positively regulates its own expression in normal and transformed cells. J Biol Chem 263:7741–7746
101. Waddington CH (1957) The Strategy of the Genes. Allen and Unwin, London
102. Winfree AT (1980) The geometry of biological time. Springer, New York
103. Wilkins AS (1997) Canalization: a molecular genetic perspective. BioEssays 19:257–262
104. Wolpert L (2002) Principles of development. Oxford University Press, Oxford New York

Books and Reviews

Meinhardt H (1982) Models of biological pattern formation. Academic, New York
Müller GB, Newman SA (2003) Origination of organismal form: beyond the gene in developmental and evolutionary biology. MIT Press, Cambridge, pp 221–239
Newman SA, Comper WD (1990) 'Generic' physical mechanisms of morphogenesis and pattern formation. Development 110: 1–18

Biological Fluid Dynamics, Non-linear Partial Differential Equations

Antonio DeSimone[1], François Alouges[2], Aline Lefebvre[2]
[1] SISSA-International School for Advanced Studies, Trieste, Italy
[2] Laboratoire de Mathématiques, Université Paris-Sud, Orsay cedex, France

Article Outline

Glossary
Definition of the Subject
Introduction
The Mathematics of Swimming
The Scallop Theorem Proved
Optimal Swimming
The Three-Sphere Swimmer
Future Directions
Bibliography

Glossary

Swimming The ability to advance in a fluid in the absence of external propulsive forces by performing cyclic shape changes.

Navier–Stokes equations A system of partial differential equations describing the motion of a simple viscous incompressible fluid (a Newtonian fluid)

$$\rho\left(\frac{\partial v}{\partial t} + (v\cdot\nabla)v\right) = -\nabla p + \eta\Delta v$$

$$\operatorname{div} v = 0$$

where v and p are the velocity and the pressure in the fluid, ρ is the fluid density, and η its viscosity. For simplicity external forces, such as gravity, have been dropped from the right hand side of the first equation, which expresses the balance between forces and rate of change of linear momentum. The second equation constrains the flow to be volume preserving, in view of incompressibility.

Reynolds number A dimensionless number arising naturally when writing Navier–Stokes equations in non-dimensional form. This is done by rescaling position and velocity with $x^* = x/L$ and $v^* = v/V$, where L and V are characteristic length scale and velocity associated with the flow. Reynolds number (Re) is defined by

$$\mathrm{Re} = \frac{VL\rho}{\eta} = \frac{VL}{\nu}$$

where $\nu = \eta/\rho$ is the kinematic viscosity of the fluid, and it quantifies the relative importance of inertial versus viscous effects in the flow.

Steady Stokes equations A system of partial differential equations arising as a formal limit of Navier–Stokes equations when $\mathrm{Re} \to 0$ and the rate of change of the data driving the flow (in the case of interest here, the velocity of the points on the outer surface of a swimmer) is slow

$$\begin{aligned} -\eta\Delta v + \nabla p &= 0 \\ \operatorname{div} v &= 0\,. \end{aligned}$$

Flows governed by Stokes equations are also called creeping flows.

Microscopic swimmers Swimmers of size $L = 1\,\mu\mathrm{m}$ moving in water ($\nu \sim 1\,\mathrm{mm}^2/\mathrm{s}$ at room temperature) at one body length per second give rise to $\mathrm{Re} \sim 10^{-6}$. By contrast, a 1 m swimmer moving in water at $V = 1\,\mathrm{m/s}$ gives rise to a Re of the order 10^6.

Biological swimmers Bacteria or unicellular organisms are microscopic swimmers; hence their swimming strategies cannot rely on inertia. The devices used for swimming include rotating helical flagella, flexible tails traversed by flexural waves, and flexible cilia covering the outer surface of large cells, executing oar-like rowing motion, and beating in coordination. Self propulsion is achieved by cyclic shape changes described by time periodic functions (*swimming strokes*). A notable exception is given by the rotating flagella of bacteria, which rely on a submicron-size rotary motor capable of turning the axis of an helix without alternating between clockwise and anticlockwise directions.

Swimming microrobots Prototypes of artificial microswimmers have already been realized, and it is hoped that they can evolve into working tools in biomedicine. They should consist of minimally invasive, small-scale self-propelled devices engineered for drug delivery, diagnostic, or therapeutic purposes.

Definition of the Subject

Swimming, i. e., being able to advance in a fluid in the absence of external propulsive forces by performing cyclic shape changes, is particularly demanding at low Reynolds numbers (Re). This is the regime of interest for micro-organisms and micro-robots or nano-robots, where hydrodynamics is governed by Stokes equations. Thus, besides the rich mathematics it generates, low Re propulsion is of great interest in biology (How do microorganism swim? Are their strokes optimal and, if so, in which sense? Have these optimal swimming strategies been selected by evolutionary pressure?) and biomedicine (can small-scale self-propelled devices be engineered for drug delivery, diagnostic, or therapeutic purposes?).

For a microscopic swimmer, moving and changing shape at realistically low speeds, the effects of inertia are negligible. This is true for both the inertia of the fluid and the inertia of the swimmer. As pointed out by Taylor [10], this implies that the swimming strategies employed by bacteria and unicellular organism must be radically different from those adopted by macroscopic swimmers such as fish or humans. As a consequence, the design of artificial microswimmers can draw little inspiration from intuition based on our own daily experience.

Taylor's observation has deep implications. Based on a profound understanding of low Re hydrodynamics, and on a plausibility argument on which actuation mechanisms are physically realizable at small length scales, Berg postulated the existence of a sub-micron scale rotary motor propelling bacteria [5]. This was later confirmed by experiment.

Introduction

In his seminal paper *Life at low Reynolds numbers* [8], Purcell uses a very effective example to illustrate the subtleties involved in microswimming, as compared to the swimming strategies observable in our mundane experience. He argues that at low Re, any organism trying to swim adopting the reciprocal stroke of a scallop, which moves by opening and closing its valves, is condemned to the frustrating experience of not having advanced at all at the end of one cycle.

This observation, which became known as the *scallop theorem*, started a stream of research aiming at finding the simplest mechanism by which cyclic shape changes may lead to effective self propulsion at small length scales. Purcell's proposal was made of a chain of three rigid links moving in a plane; two adjacent links swivel around joints and are free to change the angle between them. Thus, shape is described by two scalar parameters (the angles between adjacent links), and one can show that, by changing them independently, it is possible to swim.

It turns out that the mechanics of swimming of Purcell's three-link creature are quite subtle, and a detailed understanding has started to emerge only recently [4,9]. In particular, the direction of the average motion of the center of mass depends on the geometry of both the swimmer and of the stroke, and it is hard to predict by simple inspection of the shape of the swimmer and of the sequence of movements composing the swimming stroke. A radical simplification is obtained by looking at axisymmetric swimmers which, when advancing, will do so by moving along the axis of symmetry. Two such examples are the three-sphere-swimmer in [7], and the push-me–pull-you in [3]. In fact, in the axisymmetric case, a simple and complete mathematical picture of low Re swimming is now available, see [1,2].

The Mathematics of Swimming

This article focuses, for simplicity, on swimmers having an axisymmetric shape Ω and swimming along the axis of symmetry, with unit vector $\vec{\imath}$. The configuration, or state s of the system is described by $N+1$ scalar parameters: $s = \{x^{(1)}, \ldots, x^{(N+1)}\}$. Alternatively, s can be specified by a position c (the coordinate of the center of mass along the symmetry axis) and by N shape parameters $\xi = \{\xi^{(1)}, \ldots, \xi^{(N)}\}$. Since this change of coordinates is invertible, the generalized velocities $u^{(i)} := \dot{x}^{(i)}$ can be represented as linear functions of the time derivatives of position and shape:

$$(u^{(1)}, \ldots, u^{(N+1)})^t = A(\xi^{(1)}, \ldots, \xi^{(N)})(\dot{\xi}^{(1)}, \ldots, \dot{\xi}^{(N)}, \dot{c})^t \tag{1}$$

where the entries of the $N+1 \times N+1$ matrix A are independent of c by translational invariance.

Swimming describes the ability to change position in the absence of external propulsive forces by executing a cyclic shape change. Since inertia is being neglected, the total drag force exerted by the fluid on the swimmer must also vanish. Thus, since all the components of the total force in directions perpendicular to $\vec{\imath}$ vanish by symmetry, self-propulsion is expressed by

$$0 = \int_{\partial\Omega} \sigma n \cdot \vec{\imath} \tag{2}$$

where σ is the stress in the fluid surrounding Ω, and n is the outward unit normal to $\partial\Omega$. The stress $\sigma = \eta\left(\nabla v + (\nabla v)^t\right) - p\mathrm{Id}$ is obtained by solving Stokes equation outside Ω with prescribed boundary data $v = \bar{v}$ on $\partial\Omega$. In turn, $\bar{v}$ is the velocity of the points on the boundary $\partial\Omega$ of the swimmer, which moves according to (1).

By linearity of Stokes equations, (2) can be written as

$$\begin{aligned} 0 &= \sum_{i=1}^{N+1} \varphi^{(i)}(\xi^{(1)}, \ldots, \xi^{(N)}) u^{(i)} \\ &= A^t \Phi \cdot (\dot{\xi}^{(1)}, \ldots, \dot{\xi}^{(N)}, \dot{c})^t \end{aligned} \tag{3}$$

where $\Phi = (\varphi^{(1)}, \ldots, \varphi^{(N)})^t$, and we have used (1). Notice that the coefficients $\varphi^{(i)}$ relating drag force to velocities are independent of c because of translational invariance. The coefficient of $\dot{c}$ in (3) represents the drag force corresponding to a rigid translation along the symmetry axis at unit speed, and it never vanishes. Thus (3) can be solved for $\dot{c}$, and we obtain

$$\dot{c} = \sum_{i=1}^{N} V_i(\xi^{(1)}, \ldots, \xi^{(N)}) \dot{\xi}^{(i)} = V(\xi) \cdot \dot{\xi}\,. \tag{4}$$

Equation (4) links positional changes to shape changes through shape-dependent coefficients. These coefficients encode all hydrodynamic interactions between Ω and the surrounding fluid due to shape changes with rates $\dot{\xi}^{(1)}, \ldots, \dot{\xi}^{(N)}$.

A stroke is a closed path γ in the space S of admissible shapes given by $[0, T] \ni t \mapsto (\xi^{(1)}, \ldots \xi^{(N-1)})$. Swimming requires that

$$0 \neq \Delta c = \int_0^T \sum_{i=1}^{N} V_i \dot{\xi}^{(i)} \mathrm{d}t \tag{5}$$

i. e., that the differential form $\sum_{i=1}^{N} V_i \, \mathrm{d}\xi^{(i)}$ is not exact.

The Scallop Theorem Proved

Consider a swimmer whose motion is described by a parametrized curve in two dimensions ($N = 1$), so that (4) becomes

$$\dot{c}(t) = V(\xi(t))\dot{\xi}(t)\,, \qquad t \in \mathbb{R}\,, \tag{6}$$

and assume that $V \in L^1(S)$ is an integrable function in the space of admissible shapes and $\xi \in W^{1,\infty}(\mathbb{R}; S)$ is a Lipschitz-continuous and T-periodic function for some $T > 0$, with values in S.

Figure 1 is a sketch representing concrete examples compatible with these hypotheses. The axisymmetric case consists of a three-dimensional cone with axis along $\vec{\imath}$ and opening angle $\xi \in [0, 2\pi]$ (an *axisymmetric octopus*). A non-axisymmetric example is also allowed in this discussion, consisting of two rigid parts (*valves*), always maintaining mirror symmetry with respect to a plane (containing $\vec{\imath}$ and perpendicular to it) while swiveling around a joint contained in the symmetry plane and perpendicular to $\vec{\imath}$ (a *mirror-symmetric scallop*), and swimming parallel to $\vec{\imath}$.

Among the systems that are *not* compatible with the assumptions above are those containing helical elements with axis of rotation $\vec{\imath}$, and capable of rotating around $\vec{\imath}$ always in the same direction (call θ the rotation angle). Indeed, a monotone function $t \mapsto \theta(t)$ is not periodic.

The celebrated "scallop theorem" [8] states that, for a system like the one depicted in Fig. 1, the net displacement of the center of mass at the end of a periodic stroke will always vanish. This is due to the linearity of Stokes equation (which leads to symmetry under time reversals), and to the low dimensionality of the system (a one-dimensional periodic stroke is necessarily reciprocal). Thus, whatever forward motion is achieved by the scallop by closing its valves, it will be exactly compensated by a backward motion upon reopening them. Since the low Re world is unaware of inertia, it will not help to close the valves quickly and reopen them slowly. A precise statement and a rigorous short proof of the scallop theorem are given below.

Theorem 1 *Consider a swimmer whose motion is described by*

$$\dot{c}(t) = V(\xi(t))\dot{\xi}(t)\,, \quad t \in \mathbb{R}\,, \tag{7}$$

with $V \in L^1(S)$. *Then for every* T*-periodic stroke* $\xi \in W^{1,\infty}(\mathbb{R}; S)$, *one has*

$$\Delta c = \int_0^T \dot{c}(t)dt = 0\,. \tag{8}$$

Proof Define the primitive of V by

$$\Psi(s) = \int_0^s V(\sigma)\,\mathrm{d}\sigma \tag{9}$$

so that $\Psi'(\xi) = V(\xi)$. Then, using (7),

$$\begin{aligned}\Delta c &= \int_0^T V(\xi(t))\dot{\xi}(t)\mathrm{d}t \\ &= \int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\Psi(\xi(t))\mathrm{d}t \\ &= \Psi(\xi(T)) - \Psi(\xi(0)) = 0\end{aligned}$$

by the T-periodicity of $t \mapsto \xi(t)$.

Optimal Swimming

A classical notion of swimming efficiency is due to Lighthill [6]. It is defined as the inverse of the ratio between the average power expended by the swimmer during a stroke starting and ending at the shape $\xi_0 = (\xi_0^{(1)}, \ldots, \xi_0^{(N)})$ and the power that an external force would spend to translate the system rigidly at the same average speed $\bar{c} = \Delta c/T$:

$$\mathrm{Eff}^{-1} = \frac{\frac{1}{T}\int_0^T \int_{\partial\Omega} \sigma n \cdot v}{6\pi\eta L\bar{c}^2} = \frac{\int_0^1 \int_{\partial\Omega} \sigma n \cdot v}{6\pi\eta L(\Delta c)^2} \tag{10}$$

where η is the viscosity of the fluid, $L = L(\xi_0)$ is the effective radius of the swimmer, and time has been rescaled to a unit interval to obtain the second identity. The expression in the denominator in (10) comes from a generalized version of Stokes formula giving the drag on a sphere of radius L moving at velocity $\bar{c}$ as $6\pi\eta L\bar{c}$.

Let DN: $H^{1/2}(\partial\Omega) \to H^{-1/2}(\partial\Omega)$ be the Dirichlet to Neumann map of the outer Stokes problem, i. e., the map such that $\sigma n = DNv$, where σ is the stress in the fluid, evaluated on $\partial\Omega$, arising in response to the prescribed velocity v on $\partial\Omega$, and obtained by solving the Stokes problem outside Ω. The expended power in (10) can be written as

$$\int_{\partial\Omega} \sigma n \cdot v = \int_{\partial\Omega} DN(v) \cdot v\,. \tag{11}$$

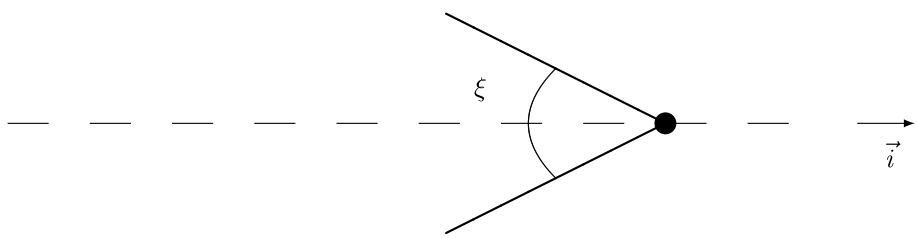

Biological Fluid Dynamics, Non-linear Partial Differential Equations, Figure 1
A mirror-symmetric scallop or an axisymmetric octopus

At a point $p \in \partial\Omega$, the velocity $v(p)$ accompanying a change of state of the swimmer can be written as a linear combination of the $u^{(i)}$

$$v(p) = \sum_{i=1}^{N+1} \mathcal{V}_i(p,\xi)u^{(i)} \tag{12}$$

$$= \sum_{i=1}^{N} \mathcal{W}_i(p,\xi)\dot{\xi}^{(i)} . \tag{13}$$

Indeed, the functions $\mathcal{V}_i$ are independent of c by translational invariance, and (4) has been used to get (13) from the line above.

Substituting (13) in (11), the expended power becomes a quadratic form in $\dot{\xi}$

$$\int_{\partial\Omega} \sigma n \cdot v = (G(\xi)\dot{\xi}, \dot{\xi}) \tag{14}$$

where the symmetric and positive definite matrix $G(\xi)$ is given by

$$G_{ij}(\xi) = \int_{\partial\Omega} DN(\mathcal{W}_i(p,\xi)) \cdot \mathcal{W}_j(p,\xi)\mathrm{d}p . \tag{15}$$

Strokes of maximal efficiency may be defined as those producing a given displacement Δc of the center of mass with minimal expended power. Thus, from (10), maximal efficiency is obtained by minimizing

$$\int_0^1 \int_{\partial\Omega} \sigma n \cdot v = \int_0^1 (G(\xi)\dot{\xi}, \dot{\xi}) \tag{16}$$

subject to the constraint

$$\Delta c = \int_0^1 V(\xi) \cdot \dot{\xi} \tag{17}$$

among all closed curves $\xi : [0,1] \to S$ in the set S of admissible shapes such that $\xi(0) = \xi(1) = \xi_0$.

The Euler–Lagrange equations for this optimization problem are

$$-\frac{\mathrm{d}}{\mathrm{d}t}(G\dot{\xi}) + \frac{1}{2}\begin{pmatrix} \left(\frac{\partial G}{\partial \xi^{(1)}}\dot{\xi}, \dot{\xi}\right) \\ \vdots \\ \left(\frac{\partial G}{\partial \xi^{(N)}}\dot{\xi}, \dot{\xi}\right) \end{pmatrix} + \lambda\left(\nabla_\xi V - \nabla_\xi^t V\right)\dot{\xi} = 0 \tag{18}$$

where $\nabla_\xi V$ is the matrix $(\nabla_\xi V)_{ij} = \partial V_i/\partial \xi_j$, $\nabla_\xi^t V$ is its transpose, and λ is the Lagrange multiplier associated with the constraint (17).

Given an initial shape ξ_0 and an initial position c_0, the solutions of (18) are in fact sub-Riemannian geodesics joining the states parametrized by (ξ_0, c_0) and $(\xi_0, c_0 + \Delta c)$ in the space of admissible states $\mathcal{X}$, see [1]. It is well known, and easy to prove using (18), that along such geodesics $(G(\gamma)\dot{\gamma}, \dot{\gamma})$ is constant. This has interesting consequences, because swimming strokes are often divided into a power phase, where $|G(\gamma)|$ is large, and a recovery phase, where $|G(\gamma)|$ is smaller. Thus, along optimal strokes, the recovery phase is executed quickly while the power phase is executed slowly.

The Three-Sphere Swimmer

For the three-sphere-swimmer of Najafi and Golestanian [7], see Fig. 2, Ω is the union of three rigid disjoint balls $B^{(i)}$ of radius a, shape is described by the distances x and y, the space of admissible shapes is $S = (2a, +\infty)^2$, and the kinematic relation (1) takes the form

$$\begin{aligned} u^{(1)} &= \dot{c} - \frac{1}{3}(2\dot{x} + \dot{y}) \\ u^{(2)} &= \dot{c} + \frac{1}{3}(\dot{x} - \dot{y}) \\ u^{(3)} &= \dot{c} + \frac{1}{3}(2\dot{y} + \dot{x}) . \end{aligned} \tag{19}$$

Consider, for definiteness, a system with $a = 0.05\,\mathrm{mm}$, swimming in water. Calling $f^{(i)}$ the total propulsive force on ball $B^{(i)}$, we find that the following relation among forces and ball velocities holds

$$\begin{pmatrix} f^{(1)} \\ f^{(2)} \\ f^{(3)} \end{pmatrix} = R(x,y)\begin{pmatrix} u^{(1)} \\ u^{(2)} \\ u^{(3)} \end{pmatrix} \tag{20}$$

where the symmetric and positive definite matrix R is known as the resistance matrix. From this last equation, using also (19), the condition for self-propulsion $f^{(1)} + f^{(2)} + f^{(3)} = 0$ is equivalent to

$$\dot{c} = V_x(x,y)\dot{x} + V_y(x,y)\dot{y} , \tag{21}$$

where

$$V_x(x,y) = \frac{Re_c \cdot (e_c \times e_y)}{Re_c \cdot (e_x \times e_y)} \tag{22}$$

$$V_y(x,y) = -\frac{Re_c \cdot (e_c \times e_x)}{Re_c \cdot (e_x \times e_y)} . \tag{23}$$

Moreover, $e_x = (-1,1,0)^t$, $e_y = (0,-1,1)^t$, $e_c = (1/3, 1/3, 1/3)^t$.

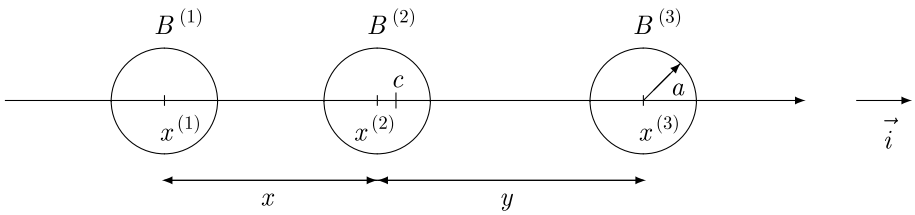

Biological Fluid Dynamics, Non-linear Partial Differential Equations, Figure 2
Swimmer's geometry and notation

Biological Fluid Dynamics, Non-linear Partial Differential Equations, Table 1
Energy consumption (10^{-12} J) for the three strokes of Fig. 3 inducing the same displacement $\Delta c = 0.01$ mm in $T = 1$ s

Optimal stroke	Small square stroke	Large square stroke
0.229	0.278	0.914

Given a stroke $\gamma = \partial\omega$ in the space of admissible shapes, condition (5) for swimming reads

$$0 \neq \Delta c = \int_0^T \left(V_x \dot{x} + V_y \dot{y}\right) \mathrm{d}t = \int_\omega \mathrm{curl}\, V(x, y) \mathrm{d}x\mathrm{d}y \tag{24}$$

which is guaranteed, in particular, if curl V is bounded away from zero. Strokes of maximal efficiency for a given initial shape (x_0, y_0) and given displacement Δc are obtained by solving Eq. (18). For $N = 2$, this becomes

$$-\frac{\mathrm{d}}{\mathrm{d}t}(G\dot{\gamma}) + \frac{1}{2}\begin{pmatrix}(\partial_x G\dot{\gamma}, \dot{\gamma}) \\ (\partial_y G\dot{\gamma}, \dot{\gamma})\end{pmatrix} + \lambda\, \mathrm{curl}\, V(\gamma)\dot{\gamma}^{\perp} = 0 \tag{25}$$

where $\partial_x G$ and $\partial_y G$ stand for the x and y derivatives of the 2×2 matrix $G(x, y)$.

It is important to observe that, for the three-sphere swimmer, all hydrodynamic interactions are encoded in the shape dependent functions $V(x, y)$ and $G(x, y)$. These can be found by solving a two-parameter family of outer Stokes problems, where the parameters are the distances x and y between the three spheres. In [1], this has been done numerically via the finite element method: a representative example of an optimal stroke, compared to two more naive proposals, is shown in Fig. 3.

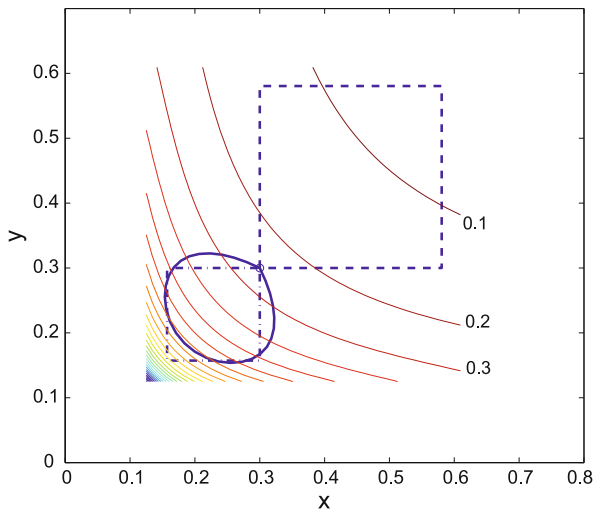

Biological Fluid Dynamics, Non-linear Partial Differential Equations, Figure 3
Optimal stroke and square strokes which induce the same displacement $\Delta c = 0.01$ mm in $T = 1$ s, and equally spaced level curves of curl V. The small circle locates the initial shape $\xi_0 = (0.3\text{ mm}, 0.3\text{ mm})$

Future Directions

The techniques discussed in this article provide a head start for the mathematical modeling of microscopic swimmers, and for the quantitative optimization of their strokes. A complete theory for axisymmetric swimmers is already available, see [2], and further generalizations to arbitrary shapes are relatively straightforward. The combination of numerical simulations with the use of tools from sub-Riemannian geometry proposed here may prove extremely valuable for both the question of adjusting the stroke to global optimality criteria, and of optimizing the stroke of complex swimmers. Useful inspiration can come from the sizable literature on the related field dealing with control of swimmers in a perfect fluid.

Bibliography

Primary Literature

1. Alouges F, DeSimone A, Lefebvre A (2008) Optimal strokes for low Reynolds number swimmers: an example. J Nonlinear Sci 18:277–302
2. Alouges F, DeSimone A, Lefebvre A (2008) Optimal strokes for low Reynolds number axisymmetric swimmers. Preprint SISSA 61/2008/M
3. Avron JE, Kenneth O, Oakmin DH (2005) Pushmepullyou: an efficient micro-swimmer. New J Phys 7:234-1–8

4. Becker LE, Koehler SA, Stone HA (2003) On self-propulsion of micro-machines at low Reynolds numbers: Purcell's three-link swimmer. J Fluid Mechanics 490:15–35
5. Berg HC, Anderson R (1973) Bacteria swim by rotating their flagellar filaments. Nature 245:380–382
6. Lighthill MJ (1952) On the Squirming Motion of Nearly Spherical Deformable Bodies through Liquids at Very Small Reynolds Numbers. Comm Pure Appl Math 5:109–118
7. Najafi A, Golestanian R (2004) Simple swimmer at low Reynolds numbers: Three linked spheres. Phys Rev E 69:062901-1–4
8. Purcell EM (1977) Life at low Reynolds numbers. Am J Phys 45:3–11
9. Tan D, Hosoi AE (2007) Optimal stroke patterns for Purcell's three-link swimmer. Phys Rev Lett 98:068105-1–4
10. Taylor GI (1951) Analysis of the swimming of microscopic organisms. Proc Roy Soc Lond A 209:447–461

Books and Reviews

Agrachev A, Sachkov Y (2004) Control Theory from the Geometric Viewpoint. In: Encyclopaedia of Mathematical Sciences, vol 87, Control Theory and Optimization. Springer, Berlin

Childress S (1981) Mechanics of swimming and flying. Cambridge University Press, Cambridge

Happel J, Brenner H (1983) Low Reynolds number hydrodynamics. Nijhoff, The Hague

Kanso E, Marsden JE, Rowley CW, Melli-Huber JB (2005) Locomotion of Articulated Bodies in a Perfect Fluid. J Nonlinear Sci 15: 255–289

Koiller J, Ehlers K, Montgomery R (1996) Problems and Progress in Microswimming. J Nonlinear Sci 6:507–541

Montgomery R (2002) A Tour of Subriemannian Geometries, Their Geodesics and Applications. AMS Mathematical Surveys and Monographs, vol 91. American Mathematical Society, Providence

Biological Models of Molecular Network Dynamics

Herbert M. Sauro
Department of Bioengineering,
University of Washington, Seattle, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Modeling Approaches
Deterministic Modeling
Stoichiometry Matrix
System Equation
Theoretical Approaches to Modeling
Negative Feedback Systems
FeedForward Systems
Positive Feedback
Future Prospects
Bibliography

Glossary

Deterministic continuous model A mathematical model where the variables of the model can take any real value and where the time evolution of the model is set by the initial conditions.

Stochastic discrete model A mathematical model where the variables of the model take on discrete values and where the time evolution of the model is described by a set of probability distributions.

Definition of the Subject

Understanding the operation cellular networks is probably one of the most challenging and intellectually exciting scientific fields today. With the availability of new experimental and theoretical techniques our understanding of the operation of cellular networks has made great strides in the last few decades. An important outcome of this work is the development of predictive quantitative models. Such models of cellular function will have a profound impact on our ability of manipulate living systems which will lead to new opportunities for generating energy, mitigating our impact on the biosphere and last but not least, opening up new approaches and understanding of important disease states such as cancer and aging.

Introduction

Cellular networks are some of the most complex natural systems we know. Even in a "simple" organism such as *E. coli*, there are at least four thousand genes with many thousands of interactions between molecules of many different sizes [11]. In a human cell the number of interactions is probably orders of magnitude larger. Why all this complexity? Presumably the earliest living organisms were much simpler than what we find today but competition for resources and the need to adapt in unfavorable conditions must have led to the development of sensory and decision-making capabilities above and beyond the basic requirements for life. What we see today in almost all living organisms are complex signaling and genetic networks whose complexity and subtlety is beyond most man-made technological systems [87].

Over the last sixty or so years, biochemists and molecular biologists have identified many of the components in living cells and have traced out many of the interactions that delineate cellular networks. What emerges is a picture that would be familiar to many control engineers

where one finds feedback loops, switches, logic gates, oscillators, filters and a myriad of other control and decision-making devices. This knowledge allows us to rationalize the complexity we find in cellular networks by functionalizing a large network into smaller well defined modules. This work is also spawning a new engineering field called synthetic biology which uses knowledge of modular organization and function to engineer new biological networks [3].

One of the important techniques that has permitted us to gain a deeper understanding of network structure and function is mathematical modeling and new theoretical approaches. These, coupled with considerable advances in measurement technology, has revolutionized our picture of how cellular networks operate. Instead of being relatively static structures as portrayed in many textbooks, cellular networks are, unsurprisingly, highly dynamic systems [42,90].

In this chapter the basic approach to mathematically modeling cellular networks will be introduced followed by a brief review of some of the properties of the commonest network modular designs that are known to exist. It should be noted that computational and theoretical models that help us understand the operational characteristics of different network configurations is now an extremely large field and it would be impossible in one short chapter to cover even a small fraction of the advances that have been made in the last few decades. There is a growing list of excellent reviews and some text books that the interested reader can use to expand on some of the topics covered here, these include excellent works by Alon [1], Arkin [6], Fell [26], Ferrell [29], Goldbetter [32], Heinrich [36], Hofmeyr [38], [40], Kacser [44], Kholodenko [46], Klipp [47], Lauffenburger [50], Sauro [67], Savageau [72], [35] and Tyson [79,80] to name but a few.

Much like human engineered systems, the organization of cellular networks is extremely hierarchical, both in time and space. At the most basic level we have a remarkably simple picture of elemental reactions such as unimolecular and bimolecular reactions. All cellular behavior is built on these elemental reactions. On their own, uni and bimolecular reactions show very simple behavior, it is only when combined into more complex structures that more interesting behavior emerges. For example, the simplest systems we can construct from elementary reactions includes catalysis by enzymes where we already see the emergence of more complex responses in the form of hyperbolic and sigmoid behavior. These more complex responses originate through mass conservation and nonlinearities in the underling bimolecular reactions. The hierarchical range of organization, however, goes much further, ranging all the way from simple enzymatic kinetics to complex signal processing encompassing hundreds of components [56]. Any understanding of network dynamics has to appreciate the different levels of organization and the characteristic dynamic properties that they bring to bear. Cellular networks offer an extremely rich source of signal processing capabilities, much of which is largely unexplored. With the rise of systems biology, this picture is changing.

Modeling Approaches

There are two primary methods for modeling cellular networks, one uses a continuous deterministic approach mainly centered on using ordinary differential equations [88] and the other is based on a discrete stochastic approach, mainly centered on using the chemical master equation [85]. Each approach has its benefits and drawbacks. For systems with large numbers of molecules (> 100), the use of differential equations is sufficient and recommended. This approach assumes that concentrations change continuously in time and space and that stochastic events do not occur.

The use of differential equations in modeling physical phenomena has a long history dating back to at least the time of Newton. As a result, a very large body of knowledge, including analytical as well as computational approaches are available to the modeling community. In consequence, differential equation models in systems biology are very popular with many software tools and analytical techniques available for their study [10]. Unfortunately, some cellular systems, particularly bacterial genetic networks, operate at very low molecule numbers. This makes the use of differential equations models for these systems problematic. At very low molecule numbers the stochastic nature of chemical kinetics comes into play and it is necessary to use a stochastic modeling approach [31]. In some cases, stochastic behavior can be markedly different from that shown in an equivalent deterministic model [60,65]. Unfortunately at this point in time, simulation tools and particular analytical tools are not as well developed which impedes progress in understanding the dynamics at the stochastic level although progress has been more rapid in recent years. In this chapter the focus will be on modeling by differential equations even though some of the interesting complexity we see in biological networks emerges as a result of stochastic behavior.

In later sections, models of genetic networks will be described using a deterministic approach. One could argue that in these cases a stochastic approach is crucial because the number of genes and particular promoter sites

is extremely low (< 4 per gene). The simulations of these models could have been carried out using a stochastic approach, however, the general results obtained would have been the same (the analytical analysis of the stochastic equivalent models could not have been repeated). The question then is when should a stochastic approach be used? There are at least three clear cases when a stochastic approach is critical to modeling biological networks because of differences in predicted behavior: 1) When single events can make a qualitative difference such as shown in the classic work by Arkin et al. on the developmental pathway of phage lambda when infecting *E. coli* [6] or when modeling free radical chemistry [54]; 2) Nonlinearities in the rate laws which can lead to stochastic focusing and thus changes in the mean reaction rates. Such focusing in turn can cause shifts in the steady state solutions or even behavior [60,65]; or 3) When noise can be measured experimentally and an appropriate stochastic model is required to explain the measurements. In many instances both stochastic and deterministic models will yield the same solutions (at least when averaged) but the availability of a mature theory of deterministic dynamics makes deterministic models in such situations much more appealing. When building models, one should always consider the purpose of the model. A model is always a simplification, the reduction of a complicated and messy reality to its essential features. A model is therefore never intended to represent reality in all its detail and efforts for example to build replica *in silico cells* miss the point of modeling. If noise is not of interest or deemed unimportant than a simpler model will suffice and if in doubt one should follow Occam's razor [4].

Deterministic Modeling

Chemical kinetics is at the heart of all dynamic behavior in biological cells. Almost all chemical transformations are mediated by either enzyme or protein factors. Any understanding of network dynamics must therefore start with a foundation in chemical kinetics. To begin with, one must define what is meant by *the rate of reaction* [55].

Consider a general reaction of the form:

$$c_1 A_1 + c_2 A_2 + \cdots \longrightarrow c_n A_n + c_{n+1} A_{n+1} + \cdots$$

where c_i represents the stoichiometric mole numbers and A_i a chemical species. The notation indicates that species A_1, A_2 etc. are transformed with mole numbers c_1 etc. to product species A_n etc. with mole numbers, c_n etc. Normally when expressing a reaction in this form signs are not explicitly given to the stoichiometries, however it is implied that the stoichiometric mole numbers c_i are **positive** for products and **negative** for reactants.

Let us introduce the notion of the *extent of reaction*, indicated by the symbol ξ. We define a change from ξ to $\xi + \mathrm{d}\xi$ in time $\mathrm{d}t$ to mean that $c_1\,\mathrm{d}\xi$ moles of A_1, $c_2\,\mathrm{d}\xi$ moles of A_2 etc., react to form $c_n\,\mathrm{d}\xi$ moles of A_n etc. By this definition we can state that for any component i the following is true in the time interval $\mathrm{d}t$:

$$\frac{\mathrm{d}n_i}{\mathrm{d}t} = c_i \frac{\mathrm{d}\xi}{\mathrm{d}t}$$

where n_i equals the amount in moles of species i. From this relation we *define* the rate of reaction, v_r, to be:

$$v_r \equiv \frac{\mathrm{d}\xi}{\mathrm{d}t} .$$

In other words

$$\frac{\mathrm{d}n_i}{\mathrm{d}t} = c_i v_r .$$

The advantage of employing the extent of reaction is that it allows us to define the rate of reaction independently of the species we use to measure the rate. This convenient property can be expressed as:

$$v_r \equiv \frac{\mathrm{d}\xi}{\mathrm{d}t} = -\frac{1}{c_1}\frac{\mathrm{d}n_1}{\mathrm{d}t} = -\frac{1}{c_2}\frac{\mathrm{d}n_2}{\mathrm{d}t} = \cdots = \frac{1}{c_n}\frac{\mathrm{d}n_n}{\mathrm{d}t} = \frac{1}{c_{n+1}}\frac{\mathrm{d}n_{n+1}}{\mathrm{d}t} = \cdots .$$

Often models are expressed in unit volume so that more commonly modelers will use the intensive form for the reaction rate:

$$\frac{\mathrm{d}C_i}{\mathrm{d}t} = c_i v_j \tag{1}$$

where C_i is the concentration of species i and v_j is the intensive rate of reaction. When building network models, this formalism allows us to separate the connectivity or structure of the network from the rates of reactions.

Mass-Action Kinetics

One of the basic concepts in chemical kinetics is the Law of Mass Action [7]. This states that the rate of a chemical reaction is proportional to the product of the reactant concentrations raised to a given power. For a simple monomolecular reaction such as:

$$A \rightarrow B .$$

The rate of reaction, v, is given by:

$$v = k_1 A$$

where A represents the concentration of reactant. v is the intensive rate of reaction. The rate of change of A is given by (note that the stoichiometry for $A = -1$, see Eq. (1)):

$$\frac{\mathrm{d}A}{\mathrm{d}t} = -k_1 A \,. \tag{2}$$

The units for k_1 are t^{-1} and for the concentration of A, moles l^{-1}. The rate of change of A therefore has units of moles $l^{-1}t^{-1}$. By convention, a rate that is proportional to a reactant raised to a first power is called first-order with respect to the reactant. Similarly, reactants raised to the zeroth power are called zero-order and reactants raised to the power of two are called second-order.

For the general reaction such as:

$$n_1 A + n_2 B + \cdots \longrightarrow m_1 P + m_2 Q + \cdots$$

the rate law has been found through empirical observation to often have the form:

$$v = k_1 A^{n_1} B^{n_2} \ldots$$

where each reactant is raised to the power of its stoichiometry. For example the reaction catalyzed by adenylate kinase that equilibrates ATP, ADP and AMP:

$$2ADP \rightarrow ATP + AMP$$

could have the rate law:

$$v = kADP^2 \,.$$

If the reaction is reversible then the rate law is appended with the reverse rate:

$$v = k_1 A^{n_1} B^{n_2} \cdots - k_2 P^{m_1} Q^{m_2} \cdots$$

$$v = k_1 ADP^2 - k_2 ATP\,AMP \,.$$

The units for the reactant and product terms must be in concentration. The units for the rate constants, k will depend on the exact form of the rate law but must be set to ensure that the rate of reaction is expressed in the units moles $l^{-1}t^{-1}$ if the intensive form is used.

Mass-action kinetics forms the most fundamental and simplest level of description in the cellular hierarchy. Beyond simple mass-action kinetics there is a wide range of aggregate rate laws, probably the most celebrated being the Michaelis–Menten enzyme rate law [20]. These rate laws make assumptions about the time scale of some reaction processes, such as the binding of substrate to enzyme, so that simplifications can be made that allow one to derive simpler kinetic descriptions. Aggregate rate laws display more complex behavior than simple mass-action kinetics, the most important of these are hyperbolic (or saturable) and sigmoid kinetics [75]. These rate laws can be used across the range of cellular processes, including gene expression, protein signaling and metabolic pathways. A variety of generalized rate laws also exist that make the choice of which rate law to use much easier [35].

Stoichiometry Matrix

When describing multiple reactions in a network, it is convenient to represent the stoichiometries in a compact form called the stoichiometry matrix, **N** [5,19,25,41]. This matrix is a m row by n column matrix where m is the number species and n the number reactions. The columns of the stoichiometry matrix correspond to the distinct chemical reactions in the network, the rows to the molecular species, one row per species. Thus the intersection of a row and column in the matrix indicates whether a certain species takes part in a particular reaction or not and, according to the sign of the element, whether there is a net loss or gain of substance, and by the magnitude, the relative quantity of substance that takes part in that reaction. The elements of the stoichiometry matrix thus concerns the relative mole amounts of chemical species that react in a particular reaction; it does not concern itself with the rate of reaction. The use of **N** has a long tradition and is the preferred notation to use [61].

For example, consider the simple chain of reactions which has five molecular species and four reactions. The four reactions are labeled, v_1 to v_4.

$$S_1 \xrightarrow{v_1} S_2 \xrightarrow{v_2} S_3 \xrightarrow{v_3} S_4 \xrightarrow{v_4} S_5 \,.$$

The stoichiometry matrix for this simple system is given by:

$$\mathbf{N} = \begin{array}{c} \\ \\ \\ \\ \\ \end{array}\begin{array}{c} \begin{array}{cccc} v_1 & v_2 & v_3 & v_4 \end{array} \\ \begin{bmatrix} -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array} \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \end{array} \tag{3}$$

Entries in the stoichiometry matrix are computed as follows. Given a species S_i and reaction v_j, the corresponding entry in the stoichiometry matrix at ij is given by the total stoichiometry for S_i on the product side minus the total

stoichiometry of S_i on the reactant side. Thus, considering species S_1 in reaction v_1, we note that the total stoichiometry on the product size is zero (no S_1 molecules are formed on the product side) and the total stoichiometry of S_1 on the reactant side is $+1$. Subtracting one from the other $(0-1)$ we obtain -1, which is entered into the stoichiometry matrix. This rather long-winded approach to computing stoichiometries avoids errors that arise when a species occurs as both reactant and product. For example, for the more complex reaction such as:

$$3A \longrightarrow 2A + B$$

the stoichiometry entry for A is $(2-3)$ that is -1, since the stoichiometry for A on the product side is two and on the reactant side is three.

The stoichiometry matrix is an important element in modeling cellular networks. It represents the connectivity of the network without reference to kinetics or regulation. Even without kinetic information, the stoichiometry matrix contains much interesting information and certain modeling approaches, such as flux balance analysis, make extensive use of the stoichiometry matrix to determine flux conservation relationships from which one can estimate fluxes through a given network [28,59,76].

System Equation

The central equation in deterministic modeling is the system equation (4). The system equation takes into account both the kinetics of the individual reactions as well as the structure of the network. When models of the time evolution of the molecular species by differential equations are built, the stoichiometry matrix is combined with the vector of reaction rates to obtain the *system equation*:

$$\frac{\mathrm{d}\mathbf{S}}{\mathrm{d}\mathbf{t}} = \mathbf{N}\mathbf{v}(\mathbf{S}) \tag{4}$$

where $\mathbf{N}$ is the $m \times n$ stoichiometry matrix and $\mathbf{v}$ is the n dimensional rate vector, whose ith component gives the rate of reaction i as a function of the species concentrations.

If we look again at the model depicting the simple chain of reactions (Eq. (3)), the system equation can be written as:

$$\frac{\mathrm{d}\mathbf{S}}{\mathrm{d}\mathbf{t}} = \mathbf{N}\mathbf{v} = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} .$$

Two particular solutions of the system equation are of interest, time course evolution of molecular species concentrations and the steady state. The steady state is defined as the solution(s) to the system equation when the rates of change of species is zero, Eq. (5)

$$\mathbf{N}\mathbf{v}(\mathbf{S}) = 0 . \tag{5}$$

Redundancies

The system equation (4) will often contain redundant equations. These redundant equations arise from conserved moieties in the network. Conserved moieties are molecular subgroups conserved during the evolution of a network [62]. The total amount of a particular moiety in a network is time invariant and is determined solely by the initial conditions imposed on the system. A typical example of a conserved moiety in a computational model is the conservation of adenine nucleotide, that is when the total amount of ATP, ADP and AMP is constant during the evolution of the system. Other examples include NAD/NADH, phosphorylation/dephosphrylation cycles, and so on. It should be noted that conserved cycles are a model time-scale abstraction because in reality all cycles turn over. However, over the relatively short-time scales used in modeling studies conserved cycles have important biological effects.

Figure 1 illustrates the simplest possible network which displays a conserved moiety, in this case the total mass, $S_1 + S_2$ is constant during the evolution of the network. Such conservation laws can have dramatic effects on the dynamical properties of models. For example, the hyperbolic behavior in the simple Michaelis–Menten rate law can be attributed to the conservation law between free enzyme and enzyme-substrate complex but more subtle effects, such as the emergence of new regulatory properties [53], can be attributed to moiety conserved cycles.

The effect of conserved moieties on the system equation 4 is to reduce the number of differential equations by the number of conservation laws in the model. In a model-

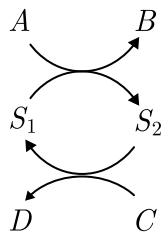

Biological Models of Molecular Network Dynamics, Figure 1
Simple Conserved cycle

ing environment, it is therefore customary to divide the set of molecular species into two groups, the dependent ($\mathbf{S}_d$) and independent set ($\mathbf{S}_i$) [37]. This division is a function of the number and kind of conserved cycles in the network. If the model has no conserved cycles then the dependent set is empty and the size of the independent set equals the number of molecular species in the model. For details on how to compute $\mathbf{S}_d$ and $\mathbf{S}_i$ the reader can consult Sauro and Ingalls [66] and Vallabhajosyula [82]. In many cases it is critical that this separation is made. For simple time course simulations the separation is not so important, but for most other analyses it is very important and for stiff integration methods highly desirable. The reason is that many numerical methods, including the stiff integrators, use the Jacobian matrix [77] as part of the numerical calculation. If the separation is not carried out, the Jacobian becomes singular which renders many analyses numerically unstable if not impossible to carry out. In addition, the separation is also very useful because it enables the number of differential equations to be reduced in number and thereby improve computational efficiency.

Many simulators and mathematical analyses do not actually solve the system equation (4) directly but instead solve the following modified system:

$$\begin{aligned} \mathbf{S}_d &= \mathbf{L}_0\mathbf{S}_i + \mathbf{T} \\ \frac{\mathrm{d}\mathbf{S}_i}{\mathrm{d}\mathbf{t}} &= \mathbf{N}_R\mathbf{v}(\mathbf{S}_i, \mathbf{S}_d)\,. \end{aligned} \tag{6}$$

In these equations, $\mathbf{S}_i$ is the vector of independent species, $\mathbf{S}_d$, the vector of dependent species, $\mathbf{L}_0$ the link matrix, $\mathbf{T}$ the total mass vector, $\mathbf{N_R}$ the reduced stoichiometry matrix and $\mathbf{v}$ the rate vector. This modified equation constitutes the most general expression for a differential equation-based model [35,39,61]. The symbolism used in equation 6 is the standard notation used by many in the Systems Biology community.

Finally, it should be noted that conservation analysis is not confined to deterministic models but is equally applicable to stochastic models and could be used to improve the computational efficiencies of stochastic model solvers.

Theoretical Approaches to Modeling

One way to study the properties of molecular networks is to build computer simulation models. Simulation is often the only recourse when the system equations that govern the dynamics are nonlinear and no analytical solution is possible. However, certain theoretical approaches allow models to be studied analytically if we confine ourselves to small perturbations and their effects. Such types of analyses are very common in the physical sciences and engineering [14]. In the late 1960s there developed a small perturbation approach for molecular networks, called Biochemical Systems Theory, metabolic control analysis or more simply, control analysis. This approach was developed independently by three different groups [26,34,36,44,69,72] and has since been expanded to deal with arbitrary network structures, time dependent changes and more recently it has been extended into the frequency domain which showed that the approach was identical to classical control analysis used by engineers. A brief description of the theory will be given here.

Control analysis is a mathematical technique that allows us to describe some property of a molecular network, such as a steady state concentration or flux, in terms of the properties of the individual parts. In control analysis, system properties are described using control coefficients, of particular interest are the concentration and flux control coefficients. These are defined as the fractional change in concentration or flux relative to the fractional change in some system parameter, usually an enzyme or gene activity:

$$\begin{aligned} C_E^J &= \frac{\mathrm{d}J}{\mathrm{d}E}\frac{E}{J} \\ C_E^S &= \frac{\mathrm{d}S}{\mathrm{d}E}\frac{E}{S}\,. \end{aligned}$$

High control coefficients indicate that the system parameter (E) has a significant effect on the concentration or flux. To make the analysis simpler, control coefficients are also usually measured at steady state. Therefore the control coefficient, C_E^S, measures the effect of E on the steady state concentration of S. Concentration control coefficients are of particular interest in understanding the dynamics of gene and protein regulatory networks where the notion of a flux is not so important. Flux control coefficients on the other hand are of particular interest when studying the dynamics of metabolic pathways where the flux might represent the consumption of an important nutrient or the export of some valuable commodity such as a drug [63].

Control analysis also defines a property of a network part, the elasticity, a local property. Elasticities describe how a reaction rate is affected by changes in the immediate environment of the enzyme or gene that is catalyzing the reaction. If there are multiple factors that can affect a process then there will be multiple elasticities, one for each factor. Often these factors are simply the concentrations of substrates and products, but they could also be regulators, pH, ion concentrations, etc. Elasticities are defined as follows:

$$\varepsilon_S^v = \frac{\partial v}{\partial S}\frac{S}{v}\,.$$

Note that the elasticities are defined using a partial derivative whereas the control coefficients are defined using total derivatives, this reflects the local and system characteristics of each of these coefficients.

The real power of control analysis is its ability to relate these two coefficients together. There are various ways to accomplish this but here the approach using the control theorems will be described [27]. The theorems of control analysis describe various relationships among the coefficients, two in particular are significant for this discussion, namely the summation and connectivity theorems.

Two summation theorems of interest include the concentration and flux summation theorems, they can be stated as follows:

$$\sum^{n} C^{J}_{E_i} = 1$$
$$\sum^{n} C^{S}_{E_i} = 0$$

where the summation of over all steps in the pathway. Operationally the concentration summation theorem states that increasing the activity of all enzymes and/or genes in a network will result in no change to the metabolite or protein concentrations. The flux summation makes a similar statement except that it indicates that the steady state flux increases in proportion to the activity changes in the enzymes and/or genes. The theorems, and particularly generalizations of them [35], have important implications for metabolic engineers. The generalized summation theorems can be used to indicate which steps in a pathway should be changed in order to increase or decrease the output of a particular pathway commodity [21,73,74].

The connectivity theorem is more subtle and relates control coefficients to elasticities:

$$\sum C^{J}_{E_i} \varepsilon^{v_i}_{S} = -1$$
$$\sum C^{S}_{E_i} \varepsilon^{v_i}_{S} = 0\,.$$

The sum is over all reactions and/or genes that the molecular species, S can influence [15,44]. From these theorems (next section) it is possible to write down equations that express control coefficients in terms of elasticities [27,68]. This allows one to understand how properties related to the parts of a network contribute to the properties of the whole network. In the next section a simple example will be given that illustrates this process.

Negative Feedback Systems

Negative feedback in molecular networks is probably one of the most celebrated regulatory mechanisms in biology. There have been numerous discussions on the role of feedback in biochemical networks ever since the discovery by Umbarger [81] of feedback inhibition in the isoleucine biosynthesis pathway and the feedback inhibition of aspartate transcarbamylase in *E. coli* by Yates and Pardee [89].

Probably the most extensive mathematical analysis of biochemical feedback was conducted by Savageau [69,70,72] and Burns and Kacser [13,44] and Othmer and Tyson [57,78] in the 1970s and Dibrov et al. in the early 1980s [22]. More recently, Cinquin and Demongeot have published an interesting review on the roles of feedback in biological systems [18].

To begin the discussion let us consider the simplest possible system, a linear chain of two reactions. These reactions could be two metabolic enzymes in sequence or a gene expressing a protein and the protein undergoing degradation. Figure 2 illustrates a model of simple gene expression where the gene product, P, negatively influences its own production. If our reaction steps were two metabolic enzymes in sequence, the negative inhibition would be generated by simple product inhibition. In either case what properties might such as simple system have? First one needs to construct a mathematical model of our system. Since there is only one floating species in the model, protein P, there will only be one differential equation. It will be assumed that the input concentration, I is fixed and that the protein product degrades into an infinite volume, both constraints will ensure that the system will have a steady state. This is a common feature of all models of molecular networks. Without these constraints the network would simply evolve to thermodynamic equilibrium which is neither interesting nor relevant to the study of biological networks. For the negative feedback one can assume a simple Hill-like inhibition, the model equation can therefore be written as:

$$\frac{\mathrm{d}P}{\mathrm{d}t} = \frac{IV_{\max}}{(1 + (P/K)^h)} - k_1 P\,.$$

$V_{\max}$ is the maximal rate of expression, h is the Hill coefficient, k_1 the degradation rate and K the inhibition constant. There are two properties that can be investigated, how long does it take for the network to reach steady state and once at steady state how robust is the steady state to parameter variation?

In order to examine the properties of this network two strategies can be used, the network can be compared with and without negative feedback or small changes can be made to the strength of the feedback and the effect observed.

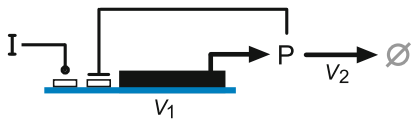

Biological Models of Molecular Network Dynamics, Figure 2
Simple Negative feedback in a genetic circuit. *P* represents the protein product from gene expression v_1. v_2 represents protein degradation. *I* represents the input signal that can activate gene expression. Product, *P*, binds to an operator on the gene repressing gene expression. Bars represent inhibition; round ends, activation. Similar conventions apply to all subsequent figures

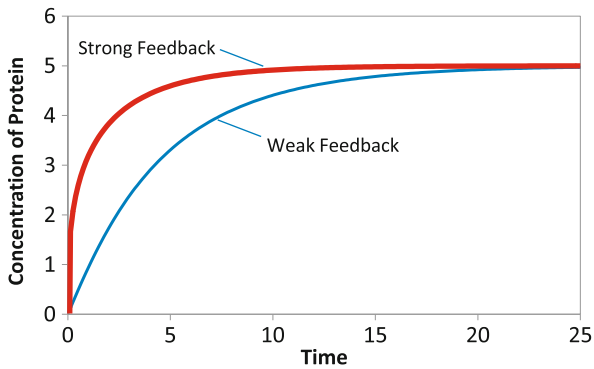

Biological Models of Molecular Network Dynamics, Figure 3
Rise times for a simple negative feedback circuit with weak and strong negative feedback. For a fair comparison, both systems have been adjusted so that they both reach the same steady state. Note that the rise time for the circuit with strong negative feedback is much faster. $k_1 = 0.1$, $I = 1$, $P = 0$ at $t = 0$, $n = 2$. Weak feedback: $K = 5$, $V_{max} = 1$. Strong feedback: $K = 0.05$, $V_{max} = 5000$

Time to Reach Steady State

The time taken to reach steady state can be done by carrying out a simulation and observing the progress curves. If one simply took the network with weak feedback and compared it with the network with strong feedback one immediate problem arises. Negative feedback reduces the steady state concentration of *P*. This means that a comparison between the two systems will be unfair because of the difference in steady state concentrations in *P*. In order to make a fair comparison the maximal rate of gene expression in the strongly regulated system must be raised until the network has the same steady state as the unregulated network [1]. This has been done in the simulations shown in Fig. 3. Note that the rise time for the circuit with negative feedback is much shorter. This can also be easily seen in the rate of change of *P* at time zero for both circuits. With weak feedback, the rate of change of *P* is 1 unit per unit time. With strong feedback, this rises to 5000 units per unit time. We can therefore conclude from this simple study that for comparable systems, negative feedback reduces the time it takes to reach steady state. This effect has been observed in synthetic networks built in *E. coli* cells. In these experiments the rise time was observed using GFP (green fluorescence protein) tagged proteins in autoregulated and non-regulated transcription units [64]. The researchers observed that negative autoregulation reduced the rise time to one fifth of the time taken for the non-regulated circuit.

Properties of the Steady State

In the previous section simple mathematical models showed that the time taken to reach steady state in the presence of negative feedback could be significantly reduced although at the expense of increasing the maximal expression rate for the gene. What about the steady state itself, what effect does negative feedback have on the steady state? This question can be answered by using control analysis. The following equations represent the summation and connectivity theorems for the simple regulated gene circuit in Fig. 2.

$$C^P_{V_{\max}} + C^P_{k_1} = 0$$
$$C^P_{V_{\max}} \varepsilon^{v_1}_P + C^P_{k_1} \varepsilon^{v_2}_P = 0 \,.$$

These two equations can be used to solve for the control coefficients in terms of the elasticities:

$$C^P_{V_{\max}} = \frac{1}{\varepsilon^{v_1}_P - \varepsilon^{v_2}_P}$$
$$C^P_{k_1} = -\frac{1}{\varepsilon^{v_1}_P - \varepsilon^{v_2}_P} \,.$$

These equations can be used to study how the steady state concentration of *P* is influenced by the strength of the feedback. To answer this one need only look at $C^P_{V_{\max}}$ and $C^P_{k_1}$. The feedback strength is given by the elasticity coefficient, $\varepsilon^{v_1}_P$, since this measures the influence that the protein product, *P*, has on gene expression. Note that the negative feedback elasticity, $\varepsilon^{v_1}_P$, is negative, that is increasing *P* decreases the rate of gene expression. If the absolute value of $\varepsilon^{v_1}_P$ is small then the strength of the negative feedback is weak, whereas if the absolute value is large then the strength of the negative feedback is strong. Consider the first case when the feedback is weak, i. e. $\varepsilon^{v_1}_P \approx 0$. The degradation rate can be assumed to be linear in *P*, so that its elasticity, $\varepsilon^{v_2}_P$ will equal one. Taking both into consideration means that the control coefficient on the gene will tend to unity and the control coefficient on protein degradation will tend to minus unity. This result means that

variation in either gene expression or protein degradation will have significant effects on the steady state concentration of protein. Consider now the alternative where there is strong feedback, in this case $\varepsilon_P^{v_1}$ will have a large absolute value and $\varepsilon_P^{v_2} = 1$ as before. If the feedback elasticity is sufficiently large, which it can be if the Hill coefficient is large, both control coefficients will tend to small values. This means that both variation in gene expression and protein degradation will have little effect on the steady state concentration of protein.

We conclude that negative feedback stabilizes the concentration of the protein by reducing the effect of parameter variation. Intuitively, this is a well known idea and engineers frequently use negative feedback to stabilize systems. However the approach that uses control analysis enables us to quantify the effect and understand the properties of negative feedback in more depth.

There are many examples in metabolic networks where feedback appears to be used to stabilize a concentration, for example the negative feedbacks that go to the enzyme phosphofructokinase can stabilize ATP concentrations in the face of varying ATP demand [48]. In addition, synthetic networks have also been constructed [9] where a comparison is made between a gene circuit with and without negative feedback. They show that with feedback the effect of noise on the system, as indicated by the expression of green fluorescence protein (GFP), is greatly reduced, thus supporting the theoretical predication that negative feedback suppresses the influence of variation in the network.

Long Range Negative Feedback

Figure 4 illustrates a long range negative feedback where instead of the product of a gene inhibiting its own production, a product further downstream of a cascade of genes feeds back a signal to the first gene. The steady state properties of this network can be quite different from a single stage network. What is critical in a multi stage system is the delay in signal transmission caused by the time it takes each genetic unit to respond to changes upstream. Consider a pulse of P_1 added to the pathway in Fig. 4. The pulse will increase the rate of degradation of P_1 but it will also stimulate the second stage to express more P_2, this in turn will cause P_3 to increase its production rate. The propagation of the pulse of P_1 down the chain will however take a finite amount of time during which the concentration of P_1 is likely to have recovered to its original steady state level due to degradation. At the same time that P_1 is recovering P_3 will be increasing and via the negative feedback will suppress the production of P_1. Therefore rather than being restored P_1 will in fact be forced to go under its nominal steady state level. The response to the signal is therefore out of phase, that is whatever changes occur upstream, the time lags in the network cause the negative feedback to restore the system at the wrong time. Given a sufficient delay and enough strength in the feedback, the inability to correctly restore perturbations can develop into an unstable situation leading to sustained oscillations in the network. Such oscillatory systems are well known in engineering and are often called phase-shift oscillators [43].

A critical factor that determines the onset of oscillations, apart from the number of stages in the cascade is the strength of the feedback. Savageau [70,71,72] showed, in a detailed analysis, that for a metabolic pathway with negative feedback, if the substrate elasticities are equal (e. g. all first-order kinetics) then the ratio of the feedback elasticity (ε_{inh}) to the output elasticity ($\varepsilon_{\text{sub}}, \varepsilon_3^4$) determined the onset of oscillations (Table 1). Similar criterion would hold for a genetic network although the detailed analysis has not yet been done. Table 1 shows that as the pathway becomes longer, less feedback inhibition is required to destabilize the pathway. This highlights the other factor that contributes to instability, the delay in routing the signal around the network.

All feedback oscillators require some device to provide amplification of a signal combined with a suitable time delay so that the signal response can go out of phase. In metabolic pathways amplification is often provided by a cooperative enzyme while the delay is provided by the intermediate steps in the pathway. In signaling pathways, amplification can be generated by covalent modification cycles [33,45] while in genetic networks, amplification can be provided by cooperative binding of the transcription factor to the promoter site. The criterion for instability is the ratio of the inhibition elasticity to the substrate elas-

Biological Models of Molecular Network Dynamics, Figure 4
Long range negative feedback comprising of three gene stages with the product of the last stage inhibiting the first

Biological Models of Molecular Network Dynamics, Fig. 5
Ring oscillator, also called the repressilator

Biological Models of Molecular Network Dynamics, Table 1
Relationship between the pathway length and the degree of feedback inhibition on the threshold for stability. ε_{inh} is the elasticity of the feedback inhibition and ε_{sub} is the elasticity of the distal step with respect to the signal. See [71] for details of calculations

Length of Pathway	Instability Threshold $-\varepsilon_{inh}/\varepsilon_{sub}$
1	stable
2	stable
3	8.0
4	4.0
5	2.9
6	2.4
7	2.1
⋮	⋮
∞	1.0

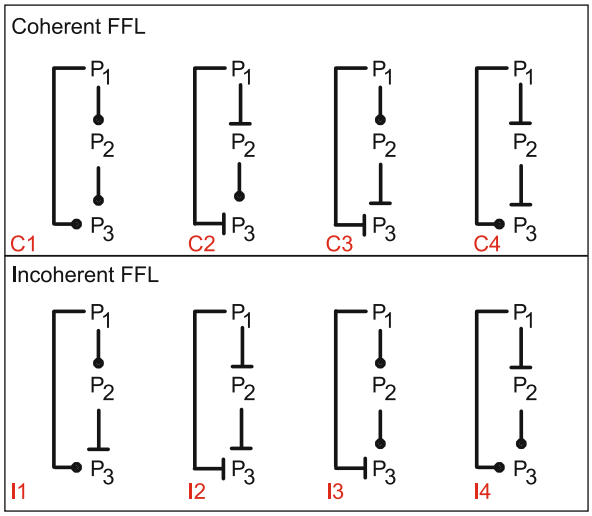

Biological Models of Molecular Network Dynamics, Fig. 6
Full complement of feedforward motifs, classified into coherent and incoherent types

ticity. If the output reaction of the pathway is governed by a saturable enzyme then it is possible to have ε_{sub} less than unity. This means that it is possible to trade cooperativity at the inhibition site with saturation at the output reaction. The modified phase-shift Goodwin model of Bliss [12] illustrates a model with no cooperativity at the inhibition site but with a degree of saturation at the output reaction by using a simple Michaelis–Menten rate law.

The ring oscillator, Fig. 5, also called the repressilator, is a type of phase-shift oscillator. This device is composed of an odd number of gene inverters connected in a closed loop. A ring oscillator has been built experimentally in *E. coli* [23]. Ring oscillators with an even number of inverters can also be used to form memory units or toggle switches. The even number of units means that the signal latches to either on or off, the final state depending on the initial conditions. Toggle circuits have also been implemented experimentally in *E. coli* [30].

FeedForward Systems

Many efforts in modeling cellular networks have been to understand the effects of negative feedback. However in recent years there has been a growing awareness of feedforward loops (FFL). Recent work in identifying motifs in genetic networks of *E. coli* and yeast has revealed the occurrence of significant numbers of feedforward loops [52] in these organisms, Fig. 7. These motifs have been classified into two broad groups, coherent and incoherent networks. Within each group four combinations of regulation can be discerned. Figure 6 illustrates all the possible combinations for the FFL motif. Figure 7 shows the relative abundance of the different kinds of feedforward motifs found in *E. coli* and yeast. This data indicates that two types predominate in both organisms, Coherent Type 1 (C1) and Incoherent Type 1 (I1). What is particularly interesting is that these two types have distinct behavioral properties.

Properties of the Coherent Type I Motif

The best way to understand the properties of the Coherent Type I feedforward network is to build a computer model and run a simulation. Figure 8 illustrates a possible genetic feedforward Coherent Type I network. Figure 9 illustrates one of the key properties of the Coherent Type I feedforward network. The output signal, P_3, is controlled by P_2 and P_1. If a pulse signal is applied to P_1, the signal travels two routes to get to P_3. If the pulse is too short, P_2 does

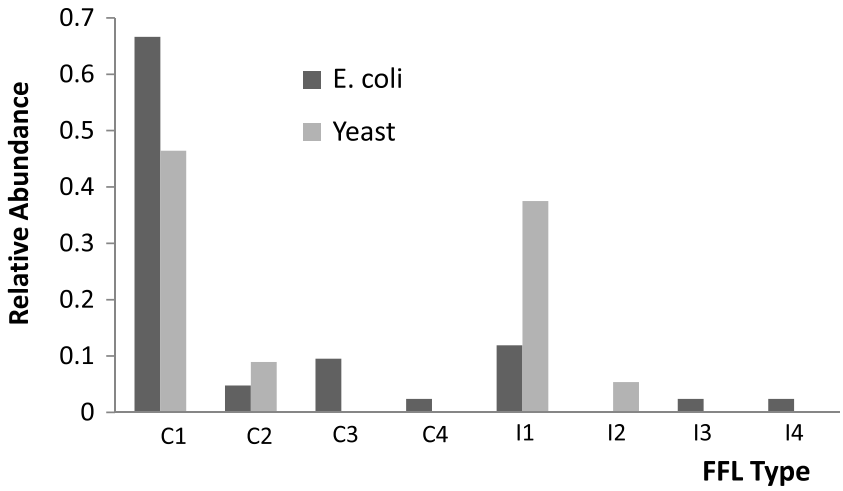

Biological Models of Molecular Network Dynamics, Fig. 7
Relative abundance of different FFL types in yeast and *E. coli*. Labels on the X-axis refer to the particular FFL motifs seen in Figure 6. Data taken from Mangan et al. [52]

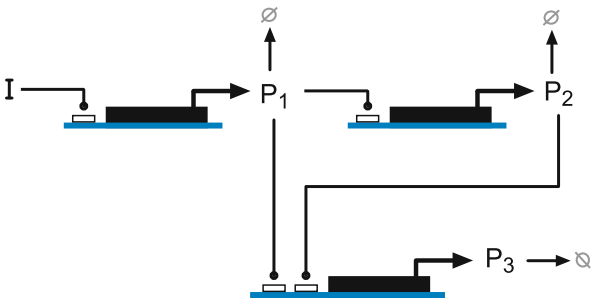

Biological Models of Molecular Network Dynamics, Fig. 8
Example of a feedforward Coherent Type I genetic network

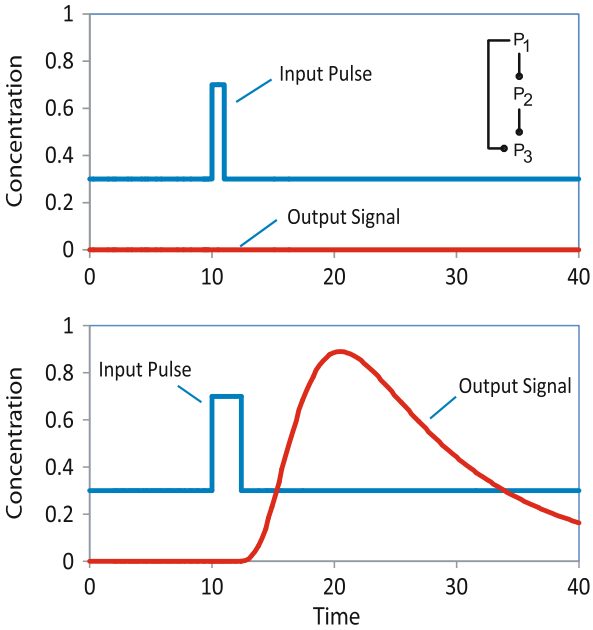

Biological Models of Molecular Network Dynamics, Fig. 9
Simulation of a Coherent Type I feedforward network. The top panel shows the effect of a narrow pulse in the concentration of P_1. The output signal, P_3, shows no response. In the lower panel, the input signal, P_1, is wider, this gives time for the feedforward loop to activate as shown by the large change in output signal, P_3. The Coherent Type I FFL will only *fire* if the input pulse is of sufficient width. This means that the network can filter out transient input pulses and only respond to input pulses of sufficient duration

not have sufficient time to increase in order to reach the threshold level at P_3, as a result, short pulses do not turn P_3 on. This effect is shown in the upper panel of Fig. 9. In sharp contrast, if the P_1 pulse width is wide enough, P_2 has sufficient time to increase so that is reaches the threshold level and activates P_3. The network therefore acts as a transient signal filter. Only signals of sufficient duration will result in activation.

Properties of the Incoherent Type I Motif

The Incoherent Type I feedforward network (Fig. 10) has a completely different response compared to the Coherent Type I network.

The Incoherent Type I network has a number of interesting behaviors, including pulse generator, concentration band detector and frequency band pass filter. The pulse generator simulation is shown in Fig. 11 where an input step function is applied to the input, P_1 and the effect on

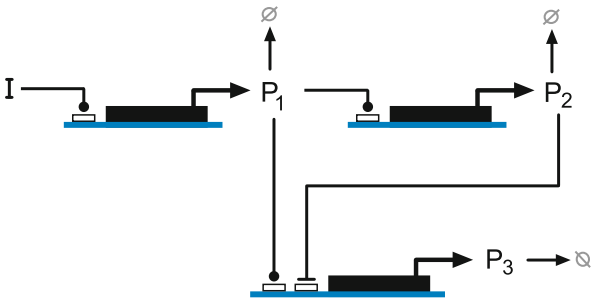

Biological Models of Molecular Network Dynamics, Fig. 10
Example of a feedforward Incoherent Type I genetic network. Note that the regulation on P_3 is antagonistic compared to the Coherent Type I network

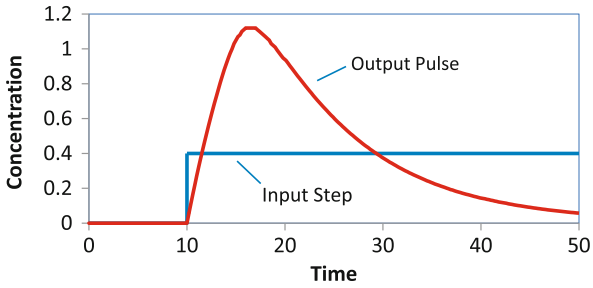

Biological Models of Molecular Network Dynamics, Fig. 11
Simulation of the pulse generation characteristics of an Incoherent Type I feedforward network. The X-axis shows the change in the input signal over time. At 10 time units, a step signal is applied, this causes the output to rise then fall in a pulse like manner. The width of the pulse can be adjusted by changing the degree of cooperativity on P_3

P_3 observed. In this case, P_3 rises rapidly then falls off in a pulse like manner even though the input signal remains on. This is due to the delay in the inhibition route, initially the inhibition is weak and the output raises, but eventually P_2 increases and begins to repress the production of P_3. The second type of behavior that the network can display is to act as a concentration band detector. The simulation shown in Fig. 12 shows the network turning on in a specific range of input signal. The position of the peak in the response can be adjusted by changing the strength of inhibition and the input threshold from P_1 to P_3.

Finally, an Incoherent Type I network can act as a band pass filter. That is the network will respond more strongly if the input signal varies at a specific range of frequencies. That is at high and low frequencies the network will not respond but at mid range frequencies it will.

A series of synthetic Incoherent Type I networks have been built in *E. coli* that illustrate the band pass behavior [24]. Further details on the properties of feedforward networks can be can also be found in Uri Alon's book [1].

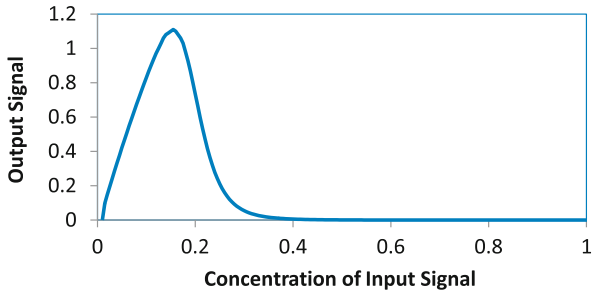

Biological Models of Molecular Network Dynamics, Fig. 12
Simulation of the concentration band detector characteristics of an Incoherent Type I feedforward network. The X-axis shows the change in the input signal while the Y-axis shows the steady state concentration in the output signal. The network only responds in a particular range of input concentration

Positive Feedback

The alternative to negative feedback is positive feedback; that is where the product of a process activates its own production. Ten percent of known transcription factors in *E. coli* are thought to act as positive feedback regulators [1]. Positive feedback shows many of the opposite properties to those of negative feedback. For example, systems with positive feedback have slow responses compared to systems without feedback and unlike negative feedback which tends to stabilize a system, positive feedback tends to destabilize. One may wonder why evolution has favored a destabilizing force since homeostasis and robustness are presumably desirable attributes to have inside a cell. What positive feedback brings is in fact robustness but via a behavior called bistability. Figure 13 shows a simple gene circuit where the gene product, P, positively influences its own production. We can easily model positive feedback using Hill like equations. A mathematical model for the system in Fig. 13 is given below:

$$
\begin{aligned}
v_1 &= k_1 + \frac{V_{\max} P^n}{15 + P^n} \\
v_2 &= k_2 P \\
\frac{\mathrm{d}P}{\mathrm{d}t} &= v_1 - v_2 \,.
\end{aligned}
\tag{7}
$$

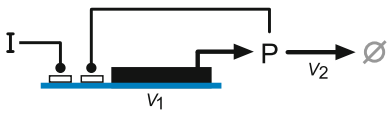

Biological Models of Molecular Network Dynamics, Fig. 13
Simple positive feedback

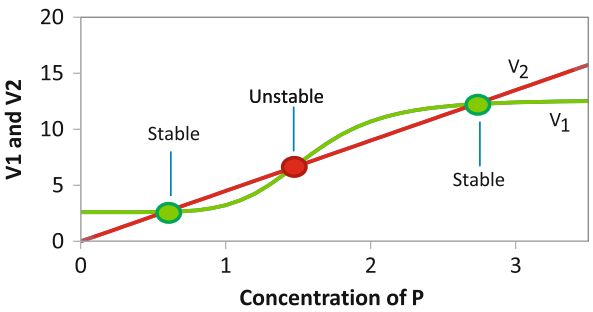

Biological Models of Molecular Network Dynamics, Fig. 14
Simple positive feedback. The plot shows v_1 and v_2 plotted against the concentration of *P*. Where the curves intersect, the system is at steady state. By examining perturbations around each steady state it is possible to determine that two of the states are stable and one is unstable

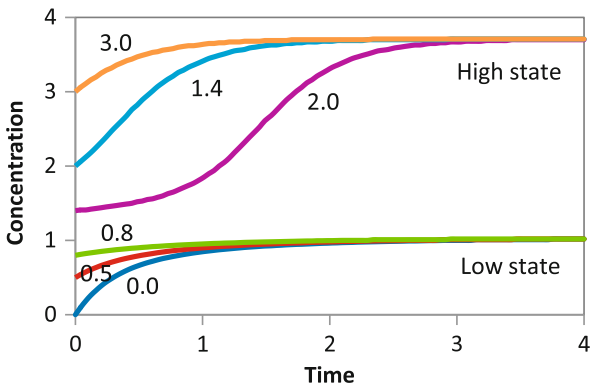

Biological Models of Molecular Network Dynamics, Fig. 15
Simulation of model in Fig. 13. Each curve displays the change in concentration for the gene product, *P* at different starting concentrations for *P*. Low concentrations of *P* settle to the low steady state, while higher starting concentrations evolve to the high state. Model given by Eqs. (7). Parameters: $k_1 = 2.6$, $k_2 = 3.2$, $V_{max} = 10$ and $n = 4$

The equation that describes gene expression, v_1, has two terms, k_1 which represents a basal level of gene expression and a second Hill like term that represents the positive feedback. The second rate, v_2, is the protein degradation rate. The net rate of change of *P* is the difference between the two. To illustrate the key behavior that the simple genetic circuit has, the rates v_1 and v_2 are plotted against the concentration of *P*, Fig. 14 shows this plot. The first thing to note is that where the two lines intersect the system is at steady state since at these points v_1 must equal v_2, which is the definition of steady state. What is interesting is that there are three possible steady states, where each steady state displays a different steady state concentration of *P*. If we focus on each steady state and examine the relative rates, v_1 and v_2, as we move away from the steady state by changing *P*, it becomes clear that the lower and upper steady states are stable, that is any changes in *P* at these points will restore *P*. However, if we were to change *P* at the middle steady state, for example increase *P*, it should be clear from the graph that v_1 will be higher than v_2, this means that more *P* is being produced which will move *P* further away from the steady state. This tells us the middle steady state is in fact unstable.

What we have here is something akin to simple light toggle switch. That is, a switch that has two stable states, which we might nominally call high (on) and low (off) and one middle state which is unstable. In a toggle switch one could imagine, very carefully, balancing the switch so that it is halfway between on and off, knowing that a slight nudge will flip the switch to either the on or off state. Positive feedback can lead to systems that display bistability. To illustrate bistability in a different way, Fig. 15 shows time course simulations of the genetic switch in Fig. 13. Six different simulations are shown, each simulation starts at a different initial concentration of *P*. At low concentrations, the time course behavior of *P* converges to a low state, at around 1.0 unit of *P*. High concentrations of *P*, on the other hand, converge to a different final state, a high state at around 3.5 units of *P*. What evolutionary selective advantage might bistability have? One clear benefit is that bistable states are very stable to perturbations in *P* such as thermal noise or transient signals that should be ignored. Once a bistable switch has settled to either the high or low state it is difficult to persuade the system to switch to the alternative state. In a noisy cellular environment, or where a definite decision needs to be made, for example, replication, sporulation or apoptosis, bistability is a useful behavior to have.

Relaxation Oscillator

In addition to providing switch like behavior, bistability can also be coupled with negative feedback to generate robust oscillators called relaxation oscillators [80]. Relaxation oscillators are commonly used in electronics where non-sinusoidal waveforms are required, for example, sawtooth or square waveforms [43]. Electronic relaxation oscillators repeatedly alternate between two states at a frequency that depends on a charging device such as a capacitor or an inductor. As a result, relaxation oscillators are characterized by two time-scales which is why they can be used to generate non-sinusoidal waveforms.

In biological networks, a relaxation oscillator operates by charging a species concentration which, upon reaching a threshold, changes the state of a bistable switch. When the switch changes state it causes the species to discharge. Once the species has discharged the bistable switch returns

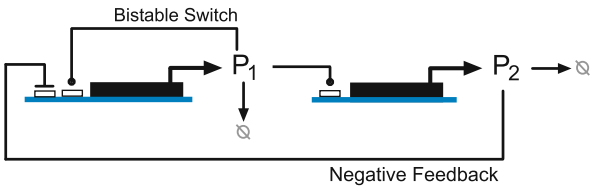

Biological Models of Molecular Network Dynamics, Fig. 16
Simple genetic circuit that behaves as a relaxation oscillator

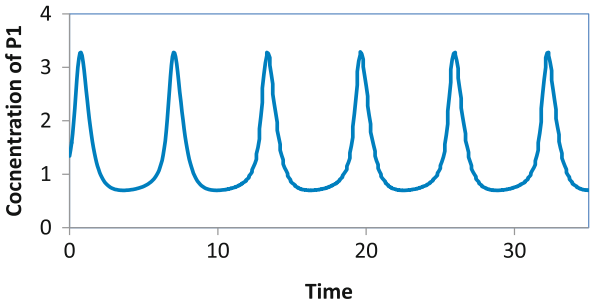

Biological Models of Molecular Network Dynamics, Fig. 17
Simulation of the relaxation model depicted in Fig. 16. Relaxation oscillators often have a characteristic *spiky* pattern that originates from the different times scales in the bistable switch and the negative feedback component

to the original state and the sequence begins again. Figure 16 shows a simple genetic relaxation oscillator with the bistable switch on the left (formed from P_1) and the negative feedback originating from the gene product P_2. One of the characteristics of a relaxation oscillator is the *spiky* appearance of the oscillations. This is due to the rapid switching of the bistable circuit which can be faster compared to the operation of the negative feedback. A simulation of the relaxation oscillator shown in Fig. 16 is given in Fig. 17.

The previous sections have covered a few of the common design patterns we find in biological networks but many others exist and most likely many more remain to be discovered. Although the focus has been on genetic networks, the exact same design patterns can be found in metabolic as well as signaling pathways [16,17,46]. This means that once an understanding of how these patterns operate it is relatively easy to translate this understanding to many different biological situations. Crucial to understanding how these networks operate and uncovering the basic design principles of how molecular network operate has required a concerted effort between experimentalists, modelers and theorists. Modeling in particular makes it easy, for both experimentalists and computationalists, to explore the dynamic behavior of complex network configurations. The availability of easy to use software such as the Systems Biology Workbench [10] makes modeling and model construction accessible to a wide range of researchers in biology.

Future Prospects

Like all good science, the development of model-based systems biology depends heavily on the ability to make measurements on real systems. In the case of model development the ideal measurements would be high resolution, the time course behavior of molecular states. In recent years there has been significant progress in this area, in particular, researchers now have the ability, through the use of fluorescent proteins for example, to measure the detailed dynamics that occur in the levels of protein in single cells. This kind of data allows model builders and experimentalists to combine their expertise in understanding molecular networks. A particularly striking example of this synergy is found in the work of Bar-Or et al. [51] and Lahav et al. [49] where the authors coupled high resolution measurements of Mdm2 and P53 protein levels in single cells with mathematical modeling. Further progress is needed however, to make more detailed measurements, in particular experimentalists still find it difficult to measure at high time resolution particular protein states (e. g. phosphorylated states) in live single cells. In addition, real-time measurements of metabolic states other than naturally fluorescent molecules such as NAD is still not possible. However recent progress in aptamer technology [8,86] may allow, for the first time, detailed measurements in real-time of both protein and metabolic states.

In addition to making measurements, validating, or more precisely, falsifying models, is still not a well developed area. One of the few researches who has taken this seriously is Tyson and Novak who have developed sophisticated computer tools to match models against experimental data to ensure the validity of models as more data is forthcoming [83].

Perhaps one of the most exiting and promising avenues for matching experimental measurements with model building is to through synthetic biology. The approach that synthetic biologists take is to build new networks in vivo, in organisms such as *E. coli*. The great advantage of synthetic biology is that networks can be studied under controlled conditions [24]. Many insights into network dynamics have been provided by using synthetic biology techniques, far more probably than any other approach [2,58,84].

Bibliography

1. Alon U (2006) An Introduction to Systems Biology: Design Principles of Biological Circuits. Chapman & Hall/Crc Mathematical

and Computational Biology Series. Chapman & Hall/CRC, Boca Raton
2. Alon U (2007) Network motifs: theory and experimental approaches. Nat Rev Genet 8(6):450–461
3. Andrianantoandro E, Basu S, Karig DK, Weiss R (2006) Synthetic biology: new engineering rules for an emerging discipline. Mol Syst Biol 2:2006–2006
4. Ariew R (1976) Ockham's Razor: A Historical and Philosophical Analysis of Ockham's Principle of Parsimony. Champaign-Urbana, University of Illinois
5. Aris R (1965) Prolegomena to the Rational Analysis of Systems of Chemical Reactions. Arch Rational Mech Anal 19:81–99
6. Arkin A, Ross J, McAdams HH (1998) Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells. Genetics 149:1633–48
7. Atkins P (2001) Physical Chemistry, 7th edn. W. H. Freeman
8. Bayer TS, Smolke CD (2005) Programmable ligand-controlled riboregulators of eukaryotic gene expression. Nat Biotechnol 23(3):337–343
9. Becskei A, Serrano L (2000) Engineering stability in gene networks by autoregulation. Nature 405:590–593
10. Bergmann FT, Vallabhajosyula RR, Sauro HM (2006) Computational Tools for Modeling Protein Networks. Current Proteomics 3(3):181–197
11. Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, et al (1997) The complete genome sequence of Escherichia coli K-12. Science 277(5331):1453–1474
12. Bliss RD, Painter PR, Marr AG (1982) Role of feedback inhibition in stabilizing the classical operon. J Theor Biol 97(2):177–193
13. Burns JA (1971) Studies on Complex Enzyme Systems. Dissertation, University of Edinburgh. http://www.sys-bio.org/BurnsThesis
14. Cannon RH (1967) Dynamics of Physical Systems. McGraw-Hill College, New York
15. Chen YD, Westerhoff HV (1986) How do Inhibitors and Modifiers of Individual Enzymes Affect Steady-state Fluxes and Concentrations in Metabolic Systems? Math Model 7:1173–1180
16. Chickarmane V, Kholodenko BN, Sauro HM (2007) Oscillatory dynamics arising from competitive inhibition and multisite phosphorylation. J Theor Biol 244(1):68–76
17. Chickarmane V, Ray A, Sauro HM, Nadim A (2007) A Model for p53 Dynamics Triggered by DNA Damage. SIAM J Appl Dynamical Syst 6(1):61–78
18. Cinquin O, Demongeot J (2002) Roles of positive and negative feedback in biological systems. C R Biol. 325:1085–1095
19. Clarke BL (1980) Stability of complex reaction networks. Adv. Chem. Phys, vol. 42. Wiley, New York
20. Cornish-Bowden A (1995) Fundamentals of Enzyme Kinetics. Portland Press, London
21. Cornish-Bowden A, Cárdenas ML (2002) Metabolic balance sheets. Nature 420(6912):129–130
22. Dibrov BF, Zhabotinsky AM, Kholodenko BN (1982) Dynamic stability of steady states and static stabilization in unbranched metabolic pathways. J Math Biol 15:51–63
23. Elowitz MB, Leibler S (2000) A synthetic oscillatory network of transcriptional regulators. Nature 403:335–338
24. Entus R, Aufderheide B, Sauro HM (2007) Design and implementation of three incoherent feed-forward motif based biological concentration sensors. Systems and Synthetic Biology. doi:10.1007/s11693-007-9008-6
25. Erdi P, Toth J (1989) Mathematical Models of Chemical Reactions. Theory and Applications of Deterministic and Stochastic Models. Manchester University Press, Manchester, Princeton University Press, Princeton
26. Fell D (1997) Understanding the Control of Metabolism. Portland Press, London
27. Fell DA, Sauro HM (1985) Metabolic Control Analysis: Additional relationships between elasticities and control coefficients. Eur J Biochem 148:555–561
28. Fell DA, Small JR (1986) Fat synthesis in adipose tissue: an examination of stoichiometric constraints. Biochem J 238:781–786
29. Ferrell JE (1996) Tripping the switch fantastic: how a protein kinase cascade can convert graded inputs into switch-like outputs. Trends in Biochemical Sciences 21:460–466
30. Gardner TS, Cantor CR, Collins JJ (2000) Construction of a genetic toggle switch in Escherichia coli. Nature 403:339–342
31. Gillespie DT (2007) Stochastic simulation of chemical kinetics. Annu Rev Phys Chem 58:35–55
32. Goldbeter A (1997) Biochemical Oscillations and Cellular Rhythms: The Molecular Bases of Periodic and Chaotic Behaviour. Cambridge University Press, Cambridge
33. Goldbeter A, Koshland DE (1984) Ultrasensitivity in biochemical systems controlled by covalent modification. Interplay between zero-order and multistep effects. J Biol Chem 259:14441–7
34. Heinrich R, Rapoport TA (1974) A Linear Steady-state Treatment of Enzymatic Chains; General Properties, Control and Effector Strength. Eur J Biochem 42:89–95
35. Heinrich R, Schuster S (1996) The Regulation of Cellular Systems. Chapman and Hall, New York
36. Heinrich R, Rapoport SM, Rapoport TA (1977) Metabolic regulation and mathematical models. Prog Biophys Molec Biol 32:1–82
37. Hofmeyr JHS (1986) Steady state modelling of metabolic pathways: a guide for the prespective simulator. Comp Appl Biosci 2:5–11
38. Hofmeyr JHS (1986) Studies in steady state modelling and control analysis of metabolic systems. Dissertation, University of Stellenbosch
39. Hofmeyr JHS (2001) Metabolic Control Analysis in a Nutshell. In: Proceedings of the Second International Conference on Systems Biology, Caltech
40. Hood L, Heath JR, Phelps ME, Lin B (2004) Systems biology and new technologies enable predictive and preventative medicine. Science 306(5696):640–643
41. Horn F, Jackson R (1972) General Mass Action Kinetics. Arch Rational Mech Anal 47:81–116
42. Ideker T, Galitski T, Hood L (2001) A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet 2:343–372
43. Jones MH (1977) A practical introduction to electronic circuits. Cambridge University Press, Cambridge
44. Kacser H, Burns JA (1973) The Control of Flux. In D. D. Davies (eds) Rate Control of Biological Processes, Symp Soc Exp Biol, vol 27, Cambridge University Press, Cambridge, pp 65–104
45. Kholodenko BN (2000) Negative feedback and ultrasensitivity can bring about oscillations in the mitogen-activated protein kinase cascades. Eur J Biochem 267:1583–1588
46. Kholodenko BN (2006) Cell-signalling dynamics in time and space. Nat Rev Mol Cell Biol 7(3):165–176

47. Klipp E, Herwig R, Kowald A, Wierling C, Lehrach H (2005) Systems Biology in Practice. Wiley-VCH, Weinheim
48. Koebmann BJ, Westerhoff HV, Snoep JL, Nilsson D, Jensen PR (2002) The Glycolytic Flux in Escherichia coli Is Controlled by the Demand for ATP. J Bacteriol 184(14):3909–3916
49. Lahav G, Rosenfeld N, Sigal A, Geva-Zatorsky N, Levine AJ, Elowitz MB, Alon U (2004) Dynamics of the p53-Mdm2 feedback loop in individual cells. Nature, Genetics 36(2):147–150
50. Lauffenburger DA (2000) Cell signaling pathways as control modules: complexity for simplicity? Proc Natl Acad Sci U S A 97:5031–3
51. Lev Bar-Or R, Maya R, Segel LA, Alon U, Levine AJ, Oren M (2000) Generation of oscillations by the p53-Mdm2 feedback loop: a theoretical and experimental study. Proc Natl Acad Sci USA 97(21):11250–11255
52. Mangan S, Itzkovitz S, Zaslaver A, Alon U (2006) The incoherent feed-forward loop accelerates the response-time of the gal system of Escherichia coli. J Mol Biol 356(5):1073–1081
53. Markevich NI, Hoek JB, Kholodenko BN (2004) Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades. J Cell Biol 164:353–9
54. Moniz-Barreto P, Fell DA (1993) Simulation of dioxygen free radical reactions. Biochem Soc Trans 21(3):256–256
55. Moore WJ (1972) Physical Chemistry. 5th edn. Longman, London, Prentice Hall, NJ
56. Oda K, Kitano H (2006) A comprehensive map of the toll-like receptor signaling network. Mol Syst Biol 2:2006–2006
57. Othmer HH (1976) The quantitative dynamics of a class of biochemical control circuits. J Math Biol 37:53–78
58. Ozbudak EM, Thattai M, Lim HN, Shraiman BI, Van Oudenaarden A (2004) Multistability in the lactose utilization network of Escherichia coli. Nature 427(6976):737–740
59. Papin JA, Stelling J, Price ND, Klamt S, Schuster S, Palsson BO (2004) Comparison of network-based pathway analysis methods. Trends Biotechnol 22(8):400–405
60. Paulsson J, Berg OG, Ehrenberg M (2000) Stochastic focusing: fluctuation-enhanced sensitivity of intracellular regulation. Proc Natl Acad Sci USA 97(13):7148–7153
61. Reder C (1988) Metabolic Control Theory: A Structural Approach. J Theor Biol 135:175–201
62. Reich JG, Selkov EE (1981) Energy metabolism of the cell. Academic Press, London
63. Ro DK, Paradise EM, Ouellet M, Fisher KJ, Newman KL, Ndungu JM, Ho KA, et al (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature 440(7086):940–943
64. Rosenfeld N, Elowitz MB, Alon U (2002) Negative autoregulation speeds the response times of transcription networks. J Mol Biol 323(5):785–793
65. Samoilov M, Plyasunov S, Arkin AP (2005) Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations. Proc Natl Acad Sci USA 102(7):2310–2315
66. Sauro HM, Ingalls B (2004) Conservation analysis in biochemical networks: computational issues for software writers. Biophys Chem 109:1–15
67. Sauro HM, Kholodenko BN (2004) Quantitative analysis of signaling networks. Prog Biophys Mol Biol. 86:5–43
68. Sauro HM, Small JR, Fell DA (1987) Metabolic Control and its Analysis: Extensions to the theory and matrix method. Eur J Biochem 165:215–221
69. Savageau MA (1972) The behaviour of intact biochemical control systems. Curr Topics Cell Reg 6:63–130
70. Savageau MA (1974) Optimal design of feedback control by inhibition: Steady-state considerations. J Mol Evol 4:139–156
71. Savageau MA (1975) Optimal design of feedback control by inhibition: dynamic considerations. J Mol Evol 5(3):199–222
72. Savageau MA (1976) Biochemical systems analysis: a study of function and design in molecular biology. Addison-Wesley, Reading
73. Schuster S, Dandekar T, Fell DA (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. Trends Biotechnol 17(2):53–60
74. Schuster S, Fell DA, Dandekar T (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nature Biotechnology 18:326–332
75. Segel IH (1975) Enzyme Kinetics: Behavior and Analysis of Rapid Equilibrium and Steady-State Enzyme Systems. Wiley-Interscience, New York
76. Stephanopoulos GN, Aristidou AA, Nielsen J (1998) Metabolic Engineering: Principles and Methodologies. Academic Press, San Diego
77. Strogatz S (2001) Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering. Perseus Books Group, Reading
78. Tyson J, Othmer HG (1978) The dynamics of feedback control circuits in biochemical pathways. In: Rosen R, Snell FM (eds) Progress in Theoretical Biology, vol 5. Academic press, New York, pp 1–62
79. Tyson JJ, Chen K, Novak B (2001) Network Dynamics And Cell Physiology. Nat Rev Mol Cell Biol 2:908–916
80. Tyson JJ, Chen KC, Novak B (2003) Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. Curr Opin Cell Biol 15:221–231
81. Umbarger HE (1956) Evidence for a Negative-Feedback Mechanism in the Biosynthesis of Leucine. Science 123:848
82. Vallabhajosyula RR, Chickarmane V, Sauro HM (2006) Conservation analysis of large biochemical networks. Bioinformatics 22(3):346–353
83. Vass M, Allen N, Shaffer CA, Ramakrishnan N, Watson LT, Tyson JJ (2004) the JigCell model builder and run manager. Bioinformatics 20(18):3680–3681
84. Voigt CA (2006) Genetic parts to program bacteria. Curr Opin Biotechnol 17(5):548–557
85. Wilkinson DJ (2006) Stochastic Modelling for Systems Biology. Chapman and Hall, New York
86. Win MN, Smolke CD (2007) A modular and extensible RNA-based gene-regulatory platform for engineering cellular function. Proc Natl Acad Sci U S A 104(36):14283–14288
87. Wolf DM, Arkin AP (2003) Motifs, modules and games in bacteria. Current Opinion in Microbiology 6:125–34
88. Wolkenhauer O, Ullah M, Wellstead P, Cho KH (2005) The dynamic systems approach to control and regulation of intracellular networks. FEBS Lett 579(8):1846–1853
89. Yates RA, Pardee AB (1956) Control of Pyrimidine Biosynthesis in Escherichia coli by a Feed-Back Mechanism. J Biol Chem 221:757–770
90. Zaslaver A, Mayo AE, Rosenberg R, Bashkin P, Sberro H, Tsalyuk M, Surette MG, et al (2004) Just-in-time transcription program in metabolic pathways. Nat Genet 36(5):486–491

Biomolecular Network Structure and Function

Lan V. Zhang[1], Frederick P. Roth[1,2]
[1] Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, USA
[2] Center for Cancer Systems Biology (CCSB), Dana-Farber Cancer Institute, Boston, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Single-Color Networks
Multi-Color Networks
Rich Network Models
Future Directions
Bibliography

Glossary

Biomolecule Any organic molecule that is produced by or essential to a living organism, sometimes specifically referring to macromolecules such as a protein or nucleic acid.

Biomolecular network a graph representation of relationships amongst a group of biomolecules. Nodes or vertices represent biomolecules. An edge or link between two vertices indicates a relationship between the corresponding biomolecules, for example, physical interaction, genetic interaction, or regulatory relationship.

Protein-protein interaction the physical association of two protein molecules with each other. A pair of proteins can interact directly with physical contact, or indirectly through other biomolecule(s), often other protein(s).

Yeast two-hybrid an experimental method to examine protein-protein interaction, in which one protein is fused to a transcriptional activation domain (the GAL4 activation domain) and the other to a DNA-binding domain (the GAL4 DNA-binding domain), and both fusion proteins are introduced into yeast. Expression of a GAL4-regulated reporter gene with the appropriate DNA-binding sites upstream of its promoter indicates that the two proteins physically interact.

Genetic interaction (epistasis) functional interaction between genes, in which the action of one gene is modified by the other gene, sometimes called the modifier gene. The gene whose phenotype is expressed is said to be ***epistatic***, while the one whose phenotype is altered or suppressed is said to be ***hypostatic***. ***Epistasis*** can either refer to this phenomenon, or more broadly to any case in which two mutations together cause a phenotype that is surprising given the phenotypes of each single mutation alone.

"Single-color" network a network with edges defined by only one type of interaction or relationship.

"Multi-color" network a network with edges defined by more than one type of interaction or relationship, with each type corresponding to a different '*color*'.

Scale-free network See **Power-law network**

Power-law network a network defined by a degree distribution which follows $P(k) \sim k^{-\gamma}$, where the probability $P(k)$ that a vertex in the network connects with k other vertices is roughly proportional to $k^{-\gamma}$. Sometimes networks that exhibit this behavior only at high degree are also called power-law. The coefficient γ seems to vary approximately between 2 and 3 for most real networks. In a power-law network, majority of the vertices have low degree (connectivity), while a small fraction of the vertices have very high degree. Highly-connected vertices are referred to as ***hubs***.

Small-world network a special type of network with: 1) short characteristic path length, such that most vertex pairs are connected to one another via only a small number of edges) and 2) high clustering coefficient, such that neighbors of a given vertex tend to be connected to one another.

Network motif a specific pattern of connected vertices and edges that occurs frequently within a given network.

Definition of the Subject

Biological research over the past century or so has been dominated by reductionism – identifying and characterizing individual biomolecules – and has enjoyed enormous success. Throughout this history, however, it has become increasingly clear that an individual biomolecule can rarely account for a discrete biological function on its own. A biological process is almost always the result of a complex interplay of relationships amongst biomolecules [5,19,50,51,67,68,90,128], and the treatment of these relationships as a graph is a natural and useful abstraction.

Broadly speaking, **a biomolecular network** is a graph representation of relationships (of which there are many types) amongst a group of biomolecules. Vertices or nodes represent biomolecules, including macromolecules such

as genes, proteins, and RNAs, or small biomolecules like amino acids, sugars, and nucleic acids. In the next few sections, we focus mostly on the macromolecules. An edge or link between two vertices indicates a relationship between the corresponding biomolecules, which could include physical interaction, genetic interaction, or a regulatory relationship (e. g., the protein product of gene A regulates the expression of gene B). This abstraction, although simplifying, converts a complex web of biological relationships into a mathematical graph, from which we can study its structural features as well as their implications on biological functions.

Aside from biomolecular networks, other biological networks also bear significance in systems biology. Examples include networks of protein domains and neuronal networks [29,86]. Although many of the features described in this chapter also apply, these networks are largely out of the scope of this chapter.

Introduction

Study of small-scale biomolecular networks has a long history, stemming from "chemical graphs" introduced in the 1700s [98], with knowledge about networks of metabolic reactions reaching a large scale by the 1970s.

A prominent early example may be found in the context of Monod and Jacob's Nobel-winning research in the 1950s, which first illuminated a transcriptional regulation system, namely control of gene expression in the *Lac* operon [58,84,85]. In their landmark work, Monod and Jacob described the operon model, in which the expression of structural genes is suppressed by the binding of the repressor to the operator gene, which in turn can be regulated by certain metabolites [58]. Regulatory relationships amongst the repressor, the metabolite, the operator gene, and the structural genes in the operon model present one of the first examples of biomolecular networks (Fig. 1). In addition, Monod and Jacob also proposed several theoretical models, in the form of biochemical or enzymatic networks, as potential mechanisms of biochemical differentiation of cells with identical genomes [85].

Pioneering work on the *lac* operon also includes studying functional implications of bionetwork structure. Building upon Jacob and Monod's work, the elegant experiments of Novick and Weiner [89] showed interesting features of beta-galactosidase induction – in a single cell, the production of lactose degrading enzymes is either switched on or off, and the cell's on or off state is inherited by subsequent generations. A mixture of the two types of cells has the intermediate overall level of enzyme production observed in a normal cell culture. This all-or-none phenomenon can be explained by positive feedback in the regulatory network of beta-galactosidase synthesis, where the intracellular concentration of inducer is boosted by an intermediate mechanism, which is also induced by the inducer (Fig. 1). As we can see from Fig. 1, the very first studies on gene regulation also show how the structure of biomolecular networks, even the small-scale pathways known at that time, can illuminate regulation of protein production.

Molecular biology has entered an era in which the nature, function and relationships of macromolecules can be

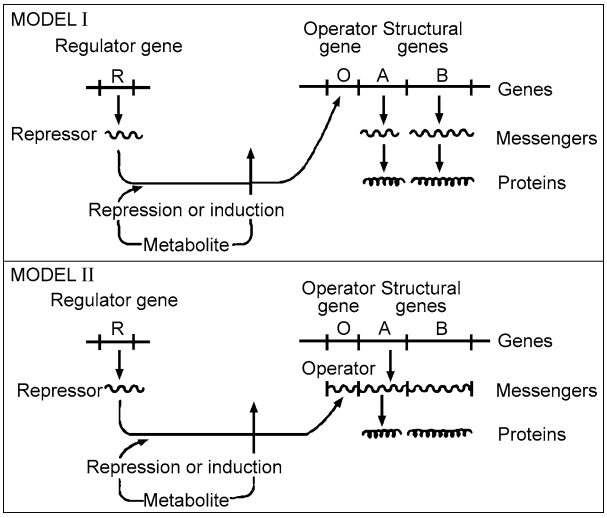

Biomolecular Network Structure and Function, Figure 1
***Lac* operon model**

studied at a much larger scale. High-throughput technology platforms such as genome sequencing, microarrays, and yeast two-hybrid screening have enabled biologists to examine how nearly all genes or proteins in a genome or proteome interact or relate to one another [40,76,100,107,108]. Collections of these genomic or proteomic data have led to networks of various relationship types – including protein-protein interaction, transcriptional regulation, sequence homology – involving thousands of proteins or genes [22,35,52,57,69,73,99,125]. This has offered an unprecedented 'bird's-eye view' of how parts of the network, each studied intensively on a local scale, connect to one another. This perspective may be useful in uncovering unknown pieces of the puzzle that together form the intricate machinery of life responsible for the behavior of the cell or the organism. The extremely large scale of these networks also presents a tremendous challenge as it is practically impossible to study each vertex or each interaction individually. A simplified abstract representation of these networks and mathematical methods permits automated study of the structural features, to identify local and global trends that relate biological macromolecules to their functional implications. Early work on large-scale bionetworks focused on **single-color networks** – networks describing a single type of relationship, while **multi-color networks** – networks combining multiple relationship types have gained increasing attention in recent years.

At an abstract level, a network of protein-protein interactions and a network of computers connected through the internet have much in common. Both have simplified representations as mathematical graphs: vertices represent proteins in the former case and computers in the latter case, and edges between vertices denote physical interactions between proteins or communication wires between computers. Concurrent with the increasing availability of large-scale biological networks, theoretical research in complex networks is advancing rapidly, with significant progress in uncovering the structural features and organizing principles of complex technological and social networks [3,18,27,88]. In a truly interdisciplinary effort, these principles were soon applied to biological networks, leading to the observation that biological networks share many global architectural features with networks from other complex systems [33,43,110,126,129], including power-law connectivity, small-world topology, and hierarchy, each discussed further below. Such resemblance suggested that many biological networks might be governed by similar 'design' principles, allowing methods proven in other large, non-biological complex systems to be used to characterize the intricate relationships between biomolecules.

While some reported global features of biomolecular networks, others sought local topological patterns recurrent in the network [5,73,83,103]. **Network motifs** are simple patterns of interconnection in networks that occur more frequently than expected in randomized networks [83,103], and have been proposed to represent basic building blocks of complex networks [73,83,103]. Different types of networks exhibit different motif profiles, providing a means for network classification [82]. Dobrin et al. later showed for the *E. coli* transcriptional network that subgraphs matching two types of transcriptional regulatory circuit motif – feed-forward and bi-fan motifs – overlap with one another and form large clusters which further coalesce into 'superclusters' [26]. Although some circuits corresponding to motifs may be capable of functioning independently [79], observation of Debrin et al., among many others, leads to the idea that instances of a network motif are often interdependent on one another as opposed to being building blocks with separable functions [138].

Cellular function is carried out through multiple types of interaction, resulting in a complex interplay of biological relationships amongst the myriad of cellular components. To have a truly global picture of the cell as a complex system, one needs to relate multiple types of interactions to one another. **Multi-color networks** are integrated networks of many relationship types, with frequent overlap between relationships of different types. For example, physically interacting proteins have expression patterns that are slightly more similar than would be expected by chance [36,60]; genes with correlated expression are more likely to be controlled by a common transcription factor [137]; and some genetic interaction types have a weak tendency to occur between homologous genes [122] (see Sect. "Single-Color Networks" for definitions of the various interaction types).

Topological study of single-color networks is paralleled in multi-color networks. Multi-color network motifs characterize relationships between different biological interaction types within local network neighborhoods [120,134,138]. Furthermore, multi-color networks also exhibit network themes – recurring higher-order interconnection patterns that encompass multiple occurrences of network motifs and reflect a common organizational principle [138]. Network themes can be tied to specific biological phenomena that represent more fundamental 'design principles', for example, the tendency for proteins within the same physical complex to be regulated by a common transcription factor [105,138]. This suggests that instead of representing network 'building blocks', network motifs should in some cases be viewed as signatures of more fundamental higher-order structures [138]. Net-

work themes also provide a more natural simplification of the otherwise complex set of relationships in an integrated network [138].

Studying the architectural features of biological networks conveys useful biological intuition and also allows us to make biological predictions. Even simple methods exploiting the predictive power of local topology have proven useful in predicting physical and genetic interactions [12,41,43,132]. Richer network models are able to incorporate more information and hence promise better predictions.

A multi-color network can be viewed as a graph in which each edge carries labels denoting the relationships between the corresponding pair of vertices. Similarly, each vertex may carry multiple labels (e. g., denoting that a particular protein is expressed in the pancreas; or that a gene is conserved in chimpanzee). A **rich network** is a network in which a vector describing potentially many vertex properties is associated with each vertex, and a vector describing potentially many edge properties is associated with each pair of vertices. By incorporating a more comprehensive set of information, these rich network models provide additional power in predicting features of genes or proteins, as well the relationships between them. In the following sections, we will survey methodology, findings, and applications for biological network models of increasing complexity, starting with single-color biomolecular networks.

As described previously, study of large-scale biomolecular networks has been enabled by the development of a variety of genomic and proteomic technologies, yielding a variety of relationship types. *Saccharomyces cerevisiae*, the budding yeast, a relatively simple single-celled eukaryote, has been an important model system, and many network studies originate from this yeast. Many core functions in yeast are conserved in humans, and many network trends observed in yeast have proven true for other organisms [99,116], leading us to believe that relationships in different organisms share similar organizational principles.

Before delving into the details of biomolecular network modeling, we add a few notes of caution. First, even the most complex network models remain a vast oversimplification of the much richer and more complex reality of cellular networks. A single 'vertex' can actually correspond to hundreds or even millions of physical protein molecules carrying a variety of post-translation modifications, and each diffusing independently throughout the cell. As important nuances relating to spatial and temporal dynamics are often lost, the network abstraction is only relevant to the extent that it proves useful. As the famous statistician George Box put it, "all models are wrong, some are useful" [71]. Second, the method of network abstraction and the data source/experimental may introduce systematic errors. For example, the so-called 'matrix' method of inferring protein interactions (which assumes that proteins interact if each binds directly or indirectly to a common 'bait' protein) will show 'over-clustering' by definition. Alternatively the 'spoke' method, in which interactions are only inferred between a bait protein and each prey protein that it 'pulls down', will result in 'under-clustering' if only a small subset of proteins are used as baits [9,35,52]. Therefore, one should take extra caution in interpreting observed network features to separate artifacts from real biological phenomena by using appropriate controls. Last, in searching network structure for trends, one often needs to compare the actual biomolecular network with appropriate random networks to ensure that the features identified are not a trivial consequence of a simpler phenomenon. For example, in searching for network motifs, randomized networks with the same **degree distribution** (see Sect. "Single-Color Networks") can be used as a control to avoid discovering motifs that are a trivial consequence of the underlying degree distribution [66,83,138].

Single-Color Networks

Definition and Categorization

The most basic biomolecular network is a single-color network, which contains only one type of biological interaction or relationship. The nature of the biological interaction or relationship defines the type of network, as it also specifies the nature of biomolecules involved. For interactions with clearly defined directionality, the corresponding network can be represented mathematically as a **directed** graph, with vertices denoting biomolecules, and edges connecting pairs of vertices whose corresponding biomolecules have directional interactions. Edge direction can represent material flow, exemplified by a **metabolite network** in which vertices are metabolites, and edges are enzyme-catalyzed reactions. As many metabolic reactions are practically irreversible under physiological conditions, we can use directed edges to represent the direction of a reaction, pointing from substrate to product. Alternatively, one can represent biochemical reactions using a directed edge to connect two enzymes such that the first enzyme's metabolic product is the second one's substrate. We refer to networks of the latter type as **metabolic enzyme networks**, to distinguish them from metabolite networks. Direction of interaction can also be derived from the inferred direction of information flow.

For example, in a **signaling pathway**, a directed edge points from one protein to another immediately downstream in signal transduction. Transcriptional regulation is another type of directed biomolecular interaction, and we can construct a **network of transcriptional regulation** from directed edges connecting transcription factors to the genes they regulate. In addition to transcriptional regulation, gene regulation in the cell also happens post-transcriptionally, for example through microRNAs (or small RNAs) which base pair with mRNAs and influence their stability or translation. Researchers have constructed **microRNA networks** by connecting microRNAs with their target mRNAs based on experimental identification at a limited scale [30,111,115] or based on computational predictions [94,124]. Post-translational regulation is get another important part of the cellular regulatory process. One example of a **post-translational regulatory network** is a **protein modification network** such as a **protein phosphorylation network**, in which each directed edge connects a kinase with its protein substrate or phosphorylation target [81,92].

Single-color networks that capture symmetric relationships can be represented by **undirected** graphs, with each vertex denoting a biomolecule, and an edge connecting vertices whose corresponding biomolecules have a symmetric relationship. An important example is a **protein network of physical interactions**, in which edges connect proteins that physically associate with each other with sufficiently small dissociation constants. Such association may be via direct physical contact, or indirect via one or more intermediates.

Genetic interaction is another important type of biomolecular relationship in studying cellular functions. It describes pairs of genes for which the phenotype of the double mutant departs from expected under some null model of independence, e. g., gene pairs for which the fitness of the double mutant deviates from the product of the fitness values for each of the two single mutants. **Synthetic sick genetic interactions** (**synthetic lethality** in the extreme case), correspond to gene pairs for which double mutant fitness is lower than expected. These synthetic sick or lethal (SSL) interactions can be represented as an undirected network. SSL interactions of ten correspond to genes with have redundant or compensatory function. They are also slightly more likely to encode proteins that interact physically [122]. If the double mutant phenotype for a pair of genes is less extreme than expected, the corresponding gene pair is said to have an antagonistic (or alleviating) genetic interaction. Although we can use undirected edges for alleviating interactions, in some cases additional information can be conveyed using **directed edges**. For example, if phenotypes arising from mutations in genes A and B differ and the double mutant phenotype most closely approximates gene A, we may draw a directed edge from gene A to gene B. With the aid of other prior biological knowledge, such directed edges can be useful in assigning order of action in biological pathways [8,65]. There exist many other types of genetic interaction [28,114].

Another relationship type is sequence homology (describing two genes with similar sequences that each descended from a common ancestor). Other biomolecular relationships may be derived from shared biomolecular properties, for example the relationship of **similar mRNA expression profile** [130], **similar phenotype** [39,45], or **similar gene function** [72]. Because there are many different properties of biomolecules, many 'similar-property' networks are possible. Above, we sampled types of biomolecular interactions or relationships used in network studies to date. However, this is not meant to be an exhaustive list, as there are many more relationship types, some unknown to us, and some for which large-scale data is not yet available. As more large-scale experimental methods are developed and applied to biological research, more biomolecular relationships can be represented as networks.

A few special types of graphs have particular relevance to the study large-scale biomolecular networks. These include **bipartite** and **quasi-bipartite** graphs. A **bipartite** graph is a graph with two disjoint vertex sets, such that no edges connect pairs of vertices within the same set. For example, a bipartite graph might consist a set of gene vertices and a set of protein vertices, together with edges that each correspond to a relationship described by "gene X encodes protein Y". If one of the two vertex sets contains within-set edges, while the other does not, the graph is called **quasi-bipartite**. Certain biomolecular networks are intrinsically quasi-bipartite. For example, transcriptional regulatory networks contain edges pointing from genes encoding transcription factors to the regulated genes. Thus, vertices may be divided into transcription factors and non-transcription factors. While the latter vertex set, by definition, does not contain any within-set links, a transcription factor may regulate the expression of another transcription factor, so that the network is quasi-bipartite. Quasi-bipartite graphs can also arise as a simple consequence of experimental design. Protein interaction networks derived from yeast two-hybrid or genetic interactions derived from systematic genetic array (SGA) analysis are two examples, where only a subset of genes have been used as 'baits' or 'queries'. In network analyses, care must be taken to avoid discoveries that are trivial conse-

quences of bipartite or quasi-bipartite properties arising from the experimental design.

Tools and Methodology

Studying global structural features of a network involves many different measures of graph properties [25]. Here we survey a few frequently used measures below.

Degree (or **connectivity**) is the most basic characteristic of a vertex. The degree of vertex V, $k(V)$ is defined as the number of edges incident on vertex V, and tells us how connected V is with other vertices in the network. In the case of directed networks, there is an **outgoing degree** (or **out-degree**) $k_{\text{out}}(V)$, denoting the number of edges that start from V, and an **incoming degree** (or **in-degree**) $k_{\text{in}}(V)$, denoting the number of edges that end at V.

The **clustering coefficient**, $C(V)$, measures the connectivity amongst the neighbors of a vertex V, and can be defined as the number of edges among V's neighbors, divided by the total possible number of edges if all the neighbors were connected to one another. The average clustering coefficient of a network characterizes the general tendency of vertices to form clusters [129].

In a network, **distance** between two vertices U and V, $d(U, V)$, is measured as the number of edges in the shortest **path** from U to V. In an undirected graph, $d(U, V)$ satisfies non-negativity, symmetry and the triangle inequality of a distance metric. In the case of directed graph, however, all edges in a path must be in the same direction. While non-negativity still holds, the distance $d(U, V)$ from U to V could be different from the distance $d(V, U)$ from V to U so that symmetry is not necessarily satisfied, nor is the triangle inequality. Hence, strictly speaking, $d(U, V)$ is not a distance metric in directed graphs. The overall connectedness of a network can be quantified by its **characteristic path length**, defined as the average distance between pairs of vertices in the network.

A **subgraph** of a graph G is a graph whose vertex and edge sets are subsets of those of G. A **component** is a subgraph such that all vertices are connected to one another by some path, and there are no connections to vertices outside of that subgraph. If a component contains a majority of a graph's vertices it is referred to as the **giant component.** Other subgraphs of interest are **cliques**, a set of completely connected vertices. Such dense subgraphs bear particular relevance to protein complexes, which often present themselves as cliques, or dense but incompletely connected subgraphs, in a protein network of physical interaction. One can 'seed' the search for such dense subgraphs in the network (as a proxy search for protein complexes) using **k-cores** – subgraphs such that each vertex has at least k neighbors within that subgraph [9].

Global Structural Features

Topological features of biomolecular networks have been subjected to extensive study (reviewed in [16]). Below we describe some features of particular relevance to network structure and function.

Power-Law Networks In a biomolecular network, some vertices are better connected than others. To characterize how connectivity varies from vertex to vertex, one can calculate the degree distribution – i. e. the distribution of $P(k)$, the fraction of vertices with degree k. The Erdös–Rényi (ER) model describes random networks in which each vertex pair is assigned an edge randomly with a probability that is fixed for all vertex pairs. In other words, each edge is assigned or not assigned according to the outcome of a 'weighted coin flip' (also known as a Bernoulli trial). For such ER networks, $P(k)$ follows a binomial distribution, but can often be approximated well and conveniently by a Poisson distribution. The degree distributions of biomolecular networks (and many other complex networks such as food webs and social networks) tend to fall off at high degree much more slowly than a Poisson, and often follow a **power law** distribution $P(k) \sim k^{-\lambda}$ (Fig. 2) [15]. The degree exponent λ in most biomolecular networks seem to have a value between 2 and 3 [15]. The most intuitive (and perhaps most useful) feature of this power law distribution is that most vertices in the network have only a few edges, while a few vertices have many edges. These highly connected vertices are called **hubs**, which are sometimes described to 'hold the network together' since removal of a hub has a particularly high tendency to break the network up into separated components. Networks with a power-law degree distribution (or sometimes networks that only follow a power-law in the high-degree tail of the distribution) are called **scale-free** networks. The term scale-free is meant to indicate that there is not a typical vertex in the network that can characterize the rest of the vertices. For some random networks, many vertices have degrees equal or close to the average degree; however, it should be noted that ER networks with a low edge probability can be easily confused with power-law networks [47]. Arguments have also been made that **scale-rich** is a more appropriate description of graphs with a power-law degree distribution [119]. For clarity, we prefer the term **power-law network**.

Networks of metabolic reactions were the first biomolecular networks to be characterized as power-

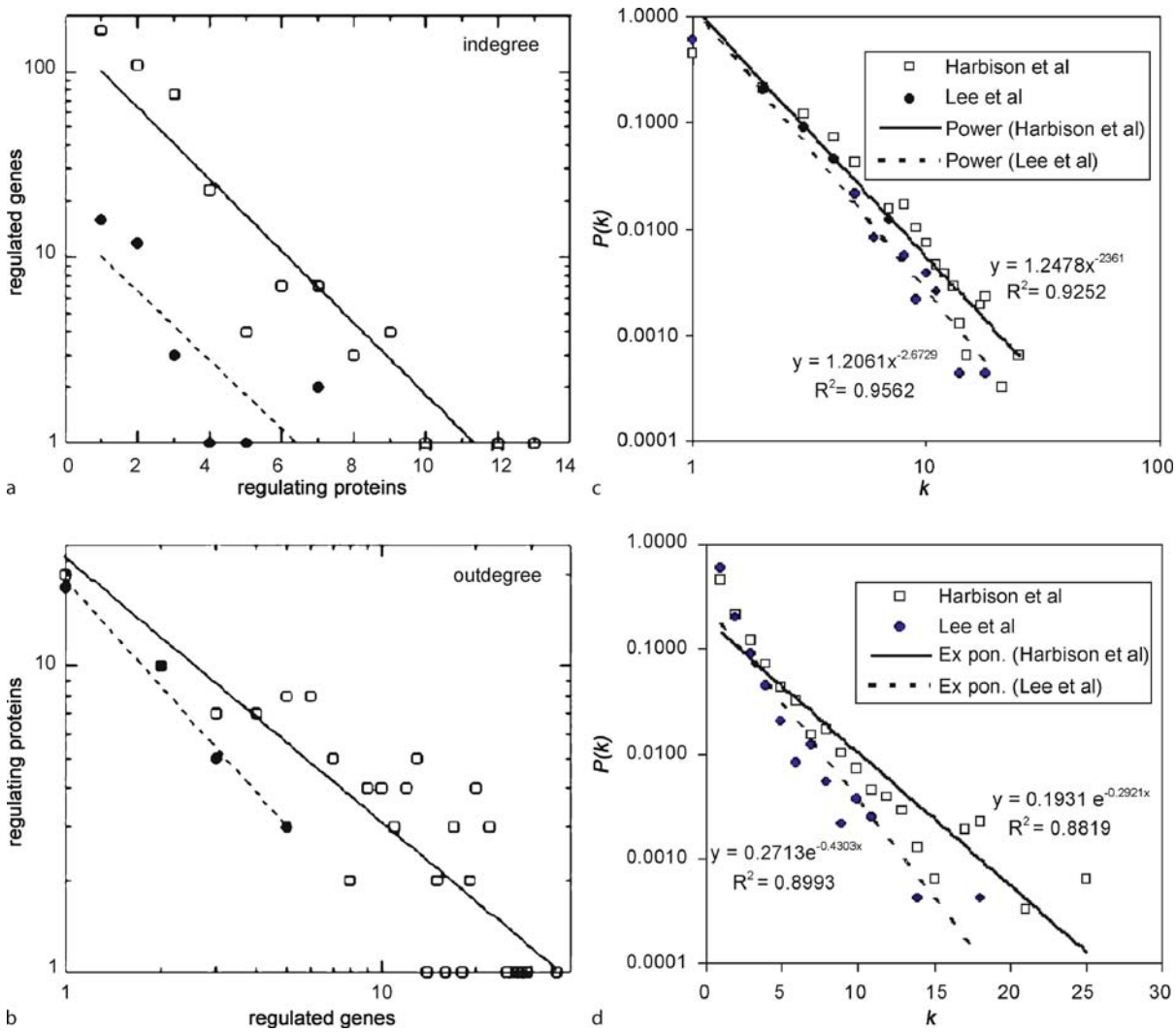

Biomolecular Network Structure and Function, Figure 2

Degree distribution of transcriptional regulatory networks. a/b incoming and outgoing degree distributions of a manually curated network; c/d incoming degree distribution of two large-scale transcriptional regulatory maps, fitted with power-law (c) and exponential (d)

law [64,127]. Jeong et al. analyzed the metabolite networks of 43 organisms across all three domains of life – bacteria, archaea, and eukaryotes, and discovered that despite significant variation in the components and pathways, all metabolite networks have similar topological scaling properties [64]. As many biochemical reactions are irreversible under physiological conditions, metabolite networks should be represented as directed graphs. The distribution of both in-degrees and out-degrees follow power-law distributions, i. e. $P(k_{\text{in}}) \sim k_{\text{in}}^{-\lambda_{\text{in}}}$, $P(k_{\text{out}}) \sim k_{\text{out}}^{-\lambda_{\text{out}}}$, and the values of both λ_{in} and λ_{out} lie within a narrow band of 2.0 to 2.4 [64]. This demonstrated that a power-law degree distribution is a common feature of cellular metabolism, with most substrates involved in one or two reactions, while a few molecules, such as coenzyme A, function as metabolic hubs by participating in several dozen reactions.

Other biomolecular networks have also been described as power-law, including networks of physical interactions between proteins, transcriptional regulatory networks, and networks of SSL genetic interactions. The protein interaction map of the yeast *Saccharomyces cerevisiae* as measured by systematic two-hybrid screens [56,125] exhibits a power-law degree distribution, as do physical interaction maps of multi-cellular organisms – the worm *C. elegans*, the fly *D. melanogaster*, and hu-

man [41,63,74,126,135]. The degree distribution of correlated expression networks based on microarray data has been shown to follow power-law [1,32]. The outgoing connections from transcription factors to target genes also exhibit power law behavior (Fig. 2b), [44,73,121]. The incoming degree distributions in these networks have been found instead to follow an exponential decay according to manually curated data (Fig. 2a), [44,121]. However, analysis of more recent large-scale data [49,73,78] suggests that power-law distribution may provide a slightly better fit (Fig. 2c,d). The resolution of this awaits further refined data and analysis, but it is clear that the distribution of in-degrees and out-degrees in a directed graph may behave differently. Similar to directed networks, the degree distribution of quasi-bipartite networks, should also be examined separately for the different vertex sets.

Closer examination of several biomolecular networks showed that their degree distributions exhibit better fit to a truncated power law than a power-law model [6,99]. A truncated power-law distribution follows a power-law distribution with an 'exponential envelope', such that power-law behavior dominates the distribution at low degrees, while the exponential function dominates the distribution at the high degree tail of the distribution. Truncated power-law behavior may arise from constraints on the number of interaction partners for high degree vertices [6]. However, biological implications of the differences between these degree distributions remain to be demonstrated in most cases. It is nevertheless clear that biomolecular networks share the common feature that a few vertices have many edges, while most vertices have only a few edges.

The ubiquity of power-law and presence of hubs in biomolecular networks prompts us to ask what advantage such network architecture might offer. Some argue that the answer lies in robustness. Intuitively, random removal of a considerable number of vertices could disintegrate a network, breaking it down into small, disconnected groups of vertices. This is true for Erdös–Rényi random networks [16,117]. Power-law networks, such as the electrical power grid, are extremely resilient to accidental vertex failures. Even if 80% of the vertices are randomly removed, the rest often still form a connected component. This is confirmed by both numerical and theoretical analyses examining how removing vertices of varying degrees impacts the average path length and size of the giant component [4,20,21,23]. As most vertices in power-law networks have small degrees, any vertex that fails is most likely one of them and hence the damage will be negligible. However, such robustness is accompanied by the weakness that deliberate attacks on hubs are much more likely to create disconnected components and hence particularly devastating. Some believe this notion may apply beyond power grids to biological networks, and we discuss this possibility further in the Applications section.

A related question is the evolutionary origin of power-law networks. Building upon Simon's early work [17,104], Barabasi and Jeong proposed a growth and preferential attachment model [15]. Under this model, the network must emerge from gradual addition of vertices and edges. Preferential attachment specifies that the newly added vertices connect to the rest of the network by preferentially attaching themselves to highly connected vertices (e. g., with a probability proportional to the degree of the target vertex). This rich-get-richer mechanism offers an explanation for hubs. In protein networks, growth and preferential attachment may result from gene duplication, which produces initially-identical proteins with the same interaction partners. The more interactions possessed by a protein, the more likely it is to acquire a new interaction through duplication of an interaction partner, resulting in the rich-get-richer phenomenon associated with preferential attachment.

Neighborhood Clustering **Neighborhood clustering** describes a network characteristic that resembles the phenomenon of neighbors of my neighbors tend to be neighbors of one another. Above we defined a clustering coefficient to quantify this local property of a graph. The average clustering coefficient of all vertices in the graph characterizes the global tendency towards **clustering** in the network [129]. Several biomolecular networks, for example, the metabolic network [96] and physical interaction networks, have a significantly higher average clustering coefficient than random networks of equivalent size and degree distribution, suggesting modular organization in addition to power-law degree distribution. A partially-determined human interactome network, however, was found to be less clustered than random networks with the same degree distribution [99]. This could be because only a subset of edges in the complete network was sampled, or it might suggest 'clustering' of protein-protein interaction networks in other organisms have been artificially boosted by previous experimental methods [99]. It is also possible that the human interactome network indeed is less clustered than other organisms. Further experimentation is required to resolve this apparent discrepancy.

Small World Another important feature of biomolecular networks is that they are frequently small-world. A small-world network is defined by two properties: 1) short characteristic path length, meaning that the length

of the shortest path between any pair of vertices tends to be very small (this is sometimes confusingly called "the small world property," despite there being two properties of a small-world network); 2) high clustering coefficient, i. e. densely connected local neighborhoods as described above [129]. Random networks following the ER model satisfy the first requirement, but not the second, whereas some regular lattices possess the second property, but not the first. Both metabolite networks and physical interaction networks have been shown to be small-world [33,110,127]. Short characteristic path length in small world communication networks allows efficient signaling between vertices, in that local perturbations can propagate rapidly to the entire network. Small-worldness seems to be evolutionarily conserved, as the metabolite network of a parasitic bacterium exhibits roughly the same average path length as observed for a large multi-cellular organism [64,127].

Disassortativity In some networks, highly connected vertices tend to link to vertices with few edges, avoiding connections to other hubs. This is the disassortative property. Disassortativity is shared by both metabolic networks and protein-protein interaction networks, as well as technological networks such as the world wide web, but not in social networks which are assortative [80,87,91]. The avoidance of hub-hub connections may come from the compartmentalization of many cellular processes [50], although other explanations have been proposed. For example, disassortativity may provide some protection against deliberate attacks to the hub vertices. As the robustness of power-law networks comes with an increased vulnerability to deliberate attacks against hubs, suppression of hub-hub connection has been suggested to alleviate or slow down propagation of such deleterious perturbations [80]. While it is difficult to know how applicable the communication network analogy is to biological networks, these ideas are certainly thought-provoking.

Network Motifs The average clustering coefficient discussed above can be expressed in terms of the number of triangles in a network relative to three times the number of paths of length two [88]. It is then interesting to ask whether there are other local patterns of connectivity that are over-represented within biological networks. Such patterns are called **network motifs**. Demonstrating enrichment of a given connectivity pattern requires counting the number of subgraphs in the network matching the pattern and comparing the count to that observed in many randomized versions of the same network [83].

Milo et al. initially proposed that network motifs represent elementary building blocks of cellular networks. The transcription regulatory network in *E. coli* and *S. cerevisiae* both exhibit enrichment for the feed-forward loop and the bi-fan (loop Fig. 3), which can be attributed to specific tasks in information processing [83,103]. In particular, the feed-forward motif (Fig. 3) involves two transcription factors, one regulating the other and both controlling the expression of a third target gene. The feed-forward motif can serve as an AND-gate that corresponds to a low-pass filter which suppresses out transient signals while preseving persistent signals, as is the case in the araBAN operon in the arabinose feed-forward loop [101].

Different types of networks exhibit distinct sets of motifs. For example, the food web exhibits a 'three-chain' motif and a bi-parallel motif, in contrast to the feed-forward motif and bi-fan motif in transcriptional regulatory networks (Fig. 3), [83]. Thus network motifs provide insight into common mechanisms of interaction within a network. The abundance profile of network motifs for a given network serves as a fingerprint for classifying networks into different superfamilies [82].

A given network motif may have many instances within a network. The relationship of network motif instances to one another poses an interesting question. Observations made in the *E. coli* transcriptional regulatory network show that instances of both the feed-forward motif and the bi-fan motif do not exist in isolation. Instead, they tend to overlap with one another and form large clusters, which then coalesce into superclusters [26]. This suggests hierarchical organization of biomolecular networks, and that motifs should not necessarily be viewed as elementary building blocks of complex networks, but might instead be thought of as signatures of higher order structures corresponding to distinct cellular processes. The fact that many motif clusters overlap with known biological functions [26] interferense between motif instances. We further discuss this notion below.

Modularity and Hierarchy Clustering of network motifs suggested presence of higher order structures in biomolecular networks [26]. The search for enrichment of larger patterns (involving more than four vertices) in large networks is computationally challenging due to the exponential computational complexity associated with enumerating increasingly large subgraphs in biomolecular networks. An alternative approach to brute force computation or sampling methods is to identify densely connected clusters, or network modules. Various clustering methods have been used to characterize modules using network topology [42,53,96,97,102,109,113]. Note that the

Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Gene regulation (transcription)			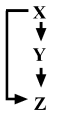		**Feed-forward loop**			**Bi-fan**
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41

Biomolecular Network Structure and Function, Figure 3
Single-color network motifs

ideal module is a collection of gene products that work together as a unit, which may plug in to perform its molecular function in different contexts. Densely connected clusters may be useful in identifying modules, but we should separate the ideal concept from its surrogate measure [50].

The concept of relatively isolated modules may seem contradictory to the power-law characteristic of biomolecular networks. Both naturally exist [95,96], however, indicating network modules do not exist in isolation, but combine at various levels to form a hierarchical network. The hierarchical structure of biomolecular networks poses a challenge in identifying modules, as there are no clear boundaries between them. However, understanding networks' hierarchical organization can also aid the module identification, at least in the transcriptional regulation network. For example, the transcriptional regulatory networks of *E. coli* and *S. cerevisiae* both exhibit a multilayer hierarchical structure [77,136], and removing global regulators in the top layers of the hierarchy decomposes the network into functionally relevant modules [77]. Recent work also identified 'origons', or 'organizer' structures within the transcriptional regulatory network that originate at a distinct class of sensor transcription factors and serve as functional modules for various environmental perturbations [13,31]. Combining topological information with other genomic or proteomic data can lead to more biologically relevant modules [9,14,55,62,65,118,123]. We describe this further in the next section.

Applications

Structural features of biomolecular networks can be used to infer organizing principles of biomolecular interactions and relationships. The power-law nature suggests robustness of biomolecular networks against random attacks, and models of network designation evolution – e. g. growth and preferential attachment – serve as additional support to the gene duplication model of protein network growth. The small-world designation suggests the potential for rapid communication within biomolecular networks, with local perturbations reaching the entire network via a small number of links, while also reflecting clustering due to modularity, e. g. as a result of protein complexes. Network motifs are local interconnection patterns that are often footprints of higher order structure (i. e., functional units composed of more vertices than are considered in network motifs). Different types of biomolecular network have distinctive profiles of the relative abundance of network motifs, so that network motifs can also be used to characterize and compare complex networks.

More practically, structural characteristics of biomolecular networks can be used to make biological inferences. Experimentally derived networks are inevitably susceptible to random as well as systematic errors, and characteristics of network structure can be used to assess confidence of edges in biomolecular networks and also make novel predictions. Goldberg et al. examined a list of *S. cerevisiae* protein-protein interactions and were able to stratify this list into many subsets of varying accuracy by exploiting the neighborhood clustering feature of small-world networks [43]. Other methods using network topology also proved useful in predicting protein interaction networks of other organisms [41], for example, by exploiting conservation of network motifs [2].

High clustering in biomolecular networks often represents physical organization of biomolecules, particularly in the case of physical protein interaction maps. Therefore, identifying densely connected subgraphs (clusters) can help detect protein complexes [10,66,97,113]. In addition, topological modules are believed to represent specific biological processes or functions [9,14,55,59,118,123]. Hence, identifying modules can also help assign biological processes or functions to genes in a network.

Features of vertices can also be inferred from network topology. As discussed above, power-law networks are robust with regard to random vertex failures, but highly sensitive to targeted attacks against hubs. Whether this notion of robustness applies beyond electrical power grids to biological networks is subject to debate. In protein networks that include co-complexed relationships, for example, many proteins are hubs by virtue of their membership in protein complexes (appearing as cliques). Re-

moval of such hubs is not highly disruptive in terms of generating disconnected components, and yet there are many essential complexes for which the removal of a single protein member disrupts the physical assembly of the whole complex and results in death for the organism. Nevertheless, genome-wide deletion analysis in *S. cerevisiae* demonstrated a correlation between degree and essentiality: only ∼10% of the genes encoding proteins with degrees less than 5 are essential, as compared to ∼60% for genes encoding proteins with 15 or more interactions [39,63,131]. Thus, protein connectivity can to some extent predict the phenotype of its corresponding gene. Recent work on metabolic networks also demonstrated that certain topology-based vertex characteristics can accurately predict knock-out phenotypes of the corresponding enzyme [133].

In the case of SSL genetic interaction networks, highly connected genes are often more critical to fitness than genes of less connectivity as they 'buffer' a large number of genes to maintain the cell's survival [122]. Upon disruption of a hub gene in the SSL network, mutations in an increased number of genes will result in fitness defect. It has been hypothesized that hub genes could be useful targets for anti-cancer drugs to specifically kill cancer cells as they often carry many additional mutations, while normal cells can be protected due to genetic buffering by the hub gene [122].

Examples of structural features and applications discussed thus far have focused on networks composed of a single type of relationship. However, as there are many biological interaction or relationship types, it is worthwhile to combine information from various biomolecular networks. In the next section we discuss multi-color networks, that represent more than one biological interaction/relationship simultaneously.

Multi-Color Networks

Definition and Categorization

Previously we described various types of single-color networks, each containing edges that correspond to a single type of relationship between biomolecules. Different types of interactions, however, do not exist independently. Instead, they are inter-related with one another and offer different perspectives on the underlying biological mechanisms. It is therefore useful to consider multiple biomolecular relationships simultaneously by integrating them into one **multi-color network** (Fig. 4). Each edge type can be represented by a color that indicates the type of biological interaction or relationship. Here we focus on multi-color networks where each vertex corresponds to a gene and its corresponding protein(s). As a multi-color network can be viewed as an overlay of multiple single-color networks, the structural features of single-color networks discussed previously still hold true for the individual component single-color networks, but additional properties of multi-color networks emerge.

Features and Implications

Correlation of Edge 'Colors' The first step in studying multi-color networks is to examine how various edge types correlate to one another. Various statistical methods can be used to test significance of correlation. Without considering network topology, one can treat pairs of vertices in the network as independent data points, and assess significance of overlap between two edge types analytically under the null hypothesis of the hypergeometric distribution. Network topology, however, can have significant impact on edge correlation. For example, as most single-color biomolecular networks contain highly-connected hubs, if many of the hubs in one network are hubs in another network, the overlay of the two single-color networks will, as a result, have considerable overlap of the two interaction type. To test edge correlation more rigorously in the context of network topology, we can generate randomized versions of the multi-color network by rewiring edges in each component single-color network independently, while preserving certain topological features, for example, degree distribution [134,138]. We can then determite the significance of correlation by calculating the fraction of random networks that contain more two-color edges than the real network. This approach can also be generalized to assess significance of overlap of more than two edge colors.

Several studies have noted the correlation between different types of biological interactions. For example, interacting proteins are more likely to have similar expression pattern [36,60]; genes with correlated expression are more likely to be controlled by a common transcription factor [137]; and SSL genetic interactions are more likely to occur between homologous genes [122]. Although each of these correlations are statistically significant, no single pair of interaction types shows perfect overlap (and indeed the level of correlation is low in each example). Thus there is dependence amongst many of the observed types of biomolecular interactions, and yet each type offers a distinct view of the same complex biological system.

'Multi-Color' Network Motifs and Themes The network motif concept, described previously for single-color networks, is extensible to a multi-color network. Multi-color network motifs characterize relationships between

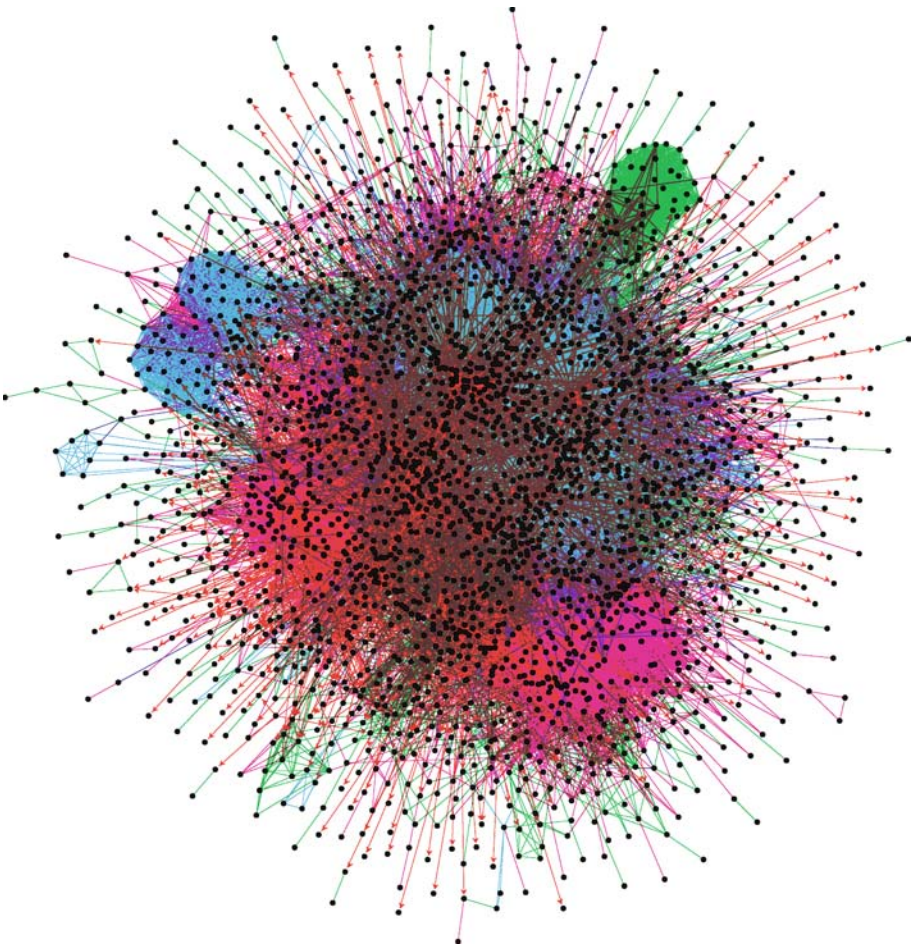

Biomolecular Network Structure and Function, Figure 4
A multi-color network

different biological interaction types within local network neighborhoods. Yeger-Lotem et al. examined network motifs in combined cellular networks of two interaction types – transcriptional regulation and protein interaction [134]. They identified three multi-color motifs of three genes – two interacting transcription factors that co-regulate a third gene, two interacting proteins co-regulated by the same transcription factor, and a 'mixed feedback loop', where a given transcription factor and the product of its target gene form a protein interaction with each other. More multi-color motifs were characterized in an integrated network of five interaction types – transcription regulation, physical interaction, SSL genetic interaction, sequence homology, and correlated mRNA expression Fig. 5 [138]. These motifs can be explained in terms of high-order network structures, or network themes (Fig. 5), which are representative of the underlying biological phenomena. One interesting example is motifs containing two vertices linked by protein interaction or correlated expression, such that both vertices connect to a third vertex either by SSL interaction, or by sequence homology. This suggests the theme that a protein and a distinct protein complex/biological process have compensatory functions, resulting in SSL relationships between the protein and any component essential to the compensatory complex/process. Many four-vertex motifs that correspond to the compensatory complexes theme (Fig. 5) further support this notion. These four-vertex motifs also suggested that some instances of the 'compensatory protein-complex/process' theme could simply be signatures of compensatory pairs of complexes or processes, rather than a single protein acting alone to compensate for loss of another process. An excellent example of the compensatory complexes theme is the pairing of endoplasmic reticulum (ER) protein-translocation subcomplex [48] and the Gim complex [37], with each

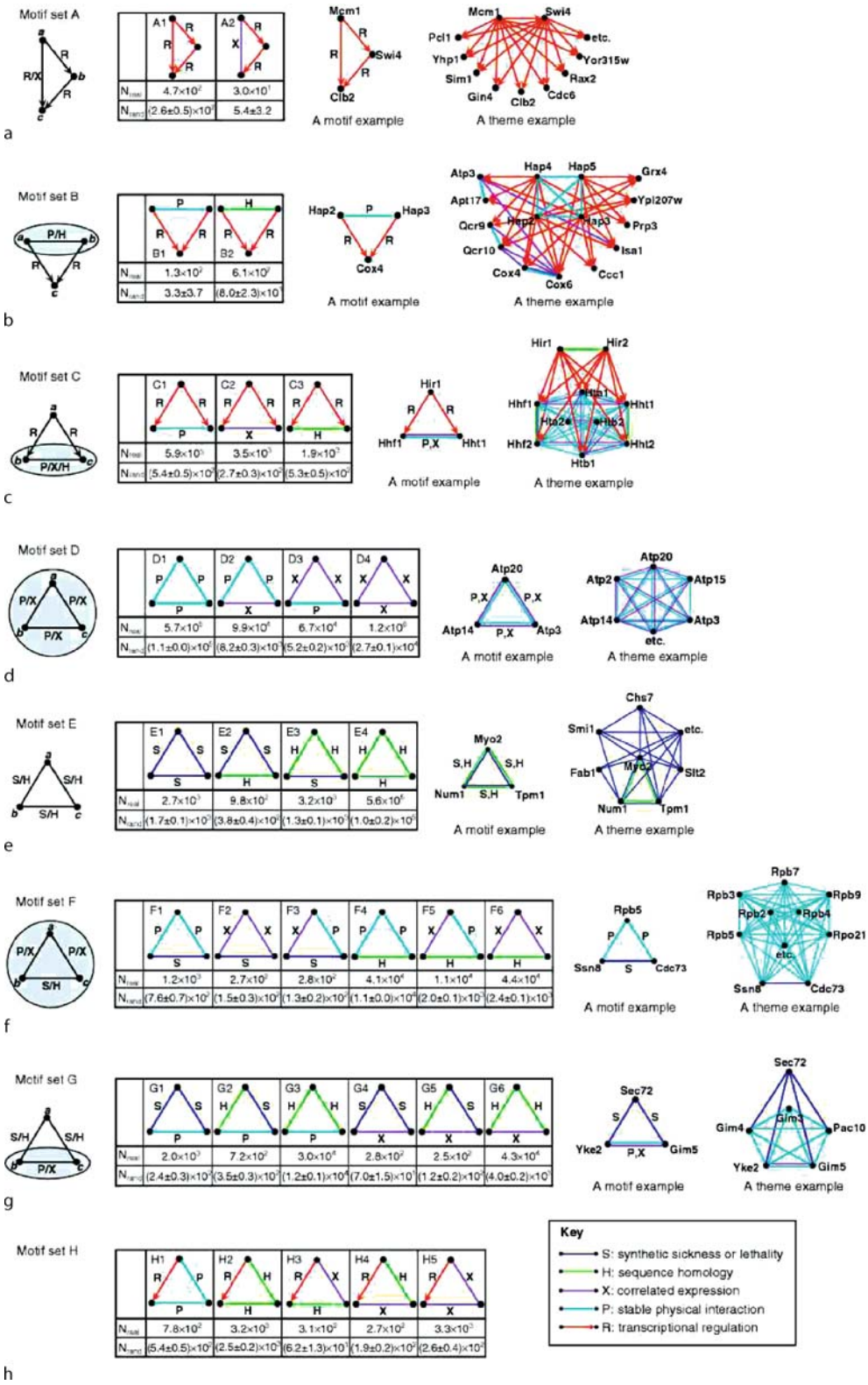

Biomolecular Network Structure and Function, Figure 5
Multi-color network motifs and themes

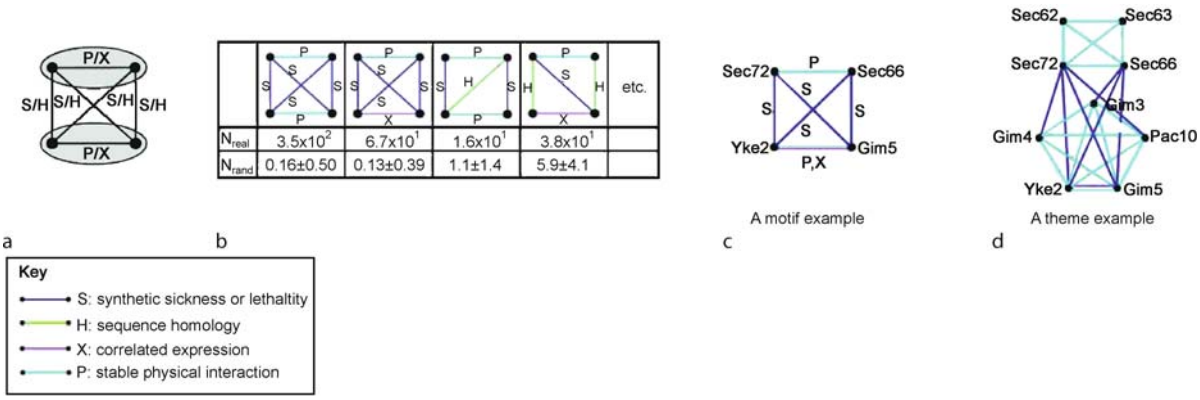

Biomolecular Network Structure and Function, Figure 5
continued

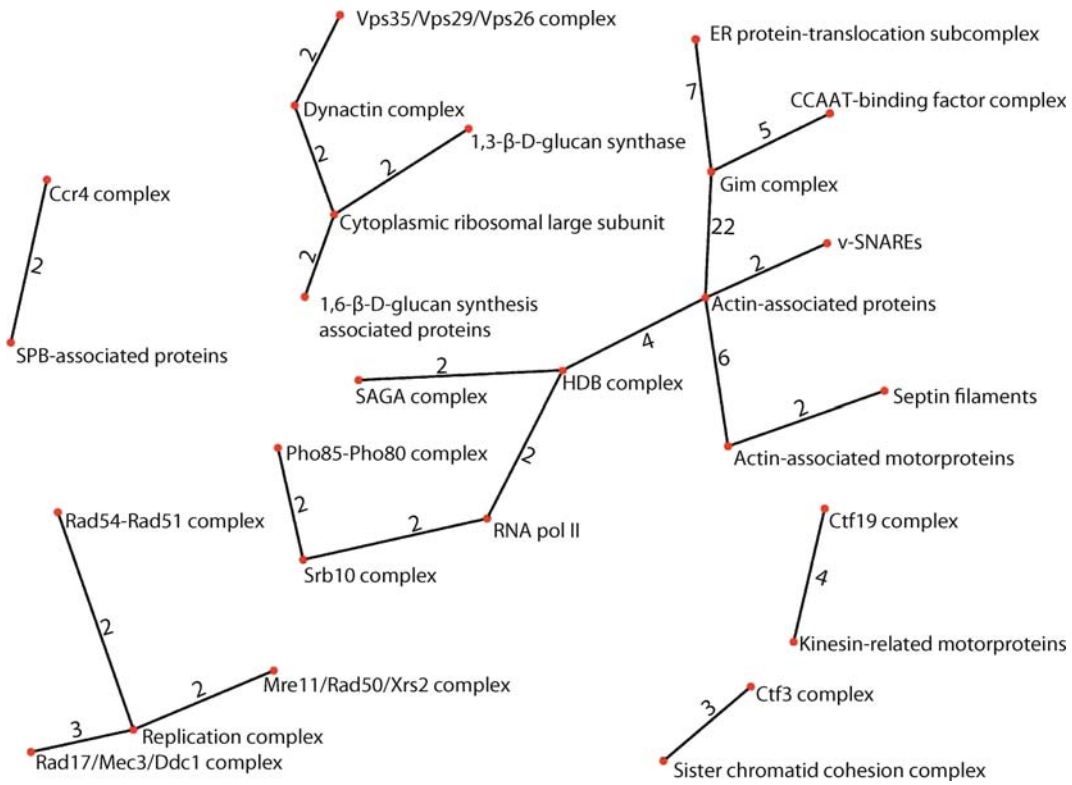

Biomolecular Network Structure and Function, Figure 6
Network themes

complex connected to the other by many SSL interactions [122,138].

Thematic Maps In the previous section, we described the idea that network motifs may in general be footprints of larger network structures. Organization of some multi-color network motifs into larger structures of network themes further supported this notion, and also confirmed that the concept of modularity discovered in single-color networks is extensible to networks of multiple interaction types. As such modules in a multi-color network can exist in various sizes, and exhibit a combina-

torial number of different connection patterns, systematically identifying them by prevalence of network motifs of a fixed size can be impracticable. Network theme representations, in which a simplified network is drawn in terms of specific network themes [138], is a convenient way of examining the system. In the example of the compensatory complex theme, we can collapse members of the same complex into one single vertex, and draw links between complexes with significantly enriched inter-complex SSL interactions [138]. The resulting thematic map of compensatory complexes (Fig. 6) serves as a guide to the 'redundant systems' within the integrated *S. cerevisiae* network [138]. Similarly, one can create maps of 'regulonic complexes' – complexes for which the corresponding genes are co-regulated – by linking each transcription factor to protein complexes that are significantly enriched for genes targeted by that transcription factor, providing a schematic view of the transcriptional circuitry in the cell [105,106,138].

Applications

Multi-color networks integrate several biomolecular interaction types and contain information that is richer than the component single-color networks taken individually. Hence they can be used to make more accurate biological inference. Correlation between biological interaction types alone can be useful, especially when a large number of interaction types are combined together. A Bayesian approach achieved considerable success in predicting physical interactions between proteins [59]. To capture the additional information from interdependency of interaction types, one can examine the conditional probability of a physical interaction given all possible 'multi-color relationships' (each corresponding to a particular configuration of single-color relationships) [59,83]. However, such an approach scales poorly as the number of interaction types grows (the number of multicolor relationships grows as 2^N, where N is the number of interaction types, so that there will be increasingly many multicolor relationships with no training examples). More sophisticated methods, such as logistic regression [12,41], decision trees, or random forests, can model the interdependency between interaction types in a more scalable manner, allowing the combination of a large number of biomolecular relationships. Studies using such methods can significantly improve the quality of predictions [59,61,93,132,139].

Multi-color network motifs put correlations of interaction types in the context of local topology, and provide potential to further improve biological inference. Specifically, one can predict a certain type of link between a given pair of vertices if its addition completes a structure matching an enriched network motif. For example, two genes with a common SSL interaction partner may have an increased probability of protein interaction, as addition of a physical interaction edge results in a match of a motif depicted in Fig. 5 (G1). Similarly, a SSL edge between two genes can complete a match to the same motif, if the two genes are indirectly connected by physical interaction followed by SSL. This 'two-hop' relationship of physical interaction and SSL proves to be a strong predictor of SSL interactions [132].

More broadly, we can also predict an interaction if its addition fits into a recurring network theme. For instance, there are many more SSL interactions between the ER-translocation complex and the Gim complex than would be expected by chance given the out-degrees of both complexes in the SSL network. However, no SSL interaction has been observed between Sec62 or Sec63, two members of the ER protein-translocation subcomplex, and any protein in the Gim complex because Sec62 and Sec63 were not used as queries in the SGA analysis [122]. We can therefore hypothesize that Sec62 or Sec63 are SSL with many members of the Gim complex [138].

Network themes represent modular organization of the network at the functional level, and can be used to predict the function or biological process in which a gene is involved. The rationale is that if a gene is part of a module where a particular function is enriched, then there is increased probability that this gene also holds the same function. For example, in the feed-forward theme depicted in Fig. 5, most of the genes regulated by both Mcm1 and Swi4 are involved in the control or execution of the cell cycle. Therefore, we can hypothesize that Yor315w, a protein about which little is known, is also involved in cell cycle regulation. Indeed, this protein is a target of cyclin-dependent kinase Cdk1, a primary regulator of the cell cycle. Predictions based on network themes may be robust to errors in the input data, since they depend on connectivity patterns in extended network neighborhoods instead of one or very few links.

It is worth emphasizing that predictions based on network motifs and themes are more suggestive than definitive. Above we focused on organizational principles that can be utilized for biological inference. Such inferences should be combined with prior information from as many sources as possible, and should be tested experimentally. In the next sections, we will touch upon rich network models that can be used for graphical data integration. We also refer readers to the chapter ► Biological Data Integration and Model Building in this book for more general approaches to data integration.

Rich Network Models

Networks of Labeled Vertices and Edges

Just as the relationships between biomolecules come in many colors, the biomolecules themselves (the vertices) also exhibit a range of properties. These include intracellular localization, mRNA expression level, cellular function, and protein complex membership, to name just a few. A network labeled vertices and edges captures multiple properties of biomolecules as well as relationships among them. Extending the concept of multi-color networks, one can build a **rich network model** using a network with vector-labeled vertices and edges. Each vertex or edge can be labeled with a vector, such that each variable in the vector represents some vertex property or edge type. The rich information captured by such a model can be used to infer types of interactions or characteristics of vertices [38].

Hypergraphs Previously we mentioned that sharing of the same or similar vertex property can be viewed as a type of biomolecular relationship, and hence represented using an edge. A **hypergraph** can be used to more explicitly represent a shared property of multiple vertices. A hypergraph is a generalization of a graph where edges can connect any number of vertices, instead of only two as in conventional graphs, and a hypergraph edge (or 'hyperedge') can be used to connect all vertices sharing the same property. When further generalized to contain multiple types of edges, hypergraphs provide an alternative to networks of labeled vertices and edges described earlier. Hypergraphs have been used, for example, to model single-color networks of protein domains [34]. However, it remains to be seen whether this provides a more useful abstraction than a rich network representation in which vertices and edges are each labeled with vectors.

Weighted Graphs and Probabilistic Networks Network models discussed thus far have mostly been discrete. For example, an edge in each single-color network either exists or does not exist. In the rich network models we discussed, variables used to describe a vertex or edge characteristic can be binary or categorical variables. However, some properties of a vertex may have continuous values, for example, a gene product may be described by its corresponding mRNA expression level. An edge property may also be continuous, e. g., affinity of a physical interaction or coefficient of correlation between mRNA abundance profiles. Such quantitative information can also be incorporated in a rich network model, simply by including continuously-valued variables within the vertex or edge vectors.

One special case of a weighted graph is a probabilistic network, in which the weight on each edge represents a posterior probability that the edge is real, given the evidence supporting that edge. Probabilistic networks are particularly useful when knowledge of relationships is uncertain. This is often the case for biological data, sometimes due to inaccuracy of measurement or varying levels of confidence in a computationally inferred relationship [7,139]. Probabilistic networks have been used to predict novel members of partially known protein complexes, for example, by sampling from networks of physical interaction from multiple data sources based on their corresponding confidence levels [7,11]. Algorithms have also been developed to identify protein complexes by finding dense subgraphs within a weighted graph [75].

Future Directions

Research on the structure and function of biomolecular networks has covered a lot of ground in the past decade. So far, we have studied global topological features such as power-law degree distributions, small-worldness, disassortativity and associated biological implications such as robustness and dense neighborhood clustering. We have also discussed the hierarchical organization of networks, with network 'structures' ranging from network motifs to dense clusters, to pairs of clusters, densely connected within each cluster by one 'color' of interaction, and connected to one another by interactions of a different color. We have also explored the roles of these structures in cellular functions; for example, the involvement of certain network motifs in information processing and the correspondence of dense clusters to protein complexes and biological processes. We also recognized the interplay of various interaction types, for example, the correspondence of pairs of clusters connected by SSL genetic interaction to protein complexes with mutually compensatory functions. Deeper understandings in the above areas certainly warrant future study. However, many currently under-explored territories remain.

The cell is a constantly changing environment, so are the molecules and processes within that environment. Network studies have been mostly limited to the union of interactions in all locations, times and environments. Cellular state, such as position in the cell cycle, is a key determinant of concentrations and status of biomolecules, and hence the interactions between them. Studies have begun to reveal the dynamic nature of biomolecular networks. For example, by overlaying mRNA expression patterns with the physical interaction network, Han et al. described two types of highly connected hubs – 'party'

hubs, which interact with most partners simultaneously; and 'date' hubs, which connect to different partners at different time [46]. Expression data have also been combined with protein interaction data to reveal cell cycle-dependent protein complexes or arrangements of protein complexes [24]. Genome-scale molecular data across a wide variety of cellular states are so far only available for mRNA expression levels [22,112]. Availability of more biomolecular interaction data under various cellular conditions will allow us to study the additional complexity of cellular processes. Such research will require adding another dimension when modeling biomolecular networks, and will likely also require new mathematical tools.

In addition to overall cellular state, intracellular environment also bear significance on cellular interactions. For example, proteins may be restricted to particular compartments, either defined by a membrane as are the nuclei or mitochondria, or in less-clearly delimited regions such as nucleoli. Such features can be viewed as vertex properties in the network, and can therefore be modeled either using a hypergraph or by labeling vertices. Cellular localization data are already available at a genomic scale [54,70], and should yield additional insights.

One major limitation of network analysis is data collection. Even with the rapid development of high-throughput genomic and proteomic experimentation, we are still far from being able to systematically characterize all biomolecular features and interactions. Quantifying them at high resolution in both space (across subcellular locations and across cell and tissue types) and time (e. g., through the course of organismal development) appears a more distant goal. More comprehensive network data will undoubtedly advance the study of the structure and function of biomolecular networks, and the knowledge can in turn aid experimental discovery. One intuitive example is to use inferences made from network study to apply limited resources to experiments that are more likely to yield discoveries. We are optimistic that experimental discovery and network analysis can form a positive-feedback loop, with each reinforcing the other, and together advanciny our understanding of biomolecular networks.

Bibliography

Primary Literature

1. Agrawal H (2002) Extreme self-organization in networks constructed from gene expression data. Phys Rev Lett 89:268702–268702
2. Albert I, Albert R (2004) Conserved network motifs allow protein-protein interaction prediction. Bioinformatics 20(18):3346–3352
3. Albert R, Barabasi AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74(47)
4. Albert R, Jeong H et al (2000) Error and attack tolerance of complex networks. Nature 406:378–382
5. Alon U (2003) Biological networks: the tinkerer as an engineer. Science 301:1866–1867
6. Amaral LA, Scala A et al (2000) Classes of small-world networks. Proc Natl Acad Sci USA 97(21):11149–52
7. Asthana S, King OD et al (2004) Predicting protein complex membership using probabilistic network reliability. Genome Res 14(6):1170–5
8. Avery L, Wasserman S (1992) Ordering gene function: the interpretation of epistasis in regulatory hierarchies. Trends Genet 8(9):312–6
9. Bader GD, Hogue CW (2002) Analyzing yeast protein-protein interaction data obtained from different sources. Nat Biotechnol 20(10):991–7
10. Bader GD, Hogue CW (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4(1):2
11. Bader JS (2003) Greedily building protein networks with confidence. Bioinformatics 19(15):1869–74
12. Bader JS, Chaudhuri A et al (2004) Gaining confidence in high-throughput protein interaction networks. Nat Biotechnol 22(1):78–85
13. Balazsi G, Barabasi AL et al (2005) Topological units of environmental signal processing in the transcriptional regulatory network of Escherichia coli. Proc Natl Acad Sci USA 102(22):7841–6
14. Bar-Joseph Z (2003) Computational discovery of gene modules and regulatory networks. Nature Biotechnol 21:1337–1342
15. Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–12
16. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5(2):101–13
17. Bornholdt S, Ebel H (2001) World Wide Web scaling exponent from Simon's 1955 model. Phys Rev E 64(3):35104
18. Bornholdt S, Schuster HG (2003) Handbook of Graphs and Networks: from the Genome to the Internet
19. Bray D (2003) Molecular networks: the top-down view. Science 301:1864–1865
20. Broder A (2000) Graph structure in the web. Comput Netw 33:309–320
21. Callaway DS, Newman MEJ et al (2000) Network robustness and fragility: percolation on random graphs. Phys Rev Lett 85:5468–5471
22. Cho RJ, Campbell MJ et al (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. Molecular Cell 2(1):65–73
23. Cohen R, Erez K et al (2000) Resilience of the Internet to random breakdowns. Phys Rev Lett 85:4626–4628
24. de Lichtenberg U, Jensen LJ et al (2005) Dynamic complex formation during the yeast cell cycle. Science 307(5710):724–7
25. Diestel R (2005) Graph Theory, 3rd edn. Springer, Heidelberg
26. Dobrin R, Beg QK et al (2004) Aggregation of topological motifs in the Escherichia coli transcriptional regulatory network. BMC Bioinformatics 5(1):10

27. Dorogovtsev SN, Mendes JF (2003). Evolution of Networks: from Biological Nets to the Internet and WWW. Oxford University Press
28. Drees BL, Thorsson V et al (2005) Derivation of genetic interaction networks from quantitative phenotype data. Genome Biol 6(4):R38
29. Fanning AS, Anderson JM (1996) Protein-protein interactions: PDZ domain networks. Curr Biol 6(11):1385–8
30. Farh KK, Grimson A et al (2005) The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. Science 310(5755):1817–21
31. Farkas IJ, Wu C et al (2006) Topological basis of signal integration in the transcriptional-regulatory network of the yeast, Saccharomyces cerevisiae. BMC Bioinformatics 7:478
32. Featherstone DE, Broadie K (2002) Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network. Bioessays 24:267–274
33. Fell DA, Wagner A (2000) The small world of metabolism. Nat Biotechnol 18(11):1121–2
34. Freudenberg J, Zimmer R et al (2002) A hypergraph-based method for unification of existing protein structure- and sequence-families. In Silico Biol 2(3):339–49
35. Gavin AC, Bosche M et al (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415(6868):141–7
36. Ge H, Liu Z et al (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nat Genet 29(4):482–6
37. Geissler S, Siegers K et al (1998) A novel protein complex promoting formation of functional alpha- and gamma-tubulin. Embo J 17(4):952–66
38. Getoor L, Rhee JT et al (2004) Understanding tuberculosis epidemiology using structured statistical models. Artif Intell Med 30(3):233–56
39. Giaever G (2002) Functional profiling of the Saccharomyces cerevisiae genome. Nature 418:387–391
40. Gietz RD, B Triggs-Raine et al (1997) Identification of proteins that interact with a protein of interest: applications of the yeast two-hybrid system. Mol Cell Biochem 172(1–2):67–79
41. Giot L (2003) A protein interaction map of Drosophila melanogaster. Science 302:1727–1736
42. Girvan M, Newman ME (2002) J Community structure in social and biological networks. Proc Natl Acad Sci USA 99:7821–7826
43. Goldberg DS, Roth FP (2003) Assessing experimentally derived interactions in a small world. Proc Natl Acad Sci USA 3:3
44. Guelzim N, Bottani S et al (2002) Topological and causal structure of the yeast transcriptional regulatory network. Nat Genet 31(1):60–3
45. Gunsalus K C, Ge H et al (2005) Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis. Nature 436(7052):861–5
46. Han JD, Bertin N et al (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430(6995):88–93
47. Han JD, Dupuy D et al (2005) Effect of sampling on topology predictions of protein-protein interaction networks. Nat Biotechnol 23(7):839–44
48. Hanein D, Matlack KE et al (1996) Oligomeric rings of the Sec61p complex induced by ligands required for protein translocation. Cell 87(4):721–32
49. Harbison CT, Gordon DB et al (2004) Transcriptional regulatory code of a eukaryotic genome. Nature 431(7004):99–104
50. Hartwell LH, Hopfield JJ et al (1999) From molecular to modular cell biology. Nature 402:C47–C52
51. Hasty J, McMillen D et al (2002) Engineered gene circuits. Nature 420:224–230
52. Ho Y, Gruhler A et al (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415(6868):180–3
53. Holme P, Huss M et al (2003) Subnetwork hierarchies of biochemical pathways. Bioinformatics 19:532–538
54. Huh WK, Falvo JV et al (2003) Global analysis of protein localization in budding yeast. Nature 425(6959):686–91
55. Ihmels J (2002) Revealing modular organization in the yeast transcriptional network. Nat Genet 31:370–377
56. Ito T (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 98:4569–4574
57. Ito T, Tashiro K et al (2000) Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc Natl Acad Sci USA 97(3):1143–7
58. Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3:318–56
59. Jansen R (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 302:449–453
60. Jansen R, Greenbaum D et al (2002) Relating whole-genome expression data with protein-protein interactions. Genome Res 12(1): 37–46
61. Jansen R, Lan N et al (2002) Integration of genomic datasets to predict protein complexes in yeast. J Struct Funct Genomics 2:71–81
62. Jansen R, Yu H et al (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 302(5644):449–53
63. Jeong H, Mason SP et al (2001) Lethality and centrality in protein networks. Nature 411(6833):41–2
64. Jeong H, Tombor B et al (2000) The large-scale organization of metabolic networks. Nature 407:651–654
65. Juvan P, Demsar J et al (2005) GenePath: from mutations to genetic networks and back. Nucleic Acids Res 33(Web Server issue):W749–52
66. King OD (2004) Comment on Subgraphs in random networks. Phys Rev E Stat Nonlin Soft Matter Phys 70(5 Pt 2):058101. author reply 058102
67. Kitano H (2002) Computational systems biology. Nature 420:206–210
68. Koonin EV, Wolf YI et al (2002) The structure of the protein universe and genome evolution. Nature 420:218–223
69. Krogan NJ, Cagney G et al (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440(7084):637–43
70. Kumar A, Agarwal S et al (2002) Subcellular localization of the yeast proteome. Genes Dev 16(6):707–19
71. Launer RL, Wilkinson GN (1979) Robustness in statistics. Academic Press, New York
72. Lee I, Date SV et al (2004) A probabilistic functional network of yeast genes. Science 306(5701):1555–8
73. Lee TI, Rinaldi NJ et al (2002) Transcriptional regulatory net-

works in Saccharomyces cerevisiae. Science 298(5594):799–804
74. Li S (2004) A map of the interactome network of the metazoan, C elegans. Science 303:590–593
75. Li W, Liu Y et al (2007) Dynamical systems for discovering protein complexes and functional modules from biological networks. IEEE/ACM Trans Comput Biol Bioinform 4(2):233–50
76. Lockhart DJ, Dong H et al (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14(13):1675–80
77. Ma HW, Buer J et al (2004) Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach. BMC Bioinformatics 5:199
78. MacIsaac KD, Wang T et al (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. BMC Bioinformatics 7:113
79. Mangan S, Itzkovitz S et al (2006) The incoherent feed-forward loop accelerates the response-time of the gal system of Escherichia coli. J Mol Biol 356(5):1073–81
80. Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. Science 296:910–913
81. Ma'ayan A, Jenkins SL et al (2005) Formation of regulatory patterns during signal propagation in a Mammalian cellular network. Science 309(5737):1078–83
82. Milo R, Itzkovitz S et al (2004) Superfamilies of evolved and designed networks. Science 303(5663):1538–42
83. Milo R, S Shen-Orr et al (2002) Network motifs: simple building blocks of complex networks. Science 298(5594):824–7
84. Monod J, Cohen-Bazire G et al (1951) The biosynthesis of beta-galactosidase (lactase) in Escherichia coli; the specificity of induction. Biochim Biophys Acta 7(4):585–99
85. Monod J, Jacob F (1961) Teleonomic mechanisms in cellular metabolism, growth, and differentiation. Cold Spring Harb Symp Quant Biol 26:389–401
86. Nadvornik P, Drozen V (1964) Models of Neurons and Neuron Networks. Act Nerv Super (Praha) 6:293–302
87. Newman MEJ (2002) Assortative mixing in networks. Phys Rev Lett 89:208701–208701
88. Newman ME, Strogatz SH et al (2001) Random graphs with arbitrary degree distributions and their applications. Phys Rev E Stat Nonlin Soft Matter Phys 64(2 Pt 2):026118
89. Novick A, Weiner M (1957) Enzyme Induction as an All-or-None Phenomenon. Proc Natl Acad Sci USA 43(7):553–66
90. Oltvai ZN, Barabasi AL (2002) Life's complexity pyramid. Science 298:763–764
91. Pastor-Satorras R, Vazquez A et al (2001) Dynamical and correlation properties of the Internet. Phys. Rev. Lett. 87:258701–258701
92. Ptacek J, Devgan G et al (2005) Global analysis of protein phosphorylation in yeast. Nature 438(7068):679–84
93. Qi Y, Klein-Seetharaman J et al (2005) Random forest similarity for protein-protein interaction prediction from multiple sources. Pac Symp Biocomput:531–42
94. Rajewsky N (2006) microRNA target predictions in animals. Nat Genet 38 Suppl:S8–13
95. Ravasz E, Barabasi AL (2003) Hierarchical organization in complex networks. Phys Rev E Stat Nonlin Soft Matter Phys 67:026112–026112
96. Ravasz E, Somera AL et al (2002) Hierarchical organization of modularity in metabolic networks. Science 297(5586):1551–5
97. Rives AW, Galitski T (2003) Modular organization of cellular networks. Proc Natl Acad Sci USA 100(3):1128–33
98. Rouvray H (1990) The Origins of Chemical Graph Theory. In: Bonchev D, Rouvray DH (eds) Chemical Graph Theory: Introduction and Fundamentals, vol 41. Gordon and Breach Science Publishers, New York
99. Rual JF, Venkatesan K et al (2005) Towards a proteome-scale map of the human protein-protein interaction network. Nature 437(7062):1173–8
100. Schena M, Shalon D et al (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270(5235):467–470
101. Schleif R (2000) Regulation of the L-arabinose operon of Escherichia coli. Trends Genet 16(12):559–65
102. Schuster S, Pfeiffer T et al (2002) Exploring the pathway structure of metabolism: decomposition into subnetworks and application to Mycoplasma pneumoniae. Bioinformatics 18:351–361
103. Shen-Orr SS, Milo R et al (2002) Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet 31(1):64–8
104. Simon HA (1955) On a class of skew distribution functions. Biometrika 42:425–440
105. Simonis N, Gonze D et al (2006) Modularity of the transcriptional response of protein complexes in yeast. J Mol Biol 363(2):589–610
106. Simonis N, van Helden J et al (2004) Transcriptional regulation of protein complexes in yeast. Genome Biol 5(5):R33
107. Smith LM, Fung S et al (1985) The synthesis of oligonucleotides containing an aliphatic amino group at the 5′ terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. Nucleic Acids Res 13(7):2399–412
108. Smith LM, Sanders JZ et al (1986) Fluorescence detection in automated DNA sequence analysis. Nature 321(6071):674–9
109. Snel B, Bork P et al (2002) The identification of functional modules from the genomic association of genes. Proc Natl Acad Sci USA 99:5890–5895
110. Sole RV, R Pastor-Satorras et al (2002) A Model of Large-Scale Proteome Evolution. Adv Complex Syst 5:43–54
111. Sood P, Krek A et al (2006) Cell-type-specific signatures of microRNAs on target mRNA expression. Proc Natl Acad Sci USA 103(8):2746–51
112. Spellman PT, Sherlock G et al (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Cell Biol 9(12):3273–3297
113. Spirin V, Mirny LA (2003) Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci USA 100(21):12123–8
114. St Onge RP, Mani R et al (2007) Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. Nat Genet 39(2):199–206
115. Stark A, Brennecke J et al (2005) Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. Cell 123(6):1133–46
116. Stelzl U, Worm U et al (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122(6):957–68
117. Strogatz SH (2001) Exploring complex networks. Nature 410(6825):268–76
118. Stuart JM, Segal E et al (2003) A gene-coexpression network

for global discovery of conserved genetic modules. Science 302:249–255
119. Tanaka R (2005) Scale-rich metabolic networks. Phys Rev Lett 94(16):168101
120. Taylor RJ, Siegel AF et al (2007) Network motif analysis of a multi-mode genetic-interaction network. Genome Biol 8(8):R160
121. Thieffry D, Huerta AM et al (1998) From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. Bioessays 20(5):433–40
122. Tong AH, Lesage G et al (2004) Global mapping of the yeast genetic interaction network. Science 303(5659):808–13
123. Tornow S, Mewes HW (2003) Functional modules by relating protein interaction networks and gene expression. Nucleic Acids Res 31:6283–6289
124. Tsang J, Zhu J et al (2007) MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. Mol Cell 26(5):753–67
125. Uetz P, Giot L et al (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403(6770):623–7
126. Wagner A (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol Biol Evol 18(7):1283–92
127. Wagner A, Fell DA (2001) The small world inside large metabolic networks. Proc Biol Sci 268(1478):1803–10
128. Wall ME, Hlavacek WS et al (2004) Design of gene circuits: lessons from bacteria. Nature Rev Genet 5:34–42
129. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440–2
130. Wen X, Fuhrman S et al (1998) Large-scale temporal gene expression mapping of central nervous system development. Proc Natl Acad Sci USA 95(1):334–339
131. Winzeler EA (1999) Functional characterization of the S cerevisiae genome by gene deletion and parallel analysis. Science 285:901–906
132. Wong SL, Zhang LV et al (2004) Combining biological networks to predict genetic interactions. Proc Natl Acad Sci USA 101(44):15682–7
133. Wunderlich Z, Mirny LA (2006) Using the Topology of Metabolic Networks to Predict Viability of Mutant Strains. Biophys J 91(6):2304–2311
134. Yeger-Lotem E, Sattath S et al (2004) Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. Proc Natl Acad Sci USA 101(16):5934–9
135. Yook SH, Oltvai ZN et al (2004) Functional and topological characterization of protein interaction networks. Proteomics 4(4):928–42
136. Yu H, Gerstein M (2006) Genomic analysis of the hierarchical structure of regulatory networks. Proc Natl Acad Sci USA 103(40):14724–31
137. Yu H, Luscombe NM et al (2003) Genomic analysis of gene expression relationships in transcriptional regulatory networks. Trends Genet 19(8):422–7
138. Zhang L, King O et al (2005) Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network. J Biol 4(2):6
139. Zhang LV, Wong SL et al (2004) Predicting co-complexed protein pairs using genomic and proteomic data integration. BMC Bioinformatics 5(1):38

Books and Reviews

Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5(2):101–13
Diestel R (2005) Graph Theory, 3rd edn. Springer, Heidelberg

Bivariate (Two-dimensional) Wavelets

Bin Han
Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Canada

Article Outline

Glossary
Definitions
Introduction
Bivariate Refinable Functions and Their Properties
The Projection Method
Bivariate Orthonormal and Biorthogonal Wavelets
Bivariate Riesz Wavelets
Pairs of Dual Wavelet Frames
Future Directions
Bibliography

Glossary

Dilation matrix A 2×2 matrix M is called *a dilation matrix* if all the entries of M are integers and all the eigenvalues of M are greater than one in modulus.

Isotropic dilation matrix A dilation matrix M is said to be *isotropic* if M is similar to a diagonal matrix and all its eigenvalues have the same modulus.

Wavelet system A wavelet system is a collection of square integrable functions that are generated from a finite set of functions (which are called wavelets) by using integer shifts and dilations.

Definitions

Throughout this article, $\mathbb{R}, \mathbb{C}, \mathbb{Z}$ denote the real line, the complex plane, and the set of all integers, respectively. For $1 \leqslant p \leqslant \infty$, $L_p(\mathbb{R}^2)$ denotes the set of all Lebesgue measurable bivariate functions f such that $\|f\|^p_{L_p(\mathbb{R}^2)} := \int_{\mathbb{R}^2} |f(x)|^p \mathrm{d}x < \infty$. In particular, the space $L_2(\mathbb{R}^2)$ of square integrable functions is a Hilbert space under the inner product

$$\langle f, g\rangle := \int_{\mathbb{R}^2} f(x)\overline{g(x)}\mathrm{d}x\,, \quad f, g \in L_2(\mathbb{R}^2)\,,$$

where $\overline{g(x)}$ denotes the complex conjugate of the complex number $g(x)$.

In applications such as image processing and computer graphics, the following are commonly used isotropic dilation matrices:

$$M_{\sqrt{2}} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \qquad Q_{\sqrt{2}} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix},$$
$$M_{\sqrt{3}} = \begin{bmatrix} 2 & -1 \\ 1 & -2 \end{bmatrix}, \qquad dI_2 = \begin{bmatrix} d & 0 \\ 0 & d \end{bmatrix}, \tag{1}$$

where d is an integer with $|d| > 1$.

Using a dilation matrix M, a bivariate M-wavelet system is generated by integer shifts and dilates from a finite set $\{\psi^1, \ldots, \psi^L\}$ of functions in $L_2(\mathbb{R}^2)$. More precisely, the set of all the basic wavelet building blocks of an M-wavelet system generated by $\{\psi^1, \ldots, \psi^L\}$ is given by

$$X_M(\{\psi^1, \ldots, \psi^L\}) := \{\psi_{j,k}^{\ell,M} : j \in \mathbb{Z},\ k \in \mathbb{Z}^2, \ell = 1, \ldots, L\}, \tag{2}$$

where

$$\psi_{j,k}^{\ell,M}(x) := |\det M|^{j/2} \psi^\ell(M^j x - k), \quad x \in \mathbb{R}^2. \tag{3}$$

One of the main goals in wavelet analysis is to find wavelet systems $X_M(\{\psi^1, \ldots, \psi^L\})$ with some desirable properties such that any two-dimensional function or signal $f \in L_2(\mathbb{R}^2)$ can be sparsely and efficiently represented under the M-wavelet system $X_M(\{\psi^1, \ldots, \psi^L\})$:

$$f = \sum_{\ell=1}^{L} \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}^2} h_{j,k}^\ell(f) \psi_{j,k}^{\ell,M}, \tag{4}$$

where $h_{j,k}^\ell : L_2(\mathbb{R}^2) \mapsto \mathbb{C}$ are linear functionals. There are many types of wavelet systems studied in the literature. In the following, let us outline some of the most important types of wavelet systems.

Orthonormal Wavelets

We say that $\{\psi^1, \ldots, \psi^L\}$ generates *an orthonormal M-wavelet basis* in $L_2(\mathbb{R}^2)$ if the system $X_M(\{\psi^1, \ldots, \psi^L\})$ is an orthonormal basis of the Hilbert space $L_2(\mathbb{R}^2)$. That is, the linear span of elements in $X_M(\{\psi^1, \ldots, \psi^L\})$ is dense in $L_2(\mathbb{R}^2)$ and

$$\langle \psi_{j,k}^{\ell,M}, \psi_{j',k'}^{\ell',M} \rangle = \delta_{\ell-\ell'} \delta_{j-j'} \delta_{k-k'}, \quad \forall\, j, j' \in \mathbb{Z},\ k, k' \in \mathbb{Z}^2,\ \ell, \ell' = 1, \ldots, L, \tag{5}$$

where δ denotes the *Dirac sequence* such that $\delta_0 = 1$ and $\delta_k = 0$ for all $k \neq 0$. For an orthonormal wavelet basis $X_M(\{\psi^1, \ldots, \psi^L\})$, the linear functional $h_{j,k}^\ell$ in (4) is given by $h_{j,k}^\ell(f) = \langle f, \psi_{j,k}^{\ell,M} \rangle$ and the representation in (4) becomes

$$f = \sum_{\ell=1}^{L} \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}^2} \langle f, \psi_{j,k}^{\ell,M} \rangle \psi_{j,k}^{\ell,M}, \quad f \in L_2(\mathbb{R}^2) \tag{6}$$

with the series converging in $L_2(\mathbb{R}^2)$.

Riesz Wavelets

We say that $\{\psi^1, \ldots, \psi^L\}$ generates *a Riesz M-wavelet basis* in $L_2(\mathbb{R}^2)$ if the system $X_M(\{\psi^1, \ldots, \psi^L\})$ is a Riesz basis of $L_2(\mathbb{R}^2)$. That is, the linear span of elements in $X_M(\{\psi^1, \ldots, \psi^L\})$ is dense in $L_2(\mathbb{R}^2)$ and there exist two positive constants C_1 and C_2 such that

$$C_1 \sum_{\ell=1}^{L} \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}^2} |c_{j,k}^\ell|^2 \leqslant \left\| \sum_{\ell=1}^{L} \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}^2} c_{j,k}^\ell \psi_{j,k}^{\ell,M} \right\|_{L_2(\mathbb{R}^2)}^2 \leqslant C_2 \sum_{\ell=1}^{L} \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}^2} |c_{j,k}^\ell|^2$$

for all finitely supported sequences $\{c_{j,k}^\ell\}_{j\in\mathbb{Z},k\in\mathbb{Z}^2,\ell=1,\ldots,L}$. Clearly, a Riesz M-wavelet generalizes an orthonormal M-wavelet by relaxing the orthogonality requirement in (5). For a Riesz basis $X_M(\{\psi^1, \ldots, \psi^L\})$, it is well-known that there exists a dual Riesz basis $\{\tilde{\psi}^{\ell,j,k} : j \in \mathbb{Z}, k \in \mathbb{Z}^2, \ell = 1, \ldots, L\}$ of elements in $L_2(\mathbb{R}^2)$ (this set is not necessarily generated by integer shifts and dilates from some finite set of functions) such that (5) still holds after replacing $\psi_{j',k'}^{\ell',M}$ by $\tilde{\psi}^{\ell',j',k'}$. For a Riesz wavelet basis, the linear functional in (4) becomes $h_{j,k}^\ell(f) = \langle f, \tilde{\psi}^{\ell,j,k} \rangle$. In fact, $\tilde{\psi}^{\ell,j,k} := \mathcal{F}^{-1}(\psi_{j,k}^{\ell,M})$, where $\mathcal{F} : L_2(\mathbb{R}^2) \mapsto L_2(\mathbb{R}^2)$ is defined to be

$$\mathcal{F}(f) := \sum_{\ell=1}^{L} \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}^2} \langle f, \psi_{j,k}^{\ell,M} \rangle \psi_{j,k}^{\ell,M}, \quad f \in L_2(\mathbb{R}^2). \tag{7}$$

Wavelet Frames

A further generalization of a Riesz wavelet is a wavelet frame. We say that $\{\psi^1, \ldots, \psi^L\}$ generates *an M-wavelet frame* in $L_2(\mathbb{R}^2)$ if the system $X_M(\{\psi^1, \ldots, \psi^L\})$ is

a frame of $L_2(\mathbb{R}^2)$. That is, there exist two positive constants C_1 and C_2 such that

$$C_1\|f\|^2_{L_2(\mathbb{R}^2)} \leqslant \sum_{\ell=1}^{L}\sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}^2} |\langle f, \psi^{\ell,M}_{j,k}\rangle|^2 \leqslant C_2\|f\|^2_{L_2(\mathbb{R}^2)}, \quad \forall\, f \in L_2(\mathbb{R}^2). \tag{8}$$

It is not difficult to check that a Riesz M-wavelet is an M-wavelet frame. Then (8) guarantees that the *frame operator* $\mathcal{F}$ in (7) is a bounded and invertible linear operator. For an M-wavelet frame, the linear functional in (4) can be chosen to be $h^{\ell}_{j,k}(f) = \langle f, \mathcal{F}^{-1}(\psi^{\ell,M}_{j,k})\rangle$; however, such functionals may not be unique. The fundamental difference between a Riesz wavelet and a wavelet frame lies in that for any given function $f \in L_2(\mathbb{R}^2)$, the representation in (4) is unique under a Riesz wavelet while it may not be unique under a wavelet frame. The representation in (4) with the choice $h^{\ell}_{j,k}(f) = \langle f, \mathcal{F}^{-1}(\psi^{\ell,M}_{j,k})\rangle$ is called the *canonical representation* of a wavelet frame. In other words, $\{\mathcal{F}^{-1}(\psi^{\ell,M}_{j,k}) : j \in \mathbb{Z}, k \in \mathbb{Z}^2, \ell = 1,\ldots,L\}$ is called the *canonical dual frame* of the given wavelet frame $X_M(\{\psi^1,\ldots,\psi^L\})$.

Biorthogonal Wavelets

We say that $(\{\psi^1,\ldots,\psi^L\},\{\tilde\psi^1,\ldots,\tilde\psi^L\})$ generates *a pair of biorthogonal M-wavelet bases* in $L_2(\mathbb{R}^2)$ if each of $X_M(\{\psi^1,\ldots,\psi^L\})$ and $X_M(\{\tilde\psi^1,\ldots,\tilde\psi^L\})$ is a Riesz basis of $L_2(\mathbb{R}^2)$ and (5) still holds after replacing $\psi^{\ell',M}_{j',k'}$ by $\tilde\psi^{\ell',M}_{j',k'}$. In other words, the dual Riesz basis of $X_M(\{\psi^1,\ldots,\psi^L\})$ has the wavelet structure and is given by $X_M(\{\tilde\psi^1,\ldots,\tilde\psi^L\})$. For a biorthogonal wavelet, the wavelet representation in (4) becomes

$$f = \sum_{\ell=1}^{L}\sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}^2} \langle f, \tilde\psi^{\ell,M}_{j,k}\rangle \psi^{\ell,M}_{j,k}, \quad f \in L_2(\mathbb{R}^2). \tag{9}$$

Obviously, $\{\psi^1,\ldots,\psi^L\}$ generates an orthonormal M-wavelet basis in $L_2(\mathbb{R}^2)$ if and only if $(\{\psi^1,\ldots,\psi^L\}, \{\psi^1,\ldots,\psi^L\})$ generates a pair of biorthogonal M-wavelet bases in $L_2(\mathbb{R}^2)$.

Dual Wavelet Frames

Similarly, we have the notion of a pair of dual wavelet frames. We say that $(\{\psi^1,\ldots,\psi^L\},\{\tilde\psi^1,\ldots,\tilde\psi^L\})$ generates *a pair of dual M-wavelet frames* in $L_2(\mathbb{R}^2)$ if each of $X_M(\{\psi^1,\ldots,\psi^L\})$ and $X_M(\{\tilde\psi^1,\ldots,\tilde\psi^L\})$ is an M-wavelet frame in $L_2(\mathbb{R}^2)$ and

$$\langle f, g\rangle = \sum_{\ell=1}^{L}\sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}^2} \langle f, \tilde\psi^{\ell,M}_{j,k}\rangle\langle \psi^{\ell,M}_{j,k}, g\rangle, \quad f, g \in L_2(\mathbb{R}^2).$$

It follows from the above identity that (9) still holds for a pair of dual M-wavelet frames. We say that $\{\psi^1,\ldots,\psi^L\}$ generates *a tight M-wavelet frame* if $(\{\psi^1,\ldots,\psi^L\}, \{\psi^1,\ldots,\psi^L\})$ generates a pair of dual M-wavelet frames in $L_2(\mathbb{R}^2)$. For a tight wavelet frame $\{\psi^1,\ldots,\psi^L\}$, the wavelet representation in (6) still holds. So, a tight wavelet frame is a generalization of an orthonormal wavelet basis.

Introduction

Bivariate (two-dimensional) wavelets are of interest in representing and processing two dimensional data such as images and surfaces. In this article we shall discuss some basic background and results on bivariate wavelets.

We denote Π_J the set of all bivariate polynomials of total degree at most J. For compactly supported functions $\psi^1,\ldots,\psi^L$, we say that $\{\psi^1,\ldots,\psi^L\}$ has J *vanishing moments* if $\langle P, \psi^\ell\rangle = 0$ for all $\ell = 1,\ldots,L$ and all polynomials $P \in \Pi_{J-1}$.

The advantages of wavelet representations largely lie in the following aspects:

1. There is *a fast wavelet transform* (FWT) for computing the wavelet coefficients $h^{\ell}_{j,k}(f)$ in the wavelet representation (4).
2. The wavelet representation has *good time and frequency localization*. Roughly speaking, the basic building blocks $\psi^{\ell,M}_{j,k}$ have good time localization and smoothness.
3. The wavelet representation is sparse. For a smooth function f, most wavelet coefficients are negligible. Generally, the wavelet coefficient $h^{\ell}_{j,k}(f)$ only depends on the information of f in a small neighborhood of the support of $\psi^{\ell,M}_{j,k}$ (more precisely, the support of $\tilde\psi^{\ell,M}_{j,k}$ if $h^{\ell}_{j,k}(f) = \langle f, \tilde\psi^{\ell,M}_{j,k}\rangle$). If f in a small neighborhood of the support of $\psi^{\ell,M}_{j,k}$ is smooth or behaves like a polynomial to certain degree, then by the vanishing moments of the dual wavelet functions, $h^{\ell}_{j,k}(f)$ is often negligible.
4. For a variety of function spaces B such as Sobolev and Besov spaces, its norm $\|f\|_B$, $f \in B$, is equivalent to certain weighted sequence norm of its wavelet coefficients $\{h^{\ell}_{j,k}(f)\}_{j\in\mathbb{Z},k\in\mathbb{Z}^2,\ell=1,\ldots,L}$.

For more details on advantages and applications of wavelets, see [6,8,14,16,21,58].

For a dilation matrix $M = dI_2$, the easiest way for obtaining a bivariate wavelet is to use the tensor product method. In other words, all the functions in the generator set $\{\psi^1, \dots, \psi^L\} \subset L_2(\mathbb{R}^2)$ take the form $\psi^\ell(x, y) = f^\ell(x)g^\ell(y)$, $x, y \in \mathbb{R}$, where f^ℓ and g^ℓ are some univariate functions in $L_2(\mathbb{R})$. However, tensor product (also called *separable*) bivariate wavelets give preference to the horizontal and vertical directions which may not be desirable in applications such as image processing ([9,16,50,58]). Also, for a non-diagonal dilation matrix, it is difficult or impossible to use the tensor product method to obtain bivariate wavelets. Therefore, nonseparable bivariate wavelets themselves are of importance and interest in both theory and application. In this article, we shall address several aspects of bivariate wavelets with a general two-dimensional dilation matrix.

Bivariate Refinable Functions and Their Properties

In order to have a fast wavelet transform to compute the wavelet coefficients in (4), the generators $\psi^1, \dots, \psi^L$ in a wavelet system are generally obtained from a refinable function ϕ via a multiresolution analysis [6,15,16,58].

Let M be a 2×2 dilation matrix. For a function ϕ in $L_2(\mathbb{R}^2)$, we say that ϕ is *M-refinable* if it satisfies the following refinement equation

$$\phi(x) = |\det M| \sum_{k \in \mathbb{Z}^2} a_k \phi(Mx - k), \quad \text{a.e. } x \in \mathbb{R}^2, \quad (10)$$

where $a : \mathbb{Z}^2 \mapsto \mathbb{C}$ is a finitely supported sequence on $\mathbb{Z}^2$ satisfying $\sum_{k \in \mathbb{Z}^2} a_k = 1$. Such a sequence a is often called *a mask* in wavelet analysis and computer graphics, or *a low-pass filter* in signal processing. Using the Fourier transform $\hat{\phi}$ of $\phi \in L_2(\mathbb{R}^2) \cap L_1(\mathbb{R}^2)$ and the Fourier series $\hat{a}$ of a, which are defined to be

$$\hat{\phi}(\xi) := \int_{\mathbb{R}^2} \phi(x) e^{-ix\cdot\xi}\, \mathrm{d}x \quad \text{and}$$
$$\hat{a}(\xi) := \sum_{k \in \mathbb{Z}^2} a_k e^{-ik\cdot\xi}, \quad \xi \in \mathbb{R}^2, \quad (11)$$

the refinement equation (10) can be equivalently rewritten as

$$\hat{\phi}(M^{\mathrm{T}}\xi) = \hat{a}(\xi)\hat{\phi}(\xi), \quad \xi \in \mathbb{R}^2, \quad (12)$$

where M^T denotes the transpose of the dilation matrix M. Since $\hat{a}(0) = 1$ and $\hat{a}$ is a trigonometric polynomial, one can define a function $\hat{\phi}$ by

$$\hat{\phi}(\xi) := \prod_{j=1}^{\infty} \hat{a}((M^{\mathrm{T}})^{-j}\xi), \quad \xi \in \mathbb{R}^2, \quad (13)$$

with the series converging uniformly on any compact set of $\mathbb{R}^2$. It is known ([4,16]) that ϕ is a compactly supported tempered distribution and clearly ϕ satisfies the refinement equation in (12) with mask a. We call ϕ the *standard refinable function* associated with mask a and dilation M. Now the generators $\psi^1, \dots, \psi^L$ of a wavelet system are often obtained from the refinable function ϕ via

$$\widehat{\psi^\ell}(M^{\mathrm{T}}\xi) := \widehat{a^\ell}(\xi)\hat{\phi}(\xi), \quad \xi \in \mathbb{R}^2, \quad \ell = 1, \dots, L,$$

for some 2π-periodic trigonometric polynomials $\widehat{a^\ell}$. Such sequences a^ℓ are called *wavelet masks* or *high-pass filters*.

Except for very few special masks a such as masks for box spline refinable functions (see [20] for box splines), most refinable functions ϕ, obtained in (13) from mask a and dilation M, do not have explicit analytic expressions and a so-called *cascade algorithm* or *subdivision scheme* is used to approximate the refinable function ϕ. Let B be a Banach space of bivariate functions. Starting with a suitable initial function $f \in B$, we iteratively compute a sequence $\{Q_{a,M}^n f\}_{n=0}^{\infty}$ of functions, where

$$Q_{a,M} f := |\det M| \sum_{k \in \mathbb{Z}^2} a_k f(M \cdot - k). \quad (14)$$

If the sequence $\{Q_{a,M}^n f\}_{n=0}^{\infty}$ of functions converges to some $f_\infty \in B$ in the Banach space B, then we have $Q_{a,M} f_\infty = f_\infty$. That is, as a fixed point of the *cascade operator* $Q_{a,M}$, f_∞ is a solution to the refinement equation (10). A cascade algorithm plays an important role in the study of refinable functions in wavelet analysis and of subdivision surfaces in computer graphics. For more details on cascade algorithms and subdivision schemes, see [4,19,21,22,23,31,36,54] and numerous references therein.

One of the most important properties of a refinable function ϕ is its smoothness, which is measured by its L_p smoothness exponent $\nu_p(\phi)$ and is defined to be

$$\nu_p(\phi) := \sup\{\nu : \phi \in W_p^\nu(\mathbb{R}^2)\}, \quad 1 \leqslant p \leqslant \infty, \quad (15)$$

where $W_p^\nu(\mathbb{R}^2)$ denotes the fractional Sobolev space of order $\nu > 0$ in $L_p(\mathbb{R}^2)$.

The notion of sum rules of a mask is closely related to the vanishing moments of a wavelet system ([16,44]). For a bivariate mask a, we say that a satisfies the *sum rules* of order J with the dilation matrix M if

$$\sum_{k \in \mathbb{Z}^2} a_{j+Mk} P(j+Mk) = \sum_{k \in \mathbb{Z}^2} a_{Mk} P(Mk), \quad \forall P \in \Pi_{J-1}. \quad (16)$$

Or equivalently, $\partial_1^{\mu_1}\partial_2^{\mu_2}\hat{a}(2\pi\gamma) = 0$ for all $\mu_1 + \mu_2 < J$ and $\gamma \in [(M^{\mathrm{T}})^{-1}\mathbb{Z}^2]\backslash\mathbb{Z}^2$, where ∂_1 and ∂_2 denote the partial derivatives in the first and second coordinate, respectively. Throughout the article, we denote $sr(a,M)$ the highest order of sum rules satisfied by the mask a with the dilation matrix M.

To investigate various properties of refinable functions, next we introduce a quantity $\nu_p(a, M)$ in wavelet analysis (see [31]). For a bivariate mask a and a dilation matrix M, we denote

$$\nu_p(a, M) := -\log_{\rho(M)}[|\det M|^{1-1/p}\rho(a, M, p)]\,, \qquad 1 \leqslant p \leqslant \infty\,, \quad (17)$$

where $\rho(M)$ denotes the *spectral radius* of M and

$$\rho(a, M, p) := \max\left\{\limsup_{n\to\infty} \|a^{n,(\mu_1,\mu_2)}\|_{\ell_p(\mathbb{Z}^2)}^{1/n} : \mu_1 + \mu_2 = sr(a, M), \mu_1, \mu_2 \in \mathbb{N} \cup \{0\}\right\},$$

where the sequence $a^{n,(\mu_1,\mu_2)}$ is defined by its Fourier series as follows: for $\xi = (\xi_1, \xi_2) \in \mathbb{R}^2$,

$$\widehat{a^{n,(\mu_1,\mu_2)}}(\xi) := (1 - \mathrm{e}^{-i\xi_1})^{\mu_1}(1 - e^{-i\xi_2})^{\mu_2} \hat{a}((M^{\mathrm{T}})^{n-1}\xi)\hat{a}((M^{\mathrm{T}})^{n-2}\xi)\cdots\hat{a}(M^{\mathrm{T}}\xi)\hat{a}(\xi)\,.$$

It is known that ([31,36]) $\nu_p(a, M) \geqslant \nu_q(a, M) \geqslant \nu_p(a, M) + (1/q - 1/p)\log_{\rho(M)}|\det M|$ for all $1 \leqslant p \leqslant q \leqslant \infty$.

Parts of the following result essentially appeared in various forms in many papers in the literature. For the study of refinable functions and cascade algorithms, see [4,10,12,15,16,22,23,25,29,31,45,48,53] and references therein. The following is from [31].

Theorem 1 *Let M be a 2×2 isotropic dilation matrix and a be a finitely supported bivariate mask. Let ϕ denote the standard refinable function associated with mask a and dilation M. Then $\nu_p(\phi) \geqslant \nu_p(a, M)$ for all $1 \leqslant p \leqslant \infty$. If the shifts of ϕ are stable, that is, $\{\phi(\cdot - k) : k \in \mathbb{Z}^2\}$ is a Riesz system in $L_p(\mathbb{R}^2)$, then $\nu_p(\phi) = \nu_p(a, M)$. Moreover, for every nonnegative integer J, the following statements are equivalent.*

1. *For every compactly supported function $f \in W_p^J(\mathbb{R}^2)$ (if $p = \infty$, we require $f \in C^J(\mathbb{R}^2)$) such that $\hat{f}(0) = 1$ and $\partial_1^{\mu_1}\partial_2^{\mu_2}\hat{f}(2\pi k) = 0$ for all $\mu_1 + \mu_2 < J$ and $k \in \mathbb{Z}^2\backslash\{0\}$, the cascade sequence $\{Q_{a,M}^n f\}_{n=0}^{\infty}$ converges in the Sobolev space $W_p^J(\mathbb{R}^2)$ (in fact, the limit function is ϕ).*
2. *$\nu_p(a, M) > J$.*

Symmetry of a refinable function is also another important property of a wavelet system in applications. For example, symmetry is desirable in order to handle the boundary of an image or to improve the visual quality of a surface ([16,21,22,32,50,58]).

Let M be a 2×2 dilation matrix and G be a finite set of 2×2 integer matrices. We say that G is *a symmetry group with respect to M* ([27,29]) if G is a group under matrix multiplication and $MEM^{-1} \in G$ for all $E \in G$. In dimension two, two commonly used symmetry groups are D_4 and D_6:

$$D_4 := \left\{\pm\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}, \pm\begin{bmatrix}1 & 0\\0 & -1\end{bmatrix}, \pm\begin{bmatrix}0 & 1\\1 & 0\end{bmatrix}, \pm\begin{bmatrix}0 & 1\\-1 & 0\end{bmatrix}\right\}$$

and

$$D_6 := \left\{\pm\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}, \pm\begin{bmatrix}0 & -1\\1 & -1\end{bmatrix}, \pm\begin{bmatrix}-1 & 1\\-1 & 0\end{bmatrix}, \pm\begin{bmatrix}0 & 1\\1 & 0\end{bmatrix}, \pm\begin{bmatrix}1 & -1\\0 & -1\end{bmatrix}, \pm\begin{bmatrix}-1 & 0\\-1 & 1\end{bmatrix}\right\}.$$

For symmetric refinable functions, we have the following result ([32, Proposition 1]):

Proposition 2 *Let M be a 2×2 dilation matrix and G be a symmetry group with respect to M. Let a be a bivariate mask and ϕ be the standard refinable function associated with mask a and dilation M. Then a is G-symmetric with center c_a: $a_{E(k-c_a)+c_a} = a_k$ for all $k \in \mathbb{Z}^2$ and $E \in G$, if and only if, ϕ is G-symmetric with center c:*

$$\phi(E(\cdot - c) + c) = \phi, \quad \forall\, E \in G \quad \textit{with}\ \ c := (M - I_2)^{-1}c_a\,.$$

In the following, let us present some examples of bivariate refinable functions.

Example 3 Let $M = 2I_2$ and $\hat{a}(\xi_1, \xi_2) := \cos^4(\xi_1/2)\cos^4(\xi_2/2)$. Then a is a tensor product mask with $sr(a, 2I_2) = 4$ and a is D_4-symmetric with center 0. The associated standard refinable function ϕ is the tensor product spline of order 4. The Catmull-Clark subdivision scheme for quadrilateral meshes in computer graphics is based on this mask a ([21]).

Example 4 Let $M = 2I_2$ and $\hat{a}(\xi_1, \xi_2) := \cos^2(\xi_1/2)\cos^2(\xi_2/2)\cos^2(\xi_1/2 + \xi_2/2)$. Then $sr(a, 2I_2) = 4$ and a is D_6-symmetric with center 0. The associated refinable function ϕ is the convolution of the three direction box spline with itself ([20]). The Loop subdivision scheme for triangular meshes in computer graphics is based on this mask a ([21]).

Another important property of a refinable function is interpolation. We say that a bivariate function ϕ is *interpolating* if ϕ is continuous and $\phi(k) = \delta_k$ for all $k \in \mathbb{Z}^2$.

Theorem 5 *Let M be a 2×2 dilation matrix and a be a finitely supported bivariate mask. Let ϕ denote the standard refinable function associated with mask a and dilation M. Then ϕ is an interpolating function if and only if $\nu_\infty(a, M) > 0$ and a is an interpolatory mask with dilation M, that is, $a_0 = |\det M|^{-1}$ and $a_{Mk} = 0$ for all $k \in \mathbb{Z}^2 \backslash \{0\}$.*

Example 6 Let $M = 2I_2$ and

$$\begin{aligned}\hat{a}(\xi_1, \xi_2) :=\ & \cos(\xi_1/2)\cos(\xi_2/2)\cos(\xi_1/2 + \xi_2/2) \\ & \big[1 + 2\cos(\xi_1) + 2\cos(\xi_2) + 2\cos(\xi_1 + \xi_2) \\ & - \cos(2\xi_1 + \xi_2) - \cos(\xi_1 + 2\xi_2) \\ & - \cos(\xi_1 - \xi_2)\big]/4\,. \end{aligned} \tag{18}$$

Then $sr(a, 2I_2) = 4$, a is D_6-symmetric with center 0 and a is an interpolatory mask with dilation $2I_2$. Since $\nu_2(a, 2I_2) \approx 2.44077$, so, $\nu_\infty(a, 2I_2) \geqslant \nu_2(a, 2I_2) - 1 \approx 1.44077 > 0$. By Theorem 5, the associated standard refinable function ϕ is interpolating. The butterfly interpolatory subdivision scheme for triangular meshes is based on this mask a (see [23]).

For more details on bivariate interpolatory masks and interpolating refinable functions, see [4,19,22,23,25,26,27,31,37,38,39,59].

The Projection Method

In applications, one is interested in analyzing some optimal properties of multivariate wavelets. The projection method is useful for this purpose.

Let r and s be two positive integers with $r \leqslant s$. Let P be an $r \times s$ real-valued matrix. For a compactly supported function ϕ and a finitely supported mask a in dimension s, we define the *projected function* $P\phi$ and *projected mask* Pa in dimension r by

$$\widehat{P\phi}(\xi) := \hat{\phi}(P^{\mathrm{T}}\xi) \quad \text{and}$$
$$\widehat{Pa}(\xi) := \sum_{j=1}^{t} \hat{a}(P^{\mathrm{T}}\xi + 2\pi\varepsilon_j)\,, \quad \xi \in \mathbb{R}^r\,, \tag{19}$$

where $\hat{\phi}$ and $\hat{a}$ are understood to be continuous and $\{\varepsilon_1, \ldots, \varepsilon_t\}$ is a complete set of representatives of the distinct cosets of $[P^{\mathrm{T}}\mathbb{R}^r]/\mathbb{Z}^s$. If P is an integer matrix, then $\widehat{Pa}(\xi) = \hat{a}(P^{\mathrm{T}}\xi)$.

Now we have the following result on projected refinable functions ([27,28,34,35]).

Theorem 7 *Let N be an $s \times s$ dilation matrix and M be an $r \times r$ dilation matrix. Let P be an $r \times s$ integer matrix such that $PN = MP$ and $P\mathbb{Z}^s = \mathbb{Z}^r$. Let $\hat{a}$ be a 2π-periodic trigonometric polynomials in s-variables with $\hat{a}(0) = 1$ and ϕ the standard N-refinable function associated with mask a. Then $sr(a, N) \leqslant sr(Pa, M)$ and $\widehat{P\phi}(M^{\mathrm{T}}\xi) = \widehat{Pa}(\xi)\widehat{P\phi}(\xi)$. That is, $P\phi$ is M-refinable with mask Pa. Moreover, for all $1 \leqslant p \leqslant \infty$,*

$$\nu_p(\phi) \leqslant \nu_p(P\phi) \quad \textit{and}$$
$$|\det M|^{1-1/p}\rho(Pa, M, p) \leqslant |\det N|^{1-1/p}\rho(a, N, p)\,.$$

If we further assume that $\rho(M) = \rho(N)$, then $\nu_p(a, N) \leqslant \nu_p(Pa, M)$.

As pointed out in [34,35], the projection method is closely related to box splines. For a given $r \times s$ (direction) integer matrix Ξ of rank r with $r \leqslant s$, the Fourier transform of its associated box spline M_Ξ and its mask a_Ξ are given by (see [20])

$$\widehat{M_\Xi}(\xi) := \prod_{k \in \Xi} \frac{1 - e^{-ik\cdot\xi}}{ik\cdot\xi} \quad \text{and}$$
$$\widehat{a_\Xi}(\xi) := \prod_{k \in \Xi} \frac{1 + e^{-ik\cdot\xi}}{2}\,, \quad \xi \in \mathbb{R}^r\,, \tag{20}$$

where $k \in \Xi$ means that k is a column vector of Ξ and k goes through all the columns of Ξ once and only once. Let $\chi_{[0,1]^s}$ denote the characteristic function of the unit cube $[0, 1]^s$. From (20), it is evident that the box spline M_Ξ is just the projected function $\Xi\chi_{[0,1]^s}$, since $\Xi\chi_{[0,1]^s} = M_\Xi$ by $\widehat{\Xi\chi_{[0,1]^s}}(\xi) = \widehat{\chi_{[0,1]^s}}(\Xi^{\mathrm{T}}\xi), \xi \in \mathbb{R}^r$. Note that M_Ξ is $2I_r$-refinable with the mask a_Ξ, since $\widehat{M_\Xi}(2\xi) = \widehat{a_\Xi}(\xi)\widehat{M_\Xi}(\xi)$.

As an application of the projection method in Theorem 7, we have ([25, Theorem 3.5.]):

Corollary 8 *Let $M = 2I_s$ and a be an interpolatory mask with dilation M such that a is supported inside $[-3, 3]^s$. Then $\nu_\infty(a, 2I_s) \leqslant 2$ and therefore, $\phi \notin C^2(\mathbb{R}^s)$, where ϕ is the standard refinable function associated with mask a and dilation $2I_s$.*

We use proof by contradiction. Suppose $\nu_\infty(a, 2I_s) > 2$. Let $P = [1, 0, \ldots, 0]$ be a $1 \times s$ matrix. Then we must have $sr(a, 2I_s) \geqslant 3$, which, combining with other assumptions on a, will force $\widehat{Pa}(\xi) = 1/2 + 9/16\cos(\xi) - 1/16\cos(3\xi)$. Since $\nu_\infty(Pa, 2) = 2$, by Theorem 7, we must have $\nu_\infty(a, 2I_s) \leqslant \nu_\infty(Pa, 2) = 2$. A contradiction. So, $\nu_\infty(a,$

$2I_s) \leqslant 2$. In particular, the refinable function in the butterfly subdivision scheme in Example 6 is not C^2.

The projection method can be used to construct interpolatory masks painlessly [27]).

Theorem 9 *Let M be an $r \times r$ dilation matrix. Then there is an $r \times r$ integer matrix H such that $M\mathbb{Z}^r = H\mathbb{Z}^r$ and $H^r = |\det M| I_r$. Let $P := |\det M|^{-1} H$. Then for any (tensor product) interpolatory mask a with dilation $|\det M| I_r$, the projected mask Pa is an interpolatory mask with dilation M and $sr(Pa, M) \geqslant sr(a, |\det M| I_r)$.*

Example 10 For $M = M_{\sqrt{2}}$ or $M = Q_{\sqrt{2}}$, we can take $H := M_{\sqrt{2}}$ in Theorem 9.

For more details on the projection method, see [25,27,28,32,34,35].

Bivariate Orthonormal and Biorthogonal Wavelets

In this section, we shall discuss the analysis and construction of bivariate orthonormal and biorthogonal wavelets.

For analysis of biorthogonal wavelets, the following result is well-known (see [9,11,16,24,26,31,46,51,53] and references therein):

Theorem 11 *Let M be a 2×2 dilation matrix and $a, \tilde{a}$ be two finitely supported bivariate masks. Let ϕ and $\tilde{\phi}$ be the standard M-refinable functions associated with masks a and $\tilde{a}$, respectively. Then $\phi, \tilde{\phi} \in L_2(\mathbb{R}^2)$ and satisfy the biorthogonality relation*

$$\langle \phi, \tilde{\phi}(\cdot - k) \rangle = \delta_k \,, \quad k \in \mathbb{Z}^2 \,, \tag{21}$$

if and only if, $\nu_2(a, M) > 0$, $\nu_2(\tilde{a}, M) > 0$, and $(a, \tilde{a})$ is a pair of dual masks:

$$\sum_{\gamma \in \Gamma_{M^{\mathrm{T}}}} \overline{\hat{a}(\xi + 2\pi\gamma)} \hat{\tilde{a}}(\xi + 2\pi\gamma) = 1 \,, \quad \xi \in \mathbb{R}^2 \,, \tag{22}$$

where $\Gamma_{M^{\mathrm{T}}}$ is a complete set of representatives of distinct cosets of $[(M^{\mathrm{T}})^{-1}\mathbb{Z}^2]/\mathbb{Z}^2$ with $0 \in \Gamma_{M^{\mathrm{T}}}$. Moreover, if (21) holds and there exist 2π-periodic trigonometric polynomials $\widehat{a^1}, \ldots, \widehat{a^{m-1}}$, $\widehat{\tilde{a}^1}, \ldots, \widehat{\tilde{a}^{m-1}}$ with $m := |\det M|$ such that

$$\mathcal{M}_{[\hat{a}, \widehat{a^1}, \ldots, \widehat{a^{m-1}}]}(\xi) \overline{\mathcal{M}_{[\hat{\tilde{a}}, \widehat{\tilde{a}^1}, \ldots, \widehat{\tilde{a}^{m-1}}]}(\xi)}^{\mathrm{T}} = I_m \,, \quad \xi \in \mathbb{R}^2 \,, \tag{23}$$

where for $\xi \in \mathbb{R}^2$,

$$\mathcal{M}_{[\hat{a}, \widehat{a^1}, \ldots, \widehat{a^{m-1}}]}(\xi) := \begin{bmatrix} \hat{a}(\xi + 2\pi\gamma_0) & \widehat{a^1}(\xi + 2\pi\gamma_0) & \ldots & \widehat{a^{m-1}}(\xi + 2\pi\gamma_0) \\ \hat{a}(\xi + 2\pi\gamma_1) & \widehat{a^1}(\xi + 2\pi\gamma_1) & \ldots & \widehat{a^{m-1}}(\xi + 2\pi\gamma_1) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{a}(\xi + 2\pi\gamma_{m-1}) & \widehat{a^1}(\xi + 2\pi\gamma_{m-1}) & \ldots & \widehat{a^{m-1}}(\xi + 2\pi\gamma_{m-1}) \end{bmatrix} \tag{24}$$

with $\{\gamma_0, \ldots, \gamma_{m-1}\} := \Gamma_{M^{\mathrm{T}}}$ and $\gamma_0 := 0$. Define $\psi^1, \ldots, \psi^{m-1}, \tilde{\psi}^1, \ldots, \tilde{\psi}^{m-1}$ by

$$\widehat{\psi^\ell}(M^{\mathrm{T}}\xi) := \widehat{a^\ell}(\xi)\hat{\phi}(\xi) \quad \textit{and}$$
$$\widehat{\tilde{\psi}^\ell}(M^{\mathrm{T}}\xi) := \widehat{\tilde{a}^\ell}(\xi)\hat{\tilde{\phi}}(\xi) \,, \quad \ell = 1, \ldots, m-1 \,. \tag{25}$$

Then $(\{\psi^1, \ldots, \psi^{m-1}\}, \{\tilde{\psi}^1, \ldots, \tilde{\psi}^{m-1}\})$ generates a pair of biorthogonal M-wavelet bases in $L_2(\mathbb{R}^2)$.

As a direct consequence of Theorem 11, for orthonormal wavelets, we have

Corollary 12 *Let M be a 2×2 dilation matrix and a be a finitely supported bivariate mask. Let ϕ be the standard refinable function associated with mask a and dilation M. Then ϕ has orthonormal shifts, that is, $\langle \phi, \phi(\cdot - k) \rangle = \delta_k$ for all $k \in \mathbb{Z}^2$, if and only if, $\nu_2(a, M) > 0$ and a is an orthogonal mask (that is, (22) is satisfied with $\hat{\tilde{a}}$ being replaced by $\hat{a}$). If in addition there exist 2π-periodic trigonometric polynomials $\widehat{a^1}, \ldots, \widehat{a^{m-1}}$ with $m := |\det M|$ such that $\mathcal{M}_{[\hat{a}, \widehat{a^1}, \ldots, \widehat{a^{m-1}}]}(\xi) \overline{\mathcal{M}_{[\hat{a}, \widehat{a^1}, \ldots, \widehat{a^{m-1}}]}(\xi)}^{\mathrm{T}} = I_m$. Then $\{\psi^1, \ldots, \psi^{m-1}\}$, defined in (25), generates an orthonormal M-wavelet basis in $L_2(\mathbb{R}^2)$.*

The masks a and $\tilde{a}$ are called *low-pass filters* and the masks $a^1, \ldots, a^{m-1}, \tilde{a}^1, \ldots, \tilde{a}^{m-1}$ are called *high-pass filters* in the engineering literature. The equation in (23) is called the *matrix extension problem* in the wavelet literature (see [6,16,46,58]). Given a pair of dual masks a and $\tilde{a}$, though the existence of $\widehat{a^1}, \ldots, \widehat{a^{m-1}}, \widehat{\tilde{a}^1}, \ldots, \widehat{\tilde{a}^{m-1}}$ (without symmetry) in the matrix extension problem in (23) is guaranteed by the Quillen–Suslin theorem, it is far from a trivial task to construct them (in particular, with symmetry) algorithmically. For the orthogonal case, the general theory of the matrix extension problem in (23) with $\tilde{a} = a$ and $\tilde{a}^\ell = a^\ell, \ell = 1, \ldots, m-1$, remains unanswered. However, when $|\det M| = 2$, the matrix extension problem is trivial. In fact, letting $\gamma \in \Gamma_{M^{\mathrm{T}}} \backslash \{0\}$ (that is, $\Gamma_{M^{\mathrm{T}}} = \{0, \gamma\}$, one defines $\widehat{a^1}(\xi) := e^{-i\eta\cdot\xi}\overline{\hat{\tilde{a}}(\xi + 2\pi\gamma)}$ and $\widehat{\tilde{a}^1}(\xi) := e^{-i\eta\cdot\xi}\overline{\hat{a}(\xi + 2\pi\gamma)}$, where $\eta \in \mathbb{Z}^2$ satisfies

$\eta \cdot \gamma = 1/2$. It is not an easy task to construct nonseparable orthogonal masks with desirable properties for a general dilation matrix. For example, for the dilation matrix $Q_{\sqrt{2}}$ in (1), it is not known so far in the literature [9] whether there is a compactly supported C^1 $Q_{\sqrt{2}}$-refinable function with orthonormal shifts. See [1,2,9, 25,27,30,32,50,51,53,57] for more details on construction of orthogonal masks and orthonormal refinable functions.

However, for any dilation matrix, "separable" orthogonal masks with arbitrarily high orders of sum rules can be easily obtained via the projection method. Let M be a 2×2 dilation matrix. Then $M = E\,\mathrm{diag}(d_1, d_2)F$ for some integer matrices E and F such that $|\det E| = |\det F| = 1$ and $d_1, d_2 \in \mathbb{N}$. Let a be a tensor product orthogonal mask with the diagonal dilation matrix $\mathrm{diag}(d_1, d_2)$ satisfying the sum rules of any preassigned order J. Then the projected mask Ea (that is, $(Ea)_k := a_{E^{-1}k}$, $k \in \mathbb{Z}^2$) is an orthogonal mask with dilation M and $sr(Ea, M) \geqslant J$. See [27, Corollary 3.4.] for more detail.

Using the high-pass filters in the matrix extension problem, for a given mask a, dual masks $\tilde{a}$ (without any guaranteed order of sum rules of the dual masks) of a can be obtained by the lifting scheme in [63] and the stable completion method in [3]. Special families of dual masks with sum rules can be obtained by the convolution method in [24, Proposition 3.7.] and [43], but the constructed dual masks generally have longer supports with respect to their orders of sum rules. Another method, which we shall cite here, for constructing all finitely supported dual masks with any preassigned orders of sum rules is the CBC (coset by coset) algorithm proposed in [5,25,26].

For $\mu = (\mu_1, \mu_2)$ and $\nu = (_1, \nu_2)$, we say $\nu \leqslant \mu$ if $\nu_1 \leqslant \mu_1$ and $\nu_2 \leqslant \mu_2$. Also we denote $|\mu| = \mu_1 + \mu_2$, $\mu! := \mu_1!\mu_2!$ and $x^\mu := x_1^{\mu_1} x_2^{\mu_2}$ for $x = (x_1, x_2)$. We denote Ω_M a complete set of representatives of distinct cosets of $\mathbb{Z}^2/[M\mathbb{Z}^2]$ with $0 \in \Omega_M$. The following result is from [25].

Theorem 13 (CBC Algorithm) *Let M be a 2×2 dilation matrix and a be an interpolatory mask with dilation M. Let J be any preassigned positive integer.*

1. *Compute the quantities h^a_μ, $|\mu| < J$, from the mask a by the recursive formula:*

$$h^a_\mu := \delta_\mu - \sum_{0 \leqslant \nu < \mu} (-1)^{|\mu-\nu|} \frac{\mu!}{\nu!(\mu-\nu)!} h^a_\nu \sum_{k \in \mathbb{Z}^2} \overline{a_k} k^{\mu-\nu} .$$

2. *For every nonzero coset $\gamma \in \Omega_M \backslash \{0\}$ and pre-selected subsets $E_\gamma \subset \mathbb{Z}^2$, construct $\tilde{a}$ on the coset $\gamma + M\mathbb{Z}^2$, with $\tilde{a}_{\gamma+Mk} = 0$ for $k \in \mathbb{Z}^2 \backslash E_\gamma$, satisfying the equation:*

$$\sum_{k \in E_\gamma} \tilde{a}_{\gamma+Mk} (\gamma + Mk)^\mu = |\det M|^{-1} h^a_\mu , \quad |\mu| < J .$$

3. *Construct $\tilde{a}$ at the zero coset $M\mathbb{Z}^2$ by*

$$\tilde{a}_{Mj} := |\det M|^{-1} \delta_k - \sum_{\gamma \in \Omega_M \backslash \{0\}} \sum_{k \in E_\gamma} \overline{a_{Mk-Mj+\gamma}} \tilde{a}_{\gamma+Mk} , \quad j \in \mathbb{Z}^2.$$

Then mask $\tilde{a}$ is a dual mask of a and satisfies the sum rules of order J with dilation M.

For a given interpolatory mask a and an arbitrary integer J, in fact any finitely supported dual mask $\tilde{a}$ of a, with $\tilde{a}$ satisfying the sum rules of order J, must come from the above CBC algorithm. The CBC algorithm also works well with symmetry. When a is symmetric, a dual mask $\tilde{a}$ with symmetry and sum rules can be easily obtained via the CBC algorithm by appropriately selecting the sets E_γ. A general CBC algorithm is given in [26, Page 33 and Theorem 3.4.] for constructing finitely supported dual masks $\tilde{a}$ with arbitrarily high orders of sum rules as long as the primal mask a possesses at least one finitely supported dual mask which may not have any order of sum rules. See [5,25,26] for more details on the CBC algorithm.

The projection method can be used to obtain some optimal properties of bivariate orthonormal and biorthogonal wavelets ([25,28,34,35]).

Theorem 14 *Let $M = dI_2$ and $(a, \tilde{a})$ be a pair of dual masks with dilation M. Let $P = [1, 0]$. Suppose that the mask a is projectable, that is, $\hat{a}(\xi, \gamma\pi/d) = 0$ for all $\xi \in \mathbb{R}$ and $\gamma = 1, \ldots, d-1$. Then $(Pa, P\tilde{a})$ is also a pair of dual masks with dilation d. If (21) holds for ϕ and $\tilde{\phi}$, where ϕ and $\tilde{\phi}$ are standard M-refinable functions with masks a and $\tilde{a}$, then (21) still holds with ϕ and $\tilde{\phi}$ being replaced by $P\phi$ and $P\tilde{\phi}$. Moreover, $\nu_p(\phi) \leqslant \nu_p(P\phi)$, $\nu_p(\tilde{\phi}) \leqslant \nu_p(P\tilde{\phi})$, $\nu_p(a, dI_2) \leqslant \nu_p(Pa, d)$ and $\nu_p(\tilde{a}, dI_2) \leqslant \nu_p(P\tilde{a}, d)$ for all $1 \leqslant p \leqslant \infty$.*

Let us present a simple example to illustrate the CBC algorithm and the projection method on construction and analysis of bivariate biorthogonal wavelets.

Example 15 Let $M = 2I_2$ and a be the interpolatory mask given in (18). Note that the mask a is projectable. Let $P = [1, 0]$. Then $\widehat{Pa}(\xi) = (2 - \cos(\xi)) \cos^4(\xi/2)$. By the CBC algorithm, we see that the shortest symmetric dual mask b of Pa such that b satisfies the sum rules of order at least one is supported on $[-4, 4]$ and is uniquely given by $\hat{b}(\xi) := \cos^2(\xi/2)[14 - 5\cos(\xi) - 2\cos(2\xi) + \cos(3\xi)]/8$.

By calculation, we have $\nu_2(b,2) \approx 0.59322$. Take $J = 2$ since $sr(b,2) = 2$. By the CBC algorithm, any D_6-symmetric dual mask $\tilde{a}$ of a, such that $\tilde{a}$ is supported inside $[-4,4]^2$ and $sr(\tilde{a}, 2I_2) = 2$, must take the form

$$\frac{1}{64}\left[\begin{array}{ccccccccc}
0 & 0 & 0 & 0 & t_2 & 0 & 1-2t_2 & 0 & t_2 \\
0 & 0 & 0 & 0 & -t_3 & t_3 & t_3 & -t_3 & 0 \\
0 & 0 & 1-2t_2 & t_3 & t_1 & t_4 & t_1 & t_3 & 1-2t_2 \\
0 & 0 & t_3 & t_4 & 28+4t_1 & 28+4t_1 & t_4 & t_3 & 0 \\
t_2 & -t_3 & t_1 & 28+4t_1 & \mathbf{10-6t_1+6t_2} & 28+4t_1 & t_1 & -t_3 & t_2 \\
0 & t_3 & t_4 & 28+4t_1 & 28+4t_1 & t_4 & t_3 & 0 & 0 \\
1-2t_2 & t_3 & t_1 & t_4 & t_1 & t_3 & 1-2t_2 & 0 & 0 \\
0 & -t_3 & t_3 & t_3 & -t_3 & 0 & 0 & 0 & 0 \\
t_2 & 0 & 1-2t_2 & 0 & t_2 & 0 & 0 & 0 & 0
\end{array}\right]$$

with $t_3 := 10 + 2t_1 + 4t_2$ and $t_4 := -30 - 6t_1 - 4t_2$, where $\tilde{a}_{(0,0)} = 10 - 6t_1 + 6t_2$. By Theorem 14 and $P\tilde{a} = b$, we have $\nu_2(\tilde{a}, 2I_2) \leqslant \nu_2(b,2) \approx 0.59322$. Take $t_1 = -17/4$ and $t_2 = -1/2$, by calculation, we have $\nu_2(\tilde{a}, 2I_2) \approx 0.56715$. So, the conditions in Theorem 11 are satisfied and $\tilde{a}$ is near best in terms of the L_2 smoothness with respect to the support of the dual mask.

Bivariate Riesz Wavelets

Wavelet-based numerical algorithms have been successfully used in numerical solutions to partial differential equations and integral equations ([8,14]). The wavelets used in such wavelet-based methods are Riesz wavelets. For analysis on Riesz wavelets, see [10,33,40,41,42,47,56]. For applications based on Riesz wavelets, see [8,14,41,49].

Theorem 16 *Let M be a 2×2 dilation matrix with $m := |\det M|$ and a be a finitely supported bivariate mask. Let $\widehat{a^1}, \ldots, \widehat{a^{m-1}}$ be some 2π-periodic trigonometric polynomials such that $\det \mathcal{M}_{[\hat{a}, \widehat{a^1}, \ldots, \widehat{a^{m-1}}]}(\xi) \neq 0$ for all $\xi \in \mathbb{R}^2$ and $\widehat{a^1}(0) = \cdots = \widehat{a^{m-1}}(0) = 0$. Let ϕ be the standard refinable function associated with mask a and dilation M. If $\nu_2(a,M) > 0$ and $\nu_2(\tilde{a}, M) > 0$, where $\hat{\tilde{a}}$ is the $(1,1)$-entry of the matrix $[\overline{\mathcal{M}_{[\hat{a}, \widehat{a^1}, \ldots, \widehat{a^{m-1}}]}(\xi)}^{\mathrm{T}}]^{-1}$, then $\{\psi^1, \ldots, \psi^{m-1}\}$, which are defined in (25), generates a Riesz M-wavelet basis in $L_2(\mathbb{R}^2)$.*

Example 17 Let $M = 2I_2$ and $\hat{a}(\xi_1, \xi_2) := \cos^2(\xi_1/2)\cos^2(\xi_2/2)\cos^2(\xi_1/2 + \xi_2/2)$ in Example 4. Then $sr(a, 2I_2) = 4$ and a is D_6-symmetric with center 0. Define ([41,60])

$$\widehat{a^1}(\xi_1,\xi_2) := \mathrm{e}^{-i(\xi_1+\xi_2)}\overline{\hat{a}(\xi_1+\pi,\xi_2)},$$
$$\widehat{a^2}(\xi_1,\xi_2) := \mathrm{e}^{-i\xi_2}\overline{\hat{a}(\xi_1,\xi_2+\pi)},$$
$$\widehat{a^3}(\xi_1,\xi_2) := e^{-i\xi_1}\overline{\hat{a}(\xi_1+\pi,\xi_2+\pi)}\,.$$

Then all the conditions in Theorem 16 are satisfied and $\{\psi^1, \psi^2, \psi^3\}$ generates a Riesz $2I_2$-wavelet basis in $L_2(\mathbb{R}^2)$ ([41]). This Riesz wavelet derived from the Loop scheme has been used in [49] for mesh compression in computer graphics with impressive performance.

Pairs of Dual Wavelet Frames

In this section, we mention a method for constructing bivariate dual wavelet frames. The following Oblique Extension Principle (OEP) has been proposed in [18] (and independently in [7]. Also see [17]) for constructing pairs of dual wavelet frames.

Theorem 18 *Let M be a 2×2 dilation matrix and $a, \tilde{a}$ be two finitely supported masks. Let $\phi, \tilde{\phi}$ be the standard M-refinable functions with masks a and $\tilde{a}$, respectively, such that $\phi, \tilde{\phi} \in L_2(\mathbb{R}^2)$. If there exist 2π-periodic trigonometric polynomials $\Theta, \widehat{a^1}, \ldots, \widehat{a^L}, \widehat{\tilde{a}^1}, \ldots, \widehat{\tilde{a}^L}$ such that $\Theta(0) = 1$, $\widehat{a^1}(0) = \cdots = \widehat{a^L}(0) = \widehat{\tilde{a}^1}(0) = \cdots = \widehat{\tilde{a}^L}(0) = 0$, and*

$$\mathcal{M}_{[\Theta(M^{\mathrm{T}}\cdot)\hat{a}, \widehat{a^1}, \ldots, \widehat{a^L}]}(\xi)\overline{\mathcal{M}_{[\hat{\tilde{a}}, \widehat{\tilde{a}^1}, \ldots, \widehat{\tilde{a}^L}]}(\xi)}^{\mathrm{T}} = \mathrm{diag}(\Theta(\xi+\gamma_0), \ldots, \Theta(\xi+\gamma_{m-1}))\,, \quad (26)$$

where $\{\gamma_0, \ldots, \gamma_{m-1}\} = \Gamma_{M^{\mathrm{T}}}$ with $\gamma_0 = 0$ in Theorem 11. Then $(\{\psi^1, \ldots, \psi^L\}, \{\tilde{\psi}^1, \ldots, \tilde{\psi}^L\})$, defined in (25), generates a pair of dual M-wavelet frames in $L_2(\mathbb{R}^2)$.

For dimension one, many interesting tight wavelet frames and pairs of dual wavelet frames in $L_2(\mathbb{R})$ have been constructed via the OEP method in the literature, for more details, see [7,17,18,24,30,52,61,62] and references therein. The application of the OEP in high dimensions is much more difficult, mainly due to the matrix extension problem in (26).

The projection method can also be used to obtain pairs of dual wavelet frames ([34,35]).

Theorem 19 *Let M be an $r\times r$ dilation matrix and N be an $s\times s$ dilation matrix with $r \leqslant s$. Let P be an $r\times s$ integer matrix of rank r such that $MP = PN$ and $P^{\mathrm{T}}(\mathbb{Z}^r\backslash[M^{\mathrm{T}}\mathbb{Z}^r]) \subseteq \mathbb{Z}^s\backslash[N^{\mathrm{T}}\mathbb{Z}^s]$. Let $\psi^1, \ldots, \psi^L, \tilde{\psi}^1, \ldots, \tilde{\psi}^L$ be compactly supported functions in $L_2(\mathbb{R}^s)$ such that*

$$\nu_2(\psi^\ell) > 0 \quad \text{and} \quad \nu_2(\tilde{\psi}^\ell) > 0 \quad \forall\ \ell = 1, \ldots, L\,.$$

If $(\{\psi^1,\ldots,\psi^L\},\{\tilde\psi^1,\ldots,\tilde\psi^L\})$ *generates a pair of dual N-wavelet frames in* $L_2(\mathbb{R}^s)$, *then* $(\{P\psi^1,\ldots,P\psi^L\},\{P\tilde\psi^1,\ldots,P\tilde\psi^L\})$ *generates a pair of dual M-wavelet frames in* $L_2(\mathbb{R}^r)$.

Example 20 Let Ξ be an $r\times s$ (direction) integer matrix such that $\Xi^{\mathrm{T}}(\mathbb{Z}^r\backslash[2\mathbb{Z}^r])\subseteq\mathbb{Z}^s\backslash[2\mathbb{Z}^2]$. Let $M=2I_r$ and $N=2I_s$. Let $\{\psi^1,\ldots,\psi^{2^s-1}\}$ be the generators of the tensor product Haar orthonormal wavelet in dimension s, derived from the Haar orthonormal refinable function $\phi:=\chi_{[0,1]^s}$. Then by Theorem 19, $\{\Xi\psi^1,\ldots,\Xi\psi^{2^s-1}\}$ generates a tight $2I_r$-wavelet frame in $L_2(\mathbb{R}^r)$ and all the projected wavelet functions are derived from the refinable box spline function $\Xi\phi=M_\Xi$.

Future Directions

There are still many challenging problems on bivariate wavelets. In the following, we only mention a few here.

1. For any 2×2 dilation matrix M, e. g., $M=Q_{\sqrt{2}}$ in (1), can one always construct a family of MRA compactly supported orthonormal M-wavelet bases with arbitrarily high smoothness (and with symmetry if $|\det M|>2$)?
2. The matrix extension problem in (23) for orthogonal masks. That is, for a given orthogonal mask a with dilation M, find finitely supported high-pass filters $a^1,\ldots,a^{m-1}$ (with symmetry if possible) such that (23) holds with $\tilde a=a$ and $\tilde a^\ell=a^\ell$.
3. The matrix extension problem in (23) for a given pair of dual masks with symmetry. That is, for a given pair of symmetric dual masks a and $\tilde a$ with dilation M, find finitely supported symmetric high-pass filters $a^1,\ldots,a^{m-1},\tilde a^1,\ldots,\tilde a^{m-1}$ such that (23) is satisfied.
4. Directional bivariate wavelets. In order to handle edges of different orientations in images, directional wavelets are of interest in applications. See [13,55] and many references on this topic.

Bibliography

Primary Literature

1. Ayache A (2001) Some methods for constructing nonseparable, orthonormal, compactly supported wavelet bases. Appl Comput Harmon Anal 10:99–111
2. Belogay E, Wang Y (1999) Arbitrarily smooth orthogonal nonseparable wavelets in $\mathbb{R}^2$. SIAM J Math Anal 30:678–697
3. Carnicer JM, Dahmen W, Peña JM (1996) Local decomposition of refinable spaces and wavelets. Appl Comput Harmon Anal 3:127–153
4. Cavaretta AS, Dahmen W, Micchelli CA (1991) Stationary subdivision. Mem Amer Math Soc 93(453)1–186
5. Chen DR, Han B, Riemenschneider SD (2000) Construction of multivariate biorthogonal wavelets with arbitrary vanishing moments. Adv Comput Math 13:131–165
6. Chui CK (1992) An introduction to wavelets. Academic Press, Boston
7. Chui CK, He W, Stöckler J (2002) Compactly supported tight and sibling frames with maximum vanishing moments. Appl Comput Harmon Anal 13:224–262
8. Cohen A (2003) Numerical analysis of wavelet methods. North-Holland, Amsterdam
9. Cohen A, Daubechies I (1993) Nonseparable bidimensional wavelet bases. Rev Mat Iberoamericana 9:51–137
10. Cohen A, Daubechies I (1996) A new technique to estimate the regularity of refinable functions. Rev Mat Iberoamericana 12:527–591
11. Cohen A, Daubechies I, Feauveau JC (1992) Biorthogonal bases of compactly supported wavelets. Comm Pure Appl Math 45:485–560
12. Cohen A, Gröchenig K, Villemoes LF (1999) Regularity of multivariate refinable functions. Constr Approx 15:241–255
13. Cohen A, Schlenker JM (1993) Compactly supported bidimensional wavelet bases with hexagonal symmetry. Constr Approx 9:209–236
14. Dahmen W (1997) Wavelet and multiscale methods for operator equations. Acta Numer 6:55–228
15. Daubechies I (1988) Orthonormal bases of compactly supported wavelets. Comm Pure Appl Math 41:909–996
16. Daubechies I (1992) Ten Lectures on Wavelets, CBMS-NSF Series. SIAM, Philadelphia
17. Daubechies I, Han B (2004) Pairs of dual wavelet frames from any two refinable functions. Constr Approx 20:325–352
18. Daubechies I, Han B, Ron A, Shen Z (2003) Framelets: MRA-based constructions of wavelet frames. Appl Comput Harmon Anal 14:1–46
19. Dahlke S, Gröchenig K, Maass P (1999) A new approach to interpolating scaling functions. Appl Anal 72:485–500
20. de Boor C, Hollig K, Riemenschneider SD (1993) Box splines. Springer, New York
21. DeRose T, Forsey DR, Kobbelt L, Lounsbery M, Peters J, Schröder P, Zorin D (1998) Subdivision for Modeling and Animation, (course notes)
22. Dyn N, Levin D (2002) Subdivision schemes in geometric modeling. Acta Numer 11:73–144
23. Dyn N, Gregory JA, Levin D (1990) A butterfly subdivision scheme for surface interpolation with tension control. ACM Trans Graph 9:160–169
24. Han B (1997) On dual wavelet tight frames. Appl Comput Harmon Anal 4:380–413
25. Han B (2000) Analysis and construction of optimal multivariate biorthogonal wavelets with compact support. SIAM Math Anal 31:274–304
26. Han B (2001) Approximation properties and construction of Hermite interpolants and biorthogonal multiwavelets. J Approx Theory 110:18–53
27. Han B (2002) Symmetry property and construction of wavelets with a general dilation matrix. Lin Algeb Appl 353:207–225
28. Han B (2002) Projectable multivariate refinable functions and biorthogonal wavelets. Appl Comput Harmon Anal 13:89–102
29. Han B (2003) Computing the smoothness exponent of a symmetric multivariate refinable function. SIAM J Matrix Anal Appl 24:693–714

30. Han B (2003) Compactly supported tight wavelet frames and orthonormal wavelets of exponential decay with a general dilation matrix. J Comput Appl Math 155:43–67
31. Han B (2003) Vector cascade algorithms and refinable function vectors in Sobolev spaces. J Approx Theory 124:44–88
32. Han B (2004) Symmetric multivariate orthogonal refinable functions. Appl Comput Harmon Anal 17:277–292
33. Han B (2006) On a conjecture about MRA Riesz wavelet bases. Proc Amer Math Soc 134:1973–1983
34. Han B (2006) The projection method in wavelet analysis. In: Chen G, Lai MJ (eds) Modern Methods in Mathematics. Nashboro Press, Brentwood, pp. 202–225
35. Han B (2008) Construction of wavelets and framelets by the projection method. Int J Appl Math Appl 1:1–40
36. Han B, Jia RQ (1998) Multivariate refinement equations and convergence of subdivision schemes. SIAM J Math Anal 29:1177–1199
37. Han B, Jia RQ (1999) Optimal interpolatory subdivision schemes in multidimensional spaces. SIAM J Numer Anal 36:105–124
38. Han B, Jia RQ (2002) Quincunx fundamental refinable functions and quincunx biorthogonal wavelets. Math Comp 71:165–196
39. Han B, Jia RQ (2006) Optimal C^2 two-dimensional interpolatory ternary subdivision schemes with two-ring stencils. Math Comp 75:1287–1308
40. Han B, Jia RQ (2007) Characterization of Riesz bases of wavelets generated from multiresolution analysis. Appl Comput Harmon Anal 23:321–345
41. Han B, Shen Z (2005) Wavelets from the Loop scheme. J Fourier Anal Appl 11:615–637
42. Han B, Shen Z (2006) Wavelets with short support. SIAM J Math Anal 38:530–556
43. Ji H, Riemenschneider SD, Shen Z (1999) Multivariate compactly supported fundamental refinable functions, duals, and biorthogonal wavelets. Stud Appl Math 102:173–204
44. Jia RQ (1998) Approximation properties of multivariate wavelets. Comp Math 67:647–665
45. Jia RQ (1999) Characterization of smoothness of multivariate refinable functions in Sobolev spaces. Trans Amer Math Soc 351:4089–4112
46. Jia RQ, Micchelli CA (1991) Using the refinement equation for the construction of pre-wavelets II: Power of two. In: Laurent PJ, Le Méhauté A, Schumaker LL (eds) Curves and Surfaces. Academic Press, New York, pp. 209–246
47. Jia RQ, Wang JZ, Zhou DX (2003) Compactly supported wavelet bases for Sobolev spaces. Appl Comput Harmon Anal 15:224–241
48. Jiang QT (1998) On the regularity of matrix refinable functions. SIAM J Math Anal 29:1157–1176
49. Khodakovsky A, Schröder P, Sweldens W (2000) Progressive geometry compression. Proc SIGGRAPH
50. Kovačević J, Vetterli M (1992) Nonseparable multidimensional perfect reconstruction filter banks and wavelet bases for $\mathbb{R}^n$. IEEE Trans Inform Theory 38:533–555
51. Lai MJ (2006) Construction of multivariate compactly supported orthonormal wavelets. Adv Comput Math 25:41–56
52. Lai MJ, Stöckler J (2006) Construction of multivariate compactly supported tight wavelet frames. Appl Comput Harmon Anal 21:324–348
53. Lawton W, Lee SL, Shen Z (1997) Stability and orthonormality of multivariate refinable functions. SIAM J Math Anal 28:999–1014
54. Lawton W, Lee SL, Shen Z (1998) Convergence of multidimensional cascade algorithm. Numer Math 78:427–438
55. Le Pennec E, Mallat S (2005) Sparse geometric image representations with bandelets. IEEE Trans Image Process 14:423–438
56. Lorentz R, Oswald P (2000) Criteria for hierarchical bases in Sobolev spaces. Appl Comput Harmon Anal 8:32–85
57. Maass P (1996) Families of orthogonal two-dimensional wavelets. SIAM J Math Anal 27:1454–1481
58. Mallat S (1998) A wavelet tour of signal processing. Academic Press, San Diego
59. Riemenschneider SD, Shen Z (1997) Multidimensional interpolatory subdivision schemes. SIAM J Numer Anal 34:2357–2381
60. Riemenschneider SD, Shen ZW (1992) Wavelets and prewavelets in low dimensions. J Approx Theory 71:18–38
61. Ron A, Shen Z (1997) Affine systems in $L_2(\mathbb{R}^d)$: the analysis of the analysis operator. J Funct Anal 148:408–447
62. Ron A, Shen Z (1997) Affine systems in $L_2(\mathbb{R}^d)$ II: dual systems. J Fourier Anal Appl 3:617–637
63. Sweldens W (1996) The lifting scheme: a custom-design construction of biorthogonal wavelets. Appl Comput Harmon Anal 3:186–200

Books and Reviews

Cabrelli C, Heil C, Molter U (2004) Self-similarity and multiwavelets in higher dimensions. Mem Amer Math Soc 170(807):1–82

Boolean Modeling of Biological Networks

Peter Krawitz[1], Ilya Shmulevich[2]
[1] Fakultät für Physik, Ludwig Maximilians Universität, München, Germany
[2] Institute for Systems Biology, Department of Bioengineering University of Washington, Seattle, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Theoretical Aspects of Boolean Networks
Boolean Models of Biological Networks
Network Inference
Future Directions
Bibliography

Glossary

Connectivity matrix In a Boolean network model, the information about *who* is interacting with *whom* is stored in the connectivity matrix. This static wiring diagram can be described by parameters from graph the-

ory, such as the degree distributions of its elements or nodes.

Activity, sensitivity Boolean functions define *how* elements of a network interact. Activity and sensitivity are parameters describing the dynamic properties of single network elements or nodes. The activity of a node measures its influence on another node and conversely, the sensitivity of a node measures how sensitive it is to random changes in one of its inputs.

Phase transition and criticality Regarding the propagation of perturbations, one can distinguish between an ordered phase, where perturbations will not influence the network dynamics and an chaotic phase, where small perturbations will increase over time. Networks that operate at the transition of these two phases, such that perturbations are neither attenuated nor amplified on average, are called critical.

Attractors In Boolean networks, dynamically stable patterns are called attractors. In their biological context, attractors may be interpreted as different cell types in models of cell differentiation or as cell states in models of cell signaling. Analyzing the influence of perturbations on attractor dynamics is thus of pivotal biological interest.

Relevant and irrelevant nodes Only the perturbation of certain elements will move the system from one attractor to another. These dynamically relevant nodes are potential lever points for system intervention.

Definition of the Subject

Boolean Networks (BN) are a class of discrete dynamical systems that can be characterized by the interaction of a set of Boolean variables. Random Boolean Networks (RBN), which are ensembles of random network structures, were first introduced by Stuart Kauffman in 1969 as a simple model class for studying dynamical properties of gene regulatory networks [1,2]. Since then, Boolean Networks have been used as generic models for dynamics of complex systems of interacting entities, such as social and economic networks, neural networks, as well as gene and protein interaction networks. Despite their conceptual simplicity, Boolean Networks exhibit complex nonlinear behaviors that are, to this day, a challenging object of investigation for theoretical physicists and mathematicians. Further, a discretization of gene expression is often regarded as an experimentally justifiable simplification [3,4], making Boolean network models attractive tools for the application oriented biologist. In this chapter, the theoretical findings for RBNs of the last decade are reviewed and the value of Boolean network models for the understanding of biological networks is illustrated.

Introduction

Since the early 1960s, the hypothesis that different dynamically stable patterns of gene expression are the basic key to understanding the diversity of functionally and morphologically different cells of a single organism began to prevail and today, a gene expression profile of a cell is often regarded as a sufficient definition of its state [5]. In 1969, long before accurate gene numbers or the very order of the magnitude of genome sizes in higher developed organisms were known, Stuart A. Kauffman introduced random Boolean networks as a generic model to study and explain how an increasing number of dynamical stable states can arise from a growing network size [1]. The now classical model introduced by Kauffman consists of n binary variables x_i (also referred to as nodes or genes), that are each connected to k randomly chosen other nodes. A Boolean function f_i that specifies a value for every possible input combination, $f_i : \{0,1\}^k \to \{0,1\}$, is assigned randomly to each node x_i. For k input variables, there are 2^{2^k} different functions. The state of a node x_i is then updated by calculating the Boolean function value of its k input nodes: $x_i(t+1) = f_i(x_{i_1}(t), \ldots, x_{i_k}(t))$.

Under synchronous updating, a wiring diagram (i. e. which nodes are connected to which other nodes) and the Boolean functions assigned to every node define a deterministic system that assigns to any network state $\mathbf{x}(t) \in \{0,1\}^n$ a successor state $\mathbf{x}(t+1) = F(\mathbf{x}(t))$, where $F : \{0,1\}^n \to \{0,1\}^n$ is a vector valued (multi-output) Boolean function that is comprised of the Boolean functions assigned to each of the nodes and that represents the state transition rules of the system. For a given initial state, a network thus defines a trajectory through the state space of size 2^n. As this system is deterministic and the state space finite, a trajectory will eventually fall on a previously visited state and periodically circle through a certain number l of states from there on. These reoccurring and therefore dynamically stable states form an *attractor* of size l and may, in the context of gene regulatory networks, be interpreted as functionally salient states. For example, in the context of cellular differentiation, such attractors may correspond to cell types or cell fates [2]. All states from which the system eventually flows to a given attractor constitute that attractor's *basin of attraction*. The entire state space is naturally divided into disjoint basins of attraction. Kauffman observed in his numerical simulations that for $k = 2$, the mean number of attractors scaled approximately with the square root of the system size – a scaling behavior that

seemed to agree with the increasing number of different cell types in higher organized organisms of larger genome sizes.

Although the scaling behavior of the mean number of attractors in such network ensembles as well as the estimated genome sizes are now outdated, Kauffman's ensemble approach to gene regulation was pioneering and is still valid [6]: define an ensemble of Boolean networks such that it fulfills certain known features of biological networks and then study random instances of these networks to learn more about general properties of these ensembles. This iterative and highly interdisciplinary process depends, on the one hand, on the input of new biological findings and might guide, on the other hand, the research of the experimentalist. The Kauffman ensembles of the parameter $k = 2$, for example, turned out to be "critical", lying at a phase transition between an ordered and a chaotic phase, in regard to the propagation of perturbations, and contributed to the hypothesis that living systems may be poised at the "edge of chaos" – a hypothesis for which experimental evidence has been mounting [7,8,9].

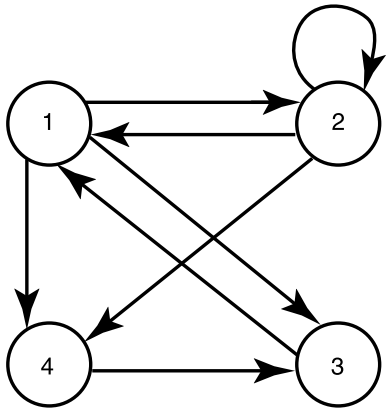

Boolean Modeling of Biological Networks, Figure 1
The connectivity graph of a Boolean network of four nodes

Boolean Modeling of Biological Networks, Table 1
Connectivity matrix of the example network

$x_i \backslash i$	1	2	3	4
x_{i_1}	2	1	1	1
x_{i_2}	3	2	4	2

Boolean Modeling of Biological Networks, Table 2
Boolean functions of the example network

x_{i_1}	x_{i_2}	f_1	f_2	f_3	f_4
0	0	0	1	0	0
0	1	0	0	1	1
1	0	1	1	1	1
1	1	1	1	0	1

Theoretical Aspects of Boolean Networks

In this section some key findings learned from the study of ensembles of Boolean networks will be illustrated. The understanding of dynamic properties of a network and how these are caused may help to identify the pivotal features in a biological Boolean network model. For a profound and detailed review of the theory of Random Boolean Networks, see also [10].

The Concept of Relevant Nodes

Given a particular starting state, the attractor to which the system will eventually flow starting from this state will be determined entirely by so-called *relevant nodes*. Some nodes of a network will always take on the same value regardless of the attractor on which the network finally settles down. These nodes establish the *frozen core* of a network. They are either updated by a constant Boolean function and are *frozen* to a fixed value right from the beginning or they freeze as the system transitions through its state space, in which case they are called *clamped*. Other nodes may have different values on different attractors, but their values are of no importance for determining the attractor to which the system will flow. These nodes are therefore called *non-frozen*, but *irrelevant*. Finally, there are the non-frozen *relevant* nodes that can be recursively defined as influencing other relevant nodes. Relevant nodes ultimately have an effect on their own state. The concept of classifying the nodes of a network into clamped, irrelevant, and relevant, regarding their importance for the attractor dynamics, is simple yet powerful for the understanding of Boolean network dynamics [11,12,13].

The example network, which is illustrated in Fig. (1) and whose connectivity matrix, Boolean functions and transition matrix are given in Tables 1–3, has two attractors, a fixed point, $(x_1x_2x_3x_4) = (1101)$ and an attractor of length two, $(0101) \rightleftarrows (1011)$. Only the nodes 1 and 2 are relevant, whereas node 3 is non-frozen but irrelevant and node 4 is a clamped node, that freezes to the value $x_4 = 1$ as soon as the system reaches an attractor.

Algorithmically, relevant nodes can often be identified by an iterative procedure:

First, links of frozen nodes to other nodes are cut, as this kind of input is not *variable* but *constant*.

Second, so-called '*degenerate*' links have to be identified and removed from the connectivity matrix. In the example network a closer look at the Boolean function of node 1 reveals that it actually depends only on its first

Boolean Modeling of Biological Networks, Table 3
Transition matrix of the example network

x				$F(\mathbf{x})$			
x_1	x_2	x_3	x_4	f_1	f_2	f_3	f_4
0	0	0	0	0	1	0	0
0	0	0	1	0	1	1	0
0	0	1	0	0	1	0	0
0	0	1	1	0	1	1	0
0	1	0	0	1	0	0	1
0	**1**	**0**	**1**	**1**	**0**	**1**	**1**
0	1	1	0	1	0	0	1
0	1	1	1	1	0	1	1
1	0	0	0	0	1	1	1
1	0	0	1	0	1	0	1
1	0	1	0	0	1	1	1
1	**0**	**1**	**1**	**0**	**1**	**0**	**1**
1	1	0	0	1	1	1	1
1	**1**	**0**	**1**	**1**	**1**	**0**	**1**
1	1	1	0	1	1	1	1
1	1	1	1	1	1	0	1

variable, which is node 2. Therefore link $3 \to 1$ can be removed from the connectivity matrix.

Third, nodes that do not influence other nodes are irrelevant and can be removed from the network.

This procedure is repeated until no further nodes can be removed from the network. In our example network node 3 can be removed after cutting link $3 \to 1$. In a second round node 4 can be removed, which leaves us with the two relevant nodes 1 and 2. Some nodes identified by this method, however, are actually not relevant, but *freeze* to a constant value, as the following example illustrates: $f_1 = x_1 \vee x_2$, $f_2 = x_1 \vee \overline{x_2}$. The properties of these self-freezing nodes were first described in [14].

The Critical Notion of Criticality

Generally, a dynamic system is said to behave in an *ordered* or *robust* manner when similar initial conditions will lead to similar trajectories in the phase (state) space. A system is called *chaotic* when even infinitesimal differences will cause an increasing divergence in the ensuing trajectories. In Boolean networks, the difference between two states $\mathbf{x}$ and $\mathbf{x}'$ is usually quantified in terms of the normalized Hamming distance η:

$$\eta(\mathbf{x}, \mathbf{x}') = \frac{1}{n} \sum_{i=1}^{n} \left| x_i - x_i' \right| . \tag{1}$$

With this metric, the evolution of the overlap between two random states, $a(t) = 1 - \eta(\mathbf{x}(t), \mathbf{x}'(t))$, can be measured over time.

In the second half of the 1980s, Derrida, Flyvbjerg and others began to study the propagation of perturbations in Boolean networks with a mean field approach that had already successfully been applied in the theory of spin glasses [15,16,17]. Instead of considering the precise effects of all nodes on their linked nodes in a certain network instance, the mean field approach only deals with the average effect on a random node. If the overlap between two states, $\mathbf{x}$ and $\mathbf{x}'$, is $a(t)$ at a given time t, the probability that a random node i will receive the same input signal in the next step, is $a(t)^k$. If the Boolean function f_i assigned to the node is generated with *bias* p, meaning that the function value of a random input vector is 1 with probability p, then the probability that two random input vectors yield the same function value is $P = p^2 + (1-p)^2$. The expected overlap of the two states after one time step is given by:

$$a(t+1) = a(t)^k + P(1 - a(t))^k . \tag{2}$$

In the limit of t and $n \to \infty$, $a(t)$ asymptotically approaches the fixed point a^*, which obeys:

$$a^* = 1 - (1 - (a^*)^k) 2p(1-p) . \tag{3}$$

The critical transition curve between the ordered phase, where the overlap approaches $a^* = 1$, and a chaotic phase, where the overlap approaches a value $a^* < 1$ is given by the relation

$$2kp(1-p) = 1 . \tag{4}$$

Figure 2 shows the avareged overlap a $(+)$ of two random states of Hamming distance $\eta(\mathbf{x}, \mathbf{x}') = 0.01$ after 200 time steps in network ensembles of $n = 1000$ nodes and connectivity $k = 3$ as a function of the bias p, as well as the results of the annealed approximation (blue line, Eq. 3). In ensembles with $p \geq 0.7887$ the initial perturbation dies out and the overlap approaches $a = 1$.

Another order parameter of the phase transition, which is based on the concept of Boolean derivatives, is the network sensitivity s or the expected sensitivity of a random node [18,19]. The partial derivative of a Boolean function f with respect to x_j is defined as

$$\frac{\partial f}{\partial x_j} = f(\mathbf{x}^{(j,0)}) \oplus f(\mathbf{x}^{(j,1)}) , \tag{5}$$

where $\oplus$ is addition modulo 2 and $\mathbf{x}^{(j,\sigma)} = (x_1, \ldots, x_{j-1}, \sigma, x_{j+1}, \ldots, x_k)$, $\sigma = 0, 1$. The $n \times n$ Jacobian matrix F'_{ij} of a network at time t is then defined as

$$F'_{ij}(\mathbf{x}(t)) = \frac{\partial f_i(\mathbf{x}(t))}{\partial x_j} . \tag{6}$$

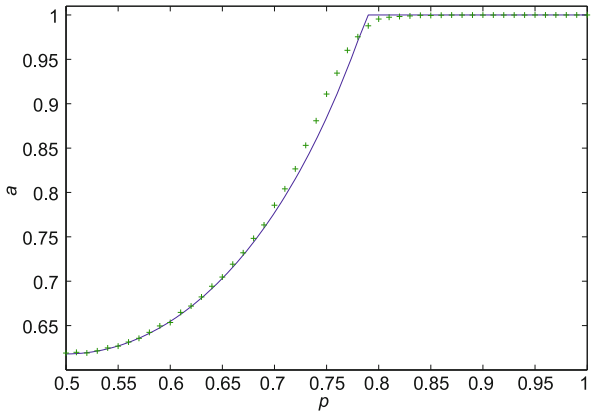

Boolean Modeling of Biological Networks, Figure 2
The averaged overlap *a*, of two random states $\mathbf{x}, \mathbf{x}'$ with $\eta(\mathbf{x}, \mathbf{x}') = 0.01$ in network ensembles $B(n, k, p)$, $n = 1000$, $k = 3$ and $p = [0.5, 1]$, after 200 time steps is indicated by (+). The *blue line* shows the annealed approximation (Eq. 3)

The behavior of a perturbation vector $\mathbf{x}$ is then determined by the successive products of the Jacobian matrices

$$Q(t) = F'(\mathbf{x}(0)) \cdot F'(\mathbf{x}(1)) \ldots . F'(\mathbf{x}(t)) \,. \tag{7}$$

An initial perturbation disappears if $Q(t)$ converges to a matrix Q^* consisting of zeros only.

In a mean field approach, the deterministic matrix $F'(\mathbf{x}(t))$ is substituted by a random matrix R of the same form. The probability for a random Boolean function to have a partial derivative equal to 1 is equal to the probability that a change in its input x_j causes a change in the function value. In an ensemble with a given bias p this happens with probability $2p(1-p)$. The mean number of 'ones' in each row is accordingly $2kp(1-p)$ and $Q(t+1)$ becomes $Q(t+1) = RQ(t)$. The percentage $q(t)$ of zeros in $Q(t)$ evolves in the limit of large n as

$$q(t+1) = \lim_{n\to\infty} \left[1 - \frac{2p(1-p)k}{n}\left(1 - q(t)\right)\right]^n \tag{8}$$

$$= e^{-(1-q(t))2p(1-p)k} \,. \tag{9}$$

By analyzing this equation for the fixed point $q^* = 1$ one finds again the critical transition between the ordered phase, where perturbations die out over time, and the chaotic phase, where even small perturbations spread, given by:

$$\left.\frac{q(t+1)}{q(t)}\right|_{q^*=1} = 2kp(1-p) > 1 \,.$$

The activity $\alpha_j^{f_i}$ of a variable x_j in a function f_i of k variables is defined as the mean value of the partial derivatives over all possible inputs of this function

$$\alpha_j^{f_i} = \frac{1}{2^k} \sum_{\mathbf{x}\in\{0,1\}^k} \frac{\partial f_i(\mathbf{x})}{\partial x_j} \,. \tag{10}$$

It gives the probability that a change in the value of this variable will cause a change in the function's value. The sensitivity s^i of a node is the sum of the activities of its variables $s^i = \sum_{j=1}^k \alpha_j^{f_i}$ and indicates how likely the function value is to change under a random change in its arguments. Critical random Boolean networks have an expected network sensitivity of $s = \frac{1}{n}\sum_{i=1}^n s^i = 1$, a perturbation in one node will on average change the value of one other node in the next time step. It is an appealing feature of the sensitivity parameter that it is independent of from the network architecture and entirely determined by the Boolean functions in a given network. The sensitivity of a Boolean model of a biological network and the activities of its nodes are easy to compute and may help conceptualize the dynamic characteristics of its elements. Using the mixing ratio of different classes of Boolean function as basis for the network dynamics the definition of criticality was further generalized in [20].

Scaling Behaviors in RBNs

With increasing computational power, the mean number and length of attractors in larger network ensembles could be studied and the original hypothesis of a power law-like scaling of the mean number of attractors in critical networks was shown to be incorrect. In 2003, Samuelson and Troein showed by analytical means that the number of attractors grows superpolynomially in critical networks [21]. This astonishing result attracted new interest in the scaling behavior of other network parameters. Kaufman, Mihaljev and Drossel showed that the number of relevant nodes, n_r, in a general class of critical networks increases in a power law manner with the network size, $n_r \sim n^{1/3}$ [20,22]. A further result of this work is that relevant nodes of random critical ensembles are only rarely regulated by more than one other relevant node – a finding whose implications on the ensemble approach will have to be carefully considered.

Another entity that increases only in critical network ensembles is the recently introduced *basin entropy*. A Boolean network classifies its state space into several basins of attraction by assigning a unique attractor to every state. The size w_ρ of basin ρ is the number of states belonging to that basin (including the attractor states, which are by definition part of the basin). Thus, the basin sizes describe the complexity of the state space partition.

In the example network, for instance, the basin of the fixed point only drains one forth of all states, whereas the rest flow into the second attractor. The basin entropy, $h = -\sum_{\rho=1}^{m} w_\rho \ln w_\rho$, where m is the total number of basins/attractors, serves as a measure for the complexity of this classification process. Equivalently, the basin entropy measures the uncertainty of a system that is started in a random state. Surprisingly, basin entropy only increases in critical network ensembles, whereas it approaches a finite limit in the ordered and highly chaotic phase [23]. This suggests that the informationally optimal partition of the state space is achieved when the system is operating at the critical boundary between the ordered and disordered phases. For biological networks this means that their informational complexity will only grow with size when their elements are organized in a critical manner.

Under asynchronous updating schemes many attractors of the synchronous scheme become unstable [24,25]. For the scaling behaviour of attractors under asynchronous updating, the mean number of attractors increases only polynomially with system size [26].

Modifications to the Classical Model

The fixed in-degree distribution and the synchronous updating scheme are clearly among the most evident shortcomings of the classical Kauffman model from a biological point of view. The discovery of topological similarities in a variety of complex networks [27] spurred interest in the dynamical implications of 'scale-free' architectures of Boolean network models in which the degree distribution follows a power law. Aldana et al. showed that Boolean networks with a scale-free topology give rise to a much broader class of networks that exhibit robust network dynamics [28]. Only recently, realistic data for in- and out degree distributions in gene regulatory networks became available through genome-wide localization analysis of transcription factors [29]. These topological findings can now be incorporated into more realistic random network ensembles.

Harris studied in an extensive literature search the frequencies of logical rules in eukaryotic gene transcription [30] and showed that canalizing functions occur far more often than randomly expected. A variable x_i is called *canalizing* for a certain value $x_i = \sigma_i \in \{0, 1\}$, if it determines the value $\sigma_j \in \{0, 1\}$ of the function $f_j(\ldots, x_i = \sigma_i, \ldots) = \sigma_j$. In the example network for instance in function f_4 both variables x_1, x_2 are canalizing for $\sigma = 1$, whereas in function f_3 none of the variables are canalizing, meaning that both variables are always necessary to determine the value of the function. Canalizing and related classes of functions were studied in the context of the sources of order and remarkable robustness observed in genetic regulatory networks [31] as well as in terms of stability [32]. The knowledge about which logical rules actually do occur in biological networks is not only important for more realistic ensembles but might also help to make an educated guess in the modeling of a concrete biological circuit, when the updating rules for a gene are uncertain.

Although the synchronous updating scheme has been known to be biologically unrealistic, it has been applied in most analytical studies so far, as it is the only one that yields deterministic state transitions. Under asynchronous updating, many attractors of the synchronous scheme become unstable and decay into other dynamically stable cycles. However, the very definition of an attractor, except for fixed points remains difficult. Further, asynchronous updating is not per se GE biologically more realistic and has to be motivated carefully in every case not to fall victim to artefacts [33].

Boolean Models of Biological Networks

Biological networks can often be modelled as logical circuits from well-known local interaction data in a straightforward way. This is clearly one of the advantages of the Boolean network approach. Though logical models may sometimes appear obvious and simplistic they help to understand the dynamic key properties of a regulatory process. They may also inspire the experimentalist to ask new question and to test them first in silico. The following three network models exemplify how Boolean networks may serve as useful in silico tools from cell differentiation processes to cell signaling.

The Segment Polarity Network of the Fruit Fly

The segment polarity genes represent the last step in the hierarchical cascade of gene families initiating the segmented body of the fruit fly *Drosophila melanogaster*. The stable maintenance of the segment polarity gene expression pattern is a crucial requirement in the further development of the embryo. The dynamics of the best characterized segment polarity genes have been studied in different modelling approaches in order to understand the stability and robustness of this pattern formation [33,34,35,36]. Albert and Othmer suggested a Boolean model of gene and protein interactions which is depicted in Fig. 3. This model correctly reproduces the characteristic expression patterns observed in the wild-type development. Further, distinct '*knock down*' or '*overexpression*' phenotypes can be simulated through the model by fixing

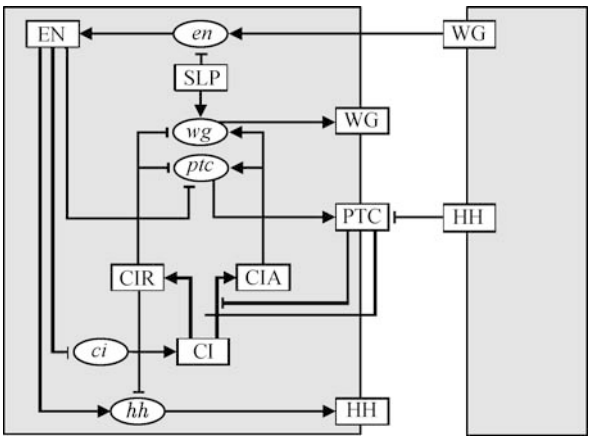

Boolean Modeling of Biological Networks, Figure 3
The network of interactions between the segment polarity genes. The *gray background layers* illustrate two neighboring cells indicating that some interactions in this network are intercelular [36]

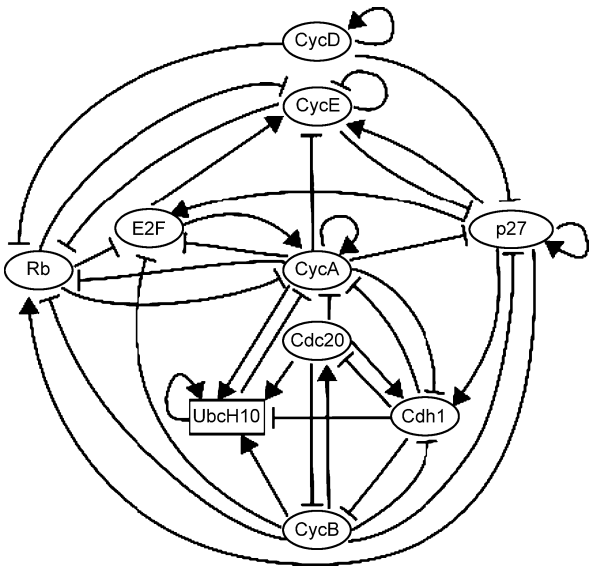

Boolean Modeling of Biological Networks, Figure 4
Logical regulatory graph for the mammalian cell cycle network. Each *node* represents the activity of a key regulatory element, whereas the *edges* represent cross-regulations. *Blunt arrows* stand for inhibitory effects, *normal arrows* for activations [37]

the state of the particular gene. An analysis of the robustness of the wild-type steady states under asynchronous updating schemes reveals that the model is very sensitive to temporal variations, whereas the wild-type development is stable as soon as certain differences in the temporal order of transcriptional and translational processes are incorporated into the updating procedure.

Control of the Cell Cycle

Proliferating eukaryotic cells perform a series of highly regulated and coordinated actions collectively referred to as the cell cycle. As cancer development is associated with a loss of control over the cell cycle, an understanding of its regulatory dynamics is of paramount importance. The integration of the biological and mostly qualitative knowledge about interacting genes and proteins into a Boolean network model allows the analysis of the dynamic regulatory properties. Fauré et al. studied the effects of different updating schemes on the dynamics of a highly abstracted cell cycle control model [37]. Under the simplistic synchronous updating scheme the network gives rise to two different attractors, a stable state, matching the quiescent cell when growth factors are lacking and a complex dynamical cycle representing the cell cycle when the cyclin D complex is present (Fig. 4).

T-Cell Receptor Signaling

T lymphocytes play a key role in the regulation of the adaptive immune system, as they are able to differentiate between self- and foreign-, potentially dangerous, antigens. T cells recognize foreign peptides that are presented by antigen presenting cells by means of the T cell receptor (TCR) and costimulatory molecules (CD4/8/28). This stimulus initiates a signaling cascade within the T cell that eventually influences its activation status. The T cell response must be tightly regulated, as a response to a pathogen that is too weak endangers the whole organism, whereas an overshooting reaction may lead to autoimmune disorders. Saez-Rodriguez et al. constructed a complex Boolean model of this signaling network from local interaction knowledge (Fig. 5). In addition to reproducing known global behaviors, the model also predicted unexpected signaling events after antibody-mediated perturbation of CD28, which could experimentally be validated [38,39].

Network Inference

While the inference of the network models described in the last section was clearly knowledge-driven and required the input of many interpreted biological experiments, network inference algorithms focus on automating or at least assisting this process. Different types of data can help reveal different aspects of the underlying network. For example, genome-wide transcription factor localization studies may reveal which elements are interacting, although they won't tell much about the logical rules representing their relationships. Time-series gene expression data can fill that gap and yield the required state transitions, in order to

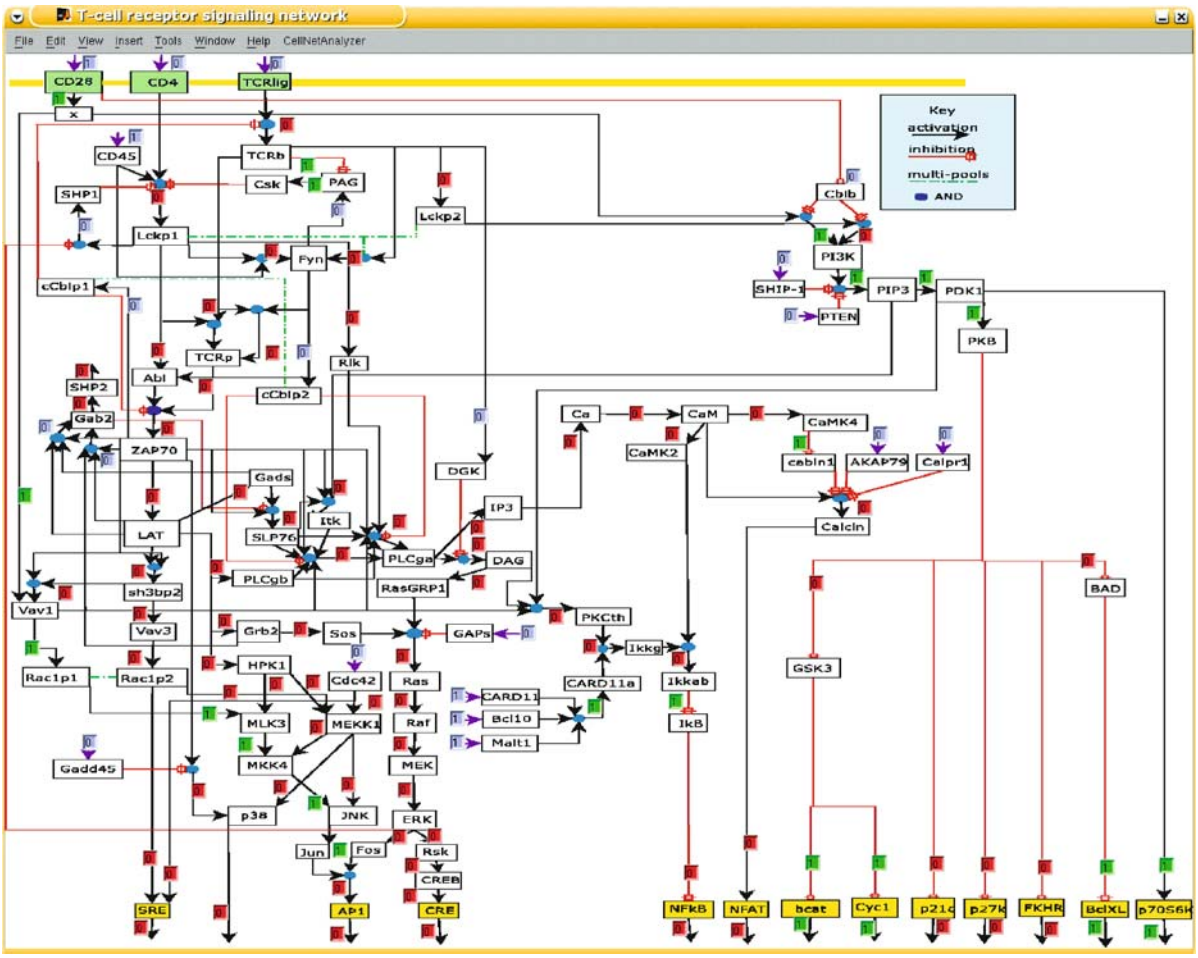

Boolean Modeling of Biological Networks, Figure 5

Logical model of T cell signaling (screenshot from *CellNetAnalyzer*). The network consisting of 96 nodes describes the signaling from the T cell receptor and its coreceptors to the key transcription factors AP-1, NFAT and NFκB, which determine T cell activation and function [38]

derive the Boolean functions. However, it is often experimentally unrealistic to obtain a sufficiently rich time-series data set necessary to infer the Boolean rules non-ambiguously.

Suppose the connectivity graph of the example network is given and both attractor profiles are known, then Boolean rules for node 1, 2 and 4 can be derived, whereas this data is consistent with several logical rules for node 3: $f_3 = \overline{x_1}x_2$, or $x_1\overline{x_2} \vee \overline{x_1}x_2$ or even $x_1\overline{x_2} \vee \overline{x_1}x_2 \vee \overline{x_1x_2}$. Further, whenever one deals with real data, one has to take into account experimental noise. Thus, inference algorithms should not necessarily aim to identify the hypothetical single correct network, but instead, either identify a set of networks that optimally fit the data or construct a network model that incorporates stochasticity. The former approach, chosen by Martin et al., was based on time series data from microarray expression profiling; the authors identified a set of networks that best fit the data by an algorithm proposed by Akutsu et al. [40] and analyzed the dynamics of this set [41]. For the latter approach, the model of Probabilistic Boolean Networks can be used – if the data does not allow a non-ambiguous choice of an updating rule, a set of probable rules can be implemented in the model [42,43,44].

Future Directions

The theory of Boolean networks has always been enriched by data from biology. The theoretical description of criticality and its implications might be given as a paradigm. Today, new findings about the scale free and modular organization of networks are waiting for a theoretical expla-

nation. The increasingly precise data derived from high troughput genome-wide measurement technologies will yield more realistic in- and out-degree distributions of biological systems. The implications of these archictectural features on the dynamics will also be numerically investigated by means of the ensemble approach. Also the effects of various update schemes on the network dynamics and how e. g. delays may help to filter out perturbations have to be further studied.

On the other hand the increasing quantity and quality of high-throughput gene and protein expression data and sophisticated algorithms for network inference provide the basis for a plethora of new biological network models, that can be analyzed by the Boolean formalism and its probabilistic extensions. Such models can then be used to make predictions about how the system would respond to various perturbations and environmental signals and, conversely, what kinds of intervention strategies would be most successful for controlling the system that is in a diseased state [45].

Bibliography

1. Kauffman SA (1969) Metabolic stability and epigenesis in randomly constructed genetic sets. J Theor Biol 22:437–467
2. Kauffman SA (1993) The Origins of Order. Oxford University Press, New York
3. Thomas R (1973) Boolean formalization of genetic control circuits. J Theor Biol 42:563–585
4. Yuh CH, Bolouri H, Bower JM, Davidson EH (2001) A logical model of cis-regulatory control in a eukaryotic system. In: Bower JM, Bolouri H (eds) Computational Modeling of Genetic and Biochemical Networks. MIT Press, Cambridge, pp 73–100
5. Jacob F, Monod J (1961) On the regulation of gene activity. Cold Spring Harbor Symposia on Quantitative Biology
6. Kauffman SA (2004) The ensemble approach to understand genetic regulatory networks. Physica A 340:733–740
7. Ramo P, Kesseli J, Yli-Harja O (2006) Perturbation avalanches and criticality in gene regulatory networks. J Theor Biol 242:164–170
8. Shmulevich I, Kauffman SA, Aldana M (2005) Eukaryotic cells are dynamically ordered or critical but not chaotic. Proc Natl Acad Sci USA 102:13439–13444
9. Nykter M, Price ND, Aldana M, Ramsey SA, Kauffman SA, Hood L, Yli-Harja O, Shmulevich I (2008) Gene Expression Dynamics in the Macrophage Exhibit Criticality. Proc Natl Acad Sci USA 105(6):1897–1900
10. Drossel B (2007) Random Boolean networks. In: Schuster HG (ed) Annual Review of Nonlinear Dynamics and Complexity, vol 1. Wiley, Weinheim
11. Bastolla U, Parisi G (1998) Relevant elements,magnetization and dynamical properties in Kauffman networks: A numerical study. Physica D 115:203–218
12. Bastolla U, Parisi G (1998) The modular structure of Kauffman networks. Physica D 115:219–233
13. Socolar JES, Kauffman SA (2003) Scaling in Ordered and Critical Random Boolean Networks. Phys Rev Lett 90:068702
14. Paul U, Kaufman V, Drossel B (2006) Properties of attractors of canalyzing random Boolean networks. Phys Rev E 73:026118
15. Derrida B, Pomeau Y (1986) Random networks of automata: A simple annealed approximation. Europhys Lett 1:45–49
16. Derrida B, Stauffer D (1986) Phase transition in two dimensional Kauffman cellular automata. Europhys Lett 2:739–745
17. Flyvbjerg H (1988) An order parameter for networks of automata. J Phys A: Math Theor 21:955–960
18. Luque B, Sole R (2000) Lyapunov exponents in random Boolean networks. Physica D 284:33–45
19. Shmulevich I, Kauffman SA (2004) Activities and Sensitivities in Boolean Network Models. Phys Rev Lett 93:048701
20. Mihaljev T, Drossel (2006) Scaling in a general class of critical random Boolean networks. Phys Rev E 74:046101
21. Samuelson B, Troein C (2003) Superpolynomial Growth in the Number of Attractors in Kauffman Networks. Phys Rev Lett 90:098701
22. Kaufman V, Mihaljev T, Drossel B (2005) Scaling in critical random Boolean networks. Phys Rev E 72:046124
23. Krawitz P, Shmulevich I (2007) Basin Entropy in Boolean Network Ensembles. Phys Rev Lett 98:158701
24. Harvey I, Bossomaier T (1997) Time out of joint: Attractors in asynchronous random Boolean networks. In: Husbands P, Harvey I (eds) Fourth European Conference on Artificial Life. MIT Press, Cambridge, pp 67–75
25. Klemm K, Bornholdt S (2005) Stable and unstable attractors in Boolean networks. Phys Rev E 72:055101
26. Greil F, Drossel B (2005) Dynamics of Critical Kauffman Networks under Asynchronous Stochastic Update. Phys Rev Lett 95:048701
27. Albert R, Barabasi AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
28. Aldana M, Cluzel P (2003) A natural class of robust networks. Proc Natl Acad Sci USA 100(15):8710–8714
29. Barrera LO, Ren B (2006) The transcriptional regulatory code of eukaryotic cells – insights from genome-wide analysis of chromatin organization and transcription factor binding. Curr Opin Cell Biol 18(3):291–298
30. Harris SE, Sawhill BK, Wuensche A, Kauffman SA (2002) A Model of Transcriptional Regulatory Networks Based on Biases in the Observed Regulation Rules. Complexity 7(4):23–40
31. Shmulevich I, Lähdesmäki H, Dougherty ER, Astola J, Zhang W (2003) The role of certain Post classes in Boolean network models of genetic networks. Proc Natl Acad Sci USA 100(19):10734–10739
32. Kauffman S, Peterson C, Samuelsson B, Troein C (2004) Genetic networks with canalyzing Boolean rules are always stable. Proc Natl Acad Sci USA 101(49):17102–17107
33. Chaves M, Sontag ED, Albert R (2006) Methods of robustness analysis for Boolean models of gene control networks. IEE Proc Syst Biol 153:154–167
34. Dassow G, Meir E, Munro EM, Odell GM (2000) The segment polarity network is a robust develomental module. Nature 406:188–192
35. Chaves M, Albert R, Sontag D (2005) Robustness and fragility of Boolean models for genetic regulatory networks. J Theor Biol 235:431–449
36. Albert R, Othmer HG (2003) The topology of the regulatory interactions predicts the expression pattern of the Drosophila segment polarity genes. J Theor Biol 223:1–18
37. Faure A, Naldi A, Chaouiya C, Thieffry D (2006) Dynamical anal-

ysis of a generic Boolean model for the control of the mammalian cell cycle. Bioinformatics 22(14):124–131
38. Saez-Rodriguez J, Simeoni L, Lindquist JA, Hemenway R, Bommhardt U, Arndt B, Haus UU, Weismantel R, Gilles ED, Klamt S, Schraven B (2007) A logical model provides insights into T cell receptor signaling. PLoS Comput Biol 3(8):1580–1590
39. Klamt S, Saez-Rodriguez J, Lindquist JA, Simeoni L, Gilles ED (2006) A methodology for the structural and functional analysis of signaling and regulatory networks. BMC Bioinf 7:56
40. Akutsu T, Miyano S, Kuhara S (2000) Inferring qualitative relations in generic networks and in metabolic pathways. Bioinformatics 16(8):727–734
41. Martin S, Zhang Z, Martino A, Faulon JL (2007) Boolean Dynamics of the Genetic Regulatory Networks Inferred from Microarray Time Series Data. Bioinformatics 23(7):866–874
42. Shmulevich I, Dougherty E, Zhang W (2002) From Boolean to Probabilistic Boolean Networks as Models of Genetic Regulatory Networks. Proc IEEE 90(11):1778–1792
43. Lähdesmäki H, Shmulevich I, Yli-Harja O (2003) On Learning Gene Regulatory Networks Under the Boolean Network Model. Mach Learn 52:147–167
44. Shmulevich I (2007) Genomic Signal Processing. Princeton University Press, Princeton
45. Datta A, Pal R, Choudhary A, Dougherty ER (2007) Control Approaches for Probabilistic Gene Regulatory Networks. IEEE Signal Process Mag 24:1:54–63

Bootstrap Percolation

Paolo De Gregorio[1], Aonghus Lawlor[2], Kenneth A. Dawson[2]
[1] Department of Chemistry and Chemical Biology, Cornell University, Ithaca, USA
[2] School of Chemistry and Chemical Biology, University College Dublin, Dublin, Ireland

Article Outline

Glossary
Definition of the Subject
Introduction
Bootstrap Percolation with $m < 2$
Sharp Thresholds: $m \geq d$ on Regular Lattices
Void Instability and Rigorous Results
Corrections to the Asymptotic Behavior in the Presence of Unstable Voids
Bethe Lattice Calculations: The Cases $m \geq 2$
Probabilistic Bootstrap Percolation
High Dimensional Lattices
k-core Percolation on Random Networks
Future Directions
Bibliography

Glossary

Bethe lattice A graph of nodes (sites) each being linked to exactly z other nodes in the graph (z is the coordination number). The Bethe lattice is an infinite object such that there is no boundary and each node has an identical topology. Given any two nodes, there is only one unique set of links connecting them, so there are no closed loops. This is a regular graph with connectivity z. The Cayley tree, also a loopless graph with connectivity z, differs from the Bethe lattice in that a finite fraction of nodes lead to dead ends at a boundary (where they have connectivity 1).

Bootstrap percolation Also known as k-core percolation, is defined as follows. Given a collection of occupied (or active) and vacant (inactive) sites on a lattice, subsets might exist with the following property: Elements in each subset must be occupied and have at least k occupied neighbors, each of those occupied neighbor in turn belonging to the same subset. Bootstrap percolation occurs if any subset exists that spans the system from one end to the other.

Cluster Any subset of occupied sites, each having k or more occupied sites belonging to the same subset. The emergence of a spanning cluster underlies the bootstrap percolation transition.

Culling process Starting from an initial configuration, culling means that occupied sites with less than k neighbors are rendered vacant, i. e. 'culled', and the same operation is iterated a sufficient number of times as to achieve a final configuration with no more sites that are candidates for culling. The culling process is a practical protocol that leads to identify all the relevant subsets (i. e. clusters) in bootstrap percolation.

Lattice A collection of points ('sites') generated from linear combinations with integral coefficients of some basis vectors. A d-dimensional lattice is generated from d basis vectors. Therefore, a d-dimensional lattice is a d-dimensional array or regularly spaced points. A lattice of finite size is such that all the integral coefficients have lower and upper bounds. The coordination number z is the number of sites (called the nearest neighbors) of equal least distance to any given site.

Random networks A network (or graph) of nodes (sites) linked to other randomly chosen nodes. The network is characterized by the distribution of the degree (the number of neighbors of a node). We will discuss bootstrap percolation (or k-core percolation) on simple random, uncorrelated networks with arbitrary degree distribution (we consider always undirected networks).

Regular lattice A regular lattice is a lattice generated from basis vectors that are Cartesian basis vectors, i. e. it is the set of points or a bounded proper subset of points in Z^d. A square lattice is the 2-dimensional regular lattice whose basis vectors are $a(1,0)$ and $a(0,1)$, a simple cubic lattice has basis vectors $a(1,0,0)$, $a(0,1,0)$ and $a(0,0,1)$. A hypercubic lattice is a regular lattice of dimensionality $d > 3$. The coordination number z of a regular lattice equals $2d$. The length a is called the lattice spacing (it is set to unity here for practical purposes).

Definition of the Subject

In bootstrap percolation, we start with a lattice configuration of occupied and vacant sites. Occupied sites that have less than a certain prescribed number of occupied neighbors are rendered vacant, and as new occupied sites are found to satisfy the same condition, these are also rendered vacant. The process is iterated until eventually no more sites can be removed (if any exist). Bootstrap percolation endeavors to determine whether any occupied sites will survive the culling process, and what the macroscopic geometrical properties of the occupied ensembles are. In essence, it is a generalization of conventional percolation which has led to many fruitful insights. A complementary view which emphasizes the dynamical aspect of the bootstrap process, treats the vacant sites as invasive units and occupied sites as inert ones. Inert sites that are surrounded by too many invasive sites will become irreversibly infected and will start themselves to invade others. At the end, the question arises as to whether the 'infection' has been contained or it has been pervasive. Bootstrap percolation was first introduced in the late 1970s as a simple model to study the properties of some magnetic materials [20,43,56]. One application that it was proposed to model was crystal field interactions, which tend to suppress magnetic moments by forcing the atoms into their singlet state- they are then in competition with exchange interactions that favor the alignment of the magnetic moments of neighboring atoms. For the material to display macroscopic magnetic properties, exchange interactions must be everywhere locally dominant. Later it became apparent that the bootstrap percolation model is quite accurate in describing certain (low temperature) properties related to quadrupolar orientational ordering of molecular solids [5], for example when both quadrupolar and isotropic molecules are present. A paradigmatic example is solid molecular $(\text{ortho-H}_2)_x(\text{para-H2})_{1-x}$, or $(\text{para-D2})_x(\text{ortho-D2})_{1-x}$ [1,5]. More recently bootstrap percolation, also called k-core percolation, has seen a resurgence of interest in connection to simple kinetic models for the glass transition [29,39,42], and as a model system for jamming of both hard-sphere and soft-spheres molecules at high packing densities [63]. Other applications have been suggested in the study of porous fluid flow through cracked rocks [2], disease spreading [10], hysteresis [58], computer storage arrays resilience to failure [41], sand-pile formation [47], spreading of alerts on distributed networks [69] and many other problems for which standard nearest-neighbor percolation has already been proposed as a model system [66]. More generally, bootstrap percolation is a useful model to address certain physical properties of a system whose local interactions are strongly contingent on the relative states of neighboring units, whereby a prevailing driving 'force' only arises when many units act cooperatively 'in phase'.

Introduction

In the decades that preceded the introduction of bootstrap (or k-core) percolation models, a new paradigm in Statistical Physics had revolutionized our understanding of complex physical systems, especially in the presence of critical behavior and in its vicinity. From Ising onward [38], simple (often lattice) models became an increasingly popular way to treat highly complex problems, analytically and later by means of computer simulations. In some rare cases highly non-trivial exact solutions [53] provided the first hint that (already in two dimensions) one would discover important connections with real critical behavior of continuous systems. Quite unexpectedly, given the simplicity of lattice models, it was later demonstrated that in the proximity of critical points a surprising variety of important and correct general results can be derived all within the framework of renormalization group calculations for lattice models [75]. It emerged that in most cases dimensionality is the only relevant variable in play for systems belonging to the same universality classes. Thus, to rely on the use of lattice models was found to be much less problematic than one would have anticipated. The universality of behavior emerging from the study of criticality for simple lattice models was understood to apply under some very general conditions about the microscopic forces, as with short ranged interactions and if contributions from multi-body interactions were unimportant. Conventional (site or bond) percolation fits perfectly in this picture [65,66], providing a way to investigate further the geometrical properties of systems with critical behavior that are characterized by the same universal exponents as those of critical phenomena. On the other hand, it rep-

resented a confirmation of the rather surprising ubiquitousness of scaling [74] universality classes.

Bootstrap percolation provides an extremely useful platform to investigate other classes of phenomena. Can we construct a simple percolation model that produces other types of universality classes? In percolation theory, models that deal with exceptions to equilibrium continuous phase transition critical behavior are known to exist. Oriented and rigidity percolation are just some examples [59]. Bootstrap percolation is a model which has the flexibility to explore an ample spectrum of conditions over the local 'interactions' while still facilitating in many cases, through its simplicity, the development of rigorous mathematical results [7,18,19,35,36,62,70,72,73]. The underlying 'interactions' can be thought of as strictly local. However, over larger scales and unlike with critical phenomena, the results are strongly dependent on the specifics of the interactions, scaling can be sometimes problematic, and the outcomes can be quite diverse. Concurrently, bootstrap percolation can be viewed as a cellular automaton where basic dynamical properties of out-of-equilibrium or slowly evolving systems can be studied [2,27,39].

Let us make one concrete example. To set the discussion into context, we can think of some of the several lattice models that have recently been proposed to describe some dynamical properties of glass-forming systems [29,39,42], or jamming scenarios [63]. These models can be of two kinds [57], either lattice gas models in which the particle number is conserved [27,39,42], or spin models of the Ising type [29] in which some simple external field drives the equilibrium configurations to have low mobility at low temperatures. The distinction is not extremely important, similarly to the situation of Ising-like models for ferromagnetic transitions versus lattice models for liquid-vapor transitions. In any case, it is postulated that some particles can only move (or spins can flip) if the number of occupied neighbors (or of up-spins) does not exceed a given threshold number. Conversely, one may say that the dynamics is facilitated if a sufficient number of vacant sites (down-spins) exist in the neighborhood of a given particle to support the required density transformation. The outcome is clear. The dynamics are constrained locally in all-or-nothing possibilities, which depend on concentration of vacant (or mobile) space around. If this vacant space is not sufficient, the dynamics does simply not progress. As a consequence collective motion can become heterogeneous and dramatically slowed [29,39,45,57].

Having such a picture in mind, we can immediately recognize the connections with bootstrap percolation (BP). In BP, the sites of the lattice can be either occupied with probability p, or vacant with probability $q = 1 - p$. Usually, multiple occupancies are not admitted in such models, and the initial probability for occupied sites is entirely random, but such conditions are of a practical nature, and variants are admitted [47]. We can refer to p as either the concentration, or particle density. The linear system size of the lattice is taken to be L. We then define a number $1 \leq k \leq z$, z being the coordination number. For convenience, let us also define $m = k - 1$ to make contact with several alternative definitions of the same parameter that are present in the literature. Then a culling process is carried out. All vacant sites as well as all sites that have k or more occupied sites among the nearest neighbors (the 'constrained' sites) are left unaltered, while all sites that have m or less occupied neighbors among their nearest neighbors are removed. Step by step, many particles are removed and the concentration of particles gradually decreases. As new space is emptied, other particles previously stable can be removed. We eventually reach a stationary situation.

There can be different scenarios for the stationary state at the end of the culling process. There can be either no particles at all, or there can be clusters of particles, one of which will eventually span the entire system (we shall see that in some cases there can be no finite-sized clusters, and the only two possibilities are that there be either no particles at all or a giant spanning cluster). The outcome depends on p, k, the lattice geometry and on the system size L. For example, for not too small L, $k > 2$ and for small values of p we can expect that all particles will be eventually removed. A systematic study can be carried out using a similar procedure to that of conventional percolation. Given the system size L, the culling process is carried out at some low value of p. It is expected that there will not be spanning clusters in the final configurations. p is gradually incremented, until at a certain value $p_c(L)$ that depends on L, 50% of the final configurations will have a spanning cluster of irremovable particles. By changing the system size L we can construct the relation $p_c(L)$ between the bootstrap percolation finite-size critical concentration and the system size. For an infinitely extended system we define $p_c = p_c(\infty)$ as the ('thermodynamic') critical fraction. Alternatively, one can fix p and let the system size vary from small to large L and identify the critical size L_c, known as the bootstrap percolation length- this is the size of system for a given p for which 50% of the configurations lead to a large spanning cluster in the final state. Then p is varied so that we can construct the relation $L_c(p)$ between p and the corresponding bootstrap percolation critical size L_c, which is of course just the inverse of $p_c(L)$. Other quantities of interest are the same of conven-

tional percolation, such as the concentration of occupied sites in the clusters surviving the culling process, or the mass factor associated to the spanning cluster.

$p_c(L)$ and $L_c(p)$ can be clearly evaluated in a systematic fashion from computer simulations. Analytically there are radically different techniques that can be used to find them, depending on the value of k and the lattice geometry. For low k the bootstrap percolation problem is essentially equivalent to conventional percolation [66], and therefore real-space renormalization group calculations are the obvious choice. As we will see, there typically is some value of k, call it k_{uv}, at which 'unstable voids' [1,2,3,7,23,43,70,72,73] emerge for any choice of $k \geq k_{uv}$. More precisely, while for $k < k_{uv}$ clusters of either finite size or spanning the system can survive the culling procedure, for any $k \geq k_{uv}$ self-sustaining (finite) clusters cannot survive the decimation process. This is typically due to the presence of voids that can become unstable when $k \geq k_{uv}$, i. e. are cores of empty sites of certain critical sizes that with high probability can nucleate and lead to a complete decimation of the particles in the system. For such models sophisticated analytical techniques have been devised and developed [7,11,18,19,60,62], in some cases leading to explicit rigorous mathematical results [35,36] or exact interpolations and predictions [23]. A third route is that of 'mean-field' calculations for Cayley trees (Bethe lattices) [20,34,49,63,64]. These have been investigated thoroughly in recent years especially because there can be values of the parameter k in the window between those values of k that can be treated with usual renormalization group techniques as in conventional percolation, and the value k_{uv} at which unstable voids are expected to arise. Some very interesting results have emerged already, in combination with simulations [63] and some generalizations of bootstrap percolation models [40,67,68]. However, there still are some controversies that will have to be clarified [54]. We expect that many of these issues will become clearer in the near future, possibly with new approaches [34] that will not rely entirely on mean-field calculations [20]. Finally, one obvious extension of bootstrap percolation is to random networks [25,30], which we discuss in Sect. "*k*-core Percolation on Random Networks" – this approach is already attracting considerable interest.

Bootstrap Percolation with $m < 2$

The case $k = 1 (m = 0)$ is trivially the same problem as conventional percolation [66], in any dimension and for any lattice geometry. In fact, only isolated occupied sites are removed, which in any case do not participate to the construction of percolation clusters in conventional nearest-neighbor percolation. It is then straightforward to recognize that, on approach to criticality, the critical fraction $p_c(L)$ is related to the system size and $p_c = p_c(\infty)$ via the relation,

$$|p_c(L) - p_c| = AL^{-1/\nu}[1 + \epsilon(L)] \qquad \lim_{L\to\infty} \epsilon(L) = 0 \tag{1}$$

while the total fraction of sites belonging to the infinite cluster vanishes with p approaching p_c from above as,

$$P_\infty \propto (p - p_c)^\beta \tag{2}$$

The exponents ν and β can be derived within real-space renormalization group calculations, or their relations derived from scaling theory, and estimated from computer simulations. They are therefore only dependent on the dimensionality d and not on the geometry of the lattice. All other exponents that are related to the mass fraction or to the percolation correlation length are dependent on the previous exponents and can be derived from those. Their values are given in this volume in the sections devoted to standard percolation.

In the case $k = 2 (m = 1)$, only dangling ends are removed from the clusters of any size. It follows that this has no effect on the relation given in Eq. (1), representing correctly the relation between the concentration of particles and the size of the spanning cluster near the percolation transition. In regard to the other exponents relating the system size to the mass fraction or the probability that a site belongs to the spanning cluster, although not apparent from usual system sizes [15], it is found that the exponent β is also universal [16,21,48]. That is, β coincides with β from standard percolation. This becomes only evident if the system sizes are very large [21,48], and it is supported by general arguments [16]. It is noticeable that scaling arguments apply appropriately to both $k = 1$ and $k = 2$, and one has the usual picture of continuous phase transitions.

Sharp Thresholds: $m \geq d$ on Regular Lattices

The case that perhaps historically has attracted most interest in bootstrap percolation is when we pass from a picture of the type well described by the theory of critical points to one where 'first-order' type transitions emerge. Typically, the precursor of such transitions is a very sharp threshold in the concentration-dependent probability to percolate, which is driven by the indirect effects of the so called 'unstable voids' [1,3,7,22,23,27,39,43,70,72,73]. These are special objects whose interior is void of particles. They have

the potential capacity to eat away all the particles in the system, provided that they manage to grow beyond a certain threshold size. We can say that if their spreading is not contained at an early time, they will almost certainly end up 'invading' or 'infecting' the entire system. Whether there is any such object in the system depends of course on the initial concentration of occupied and vacant sites, and on the system size (if the system is finite in size). Their properties will be discussed extensively in this section. One interesting aspect of these models is that their studies have been characterized by an intriguing series of twists and turns, before a clearer and more thorough picture would finally emerge.

The simplest case in which dominance of unstable voids can emerge is when $m \geq d$ ($k \geq d+1$) on regular lattices. Therefore, $m \geq 2$ for the square lattice, $m \geq 3$ for the simple cubic. However, other geometries can also also lead to the same situation, and d might not always be the relevant threshold value for separating the different classes. For example, the same picture holds for the triangular lattice (the 2-dimensional lattice with coordination number $z = 6$), whereby unstable voids will emerge if $m \geq 3$. Notice that for the triangular lattice, the threshold number for this class is 3 and it also also coincides with $z/2$, as it does for the square lattice, and in general for the hypercubic lattice.

First, we shall be discussing informally some general preliminary properties in the case $m = d$ for regular lattices. Next, some of the known results will be outlined. Finally some of them will be discussed more extensively with some derivations.

Spanning Clusters: $m = d$

Let us take the case $m = 2$ on the square lattice (i. e. $k = m + 1 = 3$). This implies that only particles with less than 3 occupied neighbors can be removed. To start with, it is useful in general to know whether clusters of finite size (i. e. non-spanning) might survive the decimation process in the box of area L^2. Let us suppose that one such cluster exists. By definition, if it is finite, the cluster must have a boundary. But if it has a boundary, this implies that there exists a non-empty set of particles in such cluster that have at least two vacant sites as their nearest neighbors. This set cannot be empty because any finite cluster is a compact object (you may think that it has at least four 'corners', or vertices). Particles in such set are therefore removed at the next iteration of the culling process. But now we have again a finite cluster with a non-empty set of 'corners'. It follows that the removal procedure cannot be stopped and the cluster will eventually be washed away. We conclude that the only possibility for a cluster to survive the culling process is that it is not self-sustaining, i. e. only spanning clusters that are connected to themselves through the lattice boundary lines may exist at the end. Similar conclusions can be derived for any dimension for the simple cubic and hypercubic lattices.

It should start to transpire already why these models are special. Given a certain system size L and some value of p, after culling terminates only two situations can happen. Either the lattice is empty, or it contains one large (massive) spanning cluster, and we collect the statistics for many equivalent realizations at L and p. We cannot, as in the case of critical points in conventional percolation, identify a broad distribution of finite-sized clusters of all lengths, whereby each sampled configuration (whether percolating or not) would be similarly distributed in any element from the wide samples. Furthermore, from the above argument it is also clear that all the 'roughness' over the boundaries of any set of connected particles in the spanning cluster is washed away during the culling process. In other words, we should expect its fractal dimensionality to be the same as the volume it is embedded in. All these qualitative considerations suggest a scenario that is very typical of first order phase transitions rather than that of continuous ones. More careful examinations are consistent with this picture. These models are without doubt a departure from standard percolation problems, and also from bootstrap percolation models with small values of m.

In Fig. 1, we see an example of the probability of a spanning cluster in the final configuration, plotted as a function of particle concentration p and for different system sizes (from theory and for a modified model that is qualitatively equivalent to the two dimensional $m = 2$ square lattice model). As we can see, the 'transition' in the probability, from essentially zero to some finite value, occurs over a very narrow range, and it displays an important size dependence. At large system sizes the sharp change is 'almost' discontinuous. Furthermore as said, whenever a spanning cluster survives the culling process, the fraction of occupied sites in it is typically finite (i. e. non-fractal), leading to a similar sharpness in the probability that any given site anywhere in the lattice belongs to the spanning cluster. The analysis of similar results in two and three dimension from computer simulations (which are limited as to the maximum system size available) led initially to the conjecture of a power-law behavior [43] for $d = 3$, in the form of Eq. (1) but with a critical concentration lower than that of the discontinuous jump, and it was not excluded that the critical concentration for an infinitely extended system would have been $p_c < 1$. This

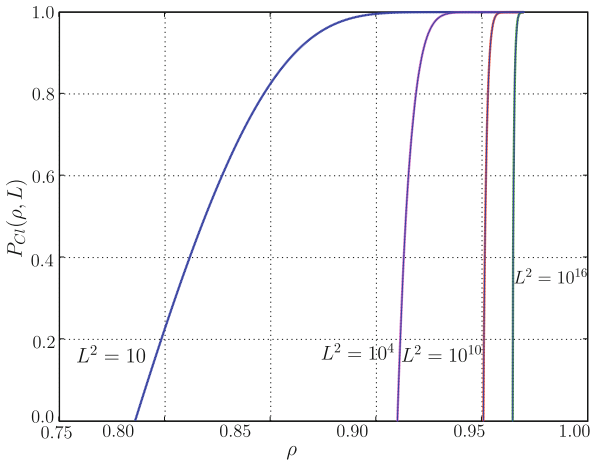

Bootstrap Percolation, Figure 1
The typical probability P_{Cl} that a spanning cluster survives the bootstrap process as a function of particle density and for different system sizes. Two dimensional modified model in the example. From *left* to *right*, system sizes correspond to $L^2 = 10$, $L^2 = 10^4$, $L^2 = 10^{10}$ and $L^2 = 10^{16}$

premature conjecture was only partially and very qualitatively consistent with the mean-field Bethe lattice behavior [20] studied theoretically (this topic will be discussed in another section). However, it was later rigorously proven that $p_c = p_c(\infty) = 1$ [7,62,72,73]. More precisely, as we will see, the probability that an unstable void develops is always small but strictly positive, for any initial value of $p < 1$ and in any dimension for regular lattices, whenever $m \geq d$. Therefore, if we suppose that the lattice is infinitely extended, we will encounter at least one unstable void with certainty.

Void Instability and Rigorous Results

In recent years more and more rigorous, theoretical and simulation results have led to an an impressively detailed understanding of these models [1,3,4,6,7,11,18,19,22,23,35,36,37,44,62,72]. Before stating some of these results we first define, for convenience, a variant model that has been described in great detail. It is called the modified bootstrap percolation model [22,23,35,36,62]. For any regular lattice, any particle can be removed if and only if in any of the coordinate directions, it is neighbored by *at least* one vacant site. This condition is similar to the case $m = d = z/2$, but with an additional requirement on how the vacant sites must be distributed around the particle (so to avoid too uneven distributions of voids). The set of particles that can be removed in modified bootstrap is a subset of those that can be removed for the standard (conventional) $m = d$ case.

We can now look into some of the known rigorous results. They are listed here as relations between the threshold concentration for the first appearance of a spanning cluster, as functions of the system size L. This is the commonest way to visualize quickly the types of finite size corrections. However, later we will discuss why for other purposes not necessarily this is the most convenient form. The inverted form of these relations will be discussed later. In all the following, $\epsilon(L)$ must be intended as an unspecified, *non-universal* (positive or negative) function that is vanishing in the limit $\lim_{L\to\infty} .B(L)$ is some positive term which is certainly bounded by two constants for sufficiently large L. We have for all regular lattices from square to hypercube,

$$p_c(L) = 1 - \frac{\pi^2}{18 \ln L}[1 + \epsilon(L)] \qquad d = 2\,,\ \ m = 2 \quad [39] \tag{3}$$

$$= 1 - \frac{\pi^2}{6 \ln L}[1 + \epsilon(L)] \qquad d = 2\,,\ \ \text{modified} \ [39] \tag{4}$$

$$= 1 - \frac{B(L)}{\ln \ln L} \qquad d = 3\,,\ \ m = 3 \quad [7, 30] \tag{5}$$

$$= 1 - \frac{\pi^2}{6 \ln \ln L}[1 + \epsilon(L)] \qquad d = 3 \ \text{modified} \ [40] \tag{6}$$

$$= 1 - \frac{B(L)}{\ln^{o(d-1)} L} \qquad \text{any} \ d \geq 2\,,\ \ m = d \quad [19] \tag{7}$$

$$= 1 - \frac{\pi^2}{6 \ln^{o(d-1)} L}[1 + \epsilon(L)] \qquad \text{any} \ d \geq 2\,,\ \ \text{modified} \ [40] \tag{8}$$

$$= 1 - \frac{B(L)}{(\ln L)^{d-1}} \qquad \text{any} \ d \geq 2\,,\ \ m = z - 2 \quad [7] \tag{9}$$

$$= 1 - \frac{B(L)}{(\ln^{o(2d-m-1)} L)^{m-d+1}} \qquad \text{any} \ d \geq 2\,, d \leq m \leq 2d - 2 \quad [19, 29] \tag{10}$$

As usual, z is the coordination number, $z = 2d$ in this case. The expression $\ln^{o(n)}$ stands for the natural logarithm iterated n times. The explicit form of $\epsilon(L)$ is expected to

vary considerably case by case, nevertheless it vanishes for large L, but it is unknown rigorously. From expressions from Eq. (3) to Eq. (10) we readily recognize the importance of finite-size effects. To see this, take for example the general expression for $m = d$ outlined in Eq. (7), which is in fact the case of major interest. The leading term in the expression is the inverse of the logarithm iterated $d-1$ times, which converges to zero extremely slowly, the higher the dimension the slower it converges. This is very different from standard power law behavior, and explains why it was not immediately recognized that these systems have a critical concentration of 1. As it transpires, we have a whole new 'universality class', peculiar to bootstrap percolation. It is especially noticeable that in many cases, the asymptotic expressions are such that the multiplicative constant is known exactly. Another issue that may not appear so obvious at first is that the forms of $\epsilon(L)$ are not rigorously known. Due to the slow convergence of the leading terms, $\epsilon(L)$ may add substantial finite-size corrections that are in addition to the iterated logarithmic part. In fact, in principle $\epsilon(L)$ in itself may converge to zero even slower than inverse logarithms, and it may be comparable to unity up until the largest sizes that are physically interesting. As we are going to see, already in two dimensions this effect is not at all negligible.

As we shall see, a more natural choice for these models is to express the critical bootstrap length in terms of the initial particle concentration, or rather of the initial concentration of empty sites $q = 1 - p$. There is a unique transformation from $q = q(L)$ to the inverse relation $L = L(q)$ that relates critical concentrations to critical sizes and viceversa. For example, we can preliminarily rewrite Eq. (7) as expressing the asymptotic result $L_c(q) \sim \exp^{o(d-1)}(A/q)$, i. e., an iterated exponential whose argument is of order $O(1/q)$. Therefore in $d = 2$ we have

$$L_c(q) \sim \exp\left(\frac{A}{q}\right) \tag{11}$$

expressing the exact dramatic divergence of the critical size when the fully filled lattice limit is approached. It is useful however to formally invert Eqs. (3) and (4) by separating out the asymptotic contribution from the corrective term, and by defining an unknown function $f(q)$ related to $\epsilon(L)$. We can do this by writing

$$\ln L = \frac{A}{q} + f(q)$$

and observing that

$$\epsilon(L) = \frac{q f(q)}{A} \tag{12}$$

must hold. Notice that $A > 0$ and therefore $f(q)$ and $\epsilon(L)$ are of equal sign. From Eqs. (3) and (4) and the properties of $\epsilon(L)$, in the limit $L \to \infty$ both $q \to 0$ and $\epsilon(L) \to 0$, and therefore,

$$L_c(q) = \exp\left[\frac{A}{q} + f(q)\right] \tag{13}$$

$$\lim_{q\to 0} q\, f(q) = 0 \tag{14}$$

As we can see, $f(q)$ may in principle be a divergent function with vanishing q, although it must always diverge slower than $1/q$. Indeed, the observed critical bootstrap percolation size is often orders of magnitude off from the expected leading asymptotic behavior [1,3,22,35]. This subtlety will be briefly discussed later.

How can we derive Eq. (11) in $d = 2$ and the corresponding iterated exponentials for $d > 2$? First, we expand more explicitly on the concept of unstable voids (or critical droplets). To illustrate an elementary example of what an unstable void is, take the trivial case $m = z - 1$, that is, all occupied sites with at least one vacant neighbor are removed. Take a large lattice with a certain distribution of particles $p = 1 - q < 1$, and p very close to 1 (i. e. q very close to zero). If $L^d \gg 1/q$, we can be certain that there would be some empty sites in the system (despite q being small). According to the rule assigned, such voids will gradually expand and eat away all the particles in the system, without any barrier that may stop them from growing. Isolated voids are therefore always 'unstable' when $m = z - 1$, they start to grow immediately as soon as the culling process is started, and they cannot be stopped. Furthermore, if starting from the same initial configurations we subdivide the system into many subsystems of volume L' such that $L'^d \ll 1/q \ll L^d$, we end up with many separated subsystems entirely filled, and a very few containing one (or more) vacant sites. After the culling process is implemented separately on each one of the subsystems, a tiny fraction of the total number of them (those with vacant sites, i. e. unstable voids) are completely empty. All the others are entirely filled by particles. We thus see how in the presence of unstable voids, for a given value of $p = 1 - q$ we can have different scenarios, depending on the system size, while at the same time for an ideal system that is infinitely large, any $p < 1$ leads to all space to be empty after bootstrap.

The situation is of course much more complex in the more interesting cases, but some basic principles are unaltered (in a crude sense). Let us consider the case of the square lattice when $m = 2$, or consider the modified

model (they behave similarly). Imagine that we have created somewhere a rectangle whose inside is completely empty (note that it is often said that such rectangle is internally spanned, but we must not confuse this notion with that of a spanning cluster of percolating particles). In Fig. 2 we have a square as an example. On the boundary lines of its perimeter there can be any number of particles. But let us see what can happen if some sites on those lines are vacant. Due the rules for the removal of particles, any vacant site on any one of those lines is a one dimensional unstable void limited to that line, because each site on the line has already one vacant neighboring site which lies in the interior of the empty square. This means that if on any of the boundary segments there is one vacant site or there are more, that segment will be vacated. If the condition is satisfied in all four directions, the empty square can engulf its boundary and become 2 linear units larger

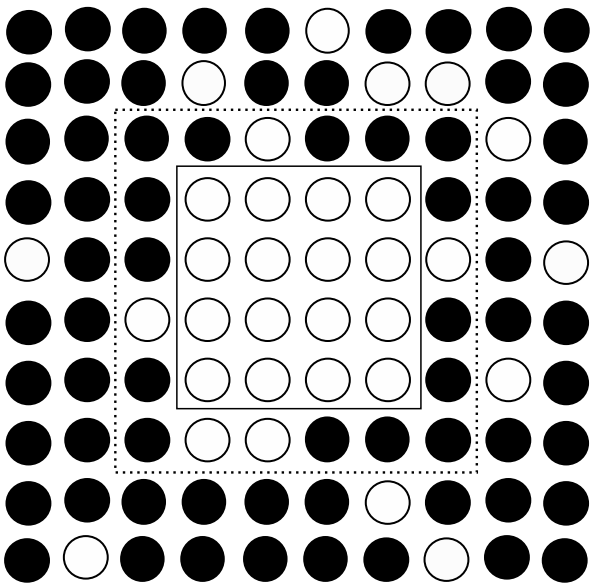

Bootstrap Percolation, Figure 2
Unstable voids in the square lattice. The internal 4 × 4 square delimited by the solid line has been vacated previously. Both in the modified and in the conventional $m = 2$ model, the square can engulf its four next sides leading to the vacation of the 6 × 6 square delimited by the *dotted line*. From there, the growth by squares in the modified model would be stopped because of the impossibility to proceed in the West direction, while the growth by squares is still allowed in the conventional model up to the vacation of the largest 10 × 10 square in the figure. Finally, allowing for the growth by elongated rectangles in the modified model, the void can momentarily expand in the North, East and South direction before all particles in the East direction can be vacated too. Differently from the squares-only process, this time the largest 10 × 10 square in the figure will be vacant at the end, also in the modified model

per side. Again, let us consider $p < 1$ but very close to 1, i. e. q positive but very small. For small squares with side length much smaller than $1/q$ it is very unlikely that any such empty square will grow further. However, for side lengths much larger than $1/q$ the opposite is true, it is almost certain that the square will grow to the next step, and to the steps after (as the side length becomes larger, it is increasingly likely that the boundary lines will be unstable with respect to the internal empty square). Voids of the first type are therefore stable, voids of the second type are expected to be unstable. Therefore when we turn our attention back to the entire lattice of size L, we can be certain that all particles will be removed provided that L is large enough so that the lattice contains one or more unstable voids (and so the lattice is sometimes said to be internally spanned by the voids. Again such terminology must not be confused here with spanning clusters of particles). In order to determine whether voids can ever become unstable even for very small q, we must determine rigorously some bounds for their probability to arise.

To quantify this argument we ask what is the probability that a vacant site can nucleate and grow indefinitely? By letting this process proceed via the formation of larger and larger squares we certainly obtain a lower bound to the nucleation probability (it is a lower bound because, more generally, nucleation can also proceed via rectangular voids of various elongations [23]). Given an empty square of side length l, the probability that any of the four external boundary lines can be vacated coincides the probability of there being at least one vacant site, which is formally given by $1 - p^l$ (it is the complement of there being l particles, which is p^l from the random measure distribution on p). The joint probability for all four sides *at the same time* is $(1 - p^l)^4$. Let us call the probability of indefinite nucleation ν, and $\nu_b < \nu$ is some lower bound. To evaluate this bound, we may start with a square of size 2×2 as the nucleus. We may write,

$$\nu_b = q^4 \prod_{k=1}^{\infty} \left(1 - p^{2k}\right)^4 \tag{15}$$

Here, k is the side length of the growing square (which is an even number at each step). Therefore,

$$\ln \nu_b = 4 \ln q + 4 \sum_{k=1}^{\infty} \ln \left(1 - p^{2k}\right) \tag{16}$$

To bound the sum we first expand the logarithm in infinite series and we exchange the order of summation. We then

use the inequality $(1-p^{2k})/p^{2k} > 2k(1-p) = 2kq$.

$$\sum_{k=1}^{\infty} \ln\left(1-p^{2k}\right) = -\sum_{k=1}^{\infty} \frac{1}{k}\frac{p^{2k}}{1-p^{2k}} > -\frac{1}{2q}\sum_{k=1}^{\infty}\frac{1}{k^2} = -\frac{\pi^2}{12q} \tag{17}$$

The last equality comes from Euler's sum which equals $\pi^2/6$ (one can also show that when $q \ll 1$ the inequality above is essentially an equality but for unimportant corrective terms). From (16) and (17) we therefore have a new bound for the overall nucleation probability, meaning,

$$\nu > q^4 \exp\left(-\frac{\pi^2}{3q}\right) \tag{18}$$

Equation (18) shows that the nucleation probability, i. e. the probability that unstable voids exist, is a strictly positive quantity for any $q > 0$. This implies that for any system of size $L^2\nu_b \gg 1$, i. e. $L \gg q^{-2}\exp(\pi^2/6q)$, we can be almost certain that no particle can survive decimation. As $q \to 0$ ($p \to 1$), the required L increases dramatically. Perhaps even more surprising is if we look at the latter expression, and compare it with Eq. (4), valid for the modified model. The two are fully consistent, with the correct constant. This is surprising because we started with a bounding argument and, apparently, we have ended up with the correct estimation of the asymptotic expression (at vanishing q) for ν and L_c. The proof that in fact the bound described above does furnish the correct asymptotics for both ν and L_c is highly non-trivial [35], and despite it being a relatively recent result it is already a cornerstone of bootstrap percolation. One needs to calculate an upper bound which is, to the leading asymptotic order, coincident with the lower bound. In regard to the conventional (non modified) $m = 2$ case, the argument employed for bounding ν starting from the probability of growing square-like unstable voids can be refined. In that case, even if one boundary line is entirely filled with particles, one or more vacancies present in the adjacent line can also lead to all particles in both lines to be vacated (this is seen with the help of Fig. 2). This necessitates some technical adjustments in the calculation above in Eq. (15), but the rest of the argument remains unchanged [35]. One gets a new bound for the square-like process that is precisely in the form (18), but with a new constant given by $\pi^2/9$. Again, by means of the identification $L_c^2 \sim \nu^{-1}$ for the critical size (part of the demonstration [35] shows that to be licit at $q \ll 1$, and simulations are consistent already at not too low q [22]), we have an expression consistent with Eq. (3). The complete proof for the upper bound is essentially equivalent to that for the modified model. In general, the modified model renders some of the calculations more straightforward, but the two models are in essence very similar.

We thus learn that Eq. (11) (or more appropriately (13)–(14)) is a rigorous result, with the calculated constants A equal to $\pi^2/6$ and $\pi^2/18$ for the modified and conventional $m = 2$ case on the square lattice, respectively. It would appear that we have at last gained a correct and complete picture, with exact rigorous mathematical results [35] that we can finally compare with computer simulations [44] with a sufficient degree of confidence. However, this is not quite as true yet [3,22]. In reality, fits that use Eq. (11) are typically consistent with estimated constants that are sensibly different [3,6] from the ones calculated within the theory (by factors as large as 2 or 3). This puzzle will be discussed in the next section. It has to do, as earlier pointed out briefly, with the more general expression in (13)–(14) and its subtle implications.

As we have seen, there is a non-zero probability to encounter unstable voids at any $q > 0$ ($p < 1$). When $q \ll 1$, voids can become unstable once they reach a side length much larger than of order $\bar{k} \sim O(1/q)$ (a more rigorous estimate lends $\bar{k} \sim O(\ln q/\ln p)$). To appreciate this statement qualitatively, we observe that the probability $1-p^k$ that a boundary line can be wiped out by the growing void is of order $1-\exp(-b) \sim 1$ if $k = b/q$, $b \gg 1$ and $q \ll 1$ (notice that the latter also implies $q \approx -\ln p$). Consequently, if on the one hand for $k \ll O(1/q)$ the voids are stable with high probability, on the other hand for $k \gg O(1/q)$ they are unstable with high probability. The length $\bar{k} = 1/q$ is a novel length not present in standard percolation models with critical points (see Fig. 3). We see that contrary to those models, we might be in serious trouble if we were to assume that we can freely re-scale the lattice spacing by any finite factor (like in a self-similar spin-block picture of the type of Kadanoff), just as long as the scaling length is much smaller than the longest length. Here instead we have the larger length-scale $L_c(q)$ which is always sensibly much greater than the length-scale $\bar{k}(q)$, for small q. On the other hand, on a practical scale, this feature can be of tremendous advantage. Indeed, it is sometimes sufficient to determine the properties of some physical quantities (e. g. correlations) up to within the distance $\bar{k}$ and no further, when $L_c \gg \bar{k}$. A good example is when we want to evaluate the nucleation probability ν. Rather than sample systematically systems of very large sizes, it is sufficient to determine whether some nuclei can grow further than $\bar{k}$ [22], and what the associated probability is for a given concentration p. This saves an enormous amount

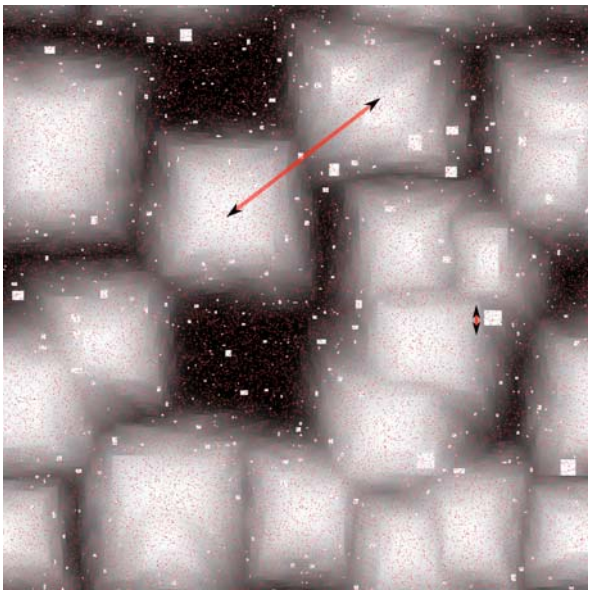

Bootstrap Percolation, Figure 3
Evolution of the bootstrap process for $k = 3$ in $2D$ ($p = 0.952$, $L = 900$). The time at which particles are removed is depicted by the color- the first ones removed are *white* and the last ones removed are *black* (initial vacancies are *red*). It is seen that the bootstrap removal process starts at nucleation centres, also called connected holes. The *larger scale* indicated in the figure is the bootstrap length $L = 1/\sqrt{\nu}$, or the typical distance between connected holes. The *shorter scale*, $\bar{k} \sim 1/(1-p)$, is the critical length above which voids will become unstable

of computer running time for values of p that correspond to very large critical bootstrap lengths.

More articulated mathematical analysis lead to bounds and exact results in higher dimensions [18,19,36]. We can consider a qualitative argument to determine the bootstrap critical system size in more than 2 dimensions. As an example, let us consider $d = 3 = m$ (also modified model). We need to establish whether voids can be unstable in this case, and this is indeed the case. Take for instance [72,73] some large cube all empty in the inside, and whose side length is of some value k. We look for a condition such that the empty cube can grow in a given direction by one step, thus absorbing any particle that might eventually lie within the boundary surfaces. In the simple case, we consider one square-like boundary surface per side at a time. The sites on one single surface form a $k \times k$ square lattice, but additionally each site is also connected to one site that lies on the interior of the cube (which is certainly empty), and one on its exterior. Thus any site within such $k \times k$ square lattice that forms the boundary surface of the cube has at least one neighboring vacant site (because the cube is empty), and only needs two more vacant neighbors to be removed. A sufficient condition is that the two required vacant sites are among the 4 neighbors that lie on the $k \times k$ square lattice itself, irrespective of the occupation state of any site from the exterior. We thus realize that if locally the boundary $k \times k$ square lattice has at least one unstable void in the $m = 2$ sense, then the entire square can be vacated. This can be repeated for any of the six sides that form the entire boundary of the cube. It follows that we have reduced the treatment of the one-step growth for the empty cube for $m = 3$ to six equivalent $m = 2$ subproblems on its surface. A problem that we already know how to solve nevertheless. For example at $q \ll 1$, we know that if the side length k is of order $\exp[O(1/q)]$, then with high probability all particles within the square-shaped surfaces are likely to be removed. But this means that the empty cube will almost certainly grow by one step, and the same will happen at the next step. We have therefore established that, very approximately, we must expect the new $\bar{k}$ to equal $\exp[O(1/q)]$ in the present case (much larger than $O(1/q)$ as for the square lattice).

All that is left is to evaluate the probability that such unstable void of the required length arises somewhere. In principle we would need an iterative formula of the type of Eq. (15), but it will suffice to determine a qualitative bound for that probability. A minimal condition for a cubical region at $q \ll 1$ to be unstable is that It must contain $\exp[O(1/q)]$ empty sites at least, in some appropriate positions. For example, if the 6 faces that constitute the frame of a cube are all empty, one can prove that the interior of the cube will be eaten away gradually, leading necessarily to an entirely empty cube of side length $\exp[O(1/q)]$, which is unstable for the argument above. This can happen with probability

$$q^{\exp[O(1/q)]} = \exp\{-\exp[O(1/q)]\ln(1/q)\}\,.$$

We then apply the usual condition that L_c^d multiplied by such probability be $O(1)$, and we get

$$L_c \sim \exp\{\exp[O(1/q)]\ln(1/q)\}\,.$$

By taking twice the logarithm we first neglect $\ln\ln(1/q)$ compared to $O(1/q)$ and we then obtain

$$q \sim \frac{1}{\ln\ln L_c}$$

which is the desired result as in Eqs. (5)–(6). Rigorous results for upper bounds require other elaborate treatments, not to be mentioned here. However, as above, in most cases a convenient scheme for the proofs requires one to employ some dimensional induction from $d-1$ (supposedly known) to d [19]. It is worth noticing that an (elaborate) induction argument on dimensionality also allows

one to derive Eq. (6) from Eq. (4) [36]. Similar arguments (alongside with their sophisticated rigorous counterparts) apply to dimensions higher than 3, giving rise to the iterated exponential (for $L_c(q)$) and iterated logarithmic forms (for $q_c(L)$) present in Eq. (7).

A word of caution however is in order here. The results so far have been illustrated assuming that d is kept fixed. From Eq. (7), which is valid for $m = d$ on the hypercubic lattice, we see that every time we raise d, the required L for attaining an arbitrarily small $q_c = 1 - p_c$ must increase, because each time we must add another iteration of the logarithmic function in the denominator. What happens then if we let d vary at the same time as varying L? One notices that in reality Eq. (7) holds not only when d is fixed, but also when it is some extremely slowly growing function of L [13,19]. And so we ask if it is always guaranteed that q_c becomes arbitrarily small when d grows too (provided that L is large enough)? This must be answered in the negative [13]. In fact, there exists some monotonic $d = d(L)$ such that in the limit of $L \to \infty$ we have $q_c = 1/2 = p_c$. It is important to notice that $d(L)$ is still a very slowly growing function of L, more precisely $d(L) = (\ln \ln L)^2 \ln \ln \ln L$, but nevertheless it is sufficient for the result. In other words, there certainly exists a severe bound on how fast L must grow when d is systematically raised, in order that we still have $p_c = 1$ for the system of infinite size. If the condition is not met, then $p_c < 1$ ($q_c > 0$) for a diverging L.

To complete this survey, we mention briefly the case $m = z - 2$ [7] where we have relations that are described in Eq. (9). The simplest case is when $m = 4$ for the simple cubic lattice. One can derive a bound by a recursion formula similar to Eq. (15), with partial probabilities of the type p^k in Eq. (15) now substituted by p^{k^2} (or more generally $p^{k^{d-1}}$ in d dimensions). To see this, consider the simple cubic model with $m = 4$ and the instability condition for cubic voids. We can reiterate an argument similar to (and yet simpler of) the argument above. If in any of the six boundary surfaces there lies at least one vacant site, then the empty cube can grow in that direction. If all six surfaces satisfy the condition, the whole cube will grow in all six directions concurrently. If the side length is given by k, the probability that one boundary surface is entirely filled with particles is given by p^{k^2}, and the complement is therefore $1 - p^{k^2}$, i. e. the probability that there lies one vacant site at least, or more. Rather than employing a full calculation (lengthy but trivial) for the bound such as the one in Eq. (15), we recognize that this situation is similar to the case $m = 2$ on the square lattice. There the critical size for the instability of the voids had to satisfy the condition $kq \gg 1$, i. e. with high probability there are vacant sites on the line. Here we have that a similar condition must hold, more precisely that $k^2 q \gg 1$ must hold. We therefore have, roughly, that $\bar{k} \sim 1/\sqrt{q}$ is the threshold size for voids' instability. By analogy with that case, we conclude that $L_c \simeq O \exp(1/\sqrt{q})$, in accordance with (9).

Additionally, some authors [9] have studied the behavior of isolated hypercubes for $m = d - 2$, for large values of d, showing that the critical concentration of empty sites sharply decays to 0 as $q_c \sim 2^{-2\sqrt{d}}/d^2$.

We conclude this survey on models with sharp thresholds by noticing that variants of bootstrap percolation may entail lengths that are diverging comparatively more severely than those outlined in this section, on approach to the critical point $q = 0$. One example is the square lattice with the asymmetric rule introduced by Gravner and Griffeath [31]. The neighborhood is extended to two additional sites, the second neighbors that lie, for example, on the horizontal directions. The total number of relevant neighbors is therefore $4 + 2 = 6$, the four nearest and two second-nearest neighbors. The value of m is taken to be $m = 3$. One finds [71] that at $q \ll 1$, $L_c \sim \exp[O(1/q) \ln^2(1/q)]$. This is the same asymptotic behavior [50,72,73] as in another asymmetric bootstrap percolation model on the square lattice introduced by Duarte [26]. Only three out of four nearest neighbors are considered each time, and particles are removed if at least 2 of those neighbors are empty. An even sharper transition was found earlier for a special variant of the bootstrap model on the square lattice [61]. In this model, occupied sites can be removed as usual, but at the same time empty sites can also become occupied. The rule is therefore not strictly irreversible. At each iteration, any given site adjusts its state so that it conforms to the occupancy state of the majority of its neighbors. The key ingredient for the model is the assignment of a bias in all cases in which there is a parity situation, i. e. two neighbors are occupied and two are vacant. There is no majority, and to determine which state a site has to take a bias in introduced towards one of the two possibilities. For example we can codify the rule so that, for all parity situations, the site is removed with probability $y > 1/2$, and it becomes occupied with probability $1 - y < 1/2$. Therefore when 2 neighboring sites are empty and 2 are occupied, at the next step the site will become empty with a higher probability than to become occupied. In this case, for every choice of $y > 1/2$, $p < 1$ and for large enough system sizes, there is a high probability that the final state will be an empty lattice. The critical system size scales [61] as $c_1(y)/q^2 < \ln L_c(q; y) < c_2(y)/q^2$, where c_1 and c_2 are two unknown constants that depend only on y. In other words, in this case $L_c \sim \exp[O(1/q^2)]$.

Corrections to the Asymptotic Behavior in the Presence of Unstable Voids

Typically, for the models that have a critical bootstrap length that grows exponentially or super-exponentially in $1/q$ a detailed comparison of computer simulations inferences with rigorous mathematical results is unsuccessful [3,44]. For the two dimensional models, conventional $m = 2$ or modified, the constant A of best fit from the expression in Eq. (11) differs sensibly from the expected values derived from Eqs. (3), (4), which are mathematically rigorous [35]. As an even more concerning issue, it has been suggested that the best fits can sometimes be obtained from different functional expressions altogether. For example in three dimensions, on the one hand we have the relation (5) that implies that we should expect the best fits to L_c to be in the form $L_c \sim \exp\exp A/q$, the closer we approach $q = 0^+$ the better the fit. Contrarily to common intuition, however, it was proven from computer simulations [44] that for the largest sizes and the smallest q available a fit in the form $L_c \sim \exp\exp A/q^2$ is far more reliable, the closer we are to $q = 0^+$ the better the fit is compared to the expected $L_c \sim \exp\exp A/q$. At first sight this could erroneously be interpreted as a mere failure of either mathematicians or simulators. However this does not need be the case. Take for example the two-dimensional case. If we look carefully at what the asymptotic results really mean, we have two equivalent ways of expressing them in general terms. One is as in Eqs. (3) and (4), and the other one is the equivalent form Eq. (13), provided that Eq. (12) holds. From the rigorous analytical arguments usually employed we have no way (as yet) to infer the exact form of either $\epsilon(L)$ or $f(q)$. We know however that $\lim_{L\to\infty} \epsilon(L) = 0$ and $\lim_{q\to 0} q\, f(q) = 0$. However, $f(q)$ might still be a diverging function of q in the limit $q \to 0^+$, albeit not as fast as $1/q$. Plus, $f(q)$ enters Eq. (13) exponentially. This is worrying because it implies that the resulting discrepancy between Eq. (13) and Eq. (11) may be in the region of orders of magnitude, and indeed it is [22].

A natural question arises. Where did we lose track of all such non-asymptotic corrections? We may start looking backward at all the derivations and see where we might have made too gross assumptions. One of the main suspects is a relation that we have used all throughout in the discussion, although without rigorous proofs of its legitimacy, but rather tacitly assuming its general validity. Given the probability ν (also called the 'connected hole density' [22,23,45]) that a vacant site can nucleate indefinitely (i. e. that it becomes an unstable void), we have repeatedly inferred that the relation $L_c^d \nu \simeq 1$ holds reliably. Indeed, this is not strictly true in general. It tacitly assumes that the successful nuclei are not correlated spatially in any important way. On the other hand, at high concentrations of particles such nuclei become very rare and therefore highly contingent on the specific distribution of particles and vacancies in their surrounding, meaning that the probability to find one nucleating center just a few lattice spacings away from another one is extremely low. Even if we were to assume that there is a typical average number n of nucleating centers that occasionally tend to stick together, the consequent correction to the usual assumption $L_c^d \nu \simeq 1$ would be just a number, a factor of order n (eventually depending slowly on p). Hardly something that would account for order of magnitudes anyway. The good news is that ν is a quantity that is directly accessible from simulations, more efficiently than L_c indeed [22,45]. It is therefore possible to compare ν to direct estimates [3,44] of $1/L_c^d$ and operatively test the relation $L_c^d \nu \simeq 1$. Wherever comparisons have been made the two are in excellent agreement, provided p is not too small. We must look elsewhere to solve the mystery.

When we estimated ν_b as a bound for ν starting from Eq. (15), we explicitly stated that we were using a sufficient condition for the candidate unstable void, i. e. that this had to be be a square all throughout the sequence. However this is not a necessary condition. In principle, even if the void cannot grow conserving its aspect ratio at some particular stage, it may still grow as a rectangle, and eventually become unstable later in the process, although this event cannot be accounted for when we estimate ν_b from Eq. (15). In Fig. 2 we have an example of a configuration that for the modified model would lead the square process to stop before the largest is vacated entirely, while the same configuration would lead to an empty final state if we allow one small elongation. One can also prove that growth by rectangular voids of different sizes and elongations encompasses (i. e. exhausts) all self-efficient transformations through which a nucleating center can engulf one by one, neighbor by neighbor, the particles in the entire lattice. In other words, if a growing void cannot wipe out all particles in the system via rectangular transformations, it cannot do so on its own in any other way. The condition of growth by rectangles is not only a sufficient one (as it was in the case of squares), but also necessary. Consequently a calculation (if available) of the type derived in Eq. (15), but dealing with rectangles instead of squares as the intermediate transition states, would gain us access to ν exactly.

For the modified model in $d = 2$ this is exactly what can be done [23]. Instead of using Eq. (15) to estimate ν_b, one can generalize the method and determine ν accurately for any density, without the need to simulate the model.

The procedure entails mapping the growth by rectangles into an explicit form that takes the form,

$$P^i_{m,n} = \sum_j \left(c^{ij}_{m,n} P^j_{m-1,n} + d^{ij}_{m,n} P^j_{m,n-1} \right) \tag{19}$$

In Eq. (19), $P^i_{m,n} = P^i_{m,n}(p)$ represents the probability that the nucleation process via rectangular growth will be found after $m+n$ steps to occupy the state i. The state i at given m, n is typically a rectangle of sides $m \times n$ that is entirely empty, but eventually for some of the lines on its contour that may be entirely occupied (which ones these are is in fact characteristic of the label i itself). The range of i's and of j's coincide. The map is constructed purposefully in such a way that the transitions from state to state can be expressed in the linear form Eq. (19). Concurrently, it has to be constructed so that the transitions coefficients $c^{ij}_{m,n} = c^{ij}_{m,n}(p)$ and $d^{ij}_{m,n} = d^{ij}_{m,n}(p)$ have explicit expressions (which are similar but not equal to the partial terms employed in the product Eq. (15) for squares). The mapping is not unique, as in the end we are going to take the limit $m, n \to \infty$, but any map in the form (19) has to satisfy the two conditions above. From the set of states with probabilities $P^i_{m,n}$ there must be one, say $i = 1$ for example, that represents an $m \times n$ rectangle completely empty. $P^1_{m,m}$ is therefore the probability for an empty square to arise after $2m$ steps, and ν will equal $\nu = \lim_{m\to\infty} P^1_{m,m}$. The map in (19) can be easily solved numerically. $P^1_{m,m}$ converges sufficiently rapidly to ν so that we can genuinely calculate numerically ν for almost any value of p, well beyond anything of physical interest.

By solving Eq. (19) we therefore have the exact ν for any particle density p for the $d = 2$ modified model. Agreement with its value from computer simulations is beyond the most optimistic expectations [23]. From $\nu \simeq 1/L_c^2$ we now have a reliable interpolation to the bootstrap critical length for any system size that is available from computer simulations, plus a prediction for ν and therefore L_c well beyond scales of even physical interest, let al.one one-day to-be computational limits. This procedure lacks the insight and elegance of compact formulas such as the one in Eq. (4), but it has the merit (at least for two dimensional models) to wash away all uncertainties related to previously irreconcilable mathematical and computational differences.

Once we have ν for any p, we can access many quantities that are normally available only via computer simulation, with the advantage that now we do not have any important limitation in regard to system sizes. Thus for example we may infer, approximately, the probability that a spanning cluster survives the culling process, given the system size L and the particle density p. We have

$$P_{Cl}(p, L) \simeq \Theta\left(1 - L^2\nu(p)\right)\left[1 - L^2\nu(p)\right] \tag{20}$$

meaning that the average probability $1 - P_{Cl}$ that a cluster does not survive the culling process is proportional to the average number of connected holes for that size. In essence, a re-elaboration of our (in)famous $L^2\nu \simeq 1$. The pre-factor with the Θ step-function (equal to 0 if its argument is negative and to 1 if positive) guarantees that P_{Cl} is positive definite. In Fig. 1 we have P_{Cl} as a function of p for four different system sizes. We have already appreciated the same figure in a previous explanatory section, and to whoever may have wondered how one could have those largest system sizes from computer simulations the trick is now revealed.

We may also try to fit $L_c \simeq 1/\sqrt{\nu}$ using the expression given in Eq. (13). In essence we thus try to infer the form of $f(q)$. Although this is only a best fit, and has no rigorous analytical basis, it is suggestive, because we can access virtually any large value of L_c from the knowledge of ν (even up to 10 to the order hundred). It is also insightful because it gives a flavor of the magnitude of the correction. An excellent fit at $q \ll 1$ bears [24], for example, $f(q)$ in the form $f(q) \approx -Bq^{-\beta}$, where B is a fitting constant of a few units, and the best estimate for β is $\beta \approx 2/3$. The modulating pre-factor in Eq. (13) given by $\exp[f(q)]$ is therefore moderating the divergence of the asymptotic contribution. From Eq. (12) one sees that by successive iterations one can get $\epsilon(L)$ (which is negative as $f(q)$ is negative), but in general the form in Eq. (4) is less convenient (and less natural) than the one in Eq. (13). More explicitly, we have that if $qf(q) \ll 1$ and $f(q) \approx -Bq^{-\beta}$ holds, we estimate $\epsilon(L) \approx -C/(\ln L)^{1-\beta}$ for diverging L, where C is a constant. Recently, Gravner and Holroyd [32] have proposed a rigorous proof that $\epsilon(L)$ indeed must approach zero no faster than $-\text{Const}/\sqrt{\ln L}$. This implies $\beta > 1/2$, consistently with the extrapolated value of 2/3 above.

Bethe Lattice Calculations: The Cases $m \geq 2$

As we have seen in the previous sections, we are always able to classify the behavior of bootstrap percolation model if either $m < 2$ (i. e. $k \leq 2$) or, for regular lattices, if $m \geq d$ (i. e. $k > d$). The former cases fall under the same universality classes of standard percolation and are valid for any lattice geometry and any dimension. The transition is continuous and characterized by universal exponents [16,21,48]. The latter class is that with a critical concentration p_c that is unity for an infinitely extended system, and is more reminiscent of first-order transition behavior. It is characterized by the absence of any genuine

power law behavior in the relevant quantities. Notice also that, for regular lattices, $d = z/2$. If we then generalize the condition $m \geq d$ to the condition $m \geq z/2$, we are allowed to include the triangular lattice [43] in the same class discussed in the two previous sections, as well as the face centered cubic lattice (fcc), as simulation results suggest [45] (it would be interesting to determine rigorously whether $m \geq z/2$ is a general requirement for belonging to the second class of models for any lattice geometry). A question arises as to whether, depending on the lattice geometry and the dimensionality, there can be other classes with behavior different from the two classes above. The rationale for this expectation has three motivations. One is the obvious observation that there are dimensions and lattice geometries for which some values of the parameter m are not covered by either class, and we cannot apply the general theoretical frameworks of above. Examples are the fcc lattice with $m = 2, 3, 4, 5$ ($k = 3, 4, 5, 6$) or the $d = 4$ hypercube lattice with $m = 2, 3$ ($k = 3, 4$). A second motivation is that as we admit conditions for culling that include an extended neighborhood larger than the set of nearest neighbors, different classes of behavior do appear to arise [63,67]. The third motivation stems from the earliest results that were obtained within bootstrap percolation calculations on Bethe lattices [20], which were unexpected in nature as will be discussed briefly. The Bethe lattices are graphs with no closed loops and therefore many geometric specific correlations associated to Euclidean lattices are washed away. On the other hand, they are often regarded as mean-field models for the associated transitions (in this case the bootstrap transitions). In this sense, they might be seen as characteristic of genuine bootstrap behavior in infinite dimensions. Depending on the lattice geometry and on the dimension, remnants of the ideal Bethe-like behavior might be observed.

In a Bethe lattice (or Cayley tree) with coordination number z, every site is linked to z other sites. In turn each one of those sites is linked to $z-1$ other sites all disconnected with each other, and so on in a hierarchical fashion. As usual, each site is initially populated by particles with random probability p. The bootstrap percolation rule is unchanged, i. e. we assign a rule so that at each iteration occupied sites that have m or less occupied sites among the z neighboring are removed, while those with $k = m + 1$ or more that are occupied remain occupied. After many iterations we check for the presence of a spanning cluster formed by the remaining particles. For $m = 0$, only isolated sites are removed, and we have as usual standard nearest neighbor percolation on the Bethe lattice, with a critical concentration $p_c = 1/(z-1)$ and a probability $P_\infty(p)$ that a site is occupied by a particle belonging to the spanning cluster that vanishes linearly in $p - p_c$, when p approaches p_c from above. Also for the case $m = 1$ $P_\infty(p)$ vanishes at p_c, but this time with a non universal exponent of 2 [20]. Naturally, the case $m = z - 1$ ($k = z$) is trivial, because for any $p < 1$ all particles can be removed one after the other. The most interesting case is when $2 \leq m < z - 1$ ($2 < k \leq z - 1$). One finds that for any given p the probability $P_\infty(p)$ that a site hosts a particle belonging to the infinite k-cluster is given by,

$$\begin{aligned} P_\infty(p) &= 0 && p < p_c \\ &\simeq a + b(p - p_c)^{\frac{1}{2}}\,, && p \geq p_c, p \simeq p_c \end{aligned} \quad . \tag{21}$$

We have therefore a discontinuity at p_c, and concurrently the order parameter has critical behavior when $p \to p_c^+$. Noticeably, $\partial P_\infty/\partial p \simeq (p - p_c)^{-1/2}$ when p approaches p_c from above. Furthermore, a correlation length diverging at p_c (identified as the susceptibility) has been found to exist [63], thus showing that the transition on the Bethe lattice for $m > 2$ is of a hybrid nature.

To proceed for the general case, we may define the probability R that a given site is *not* connected jointly with one specified neighbor to a k-cluster that may eventually span the system. This probability R must satisfy a condition of self-consistency of the type,

$$R = 1 - p + p \sum_{n=0}^{m-1} \binom{z-1}{n} R^{z-1-n}(1-R)^n \tag{22}$$

We also define $T = 1 - R$ and we can illustrate the result [20,49,64] in Eq. (21) briefly for the case $z = 4$ and $m = 2$ as a concrete explanatory example. In this case all the particles in the surviving spanning cluster must have 3 or more occupied neighbors. We can imagine a hierarchical organization of connected links in which 3 links from upward merge into one site that has one other link going downward, and that will itself merge with 2 other links from sideways, and so on. T ($= 1 - R$) is the probability that a site is connected to at least 2 among the 3 sites merging from above it, all of which are themselves connected to at least 2 of the 3 sites above, and so on indefinitely. If $T = 0$, we immediately conclude that there cannot be any spanning cluster with connectivity 3. On the other hand, if $T > 0$, the opposite is true. This is not straightforward, but it suffices to realize that if a site is occupied, and if at least 3 of the 4 nearest neighbors are each part of the substructure defined by the probability T, it follows that that particle must necessarily belong to a spanning cluster of connectivity 3. The probability T must satisfy the following relation of self-consistency,

$$T = p[T^3 + 3T^2(1-T)] \tag{23}$$

which is equivalent to Eq. (22) above for $R = 1 - T$, while the probability that a site belongs to the spanning cluster is given by,

$$P_\infty = p[T^4 + 4T^3(1 - T)] \tag{24}$$

Equation (23) always admits the solution $T = 0$ for any p. We can eliminate T on both sides and seek for the solutions with $T \neq 0$. We must solve,

$$2T^2 - 3T + \frac{1}{p} = 0 \tag{25}$$

which has as a physical solution the only real root such that $T = 1$ when $p = 1$, which is

$$T = \frac{3}{4} + \sqrt{\frac{9}{16} - \frac{1}{2p}} \tag{26}$$

As we see, for any $p < 8/9$, T has no non-zero real solutions. It has the solution $T = 3/4$ when $p = p_c = 8/9$, which shows the discontinuity explicitly. We can write Eq. (26) in the convenient form,

$$T = \frac{3}{4} + \sqrt{\frac{p - p_c}{2pp_c}} \tag{27}$$

which takes the desired form, as $p \to p_c^+$,

$$T = \frac{3}{4} + \frac{1}{p_c\sqrt{2}}\left(p - p_c\right)^{\frac{1}{2}} \tag{28}$$

From the above expression for $p \to p_c^+$ combined with Eq. (24) we can easily verify that the relation (21) holds for $z = 4$ and $m = 2$.

Probabilistic Bootstrap Percolation

Regular lattices have the disadvantage that the coordination number z is relatively small, and we do not have a very wide spectrum of choices for the parameter m. One way to circumvent this problem in planar lattices is to extend the neighborhood of influence of a given site, so as to artificially enlarge the formal coordination number at each site. In this way, the bootstrap rule for removal of one particle is set to depend on the state (occupied or vacant) of many sites, beyond the mere nearest neighbors. Mixed transitions arise in these cases [63,67], although not exactly expressible in a form quantitatively identical to that given in Eq. (21).

Another approach that was taken by [16] was to introduce a probabilistic bootstrap model. This involves a simple extension to the model by considering mixtures of particles with different bootstrap rules. The particles are again occupied with density p, but each particle is assigned its own threshold m_i as a criterion for removal during the bootstrap procedure. The author of [16] studied the case of a simple mixture of two types: m or m' with probability r and $1 - r$ respectively. The motivation is to study non-integer values of m by controlling the mixing parameter r. The model is exactly solved and the results show that we can recover the full range of behavior from a continuous transition ($\beta = 1$) to the standard discontinuous or hybrid transition ($\beta = 1/2$).

The interesting cases are the ones which interpolate between continuous $m < 3$ and discontinuous transitions $m' \geq 3$ for the simple cubic lattice.

Using mixtures of particles with different rules is a promising approach for investigating the nature of bootstrap transition on real lattices [46] in 2 and 3 dimensions where simple nearest-neighbor models do not conclusively show the hybrid transition of Eq. (21).

Another case that has been studied is the $m = 2$ case on the square lattice in a polluted environment [33], i. e. in the presence of a random fraction s of sites that are neither occupied nor empty. The occupied sites can be removed only if at least 2 neighboring sites are empty, and thus polluted sites may sometimes obstruct the culling process. It is found that even at low pollution s, if additionally $q < \sqrt{s}$ then with finite probability many particles survive the bootstrap process, no matter how large the system.

The general result in Eq. (21) is suggestive also for other reasons. An identical relation (with the same exponent) to that of Eq. (21) holds if we replace P_∞ with the average number of touching hard spheres in the proximity of the jamming transition, at a given packing fraction [52]. Similar correspondences can be made with other quantities, still bearing the same exponents. It has been natural to investigate whether similar hybrid transitions as shown in the Bethe lattice survive if we pass from the loop-less case back to bootstrap percolation on Euclidean lattices. Undoubtedly this cannot be true in general, for Eq. (21) is valid for any z and for any $m \geq 2$, while we know already from the previous sections that Eq. (21) does not hold for the square lattice when $m = 2$ and for the simple cubic lattice in the case $m = 3$. Furthermore, when very large system sizes are tested, it appears that also for the case $m = 2$ for the simple cubic lattice all exponents lie in the same universality classes of standard percolation [17] (i. e. of continuous phase transitions). The former case ($m = 2$ simple cubic) is particularly interesting. The critical concentration is estimated to lie in the vicinity of $p_c \simeq 0.573$, higher than that of the standard $m = 0$ and $m = 1$ value of $p_c \simeq 0.312$. All exponents but the exponent β (related to the vanishing of P_∞) were recognized

to be universal [4,43]. Later, results for sizes larger than tested before suggested that also β might be universal [17]. Therefore we conclude that we must rule out all nearest-neighbor bootstrap percolation models on regular lattices in $d = 2, 3$ (square and simple cubic), if we want to recover the behavior as shown in Eq. (21).

High Dimensional Lattices

Another scenario that we can envisage involves high-dimensional Euclidean lattices. It has been shown that high dimensionality expansions of the infinite dimensional Bethe lattice calculations, by means of perturbative methods [34], produce results that are qualitatively similar to Eq. (21) at sufficiently large d. In other words, for $m \geq 2$ and for high d we might encounter hybrid mixed transitions also for regular Euclidean lattices. This result has been somewhat complemented by the work of [13] for $m = d$ on hypercubic lattices as discussed towards the end of Sect. "Void Instability and Rigorous Results", whereby if d grows with L at a certain (slow) rate $d(L)$, we find $p_c = 1/2$, in contrast to $p_c = 1$ which is found for fixed d (a matter of future investigations will certainly be to determine whether p_c drops from 1 to 1/2 discontinuously for different $d(L)$). Whether such hybrid transitions can be observed in high d geometries is a matter of present investigations. Simulations in the $d = 4$ hypercubic lattice with $m = 3$ seem to suggest [54] that a discontinuity does arise at p_c, but that this is not associated with a diverging susceptibility (which is always finite at the transition). This is a reflection of the fact that the putative critical point appears to be at a concentration lower than p_c. Thus the discontinuous jump at p_c would be purely first-order in nature.

k-core Percolation on Random Networks

The Bootstrap Percolation problem has been studied on random networks, where it is commonly known as k-core percolation. The k-core is the largest connected cluster of the network (occupation density p) in which all the nodes have at least k neighbors, each of which is also in the subgraph. The k-core is found by removing all nodes with less than k neighbors- this culling process is repeated until all remaining nodes (if there are any) have at least k neighbors. The k-core extends the simpler notion of the giant connected component of a network and provides a useful description of a network in terms of successively nested cores. The first exact solutions of this problem were obtained by Chalupa et al [20]. on the Bethe lattice (an infinite regular graph) as detailed above Eq. (21). After some progress was made in determining thresholds for the emergence of a k-core in random networks [12,28,55], an exact solution was recently described by Dorogovtsev et al.[25] for uncorrelated, damaged or undamaged random networks with an arbitrary degree distribution.

We give a brief outline of the solution here. The starting point is the configuration model [14,51], an uncorrelated random network with a given degree distribution. An important point about this model is that the probability of there being more than one path between any pair of nodes is $O(n^{-1})$ where n is the number of nodes, and so the local structure of the network is tree-like. Making use of the tree-like structure of the network it is observed that the k-core coincides with the infinite $(k-1)$-ary sub-tree (in a k-ary sub-tree each node has at least k neighbors). We can therefore be certain that a randomly chosen node belongs to the k-core if at least k of its neighbors are the roots of $(k-1)$-ary sub-trees. This quantity, which we call R, is the probability that a randomly chosen node is *not* the root of a $(k-1)$-ary sub-tree. As R is defined it only describes one end of an edge, and is independent of whether the other end is in the k-core or not. If we choose an edge at random from the network, then it belongs to the k-core if both ends are the roots of $(k-1)$-ary sub-trees- and so, with a factor of $(1-R)$ for each end of the edge we can express R in terms of the number of edges in the k-core, L_k

$$\frac{L_k}{L} = (1-R)^2 \tag{29}$$

The parameter R plays the role of the order parameter in k-core percolation and this expression relates it to simple measurable quantities of the networks (statistics of edges in the k-core).

Another basic quantity of interest is the relative size of the k-core M_k:

$$M_k = p \sum_{n \geq k} \sum_{q \geq n} P(q) C_n^q R^{q-n} (1-R)^n \tag{30}$$

In this expression, $C_n^q R^{q-n}(1-R)^n$ is the probability that exactly n neighbors are roots and $q-n$ are not roots of $(k-1)$-ary trees (where C_n^q is the binomial coefficient) and $P(q)$ is the probability that a node has q neighbors.

The nodes that comprise the k-core itself constitute an uncorrelated network and so their structure is encoded in the degree distribution of the k-core $P_k(n)$ (the probability that a node in the k-core has n neighbors):

$$P_k(n) = \frac{M_k(n)}{M_k} \tag{31}$$

Order Parameter for *k*-core Percolation

In general, to find the physical quantities of interest, we must first solve for R for a given state of the network (R de-

pends on k and the damage p). A given end of an edge is the root of a $(k-1)$-ary sub-tree (with probability $(1-R)$), if at least $(k-1)$ of its children are also the roots of $(k-1)$-ary sub-trees. The exact expression is [30]:

$$1-R = p \sum_{n=k-1}^{\infty} \left[\sum_{i=n}^{\infty} \frac{(i+1)P(i+1)}{z_1} C_n^i R^{i-n} (1-R)^n \right] \tag{32}$$

where z_1 is the first moment of the degree distribution, and $(i+1)P(i+1)/z_1$ is the number of *out* edges of the node arrived at from a given edge. This equation bears some similarity to that for ordinary ($k=2$) percolation (see for example Eq. 1 of Chalupa et al [20].). As in $k=2$ percolation, it is found that there are two possible solutions: $R=1$: leading to a continuous transition, and $0 \le R < 1$: leading to a discontinuous/hybrid transition. The critical behavior for the $R<1$ case is found to be [25,30]:

$$R^* - R \propto (p - p_c)^{1/2} \tag{33}$$

where R^* is the value of order parameter at the point where the k-core vanishes. This hybrid critical behavior, characterized by a jump in the order parameter at $p_c < 1$ and a square root singularity in the percolating phase, is shared by other quantities, such as the relative size of the k-core, M_k:

$$M_k - M_k^* \propto (p - p_c)^{1/2} \tag{34}$$

As the bootstrap process removes nodes at random, the cores are destroyed starting with the highest k-core. Each k-core collapses with the critical behavior of Eq. (33).

In Fig. 4 we show the relative sizes of all the possible k-cores in networks with Poisson degree distributions ($z_1 = 10$) as a function of occupation density.

These results hold for networks with arbitrary degree distribution but z_2 must also be finite. In scale free networks (degree distribution $P(q) \propto 1/(q+c)^{-\gamma}$) with $\gamma > 3$, where z_2 is finite, the critical behavior for $k \ge 3$-core percolation is found to be the hybrid phase transition (33). For $2 < \gamma \le 3$ the transition is infinite order, similar to the case for ordinary percolation in scale-free networks.

It is worth noting that the hybrid transition indicates there should be a diverging length-scale which has not been conclusively identified – it is the mean size of finite clusters which drives the conventional continuous transition, but for tree-like network models it is not obvious what the nature of divergence is since finite clusters are not supported for $p < p_c$.

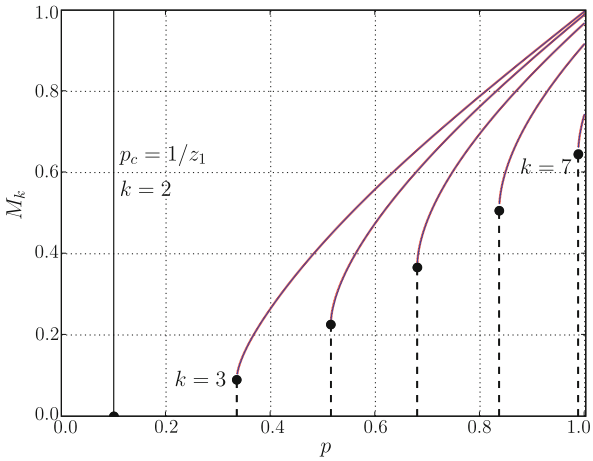

Bootstrap Percolation, Figure 4

The relative size of the k-core for a classical Erdös-Renyi random network. The mean degree $z_1 = 10$. The case $k = 2$ corresponds to ordinary percolation where there is a continuous transition at $p_c = 1/z_1$. For $3 <= k <= 7$ we see that a core emerges with a discontinuous jump in M_k (see Eq. (34))

The first clues to understanding the nature of the hybrid transition were found by studying the critical behavior of the subset of the k-core known as the corona [63]. The corona is comprised of nodes in the k-core which have exactly k neighbors in the k-core (this is the minimum number they can have). The nodes in the corona are randomly distributed throughout the k-core and for $p > p_c$ the corona is comprised of small finite clusters. It is an important observation that if a random node in the k-core is removed, then all the nodes in the corona clusters attached to that original node will be removed- leading to increasingly massive cascading removals, contingent on the removal of a single node, on approach to the transition.

The precise behavior of the mean size of clusters in the corona attached to a randomly chosen vertex in the k-core, N_{crn}, is known for random networks [25,30]. It diverges at the transition,

$$N_{crn} \propto (p - p_c)^{-1/2} \tag{35}$$

as was also found by [63] for the Bethe lattice. The picture that emerges is one of the collapse of the k-core at p_c, accompanied by a percolation transition of the corona cluster. For $p > p_c$ the corona clusters are disconnected, while for $p < p_c$ there is neither a k-core nor any corona clusters. As in ordinary percolation, we can identify a susceptibility belonging to the mean size of the clusters in the corona [30]. find that this susceptibility diverges also at the transition- indicating that the corona clusters undergo a percolation transition, growing in size as $p \to p_c^{(+)}$,

percolating at $p = p_c$, and disappearing altogether for $p < p_c$.

There have been some attempts to apply the above picture of k-core percolation to real networks, such as the Internet (at the Autonomous System level) [8]. The identification of the k-cores provides a useful insight into the structural properties and regions of increasingly centrality in these networks. Although the results are promising they are not in precise agreement with the predictions, a consequence of the fact that the theory deals only with uncorrelated networks.

Future Directions

Future work will continue to try to understand the unusual scaling behaviour, and further develop rigorous results. The existence and the consequent classification of the transitions described above, including but not limited to hybrid mixed transitions, is likely to represent the next step for many working on bootstrap percolation transitions. The development of solutions for complex networks, k-core percolation, will lead to many new directions of interest including an understanding of the effect of correlations and loops. The broad range of applications that the bootstrap percolation problem has already found will continue to inspire fruitful research in diverse areas of physics and mathematics.

Bibliography

1. Adler J (1991) Bootstrap percolation. Physica A 171(3):453–470
2. Adler J, Aharony A (1988) Diffusion percolation. I. Infinite time limit and bootstrap percolation. J Phys A: Math Gen 21(6):1387–1404
3. Adler J, Lev U (2003) Bootstrap percolation: Visualisations and applications. Braz J Phys 33:641
4. Adler J, Stauffer D (1990) Evidence for non-universal exponents in bootstrap percolation. J Phys A Math Gen 23:L1119
5. Adler J, Palmer RG, Meyer H (1987) Transmission of order in some unusual dilute systems. Phys Rev Lett 58(9):882–885
6. Adler J, Stauffer D, Aharony A (1989) Comparison of bootstrap percolation models. J Phys A Math Gen 22:L297
7. Aizenman M, Lebowitz JL (1988) Metastability effects in bootstrap percolation. J Phys A: Math Gen 21:3801–3813
8. Alvarez-Hamelin JI, Dall'Asta L, Barrat A, Vespignani A (2005) k-Core decomposition: A tool for the analysis of large scale internet graphs. eprint arXiv:cs/0504107
9. Balogh J, Bollobás B (2006) Bootstrap percolation on the hypercube. Probab Theory Relat Field 134(4):624–648
10. Balogh J, Pete G (1998) Random disease on the square grid. Random Struct Algorithm 13(3–4):409–422
11. Balogh J, Bollobás B (2003) Sharp thresholds in bootstrap percolation. Phys A: Stat Mech Appl 326(3–4):305–312
12. Balogh J, Pittel BG (2007) Bootstrap percolation on the random regular graph. Random Struct Algorithm 30(1–2):257–286
13. Balogh J, Bollobás B, Morris R (2007) Majority bootstrap percolation on the hypercube. Arxiv preprint mathCO/0702373
14. Bollobás B (2001) Random graphs. Cambridge University Press, Cambridge
15. Branco N, Dos Santos R, de Queiroz S (1984) Bootstrap percolation: a renormalization group approach. J Phys C 17:1909–1921
16. Branco NS (1993) Probabilistic bootstrap percolation. J Stat Phys 70:1035–1044
17. Branco NS, Silva CJ (1999) Universality class for bootstrap percolation with $m = 3$ on the cubic lattice. Int J Mod Phys C 10:921–930
18. Cerf R, Cirillo E (1999) Finite size scaling in three-dimensional bootstrap percolation. Ann Probab 27(4):1837–1850
19. Cerf R, Manzo F (2002) The threshold regime of finite volume bootstrap percolation. Stoch Process Appl 101:69–82
20. Chalupa J, Leath PL, Reich GR (1979) Bootstrap percolation on a bethe lattice. J Phys C: Solid State Phys 12(1):L31–L35
21. Chaves C, Koiller B (1995) Universality, thresholds and critical exponents in correlated percolation. Phys A 218(3):271–278
22. De Gregorio P, Lawlor A, Bradley P, Dawson KA (2004) Clarification of the bootstrap percolation paradox. Phys Rev Lett 93:025501
23. De Gregorio P, Lawlor A, Bradley P, Dawson KA (2005) Exact solution of a jamming transition: Closed equations for a bootstrap percolation problem. Proc Nat Acad Sci 102:5669–5673
24. De Gregorio P, Lawlor A, Dawson KA (2006) New approach to study mobility in the vicinity of dynamical arrest; exact application to a kinetically constrained model. Europhys Lett 74:287–293
25. Dorogovtsev SN, Goltsev AV, Mendes JFF (2006) k-Core organization of complex networks. Phys Rev Lett 96:040601
26. Duarte J (1989) Simulation of a cellular automat with an oriented bootstrap rule. Phys A 157(3):1075–1079
27. Ertel W, Frobröse K, Jäckle J (1988) Constrained diffusion dynamics in the hard-square lattice gas at high density. J Chem Phys 88:5027–5034
28. Fernholz D, Ramachandran V (2003) The giant k-core of a random graph with a specified degree sequence. manuscript, UT-Austin, November
29. Fredrickson GH, Andersen HC (1984) Kinetic ising model of the glass transition. Phys Rev Lett 53(13):1244–1247
30. Goltsev AV, Dorogovtsev SN, Mendes JFF (2006) k-Core (bootstrap) percolation on complex networks: Critical phenomena and nonlocal effects. Phys Rev E 73:056101
31. Gravner J, Griffeath D (1996) First passage times for threshold growth dynamics on Z^2. Ann Probab 24(4):1752–1778
32. Gravner J, Holroyd A (2008) Slow convergence in bootstrap percolation. Ann Probat 18(3):909–928
33. Gravner J, McDonald E (1997) Bootstrap percolation in a polluted environment. J Stat Phys 87(3):915–927
34. Harris A, Schwarz J (2005) $1/d$ expansion for k-core percolation. Phys Rev E 72(4):46123
35. Holroyd A (2003) Sharp metastability threshold for two-dimensional bootstrap percolation. Probab Theory Relat Field 125(2):195–224
36. Holroyd A (2006) The metastability threshold for modified bootstrap percolation in d dimensions. Electron J Probab 11:418–433
37. Holroyd A, Liggett T, Romik D (2004) Integrals, partitions and cellular automata. Trans Amer Math Soc 356:3349–3368

38. Ising E (1925) Beitrag zur Theorie des Ferromagnetismus. Z Phys 31:253–258
39. Jäckle J, Krönig A (1994) A kinetic lattice-gas model for the triangular lattice with strong dynamic correlations: I. cself-diffusion. J Phys: Condens Matter 6:7633–7653
40. Jeng M, Schwarz JM (2007) Comment on "jamming percolation and glass transitions in lattice models". Phys Rev Lett 98(12):129601
41. Kirkpatrick S, Wilcke W, Garner R, Huels H (2002) Percolation in dense storage arrays. Physica A: Stat Mech Appl 314(1–4):220–229
42. Kob W, Andersen HC (1993) Kinetic lattice-gas model of cage effects in high-density liquids and a test of mode-coupling theory of the ideal-glass transition. Phys Rev E 48:4364–4377
43. Kogut PM, Leath PL (1981) Bootstrap percolation transitions on real lattices. J Phys C 14(22):3187–3194
44. Kurtsiefer D (2003) Threshold value of three dimensional bootstrap percolation. Int J Mod Phys C 14:529
45. Lawlor A, De Gregorio P, Bradley P, Sellitto M, Dawson KA (2005) Geometry of dynamically available empty space is the key to near-arrest dynamics. Phys Rev E 72:021401
46. Lawlor A, De Gregorio P, Cellai D, Dawson KA (2008) (to be published)
47. Manna SS (1998) Abelian cascade dynamics in bootstrap percolation. Physica A: Stat Theor Phys 261:351–358
48. Medeiros M, Chaves C (1997) Universality in bootstrap and diffusion percolation. Physica A 234(3):604–610
49. Moukarzel C, Duxbury PM, Leath PL (1997) Infinite-cluster geometry in central-force networks. Phys Rev Lett 78(8):1480–1483
50. Mountford TS (1995) Critical length for semi-oriented bootstrap percolation. Stoch Process Appl 56:185–205
51. Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45:167
52. O'Hern C, Silbert L, Liu A, Nagel S (2003) Jamming at zero temperature and zero applied stress: The epitome of disorder. Phys Rev E 68(1):11306
53. Onsager L (1944) Crystal statistics. I. A two-dimensional model with an order-disorder transition. Phys Rev 65(3–4):117–149
54. Parisi G, Rizzo T (2006) On *k*-core percolation in four dimensions. Arxiv preprint cond-mat/0609777
55. Pittel B, Spencer J, Wormald N (1996) Sudden emergence of a giant *k*-core in a random graph. J Comb Theory B 67:111–151
56. Pollak M, Riess I (1975) Application of percolation theory to 2D–3D Heisenberg ferromagnets. Phys Stat Solidi (B) 69(1):K15–K18
57. Ritort F, Sollich P (2003) Glassy dynamics of kinetically constraint models. Adv Phys 52:219–342
58. Sabhapandit S, Dhar D, Shukla P (2002) Hysteresis in the random-field ising model and bootstrap percolation. Phys Rev Lett 88(19):197202
59. Sahimi M (1994) Applications of percolation theory. Taylor/Francis, London
60. Schonmann R (1990) Critical points of two-dimensional bootstrap percolation-like cellular automata. J Stat Phys 58(5):1239–1244
61. Schonmann R (1990) Finite size scaling behavior of a biased majority rule cellular automaton. Phys A 167(3):619–627
62. Schonmann R (1992) On the behavior of some cellular automata related to bootstrap percolation. Ann Probab 20(1):174–193
63. Schwarz JM, Liu AJ, Chayes LQ (2006) The onset of jamming as the sudden emergence of an infinite *k*-core cluster. Europhys Lett (EPL) 73(4):560–566
64. Sellitto M, Biroli G, Toninelli C (2005) Facilitated spin models on Bethe lattice: Bootstrap percolation, mode-coupling transition and glassy dynamics. Europhys Lett 69(4):496–502
65. Smirnov S, Werner W (2001) Critical exponents for 2D percolation. Math Res Lett 8:729–744
66. Stauffer D, Aharony A (1992) Introduction to percolation theory. Taylor/Francis, London
67. Toninelli C, Biroli G, Fisher DS (2006) Jamming percolation and glass transitions in lattice models. Phys Rev Lett 96(3):035702
68. Toninelli C, Biroli G, Fisher D (2007) Toninelli, Biroli, and Fisher reply. Phys Rev Lett 98(12):129602
69. Treaster M, Conner W, Gupta I, Nahrstedt K (2006) Contagalert: Using contagion theory for adaptive, distributed alert propagation. nca 0:126–136
70. van Enter A (1987) Proof of Straley's argument for bootstrap percolation. J Stat Phys 48(3):943–945
71. van Enter A, Hulshof T (2007) Finite-size effects for anisotropic bootstrap percolation: logarithmic corrections. J Stat Phys 128(6):1383–1389
72. van Enter ACD, Adler J, Duarte JAMS (1990) Finite size effects for some bootstrap percolation models. J Stat Phys 60:323–332
73. van Enter A, Adler J, Duarte J (1991) Finite-size effects for some bootstrap percolation models, addendum. J Stat Phys 62:505–506
74. Widom B (1974) The critical point and scaling theory. Phys 73(1):107–118
75. Wilson K (1983) The renormalization group and critical phenomena. Rev Mod Phys 55:583–600

Brain Pacemaker

Peter A. Tass[1,2,3], Oleksandr V. Popovych[1], Christian Hauptmann[1]

[1] Institute of Neuroscience and Biophysics 3 – Medicine and Virtual Institute of Neuromodulation, Research Center Jülich, Jülich, Germany

[2] Department of Stereotaxic and Functional Neurosurgery, University of Cologne, Cologne, Germany

[3] Brain Imaging Center West, Research Center Jülich, Jülich, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Standard High-Frequency Stimulation
Multisite Coordinated Reset Stimulation
Linear Multisite Delayed Feedback

Nonlinear Delayed Feedback
Proportional-Integro-Differential Feedback
Plasticity
Summary
Bibliography

Glossary

Electrical deep brain stimulation (DBS) is the standard therapy for medically refractory movements disorders, e. g., Parkinson's disease and essential tremor. It requires a surgical treatment, where depth electrodes are chronically implanted in target areas like the thalamic ventralis intermedius nucleus or the subthalamic nucleus and, as yet, an electrical high-frequency (>100 Hz) stimulation is permanently delivered via depth electrodes. More sophisticated deep brain stimulation techniques are in the process of being established for clinical use.

Delayed feedback Delayed feedback is a method for the creation of a closed-loop forcing, where some portion of the measured output signal of a system is time delayed, linearly or non-linearly processed, and fed back into the system. This approach is often used to control the dynamic behavior of complex systems. In this article delayed feedback is used to control synchronization in ensembles of coupled oscillators, e. g., neurons.

Order parameter The order parameter is a quantity characterizing a phase transition or phase change in the transformation of a complex system from one phase (state) to another. The order parameter is convenient for characterizing the onset and extent of synchronization in larger ensembles: Perfect phase synchronization corresponds to a large value of the order parameter, whereas an incoherent (desynchronized) state is associated with a small value of the order parameter. In synergetics it has been shown that the dynamics of complex systems may be governed by only a few order parameters.

Synchronization Synchronization (from Greek *syn* = the same, common and *chronos* = time) means the adjustment of rhythms of self-sustained oscillators due to their weak interaction. The interacting oscillators can be regular (periodic) or chaotic. There are several different forms of synchronization including phase, complete, generalized, and lag synchronization, etc. In this article we focus on phase synchronization. In the simplest form, the oscillators, rotating with the same frequency, become phase synchronized (phase locked) to each other, if they tend to oscillate with the same repeating sequence of relative phase angles. Put otherwise, the oscillators adjust their rhythms, while their amplitude dynamics need not be correlated.

Definition of the Subject

A brain pacemaker is a medical device that is implanted into the brain with the purpose to stimulate nervous tissue with electrical signals. Brain pacemakers are used for the therapy of patients suffering for example from Parkinson's disease, epilepsy or mental disorders. Brain stimulation is either called deep brain stimulation (DBS) if structures deeply inside the brain are targeted or Stimulation, if the electrical contacts of the stimulator are positioned on the surface of the brain. Apart from direct brain stimulation, brain pacemakers may also be used to stimulate the spinal cord (e. g. for the treatment of pain) or the vagus nerve (for the treatment of epilepsy). The electrical stimulation of the nervous system has a long history which goes back to the 19th century where first tests with cortical stimulation were documented [23]. The first intraoperative deep brain stimulation was performed by Spiegel et al. in 1947 in a patient suffering from Huntington's chorea and in the eighties DBS was introduced as a treatment for motor disorders [4,11]. DBS was approved by the FDA as a treatment for essential tremor in 1997, for Parkinson's disease in 2002, and dystonia in 2003. The treatment of severe neurological and psychiatric diseases with brain pacemakers is a rapidly growing and promising field. Novel, model-based approaches, which use methods from synergetics, nonlinear dynamics, and statistical physics, to specifically restore brain function and connectivity, demonstrate how insights into the dynamics of complex systems contribute to the development of novel therapies.

Introduction

Self-organization processes are abundant in numerous fields of the natural sciences [28,29]. For instance, the nervous system elegantly utilizes self-organization principles for motor control purposes [30,32,43,72]. A classical example of a self-organization process is synchronization of populations of interacting oscillators, which is widely observed in physics [27,29,60,81], chemistry [46], biology [102], neuroscience [31,80], and medicine [16,54,86]. In the nervous system synchronization processes are important, e. g., in the context of information processing [76] and motor control [3]. However, pathological, excessive synchronization strongly impairs brain function [16,54]. In fact, pathological synchronization processes are the hallmark of several neurological diseases like Parkinson's disease (PD) or essential tremor [2,58]. For example, Parkinsonian resting tremor is caused by a pacemaker-like

population of neurons which fires in a synchronized and periodical manner [2,77]. In contrast, in healthy subjects these neurons fire in an uncorrelated, i. e. desynchronized manner [58].

Permanent deep brain stimulation (DBS) at high frequencies (>100 Hz) is the standard therapy for medically refractory patients suffering from Parkinson's disease and essential tremor [5,6,10], see Fig. 1. High-frequency DBS has been developed empirically, mainly based on experimental results and clinical observations. The mechanism of high-frequency DBS is still a matter of debate [7]. Clinical studies showed that high-frequency DBS essentially has similar effects as observed after tissue lesioning. Compared to lesioning, DBS is reversible and has a lower rate of side effects [71,82]. However, in spite of many beneficial effects, in some patients DBS may not help, or may cause side effects, or the therapeutic effects may disappear over time [14,20,66,82,101]. With the objective of finding better tolerated and more effective DBS techniques, a model-based development of novel stimulation methods has been initiated [37,38,39,41,62,63,64,86,91,93,94,98], see Fig. 2. In these studies, relevant neuronal target populations were modeled mathematically and stimulation techniques have been developed utilizing principles from nonlinear dynamics and statistical physics [86].

One goal of this approach is to control the pathological neural dynamics appropriately in order to achieve a mild and efficient relief of symptoms [86]. The second, more ambitious goal is to stimulate in a way that the formerly affected neuronal populations unlearn their pathological connectivity and, hence, their tendency to produce pathological synchronization [97]. Put otherwise, the very goal of this approach is to induce long-lasting therapeutic effects which outlast the cessation of stimulation [36,83,97]. To this end, stimulation algorithms have been developed and optimized to exploit dynamic self-organization principles and plasticity rules [36,41,83,96,97].

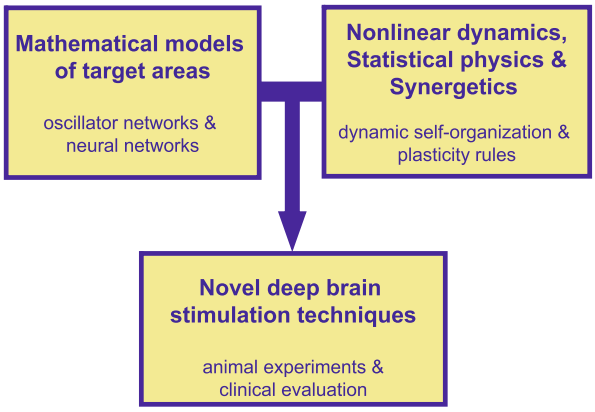

Brain Pacemaker, Figure 2
***Model-based development of novel deep brain stimulation techniques*: Along the lines of a top-down approach target areas for deep brain stimulation are modeled by means of oscillator networks and physiology and anatomy based neural networks. Methods from nonlinear dynamics, statistical physics, and synergetics are employed to develop stimulation techniques which specifically utilize dynamic self-organization principles and plasticity rules. Experimental feedback from both animal experiments and clinical evaluation serves to validate, falsify or modify theoretical assumptions and predictions. This iterative approach aims at steadily improving the mathematically designed stimulation techniques and, hence, at establishing superior therapies**

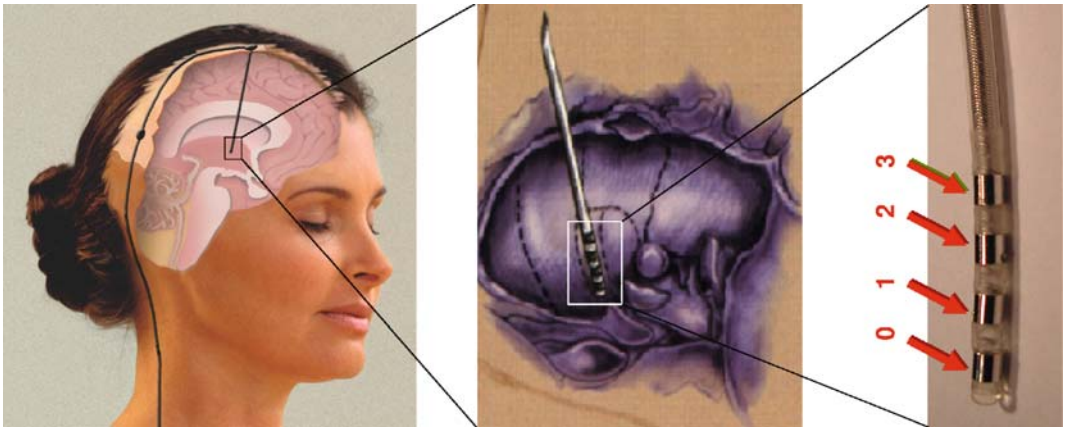

Brain Pacemaker, Figure 1
Standard DBS setup. A depth electrode is implanted into the target structure (e. g. the *subthalamic nucleus*). The electrode is subcutaneously connected with the generator of the high-frequency stimulation signal (not shown in this image). The stimulation signal is delivered through one or more of the four stimulation contacts labeled from 0 to 3

Three novel, particularly promising stimulation techniques have been developed in the past: The multisite coordinated reset stimulation (MCRS) [86,91,93,94], the multisite coordinated delayed feedback stimulation (MDFS) [37,38,39], and the nonlinear delayed feedback stimulation (NDFS) [62,63,64]. These techniques have the common objective of reducing the synchronized activity of the target population by reestablishing a normal physiological activity in a highly synchronized population of neurons.

MCRS uses short pulse trains to reset sub-populations of the neuronal network subsequently which induces a desynchronized state [86,91,93,94]. Here, the stimulation was applied through a small number of stimulation sites which were equally spaced within the neuronal population. MCRS induced desynchronization is achieved by utilizing self-organization principles, in particular, the slaving principle induced by the pathological neuronal interactions (i. e. interactions which have the potential to induce a pathological synchronization) [91,93,94]. MDFS [37,38,39] and NDFS [62,63,64] use delayed feedback for stabilizing a desynchronized state which was intended to be as close to the physiological mode of action as possible. Here, the local field potential (LFP) of the target population is measured, amplified, delayed, and fed back into the ensemble.

However, a novel, clinically extremely relevant aspect comes into play by considering the effect of such stimulation techniques on synaptic plasticity [97]. It has been shown experimentally, that synaptic plasticity enhances the pathological synchronization [59]. From the kindling phenomenon in the context of epilepsy it is well-known that neural networks may learn pathological strong interactions [56,79]. The novel desynchronizing stimulation protocols are designed to invert this pathological process, so that the affected neuronal populations unlearn their pathological connectivity, and physiological neuronal activity is re-established on a long-term basis. In a nutshell, the novel stimulation techniques aim at a well directed employment of fundamental principles of dynamic brain action to induce long-lasting therapeutic effects.

Standard High-Frequency Stimulation

High-frequency deep brain stimulation (HFDBS) is the standard therapy for patients suffering from medically refractory PD or essential tremor [5,6]. To this end, depth electrodes are chronically implanted in the thalamic ventralis intermedius nucleus or the subthalamic nucleus [5,6] and a permanent high-frequency (> 100 Hz) periodic pulse train stimulation is applied (Fig. 1). HFDBS has been developed empirically, mainly based on intraoperative observations. HFDBS strongly alters the neuronal firing and mimics the effect of tissue lesioning, e. g., by suppressing neuronal firing, which, in turn, suppresses the peripheral tremor [6,19,53,101]. However, as yet, the mechanism of HFDBS is not sufficiently understood [53].

During stimulation, high-frequency DBS seems to induce a regular bursting mode [9]. After a reduction of stimulation artifacts, robust bursting activity in subthalamic nucleus (STN) neurons was observed in slices from naive or resperine-treated rats [8]. After offset of stimulation, a blockade of activity, i. e. a non-specific suppression of the neuronal activity in the target structures through a depolarization blockade which did not require synaptic transmission was observed [8].

Other hypotheses are that high-frequency DBS applied to a PD patient mimics the effect of tissue lesioning and appears to block neuronal activity in relevant target areas during stimulation [6]. The most probable hypothesis was offered recently by Benabid et al. [7], in which a mixture of different mechanisms was discussed. The contributing mechanisms resulting in the observed effects of high-frequency DBS might be membrane inhibition, jamming, excitation of excitatory and inhibitory afferents, excitation of efferents and plasticity [7].

To precisely evaluate the contribution of these different mechanisms, spatially extended multi-compartment neuron models were used to demonstrate the effects of extracellular stimulation on the different structures of the stimulated neuronal population [26]. Depending on the stimulation amplitude and the shape of the stimulation pulses, either the cells were activated directly or fibers mediating excitatory or strong inhibitory action were activated [26]. Modeling studies indicate that already at the level of single neurons, the activation of a larger number of structures can take place with different and possibly conflicting impacts on the single neuron dynamics [26]. The collective dynamics of neuronal populations further adds aspects which are important for the creation of synchronized activity: Cells responding differently to external inputs like somatosensory stimulation or stimulation due to active movements are present in the target tissue together with so called no-response cells [47]. High-frequency stimulation has a complex impact on these structures [6,74].

Recent experimental and modeling studies indicate that the *globus pallidum interior* GPi – one structure of the basal ganglia – might be strongly involved in the mechanisms of DBS [21,35,52,55,69]. The results of modeling studies indicate that under Parkinsonian conditions the rhythmic inhibition from GPi to the thalamus com-

promises the ability of thalamocortical relay cells to respond to depolarizing inputs, such as sensorimotor signals. High frequency stimulation of STN regularizes GPi firing, and this restores the responsiveness of the thalamus [69].

HF DBS is reversible and has a much lower rate of side effects than lesioning with thermocoagulation [71]. Although HF DBS is the golden standard for the therapy of medically refractory movement disorders, there are still limitations of HF DBS: On the one hand HF DBS may cause adverse effects like dysarthria, dysesthesia, cerebellar ataxia, and memory decline [20,66,101]. On the other hand HF DBS may be ineffective or its therapeutic effect may wear off over time [45,66]. For instance, 11–15% of PD patients have unsatisfactory outcomes although their depth electrodes are properly placed [48].

Multisite Coordinated Reset Stimulation

To study the impact of pulsatile stimuli on single oscillators and, in particular, populations of oscillators in a biologically more realistic setting, it was necessary to take into account random forces [84,85,86]. To this end, a stochastic concept of phase resetting has been developed for populations of non-interacting [84,85] as well as interacting [86] oscillators in the presence of noise. In this approach limit-cycle oscillators were approximated by phase oscillators [34], so that the pulsatile stimulation only affects the oscillators' phases. If a single-pulse of the right intensity and duration is delivered to the population in the stable synchronized state, it causes an at least temporary desynchronization provided it hits the population at a vulnerable phase. Theoretically, single-pulse stimulation has also been studied in more complex networks, for instance, networks of coupled phase oscillators with inertia, modeling dendritic dynamics [15,50]. Based on the stochastic phase resetting theory and utilizing a phase oscillator as a model for a single neuron [34], demand-controlled single-pulse deep brain stimulation has been suggested for the therapy of movement disorders like Parkinson's disease or essential tremor [86,87].

However, there are drawbacks to single-pulse stimulation which decisively limit its applicability [89,104]: First, if the mutual coupling is not weak, the vulnerable phase range we have to hit in order to cause an effective desynchronization is only a small fraction (e. g., 5%) of a period of the collective oscillation. Second, the critical stimulation parameters required to achieve a good desynchronization depend on the dynamical state of the population. Thus, different stimulation parameters have to be used if the cluster is not in its stable synchronized state.

To overcome the limitations of single-pulse stimulation, double-pulse [88,89] stimulation has been proposed: Two qualitatively different stimuli are successively delivered. The first, stronger pulse resets (restarts) the collective oscillation irrespective of the initial state of the population. The second, weaker pulse is applied after a fixed time delay, where it hits the cluster in its vulnerable state and, hence, causes a desynchronization. There are different variants of double-pulse stimulation, depending on the type of stimuli used to achieve a reset or a desynchronization [90,91,92,93]. For instance, the first resetting pulse can be replaced by a brief high-frequency pulse train [90] or by a softly resetting low-frequency pulse train [91,92].

Although double-pulse stimulation causes a desynchronization irrespective of the initial dynamical state at which this stimulus is delivered, there are still limitations which may hinder an application to a biological system [94,95]: On the one hand, double-pulse stimulation requires a calibration. On the other hand, double-pulse stimulation is only effective if the system parameters are reasonably stable. The required quasi-stationarity of the system parameters combined with the possibly time consuming calibration may cause problems when applied to a real biological system, where fluctuations of model parameters are inevitable.

To provide a stimulation technique which is robust with respect to system parameters and which does not require calibration, multisite coordinated reset stimulation has been developed [94,95]. The idea behind this approach is to abstain from achieving a perfect desynchronization by a well-calibrated stimulus. Rather, by means of a robust and comparably mild stimulus the stimulated population is shifted into a dynamical state which is not the desired desynchronized state, but sufficiently close to it. Close in the sense that due to the pathologically strong coupling the population automatically relaxes into the desired desynchronized state. This approach essentially exploits the pathological tendency of the neuronal population to establish a synchronized state. Accordingly, multisite coordinated reset stimulation is in a way comparable to Asian martial arts, where ideally a minimal amount of energy (i. e. the multisite coordinated reset) is invested to control the adversary by utilizing the adversary's own energy (i. e. the neurons' pathological strong coupling).

The scheme of the stimulation setup is presented in Fig. 3 (left plot). Several stimulation sites are placed within the target network and weak resetting stimulation signals are administered via these stimulation sites. In this way the oscillatory population is split into several sub-populations, where each of them is assigned to the corresponding stimulation site and receiving the stimulation signal

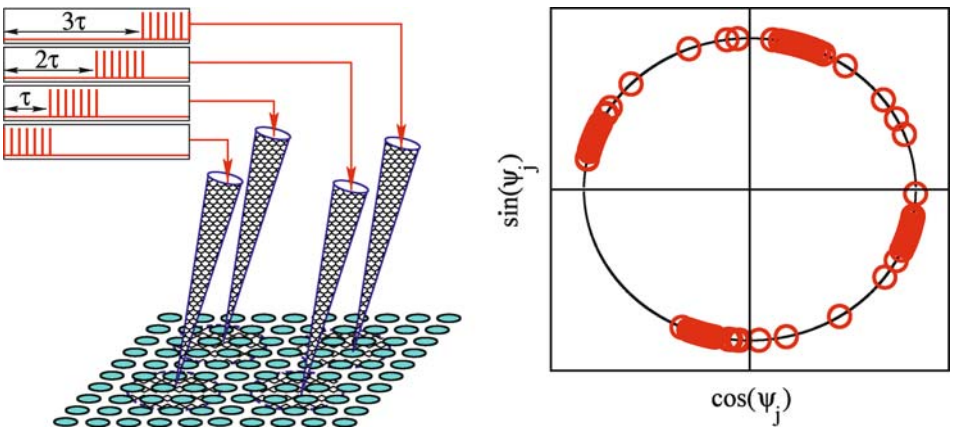

Brain Pacemaker, Figure 3

Stimulation setup of the multisite coordinated reset stimulation method. Brief and mild resetting stimuli are administered at different sites at subsequent times (*left plot*) and cause an effective splitting of the stimulated population into several sub-populations such that their phases ψ_j form phase clusters distributed uniformly (or close to that) over the period of oscillations (*right plot*)

mostly from that stimulation site. Multi-site coordinated reset stimulation means that a synchronized population of neurons is stimulated with a sequence of brief resetting stimuli (typically brief high-frequency (HF) stimulus trains) via the different sites. The delay between the subsequent resetting stimuli can be chosen as $\tau = T/n$ with respect to that at the preceding site, where n is the number of the stimulation sites, T approximates the mean period of the collective dynamics of the synchronized oscillators, and n is the number of stimulation sites [94,95].

The subsequent reset of the different sub-populations induces a so-called cluster state, i. e. the whole population splits into n sub-populations which differ with respect to their mean phase. This effect is illustrated in Fig. 3 (right plot), where a snapshot of the distribution of the phases ψ_j of stimulated oscillators is shown in the $(\cos(\psi_j), \sin(\psi_j))$ plane after coordinated reset stimulation. The phases of the population of oscillators stimulated via, e. g., four sites form four phase clusters distributed uniformly (or close to that) over the unit circle. To estimate the extent and type of synchronization of the whole population of N oscillators, the cluster variables

$$Z_m(t) = R_m(t)\mathrm{e}^{\mathrm{i}\Psi_m(t)} = \frac{1}{N}\sum_{j=1}^{N}\mathrm{e}^{\mathrm{i}m\psi_j(t)}\,, \qquad (1)$$

can be used, where $R_m(t)$ and $\Psi_m(t)$ are the corresponding real amplitude and real phase, where $0 \le R_m(t) \le 1$ for all time t [12,86]. Cluster variables are convenient for characterizing synchronized states of different types: Perfect in-phase synchronization corresponds to $R_1 = 1$, whereas an incoherent state, with uniformly distributed phases, is associated with $R_m = 0$ $(m = 1, 2, 3, \ldots)$. $R_1 = 0$ combined with large R_m is indicative of an m-cluster state consisting of m distinct and equally spaced clusters, where within each cluster all oscillators have similar phase. In Fig. 3 (right plot), for instance, $R_1 \approx 0.02$, whereas $R_4 \approx 0.87$ indicating a four-cluster state induced by the multisite coordinated reset stimulation.

From the cluster state the neurons typically relax to a uniformly desynchronized state before they revert back to the in-phase synchronized state, if left unperturbed. To understand how a stimulus-induced clustering leads to an effective desynchronization, the dynamics of the leading modes $Z_1, Z_2, \ldots$ can be considered. When the coupling among oscillators becomes sufficiently large, e. g., it exceeds a certain critical value, Z_1 from (1) becomes an *order parameter* [46], which according to the slaving principle [29] governs the dynamics of the other, stable modes Z_m $(m = 2, 3, \ldots)$ on the center manifold [61]: The order parameter Z_1 acts on a slow time scale, whereas the stable modes Z_m act on a fast time scale and relax to values given by the order parameter Z_1 [29,103]. In a system with large number of oscillators this relationship reads [86]:

$$R_m \propto R_1^{\nu} \quad \text{with} \quad \nu \ge 2\,, \quad m = 2, 3, 4, \ldots \qquad (2)$$

Hence, to maintain a desynchronized neuronal firing, multisite coordinated reset stimuli have to be administered repetitively.

Multisite coordinated reset stimulation exploits transient responses which are due to the oscillators' (pathologically strong) interactions. The general stimulation protocol of the coordinated reset stimulation is illustrated in Fig. 4. Here, the collective dynamics is visualized by considering the collective firing of the neurons. A single firing/bursting model neuron fires/bursts whenever its phase

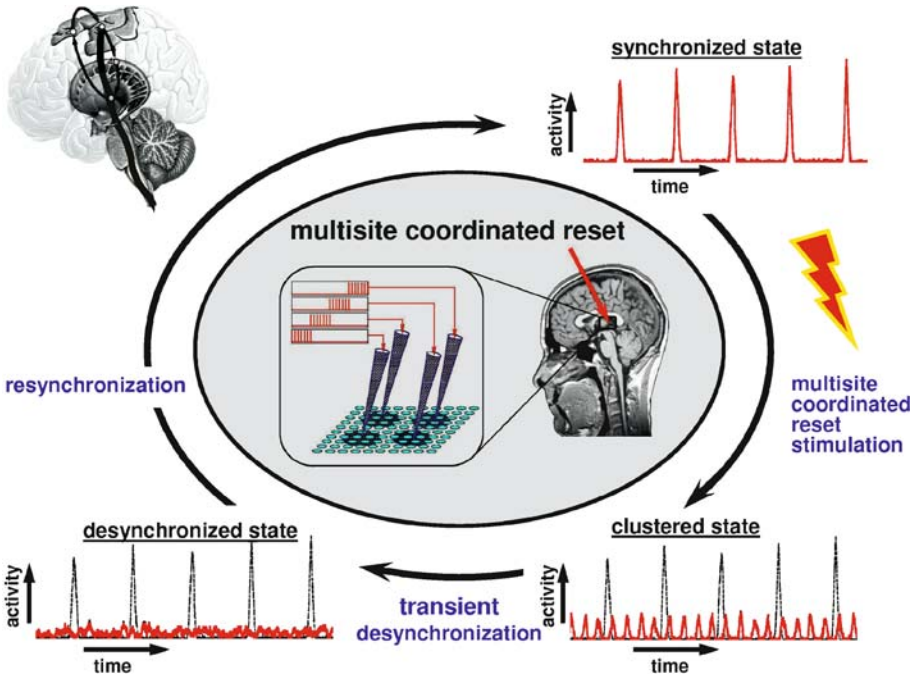

Brain Pacemaker, Figure 4
A general scheme of the stimulation with multisite coordinated reset. Desynchronized firing of neurons is maintained by repetitive administration of multisite coordinated reset stimuli intermingled with epochs of no stimulation

is close to zero (modulo 2π) [17,25,33,46,86]. The collective firing can be illustrated with the *relative number of neurons producing an action potential or burst* at time t given by

$$n_{\text{fire}}(t) = \frac{\text{number of neurons with } \cos\psi_j > 0.99}{N} . \quad (3)$$

$0 \leq n_{\text{fire}}(t) \leq 1$ for all t. $n_{\text{fire}}(t) = 0$ means that no neuron fires/bursts, while all neurons fire/burst at time t if $n_{\text{fire}}(t) = 1$. Varying the threshold parameter 0.99 in a reasonable range does not change the results. As shown in Fig. 4, stimulation starts when the neurons are synchronized and the collective firing demonstrates high-amplitude rhythmic oscillations (upper-right insert in Fig. 4). After a few periods of stimulation the oscillatory population is set in a cluster state (bottom-right insert in Fig. 4). Then the stimulation is switched off and the ensemble returns to a synchronized state passing on this way through a uniform desynchronized state (bottom-left insert in Fig. 4). And the procedure is repeated such that the ensemble is kept in a transient desynchronized state. The relaxation to a cluster state is due to the system being attracted by the center manifold as characterized by Eq. (2). By imposing a cluster state, the stimulation does only half of the desynchronizing work. The rest, namely approaching a uniformly desynchronized state, is done by the system itself. In this way the coupling, which causes the synchronization, is used for improving the desynchronizing effect. In the course of the post-stimulus transient R_1 and according to (2) also $R_2, R_3, \ldots$ recover again. The system finally reaches its stable in-phase synchronized state again. In summary, by shifting the system into an unstable cluster state, the system reacts by automatically running through a desynchronized state. Finally, the system reverts back to the synchronized state, if left unperturbed.

The effectively desynchronizing multisite coordinated reset can be used to block the resynchronization. For this, the repetitive stimulus administration can be organized either regardless of the state of the stimulated ensemble (open-loop control) or in a demand-controlled way (closed-loop control), where the following three different control strategies can be utilized:

(i) *Periodic administration of coordinated reset stimuli*: The most simple, open-loop type of stimulation is a periodic administration of coordinated reset stimuli.
(ii) *Demand-controlled timing of the administration of identical stimuli*: Whenever the population tends to resynchronize, the same stimulus is administered

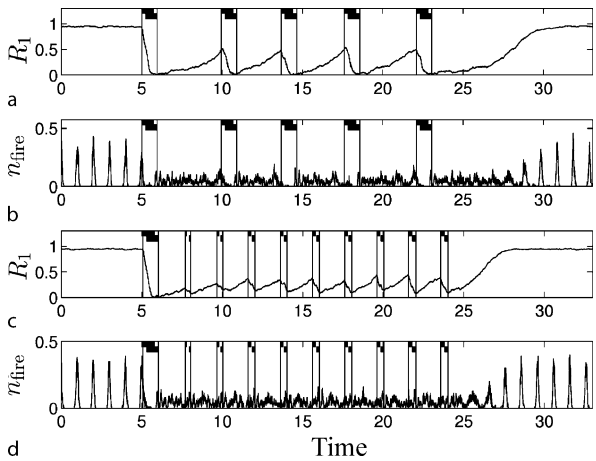

Brain Pacemaker, Figure 5
Time course of R_1 from (1) (a, c) and of n_{fire} from (3) (b, d) during different types of stimulation. *Demand-controlled timing of stimulus administration* (a, b): As soon as the amplitude R_1 of the recovering order parameter reaches the value of 0.5, the stimulus is administered again. *Periodical stimulation with demand-controlled length of HF pulse train* (c, d): The stimulus is administered periodically, where the length of the HF pulse trains is adapted to R_1 according to (4) with $M_{\max} = 15$ and $M_{\min} = 0$

(Fig. 5). The stronger the synchronization among the neurons is, the more often a stimulus has to be administered to maintain an uncorrelated firing. In an experimental application one has to observe the synchronized oscillation during a sufficiently long period of time in order to perform a frequency analysis which yields the period T of the population in the absence of stimulation and, thus, the critical stimulation parameter τ (the time delay between the two pairs of HF pulse trains, see Fig. 3).

(iii) *Periodically administered HF pulse trains of demand-controlled length*: The stimuli are periodically administered with offset times $t_k = k\nu T$, where $k = 0, 1, 2, 3, \ldots$ is the index labeling the different stimuli, $T = \tilde{T} + \varepsilon$ is a time interval in the range of the period $\tilde{T}$ of the population without stimulation, and ν is a small integer such as 2 or 3. This means that a 1:ν entrainment of the four sub-populations is performed, where the spontaneous frequency of the neurons is approximately ν times larger compared to the frequency of stimulus administration. The smaller $|\varepsilon|$, the smaller is the stimulation strength necessary to achieve an entrainment.

The closed-loop variants (ii) and (iii) require that the ensembles activity can be measured appropriately. Here, either the start times of identical coordinated reset stimuli or the length of periodically administered stimuli are calculated from the values of R_1. For example, in the case (ii) the stimulation is started if R_1 becomes larger than a certain threshold (Fig. 5, upper two plots), whereas in the case (iii) the stimulation period is longer for larger values of R_1 measured at the onset of the stimulation (Fig. 5, bottom two plots). In the latter case the length of the HF pulse trains increases linearly between a minimal value $M_{\min}$ and a maximal value $M_{\max}$ of single-pulses (except for rounding), where the latter is initially used for desynchronizing the fully synchronized population. R_1 is measured at times $t'_k = t_k - t_{\max}$, where $t_{\max}$ is the maximal duration of a HF pulse train (containing $M_{\max}$ single-pulses). $R_1(t'_k)$ determines the number of pulses of the HF pulse trains administered via each of the stimulation sites of the kth stimulus according to

$$M_k = \min\left\{\left[\frac{R_1(t'_k)(M_{\max} - M_{\min})}{R_1(t_0)}\right]_{\mathbb{Z}} + M_{\min}, M_{\max}\right\}, \tag{4}$$

where $k = 0, 1, 2, 3, \ldots$, $[x]_{\mathbb{Z}}$ stands for rounding x to the nearest integer, and $\min\{x_1, x_2\}$ stands for the minimum of $\{x_1, x_2\}$. The kth stimulus ends precisely at time $t_k = k\nu T$, whereas it starts somewhere between t'_k (for $M_k = M_{\max}$) and t_k (for $M_k = M_{\min} = 0$), depending on its duration. If the suppression of R_1 is not sufficient one may (i) choose a larger intensity of stimulation, (ii) increase $M_{\min}$, (iii) administer the stimuli at a higher rate, i. e. decrease ν, so that the inter-stimulus interval $t_{k+1} - t_k = \nu T$ gets smaller, and/or (iv) increase the duration of each single-pulse of the pulse trains. The feedback value of R_1 can also be evaluated before time t'_k, especially in case of a slow order parameter dynamics (i. e. when the synchronization is weak with respect to the noise). One could also use the mean of R_1 in a period of evaluation.

Applying the standard, permanent HF pulse train stimulation [5,10] (in a first approximation) corresponds to stimulating each neuron with the same HF pulse train. During a permanent HF stimulation a high-frequency entrainment of the order parameter Z_1 captures Z_1 in a small portion of the complex plane [90], so that the individual neurons' firing is stopped, but no desynchronization occurs. In contrast, during stimulation R_1 can be even larger compared to its pre-stimulus level, and after stimulation the synchronous firing continues immediately. To suppress the firing with such a simple pulse train persistently, it has to be administered permanently. The number of single-pulses used to suppress the firing in the case of the standard permanent HF pulse train stimulation is

about 5 and 8 times larger than that used for blocking the resynchronization in Figs. 5a,b and 5c,d, respectively. This illustrates the effectiveness of the demand-controlled multisite coordinated reset stimulation. The latter can effectively desynchronize stimulated oscillators with a significantly smaller amount of stimulation current compared to the standard permanent HF pulse-train stimulation. Coordinated reset stimulation has been experimentally verified in a hybrid neuroelectronic system of paddlefish electroreceptors [57].

Linear Multisite Delayed Feedback

Similarly as in the case of the coordinated reset stimulation, the linear multisite delayed feedback [37,38,39,40] is administered via several stimulation sites, e. g. via four sites as illustrated in Fig. 6 (left plot). Individual stimulation signals $S_m(t)$ administered via each of the stimulation sites are derived from the delayed mean field of the stimulated ensemble using different time delays for different stimulation signals. The mean field characterizes the collective macroscopic dynamics of the oscillators and can be viewed as the ensemble average of the signals of individual oscillators $Z(t) = N^{-1}\sum_{j=1}^{N} z_j(t)$, where $z_j(t)$, $j = 1, \ldots, N$ are the signals of the individual elements of the ensemble.

For n stimulation sites, the stimulation signals are calculated as $S_m(t) = KZ(t - \tau_m)$, $m = 1, \ldots, n$, where the values of delay τ_m, for example, for $n = 4$ are calculated from the following relation

$$\tau_m = \frac{11 - 2(m-1)}{8}\tau\,, \quad m = 1, 2, 3, 4\,. \tag{5}$$

The delays τ_m are symmetrically distributed with respect to the main delay τ, where the smallest time delay between neighboring stimulation sites is chosen as $\tau/4$. In the case $\tau = T$ (mean period of the ensemble), the delays τ_m are uniformly distributed over the mean period T. In another realization, instead of four delays τ_m, $m = 1, \ldots, 4$ one can use only two of them, e. g., τ_1 and τ_2. One can put $\tau_3 = \tau_1$ and $\tau_4 = \tau_2$, where the polarity of the stimulation signals $S_3(t)$ and $S_4(t)$ is reversed: $S_3(t) = -S_1(t)$ and $S_4(t) = -S_2(t)$. Assuming that the mean field of the ensemble uniformly oscillates with period $T = \tau$, the alternating polarity of the signal corresponds to a shift in time by half a period. Therefore, under this condition the stimulation signal $S_3(t) = -S_1(t) = -KZ(t - \tau_1)$ approximates the stimulation signal $S_1(t + \tau/2)$ which is shifted in time by half of the period, which, in turn, is equal to $KZ(t - \tau_3)$, where τ_3 is calculated according to Eq. (5). Analogous arguments are applicable to the stimulation signal $S_4(t) = -S_2(t) = -KZ(t - \tau_2)$.

If the phase $\Psi(t)$ of the mean field $Z(t)$ uniformly rotates with a constant frequency $\Omega = 2\pi/\tau$, the phases $\Psi_m(t) = \Psi(t - \tau_m)$ of the stimulation signals $S_m(t)$ are distributed uniformly over the unit circle as illustrated in

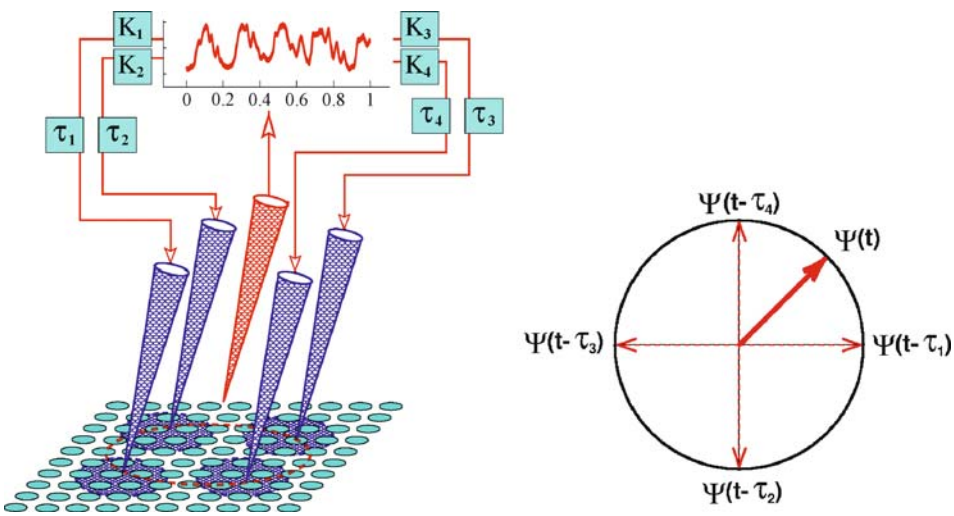

Brain Pacemaker, Figure 6

(*Left plot*) The macroscopic activity (mean field) of the controlled neuronal population is measured, delayed, amplified and fed back in a spatially coordinated way via several stimulation sites using different delays for different stimulation sites. (*Right plot*) Distribution of the delayed phases $\Psi(t - \tau_m)$ of the stimulation signals $S_m(t)$ of the multisite delayed feedback for four stimulation sites and for $\tau = T$ in Eq. (5)

Fig. 6 (right plot). Then the phases $\psi_j(t)$ of the stimulated neuronal sub-population assigned to the stimulation site m are attracted to the phase $\Psi(t-\tau_m)$ of the corresponding stimulation signal. Hence, the phases of all oscillators stimulated with the multisite delayed feedback become redistributed symmetrically on the circle $(0, 2\pi)$ in a cluster state. The order parameter $R_1(t)$ gets thus minimized. Depending on the value of delay τ, the stimulation can induce different cluster states in the stimulated ensemble, where the corresponding order parameter R_m attains large values.

As shown in Fig. 7, the in-phase synchronization in the stimulated ensemble is effectively suppressed (for time $t > 200$, where both coupling and stimulation are switched on), where the order parameter $R_1(t) = |Z_1(t)|$ attains small values (Fig. 7, red curve). This indicates a symmetrical redistribution of the phases $\psi_j(t)$ over the unit circle. For the parameter τ close to the mean period T of the stimulation-free ensemble a four cluster state is induced by the stimulation, where the amplitudes of the cluster variables $|Z_1|$ and $|Z_2|$ are small, whereas $|Z_4|$ is relatively large (Fig. 7a). In the subplot, where four trajectories from each of the stimulated sub-populations are depicted, the emerging four-cluster state induced by linear multi-site delayed feedback is illustrated. For τ closer to, for example, $2T$ the stimulation induces a two-cluster state, where $|Z_1|$ is small, whereas $|Z_2|$ and $|Z_4|$ are large

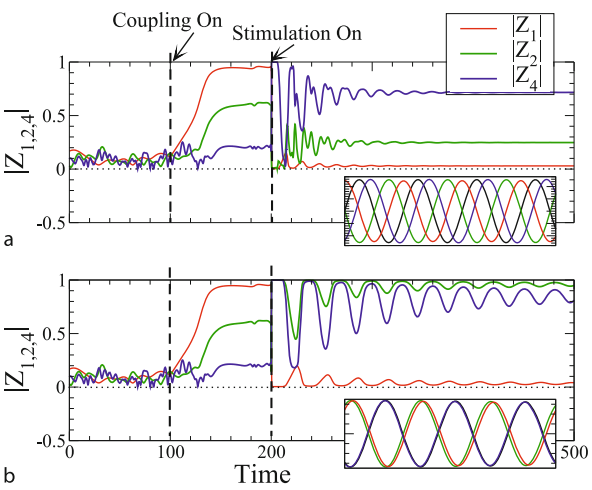

Brain Pacemaker, Figure 7

Control of synchronization by multisite delayed feedback stimulation. Time courses of the the amplitudes of the cluster variables $|Z_1|$, $|Z_2|$, and $|Z_4|$ are depicted by *red*, *green*, and *blue curves*, respectively. In the subplots four trajectories from each of four stimulated sub-populations assigned to each of four different stimulation sites are shown for $t \in (320, 340)$. Parameter $\tau = T$ in a and $\tau = 2T$ in b

(Fig. 7b). The oscillators thus split into two clusters, which is also illustrated in the subplot in (Fig. 7b).

Linear multi-site delayed feedback robustly suppresses the in-phase synchronization as shown in Fig. 8a, where the order parameter R_1 attains small values for a broad range of parameters τ and K. On the other hand, depending on the values of parameters, either the second order parameter R_2 (Fig. 8b) attains relatively large values (e. g., for $\tau \approx 2T$) indicating a two-cluster state, or, if the values of R_2 become small (e. g., for $\tau \approx T$), the relatively large values of the fourth order parameter R_4 (Fig. 8c) indicate a four-cluster state. Therefore, the whole stimulated population splits into two or four distinct sub-populations. Within the phase clusters the individual oscillators have phases close to each other, while the different phase clusters are equidistantly distributed within the cycle. Hence, depending on the values of the parameters τ and K, linear multi-site delayed feedback with four stimulation sites may cause either a two-cluster state, where R_1 is close to zero and R_2 is maximal, or a four-cluster state, where both R_1 and R_2 are close to zero, but R_4 is maximal. The cluster states become less pronounced and the phases redistribute on the circle even more uniformly if a local coupling as well as spatially decaying profile of the current spread is taken into account [37,40].

In Fig. 8d a similar two-parameter diagram for the averaged order parameter $\langle R_1(t)\rangle$ is presented for *linear single-site delayed feedback*, suggested for synchronization control in Refs. [67,68]. The stimulation is performed via one stimulation electrode in such a way that all oscillators of the ensemble (in a first approximation) receive the same stimulation signal $S(t)$. In this case the stimulation signal $S(t)$ attains the form $S(t) = KZ(t-\tau)$. For the stimulation with linear single-site delayed feedback in a corresponding two-parameter diagram (Fig. 8d) islands of perfect desynchronization are complemented by areas of stimulation-enhanced synchronization. In the limit $N \to \infty$ the order parameter $R_1 = 0$ in the desynchronization regions, where the phases are uniformly distributed on the circle $(0, 2\pi)$ [67,68]. This is the state of complete desynchronization, where the stimulated oscillators rotate with different frequencies indicating an absence of any cluster state whatsoever.

The important property of the stimulation with the linear multi- and single-site delayed feedback is the inherited demand-controlled character of the methods. As soon as the desired desynchronized state is achieved, the value of the order parameter $R_1(t)$, i. e., the amplitude of the mean field becomes small. Along with the order parameter, in the desynchronized state the amplitude of the stimulation signal $S(t)$ vanishes as well. The stimulation with

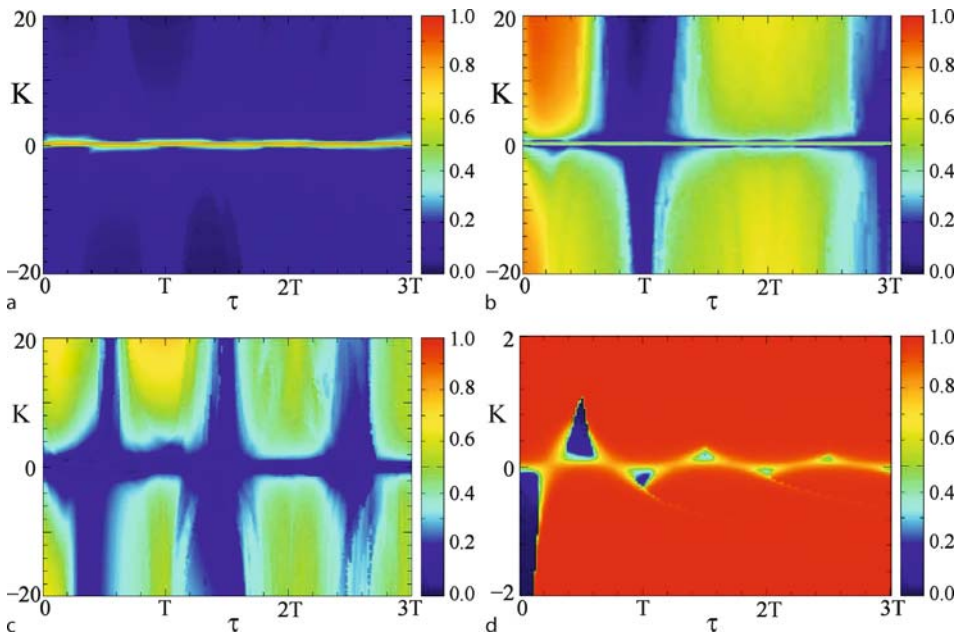

Brain Pacemaker, Figure 8

Impact of the multisite delayed feedback stimulation versus parameters τ and stimulus amplification K. The time-averaged order parameters $\langle R_1 \rangle$, $\langle R_2 \rangle$, and $\langle R_4 \rangle$ are depicted in a, b, and c, respectively, and are encoded in color ranging from 0 (*blue*) to 1 (*red*). In plot d the impact of the linear single-site delayed feedback on the oscillatory population is illustrated, where the values of the order parameter $\langle R_1 \rangle$ are depicted in color versus parameters τ and K

linear multi- and single-site delayed feedback thus represents a noninvasive control method for the desynchronization of coupled oscillators. The stimulated ensemble is then subjected to a highly effective control at a minimal amount of stimulation force.

Nonlinear Delayed Feedback

As for the case of the linear single-site delayed feedback, for the stimulation with nonlinear delayed feedback only one registering and one stimulating site is required, see Fig. 9a. All stimulated oscillators receive the same stimulation signal $S(t)$ which is constructed from the measured mean field of the ensemble. It is assumed that the measured mean field $Z(t)$ of the ensemble has the form of a complex analytic signal $Z(t) = X(t) + iY(t)$, where $X(t)$ and $Y(t)$ are the real and imaginary parts of $Z(t)$, respectively. In the case if only the real part $X(t)$ of the mean field is measured, the imaginary part can be calculated, e. g., with the help of Hilbert transform [60]. The stimulation signal is then constructed by a nonlinear combination of a delayed complex conjugate mean field with the instantaneous mean field [62,63,64,98]

$$S(t) = KZ^2(t)Z(t-\tau)\,, \tag{6}$$

where K is a stimulus amplification parameter, τ is a time delay, and the asterisk denotes complex conjugacy.

The desynchronizing effect of the stimulation with nonlinear delayed feedback is illustrated in Fig. 9b,c, where the time courses of the order parameter $R_1(t)$ (red curves) and the amplitude of the stimulation signal $|S(t)|$ (black curves) are plotted for two values of delay $\tau = T/2$ (Fig. 9b) and $\tau = T$ (Fig. 9c), where T is the mean period of the stimulation-free population. The onset of stimulation at $t = 550$ results in desynchronization of the stimulated oscillators and the order parameter reaches the values of approximately the same order of magnitude as in the uncoupled regime ($t < 400$). This indicates a high level of desynchronization. The stimulation does not destroy the normal oscillatory activity of the individual phase oscillators. In the insets in Figs. 9b and 9c individual trajectories of two selected oscillators of the stimulated ensemble are plotted. The stimulated oscillators rotate with different individual frequencies just as in the coupling- and stimulation-free regime.

As soon as a desynchronized state is achieved, the stimulation force declines and the stimulated system is subjected to a highly effective control with a minimal amount of stimulation force. Also, as soon as a resynchronization occurs, the mean field starts to exhibit large-am-

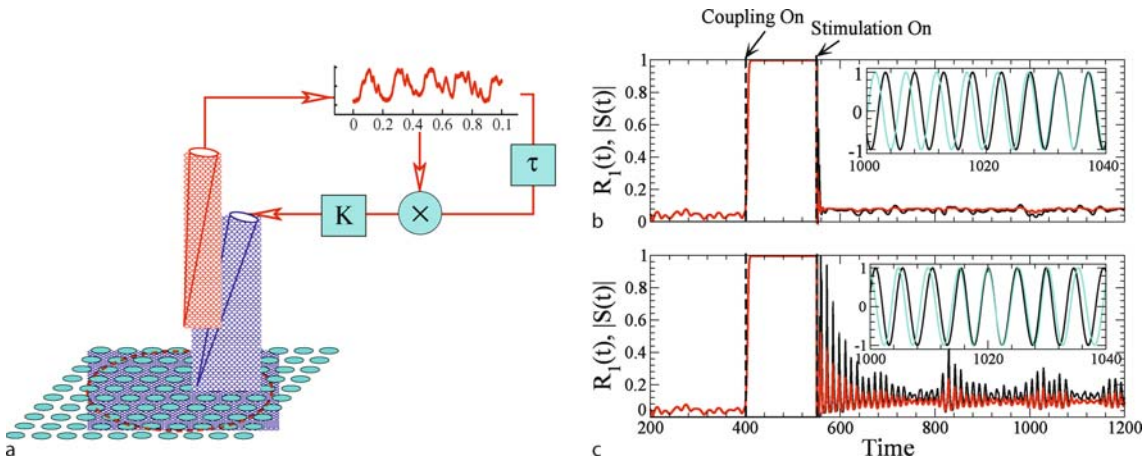

Brain Pacemaker, Figure 9

a The macroscopic activity (mean field) of the controlled population is measured, delayed, nonlinearly combined with the instantaneous mean field, amplified, and fed back via a single stimulation site. b, c Desynchronization of strongly synchronized oscillators by nonlinear delayed feedback. Time courses of the order parameter $R_1(t)$ (*red curves*) and the amplitude of the stimulation signal $|S(t)|$ (*black curves*) are plotted for delay b $\tau = T/2$ and c $\tau = T$. In the subplots two selected trajectories of oscillators are depicted in the stimulated regime. T is the mean period of the stimulation-free ensemble

plitude oscillations and the stimulation signal increases its amplitude and brings the ensemble back to a desynchronized state. This demand-controlled character of the nonlinear delayed feedback is particularly well illustrated in Fig. 9c, where the onsets of resynchronization (increase of $R_1(t)$, red curve) at times around $t \approx 850$, 1050, and 1200 lead to an increase of the amplitude of the stimulation signal $|S(t)|$ (black curve), which in turn results in a suppression of the resynchronization.

The impact of the nonlinear delayed feedback on the stimulated oscillators is twofold. On one hand, the stimulation can effectively desynchronize even strongly interacting oscillators for a range of large values of the stimulus amplification K, see Fig. 10, left plot. This effect is very robust with respect to the variation of the delay τ and, as a result, with respect to the variation of the mean frequency Ω of the stimulated ensemble. On the other hand, in a weakly coupled ensemble the stimulation can induce synchronization in island-like regions of small values of the stimulus amplification K complemented by domain of desynchronization, see Fig. 10, right plot.

An increase of the stimulus amplification parameter K results in a gradual decay of the order parameter R_1, which indicates an onset of desynchronization in the stimulated

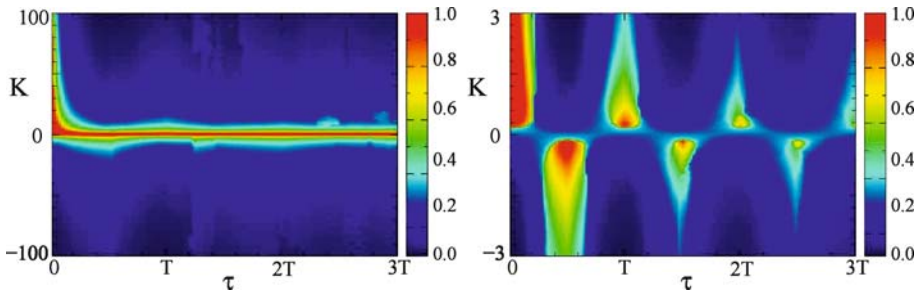

Brain Pacemaker, Figure 10

Nonlinear delayed feedback: Stimulation-induced desynchronization (*left plot*) and synchronization (*right plot*) in ensembles of strongly (*left plot*) or weakly (*right plot*) coupled oscillators. Note, nonlinear delayed feedback is a robust desynchronization technique, which enables to effectively counteract intermediate and strong coupling (*left plot*). However, in ensembles with weak coupling nonlinear delayed feedback with particularly chosen parameters may also enhance synchronization (*right plot*). The time-averaged values of the order parameter $\langle R_1 \rangle$ are encoded in color ranging from *red* (synchronization) to *blue* (desynchronization) versus delay τ and stimulus amplification K

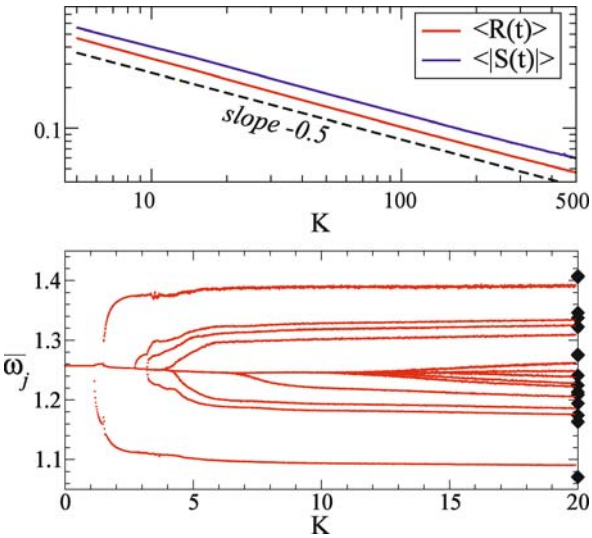

Brain Pacemaker, Figure 11
Time-averaged order parameter $\langle R_1 \rangle$ and amplitude of the stimulation signal $\langle |S(t)| \rangle$ (*upper plot, log-log scale*) and the observed individual frequencies $\overline{\omega}_j$ of the stimulated oscillators (*bottom plot*) versus the stimulus amplification parameter K. *Black diamonds* depict values of the natural frequencies of the stimulated oscillators

ensemble. Simultaneously, the amplitude of the stimulation signal $|S(t)|$ decays as well, indicating a demand-controlled character of the nonlinear delayed feedback stimulation. For a fixed delay $\tau > 0$ the order parameter and the amplitude of the stimulation signal decay according to the following power law as $|K|$ increases:

$$R_1 \sim |K|^{-1/2} , \quad |S| \sim |K|^{-1/2} , \tag{7}$$

see Fig. 11, upper plot. The desynchronization transition for increasing K also manifests itself in a sequence of frequency-splitting bifurcations, where the observed individual frequencies $\overline{\omega}_j = \langle \dot{\psi}_j \rangle$ of the stimulated oscillators split, one after another from the mean frequency Ω as K increases and approach the natural frequencies of the unperturbed oscillators (black diamonds in Fig. 11, bottom plot).

For large values of K all stimulated oscillators rotate with different frequencies which approach the natural frequencies ω_j of the oscillators. The oscillators thus exhibit a uniform desynchronized dynamics without any kind of cluster states. Simultaneously, depending on the values of the delay τ, the nonlinear delayed feedback can significantly change the mean frequency Ω, i. e., the frequency of the mean field $Z(t)$ of the stimulated ensemble [62,63]. The macroscopic dynamics can thus be either accelerated or slowed down, whereas the individual dynamics remains close to the original. This opens an approach to frequency control of the oscillatory population stimulated with nonlinear delayed feedback.

Proportional-Integro-Differential Feedback

For a particularly difficult situation, where the measurement and stimulation are not possible at the same time and at the same place, there is another control method which is based on a proportional-integro-differential (PID) feedback. The scheme of this stimulation protocol is sketched in Fig. 12a. The controlled ensemble of N coupled oscillators is split into two separate sub-populations of N_1 and $N_2 = N - N_1$ oscillators, one being exclusively measured and the other being exclusively stimulated. In this way a separate stimulation-registration setup is realized, where the recording and stimulating sites are spatially separated and the measured signal is not corrupted by stimulation artifacts. The observed signal is considered to be the mean field $Z^{(1)}$ of the measured sub-population. Below the main attention will be paid to the proportional-differential (PD) feedback only (for more details, see [65]). Then, the stimulation signal $S(t)$ administered to the second, stimulated sub-population is constructed as

$$S(t) = PZ^{(1)}(t) + D\dot{Z}^{(1)}(t) , \tag{8}$$

where the parameters P and D define the strengths of the proportional and differential feedback, respectively.

The effect of the stimulation with PD feedback is illustrated in Fig. 12b. As the strength of the feedback (parameters P and D) increases the stimulation results in a complete desynchronization of both, measured and stimulated sub-populations. The threshold of the onset of desynchronization depends on the relative splitting $N_1 : N_2$ of the oscillators between sub-populations and on the mean frequency Ω: The threshold is larger for smaller number of oscillators N_2 in the stimulated populations or for larger frequency Ω. The later dependence can be eliminated if an integral component is included in the stimulation signal, see [65]. Moreover, if the coupling in the ensemble is rather weak, the desynchronization can be achieved by applying the proportional feedback only. In contrast, in the case of strong coupling the stimulation signal must also contain the differential feedback for robust desynchronization. As illustrated in the 2-parameter diagrams in Fig. 12b (bottom plots), there exists a certain threshold in parameters P and D values, where the stimulation with PID feedback desynchronizes both sub-populations in the target ensemble: Stimulated sub-population (Fig. 12b, bottom right plot) and also measured, non-stimulated sub-population (Fig. 12b, bottom left plot).

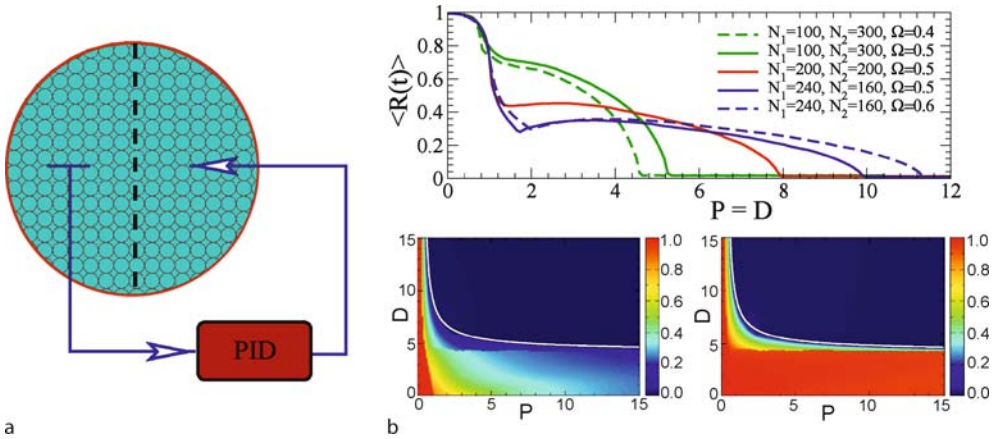

Brain Pacemaker, Figure 12

PID Control: a The mean field is measured in one part of the controlled ensemble and, after processing according to proportional-integro-differential (PID) algorithm, is administered to the other part of the ensemble. b *Upper plot*: The order parameter of the whole population versus the strength of the PD feedback (with $P = D$) for different splitting $N_1 : N_2$ and different mean frequencies Ω. *Bottom plots*: The values of the order parameter (encoded in color) of the measured sub-population (*left*) and stimulated sub-population (*right*) versus stimulation parameters P and D. The white curve is the parameter threshold of the onset of desynchronization in sub-populations obtained in [65]

In this sense the PID feedback stimulation method appears to be very effective even in such a complicated stimulation protocol of separate stimulation-registration setup.

Plasticity

Plasticity is a fundamental property of the nervous system: In order to learn and to adapt to sensory inputs, neurons continuously adapt the strength of their synaptic connections in relation to the mutual timing properties of their firing or bursting [1,13,18,22,42,44,51,78,100,105]. However, plasticity may not only lead to desired learning and optimization processes. Rather neuronal populations can learn pathologically strong interactions which may lead, e. g., to the emergence of epilepsies [56,79]. This is well-known from the so-called kindling phenomenon [24], where preparatory stimulation induces the spontaneous production of epileptic seizures without gross morphological changes [56].

Recently, theoretical studies have initiated an approach which is targeted on unlearning pathologically strong synaptic interactions by desynchronizing brain stimulation [36,49,83,97]. This approach exploits plasticity in two different ways: On the one hand, due to plasticity desynchronizing stimulation may decrease the strength of the neurons' synapses by decreasing the rate of coincidences. On the other hand, neuronal networks with synaptic plasticity may exhibit bi- or multistability [36,49,73,83,97]. Accordingly, by decreasing the mean synaptic weight, desynchronizing stimulation may shift a neuronal population from a stable synchronized (pathological) state to a stable desynchronized (healthy) state, where the neuronal population remains thereafter, if left unperturbed. In Fig. 13 an exemplary simulation of a network of 200 model neurons is displayed, for further details concerning the mathematical model we refer to [36,83]. Induced by appropriate stimulation protocols a switching between the different stable states is realizable. Starting from a desynchronized state, associated with physiological system dynamics, low-frequency stimulation can induce a kindling of the interneuronal connectivities and causes a stabilization of the synchronized state. After stimulation offset, the system remains in the pathological state, see Fig. 13. In contrast, desynchronizing (MCR) stimulation results in an anti-kindling of the pathological connectivities and finally the physiological state is reestablished, see Fig. 13.

From a mathematical point of view, in a first approximation this situation may be illustrated by considering a double well potential, where each minimum corresponds to a stable attractor, surrounded by a basin of attraction (Fig. 14). The strongly synchronized state (*Pathol.* in Fig. 14) serves as a model for a disease state, whereas the uncorrelated or weakly synchronized state (*Healthy* in Fig. 14) is used as a model for a healthy state. As soon as the system, i. e. the neuronal population (illustrated by the ball in Fig. 14), enters a particular basin of attraction, it gets attracted by the corresponding attractor, so that it relaxes towards the corresponding minimum of the potential.

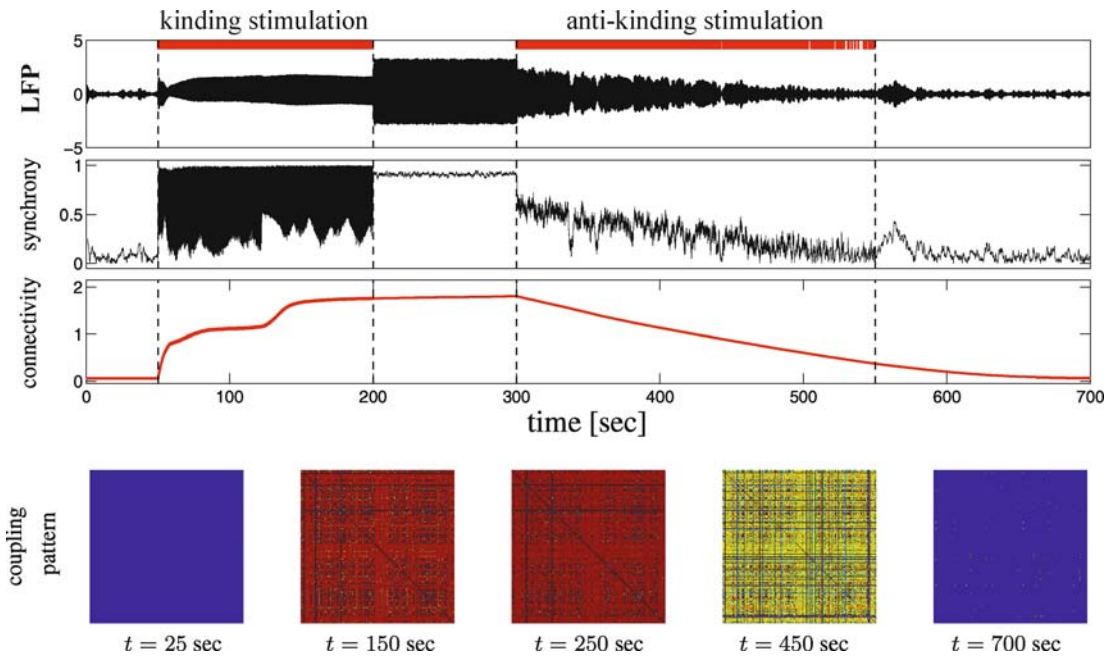

Brain Pacemaker, Figure 13

Effects of kindling and anti-kindling stimulation on a population of model neurons. Low-frequency stimulation is applied between 50 to 200 seconds and multi-site coordinated reset stimulation is applied between 300 to 550 seconds. The local field potential, the level of synchronization within the network and the mean connectivity is plotted (from *top* to *bottom*). Below five patterns representing the coupling topology of the network at different times are plotted. *Blue* (*red*) colors represent low (high) values of the interneuronal connectivity

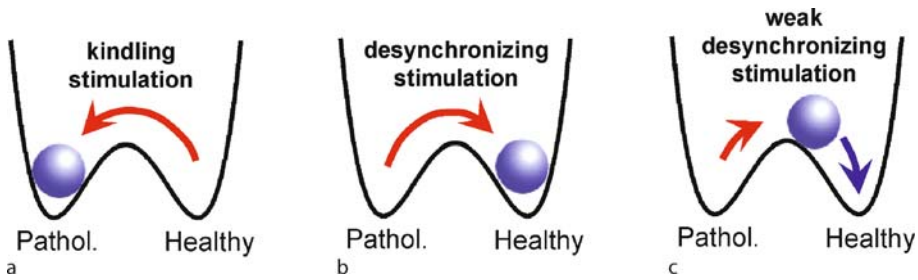

Brain Pacemaker, Figure 14

Anti-kindling and kindling stimulation can move the neuronal population from one attractor to another: Schematic plot of the attractors symbolizing pathological (*Pathol.*) or healthy (*Healthy*) dynamical model states. a Periodic, kindling stimulation (*red trajectory*) shifts the population from a healthy, desynchronized state (*Healthy*) to a pathological, synchronized state (*Pathol.*). b Conversely, desynchronizing stimulation shifts the population from a pathological state (*Pathol.*) to the healthy uncorrelated state (*Healthy*). b This anti-kindling is achieved by a desynchronizing stimulation of sufficient duration, so that after stimulus offset the population is close to the healthy state (*red trajectory*). c Alternatively, the same long-term anti-kindling effect can be achieved with a brief desynchronizing stimulation, which shifts the population to an intermediate state, which may still be connected with pronounced synchrony. However, since the intermediate state (*blue ball*) lies within the basin of attraction of a healthy state (*Healthy*) the population spontaneously relaxes to the healthy state without any further intervention (*blue trajectory*)

Appropriate stimulation protocols may shift the neuronal population from one state to another. Kindling stimulation of appropriate duration shifts the neuronal population from a desynchronized state close to a strongly synchronized state or at least into the basin of attraction of such a state (Fig. 14a, red trajectory from *Healthy* to *Pathol.*). Conversely, anti-kindling can be achieved by means of a desynchronizing stimulation which shifts the neuronal population close to the desynchronized state [36,83,97] (Fig. 14c, red trajectory from *Pathol.* to *Healthy*). However, with respect to the long-term anti-kindling outcome, it is even sufficient to move the neuronal

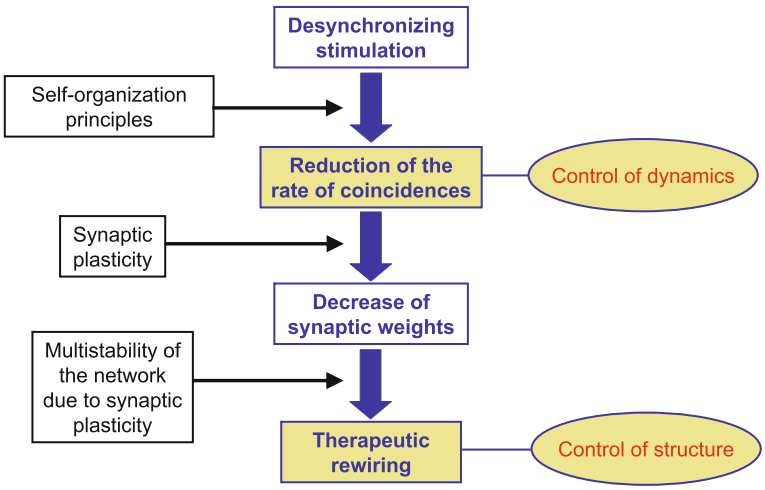

Brain Pacemaker, Figure 15
***Schematic illustration of how desynchronizing stimulation may cause an unlearning of pathological connectivity*: Desynchronizing stimulation reduces the overall rate of coincidences in the neuronal population. This is effectively achieved by using dynamic self-organization principles. The reduction of the rate of coincidences, in turn, reduces the synaptic weights. Because of the multistability not only strongly synchronized states, but also uncorrelated and weakly synchronized states are stable. Accordingly, the neuronal population stably remains in a desynchronized or weakly synchronized state, after having been shifted into the basin of attraction of that state by means of desynchronizing stimulation. Hence, a suitable control of the dynamics of the network leads to a modulation of its connectivity**

population from the synchronized state just into the basin of attraction of the desynchronized state (Fig. 14b, red trajectory from *Pathol.* to the intermediate state). After stimulus offset there may still be pronounced synchrony, but being captured within the basin of attraction of the desynchronized state, the neuronal population spontaneously relaxes to the desynchronized state, without further intervention (Fig. 14c, blue trajectory from the intermediate state to *Healthy*). It has to be considered that the short as well as the long desynchronizing stimulation in this schematic illustration have the same long-term anti-kindling outcome.

In PD neuronal populations of the basal ganglia are strongly synchronized [9,70,99] and synaptic plasticity results in a further amplification of the synchronized activity by a strengthening of the synaptic connections [59]. Properly designed electrical stimulation may be used to break this vicious circle and to induce an anti-kindling [36,83,97], which finally might reestablish the normal level of connectivity, associated with a mostly uncorrelated neuronal activity (Fig. 15).

Summary

High-frequency deep brain stimulation is the golden standard for the treatment of medically refractory movement disorders [5,101]. To overcome limitations of this type of therapy, a model-based development of novel deep brain stimulation techniques has been initiated [86]. The ingredients of this approach are dynamic neuronal self-organization principles combined with fundamental plasticity rules of the nervous system (Fig. 15) [36,83,96,97]. In forthcoming studies the mathematical modeling will be refined to incorporate further anatomical and physiological details, for instance, contributions of glial cells [75]. By the same token, control techniques will be optimized and further developed. As currently done for multisite coordinated reset stimulation, clinical studies will be performed to evaluate the therapeutic effects of the novel stimulation techniques. This interdisciplinary endeavor is performed to finally provide superior therapies for patients with neurological or psychiatric diseases.

Bibliography

1. Abbott L, Nelson S (2000) Synaptic plasticity: Taming the beast. Nat Neurosci 3:1178–1183
2. Alberts WW, Wright EJ, Feinstein B (1969) Cortical potentials and parkinsonian tremor. Nature 221:670–672
3. Andres F, Gerloff C (1999) Coherence of sequential movements and motor learning. J Clin Neurophysiol 16(6):520–527
4. Benabid A, Pollak P, Louveau A, Henry S, de Rougemont JJ (1987) Combined (thalamotomy and stimulation) stereotac-

tic surgery of the VIM thalamic nucleus for bilateral Parkinson disease. Appl Neurophysiol 50(1-6):344–346
5. Benabid AL, Pollak P, Gervason C, Hoffmann D, Gao DM, Hommel M, Perret JE, de Rougemount J (1991) Longterm suppression of tremor by chronic stimulation of ventral intermediate thalamic nucleus. Lancet 337:403–406
6. Benabid AL, Benazzous A, Pollak P (2002) Mechanisms of deep brain stimulation. Mov Disord 17:73–74
7. Benabid A-L, Wallace B, Mitrofanis J, Xia R, Piallat B, Chabardes S, Berger F (2005) A putative gerneralized model of the effects and mechanism of action of high frequency electrical stimulation of the central nervous system. Acta neurol Belg 105:149–157
8. Beurrier C, Bioulac B, Audin J, Hammond C (2001) High-frequency stimulation produces a transient blockade of voltage-gated currents in subthalamic neurons. J Neurophysiol 85(4):1351–1356
9. Beurrier C, Garcia L, Bioulac B, Hammond C (2002) Subthalamic nucleus: A clock inside basal ganglia? Thalamus Relat Syst 2:1–8
10. Blond S, Caparros-Lefebvre D, Parker F, Assaker R, Petit H, Guieu J-D, Christiaens J-L (1992) Control of tremor and involuntary movement disorders by chronic stereotactic stimulation of the ventral intermediate thalamic nucleus. J Neurosurg 77:62–68
11. Brice J, McLellan L (1980) Suppression of intention tremor by contingent deep-brain stimulation. Lancet 1(8180):1221–1222
12. Daido H (1992) Order function and macroscopic mutual entrainment in uniformly coupled limit-cycle oscillators. Prog Theor Phys 88:1213–1218
13. Debanne D, Gahweiler B, Thompson S (1998) Long-term synaptic plasticity between pairs of individual CA3 pyramidal cells in rat hippocampus slice cultures. J Physiol 507:237–247
14. Deuschl G, Schade-Brittinger C, Krack P, Volkmann J, Schäfer H, Bötzel K, Daniels C, Deutschländer A, Dillmann U, Eisner W, Gruber D, Hamel W, Herzog J, Hilker R, Klebe S, Kloß M, Koy J, Krause M, Kupsch A, Lorenz D, Lorenzl S, Mehdorn H, Moringlane J, Oertel W, Pinsker M, Reichmann H, Reuß A, Schneider G-H, Schnitzler A, Steude U, Sturm V, Timmermann L, Tronnier V, Trottenberg T, Wojtecki L, Wolf E, Poewe W, Voges J (2006) A randomized trial of deep-brain stimulation for parkinsons disease. N Engl J Med 355:896–908
15. Dolan K, Majtanik M, Tass P (2005) Phase resetting and transient desynchronization in networks of globally coupled phase oscillators with inertia. Physica D 211:128–138
16. Elble RJ, Koller WC (1990) Tremor. John Hopkins University Press, Baltimore
17. Ermentrout B, Kopell N (1991) Multiple pulse interactions and averaging in systems of coupled neural assemblies. J Math Biol 29:195–217
18. Feldman D (2000) Timing-based LTP and LTD at vertical inputs to layer II/III pyramidal cells in rat barrel cortex. Neuron 27:45–56
19. Filali M, Hutchison W, Palter V, Lozano A, Dostrovsky JO (2004) Stimulation-induced inhibition of neuronal firing in human subthalamic nucleus. Exp Brain Res 156:274–281
20. Freund H-J (2005) Long-term effects of deep brain stimulation in parkinsons disease. Brain 128:2222–2223
21. Garcia L, D'Alessandro G, Fernagut P-O, Bioulac B, Hammond C (2005) Impact of high-frequency stimulation parameters on the pattern of discharge of subthalamic neurons. J Neurophysiol 94:3662–3669
22. Gerstner W, Kempter R, van Hemmen J, Wagner H (1996) A neuronal learning rule for sub-millisecond temporal coding. Nature 383:76–78
23. Gildenberg P (2005) Evolution of neuromodulation. Stereotactic Funct Neurosurg 83:71–79
24. Goddar G (1967) Development of epileptic seizures through brain stimulation at low intensity. Nature 214:1020–1021
25. Grannan ER, Kleinfeld D, Sompolinsky H (1993) Stimulus-dependent synchronization of neuronal assemblies. Neural Comp 5:550–569
26. Grill WM, McIntyre CC (2001) Extracellular excitation of central neurons: Implications for the mechanisms of deep brain stimulation. Thalamus Relat Syst 1:269–277
27. Haken H (1970) Laser theory vol XXV/2C. In: Flügge S (ed) Encyclopedia of physics. Springer, Berlin
28. Haken H (1977) Synergetics. An introduction. Springer, Berlin
29. Haken H (1983) Advanced synergetics. Springer, Berlin
30. Haken H (1996) Principles of brain functioning. A synergetic approach to brain activity, behavior, and cognition. Springer, Berlin
31. Haken H (2002) Brain dynamics. Synchronization and activity patterns in pulse-coupled neural nets with delays and noise. Springer, Berlin
32. Haken H, Kelso J, Bunz H (1985) A theoretical model of phase transitions in human hand movements. Biol Cybern 51:347–356
33. Hansel D, Mato G, Meunier C (1993) Phase dynamics of weakly coupled Hodgkin–Huxley neurons. Europhys Lett 23:367–372
34. Hansel D, Mato G, Meunier C (1993) Phase reduction and neuronal modeling. Concepts Neurosci 4(2):193–210
35. Hashimoto T, Elder C, Okun M, Patrick S, Vitek J (2003) Stimulation of the subthalamic nucleus changes the firing pattern of pallidal neurons. J Neurosci 23(5):1916–1923
36. Hauptmann C, Tass PA (2007) Therapeutic rewiring by means of desynchronizing brain stimulation. Biosystems 89:173–181
37. Hauptmann C, Popovych O, Tass PA (2005) Delayed feedback control of synchronization in locally coupled neuronal networks. Neurocomputing 65–66:759–767
38. Hauptmann C, Popovych O, Tass PA (2005) Effectively desynchronizing deep brain stimulation based on a coordinated delayed feedback stimulation via several sites: A computational study. Biol Cybern 93:463–470
39. Hauptmann C, Popovych O, Tass PA (2005) Multisite coordinated delayed feedback for an effective desynchronization of neuronal networks. Stoch Dyn 5(2):307–319
40. Hauptmann C, Omelchenko O, Popovych OV, Maistrenko Y, Tass PA (2007) Control of spatially patterned synchrony with multisite delayed feedback. Phys Rev E 76:066209
41. Hauptmann C, Popovych O, Tass P (2007) Desynchronizing the abnormally synchronized neural activity in the subthalamic nucleus: A modeling study. Expert Rev Med Devices 4(5):633–650
42. Hebb D (1949) The organization of behavior. Wiley, New York
43. Kelso J (1995) Dynamic patterns: The self-organization of brain and behavior. MIT Press, Cumberland
44. Kilgard M, Merzenich M (1998) Cortical map reorganization enabled by nucleus basalis activity. Science 279:1714–1718
45. Kumar R, Lozano A, Sime E, Lang A (2003) Long-term follow-

up of thalamic deep brain stimulation for essential and parkinsonian tremor. Neurology 61:1601–1604
46. Kuramoto Y (1984) Chemical oscillations, waves, and turbulence. Springer, Berlin
47. Lenz F, Kwan H, Martin R, Tasker R, Dostrovsky J, Lenz Y (1994) Single unit analysis of the human ventral thalamic nuclear group. Tremor-related activity in functionally identified cells. Brain 117:531–543
48. Limousin P, Speelman J, Gielen F, Janssens M (1999) Multicentre European study of thalamic stimulation in parkinsonian and essential tremor. J Neurol Neurosurg Psychiatry 66(3):289–296
49. Maistrenko Y, Lysyansky B, Hauptmann C, Burylko O, Tass P (2007) Multistability in the Kuramoto model with synaptic plasticity. Phys Rev E 75:066207
50. Majtanik M, Dolan K, Tass P (2006) Desynchronization in networks of globally coupled neurons with dendritic dynamics. J Biol Phys 32:307–333
51. Markram H, Lübke J, Frotscher M, Sakmann B (1997) Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. Science 275:213–215
52. McIntyre C, Grill W, Sherman D, Thakor N (2004) Cellular effects of deep brain stimulation: Model-based analysis of activation and inhibition. J Neurophysiol 91:1457–1469
53. McIntyre CC, Savasta M, Goff KKL, Vitek J (2004) Uncovering the mechanism(s) of action of deep brain stimulation: Activation, inhibition, or both. Clin Neurophysiol 115:1239–1248
54. Milton J, Jung P (eds) (2003) Epilepsy as a dynamics disease. Springer, Berlin
55. Miocinovic S, Parent M, Butson C, Hahn P, Russo G, Vitek J, McIntyre C (2006) Computational analysis of subthalamic nucleus and lenticular fasciculus activation during therapeutic deep brain stimulation. J Neurophysiol 96:1569–1580
56. Morimoto K, Fahnestock M, Racine R (2004) Kindling and status epilepticus models of epilepsy: Rewiring the brain. Prog Neurobiol 73:1–60
57. Neiman A, Russell D, Yakusheva T, DiLullo A, Tass PA (2007) Response clustering in transient stochastic synchronization and desynchronization of coupled neuronal bursters. Phys Rev E 76:021908
58. Nini A, Feingold A, Slovin H, Bergmann H (1995) Neurons in the globus pallidus do not show correlated activity in the normal monkey, but phase-locked oscillations appear in the MPTP model of parkinsonism. J Neurophysiol 74:1800–1805
59. Nowotny T, Zhigulin V, Selverston A, Abarbanel H, Rabinovich M (2003) Enhancement of synchronization in a hybrid neural circuit by spike-timing dependent plasticity. J Neurosci 23:9776–9785
60. Pikovsky A, Rosenblum M, Kurths J (2001) Synchronization, a universal concept in nonlinear sciences. Cambridge University Press, Cambridge
61. Pliss V (1964) Principal reduction in the theory of stability of motion. Izv Akad Nauk SSSR Math Ser 28:1297–1324
62. Popovych OV, Hauptmann C, Tass PA (2005) Effective desynchronization by nonlinear delayed feedback. Phys Rev Lett 94:164102
63. Popovych OV, Hauptmann C, Tass PA (2006) Control of neuronal synchrony by nonlinear delayed feedback. Biol Cybern 95:69–85
64. Popovych OV, Hauptmann C, Tass PA (2006) Desynchronization and decoupling of interacting oscillators by nonlinear delayed feedback. Int J Bif Chaos 16(7):1977–1987
65. Pyragas K, Popovych OV, Tass PA (2007) Controlling synchrony in oscillatory networks with a separate stimulation-registration setup. Europhys Lett 80:40002
66. Rodriguez-Oroz M, Obeso J, Lang A, Houeto J, Pollak P, Rehncrona S, Kulisevsky J, Albanese A, Volkmann J, Hariz M, Quinn N, Speelman J, Guridi J, Zamarbide I, Gironell A, Molet J, Pascual-Sedano B, Pidoux B, Bonnet A, Agid Y, Xie J, Benabid A, Lozano A, Saint-Cyr J, Romito L, Contarino M, Scerrati M, Fraix V, Blercom NV (2005) Bilateral deep brain stimulation in Parkinsons disease: A multicentre study with 4 years follow-up. Brain 128:2240–2249
67. Rosenblum MG, Pikovsky AS (2004) Controlling synchronization in an ensemble of globally coupled oscillators. Phys Rev Lett 92:114102
68. Rosenblum MG, Pikovsky AS (2004) Delayed feedback control of collective synchrony: An approach to suppression of pathological brain rhythms. Phys Rev E 70:041904
69. Rubin J, Terman D (2004) High frequency stimulation of the subthalamic nucleus eliminates pathological thalamic rhythmicity in a computational model. J Comput Neurosci 16:211–235
70. Schnitzler A, Timmermann L, Gross J (2006) Physiological and pathological oscillatory networks in the human motor system. J Physiol Paris 99(1):3–7
71. Schuurman PR, Bosch DA, Bossuyt PM, Bonsel GJ, van Someren EJ, de Bie RM, Merkus MP, Speelman JD (2000) A comparison of continuous thalamic stimulation and thalamotomy for suppression of severe tremor. N Engl J Med 342:461–468
72. Schöner G, Haken H, Kelso J (1986) A stochastic theory of phase transitions in human hand movement. Biol Cybern 53:247–257
73. Seliger P, Young S, Tsimring L (2002) Plasticity and learning in a network of coupled phase oscillators. Phys Rev E 65:041906
74. Shen K, Zhu Z, Munhall A, Johnson SW (2003) Synaptic plasticity in rat subthalamic nucleus induced by high-frequency stimulation. Synapse 50:314–319
75. Silchenko A, Tass P (2008) Computational modeling of paroxysmal depolarization shifts in neurons induced by the glutamate release from astrocytes. Biol Cybern 98:61–74
76. Singer W (1989) Search for coherence: A basic principle of cortical self-organization. Concepts Neurosci 1:1–26
77. Smirnov D, Barnikol U, Barnikol T, Bezruchko B, Hauptmann C, Bührle C, Maarouf M, Sturm V, Freund H-J, Tass P (2008) The generation of parkinsonian tremor as revealed by directional coupling analysis. Europhys Lett 83:20003
78. Song S, Miller K, Abbott L (2000) Competitive Hebian learning through spike-timing-dependent synaptic plasticity. Nat Neurosci 3(9):919–926
79. Speckmann E, Elger C (1991) The neurophysiological basis of epileptic activity: A condensed overview. Epilepsy Res Suppl 2:1–7
80. Steriade M, Jones EG, Llinas RR (1990) Thalamic oscillations and signaling. Wiley, New York
81. Strogatz SH (2003) Sync: The emerging science of spontaneous order. Hyperion, New York
82. Tasker RR (1998) Deep brain stimulation is preferable to thalamotomy for tremor suppression. Surg Neurol 49:145–154
83. Tass P, Hauptmann C (2007) Therapeutic modulation of

synaptic connectivity with desynchronizing brain stimulation. Int J Psychophysiol 64:53–61
84. Tass PA (1996) Phase resetting associated with changes of burst shape. J Biol Phys 22:125–155
85. Tass PA (1996) Resetting biological oscillators – a stochastic approach. J Biol Phys 22:27–64
86. Tass PA (1999) Phase resetting in medicine and biology: Stochastic modelling and data analysis. Springer, Berlin
87. Tass PA (2000) Stochastic phase resetting: A theory for deep brain stimulation. Prog Theor Phys Suppl 139:301–313
88. Tass PA (2001) Desynchronizing double-pulse phase resetting and application to deep brain stimulation. Biol Cybern 85:343–354
89. Tass PA (2001) Effective desynchronization by means of double-pulse phase resetting. Europhys Lett 53:15–21
90. Tass PA (2001) Effective desynchronization with a resetting pulse train followed by a single-pulse. Europhys Lett 55:171–177
91. Tass PA (2002) Desynchronization of brain rhythms with soft phase-resetting techniques. Biol Cybern 87:102–115
92. Tass PA (2002) Effective desynchronization with a stimulation technique based on soft phase resetting. Europhys Lett 57:164–170
93. Tass PA (2002) Effective desynchronization with bipolar double-pulse stimulation. Phys Rev E 66:036226
94. Tass PA (2003) A model of desynchronizing deep brain stimulation with a demand-controlled coordinated reset of neural subpopulations. Biol Cybern 89:81–88
95. Tass PA (2003) Desynchronization by means of a coordinated reset of neural sub-populations – a novel technique for demand-controlled deep brain stimulation. Prog Theor Phys Suppl 150:281–296
96. Tass PA, Hauptmann C (2006) Therapeutic rewiring by means of desynchronizing brain stimulation. Nonlinear Phenom Complex Syst 9(3):298–312
97. Tass PA, Majtanik M (2006) Long-term anti-kindling effects of desynchronizing brain stimulation: A theoretical study. Biol Cybern 94:58–66
98. Tass PA, Hauptmann C, Popovych OV (2006) Development of therapeutic brain stimulation techniques with methods from nonlinear dynamics and statistical physics. Int J Bif Chaos 16(7):1889–1911
99. Timmermann L, Florin E, Reck C (2007) Pathological cerebral oscillatory activity in parkinson's disease: A critical review on methods, data and hypotheses. Expert Rev Med Devices 4(5):651–661
100. van Hemmen J (2001) Theory of synaptic plasticity. In: Moss F, Gielen S (eds) Handbook of biologicla physics, vol 4. Elsevier, Amsterdam, pp 771–823
101. Volkmann J (2004) Deep brain stimulation for the treatment of Parkinson's disease. J Clin Neurophysiol 21:6–17
102. Winfree A (1980) The geometry of biological time. Springer, Berlin
103. Wunderlin A, Haken H (1975) Scaling theory for nonequilibrium systems. Z Phys B 21:393–401
104. Zhai Y, Kiss IZ, Tass PA, Hudson JL (2005) Desynchronization of coupled electrochemical oscillators with pulse stimulations. Phys Rev E 71:065202R
105. Zhou Q, Tao H, Poo M (2003) Reversal and stabilization of synaptic modifications in a developing visual system. Science 300:1953–1957

Branching Processes

Mikko J. Alava[1], Kent Bækgaard Lauritsen[2]
[1] Department of Engineering Physics, Espoo University of Technology, Espoo, Finland
[2] Research Department, Danish Meteorological Institute, Copenhagen, Denmark

Article Outline

Glossary
Definition of the Subject
Introduction
Branching Processes
Self-Organized Branching Processes
Scaling and Dissipation
Network Science and Branching Processes
Conclusions
Future Directions
Acknowledgments
Bibliography

Glossary

Markov process A process characterized by a set of probabilities to go from a certain state at time t to another state at time $t+1$. These transition probabilities are independent of the history of the process and only depend on a fixed probability assigned to the transition.

Critical properties and scaling The behavior of equilibrium and many non-equilibrium systems in steady states contain critical points where the systems display scale invariance and the correlation functions exhibit an algebraic behavior characterized by so-called critical exponents. A characteristics of this type of behavior is the lack of finite length and time scales (also reminiscent of fractals). The behavior near the critical points can be described by scaling functions that are universal and that do not depend on the detailed microscopic dynamics.

Avalanches When a system is perturbed in such a way that a disturbance propagates throughout the system one speak of an avalanche. The local avalanche dynamics may either conserve energy (particles) or dissipate energy. The avalanche may also loose energy when it reaches the system boundary. In the neighborhood of a critical point the avalanche distribution is described by a power-law distribution.

Self-organized criticality (SOC) SOC is the surprising "critical" state in which many systems from physics to biology to social ones find themselves. In physics jar-

gon, they exhibit scale-invariance, which means that the dynamics – consisting of avalanches – has no typical scale in time or space. The really necessary ingredient is that there is a hidden, fine-tuned balance between how such systems are driven to create the dynamic response, and how they dissipate the input ("energy") to still remain in balance.

Networks These are descriptions of interacting systems, where in graph theoretical language nodes or vertices are connected by links or edges. The interesting thing can be the structure of the network, and its dynamics, or that of a process on the top of it like the spreading of computer viruses on the Internet.

Definition of the Subject

Consider the fate of a human population on a small, isolated island. It consists of a certain number of individuals, and the most obvious question, of importance in particular for the inhabitants of the island, is whether this number will go to zero. Humans die and reproduce in steps of one, and therefore one can try to analyze this fate mathematically by writing down what is called master equations, to describe the dynamics as a "branching process" (BP). The branching here means that if at time $t = 0$ there are N humans, at the next step $t = 1$ there can be $N - 1$ (or $N + 1$ or $N + 2$ if the only change from $t = 0$ was that a pair of twins was born). The outcome will depend in the simplest case on a "branching number", or the number of offspring λ that a human being will have [1,2,3,4].

If the offspring created are too few, then the population will decay, or reach an "absorbing state" out of which it will never escape. Likewise if they are many (the Malthusian case in reality, perhaps), exponential growth in time will ensue in the simplest case. In between, there is an interesting twist: a phase transition that separates these two outcomes at a critical value of λ_c. As is typical of such transitions in statistical physics one runs into scale-free behavior. The lifetime of the population suddenly has no typical scale, and its total size will be a stochastic quantity, described by a probability distribution that again has no typical scale exactly at λ_c.

The example of a small island also illustrates the many different twists that one can find in branching processes. The population can be "spatially dispersed" such that the individuals are separated by distance. There are in fact two interacting populations, called "male" and "female", and if the size of one of the populations becomes zero the other one will die out soon as well. The people on the island eat, and there is thus a hidden variable in the dynamics, the availability of food. This causes a history effect which makes the dynamics of human population what is called "non-Markovian". Imagine as above, that we look at the number of persons on the island at discrete times. A Markovian process is such that the probabilities to go from a state (say of) N to state $N + \delta$ depends only on the fixed probability assigned to the "transition" $N \to N + \delta N$. Clearly, any relatively faithful description of the death and birth rates of human beings has to consider the average state of nourishment, or whether there is enough food for reproduction.

Introduction

Branching processes are often perfect models of complex systems, or in other words exhibit deep complexity themselves. Consider the following example of a one-dimensional model of activated random walkers [5]. Take a line of sites x_i, $i = 1 \ldots L$. Fill the sites randomly to a certain density $n = N/L$, where N is the pre-set number of individuals performing the activated random walk. Now, let us apply the simple rule, that if there are two or more walkers at the same x_j, two of them get "activated" and hop to $j - 1$ or $j + 1$, again at random. In other words, this version of the drunken bar-hoppers problem has the twist that they do not like each other.

If the system is "periodic" or $i = 1$ is connected to $i = L$, then the dynamics is controlled by the density n. For a critical value n_c (estimated by numerical simulations to be about 0.9488… [6]) a phase transition takes place, such that for $n < n_c$ the asymptotic state is the "absorbing one", where all the walkers are immobilized since $N_i = 1$ or 0. In the opposite case for $n > n_c$ there is an active phase such that (in the infinite L-limit) the activity persists forever. This particular model is unquestionably non-Markovian if one only considers the number of active walkers or their density ρ. One needs to know the full state of N_i to be able to write down exact probabilities for how ρ changes in a discrete step of time.

The most interesting things happen if one changes the one-dimensional lattice by adapting two new rules. If a walker walks out (to $i = 0$ or $i = L + 1$), it disappears. Second, if there are no active ones ($\rho = 0$), one adds one new walker randomly to the system. Now the activity $\rho(t)$ is always at a marginal value, and the long-term average of n becomes a (possibly L-dependent) constant such that the first statement is true, i. e. the system becomes critical. With these rules, the model of activated random walkers is also known as the Manna model after the Indian physicist [5], and it exhibits the phenomenon dubbed "Self-Organized Criticality" (SOC) [7]. Figure 1 shows an example of the dynamics by using what is called an "activity plot",

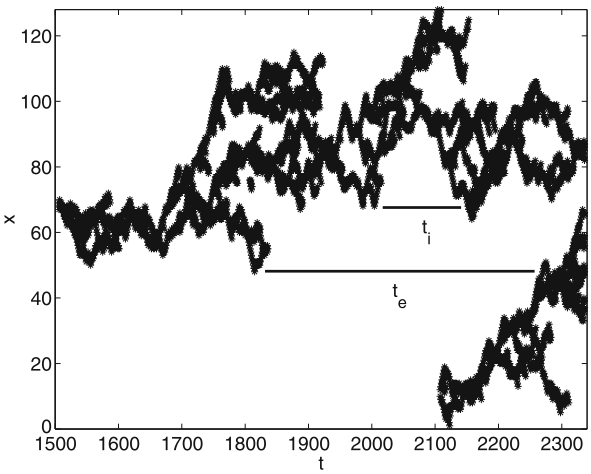

Branching Processes, Figure 1

We follow the "activity" in a one-dimensional system of random activated walkers. The walkers stroll around the *x*-axis, and the pattern becomes in fact scale-invariant. This system is such that some of the walkers disappear (by escaping through open boundaries) and to maintain a constant density new ones are added. One question one may ask is what is the waiting time, that another (or the same) walker gets activated at the same location after a time of inactivity. As one can see from the figure, this can be a result of old activity getting back, or a new "avalanche" starting from the addition of an outsider (courtesy of Lasse Laurson)

where those locations x_i are marked both in space and time which happen to contain just-activated walkers. One can now apply several kinds of measures to the system, but the figure already hints about the reasons why these simple models have found much interest. The structure of the activity is a self-affine fractal (for discussions about fractals see other reviews in this volume).

The main enthusiasm about SOC comes from the avalanches. These are in other words the bursts of activity that separate quiescent times (when $\rho = 0$). The silence is broken by the addition of a particle or a walker, and it creates an integrated quantity (volume) of activity, $s = \int_0^T \rho(t)\mathrm{d}t$, where for $0 < t < T$, $\rho > 0$ and $\rho = 0$ at the endpoints t and T. The original boost to SOC took place after Per Bak, Chao Tang, and Kay Wiesenfeld published in 1987 a highly influential paper in the premium physics journal Physical Review Letters, and introduced what is called the Bak–Tang–Wiesenfeld (BTW) sandpile model – of which the Manna one is a relative [7]. The BTW and Manna models and many others exhibit the important property that the avalanche sizes s have a *scale-free* probability distribution but note that this simple criticality is not always true, not even for the BTW model which shows so-called multiscaling [8,9]. Scale-free probability distributions are usually written as

$$P(s) \sim s^{-\tau_s} f_s(s/L^{D_s}) \,, \tag{1}$$

here all the subscripts refer to the fact that we look at avalanches. τ_s and D_s define the avalanche exponent and the cut-off exponent, respectively. f_s is a cut-off function that together with D_s includes the fact that the avalanches are restricted somehow by the system size (in the one-dimensional Manna model, if s becomes too large many walkers are lost, and first n drops and then ρ goes to zero). Similar statements can be made about the avalanche durations (T), area or support A, and so forth [10,11,12,13,14].

The discovery of simple-to-define models, yet of very complex behavior has been a welcome gift since there are many, many phenomena that exhibit apparently scale"free statistics and/or bursty or intermittent dynamics similar to SOC models. These come from natural sciences – but not only, since economy and sociology are also giving rise to cases that need explanations. One particular field where these ideas have found much interest is the physics of materials ranging from understanding earthquakes to looking at the behavior of vortices in superconductors. The Gutenberg–Richter law of earthquake magnitudes is a power-law, and one can measure similar statistics from fracture experiments by measuring acoustic emission event energies from tensile tests on ordinary paper [15]. The modern theory of networks is concerned with graph or network structures which exhibit scale-free characteristics, in particular the number of neighbors a vertex has is often a stochastic quantity exhibiting a power-law-like probability distribution [16,17]. Here, branching processes are also finding applications. They can describe the development of network structures, e. g., as measured by the degree distribution, the number of neighbors a node (vertex) is connected to, and as an extension of more usual models give rise to interesting results when the dynamics of populations *on networks* are being considered.

Whenever comparing with real, empirical phenomena models based on branching processes give a paradigm on two levels. In the case of SOC this is given by a combination of "internal dynamics" and "an ensemble". Of these, the first means e. g. that activated walkers or particles of type A are moved around with certain rules, and of course there is an enormous variety of possible models. e. g. in the Manna-model this is obvious if one splits the walkers into categories A (active) and B (passive). Then, there is the question of how the balance (assuming this balance *is* true) is maintained. The SOC models do this by a combination of dissipation (by e. g. particles dropping off the edge of a system) and drive (by addition of B's), where the rates are chosen actually pretty carefully. In the case of growing

networks, one can similarly ask what kind of growth laws allow for scale-free distributions of the node degree.

For the theory of complex systems branching processes thus give rise to two questions: what classes of models are there? What kinds of truly different ensembles are there? The theory of reaction-diffusion systems has tried to answer to the first question since the 70's, and the developments reflect the idea of universality in statistical physics (see e. g. the article of Dietrich Stauffer in this volume). There, this means that the behavior of systems at "critical points" – such as defined by the λ_c and n_c from above – follows from the dimension at hand and the "universality class" at hand. The activated walkers' model, it has recently been established, belongs to the "fixed energy sandpile" one, and is closely related to other seemingly far-fetched ones such as the depinning/pinning of domain walls in magnets (see e. g. [18]). The second question can be stated in two ways: forgetting about the detailed model at hand, when can one expect complex behavior such as power-law distributions, avalanches etc.? Second, and more technically, can one derive the exponents such as τ_s and D_s from those of the same model at its "usual" phase transition?

The theory of branching processes provide many answers to these questions, and in particular help to illustrate the influence of boundary conditions and modes of driving on the expected behavior. Thus, one gets a clear idea of the kind of complexity one can expect to see in the many different kinds of systems where avalanches and intermittency and scaling is observed.

Branching Processes

The mathematical branching process is defined for a set of objects that do not interact. At each iteration, each object can give rise to new objects with some probability p (or in general it can be a set of probabilities). By continuing this iterative process the objects will form what is referred to as a cluster or an avalanche. We can now ask questions of the following type: Will the process continue for ever? Will the process maybe die out and stop after a finite number of iterations? What will the average lifetime be? What is the average size of the clusters? And what is (the asymptotic form of) the probability that the process is active after a certain number of iterations, etc.

BP Definition

We will consider a simple discrete BP denoted the Galton–Watson process. For other types of BPs (including processes in continuous time) we refer to the book by Harris [1]. We will denote the number of objects at generation n as: $z_0, z_1, z_2, \ldots, z_n, \ldots$. The index n is referred to as time and time $t = 0$ corresponds to the 0th generation where we take $z_0 = 1$, i. e., the process starts from one object. We will assume that the transition from generation n to $n + 1$ is given by a probability law that is independent of n, i. e., it is assumed that it forms a Markov process. And finally it will be assumed that different objects do not interact with one another.

The probability measure for the process is characterized by probabilities as follows: p_k is the probability that an object in the nth generation will have k offsprings in the $n + 1$th generation. We assume that p_k is independent of n. The probabilities p_k thus reads $p_k = \mathrm{Prob}(z_1 = k)$ and fulfills:

$$\sum_k p_k = 1 \,. \tag{2}$$

Figure 2 shows an example of the tree structure for a branching process and the resulting avalanche.

One can define the probability generating function $f(s)$ associated to the transition probabilities:

$$f(s) = \sum_k p_k s^k \,. \tag{3}$$

The first and second moments of the number z_1 are denoted as:

$$m = Ez_1 \,, \qquad \sigma^2 = \mathrm{Var}\, z_1 \,. \tag{4}$$

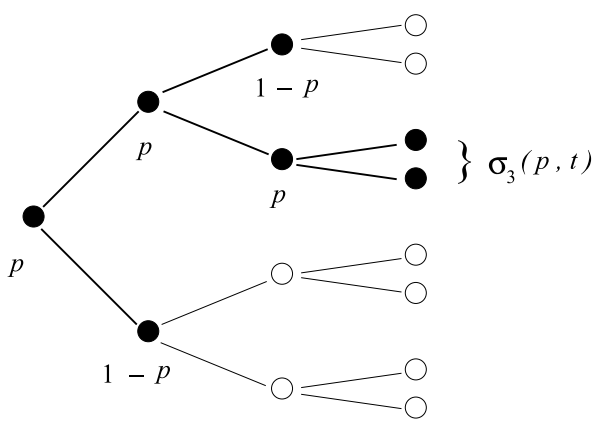

Branching Processes, Figure 2
Schematic drawing of an avalanche in a system with a maximum of $n = 3$ avalanche generations corresponding to $N = 2^{n+1} - 1 = 15$ sites. Each black site relaxes with probability p to two new black sites and with probability $1 - p$ to two white sites (i. e., $p_0 = 1 - p$, $p_1 = 0$, $p_2 = p$). The black sites are part of an avalanche of size $s = 7$, whereas the active sites at the boundary yield a boundary avalanche of size $\sigma_3(p, t) = 2$

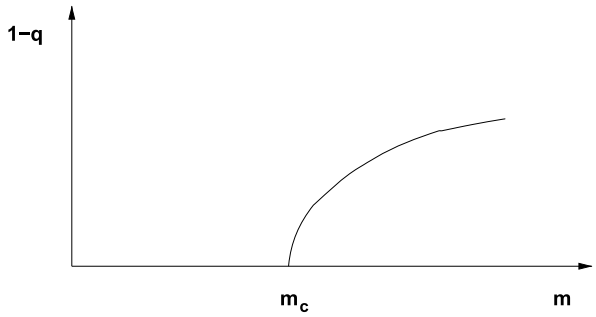

Branching Processes, Figure 3
Schematic drawing of the behavior of the probability of extinction q, of the branching process. The quantity $1-q$ is similar to the order parameter for systems in equilibrium at their critical points and the quantity m_c is referred to as the critical point for the BP

By taking derivatives of the generating function at $s = 1$ it follows that:

$$Ez_n = m^n \,, \qquad \mathrm{Var}\, z_n = n\sigma^2 \,. \tag{5}$$

For the BP defined in Fig. 2, it follows that $m = 2p$ and $\sigma^2 = 4p(1-p)$.

An important quantity is the probability for extinction q. It is obtained as follows:

$$q = P(z_n \to 0) = P(z_n = 0 \text{ for some } n) = \lim_n P(z_n = 0). \tag{6}$$

It can be shown that for $m \le 1$: $q = 1$; and for $m > 1$: there exists a solution that fulfills $q = f(q)$, where $0 < q < 1$ [1]. It is possible to show that $\lim_n P(z_n = k) = 0$, for $k = 1, 2, 3, \ldots$, and that $z_n \to 0$ with probability q, and $z_n \to \infty$, with probability $1-q$. Thus, the sequence $\{z_n\}$ does not remain positive and bounded [1].

The quantity $1-q$ is similar to an order parameter for systems in equilibrium and its behavior is schematically shown in Fig. 3. The behavior around the value $m_c = 1$, the so-called critical value for the BP (see below), can in analogy to second order phase transitions in equilibrium systems be described by a critical exponent β defined as follows ($\beta = 1$, cf. [1]):

$$\text{Prob(survival)} = \begin{cases} 0 \,, & m \le m_c \\ (m - m_c)^\beta \,, & m > m_c. \end{cases} \tag{7}$$

Avalanches and Critical Point

We will next consider the clusters, or avalanches, in more detail and obtain the asymptotic form of the probability distributions. The size of the cluster is given by the sum $z = z_1 + z_2 + z_3 + \ldots$. One can also consider other types of clusters, e. g., the activity σ of the boundary (of a finite tree) and define the boundary avalanche (cf. Fig. 2).

For concreteness, we consider the BP defined in Fig. 2. The quantities $P_n(s,p)$ and $Q_n(\sigma,p)$ denote the probabilities of having an avalanche of size s and boundary size σ in a system with n generations. The corresponding generating functions are defined by [1]

$$f_n(x,p) \equiv \sum_s P_n(s,p) x^s \tag{8}$$

$$g_n(x,p) \equiv \sum_\sigma Q_n(\sigma,p) x^\sigma \,. \tag{9}$$

Due to the hierarchical structure of the branching process, it is possible to write down recursion relations for $P_n(s,p)$ and $Q_n(\sigma,p)$:

$$f_{n+1}(x,p) = x\left[(1-p) + p f_n^2(x,p)\right], \tag{10}$$

$$g_{n+1}(x,p) = (1-p) + p g_n^2(x,p) \,, \tag{11}$$

where $f_0(x,p) = g_0(x,p) = x$. The avalanche distribution $D(s)$ is determined by $P_n(s,p)$ by using the recursion relation (10). The stationary solution of Eq. (10) in the limit $n \gg 1$ is given by

$$f(x,p) = \frac{1 - \sqrt{1 - 4x^2 p(1-p)}}{2xp} \,. \tag{12}$$

By expanding Eq. (12) as a series in x, comparing with the definition (8), and using Stirling's formula for the high-order terms, we obtain for sizes such that $1 \ll s \lesssim n$ the following behavior:

$$P_n(s,p) = \frac{\sqrt{2(1-p)/\pi p}}{s^{3/2}} \exp\left(-s/s_c(p)\right) \,. \tag{13}$$

The cutoff $s_c(p)$ is given by $s_c(p) = -2/\ln[4p(1-p)]$. As $p \to 1/2$, $s_c(p) \to \infty$, thus showing explicitly that the critical value for the branching process is $p_c = 1/2$ (i. e., $m_c = 1$), and that the mean-field avalanche exponent, cf. Eq. (1), for the critical branching process is $\tau = 3/2$.

The expression (13) is only valid for avalanches which are not effected by the finite size of the system. For avalanches with $n \lesssim s \lesssim N$, it is possible to solve the recursion relation (10), and then obtain $P_n(s,p)$ for $p \ge p_c$ by the use of a Tauberian theorem [19,20,21]. By carrying out such an analysis one obtains after some algebra $P_n(s,p) \approx A(p) \exp\left(-s/s_0(p)\right)$, with functions $A(p)$ and $s_0(p)$ which can not be determined analytically. Nevertheless, we see that for any p the probabilities $P_n(s,p)$ will decay exponentially. One can also calculate the asymptotic form of $Q_n(\sigma,p)$ for $1 \ll \sigma \lesssim n$ and $p \ge p_c$ by the use of

a Tauberian theorem [19,20,21]. We will return to study this in the next section where we will also define and investigate the distribution of the time to extinction.

Self-Organized Branching Processes

We now return to discussing the link between self-organized criticality as mentioned in the introduction and branching processes. The simplest theoretical approach to SOC is mean-field theory [22], which allows for a qualitative description of the behavior of the SOC state. Mean-field exponents for SOC models have been obtained by various approaches [22,23,24,25] and it turns out that their values (e. g., $\tau = 3/2$) are the same for all the models considered thus far. This fact can easily be understood since the spreading of an avalanche in mean-field theory can be described by a front consisting of non-interacting particles that can either trigger subsequent activity or die out. This kind of process is reminiscent of a branching process. The connection between branching processes and SOC has been investigated, and it has been argued that the mean-field behavior of sandpile models can be described by a critical branching process [26,27,28].

For a branching process to be critical one must fine tune a control parameter to a critical value. This, by definition, cannot be the case for a SOC system, where the critical state is approached dynamically without the need to fine tune any parameter. In the so-called self-organized branching-process (SOBP), the coupling of the local dynamical rules to a global condition drives the system into a state that is indeed described by a critical branching process [29]. It turns out that the mean-field theory of SOC models can be exactly mapped to the SOBP model.

In the mean-field description of the sandpile model ($d \to \infty$) one neglects correlations, which implies that avalanches do not form loops and hence spread as a branching process. In the SOBP model, an avalanche starts with a single active site, which then relaxes with probability p, leading to two new active sites. With probability $1 - p$ the initial site does not relax and the avalanche stops. If the avalanche does not stop, one repeats the procedure for the new active sites until no active site remains. The parameter p is the probability that a site relaxes when it is triggered by an external input. For the SOBP branching process, there is a critical value, $p_c = 1/2$, such that for $p > p_c$ the probability to have an infinite avalanche is non-zero, while for $p < p_c$ all avalanches are finite. Thus, $p = p_c$ corresponds to the critical case, where avalanches are power law distributed.

In this description, however, the boundary conditions are not taken into account – even though they are crucial for the self-organization process. The boundary conditions can be introduced in the problem in a natural way by allowing for no more than n generations for each avalanche. Schematically, we can view the evolution of a single avalanche of size s as taking place on a tree of size $N = 2^{n+1} - 1$ (see Fig. 2). If the avalanche reaches the boundary of the tree, one counts the number of active sites σ_n (which in the sandpile language corresponds to the energy leaving the system), and we expect that p decreases for the next avalanche. If, on the other hand, the avalanche stops before reaching the boundary, then p will slightly increase. The number of generations n can be thought of as some measure of the linear size of the system.

The above avalanche scenario is described by the following dynamical equation for $p(t)$:

$$p(t+1) = p(t) + \frac{1 - \sigma_n(p,t)}{N} , \tag{14}$$

where σ_n, the size of an avalanche reaching the boundary, fluctuates in time and hence acts as a stochastic driving force. If $\sigma_n = 0$, then p increases (because some energy has been put into the system without any output), whereas if $\sigma_n > 0$ then p decreases (due to energy leaving the system). Equation (14) describes the global dynamics of the SOBP, as opposed to the local dynamics which is given by the branching process. One can study the model for a fixed value of n, and then take the limit $n \to \infty$. In this way, we perform the long-time limit before the "thermodynamic" limit, which corresponds exactly to what happens in sandpile models.

We will now show that the SOBP model provides a mean-field theory of self-organized critical systems. Consider for simplicity the sandpile model of activated random walkers from the Introduction [5]: When a particle is added to a site z_i, the site will relax if $z_i = 1$. In the limit $d \to \infty$, the avalanche will never visit the same site more than once. Accordingly, each site in the avalanche will relax with the same probability $p = P(z = 1)$. Eventually, the avalanche will stop, and $\sigma \geq 0$ particles will leave the system. Thus, the total number of particles $M(t)$ evolves according to

$$M(t+1) = M(t) + 1 - \sigma . \tag{15}$$

The dynamical Eq. (14) for the SOBP model is recovered by noting that $M(t) = NP(z = 1) = Np$. By taking the continuum time limit of Eq. (14), it is possible to obtain the following expression:

$$\frac{\mathrm{d}p}{\mathrm{d}t} = \frac{1 - (2p)^n}{N} + \frac{\eta(p,t)}{N} , \tag{16}$$

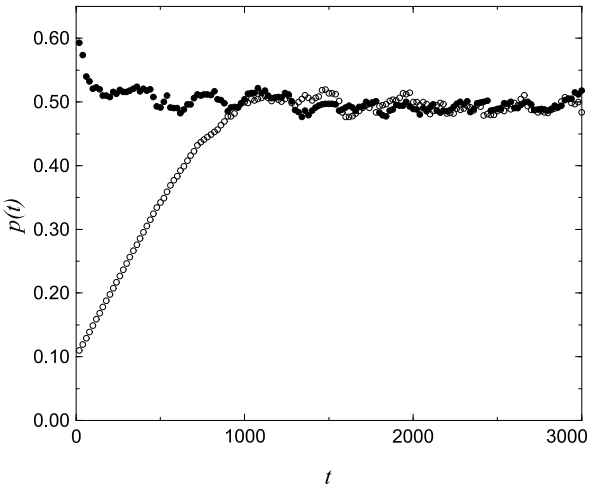

Branching Processes, Figure 4
The value of p as a function of time for a system with $n = 10$ generations. The two curves refer to two different initial conditions, above p_c(●) and below p_c(○). After a transient, the control parameter $p(t)$ reaches its critical value p_c and fluctuates around it with short-range correlations

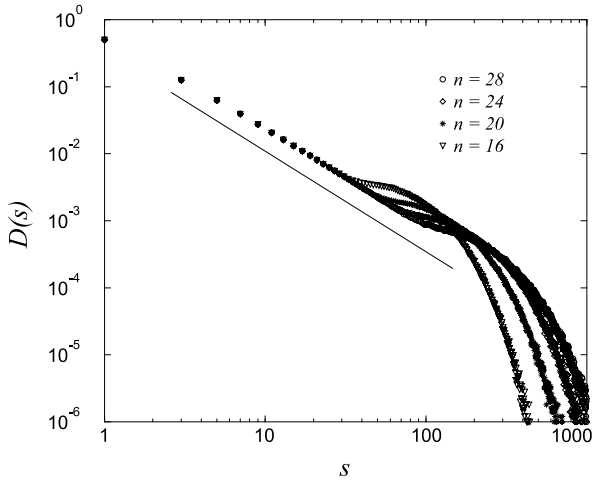

Branching Processes, Figure 5
Log–log plot of the avalanche distribution $D(s)$ for different system sizes. The number of generations n increases from left to right. A line with slope $\tau = 3/2$ is plotted for reference, and it describes the behavior of the data for intermediate s values, cf. Eq. (18). For large s, the distributions fall off exponentially

where $\eta \equiv \langle \sigma_n \rangle - \sigma_n = (2p)^n - \sigma_n(p,t)$ describes the fluctuations in the steady state. Thus, η is obtained by measuring σ_n for each realization of the process. Without the last term, Eq. (16) has a fixed point ($\mathrm{d}p/\mathrm{d}t = 0$) for $p = p_c = 1/2$. On linearizing Eq. (16), one sees that the fixed point is attractive, which demonstrates the self organization of the SOBP model since the noise η/N will have vanishingly small effect in the thermodynamic limit [29].

Figure 4 shows the value of p as a function of time. Independent of the initial conditions, one finds that after a transient $p(t)$ reaches the self-organized state described by the critical value $p_c = 1/2$ and fluctuates around it with short-range correlations (of the order of one time unit). By computing the variance of $p(t)$, one finds that the fluctuations can be very well described by a Gaussian distribution, $\phi(p)$ [29]. In the limit $N \to \infty$, the distribution $\phi(p)$ approaches a delta function, $\delta(p - p_c)$.

Avalanche Distributions

Figure 5 shows the avalanche size distribution $D(s)$ for different values of the number of generations n. One notices that there is a scaling region ($D(s) \sim s^{-\tau}$ with $\tau = 3/2$), whose size increases with n, and characterized by an exponential cutoff. This power-law scaling is a signature of the mean-field criticality of the SOBP model. The distribution of active sites at the boundary, $D(\sigma)$, for different values of the number of generations falls off exponentially [29].

In the limit where $n \gg 1$ one can obtain various analytical results and e. g. calculate the avalanche distribution $D(s)$ for the SOBP model. In addition, one can obtain results for finite, but large, values of n. The distribution $D(s)$ can be calculated as the average value of $P_n(s,p)$ with respect to the probability density $\phi(p)$, i. e., according to the formula

$$D(s) = \int_0^1 \mathrm{d}p\, \phi(p)\, P_n(s,p)\,. \tag{17}$$

The simulation results in Fig. 4 yield that $\phi(p)$ for $N \gg 1$ approaches the delta function $\delta(p - p_c)$. Thus, from Eqs. (13) and (17) we obtain the power-law behavior

$$D(s) = \sqrt{\frac{2}{\pi}}\, s^{-\tau}\,, \tag{18}$$

where $\tau = 3/2$, and for $s \gtrsim n$ we obtain an exponential cutoff $\exp(-s/s_0(p_c))$. These results are in complete agreement with the numerical results shown in Fig. 5. The deviations from the power-law behavior (18) are due to the fact that Eq. (13) is only valid for $1 \ll s \lesssim n$. One can also calculate the asymptotic form of $Q_n(\sigma,p)$ for $1 \ll \sigma \lesssim n$ and $p \geq p_c$ by the use of a Tauberian theorem [19,20,21]; the result show that the boundary avalanche distribution is

$$D(\sigma) = \int_0^1 \mathrm{d}p\, \phi(p)\, Q_n(\sigma,p) = \frac{8}{n^2}\, \exp\left(-2\sigma/n\right)\,, \tag{19}$$

which agrees with simulation results for $n \gg 1$, cf. [29].

The avalanche lifetime distribution $L(t) \sim t^{-y}$ yields the probability to have an avalanche that lasts for a time t. For a system with m generations one obtains $L(m) \sim m^{-2}$ [1]. Identifying the number of generations m of an avalanche with the time t, we thus obtain the mean-field value $y = 2$, in agreement with simulations of the SOBP model [29]. In summary, the self-organized branching process captures the physical features of the self-organization mechanism in sandpile models. By explicitly incorporating the boundary conditions it follows that the dynamics drives the system into a stationary state, which in the thermodynamic limit corresponds to the critical branching process.

Scaling and Dissipation

Sometimes it can be difficult to determine whether the cutoff in the scaling is due to finite-size effects or due to the fact that the system is not *at* but rather only *close to* the critical point. In this respect, it is important to test the robustness of critical behavior by understanding which perturbations destroy the critical properties. It has been shown numerically [30,31,32] that the breaking of the conservation of particle numbers leads to a characteristic size in the avalanche distributions. We will now allow for dissipation in branching processes and show how the system self-organizes into a sub-critical state. In other words, the degree of nonconservation is a relevant parameter in the renormalization group sense [33].

Consider again the two-state model introduced by Manna [5]. Some degree of nonconservation can be introduced in the model by allowing for energy dissipation in a relaxation event. In a continuous energy model this can be done by transferring to the neighboring sites only a fraction $(1-\epsilon)$ of the energy lost by the relaxing site [30]. In a discrete energy model, such as the Manna two-state model, one can introduce dissipation as the probability ϵ that the two particles transferred by the relaxing site are annihilated [31]. For $\epsilon = 0$ one recovers the original two-state model.

Numerical simulations [30,31] show that different ways of considering dissipation lead to the same effect: a characteristic length is introduced into the system and the criticality is lost. As a result, the avalanche size distribution decays not as a pure power law but rather as

$$D(s) \sim s^{-\tau}\, h_s(s/s_c)\,. \tag{20}$$

Here $h_s(x)$ is a cutoff function and the cutoff size scales as

$$s_c \sim \epsilon^{-\varphi}\,. \tag{21}$$

The size s is defined as the number of sites that relax in an avalanche. We define the avalanche lifetime T as the number of steps comprising an avalanche. The corresponding distribution decays as

$$D(T) \sim T^{-y}\, h_T(T/T_c)\,, \tag{22}$$

where $h_T(x)$ is another cutoff function and T_c is a cutoff that scales as

$$T_c \sim \epsilon^{-\psi}\,. \tag{23}$$

The cutoff or "scaling" functions $h_s(x)$ and $h_T(x)$ fall off exponentially for $x \gg 1$.

To construct the mean-field theory one proceeds as follows [34]: When a particle is added to an arbitrary site, the site will relax if a particle was already present, which occurs with probability $p = P(z = 1)$, the probability that the site is occupied. If a relaxation occurs, the two particles are transferred with probability $1-\epsilon$ to two of the infinitely many nearest neighbors, or they are dissipated with probability ϵ (see Fig. 6).

The avalanche process in the mean-field limit is a branching process. Moreover, the branching process can be described by the *effective* branching probability

$$\tilde{p} \equiv p(1-\epsilon)\,, \tag{24}$$

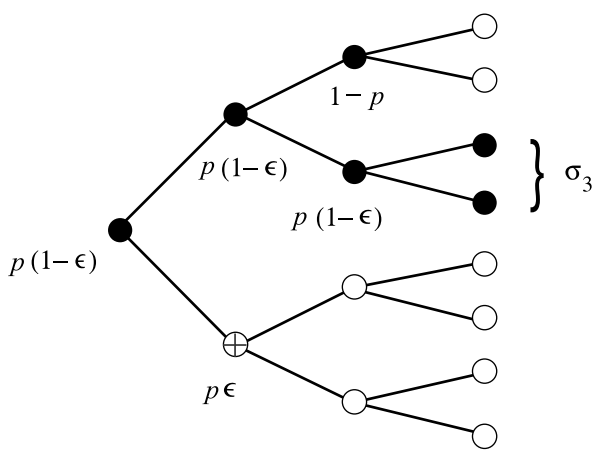

Branching Processes, Figure 6
Schematic drawing of an avalanche in a system with a maximum of $n = 3$ avalanche generations corresponding to $N = 2^{n+1} - 1 = 15$ sites. Each black site (●) can relax in three different ways: (i) with probability $p(1-\epsilon)$ to two new black sites, (ii) with probability $1-p$ the avalanche stops, and (iii) with probability $p\epsilon$ two particles are dissipated at a black site, which then becomes a marked site (⊕), and the avalanche stops. The black sites are part of an avalanche of size $s = 6$, whereas the active sites at the boundary yield $\sigma_3(p, t) = 2$. There was one dissipation event such that $\kappa = 2$

where $\tilde{p}$ is the probability to create two new active sites. We know that there is a critical value for $\tilde{p} = 1/2$, or

$$p = p_c \equiv \frac{1}{2(1-\epsilon)} \,. \tag{25}$$

Thus, for $p > p_c$ the probability to have an infinite avalanche is non-zero, while for $p < p_c$ all avalanches are finite. The value $p = p_c$ corresponds to the critical case where avalanches are power law distributed.

The Properties of the Steady State

To address the self-organization, consider the evolution of the total number of particles $M(t)$ in the system after each avalanche:

$$M(t+1) = M(t) + 1 - \sigma(p,t) - \kappa(p,t) \,. \tag{26}$$

Here σ is the number of particles that leave the system from the boundaries and κ is the number of particles lost by dissipation. Since (cf. Sect. "Self–Organized Branching Processes") $M(t) = NP(z=1) = Np$, we obtain an evolution equation for the parameter p:

$$p(t+1) = p(t) + \frac{1 - \sigma(p,t) - \kappa(p,t)}{N} \,. \tag{27}$$

This equation reduces to the SOBP model for the case of no dissipation ($\kappa = 0$). In the continuum limit one obtains [34]

$$\frac{\mathrm{d}p}{\mathrm{d}t} = \frac{1}{N}\left[1 - (2p(1-\epsilon))^n - p\epsilon H(p(1-\epsilon))\right] + \frac{\eta(p,t)}{N} \,. \tag{28}$$

Here, we defined the function $H(p(1-\epsilon))$, that can be obtained analytically, and introduced the function $\eta(p,t)$ to describe the fluctuations around the average values of σ and κ. It can be shown numerically that the effect of this "noise" term is vanishingly small in the limit $N \to \infty$ [34].

Without the noise term one can study the fixed points of Eq. (28) and one finds that there is only one fixed point,

$$p^* = 1/2 \,, \tag{29}$$

independent of the value of ϵ; the corrections to this value are of the order $O(1/N)$. By linearizing Eq. (28), it follows that the fixed point is attractive. This result implies that the SOBP model self-organizes into a state with $p = p^*$. In Fig. 7 is shown the value of p as a function of time for different values of the dissipation ϵ. We find that independent of the initial conditions after a transient $p(t)$ reaches the self-organized steady-state described by the fixed point value $p^* = 1/2$ and fluctuates around it with short-range correlations (of the order of one time unit). The fluctuations around the critical value decrease with the system size as $1/N$. It follows that in the limit $N \to \infty$ the distribution $\phi(p)$ of p approaches a delta function $\phi(p) \sim \delta(p - p^*)$.

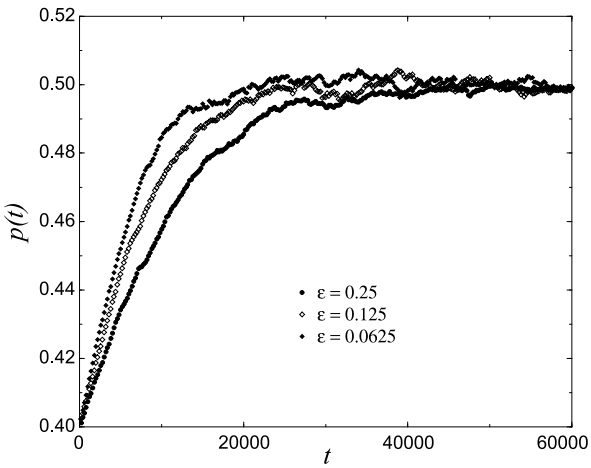

Branching Processes, Figure 7
The value of the control parameter $p(t)$ as a function of time for a system with different levels of dissipation. After a transient, $p(t)$ reaches its fixed-point value $p^* = 1/2$ and fluctuates around it with short-range time correlations

By comparing the fixed point value (29) with the critical value (25), we obtain that in the presence of dissipation ($\epsilon > 0$) the self-organized steady-state of the system is *subcritical*. Figure 8 is a schematic picture of the phase diagram of the model, including the line $p = p_c$ of critical behavior (25) and the line $p = p^*$ of fixed points (29). These two lines intersect only for $\epsilon = 0$.

Avalanche and Lifetime Distributions

In analogy to the results in Sect. "Self–Organized Branching Processes", we obtain similar formulas for the avalanche size distributions but with $\tilde{p}$ replacing p. As a result we obtain the distribution

$$D(s) = \sqrt{\frac{2}{\pi}}\,\frac{1+\epsilon+\ldots}{s^\tau}\,\exp\left(-s/s_c(\epsilon)\right) \,. \tag{30}$$

We can expand $s_c(\tilde{p}(\epsilon)) = -2/\ln\left[4\tilde{p}(1-\tilde{p})\right]$ in ϵ with the result

$$s_c(\epsilon) \sim \frac{2}{\epsilon^\varphi} \,, \quad \varphi = 2 \,. \tag{31}$$

Furthermore, the mean-field exponent for the critical branching process is obtained setting $\epsilon = 0$, i. e.,

$$\tau = 3/2 \,. \tag{32}$$

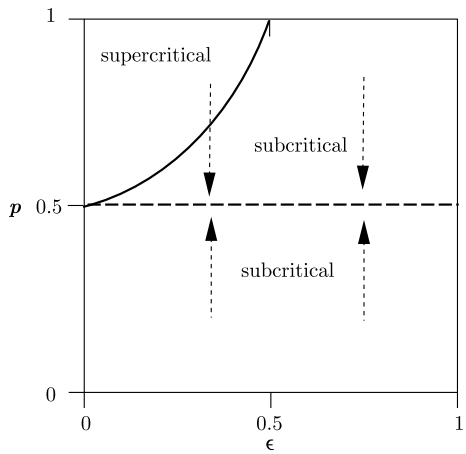

Branching Processes, Figure 8
Phase diagram for the SOBP model with dissipation. The dashed line shows the fixed points $p^* = 1/2$ of the dynamics, with the flow being indicated by the arrows. The solid line shows the critical points, cf. Eq. (25)

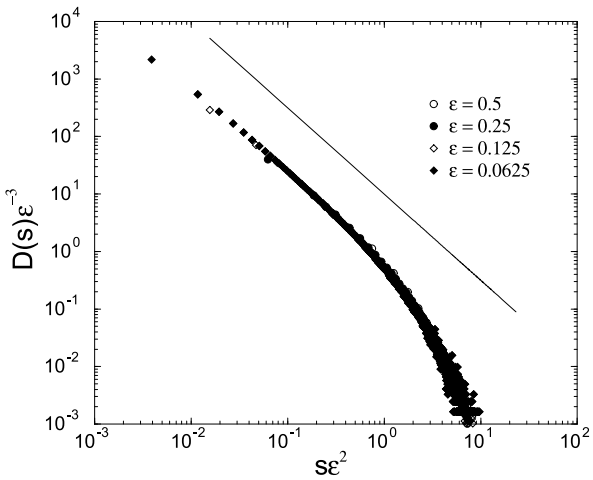

Branching Processes, Figure 9
Log–log plot of the avalanche distribution $D(s)$ for different levels of dissipation. A line with slope $\tau = 3/2$ is plotted for reference, and it describes the behavior of the data for intermediate s values, cf. Eq. (30). For large s, the distributions fall off exponentially. The data collapse is produced according to Eq. (30)

These results are in complete agreement with the SOBP model and the simulation of $D(s)$ for the SOBP model with dissipation (cf. Fig. 9). The deviations from the power-law behavior (30) are due to the fact that Eq. (13) is only valid for $1 \ll s \lesssim n$.

Next, consider the avalanche lifetime distribution $D(T)$ characterizing the probability to obtain an avalanche which spans m generations. It can be shown that the result

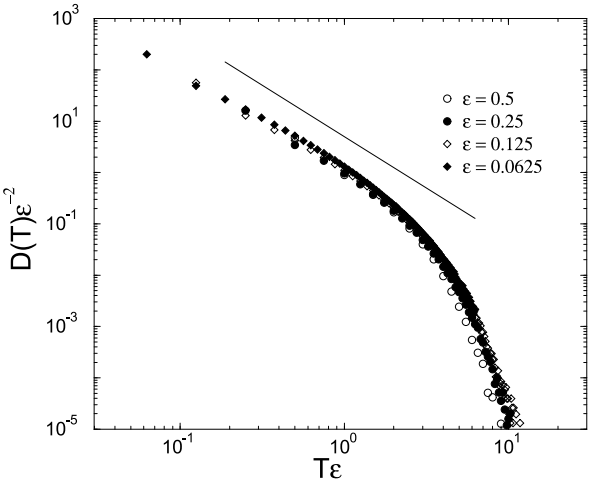

Branching Processes, Figure 10
Log–log plot of the lifetime distribution $D(T)$ for different levels of dissipation. A line with slope $y = 2$ is plotted for reference. Note the initial deviations from the power law for $\epsilon = 0$ due to the strong corrections to scaling. The data collapse is produced according to Eq. (33)

can be expressed in the scaling form [1,34]

$$D(T) \sim T^{-y} \exp(-T/T_c) \,, \tag{33}$$

where

$$T_c \sim \epsilon^{-\psi} \,, \quad \psi = 1 \,. \tag{34}$$

The lifetime exponent y was defined in Eq. (22), wherefrom we confirm the mean-field result

$$y = 2 \,. \tag{35}$$

In Fig. 10, we show the data collapse produced by Eq. (33) for the lifetime distributions for different values of ϵ.

In summary, the effect of dissipation on the dynamics of the sandpile model in the mean-field limit ($d \to \infty$) is described by a branching process. The evolution equation for the branching probability has a single attractive fixed point which in the presence of dissipation is not a critical point. The level of dissipation ϵ therefore acts as a relevant parameter for the SOBP model. These results show, in the mean-field limit, that criticality in the sandpile model is lost when dissipation is present.

Network Science and Branching Processes

An interesting mixture of two kinds of "complex systems" is achieved by considering what kind of applications one can find of branching processes applied to or on networks.

Here, network is a loose term in the way it is often used – as in the famous "six degrees" concept of how the social interactions of human beings can be measured to have a "small world" character [16,17,35]. These structures are usually thought of in terms of graphs, with vertices forming a set G, and edges a set E such that if $e_{gg'} \in E$ it means that it connects two sites g and $g' \in G$. An edge or link can either be symmetric or asymmetric. Another article in this volume discusses in depth the modern view on networks, providing more information. The Structural properties of networks heterogeneous networks are most usually measured by the degree k_i of site $i \in G$, i. e. the number of nearest neighbors the vertex i is connected to. One can go further by defining weights w_{ij} for edges e_{ij} [36,37,38].

The interest in heterogeneous networks is largely due to two typical properties one can find. In a scale-free network the structure is such that the probability distribution has a power-law form $P(k) \sim k^{-\alpha}$ up to a size-dependent cut-off $k_{\max}$. One can now show using mean-field-like reference models (such as the "configuration model", which has no structural correlations and a prefixed $P(k)$), that it follows that these networks tend to have a "small world" character, that the average vertex-vertex distance is only logarithmic in the number of vertices N: $d(G) \sim \ln N$. This property is correlated with the fact that the various moments of k depend on α. For $\langle k \rangle$ to remain finite, $\alpha > 2$ is required. In many empirically measured networks it appears that $\alpha < 3$, which implies that $\langle k^2 \rangle$ diverges (we are assuming, that $P(k)$ is cut off at a maximum degree $k_{\max}$ for which $\int_{k_{\max}}^{\infty} \sim 1/N$, from which the divergences follow) [16,17,35].

The first central question of the two we pose here is, what happens to branching processes on the top of such networks? Again, this can be translated into an inquiry about the "phase diagram" – given a model – and its particulars. We would like to understand what is the critical threshold λ_c, and what happens to sub-critical and super-critical systems (with appropriate values of λ). The answers may of course depend on the model at hand, i. e. the universality classes are still an important issue.

One easy analytical result for un-correlated (in particular tree-like) networks is that activity will persist in the branching process if the number of offspring is at least one for an active "individual" or vertex. Consider now one such vertex, i. Look at the neighbors of neighbors. There are of the order $\mathrm{O}\left(k_i \times \langle k_{\text{neighbor of } i} \rangle - 1\right)$ of these, so $\lambda_c \sim 1/\langle k^2 \rangle$. This demonstrates the highly important result, that the *critical threshold may vanish* in heterogeneous networks, for $\alpha \leq 3$ in the example case.

Pastor-Satorras and Vespignani were the first to point out this fact for the Susceptible-Infected-Susceptible (SIS) model [39]. The analysis is relatively easy to do using mean-field theory for ρ^k, the average activity for the SIS branching process on nodes of degree k. The central equation reads

$$\partial_t \rho^k(t) = -\rho^k(t) + \lambda k \left[1 - \rho^k(t)\right] \Theta_k(\lambda) \,. \tag{36}$$

Here Θ measures the probability that at least one of the neighbors of a vertex of degree k is infected. A MF-treatment of Θ, excluding degree correlations, makes it possible to show that the SIS-model may have a zero threshold, and also establishes other interesting properties as $\langle \rho_k \rangle$. As is natural, the branching process concentrates on vertices with higher degrees, that is ρ_k increases with k.

The consequences of a zero epidemic threshold are plentiful and important [40]. The original application of the theory was to the spreading of computer viruses: the Internet is (on the Autonomous System level) a scale-free network [41]. So is the network of websites. The outcome even for subcritical epidemics is a long life-time, and further complications ensue if the internal dynamics of the spreading do not follow Poissonian statistics of time (e. g. the waiting time before vertex i sends a virus to a neighbor), but a power-law distribution [42]. Another highly topical issue is the behavior and control of human-based disease outbreaks. How to vaccinate against and isolate a dangerous virus epidemy depends on the structure of the network on which the spreading takes place, and on the detailed rules and dynamics. An intuitively easy idea is to concentrate on the "hubs", on the most connected vertices of the network [43]. This can be an airport through which travellers acting as carriers move, or it can – as in the case of sexually transmitted viruses such as HIV – be the most active individuals. There are indeed claims that the network of sexual contacts is scale-free, with a small exponent α. Of interest here is the recent result that for $\lambda > \lambda_c$ and $\alpha < 3$ the spreading of branching processes deviates from usual expectations. The outgrowth covers a finite fraction of the network in a short time (vanishing in the large-N limit), and the growth in time of the infected nodes becomes polynomial, after an initial exponential phase [44].

Branching processes can also be applied to network structure. The "standard model" of growing networks is the Barabasi–Albert model. In it, from a small seed graph one grows an example network by so-called preferential attachment. That is, there are two mechanisms that operate: i) new vertices are added, and ii) old ones grow more links (degree increases) by getting linked to new ones. There is a whole zoo of various network growth models, but we next overview some ideas that directly connect to

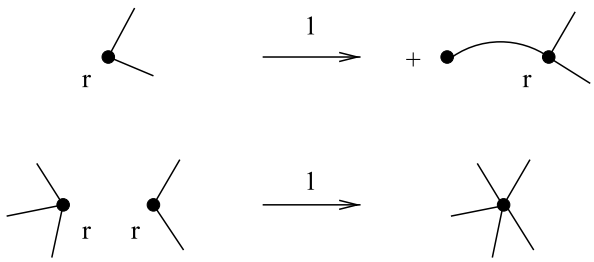

Branching Processes, Figure 11
Two processes creating a growing or steady-state network: addition and merging of vertices

branching processes. These apply to cases in which vertices disappear by combining with other vertices, are generated from thin air, or split into "sub-vertices" such that old links from the original one are retained. One important mechanism, operative in cellular protein interaction networks, is the duplication of vertices (a protein) with slight changes to the copy's connectivity from the parent vertex.

Figure 11 illustrates an example where the structure in the steady-state can be fruitfully described by similar mathematics as in other similar cases in networks. The basic rules are two-fold: i) a new vertex is added to the network, ii) two randomly chosen vertices merge. This kind of mechanisms are reminiscent of aggregation processes, of e. g. sticky particles that form colloidal aggregates in fluids. The simple model leads to a degree distribution of the asymptotic form $P(k) \sim k^{-3/2}$, reminiscent of the mean-field branching process result for the avalanche size distribution (where $\tau = 3/2$, cf. Sect. "Branching Processes") [45,46]. The mathematical tool here is given by rate equations that describe the number of vertices with a certain degree, k. These are similar to those that one can write for the size of an avalanche in other contexts. It is an interesting question of what kind of generalizations one can find by considering cases where the edges have weights, and the elementary processes are made dependent in some way on those [47].

It is worth noting that the aggregation/branching process description of network structure develops easily deep complexity. This can be achieved by choosing the nodes to be joined with an eye to the graph geometry, and/or splitting vertices with some particular rules. In the latter case, a natural starting point is to maintain all the links of the original vertex and to distribute them among the descendant vertices. Then, one has to choose whether to link those to each other or not – either choice influences the local correlations. The same is naturally true, if one considers reactions of the type $k + x' \to k''$, which originate from joining two neighboring vertices with degrees k and x'. The likelihood of this rate is dependent on the conditional probability of a vertex with degree k having a neighbor with x'. Such processes are dependent on structural properties of the networks, which themselves change in time, thus the theoretical understanding is difficult due to the correlations. As a final note, in some cases the merging of vertices produces to the network a *giant component*, a vertex which has a finite fraction of the total mass (or, links in the whole). This is analogous to avalanching systems which are "weakly first-order", in other words exhibit statistics which has a power-law, scale-free part and then a separate, singular peak. Often, this would be a final giant avalanche.

Conclusions

In this short overview we have given some ideas of how to understand complexity via the tool of branching processes. The main issue has been that they are an excellent means of understanding "criticality" and "complexity" in many systems. We have concentrated on two particular applications, SOC and networks, to illustrate this. Many other important fields where BP-based ideas find use have been left out, from biology and the dynamics of species and molecules to geophysics and the spatial and temporal properties of say earthquakes. An example is the so-called ETAS model used for their modeling (see [48,49,50]).

Future Directions

The applications of branching processes to complex systems continue on various fronts. One can predict interesting developments in a number of cases. First, as indicated the dynamics of BP's on networks are not understood very well at all when it comes to the possible scenarios. A simple question is, whether the usual language of absorbing state phase transitions applies, and if not why – and what is the effect of the heterogeneous geometry of complex networks. In many cases the structure of a network is a dynamic entity, which can be described by generalized branching processes and there has been so far relatively little work into this direction.

The inclusion of spatial effects and temporal memory dynamics is another interesting and important future avenue. Essentially, one searches for complicated variants of usual BP's to be able to model avalanching systems or cases where one wants to compute the typical time to reach an absorbing state, and the related distribution. Or, the question concerns the supercritical state ($\lambda > \lambda_c$) and the spreading from a seed. As noted in the networks section this can be complicated by the presence of non-Poissonian temporal statistics. Another exciting future task is the de-

scription of non-Markovian phenomena, as when for instance the avalanche shape is non-symmetrical [51]. This indicates that there is an underlying mechanism which needs to be incorporated into the BP model.

Acknowledgments

We are grateful to our colleague Stefano Zapperi with whom we have collaborated on topics related to networks, avalanches, and branching processes. This work was supported by the Academy of Finland through the Center of Excellence program (M.J.A.) and EUMETSAT's GRAS Satellite Application Facility (K.B.L.).

Bibliography

Primary Literature

1. Harris TE (1989) The Theory of Branching Processes. Dover, New York
2. Kimmel M, Axelrod DE (2002) Branching Processes in Biology. Springer, New York
3. Athreya KB, Ney PE (2004) Branching Processes. Dover Publications, Inc., Mineola
4. Haccou P, Jagers P, Vatutin VA (2005) Branching Processes: Variation, Growth, and Extinction of Populations. Cambridge University Press, Cambridge
5. Manna SS (1991) J Phys A 24:L363. In this two-state model, the energy takes the two stable values, $z_i = 0$ (empty) and $z_i = 1$ (particle). When $z_i \geq z_c$, with $z_c = 2$, the site relaxes by distributing two particles to two randomly chosen neighbors
6. Dickman R, Alava MJ, Munoz MA, Peltola J, Vespignani A, Zapperi S (2001) Phys Rev E64:056104
7. Bak P, Tang C, Wiesenfeld K (1987) Phys Rev Lett 59:381; (1988) Phys Rev A 38:364
8. Tebaldi C, De Menech M, Stella AL (1999) Phys Rev Lett 83:3952
9. Stella AL, De Menech M (2001) Physica A 295:1001
10. Vespignani A, Dickman R, Munoz MA, Zapperi S (2000) Phys Rev E 62:4564
11. Dickman R, Munoz MA, Vespignani A, Zapperi S (2000) Braz J Phys 30:27
12. Alava M (2003) Self-Organized Criticality as a Phase Transition. In: Korutcheva E, Cuerno R (eds) Advances in Condensed Matter and Statistical Physics. arXiv:cond-mat/0307688,(2004) Nova Publishers, p 45
13. Lubeck S (2004) Int J Mod Phys B18:3977
14. Jensen HJ (1998) Self-Organized Criticality. Cambridge University Press, Cambridge
15. Alava MJ, Nukala PKNN, Zapperi S (2006) Statistical models of fracture. Adv Phys 55:349–476
16. Albert R, Barabási AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47
17. Dorogovtsev SN, Mendes JFF (2003) Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford University Press, Oxford; (2002) Adv Phys 51:1079; (2004) arXiv:cond-mat/0404593
18. Bonachela JA, Chate H, Dornic I, Munoz MA (2007) Phys Rev Lett 98:115702
19. Feller W (1971) An Introduction to Probability Theory and its Applications. vol 2, 2nd edn. Wiley, New York
20. Asmussen S, Hering H (1983) Branching Processes. Birkhäuser, Boston
21. Weiss GH (1994) Aspects and Applications of the Random Walk. North-Holland, Amsterdam
22. Tang C, Bak P (1988) J Stat Phys 51:797
23. Dhar D, Majumdar SN (1990) J Phys A 23:4333
24. Janowsky SA, Laberge CA (1993) J Phys A 26:L973
25. Flyvbjerg H, Sneppen K, Bak P (1993) Phys Rev Lett 71:4087; de Boer J, Derrida B, Flyvbjerg H, Jackson AD, Wettig T (1994) Phys Rev Lett 73:906
26. Alstrøm P (1988) Phys Rev A 38:4905
27. Christensen K, Olami Z (1993) Phys Rev E 48:3361
28. García-Pelayo R (1994) Phys Rev E 49:4903
29. Zapperi S, Lauritsen KB, Stanley HE (1995) Phys Rev Lett 75:4071
30. Manna SS, Kiss LB, Kertész J (1990) J Stat Phys 61:923
31. Tadić B, Nowak U, Usadel KD, Ramaswamy R, Padlewski S (1992) Phys Rev A 45:8536
32. Tadic B, Ramaswamy R (1996) Phys Rev E 54:3157
33. Vespignani A, Zapperi S, Pietronero L (1995) Phys Rev E 51:1711
34. Lauritsen KB, Zapperi S, Stanley HE (1996) Phys Rev E 54:2483
35. Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45:167
36. Yook SH, Jeong H, Barabasi AL, Tu Y (2001) Phys Rev Lett 86:5835
37. Barrat A, Barthelemy M, Pastor-Satorras R, Vespignani A (2004) Proc Natl Acad Sci USA 101:3747
38. Barrat A, Barthelemy M, Vespignani A (2004) Phys Rev Lett 92:228701
39. Pastor-Satorras R, Vespignani A (2001) Phys Rev Lett 86:3200
40. Dorogovtsev SN, Goltsev AV, Mendes JFF (2007) arXiv:cond-mat/0750.0110
41. Pastor-Satorras R, Vespignani A (2004) Evolution and Structure of the Internet: A Statistical Physics Approach. Cambridge University Press, Cambridge
42. Vazquez A, Balazs R, Andras L, Barabasi AL (2007) Phys Rev Lett 98:158702
43. Colizza V, Barrat A, Barthelemy M, Vespignani A (2006) PNAS 103:2015
44. Vazquez A (2006) Phys Rev Lett 96:038702
45. Kim BJ, Trusina A, Minnhagen P, Sneppen K (2005) Eur Phys J B43:369
46. Alava MJ, Dorogovtsev SN (2005) Phys Rev E 71:036107
47. Hui Z, Zi-You G, Gang Y, Wen-Xu W (2006) Chin Phys Lett 23:275
48. Ogata Y (1988) J Am Stat Assoc 83:9
49. Saichev A, Helmstetter A, Sornette D (2005) Pure Appl Geophys 162:1113
50. Lippidello E, Godano C, de Arcangelis L (2007) Phys Rev Lett 98:098501
51. Zapperi S, Castellano C, Colaiori F, Durin G (2005) Nat Phys 1:46

Books and Reviews

Albert R, Barabási AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47

Asmussen S, Hering H (1883) Branching Processes. Birkhäuser, Boston

Athreya KB, Ney PE (2004) Branching Processes. Dover Publications, Inc., Mineola
Feller W (1971) An Introduction to Probability Theory and its Applications. vol 2, 2nd edn. Wiley, New York
Haccou P, Jagers P, Vatutin VA (2005) Branching Processes: Variation, Growth, and Extinction of Populations. Cambridge University Press, Cambridge
Harris TE (1989) The Theory of Branching Processes. Dover, New York
Jensen HJ (1998) Self-Organized Criticality. Cambridge University Press, Cambridge
Kimmel M, Axelrod DE (2002) Branching Processes in Biology. Springer, New York
Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45:167
Weiss GH (1994) Aspects and Applications of the Random Walk. North-Holland, Amsterdam

Brittle Tectonics: A Non-linear Dynamical System

Christopher H. Scholz
Lamont-Doherty Earth Observatory,
Columbia University, New York, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Scaling Relations: Self-similarity
Earthquake and Fault Populations: Self-organization
Models of the System
Future Directions
Bibliography

Glossary

Ductile shear zone A quasi-planar tabular zone of localized shear deformation in the semi-brittle to fully plastic regimes.

Earthquake Dynamically running shear instability on a fault.

Fault A shear crack with friction between its interfaces.

Mylonite A metamorphic rock with a fabric produced by shear deformation.

Suprafault The shear relaxation structure that includes a fault and its associated ductile shear zone.

Definition of the Subject

Brittle deformation is the primary mode of deformation of Earth's crust. At the long timescale it is manifested by faulting, and on the short timescale by earthquakes. It is one of the best-known examples of a system exhibiting self-organized criticality. A full understanding of this system is essential to the evaluation of earthquake hazard.

Introduction

The upper part of Earth's crust is brittle and under a state of all-round compression. It responds to deformation by faulting: the formation and propagation of shear cracks. The crack walls support normal stresses and hence fault propagation must overcome not only the rupture resistance of the fault tips but friction between its interior interfaces. This friction is usually velocity weakening, such that any slippage results in stick-slip instability. The resulting dynamically running crack-like shear instability radiates elastic waves, producing the shaking known as an earthquake. Thus brittle tectonics separates into two greatly different timescales: the long timescale of the faulting and the short timescale of the earthquakes. Faults propagate quasi-statically at rates set by tectonic driving rates, whereas earthquakes propagate at a rate set by the shear velocity of the rock. Almost all displacement on continental faults occurs by earthquakes, so that there is a coupling between the two timescales.

This system is illustrated in Fig. 1a. The term *fault* is usually restricted to the rupture of the brittle part, but large faults continue downwards as ductile shear zones, which, in the case of major faults, may extend as deep as the crust-mantle boundary. We will refer to the entire shear release structure, including both the fault and its corresponding ductile shear zone as the *suprafault*.

The division between brittle and ductile behavior occurs at what is usually called the brittle-ductile transition. This transition is gradual, with a change in deformation mechanism from brittle crack propagation and friction to fully plastic deformation through a region of mixed behavior known as semi-brittle. For our purposes, however, it is more useful to divide the suprafault into two regimes, of velocity weakening and velocity strengthening rheologies, at a depth S. These regions are predicted by a rate/state variable friction law that has been established in the laboratory (for a review, see [25]). Here, we also refer to the ductile regime as velocity strengthening because the creep law for velocity strengthening friction is the same form as for ductile creep. S corresponds to the upper onset of the brittle-ductile transition and it varies regionally because it depends on temperature (heat flow) and strain rate. For faults in continental crust it varies between about 10–20 kms. Earthquakes can only nucleate above S, so this constitutes the *seismogenic zone*. When earthquakes be-

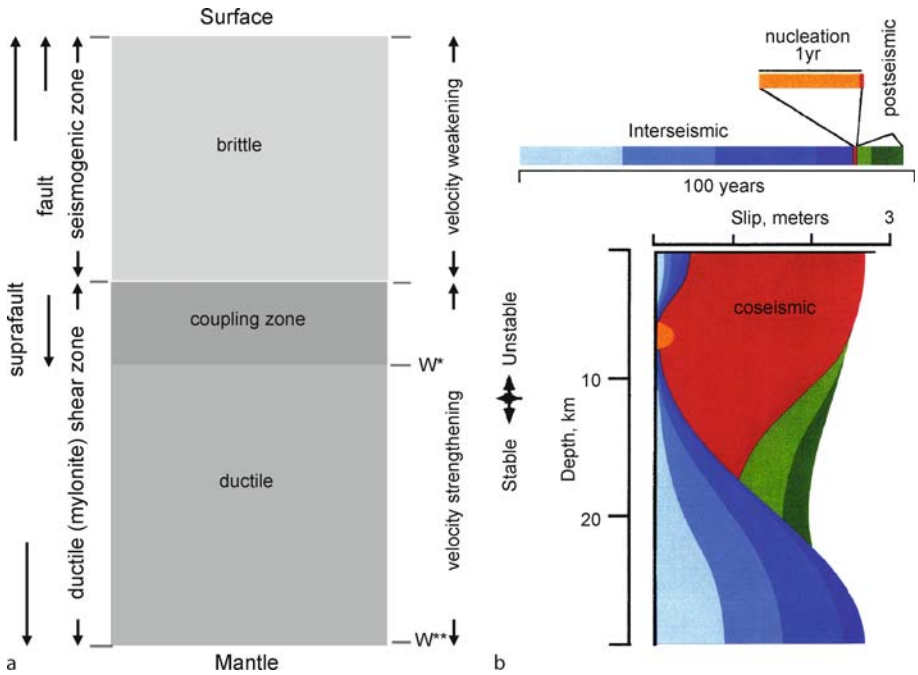

Brittle Tectonics: A Non-linear Dynamical System, Figure 1

a A schematic longitudinal section of the rheological components of a crustal suprafault. b A simulation of a model with the properties of (a), shown throughout a seismic cycle (after Tse and Rice [37])

come large, they can, owing to the very high strain rates at their rupture tips, propagate below S to W^*, which defines the maximum width of earthquakes. Similarly, the maximum width of a suprafault is W^{**}.

Although the ductile region is velocity (strain-rate) strengthening, it is also, seemingly paradoxically, strain softening, which leads to strain localization into narrow shear zones. As the rock shears, it gradually reorganizes internally to form a mylonitic fabric, which facilitates further shearing. With finite strain, the mylonitic fabric develops further and the mylonitic foliation rotates closer to the plane of shear, progressively reducing the rock's resistance to shear in that particular direction and plane. It is this mechanism of *fabric weakening* that produces the shear localization.

The system has two characteristic length scales, W^* and W^{**}. An earthquake nucleates within the seismogenic zone and initially propagates in all directions along its perimeter, acting as a 3D crack. When its dimension exceeds W^*, it has breached the free surface and penetrated to W^* and then is prohibited from growing deeper. It is then restricted to propagating in the horizontal direction, acting as a 2D crack. Thus there is a symmetry breakage at the dimension W^*. *Small* earthquakes, with dimensions smaller than W^*, are not self-similar with *large* earthquakes, those with lengths larger than W^*. The same occurs for suprafaults at the dimension W^{**}.

The coupling of the system is illustrated in Fig. 1b. This simulation shows the deformation of a strike slip suprafault (one in which the slip is horizontal), over the seismic cycle, i. e. the recurrence time of a large earthquake. The recurrence time depends on the long-term (geologic) slip rate of the suprafault. During the *interseismic* period, the ductile region shears at a constant rate and gradually propagates upwards. This loads the upper region until nucleation occurs above S as a quasi-static crack that eventually goes dynamic, resulting in an earthquake that propagates over S and down to W^*. The ductile region is instantaneously reloaded by the downward propagation of the earthquake. This results in *postseismic* relaxation that exponentially decays over a period of years to decades.

Scaling Relations: Self-similarity

Earthquakes

The primary scaling relation for cracks is that between the shear displacement (slip), D, for faults or ΔD for earthquakes, and the dimension L of the rupture. The ratio of these parameters is a stress-drop $\Delta\sigma$ normalized by an

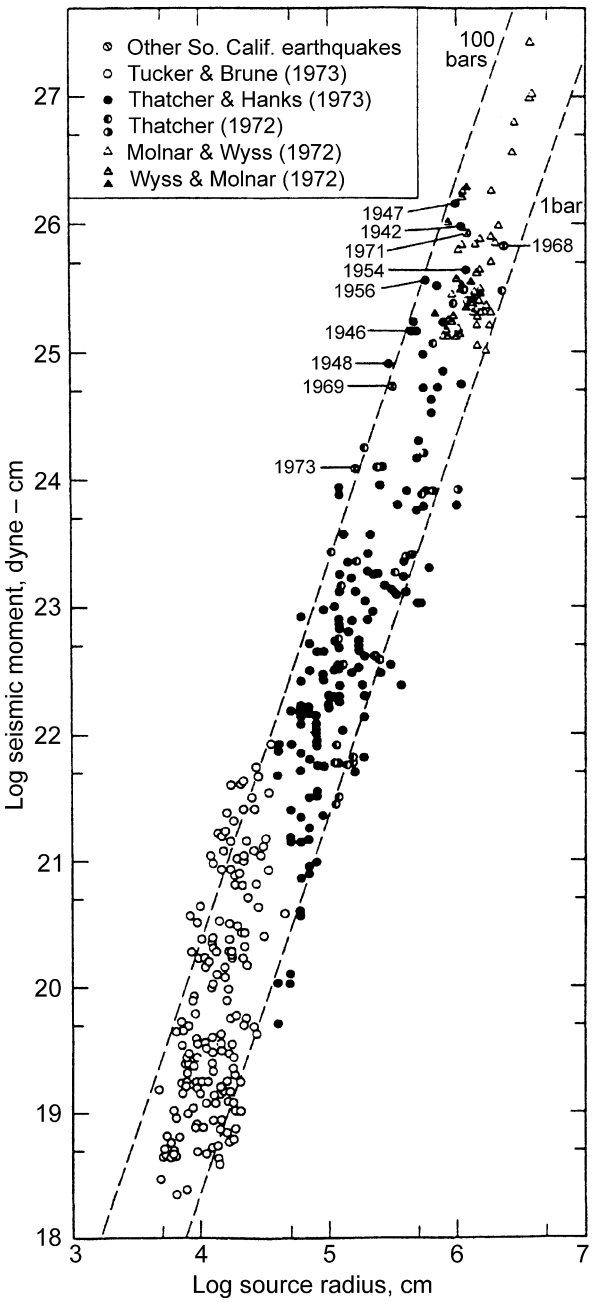

Brittle Tectonics: A Non-linear Dynamical System, Figure 2
A collection of data for small earthquakes showing the relationship between seismic moment $M_0 = \mu \Delta DA$ and source radius. *Dashed lines* are of constant stress drop. From Hanks [15]

elastic constant. Figure 2 shows data for small earthquakes, in which seismic moment $M_0 = \mu \Delta DA$ is plotted vs. L (μ is the shear modulus and A is rupture area). Because these are 3D cracks the L^3 lines are loci of equal stress-drop. Although there is considerable scatter, the data follow this trend. Thus, while $\Delta\sigma$ varies by about 2 orders of magnitude among earthquakes, it is scale-invariant. These earthquakes are thus self-similar.

A plot of slip vs. length for large earthquakes is shown in Fig. 3. What is seen is that initially ΔD increases linearly with L and, following a long crossover, becomes constant. From the static analysis of a 2D crack, we would expect, for constant stress-drop, that ΔD be proportional to W^*, and hence constant on a plot like this, but this is seen only for the rare earthquakes with aspect ratios greater than 10. A static scaling analysis suggests that the crossover be complete by an aspect ratio of 4. A numerical model of dynamic shear cracking produced the same delayed crossover [30]. Further study of the same model suggests that the long rollover results from significant slip below the seismogenic depth [31]. The scaling of large earthquakes thus exhibit two self-similar regimes, neither of which is self-similar with small earthquakes.

The two ΔD scales in Fig. 3, for interplate and intraplate earthquakes, illustrate the rate effect on stress drop. Earthquakes on intraplate faults, for which geological sliding rates are several orders of magnitude less than for interplate faults, have, on average, stress drops about three times larger than interplate earthquakes. Such fault healing is expected from the rate/state friction constitutive law [4].

Faults

The displacement profiles for some faults in the same rock type, ranging in length from 690 to 2200 m, are shown, normalized by length, in Fig. 4. The data collapse shows the self-similarity of these profiles. Displacement amplitude scales linearly with length, whereas the linear displacement tapers as the tips are approached are scale invariant. A global plot of D vs. L is shown in Fig. 5. There is, similar to earthquakes, considerable scatter, but the overall trend shows a linear relationship between D and L. Here more can be said about the origin of the scatter than in the case of earthquakes. Those faults with D/L ratios less than 10^{-2} are in soft sedimentary rock at shallow depth; those with greater D/L are in strong metamorphic or igneous rock. Faults of length greater than 5 km extend increasingly deeper into the crust and exhibit an increase in D/L (stress-drop) owing to the increase of strength with pressure. Scatter is also produced by the interaction of faults through their stress fields, which distorts their displacement profiles (see pp. 126–129 in [27]). These effects of lithology and fault interaction also occur for the displacement tapers near the fault tips, which are another measure of strength [28]. In a study in which the faults are in a sin-

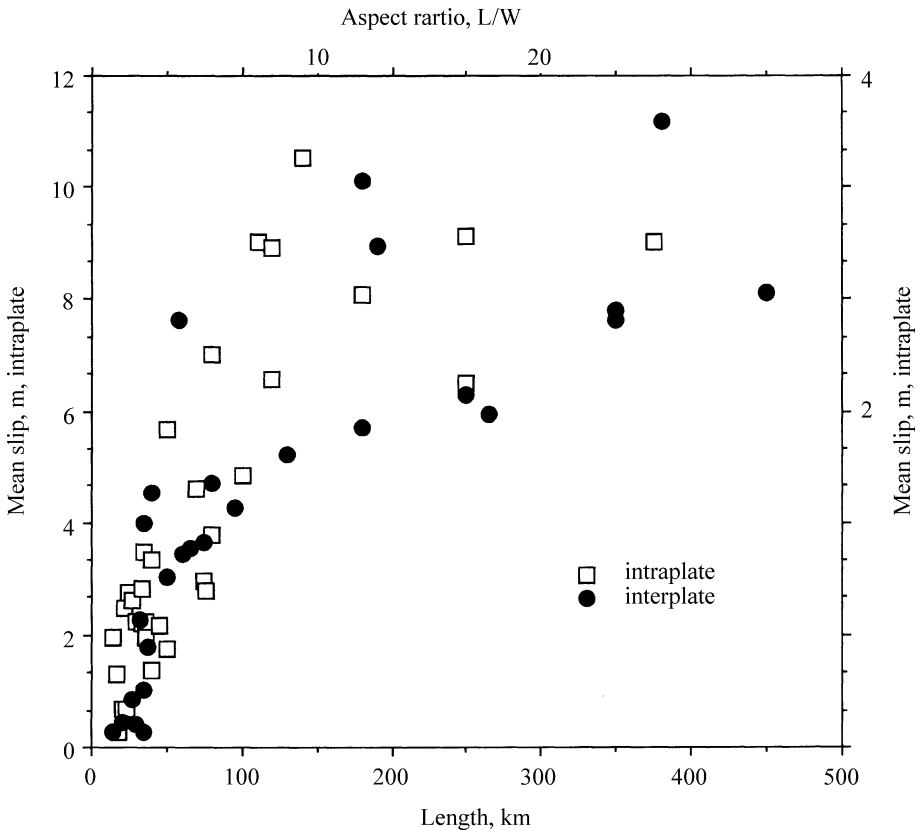

Brittle Tectonics: A Non-linear Dynamical System, Figure 3
Mean slip vs. rupture length for large earthquakes. Two scaling regimes are observed. Slip increases approximately linearly with *L* up to an aspect ratio of about 10, and is independent of length thereafter. After Scholz [22]

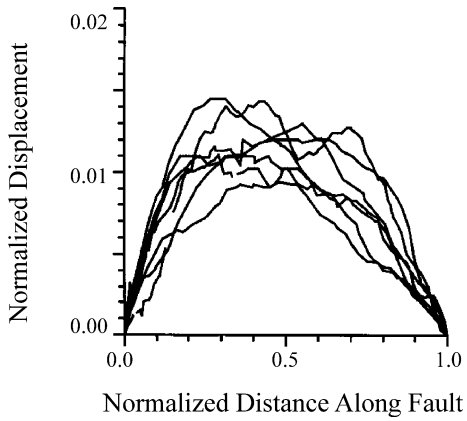

Brittle Tectonics: A Non-linear Dynamical System, Figure 4
Displacement profiles along faults of different length in the Volcanic Tablelands of eastern California, normalized to length. Data from Dawers et al. [13]

gle lithology and were selected to be isolated [13], the scatter in both D/L and fault tip taper was correspondingly reduced (as in Fig. 4).

We do not have data for long enough faults to observe a crossover from 3D to 2D scaling, as we did for earthquakes. Because $W^{**} \sim 40$ km, we would need data from faults of length exceeding 400 km to clearly see this effect. Although faults of that length and longer exist, they are predominately plate boundaries, which have no tips because their ends connect to other plate boundaries and are stress relieved by them. Therefore the scaling of Fig. 5 does not apply to them. The San Andreas fault in California, for example, terminates at its southern end at a spreading ridge. Its net slip is more than an order of magnitude greater than would be expected from simple application of the scaling shown in Fig. 5.

Faults, like earthquakes, are self-similar. The scaling parameters relating D and L differ greatly, however. For faults it is $\sim 10^{-2}$ and for earthquakes, 10^{-4}–10^{-5}. Both represent stress-drops. For faults, it is the stress-drop from the fracture strength of intact rock to the residual frictional

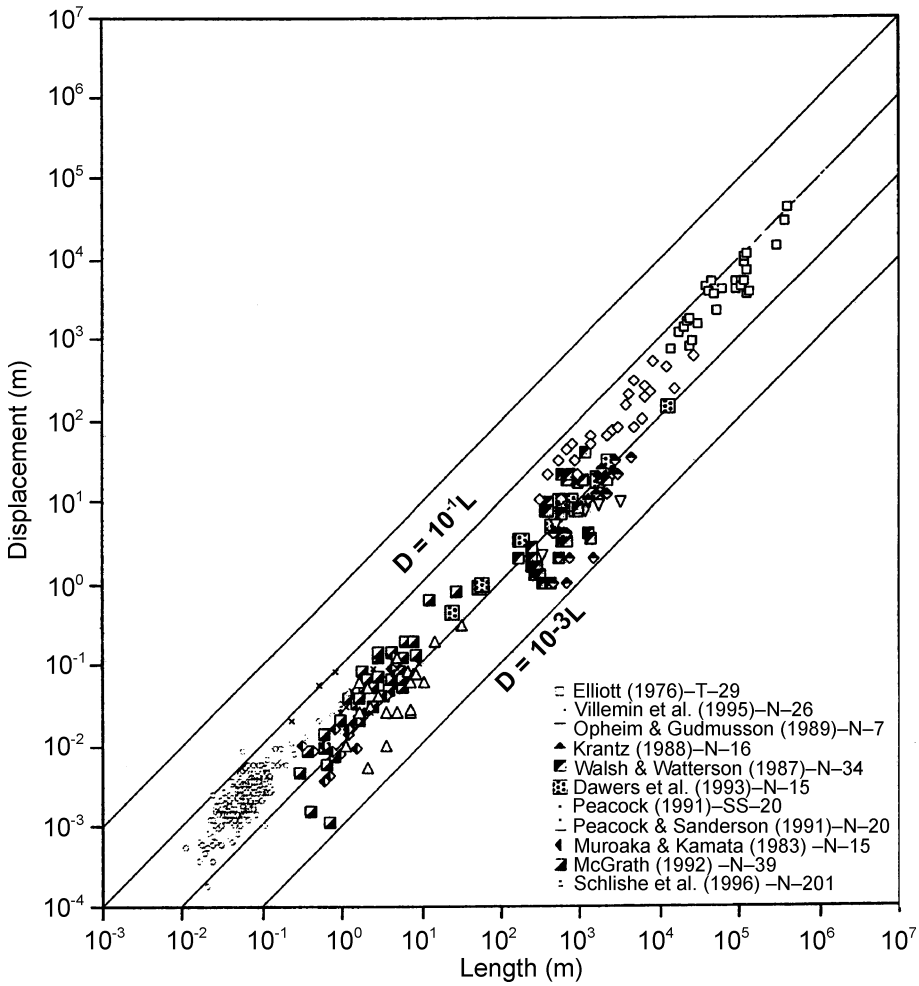

Brittle Tectonics: A Non-linear Dynamical System, Figure 5
Maximum slip vs. length for fault data sets from various tectonic settings. Modified from Schlische et al. [20]

strength. For earthquakes, it is the stress-drop from static to dynamic friction.

Earthquake and Fault Populations: Self-organization

Earthquakes and faults both self-organize into populations with well defined statistical properties, thus exhibiting complexity in the physics of this system.

Earthquakes

The frequency $N(M)$ of earthquakes greater than magnitude M has long been known to obey the Gutenberg–Richter law, $\log N(M) = a - bM$ in which $b \approx 1$ is found universally for small earthquakes. In laboratory experiments b is found to decrease with stress [21]. The fact that b is narrowly limited in the Earth's crust may be because stress in the crust is limited to a narrow range by the frictional strength of faults [36]. M is a logarithmic measure of energy E or seismic moment, M_0. In terms of M_0 this relation is $N(M_0) = aM_0^{-B}$ where $B = 2/3b$. However, as shown in Fig. 6, in the global catalog of earthquakes there is a crossover from $b = 1$ to $b = 1.5$ at about $M7.5$. This corresponds to a rupture dimension of about 50 km, which is an average value of W^* for subduction zones, which dominate this data set. For a subset of continental earthquakes, the same crossover is observed at $M6.5$, corresponding to a rupture length of 10–20 km, similar to W^* for crustal faults. Thus in terms of a linear measure of earthquake size like E or M_0, earthquakes have a power law size distribution. Where there is a breakdown of self-similarity between small and large earthquakes there is a corresponding change in the exponent B in this distribution.

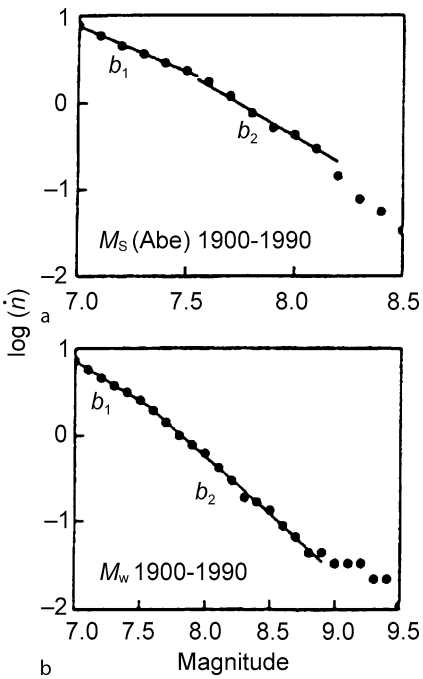

Brittle Tectonics: A Non-linear Dynamical System, Figure 6
The frequency size distribution from a global cataloger of Pacheco and Sykes [18] for two magnitude scales. From Pacheco et al. [19]

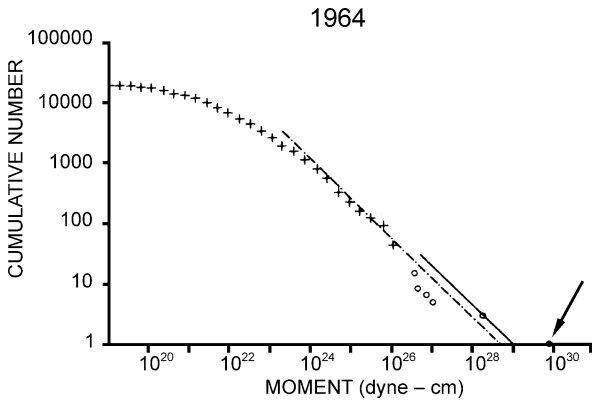

Brittle Tectonics: A Non-linear Dynamical System, Figure 7
Distribution of small earthquakes in the rupture zone of the 1964 Alaskan *M*9.4 earthquake, normalized to the recurrence time of that earthquake. An *arrow* indicates the 1964 earthquake. It is about 1 ½ times larger than an extrapolation of the small earthquakes would indicate. The roll off at $M_0 < 3 \times 10^{23}$ dyne cm is cause by a lack of perceptibility of smaller events. From Davison and Scholz [11]

The size distribution discussed above applies to regions large enough to include many active faults or plate boundary segments. If we study instead a single fault or plate boundary segment that ruptures in a single large earthquake and consider earthquakes that occur within a single seismic cycle of that earthquake, a different picture emerges. An example is shown in Fig. 7. There we see that the small earthquakes during that interval obey a power law with $B = 2/3$, as usual, but that the single large earthquake is more than an order of magnitude greater than expected from extrapolation of the small events. This is another manifestation of large and small earthquakes belonging to different fractal sets.

Faults

Fault data are usually obtained by geological mapping. Because mapping has a dynamic range limited to about one order of magnitude, it is difficult to convincingly demonstrate a power law size distribution with this method. When the results of mapping at several scales are combined [29], the case for a power law is strengthened, but the exponent is poorly determined. This problem was overcome with the data in Fig. 8a, from a Magellan SAR image of normal faults on the plains of Venus. This image has a far greater dynamic range, and unequivocally demonstrates a power law distribution, with an exponent −2. This is a frequency plot, rather than a cumulative plot such as shown in Fig. 7.

On close examination, faults are found to be not a continuous surface, but a complex of myriad strands or subfaults. A cumulative size distribution of the subfaults of the San Andreas fault is shown in Fig. 8b, where they are also shown to obey a power law. The upper fractal limit of the subfaults appears to be W^*.

The size distributions of earthquakes and faults are not unrelated. If one equates faults with large earthquakes and subfaults with small earthquakes, then by convolving over the fault distributions of Fig. 8 one can obtain the earthquake distributions of Figs. 6 and 7 [24,26].

Not all fault distributions are power law, however. Figure 9 shows the size distribution of faults on the flanks of the East Pacific Rise, obtained by sonar swath mapping. These clearly obey an exponential rather than a power law distribution. One can also find examples of periodic distributions, such as the regularly spaced faults of the Basin and Range province of the Western US. We show later that three regimes, characterized by power law, exponential, and periodic distributions, evolve with increasing strain.

Models of the System

The first model of the earthquake system was a bench-top apparatus of a string of weights connected by springs and pulled along a frictional surface [6]. Their main finding

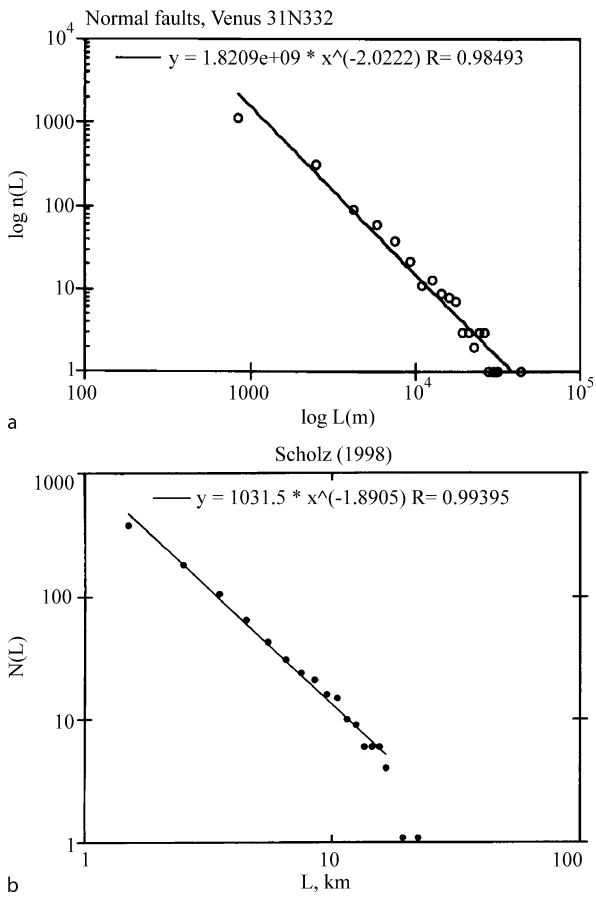

Brittle Tectonics: A Non-linear Dynamical System, Figure 8
Power law size distributions of faults. a Frequency-length distribution for faults on the plains of Venus from a Magellan SAR image [23]. b Cumulative length distribution of subfaults of the San Andreas fault [26]

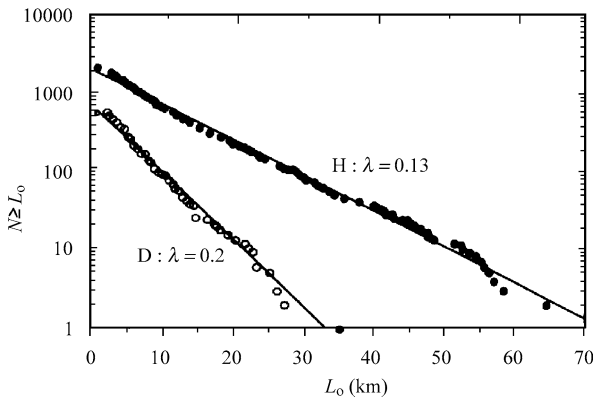

Brittle Tectonics: A Non-linear Dynamical System, Figure 9
The distribution of faults at two places on the flanks of mid-ocean ridges. These distributions are exponential, rather than power law. From Cowie et al. [9]

was that this model obeyed the Gutenberg–Richter law. Interest in this result was revived by the paper of Bak, Tang, and Wiesenfeld [3], which introduced the concept of *self-organized criticality*.

Bak et al. [3] illustrated this concept with a cellular automata model of a sandpile, in which the cells are governed by a nearest neighbor rule for avalanches and the system is driven by the slow random addition of grains to the pile. The result was a power law distribution of avalanches and a $1/f$ energy flux. From this they hypothesized that systems containing many interacting elements may exhibit general statistical properties that are described by power laws. They called this behavior self-organized criticality (SOC). Although this concept is intuitively appealing, there is no clear-cut definition of what SOC is.

Because the Gutenberg–Richter law is the most famous of such power laws, adaptations of this cellular automata model to earthquakes soon followed (e. g. [2,5]). Carlson and Langer [7] explored a numerical spring-slider model that contained the full dynamics of the slider blocks. All these models reproduced the Gutenberg–Richter law and also explored other aspects of the system, for example, in the case of Brown et al. [5], the distribution of recurrence times of large earthquakes. Olami et al. [17] explored a non-conservative spring block model, and Christensen and Olami [8] showed how Omori's law of aftershocks might be explained without having to appeal to an added viscosity as did Burridge and Knopoff. Sornette and Virieux [32], using a mean field theory approach, showed that the scaling difference between small and large earthquakes results in the change in b value shown in Fig. 6. Extended discussion of these models and their results can be found in Turcotte [38] and Jensen [16].

The study of the self-organization of faults was done first in an analog model of continental deformation by Davy et al. [12], who found a geometrically fractal fault distribution. Cowie et al. [10] studied a model simulating a brittle sheet undergoing stretching. For large strains this model converges on a single system size crack, but at small strains power law size crack distributions were observed.

Spyropoulos et al. [34] expanded on this with a spring-block model of the stretching of a brittle sheet supported by a ductile substrate. They found an evolution of the crack population with strain, from power law to exponential distributions and finally to a periodic distribution of system size cracks. In this model, the brittle layer contains only nearest neighbor interactions but the ductile substrate provides longer-range force interactions. It is the longer-range interactions that result in the evolution of the system and is what differs this model from the sandpile family of models. The evolution from a power law

distribution to an exponential distribution of cracks has been demonstrated with physical models [1,33]. It has also been demonstrated in the field, for faults in the Asal rift of Ethiopia [14].

Future Directions

Our understanding of brittle tectonics as a dynamical system is still very sketchy. Approaches to it tend to be piecewise, which leads to self-limiting results. For example, although it is now well known that earthquakes interact through their stress fields (e. g. [35]), this phenomenon is modeled by assuming that the background 'tectonic' stress field is uniform, thus ignoring the stress fields of faults, which also interact. The interaction of faults has so far been studied in only the simplest case, that of sub-parallel normal faults. For strike-slip faults, in which the fault tip stresses are asymmetric, more interesting possibilities exist, as yet unexplored. The San Andreas fault system in California is a complex of faults several hundred km wide that somehow act together to accommodate most of the Pacific-North America plate motion. How does this work? How do the faults interact in such a way to preserve continuity of plate motion along strike, as geodetic data indicate? How does the interaction between faults lead to the geometric evolution of fault patterns? Are fault slip rates stationary, as often assumed, or do they vary at intermediate time scales between that of earthquakes and that of the faults themselves? If so, can this explain the temporal clustering of earthquakes that seem to occur over spatial distances long compared to the stress fields of earthquakes? These are but a few of the problems one can imagine studying in the future, once the idea is accepted that the system must be considered in its entirety before being broken into digestible chunks.

Bibliography

Primary Literature

1. Ackermann RV, Schlische RW, Withjack MO (2001) The geometric and statistical evolution of normal fault systems: an experimental study of the effects of mechanical layer thickness on scaling laws. J Struct Geol 23:1803–1819
2. Bak P, Tang C (1989) Earthquakes as a self-organized critical phenomenon. J Geophys Res 94:15635–15637
3. Bak P, Tang C, Weisenfeld K (1987) Self-organized criticality: An explanation of 1/*f* noise. Phys Rev Lett 59:381–384
4. Beeler NM, Hickman SH, Wong TF (2001) Earthquake stress drop and laboratory-inferred interseismic strength recovery. J Geophys Res-Solid Earth 106:30701–30713
5. Brown SR, Scholz CH, Rundle JB (1991) A simplified spring-block model of earthquakes. Geophys Res Lett 18:215–218
6. Burridge R, Knopoff L (1967) Model and theoretical seismicity. Bull Seism Soc Am 57:341–362
7. Carlson JM, Langer JS (1989) Properties of earthquakes generated by fault dynamics. Phys Rev Lett 62:2632–2635
8. Christensen K, Olami Z (1992) Variation of the Gutenberg–Richter B Values and Nontrivial Temporal Correlations in a Spring-Block Model for Earthquakes. J Geophys Res-Solid Earth 97:8729–8735
9. Cowie PA, Scholz CH, Edwards M, Malinverno A (1993) Fault strain and seismic coupling on midocean ridges. J Geophys Res-Solid Earth 98:17911–17920
10. Cowie PA, Sornette D, Vanneste C (1995) Multifractal scaling properties of a growing fault population. Geophys J Int 122:457–469
11. Davison F, Scholz C (1985) Frequency-moment distribution of earthquakes in the Aleutian Arc: A test of the characteristic earthquake model. Bull Seismol Soc Am 75:1349–1362
12. Davy P, Sornette A, Sornette D (1990) Some consequences of a proposed fractal nature of continental faulting. Nature 348:56–58
13. Dawers NH, Anders MH, Scholz CH (1993) Growth of normal faults – displacement-length scaling. Geology 21:1107–1110
14. Gupta A, Scholz CH (2000) Brittle strain regime transition in the Afar depression: implications for fault growth and seafloor spreading. Geology 28:1087–1090
15. Hanks TC (1977) Earthquake Stress Drops, Ambient Tectonic Stresses and Stresses That Drive Plate Motions. Pure Appl Geophys 115:441–458
16. Jensen HJ (1998) Self-organized criticality: Emergent complex behavior in physical and biological systems. Cambridge Univ. Press, Cambridge
17. Olami Z, Feder HJS, Christensen K (1992) Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes. Phys Rev Lett 68:1244–1247
18. Pacheco JF, Sykes LR (1992) Seismic moment catalog of large shallow earthquakes, 1900 to 1989. Bull Seismol Soc Am 82:1306–1349
19. Pacheco JF, Scholz CH, Sykes LR (1992) Changes in frequency-size relationship from small to large earthquakes. Nature 355:71–73
20. Schlische RW, Young SS, Ackermann RV, Gupta A (1996) Geometry and scaling relations of a population of very small rift-related normal faults. Geology 24:683–686
21. Scholz CH (1968) The frequency-magnitude relation of microfracturing in rock and its relation to earthquakes. Bull Seismol Soc Am 58:399–415
22. Scholz CH (1994) A reappraisal of large earthquake scaling. Bull Seismol Soc Am 84:215–218
23. Scholz CH (1997) Earthquake and fault populations and the calculation of brittle strain. Geowissenschaften 3–4:124–130
24. Scholz CH (1997) Size distributions for large and small earthquakes. Bull Seismol Soc Am 87:1074–1077
25. Scholz CH (1998) Earthquakes and friction laws. Nature 391:37–42
26. Scholz CH (1998) A further note on earthquake size distributions. Bull Seismol Soc Am 88:1325–1326
27. Scholz CH (2002) The mechanics of earthquakes and faulting, 2nd edn. Cambridge University Press, Cambridge
28. Scholz CH, Lawler TM (2004) Slip tapers at the tips of faults and earthquake ruptures. Geophys Res Lett 31:L21609, doi:10.1029/2004GL021030

29. Scholz CH, Dawers NH, Yu JZ, Anders MH (1993) Fault growth and fault scaling laws – preliminary-results. J Geophys Res-Solid Earth 98:21951–21961
30. Shaw BE, Scholz CH (2001) Slip-length scaling in large earthquakes: observations and theory and implications for earthquake physics. Geophys Res Lett 28:2995–2998
31. Shaw BE, Wesnouski SG (2008) Slip-length Scaling in large earthquakes: The role of deep penetrating slip below the seismogenic layer. Bull Seismol Soc Am 98:1633–1641
32. Sornette D, Virieux J (1992) Linking short-timescale deformation to long-timescale tectonics. Nature 357:401–403
33. Spyropoulos C, Griffith WJ, Scholz CH, Shaw BE (1999) Experimental evidence for different strain regimes of crack populations in a clay model. Geophys Res Lett 26:1081–1084
34. Spyropoulos C, Scholz CH, Shaw BE (2002) Transition regimes for growing crack populations. Phys Rev E 65:056105, doi:10.1103/PhysRevE.65.056105
35. Stein RS (1999) The role of stress transfer in earthquake occurrence. Nature 402:605–609
36. Townend J, Zoback MD (2000) How faulting keeps the crust strong. Geology 28:399–402
37. Tse S, Rice J (1986) Crustal earthquake instability in relation to the depth variation of frictional slip properties. J Geophys Res 91:9452–9472
38. Turcotte DL (1999) Seismicity and self-organized criticality. Phys Earth Planet Inter 111:275–293

Books and Reviews

Sornette D (2003) Critical phenomena in natural systems: Chaos, fractals, self-organization, and disorder. Springer, Berlin
Turcotte DL (1997) Fractals and chaos in geology and geophyics. Cambridge, New York

Business Policy and Strategy, System Dynamics Applications to

James M. Lyneis
Worcester Polytechnic Institute, Worcester, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Using System Dynamics Models in Policy and Strategy Formulation and Implementation
Theory Development – Understanding the Drivers of Business Dynamics
Applications and Case Examples
Future Directions
Bibliography

Glossary

Business policy and strategy A firm's business strategy defines how and where it competes, and its approach to doing so. A business strategy typically specifies a firm's goals, the products and services offered and the markets served, and the basis for competing (price, service, quality, etc.). A strategy may also define the organization structure, systems and policies which implement the strategy. In addition, firm's will have systems and policies which focus on operations and functions, and are not truly "strategic" in nature. Nevertheless, these operational policies can be important in determining business performance.

Business dynamics Business dynamics is the study of how the structure of a business (or a part of the business), the policies it follows, and its interactions with the outside world (customers, competitors, suppliers) determine its performance over time. Business structure consists of feedback loops surrounding the stocks and flows of resources, customers, and competitive factors that cause change over time; business policies are important components of these feedback loops. Business dynamics is a means of determining the likely performance that will result from alternative business policies and strategies.

Definition of the Subject

System dynamics has long been applied to problems of business performance. These applications range from operational/functional performance to overall strategic performance. Beginning with its founding at MIT's Sloan School of Management in 1957, an important focus of research, teaching, and application has been on understanding why companies and markets exhibit cycles, or underperform competitors in terms of growth or profitability. The original publication in the field was Forrester's Industrial Dynamics [26], which not only laid the theoretical foundations for the field, but also provided an understanding of the causes of instability in supply chains. Since that initial work, research and application has been widespread. It has addressed the dynamics underlying instability in manufacturing and service organizations, the processes which encourage or inhibit growth, the dynamics of research organizations, and the causes of cost and schedule overruns on individual projects. It has been applied in many industries, from manufacturing to high-tech to financial services and utilities, both by academics and consultants. Business theory and applications are taught at many universities, including but not limited to MIT, London Business School and others in England, Bergen (Nor-

way), Manheim and Stuttgart (Germany) (see [62,72] for more details). Business policy and strategy has and will continue to be one of the major application areas for system dynamics.

Introduction

Business strategy, sometimes called simply 'policy' or 'strategy', is primarily concerned with how and where firm's choose to compete. It includes such decisions as setting goals, selecting which products and services to offer in which markets, establishing the basis for competing (price, service, quality, etc.), determining the organization structure, systems and policies to accomplish the strategy, and designing policies for steering that strategy continually into the future. Academic and applied research on business strategy developed separately from system dynamics. That research, while widely disparate, has largely focused on static assessments and tools. For example, cross-sectional studies of many companies attempt to identify key differences that determine success or failure as a guide to management; "strategic frameworks" (for example, learning curves, growth share matrices, Porter's five forces [79,80]) assist managers in framing strategy and intuitively assessing performance over time; scenario planning helps managers visualize alternative futures; the resource-based view of the firm and core competencies [81] help managers identify how resources and capabilities determine the best way to compete. While these tools provide valuable insights and frameworks, they leave the connection between a firm's business strategy and the evolution of its performance over time to the intuition of managers – while traditional business strategy addresses the starting point and the desired end point, and the mechanisms that might allow the firm to transition between the two, the ability of those mechanisms to achieve that transition, and the path for getting between the two, is left unanswered.

Academic and applied research on operational and functional performance has similarly developed separately from system dynamics. Although it is difficult to generalize, this research is again typically static in nature and/or focused on the detailed management of a part of the organization over a relatively short period of time (for example, optimization of production scheduling during a month, quarter or year; optimal inventory management during a quarter or year). While this detailed management is necessary for running a business, it often overlooks the longer run implications of the policies established to manage the business in the short run, and of the impacts of one part of the business on other parts.

In contrast, system dynamics addresses how structure (feedback loops, stocks and flows) and policies determine performance over time – how does the firm, or a part of the firm, get from its current state to some future state. Evolving from this structural theory, system dynamicists have studied why firms and industries exhibit instability and cycles, and why firms grow or decline. Two real examples of problematic behavior over time are shown in Fig. 1. The example on the left shows the pattern of orders for commercial jet aircraft – a system dynamicist would try to understand why the orders are cyclical, and what can be done to make them less so (or to take advantage of the cycles by forecasting that cyclicality); the example on the right shows market shares of major players in a recently deregulated telecom market – a system dynamicist, working for the incumbent telecom, would try to understand the causes of market share loss and what can be done to reverse that loss.

From its beginnings in the late 1950s, system dynamics has been used to progressively develop structural theories to explain instability in supply chains [26], cycles of growth [28], boom and bust in product sales [100], and cost and schedule overrun on projects [54], to mention just a few. While system dynamics has at times borrowed, or in some cases reinvented, concepts from business policy and strategy, this structural theory development has until recently evolved largely independently of traditional business research and practice. There is, however, a great deal of potential synergy between traditional business research, particularly strategy research, and system dynamics that is increasingly being exploited.

This paper surveys the application of system dynamics to business policy and strategy. The next section discusses the role of system dynamics models in policy and strategy formulation and implementation. Topics include how system dynamics fits into the typical policy and strategy formulation process, how system dynamics offers synergies with more traditional approaches, and how to conduct a modeling effort in order to enhance the implementation of any changes in policy or strategy. In Sect. Theory Development – Understanding the Drivers of Business Dynamics, system dynamics contribution to theory development is discussed – what are the structures underlying common business problems, and how can performance be improved? For example, what creates instability in supply chains, or "boom and bust" in product sales, and how can these behaviors be changed? Finally, in Sect. Applications and Case Examples applications to real world situations are presented – case studies that illustrate the value and impact of using system dynamics for policy and strategy development and implementation in specific firms and in-

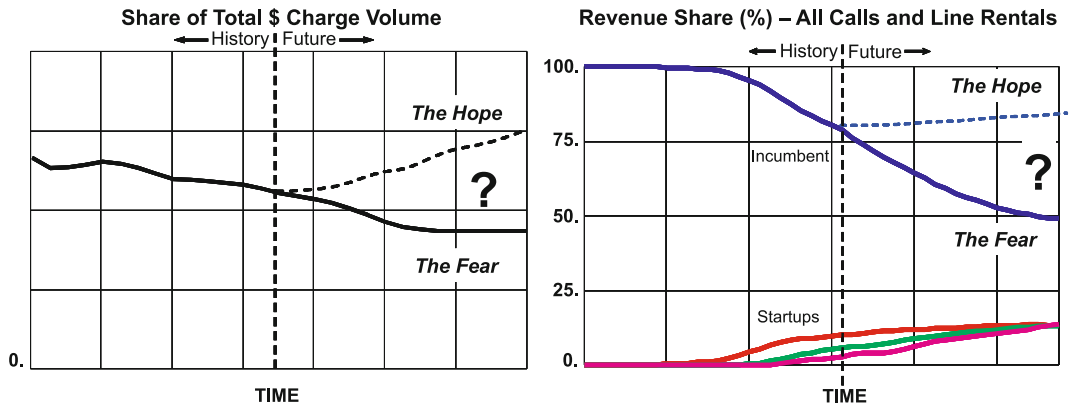

Business Policy and Strategy, System Dynamics Applications to, Figure 1
Examples of problematic behavior over time

dustries. As there has been a substantial amount of work done in this area, I must be selective, trying to touch on major themes and a representative sampling of work. Inevitably this will reflect my personal experiences, and my apologies to others that I have omitted either intentionally or unintentionally.

Using System Dynamics Models in Policy and Strategy Formulation and Implementation

Role of System Dynamics Models in Policy and Strategy Formulation

There is general agreement among system dynamics modelers on the role that models play in the policy and strategy formulation process. This role has been depicted diagrammatically and described by Morecroft [65], Sterman [100], and Dyson et al. [23]. The role of the model in policy and strategy formulation is to act as a "virtual world" – a simpler and transparent version of the "real world". The model serves as a vehicle for testing our understanding of the causes of behavior in the real world, and as a laboratory for experimentation with alternative policies and/or strategies.

One version of that role is shown in Fig. 2. In many cases, the process starts with the definition of a problem – an aspect of behavior that is problematic or threatening. This might be a decline in market share or profitability, or the threat posed by a new competitive product or service, as illustrated in the example of Fig. 1. Sometimes the problem can be expressed in terms of achieving business goals or objectives in the future. As illustrated in Fig. 2, the overall policy/strategy management process can be divided into three components: analysis, planning, and control. "Analysis" is usually triggered by a significant and/or persistent deviation between actual and expected performance. It involves the iterative structuring, testing and refinement of an organization's understanding of its operational or strategic problems and of the options open to it to deal with the performance gap. The model is the vehicle for this analysis – does our understanding of the system as reflected in model equations in fact produce behavior consistent with the observed problem, and if not, how can our understanding of structure be made more consistent with reality? The process of evaluating alternative policies/strategies often sheds new light on the problems faced by an organization or reveals the need for further analyzes. Note that the "modeling" cycle is iterative and compares simulated behavior to actual performance – the scientific method applied to strategy. Simulation of the model shows how business structure, policies, and external events together caused the past performance of the firm, and how future performance will evolve if structure, policies, and external events differ. The next phase, "planning", is also an iterative process; it involves the evaluation, selection, and implementation of policies/strategies – some authors refer to this as "rehearsing" strategy. Evaluation of alternative policies/strategies depends not only on projected accomplishment of organizational goals, but also on the realities of current performance. The existing operational policies or existing strategy (vision, mission, strategic objectives) and goals are subject to refinement, as required, based on the successes and problems encountered, and in response to changing conditions.

A third phase of the policy/strategy formulation process is here called "control". On-going policy/strategy *management* involves the continual, systematic monitoring of performance and the effective feeding back of successes, problems, threats, opportunities, experience, and lessons learned to the other components of the policy/strategy management process. The control phase is

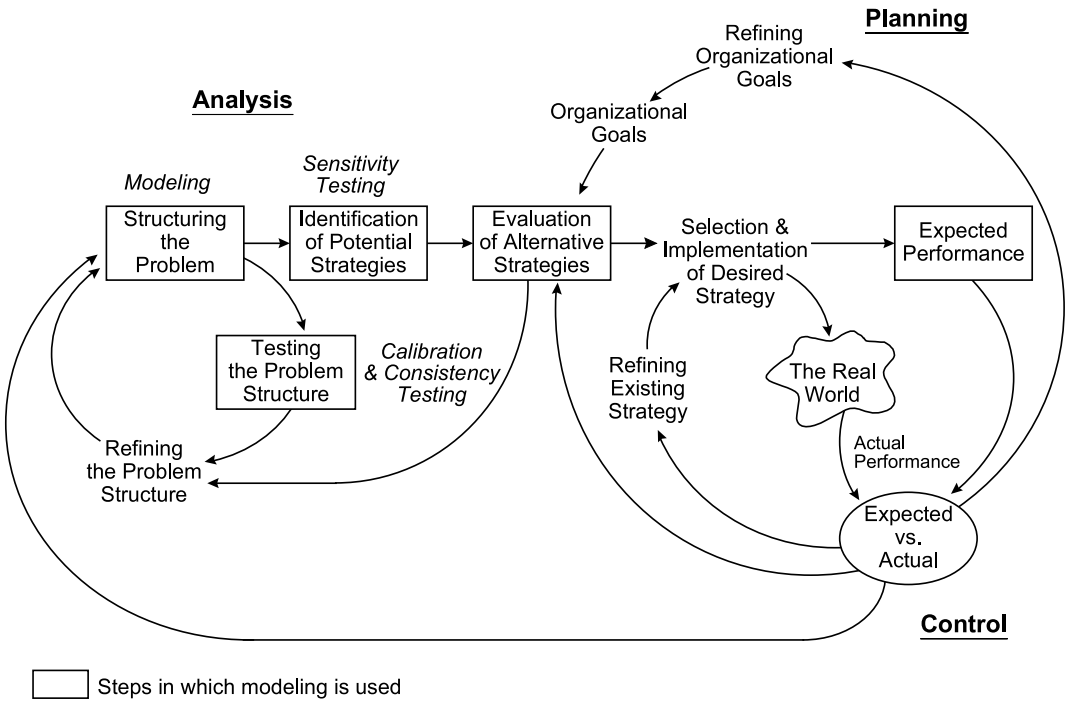

Business Policy and Strategy, System Dynamics Applications to, Figure 2
Policy/strategy management process and the role of system dynamics models [52]

where organizations continue to learn. *The model provides an essential element to the control process – a forecast of expected performance against which actual performance can be monitored on a regular basis*. Deviations provide a signal for additional analysis: Has the policy/strategy been implemented effectively? Have conditions about the external environment changed? Are competitors acting differently than expected? Has the structure of the system changed? The model provides a means of assessing the likely causes of the deviation, and thereby provides an early warning of the need to act.

Synergy Between Traditional Strategy and System Dynamics

While Fig. 2 illustrates how system dynamics models fit into the policy/strategy formulation *process*, there is also a synergy between system dynamics models and traditional *strategy frameworks and concepts*. Figure 3 illustrates the factors that drive business performance from a system dynamics perspective. Starting with resources, a firm's resources determine its product attractiveness; a firm's market share is based on that attractiveness compared to the attractiveness of competitor products; market share drives customer orders, which in turn generates profits and cash flow to finance the acquisition of additional resources for further growth – thereby completing a growth-producing feedback around the outside of the figure (or as in the example of the telecom in Fig. 1, "growth" in the downward direction for the incumbent). However, the acquisition of additional resources can constrain future growth. To the extent increased resources increase costs, then profits and cash flow are reduced, and/or prices may need to increase. Both constrain growth (as might happen for the startups in the telecom example of Fig. 1).

There are a number of places in Fig. 3 where the system dynamics approach can be, and has been, connected to traditional strategy research and practice. For example, concepts such as learning curves, economics of scale, and economies of scope define possible connections between resources and costs – system dynamics models typically represent these connections. Figure 3 shows a number of factors external to the firm: market demand, competitor product attractiveness, technology, and supplier inputs. Strategy frameworks such as "five forces" and visioning approaches such as "scenario-based planning" [112], provide methods for thinking through these inputs – system dynamics models determine the consequences of alternative assumptions for the performance of the firm (note that system dynamics models also often internally represent structural dynamics of competi-

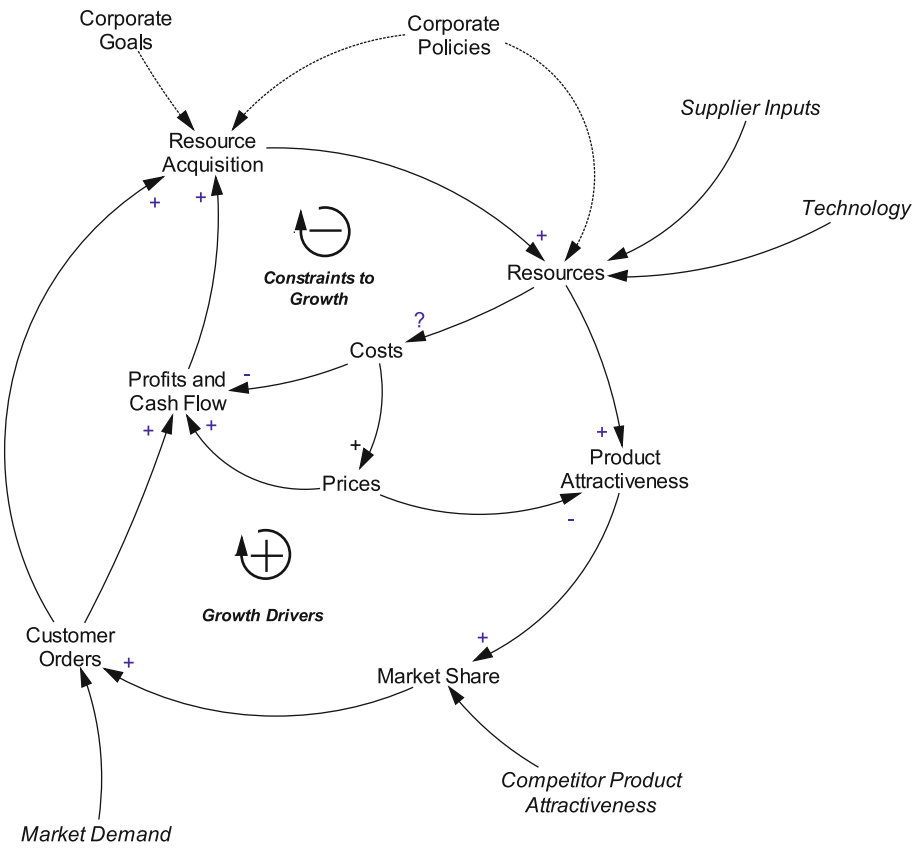

Business Policy and Strategy, System Dynamics Applications to, Figure 3
Drivers of business performance (adapted from [49])

tors, suppliers, and the market as appropriate to explain the behaviors and issues of interest, rather than specifying them as exogenous inputs). For example, Fig. 4 shows a "sector" diagram of the major components of a strategy model developed by a consulting company for a telecom company dealing with loss of market share as in Fig. 1 above (from [35], originally developed by Lyneis). The model not only represents factors internal to the dynamics of the telecom firm, but also factors related to the internal dynamics of competitors, regulatory responses to telecom and competitor performance, financial market responses to telecom performance, and market reactions to telecom and competitor competitive position. This sector diagram connects to a number of the traditional strategy frameworks. In addition to these general connections, a number of system dynamics papers have detailed specific connections to strategy concepts: learning curves and product portfolios [59]; duopoly competition [97]; diversification [31,68]; and industry structure and evolution [93].

Many of the connections between system dynamics and traditional strategy practice and research are discussed in Warren's *Competitive Strategy Dynamics* [105] and *Strategic Management Dynamics* [108]. More importantly, Warren's book details and expands on the connections between the system dynamics concepts of structure, particularly the concepts of stocks and flows, and the well-established resource-based view (RBV) of strategy and performance. (See [63], and [30], for explanations of the development of RBV; a managerial explanation of how this theoretical perspective can be applied can be found in Chap. 5 in [36]. Again, system dynamics provides a means of simulating the consequences of alternative resource acquisition and allocation strategies on firm performance. As such, there would seem to be a strong synergy between system dynamics and this strategy approach.

In summary, while system dynamics and traditional strategy approaches developed largely independently, the potential synergies between the two are significant. Until recently, few researchers and practitioners have made the

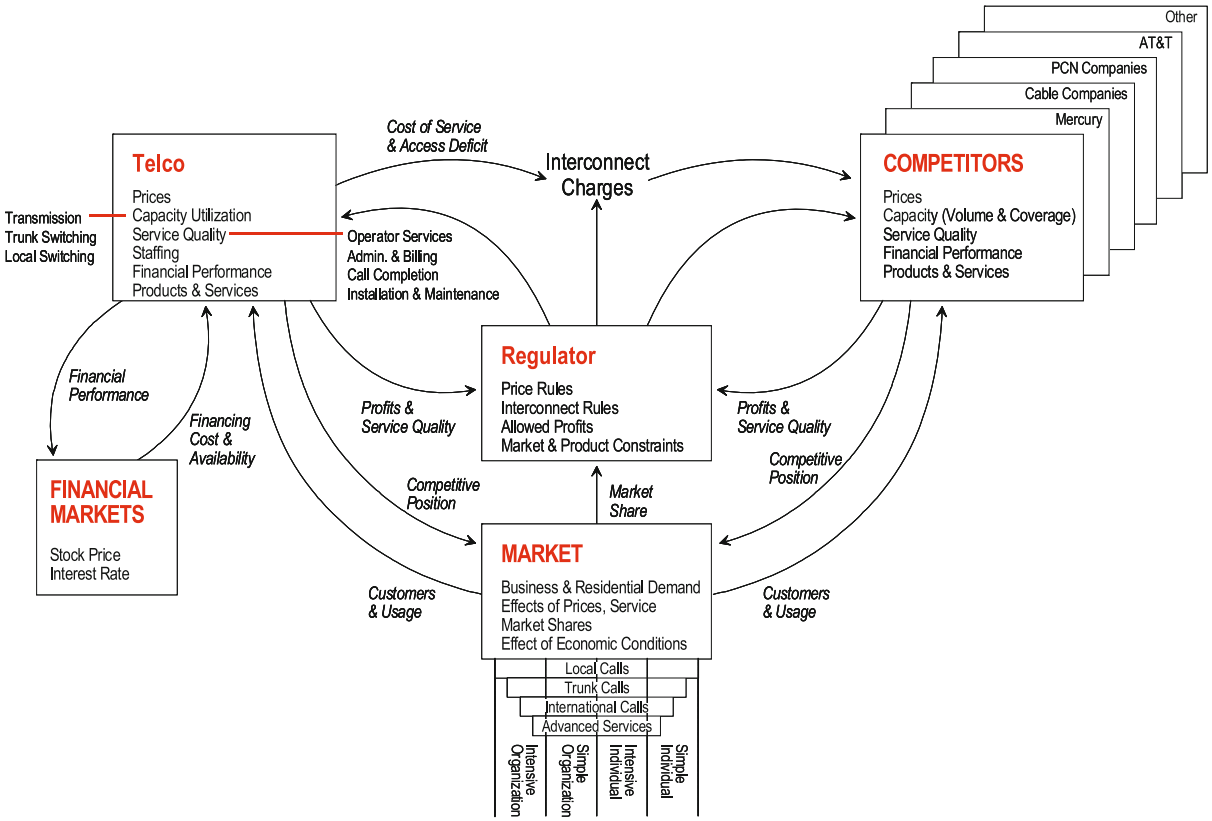

Business Policy and Strategy, System Dynamics Applications to, Figure 4
"Sector" diagram from a typical strategy model

effort to cross disciplines and exploit this synergy. More effort and publication are needed to demonstrate areas of synergy and get system dynamics into the mainstream of business strategy research and ultimately practice.

Working with Management Teams to Achieve Implementation

While Fig. 2 depicts the overall role of the model in strategy management, the approach to developing and using the model itself involves an iterative, multi-phased process. That process has evolved over time as practitioners and researchers have learned from experience. Since its inception, system dynamics has been concerned with having an impact on business decisions. Jay Forrester, the founder of the field, stressed the importance of working on important problems – those that affect the success or failure of firms – and generating solutions that are relevant to those problems. Ed Roberts, one of the first researchers and practitioners in the field, was also involved in early research and experimentation on organizational change and how models can fit into that process [90,92]. The emphasis on having an impact has remained a central tenet of system dynamics, and over the years system dynamics practitioners have developed and refined methods of working with managers to not only solve problems, but also to enhance the likelihood of those solutions being implemented.

Starting with Roberts and his consulting firm Pugh-Roberts Associates (now a part of the PA Consulting Group), a number of researchers and practitioners have contributed to the evolving approach of working with managers to affect organizational change. These include, but are not limited to, Coyle and his group originally at Bradford University in the UK [14,15], Morecroft and his group at the London Business School [65,71], Richmond [88], Sterman [100] and Hines [37] at MIT, the "group model building" approach ([87,104]; ► Group Model Building and Peter Senge's organizational learning [95]. While there are some differences in emphasis and details, there is in general agreement on the high-level process of using system dynamics to affect corporate strategy development and change. In this section, I describe that process, discuss some specific areas where there is some

divergence in practice, and end with some examples to the approach in practice.

In the early years, the approach of most system dynamicists to consulting was heavy on "product" and light on "process". Like management science in general, many in system dynamics took the view that as experts we would solve the client's problem for him, and present him with the solution. Practitioners gradually recognized that elegant solutions did not necessarily lead to implementation, and consulting styles changed to include increased client involvement [90,109]. At the same time, the "product" was evolving to meet the needs of clients (and to take advantage of the increased power of computers). Practitioners evolved from the smaller, policy-based models which characterized the original academic approach to more detailed models, along with the use of numerical time series data to calibrate these models and determine the expected numerical payoff to alternative strategies [52]. In addition, during the 1980s academic research began to focus more on "process:" the use of models in support of business strategy and on more effective ways to involve the client in the actual building of the model [65,66,87,88,104].

System dynamics practitioners now generally agree on a four-phased approach to accomplish these objectives:

1. Structuring the Problem
2. Developing an Initial Model and Generating Insight
3. Refining and Expanding the Model, and Developing a Strategy
4. On-going Strategy Management and Organizational Learning.

In practice, there is sometimes a fifth phase of work. Often modelers simplify a final project model in order to capture the core feedback loops that lie behind observed dynamics. Many of the generic structures discussed later in this paper arose in this way. In other cases, modelers will create management "games" and/or learning labs from the final project model.

As discussed below, the relative mix between "product" (detail complexity and calibration of model) and "process" (degree of involvement of client in model development and use) is perhaps the main difference in the style and approach of different practitioners of system dynamics. For those new to system dynamics, or for those seeking good examples, there is an excellent example of this process, including model development, by Kunc and Morecroft [44].

Phase 1 – Structuring the Problem The purpose of the first phase of analysis is to clearly define the problem of interest (either a past problem behavior or a desired future trajectory), the likely causes of that problem (or desired trajectory), and any constraints that may arise in implementing a solution. It identifies the performance objectives of the organization and possible solutions – all to be rigorously tested in later phases. Similar to more traditional policy and strategy approaches, during this phase the consultant team reviews company documents, the business press, and available company data, and interviews company managers, and possibly customers and competitors. During this phase, the hypothesized drivers of business performance are identified: what compels customers to buy this product, what compels them to buy from one supplier versus another, what drives the internal acquisition and allocation of resources, what major externalities affect the business (e. g., the economy, regulations, etc.), perhaps drawing on frameworks such as "five forces" and SWOT analyzes. More importantly, these drivers are linked in a cause-effect model to form a working hypothesis of the reasons for company behavior. This hypothesis formation builds heavily on the tools and techniques of what is now commonly called "systems thinking":

1. Behavior-over-time graphs (reference modes) – Graphs of problematic behavior over time, often with objectives for future performance highlighted (using actual data where readily available).
2. Causal-loop and mixed causal, stock-flow diagramming as a diagrammatic hypothesis of the causes of problematic behavior.
3. System archetypes, or common generic problem behaviors and structures observed over and over again in different businesses, as a means of identifying structure (see for example [95] and [43]); and
4. Mental simulation – does the hypothesis embodied in the conceptual model seem capable of explaining the observed problem(s)? Mental simulation is also used to identify the possible impact of alternative courses of action.

Note that the exercise to this point, as commonly practiced, is almost entirely qualitative. Warren [105,108] introduces quantitative dimensions even in this phase.

Phase 2 – Developing an Initial Model and Generating Insight The power of system dynamics comes from building and analyzing formal computer models. This is best done in two steps. In the first, a small, insight-based model is developed to understand the dynamics of the business so as to generate insights into the direction of actions needed to improve behavior. The small, insight-based model is also the next logical progression beyond "systems thinking" in the education of the client in the

methods and techniques of system dynamics modeling. In the second quantitative modeling step (Phase 3 below), a more detailed version of the first model is developed, and is often calibrated to historical data. Its purpose is to quantify the actions needed, to assure that the model accurately reflects all relevant knowledge, and to sell others.

Small models (anywhere from 20–100 equations) make it much easier to understand the relationship between structure and behavior: how is it that a particular set of positive feedback loops, negative feedback loops, and stocks and delays interact to create the behavior shown in the simulation output? This can only be determined by experimentation and analysis, which is very difficult with large models. The focus of the first model is on insight generation, communication, and learning, rather than determining a specific shift in strategic direction or investment. These models can help managers improve their intuition (mental models) about the nature of their business, and thereby to better understand the rationale behind more detailed strategies that evolve in later phases of model development.

In summary, Phase 2 delivers:

- A small model which recreates the observed pattern of behavior or hypothesized future behavior (and is roughly right quantitatively);
- Analysis and understanding of the principal causes of that pattern of behavior;
- Ideas of high leverage areas that could improve behavior into the future; and
- Recommendations as to where additional detail will improve the strategy advice, or will make the results of the model more usable and/or easier to accept by others.

Phase 3 – Refining and Expanding the Model, and Developing a Strategy The final phase of model development entails the iterative expansion of the model to include more detail, and often calibration to historical data, as deemed appropriate for the situation. One progressively adds detail and structure, initially to make the process manageable, and then as necessary to correct discrepancies between simulated output and data, or to add policy handles and implementation constraints. Further, model development is likely to continue in the "on-going learning" phase as additional structure and/or detail is required to address new issues that arise. The purpose of this more elaborate modeling phase is to:

1. *Assure that the model contains all of the structure necessary to create the problem behavior*. Conceptual models, and even small, insight-based models, can miss dynamically important elements of structure, often because without data, the reference mode is incomplete or inaccurate (see [52] for examples of this).
2. *Accurately price out the cost-benefit of alternative choices*. Strategic moves often require big investments, and "worse-before-better" solutions. Knowing what is involved, and the magnitude of the risks and payoff, will make sticking with the strategy easier during implementation. Understanding the payoff and risks requires quantifying as accurately as possible the strengths of relationships.
3. *Facilitate strategy development and implementation*. Business operates at a detail level – information is often assembled at this level, and actions must be executed at that level. Therefore, the closer model information needs and results can be made to the normal business lines and planning systems of the company, the easier strategy development and implementation will be. And,
4. *Sell the results to those not on the client's project team*. Few, if any, managers can dictate change – most often, change requires consensus, cooperation, and action by others. The "selling" of results may be required for a number of reasons. If, as in the optimal client situation, consensus among key decision-makers is achieved because they are all a part of the project team, then the only "selling" may be to bring on board those whose cooperation is needed to implement the change. Under less optimal client circumstances where the project is executed by advisors to key decision-makers, or by a support function such as strategy or planning, then selling to decision maker(s) and to other functions will be required.

There are two important elements in this phase of work: adding detail to the model; and possibly calibrating it to historical data. Adding detail to the small, insight-based model usually involves some combination of: (1) disaggregation of products, staff, customers, etc.; (2) adding cause and effect relationships, and feedback loops, often where the more detailed disaggregation requires representing allocations, etc., but also to represent additional feedback effects that may seem secondary to understanding key dynamics, but may come into play under alternative scenarios, or may later help to "prove" the feedbacks were not important; (3) including important external inputs, typically representing the economy, regulatory changes, etc.; and (4) adding detailed financial sectors, which entail numerous equations, with important feedback from profitability and cash flow to ability to invest, employment levels, pricing, and so on. Calibration is the iterative process

of adjusting model parameters, and revising structure, to achieve a better correspondence between simulated output and historical data. Whereas the Phase 2 model primarily relies on our store of knowledge and information about cause-effect structure, the Phase 3 model relies on our store of information about what actually happened over time.

In summary, Phase 3 delivers:

- An internally consistent data base of strategic information;
- A detailed, calibrated model of the business issue;
- A rigorous explanation and assessment of the causes of performance problems;
- Analyzes in support of strategic and/or tactical issues;
- Specific recommendations for actions; and
- Expectations regarding the performance of the business under the new strategy, and the most likely scenario.

Phase 4 – On-going Strategy Management System and Organizational Learning True strategy management ("control") involves the on-going, systematic monitoring of performance and the effective feeding back of successes, problems, threats, opportunities, experience, and lessons learned to the other components of the strategy management process. The control phase is where organizations continue to learn. *The model provides an essential element to the control process – a forecast of expected performance against which actual performance can be monitored on a regular basis*. Deviations provide a signal for additional analysis: Has the strategy been implemented effectively? Have conditions about the external environment changed? Are competitors acting differently than expected? Has the structure of the system changed? The model provides a means of assessing the likely causes of the deviation, and thereby provides an early warning of the need to act. This feedback is only possible with a detailed, calibrated model.

Differences in Emphasis and Style

While there is general agreement among system dynamics practitioners regarding the role of models in the strategy development process and of the basic steps in that process as described above, there are some differences in emphasis and style regarding: (1) the use of causal-loop diagrams (CLDs) vs. stock-flow (SF) diagrams; (2) whether you can stop after Phase 1 (i. e., after the "qualitative" phase of work); (3) is calibration necessary and/or cost effective; and (4) how much model detail is desirable.

CLDs vs. SF There are disagreements within the field about the value of causal loop diagramming (versus stock-flow diagrams). Causal-loop diagrams focus on the feedback loop structure that is believed to generate behavior; stock-flow diagrams also include key stocks and flows, and in the extreme correspond one-to-one with complete model equations. In my view, there is no "right" answer to this debate. The most important point is that in Phase 1 diagramming one is trying to develop a dynamic hypothesis that can explain the problem behavior, and that can form the basis of more detailed diagramming and modeling – whether that dynamic hypothesis is a CLD, a stock-flow diagram with links and loops labeled, or some combination depends in part on:

- Personal style and experience – some people, Jay Forrester perhaps being the best example, seem to always start with the key stocks and flows and work from there; Kim Warren [106] also argues for this approach as an effective means of connecting to the way managers view the problem and to the data;
- The structure of the system – some systems have "obvious" key chains of stocks and flows, and so starting there makes the most sense (for example, the aging chains in the urban dynamics model [29], the rework cycle on projects, and inventory control systems); other systems, without critical chains of stocks and flows, may be easier to address starting with CLDs;
- Whether or not you are doing the model for yourself or with a group – especially if the group is not conversant with the basics of system dynamics, starting with CLDs is easier, and it's also easier to brainstorm with CLDs (which is different than developing a dynamic hypothesis); but again, it's personal preference and nature of system as well. In practice, I have found that CLDs alone, or a mixed stock-flow/causal diagram, are extremely valuable for eliciting ideas in a group setting about the cause-effect structure of the business, and later for explaining the dynamics observed in simulation output. However, one cannot build a model literally from a causal diagram, and either explicit or implicit translation is required.

Qualitative vs. Quantitative Modeling Some practitioners of system dynamics believe that strategic insights can sometimes be obtained after the first phase of work, after the dynamic hypothesis and mental simulation (and note that much of traditional strategy practice relies on such qualitative insights). Coyle [17,18] argues for this; the popularity of "systems thinking" engendered by Senge's work [95] has spawned a number of practitioners that use

only qualitative modeling [52]. Coyle's views generated a strong counter response from [41]. Wolstenholme [110] provides a history and discusses his view on the advantages and disadvantages of each approach. My own view is that while Phase 1 and the systems thinking that is a key part of it are a necessary start, it should not be the end point. Two problems limit its effectiveness in supporting business strategy. First, simple causal diagrams represented by system archetypes, while useful pedagogically, take a very narrow view of the situation (typically, one or two feedback loops). In reality, more factors are likely to affect performance, and it is therefore dangerous to draw policy conclusions from such a limited view of the system. A more complete representation of the problem considers more feedback effects and distinguishes stocks from flows, but introduces the second problem: research has shown that the human mind is incapable of drawing the correct dynamic insights from mental simulations on a system with more than two or three feedback loops [78,98]. In fact, without the rigor and check of a formal simulation model, a complex causal diagram might be used to argue any number of different conclusions. In addition to overcoming these limitations, as discussed below, formal modeling adds significant value to the development and implementation of effective business strategies. Warren (p. 347 in [107]) also stresses need to focus on quantitative behavior to achieve management consensus.

Need for Data/Validation The necessity of obtaining numerical data and calibrating model output to that data is also questioned by some practitioners. While I agree that curve fitting via exogenous variables is not a useful endeavor, proper calibration is an important part of the scientific method that involves systematically comparing simulation output to data, identifying causes of error, and correcting discrepancies by improving first the structure of the model and then its parametrization. In some cases, discrepancies are ignored because they are deemed to be caused by factors irrelevant to the problem of interest, or may be "fixed" by exogenous factors if these are deemed significant by the client and are consistent with the remaining model structure and calibration. As Homer [38,39] argues, the use of historical data and calibration is essential to scientific modeling.

In some cases, organizations lack the data on key factors felt to be essential to the dynamic performance of the business, and by implication essential to sound strategic management of the business. The modeling process can highlight these short-comings and, in the short-term, substitute educated assumptions for this data. In the longer-term, companies can be encouraged to acquire this important data (and substitute it for much of the unimportant information companies generally pore over).

Accurate calibration can greatly enhance confidence in a model. This can be especially important when trying to convince others of the appropriateness of actions a management team is going to take, or to demonstrate to others the need to take action themselves based on the results of the model. Calibration can also be important for other reasons: (1) numerical accuracy is often necessary to evaluate the relative cost and benefits of changes in strategy, or to assess short-term costs before improvements occur; (2) calibration often uncovers errors in the data or other models, especially incomplete or incorrect mental models that form the basis for the dynamic hypothesis (see [52] for examples); and (3) the "control" feedback in the fourth phase of the strategy management process is only possible with a detailed, calibrated model.

Level of Detail and Model Complexity Some practitioners argue that large, complex models should be avoided, for a number of reasons: they can be even more like black boxes; they can be difficult to understand (not only for the non-modelers, but even the modelers); and they are costly to develop. Morecroft argues that a detailed model "loses its agility and becomes less effective as a basis for argument". (p. 227 in [65]) In practice, the first two issues can be avoided and/or minimized by executing the model development in three phases as discussed above. This allows the client to grow slowly with the concepts, and it allows the modeling team to develop a solid understanding of the model. The third problem is generally not an issue if you are working on significant problems – in my view the cost of the consulting engagement is trivial relative to the expected payoff. While I believe that the client obtains value, regardless of when you stop, strategy consulting is one case where the "80/20 rule" does not apply – the client does not get 80% of the value for 20% of the cost (which would be essentially at the end of Phase 1). In part this is a function of what I view as the objective of the project – providing tools, strategic analyzes, and advice in support of an important strategic and/or investment decision. In this situation, the "value" is back-end loaded. Finally, effective strategy management is only possible with a detailed, calibrated model.

In addition, detail and calibration are often necessary to sell the model to others. In many situations, everyone who may have an input to a strategic decision or be necessary for successful implementation cannot be a part of the client team. As surprising as it may seem, the selling of re-

sults (as opposed to understanding) is easier to accomplish with a detailed, calibrated model than with a small model. First, the numerical accuracy gives the model face validity. Second, a detailed model more often allows the modeler to counter the "have you considered (insert pet theory)?" criticism. I have often found that when you start explaining the model to others, they respond by asking "Have you considered this feedback? Or this effect?" And if you have not, that ends the discussion. Even though you may think that feedback or that effect may not have any impact, if it is not included in the model you cannot say "Yes, we looked at that and it did not have any impact", and explain why. If it is not in the model the critic can argue that your results would be changed by the inclusion of their pet theory. One has a hard time countering that assertion without a convincing argument based on simulation results. Finally, a detailed, calibrated model helps tell a convincing story. The simulation output, which corresponds closely to the data, can be used to explain (again with output) why, for example, a loss of market share occurred. How price relative to the competitions' price was the key factor, and/or how the factors affecting share changed over time. The simulation output can and should be tied to specific events. We have found that an explanation like this is compelling, and is important in enhancing the credibility of the model and the modeler.

The benefits of large, complex models in a consulting setting are also noted by Winch [111]. He specifically finds that "For the executive team to have confidence in the impartiality of the model, each person must feel it captures the detailed pressures and processes of his or her own sphere of responsibility yet produces a holistic view of the organization". (pp. 295–6 in [111]), and that the model was essential to getting everyone to agree: "The process of building system dynamics models, in each case ostensibly as a forecasting and evaluation tool, enabled the managers eventually to develop a shared view, which formed the basis for formulating and agreeing upon a final strategy". (p. 298).

Process Examples

There are a number of published examples that support the four-phase process of applying system dynamics to business strategy:

- Lyneis [52] provides not only a more fully developed description of the detailed, calibrated-model Pugh–Roberts approach, but also illustrates its application to the credit card and airline manufacturing industries.
- Morecroft et al. [70] describe how a model was created and used to stimulate debate and discussion about growth management in a biotechnology startup firm. The paper highlights several novel features about the *process* used for capturing management team knowledge. A heavy emphasis was placed on mapping the operating structure of the factory and distribution channels. Qualitative modeling methods (structural diagrams, descriptive variable names, "friendly" algebra) were used to capture the management team's descriptions of the business. Simulation scenarios were crafted to stimulated debate about strategic issues such as capacity allocation, capacity expansion, customer recruitment, customer retention, and market growth, and to engage the management team in using the computer to design strategic scenarios. The article concludes with comments on the impact of the project.
- Winch [111] examines the role that building and using a system dynamics model plays in developing consensus within management teams facing key strategic decisions: A shared view emerges within the team as individual views of the company, its industry, and the socioeconomic climate are articulated and compared. Examples are given based on two actual consulting assignments in which differing views concerning the competitive environment and the general business outlook initially pointed to quite different strategies. The emergence of consensus was considered a major benefit in addition to the forecasts and quantitative evaluations the model provided. In its analysis and examples, this article emphasizes both the "hard" benefits of forecasts and an objective framework for quantitative evaluations and the "soft" benefits of building consensus within management teams.
- Coyle [15,16] also has an approach that he discusses, with emphasis on CLDs (he terms these "influence diagrams", and his group was instrumental in initial use of this technique).
- Snabe and Grossler [94] show how modeling can be supportive for strategy implementation in organizations and illustrate with a detailed case study from a high-tech company.
- A special issue of the Journal of the Operational Research Society on System Dynamics for Policy, Strategy, and Management, edited by Coyle and Morecroft [19], contains a number of papers which in part discuss consulting process issues [21,110,112] among others).
- The special issue Fall 2001 of *System Dynamics Review* on consulting practice contains papers by Thompson [102], Campbell [11], and Backus et al. [5] that focus on the consulting process.

Theory Development – Understanding the Drivers of Business Dynamics

Another important contribution of system dynamics to business policy and strategy formulation is the development of structural theories to explain commonly observed patterns of behavior. Theory development provides us with: an understanding of the basic drivers of business dynamics; insights, enhanced mental models, and policy guidelines for improved performance; and building blocks of tested model equations for real applications (equations for the model must be provided for models to add to our base of theory). System dynamicists have developed structural theories to explain the basic patterns of business dynamics: (1) cycles and instability; (2) productivity and eroding performance; (3) life cycles; and (4) growth. Each is discussed in turn below.

Cycles and Instability: Stock Management, Supply Chains, and Manufacturing Systems

The very first applications of system dynamics were to understanding the tendencies of production-distribution systems, or "supply chains", toward cycles and instability; these applications remain important to this day [25,26]. For example, the "Beer Game", now distributed by the System Dynamics Society, was developed and refined at MIT beginning in the early 1960s and remains one of the most popular introductions to both system dynamics principles, and to supply chain issues.

Supply chains are an important component of all industrialized societies. They exist in any industry where goods are produced and distributed to consumers, for example, food and beverage production and distribution, or manufactured goods such as automobiles and appliances. Supply chains exhibit a classic behavior pattern which has impacts not only on the individual company, but also for the economy as a whole: as one moves up the supply chain from the end user, any variation in orders from the end user are progressively *amplified* and delayed at each additional stage in the chain – factory variations are greater than customer variations; raw materials production variations are greater than factory variations (see Chap. 17 and 20 in [100] for real world examples of this behavior). This behavior is also sometimes referred to as the "bull-whip" effect.

Figure 5 illustrates the structure of one stage of a typical supply chain and the causes of amplification: a stock of inventory is depleted by shipments (here assumed equal to demand) and replenished by production completions (or more generally, shipments from a supplier); the stock of goods in production (or goods being assembled and shipped by a supplier) are increased by production and reduced, after the production (and/or shipping) delay, by production completions. This structure has a tendency to "amplify" any changes in demand – that is, "production" (or orders and reorders, depending on the system) increase or decrease more than any increase or decrease in demand, and tend to lag changes in demand. For example, in Fig. 5, when demand increases, even if production increases immediately inventory falls because production completions are delayed by the production (and/or shipping) delay. Therefore, production must increase higher than demand (amplification) in order to rebuild inventories. In addition to production and shipping delays, inventory might also fall because of delays caused by smoothing information about demand (such that production changes lag changes in demand). Production further increases above demand because of the need to increase inventories and production or supply lines to higher target levels. Intuitively, and verified by simulations, amplification is greater if: desired inventories are larger; production/transit delays are longer; and/or responses to inventory gaps are more aggressive (smaller adjustment time constant, as discussed below).

This basic structure in Fig. 5 also illustrates the "stock management" problem. In Fig. 5, the stock of finished goods inventory must be managed in order to serve customer demand in a timely fashion. Figure 6 details the structure typically used to control stocks, one of the most used and important structures in system dynamics:

$$\text{Production} = \text{Expected Demand} + \text{Inventory Correction} + \text{Goods In Process Correction}$$

$$\text{Inventory Correction} = (\text{Desired Inventory} - \text{Inventory})/\text{Time to Correct}$$

$$\text{Inventory Desired Inventory} = \text{Expected Demand} \times \text{Months Coverage Goal}$$

$$\text{Goods In Process Correction} = (\text{Desired Goods In Production} - \text{Goods In Production})/\text{Time to Correct Inventory}$$

$$\text{Desired Goods In Production} = \text{Expected Demand} \times \text{Production Time}$$

Stock management is complicated by delays in replenishing the stock, here a production delay. Depending on the pattern of demand, there is often a tradeoff between amplification and variations in inventory – less aggressive responses (longer time to correct inventory) generally reduce amplification but cause greater variations in inventory (and therefore may necessitate higher target levels to

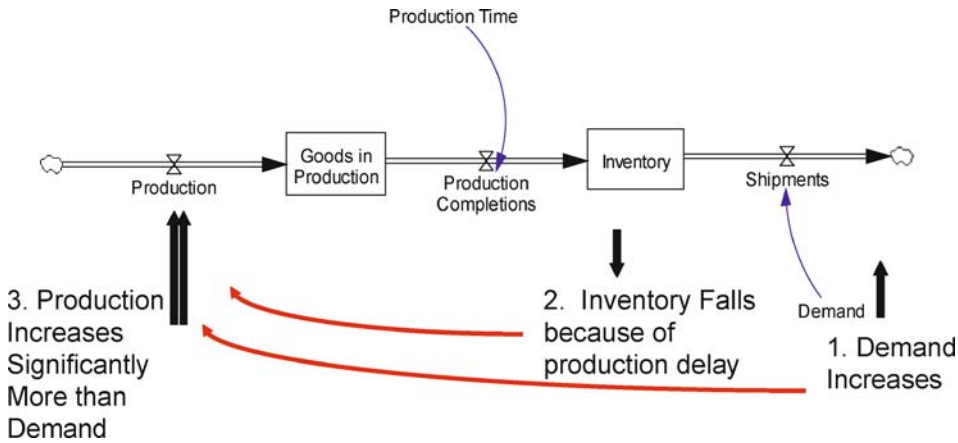

Business Policy and Strategy, System Dynamics Applications to, Figure 5
Structure and causes of amplification in one stage of a supply chain

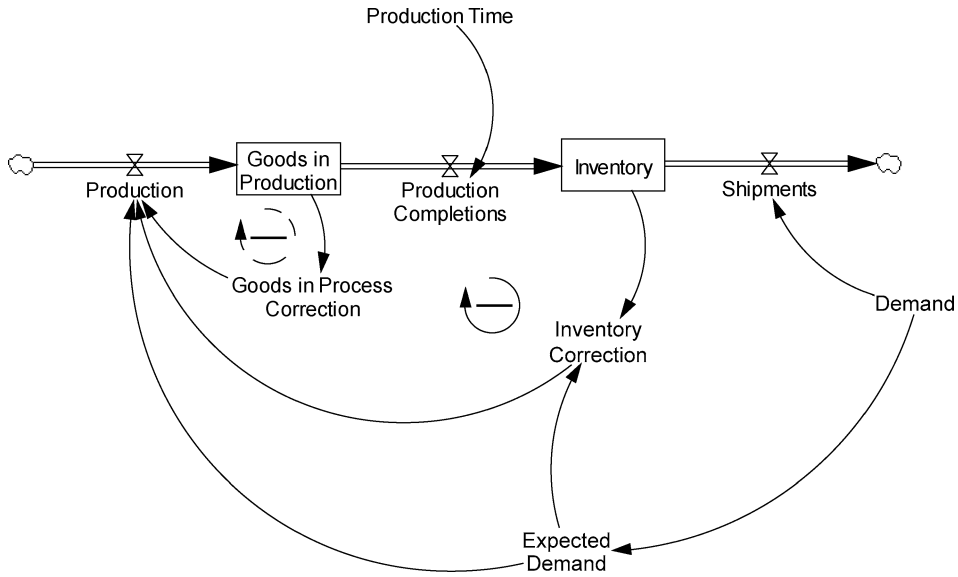

Business Policy and Strategy, System Dynamics Applications to, Figure 6
Stock management structure

reduce the likelihood of stockouts); more aggressive responses (shorter time to correct inventory) increase amplification and demands on manufacturing and suppliers, and potentially costs, but can result in more stable inventory levels. However, under some demand patterns and production conditions, aggressive responses can increase both amplification and inventory instability. While as noted below structural changes can significantly improve the overall performance of stock management and supply chain systems, nevertheless this fundamental tradeoff between amplification and inventory levels will remain. The "optimal" solution will vary by firm, and over time as the inherent pattern of demand changes. These dynamics and tradeoffs are discussed in depth in [49,100].

In a typical supply chain, there are multiple stages connected in series, for example, in the automotive industry: dealers, car manufacturers/assemblers, machine tool producers, parts manufacturers, raw material suppliers (with potentially several stock management stages in some of these main categories). The upstream stages suffer greater amplification than the downstream stages. In the main, this occurs because each stage uses as its demand signal the orders from the prior stage, which are amplified by that stage's stock management policies as dis-

cussed above. Other reasons for increased upstream amplification include [2]: (1) in determining "expected demand", each stage tends to extrapolate trends in orders from the prior stage; (2) order batching; (3) price fluctuations (in response to inventory levels); and (4) shortage gaming (ordering more than you really need to get a higher share of the rationed goods from the supplier; see [52] for an example in the aircraft industry). System dynamics analyzes have identified a number of structural changes which can improve supply chain and stock management performance: (1) reduce delays; (2) reduce inventory; and (3) share information (for example, if upstream stages are aware of the downstream end user customer demand pattern, they can use that information rather than the amplified orders from the prior stage as the basis for their decisions, and at least partially avoid amplification [20].

Applications of system dynamics to supply chain management, and production management, remain an important area of research and applications. Akkermans and Daellaert [1], in an article entitled "The Rediscovery of Industrial Dynamics: The Contribution of System Dynamics to Supply Chain Management in a Dynamic and Fragmented World", provide an excellent survey of supply chain management and system dynamics potential role in moving that field forward. Additional work in this area includes Morecroft's original analysis of the dynamics created by MRP systems ([64]); Gonçalves doctoral dissertation [32], some of which is summarized in [33,34]; Anderson and Fine [3] on capital equipment supply cycles, and Zahn et al. [114] on flexible assembly systems. Each of these discusses variations on the basic stock/production/supply chain management systems, and provides references for further research.

In addition to inventories, firms need to manage other stocks and resources, including raw materials, employees, capital equipment, and so on; the stock management structure described above for inventory applies to these other stocks as well. The management of stocks and resources is central to dynamic and strategy problems in many industries. First, the management of one stock often influences the ability to manage other stocks (for example, capital equipment and employees determine production). Not only does this interdependency create constraints, the additional negative feedback control in managing resources is another source of cyclical behavior (see Chap. 19 in [100], and Chap. 5 of [69]). Second, in addition to the stocks of resources, production is affected by the productivity of those resources. Dynamic drivers of productivity, such as experience and fatigue, are discussed in the next section.

While the negative control feedbacks described above are central to the observed cyclical behavior of supply chains and resource-based firms, an additional negative feedback through the market adds a further source of instability. This dynamic is perhaps clearest in commodity-based industries, which have also been extensively modeled by system dynamicists as first summarized by Meadows [58]. As illustrated in Fig. 7, these models integrate the supply chain with the dynamics created by supply and demand – in a typical commodity system, there are three major negative feedback loops: two supply feedbacks (one through production, often representing the resource labor, and one through the resource capacity), and one demand feedback (for example, an increase in inventory causes prices to fall, which increases demand and leads to a decrease in inventory from what it otherwise would be). Commodity industries typically exhibit behaviors that include cycles of two periodicities, one determined primarily by the shorter production feedback loop and another longer cycle driven by the capacity loop (see Chap. 20 in [100] for both detailed equations and for examples of the structure applied to the livestock and paper industries). The demand feedback loop, however, can play a role in the dynamics as well – if the demand feedback is strong and with a short delay, then demand corrections occur before the supply feedbacks operate and system stability is improved; however, if the magnitude of the delay in the demand loop is similar to the magnitude of the delays in either of the supply loops, the intensity of the corresponding cycle is increased as two negative feedback loops are both independently acting to "solve" the inventory problem. In addition, commodity industries, where they involve a depletable resource such as oil, can experience long-term resource depletion dynamics [101].

In conclusion, manufacturing and supply chain dynamics are central to many of the behaviors observed in businesses (see Chap. 20 in [100] for real world examples of these cycles). The supply chain, stock management, resource management, and commodity structures discussed above are therefore important components of many system dynamics models developed to support business policy and strategy. In some cases, the firm can change policies to reduce the severity of these cycles; in other cases, especially where the cycles are driven primarily by industry dynamics, the individual firm can use the enhanced understanding and forecasting of these cycles for more strategic decisions such as new product introduction and capacity planning (as in the commercial aircraft market case illustrated in Fig. 1 and discussed in Sect. Applications and Case Examples).

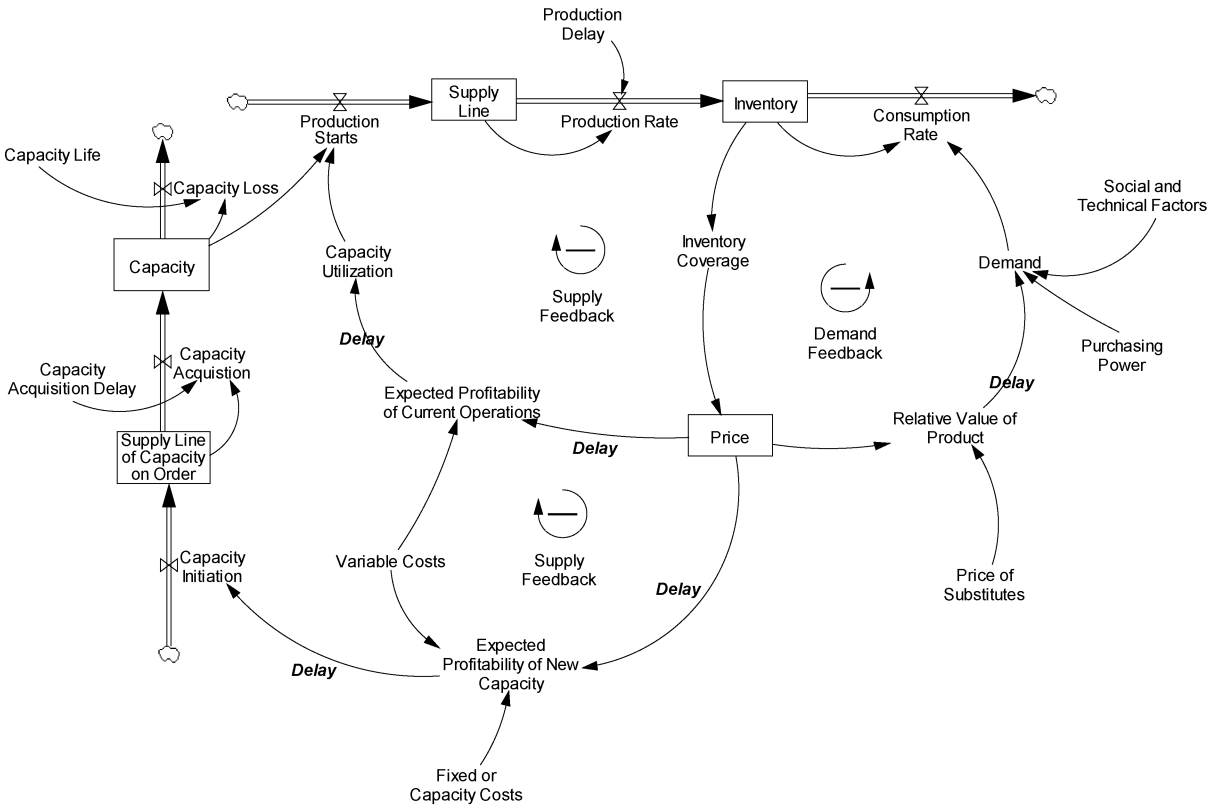

Business Policy and Strategy, System Dynamics Applications to, Figure 7
Commodity industry dynamics showing three controlling feedbacks (adopted from [100])

Productivity and Eroding Performance: Service Industry Dynamics

Service-based firms (e. g. professional services, transportation, catalog and online shopping, etc.), and the service arms of manufacturing-based organizations, have a somewhat different set of structural dynamics. Firms in these industries have a number of characteristics that make them more difficult to manage than more manufacturing intensive industries: (1) their product is difficult if not impossible to inventory, and so something else must buffer changes in demand; (2) they are particularly dependent on the performance of people (although the productivity of resources is also important to manufacturing-based businesses as well); and (3) the performance of the system can be harder to detect, and so they are much more subject to a gradual erosion in performance and goals. "The major recurring problems observed in service industry – erosion of service quality, high turnover, and low profitability – can be explained by the organization's response to changes in work pressure." (see p. 28 of [75]).

One of the primary distinguishing features of service-based firms is that their end product is people-dependent and cannot be inventoried. While there may be inventories of products that support delivery of the service, that delivery must be performed based on current resources. As a result, work or order backlogs are the stock that buffers demand from "production". A simplified example is shown in Fig. 8. Customer orders (demand) fill an order backlog, which is depleted by order fulfillment. Order fulfillment is based on the firm's service "capacity" and the amount of time spent per order – for the same capacity, order fulfillment will be greater if less time is spent per order (although service quality may suffer). Capacity is dependent upon people, overtime, and productivity, as discussed further below. Employees are increased or decreased based on desired capacity, which in turn depends on order backlog relative to the firm's service standard (time per order). Work pressure depends on desired capacity relative to the current stock of employees – to the extent the number of employees does not increase as needed, work pressure builds which can result in increases in effective capacity

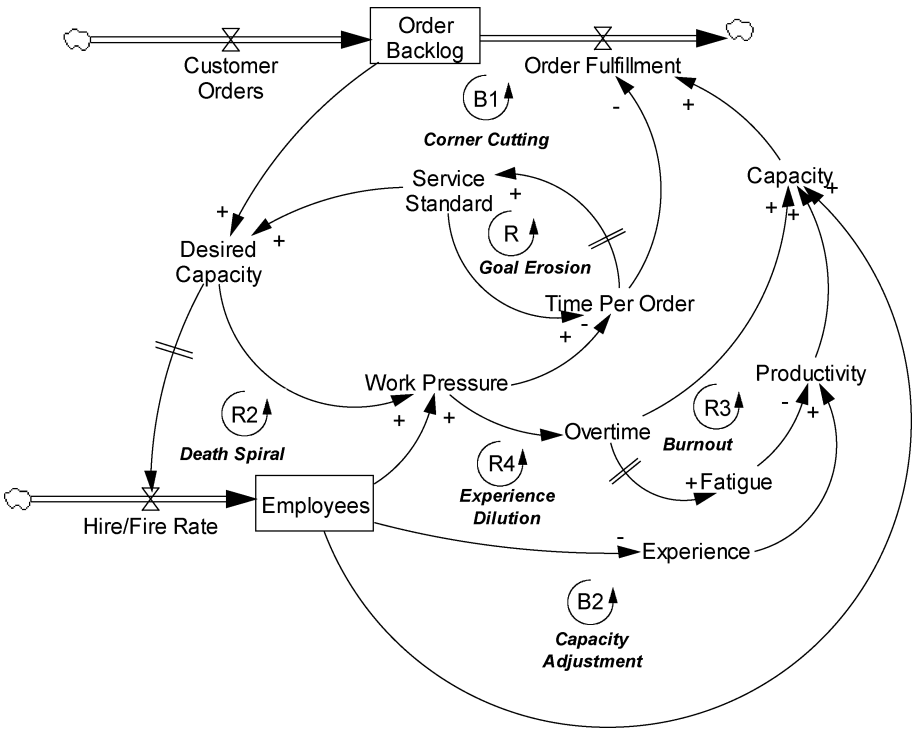

Business Policy and Strategy, System Dynamics Applications to, Figure 8
Drivers of service business dynamics

via overtime, or reductions in time spent per order. Time spent per order depends on the firm's service standard for time spent per order, modified by work pressure – if work pressure is high, time spent can be reduced. The service standard often responds to actual performance.

Another important characteristic of service supply firms shown in Fig. 8 is that their capacity is particularly dependent on the performance of people, both in numbers and in productivity. While productivity is also a factor in manufacturing systems, the sensitivity of performance to people factors is generally less than in service-based firms. Therefore, models of such firms generally represent in some detail the factors that drive productivity, including: (1) skill and experience, often using an aging chain or "rookie-pro" structure [100,113]; (2) fatigue from sustained amounts of overtime; and (3) work intensity increasing productivity, but with "haste-makes-waste" impacts on errors and rework (not shown in Fig. 8). In these situations, when demand is growing there are considerable short-term forces which reduce productivity and cause a deterioration in service quality: adding people reduces productivity because of "experience dilution"; working overtime increases fatigue and reduces productivity; pressures to work more intensely increase errors and cause additional work. As shown in Fig. 8, these productivity effects form reinforcing feedback loops which can drive down a service system's performance: an increase in work backlog and desired capacity causes the firm to hire more people; experience levels and productivity decline as a result, thereby reducing order fulfillment below what it otherwise would have been; order backlog does not fall as much as expected, necessitating additional capacity, further hires, and decreased experience; this completes the "experience dilution" R4 loop. The "burnout" loop through overtime and fatigue is similarly a reinforcing loop (R3).

Oliva [75] shows that how management responds to these work pressure problems can determine the long-term success or failure of a service-based organization, largely as a result of the third characteristic of such systems: performance of the system can be harder to detect, and so they are much more subject to a gradual erosion in performance and goals. Oliva demonstrates that if the firm reduces its service standards (goals) in response to deteriorating performance (loop R1 goal erosion), a death spiral can ensue in which the declining goals cause the firm to add fewer people, which locks in a situation of excessive work pressure and further declining performance (thereby

completing the "death spiral" loop R2). Unless there is a cyclical downturn in demand which alleviates the pressure, a firm's service performance will gradually erode until competition captures the market. He further discusses solutions to these problems, including buffers and faster response. Oliva's work, together with applications noted below, suggest that a service company should hire steadily rather than in spurts to avoid problems of inexperience, should hire enough workers to avoid overwork and a drift to low standards, and (in the case of equipment service) should give preventive maintenance high priority to avoid a spiral of equipment failures.

The resultant financial pressures engendered by the dynamics described above often drive service organizations to investments in process improvement and other cost containment initiatives to seek efficiency gains. Such investments, while offering perhaps the only long-term solution to remaining competitive, cause short-term workloads that further increase the demands on service personnel. This is demonstrated in the work of Repenning and Kaufman [83], and Repenning and Sterman [84,85].

Akkermans and Vos [2], and Anderson et al. [4] have studied the extent to which service industries have multistage supply chains similar to manufacturing industries, albeit with backlogs rather than inventories. Akkermans and Vos demonstrate that "inventory" cycles in service chains manifest themselves in terms of order backlog and workload cycles, and that while some of the causes of amplification existent in product supply chains apply to service supply chains (demand signaling and pricing), others, particularly those related to inventory management, do not (order batching, shortage gaming). They find that the real drivers of amplification in service supply chains come from the interactions of workloads, process quality, and rework. Because of delays in hiring and firing, capacity is slow to respond to changes, and is likely to exacerbate cycles. Anderson et al. [4] find that the bullwhip effect may or may not occur in service supply chains, depending on the policies used to manage each stage. However, when it does occur, they find that the systemic improvements that can often be achieved in physical supply chains by locally applied policies (e. g., reducing delay times and sharing information) do not have as many parallels in service chains. Instead service supply chains are characterized by numerous tradeoffs between improving local performance and improving system performance.

The modeling of service delivery has also had a long history in system dynamics, though the number of published works is more modest than in other areas. Much of the early work was in the area of health care and education [46], Later works of note include models of People Express Airlines [98], Hanover Insurance claims processing [96]. NatWest Bank lending [74], and DuPont chemical plant equipment maintenance [12]. Homer [40] presents a case application for a major producer of equipment for semiconductor manufacturing that demonstrates many of the structures and policy issues enumerated above. These works incorporate the basic dynamic theory discussed above and illustrated in Fig. 8, and add another set of important structural theories to the building blocks for business strategy applications (note that the modeling of various effects on productivity is much more extensive in the area of project modeling, as discussed in [55]).

Life Cycles of Products and Diffusion

Another important pattern of behavior characteristic of many firms (or subsets of firms) is that of a life cycle (for the flow) and S-shaped pattern for the stock, as illustrated in Fig. 9: a gradual increase from a low level up to a peak, followed by a gradual decline either to zero or to some replacement level (sometimes referred to as "boom and bust" behavior). Sterman [100] and Oliva et al. [76] provide some real world examples of this behavior. The example shown is common for the sales of new products: the flow represents people becoming customers, and the stock, customers.

The structure which creates this "boom and bust" dynamics is shown in Fig. 10. In the marketing literature this structure is referred to as the "Bass Diffusion Model" after its original proponent [8]. The structure consists of three feedback loops: a reinforcing "word of mouth" loop that dominates behavior in the first half of customer sales growth; a balancing "market saturation" loop that constrains and eventually shuts down growth as the number of potential customers falls to zero; and another balancing loop "advertising saturation", which represents other means of stimulating awareness, such as advertising, direct sales efforts, and media reports. These other channels are usually assumed to be proportional to the size of the pool of potential customers, and therefore initially stimulate the flow of "becoming customers" but then decline over time as the pool is depleted.

The dynamics of this structure, extensions to it (for example, loss of customers, competition, repeat sales), and policy implications are discussed in depth in Chap. 9 in [100] and Chap. 6 in [69]. This structure forms the basis of many system dynamics models that represent product sales, customer development, and the diffusion of innovations. Published examples include the work of Milling [60,61] and Maier [57] on the management of innovation diffusions and Oliva et al. [76] on boom and bust

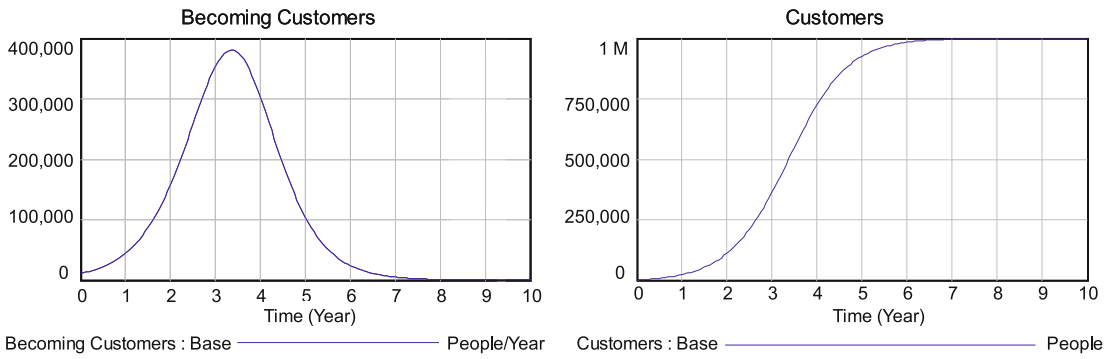

Business Policy and Strategy, System Dynamics Applications to, Figure 9
Life cycle behavior mode ("Boom and Bust")

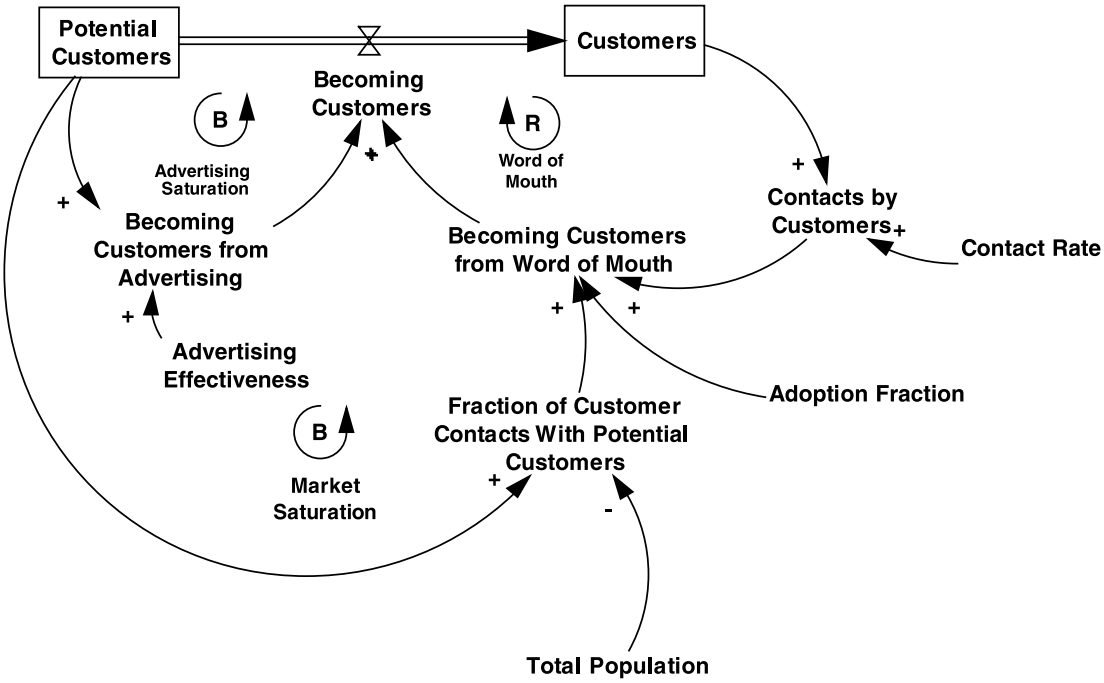

Business Policy and Strategy, System Dynamics Applications to, Figure 10
Basic structure generating boom and bust dynamics

in e-commerce. Milling discusses the ► Diffusion of Innovations, System Dynamics Analysis of the in more depth.

Growth Dynamics

Growth is fundamentally a dynamic process, and therefore it is no surprise that since its early days system dynamicists have shown an interest in the dynamics of corporate growth. Forrester [27,28], Packer [77], Lyneis [48,49], Morecroft [67], Morecroft and Lane [70] and others studied corporate growth in the field's early years. More recently, the People Express [98] and B&B ("Boom and Bust") flight simulators [78] illustrate the field's interest in growth dynamics.

In his 1964 article, Forrester identified the range of possible growth patterns (see Fig. 11): smooth, steady growth; growth with repeated setbacks; stagnation; and decline. Examples of these patterns can be found in many real world industries, as illustrated for the computer industry in Fig. 11 (see also [50] for examples). In his classic article "Market Growth as Influenced by Capital Investment", Forrester detailed the three types of feedback loops which can create the range of possible growth patterns. These are illustrated in Fig. 11 (the equations for

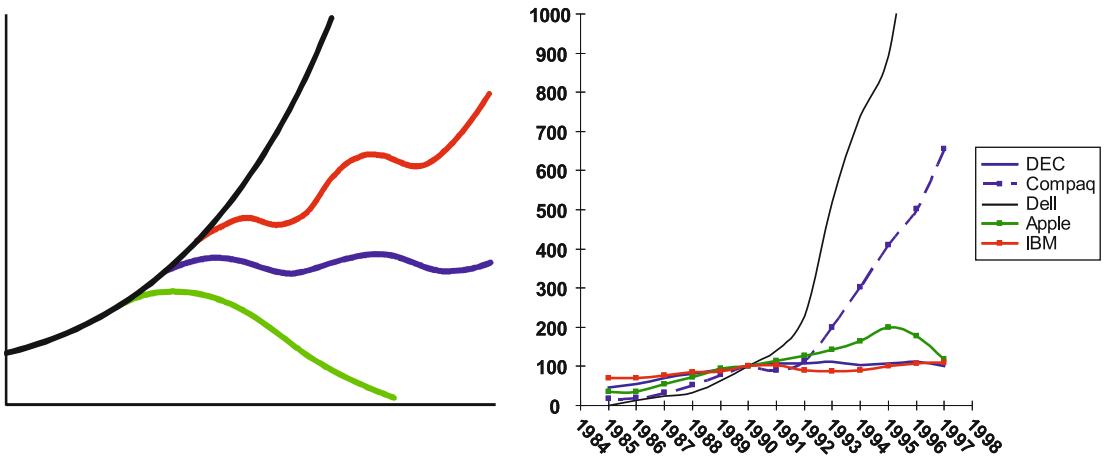

Business Policy and Strategy, System Dynamics Applications to, Figure 11
Stylized patterns of growth and examples from the computer hardware industry

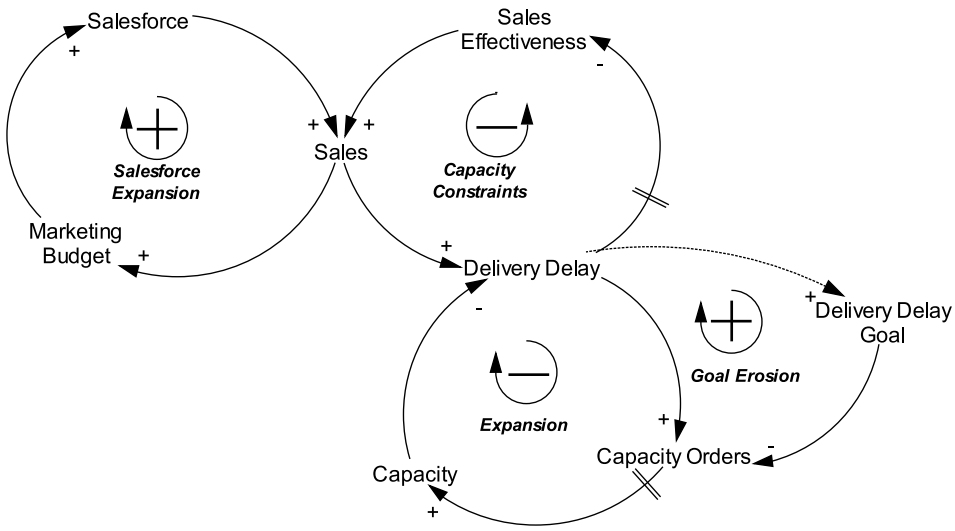

Business Policy and Strategy, System Dynamics Applications to, Figure 12
Feedback loops creating observed patterns of growth (adapted from Forrester 1968)

this model are provided in the original Forrester article; the model is also presented and discussed and analyzed in detail in Chap. 15 in [100], and Chap. 7 in [69]).

On the left in Fig. 12 is the reinforcing "salesforce expansion" loop: the salesforce generates sales, a portion of those sales are allocated to future marketing budgets, which allows an increase in the size of the salesforce and a further increase in sales. The salesforce expansion loop in isolation can create smooth growth forever (until the market is saturated). However, assuming a fixed capacity, the balancing "capacity constraints" loop activates: if sales exceed capacity, delivery delay increases such that, after a delay, sales effectiveness falls and sales decline. The goal of the loop is to equate sales and capacity, and the two loops together can produce growth followed by stagnation (with fluctuations caused by delays in the balancing loop). In response to increasing delivery delay, however, firms often increase capacity ("capacity expansion" loop): when delivery delay exceeds the firm's delivery delay goal, capacity orders increase, which after a delay increases capacity and thereby reduces delivery delay; the goal of this loop is to equate delivery delay to the firm's delivery delay goal. However, once delivery delay is reduced, sales effectiveness and sales increase, thereby stimulating additional salesforce expansion, such that the growth with setbacks pattern of behavior can result. The final loop shown in

Fig. 12 is the reinforcing "goal erosion" loop: if the firm's delivery delay goal responds to the actual delivery delay performance, a downward spiral can ensue – the goal increases, less expansion occurs than had before, capacity is less than needed, delivery delay increases, the goal is further increased, and so on (this loop is similar to the service standard goal erosion loop discussed above). The goal erosion loop can create the decline dynamics illustrated in Fig. 11 (although in actual practice the decline would likely occur over a much more extended period than shown).

In actual practice, there are numerous positive feedback loops through resources that might stimulate growth. These loops are listed below, and many are discussed and diagrammed in Chap. 10 in [100]. In each of these loops, an increase in sales causes management actions and/or investments that further increase the resource and sales:

- Sales channels – sales capability (which might include salesforce as discussed above, or retail stores), advertising, word-of-mouth contagion (as in the diffusion model), media hype (sales create media exposure which attracts potential customers and more sales)
- Price – Product attractiveness channels (operationalizing the link between resources and costs in Fig. 2 – spreading of fixed costs over more units, thereby lowering unit costs; learning curves; economies of scale; economies of scope; investments in process improvements)
- Market channels which increase the pool of potential customers – network effects (the more people using cell phones the greater their value), development of complementary goods (software applications for computers)
- Product investment channels – product improvement, new products
- Market power channels – over suppliers, over labor, over customers, cost of capital.

With all these positive feedback loops, how can anyone fail? In fact, there are also numerous constraints to growth, including: depletion of the pool of potential customers as discussed in the last section; growth of competition; delays in acquiring production capacity and/or service capacity; limits to financial capital (which can increase delays or limit acquiring productive assets); and increases in organizational size, complexity and administrative overheads (which might make the resources – costs loop in Fig. 2 revert to a positive connection, thereby constraining growth). Structures for representing these constraints are provided in the earlier references to this section, especially [49,69,100], and form the building blocks for many of the practical applications of system dynamics to business growth strategy.

As system dynamicists have long recognized, managing growth is one of the more challenging management tasks. It entails fostering the positive, reinforcing feedback loops while simultaneously relaxing the constraining, negative feedback loops. While it is difficult to generalize without sounding platitudinous, a number of important lessons have emerged from studies of growth. Lyneis [50] discusses these in more detail, and references other work:

Lesson 1 You won't achieve what you don't try for (if a firm is timid in its growth objectives, it will be timid in the acquisition of resources – balancing loops through for example delivery delay will then drive sales growth to the firm's resource growth). Corollary 1: Don't mistake forecasts for reality (a firm may be continually surprised by how accurate their sales forecasts are, because if resources are based on these forecasts, the balancing loops will drive sales to those resources). Corollary 2: Provide sufficient buffers and contingencies (these help minimize the risks that the balancing loops will become dominant).

Lesson 2 Avoid the temptation to reduce objectives in the face of performance problems (the "goal erosion" loop in Figs. 8 and 12).

Lesson 3 In a world of limited resources, something must limit growth. Proactively managing these limits is a key factor affecting performance. For example, if financial constraints are limiting expansion, with resultant delivery constraints on sales, why not increase prices to limit sales, and use the extra cash to finance expansion?

Lesson 4 Make an effort to account for delays, especially in market response (for example, improvements in service will take a while to manifest themselves in improved sales, so avoid the temptation to cut back productive resources that will later be needed).

Lesson 5 Account for short-term productivity losses such as fatigue and experience dilution in resource expansion decisions (in the short-term, you may be getting less capacity than you think).

Lesson 6 Account for likely competitor responses in taking actions (it's easy for competitors to follow price changes, and trigger a price war; improvements in other components of product attractiveness are harder to detect and replicate).

Applications and Case Examples

The process and structural developments discussed in the last sections have formed the basis of numerous applications of system dynamics in support of business strategy.

In turn, these applications have provided the practice field through which the process has been refined, and the structural models, insights, and tools validated. While most of the published "real world" applications of system dynamics do not provide details of the models, they do nevertheless provide diagrams which show the nature of the model, representative results and policy conclusions, and the role the models played in business strategy formulation.

Space limitations preclude covering specific applications in depth. Therefore, I have chosen to reference applications in a number of different areas so readers can find references to the literature in their particular area of interest. Before getting to that, however, there are a couple of general references worthy of note: Roberts [91] covers many of the early published applications of system dynamics, with sections on manufacturing, marketing, research and development, and management and financial control. Coyle [14,15,16] touches on many of the applications initiated by his team at the University of Bradford, including work in the defense, natural resources, and utility industries. Richardson [86] provides an edited collection of academic journal articles containing some of the best work in system dynamics for business (and public) policy from its early years to the 1990s. Beyond these general compendiums, business strategy applications can perhaps best be classified by the industry of interest.

First, there have been a number of industry-wide models. The purpose of these models is typically to understand the drivers of change in the industry, and to forecast demand for use in other planning models. These include:

- Aircraft market as illustrated in Fig. 1 above [47,53]
- Health care market [42,103]
- Oil market [69,71,101]
- Shipping market [82].

Second, in addition to these industry-wide models, multiple applications to specific firms (which might also include some industry and/or competitive modeling), have been done in the following industry sectors:

- Utilities/Regulation – In the electric industry, work by Ford [24], Lyneis [51], Bunn and Larson [10,45] Ford covers some of this work elsewhere in ► System Dynamics Models of Environment, Energy and Climate Change.
- Telecoms [35]
- Financial Services – work for MasterCard [52] and in the insurance industry [7,13,96,102].

Finally, applications to specific types of firms, particularly small and medium size enterprises [9].

For particular examples of where system dynamics has had an impact in changing or forming business strategy, the MasterCard application described by Lyneis [52] and the General Motors OnStar application described by Barabba et al. [6] are noteworthy. These applications provide some detail about the model structure and describe how the modeling process changed management intuition and thereby led to significant shifts in business strategy. The MasterCard model represents growth and competitive dynamics in some detail, and is used to illustrate the multi-stage development process detailed in Sect. Using System Dynamics Models in Policy and Strategy Formulation and Implementation above. The OnStar example describes the process of modeling an industry that does not yet exist. The model itself builds from the diffusion model discussed above, with significant elaboration of potential customers, provision of service, alliances, dealers, and financial performance. The paper details the significant role that the system dynamics model played in reshaping GM's strategy.

Future Directions

System dynamics has made significant theoretical and practical contributions to business strategy. These contributions fall into two general categories: first, the process through which models are developed, working with management teams to enhance model validity and implementation; and second, understanding of the business structures and policies which cause observed problem behavior within firms and industries. Nevertheless, the impact of system dynamics on business strategy has been modest – relatively few firms use system dynamics. In my view, several factors contribute to this slow uptake. These are listed below, with possible actions that could be taken to alleviate.

1. Knowledge of system dynamics and its potential in strategy formulation is limited. Senior executives of firm's are unaware of system dynamics, or its potential [106]. In part this is because system dynamics is taught at relatively few business schools. Over time, this problem will be solved, but it will take a long time. But more importantly, researchers and practitioners of system dynamics rarely publish their work in publications which are widely read by senior management, such as the Harvard Business Review. This is a problem which can be addressed in the near future if researchers and practitioners made the effort to communicate with this market.
2. System dynamics is not well connected with traditional strategy research and practice. As noted earlier, system

dynamics and traditional strategy developed largely independently. While the potential synergies between the two are significant, until recently few researchers and practitioners have made the effort to cross disciplines. More effort and publication are needed to demonstrate areas of synergy and get system dynamics into the mainstream of business strategy research and ultimately practice. Warren [108] makes a start on this.

3. System dynamics is hard. Building system dynamics models, calibrating them to data, and analyzing their behavior to improve business strategy requires significant skill and experience. This has traditionally been developed via an apprenticeship program, either in university or consulting firms. Much more can be done to hasten this process. First, tried and true model structures that can be used as building blocks for models must be better documented, and gaps closed. While the theoretical basis for business dynamics described above is a good start, and reflects structures which all practicing system dynamicists should know, the underlying models and building blocks are widely dispersed and difficult to access. For example, the models in [100] are contained within a 900 page introductory textbook; the models in [49] are in old software and the book is out of print. In both cases, the models are only starting points and much unpublished work has occurred that expands these introductory models to make them more relevant and directly applicable to real business strategy problems. Efforts could and should be made to expand the library of business strategy building blocks. In addition, the process of building models and working with managers needs to be better documented and disseminated, both in formal courses and in published works, so as to facilitate the apprenticeship learning process.

Bibliography

1. Akkermans HA, Daellaert N (2005) The Rediscovery of Industrial Dynamics: The Contribution of System Dynamics to Supply Chain Management in a Dynamic and Fragmented World. Syst Dyn Rev 21(3):173–186
2. Akkermans HA, Vos B (2003) Amplification in service supply chains: an exploratory case study from the telecom industry. Prod Oper Manag 12(2):204–223
3. Anderson E, Fine C (1999) Business Cycles and Productivity in Capital Equipment Supply Chains. In: Tayur et al(eds) Quantitative Models for Supply Chain Management. Kluwer Academic Publishers, Norwell
4. Anderson E, Morrice D, Lundeen G (2005) The 'physics' of capacity and backlog management in service and custom manufacturing supply chains. Syst Dyn Rev 21(3):187–216
5. Backus G, Schwein MT, Johnson ST, Walker RJ (2001) Comparing expectations to actual events: the post mortem of a Y2K analysis. Syst Dyn Rev 17(3):217–235
6. Barabba V, Huber C, Cooke F, Pudar N, Smith J, Paich M (2002) A Multimethod Approach for Creating New Business Models: The General Motors OnStar Project. Interfaces 32(1):20–34
7. Barlas Y, Cırak K, Duman E (2000) Dynamic simulation for strategic insurance management. Syst Dyn Rev 16(1):43–58
8. Bass FM (1969) New product growth model for consumer durables. Manag Sci 15:215–227
9. Bianchi C (2002) Editorial to Special Issue on Systems Thinking and System Dynamics in Small-Medium Enterprises. Syst Dyn Rev 18(3):311–314
10. Bunn DW, Larsen ER (eds) (1997) Systems Modeling for Energy Policy. Wiley, Chichester
11. Campbell D (2001) The long and winding (and frequently bumpy) road to successful client engagement: one team's journey. Syst Dyn Rev 17(3):195–215
12. Carroll JS, Sterman JD, Marcus AA (1998) Playing the maintenance game: How mental models drive organizational decisions. In: Halpern JJ, Stern RN (eds) Nonrational Elements of Organizational Decision Making. Cornell University Press, Ithaca
13. Doman A, Glucksman M, Mass N, Sasportes M (1995) The dynamics of managing a life insurance Company. Syst Dyn Rev 11(3):219–232
14. Coyle RG (1996) System Dynamics Modelling: A Practical Approach. Chapman and Hall, London
15. Coyle RG (1997) System Dynamics at Bradford University: A Silver Jubilee Review. Syst Dyn Rev 13(4):311–321
16. Coyle RG (1998) The Practice of System Dynamics: Milestones, Lessons and Ideas From 30 Years Experience. Syst Dyn Rev 14(4):343–365
17. Coyle RG (2000) Qualitative and Quantitative Modeling in System Dynamics: Some Research Questions. Syst Dyn Rev 16(3):225–244
18. Coyle RG (2001) Rejoinder to Homer and Oliva. Syst Dyn Rev 17(4):357–363
19. Coyle RG, Morecroft JDW (1999) System Dynamics for Policy, Strategy and Management Education. J Oper Res Soc 50(4)
20. Croson R, Donohue K (2005) Upstream versus downstream information and its impact on the bullwhip effect. Syst Dyn Rev 21(3):187–216
21. Delauzun F, Mollona E (1999) Introducing system dynamics to the BBC World Service: an insider perspective. J Oper Res Soc 50(4):364–371
22. Doman A, Glucksman M, Mass N, Sasportes M (1995) The dynamics of managing a life insurance company. Syst Dyn Rev 11(3):219–232
23. Dyson RG, Bryant J, Morecroft J, O'Brien F (2007) The Strategic Development Process. In: O'Brien FA, Dyson RG (eds) Supporting Strategy. Wiley, Chichester
24. Ford AJ (1997) System Dynamics and the Electric Power Industry. Syst Dyn Rev 13(1):57–85
25. Forrester JW (1958) Industrial dynamics: a major breakthrough for decision makers. Harv Bus Rev 36(4):37–66
26. Forrester JW (1961) Industrial Dynamics. MIT Press, Cambridge (now available from Pegasus Communications, Waltham)
27. Forrester JW (1964) Common Foundations Underlying Engineering and Management. IEEE Spectr 1(9):6–77
28. Forrester JW (1968) Market growth as influenced by capital investment. Industrial Management Review 9(2):83–105.

Reprinted in Forrester JW (1975) Collected Papers of Jay W Forrester. Pegasus Communications, Waltham
29. Forrester JW (1969) Urban Dynamics. Pegasus Communications, Waltham
30. Foss NJ (ed) (1997) Resources, Firms and Strategies. Oxford University Press, Oxford
31. Gary MS (2005) Implementation Strategy and Performance Outcomes in Related Diversification. Strateg Manag J 26:643–664
32. Gonçalves PM (2003) Demand Bubbles and Phantom Orders in Supply Chains. Unpublished Dissertation Sloan School of Management. MIT, Cambridge
33. Gonçalves PM (2006) The Impact of Customer Response on Inventory and Utilization Policies. J Bus Logist 27(2):103–128
34. Gonçalves P, Hines J, Sterman J (2005) The Impact of Endogenous Demand on Push-Pull Production Systems. Syst Dyn Rev 21(3):187–216
35. Graham AK, Walker RJ (1998) Strategy Modeling for Top Management: Going Beyond Modeling Orthodoxy at Bell Canada. Proceedings of the 1998 International System Dynamics Conference, Quebec
36. Grant RM (2005) Contemporary Strategy Analysis, 5th edn. Blackwell, Cambridge
37. Hines JH, Johnson DW (1994) Launching System Dynamics. Proceedings of the 1994 International System Dynamics Conference, Business Decision-Making, Stirling
38. Homer JB (1996) Why We Iterate: Scientific Modeling In Theory and Practice. Syst Dyn Rev 12(1):1–19
39. Homer JB (1997) Structure, Data, and Compelling Conclusions: Notes from the Field. Syst Dyn Rev 13(4):293–309
40. Homer JB (1999) Macro- and Micro-Modeling of Field Service Dynamics. Syst Dyn Rev 15(2):139–162
41. Homer J, Oliva R (2001) Maps and Models in System Dynamics: A Response to Coyle. Syst Dyn Rev 17(4):347–355
42. Homer J, Hirsch G, Minniti M, Pierson M (2004) Models for Collaboration: How System Dynamics Helped a Community Organize Cost-Effective Care For Chronic Illness. Syst Dyn Rev 20(3):199–222
43. Kim DH, Lannon C (1997) Applying Systems Archetypes. Pegasus Communications Inc, Waltham
44. Kunc M, Morecroft J (2007) System Dynamics Modelling for Strategic Development. In: O'Brien FA, Dyson RG (eds) Supporting Strategy. Wiley, Chichester
45. Larsen ER, Bunn DW (1999) Deregulation in electricity: understanding strategic and regulatory risks. J Oper Res Soc 50(4):337–344
46. Levin G, Roberts EB, Hirsch GB (1976) The Dynamics of Human Service Delivery. Ballinger, Cambridge
47. Liehr M, Großler A, Klein M, Milling PM (2001) Cycles in the sky: understanding and managing business cycles in the airline market. Syst Dyn Rev 17(4):311–332
48. Lyneis JM (1975) Designing Financial Policies to Deal With Limited Financial Resources. Financ Manag 4(1)
49. Lyneis JM (1980) Corporate Planning and Policy Design: A System Dynamics Approach. M.I.T. Press, Cambridge
50. Lyneis JM (1998) Learning to Manage Growth: Lessons From a Management Flight Simulator, Proceedings of the 1998 International System Dynamics Conference (Plenary Session), Quebec City, 1998
51. Lyneis JM (1997) Preparing for a Competitive Environment: Developing Strategies for America's Electric Utilities. In: Bunn DW, Larsen ER (eds) Systems Modeling for Energy Policy. Wiley, Chichester
52. Lyneis JM (1999) System dynamics for business strategy: a phased approach. Syst Dyn Rev 15(1):37–70
53. Lyneis JM (2000) System Dynamics for Market Forecasting and Structural Analysis. Syst Dyn Rev 16(1):3–25
54. Lyneis JM, Cooper KG, Els SA (2001) Strategic Management of Complex Projects: A Case Study Using System Dynamics. Syst Dyn Rev 17(3):237–260
55. Lyneis JM, Ford DN (2007) System Dynamics Applied to Project Management: A Survey, Assessment, and Directions for Future Research. Syst Dyn Rev 23(2/3):157–189
56. Magee J (1958) Production Planning and Inventory Control. McGraw-Hill, London
57. Maier FH (1998) New product diffusion models in innovation management – a system dynamics perspective. Syst Dyn Rev 14(4):285–308
58. Meadows DL (1970) Dynamics of commodity production cycles. Pegasus Communications, Waltham
59. Merten PP, Loffler R, Wiedmann KP (1987) Portfolio Simulation: A Tool to Support Strategic Management. Syst Dyn Rev 3(2):81–101
60. Milling PM (1996) Modeling innovation processes for decision support and management Simulation. Syst Dyn Rev 12(3):211–23
61. Milling PM (2002) Understanding and managing innovation processes. Syst Dyn Rev 18(1):73–86
62. Milling PM (2007) A Brief History of System Dynamics in Continental Europe. Syst Dyn Rev 23(2/3):205–214
63. Montgomery CA (ed) (1995) Resource-Based and Evolutionary Theories of the Firm. Kluwer, Boston
64. Morecroft JDW (1983) A Systems Perspective on Material Requirements Planning. Decis Sci 14(1):1–18
65. Morecroft JDW (1984) Strategy Support Models. Strateg Manag J 5:215–229
66. Morecroft JDW (1985) The Feedback View of Business Policy and Strategy. Syst Dyn Rev 1(1):4–19
67. Morecroft JDW (1986) The Dynamics of a Fledgling High-Technology Growth Market. Syst Dyn Rev 2(1):36–61
68. Morecroft JDW (1999) Management attitudes, learning and scale in successful diversification: a dynamic and behavioural resource system view. J Oper Res Soc 50(4):315–336
69. Morecroft JDW (2007) Strategic Modelling and Business Dynamics: A Feedback Systems View. Wiley, Chichester
70. Morecroft JDW, Lane DC, Viita PS (1991) Modeling growth strategy in a biotechnology startup firm. Syst Dyn Rev 7(2):93–116
71. Morecroft JDW, van der Heijden KAJM (1994) Modeling the Oil Producers: Capturing Oil Industry Knowledge in a Behavioral Simulation Model. In: Morecroft JDW, Sterman JD (eds) Modeling for Learning Organizations. Productivity Press, Portland
72. Morecroft JDW, Wolstenholme E (2007) System Dynamics in the UK: A Journey from Stirling to Oxford and Beyond. Syst Dyn Rev 23(2/3):205–214
73. O'Brien FA, Dyson RG (eds) (2007) Supporting Strategy. Wiley, Chichester
74. Oliva R (1996) A Dynamic Theory of Service Delivery: Implications for Managing Service Quality. Ph.D. Thesis, Sloan School of Management, Massachusetts Institute of Technology

75. Oliva R (2001) Tradeoffs in responses to work pressure in the service industry. Calif Manag Rev 43(4):26–43
76. Oliva R, Sterman JD, Giese M (2003) Limits to growth in the new economy: exploring the 'get big fast' strategy in e-commerce. Syst Dyn Rev 19(2):83–117
77. Packer DW (1964) Resource Acquisition in Corporate Growth. M.I.T. Press, Cambridge
78. Paich M, Sterman JD (1993) Boom, Bust, and Failures to Learn in Experimental Markets. Manag Sci 39(12)
79. Porter ME (1980) Competitive Strategy. The Free Press, New York
80. Porter ME (1985) Competitive Advantage. The Free Press, New York
81. Prahalad CK, Hamel G (1990) The Core Competence of the Corporation. Harv Bus Rev (May-June):71–91
82. Randers J, Göluke U (2007) forecasting turning points in shipping freight rates: lessons from 30 years of practical effort. Syst Dyn Rev 23(2/3):253–284
83. Repenning N, Kofmann F (1997) Unanticipated Side Effects of Successful Quality Programs: Exploring a Paradox of Organizational Improvement. Manag Sci 43(4)
84. Repenning NP, Sterman JD (2001) Nobody Ever Gets Credit for Fixing Problems that Never Happened: Creating and Sustaining Process Improvement. Calif Manag Rev 43(4):64–88
85. Repenning NP, Sterman JD (2002) Capability traps and self-confirming attribution errors in the dynamics of process improvement. Adm Sci Q 47:265–295
86. Richardson GP (1996) Modelling for Management: Simulation in Support of Systems Thinking. Dartsmouth Publishing Company, Aldershot
87. Richardson GP, Andersen DF (1995) Teamwork in Group Model Building. Syst Dyn Rev 11(2):113–138
88. Richmond B (1997) The Strategic Forum: Aligning Objectives, Strategy and Process. Syst Dyn Rev 13(2):131–148
89. Risch J, Troyano-Bermúdez L, Sterman J (1995) Designing Corporate Strategy with System Dynamics: A Case Study in the Pulp and Paper Industry. Syst Dyn Rev 11(4):249–274
90. Roberts EB (1977) Strategies for the Effective Implementation of Complex Corporate Models. Interfaces 8(1):26–33
91. Roberts EB (ed) (1978) Managerial Applications of System Dynamics. The MIT Press, Cambridge
92. Roberts EB (2007) Making System Dynamics Useful: A Personal Memoir. Syst Dyn Rev 23(2/3):119–136
93. Rockart SF, Lenox MJ, Lewin AY (2007) Interdependency, Competition, and Industry Dynamics. Manag Sci 53(4):599–615
94. Snabe B, Grossler A (2006) System Dynamics Modelling for Strategy Implementation—Case Study and Issues. Syst Res Behav Sci 23:467–481
95. Senge PM (1990) The Fifth Discipline: The Art and Practice of The Learning Organization. Doubleday, New York
96. Senge PM, Sterman JD (1992) Systems thinking and organizational learning: Acting locally and thinking globally in the organization of the future. Eur J Oper Res 59(1):137–150
97. Sice P, Mosekilde E, Moscardini A, Lawler K, French I (2000) Using system dynamics to analyse interactions in duopoly competition. Syst Dyn Rev 16(2):113–133
98. Sterman JD (1989) Modeling management behavior: misperceptions of feedback in a dynamic decision making experiment. Manag Sci 35:321–339
99. Sterman JD (1989) Strategy Dynamics: the Rise and Fall of People Express. Memorandum D-3939-1 Sloan School of Management, M.I.T
100. Sterman JD (2000) Business Dynamics: Systems Thinking and Modeling for a Complex World. McGraw-Hill, New York
101. Sterman JD, Richardson G, Davidsen P (1988) Modeling the estimation of petroleum resources in the United States. Technol Forecast Soc Chang 33(3):219–249
102. Thompson JP (1999) Consulting approaches with system dynamics: three case studies. Syst Dyn Rev 15(1):71–95
103. Thompson JP (2006) Making sense of US health care system dynamics. Proceedings of the 2006 International System Dynamics Conference, Nijmegen, The Netherlands
104. Vennix JAM (1996) Group Model Building: Facilitating Team Learning Using System Dynamics Field. Wiley, Chichester
105. Warren KD (2002) Competitive Strategy Dynamics. Wiley, Chichester
106. Warren KD (2004) Why Has Feedback Systems Thinking Struggled to Influence Strategy and Policy Formulation? Suggestive Evidence, Explanations and Solutions. Syst Res Behav Sci 21:331–347
107. Warren KD (2005) Improving Strategic Management With the Fundamental Principles of System Dynamics. Syst Dyn Rev 21(4):329–350
108. Warren KD (2007) Strategic Management Dynamics. Wiley, Chichester
109. Weil HB (1980) The Evolution of an Approach for Achieving Implemented Results from System Dynamics Projects. In: Randers J (ed) Elements of the System Dynamics Method. MIT Press, Cambridge
110. Wolstenholme EF (1999) Qualitative vs. Quantitative Modelling: The Evolving Balance. J Oper Res Soc 50(4):422–428
111. Winch GW (1993) Consensus building in the planning process: benefits from a "hard" modeling approach. Syst Dyn Rev 9(3):287–300
112. Winch GW (1999) Dynamic Visioning for Dynamic Environments. J Oper Res Soc 50(4):354–361
113. Winch GW (2001) Management of the "skills inventory" in times of major change. Syst Dyn Rev 17(2):151–159
114. Zahn E, Dillerup R, Schmid U (1998) Strategic evaluation of flexible assembly systems on the basis of hard and soft decision criteria. Syst Dyn Rev 14(4):263–284

C

Carbon Nanotubes, Thermo-mechanical and Transport Properties of

H. Rafii-Tabar[1,2]
[1] Department of Medical Physics and Biomedical Engineering, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[2] Department of Nano-science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

Article Outline

Glossary
Definition of the Subject
Introduction
Geometry of SWCNT, MWCNT and SWCNH
Simulation at Nano-Scale
Modeling Fluid Transport and Gas Storage Properties of Nanotubes
Modeling the Mechanical Properties of Nanotubes
Modeling the Thermal Properties of Nanotubes
Concluding Remarks
Bibliography

Glossary

Scanning tunneling microscope (STM) A nano-technology instrument capable of imaging the topography of conducting surfaces with atomistic resolution. It can also be used to manipulate individual atoms and molecules, and construct nano-scale structures.

Atomic force microscope (AFM) A nano-technology instrument for investigation of material surfaces on atomistic and molecular scales. It can be used to map the topography of non-conducting surfaces, by sensing the inter-atomic forces, and produce three-dimensional images of these surfaces.

Spintronics Refers to the field of spin-based electronics rather than charge-based electronics. It will lead to a new type of device that is based on the use of electron spin for transfer and storage of information.

Brillouin zone In solid state physics several Brillouin zones can be defined. The first zone is defined as the Wigner–Seitz primitive cell of the reciprocal lattice. The nth Brillouin zone is defined as the set of points that are reached from the origin by crossing $(n-1)$ Bragg planes.

Monte Carlo (MC) method In computational modeling, this method provides a probabilistic scheme for solving a variety of problems by employing powerful sampling techniques. In nano-science and condensed matter physics, one application of this method is for computing the minimum energy state of a nano-structure.

Stochastic dynamics (SD) method This refers to the computer simulation method wherein the Langevin equation of motion, describing the random behavior of a particle, is solved as opposed to the deterministic MD method in which Newton's equations of motion are solved.

Nano-electromechanical systems (NEMS) These are nano-technology based systems that are the smaller versions of the micro-electromechanical systems (MEMS). They are capable of measuring small displacements, sensing minute amount of substances, and performing rotary motions. NEMS can be constructed via either the top-down approach, i. e., via miniaturization of the micro-scale devices, or via the bottom-up approach, i. e., by positioning single atoms or molecules so that a complex and functional nano-system is built from below the nano-scale.

Ab initio approach This is the first-principles approach to the computation of the properties, especially the electronic-structure properties, of nano-scale systems using quantum-mechanical concepts and methods. In this method, the structure of a molecule, for instance, is obtained purely from a knowledge of its composition by solving the Schrödinger equation.

Definition of the Subject

Carbon nanotubes form the fourth allotrope of crystalline carbon after graphite, diamond, and a variety of caged-like fullerene molecules, and were discovered in the early 1990s. Their mechanical properties make them stronger than steel, and their thermal conductivity is faster than copper. They have very exotic electronic-conduction properties, namely, by changing their geometry, or introducing topological defects into their structure, their electronic conductance can change from metals to semiconductors. They can also be used to store gases and transport fluids. Furthermore, nano-scale robots, machines, and sensors can be constructed from them, and these can be used to deliver drugs to specific locations in the body, or detect individual cancer cells, or be used as molecular filters to separate minute particles from the environment. Carbon nanotubes are referred to as the fabric of nanotechnology, and will play a central role in the future development of this technology. Understanding the properties of nanotubes, via computational simulation studies, has been one of the most intensive areas of research in physical sciences during the past 20 years.

Introduction

The current scientific-technological arena is distinguished by the presence of *four* distinct and highly complex disciplines, namely:

1) information science, underlying the fields of information technology, high performance computing, and computational science,
2) molecular genetics and molecular biology, underlying the fields of genetic engineering and bio-technology,
3) neuro and cognitive sciences, and their related fields of neural- and brain-based technologies,
4) nano-science, and its related fields of nano-technology and molecular-scale engineering, making atom-by-atom manipulation of the physical, biological and smart matter possible.

These four areas are strongly correlated, and the post-modern era in science and technology will be characterized by research efforts to establish, step-by-step, a close synergy among these fields, leading to their total integration within a *unified* and holistic framework, promoting the convergence of all these seemingly separate and compartmentalized branches and paving the way for the emergence of a *convergent* technology [49]. Such a technology will produce advanced man-made machines, devices, components, and materials that carry some of the most essential characteristics of the biological and smart systems, such as self-repair, self-assembly, self-correction of internal faults, re-production, ability to establish communication, and adaptation to unfamiliar environments.

The convergence takes place at the nano-scale (1–100 nanometers) since the main building blocks of the physical, biological and smart matter, i. e., the physical and biological *nano-structures*, and the laws governing their evolution, are formed at this scale. It is, therefore, no exaggeration to state that nano-scale science and nano-scale technology [17] represent the main components of the 21st century science and technology. It is now accepted that nano-science can be defined as *the study of structures, dynamics, and properties of systems wherein one or more of spatial dimensions varies in the nanoscopic range*. At this scale, the dynamics and properties of systems are distinctly different, very often in quite unforeseen ways, from those in microscopic and macroscopic systems.

Nanoscopic structures operate on very reduced time and energy scales, and are constructed from a *countable* number of atoms and molecules. Their sizes are located between those of molecules and micro-structures, and their distinguishing feature is their high aspect, or surface-to-volume, ratio. Nano-technology represents the combination of atomic and molecular level techniques and instrumentations for constructing nano-scale devices, systems, and components from the *assemblies* of individual nano-structures. On the other hand, nano-technology is an *enabling* technology, injecting nano-structures into micro- and macro-scale, devices, materials, and systems to enhance, or radically change, their electronic, mechanical, optical and thermal properties. For instance, nano-structured materials, i. e., materials with nano-sized grains, or materials injected with nano-grains, can show very different mechanical, thermal, electronic and optical properties. For instance, it is known that nano-structured Fe, Cu and Ni have electrical resistances respectively 55%, 15% and 35% higher than the coarse-grained polycrystalline samples [38].

Nano-structures can be assembled by a precise positioning of the atoms at specified locations, using such highly sensitive devices as the scanning tunneling microscope (STM) [9], and the atomic force microscope (AFM) [8]. These devices can provide access to the detailed topography, crystal structure, and the electronic-structure maps at material surfaces and of individual nano-structures. Their use has led to the design and fabrication of devices, and materials, that manifest novel physical, chemical and biological properties.

Among the multitude of nano-structures currently used in nano-technology, carbon nano-structures occupy a central position, with wide-ranging applications in prac-

tically all the nano-technology sub-fields. Carbon is a fundamental element, crystallizing in several allotropes. In the 1970s, *diamond* and *graphite* were the only known forms of condensed carbon, and a third allotrope, the cage-like *fullerene* molecules, was discovered and synthesised in macroscopic amounts by mid 1980s. In 1990s, the fourth allotrope, the multi-walled carbon *nanotube* (MWCNT), and the single-walled carbon nanotube (SWCNT) were synthesized [20,21]. Several growth techniques, namely, the arc-discharge, the laser ablation, and recently the chemical vapor deposition method, have been developed. The appearance of nanotubes has led to the emergence of new and very active areas of research within several fields; in computational and experimental nano-science, industrial nano-technology, theoretical and experimental condensed matter physics and chemistry, materials science and engineering, bio-physics, medical nano-technology, molecular genetics, nano-biotechnology, information technology device fabrication, optics, electro-mechanical systems and electronics, carbon nanotubes are viewed as highly relevant nano-structures with extensive potential applications.

Originally, the carbon nanotubes were of *multi-walled* variety, with outer diameters in the range of 4–30 nm and lengths of up to 1 μm [21]. These consisted of seamless concentric cylindrical shells, separated by 3.4 Å. Each shell was a rolled-up version of a two-dimensional graphene sheet. The *single-walled* nanotubes were synthesized later [6,22], and consisted of single graphene sheets with a typical diameter on the order of 1.4 nm, similar to the diameter of a C_{60} molecule, and lengths usually on microscopic orders [21]. It was also found that these SWCNTs can bundle together to form a *rope*, normally patterned as a hexagonal array, with a morphology very similar to porous materials, and membranes, with nanometer spaces available both inside the nanotubes and in the interstitial channels between them that can be utilized for *storage* of adsorbed gases, and *flow* of fluids, turning them into filtering channels and molecular sieves.

Another form of SWCNTs, called single-walled carbon *nanohorn* (SWCNH), has also been synthesized [23]. This horn-shaped material has a closed structure, with a cylindrical part and a horn-tip part. Its internal space is not normally available as a storage medium, but heat-treatment within an oxygen environment promotes the appearance of windows on the its walls, allowing for gas and liquid particle transfer to the interior.

Nanotubes incorporating fullerenes, such C_{60} molecules, have also been synthesized [55], and are referred to as *peapods*. The internal space of the SWCNT is filled with a regularly-arranged formation of fullerene molecules [21]. Peapods, when heated to 1000–1200°C, transform into double-walled carbon nanotubes (DWCNTs), as the encapsulated molecules coalesce [21]. Peapods provide a good model of nano-platforms for targeted drug delivery.

As will be discussed later, an SWCNT is identified by two *chiral* indices (n, m) that determine its diameter, and depending on their values, three *different* classes of nanotube are observed. These are the $(n, 0)$, or the *zigzag*, nanotube, the (n, n), or the *armchair* nanotube, and the $(2n, n)$, or the general *chiral* nanotube. They have very different *electronic* conduction properties. The graphene sheet is a semi-metal with a zero band-gap. The electronic states of an infinitely long nanotube are continuous in its axial direction, but are quantized along its circumference. For the (n, n) armchair nanotubes, there always exist electronic states that cross the corner points of the first Brillouin zone, making these nanotubes a *metallic* material. For the (n, m) nanotubes, if $(n - m)/3 \neq$ an integer, the electronic states miss the corner points, and these nanotubes behave as semi-conductors with a band-gap that scales with the diameter d of the tube as $1/d$, and which is on the order of 0.5 eV. If $(n - m)/3 =$ an integer, however, certain electronic states are located on the corner points, and these nanotubes behave as semi-metals, but become small-band semi-conductors due to the curvature-induced rehybridization [14]. These remarkable electronic-conduction properties, that are linked with the geometry of the material, are also very partial to local deformations in the geometry. Recent exploitation of the electronic properties of carbon nanotubes includes the field of spin-electronics, or *spintronics* that utilizes the spin of the electrons to transfer and store of information.

Nanotubes have also extraordinary mechanical, thermal, mass-transport, and gas-storage properties. For example, their Young's modulus, being a measure of their stiffness, is estimated to be of the order of 1.5 to nearly 5 TPa, and their tensile strength is far higher than that of steel. They are, therefore, strong candidates as functional units in molecular-scale machinery, and highly complex nano-electromechanical systems (NEMS), and as probe *tips* in STM and AFM [56]. Their high thermal conductivity, far exceeds that of diamond. They can also act as media for the transport and storage of various types of gases and fluids.

The investigation into nanotube properties has prompted an intensive experimental and theoretical/computational research, leading to major discoveries, and the appearance of several thousand patents and publications in fields of basic sciences, engineering, and medicine. One of the most active areas has involved the use of predic-

tive computational modeling of their mechanical, thermal and mass transport properties [44]. These computational modelings and simulations have been performed via two distinct methodologies:

1) using quantum mechanical methods, such as the density functional theory (DFT) of atoms and molecules [36] that provides an *ab initio* approach to computation of the properties of nano-scale systems composed of several *ten* to several *hundred* atoms,
2) using classical statistical mechanics methods, such as the molecular dynamics (MD), Monte Carlo (MC) and stochastic dynamics (SD) simulation methods [2] that can handle nano-scale structures composed of several *thousand* to several *billion* atoms. These simulation methods require the use of phenomenological inter-atomic potentials to model the energetics of the system. These potentials play a very crucial role, and the more accurate they are, the closer the simulation results approach the experimental data. A good deal of efforts have been focused to develop highly accurate inter-atomic potentials for different classes of materials, particularly the covalently-bonded ones. Some of these potentials are *many-body* potentials, and most of the simulations concerning carbon nanotubes have used these many-body potentials.

Research into the properties of carbon nanotubes has led to the appearance of several informative reviews, among which one can mention those by [1,5,16,40,43,45]. In addition, an impressive number of encyclopedias and handbooks [7,34,48,51] covering most sub-fields of this topic are also available.

In this review I have surveyed the field of computational modeling of the thermo-mechanical, transport and storage properties of nanotubes. In the interest of saving space, I have only described, albeit very briefly, one of the computational methods that has been extensively employed in these research studies, namely the MD simulation method. The interested reader can find a very good summary of the other methods in [43]. An extensive part of the modeling-based research into the mechanical properties has employed concepts from the field of continuum elasticity theory, using such structures such as *curved plates, shells, vibrating rods, and bent beams*. An easy-to-follow introduction to these topics can found in my book [44].

The organization of this review is as follows. In Sect. "Geometry of SWCNT, MWCNT and SWCNH", the description of the geometrical structure of the SWCNTs, MWCNTs and the SWCNHs is briefly considered. In Sect. "Simulation at Nano-Scale" the basic principles of the classical MD simulation method is presented. In Sect. "Modeling Fluid Transport and Gas Storage Properties of Nanotubes", the flow of fluids and the storage of gases in nanotubes are considered, while in Sect. "Modeling the Mechanical Properties of Nanotubes", the mechanical properties are considered. Section. "Modeling the Thermal Properties of Nanotubes" summarizes the thermal conductivity and specific heat properties of nanotubes, and Sect. "Concluding Remarks" presents an overall summary together with the future directions.

Geometry of SWCNT, MWCNT and SWCNH

The SWCNT

We can construct an SWCNT by rolling a 2D graphene sheet into a cylinder. Consider a lattice point O as the origin in a graphene sheet, as shown in Fig. 1. The 2D Bravais lattice vector of the graphene sheet $\mathbf{C}_h$, referred to as the *chiral* vector, can be used to reach any other equivalent point Z on the sheet. The vector $\mathbf{C}_h$ is constructed from the pair of unit-cell vectors $\mathbf{a}_1$ and $\mathbf{a}_1$ as

$$\mathbf{C}_h = n\mathbf{a}_1 + m\mathbf{a}_2 \,, \tag{1}$$

where m and n are a pair of integers, and

$$\mathbf{a}_1 = (\frac{\sqrt{3}}{2}, \frac{1}{2})a \,, \quad \mathbf{a}_2 \quad = (\frac{\sqrt{3}}{2}, -\frac{1}{2})a \,, \tag{2}$$

where $a = 2.46$ Å is the lattice constant of graphite

$$a = \sqrt{3}a_{\text{C–C}} \,, \tag{3}$$

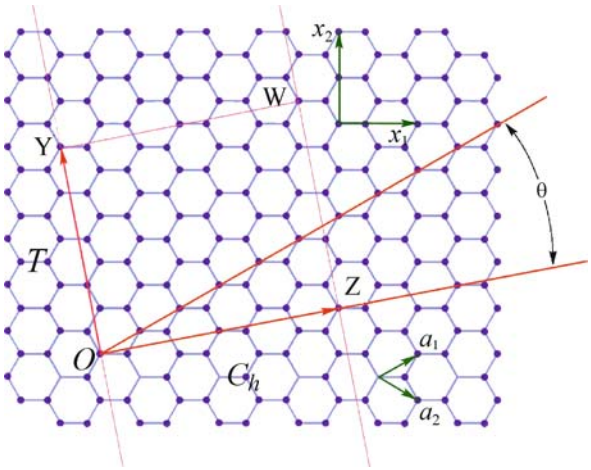

Carbon Nanotubes, Thermo-mechanical and Transport Properties of, Figure 1
The geometry of a two-dimensional graphene sheet, showing the relevant vectors that characterize a single-walled carbon nanotube (SWCNT). Figure based on [15]

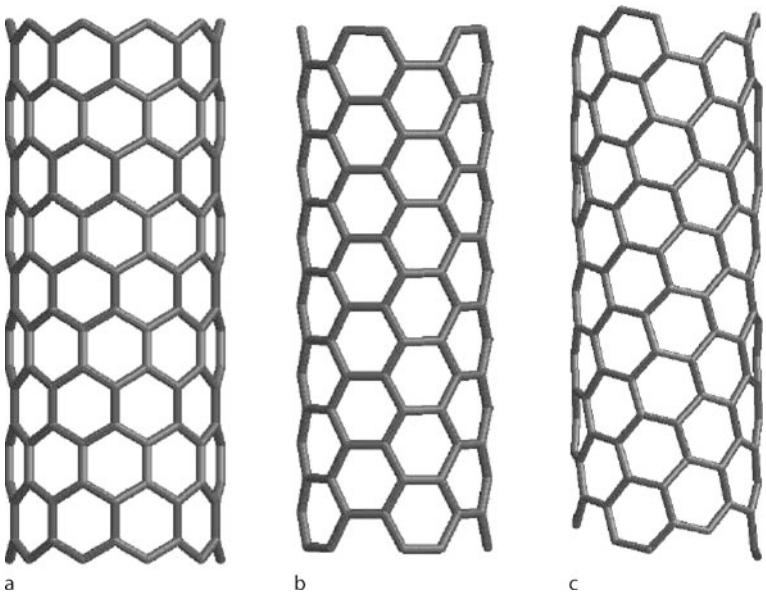

Carbon Nanotubes, Thermo-mechanical and Transport Properties of, Figure 2
The outlines of three types of nanotube: a a (10,0) zig-zag nanotube; b a (5,5) armchair nanotube; c a (7,3) general chiral nanotube

and $a_{C\text{-}C}$ is the carbon-carbon bond length. The angle θ that $\mathbf{C}_h$ makes with the zigzag axis of the sheet is referred to as the chiral angle, and (n, m) are called the chiral indices. If the sheet is rolled, and the point characterized by the pair of values of (n, m) coincides with the origin O, then an (n, m) nanotube is generated. The *zigzag* direction corresponds to $\theta = 0$, and when the direction of rolling is along this axis, a zigzag nanotube is formed. When $\theta = \pi/6$, the direction is called the *armchair* direction, and rolling along this direction generates an armchair nanotube. A nanotube generated for any other value $0 < \theta < \pi/6$ is referred to as a general chiral nanotube. Figure 2 shows the schematic representations of these three types of nanotube.

The circumference of the nanotube, i. e., the length of the chiral vector, is given by [15]

$$L = |\mathbf{C}_h| = a(n^2 + m^2 + nm)^{\frac{1}{2}} , \tag{4}$$

and, hence, the diameter is

$$d_t = \frac{L}{\pi} = \frac{(n^2 + m^2 + nm)^{\frac{1}{2}}}{\pi} a . \tag{5}$$

Other relationships connecting the chiral angle with the chiral indices are [15]

$$\sin\theta = \frac{\sqrt{3}m}{2(n^2 + m^2 + nm)^{\frac{1}{2}}} ,$$

$$\cos\theta = \frac{2n + m}{2(n^2 + m^2 + nm)^{\frac{1}{2}}} ,$$

$$\tan\theta = \frac{\sqrt{3}m}{2n + m} . \tag{6}$$

Hence, if $n = m$, then $\theta = \pi/6$ and the resulting nanotube is an (n, n), or an armchair, nanotube. If $m = 0$, then $\theta = 0$, and an $(n, 0)$, or a zigzag, nanotube is obtained. The nanotubes in the general chiral category are $(2n, n)$.

The rectangle bounded by the chiral vector $\mathbf{C}_h$ and the vector $\mathbf{T}$, i. e., OYWZ, where Y is the first lattice point through which the vector passes, is the unit-cell of the nanotube. This vector is written as [15]

$$\mathbf{T} = t_1\mathbf{a}_1 + t_2\mathbf{a}_2 , \tag{7}$$

where t_1 and t_2 are a pair of integers, related to m and n via

$$t_1 = \frac{2m + n}{d_R} , t_2 = -\frac{2n + m}{d_R} , \tag{8}$$

and d_R denotes the highest common divisor (HCD) of $(2n + m, 2m + n)$, given by

$$d_R = \left\{ \begin{array}{ll} = d & \text{if } (n - m) \text{ not a multiple of } 3d , \\ = 3d & \text{if } (n - m) \text{ multiple of } 3d , \end{array} \right\} \tag{9}$$

where d is the HCD of (n, m). The length T of the vector $\mathbf{T}$ is given by

$$T = |\mathbf{T}| = \frac{\sqrt{3}L}{d_R} . \tag{10}$$

The unit-cell of the nanotube contains $2N$ atoms, where

$$N = \frac{2(n^2 + m^2 + nm)}{d_R} . \tag{11}$$

Carbon nanotubes can be capped at both ends [15]. This can be done by bisecting a C_{60} molecule into two hemispheres and joining these with a cylindrical nanotube of the same diameter as the molecule.

An SWCNT is a low-energy structure with a distinct topology, brought about by distorting the geometry of the 2D graphene sheet into the third dimension. The act of rolling the sheet calls for the expenditure of a modest amount of strain, or curvature, energy, which is the difference between the total energy of a carbon atom in an SWCNT and that in a graphene sheet. In nanotubes with a diameter less than 1 nm, this energy penalty is more significant since the bond angles deviate far below the ideal 120° angles in the graphene sheet.

The MWCNT

Two forms of MWCNTs have been found in experiments. These are a thermodynamically-stable system composed of nested coaxial cylinders [20], and a metastable type with a scroll structure [26]. The transition from the latter type to the nested type is attributed [26] to the presence of the dislocation-like defects in the scroll type. The inter-shell gap in an MWCNT is estimated to be ≈3.4 Å [50], very close to the inter-planar gap in the graphite crystal. Several other values of this gap, in the range 3.42 Å to 3.75 Å, have also been reported [25], with the gap increasing when the nanotube diameter decreases. The simplest type of an MWCNT is a DWCNT. To construct a DWCNT, the chiral vector of the inner SWCNT, with indices (n_1, m_1), is related to the chiral vector of the outer SWCNT, with indices (n_2, m_2). These indices are related via

$$m_2^2 + m_2 n_2 + (n_2^2 - \kappa(n_1, m_1)) = 0 , \tag{12}$$

where

$$\kappa(n_1, m_1) = \left[\frac{2\pi(r_{t1} + r_g)}{a}\right]^2 , \tag{13}$$

and r_{t1} and r_g are respectively the radius of the inner nanotube, and the inter-shell gap. Then n_2 and m_2 can be obtained by solving (12). For example, for a (9,6) inner nanotube, and $n_2 = 15$, we find $m_2 = 9.96$ which rounds to $m_2 = 10$ and, therefore, the nanotube (9,6) can be nested inside the nanotube (15,10).

The SWCNH

These objects are distinguished by their horn shape. Pure SWCNH samples are relatively easier to produce than pure SWCNT samples [33]. SWCNHs are always produced with closed tips, incorporating pentagons into their hexagonal lattices to close the ends. They consist of a tubule part, and a conical cap whose average angle is 20°, implying that the caps contain five pentagon rings and several hexagon rings. The average length of SWCNHs is 30-50 nm, and the separation between neighboring SWCNH is about 0.35 nm. The average diameter of the tubular parts is 2-3 nm, larger than the 1.4 nm diameter of a typical SWCNT. The interstitial spaces in an SWCNH assembly provide the external micropore and mesopore spaces [33,35], where pores with a width of less than 2 nm are called micropores, and those with a width between 2 and 50 nm are called mesopores. The width of the internal pores in SWCNHs is close to the critical size of 2 nm [33]. Oxidation produces windows on the walls of a closed SCWNH [35], making available 11% and 36% of the closed spaces at $T = 573$ K and 623 K respectively [32].

Simulation at Nano-Scale

The time-evolution of the energetics and dynamics of complex nano-structures and nano-processes, involving several billions of atoms, can be accurately studied via simulation methods that use concepts from classical statistical mechanics. The information gained from these simulations can be profitably used in the design of nano-scale components, nano-structured materials, and structures dominated by nano-interfaces. Furthermore, the construction of functional assemblies of nano-structures requires a deep understanding of the interaction between individual nano-structures, and this can be handled by the use of such simulation methods. We briefly consider the essential elements of one of the most popular simulation methods at the nano-scale, namely the MD simulation method.

MD Simulation Method

Classical MD simulation [2,18,46] studies the motion of individual atoms in an assembly of N atoms or molecules employing either the Newtonian, or the stochastic dynamics, when the initial spatial configuration and velocities of these atoms are given. In MD, the N-atom structure is confined to a simulation cell of volume V, and this cell is replicated in all dimensions, generating the periodic images of itself and of the N atoms. This periodic boundary condition (PBC) is necessary to compensate for the undesirable effects of the artificial surfaces associated with the finite

size of the simulated system. The energetics of the system are modeled by two- or many-body inter-atomic potentials $H_I(r_{ij})$, and in the simulation the force experienced by each atom is obtained from

$$\mathbf{F}_i = -\sum_{j>i} \nabla_{\mathbf{r}_i} H_I(r_{ij}) \,, \tag{14}$$

where r_{ij} is the distance between atoms i and j. The $3N$ coupled equations of motion are integrated numerically via a variety of numerical techniques, all based on the finite-difference method. One popular integration scheme is the velocity Verlet algorithm [2], whereby the positions $\mathbf{r}_i$ and velocities $\mathbf{v}_i$ of the atoms of mass m_i are updated at each simulation time step $\mathrm{d}t$ according to

$$\begin{aligned}
\mathbf{r}_i(t+\mathrm{d}t) &= \mathbf{r}_i(t) + \mathbf{v}_i(t)\mathrm{d}t + \frac{1}{2}(\mathrm{d}t)^2 \frac{\mathbf{F}_i(t)}{m_i} \,, \\
\mathbf{v}_i(t+\frac{1}{2}\mathrm{d}t) &= \mathbf{v}_i(t) + \frac{1}{2}\mathrm{d}t \frac{\mathbf{F}_i(t)}{m_i} \,, \\
\mathbf{v}_i(t+\mathrm{d}t) &= \mathbf{v}_i(t+\frac{1}{2}\mathrm{d}t) + \frac{1}{2}\mathrm{d}t \frac{\mathbf{F}_i(t+\mathrm{d}t)}{m_i} \,.
\end{aligned} \tag{15}$$

At each $\mathrm{d}t$, the exact *instantaneous* values of thermodynamical observables, such as the temperature and pressure

$$\begin{aligned}
T_{\mathrm{ins}} &= \frac{1}{3Nk_{\mathrm{B}}} \sum_i^N \frac{|\,\mathbf{p}_i^2\,|}{m_i} \,, \\
P_{\mathrm{ins}} &= \frac{1}{3V} \left(\sum_i^N \frac{|\,\mathbf{p}_i^2\,|}{m_i} + \sum_{i=1}^N \sum_{j>i}^N \mathbf{r}_{ij}.\mathbf{F}_{ij} \right) \,.
\end{aligned} \tag{16}$$

can be computed, where $\mathbf{p}_i$ is momentum of particle i, k_{B} is the Boltzmann constant and $\mathbf{F}_{ij}$ is the force experienced by atom i due to atom j. These instantaneous data then allow for the computation of time-averaged values at the conclusion of the simulation.

Constant-Temperature MD: The Nosé–Hoover Method

MD simulations are generally performed on closed isothermal systems, represented in statistical mechanics by canonical ensembles wherein N, V, and the temperature T of the members are all fixed [37]. A constant-temperature MD simulation can be realized in a variety of ways [42], and a method that generates the canonical ensemble distribution in *both* the configuration and momentum parts of the phase space was proposed by Nosé [42]. Its application leads to a modification of the equations of motion (15) to the following forms [59]

$$\begin{aligned}
\mathbf{r}_i(t+\mathrm{d}t) &= \mathbf{r}_i(t) + \mathbf{v}_i(t)\mathrm{d}t \\
&\quad + \frac{1}{2}\mathrm{d}t^2 \left[\frac{\mathbf{F}_i(t)}{m_i} - \eta(t)\mathbf{v}_i(t) \right] \,, \\
\mathbf{v}_i(t+\frac{\mathrm{d}t}{2}) &= \mathbf{v}_i(t) + \frac{\mathrm{d}t}{2} \left[\frac{\mathbf{F}_i(t)}{m_i} - \eta(t)\mathbf{v}_i(t) \right] \,, \\
\eta(t+\frac{\mathrm{d}t}{2}) &= \eta(t) + \frac{\mathrm{d}t}{2\,Q} \left[\sum_i^N m_i \mathbf{v}_i^2(t) - g k_{\mathrm{B}}\, T \right] \,, \\
\eta(t+\mathrm{d}t) &= \eta(t+\frac{\mathrm{d}t}{2}) \\
&\quad + \frac{\mathrm{d}t}{2\,Q} \left[\sum_i^N m_i \mathbf{v}_i^2(t+\frac{\mathrm{d}t}{2}) - g k_{\mathrm{B}}\, T \right] \,, \\
\mathbf{v}_i(t+\mathrm{d}t) &= \frac{2}{2+\eta(t+\mathrm{d}t)\,\mathrm{d}t} \\
&\quad \cdot \left[\mathbf{v}_i(t+\frac{\mathrm{d}t}{2}) + \mathrm{d}t \frac{\mathbf{F}_i(t+\mathrm{d}t)}{2\,m_i} \right] \,.
\end{aligned} \tag{17}$$

where Q is given by

$$Q = g\, k_{\mathrm{B}}\, T\, \tau^2 \,, \tag{18}$$

and η is the friction coefficient of the heat bath connected to the system, τ is the relaxation time of this bath, normally of the same order of magnitude as $\mathrm{d}t$, and $g = 3(N-1)$ is the number of degrees of freedom. The parameter τ controls the speed with which the bath damps down the fluctuations in the temperature.

Modeling Fluid Transport and Gas Storage Properties of Nanotubes

Fluid flow and gas storage in nano-scale objects, like carbon nanotubes, is significantly different as compared with the corresponding activities in microscopic and macroscopic structures. For instance, fluid flow in a nano- scale machine is not similar to the flow in a large-scale machine, since in the latter case, there is no need to consider the atomic structure of the fluid, and it can be characterized by its viscosity, density and other bulk properties [61]. Furthermore, in large-scale objects, the so-called *no-slip* boundary condition is often invoked, according to which the fluid velocity is negligibly small at the fluid-wall interface.

Reduction in length scales immediately introduces new phenomena, such as *diffusion*, into the system. Moreover, for flow, or storage, in nano-scale objects, the dynamics of the walls, the fluid, or the gas, and their mutual interaction, must all be considered. We note that the dynamics of the walls can be strongly size-dependent.

An added difficulty is that, at nano-scale, standard classical notions such as pressure and viscosity may also not have a clear definition. For example, the surface area of a nano-scale object, such as a nanotube, may be ambiguous. Notwithstanding these difficulties, modeling fluid flow and gas storage in nanotubes has been an active area of research, with practical consequences. In this section we consider a set of studies, each highlighting one aspect of transport and storage properties of nanotubes.

Modeling Fluid Flow

Let us consider the MD simulation of the flow of fluids, composed of Ar and He particles, through an SWCNT [61]. In simulations, the fluid atoms were always dynamic, whereas for the nanotube, the first 10 and the last 10 rings were always frozen, and the rings in the middle section were treated either as static or as dynamic rings, giving rise to either a static or a dynamic nanotube. The significant result was that the dynamic nanotube slowed down the fluid particles faster than the static nanotube. It took about 42 ps for the fluid velocity to slow down to 2.5% of its initial velocity in the static nanotube, while in the dynamic nanotube this time was 15 ps. Furthermore, the motion of the nanotube perturbed the motion of the adjacent fluid atoms, causing faster randomization and promoting hard collisions with the walls that slowed down the fluid steadily until it essentially came to a stop. Similar results were also obtained for the Ar fluid, whose atoms are some 10 times heaver than He atoms. In this case, it was found, however, that the velocity damping was even more pronounced. It was also found that the higher the fluid density, the faster it slowed down.

Another MD simulation has addressed the rapid imbibition of oil (decane) in a (13,13) SWCNT, at temperature $T = 298$ K [58], modeling the SWCNT as a rigid cylinder. The computed fluid density ρ inside the nanotube showed that the nanotube first filled rapidly with the low density fluid, followed by the higher density fluid at lower speeds tending to ≈ 150 m/s for $\rho > 0.2$. Such speeds are comparable to the speed of sound in air (340 m/s). It was found that, the flow terminated at the end of the nanotube, i. e., although the nanotube was open, its end acted as a barrier to further flow. It was also found that the imbibition was much faster than predicted by the classical Washburn equation which relates the penetration length of the fluid, in a macroscopic tube, to its radius, the surface tension, the viscosity and the time.

The effect of confinement of liquid water in a nanotube on its vibrational and rotational spectra has been also been investigated via MD simulations [30] for four different-sized nanotubes (6,6), (8,8), (10,10) and (12,12). The liquid water density was an input. The significant aspect of the results was the presence of a frequency band, between 3640 and 3690 cm^{-1}, which is absent in bulk water. This was referred to as a vibration frequency, and its position shifted to smaller values when the nanotube radius increased. It was concluded that this *vibration band was purely an effect of confinement*. Adsorption line shape in the domain between 300 and 1000 cm^{-1}, that corresponds to the rotational motion of the molecules, was also obtained, and again a shift to lower frequencies was observed compared with those in the bulk. Hence, there was a direct connection between the radius of the nanotube and some of the observed frequency shifts.

Transport of water through a solvated carbon nanotube has also been modeled [19]. The central channel of the nanotube is strongly hydrophobic in character. It was, however, found that the initially empty channel rapidly filled up with water from the surrounding container, and was occupied during the entire simulation run. The water density inside the cylinder was found to exceed the bulk density. Since the water molecules interact weakly with the carbon atoms, the continuous hydration of the nanotube interior was unexpected.

Modeling Gas Storage

Gas storage can take place in single nanotubes, as well as in their bundles, where three adsorption sites are available, namely the interstitial channels between the SWCNTs, the outer surfaces of the nanotubes in the bundle, and the groove channels, i. e., the wedge-shaped spaces on the outer surface of the bundle where two SWCNTs meet [60]. Simulation results indicate that while H_2, He and Ne atoms can adsorb in the interstitial channels, other types of atoms are too large to fit into these spaces. The clarification of adsorption sites and the adsorbed amount has been the focal point of research in this area.

Hydrogen Storage The storage of atomic and molecular hydrogen has occupied a very prominent position in the field of gas adsorption in nanotubes. The adsorption of para-hydrogen in (9,9) and (18,18) SWCNTs and their bundles has been modeled at several temperatures [62]. The inter-nanotube spacing within a bundle, called the van der Waals gap, is approximately 3.2 Å, measured from the center of the nanotube walls. The gap g is defined as ($g = a - D$), where a is the lattice spacing, and D is the diameter of the nanotube. The total storage in the bundle is a sum of interior and interstitial adsorptions. The computed isotherms at $T = 77$ K show that, in the low pres-

sure range, the (9,9) nanotube bundle gives a higher gravimetric and volumetric densities than the (18,18) nanotube bundle, since the hydrogen-nanotube interaction in the former case is stronger. At pressures above 5 and 10 atm, the situation is reversed. The volume of a (9,9) nanotube is of such a size that it can hold one layer of adsorbed molecules on the interior surface and a column of molecules in the center of the nanotube, whereas the volume of the (18,18) nanotube can accommodate three layers of hydrogen in three concentric rings and a column in the center of the nanotube.

It was found that the interstitial adsorption formed 14% of the total adsorption in the (18,18) nanotube bundle, in the (9,9) nanotube bundle this proportion was less than 1%. Increasing the temperature from $T = 77$ to 298 K lowered the adsorbed amount in both the (9,9) and the (18,18) bundles by approximately a factor of 5 at 100 atm. The computed adsorption isotherms on the external and internal surfaces of an isolated (18,18) SWCNT showed that it adsorbed slightly more than the (9,9) SWCNT at $T = 77$ K, except at the lowest pressures. A significant portion of the total adsorption took place on the external surfaces.

Quantum-mechanical based computation of adsorption of hydrogen atoms in (5,5) and (10,10) SWCNTs and in a (5,5)@(10,10) DWCNT has also been performed [27]. The results show that there were two chemisorption sites on the exterior and the interior of the nanotube, and that a form of H_2 *molecule* formed within the empty pore inside the nanotube. Furthermore, hydrogen capacity increased linearly with the diameter of the nanotube. The results for the DWCNT showed that the SWCNTs were better for higher hydrogen storage than the DWCNT.

Another study on the physisorption of *molecular* hydrogen H_2 in SWCNT bundles [64] has shown that small-diameter bundles are preferable for storage, and that the delamination of the nanotube bundle increases the gravimetric storage capacity.

In another simulation [63], the question of optimization of the gap in a bundle of SWCNTs for maximum storage of H_2 has been addressed for two bundle geometries, namely, a square bundle and a triangular bundle, each bundle composed of four nanotubes of types (9,9), (12,12), (18,18). The gap was varied to obtain the optimum separation for adsorption. The results on the gravimetric and volumetric densities, as a function of the gap, for $T = 298$ K and $P = 50$ atm showed that for $g = 3.2$ Å, i. e., the smallest gap, much of the volume and surface area in the bundles were unavailable for adsorption in both packing geometries. Increasing this gap allowed adsorption to take place on the external surface of the nanotubes. It was found that, the optimum value of g, as a function of temperature, was $g = 6$ Å at $T = 298$ K, and $g = 9$ Å at $T = 77$ K in a *triangular* bundle of the (9,9) nanotubes.

The adsorption of H_2 molecules in *charged* (9,9) SWCNTs, and their two-dimensional rhombic-shape bundles, has also been investigated [53]. The results for isolated SWCNTs showed more second layer adsorptions in the charged nanotubes, as compared with the uncharged nanotubes, and that the difference in adsorption was about 10%–20% at $T = 298$ K and 15%–30% at $T = 77$ K. Furthermore, it was noticed that the negatively charged nanotubes adsorbed more than the positively charged nanotubes at $T = 77$ K. At the higher temperature of $T = 298$ K, there was no observable difference. Volumetric densities for the adsorption in the bundles of charged and neutral nanotubes, as a function of the van der Waals gap, showed that the charged bundle had an enhanced adsorption of H_2 because of charge-multipole interactions. The results from these simulations imply that an enhancement of the hydrogen storage capacity of nanotubes can be achieved by using electric fields and gradients.

Adsorption of Other Gases The quantum ground-state energy of He atoms, as well as the isosteric specific heat of low-density He gas, adsorbed on the external groove sites in a bundle of 37 nanotubes have been computed [52]. The energy was found to be -22.7 meV, in close agreement with the experimental value, but larger than the computed energy for the He atom adsorbed in the interstitial channels, which is -27.9 meV. The data on specific heat for adsorption in the grooves were computed, and were seen to be very different from the data for adsorption in the interstitial positions.

The adsorption of Xe atoms in (10,10) SWCNTs, and their hexagonal bundle, at $T = 95$ K was modeled [54]. The results showed that the rather strong interaction between highly polarizable Xe atoms and the nanotubes led to significant adsorption even at very low pressures. The adsorption isotherms for internal and external surfaces showed a significant difference. Adsorption in the bundle of SWCNTs was also considered, and as the interstitial space was only 2.9 Å wide, for a large bundle, adsorption was expected to take place only in the interior of the nanotubes.

Adsorption of NO_2, O_2, NH_3, CO_2, H_2O gas molecules on (10,0), (17,0), (5,5) and (10,10) SWCNTs has also been obtained [66]. Different adsorption sites, such as the top of a carbon atom, the top of the center of a C-C bond, and the top of a carbon hexagon, were considered. The results showed that, in general most of the gases were weakly bound to the nanotubes, i. e., they were physisorbed. Fur-

thermore, most of the molecules were charge donors with a small charge transfer, whereas O_2 and NO_2 were charge acceptors. The influence of adsorption on the electronic properties of SWCNTs was also investigated, and it was found that the adsorption of NH_3 and NO_2 did not significantly affect either the conduction or the valence band of a (10,0) nanotube. The density of states (DOS) of the nanotube decorated with NH_3 was very close to that of a pure nanotube. This behavior was also observed for all charge donor molecules, i. e., N_2, H_2O, CO_2 etc. The adsorption of NO_2 significantly changed the shape of DOS, implying that semi-conducting SWCNTs can be turned into p-type conductors following adsorption of NO_2 or O_2.

Another study [12] has considered the influence of lattice dilation, due to adsorption of He, Ne, H_2, D_2, Ar and CH_4 gases in the interstitial channels of the SWCNT triangular bundles, on the further uptake of these gases. The basic idea was that, lattice dilation allows small molecules to significantly enhance their binding energies in the interstitial channels without a substantial increase in the inter-nanotube interaction energy. The results revealed that the chemical potential for H_2 was about 200 K lower than the value in the absence of dilation, i. e., there was a greater tendency for the adsorption in a dilated medium than in an undilated one. In the case of D_2, the lattice dilated less than in the case of H_2. Also the dilation in the He and Ne adsorptions was less than 0.5%. For H_2, Ar and CH_4 the consequences of dilation were significant.

Adsorption of Gases in SWCNH Assemblies Since SWCNHs are closed structures, access to their interior must first be achieved by opening potential entry points. The most common method for pore opening is heat treatment in oxygen [3], and oxidization at $T = 573$ and 623 K respectively leads to the opening of 11% and 36% of the spaces. Assemblies of SWCNHs also possess interstitial channels.

Adsorption of nitrogen, at $T = 77$ K was used to estimate the volumetric porosity of aggregates of heat-treated SWCNHs [32]. The interaction of the corn (apex) parts of the horns was not taken into account, and only the contribution of tube parts was considered. It was found that the interstitial site in the middle of three SWCNHs provided the strongest site. However, since only a one-dimensional array of N_2 molecules could be packed at this site the adsorption capacity was limited there. Strong adsorption could also occur on the sites within the SWCNHs near their walls, leading to the formation of a monolayer. A third site, of weaker type, also located in the interior of SWCNHs, was also identified. Therefore, three sites were found for adsorption.

Adsorption of N_2 in the internal pore space of *individual* SWCNHs has been addressed [35], with the interaction of the corn part of the SWCNH included. Adsorption isotherms were obtained for different diameters D of the tube part of the SWCNH, and since the SWCNHs were partially oxidized at $T = 693$ K, almost all of the pore spaces were available for adsorption. The computed isotherms showed two gradual steps around the normalized pressures of 10^{-5} and 10^{-1}, and the smaller the value of D, the shaper was the step. The average pore width w of the internal space was found to be $w = 29$ Å. Experimental images of three regions, namely the tip, the neck and the tube parts of the SWCNHs showed that the adsorption in these regions was considerably different, and that it began at the tip and then moved to the neck space, and in the tube space the molecules were adsorbed on the internal wall in multilayers.

Modeling the Mechanical Properties of Nanotubes

Modeling the mechanical properties of carbon nanotubes occupies a central position in research on nanotubes, and computational simulation in this field has employed highly sophisticated atomistic and continuum-elasticity models. These studies reveal that nanotubes have high tensile strength, large bending flexibility and high aspect ratios, properties that make them an ideal superstrong material. Research has focused on the elastic constants, Poisson's ratio and Young's modulus, and their dependence on the diameter and chirality of the nanotubes. A very interesting part of this research has used well-established continuum-based theories of curved plates, thin shells, bending beams and vibrating rods. An extensive presentation of these topics, as applied to nanotubes, can be found in [44]. These theories have been used either as independent computational tools to model the elastic properties and deformation modes of nanotubes, or in conjunction with the atomistic models to interpret the results obtained from them. *A key issue in this respect is the identification of the total energy of an atomistic system with the elastic strain energy of an equivalent continuum system.*

Some research problems pose a challenge to both types of the modeling. Large MWCNTs are a case in point, as their sizes present a challenge to MD simulations, and the continuum mechanics is most successful in the limit of a thin-shell. A very fruitful approach would be a *multiscale* modeling that couples the continuum, the atomistic and the quantum description of the nanotubes within a unified model, and until such models are developed, a pragmatic attitude would be the application of both

atomistic and continuum modeling to extract as much useful information as possible.

Structural Deformations of Nanotubes

Modeling the axial compression, bending and torsion modes of a (7,7) SWCNT has been performed via both MD simulations, and the continuum elasticity theory of plates [65]. Let us consider the axial compression modeling, whereby a compression was applied to the ends of the nanotube by shifting the end atoms along the axis in small steps, and then relaxing the nanotube while keeping each end constrained. The variation of the strain with the strain energy showed a set of discontinuities that were associated with the rather exotic shape changes undergone by the nanotube. The evolution of this deformation mode within the continuum theory was examined via the expression for the deformation energy of a curved plate. Using the MD generated information as input, the curved plate parameters, i. e., its flexural rigidity and in-plane stiffness, that appear in the continuum energy expression were obtained. It was found that at a critical level of imposed strain, and employing the plate parameters, the nanotube buckled sideways, very similar to what happened in the MD simulation, and that the lowest value of the strain was close to that obtained in the MD simulation. The overall results from all the simulations showed that nanotubes could sustain extreme strains without developing any brittle or plastic transitions or dislocations, and beyond the range covered by Hooke's law, their behavior can be described by continuum-based modeling.

The mechanism whereby uniaxially strained (5,5) and (10,10) nanotubes release their stored strain energy has also been investigated [11]. The nanotube was subjected to a 10% tensile strain, and following equilibration at 1800 K, the first signs of mechanical yielding appeared. This yield was accompanied by the formation of *topological* defects, beyond a critical tension. The first defect to appear was a 90° rotation of a C–C bond about its center which transformed four hexagons into a double pentagon-heptagon (5-7-7-5) defect. This bond rotation is referred to as the Stone–Wales transformation. The energy barriers to the bond rotation that produced the defect were calculated, and it was found that for both types of nanotubes it decreased with an increase in the strain. Furthermore, the results showed that by annealing the (5,5) nanotube, subject to 10% strain, under the condition of high temperature ($T = 2500$ K), the (5-7-7-5) defect was reversible and the nanotube recovered its original hexagonal network.

In a further simulation [10], a (10,10) nanotube containing an initial (5-7-7-5) defect was strained up to 15%, and it was observed that octagonal defects and higher order rings appeared, i. e., the initial defect had acted as a nucleation center for other defects. The simulation showed that armchair nanotubes could display a ductile behavior via plastic flow of the (5-7) dislocation cores. These topological defects could change the chirality of the nanotube, and when dislocations covered a sizable distance of the nanotube wall, they left behind a nanotube whose chirality changed according to the dislocation rules, i. e., $(n, n) \rightarrow (n, n-1) \rightarrow (n, n-2) \ldots$, i. e., an armchair nanotube transformed into a zigzag geometry.

The effect of hydrogen storage on the deformation properties of (10,10) and (17,0) SWCNTs has also been investigated [67]. The SWCNTs were pre-stored with 4.17 or 8.34 wt % H_2 gas, and the maximum strain and maximum tensile force as a function of hydrogen wt %, and temperature were obtained, showing that both the maximum strain and maximum force decreased with the storage of H_2, and that the reduction at $T = 600$ K was much bigger than that at $T = 300$ K. The reductions were attributed to the competition between the C–C bonds and the C–H bonds. The effect of storage on the (17,0) nanotube did not seem to be as significant as on the (10,10) nanotube.

Elastic Properties of Nanotubes

Elastic properties of nanotubes, i. e., their elastic constants, Young's, bulk and shear moduli, have been computed [28,29] for a number of SWCNTs and MWCNTs. An inspection of the results shows that the shear and bulk moduli of SWCNTs and MWCNTs of different diameters were respectively around 0.4 TPa and 0.75 TPa, and Young's modulus for these nanotubes was around 1 TPa, comparing well with the experimental value of 1.3 TPa. Comparison of the SWCNT and MWCNT results shows that:

1) the elastic moduli change very little with the number of walls,
2) the inter-shell van der Waals interaction does not affect significantly the elastic moduli of multi-walled nanotubes,
3) there is a large anisotropy in elastic properties of both SWCNTs and MWCNTs,
4) the elastic properties are basically the same for all nanotubes with radii greater than 1 nm.

An MD simulation [41] has also computed Young's modulus and Possion's ratio ν of the SWCNTs, giving

$$\begin{aligned} E &= 1.20\,\text{TPa}\,, \quad \nu = 0.27 \text{ for a (10,10) nanotube}\,, \\ E &= 1.10\,\text{TPa}\,, \quad \nu = 0.28 \text{ for a (5,5) nanotube}\,, \end{aligned} \tag{19}$$

Another MD simulation [24], using an energy approach (EA) and a force approach (FA), has computed the axial E_z and the rotational E_θ Young's moduli of a set of (n, n) SWCNTs, for $n = 6$ to 19, and has found that E_z and E_θ had little dependence on the radius of the nanotubes in both approaches, and that both moduli were very close. The average values for different nanotubes in both approaches were

$$E_z = E_\theta = \begin{cases} 1.344 \pm 0.02 & \text{TPa (EA)} , \\ 1.238 \pm 0.01 & \text{TPa (FA)} . \end{cases} \tag{20}$$

The bulk moduli of (6,6), (10,10) and (8,4) SWCNT hexagonal bundles subject to hydrostatic pressure have also been computed [47], giving the bulk modulus of the bundle K_{bund}, and the individual nanotubes K_{tube} as

$$\begin{aligned} K_{\text{bund}} &= 37\,\text{GPa} , \\ K_{\text{tube}} &= 230\,\text{GPa} . \end{aligned} \tag{21}$$

Modeling the Thermal Properties of Nanotubes

Thermal stability of nanotubes is crucial to the safe operation of nano-scale devices. Measurements of the specific heat and thermal conductivity of microscopic *mats*, covered with compressed ropes of nanotubes have been made, providing information on the ensemble-average thermal properties of bulk-phase materials, rather than on individual nanotubes. Equilibrium and non-equilibrium types of MD simulations have also been applied. In the former case, the equilibrium time-correlation functions of the heat flux operator are computed and are used in the Green–Kubo relation to obtain the thermal conductivity. In the latter method, hot and cold reservoirs are coupled to the two ends of the nanotube, and the thermal conductivity is obtained via the average heat flux .

Thermal Conductivity of Nanotubes

In an MD simulation [4], combining both equilibrium and non-equilibrium methods, the temperature-dependence of the thermal conductivity λ of an isolated (10,10) nanotube has been computed, showing that at $T = 100$ K the nanotube displayed an unusually high value $\lambda = 37\,000$ W/m K, far exceeding that of pure diamond, i. e., $\lambda = 3320$ W/m K. Comparing the thermal conductivity of an isolated nanotube with those of an isolated graphene sheet and bulk graphite showed that the nanotube had a very similar heat transport properties as those of a hypothetical isolated graphene sheet. The results on the three carbon structures showed that below $T = 300$ K, the thermal conductivity of the graphene sheet was higher than that of the nanotube, and between $T = 300$ K and 400 K they were nearly equal. For graphite, the thermal conductivity of the basal plane also peaked near the $T = 100$ K point, and its value in the range $T = 200$ K to 400 K was always smaller than the values for other two structures.

Another MD simulation [13] investigated the dependence of λ on the nanotube length, and the presence of defects and vacancies, showing that after an initial rise in λ, up to a nanotube length of $L = 100$ Å, it converged to a fairly constant value $\lambda = 29\,800$ W/m K along the nanotube axis for L in the range 100 to 500 Å. This was similar to the value obtained by [4]. In the computations, the thickness of the nanotube was taken to be 1 Å. The results on the variation of λ with the vacancy concentration showed that it decreased smoothly from $\lambda = 14\,000$ W/m K for a concentration of 0.002 to $\lambda = 4000$ W/m K for a concentration of 0.01. Furthermore, the role of Stone–Wales (5-7-7-5) defects was examined, showing that λ decreased from $\lambda = 35000$ W/m K at concentration of 0.5 (on the 1/1000 scale) to about $\lambda = 12\,000$ W/m K for a concentration of 3. Both the rate of decrease and its absolute amount were less than the case with vacancies.

For a bundle of (10,10) nanotubes, it was found that $\lambda = 9500$ W/m K along the axis, very close to that of the in-plane bulk graphite at $\lambda = 10\,000$ W/m K, while perpendicular to the axis, $\lambda = 56$ W/m K, similar to the out-of-plane conductivity in graphite at $\lambda = 55$W/m K.

Specific Heat of Carbon Nanotubes

The low-temperature specific heats C_V of individual SWCNTs, and of their bundles composed of n SWCNTs of type (9,9), with $n = 1, 2, \ldots, 7$, and of MWCNTs have been computed [39], based on the phonon contribution to C_V. The results showed three different patterns for C_V of an individual SWCNT for $T < 100$ K, namely at very low temperatures,

$$C_V \propto T^{\frac{1}{2}} , \tag{22}$$

while with an increase in temperature,

$$C_V \propto T , \tag{23}$$

and above $T \approx 5$ K, the optical phonons began to make a contribution. Furthermore, it was found that, the $T^{\frac{1}{2}}$ behavior diminished when nanotubes were added to the bundle, and for a bundle composed of 7 nanotubes, this form of dependency on T was no longer observable and $C_V(T)$ approached the experimental value.

For MWCNTs, the specific heat of a five-walled nanotube, (5,5)@(10,10)@(15,15)@(20,20)@(25,25), was calculated, showing that by starting with the (5,5) nanotube and adding more shells to it, the $T^{\frac{1}{2}}$ part diminished and disappeared and was replaced by a linear dependence on T. If the number of shells were increased, C_V tended to that of graphite. It should be remarked that the variation of C_V with T for MWCNTs differed significantly from the experimental behavior.

Concluding Remarks

In this survey we have reviewed the modeling studies on the mechanical, thermal and transport properties of carbon nanotubes. We have seen that in modeling the fluid flow correctly, the motion of both the fluid and the walls, and their mutual coupling must be considered. In this field, the future research must address the use of such classical concepts as *viscosity* and *pressure* that are not clearly defined at nano-scales. Furthermore, flow-induced structural instability poses another important research problem.

In the storage of gases, we have seen that adsorption can take in both the internal spaces and in the interstitial channels, and that SWCNTs provided a better storage medium for hydrogen than the MWCNTs. Furthermore, the small-diameter ropes were more preferable than large-diameter ones. The future research should help clarify the question of volume storage capacity of nanotubes and the ways this capacity could be improved as the data obtained in different studies are not yet unitary. They indicate, however, that the adsorption in nanotubes is higher than in most carbonaceous materials.

In the studies of mechanical and elastic properties, both the atomistic and continuum theories have been used successfully in providing deep insights into the deformation and elastic properties of nanotubes. as well as providing quantitative estimates of the pertinent variables. Studies focusing on obtaining Young's modulus of the nanotubes have shown that the estimates obtained depended on the theoretical models or the inter-atomic potentials used. Computed values ranged from 1 TPa to 5.5 TPa for SWCNTs, and for the MWCNTs an average experimental value of about 1.8 TPa seemed to be the accepted result. Future research could employ better continuum-based models to help reduce the number of very time-consuming and costly MD simulations that are currently employed to obtain the estimates of mechanical and elastic properties.

Finally, we have seen that the thermal conductivity of the SWCNT as a function of temperature is unusually high, implying that the mean free path of phonons was quite large in nanotubes. The presence of defects and vacancies also substantially affected this conductivity. In the computation of the specific heat, the main finding concerned the significant contribution of phonons to the specific heat as compared with the electrons, and in the low temperature regime, several important scaling laws on the behavior of the specific heat have been derived. The future research could consider such problems as the computation of the thermal properties of nanotubes containing defects, or decorated with adsorbed molecules, or nano-scale devices composed of carbon nanotubes.

Bibliography

1. Ajayan PM, Ebbesen TW (1997) Nanometre-size tubes of carbon. Rep Prog Phys 60:1025–1062
2. Allen MP, Tildesley DJ (1987) Computer simulation of liquids. Clarendon Press, Oxford
3. Bekyarova E, Kaneko K, Kasuya D, Murata K, Yudasaka M, Iijima S (2002) Oxidation and porosity evaluation of budlike single-wall carbon nanohorn aggregates. Langmuir 18:4138–4141
4. Berber S, Kwon YK, Tománek D (2000) Unusually high thermal conductivity of carbon nanotubes. Phys Rev Lett 84:4613–4616
5. Berbholc J, Brenner D, Nardelli MB, Meunier V, Roland C (2002) Mechanical and electrical properties of nanotubes. Annu Rev Mater Res 32:347–75
6. Bethune DS, Kiang CH, de Vries MD, Gorman G, Savoy R, Vazquez J, Beyers R (1993) Cobalt-catalysed growth of carbon nanotubes with single-atomic-layer walls. Nature 363:605–607
7. Bhushan B (ed) (2004) Springer handbook of nanotechnology. Springer, Berlin
8. Binnig G, Quate CF, Gerber C (1986) Atomic force microscope. Phys Rev Lett 56:930–933
9. Binnig G, Rohrer H (1982) Scanning tunnelling microscopy. Helv Phys Acta 55:726–735
10. Buongiorno Nardelli M, Yakobson BI, Bernholc J (1998) Brittle and ductile behaviour in carbon nanotubes. Phys Rev Lett 81:4656–4659
11. Buongiorno Nardelli M, Yakobson BI, Bernholc J (1998) Mechanism of strain release in carbon nanotubes. Phys Rev B 57:R4277–R4280
12. Calbi MM, Toigo F, Cole MW (2001) Dilation-induced phases of gases adsorbed within a bundle of carbon nanotubes. Phys Rev Lett 86:5062–5065
13. Che J, Cagin T, Goddard WA III (2000) Thermal conductivity of carbon nanotubes. Nanotechnology 11:65–69
14. Dai H (2002) Carbon nanotubes: opportunities and challenges. Surf Sci 500:218–241
15. Dresselhaus MS, Dresselhaus G, Saito R (1995) Physics of Carbon Nanotubes. Carbon 33:883–891
16. Dresselhaus MS, Dresselhaus G, Saito R, Jorio A (2005) Raman spectroscopy of carbon nanotubes. Phys Rep 409:47–99
17. Drexler KE (1992) Nanosystems: Molecular Machinery, Manufacturing and Computation. Wiley, New York
18. Haile JM (1992) Molecular dynamics simulation: Elementary methods. Wiley, New York
19. Hummer G, Rasaiah JC, Noworyta JP (2001) Water conduction through the hydrophobic channel of a carbon nanotube. Nature 414:188–190

20. Iijima S (1991) Helical microtubules of graphitic carbon. Nature 354:56–58
21. Iijima S (2002) Carbon nanotubes: past, present, and future. Physica B 323:1–5
22. Iijima S, Ichihashi T (1993) Single-shell carbon nanotubes of 1-nm diameter. Nature 363:603–605
23. Iijima S, Yudasaka M, Yamada R, Bandow S, Suenaga K, Kokai F, Takahashi K (1999) Nano-aggregates of single-walled graphitic carbon nano-horns. Chem Phys Lett 309:165–170
24. Jin Y, Yuan G (2003) Simulation of elastic properties of single-walled carbon nanotubes. Compos Sci Technol 63:1507–1515
25. Kiang CH, Endo M, Ajayan PM, Dresselhaus G, Dresselhaus MS (1998) Size effects in carbon nanotubes. Phys Rev Lett 81:1869–1872
26. Lavin JG, Subramoney S, Ruoff RS, Berber S, Tománek D (2002) Scrolls and nested tubes in multiwall carbon nanotubes. Carbon 40:1123–1130
27. Lee SM, An KH, Lee YH, Seifert G, Frauencheim T (2001) A hydrogen storage mechanism in single-walled carbon nanotubes. J Am Chem Soc 123:5059–5063
28. Lu JP (1997) Elastic properties of carbon nanotubes and nanoropes. Phys Rev Lett 79:1297–1300
29. Lu JP (1997) Elastic properties of single and multilayered nanotbes. J Phys Chem Solids 58:1649–1652
30. Martí J, Gordillo MC (2001) Effects of confinement on the vibrational spectra of liquid water adsorbed in carbon nanotubes. Phys Rev B 63:165430-1–165430-5
31. Mizel A, Benedict LX, Cohen ML, Louie SG, Zettl A, Budraa NK, Beyermann WP (1999) Analysis of the low-temperature specific heat of multiwalled carbon nanotubes and carbon nanotube ropes. Phys Rev B 60:3264–3270
32. Murata K, Kaneko K, Steele WA, Kokai F, Takahashi K, Kasuya D, Hirahara K, Yudasaka M, Iijima S (2001) Molecular potential structures of heat-treated single-wall carbon nanohorn assemblies. J Phys Chem B 105:10210–10216
33. Murata K, Kaneko K, Steele WA, Kokai F, Takahashi K, Kasuya D, Hirahara K, Yudasaka M, Iijima S (2001) Porosity evaluation of intrinsic intraparticle nanopores of single wall carbon nanohorn. Nano Lett 1:197–199
34. Nalwa HS (ed) (2004) Encyclopedia of nanoscience and nanotechnology, vol 1–10. American Scientific Publishers, Cambridge
35. Ohba T, Murata K, Kaneko K, Steele WA, Kokai F, Takahashi K, Kasuya K, Yudasaka M, Iijima S (2001) N_2 adsorption in an internal nanopore space of single-walled carbon nanohorn: GCMC simulation and experiment. Nano Lett 1:371–373
36. Parr RG, Yang W (1989) Density functional theory of atoms and molecules. Oxford University Press, Oxford
37. Pathria RK (1972) Statistical mechanics. Pergamon, Oxford
38. Pekala K, Pekala M (1995) Low temperature transport properties of nanocrystalline Cu, Fe and Ni. NanoStructured Mater 6:819–822
39. Popov VN (2002) Low-temperature specific heat of nanotube systems. Phys Rev B 66:153408-1–153408-4
40. Popov VN (2004) Carbon nanotubes: properties and applications. Math Sci Eng R 43:61–102
41. Prylutskyy YI, Durov SS, Ogloblya OV, Buzaneva EV, Scharff P (2000) Molecular dynamics simulation of mechanical, vibrational and electronic properties of carbon nanotubes. Comp Math Sci 17:352–355
42. Rafii-Tabar H (2000) Modeling the nano-scale phenomena in condensed matter physics via computer-based numerical simulations. Phys Rep 325:239–310
43. Rafii-Tabar H (2004) Computational modeling of thermo-mechanical and transport properties of carbon nanotubes. Phys Rep 390:235–452
44. Rafii-Tabar H (2008) Computational physics of carbon nanotubes. Cambridge University Press, Cambridge
45. Rao CNR, Satishkumar BC, Govindaraj A, Nath M (2001) Nanotubes. Chem Phys Chem 2:78–105
46. Rapaport DC (1995) The art of molecular dynamics simulation. Cambridge University Press, Cambridge
47. Reich S, Thomsen C, Ordejón P (2002) Elastic properties of carbon nanotubes under hydrostatic pressure. Phys Rev B 65:153407-1–153407-4
48. Rieth M, Schommers W (ed) (2007) Handbook of theoretical and computational nanotechnology, vol 1–10. American Scientific Publishers, Cambridge
49. Roco MC, Bainbridge WS (2002) Convergent technologies for improving human performance. NSF/DOC-sponsored Report. WTEC Inc, Arlington
50. Saito Y, Yoshikawa T, Bandow S, Tomita M, Hayashi T (1993) Interlayer spacings in carbon nanotubes. Phys Rev B 48:1907–1909
51. Schwarz JA, Contescu CI, Putyera K (ed) (2004) Dekker encyclopedia of nanoscience and nanotechnology, vol 1–5. Marcel Dekker Inc, New York
52. Siber A (2002) Adsorption of He atoms in external grooves of single-walled carbon nanotube bundles. Phys Rev B 66:205406-1–205406-6
53. Simonyan VV, Diep P, Johnson JK (1999) Molecular simulation of hydrogen adsorption in charged single-walled carbon nanotubes. J Chem Phys 111:9778–9783
54. Simonyan VV, Johnson JK, Kuznetsova A, Yates Jr JT (2001) Molecular simulation of xenon adsorption on single-walled carbon nanotubes. J Chem Phys 114:4180–4185
55. Smith BW, Monthoux M, Luzzi DE (1998) Encapsulated C_{60} in carbon nanotubes. Nature 396:323–324
56. Stevens RMD, Frederick NA, Smith BL, Morse DE, Stucky GD, Hansma PK (2000) Carbon nanotubes as probes for atomic force microscopy. Nanotechnology 11:1–5
57. Stojkovic D, Zhang P, Crespi VH (2001) Smallest nanotube: breaking the symmetry of sp^3 bonds in tubular geometries. Phys Rev Lett 87:125502–125505
58. Supple S, Quirke N (2003) Rapid imbibition of fluids in carbon nanotubes. Phys Rev Lett 90:214501-1–214501-4
59. Sutton AP, Pethica JB, Rafii-Tabar H, Nieminen JA (1994) Mechanical properties of metals at nanometre scale. In: Pettifor DG, Cottrell AH (eds) Electron theory in alloy design. Institute of Materials, London, pp 191–233
60. Talapatra S, Zambano AZ, Weber SE, Migone AD (2000) Gases do not adsorb on the interstitial channels of closed-ended single-walled carbon nanotube bundles. Phys Rev Lett 85:138–141
61. Tuzun RE, Noid DW, Sumpter BG, Merkle RC (1996) Dynamics of fluid flow inside carbon nanotubes. Nanotechnology 7:241–246
62. Wang Q, Johnson JK (1999) Molecular simulation of hydrogen adsorption in single-walled carbon nanotubes and idealised carbon slit pore. J Chem Phys 110:577–586
63. Wang Q, Johnson JK (1999) Optimisation of carbon nanotube arrays for hydrogen adsorption. J Phys Chem B 103:4809–4813

64. Williams KA, Eklund PC (2000) Monte Carlo simulation of H_2 physisorption in finite-diameter carbon nanotube ropes. Chem Phys Lett 320:352–358
65. Yakobson BI, Brabec CJ, Bernholc J (1996) Nanomechanics of carbon tubes: instabilities beyond linear response. Phys Rev Lett 76:2511–2514
66. Zhao J, Buldum A, Han J, Lu JP (2002) Gas molecule adsorption in carbon nanotubes and nanotube bundles. Nanotechnology 13:195–200
67. Zhou LG, Shi SQ (2002) Molecular dynamic simulations on tensile mechanical properties of single-walled carbon nanotubes with and without hydrogen storage. Comp Math Sci 23:166–174

Catastrophe Theory

Werner Sanns
University of Applied Sciences, Darmstadt, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Example 1: The Eccentric Cylinder on the Inclined Plane
Example 2: The Formation of Traffic Jam
Unfoldings
The Seven Elementary Catastrophes
The Geometry of the Fold and the Cusp
Further Applications
Future Directions
Bibliography

Glossary

Singularity Let $f\colon \mathbb{R}^n \to \mathbb{R}^m$ be a differentiable map defined in some open neighborhood of the point $p \in \mathbb{R}^n$ and $J_p f$ its Jacobian matrix at p, consisting of the partial derivatives of all components of f with respect to all variables. f is called singular in p if rank $J_p f < \min\{n, m\}$. If rank $J_p f = \min\{m, n\}$, then f is called regular in p. For $m = 1$ (we call the map $f\colon \mathbb{R}^n \to \mathbb{R}$ a differentiable function) the definition implies: A differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$ is singular in $p \in \mathbb{R}^n$, if grad $f(p) = 0$. The point p, where the function is singular, is called a singularity of the function. Often the name "singularity" is used for the function itself if it is singular at a point p. A point where a function is singular is also called critical point. A critical point p of a function f is called a degenerated critical point if the Hessian (a quadratic matrix containing the second order partial derivatives) is singular at p; that means its determinant is zero at p.

Diffeomorphism A diffeomorphism is a bijective differentiable map between open sets of $\mathbb{R}^n$ whose inverse is differentiable, too.

Map germ Two continuous maps $f\colon U \to \mathbb{R}^k$ and $g\colon V \to \mathbb{R}^k$, defined on neighborhoods U and V of $p \in \mathbb{R}^n$, are called equivalent as germs at p if there exists a neighborhood $W \subset U \cap V$ on which both coincide. Maps or functions, respectively, that are equivalent as germs can be considered to be equal regarding local features. The equivalence classes of this equivalence relation are called germs. The set of all germs of differentiable maps $\mathbb{R}^n \to \mathbb{R}^k$ at a point p is named $\varepsilon_p(n, k)$. If p is the origin of $\mathbb{R}^n$, one simply writes $\varepsilon(n, k)$ instead of $\varepsilon_0(n, k)$. Further, if $k = 1$, we write $\varepsilon(n)$ instead of $\varepsilon(n, 1)$ and speak of function germs (also simplified as "germs") at the origin of $\mathbb{R}^n$. $\varepsilon(n, k)$ is a vector space and $\varepsilon(n)$ is an algebra, that is, a vector space with a structure of a ring. The ring $\varepsilon(n)$ contains a unique maximal ideal $\mu(n) = \{f \in \varepsilon(n) | f(0) = 0\}$. The ideal $\mu(n)$ is generated by the germs of the coordinate functions $x_1, \ldots, x_n$. We use the form $\mu(n) = \langle x_1, \ldots, x_n \rangle \varepsilon(n)$ to emphasize, that this ideal is generated over the ring $\varepsilon(n)$. So a function germ in $\mu(n)$ is of the form $\sum_{i=1}^n a_i(x) \cdot x_i$, with certain function germs $a_i(x) \in \varepsilon(n)$.

r-Equivalence Two function germs $f, g \in \varepsilon(n)$ are called r-equivalent if there exists a germ of a local diffeomorphism $h\colon \mathbb{R}^n \to \mathbb{R}^n$ at the origin, such that $g = f \circ h$.

Unfolding An unfolding of a differentiable function germ $f \in \mu(n)$ is a germ $F \in \mu(n + r)$ with $F|\mathbb{R}^n = f$ (here $|$ means the restriction. The number r is called an unfolding dimension of f. An unfolding F of a germ f is called universal if every other unfolding of f can be received by suitable coordinate transformations, "morphisms of unfoldings" from F, and the number of unfolding parameters of F is minimal (see "codimension").

Unfolding morphism Suppose $f \in \varepsilon(n)$ and $F\colon \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}$ and $G\colon \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}$ be unfoldings of f. A right-morphism from F to G, also called unfolding morphism, is a pair (Φ, α), with $\Phi \in \varepsilon(n + k, n + r)$ and $\alpha \in \mu(k)$, such that:

1. $\Phi|\mathbb{R}^n = \mathrm{id}(\mathbb{R}^n)$, that is $\Phi(x, 0) = (x, 0)$,
2. If $\Phi = (\phi, \psi)$, with $\phi \in \varepsilon(n + k, n)$, $\psi \in \varepsilon(n + k, r)$, then $\psi \in \varepsilon(k, r)$,
3. For all $(x, u) \in \mathbb{R}^n \times \mathbb{R}^k$ we get $F(x, u) = G(\Phi(x, u)) + \alpha(u)$.

Catastrophe A catastrophe is a universal unfolding of a singular function germ. The singular function germs are called organization centers of the catastrophes.

Codimension The codimension of a singularity f is given by $\operatorname{codim}(f) = \dim_{\mathbb{R}} \mu(n)/\langle \partial_x f \rangle$ (quotient space). Here $\langle \partial_x f \rangle$ is the Jacobian ideal generated by the partial derivatives of f and $\mu(n) = \{f \in \varepsilon(n) | f(0) = 0\}$. The codimension of a singularity gives of the minimal number of unfolding parameters needed for the universal unfolding of the singularity.

Potential function Let $f: \mathbb{R}^n \to \mathbb{R}^n$ be a differentiable map (that is, a differentiable vector field). If there exists a function $\varphi: \mathbb{R}^n \to \mathbb{R}$ with the property that $\operatorname{grad} \varphi = f$, then f is called a gradient vector field and φ is called a potential function of f.

Definition of the Subject

Catastrophe theory is concerned with the mathematical modeling of sudden changes – so called "catastrophes" – in the behavior of natural systems, which can appear as a consequence of continuous changes of the system parameters. While in common speech the word catastrophe has a negative connotation, in mathematics it is neutral.

You can approach catastrophe theory from the point of view of differentiable maps or from the point of view of dynamical systems, that is, differential equations. We use the first case, where the theory is developed in the mathematical language of maps and functions (maps with range $\mathbb{R}$). We are interested in those points of the domain of differentiable functions where their gradient vanishes. Such points are called the "singularities" of the differentiable functions.

Assume that a system's behavior can be described by a potential function (see Sect. "Glossary"). Then the singularities of this function characterize the equilibrium points of the system under consideration. Catastrophe theory tries to describe the behavior of systems by local properties of corresponding potentials. We are interested in local phenomena and want to find out the qualitative behavior of the system independent of its size.

An important step is the classification of catastrophe potentials that occur in different situations. Can we find any common properties and unifying categories for these catastrophe potentials? It seems that it might be impossible to establish any reasonable criteria out of the many different natural processes and their possible catastrophes. One of the merits of catastrophe theory is the mathematical classification of simple catastrophes where the model does not depend on too many parameters.

Classification is only one of the mathematical aspects of catastrophe theory. Another is stability. The stable states of natural systems are the ones that we can observe over a longer period of time. But the stable states of a system, which can be described by potential functions and their singularities, can become unstable if the potentials are changed by perturbations. So stability problems in nature lead to mathematical questions concerning the stability of the potential functions.

Many mathematical questions arise in catastrophe theory, but there are also other kinds of interesting problems, for example, historical themes, didactical questions and even social or political ones. How did catastrophe theory come up? How can we teach catastrophe theory to the students at our universities? How can we make it understandable to non-mathematicians? What can people learn from this kind of mathematics? What are its consequences or insights for our lives?

Let us first have a short look at mathematical history. When students begin to study a new mathematical field, it is always helpful to learn about its origin in order to get a good historical background and to get an overview of the dependencies of inner-mathematical themes. Catastrophe theory can be thought of as a link between classical analysis, dynamical systems, differential topology (including singularity theory), modern bifurcation theory and the theory of complex systems. It was founded by the French mathematician René Thom (1923–2002) in the sixties of the last century. The name 'catastrophe theory' is used for a combination of singularity theory and its applications. In the year 1972 Thom's famous book "Stabilité structurelle et morphogénèse" appeared. Thom's predecessors include Marston Morse, who developed his theory of the singularities of differentiable functions (Morse theory) in the thirties, Hassler Whitney, who extended this theory in the fifties to singularities of differentiable maps of the plane into the plane, and John Mather (1960), who introduced algebra, especially the theory of ideals of differentiable functions, into singularity theory. Successors to Thom include Christopher Zeeman (applications of catastrophe theory to physical systems), Martin Golubitsky (bifurcation theory), John Guckenheimer (caustics and bifurcation theory), David Schaeffer (shock waves) and Gordon Wassermann (stability of unfoldings). From the point of view of dynamical systems, the forerunners of Thom are Jules Henry Poincaré (1854–1912) who looked at stability problems of dynamical systems, especially the problems of celestial mechanics, Andrei Nikolaevic **K**olmogorov, Vladimir I. **A**rnold and Jürgen **M**oser (KAM), who influenced catastrophe theory with their works on stability problems (KAM-theorem).

From the didactical point of view, there are two main positions for courses in catastrophe theory at university level: Trying to teach the theory as a perfect axiomatic system consisting of exact definitions, theorems and proofs

or trying to teach mathematics as it can be developed from historical or from natural problems (see [9]). In my opinion the latter approach has a more lasting effect, so there is a need to think about simple examples that lead to the fundamental ideas of catastrophe theory. These examples may serve to develop the theory in a way that starts with intuitive ideas and goes forward with increasing mathematical precision.

When students are becoming acquainted with catastrophe theory, a useful learning tool is the insight that continuous changes in influencing system parameters can lead to catastrophes in the system behavior. This phenomenon occurs in many systems. Think of the climate on planet earth, think of conflicts between people or states. Thus, they learn that catastrophe theory is not only a job for mathematical specialists, but also a matter of importance for leading politicians and persons responsible for guiding the economy.

Introduction

Since catastrophe theory is concerned with differentiable functions and their singularities, it is a good idea to start with the following simple experiment: Take a piece of wire, about half a meter in length, and form it as a 'curve' as shown in Fig. 1. Take a ring or screw-nut and thread the wire through it so that the object can move along the wire easily.

The wire represents the model of the graph of a potential function, which determines the reactions of a system by its gradient. This means the ring's horizontal position represents the actual value of a system parameter x that characterizes the state of the system, for example, x might be its position, rotation angle, temperature, and so on. This parameter x is also called the state parameter. The form of the function graph determines the behavior of the system, which tries to find a state corresponding to a minimum of the function (a zero of the gradient of f) or to stay in such a 'stable' position. (Move the ring to any position by hand and then let it move freely.) Maxima of the potential function are equilibrium points too, but they are not stable. A small disturbance makes the system leave such an equilibrium. But if the ring is in a local minimum, you may disturb its position by slightly removing it from that position. If this disturbance is not too big, the system will return to its original position at the minimum of the function when allowed to move freely. The minimum is a stable point of the function, in the sense that small disturbances of the system are corrected by the system's behavior, which tries to return to the stable point. Observe that you may disturb the system by changing its position by hand (the position of the ring) or by changing the form of the potential function (the form of the wire). If the form of the potential function is changed, this corresponds to the change of one or more parameters in the defining equation for the potential. These parameters are called external parameters. They are usually influenced by the experimenter. If the form of the potential (wire) is changed appropriately (for example, if you pull upward at one side of the wire), the ring can leave the first disappearing minimum and fall into the second stable position. This is the way catastrophes happen: changes of parameters influence the potentials and thus may cause the system to suddenly leave a stable position.

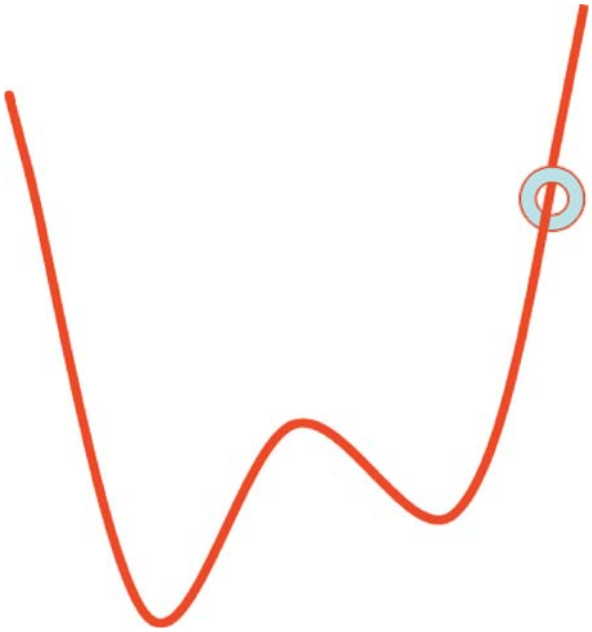

Catastrophe Theory, Figure 1
Ring on the wire

Some typical questions of catastrophe theory are:

- How can one find a potential function that describes the system under consideration and its catastrophic jumps?
- What does this potential look like locally? How can we describe the changes in the potential's parameters?
- Can we classify the potentials into simple categories?
- What insights can such a classification give us?

Now we want to consider two simple examples of systems where such sudden changes become observable. They are easy to describe and lead to the simplest forms of catastrophes which Thom listed in the late nineteen-sixties. The first is a simple physical model of an eccentric cylinder on

an inclined plane. The second example is a model for the formation of a traffic jam.

Example 1: The Eccentric Cylinder on the Inclined Plane

We perform some experiments with a cylinder or a wide wheel on an inclined plane whose center of gravity is eccentric with respect to its axes (Fig. 2).

You can build such a cylinder from wood or metal by taking two disks of about 10 cm in radius. Connect them with some sticks of about 5 cm in length. The eccentric center of gravity is generated by fixing a heavy piece of metal between the two discs at some distance from the axes of rotation. Draw some lines on the disks, where you can read or estimate the angle of rotation. (Hint: It could be interesting for students to experimentally determine the center of gravity of the construction.)

If mass and radius are fixed, there are two parameters which determine the system: The position of the cylinder can be specified by the angle x against a zero level (that is, against the horizontal plane; Fig. 4). This angle is the "inner" variable of the system, which means that it is the "answer" of the cylinder to the experimenter's changes of the inclination u of the plane, which is the "outer" parameter. In this example you should determine experimentally the equilibrium positions of the cylinder, that is, the angle of rotation where the cylinder stays at rest on the inclined plane (after releasing it cautiously and after letting the system level off a short time). We are interested in the stable positions, that is, the positions where the cylinder returns after slight disturbances (slightly nudging it or slightly changing the inclination of the plane) without rolling down the whole plane. The search for the equilibrium position x varies with the inclination angle u. You can make a table and a graphic showing the dependence of x and u.

The next step is to find an analytical approach. Refer to Fig. 3: Point P is the supporting point of the cylinder on the plane. Point A marks the starting position of the center of gravity. When the wheel moves and the distance between M and A is big enough, this center can come into two equilibrium positions S_1 and S_2, which lie vertically over P.

Now, what does the potential function look like? The variables are:

***R*:** radius of the discs,
***u*:** angle of the inclined plane against the horizontal plane measured in radians,
***x*:** angle of rotation of the cylinder measured against the horizontal plane in radians,
***r*:** distance between the center of gravity S of the cylinder from the axes marked by point M,
x_0 (resp. u_0): angles where the cylinder suddenly starts to roll down the plane.

Refer to Fig. 4. If the wheel is rolled upward by hand from the lower position on the right to the upper position, the center of gravity moves upward with the whole wheel, or

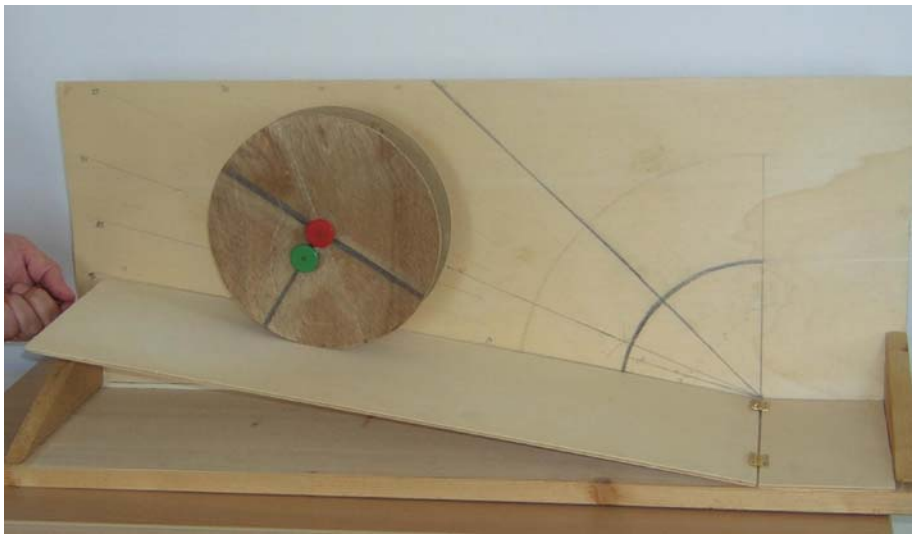

Catastrophe Theory, Figure 2
The eccentric cylinder on the inclined plane. The inclination of the plane can be adjusted by hand. The position of the disc center and the center of gravity are marked by colored buttons

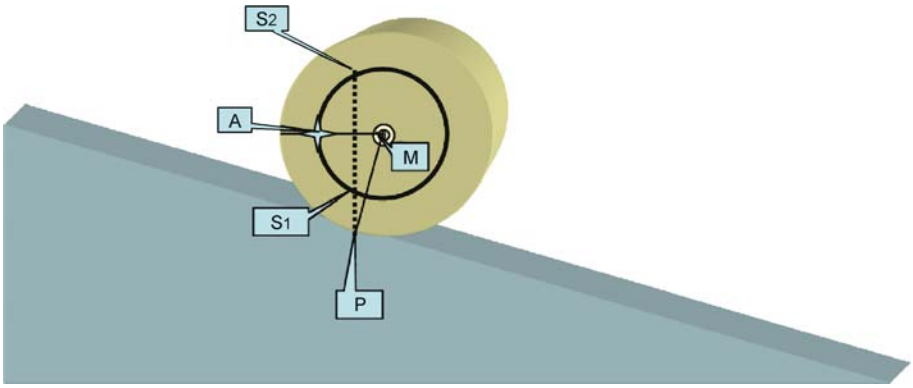

Catastrophe Theory, Figure 3
Two equilibrium positions (S_1 and S_2) are possible with this configuration. If the point A lies "too close" to M, there is no equilibrium position possible at all when releasing the wheel, since the center of gravity can never lie vertically over P

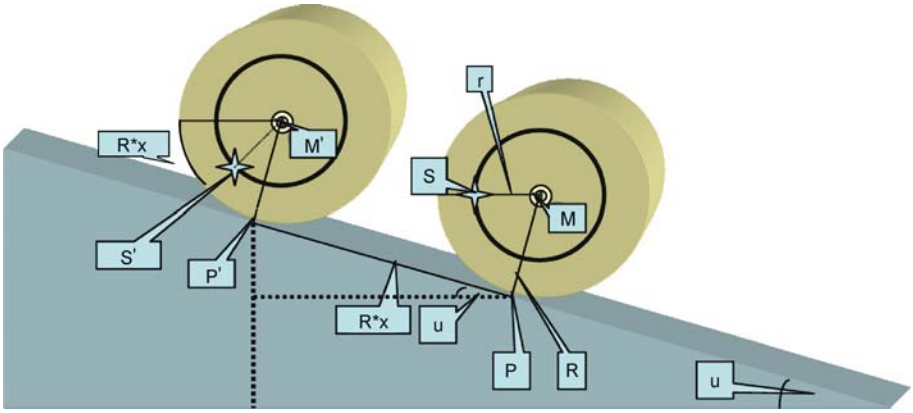

Catastrophe Theory, Figure 4
The movement of S is a composite of the movement of the cylinder represented by the movement of point M and the rotation movement with angle *x*

as the midpoint M or the point P. But point S also moves downward by rotation of the wheel with an amount of $r \cdot \sin(x)$.

The height gained is

$$h = R \cdot x \cdot \sin(u) - r \cdot \sin(x) \ . \tag{1}$$

From physics one knows the formula for the potential energy E_{pot} of a point mass. It is the product $m \cdot g \cdot h$, where $m =$ mass, $g =$ gravitational constant, $h =$ height, where the height is to be expressed by the parameters x and u and the given data of the cylinder.

To a fixed angle u_1 of the plane, by formula (1), there corresponds the following potential function:

$$f(x) = E_{\text{pot}} = m \cdot g \cdot h = m \cdot g \cdot (R \cdot x \cdot \sin(u_1) - r \cdot \sin(x)) \ .$$

If the inclination of the plane is variable, the angle u of inclination serves as an outer parameter and we obtain a whole family of potential functions as a function F in two variables x and u:

$$F(x, u) = F_u(x) = m \cdot g \cdot (R \cdot x \cdot \sin(u) - r \cdot \sin(x)) \ .$$

Thus, the behavior of the eccentric cylinder is described by a family of potential functions. The extrema of the single potentials characterize the equilibria of the system. The saddle point characterizes the place where the cylinder suddenly begins to roll down due to the slightest disturbance, that is, where the catastrophe begins (see Fig. 5).

Example. The data are $m = 1\,\text{kg}$, $g = 10\,\text{m/sec}^2$, $R = 0.1\,\text{m}$, $r = 0.025\,\text{m}$.

For constant angle of inclination u, each of the functions $f(x) = F(x, u = \text{const})$ is a section of the function F.

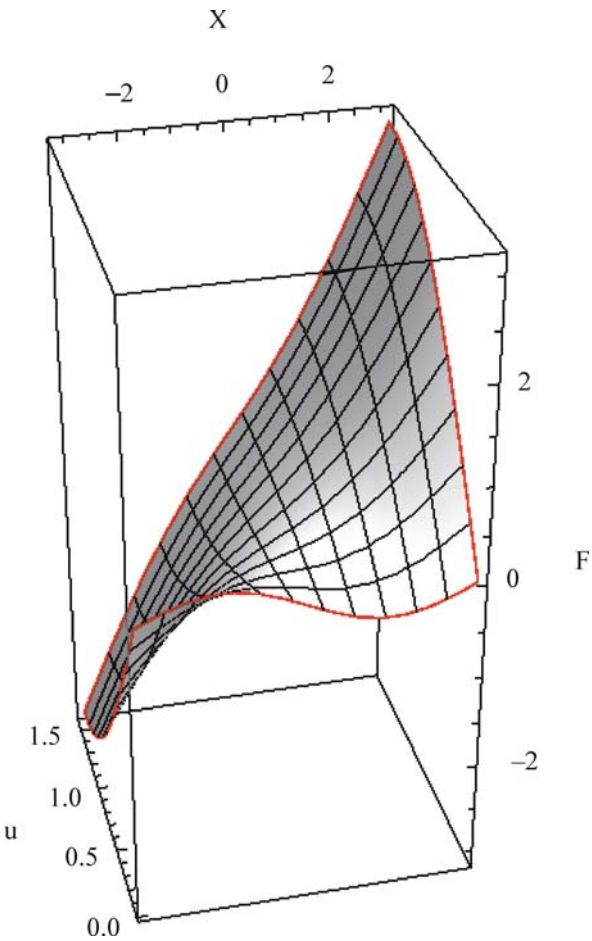

Catastrophe Theory, Figure 5
Graph of $F(x, u)$ for a concrete cylinder

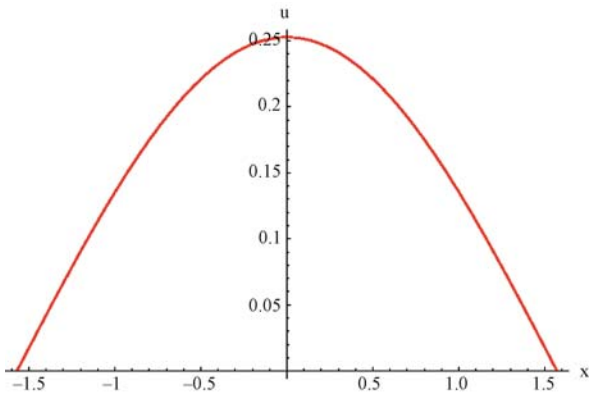

Catastrophe Theory, Figure 6
Curve in x–u-space, for which $F(x, u)$ has local extrema

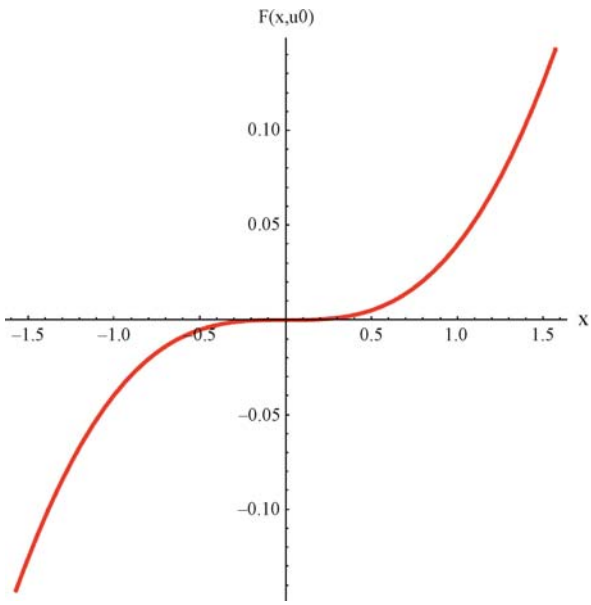

Catastrophe Theory, Figure 7
Graph of the section $F(x, u_0)$

To determine the minima of each of the section graphs, we calculate the zeros of the partial derivative of F with respect to the inner variable x. In our particular example above, the result is $\partial_x F = -0.25\cos(x) + \sin(u)$. The solution for $\partial_x F = 0$ is $u = \arcsin(0.25\cos(x))$ (see Fig. 6).

Calculating the point in x–u-space at which both partial derivatives $\partial_x F$ and $\partial_{xx} F$ vanish (saddle point of the section) and using only positive values of angle u gives the point $(x_0, u_0) = (0, 0.25268)$. In general: $x_0 = 0$ and $u_0 = \arcsin(r/R)$. In our example this means the catastrophe begins at an inclination angle of $u_0 = 0.25268$ (about 14.5 degrees). At that angle, sections of F with $u > u_0$ do not possess any minimum where the system could stably rest. Making u bigger than u_0 means that the cylinder must leave its stable position and suddenly roll down the plane. The point u_0 in the u-parameter space is called catastrophe point.

What is the shape of the section of F at u_0? Since in our example

$$f(x) = F(x, u_0) = 0.25x - 0.25\sin(x) \ ,$$

the graph appears as shown in Fig. 7.

The point $x = 0$ is a degenerated critical point of $f(x)$, since $f'(x) = f''(x) = 0$. If we expand f into its Taylor polynomial $t(x)$ around that point we get

$$t(x) = 0.0416667x^3 - 0.00208333x^5 + 0.0000496032x^7 \\ -6.88933 \times 10^{-7}x^9 + O[x]^{11} \ .$$

Catastrophe theory tries to simplify the Taylor polynomials of the potentials locally about their degenerated critical points. This simplification is done by coordinate transformations, which do not change the structure of the singularities. These transformations are called diffeomorphisms (see Sect. "Glossary"). A map $h\colon U \to V$, with U and V open in $\mathbb{R}^n$, is a local diffeomorphism if the Jacobian determinant $\det(J_p h) \neq 0$ at all points p in U. Two functions f and g, both defined in some neighborhood of the origin of $\mathbb{R}^n$, are called r-equivalent if there exists a local diffeomorphism $h\colon U \to V$ at the origin, such that $g = f \circ h$.

The Taylor polynomial $t(x)$ in our example starts with a term of 3rd order in x. Thus, there exists a diffeomorphism (a non-singular coordinate transformation), such that in the new coordinates t is of the form x^3. To see this, we write $t(x) = x^3 \cdot a(x)$, with $a(0) \neq 0$. We define $h(x) = x \cdot a(x)^{1/3}$. Note that the map h is a local diffeomorphism near the origin since $h'(0) = a(0)^{1/3} \neq 0$. (Hint: Modern computer algebra systems, such as Mathematica, are able to expand expressions like $a(x)^{1/3}$ and calculate the derivative of h easily for the concrete example.)

Thus for the eccentric cylinder the intersection of the graph of the potential $F(x, u)$ with the plane $u = u_0$ is the graph of a function, whose Taylor series at its critical point $x = 0$ is $t = x^3 \circ h$, in other words: t is r-equivalent to x^3.

Before we continue with this example and its relation to catastrophe theory, let us introduce another example which leads to a potential of the form x^4. Then we will have two examples which lead to the two simplest catastrophes in René Thom's famous list of the seven elementary catastrophes.

Example 2: The Formation of Traffic Jam

In the first example we constructed a family F of potential functions directly from the geometric properties of the model of the eccentric cylinder. We then found its critical points by calculating the derivatives of F. In this second example the method is slightly different. We start with a partial differential equation which occurs in traffic flow modeling. The solution surface of the initial value problem will be regarded as the surface of zeros of the derivative of a potential family. From this, we then calculate the family of potential functions. Finally, we will see that the Taylor series of a special member of this family, the one which belongs to its degenerated critical point, is equivalent to x^4.

Note: When modeling a traffic jam we use the names of variables that are common in literature about this subject. Later we will rename the variables into the ones commonly used in catastrophe theory.

Let $u(x, t)$ be the (continuous) density of the traffic at a street point x at time t and let $a(u)$ be the derivative of the traffic flow $f(u)$. The mathematical modeling of simple traffic problems leads to the well known traffic flow equation:

$$\frac{\partial u(x,t)}{\partial t} + a(u)\frac{\partial u(x,t)}{\partial x} = 0 \,.$$

We must solve the Cauchy problem for this quasilinear partial differential equation of first order with given initial density $u_0 = u(x, 0)$ by the method of characteristics. The solution is constant along the (base) characteristics, which are straight lines in xt-plane with slope $a(u_0(y))$ against the t-axes, if y is a point on the x-axes, where the characteristic starts at $t = 0$. For simplicity we write "a" instead of $a(u_0(y))$. Thus the solution of the initial value problem is $u(x, t) = u_0(x - a \cdot t)$ and the characteristic surface $S = \{(x, t, u) | u - u_0(x - a \cdot t) = 0\}$ is a surface in xtu-space. Under certain circumstances S may not lie uniquely above the xt-plane but may be folded as is shown in Fig. 8. There is another curve to be seen on the surface: The border curve of the fold. Its projection onto the xt-plane is a curve with a cusp, specifying the position and the beginning time of the traffic jam.

The equations used in the example, in addition to the traffic flow equation, can be deduced in traffic jam modeling from a simple parabolic model for a flow-density relation (see [5]). The constants in the following equations result from considerations of maximal possible density and an assumed maximal allowed velocity on the street. In our example a model of a road of 2 km in length and maximal traffic velocity of 100 km/h is chosen. Maximal traffic density is $u_{\max} = 0.2$. The equations are $f(u) = 27.78u - 138.89u^2$ and $u(x, 0) = u_0(x) = 0.1 + 0.1 \cdot \mathrm{Exp}(-((x - 2000)/700)^2)$. The latter was constructed to simulate a slight increase of the initial density u_0 along the street at time $t = 0$. The graph of u_0 can be seen as the front border of the surface.

To determine the cusp curve shown in Fig. 8, consider the following parameterization Φ of the surface S and the projection map $\pi\colon S \to \mathbb{R}^2$:

$$\begin{aligned} \Phi\colon \mathbb{R} \times \mathbb{R}_{\geq 0} &\to S \subset \mathbb{R}^3 \\ (x, t) &\mapsto (x + a(u_0(x)) \cdot t, t, u_0(x)) \,. \end{aligned}$$

The Jacobian matrix of $\pi \circ \Phi$ is

$$\begin{aligned} J(\pi \circ \Phi)(x, t) &= \begin{pmatrix} \dfrac{\partial(\pi \circ \Phi)_1}{\partial x} & \dfrac{\partial(\pi \circ \Phi)_1}{\partial t} \\ \dfrac{\partial(\pi \circ \Phi)_2}{\partial x} & \dfrac{\partial(\pi \circ \Phi)_2}{\partial t} \end{pmatrix} \\ &= \begin{pmatrix} 1 + a(u_0(x))' \cdot t & a(u_0(x)) \\ 0 & 1 \end{pmatrix} . \end{aligned}$$

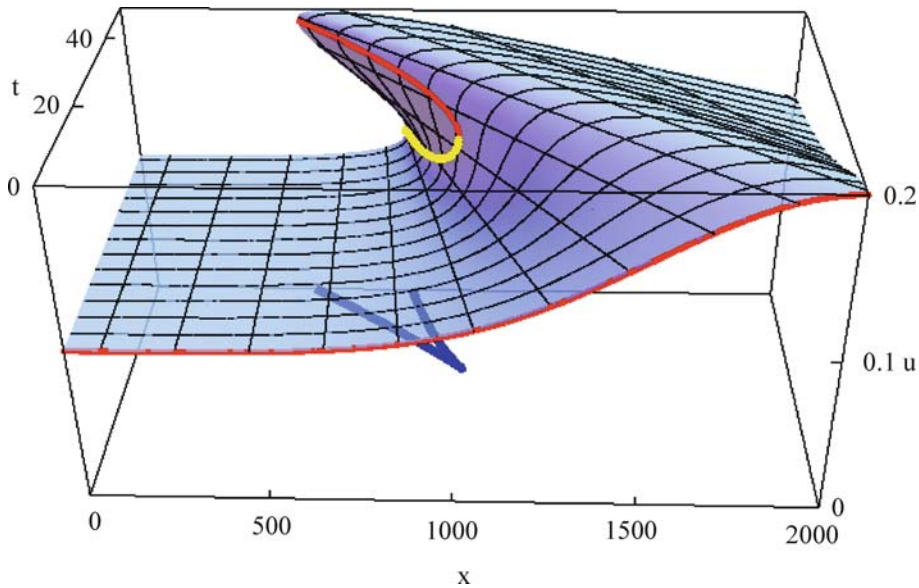

Catastrophe Theory, Figure 8
Characteristic surface *S*, initial density, fold curve and cusp curve

It is singular (that is, its determinate vanishes) if $1 + (a(u_0(x)))' \cdot t = 0$. From this we find the curve $c\colon \mathbb{R} \to \mathbb{R}^2$

$$c(x) = \left(x, -\frac{1}{a'(u_0(x)) \cdot u_0'(x)}\right)$$

which is the pre-image of the cusp curve and the cusp curve in the Fig. 8 is the image of c under $\pi \circ \Phi$.

Whether the surface folds or not depends on both the initial density u_0 and the properties of the flow function $f(u)$. If such folding happens, it would mean that the density would have three values at a given point in xt-space, which is a physical impossibility. Thus, in this case, no continuous solution of the flow equation can exist in the whole plane. "Within" the cusp curve, which is the boundary for the region where the folding happens, there exists a curve with its origin in the cusp point. Along this curve the surface must be "cut" so that a discontinuous solution occurs. The curve within the cusp region is called a shock curve and it is characterized by physical conditions (jump condition, entropy condition). The density makes a jump along the shock.

The cusp, together with the form of the folding of the surface, may give an association to catastrophe theory. One has to find a family of potential functions for this model, that is, functions $F_{(x,t)}(u)$, whose negative gradient, that is, its derivative by u (the inner parameter), describes the characteristic surface S by its zeroes. The family of potentials is given by

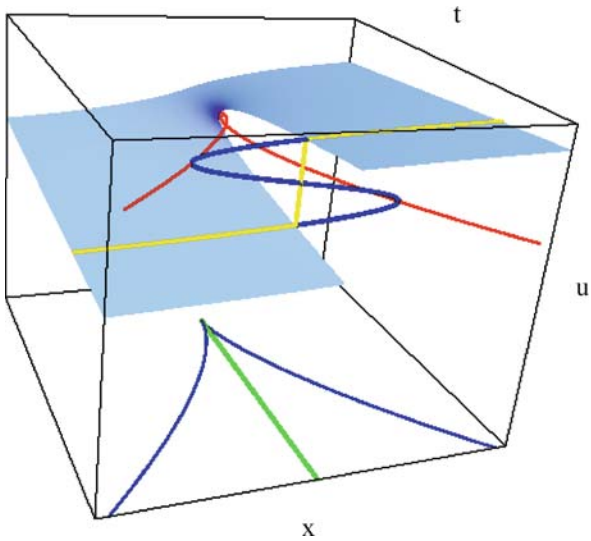

Catastrophe Theory, Figure 9
The solution surface is "cut" along the shock curve, running within the cusp region in the *xt*-plane

$$F(x,t,u) = -t \cdot \left(u \cdot a(u) - \int a(u)\mathrm{d}u\right) - \int_0^{x-a(u)\cdot t} u_0(x)\mathrm{d}x$$

(see [6]). Since

$$-\mathrm{grad}F_{x,t}(u) = -\frac{\partial F}{\partial u} = 0 \Leftrightarrow u - u_0(x - a(u) \cdot t) = 0$$

the connection of F and S is evident. One can show that

a member f of the family F, the function $f(u) = F(\xi, \tau, u)$, is locally in a traffic jam formation point (ξ, τ) of the form $f(u) = u^4$ (after Taylor expansion and after suitable coordinate transformations, similar to the first example).

The complicated function terms of f and F thus can be replaced in qualitative investigations by simple polynomials. From the theorems of catastrophe theory, more properties of the models under investigation can be discovered.

Now we want to use the customary notation of catastrophe theory. We shall rename the variables: u is the inner variables, which in catastrophe theory usually is x, and x, t (the outer parameters) get the names u and v respectively. Thus $F = F(x, u, v)$ is a potential family with one inner variable (the state variable x) and two outer variables (the control variables u and v).

Unfoldings

In the first examples above we found a family $F(x, u)$ of potential functions for the eccentric cylinder. For u fixed we found single members of that family. The member which belonged to the degenerated critical point of F, that is, to the point (x, u) where $\partial_x F = \partial_{xx} F = 0$, turned out to be equivalent to x^3. In the second example the corresponding member is equivalent to x^4 (after renaming the variables). These two singularities are members of potential families which also can be transformed into simple forms.

In order to learn how such families, in general, arise from a single function germ, let us look at the singularity $f(x) = x^3$. If we add to it a linear "disturbance" $u \cdot x$, where the factor (parameter) u may assume different values, we qualitatively have one of the function graphs in Fig. 10.

The solid curve in Fig. 10 represents the graph of $f(x) = x^3$. The dashed curve is the graph of $g(x) = x^3 - u \cdot x$; $u > 0$. The dotted line is $h(x) = x^3 + u \cdot x$; $u > 0$.

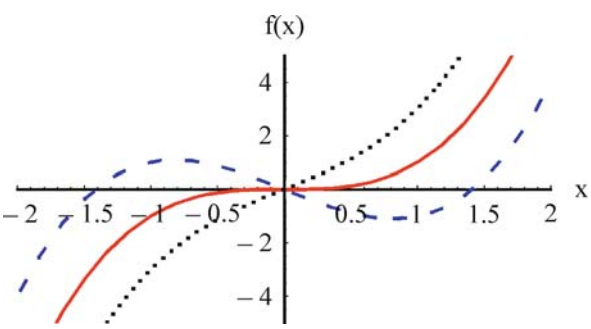

Catastrophe Theory, Figure 10
Perturbations of x^3

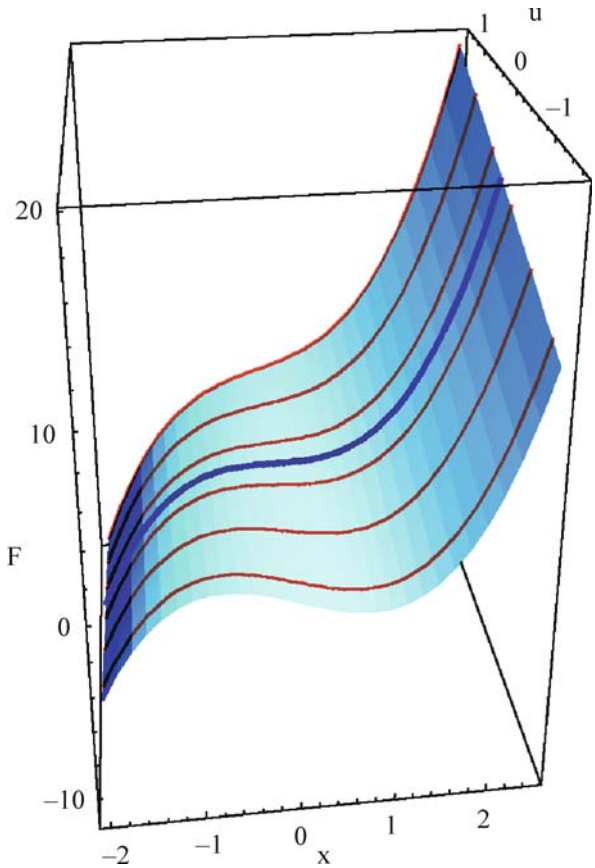

Catastrophe Theory, Figure 11
The function $f(x) = x^3$ embedded in a function family

Please note: While for positive parameter values, the disturbed function has no singularity, in the case of negative u-values a relative maximum and a relative minimum exist. The origin is a singular point only for $u = 0$, that is, for the undisturbed function f.

We can think of $f(x) = x^3$ as a member of a function family, which contains disturbances with linear terms.

The function $f(x) = x^3$ changes the type of singularity with the addition of an appropriate small disturbing function, as we have seen. Therefore f is called "structurally unstable". But we have also learned that $f(x) = x^3$ can be seen as a member of a whole family of functions $F(x, u) = x^3 + u \cdot x$, because $F(x, 0) = f(x)$. This family is stable in the sense that all kinds of singularities of the perturbed function are included by its members. This family is only one of the many possibilities to "unfold" the function germ $f(x) = x^3$ by adding disturbing terms. For example $F(x, u, v) = x^3 + u \cdot x^2 + v \cdot x$ would be another such unfolding (see Sect. "Glossary") of f. But this unfolding has more parameters than the minimum

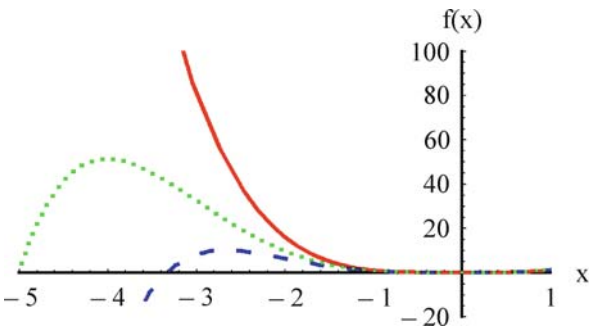

Catastrophe Theory, Figure 12
The function $f(x) = x4 + u \cdot x5$, $u = 0$ (*solid*), $u = 0.2$ (*dotted*) and $u = 0.3$ (*dashed*)

needed. The example of a traffic jam leads to an unfolding of $f(x) = x^4$ which is $F(x, u, v) = x^4 + u \cdot x^2 + v \cdot x$.

How can we find unfoldings, which contain all possible disturbances (all types of singularities), that is, which are stable and moreover have a minimum number of parameters?

Examine the singularity x^4 (Fig. 12). It has a degenerated minimum at the origin. We disturb this function by a neighboring polynomial of higher degree, for example, by $u \cdot x^5$, where the parameter u with small values (here and in what follows "small" means small in amount), ensures that the perturbed function is "near" enough to x^4.

For $f(x) = x^4 + u \cdot x^5$, we find that the equation $f'(x) = 0$ has the solutions $x_1 = x_2 = x_3 = 0$, $x_4 = -4/(5u)$. To the threefold zero of the derivative, that is, the threefold degenerated minimum, another extremum is added which is arbitrarily far away from the origin for values of u that are small enough in amount due to the term $-4/(5u)$. This is so, because the term $-4/(5u)$ increases as u decreases. (This is shown in Fig. 12 with $u = 0$, $u = 0.2$ and $u = 0.3$). The type of the singularity at the origin thus is not influenced by neighboring polynomials of the form $u \cdot x^5$.

If we perturb the polynomial x^4 by a neighboring polynomial of lower degree, for example, $u \cdot x^3$, then for small amounts of u a new minimum is generated arbitrarily close to the existing singularity at the origin.

If disturbed by a linear term, the singularity at the origin can even be eliminated. The only real solution of $h'(x) = 0$, where $h(x) = x^4 + ux$, is $x = -(u/4)^{(1/3)}$.

Note: Only for $u = 0$, that is, for a vanishing disturbance, is h singular at the origin. For each $u \neq 0$ the linear disturbance ensures that no singularity at the origin occurs.

The function $f(x) = x^4$ is structurally unstable. To find a stable unfolding for the singularity $f(x) = x^4$, we must add terms of lower order. But we need not take into account all terms $x^4 + u \cdot x^3 + v \cdot x^2 + w \cdot x + k$, since the absolute term plays no role in the computation of singular points, and each polynomial of degree 4 can be written by a suitable change of coordinates without a cubic term. Similarly, a polynomial of third degree after a coordinate transformation can always be written without a quadratic term. The "Tschirnhaus-transformation": $x \mapsto x - a_1/n$, if a_1 is the coefficient of x^{n-1} in a polynomial $p(x) = x^n + a_1 x^{n-1} + \ldots + a_{n-1}x + a_n$ of degree n, leads to a polynomial $q(x) = x^n + b_1 x^{n-2} + \ldots + b_{n-2}x + b_{n-1}$ without a term of degree $n - 1$. Indeed, the expression $F(x, u, v) = x^4 + u \cdot x^2 + v \cdot x$ is, as we shall see later, the "universal" unfolding of x^4. Here an unfolding F of a germ f is called universal, if every other unfolding of f can be received by suitable coordinate transformations "morphism of unfoldings" from F, and the number of unfolding parameters of F is minimal.

The minimal number of unfolding parameters needed for the universal unfolding F of the singularity f is called the codimension of f. It can be computed as follows:

$$\mathrm{codim}(f) = \dim_{\mathbb{R}} \mu(n)/\langle \partial_x f \rangle \ .$$

Here $\langle \partial_x f \rangle$ is the Jacobian ideal generated by the partial derivatives of f and $\mu(n) = \{f \in \varepsilon(n) | f(0) = 0\}$. $\varepsilon(n)$ is the vector space of function germs at the origin of $\mathbb{R}^n$. The quotient $\mu(n)/\langle \partial_x f \rangle$ is the factor space.

What is the idea behind that factor space? Remember, that the mathematical description of a plane (a two dimensional object) in space (3 dimensional) needs only one equation, while the description of a line (1 dimensional) in space needs two equations. The number of equations that are needed for the description of these geometrical objects is determined by the difference between the dimensions of the surrounding space and the object in it. This number is called the codimension of the object. Thus the codimension of the plane in space is 1 and the codimension of the line in space is 2. From linear algebra it is known that the codimension of a subspace W of a finite dimensional vector space V can be computed as $\mathrm{codim}\, W = \dim V - \dim W$. It gives the number of linear independent vectors that are needed to complement a basis of W to get a basis of the whole space V. Another well known possibility for the definition is $\mathrm{codim}\, W = \dim V/W$. This works even for infinite dimensional vector spaces and agrees in finite dimensional cases with the previous definition. In our example $V = \mu(n)$ and W should be the set O of all functions germs, which are r-equivalent to f (O is the orbit of f under the action of the group of local diffeomorphisms preserving the origin of $\mathbb{R}^n$). But this is an infinite dimen-

sional manifold, not a vector space. Instead, we can use its tangent space $T_f O$ in the point f, since spaces tangent to manifolds are vector spaces with the same dimension as the manifold itself and this tangent space is contained in the space tangent to $\mu(n)$ in an obvious way. It turns out, that the tangent space $T_f O$ is $\langle \partial_x f \rangle$. The space tangent to $\mu(n)$ agrees with $\mu(n)$, since it is a vector space. The details for the computations together with some examples can be found in the excellent article by M. Golubitsky (1978).

The Seven Elementary Catastrophes

Elementary catastrophe theory works with families (unfoldings) of potential functions

$$F\colon \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R} \quad (\text{with } k \le 4)$$
$$(x, u) \mapsto F(x, u)\ .$$

$\mathbb{R}^n$ is called the state space, $\mathbb{R}^k$ is called the parameter space or control space. Accordingly $x = (x_1, \ldots, x_n)$ is called a state variable or endogenous variable and $u = (u_1, \ldots, u_k)$ is called a control variable (exogenous variable, parameter).

We are interested in the singularities of F with respect to x and the dependence of F with respect to the control variables u. We regard F as the unfolding of a function germ f and want to classify F.

In analogy to the r-equivalence of function germs, we call two unfoldings F and G of the same function germ r-equivalent as unfoldings if there exists an unfolding-morphism (see Sect. "Glossary") from F to G. An unfolding F of f is called versal (or stable), if each other unfolding of f is right-equivalent as unfolding to F, that is, if there exists a right-morphism between the unfoldings. A versal unfolding with minimal unfolding dimensions is called universal unfolding of f.

Thus we have the following theorem (all proofs of the theorems cited in this article can be found in [1,2] or [12].)

Theorem 1 (Theorem on the existence of a universal unfolding) *A singularity f has a universal unfolding iff $codim(f) = k < \infty$.*

Examples (see also the following theorem with its examples):

$$\text{If } f(x) = x^3 \quad \text{it follows}$$
$$F(x, u, v) = x^3 + ux^2 + vx \quad \text{is versal,}$$
$$F(x, u) = x^3 + ux \quad \text{is universal,}$$
$$F(x, u) = x^3 + ux^2 \quad \text{is not versal}\ .$$

The following theorem states how one can find a universal unfolding of a function germ:

Theorem 2 (Theorem on the normal form of universal unfoldings) *Let $f \in \mu(n)$ be a singularity with $codim(f) = k < \infty$. Let $u = (u_1, \ldots, u_k)$ be the parameters of the unfolding. Let $b_i(x)$, $i = 1, \ldots, k$, be the elements of $\mu(n)$, whose cosets modulo $\langle \partial_x f \rangle$ generate the vector space $\mu(n)/\langle \partial_x f \rangle$. Then*

$$F(x, u) = f(x) + \sum_{i=1}^{k} u_i b_i(x)$$

is a universal unfolding of f.

Examples:

1. Consider $f(x) = x^4$. Here $n = 1$ and $\mu(n) = \mu(1) = \langle x \rangle$. The derivative of f is $4x^3$, so $\langle \partial_x f \rangle = \langle x^3 \rangle$ and we get $\langle x \rangle / \langle x^3 \rangle = \langle x, x^2 \rangle$. Thus $F(x, u_1, u_2) = x^4 + u_1 x^2 + u_2 x$ is the universal unfolding of f.
2. Consider $f(x, y) = x^3 + y^3$. Here $n = 2$ and $\mu(n) = \mu(2) = \langle x, y \rangle$. The partial derivatives of f with respect to x and y are $3x^2$ and $3y^2$, thus $\langle \partial_x f \rangle = \langle x^2, y^2 \rangle$ and we get $\langle x, y \rangle / \langle x^2, y^2 \rangle = \langle x, xy, y \rangle$. Therefore $F(x, y, u_1, u_2, u_3) = x^3 + y^3 + u_1 \cdot x \cdot y + u_2 \cdot x + u_3 \cdot y$ is the universal unfolding of f.

Hint: There exists a useful method for the calculation of the quotient space which you will find in the literature as "Siersma's trick" (see for example [7]).

We now present René Thom's famous list of the seven elementary catastrophes.

Classification Theorem (Thom' s List)

Up to addition of a non degenerated quadratic form in other variables and up to multiplication by ± 1 a singularity f of codimension k ($1 \le k \le 4$) is right-equivalent to one of the following seven:

f	codim f	universal unfolding	name
x^3	1	$x^3 + ux$	fold
x^4	2	$x^4 + ux^2 + vx$	cusp
x^5	3	$x^5 + ux^3 + vx^2 + wx$	swallowtail
$x^3 + y^3$	3	$x^3 + y^3 + uxy + vx + wy$	hyperbolic umbilic
$x^3 - xy^2$	3	$x^3 - xy^2 + u(x^2 + y^2) + vx + wy$	elliptic umbilic
x^6	4	$x^6 + ux^4 + vx^3 + wx^2 + tx$	butterfly
$x^2y + y^4$	4	$x^2y + y^4 + ux^2 + vy^2 + wx + ty$	parabolic umbilic

In a few words: The seven elementary catastrophes are the seven universal unfoldings of singular function germs

of codimension k ($1 \leq k \leq 4$). The singularities itself are called organization centers of the catastrophes.

We will take a closer look at the geometry of the simplest of the seven elementary catastrophes: the fold and the cusp. The other examples are discussed in [9].

The Geometry of the Fold and the Cusp

The fold catastrophe is the first in Thom's list of the seven elementary catastrophes. It is the universal unfolding of the singularity $f(x) = x^3$ which is described by the equation $F(x, u) = x^3 + u \cdot x$. We are interested in the set of singular points of F relative to x, that is, for those points where the first partial derivative with respect to x (the system variable) vanishes. This gives the parabola $u = -3x^2$. Inserting this into F gives a space curve.

Figure 13 shows the graph of $F(x, u)$. The curve on the surface joins the extrema of F and the projection curve represents their x, u values. There is only one degenerated critical point of F (at the vertex of the Parabola), that is, a critical point where the second derivative by x also vanishes. The two branches of the parabola in the xu-plane give the positions of the maxima and the minima of F, respectively. At the points belonging to a minimum, the system is stable, while in a maximum it is unstable. Projecting x–u space, and with it the parabola, onto parameter space (u-axis), one gets a straight line shown within the interior of the parabola, which represents negative parameter values of u. There, the system has a stable minimum (and an unstable maximum, not observable in nature). For parameter values $u > 0$ the system has no stability points at all. Interpreting the parameter u as time, and "walking along" the u-axis, the point $u = 0$ is the beginning or the end of a stable system behavior. Here the system shows catastrophic behavior. According to the interpretation of the external parameter as time or space, the morphology of the fold is a "beginning," an "end" or a "border", where something new occurs.

What can we say about the fold catastrophe and the modeling of our first example, the eccentric cylinder? We have found that the potential related to the catastrophe point of the eccentric cylinder is equivalent to x^3, and $F(x, u) = x^3 + ux$ is the universal unfolding. This is independent of the size of our 'machine'. The behavior of the machine should be qualitatively the same as that of other machines which are described by the fold catastrophe as a mechanism. We can expect that the begin (or end) of a catastrophe depends on the value of a single outer parameter. Stable states are possible (for $u < 0$ in the standard fold. In the untransformed potential function of our model of the eccentric cylinder, this value is $u = 0.25268$)

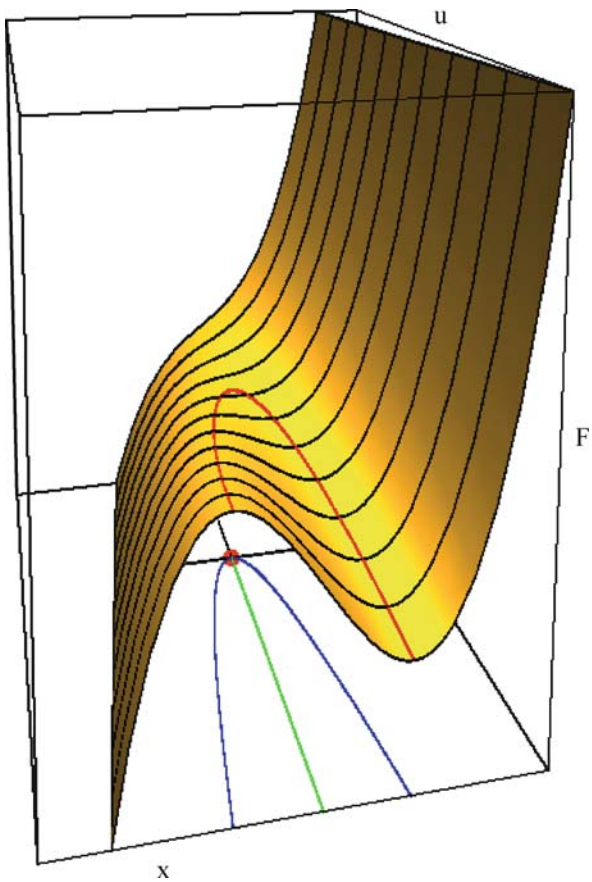

Catastrophe Theory, Figure 13
Graph of $F(x, u) = x^3 + u \cdot x$

corresponding to the local minima of $F(., u)$. This means that it is possible for the cylinder to stay at rest on the inclined plane. There are unstable equilibria (local maxima of $F(., u)$ and the saddle point, too). The cylinder can be turned such that the center of gravity lies in the upper position over the supporting point (see Fig. 3). But a tiny disturbance will make the system leave this equilibrium.

Let us now turn to the cusp catastrophe. The cusp catastrophe is the universal unfolding of the function germ $f(x) = x^4$. Its equation is $F(x, u, v) = x^4 + u \cdot x^2 + v \cdot x$. In our second example (traffic jam), the function f is equivalent to the organization center of the cusp catastrophe, the second in René Thom's famous list of the seven elementary catastrophes. So the potential family is equivalent to the universal unfolding $F(x, u, v) = x^4 + u \cdot x^2 + v \cdot x$ of the function x^4. We cannot draw such a function in three variables in three-space, but we can try to draw sections giving u or v certain constant values. We want to look at its catastrophe surface S, that is, the set of those points in (x, u, v)-

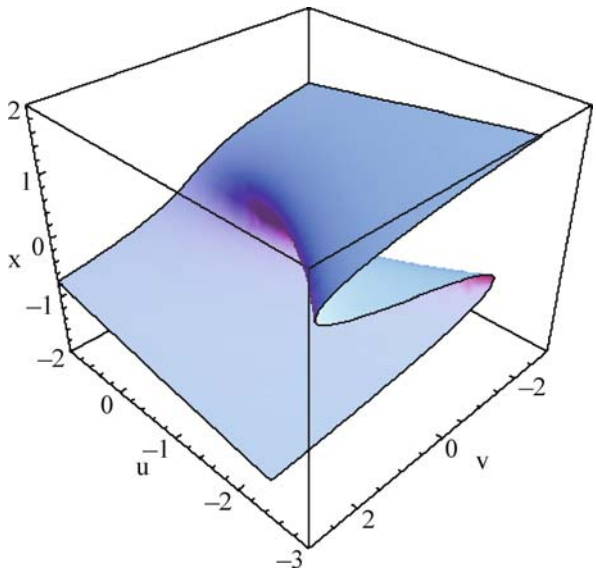

Catastrophe Theory, Figure 14
The catastrophe manifold of the cusp catastrophe

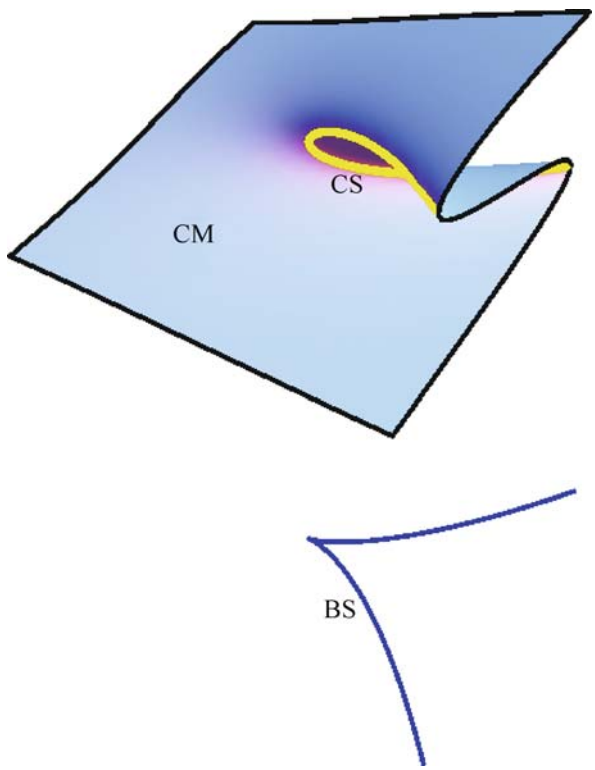

Catastrophe Theory, Figure 15
The catastrophe manifold (CM), catastrophe set (CS) and bifurcation set (BS) of the cusp catastrophe

space, where the partial derivative of F with respect to the system variable x is zero. It is a surface in three-space called catastrophe manifold, stability surface or equilibrium surface, since the surface S describes the equilibrium points of the system.

If S is not folded over a point in u, v-space there is exactly one minimum of the potential F. If it is folded in three sheets, there are two minima (corresponding to the upper and lower sheet of the fold) and a maximum (corresponding to the middle sheet). Stable states of the system belong to the stable minima, that is, to the upper and lower sheets of the folded surface. Other points in three-space that do not lie on the surface S correspond to the states of the system that are not equilibrium. The system does not rest there. Catastrophes do occur when the system jumps suddenly from one stable state to another stable state, that is, from one sheet to the other. There are two principal possibilities, called "conventions", where this jump can happen. The "perfect delay convention" says that jumps happen at the border of the folded surface. The "Maxwell convention" says that jumps can happen along a curve in uv-space (a "shock-curve") which lays inside the cusp area. This curve consists of those points (u, v) in parameter space, where $F(., u, v)$ has two critical points with the same critical value. In our traffic jam model a shock curve is a curve in xt-space that we can interpret dynamically as the movement of the jam formation front along the street (see Fig. 9). The same figure shows what we call the morphology of the cusp: according to the convention, the catastrophe manifold with its fold or its cut can be interpreted as a fault or slip (as in geology) or a separation (if one of the parameter represents time). In our traffic jam model, there is a "line" (actually a region) of separation (shock wave) between the regions of high and low traffic density.

Besides the catastrophe manifold (catastrophe surface) there are two other essential terms: catastrophe set and bifurcation set (see Fig. 15). The catastrophe set can be viewed as the curve which goes along the border of the fold on the catastrophe manifold. It is the set of degenerated critical points and is described mathematically as the set $\{(x, u, v) | \partial_x F = \partial_x^2 F = 0\}$. Its projection into the uv-parameter space is the cusp curve, which is called the bifurcation set of the cusp catastrophe.

The cusp catastrophe has some special properties which we discuss now with the aid of some graphics. If the system under investigation shows one or more of these properties, the experimenter should try to find a possible cusp potential that accompanies the process.

The first property is called divergence. This property can be shown by the two curves (black and white) on the

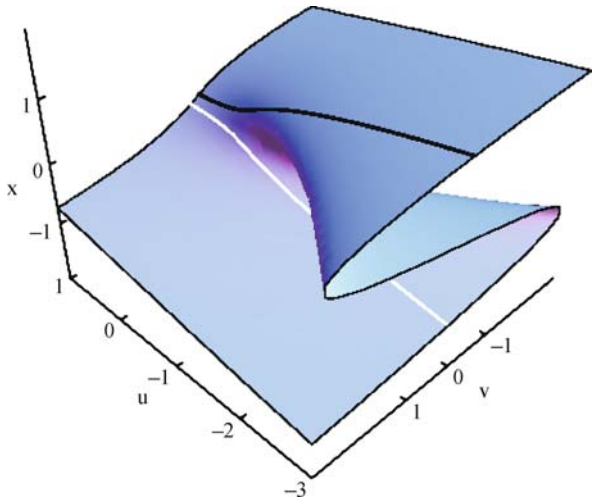

Catastrophe Theory, Figure 16
The divergence property of the cusp catastrophe

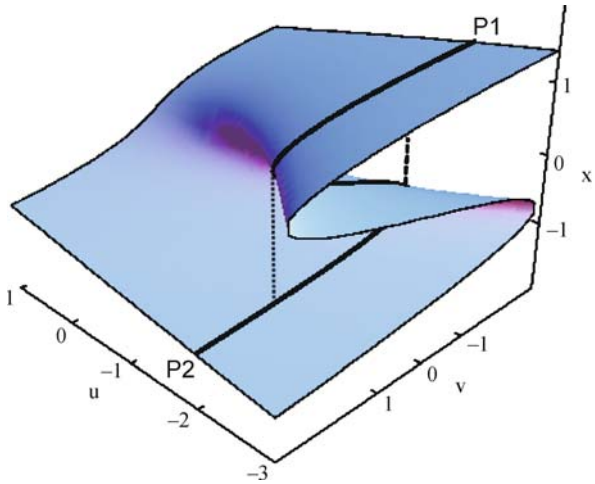

Catastrophe Theory, Figure 17
The hysteresis property of the cusp catastrophe

catastrophe manifold which describes the system's equilibrium points. Both curves initially start at nearby positions on the surface, that is, they start at nearby stable system states. But the development of the system can proceed quite differently. One of the curves runs on the upper sheet of the surface while the other curve runs on the lower sheet. The system's stability positions are different and thus its behavior is different.

The next property is called hysteresis. If the system development is running along the path shown in Fig. 17 from P1 to P2 or vice versa, the jumps in the system behavior, that is, the sudden changes of internal system variables, occur at different parameter constellations, depending on

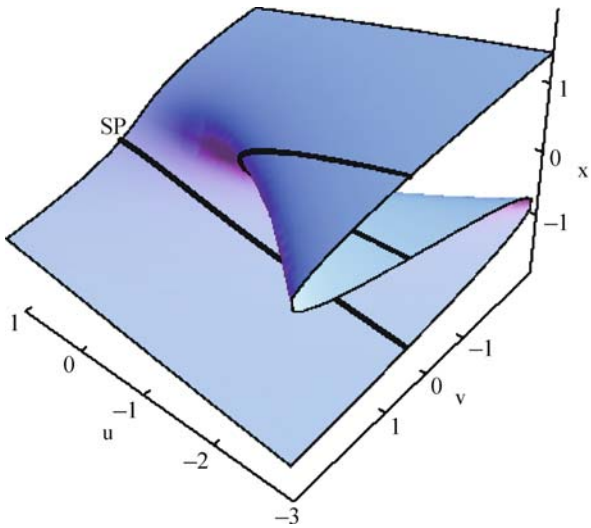

Catastrophe Theory, Figure 18
The bifurcation property of the cusp catastrophe

the direction the path takes, from the upper to the lower sheet or vice versa. Jumps upward or downward happen at the border of the fold along the dotted lines. The name of this property of the cusp comes from the characteristic hysteresis curve in physics, occurring for example at the investigation of the magnetic properties of iron.

Figure 18 shows the bifurcation property of the cusp catastrophe: the number of equilibrium states of the system splits from one to three along the path beginning at the starting point (SP) and running from positive to negative u values as shown. If the upper and the lower sheet of the fold correspond to the local minima of F and the middle sheet corresponds to the local maxima, then along the path shown, one stable state splits into two stable states and one unstable state.

Further Applications

Many applications of catastrophe theory are attributable to Christopher Zeeman. For example, his catastrophe machine, the "Zeeman wheel," is often found in literature. This simple model consists of a wheel mounted flat against a board and able to turn freely. Two elastics are attached at one point (Fig. 19, point B) close to the periphery of the wheel. One elastic is fixed with its second end on the board (point A). The other elastic can be moved with its free end in the plane (point C).

Moving the free end smoothly, the wheel changes its angle of rotation smoothly almost everywhere. But at certain positions of C, which can be marked with a pencil, the wheel suddenly changes this angle dramatically. Join-

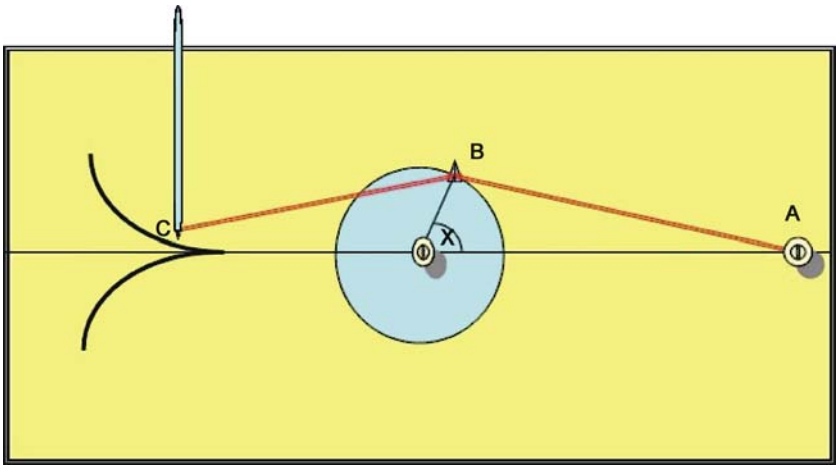

Catastrophe Theory, Figure 19
The Zeeman wheel

ing the marked points where these jumps occur, you will find a cusp curve in the plane. In order to describe this behavior, the two coordinates of the position of point C serve as control parameters in the catastrophe model. The potential function for this model results from Hook's law and simple geometric considerations. Expanding the potential near its degenerated critical point into its Taylor series, and transforming the series with diffeomorphisms similar to the examples we have already calculated, this model leads to the cusp catastrophe. Details can be found in the book of Poston and Stewart [7] or Saunders (1986).

Another example of an application of catastrophe theory is similar to the traffic jam model we used here. Catastrophe theory can describe the formation of shockwaves in hydrodynamics. Again the cusp catastrophe describes this phenomenon. The mathematical frame is given in the works of Lax [6], Golubitsky and Schaeffer[3] and Guckenheimer [4].

In geometrical optics light rays are investigated which are reflected on smooth surfaces. Caustics are the envelopes of the light rays. They are the bright patterns of intensive light, which can be seen, for example, on a cup of coffee when bright sunlight is reflected on the border of the cup. Catastrophe theory can be applied to light caustics, because the light rays obey a variational principle. According to Fermat's principle, light rays travel along geodesics and the role of the potential functions in catastrophe theory is played by geodesic distance functions. Caustics are their bifurcation sets. The calculations are given for example in the work of Sinha [11].

A purely speculative example, given by Zeeman, is an aggression model for dogs. Suppose that fear and rage of

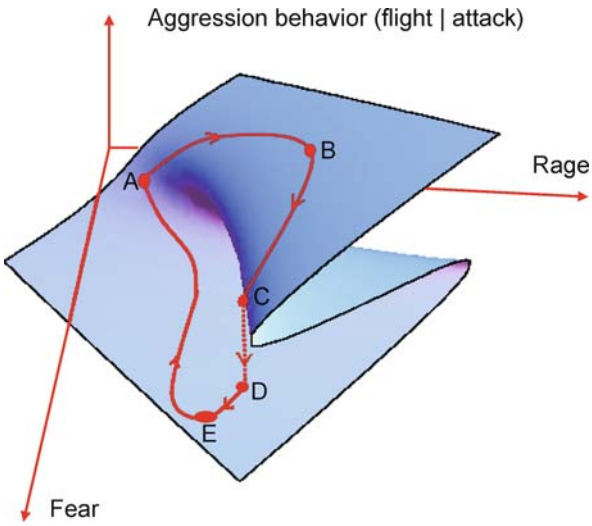

Catastrophe Theory, Figure 20
A dog's aggression behavior depending on its rage and fear

a dog can be measured in some way from aspects of its body language (its attitude, ears, tail, mouth, etc.). Fear and rage are two conflicting parameters, so that a catastrophe in the dog's behavior can happen. Look at Fig. 20.

The vertical axis shows the dog's aggression potential, the other axes show its fear and rage. Out of the many different ways the dog can behave, we choose one example. Suppose we start at point A on the catastrophe manifold (behavioral surface). Maybe the dog is dozing in the sun without any thought of aggression. The dog's aggression behavior is neutral (point A). Another dog is coming closer and our dog, now awakened, is becoming more and

more angry by this breach of his territory. Moving along the path on the behavior surface, the dog is on a trajectory to attack his opponent. But it suddenly notices that the enemy is very big (point B). So its fear is growing, its rage diminishes. At point C the catastrophe happens: Attack suddenly changes to flight (point D). This flight continues until the dog notices that the enemy does not follow (point E). Fear and rage both get smaller until the dog becomes calm again.

The example can be modified a little bit to apply to conflicts between two states. The conflicting factors can be the costs and the expected gain of wars, and so on.

Among the variety of possible examples, those worthy of mention include investigations of the stability of ships, the gravitational collapse of stars, the buckling beam (Euler beam), the breaking of ocean waves, and the development of cells in biology. The latter two problems are examples for the occurrence of higher catastrophes, for example the hyperbolic umbilic (see Thom's List of the seven elementary catastrophes).

Future Directions

When catastrophe theory came up in the nineteen-sixties, much enthusiasm spread among mathematicians and other scientists about the new tool that was expected to explain many of the catastrophes in natural systems. It seemed to be the key for the explanation of discontinuous phenomena in all sciences. Since many applications were purely qualitative and speculative, a decade later this enthusiasm ebbed away and some scientists took this theory for dead. But it is far from that. Many publications in our day show that it is still alive. The theory has gradually become a 'number producing' theory, so that it is no longer perceived as purely qualitative. It seems that it will be applied the sciences increasingly, producing numerical results. Thom's original ideas concerning mathematical biology seem to have given the basis for the trends in modern biology. This is probably the most promising field of application for catastrophe theory. New attempts are also made, for example, in the realm of decision theory or in statistics, where catastrophe surfaces may help to clarify statistical data.

From the point of view of teaching and learning mathematics, it seems that catastrophe theory is becoming and increasingly popular part of analysis courses at our universities. Thus, students of mathematics meet the basic ideas of catastrophe theory within their first two years of undergraduate studies. Perhaps in the near future its basic principles will be taught not only to the mathematical specialists but to students of other natural sciences, as well.

Bibliography

Primary Literature

1. Bröcker T (1975) Differentiable germs and catastrophes. London Mathem. Soc. Lecture Notes Series. Cambridge Univ Press, Cambridge
2. Golubitsky M, Guillemin V (1973) Stable mappings and their singularities. Springer, New York, GTM 14
3. Golubitsky M, Schaeffer DG (1975) Stability of shock waves for a single conservation law. Adv Math 16(1):65–71
4. Guckenheimer J (1975) Solving a single conservation law. Lecture Notes, vol 468. Springer, New York, pp 108–134
5. Haberman R (1977) Mathematical models – mechanical vibrations, population dynamics, and traffic flow. Prentice Hall, New Jersey
6. Lax P (1972) The formation and decay of shockwaves. Am Math Monthly 79:227–241
7. Poston T, Stewart J (1978) Catastrophe theory and its applications. Pitman, London
8. Sanns W (2000) Catastrophe theory with mathematica – a geometric approach. DAV, Germany
9. Sanns W (2005) Genetisches Lehren von Mathematik an Fachhochschulen am Beispiel von Lehrveranstaltungen zur Katastrophentheorie. DAV, Germany
10. Saunders PT (1980) An introduction to catastrophe theory. Cambridge University Press, Cambridge. Also available in German: Katastrophentheorie. Vieweg (1986)
11. Sinha DK (1981) Catastrophe theory and applications. Wiley, New York
12. Wassermann G (1974) Stability of unfoldings. Lect Notes Math, vol 393. Springer, New York

Books and Reviews

Arnold VI (1984) Catastrophe theory. Springer, New York

Arnold VI, Afrajmovich VS, Il'yashenko YS, Shil'nikov LP (1999) Bifurcation theory and catastrophe theory. Springer

Bruce JW, Gibblin PJ (1992) Curves and singularities. Cambridge Univ Press, Cambridge

Castrigiano D, Hayes SA (1993) Catastrophe theory. Addison Wesley, Reading

Chillingworth DRJ (1976) Differential topology with a view to applications. Pitman, London

Demazure M (2000) Bifurcations and catastrophes. Springer, Berlin

Fischer EO (1985) Katastrophentheorie und ihre Anwendung in der Wirtschaftswissenschaft. Jahrb f Nationalök u Stat 200(1):3–26

Förster W (1974) Katastrophentheorie. Acta Phys Austr 39(3):201–211

Gilmore R (1981) Catastrophe theory for scientists and engineers. Wiley, New York

Golubistky M (1978) An introduction to catastrophe theory. SIAM 20(2):352–387

Golubitsky M, Schaeffer DG (1985) Singularities and groups in bifurcation theory. Springer, New York

Guckenheimer J (1973) Catastrophes and partial differential equations. Ann Inst Fourier Grenoble 23:31–59

Gundermann E (1985) Untersuchungen über die Anwendbarkeit der elementaren Katastrophentheorie auf das Phänomen "Waldsterben". Forstarchiv 56:211–215

Jänich K (1974) Caustics and catastrophes. Math Ann 209:161–180

Lu JC (1976) Singularity theory with an introduction to catastrophe theory. Springer, Berlin
Majthay A (1985) Foundations of catastrophe theory. Pitman, Boston
Mather JN (1969) Stability of C^{∞}-mappings. I: Annals Math 87:89–104; II: Annals Math 89(2):254–291
Poston T, Stewart J (1976) Taylor expansions and catastrophes. Pitman, London
Stewart I (1977) Catastrophe theory. Math Chronicle 5:140–165
Thom R (1975) Structural stability and morphogenesis. Benjamin Inc., Reading
Thompson JMT (1982) Instabilities and catastrophes in science and engineering. Wiley, Chichester
Triebel H (1989) Analysis und mathematische Physik. Birkhäuser, Basel
Ursprung HW (1982) Die elementare Katastrophentheorie: Eine Darstellung aus der Sicht der Ökonomie. Lecture Notes in Economics and Mathematical Systems, vol 195. Springer, Berlin
Woodcock A, Davis M (1978) Catastrophe theory. Dutton, New York
Zeeman C (1976) Catastrophe theory. Sci Am 234(4):65–83

Cell Biology: Networks, Regulation and Pathways

Gašper Tkačik[1], William Bialek[1,2]
[1] Joseph Henry Laboratories of Physics, Lewis–Sigler Institute for Integrative Genomics, Princeton University, Princeton, USA
[2] Princeton Center for Theoretical Physics, Princeton University, Princeton, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Biological Networks and Their Building Blocks
Models of Biological Networks
Network Properties and Operating Principles
Future Directions
Acknowledgments
Bibliography

Glossary

Dynamical system is a set of components the properties of which (e. g. their quantity, activity level etc.) change in time because the components interact among themselves and are also influenced by external forces.

Network node is a constituent component of the network, in biological networks most often identified with a molecular species.

Interaction is a connection between network nodes; in biological networks an interaction means that two nodes chemically react, regulate each other, or effectively influence each other's activities. Interactions are mostly pairwise, but can be higher-order as well; they can be directed or undirected, and are usually characterized by an interaction strength.

Network is a system of interacting nodes. A network can be represented mathematically as a graph, where vertices denote the nodes and edges denote the interactions. Biological networks often are understood to be dynamical systems as well, because the activities of network nodes evolve in time due to the graph of interactions.

Network state is the vector of activities of all nodes that fully characterizes the network at any point in time; since a biological network is a dynamical system, this state generally changes through time according to a set of dynamical equations.

Biological function refers to the role that a specific network plays in the life of the organism; the network can be viewed as existing to perform a task that enables the cell to survive and reproduce, such as the detection or transduction of a specific chemical signal.

Pathway is a subset of nodes and interactions in a network along which information or energy and matter flow in a directed fashion; pathways can be coupled through interactions or unwanted cross-talk.

Curse of dimensionality is the rapid increase of complexity encountered when analyzing or experimentally observing network states, as more and more network nodes are added. If there are N network nodes each of which only has two states (for example *on* and *off*), the number of states that the network can be in grows as 2^N.

Design principle is an (assumed) constraint on the network architecture, stating that a biological network, in addition to performing a certain function, implements that function in a particular way, usually to maximize or minimize some further objective measure, for instance robustness, information transmission, or designability.

Definition of the Subject

In cell biology, networks are systems of interacting molecules that implement cellular functions, such as the regulation of gene expression, metabolism or intracellular signaling. While on a molecular level a biological network is a mesh of chemical reactions between, for example, enzymes and their substrates, or DNA-binding proteins and the genes that they regulate, the collective effect of these reactions can often be thought of as the enabling and regulat-

ing the flow of matter and energy (in metabolic networks), or of information (in signaling and transcriptional regulatory networks). The field is concerned primarily with the description and properties of such flows and with their emergence from network constituent parts – the molecules and their physical interactions. An important focus is also the question of how network function and operating principles can be inferred despite the limited experimental access to network states and building blocks.

Introduction

Biological network has come to mean a system of interacting molecules that jointly perform cellular tasks such as the regulation of gene expression, information transmission, or metabolism [28]. Specific instances of biological networks include, for example, the DNA and DNA binding proteins comprising the transcriptional regulatory network; signaling proteins and small molecules comprising various signaling networks; or enzymes and metabolites comprising the metabolic network. Two important assumptions shape our current understanding of such systems: first, that the biological networks have been under selective evolutionary pressure to perform specific cellular functions in a way that furthers the overall reproductive success of the individual; and second, that these functions often are not implemented on a microscopic level by single molecules, but are rather a collective property of the whole interaction network. The question of how complex behavior emerges in a network of (simple) nodes under a functional constraint is thus central [144].

To start off with a concrete example, consider chemotaxis in the bacterium *Escherichia coli* [16,40], one of the paradigmatic examples of signal transduction. This system is dedicated to steering the bacteria towards areas high in nutrient substances and away from repellents. Chemoeffector molecules in the solution outside the bacterium bind to receptor molecules on the cell surface, and the resulting structural changes in the receptors are relayed in turn by the activities of the intracellular signaling proteins to generate a control signal for molecular motors that drive the bacterial flagella. The chemotactic network consists of about 10 nodes (here, signaling proteins), and the interactions between the nodes are the chemical reactions of methylation or phosphorylation. Notable features of this system include its extreme sensitivity, down to the limits set by counting individual molecules as they arrive at the cell surface [17], and the maintenance of this sensitivity across a huge dynamic range, through an adaptation mechanism that provides nearly perfect compensation of background concentrations [27]. More recently it has been appreciated that aspects of this functionality, such as perfect adaptation, are also robust against large variations in the concentrations of the network components [6].

Abstractly, different kinds of signaling proteins, such those in chemotaxis, can be thought of as the building blocks of a network, with their biochemical interactions forming the wiring diagram of the system, much like the components and wiring diagram of, for instance, a radio receiver. In principle, these wiring diagrams are hugely complex; for a network composed of N species, there are $\sim C_k^N$ possible connections among any set of k components, and typically we don't have direct experimental guidance about the numbers associated with each 'wire.' One approach is to view this as giant fitting problem: once we draw a network, there is a direct translation of this graph into dynamical equations, with many parameters, and we should test the predictions of these dynamics against whatever data are available to best determine the underlying parameters. Another approach is to ask whether this large collection of parameters is special in any way other than that it happens to fit the data – are there principles that allow us to predict how these systems *should* work? In the context of chemotaxis, we might imagine that network parameters have been selected to optimize the average progress of bacteria up the chemical gradients of nutrients, or to maximize the robustness of certain functions against extreme parameter variations. These ideas of design principles clearly are not limited to bacterial chemotaxis.

An important aspect of biological networks is that the same components (or components that have an easily identifiable evolutionary relationship) can be (re)used in different modules or used for the same function in a different way across species, as discussed for example by Rao et al. [118] for the case of bacterial chemotaxis. Furthermore, because evolutionary selection depends on function and not directly on microscopic details, different wiring diagrams or even changes in components themselves can result in the same performance; evolutionary process can gradually change the structure of the network as long as its function is preserved; as an example see the discussion of transcriptional regulation in yeast by Tanay et al. [148]. On the other hand, one can also expect that signal processing problems like gain control, noise reduction, ensuring (bi)stability etc, have appeared and were solved repeatedly, perhaps even in similar ways across various cellular functions, and we might be able to detect the traces of their commonality in the network structure, as for example in the discussion of local connectivity in bacterial transcriptional regulation by Shen–Orr et al. [136]. Thus there are reasons to believe that in addition to design prin-

ciples at the network level, there might also be local organizing principles, similar to common wiring motifs in electronic circuitry, yet still independent of the identity of the molecules that implement these principles.

Biological networks have been approached at many different levels, often by investigators from different disciplines. The basic wiring diagram of a network – the fact that a kinase phosphorylates these particular proteins, and not all others, or that a transcription factor binds to the promoter regions of particular genes – is determined by classical biochemical and structural concepts such as binding specificity. At the opposite extreme, trying to understand the collective behavior of the network as a whole suggests approaches from statistical physics, often looking at simplified models that leave out many molecular details. Analyses that start with design principles are yet a different approach, more in the 'top–down' spirit of statistical physics but leaving perhaps more room for details to emerge as the analysis is refined. Eventually, all of these different views need to converge: networks really are built out of molecules, their functions emerge as collective behaviors, and these functions must really be functions of use to the organism. At the moment, however, we seldom know enough to bridge the different levels of description, so the different approaches are pursued more or less independently, and we follow this convention here. We will start with the molecular building blocks, then look at models for networks as a whole, and finally consider design principles. We hope that this sequence doesn't leave the impression that we actually know how to build up from molecules to function!

Before exploring our subject in more detail, we take a moment to consider its boundaries. Our assignment from the editors was to focus on phenomena at the level of molecular and cellular biology. A very different approach attempts to create a 'science of networks' that searches for common properties in biological, social, economic and computer networks [104]. Even within the biological world, there is a significant divide between work on networks in cell biology and networks in the brain. As far as we can see this division is an artifact of history, since there are many issues which cut across these different fields. Thus, some of the most beautiful work on signaling comes from photoreceptors, where the combination of optical inputs and electrical outputs allowed, already in the 1970s, for experiments with a degree of quantitative analysis that even today is hard to match in systems which take chemical inputs and give outputs that modulate the expression levels of genes [14,121]. Similarly, problems of noise in the control of gene expression have parallels in the long history of work on noise in ion channels, as we have discussed elsewhere [156], and the problems of robustness have also been extensively explored in the network of interactions among the multiple species of ion channels in the membrane [51,88]. Finally, the ideas of collective behavior are much better developed in the context of neural networks than in cellular networks, and it is an open question how much can be learned by studying these different systems in the same language [151].

Biological Networks and Their Building Blocks

Genetic Regulatory Networks

Cells constantly adjust their levels of gene expression. One central mechanism in this regulatory process involves the control of transcription by proteins known as transcription factors (TFs), which locate and bind short DNA sequences in the regulated genes' promoter or enhancer regions. A given transcription factor can regulate either a few or a sizable proportion of the genes in a genome, and a single gene may be regulated by more than one transcription factor; different transcription factors can also regulate each other [166].

In the simplest case of a gene regulated by a single TF, the gene might be expressed whenever the factor – in this case called an activator – is bound to the cognate sequence in the promoter (which corresponds to the situation when the TF concentration in the nucleus is high), whereas the binding of a repressor would shut a normally active gene down. The outlines of these basic control principles were established long ago, well before the individual transcription factors could be isolated, in elegant experiments on the *lactose* operon of *Escherichia coli* [69] and even simpler model systems such as phage λ [115]. To a great extent the lessons learned from these experiments have provided the framework for understanding transcriptional control more generally, in prokaryotes [114], eukaryotes [75], and even during the development of complex multicellular organisms [8].

The advent of high throughput techniques for probing gene regulation has extended our reach beyond single genes. In particular, microarrays [30] and the related data analysis tools, such as clustering [36], have enabled researchers to find sets of genes, or *modules*, that are *coexpressed*, i. e. up- or down-regulated in a correlated fashion when the organism is exposed to different external conditions, and are thus probably regulated by the same set of transcription factors. Chromatin immunoprecipitation (ChIP) assays have made it possible to directly screen for short segments of DNA that known TFs bind; using microarray technology it is then possible to locate the intergenic regions which these segments belong to, and hence

find the regulated genes, as has recently been done for the *Saccharomyces cerevisiae* DNA-TF interaction map [86].

These high throughput experimental approaches, combined with traditional molecular biology and complemented by sequence analysis and related mathematical tools [139], provide a large scale, topological view of the transcriptional regulatory network of a particular organism, where each link between two nodes (genes) in the regulatory graph implies either activation or repression [5]. While useful for describing causal interactions and trying to predict responses to mutations and external perturbations [89], this picture does not explain how the network operates on a physical level: it lacks dynamics and specifies neither the strengths of the interactions nor how all the links converging onto a given node jointly exercise control over it. To address these issues, representative wild-type or simple synthetic regulatory elements and networks consisting of a few nodes have been studied extensively to construct quantitative models of the network building blocks.

For instance, combinatorial regulation of a gene by several transcription factors that bind and interact on the promoter has been considered by Buchler et al. [31] as an example of (binary) biological computation and synthetic networks implementing such computations have been created [56,170]. Building on classical work describing allosteric proteins such as hemoglobin, thermodynamic models have been used with success to account for combinatorial interactions on the operator of the λ phage [2]. More recently Bintu et al. [24,25] have reviewed the equilibrium statistical mechanics of such interactions, Setty et al. [134] have experimentally and systematically mapped out the response surface of the *lac* promoter to combinations of its two regulatory inputs, cAMP and IPTG, and Kuhlman et al. [85] have finally provided a consistent picture of the known experimental results and the thermodynamic model for the combinatorial regulation of the lactose operon. There have also been some successes in eukaryotic regulation, where Schroeder et al. [132] used thermodynamically motivated models to detect clusters of binding sites that regulate the gap genes in morphogenesis of the fruit fly.

Gene regulation is a dynamical process composed of a number of steps, for example the binding of TF to DNA, recruitment of transcription machinery and the production of the messenger RNA, post-transcriptional regulation, splicing and transport of mRNA, translation, maturation and possible localization of proteins. While the extensive palette of such microscopic interactions represents a formidable theoretical and experimental challenge for each detailed study, on a network level it primarily induces three effects. First, each node – usually understood as the amount of gene product – in a graph of regulatory interactions is really not a single dynamical variable, but has a nontrivial internal state representing the configuration on the associated promoter, concentration of the corresponding messenger RNA etc.; the relation of these quantities to the concentration of the output protein is not necessarily straightforward, as emphasized in recent work comparing mRNA and protein levels in yeast [46]. Second, collapsing multiple chemical species onto a single node makes it difficult to include non-transcriptional regulation of gene expression in the same framework. Third, the response of the target gene to changes in the concentrations of its regulators will be delayed and extended in time, as in the example explored by Rosenfeld and Alon [123].

Perhaps the clearest testimonies to the importance of dynamics in addition to network topology are provided by systems that involve regulatory loops, in which the output of a network feeds back on one of the inputs as an activator or repressor. McAdams and Shapiro [99] have argued that the time delays in genetic regulatory elements are essential for the proper functioning of the phage λ switch, while Elowitz and Leibler [38] have created a synthetic circuit made up of three mutually repressing genes (the "repressilator"), that exhibits spontaneous oscillations. Circadian clocks are examples of naturally occurring genetic oscillators [171].

In short, much is known about the skeleton of genetic regulatory interactions for model organisms, and physical models exist for several well studied (mostly prokaryotic) regulatory elements. While homology allows us to bridge the gap between model organisms and their relatives, it is less clear how and at which level of detail the knowledge about regulatory elements must be combined into a network to explain and predict its function.

Protein–Protein Interaction Networks

After having been produced, proteins often assemble into complexes through direct contact interactions, and these complexes are functionally active units participating in signal propagation and other pathways. Proteins also interact through less persistent encounters, as when a protein kinase meets its substrate. It is tempting to define a link in the network of protein–protein interactions by such physical associations, and this is the basis of several experimental methods which aim at a genome-wide survey of these interactions. Although starting out being relatively unreliable (with false positive rates of up to 50%), high throughput techniques like the yeast two hybrid assay [68,161] or mass spectrometry [45,61] are providing data of increasing quality about protein–protein in-

teractions, or the "interactome" [84]. While more reliable methods are being developed [5] and new organisms are being analyzed in this way [49,91,125], the existing interaction data from high throughput experiments and curated databases has already been extensively studied.

Interpretation of the interactions in the protein network is tricky, however, due to the fact that different experimental approaches have various biases – for example, mass spectrometry is biased towards detecting interactions between proteins of high abundance, while two hybrid methods seem to be unbiased in this regard; on the other hand, all methods show some degree of bias towards different cellular localizations and evolutionary novelty of the proteins. Assessing such biases, however, currently depends not on direct calibration of the methods themselves but on comparison of the results with manually curated databases, although the databases surely have their own biases [70]. It is reassuring that the intersection of various experimental results shows significantly improved agreement with the databases, but this comes at the cost of a substantial drop in coverage of the proteome [100].

In contrast to the case of transcriptional regulation, the relationship between two interacting proteins is symmetric: if protein A binds to protein B, B also binds to A, so that the network is described by an undirected graph. Most of the studies have been focused on binary interactions that yeast two hybrid and derived approaches can probe, although spectrometry can detect multiprotein complexes as well. Estimates of number of links in these networks vary widely, even in the yeast *Saccharomyces cerevisiae*: Krogan et al. [84] directly measure around 7100 interactions (between 2700 proteins), while Tucker et al. [158] estimate the total to be around 13 000–17 000, and von Mering et al. [100] would put the lower estimate at about 30 000. Apart from the experimental biases that can influence such estimates and have been discussed already, it is important to realize that each experiment can only detect interactions between proteins that are expressed under the chosen external conditions (e. g. the nutrient medium); moreover, interactions can vary from being transient to permanent, to which various measurement methods respond differently. It will thus become increasingly important to qualify each interaction in a graph by specifying how it depends on context in which the interaction takes place.

Proteins ultimately carry out most of the cellular processes such as transcriptional regulation, signal propagation and metabolism, and these processes can be modeled by their respective network and dynamical system abstractions. In contrast, the interactome is not a dynamical system itself, but instead captures specific reactions (like protein complex assembly) and structural and/or functional relations that are present in all of the above processes. In this respect it has an important practical role of annotating currently unknown proteins through 'guilt by association,' by tying them into complexes and processes with a previously known function.

Metabolic Networks

Metabolic networks organize our knowledge about anabolic and catabolic reactions between the enzymes, their substrates and co-factors (such as ATP), by reducing the set of reactions to a graph representation where two substrates are joined by a link if they participate in the same reaction. For model organisms like the bacterium *Escherichia coli* the metabolic networks have been studied in depth and are publicly available [77,78], and an increasing number of analyzed genomes offers sufficient sampling power to make statistical statements about the network properties across different domains of life [72].

Several important features distinguish metabolic from protein–protein interaction and transcriptional regulation networks. First, for well studied systems the coverage of metabolic reactions is high, at least for the central routes of energy metabolism and small molecule synthesis; notice that this is a property of our knowledge, not a property of the networks (!). Second, cellular concentrations of metabolites usually are much higher than those of transcription factors, making the stochasticity in reactions due to small molecular counts irrelevant. Third, knowledge of the stoichiometry of reactions allows one to directly write down a system of first order differential equations for the metabolite fluxes [60], which in steady state reduces to a set of linear constraints on the space of solutions. These chemical constraints go beyond topology and can yield strong and testable predictions; for example, Ibarra et al. [66] have shown how computationally maximizing the growth rate of *Escherichia coli* within the space of allowed solutions given by flux balance constraints can correctly predict measurable relationships between oxygen and substrate uptake, and that bacteria can be evolved towards the predicted optimality for growth conditions in which the response was initially suboptimal.

Signaling Networks

Signaling networks consist of receptor and signaling proteins that integrate, transmit and route information by means of chemical transformations of the network constituents. One class of such transformations, for example, are post–translational modifications, where targets are phosphorylated, methylated, acetylated, ... on spe-

cific residues, with a resulting change in their enzymatic (and thus signaling) activity. Alternatively, proteins might form stable complexes or dissociate from them, again introducing states of differential activity. The ability of cells to modify or tag proteins (possibly on several residues) can increase considerably the cell's capacity to encode its state and transmit information, assuming that the signaling proteins are highly specific not only for the identity but also the modification state of their targets; for a review see [110].

Despite the seeming overlap between the domains of protein–protein network and signaling networks, the focus of the analysis is substantially different. The interactome is simply a set of possible protein–protein interactions and thus a topological (or connectivity) map; in contrast, signaling networks aim to capture signal transduction and therefore need to establish a causal map, in which the nature of the protein–protein interaction, its direction and timescale, and its quantitative effect on the activity of the target protein matter. As an example, see the discussion by Kolch et al. [83] on the role of protein–protein interactions in MAPK signaling cascade.

Experiments on some signaling systems, such as the *Escherichia coli* chemotactic module, have generated enough experimental data to require detailed models in the form of dynamical equations. Molecular processes in a signaling cascade extend over different time scales, from milliseconds required for kinase and phosphatase reactions and protein conformational changes, to minutes or more required for gene expression control, cell movement and receptor trafficking; this fact, along with the (often essential) spatial effects such as the localization of signaling machinery and diffusion of chemical messengers, can considerably complicate analyses and simulations.

Signaling networks are often factored into pathways that have specific inputs, such as the ligands of the G protein coupled receptors on the cell surface, and specific outputs, as with pathways that couple to the transcriptional regulation apparatus or to changes in the intracellular concentration of messengers such as calcium or cyclic nucleotides. Nodes in signaling networks can participate in several pathways simultaneously, thus enabling signal integration or potentially inducing damaging "crosstalk" between pathways; how junctions and nodes process signals is an area of active research [74].

The components of signaling networks have long been the focus of biochemical research, and genetic methods allow experiments to assess the impact of knocking out or over-expressing particular components. In addition, several experimental approaches are being designed specifically for elucidating signaling networks. Ab-chips localize various signaling proteins on chips reminiscent of DNA microarrays, and stain them with appropriate fluorescent antibodies [105]. Multicolor flow cytometry is performed on cells immuno-stained for signaling protein modifications and hundreds of single cell simultaneous measurements of the modification state of pathway nodes are collected [113]. Indirect inference of signaling pathways is also possible from genomic or proteomic data.

One well studied signal transduction system is the mitogen activated protein kinase (MAPK) cascade that controls, among other functions, cell proliferation and differentiation [32]. Because this system is present in all eukaryotes and its structural components are used in multiple pathways, it has been chosen as a paradigm for the study of specificity and crosstalk. Similarly, the TOR system, identified initially in yeast, is responsible for integrating the information on nutrient availability, growth factors and energy status of the cell and correspondingly regulating the cell growth [95]. Another interesting example of signal integration and both intra- and inter-cellular communication is observed in the quorum sensing circuit of the bacterium *Vibrio harveyi*, where different kinds of species- and genus-specific signaling molecules are detected by their cognate receptors on the cell surface, and the information is fed into a common *Lux* phosphorelay pathway which ultimately regulates the quorum sensing genes [165].

Models of Biological Networks

Topological Models

The structural features of a network are captured by its connectivity graph, where interactions (reactions, structural relations) are depicted as the links between the interacting nodes (genes, proteins, metabolites). Information about connectivity clearly cannot and does not describe the network behavior, but it might influence and constrain it in revealing ways, similar to effect that the topology of the lattice has on the statistical mechanics of systems living on it.

Theorists have studied extensively the properties of regular networks and random graphs starting with Erdös and Rényi in 1960s. The first ones are characterized by high symmetry inherent in a square, triangular, or all-to-all (mean field) lattice; the random graphs were without such regularity, constructed simply by distributing K links at random between N nodes. The simple one–point statistical characterization that distinguishes random from regular networks looks at the node degree, that is the probability $P(k)$ that any node has k incoming and/or outgoing links. For random graphs this distribution is Poisson,

meaning that most of the nodes have degrees very close to the mean, $\langle k \rangle = \sum_k k\,P(k)$, although there are fluctuations; for regular lattices every node has the same connectivity to its neighbors.

The first analyses of the early reconstructions of large metabolic networks revealed a surprising "scale free" node degree distribution, that is $P(k) \sim k^{-\gamma}$, with γ between 2 and 3 for most networks. For the physics community, which had seen the impact of such scale invariance on our understanding of phase transitions, these observations were extremely suggestive. It should be emphasized that for many problems in areas as diverse as quantum field theory, statistical mechanics and dynamical systems, such scaling relations are much more than curiosities. Power laws relating various experimentally observable quantities are exact (at least in some limit), and the exponents (here, γ) really contain everything one might want to know about the nature of order in the system. Further, some of the first thoughts on scaling emerged from phenomenological analyses of real data. Thus, the large body of work on scaling ideas in theoretical physics set the stage for people to be excited by the experimental observation of power laws in much more complex systems, although it is not clear to us whether the implied promise of connection to a deeper theoretical structure has been fulfilled. For divergent views on these matters see Barabási et al. [10] and Keller et al. [81].

The most immediate practical consequence of a scale free degree distribution is that – relative to expectations based on random graphs – there will be an over-representation of nodes with very large numbers of links, as with pyruvate or co-enzyme A in metabolic networks [72,163]. These are sometimes called hubs, although another consequence of a scale free distribution is that there is no 'critical degree of connection' that distinguishes hubs from non-hubs. In the protein–protein interaction network of *Saccharomyces cerevisiae*, nodes with higher degree are more likely to represent essential proteins [73], suggesting that node degree does have some biological meaning. On the theoretical side, removal of a sizable fraction of nodes from a scale free network will neither increase the network diameter much, nor partition the network into equally sized parts [3], and it is tempting to think that this robustness is also biologically significant. The scale free property has been observed in many non-biological contexts, such as the topology of social interactions, World Wide Web links, electrical power grid connectivity … [144]. A number of models have been proposed for how such scaling might arise, and some of these ideas, such as growth by preferential attachment, have a vaguely biological flavor [11,12]. Finding the properties of networks that actually discriminate among different mechanisms of evolution or growth turns out to be surprisingly subtle [173].

Two other revealing measures are regularly computed for biological networks. The mean path length, $\langle l \rangle$, is the shortest path between a pair of nodes, averaged over all pairs in the graph, and measures the network's overall 'navigability.' Intuitively, short path lengths correspond to, for example, efficient or fast flow of information and energy in signaling or metabolic networks, quick spread of diseases in a social network and so on. The clustering coefficient of a node i is defined as $C_i = 2n_i/k_i(k_i - 1)$, where n_i is the number of links connecting the k_i neighbors of node i to each other; equivalently, C_i is the ratio between the number of triangles passing through two neighbors of i and node i itself, divided by the maximum possible number of such triangles. Random networks have low path lengths and low clustering coefficients, whereas regular lattices have long path lengths and are locally clustered. Watts and Strogatz [167] have constructed an intermediate regime of "small world" networks, where the regular lattice has been perturbed by a small number of random links connecting distant parts of the network together. These networks, although not necessarily scale free, have short path lengths and high clustering coefficients, a property that was subsequently observed in metabolic and other biological networks as well [163].

A high clustering coefficient suggests the existence of densely connected groups of nodes within a network, which seems contradictory to the idea of scale invariance, in which there is no inherent group or cluster size; Ravasz et al. [120] addressed this problem by introducing hierarchical networks and providing a simple construction for synthetic hierarchical networks exhibiting both scale free and clustering behaviors. Although there is no unique scale for the clusters, clusters will appear at any scale one chooses to look at, and this is revealed by the scaling of clustering coefficient $C(k)$ with the node degree k, $C(k) \sim k^{-1}$, on both synthetic as well as natural metabolic networks of organisms from different domains of life [120]. Another interesting property of some biological networks is an anti-correlation of node degree of connected nodes [96], which we can think of as a 'dissociative' structure; in contrast, for example, with the associative character of social networks, where well connected people usually know one another.

As we look more finely at the structure of the graph representing a network, there is of course a much greater variety of things to look at. For example, Spirin and Mirny [142] have focused on high clustering coefficients as a starting point and devised algorithms to search for modules, or densely connected subgraphs within the yeast

protein–protein interaction network. Although the problem has combinatorial complexity in general, the authors found about 50 modules (of 5–10 proteins in size, some of which were unknown at the time) that come in two types: the first represents dynamic functional units (e. g. signaling cascades), and the second protein complexes. A similar conclusion was reached by Han et al. [57], after having analyzed the interactome in combination with the temporal gene expression profiles and protein localization data; the authors argue that nodes of high degree can sit either at the centers of modules, which are simultaneously expressed ("party hubs"), or they can be involved in various pathways and modules at different times ("date hubs"). The former kind is at a lower level of organization, whereas the latter tie the network into one large connected component.

Focusing on even a smaller scale, Shen-Orr et al. [136] have explored motifs, or patterns of connectivity of small sets of nodes that are over-represented in a given network compared to the randomized networks of the same degree distribution $P(k)$. In the transcriptional network of the bacterium *E. coli*, three such motifs were found: feed forward loops (in which gene X regulates Y that regulates Z, but X directly regulates Z as well), single input modules (where gene X regulates a large number of other genes in the same way and usually auto-regulates itself), and dense overlapping regulons (layers of overlapping interactions between genes and a group of transcription factors, much denser than in randomized networks). The motif approach has been extended to combined network of transcriptional regulation and protein–protein interactions [169] in yeast, as well as to other systems [101].

At the risk of being overly pessimistic, we should conclude this section with a note of caution. It would be attractive to think that a decade of work on network topology has resulted in a coherent picture, perhaps of the following form: on the smallest scale, the nodes of biological networks are assembled into motifs, these in turn are linked into modules, and this continues in a hierarchical fashion until the entire network is scale free. As we will discuss again in the context of design principles, the notion of such discrete substructure – motifs and modules – is intuitively appealing, and some discussions suggest that it is essential either for the function or the evolution of networks. On the other hand, the evidence for such structure usually is gathered with reference to some null model (e. g., a random network with the same $P(k)$), so we don't even have an absolute definition of these structures, much less a measure of their sufficiency as a characterization of the whole system; for attempts at an absolute definition of modularity see Ziv et al. [174] and Hofman and Wiggins [62]. Similarly, while it is appealing to think about scale free networks, the evidence for scaling almost always is confined to less than two decades, and in practice scaling often is not exact. It is then not clear whether the idealization of scale invariance captures the essential structure in these systems.

Boolean Networks

A straightforward extension of the topological picture that also permits the study of network dynamics assumes that the entities at the nodes – for example, genes or signaling proteins – are either 'on' or 'off' at each moment of time, so that for node i the state at time t is $\sigma_i(t) \in \{0, 1\}$. Time is usually discretized, and an additional prescription is needed to implement the evolution of the system: $\sigma_i(t+1) = f_i(\{\sigma_\mu(t)\})$, where f_i is a function that specifies how the states of the nodes μ that are the inputs to node i in the interaction graph combine to determine the next state at node i. For instance, f_A might be a Boolean function for gene A, which needs to have its activator gene B present and repressor gene C absent, so that $\sigma_A(t+1) = \sigma_B(t) \wedge \bar{\sigma}_C(t)$. Alternatively, f might be a function that sums the inputs at state t with some weights, and then sets $\sigma_i = 1(0)$ if the result is above (below) a threshold, as in classical models of neural networks.

Boolean networks are amenable both to analytical treatment and to efficient simulation. Early on, Kauffman [80] considered the family of random boolean networks. In these models, each node is connected at random to K other nodes on average, and it computes a random Boolean function of its inputs in which a fraction ρ of the 2^K possible input combinations leads to $\sigma_i(t+1) = 1$. In the limit that the network is large, the dynamics are either regular (settling into a stable fixed cycle) or chaotic, and these two phases are separated by a separatrix $2\rho(1-\rho)K = 1$ in the phase space (ρ, K).

Aldana and Cluzel [4] have shown that for connectivities of $K \sim 20$ that could reasonably be expected in e. g. transcriptional regulatory networks, the chaotic regime dominates the phase space. They point out, however, that if the network is scale free, there is no 'typical' K as the distribution $P(k) \sim k^{-\gamma}$ does not have a well-defined mean for $\gamma \leq 3$ and the phase transition criterion must be restated. It turns out, surprisingly, that regular behavior is possible for values of γ between 2 and 2.5, observed in most biological networks, and this is exactly the region where the separatrix lies. Scale free architecture, at least for Boolean networks, seems to prevent chaos.

Several groups have used Boolean models to look at specific biological systems. Thomas [150] has established a theoretical framework in which current states of the

genes (as well as the states in the immediate past) and the environmental inputs are represented by Boolean variables that evolve through the application of Boolean functions. This work has been continued by, for example, Sanchez and Thieffry [128] who analyzed the gap-gene system of the fruit fly *Drosophila* by building a Boolean network that generates the correct levels of gene expression for 4 gap genes in response to input levels of 3 maternal morphogens with spatially varying profiles stretched along the anterior-posterior axis of the fly embryo. Interestingly, to reproduce the observed results and correctly predict the known *Drosophila* segmentation mutants, the authors had to introduce generalized Boolean variables that can take more than two states, and have identified the smallest necessary number of such states for each gene.

In a similar spirit, Li et al. [91] studied the skeleton of the budding yeast cell cycle, composed of 11 nodes, and a thresholding update rule. They found that the topology of this small network generates a robust sequence of transitions corresponding to known progression through yeast cell-cycle phases G1 (growth), S (DNA duplication), G2 (pause) and M (division), triggered by a known 'cell-size checkpoint.' This progression is robust, in the sense that the correct trajectory is the biggest dynamical attractor of the system, with respect to various choices of update rules and parameters, small changes in network topology, and choice of triggering checkpoints.

The usefulness of Boolean networks stems from the relative ease of implementation and simple parametrization of network topology and dynamics, making them suitable for studying medium or large networks. In addition to simplifying the states at the nodes to two (or more) discrete levels, which is an assumption that has not been clearly explored, one should be cautious that the discrete and usually synchronous dynamics in time can induce unwanted artifacts.

Probabilistic Models

Suppose one is able to observe simultaneously the activity levels of several proteins comprising a signaling network, or the expression levels of a set of genes belonging to the same regulatory module. Because they are part of a functional whole, the activity levels of the components will be correlated. Naively, one could build a network model by simply computing pairwise correlation coefficients between pairs of nodes, and postulating an interaction, and therefore a link, between the two nodes whenever their correlation is above some threshold. However, in a test case where A $\rightarrow$ B $\rightarrow$ C (gene A induces B which induces C), one expects to see high positive correlation among all three elements, even though there is no (physical) interaction between A and C. Correlation therefore is not equal to interaction or causation. Constructing a network from the correlations in this naive way also does not lead to a generative model that would predict the probabilities for observing different states of the network as a whole. Another approach is clearly needed; see Markowetz and Spang [94] for a review.

In the simple case where the activity of a protein/gene i can either be 'on' ($\sigma_i = 1$) or 'off' ($\sigma_i = 0$), the state of a network with N nodes will be characterized by a binary word of N bits, and because of interaction between nodes, not all these words will be equally likely. For example, if node A represses node B, then combinations such as $1_A 0_B \ldots$ or $0_A 1_B \ldots$ will be more likely than $1_A 1_B \ldots$. In the case of deterministic Boolean networks, having node A be 'on' would imply that node B is 'off' with certainty, but in probabilistic models it only means that there is a positive *bias* for node B to be 'off,' quantified by the probability that node B is 'off' given that the state of node A is known. Having this additional probabilistic degree of freedom is advantageous, both because the network itself might be noisy, and because the experiment can induce errors in the signal readout, making the inference of deterministic rules from observed binary patterns an ill-posed problem.

Once we agree to make a probabilistic model, the goal is to find the distribution over all network states, which we can also think of as the joint distribution of all the N variable that live on the nodes of the network, $P(\sigma_1, \ldots, \sigma_N | C)$, perhaps conditioned on some fixed set of environmental or experimental factors C. The activities of the nodes, σ_i, can be binary, can take on a discrete set of states, or be continuous, depending on our prior knowledge about the system and experimental and numerical constraints. Even for a modest N, experiments of realistic scale will not be enough to directly estimate the probability distribution, since even with binary variable the number of possible states, and hence the number of parameters required to specify the general probability distribution, grows as $\sim 2^N$. Progress thus depends in an essential way on simplifying assumptions.

Returning to the three gene example A $\rightarrow$ B $\rightarrow$ C, we realize that C depends on A only through B, or in other words, C is *conditionally independent* of A and hence no interaction should be assigned between nodes A and C. Thus, the joint distribution of three variables can be factorized,

$$P(\sigma_{\mathrm{A}}, \sigma_{\mathrm{B}}, \sigma_{\mathrm{C}}) = P(\sigma_{\mathrm{C}} | \sigma_{\mathrm{B}}) P(\sigma_{\mathrm{B}} | \sigma_{\mathrm{A}}) P(\sigma_{\mathrm{A}}).$$

One might hope that, even in a large network, these sorts

of conditional independence relations could be used to simplify our model of the probability distribution. In general this doesn't work, because of feedback loops which, in our simple example, would include the possibility that C affects the state of A, either directly or through some more circuitous path. Nonetheless one can try to make an approximation in which loops either are neglected or (more sensibly) taken into account in some sort of average way; in statistical mechanics, this approximation goes back at least to the work of Bethe [19].

In the computer science and bioinformatics literature, the exploitation of Bethe-like approximations has come to be known as 'Bayesian network modeling' [43]. In practice what this approach does is to search among possible network topologies, excluding loops, and then for fixed topology one uses the conditional probability relationships to factorize the probability distribution and fit the tables of conditional probabilities at each node that will best reproduce some set of data. Networks with more links have more parameters, so one must introduce a trade-off between the quality of the fit to the data and this increasing complexity. In this framework there is thus an explicit simplification based on conditional independence, and an implicit simplification based on a preference for models with fewer links or sparse connectivity.

The best known application of this approach to a biological network is the analysis of the MAPK signaling pathway in T cells from the human immune system [127]. The data for this analysis comes from experiments in which the phosophorylated states of 11 proteins in the pathway are sampled simultaneously by immunostaining [113], with hundreds of cells sampled for each set of external conditions. By combining experiments from multiple conditions, the Bayesian network analysis was able to find a network of interactions among the 11 proteins that has high overlap with those known to occur experimentally.

A very different approach to simplification of probabilistic models is based on the maximum entropy principle [71]. In this approach one views a set of experiments as providing an estimate of some set of correlations, for example the $\sim N^2$ correlations among all pairs of elements in the network. One then tries to construct a probability distribution which matches these correlations but otherwise has as little structure – as much entropy – as possible. We recall that the Boltzmann distribution for systems in thermal equilibrium can be derived as the distribution which has maximum entropy consistent with a given average energy, and maximum entropy modeling generalizes this to take account of other average properties. In fact one can construct a hierarchy of maximum entropy distributions which are consistent with higher and higher orders of correlation [130]. Maximum entropy models for binary variables that are consistent with pairwise correlations are exactly the Ising models of statistical physics, which opens a wealth of analytic tools and intuition about collective behavior in these systems.

In the context of biological networks (broadly construed), recent work has shown that maximum entropy models consistent with pairwise correlations are surprisingly successful at describing the patterns of activity among populations of neurons in the vertebrate retina as it responds to natural movies [131,153]. Similar results are obtained for very different retinas under different conditions [137], and these successes have touched a flurry of interest in the analysis of neural populations more generally. The connection to the Ising model has a special resonance in the context of neural networks, where the collective behavior of the Ising model has been used for some time as a prototype for thinking about the dynamics of computation and memory storage [64]; in the maximum entropy approach the Ising model emerges directly as the least structured model consistent with the experimentally measured patterns of correlation among pairs of cells. A particularly striking result of this analysis is that the Ising models which emerge seem to be poised near a critical point [153]. Returning to cell biology, the maximum entropy approach has also been used to analyze patterns of gene expression in yeast [90] as well as to revisit the MAPK cascade [151].

Dynamical Systems

If the information about a biological system is detailed enough to encompass all relevant interacting molecules along with the associated reactions and estimated reaction rates, and the molecular noise is expected to play a negligible role, it is possible to describe the system with rate equations of chemical kinetics. An obvious benefit is the immediate availability of mathematical tools, such as steady state and stability analyses, insight provided by nonlinear dynamics and chaos theory, well developed numerical algorithms for integration in time and convenient visualization with phase portraits or bifurcation diagrams. Moreover, analytical approximations can be often exploited productively when warranted by some prior knowledge, for example, in separately treating 'fast' and 'slow' reactions. In practice, however, reaction rates and other important parameters are often unknown or known only up to order-of-magnitude estimations; in this case the problem usually reduces to the identification of phase space regions where the behavior of the system is qualitatively the same, for example, regions where the system exhibits limit-cycle oscil-

lations, bistability, convergence into a single steady state etc.; see Tyson et al. [159] for a review. Despite the difficulties, deterministic chemical kinetic models have been very powerful tools in analyzing specific network motifs or regulatory elements, as in the protein signaling circuits that achieve perfect adaptation, homeostasis, switching and so on, described by Tyson et al. [160], and more generally in the analysis of transcriptional regulatory networks as reviewed by Hasty et al. [59].

In the world of bacteria, some of the first detailed computer simulations of the chemotaxis module of *Escherichia coli* were undertaken by Bray et al. [29]. The signaling cascade from the Tar receptor at the cell surface to the modifications in the phosphorylation state of the molecular motor were captured by Michaelis–Menten kinetic reactions (and equilibrium binding conditions for the receptor), and the system of equations was numerically integrated in time. While slow adaptation kinetics was not studied in this first effort, the model nevertheless qualitatively reproduces about 80 percent of examined chemotactic protein deletion and overexpression mutants, although the extreme sensitivity of the system remained unexplained.

In eukaryotes, Novak and Tyson [107] have, for instance, constructed an extensive model of cell cycle control in fission yeast. Despite its complexity (~ 10 proteins and ~ 30 rate constants), Novak and colleagues have provided an interpretation of the system in terms of three interlocking modules that regulate the transitions from G1 (growth) into S (DNA synthesis) phase, from G2 into M (division) phase, and the exit from mitosis, respectively. The modules are coupled through cdc2/cdc13 protein complex and the system is driven by the interaction with the cell size signal (proportional to the number of ribosomes per nucleus). At small size, the control circuit can only support one stable attractor, which is the state with low cdc2 activity corresponding to G1 phase. As the cell grows, new stable state appears and the system makes an irreversible transition into S/G2 at a bifurcation point, and, at an even larger size, the mitotic module becomes unstable and executes limit cycles in cdc2-cdc13 activity until the M phase is completed and the cell returns to its initial size. The basic idea is that the cell, driven by the the size readout, progresses through robust cell states created by bistability in the three modules comprising the cell cycle control – in this way, once it commits to a transition from G2 state into M, small fluctuations will not flip it back into G2. The mathematical model has in this case successfully predicted the behaviors of a number of cell cycle mutants and recapitulated experimental observations collected during 1970s and 1980s by Nurse and collaborators [108].

The circadian clock is a naturally occurring transcriptional module that is particularly amenable to dynamical systems modeling. Leloup and Goldbeter [87] have created a mathematical model of a mammalian clock (with ~ 20 rate equations) that exhibits autonomous sustained oscillations over a sizable range of parameter values, and reproduces the entrainment of the oscillations to the light–dark cycles through light-induced gene expression. The basic mechanism that enables the cyclic behavior is negative feedback transcriptional control, although the actual circuit contains at least two coupled oscillators. Studying circadian clock in mammals, the fruit fly *Drosophila* or *Neurospora* is attractive because of the possibility of connecting a sizable cataloge of physiological disorders in circadian rhythms to malfunctions in the clock circuit and direct experimentation with light-dark stimuli [171]. Recent experiments indicate that at least in cyanobacteria the circadian clock can be reconstituted from a surprisingly small set of biochemical reactions, without transcription or translation [102,157], and this opens possibilities for even simpler and highly predictive dynamical models [126].

Dynamical modeling has in addition been applied to many smaller systems. For example, the construction of a synthetic toggle switch [44], and the 'repressilator' – oscillating network of three mutually repressing genes [38] – are examples where mathematical analysis has stimulated the design of synthetic circuits. A successful reaction-diffusion model of how localization and complex formation of Min proteins can lead to spatial limit cycle oscillations (used by *Escherichia coli* to find its division site) was constructed by Huang et al. [65]. It remains a challenge, nevertheless, to navigate in the space of parameters as it becomes ever larger for bigger networks, to correctly account for localization and count various forms of protein modifications, especially when the signaling networks also couple to transcriptional regulation, and to find a proper balance between models that capture all known reactions and interactions and phenomenological models that include coarse-grained variables.

Stochastic Dynamics

Stochastic dynamics is in principle the most detailed level of system description. Here, the (integer) count of every molecular species is tracked and reactions are drawn at random with appropriate probabilities per unit time (proportional to their respective reaction rates) and executed to update the current tally of molecular counts. An algorithm implementing this prescription, called the stochastic simulation algorithm or SSA, was devised by Gille-

spie [47]; see Gillespie [48] for a review of SSA and a discussion of related methods. Although slow, this approach for simulating chemical reactions can be made exact. In general, when all molecules are present in large numbers and continuous, well-mixed concentrations are good approximations, the (deterministic) rate dynamics equations and stochastic simulation give the same results; however, when molecular counts are low and, consequently, the stochasticity in reaction timing and ordering becomes important, the rate dynamics breaks down and SSA needs to be used. In biological networks and specifically in transcriptional regulation, a gene and its promoter region are only present in one (or perhaps a few) copies, while transcription factors that regulate it can also be at nanomolar concentrations (i. e. from a few to a few hundred molecules per nucleus), making stochastic effects possibly very important [97,98].

One of the pioneering studies of the role of noise in a biological system was a simulation of the phage λ lysis-lysogeny switch by Arkin et al. [7]. The life cycle of the phage is determined by the concentrations of two transcription factors, *cI* (lambda repressor) and *cro*, that compete for binding to the same operator on the DNA. If *cI* is prevalent, the phage DNA is integrated into the host's genome and no phage genes except for *cI* are expressed (the lysogenic state); if *cro* is dominant, the phage is in lytic state, using cell's DNA replication machinery to produce more phages and ultimately lyse the host cell [115]. The switch is bistable and the fate of the phage depends on the temporal and random pattern of gene expression of two mutually antagonistic transcription factors, although the balance can be shifted by subjecting the host cell to stress and thus flipping the toggle into lytic phase. The stochastic simulation correctly reproduces the experimentally observed fraction of lysogenic phages as a function of multiplicity-of-infection. An extension of SSA to spatially extended models is possible.

Although the simulations are exact, they are computationally intensive and do not offer any analytical insight into the behavior of the solutions. As a result, various theoretical techniques have been developed for studying the effects of stochasticity in biological networks. These are often operating in a regime where the deterministic chemical kinetics is a good approximation, and noise (i. e. fluctuation of concentrations around the mean) is added into the system of differential equations as a perturbation; these Langevin methods have been useful for the study of noise propagation in regulatory networks [76,111,149]. The analysis of stochastic dynamics is especially interesting in the context of design principles which consider the reliability of network function, to which we return below.

Network Properties and Operating Principles

Modularity

Biological networks are said to be modular, although the term has several related but nevertheless distinct meanings. Their common denominator is the idea that there exist a partitioning of the network nodes into groups, or modules, that are largely independent of each other and perform separate or autonomous functions. Independence can be achieved through spatial isolation of the module's processes or by chemical specificity of its components. The ability to extract the module from the cell and reconstitute it in vitro, or transplant it to another type of cell is a powerful argument for the existence of modularity [58]. In the absence of such strong and laborious experimental verifications, however, measures of modularity that depend on a particular network model are frequently used.

In topological networks, the focus is on the module's independence: nodes within a module are densely connected to each other, while inter-modular links are sparse [57,120,142] and the tendency to cluster is measured by high clustering coefficients. As a caveat to this view note that despite their sparseness the inter-module links could represent strong dynamical couplings. Modular architecture has been studied in Boolean networks by Kashtan and Alon [79], who have shown that modularity can evolve by mutation and selection in a time-varying fitness landscape where changeable goals decompose into a set of fixed subproblems. In the example studied they computationally evolve networks implementing several Boolean formulae and observe the appearance of a module – a circuit of logical gates implementing a particular Boolean operator (like XOR) in a reusable way. This work makes clear that modularity in networks is plausibly connected to modularity in the kinds of problems that these networks were selected to solve, but we really know relatively little about the formal structure of these problems.

There are also ways of inferring a form of modularity directly without assuming any particular network model. Clustering tools partition genes into co-expressed groups, or clusters, that are often identified with particular modules [36,133,140]. Ihmels et al. [67] have noted that each node can belong to more than one module depending on the biological state of the cell, or the context, and have correspondingly reexamined the clustering problem. Elemento et al. [37] have recently presented a general information theoretic approach to inferring regulatory modules and the associated transcription factor binding sites from various kinds of high-throughput data. While clustering methods have been widely applied in the exploration of

gene expression, it should be emphasized that merely finding clusters does not by itself provide evidence for modularity. As noted above, the whole discussion would be much more satisfying if we had independent definitions of modularity and, we might add, clearly stated alternative hypotheses about the structure and dynamics of these networks.

Focusing on the functional aspect of the module, we often observe that the majority of the components of a system (for instance, a set of promoter sites or a set of genes regulating motility in bacteria) are conserved together across species. These observations support the hypothesis that the conserved components are part of a very tightly coupled sub-network which we might identify as a module. Bioinformatic tools can then use the combined sequence and expression data to give predictions about modules, as reviewed by Siggia et al. [139]. Purely phylogenetic approaches that infer module components based on inter-species comparisons have also been productive and can extract candidate modules based only on phylogenetic footprinting, that is, studying the presence or absence of homologous genes across organisms and correlating their presence with hand annotated phenotypic traits [141].

Robustness

Robustness refers to a property of the biological network such that some aspect of its function is not sensitive to perturbations of network parameters, environmental variables (e. g. temperature), or initial state; see de Visser et al. [162] for a review of robustness from an evolutionary perspective and Goulian [53] for mechanisms of robustness in bacterial circuits. Robustness encompasses two very different ideas. One idea has to do with a general principle about the nature of explanation in the quantitative sciences: qualitatively striking facts should not depend on the fine tuning of parameters, because such a scenario just shifts the problem to understanding why the parameters are tuned as they are. The second idea is more intrinsic to the function of the system, and entails the hypothesis that cells cannot rely on precisely reproducible parameters or conditions and must nonetheless function reliably and reproducibly.

Robustness has been studied extensively in the chemotactic system of the bacterium *Escherichia coli*. The systematic bias to swim towards chemoattractants and away from repellents can only be sustained if the bacterium is sensitive to the spatial gradients of the concentration and not to its absolute levels. This discriminative ability is ensured by the mechanism of perfect adaptation, with which the proportion of bacterial straight runs and tumbles (random changes in direction) always returns to the same value in the absence of gradients [27]. Naively, however, the ability to adapt perfectly seems to be sensitive to the amounts of intracellular signaling proteins, which can be tuned only approximately by means of transcriptional regulation. Barkai and Leibler [13] argued that there is integral feedback control in the chemotactic circuit which makes it robust against changes in these parameters, and Alon et al. [6] showed experimentally that precision of adaptation truly stays robust, while other properties of the systems (such as the time to adapt and the steady state) show marked variations with intracellular signaling protein concentrations.

One seemingly clear example of robust biological function is embryonic development. We know that the spatial structure of the fully developed organism follows a 'blueprint' laid out early in development as a spatial pattern of gene expression levels. von Dassow et al. [34] studied one part of this process in the fruit fly *Drosophila*, the 'segment polarity network' that generates striped patterns of expression. They considered a dynamical system based on the wiring diagram of interactions among a small group of genes and signaling molecules, with ~ 50 associated constants parametrizing production and degradation rates, saturation response and diffusion, and searched the parameter space for solutions that reproduce the known striped patterns. They found that, with their initial guess at network topology, such solutions do not exist, but adding a particular link – biologically motivated though unconfirmed at the time – allowed them to find solutions by random sampling of parameter space. Although they presented no rigorous measure for the volume of parameter space in which correct solutions exist, it seems that a wide variety of parameter choices and initial conditions indeed produce striped expression patterns, and this was taken to be a signature of robustness.

Robustness in dynamical models is the ability of the biological network to sustain its trajectory through state space despite parameter or state perturbations. In circadian clocks the oscillations have to be robust against both molecular noise inherent in transcriptional regulation, examined in stochastic simulations by Gonze et al. [52], as well as variation in rate parameters [143]; in the latter work the authors introduce integral robustness measures along the trajectory in state space and argue that the clock network architecture tends to concentrate the fragility to perturbations into parameters that are global to the cell (maximum overall translation and protein degradation rates) while increasing the robustness to processes specific to the circadian oscillator. As was mentioned earlier, robustness to state perturbations was demonstrated by Li et al. [91] in

the threshold binary network model of the yeast cell cycle, and examined in scale-free random Boolean networks by Aldana and Cluzel [4].

As with modularity, robustness has been somewhat resistant to rigorous definitions. Importantly, robustness has always been used as a relational concept: function X is robust to variations in Y. An alternative to robustness is for the organism to exert precise control over Y, perhaps even using X as a feedback signal. This seems to be how neurons stabilize a functional mix of different ion channels [93], following the original theoretical suggestion of LeMasson et al. [88]. Pattern formation during embryonic development in *Drosophila* begins with spatial gradients of transcription factors, such as Bicoid, which are established by maternal gene expression, and it has been assumed that variations in these expression levels are inevitable, requiring some robust readout mechanism. Recent measurements of Bicoid in live embryos, however, demonstrate that the absolute concentrations are actually reproducible from embryo to embryo with $\sim 10\%$ precision [54]. While there remain many open questions, these results suggest that organisms may be able to exert surprisingly exact control over critical parameters, rather than having compensation schemes for initially sloppy mechanisms. The example of ion channels alerts us to the possibility that cells may even 'know' which combinations of parameters are critical, so that variations in a multidimensional parameter space are large, but confined to a low dimensional manifold.

Noise

A dynamical system with constant reaction rates, starting repeatedly from the same initial condition in a stable environment, always follows a deterministic time evolution. When the concentrations of the reacting species are low enough, however, the description in terms of time (and possibly space) dependent concentration breaks down, and the stochasticity in reactions, driven by random encounters between individual molecules, becomes important: on repeated trials from the same initial conditions, the system will trace out different trajectories in the state space. As has been pointed out in the section on stochastic dynamics, biological networks in this regime need to be simulated with the Gillespie algorithm [47], or analyzed within approximate schemes that treat noise as perturbation of deterministic dynamics. Recent experimental developments have made it possible to observe this noise directly, spurring new research in the field. Noise in biological networks fundamentally limits the organism's ability to sense, process and respond to environmental and internal signals, suggesting that analysis of noise is a crucial component in any attempt to understand the design of these networks. This line of reasoning is well developed in the context of neural function [20], and we draw attention in particular to work on the ability of the visual system to count single photons, which depends upon the precision of the G-protein mediated signaling cascade in photo receptors; see, for example, [117].

Because transcriptional regulation inherently deals with molecules, such as DNA and transcription factors, that are present at low copy numbers, most noise studies were carried out on transcriptional regulatory elements. The availability of fluorescent proteins and their fusions to wild type proteins have been the crucial tools, enabling researchers to image the cells expressing these probes in a controllable manner, and track their number in time and across the population of cells. Elowitz et al. [39] pioneered the idea of observing the output of two identical regulatory elements driving the expression of two fluorescent proteins of different colors, regulated by a common input in a single *Escherichia coli* cell. In this 'two-color experiment,' the correlated fluctuations in both colors must be due to the *extrinsic* fluctuations in the common factors that influence the production of both proteins, such as overall RNA polymerase or transcription factor levels; on the other hand, the remaining, uncorrelated fluctuation is due to the *intrinsic* stochasticity in the transcription of the gene and translation of the messenger RNA into the fluorescent protein from each of the two promoters [147]. Ozbudak et al. [109] have studied the contributions of stochasticity in transcription and translation to the total noise in gene expression in prokaryotes, while Pedraza and van Oudenaarden [112] and Hooshangi et al. [63] have looked at the propagation of noise from transcription factors to their targets in synthetic multi-gene cascades. Rosenfeld et al. [124] have used the statistics of binomial partitioning of proteins during the division of *Escherichia coli* to convert their fluorescence measurements into the corresponding absolute protein concentrations, and also were able to observe the dynamics of these fluctuations, characterizing the correlation times of both intrinsic and extrinsic noise.

Theoretical work has primarily been concerned with disentangling and quantifying the contributions of different steps in transcriptional regulation and gene expression to the total noise in the regulated gene [111,146,149], often by looking for signatures of various noise sources in the behavior of the measured noise as a function of the mean expression level of a gene. For many of the examples studied in prokaryotes, noise seemed to be primarily attributable to the production of proteins in bursts from single messenger RNA molecules, and to pulsatile and ran-

dom activation of genes and therefore bursty translation into mRNA [50]. In yeast [26,119] and in mammalian cells [116] such stochastic synthesis of mRNA was modeled and observed as well. Simple scaling of noise with the mean was observed in $\sim$40 yeast proteins under different conditions by Bar-Even et al. [9] and interpreted as originating in variability in mRNA copy numbers or gene activation.

Bialek and Setayeshgar [22] have demonstrated theoretically that at low concentrations of transcriptional regulator, there should be a lower bound on the noise set by the basic physics of diffusion of transcription factor molecules to the DNA binding sites. This limit is independent of (possibly complex, and usually unknown) molecular details of the binding process; as an example, cooperativity enhances the 'sensitivity' to small changes in concentration, but doesn't lower the physical limit to noise performance [23]. This randomness in diffusive flux of factors to their 'detectors' on the DNA must ultimately limit the precision and reliability of transcriptional regulation, much like the randomness in diffusion of chemoattractants to the detectors on the surface of *Escherichia coli* limits its chemotactic performance [17]. Interestingly, one dimensional diffusion of transcription factors along the DNA can have a big impact on the speed with which TFs find their targets, but the change in noise performance that one might expect to accompany these kinetic changes is largely compensated by the extended correlation structure of one dimensional diffusion [152]. Recent measurements of the regulation of the *hunchback* gene by Bicoid during early fruit fly development by Gregor et al. [54] have provided evidence for the dominant role of such input noise, which coexists with previously studied output noise in production of mRNA and protein [156]. These results raise the possibility that initial decisions in embryonic development are made with a precision limited by fundamental physical principles.

Dynamics, Attractors, Stability and Large Fluctuations

The behavior of a dynamical system as the time tends to infinity, in response to a particular input, is interesting regardless of the nature of the network model. Both discrete and continuous, or deterministic and noisy, systems can settle into a number of fixed points, exhibit limit-cycle oscillations, or execute chaotic dynamics. In biological networks it is important to ask whether these qualitatively different outcomes correspond to distinct phenotypes or behaviors. If so, then a specific stable gene expression profile in a network of developmental genes, for example, encodes that cell's developmental fate, as the amount of lambda repressor encodes the state of lysis vs lysogeny switch in the phage. The history of the system that led to the establishment of a specific steady state would not matter as long as the system persisted in the same attractor: the dynamics could be regarded as a 'computation' leading to the final result, the identity of the attractor, with the activities of genes in this steady state in turn driving the downstream pathways and other modules; see Kauffman [80] for genetic networks and Hopfield [64] for similar ideas in neural networks for associative memory. Alternatively, such partitioning into transient dynamics and 'meaningful' steady states might not be possible: the system must be analyzed as a whole while it moves in state space, and parts of it do not separately and sequentially settle into their attactors.

It seems, for example, that qualitative behavior of the cell cycle can be understood by progression through well-defined states or checkpoints: after transients die away, the cell cycle proteins are in a 'consistent' state that regulates division or growth related activities, so long as the conditions do not warrant a new transition into the next state [33,103]. In the fruit fly *Drosophila* development it has been suggested that combined processes of diffusion and degradation first establish steady-state spatial profiles of maternal morphogens along the major axis of the embryo, after which this stable 'coordinate system' is read out by gap and other downstream genes to generate the body segments. Recent measurements by Gregor et al. [55] have shown that there is a rich dynamics in the Bicoid morphogens concentration, prompting Bergmann et al. [18] to hypothesize that perhaps downstream genes read out and respond to morphogens even before the steady state has been reached. On another note, an interesting excitable motif, called the "feedback resistor," has been found in HIV Tat system – instead of having a bistable switch like the λ phage, HIV (which lacks negative feedback capability) implements a circuit with a single stable 'off' lysogenic state, that is perturbed in a pulse of trans activation when the virus attacks. The pulse probably triggers a threshold-crossing process that drives downstream events, and subsequently decays away; the feedback resistor is thus again an example of a dynamic, as opposed to the steady-state, readout [168]. Excitable dynamics are of course at the heart of the action potential in neurons, which results from the coupled dynamics of ion channel proteins, and related dynamical ideas are now emerging other cellular networks [145].

If attractors of the dynamical system correspond to distinct biological states of the organism, it is important to examine their stability against noise-induced spontaneous flipping. Bistable elements are akin to the 'flip-flop'

switches in computer chips – they form the basis of cellular (epigenetic) memory. While this mechanism for remembering the past is not unique – for example, a very slow, but not bistable, dynamics will also retain 'memory' of the initial condition through protein levels that persist on a generation time scale [138], it has the potential to be the most stable mechanism. The naturally occurring bistable switch of the λ phage was studied using stochastic simulation by Arkin et al. [7], and a synthetic toggle switch was constructed in *Escherichia coli* by Gardner et al. [44]. Theoretical studies of systems where large fluctuations are important are generally difficult and restricted to simple regulatory elements, but Bialek [21] has shown that a bistable switch can be created with as few as tens of molecules yet remain stable for years. A full understanding of such stochastic switching brings in powerful methods from statistical physics and field theory [122,129,164], ultimately with the hope of connecting to quantitative experiments [1].

Optimization Principles

If the function of a pathway or a network module can be quantified by a scalar measure, it is possible to explore the space of networks that perform the given function optimally. An example already given was that of maximizing the growth rate of the bacterium *Escherichia coli*, subject to the constraints imposed by the known metabolic reactions of the cell; the resulting optimal joint usage of oxygen and food could be compared to the experiments [66]. If enough constraints exist for the problem to be well posed, and there is sufficient reason to believe that evolution drove the organism towards optimal behavior, optimization principles allow us to both tune the otherwise unknown parameters to achieve the maximum, and also to compare the wild type and optimal performances.

Dekel and Alon [35] have performed the cost/benefit analysis of expressing *lac* operon in bacteria. On one hand *lac* genes allow *Escherichia coli* to digest lactose, but on the other there is the incurred metabolic cost to the cell for expressing them. That the cost is not negligible to the bacterium is demonstrated best by the fact that it shuts off the operon if no lactose is present in the environment. The cost terms are measured by inducing the *lac* operon with changeable amount of IPTG that provides no energy in return; the benefit is measured by fully inducing *lac* with IPTG and supplying variable amounts of lactose; both cost and benefit are in turn expressed as the change in the growth rate compared to the wild-type grown at fixed conditions. Optimal levels of *lac* expression were then predicted as a function of lactose concentration and bacteria were evolved for several hundred generations to verify that evolved organisms lie close to the predicted optimum.

Zaslaver et al. [172] have considered a cascade of amino-acid biosynthesis reactions in *Escherichia coli*, catalyzed by their corresponding enzymes. They have then optimized the parameters of the model that describes the regulation of enzyme gene expression, such that the total metabolic cost for enzyme production was balanced against the benefit of achieving a desired metabolic flux through the biosynthesis pathway. The resulting optimal on-times and promoter activities for the enzymes were compared to the measured activities of amino-acid biosynthesis promoters exposed to different amino-acids in the medium. The authors conclude that the bacterium implements a 'just-in-time' transcription program, with enzymes catalyzing initial steps in the pathway being produced from strong and low-latency promoters.

In signal transduction networks the definition of an objective function to be maximized is somewhat more tricky. The ability of the cell to sense its environment and make decisions, for instance about which genes to up- or down-regulate, is limited by several factors: scarcity of signals coming from the environment, perhaps because of the limited time that can be dedicated to data collection; noise inherent in the signaling network that degrades the quality of the detected signal; (sub-)optimality of the decision strategy; and noise in the effector systems at the output. A first idea would be to postulate that networks are designed to lower the noise, and intuitively the ubiquity of mechanisms such as negative feedback [15,53] is consistent with such an objective. There are various definitions for noise, however, which in addition are generally a function of the input, raising serious issues about how to formulate a principled optimization criterion.

When we think about energy flow in biological systems, there is no doubt that our thinking must at least be consistent with thermodynamics. More strongly, thermodynamics provides us with notions of efficiency that place the performance of biological systems on an absolute scale, and in many cases this performance really is quite impressive. In contrast, most discussions of information in biological systems leave "information" as a colloquial term, making no reference to the formal apparatus of information theory as developed by Shannon and others more than fifty years ago [135]. Although many aspects of information theory that are especially important for modern technology (e. g., sophisticated error-correcting codes) have no obvious connection to biology, there is something at the core of information theory that is vital: Shannon proved that if we want to quantify the intuitive concept that "x provides information about y," then there

is only one way to do this that is guaranteed to work under all conditions and to obey simple intuitive criteria such as the additivity of independent information. This unique measure of "information" is Shannon's mutual information. Further, there are theorems in information theory which, in parallel to results in thermodynamics, provide us with limits to what is possible and with notions of efficiency.

There is a long history of using information theoretic ideas to analyze the flow of information in the nervous system, including the idea that aspects of the brain's coding strategies might be chosen to optimize the efficiency of coding, and these theoretical ideas have led directly to interesting experiments. The use of information to think about cellular signaling and its possible optimization is more recent [154,175]. An important aspect of optimizing information flow is that the input/output relations of signaling devices must be matched to the distribution of inputs, and recent measurements on the control of *hunchback* by Bicoid in the early fruit fly embryo [54] seem remarkably consistent with the (parameter free) predictions from these matching relations [155].

In the context of neuroscience there is a long tradition of forcing the complex dynamics of signal processing into a setting where the subject needs to decide between a small set of alternatives; in this limit there is a well developed theory of optimal Bayesian decision making, which uses prior knowledge of the possible signals to help overcome noise intrinsic to the signaling system; Libby et al. [92] have recently applied this approach to the *lac* operon in *Escherichia coli*. The regulatory element is viewed as an inference module that has to 'decide,' by choosing its induction level, if the environmental lactose concentration is high or low. If the bacterium detects a momentarily high sugar concentration, it has to discriminate between two situations: either the environment really is at low overall concentration but there has been a large fluctuation; or the environment has switched to a high concentration mode. The authors examine how plausible regulatory element architectures (e. g. activator vs repressor, cooperative binding etc.) yield different discrimination performance. Intrinsic noise in the *lac* system can additionally complicate such decision making, but can be included into the theoretical Bayesian framework.

The question of whether biological systems are optimal in any precise mathematical sense is likely to remain controversial for some time. Currently opinions are stronger than the data, with some investigators using 'optimized' rather loosely and others convinced that what we see today is only a historical accident, not organizable around such lofty principles. We emphasize, however, that attempts to formulate optimization principles require us to articulate clearly what we mean by "function" in each context, and this is an important exercise. Exploration of optimization principles also exposes new questions, such as the nature of the distribution of inputs to signaling systems, that one might not have thought to ask otherwise. Many of these questions remain as challenges for a new generation of experiments.

Evolvability and Designability

Kirschner and Gerhart [82] define *evolvability* as an organism's capacity to generate heritable phenotypic variation. This capacity may have two components: first, to reduce the lethality of mutations, and second, to reduce the number of mutations needed to produce phenotypically novel traits. The systematic study of evolvability is hard because the genotype-to-phenotype map is highly nontrivial, but there have been some qualitative observations relevant to biological networks. Emergence of *weak linkage* of processes, such as the co-dependence of transcription factors and their DNA binding sites in metazoan transcriptional regulation, is one example. Metazoan regulation seems to depend on combinatorial control by many transcription factors with weak DNA-binding specificities and the corresponding binding sites (called cis-regulatory modules) can be dispersed and extended on the DNA. This is in stark contrast to the strong linkage between the factors and the DNA in prokaryotic regulation or in metabolism, energy transfer or macromolecular assembly, where steric and complementarity requirements for interacting molecules are high. In protein signaling networks, strongly conserved but flexible proteins, like calmodulin, can bind weakly to many other proteins, with small mutations in their sequence probably affecting such binding and making the establishment of new regulatory links possible and perhaps easy.

Some of the most detailed attempts to follow the evolution of network function have been by Francois and coworkers [41,42]. In their initial work they showed how simple functional circuits, performing logical operations or implementing bistable or oscillatory behavior, can be reliably created by a mutational process with selection by an appropriate fitness function. More recently they have considered fitness functions which favor spatial structure in patterns of gene expression, and shown how the networks that emerge from dynamics in this fitness landscape recapitulate the outlines of the segmentation networks known to be operating during embryonic development.

Instead of asking if there *exists* a network of nodes such that they perform a given computation, and if it can be

found by mutation and selection as in the examples above, one can ask how many network topologies perform a given computation. In other words, one is asking whether there is only one (fine tuned?) or many topologies or solutions to a given problem. The question of how many network topologies, proxies for different genotypes, produce the same dynamics, a proxy for phenotype, is a question of designability, a concept originally proposed to study the properties of amino-acid sequences comprising functional proteins, but applicable also to biological regulatory networks [106]. The authors examine three- and four-node binary networks with threshold updating rule and show that all networks with the shared phenotype have a common 'core' set of connections, but can differ in the variable part, similar to protein folding where the essential set of residues is necessary for the fold, with numerous variations in the nonessential part.

Future Directions

The study of biological networks is at an early stage, both on the theoretical as well as on the experimental side. Although high-throughput experiments are generating large data sets, these can suffer from serious biases, lack of temporal or spatial detail, and limited access to the component parts of the interacting system. On a theoretical front, general analytical insights that would link dynamics with network topology are few, although for specific systems with known topology computer simulation can be of great assistance. There can be confusion about which aspects of the dynamical model have biological significance and interpretation, and which aspects are just 'temporary variables' and the 'envelope' of the proverbial back-of-the-envelope calculations that cells use to perform their biological computations on; which parts of the trajectory are functionally constrained and which ones could fluctuate considerably with no ill-effects; how much noise is tolerable in the nodes of the network and what is its correlation structure; or how the unobserved, or 'hidden,' nodes (or their modification/activity states) influence the network dynamics.

Despite these caveats, cellular networks have some advantages over biological systems of comparable complexity, such as neural networks. Due to technological developments, we are considerably closer to the complete census of the interacting molecules in a cell than we are generally to the picture of connectivity of the neural tissue. Components of the regulatory networks are simpler than neurons, which are capable of a range of complicated behaviors on different timescales. Modules and pathways often comprise smaller number of interacting elements than in neural networks, making it possible to design small but interesting synthetic circuits. Last but not least, sequence and homology can provide strong insights or be powerful tools for network inference in their own right.

Those of us who come from the traditionally quantitative sciences, such as physics, were raised with experiments in which crucial elements are isolated and controlled. In biological systems, attempts at such isolation may break the regulatory mechanisms that are essential for normal operation of the system, leaving us with a system which is in fact more variable and less controlled than we would have if we faced the full complexity of the organism. It is only recently that we have seen the development of experimental techniques that allow fully quantitative, real time measurements of the molecular events inside individual cells, and the theoretical framework into which such measurements will be fit still is being constructed. The range of theoretical approaches being explored is diverse, and it behooves us to search for those approaches which have the chance to organize our understanding of many different systems rather than being satisfied with models of particular systems. Again, there is a balance between the search for generality and the need to connect with experiments on specific networks. We have tried to give some examples of all these developments, hopefully conveying the correct combination of enthusiasm and skepticism.

Acknowledgments

We thank our colleagues and collaborators who have helped us learn about these issues: MJ Berry, CG Callan, T Gregor, JB Kinney, P Mehta, SE Palmer, E Schneidman, JJ Hopfield, T Mora, S Setayeshgar, N Slonim, GJ Stephens, DW Tank, N Tishby, A Walczak, EF Wieschaus, CH Wiggins and NS Wingreen. Our work was supported in part by NIH grants P50 GM071508 and R01 GM077599, by NSF Grants IIS–0613435 and PHY–0650617, by the Swartz Foundation, and by the Burroughs Wellcome Fund.

Bibliography

1. Acar M, Becksei A, van Oudenaarden A (2005) Enhancement of cellular memory by reducing stochastic transitions. Nature 435:228–232
2. Ackers GK, Johnson AD, Shea MA (1982) Quantitative model for gene regulation by lambda phage repressor. Proc Natl Acad Sci (USA) 79(4):1129–33
3. Albert R, Jeong H, Barabasi AL (2000) Error and attack tolerance of complex networks. Nature 406(6794):378–82
4. Aldana M, Cluzel P (2003) A natural class of robust networks. Proc Natl Acad Sci (USA) 100:8710–4
5. Alm E, Arkin AP (2003) Biological networks. Curr Opin Struct Biol 13(2):193–202

6. Alon U, Surette MG, Barkai N, Leibler S (1999) Robustness in bacterial chemotaxis. Nature 397(6715):168–71
7. Arkin A, Ross J, McAdams HH (1998) Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected escherichia coli cells. Genetics 149(4):1633–48
8. Arnosti DN, Kulkarni MM (2005) Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? J Cell Biochem 94(5):890–8
9. Bar-Even A, Paulsson J, Maheshri N, Carmi M, O'Shea E, Pilpel Y, Barkai N (2006) Noise in protein expression scales with natural protein abundance. Nat Genet 38(6):636–43
10. Barabási AL (2002) Linked: The New Science of Networks. Perseus Publishing, Cambridge
11. Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–12
12. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5(2): 101–13
13. Barkai N, Leibler S (1997) Robustness in simple biochemical networks. Nature 387(6636):913–7
14. Baylor DA, Lamb TD, Yau KW (1979) Responses of retinal rods to single photons. J Physiol (Lond) 288:613–634
15. Becskei A, Serrano L (2000) Engineering stability in gene networks by autoregulation. Nature 405(6786):590–3
16. Berg HC (1975) Chemotaxis in bacteria. Annu Rev Biophys Bioeng 4(00):119–36
17. Berg HC, Purcell EM (1977) Physics of chemoreception. Biophys J 20(2):193–219
18. Bergmann S, Sandler O, Sberro H, Shnider S, Schejter E, Shilo BZ, Barkai N (2007) Pre-steady-state decoding of the bicoid morphogen gradient. PLoS Biol 5(2):e46
19. Bethe HA (1935) Statistical theory of superlattices. Proc R Soc London Ser A 150:552–575
20. Bialek W (1987) Physical limits to sensation and perception. Annu Rev Biophys Biophys Chem 16:455–78
21. Bialek W (2001) Stability and noise in biochemical switches. Adv Neur Info Proc Syst 13:103
22. Bialek W, Setayeshgar S (2005) Physical limits to biochemical signaling. Proc Natl Acad Sci (USA) 102(29):10040–5
23. Bialek W, Setayeshgar S (2006) Cooperativity, sensitivity and noise in biochemical signaling. arXiv.org:q-bio.MN/0601001
24. Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Kuhlman T, Phillips R (2005a) Transcriptional regulation by the numbers: applications. Curr Opin Genet Dev 15(2): 125–35
25. Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Phillips R (2005b) Transcriptional regulation by the numbers: models. Curr Opin Genet Dev 15(2):116–24
26. Blake WJ, Kaern M, Cantor CR, Collins JJ (2003) Noise in eukaryotic gene expression. Nature 422(6932):633–7
27. Block SM, Segall JE, Berg HC (1983) Adaptation kinetics in bacterial chemotaxis. J Bacteriol 154(1):312–23
28. Bray D (1995) Protein molecules as computational elements in living cells. Nature 376(6538):307–12
29. Bray D, Bourret RB, Simon MI (1993) Computer simulation of the phosphorylation cascade controlling bacterial chemotaxis. Mol Biol Cell 4(5):469–82
30. Brown PO, Botstein D (1999) Exploring the new world of the genome with DNA microarrays. Nat Genet 21(1 Suppl): 33–7
31. Buchler NE, Gerland U, Hwa T (2003) On schemes of combinatorial transcription logic. Proc Natl Acad Sci (USA) 100(9): 5136–41
32. Chang L, Karin M (2001) Mammalian map kinase signalling cascades. Nature 410(6824):37–40
33. Chen KC, Csikasz-Nagy A, Gyorffy B, Val J, Novak B, Tyson JJ (2000) Kinetic analysis of a molecular model of the budding yeast cell cycle. Mol Biol Cell 11(1):369–91
34. von Dassow G, Meir E, Munro EM, Odell GM (2000) The segment polarity network is a robust developmental module. Nature 406(6792):188–92
35. Dekel E, Alon U (2005) Optimality and evolutionary tuning of the expression level of a protein. Nature 436(7050):588–92
36. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci (USA) 95(25):14863–8
37. Elemento O, Slonim N, Tavazoie S (2007) A universal framework for regulatory element discovery across all genomes and data types. Mol Cell 28(2):337–50
38. Elowitz MB, Leibler S (2000) A synthetic oscillatory network of transcriptional regulators. Nature 403(6767):335–8
39. Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. Science 297(5584):1183–6
40. Falke JJ, Bass RB, Butler SL, Chervitz SA, Danielson MA (1997) The two-component signaling pathway of bacterial chemotaxis: a molecular view of signal transduction by receptors, kinases, and adaptation enzymes. Annu Rev Cell Dev Biol 13:457–512
41. Francois P, Hakim V (2004) Design of genetic networks with specified functions by evolution in silico. Proc Natl Acad Sci (USA) 101(2):580–5
42. Francois P, Hakim V, Siggia ED (2007) Deriving structure from evolution: metazoan segmentation. Mol Syst Bio 3: Article 154
43. Friedman N (2004) Inferring cellular networks using probabilistic graphical models. Science 303(5659):799–805
44. Gardner TS, Cantor CR, Collins JJ (2000) Construction of a genetic toggle switch in *Escherichia coli*. Nature 403(6767): 339–42
45. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415(6868):141–7
46. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS (2003) Global analysis of protein expression in yeast. Nature 425(6959):737–41
47. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. J Phys Chem 81:2340–2361
48. Gillespie DT (2007) Stochastic simulation of chemical kinetics. Annu Rev Phys Chem 58:35–55
49. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley J R L, White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets

RA, McKenna MP, Chant J, Rothb erg JM (2003) A protein interaction map of *Drosophila melanogaster*. Science (5651): 1727–36
50. Golding I, Paulsson J, Zawilski SM, Cox EC (2005) Real-time kinetics of gene activity in individual bacteria. Cell 123(6): 1025–36
51. Goldman MS, Golowasch J, Marder E, Abbott LF (2001) Global structure robustness and modulation of neural models. J Neurosci 21:5229–5238
52. Gonze D, Halloy J, Goldbeter A (2002) Robustness of circadian rhythms with respect to molecular noise. Proc Natl Acad Sci (USA) 99(2):673–8
53. Goulian M (2004) Robust control in bacterial regulatory circuits. Curr Opin Microbiol 7(2):198–202
54. Gregor T, Tank DW, Wieschaus EF, Bialek W (2007a) Probing the limits to positional information. Cell 130(1):153–64
55. Gregor T, Wieschaus EF, McGregor AP, Bialek W, Tank DW (2007b) Stability and nuclear dynamics of the bicoid morphogen gradient. Cell 130(1):141–52
56. Guet CC, Elowitz MB, Hsing W, Leibler S (2002) Combinatorial synthesis of genetic networks. Science 296(5572):1466–70
57. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP, Vidal M (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430(6995): 88–93
58. Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402(6761 Suppl):C47–52
59. Hasty J, McMillen D, Isaacs F, Collins JJ (2001) Computational studies of gene regulatory networks: in numero molecular biology. Nat Rev Genet 2(4):268–79
60. Heinrich R, Schuster S (1996) The Regulation of Cellular Systems. Chapman and Hall, New York
61. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature 415(6868):180–3
62. Hofman J, Wiggins C (2007) A bayesian approach to network modularity. arXiv.org:07093512
63. Hooshangi S, Thiberge S, Weiss R (2005) Ultrasensitivity and noise propagation in a synthetic transcriptional cascade. Proc Natl Acad Sci (USA) 102(10):3581–6
64. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci (USA) 79(8):2554–8
65. Huang KC, Meir Y, Wingreen NS (2003) Dynamic structures in *Escherichia coli:* spontaneous formation of mine rings and mind polar zones. Proc Natl Acad Sci (USA) 100(22):12724–8
66. Ibarra RU, Edwards JS, Palsson BO (2002) *Escherichia coli* K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature 420(6912):186–9
67. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N (2002) Revealing modular organization in the yeast transcriptional network. Nat Genet 31(4):370–7
68. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci (USA) 98(8):4569–74
69. Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3:318–56
70. Jansen R, Gerstein M (2004) Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. Curr Opin Microbiol 7(5):535–45
71. Jaynes ET (1957) Information theory and statistical mechanics. Phys Rev 106:62–79
72. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL (2000) The large-scale organization of metabolic networks. Nature 407(6804):651–4
73. Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411(6833):41–2
74. Jordan JD, Landau EM, Iyengar R (2000) Signaling networks: the origins of cellular multitasking. Cell 103(2):193–200
75. Kadonaga JT (2004) Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. Cell 116(2):247–57
76. van Kampen NG (2007) Stochastic Processes in Physics and Chemistry. Elsevier, Amsterdam
77. Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at genomenet. Nucleic Acids Res 30(1):42–6
78. Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrini-Toole A, Bonavides C, Gama-Castro S (2002) The ecocyc database. Nucleic Acids Res 30(1):56–8
79. Kashtan N, Alon U (2005) Spontaneous evolution of modularity and network motifs. Proc Natl Acad Sci (USA) 102(39):13773–8
80. Kauffman SA (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor Biol 22(3):437–67
81. Keller EF (2005) Revisiting "scale-free" networks. Bioessays 27(10):1060–8
82. Kirschner M, Gerhart J (1998) Evolvability. Proc Natl Acad Sci (USA) 95(15):8420–7
83. Kolch W (2000) Meaningful relationships: the regulation of the ras/raf/mek/erk pathway by protein interactions. Biochem J 351 Pt 2:289–305
84. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature 440(7084):637–43
85. Kuhlman T, Zhang Z, Saier J M H, Hwa T (2007) Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. Proc Natl Acad Sci (USA) 104(14):6043–8
86. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young

RA (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 298(5594):799–804
87. Leloup JC, Goldbeter A (2003) Toward a detailed computational model for the mammalian circadian clock. Proc Natl Acad Sci (USA) 100(12):7051–6
88. LeMasson G, Marder E, Abbott LF (1993) Activity-dependent regulation of conductances in model neurons. Science 259:1915–1917
89. Levine M, Davidson EH (2005) Gene regulatory networks for development. Proc Natl Acad Sci (USA) 102(14):4936–42
90. Lezon TR, Banavar JR, Cieplak M, Maritan A, Federoff NV (2006) Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. Proc Natl Acad Sci (USA) 103:19033–19038
91. Li F, Long T, Lu Y, Ouyang Q, Tang C (2004) The yeast cell-cycle network is robustly designed. Proc Natl Acad Sci (USA) 101(14):4781–6
92. Libby E, Perkins TJ, Swain PS (2007) Noisy information processing through transcriptional regulation. Proc Natl Acad Sci (USA) 104(17):7151–6
93. Marder E, Bucher D (2006) Variability, compensation and homeostasis in neuron and network function. Nature Rev Neurosci 7:563–574
94. Markowetz F, Spang R (2007) Inferring cellular networks – a review. BMC Bioinformatics 8: 6-S5
95. Martin DE, Hall MN (2005) The expanding tor signaling network. Curr Opin Cell Biol 17(2):158–66
96. Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. Science 296(5569):910–3
97. McAdams HH, Arkin A (1997) Stochastic mechanisms in gene expression. Proc Natl Acad Sci (USA) 94(3):814–9
98. McAdams HH, Arkin A (1999) It's a noisy business! Genetic regulation at the nanomolar scale. Trends Genet 15(2):65–9
99. McAdams HH, Shapiro L (1995) Circuit simulation of genetic networks. Science 269(5224):650–6
100. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417(6887):399–403
101. Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S, Ayzenshtat I, Sheffer M, Alon U (2004) Superfamilies of evolved and designed networks. Science 303(5663):1538–42
102. Nakajima M, Imai K, Ito H, Nishiwaki T, Murayama Y, Iwasaki H, Oyama T, Kondo T (2005) Reconstitution of circadian oscillation of cyanobacterial kaic phophorylation in vitro. Science 308:414–415
103. Nasmyth K (1996) At the heart of the budding yeast cell cycle. Trends Genet 12(10):405–12
104. Newman M, Watts D, Barabási AL (2006) The Structure and Dynamics of Networks. Princeton University Press, Princeton
105. Nielsen UB, Cardone MH, Sinskey AJ, MacBeath G, Sorger PK (2003) Profiling receptor tyrosine kinase activation by using ab microarrays. Proc Natl Acad Sci (USA) 100(16):9330–5
106. Nochomovitz YD, Li H (2006) Highly designable phenotypes and mutational buffers emerge from a systematic mapping between network topology and dynamic output. Proc Natl Acad Sci (USA) 103(11):4180–5
107. Novak B, Tyson JJ (1997) Modeling the control of DNA replication in fission yeast. Proc Natl Acad Sci (USA) 94(17): 9147–52
108. Nurse P (2001) Cyclin dependent kinases and cell cycle control. Les Prix Nobel
109. Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A (2002) Regulation of noise in the expression of a single gene. Nat Genet 31(1):69–73
110. Papin JA, Hunter T, Palsson BO, Subramaniam S (2005) Reconstruction of cellular signalling networks and analysis of their properties. Nat Rev Mol Cell Biol 6(2):99–111
111. Paulsson J (2004) Summing up the noise in gene networks. Nature 427(6973):415–8
112. Pedraza JM, van Oudenaarden A (2005) Noise propagation in gene networks. Science 307(5717):1965–9
113. Perez OD, Nolan GP (2002) Simultaneous measurement of multiple active kinase states using polychromatic flow cytometry. Nat Biotechnol 20(2):155–62
114. Ptashne M (2001) Genes and Signals. CSHL Press, Cold Spring Harbor, USA
115. Ptashne M (2004) A Genetic Switch: Phage lambda Revisited. CSHL Press
116. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S (2006) Stochastic mRNA synthesis in mammalian cells. PLoS Biol 4(10):e309
117. Ramanathan S, Detwiler PB, Sengupta AM, Shraiman BI (2005) G-protein-coupled enzyme cascades have intrinsic properties that improve signal localization and fidelity. Biophys J 88(5):3063–71
118. Rao CV, Kirby JR, Arkin AP (2004) Design and diversity in bacterial chemotaxis: a comparative study in *Escherichia coli* and *Bacillus subtilis*. PLoS Biol 2(2):E49
119. Raser JM, O'Shea EK (2005) Noise in gene expression: origins, consequences, and control. Science 309(5743):2010–3
120. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. Science 297(5586):1551–5
121. Rieke F, Baylor DA (1998) Single photon detection by rod cells of the retina. Rev Mod Phys 70:1027–1036
122. Roma DM, O'Flanagan R, Ruckenstein AE, Sengupta AM (2005) Optimal path to epigenetic swithcing. Phys Rev E 71: 011902
123. Rosenfeld N, Alon U (2003) Response delays and the structure of transcription networks. J Mol Biol 329(4):645–54
124. Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB (2005) Gene regulation at the single-cell level. Science 307(5717):1962–5
125. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M (2005) Towards a proteome-scale map of the human protein–protein interaction network. Nature 437(7062): 1173–8
126. Rust MJ, Markson JS, Lane WS, Fisher DS, O'Shea EK (2007) Ordered phosphorylation giverns oscillation of a three-protein circadian clock. Science 318:809–812
127. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP (2005) Causal protein-signaling networks derived from multiparameter single-cell data. Science 308(5721):523–9

128. Sanchez L, Thieffry D (2001) A logical analysis of the *Drosophila* gap-gene system. J Theor Biol 211(2):115–41
129. Sasai M, Wolynes PG (2003) Stochastic gene expression as a many-body problem. Proc Natl Acad Sci (USA) 100(5): 2374–9
130. Schneidman E, Still S, Berry II MJ, Bialek W (2003) Network information and connected correlations. Phys Rev Lett 91(23):238701
131. Schneidman E, Berry II MJ, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. Nature 440(7087):1007–12
132. Schroeder MD, Pearce M, Fak J, Fan H, Unnerstall U, Emberly E, Rajewsky N, Siggia ED, Gaul U (2004) Transcriptional control in the segmentation gene network of *Drosophila*. PLoS Biol 2(9):E271
133. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 34(2):166–76
134. Setty Y, Mayo AE, Surette MG, Alon U (2003) Detailed map of a cis-regulatory input function. Proc Natl Acad Sci (USA) 100(13):7702–7
135. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423 & 623–656
136. Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat Genet 31(1):64–8
137. Shlens J, Field GD, Gauthier JL, Grivich MI, Petrusca D, Sher A, Litke AM, Chichilnisky EJ (2006) The structure of multi-neuron firing patterns in primate retina. J Neurosci 26(32):8254–66
138. Sigal A, Milo R, Cohen A, Geva-Zatorsky N, Klein Y, Liron Y, Rosenfeld N, Danon T, Perzov N, Alon U (2006) Variability and memory of protein levels in human cells. Nature 444(7119):643–6
139. Siggia ED (2005) Computational methods for transcriptional regulation. Curr Opin Genet Dev 15(2):214–21
140. Slonim N, Atwal GS, Tkačik G, Bialek W (2005) Information-based clustering. Proc Natl Acad Sci (USA) 102(51):18297–302
141. Slonim N, Elemento O, Tavazoie S (2006) Ab initio genotype-phenotype association reveals intrinsic modularity in genetic networks. Mol Syst Biol 2 (2006) 0005
142. Spirin V, Mirny LA (2003) Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci (USA) 100(21):12123–8
143. Stelling J, Gilles ED, Doyle 3rd FJ (2004) Robustness properties of circadian clock architectures. Proc Natl Acad Sci (USA) 101(36):13210–5
144. Strogatz SH (2001) Exploring complex networks. Nature 410(6825):268–76
145. Süel GM, Garcia-Ojalvo J, Liberman L, Elowitz MB (2006) An excitable gene regulatory circuit induces transient cellular differentiation. Nature 440:545–550
146. Swain PS (2004) Efficient attenuation of stochasticity in gene expression through post-transcriptional control. J Mol Biol 344(4):965–76
147. Swain PS, Elowitz MB, Siggia ED (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc Natl Acad Sci (USA) 99(20):12795–800
148. Tanay A, Regev A, Shamir R (2005) Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. Proc Natl Acad Sci (USA) 102(20):7203–8
149. Thattai M, van Oudenaarden A (2001) Intrinsic noise in gene regulatory networks. Proc Natl Acad Sci (USA) 98(15):8614–9
150. Thomas R (1973) Boolean formalization of genetic control circuits. J Theor Biol 42(3):563–85
151. Tkačik G (2007) Information Flow in Biological Networks. Dissertation, Princeton University, Princeton
152. Tkačik G, Bialek W (2007) Diffusion, dimensionality and noise in transcriptional regulation. arXiv.org:07121852 [q-bio.MN]
153. Tkačik G, Schneidman E, Berry II MJ, Bialek W (2006) Ising models for networks of real neurons. arXiv.org: q-bio.NC/0611072
154. Tkačik G, Callan Jr CG, Bialek W (2008) Information capacity of genetic regulatory elements. Phys Rev E 78:011910. arXiv.org:0709.4209. [q-bioMN]
155. Tkačik G, Callan Jr CG, Bialek W (2008) Information flow and optimization in transcriptional regulation. Proc Natl Acad Sci 105(34):12265–12270. arXiv.org:0705.0313. [q-bio.MN]
156. Tkačik G, Gregor T, Bialek W (2008) The role of input noise in transcriptional regulation. PLoS One 3, e2774 arXiv.org:q-bioMN/0701002
157. Tomita J, Nakajima M, Kondo T, Iwasaki H (2005) No transcription-translation feedback in circadian rhythm of kaic phosphorylation. Science 307:251–254
158. Tucker CL, Gera JF, Uetz P (2001) Towards an understanding of complex protein networks. Trends Cell Biol 11(3):102–6
159. Tyson JJ, Chen K, Novak B (2001) Network dynamics and cell physiology. Nat Rev Mol Cell Biol 2(12):908–16
160. Tyson JJ, Chen KC, Novak B (2003) Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. Curr Opin Cell Biol 15(2):221–31
161. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature 403(6770):623–7
162. de Visser JA, Hermisson J, Wagner GP, Ancel Meyers L, Bagheri-Chaichian H, Blanchard JL, Chao L, Cheverud JM, Elena SF, Fontana W, Gibson G, Hansen TF, Krakauer D, Lewontin RC, Ofria C, Rice SH, von Dassow G, Wagner A, Whitlock MC (2003) Perspective: Evolution and detection of genetic robustness. Evolution Int J Org Evolution 57(9):1959–72
163. Wagner A, Fell DA (2001) The small world inside large metabolic networks. Proc Biol Sci 268(1478):1803–10
164. Walczak AM, Sasai M, Wolynes PG (2005) Self-consistent proteomic field theory of stochastic gene switches. Biophys J 88(2):828–50
165. Waters CM, Bassler BL (2005) Quorum sensing: cell-to-cell communication in bacteria. Annu Rev Cell Dev Biol 21: 319–46
166. Watson JD, Baker TA, Beli SP, Gann A, Levine M, Losick R (2003) Molecular Biology of the Gene: 5th edn. Benjamin Cummings, Menlo Park
167. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440–2
168. Weinberger LS, Shenk T (2007) An hiv feedback resistor: autoregulatory circuit deactivator and noise buffer. PLoS Biol 5(1):e9
169. Yeger-Lotem E, Sattath S, Kashtan N, Itzkovitz S, Milo R, Pinter RY, Alon U, Margalit H (2004) Network motifs in integrated cellular networks of transcription-regulation and protein-

protein interaction. Proc Natl Acad Sci (USA) 101(16):5934–5939
170. Yokobayashi Y, Weiss R, Arnold FH (2002) Directed evolution of a genetic circuit. Proc Nat Acad Sci (USA) 99(26):16587–91
171. Young MW, Kay SA (2001) Time zones: a comparative genetics of circadian clocks. Nat Rev Genet 2(9):702–15
172. Zaslaver A, Mayo AE, Rosenberg R, Bashkin P, Sberro H, Tsalyuk M, Surette MG, Alon U (2004) Just-in-time transcription program in metabolic pathways. Nat Genet 36(5):486–91
173. Ziv E, Koytcheff R, Middendorf M, Wiggins C (2005a) Systematic identification of statistically significant network measures. Phys Rev E 71:016110
174. Ziv E, Middendorf M, Wiggins C (2005b) Information theoretic approach to network modularity. Phys Rev E 71:046117
175. Ziv E, Nemenman I, Wiggins CH (2006) Optimal signal processing in small stochastic biochemical networks. arXiv.org:q-bio/0612041

Cellular Automata as Models of Parallel Computation

Thomas Worsch
Lehrstuhl Informatik für Ingenieure und Naturwissenschaftler, Universität Karlsruhe, Karlsruhe, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Time and Space Complexity
Measuring and Controlling the Activities
Communication in CA
Future Directions
Bibliography

Glossary

Cellular automaton The classical fine-grained parallel model introduced by John von Neumann.

Hyperbolic cellular automaton A cellular automaton resulting from a tessellation of the hyperbolic plane.

Parallel Turing machine A generalization of Turing's classical model where several control units work cooperatively on the same tape (or set of tapes).

Time complexity Number of steps needed for computing a result. Usually a function $t\colon \mathbb{N}_+ \to \mathbb{N}_+$, $t(n)$ being the maximum ("worst case") for any input of size n.

Space complexity Number of cells needed for computing a result. Usually a function $s\colon \mathbb{N}_+ \to \mathbb{N}_+$, $s(n)$ being the maximum for any input of size n.

State change complexity Number of proper state changes of cells during a computation. Usually a function $sc\colon \mathbb{N}_+ \to \mathbb{N}_+$, $sc(n)$ being the maximum for any input of size n.

Processor complexity Maximum number of control units of a parallel Turing machine which are simultaneously active during a computation. Usually a function $sc\colon \mathbb{N}_+ \to \mathbb{N}_+$, $sc(n)$ being the maximum for any input of size n.

$\mathbb{N}_+$ The set $\{1, 2, 3, \ldots\}$ of positive natural numbers.

$\mathbb{Z}$ The set $\{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$ of integers.

Q^G The set of all (total) functions from a set G to a set Q.

Definition of the Subject

This article will explore the properties of cellular automata (CA) as a parallel model.

The Main Theme

We will first look at the standard model of CA and compare it with Turing machines as the standard sequential model, mainly from a computational complexity point of view. From there we will proceed in two directions: by removing computational power and by adding computational power in different ways in order to gain insight into the importance of some ingredients of the definition of CA.

What Is Left Out

There are topics which we will not cover although they would have fit under the title.

One such topic is *parallel algorithms* for CA. There are algorithmic problems which make sense only for parallel models. Probably the most famous for CA is the so-called *Firing Squad Synchronization Problem*. This is the topic of Umeo's article (► Firing Squad Synchronization Problem in Cellular Automata), which can also be found in this encyclopedia.

Another such topic in this area is the Leader election problem. For CA it has received increased attention in recent years. See the paper by Stratmann and Worsch [29] and the references therein for more details.

And we do want to mention the most exciting (in our opinion) CA algorithm: Tougne has designed a CA which, starting from a single point, after t steps has generated the discretized circle of radius t, for all t; see [5] for this gem.

There are also models which generalize standard CA by making the cells more powerful. Kutrib has introduced *push-down cellular automata* [14]. As the name indicates,

in this model each cell does not have a finite memory but can make use of a potentially unbounded stack of symbols.

The area of nondeterministic CA is also not covered here. For results concerning formal language recognition with these devices refer to ▶ Cellular Automata and Language Theory.

All these topics are, unfortunately, beyond the scope of this article.

Structure of the Paper

The core of this article consists of four sections:

Introduction: The main point is the standard definition of Euclidean deterministic synchronous cellular automata. Furthermore, some general aspects of parallel models and typical questions and problems are discussed.

Time and space complexity: After defining the standard computational complexity measures, we compare CA with different resource bounds. The comparison of CA with the Turing Machine (TM) gives basic insights into their computational power.

Measuring and controlling activities: There are two approaches to measure the "amount of parallelism" in CA. One is an additional complexity measured directly for CA, the other via the definition of so-called parallel Turing machines. Both are discussed.

Communication: Here we have a look at "variants" of CA with communication structures other than the one-dimensional line. We sketch the proofs that some of these variants are in the second machine class.

Introduction

In this section we will first formalize the classical model of cellular automata, basically introduced by von Neumann [21]. Afterwards we will recap some general facts about parallel models.

Definition of Cellular Automata

There are several equivalent formalizations of CA and of course one chooses the one most appropriate for the topics to be investigated. Our point of view will be that each CA consists of a regular arrangement of basic processing elements working in parallel while exchanging data.

Below, for each of the words *regular*, *basic*, *processing*, *parallel* and *exchanging*, we first give the standard definition for clarification. Then we briefly point out possible alternatives which will be discussed in more detail in later sections.

Underlying Grid A Cellular Automaton (CA) consists of a set G of cells, where each cell has at least one *neighbor* with which it can exchange data. Informally speaking one usually assumes a "regular" arrangement of cells and, in particular, identically shaped neighborhoods.

For a *d-dimensional CA*, $d \in \mathbb{N}_+$, one can think of $G = \mathbb{Z}^d$. Neighbors are specified by a finite set N of coordinate differences called the *neighborhood*. The cell $i \in G$ has as its neighbors the cells $i + n$ for all $n \in N$. Usually one assumes that $0 \in N$. (Here we write 0 for a vector of d zeros.)

As long as one is not specifically interested in the precise role N is playing, one may assume some standard neighborhood: The *von Neumann neighborhood* of radius r is $N^{(r)} = \{(k_1, \ldots, k_d) \mid \sum_j |k_j| \leq r\}$ and the *Moore neighborhood* of radius r is $M^{(r)} = \{(k_1, \ldots, k_d) \mid \max_j |k_j| \leq r\}$.

The choices of G and N determine the structure of what in a real parallel computer would be called the "communication network". We will usually consider the case $G = \mathbb{Z}^d$ and assume that the neighborhood is $N = N^{(1)}$.

Discussion The structure of connections between cells is sometimes defined using the concept of *Cayley graphs*. Refer to the article by Ceccherini–Silberstein (▶ Cellular Automata and Groups), also in this encyclopedia, for details.

Another approach is via regular tessellations. For example the 2-dimensional Euclidean space can be tiled with copies of a square. These can be considered as cells, and cells sharing an edge are neighbors. Similarly one can tile, e. g., the *hyperbolic plane* with copies of a regular k-gon. This will be considered to some extent in Sect. "Communication in CA". A more thorough exposition can be found in the article by Margenstern (▶ Cellular Automata in Hyperbolic Spaces) also in this encyclopedia.

CA resulting, for example, from tessellations of the 2-dimensional Euclidean plane with triangles or hexagons are considered in the article by Bays (▶ Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations) also in this encyclopedia.

Global and Local Configurations The basic processing capabilities of each cell are those of a finite automaton. The set of possible states of each cell, denoted by Q, is finite. As inputs to be processed each cell gets the states of all the cells in its neighborhood.

We will write Q^G for the set of all functions from G to Q. Thus each $c \in Q^G$ describes a possible global state of the whole CA. We will call these *c (global) configurations*.

On the other hand, functions $\ell\colon N \to Q$ are called *local configurations*. We say that in a configuration c cell i *observes* the local configuration $c_{i+N}\colon N \to Q$ where $c_{i+N}(n) = c(i+n)$. A cell gets its currently observed local configuration as input. It remains to be defined how they are processed.

Dynamics The dynamics of a CA are defined by specifying the local dynamics of a single cell and how cells "operate in parallel" (if at all). In the first sections we will consider the classical case:

- A *local transition function* is a function $f\colon Q^N \to Q$ prescribing for each local configuration $\ell \in Q^N$ the next state $f(\ell)$ of a cell which currently observes ℓ in its neighborhood. In particular, this means that we are considering *deterministic* behavior of cells.
- Furthermore, we will first concentrate on CA where all cells are working *synchronously*: The possible transitions from one configuration to the next one in one global step of the CA can be described by a function $F\colon Q^G \to Q^G$ requiring that *all* cells make one state transition: $\forall i \in G\colon F(c)(i) = f(c_{i+N})$.

For alternative definitions of the dynamic behavior of CA see Sect. "Measuring and Controlling the Activities".

Discussion Basically, the above definition of CA is the standard one going back to von Neumann [21]; he used $G = \mathbb{Z}^2$ and $N = \{(-1,0),(1,0),(0,-1),(0,1)\}$ for his construction. But for all of the aspects just defined there are other possibilities, some of which will be discussed in later sections.

Finite Computations on CA

In this article we are interested in using CA as devices for computing, given some finite input, in a finite number of steps a finite output.

Inputs As the prototypical examples of problems to be solved by CA and other models we will consider the recognition of formal languages. This has the advantages that the inputs have a simple structure and, more importantly, the output is only one bit (accept or reject) and can be formalized easily.

A detailed discussion of CA as formal language recognizers can be found in the article by Kutrib (▶ Cellular Automata and Language Theory).

The *input alphabet* will be denoted by A. We assume that $A \subset Q$. In addition there has to be a special state $q \in Q$ which is called a *quiescent state* because it has the property that for the quiescent local configuration $\ell_q\colon N \to Q\colon n \mapsto q$ the local transition function must specify $f(\ell_q) = q$.

In the literature two input modes are usually considered.

- Parallel input mode: For an input $w = x_1 \cdots x_n \in A^n$ the initial configuration c_w is defined as

$$c_w(i) = \begin{cases} x_j & \text{iff } i = (j,0,\ldots,0) \\ q & \text{otherwise}\,. \end{cases}$$

- Sequential input mode: In this case, all cells are in state q in the initial configuration. But cell $(0,\ldots,0)$ is a designated *input cell* and acts differently from the others. It works according to a function $g\colon Q^N \times (A \sim \cup \{q\})$. During the first n steps the input cell gets input symbol x_j in step j; after the last input symbol it always gets q. CA using this type of input are often called *iterative arrays (IA)*.

Unless otherwise noted we will always assume parallel input mode.

Conceptually the difference between the two input modes has the same consequences as for TM. If input is provided sequentially it is meaningful to have a look at computations which "use" less than n cells (see the definition of space complexity later on).

Technically, some results occur only for the parallel input mode but not for the sequential one, or vice versa. This is the case, for example, when one looks at devices with small time bounds like n or $n + \sqrt{n}$ steps. But as soon as one considers more generally $\Theta(n)$ or more steps and a space complexity of at least $\Theta(n)$, both CA and IA can simulate each other in linear time:

- An IA can first read the complete input word, storing it in successive cells, and then start simulating a CA, and
- A CA can shift the whole word to the cell holding the first input symbol and have it act as the designated input cell of an IA.

Outputs Concerning output, one usually defines that a CA has finished its work whenever it has reached a *stable configuration* c, i. e., $F(c) = c$. In such a case we will also say that the CA *halts* (although formally one can continue to apply F). An input word $w \in A^+$ is *accepted*, iff the cell $(1,0,\ldots,0)$ which got the first input symbol, is in an *accepting* state from a designated finite subset $F_+ \subset Q$ of states. We write $L(C)$ for the set of all words $w \in A^+$ which are accepted by a CA C.

For the sake of simplicity we will assume that all deterministic machines under consideration halt for all inputs. E. g., the CA as defined above always reaches a stable configuration.

The sequence of all configurations from the initial one for some input w to the stable one is called the *computation* for input w.

Discussion For 1-dimensional CA the definition of parallel input is the obvious one. For higher-dimensional CA, say $G = \mathbb{Z}^2$, one could also think of more compact forms of input for one-dimensional words, e. g., inscribing the input symbols row by row into a square with side length $\lceil \sqrt{n} \rceil$. But this requires extra care. Depending on the formal language to be accepted, the special way in which the symbols are input might provide additional information which is useful for the language recognition task at hand.

Since a CA performs work on an infinite number of bits in each step, it would also be possible to consider inputs and outputs of infinite length, e. g., as representations of all real numbers in the interval $[0, \ldots, 1]$. There is much less literature about this aspect; see for example chapter 11 of [8].

It is not too surprising that this area is also related to the view of CA as dynamical systems (instead of computers); see the contributions by Formenti (▶ Chaotic Behavior of Cellular Automata) and Kůrka (▶ Topological Dynamics of Cellular Automata).

Example: Recognition of Palindromes As an example that will also be useful later, consider the formal language L_{pal} of palindromes of odd length:

$$L_{\text{pal}} = \{vxv^R \mid v \in A^* \wedge x \in A\} .$$

(Here v^R is the mirror image of v.) For example, if A contains all Latin letters `saippuakauppias` belongs to L_{pal} (the Finnish word for a soap dealer).

It is known that each $\mathbb{T}-$TM with only one tape and only one head on it (see Subsect. "Turing Machines" for a quick introduction) needs time $\Omega(n^2)$ for the recognition of at least some inputs of length n belonging to L_{pal} [10].

We will sketch a CA recognizing L_{pal} in time $\Theta(n)$. As the set of states for a single cell we use $Q = A \cup \{\sqcup\} \cup Q_l \times Q_r \times Q_v \times Q_{lr}$, basically subdividing each cell in 4 "registers", each containing a "substate". The substates from $Q_l = \{ <\mathtt{a}, <\mathtt{b}, \sqcup\}$ and $Q_r = \{\mathtt{a}>, \mathtt{b}>, \sqcup\}$ are used to shift input symbols to the left and to the right respectively. In the third register a substate from $Q_v = \{+, -\}$ indicates the results of comparisons. In the fourth register substates from $Q_{lr} = \{>, <, <+>, <->, \sqcup\}$ are used to realize "signals" > and < which identify the middle cell and distribute the relevant overall comparison result to all cells.

As accepting states one chooses those, whose last component is <+>: $F_+ = Q_l \times Q_r \times Q_v \times \{<+>\}$.

There is a total of $3 + 3 \cdot 3 \cdot 2 \cdot 5 = 93$ states and for a complete definition of the local transition function one would have to specify $f(x, y, z)$ for $93^3 = 804357$ triples of states. We will not do that, but we will sketch some important parts. In the first step the registers are initialized. For all $x, y, z \in A$:

$\ell(-1)$	$\ell(0)$	$\ell(1)$	$f(\ell)$
$\sqcup$	y	z	$(<y, y>, +, >)$
x	y	z	$(<y, y>, +, \sqcup)$
x	y	$\sqcup$	$(<y, y>, +, <)$
$\sqcup$	y	$\sqcup$	$(<y, y>, +, <+>)$

In all later steps, if

$$\ell(-1) = \boxed{\begin{matrix} <x_l \\ x_r> \\ v_l \\ d_l \end{matrix}} \quad \ell(0) = \boxed{\begin{matrix} <y_l \\ y_r> \\ v_m \\ d_m \end{matrix}} \quad \ell(1) = \boxed{\begin{matrix} <z_l \\ z_r> \\ v_r \\ d_r \end{matrix}}$$

then

$$f(\ell) = \boxed{\begin{matrix} <z_l \\ x_r> \\ v'_m \\ d'_m \end{matrix}}$$

Here, of course, the new value of the third register is computed as

$$v'_m = \begin{cases} + & \text{if } v_m = + \wedge z_l = x_r \\ - & \text{otherwise} . \end{cases}$$

We do not describe the computation of d'_m in detail. Figure 1 shows the computation for the input `babbbab`, which is a palindrome. Horizontal double lines separate configurations at subsequent time steps. Registers in state $\sqcup$ are simply left empty.

As can be seen there is a triangle in the space time diagram consisting of the n input cells at time $t = 0$ and shrinking at both ends by one cell in each subsequent step where "a lot of activity" can happen due to the shifting of the input symbols in both directions.

Clearly a two-head TM can also recognize L_{pal} in linear time, by first moving one head to the last symbol and then synchronously shifting both heads towards each other comparing the symbols read.

Informally speaking in this case the ability of multi-head TM to transport a small amount of information over

···		b	a	b	b	b	a	b		···
···		<b	<a	<b	<b	<b	<a	<b		···
···		b>	a>	b>	b>	b>	a>	b>		···
···		+	+	+	+	+	+	+		···
···		>						<		···
···		<a	<b	<b	<b	<a	<b			···
···			b>	a>	b>	b>	b>	a>		···
···		-	+	-	+	-	+	-		···
···			>				<			···
···		<b	<b	<b	<a	<b				···
···				b>	a>	b>	b>	b>		···
···		-	-	-	+	-	-	-		···
···				>		<				···
···		<b	<b	<a	<b					···
···					b>	a>	b>	b>		···
···		-	-	-	+	-	-	-		···
···					<+>					···
···		<b	<a	<b						···
···						b>	a>	b>		···
···		-	-	-	+	-	-	-		···
···				<+>	<+>	<+>				···
···		<a	<b							···
···							b>	a>		···
···		-	-	-	-	-	-	-		···
···			<+>	<+>	<+>	<+>	<+>			···
···		<b								···
···								b>		···
···		-	-	-	-	-	-	-		···
···		<+>	<+>	<+>	<+>	<+>	<+>	<+>		···
···										···
···										···
···		-	-	-	-	-	-	-		···
···		<+>	<+>	<+>	<+>	<+>	<+>	<+>		···

Cellular Automata as Models of Parallel Computation, Figure 1
Recognition of a palindrome; the last configuration is stable and the cell which initially stored the first input symbol is in an accepting state

a long distance in one step can be "compensated" by CA by shifting a large amount of information over a short distance. We will see in Theorem 3 that this observation can be generalized.

Complexity Measures: Time and Space For one-dimensional CA it is straightforward to define their time and space complexity. We will consider only worst-case complexity. Remember that we assume that all CA halt for all inputs (reach a stable configuration).

For $w \in A^+$ let $\text{time}'(w)$ denote the smallest number τ of steps such that the CA reaches a stable configuration after τ steps when started from the initial configuration for input w. Then

$$\text{time}\colon \mathbb{N}_+ \to \mathbb{N}_+ \colon n \mapsto \max\{\text{time}'(w) \mid w \in A^n\}$$

is called the *time complexity* of the CA.

Similarly, let $\text{space}'(w)$ denote the total number of cells which are not quiescent in at least one configuration occurring during the computation for input w. Then

$$\text{space}\colon \mathbb{N}_+ \to \mathbb{N}_+ \colon n \mapsto \max\{\text{space}'(w) \mid w \in A^n\}$$

is called the *space complexity* of the CA. If we want to mention a specific CA C, we indicate it as an index, e. g., time_C.

If s and t are functions $\mathbb{N}_+ \to \mathbb{N}_+$, we write $\text{CA} - \textsc{Spc}(s) - \textsc{Time}(t)$ for the set of formal languages which can be accepted by some CA C with $\text{space}_C \le s$ and $\text{time}_C \le t$, and analogously $\text{CA} - \textsc{Spc}(s)$ and $\text{CA} - \textsc{Time}(t)$ if only one complexity measure is bounded. Thus we only look at upper bounds. For a whole set of functions $\mathcal{T}$, we will use the abbreviation

$$\text{CA} - \textsc{Time}(\mathcal{T}) = \bigcup_{t \in \mathcal{T}} \text{CA} - \textsc{Time}(f)\,.$$

Typical examples will be $\mathcal{T} = \mathrm{O}(n)$ or $\mathcal{T} = \mathrm{Pol}(n)$, where in general $\mathrm{Pol}(f) = \bigcup_{k \in \mathbb{N}_+} \mathrm{O}(f^k)$.

Resource bounded complexity classes for other computational models will be noted similarly. If we want to make the dimension of the CA explicit, we write $\mathbb{Z}^d - \text{CA} - \ldots$; if the prefix $\mathbb{Z}^d$ is missing, $d = 1$ is to be assumed.

Throughout this article n will always denote the length of input words. Thus a time complexity of $\mathrm{Pol}(n)$ simply means polynomial time, and similarly for space, so that $\text{TM} - \textsc{Time}(\mathrm{Pol}(n)) = \mathbf{P}$ and $\text{TM} - \textsc{Spc}(\mathrm{Pol}(n)) = \mathbf{PSPACE}$.

Discussion For higher-dimensional CA the definition of space complexity requires more consideration. One possibility is to count the number of cells used during the computation. A different, but sometimes more convenient approach is to count the number of cells in the smallest hyper-rectangle comprising all used cells.

Turing Machines

For reference, and because we will consider a parallel variant, we set forth some definitions of Turing machines.

In general we allow Turing machines with k work tapes and h heads on each of them. Each square carries a symbol from the *tape alphabet* B which includes the blank symbol $\square$. The control unit (CU) is a finite automaton with *set of states* S. The possible actions of a deterministic TM are described by a function $f\colon S \times B^{kh} \to S \times B^{kh} \times D^{kh}$, where $D = \{-1, 0, +1\}$ is used for indicating the direction of movement of a head.

If the machine reaches a situation in which $f(s, b_1, \ldots, b_{kh}) = (s, b_1, \ldots, b_{kh}, 0, \ldots, 0)$ for the current state s and the currently scanned symbols $b_1, \ldots, b_{kh}$, we say that it halts.

Initially a word w of length n over the input alphabet $A \subset B$ is written on the first tape on squares $1, \ldots, n$, all other tape squares are empty, i. e., carry the $\square$. An input is

accepted if the CU halts in an accepting state from a designated subset $F_+ \subset S$. $L(T)$ denotes the formal language of all words accepted by a TM T.

We write $k\mathbb{T}h - \mathrm{TM} - \textsc{Spc}(s) - \textsc{Time}(t)$ for the class of all formal languages which can be recognized by TM T with k work tapes and h heads on each of them which have a space complexity $\mathrm{space}_T \leq s$ and $\mathrm{time}_T \leq t$. If k and/or h is missing, 1 is assumed instead. If the whole prefix $k\mathbb{T}h$ is missing, $\mathbb{T}$ is assumed. If arbitrary k and h are allowed, we write $*\mathbb{T}*$.

Sequential Versus Parallel Models

Today quite a number of different computational models are known which intuitively look as if they are parallel. Several years ago van Emde Boas [6] observed that many of these models have one property in common. The problems that can be solved in polynomial *time* on such a model $\mathcal{P}$, coincide with the problems that can be solved in polynomial *space* on Turing machines:

$$\mathcal{P} - \textsc{Time}(\mathrm{Pol}(n)) = \mathrm{TM} - \textsc{Spc}(\mathrm{Pol}(n)) = \textbf{PSPACE}\,.$$

Here we have chosen $\mathcal{P}$ as an abbreviation for "parallel" model. Models $\mathcal{P}$ satisfying this equality are by definition the members of the so-called *second machine class*.

On the other hand, the *first machine class* is formed by all models S satisfying the relation

$$S - \textsc{Spc}(s) - \textsc{Time}(t) = \mathrm{TM} - \textsc{Spc}(\Theta(s)) - \textsc{Time}(\mathrm{Pol}(t))$$

at least for some reasonable functions s and t. We deliberately avoid making this more precise. In general there is consensus on which models are in the these machine classes.

We do want to point out that the naming of the two machine classes does *not* mean that they are different or even disjoint. This is not known. For example, if **P = PSPACE**, it might be that the classes coincide. Furthermore there are models, e. g., Savitch's NLPRAM [25], which might be in neither machine class.

Another observation is that in order to possibly classify a machine model it obviously has to have something like "time complexity" and/or "space complexity". This may sound trivial, but we will see in Subsect. "Parallel Turing Machines" that, for example, for so-called parallel Turing machines with several work tapes it is in fact not.

Time and Space Complexity

Comparison of Resource Bounded One-Dimensional CA

It is clear that time and space complexity for CA are Blum measures [2] and hence infinite hierarchies of complexity classes exist. It follows from the more general Theorem 9 for parallel Turing machines that the following holds:

Theorem 1 *Let s and t be two functions such that s is fully CA space constructable in time t and t is CA computable in space s and time t. Then:*

$$\bigcup_{\gamma \notin \mathrm{O}(1)} \mathrm{CA} - \textsc{Spc}(\Theta(s/\gamma)) - \textsc{Time}(\Theta(t/\gamma)) \subsetneq \mathrm{CA} - \textsc{Spc}(\mathrm{O}(s)) - \textsc{Time}(\mathrm{O}(t))$$

$$\mathrm{CA} - \textsc{Spc}(o(s)) - \textsc{Time}(o(t)) \subsetneq \mathrm{CA} - \textsc{Spc}(\mathrm{O}(s)) - \textsc{Time}(\mathrm{O}(t))$$

$$\mathrm{CA} - \textsc{Time}(o(t)) \subsetneq \mathrm{CA} - \textsc{Time}(\mathrm{O}(t))\,.$$

The second and third inclusion are simple corollaries of the first one. We do not go into the details of the definition of CA constructibility, but note that for hierarchy results for TM one sometimes needs analogous additional conditions. For details, interested readers are referred to [3,12,17].

We not that for CA the situation is better than for deterministic TM: there one needs $f(n)\log f(n) \in \mathrm{o}(g(n))$ in order to prove $\mathrm{TM} - \textsc{Time}(f) \subsetneq \mathrm{TM} - \textsc{Time}(g)$.

Open Problem 2 For the proper inclusions in Theorem 1 the construction used in [34] really needs to increase the space used in order to get the time hierarchy. It is an open problem whether there also exists a time hierarchy if the space complexity is *fixed*, e. g., as $s(n) = n$. It is even an open problem to prove or disprove that the inclusion

$$\mathrm{CA} - \textsc{Spc}(n) - \textsc{Time}(n) \subseteq \mathrm{CA} - \textsc{Spc}(n) - \textsc{Time}(2^{\mathrm{O}(n)})\,.$$

is proper or not. We will come back to this topic in Subsect. "Parallel Turing Machines" on parallel Turing machines.

Comparison with Turing Machines

It is well-known that a TM with one tape and one head on that tape can be simulated by a one-dimensional CA. See for example the paper by Smith [27]. But even multi-tape TM can be simulated by a one-dimensional CA without any significant loss of time.

Theorem 3 *For all space bounds $s(n) \geq n$ and all time bounds $t(n) \geq n$ the following holds for one-dimensional CA and TM with an arbitrary number of heads on its tapes:*

$$*\,\mathbb{T}\,* - \mathrm{TM} - \textsc{Spc}(s) - \textsc{Time}(t) \subseteq \mathrm{CA} - \textsc{Spc}(s) - \textsc{Time}(\mathrm{O}(t))\,.$$

Sketch of the Simulation We first describe a simulation for $1\mathbb{T}1 - \mathrm{TM}$.

In this case the actions of the TM are of the form $s, b \to s', b', d$ where $s, s' \in S$ are old and new state, $b, b' \in B$ old and new tape symbol and $d \in \{-1, 0, +1\}$ the direction of *head* movement.

The simulating CA uses three substates in each cell, one for a TM state, one for a tape symbol, and an additional one for shifting tape symbols: $Q = Q_S \times Q_T \times Q_M$. We use $Q_S = S \cup \{\sqcup\}$ and a substate of $\sqcup$ means, that the cell does not store a state. Similarly $Q_T = B \cup \{ < \circ, \circ > \}$ and a substate of $< \circ$ or $\circ >$ means that there is no symbol stored but a "hole" to be filled with an adjacent symbol. Substates from $Q_S = B \times \{ <, > \}$, like $< b$ and $b >$, are used for shifting symbols from one cell to the adjacent one to the left or right.

Instead of moving the state one cell to the right or left whenever the TM moves its head, the tape contents as stored in the CA are shifted in the opposite direction. Assume for example that the TM performs the following actions:

- $\mathtt{s0,d} \to \mathtt{s1}, \mathtt{d}', +1$
- $\mathtt{s1,e} \to \mathtt{s2}, \mathtt{e}', -1$
- $\mathtt{s2,d}' \to \mathtt{s3}, \mathtt{d}'', -1$
- $\mathtt{s3,c} \to \mathtt{s4}, \mathtt{c}', +1$

Figure 2 shows how shifting the tape in direction d can be achieved by sending the current symbol in that direction and sending a "hole" $\circ$ in the opposite direction $-d$. It should be clear that the required state changes of each cell depend only on information available *in its neighborhood*.

A consequence of this approach to incrementally shift the tape contents is that it takes an arbitrary large number of steps until all symbols have been shifted. On the other hand, after only *two* steps the cell simulating the TM control unit has information about the next symbol visited and can simulate the next TM step and initialize the next tape shift.

Clearly the same approach can be used if one wants to simulate a TM with several tapes, each having one head. For each additional tape the CA would use two additional registers analogously to the middle and bottom row used in Fig. 2 for one tape. Stoß [28] has proved that $k\mathbb{T}\,h-$ TM (h heads on each tape) can be simulated by $(kh)\mathbb{T}-$ TM (only one head on each tape) in linear time. Hence there is nothing left to prove.

Discussion As one can see in Fig. 2, in every second step one signal is sent to the left and one to the right. Thus, if the TM moves its head a lot and if the tape segment which has to be shifted is already long, many signals are traveling simultaneously.

...		a	b	c	s0 d	e	f	g		...
...		a	b	c	s1 o> <d'	e	f	g		...
...		a	b	d' <c	s1 e	o>	f	g		...
...		a	c <c	d'	s2 <o e'>	f	o>	g		...
...		b <a	c	<o	s2 d'	e' f>	g	o>		...
...	a	b	<o	c	s3 <o d''>	e'	f g>			...
...	a	<o	b	<o	s3 c	d'' e'>	f	g		...
...	<o	a	<o	b	s4 o> <c'	d''	e' f>	g		...

Cellular Automata as Models of Parallel Computation, Figure 2
Shifting tape contents step by step

In other words, the CA transports "a large amount of information over a short distance in one step". Theorem 3 says that this ability is at least as powerful as the ability of multi-head TM to transport "a small amount of information over a long distance in one step".

Open Problem 4 The question remains whether some kind of converse also holds and in Theorem 3 an $=$ sign would be correct instead of the $\subseteq$, or whether CA are more powerful, i. e., a $\subsetneq$ sign would be correct. This is not known.

The best simulation of CA by TM that is known is the obvious one: states of neighboring cells are stored on adjacent tape squares. For the simulation of one CA step the TM basically makes one sweep across the complete tape segment containing the states of all non-quiescent cells updating them one after the other. As a consequence one gets

Theorem 5 *For all space bounds $s(n) \geq n$ and all time bounds $t(n) \geq n$ holds:*

$$\begin{aligned} \text{CA}-\textsc{Spc}(s)-\textsc{Time}(t) &\subseteq \text{TM}-\textsc{Spc}(s)-\textsc{Time}(\mathrm{O}(s \cdot t)) \\ &\subseteq \text{TM}-\textsc{Spc}(s)-\textsc{Time}\left(\mathrm{O}\left(t^2\right)\right) . \end{aligned}$$

The construction proving the first inclusion needs only a one-head TM, and no possibility is known to take advantage of more heads. The second inclusion follows from the observation that in order to use an initially blank tape square, a TM must move one of its heads there, which requires time. Thus $s \in \mathrm{O}(t)$.

Taking Theorems 3 and 5 together, one immediately gets

Corollary 6 *Cellular automata are in the first machine class.*

And it is not known that CA are in the second machine class. In this regard they are "more like" sequential models. The reason for this is the fact, the number of active processing units only grows polynomially with the number of steps in a computation. In Sect. "Communication in CA" variations of the standard CA model will be considered, where this is different.

Measuring and Controlling the Activities

Parallel Turing Machines

One possible way to make a parallel model from Turing machines is to allow several control units (CU), but with all of them working on the same tape (or tapes). This model can be traced back at least to a paper by Hemmerling [9] who called it *Systems of Turing automata*. A few years later Wiedermann [32] coined the term *Parallel Turing machine* (PTM).

We consider only the case where there is only one tape and each of the control units has only one head on that tape. As for sequential TM, we usually drop the prefix $1\mathbb{T}1$ for PTM, too. Readers interested in the case of PTM with multi-head CUs are referred to [33].

PTM with One-Head Control Units The specification of a PTM includes a tape alphabet B with a blank Cybil $\square$ and a set S of possible states for each CU. A PTM starts with one CU on the first input symbol as a sequential $1\mathbb{T}1-$TM. During the computation the number of control units may increase and decrease, but all CUs always work cooperatively on one common tape.

The idea is to have the CUs act independently unless they are "close" to each other, retaining the idea of only local interactions, as in CA.

A *configuration* of a PTM is a pair $c = (p, b)$. The mapping $b\colon \mathbb{Z} \to B$ describes the contents of the tape. Let 2^S denote the power set of S. The mapping $p\colon \mathbb{Z} \to 2^S$ describes for each tape square i the set of states of the finite automata currently visiting it. In particular, this formalization means that it is not possible to distinguish two automata on the same square and in the same state: the idea is that because of that they will always behave identically and hence need not be distinguished.

The mode of operation of a PTM is determined by the *transition function* $f\colon 2^Q \times B \to 2^{Q\times D} \times B$ where D is the set $\{-1, 0, 1\}$ of possible movements of a control unit. In order to compute the successor configuration $c' = (p', b')$ of a configuration $c = (p, b)$, f is simultaneously computed for all tape positions $i \in \mathbb{Z}$. The arguments used are the set of states of the finite automata currently visiting square i and its tape symbol. Let $(M'_i, b'_i) = f(p(i), b(i))$. Then the new symbol on square i in configuration c' is $b'(i) = b'_i$. The set of finite automata on square i is replaced by a new set of finite automata (defined by $M'_i \subseteq Q \times D$) each of which changes the tape square according to the indicated direction of movement. Therefore $p'(i) = \{q \mid (q, 1) \in M'_{i-1} \vee (q, 0) \in M'_i \vee (q, -1) \in M'_{i+1}\}$. Thus f induces a global transition function F mapping global configurations to global configurations.

In order to make the model useful (and to come up to some intuitive expectations) it is required, that CUs cannot arise "out of nothing" and that the symbol on a tape square can change only if it is visited by at least one CU. In other words we require that $\forall\, b \in B\colon f(\emptyset, b) = (\emptyset, b)$.

Observe that the number of finite automata on the tape may change during a computation. Automata may vanish, for example if $f(\{s\}, b) = (\emptyset, b)$ and new automata may be generated, for example if $f(\{s\}, b) = (\{(q, 1), (q', 0)\}, b)$.

For the recognition of formal languages we define the *initial configuration* c_w for an input word $w \in A^+$ as the one in which w is written on the otherwise blank tape on squares $1, 2, \ldots, |w|$, and in which there exists exactly one finite automaton in an initial state q_0 on square 1.

A configuration (p, b) of aPTM is called *accepting* iff it is stable (i. e. $F((p, b)) = (p, b)$) and $p(1) \subseteq F_+$. The *language $L(P)$ recognized by a* PTM P is the set of input words, for which it reaches an accepting configuration.

Complexity Measures for PTM Time complexity of a PTM can be defined in the obvious way.

For space complexity, one counts the total number of tape squares which are used in at least one configuration. Here we call a tape square i *unused* in a configuration $c = (p, b)$ if $p(i) = \emptyset$ and $b(i) = \square$; otherwise it is used.

What makes PTM interesting is the definition of its *processor complexity*. Let $\mathrm{proc}'(w)$ denote the maximum number of CU which exist simultaneously in a configuration occurring during the computation for input w and define $\mathrm{proc}\colon \mathbb{N}_+ \to \mathbb{N}_+ : n \mapsto \max\{\mathrm{proc}'(w) \mid w \in A^n\}$. For complexity classes we use the notation $\mathrm{PTM}-\textsc{Spc}(s)-\textsc{Time}(t)-\textsc{Proc}(p)$ etc.

The processor complexity is one way to measure (an upper bound on) "how many activities" happened simultaneously. It should be clear that at the lower end one has the case of constant $\mathrm{proc}(n) = 1$, which means that the PTM is in fact (equivalent to) a sequential TM. The other extreme is to have CUs "everywhere". In that case

$\mathrm{proc}(n) \in \Theta(\mathrm{space}(n))$, and one basically has a CA. In other words, processor complexity measures the mount of parallelism of a PTM.

Theorem 7 *For all space bounds s and time bounds t:*

$$\begin{aligned}&\mathrm{PTM} - \textsc{Spc}(s) - \textsc{Time}(t) - \textsc{Proc}(1)\\ &\qquad = \mathrm{TM} - \textsc{Spc}(s) - \textsc{Time}(t)\\ &\mathrm{PTM} - \textsc{Spc}(s) - \textsc{Time}(t) - \textsc{Proc}(s)\\ &\qquad = \mathrm{CA} - \textsc{Spc}(s) - \textsc{Time}(\mathrm{O}(t))\,.\end{aligned}$$

Under additional constructibility conditions it is even possible to get a generalization of Theorem 5:

Theorem 8 *For all functions $s(n) \geq n$, $t(n) \geq n$, and $h(n) \geq 1$, where h is fully* PTM *processor constructable in space s, time t, and with h processors, holds:*

$$\begin{aligned}&\mathrm{CA} - \textsc{Spc}(\mathrm{O}(s)) - \textsc{Time}(\mathrm{O}(t))\\ &\subseteq \mathrm{PTM} - \textsc{Spc}(\mathrm{O}(s)) - \textsc{Time}(\mathrm{O}(st/h)) - \textsc{Proc}(\mathrm{O}(h)).\end{aligned}$$

Decreasing the processor complexity indeed leads to the expected slowdown.

Relations Between PTM Complexity Classes (part 1) The interesting question now is whether different upper bounds on the processor complexity result in different computational power. In general that is not the case, as PTM with only one CU are TM and hence computationally universal. (As a side remark we note that therefore processor complexity cannot be a Blum measure. In fact that should be more or less clear since, e. g., deciding whether a second CU will ever be generated might require finding out whether the first CU, i. e., a TM, ever reaches a specific state.)

In this first part we consider the case where two complexity measures are allowed to grow in order to get a hierarchy. Results which only need one growing measure are the topic of the second part.

First of all it turns out that for fixed processor complexity between $\log n$ and $s(n)$ there is a space/time hierarchy:

Theorem 9 *Let s and t be two functions such that s is fully* PTM *space constructable in time t and t is* PTM *computable in space s and time t and let $h \geq \log$. Then:*

$$\begin{aligned}&\bigcup_{\gamma \notin \mathrm{O}(1)} \mathrm{PTM} - \textsc{Spc}(\Theta(s/\gamma)) - \textsc{Time}(\Theta(t/\gamma))\\ &\qquad - \textsc{Proc}(\mathrm{O}(h))\\ &\subsetneq \mathrm{PTM} - \textsc{Spc}(\mathrm{O}(s)) - \textsc{Time}(\mathrm{O}(t)) - \textsc{Proc}(\mathrm{O}(h))\,.\end{aligned}$$

The proof of this theorem applies the usual idea of diagonalization. Technical details can be found in [34].

Instead of keeping processor complexity fixed and letting space complexity grow, one can also do the opposite. As for analogous results for TM, one needs the additional restriction to one *fixed tape alphabet*. One gets the following result, where the complexity classes carry the additional information about the size of the tape alphabet.

Theorem 10 *Let s, t and h be three functions such that s is fully* PTM *space constructable in time t and with h processors, and that t and h are* PTM *computable in space s and time t and with h processors such that in all cases the tape is not written. Let $b \geq 2$ be the size of the tape alphabet. Then:*

$$\begin{aligned}&\bigcup_{\gamma \notin \mathrm{O}(1)} \mathrm{PTM} - \textsc{Spc}(s) - \textsc{Time}\Theta(t/\gamma) - \textsc{Proc}(h/\gamma)\\ &\qquad - \textsc{Alph}(b)\\ &\subsetneq \mathrm{PTM} - \textsc{Spc}(s) - \textsc{Time}(\Theta(st)) - \textsc{Proc}(\Theta(h))\\ &\qquad - \textsc{Alph}(b)\,.\end{aligned}$$

Again we do not go into the details of the constructibility definitions which can be found in [34]. The important point here is that one can prove that increasing time and processor complexity by a non-constant factor does increase the (language recognition) capabilities of PTM, even if the space complexity is fixed, provided that one does not allow any changes to the tape alphabet. In particular the theorem holds for the case $\mathrm{space}(n) = n$.

It is now interesting to reconsider Open Problem 2. Let's *assume* that

$$\mathrm{CA} - \textsc{Spc}(n) - \textsc{Time}(n) = \mathrm{CA} - \textsc{Spc}(n) - \textsc{Time}(2^{\mathrm{O}(n)})\,.$$

One may choose $\gamma(n) = \log n$, $t(n) = 2^{n/\log n}$ and $h(n) = n$ in Theorem 10. Using that together with Theorem 7 the assumption would give rise to

$$\begin{aligned}&\mathrm{PTM} - \textsc{Spc}(n) - \textsc{Time}\left(\frac{2^{n/\log n}}{\log n}\right) - \textsc{Proc}\left(\frac{n}{\log n}\right)\\ &\qquad - \textsc{Alph}(b)\\ &\subsetneq \mathrm{PTM} - \textsc{Spc}(n) - \textsc{Time}(n2^{n/\log n}) - \textsc{Proc}(n)\\ &\qquad - \textsc{Alph}(b)\\ &= \mathrm{PTM} - \textsc{Spc}(n) - \textsc{Time}(n) - \textsc{Proc}(n)\\ &\qquad - \textsc{Alph}(b)\,.\end{aligned}$$

If the polynomial time hierarchy for n-space bounded CA collapses, then there are languages which cannot be recognized by PTM in almost exponential time with $n/\log n$ processors but which can be recognized by PTM with n processors in linear time, *if the tape alphabet is fixed*.

Relations Between PTM Complexity Classes (part 2) One can get rid of the fixed alphabet condition by using a combinatorial argument for a specific formal language (instead of diagonalization) and even have not only the space but also the processor complexity fixed and still get a time hierarchy. The price to pay is that the range of time bounds is more restricted than in Theorem 10.

Consider the formal language

$$L_{\mathrm{vv}} = \left\{ v\mathtt{c}^{|v|}v \mid v \in \{\mathtt{a},\mathtt{b}\}^+ \right\} .$$

It contains all words which can be divided into three segments of equal length such that the first and third are identical. Intuitively whatever type of machine is used for recognition, it is unavoidable to "move" the complete information from one end to the other. L_{vv} shares this feature with L_{pal}.

Using a counting argument inspired by Hennie's concept of crossing sequences [10] applied to L_{pal}, one can show:

Lemma 11 ([34]) *If P is a PTM recognizing L_{vv}, then* $\mathrm{time}_P^2 \cdot \mathrm{proc}_P \in \Omega(n^3/\log^2 n)$.

On the other hand, one can construct a PTM recognizing L_{vv} with processor complexity n^a for sufficiently nice a:

Lemma 12 ([34]) *For each $a \in \mathbb{Q}$ with $0 < a < 1$ holds:*

$$L_{\mathrm{vv}} \in \mathrm{PTM}\text{–}\textsc{Spc}(n)\text{–}\textsc{Time}\left(\Theta\left(n^{2-a}\right)\right)\text{–}\textsc{Proc}\left(\Theta\left(n^{a}\right)\right) .$$

Putting these lemmas together yields another hierarchy theorem:

Theorem 13 *For rational numbers $0 < a < 1$ and $0 < \varepsilon < 3/2 - a/2$ holds:*

$$\begin{aligned} &\mathrm{PTM} - \textsc{Spc}(n) - \textsc{Time}\left(\Theta\left(n^{3/2-a/2-\varepsilon}\right)\right) \\ &\qquad\qquad - \textsc{Proc}\left(\Theta\left(n^{a}\right)\right) \\ &\subsetneq \mathrm{PTM} - \textsc{Spc}(n) - \textsc{Time}\left(\Theta\left(n^{2-a}\right)\right) \\ &\qquad\qquad - \textsc{Proc}\left(\Theta\left(n^{a}\right)\right) . \end{aligned}$$

Hence, for a close to 1 a "small" increase in time by some $n^{\varepsilon'}$ suffices to increase the recognition power of PTM while the processor complexity is fixed at n^a and the space complexity is fixed at n as well.

Open Problem 14 For the recognition of L_{vv} there is a gap between the lower bound of $\mathrm{time}_P^2 \cdot \mathrm{proc}_P \in \Omega(n^3/\log^2 n)$ in Lemma 11 and the upper bound of $\mathrm{time}_P^2 \cdot \mathrm{proc}_P \in \mathrm{O}(n^{4-a})$ in Lemma 12. It is not known whether the upper or the lower bound or both can be improved.

An even more difficult problem is to prove a similar result for the case $a = 1$, i. e., cellular automata, as mentioned in Open Problem 2.

State Change Complexity

In CMOS technology, what costs most of the energy is to make a *proper state change*, from zero to one or from one to zero. Motivated by this fact Vollmar [30] introduced the state change complexity for CA.

There are two variants based on the same idea: Given a halting CA computation for an input w and a cell i one can count the number of time points τ, $1 \le \tau \le \mathrm{time}'(w)$, such that cell i is in different states at times $\tau - 1$ and τ. Denote that number by $\mathrm{change}'(w, i)$. Define

$$\mathrm{maxchg}'(w) = \max_{i \in G} \mathrm{change}'(w, i) \quad \text{and}$$
$$\mathrm{sumchg}'(w) = \sum_{i \in G} \mathrm{change}'(w, i)$$

and

$$\mathrm{maxchg}(n) = \max\{\mathrm{maxchg}'(w) \mid w \in A^n\}$$
$$\mathrm{sumchg}(n) = \max\{\mathrm{sumchg}'(w) \mid w \in A^n\} .$$

For the language L_{vv} which already played a role in the previous subsection one can show:

Lemma 15 *Let $f(n)$ be a non-decreasing function which is not in $\mathrm{O}(\log n)$, i. e., $\lim_{n\to\infty} \log n/f(n) = 0$. Then any CA C recognizing L_{vv} makes a total of at least $\Omega(n^2/f(n))$ state changes in the segment containing the n input cells and n cells to the left and to the right of them.*

In particular, if $\mathrm{time}_C \in \Theta(n)$, then $\mathrm{sumchg}_C \in \Omega(n^2/f(n))$.

Furthermore $\mathrm{maxchg}_C \in \Omega(n/f(n))$.

In the paper by Sanders et al. [24] a generalization of this lemma to d-dimensional CA is proved.

Open Problem 16 While the processor complexity of PTM measures how many activities happen simultaneously "across space", state change complexity measures how many activities happen over time. For both cases we have made use of the same formal language in proofs. That might be an indication that there are connections between the two complexity measures. But no non-trivial results are known until now.

Asynchronous CA

Until now we have only considered one global mode of operation: the so-called *synchronous* case, where in each

global step of the CA all cells must update their states synchronously. Several models have been considered where this requirement has been relaxed.

Generally speaking, asynchronous CA are characterized by the fact that in one global step of the CA *some* cells are active and do update their states (all according to the same local transition function) while others do nothing, i. e., remain in the same state as before. There are then different approaches to specify some restrictions on which cells may be active or not.

Asynchronous update mode. The simplest possibility is to not quantify anything and to say that a configuration c' is a legal successor of configuration c, denoted $c \vdash c'$, iff for all $i \in G$ one has $c'(i) = c(i)$ or $c'(i) = f(c_{i+N})$.

Unordered sequential update mode. In this special case it is required that there is only one active cell in each global step, i. e., $\mathrm{card}(\{i | c'(i) \neq c(i)\}) \leq 1$.

Since CA with an asynchronous update mode are no longer deterministic from a global (configuration) point of view, it is not completely clear how to define, e. g., formal language recognition and time complexity. Of course one could follow the way it is done for nondeterministic TM. To the best of our knowledge this has not considered for asynchronous CA. (There are results for general nondeterministic CA; see for example ► Cellular Automata and Language Theory.)

It should be noted that Nakamura [20] has provided a very elegant construction for simulating a CA C_s with synchronous update mode on a CA C_a with one of the above asynchronous update modes. Each cell stores the "current" and the "previous" state of a C_s-cell before its last activation and a counter value T modulo 3 ($Q_a = Q_s \times Q_s \times \{0, 1, 2\}$). The local transition function f_a is defined in such a way that an activated cell does the following:

- T always indicates how often a cell has already been updated modulo 3.
- If the counters of all neighbors have value T or $T + 1$, the current C_s-state of the cell is remembered as previous state and a new current state is computed according to f_s from the current and previous C_s-states of the neighbors; the selection between current and previous state depends on the counter value of that cell. In this case the counter is incremented.
- If the counter of at least one neighboring cell is at $T - 1$, the activated cell keeps its complete state as it is.

Therefore, if one does want to gain something using asynchronous CA, their local transition functions would have to be designed for that specific usage.

Recently, interest has increased considerably in CA where the "degree of (a-)synchrony" is quantified via probabilities. In these cases one considers CA with only a finite number of cells.

Probabilistic update mode. Let $0 \leq \alpha \leq 1$ be a probability. In probabilistic update mode each legal global step $c \vdash c'$ of the CA is assigned a probability by requiring that each cell i independently has a probability α of updating its state.

Random sequential update mode. This is the case when in each global step one of the cells in G is chosen with even probability and its state updated, while all others do not change their state. CA operating in this mode are called *fully asynchronous* by some authors.

These models can be considered special cases of what is usually called *probabilistic* or *stochastic CA*. For these CA the local transition function is no longer a map from Q^N to Q, but from Q^N to $[0; 1]^Q$. For each $\ell \in Q^N$ the value $f(\ell)$ is a probability distribution for the next state (satisfying $\sum_{q \in Q} f(\ell)(q) = 1$). There are only very few papers about formal language recognition with probabilistic CA; see [18].

On the other hand, probabilistic update modes have received some attention recently. See for example [22] and the references therein. Development of this area is still at its beginning. Until now, specific local rules have mainly been investigated; for an exception see [7].

Communication in CA

Until now we have considered only one-dimensional Euclidean CA where one bit of information can reach $O(t)$ cells in t steps. In this section we will have a look at a few possibilities for changing the way cells communicate in a CA.

First we have a quick look at CA where the underlying grid is $\mathbb{Z}^d$. The topic of the second subsection is CA where the cells are connected to form a tree.

Different Dimensionality

In $\mathbb{Z}^d$ − CA with, e. g., von Neumann neighborhood of radius 1, a cell has the potential to influence $O(t^d)$ cells in t steps. This is a polynomial number of cells. It comes as no

surprise that

$$\mathbb{Z}^d - \text{CA} - \textsc{Spc}(s) - \textsc{Time}(t) \subseteq \text{TM} - \textsc{Spc}(\text{Pol}(s)) - \textsc{Time}(\text{Pol}(t))$$

and hence $\mathbb{Z}^d - \text{CA}$ are in the first machine class. One might only wonder why for the TM a space bound of $\Theta(s)$ might not be sufficient. This is due to the fact that the shape of the cells actually used by the CA might have an "irregular" structure and that the TM has to perform some bookkeeping or simulate a whole (hyper-)rectangle of cells encompassing all that are really used by the CA.

Trivially, a d-dimensional CA can be simulated on a d'-dimensional CA, where $d' > d$. The question is how much one loses when *decreasing* the dimensionality. The currently known best result in this direction is by Scheben [26]:

Theorem 17 *It is possible to simulate a d'-dimensional CA with running time t on a d-dimensional CA, $d < d'$ with running time and space* $\mathrm{O}(t^{2^{\mathrm{ld}\lceil d'/d\rceil}})$.

It should be noted that the above result is not directly about language recognition; the redistribution of input symbols needed for the simulation is not taken into account. Readers interested in that as well are referred to [1].

Open Problem 18 Try to find simulations of lower on higher dimensional CA which somehow make use of the "higher connectivity" between cells. It is probably much too difficult or even impossible to hope for general speedups. But efficient use of space (small hypercubes) for computations without losing time might be achievable.

Tree CA and hyperbolic CA

Starting from the root of a full binary tree one can reach an exponential number 2^t of nodes in t steps. If there are some computing capabilities related to the nodes, there is at least the possibility that such a device might exhibit some kind of strong parallelism.

One of the earliest papers in this respect is Wiedermann's article [31] (unfortunately only available in Slovak). The model introduced there one would now call parallel Turing machines, where a tape is not a linear array of cells, but where the cells are connected in such a way as to form a tree. A proof is sketched, showing that these devices can simulate PRAMs in linear time (assuming the so-called logarithmic cost model). PRAMs are in the second machine class.

So, indeed in some sense, trees are powerful. Below we first quickly introduce a **PSPACE**-complete problem which is a useful tool in order to prove the power of computational models involving trees. A few examples of such models are considered afterwards.

Quantified Boolean Formula The instances of the problem *Quantified Boolean Formula* (QBF, sometimes also called QSAT) have the structure

$$Q_1 x_1 Q_2 x_2 \cdots Q_k x_k \colon F(x_1, \ldots, x_k)\,.$$

Here $F(x_1, \ldots, x_k)$ is a Boolean formula with variables $x_1, \ldots, x_k$ and connectives $\wedge$, $\vee$ and $\neg$. Each Q_j is one of the quantifiers $\forall$ or $\exists$. The problem is to decide whether the formula is true under the obvious interpretation. This problem is known to be complete for **PSPACE**. All known TM, i. e., all deterministic sequential algorithms, for solving QBF require exponential time.

Thus a proof that QBF can be solved by some model $\mathcal{M}$ in polynomial time (usually) implies that all problems in **PSPACE** can be solved by $\mathcal{M}$ in polynomial time. Often this can be paired with the "opposite" results that problems that can be solved in polynomial time on $\mathcal{M}$ are in **PSPACE**, and hence $\text{TM} - \textbf{PSPACE} = \mathcal{M} - \textbf{P}$.

This, of course, is not the case only for models "with trees"; see [6] for many alternatives.

Tree CA A tree CA (TCA for short) working on a full d-ary tree can be defined as follows: There is a set of states Q. For the root there is a local transition function $f_0 \colon (A \cup \{\Box\}) \times Q \times \mathbb{Q}^d \to Q$, which uses an input symbol (if available), the root cell's own state and those of the d child nodes to compute the next state of the node. And there are d local transition functions $f_i \colon Q \times Q \times Q^d \to Q$, where $1 \le i \le d$. The ith child of a node uses f_i to compute its new state depending the state of its parent node, its own state and the states of the d child nodes. For language recognition, input is provided sequentially to the root node during the first n steps and a blank symbol $\Box$ afterwards. A word is accepted if the root node enters an accepting state from a designated subset $F_+ \subset Q$.

Mycielski and Niwiński [19] were the first to realize that sequential polynomial reductions can be carried out by TCA and that QBF can be recognized by tree CA as well: A formula to be checked with k variables is copied and distributed to 2^k "evaluation cells". The sequences of left/right choices of the paths to them determine a valuation of the variables with zeros and ones. Each evaluation cell uses the subtree below it to evaluate $F(x_1, \ldots, x_k)$ accordingly. The results are propagated up to the root. Each cell in level i below the root, $1 \le i \le k$ above an evalua-

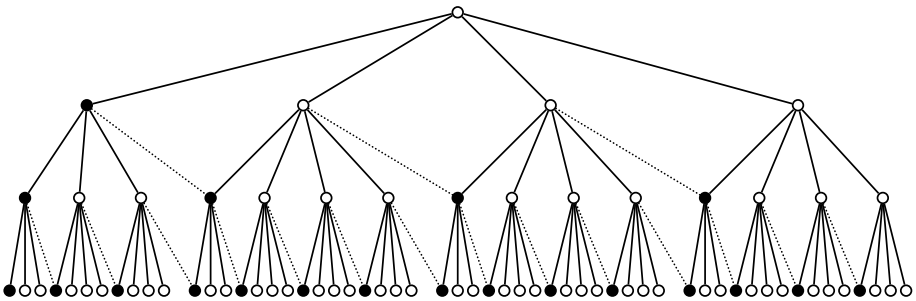

Cellular Automata as Models of Parallel Computation, Figure 3
The first levels of a tree of cells resulting from a tiling of the hyperbolic plane with 6-gons

tion cell combines the results using $\vee$ or $\wedge$, depending on whether the ith quantifier of the formula was $\exists$ or $\forall$.

On the other hand, it is routine work to prove that the result of a TCA running in polynomial time can be computed sequentially in polynomial space by a depth first procedure. Hence one gets:

Theorem 19

$$\text{TCA} - \textsc{Time}(\text{Pol}(n)) = \textbf{\textit{PSPACE}}$$

Thus tree cellular automata are in the second machine class.

Hyperbolic CA Two-dimensional CA as defined in Sect. "Introduction" can be considered as arising from the tessellation of the Euclidean plane $\mathbb{Z}^2$ with squares. Therefore, more generally, sometimes CA on a grid $G = \mathbb{Z}^d$ are called Euclidean CA. Analogously, some *Hyperbolic CA* arise from tessellation of the hyperbolic plane with some regular polygon. They are covered in depth in a separate article (▶ Cellular Automata in Hyperbolic Spaces).

Here we just consider one special case: The two-dimensional hyperbolic plane can be tiled with copies of the regular 6-gon with six right angles. If one considers only one quarter and draws a graph with the tiles as nodes and links between those nodes who share a common tile edge, one gets the graph depicted in Fig. 3. Basically it is a tree with two types of nodes, *black* and *white* ones, and some "additional" edges depicted as dotted lines. The root is a white node. The first child of each node is black. All other children are white; a black node has 2 white children, a white node has 3 white children.

For *hyperbolic CA* (HCA) one uses the formalism analogously to that described for tree CA. As one can see, basically HCA are trees with some additional edges. It is therefore not surprising that they can accept the languages from **PSPACE** in polynomial time. The inverse inclusion is also proved similarly to the tree case. This gives:

Theorem 20

$$\text{HCA} - \textsc{Time}(\text{Pol}(n)) = \textbf{\textit{PSPACE}}$$

It is also interesting to have a look at the analogs of **P**, **PSPACE**, and so on for hyperbolic CA. Somewhat surprisingly Iwamoto et al. [11] have shown:

Theorem 21

$$\begin{aligned}\text{HCA} - \textsc{Time}(\text{Pol}(n)) &= \text{HCA} - \textsc{Spc}(\text{Pol}(n))\\ &= \text{NHCA} - \textsc{Time}(\text{Pol}(n)) = \text{NHCA} - \textsc{Spc}(\text{Pol}(n))\end{aligned}$$

where NHCA *denotes nondeterministic hyperbolic CA. The analogous equalities hold for exponential time and space.*

Outlook There is yet another possibility for bringing trees into play: trees of configurations. The concept of *alternation* [4] can be carried over to cellular automata. Since there are several active computational units, the definitions are little bit more involved and it turns out that one has several possibilities which also result in models with slightly different properties. But in all cases one gets models from the second machine class. For results, readers are referred to [13] and [23].

On the other end there is some research on what happens if one restricts the possibilities for communication between neighboring cells: Instead of getting information about the complete states of the neighbors, in the extreme case only *one bit* can be exchanged. See for example [15].

Future Directions

At several points in this paper we have pointed out open problems which deserve further investigation. Here, we want to stress three areas which we consider particularly interesting in the area of "CA as a parallel model".

Proper Inclusions and Denser Hierarchies

It has been pointed out several times, that inclusions of complexity classes are not known to be proper, or that gaps between resource bounds still needed to be "large" in order to prove that the inclusion of the related classes is a proper one. For the foreseeable future it remains a wide area for further research. Most probably new techniques will have to be developed to make significant progress.

Activities

The motivation for considering state change complexity was the energy consumption of CMOS hardware. It is known that irreversible physical computational processes must consume energy. This seems not to be the case for reversible ones. Therefore *reversible* CA are also interesting in this respect. The definition of reversible CA and results for them are the topic of the article by Morita (► Reversible Cellular Automata), also in this encyclopedia. Surprisingly, all currently known simulations of irreversible CA on reversible ones (this is possible) exhibit a *large* state change complexity. This deserves further investigation.

Also the examination of CA which are "reversible on the computational core" has been started only recently [16]. There are first surprising results; the impacts on computational complexity are unforeseeable.

Asynchronicity and Randomization

Randomization is an important topic in sequential computing. It is high time that this is also investigated in much more depth for cellular automata. The same holds for cellular automata where not all cells are updating their states synchronously. These areas promise a wealth of new insights into the essence of fine-grained parallel systems.

Bibliography

1. Achilles AC, Kutrib M, Worsch T (1996) On relations between arrays of processing elements of different dimensionality. In: Vollmar R, Erhard W, Jossifov V (eds) Proceedings Parcella '96, no. 96 in Mathematical Research. Akademie, Berlin, pp 13–20
2. Blum M (1967) A machine-independent theory of the complexity of recursive functions. J ACM 14:322–336
3. Buchholz T, Kutrib M (1998) On time computability of functions in one-way cellular automata. Acta Inf 35(4):329–352
4. Chandra AK, Kozen DC, Stockmeyer LJ (1981) Alternation. J ACM 28(1):114–133
5. Delorme M, Mazoyer J, Tougne L (1999) Discrete parabolas and circles on 2D cellular automata. Theor Comput Sci 218(2):347–417
6. van Emde Boas P (1990) Machine models and simulations. In: van Leeuwen J (ed) Handbook of Theoretical Computer Science, vol A. Elsevier Science Publishers and MIT Press, Amsterdam, chap 1, pp 1–66
7. Fatès N, Thierry É, Morvan M, Schabanel N (2006) Fully asynchronous behavior of double-quiescent elementary cellular automata. Theor Comput Sci 362:1–16
8. Garzon M (1991) Models of Massive Parallelism. Texts in Theoretical Computer Science. Springer, Berlin
9. Hemmerling A (1979) Concentration of multidimensional tape-bounded systems of Turing automata and cellular spaces. In: Budach L (ed) International Conference on Fundamentals of Computation Theory (FCT '79). Akademie, Berlin, pp 167–174
10. Hennie FC (1965) One-tape, off-line Turing machine computations. Inf Control 8(6):553–578
11. Iwamoto C, Margenstern M (2004) Time and space complexity classes of hyperbolic cellular automata. IEICE Trans Inf Syst E87-D(3):700–707
12. Iwamoto C, Hatsuyama T, Morita K, Imai K (2002) Constructible functions in cellular automata and their applications to hierarchy results. Theor Comput Sci 270(1–2):797–809
13. Iwamoto C, Tateishi K, Morita K, Imai K (2003) Simulations between multi-dimensional deterministic and alternating cellular automata. Fundamenta Informaticae 58(3/4):261–271
14. Kutrib M (2008) Efficient pushdown cellular automata: Universality, time and space hierarchies. J Cell Autom 3(2):93–114
15. Kutrib M, Malcher A (2006) Fast cellular automata with restricted inter-cell communication: Computational capacity. In: Navarro YKG, Bertossi L (ed) Proceedings Theor Comput Sci (IFIP TCS 2006), pp 151–164
16. Kutrib M, Malcher A (2007) Real-time reversible iterative arrays. In: Csuhaj-Varjú E, Ésik Z (ed) Fundametals of Computation Theory 2007. LNCS vol 4639. Springer, Berlin, pp 376–387
17. Mazoyer J, Terrier V (1999) Signals in one-dimensional cellular automata. Theor Comput Sci 217(1):53–80
18. Merkle D, Worsch T (2002) Formal language recognition by stochastic cellular automata. Fundamenta Informaticae 52(1–3):181–199
19. Mycielski J, Niwiński D (1991) Cellular automata on trees, a model for parallel computation. Fundamenta Informaticae XV:139–144
20. Nakamura K (1981) Synchronous to asynchronous transformation of polyautomata. J Comput Syst Sci 23:22–37
21. von Neumann J (1966) Theory of Self-Reproducing Automata. University of Illinois Press, Champaign. Edited and completed by Arthur W. Burks
22. Regnault D, Schabanel N, Thierry É (2007) Progress in the analysis of stochastic 2d cellular automata: a study of asychronous 2d minority. In: Csuhaj-Varjú E, Ésik Z (ed) Fundametals of Computation Theory 2007. LNCS vol 4639. Springer, Berlin, pp 376–387
23. Reischle F, Worsch T (1998) Simulations between alternating CA, alternating TM and circuit families. In: MFCS'98 satellite workshop on cellular automata, pp 105–114
24. Sanders P, Vollmar R, Worsch T (2002) Cellular automata: Energy consumption and physical feasibility. Fundamenta Informaticae 52(1–3):233–248
25. Savitch WJ (1978) Parallel and nondeterministic time complexity classes. In: Proc. 5th ICALP, pp 411–424
26. Scheben C (2006) Simulation of d'-dimensional cellular automata on d-dimensional cellular automata. In: El Yacoubi S, Chopard B, Bandini S (eds) Proceedings ACRI 2006. LNCS, vol 4173. Springer, Berlin, pp 131–140
27. Smith AR (1971) Simple computation-universal cellular spaces. J ACM 18(3):339–353

28. Stoß HJ (1970) *k*-Band-Simulation von *k*-Kopf-Turing-Maschinen. Computing 6:309–317
29. Stratmann M, Worsch T (2002) Leader election in *d*-dimensional CA in time $diam \cdot \log(diam)$. Future Gener Comput Syst 18(7):939–950
30. Vollmar R (1982) Some remarks about the 'efficiency' of polyautomata. Int J Theor Phys 21:1007–1015
31. Wiedermann J (1983) Paralelný Turingov stroj – Model distribuovaného počítača. In: Gruska J (ed) Distribouvané a parallelné systémy. CRC, Bratislava, pp 205–214
32. Wiedermann J (1984) Parallel Turing machines. Tech. Rep. RUU-CS-84-11. University Utrecht, Utrecht
33. Worsch T (1997) On parallel Turing machines with multi-head control units. Parallel Comput 23(11):1683–1697
34. Worsch T (1999) Parallel Turing machines with one-head control units and cellular automata. Theor Comput Sci 217(1):3–30

Cellular Automata, Classification of

Klaus Sutner
Carnegie Mellon University, Pittsburgh, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Reversibility and Surjectivity
Definability and Computability
Computational Equivalence
Conclusion
Bibliography

Glossary

Cellular automaton For our purposes, a (one-dimensional) cellular automaton (CA) is given by a local map $\rho : \Sigma^w \to \Sigma$ where Σ is the underlying alphabet of the automaton and w is its width. As a data structure, suitable as input to a decision algorithm, a CA can thus be specified by a simple lookup table. We abuse notation and write $\rho(x)$ for the result of applying the global map of the CA to configuration $x \in \Sigma^{\mathbb{Z}}$.

Wolfram classes Wolfram proposed a heuristic classification of cellular automata based on observations of typical behaviors. The classification comprises four classes: evolution leads to trivial configurations, to periodic configurations, evolution is chaotic, evolution leads to complicated, persistent structures.

Undecidability It was recognized by logicians and mathematicians in the first half of the 20th century that there is an abundance of well-defined problems that cannot be solved by means of an algorithm, a mechanical procedure that is guaranteed to terminate after finitely many steps and produce the appropriate answer. The best known example of an undecidable problem is Turing's Halting Problem: there is no algorithm to determine whether a given Turing machine halts when run on an empty tape.

Semi-decidability A problem is said to be semi-decidable or computably enumerable if it admits an algorithm that returns "yes" after finitely many steps if this is indeed the correct answer. Otherwise the algorithm never terminates. The Halting Problem is the standard example for a semi-decidable problem. A problem is decidable if, and only if, the problem itself and its negation are semi-decidable.

Universality A computational device is universal it is capable of simulating any other computational device. The existence of universal computers was another central insight of the early days of computability theory and is closely related to undecidability.

Reversibility A discrete dynamical system is reversible if the evolution of the system incurs no loss of information: the state at time t can be recovered from the state at time $t + 1$. For CAs this means that the global map is injective.

Surjectivity The global map of a CA is surjective if every configuration appears as the image of another. By contrast, a configuration that fails to have a predecessor is often referred to as a Garden-of-Eden.

Finite configurations One often considers CA with a special quiescent state: the homogeneous configuration where all cells are in the quiescent state is required to be fixed point under the global map. Infinite configurations where all but finitely many cells are in the quiescent state are often called finite configurations. This is somewhat of a misnomer; we prefer to speak about configurations with finite support.

Definition of the Subject

Cellular automata display a large variety of behaviors. This was recognized clearly when extensive simulations of cellular automata, and in particular one-dimensional CA, became computationally feasible around 1980. Surprisingly, even when one considers only elementary CA, which are constrained to a binary alphabet and local maps involving only nearest neighbors, complicated behaviors are observed in some cases. In fact, it appears that most behaviors observed in automata with more states and larger neighborhoods already have qualitative analogues in the realm of elementary CA. Careful empirical studies lead Wolfram to suggest a phenomenological classification of CA based

on the long-term evolution of configurations, see [68,71] and Sect. "Introduction". While Wolfram's four classes clearly capture some of the behavior of CA it turns out that any attempt at formalizing this taxonomy meets with considerable difficulties. Even apparently simple questions about the behavior of CA turn out to be algorithmically undecidable and it is highly challenging to provide a detailed mathematical analysis of these systems.

Introduction

In the early 1980's Wolfram published a collection of 20 open problems in the the theory of CA, see [69]. The first problem on his list is "What overall classification of cellular automata behavior can be given?" As Wolfram points out, experimental mathematics provides a first answer to this problem: one performs a large number of explicit simulations and observes the patterns associated with the long term evolution of a configuration, see [67,71]. Wolfram proposed a classification that is based on extensive simulations in particular of one-dimensional cellular automata where the evolution of a configuration can be visualized naturally as a two-dimensional image. The classification involves four classes that can be described as follows:

- *W1*: Evolution leads to homogeneous fixed points.
- *W2*: Evolution leads to periodic configurations.
- *W3*: Evolution leads to chaotic, aperiodic patterns.
- *W4*: Evolution produces persistent, complex patterns of localized structures.

Thus, Wolfram's first three classes follow closely concepts from continuous dynamics: fixed point attractors, periodic attractors and strange attractors, respectively. They correspond roughly to systems with zero temporal and spatial entropy, zero temporal entropy but positive spatial entropy, and positive temporal and spatial entropy, respectively. *W4* is more difficult to associate with a continuous analogue except to say that transients are typically very long. To understand this class it is preferable to consider CA as models of massively parallel computation rather than as particular discrete dynamical systems. It was conjectured by Wolfram that *W4* automata are capable of performing complicated computations and may often be computationally universal. Four examples of elementary CA that are typical of the four classes are shown in Fig. 1. Li and Packard [32,33] proposed a slightly modified version of this hierarchy by refining the low classes and in particular Wolfram's *W2*. Much like Wolfram's classification, the Li–Packard classification is concerned with the asymptotic behavior of the automaton, the structure and behavior of the limiting configurations. Here is one version of the Li–Packard classification, see [33].

- *LP1*: Evolution leads to homogeneous fixed points.
- *LP2*: Evolution leads to non-homogeneous fixed points, perhaps up a to a shift.
- *LP3*: Evolution leads to ultimately periodic configurations. Regions with periodic behavior are separated by domain walls, possibly up to a shift.
- *LP4*: Configurations produce locally chaotic behavior. Regions with chaotic behavior are separated by domain walls, possibly up to a shift.
- *LP5*: Evolution leads to chaotic patterns that are spatially unbounded.
- *LP6*: Evolution is complex. Transients are long and lead to complicated space-time patterns which may be non-monotonic in their behavior.

By contrast, a classification closer to traditional dynamical systems theory was introduced by Kůrka, see [27,28]. The classification rests on the notions of equicontinuity, sensitivity to initial conditions and expansivity. Suppose x is a point in some metric space and f a map on that space. Then f is *equicontinuous* at x if

$$\forall\, \varepsilon > 0\, \exists\, \delta > 0\, \forall\, y \in B_\delta(x), n \in \mathbb{N}\, (d(f^n(x), f^n(y)) < \varepsilon)$$

where $d(.,.)$ denotes a metric. Thus, all points in a sufficiently small neighborhood of x remain close to the iterates of x for the whole orbit. Global equicontinuity is a fairly strong condition, it implies that the limit set of the automaton is reached after finitely many steps. The map is *sensitive* (to initial conditions) if

$$\forall\, x, \varepsilon > 0\, \exists\, \delta > 0\, \forall\, y \in B_\delta(x) \exists\, n \in \mathbb{N} \quad (d(f^n(x), f^n(y)) \geq \varepsilon)\,.$$

Lastly, the map is *positively expansive* if

$$\exists\, \varepsilon > 0\, \forall\, x \neq y\, \exists\, n \in \mathbb{N}\, (d(f^n(x), f^n(y)) \geq \varepsilon)\,.$$

Kůrka's classification then takes the following form.

- *K1*: All points are equicontinuous under the global map.
- *K2*: Some but not all points are equicontinuous under the global map.
- *K3*: The global map is sensitive but not positively expansive.
- *K4*: The global map is positively expansive.

This type of classification is perfectly suited to the analysis of uncountable spaces such as the Cantor space $\{0, 1\}^{\mathbb{N}}$

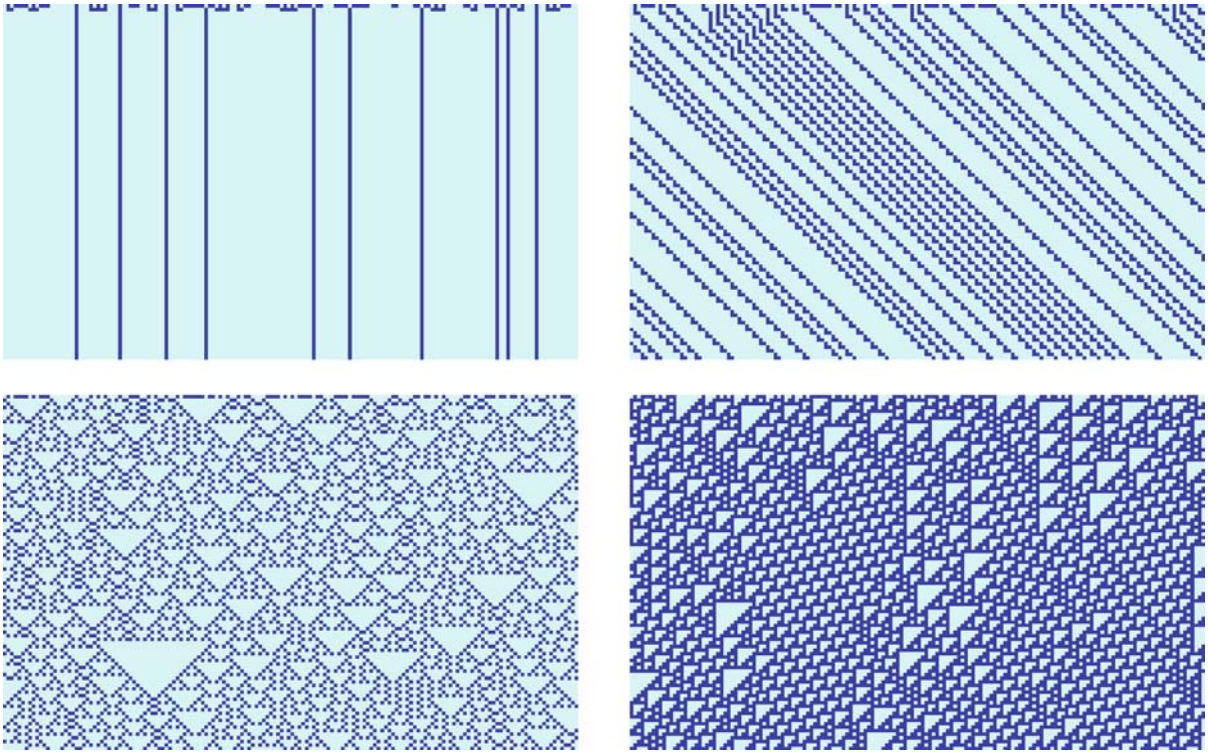

Cellular Automata, Classification of, Figure 1
Typical examples of the behavior described by Wolfram's classes among elementary cellular automata

or the full shift space $\Sigma^{\mathbb{Z}}$ which carry a natural metric structure. For the most part we will not pursue the analysis of CA by topological and measure theoretic means here and refer to ► Topological Dynamics of Cellular Automata in this volume for a discussion of these methods. See Sect. "Definability and Computability" for the connections between topology and computability.

Given the apparent complexity of observable CA behavior one might suspect that it is difficult to pinpoint the location of an arbitrary given CA in any particular classification scheme with any precision. This is in contrast to simple parameterizations of the space of CA rules such as Langton's λ parameter that are inherently easy to compute. Briefly, the λ value of a local map is the fraction of local configurations that map to a non-zero value, see [29,33]. Small λ values result in short transients leading to fixed points or simple periodic configurations. As λ increases the transients grow longer and the orbits become more and more complex until, at last, the dynamics become chaotic. Informally, sweeping the λ value from 0 to 1 will produce CA in *W1*, then *W2*, then *W4* and lastly in *W3*. The last transition appears to be associated with a threshold phenomenon. It is unclear what the connection between Langton's λ-value and computational properties of a CA is, see [37,46]. Other numerical measures that appear to be loosely connected to classifications are the mean field parameters of Gutowitz [20,21], the Z-parameter by Wuensche [72], see also [44]. It seems doubtful that a structured taxonomy along the lines of Wolfram or Li–Packard can be derived from a simple numerical measure such as the λ value alone, or even from a combination of several such values. However, they may be useful as empirical evidence for membership in a particular class.

Classification also becomes significantly easier when one restricts one's attention to a limited class of CA such as additive CA, see ► Additive Cellular Automata. In this context, additive means that the local rule of the automaton has the form $\rho(\vec{x}) = \sum_i c_i x_i$ where the coefficients as well as the states are modular numbers. A number of properties starting with injectivity and surjectivity as well as topological properties such as equicontinuity and sensitivity can be expressed in terms of simple arithmetic conditions on the rule coefficients. For example, equicontinuity is equivalent to all prime divisors of the modulus m dividing all coefficients c_i, $i > 1$, see [35] and the references therein. It is also noteworthy that in the linear case meth-

ods tend to carry over to arbitrary dimensions; in general there is a significant step in complexity from dimension one to dimension two.

No claim is made that the given classifications are complete; in fact, one should think of them as prototypes rather than definitive taxonomies. For example, one might add the class of *nilpotent* CA at the bottom. A CA is nilpotent if all configurations evolve to a particular fixed point after finitely many steps. Equivalently, by compactness, there is a bound n such that all configurations evolve to the fixed point in no more than n steps. Likewise, we could add the class of *intrinsically universal* CA at the top. A CA is intrinsically universal if it is capable of simulating all other CA of the same dimension in some reasonable sense. For a fairly natural notion of simulation see [45]. At any rate, considerable effort is made in the references to elaborate the characteristics of the various classes. For many concrete CA visual inspection of the orbits of a suitable sample of configurations readily suggests membership in one of the classes.

Reversibility and Surjectivity

A first tentative step towards the classification of a dynamical systems is to determine its reversibility or lack thereof. Thus we are trying to determine whether the evolution of the system is associated with loss of information, or whether it is possible to reconstruct the state of the system at time t from its state at time $t+1$. In terms of the global map of the system we have to decide injectivity. Closely related is the question whether the global map is surjective, i. e., whether there is no Garden-of-Eden: every configuration has a predecessor under the global map. As a consequence, the limit set of the automaton is the whole space. It was shown of Hedlund that for CA the two notions are connected: every reversible CA is also surjective, see [24], ► Reversible Cellular Automata. As a matter of fact, reversibility of the global map of a CA implies openness of the global map, and openness implies surjectivity. The converse implications are both false. By a well-known theorem by Hedlund [24] the global maps of CA are precisely the continuous maps that commute with the shift. It follows from basic topology that the inverse global map of a reversible CA is again the global map of a suitable CA. Hence, the predecessor configuration of a given configuration can be reconstructed by another suitably chosen CA. For results concerning reversibility on the limit set of the automaton see [61].

From the perspective of complexity the key result concerning reversible systems is the work by Lecerf [30] and Bennett [7]. They show that reversible Turing machines can compute any partial recursive function, modulo a minor technical problem: In a reversible Turing machine there is no loss of information; on the other hand even simple computable functions are clearly irreversible in the sense that, say, the sum of two natural numbers does not determine these numbers uniquely. To address this issue one has to adjust the notion of computability slightly in the context of reversible computation: given a partial recursive function $f : \mathbb{N} \to \mathbb{N}$ the function $\hat{f}(x) = \langle x, f(x)\rangle$ can be computed by a reversible Turing machine where $\langle .,. \rangle$ is any effective pairing function. If f itself happens to be injective then there is no need for the coding device and f can be computed by a reversible Turing machine directly. For example, we can compute the product of two primes reversibly. Morita demonstrated that the same holds true for one-dimensional cellular automata [38,40,62], ► Tiling Problem and Undecidability in Cellular Automata: reversibility is no obstruction to computational universality. As a matter of fact, any irreversible cellular automaton can be simulated by a reversible one, at least on configurations with finite support. Thus one should expect reversible CA to exhibit fairly complicated behavior in general.

For infinite, one-dimensional CA it was shown by Amoroso and Patt [2] that reversibility is decidable. Moreover, it is decidable if the the global map is surjective. An efficient practical algorithm using concepts of automata theory can be found in [55], see also [10,14,23]. The fast algorithm is based on interpreting a one-dimensional CA as a deterministic transducer, see [6,48] for background. The underlying semi-automaton of the transducer is a de Bruijn automaton $\mathcal{B}$ whose states are words in Σ^{w-1} where Σ is the alphabet of the CA and w is its width. The transitions are given by $ax \xrightarrow{c} xb$ where $a, b, c \in \Sigma$, $x \in \Sigma^{w-2}$ and $c = \rho(axb)$, ρ being the local map of the CA. Since $\mathcal{B}$ is strongly connected, the product automaton of $\mathcal{B}$ will contain a strongly connected component C that contains the diagonal D, an isomorphic copy of $\mathcal{B}$. The global map of the CA is reversible if, and only if, $C = D$ is the only non-trivial component. It was shown by Hedlund [24] that surjectivity of the global map is equivalent with local injectivity: the restriction of the map to configurations with finite support must be injective. The latter property holds if, and only if, $C = D$ and is thus easily decidable. Automata theory does not readily generalize to words of dimensions higher than one. Indeed, reversibility and surjectivity in dimensions higher than one are undecidable, see [26] and ► Tiling Problem and Undecidability in Cellular Automata in this volume for the rather intricate argument needed to establish this fact.

While the structure of reversible one-dimensional CA is well-understood, see ► Tiling Problem and Undecid-

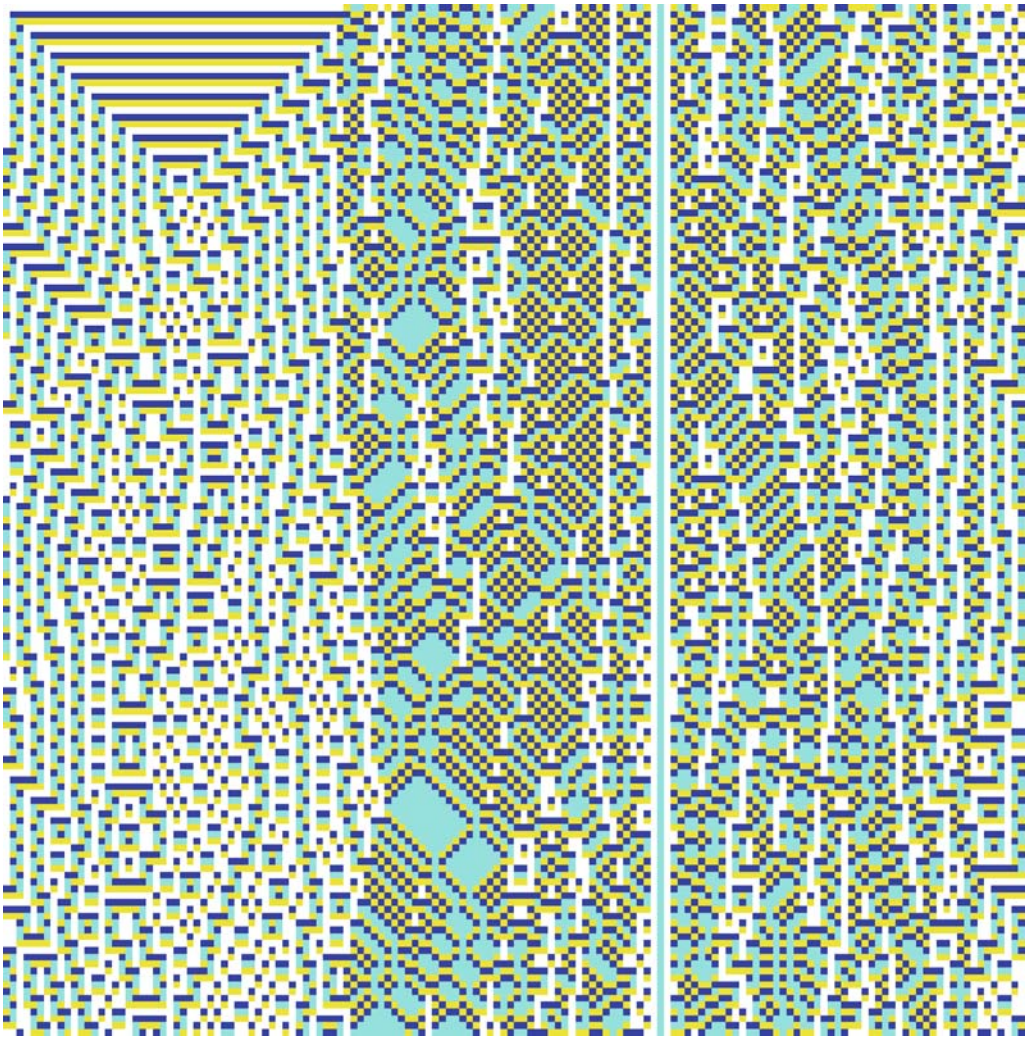

Cellular Automata, Classification of, Figure 2
A reversible automaton obtained by applying Fredkin's construction to the irreversible elementary CA 77

ability in Cellular Automata, [16], and while there is an efficient algorithm to check reversibility, few methods are known that allow for the construction of interesting reversible CA. There is a noteworthy trick due to Fredkin that exploits the reversibility of the Fibonacci equation $X_{n+1} = X_n + X_{n-1}$. When addition is interpreted as exclusive or this can be used to construct a second-order CA from any given binary CA; the former can then be recoded as a first-order CA over a 4-letter alphabet. For example, for the open but irreversible elementary CA number 90 we obtain the CA shown in Fig. 2.

Another interesting class of reversible one-dimensional CA, the so-called *partitioned cellular automata (PCA)*, is due to Morita and Harao, see [38,39,40]. One can think of a PCA as a cellular automaton whose cells are divided into multiple tracks; specifically Morita uses an alphabet of the form $\Sigma = \Sigma_1 \times \Sigma_2 \times \Sigma_3$. The configurations of the automaton can be written as (X, Y, Z) where $X \in {\Sigma_1}^{\mathbb{Z}}$, $Y \in {\Sigma_2}^{\mathbb{Z}}$ and $Z \in {\Sigma_3}^{\mathbb{Z}}$. Now consider the *shearing map* σ defined by $\sigma(X, Y, Z) = (\mathsf{RS}(X), Y, \mathsf{LS}(Z))$ where RS and LS denote the right and left shift, respectively. Given any function $f : \Sigma \to \Sigma$ we can define a global map $f \circ \sigma$ where f is assumed to be applied point-wise. Since the shearing map is bijective, the CA will be reversible if, and only if, the map f is bijective. It is relatively easy to construct bijections f that cause the CA to perform particular computational tasks, even when a direct construction appears to be entirely intractable.

Definability and Computability

Formalizing Wolfram's Classes

Wolfram's classification is an attempt to categorize the complexity of the CA by studying the patterns observed during the long-term evolution of all configurations. The first two classes are relatively easy to observe, but it is difficult to distinguish between the last two classes. In particular *W4* is closely related to the kind of behavior that would be expected in connection with systems that are capable of performing complicated computations, including the ability to perform universal computation; a property that is notoriously difficult to check, see [52]. The focus on the full configuration space rather than a significant subset thereof corresponds to the worst-case approach well-known in complexity theory and is somewhat inferior to an average case analysis. Indeed, Baldwin and Shelah point out that a product construction can be used to design a CA whose behavior is an amalgamation of the behavior of two given CA, see [3,4]. By combining CA in different classes one obtains striking examples of the weakness of the worst-case approach. A natural example of this mixed type of behavior is elementary CA 184 which displays class II or class III behavior, depending on the initial configuration. Another basic example for this type of behavior is the well-studied elementary CA 30, see Sect. "Conclusion".

Still, for many CA a worst-case classification seems to provide useful information about the structural properties of the automaton. The first attempt at formalizing Wolfram's class was made by Culik and Yu who proposed the following hierarchy, given here in cumulative form, see [11]:

- *CY1*: All configurations evolve to a fixed point.
- *CY2*: All configurations evolve to a periodic configuration.
- *CY3*: The orbits of all configurations are decidable.
- *CY4*: No constraints.

The Culik–Yu classification employs two rather different methods. The first two classes can be defined by a simple formula in a suitable logic whereas the third (and the fourth in the disjoint version of the hierarchy) rely on notions of computability theory. As a general framework for both approaches we consider *discrete dynamical systems*, structures of the form $\mathcal{A} = \langle C, \rightarrow \rangle$ where $C \subseteq \Sigma^{\mathbb{Z}}$ is the space of *configurations* of the system and $\rightarrow$ is the "next configuration" relation on C. We will only consider the deterministic case where for each configuration x there exists precisely one configuration y such that $x \rightarrow y$. Hence we are really dealing with algebras with one unary function, but iteration is slightly easier to deal with in the relational setting. The structures most important in this context are the ones arising from a CA. For any local map ρ we consider the structure $\mathcal{A}_\rho = \langle C, \rightarrow \rangle$ where the next configuration relation is determined by $x \rightarrow \rho(x)$.

Using the standard language of first order logic we can readily express properties of the CA in terms of the system $\mathcal{A}_\rho$. For example, the system is reversible, respectively surjective, if the following assertions are valid over $\mathcal{A}$:

$$\forall\, x, y, z\, (x \rightarrow z \text{ and } y \rightarrow z \text{ implies } x = y)\,,$$
$$\forall\, x\, \exists\, y\, (y \rightarrow x)\,.$$

As we have seen, both properties are easily decidable in the one-dimensional case. In fact, one can express the basic predicate $x \rightarrow y$ (as well as equality) in terms of finite state machines on infinite words. These machines are defined like ordinary finite state machines but the acceptance condition requires that certain states are reached infinitely and co-infinitely often, see [8,19]. The emptiness problem for these automata is easily decidable using graph theoretic algorithms. Since regular languages on infinite words are closed under union, complementation and projection, much like their finite counterparts, and all the corresponding operations on automata are effective, it follows that one can decide the validity of first order sentences over $\mathcal{A}_\rho$ such as the two examples above: the model-checking problem for these structures and first order logic is decidable, see [34]. For example, we can decide whether there is a configuration that has a certain number of predecessors. Alternatively, one can translate these sentences into monadic second order logic of one successor, and use well-known automata-based decision algorithms there directly, see [8]. Similar methods can be used to handle configurations with finite support, corresponding to weak monadic second order logic. Since the complexity of the decision procedure is non-elementary one should not expect to be able to handle complicated assertions. On the other hand, at least for weak monadic second order logic practical implementations of the decision method exist, see [17]. There is no hope of generalizing this approach as the undecidability of, say, reversibility in higher dimensions demonstrates.

Write $x \xrightarrow{t} y$ if x evolves to y in exactly t steps, $x \xrightarrow{+} y$ if x evolves to y in any positive number of steps and $x \xrightarrow[t]{*} y$ if x evolves to y in any number of steps. Note that $\xrightarrow{t}$ is definable for each fixed t, but $\xrightarrow{*}$ fails to be so definable in first order logic. This is in analogy to the undefinability of path existence problems in the first order theory of graphs, see [34]. Hence it is natural to extend our language so we can express iterations of the global map, either by adding

transitive closures or by moving to some limited system of higher order logic over $\mathcal{A}_\rho$ where $\xrightarrow{*}$ is definable, see [8].

Arguably the most basic decision problem associated with a system $\mathcal{A}$ that requires iteration of the global map is the *Reachability Problem*: given two configurations x and y, does the evolution of x lead to y? A closely related but different question is the *Confluence Problem*: will two configurations x and y evolve to the same limit cycle? Confluence is an equivalence relation and allows for the decomposition of configuration space into limit cycles together with their basins of attraction. The Reachability and Confluence Problem amount to determining, given configurations x and y, whether

$$x \xrightarrow{*} y,$$

$$\exists z\,(x \xrightarrow{*} z \text{ and } y \xrightarrow{*} z)\,,$$

respectively. As another example, the first two Culik–Yu class can be defined like so:

$$\forall x\,\exists z\,(x \xrightarrow{*} z \text{ and } z \to z),$$

$$\forall x\,\exists z\,(x \xrightarrow{*} z \text{ and } z \xrightarrow{+} z)\,.$$

It is not difficult to give similar definitions for the lower Li–Packard classes if one extends the language by a function symbol denoting the shift operator.

The third Culik–Yu class is somewhat more involved. By definition, a CA lies in the third class if it admits a global decision algorithm to determine whether a given configuration x evolves to another given configuration y in a finite number of steps. In other words, we are looking for automata where the Reachability Problem is algorithmically solvable. While one can agree that *W4* roughly translates into undecidability and is thus properly situated in the hierarchy, it is unclear how chaotic patterns in *W3* relate to decidability. No method is known to translate the apparent lack of tangible, persistent patterns in rules such as elementary CA 30 into decision algorithms for Reachability. There is another, somewhat more technical problem to overcome in formalizing classifications. Recall that the full configuration space is $C = \Sigma^{\mathbb{Z}}$. Intuitively, given $x \in C$ we can effectively determine the next configuration $y = \rho(x)$. However, classical computability theory does not deal with infinitary objects such as arbitrary configuration so a bit of care is needed here. The key insight is that we can determine arbitrary finite segments of $\rho(x)$ using only finite segments of x (and, of course, the lookup table for the local map). There are several ways to model computability on $\Sigma^{\mathbb{Z}}$ based on this idea of finite approximations, we refer to [66] for a particularly appealing model based on so-called type-2 Turing machines; the reference also contains many pointers to the literature as well as a comparison between the different approaches. It is easy to see that for any CA the global map ρ as well as all its iterates ρ^t are computable, the latter uniformly in t. However, due to the finitary nature of all computations, equality is not decidable in type-2 computability: the unequal operator $U_0(x, y) = 0$ if $x \neq y$, $U_0(x, y)$ undefined otherwise, is computable and thus unequality is semi-decidable, but the stronger $U_0(x, y) = 0$ if $x \neq y$, $U_0(x, y) = 1$, otherwise, is not computable. The last result is perhaps somewhat counterintuitive, but it is inevitable if we strictly adhere to the finite approximation principle.

In order to avoid problems of this kind it has become customary to consider certain subspaces of the full configuration space, in particular C_{fin}, the collection of configurations with finite support, C_{per}, the collection of spatially periodic configurations and C_{ap}, the collection of almost periodic configurations of the form $\ldots uuuwvvv \ldots$ where u, v and w are all finite words over the alphabet of the automaton. Thus, an almost periodic configuration differs from a configuration of the form ${}^{\omega}u\,v^{\omega}$ in only finitely many places. Configurations with finite support correspond to the special case where $u = v = 0$ is a special quiescent symbol and spatially periodic configurations correspond to $u = v$, $w = \varepsilon$. The most general type of configuration that admits a finitary description is the class C_{rec} of recursive configurations, where the assignment of state to a cell is given by a computable function.

It is clear that all these subspaces are closed under the application of a global map. Except for C_{fin} there are also closed under inverse maps in the following sense: given a configuration y in some subspace that has a predecessor x in C_{all} there already exists a predecessor in the same subspace, see [55,58]. This is obvious except in the case of recursive configurations. The reference also shows that the recursive predecessor cannot be computed effectively from the target configuration. Thus, for computational purposes the dynamics of the cellular automaton are best reflected in C_{ap}: it includes all configuration with finite support and we can effectively trace an orbit in both directions. It is not hard to see that C_{ap} is the least such class. Alas, it is standard procedure to avoid minor technical difficulties arising from the infinitely repeated spatial patterns and establish classifications over the subspace C_{fin}. There is a arguably not much harm in this simplification since C_{fin} is a dense subspace of C_{all} and compactness can be used to lift properties from C_{fin} to the full configuration space.

The Culik–Yu hierarchy is correspondingly defined over C_{fin}, the class of all configurations of finite support. In this setting, the first three classes of this hierarchy are

undecidable and the fourth is undecidable in the disjunctive version: there is no algorithm to test whether a CA admits undecidable orbits. As it turns out, the CA classes are complete in their natural complexity classes within the arithmetical hierarchy [50,52]. Checking membership in the first two classes comes down to performing an infinite number of potentially unbounded searches and can be described logically by a Π_2 expression, a formula of type $\forall\, x\, \exists\, y\, R(x, y)$ where R is a decidable predicate. Indeed, *CY1* and *CY2* are both Π_2-complete. Thus, deciding whether all configurations on a CA evolve to a fixed point is equivalent to the classical problem of determining whether a semi-decidable set is infinite. The third class is even less amenable to algorithmic attack; one can show that *CY3* is Σ_3-complete, see [53]. Thus, deciding whether all orbits are decidable is as difficult as determining whether any given semi-decidable set is decidable. It is not difficult to adjust these undecidability results to similar classes such as the lower levels of the Li–Packard hierarchy that takes into account spatial displacements of patterns.

Effective Dynamical Systems and Universality

The key property of CA that is responsible for all these undecidability results is the fact that CA are capable of performing arbitrary computations. This is unsurprising when one defines computability in terms of Turing machines, the devices introduced by Turing in the 1930's, see [47,63]. Unlike the Gödel–Herbrand approach using general recursive functions or Church's λ-calculus, Turing's devices are naturally closely related to discrete dynamical systems. For example, we can express an instantaneous description of a Turing machine as a finite sequence

$$a_{-l}\, a_{-l+1} \ldots a_{-1}\; p\; a_1 a_2 \ldots a_r$$

where the a_i are tape symbols and p is a state of the machine, with the understanding that the head is positioned at a_1 and that all unspecified tape cells contain the blank symbol. Needless to say, these Turing machine configurations can also be construed as finite support configurations of a one-dimensional CA. It follows that a one-dimensional CA can be used to simulate an arbitrary Turing machine, hence CA are computational universal: any computable function whatsoever can already be computed by a CA.

Note, though, that the simulation is not entirely trivial. First, we have to rely on input/output conventions. For example, we may insist that objects in the input domain, typically tuples of natural numbers, are translated into a configuration of the CA by a primitive recursive coding function. Second, we need to adopt some convention that determines when the desired output has occurred: we follow the evolution of the input configuration until some "halting" condition applies. Again, this condition must be primitive recursively decidable though there is considerable leeway as to how the end of a computation should be signaled by the CA. For example, we could insist that a particular cell reaches a special state, that an arbitrary cell reaches a special state, that the configuration be a fixed point and so forth. Lastly, if and when a halting configuration is reached, we a apply a primitive recursive decoding function to obtain the desired output.

Restricting the space to configurations that have finite support, that are spatially periodic, and so forth, produces an *effective dynamical system*: the configurations can be coded as integers in some natural way, and the next configuration relation is primitive recursive in the sense that the corresponding relation on code numbers is so primitive recursive. A classical example for an effective dynamical system is given by selecting the instantaneous descriptions of a Turing machine M as configurations, and one-step relation of the Turing machine as the operation of C. Thus we obtain a system $\mathcal{A}_M$ whose orbits represent the computations of the Turing machine. Likewise, given the local map ρ of a CA we obtain a system $\mathcal{A}_\rho$ whose operation is the induced global map. While the full configuration space C_{all} violates the effectiveness condition, any of the spaces C_{per}, C_{fin}, C_{ap} and C_{rec} will give rise to an effective dynamical system. Closure properties as well as recent work on the universality of elementary CA 110, see Sect. "Conclusion", suggests that the class of almost periodic configurations, also known as backgrounds or wallpapers, see [9,58], is perhaps the most natural setting. Both C_{fin} and C_{ap} provide a suitable setting for a CA that simulates a Turing machine: we can interpret $\mathcal{A}_M$ as a subspace of $\mathcal{A}_\rho$ for some suitably constructed one-dimensional CA ρ; the orbits of the subspace encode computations of the Turing machine. It follows from the undecidability of the Halting Problem for Turing machines that the Reachability Problem for these particular CA is undecidable.

Note, though, that orbits in $\mathcal{A}_M$ may well be finite, so some care must be taken in setting up the simulation. For example, one can translate halting configurations into fixed points. Another problem is caused by the worst-case nature of our classification schemes: in Turing machines and their associated systems $\mathcal{A}_M$ it is only behavior on specially prepared initial configurations that matters, whereas the behavior of a CA depends on all configurations. The behavior of a Turing machine on all instantaneous descriptions, rather than just the ones that can occur during a legitimate computation on some actual input, was first studied by Davis, see [12,13], and also Hooper [25].

Call a Turing machine *stable* if it halts on any instantaneous description whatsoever. With some extra care one can then construct a CA that lies in the first Culik–Yu class, yet has the same computational power as the Turing machine. Davis showed that every total recursive function can already be computed by a stable Turing machine, so membership in *CY1* is not an impediment to considerable computational power. The argument rests on a particular decomposition of recursive functions. Alternatively, one directly manipulate Turing machines to obtain a similar result, see [49,53]. On the other hand, unstable Turing machines yield a natural and coding-free definition of universality: a Turing machine is *Davis-universal* if the set of all instantaneous description on which the machine halts is Σ_1-complete.

The mathematical theory of infinite CA is arguably more elegant than the actually observable finite case. As a consequence, classifications are typically concerned with CA operating on infinite grids, so that even a configuration with finite support can carry arbitrarily much information. If we restrict our attention to the space of configurations on a finite grid a more fine-grained analysis is required. For a finite grid of size n the configuration space has the form $C_n = [n] \to \Sigma$ and is itself finite, hence any orbit is ultimately periodic and the Reachability Problem is trivially decidable. However, in practice there is little difference between the finite and infinite case. First, computational complexity issues make it practically impossible to analyze even systems of modest size. The Reachability Problem for finite CA, while decidable, is PSPACE-complete even in the one-dimensional case. Computational hardness appears in many other places. For example, if we try to determine whether a given configuration on a finite grid is a Garden-of-Eden the problem turns out to be NLOG-complete in dimension one and $\mathbb{NP}$-complete in all higher dimensions, see [56].

Second, it stands to reason that the more interesting classification problem in the finite case takes the following parameterized form: given a local map together with boundary conditions, determine the behavior of ρ on all finite grids. Under periodic boundary conditions this comes down to the study of C_{per} and it seems that there is little difference between this and the fixed boundary case. Since all orbits on a finite grid are ultimately periodic one needs to apply a more fine-grained classification that takes into account transient lengths. It is undecidable whether all configurations on all finite grids evolve to a fixed point under a given local map, see [54]. Thus, there is no algorithm to determine whether

$$\langle C_n, \to \rangle \models \forall\, x\, \exists\, z\, (x \overset{*}{\to} z \text{ and } z \to z)$$

for all grid sizes n. The transient lengths are trivially bounded by k^n where k is the size of the alphabet of the automaton. It is undecidable whether the transient lengths grow according to some polynomial bound, even when the polynomial in question is constant.

Restrictions of the configuration space are one way to obtain an effective dynamical system. Another is to interpret the approximation-based notion of computability on the full space in terms of topology. It is well-known that computable maps $C_{\mathsf{all}} \to C_{\mathsf{all}}$ are continuous in the standard product topology. The clopen sets in this topology are the finite unions of cylinder sets where a cylinder set is determined by the values of a configuration in finitely many places. By a celebrated result of Hedlund the global maps of a CA on the full space are characterized by being continuous and shift-invariant. Perhaps somewhat counter-intuitively, the decidable subsets of C_{all} are quite weak, they consist precisely of the clopen sets. Now consider a partition of C_{all} into finitely many clopen sets $C_0, C_2, \ldots, C_{n-1}$. Thus, it is decidable which block of the partition a given point in the space belongs to. Moreover, Boolean operations on clopen sets as well as application of the global map and the inverse global map are all computable. The partition affords a natural projection $\pi : C_{\mathsf{all}} \to \Sigma_n$ where $\Sigma_n = \{0, 1, \ldots, n-1\}$ and $\pi(x) = i$ iff $x \in C_i$. Hence the projection translates orbits in the full space C_{all} into a class W of ω-words over Σ_n, the symbolic orbits of the system. The Cantor space $\Sigma_n^{\mathbb{Z}}$ together with the shift describes all logically possible orbits with respect to the given partition and W describes the symbolic orbits that actually occur in the given CA. The shift operator corresponds to an application of the global map of the CA. The finite factors of W provide information about possible finite traces of an orbit when filtered through the given partition. Whole orbits, again filtered through the partition, can be described by ω-words. To tackle the classification of the CA in terms of W it was suggested by Delvenne et al., see [15], to refer to the CA as *decidable* if there it is decidable whether W has non-empty intersection with a ω-regular language. Alas, decidability in this sense is very difficult, its complexity being Σ_1^1-complete and thus outside of the arithmetical hierarchy. Likewise it is suggested to call a CA *universal* if the problem of deciding whether the cover of W, the collection of all finite factors, is Σ_1-complete, in analogy to Davis-universality.

Computational Equivalence

In recent work, Wolfram suggests a so-called *Principle of Computational Equivalence*, or PCE for short, see [71],

p. 717. PCE states that most computational processes come in only two flavors: they are either of a very simple kind and avoid undecidability, or they represent a universal computation and are therefore no less complicated than the Halting Problem. Thus, Wolfram proposes a zero-one law: almost all computational systems, and thus in particular all CA, are either as complicated as a universal Turing machine or are computationally simple. As evidence for PCE Wolfram adduces a very large collection of simulations of various effective dynamical systems such as Turing machines, register machines, tag systems, rewrite systems, combinators, and cellular automata. It is pointed out in Chap. 3 of [71], that in all these classes of systems there are surprisingly small examples that exhibit exceedingly complicated behavior – and presumably are capable of universal computation. Thus it is conceivable that universality is a rather common property, a property that is indeed shared by all systems that are not obviously simple. Of course, it is often very difficult to give a complete proof of the computational universality of a natural system, as opposed to carefully constructed one, so it is not entirely clear how many of Wolfram's examples are in fact universal. As a case in point consider the universality proof of Conway's Game of Life, or the argument for elementary CA 110. If Wolfram's PCE can be formally established in some form it stands to reason that it will apply to all effective dynamical systems and in particular to CA. Hence, classifications of CA would be rather straightforward: at the top there would be the class of universal CA, directly preceded by a class similar to the third Culik–Yu class, plus a variety of subclasses along the lines of the lower Li–Packard classes.

The corresponding problem in classical computability theory was first considered in the 1930's by Post and is now known as Post's Problem: is there a semi-decidable set that fails to be decidable, yet is not as complicated as the Halting Set? In terms of Turing degrees the problem thus is to construct a semi-decidable set A such that $\emptyset <_{\mathrm{T}} A <_{\mathrm{T}} \emptyset'$, or to rule out the existence of any such set, see [31,47,52] for background on Turing degrees in general and semi-decidable degrees in particular. Post's Problem resisted all attempts at resolution until Friedberg and Muchnik independently and almost simultaneously discovered a way to construct a set of intermediate complexity, see [18,41]. The construction is based on the idea of a so-called priority argument and is significantly more complicated than any construction of semi-decidable sets previously known [52]. Indeed, priority arguments have since become the hallmark of computability theory and have even engendered some criticism as being so very technical that, occasionally, the proofs seem to attract more attention than the theorems being established, see [65]. Be that as it may, it is striking how much more artificial and ad hoc intermediate sets are, as compared to natural examples such as the theory of the reals (decidable) or of Diophantine equations (equivalent to the Halting Problem). No natural examples of intermediate semi-decidable sets are known to date.

Nonetheless, given an intermediate set A one can construct a one-dimensional CA whose Reachability Problem has the same degree as A. This suggests a degree-based classification: given any computably enumerable degree $\mathbf{d}$, define the class $\mathbb{C}_{\mathbf{d}}$ to consist of all CA whose Reachability Problem has degree exactly $\mathbf{d}$, see [57,59]. The degree classification is non-trivial in the sense that every class is non-empty. Note that the first three Culik–Yu classes are all contained in $\mathbb{C}_{\mathbf{0}}$ whereas $\mathbb{C}_{\mathbf{0}'}$ comprises all computationally universal CA. Unsurprisingly, it is again undecidable whether a CA belongs to any particular class. At the bottom end of the hierarchy it is Σ_3-complete to determine membership in $\mathbb{C}_{\mathbf{0}}$; at the top end it is Σ_4-complete to determine membership in $\mathbb{C}_{\mathbf{0}'}$. Thus, it is easier to determine decidability than universality. In general, deciding membership in $\mathbb{C}_{\mathbf{d}}$ is Σ_3^d-complete for any semi-decidable degree $\mathbf{d}$. Similar results hold for the analogous cumulative classes $\mathbb{C}_{\leq \mathbf{d}} = \bigcup_{\mathbf{e} \leq \mathbf{d}} \mathbb{C}_{\mathbf{e}}$.

Unlike the Culik–Yu classification, the structure of the degree classification between $\mathbb{C}_{\mathbf{0}}$ and $\mathbb{C}_{\mathbf{0}'}$ is exceedingly complicated. For example, the proof of the Friedberg–Muchnik theorem shows that there are incomparable semi-decidable degrees $\mathbf{d}_1$ and $\mathbf{d}_2$. Hence there is are CA whose orbits are undecidable but not as complicated as the Halting Problem. Indeed, complete knowledge of the orbits of one of the two CA will not help in deciding membership in the orbits of the other. Another surprising result in the theory of computably enumerable degrees is Sack's Density Theorem, see [52]: between any two computably enumerable degrees $\mathbf{d}_1 < \mathbf{d}_2$ there lies a third: $\mathbf{d}_1 < \mathbf{d} < \mathbf{d}_2$. Thus, between any two CA of strictly increasing complexity there is an infinite and dense hierarchy of other CA. The computably enumerable degrees form a semi-lattice, so it is natural to try to understand the complexity of the structure by analyzing its first order theory. It is well-known that the Σ_1-theory of this semi-lattice is decidable. However, the reason for this decidability result lies in the fact that any countable partial order can be embedded into the semi-lattice so that the relative computational strength of cellular automata is indeed arbitrarily complicated. On the other hand, the full theory of the semi-lattice of semi-decidable degrees is known to be highly undecidable, see [22]; its degree is $\emptyset^{(\omega)}$. One might hope that restriction to reversible CA would simplify the

situation somewhat. Somewhat surprisingly it turns out that each class $\mathbb{C}_{\mathbf{d}}$ already contains an irreversible CA, see [60], so the same difficulties arise in the classification of reversible CA as in the classification of ordinary CA.

While reachability is arguably the most basic relation between configurations, similar difficulties also arise with confluence. As a matter of fact, one can construct a CA whose Reachability Problem has complexity some arbitrarily chosen computably enumerable degree $\mathbf{d}_1$ while the Confluence Problem for the same CA has degree $\mathbf{d}_2$, another arbitrarily chosen computably enumerable degree. Thus, a classification according to reachability is entirely independent of a confluence-based classification.

How do these results relate to PCE? Wolfram would not accept any of the intermediate classes of CA as a counterexample to PCE. The argument is that though intermediate degrees exist, their construction is critically linked to universal computation. While the universal computation is invisible when only the output of the system is observed, the associated computational process includes the whole computation and is thus universal. As a case in point, consider the standard Friedberg–Muchnik construction for an intermediate semi-decidable set A. The construction actually builds two semi-decidable sets A and B that are mutually incomparable with respect to Turing reducibility. Only A is output and B remains hidden. However, even ignoring all the intricate technical details of the whole construction, if we consider both A and B as output then the computation is indeed universal: the disjoint union $A \oplus B$ is Σ_1-complete, see [51]. It remains to be seen if similar arguments can be put forth in connection with priority-free constructions of intermediate degrees or if natural examples of intermediate sets can be found. At any rate, by considering only the reachability relation instead of a whole segment of the orbit we also achieve information-hiding, much as in the classical Friedberg–Muchnik construction.

Conclusion

Classification schemes of cellular automata based on the long-term evolution of pattern are typically undecidable, even if the property in question can be expressed in a fairly week system. While it is easy to construct examples of CA in particular classes it is usually very difficult to establish the position of a given CA in a particular classification. An excellent example for the difficulty of analyzing a given CA is Cook's proof of the universality of elementary CA number 110 whose local rule is given by $\rho(x, y, z) = (\overline{x} \wedge y \wedge z) \oplus y \oplus z$ where $\oplus$ denotes exclusive or, see [9], ► Cellular Automata, Universality of. The argument shows that cyclic tag systems, which are known to be complete, can be simulated by elementary CA 110 provided one allows an almost periodic background. Recent work by Turlough and Woods has shown that the whole simulation can be effected with only a polynomial slow-down, see [42,43]. This result suggests that the appropriate setting for classifications is the space of almost periodic configurations rather than finite ones.

In light of the successful analysis of elementary CA 110 it is tempting to ask about the classification of elementary CA 30. Figure 3 shows a segment of the orbit of a one-point seed configuration under rule 30. It is striking how chaotic and apparently random the image is. As a matter of fact, rule 30 has been used for many years as the default random number generator in the commercial computer algebra system Mathematica, see [70]. The underlying local map is simply $\rho(x, y, z) = x \oplus (y \vee z)$. Alas, there appear to be no structures in the evolution of configurations under rule 30 such as "moving particles" that might be exploited in a universality argument along the lines of rule 110. On the other hand, it is unclear how a decision procedure for reachability could be developed. This makes it tempting to conjecture that rule 30 in C_{ap} might be a member of one of the intermediate classes $\mathbb{C}_{\mathbf{d}}$, though at present there seems to be no way to either establish or refute this conjecture.

While undecidability results rule out the possibility of automatic classification mechanisms there is still ample room for the development of sufficient criteria for membership in certain classes, see [1,64,72]. For example, a proof of computational universality in a CA that has not been artificially constructed to simulate some other device often rests on the presence of "particles" or "gliders" that can be used to send "signals" between spatially separated locations. Moreover, one has to be able to process these signals much in the way of Boolean logic gates, to store state and so forth. A good example for complicate interactions between signals are the various solutions to the firing squad problem, albeit not in the context of simulating arbitrary computations; see Fig. 4, [36]. A more recent example is Cook's ingenious method of using natural gliders in elementary CA 110 to implement a cyclic tag system in C_{ap}, thereby establishing computational universality of rule 110, see [9]. Notable here is the fact that the automaton was fixed from the start and the the appropriate coding mechanisms had to be developed in a very constrained environment. This is in stark contrast to other hardness arguments where the CA is carefully constructed to display the desired behavior. Careful visual inspection of rule 110 orbits was a crucial component in Cook's proof, it is difficult to imagine that the result could have been established in a purely combinatorial or algebraic fashion. One

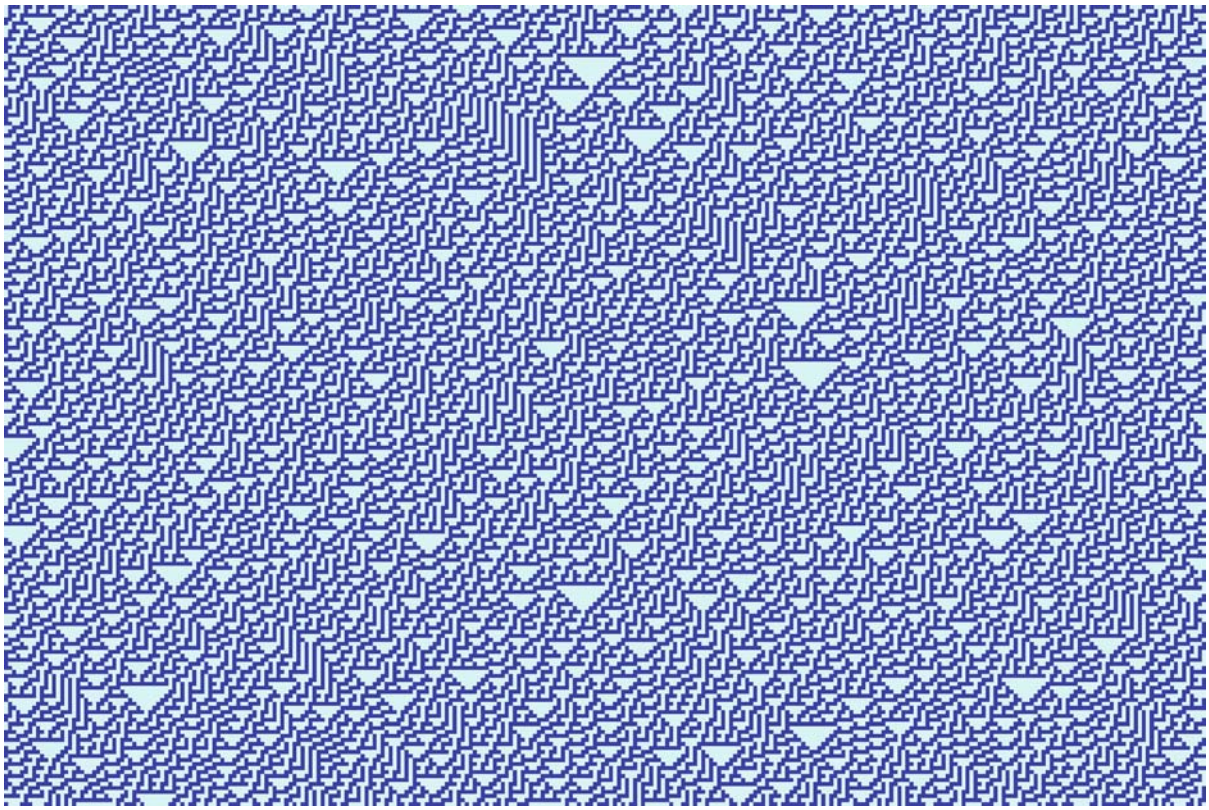

Cellular Automata, Classification of, Figure 3
A pseudo-random pattern generated by elementary CA 30

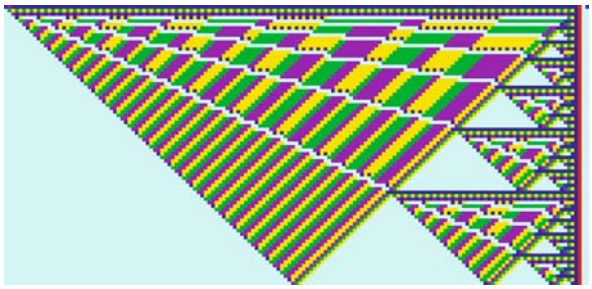

Cellular Automata, Classification of, Figure 4
Interacting signals in Mazoyer's optimal solution to the firing squad problem

can envision an interactive software system that helps to tackle some algorithmically unsolvable classification problems in special cases, much as Baumslag's Magnus project in group theory, see [5].

Bibliography

1. Adamatzky A (1994) Identification of Cellular Automata. Taylor & Francis, London
2. Amoroso S, Patt YN (1972) Decision procedures for surjectivity and injectivity of parallel maps for tesselation structures. J Comput Syst Sci 6:448–464
3. Baldwin JT (2002) Computation versus simulation. http://www.math.uic.edu/~jbaldwin/pub/cafom.ps. Accessed May 2007
4. Baldwin JT, Shelah S (2000) On the classifiability of cellular automata. Theor Comput Sci 230(1–2):117–129
5. Baumslag G () Magnus. http://caissny.org/. Accessed May 2007
6. Beal M-P, Perrin D (1997) Symbolic dynamics and finite automata. In: Rozenberg G, Salomaa A (eds) Handbook of Formal Languages. Springer, Berlin
7. Bennett CH (1973) Logical reversibility of computation. IBM J Res Dev 17:525–532
8. Börger E, Grädel E, Gurevich Y (2001) The Classical Decision Problem. Springer, Berlin
9. Cook M (2004) Universality in elementary cellular automata. Complex Syst 15(1):1–40
10. Culik K (1987) On invertible cellular automata. Complex Syst 1(6):1035–1044
11. Culik K, Sheng Y (1988) Undecidability of CA classification schemes. Complex Syst 2(2):177–190
12. Davis M (1956) A note on universal Turing machines. Shannon CE, McCarthy J (eds) Automata Studies. Annals of Mathematics Studies, vol 34. Princeton University Press, Princeton, pp 167–175
13. Davis M (1957) The definition of universal Turing machines. Proc Am Math Soc 8:1125–1126
14. Delorme M, Mazoyer J (1999) Cellular Automata: A Parallel

Model. Mathematics and Its Applications, vol 460. Kluwer, Dordrecht
15. Delvenne J-C, Kůrka P, Blondel V (2006) Decidability and universality in symbolic dynamical systems. Fundamenta Informaticae 74(4):463–490
16. Durand-Lose J (2001) Representing reversible cellular automata with reversible block automata. Disc Math Theor Comp Sci Proc AA:145–154
17. Elgaard J, Klarlund N, Møller A (1998) MONA 1.x: new techniques for WS1S and WS2S. In: Proc. 10th International Conference on Computer-Aided Verification, CAV '98. LNCS, vol 1427. Springer, Berlin, pp 516–520
18. Friedberg RM (1957) Two recursively enumerable sets of incomparable degrees of unsolvability. Proc Natl Acad Sci USA 43:236–238
19. Grädel E, Thomas W, Wilke T (eds) (2002) Automata, Logics, and Infinite Games. LNCS, vol 2500. Springer, Berlin
20. Gutowitz H (1996) Cellular automata and the sciences of complexity, part I. Complexity 1(5):16–22
21. Gutowitz H (1996) Cellular automata and the sciences of complexity, part II. Complexity 1(6)
22. Harrington L, Shelah S (1982) The undecidability of the recursively enumerable degrees. Bull Amer Math Soc 6:79–80
23. Head T (1989) Linear CA: injectivity from ambiguity. Complex Syst 3(4):343–348
24. Hedlund GA (1969) Endomorphisms and automorphisms of the shift dynamical system. Math Syst Theory 3:320–375
25. Hooper PK (1966) The undecidability of the Turing machine immortality problem. J Symb Log 31(2):219–234
26. Kari J (1990) Reversibility of 2D cellular automata is undecidable. Physica D 45:397–385
27. Kůrka P (1997) Languages, equicontinuity and attractors in cellular automata. Ergod Th Dyn Syst 17:417–433
28. Kůrka P (2003) Topological and Symbolic Dynamics. Cours Spécialisés, vol 11. Societe Mathematique de France, Paris
29. Langton CG (1990) Computation at the edge of chaos. Physica D 42:12–37
30. Lecerf Y (1963) Machine de Turing réversible. Insolubilité récursive en $n \in N$ de l'équation $u = \theta^n u$, où θ est un "isomorphisme de codes". C R Acad Sci Paris 257:2597–2600
31. Lerman M (1983) Degrees of Unsolvability. Perspectives in Mathematical Logic. Springer, Berlin
32. Li W, Packard N (1990) The structure of the elementary cellular automata rule space. Complex Syst 4(3):281–297
33. Li W, Packard N, Langton CG (1990) Transition phenomena in CA rule space. Physica D 45(1–3):77–94
34. Libkin L (2004) Elements of Finite Model Theory. Springer, Berlin
35. Manzini G, Margara L (1999) A complete and efficiently computable topological classification of D-dimensional linear cellular automata over Z_m. Theor Comput Sci 221(1–2):157–177
36. Mazoyer J (1987) A six state minimal time solution to the firing squad synchronization problem. Theor Comput Sci 50:183–238
37. Mitchell M, Crutchfield JP, Hraber PT (1994) Evolving cellular automata to perform computations: Mechanisms and impediments. Physica D (75):361–369
38. Morita K (1994) Reversible cellular automata. J Inf Process Soc Japan 35:315–321
39. Morita K (1995) Reversible simulation of one-dimensional irreversible cellular automata. Theor Comput Sci 148:157–163
40. Morita K, Harao M (1989) Computation universality of 1 dimensional reversible (injective) cellular automata. Trans Inst Electron, Inf Commun Eng E 72:758–762
41. Muchnik AA (1956) On the unsolvability of the problem of reducibility in the theory of algorithms. Dokl Acad Nauk SSSR 108:194–197
42. Neary R, Woods D (2006) On the time complexity of 2-tag systems and small universal turing machines. In: FOCS, pp 439–448. IEEE Comput Soc, Berkeley
43. Neary R, Woods D (2006) Small fast universal turing machines. Theor Comput Sci 362(1–3):171–195
44. Oliveira G, Oliveira P, Omar N (2001) Definition and application of a five-parameter characterization of one-dimensional cellular automata rule space. Artif Life 7(3):277–301
45. Ollinger N (2003) The intrinisic universality problem of one-dimensional cellular automata. In: Alt H, Habib M (eds) Proc. STACS. LNCS, vol 2607. Springer, Berlin, pp 632–641,
46. Packard NH (1988) Adaptation towards the Edge of Chaos. In: Dynamic Patterns in Complex Systems. World Scientific, Singapore, pp 29–301
47. Rogers H (1967) Theory of Recursive Functions and Effective Computability. McGraw Hill, New York
48. Rozenberg G, Salomaa A (1997) Handbook of Formal Languages. Springer, Berlin
49. Shepherdson JC (1965) Machine configuration and word problems of given degree of unsolvability. Z Math Logik Grundl Math 11:149–175
50. Shoenfield JR (1967) Mathematical Logic. Addison Wesley
51. Soare RI (1972) The Friedberg-Muchnik theorem re-examined. Can J Math 24:1070–1078
52. Soare RI (1987) Recursively Enumerable Sets and Degrees. Perspectives in Mathematical Logic. Springer, Berlin
53. Sutner K (1989) A note on Culik-Yu classes. Complex Syst 3(1):107–115
54. Sutner K (1990) Classifying circular cellular automata. Physica D 45(1–3):386–395
55. Sutner K (1991) De Bruijn graphs and linear cellular automata. Complex Syst 5(1):19–30
56. Sutner K (1995) The complexity of finite cellular automata. J Comput Syst Sci 50(1):87–97,
57. Sutner K (2002) Cellular automata and intermediate reachability problems. Fundamentae Informaticae 52(1–3):249–256
58. Sutner K (2003) Almost periodic configurations on linear cellular automata. Fundamentae Informaticae 58(3,4):223–240
59. Sutner K (2003) Cellular automata and intermediate degrees. Theor Comput Sci 296:365–375
60. Sutner K (2004) The complexity of reversible cellular automata. Theor Comput Sci 325(2):317–328
61. Taati S (2007) Cellular automata reversible over limit set. J Cell Autom 2(2):167–177
62. Toffoli T, Margolus N (1990) Invertible cellular automata: A review. Physica D 45:229–253
63. Turing AM (1936) On computable numbers, with an application to the Entscheidungsproblem. P Lond Math Soc 42: 230–65
64. Vorhees B (1996) Computational Analysis of One-Dimensional Cellular Automata. World Scientific, Singapore
65. Wang H (1993) Popular Lectures on Mathematical Logic. Dover Publications, Dover, New York
66. Weihrauch K (2000) Computable Analysis. EATCS Monographs. Springer, Berlin

67. Wolfram S (1984) Computation theory of cellular automata. Comm Math Phys 96(1):15–57
68. Wolfram S (1984) Universality and complexity in cellular automata. Physica D 10:1–35
69. Wolfram S (1985) Twenty problems in the theory of cellular automata. Phys Scr T9:170–183
70. Wolfram S (2002) The Mathematica Book. Cambridge University Press, Cambridge
71. Wolfram S (2002) A New Kind of Science. Wolfram Media, Champaign
72. Wuensche A (1999) Classifying cellular automata automatically. Complexity 4(3):47–66

Cellular Automata, Emergent Phenomena in

James E. Hanson
IBM T.J. Watson Research Center,
Yorktown Heights, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Synchronization
Domains in One Dimension
Particles in One Dimension
Emergent Phenomena in Two and Higher Dimensions
Future Directions
Bibliography

Glossary

Cellular automaton A spatially-extended dynamical system in which spatially-discrete cells take on discrete values, and evolve according to a spatially-localized discrete-time update rule.

Emergent phenomenon A phenomenon that arises as a result of a dynamical system's intrinsic dynamical behavior.

Domain A spatio-temporal region of a cellular automation that conforms to a specific pattern.

Particle A spatially-localized region of a cellular automaton that exists as a boundary or defect in a domain, and persists for a significant amount of time.

Definition of the Subject

In a dynamical system, an "emergent" phenomenon is one that arises out of the system's own dynamical behavior, as opposed to being introduced from outside. Emergent phenomena are ubiquitous in the natural world; as just one example, consider a shallow body of water with a sandy bottom. It often happens that small ridges form in the sand. These ridges emerge spontaneously, have a characteristic size and shape, and move across the bottom in a characteristic way – all due to the interaction of the sand and the water.

In cellular automata (CA), the system's state consists of an N-dimensional array of discrete cells that take on discrete values and the dynamics is given by a discrete time update rule (see below). The "phenomena" that emerge in CA therefore necessarily consist of spatio-temporal patterns and/or statistical regularities in the cell values. Therefore, the study of emergent phenomena is CA is the study of the spatio-temporal patterns and statistical regularities that arise spontaneously in cellular automata.

Introduction

The study of emergent phenomena in cellular automata dates back at least to the beginnings of the modern era of CA investigation inaugurated by Stephen Wolfram and collaborators. Indeed, it was a central theme of the landmark paper that introduced the four "Wolfram classes" [15] shown in Fig. 1. Ever since, emergent phenomena have been the driving force behind a great deal of CA research.

To be genuinely emergent, a phenomenon must arise out of configurations in which it is not present; and furthermore, to be of any significance, it must do so with nonvanishing likehood, and persist for a measurable amount of time. Thus the proper study of emergent phenomena in CA excludes from consideration a broad subcategory of systems in which the initial condition and update rule are chosen a priori to exhibit some particular structural feature (lattice gases are a representative example). The fact that such systems are CA is an implementation detail; the CA is merely a substrate or means for the simulation of higher-order structures. Note also that the essential issue is not whether the phenomena were intentionally designed into the CA rule; it is whether they arise naturally with any degree of frequency from configurations in which they are not present.

Notation and Terminology

A **cellular automaton** (CA) consists of a discrete N-dimensional array of sites or **cells** and a discrete-time **local update rule** ϕ applied to all cells in parallel.

The location of a cell is given by the N integer-valued coordinates $\{i, j, k, \ldots\}$. Cells take on values in a discrete set or **alphabet**, conventionally written $0, 1, \ldots, k-1$,

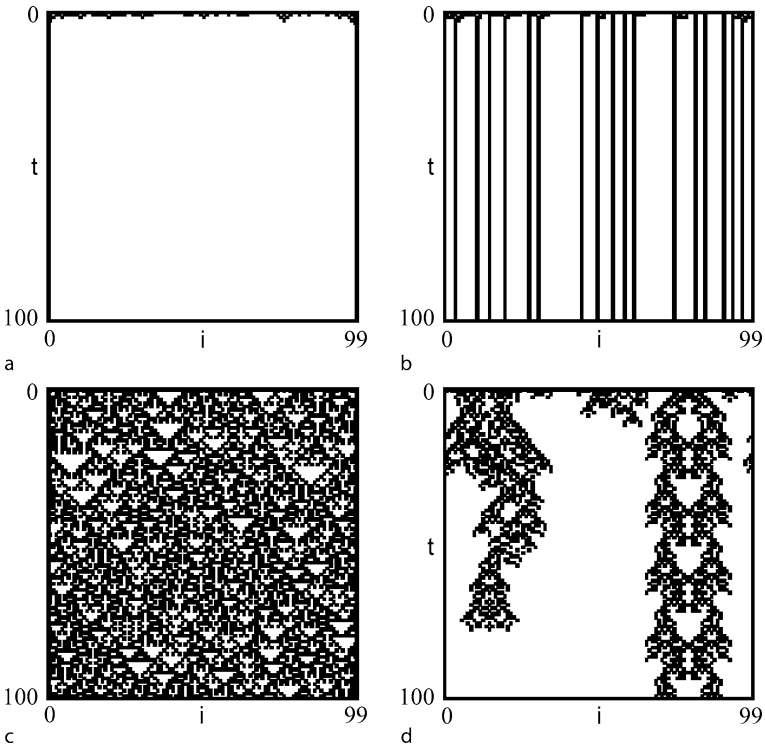

Cellular Automata, Emergent Phenomena in, Figure 1
Examples of Wolframs four qualitative classes. a Class 1: Spatio-temporally uniform configuration of ECA 32. b Class 2: Separated simple or periodic structures of ECA 44. c Class 3: Chaotic space-time pattern of ECA 90. d Class 4: Complex localized structures of Binary radius-2 CA 1771476584. In all cases the initial condition is random. In this and subsequent figures, cells with value 0 are shown as *white squares*, cells with value 1 are *black*

with k the **alphabet size**. An assignment of values to cells is called the **configuration** of those cells. The value 0 is sometimes treated as a special "quiescent" value, particularly in rules that obey the **quiescence condition** $\phi(\ldots 0 \ldots) = 0$.

The local update rule determines the value of a cell at time $t+1$ as a function of the values at time t of the cells around it. Typical neighborhoods are symmetrical, centered on the cell to be updated, and are parametrized by the **radius** r, which is the greatest distance from the center cell to any cell in the neighborhood. An assignment of values to the cells in a neighborhood is called a **parent neighborhood**, denoted by η, and the value $\phi(\eta)$ to which that parent neighborhood is mapped under the local update rule is its **child value**. The set of ordered pairs $\{\eta, \phi(\eta)\}$ is the **rule table**. The **speed of light** of a CA is the maximal rate at which information about a cell's value may travel; in general it is given by the radius r.

In two dimensions there are two common alternatives for the neighborhood's shape: the **von Neumann** neighborhood includes the center cell and its four neighbors up, down, left, and right; and the **Moore** neighborhood, which also includes the four cells diagonally adjacent to the center cell.

The so-called **elementary** cellular automata (ECA) are one-dimensional CA with $k = 2$, $r = 1$; a cell is denoted σ^i, takes on values in $\{0, 1\}$ and evolves over time according to the rule $\sigma^i_{t+1} = \phi(\sigma^{i-1}_t, \sigma^i_t, \sigma^{i+1}_t)$. A neighborhood consists of three consecutive cells, so there are 8 distinct parent neighborhoods and 256 different rule tables. It is convenient to refer to an elementary CA by its **rule number**, which is determined as follows. The different parent neighborhoods η are regarded as numbers in base k and are arranged in decreasing numerical order, from left to right. Immediately beneath each parent neighborhood its child value $\phi(\eta)$ is written. The rule number is obtained by regarding the sequence of child symbols as another number, again in base k. This numbering scheme may be used for one-dimensional CA with any k and r, and may be extended to higher-dimensional CA by the adoption of a convention for assigning numerical values to parent neighborhoods.

Different formulations of the local update rule are possible for CA in which symmetry or other constraints are

present. For example, one important subclass of CA rules are the **totalistic** rules, in which the child value depends only on the sum of the values in the parent neighborhood, not on their positions. Totalistic rules may also be assigned a rule number, by writing down the different possible sums of cell values in the parent neighborhood in order, writing the child cell beneath each such sum, and interpreting the sequence of child cells as a number.

In describing patterns in one-dimensional configurations, it is convenient to adopt a simplified form of regular expression notation, as follows:

- symbols $0, 1, \ldots, k-1$ denote literal cell values
- the symbol Σ denotes a "wild card" that may take on any value in the alphabet
- the expression x^* denotes any number of repetitions of the pattern x
- $[\ldots]$ denotes grouping
- concatenation denotes spatial adjacency.

For example, 0^* represents any number of consecutive 0s, while $[10]^*1$ is any configuration consisting of some number of repetitions of the pattern 10 followed by a 1: e. g., 101, 10101, 1010101, and so forth.

Synchronization

Possibly the simplest type of emergent phenomenon in CA is **synchronization**, which is the growth of spatial regions in which all cells have the same value. A synchronized region remains synchronized over time (except possibly at its borders) and it may either temporally invariant (i. e., the cell values to not change in time) or periodic (the cells all cycle together through the same temporal sequence of values). The temporal periodicity in the latter case is not greater than the alphabet size k.

About the trivial case in which the CA rule maps all neighborhoods to the same value (e. g., ECA 0 or ECA 255), there is little to be said. However, other cases exist in which the synchronized regions emerge only gradually. Characteristic examples in one dimension are shown in Fig. 1a and c. It is evident from these examples that any initial condition can be roughly, but usefully, described in terms of four patterns: (a) pattern 0^*, which represents the synchronized regions; (b,c) boundary regions $0^*\Sigma^*$ and Σ^*0^*; and (d) Σ^* for the interior of the non-synchronized regions. The behavior of the boundary regions determines whether the synchronized regions grow or shrink. For example, in ECA 32 (Fig. 1a), the parent neighborhoods in the boundary region are $\eta = \{0\Sigma\Sigma, \Sigma\Sigma 0\}$, all of which have child value 0; this means that the synchronized region grows as fast as is possible. Also not that since

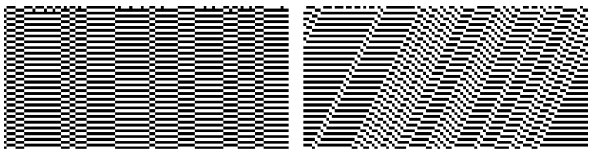

Cellular Automata, Emergent Phenomena in, Figure 2
Synchronization and phase defects: a ECA 55. b ECA 17

the only parent neighborhood that is *not* mapped to 0 is $\eta = 101$. the time taken for a given configuration in ECA 32 to reach a globally synchronized state is governed by the length of the longest region of pattern $[10]^*1$.

In general, the growth (or shrinkage) of synchronized regions is determined by the aggregate behavior of the neighborhoods that occur its boundaries; if they recede from each other, the region will grow. The boundaries need not move at the speed of light; the left and right boundaries need not move at the same speed; and their motion need not be perfectly uniform over time.

Figure 2a shows ECA 55, in which synchronized regions with temporal period $p = 2$ emerge from random initial conditions. Note, however, that multiple distinct synchronized regions persist indefinitely. This is an example of a **temporal phase defect**, which is a boundary between spatio-temporal regions that have the same overall pattern, but one of which is *ahead* of the other in time.

In general, phase defects need not be stationary: an example is shown in Fig. 2b. Also note that for CA with $k > 2$ it is possible for several different synchronized patterns to emerge and coexist. For example, consider a CA with $k = 3$ in which the pattern 0^* is temporally invariant, while 1^* and 2^* are mapped into each other to form a period-2 cycle.

Domains in One Dimension

Synchronization is a special case of a more general emergent phenomenon, the **domain**. A domain is spatial region that conforms to some specific pattern which persists over time. As has been seen in the case of synchronization, the emergence of a domain is governed by the behavior of its boundaries.

An important subclass of domain is the **regular domain**, in which the spatial pattern may be expressed in terms of a regular language (or equivalently, a finite state machine) [11]. As defined in [9], a regular domain has two properties: all spatial sequences of cells in the domain are in a given regular language; and (2) the set of all sequences in that regular language is itself temporally invariant or periodic. Regular domains are a powerful tool for identify-

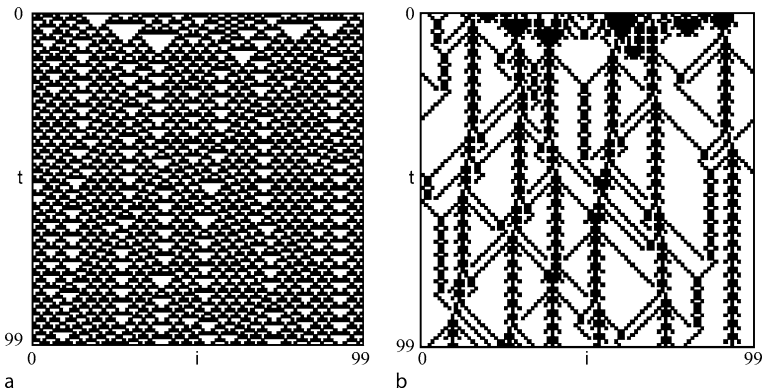

Cellular Automata, Emergent Phenomena in, Figure 3
Raw and domain-filtered space-time diagrams of ECA 54

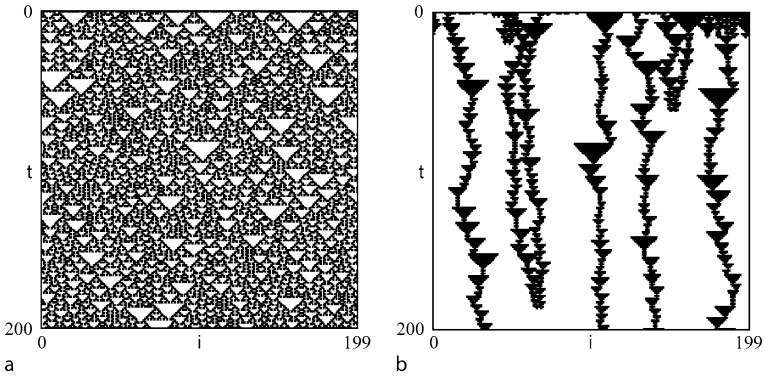

Cellular Automata, Emergent Phenomena in, Figure 4
Raw and domain-filtered ECA 18

ing and analyzing emergent phenomena in CA of one dimension. Generalization to two or more dimensions has proven challenging, though [12] made a significant step in that direction.

In studying domains in CA, it is useful to pass the space-time data through a **domain filter** to help visualize them. A domain filter, which may be constructed for any regular domain, maps every cell that is in the domain to a chosen value (0, say) and maps all cells not in the domain to other values in a prescribed way. Multi-domain filters may be constructed in a similar fashion, to map cells in any of a set of distinct domains $\Lambda_1, \Lambda_2, \ldots$ onto distinct values $\sigma_1, \sigma_2, \ldots$. See [3] for details.

An illustrative example is ECA 54, shown in Fig. 3. On the left is the unfiltered data; and on the right, the same dat after passing through the domain filter for ECA 54's primary domain. The domain has temporal period $p = 2$ and alternates between patterns [0001] and [110] The two patterns line up to form the interlocking white and black "T" shapes visible in the unfiltered data.

As the filtered plot clearly shows, the cells *not* in the domain have patterns of their own; this will be discussed in the next section. For now, it is sufficient to note that, in addition to the *temporal* phase defects seen in the emergence of temporally periodic synchronized regions, domains with nontrivial spatial structure may also show *spatial* phase defects, in which the pattern, in effect, skips or slips by a few cells.

The spatial regions that make up a domain may themselves contain disorder; such domains are called **chaotic**. ECA 90 is the archetypical example of this see Fig. 1c. From a random initial condition, ECA 90 quickly evolves so that entire configuration is in the domain $[0\Sigma]^*$. ECA 18 see Fig. 4a, attempts to do the same, except that the global synchronization is frustrated by long-lived spatial phase defects. This is clearly visible in the filtered space-time diagram shown in Fig. 4b. In this case the boundaries of the domain are inherently ambiguous: the pattern $[0\Sigma]^*[00]^*[\Sigma 0]^*$ contains exactly one spatial phase defect, but it may be regarded as lying anywhere in the cen-

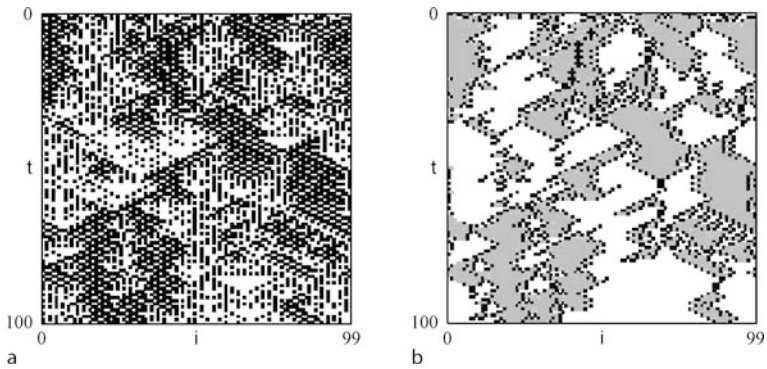

Cellular Automata, Emergent Phenomena in, Figure 5
Multiple coexisting chaotic domains

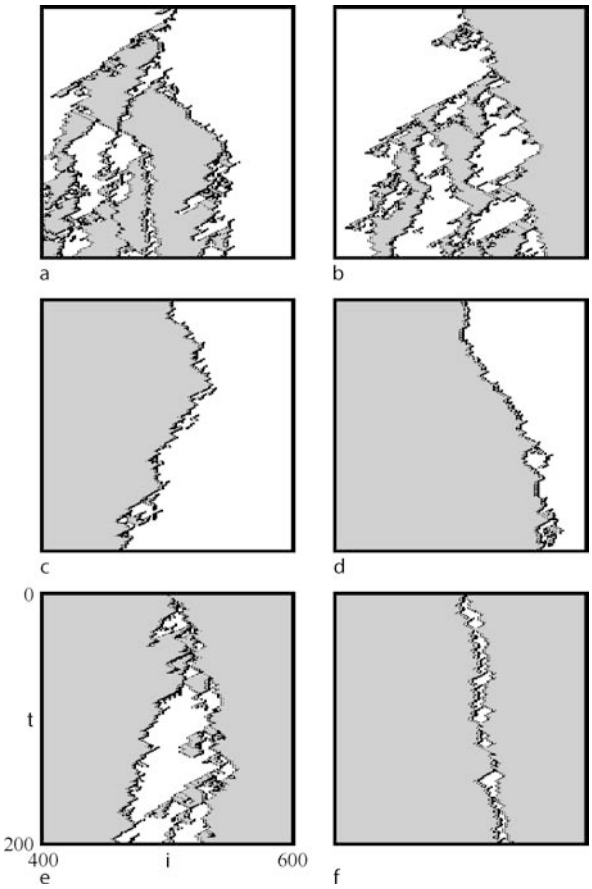

Cellular Automata, Emergent Phenomena in, Figure 6
Domain interfaces in the CA of Fig. 5

tral [00]* region. The filter used maps all cells in regions that contain a spatial phase defect to 1s.

A single CA may support the emergence of multiple different domain patterns. In many cases one domain dominates and will eventually take over. But this is not always true. An interesting case in which two domains, both chaotic, compete on roughly equal status, is binary radius-2 rule 2614700074, shown in Fig. 5. The two domains have patterns $\Lambda_0 = [0\Sigma]^*$ and $\Lambda_1 = [110\Sigma]^*$, respectively. In the filtered plot, cells in Λ_0 are shown in white, cells in Λ_1 are gray, and all other cells are black. It appears that by about $t = 200$ Λ_0 appears to be winning, but in fact, by about $t = 700$, the entire configuration was in Λ_1, where it remained indefinitely. Depending on the initial condition, one or the other domain was always found to eventually take over with Λ_0 winning about 80% of the time.

The coexistence of multiple domains, each with its own spatial structure, gives rise to a large number of possible interfaces. in general, the number of distinct interface types is governed by the complexity of the pattern in each domain; for 2614700074 it turns out that there are 8 distinct possibilities. Six of these show qualitatively distinct behavior, and are plotted (in filtered form only) in Fig. 6. Note that of the six interfaces, two show a quickly growing region in which defects continually multiply, three of them appear to remain spatially localized, and one (at bottom left) is ambiguous.

Particles in One Dimension

An immediate consequence of the emergence of domains is the simultaneous emergence of boundaries between them. These boundaries may be phase defects, as mentioned in Sect. "Synchronization", but they may also take the form of **particles**. A particle is a small region of cells that separates two domains, persists for a relatively long period of time and remains spatially localized. Particles may be stationary or may move; they may themselves exhibit a pattern that is temporally invariant, periodic, or even disordered.

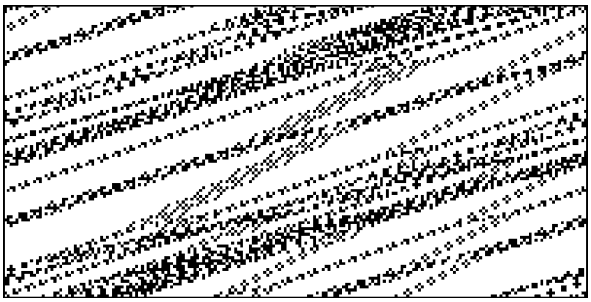

Cellular Automata, Emergent Phenomena in, Figure 7
Examples of solitons in the one-dimensional Filtering Rule

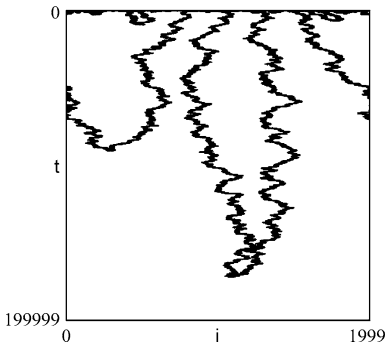

Cellular Automata, Emergent Phenomena in, Figure 8
Long-term behavior of ECA 18

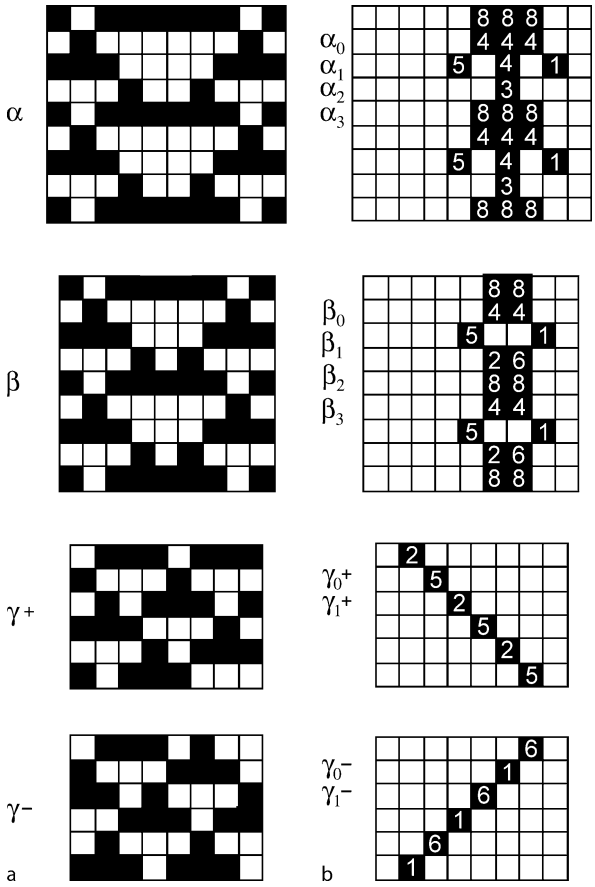

Cellular Automata, Emergent Phenomena in, Figure 9
Fundamental particles in ECA54

Solitons

An interesting type of particle emerges in the so-called **soliton** CA, shown in Fig. 7. These CA rules received their name in analogy with the solitons of fluid dynamics, which are solitary traveling waves with the interesting property that two solitons may collide, interact, and pass safely through each other, ultimately recovering their original form as if no collision had taken place. In soliton CA, something similar occurs.

In the simplest case, $k = 2$, the quiescence condition holds with the usual quiescent symbol 0. The solitons or particles embedded in a large lattice of 0s are finite sequences of 1s and 0s that are both temporally periodic (up to a spatial shift) and can collide and pass through each other without being destroyed. A particle consists of a finite sequence of basic strings of length $r + 1$ (where r is the CA radius). The leftmost cell of a particle is always a nonquiescent cell. A particle is bounded on the right by a sequence of $r + 1$ quiescent cells. Under the action of the CA rule, a particle may move to the left or right, may grow or shrink, but ultimately will come back to its original configuration after a finite time p – though possibly shifted by some number of cells. The ratio of the shift and temporal period p determines the particle's velocity V defined in the obvious way: $V =$ (spatial shift)/(temporal period).

A particle may even temporarily split into two or more smaller particles, so long as eventually they rejoin to form the original configuration. And, as the name implies, two particles with different velocities may collide and pass through each other without being destroyed.

Particles and Defects Defined by Domains

Given the wide variety of domains that arise in CA, the resultant variety of particles that they support is apparently limitless. However, two simple examples may suffice to illustrate these phenomena: ECA 18 and ECA 54, both of which were discussed in the previous section.

Particles in ECA 18

The spatial phase defects that occur in the domain of ECA 18 (see Fig. 4b) appear, on casual inspection, to be moving more or less at random. It turns out that to a very good approximation, an isolated defect performs a random walk

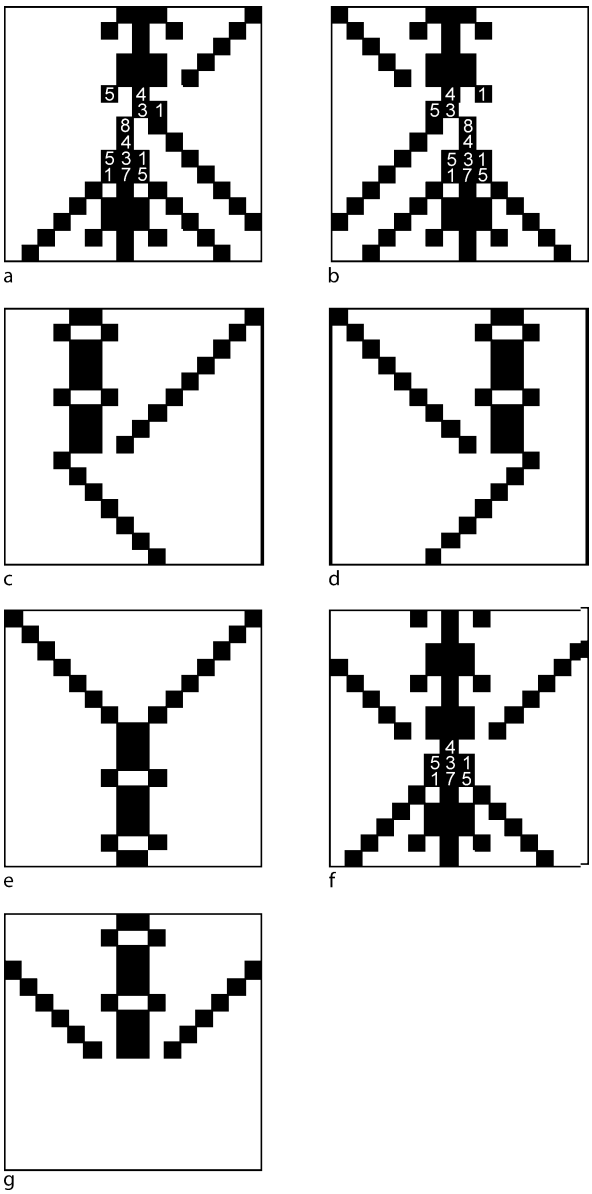

Cellular Automata, Emergent Phenomena in, Figure 10
Pairwise interactions between fundamental particles in ECA 54

on the lattice [4,7]. When two of them meet, they mutually annihilate. This behavior is purely deterministic, of course; it is caused entirely by the iterated action of the update rule on the initial condition. In effect, the disorder in the domains is causing disorder in the motion of their boundaries. For small systems, and *eventually* on all systems, finite-size effects cause departures from statistical randomness; but otherwise, except for a few highly atypical system-sizes, the defects' behavior is statistically indistinguishable from random motion. Figure 8 shows the

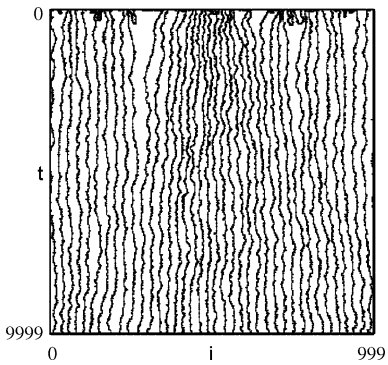

Cellular Automata, Emergent Phenomena in, Figure 11
Long-term behavior of ECA 54

long-term behavior of a random initial condition on a relatively large lattice.

Particles in ECA 54

ECA 54 represents an interesting case which can serve to illustrate many the emergent phenomena in one dimensional CA [1,10]. The primary domain gives rise to the so-called "fundamental particles" α, β and γ, shown in Fig. 9. The unfiltered space-time diagrams are shown on the left, and their filtered counterparts on the right. The interactions between the fundamental particles are shown in Fig. 10. In the filtered figures, the numbers inscribed in the black squares are the different outputs of the domain filter; each different sequence of numbers represents a different way in which the domain pattern has been violated.

The long-term behavior of the particles can be seen in Fig. 11. The βs decay relatively quickly, leaving only αs and γs – except for rare cases where a β is created by the interaction in Fig. 10e and persists for a short while, and rarer cases where some other pattern is momentarily present. (Note that the scale of the figure is so compressed that only the αs are visible.) It appears, and is borne out by numerical experiments, that the number of αs decays extremely slowly, and that the system settles into a state in which the αs are roughly equidistant, but move back and forth slightly in a disordered way. Unlike the case of ECA 18, the domains are not disordered, so the particle motion cannot be caused by disorder in the domain. Instead, it comes from the α–γ interactions.

Emergent Phenomena in Two and Higher Dimensions

As might be expected, the emergent phenomena in CA of more than one spatial dimension are at once richer

and less systematically studied. All of the phenomena that are observed in one dimension have their analogues in higher dimensions: domains and particle abound. In 2 or more dimensions, "particle" is no longer synonymous with "boundary"; one sees particles that are entirely surrounded by a domain, and spatially-extended boundaries that separate domains. Fundamentally new types of emergent phenomena appear as well.

Domains, Particles, and Interfaces

Many of the coherent structures found to exist in Conway's famous Game of Life can be observed to arise spontaneously from random initial conditions, so they properly fall into the category of emergent phenomena. In Fig. 12 a configuration of $100x100$ cells is shown at four successive times $t = 0, 50, 900, 1350$. From the random initial condition, a background pattern of 0s quickly emerges, against which there exist a rich variety of particles and disordered structures. By $t = 1350$ the configuration has settled to its final state, in which only a few particles remain, all of which are stationary and have temporal period $p = 1$ or $p = 2$. At intermediate times, various moving structures may be identified: see, for example, the "glider" at $t = 900$, about halfway between the center and the top. In moving about, these inevitably collide with each other or with the stationary particles, eventually leading to the final state.

Interestingly enough, a minor variation on the rule gives rise to the patterns shown in Fig. 13. Small regions of horizontal or vertical stripes emerge quickly. Boundaries between them settle down. By $t = 100$, a few non-striped areas persist, along with a few "dotted lines" that take the place of a stripe, and in which the "dots" oscillate. The non-striped areas eventually all disappear. The dotted lines persist indefinitely.

As these examples suggest, 2-dimensional CA support the emergence of synchronized regions, "domains", and particles in close analogy to 1-D CA. The striped regions in Fig. 13 are an example of a two-dimensional, temporally-invariant domain.

Fundamentally new features also appear in two and higher dimensions as well. The most obvious of these is the spatially-extended **interface** or boundary between two adjacent domains. Unlike the one-dimensional case, in

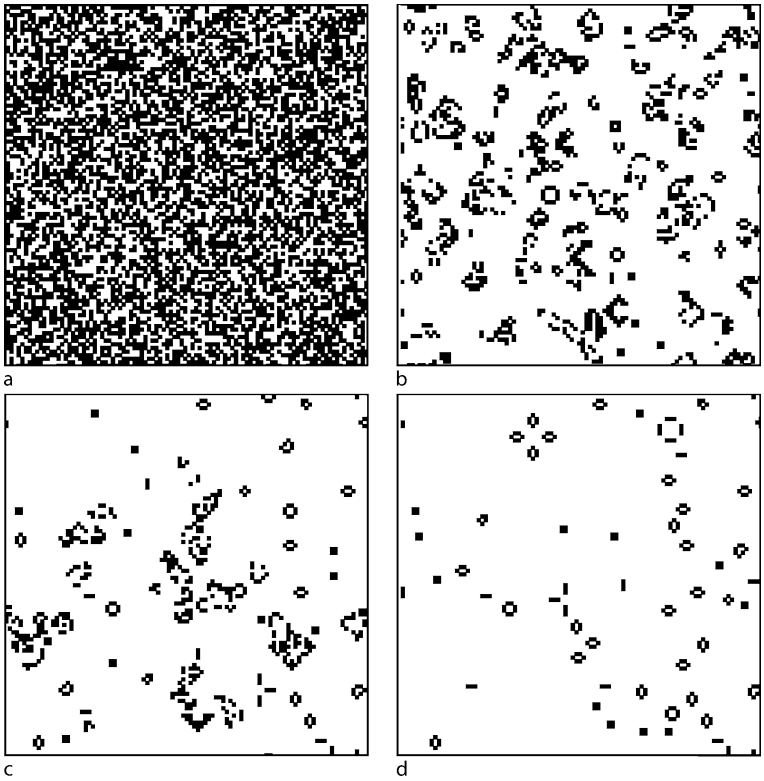

Cellular Automata, Emergent Phenomena in, Figure 12
Conway's Game of Life, starting from a random initial condition. a $t = 0$. b $t = 50$. c $t = 900$. d $t = 1350$

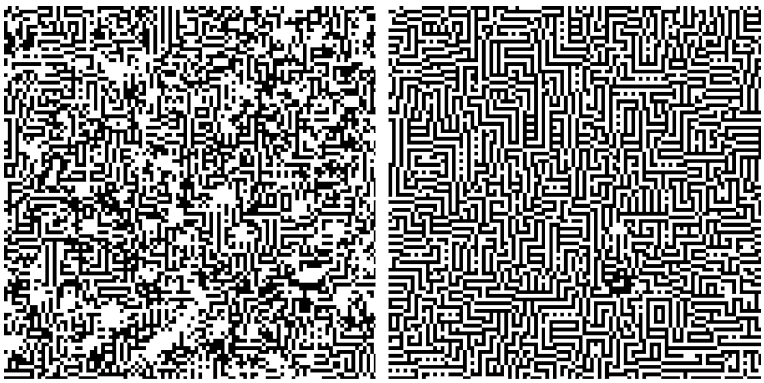

Cellular Automata, Emergent Phenomena in, Figure 13
Variant on Conway's Game of Life, starting from the same random initial condition as in Fig. 12. a $t = 10$. b $t = 100$

which particles and interfaces are more or less the same thing, interfaces in two dimensions are themselves one-dimensional. A characteristic example is seen in the *voting rule*, a 2-D binary CA with von Neumann neighborhood, in which the child cell is determined by the majority of the the local update rule maps a the child cell is equal to the value held by the majority of cells in the parent neighborhood, or if the vote is a tie, by a 0. Figure 14a shows a snapshot at $t = 50$ of the voting rule starting from a random initial condition. The system has organized itself into regions of two domain patterns. The pattern has stabilized by this time and does not change thereafter.

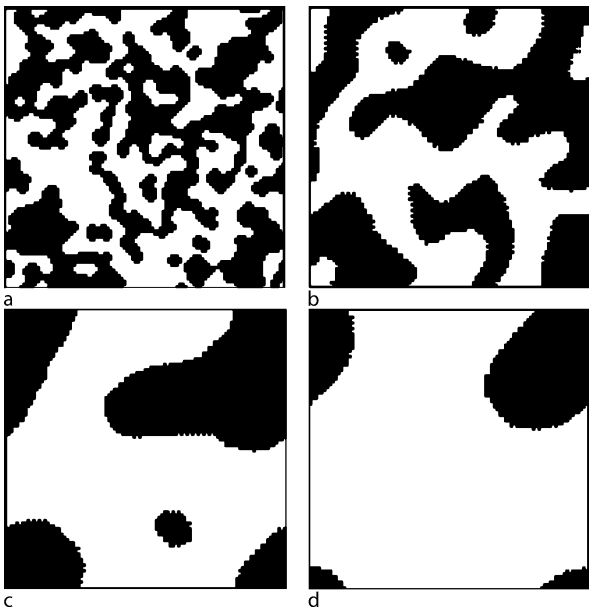

Cellular Automata, Emergent Phenomena in, Figure 14
Two variants of voter rule. a Voter rule at $t = 50$. This configuration is time-invariant. b Voter rule with random tie-breaking at $t = 50$. c Voter rule with random tie- breaking at $t = 250$. d Voter rule with random tie-breaking at $t = 750$

A stochastic variation on the voting rule uses a random variable to break tie votes, resulting it patterns such as Fig. 14b, c and d. Over time, the long boundaries gradually straighten, and small regions of one domain embedded in the other gradually shrink.

A number of extensive tours of patterns observed in selected 2-d CA may be found online; see, for example, [8,14].

Spiral Waves

Another important class of patterns in 2-D CA are expanding wavelike patterns, as shown in Fig. 15. These are typical of the class of rules called **cyclic CA** [5], and generally evolve to configurations of spirals (as shown). These patterns are not domains in the usual sense, because they have a geometric center. The shape of the spiral is closely related to the shape of the parent neighborhood. Starting from a random initial condition, eventually some number of centers form out from which the spiral waves emanate.

Quasiperiodicity

The final phenomenon to be mentioned here is an intriguing form of emergent phenomenon fundamentally different from what has been discussed above: the emergence of quasiperiodic oscillations in coarse statistical properties of the configuration (such as, percentage of 1s). [2,6] The evidence consists of **return maps**, in which the fraction m_t of 1s at time t, is plotted against the fraction m_{t+1} at time $t + 1$. A synchronized system would show a return map consisting of a single point: $m_t = m_{t+1}$. A periodic sys-

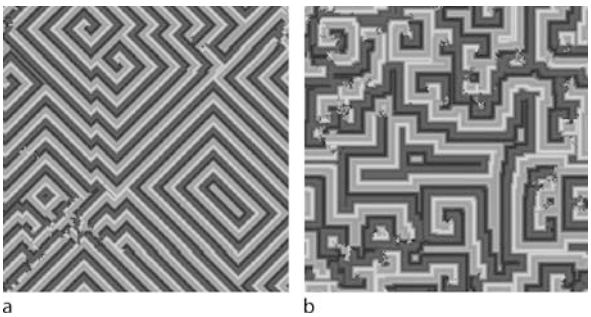

Cellular Automata, Emergent Phenomena in, Figure 15
Spiral waves. a Cyclic CA with $k = 16$, von Neumann neighborhood. b Cyclic CA with $k = 16$, Moore neighborhood

tem would show a sequence of points for the different values of m at the different temporal phases of the sequence, and would have $m_t = m_{t+p}$, where p is the period. The observed return plots, however, showed the characteristic shape of quasiperiodic behavior in nonlinear dynamical systems, which is a sequence of points that eventually map out a roughly continuous, closed curve in the plane. This quasiperiodic behavior was found to occur only in CA of dimension $N > 3$.

Future Directions

This short survey has only been able to hint at the vast wealth of emergent phenomena that arise in CA. Much work yet remains to be done, in classifying the different structures, identifying general laws governing their behavior, and determining the the causal mechanisms that lead them to arise.

For example, there are as yet no general techniques for determining whether a given domain is stable in a given CA; for characterizing the set of initial conditions that will eventually give rise to it; or for working out the particles that it supports. In CA or two or more dimensions, a large body of *descriptive* results are available, but these are more frequently anecdotal than systematic. A significant barrier to progress has been the lack of good mathematical techniques for identifying, describing, and classifying domains. One promising development in this area is an information-theoretic filtering technique that can operate on configurations of any dimension [13].

Bibliography

Primary Literature

1. Boccara N, Nasser J, Roger M (1991) Particlelike structures and their interactions in spatio-temporal patterns generated by one-dimensional deterministic cellular automaton rules. Phys Rev A 44:866
2. Chate H, Manneville P (1992) Collective behaviors in spatially extended systems with local interactions and synchronous updating. Profress Theor Phys 87:1
3. Crutchfield JP, Hanson JE (1993) Turbulent pattern bases for cellular automata. Physica D 69:279
4. Eloranta K, Nummelin E (1992) The kink of cellular automaton rule 18 performs a random walk. J Stat Phys 69:1131
5. Fisch R, Gravner J, Griffeath D (1991) Threshold-range scaling of excitable cellular automata. Stat Comput 1:23–39
6. Gallas J, Grassberger P, Hermann H, Ueberholz P (1992) Noisy collective behavior in deterministic cellular automata. Physica A 180:19
7. Grassberger P (1984) Chaos and diffusion in deterministic cellular automata. Phys D 10:52
8. Griffeath D (2008) The primordial soup kitchen. http://psoup.math.wisc.edu/kitchen.html
9. Hanson JE, Crutchfield JP (1992) The attractor-basin portrait of a cellular automaton. J Stat Phys 66:1415
10. Hanson JE, Crutchfield JP (1997) Computational mechanics of cellular automata: An example. Physica D 103:169
11. Hopcroft JE, Ullman JD (1979) Introduction to Automata Theory, Languages, and Computation. Addison-Wesley, Reading
12. Lindgren K, Moore C, Nordahl M (1998) Complexity of two-dimensional patterns. J Stat Phys 91:909
13. Shalizi C, Haslinger R, Rouquier J, Klinker K, Moore C (2006) Automatic filters for the detection of coherent structure in spatiotemporal systems. Phys Rev E 73:036104
14. Wojtowicz M (2008) Mirek's cellebration. http://www.mirekw.com/ca/
15. Wolfram S (1984) Universality and complexity in cellular automata. Physica D 10:1

Books and Reviews

Das R, Crutchfield JP, Mitchell M, Hanson JE (1995) Evolving Globally Synchronized Cellular Automata. In: Eshelman LJ (ed) Proceedings of the Sixth International Conference on Genetic Algorithms. Morgan Kaufmann, San Mateo

Gerhardt M, Schuster H, Tyson J (1990) A cellular automaton model of excitable medai including curvature and dispersion. Science 247:1563

Gutowitz HA (1991) Transients, Cycles, and Complexity in Cellular Automata. Phys Rev A 44:R7881

Henze C, Tyson J (1996) Cellular automaton model of three-dimensional excitable media. J Chem Soc Faraday Trans 92:2883

Hordijk W, Shalizi C, Crutchfield J (2001) Upper bound on the products of particle interactions in cellular automata. Physica D 154:240

Iooss G, Helleman RH, Stora R (ed) (1983) Chaotic Behavior of Deterministic Systems. North-Holland, Amsterdam

Ito H (1988) Intriguing Properties of Global Structure in Some Classes of Finite Cellular Automata. Physica 31D:318

Jen E (1986) Global Properties of Cellular Automata. J Stat Phys 43:219

Kaneko K (1986) Attractors, Basin Structures and Information Processing in Cellular Automata. In: Wolfram S (ed) Theory and Applications of Cellular Automata. World Scientific, Singapore, pp 367

Langton C (1990) Computation at the Edge of Chaos: Phase transitions and emergent computation. Physica D 42:12

Lindgren K (1987) Correlations and Random Information in Cellular Automata. Complex Syst 1:529
Lindgren K, Nordahl M (1988) Complexity Measures and Cellular Automata. Complex Syst 2:409
Lindgren K, Nordahl M (1990) Universal Computation in Simple One-Dimensional Cellular Automata. Complex Syst 4:299
Mitchell M (1998) Computation in Cellular Automata: A Selected Review. In: Schuster H, Gramms T (eds) Nonstandard Computation. Wiley, New York
Packard NH (1984) Complexity in Growing Patterns in Cellular Automata. In: Demongeot J, Goles E, Tchuente M (eds) Dynamical Behavior of Automata: Theory and Applications. Academic Press, New York
Packard NH (1985) Lattice Models for Solidification and Aggregation. Proceedings of the First International Symposium on Form, Tsukuba
Pivato M (2007) Defect Particle Kinematics in One-Dimensional Cellular Automata. Theor Comput Sci 377:205–228
Weimar J (1997) Cellular automata for reaction-diffusion systems. Parallel Comput 23:1699
Wolfram S (1984) Computation Theory of Cellular Automata. Comm Math Phys 96:15
Wolfram S (1986) Theory and Applications of Cellular Automata. World Scientific Publishers, Singapore
Wuensche A, Lesser MJ (1992) The Global Dynamics of Cellular Automata. Santa Fe Institute Studies in the Science of Complexity, Reference vol 1. Addison-Wesley, Redwood City

Cellular Automata and Groups

Tullio Ceccherini-Silberstein[1],
Michel Coornaert[2]
[1] Dipartimento di Ingegneria, Università del Sannio, Benevento, Italy
[2] Institut de Recherche Mathématique Avancée, Université Louis Pasteur et CNRS, Strasbourg, France

Article Outline

Glossary
Definition of the Subject
Introduction
Cellular Automata
Cellular Automata with a Finite Alphabet
Linear Cellular Automata
Group Rings and Kaplansky Conjectures
Future Directions
Bibliography

Glossary

Groups A *group* is a set G endowed with a binary operation $G \times G \ni (g, h) \mapsto gh \in G$, called the *multiplication*, that satisfies the following properties: (i) for all g, h and k in G, $(gh)k = g(hk)$ (associativity); (ii) there exists an element $1_G \in G$ (necessarily unique) such that, for all g in G, $1_G g = g 1_G = g$ (existence of the identity element); (iii) for each g in G, there exists an element $g^{-1} \in G$ (necessarily unique) such that $gg^{-1} = g^{-1}g = 1_G$ (existence of the inverses).

A group G is said to be *Abelian* (or *commutative*) if the operation is commutative, that is, for all $g, h \in G$ one has $gh = hg$.

A group F is called *free* if there is a subset $S \subset F$ such that any element g of F can be uniquely written as a *reduced word* on S, i. e. in the form $g = s_1^{\alpha_1} s_2^{\alpha_2} \cdots s_n^{\alpha_n}$, where $n \geq 0$, $s_i \in S$ and $\alpha_i \in \mathbb{Z} \setminus \{0\}$ for $1 \leq i \leq n$, and such that $s_i \neq s_{i+1}$ for $1 \leq i \leq n-1$. Such a set S is called a *free basis* for F. The cardinality of S is an invariant of the group F and it is called the *rank* of F.

A group G is *finitely generated* if there exists a finite subset $S \subset G$ such that every element $g \in G$ can be expressed as a product of elements of S and their inverses, that is, $g = s_1^{\epsilon_1} s_2^{\epsilon_2} \cdots s_n^{\epsilon_n}$, where $n \geq 0$ and $s_i \in S$, $\epsilon_i = \pm 1$ for $1 \leq i \leq n$. The minimal n for which such an expression exists is called the *word length* of g with respect to S and it is denoted by $\ell(g)$. The group G is a (discrete) metric space with the distance function $d\colon G \times G \to \mathbb{R}_+$ defined by setting $d(g, g') = \ell(g^{-1}g')$ for all $g, g' \in G$. The set S is called a *finite generating subset* for G and one says that S is *symmetric* provided that $s \in S$ implies $s^{-1} \in S$.

The *Cayley graph* of a finitely generated group G w.r. to a symmetric finite generating subset $S \subset G$ is the (undirected) graph $Cay(G, S)$ with vertex set G and where two elements $g, g' \in G$ are joined by an edge if and only if $g^{-1}g' \in S$.

A group G is *residually finite* if the intersection of all subgroups of G of finite index is trivial.

A group G is *amenable* if it admits a *right-invariant mean*, that is, a map $\mu\colon \mathcal{P}(G) \to [0, 1]$, where $\mathcal{P}(G)$ denotes the set of all subsets of G, satisfying the following conditions: (i) $\mu(G) = 1$ (*normalization*); (ii) $\mu(A \cup B) = \mu(A) + \mu(B)$ for all $A, B \in \mathcal{P}(G)$ such that $A \cap B = \varnothing$ (*finite additivity*); (iii) $\mu(Ag) = \mu(A)$ for all $g \in G$ and $A \in \mathcal{P}(G)$ (*right-invariance*).

Rings A *ring* is a set R equipped with two binary operations $R \times R \ni (a, b) \mapsto a + b \in R$ and $R \times R \ni (a, b) \mapsto ab \in R$, called the *addition* and the *multiplication*, respectively, such that the following properties are satisfied: (i) R, with the addition operation, is an Abelian group with identity element 0, called the *zero element*, (the inverse of an element $a \in R$ is denoted by $-a$); (ii) the multiplication is associative and admits an

identity element 1, called the *unit element*; (iii) multiplication is distributive with respect to addition, that is, $a(b+c) = ab + ac$ and $(b+c)a = ba + ca$ for all a, b and $c \in R$.
A ring R is *commutative* if $ab = ba$ for all $a, b \in R$.
A *field* is a commutative ring $\mathbb{K} \neq \{0\}$ where every non-zero element $a \in \mathbb{K}$ is invertible, that is there exists $a^{-1} \in \mathbb{K}$ such that $aa^{-1} = 1$.
In a ring R a non-trivial element a is called a *zero-divisor* if there exists a non-zero element $b \in R$ such that either $ab = 0$ or $ba = 0$.
A ring R is *directly finite* if whenever $ab = 1$ then necessarily $ba = 1$, for all $a, b \in R$. If the ring $M_d(R)$ of $d \times d$ matrices with coefficients in R is directly finite for all $d \geq 1$ one says that R is *stably finite*.
Let R be a ring and let G be a group. Denote by $R[G]$ the set of all formal sums $\sum_{g\in G} \alpha_g g$ where $\alpha_g \in R$ and $\alpha_g = 0$ except for finitely many elements $g \in G$. We define two binary operations on $R[G]$, namely the addition, by setting

$$\left(\sum_{g\in G} \alpha_g g\right) + \left(\sum_{h\in G} \beta_h h\right) = \sum_{g\in G} (\alpha_g + \beta_g) g\,,$$

and the multiplication, by setting

$$\left(\sum_{g\in G} \alpha_g g\right)\left(\sum_{h\in G} \beta_h h\right) = \sum_{g,h\in G} \alpha_g \beta_h gh$$
$$\equiv_{k=gh} \sum_{g,k\in G} \alpha_g \beta_{g^{-1}k} k.$$

Then, with these two operations, $R[G]$ becomes a ring; it is called the *group ring* of G with coefficients in R.

Cellular automata Let G be a group, called the *universe*, and let A be a set, called the *alphabet*. A *configuration* is a map $x\colon G \to A$. The set A^G of all configurations is equipped with the right action of G defined by $A^G \times G \ni (x, g) \mapsto x^g \in A^G$, where $x^g(g') = x(gg')$ for all $g' \in G$.
A *cellular automaton* over G with coefficients in A is a map $\tau\colon A^G \to A^G$ satisfying the following condition: there exists a finite subset $M \subset G$ and a map $\mu\colon A^M \to A$ such that $\tau(x)(g) = \mu(x^g|_M)$ for all $x \in A^G$, $g \in G$, where $x^g|_M$ denotes the restriction of x^g to M. Such a set M is called a *memory set* and μ is called a *local defining map* for τ.
If $A = V$ is a vector space over a field $\mathbb{K}$, then a cellular automaton $\tau\colon V^G \to V^G$, with memory set $M \subset G$ and local defining map $\mu\colon V^M \to V$, is said to be *linear* provided that μ is linear.
Two configurations $x, x' \in A^G$ are said to be *almost equal* if the set $\{g \in G; x(g) \neq x'(g)\}$ at which they differ is finite. A cellular automaton is called *pre-injective* if whenever $\tau(x) = \tau(x')$ for two almost equal configurations $x, x' \in A^G$ one necessarily has $x = x'$.
A *Garden of Eden configuration* is a configuration $x \in A^G \setminus \tau(A^G)$. Clearly, GOE configurations exist if and only if τ is not surjective.

Definition of the Subject

A cellular automaton is a self-mapping of the set of configurations of a group defined from local and invariant rules. Cellular automata were first only considered on the n-dimensional lattice group $\mathbb{Z}^n$ and for configurations taking values in a finite alphabet set but they may be formally defined on any group and for any alphabet. However, it is usually assumed that the alphabet set is endowed with some mathematical structure and that the local defining rules are related to this structure in some way. It turns out that general properties of cellular automata often reflect properties of the underlying group. As an example, the Garden of Eden theorem asserts that if the group is amenable and the alphabet is finite, then the surjectivity of a cellular automaton is equivalent to its pre-injectivity (a weak form of injectivity). There is also a linear version of the Garden of Eden theorem for linear cellular automata and finite-dimensional vector spaces as alphabets. It is an amazing fact that famous conjectures of Kaplansky about the structure of group rings can be reformulated in terms of linear cellular automata.

Introduction

The goal of this paper is to survey results related to the Garden of Eden theorem and the surjunctivity problem for cellular automata.

The notion of a cellular automaton goes back to John von Neumann [37] and Stan Ulam [34]. Although cellular automata were firstly considered only in theoretical computer science, nowadays they play a prominent role also in physics and biology, where they serve as models for several phenomena (► Cellular Automata Modeling of Physical Systems, ► Chaotic Behavior of Cellular Automata), and in mathematics. In particular, cellular automata are studied in ergodic theory (► Entropy in Ergodic Theory, ► Ergodic Theory: Basic Examples and Constructions, ► Ergodic Theory of Cellular Automata, ► Ergodicity and Mixing Properties) and in the theory of dynamical systems (► Topological Dynamics of Cellular Automata, ► Symbolic Dynamics), in functional and har-

monic analysis (▶ Spectral Theory of Dynamical Systems), and in group theory.

In the classical framework, the universe U is the lattice $\mathbb{Z}^2$ of integer points in Euclidean plane and the alphabet A is a finite set, typically $A = \{0, 1\}$. The set $A^U = \{x\colon U \to A\}$ is the configuration space, a map $x\colon U \to A$ is a configuration and a point $(n, m) \in U$ is called a cell. One is given a neighborhood M of the origin $(0, 0) \in U$, typically, for some $r > 0$, $M = \{(n, m) \in \mathbb{Z}^2\colon |n| + |m| \leq r\}$ (von Neumann r-ball) or $M = \{(n, m) \in \mathbb{Z}^2\colon |n|, |m| \leq r\}$ (Moore's r-ball) and a local map $\mu\colon A^M \to A$. One then "extends" μ to the whole universe obtaining a map $\tau\colon A^U \to A^U$, called a cellular automaton, by setting $\tau(x)(n, m) = \mu(x(n + s, m + t)_{(s,t)\in M})$. This way, the value $\tau(x)(n, m) \in A$ of the configuration x at the cell $(n, m) \in U$ only depends on the values $x(n+s, m+s)$ of x at its neighboring cells $(x + s, y + t) \equiv (x, y) + (s, t) \in (x, y) + M$, in other words, τ is $\mathbb{Z}^2$-equivariant. M is called a *memory set* for τ and μ a *local defining map*.

In 1963 E.F. Moore proved that if a cellular automaton $\tau\colon A^{\mathbb{Z}^2} \to A^{\mathbb{Z}^2}$ is surjective then it is also pre-injective, a weak form of injectivity. Shortly later, John Myhill proved the converse to Moore's theorem. The equivalence of surjectivity and pre-injectivity of cellular automata is referred to as the Garden of Eden theorem (briefly GOE theorem), this biblical terminology being motivated by the fact that it gives necessary and sufficient conditions for the existence of configurations x that are not in the image of τ, i. e. $x \in A^{\mathbb{Z}^2} \setminus \tau(A^{\mathbb{Z}^2})$, so that, thinking of $(\tau, A^{\mathbb{Z}^2})$ as a discrete dynamical system, with τ being the time, they can appear only as "initial configurations".

It was immediately realized that the GOE theorem was holding also in higher dimension, namely for cellular automata with universe $U = \mathbb{Z}^d$, the lattice of integer points in the d-dimensional space. Then, Machì and Mignosi [27] gave the definition of a cellular automaton over a finitely generated group and extended the GOE theorem to the class of groups G having sub-exponential growth, that is for which the growth function $\gamma_G(n)$, which counts the elements $g \in G$ at "distance" at most n from the unit element 1_G of G, has a growth weaker than the exponential, in formulæ, $\lim_{n\to\infty} \sqrt[n]{\gamma_G(n)} = 1$. Finally, in 1999 Ceccherini-Silberstein, Machì and Scarabotti [9] extended the GOE theorem to the class of amenable groups. It is interesting to note that the notion of an amenable group was also introduced by von Neumann [36]. This class of groups contains all finite groups, all Abelian groups, and in fact all solvable groups, all groups of sub-exponential growth and it is closed under the operation of taking subgroups, quotients, directed limits and extensions. In [27] two examples of cellular automata with universe the free group F_2 of rank two, the prototype of a non-amenable group, which are surjective but not pre-injective and, conversely, pre-injective but not surjective, thus providing an instance of the failure of the theorems of Moore and Myhill and so of the GOE theorem. In [9] it is shown that this examples can be extended to the class of groups, thus necessarily non-amenable, containing the free group F_2. We do not know whether the GOE theorem only holds for amenable groups or there are examples of groups which are non-amenable and have no free subgroups: by results of Olshanskii [30] and Adyan [1] it is know that such class is non-empty. In 1999 Misha Gromov [20], using a quite different terminology, reproved the GOE for cellular automata whose universes are infinite amenable graphs Γ with a dense pseudogroup of holonomies (in other words such Γs are rich in symmetries). In addition, he considered not only cellular automata from the full configuration space A^Γ into itself but also between subshifts $X, Y \subset A^\Gamma$. He used the notion of entropy of a subshift (a concept hidden in the papers [27] and [9].

In the mid of the fifties W. Gottschalk introduced the notion of surjunctivity of maps. A map $f\colon X \to Y$ is surjunctive if it is surjective or not injective. We say that a group G is surjunctive if all cellular automata $\tau\colon A^G \to A^G$ with finite alphabet are surjunctive. Lawton [18] proved that residually finite groups are surjunctive. From the GOE theorem for amenable groups [9] one immediately deduce that amenable groups are surjunctive as well. Finally Gromov [20] and, independently, Benjamin Weiss [38] proved that all sofic groups (the class of sofic groups contains all residually finite groups and all amenable groups) are surjunctive. It is not known whether or not all groups are surjunctive.

In the literature there is a notion of a linear cellular automaton. This means that the alphabet is not only a finite set but also bears the structure of an Abelian group and that the local defining map μ is a group homomorphism, that is, it preserves the group operation. These are also called *additive cellular automata* (▶ Additive Cellular Automata).

In [5], motivated by [20], we introduced another notion of linearity for cellular automata. Given a group G and a vector space V over a (not necessarily finite) field $\mathbb{K}$, the configuration space is V^G and a cellular automaton $\tau\colon V^G \to V^G$ is linear if the local defining map $\mu\colon V^B \to V$ is $\mathbb{K}$-linear. The set $\mathrm{LCA}(V, G)$ of all linear cellular automata with alphabet V and universe G naturally bears a structure of a ring.

The finiteness condition for a set A in the classical framework is now replaced by the finite dimensionality of V. Similarly, the notion of entropy for subshifts $X \subset A^G$

is now replaced by that of mean-dimension (a notion due to Gromov [20]). In [5] we proved the GOE theorem for linear cellular automata $\tau: V^G \to V^G$ with alphabet a finite dimensional vector space and with G an amenable group. Moreover, we proved a linear version of Gottschalk surjunctivity theorem for residually finite groups.

In the same paper we also establish a connection with the theory of group rings. Given a group G and a field $\mathbb{K}$, there is a one-to-one correspondence between the elements in the group ring $\mathbb{K}[G]$ and the cellular automata $\tau: \mathbb{K}^G \to \mathbb{K}^G$. This correspondence preserves the ring structures of $\mathbb{K}[G]$ and $\mathrm{LCA}(\mathbb{K}, G)$. This led to a reformulation of a long standing problem, raised by Irving Kaplansky [23], about the absence of zero-divisors in $\mathbb{K}[G]$ for G a torsion-free group, in terms of the pre-injectivity of all $\tau \in \mathrm{LCA}(\mathbb{K}, G)$.

In [6] we proved the linear version of the Gromov–Weiss surjunctivity theorem for sofic groups and established another application to the theory of group rings. We extended the correspondence above to a ring isomorphism between the ring $\mathrm{Mat}_d(\mathbb{K}[G])$ of $d \times d$ matrices with coefficients in the group ring $\mathbb{K}[G]$ and $\mathrm{LCA}(\mathbb{K}^d, G)$. This led to a reformulation of another famous problem, raised by Irving Kaplansky [24] about the structure of group rings. A group ring $\mathbb{K}[G]$ is stably finite if and only if, for all $d \geq 1$, all linear cellular automata $\tau: (\mathbb{K}^d)^G \to (\mathbb{K}^d)^G$ are surjunctive. As a byproduct we obtained another proof of the fact that group rings over sofic groups are stably finite, a result previously established by G. Elek and A. Szabó [11] using different methods.

The paper is organized as follows. In Sect. "Cellular Automata" we present the general definition of a cellular automaton for any alphabet and any group. This includes a few basic examples, namely Conway's Game of Life, the majority action and the discrete Laplacian. In the subsequent section we specify our attention to cellular automata with a finite alphabet. We present the notions of Cayley graphs (for finitely generated groups), of amenable groups, and of entropy for G-invariant subsets in the configuration space. This leads to a description of the setting and the statement of the Garden of Eden theorem for amenable groups. We also give detailed expositions of a few examples showing that the hypotheses of amenability cannot, in general, be removed from the assumption of this theorem. We also present the notion of surjunctivity, of sofic groups and state the surjunctivity theorem of Gromov and Weiss for sofic groups. In Sect. "Linear Cellular Automata" we introduce the notions of linear cellular automata and of mean dimension for G-invariant subspaces in V^G. We then discuss the linear analogue of the Garden of Eden theorem and, again, we provide explicit examples showing that the assumptions of the theorem (amenability of the group and finite dimensionality of the underlying vector space) cannot, in general, be removed. Finally we present the linear analogue of the surjunctivity theorem of Gromov and Weiss for linear cellular automata over sofic groups. In Sect. "Group Rings and Kaplansky Conjectures" we give the definition of a group ring and present a representation of linear cellular automata as matrices with coefficients in the group ring. This leads to the reformulation of the two long standing problems raised by Kaplansky about the structure of group rings.

Finally, in Sect. "Future Directions" we present a list of open problems with a description of more recent results related to the Garden of Eden theorem and to the surjunctivity problem.

Cellular Automata

The Configuration Space

Let G be a group, called the *universe*, and let A be a set, called the *alphabet* or the *set of states*. A *configuration* is a map $x: G \to A$. The set A^G of all configurations is equipped with the right action of G defined by $A^G \times G \ni (x, g) \mapsto x^g \in A^G$, where $x^g(g') = x(gg')$ for all $g' \in G$.

Cellular Automata

A *cellular automaton* over G with coefficients in A is a map $\tau: A^G \to A^G$ satisfying the following condition: there exists a finite subset $M \subset G$ and a map $\mu: A^M \to A$ such that

$$\tau(x)(g) = \mu(x^g|_M) \tag{1}$$

for all $x \in A^G$, $g \in G$, where $x^g|_M$ denotes the restriction of x^g to M. Such a set M is called a *memory set* and μ is called a *local defining map* for τ.

It follows directly from the definition that every cellular automaton $\tau: A^G \to A^G$ is *G-equivariant*, i. e., it satisfies

$$\tau(x^g) = \tau(x)^g \tag{2}$$

for all $g \in G$ and $x \in A^G$.

Note that if M is a memory set for τ, then any finite set $M' \subset G$ containing M is also a memory set for τ. The local defining map associated with such an M' is the map $\mu': A^{M'} \to A$ given by $\mu' = \mu \circ \pi$, where $\pi: A^{M'} \to A^M$ is the restriction map. However, there exists a unique memory set M_0 of minimal cardinality. This memory set M_0 is called the *minimal* memory set for τ.

We denote by $\mathrm{CA}(G, A)$ the set of all cellular automata over G with alphabet A.

Examples

Example 1 (Conway's Game of Life [3]) The most famous example of a cellular automaton is the *Game of Life* of John Horton Conway. The set of states is $A = \{0, 1\}$. State 0 corresponds to *absence of life* while state 1 indicates *life*. Therefore passing from 0 to 1 can be interpreted as *birth*, while passing from 1 to 0 corresponds to *death*.

The universe for Life is the group $G = \mathbb{Z}^2$, that is, the free Abelian group of rank 2. The minimal memory set is $M = \{-1, 0, 1\}^2 \subset \mathbb{Z}^2$. The set M is the *Moore neighborhood* of the origin in $\mathbb{Z}^2$. It consists of the origin $(0, 0)$ and its eight neighbors $\pm(1, 0), \pm(0, 1), \pm(1, 1), \pm(-1, 1)$. The corresponding local defining map $\mu\colon A^M \to A$ given by

$$\mu(y) = \begin{cases} 1 & \text{if} \begin{cases} \sum_{s\in S} y(s) = 3 \\ \text{or} \\ \sum_{s\in S} y(s) = 4 \text{ and } y((0,0)) = 1\,, \end{cases} \\ 0 & \text{otherwise}\,. \end{cases}$$

Example 2 (The majority action [21]) Let G be a group, M a finite subset of G, and $A = \{0, 1\}$. The automaton $\tau\colon A^G \to A^G$ with memory set M and local defining map $\mu\colon A^M \to A$ given by

$$\mu(y) = \begin{cases} 1 & \text{if } \sum_{m\in M} y(m) > \frac{|M|}{2}\,, \\ 0 & \text{if } \sum_{m\in M} y(m) < \frac{|M|}{2}\,, \\ y(1_G) & \text{if } \sum_{m\in M} y(m) = \frac{|M|}{2}\,, \end{cases}$$

for all $y \in A^M$, is the *majority action automaton* associated with G and M.

Example 3 Let G be a group and let A be any alphabet. Let $f\colon A \to A$ be a map and consider the map $\tau_f\colon A^G \to A^G$ defined by setting $\tau_f(x)(g) = f(x(g))$ for all $x \in A^G$ and $g \in G$. Then τ_f is a cellular automaton with memory set $M = \{1_G\}$ and the local defining map $\mu\colon A^M \to A$ given by $y \mapsto f(y(1_G))$ for all $y \in A^M$. When $f = \iota_A$ is the identity map on A then $\tau_{\iota_A} =: I$ is the identity map on A^G. On the other hand, given $c \in A$, if $f = f_c$ is the constant map given by $f_c(a) = c$ for all $a \in A$, then τ_{f_c} is the *constant* cellular automaton defined by $\tau_{f_c}(x) = x_c$ for all $x \in A^G$, where $x_c(g) = c$ for all $g \in G$.

Example 4 (The discrete Laplacian) Let G be a group, S a finite subset of G not containing 1_G, and $A = \mathbb{R}$, the field of *real numbers*. The (linear) map $\Delta_S\colon \mathbb{R}^G \to \mathbb{R}^G$ defined by

$$\Delta_S(x)(g) = x(g) - \frac{1}{|S|}\sum_{s\in S} x(gs)$$

is a cellular automaton over G with memory set $M = S \cup \{1_G\}$ and local defining map $\mu\colon \mathbb{R}^M \to \mathbb{R}$ given by $\mu(y) = y(1_G) - \frac{1}{|S|}\sum_{s\in S} y(s)$, for all $y \in \mathbb{R}^M$. It is called the *Laplacian* or *Laplace operator* on G relative to S.

Let $\tau_1, \tau_2 \in \mathrm{CA}(G, A)$ be two cellular automata (with memory sets M_1 and M_2, respectively). It is easy to see that their composition $\tau_1 \circ \tau_2$, defined by $[\tau_1 \circ \tau_2](x) = \tau_1(\tau_2(x))$ for all $x \in A^G$, is a cellular automaton (admitting $M = M_1M_2$ as a memory set). Since the identity map $I\colon A^G \to A^G$ is a cellular automaton, it follows that $\mathrm{CA}(G, A)$ is a monoid for the composition operation.

Cellular Automata with a Finite Alphabet

The Configuration Space as a Metric Space

Let G be a countable group, e. g., a finitely generated group (see Subsect. "Cayley Graphs") and let A be a finite alphabet with at least two elements.

The set A^G of all configurations can be equipped with a metric space structure as follows. Let $\emptyset = E_1 \subset E_2 \subset \cdots \subset E_n \subset E_{n+1} \subset \cdots$ be an increasing sequence of finite subsets of G such that $\cup_{n\geq 1} E_n = G$. Then, given any two configurations $x, x' \in A^G$, we set:

$$d(x, x') = 1/\sup\left\{n \in \mathbb{N} : x|_{E_n} = x'|_{E_n}\right\} \tag{3}$$

(we use the convention that $1/\infty = 0$). In this way, A^G becomes a compact totally disconnected space homeomorphic to the middle third Cantor set.

We then have Hedlund's topological characterization of cellular automata.

Theorem 1 (Hedlund) *Suppose that A is a finite set. A map $\tau\colon A^G \to A^G$ is a cellular automaton if and only if it is continuous and G-equivariant.*

Corollary 1 *Suppose that A is a finite set. Let $\tau\colon A^G \to A^G$ be a bijective cellular automaton. Then the inverse map $\tau^{-1}\colon A^G \to A^G$ is also a cellular automaton.*

The Garden of Eden Theorem of Moore and Myhill

Let G be any group and A be any alphabet. Let $F \subset G$ be a finite subset. A *pattern* with *support* F is a map $p\colon A^F \to A$.

Let now $\tau\colon A^G \to A^G$ be a cellular automaton. One says that τ is *surjective* if $\tau(A^G) = A^G$. One often thinks of τ as describing time evolution: if $x \in A^G$ is the configuration of the universe at time t, then $\tau(x)$ is the configuration of the universe at time $t + 1$. An *initial configuration* is a configuration at time $t = 0$. A configuration x which is not in the image of τ, namely such that

$x \in A^G \setminus \tau(A^G)$, is called a *Garden of Eden* (briefly *GOE*) *configuration*. This biblical terminology is motivated by the fact that GOE configurations may only appear as initial configurations. Analogously, a pattern p with support $F \subset G$ is called a *GOE pattern* if $p \neq \tau(x)|_F$ for all $x \in A^G$. Using topological methods it is easy to see that, when the alphabet is finite, the existence of GOE patterns for τ is equivalent to the existence of GOE configurations for τ, i. e., to the non-surjectivity of τ.

One says that τ is *injective* if, for $x, x' \in A^G$, one has $x = x'$ whenever $\tau(x) = \tau(x')$.

Two configurations $x, x' \in A^G$ are *almost equal*, and we write $x \sim_{a.e.} x'$, if they coincide outside a finite subset of G, namely $|\{g \in G : x(g) \neq x'(g)\}| < \infty$. Finally, using terminology introduced by Gromov, one says that τ is *pre-injective* if, for $x, x' \in A^G$ s.t. $x \sim_{a.e.} x'$, one has $x = x'$ whenever $\tau(x) = \tau(x')$.

Two patterns p, p' with the same support F are *mutually erasable* if they are distinct and whenever $x, x' \in A^G$ are two configurations which extend in the same way p and p' outside of F (i. e. $x|_F = p$, $x'|_F = p'$ and $x|_{G\setminus F} = x'|_{G\setminus F}$), then $\tau(x) = \tau(x')$. The non-existence of mutually erasable patterns is equivalent to the pre-injectivity of the cellular automaton. Finally note that injectivity implies pre-injectivity (but the converse is false, in general).

The following is the celebrated Garden of Eden theorem of Moore and Myhill.

Theorem 2 (Moore and Myhill) *Let $\tau \in CA(\mathbb{Z}^2, A)$ be a cellular automaton with coefficients in a finite set A. Then τ is surjective if and only if it is pre-injective.*

The necessary condition is due to Moore, the converse implication to Myhill.

As Conway's Game of Life is concerned, we have that this cellular automaton is clearly not pre-injective (the constant dead configuration and the configuration with only one live cell have the same image) and by the previous theorem it is not surjective either. We mention that the non-surjectivity of the Game of Life is not trivial: the smallest GOE pattern known up to now has as a support a rectangle 13×12 with 81 live cells.

Cayley Graphs

A group G is said to be *finitely generated* if there exists a finite subset $S \subset G$ such that every element $g \in G$ can be expressed as a product of elements of S and their inverses, that is, $g = s_1^{\epsilon_1} s_2^{\epsilon_2} \cdots s_n^{\epsilon_n}$, where $n \geq 0$ and $s_i \in S, \epsilon_i = \pm 1$ for $1 \leq i \leq n$. The minimal n for which such an expression exists is called the *word length* of g with respect to S and it is denoted by $\ell(g)$. The group G is a (discrete) metric space with the distance function $d\colon G \times G \to \mathbb{R}_+$ defined by setting $d(g, g') = \ell(g^{-1}g')$ for all $g, g' \in G$. The set S is called a *finite generating subset* for G and one says that S is *symmetric* provided that $s \in S$ implies $s^{-1} \in S$.

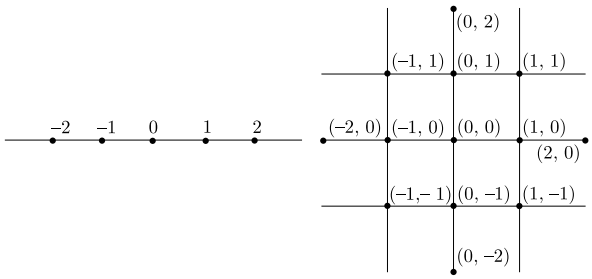

Cellular Automata and Groups, Figure 1
The ball $B(2, 1_G)$ in $\mathbb{Z}$ and in $\mathbb{Z}^2$, respectively

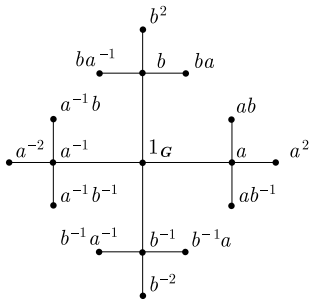

Cellular Automata and Groups, Figure 2
The ball $B(2, 1_G)$ in F_2

Suppose that G is finitely generated and let S be a symmetric finite generating subset of G. The *Cayley graph* of G w.r.t. S is the (undirected) graph $Cay(G, S)$ with vertex set G and two elements $g, g' \in G$ are joined by an edge if and only if $g^{-1}g' \in S$. The group G becomes a (discrete) metric space by introducing the distance $d\colon G \times G \to \mathbb{R}_+$ defined by $d(g, g') = \ell(g^{-1}g')$ for all $g, g' \in G$. Note that the distance d coincides with the graph distance on the Cayley graph $Cay(G, S)$. For $g \in G$ and $n \in \mathbb{N}$ we denote by $B(n, g) = \{g' \in G\colon d(g, g') \leq n\}$ the *ball* of *radius* n with *center* g.

Amenable Groups

The notion of an amenable group is also due to J. von Neumann [36]. Let G be a group and denote by $\mathcal{P}(G)$ the set of all subsets of G. The group G is said to be *amenable* if there exists a *right-invariant mean*, that is, a map $\mu\colon \mathcal{P}(G) \to [0, 1]$ satisfying the following conditions:

1. $\mu(G) = 1$ (*normalization*);

2. $\mu(A \cup B) = \mu(A) + \mu(B)$ for all $A, B \in \mathcal{P}(G)$ such that $A \cap B = \varnothing$ (*finite additivity*);
3. $\mu(Ag) = \mu(A)$ for all $g \in G$ and $A \in \mathcal{P}(G)$ (*right-invariance*).

We mention that if G is amenable, namely such a right-invariant mean exists, then also left-invariant and in fact even bi-invariant means do exist. The class of amenable groups includes, in particular, all finite groups, all Abelian groups (and, more generally, all solvable groups), and all finitely generated groups of subexponential growth. It is closed under the operations of taking subgroups, taking factors, taking extensions and taking direct limits.

It was observed by von Neumann himself [36] that the free group F_2 based on two generators is non-amenable. Therefore, all groups containing a subgroup isomorphic to the free group F_2 are non-amenable as well.

However, there are examples of non-amenable groups which do not contain subgroups isomorphic to F_2. The first examples are due to Olshanski [30]; later Adyan [1] showed that also the free Burnside groups $B(m, n) = \langle s_1, s_2, \ldots, s_m : w^n \rangle$ of rank $m \geq 2$ and odd *exponent* $n \geq 665$ are non-amenable.

It follows from a result of Følner [16] that a countable group G is amenable if and only if it admits a *Følner sequence*, i. e., a sequence $(F_n)_{n\in\mathbb{N}}$ of non-empty finite subsets of G such that

$$\lim_{n\to\infty} \frac{|F_n \triangle gF_n|}{|F_n|} = 0 \quad \text{for all } g \in G, \tag{4}$$

where $F_n \triangle gF_n = (F_n \cup gF_n) \setminus (F_n \cap gF_n)$ is the symmetric difference of F_n and gF_n.

For instance, for $G = \mathbb{Z}$ one can take as Følner sets the intervals $[-n, n]$ where $[-n, n] = \{-n, -n+1, \ldots, -1, 0, 1, \ldots, n\}$, $n \in \mathbb{N}$. Analogously, for $G = \mathbb{Z}^2$ one can take as Følner sets the squares $F_n = [-n, n] \times [-n, n]$.

Suppose that G is countable and amenable, and fix a Følner sequence $(F_n)_{n\in\mathbb{N}}$. Let A be a finite alphabet. A subset $X \subset A^G$ is said to be *G-invariant* if $x \in X$ implies that $x^g \in X$ for all $g \in G$. The *entropy* $\mathrm{ent}(X)$ of a G-invariant subset $X \subset A^G$ is defined by

$$\mathrm{ent}(X) = \lim_{n\to\infty} \frac{\log |X_{F_n}|}{|F_n|} \tag{5}$$

where, for any subset $F \subset G$

$$X_F = \{x|_F : x \in X\} \tag{6}$$

denotes the set of restrictions to F of all configurations in X.

By using a result of Ornstein and Weiss [31], it can be shown that the above limit in (5) exists and does not depend on the particular Følner sequence $(F_n)_{n\in\mathbb{N}}$.

One clearly has $\mathrm{ent}(A^G) = \log|A|$ and $\mathrm{ent}(X) \leq \mathrm{ent}(Y)$ if $X \subset Y$ are G-invariant subsets of A^G.

Theorem 3 (Ceccherini-Silberstein, Machì and Scarabotti [9]) *Let G be a countable amenable group and let A be a finite set. Let $\tau : A^G \to A^G$ be a cellular automaton. The following are equivalent:*

(a) τ is surjective (i. e. there are no GOE configurations);
(b) τ is pre-injective;
(c) $\mathrm{ent}(\tau(A^G)) = \log|A|$.

Example 5 Let G be a group. Let M be a finite subset of G with at least three elements. Let $A = \{0, 1\}$ and consider the majority action cellular automaton $\tau : A^G \to A^G$ associated with G and M (see Subsect. "Cellular Automata"). Clearly τ is not pre-injective. Indeed the configurations $x_1, x_2 \in A^G$ defined by $x_1(g) = 0$ for all $g \in G$ and

$$x_2(g) = \begin{cases} 1 & \text{if } g = 1_G \\ 0 & \text{otherwise} \end{cases}$$

are almost equal and $\tau(x_2) = x_1 = \tau(x_1)$. By applying Theorem 3 we deduce that τ is not surjective when G is a countable amenable group.

In the example below we show that for the non-amenable group F_2, the free group of rank two, the implication (a) $\Rightarrow$ (b) fails to hold. In Example 9 in Sect. "Linear Cellular Automata" we show that also the converse implication fails to hold, in general, for cellular automata over F_2.

Example 6 Let $G = F_2$ be the free group on two generators a and b. Take $A = \{0, 1\}$ and $M = \{a, a^{-1}, b, b^{-1}\} = S$. Consider the majority action cellular automaton $\tau : A^G \to A^G$ associated with G and M. As observed above, τ is not pre-injective. However, τ is surjective. To see this, let $z \in A^G$. Let us show that there exists $x \in A^G$ such that $\tau(x) = z$. We first set $x(1_G) = 0$. For $g \in G$ such that $g \neq 1_G$, denote by $g' \in G$ the unique element such that $\ell(g') = \ell(g) - 1$ and $g = g's'$ for some $s' \in S$. Then set $x(g) = z(g')$. We clearly have $\tau(x) = z$. This shows that τ is surjective (see Fig. 3).

Recently, Laurent Bartholdi [2] proved that if G is a non-amenable group, then there exists a cellular automaton $\tau : A^G \to A^G$ with finite alphabet which is surjective but not pre-injective. In other words, the implication "(a) τ is surjective $\Rightarrow$ (b) τ is pre-injective" in Theorem 3 (which corresponds to the generalization of Moore's theorem)

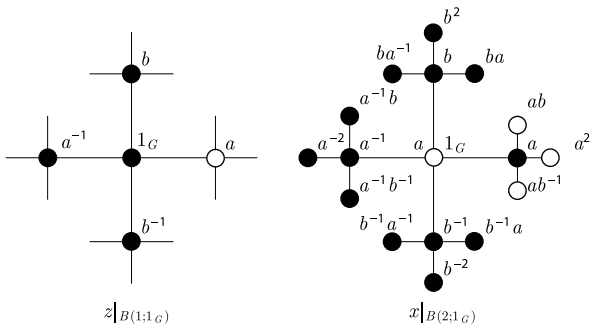

Cellular Automata and Groups, Figure 3
The construction of $x \in A^G$ such that $\tau(x) = z$

holds true only if the group G is amenable. In particular, the Garden of Eden Theorem (Theorem 3) holds true if and only if the universe G is amenable. This clearly gives a new characterization of amenability for groups in terms of cellular automata.

However, up to now, it is not known whether the validity of the implication (b) $\Rightarrow$ (a) in Theorem 3 (which corresponds to the generalization of Myhill's theorem) holds true only if the group G is amenable.

Surjunctivity

A group G is said to be *surjunctive* (a terminology due to Gottschalk [17]) if given any finite alphabet A, every injective cellular automaton $\tau: A^G \to A^G$ is surjective. In other words, uniqueness implies existence for solutions of the equation $y = \tau(x)$. This property is reminiscent of several other classes of mathematical objects for which injective endomorphisms are always surjective (finite sets, finite-dimensional vector spaces, Artinian modules, complex algebraic varieties, co-Hopfian groups, etc.).

Recall that a group G is said to be *residually finite* if for every element $g \neq 1_G$ in G there exist a finite group F and a homomorphism $h: G \to F$ such that $h(g) \neq 1_F$. This amounts to saying that the intersection of all subgroups of finite index of G is reduced to the identity element. From a dynamical viewpoint we have the following characterization of residual finiteness. Given a finite set A a configuration $x \in A^G$ is said to be *periodic* if its *G-orbit* $\{x^g : g \in G\} \subset A^G$ is finite. Then G is residually finite if and only if the set of periodic configurations is dense in A^G.

The class of residually finite groups is quite large. For instance, every finitely generated subgroup of $\mathrm{GL}_n(\mathbb{C})$, the group of n by n invertible matrices over the complex numbers, is residually finite. However, there are finitely generated amenable groups which are not residually finite.

Lawton [17,18] proved that residually finite groups are surjunctive.

From Theorem 3 one immediately deduces the following

Corollary 2 *Amenable groups are surjunctive.*

Note that the implication "surjectivity $\Rightarrow$ injectivity" fails to hold, in general, for cellular automata with finite alphabet over amenable groups, even for $G = \mathbb{Z}$. Take, for instance $A = \{0, 1\}$ and $\tau: A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ defined by $\tau(x)(n) = x(n+1) - x(n)$ for all $x \in A^{\mathbb{Z}}$ and $n \in \mathbb{Z}$. This cellular automaton is surjective but not injective. See also Example 8 below.

Sofic Groups

Let S be a set. An *S-labeled graph* is a triple (Q, E, λ), where Q is a set, E is a symmetric subset of $Q \times Q$ and λ is a map from E to S. The set Q is the set of *vertices*, E is the set of *edges* and $\lambda: E \to S$ is the *labeling map* of the S-labeled graph (Q, E, λ). We shall view every subgraph of a labeled graph as a labeled graph in the obvious way. Also, for $r \in \mathbb{R}$ and $q \in Q$, we denote by $B(q, r) = \{q' \in Q : d(q, q') \leq r\}$ the *ball* of *radius* r centered at q (here d denotes the *graph distance* in Q).

Let (Q, E, λ) and (Q', E', λ') be S-labeled graphs. Two vertices $q \in Q$ and $q' \in Q'$ are said to be *r-equivalent*, and we write $q \sim_r q'$, if the balls $B(q, r)$ and $B(q', r)$ are isomorphic as labeled graphs (i. e. there is a bijection $\varphi: B(q, r) \to B(q', r)$ sending q to q' such that $(q_1, q_2) \in E \cap (B(q, r) \times B(q, r))$ if and only if $(\varphi(q_1), \varphi(q_2)) \in E' \cap (B(q', r) \times B(q', r))$ and $\lambda(q_1, q_2) = \lambda'(\varphi(q_1), \varphi(q_2))$.

Let G be a finitely generated group and let S be a finite symmetric ($S = S^{-1}$) generating subset of G. We denote by $Cay(G, S)$ the *Cayley graph* of G with respect to S. Its vertex set is G and $(g, g') \in G \times G$ is an edge if $s := g^{-1}g' \in S$, if this is the case, its label is $\lambda(g, g') = s$.

The group G is said to be *sofic* if for all $\varepsilon > 0$ and $r \in \mathbb{N}$ there exists a finite S-labeled graph (Q, E, λ) such that the set $Q(r) \subset Q$ defined by $Q(r) = \{q \in Q : q \sim_r 1_G\}$ (here 1_G is considered as a vertex in $Cay(G, S)$) satisfies

$$|Q(r)| \geq (1 - \varepsilon)|Q|. \tag{7}$$

It can be shown (see [38]) that this definition does not depend on the choice of S and that it can be extended as follows. A (not necessarily finitely generated) group G is said to be *sofic* if all of its finitely generated subgroups are sofic.

Sofic groups were introduced by M. Gromov in [20]. The sofic terminology is due to B. Weiss [38]. The class of sofic groups contains, in particular, all residually finite

groups, all amenable groups, and it is closed under direct products, free products, taking subgroups and extensions by amenable groups, and taking direct limits [12].

The following generalizes Lawton's result mentioned in Subsect. "Surjunctivity" as well as Corollary 2.

Theorem 4 (Gromov and Weiss [20,38]) *Sofic groups are surjunctive.*

We end this section by mentioning that there is no known example of a non-surjunctive group nor even of a non-sofic group up to now.

Linear Cellular Automata

Let G be a group and V be a vector space over a field $\mathbb{K}$. The configuration space $V^G = \{x\colon G \to V\}$ is a vector space over $\mathbb{K}$. Simply set $(x+y)(g) = x(g) + y(g)$ and $(\lambda x)(g) = \lambda x(g)$ for all $x, y \in V^G$, $g \in G$ and $\lambda \in \mathbb{K}$. The zero vector is the zero configuration $\mathbf{0}(g) = 0$ for all $g \in G$. The *support* of a configuration $x \in A^G$ is the subset $\mathrm{supp}(x) = \{g \in G\colon x(g) \neq 0\} \subset G$. We denote by $V[G] = \{x \in V^G\colon x \sim_{a.e.} \mathbf{0}\}$ the subspace of all configurations with *finite* support.

A linear cellular automaton is a cellular automaton $\tau\colon V^G \to V^G$ which is a linear map, that is, $\tau(x+y) = \tau(x) + \tau(y)$ and $\tau(\lambda x) = \lambda\tau(x)$ for all $x, y \in V^G$ and $\lambda \in \mathbb{K}$. This is equivalent to the linearity of the local defining map $\mu\colon V^M \to V$. We denote by $\mathrm{LCA}(G, V)$ the space of all linear cellular automata over G with coefficients in V.

Example 7 The Laplacian (cf. Example 4) is a linear cellular automaton.

Remark If the field is finite, so that necessarily $|\mathbb{K}| = p^n$ with p a prime number, and V has finite dimension over $\mathbb{K}$, then the vector space V is also finite.

It is easy to see that if $\tau \in \mathrm{LCA}(G, V)$ and $x \in V^G$ has finite support, then $\tau(x)$ also has finite support (in fact $\mathrm{supp}(\tau(x)) \subset \mathrm{supp}(x)M$). In other words $\tau(V[G]) \subset V[G]$. We denote by $\tau_0 = \tau|_{V[G]}\colon V[G] \to V[G]$ the restriction of τ to $V[G]$. We then have the following characterization of pre-injectivity for linear cellular automata.

Proposition 1 *The linear cellular automaton $\tau \in \mathrm{LCA}(G, V)$ is pre-injective if and only if $\tau_0\colon V[G] \to V[G]$ is injective.*

Note that if G is countable, V^G admits also a structure of a *metric space* (the distance function (3) is defined in the same way). Then $V[G]$ is dense in V^G with respect to the topology induced by the distance (3). However, V^G is no longer a compact space so that many topological arguments based on compactness need to be obtained with an alternative method. As an example, the following linear analogue of Corollary 2 needs an appropriate proof.

Theorem 5 ([6]) *Let G be a countable group and let V be a finite-dimensional vector space over a field $\mathbb{K}$. Suppose that $\tau\colon V^G \to V^G$ is a linear cellular automaton.*

(i) $\tau(V^G)$ is a closed subspace of V^G.
(ii) If τ is bijective then the inverse map $\tau^{-1}\colon V^G \to V^G$ is also a linear cellular automaton.

Mean Dimension and the GOE Theorem

Let G be a countable amenable group and V a finite-dimensional vector space over a field $\mathbb{K}$. Fix a Følner sequence $(F_n)_{n\in\mathbb{N}}$ for G. The *mean dimension* of a G-invariant vector subspace $X \subset V^G$, which plays the role of entropy used in the finite alphabet case, is the non-negative number

$$\mathrm{mdim}(X) = \lim_{n\to\infty} \frac{\dim(X_{F_n})}{|F_n|}\,, \tag{8}$$

where X_{F_n} is defined as in (6).

The result of Ornstein and Weiss [31] already mentioned above implies that the limit (8) exists and does not depend on the particular choice of the Følner sequence $(F_n)_{n\in\mathbb{N}}$ for G.

Note that it immediately follows from this definition that $\mathrm{mdim}(V^G) = \dim(V)$ and that $\mathrm{mdim}(X) \leq \mathrm{mdim}(Y) \leq \dim(V)$ for all G-invariant vector subspaces $X \subset Y$ of V^G.

The linear analogue of the Garden of Eden theorem for linear cellular automata states as follows.

Theorem 6 (Ceccherini-Silberstein and Coornaert, [5]) *Let V be a finite-dimensional vector space over a field $\mathbb{K}$ and let G be a countable amenable group. Let $\tau\colon V^G \to V^G$ be a linear cellular automaton. Then the following are equivalent.*

(a) τ is surjective (i. e. there are no GOE configurations);
(b) τ is pre-injective;
(c) $\mathrm{mdim}(\tau(V^G)) = \dim(V)$.

As an application we present the following example.

Example 8 Let G be a finitely generated group. Let $S \subset G$ be a finite generating subset (not necessarily symmetric) such that $1_G \notin S$, and denote by $\Delta_S\colon \mathbb{R}^G \to \mathbb{R}^G$ the corresponding Laplacian (cf. Example 4). It follows from the Maximum Principle, see also Proposition 6.4 in [5], that if the group G is infinite, then the linear cellular automaton

Δ_S is pre-injective (though not injective since the constant configurations are in the kernel of Δ_S).

Thus, as a consequence of Theorem 6, we deduce that Δ_S is also surjective if G is an infinite amenable group.

Actually, one has that Δ_S is always surjective whenever G is infinite. Indeed, denoting by $P_S = I - \Delta_S$ the *Markov operator* associated with S, one has that if G is non-amenable then G is *transient* i. e. the series $\sum_{n=0}^{\infty}(P_S)^n$ converges [39] (in fact, by a profound result of N. Varopoulos (1986, see, e. g. [35]), G is transient if and only if it has no finite index subgroup isomorphic to either $\mathbb{Z}$ or $\mathbb{Z}^2$). We denote by $\mathcal{G}_S$ the sum of this series. It is called the *Green operator* of G.

But then, for $f \in \mathbb{R}[G]$ the function $g = \mathcal{G}_S f \in \mathbb{R}^G$ clearly satisfies $\Delta_S g = (I - P_S)g = f$. This shows that $\Delta_S(\mathbb{R}^G) \supset \mathbb{R}[G]$ and, by virtue of Theorem 5 (i) and the density of $\mathbb{R}[G]$ in $\mathbb{R}^G$, one has indeed $\Delta_S(\mathbb{R}^G) = \mathbb{R}^G$, that is, Δ_S is surjective. We thank Vadim Kaimanovich and Nic Varopoulos for clarifying this point to us.

In the example below we show that the implication (b) $\Rightarrow$ (a) in Theorem 6 fails to hold, in general, for linear cellular automata over the free group of rank two. Note that if the field is finite, then this example also provides an instance of the failure of the implication (b) $\Rightarrow$ (a) in Theorem 3 when $G = F_2$.

Example 9 Let $G = F_2$ be the free group on two generators a and b. Let $\mathbb{K}$ be a field and set $V = \mathbb{K}^2$. Consider the endomorphisms p and q of V defined by $p(\alpha, \beta) = (\alpha, 0)$ and $q(\alpha, \beta) = (\beta, 0)$ for all $(\alpha, \beta) \in V$. Let $\tau\colon V^G \to V^G$ be the linear cellular automaton, with memory set $M = \{a, b, a^{-1}, b^{-1}\}$, given by

$$\tau(x)(g) = p(x(ga)) + q(x(gb)) + p(x(ga^{-1}) + q(x(gb^{-1}))$$

for all $x \in V^G$, $g \in G$. The image of τ is contained in $(\mathbb{K} \times \{0\})^G$. Therefore τ is not surjective. Let us show that τ is pre-injective. Assume that there is an element $x_0 \in V^G$ with non-empty finite support $\Omega \subset G$ such that $\tau(x_0) = 0$. Consider a vertex $g_0 \in \Omega$ at maximal distance n_0 from the identity in the Cayley graph of G. The vertex g_0 has at least three adjacent vertices at distance $n_0 + 1$ from the identity. It follows from the definition of τ that $\tau(x_0)$ does not vanish at (at least) one of these three vertices. This gives a contradiction. Thus τ is pre-injective.

The following, which is a linear version of Example 6, provides an instance of the failure of the implication (a) $\Rightarrow$ (b) in Theorem 6 when $G = F_2$.

Example 10 Let $G = F_2$ be the free group on two generators a and b. Let $\mathbb{K}$ be a field and set $V = \mathbb{K}^2$. Consider the endomorphisms p' and q' of V defined by $p'(\alpha, \beta) = (\alpha, 0)$ and $q'(\alpha, \beta) = (0, \alpha)$ for all $(\alpha, \beta) \in V$.

Let $\tau\colon V^G \to V^G$ be the $\mathbb{K}$-linear cellular automaton, with memory set $S = \{a, b, a^{-1}, b^{-1}\}$, given by

$$\tau(x)(g) = p'(x(ga)) + p'(x(ga^{-1})) + q'(x(gb)) + q'(x(gb^{-1}))$$

for all $x \in V^G$ and $g \in G$.

Consider the configuration $x_0 \in V^G$ defined by

$$x_0(g) = \begin{cases} (0,1) & \text{if } g = 1_G \\ (0,0) & \text{otherwise}\,. \end{cases}$$

Then, x_0 is almost equal to 0 and $\tau(x_0) = 0$. This shows that τ is not pre-injective (cf. Proposition 1).

However, τ is surjective. To see this, let $z = (z_1, z_2) \in \mathbb{K}^G \times \mathbb{K}^G = V^G$. Let us show that there exists $x \in V^G$ such that $\tau(x) = z$. We define $x(g)$ by induction on the graph distance, which we denote by $|g|$, of $g \in G$ from 1_G in the Cayley graph of G. We first set $x(1_G) = (0, 0)$.

Then, for $s \in S$ we set

$$x(s) = \begin{cases} (z_1(1_G), 0) & \text{if } s = a \\ (z_2(1_G), 0) & \text{if } s = b \\ (0,0) & \text{otherwise}\,. \end{cases}$$

Suppose that $x(g)$ has been defined for all $g \in G$ with $|g| \le n$, for some $n \ge 1$. For $g \in G$ with $|g| = n$, let $g' \in G$ and $s' \in S$ be the unique elements such that $|g'| = n-1$ and $g = g's'$. Then, for $s \in S$ with $s's \neq 1_G$, we set

$$x(gs) = \begin{cases} (z_1(g) - x_1(g'), 0) & \text{if } s' \in \{a, a^{-1}\} \text{ and } s = s' \\ (z_2(g), 0) & \text{if } s' \in \{a, a^{-1}\} \text{ and } s = b \\ (z_1(g), 0) & \text{if } s' \in \{b, b^{-1}\} \text{ and } s = a \\ (z_2(g) - x_2(g'), 0) & \text{if } s' \in \{b, b^{-1}\} \text{ and } s = s' \\ (0,0) & \text{otherwise}\,. \end{cases}$$

Then one easily checks that $\tau(x) = z$. This shows that τ is surjective.

We now show that, for any group, both implications of the equivalence (a) $\Leftrightarrow$ (b) in Theorem 6 fail to hold, in general, when the vector space V is infinite-dimensional.

Example 11 Let V be an infinite-dimensional vector space over a field $\mathbb{K}$ and let G be any group. Let us choose a basis B for V. Every map $\alpha\colon B \to B$ uniquely extends to a linear map $\widetilde{\alpha}\colon V \to V$. The product map $\tau = \widetilde{\alpha}^G\colon V^G \to V^G$ is a linear cellular automaton with memory set $M = \{1_G\}$ and local defining map $\widetilde{\alpha}$. Since B

is infinite, we can find a map $\alpha\colon B \to B$ which is surjective but not injective (resp. injective but not surjective). Clearly, the associated linear cellular automaton τ is surjective but not pre-injective (resp. injective but not surjective).

We say that a group G is *L-surjunctive* if for any field $\mathbb{K}$ and for any finite dimensional vector space V over $\mathbb{K}$, every injective linear cellular automaton $\tau\colon V^G \to V^G$ is surjective.

The following is the linear analogue of the Gromov–Weiss theorem (Theorem 4).

Theorem 7 ([6]) *Sofic groups are L-surjunctive.*

Group Rings and Kaplansky Conjectures

Irving Kaplansky [23,24] posed some famous problems in the theory of group rings. Here we establish some connections between these problems and the theory of linear cellular automata.

Group Rings

Let G be a group and $\mathbb{K}$ a field. A natural basis for $\mathbb{K}[G]$, the subspace of finitely supported configurations in $\mathbb{K}^G$, is given by $\{\delta_g\colon g \in G\}$, where $\delta_g\colon G \to \mathbb{K}$ is defined by $\delta_g(g) = 1$ and $\delta_g(g') = 0$ if $g' \neq g$. Also, $\mathbb{K}[G]$ can be endowed with a ring structure by defining the *convolution* product xy of two elements $x, y \in \mathbb{K}[G]$ by setting, for all $g \in G$,

$$[xy](g) = \sum_{h\in G} x(h)y(h^{-1}g). \tag{9}$$

One has $\delta_g\delta_h = \delta_{gh}$ and $\delta_h x = x^{h^{-1}}$ for all $g, h \in G$ and $x \in \mathbb{K}[G]$, in particular, δ_{1_G} is the unit element in $\mathbb{K}[G]$. The ring $\mathbb{K}[G]$ is called the *group ring* of G with coefficients in $\mathbb{K}$. Note that the product map $(x, y) \to xy$ is $\mathbb{K}$-linear so that $\mathbb{K}[G]$ is in fact a $\mathbb{K}$-algebra.

Note also that (9) makes sense for $x, y \in K^G$ and at least one is finitely supported. Moreover, the group G, via the map $G \ni g \mapsto \delta_g \in \mathbb{K}[G]$, can be identified with a subgroup of the group of invertible elements in $\mathbb{K}[G]$. This way, every element x of $\mathbb{K}[G]$ can be uniquely expressed as $x = \sum_{g\in G} x(g)g$.

The Matrix Representation of Linear Cellular Automata

Let $d \geq 1$ be an integer. Denote by $\mathrm{Mat}_d(\mathbb{K}[G])$ the $\mathbb{K}$-algebra of $d \times d$ matrices with coefficients in $\mathbb{K}[G]$. For $x = (x_1, x_2, \ldots, x_d) \in (\mathbb{K}^G)^d$ and $\alpha = (\alpha_{ij})_{i,j=1}^d \in \mathrm{Mat}_d(\mathbb{K}[G])$, we define $x\alpha = (y_1, y_2, \ldots, y_d) \in (\mathbb{K}^G)^d$ by setting $y_j = \sum_{i=1}^d x_i\alpha_{ij}$ for all $j = 1, 2, \ldots, d$, where $x_i\alpha_{ij}$ is the convolution product of $x_i \in \mathbb{K}^G$ and $\alpha_{ij} \in \mathbb{K}[G]$ defined using (1).

The map $\mathrm{Mat}_d(\mathbb{K}[G]) \ni \alpha \mapsto \overline{\alpha} \in \mathrm{Mat}_d(\mathbb{K}[G])$, where $\overline{\alpha}_{ij}(g) = \alpha_{ji}(g^{-1})$ for all $g \in G$ and $i, j = 1, 2, \ldots, d$ is an *anti-involution* of the algebra $\mathrm{Mat}_d(\mathbb{K}[G])$, since $\overline{\overline{\alpha}} = \alpha$ and $\overline{\alpha\beta} = \overline{\beta}\,\overline{\alpha}$, for all $\alpha, \beta \in \mathrm{Mat}_d(\mathbb{K}[G])$.

Let $\alpha \in \mathrm{Mat}_d(\mathbb{K}[G])$ and define the map $\tau_\alpha : (\mathbb{K}^d)^G \to (\mathbb{K}^d)^G$ by setting

$$\tau_\alpha(x) = x\overline{\alpha}$$

for all $x = (x_1, x_2, \ldots, x_d) \in (\mathbb{K}^G)^d = (\mathbb{K}^d)^G$.

Theorem 8 ([6]) *For $\alpha \in \mathrm{Mat}_d(\mathbb{K}[G])$ the map $\tau_\alpha\colon (K^d)^G \to (\mathbb{K}^d)^G$ is a linear cellular automaton. Moreover, the map $\mathrm{Mat}_d(\mathbb{K}[G]) \ni \alpha \to \tau_\alpha \in \mathrm{LCA}(G, \mathbb{K}^d)$ is an isomorphism of $\mathbb{K}$-algebras.*

Remark When $G = \mathbb{Z}$ and $\mathbb{K}$ is a finite field, linear cellular automata $\tau_\alpha \in \mathrm{LCA}(\mathbb{Z}, \mathbb{K}^d)$ with $\alpha \in \mathrm{Mat}_d(\mathbb{K}[\mathbb{Z}])$, are called *convolutional encoders* in Sect. 1.6 of [26].

Zero-Divisors in Group Rings

Let R be a ring. A non zero element $a \in R$ is said to be a *left* (resp. *right*) *zero-divisor* provided there exists a non zero element $b \in R$ such that $ab = 0$ (resp. $ba = 0$).

The following result relates the notion of a zero-divisor in a group ring $\mathbb{K}[G]$ with the pre-injectivity of one-dimensional linear cellular automata over the group G. We use the same notation as in Theorem 8 (here $d = 1$).

Lemma 1 ([5]) *Let G be a group and let $\mathbb{K}$ be a field. An element $\alpha \in \mathbb{K}[G]$ is a left zero-divisor if and only if the linear cellular automaton $\tau_\alpha\colon \mathbb{K}^G \to \mathbb{K}^G$ is not pre-injective.*

Let G be a group and suppose that it contains an element g_0 of finite order $n \geq 2$. Then we have

$$(1_G - g_0)\left(1_G + g_0 + g_0^2 + \cdots + g_0^{n-1}\right) = 0\,,$$

showing that $\mathbb{K}[G]$ has zero-divisors.

A group is *torsion-free* if it has no non-trivial element of finite order. *Kaplansky zero-divisor problem* [23] asks whether $\mathbb{K}[G]$ has no zero-divisors whenever G is torsion-free. In virtue of Lemma 1 and Theorem 8 we can state it as follows (see [5]).

Problem 1 (Kaplansky zero-divisor problem reformulated in terms of cellular automata) *Let G be a torsion-free group and let $\mathbb{K}$ be a field. Is it true that every non-identically-zero linear cellular automaton $\tau\colon \mathbb{K}^G \to \mathbb{K}^G$ is pre-injective?*

The zero-divisor problem is known to have an affirmative answer for a wide class of groups including the free groups, the free Abelian groups, the fundamental groups of surfaces and the braid groups B_n.

Combining Lemma 1 with Theorem 6 we deduce the following.

Corollary 3 *Let G be a countable amenable group and let $\mathbb{K}$ be a field. Suppose that $\mathbb{K}[G]$ has no zero-divisors. Then every non-identically-zero linear cellular automaton $\tau : \mathbb{K}^G \to \mathbb{K}^G$ is surjective.*

The class of *elementary amenable* groups is the smallest class of groups containing all finite and all Abelian groups that is closed under taking extensions and directed unions. It is known, see Theorem 1.4 in [25], that if G is a torsion-free elementary amenable group, then $\mathbb{K}[G]$ has no zero-divisor for any field $\mathbb{K}$. As a consequence, the conclusion of Corollary 3 holds for all torsion-free elementary amenable groups.

Stable Finiteness of Group Rings

Recall that a ring R with identity element 1_R is said to be *directly finite* if one-sided inverses in R are also two-sided inverses, i. e., $ab = 1_R$ implies $ba = 1_R$ for $a, b \in R$. The ring R is said to be *stably finite* if the ring $M_d(R)$ of $d \times d$ matrices with coefficients in R is directly finite for all integers $d \geq 1$.

Commutative rings and finite rings are obviously directly finite. Also observe that if elements a and b of a ring R satisfy $ab = 1_R$ then $(ba)^2 = ba$, that is, ba is an idempotent. Therefore if the only idempotent of R are 0_R and 1_R (e. g. if R has no zero-divisors) then R is directly finite. The ring of endomorphisms of an infinite-dimensional vector space yields an example of a ring which is not directly finite.

Kaplansky [24] observed that, for any group G and any field $\mathbb{K}$ of characteristic 0, the group ring $\mathbb{K}[G]$ is stably finite and asked whether this property remains true for fields of characteristic $p > 0$. We have that this holds for L-surjunctive groups.

Using the matrix representation of linear cellular automata (Theorem 8) one has the following characterization of L-surjunctivity.

Theorem 9 *For a group G, a field $\mathbb{K}$, and an integer $d \geq 1$ the following conditions are equivalent:*

(a) Every injective linear cellular automaton $\tau : (\mathbb{K}^d)^G \to (\mathbb{K}^d)^G$ is surjective;
(b) The ring $\mathrm{Mat}_d(\mathbb{K}[G])$ is directly finite.

As a consequence, a group G is L-surjunctive if and only if the group ring $\mathbb{K}[G]$ is stably finite for any field $\mathbb{K}$.

From Theorem 7 and Theorem 9 we deduce the following result previously established by G. Elek and A. Szabó using different methods.

Corollary 4 ([6,11]) *Let G be a sofic group and $\mathbb{K}$ any field. Then the group ring $\mathbb{K}[G]$ is stably finite.*

Future Directions

We indicate some open problems related to the topics that we have treated in this article.

Garden of Eden Theorem

As we mentioned in the Subsect. "Amenable Groups", it would be interesting to determine whether the Myhill theorem (i. e. the implication (b) $\Rightarrow$ (a) in Theorem 3), which holds for amenable groups, but fails to hold for groups containing the free group F_2 (cf. Example 9), holds or not for the non-amenable groups with no free subgroups (such as the free Burnside groups $B(m, n)$, $m \geq 665$ odd, see [1]). Note that a negative answer would give another new characterization of amenability for groups.

Problem 2 *Determine whether the Myhill theorem (pre-injectivity implies surjectivity) for cellular automata with finite alphabet holds only for amenable groups.*

It turns out that Bartholdi's cellular automata ([2], see Subsect. "Amenable Groups") are not linear, so that the question whether the linear GOE theorem (Theorem 6) holds also for non-amenable groups remains an open problem.

Problem 3 *Determine whether the GOE theorem for linear cellular automata over finite-dimensional vector spaces holds only for amenable groups or not. More precisely, determine whether Moore's theorem and/or Myhill's theorem for linear cellular automata over finite dimensional vector spaces hold only for amenable groups or not.*

In [7] we generalized the GOE theorem to linear cellular automata over semisimple modules (over a, not necessarily commutative, ring R) of finite length with universe an amenable group. A vector space is a (semisimple) left module over a field. The finite length condition for modules (which corresponds to the ascending chain condition (Noetherianity) and to the descending chain condition (Artinianity)) is the natural analogue of the notion of finite dimension for vector spaces.

The Garden of Eden theorem can be also generalized by looking at subshifts. Given an alphabet A and a countable group G a *subshift* $X \subset A^G$ is a subset which is closed

(in the topology induced by the metric (3)) and G-invariant (if $x \in X$ then $x^g \in X$ for all $g \in G$). If G is amenable and $X \subset A^G$ is a subshift the quantity (5) is the entropy of X. Given two subshifts $X, Y \subset A^G$ we define a cellular automaton $\tau\colon X \to Y$ as the restriction to X of a cellular automaton $\overline{\tau}\colon A^G \to A^G$ such that $\overline{\tau}(X) \subset Y$.

Problem 4 *Let G be a countable amenable group, A a finite alphabet and $X, Y \subset A^G$ two subshifts with* $\mathrm{ent}(X) = \mathrm{ent}(Y)$. *Prove, under suitable conditions, the GOE theorem for cellular automata $\tau\colon X \to Y$.*

We mention that the GOE theorem for subshifts over amenable groups fails to hold in general with no additional hypotheses on the subshifts.

Let $G = \mathbb{Z}$ be the infinite cyclic group and let A be a finite alphabet. For $n, m \in \mathbb{Z}$ we set $[n, m] = \{x \in \mathbb{Z}\colon n \leq x \leq m\}$. Also, we denote by $A^* = \{a_1 a_2 \cdots a_n\colon a_i \in A, 1 \leq i \leq n, n \in \mathbb{N}\}$ the set of all *words* over A. The *length* of a word $w = a_1 a_2 \cdots a_n \in A^*$ is $\ell(w) = n$. Given a subset $X \subset A^G$ we denote by $W(X) = \{w \in A^*\colon \exists x \in X, \text{ s.t. } w = x|_{[0,\ell(w)]}\}$ the *language* associated with X. Then one says that a subshift $X \subset A^{\mathbb{Z}}$ is irreducible if, for any two subwords $w_1, w_2 \in W(X)$ there exists $w_3 \in A^*$ (necessarily in $W(X)$) such that $w_1 w_3 w_2 \in W(X)$.

Also it can be shown that, for a subshift $X \subset A^{\mathbb{Z}}$, there exists a set $\mathcal{F} \subset A^*$ of so-called *forbidden words*, such that setting

$$X_{\mathcal{F}} := \{x \in A^{\mathbb{Z}}\colon x^g|_{[0,n]} \notin \mathcal{F}, \forall n \in \mathbb{N}, \forall g \in \mathbb{Z}\}\,,$$

one has $X = X_{\mathcal{F}}$. Then a subshift $X \subset A^{\mathbb{Z}}$ is of *finite type* if $X = X_{\mathcal{F}}$ for some *finite* set $\mathcal{F} \subset A^*$. Finally, X is *sofic* (same etymology but a different meaning from that used for groups in Subsect. "Sofic Groups") if $X = \tau(Y)$ for some cellular automaton $\tau\colon A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ and some subshift $Y \subset A^{\mathbb{Z}}$ of finite type.

In [13], F. Fiorenzi considered cellular automata on irreducible subshifts of finite type inside $A^{\mathbb{Z}}$(A finite). She proved the GOE theorem for such cellular automata and provided examples of cellular automata on subshifts of finite type which are not irreducible and for which both implications of the GOE theorem fail to hold. She also provided an example of a cellular automaton on an irreducible subshift which is sofic but not of finite type for which the Moore theorem (namely the implication (a) $\Rightarrow$ (b) in Theorem 3) fails to hold. Note that it is well known (cf. 2.1 in [13]) that, under the same hypotheses, the Myhill theorem (namely the implication (b) $\Rightarrow$ (a) in Theorem 3) always holds true ($G = \mathbb{Z}$).

More generally, for groups G other than the integers $\mathbb{Z}$, one needs appropriate notions of *irreducibility* for the subsets $X \subset A^G$ as investigated by Fiorenzi [14,15].

Surjunctivity

Problem 5 *Prove or disprove that all groups are surjunctive.*

A positive answer to the previous problem could be derived by positively answering to the following

Problem 6 *Prove or disprove that all groups are sofic.*

By considering the linear analogue of Problem 5 we have

Problem 7 (Kaplansky's conjecture on stable finiteness of group rings) *Prove or disprove that all groups are L-surjunctive. Equivalently (cf. Theorem 9), prove or disprove that the group ring $\mathbb{K}[G]$ is stably finite for any group G and any field $\mathbb{K}$.*

Also, one could look for surjunctivity results for cellular automata with alphabets other than the finite ones and the finite dimensional vector spaces. A ring is called *left Artinian* (see [40]) if it satisfies the *descending condition on left ideals*, namely every decreasing sequence $R \supset I_1 \supset I_2 \cdots \supset I_n \supset I_{n+1} \supset \cdots$ of left ideals eventually stabilizes (there exists n_0 such that $I_n = I_{n_0}$ for all $n \geq n_0$).

In [4] we showed that if G is a residually finite group and M is an Artinian left module over a ring R (e. g. if M is a finitely generated left module over a left Artinian ring R), then every injective R-linear cellular automaton $\tau\colon M^G \to M^G$ is surjective.

In [8] we showed that if G is a sofic group (thus a weaker condition than being residually finite) and M is a left module of finite length over a ring R (thus a stronger condition than being just Artinian), then every injective R-linear cellular automaton $\tau\colon M^G \to M^G$ is surjective. As a consequence (cf. Theorem 9) we have that the group ring $R[G]$ is stably finite for any left (or right) Artinian ring R and any sofic group G.

It is therefore natural to consider the following generalization of Problem 7.

Problem 8 *Prove or disprove that the group ring $R[G]$ is stably finite for any group G and any left (or right) Artinian ring R.*

Problem 9 (Kaplansky's zero divisor conjecture for group rings) *Prove or disprove that if G is torsion-free then any non-identically zero linear cellular automaton $\tau\colon \mathbb{K}^G \to \mathbb{K}^G$, where $\mathbb{K}$ is a field, is preinjective. Equivalently (cf. Lemma 1 and Theorem 8), prove or disprove that if G is torsion-free, the group ring $\mathbb{K}[G]$ has no zero divisors.*

Bibliography

1. Adyan SI (1983) Random walks on free periodic groups. Math USSR Izvestiya 21:425–434

2. Bartholdi L (2008) A converse to Moore's and Hedlund's theorems on cellular automata. J Eur Math Soc (to appear). Preprint arXiv:0709.4280
3. Berlekamp ER, Conway JH, Guy RK (1982) Winning ways for your mathematical plays, vol 2, Chap. 25. Academic Press, London
4. Ceccherini-Silberstein T, Coornaert M (2007) On the surjunctivity of Artinian linear cellular automata over residually finite groups. In: Geometric Group Theory. Trends in Mathematics. Birkhäuser, Basel, pp 37–44
5. Ceccherini-Silberstein T, Coornaert M (2006) The garden of eden theorem for linear cellular automata. Ergod Th Dynam Syst 26:53–68
6. Ceccherini-Silberstein T, Coornaert M (2007) Injective linear cellular automata and sofic groups. Israel J Math 161:1–15
7. Ceccherini-Silberstein T, Coornaert M (2008) Amenability and linear cellular automata over semisimple modules of finite length. Comm Algebra 36:1320–1335
8. Ceccherini-Silberstein T, Coornaert M (2007) Linear cellular automata over modules of finite length and stable finiteness of group rings. J Algebra 317:743–758
9. Ceccherini-Silberstein TG, Machì A, Scarabotti F (1999) Amenable groups and cellular automata. Ann Inst Fourier 49:673–685
10. Ceccherini-Silberstein TG, Fiorenzi F, Scarabotti F (2004) The garden of eden theorem for cellular automata and for symbolic dynamical systems. In: Random walks and geometry (Vienna 2001), pp 73–108. de Gruyter, Berlin
11. Elek G, Szabó A (2004) Sofic groups and direct finiteness. J Algebra 280:426–434
12. Elek G, Szabó A (2006) On sofic groups. J Group Theor 9: 161–171
13. Fiorenzi F (2000) The Garden of eden theorem for sofic shifts. Pure Math Appl 11(3):471–484
14. Fiorenzi F (2003) Cellular automata and strongly irreducible shifts of finite type. Theor Comput Sci 299(1–3):477–493
15. Fiorenzi F (2004) Semistrongly irreducible shifts. Adv Appl Math 32(3):421–438
16. Følner E (1955) On groups with full Banach mean value. Math Scand 3:245–254
17. Gottschalk W (1973) Some general dynamical systems. In: Recent advances in topological dynamics. Lecture Notes in Math, vol 318. Springer, Berlin, pp 120–125
18. Gottschalk WH, Hedlund GA (1955) Topological dynamics. In: American Mathematical Society Colloquium Publications, vol 36. American Mathematical Society, Providence
19. Greenleaf FP (1969) Invariant means on topological groups and their applications. Van Nostrand, New York
20. Gromov M (1999) Endomorphisms of symbolic algebraic varieties. J Eur Math Soc 1:109–197
21. Ginosar Y, Holzman R (2000) The majority action on infinite graphs: Strings and puppets. Discret Math 215:59–71
22. Hungerford TW (1987) Algebra, graduate texts in mathematics. Springer, New York
23. Kaplansky I (1957) Problems in the theory of rings. Report of a conference on linear algebras, June, 1956, pp 1–3. National Academy of Sciences-National Research Council, Washington, Publ. 502
24. Kaplansky I (1969) Fields and rings. Chicago Lectures in Math. Univ. of Chicago Press, Chicago
25. Kropoholler PH, Linnell PA, Moody JA (1988) Applications of a new K-theoretic theorem to soluble group rings. Proc Amer Math Soc 104:675–684
26. Lind D, Marcus B (1995) An introduction to symbolic dynamics and coding. Cambridge University Press, Cambridge
27. Machì A, Mignosi F (1993) Garden of eden configurations for cellular automata on Cayley graphs of groups. SIAM J Discret Math 6:44–56
28. Moore EF (1963) Machine models of self-reproduction. Proc Symp Appl Math AMS 14:17–34
29. Myhill J (1963) The converse of Moore's garden of eden theorem. Proc Amer Math Soc 14:685–686
30. Ol'shanskii AY (1980) On the question of the existence of an invariant mean on a group. Uspekhi Mat Nauk 35(214)4: 199–200
31. Ornstein DS, Weiss B (1987) Entropy and isomorphism theorems for actions of amenable groups. J Anal Math 48:1–141
32. Passman DS (1985) The algebraic structure of group rings. Reprint of the 1977 original. Robert E. Krieger Publishing, Melbourne, FL
33. Paterson A (1988) Amenability, AMS mathematical surveys and monographs, vol 29. American Mathematical Society, Providence
34. Ulam S (1952) Processes and Transformations. Proc Int Cong Math 2:264–275
35. Varopoulos NT, Saloff-Coste L, Coulhon T (1992) Analysis and geometry on groups. Cambridge University Press, Cambridge
36. von Neumann J (1930) Zur allgemeinen Theorie des Maßes. Fun Math 13:73–116
37. von Neumann J (1966) In: Burks A (ed) The theory of self-reproducing automata. University of Illinois Press, Urbana, London
38. Weiss B (2000) Sofic groups and dynamical systems, (Ergodic theory and harmonic analysis, Mumbai, 1999). Sankhya Ser A 62:350–359
39. Woess W (2000) Random walks on infinite graphs and groups, Cambridge Tracts in Mathematics 138. Cambridge University Press, Cambridge
40. Zariski O, Samuel P (1975) Commutative algebra. vol 1 With the cooperation of IS Cohen. Corrected reprinting of the 1958 edition. Graduate Texts in Mathematics, No. 28. Springer, New York

Cellular Automata in Hyperbolic Spaces

Maurice Margenstern
Université Paul Verlaine, Metz, France

Article Outline

Glossary
Definition of the Subject
Introduction
The Locating Problem in Hyperbolic Tilings
Implementation of Cellular Automata in Hyperbolic Spaces
Complexity of Cellular Automata in Hyperbolic Spaces
On Specific Problems of Cellular Automata

Universality in Cellular Automata in Hyperbolic Spaces
The Connection with Tiling Problems
Future Directions
Acknowledgments
Bibliography

Glossary

Fibonacci sequence A sequence of natural integers, denoted by f_n and defined by the recurrent equation $f_{n+2} = f_{n+1} + f_n$, for all $n \in \boldsymbol{N}$, and by the initial values $f_0 = f_1 = 1$.

Hyperbolic geometry This geometry was discovered independently by both Nikolaj Lobachevsky and Jànos Bolyai around 1830. This geometry satisfies the axioms of Euclidean geometry, the axiom of parallels being excepted and replaced by the following one: through a point out of a line, there are exactly two parallels to the line. In this geometry, there are also lines which never meet: they are called **non-secant**. They are characterized by the existence, for any couple of such lines, of a unique common perpendicular. Also, in this geometry, the sum of the interior angles of a triangle is always less than π. The difference to π defines the area of the triangle. In hyperbolic geometry, distances are absolute: there is no notion of similarity. See also **Poincaré's disc**.

Pentagrid The tiling $\{5, 4\}$, with five sides and four tiles around a vertex. The angles are right angles.

Poincaré's disc A model of the hyperbolic plane inside the Euclidean plane. The points are the points which are interior to a fixed disc D. The lines are the trace in D of diameters or circles which are orthogonal to the border of D. The model was first found by Beltrami and then by Poincaré who also devised the half-plane model also called after his name. The half-plane model is a conformal image of the disc model.

Ternary heptagrid The tiling $\{7, 3\}$ with seven sides and three tiles around a vertex.

Tessellation A particular case of a finitely generated tiling. It is defined by a polygon and by its reflections in its sides and, recursively, of the images in their sides.

Tiling A partition of a geometric space; the closure of the elements of the partition are called the **tiles**. An important case is constituted by **finitely generated** tilings: there is a finite set of tiles G such that any tile is a copy of an element of G.

Tiling $\{p, q\}$ This is tessellation based on the regular polygon with p sides and with vertex angle $2\pi/q$.

Invariant group of a tiling A group of transformations which define a **bijection** on the set of tiles. Usually, in a geometrical space, they are required to belong to the group of isometries of the space.

Definition of the Subject

Cellular Automata in Hyperbolic Spaces, in short **hyperbolic cellular automata** (abbreviated HCA), consists in the implementation of Cellular Automata in the environment of a regular tiling of a hyperbolic space.

The first implementation of such an object appeared in a paper by the present author and Kenichi Morita in 1999, see [18]. In this first paper, a first solution to the location problem in this context was given. The paper also focused on one advantage of this implementation: it allows to solve **NP**-problems in polynomial time. In 2000, a second paper appeared, by the present author, where a decisive solution of the location problem was given.

The study of HCA's is a new domain in computer science, at the border of mathematics and physics. They involve hyperbolic geometry as well as elementary arithmetics and algebra with the connections of polynomials with matrices, and also some theory of fields. They also involve the theory of formal languages in connection with their properties of elementary arithmetics.

To be a melting pot of such different techniques is already something which is very interesting.

But the new field has very striking properties. Their complexity classes offer a very different landscape than that of the classical theory of complexity based on the Turing machine. They also provide a bridge between the classical theory of computation and super-Turing computations.

HCA's are a novel object with rich properties: they inherit the richness of the infinitely many regular tilings which live in the hyperbolic plane. We are at the beginning of the study and still, there are a lot of surprising results. HCA's might appear as successful as their Euclidean relatives in various domains as astrophysics, nuclear physics and computer science.

For many results indicated in this paper, we quote the book [15], where the results and its proof can be found there.

Introduction

Before the appearance of HCA's, there were a few papers on possible implementations of cellular automata in abstract contexts, especially in the case of Caley graphs, see [25]. However, as infinitely many tilings of the hyperbolic plane are not Caley graphs of a their invariant group, this method cannot solve the problem in full generality. The difficulty was the location of the tiles, the **locating**

problem. The problem is already difficult in the simple case of tessellations. Note that there are infinitely many of them in the hyperbolic plane.

The study appeared to be possible thanks to a partial solution to the locating problem, see [18,19]. A decisive step was done in [7], where the already mentioned mixing of various techniques appear. This first solution in the case of a particular tiling, the pentagrid, is dealt with in Sect. "The Locating Problem in Hyperbolic Tilings". A significant advance was performed at the occasion of the meetings organized in 2002 for the bicentury of the birth of Jànos Bolyai, the co-inventor with Nikolaj Lobachevsky of hyperbolic geometry, and also at the occasion of SCI'2002. At this conference, seven papers were presented on the topic of this article, and they had a strong impact on the later development.

This introduction should contain a paragraph on hyperbolic geometry. If the reader is not familiar with this geometry and has some time, we recommend to him/her the first chapter of [15], or any other book introducing hyperbolic geometry. For a reader which is not familiar and who has no time, we recommend the following solution. First, forget everything of Euclidean geometry and try to remember the few elements given in the glossary. Don't worry, the Euclidean objects will always be the first thing to come to your mind and, most often, they will be misleading. Second, never forget that in traveling over hyperbolic spaces, you are in the situation of a pilot of a plane flying with instruments only. You can see nothing in the usual sense of these words and, sorry to repeat it again, the usual intuition is misleading. The best introduction is to imagine that when you venture into the hyperbolic plane, always keep with you the Ariadne thread of the way backwards. Otherwise you will definitely be lost. With this precaution, you will never regret the trip. The landscape changes very quickly and you are always fascinated by its unbelievable beauty.

The Locating Problem in Hyperbolic Tilings

The Classical Case of the Pentagrid

The method introduced in [18] consists of constructing a bijection between the tiling and a tree, the **spanning tree** of the tiling. The tree is constructed in a recursive way, defined as follows, also see Fig. 1.

Initial step: P_0 is the root of the tree; it is called the **leading pentagon** of the quarter $\mathcal{Q}_0$ it is defined by its sides 1 and 5;

Induction step: Let P be the current pentagon;
if P is the leading pentagon of a quarter $\mathcal{Q}$, see P_0 in Fig. 1, the complement of P in $\mathcal{Q}$ splits into two quarters, $\mathcal{R}_1$ and $\mathcal{R}_3$ and a remaining region, $\mathcal{R}_3$ which we call a **strip**;
if P is the leading pentagon of a strip S, see P_1 in Fig. 1, the complement of P in S splits into a quarter S_1 and again a strip, S_2.

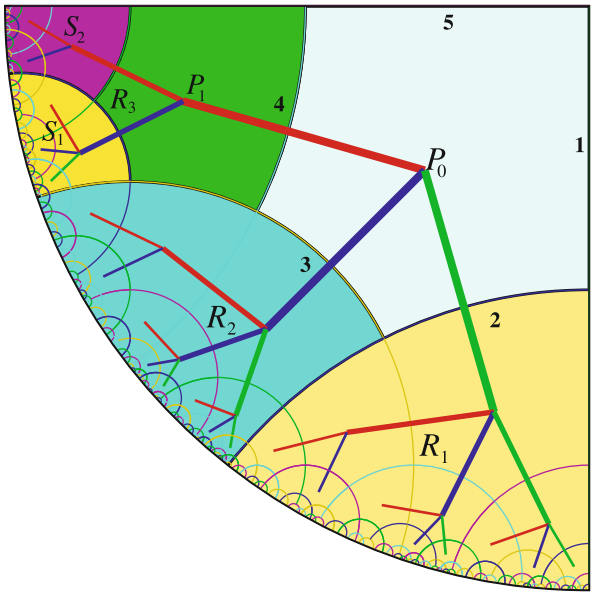

Cellular Automata in Hyperbolic Spaces, Figure 1
The pentagrid: regular pentagons with vertex angle $\pi/2$

As proved in [15], the set of tiles attached to the tree generated in this way, the **leading pentagons** of the above algorithm, is exactly the set of pentagons contained in the quarter $\mathcal{Q}_0$.

With [7,15], a new ingredient is brought in: number the nodes of the tree from the root, to which we attach 1, and then level by level, from left to right on each level, see Fig. 2. As already noticed in [18], the number of nodes of the tree which spans the tiling of a quarter which are on the same level n is f_{2n+1}, where $\{f_n\}_{n\in N}$ with $f(0) = f(1) = 1$. For this reason, the spanning tree of the pentagrid is called the **standard Fibonacci tree**, illustrated by Fig. 2.

The above splitting induces a particular structure on the standard Fibonacci tree. Define **white nodes** as nodes which have three sons and **black nodes** as nodes which have two sons. Black and white are the two possible values of the **status** of a node. Then, there is a **rule** to define the status of the sons of a node. We can write them as follows, in self-explained notations:

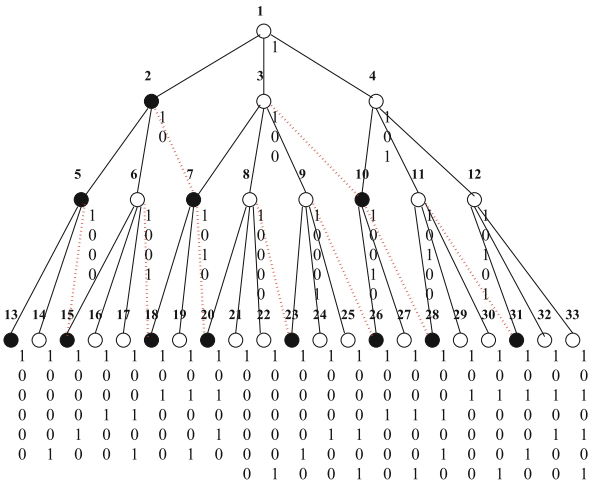

Cellular Automata in Hyperbolic Spaces, Figure 2
The Fibonacci tree

$$W \longrightarrow BWW$$
$$B \longrightarrow BW$$

Now, let us represent the numbers attached to the nodes of the Fibonacci tree in the numeration basis defined by the Fibonacci sequence itself, starting from f_1. The representation is not unique. Choose the longest representation with respect to the lexicographic order and call it the **coordinate** of the node to which the corresponding number is attached.

First, we have that the set of coordinates is a **regular language**, which is a corollary of a well known theorem, see [2]. Now we have a more interesting property, which was noticed in [7] and which we call the **preferred son property**. Let $\alpha_k..\alpha_0$ be the coordinate of a node v of the standard Fibonacci tree, with α_0 as the lightest digit of the representation. The property says that for each node v of the standard Fibonacci tree, with coordinate $\alpha_k..\alpha_0$, there is exactly one son of v whose coordinate is $\alpha_k..\alpha_0 00$. This son is called **preferred**. Moreover, there is a rule to find out the preferred son from the status of a node: in a black node, the preferred son is the black son; in a white node, the preferred son is the middle one.

Generalization: The Splitting Method

The generalization was first announced in [8]. It was then presented in [9], with a new visit to Poincaré's theorem, at the occasion of the second century of the birth of Jànos Bolyai.

The method defines a **basis of splitting** and then, the notion of a **combinatoric tiling**. Two important consequences can be derived from these very definitions to which we turn now.

Let $S_0, \ldots, S_k$ be finitely many parts of some geometric metric space X which are supposed to be closed, with non-empty interior, unbounded and simply connected. Consider also finitely many closed simply connected bounded sets $P_1, \ldots, P_h$ with $h \leq k$. Say that the S_i's and P_ℓ's constitute a **basis of splitting** *if and only if: (i) X splits into finitely many copies of S_0, (ii) any S_i splits into one copy of some P_ℓ, the* **leading tile** *of S_i, and finitely many copies of S_j's, where* **copy** *means an* **isometric image**, *and where, in the condition (ii), the copies may be of different S_j's, S_i possibly included. As usual, it is assumed that the interiors of the copies of P_ℓ and of the copies of the S_j's are pairwise disjoint. The set S_0 is called the* **head** *of the basis and the P_ℓ's are called the* **generating tiles** *and the S_i's are called the* **regions** *of the splitting.*

On the example of the pentagrid, a basis of splitting is given by a quarter $\mathcal{Q}$ and a strip S. When there is a basis of a splitting, we then define:

Say that a tiling of X is **combinatoric** *if X has a basis of splitting and if the spanning tree of the splitting yields exactly the restriction of the tiling to the head S_0 of the basis.*

In [15,18], the pentagrid is proven to be combinatoric. A lot of other tilings of the hyperbolic plane are combinatoric. In particular, all the tilings $\{p, q\}$, with $q \geq 4$, possess this property. In higher dimensions, the following tilings were proved combinatoric: the 3D tiling $\{5, 3, 4\}$, based on Poincaré's dodecahedron, see [15,21] and the 4D tiling $\{5, 3, 3, 4\}$, based on the 120-cell, see [11,15].

Once a tiling is combinatoric, from the definition of its basis of splitting, we can derive a square matrix M called the **matrix of the splitting**, see [8,15]. Its lines indicate, for each column, the number of copies of S_j's entering in the splitting of S_i. The **polynomial of the splitting** is the characteristic polynomial of M, divided by the greatest power of X it contains as a factor. In our cases, this polynomial has a greatest real root. The polynomial of the splitting induces a recurrent equation which defines the **sequence of the splitting** with appropriate initial values. The maximal representations of numbers in the basis defined by the sequence of the splitting constitute the **language of the splitting**. As proved in [15,20], the language of the splitting of the tilings $\{p, q\}$ is regular when $q \geq 4$ and $p \geq 4$.

Implementation of Cellular Automata in Hyperbolic Spaces

The implementation of HCA's is induced by the results mentioned in the previous section.

But first, let us go back to the general definition of CA's. Three conditions must be fulfilled by a set of cells to be called a cellular automaton. The cells of the automaton must uniformly be distributed in the considered space. The neighborhood of each cell is defined in a uniform way. At each top of the discrete clock, all the cells update their own state according to the same function applied to the state of the cell and the sequence of states of its neighbors.

To implement cellular automata, we have to satisfy these three requirements.

The first two conditions are easily satisfied in a tessellation. Note that this is the standard frame for CA's in the Euclidean plane and in the 3D Euclidean space.

The third condition already requires that we have a system of coordinates for the tiles at our disposal. More than three centuries after Descartes' discovery of the system of coordinates which everybody uses for the Euclidean plane, this condition is trivially fulfilled. This is not only the three-century usage. This is also the case because the mathematical structure of the group of displacements which leaves the considered tessellations of the Euclidean plane globally invariant is a very simple structure.

The situation is very different in the case of hyperbolic spaces. Before [7], there was no convenient, or at least fast, procedure to define the coordinates of the tiles in a way which is in connection with the geometrical properties of the tiling.

Now, the splitting method gives such a solution. First, it effectively exhibits a tree which generates the tiling. As Gormov pointed out, [3], hyperbolic spaces are characterized by a tree structure. Second, it provides **fast** algorithms to handle these coordinates. By fast, we mean that the basic algorithms we need are **linear** in time with respect to the coordinate of the initial point. Note that nobody really matters with the fact that addition of vectors in Euclidean coordinates is linear while multiplication of coordinates by a scalar is not. Here, we have no addition, no multiplication, no nice formula. We have algorithms only, but they turn out to work in the best time.

The result of these considerations is that the directions, **north**, **south**, **east** and **west** which play a nice role in the Euclidean case no longer exist. In fact, we have infinitely many directions, each of which defines an essential direction in the space: if you follow other directions, you will never go to the area covered by this one. Of course, an infinite amount of information is ruled out in computer science. And so, we replace this basic indetermination of the direction by the direction of the **father**. Of course, we are led to a root and a central cell, but nobody complains about using an origin in the Euclidean case. Moreover, as shown in [12,15], it is also possible in the case of tessellations of the hyperbolic plane to get rid of the origin. We just mention this point, here, and refer the interested reader to the quoted papers for a closer study.

Once again, we illustrate how we proceed by the case of the pentagrid. It is repeated in the case of the ternary heptagrid, see [12,14] and in the case of the 3D tiling of Poincaré's dodecahedron, see [13].

For the implementation, we first fix a basis of splitting and the representation of the tiling. As indicated in [7,15], there are a lot of choices with the same basis of splitting. Moreover, in the case of the pentagrid and of the ternary heptagrid in which the standard Fibonacci tree is also a spanning tree, we have the choice between using the Fibonacci sequence, as we did in Sect. "The Locating Problem in Hyperbolic Tilings", or using the basis derived from the polynomial of the splitting. The difference is that the Fibonacci sequence is defined by the golden mean $(1+\sqrt{5})/2$, while the sequence of the splitting is defined by the square of the golden mean, $(3+\sqrt{5})/2$.

Let us go on with the Fibonacci sequence, as is it used in the majority of papers.

The preferred son property allows us to compute very easily the coordinates of the neighbors of a cell ν from the coordinate c of ν: the computation is linear in time with respect to the length of c, see [10,15]. Similarly, the path from a cell c to the root of its tree can be computed in a linear time with respect to the length of c, again see [10,15]. As a simple example, if $c = \alpha_k \ldots \alpha_1\alpha_0$, the father of ν has for number $\mathcal{A}(\alpha_k..\alpha_2) + \alpha_1$, where $\mathcal{A}(\beta_h \ldots \beta_0)$ computes the number m whose coordinate is $\beta_\mathrm{h} \ldots \beta_0$.

What we indicated up to now fixes the coordinates for a cell whose supporting tile is in a given quarter. Now, it is enough to number the five initial quarters which lie around the central pentagon in order to completely define the coordinates of a cell. The central pentagon has 0 as an unique coordinate. All other cells are defined by two numbers: (α, ν). The first number, α is in $\{1..5\}$ and defines the quarter. The second number, ν, defines the tile in the indicated quarter. Together with its coordinate, a cell is associated with other data: the status of its supporting tile, and the indication of which side is shared with its father. On one hand, note that the coordinate is a hardware feature: it is never known by the cell and it cannot: it does not have a bounded size. Note that this is the same for CA's in Euclidean spaces. On the other hand, the status of the supporting node can be known by the cell. As shown in [10], one can define rules for a cellular automaton to dispatch this information. As it is a finite information which can be provided by the hardware, we may assume that the cell knows it.

Complexity of Cellular Automata in Hyperbolic Spaces

Now, we are ready to give the results about the complexity classes of HCA's.

SAT and NP-complete Problems

In [18], HCA's are proved to be able to solve **SAT** in polynomial time. Historically, the possibility to solve **NP**-complete problems in the hyperbolic plane was first announced in [24]. Although the authors of [18] were not aware of paper [24], the latter paper does not involve cellular automata and does not provide a precise description of how **SAT** can be solved in the new frame. On the contrary, [18] describes an HCA which is able to solve the problem. In [18], the computation is estimated as quadratic. In fact it can be proved to be linear in the size of the input.

The solution for **SAT** is easy: it makes use of a Fibonacci tree, in which only two nodes are selected among the sons of a node. Each level represents the possible assignment of true and false values to the variable indexed by this level. The computation of all possible assignments until the level n, where n is the number of variables is triggered at initial time. Once it is reached, the information comes back to the root from the leaves of the tree, i. e. the nodes which are on the level n of the tree: each node computes the OR on the values of its left-hand side and right-hand side sons. Accordingly, the root gives **true** if and only if there is a branch from it to a leaf along which the value is always **true**.

From this, applying classical tools of the theory of complexity, we obtain that any **NP**-complete problem can be solved in polynomial time by an appropriate cellular automaton of the hyperbolic plane.

P = NP in the Hyperbolic Plane

From what we have seen previously, we have that the classical class **NP** is contained in the class of HCA's which work in polynomial time, denoted by $\mathbf{P}_h$. Now, it is also possible to define $\mathbf{NP}_h$ for HCA's, taking the classical definition of non-deterministic computations in polynomial time.

As shown in [6], it turns out that $\mathbf{P}_h = \mathbf{NP}_h$. The key point is that the computation of a non-deterministic Turing machine in time $O(t(n))$, with $t(n) \geq n$, can be computed by a deterministic HCA in time $O(t^2(n))$.

From this theorem, the following surprising result can easily be derived, see [6]:

$$\mathbf{P}_h = \mathbf{NP}_h = \mathbf{PSPACE}$$

where **PSPACE** is the classical class of functions computed in polynomial space by a Turing machine.

Of course, in these results, a basic ingredient is the possibility, given by the hyperbolic plane, to occupy a working space of exponential area within a polynomial time. The above process for solving **SAT** is a basic example of such a possibility.

Other Parts of the Complexity Hierarchy of HCA's

In fact, if we look at the hierarchy of complexity classes for HCA's, we get a landscape which is very different from the classical situation.

We have the following situation, described in [5]:

$$\begin{aligned}\mathbf{DLOG}_h = \mathbf{NLOG}_h = \mathbf{P}_h = \mathbf{NP}_h = \mathbf{PSPACE}\\ \subsetneq \mathbf{PSPACE}_h = \mathbf{EXPTIME}_h = \mathbf{NEXPTIME}_h\\ = \mathbf{EXPSPACE}\end{aligned}$$

We notice that compared to the Euclidean analogs, the hyperbolic hierarchy seems to be very flat. As, by construction, $\mathbf{P}_h \subsetneq \mathbf{EXPTIME}_h$, there are indeed two classes on which the hierarchy concentrates.

We also have $\mathbf{NP}_h \subsetneq \mathbf{AP}_h$, unless $\mathbf{PSPACE} = \mathbf{NEXPTIME}$, where $\mathbf{AP}_h$ denotes the class of alternate HCA's. As with classical machines, an alternate HCA is defined on the set of configurations of a non-deterministic HCA. In the tree of these configurations, certain nodes are called existential, others are called universal. At an existential node, the node is accepting if and only if it has at least one accepting child. At a universal node, the node is accepting if and only if all its children are accepting. The result about $\mathbf{AP}_h$ indicates a similar situation with the Euclidean classes where $\mathbf{P} \subsetneq \mathbf{AP}$, unless $\mathbf{P} = \mathbf{PSPACE}$. Accordingly, we may expect that alternating HCA's should be more powerful than HCA's, either deterministic or non-deterministic.

On Specific Problems of Cellular Automata

Synchronization of an HCA

Although no paper is especially devoted to this problem, we mention it because it has an analog to the standard problem of the firing squad in one-dimensional CA's, and we shall use it in the next section.

In fact, as mentioned more or less explicitly in papers devoted to HCA, see [5,16], for instance, it is very easy to synchronize a disc or a sector inside a disc, defined by a tree rooted at the center of the disc. The idea is simply to simulate any classical algorithm of synchronization of a one-dimensional CA on each branch of the tree or each radius of the disc.

The synchronization is linear in the radius of the disc or the height of the tree.

Communications between HCA's

Another problem, more specific to HCA's is the communication between HCA's, possibly distant ones. Two papers study the problem in different settings, see [12] and [14].

In [12] the question is: how to establish a contact between two cells of a HCA, possibly distant ones? The paper provides a solution based on a new system of coordinates in which there is not an origin. The new system is based on the possibility to represent the hyperbolic plane as a union of growing quarters. We fix such a sequence in an appropriate way. Each term of the sequence is a Fibonacci tree, indexed by an integer n, and it contains all the trees indexed by m when $m > n$. Inside a given Fibonacci tree, we use the standard system of coordinates, indicated in Sect. "The Locating Problem in Hyperbolic Tilings". In the construction, the roots of the mentioned trees belong to a line δ. It is not difficult to see that sending signals on δ makes it possible for the cell to establish a contact in a linear time with respect to their mutual distance. Once the signals of each cell reach δ, they send in both directions a new signal at speed 1/2. A pair of such signals meet, which triggers both the communication signal and a killing signal, at speed 1, to erase the now useless signals sent in the wrong directions.

In [14], another problem is considered. This time all cells may dispatch messages, and each cell forwards the messages it receives and to which it does not want to reply. Accordingly, the same cell may be an emitter of messages, a receiver of messages, and a relay in the message system. The idea is to use the tree property to be in bijection with the tiling as follows: each emitting cell considers that it is the center of the hyperbolic plane, and the message is accompanied by an address which is updated by the relays and which is the address in the tree whose root is the sender of the message. This allows any receiver willing to answer the message to send it to the right emitter. Again, the complexity of the computation is linear in the mutual distance of a sender and a receiver.

Universality in Cellular Automata in Hyperbolic Spaces

Of course, from the existence of universal cellular automata on the line, we conclude that there are universal HCA's. This means that there are HCA's which are able to simulate any universal device, a Turing machine for instance.

Paradoxically, the study of universal HCA's has very few results. There is a universal HCA in the hyperbolic plane with 9 states, see [22], a recent result improving [4] where the HCA had 22 states. There is a universal HCA in the hyperbolic 3D space with 5 states, see [13]. At last, there is an intrinsically universal HCA in the hyperbolic plane, see [16].

Universal HCA's with a Small Number of States

First, we have to notice that the just mentioned universal HCA's with a small number of states are in fact **weakly** universal HCA's. The term **weak** refers to two conditions:

- the HCA needs an infinite initial configuration;
- the initial configuration is ultimately periodic.

Note that these conditions are standardly used with ordinary CA's where universality with a small number of states is reached with such conditions in many cases.

The second condition requires some explanation. In the context of a hyperbolic space, the notion of periodicity is not as clear as it is in the Euclidean case. Accordingly, we mean, by ultimate periodicity that at a large, *i. e.* outside a big enough domain, the configuration is globally invariant under a shift.

The above mentioned universal HCA's with a small number of states are obtained by a similar construction. They both simulate a railway circuit with the kind of switches, described by [26]. While in [26] a Turing machine is simulated, in [4,22] and in [13], we simulate a register machine. It can be remarked that the bigger number of states in the hyperbolic plane is due to the management of **crossings** of tracks which can be simply replaced by **bridges** in the hyperbolic 3D space. Moreover, as a cell in the hyperbolic 3D space has 12 neighbors, there are much more combinations of states which can be used to differentiate the relevant steps of the computation.

An Intrinsically Universal HCA

The intrinsically universal HCA is required to simulate any HCA in the same space. Of course, both the simulating HCA and the simulated one are required to work starting from finite configurations only.

In [16], two ingredients are used to achieve the simulation. One ingredient is the synchronization algorithm mentioned in Sect. "On Specific Problems of Cellular Automata". The second is the construction of **scaled** trees. The construction consists of building a new Fibonacci tree inside the tiling, but with a constant distance k between two consecutive nodes on a same branch. It is not difficult to construct such a tree, which is illustrated by Fig. 3.

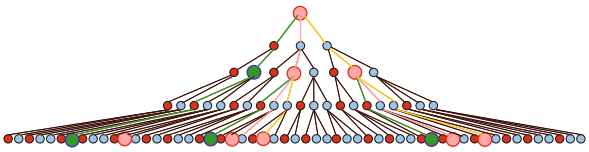

Cellular Automata in Hyperbolic Spaces, Figure 3
A scaled tree by a factor 2

The constant k is computed in such a way that a disc of radius k contains both an encoding of the initial configuration of the HCA to be simulated, say A, and an encoding of the transition table of A. Figure 4 illustrates the mechanism of propagation of the scaled tree.

Each step of the simulated HCA A is simulated by a **cycle** of steps of the simulating HCA U. The number of steps of U in a cycle is not constant. It may be increasing, especially if the simulated configuration is growing during its own computation. The synchronization algorithm of Sect. "On Specific Problems of Cellular Automata" is used to delimit the stages into which a cycle is split. These stages are the reception of the current states of the neighbors of the simulated cell of A, for each simulating cell of U. When this is achieved, possibly at different times for each simulating cell, the new state is determined and it is installed in the appropriate region, controlled by the simulating cell. When this is performed, the cell waits until it is informed by its simulating sons that their step of computation is completed. When this is the case, the cell informs its father in the scaled tree that it finished its computation. Accordingly, when the central cell receives the message of completion from all its neighbors of the scaled tree, it knows that the computation of this step of A is finished. Then the comparison with the previous configuration is performed, thanks to a synchronization. Depending on the result of the comparison, the computation is stopped if there was no difference, or it goes on, when a difference was noticed.

The Connection with Tiling Problems

As usual, cellular automata have deep connections with tilings.

This is probably the case with HCA's, although, up to now, the single connection is the possibility to implement them in the tilings, thanks to the coordinate system.

However, this system itself appeared to be useful in order to investigate the properties of tilings in the hyperbolic plane and in the hyperbolic spaces of higher dimensions, namely the dimensions 3 and 4.

Indeed, the splitting method could be applied to the tiling $\{5, 3, 4\}$ of the hyperbolic 3D space, see [15,21]. It

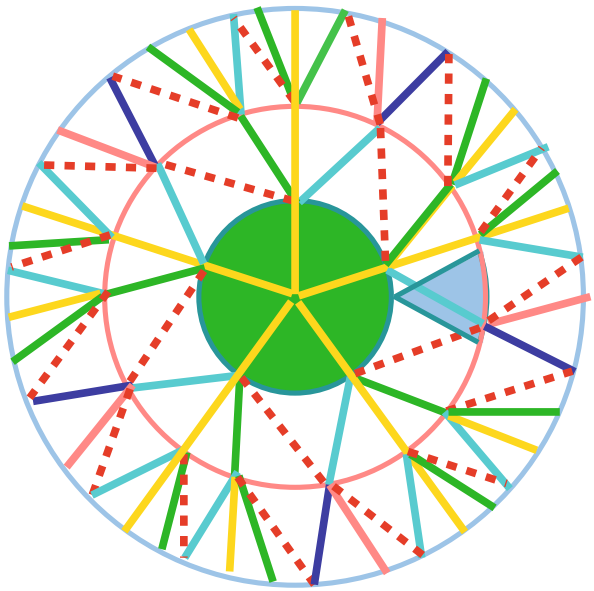

Cellular Automata in Hyperbolic Spaces, Figure 4
Propagation of a scaled tree

turned out to be possible to use an old tool of the late 19th century, Schlegel diagrams, to both represent the tiles and the construction of the tiling as a process which is infinite in time. The application of the splitting method revealed an interesting property. The language of the splitting of this tiling provides us with a **natural** example of a language which is neither rational nor context free. As a corollary, the algorithm to compute the path from a tile to the root of its tree is cubic in time with respect to the size of the coordinate of the cell supported by the tile.

The splitting method could also be applied to the tiling $\{5, 3, 3, 4\}$ of the hyperbolic 4D space. It provides us with a simple system of coordinates to explore this tiling which is the natural extension of the tiling $\{5, 3, 4\}$ of the hyperbolic 3D space. Note that the same process which allows to go from the pentagon with right angles to Poicaré's dodecahedron also allows to go from that dodecahedron to the 120-cell. This process is called **orthogonal completion** in [11,15]. Together with an appropriate notion of interior and exterior, it allows to get a correct orientation in the hyperbolic 4D space and to correctly use the dimensional analogy with the spaces of lower dimension.

There are also important applications to tilings of the hyperbolic plane. As already mentioned about the communication between cells of a HCA, these algorithms benefit from the study of the tilings which are made possible by using the splitting method. The new results which are obtained in an algorithmic way are indicated in [15], in particular in its Chapter 4. An important application to the

study of tilings in the hyperbolic plane is given by the recent result given in [17]: there, it is proved that the tiling problem of the hyperbolic plane is undecidable.

Future Directions

There are still a lot of problems to investigate about cellular automata in hyperbolic spaces. Let us look at a few of them.

In the classical study of cellular automata in an Euclidean space, there are well known theorems about the characterization of cellular automata in terms of mappings over the space of configurations. There are also well known studies about properties of the global transition function, which transforms a given configuration into the next one which is obtained, for each cell, by application of the transition table of the automaton. What can be said in the hyperbolic spaces? Are these theorems, or their adaptions, still valid in this new setting?

Another direction is a closer study of HCA in the hyperbolic 4D space. The foundation is ready: it is enough, now, to pave the way.

In the very beginning of this article, we mentioned possible applications. We think that this is also a promising direction for future investigations. A few applications were recently performed, see [1,23], for instance, but nothing significant enough to enter this article. We think that what was said in Sect. "Communications Between HCA's" is a serious foundation for future developments.

At last, the connection with tilings gives us the hope that the method initiated by the implementation of cellular automata in hyperbolic spaces will still help to improve the study of tilings in hyperbolic spaces.

Acknowledgments

The author is particularly in debt to Andrew Adamatzky for giving him the task to write this article.

Bibliography

Primary Literature

1. Chelghoum K, Margenstern M, Martin B, Pecci I (2004) Celluladr automata in the hyperbolic plane: proposal for a new environment. In: Proceedings of ACRI'2004, Amsterdam, 25–27 October 2004. Lecture Notes in Computer Sciences, vol 3305. Springer, Berlin, pp 678–687
2. Fraenkel AS (1985) Systems of numerations. Amer Math Mon 92:105–114
3. Gromov M (1981) Groups of polynomial growth and expanding maps. Publ Math IHES 53:53–73
4. Herrmann F, Margenstern M (2003) A universal cellular automaton in the hyperbolic plane. Theor Comput Sci 296:327–364
5. Iwamoto C, Margenstern M (2003) A survey on the Complexity Classes in Hyperbolic Cellular Automata. In: Proceedings of SCI'2003, V, pp 31–35
6. Iwamoto C, Margenstern M, Morita K, Worsch T (2002) Polynomial-time cellular automata in the hyperbolic plane accept accept exactly the PSPACE Languages. SCI'2002, Orlando, pp 411–416
7. Margenstern M (2000) New tools for cellular automata in the hyperbolic plane. J Univers Comput Sci 6(12):1226–1252
8. Margenstern M (2002) A contribution of computer science to the combinatorial approach to hyperbolic geometry. In: SCI'2002, Orlando, USA, 14–19 July 2002
9. Margenstern M (2002) Revisiting Poincaré's theorem with the splitting method. In: Bolyai'200, International Conference on Geometry and Topology, Cluj-Napoca, Romania, 1–3 October 2002
10. Margenstern M (2003) Implementing Cellular Automata on the Triangular Grids of the Hyperbolic Plane for New Simulation Tools. ASTC'2003, Orlando, 29 March–4 April
11. Margenstern M (2004) The Tiling of the Hyperbolic 4D Space by the 120-cell is Combinatoric. J Univers Comput Sci 10(9):1212–1238
12. Margenstern M (2006) A new way to implement cellular automata on the penta- and heptagrids. J Cell Autom 1(1): 1–24
13. Margenstern M (2007) A universal cellular automaton with five states in the 3D hyperbolic space. J Cell Autom 1(4):317–351
14. Margenstern M (2007) On the communication between cells of a cellular automaton on the penta- and heptagrids of the hyperbolic plane. J Cell Autom 1(3):213–232
15. Margenstern M (2007) Cellular Automata in Hyperbolic Spaces, vol 1: Theory. Old City Publishing, Philadelphia, p 422
16. Margenstern M (2008) A Uniform and Intrinsic Proof that there are Universal Cellular Automata in Hyperbolic Spaces. J Cell Autom 3(2):157-180
17. Margenstern M (2008) The domino problem of thehyperbolic plane is undecidable. Theor Comput Sci (to appear)
18. Margenstern M, Morita K (1999) A Polynomial Solution for 3-SAT in the Space of Cellular Automata in the Hyperbolic Plane. J Univers Comput Syst 5:563–573
19. Margenstern M, Morita K (2001) NP problems are tractable in the space of cellular automata in the hyperbolic plane. Theor Comput Sci 259:99–128
20. Margenstern M, Skordev G (2003) The tilings $\{p, q\}$ of the hyperbolic plane are combinatoric. In: SCI'2003, V, pp 42–46
21. Margenstern M, Skordev G (2003) Tools for devising cellular automata in the hyperbolic 3D space. Fundamenta Informaticae 58(2):369–398
22. Margenstern M, Song Y (2008) A new universal cellular automaton on the pentagrid. AUTOMATA'2008, Bristol, UK, 12–14 June 2008
23. Martin B (2005) VirHKey: a VIRtual Hyperbolic KEYboard with gesture interaction and visual feedback for mobile devices. In: MobileHCI'05, September, Salzburg, Austria
24. Morgenstein D, Kreinovich V (1995) Which algorithms are feasible and which are not depends on the geometry of space-time. Geombinatorics 4(3):80–97
25. Róka Z (1994) One-way cellular automata on Cayley Graphs. Theor Comput Sci 132:259–290
26. Stewart I (1994) A Subway Named Turing. Math Recreat Sci Am 90–92

Books and Reviews

Alekseevskij DV, Vinberg EB, Solodovnikov AS (1993) Geometry of spaces of constant curvature. In: Vinberg EB (ed) Geometry II, Encyclopedia of Mathematical Sciences, vol 29. Springer, Berlin

Berlekamp ER, Conway JH, Guy RK (1982) Winning Ways for Your Mathematical Plays. Academic Press

Bonola R (1912) Non-Euclidean Geometry. Open Court Publishing Company (also, (1955) Dover, New York)

Codd EF (1968) Cellular Automata. Academic Press, New York

Coxeter HSM (1969) Introduction to Geometry. Wiley, New York

Coxeter HSM (1974) Regular Complex Polytopes. Cambridge University Press, Cambridge

Delorme M, Mazoyer J (eds) (1999) Cellular automata, a parallel model. Kluwer, p 460

Epstein DBA, Cannon JW, Holt DF, Levi SVF, Paterson MS, Thurston WP (1992) Word Processing in Groups. Jones and Barlett, Boston

Grünbaum B, Shephard GS (1987) Tilings and Patterns. Freeman, New York

Gruska J (1997) Foundations of computing. International Thomson Computer Press

Knuth DE (1998) The Art of Computer Programming, vol II: Seminumerical algorithms. Addison-Wesley

Meschkowski H (1964) Noneuclidean Geometry. Translated by Shenitzer A. Academic Press, New York

Millman RS, Parker GD (1981) Geometry, a metric approach with models. Springer

Minsky ML (1967) Computation: Finite and Infinite Machines. Prentice-Hall, Englewood Cliffs

Ramsay A, Richtmyer RD (1995) Introduction to Hyperbolic Geometry. Springer

Toffoli T, Margolus N (1987) Cellular automata machines. MIT Press, Cambridge

von Neuman J (1966) Theory of self-reproducing automata. Edited and completed by A.W. Burks. The University of Illinois Press, Urbana

Wolfram S (1994) Cellular Automata and Complexity. Addison-Wesley

Wolfram S (2002) A New Kind of Science. Wolfram Media

Cellular Automata and Language Theory

Martin Kutrib
Institut für Informatik, Universität Giessen,
Giessen, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Cellular Language Acceptors
Tools and Techniques
Computational Capacities
Closure Properties
Decidability Problems
Future Directions
Bibliography

Glossary

Cellular automaton A (one-dimensional) cellular automaton is a linear array of cells which are connected to both of their nearest neighbors. The total number of cells in the array is determined by the input data. They are exactly in one of a finite number of states, which is changed according to local rules depending on the current state of a cell itself and the current states of its neighbors. The state changes take place simultaneously at discrete time steps. The input mode for cellular automata is called parallel. One can suppose that all cells fetch their input symbol during a pre-initial step.

Iterative array Basically, iterative arrays are cellular automata whose leftmost cell is distinguished. This so-called communication cell is connected to the input supply and fetches the input sequentially. The cells are initially empty, that is, in a special quiescent state.

Formal language The data on which the devices operate are strings built from input symbols of a finite set or alphabet. A subset of strings over a given alphabet is a formal language.

Signal Signals are used to transmit and encode information in cellular automata. If a cell changes to the state of its neighbor after some k time steps, and if subsequently its neighbors and their neighbors do the same, then the basic signal moves with speed $\frac{1}{k}$ through the array. With the help of auxiliary signals, rather complex signals can be established.

Closure property Closure properties of families of formal languages indicate their robustness under certain operations. A family of formal languages is closed under some operation, if any application of the operation on languages from the family yields again a language from the family.

Turing machine A Turing machine is the simplest form of a universal computer. It captures the idea of an effective procedure or algorithm. At any time the machine is in any one of a finite number of states. It is equipped with an infinite tape divided into cells and a read-write head scanning a single cell. Each cell may contain a symbol from a finite set or alphabet. Initially, the finite input is written in successive cells. All other cells are empty. Dependent on a list of instructions, which serve as the program for the machine, the action is determined completely by the current state and the symbol currently scanned by the head. The action

comprises the symbol to be written on the current cell, the new state of the machine, and the information of whether the head should move left or right.

Decidability A formal problem with two alternatives is decidable, if there is an algorithm or a Turing machine that solves it and halts on all inputs. That is, given an encoding of some instance of the problem, the algorithm or Turing machine returns the correct answer *yes* or *no*. The problem is semidecidable, if the algorithm halts on all instances for which the answer is *yes*.

Definition of the Subject

One of the cornerstones in the theory of automata is the early result of John von Neumann, who solved the logical problem of nontrivial self-reproduction. He employed a mathematical device which is a multitude of interconnected identical finite-state machines operating in parallel to form a larger machine. He showed that it is logically possible for such a nontrivial computing device to replicate itself ad infinitum [97]. Such devices are commonly called cellular automata (abbreviated, CA), and can be considered as homogeneously structured models for massively parallel computing systems. The global behavior of cellular automata is achieved by local interactions only. While the underlying rules are quite simple, the global behavior may be rather complex. In general, it is unpredictable.

The data supplied to CAs can be arranged as strings of symbols. Instances of problems to solve can be encoded as strings with a finite number of different symbols. Furthermore, complex answers to problems can be encoded as binary sequences such that the answer is computed bit by bit. In order to compute one piece of the answer, the set of possible inputs is split into two sets associated with the binary outcome. From this point of view, the computational capabilities of CAs are studied in terms of string acceptance, that is, the determination to which of the two sets a given string belongs. These investigations are with respect to and with the methods of language theory. They originated in Stephen N. Cole [18,19] and Alvy R. Smith [76,80]. Over the years substantial progress has been achieved, but there are still some basic open problems with deep relations to other fields. So, exploring the capabilities of cellular automata may benefit the understanding of the nature of parallelism and nondeterminism.

Introduction

In general, the specification of a cellular automaton includes the type and specification of the cells, their interconnection scheme (which can imply a dimension of the system), the local rules which are formalized as local transition function, and the input and output modes. With an eye towards language acceptance, we consider one-dimensional synchronous devices with nearest neighbor connections whose cells are deterministic finite-state machines. They are commonly called cellular automata in case of parallel input mode, and iterative arrays (abbreviated, IA) if the input mode is sequential. If each cell is connected to only one of its neighbors, say to the right one, then the flow of information through the array is from right to left. The corresponding device is a one-way cellular automaton (abbreviated, OCA). In any case the number of cells is determined by the length of the input string; there is one cell per symbol. If the input is in parallel, then all cells fetch their input symbol during a pre-initial step. To this end, the set of symbols has to be a subset of the set of states. Sometimes for practical reasons and for the design of systolic algorithms, a sequential input mode is more convenient than the parallel one. In iterative arrays the leftmost cell is distinguished to be the communication cell. It is equipped with a one-way read-only input tape.

In order to obtain the binary answer of a system we have to overcome a problem with the end of the computation. It follows from the definitions that the machines never halt. A way to cope with the situation is to define a predicate on configurations. The answer depends on whether a configuration satisfies the predicate or not. Here we apply the common predicate that requires a border cell or the communication cell to be in some state designated to be an accepting state. Further predicates are studied, e. g., in [38,82], while more general input modes are considered in [57]. Due to the bounded number of cells in a computation, the time complexity is exponentially bounded. After exceeding the bound, the computation runs into a loop and is rather useless. With respect to the wide language classes obeying an exponential or polynomial time bound, parallel devices cannot take advantage of their large number of processing elements. They are just a factor to be multiplied with the size of the input. So, there is a particular interest in fast computations, that is, in real-time and linear-time computations. Real time is determined by the shortest time necessary for nontrivial computations, whereas linear time is real time multiplied by an arbitrary but fixed constant greater than or equal to one. In addition, we consider general computations without time bounds which, actually, are exponentially time bounded. The following well-known example from [17] joins several notions. It uses signals in order to construct the mapping $n \longmapsto 2^n$ in time; i. e., the leftmost cell recognizes the time steps 2^n, $n \geq 1$. The time constructor is then extended to a real-time CA and IA.

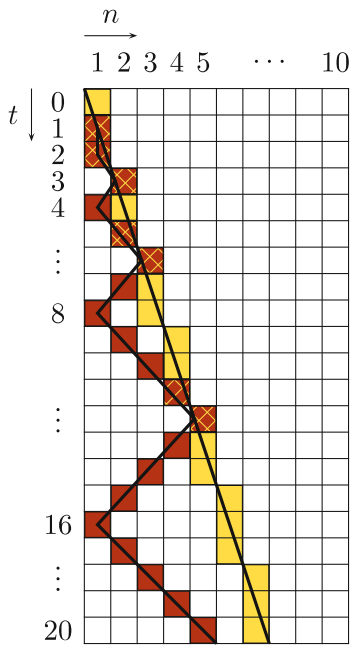

Cellular Automata and Language Theory, Figure 1
Space-time diagram showing signals of a real-time two-way acceptor of the language $\{a^{2^n} \mid n \geq 1\}$

Example 1 The unary language consisting of the strings $\{a^{2^n} \mid n \geq 1\}$ is accepted by IAs as well as by CAs in real time.

At initial time the leftmost cell emits a signal which moves with speed $\frac{1}{3}$ to the right; i. e., the signal alternates from moving one cell to the right and staying for two time steps in a cell (see Fig. 1). In addition, another signal is emitted which moves with maximal speed (speed 1). It bounces between the slow signal and the leftmost cell. It is easy to see that the signal passes through the leftmost cell exactly at the time steps 2^n, $n \geq 1$. Finally, an iterative array can accept its input when the last input symbol is read at one of these time steps. Similarly, in a CA computation the rightmost cell can emit a third signal which moves with maximal speed to the left. This signal arrives at the leftmost cell at the time step which corresponds to the length of the input. If it meets the bouncing signal at its arrival, the input is accepted. □

More results about mappings that are constructible in the above sense can be found in [13,63,93]. In [27,94] the series of prime numbers is constructed in real-time devices.

The investigations of iterative arrays and cellular automata as cellular language acceptors originated in [18,19] and [76,80]. In [18,19], it is shown that the family of languages accepted by real-time IAs is closed under intersection, union, and complementation, but is not closed under concatenation and reversal; and in [76,80], where among other results the identity of the sequential complexity class DSPACE(n) (i. e., the class of languages accepted by deterministic Turing machines whose tape is bounded by the length of the input n) and the family of languages accepted by CAs without time bound is shown. A long-standing open problem is the question whether or not one-way information flow is a strict weakening of two-way information flow for unbounded time. In [16,32] it is proved that a PSPACE-complete language is accepted by OCAs, from which one can draw conclusions about the hardness of sequential OCA simulations. Furthermore, in the same papers strong closure properties are derived for the family of OCA languages. In addition, it is a proper superset of the context-free languages which, in turn, are of great practical relevance.

The proofs are based on characterizations of the parallel language families by certain types of customized sequential machines. Such machines have been developed for all classes of acceptors which are here under consideration [32,36,37]. In particular, speed-up theorems are given that allow to speed up the time beyond real time linearly. Therefore, linear-time computations can be sped up close to real time. Nevertheless, for OCAs and IAs linear time is strictly more powerful than real time. The problem is still open for CAs. In fact, it is an open question whether real-time CAs are strictly weaker than unbounded time CAs. If both classes coincide, then a PSPACE-complete language would be accepted in polynomial time! Apart from that it is known that linear-time CAs can be simulated by unbounded time OCAs [16,32].

The rest of the article is organized as follows. In the following section (Sect. "Cellular Language Acceptors"), some basic notions and formal definitions are given. The problem in connection with the end of computations is discussed in more detail, and honest time complexities are derived informally. Then, in Sect. "Tools and Techniques" selected tools and techniques are presented that can be applied to prove or to disprove that certain languages are accepted by certain devices. In particular, it is shown that two-way devices can simulate the data structure *stack* without loss of time. It is often much harder to disprove the acceptance of languages than to prove it, since the technique of a suitable construction is trivially not applicable. Some techniques based on counting and pumping arguments are shown and applied in order to obtain witness languages. In Sect. "Computational Capacities" computational capacity aspects are investigated. A basic hierarchy of language families defined by cellular automata and iterative arrays is established. The levels

are compared with well-known linguistic families. Next, Sect. "Closure Properties" is devoted to exploring the closure properties of the language families in question. For example, it turns out that the languages accepted by unbounded time one-way and two-way cellular automata as well as iterative arrays share the strong closure properties of the class DSPACE(n) which characterizes the deterministic context-sensitive languages. Decidability problems are considered in Sect. "Decidability Problems". By reductions of Turing machine problems, it follows that almost all interesting properties are undecidable. In fact, they are not even semidecidable for the weakest devices in question. This emphasizes once more the power gained in parallelism even if the number of cells is bounded. In general, their behavior is unpredictable. Finally, in Sect. "Future Directions", some future directions of cellular language acceptors are discussed.

Cellular Language Acceptors

We denote the set of nonnegative integers by $\mathbb{N}$. In connection with formal languages, strings are called *words*. Let A^* denote the set of all words over a finite alphabet A. The *empty word* is denoted by λ, and we set $A^+ = A^* - \{\lambda\}$. For the *reversal of a word* w we write w^R and for its *length* we write $|w|$. For the *number of occurrences* of a symbol a in w we use the notation $|w|_a$. We use $\subseteq$ for *inclusions* and $\subset$ for *strict inclusions*. In order to avoid technical overloading in writing, two languages L and L' are considered to be equal, if they differ at most by the empty word, i. e., $L - \{\lambda\} = L' - \{\lambda\}$. Throughout the article, two automata or grammars are said to be *equivalent* if and only if they accept or generate the same language.

The cells of a cellular automaton are identified by positive integers. In order to handle the application of the local transition function to the border cells, we assume that the missing neighbors are in a permanent so-called border state. A formal definition is:

Definition 2 *A* two-way cellular automaton *(CA) is a system* $\langle S, \delta, \#, A, F\rangle$, *where*

1. S *is the finite, nonempty set of* cell states,
2. $\# \notin S$ *is the permanent* boundary state,
3. $A \subseteq S$ *is the nonempty set of* input symbols,
4. $F \subseteq S$ *is the set of* accepting states, *and*
5. $\delta : (S \cup \{\#\}) \times S \times (S \cup \{\#\}) \to S$ *is the* local transition function.

If the flow of information is restricted to one-way, the resulting device is a one-way cellular automaton (abbreviated, OCA). In such devices, the next state of each cell depends on the state of the cell itself and the state of its immediate neighbor to the right.

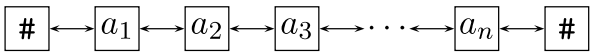

Cellular Automata and Language Theory, Figure 2
A (two-way) cellular automaton

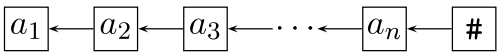

Cellular Automata and Language Theory, Figure 3
A one-way cellular automaton

A *configuration* of a cellular automaton $\langle S, \delta, \#, A, F\rangle$ at time $t \geq 0$ is a description of its global state, which is formally a mapping $c_t\{1, \ldots, n\} \to S$, for $n \geq 1$. The configuration at time 0 is defined by the given input $w = a_1 \cdots a_n \in A^+$. We set $c_0(i) = a_i$, for $1 \leq i \leq n$. Configurations may be represented as words over the set of cell states in their natural ordering. For example, the initial configuration for w is represented by $\#a_1\, a_2 \cdots a_n\#$. Successor configurations are computed according to the global transition function Δ. Let c_t, $t \geq 0$, be a configuration with $n \geq 2$, then its successor c_{t+1} is defined as follows:

$$c_{t+1} = \Delta(c_t) \iff \begin{cases} c_{t+1}(1) = \delta(\#, c_t(1), c_t(2)) \\ c_{t+1}(i) = \delta(c_t(i-1), c_t(i), c_t(i+1)), \\ \quad i \in \{2, \ldots, n-1\} \\ c_{t+1}(n) = \delta(c_t(n-1), c_t(n), \#) \end{cases}$$

for CAs, and

$$c_{t+1} = \Delta(c_t) \iff \begin{cases} c_{t+1}(i) = \delta(c_t(i), c_t(i+1)), \\ \quad i \in \{1, \ldots, n-1\} \\ c_{t+1}(n) = \delta(c_t(n), \#) \end{cases}$$

for OCAs. For $n = 1$, the next state of the sole cell is $\delta(\#, c_t(1), \#)$. Thus, Δ is induced by δ.

A computation can be represented as a space-time diagram, where each row is a configuration and the rows appear in chronological ordering.

In order to define iterative arrays formally we have to provide an initial (quiescent) state for the cells. We assume that once the whole input is consumed an end-of-input symbol is supplied permanently.

Definition 3 *An* iterative array *(IA) is a system* $\langle S, \delta, \delta_0, s_0, \#, \triangleleft, A, F\rangle$, *where*

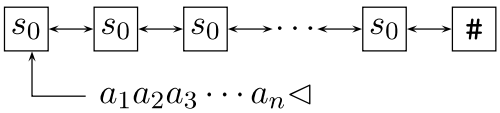

Cellular Automata and Language Theory, Figure 4
An iterative array

1. *S is the finite, nonempty set of* cell states,
2. $s_0 \in S$ *is the* quiescent state,
3. $\# \notin S$ *is the permanent* boundary state,
4. $\triangleleft \notin A$ *is the* end-of-input symbol,
5. *A is the finite, nonempty set of* input symbols,
6. $F \subseteq S$ *is the set of* accepting states,
7. $\delta : S \times S \times (S \cup \{\#\}) \to S$ *is the* local transition function for non-communication cells *satisfying* $\delta(s_0, s_0, s_0) = \delta(s_0, s_0, \#) = s_0$,
8. $\delta_0 : (A \cup \{\triangleleft\}) \times S \times (S \cup \{\#\}) \to S$ *is the* local transition function for the communication cell.

A configuration of an iterative array $\langle S, \delta, \delta_0, s_0, \#, \triangleleft, A, F \rangle$ at time $t \geq 0$ is a pair (w_t, c_t), where $w_t \in A^*$ is the remaining input sequence and $c_t : \{1, \ldots, n\} \to S$, for $n \geq 1$, is a mapping that maps the single cells to their current states. The configuration (w_0, c_0) at time 0 is defined by the input word w_0 and the mapping $c_0(i) = s_0$, $1 \leq i \leq n$. The global transition function Δ is induced by δ and δ_0 as follows: Let (w_t, c_t), $t \geq 0$, be a configuration and $i \in \{2, \ldots, n-1\}$. Then

$$(w_{t+1}, c_{t+1}) = \Delta((w_t, c_t)) \iff \begin{cases} c_{t+1}(1) = \delta_0(a, c_t(1), c_t(2)) \\ c_{t+1}(i) = \delta(c_t(i-1), c_t(i), c_t(i+1)) \\ c_{t+1}(n) = \delta(c_t(n-1), c_t(n), \#) \end{cases}$$

where $a = \triangleleft$, $w_{t+1} = \lambda$ if $w_t = \lambda$, and $a = a_1$, $w_t + 1 = a_2 \cdots a_n$ if $w_t = a_1 \cdots a_n$.

An input w is accepted by a CA, OCA, or an IA $\mathcal{M}$ if at some time i during its course of computation the leftmost cell enters an accepting state. The *language accepted by* $\mathcal{M}$ is denoted by $L(\mathcal{M})$. Let $t : \mathbb{N} \to \mathbb{N}$, $t(n) \geq n$ ($t(n) \geq n+1$ for IAs) be a mapping. If all $w \in L(\mathcal{M})$ are accepted with at most $t(|w|)$ time steps, then $L(\mathcal{M})$ is said to be of time complexity t.

Observe that time complexities do not have to meet any further conditions. This general treatment is made possible by the way of acceptance. An input w is accepted if the leftmost cell enters an accepting state at some time $i \leq t(|w|)$. But what if afterwards, a final configuration has been reached? Subsequent states of the leftmost cell are not relevant.

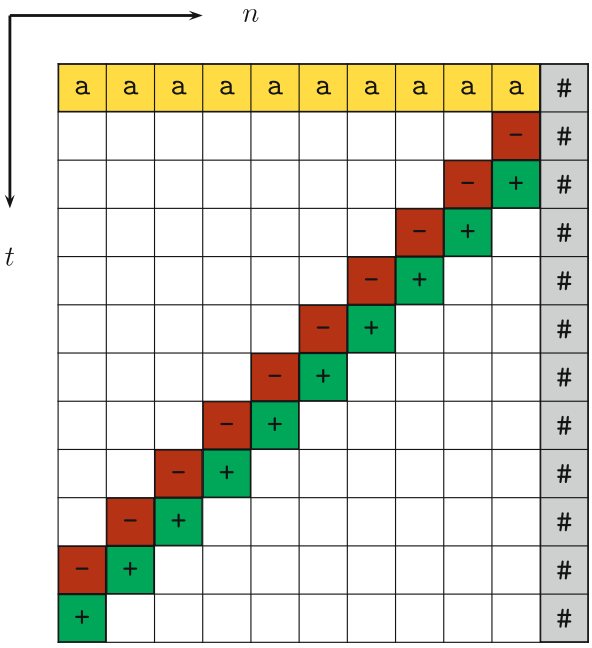

Cellular Automata and Language Theory, Figure 5
An OCA accepting any unary language. Here + is an accepting and — a non-accepting state

Following the different approach to gather the result of a computation at time step $t(|w|)$ by the outside world does not yield the desired outcome in general. In this case, the intrinsic computation may be hidden in the determination of the time step $t(|w|)$. That is, computational power may be added from the outside world. For example, let $L \in \{a\}^+$ be an arbitrary language. Then L is accepted by some OCA with time complexity

$$t(n) = \begin{cases} n & \text{if} \quad a^n \notin L \\ n+1 & \text{if} \quad a^n \in L \end{cases},$$

where the local transition function is easily designed to realize the behavior depicted in the space-time diagram of Fig. 5.

So, it is reasonable to consider only such time complexities t that allow the leftmost cell to recognize the time step $t(n)$. For example, the identity $t(n) = n$ is an honest time complexity for OCAs and CAs. A signal which is initially emitted by the rightmost cell and moves with maximal speed arrives at the leftmost cell exactly at time step n. By slowing down the signal to speed $\frac{x}{y}$, i. e., the signal alternating moves x cells to the left and stays for $y - x$ time steps in a cell, it is seen that the time complexities $\frac{x}{y} \cdot n$, for any positive integers x, y, are also honest. Another example are exponential time complexities $t(n) = k^n$, for any in-

teger $k \geq 2$. Without going too deep into technical details, a corresponding device can be set up as a $k$ary counter. The rightmost cell simulates the least significant digit and adds one to the counter at every time step. The neighboring cell to the left observes when a carry-over appears, increases its own digit and so on. Then, the leftmost cell produces a first carry-over exactly at time step k^n.

The family of languages that are accepted by IAs (and CAs, OCAs) with time complexity t is denoted by $\mathcal{L}_t(\mathrm{IA})$ (and $\mathcal{L}_t(\mathrm{CA})$, $\mathcal{L}_t(\mathrm{OCA})$, respectively). The index is omitted for arbitrary time. Actually, arbitrary time is exponential time due to the space bound. If t is the function $n+1$ (the function n), acceptance is said to be in *real time* and we write $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$ ($\mathcal{L}_{\mathrm{rt}}(\mathrm{CA})$, $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$). Since for nontrivial computations an IA has to read at least one end-of-input symbol, real time has to be defined as $(n+1)$-time. The *linear-time* languages $\mathcal{L}_{\mathrm{lt}}(\mathrm{IA})$ are defined according to $\mathcal{L}_{\mathrm{lt}}(\mathrm{IA}) = \bigcup_{k\in\mathbb{Q},k\geq 1} \mathcal{L}_{k\cdot n}(\mathrm{IA})$, and similarly for CAs and OCAs.

Tools and Techniques

An elementary technique in automata theory is the usage of multiple tracks. Basically, this means to consider the state set as Cartesian product of some smaller sets. Each component of a state is called a *register*, and the same register of all cells together forms a *track*.

The first goal of this section is to show how to simulate pushdown stores, i. e., stores obeying the principle *last in first out*, by IAs and CAs in real time. Assume without loss of generality that at most one symbol is pushed onto or popped from the stack at each time step. We distinguish one cell that simulates the top of the pushdown store. It suffices to use three additional tracks for the simulation. Let the three pushdown registers of each cell be numbered one, two and three from top to bottom. Each cell prefers to have only the first two registers filled. The third register is used as a buffer. In order to reach that charge it obeys the following rules (cf. Fig. 6).

a) If all three registers of its left (upper) neighbor are filled, it takes over the symbol from the third register of the neighbor and stores it in its first register. The old contents of the first and second registers are shifted to the second and third register.

b) If there is only the first register of its left (upper) neighbor filled, the cell erases its first register and shifts the contents of the second and third registers to the first and second register. Observe that the erased symbol is taken over by the left neighbor.

c) Possibly, more than one of these actions are superimposed.

From the simulation, it follows immediately that real-time IAs as well as real-time CAs accept all languages accepted by sequential pushdown automata as long as they work in real time.

Theorem 4 *Given some real-time deterministic pushdown automaton, an equivalent real-time IA and CA can effectively be constructed, i. e., every real-time deterministic context-free language belongs to the families $\mathcal{L}_{rt}(IA)$ and $\mathcal{L}_{rt}(CA)$.*

Now we turn to a technique for disproving that languages are accepted. In general, the method is based on equivalence classes which are induced by formal languages. If some language induces a number of equivalence classes which exceeds the number of classes distinguishable by a certain device, then the language is not accepted by that device. First we give the definition of an equivalence relation which applies to real-time IAs.

Definition 5 *Let $L \subseteq A^*$ be a language and $l \geq 1$ be a constant. Two words $w \in A^*$ and $w' \in A^*$ are l-equivalent with respect to L if and only if*

$$wu \in L \iff w'u \in L$$

for all $u \in A^*$, $|u| \leq l$. The number of l-equivalence classes with respect to L is denoted by $E(L, l)$.

In [19] the following upper bound for the number of equivalence classes distinguishable by real-time IA is derived.

Lemma 6 *If $L \in \mathcal{L}_{rt}(IA)$, then there exists a constant $p \geq 1$ such that $E(l, L) \leq p^l$.*

Proof Let $\mathcal{M}$ be a real-time IA with state set S. In order to determine an upper bound for the number of l-equivalence classes with respect to $L(\mathcal{M})$, we consider the possible configurations of $\mathcal{M}$ after reading all but l input symbols. The remaining computation depends on the last l input symbols and the states of the cells $1, \ldots, l+2$. For the $l+2$ states there are at most $|S|^{l+2}$ different possibilities. Setting $p = |S|^3$, we derive $|S|^{l+2} \leq |S|^{3l} = p^l$, and obtain at most p^l different possibilities. Since the number of equivalence classes is not affected by the last l input symbols, there are at most p^l equivalence classes. □

Example 7 The language

$$L = \{\&x_k\&\cdots\&x_1?y_1\&\cdots\&y_k\& \mid k \geq 1, x_i^R = y_i z_i \text{ and } x_i, y_i, z_i \in \{a, b\}^*\}$$

does not belong to $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$.

For a pair of different prefixes $w = \&x_k\&\cdots\&x_1?$ and $w' = \&x'_k\&\cdots\&x'_1?$ with $|x_i| = |x'_i| = k$, $1 \leq i \leq k$, there exists at least one $1 \leq j \leq k$ such that $x_j \neq x'_j$. This

Cellular Automata and Language Theory, Figure 6
Principle of a pushdown store simulation. Subfigures are in row-major order

implies $w\,\&^{j-1}\,x_j^R\,\&^{k-j+1} \in L$ and $w'\,\&^{j-1}\,x_j^R\,\&^{k-j+1} \notin L$. Since there are 2^{k^2} different prefixes of the given form, language L induces at least 2^{k^2} classes.

On the other hand, if L would be accepted by some real-time IA, then by Lemma 6 there is a constant $p \geq 1$ such that $E(L,2k) \leq p^{2k}$. Since L is infinite, we may choose k large enough such that $2^{k^2} > p^{2k}$, which is a contradiction. □

Now we change to real-time OCAs. The next result is a tool which allows us to show that languages do not belong to the family $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$. It is based on pumping arguments for cyclic strings [68].

Lemma 8 *Let L be a real-time OCA language. Then there exists a constant $p \geq 1$ such that any pair of a word w and an integer k that meets the condition $w^k \in L$ and $k > p^{|w|}$ implies that there is some $1 \leq q \leq p^{|w|}$ such that $w^{k+jq} \in L$, for all $j \geq 0$.*

Proof For a given real-time OCA $\mathcal{M} = \langle S, \delta, \#, A, F \rangle$, we set $p = |S|^2$. Let $w^k \in L(\mathcal{M})$, where $k > p^{|w|}$. Then we consider an accepting computation of $\mathcal{M}$ on input w^k. The initial configuration is represented by $\#w^k\#$. Clearly, a cyclic left part of some configuration leads again to a cyclic left part, though the new left part gets one cell shorter at any time step. Therefore, after $|w|$ time steps the left part of the configuration which still may influence the overall computation result is represented by $\#w_1^{k-1}s_1$, where $|w_1| = |w|$ and $s_1 \in S$. After another $|w|$ time steps, we obtain $\#w_2^{k-2}s_2$, where $|w_2| = |w|$ and $s_2 \in S$. In general, the relevant part of a configuration at time $i \cdot |w|$, $1 \leq i \leq k$, is represented by

$\# w_i^{k-i} s_i$, where $|w_i| = |w|$ and $s_i \in S$. In addition, state s_k is an accepting one.

Since the number of different words w_i is bounded by $|S|^{|w|}$, for $w_i s_i$ there are at most

$$|S|^{|w|+1} \leq |S|^{2\left\lceil \frac{|w|+1}{2} \right\rceil} \leq p^{|w|}$$

different possibilities. Now $k > p^{|w|}$ implies that $w_i s_i = w_l s_l$ for some $1 \leq i < l \leq k$. Therefore, there is a loop and, for $q = l - i$, the word w^{k+jq} is accepted, for any $j \geq 0$. □

Example 9 The language $L = \{a^{2^n} \mid n \geq 1\}$ is not accepted by any real-time OCA.

Contrarily assume there is a real-time OCA accepting L. Then we set $w = a, k = 2^p$ and observe that the conditions of Lemma 8 are met. Therefore, a^{2^p+q} as well as a^{2^p+2q} belong to L. If $2^p + q$ is not a power of two, we obtain a contradiction. So, let $2^p + q = 2^{p+r}$, for some $r \geq 1$. We derive $2^{p+r} < 2^{p+r} + q = 2^{p+r} + 2^{p+r} - 2^p = 2^{p+r+1} - 2^p < 2^{p+r+1}$, and conclude that $2^p + 2q$ is strictly in between two consecutive powers of two. Hence, a^{2^p+2q} does not belong to L. □

Next we come back to equivalence classes. By a similar idea as for iterative arrays an equivalence relation can be defined such that the number of equivalence classes distinguishable by real-time OCAs is bounded. But due to the nature of OCA computations, both the prefixes as well as the suffixes of inputs have to be regarded [84,85]. We continue with the equivalence relation:

Definition 10 *Let $L \subseteq A^*$ be a language and $X \subseteq A^*$ and $Y \subseteq A^*$ be two sets of words. Two words $w \in A^*$ and $w' \in A^*$ are (L, X, Y)-equivalent if and only if*

$$xwy \in L \iff xw'y \in L$$

for all $x \in X$ and $y \in Y$.

The next step is to derive an upper bound for the number of equivalence classes which can be distinguished by some real-time OCA.

Theorem 11 *Let $L \subseteq A^*$ be a real-time OCA language and $X = A^{m_1}$ and $Y = A^{m_2}$ be two sets of words for positive integers m_1 and m_2. Then there exists a constant $p \geq 1$ such that the number N of (L, X, Y)-equivalence classes is bounded by*

$$N \leq p^{|X|} p^{(m_2+1)|Y|} .$$

Proof For any word w and any $x \in X$ and $y \in Y$ the input xwy implies exactly one real-time OCA configuration at time $|w|$ (see Fig. 7). At this time only the leftmost $|xy| + 1$ cells can still influence the overall computation result, i. e., the cells $1, \ldots, |xy| + 1$. The cells $1, \ldots, |x|$ are not affected by the cells $|xw| + 1, \ldots, |xwy|$ up to time $|w|$. So, they are not affected by y up to that time. Furthermore, the cells $|x| + 1, \ldots, |xy| + 1$ are not affected by the cells $1, \ldots, |x|$, that is, not by x. On the other hand, different w may cause different configurations for x and y, where x and y are independent of each other.

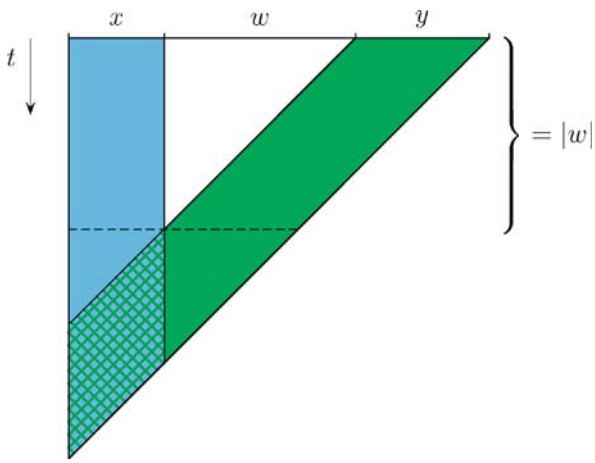

Cellular Automata and Language Theory, Figure 7
Dependencies in the proof of Theorem 11

It follows that w and w' are equivalent, if the cells $1, \ldots, |xy| + 1$ at time step $|w|$ respectively $|w'|$ are in the same states for all $x \in X$ and $y \in Y$. For any $y \in Y$, there are $|S|^{m_2} + 1$ different configurations for the cells $|x| + 1, \ldots, |xy| + 1$. So, there are altogether at most $|S|^{(m_2+1)|Y|}$ different possibilities to distinguish the words w.

The possibilities for the cells $1, \ldots, |x|$ are as follows. For a word $x = x_{m_1} \ldots x_1$ the state s_i of cells $1 \leq i \leq m_1$ at time $|w|$ depends only on $x_i, \ldots, x_1$ and w. So, for s_i there are at most $|S|^{|A|^i}$ different possibilities to distinguish the words w. Together we obtain

$$\prod_{i=1}^{m_1} |S|^{|A|^i} < |S|^{|A|^{m_1+1}}$$

possibilities. For $p = |S|^{|A|}$ this implies $|S|^{|A|^{m_1+1}} = p^{|A|^{m_1}} = p^{|X|}$. Thus, the number of equivalence classes is bounded by $p^{|X|} p^{(m_2+1)|Y|}$. □

As an example we show that the language $L_d \subset \{0, 1, (,), |\}^+$ whose words are of the form

$$x(x_1|y_1) \cdots (x_n|y_n)y ,$$

where $x, x_i, y, y_i \in \{0,1\}^*$, for $1 \leq i \leq n$, and $(x|y) = (x_i|y_i)$ for at least one $i \in \{1, \ldots, n\}$, is not a real-time OCA language.

It can be thought of as a dictionary. The task for the OCA is to check whether the pair $(x|y)$ occurs in the dictionary or not.

Example 12 Let $X = Y = \{0,1\}^*$. Two words $w = (x_1|y_1)\cdots(x_n|y_n)$ and $w' = (x'_1|y'_1)\cdots(x'_m|y'_m)$ are (L_d, X, Y)-equivalent if and only if $\{(x_1|y_1),\ldots,(x_n|y_n)\} = \{(x'_1|y'_1),\ldots,(x'_m|y'_m)\}$.

First assume that the two sets are equal. Let $x \in X$ and $y \in Y$, then $xwy \in L_d$ implies $(x|y) = (x_i|y_i)$, for some i. Since the two sets are equal, we have $(x|y) = (x'_j|y'_j)$, for some j. Therefore, $xwy \in L_d$ implies $xw'y \in L_d$ and vice versa, i. e., w and w' are (L_d, X, Y)-equivalent.

Now assume the two sets are different. Without loss of generality, we can assume that there exist $x \in X$ and $y \in Y$ with $(x|y) = (x_i|y_i)$, for some i, but $(x|y) \neq (x'_j|y'_j)$, for all $j = 1,\ldots,m$. Then $xwy \in L_d$ but $xw'y \notin L_d$ and, thus, w and w' are not (L_d,X,Y)-equivalent.

In order to derive a lower bound for the number of (L_d, X, Y) equivalence classes induced by L_d, let $m_1, m_2 \geq 1$, $X = \{0,1\}^{m_1}$ and $Y = \{0,1\}^{m_2}$. Then the number N of (L_d,X,Y)-equivalence classes is at least $2^{2^{m_1+m_2}}$.

For fixed $m_1 \geq 1$ and $m_2 \geq 1$, we consider all words of the form $(x_1|y_1)\cdots(x_k|y_k)$ with $x_i \in \{0,1\}^{m_1}$ and $y_i \in \{0,1\}^{m_2}$, for all $i \in \{1,\ldots,k\}$, and $(x_i|y_i) \neq (x_j|y_j)$, for $i \neq j$. These words are said to be of type (m_1,m_2). Following the argumentation above, two words are equivalent if and only if the sets of subwords are equal. There are $2^{2^{m_1+m_2}}$ words of type (m_1,m_2) which belong to different equivalence classes with respect to L_d, $X = \{0,1\}^{m_1}$ and $Y = \{0,1\}^{m_2}$.

Now assume that L_d is accepted by some real-time OCA. For all $m_1,m_2 \geq 1$ there is a constant p such that the number of equivalence classes is at most $p^{2^{m_1}}\, p^{(m_2+1)2^{m_2}}$. For large enough m_1 and m_2 we obtain a contradiction to the lower bound $2^{2^{m_1+m_2}}$. □

Helpful tools of a different nature are speed-up theorems. Strong results are obtained in [34,36], where the parallel language families are characterized by certain types of customized sequential machines. Such machines have been developed for all classes of acceptors which are here under consideration. In particular, speed-up theorems are given that allow to speed up the time beyond real time linearly. Therefore, linear-time computations can be sped up close to real time. The question whether or not real time is strictly weaker than linear time is discussed in detail later.

Theorem 13 *Let $\mathcal{M}$ be a CA, OCA, or IA obeying time complexity $rt + r(n)$, where $r\colon \mathbb{N} \to \mathbb{N}$ is a mapping and rt denotes real time. Then for all $k \geq 1$ an equivalent device $\mathcal{M}'$ of the same type obeying time complexity $rt + \lfloor \frac{r(n)}{k} \rfloor$ can effectively be constructed.*

The next two examples are frequently used applications of the linear speed-up theorem.

Example 14 Let $k_0 \geq 1$ and $\mathcal{M}$ be a device in question with time complexity $rt + k_0$. Then there is an equivalent real-time device $\mathcal{M}'$ of the same type. It suffices to set $k = k_0 + 1$ and to apply Theorem 13 in order to obtain $rt + \lfloor \frac{k_0}{k} \rfloor = rt + \lfloor \frac{k_0}{k_0+1} \rfloor = rt$ for the time complexity of $\mathcal{M}'$. □

Example 15 Let $k_0 \geq 1$ and $\mathcal{M}$ be a device in question with time complexity $rt + k_0 \cdot rt$. Then for all rational numbers $\varepsilon > 0$ there is an equivalent device $\mathcal{M}'$ of the same type with time complexity $\lfloor (1+\epsilon)\cdot rt \rfloor$. We set $k = \lceil \frac{k_0}{\epsilon} \rceil$ and apply Theorem 13 in order to obtain $rt + \left\lfloor \frac{k_0 \cdot rt}{\lceil k_0/\epsilon \rceil} \right\rfloor \leq rt + \left\lfloor \frac{k_0 \cdot rt}{k_0/\epsilon} \right\rfloor = rt + \lfloor \epsilon \cdot rt \rfloor = \lfloor (1+\epsilon)\cdot rt \rfloor$. □

Computational Capacities

In this section we explore the computational capacities of real-time, linear-time, and unbounded time devices. The goal is to establish a hierarchy of language families and to compare the levels with well-known linguistic families of the Chomsky hierarchy language family. The properness of some inclusions are long-standing open problems with deep relations to sequential complexity problems. In order to establish the hierarchy we start at the upper and lower end. Straightforward constructions of linearly space-bounded Turing machines from IAs, of IAs from CAs, and of CAs from linearly space-bounded Turing machines show the following lemma [80].

Lemma 16 *The families $\mathcal{L}(CA)$ and (IA) are identical with the deterministic context-sensitive languages, i. e., with the complexity class DSPACE(n).*

At the lower end we consider the regular languages. Since already the communication cell is a deterministic finite-state machine, clearly, the regular languages are a subset of $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$. In Theorem 4 the stronger result that any real-time deterministic context-free language belongs to $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$ has been shown. So, the properness of the following inclusion is obvious.

Corollary 17 *The regular languages are a proper subset of $\mathcal{L}_{rt}(IA)$.*

The question arises whether the real-time condition of Theorem 4 can be relaxed in order to obtain a larger sub-family of the context-free languages which is accepted

by real-time IAs. The answer is negative. The following lemma marks a sharp boundary between context-free languages acceptable and non-acceptable by real-time IAs. Recall that linear context-free languages are accepted by nondeterministic one-turn pushdown automata. Since here we deal with deterministic devices, the alternative grammar characterization is more suitable for our purposes.

A grammar $\mathcal{G} = \langle N,T,S,P \rangle$ is said to be *linear context free*, if the productions in P are of the forms $(X \to a)$, $(X \to Ya)$, or $(X \to aY)$, where $X,Y \in N$ are nonterminals and $a \in T$ is a terminal symbol (e. g. [72]).

Lemma 18 *There exists a deterministic, linear context-free language that does not belong to $\mathcal{L}_{rt}(IA)$.*

Proof Clearly, language

$$L = \{\&x_k\&\cdots\&x_1?y_1\&\cdots\&y_k\& \mid k \geq 1, x_i^R = y_i z_i \text{ and } x_i, y_i, z_i \in \{a,b\}^*\}$$

is deterministic and linear context free. By Example 7 it does not belong to $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$. □

We turn to real-time OCAs. Similar as above, the linear context-free languages serve as sub-family of the context-free languages which is contained in $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$ [76].

Theorem 19 *Given some linear context-free grammar, an equivalent real-time OCA can effectively be constructed, i. e., every linear context-free language belongs to $\mathcal{L}_{rt}(OCA)$.*

Proof Let $\mathcal{G} = \langle N, T, S, P\rangle$ be a linear context-free grammar, and $w = a_1 a_2 \cdots a_n$ be a word in $L(\mathcal{G})$. If $X \Rightarrow^* a_i \cdots a_j$, $1 \leq i < j \leq n$, then there exists a $Y \in N$ such that either $(Y \Rightarrow^* a_i \cdots a_{j-1} \wedge X \Rightarrow Ya_j)$ or $(Y \Rightarrow^* a_{i+1} \cdots a_j \wedge X \Rightarrow a_i Y)$. Based on this fact we define sets of nonterminals for the given word w as follows:

$$\begin{aligned} N(i,i) &= \{X \in N \mid (X \to a_i) \in P\}\,, \quad 1 \leq i \leq n \\ N(i,j) &= N_1(i,j) \cup N_2(i,j)\,, \quad 1 \leq i < j \leq n\,, \\ \text{where } N_1(i,j) &= \{X \in N \mid (X \to Ya_j) \in P \\ &\quad \wedge\, Y \in N(i,j-1)\} \\ \text{and } N_2(i,j) &= \{X \in N \mid (X \to a_i Y) \in P \\ &\quad \wedge\, Y \in N(i+1,j)\}\,. \end{aligned}$$

So, the word $a_1 \cdots a_n$ is generated by $\mathcal{G}$ if and only if $S \in N(1,n)$.

A real-time OCA $\mathcal{M} = \langle S', \delta, \#, T, F\rangle$ accepting $L(\mathcal{G})$ behaves as follows. It analyzes its input $a_1 \cdots a_n$ in such a way that the cells successively compute certain sets $N(i,j)$. In the first step, all cells $1 \leq i \leq n$ compute $N(i,i)$ and (a_i, a_i). In subsequent steps $k \geq 2$, they compute $N(i, i+k-1)$ and (a_i, a_{i+k-1}) if $k \leq n+1-i$, and keep their state otherwise. The new values can be determined by the state of the cell itself and the state of its right neighbor, i. e., $N(i, i+k-2)$, (a_i, a_{i+k-1}) and $N(i+1, i+k-1)$, (a_{i+1}, a_{i+k}). Thus, the leftmost cell computes $N(1,n)$ at time step n. □

Corollary 20 *The regular languages are a proper subset of $\mathcal{L}_{rt}(OCA)$.*

Again, the question arises whether the condition of Theorem 19 can be relaxed in order to obtain a larger sub-family of the context-free languages that is accepted by real-time OCAs. Again, the answer is negative.

Lemma 21 *There exists a two-linear context-free language that does not belong to $\mathcal{L}_{rt}(OCA)$.*

Proof Consider the linear language $L = \{a^n b^n \mid n \geq 1\} \cup \{a^n bvab^n \mid n \geq 1, v \in \{a,b\}^*\}$. By counting arguments, in [85] it is shown that the two-linear concatenation LL is not accepted by any real-time OCA. □

Beginning with the regular languages at the bottom of the hierarchy, next we deal with the proper supersets given by real-time deterministic and linear context-free languages. Since both families are known to be incomparable with respect to inclusion, the hierarchy splits into two strands. Unfortunately we have to continue with these two strands for a moment.

Theorem 22 *The families $\mathcal{L}_{rt}(OCA)$ and $\mathcal{L}_{rt}(IA)$ are incomparable.*

Proof By Lemma 18 there exists a linear context-free language not accepted by any real-time IA, but by Theorem 19 it belongs to a real-time OCA.

Conversely, one can show that the two-linear language of Lemma 21 belongs to $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$. □

The next step is to go beyond $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$ and $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$. Structurally this concerns CAs, but first we increase the time complexity. For the proof of infinite hierarchies in between real time and linear time, it is necessary to control the lengths of words with respect to some internal substructures. The following notion of constructibility expresses the idea that the length of a word relative to the length of a subword should be computable. To this end, a function $f\colon \mathbb{N} \to \mathbb{N}$ is said to be *OCA-constructible*, if there exist an λ-free homomorphism h and a language $L \in \mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$ such that $h(L) = \{a^{f(n)-n} b^n \mid n \geq 1\}$. Since constructible functions describe the length of the whole word dependent on the length of a subword, it is obvious that each constructible function must be greater than or equal to the identity. At a glance, this notion of constructibility might look somehow unusual or restrictive. However, λ-free homomorphisms are very powerful,

so that the family of (in this sense) constructible functions is very rich. Examples and the next theorem are presented in [45].

Theorem 23 *Let $r_1, r_2 \colon \mathbb{N} \to \mathbb{N}$ be two increasing functions. If $r_2 \cdot \log(r_2) \in o(r_1)$ and r_1^{-1} is OCA constructible, then $\mathcal{L}_{n+r_2(n)}(OCA) \subset \mathcal{L}_{n+r_1(n)}(OCA)$.*

Example 24 Let $0 \le p < q \le 1$ be two rational numbers. Clearly, $n^p \cdot \log(n^p)$ is of order $o(n^q)$. Moreover, the inverse of n^q is OCA-constructible. Thus, an application of Theorem 23 yields the strict inclusion

$$\mathcal{L}_{n+n^p}(\mathrm{OCA}) \subset \mathcal{L}_{n+n^q}(\mathrm{OCA}) \,. \qquad \square$$

Example 25 Let $i < j$ be two positive integers, then $\log^{[j]}(n) \cdot \log^{[j+1]}(n)$ is of order $o(\log^{[i]}(n))$. Since the inverse of $\log^{[i]}(n)$ is OCA-constructible, we obtain the strict inclusion

$$\mathcal{L}_{n+\log^{[j]}(n)}(\mathrm{OCA}) \subset \mathcal{L}_{n+\log^{[i]}(n)}(\mathrm{OCA}) \,. \qquad \square$$

Similar results are known for iterative arrays. The infinite strict hierarchies in the range between real time and linear time are established in [9]. The constructibility is defined differently. A strictly increasing function $f \colon \mathbb{N} \to \mathbb{N}$ is *IA-constructible*, if there exists an IA with infinitely many cells to the right, such that the leftmost cell enters some state from a distinguished subset of states exactly at all time steps $f(i)$, $1 \le i$. Example 1 shows that the function 2^n is IA-constructible.

Theorem 26 *Let $r_1, r_2 \colon \mathbb{N} \to \mathbb{N}$ be two increasing functions. If $r_2 \in o(r_1)$ and r_1^{-1} is IA-constructible, then $\mathcal{L}_{n+r_2(n)}(IA) \subset \mathcal{L}_{n+r_1(n)}(IA)$.*

The following example is based on natural functions.

Example 27 Since the family of IA-constructible functions is closed under composition and contains 2^n and n^k, $k \ge 1$, the functions $\log^{[i]}(n)$, $i \ge 1$, and $\sqrt[k]{n}$ are inverses of constructible functions. Actually, the inverses of 2^n and n^k are $\lceil \log(n) \rceil$ and $\lceil n^{\frac{1}{k}} \rceil$ but for convenience we simplify the notation. Therefore, an application to the hierarchy theorem yields

$$\mathcal{L}_{\mathrm{rt}}(\mathrm{IA}) \subset \cdots \subset \mathcal{L}_{n+\log^{[i+1]}(n)}(\mathrm{IA}) \subset \mathcal{L}_{n+\log^{[i]}(n)}(\mathrm{IA}) \subset \cdots \subset \mathcal{L}_{\mathrm{lt}}(\mathrm{IA})$$

and

$$\mathcal{L}_{\mathrm{rt}}(\mathrm{IA}) \subset \cdots \subset \mathcal{L}_{n+n^{\frac{1}{i+1}}}(\mathrm{IA}) \subset \mathcal{L}_{n+n^{\frac{1}{i}}}(\mathrm{IA}) \subset \cdots \subset \mathcal{L}_{\mathrm{lt}}(\mathrm{IA}) \,,$$

or in combination, e. g.,

$$\begin{aligned}\mathcal{L}_{\mathrm{rt}}(\mathrm{IA}) &\subset \cdots \subset \mathcal{L}_{n+(\log^{[j+1]}(n))^{\frac{1}{i+1}}}(\mathrm{IA}) \\ &\subset \mathcal{L}_{n+(\log^{[j+1]}(n))^{\frac{1}{i}}}(\mathrm{IA}) \subset \cdots \subset \mathcal{L}_{n+(\log^{[j]}(n))^{\frac{1}{i+1}}}(\mathrm{IA}) \\ &\subset \mathcal{L}_{n+(\log^{[j]}(n))^{\frac{1}{i}}}(\mathrm{IA}) \subset \cdots \subset \mathcal{L}_{\mathrm{lt}}(\mathrm{IA}) \,.\end{aligned}$$

$\square$

Example 9 reveals a unary language not accepted by any real-time OCA. Note that the witness language $\{a^{2^n} \mid n \ge 1\}$ is not context free. Next, we generalize this observation and show that even massively parallel real-time OCAs cannot accept more unary languages than a single deterministic finite-state machine [73].

Theorem 28 *Each unary real-time OCA language is regular.*

Proof Let $A = \{a\}$ be an alphabet, $L \subseteq A^+$ a language, and $\mathcal{M} = \langle S, \delta, \#, A, F \rangle$ be an OCA accepting L in real time. We construct an equivalent deterministic finite-state machine $\mathcal{E}$ with state set $S \times S$, initial state s_0, set of accepting states F', and transition function δ' as follows.

$$\begin{aligned} &s_0 = (\#, a) \,, \qquad F' = \{(s_1, s_2) \in S \times S \mid s_1 \in F\} \,, \\ &\delta'((s_1, s_2), a) = (\delta(s_2, s_1), \delta(s_2, s_2)) \,, \\ &\text{for all } (s_1, s_2) \in S \times S \end{aligned}$$

In order to give evidence that the construction is correct, we depict two correspondent computations in Fig. 8. $\square$

Next we can join the two strands of the hierarchy again. The superfamily is $\mathcal{L}_{\mathrm{rt}}(\mathrm{CA})$.

Theorem 29 *The family $\mathcal{L}_{rt}(OCA)$ is properly included in $\mathcal{L}_{rt}(CA)$.*

Proof The inclusion follows for structural reasons. For the properness we argue as follows. Since the language $L = \{a^{2^n} \mid n \ge 1\}$ is not regular, it does not belong to $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$. On the other hand, Example 1 shows that it belongs to $\mathcal{L}_{\mathrm{rt}}(\mathrm{CA})$. $\square$

Theorem 30 *The family $\mathcal{L}_{rt}(IA)$ is properly included in $\mathcal{L}_{rt}(CA)$.*

Proof First we give evidence of the inclusion $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA}) \subseteq \mathcal{L}_{\mathrm{rt}}(\mathrm{CA})$. A real-time CA can be set up in such a way that it shifts its input successively to the left on an additional track. So, the leftmost cell receives the input symbol by symbol. Therefore, the CA can simulate the iterative array, where the leftmost cell simulates the communication cell.

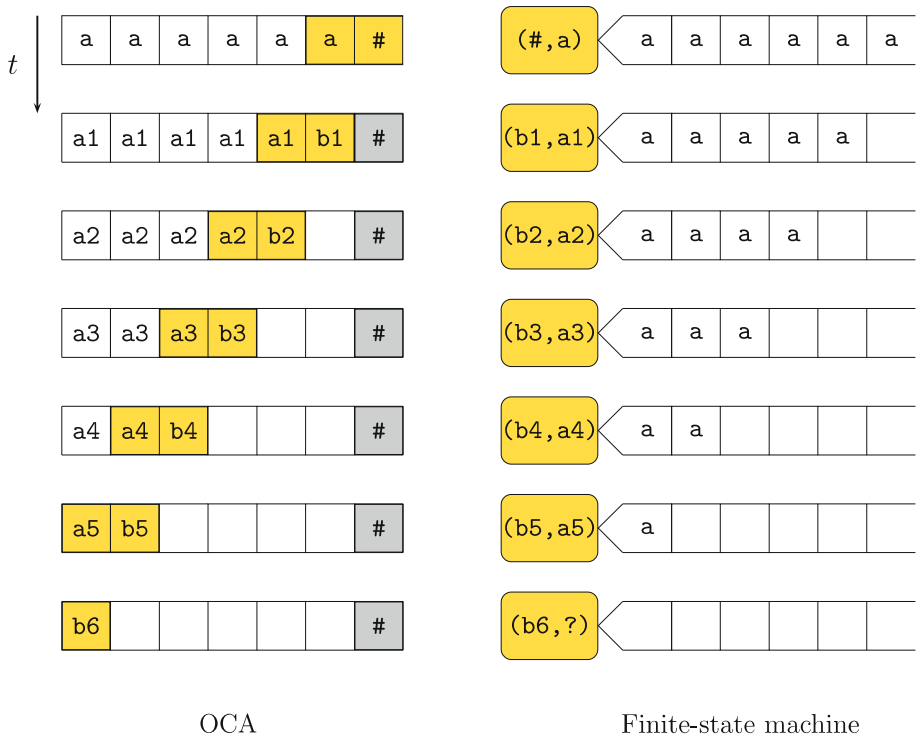

Cellular Automata and Language Theory, Figure 8
A finite-state machine simulating a real-time OCA on unary input

The properness of the inclusion follows by the inclusion $\mathcal{L}_{rt}(\text{OCA}) \subset \mathcal{L}_{rt}(\text{CA})$ and the incomparability of $\mathcal{L}_{rt}(\text{OCA})$ and $\mathcal{L}_{rt}(IA)$. □

Once we know that, in general, a real-time CA language cannot be accepted by any real-time OCA, the question arises of how much time is necessary for that, if possible at all. The next result gives an upper bound for the time [17,95]. Admittedly, to this end the input has to be reversed. Alternatively, one could reverse the neighborhood of the cells in an OCA. Then the rightmost cell indicates the result of the computation. In this case the input could remain as it is. In any case, the condition cannot be relaxed since it is an open problem whether the corresponding language families are closed under reversal.

Theorem 31 *A language is accepted by a linear-time OCA if and only if its reversal is accepted by a CA in real time.*

Proof Let $\mathcal{M}$ be a real-time CA. The cells of a linear-time OCA $\mathcal{M}'$ accepting $L(\mathcal{M})^R$ collect the information necessary to simulate one transition of $\mathcal{M}$, in an intermediate step. Therefore, the first step of $\mathcal{M}$ is simulated in the second step of $\mathcal{M}'$. We obtain a behavior as depicted in Fig. 9.

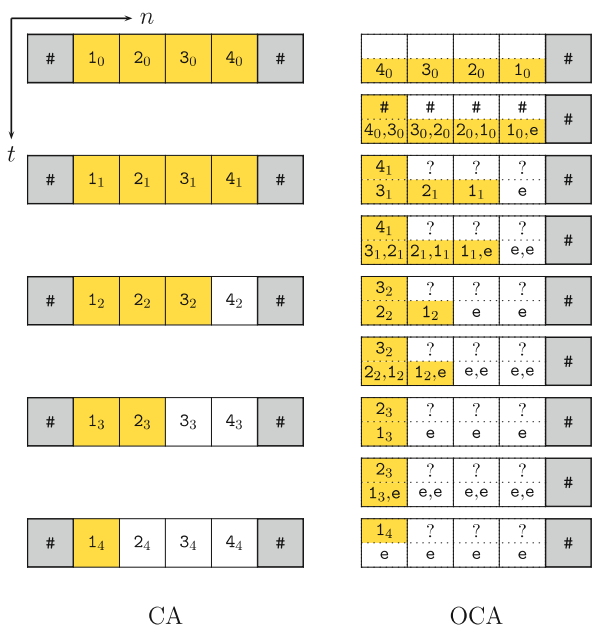

Cellular Automata and Language Theory, Figure 9
Intermediate steps in the construction of the proof of Theorem 31

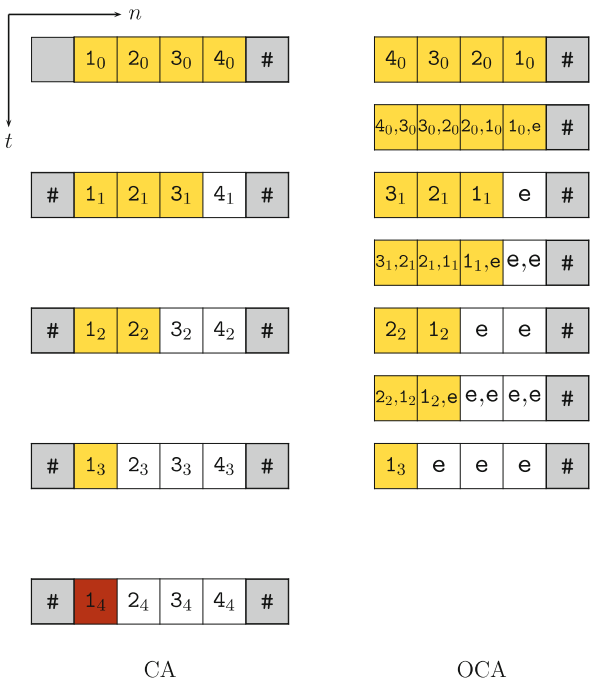

Cellular Automata and Language Theory, Figure 10
Example of a linear-time OCA simulation of a real-time CA computation on reversed input

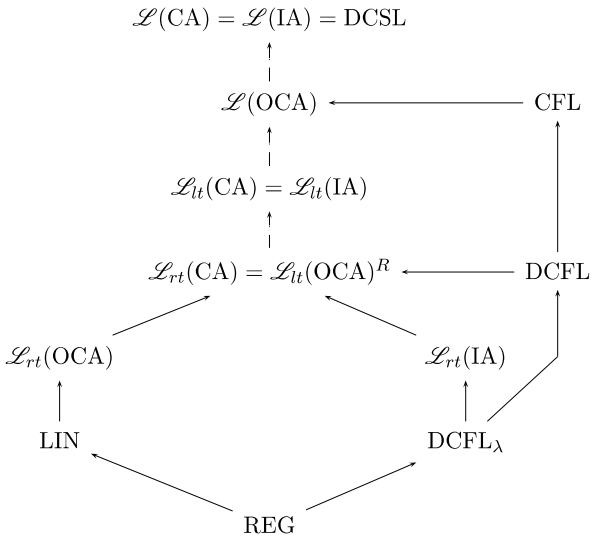

Cellular Automata and Language Theory, Figure 11
Basic hierarchy of language families. A solid arrow indicates a proper inclusion and a dashed arrow an inclusion. In addition, the linear languages (LIN) are properly included in the context-free languages (CFL). Deterministic and real-time deterministic context-free languages are denoted by DCFL and DCFL_λ, regular languages by REG, and deterministic context-sensitive languages by DCSL

Altogether, $\mathcal{M}'$ cannot simulate the last step of $\mathcal{M}$. So, the construction has to be extended slightly. Each cell has an extra register that is used to simulate transitions of $\mathcal{M}$ under the assumption that the cell is the leftmost one (see Fig. 10). The transitions of the real leftmost cell now correspond to the missing transitions of the previous simulation. □

Climbing up one level, the hierarchy continues with two-way linear-time devices.

Theorem 32 *A language is accepted by a linear-time IA if and only if it is accepted by a linear-time CA.*

Proof The inclusion $\mathcal{L}_{\mathrm{lt}}(\mathrm{IA}) \subseteq \mathcal{L}_{\mathrm{lt}}(\mathrm{CA})$ follows by the proof of Theorem 30.

Conversely, a linear-time iterative array can simulate a linear-time CA as follows. In the first phase, it reads the input and stores it successively in its cells. Next, the iterative array starts a time optimal firing squad synchronization algorithm (see for example [62,98]). That is, an algorithm which is initiated by the communication cell, and which synchronizes the n cells within $2n - 2$ time steps. Finally, all cells start the simulation of the CA at the same time. Clearly, the iterative array obeys a linear time bound if the cellular automaton does. □

The next inclusion does not follow for structural reasons. It is proved in [16,32] in terms of simulations of equivalent sequential machines. The properness is an open problem.

Theorem 33 *Each linear-time CA language belongs to the family $\mathcal{L}(OCA)$.*

The family $\mathcal{L}(\mathrm{OCA})$ is very powerful. It contains the context-free languages as well as a PSPACE-complete language [16,32]. Altogether we obtained the hierarchy depicted in Fig. 11, where the only known proper inclusions are at the lower end. Nevertheless, even the real-time OCA languages contain important families, e. g., the Dyck languages [73] and the bracketed context-free languages [26]. Furthermore, the non-semilinear language $\{(a^i b)^* \mid i \geq 0\}$ [73] and the inherently ambiguous language $\{a^i b^j c^k \mid i = j \text{ or } j = k \text{ for } i, j, k \geq 1\}$ [80] belong to $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$. On the other hand, the context-free languages have been shown to be incomparable with $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$. Whether or not they are a subset of the family $\mathcal{L}_{\mathrm{rt}}(\mathrm{CA})$ is an open question raised in [80].

We conclude this section with a real-time OCA based characterization of the recursively enumerable languages and its implication to incomparability [58].

Lemma 34 *A language is recursively enumerable if and only if there exists a homomorphism h and a language $L' \in \mathcal{L}_{\mathrm{rt}}(OCA)$ such that $L = h(L')$. The same holds for $\mathcal{L}_{\mathrm{rt}}(IA)$.*

Proof It is known that the recursively enumerable languages can be represented by $h(L_1 \cap L_2)$, where h is a homomorphism and L_1, L_2 are either linear context-free languages [4] or real-time deterministic context-free languages [29]. In fact, in the latter paper deterministic pushdown automata are constructed that use λ-moves. But the constructions can easily be modified to obtain real-time pushdown automata. So, the assertions follow since $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$ contains the linear context-free languages, $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$ contains the real-time deterministic context-free languages, and both families are closed under intersection (Theorem 39). □

The homomorphic characterization of the recursively enumerable languages reveals the following general incomparability result.

Theorem 35 *Both families $\mathcal{L}_{rt}(OCA)$ and $\mathcal{L}_{rt}(IA)$ are incomparable to each language family $\mathcal{L}$ that contains the context-free languages, is itself properly contained in the recursively enumerable languages, and is closed under homomorphism.*

Proof Since there is a context-free language not belonging to $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$, we obtain $\mathcal{L} \not\subseteq \mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$. Conversely, if $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA}) \subseteq \mathcal{L}$, then the recursively enumerable languages are a subset of $\mathcal{L}$ by Lemma 34. This contradicts the proper containment in the recursively enumerable languages. The same argumentation applies to $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$. □

Example 36 The families $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$ and $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$ are incomparable to the languages of indexed grammars, certain grammars with regulated rewriting, certain contextual grammars, and certain Lindenmayer systems, e. g., ET0L systems. □

Closure Properties

Closure properties of families of formal languages indicate their robustness under certain operations. A family of languages is *closed under some operation*, if any application of the operation on languages from the family yields again a language from the family. It is *effectively closed* if the result of the operation can be constructed from the given language(s). The knowledge of closure properties often reveals insights in particular features, structures, and capabilities. Moreover, positive closure properties are a useful tool for modular constructions and decompositions in a natural way. Negative properties may serve, for example, as valuable basis for extensions. That is, for building the smallest family of languages which is closed under the operation and contains the family in question. In any case closure properties are filigree tools for dealing with language families.

Boolean Operations

The operations union, intersection, complementation and set difference are commonly called *Boolean operations* or *elementary operations*.

For deterministic devices, the closure under complementation is often shown by interchanging accepting and non-accepting states. But, in general, this requires halting computations.

Theorem 37 *For $X \in \{CA, OCA, IA\}$, the families $\mathcal{L}_{rt}(X)$ and $\mathcal{L}_{lt}(X)$ are effectively closed under complementation.*

Proof We show the closure exemplarily for linear-time OCAs. The other proofs are similar. The reason that the complementary device cannot be constructed by simply interchanging accepting and non-accepting states is that an input is accepted when the leftmost cell enters an accepting state at some arbitrary time step. So, in general, the leftmost cell will enter accepting as well as non-accepting states during a computation. To cope with this problem we modify a given linear-time OCA $\mathcal{M}$ in the following way. At initial time a signal is emitted by the rightmost cell. The signal moves on an extra track with speed $\frac{1}{2}$ to the left. It will arrive at the leftmost cell at twice real time, i. e., linear time. This is the time step at which we wish to make the final decision whether to accept or to reject the input. To this end, the leftmost cell has to remember if it has entered an accepting state at some time before. So, we use a copy S' of the state set S of $\mathcal{M}$, and modify the local transition function to drive the leftmost cell into a state of S' when it enters an accepting state. Subsequently, the normal behavior of the leftmost cell is simulated, except that states of S' are used instead of states of S. Now the modified automaton $\mathcal{M}'$ accepts input if and only if the leftmost cell is in some state of S' when the signal arrives. In order to accept the complement of $L(\mathcal{M}) = L(\mathcal{M}')$, it suffices to let the automaton accept input if and only if the leftmost cell is in some state of S when the signal arrives. □

Devices without any time bounds are, in fact, exponentially time bounded due to the space limitation.

Lemma 38 *For $X \in \{CA, OCA, IA\}$, the families $\mathcal{L}(X)$ are effectively closed under complementation.*

Proof The construction of Theorem 37 is modified as follows. Some device in question with n cells and state set S may run through at most $|S|^n$ different configurations until its behavior becomes cyclic. So, it suffices to check whether the input is accepted during the first $|S|^n$ time steps. As discussed briefly in Sect. "Cellular Language Acceptors", exponential time complexities are honest. That is, the leftmost cell can recognize the time step $|S|^n$. To this end, an

$|S|$ary counter is set up on an extra track. The rightmost cell simulates the least significant digit and adds one to the counter at every time step. The neighboring cell to the left observes when a carry-over appears, increases its own digit and so on. The time step at which the first carry-over appears in the leftmost cell is the desired one. □

Now we turn to intersection, union, and set difference.

Theorem 39 *For $X \in \{CA, OCA, IA\}$, the families $\mathcal{L}_{rt}(X)$, $\mathcal{L}_{lt}(X)$, and (X) are effectively closed under intersection, union, and set difference.*

Proof In order to prove the closures, the well-known two-track technique is applicable. That is, on two different tracks acceptors for the languages in question are simulated independently of each other. By the same construction as in the previous proof, the leftmost cell can remember whether the single tracks accept or not. The accepting states are composed by the accepting and non-accepting components for the single tracks in the usual way. For example, to show intersection both components have to be accepting ones. □

Reversal

The closure under reversal is of crucial importance. It is an open problem for $\mathcal{L}_{\mathrm{rt}}(\mathrm{CA})$ and, equivalently, for $\mathcal{L}_{\mathrm{lt}}(\mathrm{OCA})$. Moreover, it is linked with the open closure property under concatenation for the same family and, hence, with the question whether linear-time CAs are more powerful than real-time CAs. So, it remains open whether the computational capacities of CAs differ if the rightmost or the leftmost cell indicates acceptance.

Theorem 40 *The family $\mathcal{L}_{rt}(OCA)$ is effectively closed under reversal.*

Proof Let $\mathcal{M}$ be some real-time OCA. In order to obtain a real-time OCA $\mathcal{M}'$ for the language $L(\mathcal{M})^R$, the arguments of the local transition function are interchanged. That is, $\delta'(s_2, s_1) = s_3$ if $\delta(s_1, s_2) = s_3$. In addition, we have to pay special attention to the boundary state. The further construction is as in the proof of Theorem 31, with the exception that we do not need to collect information in intermediate steps, since the information flow is one-way in both devices. (see Fig. 12). □

As mentioned above, the closure under reversal can be interpreted in two different ways. On one hand, one can construct an OCA that accepts the reversal of a given OCA language. On the other hand, one can construct an OCA that accepts the same language with the rightmost cell. The latter point of view yields immediately the closure under reversal of two-way linear-time cellular automata.

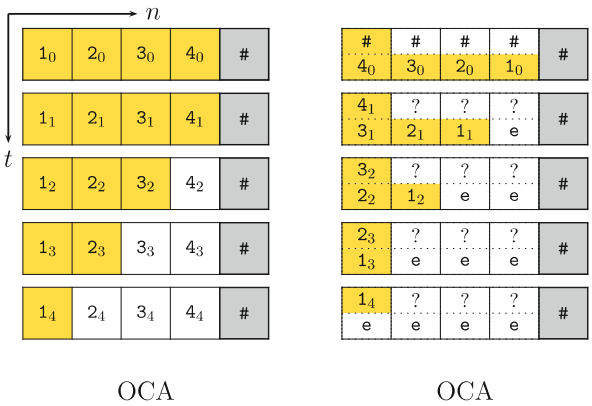

Cellular Automata and Language Theory, Figure 12
Construction showing the closure of real-time OCA languages under reversal

Theorem 41 *The family $\mathcal{L}_{lt}(CA) = \mathcal{L}_{lt}(IA)$ is effectively closed under reversal.*

Proof Let $\mathcal{M}$ be some linear-time CA. In order to obtain a linear-time CA $\mathcal{M}'$ for the language $L(\mathcal{M})^R$, the first and third arguments of the local transition function are interchanged. That is, $\delta'(s_3, s_2, s_1) = s_4$ if $\delta(s_1, s_2, s_3) = s_4$. The resulting device accepts $L(\mathcal{M})^R$ with the rightmost cell. Then the result is sent as a signal to the leftmost cell. Altogether, $\mathcal{M}'$ still obeys a linear time bound. □

The closure under reversal of the devices without time bounds follows from the known closures of the characterizing linguistic language family, i. e., the deterministic context-sensitive languages DSPACE(n).

Corollary 42 *The family $\mathcal{L}(CA) = \mathcal{L}(IA)$ is effectively closed under reversal.*

In [16,32] unbounded time OCAs are simulated by a variant of unbounded time one-way iterative arrays and vice versa. Moreover, it is shown that the family of accepted languages forms an AFL, i. e., an *abstract family of languages*, (e. g. [72]) which is in addition closed under reversal.

Lemma 43 *The family $\mathcal{L}(OCA)$ is effectively closed under reversal.*

In order to show negative closure properties it is sometimes convenient to have witness languages not belonging to the family in question. By Example 7 the language

$$L = \{\&x_k\&\cdots\&x_1?y_1\&\cdots\&y_k\& \mid k \geq 1, x_i^R = y_i z_i \text{ and } x_i, y_i, z_i \in \{a, b\}^*\}$$

is known not to belong to $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$.

Theorem 44 *The family $\mathcal{L}_{rt}(IA)$ is not closed under reversal.*

Proof It suffices to show that L^R belongs to $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$. To this end, a real-time deterministic pushdown automaton accepting L^R is easily constructed. Then by Theorem 4 the containment $L^R \in \mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$ follows. □

If the answer to the open reversal closure of $\mathcal{L}_{\mathrm{rt}}(\mathrm{CA})$ is negative, we have to deal with two different language families. Since the properness of the inclusion $\mathcal{L}_{\mathrm{rt}}(\mathrm{CA}) \subseteq \mathcal{L}_{\mathrm{lt}}(\mathrm{CA})$ is also open, the problem gains in importance. A negative answer of the former problem would imply a proper inclusion. A language $L \in \mathcal{L}_{\mathrm{rt}}(\mathrm{CA})$ whose reversal does not belong to $\mathcal{L}_{\mathrm{rt}}(\mathrm{CA})$ may serve as witness since $\mathcal{L}_{\mathrm{lt}}(\mathrm{CA})$ is closed under reversal by Theorem 41. In fact, the following stronger relation is shown in [33] by a long proof.

Theorem 45 *The family $\mathcal{L}_{rt}(CA)$ is closed under reversal if and only if $\mathcal{L}_{rt}(CA)$ and $\mathcal{L}_{lt}(CA)$ are identical.*

Concatenation

Concerning the closure properties under concatenation, the situation is similar to reversal. Either the properties are trivial due to the characterizations by well-known language families, or they are negative, or open problems. For devices without time bounds we have, again, the closures of the deterministic context-sensitive languages DSPACE(n).

Corollary 46 *The family $\mathcal{L}(CA) = \mathcal{L}(IA)$ is effectively closed under concatenation.*

Since any AFL is closed under concatenation and, as mentioned before, $\mathcal{L}(\mathrm{OCA})$ is an AFL [16,32] whose closure properties are shown by simulations, the next lemma follows immediately.

Lemma 47 *The family $\mathcal{L}(OCA)$ is effectively closed under concatenation.*

In order to show that $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$ is not closed under concatenation, once more the witness is

$$L = \{\&x_k\& \cdots \&x_1?y_1\& \cdots \&y_k\& \mid k \geq 1, x_i^R = y_i z_i \text{ and } x_i, y_i, z_i \in \{a,b\}^*\}\,.$$

Theorem 48 *The family $\mathcal{L}_{rt}(IA)$ is not closed under concatenation.*

Proof In contrast to the assertion, assume that $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$ is closed under concatenation. The language

$$L_1 = \{\&x(\&\{a,b\}^+)^k?\&^k y\& \mid x^R = yz,\ x,y,z \in \{a,b\}^*\}$$

is clearly a real-time deterministic context-free and, thus, a real-time IA language. The language

$$L_2 = (\&\{a,b\}^+)^*$$

is regular and, therefore, accepted by some real-time IA, too. Due to the assumption, the language

$$L_3 = L_2 L_1 \&^*$$

belongs also to $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$. Next, the language

$$L_4 = \{(\&\{a,b\}^+)^k?(\{a,b\}^*\&)^k \mid k \geq 1\}$$

is a real-time deterministic context-free language and therefore accepted by some real-time IA. Finally, from the closure under intersection we obtain

$$L_5 = L_3 \cap L_4 \in \mathcal{L}_{\mathrm{rt}}(\mathrm{IA})\,.$$

However, the language $L_5 \subset L$ contains the words used to show $L \notin \mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$. We conclude $L_5 \notin \mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$, a contradiction. □

The question whether or not the family $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$ is closed under concatenation was open for a long time. It has been solved negatively in [84]. Here we utilize the language L_d of Example 12. Recall that $L_d \subset \{0,1,(,),|\}^+$ is the language whose words are of the form

$$x(x_1|y_1)\cdots(x_n|y_n)y\,,$$

where $x, x_i, y, y_i \in \{0,1\}^*$, for $1 \leq i \leq n$, and $(x|y) = (x_i|y_i)$, for at least one $i \in \{1,\ldots,n\}$.

The following example will be helpful.

Example 49 The language $L_c = \{w \bullet w \mid w \in \{0,1\}^+\}$ is a real-time OCA language.

An acceptor for L_c has several tracks. On one track the input is preserved. On two different tracks the input is successively shifted to the left. The shifting is in such a way that a symbol moves to the left one cell per time step until it passes through the center cell with input $\bullet$. Subsequently, it moves to the left every other time step. In order to achieve this slow-down, the second track is used. Symbols are received in the first register, then shifted to the second one and, finally, send to the left neighbor. In addition, at initial time a signal is emitted at the right border. When the signal has passed through the center cell, it starts to compare the original input symbol with the symbol to be shifted out of the cell next.

Let the input be $a_m \cdots a_1 \bullet a'_m \cdots a'_1$. At time step $m+1+i$, $1 \leq i \leq m$, the signal arrives in the cell with input a_i. The symbol a'_i takes $m+1-i$ time steps to arrive at the

center, and is to be shifted out of the cell carrying the original input a_i at time $m+1-i+2i$. So, the signal compares a'_i with a_i as required. □

Theorem 50 *The family $\mathcal{L}_{rt}(OCA)$ is not closed under concatenation.*

Proof Consider the language L_1 whose words are of the form

$$x(x_1|y_1)\cdots(x_n|y_n)(x|,$$

where $x, x_i, y_i \in \{0,1\}^*$, for $1 \le i \le n$.

Similar to Example 49 we obtain a simple construction of a real-time OCA accepting L_1. For symmetry reasons the language L_2 whose words are of the form

$$y)(x_1|y_1)\cdots(x_n|y_n)y\ ,$$

where $y, x_i, y_i \in \{0,1\}^*$, for $1 \le i \le n$, is a real-time OCA language, too.

However, the concatenation L_1L_2 equals L_d, which does not belong to $\mathcal{L}_{rt}(OCA)$. □

The question whether or not one of the families $\mathcal{L}_{rt}(CA) = \mathcal{L}_{lt}^R(OCA)$ or $\mathcal{L}_{lt}(CA) = \mathcal{L}_{lt}(IA)$ is closed under concatenation is another famous open problem in this field. Nevertheless, it is shown in [33] that the closure of $\mathcal{L}_{rt}(CA)$ under reversal implies its closure under concatenation. Since in this case we obtain $\mathcal{L}_{rt}(CA) = \mathcal{L}_{lt}(CA)$, the family of linear-time CA languages were also closed under concatenation.

Theorem 51 *If the family $\mathcal{L}_{rt}(CA)$ is closed under reversal, then it is closed under concatenation.*

Proof If $\mathcal{L}_{rt}(CA)$ is closed under reversal, then by Theorem 45 we have the identity $\mathcal{L}_{rt}(CA) = \mathcal{L}_{lt}(CA)$. So, it suffices to construct a linear-time CA $\mathcal{M}$ for the concatenation of two real-time CA languages L_1 and L_2. By the closure under reversal, Theorem 31, and the speed-up theorems, there are $2n$-time OCAs $\mathcal{M}_1$ and $\mathcal{M}_2$ for L_1^R and L_2.

During a first phase, automaton $\mathcal{M}$ reverses its input, say $a_1 \cdots a_n$, on an extra track. This takes n time steps.

During a second phase, automaton $\mathcal{M}$ simulates automaton $\mathcal{M}_1$ on $a_n \cdots a_1$ and $\mathcal{M}_2$ on $a_1 \cdots a_n$ in parallel. When some cell enters an accepting state during the simulations, the cell is marked on the corresponding track. If a cell at position $n+1-i$ carrying the input symbol a_i is marked by the simulation of $\mathcal{M}_1$, we have $a_i \cdots a_1 \in L_1^R$ and, thus, $a_1 \cdots a_i \in L_1$. If a cell at position i carrying the input symbol a_i is marked by the simulation of $\mathcal{M}_2$, we have $a_i \cdots a_n \in L_2$. So, the input $a_1 \cdots a_n$ belongs to the concatenation $L_1\ L_2$ if and only if $\mathcal{M}_1$ marks a cell at position $n+1-i$ and $\mathcal{M}_2$ a cell at position $i+1$, for $1 \le i < n$. In order to check this condition, $\mathcal{M}$ reverses the result of the simulation of $\mathcal{M}_1$. Now a cell at position i is marked if and only if previously a cell at position $n+1-i$ was marked. Therefore, it suffices to verify by a signal whether two adjacent cells are marked. □

Cellular Automata and Language Theory, Table 1
Summary of closure properties. Concatenation REG denotes the concatenation with regular languages at the right, REG concatenation at the left, *hom* denotes homomorphisms, *gsm* generalized sequential machine mappings, and inj. length-pres. abbreviates injective length-preserving. A + indicates closure, a — non-closure, and a question mark an open problem

	$\mathcal{L}_{rt}$ (OCA)	$\mathcal{L}_{rt}$ (IA)	$\mathcal{L}_{rt}$ (CA)	$\mathcal{L}_{lt}$ (CA)	$\mathcal{L}$ (OCA)	$\mathcal{L}$ (CA)
$\cup, \cap$	+	+	+	+	+	+
complementation, —	+	+	+	+	+	+
reversal	+	—	?	+	+	+
concatenation	—	—	?	?	+	+
λ-free iteration	—	—	?	?	+	+
concatenation REG	+	+	+	?	+	+
REG concatenation	+	—	?	?	+	+
marked concatenation	+	+	+	+	+	+
marked λ-free iteration	+	+	+	+	+	+
hom^{-1}	+	+	+	+	+	+
deterministic gsm^{-1}	+	+	+	+	+	+
gsm^{-1}	—	?	?	?	+	+
inj. length-pres. *hom*	+	+	+	+	L_2	+
λ-free *hom*	—	—	?	?	+	+
λ-free *gsm*	—	—	?	?	+	+
λ-free substitution	—	—	?	?	+	+
hom	—	—	—	—	—	—

The concatenation closure for *unary* real-time CA languages has been solved in the affirmative [33].

Table 1 summarizes some closure properties of the language families in question.

Decidability Problems

It is well known that all nontrivial decidability problems for Turing machines are undecidable [69]. Moreover, many of them are not even semidecidable, e. g., neither finiteness nor infiniteness. Now we turn to explore undecidable properties for cellular automata and iterative arrays. Most of the early results are shown in [73] by reductions of the the Post Correspondence Problem. In terms of trellis automata the undecidability of emptiness, equivalence, and universality is derived in [22]. Here we present improved results that show the *non-semidecidability* of the

properties. Almost all results in this section are obtained by Andreas Malcher in [58,59].

In [30] large Turing machine computations have been encoded in small grammars. These encodings and variants thereof are of tangible advantage for our purposes. To this end, we consider *valid computations of Turing machines*. Roughly speaking, these are histories of accepting Turing machine computations. It suffices to consider deterministic Turing machines with a single tape and a single read-write head. Without loss of generality and for technical reasons, one can assume that any accepting computation has at least three and, in general, an odd number of steps. Therefore, it is represented by an even number of configurations. Moreover, it is assumed that the Turing machine cannot print blanks, and that a configuration is halting if and only if it is accepting.

Let S be the state set of some Turing machine $\mathcal{M}$, where s_0 is the initial state, $T \cap S = \emptyset$ is the tape alphabet containing the blank symbol, $A \subset T$ is the input alphabet, and $F \subseteq S$ is the set of accepting states. Then a configuration of $\mathcal{M}$ can be written as a word of the form T^*ST^* such that $t_1 \cdots t_i\, s\, t_{i+1} \cdots t_n$ is used to express that $\mathcal{M}$ is in state s, scanning tape symbol t_{i+1}, and t_1 to t_n is the support of the tape inscription. The set of valid computations VALC($\mathcal{M}$) is now defined to be the set of words of the form $w_1 \$ w_3 \$ \cdots \$ w_{2k-1} ¢\, w_{2k}^R \$ \cdots \$ w_4^R \$ w_2^R$, where w_i are configurations, \$ and ¢ are symbols not appearing in w_i, w_1 is an initial configuration of the form s_0A^*, w_{2k} is an accepting configuration of the form $T^*\ FT^*$, and w_{i+1} is the successor configuration of w_i, for $1 \le i \le 2k$. The set of *invalid computations* INVALC($\mathcal{M}$) is the complement of VALC($\mathcal{M}$) with respect to the coding alphabet $\{\$, ¢\} \cup T \cup S$. The following lemma shows some of the important properties of valid computations.

Lemma 52 *Let $\mathcal{M}$ be some Turing machine.*

1. *$L(\mathcal{M})$ is empty if and only if VALC($\mathcal{M}$) is empty.*
2. *$L(\mathcal{M})$ is finite if and only if VALC($\mathcal{M}$) is finite.*
3. *$L(\mathcal{M})$ is finite if and only if VALC($\mathcal{M}$) is context free.*
4. *$L(\mathcal{M})$ is finite if and only if INVALC($\mathcal{M}$) is regular.*
5. *VALC($\mathcal{M}$) can be represented by the intersection of two real-time deterministic, linear context-free languages, such that both deterministic pushdown automata and both linear context-free grammars can effectively be constructed from $\mathcal{M}$.*
6. *INVALC($\mathcal{M}$) is a linear context-free language, such that its grammar can effectively be constructed from $\mathcal{M}$.*

Proof Assertions 1 and 2 are immediate observations. In order to show assertion 3 assume that $L(\mathcal{M})$ is finite. Then VALC($\mathcal{M}$) is finite and, clearly, context free. If conversely $L(\mathcal{M})$ is infinite, then an application of the pumping lemma shows that VALC($\mathcal{M}$) is not context free [30]. Now, assertion 4 is shown as follows. If $L(\mathcal{M})$ is finite, then VALC($\mathcal{M}$) is finite. Therefore, INVALC($\mathcal{M}$) is co-finite and, thus, regular. Conversely, if INVALC($\mathcal{M}$) is regular, then VALC($\mathcal{M}$) is regular since the regular languages are closed under complementation. Therefore, VALC($\mathcal{M}$) is context free which implies that $L(\mathcal{M})$ is finite. The two deterministic, linear context-free languages for assertion 5 are constructed in [4]. Assertion 6 has been shown in [30] for a similar definition of invalid computations. The proof can easily be adapted. □

Lemma 53 *Let $\mathcal{M}$ be a Turing machine. Then both languages VALC($\mathcal{M}$) and INVALC($\mathcal{M}$) are accepted by real-time OCAs as well as by real-time IAs.*

Proof Given some Turing machine $\mathcal{M}$, by Lemma 52 item 6 we can effectively construct a linear context-free grammar for the language INVALC($\mathcal{M}$). Due to Theorem 19, a real-time OCA accepting INVALC($\mathcal{M}$) can be constructed from the grammar. Since $\mathcal{L}_{\mathrm{rt}}$(OCA) is effectively closed under complementation, we obtain a real-time OCA that accepts the valid computations of $\mathcal{M}$.

Similarly, by Lemma 52 item 5 we can effectively construct two real-time deterministic pushdown automata whose intersection represents the language VALC($\mathcal{M}$). Due to Theorem 4, two real-time IAs can be constructed from the pushdown automata. Since the family $\mathcal{L}_{\mathrm{rt}}$(IA) is effectively closed under intersection and complementation, we obtain real-time IAs that accept the valid and the invalid computations of $\mathcal{M}$. □

Now we are prepared to reduce the finiteness and infiniteness problems of Turing machines to some of the decidability problems in question.

Theorem 54 *For any language family that effectively contains $\mathcal{L}_{rt}$(OCA) or $\mathcal{L}_{rt}$(IA), emptiness, universality, finiteness, infiniteness, context-freeness, and regularity are not semidecidable.*

Proof Given some Turing machine $\mathcal{M}$, by Lemma 53 we can effectively construct a real-time OCA and a real-time IA for the language VALC($\mathcal{M}$). Therefore, by Lemma 51 item 1, if emptiness were semidecidable for real-time OCAs or IAs, then emptiness is semidecidable for Turing machines, too.

Since $L(\mathcal{M}) = A^*$ if and only if the complement of $L(\mathcal{M})$ is empty, the non-semidecidability of universality follows from the effective closure under complementation and the non-semidecidability of emptiness.

In the same way as emptiness and universality the non-semidecidability of finiteness and infiniteness follows.

Since we have constructed real-time OCAs and IAs for INVALC($\mathcal{M}$) as well as for VALC($\mathcal{M}$), the finiteness problem for Turing machines immediately reduces to the context-freeness and to the regularity problem for $\mathcal{L}_{\mathrm{rt}}$(OCA) and $\mathcal{L}_{\mathrm{rt}}$(IA) by Lemma 52 items 3 and 4. □

Theorem 55 *For any language family that effectively contains $\mathcal{L}_{rt}(OCA)$ or $\mathcal{L}_{rt}(IA)$ equivalence and inclusion are not semidecidable.*

Proof It is easy to construct a real-time OCA and a real-time IA that accept the empty language. So, the semidecidability of equivalence would imply the semidecidability of emptiness. Since $L(\mathcal{M}) = L(\mathcal{M}')$ if and only if $L(\mathcal{M}) \subseteq L(\mathcal{M}')$ and $L(\mathcal{M}') \subseteq L(\mathcal{M})$, inclusion is not semidecidable either. □

Next the question arises whether some structural properties of cellular language acceptors are (semi)decidable. For example, whether or not a real-time two-way language is a real-time one-way language. We also compare sequential input mode and two-way information flow with parallel input mode and one-way information flow from a decidability point of view. The questions turn out to be not even semidecidable. So the resources inherent in the structures of cellular automata seem to be fairly different.

Let $\mathcal{M}$ be some Turing machine. We consider the language

$$L_{\mathcal{M}} = \{w^{|w|!} \mid w \in \triangleright\mathrm{VALC}(\mathcal{M})\triangleleft\} ,$$

where $\triangleright$ and $\triangleleft$ are new symbols not appearing in VALC($\mathcal{M}$), and deal with its acceptance.

Lemma 56 *Given some Turing machine $\mathcal{M}$, a real-time IA accepting $L_{\mathcal{M}}$ can effectively be constructed from $\mathcal{M}$, i. e., the language $L_{\mathcal{M}}$ belongs to $\mathcal{L}_{rt}(IA)$.*

Proof We set B to be the alphabet of VALC($\mathcal{M}$), and represent $L_{\mathcal{M}}$ as intersection of the three languages

$$L_{\mathcal{M},1} = \{w \mid w \in (\triangleright\mathrm{VALC}(\mathcal{M})\triangleleft)^*\} ,$$
$$L_{\mathcal{M},2} = \{w^i \mid w \in \triangleright B^*\triangleleft,\ i \geq 2,\ i \quad \text{even}\}, \text{and}$$
$$L_{\mathcal{M},3} = \{wu \mid w \in \triangleright B^*\triangleleft,\ u \in (\triangleright B^*\triangleleft)^*, |wu|_{\triangleright} = |w|!\}.$$

Since $\mathcal{L}_{\mathrm{rt}}$(IA) is closed under intersection, it remains to be shown that each of the languages is accepted by some real-time IA. Language $L_{\mathcal{M},1}$ is the marked iteration of $\triangleright$VALC($\mathcal{M}$) followed by a single $\triangleleft$, where $\triangleleft$ is the marking symbol. Since VALC($\mathcal{M}$) $\in \mathcal{L}_{\mathrm{rt}}$(IA) and due to its closure under concatenation with single symbols and under marked iteration [73], we obtain $L_{\mathcal{M},1} \in \mathcal{L}_{\mathrm{rt}}$(IA).

The copy language $\{vv \mid v \in B^*\}$ belongs to $\mathcal{L}_{\mathrm{rt}}$(IA) [19]. So, $L = \{vv \mid v \in \triangleright B^*\triangleleft\}$ is accepted by some real-time IA. Since $\mathcal{L}_{\mathrm{rt}}$(IA) is closed under marked iteration and intersection, the languages L^* and $\triangleright B^* \triangleleft L^* \triangleright B^*\triangleleft$ as well as their intersection are accepted by some real-time IA. Here the marking is hidden in a regular structure. The new word starts after the second symbol $\triangleleft$, respectively. Clearly, the intersection equals $L_{\mathcal{M},2}$.

The construction of a real-time IA accepting the language $L_{\mathcal{M},3}$ makes use of the IA-constructibility of the factorials (cf. Example 1 and the discussion before Theorem 26). Recall, that this means that there is an IA which indicates by states of the leftmost cell the time steps $n!$, for $n \geq 1$. For further results on IA-constructibility we refer to [13,27,63].

Since all constructions and closures are effective, the assertion follows. □

Lemma 57 *Let $\mathcal{M}$ be some Turing machine. Then $L_{\mathcal{M}}$ belongs to $\mathcal{L}_{rt}(OCA)$ if and only if $L(\mathcal{M})$ is finite.*

Proof We observe that a finite language is regular and that the regular languages are accepted by real-time OCAs. Conversely, if $L(\mathcal{M})$ is infinite, then we apply Lemma 8 in order to show $L_{\mathcal{M}} \notin \mathcal{L}_{\mathrm{rt}}$(OCA). Assume the contrary, and let p be the constant of Lemma 8. Clearly, due to the infinity of $L(\mathcal{M})$ there is some $w \in \triangleright\mathrm{VALC}(\mathcal{M})\triangleleft$ such that $|w|! > p^{|w|}$. We conclude $w^{|w|!} \in L_{\mathcal{M}}$, and the conditions of Lemma 8 are met with $k = |w|!$. Therefore, there is some $1 \leq q \leq p^{|w|}$ such that $w^{|w|!+q} \in L_{\mathcal{M}}$. But $|w|! < |w|!+q < (|w|+1)!$ and, thus, $|w|!+q$ is not a factorial which implies the contradiction $w^{|w|!+q} \notin L_{\mathcal{M}}$. □

Now we may obtain the next (un)decidability result.

Theorem 58 *For any language family $\mathcal{L}$ that effectively contains $\mathcal{L}_{rt}(IA)$, it is not semidecidable whether $L \in \mathcal{L}$ is a real-time OCA language.*

Proof If the problem in question were semidecidable, then the finiteness for Turing machines is also semidecidable. To this end, given some Turing machine $\mathcal{M}$ we construct a real-time IA for the language $L_{\mathcal{M}}$ according to Lemma 56. If $L_{\mathcal{M}}$ is accepted by some real-time OCA, then $L_{\mathcal{M}}$ is finite by Lemma 57. This implies the finiteness of VALC($\mathcal{M}$) and, thus, the finiteness of $L(\mathcal{M})$. □

Since the families $\mathcal{L}_{\mathrm{rt}}$(OCA) and $\mathcal{L}_{\mathrm{rt}}$(IA) are incomparable with respect to set inclusion, there is a natural interest to know whether the incomparability is also with respect to the decidability in question. Therefore, we turn to the converse question of Theorem 58, i. e., whether it is (semi)decidable that a real-time OCA language is accepted by some real-time IA. First we need some preliminaries.

A Turing machine $\mathcal{M}$ is converted into a Turing machine $\mathcal{M}'$ such that the input alphabet A' of $\mathcal{M}'$ contains at least two symbols and, furthermore, $\mathcal{M}'$ accepts any input of length n if and only if there is at least one input of length n accepted by $\mathcal{M}$. Clearly, this conversion is always effectively possible. Extending a frequently used witness language, we set

$$L'_{\mathcal{M}} = \{\&x_k\&\cdots\&x_1?y_1\&\cdots\&y_k\& \mid k \geq 1, x_i^R = y_i z_i \text{ and } y_i z_i \in (A')^* \text{ and } x_i \in \mathrm{VALC}(\mathcal{M}')^R\},$$

where & and ? are new symbols not appearing in $\mathrm{VALC}(\mathcal{M}')$.

Lemma 59 *Let $\mathcal{M}$ be some Turing machine. Then $L'_{\mathcal{M}}$ belongs to $\mathcal{L}_{rt}(IA)$ if and only if $L(\mathcal{M})$ is finite.*

Proof If $L(\mathcal{M})$ is finite, then so is $L(\mathcal{M}')$ and, thus, $\mathrm{VALC}(\mathcal{M}')^R$ is finite, say $\mathrm{VALC}(\mathcal{M}')^R = \{v_1, v_2, \ldots, v_r\}$. A real-time deterministic pushdown automaton accepting $L'_{\mathcal{M}}$ has r different stack symbols representing the elements in $\{v_1, v_2, \ldots, v_r\}$. It reads the input until the ? appears. For any occurring $v \in \mathrm{VALC}(\mathcal{M}')^R$ the corresponding stack symbol is pushed onto the stack. After reading the ?, the pushdown automaton matches each y_i with the suffix of the $v \in \mathrm{VALC}(\mathcal{M}')^R$ which is identified by the symbol at the top of the stack. By Theorem 4, we obtain $L'_{\mathcal{M}} \in \mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$.

Now let $L(\mathcal{M})$ be infinite. Then $L(\mathcal{M}')$, $\mathrm{VALC}(\mathcal{M}')^R$, and $L'_{\mathcal{M}}$ are infinite, as well. In order to show that in this case $L'_{\mathcal{M}}$ does not belong to $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$ we apply Lemma 6 as follows. Assume in contrast to the assertion that $L'_{\mathcal{M}}$ is accepted by some real-time IA with state set S. Every $v \in \mathrm{VALC}(\mathcal{M}')^R$ has a suffix of the form $\$\, u\, s_0$, where s_0 is the initial state and $u = input(v)$ is the reversal of the input. Let $|u| = k$. Moreover, due to the construction of $\mathcal{M}'$, for every input word with length k there is an element in $\mathrm{VALC}(\mathcal{M}')^R$. We denote the set of these elements by $V(k)$ and conclude $|V(k)| = |A'|^k$. For two different prefixes $w = \&\, x_k\, \&\cdots\&\, x_1\, ?$ and $w' = \&\, x'_k\, \&\cdots\&\, x'_1\, ?$ with $x_i, x'_i \in V(k)$, $1 \leq i \leq k$, there exists at least one $1 \leq j \leq k$ such that $x_j \neq x'_j$. Therefore, $w\, \&^{j-1}\, s_0\, input(x_j)^R\, \&^{k-j+1} \in L'_{\mathcal{M}}$ and $w'\, \&^{j-1}\, s_0\, input(x_j)^R\, \&^{k-j+1} \notin L'_{\mathcal{M}}$. Since the number of such prefixes is $|A'|^{k^2}$ and $|A'| \geq 2$, we obtain at least 2^{k^2} different $2k$-equivalence classes with respect to $L'_{\mathcal{M}}$. On the other hand, there is a constant $p \geq 1$ such that $E(L'_{\mathcal{M}}, 2k) \leq p^{2k}$. Since $L(\mathcal{M})$ is infinite, we may choose k large enough such that $2^{k^2} > p^{2k}$, which is a contradiction. □

Lemma 60 *Given some Turing machine $\mathcal{M}$, a real-time OCA accepting $L'_{\mathcal{M}}$ can effectively be constructed from $\mathcal{M}$, i. e., the language $L'_{\mathcal{M}}$ belongs to $\mathcal{L}_{rt}(OCA)$.*

Proof The language $L'_{\mathcal{M}}$ can be represented as the intersection of L_1 and L_2, where $L_1 = \{\&x_k\&\cdots\&x_1?y_1\&\cdots\&y_k\& \mid k \geq 1, x_i^R = y_i z_i \text{ and } x_i, y_i, z_i \in (A')^*\}$ and $L_2 = (\&\mathrm{VALC}(\mathcal{M}')^R)^*?((A')^*\&)^*$. Since L_1 is a linear context-free language, it belongs to $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$. The family $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$ contains $\mathrm{VALC}(\mathcal{M}')$, is closed under reversal, marked iteration and right concatenation with regular sets [73]. Therefore, L_2 belongs to $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$, as well. From its closure under intersection we derive $L'_{\mathcal{M}} \in \mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$. □

Similarly to Theorem 58 we obtain the next undecidability of a structural property.

Theorem 61 *For any language family $\mathcal{L}$ that effectively contains $\mathcal{L}_{rt}(OCA)$ it is not semidecidable whether $L \in \mathcal{L}$ is a real-time IA language.*

Proof If the problem in question were semidecidable, then also the finiteness for Turing machines. To this end, given some Turing machine $\mathcal{M}$, we construct a real-time OCA for the language $L'_{\mathcal{M}}$ according to Lemma 60. If $L'_{\mathcal{M}}$ is accepted by some real-time IA, then $L(\mathcal{M})$ is finite by Lemma 59. □

In general, a family $\mathcal{L}$ of languages possesses a *pumping lemma in the narrow sense* if for each $L \in \mathcal{L}$ there exists a constant $n \geq 1$ computable from L such that each $z \in L$ with $|z| > n$ admits a factorization $z = uvw$, where $|v| \geq 1$ and $u'v^i w' \in L$, for infinitely many $i \geq 0$. The prefix u' and the suffix w' depend on u, w and i.

Theorem 62 *Any language family whose word problem is semidecidable and that effectively contains $\mathcal{L}_{rt}(OCA)$ or $\mathcal{L}_{rt}(IA)$ does not possess a pumping lemma (in the narrow sense).*

Proof Let $\mathcal{M}$ be a real-time OCA or IA and assume there is a pumping lemma. Clearly, $L(\mathcal{M})$ is infinite if and only if it contains some w with $|w| > n$. So, we can semidecide infiniteness by first computing n and then verifying for all words longer than n whether they belong to $L(\mathcal{M})$. If at least for one word the answer is in the affirmative, then by pumping infinitely many words belong to $L(\mathcal{M})$. □

Example 63 The families $\mathcal{L}_{\mathrm{rt}}(\mathrm{OCA})$ and $\mathcal{L}_{\mathrm{rt}}(\mathrm{IA})$ itself as well as, e. g., the families $\mathcal{L}_{\mathrm{rt}}(\mathrm{CA})$, $\mathcal{L}_{\mathrm{lt}}(\mathrm{OCA})$, $\mathcal{L}_{\mathrm{lt}}(\mathrm{CA}) = \mathcal{L}_{\mathrm{lt}}(\mathrm{IA})$, and $\mathcal{L}(\mathrm{CA}) = \mathcal{L}(\mathrm{IA}) = \mathrm{DSPACE}(n)$ do not possess a pumping lemma. □

Theorem 64 *There is no minimization algorithm converting some CA, OCA or IA with arbitrary time complexity to an equivalent automaton of the same type with a minimal number of states.*

Proof For a given input alphabet A, we consider a minimal CA or OCA accepting the empty language. It has $|A|$ states and no accepting states. Assume there is a minimization algorithm. Then we can minimize an arbitrary CA and OCA and check whether the result has $|A|$ states and no accepting states. In this case the accepted language is empty. If the minimal automaton has $|A|$ states and at least one accepting state, there is an input such that the leftmost cell is initially accepting. So, the input is accepted and the accepted language is not empty. Hence emptiness is decidable, which is a contradiction to Theorem 54. Similar arguments apply for IAs. □

It remains to be mentioned that there is a nontrivial decidable property of (unbounded) cellular automata. It is known that injectivity of the global transition function is equivalent to the reversibility of the automaton. It is shown in [3] that global reversibility is decidable for one-dimensional CAs, whereas the problem is undecidable for higher dimensions [41].

Future Directions

The investigation of cellular language acceptors obeying a linear space bound reveals the hierarchy of language families in between the regular and the deterministic context-sensitive languages established in Sect. "Computational Capacities" (see Fig. 11). If the space bound is omitted, that is, if there is a potentially unlimited number of cells, then computation universality is achieved by direct simulation of Turing machines [78]. In particular, the universality can be achieved in spite of additional structural and computational limitations [2,60,64,66]. Similarly, some space bound supposed for cellular language acceptors does not yield to new language families. A Turing machine sweeping back and forth over the nonempty part of the tape can simulate the parallel device obeying the same space bound.

On the other hand, if the cellular language acceptor is simultaneously $s(n)$ space and $t(n)$ time bounded, a Turing machine simulation takes $s(n) \cdot t(n)$ time. A challenging question for further investigations is to identify languages and language classes for which homogeneously structured massive parallelism can significantly decrease the time complexity of sequential devices. Of particular interest are languages which allow a maximal saving. That is, for a sequential time complexity $t(n)$, the parallel time complexity is bounded by $t(n)/s(n)$, where $s(n)$ is the parallel space complexity. For example, in case of unary real-time languages, OCAs cannot do better than deterministic finite-state machine. Conversely, it is well known that any one-tape Turing machine takes at least $O(n^2)$ time to accept the language of palindromes $\{w \mid w = w^R, w \in \{a, b\}^*\}$. Since it is a linear context-free language, it is accepted by some real-time OCA, achieving the maximal saving in time.

From a more general point of view, central questions for future studies concern the power of additional limited resources at the disposal of time or space bounded computations. For example, nondeterminism, dimensions, the number of bits communicated to neighboring cells, or the restriction to reversible computations, all these can be seen as limited resources. We discuss some approaches in more detail.

Traditionally, nondeterministic devices have been viewed as having as many nondeterministic guesses as time steps. The studies of this concept of unlimited nondeterminism led, for example, to the famous open LBA-problem or the unsolved question whether or not P equals NP. In order to gain further understanding of the nature of nondeterminism, in [28,44] it has been viewed as an additional limited resource. In [11,46,52] cellular automata started to be considered from this point of view.

In classical computations the states of the neighboring cells are communicated in one time step. That is, the number of bits exchanged is determined by the number of states. A natural and interesting restriction is to limit the number of bits to some constant being independent of the number of states. Iterative arrays with restricted inter-cell communication have been investigated in [93,94], where algorithmic design techniques for sequence generation are shown. In particular, several important infinite, non-regular sequences such as exponential or polynomial, Fibonacci and prime sequences can be generated in real time. Connectivity recognition problems are dealt with in [92], whereas in [99] the computational capacity of one-way cellular automata with restricted inter-cell communication is considered. First results concerning formal language aspects of IAs with restricted inter-cell communication are shown in [53,54].

Finally, we turn to reversibility. Reversibility in the context of computing devices means that deterministic computations are also backward deterministic. Roughly speaking, in a reversible device no information is lost and every configuration occurring in any computation has at most one predecessor. Many different formal models have been studied in connection with reversibility. An early result on general reversible CAs is the possibility to make any CA, possibly irreversible, reversible by increasing the dimension. In detail, in [91] it is shown that any k-dimensional CA can be embedded into a $(k+1)$-dimensional reversible CA. This result has significantly been improved by showing how to make irreversible one-dimensional CAs

reversible without increasing the dimension [65]. Furthermore, it is known that even reversible one-dimensional one-way CAs are computationally universal [64,66]. These results concern cellular automata with unbounded space. Moreover, in order to obtain a reversible device the neighborhood as well as the time complexity may be increased. In [23] it is shown that the neighborhood of a reversible CA is at most $n-1$ when the given reversible CA has n states. Additionally, this upper bound is shown to be tight. Cellular language acceptors with bounded space that are reversible on the core of computation, that is, from initial configuration to the configuration given by the time complexity, are introduced in [55,56]. At first glance, such a setting should simplify matters. However, it is quite the contrary, and such real-time reversibility is undecidable. There are many properties and relations still to be discovered in this setting.

Bibliography

Primary Literature

1. Achilles AC, Kutrib M, Worsch T (1996) On relations between arrays of processing elements of different dimensionality. In: Parallel processing by cellular automata and arrays (PARCELLA 1996). Mathematical Research 96. Akademie Verlag, Berlin, pp 13–20
2. Albert J, Čulik II K (1987) A simple universal cellular automaton and its one-way and totalistic version. Complex Systems 1: 1–16
3. Amoroso S, Patt YN (1972) Decision procedures for surjectivity and injectivity of parallel maps for tesselation structures. J Comput System Sci 6:448–464
4. Baker BS, Book RV (1974) Reversal-bounded multipushdown machines. J Comput System Sci 8:315–332
5. Bleck B, Kröger H (1992) Cellular algorithms. In: Advances in parallel computing, vol 2. JAI Press, London, pp 115–143
6. Bucher W, Čulik II K (1984) On real time and linear time cellular automata. RAIRO Inform Théor 18:307–325
7. Buchholz T, Klein A, Kutrib M (1998) On time reduction and simulation in cellular spaces. Int J Comput Math 71:459–474
8. Buchholz T, Klein A, Kutrib M (1999) Iterative arrays with a wee bit alternation. In: Fundamentals of Computation Theory (FCT 1999). LNCS, vol 1684. Springer, Berlin, pp 173–184
9. Buchholz T, Klein A, Kutrib M (2000) Iterative arrays with small time bounds. In: Mathematical Foundations of Computer Science (MFCS 1998). LNCS, vol 1893. Springer, Berlin, pp 243–252
10. Buchholz T, Klein A, Kutrib M (2000) Real-time language recognition by alternating cellular automata. In: Theoretical Computer Science (TCS 2000). LNCS, vol 1872. Springer, Berlin, pp 213–225
11. Buchholz T, Klein A, Kutrib M (2002) On interacting automata with limited nondeterminism. Fund Inform 52:15–38
12. Buchholz T, Klein A, Kutrib M (2003) Iterative arrays with limited nondeterministic communication cell. In: Words, Languages and Combinatorics III (WLC 2000), World Scientific Publishing, Singapore, pp 73–87
13. Buchholz T, Kutrib M (1997) Some relations between massively parallel arrays. Parallel Comput 23(11):1643–1662
14. Buchholz T, Kutrib M (1998) On time computability of functions in one-way cellular automata. Acta Inf 35:329–352
15. Chang JH, Ibarra OH, Palis MA (1987) Parallel parsing on a one-way array of finite-state machines. IEEE Trans Comp C-36: 64–75
16. Chang JH, Ibarra OH, Vergis A (1988) On the power of one-way communication. J ACM 35:697–726
17. Choffrut C, Čulik II K (1984) On real-time cellular automata and trellis automata. Acta Inf 21:393–407
18. Cole SN (1966) Real-time computation by n-dimensional iterative arrays of finite-state machines. In: IEEE Symposium on Switching and Automata Theory (SWAT 1966). IEEE Press, New York, pp 53–77
19. Cole SN (1969) Real-time computation by n-dimensional iterative arrays of finite-state machines. IEEE Trans Comp C-18(4): 349–365
20. Čulik II K, Fris I (1985) Topological transformations as a tool in the design of systolic networks. Theoret Comp Sci 37:183–216
21. Čulik II K, Gruska J, Salomaa A (1984) Systolic trellis automata I. Int J Comp Math 15:195–212
22. Čulik II K, Gruska J, Salomaa A (1986) Systolic trellis automata: Stability, decidability and complexity. Inf Control 71:218–230
23. Czeizler E, Kari J (2005) A tight linear bound on the neighborhood of inverse cellular automata. In: Automata, Languages and Programming (ICALP 2005). LNCS, vol 3580. Springer, Berlin, pp 410–420
24. Delorme M, Mazoyer J (2004) Real-time recognition of languages on an two-dimensional archimedean thread. Theor Comp Sci 322:335–354
25. Dubacq JC, Terrier V (2002) Signals for cellular automata in dimension 2 or higher. In: Theoretical Informatics (LATIN 2002). LNCS, vol 2286. Springer, Berlin, pp 451–464
26. Dyer CR (1980) One-way bounded cellular automata. Inf Control 44:261–281
27. Fischer PC (1965) Generation of primes by a one-dimensional real-time iterative array. J ACM 12:388–394
28. Fischer PC, Kintala CMR (1979) Real-time computations with restricted nondeterminism. Math. Syst Theory 12:219–231
29. Ginsburg S, Greibach SA, Harrison MA (1967) One-way stack automata. J ACM 14:389–418
30. Hartmanis J (1967) Context-free languages and Turing machine computations. Proc Symp App Math 19:42–51
31. Höllerer WO, Vollmar R (1975) On forgetful cellular automata. J Comp Syst Sci 11:237–251
32. Ibarra OH, Jiang T (1987) On one-way cellular arrays. SIAM J Comp 16:1135–1154
33. Ibarra OH, Jiang T (1988) Relating the power of cellular arrays to their closure properties. Theor Comp Sci 57:225–238
34. Ibarra OH, Kim SM, Moran S (1985) Sequential machine characterizations of trellis and cellular automata and applications. SIAM J Comp 14:426–447
35. Ibarra OH, Kim SM (1984) Characterizations and computational complexity of systolic trellis automata. Theor Comp Sci 29:123–153
36. Ibarra OH, Palis MA (1985) Some results concerning linear iterative (systolic) arrays. J Paral Distrib Comp 2:182–218
37. Ibarra OH, Palis MA (1988) Two-dimensional iterative arrays: Characterizations and applications. Theor Comp Sci 57:47–86
38. Ibarra OH, Palis MA, Kim SM (1985) Fast parallel language

recognition by cellular automata. Theor Comp Sci 41(2–3):231–246
39. Imai K, Morita K (1996) Firing squad synchronization problem in reversible cellular automata. Theor Comp Sci 165(2):475–482
40. Iwamoto C, Hatsuyama T, Morita K, Imai K (2002) Constructible functions in cellular automata and their applications to hierarchy results. Theor Comp Sci 270:797–809
41. Kari J (1994) Reversibility and surjectivity problems of cellular automata. J Comp Syst Sci 48(1):149–182
42. Kasami T, Fuji M (1968) Some results on capabilities of one-dimensional iterative logical networks. Elect Commun Japan 51-C:167–176
43. Kim S, McCloskey R (1990) A characterization of constant-time cellular automata computation. Phys D 45:404–419
44. Kintala CMR (1977) Computations with a Restricted Number of Nondeterministic Steps. Ph D thesis, Pennsylvania State University, University Park
45. Klein A, Kutrib M (2003) Fast one-way cellular automata. Theor Comp Sci 1–3:233–250
46. Klein A, Kutrib M (2007) Cellular devices and unary languages. Fund Inf 78:343–368
47. Kosaraju SR (1975) Speed of recognition of context-free languages by array automata. SIAM J Comp 4:331–340
48. Krithivasan K, Mahajan M (1995) Nondeterministic, probabilistic and alternating computations on cellular array models. Theor Comp Sci 143:23–49
49. Kutrib M (1999) Pushdown cellular automata. Theor Comp Sci 215(1–2):239–261
50. Kutrib M (2001) Efficient universal pushdown cellular automata and their application to complexity. In: Machines, Computations, and Universality (MCU 2001). LNCS, vol 2055. Springer, Berlin, pp 252–263
51. Kutrib M, Löwe JT (2002) Massively parallel fault tolerant computations on syntactical patterns. Fut Gener Comp Syst 18:905–919
52. Kutrib M, Löwe JT (2003) Space- and time-bounded nondeterminism for cellular automata. Fund Inf 58(2003):273–293
53. Kutrib M, Malcher A (2006) Fast cellular automata with restricted inter-cell communication: Computational capacity. In: Theoretical Computer Science (IFIP TCS2006). IFIP 209. Springer, Berlin, pp 151–164
54. Kutrib M, Malcher A (2006) Fast iterative arrays with restricted inter-cell communication: Constructions and decidability. In: Mathematical Foundations of Computer Science (MFCS 2006). LNCS, vol 4162. Springer, Berlin, pp 634–645
55. Kutrib M, Malcher A (2008) Fast reversible language recognition using cellular automata. Inform Comput 206(9–10):1142–1151
56. Kutrib M, Malcher A (2007) Real-time reversible iterative arrays. In: Fundamentals of Computation Theory 2007 (FCT 2007). LNCS, vol 4693. Springer, Berlin, pp 376–387
57. Kutrib M, Worsch T (1994) Investigation of different input modes for cellular automata. In: Parallel Processing by Cellular Automata and Arrays (PARCELLA 1994). Mathematical Research 81, Akademie Verlag, Berlin, pp 141–150
58. Malcher A (2002) Descriptional complexity of cellular automata and decidability questions. J Autom Lang Comb 7:549–560
59. Malcher A (2004) On the descriptional complexity of iterative arrays. IEICE Trans Inf Syst E87-D(3):721–725
60. Martin B (1994) A universal cellular automaton in quasi-linear time and its S-m-n form. Theor Comp Sci 123(2):199–237
61. Matamala M (1997) Alternation on cellular automata. Theor Comp Sci 180:229–241
62. Mazoyer J (1987) A six-state minimal time solution to the firing squad synchronization problem. Theor Comp Sci 50:183–238
63. Mazoyer J, Terrier V (1999) Signals in one-dimensional cellular automata. Theor Comp Sci 217:53–80
64. Morita K (1992) Computation-universality of one-dimensional one-way reversible cellular automata. Inf Proc Lett 42:325–329
65. Morita K (1995) Reversible simulation of one-dimensional irreversible cellular automata. Theor Comp Sci 148:157–163
66. Morita K, Harao M (1989) Computation universality of one dimensional reversible injective cellular automata. Trans IEICE E72 pp 758–762
67. Morita K, Ueno S (1994) Parallel generation and parsing of array languages using reversible cellular automata. Int J Pattern Recog Artif Int 8:543–561
68. Nakamura K (1999) Real-time language recognition by one-way and two-way cellular automata. In: Mathematical Foundations of Computer Science (MFCS 1999). LNCS, vol 1672. Springer, Berlin, pp 220–230
69. Rice HG (1953) Classes of recursively enumerable sets and their decision problems. Trans Amer Math Soc 89:25–59
70. Rosenfeld A (1979) Picture Languages. Academic Press, New York
71. Rosenstiehl P, Fiksel JR, Holliger A (1972) Intelligent graphs: Networks of finite automata capable of solving graph problems. In: Graph Theory and Computing. Academic Press, New York, pp 219–265
72. Salomaa A (1973) Formal Languages. Academic Press, Orlando
73. Seidel SR (1979) Language recognition and the synchronization of cellular automata. Technical Report 79–02, Department of Computer Science, University of Iowa, Iowa City
74. Seiferas JI (1977) Iterative arrays with direct central control. Acta Inf 8:177–192
75. Seiferas JI (1977) Linear-time computation by nondeterministic multidimensional iterative arrays. SIAM J Comp 6:487–504
76. Smith III AR (1970) Cellular automata and formal languages. In: IEEE Symposium on Switching and Automata Theory (SWAT 1970). IEEE Press, New York, pp 216–224
77. Smith III AR (1971) Cellular automata complexity trade-offs. Inf Control 18:466–482
78. Smith III AR (1971) Simple computation–universal cellular spaces. J ACM 18:339–353
79. Smith III AR (1971) Two-dimensional formal languages and pattern recognition by cellular automata. In: IEEE Symposium on Switching and Automata Theory (SWAT 1971). IEEE Press, New York, pp 144–152
80. Smith III AR (1972) Real-time language recognition by one-dimensional cellular automata. J Comp Syst Sci 6:233–253
81. Smith III AR (1976) Introduction to and survey of polyautomata theory. In: Automata, Languages, Development. North-Holland, Amsterdam, pp 405–422
82. Sommerhalder R, van Westrhenen SC (1983) Parallel language recognition in constant time by cellular automata. Acta Inf 19:397–407
83. Terrier V (1994) Language recognizable in real time by cellular automata. Complex Syst 8:325–336
84. Terrier V (1995) On real time one-way cellular array. Theor Comp Sci 141:331–335

85. Terrier V (1996) Language not recognizable in real time by one-way cellular automata. Theor Comp Sci 156:281–287
86. Terrier V (1999) Two-dimensional cellular automata recognizer. Theor Comp Sci 218:325–346
87. Terrier V (2003) Two-dimensional cellular automata and deterministic on-line tesselation automata. Theor Comp Sci 301:167–187
88. Terrier V (2004) Two-dimensional cellular automata and their neighborhoods. Theor Comp Sci 312:203–222
89. Terrier V (2006) Closure properties of cellular automata. Theor Comp Sci 352:97–107
90. Terrier V (2006) Low complexity classes of multidimensional cellular automata. Theor Comp Sci 369:142–156
91. Toffoli T (1977) Computation and construction universality of reversible cellular automata. J Comp Syst Sci 15:213–231
92. Umeo H (2001) Linear-time recognition of connectivity of binary images on 1-bit inter-cell communication cellular automaton. Parallel Comp 27:587–599
93. Umeo H, Kamikawa N (2002) A design of real-time non-regular sequence generation algorithms and their implementations on cellular automata with 1-bit inter-cell communications. Fund Inf 52:257–275
94. Umeo H, Kamikawa N (2003) Real-time generation of primes by a 1-bit-communication cellular automaton. Fund Inf 58:421–435
95. Umeo H, Morita K, Sugata K (1982) Deterministic one-way simulation of two-way real-time cellular automata and its related problems. Inf Proc Lett 14:158–161
96. Vollmar R (1981) On cellular automata with a finite number of state changes. Computing 3:181–191
97. von Neumann J (1966) Theory of Self-Reproducing Automata. In: Burks AW (ed) University of Illinois Press, Champaign
98. Waksman A (1966) An optimum solution to the firing squad synchronization problem. Inf Control 9:66–78
99. Worsch T (2000) Linear time language recognition on cellular automata with restricted communication. In: Theoretical Informatics (LATIN 2000). LNCS, vol 1776. Springer, Berlin, pp 417–426

Books and Reviews

Delorme M, Mazoyer J (eds) (1999) Cellular Automata – A Parallel Model. Kluwer, Dordrecht

Cellular Automata with Memory

Ramón Alonso-Sanz
ETSI Agrónomos (Estadística), Universidad Politécnica de Madrid, Madrid, Spain

Article Outline

Glossary
Definition
Introduction
Average Memory
Other Memories
CA with Three Statesk
Reversible CA
Heterogeneous CA
Structurally Dynamic CA
Memory in Other Discrete Time Contexts
Future Directions
Bibliography

Glossary

Cellular automata Cellular Automata (CA) are discrete, spatially explicit extended dynamic systems composed of adjacent cells characterized by an internal state whose value belongs to a finite set. The updating of these states is made simultaneously according to a common local transition rule involving only a neighborhood of each cell.

Memory Standard CA are ahistoric (memoryless): i. e., the new state of a cell depends on the neighborhood configuration only at the preceding time step. The standard framework of CA can be extended by the consideration of all past states (history) in the application of the CA rules by implementing memory capabilities in cells and links when topology is dynamic.

Definition

Cellular Automata (CA) are discrete, spatially explicit extended dynamic systems. A CA system is composed of adjacent cells characterized by an internal state whose value belongs to a finite set. The updating of these states is made simultaneously according to a common local transition rule involving only a neighborhood of each cell. Thus, if $\sigma_i^{(T)}$ is taken to denote the value of cell i at time step T, the site values evolve by iteration of the mapping: $\sigma_i^{(T+1)} = \phi\big(\{\sigma_j^{(T)}\} \in \mathcal{N}_i\big)$, where ϕ is an arbitrary function which specifies the cellular automaton *rule* operating on $\mathcal{N}_i$, i. e. the set of cells in the neighborhood of the cell i. Last but not least, standard CA are ahistoric (memoryless): i. e., the new state of a cell depends on the neighborhood configuration only at the preceding time step.

The standard framework of CA can be extended by the consideration of all past states (history) in the application of the CA rules by implementing memory capabilities in cells: $\sigma_i^{(T+1)} = \phi\big(\{s_j^{(T)}\} \in \mathcal{N}_i\big)$, with $s_j^{(T)} = s\big(\sigma_j^{(1)}, \ldots, \sigma_j^{(T-1)}, \sigma_j^{(T)}\big)$ being a state function of the series of states of the cell j up to time-step T. Thus, in CA with memory here: while the transition functions φ of the CA remain unaltered, historic memory of all past iterations is retained by featuring each cell by a summary of its past states.

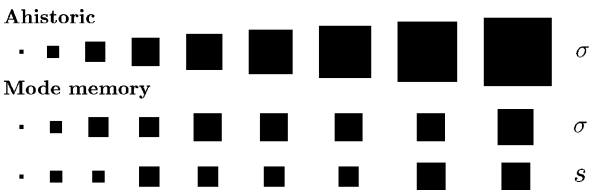

Cellular Automata with Memory, Figure 1
The *speed of light* starting from a single active cell

The memory mechanism considered here is different from that of other CA with memory reported in the literature (e. g., p. 7 in [1], p. 43 in [35], p. 118 in [56]). Typically, these are higher-order-in-time rules that incorporate memory into the transition rule, determining the new configuration in terms of the configurations at previous time-steps. Thus, in *second order in time* rules: $\sigma_i^{(T+1)} = \Phi\big(\{\sigma_j^{(T)}\} \in \mathcal{N}_i, \{\sigma_j^{(T-1)}\} \in \mathcal{N}_i\big)$. CA with memory in cells are cited in [58], but just to state that "CA with memory in cells would result in a qualitatively different behavior". Some authors [55], define rules with *memory* as those with dependence in ϕ on the state of the cell to be updated. So, one-dimensional rules with no *memory*, take the form: $\sigma_i^{(T+1)} = \phi\big(\sigma_{i-1}^{(T)}, \sigma_{i+1}^{(T)}\big)$. Our use of the term memory is not any of these.

Introduction

As a simple example, in the two-dimensional, two-state automaton labeled ahistoric in Fig. 1, a cell becomes (or remains) alive if any cell in its nearest neighborhood is alive, but becomes (or remains) dead on the contrary case. The initial single perturbation in Fig. 1 spreads as fast as possible, i. e. at the *speed of light*.

The lower series of patterns in Fig. 1 shows the effect of featuring cells by their most frequent state, i. e. *mode* memory: $s_i^{(T)} = \text{mode}(\sigma_i^{(1)}, \ldots, \sigma_i^{(T)})$ (with $s_i^{(T)} = \sigma_i^{(T)}$ in case of a tie) on the *speed of light*. Memory has a characteristic inertial effect.

Average Memory

Cells can be featured by a weighted mean value of all their previous states through powers of some parameter $\alpha \in [0, 1]$ acting as a memory factor. Thus, at every time-step T, the weighted mean of the states up to T is computed for every cell i:

$$m_i^{(T)}(\sigma_i^{(1)}, \ldots, \sigma_i^{(T)}) = \frac{\sigma_i^{(T)} + \sum_{t=1}^{T-1} \alpha^{T-t} \sigma_i^{(t)}}{1 + \sum_{t=1}^{T-1} \alpha^{T-t}} \equiv \frac{\omega_i^{(T)}}{\Omega(T)}$$

and then, the featuring states s are obtained by rounding the m ones to 1 for $m > 0.5$ and to 0 for $m < 0.5$. If m is exactly 0.5, then the last state is assigned ($s_i^{(T)} = \sigma_i^{(T)}$). This memory mechanism is *accumulative* in their demand of knowledge of past history: to calculate the memory charge $\omega_i^{(T)}$ it is not necessary to know the whole $\{\sigma_i^{(t)}\}$ series, while it can be sequentially calculated as: $\omega_i^{(T)} = \alpha\omega_i^{(T-1)} + \sigma_i^{(T)}$. It is, $\Omega(T) = (\alpha^T - 1)/(\alpha - 1)$.

The choice of the memory factor α simulates the long-term or remnant memory effect: the limit case $\alpha = 1$ corresponds to memory with equally weighted records (*full* memory, equivalent to *mode* if $k = 2$), whereas $\alpha \ll 1$ intensifies the contribution of the most recent states and diminishes the contribution of the past ones (short term working memory). The choice $\alpha = 0$ leads to the ahistoric model.

In the most unbalanced scenario up to T, i. e.: $\sigma_i^{(1)} = \cdots = \sigma_i^{(T-1)} \neq \sigma_i^{(T)}$, it is:

$$m(0, 0, \ldots, 0, 1) = \frac{1}{2} \Rightarrow \frac{\alpha - 1}{\alpha^T - 1} = \frac{1}{2}$$
$$m(1, 1, \ldots, 1, 0) = \frac{1}{2} \Rightarrow \frac{\alpha^T - \alpha}{\alpha^T - 1} = \frac{1}{2}\,.$$

Thus, memory is only operative if α is greater than a critical α_T that verifies:

$$\alpha_T^T - 2\alpha_T + 1 = 0\,, \tag{1}$$

in which case cells will be featured at T with state values different to the last one. Initial operative values are: $\alpha_3 = 0.61805$, $\alpha_4 = 0.5437$. When $T \to \infty$, Eq. 1 becomes: $-2\alpha_\infty + 1 = 0$, thus, in the $k = 2$ scenario, α-memory is not effective if $\alpha \leq 0.5$.

A Worked Example: The Parity Rule

The so-called *parity* rule states: cell alive if the number of neighbors is odd, dead on the contrary case. Figure 2 shows the effect of memory on the parity rule starting from a single live cell in the Moore neighborhood. In accordance with the above given values of α_3 and α_4: (*i*) The pattern at $T = 4$ is the ahistoric one if $\alpha \leq 0.6$, altered when $\alpha \geq 0.7$, and (*ii*) the patterns at $T = 5$ for $\alpha = 0.54$ and $\alpha = 0.55$ differ.

Not low levels of memory tend to freeze the dynamics since the early time-steps, e. g. over 0.54 in Fig. 2. In the particular case of full memory small oscillators of short

► **Cellular Automata with Memory, Figure 2**
The 2D parity rule with memory up to $T = 15$

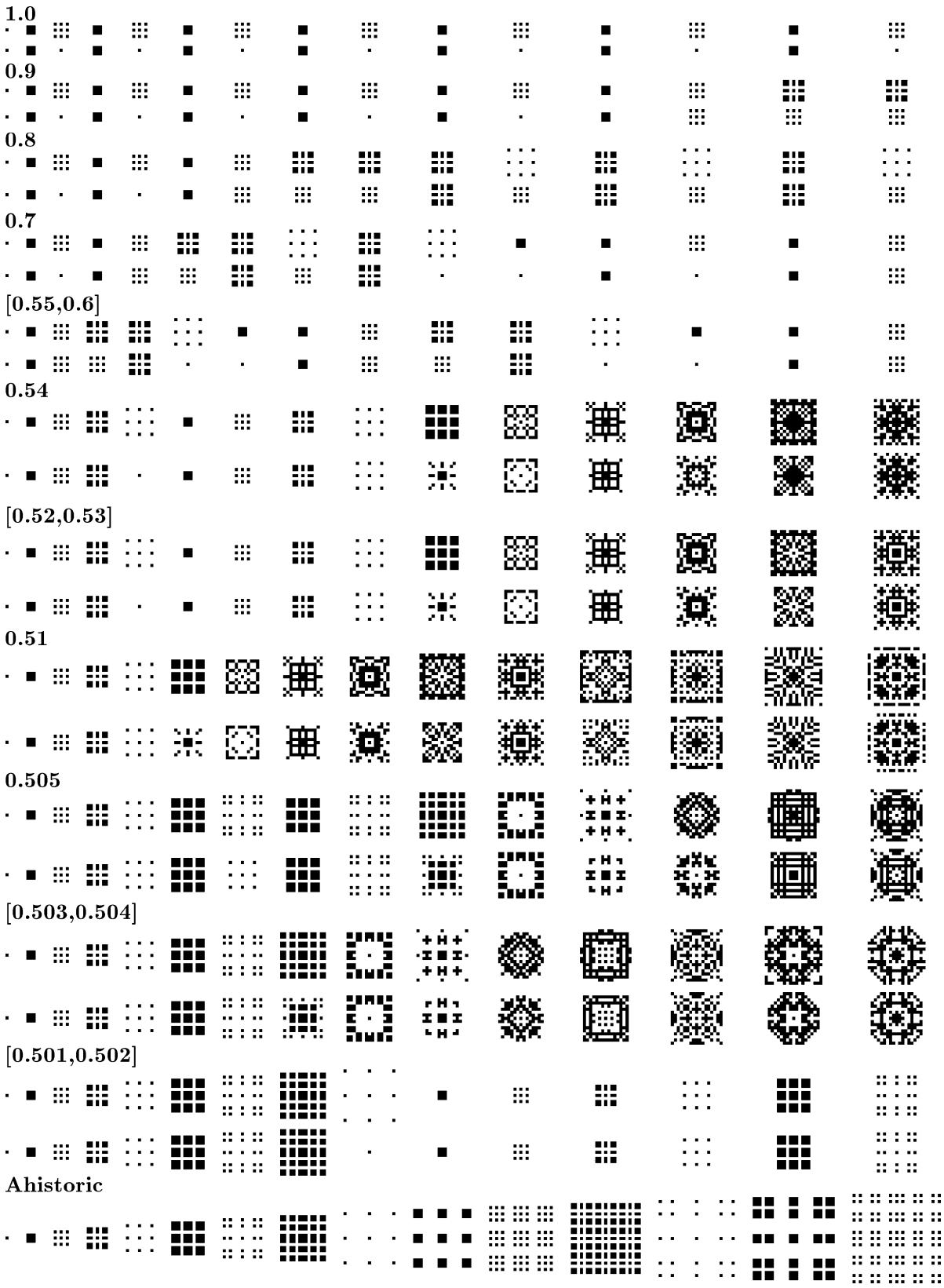
1.0
0.9
0.8
0.7
[0.55,0.6]
0.54
[0.52,0.53]
0.51
0.505
[0.503,0.504]
[0.501,0.502]
Ahistoric

range in time are frequently generated, such as the period-two oscillator that appears as soon as at $T = 2$ in Fig. 2. The group of evolution patterns shown in the [0.503,0.54] interval of α variation of Fig. 2, is rather unexpected to be generated by the parity rule, because they are too *sophisticated* for this simple rule. On the contrary, the evolution patterns with very small memory, $\alpha = 0.501$, resemble those of the ahistoric model in Fig. 2. But this similitude breaks later on, as Fig. 3 reveals: from $T = 19$, the parity rule with minimal memory evolves producing patterns notably different to the ahistoric ones. These patterns tend to be framed in squares of size not over $T \times T$, whereas in the ahistoric case, the patterns tend to be framed in $2T \times 2T$ square regions, so even minimal memory induces a very notable reduction in the affected cell area in the scenario of Fig. 2. The patterns of the featured cells tend not to be far to the actual ones, albeit examples of notable divergence can be traced in Fig. 2. In the particular case of the minimal memory scenario of Fig. 2, that of $\alpha = 0.501$, memory has no effect up to $T = 9$, when the pattern of featured live cells reduces to the initial one; afterward both evolutions are fairly similar up to $T = 18$, but at this time step both kinds of patterns notably differs, and since then the evolution patterns in Fig. 3 notably diverge from the ahistoric ones.

To give consideration to previous states (historic memory) in two-dimensional CA tends to confine the disruption generated by a single live cell. As a rule, full memory tends to generate oscillators, and less historic information retained, i. e. smaller α value, implies an approach to the ahistoric model in a rather smooth form. But the transition which decreases the memory factor from $\alpha = 1.0$ (full memory) to $\alpha = 0.5$ (ahistoric model), is not always regular, and some kind of *erratic* effect of memory can be traced.

The inertial (or conservating) effect of memory dramatically changes the dynamics of the semitotalistic LIFE rule. Thus, (*i*) the *vividness* that some small clusters exhibit in LIFE, has not been detected in LIFE with memory. In particular, the *glider* in LIFE does not *glide* with memory, but stabilizes very close to its initial position as the *tub* , (*ii*) as the size of a configuration increases, often live clusters tend to persist with a higher number of live cells in LIFE with memory than in the ahistoric formulation, (*iii*) a single *mutant* appearing in a stable *agar* can lead to its destruction in the ahistoric model, whereas its effect tends to be restricted to its proximity with memory [26].

One-Dimensional CA *Elementary* rules are one-dimensional, two-state rules operating on nearest neighbors. Following Wolfram's notation, these rules are characterized by a sequence of binary values (β) associated with each of the eight possible triplets

$$\left(\sigma_{i-1}^{(T)}, \sigma_i^{(T)}, \sigma_{i+1}^{(T)}\right) :$$

111	110	101	100	011	010	001	000
β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8

The rules are conveniently specified by their *rule number* $\mathcal{R} = \sum_{i=1}^{8} \beta_i 2^{8-i}$. *Legal* rules are *reflection symmetric* ($\beta_5 = \beta_2, \beta_7 = \beta_4$), and *quiescent* ($\beta_8 = 0$), restrictions that leave 32 possible *legal* rules.

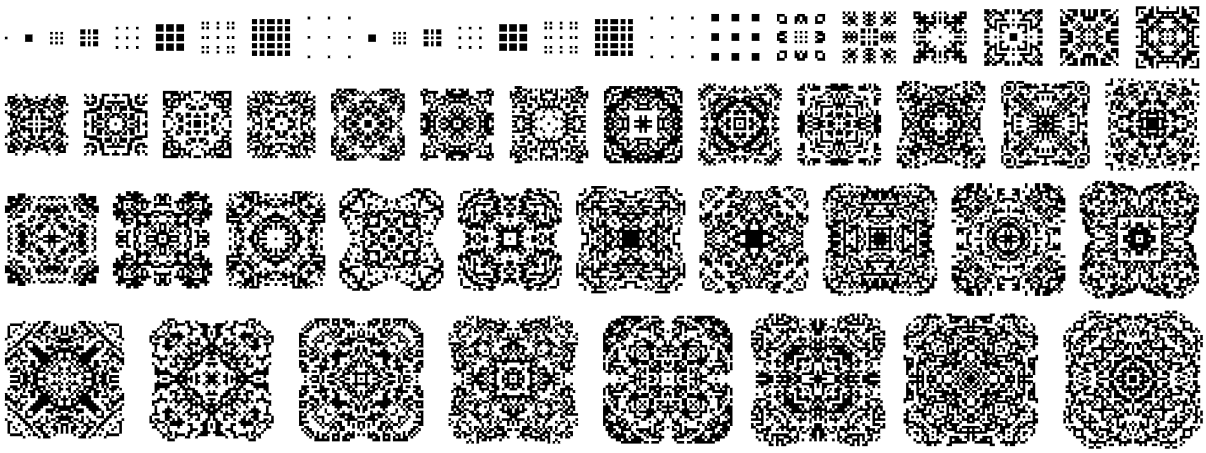

Cellular Automata with Memory, Figure 3
The 2D parity rule with $\alpha = 0.501$ memory starting from a single site live cell up to $T = 55$

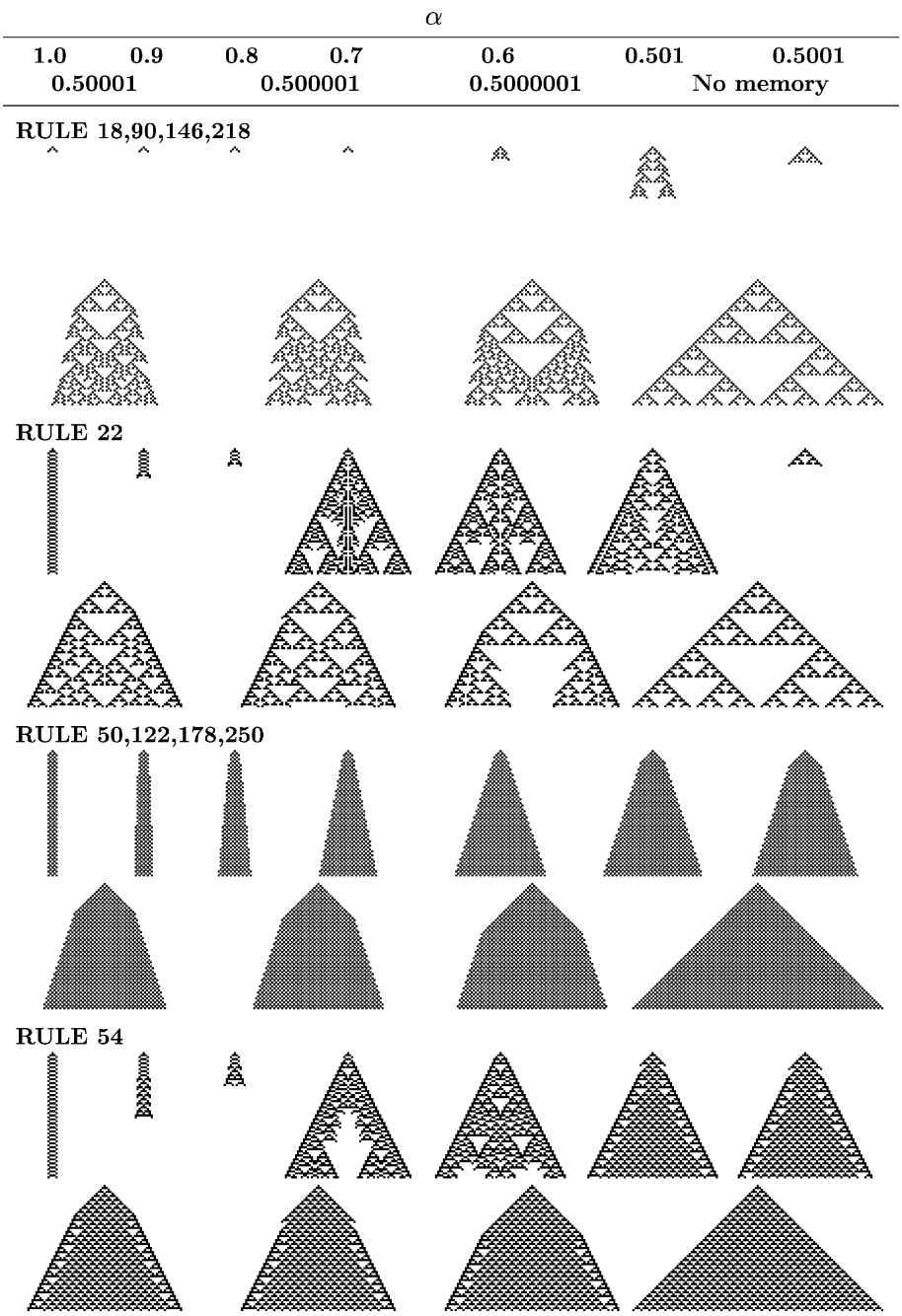

Cellular Automata with Memory, Figure 4
Elementary, legal rules with memory from a single site live cell

Figure 4 shows the spatio-temporal patterns of legal rules affected by memory when starting from a single live cell [17]. Patterns are shown up to $T = 63$, with the memory factor varying from 0.6 to 1.0 by 0.1 intervals, and adopting also values close to the limit of its effectivity: 0.5. As a rule, the transition from the $\alpha = 1.0$ (fully historic) to the ahistoric scenario is fairly gradual, so that the patterns become more expanded as less historic memory is retained (smaller α). Rules 50, 122, 178,250, 94, and 222,254 are paradigmatic of this smooth evolution. Rules 222 and

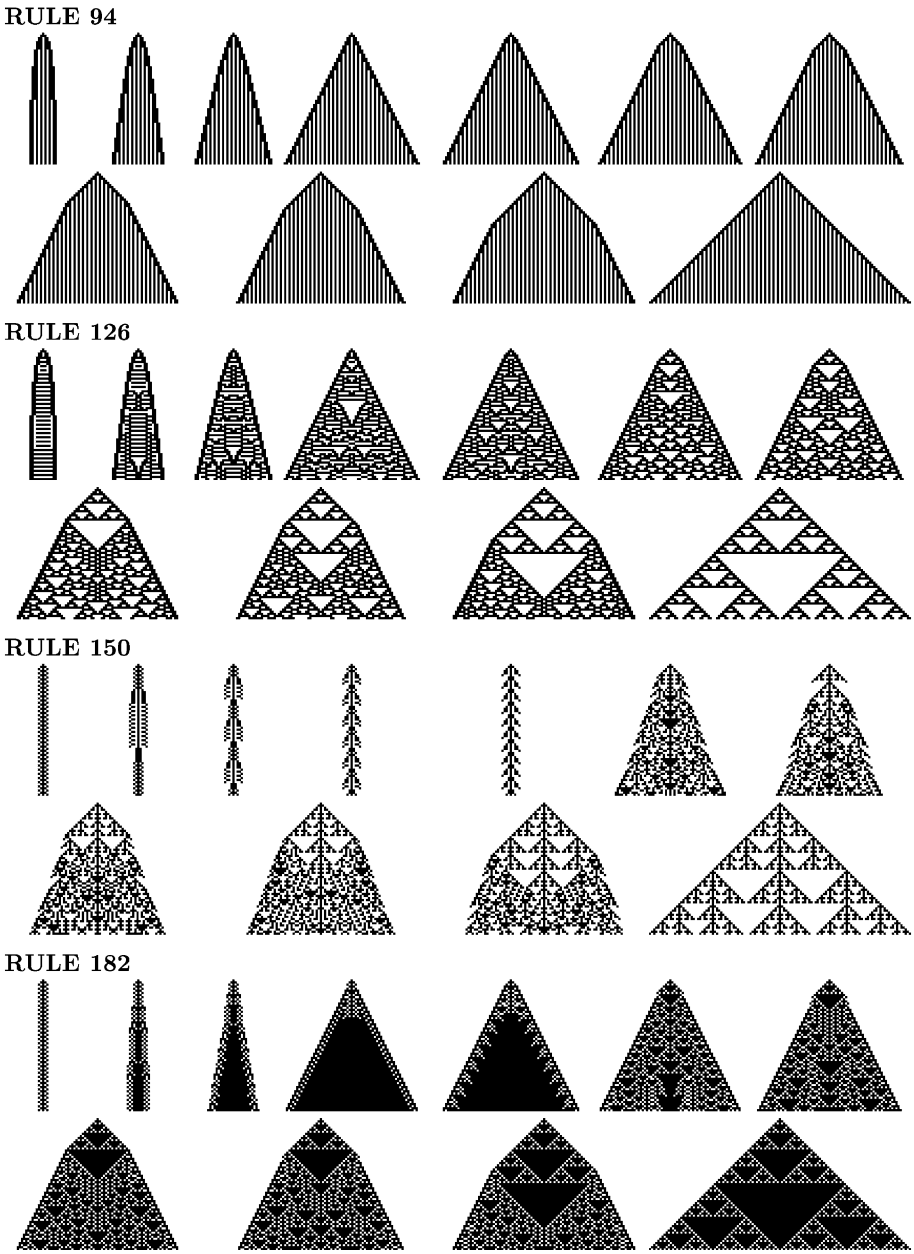

Cellular Automata with Memory, Figure 4
(***continued***)

254 are not included in Fig. 4 as they evolve as rule 94 but with the *inside* of patterns full of active cells. Rules 126 and 182 also present a gradual evolution, although their patterns with high levels of memory models hardly resemble the historic ones. Examples without a smooth effect of memory are also present in Fig. 4: (*i*) rule 150 is sharply restrained at $\alpha = 0.6$, (*ii*) the important rule 54 extinguish in [0.8,0.9], but not with full memory, (*iii*) the rules in the group {18,90,146,218} become extinct from $\alpha = 0.501$. Memory kills the evolution for these rules already at $T = 4$ for α values over α_3 (thus over 0.6 in Fig. 4): after $T = 3$ all the cells, even the two outer cells alive at $T = 3$, are featured as dead, and (*iv*) rule 22 becomes extinct for $\alpha = 0.501$, not in 0.507, 0.6, and 0.7, again extinguish at 0.8 and 0.9, and finally generate an oscillator with full memory. It has been argued that rules

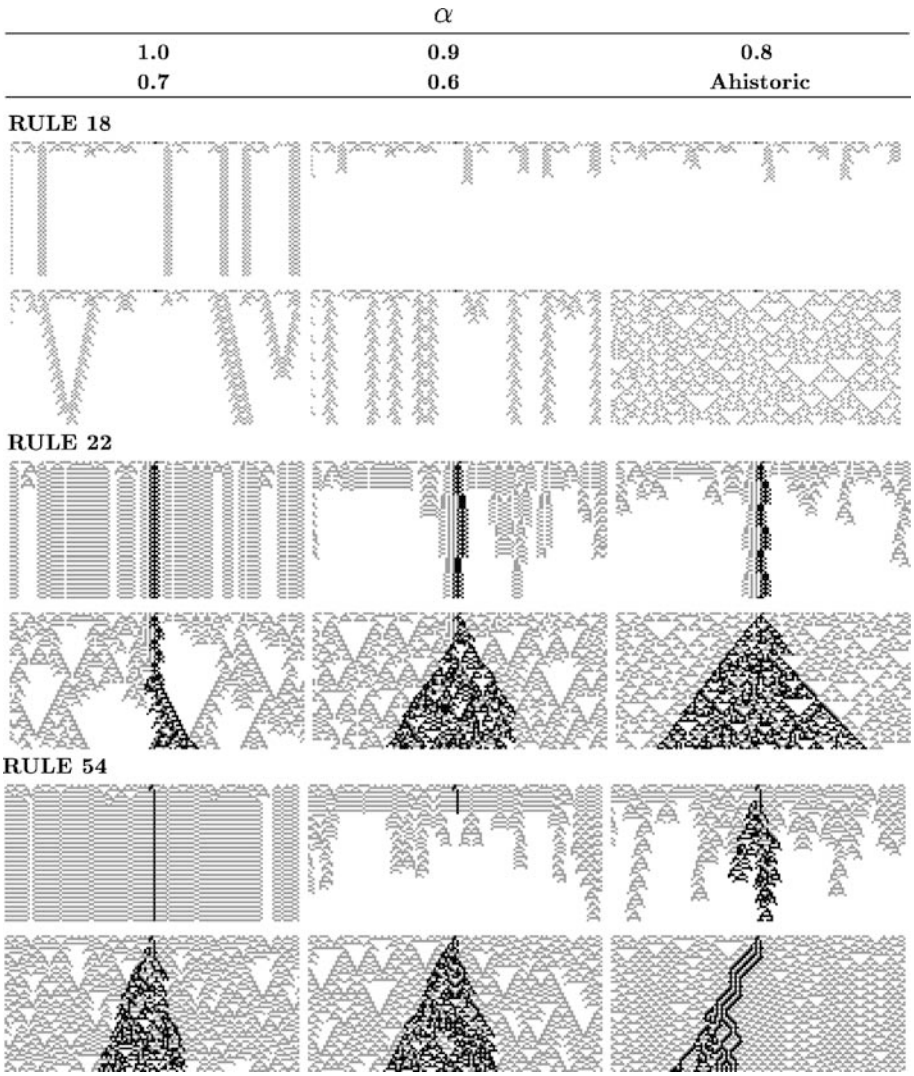

Cellular Automata with Memory, Figure 5
Elementary, legal rules with memory starting at random

18, 22, 122, 146 and 182 *simulate* Rule 90 in that their behavior coincides when restricted to certain spatial subsequences. Starting with a single site live cell, the coincidence fully applies in the historic model for rules 90, 18 and 146. Rule 22 shares with these rules the extinction for high α values, with the notable exception of no extinction in the fully historic model. Rules 122 and 182 diverge in their behavior: there is a gradual decrease in the width of evolving patterns as α is higher, but they do not reach extinction.

Figure 5 shows the effect of memory on legal rules when starting at random: the values of sites are initially uncorrelated and chosen at random to be 0 (*blank)* or 1 (*gray*) with probability 0.5. Differences in patterns resulting from reversing the center site value are shown as *black* pixels. Patterns are shown up to $T = 60$, in a line of size 129 with periodic boundary conditions imposed on the edges. Only the nine legal rules which generate non-periodic patterns in the ahistoric scenario are significantly affected by memory. The patterns with inverted triangles dominate the scene in the ahistoric patterns of Fig. 5, a common appearance that memory tends to eliminate.

History has a dramatic effect on Rule 18. Even at the low value of $\alpha = 0.6$, the appearance of its spatio-temporal pattern fully changes: a number of isolated periodic structures are generated, far from the distinctive inverted triangle world of the ahistoric pattern. For $\alpha = 0.7$, the live structures are fewer, advancing the extinction found in

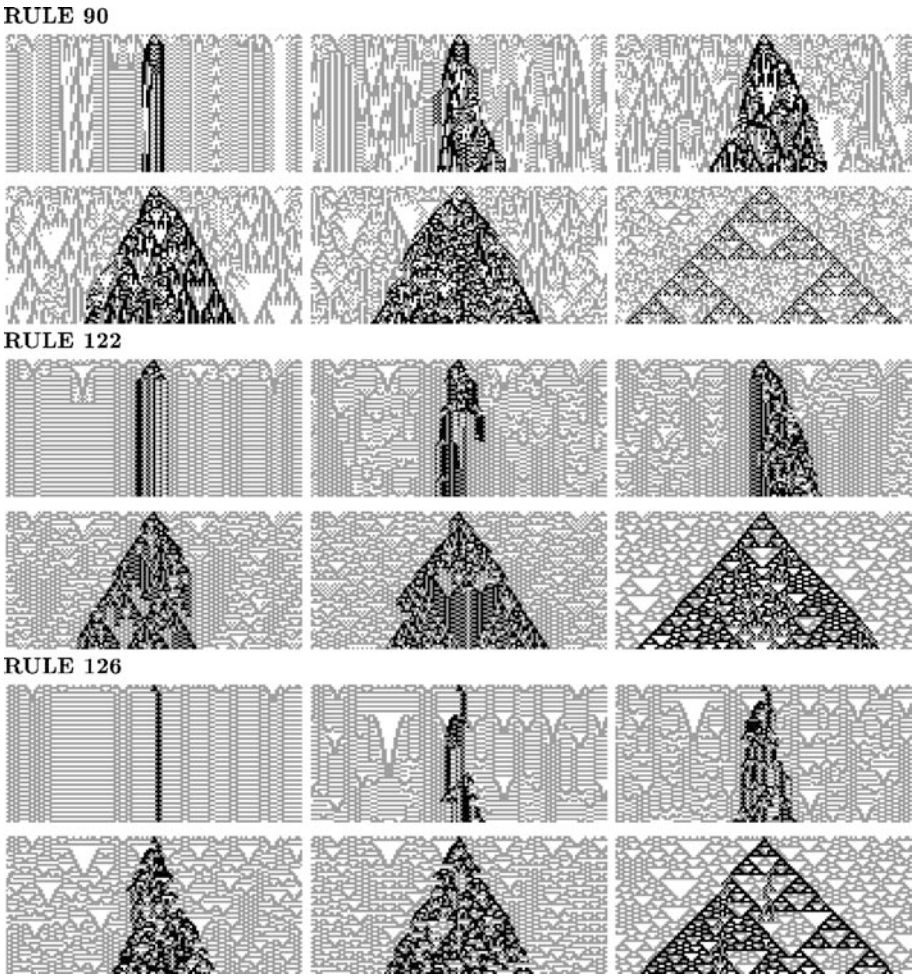

Cellular Automata with Memory, Figure 5
(***continued***)

[0.8,0.9]. In the fully historic model, simple periodic patterns survive.

Rule 146 is affected by memory in much the same way as Rule 18 because their binary codes differ only in their β_1 value. The spatio-temporal of rule 182 and its equivalent Rule 146 are reminiscent, though those of Rule 182 look like a *negatives* photogram of those of Rule 146.

The effect of memory on rule 22 and the *complex* rule 54 is similar. Their spatio-temporal patterns in $\alpha = 0.6$ and $\alpha = 0.7$ keep the essential of the ahistoric, although the inverted triangles become enlarged and tend to be more sophisticated in their *basis*. A notable discontinuity is found for both rules ascending in the value of the memory factor: in $\alpha = 0.8$ and $\alpha = 0.9$ only a few simple structures survive. But unexpectedly, the patterns of the fully historic scenario differ markedly from the others, showing a high degree of synchronization.

The four remaining chaotic legal rules (90, 122, 126 and 150) show a much smoother evolution from the ahistoric to the historic scenario: no pattern evolves either to full extinction or to the preservation of only a few isolated persistent propagating structures (solitons). Rules 122 and 126, evolve in a similar form, showing a high degree of synchronization in the fully historic model.

As a rule, the effect of memory on the differences in patterns (DP) resulting from reversing the value of its initial center site is reminiscent of that on the spatio-temporal patterns, albeit this very much depends on the actual simulation run. In the case of rule 18 for example, damage is not present in the simulation of Fig. 5. The group of rules 90, 122, 126 and 150 shows a, let us say *canonical*, fairly gradual evolution from the ahistoric to the historic scenario, so that the DP appear more constrained as more historic memory is retained, with no extinction for any α

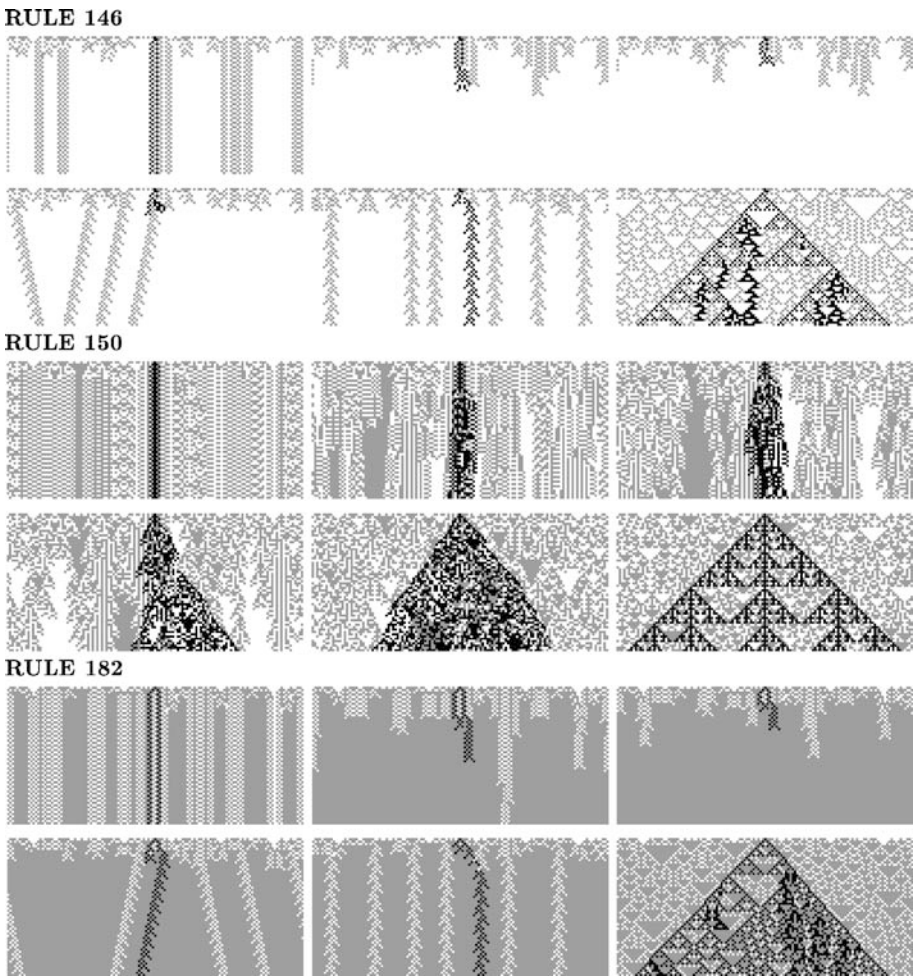

Cellular Automata with Memory, Figure 5
(***continued***)

value. Figure 6 shows the evolution of the fraction ρ_T of sites with value 1, starting at random ($\rho_0 = 0.5$). The simulation is implemented for the same rules as in Fig. 4, but with notably wider lattice: $N = 500$. A visual inspection of the plots in Fig. 6, ratifies the general features observed in the patterns in Fig. 5 regarding density. That also stands for damage spreading: as a rule, memory depletes the damaged region.

In one-dimensional $r = 2$ CA, the value of a given site depends on values of the nearest and next-nearest neighbors. *Totalistic* $r = 2$ rules with memory have the form: $\sigma_i^{(T+1)} = \phi\left(s_{i-2}^{(T)} + s_{i-1}^{(T)} + s_i^{(T)} + s_{i+1}^{(T)} + s_{i+2}^{(T)}\right)$. The effect of memory on these rules follows the way traced in the $r = 1$ context, albeit with a rich casuistic studied in [14].

Probabilistic CA So far the CA considered are deterministic. In order to study perturbations to deterministic CA as well as transitional changes from one deterministic CA to another, it is natural to generalize the deterministic CA framework to the probabilistic scenario. In the elementary scenario, the β are replaced by probabilities

$$p = P\left(\sigma_i^{(T+1)} = 1 \Big/ \sigma_{i-1}^{(T)}, \sigma_i^{(T)}, \sigma_{i+1}^{(T)}\right) :$$

111	110	101	100	011	010	001	000	
β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	$\beta \in \{0, 1\}$
p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	$p \in [0, 1]$

As in the deterministic scenario, memory can be embedded in probabilistic CA (PCA) by featuring cells by a summary of past states s_i instead of by their last state σ_i: $p = P\left(\sigma_i^{(T+1)} = 1/s_{i-1}^{(T)}, s_i^{(T)}, s_{i+1}^{(T)}\right)$. Again, memory is embedded into the characterization of cells but not in the construction of the stochastic transition rules, as done in the canonical approach to PCA. We have

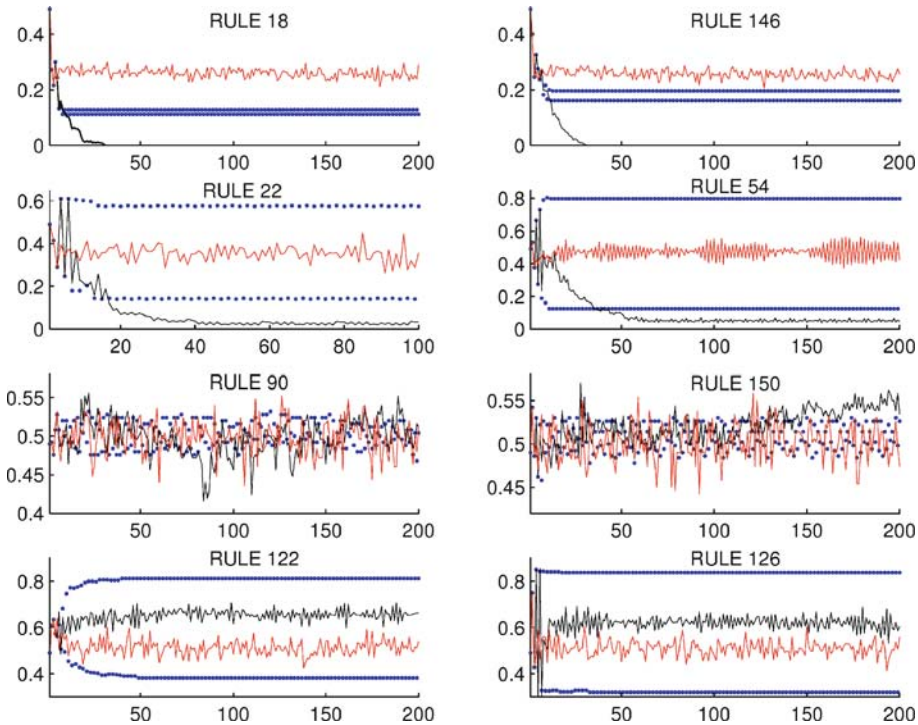

Cellular Automata with Memory, Figure 6
Evolution of the density starting at random in elementary legal rules. Color code: *blue* → full memory, *black* → $\alpha = 0.8$, *red* → ahistoric model

explored the effect of memory on three different subsets $(0, p_2, 0, p_4, p_2, 0, p_4, 0)$, $(0, 0, p_3, 1, 0, p_6, 1, 0)$, and $(p_1, p_2, p_1, p_2, p_2, 0, p_2, 0)$ in [9].

Other Memories

A number of average-like memory mechanisms can readily be proposed by using weights different to that implemented in the α-memory mechanism: $\delta(t) = \alpha^{T-t}$. Among the plausible choices of δ, we mention the weights $\delta(t) = t^c$ and $\delta(t) = c^t$, $c \in \mathbb{N}$, in which the larger the value of c, the more heavily is the recent past taken into account, and consequently the closer the scenario to the ahistoric model [19,21]. Both weights allow for integer-based arithmetics (*à la* CA) comparing $2\omega^{(T)}$ to $2\Omega(T)$ to get the featuring states s (a clear computational advantage over the α-based model), and are accumulative in respect to charge: $\omega_i^{(T)} = \omega_i^{(T-1)} + T^c \sigma_i^{(T)}$, $\omega_i^{(T)} = \omega_i^{(T-1)} + c^T \sigma_i^{(T)}$. Nevertheless, they share the same drawback: powers *explode*, at high values of t, even for $c = 2$.

Limited trailing memory would keep memory of only the last τ states. This is implemented in the context of average memory as: $\omega_i^{(T)} = \sum_{t=\top}^{T} \delta(t)\sigma_i^{(t)}$, with $\top = \max(1, T - \tau + 1)$. Limiting the trailing memory would approach the model to the ahistoric model ($\tau = 1$). In the geometrically discounted method, such an effect is more appreciable when the value of α is high, whereas at low α values (already close to the ahistoric model when memory is not limited) the effect of limiting the trailing memory is not so important. In the $k = 2$ context, if $\tau = 3$, provided that $\alpha > \alpha_3 = 0.61805$, the memory mechanism turns out to be that of selecting the mode of the last three states: $s_i^{(T)} = mode\left(\sigma_i^{(T-2)}, \sigma_i^{(T)}, \sigma_i^{(T-1)}\right)$, i. e. the elementary rule 232.

Figure 7 shows the effect of this kind of memory on legal rules. As is known, history has a dramatic effect on Rules 18, 90, 146 and 218 as their pattern dies out as early as at $T = 4$. The case of Rule 22 is particular: two *branches* are generated at $T = 17$ in the historic model; the patterns of the remaining rules in the historic model are much reminiscent of the ahistoric ones, but, let us say, *compressed.* Figure 7 shows also the effect of memory on some relevant quiescent asymmetric rules. Rule 2 *shifts* a single site live cell one space at every time-step in the ahistoric model; with the pattern dying at $T = 4$. This evolution is common

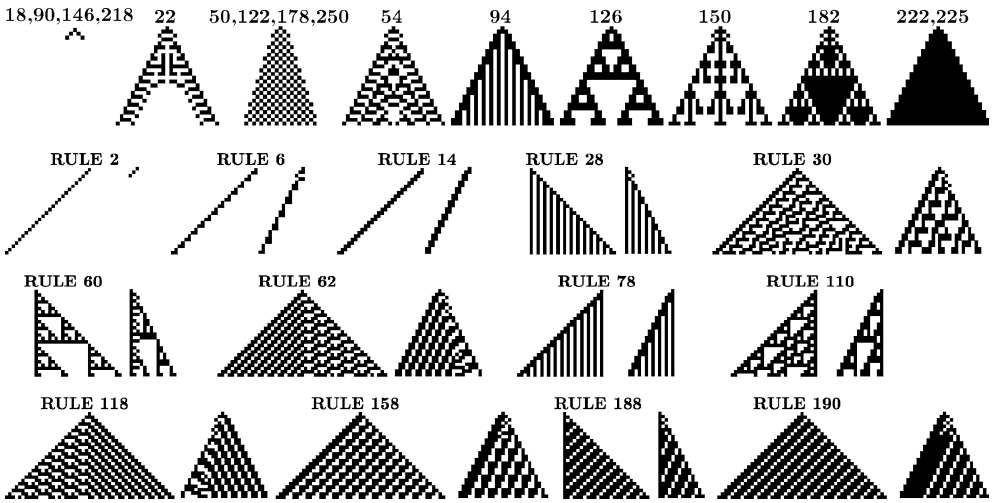

Cellular Automata with Memory, Figure 7
Legal (first row of patterns) and quiescent asymmetric elementary rules significantly affected by the *mode* of the three last states of memory

to all rules that just shift a single site cell without increasing the number of living cells at $T = 2$, this is the case of the important rules 184 and 226. The patterns generated by rules 6 and 14 are *rectified* (in the sense of having the lines in the spatio-temporal pattern slower slope) by memory in such a way that the total number of live cells in the historic and ahistoric spatio-temporal patterns is the same. Again, the historic patterns of the remaining rules in Fig. 7 seem, as a rule, like the ahistoric ones compressed [22].

Elementary rules (ER, noted f) can in turn act as memory rules:

$$s_i^{(T)} = f\big(\sigma_i^{(T-2)}, \sigma_i^{(T)}, \sigma_i^{(T-1)}\big)$$

Figure 8 shows the effect of ER memories up to $R = 125$ on rule 150 starting from a single site live cell up to $T = 13$. The effect of ER memories with $R > 125$ on rule 150 as well as on rule 90 is shown in [23]. In the latter case, complementary memory rules (rules whose rule number adds 255) have the same effect on rule 90 (regardless of the role played by the three last states in ϕ and the initial configuration). In the ahistoric scenario, Rules 90 and 150 are *linear* (or *additive*): i. e., any initial pattern can be decomposed into the superposition of patterns from a single site seed. Each of these configurations can be evolved independently and the results superposed (module two) to obtain the final complete pattern. The additivity of rules 90 and 150 remains in the historic model with linear memory rules.

Figure 9 shows the effect of elementary rules on the 2D parity rule with von Neumann neighborhood from a singe site live cell. This figure shows patterns from $T = 4$, being the three first patterns: . The consideration of CA rules as memory induces a fairly unexplored *explosion* of new patterns.

CA with Three Statesk

This section deals with CA with three possible values at each site ($k = 3$), noted $\{0, 1, 2\}$, so the rounding mechanism is implemented by comparing the unrounded weighted mean m to the hallmarks 0.5 and 1.5, assigning the last state in case on an equality to any of these values. Thus,

$$s^T = 0 \text{ if } m^T < 0.5, s^T = 1 \text{ if } 0.5 < m^T < 1.5, s^T = 2 \text{ if } m^T > 1.5, \text{ and } s^T = \sigma^T \text{ if } m^T = 0.5 \text{ or } m^T = 1.5\,.$$

In the most unbalanced cell dynamics, historic memory takes effect after time step T only if $\alpha > \alpha_T$, with $3\alpha_T^T - 4\alpha_T + 1 = 0$, which in the temporal limit becomes $-4\alpha_* + 1 = 0 \Leftrightarrow \alpha_* = 0.25$. In general, in CA with k states (termed from 0 to $k-1$), the characteristic equation at T is $(2k-3)\alpha_T^T - (2k-1)\alpha_T + 1 = 0$, which becomes $-2(k-1)\alpha_* + 1 = 0$ in the temporal limit. It is then concluded that memory does not affect the scenario if $\alpha \le \alpha_*(k) = 1/(2(k-1))$.

We study first *totalistic* rules: $\sigma_i^{(T+1)} = \phi\big(\sigma_{i-1}^{(T)} + \sigma_i^{(T)} + \sigma_{i+1}^{(T)}\big)$, characterized by a sequence of ternary values (β_s) associated with each of the seven possible values of the sum (s) of the neighbors: ($\beta_6, \beta_5, \beta_4, \beta_3, \beta_2,$

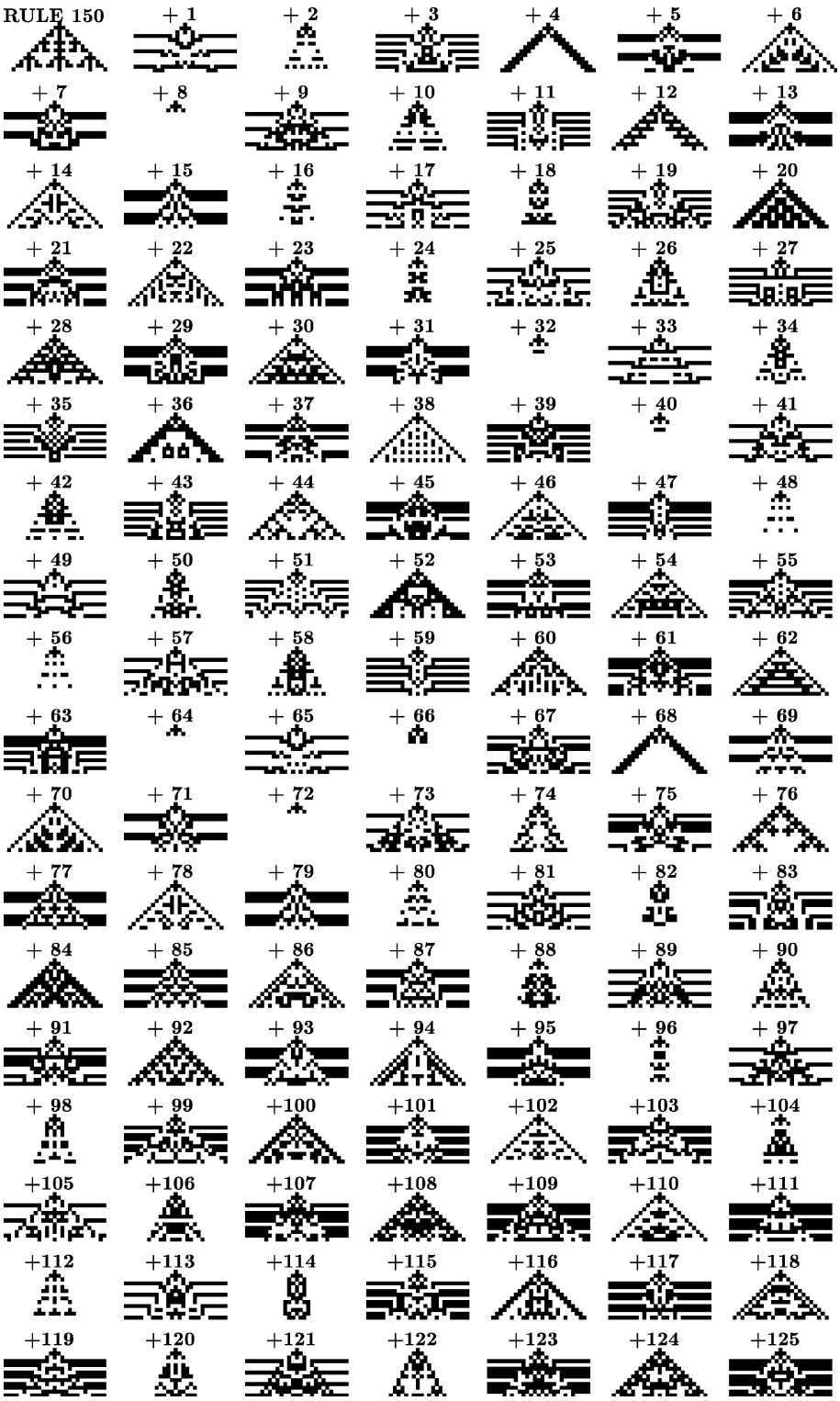

Cellular Automata with Memory, Figure 8
The Rule 150 with elementary rules up to $R = 125$ as memory

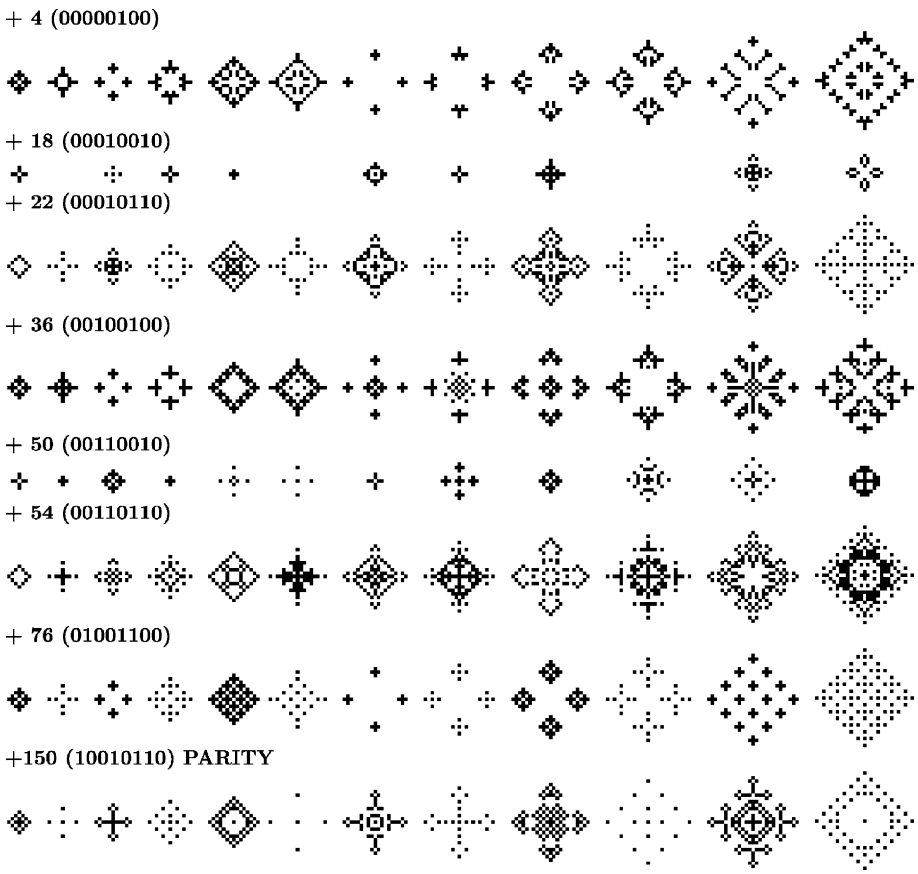

Cellular Automata with Memory, Figure 9
The parity rule with elementary rules as memory. Evolution from $T = 4 - 15$ in the Neumann neighborhood starting from a singe site live cell

β_1, β_0), with associated rule number $\mathcal{R} = \sum_{s=0}^{6} \beta_s 3^s \in [0, 2186]$.

Figure 10 shows the effect of memory on quiescent ($\beta_0 = 0$) *parity* rules, i. e. rules with β_1, β_3 and β_5 non null, and $\beta_2 = \beta_4 = \beta_6 = 0$. Patterns are shown up to $T = 26$. The pattern for $\alpha = 0.3$ is shown to test its proximity to the ahistoric one (recall that if $\alpha \leq 0.25$ memory takes no effect). Starting with a single site seed it can be concluded, regarding proper three-state rules such as those in Fig. 10, that: *(i)* as an overall rule the patterns become more expanded as less historic memory is retained (smaller α). This characteristic *inhibition of growth* effect of memory is traced on rules 300 and 543 in Fig. 10, *(ii)* the transition from the fully historic to the ahistoric scenario tends to be gradual in regard to the amplitude of the spatio-temporal patterns, although their composition can differ notably, even at close α values, *(iii)* in contrast to the two-state scenario, memory fires the pattern of some three-state rules that die out in the ahistoric model, and no rule with memory dies out. Thus, the effect of memory on rules 276, 519, 303 and 546 is somewhat unexpected: they die out at $\alpha \leq 0.3$ but at $\alpha = 0.4$ the pattern expands, the expansion being inhibited (in Fig. 10) only at $\alpha \geq 0.8$. This activation under memory of rules that die at $T = 3$ in the ahistoric model is unfeasible in the $k = 2$ scenario.

The features in the evolving patterns starting from a single seed in Fig. 10 are qualitatively reflected starting at random as shown with rule 276 in Fig. 11, which is also *activated* (even at $\alpha = 0.3$) when starting at random. The effect of average memory (α and integer-based models, unlimited and limited trailing memory, even $\tau = 2$) and that of the mode of the last three states has been studied in [21].

When working with more than three states, it is an inherent consequence of averaging the tendency to *bias* the featuring state to the *mean* value: 1. That explains the *red-shift* in the previous figures. This led us to focus on a much more fair memory mechanism: the *mode*, in what follows.

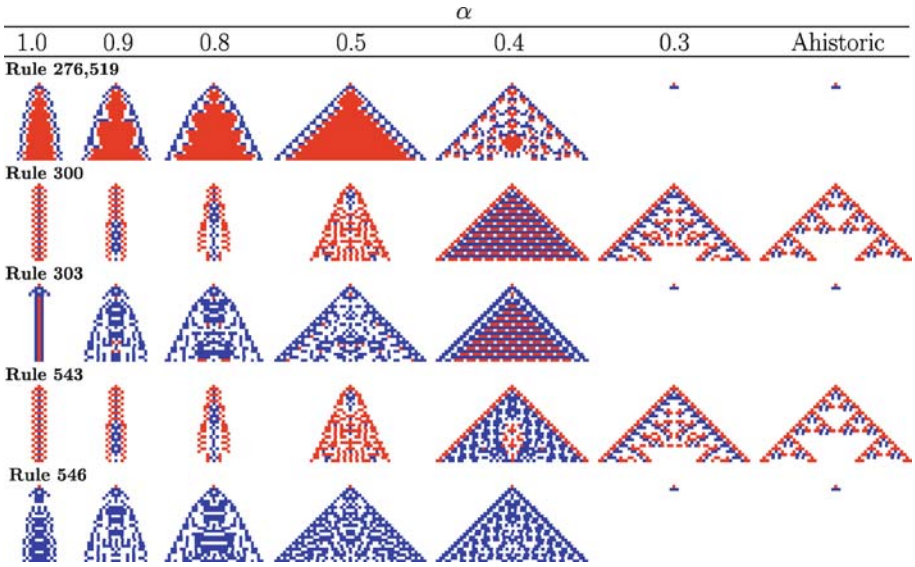

Cellular Automata with Memory, Figure 10
Parity $k = 3$ rules starting from a single $\sigma = 1$ seed. The red cells are at state 1, the blue ones at state

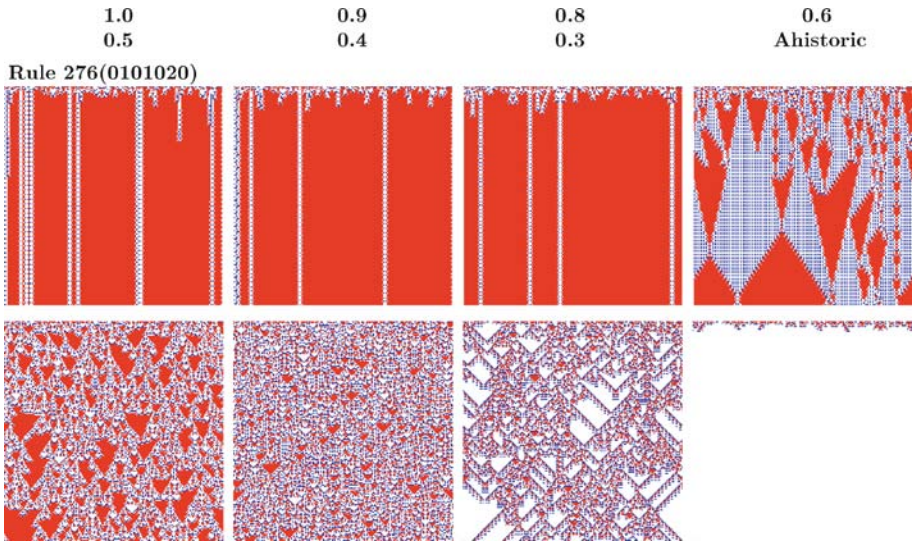

Cellular Automata with Memory, Figure 11
The $k = 3, \mathcal{R} = 276$ rule starting at random

Mode memory allows for manipulation of pure symbols, avoiding any computing/arithmetics.

In *excitable* CA, the three states are featured: resting *0*, excited *1* and refractory *2*. State transitions from excited to refractory and from refractory to resting are unconditional, they take place independently on a cell's neighborhood state: $\sigma_i^{(T)} = 1 \rightarrow \sigma_i^{(T+1)} = 2$, $\sigma_i^{(T)} = 2 \rightarrow \sigma_i^{(T+1)} = 0$. In [15] the excitation rule adopts a Pavlovian phenomenon of defensive inhibition: when strength of stimulus applied exceeds a certain limit the system 'shuts down', this can be naively interpreted as an inbuilt protection of energy loss and exhaustion. To simulate the phenomenon of defensive inhibition we adopt interval excitation rules [2], and a resting cell becomes excited only if one or two of its neighbors are excited: $\sigma_i^{(T)} = 0 \rightarrow \sigma_i^{(T)} = 1$ if $\sum_{j \in \mathcal{N}_i} \left(\sigma_j^{(T)} = 1\right) \in \{1, 2\}$ [3].

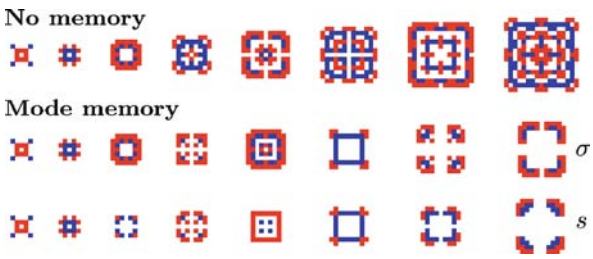

Cellular Automata with Memory, Figure 12
Effect of mode memory on the defensive inhibition CA rule

Figure 12 shows the effect of mode of the last three time steps memory on the defensive-inhibition CA rule with the Moore neighborhood, starting from a simple configuration. At $T = 3$ the outer excited cells in the actual pattern are not featured as excited but as resting cells (twice resting versus one excited), and the series of evolving patterns with memory diverges from the ahistoric evolution at $T = 4$, becoming less expanded. Again, memory tends to restrain the evolution.

The effect of memory on the *beehive* rule, a totalistic two-dimensional CA rule with three states implemented in the hexagonal tessellation [57] has been explored in [13].

Reversible CA

The second-order in time implementation based on the subtraction modulo of the number of states (noted $\ominus$): $\sigma_i^{(T+1)} = \phi(\sigma_j^{(T)} \in \mathcal{N}_i) \ominus \sigma_i^{(T-1)}$, readily reverses as: $\sigma_i^{(T-1)} = \phi(\sigma_j^{(T)} \in \mathcal{N}_i) \ominus \sigma_i^{(T+1)}$. To preserve the reversible feature, memory has to be endowed only in the pivotal component of the rule transition, so: $\sigma_i^{(T-1)} = \phi(s_j^{(T)} \in \mathcal{N}_i) \ominus \sigma_i^{(T+1)}$.

For reversing from T it is necessary to know not only $\sigma_i^{(T)}$ and $\sigma_i^{(T+1)}$ but also $\omega_i^{(T)}$ to be compared to $\Omega(T)$, to obtain:

$$s_i^{(T)} = \begin{cases} 0 & \text{if} \quad 2\omega_i^{(T)} < \Omega(T) \\ \sigma_i^{(T+1)} & \text{if} \quad 2\omega_i^{(T)} = \Omega(T) \\ 1 & \text{if} \quad 2\omega_i^{(T)} > \Omega(T)\,. \end{cases}$$

Then to progress in the reversing, to obtain $s_i^{(T-1)} = \text{round}\big(\omega_i^{(T-1)}/\Omega(T-1)\big)$, it is necessary to calculate $\omega_i^{(T-1)} = \big(\omega_i^{(T)} - \sigma_i^{(T)}\big)/\alpha$. But in order to avoid dividing by the memory factor (recall that operations with real numbers are not exact in computer arithmetic), it is preferable to work with $\gamma_i^{(T-1)} = \omega_i^{(T)} - \sigma_i^{(T)}$, and to compare these values to $\Gamma(T-1) = \sum_{t=1}^{T-1} \alpha^{T-t}$. This leads to:

$$s_i^{(T-1)} = \begin{cases} 0 & \text{if} \quad 2\gamma_i^{(T-1)} < \Gamma(T-1) \\ \sigma_i^{(T)} & \text{if} \quad 2\gamma_i^{(T-1)} = \Gamma(T-1) \\ 1 & \text{if} \quad 2\gamma_i^{(T-1)} > \Gamma(T-1)\,. \end{cases}$$

In general: $\gamma_i^{(T-\tau)} = \gamma_i^{(T-\tau+1)} - \alpha^{\tau-1}\sigma_i^{(T-\tau+1)}$, $\Gamma(T-\tau) = \Gamma(T-\tau+1) - \alpha^{\tau-1}$.

Figure 13 shows the effect of memory on the reversible parity rule starting from a single site live cell, so the scenario of Figs. 2 and 3, with the reversible qualification. As expected, the simulations corresponding to $\alpha = 0.6$ or below shows the ahistoric pattern at $T = 4$, whereas memory leads to a pattern different from $\alpha = 0.7$, and the pattern at $T = 5$ for $\alpha = 0.54$ and $\alpha = 0.55$ differ. Again, in the reversible formulation with memory, *(i)* the configuration of the patterns is notably altered, *(ii)* the speed of diffusion of the area affected are notably reduced, even by minimal memory ($\alpha = 0.501$), *(iii)* high levels of memory tend to freeze the dynamics since the early time-steps.

We have studied the effect of memory in the reversible formulation of CA in many scenarios, e. g., totalistic, $k = r = 2$ rules [7], or rules with three states [21].

Reversible systems are of interest since they preserve information and energy and allow unambiguous backtracking. They are studied in computer science in order to design computers which would consume less energy [51]. Reversibility is also an important issue in fundamental physics [31,41,52,53]. Geraldt 't Hooft, in a speculative paper [34], suggests that a suitably defined deterministic, local reversible CA might provide a viable formalism for constructing field theories on a Planck scale. Svozil [50] also asks for changes in the underlying assumptions of current field theories in order to make their discretization appear more CA-like. Applications of reversible CA with memory in cryptography are being scrutinized [30,42].

Heterogeneous CA

CA on networks have arbitrary connections, but, as proper CA, the transition rule is identical for all cells. This generalization of the CA paradigm addresses the intermediate class between CA and Boolean networks (BN, considered in the following section) in which, rules may be different at each site.

In networks two topological ends exist, random and regular networks, both display totally opposite geometric properties. Random networks have lower clustering coefficients and shorter average path length between nodes commonly known as *small world* property. On the other

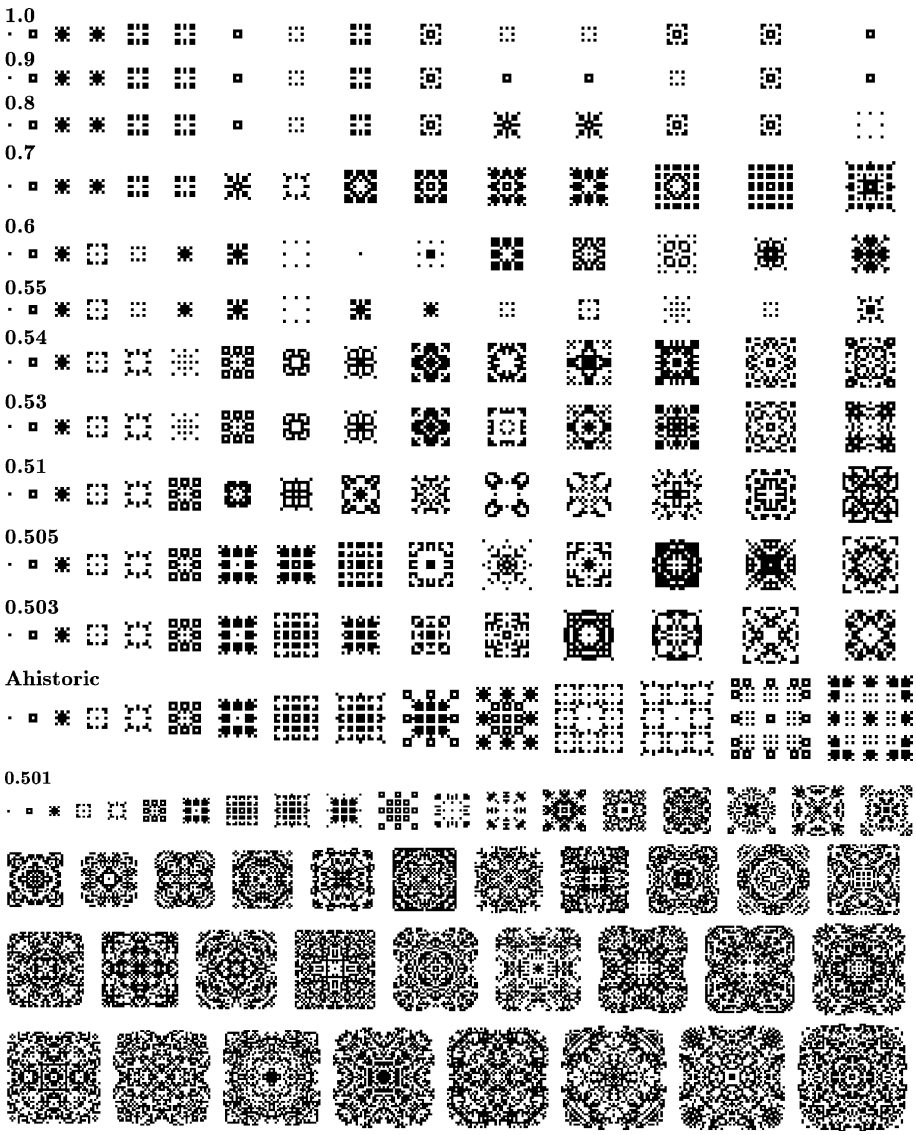

Cellular Automata with Memory, Figure 13
The reversible parity rule with memory

hand, regular graphs, have a large average path length between nodes and high clustering coefficients.

In an attempt to build a network with characteristics observed in real networks, a large clustering coefficient and a *small world* property, Watts and Strogatz (WS, [54]) proposed a model built by randomly rewiring a regular lattice. Thus, the WS model interpolates between regular and random networks, taking a single new parameter, the random rewiring degree, i. e.: the probability that any node redirects a connection, randomly, to any other. The WS model displays the high clustering coefficient common to regular lattices as well as the *small world* property (the *small world* property has been related to faster flow in the information transmission). The long-range links introduced by the randomization procedure dramatically reduce the diameter of the network, even when very few links are rewired.

Figure 14 shows the effect of memory and topology on the parity rule with four inputs in a lattice of size 65×65 with periodic boundary conditions, starting at random. As expected, memory depletes the Hamming distance between two consecutive patterns in relation to the *ahistoric*

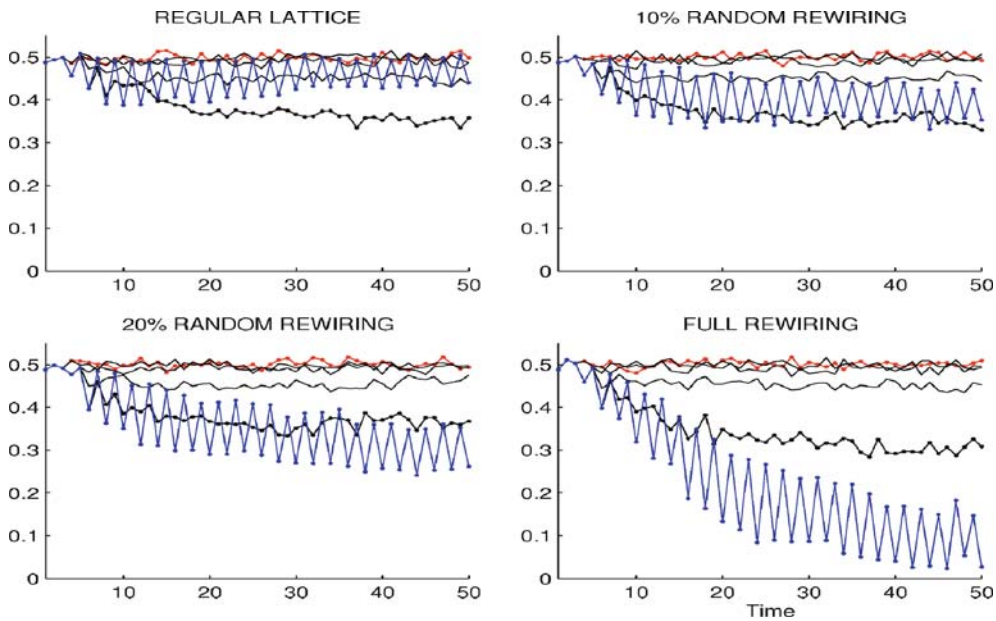

Cellular Automata with Memory, Figure 14
The parity rule with four inputs: effect of memory and random rewiring. Distance between two consecutive patterns in the ahistoric model (*red*) and memory models of α levels: 0.6,0.7.0.8, 0.9 (*dotted*) and 1.0 (*blue*)

model, particularly when the degree of rewiring is high. With full memory, quasi-oscillators tend to appear. As a rule, the higher the curve the lower the memory factor α, but in the particular case of a regular lattice (and lattice with 10% of rewiring), the evolution of the distance in the full memory model turns out rather atypical, as it is maintained over some memory models with lower α parameters.

Figure 15 shows the evolution of the damage spread when reversing the initial state of the 3×3 central cells in the initial scenario of Fig. 14. The fraction of cells with the state reversed is plotted in the regular and 10% of rewiring scenarios. The plots corresponding to higher rates of rewiring are very similar to that of the 10% case in Fig. 15. Damage spreads fast very soon as rewiring is present, even in a short extent.

Boolean Networks

In Boolean Networks (BN,[38]), instead of what happens in canonical CA, cells may have arbitrary connections and rules may be different at each site. Working with totalistic rules: $\sigma_i^{(T+1)} = \phi_i(\sum_{j \in \mathcal{N}_i} s_j^{(T)})$.

The main features on the effect of memory in Fig. 14 are preserved in Fig. 16: (*i*) the ordering of the *historic* networks tends to be stronger with a high memory factor, (*ii*) with full memory, quasi-oscillators appear (it seems that full memory tends to induce *oscillation*), (*iii*) in the particular case of the regular graph (and a lesser extent in the networks with low rewiring), the evolution of the full memory model turns out rather atypical, as it is maintained over some of those memory models with lower α parameters. The relative Hamming distance between the *ahistoric* patterns and those of *historic* rewiring tends to be fairly constant around 0.3, after a very short initial transition period.

Figure 17 shows the evolution of the damage when reversing the initial state of the 3×3 central cells. As a rule in every frame, corresponding to increasing rates of random rewiring, the higher the curve the lower the memory factor α. The *damage vanishing* effect induced by memory does result apparently in the regular scenario of Fig. 17, but only full memory controls the damage spreading when the rewiring degree is not high, the dynamics with the remaining α levels tend to the damage propagation that characterizes the *ahistoric* model. Thus, with up to 10% of connections rewired, full memory notably controls the spreading, but this control capacity tends to disappear with a higher percentage of rewiring connections. In fact, with rewiring of 50% or higher, neither full memory seems to be very effective in altering the final rate of damage, which tends to reach a plateau around 30% regardless of

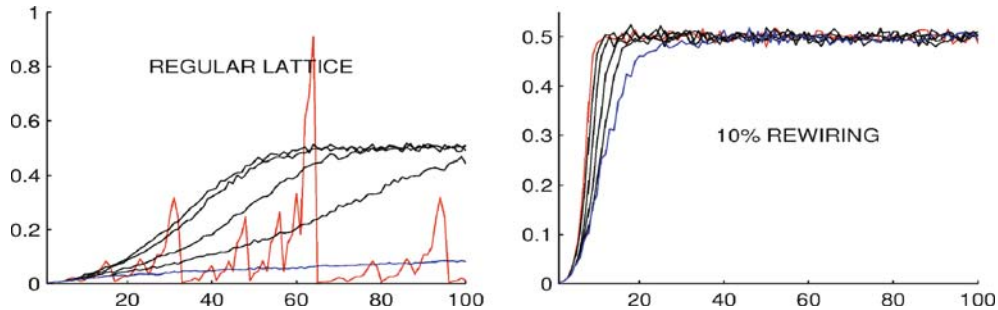

Cellular Automata with Memory, Figure 15
Damage up to $T = 100$ in the parity CA of Fig. 14

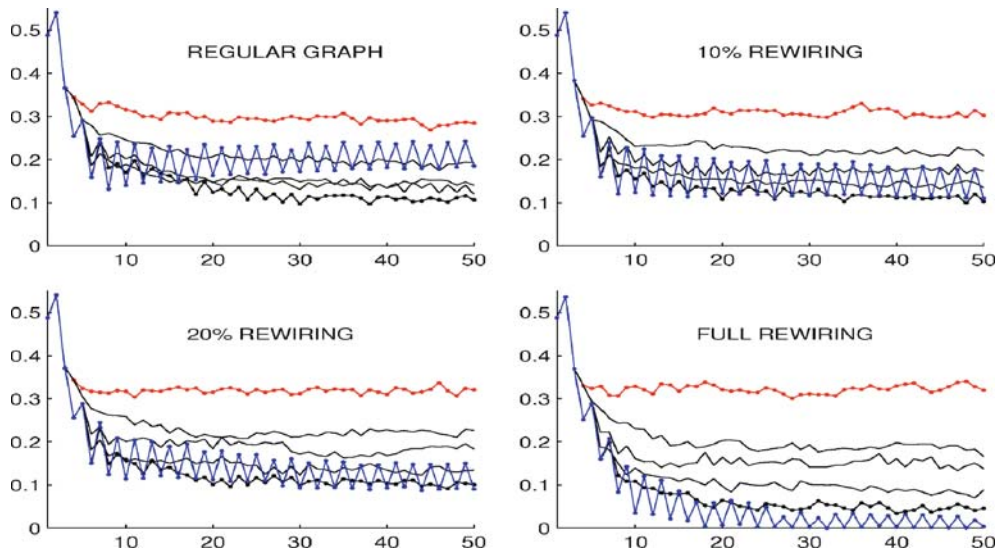

Cellular Automata with Memory, Figure 16
Relative Hamming distance between two consecutive patterns. Boolean network with totalistic, $K = 4$ rules in the scenario of Fig. 14

scenario. A level notably coincident with the percolation threshold in site percolation in the simple cubic lattice, and the critical point for the nearest neighbor Kaufmann model on the square lattice [49]: 0.31.

Structurally Dynamic CA

Structurally dynamic cellular automata (SDCA) were suggested by Ilachinski and Halpern [36]. The essential new feature of this model was that the connections between the cells are allowed to change according to rules similar in nature to the state transition rules associated with the conventional CA. This means that given certain conditions, specified by the *link transition rules*, links between rules may be created and destroyed; the neighborhood of each cell is dynamic, so, state and link configurations of an SDCA are both dynamic and continually interacting.

If cells are numbered 1 to N, their connectivity is specified by an $N \times N$ connectivity matrix in which $\lambda_i j = 1$ if cells i and j are connected; 0 otherwise. So, now: $\mathcal{N}_i^{(T)} = \{j/\lambda_{ij}^{(T)} = 1\}$ and $\sigma_i^{(T+1)} = \phi(\sigma_j^{(T)} \in \mathcal{N}_i^{(T)})$. The geodesic *distance* between two cells i and j, δ_{ij}, is defined as the number of links in the shortest path between i and j. Thus, i and j are *direct* neighbors if $\delta_{ij} = 1$, and are *next-nearest* neighbors if $\delta_i j = 2$, so $\mathcal{N}\mathcal{N}_i^{(T)} = \{j/\delta_{ij}^{(T)} = 2\}$. There are two types of link transition functions in an SDCA: *couplers* and *decouplers*, the former add new links, the latter remove links. The coupler and decoupler set determines the link transition rule: $\lambda_{ij}^{(T+1)} = \psi(l_{ij}^{(T)}, \sigma_i^{(T)}, \sigma_j^{(T)})$.

Instead of introducing the formalism of the SDCA, we deal here with just one example in which the decoupler rule removes all links connected to cells in which both values are zero ($\lambda_{ij}^{(T)} = 1 \rightarrow \lambda_{ij}^{(T+1)} = 0$ iff $\sigma_i^{(T)} + \sigma_j^{(T)} = 0$)

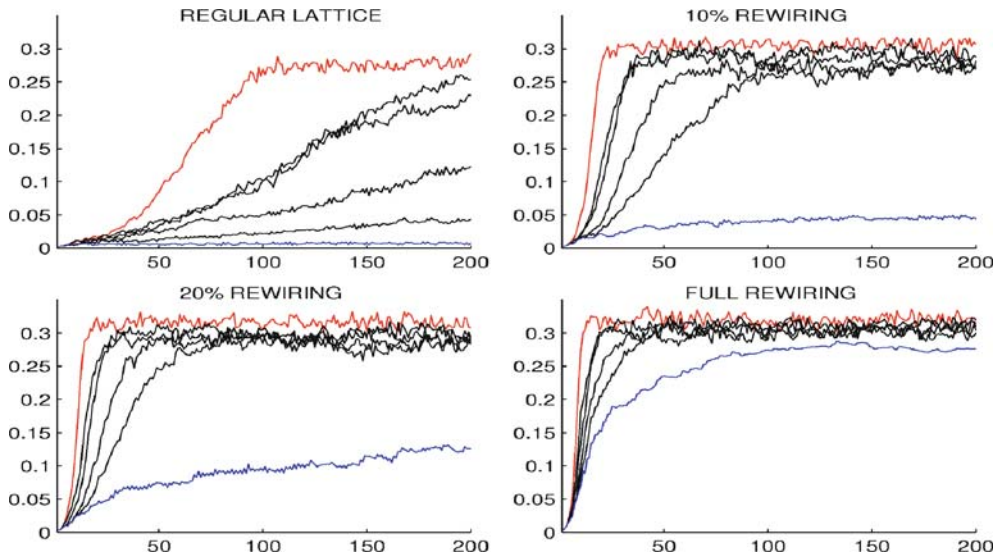

Cellular Automata with Memory, Figure 17
Evolution of the damage when reversing the initial state of the 3 × 3 central cells in the scenario of Fig. 16

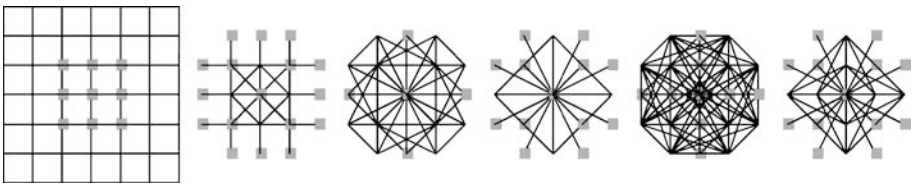

Cellular Automata with Memory, Figure 18
The SDCA described in text up to $T = 6$

and the coupler rule adds links between all next-nearest neighbor sites in which both values are one ($\lambda_{ij}^{(T)} = 0 \rightarrow \lambda_i^{(T+1)} j = 1$ iff $\sigma_i^{(T)} + \sigma_j^{(T)} = 2$ and $j \in \mathcal{N}\mathcal{N}_i^{(T)}$). The SDCA with these transition rules for connections, together with the *parity* rule for mass states, is implemented in Fig. 18, in which the initial Euclidean lattice with four neighbors (so the generic cell □ has eight next-nearest neighbors: ❖) is seeded with a 3 × 3 block of ones. After the first iteration, most of the lattice structure has decayed as an effect of the decoupler rule, so that the active value cells and links are confined to a small region. After $T = 6$, the link and value structures become periodic, with a periodicity of two.

Memory can be embedded in links in a similar manner as in state values, so the link between any two cells is featured by a mapping of its previous link values: $l_{ij}^{(T)} = l(\lambda_{ij}^{(1)}, \ldots, \lambda_{ij}^{(T)})$. The *distance* between two cells in the historic model (d_{ij}), is defined in terms of l instead of λ values, so that i and j are *direct* neighbors if $d_{ij} = 1$, and are *next-nearest* neighbors if $d_{ij} = 2$. Now: $N_i^{(T)} = \{j/d_{ij}^{(T)} = 1\}$, and $NN_i^{(T)} = \{j/d_{ij}^{(T)} = 2\}$. Generalizing the approach to embedded memory applied to states, the unchanged transition rules (ϕ and ψ) operate on the featured link and cell state values: $\sigma_i^{(T+1)} = \phi(s_j^{(T)} \in N_i)$, $\lambda_{ij}^{(T+1)} = \psi(l_i^{(T)} j, s_i^{(T)}, s_j^{(T)})$.

Figure 19 shows the effect of α-memory on the cellular automaton above introduced starting as in Fig. 18. The effect of memory on SDCA in the hexagonal and triangular tessellations is scrutinized in [11].

A plausible wiring dynamics when dealing with excitable CA is that in which the decoupler rule removes all links connected to cells in which both values are at refractory state ($\lambda_{ij}^{(T)} = 1 \rightarrow \lambda_{ij}^{(T+1)} = 0$ iff $\sigma_i^{(T)} = \sigma_j^{(T)}$=2) and the coupler rule adds links between all next-nearest neighbor sites in which both values are excited ($\lambda_{ij}^{(T)} = 0 \rightarrow \lambda_{ij}^{(T+1)} = 1$ iff $\sigma_i^{(T)} = \sigma_j^{(T)} = 1$ and $j \in \mathcal{N}\mathcal{N}_i^{(T)}$).

In the SDCA in Fig. 20, the transition rule for cell states is that of the generalized defensive inhibition rule: resting cell is excited if its ratio of excited and connected to the cell neighbors to total number of connected neighbors lies in

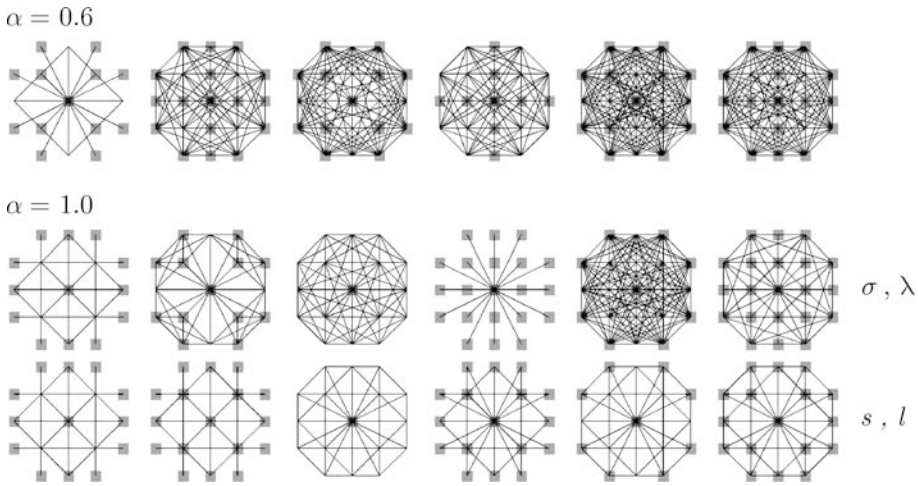

Cellular Automata with Memory, Figure 19
The SD cellular automaton introduced in text with weighted memory of factor α. Evolution from $T = 4$ up to $T = 9$ starting as in Fig. 18

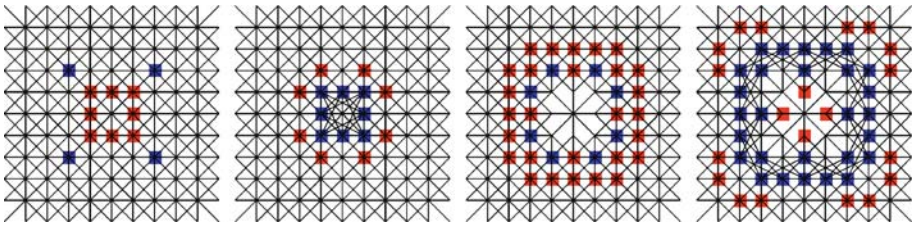

Cellular Automata with Memory, Figure 20
The $k = 3$ SD cellular automaton described in text, up to $T = 4$

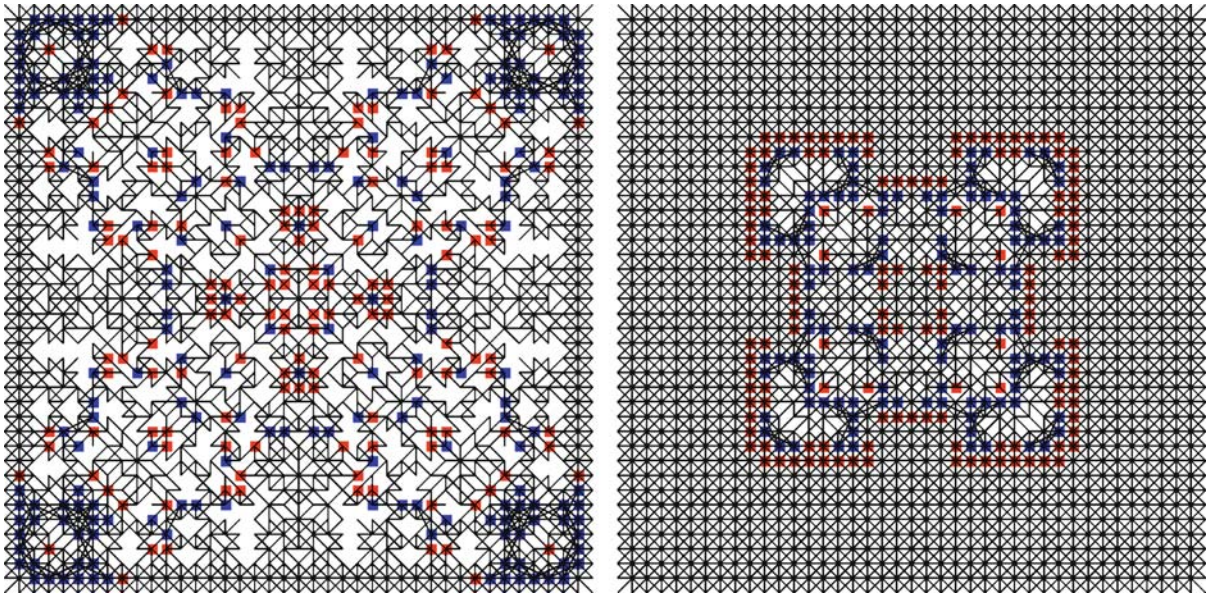

Cellular Automata with Memory, Figure 21
The SD cellular automaton starting as in Fig. 20 at $T = 20$, with no memory (*left*) and mode memory in both cell states and links

the interval [1/8,2/8]. The initial scenario of Fig. 20 is that of Fig. 12 with the wiring network revealed, that of an Euclidean lattice with eight neighbors, in which, the generic cell □ has 16 next-nearest neighbors: ▣. No decoupling is verified at the first iteration in Fig. 20, but the excited cells generate new connections, most of them lost, together with some of the initial ones, at $T = 3$. The excited cells at $T = 3$ generate a *crown* of new connections at $T = 4$. Figure 21 shows the ahistoric and mode memory patterns at $T = 20$. The figure makes apparent the preserving effect of memory.

The Fredkin's reversible construction is feasible in the SDCA scenario extending the $\ominus$ operation also to links: $\lambda_{ij}^{(T+1)} = \psi(\lambda_{ij}^{(T)}, \sigma_i^{(T)}, \sigma_j^{(T)}) \ominus \lambda_{ij}^{(T-1)}$. These automata may be endowed with memory as: $\sigma_i^{(T+1)} = \phi\big(s_j^{(T)} \in N_i^{(T)}\big) \ominus \sigma_i^{(T-1)}, \lambda_{ij}^{(T+1)} = \psi\big(l_{ij}^{(T)}, s_i^{(T)}, s_j^{(T)}\big) \ominus \lambda_{ij}^{(T-1)}$ [12].

The SDCA seems to be particularly appropriate for modeling the human brain function—updating links between cells imitates variation of synaptic connections between neurons represented by the cells—in which the relevant role of memory is apparent. Models similar to SDCA have been adopted to build a dynamical network approach to quantum space-time physics [45,46]. Reversibility is an important issue at such a fundamental physics level. Technical applications of SDCA may also be traced [47]. Anyway, besides their potential applications, SDCA with memory have an aesthetic and mathematical interest on their own [1,35]. Nevertheless, it seems plausible that further study on SDCA (and Lattice Gas Automata with dynamical geometry [40]) with memory should turn out to be profitable.

Memory in Other Discrete Time Contexts

Continuous-Valued CA

The mechanism of implementation of memory adopted here, keeping the transition rule unaltered but applying it to a function of previous states, can be adopted in any spatialized dynamical system. Thus, historic memory can be embedded in:

- *Continuous−valued* CA (or *Coupled Map Lattices* in which the state variable ranges in $\mathcal{R}$, and the transition rule φ is a continuous function [37]), just by considering m instead of σ in the application of the updating rule: $\sigma_i^{(T+1)} = \varphi\big(m_j^{(T)} \in \mathcal{N}_i^{(T)}\big)$. An elementary CA of this kind with memory would be [20]: $\sigma_i^{(T+1)} = \frac{1}{3}\big(m_{i-1}^{(T)} + m_i^{(T)} + m_{i+1}^{(T)}\big)$.
- *Fuzzy* CA, a sort of continuous CA with states ranging in the real [0,1] interval. An illustration of the effect of memory in fuzzy CA is given in [17]. The illustration operates on the elementary rule 90 : $\sigma_i^{(T+1)} = \big(\sigma_{i-1}^{(T)} \wedge (\neg\sigma_{i+1}^{(T)})\big) \vee ((\neg\sigma_{i-1}^{(T)}) \wedge \sigma_{i+1}^{(T)})$, which after fuzzification $(a \vee b \to \min(1, a+b), a \wedge b \to ab, \neg a \to 1-a)$ yields: $\sigma_i^{(T+1)} = \sigma_{i-1}^{(T)} + \sigma_{i+1}^{(T)} - 2\sigma_{i-1}^{(T)}\sigma_{i+1}^{(T)}$; thus incorporating memory: $\sigma_i^{(T+1)} = m_{i-1}^{(T)} + m_{i+1}^{(T)} - 2m_{i-1}^{(T)}m_{i+1}^{(T)}$.
- *Quantum* CA, such, for example, as the simple 1D quantum CA models introduced in [32]:

$$\sigma_j^{(T+1)} = \frac{1}{N^{1/2}}\big(i\delta\sigma_{j-1}^{(T)} + \sigma_j^{(T)} + i\delta^*\sigma_{j+1}^{(T)}\big) ,$$

which would become with memory [20]:

$$\sigma_j^{(T+1)} = \frac{1}{N^{1/2}}\big(i\delta m_{j-1}^{(T)} + m_j^{(T)} + i\delta^* m_{j+1}^{(T)}\big) .$$

Spatial Prisoner's Dilemma

The Prisoner's Dilemma (PD) is a game played by two players (A and B), who may choose either to cooperate (C or 1) or to defect (D or 0). Mutual cooperators each score the *reward* R, mutual defectors score the *punishment* P ; D scores the *temptation* $\mathcal{T}$ against C, who scores S (*sucker's* payoff) in such an encounter. Provided that $\mathcal{T} > R > P > S$, mutual defection is the only equilibrium strategy pair. Thus, in a single round both players are to be *penalized* instead of both *rewarded*, but cooperation may be rewarded in an iterated (or spatial) formulation. The game is simplified (while preserving its essentials) if $P = S = 0$. Choosing $R = 1$, the model will have only one parameter: the *temptation* $\mathcal{T}$=b.

In the spatial version of the PD, each player occupies at a site (i,j) in a 2D lattice. In each generation the payoff of a given individual ($p_{i,j}^{(T)}$), is the sum over all interactions with the eight nearest neighbors and with its own site. In the next generation, an individual cell is assigned the decision ($d_{i,j}^{(T)}$) that received the highest payoff among all the cells of its Moore's neighborhood. In case of a tie, the cell retains its choice. The spatialized PD (SPD for short) has proved to be a promising tool to explain how cooperation can hold out against the ever-present threat of exploitation [43]. This is a task that presents problems in the classic *struggle for survival* Darwinian framework.

When dealing with the SPD, memory can be embedded not only in choices but also in rewards. Thus, in the *historic* model we dealt with, at T: (i) the payoffs coming from previous rounds are accumulated ($\pi_{i,j}^{(T)}$), and (ii) players are featured by a summary of past decisions ($\delta_{i,j}^{(T)}$). Again, in each round or generation, a given cell plays with each of the eight neighbors and itself, the decision δ in the cell of the neighborhood with the highest π being adopted. This approach to modeling memory

Cellular Automata with Memory, Table 1

Choices at $T = 1$ and $T = 2$; accumulated payoffs after $T = 1$ and $T = 2$ starting from a single defector in the SPD. $b > 9/8$

$\mathbf{d}^{(1)} = \delta^{(1)}$				
1	1	1	1	1
1	1	1	1	1
1	1	0	1	1
1	1	1	1	1
1	1	1	1	1

$\mathbf{p}^{(1)} = \pi^{(1)}$				
9	9	9	9	9
9	8	8	8	9
9	8	8b	8	9
9	8	8	8	9
9	9	9	9	9

$\mathbf{d}^{(2)} = \delta(2)$						
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	0	0	0	1	1
1	1	0	0	0	1	1
1	1	0	0	0	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1

$\pi^{(2)} = \alpha\mathbf{p}^{(1)} + \mathbf{p}^{(2)}$						
$9\alpha + 9$	$9\alpha + 9$	$9\alpha + 9$	$9\alpha + 9$	$9\alpha + 9$	$9\alpha + 9$	$9\alpha + 9$
$9\alpha + 9$	$9\alpha + 8$	$9\alpha + 7$	$9\alpha + 6$	$9\alpha + 7$	$9\alpha + 8$	$9\alpha + 9$
$9\alpha + 9$	$9\alpha + 9$	$8\alpha + 5b$	$8\alpha + 3b$	$8\alpha + 5b$	$9\alpha + 9$	$9\alpha + 9$
$9\alpha + 9$	$9\alpha + 9$	$8\alpha + 3b$	8α	$8\alpha + 3b$	$9\alpha + 9$	$9\alpha + 9$
$9\alpha + 9$	$9\alpha + 9$	$8\alpha + 5b$	$8\alpha + 3b$	$8\alpha + 5b$	$9\alpha + 9$	$9\alpha + 9$
$9\alpha + 9$	$9\alpha + 8$	$9\alpha + 7$	$9\alpha + 6$	$9\alpha + 7$	$9\alpha + 8$	$9\alpha + 9$
$9\alpha + 9$	$9\alpha + 9$	$9\alpha + 9$	$9\alpha + 9$	$9\alpha + 9$	$9\alpha + 9$	$9\alpha + 9$

has been rather neglected, the usual being that of designing strategies that specify the choice for every possible outcome in the sequence of historic choices recalled [33,39].

Table 1 shows the initial scenario starting from a single defector if $8b > 9 \Leftrightarrow b > 1.125$, which means that neighbors of the initial defector become defectors at $T = 2$.

Nowak and May paid particular attention in their seminal papers to $b = 1.85$, a high but not excessive temptation value which leads to complex dynamics. After $T = 2$, defection can advance to a 5×5 square or be restrained as a 3×3 square, depending on the comparison of $8\alpha + 5 \times 1.85$ (the maximum π value of the recent defectors) with $9\alpha + 9$ (the π value of the non-affected players). As $8\alpha + 5 \times 1.85 = 9\alpha + 9 \rightarrow \alpha = 0.25$, i. e., if $\alpha > 0.25$, defection remains confined to a 3×3 square at $T = 3$. Here we see the paradigmatic effect of memory: it tends to avoid the spread of defection.

If memory is limited to the last three iterations: $\pi_{i,j}^{(T)} = \alpha^2 p_{i,j}^{(T-2)} + \alpha p_{i,j}^{(T-1)} + p_{i,j}^{(T)}$, $m_{i,j}^{(T)} = \left(\alpha^2 d_{i,j}^{(T-2)} + \alpha d_{i,j}^{(T-1)} + d_{i,j}^{(T)}\right)/(\alpha^2 + \alpha + 1)$, $\Rightarrow \delta_{i,j}^{(T)} = \mathrm{round}(m_{i,j}^{(T)})$, with assignations at $T = 2$: $\pi_{i,j}^{(2)} = \alpha\pi_{i,j}^{(1)} + \pi_{i,j}^{(2)}$, $\delta_{i,j}^{(2)} = d_{i,j}^{(2)}$.

Memory has a dramatic restrictive effect on the advance of defection as shown in Fig. 22. This figure shows the frequency of cooperators (f) starting from a single defector and from a random configuration of defectors in a lattice of size 400×400 with periodic boundary conditions when $b = 1.85$. When starting from a single defector, f at time step T is computed as the frequency of cooperators within the square of size $(2(T-1)+1)^2$ centered on the initial D site. The ahistoric plot reveals the convergence of f to 0.318, (*which seems to be the same value regardless of the initial conditions* [43]). Starting from a single defector (a), the model with small memory ($\alpha = 0.1$) seems to reach a similar f value, but sooner and in a smoother way. The plot corresponding to $\alpha = 0.2$ still shows an early decay in f that leads it to about 0.6, but higher memory factor values lead f close to or over 0.9 very

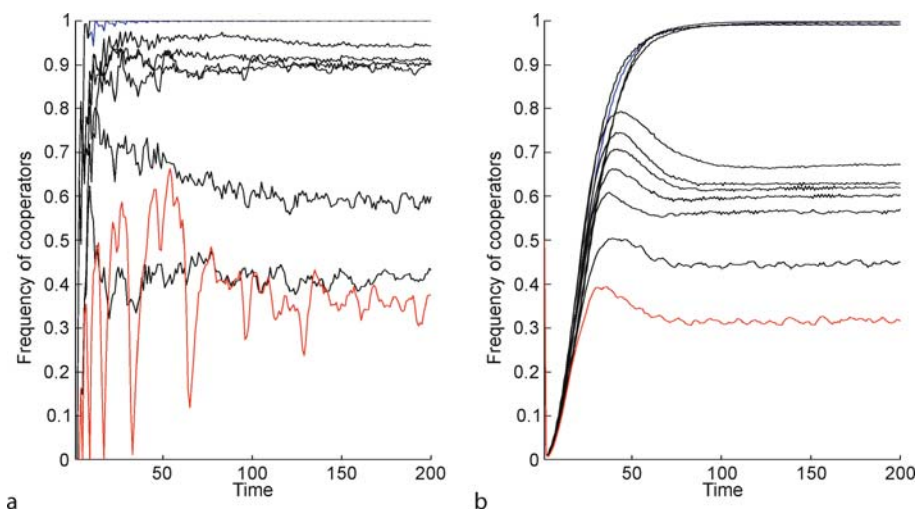

Cellular Automata with Memory, Figure 22

Frequency of cooperators (f) with memory of the last three iterations. a starting from a single defector, b starting at random ($f(1) = 0.5$). The *red curves* correspond to the ahistoric model, the *blue ones* to the full memory model, the remaining curves to values of α from 0.1 to 0.9 by 0.1 intervals, in which, as a rule, the higher the α the higher the f for any given T

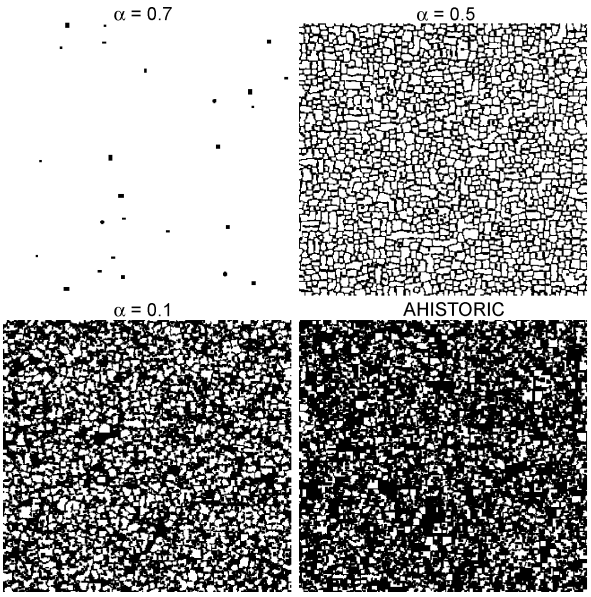

Cellular Automata with Memory, Figure 23
Patterns at $T = 200$ starting at random in the scenario of Fig. 22b

soon. Starting at random (b), the curves corresponding to $0.1 \leq \alpha \leq 0.6$ (thus with no memory of choices) do mimic the ahistoric curve but with higher f, as $\alpha \geq 0.7$ (also memory of choices) the frequency of cooperators grows monotonically to reach almost full cooperation: D persists as scattered unconnected small oscillators (*D-blinkers*), as shown in Fig. 23. Similar results are found for any temptation value in the parameter region $0.8 < b < 2.0$, in which spatial chaos is characteristic in the ahistoric model. It is then concluded that short-type memory supports cooperation.

As a natural extension of the described binary model, the 0-1 assumption underlying the model can be relaxed by allowing for *degrees* of cooperation in a continuous-valued scenario. Denoting by x the degree of cooperation of player A and by y the degree of cooperation of the player B, a consistent way to specify the pay-off for values of x and y other than zero or one is to simply interpolate between the extreme payoffs of the binary case. Thus, the payoff that the player A receives is:

$$G_A(x, y) = (x, 1 - x) \begin{pmatrix} R & S \\ \mathcal{T} & P \end{pmatrix} \begin{pmatrix} y \\ 1 - y \end{pmatrix} .$$

In the continuous-valued historic formulation it is $\delta \equiv m$, including $\delta_{i,j}^{(2)} = \left(\alpha d_{i,j}^{(1)} + d_{i,j}^{(2)}\right)/(\alpha + 1)$. Table 2 illustrates the initial scenario starting from a single (full) defector. Unlike in the binary model, in which the initial defector never becomes a cooperator, the initial defector cooperates with degree $\alpha/(1 + \alpha)$ at $T = 3$: its neighbors which received the highest accumulated payoff (those in the corners with $\pi^{(2)} = 8\alpha + 5b > 8b\alpha$), achieved this mean degree of cooperation after $T = 2$. Memory dramatically constrains the advance of defection in a smooth way, even for the low level $\alpha = 0.1$. The effect appears much more homogeneous compared to the binary model, with no special case for high values of α, as memory on decisions is always operative in the continuous-valued model [24]. The effect of unlimited trailing memory on the SPD has been studied in [5,6,7,8,9,10].

Discrete-Time Dynamical Systems

Memory can be embedded in any model in which time plays a dynamical role. Thus, Markov chains $\mathbf{p}'_{T+1} = \mathbf{p}'_T\mathbf{M}$ become with memory: $\mathbf{p}'_{T+1} = \boldsymbol{\imath}'_T\mathbf{M}$ with $\boldsymbol{\imath}_T$ being a weighted mean of the probability distributions up to T: $\boldsymbol{\imath}_T = \pi(\mathbf{p}_1, \ldots, \mathbf{p_T})$. In such scenery, even a minimal incorporation of memory notably alters the evolution of $\mathbf{p}$ [23]. Last but not least, conventional, non-spatialized, *discrete dynamical systems* become with memory: $x_{T+1} = f(m_T)$ with m_T being an average of past values. As an overall rule, memory leads the dynamics a fixed point of the map f [4].

We will introduce an example of this in the context of the PD game in which players follow the so-called Paulov strategy: a Paulov player cooperates if and only if both players opted for the same alternative in the previous move. The name Paulov stems from the fact that this strategy embodies an almost reflex-like response to the payoff: it repeats its former move if it was rewarded by $\mathcal{T}$ or R, but switches behavior if it was punished by receiving only P or S. By coding cooperation as 1 and defection as 0, this strategy can be formulated in terms of the choices x of Player A (Paulov) and y of Player B as: $x^{(T+1)} = 1 - \mid x^{(T)} - y^{(T)} \mid$. The Paulov strategy has proved to be very successful in its contests with other strategies [44]. Let us give a simple example of this: suppose that Player B adopts an Anti-Paulov strategy (which cooperates to the extent Paulov defects) with $y^{(T+1)} = 1 - \mid 1 - x^{(T)} - y^{(T)} \mid$. Thus, in an iterated Paulov-Anti-Paulov (PAP) contest, with $T(x, y) = \big(1 - |x - y|, 1 - |1 - x - y|\big)$, it is $T(0, 0) = T(1, 1) = (1, 0)$, $T(1, 0) = (0, 1)$, and $T(0, 1) = (0, 1)$, so that (0,1) turns out to be immutable. Therefore, in an iterated PAP contest, Paulov will always defect, and Anti-Paulov will always cooperate. Relaxing the 0-1 assumption in the standard formulation of the PAP contest, *degrees* of cooperation can be considered in a continuous-valued scenario. Now x and

Cellular Automata with Memory, Table 2

Weighted mean degrees of cooperation after $T = 2$ and degree of cooperation at $T = 3$ starting with a single defector in the continuous-valued SPD with $b = 1.85$

$\delta^{(2)}$				
1	1	1	1	1
1	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	1
1	$\frac{\alpha}{1+\alpha}$	0	$\frac{\alpha}{1+\alpha}$	1
1	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	1
1	1	1	1	1

$\mathbf{d}^{(3)}(\alpha < 0.25)$						
1	1	1	1	1	1	1
1	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	1
1	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	1
1	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	1
1	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	1
1	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	1
1	1	1	1	1	1	1

$\mathbf{d}^{(3)}(\alpha < 0.25)$						
1	1	1	1	1	1	1
1	1	1	1	1	1	1
1	1	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$		1
1	1	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$		1
1	1	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$		1
1	1	1	1	1	1	1
1	1	1	1	1	1	1

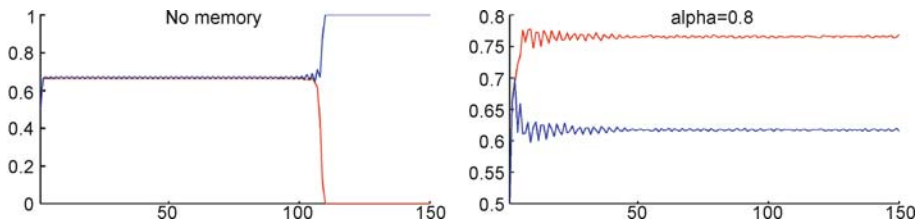

Cellular Automata with Memory, Figure 24

Dynamics of the mean values of x (*red*) and y (*blue*) starting from any of the points of the 1×1 square

y will denote the degrees of cooperation of players A and B respectively, with both x and y lying in [0,1].

In this scenario, not only (0,1) is a fixed point, but also $T(0.8, 0.6) = (0.8, 0.6)$. Computer implementation of the iterated PAP tournament turns out to be fully disrupting of the theoretical dynamics. The errors caused by the finite precision of the computer floating point arithmetics (a common problem in dynamical systems working modulo 1) make the final fate of every point to be (0,1). With no exceptions: even the *theoretically* fixed point (0.8,0.6) ends up as (0,1) in the computerized implementation.

A natural way to incorporate older choices in the strategies of decision is to feature players by a summary (m) of their own choices farther back in time. The PAP contest becomes in this way: $x^{(T+1)} = 1 - \mid m_x^{(T)} - m_y^{(T)} \mid$, $y^{(T+1)} = 1 - \mid 1 - m_x^{(T)} - m_y^{(T)} \mid$. The simplest historic extension results in considering only the two last choices: $m(z^{(T-1)}, z^{(T)}) = (\alpha z^{(T-1)} + z^{(T)})/(\alpha + 1)$ (z stands for both x and y) [10].

Figure 24 shows the dynamics of the mean values of x and y starting from any of the 101×101 lattice points of the 1×1 square with sides divided by 0.01 intervals. The dynamics in the ahistoric context are rather striking: immediately, at $T = 2$, both $\overline{x}$ and $\overline{y}$ increase from 0.5 up to app. $0.66 (\simeq 2/3)$, a value which remains stable up to app. $T = 100$, but soon after Paulov cooperation plummets, with the corresponding firing of cooperation of Anti-Paulov: finite precision arithmetics leads every point to (0,1). With memory, Paulov not only keeps a permanent mean degree cooperation but it is higher than that of Anti-Paulov; memory tends to lead the overall dynamics to the ahistoric (theoretically) fixed point (0.8, 0.6).

Future Directions

Embedding memory in states (and in links if wiring is also dynamic) broadens the spectrum of CA as a tool for modeling, in a fairly natural way of easy computer implementation. It is likely that in some contexts, a transition rule with memory could match the *correct* behavior of the CA system of a given complex system (physical, biological, social and so on). A major impediment in modeling with CA stems from the difficulty of utilizing the CA complex behavior to exhibit a particular behavior or perform a particular function. Average memory in CA tends to inhibit *complexity*, inhibition that can be modulated by varying the depth of memory, but memory not of average type opens a notable new perspective in CA. This could mean a potential advantage of CA with memory over standard CA as a tool for modeling. Anyway, besides their potential applications, CA with memory (CAM) have an aesthetic and mathematical interest on their own.

Thus, it seems plausible that further study on CA with memory should turn out profitable, and, maybe that as a result of a further rigorous study of CAM it will be possible to paraphrase T. Toffoli in presenting CAM—as an alternative to (rather than an approximation of) integro-differential equations in modeling— phenomena with memory.

Bibliography

Primary Literature

1. Adamatzky A (1994) Identification of Cellular Automata. Taylor and Francis
2. Adamatzky A (2001) Computing in Nonlinear Media and Automata Collectives. IoP Publishing, London
3. Adamatzky A, Holland O (1998) Phenomenology of excitation in 2-D cellular automata and swarm systems. Chaos Solit Fract 9:1233–1265
4. Aicardi F, Invernizzi S (1982) Memory effects in Discrete Dynamical Systems. Int J Bifurc Chaos 2(4):815–830
5. Alonso-Sanz R (1999) The Historic Prisoner's Dilemma. Int J Bifurc Chaos 9(6):1197–1210
6. Alonso-Sanz R (2003) Reversible Cellular Automata with Memory. Phys D 175:1–30
7. Alonso-Sanz R (2004) One-dimensional, $r = 2$ cellular automata with memory. Int J Bifurc Chaos 14:3217–3248
8. Alonso-Sanz R (2004) One-dimensional, $r = 2$ cellular automata with memory. Int J BifurcChaos 14:3217–3248
9. Alonso-Sanz R (2005) Phase transitions in an elementary probabilistic cellular automaton with memory. Phys A 347:383–401 Alonso-Sanz R, Martin M (2004) Elementary Probabilistic Cellular Automata with Memory in Cells. Sloot PMA et al (eds) LNCS, vol 3305. Springer, Berlin, pp 11–20
10. Alonso-Sanz R (2005) The Paulov versus Anti-Paulov contest with memory. Int J Bifurc Chaos 15(10):3395–3407
11. Alonso-Sanz R (2006) A Structurally Dynamic Cellular Automaton with Memory in the Triangular Tessellation. Complex Syst 17(1):1–15. Alonso-Sanz R, Martin, M (2006) A Structurally Dynamic Cellular Automaton with Memory in the Hexagonal Tessellation. In: El Yacoubi S, Chopard B, Bandini S (eds) LNCS, vol 4774. Springer, Berlin, pp 30-40
12. Alonso-Sanz R (2007) Reversible Structurally Dynamic Cellular Automata with Memory: a simple example. J Cell Automata 2:197–201
13. Alonso-Sanz R (2006) The Beehive Cellular Automaton with Memory. J Cell Autom 1:195–211
14. Alonso-Sanz R (2007) A Structurally Dynamic Cellular Automaton with Memory. Chaos Solit Fract 32:1285–1295
15. Alonso-Sanz R, Adamatzky A (2008) On memory and structurally dynamism in excitable cellular automata with defensive inhibition. Int J Bifurc Chaos 18(2):527–539
16. Alonso-Sanz R, Cardenas JP (2007) On the effect of memory in Boolean networks with disordered dynamics: the $K = 4$ case. Int J Modrn Phys C 18:1313–1327
17. Alonso-Sanz R, Martin M (2002) One-dimensional cellular automata with memory: patterns starting with a single site seed. Int J Bifurc Chaos 12:205–226
18. Alonso-Sanz R, Martin M (2002) Two-dimensional cellular automata with memory: patterns starting with a single site seed. Int J Mod Phys C 13:49–65
19. Alonso-Sanz R, Martin M (2003) Cellular automata with accumulative memory: legal rules starting from a single site seed. Int J Mod Phys C 14:695–719
20. Alonso-Sanz R, Martin M (2004) Elementary cellular automata with memory. Complex Syst 14:99–126
21. Alonso-Sanz R, Martin M (2004) Three-state one-dimensional cellular automata with memory. Chaos, Solitons Fractals 21:809–834
22. Alonso-Sanz R, Martin M (2005) One-dimensional Cellular Automata with Memory in Cells of the Most Frequent Recent Value. Complex Syst 15:203–236
23. Alonso-Sanz R, Martin M (2006) Elementary Cellular Automata with Elementary Memory Rules in Cells: the case of linear rules. J Cell Autom 1:70–86
24. Alonso-Sanz R, Martin M (2006) Memory Boosts Cooperation. Int J Mod Phys C 17(6):841–852
25. Alonso-Sanz R, Martin MC, Martin M (2000) Discounting in the Historic Prisoner's Dilemma. Int J Bifurc Chaos 10(1):87–102
26. Alonso-Sanz R, Martin MC, Martin M (2001) Historic Life Int J Bifurc Chaos 11(6):1665–1682
27. Alonso-Sanz R, Martin MC, Martin M (2001) The Effect of Memory in the Spatial Continuous-valued Prisoner's Dilemma. Int J Bifurc Chaos 11(8):2061–2083
28. Alonso-Sanz R, Martin MC, Martin M (2001) The Historic Strategist. Int J Bifurc Chaos 11(4):943–966
29. Alonso-Sanz R, Martin MC, Martin M (2001) The Historic-Stochastic Strategist. Int J Bifurc Chaos 11(7):2037–2050
30. Alvarez G, Hernandez A, Hernandez L, Martin A (2005) A secure scheme to share secret color images. Comput Phys Commun 173:9–16
31. Fredkin E (1990) Digital mechanics. An informal process based on reversible universal cellular automata. Physica D 45:254–270
32. Grössing G, Zeilinger A (1988) Structures in Quantum Cellular Automata. Physica B 15:366
33. Hauert C, Schuster HG (1997) Effects of increasing the number of players and memory steps in the iterated Prisoner's Dilemma, a numerical approach. Proc R Soc Lond B 264:513–519
34. Hooft G (1988) Equivalence Relations Between Deterministic and Quantum Mechanical Systems. J Statistical Phys 53(1/2):323–344
35. Ilachinski A (2000) Cellular Automata. World Scientific, Singapore
36. Ilachinsky A, Halpern P (1987) Structurally dynamic cellular automata. Complex Syst 1:503–527
37. Kaneko K (1986) Phenomenology and Characterization of coupled map lattices, in Dynamical Systems and Sigular Phenomena. World Scientific, Singapore
38. Kauffman SA (1993) The origins of order: Self-Organization and Selection in Evolution. Oxford University Press, Oxford
39. Lindgren K, Nordahl MG (1994) Evolutionary dynamics of spatial games. Physica D 75:292–309
40. Love PJ, Boghosian BM, Meyer DA (2004) Lattice gas simulations of dynamical geometry in one dimension. Phil Trans R Soc Lond A 362:1667
41. Margolus N (1984) Physics-like Models of Computation. Physica D 10:81–95

42. Martin del Rey A, Pereira Mateus J, Rodriguez Sanchez G (2005) A secret sharing scheme based on cellular automata. Appl Math Comput 170(2):1356–1364
43. Nowak MA, May RM (1992) Evolutionary games and spatial chaos. Nature 359:826
44. Nowak MA, Sigmund K (1993) A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. Nature 364:56–58
45. Requardt M (1998) Cellular networks as models for Plank-scale physics. J Phys A 31:7797; (2006) The continuum limit to discrete geometries, arxiv.org/abs/math-ps/0507017
46. Requardt M (2006) Emergent Properties in Structurally Dynamic Disordered Cellular Networks. J Cell Aut 2:273
47. Ros H, Hempel H, Schimansky-Geier L (1994) Stochastic dynamics of catalytic CO oxidation on Pt(100). Pysica A 206:421–440
48. Sanchez JR, Alonso-Sanz R (2004) Multifractal Properties of R90 Cellular Automaton with Memory. Int J Mod Phys C 15:1461
49. Stauffer D, Aharony A (1994) Introduction to percolation Theory. CRC Press, London
50. Svozil K (1986) Are quantum fields cellular automata? Phys Lett A 119(41):153–156
51. Toffoli T, Margolus M (1987) Cellular Automata Machines. MIT Press, Massachusetts
52. Toffoli T, Margolus N (1990) Invertible cellular automata: a review. Physica D 45:229–253
53. Vichniac G (1984) Simulating physics with Cellular Automata. Physica D 10:96–115
54. Watts DJ, Strogatz SH (1998) Collective dynamics of Small-World networks. Nature 393:440–442
55. Wolf-Gladrow DA (2000) Lattice-Gas Cellular Automata and Lattice Boltzmann Models. Springer, Berlin
56. Wolfram S (1984) Universality and Complexity in Cellular Automata. Physica D 10:1–35
57. Wuensche A (2005) Glider dynamics in 3-value hexagonal cellular automata: the beehive rule. Int J Unconv Comput 1:375–398
58. Wuensche A, Lesser M (1992) The Global Dynamics of Cellular Automata. Addison-Wesley, Massachusetts

Books and Reviews

Alonso-Sanz R (2008) Cellular Automata with Memory. Old City Publising, Philadelphia (in press)

Cellular Automata Modeling of Complex Biochemical Systems

Lemont B. Kier[1], Paul G. Seybold[2]
[1] Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, USA
[2] Wright State University, Dayton, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Models and Simulations with Cellular Automata
The Grid
The Cells
The Rules
Examples
Future Directions
Bibliography

Glossary

Agents Ingredients under study occupying the cells in the grid.

Asynchronous Each cell in turn, at random responds to a rule.

Cells Sections of the grid in which the agents exist.

Gravity The relative relationship of two agents within the frame of a grid with opposite bound edges.

Grid The frame containing the cells and the agents.

Movement The simulation of agent movement between cells accomplished by the disappearance of an agent in a cell and the appearance of that agent in an adjacent cell.

Neighborhood The cells immediately touching a given cell in the grid.

Rules Statements of actions to be taken by an agent under certain conditions. They may take the form of a probability of such an action.

Synchronous All cells in a grid exercise a simultaneous response to a rule.

Definition of the Subject

Cellular automata are discrete, agent-based models that can be used for the simulation of complex systems [1]. They are composed of:

- A grid of cells.
- A set of ingredients called agents that can occupy the cells.
- A set of local rules governing the behaviors of the agents.
- Specified initial conditions.

Once these components are defined, a simulation can be carried out. During the simulation the system evolves via a series of discrete time-steps, or iterations, in which the rules are applied to all of the ingredients of the system and the configuration of the system is regularly updated. A striking feature of the cellular automata (CA) models is that they treat not only the ingredients, or agents, of the model as discrete entities, as do the traditional models of

physics and chemistry, but in the CA models time (iterations) and space (the cells) are also regarded as discrete, in contrast to the continuous forms assumed for these variables in the traditional, equation-based physical models. In practice, this distinction usually makes little difference, since the traditional continuous results appear naturally as limiting cases of the discrete CA analysis.

Introduction

Cellular automata (CA) were first proposed by the mathematical physicist John von Neumann [2] and the mathematician Stanislaw Ulam [3,4] more than a half century ago, and similar ideas were suggested at about the same time, in the 1940s, by the German engineer Konrad Zuse. Von Neumann's interest was in the construction of "self-reproducing automata". His original idea was to construct a series of mechanical devices or "automata" that would gather and assemble the parts necessary to reproduce themselves. A suggestion by Ulam led him to consider more abstract systems consisting of grids with moving agents, operating under sets of rules. The first such system proposed by von Neumann consisted of a two-dimensional grid of square cells, each having a set of possible states, along with a set of rules. With the development of modern digital computers, it has became increasingly clear that these very abstract ideas could in fact be usefully applied to the examination of real physical and biological systems, with interesting and informative results.

A number of research groups have developed different realizations of the CA paradigm for the study of a broad range of physical [5,6], chemical [7], biological [8,9], medical [10,11] and even sociological [12,13,14] phenomena. These models have contributed important new insights regarding the deeper, often hidden, factors underlying a host of complex phenomena. These diverse CA studies have been especially important in treating the often-surprising behaviors of systems where large numbers of diverse interactions between the system agents serve to hide the general patterns involved and, in addition, render the conventional, differential-equation-based methods difficult to implement or ineffective—i. e., complex systems.

Models and Simulations with Cellular Automata

Cellular automata can be used to both model and to simulate real systems. A model is a substitute, usually greatly simplified, for what is the real thing. A model should capture from the real system and display in some revealing way its most important or interesting features. Where possible it should capture the essence of the system without being overly cumbersome or complicated. Many models in the physical sciences take the form of mathematical relationships, equations connecting some property with other parameters of the system. Some of these relationships are quite simple, while many mathematical relationships are more complicated, and rely on the techniques of calculus to describe the rates of change of the quantities involved. Overall, mathematical models have been exceedingly successful in depicting the broad outlines of an enormously diverse variety of phenomena in nature.

Simulations are active imitations of real things, and there are generally two different types of simulations, with different aims. In one approach a simulation is merely designed to match a certain behavior, often in a very limited context. Thus a mechanical noisemaker may simulate a desired sound, and does so through a very different mechanism than the real sound-maker. Such a simulation reveals little or nothing about the features of the original system, and is not intended to do so. Only the outcome, to some extent, matches reality. A hologram may look like a real object, but it is constructed from interfering light waves. A second type of simulation is more ambitious. It attempts to mimic at least some of the key features of the system under study, with the intent of gaining insight into how the system operates.

The Grid

The grid in a CA model may contain a single cell, or more commonly a larger collection. The grid might be one-, two-, or three-dimensional in form, although most studies have used two-dimensional grids. A moving agent may encounter an edge or boundary during its response to the rules. Three general types of two-dimensional grids are considered relating to the boundaries: (i) a box, (ii) a cylinder, and (iii) a torus. In the box grid moving agents encounter boundaries on all four sides; in the cylinder they encounter only top and bottom boundaries; and in the torus no boundaries restrict the agent movements. An illustration of a $7 \times 7 = 49$-cell grid of square cells occupied by two different types of agents, A and B, is shown in Fig. 1.

The nature of the grid type employed will normally depend on the boundary characteristics of the system of interest. For some systems, e. g., when the agents are either stationary or confined, a box grid is perfectly suitable. In other cases, one may need only a constraining top and bottom (or right and left sides), and a cylindrical grid will be used. The torus effectively simulates a small segment of a larger, unrestricted system by allowing cells to move off the top edge and appear at the bottom, or move off the bottom edge and appear at the top.

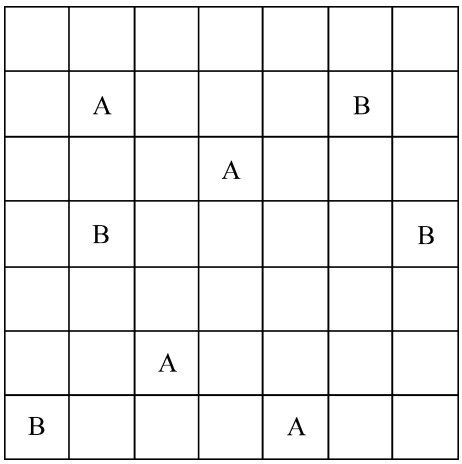

Cellular Automata Modeling of Complex Biochemical Systems, Figure 1
A two dimensional cellular automata grid with two different agents

The Cells

Structure

The cells can take a variety of shapes; they can be triangles, squares, hexagons or other shapes on the two-dimensional grid, with square cells being most common. Each cell in the grid can normally exist in a number of distinct states which define the occupancy of the cell. The cell can be empty or contain a specific agent, representing a particle, a type of molecule or isomer, a molecular electronic state, or some other entity pertinent to the study in question. The choice of the cell shape is based on the objective of the study. In the case of studies of water-related phenomena, for example, square cells are especially advantageous since water molecules, are quadravalent with respect to their participation in intermolecular hydrogen bonding. An individual water molecule can employ two hydrogen atoms and two lone pairs of electrons to form hydrogen bonds with its neighbors. This leads to the tetrahedral configuration found in ice, a structure that is retained to some extent in the liquid state. The four faces of a square cell thus correspond to the four hydrogen-bonding opportunities of a water molecule.

The interactions of an agent with other agents take place at the cell edges. Originally, cellular automata models routinely assumed that all of the edges of a given agent should obey the same rules. More recently, the idea of a variegated cell, in which each edge can have its own independent rules for interaction with other agents, has been introduced and shown to have considerable value in modeling. Examples of some types of variegated cells are shown in Fig. 2.

Cell Neighborhoods

The movements and other actions of an agent on the grid are governed by rules, that depend only on the nature of the cells in close proximity to the agent. This proximate environment of a cell is called its neighborhood. The most common neighborhood used in two-dimensional cellular automata studies is called the von Neumann neighborhood, after the pioneer of the CA method. This neighborhood for a cell A refers to the four B cells adjoining its four faces (see Fig. 3). Another common neighborhood is the Moore neighborhood referring to the eight cells completely surrounding cell A, including those cells on the diagonals.

The Rules

Several different types of rules govern the behaviors of the agents on the grid and thereby the subsequent evolutions of the CA systems. The key features of all these rules are that they are local, involving only an agent itself and possibly those agents in its immediate neighborhood, and that they are uniformly applied throughout a CA study.

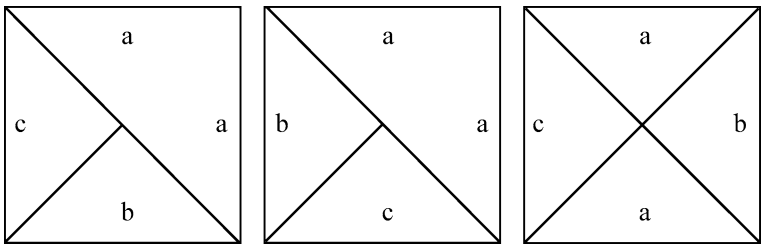

Cellular Automata Modeling of Complex Biochemical Systems, Figure 2
Cellular automata variegated cells

		B		
	B	A	B	
		B		

Cellular Automata Modeling of Complex Biochemical Systems, Figure 3
The von Neumann neighborhood

Movement Rules

Much of the dynamic character of cellular automata models is developed through the movements of the agents about the grid. During each time-step interval, or iteration, in the CA simulation an agent on the grid has the possibility of moving vertically or horizontally to an adjacent, unoccupied cell. In the absence of further restrictions, a free agent would therefore, over time, perform a random walk about the grid. Normally, however, there are other agents on the grid and their presence will influence the motion of the first agent. During each iteration the movement of every agent on the grid is computed based on rules that involve the status of its neighboring cells, i. e., whether these cells are empty or occupied, and, if occupied, by what types of agents. Deterministic cellular automata use a fixed set of rules, the values of which are immutable and uniformly applied to the agents. In probabilistic, or stochastic, cellular automata, the movements of the agent are based on probabilistic rules, embodied as probabilities of moving or not moving during each iteration.

Free Moving Probability P_m

The free moving probability $P_m(A)$ defines the probability that an agent A in a cell i will move to one of the four adjacent cells, j, in its von Neumann neighborhood if that space is unoccupied. An example would be the agent in cell A which might move to any of the unoccupied neighboring B cells. This probability is usually set at $P_m = 1.0$, which means that a movement in one of the allowed directions always happens. However, in some cases P_m can be set to lower values if certain species in the CA simulation are to be regarded as moving more slowly than others.

Joining Parameter J

The first of the two trajectory or interaction rules is the joining trajectory parameter, $J(AB)$, which defines the propensity of movement of an agent A toward or away from a second agent B when the two are separated by a vacant cell. It thus involves the extended von Neumann neighborhood of agent A, and has the effect of adding a short-range attraction or repulsion component to the interaction between agents A and B. J is a non-negative real number. When $J = 1$, species A has the same probability of movement toward or away from B as when the B cell is not present. When J is greater than 1, agent A has a greater probability of movement toward a B agent than when agent B is absent, simulating, in effect, a degree of short-range attraction. When J lies between 0 and 1, agent A has a lower probability of such movement, and this can be considered as a degree of mutual repulsion. When $J = 0$, agent A cannot move toward B at all.

Breaking Probability P_B

The second trajectory or interaction rule is the breaking probability, P_B. This parameter, in effect, assigns a persistence to the encounter between two agents that are in contact, i. e., adjacent to each other on the grid. The breaking rule assigns the probability $P_B(AB)$ that an agent A, adjacent to an agent B, will break apart from B. The value for P_B necessarily lies within the range 0 to 1. Low values of P_B imply a strong cohesion between A and B, whereas high values indicate little cohesion. Thus if $P_B = 0$, the agents will not separate from each other, and if $P_B = 1$ they have no tendency to adhere to one another. If P_B lies between these values there is an intermediate tendency to break apart. When molecule A is bordered two agents, B and C, the simultaneous probability of A breaking away is given by the product $P_B(AB) * P_B(AC)$. If agent A has three adjacent agents (B, C, and D), the simultaneous breaking probability of agent A, the probability that it will move to the remaining adjacent empty cell, is $P_B(AB) * P_B(AC) * P_B(AD)$. If agent A is surrounded by four agents, it cannot move.

Transition Rules

Transitions occur constantly in nature; molecules change from one tautomeric form to another, radioactive nuclei decay to form other nuclei, acids dissociate, proteins al-

ter their shapes, molecules undergo transitions between electronic states, chemicals react to form new species, and so forth. Transition rules allow the simulation of these changes. They govern the probability that during each iteration of the simulation an agent will transform to a different type of agent. If $P_T(AB) = 1.0$ the transition $A \rightarrow B$ is certain to occur; if $P_T(AB) = 0.0$, it will never occur. But if, for example, $P_T(AB) = 0.5$, then during each iteration there will be a 50% chance that the transition $A \rightarrow B$ will occur. The first two cases can be considered deterministic, since they do not allow for different outcomes. The third case is stochastic, however, since it allows different outcomes, the agent might remain unchanged or it might transform to a different state. The transition probabilities may, in some cases, depend on the conditions prevailing in neighboring cells. For example, the transformation probability $P_T(AB)$ might depend on the occupancies of neighboring cells. In reaction simulations two agents A and B that come in contact on the grid will have a probability $P_R(AB)$ of reacting, or transforming, to other species, say, C and D, during such an encounter. In this case the reaction probability $P_R(AB)$ defines the probability that the reaction $A + B \rightarrow C + D$ will occur when A and B encounter one another in the course of their motions. If $P_R(AB) = 1$ the reaction will take place on every encounter, but if $P_R(AB) = 0.1$, for example, only 10% of such encounters will lead to reaction.

Relative Gravity Rules

The simulation of a gravity effect can be introduced into a cellular automaton model in two different ways. Separation phenomena like the de-mixing of immiscible liquids can be simulated using a relative gravity rule. For this, a boundary condition is first imposed at the upper and lower edges of the grid to apply vertical limits on the motions of the agents (a cylindrical grid). The differential effect of gravity on different agents A and B is simulated by introducing reciprocal rules governing their tendencies to exchange positions when they come together. When one agent moves to a position on top of the other the rules are applied. The first rule, $G_R(AB)$, applies when A is above B and is the probability that agent A will exchange places with agent B, so that A will appear below, and B above. The complementary rule is $G_R(BA)$, which expresses the probability that molecule B, originally above A, will exchange positions with A and end up below.

When $G_R(AB)$ is greater than $G_R(BA)$ there will be an overall tendency for the A agents to congregate below the B agents, and when $G_R(AB)$ is less than $G_R(BA)$ the A agents will tend toward the upper part of the grid. In the first case the A's can be thought of as forming a more dense liquid than the B's, and in the latter case, a less dense liquid. The G_R rules are probabilities that the events will occur.

Absolute Gravity Rule

In other simulations an absolute gravity rule, denoted by $G_A(A)$, is more appropriate. This rule favors motion in a preferred direction. For example, one might wish to simulate the motions of different gas molecules, some heavier than others, in a gravitational field. The value $G_A(A) = 0$ is the neutral value, so that the movement probabilities are equal in all four directions. Values greater than $G_A(A) = 0$ increase the likelihood of downward movements. Thus a value of $G_A(A) = 0.2$ would impose a slight tendency of the agents of species A to move downward on the grid, and $G_A(A) = 0.5$ would impose a much stronger tendency.

Cell Rotation Rules

In those cases where variegated agents are used it is necessary to insure that there exists a balanced representation of the possible rotational states of these agents on the grid. To accomplish this, the variegated cells are rotated randomly, by 90°, −90°, +180°, or −180°, during every iteration of the run. Only free cells rotate; when a variegated agent has a neighboring agent in its von Neumann neighborhood it does not rotate.

Application of the Rules

A complete time-step (iteration) in a CA model involves the application of all of the applicable rules to all of the agents on the grid. During an iteration the movement rules can be applied either simultaneously (synchronously) or sequentially (asynchronously). Synchronous application of the governing movement rules for a CA simulation, as outlined above, can lead in some instances to conflicts, e. g., the assigning of two agents to move to the same empty cell. As a result, synchronous rule application may not be practical for cellular automata studies of some kinds. In asynchronous application of the rules, in contrast, agents are selected in random order for application of the movement rules, and such potential conflicts are avoided.

Examples

Structure of Bulk Water

The physical properties of water, such as its viscosity, heat capacity, heat of vaporization, surface tension, dielectric

constant, etc., are commonly related to temperature under normal pressure. It is accepted that underlying changes in the structure of water are responsible for the changes in the properties, although the precise physical basis for this relationship has not been established. Cellular automata models of bulk water in the liquid state at different temperatures produce a series of structural patterns, differing in the extent of binding between the individual water molecules. The fractions of water molecules unbound, bound to one, two, three or four neighbors, designated, f_0, f_1, f_2, f_3, and f_4, can be employed as structural descriptors of bulk liquid water at various temperatures. The significance of these descriptors can be evaluated by examining the correlations of these descriptors with these physical properties. A linkage between the J and P_B rules for water models has been postulated and tested leading to the relationship [15]:

$$\text{Log}\, J = -1.5P_B + 0.6\,. \tag{1}$$

The temperature of the water has been found to relate with the P_B value as

$$\text{T}\left(^\circ\text{C}\right) = 100P_B\,. \tag{2}$$

Using the f_x parameters as structure parameters in regression analysis, correlations arise between them and a variety of physical properties. Several relationships are given in Table 1, where r^2 is the coefficient of determination and s is the standard error of the fit. The observed relationships suggest that the bulk water f_x descriptors encode information allowing them to be used in correlations with a variety of physical properties. This reveals possibilities of using these descriptors in analyzing properties and behavior entirely on the basis of structure rather than using an inter-connected web of physical properties. The relationship between the f_x descriptors and the temperature (P_B value) is shown in Fig. 4.

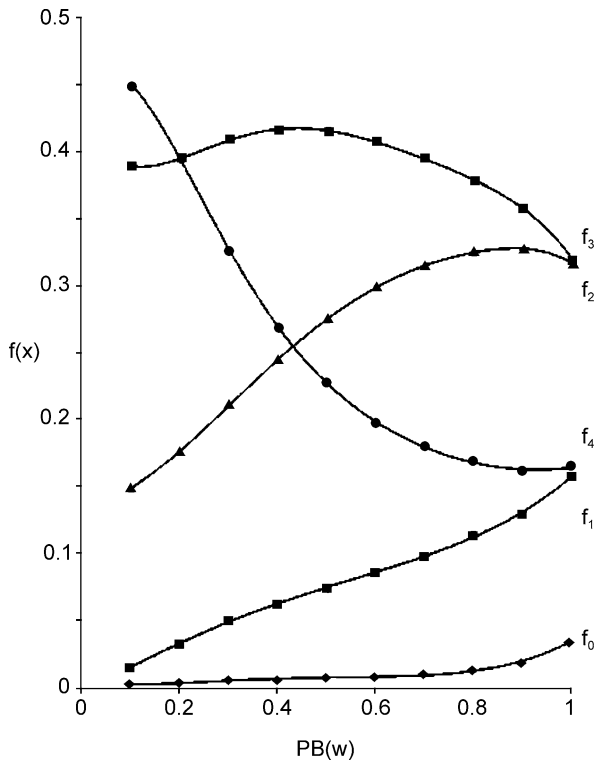

Cellular Automata Modeling of Complex Biochemical Systems, Figure 4
The calculated distribution of water bonding states, *f(x)*, and the modeled temperature using *PB(W)* values

Solution Phenomena

Solute Rules The solubility of a compound in a particular solvent reflects the intermolecular interactions between the molecular species involved. Thus the breaking and joining rules pertaining to solute-solute and solute-solvent interactions may be assigned on the basis of some anticipation of the magnitude of these intermolecular terms. For the solute (S), a large value of J(SS) encodes a high tendency for its molecules to be attracted to one another.

Cellular Automata Modeling of Complex Biochemical Systems, Table 1
Equations relating f_x values to properties of water

Property	Equation	r^2	s
Temperature	$T\,(^\circ\text{C}) = -490.28f_0 + 622.60f_1 + 4.46$	0.996	1.90
Heat capacity	$C_p\,[\text{Cal/g}/^\circ\text{C}] = 1.0478 - 0.0488f_2 - 0.0620f_3 - 0.0446f_4$	0.995	0.0002
Surface tension	$\gamma\,(\text{dynes/cm}) = -93.72f_1 + 75.33$	0.996	1.90
Vapor pressure	$\text{Log}\,P_v\,(\text{mm Hg}) = -24.30f_0 + 15.64f_1 + 0.90$	0.997	0.035
Compressibility	$\kappa\,(10^6/\text{Bar}) = 79.60 - 43.61f_2 - 39,57f_3 - 30.32f_4$	0.991	0.190
Dielectric constant	$\varepsilon = -178.88f_1 + 55.84$	0.994	0.735
Viscosity	$\eta\,(\text{centipoises}) = 1.439f_4 + 0.202$	0.965	0.045

The companion rule, $P_B(SS)$, encodes the ease or difficulty of a pair of joined solute molecules to break apart. To illustrate the cooperativity of these two rules for a solute, a molecule with rules $J(SS) = 2.0$ and $P_B(SS) = 0.2$ would be expected to have a strong tendency to remain as an insoluble aggregate or crystal. Such a solute would likely have a high melting point. In contrast, a solute with $J(SS) = 0.5$ and $P_B(SS) = 0.8$ would be expected to have a lower melting point and be more soluble. The choice of these rules govern the general behavior of solute molecules toward themselves [16].

The Hydrophobic Effect The hydrophobic effect refers to the action of relatively non-polar (hydrophobic) substances on the organization of the water molecules in their vicinity. A common expression is that water becomes "more structured" or organized in the neighborhood of a hydrophobic solute. This phenomenon has been simulated using a cellular automata model [17]. In this model the breaking probability of water-solute molecule pairs, $P_B(WS)$, was systematically increased, thus encoding an increasing hydrophobicity of the solute and a decreasing probability that the solute molecules would associate with the water molecules.

It was observed that low $P_B(WS)$ values, representing strong water-solute attractions, produced solution configurations in which the solute molecules were heavily surrounded by water molecules. Conversely, at high $P_B(WS)$ values most of the solute molecules were found outside of the water clusters and within the solution cavities. These configurations leave the water clusters relatively free of solute and hence more internally structured or organized. These models thus reflect the molecular system level conditions present in the hydrophobic effect. These results agree with molecular dynamics simulations and the interpretation presented agrees with recent experimental evidence and models proposed for the hydrophobic effect.

Solute Dissolution Dissolution refers to the breaking apart of a solute aggregate within a solvent to form a solution. Cellular automata simulations of the dissolution process have been carried out in which the solute aggregates started as solid blocks surrounded by water [18]. The solute species were endowed with specific $P_B(SS)$, $J(SS)$, $P_B(WS)$ and $J(WS)$ rules. The system attributes recorded during the ensuing system evolutions were the fraction $f_0(S)$ of solutes unbound to other solute molecules, the average number of solute-solute joined faces $T(S)$, and the average distance $D(S)$ that solute molecules traveled from the center of the block at a specific iteration. The $f_0(S)$ values were interpreted as representing the extent of dissolution of the solute, the decrease in the $T(S)$ values were used to characterize the extent of disruption of the solute block, and the $D(S)$ values quantified the extent of diffusion of the solutes into the surrounding water.

The extent and rate of the solute block disruption, $T(S)$, is primarily a function of the $P_B(S)$ rule, with secondary influence from the $P_B(WS)$ rule. A high value of $P_B(S)$ implies a high probability of solute molecules separating from each other, and hence a strong tendency toward crystal disruption. The f_0 fraction serves as a measure of the extent of solute solubility. The extent of dissolution depends upon both the $P_B(S)$ and the $P_B(WS)$ rules. High values of $P_B(S)$ and low values of $P_B(WS)$ promote an extensive degree of dissolution. A low value of $P_B(WS)$ characterizes a solute that is relatively hydrophilic. The simulations also implied that solutes with high $P_B(WS)$ values should diffuse more rapidly than those with low values of equal size. It was also observed that simulations of higher temperatures led to faster disruption of the block of solute, and more extensive diffusion of the solute molecules through the bulk water, in accord with the normal expectations.

An unexpected observation arises from the graphical display of the disintegration of the solute block of cells. The early stages of the disruption occurred as a series of intrusions of cavities, rather than water molecules, into the block. The cavities roamed throughout the block, behaving as "particles". The entrance of water into the block structure appeared much later, after significant disruption and loss of solute molecules from the block. This behavior is shown in Fig. 5.

Diffusion in Water

Models of the diffusion process in water have been reported. One study revealed that hydrophobic solutes diffuse faster than hydrophilic ones, comparing molecules of similar size [19]. A model of diffusion as a function of water temperature produced the expected result of greater diffusion with higher temperature. A series of dilute solutions were modeled where the relative polarity of a solute S_1, was varied. A second solute, S_2, was introduced at the center of the grid. The dynamics showed that the solute, S_2, diffused faster when there is no co-solute and when the polarity of S_2 is low. The presence of the co-solute revealed that the diffusion of S_2 was fastest when the co-solvent, S_1 was hydrophobic.

In another study, the grid was divided into two halves, the upper half, containing a solution of a non-polar solute while the lower half contained a solution of a polar solute [20]. Between these two halves, a thin layer of cells

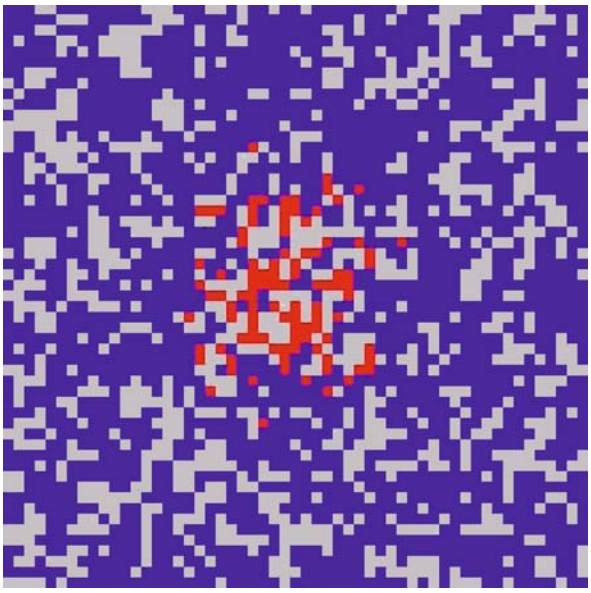

Cellular Automata Modeling of Complex Biochemical Systems, Figure 5
A model of the dissolution of a crystal in water

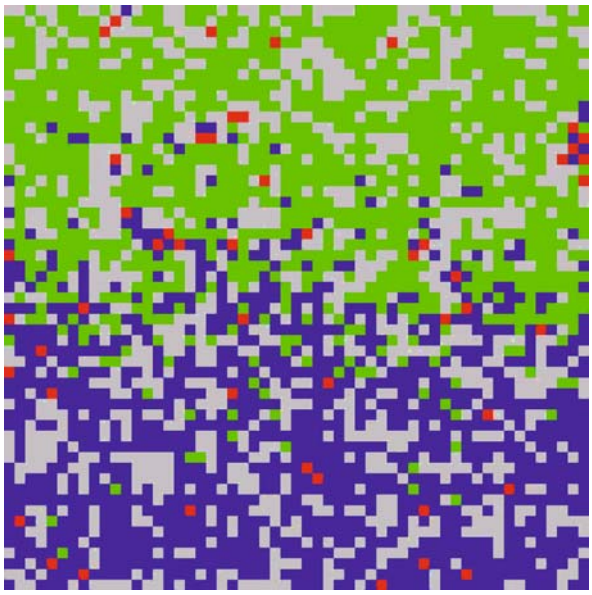

Cellular Automata Modeling of Complex Biochemical Systems, Figure 6
A model of the de-mixing of immiscible liquids

simulated a solution made up of a solute of intermediate polarity. The dynamics created a model in which the solute in the middle layer preferentially diffused into the upper layer containing the non-polar solute. A second study simulated an aqueous solution of intermediate polarity in the middle of the grid. Surrounding this solution were regions with highly polar, intermediately polar and non-polar solutes in solution. A fourth region around the center was pure water. The solute in the central section of the grid was allowed to freely diffuse. It diffused more rapidly into the pure water, however it secondarily preferred the more non-polar solution quadrant. The diffusion of a solute is modeled to be faster when it is hydrophobic and when co-solutes are also hydrophobic. The rate of diffusion of a solute in water was studied as a function of the water temperature and the hydropathic state of the solute.

Oil-Water Demixing

Models of the separation of two immiscible liquids and the partitioning of a solute between them have been reported [21]. The emerging configuration is an immiscible, two phase system. An interface formed with a greater concentration of water in the lower half while the second liquid dominated the upper half. The interface was unorganized with large "fingers" of cells from each half projecting into the other half of the grid. Figure 6 shows the interface in this model. This interface structure and behavior has been observed using other modeling paradigms. The simulated events of the demixing process present an intriguing model of a phenomenon that might be difficult to examine experimentally.

Another study tested the ability of this model to simulate the partitioning of solute molecules between the two phases, governed by their relative hydrophobicity. The addition of a small number of solute molecules was made to the initial, random mixture. As the dynamics proceeded, it was observed that the solute was associating with the patches of solvent to which it had the closest parameter-governed affinity. As a relatively stable configuration developed, the ratio of the solute molecules among the two phases became relatively constant. This ratio is the partition coefficient of the solute between the two phases. The dominant rules influencing the partition coefficient were the $P_B(WS_2)$ and $P_B(S_1S_2)$, the affinity of the solute for the two liquids, where S_2 is the solute molecule.

Observing the course of the dynamics there is a constantly changing pattern from the random configuration at the outset to the eventual formation of a disturbed interface and separated compartments of the two solvents. The solute molecules move rapidly to the patches in which the rules have ordained an affinity. The solute molecules have essentially partitioned themselves among the patches long before the two phases and the interface have formed.

Chemical Kinetics

Acid Dissociation The dissociation of an acid and the influence of the environment on this process was the subject of a recent study [22]. A cell modeling an organic acid molecule was divided into two parts, one face representing the carboxyl group, Y, and the other three faces, X, representing the non-dissociating, non-polar parts of the acid molecule. The strength of the acid, i. e. its propensity to dissociate, was governed by the probability rule, P_D. The hydronium ion was endowed with greater mobility since it is known to move very rapidly from one oxygen to another within the hydrogen bonded system of water. This was accomplished by allowing any of the four possible neighboring water molecules, relative to H to exchange positions with the hydronium cell, H.

An initial test of the model was to vary the P_D value and monitor the concentration of products. As expected, an increase in the P_D rule produced an increase in the calculated acid dissociation constant, K_a. A second study examined the influence of acid concentration on the observed properties. As expected, the K_a was approximately constant over a modest concentration range. A study on solution environment influences modeled the presence of another molecule in the solution. This co-solute was endowed with an attribute of non-dissociation. It's hydrophobicity was varied in a series of studies. The dissociation of the acid decreased when the hydrophobicity of the co-solute decreased. The interaction of two acids of different strength was also simulated using the same basic model. The observed dissociations revealed a strong and unequal influence of the two acids on each other. Both acids exhibit a suppression in their dissociations relative to their behavior in pure solution. The weaker acid is significantly more suppressed than the stronger acid. The decrease in dissociation of the two acids in a mixture cannot be readily calculated from the acid concentrations and their individual dissociation constants because of the complicating influences of ionic solvation effects on the water structure plus temperature factors.

First-Order Kinetics Many important natural processes ranging from nuclear decay to unimolecular chemical reactions, are first-order, or can be approximated as first-order [23]. This means that these processes depend only on the concentration to the first power of the transforming species itself. A cellular automaton model for such a system takes on an especially simple form, since rules for the movements of the ingredients are unnecessary and only transition rules for the inter-converting species need to be specified. Recent work described such a general cellular automaton model for first-order kinetics and tested its ability to simulate a number of classic first-order phenomena.

The prototype first-order transition is radioactive decay A → B, in which the concentration [A] of a species A decreases according to the rule that each A ingredient has a probability $P_t(\mathrm{AB})$ per unit time (here, per iteration) of converting to some other form B. For small numbers of A ingredients the actual decay curve observed for [A] is rather jagged and only roughly exponential, as a result of the irregular decays expected in this very finite, stochastic system. However, as the number of decaying ingredients is increased the decay curve approaches the smooth exponential fall-off expected for a deterministic system obeying the rate equation

$$\frac{\mathrm{d}[A]}{\mathrm{d}t} = -k[A] \,. \tag{3}$$

When a reverse transition probability $P_t(\mathrm{BA})$ for the transition B → A is included the model simulates the first-order equilibrium:

$$\mathrm{A} \leftrightarrow \mathrm{B} \,.$$

Here too, the finite size of the system causes notable fluctuations, in this case in the value of the equilibrium constant K, which fluctuates with time about the deterministic value

$$\mathrm{K} = \frac{P_t(\mathrm{AB})}{P_t(\mathrm{BA})} \,. \tag{4}$$

As an example, for 10 trials with 400 ingredients taking $P_t(\mathrm{AB}) = 0.05$ and $P_t(\mathrm{BA}) = 0.04$, we found K= 1.27 with a standard deviation of 0.13, compared to the deterministic value of 1.25. As a further test of the model one can ask whether it is ergodic in the sense that the average of K over time, i. e., for a single system observed for a long time after reaching equilibrium, is equal to the average K for identical systems taken at a particular time in a large number of trials. When this was tested for 1000 time steps (separated by 100 iterations) vs. 1000 trials the results were statistically identical, indicating that the first-order cellular automaton model is sensibly ergodic [2].

The first-order model can also be used to examine sequences of transformations of the form A → B → C . . . . For the simple example A → B → C the concentration of the initial reactant A falls exponentially, that of the intermediate species B rises then falls, and that of C builds up as it is fed from B. These time-dependent changes are illustrated in Fig. 7 using specific transformation probabilities.

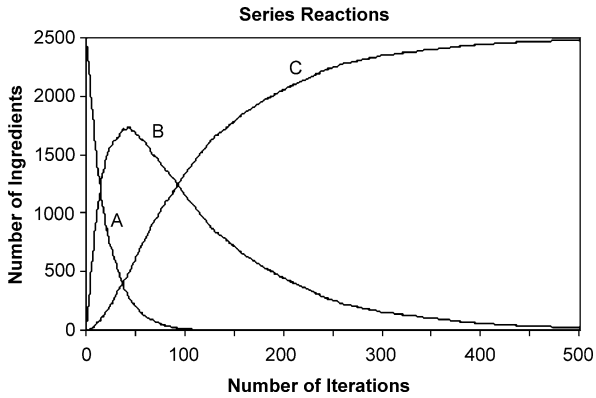

Cellular Automata Modeling of Complex Biochemical Systems, Figure 7
Typical time-dependent variations of the ingredient populations for a two-step reaction A → B → C

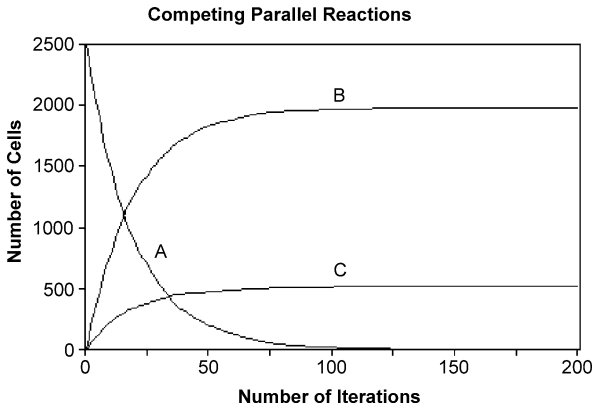

Cellular Automata Modeling of Complex Biochemical Systems, Figure 8
Typical variations in the ingredient populations for competing reaction A → B → C

Kinetic and Thermodynamic Reaction Control Parallel competing reactions

$$\begin{array}{l} A \rightarrow B \\ \downarrow \\ C \end{array}$$

can also be simulated. An especially interesting example occurs when the reactions are reversible [24]:

$$A \leftrightarrow B \quad \text{and} \quad A \leftrightarrow C\,.$$

With properly chosen transformation probabilities, this model can be utilized to examine the conditions governing thermodynamic and kinetic control of reactions. An illustration is shown in Fig. 8 for the conditions $P_t(AB) = 0.01$, $P_t(AC) = 0.001$, $P_t(BA) = 0.02$, and $P_t(CA) = 0.0005$, using a set of 10 000 ingredients. Starting with all species A, in the initial stages of the reaction the kinetically-favored product B is produced in excess, whereas at later times the thermodynamically-favored product C gains dominance. The cellular automata model is in good agreement with those found in a deterministic, numerical solution for the same conditions. For example, the cellular automata model yields final equilibrium concentrations for species B and C of $[B] = 0.1439 \pm 0.0038$ and $[C] = 0.5695 \pm 0.0048$ compared to reported deterministic values of 0.14 and 0.571, respectively.

Excited-State Kinetics Another important application of the first-order model is to the examination of the ground and excited state kinetics of atoms and molecules [25]. One illustration of the excited-state cellular automata model is the dynamics of the excited-state transitions of oxygen atoms [26]. The oxygen atom has a ^3P ground state and ^1S and ^1D excited states. Emissions from the latter two excited states play an important role in the dramatic light displays—the *Aurora Borealis*, or "Northern lights"—seen under certain conditions in the northern polar skies, and similar emissions have been detected in the atmospheres of Mars and Venus. The excited states are believed to be mainly produced by dissociative recombinations of ionized oxygen molecules and electrons generated in the atmosphere by ultraviolet bombardment during the daylight hours:

$$O_2^+ + e^- \rightarrow O^* + O^{**}\,.$$

In this, the species O^* and O^{**} are unspecified atomic oxygen states, which may be any of the species ^3P, ^1S, or ^1D. The most prominent feature in the atmospheric displays is normally the green spin-allowed $^1S \rightarrow {}^1D$ transition appearing at 5577 angstroms.

Using transition probabilities taken from the compilation of Okabe [27] we have simulated the dynamics associated with these atomic transitions under both pulse and steady-state conditions. For the pulse simulations two starting conditions were examined: the first in which all ingredients started in the upper ^1S excited state, and the second in which the ingredients started in a distribution believed characteristic of that produced by the dissociative recombination process shown above. The simulations yield excited state lifetimes and luminescence quantum yields consistent with the experimental observations for these properties.

Second-Order Kinetics Several groups have developed cellular automata models for particular reaction-diffusion

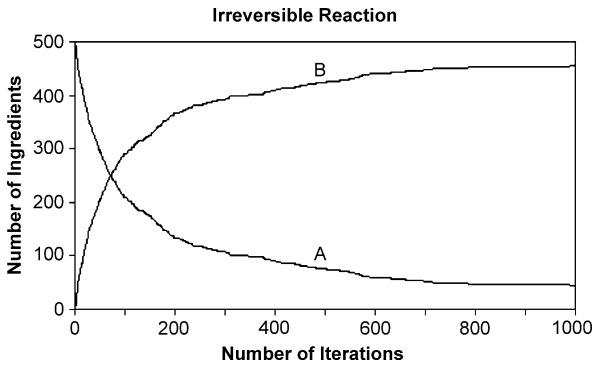

Cellular Automata Modeling of Complex Biochemical Systems, Figure 9
Illustration of the variation with time of the ingredient populations for an irreversible reaction A → B

systems. In particular, the Belousov–Zhabotinsky oscillating reaction has been examined in a number of studies. Attention has also been directed to the simpler A + B → C reaction, using both lattice-gas models and a generalized Margolus diffusion approach. We have recently developed a simple, direct cellular automaton model [28] for hard-sphere bimolecular chemical reactions of the form

$$\mathrm{A} + \mathrm{B} \rightarrow \mathrm{C} + \mathrm{D}\,.$$

As before, the different species are assigned different colors in the visualization. In this model the reactant and product species diffuse about the grid in random walks. When the species A and B encounter each other (come to adjacent cells) on the grid the probability that these species transform to C and D is determined by an assigned reaction probability $P_r(\mathrm{AB})$. The simulations take place on a toroidal space such that ingredients leaving the grid on one side appear at the opposite edge. Initially the ingredients are placed randomly on the grid.

The production of species C over time for starting counts of 100 A and 200 B ingredients on a $100 \times 100 = 10\,000$ cell grid provides an example shown in Fig. 9. The expected second-order rate law

$$\frac{\mathrm{d}[C]}{\mathrm{d}t} = k[A][B] \tag{5}$$

is found to be obeyed in the simulations, subject to fluctuations expected for a system containing a finite number of reacting ingredients. When the results from a number of trials are combined to mimic, in effect, the results for a much larger system, the fluctuations become relatively small and the results approach the deterministic forms.

When a back reaction C + D → A + B with probability $P_r(\mathrm{CD})$ is included in the automaton rules the equilibrium

$$\mathrm{A} + \mathrm{B} \equiv \mathrm{C} + \mathrm{D}$$

can be simulated. Once the system has stabilized from its initial non-equilibrium concentrations, fluctuations about the equilibrium concentrations occur over time, and the relative size of these fluctuations decreases as the number of ingredients increases. A variation of this theme occurs in a pseudo first-order reaction. This type of reaction involves an actual second-order reaction A + B → C + D in which one of the reactants, say B, is present in sufficient excess that its variation becomes effectively unnoticeable. Simulations with 50 A ingredients and 1000 B ingredients bear out the expectation that the reaction kinetics are such that the rate of production of C and D appears to depend only on the concentration of A.

Enzyme Reactions

A recent study on the kinetics of an enzyme reaction considered the Michaelis–Menten model [29]. The rules selected for this reaction included a probability of the conversion, P_c, of an enzyme-substrate pair, ES, to an enzyme-product pair, EP. The Michaelis–Menten model was observed and characteristic Lineweaver–Burk plots were found from the model. The systematic variation of the hydrophobicity of substrates and products showed that a lower affinity between a substrate and water leads to a greater extent of the reaction at a common point along the reaction progress curve. This influence is greater than the influence of the affinity between the substrate and the enzyme. The water-substrate affinity appears to primarily influence the concentration of the ES complex at the observed point along the reaction progress curve. A low affinity between water and substrate favors a high ES concentration at this point. A hydrophilic substrate appears to be more entrapped in the water continuum, hence to be less available to the enzyme. It was also observed that an accumulation of product molecules around the enzymes coincides with a decline in the reaction rate. A biochemical network was reported to be modeled using cellular automata [30]. A protein kinase was modeled and studied using various concentrations of substrates and changes in enzyme competence.

An Anticipatory Model

An anticipatory enzyme system has been modeled using the dynamic characteristics of cellular automata [31]. A concentration of an intermediate product influences the competence of an enzyme down stream. This anticipation of the future event creates a condition in which the con-

centration of a later substrate is suppressed, a property characteristic of the system. The dynamics revealed concentrations over time, influenced by the presence or absence of a feed-forward or pre-adaptation state in the system. The concentration of A steadily diminishes as successive concentrations of B, C and D rise and fall at the same levels. The concentration of E rises at the end of the run, eventually becoming the only ingredient in the system. The concentration of D is approximately 0.25 in a non-anticipatory model. In contrast, with an anticipatory or feed-forward step in the system there is created an additional amount of enzyme specific for substrate D. This enzyme, $e_{2,4}$, is available at a future time to catalyze the conversion of D to E. This creates a property of the system in which the concentration of ingredient, D, is not allowed to accumulate to its normal level. The concentration of D in an anticipatory model is approximately 0.13, about one half of the D concentrations for the non-anticipatory models. The concentration of B therefore serves as a predictor of the concentration of D at a later time.

Micelle Formation

A micelle is a structure formed from the close interaction of hydrophobic fragments of amphiphiles plus the electrostatic encounters with the surrounding water. Typically they often assume a spherical structure with the non-polar fragments in the interior and the polar fragments on the periphery, in aqueous solution. The formation of these structures is a dynamic process which has been modeled using cellular automata. The model of an amphiphile was created by treating each face of a square automaton cell as an independent structure [32]. Each face of this variegated cell can have its own set of $P_B(X)$ and $J(X)$ values. For the micelle study three of the faces were considered as equivalent and were endowed with rules modeling a hydrophobic or non-polar part of the amphiphile. The other face was treated as a polar fragment of the molecule and assigned characteristic rules. The outcome of the dynamics was the creation of structures in which the non-polar fragments were in the interior of an aggregation of cells while the polar fragment lay on the periphery. The interpretation of these organized clusters is that they model a micelle. The dominant influence on the formation of these structures is the extent of non-polar character of the designated three sides of the cell. Of secondary influence is the polarity of the remaining face of the cell. If this is too polar, the micelle formation is retarded. Both of these influences produce models that agree with experiment. Other studies on these variegated cells depicting an amphiphile revealed a temperature effect on the critical micelle concentration (cmc) which was minimal at about $P_B(W) = 0.25$, corresponding to experiment. The onset of the cmc was modeled and shown to be dependent upon a modestly polar fragment of the amphiphile.

Membrane Permeability

An extension of the micelle and diffusion models was a simulation of the diffusion of a solute through a layer of hydrophobic cells simulating a membrane separating two water compartments [33]. A membrane layer five cells wide membrane was positioned on a grid between two water compartments. The membrane cells were endowed with a $P_B(WS)$ rule making them hydrophobic. The membrane cells could only move about within the layer according to their rule response. The two water cell compartments on either side were assigned identical rules but were colored differently in order to monitor their origins after some movement into and through the membrane layer. The dynamics revealed that water molecules from both compartments pass into and through the membrane as expected. To model the behavior of a solute in this environment, a few cells simulating solute molecules were positioned randomly near the lower edge of the membrane surface. These cells were endowed with rules making them hydrophilic. As the dynamics proceeded, it was observed that more water molecules from the upper compartment passed into and through the membrane than water from the lower compartment. Since the solute molecules were hydrophilic, the membrane was relatively impervious to their passage. The behavior of this model is in agreement with experimental observations collectively referred to as an osmotic effect. This model of diffusion was solute concentration dependent.

As the hydrophobicity of the solute increased, it was observed that an increasing number of solute particles passed through the membrane from the lower aqueous compartment. There was no accumulation of solute molecules within the membrane. At a level of hydrophobicity midway on the scale, i. e. about $P_B(WS) = 0.5$, there was a very abrupt change in this behavior. At this critical hydrophobicity the number of solute molecules passing through the membrane dropped sharply. At higher $P_B(WS)$ values the number of cells passing through the membrane fell to nearly zero. At this critical hydrophobicity, the accumulation of solute molecules in the membrane increased sharply.

Chromatographic Separation

Models of chromatographic separation have been reported, derived from cellular automata [34]. The solvent

cells were randomly distributed over the grid at the initiation of each run. The stationary phase, designated B is simulated by the presence of cells randomly distributed over the grid, replacing W cells. These B cells are immobile and are positioned at least 3 cells from another B cell. The solute cells, usually simulating two different compounds are represented by a few cells each. The movements of the solutes and mobile phase in the grid were governed by rules denoting the joining and breaking of like or unlike cells. The position of each solute cell was recorded at a row in the column after a certain number of iterations. The position of the peak maximum was determined by summing the number of cells found in groups of ten rows on the column then plotting these averages against the iteration time.

The gravity parameter for each ingredient in the simulation defines the flow rate. The polarity of the solvent, W, is encoded in the relative self-association experienced. This is governed by the rules, P_B(WW) and J(WW). The migration of the solutes was found to be faster when the solvent was non-polar. Another study modeled the influence of the relative affinities of solutes for the stationary phase, B. This affinity is encoded in the parameters, P_B(SB) and J(SB). High values of P_B(SB) and low values of J(SB) denote a weak affinity. A single solute was used in this study with five different sets of parameters. These studies revealed that solutes with a greater affinity for the stationary phase migrated at a slower rate. These parameters characterize the structural differences among solutes that give rise to different migratory rates and separations in chromatography.

Modeling Biochemical Networks

The Network Dynamic evolutionary networks have recently been recognized as a universal approach to complex systems including biochemical systems. Network topology is generally used in characterizing these, focusing on their connectivity, neighborhood and distance relationships. Network complexity has also been recently quantitatively characterized [35]. The large size of the metabolic, protein, and gene regulatory networks makes impractical many of the traditional methods for dynamic modeling. A study on the MAPK signaling cascade has recently introduced the potential of cellular automata as a basic method for the dynamic modeling of networks for biological and medical applications [30].

The CA Modeling Design Each molecule involved in the MAPK pathway, Fig. 10, was represented by a number of cells in the CA grid. The numbers chosen reflect the relative concentration of that protein. Each of the cells representing all other proteins move about freely in the grid. They may encounter each other but this has no consequence. The only encounters that have a consequence are those between a specific protein (substrate) and a specific enzyme, as shown in the network. When such an encounter occurs, there is modeled a complex (enzyme-substrate). This complex has an assigned probability of converging to a new complex (enzyme-product). Following this there is a probability assigned for the separation of these two species.

The studies of the MAPK cascade were performed using a CA grid of 100 by 100 cells. Each model was obtained as the average of 50 runs, each of which included 5000 iterations. A network to be studied was represented by groups of CA cells, each group representing one of the network species. The number of cells in each group reflects the relative concentrations of each network ingredient. We have systematically altered the initial concentrations of several proteins (MAPKKK, MAPKK, and MAPK) and the competencies of several enzymes (MAPKK- and MAPK-proteases, and the hypothetical enzymes E1 and E2 that affect the forward and reverse reactions of activation and deactivation of MAPKKK). The basic variable was the concentration of MAPKKK, which was varied within a 25-fold range from 20 to 500 cells. The concentrations of MAPKK and MAPK were kept constant (500 or 250 cells) in most of the models. The four enzymes, denoted by E1, E2, E3, and E4, were represented in the CA grid by 50 cells each. In one series of models, we kept the transition probabilities of three of the enzymes the same, ($P = 0.1$), and varied the probability of the fourth enzyme within the 0 to 1 range. In another series, *all* enzyme probabilities were kept constant, whereas the concentrations of substrates were varied. The last series varied both substrate concentrations and enzyme propensities. Recorded were the variations in the concentrations of the three substrates MAPKKK, MAPKK, and MAPK, and the products MAPKKK*, MAPKK-P, MAPKK-PP, MAPK-P, and MAPK-PP.

Modeling Enzymes Network Activity Upgrading or downgrading enzymes activity is one of the typical ways the cell reacts to stress and interactions with pathogens. A systematic study was made of the variations of one of the four enzymes E1 to E4 at constant concentrations of the substrates MAPKKK, MAPKK and MAPK, and constant propensity of the other three enzymes. This is illustrated in Fig. 11 with the variation of the MAPKK-protease (E3 enzyme in Fig. 10), which reverses the two-step reaction of MAPKK phosphorilation. It is shown that the concentration of the MAPKK- and MAPK-diphos-

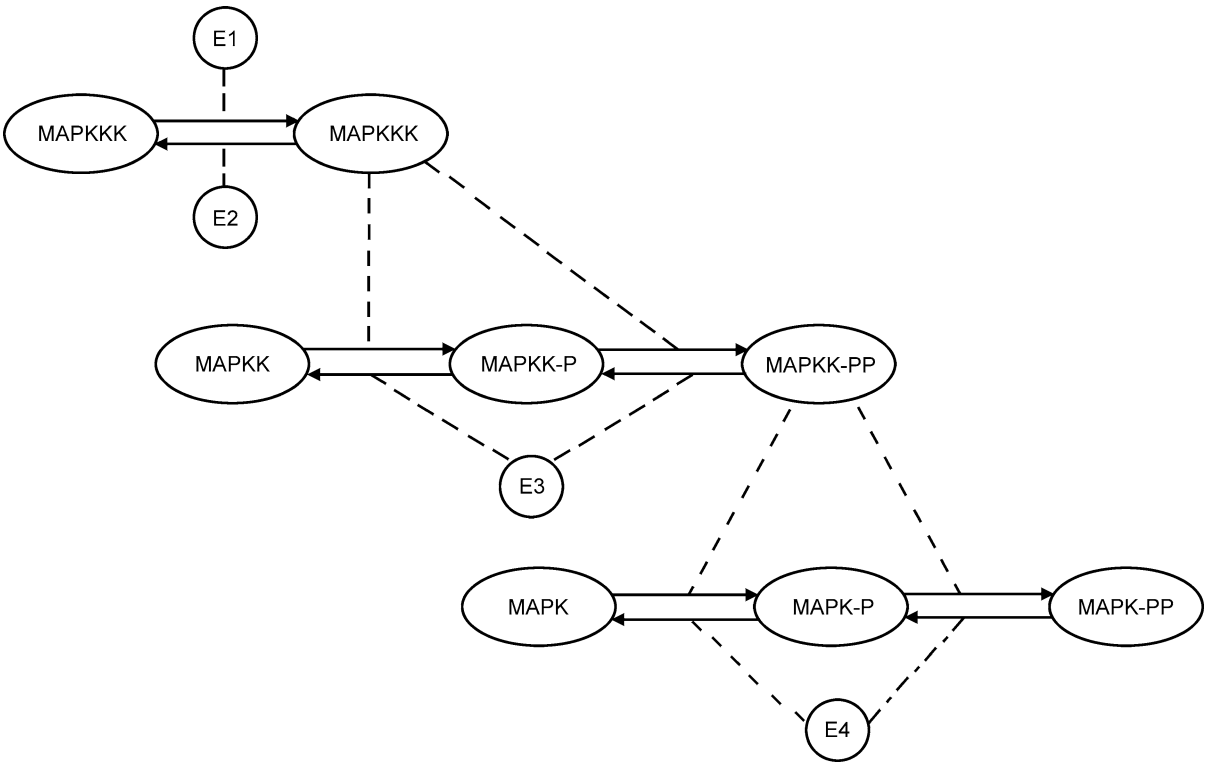

Cellular Automata Modeling of Complex Biochemical Systems, Figure 10
The MAPK signaling cascade. The substrates and products are represented by *oval contours*, reactions by *arrows*, and the catalysts action by *dashed lines*. E3 and E4 are MAPKK-protease and MAPK-protease, respectively. P stands for phosphate, PP for diphosphate

phates (marked as E and H respectively) passes through a maximum near relatively low enzyme transition probability ($P \approx 0.02$). At the point of its maximum, the concentration of MAPK-PP reaches over 80% of its maximum, whereas that of MAPKK-PP is slightly over 50%. This shows the potential for a strong influence on the concentrations of the two diphosphates in the MAPK cascade by inhibiting the MAPKK-protease. In contrast, the level of steady-state concentrations of the two monophosphates (marked by D and G in Fig. 11) is not sensitive to the activity of the enzyme modeled, except for the extreme case of very strong inhibition ($P \rightarrow 0.001$).

Future Directions

Applications of cellular automata as a modeling paradigm in chemistry is in its infancy. Literature searches show little attention to the use of this method to study chemical phenomena. Chemistry is about dynamic systems, thus a rich harvest of information is possible using this paradigm. This potential is illustrated here where primarily solution systems are modeled. A broader use is clearly on the horizon.

The study of bulk water structure is one example of this potential. It points to a way of encoding the structure of an evanescent substance with some degree of validity. A recent example of this potential is a series of studies of the influence of protein surface amino acid side chains on the structure of nearby water [36]. Because of the varying hydropathic states of the surface side chains, a series of passageways (chreodes) was postulated to exist through the water. These chreodes were invoked in the facilitated 2-D diffusion of ligands and substrates to receptors and enzyme active sites. The interference of these chreodes was proposed in a theory of the action of general anesthetics. As an extension of this theory, the induction of sleep was proposed to arise from the interference of these chreodes by elemental nitrogen accumulated through respiration over several hours. The remarkable correspondence of kinetic and enzyme CA models with reality is excellent and foretells the use of these in many applications. The treatment of networks, foretells of another area of great

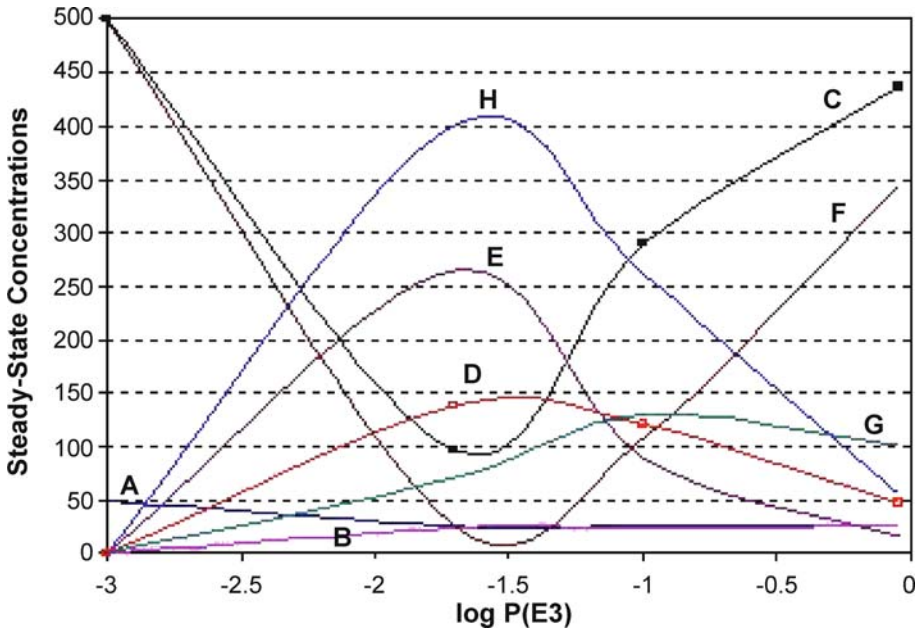

Cellular Automata Modeling of Complex Biochemical Systems, Figure 11
Influence of the MAPKK-protease propensity *P*(E3) on the steady-state concentrations of the MAPK cascade species (A = MAPKKK, B = MAPKKK*, C = MAPKK, D = MAPKK-P, E = MAPKK-PP, F = MAPK, F = MAPK-P, H = MAPK-PP). Enzyme propensities *P*(E1) = *P*(E2) = *P*(E4) = 0.1, substrate initial concentrations $[A_o] = 50$, $[C_o] = [F_o] = 500$

possibilities of models with CA. The beginning of this approach is shown above with the MAPK mode. There is success here that is not achievable with classical differential equations.

A noteable advantage and a golden opportunity exists for any CA chemical model. That is the didactic value of any CA model. You can see the dissolution of a crystal, the changes of concentration in a CA enzyme model, the changes in structure as the temperature changes, and many more. A student will never forget a CA model of these while an equation is quietly forgotten.

New aspects of CA models are on the horizon. Multidimensional models, nested hierarchies, multi-grid assemblies, all are being explored and will soon surface as viable methods rich in information.

Bibliography

Primary Literature

1. Kier LB, Seybold PG, Cheng C-K (2005) Modeling chemical systems using cellular automata. Springer, Dordrecht
2. Von Neumann J (1966) Burks A (ed) Theory of self-reproducing automata. University of Illinois Press, Champaign
3. Ulam SM (1952) Proc Int Congr Math 2:264, held in 1950
4. Ulam SM (1976) Adventures of a mathematician. Charles Scribner's Sons, New York
5. Tofolli T, Margolas N (1987) Cellular automata machines: A new environment for modeling. MIT Press, Cambridge
6. Wolfram S (2002) A new kind of science. Wolfram Media, Champaign
7. Kapral R, Showalter K (1995) Chemical waves and patterns. Kluwer, Boston
8. Ermentrout GB, Edelstein-Keshet L (1993) J Theoret Biol 160: 97–133
9. Grimm V, Revilla E, Berger U et al (2005) Science 310:987–991
10. Wu-Pong S, Cheng C-K (1999) Pharmacokinetic simulations using cellular automata in a pharmacokinetics course. Am J Pharm Educ 63:52–55
11. Moreira N (2006) In pixels and in health. Sci News Jan 21:40–44
12. Batty M (2005) Cities and complexity: Understanding cities with cellular automata, agent-based models, and fractals. MIT Press, Cambridge
13. Kohler TA, Gumerman GJ (2000) Dynamics in human and primate societies: Agent-based modelling of social and spatial processes. Oxford University Press, New York
14. White R (2005) Modelling multi-scale processes in a cellular automata framework. In: Portugali J (ed) Complex artificial environments. Springer, New York, pp 165–178
15. Kier LB, Cheng C-K (1994) A cellular automata model of water. J Chem Inf Comp Sci 34:647–654
16. Kier LB, Cheng C-K (1994) A cellular automata model of an aqueous solution. J Chem Inf Comp Sci 34:1334–1341
17. Kier LB, Cheng C-K, Testa B (1995) A cellular automata model of the hydrophobic effect. Pharm Res 12:615–622
18. Kier LB, Cheng C-K (1995) A cellular automata model of dissolution. Pharm Res 12:1521–1528

19. Kier LB, Cheng C-K, Testa B (1997) A cellular automata model of diffusion in aqueous systems. J Pharm Sci 86:774–781
20. Kier LB, Cheng C-K, Seybold PG (2001) A cellular automata model of aqueous systems. Revs Comput Chem 17:205–238
21. Cheng C-K, Kier LB (1995) A cellular automata model of oil-water partitioning. J Chem Inf Comput Sci 35:1054–1061
22. Kier LB, Cheng C-K, Tute M, Seybold PG (1998) A cellular automata model of acid dissociation. J Chem Inf Comp Sci 38:271–278
23. Seybold PG, Kier LB, Cheng C-K (1997) J Chem Inf Comput Sci 37:386–391
24. Neuforth A, Seybold PG, Kier LB, Cheng C-K (2000) Cellular automata models of kinetically and thermodynamically controlled reactions, vol A. Int J Chem Kinetic 32:529–534
25. Seybold PG, Kier LB, Cheng C-K (1998) Stochastic cellular automata models of molecular excited state dynamics. J Phys Chem A 102:886–891
26. Seybold PG, Kier LB, Cheng C-K (1999) Aurora Borealis: Stochastic cellular automata simulation of the excited state dynamics of Oxygen atoms. Int J Quantum Chem 75:751–756
27. Okabe H (1978) Photochemistry of small molecules. Wiley, New York, p 370
28. Moore J, Seybold PG; To be published personal correspondence
29. Kier LB, Cheng C-K, Testa B (1996) A cellular automata model of enzyme kinetics. J Molec Graph 14:227–234
30. Kier LB, Bonchev D, Buck G (2005) Modeling biochemical networks: A cellular automata approach. Chem Biodiv 2:233–43
31. Kier LB, Cheng C-K (2000) A cellular automata model of an anticipatory system. J Molec Graph Model 18:29–35
32. Kier LB, Cheng C-K, Testa B (1996) Cellular automata model of micelle formation. Pharm Res 13:1419–1426
33. Kier LB Cheng C-K (1997) A cellular automata model of membrane permeability. J Theor Biol 186:75–85
34. Kier LB, Cheng C-K, Karnes HT (2000) A cellular automata model of chromatography. Biomed Chrom 14:530–539
35. Bonchev D (2003) Complexity of protein–protein interaction networks, complexes and pathways. In: Conn M (ed) Handbook of proteomics methods. Humana, New York, pp 451–462
36. Kier LB (2007) Water as a complex system: Its role in ligand diffusion, general anesthesia, and sleep. Chem Biodiv 4:2473–2479

Selected Bibliography of Works on Cellular Automata

At this time thousands of scientific articles have been published describing cellular automata studies of topics ranging from applications dealing with physical and biological systems to investigations of traffic control and topics in the social sciences. It would be impossible to describe all of these studies within a limited space, but it may be useful to provide a short list of representative investigations on a limited variety of topics, permitting starting points for readers who might wish to further examine applications in these more narrow subjects. Below we give a short selection of publications, some of which, although not explicitly referring to CA, cover the same approach or a related approach.

Artificial Life

Adami C (1998) An introduction to artificial life. Springer, New York
Langton CG, Farmer JD, Rasmussen S, Taylor C (1992) Artificial life, vol II. Addison-Wesley, Reading

Biological Applications (General)

Maini PK, Deutsch A, Dormann S (2003) Cellular automaton modeling of biological pattern formation. Birkhäuser, Boston
Sigmund K (1993) Games of life: Explorations in ecology, evolution, and behaviour. Oxford University Press, New York
Solé R, Goodman B (2000) Signs of life: How complexity pervades biology. Basic Books, New York; A tour-de-force general introduction to biological complexity, with many examples

Books

Chopard B, Droz M (1998) Cellular automata modeling of physical systems. Cambridge University Press, Cambridge
Gaylord RJ, Nishidate K (1996) Modeling nature: Cellular automata simulations with Mathematica®. Telos, Santa Clara
Griffeath D, Moore C (2003) New constructions in cellular automata. In: Santa Fe Institute Studies in the Sciences of Complexity Proceedings. Oxford University Press, New York
Ilachinski A (2001) Cellular automata: A discrete universe. World Scientific, Singapore
Kier LB, Seybold PG, Cheng C-K (2005) Modeling chemical systems using cellular automata. Springer, Dodrecht
Manneville P, Boccara N, Vishniac GY, Bidaux R (1990) Cellular automata and modeling of complex physical systems. Springer, New York, pp 57–70
Schroeder M (1991) Fractals, chaos, power laws. WH Freeman, New York
Toffoli T, Margolus N (1987) Cellular automata machines: A new environment for modeling. MIT Press, Cambridge
Wolfram S (1994) Cellular automata and complexity: Collected papers. Westview Press, Boulder
Wolfram S (2002) A new kind of science. Wolfram Media, Champaign

Emergent Properties

Gruner D, Kapral R, Lawniczak AT (1993) Nucleation, domain growth, and fluctuations in a bistable chemical system. J Chem Phys 96:2762–2776
Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Nat Acad Sci USA 79:2554–2558
Kauffman S (1984) Emergent properties in random complex automata. Physica D 10:145–156
Ottino JM (2004) Engineering complex systems. Nature 427:399

Evolution

Farmer JD, Kauffman SA, Packard NH (1986) Autocatalytic replication of polymers. Physica D 22:50–67
Solé RV, Bascompté J, Manrubia SC (1996) Extinctions: Bad genes or weak chaos? Proc Roy Soc Lond B 263:1407–1413
Solé RV, Manrubia SC (1997) Criticality and unpredictability in macroevolution. Phys Rev E 55:4500–4508
Solé RV, Manrubia SC, Benton M, Bak P (1997) Self-similarity of extinction statistics in the fossil record. Nature 388:764–767

Solé RV, Montoya JM, Erwin DH (2002) Recovery from mass extinction: Evolutionary assembly in large-scale biosphere dynamics. Phil Trans Royal Soc 357:697–707

Excited State Phenomena

Seybold PG, Kier LB, Cheng C-K (1998) Stochastic cellular automata models of molecular excited-state dynamics. J Phys Chem A 102:886–891; Describes general cellular automata models of molecular excited states

Seybold PG, Kier LB, Cheng C-K (1999) Aurora Borealis: Stochastic cellular automata simulations of the excited-state dynamics of Oxygen atoms. Int J Quantum Chem 75:751–756; This paper examines the emissions and excited-state transitions of atomic Oxygen responsible for some of the displays of the Aurora Borealis

First-Order Chemical Kinetics

Hollingsworth CA, Seybold PG, Kier LB, Cheng C-K (2004) First-order stochastic cellular automata simulations of the Lindemann mechanism. Int J Chem Kinetic 36:230–237

Neuforth A, Seybold PG, Kier LB, Cheng C-K (2000) Cellular automata models of kinetically and thermodynamically controlled reactions. Int J Chem Kinetic 32:529–534; A study of kinetic and thermodynamic reaction control

Seybold PG, Kier LB, Cheng C-K (1997) Simulation of first-order chemical kinetics using cellular automata. J Chem Inf Comput Sci 37:386–391; This paper illustrates a number of first-order cellular automata models

Fluid Flow

Malevanets A, Kapral R (1998) Continuous-velocity lattice-gas model for fluid flow. Europhys Lett 44:552

Game of Life

Alpert M (1999) Not just fun and games. Sci Am 40:42; A profile of John Horton Conway

Gardner M (1970) The fantastic combinations of John Conway's new solitaire game "life". Sci Amer 223:120–123

Gardner M (1971) On cellular automata, self-reproduction, the Garden of Eden and the game of "life". Sci Amer 224:112–117"

Note: There are many examples on the web of applets that allow you to play the Game of Life. Since these come and go, you are urged to locate them using a search engine.

Geology

Barton CC, La Pointe PR (1995) Fractals in petroleum geology and earth processes. Plenum, New York

Turcotte DL (1997) Fractals and chaos in geology and geophysics, 2nd edn. Cambridge University Press, New York

Historical Notes

Ulam SM (1952) Random processes and transformations. Proc Int Congr Math 2:264, held in 1950

Ulam SM (1976) Adventures of a mathematician. Charles Scribner's Sons, New York

Von Neumann J (1966) Burks A (ed) Theory of self-replicating automata. University of Illinois Press, Urbana

Zuse K (1982) The computing universe. Int J Theoret Phys 21:589

Liquid Phase Interactions

Cheng C-K, Kier LB (1995) A cellular automata model of oil-water partitioning. J Chem Info Comput Sci 35:1054–1059

Kier LB, Cheng C-K, Testa B (1996) A cellular automata model of micelle formation. Pharm Res 13:1419–1422

Malevanets A, Kapral R (1999) Mesoscopic model for solvent dynamics. J Chem Phys 110:8605–8613

Oscillations

Chavez F, Kapral R (2002) Oscillatory and chaotic dynamics in compartmentalized geometries. Phys Rev E 65:056203

Chavez F, Kapral R, Rousseau G, Glass L (2001) Scroll waves in spherical shell geometries. Chaos 11:757

Goryachev A, Strizhak P, Kapral R (1997) Slow manifold structure and the emergence of mixed-mode oscillations. J Chem Phys 107:2881

Hemming C, Kapral R (2002) Phase front dynamics in inhomogeneously forced oscillatory systems. Physica A 306:199

Pattern Formation

Kapral R, Showalter K (1994) Chemical waves and patterns. Kluwer, Dordrecht

Parrish JK, Edelstein-Keshet L (1999) Complexity, pattern, and evolutionary trade-offs in animal aggregation. Science 284:99–101

Veroney JP, Lawniczak AT, Kapral R (1996) Pattern formation in heterogeneous media. Physica D 99:303–317

Physics Applications

Signorini J (1990) Complex computing with cellular automata. In: Manneville P, Boccara N, Vishniac GY, Bidaux R (eds) Cellular automata and modeling of complex physical systems. Springer, New York, pp 57–70

Toffoli T (1984) Cellular automata as an alternative (rather than an approximation of) differential equations in modeling physics. Physica D 10:117–127

Vichniac GY (1984) Simulating physics with cellular automata. Physica D 10:96–116

Population Biology and Ecology

Bascompté J, Solé RV (1994) Spatially-induced bifurcations in single species population dynamics. J Anim Ecol 63:256–264

Bascompté J, Solé RV (1995) Rethinking complexity: Modelling spatiotemporal dynamics in ecology. Trends Ecol Evol 10:361–366

Bascompté J, Solé RV (1996) Habitat fragmentation, extinction thresholds in spatialy explicit models. J Anim Ecol 65:465

Deutsch A, Lawniczak AT (1999) Probabilistic lattice models of collective motion, aggregation: From individual to collective dynamics. J Math Biosci 156:255–269

Fuks H, Lawniczak AT (2001) Individual-based lattice models for the spatial spread of epidemics. Discret Dyn Nat Soc 6(3):1–18

Gamarra JGP, Solé RV (2000) Bifurcations, chaos in ecology: Lynx returns revisited. Ecol Lett 3:114–121

Levin SA, Grenfell B, Hastings A, Perelson AS (1997) Mathematical, computational challenges in population biology, ecosystems science. Science 275:334–343
Montoya JM, Solé RV (2002) Small world patterns in food webs. J Theor Biol 214:405–412
Nowak MA, Sigmund K (2004) Population dynamics in evolutionary ecology. In: Keinan E, Schechter I, Sela M (eds) Life sciences for the 21st century. Wiley-VCH, Cambridge, pp 327–334
Solé RV, Alonso D, McKane A (2000) Connectivity, scaling in S-species model ecosystems. Physica A 286:337–344
Solé RV, Manrubia SC, Kauffman S, Benton M, Bak P (1999) Criticality, scaling in evolutionary ecology. Trends Ecol Evol 14:156–160
Solé RV, Montoya JM (2001) Complexity, fragility in ecological networks. Proc Roy Soc 268:2039–2045

Random Walks

Berg HC (1983) Random walks in biology. Princeton University Press, Princeton
Hayes B (1988) How to avoid yourself. Am Sci 86:314–319
Lavenda BH (1985) Brownian motion. Sci Am 252(2):70–85
Shlesinger MF, Klafter J (1989) Random walks in liquids. J Phys Chem 93:7023–7026
Slade G (1996) Random walks. Am Sci 84:146–153
Weiss GH (1983) Random walks, their applications. Am Sci 71:65–71

Reviews

Kapral R, Fraser SJ (2001) Chaos, complexity in chemical systems. In: Moore JH, Spencer ND (eds) Encyclopedia of chemical physics, physical chemistry, vol III. Institute of Physics Publishing, Philadelphia, p 2737
Kier LB, Cheng C-K, Testa (1999) Cellular automata models of biochemical phenomena. Future Generation Compute Sci 16:273–289
Kier LB, Cheng C-K, Seybold PG (2000) Cellular automata models of chemical systems. SAR QSAR Environ Res 11:79–102
Kier LB, Cheng C-K, Seybold PG (2001) Cellular automata models of aqueous solution systems. In: Lipkowitz KM, Boyd DB (eds) Reviews in computational chemistry, vol 17. Wiley-VCH, New York, pp 205–225
Turcotte DL (1999) Self-organized criticality. Rep Prog Phys 62:1377–1429
Wolfram S (1983) Cellular automata. Los Alamos Sci 9:2–21

Second-Order Chemical Kinetics

Boon JP, Dab D, Kapral R, Lawniczak AT (1996) Lattice-gas automata for reactive systems. Phys Rep 273:55–148
Chen S, Dawson SP, Doolen G, Jenecky D, Lawiczak AT (1995) Lattice methods for chemically reacting systems. Comput Chem Engr 19:617–646

Self-Organized Criticality

Bak P (1996) How nature works. Springer, New York
Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality: An explanation for 1/f noise. Phys Rev Lett 59:381–384; A classic paper introducing the "sandpile" cellular automaton
Turcotte DL (1999) Self-organized criticality. Rep Prog Phys 62:1377–1429

Social Insect Behavior

Cole BJ (1991) Short-term activity cycles in ants: Generation of periodicity by worker inaction. Amer Nat 137:144–259
Cole BJ (1996) Mobile cellular automata models of ant behavior: Movement activity of Leptothorax Allardycei. Amer Nat 148:1–15
Deneubourg J-L, Goss S, Franks NR, Pasteels JM (1989) The blind leading the blind: Modeling chemically mediated Army ant raid patterns. J Insect Behav 2:719–772
Goss S, Deneubourg J-L (1988) Autocatalysis as a source of synchronized rhythmical activity in social insects. Insectes Sociaux 35:310–315
Solé RV, Miramontes O, Goodwin BC (1993) Oscillations, chaos in ant societies. J Theoret Biol (1993) 161:343–357

Social Sciences

Gaylord RJ (1998) D'Andria LJ (1998) Simulating society: A Mathematica toolkit for modeling socioeconomic behavior. Springer/Telos, New York
Mandelbrot BB (1982) The fractal geometry of nature. Freeman, San Francisco

Traffic Rules

Huang P-H, Kong L-J, Liu M-R (2002) A study of a main-road cellular automata traffic flow model. Chin Phys 11:678–683
Nagel K, Wolf DE, Wagner P, Simon P (1998) Two-lane rules for cellular automata: A systematic approach. Phys Rev E 58:1425–1437

Water

Kier LB, Cheng C-K (1994) A cellular automata model of water. J Chem Info Comput Sci 34:647

Cellular Automata Modeling of Physical Systems

Bastien Chopard
Computer Science Department, University of Geneva, Geneva, Switzerland

Article Outline

Glossary
Definition of the Subject
Introduction
Definition of a Cellular Automata
Limitations, Advantages, Drawbacks, and Extensions
Applications
Future Directions
Bibliography

Cellular automata offer a powerful modeling framework to describe and study physical systems composed of interacting components. The potential of this approach is demonstrated in the case of applications taken from various fields of physics, such as reaction-diffusion systems, pattern formation phenomena, fluid flows and road traffic models.

Glossary

BGK models Lattice Boltzmann models where the collision term Ω is expressed as a deviation from a given local equilibrium distribution $f^{(0)}$, namely $\Omega = (f^{(0)} - f)/\tau$, where f is the unknown particle distribution and τ a relaxation time (which is a parameter of the model). BGK stands for Bhatnager, Gross and Krook who first considered such a collision term, but not specifically in the context of lattice systems.

CA Abbreviation for cellular automata or cellular automaton.

Cell The elementary spatial component of a CA. The cell is characterized by a state whose value evolves in time according to the CA rule.

Cellular automaton System composed of adjacent cells or sites (usually organized as a regular lattice) which evolves in discrete time steps. Each cell is characterized by an internal state whose value belongs to a finite set. The updating of these states is made in parallel according to a local rule involving only a neighborhood of each cell.

Conservation law A property of a physical system in which some quantity (such as mass, momentum or energy) is locally conserved during the time evolution. These conservation laws should be included in the microdynamics of a CA model because they are essential ingredients governing the macroscopic behavior of any physical system.

Collision The process by which the particles of a LGA change their direction of motion.

Continuity equation An equation of the form $\partial_t \rho + \operatorname{div} \rho \mathbf{u} = 0$ expressing the mass (or particle number) conservation law. The quantity ρ is the local density of particles and $\mathbf{u}$ the local velocity field.

Critical phenomena The phenomena which occur in the vicinity of a continuous phase transition, and are characterized by very long correlation length.

Diffusion A physical process described by the equation $\partial_t \rho = D\nabla^2 \rho$, where ρ is the density of a diffusing substance. Microscopically, diffusion can be viewed as a random motion of particles.

DLA Abbreviation of Diffusion Limited Aggregation. Model of a physical growth process in which diffusing particles stick on an existing cluster when they hit it. Initially, the cluster is reduced to a single seed particle and grows as more and more particles arrive. A DLA cluster is a fractal object whose dimension is typically 1.72 if the experiment is conducted in a two-dimensional space.

Dynamical system A system of equations (differential equations or discretized equations) modeling the dynamical behavior of a physical system.

Equilibrium states States characterizing a closed system or a system in thermal equilibrium with a heat bath.

Ergodicity Property of a system or process for which the time-averages of the observables converge, in a probabilistic sense, to their ensemble averages.

Exclusion principle A restriction which is imposed on LGA or CA models to limit the number of particles per site and/or lattice directions. This ensures that the dynamics can be described with a cellular automata rule with a given maximum number of bits. The consequence of this exclusion principle is that the equilibrium distribution of the particle numbers follows a Fermi–Dirac-like distribution in LGA dynamics.

FHP model Abbreviation for the Frisch, Hasslacher and Pomeau lattice gas model which was the first serious candidate to simulate two-dimensional hydrodynamics on a hexagonal lattice.

Fractal Mathematical object usually having a geometrical representation and whose spatial dimension is not an integer. The relation between the size of the object and its "mass" does not obey that of usual geometrical objects. A DLA cluster is an example of a fractal.

Front The region where some physical process occurs. Usually the front includes the locations in space that are first affected by the phenomena. For instance, in a reaction process between two spatially separated reactants, the front describes the region where the reaction takes place.

HPP model Abbreviation for the Hardy, de Pazzis and Pomeau model. The first two-dimensional LGA aimed at modeling the behavior of particles colliding on a square lattice with mass and momentum conservation. The HPP model has several physical drawbacks that have been overcome with the FHP model.

Invariant A quantity which is conserved during the evolution of a dynamical system. Some invariants are imposed by the physical laws (mass, momentum, energy) and others result from the model used to describe physical situations (spurious, staggered invariants). Collisional invariants are constant vectors in the space where the Chapman–Enskog expansion is per-

formed, associated to each quantity conserved by the collision term.

Ising model Hamiltonian model describing the ferromagnetic paramagnetic transition. Each local classical spin variables $s_i = \pm 1$ interacts with its neighbors.

Isotropy The property of continuous systems to be invariant under any rotations of the spatial coordinate system. Physical quantities defined on a lattice and obtained by an averaging procedure may or may not be isotropic, in the continuous limit. It depends on the type of lattice and the nature of the quantity. Second-order tensors are isotropic on a 2D square lattice but fourth-order tensors need a hexagonal lattice.

Lattice The set of cells (or sites) making up the spatial area covered by a CA.

Lattice Boltzmann model A physical model defined on a lattice where the variables associated to each site represent an average number of particles or the probability of the presence of a particle with a given velocity. Lattice Boltzmann models can be derived from cellular automata dynamics by an averaging and factorization procedure, or be defined per se, independently of a specific realization.

Lattice gas A system defined on a lattice where particles are present and follow given dynamics. Lattice gas automata (LGA) are a particular class of such a system where the dynamics are performed in parallel over all the sites and can be decomposed in two stages: (i) propagation: the particles jump to a nearest-neighbor site, according to their direction of motion and (ii) collision: the particles entering the same site at the same iteration interact so as to produce a new particle distribution. HPP and FHP are well-known LGA.

Lattice spacing The separation between two adjacent sites of a regular lattice. Throughout this book, it is denoted by the symbol Δ_{r}.

LB Abbreviation for Lattice Boltzmann.

LGA Abbreviation for Lattice Gas Automaton. See lattice gas model for a definition.

Local equilibrium Situation in which a large system can be decomposed into subsystems, very small on a macroscopic scale but large on a microscopic scale such that each sub-system can be assumed to be in thermal equilibrium. The local equilibrium distribution is the function which makes the collision term of a Boltzmann equation vanish.

Lookup table A table in which all possible outcomes of a cellular automata rule are pre-computed. The use of a lookup table yields a fast implementation of a cellular automata dynamics since however complicated a rule is, the evolution of any configuration of a site and its neighbors is directly obtained through a memory access. The size of a lookup table grows exponentially with the number of bits involved in the rule.

Margolus neighborhood A neighborhood made of two-by-two blocks of cells, typically in a two-dimensional square lattice. Each cell is updated according to the values of the other cells in the same block. A different rule may possibly be assigned dependent on whether the cell is at the upper left, upper right, lower left or lower right location. After each iteration, the lattice partition defining the Margolus blocs is shifted one cell right and one cell down so that at every other step, information can be exchanged across the lattice. Can be generalized to higher dimensions.

Microdynamics The Boolean equation governing the time evolution of a LGA model or a cellular automata system.

Moore neighborhood A neighborhood composed of the central cell and all eight nearest and next-nearest neighbors in a two-dimensional square lattice. Can be generalized to higher dimensions.

Multiparticle models Discrete dynamics modeling a physical system in which an arbitrary number of particles is allowed at each site. This is an extension of an LGA where no exclusion principle is imposed.

Navier–Stokes equation The equation describing the velocity field $\mathbf{u}$ in a fluid flow. For an incompressible fluid ($\partial_t \rho = 0$), it reads

$$\partial_t \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} = -\frac{1}{\rho}\nabla P + \nu \nabla^2 \mathbf{u}$$

where ρ is the density and P the pressure. The Navier–Stokes equation expresses the local momentum conservation in the fluid and, as opposed to the Euler equation, includes the dissipative effects with a viscosity term $\nu\nabla^2\mathbf{u}$. Together with the continuity equation, this is the fundamental equation of fluid dynamics.

Neighborhood The set of all cells necessary to compute a cellular automaton rule. A neighborhood is usually composed of several adjacent cells organized in a simple geometrical structure. Moore, von Neumann and Margolus neighborhoods are typical examples.

Occupation numbers Boolean quantities indicating the presence or absence of a particle in a given physical state.

Open system A system communicating with the environment by exchange of energy or matter.

Parallel Refers to an action which is performed simultaneously at several places. A parallel updating rule corresponds to the updating of all cells at the same time as

if resulting from the computations of several independent processors.

Partitioning A technique consisting of dividing space in adjacent domains (through a partition) so that the evolution of each block is uniquely determined by the states of the elements within the block.

Phase transition Change of state obtained when varying a control parameter such as the one occurring in the boiling or freezing of a liquid, or in the change between ferromagnetic and paramagnetic states of a magnetic solid.

Propagation This is the process by which the particles of a LGA are moved to a nearest neighbor, according to the direction of their velocity vector $\mathbf{v_i}$. In one time step Δ_t the particle travel from cell $\mathbf{r}$ to cell $\mathbf{r} + \mathbf{v_i}\Delta_t$ where $\mathbf{r} + \mathbf{v_i}\Delta_t$ is the nearest neighbor in lattice direction i.

Random walk A series of uncorrelated steps of length unity describing a random path with zero average displacement but characteristic size proportional to the square root of the number of steps.

Reaction-diffusion systems Systems made of one or several species of particles which diffuse and react among themselves to produce some new species.

Scaling hypothesis A hypothesis concerning the analytical properties of the thermodynamic potentials and the correlation functions in a problem invariant under a change of scale.

Scaling law Relations among the critical exponents describing the power law behaviors of physical quantities in systems invariant under a change of scale.

Self-organized criticality Concept aimed at describing a class of dynamical systems which naturally drive themselves to a state where interesting physics occurs at all scales.

Site Same as a cell, but preferred terminology in LGA and LB models.

Spatially extended systems Physical systems involving many spatial degrees of freedom and which, usually, have rich dynamics and show complex behaviors. Coupled map lattices and cellular automata provides a way to model spatially extended systems.

Spin Internal degree of freedom associated to particles in order to describe their magnetic state. A widely used case is the one of classical Ising spins. To each particle, one associates an "arrow" which is allowed to take only two different orientations, up or down.

Time step Interval of time separating two consecutive iterations in the evolution of a discrete time process, like a CA or a LB model. Throughout this work the time step is denoted by the symbol Δ_t.

Universality The phenomenon whereby many microscopically different systems exhibit a critical behavior with quantitatively identical properties such as the critical exponents.

Updating operation consisting of assigning a new value to a set of variables, for instance those describing the states of a cellular automata system. The updating can be done in parallel and synchronously as is the case in CA dynamics or sequentially, one variable after another, as is usually the case for Monte–Carlo dynamics. Parallel, asynchronous updating is less common but can be envisaged too. Sequential and parallel updating schemes may yield different results since the interdependencies between variables are treated differently.

Viscosity A property of a fluid indicating how much momentum "diffuses" through the fluid in a inhomogeneous flow pattern. Equivalently, it describes the stress occurring between two fluid layers moving with different velocities. A high viscosity means that the resulting drag force is important and low viscosity means that this force is weak. Kinematic viscosity is usually denoted by ν and dynamic viscosity is denoted by $\eta = \nu\rho$ where ρ is the fluid density.

von Neumann neighborhood On a two-dimensional square lattice, the neighborhood including a central cell and its nearest neighbors north, south, east and west.

Ziff model A simple model describing adsorption–dissociation–desorption on a catalytic surface. This model is based upon some of the known steps of the reaction A–B_2 on a catalyst surface (for example CO–O_2).

Definition of the Subject

The computational science community has always been faced with the challenge of bringing efficient numerical tools to solve problems of increasing difficulty. Nowadays, the investigation and understanding of the so-called complex systems, and the simulation of all kinds of phenomena originating from the interaction of many components are of central importance in many area of science.

Cellular automata turn out to be a very fruitful approach to address many scientific problems by providing an efficient way to model and simulate specific phenomena for which more traditional computational techniques are hardly applicable.

The goal of this article is to provide the reader with the foundation of this approach, as well as a selection of simple applications of the cellular automata approach to the modeling of physical systems. We invite the reader

to consult the web site http://cui.unige.ch/~chopard/CA/Animations/img-root.html in order to view short movies about several of the models discussed in this article.

Introduction

Cellular automata (hereafter termed CA) are an idealization of the physical world in which space and time are of discrete nature. In addition to space and time, the physical quantities also take only a finite set of values. Since it has been proposed by von Neumann in the late 1940s, the cellular automata approach has been applied to a large range of scientific problems (see for instance [4,10,16,35,42,51]). International conferences (e. g. ACRI) and dedicated journals (J. of Cellular Automata) also describe current developments.

When von Neumann developed the concept of CA, its motivation was to extract the abstract (or algorithmic) mechanisms leading to self-reproduction of biological organisms [6].

Following the suggestions of S. Ulam, von Neumann addressed this question in the framework of a fully discrete universe made up of simple cells. Each cell was characterized by an internal state, which typically consists of a finite number of information bits. Von Neumann suggested that this system of cells evolves, in discrete time steps, like simple automata which only know of a simple recipe to compute their new internal state. The rule determining the evolution of this system is the same for all cells and is a function of the states of the cell itself and its neighbors.

Similarly to what happens in any biological system, the activity of the cells takes place simultaneously. The same clock is assumed to drive the evolution of every cell and the updating of their internal state occurs synchronously.

Such a fully discrete dynamical system (cellular space), as invented by von Neumann, is now referred to as a cellular automaton.

Among the early applications of CA, the *game of life* [19] is famous. In 1970, the mathematician John Conway proposed a simple model leading to complex behaviors. He imagined a two-dimensional square lattice, like a checkerboard, in which each cell can be either alive (state one) or dead (state zero). The updating rule of the game of life is as follows: a dead cell surrounded by exactly three living cells gets back to life; a living cell surrounded by less than two or more than three neighbors dies of isolation or overcrowdness. Here, the surrounding cells correspond to the neighborhood composed of the four nearest cells (North, South, East and West), plus the four second nearest neighbors, along the diagonals. It turns out that the game of life automaton has an unexpectedly rich behavior. Complex structures emerge out of a primitive "soup" and evolve so as to develop some skills that are absent of the elementary cells (see Fig. 1).

The game of life is a cellular automata capable of universal computations: it is always possible to find an initial configuration of the cellular space reproducing the behavior of any electronic gate and, thus, to mimic any computation process. Although this observation has little practical interest, it is very important from a theoretical point of view since it assesses the ability of CAs to be a nonrestrictive computational technique.

A very important feature of CAs is that they provide simple models of complex systems. They exemplify the fact that a collective behavior can emerge out of the sum of many, simply interacting, components. Even if the basic and local interactions are perfectly known, it is possible that the global behavior obeys new laws that are not obviously extrapolated from the individual properties, as if the whole were more than the sum of all the parts. These properties make cellular automata a very interesting approach to model physical systems and in particular to simulate complex and nonequilibrium phenomena.

The studies undertaken by S. Wolfram in the 1980s [50,51] clearly establishes that a CA may exhibit many of the behaviors encountered in continuous systems, yet in a much simpler mathematical framework. A further step is to recognize that CAs are not only behaving similarly to some dynamical processes, they can also represent an actual model of a given physical system, leading to macroscopic predictions that could be checked experimentally.

This fact follows from statistical mechanics which tells us that the macroscopic behavior of many systems is often only weakly related to the details of its microscopic reality. Only symmetries and conservation laws survive to the change of observation level: it is well known that the flows of a fluid, a gas or even a granular media are very similar at a macroscopic scale, in spite of their different microscopic nature.

When one is interested in the global or macroscopic properties of a system, it is therefore a clear advantage to invent a much simpler microscopic reality, which is more appropriate to the available numerical means of investigation.

An interesting example is the FHP fluid model proposed by Frisch, Hasslacher and Pomeau in 1986 [18] which can be viewed as a fully discrete molecular dynamic and yet behaves as predicted by the Navier–Stokes equation when the observation time and length scales are much larger than the lattice and automaton time step.

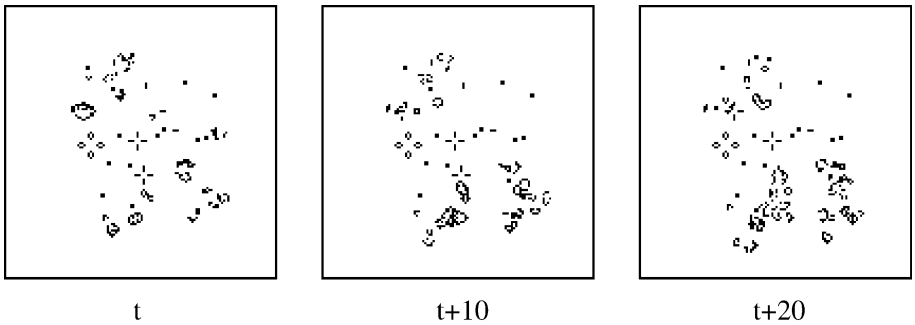

Cellular Automata Modeling of Physical Systems, Figure 1
The game of life automaton. *Black dots* represent living cells whereas dead cells are *white*. The figure shows the evolution of a random initial configuration and the formation of spatial structures, with possibly some emerging functionalities. The figure shows the evolution of some random initial configurations

A cellular automata model can then be seen as an idealized universe with its own microscopic reality but, nevertheless, with the same macroscopic behavior as given in the real system.

The cellular automata paradigm presents nevertheless some weaknesses inherent to its discrete nature. In the early 1990s, Lattice Boltzmann (LB) models were proposed to remedy some of these problems, using real-valued states instead of Boolean variables. It turns out that LB models are indeed a very powerful approach which combines numerical efficiency with the advantage of having a model whose microscopic components are intuitive. LB-fluids are more and more used to solve complex flows such as multi-component fluids or complicated geometries problems. See for instance [10,40,41,49] for an introduction to LB models.

Definition of a Cellular Automata

In order to give a definition of a cellular automaton, we first present a simple example, the so-called parity rule. Although it is very basic, the rule we discuss here exhibits a surprisingly rich behavior. It was proposed initially by Edward Fredkin in the 1970s [3] and is defined on a two-dimensional square lattice.

Each site of the lattice is a cell which is labeled by its position $\mathbf{r} = (i, j)$ where i and j are the row and column indices. A function $\psi(\mathbf{r}, t)$ is associated with the lattice to describe the state of each cell $\mathbf{r}$ at iteration t. This quantity can be either 0 or 1.

The cellular automata rule specifies how the states $\psi(\mathbf{r}, t+1)$ are to be computed from the states at iteration t. We start from an initial condition at time $t = 0$ with a given configuration of the values $\psi(\mathbf{r}, t = 0)$ on the lattice. The state at time $t = 1$ will be obtained as follows

(1) Each site $\mathbf{r}$ computes the sum of the values $\psi(\mathbf{r}', 0)$ on the four nearest neighbor sites $\mathbf{r}'$ at north, west, south, and east. The system is supposed to be periodic in both i and j directions (like on a torus) so that this calculation is well defined for all sites.
(2) If this sum is even, the new state $\psi(\mathbf{r}, t = 1)$ is 0 (white) and, else, it is 1 (black).

The same rule (steps 1 and 2) is repeated over to find the states at time $t = 2, 3, 4, \ldots$.

From a mathematical point of view, this cellular automata parity rule can be expressed by the following relation

$$\begin{aligned}\psi(i, j, t+1) = &\psi(i+1, j, t) \oplus \psi(i-1, j, t) \\ &\oplus \psi(i, j+1, t) \oplus \psi(i, j-1, t)\end{aligned} \tag{1}$$

where the symbol $\oplus$ stands for the exclusive OR logical operation. It is also the sum modulo 2: $1 \oplus 1 = 0 \oplus 0 = 0$ and $1 \oplus 0 = 0 \oplus 1 = 1$.

When this rule is iterated, very nice geometric patterns are observed, as shown in Fig. 2. This property of generating complex patterns starting from a simple rule is generic of many cellular automata rules. Here, complexity results from some spatial organization which builds up as the rule is iterated. The various contributions of successive iterations combine together in a specific way. The spatial patterns that are observed reflect how the terms are combined algebraically.

A computer implementation of this CA can be given. In Fig. 3 we propose as, an illustration, a Matlab program (the reader can also consider Octave which is a free version of Matlab).

On the basis of this example we now give a definition of a cellular automata. Formally, a cellular automata is a tuple $(A, \Psi, R, \mathcal{N})$ where

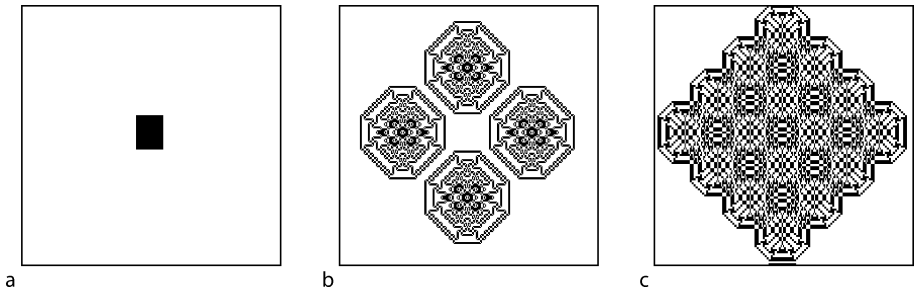

Cellular Automata Modeling of Physical Systems, Figure 2
The ⊕ rule (or parity rule) on a 256 × 256 periodic lattice. a Initial configuration; b and c configurations after $t_b = 93$ and $t_c = 110$ iterations, respectively

```
nx=128; ny=128;  % size of the domain: 128x128
a=zeros(nx,ny);  % the states are first initialized to 0

north=[ 2:nx,1];     % vectors to access the neighbors
south=[nx, 1:nx-1]; % corresponding to a cyclic permutation
east=[2:ny,1];       % of 1:nx  or 1:ny
west=[ny,1:ny-1];

% a central patch  is initialized with 1's
a(nx/2-3:nx/2+2, ny/2-4:ny/2+3)=1;

for t=1:65   % let us do 65 iterations
     pcolor(a)       % build a graphical representation
     axis off
     axis square
     shading flat
     drawnow         % display it
     somme=a(north,:) + a(south,:) + a(:,west) + a(:,east);
     a=mod(somme,2);
end
```

Cellular Automata Modeling of Physical Systems, Figure 3
A example of a Matlab program for the parity rule

(i) A is a regular lattice of cells covering a portion of a d-dimensional space.
(ii) A set $\Psi(\mathbf{r}, t) = \{\Psi^{(1)}(\mathbf{r}, t), \Psi^{(2)}(\mathbf{r}, t), \ldots, \Psi^{(m)}(\mathbf{r}, t)\}$ of m Boolean variables attached to each site $\mathbf{r}$ of the lattice and giving the local state of the cells at time t.
(iii) A set R of rules, $R = \{R^{(1)}, R^{(2)}, \ldots, R^{(m)}\}$, which specifies the time evolution of the states $\Psi(\mathbf{r}, t)$ in the following way

$$\Psi^{(j)}(\mathbf{r}, t + \Delta_t) = R^{(j)}(\Psi(\mathbf{r}, t), \Psi(\mathbf{r} + \mathbf{v_1}, t), \Psi(\mathbf{r} + \mathbf{v_2}, t), \ldots, \Psi(\mathbf{r} + \mathbf{v_q}, t)) \quad (2)$$

where $\mathbf{r} + \mathbf{v_k}$ designate the cells belonging to the neighborhood $\mathcal{N}$ of cell $\mathbf{r}$.

In the above definition, the rule R is identical for all sites and is applied simultaneously to each of them, leading to synchronous dynamics. As the number of configurations of the neighborhood is finite, it is common to precompute all the values of R in a lookup table. Otherwise, an algebraic expression can be used and evaluated at each iteration, for each cell, as in Eq. (1).

It is important to notice that the rule is *homogeneous*, that is it cannot not depend explicitly on the cell position $\mathbf{r}$. However, spatial (or even temporal) inhomogeneities can be introduced anyway by having some $\Psi_j(\mathbf{r})$ systematically

1 in some given locations of the lattice to mark particular cells on which a different rule applies. Boundary cells are a typical example of spatial inhomogeneities. Similarly, it is easy to alternate between two rules by having a bit which is 1 at even time steps and 0 at odd time steps.

The neighborhood $\mathcal{N}$ of each cell (i. e. the spatial region around each cell used to compute the next state) is usually made of its adjacent cells. It is often restricted to the nearest or next to nearest neighbors, otherwise the complexity of the rule is too large. For a two-dimensional cellular automaton, two neighborhoods are often considered: the von Neumann neighborhood which consists of a central cell (the one which is to be updated) and its four geographical neighbors North, West, South, and East. The Moore neighborhood contains, in addition, the second nearest neighbor North-East, North-West, South-West, and South-East, that is a total of nine cells. Another interesting neighborhood is the Margolus neighborhood briefly described in the glossary.

According to the above definition, a cellular automaton is deterministic. The rule **R** is some well-defined function and a given initial configuration will always evolve identically. However, it may be very convenient for some applications to have a certain degree of randomness in the rule. For instance, it may be desirable that a rule selects one outcome among several possible states, with a probability p. Cellular automata whose updating rule is driven by some external probabilities are called *probabilistic* cellular automata. On the other hand, those which strictly comply with the definition given above, are referred to as *deterministic* cellular automata.

Probabilistic cellular automata are a very useful generalization because they offer a way to adjust the parameters of a rule in a continuous range of values, despite the discrete nature of the cellular automata world. This is very convenient when modeling physical systems in which, for instance, particles are annihilated or created at some given rate.

Limitations, Advantages, Drawbacks, and Extensions

The interpretation of the cellular automata dynamics in terms of simple "microscopic" rules offers a very intuitive and powerful approach to model phenomena that are very difficult to include in more traditional approaches (such as differential equations). For instance, boundary conditions are often naturally implemented in a cellular automata model because they have a natural interpretation at this level of description (e. g. particles bouncing back on an obstacle).

Numerically, an advantage of the CA approach is its simplicity and its adequation to computer architectures and parallel machines. In addition, working with Boolean quantities prevent numerical instabilities since an exact computation is made. There is no truncation or approximation in the dynamics itself. Finally, a CA model is an implementation of an N-body system where all correlations are taken into account, as well as all spontaneous fluctuations arising in a system made up of many particles.

On the other hand, cellular automata models have several drawbacks related to their fully discrete nature. An important one is the statistical noise requiring systematic averaging processes. Another one is the little flexibility to adjust parameters of a rule in order to describe a wider range of physical situations.

The Lattice Boltzmann approach solves several of the above problems. On the other hand, it may be numerically unstable and, also, requires some hypotheses of molecular chaos which reduces the some of the richness of the original CA dynamics [10].

Finally, we should remark that the CA approach is not a rigid framework but should allow for many extensions according to the problem at hand. The CA methodology is a philosophy of modeling where one seeks a description in terms of simple but essential mechanisms. Its richness and interest comes from the microscopic contents of its rule for which there is, in general, a clear physical or intuitive interpretation of the dynamics directly at the level of the cell.

Einstein's quote "Everything should be made as simple as possible, but not simpler" is a good illustration of the CA methodology to the modeling of physical systems.

Applications

Many physical situations, like fluid flows, pattern formation, reaction-diffusion processes, nucleation-aggregation growth phenomena, phase transition, population dynamics, or traffic processes are very well suited to the cellular automata approach because both space and time play a central role in the behavior of these systems.

Below we describe several applications which illustrate the potential of the approach and that can be extended in order to address a wide class of scientific problems.

A Growth Model

A simple class of cellular automata rules consists of the so-called *majority rules*. The updating selects the new state of each cell so as to conform to the value currently held by the majority of the neighbors. Typically, in these majority rules, the state is either 0 or 1.

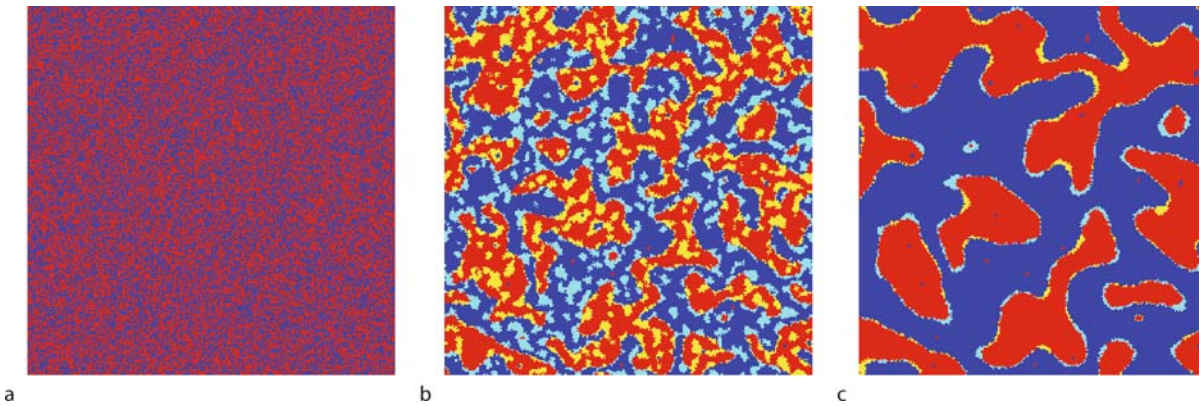

Cellular Automata Modeling of Physical Systems, Figure 4
Evolution of the twisted majority. The inherent "surface tension" present in the rule tends to separate the *red phases* $s = 1$ from the blue phase $s = 0$. The snapshots a, b and c correspond to $t = 0$, $t = 72$, and $t = 270$ iterations, respectively. The other colors indicate how "capes" have been eroded and "bays" filled: *light blue* shows the blue regions that have been eroded during the last few iterations and *yellow* marks the red regions that have been filled

A very interesting behavior is observed with the twisted majority rule proposed by G. Vichniac [44]: in two-dimensions, each cell considers its Moore neighborhood (i. e. itself plus its eight nearest neighbors) and computes the sum of the cells having a value 1. This sum can be any value between 0 and 9. The new state $s(t+1)$ of each cell is then determined from this local sum, according to the following table

$$\begin{array}{rcccccccccc} \text{sum}(t) & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ s(t+1) & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \end{array} \tag{3}$$

As opposed to the plain majority rule, here, the two middle entries of the table have been swapped. Therefore, when there is a slight majority of 1 around a cell, it turns to 0. Conversely, if there is a slight majority of 0, the cell becomes 1.

Surprisingly enough this rule describes the interface motion between two phases, as illustrated in Fig. 4. Vichniac has observed that the normal velocity of the interface is proportional to its local curvature, as required by the Allen–Cahn [21] equation. Of course, due to its local nature, the rule cannot detect the curvature of the interface directly. However, as the rule is iterated, local information is propagated to the nearest neighbors and the radius of curvature emerges as a collective effect.

This rule is particularly interesting when the initial configuration is a random mixture of the two phases, with equal concentration. Otherwise, some pathological behaviors may occur. For instance, an initial square of 1's surrounded by zero's will not evolve: 90-degree angles are not eroded and remain stable structures.

Ising-Like Dynamics

The Ising model is extensively used in physics. Its basic constituents are spins s_i which can be in one of two states: $s_i \in \{-1, 1\}$. These spins are organized on a regular lattice in d-dimensions and coupled in the sense that each pair (s_i, s_j) of neighbor spins contributes an amount $-Js_is_j$ to the energy of the system. Intuitively, the dynamics of such a system is that a spin flips ($s_i \to -s_i$) if this is favorable in view of the energy of the local configuration.

Vichniac [44], in the 1980s, proposed a CA rule, called Q2R, simulating the behavior of Ising spin dynamics. The model is as follows:

We consider a two-dimensional square lattice such that each site holds a spin s_i which is either up ($s_i = 1$) or down ($s_i = 0$) (instead of ± 1). The coupling between spins is assumed to come from the von Neumann neighborhood (i. e. north, west, south, and east neighbors).

In this simple model, the spins will flip (or not flip) during their discrete time evolution according to a local energy conservation principle. This means we are considering a system which cannot exchange energy with its surroundings. The model will be a microcanonical cellular automata simulation of Ising spin dynamics, without a temperature but with a critical energy.

A spin s_i can flip at time t to become $1 - s_i$ at time $t + 1$ if and only if this move does not cause any energy change. Accordingly, spin s_i will flip if the number of its neighbors with spin up is the same as the number of its neighbors with spin down. However, one has to remember that the motion of all spins is simultaneous in a cellular automata. The decision to flip is based on the assumption

that the neighbors are not changing. If they are allowed to flip too, (because they obey the same rule), then energy may not be conserved.

A way to cure this problem is to split the updating in two phases and consider a partition of the lattice in odd and even sites (e. g. the white and black squares of a chessboard in 2D): first, one flips the spins located at odd positions, according to the configuration of the even spins. In the second phase, the even sublattice is updated according to the odd one. The spatial structure (defining the two sublattices) is obtained by adding an extra bit b to each lattice site, whose value is 0 for the odd sublattice and 1 for the even sublattice. The flipping rule described earlier is then regulated by the value of b. It takes place only for those sites for which $b = 1$. Of course, the value of b is also updated at each iteration according to $b(t+1) = 1 - b(t)$, so that at the next iteration, the other sublattice is considered. In two-dimensions, the Q2R rule can be then expressed by the following expressions

$$s_{ij}(t+1) = \begin{cases} 1 - s_{ij}(t) & \text{if } b_{ij} = 1 \text{ and } s_{i-1,j} \\ & +s_{i+1,j} + s_{i,j-1} + s_{i,j+1} = 2 \\ s_{ij}(t) & \text{otherwise} \end{cases} \tag{4}$$

and

$$b_{ij}(t+1) = 1 - b_{ij}(t) \tag{5}$$

where the indices (i, j) label the Cartesian coordinates and $s_{ij}(t = 0)$ is either one or zero.

The question is now how well does this cellular automata rule perform to describe an Ising model? Figure 5 shows a computer simulation of the Q2R rule, starting from an initial configuration with approximately 11% of spins $s_{ij} = 1$ (Fig. 5a). After a transient phase (figures b and c), the system reaches a stationary state where domains with "up" magnetization (white regions) are surrounded by domains of "down" magnetization (black regions).

In this dynamics, energy is exactly conserved because that is the way the rule is built. However, the number of spins down and up may vary. In the present experiment, the fraction of spins up increases from 11% in the initial state to about 40% in the stationary state. Since there is an excess of spins down in this system, there is a resulting macroscopic magnetization.

It is interesting to study this model with various initial fractions ρ_s of spins up. When starting with a random initial condition, similar to that of Fig. 5a, it is observed that, for many values of ρ_s, the system evolves to a state where there is, in the average, the same amount of spin down and up, that is no macroscopic magnetization. However, if the initial configuration presents a sufficiently large excess of one kind of spins, then a macroscopic magnetization builds up as time goes on. This means there is a phase transition between a situation of zero magnetization and a situation of positive or negative magnetization.

It turns out that this transition occurs when the total energy E of the system is low enough (a low energy means that most of the spins are aligned and that there is an excess of one species over the other), or more precisely when E is smaller than a critical energy E_c. In that sense, the Q2R rule captures an important aspect of a real magnetic system, namely a non-zero magnetization at low energy (which can be related to a low temperature situation) and a transition to a nonmagnetic phase at high energy.

However, Q2R also exhibits unexpected behavior that is difficult to detect from a simple observation. There is a breaking of ergodicity: a given initial configuration of energy E_0 evolves without visiting completely the region of the phase space characterized by $E = E_0$.

This is illustrated by the following simple 1D example, where a ring of four spins with periodic boundary condition are considered.

$$\begin{aligned} t &: \ 1001 \\ t+1 &: \ 1100 \\ t+2 &: \ 0110 \\ t+3 &: \ 0011 \\ t+4 &: \ 1001 \end{aligned} \tag{6}$$

After four iterations, the system cycles back to its original state. The configuration of this example has $E_0 = 0$. As we observed, it never evolves to 0111, which is also a configuration of zero energy. This nonergodicity means that not only energy is conserved during the evolution of the automaton, but also another quantity which partitions the energy surface in independent regions.

Competition Models and Cell Differentiation

In Sect. "A Growth Model" we have discussed a majority rule in which the cells imitate their neighbors. In some sense, this corresponds to a cooperative behavior between the cells. A quite different situation can be obtained if the cells obey competitive dynamics. For instance, we may imagine that the cells compete for some resources at the expense of their nearest neighbors. A winner is a cell of state 1 and a loser a cell of state 0. No two winner cells can be neighbors and any loser cell must have at least one winner neighbor (otherwise nothing would have prevented it to also win).

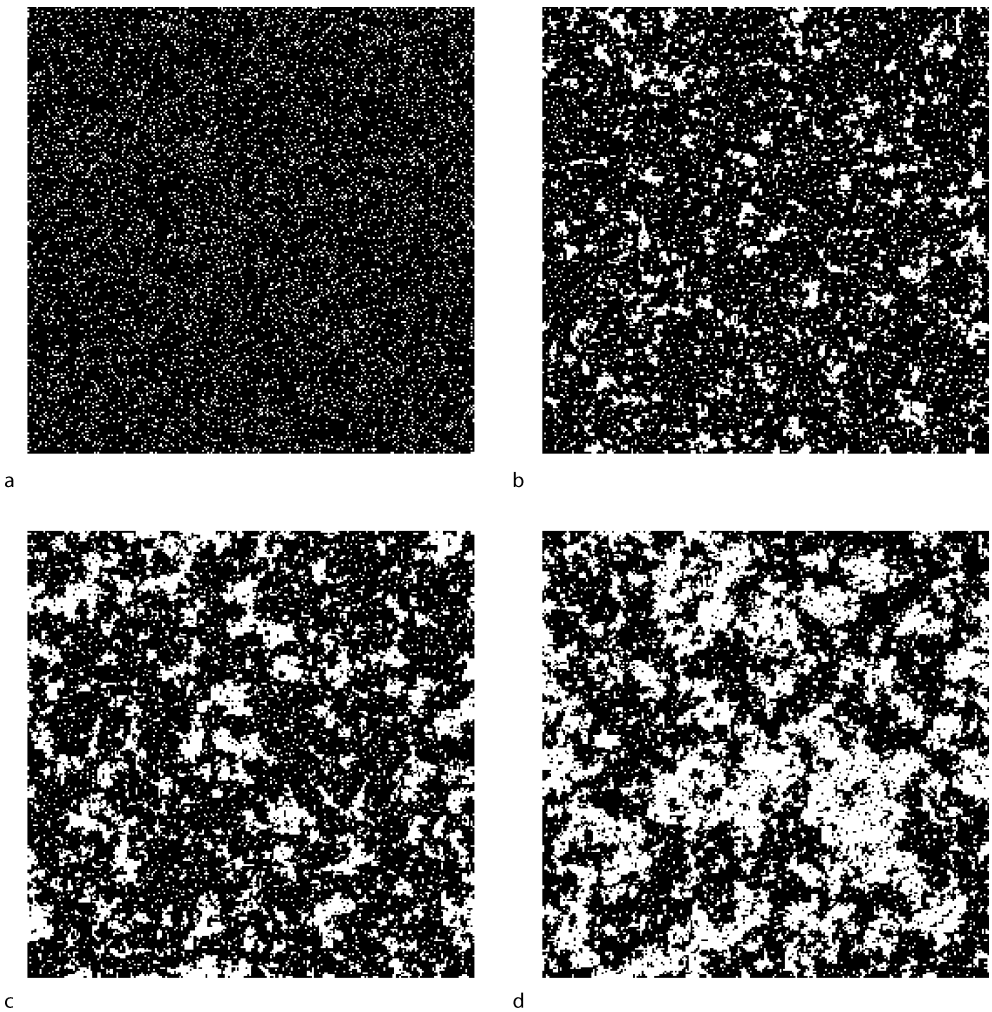

Cellular Automata Modeling of Physical Systems, Figure 5
Evolution of a system of spins with the Q2R rule. *Black* represents the spins down $s_{ij} = 0$ and *white* the spins up $s_{ij} = 1$. The four images a, b, c, and d show the system at four different times $t_a = 0 < t_b \ll t_{cd}$

It is interesting to note that this problem has a direct application in biology, to study cell differentiation. It has been observed in the development of Drosophila that about 25% of the cells forming the embryo evolve to the state of neuroblast, while the remaining 75% do not. How can we explain this differentiation and the observed fraction since, at the beginning of the process all cells can be assumed equivalent? A possible mechanism [28] is that some competition takes place between the adjacent biological cells. In other words, each cell produces some substance S but the production rate is inhibited by the amount of S already present in the neighboring cells. Differentiation occurs when a cell reaches a level of S above a given threshold.

The competition CA model we propose to describe this situation is the following. Because of the analogy with the biological system, we shall consider a hexagonal lattice, which is a reasonable approximation of the cell arrangement observed in the Drosophila embryo (see Fig. 6). We assume that the values of S can be 0 (inhibited) or 1 (active) in each lattice cell.

- A $S = 0$ cell will grow (i. e. turn to $S = 1$) with probability p_{grow} provided that all its neighbors are 0. Otherwise, it stays inhibited.
- A cell in state $S = 1$ will decay (i. e. turn to $S = 0$) with probability p_{decay} if it is surrounded by at least one active cell. If the active cell is isolated (all the neighbors are in state 0) it remains in state 1.

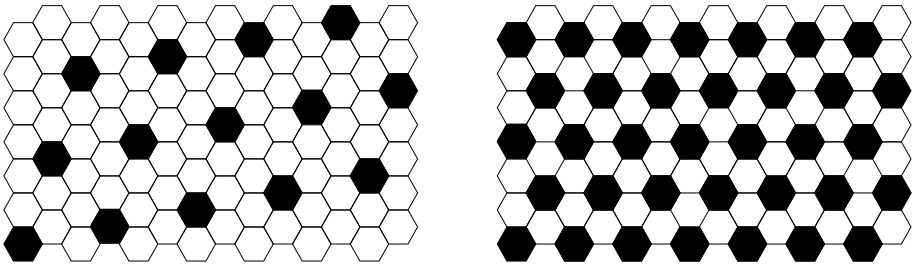

Cellular Automata Modeling of Physical Systems, Figure 6
The hexagonal lattice used for the competition-inhibition CA rule. *Black cells* are cells of state 1 (winners) and *white cells* are cells of state 0 (losers). The two possible final states with a fully regular structure are illustrated with density 1/3 and 1/7 of a winner, respectively

The evolution stops (stationary process) when no $S = 1$ cell feels any more inhibition from its neighbors and when all $S = 0$ cells are inhibited by their neighborhood. Then, with our biological interpretation, cells with $S = 1$ are those which will differentiate.

What is the expected fraction of these $S = 1$ cells in the final configuration? Clearly, from Fig. 6, the maximum value is 1/3. According to the inhibition condition we imposed, this is the close-packed situation on the hexagonal lattice. On the other hand, the minimal value is 1/7, corresponding to a situation where the lattice is partitioned in blocks with one active cell surrounded by six inhibited cells. In practice we do not expect any of these two limits to occur spontaneously after the automaton evolution. On the contrary, we should observe clusters of close-packed active cells surrounded by defects, i. e. regions of low density of active cells.

CA simulations show indeed that the final fraction s of active cells is a mix of the two limiting situations of Fig. 6

$$.23 \leq s \leq .24$$

almost irrespectively of the values chosen for p_{anihil} and p_{growth}.

This is exactly the value expected from the biological observations made on the Drosophila embryo. Thus, cell differentiation can be explained by a geometrical competition without having to specify the inhibitory couplings between adjacent cell and the production rate (i. e. the values of p_{anihil} and p_{growth}): the result is quite robust against any possible choices.

Traffic Models

Cellular automata models for road traffic have received a great deal of interest during the past few years (see [13,32,33,36,37,47,48,52] for instance).

One-dimensional models for single lane car motions are quite simple and elegant. The road is represented as a line of cells, each of them being occupied or not by a vehicle. All cars travel in the same direction (say to the right). Their positions are updated synchronously. During the motion, each car can be at rest or jump to the nearest neighbor site, along the direction of motion. The rule is simply that a car moves only if its destination cell is empty. This means that the drivers do not know whether the car in front will move or will be blocked by another car. Therefore, the state of each cell s_i is entirely determined by the occupancy of the cell itself and its two nearest neighbors s_{i-1} and s_{i+1}. The motion rule can be summarized by the following table, where all eight possible configurations $(s_{i-1}s_is_{i+1})_t \to (s_i)_{t+1}$ are given

$$\underbrace{(111)}_{1}\ \underbrace{(110)}_{0}\ \underbrace{(101)}_{1}\ \underbrace{(100)}_{1}\ \underbrace{(011)}_{1}\ \underbrace{(010)}_{0}\ \underbrace{(001)}_{0}\ \underbrace{(000)}_{0}\,. \tag{7}$$

This cellular automaton rule turns out to be Wolfram rule 184 [50,52].

These simple dynamics capture an interesting feature of real car motion: traffic congestion. Suppose we have a low car density ρ in the system, for instance something like

$$\ldots 00100000100100000010\ldots\,. \tag{8}$$

This is a *free* traffic regime in which all the cars are able to move. The average velocity $\langle v\rangle$ defined as the number of motions divided by the number of cars is then

$$\langle v_{\text{f}}\rangle = 1 \tag{9}$$

where the subscript f indicates a free state. On the other hand, in a high density configuration such as

$$\ldots 110101110101101110\ldots\,. \tag{10}$$

only six cars over 12 will move and $\langle v \rangle = 1/2$. This is a partially jammed regime.

If the car positions were uncorrelated, the number of moving cars (i. e. the number of particle-hole pairs) would be given by $L\rho(1-\rho)$, where L is the system size. Since the number of cars is ρL, the average velocity would be

$$\langle v_{\text{uncorrel}} \rangle = 1 - \rho \,. \tag{11}$$

However, in this model, the car occupancy of adjacent sites is highly correlated and the vehicles cannot move until a hole has appeared in front of them. The car distribution tries to self-adjust to a situation where there is one spacing between consecutive cars. For densities less than one-half, this is easily realized and the system can organize to have one car every other site.

Therefore, due to these correlations, Eq. (11) is wrong in the high density regime. In this case, since a car needs a hole to move to, we expect that the number of moving cars simply equals the number of empty cells [52]. Thus, the number of motions is $L(1-\rho)$ and the average velocity in the jammed phase is

$$\langle v_j \rangle = \frac{1-\rho}{\rho} \,. \tag{12}$$

From the above relations we can compute the so-called fundamental flow diagram, i. e. the relation between the flow of cars $\rho\langle v \rangle$ as a function of the car density ρ: for $\rho \le 1/2$, we use the free regime expression and $\rho\langle v \rangle = \rho$. For densities $\rho > 1/2$, we use the jammed expression and $\rho\langle v \rangle = 1-\rho$. The resulting diagram is shown in Fig. 7. As in real traffic, we observe that the flow of cars reaches a maximum value before decreasing.

A richer version of the above CA traffic model is due to Nagel and Schreckenberg [33,47,48]. The cars may have several possible velocities $u = 0, 1, 2, \ldots, u_{\max}$. Let u_i be the velocity of car i and d_i the distance, along the road, separating cars i and $i+1$. The updating rule is:

- The cars accelerate when possible: $u_i \to u_i' = u_i + 1$, if $u_i < u_{\max}$.
- The cars slow down when required: $u_i' \to u_i'' = d_i - 1$, if $u_i' \ge d_i$.
- The cars have a random behavior: $u_i'' \to u_i''' = u_i'' - 1$, with probability p_i if $u_i'' > 0$.
- Finally, the cars move u_i''' sites ahead.

This rule captures some important behaviors of real traffic on a highway: velocity fluctuations due to a non-deterministic behavior of the drivers, and "stop-and-go" waves observed in a high-density traffic regime.

We refer the reader to recent literature for the new developments of this topic. See for instance [24,25].

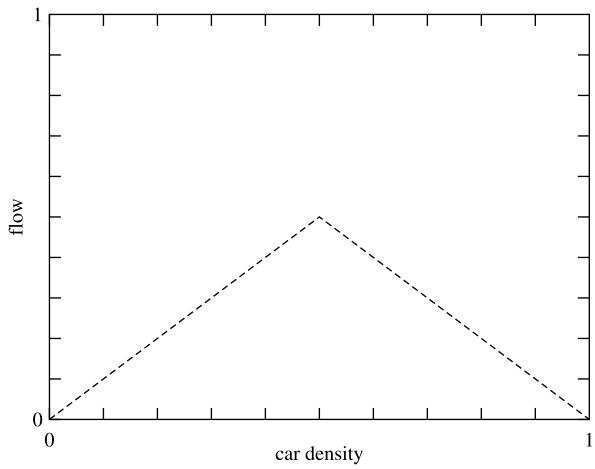

Cellular Automata Modeling of Physical Systems, Figure 7
Traffic flow diagram for the simple CA traffic rule

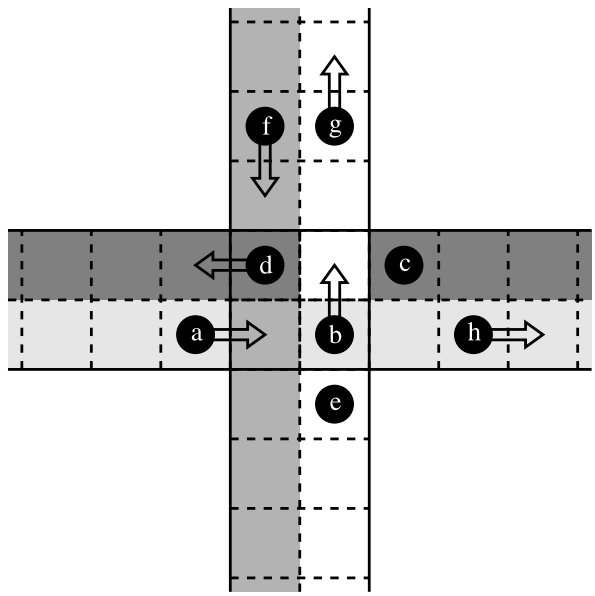

Cellular Automata Modeling of Physical Systems, Figure 8
The four central cells represent a roundabout which is traveled counterclockwise. The gray levels indicate the different traffic lanes: *white* is a northbound lane, *light gray* an eastbound lane, *gray* a southbound lane and, finally, *dark gray* is a westbound lane. The *dots* labeled *a*, *b*, *c*, *d*, *e*, *f*, *g*, and *h* are cars which will move to the destination cell indicated by the arrows, as determined by some local decision rule. Cars without an arrow are forbidden to move

Note that a street network can also be described using a CA. A possible approach is to model a road intersection as a roundabout. Cars in the roundabout have priority over those willing to enter. Figure 8 illustrates a simple four-way road junction.

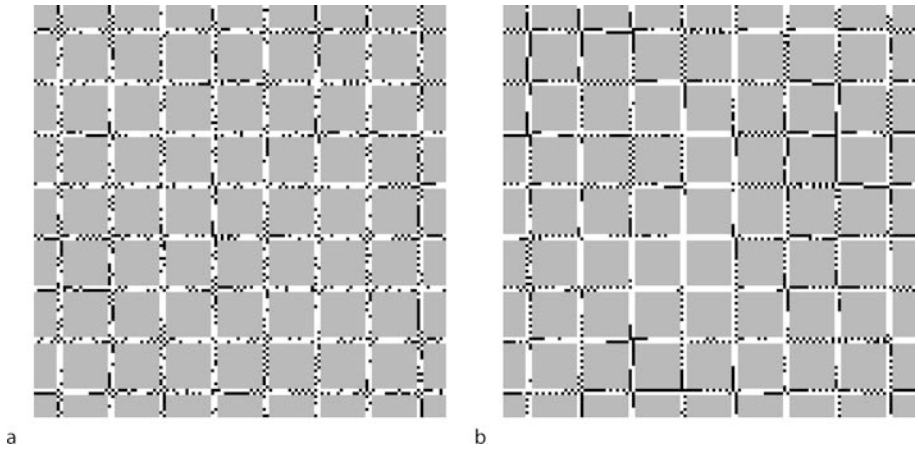

Cellular Automata Modeling of Physical Systems, Figure 9
Traffic configuration after 600 iterations, for a car density of 30%. Streets are *white*, buildings *gray* and the *black pixels* represent the cars. Situation a corresponds to the roundabout junctions, whereas image b mimics the presence of traffic lights. In the second case, queues are more likely to form and the global mobility is less than in the first case

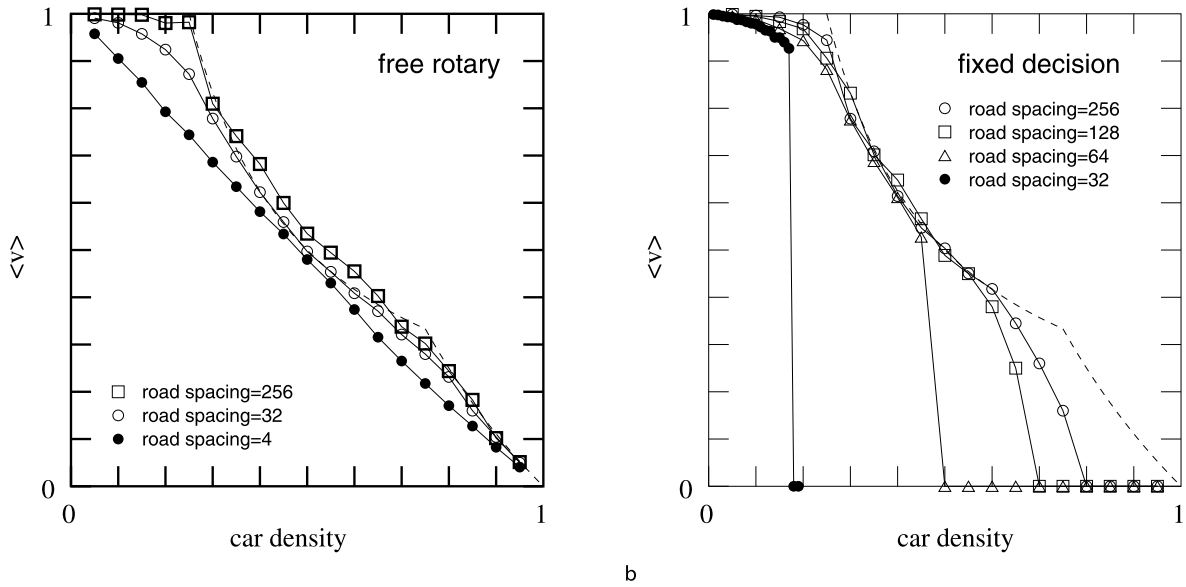

Cellular Automata Modeling of Physical Systems, Figure 10
Average velocity versus average density for the cellular automata street network, for a time-uncorrelated turning strategies and b a fixed driver decision. The different curves correspond to different distances *L* between successive road junctions. The *dashed line* is the analytical prediction (see [13]). Junction deadlock is likely to occur in b, resulting in a completely jammed state

Traffic lights can also be modeled in a CA by stopping, during a given number of iterations, the car reaching a given cell. Figure 9 illustrates CA simulation of a Manhattan-like city in which junctions are either controlled by a roundabout or by a traffic light. In both cases, the destination of a car reaching the junction is randomly chosen.

Figure 10 shows the fundamental flow diagram obtained with the CA model, for a Manhattan-like city governed by roundabouts separated by a distance L.

CA modeling of urban traffic has been used for real situations by many authors (see for instance [11]) and some cities in Europe and USA use the CA approach as a way to manage traffic.

Note finally that crowd motion is also addressed within the CA framework. Recent results [7,29] show that the approach is quite promising to plan evacuation strategies and reproduce several of the motion patterns observed in real crowds.

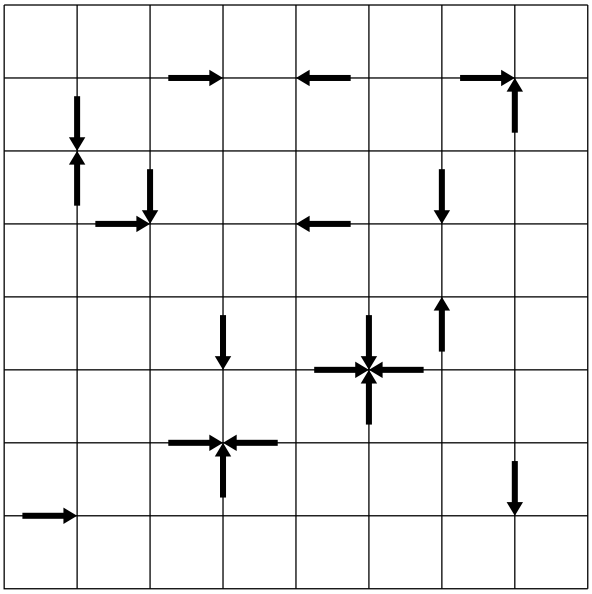

Cellular Automata Modeling of Physical Systems, Figure 11
Example of a configuration of HPP particles

A Simple Gas: The HPP Model

The HPP rule is a simple example of an important class of cellular automata models: lattice gas automata (LGA). The basic ingredient of such models are point particles that move on a lattice, according to appropriate rules so as to mimic fully discrete "molecular dynamics."

The HPP lattice gas automata is traditionally defined on a two-dimensional square lattice. Particles can move along the main directions of the lattice, as shown in Fig. 11. The model limits to 1 the number of particles entering a given site with a given direction of motion. This is the exclusion principle which is common in most LGA (LGA models without the exclusion principle are called multi-particle models [10]).

With at most one particle per site and direction, four bits of information at each site are enough to describe the system during its evolution. For instance, if at iteration t site $\mathbf{r}$ has the following state $s(\mathbf{r}, t) = (1011)$, it means that three particles are entering the site along direction 1, 3, and 4, respectively.

The cellular automata rule describing the evolution of $s(\mathbf{r}, t)$ is often split into two steps: collision and motion (or propagation). The collision phase specifies how the particles entering the same site will interact and change their trajectories. During the propagation phase, the particles actually move to the nearest neighbor site they are traveling to. This decomposition into two phases is a quite convenient way to partition the space so that the collision rule is purely local.

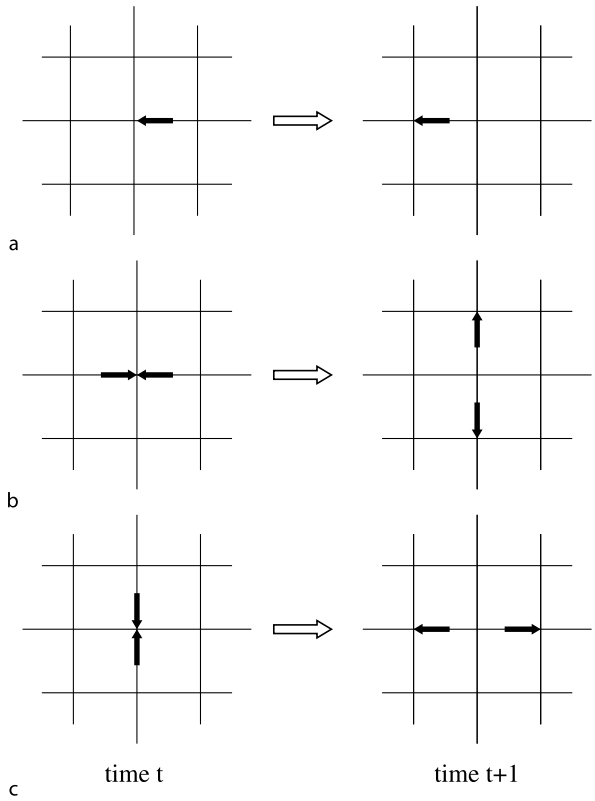

Cellular Automata Modeling of Physical Systems, Figure 12
The HPP rule: a a single particle has a ballistic motion until it experiences a collision; b and c the two nontrivial collisions of the HPP model: two particles experiencing a head on collision are deflected in the perpendicular direction. In the other situations, the motion is ballistic, that is the particles are transparent to each other when they cross the same site

Figure 12 illustrates the HPP rules. According to our Boolean representation of the particles at each site, the collision part for the two head on collisions are expressed as

$$(1010) \rightarrow (0101) \qquad (0101) \rightarrow (1010) \tag{13}$$

all the other configurations being unchanged. During the propagation phase, the first bit of the state variable is shifted to the east neighbor cell, the second bit to the north and so on.

The aim of this rule is to reproduce some aspect of the real interactions between particles, namely that momentum and particle number are conserved during a collision. From Fig. 12, it is easily checked that these properties are obeyed: a pair of zero momentum particles along a given direction is transformed into another pair of zero momentum along the perpendicular axis.

It is easy to express the HPP model in a mathematical form. For this purpose, the so-called occupation numbers $n_i(\mathbf{r}, t)$ are introduced for each lattice site $\mathbf{r}$ and each time step t. The index i labels the lattice directions (or the possible velocities of the particles). In the HPP model, the lattice has four directions (North, West, South, and East) and i runs from 1 to 4.

By definition and due to the exclusion principle, the n_i's are Boolean variables

$$n_i(\mathbf{r}, t) = \begin{cases} 1 & \text{if a particle is entering site } \mathbf{r} \text{ at time } t \\ & \text{along lattice direction } i \\ 0 & \text{otherwise .} \end{cases}$$

From this definition it is clear that, for HPP, the n_i's are simply the components of the state s introduced above

$$s = (n_1, n_2, n_3, n_4) .$$

In an LGA model, the microdynamics can be naturally expressed in terms of the occupation numbers n_i as

$$n_i(\mathbf{r} + \mathbf{v_i}\Delta_t, t + \Delta_t) = n_i(\mathbf{r}, t) + \Omega_i(n(\mathbf{r}, t)) \qquad (14)$$

where $\mathbf{v_i}$ is a vector denoting the speed of the particle in the ith lattice direction. The function Ω is called the collision term and it describes the interaction of the particles which meet at the same time and same location.

Note that another way to express Eq. (14) is through the so-called collision and propagation operators C and P

$$n(t + \Delta_t) = PCn(t) \qquad (15)$$

where $n(t)$ describe the set of values $n_i(\mathbf{r}, t)$ for all i and $\mathbf{r}$. The quantities C and P act over the entire lattice. They are defined as

$$(Pn)_i(\mathbf{r}) = n_i(\mathbf{r} - \mathbf{v_i}\Delta_t) \qquad (Cn)_i(\mathbf{r}) = n_i(\mathbf{r}) + \Omega_i .$$

More specifically, for the HPP model, it can be shown [10] that the collision and propagation phase can be expressed as

$$\begin{aligned} & n_i(\mathbf{r} + \mathbf{v_i}\Delta_t, t + \Delta_t) \\ & \quad = n_i - n_i n_{i+2}(1 - n_{i+1})(1 - n_{i+3}) \\ & \qquad + n_{i+1} n_{i+3}(1 - n_i)(1 - n_{i+2}) . \end{aligned} \qquad (16)$$

In this equation, the values $i + m$ are wrapped onto the values 1 to 4 and the right-hand term is computed at position $\mathbf{r}$ and time t.

The HPP rule captures another important ingredient of the microscopic nature of a real interaction: invariance under time reversal. Figure 12b and c show that, if at some given time, the directions of motion of all particles are reversed, the system will just trace back its own history. Since the dynamics of a deterministic cellular automaton is exact, this fact allows us to demonstrate the properties of physical systems to return to their original situation when all the particles reverse their velocity.

Figure 13 illustrates the time evolution of an HPP gas initially confined in the left compartment of a container. There is an aperture on the wall of the compartment and the gas particles will flow so as to fill the entire space available to them. In order to include a solid boundary in the system, the HPP rule is modified as follows: when a site is a wall (indicated by an extra bit), the particles no longer experience the HPP collision but bounce back from where they came. Therefore, particles cannot escape a region delimited by such a reflecting boundary.

If the system of Fig. 13 is evolved, it reaches an equilibrium after a long enough time and no macroscopic trace of its initial state is visible any longer. However, no information has been lost during the process (no numerical dissipation) and the system has the memory of where it comes from. Reversing all the velocities and iterating the HPP rule makes all particles go back to the compartment in which they were initially located.

Reversing the particle velocity can be described by an operator R which swaps occupation numbers with opposite velocities

$$(Rn)_i = n_{i+2} .$$

The reversibility of HPP stem from the fact that the collision and propagation operators obey

$$PRP = R \qquad CRC = R .$$

Thus

$$(PC)^n PR(PC)^n = R$$

which shows that the system will return to its initial state (though with opposite velocities) if the first n iterations are followed by a velocity change $\mathbf{v_i} \to -\mathbf{v_i}$, a propagation and again n iterations.

This time-reversal behavior is only possible because the dynamics are perfectly exact and no numerical errors are present in the numerical scheme. If one introduces externally some errors (for instance, one can add an extra particle in the system) before the direction of motion of each particle is reversed, then reversibility is lost.

Note that this property has inspired a new symmetric cryptography algorithm called Crystal [30], which exploits and develops the existing analogy between discrete physical models of particles and the standard diffusion-confusion paradigm of cryptography proposed by Shannon [39].

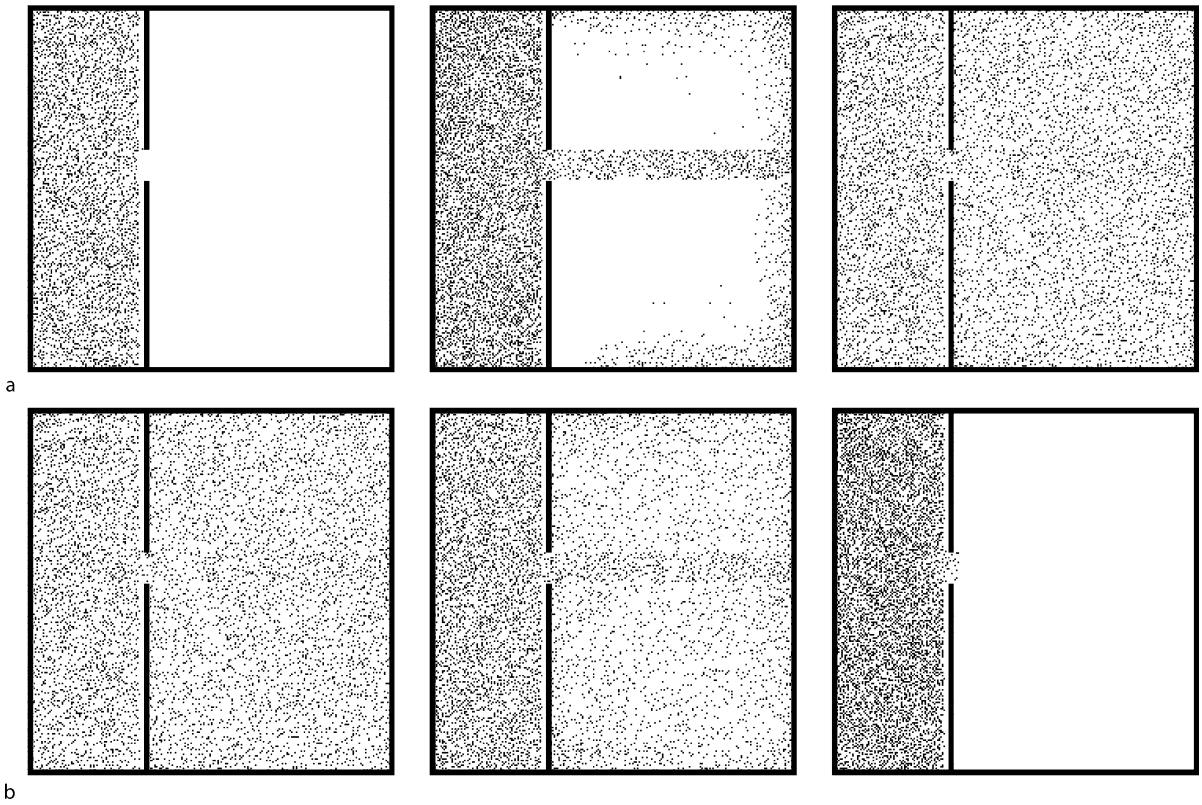

Cellular Automata Modeling of Physical Systems, Figure 13
Time evolution of an HPP gas. a From the initial state to equilibrium. b Illustration of time reversal invariance: in the rightmost image of a, the velocity of each particle is reversed and the particles naturally return to their initial position

The FHP Model

The HPP rule is interesting because it illustrates the basic ingredients of LGA models. However, the capability of this rule to model a real gas of particles is poor, due to a lack of isotropy and spurious invariants. A remedy to this problem is to use a different lattice and a different collision model.

The FHP rule (proposed by Frisch, Hasslacher, and Pomeau [18] in 1986) was the first CA whose behavior was shown to reproduce, within some limits, a two-dimensional fluid.

The FHP model is an abstraction, at a microscopic scale, of a fluid. It is expected to contain all the salient features of a real fluid. It is well known that the continuity and Navier–Stokes equations of hydrodynamics express the local conservation of mass and momentum in a fluid. The detailed nature of the microscopic interactions does not affect the form of these equations but only the values of the coefficients (such as the viscosity) appearing in them. Therefore, the basic ingredients one has to include in the microdynamics of the FHP model is the conservation of particles and momentum after each updating step. In addition, some symmetries are required so that, in the macroscopic limit, where time and space can be considered as continuous variables, the system be isotropic.

As in the case of the HPP model, the microdynamics of FHP is given in terms of Boolean variables describing the occupation numbers at each site of the lattice and at each time step (i. e. the presence or the absence of a fluid particle). The FHP particles move in discrete time steps, with a velocity of constant modulus, pointing along one of the six directions of the lattice.

Interactions take place among particles entering the same site at the same time and result in a new local distribution of particle velocities. In order to conserve the number of particles and the momentum during each interaction, only a few configurations lead to a nontrivial collision (i. e. a collision in which the directions of motion have changed). For instance, when exactly two particles enter the same site with opposite velocities, both of them are deflected by 60 degrees so that the output of the collision is

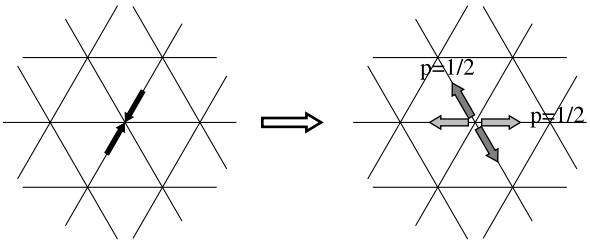

Cellular Automata Modeling of Physical Systems, Figure 14
The two-body collision in the FHP model. On the *right part of the figure*, the two possible outcomes of the collision are shown in *dark* and *light gray*, respectively. They both occur with probability one-half

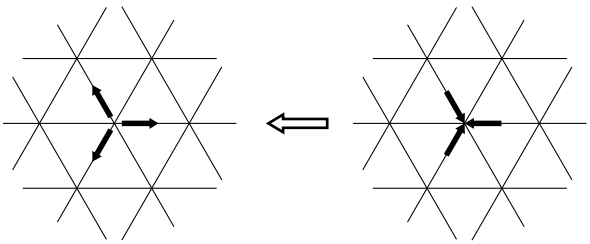

Cellular Automata Modeling of Physical Systems, Figure 15
The three-body collision in the FHP model

still a zero momentum configuration with two particles. As shown in Fig. 14, the deflection can occur to the right or to the left, indifferently. For symmetry reasons, the two possibilities are chosen randomly, with equal probability.

Another type of collision is considered: when exactly three particles collide with an angle of 120 degrees between each other, they bounce back (so that the momentum after collision is zero, as it was before collision). Figure 15 illustrates this rule.

For the simplest case we are considering here, all interactions come from the two collision processes described above. For all other configurations (i. e. those which are not obtained by rotations of the situations given in Figs. 14 and 15) no collision occurs and the particles go through as if they were transparent to each other.

Both two- and three-body collisions are necessary to avoid extra conservation laws. The two-particle collision removes a pair of particles with a zero total momentum and moves it to another lattice direction. Therefore, it conserves momentum along each line of the lattice. On the other hand, three-body interactions deflect particles by 180 degrees and cause the net momentum of each lattice line to change. However, three-body collisions conserve the number of particles within each lattice line.

The FHP model has played a central role in computational physics because it can be shown (see for instance [10]) that the density ρ, defined as the average number of particles at a given lattice site and $\mathbf{u}$ the average velocity of these particles, obey Navier–Stokes equation

$$\partial_t \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} = -\frac{1}{\rho}\nabla p + \nu \nabla^2 \mathbf{u} \tag{17}$$

where $p = c_s^2 \rho$ is the scalar pressure, with c_s the speed of sound and ν is the kinematic viscosity. Note that here both ν and c_s are quantities that emerge from the FHP dynamics. The speed of sound reflects the lattice topology whereas the viscosity reflects the details of the collision process.

As an illustration, Fig. 16 shows the propagation of a density wave in a FHP model. Figure 17 shows the eddies that form when a FHP-fluid flow against an obstacle.

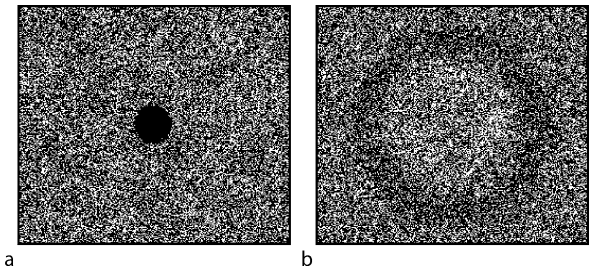

Cellular Automata Modeling of Physical Systems, Figure 16
Development of a sound wave in a FHP gas, due to an over particle concentration in the middle of the system

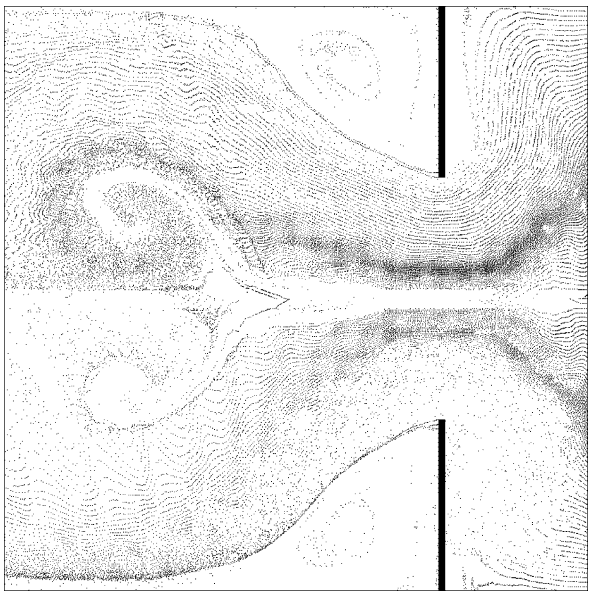

Cellular Automata Modeling of Physical Systems, Figure 17
Flow pattern from a simulation of a FHP model

Cellular Automata Modeling of Physical Systems, Figure 18
A CA snow transport and deposition model

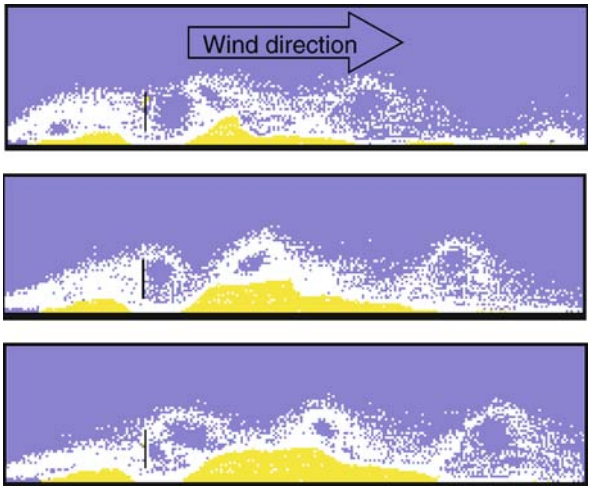

Cellular Automata Modeling of Physical Systems, Figure 19
A lattice Boltzmann snow transport and deposition model. The three panels show the time evolution of the snow deposit (*yellow*) past a fence. Airborne snowflakes are shown as *white dots*

More complex models can be built by adding new processes on top of a FHP-fluid. For instance, Fig. 18 shows the result of a model of snow transport and deposition by wind. In addition to the wind flow, obtained from a FHP model, snow particles are traveling due to the combined effect of wind and gravity. Upon reaching the ground, they pile up (possible after toppling) so as to form a new boundary condition for the wind. In Fig. 19 an extension of the model (using the lattice Boltzmann approach described in Sect. "Lattice Boltzmann Models") shows how a fence with ground clearance creates a snow deposit.

Lattice Boltzmann Models

Lattice Boltzmann models (LBM) are an extension of the CA-fluid described in the previous section. The main conceptual difference is that in LBM, the CA state is no longer Boolean numbers n_i but real-valued quantity f_i for each lattice directions i. Instead of describing the presence or absence of a particle, the interpretation of f_i is the density distribution function of particles traveling in lattice directions i.

As with LGA (see Eq. (16)), the dynamics of LBM can be expressed in terms of an equation for the f_i's. The fluid quantities such as the density ρ and velocity field $\mathbf{u}$ are obtained by summing the contributions from all lattice directions

$$\rho = \sum_i f_i \qquad \rho\mathbf{u} = \sum_i f_i \mathbf{v_i}$$

where $\mathbf{v_i}$ denotes the possible particle velocities in the lattice.

From a numerical point of view, the advantages of suppressing the Boolean constraint are several: less statistical noise, more numerical accuracy and, importantly, more flexibility to choose the lattice topology, the collision operator and boundary conditions. Thus, for many practical applications, the LBM approach is preferred to the LGA one.

The so-called BGK (or "single-time relaxation") LBM

$$f_i(\mathbf{r} + \Delta_t \mathbf{v_i}, t + \Delta_t) = f_i(\mathbf{r}, t) + \frac{1}{\tau}\left(f_i^{\mathrm{eq}}(\rho, \mathbf{u}) - f_i\right) \quad (18)$$

where f^{eq} is a given function, has been used extensively in the literature to simulate complex flows. The method is now recognized as a serious competitor of the more traditional approach based on the computer solution of the Navier–Stokes partial differential equation.

It is beyond the scope of this article to discuss the LBM approach in more detail. We refer the reader to several textbooks on this topic [8,10,40,41,49]. In addition to some rather technical aspects, one advantage of the LBM over the Navier–Stokes equation is its extended range of validity when the Knudsen number is not negligible (e. g. in microflows) [2].

Note finally that the LBM approach is also applicable to reaction-diffusion processes [1] or wave equation [10] by simply choosing an appropriate expression for f_i^{eq}.

Diffusion Processes

Diffusive phenomena and reaction processes play an important role in many areas of physics, chemistry, and biology and still constitute an active field of research. Systems in which reactive particles are brought into contact by a diffusion process and transform, often give rise to very complex behaviors. Pattern formation [31,34], is a typical example of such a behavior. CA provide an interesting framework to describe reaction-diffusion phenomena.

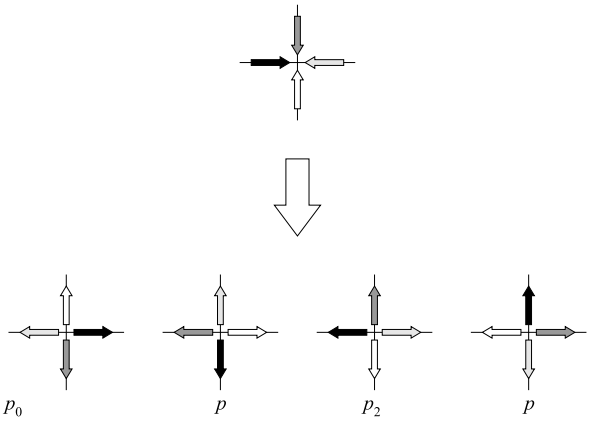

Cellular Automata Modeling of Physical Systems, Figure 20
How the entering particles are deflected at a typical site, as a result of the diffusion rule. The four possible outcomes occur with respective probabilities p_0, p_1, p_2, and p_3. The figure shows four particles, but the mechanism is data-blind and any one of the arrows can be removed when fewer entering particles are present

The HPP rule we discussed in the Sect. "A Simple Gas: The HPP Model" can be easily modified to produce many synchronous random walks. Random walk is well known to be the microscopic origin of a diffusion process.

Thus, instead of experiencing a mass and momentum conserving collision, each particle now selects, at random, a new direction of motion among the possible values permitted by the lattice. Since several particles may enter the same site (up to four, on a two-dimensional square lattice), the random change of directions should be such that there are never two or more particles exiting a site in the same direction. This would otherwise violate the exclusion principle.

The solution is to shuffle the directions of motion or, more precisely, to perform a random permutation of the velocity vectors, independently at each lattice site and each time step. Figure 20 illustrates this probabilistic evolution rule for a 2D square lattice.

It can be shown that the quantity ρ defined as the average number of particle at site $\mathbf{r}$ and time t obeys the diffusion equation [10]

$$\partial_t \rho + \operatorname{div}\left[-D \operatorname{grad} \rho\right] = 0$$

where D is the diffusion constant whose expression is

$$D = \frac{\Delta_r^2}{\Delta_t}\left(\frac{1}{4(p+p_2)} - \frac{1}{4}\right) = \frac{\Delta_r^2}{\Delta_t}\left(\frac{p+p_0}{4[1-(p+p_0)]}\right) \tag{19}$$

where Δ_t and Δ_r are the time step and lattice spacing, respectively. For the one- and three-dimensional cases, a similar approach can be developed [10].

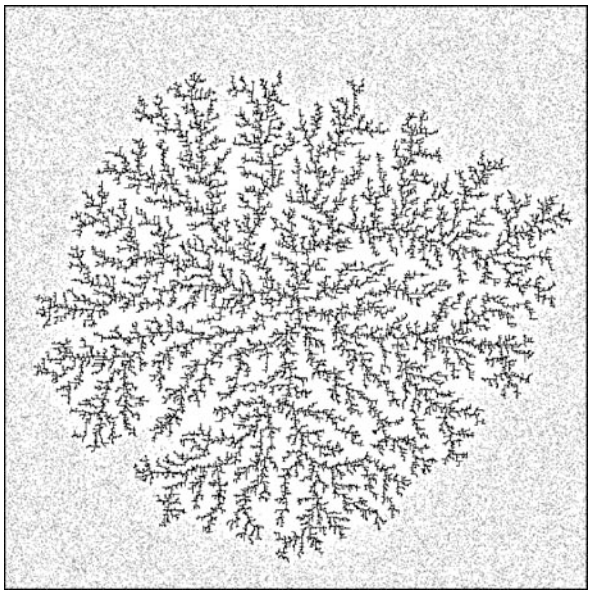

Cellular Automata Modeling of Physical Systems, Figure 21
Two-dimensional cellular automata DLA-like cluster (*black*), obtained with $p_s = 1$, an aggregation threshold of 1 particle and a density of diffusing particle of 0.06 per lattice direction. The gray dots represent the diffusing particles not yet aggregated. The fractal dimension is found to be $d_f = 1.78$

As an example of the use of the present random walk cellular automata rule, we discuss an application to growth processes. In many cases, growth is governed by an aggregation mechanism: like particles stick to each other as they meet and, as a result, form a complicated pattern with a branching structure.

A prototype model of aggregation is the so-called DLA model (diffusion-limited aggregation), introduced by Witten and Sander [46] in the early 1980s. Since its introduction, the DLA model has been investigated in great detail. However, diffusion-limited aggregation is a far from equilibrium process which is not described theoretically by first principle only. Spatial fluctuations that are typical of the DLA growth are difficult to take into account and a numerical approach is necessary to complete the analysis.

DLA-like processes can be readily modeled by our diffusion cellular automata, provided that an appropriate rule is added to take into account the particle-particle aggregation. The first step is to introduce a rest particle to represent the particles of the aggregate. Therefore, in a two-dimensional system, a lattice site can be occupied by up to four diffusing particles, or by one "solid" particle.

Figure 21 shows a two-dimensional DLA-like cluster grown by the cellular automata dynamics. At the beginning of the simulation, one or more rest particles are intro-

Cellular Automata Modeling of Physical Systems, Figure 22
Automata implementation of the $A + B \rightarrow C$ reaction process. The reaction takes place with probability k. The Boolean quantity ν determines in which direction the C particle is moving after its creation

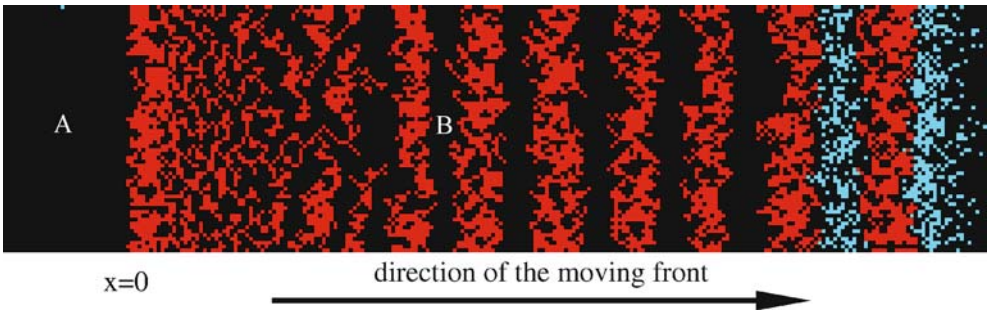

Cellular Automata Modeling of Physical Systems, Figure 23
Example of the formation of Liesegang bands in a cellular automata simulation. The *red bands* correspond to the precipitate which results from the $A + B$ reaction front (in *blue*)

duced in the system to act as aggregation seeds. The rest of the system is filled with particles with average concentration ρ. When a diffusing particle becomes nearest neighbor to a rest particle, it stops and sticks to it by transforming into a rest particle. Since several particles can enter the same site, we may choose to aggregate all of them at once (i. e. a rest particle is actually composed of several moving particles), or to accept the aggregation only when a single particle is present.

In addition to this question, the sticking condition is important. If any diffusing particle always sticks to the DLA cluster, the growth is very fast and can be influenced by the underlying lattice anisotropy. It is therefore more appropriate to stick with some probability p_s.

Reaction-Diffusion Processes

A reaction term can be added on top of the CA diffusion rule. For the sake of illustration let us consider a process such as

$$A + B \xrightarrow{K} C \tag{20}$$

where A, B, and C are different chemical species, all diffusing in the same solvent, and K is the reaction constant. To account for this reaction, one can consider the following mechanism: at the "microscopic" level of the discrete lattice dynamics, all the three species are first governed by a diffusion rule. When an A and a B particle enter the same site at the same time, they disappear and form a C particle.

Of course, there are several ways to select the events that will produce a C when more than one A or one B are simultaneously present at a given site. Also, when Cs already exist at this site, the exclusion principle may prevent the formation of new ones. A simple choice is to have A and B react only when they perform a head-on collision and when no Cs are present in the perpendicular directions. Figure 22 displays such a process.

Other rules can be considered if we want to enhance the reaction (make it more likely) or to deal with more complex situations ($2A + B \rightarrow C$, for instance).

A parameter k can be introduced to tune the reaction rate K by controlling the probability of a reaction taking place. Using an appropriate mathematical tool [10], one can show that the idealized microscopic reaction-diffusion behavior implemented by the CA rule obeys the expected partial differential equation

$$\partial_t \rho_A = D\nabla^2 \rho_A - K\rho_A \rho_B \tag{21}$$

provided k is correctly chosen [10].

As an example of a CA reaction-diffusion model, we show in Fig. 23 the formation of the so-called Liesegang patterns [22].

Liesegang patterns are produced by precipitation and aggregation in the wake of a moving reaction front. Typ-

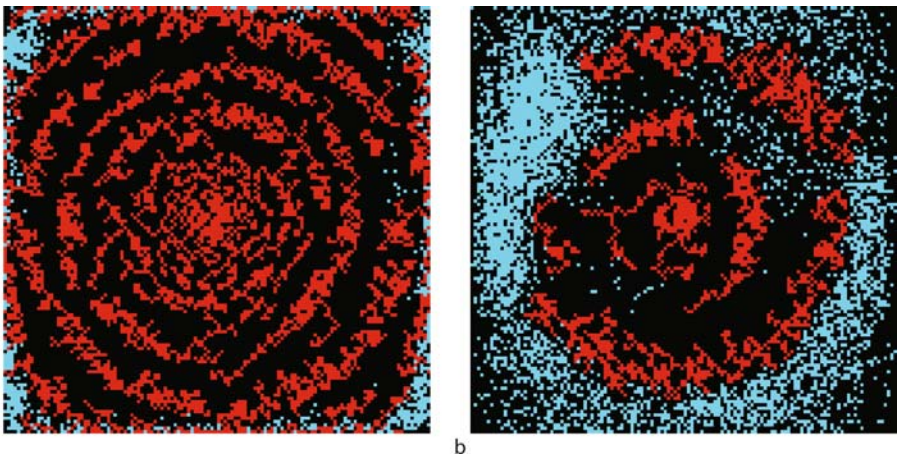

Cellular Automata Modeling of Physical Systems, Figure 24
Formation of Liesegang rings in a cellular automata simulation. The *red spots* correspond to the precipitate created by the $A + B$ reaction front (in *blue*)

ically, they are observed in a test tube containing a gel in which a chemical species B (for example $AgNO_3$) reacts with another species A (for example HCl). At the beginning of the experiment, B is uniformly distributed in the gel with concentration b_0. The other species A, with concentration a_0 is allowed to diffuse into the tube from its open extremity. Provided that the concentration a_0 is larger than b_0, a reaction front propagates in the tube. As this $A + B$ reaction goes on, formation of consecutive bands of precipitate (AgCl in our example) is observed in the tube, as shown in Fig. 23. Although this figure is from a computer simulation [12], it is very close to the picture of a real experiment.

Figure 24 shows the same process but in a different geometry. Species A is added in the middle of a 2D gel and diffuses radially. Rings (a) or spirals (b) result from the interplay between the reaction front and the solidification.

Excitable Media

Excitable media are other examples of reaction processes where unexpected space-time patterns are created. As opposed to the reaction-diffusion models discussed above, diffusion is not considered here explicitly. It is assumed that reaction occurs between nearest neighbor cells, making the transport of species unnecessary. The main focus is on the description of chemical waves propagating in the system much faster and differently than any diffusion process would produce.

An excitable medium is basically characterized by three states [5]: the resting state, the excited state, and the refractory state. The resting state is a stable state of the system. But a resting state can respond to a local perturbation and become excited. Then, the excited state evolves to a refractory state where it no longer influences its neighbors and, finally, returns to the resting state.

A generic behavior of excitable media is to produce chemical waves of various geometries [26,27]. Ring and spiral waves are a typical pattern of excitations. Many chemical systems exhibit an excitable behavior. The Selkov model [38] and the Belousov–Zhabotinsky reaction are examples. Chemical waves play an important role in many biological processes (nervous systems, muscles) since they can mediate the transport of information from one place to another.

The Greenberg–Hasting model is an example of a cellular automata model of an excitable media. This rule, and its generalization, have been extensively studied [17,20].

The implementation we propose here for the Greenberg–Hasting model is the following: the state $\psi(\mathbf{r}, t)$ of site $\mathbf{r}$ at time t takes its value in the set $\{0, 1, 2, \ldots, n-1\}$. The state $\psi = 0$ is the resting state. The states $\psi = 1, \ldots, n/2$ (n is assumed to be even) correspond to excited states. The rest, $\psi = n/2 + 1, \ldots, n-1$ are the refractory states. The cellular automata evolution rule is the following:

1. If $\psi(\mathbf{r}, t)$ is excited or refractory, then $\psi(\mathbf{r}, t+1) = \psi(\mathbf{r}, t) + 1 \mod n$.
2. If $\psi(\mathbf{r}, t) = 0$ (resting state) it remains so, unless there are at least k excited sites in the Moore neighborhood of site $\mathbf{r}$. In this case $\Psi(\mathbf{r}, t+1) = 1$.

The n states play the role of a clock: an excited state evolves

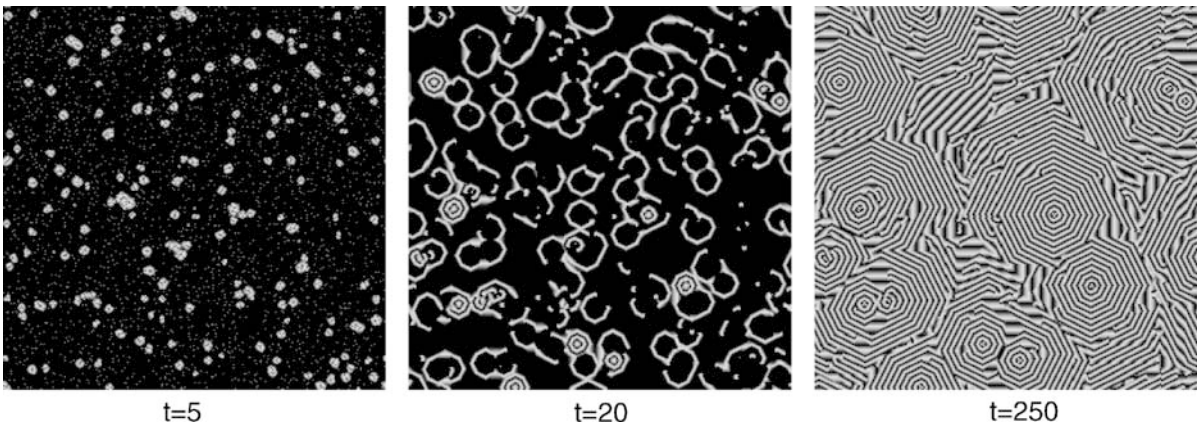

Cellular Automata Modeling of Physical Systems, Figure 25
Excitable medium: evolution of a configuration with 5% of excited states $\Psi = 1$, and 95% of resting states (*black*), for $n = 8$ and $k = 3$

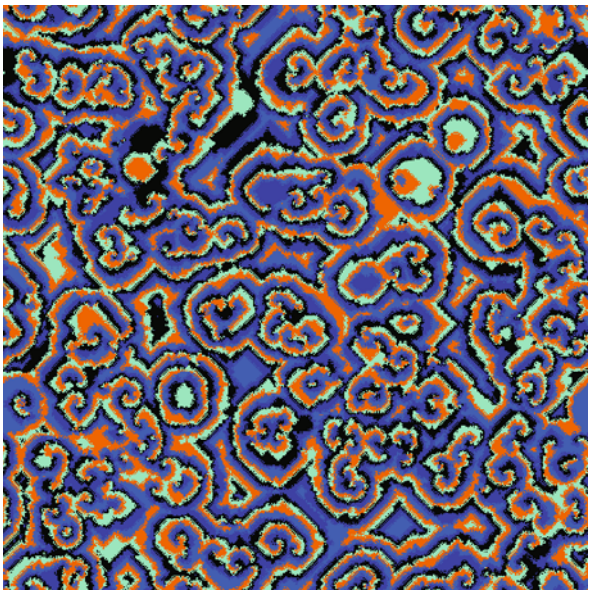

Cellular Automata Modeling of Physical Systems, Figure 26
The tube-worms rule for an excitable media

through the sequence of all possible states until it returns to 0, which corresponds to a stable situation.

The behavior of this rule is quite sensitive to the value of n and the excitation threshold k. Figure 25 shows the evolution of this automaton for a given set of the parameters n and k. The simulation is started with a uniform configuration of resting states, perturbed by some excited sites randomly distributed over the system. Note that if the concentration of perturbation is low enough, excitation dies out rapidly and the system returns to the rest state. Increasing the number of perturbed states leads to the formation of traveling waves and self-sustained oscillations may appear in the form of ring or spiral waves.

The Greenberg–Hasting model has some similarity with the "tube-worms" rule proposed by Toffoli and Margolus [42]. This rule is intended to model the Belousov–Zhabotinsky reaction and is as follows. The state of each site is either 0 (refractory) or 1 (excited) and a local timer (whose value is 3, 2, 1, or 0) controls the refractory period. Each iteration of the rule can be expressed by the following sequence of operations: (i) where the timer is zero, the state is excited; (ii) the timer is decreased by 1 unless it is 0; (iii) a site becomes refractory whenever the timer is equal to 2; (iv) the timer is reset to 3 for the excited sites which have two, or more than four, excited sites in their Moore neighborhood.

Figure 26 shows a simulation of this automaton, starting from a random initial configuration of the timers and the excited states. We observe the formation of spiral pairs of excitations. Note that this rule is very sensitive to small modifications (in particular to the order of operations (i) to (iv)).

Another rule which is also similar to Greenberg–Hasting and Margolus–Toffoli tube-worm models is the so-called forest-fire model. This rule describes the propagation of a fire or, in a different context, may also be used to mimic contagion in the case of an epidemic. Here we describe the case of a forest-fire rule.

The forest-fire rule is a probabilistic CA defined on a d-dimensional cubic lattice. Initially, each site is occupied by either a tree, a burning tree, or is empty. The state of the system is parallel updated according to the following rule: (1) a burning tree becomes an empty site; (2) a green

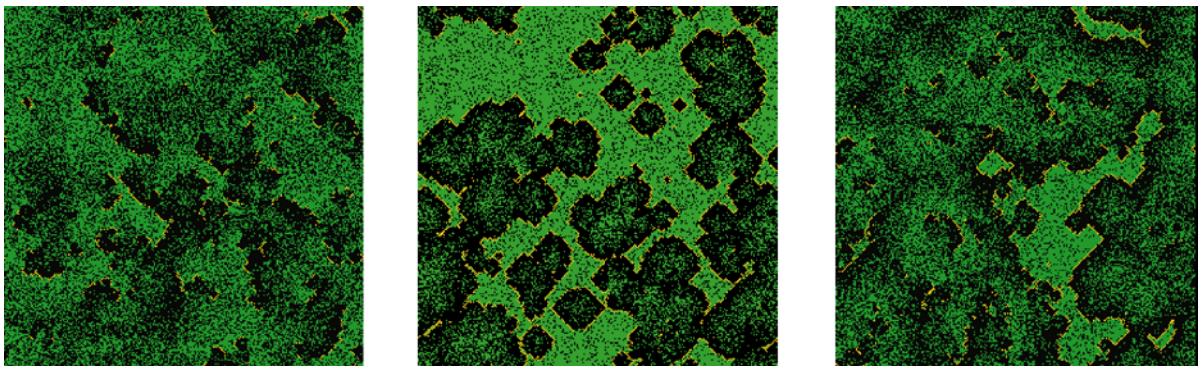

Cellular Automata Modeling of Physical Systems, Figure 27
The forest fire rule: *green sites* correspond to a grown tree, *black pixels* represent burned sites and the *yellow color* indicates a burning tree. The snapshots given here represent three situations after a few hundred iterations. The parameters of the rule are $p = 0.3$ and $f = 6 \times 10^{-5}$

tree becomes a burning tree if at least one of its nearest neighbors is burning; (3) at an empty site, a tree grows with probability p; (4) A tree without a burning neighbor becomes a burning tree with probability f (so as to mimic an effect of lightning).

Figure 27 illustrates the behavior of this rule, in a two-dimensional situation. Provided that the time scales of tree growth and burning down of forest clusters are well separated (i. e. in the limit $f/p \to 0$), this model has self-organized critical states [15]. This means that in the steady state, several physical quantities characterizing the system have a power law behavior.

Surface Reaction Models

Other important reaction models that can be described by a CA are surface-reaction models where nonequilibrium phase transitions can be observed. Nonequilibrium phase transitions are an important topic in physics because no general theory is available to describe such systems and most of the known results are based on numerical simulations.

The so-called Ziff model [54] gives an example of the reaction A–B_2 on a catalyst surface (for example CO–O_2).

The system is out of equilibrium because it is an open system in which material continuously flows in and out. However, after a while, it reaches a stationary state and, depending on some control parameters, may be in different phases.

The basic steps are

- A gas mixture with concentrations X_{B_2} of B_2 and X_A of A sits above a surface on which it can be adsorbed. The surface is divided into elementary cells and each cell can adsorb one atom only.
- The B species can be adsorbed only in the atomic form. A molecule B_2 dissociates into two B atoms only if two adjacent cells are empty. Otherwise the B_2 molecule is rejected.
- If two nearest neighbor cells are occupied by different species they chemically react and the product of the reaction is desorbed. In the example of the CO–O_2 reaction, the desorbed product is a CO_2 molecule.

This final desorption step is necessary for the product to be recovered and for the catalyst to be regenerated. However, the gas above the surface is assumed to be continually replenished by fresh material so that its composition remains constant during the whole evolution.

It is found by *sequential* numerical simulation [54] that a reactive steady state occurs only in a window defined by

$$X_1 < X_A < X_2$$

where $X_1 = 0.389 \pm 0.005$ and $X_2 = 0.525 \pm 0.001$ (provided that $X_{B_2} = 1 - X_A$). This situation is illustrated in Fig. 28, though for the corresponding cellular automata dynamics and $X_{B_2} \neq 1 - X_A$.

Outside this window of parameter, the steady state is a "poisoned" catalyst of pure A (when $X_A > X_2$) or pure B (when $X_A < X_1$). For $X_1 < X_A < X_2$, the coverage fraction varies continuously with X_A and one speaks of a continuous (or second-order) nonequilibrium phase transition. At $X_A = X_2$, the coverage fraction varies discontinuously with X_A and one speaks of a discontinuous (or first-order) nonequilibrium phase transition. Figure 30 displays this behavior.

The asymmetry of behavior at X_1 and X_2 comes from the fact that A and B atoms have a different adsorption rule: two vacant adjacent sites are necessary for B to stick on the surface, whereas one empty site is enough for A.

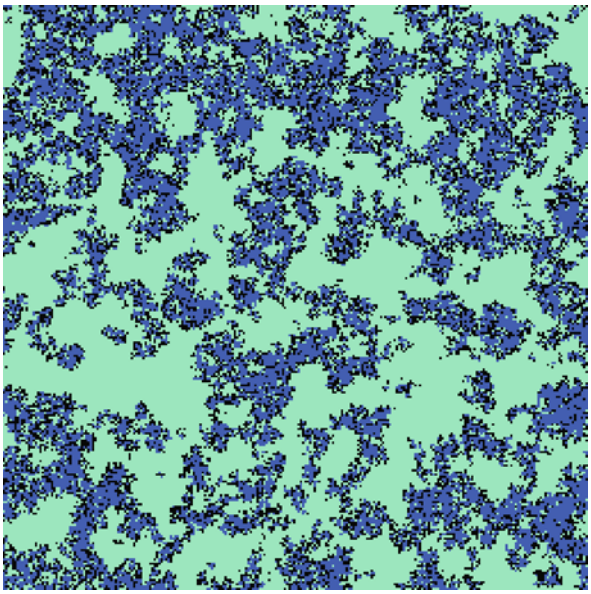

Cellular Automata Modeling of Physical Systems, Figure 28
Typical microscopic configuration in the stationary state of the CA Ziff model, where there is coexistence of the two species over time. The simulation corresponds to the generalized model described by rules R1, R2, R3, and R4 below. The *blue* and *green dots* represent, respectively, the *A* and *B* particles, whereas the empty sites are *black*

In a CA approach the elementary cells of the catalyst are mapped onto the cells of the automaton. In order to model the different processes, each cell j can be in one of four different states, denoted $|\psi_j\rangle = |0\rangle, |A\rangle, |B\rangle$ or $|C\rangle$.

The state $|0\rangle$ corresponds to an empty cell, $|A\rangle$ to a cell occupied by an atom A, and $|B\rangle$ to a cell occupied by an atom B. The state $|C\rangle$ is artificial and represents a precursor state describing the conditional occupation of the cell by an atom B. Conditional means that during the next evolution step of the automaton, $|C\rangle$ will become $|B\rangle$ or $|0\rangle$ depending upon the fact that a nearest neighbor cell is empty and ready to receive the second B atom of the molecule B_2. This conditional state is necessary to describe the dissociation of B_2 molecules on the surface.

The main difficulty when implementing the Ziff model with a fully synchronous updating scheme is to ensure that the correct stoichiometry is obeyed. Indeed, since all atoms take a decision at the same time, the same atom could well take part in a reaction with several different neighbors, unless some care is taken.

The solution to this problem is to add a vector field to every site in the lattice [53], as shown in Fig. 29. A vector field is a collection of arrows, one at each lattice site, that can point in any of the four directions of the lattice. The directions of the arrows at each time step are assigned randomly. Thus, a two-site process is carried out only on those pairs of sites in which the arrows point toward each other (matching nearest-neighbor pairs (MNN)). This concept of *reacting matching pairs* is a general way to partition the parallel computation in local parts.

In the present implementation, the following generalization of the dynamics is included: an empty site remains empty with some probability. One has then two control parameters to play with: X_A and X_{B_2} that are the arrival probability of an A and a B_2 molecule, respectively.

Thus, the time evolution of the CA is given by the following set of rules, fixing the state of the cell j at time $t+1$, $|\psi_j\rangle(t+1)$, as a function of the state of the cell j and its nearest neighbors (von Neumann neighborhood) at time t. Rules R1, R4 describe the adsorption–dissociation mechanism while rules R2, R3 (illustrated in Fig. 29) describe the reaction–desorption process.

R1: If $|\psi_j\rangle(t) = |0\rangle$ then

$$|\psi_j\rangle(t+1) = \begin{cases} |A\rangle & \text{with probability } X_A \\ |C\rangle & \text{with probability } X_{B_2} \\ |0\rangle & \text{with probability } 1 - X_A - X_{B_2} \end{cases} \tag{22}$$

R2: If $|\psi_j\rangle(t) = |A\rangle$ then

$$|\psi_j\rangle(t+1) = \begin{cases} |0\rangle & \text{if the MNN of } j \\ & \text{was in the state } |B\rangle \text{ at time } t \\ |A\rangle & \text{otherwise} \end{cases} \tag{23}$$

R3: If $|\psi_j\rangle(t) = |B\rangle$ then

$$|\psi_j\rangle(t+1) = \begin{cases} |0\rangle & \text{if the MNN of } j \\ & \text{was in the state } |A\rangle \text{ at time } t \\ |B\rangle & \text{otherwise} \end{cases} \tag{24}$$

R4: If $|\psi_j\rangle(t) = |C\rangle$ then

$$|\psi_j\rangle(t+1) = \begin{cases} |B\rangle & \text{if MNN is in the state } |C\rangle \\ & \text{at time } t \\ |0\rangle & \text{otherwise} \end{cases} \tag{25}$$

Figure 28 shows typical stationary configurations obtained with a cellular automata version of the Ziff model. At time $t = 0$, all the cells are empty and a randomly prepared mixture of gases with fixed concentrations X_A and X_{B_2} sits on top of the surface. The rules are iterated until

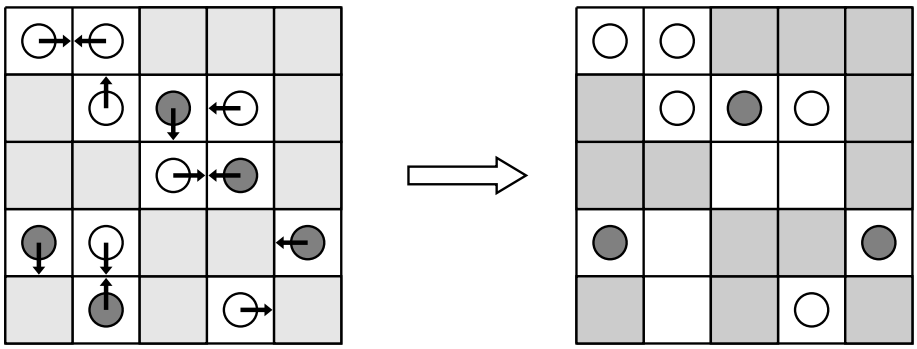

Cellular Automata Modeling of Physical Systems, Figure 29
Illustration of rules R2 and R3. The *arrows* select which neighbor is considered for a reaction. *Dark* and *white particles* represent the *A* and *B* species, respectively. The *shaded region* corresponds to cells that are not relevant to the present discussion such as, for instance, cells occupied by the intermediate *C* species

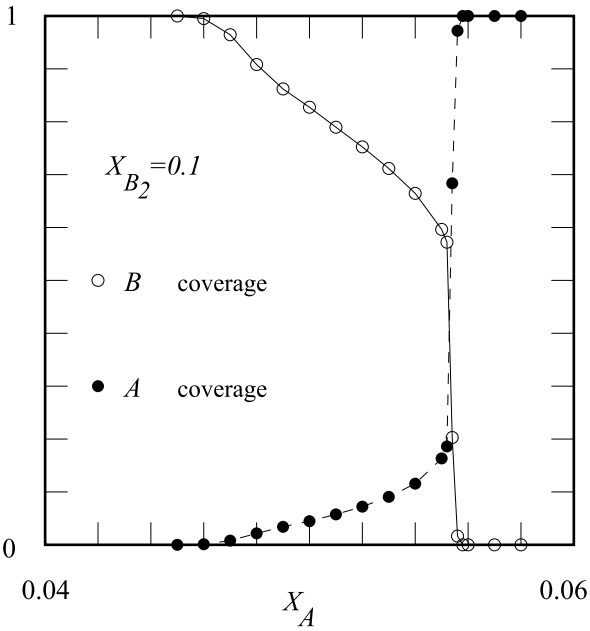

Cellular Automata Modeling of Physical Systems, Figure 30
Stationary state phase diagram corresponding to the CA Ziff model

a stationary state is reached. The stationary state is a state for which the mean coverage fractions X_A^a and X_B^a of atoms of type A or B does not change in time, although microscopically the configurations of the surface changes.

The phase diagram obtained for this generalized CA Ziff model is given in Fig. 30, with the value $X_{B_2} = 0.1$. This phase diagram is topologically similar to the sequential updating case (with $X_{B_2} = 1 - X_A$) since we observe a first and a second order transition surrounding a region of coexistence of both species. However, the locations of the critical points are different, illustrating the nonuniversal character of these quantities.

Future Directions

Cellular Automata are clearly an active field of research. CA are easy to implement on personal computers or on large-scale parallel machines. They are very well suited to discuss many complex systems and dynamical phenomena with space inhomogeneities. Thus, CA provide a very interesting framework for modeling and simulating various processes, for a wide range of application domains where space and time are playing a central role. One can expect that such developments will continue in the future.

In addition, to provide numerical predictions, CA also offer a very intuitive and efficient language to describe various processes or systems. In a period where interdisciplinary applications are very important, it is crucial to be able to communicate between different scientific communities and to develop models that incorporate the knowledge of different fields. Partial differential equations are often an obstacle for several researchers outside Mathematics or Physics. CA then provide a different tool to describe new systems, for instance in ecology, social behavior or in many fields for which a standard mathematical framework is still missing.

The concept of CA can certainly be extended to correspond better to the future scientific challenges. The Lattice Boltzmann method is a well-known example where the Boolean nature of the CA state has been removed. As a result, powerful flow solvers have been obtained.

The nature of the cellular space can also be extended. Graphs are more and more studied for their ability to describe many organizations, whether social, economical, or biological. Therefore, a natural evolution of CA is to be

able to include irregular and/or dynamic lattices of cells, so as to capture the behaviors that come from specificities of the space topology.

Another important direction in computational science are the so-called multiscale and multiscience simulations, aiming at coupling several processes spanning very different spatial or temporal scales. Such a combination of processes is for instance quite common in the simulation of biomedical problems. Recently, the concept of Complex Automaton [23] was proposed as a way to integrate in one single simulation many different components, each described by a different CA.

Bibliography

Primary Literature

1. Alemani D, Chopard B, Buffle J, Galceran J (2006) Two grid refinement methods in the Lattice Boltzmann framework for reaction-diffusion processes. Phys Chem Chem Phys 8:35
2. Ansumali S, Karlin I, Arcidiacono S, Abbas A, Prasianakis N (2007) Hydrodynamics beyond Navier–Stokes: Exact solution to the lattice boltzmann hierarchy. Phys Rev Lett 98:124502
3. Banks E (1971) Information processing and transmission in cellular automata. Tech rep., MIT, MAC TR-81
4. Boon JP (ed) (1992) Advanced Research Workshop on Lattice Gas Automata Theory, Implementations, and Simulation. J Stat Phys 68(3/4):347–672
5. Boon JP, Dab D, Kapral R, Lawniczak A (1996) Lattice gas automata for reactive systems. Phys Rep 273:55–148
6. Burks A (1970) Von Neumann's self-reproducing automata. In: Burks A (ed) Essays on Cellular Automata. University of Illinois Press, Chicago, pp 3–64
7. Burstedde C, Klauck K, Schadschneider A, Zittartz J (2001) Simulation of pedestrian dynamics using a two-dimensional cellular automaton. Physica A 295:506–525
8. Chen S, Doolen G (1998) Lattice Boltzmann methods for fluid flows. Annu Rev Fluid Mech 30:329
9. Chopard B, Droz M (1987) Cellular automata approach to non equilibrium phase transitions in a surface reaction model: static and dynamic properties. J Phys A 21:205
10. Chopard B, Droz M (1998) Cellular Automata Modeling of Physical Systems. Cambridge University Press, Cambridge
11. Chopard B, Dupuis A (2003) Cellular automata simulations of traffic: a model for the city of geneva. Netw Spat Econ 3:9–21
12. Chopard B, Luthi P, Droz M (1994) Reaction-diffusion cellular automata model for the formation of Liesegang patterns. Phys Rev Lett 72(9):1384–1387
13. Chopard B, Luthi PO, Queloz PA (1996) Cellular automata model of car traffic in two-dimensional street networks. J Phys A 29:2325–2336
14. Doolen G (ed) (1990) Lattice Gas Method for Partial Differential Equations. Addison-Wesley, Redwood City
15. Drossel B, Schwabl F (1992) Self-organized critical forest-fire model. Phys Rev Lett 69:1629
16. Farmer D, Toffoli T, Wolfram S (eds) (1984) Cellular Automata. Proceedings of an Interdisciplinary Workshop, Los Alamos. Physica D, vol 10. North-Holland, Amsterdam
17. Fisch R, Gravner J, Griffeath D (1991) Threshold-range scaling of excitable cellular automata. Stat Comput 1:23
18. Frisch U, Hasslacher B, Pomeau Y (1986) Lattice-gas automata for the navier–stokes equation. Phys Rev Lett 56:1505
19. Gardner M (1970) The fantastic combinations of john conway's new solitaire game life. Sci Am 220(4):120
20. Gravner J, Griffeath D (1993) Threshold grouse dynamics. Trans Amer Math Soc 340:837
21. Gunton J, Droz M (1983) Introduction to the Theory of Metastable and Unstable States. Springer, Berlin
22. Henisch H K (1988) Crystals in Gels and Liesegang Rings. Cambridge University Press, Cambridge
23. Hoekstra A, Lorenz E, Falcone JL, Chopard B (2007) Towards a complex automata framework for multi-scale modeling: Formalism and the scale separation map. In: Shi Y et al (ed) Computational Sciences ICCS 2007. LNCS, vol 4487. Springer, Berlin, pp 922–939
24. Kanai M, Nishinari K, Tokihiro T (2005) Stochastic optimal velocity model and its long-lived metastability. Phys Rev E 72:035102(R)
25. Kanai M, Nishinari K, Tokihiro T (2006) Stochastic cellular automaton model for traffic flow. In: Yacoubi SE, Chopard B, Bandini S (eds) Cellular Automata: 7th ACRI conference. LNCS, vol 4173. Springer, Berlin, pp 538–547
26. Kapral R, Showalter K (eds) (1995) Chemical Waves and Patterns. Kluwer, Dordrecht
27. Keener J, Tyson J (1992) The dynamics of scroll waves in excitable media. SIAM Rev. 34:1–39
28. Luthi P O, Preiss A, Ramsden JJ, Chopard B (1998) A cellular automaton model for neurogenesis in drosophila. Physica D 118:151–160
29. Marconi S, Chopard B (2002) A multiparticle lattice gas automata for a crowd. In: Bardini S et al (ed) Proceedings of ACRI 2002 Geneva, Oct, 2002. Lecture notes in computer science, vol 2493. Springer, Berlin, p 230
30. Marconi S, Chopard B (2006) Discrete physics, cellular automata and cryptography. In: Yacoubi SE, Chopard B, Bandini S (eds) Cellular Automata: 7th ACRI conference. LNCS, vol 4173. Springer, Berlin, pp 617–626
31. Muray J (1990) Mathematical Biology. Springer, Berlin
32. Nagel K, Herrmann H (1993) Deterministic models for traffic jams. Physica A 199:254
33. Nagel K, Schreckenberg M (1992) Cellular automaton model for freeway traffic. J Physique I 2:2221
34. Pearson JE (1993) Complex patterns in a simple system. Science 261:189–192
35. Rothman D, Zaleski S (1997) Lattice-Gas Cellular Automata: Simple Models of Complex Hydrodynamics. Collection Aléa. Cambridge University Press, Cambridge
36. Schadschneider A, Schreckenberg M (1993) Cellular automaton models and traffic flow. J Phys A 26:L679
37. Schreckenberg M, Schadschneider A, Nagel K, Ito N (1995) Discrete stochastic models for traffic flow. Phys. Rev. E 51:2939
38. Selkov E (1968) Self-oscillation in glycolysis: A simple kinetic model. Eur J Biochem 4:79
39. Shannon C (1949) Communication theory of secrecy systems. Bell Syst Tech J 28:656–715
40. Succi S (2001) The Lattice Boltzmann Equation, For Fluid Dynamics and Beyond. Oxford University Press, Oxford
41. Sukop M, Thorne D (2005) Lattice Boltzmann Modeling: an Introduction for Geoscientists and Engineers. Springer, Berlin

42. Toffoli T, Margolus N (1987) Cellular Automata Machines: a New Environment for Modeling. MIT Press, Cambridge
43. Tolman S, Meakin P (1989) Off-lattice and hypercubic-lattice models for diffusion-limited aggregation in dimension 2–8. Phys Rev A 40:428–37
44. Vichniac G (1984) Simulating physics with cellular automata. Physica D 10:96–115
45. Vicsek T (1989) Fractal Growth Phenomena. World Scientific, Singapore
46. Witten T, Sander L (1983) Diffusion-limited aggregation. Phys Rev B 27:5686
47. Wolf D, Schreckenberg M et al (eds) Traffic and Granular Flow '97. Springer, Singapore
48. Wolf D, Schreckenberg M, Bachem A (eds) (1996) Traffic and Granular Flow. World Scientific, Singapore
49. Wolf-Gladrow D A (2000) Lattice-Gas Cellular Automata and Lattice Boltzmann Models: an Introduction. Lecture Notes in Mathematics, vol 1725. Springer, Berlin
50. Wolfram S (1986) Theory and Application of Cellular Automata. World Scientific, Singapore
51. Wolfram S (1994) Cellular Automata and Complexity. Addison-Wesley, Reading
52. Yukawa S, Kikuchi M, Tadaki S (1994) Dynamical phase transition in one-dimensional traffic flow model with blockage. J Phys Soc Jpn 63(10):3609–3618
53. Ziff R, Fichthorn K, Gulari E (1991) Cellular automaton version of the ab_2 reaction model obeying proper stoichiometry. J Phys. A 24:3727
54. Ziff R, Gulari E, Barshad Y (1986) Kinetic phase transitions in an irreversible surface-reaction model. Phys Rev Lett 56:2553

Books and Reviews

Chopard B, Droz M (1998) Cellular AutomataModeling of Physical Systems. Cambridge University Press, Cambridge
Deutsch A, Dormann S (2005) Cellular Automaton Modeling of Biological Pattern Formation. Birkhäuser, Basel
Gaylord RJ, Nishidate K (1996) Modeling Nature with Cellular Automata using Mathematica. Springer, Berlin
Ilachinski A (2001) Cellular Automata: a discrete universe. World Scientific, Singapore
Rivet JP, Boon JP (2001) Lattice Gas Hydrodynamics. Cambridge University Press, Cambridge
Rothman D, Zaleski S (1994) Lattice-gas models of phase separation: interface, phase transition and multiphase flows. Rev Mod Phys 66:1417–1479
Weimar JR (1998) Simulation with Cellular Automata. Logos, Berlin
Wolfram S (2002) A new kind of science. Wolfram Sciences, Champaign

Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations

Carter Bays
Department of Computer Science and Engineering, University of South Carolina, Columbia, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Two Dimensional Cellular Automata in the Triangular Grid
The Hexagonal Grid
The Pentagonal Grid
Programming Tips
Future Directions
Bibliography

Typically, cellular automata ("CA") are defined in Cartesian space (e. g. a square grid). Here we explore characteristics of CA in triangular and other non-cartesian grids. Methods for programming CA for these non-cartesian grids are briefly discussed.

Glossary

Cellular automaton (CA) a structure comprising a grid with individual cells that can have two or more states; these cells evolve in discrete time units and are governed by a rule, which usually involves neighbors of each cell.

Game of life a particular cellular automaton discovered by John Conway in 1968.

Neighbor a neighbor of cell "x" is typically a cell that is in close proximity to (frequently touching) cell "x".

Oscillator a periodic shape within a specific cellular automaton rule.

Glider a translating oscillator that moves across the grid of a CA.

Generation the discrete time unit which depicts the evolution of a cellular automaton.

Rule determines how each individual cell within a cellular automaton evolves.

Definition of the Subject

A tessellation or tiling is composed of a specific shape that is repeated endlessly in a plane, with no gaps or overlaps. Examples of simple tessellations are the square grid, the triangular grid (a plane completely covered by identical triangles), etc. Hereafter, we shall also use "grid" when refering to tessellations.

Cellular automata (CA) can be explained most effectively with an example. Start with an infinite grid of squares; each square represents a cell, which is either "alive" or "dead". Time progresses in discrete units called "generations"; at every generation we evaluate simultane-

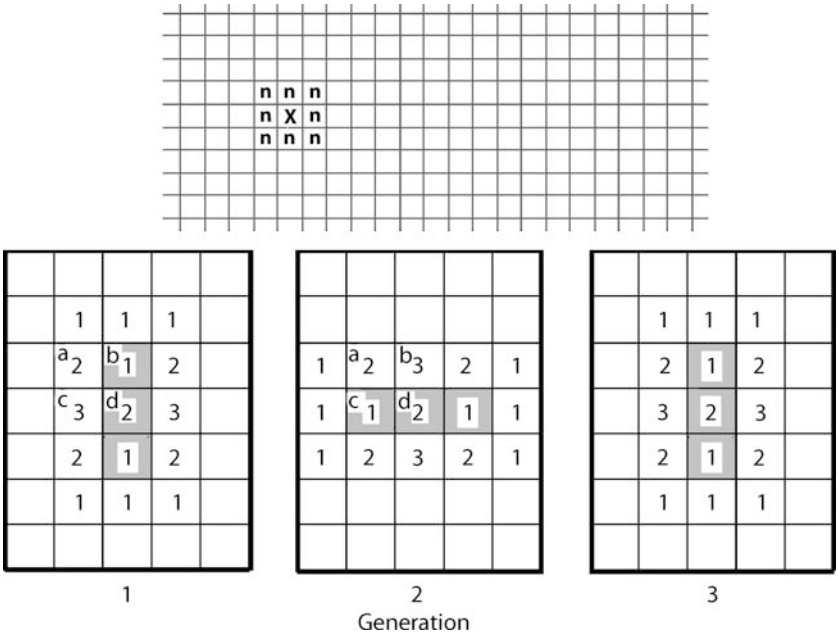

Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 1

***Top*: Each cell in a grid has 8 "neighbors". The cells containing "n" are neighbors of the cell containing the "X". Any cell in the grid can be either "dead" or "alive". *Bottom*: Here we have outlined a specific area of what is presumably a much larger grid. At the *left* we have installed an initial shape. *Shaded* cells are alive; all others are dead. The number within each cell gives the quantity of live neighbors for that cell. (Cells containing no numbers have zero live neighbors.) Depicted are three generations, starting with the configuration at generation 1. Generations 2 then 3 show the result when we apply the following cellular automata rule: "Live cells with exactly 2 or 3 live neighbors remain alive (otherwise they die); dead cells with exactly 3 live neighbors come to life (otherwise they remain dead)". Let us now evaluate the transition from generation 1 to generation 2. In our diagram, cell "a" is dead. Since it does not have exactly 3 live neighbors, it remains dead. Cell "b" is alive, but it needs exactly 2 or 3 live neighbors to remain alive; since it only has 1, it dies. Cell "c" is dead; since it has exactly 3 live neighbors, it comes to life. And cell "d" has 2 live neighbors; hence it will remain alive. And so on. Notice that the form repeats every two generations. Such forms are called oscillators**

ously the fate for each cell at the next generation by examining neighboring cells (called "neighbors") – in this case, we shall consider as neighbors any cell touching the candidate cell (eight neighbors in all). This is sometimes called the Moore neighborhood. We apply a "rule" to determine the next generation status of our candidate cell. For example, our rule might state, (a) "If our candidate cell is currently alive, then it will remain alive next generation if it touches either two or three live neighbors, otherwise it dies", and (b) "If our candidate cell is not alive then it will come to life next generation if and only if it is touching exactly three live neighbors."

Figure 1 illustrates a simple configuration to which this CA rule has been applied. Notice that this particular object repeats itself indefinitely. Such an object is called an "oscillator"; this particular oscillator has a "period" of two. Other configurations can have much larger periods, or can behave in a more chaotic fashion. Motionless patterns can be thought of as oscillators whose period is one.

Needless to say, there are a huge number of rules that can be applied, and each rule will cause a distinct action. The rule given above – the most famous cellular automaton of all – specifies the "Game of Life", discovered by John Horton Conway in 1968. Game of Life (GL) rules must satisfy the following informal criteria.

1. All neighbors must be touching the candidate cell and all are treated the same.
2. There must exist at least one translating oscillator (called a "glider").
3. Random configurations must eventually stabilize into zero or more oscillators..

For a more formal description of GL rule requirements see [1]. It is important to note that CA can be represented in one, two, three or higher dimensions, but most work has been done in one or two. Furthermore, neighbors can be defined in many ways; for example we might only consider as neighbors those cells touching the sides of

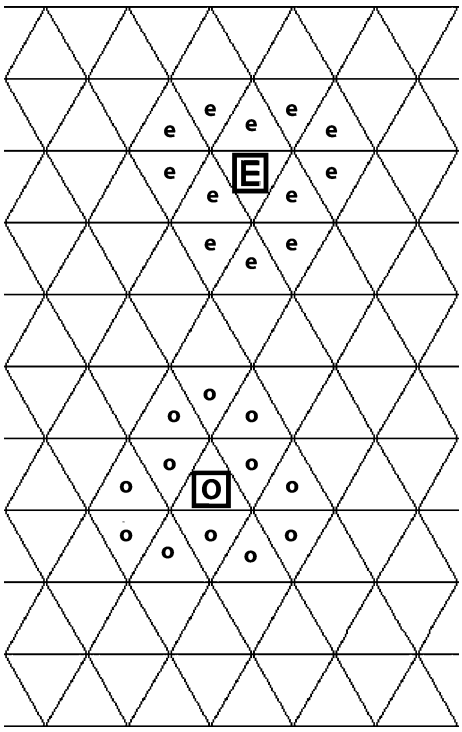

Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 2
The neighborhoods for cells in the triangular grid. Note that the candidate cells can have two orientations – "E" and "O". The neighbors are indicated by "e" and "o" respectively

a candidate cell and not the corners. Or we might expand our neighborhood to include cells within a given distance of a candidate cell. This is typically done for one-dimensional CA.

Some Convenient Notation for Describing CA Rules

We shall write CA rules using the following notation,

$$E_1, E_2, \ldots / F_1, F_2, \ldots$$

where the E_i specify the number of live neighbors required to keep a living cell alive, and the F_i give the number required to bring a non-living cell to life. The E_i and F_i will be listed in ascending order; hence if $i > j$ then $E_i > E_j$ etc. Thus the rule for Conway's Game of Life is written 2,3/3.

We shall also use a convenient shorthand when appropriate: E_i–E_j denotes $E_i, E_{i+1}, \ldots, E_{i+j-i}$ etc. Thus, 2,3,4,5,6/2,3,4 can also be written 2–6/2–4.

Introduction

Almost all CA research in two dimensions has been done using rectangular (Cartesian) coordinates, and hence typically utilizes the square grid. But there is no reason to limit ourselves to this tessellation; the number of different possible grids is almost endless. Here we shall briefly investigate CA behavior in only three – triangular, hexagonal and pentagonal.

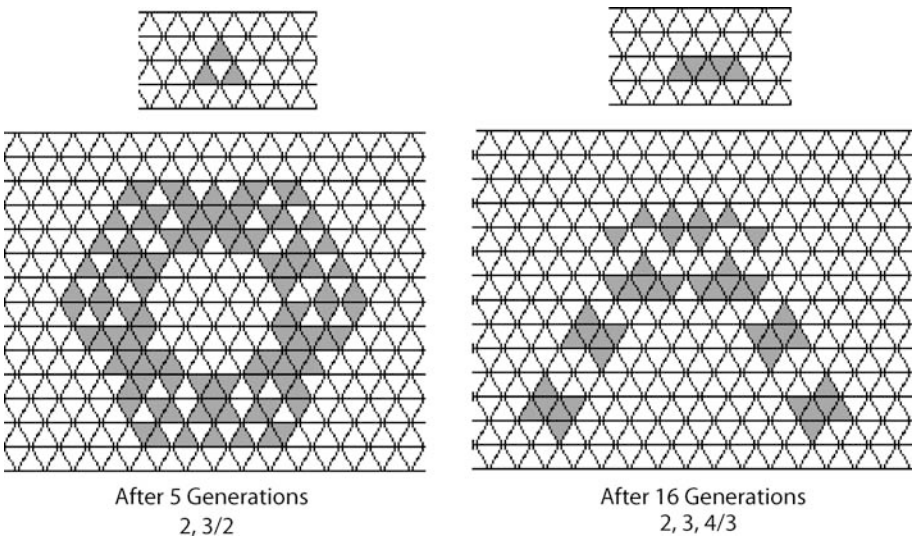

Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 3
Examples of expanding rules. The starting configurations are at the top

Two Dimensional Cellular Automata in the Triangular Grid

Throughout this article we shall consider as neighbors only those cells that touch the candidate cell; hence for a grid composed of triangles, each cell would have 12 neighbors (Fig. 2). These non-cartesian grids for CA have been investigated from time to time; most notably by Preston [3] and Bays [1,2]. Recently work relating to hexagonal CA has appeared on the internet occasionally.

With 12 touching neighbors instead of 8 (as in the square grid), we can write more than 16 million distinct rules, most of which are probably of only marginal interest. Many however exhibit behavior worthy of investigation.

Some rules will generate a continually expanding collection of live cells – we shall call such rules "expanding" or "unstable" rules. Thus 2,3/2; 2,3/3,4; 2,3,4/3 each produce an ever increasing area of live cells – even with extremely small starting configurations (Fig. 3). A few expanding rules "barely" expand; i. e. several generations are required and the initial live configuration must be fairly large in order to observe instability. For example 2,3,6/4,5 can produce unbounded growth, while 2,3,8/4,5 always eventually stabilizes. The fate of configurations under 2,3,7/4,5 is uncertain, but the rule appears to produce unbounded growth. Many rules will ultimately lead to a stable pattern (Fig. 4), or no live cells at all.

For some rules we can start with bounded forms whose innards churn endlessly forever; these rules can, for example, be used to generate random numbers (Fig. 5). Such rules differ somewhat from expanding rules in that all finite patterns are bounded and will not expand indefinitely, but an infinite grid of random live cells will never stabilize.

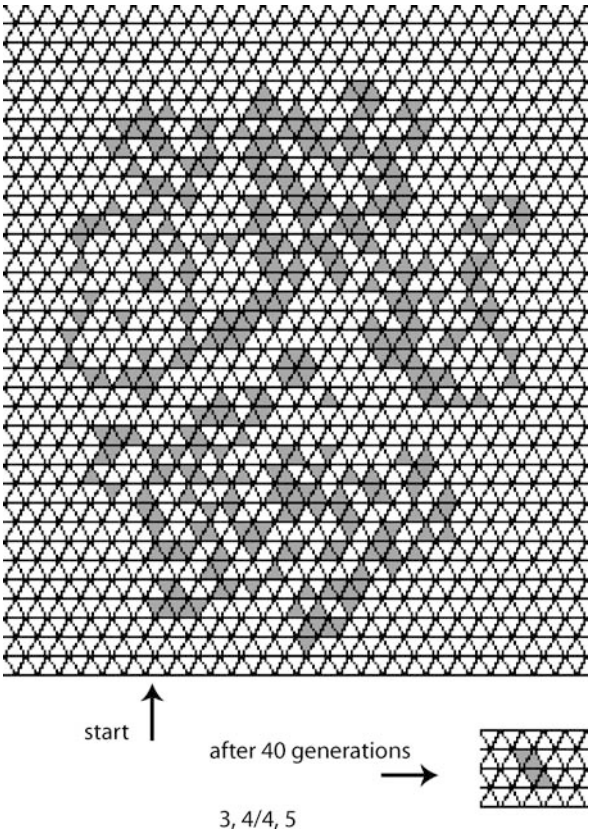

Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 4
An example of a stable rule. The starting random configuration eventually stabilizes into the shape shown at the *lower right*; interestingly this shape happens to be an oscillator with a period of two

Game of Life Rules in the Triangular Grid

As mentioned above, the most famous GL rule is Conway's game, which utilizes a square grid. But GL rules are not limited to squares; quite a few exist in the triangular grid. Among these are 4,5,6/4; 3,4/4,5; 4,5/4,5,6; 2,3/4,5; 3,4/4,5,6; 2/3; 2,4/4,6; 3,5/4; 2,4,6/4,6; 2,7/3; 2,7,8/3. Further information about these and other rules can be found in [1] and [2].

The Hexagonal Grid

The neighborhood for the hexagonal grid is only half the size of that for the triangular grid and is symmetric – each neighbor is identical in the manner of contact with the cell in question. This symmetry can be important for some applications. Unfortunately, the hexagonal grid has a limited number of possible rules – there are only about 4000, many of which are of little interest. For many years past attempts to find a GL rule in the hexagonal grid have failed, although gliders were discovered by defining rules where the spatial relationship between neighbors was a factor [3]. Recently however the GL rule 3/2 was discovered. It supports the glider shown in Fig. 12. Another glider has also turned up; its rule is 3/2,4,5. Unfortunately this rule is not a GL rule, as it will very slowly exhibit unbounded growth, given a sufficiently large starting pattern.

The Pentagonal Grid

Regular pentagons cannot be formed into a grid, but by varying the angles and side lengths, we can create several tessellations from identical convex pentagons. A classification system has been devised, wherein 14 different types of tilings have been identified (Fig. 14). Of these, 12 are

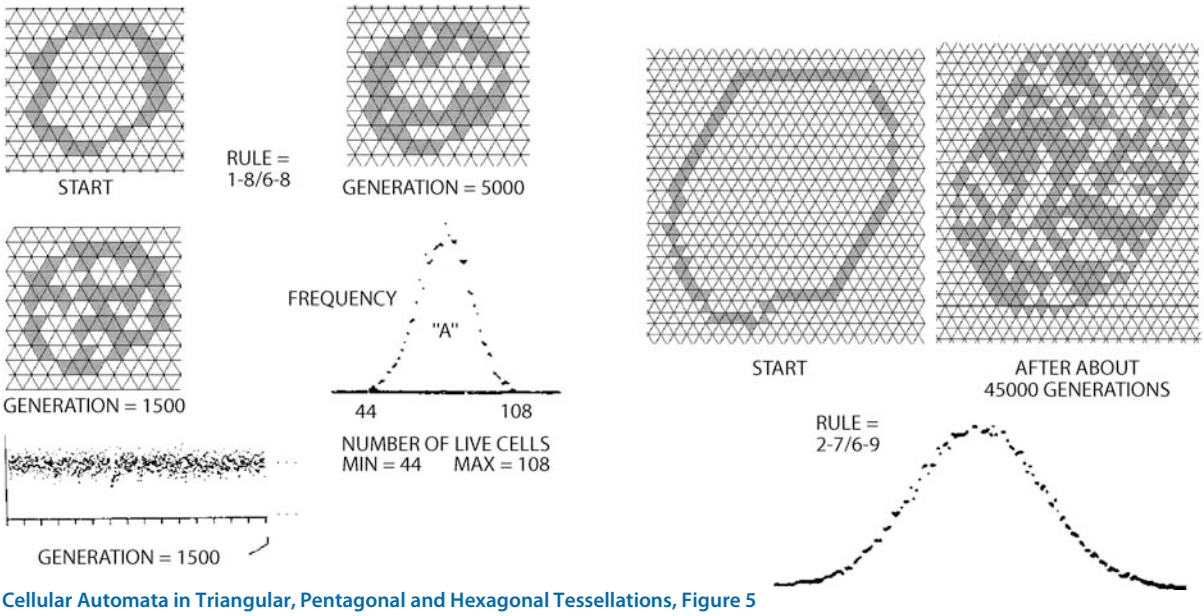

Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 5

Examples of bounded rules that churn endlessly. The total number of live cells can be employed as pseudorandom numbers that approximate a normal distribution. Many candidate rules can be used. Naturally the randomlike patterns eventually repeat, but with a sufficiently large initial shape, the period will be quite large. The plot at the *lower left* gives the number of live cells at each generation. These values exhibit a normal distribution (plot "A"). Note however that there are some gaps. This is because the rule 1-8/6-8 tends to have "clumps" of living (and fairly large "holes" of non-living) cells. Hence, before using this technique for generating random numbers, the candidate rule should be carefully investigated

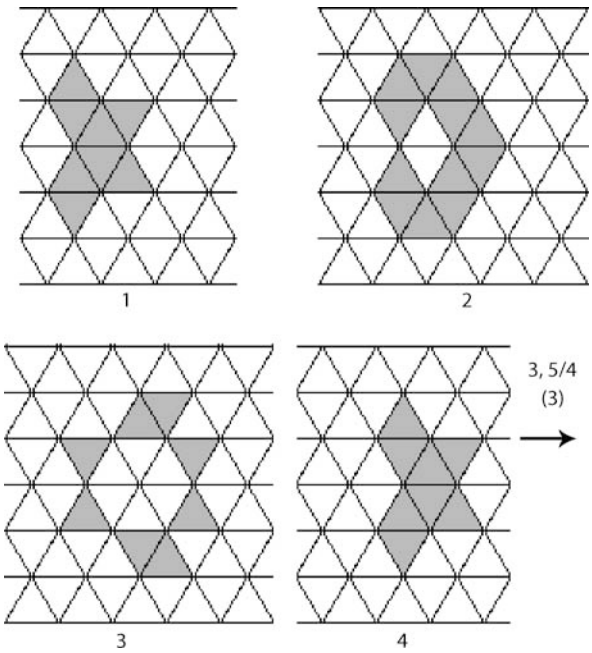

Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 6

A simple glider for the GL rule 3,5/4. It has a period of three (indicated in parentheses) after which it will have moved one cell to the right. Many gliders are not this well behaved, with much longer periods and irregular structure (see next figure)

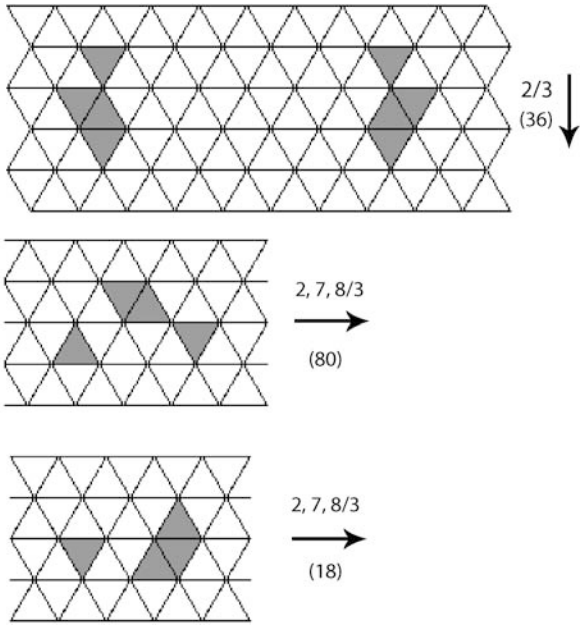

Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 7

Some gliders exhibit rather spectacular evolution. The period 80 2,7,8/3 glider, swells to 60 live cells during its swaggering trip across the grid, and at the 81st generation, will have moved 12 cells to the right. The gliders move in the direction given by the arrows. It should be noted that gliders have also been found for non-GL rules but since these rules are unstable they have not been investigated

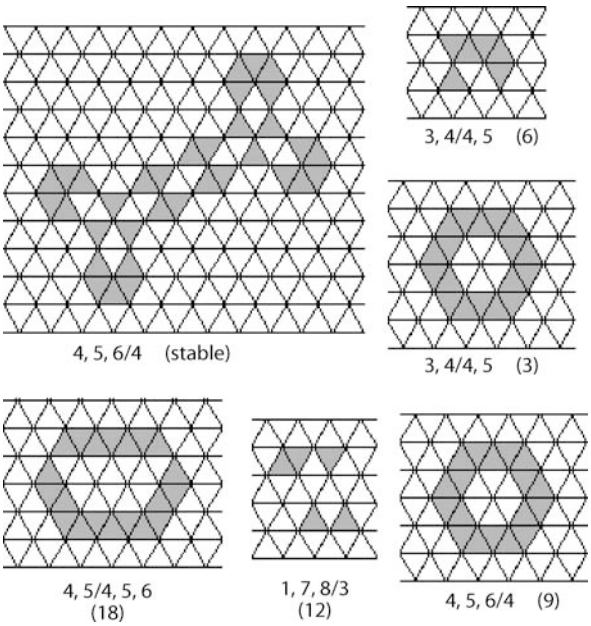

Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 8
Hundreds of oscillators exist for the GL (and other) rules in the triangular grid. A few interesting ones are illustrated here. The stationary 4,5,6/4 form is representative of an infinite number of such objects that can be created for this rule by the careful positioning of live cells. The different oscillators at the lower right happen to share one identical shape. The oscillator at the *upper right* "rotates" clockwise, as does the period 12 oscillator at the bottom. Unfortunately rule 1,7,8/3 is not a GL rule

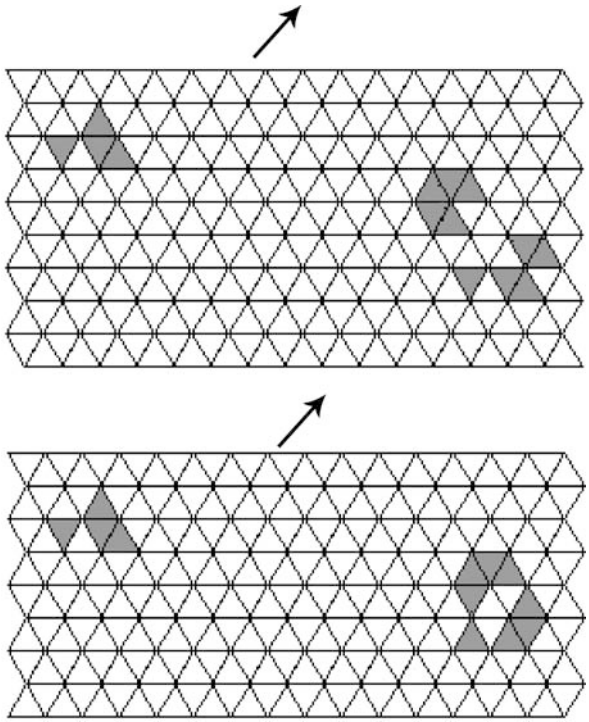

Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 9
The GL rule 2,7,8/3 is of special interest. It is the only known GL rule besides Conway's rule that supports a "glider gun" – a configuration that spews out an endless stream of gliders. In fact, there are probably several such patterns under that rule. Here we illustrate two guns; the top one generates period 18 gliders and the bottom one creates period 80 gliders. These configurations move in the direction shown, sending a stream of gliders out behind them (see next figure)

topologically distinct; these twelve varieties will behave in different ways under CA rules. We have chosen to investigate one of the most pleasing, the "Cairo tiling", so named because of its alleged use in parts of that city. It's appeal derives from the fact that the pentagons are both equilateral and isoseles.

Under the Cairo grid, there are rules that behave in the manner already described for the triangular grid – some rules expand, some stabilize, others contain a bounded, churning mass, etc. Interestingly, a GL rule has been discovered; its glider is depicted in Fig. 15. There is much opportunity for discovery in this and other pentagonal grids, as very little work has been done.

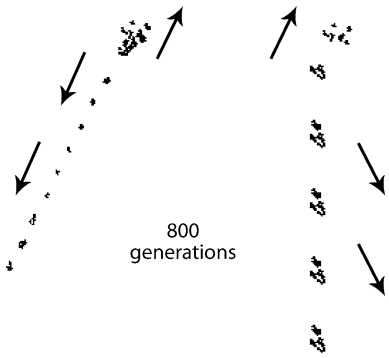

Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 10
After 800 generations, the two guns will have produced the output shown. Motion is in the direction given by the arrows. The gun at the left yields period 18 gliders, one every 80 generations, and the gun at the right produces a period 80 glider every 160 generations

Programming Tips

We can speed up the scan of any grid by storing within each cell its current number of live neighbors along with a tag that indicates whether it is alive or not. Thus when we scan the entire grid for the next generation, we update

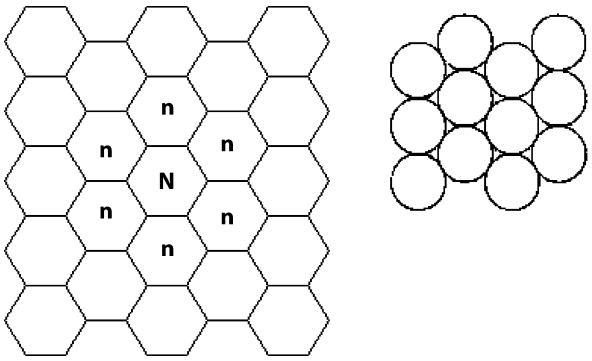

Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 11
The symmetric hexagonal neighborhood. This rather "natural" grid can also be illustrated with *circles* (*upper right*) and, just as the square grid can be expanded to cubes in 3 dimensions, the hexagonal grid lends itself to "densely packed spheres" in three dimensions, where each sphere has exactly 12 touching neighbors

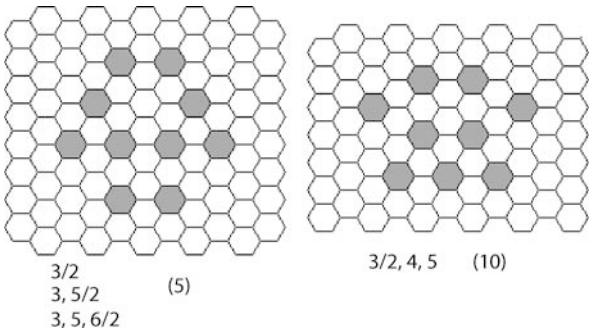

Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 12
At least two gliders have been found. The GL rule 3/2 supports a period 5 glider and the non-GL rule 3/2,4,5 supports a period 10 glider. Note that the 3/2 glider also works for GL rules 3,5/2 and 3,5,6/2

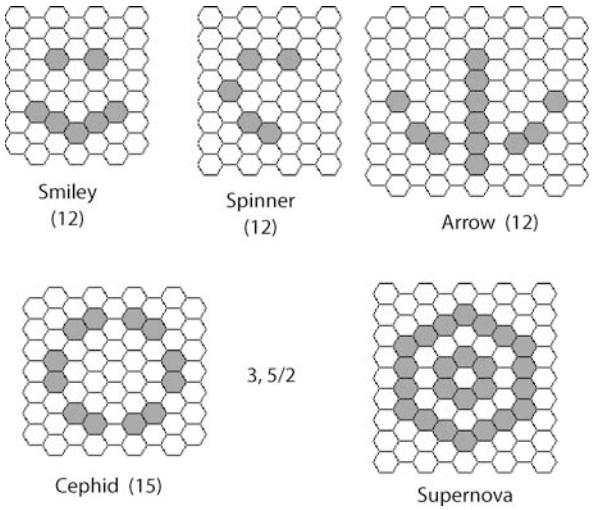

Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 13
Several interesting oscillators have been discovered for GL rule 3/2. They have been given rather whimsical names, a custom dating back to the early days of Conway's rule. After 65 generations the "supernova" pattern leaves a period 3 "neutron star" remnant. These patterns also work under rules 3,5/2 and 3,5,6/2

the status of cells that have changed since last generation (by examining their new neighbor counts) and, for each cell whose status has changed, we fix the neighbor counts for its neighbors; these cells are candidates for updating at the next generation iteration. This method employs two arrays – a "current" array, *A*, and a "next" array, *B*. And, rather than moving *B* back to *A* for the next iteration, we switch between the two; i. e. if array *A* is the array we are examining, then we copy it into array *B*, changing the status and neighbor counts of cells as needed. Array *B* then becomes array *A* for the next generation, etc. This trick allows us to rapidly scan over all non-changed cells. For further speed we can utilize hashing techniques and only store cells whose status is going to change. For this method, the speed of evaluation will depend only upon the number of cells that change between generations, and not the size of the grid, nor the total number of live cells. Furthermore, with a clever plotting algorithm, we can get away with replotting only those changed cells and not the entire grid.

We can program practically any grid or tiling in rectangular (square) coordinates by using templates to locate the neighbor cells as depicted in Fig. 16. The operation of finding the correct neighbors via templates adds a very small amount of time to the overall "next generation" evaluation; hence we would expect calculations on any type of grid to execute almost as fast as on the standard square grid.

Future Directions

The triangular grid yields 12 touching neighbors and hence an ample supply of rules to investigate – many more than the 8 neighbor square grid. The hexagonal grid affords a more natural approach to CA than does the traditional 8 neighbor square grid, since neighbors all touch in the same way. Furthermore, when we expand this grid into three dimensions, we obtain a universe of dense packed spheres, which probably gives the best methodology for emulating 3D applications, as each cell has 12 touching neighbors and all touch in the same way. The fact that

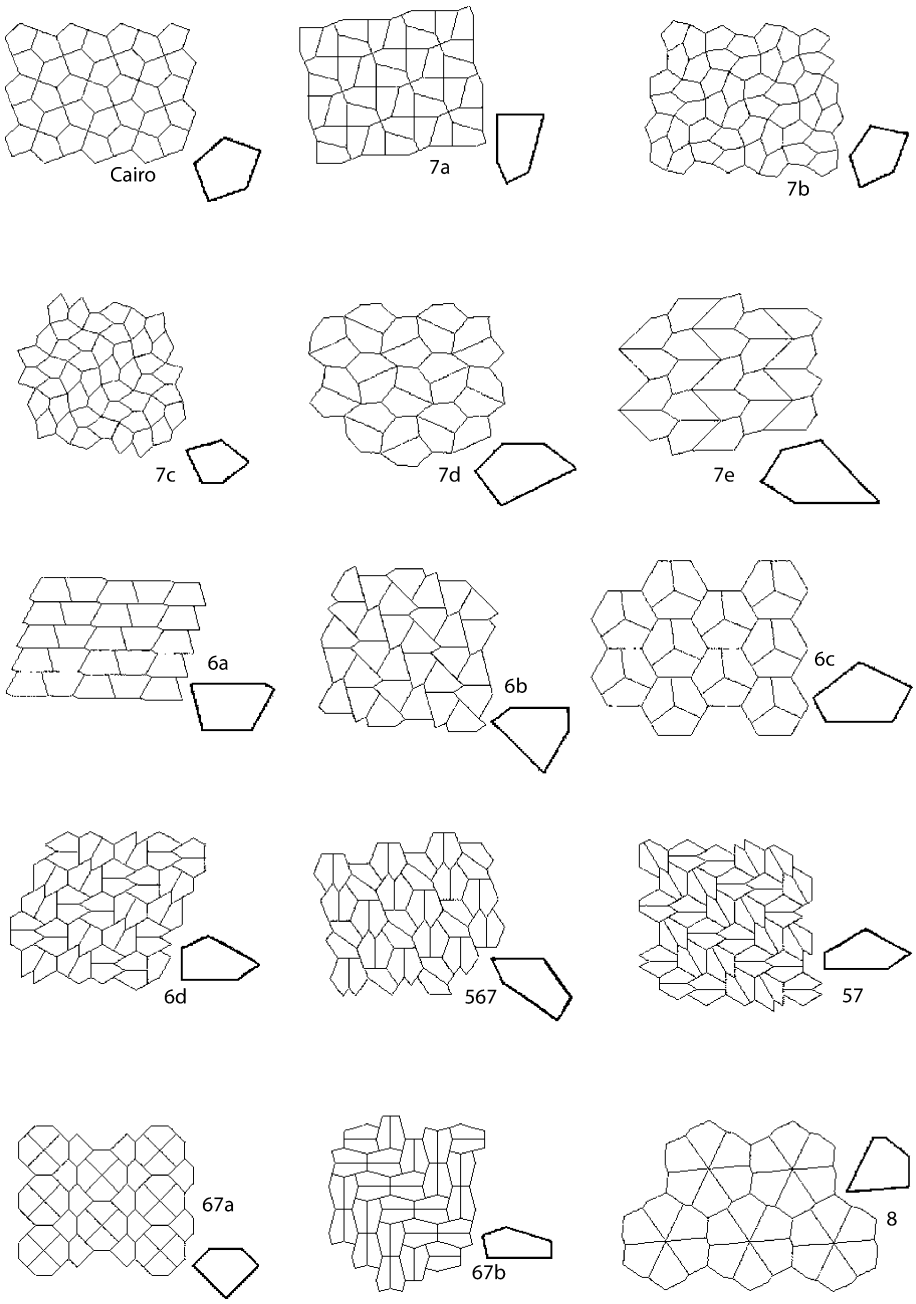

Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 14

The 14 distinct convex pentagonal tilings [see 1,5]. They are based upon certain relationships between the angles and lengths of the sides of the particular pentagon that constitutes the tiling. A sample of the pentagon for that tiling is displayed at the right of each. The tilings have been arranged to depict the number of touching neighbors for each cell. Where more than one number is given, there are some cells with each of those neighbor counts. For example the "67b" tiling is the second tiling where some cells have 6 neighbors and others 7. The Cairo tiling is at the *upper left* and is topologically equivalent to 7a and 7b. Note that 7c and 7d are also topologically equivalent

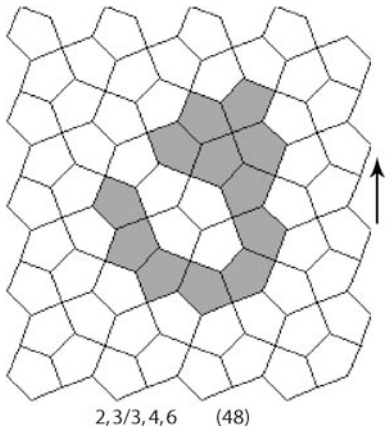

Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 15
The GL rule 2,3/3,4,6 supports the period 48 glider shown. It is asymmetric, though the second half of its period is a mirror image of the first. This characteristic is common amongst many gliders

GL rules have been found in a pentagonal grid undoubtedly means that other such rules can probably be found in many different tessellations – pentagonal and otherwise. The ultimate conclusion is that there is room for much work in the area of non-cartesian CA.

Bibliography

1. Bays C (2005) A Note on the Game of Life in Hexagonal and Pentagonal Tessellations. Complex Syst 15:245–252
2. Bays C (1994) Cellular Automata in the Triangular Tessellation. Complex Syst 8:127–150
3. Preston K Jr, Duff MJB (1984) Modern Cellular Automata. Plenum Press, New York
4. Sugimoto T, Ogawa T (2000) Tiling Problem of Convex Pentagon. Forma 15:75–79
5. Wolfram S (2002) A New Kind of Science. Wolfram Media, Champaign

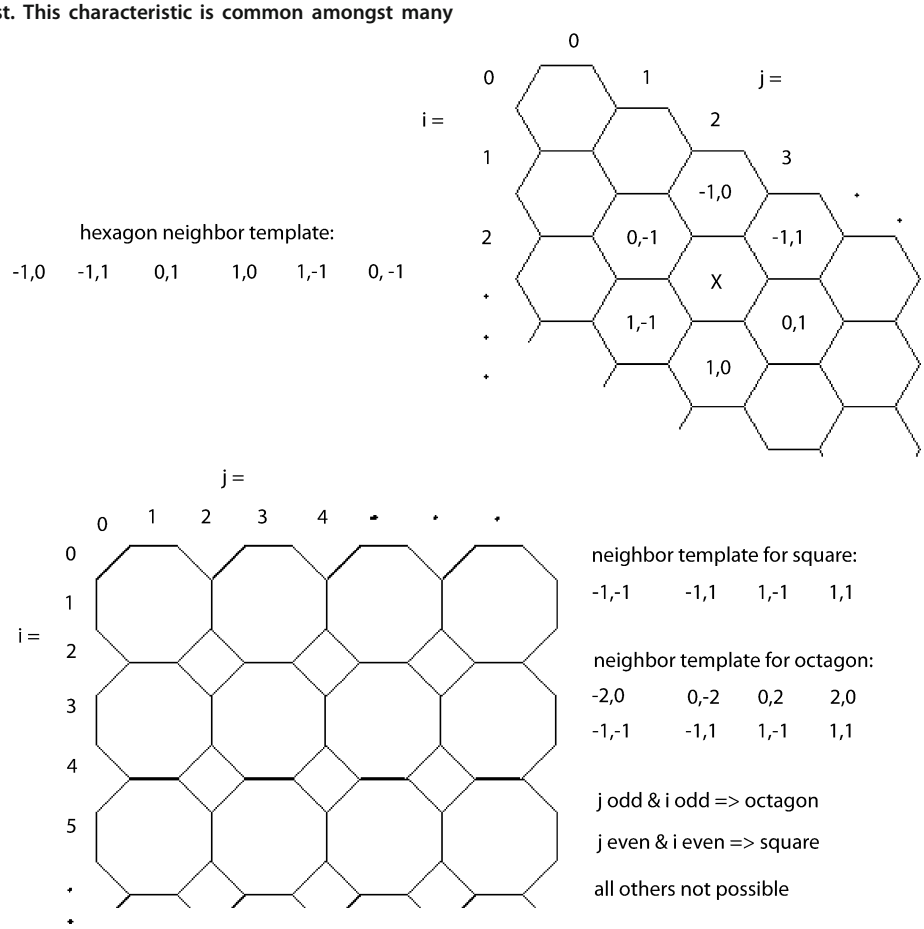

Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations, Figure 16
Templates can be used to simulate any grid with rectangular coordinates. For example, if we are evaluating the neighbors for a hexagonal cell at (i, j) (see "X") they would be found at $(i-1, j)$; $(i-1, j+1)$; $(i, j+1)$; etc. We can even simulate grids made up of different types of polygons. Here, we determine the polygon type by examining the subscripts of the cell in question. Of course, appropriate graphics procedures must be employed in order to view our grid

Cellular Automata, Universality of

Jérôme Durand-Lose
Laboratoire d'Informatique Fondamentale d'Orléans, Université d'Orléans, Orléans, France

Article Outline

Glossary
Definition of the Subject
Introduction
Computational Universality
Intrinsic Universality
Advanced Topics
Future Directions
Bibliography

Glossary

Cellular automata They are dynamical systems that are continuous, local, parallel, synchronous and space and time uniform. Cellular automata are used to model phenomena where the space can be regularly partitioned and where the same rules are used everywhere, for example: flow dynamics or percolation in physics, systolic arrays in computer science, epidemics in biology…

The configurations are infinite arrays of cells. Each cell has a state chosen inside a finite set. The dynamics is given by replacing the state of each cell according to it and the states of the cells at a bounded distance. Since there are finitely many neighboring cells, there are finitely many state patterns/inputs. The mapping to the new state is called the local function. The same local function is used for all the cells. They are all updated simultaneously.

Computational universality Computability is defined by Turing machine, μ-recursive functions or λ-calculus. All these approaches (and many more) end up defining the same set of functions over $\mathbb{N}$ (or on words, i. e. finite sequences over a finite alphabet): the computable functions. They defined, according to the Church–Turing thesis, what can be computed by any reasonable device.

A machine is computation universal if it is able to compute any computable function (indicated as a part of the entry). This corresponds also to the common approach of computer, the hardware is universal and the program to be executed is stored in main memory (like the data to process) and is part of the input as far as the hardware / operating system is concerned.

Intrinsic universality It is the capability to simulate any machine in a class of machine. If one think of Turing machines or an equivalent model of computation, this folds back to the classical computational universality. The interest of this notion lays with machines that are not equivalent to Turing machines. This is the case of cellular automata: they update infinite configurations, there are uncountably many possible configurations thus they cannot be encoded in a countable set, say $\mathbb{N}$.

An intrinsically universal CA "represents" all the CA since it can exhibit any phenomenon any other one can.

Definition of the Subject

Cellular automata (CA) and the subject are briefly defined before two kinds of universality are considered: computational universality and intrinsic universality. A more involving section on advanced topics ends this chapter.

Computational universality deals with the capability to carry out any computation as defined by Turing machines (in computability Theory) while *intrinsic universality* deals with the capability to simulate any other machine of the same class (here cellular automata). This distinction is fundamental here because while computational universality refers to finite inputs and relates to our understanding of computing with computers, intrinsic universality encompasses infinite configurations and relates to our understanding of the physical world. These universalities are presented as simply as possible and an example of universal CA is presented in each case. The last section is devoted to the history and advanced topics such as various definitions of simulation between CA, restriction to reversible CA and different underlying lattices.

Introduction

A cellular automaton is a discrete dynamical system composed of regularly displayed cells which satisfies the following properties:

1. *Local finiteness*: the set of states available for any cell is finite and there is finitely many cells in any bounded region of space;
2. *Locality of computation*: the next state of a cell only depends on the cells around it. There is no global data nor unbounded effect;
3. *Uniformity* in both space and time: the dynamics of the cells are identical (space) and never change (time);
4. *Parallelism*: all cells are updated at each iteration; and
5. *Synchrony*: all cells are iterated at the same instant.

Local finiteness and locality of computation ensure that the next state of any cell can be computed with finite information. This and uniformity ensure that a finite description exists. Parallelism and synchrony ensure that the system is deterministic.

Locality of computation also means that a finite part of a configuration can be isolated in order to see how its central part evolves. It also means that the system is continuous according to the product topology [21,48]. Uniformity also means that the date is not relevant and that if, starting from some local pattern, a phenomenon appends, when the same pattern appears again, the same phenomenon appends, whatever the iteration and location.

Universality. Being universal is somehow to represent all, to be capable to achieve anything possible. This notion is twofold: on the one hand it can be *absolute*, not related to anything in particular, and on the other hand, it can be *relative* to a specific domain. In the case of cellular automata, these two cases are: computational universality and intrinsic universality.

The first one deals with the capability to compute any computable function and relates to simulating Turing machines. Computability Theory and the Church–Turing thesis tell that the set of computable functions does not rely on one specific computing system. On the many textbooks on the subject, (see Part 2 in [51]) is among the best ones for a computer scientist approach.

The second one deals with the capability to simulate any other CA starting from any initial configuration. The distinction is real because since CA do not halt, the notion of the result of a computation is quite meaningless and because CA handle infinite configurations. Even if Turing machines with infinite entries are considered, they only update a limited part of the tape in finite time while CA update the whole (infinite) configuration at each iteration! Since there are computation universal CA in dimension one and simulation is expected to preserve this property, all intrinsically universal CA are also computation universal.

For more information on cellular automata and universality, the reader might be interested in the following surveys and books: [11,20,24,26,58,61].

Outline of this chapter. Section "Computational Universality" deals with computational universality and Sect. "Intrinsic Universality" deals with intrinsic universality. In each of these sections, definition and results as well as the construction of an example of a universal CA are provided.

Section "Advanced Topics" starts with some history on the subject and then presents advanced and more involving results on existing definitions of simulations between CA, results on the restriction to the reversible CA and on variations on different underlying lattices. Section "Future Directions" presents some insight on future researches.

Formal Definition of a Cellular Automaton

A configuration is a made of cells regularly displayed on $\mathbb{Z}^d$; d is called the *dimension*. Each cell is in a state chosen among a finite set of states Q. A *configuration* is then an element of the set:

$$C = Q^{\mathbb{Z}^d} ,$$

which is referred to as the *set of configurations*. The computation is local, each cell evolves according to the states of the cells whose coordinates differ by at most r (*radius*) on any coordinate. The local evolution is given by a *local transition function* f that maps the states of the cell and the neighboring ones to the new state of the cell. The *global transition function*, $\mathcal{G}$, maps configurations into configurations; each state is replaced by the image of the states of the cell and the neighboring ones by the local transition function. The subset of $\mathbb{Z}^d$, $\mathcal{N} = [\![-r, r]\!]^d$ is called the *(complete) neighborhood*. It represents the relative positions of neighboring cells. The neighborhood is not necessarily of this form, in fact it can be any finite subset of $\mathbb{Z}^d$. The local and global functions are defined by:

$$\begin{array}{ccc} Q^{\mathcal{N}} & \overset{f}{\rightarrow} & Q \end{array} \qquad \begin{array}{ccl} C & \overset{\mathcal{G}}{\rightarrow} & C \\ c & \mapsto & \mathcal{G}(c) \text{ s.t. } \forall x \in \mathbb{Z}^d , \\ & & (\mathcal{G}(c))_x = f(c|_{x+\mathcal{N}}) , \end{array} \tag{1}$$

where $c|_{x+\mathcal{N}}$ denotes the restriction of the configuration c to the positions in $x + \mathcal{N}$.

Definition 1 A *cellular automaton* (CA) is designated by: (d, Q, r, f). When the (finite) neighborhood does not correspond to some $[\![-r, r]\!]^d$, this is emphasized by using $\mathcal{N}$ instead of r. The *space-time diagram*, $\mathcal{D}\colon \mathbb{Z}^d \times \mathbb{N} \to Q$, or orbit of a CA is just the infinite sequence of the configurations as the CA is iterated.

Examples of space-time diagrams are provided on Figs. 2 (left) and 4. For 1-dimensional CA, the space-time diagram can be seen as a tiling of the plane. In dimension 2, each configuration can be considered as a tiling. This approach leads to many undecidability results. It is developed in the ► Tiling Problem and Undecidability in Cellular Automata.

Definition 2 If any, the *quiescent state* of a CA, $q_\#$, is a distinguished state such that the uniform configuration $\overline{q_\#}$ is map onto itself (which is equivalent to $f(q_\#, q_\#, \ldots, q_\#) = q_\#$). A configuration is *finite* if only finitely many cells are not in the quiescent state.

Computational Universality

In this section, cellular automaton are proved to be able to perform any computation in the acceptation of computability Theory and thus that there exist computation universal CA. Only useful concepts and definitions are presented.

The computable functions can be defined by μ-recursion, λ-calculus, Turing machines or any other equivalent model. The Church–Turing thesis asserts that one gets the same functions up to some encoding/representation with any reasonable mean of computation. This is important since, for example, recursive functions address functions over natural integers while Turing machines deal with computations over words (finite sequences over a finite set of symbols). A model of computation is *computation universal* if any computable function can be computed by an instance/machine of the model. A machine is *computation universal* if it can compute any computable function as long as it is provided with the description of the function to compute along with the corresponding input. Computability Theory guaranties that such a universal machine exists.

The typical techniques to prove that a model of computation, here cellular automata, is computation universal are by:

Induction like recursion Theory. This would be proving that some functions over integer (e. g. $n \mapsto n+1$) are computable and then that the functions computable in the model are closed under some operations (e. g. composition);

Simulation of a generic instance of a computation universal model of computation. This would be to prove that any, say, Turing machine could be simulated by a cellular automata; and

Simulation of a computation universal instance of a computation universal model of computation.

These approaches are very different. The first one relies on defining functions but is not concerned by the way they are computed. The second one deals with effective means of computation (allowing the implementation of algorithms and the measure of complexity). The third one is the modern vision of the all-purposes computer: a laptop can do anything from picture manipulation to music playing going through text edition and programming. There is only one hardware and according to the need, one uses a program or another. This is the duality of data and code in modern computers.

Generally, the second case is more simple to tackle. There are mainly two cases when a specific universal machine is transformed:

- When the situation is so complex that using a specific machine with few instructions becomes easier; and
- To get a computation universal CA with specific properties or qualities, typically as few states as possible.

The latter is done by designing a special simulation with a good property and then applying it to a particularly well chosen instance.

The term "computation universal" is not synonymous of "Turing equivalent" which means that whatever computes the model can be computed by, say, a Turing machine. In the case of cellular automata, this is meaningless for two reasons:

- The output is not defined since as a dynamical system, a CA never stops; and
- When considering infinite configurations, there is just no way to manipulate them with Turing machines (uncountably many configurations for countably many words).

The first reason leads to a notion of simulation in an not-ending computation rather than an input-output function definition approach. The second reason can be bypassed when considering restrictions of configurations: finite or periodic. Another canonical thing is to consider models of computation able to handle uncountably many different inputs... like cellular automata. This idea leads to intrinsic universality presented in Sect. "Intrinsic Universality".

A Computation Universal Cellular Automaton

Here Turing machines and their executions are defined and a simulation by cellular automata is provided.

A Turing machine is a very simple device: a finite automaton that can read and write on an unbounded memory (indexed by $\mathbb{N}$). The memory is organized as an infinite sequence of cells called the *tape*. Each cell has a value from a finite set of *symbols* (the alphabet). Only a finite part of the tape is not empty at any step of the computation. The automaton is equipped with a head that can read or write a single cell of the tape and move the head one position forward of backward on the tape.

Definition 3 A *Turing machine* is defined by $(\Sigma, \#, Q, q_i, \delta)$, where

- Σ is a finite *alphabet*;
- $\# \in \Sigma$ is a special symbol use to indicate *empty* part of the tape;
- Q is the *set of states* of the automaton;
- $q_i \in Q$ is the *initial state*, and
- $\delta\colon Q \times \Sigma \to Q \times \Sigma \times \{\leftarrow, \rightarrow\}$ is the *transition function*.

δ	a	b	#
q_i	q_i, b, $\rightarrow$	q_i, a, $\rightarrow$	r, #, $\leftarrow$
r	r, a, $\leftarrow$	r, b, $\leftarrow$	r, #, $\leftarrow$

x	y	z	$f(x,y,z)$
α	q_i, a	β	b
α	q_i, b	β	a
q_i, a	α	β	q_i, α
q_i, b	α	β	q_i, α
α	q_i, #	β	#
α	β	q_i, #	r, β
α	r, β	γ	β
α	β	r, γ	r, β

$\alpha, \beta, \gamma \in \{\mathtt{a}, \mathtt{b}, \#\}$

	b	
α	q_i, a	β

Cellular Automata, Universality of, Figure 1
Transition function of a Turing machine and the simulating CA

The transition function works as follows: in a state q, reading a symbol a, if $\delta(q, \mathtt{a}) = (r, \mathtt{b}, \rightarrow)$ then the new state is r, the symbol b is written instead of a and the head moves one step on the right/forward.

Definition 4 A *computation* of a TM starts with the input written on the tape (completed by #s) and the automaton in state q_i. The computation goes on as defined by the transition function δ. The computation ends when the head tries to leave the tape on the left. The result is what is written on the tape.

This is not the usual definition which involves an halting state, although it is correct. This formalization stresses that stopping is somehow an incident from the dynamical system point of view. As far as cellular automata are concerned, they do not stop; an operator might stop a CA when some condition is fulfilled but this is external to the CA.

The Turing machine considered as an example is very simple: starting on a word on $\{\mathtt{a}, \mathtt{b}\}$, it replaces each a by a b and vice-versa. The only symbols on the tape are a, b and #. There are only two states: q_i and r. The transition function is given on the left of Fig. 1.

This Turing machine is simulated by a CA of radius 1 (i. e. only cells at distance at most 1 are taken into account for computing the next state of a cell). The set of states of the CA is $\Sigma \cup \Sigma \times Q$, that is, a tape symbol alone or together with a state of the TM. The CA has 9 states, the table of its local function has $729 (= 9^3)$ entries! In the table on the right of Fig. 1 only the cases where the state of the central cell changes are indicated. On the far right, the first transition rule is represented as it appears on space-time diagrams: the bottom line represents the cell and its two closest neighbors and on top the next state of the cell (at the next iteration). In all the space-time diagrams, time is evolving upward.

The dynamics is presented on Fig. 2. On the left the whole computation of the Turing machine on the entry aab is given; the output of the function is bba as written on the tape when the machine stops. On the right, the corresponding iterations of the simulating CA are displayed. Since the CA works on a bi-infinite lattice, the configuration is completed with # on the left. As mentioned, a CA never stops so that at some point the computation has been carried out but the system goes on.

A "halting" condition has to be provided, especially when one remembers that the halting of a Turing machine (the famous *Halting problem*) is not decidable. Usually, something very simple to test is chosen, for example that some state appears somewhere. Here this would be the first time the closest #-cell on the left is not in state #.

This example can easily be extended to a general method to simulate any Turing machine or specifically a computation universal one. Starting from a computation universal TM, a computation universal CA is generated.

Other Ways to Achieve Computational Universality

Among the various systems achieving computational universality that have been used to prove computational universality of CA, a brief classification can be made.

Machines. (Turing machines, counter automata and random access machines) These systems are very simple, an automaton together with a memory. In general, the whole evolution/orbit of the Turing machine is encoded inside the space-time diagram of the cellular automata like in the example. For counter automata [36], there are finitely many counters but they can hold any natural integer, any counter can be accessed at any time. Random access machines are like counter automata but with infinitely many register and an indirect access mode which allows to access any register. The random access machines model is the closest to modern computer architecture, but to sim-

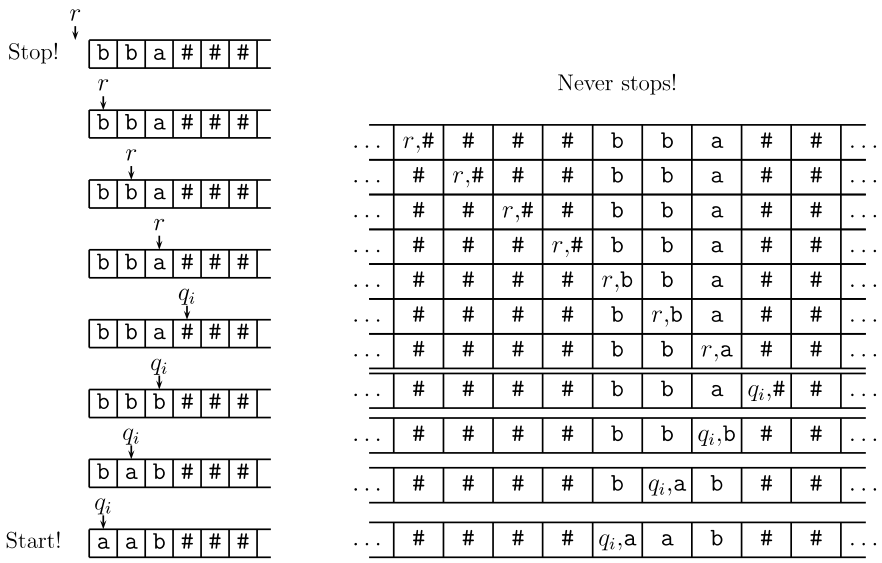

Cellular Automata, Universality of, Figure 2
Iterations of a Turing machine and of the simulating CA

ulate it, one has to consider indirect addressing and infinitely many register. Counter automata are more simple to simulate: since there exists a computation universal 2-counter automaton, only 6 instructions have to be implemented.

Boolean circuits. The idea is to encode Boolean logic and then to say that any value can be encoded in binary and any function can then be computed with the binary encoding. For example, the transition function of a TM (resp. 2-counter automata) can be encoded and the value of the tape (resp. the counters) stored in an infinite memories. This is generally done by providing a way to encode bits, then logical gates and finally wiring to connect them. It is usually done in dimension 2, since ensuring a correct wire crossing is more complicated in dimension 1. A typical example of this is the computational universality of the Game of Life [6,19]. This is developed in the ► Gliders in Cellular Automata.

Rewriting systems. These systems work on words by removing some sub-word and adding some other sub-word. Starting with a word encoding an entry, the system is expected to stop with the output encoded in the final word. Here are some examples of computation universal systems:

- Type-0 grammar in Chomsky's hierarchy: one may replace sub-words by others according to rewriting rules; and
- Tag-systems: a prefix of the word is removed and according to it and some rules, a suffix is added.

The proof of the universality of the 2-states 3-neighbors 1-dimension CA referred as the elementary cellular automaton 110 by Cook [9] is done with some tag-system. The construction relies on an intermediate level of simulation that ensures signal transmission and updating. It is too involving and lengthy to be presented here. Signals and particles/solitons [1,53] are often used to carry bits/information around.

Consequences of Computational Universality

To have computation universality provides two things. On the positive side:

- Any computation can be carried out, so that if something has to be computed, whether as a final result or to use in a broader scheme, it can be done;
- Since any computation may take place, complex and interesting behaviors may appear;

and on negative side: many questions one may ask about the system become undecidable. The second point comes from the undecidability of the *Halting problem* (whether a given computation of a TM halts). Here are a few examples of this – many more can be found in, for example the ► Tiling Problem and Undecidability in Cellular Automata –, given a CA and a finite initial configuration:

- Will some state ever appear?
- Will the CA ever enter a stable configuration?
- Will the configuration grow infinitely?
- Will some given configuration be reached?

"Quality" of a Computation Universal Cellular Automata

Computability Theory is rather disappointing: one the one side what is computing and computable becomes clear but on the other side it mostly provides negative results like this and that are not computable. In this subsection slight differences on computability/simulation are addressed as well as complexity (of computing) issues.

For Turing machines, the written part of the tape is always finite. The tape is *potentially infinite*, it is extended by # as much as needed, it is never infinite. For computing with a CA, it can be assumed that only a finite part of the space is used for computing since otherwise it would take an infinite time for states to interact (it takes $l/2r$ for cells at distance l to interact). There is a definition of finiteness for CA (Def. 2): outside a finite part is it the quiescent state $q_{\#}$ (which plays the same role as #). When an infinite configuration (even if it is periodic after some point) is used to achieve computational universality, one talks about *weak computational universality*; more insights on the weakness notion on Turing machines can be found in [62].

Another important criterion is the one of the efficiency of computation. For example, the simulating a Turing machine by a 2-counter automaton suffers an exponential slowdown so that one may not be interested in such a computing device or any device where computational universality derives from a 2-counter automaton. For example, the proof of the computational universality of rule 110 by Cook [9] has an exponential slowdown but Neary and Wood [42] proved that in fact that it can be done with a polynomial slowdown. Polynomial slowdown is the classic simulation mode between reasonable models of computation (and this one of the key to the definition – and stability – of the complexity class P –polynomial time solvable problems).

Another thing worth mentioning is that CA are inherently parallel, the universality proofs more or less directly leads to the Turing machines which is the canonical sequential model. Various approaches have been made to consider CA as a computing system on its own, independently of any other model of computation, for example:

- *Iterative automaton* where the input is given symbol by symbol to a distinguished cell;
- As word recognizers with only one cell per symbol [55] (this cannot be computation universal because the memory is bounded);
- If finitely many cells are considered but they are all equipped with a stack, then any computation can be made [27];
- A algorithmic on CA based on signals has been developed [12,32,35]; and
- Martin devised an intrinsically universal CA together $S-m-n$ theorem for CA (as computing devices over infinite configurations) to provide a *acceptable programming system* point of view [31].

This directly links to the notion of universality developed in the next section.

Intrinsic Universality

Definition

Previous section deals with universality as the capability to perform any computation as defined in computability Theory. This theory deals with numbers/words and there exist only countably many configurations. But there are uncountably many configurations for any CA (unless, of course, if there is only one state). It is worth inquiring about another kind of universality that would take this into account and not involve any other model of computation. (It can thus be used for any class of dynamical systems.) It is some kind of *inner-universality*. The idea behind *intrinsic universality* is somehow the counterpart of universality for a Turing machine: being able to simulate any other CA (of the same dimension) on any (infinite) configuration.

Definition 5 A cellular automaton $\mathcal{A}$ *simulates* another CA $\mathcal{B}$ if there is an injective (one-to-one) function, ι, from the configurations of $\mathcal{B}$ to the ones of $\mathcal{A}$ and an integer, τ, such that the following diagram commutes:

$$\begin{array}{ccc} C_{\mathcal{B}} & \xrightarrow{\iota} & C_{\mathcal{A}} \\ {\scriptstyle \mathcal{G}_{\mathcal{B}}}\downarrow & & \downarrow{\scriptstyle \mathcal{G}_{\mathcal{A}}^{\tau}} \\ C_{\mathcal{B}} & \xrightarrow{\iota} & C_{\mathcal{B}} \end{array}$$

A cellular automaton is *intrinsically universal* if it can simulate any other CA (of the same dimension).

The injectivity ensures that $\mathcal{B}$-configurations can be distinguished when mapped into $\mathcal{A}$-configurations. From this definition, it directly comes that:

$$\forall n \in \mathbb{N}\,, \quad \iota \circ \mathcal{G}_{\mathcal{B}}^{n} = \mathcal{G}_{\mathcal{A}}^{n.\tau} \circ \iota\,.$$

In an infinite run, $\mathcal{A}$ generates one iteration of $\mathcal{B}$ every τ iterations. Since the composition of injective functions is injective, the simulation relation is transitive. The definition is illustrated by the construction of an example in the rest of this section.

Other definitions of simulation can be found in Subsect. "Defining Simulation Among CA (For Intrinsic Universality)". In every case, a different intrinsic universality is generated.

The Way It Usually Works

It is assumed that the simulated CA has radius 1 and is 1-dimensional (how to deal with higher dimension is explained at the end of this section). If it is not the case, it is easy to simulate it by one of radius 1 (and then to use transitivity). Let $\mathcal{A} = (1, Q, r, f)$ be such that the radius is greater than 1. The idea is to group cells r by r and define $\mathcal{B} = (1, Q^r, 1, f_{\mathcal{B}})$ such that $f_{\mathcal{B}}$ is:

$$\begin{aligned}
&\forall (c_1, c_2, \ldots, c_r), (c_{r+1}, c_{r+2}, \ldots, c_{2r}),\\
&(c_{2r+1}, c_{2r+2}, \ldots, c_{3r}) \in (Q^r)^3,\\
&f_{\mathcal{B}}((c_1, c_2, \ldots, c_r), (c_{r+1}, c_{r+2}, \ldots, c_{2r}),\\
&(c_{2r+1}, c_{2r+2}, \ldots, c_{3r}))\\
&\qquad = (f(c_1, c_2, \ldots, c_{2r+1}), f(c_2, c_3, \ldots, c_{2r+2}),\\
&\qquad\qquad \ldots, f(c_r, c_{r+1}, \ldots, c_{3r}))\,.
\end{aligned}$$

The injection ι is the canonical injection where states are grouped r by r:

$$\forall c \in C\,, \forall i \in \mathbb{Z}\,, (\iota(c))_i = (c_{ir}, c_{ir+1}, \ldots, c_{(i+1)r-1})\,.$$

This simulation is just a rescaling of space. From now on, only radius 1 CA are considered. The construction of an intrinsically universal CA uses two scales:

- *Meta-cells scale* to manipulate the states and the transition function of the simulated CA; and
- *Bit scale* to implement the meta-cells.

A meta-cell gathers copies of the states of its two closest neighbors. Then it has to use the transition function to compute the new state. The transition function is tabulated in the form of a sequence of blocks $(\mathtt{a}, \mathtt{b}, \mathtt{c}, f(\mathtt{a}, \mathtt{b}, \mathtt{c}))$. (It could also have been be represented by, e. g., a boolean circuit with states binary encoded as in [45].)

Meta-Cells Scale Each meta-cell holds one block/entry of the transition table. The entries are infinitely repeating on both side and are endlessly shifting on the left so that in one period, the block corresponding to the update eventually appears. When one period has passed a new cycle starts. The architecture of meta-cells is presented on Fig. 3.

In the upper part of Fig. 4, a space-time diagram generated by some CA is given as well as the transition function of the CA in terms of a relation inside the space-time diagram between three states (at an iteration) and the state above (at the next iteration). In the lower part of Fig. 4, the simulation, at the meta-cell scale is presented. Each meta-cell corresponds to the presentation in Fig. 3. The transition table is set inside the initial configuration. The starting index is used by the meta-cell to detect the end of the cycle.

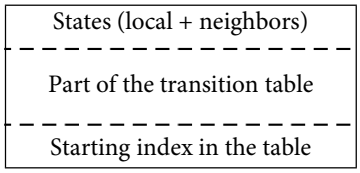

Cellular Automata, Universality of, Figure 3
Meta-cell

As it can be seen on Fig. 4, on the upper part of each cell; at start (lowest row) there are three copies of the state of the simulated cell then, copies are exchanged with the neighbors (first simulating iteration). Then the transition table starts moving. As soon as the entry is found in the table, the three states are replaced by three copies of the new state. The meta-cell then waits for its index to appear again to start this cycle again. The synchrony and uniformity of CA ensures that all meta-cells stay synchronized.

Bits Scale Since the encoding and simulation must work for any CA, it must be able to handle any set of states. The set of states of any CA if finite but unbounded, thus it is impossible to use a common set of states. A state a is thus binary encoded (denoted $(a)_2$). This way if the set of states is larger, then the meta-cell is larger, composed of more elementary cells.

On Fig. 5, a meta-cell is given and the layers named. The set of states on each layer for elementary cells are also given. Some layers are added:

- Structure: to delimit the different parts of the meta-cell,
- Control: to drive the movement, copy and test of bits, and
- Left and right: to carry bits around (to exchange states and to shift the transition table).

The left and right layers carry the bits very simply: unless noted otherwise the new value in left (resp. right) is the one that on the left (resp. right) layer of the right (resp. left) cell ensuring a left (resp. right) shift on the layer. When the simulation starts, s_{-1} and s_1 are equal to s_0 and (a, b, c) is equal to (a_0, b_0, c_0).

The meta-cell implementation works as follows: there is a single value in the control layer. It moves forth and back on the entire meta-cell like a signal. It manages bits from various layers and changes its state accordingly. The construction is only sketched; the states mentioned below are in an intermediate level between meta-cells and elementary cells. The universal CA is not detailed because although this is not complicated it would be quite lengthy and not very informative.

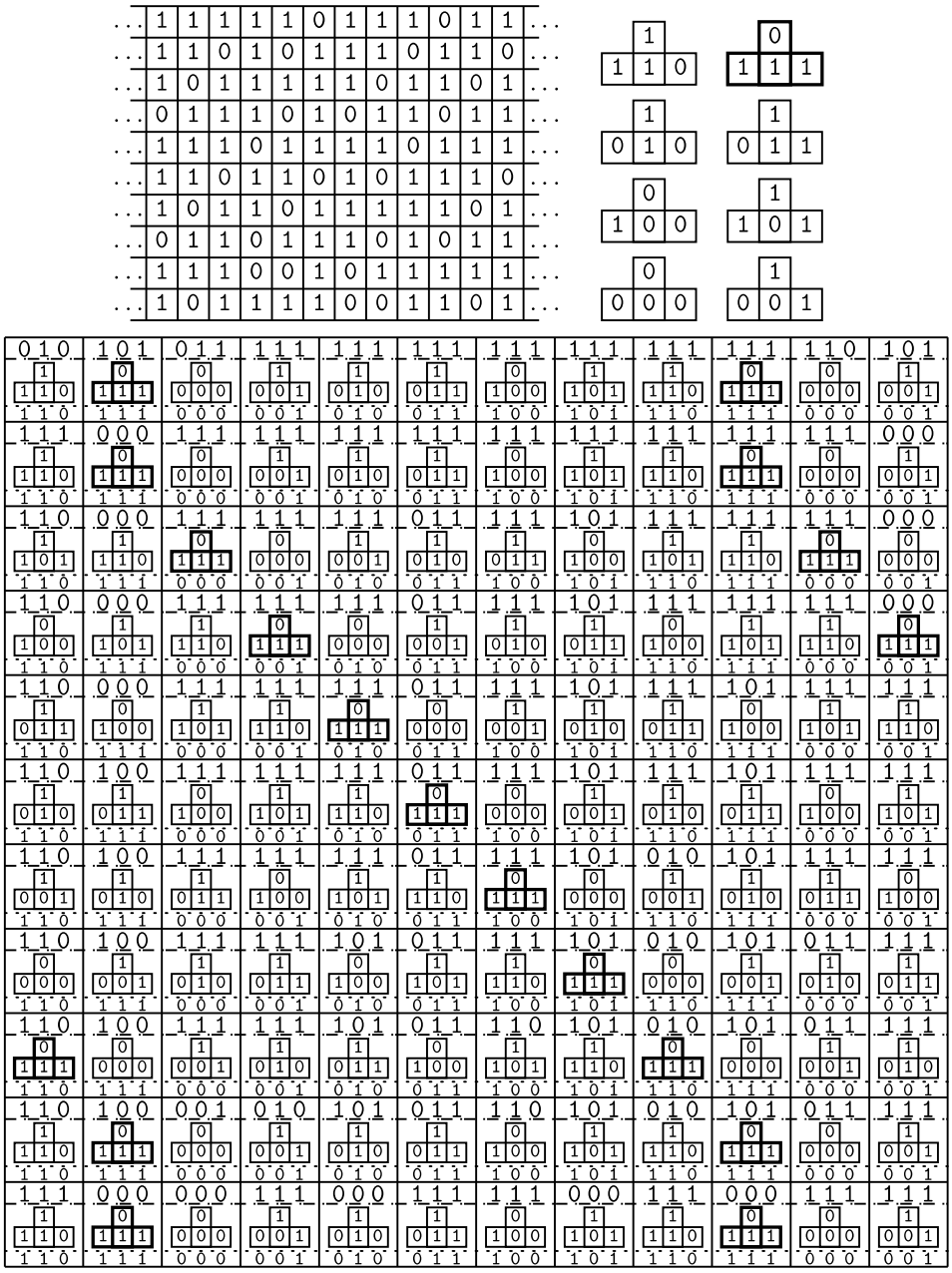

Cellular Automata, Universality of, Figure 4
Iteration of a CA and simulation at meta-cell scale of the first iteration

Starting in state q_0, a meta-cell first carries out the states exchange with the meta-cell on the left (the meta-cell on the right takes care of the other exchange) bit by bit using the left and right layers. In state q_1, the transition table is shifted by one entry. In state q_2, it moves through the entire meta-cell and checks whether the state and rule-in layers are identical. If the layers agree, the right transition is found and the state layer can be updated (the bits are copied) then it enters q_3 otherwise it restarts in q_1. In state q_3, the meta-cell has been updated; the transition

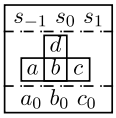

$*$	$-$	$\cdots$	$-$	$+$	$-$	$\cdots$	$-$	$+$	$-$ $\cdots$ $-$	structure $\in \{*, +, -\}$
q_0	$-$	$\cdots$	$-$		$-$	$\cdots$	$-$		$-$ $\cdots$ $-$	control $\in Q_{\mathcal{U}} \cup \{-\}$
0	$\cdots$	0		0	$\cdots$	0		0	$\cdots$ 0	left $\in \{0,1\}$
0	$\cdots$	0		0	$\cdots$	0		0	$\cdots$ 0	right $\in \{0,1\}$
$(s_1)_2$				$(s_0)_2$				$(s_{-1})_2$		state $\in \{0,1\}$
$(d)_2$				$(d)_2$				$(d)_2$		rule-out $\in \{0,1\}$
$(a)_2$				$(b)_2$				$(c)_2$		rule-in $\in \{0,1\}$
$(a_0)_2$				$(b_0)_2$				$(c_0)_2$		index $\in \{0,1\}$

Cellular Automata, Universality of, Figure 5
Binary encoding of a meta-cell with elementary cells

table is still shifted until a full shift has been done. This is indicated by identical rule-in and index layers. In this the case it enters q_0 and the simulation of a new iteration starts.

Various technical things like delays are added to keep meta-cells synchronized. The number of iterations needed to simulate an iteration is roughly speaking the length of a cell multiplied by the length of the transition table: $\log(|Q|).|Q|^3$.

Higher dimension. It is treated exactly in the same way. One layer gathers the information, another one deals with the transition function. The transition table shifting is done on one dimension only. Extra dimensions can be used to accelerate the process or to design circuitry for a more efficient encoding of the transition function.

Intrinsically universal CA are computation universal by composition (as long as the definitions are robust enough). There is no notion of semi-weakly simulation since the whole configuration has to be used to encode an infinite configuration.

Advanced Topics

This section gathers a brief history of the subject and various results on specific approaches. It is more involving and targeted to a learned reader.

A Bit of History

Since there is a lot of places where the history of CA is presented (as well as other chapters of this encyclopedia, the reader might be interested in the following survey: [50]), only the universality part is developed here.

Cellular automaton were introduced in the 50's by Ulam and von Neumann [59] to study self-reproduction [3,7,60]. Computational universality was used just to prove that any pattern can be built. Universality was investigated and proved without any explicit distinction between computational and intrinsic universality before [5].

The most famous CA is certainly Conway's Game of Life [19] from the early 70s. This 2-dimensional CA is both computation universal [6] and intrinsically universal [14].

Quest for Small Universal CA There have been an ongoing search for more than four decades for universal CA as small as possible. The interest is to know whether "small" CA are more simple and thus can be handled or there is no gap in complexity. Table 1 sums up results in this quest. Intrinsically universal CA are also computation universal but the converse is not true.

These results were achieved with various constructions. Some use quite different definitions of CA for commodity (but there are indeed CA):

- As partitioned CA [38], and
- As CA with Margolus's neighborhood (or partitioning CA): [30,58] (billiards) and [10] (spin model in Physics).

Computing universality can also be defined by the computational complexity of the sets of orbits of a CA as well as the reachability relation, relating to the Turing degrees of undecidability [54].

Defining Simulation Among CA (For Intrinsic Universality)

The definition used in Sect. "Computational Universality" is not the only existing one. Many papers prove results on simulation without providing a formalized definition of it, but by considering the construction anyone would say that it is a simulation, in an empirical fashion. There is no absolute definition commonly accepted and none contradicts the intuition of simulation. In Table 1, no distinction on the simulation is made and most of the time the construction fits in more than one definition.

Most of the presented definitions can be amended in order to cover cases where not all iterations are covered,

Cellular Automata, Universality of, Table 1
Historic bounds on computation and intrinsically universal CA

Year	Reference	Dimension	$\lvert\mathcal{N}\rvert$	$\lvert Q\rvert$	Computation	Intrinsic
1966	Neumann [60]	2	5	29	✘	✘
1968	Codd [8]	2	5	8	✘	✘
1970	Banks [5]	2	9	2	✘	✘
		1	3	18	✘	✘
		1	5	2	✘	✘
1971	Smith III [52]	2	7	7	✘	
1987	Albert and Culik [2]	1	3	14	✘	✘
1990	Lindgren and Nordahl [28]	1	3	7	✘	
2002	Ollinger [45]	1	3	6	✘	✘
2004	Cook [9]	1	3	2	✘	
2006	Richard [47]	1	3	4	✘	✘

say for example only the one out of 3. In the following, $\mathcal{A} = (Q_{\mathcal{A}}, r_{\mathcal{A}}, f_{\mathcal{A}})$ and $\mathcal{B} = (Q_{\mathcal{B}}, r_{\mathcal{B}}, f_{\mathcal{B}})$ denote cellular automata of the same dimension d.

Embedding of Hertling The term *embedding* is to be understood as simulation. It was introduced in order to prove that CA that are not onto cannot be simulated by onto (and specially reversible) ones of the same dimension.

Definition 6 (Embedding [23]) A mapping $\mu\colon Q_{\mathcal{A}}^{\mathbb{Z}^d} \to Q_{\mathcal{A}}^{\mathbb{Z}^d}$ is said to be a *morphism* if and only if for any shift of $\mathbb{Z}^d, \sigma_{\mathcal{A}}$, there exists a shift of $\mathbb{Z}^d, \sigma_{\mathcal{B}}$, such that: $\mu \circ \sigma_{\mathcal{A}} = \sigma_{\mathcal{B}} \circ \mu$.

$\mathcal{A}$ can be embedded in $\mathcal{B}$ if there are mapping $\mu\colon Q_{\mathcal{A}}^{\mathbb{Z}^d} \to Q_{\mathcal{A}}^{\mathbb{Z}^d}$ and $\nu\colon Q_{\mathcal{A}}^{\mathbb{Z}^d} \to Q_{\mathcal{A}}^{\mathbb{Z}^d}$ and an integer k such that:

$$\forall t \in \mathbb{N}\,, \quad \mathcal{G}_{\mathcal{A}}^{t} = \nu \circ \mathcal{G}_{\mathcal{B}}^{kt} \circ \mu\,. \tag{2}$$

The embedding is *strong* if μ is a continuous morphism, *weak* if it is a just a morphism and *set-theoretic* otherwise (i. e. not a morphism).

The interest of the morphism is to enforce the respect of the structure of $\mathbb{Z}^d$. Hertling proved, using the Axiom of Choice (hence the qualification) that any CA can be set-theoretically embedded in the one dimensional CA that does nothing but shift the value on the left. This is of course highly nonconstructive and contradict the intuition of what simulation could be.

Comparing to Def. 5, using ν to come back allows to have some garbage produced and discarded by ν.

Grouping Relation of Mazoyer and Rappaport This definition of a grouping relation was introduced to focus on the importance of space and time structure. It allows to consider iterations once in a while, periodically.

Definition 7 (Grouping [33,34]) A cellular automaton $\mathcal{A}$ is a *sub-automaton* of $\mathcal{B}$ (denoted $\mathcal{A} \subseteq \mathcal{B}$) if there is an injection $\phi\colon Q_{\mathcal{A}} \to Q_{\mathcal{B}}$ such that:

$$\overline{\phi} \circ \mathcal{G}_{\mathcal{A}} = \mathcal{G}_{\mathcal{B}} \circ \overline{\phi}\,,$$

where $\overline{\phi}$ is the component-wise extension of ϕ to $\mathcal{A}$-configurations. The nth grouping of an automaton is defined by grouping the cells n by n and consider only every nth iteration, $\mathcal{A}^n = (d, Q_{\mathcal{A}}^n, r_{\mathcal{A}}, f_{\mathcal{A}}^{(n)})$. The function $\mathcal{G}_{\mathcal{A}^n}$ is the nth iterate of $\mathcal{G}_{\mathcal{A}}$. Since the cells are grouped by n, the radius is not changed. The grouping relation is defined by:

$$\mathcal{A} \leq_{\text{grouping}} \mathcal{B} \iff \exists n, m \in \mathbb{N}\,,\ \ 0 < n, m\,,\ \ \mathcal{A}^n \subseteq \mathcal{B}^m\,.$$

This is a stronger form of simulation where the space-time diagrams should be included; up to some rescaling on both side, all the space-time diagrams of $\mathcal{A}$ should be exactly (up to an injection) generated. There is no shift involved and the space-time ratio should be preserved.

This relation is a pre-order. The authors proved that there is a bottom equivalence class for CA: the CA with only one state (and one configuration). The nilpotent ones (after a fixed number of iterations any configuration is turned into the same one: only quiescent state) are just above. They also proved that there is an unbounded infinite ascending chain, so that there is no top and thus no intrinsically universal CA for this definition.

Rescaling of Ollinger The previous definition is on the one hand interesting because it relays on space-time diagrams and a natural operation, grouping, over them, but on the other hand, there is no intrinsically universal CA, which have been provided for other definition. The following definition is a weakening of the first one that allows intrinsically universal CA.

Definition 8 (Rescaling [44]) For any k in $\mathbb{Z}^d$, let σ^k denotes the *shift* by k over configurations. For any m in $\mathbb{Z}^d$, let o^m denotes the *packing* of cells into packs of size m; it is a mapping from $Q^{\mathbb{Z}^d}$ into $(Q^m)^{\mathbb{Z}^d}$ (o^{-m} is the inverse, the unpacking function).

For $n, k \in \mathbb{Z}^d$ and $n \in \mathbb{N}$, $0 < n$, the $\langle m, n, k\rangle$-rescaling of $\mathcal{A}$ is the cellular automaton $\mathcal{A}^{\langle m,n,k\rangle}$ such that:

$$\mathcal{G}_{\mathcal{A}^{\langle m,n,k\rangle}} = \sigma^k \circ o^m \circ \mathcal{G}^n_{\mathcal{A}} \circ o^{-m} .$$

A cellular automaton $\mathcal{A}$ is simulated by $\mathcal{B}$ if there exists a rescaling of $\mathcal{A}$ which is a sub-automaton of a rescaling of $\mathcal{B}$. (The sub-automaton relation is defined as in Def. 7.)

This definition allows to include a shift and to treat independently the size of the blocks of cells and the iteration step. Comparing to the grouping definition (Def. 7), this allows to have enough time to locally mix information and compute the next state. Intrinsically universal CA exist (also with a meta-cell approach) and intrinsic universality of a 1-d CA is undecidable [46]. This result is still true on captive CA (the transition function may only output a state that is in the input) even though as the number of states grows larger, almost all captive CA is intrinsically universal [56,57].

Reversible Case

The reversible subset of cellular automaton (CA such that the global function is invertible, its inverse is then the one of a CA) also contains computation universal CA [13,38,40]. This topic is developed in the ► Reversible Cellular Automata.

There exist reversible CA that are intrinsically universal inside reversible CA [15,16,18]. This means able to simulate any other reversible CA (of the same dimension) and not just any CA. It is a strong embedding (Def. 6) and does not contradict Hertling's results [23] that non surjective CA cannot be simulated by reversible one.

Morita proved the any CA can be simulated over finite configurations by a reversible in [37,39] but garbage is produced in order to ensure reversibility and the simulation time varies as the simulation goes on.

There is also a particular result [17] including the simulation of the non-reversible CA, but the simulation goes by a different definition. It is centered (like the usual metric) and the simulated iterated configurations are displayed on parabolas. This twisting of space yields an infinite space to store the information for reversibility. It is not possible to recover a simulated iteration from finitely many simulating ones! The whole simulating space-time diagram is needed.

Variations on CA

Changing the Underlying Space Other 2-dimensional spaces have been considered. There exist reversible computation universal CA both on triangular lattices [25] and hexagonal lattice [41].

Róka studied simulation between CA on different lattices in a very general way: lattices are Cayley graph (it corresponds to a group, the arrows corresponds to generators). She proved the existence of simulations in the case of the existence of an homomorphism with a finite kernel and that all bi-dimensional planar structure are equivalent to $\mathbb{Z}^2$ [49].

There exist computation universal CA [22] and intrinsically universal CA [29] on the hyperbolic plane. This is more developed in the ► Cellular Automata in Hyperbolic Spaces.

Intrinsic Universality Among Quantum Cellular Automaton In the past decade, quantum computation theory has been tremendously developing. It relies on unitary gates which are of course reversible. The results on reversible CA has been "naturally" extended, for example, there is a 1-dimension Quantum CA which is intrinsically universal (among quantum CA) [4]. For more on the topic, please refer to the ► Quantum Cellular Automata.

Variable Neighborhood A new approach is to fix the states and the transition function and to have only the neighborhood (i. e. the relative localization of the entries of the transition function) varying. Somehow it can be considered that the neighborhood is not defined by the CA but by the configuration. Some simulation results exists [43,63]. The transition function of simulated CA is not given inside the simulating configuration but by the simulating neighborhood.

Future Directions

As mentioned just above, understanding the role played by the neighborhood in computing might be very enlightening.

It is known that 2 states and 2 neighbors is enough for computing with polynomial slowdown. It might be interesting to find constructions with limited slowdown and very concise encoding. Since CA are inherently parallel while Turing machines are sequential, a computing (and complexity) theory and algorithm that incorporate the parallelism of CA (local, uniform and synchronous) is worth enquiring.

As far as intrinsic universality is concerned, it relies on simulation between CA. The various definitions have to be linked and investigated.

Bibliography

1. Adamatzky A (ed) (2002) Collision based computing. Springer
2. Albert J, Čulik K II (1987) A simple universal cellular automaton and its one-way and totalistic version. Complex Syst 1:1–16
3. Arbib MA (1966) Simple self-reproducing universal automata. Inf Control 9(2):177–189
4. Arrighi P, Fargetton R (2007) Intrinsically universal one-dimensional quantum cellular automata. arXiv:0704.3961
5. Banks ER (1970) Universality in cellular automata. In: Eleventh annual symposium on switching and automata theory. IEEE, pp 194–215
6. Berlekamp E, Conway J, Guy R (1982) Winning ways for your mathematical plays (games in particular), vol 2. Academic Press
7. Burks A (1970) Essays on cellular automata. Univ of Illinois Press
8. Codd E (1968) Cellular automata. Academic Press
9. Cook M (2004) Universality in elementary cellular automata. Complex Syst 15:1–40
10. Cordero P, Goles E, Hernández G (1992) Q2R + Q2R as a universal billiard. Int J Modern Phys C 3(2):251–266
11. Čulik K II, Pachl J, Yu S (1990) Computation theoretic aspects of cellular automata. Phys D 45(1–3):357–378
12. Delorme M, Mazoyer J (2002) Signals on cellular automata. In: Adamatzky A (ed) Collision-based computing. Springer, pp 234–275
13. Dubacq J-C (1995) How to simulate Turing machines by invertible 1d cellular automata. Int J Found Comput Sci 6(4):395–402
14. Durand B, Róka Z (1998) The game of life: Universality revisited. In: Delorme M, Mazoyer J (eds) Cellular Automata: A parallel model. Math Appl, vol 460. Kluwer, Dordrecht, pp 51–74
15. Durand-Lose J (1995) Reversible cellular automaton able to simulate any other reversible one using partitioning automata. In: LATIN'95. LNCS, vol 911. Springer, Berlin, pp 230–244
16. Durand-Lose J (1997) Intrinsic universality of a 1-dimensional reversible cellular automaton. In: STACS'97. LNCS, vol 1200. Springer, pp 439–450
17. Durand-Lose J (2000) Reversible space-time simulation of cellular automata. Theor Comput Sci 246(1–2):117–129
18. Durand-Lose J (2002) Computing inside the billiard ball model. In: Adamatzky A (ed) Collision-based computing. Springer, pp 135–160
19. Gardner M (1970) Mathematical games. The fantastic combinations of John Conway's new solitaire game "life". Sci Am 223:120–123
20. Gutowitz H (ed) (1991) Cellular automata, theory and experimentation. MIT/North-Holland
21. Hedlund GA (1969) Endomorphism and automorphism of the shift dynamical system. Math Syst Theory 3:320–375
22. Herrmann F, Margenstern M (2003) A universal cellular automaton in the hyperbolic plane. Theor Comput Sci 296:327–364
23. Hertling P (1998) Embedding cellular automata into reversible ones. In: Calude C, Casti J, Dinneen M (eds) Unconventional models of computation. DMTCS. Springer
24. Ilachinski A (2001) Cellular automata – A discrete universe. World Scientific
25. Imai K, Morita K (1998) A computation-universal two-dimensional 8-state triangular reversible cellular automaton. In: Margenstern M (ed) Universal machines and computations (UCM 98), vol 2. Université de Metz, pp 90–99
26. Kari J (2005) Theory of cellular automata: A survey. Theor Comput Sci 334:3–33
27. Kutrib M (2001) Efficient universal pushdown cellular automata and their application to complexity. In: Margenstern M, Rogozhin Y (eds) Machines, computations, and universality (MCU'01), Chisinau, Moldavia, 23–27 May 2001. LNCS, vol 2055. Springer, Berlin, pp 252–263
28. Lindgren K, Nordahl MG (1990) Universal computation in simple one-dimensional cellular automata. Complex Syst 4:299–318
29. Margenstern M (2006) An algorithm for buiding inrinsically universal automata in hyperbolic spaces. In: Arabnia HR, Burgin M (eds) International conference on foundations of computer science (FCS'06). pp 3–9
30. Margolus N (2002) Universal cellular automata based on the collisions of soft spheres. In: Adamatzky A (ed) Collision-based computing. Springer, pp 107–134
31. Martin B (1994) A universal cellular automaton in quasi-linear time and its S-n-m form. Theor Comput Sci 123:199–237
32. Mazoyer J (1996) Computations on one dimensional cellular automata. Ann Math Artif Intell 16:285–309
33. Mazoyer J, Rapaport I (1998) Inducing an order on cellular automata by a grouping operation. In: 15th annual symposium on theoretical aspects of computer science (STACS'98). LNCS, vol 1373. Springer, Berlin, pp 116–127
34. Mazoyer J, Rapaport I (1999) Inducing an order on cellular automata by a grouping operation. Discet Appl Math 218:177–196
35. Mazoyer J, Terrier V (1999) Signals in one-dimensional cellular automata. Theor Comput Sci 217(1):53–80
36. Minsky M (1967) Finite and infinite machines. Prentice Hall
37. Morita K (1992) Any irreversible cellular automaton can be simulated by a reversible one having the same dimension. Technical report of the IEICE Comp 92-45(1992-10):55–64
38. Morita K (1992) Computation-universality of one-dimensional one-way reversible cellular automata. Inform Process Lett 42:325–329
39. Morita K (1995) Reversible simulation of one-dimensional irreversible cellular automata. Theor Comput Sci 148:157–163
40. Morita K, Harao M (1989) Computation universality of one-dimensional reversible (injective) cellular automata. Trans IEICE E 72(6):758–762
41. Morita K, Margenstern M, Imai K (1998) Universality of reversible hexagonal cellular automata. In: MFCS'98 Satellite workshop on frontiers between decidability and undecidability
42. Neary T, Woods D (2006) P-completeness of cellular automa-

ton rule 110. In: Bugliesi M, Preneel B, Sassone V, Wegener I (eds) International colloquium on automata languages and programming (ICALP'06). LNCS, vol 4051(1). Springer, Berlin, pp 132–143
43. Nishio H (2007) Changing the neighborhood of cellular automata. In: Durand-Lose J, Margenstern M (eds) Machine, computations and universality (MCU'07). LNCS, vol 4664. Springer, Berlin, pp 255–266
44. Ollinger N (2001) Two-states bilinear intrinsically universal cellular automata. In: FCT'01. LNCS, vol 2138. Springer, pp 369–399
45. Ollinger N (2002) The quest for small universal cellular automata. In: ICALP'02. LNCS, vol 2380. Springer, pp 318–329
46. Ollinger N (2003) The intrinsic universality problem of one-dimensional cellular automata. In: STACS'03. LNCS, vol 2607. Springer, Berlin, pp 632–641
47. Richard G (2006) A particular universal cellular automaton. oai:hal.ccsd.cnrs.fr:ccsd-00095821_v1
48. Richardson D (1972) Tessellations with local transformations. J Comput Syst Sci 6:373–388
49. Róka Z (1999) Simulation between cellular automata on cayley graphs. Theor Comput Sci 225:81–111
50. Sarkar P (2000) A brief history of cellular automata. ACM Comput Surv 32(1):80–107
51. Sipser M (1997) Introduction to the theory of computation. PWS Publishing Co, Boston
52. Smith AR III (1971) Simple computation-universal cellular spaces. J ACM 18(3):339–353
53. Steiglitz K, Kamal I, Watson A (1988) Embedding computation in one-dimensional automata by phase coding solitons. IEEE Trans Comput 37(2):138–145
54. Sutner K (2005) Universality and cellular automata. In: Margenstern M (ed) Machines, computations, and universality (MCU'04). LNCS, vol 3354. Springer, Berlin, pp 50–59
55. Terrier V (1996) Language not recognizable in real time by one-way cellular automata. Theor Comput Sci 156:281–287
56. Theyssier G (2004) Captive cellular automata. In: Fiala J, Koubek V, Kratochvíl J (eds) Mathematical foundations of computer science (MFCS'04), 29th international symposium, Prague, Czech Republic, 22–27 August. LNCS, vol 3153. Springer, Berlin, pp 427–438
57. Theyssier G (2005) How common can be universality for cellular automata? In: Diekert V, Durand B (eds) 22nd annual symposium on theoretical aspects of computer science (STACS'05), Stuttgart, Germany, 24–26 February. LNCS, vol 3404. Springer, Berlin, pp 121–132
58. Toffoli T, Margolus N (1987) Cellular automata machine – A new environment for modeling. MIT press, Cambridge
59. Ulam S (1952) Random processes and transformations. In: International congress of mathematics 1950, vol 2. pp 264–275
60. von Neumann J (1966) Theory of self-reproducing automata. University of Illinois Press, Urbana
61. Wolfram S (2002) A New kind of Science. Wolfram Media
62. Woods D, Neary T (2007) Small semi-waekly universal turing machines. In: Durand-Lolse J, Margenstern M (eds) Machines, computations and universality (MCA'07). LNCS, vol 4664. Springer, Berlin, pp 246–257
63. Worsch T, Nishio H (2007) Variations on neighborhoods in ca. In: Moreno-Diaz R (ed) EUROCAST'07. LNCS, vol 4739. Springer, Berlin, pp 581–588

Cellular Automaton Modeling of Tumor Invasion

Haralambos Hatzikirou[1], Georg Breier[2], Andreas Deutsch[1]
[1] Center for Information Services and High Performance Computing, Technische Universität Dresden, Dresden, Germany
[2] Institute of Pathology, Medical Faculty Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Cellular Automata
Models of Tumor Invasion
Discussion
Acknowledgments
Bibliography

Glossary

Cadherins Important class of transmembrane proteins. They play a significant role in cell-cell adhesion, ensuring that cells within tissues are bound together.

Chemotaxis Motion response to chemical concentration gradients of a diffusive chemical substance.

Extracellular matrix (ECM) Components that are extracellular and composed of secreted fibrous proteins (e. g. collagen) and gel-like polysaccharides (e. g. glycosaminoglycans) binding cells and tissues together.

Fiber tracts Bundle of nerve fibers having a common origin, termination, and function and especially one within the spinal cord or brain.

Haptotaxis Directed motion of cells along adhesion gradients of fixed substrates in the ECM, such as integrins.

"Slime trail motion" Cells secrete a non-diffusive substance; concentration gradients of the substance allow the cells to migrate towards already explored paths.

Somatic evolution Darwinian-type evolution that occurs on soma (as opposed to germ) cells and characterizes cancer progression [1].

Definition of the Subject

Cancer cells acquire characteristic traits in a step-wise manner during carcinogenesis. Some of these traits are autonomous growth, induction of angiogenesis, invasion and metastasis. In this chapter, the focus is on one of the

late stages of tumor progression: tumor invasion. Tumor invasion has been recognized as a complex system, since its behavior emerges from the combined effect of tumor cell-cell and cell-microenvironment interactions. Cellular automata (CA) provide simple models of self-organizing complex systems in which collective behavior can emerge out of an ensemble of many interacting "simple" components. Recently, cellular automata have been used to gain a deeper insight in tumor invasion dynamics. In this chapter, we briefly introduce cellular automata as models of tumor invasion and we critically review the most prominent CA models of tumor invasion.

Introduction

Cancer describes a group of genetic and epigenetic diseases characterized by uncontrolled growth of cells, leading to a variety of pathological consequences and frequently death. Cancer has long been recognized as an evolutionary disease [2]. Cancer progression can be depicted as a sequence of traits or phenotypes that cells have to acquire if a neoplasm (benign tumor) is to become an invasive and malignant cancer. A phenotype refers to any kind of observed morphology, function or behavior of a living cell. Hanahan and Weinberg [3] have identified six cancer cell phenotypes: unlimited proliferative potential, environmental independence for growth, evasion of apoptosis, angiogenesis, invasion and metastasis.

In this chapter, we focus on the invasive phase of tumor growth. Invasion is the main feature that allows a tumor to be characterized as malignant. The progression of a benign tumor and delimited growth to a tumor that is invasive and potentially metastatic is the major cause of poor clinical outcome in cancer patients, in terms of therapy and prognosis. Understanding tumor invasion could potentially lead to the design of novel therapeutical strategies. However, despite the immense amounts of funds invested in cancer research, the intracellular and extracellular dynamics that govern tumor invasiveness in vivo remain poorly understood.

Biomedically, invasion involves the following tumor cell *processes*:

- Tumor cell migration, which is a result of down-regulation of cadherins, that is loss of cell-cell adhesion,
- Tumor cell-extracellular matrix (ECM) interactions, such as cell-ECM adhesion, and ECM degradation or remodeling, by means of proteolysis. These processes allow for the penetration of the migrating tumor cells into host tissue barriers, such as basement and interstitial stroma [4], and
- Tumor cell proliferation.

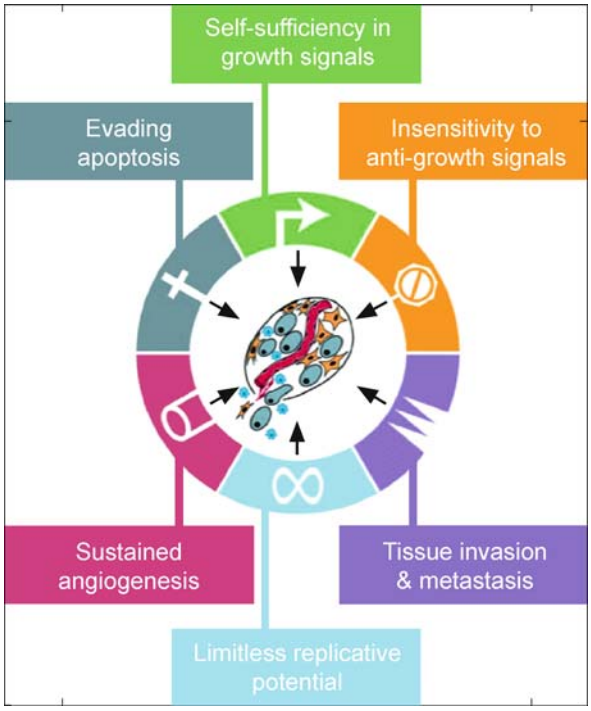

Cellular Automaton Modeling of Tumor Invasion, Figure 1
Hanahan and Weinberg have identified six possible types of cancer cell phenotypes: unlimited proliferative potential, environmental independence for growth, evasion of apoptosis, angiogenesis, invasion and metastasis (Reprinted from [3], with permission from the authors)

Tumor invasion facilitates the emergence of metastases, i. e. the spread of cancer cells to another part of the body and the formation of secondary tumors. Tumor invasion comprises a central aspect in cancer progression. However, invasive phenomena occur not only in pathological cases of malignant tumors but also during normal morphogenesis and wound healing.

Cancer research has been directed towards the understanding of tumor invasion dynamics and its implications in treatment design. In particular, research concentrates along the following *problems*:

Invasive tumor morphology: A wealth of empirical evidence links disease progression with tumor morphology [5]. The tumor morphology can indicate the degree of a tumor's malignancy. In particular, it is experimentally and clinically observed that a morphological instability is related to invasive solid tumors, producing finger-like spatial patterns. The question is which molecular and cellular mechanisms are responsible for this spatial pattern formation.

Cell migration and influence of the ECM: Important aspects of invading tumors are cell motion and the effect of the surrounding environment, especially the ECM [4].

Metabolism and acidosis: The multi-step process of carcinogenesis is often described by somatic evolution, wherein phenotypic properties are retained or lost depending on their contribution to the individual tumor cell survival and reproductive potential. One of the most prominent phenotypic changes involves the anaerobic glucose metabolism (glycolysis). A side-product of this metabolic activity is the production of H^+ ions that increase the pH of the tumor's microenvironment (acidosis). This gives rise to the questions: (i) Why does tumor evolution lead to this kind of metabolism, which is energetically deficient in comparison with the aerobic one? (ii) What are the advantages for the tumor? (iii) How do glycolytic tumor cells influence tumor invasion?

Emergence of invasion: Typically, tumor invasion appears during the late stages of carcinogenesis. Of ultimate importance is the question what are the mechanisms and the environmental conditions that trigger the progression from benign neoplasms to malignant invasive tumors.

Robustness: There are several questions concerning the stability and the resistance of tumor invasion such as: (i) Why are malignant tumor so robust (resistant) to perturbations (i. e. therapies)? (ii) Is it possible to design intelligent therapies (at the cellular level) that disturb the tumor's robustness? (iii) How can we investigate the tumor's robustness?

Mathematical modeling and analysis provide invaluable tools towards answering the above questions. Tumor invasion involves processes which occur at different spatio-temporal scales, including processes at the subcellular, cellular and tissue level. Mathematical models allow description and linking of these levels. One can distinguish *molecular*, *cellular* and *tissue* scales, respectively [6,7]:

- The molecular scale refers to phenomena at the subcellular level and concentrates on molecular interactions and resulting phenomena, such as alterations of signaling cascades and cell cycle control, gene mutations, etc. In tumor invasion, the down-regulation of cadherins provides an example of a molecular process.
- The cellular scale refers to cellular interactions and therefore to the most prominent dynamics of cell populations, e. g. adhesion, contact inhibition, chemotaxis etc.
- The tissue scale focuses on tissue level processes taking into account macroscopic quantities, such as volumes, flows etc. Continuum phenomena include cell convection and diffusion of nutrients and chemical factors, mechanical stress and the diffusion of metastases.

For example, genetic alterations may lead to invasive cells (molecular scale) that are able to migrate (cellular scale) and interact with diffusible or non-diffusible signals (tissue scale). Models that deal with phenomena at multiple scales are called multi-scaled.

Recently, a variety of mathematical models have been proposed to analyze different aspects of tumor invasion. Deterministic macroscopic models are used to model the spatio-temporal growth of tumors, usually assuming that tumor invasion is a wave propagation phenomenon [8,9,10,11,12]. Computational investigations of the invasiveness of glioma tumors illustrate that the ratio of tumor growth and spatial anisotropy in cell motility can quantify the degree of tumor invasiveness [13,14]. Whilst these models are able to capture the tumor structure at the tissue level, they fail to describe the tumor at the cellular and the sub-cellular levels. Lately, multi-scale approaches attempt to describe and predict invasive tumor morphologies, growth and phenotypical heterogeneity [15,16].

Cellular automata (CA), and more generally cell-based models, provide an alternative modeling approach, where a micro-scale investigation is allowed through a stochastic description of the dynamics at the cellular level [17]. In particular, CA define an appropriate modeling framework for tumor invasion since they allow for the following:

- CA rules can mimic the processes at the cellular level. This fact allows for the modeling of an abundance of experimental data that refer to cellular and sub-cellular processes related to tumor invasion.
- The discrete nature of CA can be exploited for investigations of the boundary layer of a tumor. Bru et al. [18] have analyzed the fractal properties of tumor surfaces (calculated by means of fractal scaling analysis) which can be compared with corresponding CA simulations to gain a better understanding of the tumor phenomenon. In addition, the discrete structure of CA facilitates the implementation of complicated environments without any of the computational problems characterizing the simulation of continuous models.
- Motion of tumor cells through heterogeneous media (e. g. ECM) involves phenomena at various spatial and temporal scales [19]. These cannot be captured in a purely macroscopic modeling approach. Alternatively, discrete microscopic models, such as CA, can in-

corporate different spatio-temporal scales and they are well-suited for simulating such phenomena.
- CA are paradigms of parallelizable algorithms. This fact makes them computationally efficient.

In the following section, we provide a definition of CA. In Sect. "Models of Tumor Invasion", we review the existing CA models for central processes of tumor invasion. Finally, the discussion, we critically discuss the use of CAs in tumor invasion modeling and we identify future research questions related to tumor invasion.

Cellular Automata

The notion of a *cellular automaton* originated in the works of John von Neumann (1903–1957) and Stanislaw Ulam (1909–1984). Cellular automata may be viewed as simple models of self-organizing complex systems in which collective behavior can emerge out of an ensemble of many interacting "simple" components. In complex systems, even if the basic and local interactions are perfectly known, it is possible that the global behavior obeys new laws that cannot be obviously extrapolated from the individual properties, as if the whole is more than the sum of the parts. This property makes cellular automata a very interesting approach to model complex systems in physics, chemistry and biology (examples are introduced in [17,20]). A CA can be defined as a 4-tuple $(\mathcal{L}, S, \mathcal{N}, \mathcal{F})$, where:

- $\mathcal{L}$ is an infinite regular *lattice* of nodes (discrete space),
- S is a finite set of *states* (discrete states); each cell $i \in \mathcal{L}$ is assigned a state $s \in S$,
- $\mathcal{N}$ is a finite set of *neighbors*,
- $\mathcal{F}$ is a deterministic or probabilistic map

$$\mathcal{F} : S^{|\mathcal{N}|} \to S \tag{1}$$

$$\{s_i\}_{i \in \mathcal{N}} \mapsto s \,, \tag{2}$$

which assigns a new state to a node depending on the state of all its neighbors indicated by $\mathcal{N}$ (*local rule*).

The evolution of a CA is defined by applying the function $\mathcal{F}$ synchronously to all nodes of the lattice $\mathcal{L}$ (*homogeneity* in space and time).

The above features can be extended, giving rise to several variants of the classical CA notion [21]. Some of these are:

Asynchronous CA: in such CA, the restriction of simultaneous update of all the nodes is revoked, allowing for asynchronous update.

Non-homogeneous CA: this variation allows the transition rules to depend on node position. Agent-based models are "relatives" of CA that lost the homogeneity property, i. e. each individual-particle may have its own set of rules.

Coupled-map lattices: in this case the constraint of discrete state space is withdrawn, i. e. the state variables are assumed to be continuous. An important type of coupled-map lattices are the so-called Lattice Boltzmann models [22].

Structurally dynamic CA: in these systems, the underlying lattice is no longer a passive static object but becomes a dynamic component. Therefore, the lattice, structure evolves depending on the values of the nodes state variables.

Models of Tumor Invasion

This section reviews the existing cellular automata models of tumor invasion. Categorizing these models is a nontrivial task. Moreover, existing CA models describe tumor invasion at more than one scale (sub-cellular, cellular and tissue). In this review, we distinguish models that analyze: (i) the invasive morphology, (ii) tumor cell migration and the influence of the ECM, (iii) metabolism and acidosis, and (iv) the emergence of tumor invasion.

Invasive Tumor Morphology

The tumor morphology arising from the spatial pattern formation of the tumor cell population has been recognized as a very important aspect of tumor growth. Several researchers have attempted to reveal the mechanisms of spatial pattern formation of invasive tumors. Here, we present the most representative CA models for the invasive tumor morphology.

Effects of Directed Cell Motion Sander et al. [23] developed a CA model to investigate the branching morphology of invasive brain tumors. In the model tumor cell motion is influenced by two key processes: (i) chemotaxis, and (ii) "slime trail following". A typical example of a slime trail following mechanism is found in the motion of certain myxobacteria [24].

The authors show that the branching morphology of tumors can be explained as a result of chemotaxis and "slime trail following". In particular, simulations reproduce the branching pattern formation observed in vitro cultures of glioma cells. However, the assumption of slime trail following has not been proven biologically as yet.

Spatial Structure of Invasive Tumors Anderson [15,25] proposed a model to examine the effects of tumor cell heterogeneity (at the genetic level) on the spatial morphol-

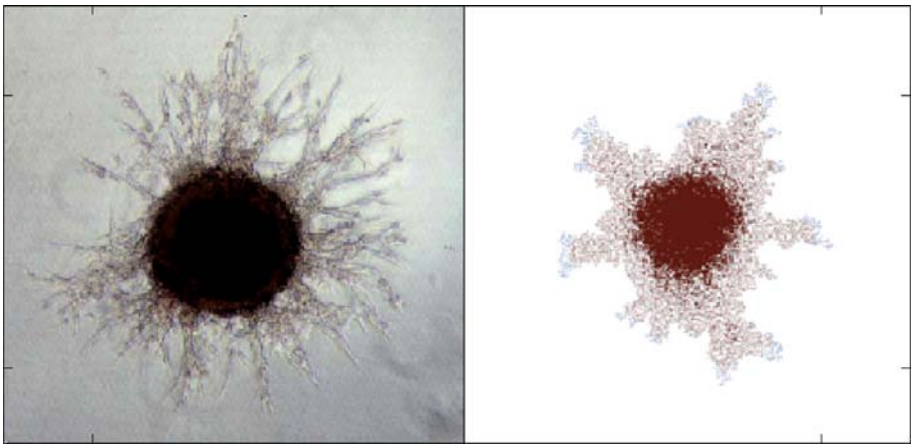

Cellular Automaton Modeling of Tumor Invasion, Figure 2
***Left*: Microscopy image of a multicellular tumor spheroid, exhibiting an extensive branching system that rapidly expands into the surrounding extracellular matrix gel. These branches consist of multiple invasive cells. (Reprinted from [26] with permission). *Right*: Simulation of Anderson's model [15] reproducing the experimentally observed morphology of invasive tumors**

ogy and to analyze the importance of cell-cell and cell-ECM adhesion. The model assumes a non-diffusible, fixed configuration of ECM. The extracellular matrix can be degraded by diffusible enzymes, such as metallo-proteinases, produced by tumor cells. Moreover, cells are allowed to mutate and evolve their phenotype from proliferative to invasive. Finally, an oxygen concentration field plays the role of nutrients in the model.

Simulations of the model show that: (i) the ECM heterogeneity is mainly responsible for the tumor branching morphology (Fig. 2), (ii) cell-cell adhesion plays an important role only in the early stages of tumor development, (iii) invasive tumor cells are located at the boundary of the tumor, and (iv) the tumor is a phenotypically heterogeneous object.

Tumor Cell Migration and the Influence of the Extracellular Matrix

Cell migration and cell-ECM interactions are two of the most crucial invasion-related processes. Cellular automata provide an appropriate framework to model and analyze the effect of cell motility and cell-environment interactions of tumor cell migration.

The Role of Cell-Cell and Cell-ECM Adhesion Turner and Sheratt [27] proposed a cellular Potts model [28] to investigate how cell-cell and cell-ECM adhesion influence the tumor invasion depth and tumor morphology. A cellular Potts model can be viewed as an extension of the CA idea allowing to analyze phenomena that take into account specific cell shapes. Cells are assumed to move according to intercellular adhesive interactions and haptotactical gradients. Moreover, cells are allowed to proliferate, while mitotic probabilities depend on the strength of the adhesive interaction. Finally, cells are assumed to secrete proteolytic enzymes that degrade the ECM.

The authors show that adhesive dynamics can explain the "fingering" patterns observed in their simulations. Moreover, the authors demonstrate that the width of the invasion zone depends less on cell-cell adhesion and more on cell-ECM adhesion facilitated by haptotaxis and proteolysis.

Cellular Mechanisms of Glioma Cell Migration In the work of Aubert et al. [29], a CA model is introduced that allows for the investigation of tumor cell migration, based on experimentally observed density profiles of glioma cell cultures. The goal is to identify the mechanisms of tumor (glioma) cell motion, which play a crucial role in tumor invasion. The authors do not consider proliferation of tumor cells. Only the influence of tumor cell migration and intercellular interactions are studied. The authors introduce and test two distinct cell mechanisms: (i) cell-cell adhesion, and (ii) a kind of "inertia" in cell motion, i. e. the cells tend to maintain the direction of their motion.

The authors carefully scale the model according to the experimental setup and calibrate the corresponding model parameters. The simulation results indicate that cell-cell adhesion can explain the experimental results. It is concluded that cell-cell adhesion is an important process in glioma cell migration.

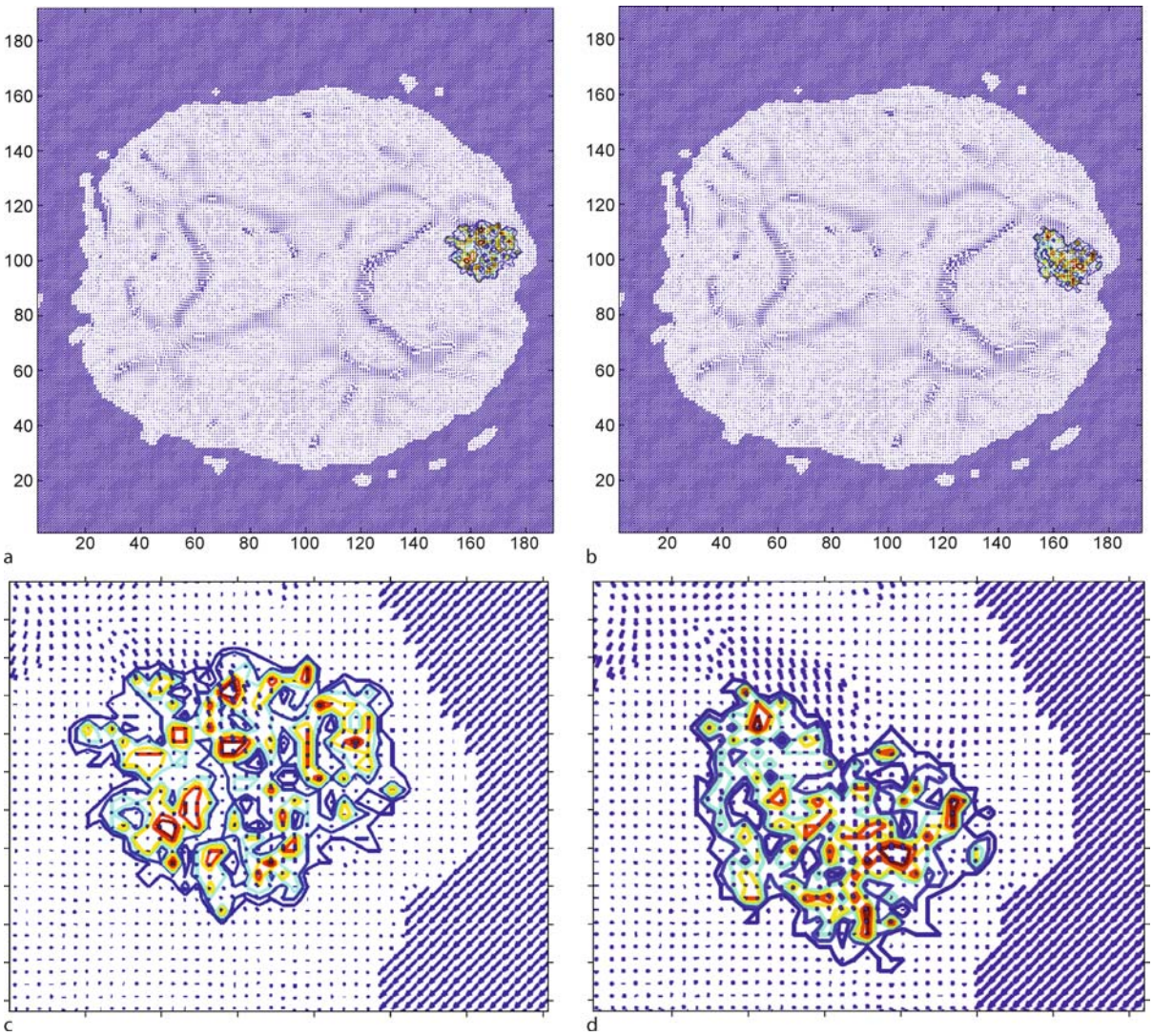

Cellular Automaton Modeling of Tumor Invasion, Figure 3
The effect of the brain's fiber tract on glioma growth. a A simulation is shown without taking into account the influence of fiber tracts. b The fiber tracts in the brain strongly drive the evolution of the tumor growth. c–d Figures display a close up of the tumor area of the top a–b simulations (Reprinted from [31])

Effects of Fiber Tracts on Glioma Invasion Wurzel et al. [30] model glioma tumor invasion with a lattice-gas cellular automaton (LGCA) [17]. The authors address the question of how fiber tracts found in the brain's white matter influence the spatio-temporal evolution and the invading front morphology of glioma tumors. Cells are assumed to move, proliferate and undergo apoptosis according to corresponding stochastic processes. Fiber tracts are represented as a local gradient field that enhances cell motion in a specific direction.

The authors develop and analyze different scenarios of fiber tract influence. A gradient field may increase the speed of the invading tumor front. For high field intensities the model predicts the formation of cancer islets at distances away from the main tumor bulk. The simulated invasion patterns qualitatively resemble clinical observations.

Effect of Heterogeneous Environments on Tumor Cell Migration Hatzikirou et al. [31] developed a LGCA model to investigate the influence of heterogeneous environments on tumor cell dispersal. This model is a simplified version of [30] which facilitates the mathematical analysis. In this study no proliferation or death of cells

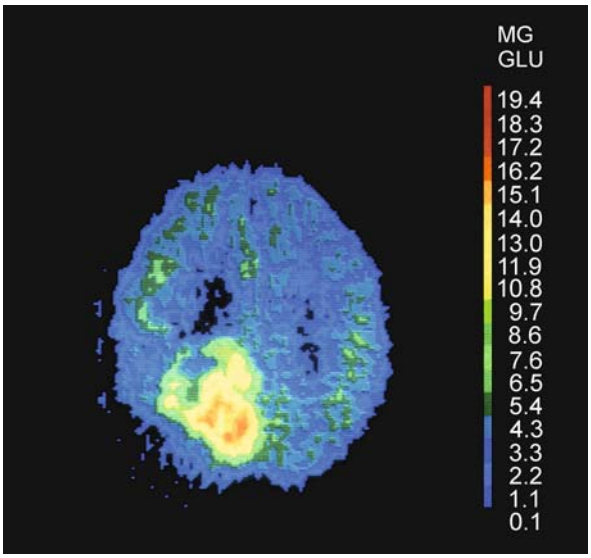

Cellular Automaton Modeling of Tumor Invasion, Figure 4
Typically, tumors exhibit abnormal levels of glucose metabolism. Positron emission imaging (PET) techniques localize the regions of abnormal glycolytic activity and identify the tumor locus

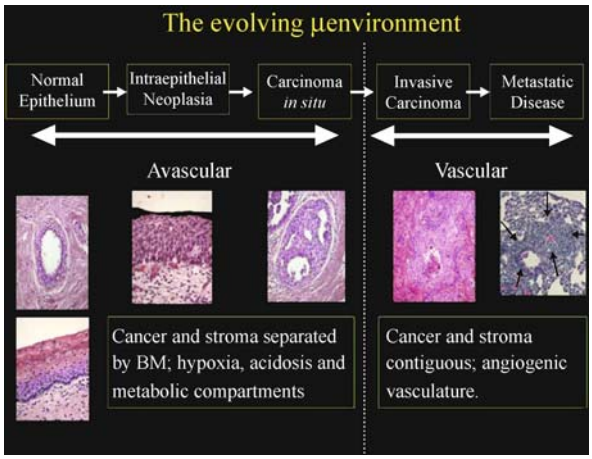

Cellular Automaton Modeling of Tumor Invasion, Figure 5
The evolving microenvironment of breast cancer. The multiple stages of breast carcinogenesis are shown progressing from left to right, along with histological representations of these stages. As indicated, the pre-invasive stages occur in an avascular environment, whereas cancer cells have direct access to vasculature following invasion. (Reprinted from [33])

is considered. The authors distinguish two kinds of cell-ECM interactions (i) cell-ECM adhesion leading to haptotactical motion along integrin concentration gradients (environment with directional information) and (ii) contact guidance that promotes the alignment along ECM pores or fibres as seen in Fig. 3 (environment with orientational information).

In this study, it is investigated the impact of both types of cell-ECM interaction on tumor cell motion. In particular, macroscopic dispersal measures (such as mean cell flux) depending on cellular and environmental parameters are calculated. Accordingly, the models allow for prediction of cell motion in different environments.

Metabolism and Acidosis

In the course of cancer progression, tumor cells undergo several phenotypic changes in terms of motility, metabolism and proliferative rates. In particular, it is important to analyze the effect of the anaerobic metabolism of tumor cells and the acidification of the environment (as a side-product of glycolysis) on tumor invasion (Fig. 4).

Patel et al. [32] proposed a model of tumor growth to examine the roles of native tissue vascularity and anaerobic metabolism on the growth and invasion efficacy of tumors. The model assumes a vascularized host tissue. Anaerobic metabolism involves the consumption of glucose and the production of H^+ ions, leading to the acidification of the local tissue. The vascular network allows for the absorption of H^+ ions. Cells are assumed to be proliferative and non-motile. The pH level, i. e. is the H^+ concentration, and the glucose concentration determine the survival and death of the cells.

Simulations of the model show: (i) high tumor H^+ ion production favors tumor invasion by the acidification of the neighboring host tissue, and (ii) there is an optimal density of microvessels that maximizes tumor growth and invasion, by minimizing the acidification effects on tumor cell proliferation (absorption of H^+ ions) and maximizing the negative effect of H^+ ions on the neighboring tissue.

Emergence of Tumor Invasion

Recently, several models have been proposed that concentrate on the evolutionary dynamics of tumors (Fig. 5). The main goal of these models is to understand under which environmental conditions particular phenotypes appear. Here, we review those models that focus on the mechanisms that allow the emergence of invasive behavior.

Influence of Metabolic Changes Smallbone et al. [34] developed an evolutionary CA model to investigate the cell-microenvironmental interactions that mediate somatic evolution of cancer cells. In particular, the authors investigate the sequence of tumor phenotypes that ultimately leads to invasive behavior. The model considers

three phenotypes, (i) the hyperplastic phenotype that allows growth away from the basement membrane, (ii) the glycolytic phenotype that allows anaerobic metabolism (the "fuel" is glucose), and (iii) the acid-resistant phenotype that enables the cell to survive in low pH. Cells are allowed to proliferate, die or adapt (change their phenotype). No cell motion is explicitly considered.

The model predicts three phases of somatic evolution: (i) Initially, cell survival and proliferation are dependent on the oxygen concentration. (ii) When the oxygen becomes scarce, the glycolytic phenotype confers a significant proliferative advantage. (iii) The side-products of glycolysis, e. g. galactic acid, increase the microenvironmental pH and promote the selection of acid-resistant phenotypes. The latter cell type is able to invade the neighboring tissue since it takes advantage of the death of host cells, due to acidification, and proliferates using the available free space.

The Game of Invasion Recently, Basanta et al. [35] have developed a game theory inspired CA that addresses the question of how invasive behavior emerges during tumor progression (see also [36]). The authors study the circumstances under which mutations that confer increased motility to cells can spread through a tumor composed of rapidly proliferating cells. The model assumes the existence of only two phenotypes: "proliferative" (high division rate and no motility) and "migratory" (low division rate and high motility). Mutations are allowed for by the random change of phenotypes. Nutrients are assumed to be uniformly distributed over the lattice.

Simulations show that low nutrient conditions confer a reproductive advantage to motile cells over the proliferative ones. The model suggests novel ideas for therapeutic strategies, e. g. by increasing the oxygen supply around the tumor to favor the reproduction of proliferative cells over the migrating ones. This is not necessarily a therapy since there are benign tumors that are life threatening even if they do not become invasive. Despite that, in most cases a growing but non-aggressive tumor will have a much better prognosis than a smaller but invasive one.

Discussion

In this review, we have focused on one of the most important aspects of cancer progression: tumor invasion. The main processes involved in tumor invasion are related to tumor cell migration, cell-ECM interactions, especially ECM degradation/remodeling and tumor cell proliferation. These processes are evolving at different scales, e. g. cell-ECM adhesion is the response of tumor cells to ECM integrins (molecular level) leading to a haptotactical cell motion (cellular level) and influencing the tumor morphology (macroscopic level). Therefore, in order to understand tumor invasion dynamics, it is important to use mathematical tools that allow for modeling sub-cellular or cellular processes and to analyze the emergent macroscopic behavior. Individual-based models, especially CA, are well-suited for this task. Moreover, some types of CA models, such as lattice-gas cellular automata [17,31], facilitate analytical investigations allowing for deeper insight into the modeled phenomena.

In this chapter, we reviewed the existing CA models of tumor invasion. The presented models explore central aspects of tumor invasion. Some of the models are in good agreement with biomedical observations for in-vitro and in-vivo tumors. In the following, we list the most interesting biological insights that can be gained from the reviewed models:

- The significance of hypoxia in the process of tumor progression: Activation of glycolysis and acidification of the host tissue facilitate tumor invasion. Low nutrient conditions, such as hypoxia, may trigger invasive behaviors.
- Cell-cell adhesion: It is evident that intercellular adhesion has a great impact in the early stages of tumor growth. However, in tumor invasion the role of cell-cell adhesion is minor, since mainly the cell-ECM interactions appear to dictate the tumor cell behavior.
- Cell-ECM adhesion: This is an important process for tumor invasion. In particular, the heterogeneous structure of the ECM strongly influences the spatial morphology of invasive tumors.

Mathematical modeling offers potentially significant insight into tumor invasion. Several crucial questions have not been adequately addressed so far by modeling efforts:

Branching morphology: Several mechanisms have been proposed that lead to branching patterns, e. g. diffusion-limited aggregation, the interplay of cell-cell and cell-ECM adhesion, as well as chemotaxis or slime trail following motion. However, biologists and modelers have not yet identified a unique mechanism that drives the branching morphology of invasive tumors.

Go or grow: The mechanisms of invasive tumor cell migration are still not understood. Recently, Fedotov et al. [37] have analyzed the effect of a postulated migration/proliferation dichotomy on cell migration.

Emergence: Concerning the emergence of invasion in tumor progression little is known. Mechanisms related to tumor cell motion and other cell processes, such as proliferation (migration/proliferation dichotomy),

may play an important role for the dominance of invasive phenotypes [35,38].

Angiogenesis: Another open issue is the influence of angiogenesis and vasculogenesis on tumor invasion. Despite significant efforts to describe the mechanisms of angio- and vasculogenesis, little is known about the effect of these processes on tumor invasion [32].

Robustness: The identification of cellular mechanisms that are responsible for tumor robustness remains significant challenge.

Finally, for clinical purposes, future models should be able to provide accurate and quantitative predictions. Simplified models considering only the essential ingredients for tumor growth, and especially tumor invasion, but validated with actual clinical data may be helpful in this regard. We sincerely hope that a more profound knowledge of important tumor characteristics, such as tumor invasion, will eventually lead to the design of more effective therapeutic strategies.

Acknowledgments

We are grateful to D. Basanta, L. Brusch, A. Chauviere, E. Flach and F. Peruani for the comments and the fruitful discussions. We acknowledge support from the systems biology network HepatoSys of the German Ministry for Education and Research through grant 0313082J. Andreas Deutsch is a member of the DFG Research Center for Regenerative Therapies Dresden – Cluster of Excellence – and gratefully acknowledges support by the Center. The research was supported in part by funds from the EU Marie Curie Network "Modeling, Mathematical Methods and Computer Simulation of Tumor Growth and Therapy" (EU-RTD IST-2001-38923). Finally, the authors would like to thank for the financial support of the Gottfried Daimler- and Karl Benz foundation through the project "Biologistics: From bio-inspired engineering of complex logistical systems until nanologistics" (25-02/07).

Bibliography

1. Bodmer W (1997) Somatic evolution of cancer cells. J R Coll Physicians Lond 31(1):82–89
2. Nowell PC (1976) The clonal evolution of tumor cell populations. Science 4260 194:23–28
3. Hanahan D, Weinberg R (2000) The hallmarks of cancer. Cell 100:57–70
4. Friedl P (2004) Prespecification and plasticity: shifting mechanisms of cell migration. Curr Opin Cell Biol 16(1):14–23
5. Sanga S, Frieboes H, Zheng X, Gatenby R, Bearer E, Cristini V (2007) Predictive oncology: multidisciplinary, multi-scale in-silico modeling linking phenotype, morphology and growth. Neuroim 37(1):120–134
6. Hatzikirou H, Deutsch A, Schaller C, Simon M, Swanson K (2005) Mathematical modelling of glioblastoma tumour development: a review. Math Mod Meth Appl Sc 15(11):1779–1794
7. Preziozi L (ed) (2003) Cancer modelling and simulation. Chapman & Hall CRC Press
8. Marchant BP, Norbury J, Perumpanani AJ (2000) Traveling shock waves arising in a model of malignant invasion. SIAM J Appl Math 60(2):263–276
9. Perumpanani AJ, Sherratt JA, Norbury J, Byrne HM (1996) Biological inferences from a mathematical model of malignant invasion. Invas Metast 16:209–221
10. Perumpanani AJ, Sherratt JA, Norbury J, Byrne HM (1999) A two parameter family of travelling waves with a singular barrier arising from the modelling of extracellular matrix mediated cellular invasion. Phys D 126:145–159
11. Sherratt JA, Nowak MA (1992) Oncogenes, anti-oncogenes and the immune response to cancer: a mathematical model. Proc Roy Soc Lond B 248:261–271
12. Sherratt JA, Chaplain MAJ (2001) A new mathematical model for avascular tumour growth. J Math Biol 43:291–312
13. Swanson KR, Alvord EC, Murray J (2002) Quantifying efficacy of chemotherapy of brain tumors (gliomas) with homogeneous and heterogeneous drug delivery. Acta Biotheor 50:223–237
14. Jbabdi S, Mandonnet E, Duffau H, Capelle L, Swanson K, Pelegrini-Issac M, Guillevin R, Benali H (2005) Simulation of anisotropic growth of low-grade gliomas using diffusion tensor imaging. Magn Res Med 54:616–624
15. Anderson A, Weaver A, Cummings P, Quaranta V (2006) Tumor morphology and phenotypics evolution driven by selective pressure from the microenvironment. Cell 127:905–915
16. Frieboes H, Lowengrub J, Wise S, Zheng X, Macklin P, Bearer E, Cristini V (2007) Computer simulation of glioma growth and morphology. Neuroim 37(1):59–70
17. Deutsch A, Dormann S (2005) Cellular Automaton Modeling of Biological Pattern Formation. Birkhauser, Boston
18. Bru A, Albertos S, Subiza JL, Lopez Garcia-Asenjo J, Bru I (2003) The universal dynamics of tumor growth. Bioph J 85:2948–2961
19. Lesne A (2007) Discrete vs continuous controversy in physics. Math Struct Comp Sc 17:185–223
20. Chopard B, Dupuis A, Masselot A, Luthi P (2002) Cellular automata and lattice Boltzmann techniques: an approach to model and simulate complex systems. Adv Compl Syst 5(2):103–246
21. Moreira J, Deutsch A (2002) Cellular automaton models of tumour development: a critical review. Adv Compl Syst 5:1–21
22. Succi S (2001) The lattice Boltzmann equation: for fluid dynamics and beyond. Series Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford
23. Sander LM, Deisboeck TS (2002) Growth patterns of microscopic brain tumours. Phys Rev E 66:051901
24. Wolgemuth CW, Hoiczyk E, Kaiser D, Oster GF (2002) How myxobacteria glide. Curr Biol 12(5):369–377
25. Anderson ARA (2005) A hybrid model of solid tumour invasion: the importance of cell adhesion. Math Med Biol 22:163–186
26. Habib S, Molina-Paris C, Deisboeck TS (2003) Complex dynamics of tumors: modeling an emerging brain tumor system with coupled reaction-diffusion equations. Phys A 327:501–524
27. Turner S, Sherratt JA (2002) Intercellular adhesion and cancer invasion: A discrete simulation using the extended Potts model. J Theor Biol 216:85–100

28. Graner F, Glazier J (1992) Simulation of biological cell sorting using a two-dimensional extended Potts Model. Phys Rev Lett 69:2013–2016
29. Aubert M, Badoual M, Freol S, Christov C, Grammaticos B (2006) A cellular automaton model for the migration of glioma cells. Phys Biol 3:93–100
30. Wurzel M, Schaller C, Simon M, Deutsch A (2005) Cancer cell invasion of normal brain tissue: Guided by Prepattern? J Theor Med 6(1):21–31
31. Hatzikirou H, Deutsch A (2008) Cellular automata as microscopic models of cell migration in heterogeneous environments. Curr Top Dev Biol 81:401–434
32. Patel A, Gawlinski E, Lemieux S, Gatenby R (2001) Cellular automaton model of early tumor growth and invasion: the effects of native tissue vascularity and increased anaerobic tumor metabolism. J Theor Biol 213:315–331
33. Gillies RJ, Gatenby RA (2007) Hypoxia and adaptive landscapes in the evolution of carcinogenesis. Canc Metast Rev 26:311–317
34. Smallbone K, Gatenby R, Gillies R, Maini P, Gavaghan D (2007) Metabolic changes during carcinogenesis: Potential impact on invasiveness. J Theor Biol 244:703–713
35. Basanta D, Hatzikirou H, Deutsch A (2008) The emergence of invasiveness in tumours: a game theoretic approach. Eur Phys J B 63:393–397
36. Basanta D, Simon M, Hatzikirou H, Deutsch A (2009) An evolutionary game theory perspective elucidates the role of glycolysis in tumour invasion. Cell Prolif (to appear)
37. Fedotov S, Iomin A (2007) Migration and proliferation dichotomy in tumor-cell invasion. Phys Rev Let 98:118101–4
38. Hatzikirou H, Basanta B, Simon M, Schaller C, Deutsch A (2009) "Go or Grow": the key to the emergence of invasion in tumor progression? (under submission)

Cellular Computing

Christof Teuscher
Los Alamos National Laboratory, Los Alamos, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Computation and Computability
The Machine as a Cell
Cellular Computing Hardware
The Cell as a Machine
Future Directions
Bibliography

Glossary

Cell The biological cell is the smallest self-contained, self-maintaining, and self-reproducing unit of all living organisms. Various computing paradigms were inspired by the biological cell.

Molecular computing A subfield of cellular computing, where the molecules instead of the cell play a central functional role.

Computing The science that deals with the manipulation of symbols. Also refers to the processes carried out by real or abstract computers.

Computation Synonymous with information processing or also algorithm. Computations can for example be performed by abstract machines, real computer hardware, or biological systems. The abstract concept of the Turing machine separates the class of computable from the class of non-computable functions.

Parallel computing Parallel computing involves the execution of a task on multiple processors with the goal to speed-up the execution process by dividing up the task into smaller sub-tasks that can be executed simultaneously.

Definition of the Subject

The field of *cellular computing* (abbreviated, CC) defines both a general computing framework and a discipline concerned with the analysis, modeling, and engineering of real cellular processes for the purpose of computation. The biological cell, discovered and coined by R. Hooke in 1665, is the smallest self-contained, self-maintaining, and self-reproducing unit of all living organisms. Its understanding and modeling is crucial both for the understanding of life and for the ability to use, control, and modify its complex bio-chemical processes to perform specific functions for the purpose of *in vivo* or *in vitro* computation. The cellular metaphor has inspired and influenced numerous, both abstract computing models and *in silico* implementations, such as *cellular automata* (abbreviated, CA), *membrane systems* (or P systems), or *Field Programmable Gate Arrays* (abbreviated, FPGAs), with the main purpose to solve algorithmic problems in alternative ways. The cellular computing approach is appealing because the cell provides a convenient level of abstraction and functionality, is reasonably simple with enough abstractions, and can be used as a building block to compose more complex systems. On the other hand, a wide range of computational approaches are used to model and understand the functioning of the inter- and intra-cellular processes of biological cells on its various levels of complexity. The cell's bio-chemical processes can either directly be interpreted as computations or they can be modified for the specific purposes of computations through bio-engineering methods. The hope in using biological cells as computing de-

vices is to (1) ultimately go beyond complexity and fabrication limits currently encountered with traditional silicon circuits, (2) to embed computing capabilities into living organisms for autonomous therapeutic and diagnostic applications, and (3) to realize integrated biological sensing applications that use living organisms, such as for example plants and bacteria, as intelligent sensors.

Introduction

The English naturalist Robert Hooke is commonly accredited for the discovery of the cell [7] and the coining of its term, which he borrowed from the Latin word *cella*, that designates a small chamber. Hooke's work was published in 1665 in his book *Micrographia* [45], which contained observations he made with a very rudimentary microscope of dead cork cells. Since then, tremendous discoveries have been made around the cell, ranging from the pioneering work of German anatomist Walther Flemming [38,75] on cell division (mitosis) to the discovery of the DNA double helix structure by Watson and Crick in 1953 [117]. Nowadays, a sheer endless amount of knowledge and data on cells is available, originating from genetics, biochemistry, and molecular dynamics research. Nevertheless, the computational modeling at the molecular level of complete cells has been notoriously difficult because of the enormous complexity of the real cell, a lack of qualitative understanding, and the large number of unknowns in the models.

With the advent of modern computer science in the 1940s under pioneering fathers such as John von Neumann and Alan M. Turing, biology and computer science started to grow closer together. Biologists increasingly used computers to analyze experimental data and to model real biological organisms, while computer scientists began to draw inspiration from biology with the essential goal to engineer better and more "lifelike" machines. Using computer science, mathematics, statistics, machine learning, and artificial intelligence as tools to solve real problems in biology is commonly known under the term *bioinformatics* [11], while *computational biology* is more concerned with the exploration and discovery of new knowledge, for example by testing specific hypotheses by means of computer models. *Systems biology* [48], on the other hand, focuses on the system level and how the participating biological components interact. This is typically done in a combined approach, which involves theory, computational modeling, and experiments.

Living organisms are complex systems exhibiting a range of desirable characteristics, such as evolution, adaptation, and fault tolerance, that have proved difficult

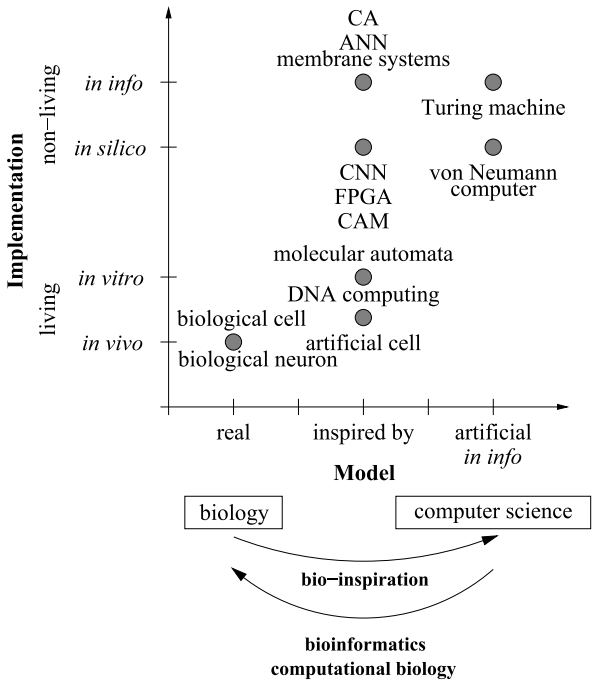

Cellular Computing, Figure 1
Illustration of the various possibilities to implement cellular models that range from realistic to abstract. Drawing inspiration from biology to build better computing machines is known under the term *bio-inspiration*, while using machines to address problems in biology is known as *bioinformatics* and *computational biology*

to realize using traditional engineering methodologies. Drawing inspiration from biology to design better and both more "lifelike" machines and algorithms is generally known under the term *biologically-inspired computer science* [39,59,90] (also *nature-inspired*, *bio-mimetic*, or *biologically-motivated*). Typically, this approach involves simulations, algorithms, and classical silicon-based circuits, but not living matter. The goal is not to copy nature, but to solely draw inspiration from it, while the field of *artificial life* [52] is more concerned with understanding life by building it. Figure 1 illustrates the various possibilities to implement cellular models that range from realistic to abstract.

The broad and very interdisciplinary field of *cellular computing*, in its natural and artificial dimension, defines both a

- General computing framework [89] and a
- Discipline concerned with the analysis, modeling, and engineering of real cellular processes for the purpose of computation [8].

As opposed to pure bio-inspired computing, cellular computing thus also includes the use of real biological cells for information processing. Cellular computing encompasses abstract theory, simulations, models, experiments, and includes both the paradigm of "the cell as a machine" and "the machine as a cell." Sometimes, cellular computing is used interchangeably with *molecular computing* [86], however, while molecular computing is concerned with the information processing in which molecules play a central functional role, cellular computing looks at the cell as the functional building block for more complex systems. Naturally, molecules are involved in cellular information processing as well, so we therefore consider molecular computing as a subset of cellular computing.

The birth hour of *connectionism* in the 1940s can be considered as the first occurrence of looking at cells as automata and modeling them by digital systems. The well known McCulloch–Pitts neuron [60] simulated a wealth of research in neural information processing and modeling that continues with increasing activity until today [9]. In 1948 (but first published in 1969, reprinted in [111]), Alan Turing proposed his own connectionist ideas [103,110], which were highly influenced by the digital systems paradigms. His neurons were simple NAND (not AND) logical gates, which would thus allow in principle to compute any logical function, provided enough neurons are available and that they can be interconnected in arbitrary ways.

The concept of molecular automata and of cells seen as information processing devices appeared around the 1960s. In their studies of the control of gene expression and the synthesis of proteins, Jacob and Monod [46,62], Davis [26], and others emphasized deterministic operations at the level of DNA, RNA, and ribosomes. Early work on protein synthesis, for example by Warner et al. [116], suggested that ribosomes move along the messenger RNA, not unlike the head reading the tape of a Turing machine. In 1961, Pattee [71] used the concept of the Turing machine computation to explain the generation of growing macromolecular sequences and rhetorically asked in the title of a section of his paper "Can molecules compute?" Stahl and Goheen [95] further pioneered the idea of seeing the cell as an information processor and Turing machine. They considered enzymes as computational primitives whose operations are simulated by a Turing machine, and are therefore computable in the formal sense of the term [27]. At about the same time, Sugita considered some aspects of molecular systems by using a logical circuit equivalent [98,99,100]. He states: "The mechanism of life shall be clarified by using electronic analogues, because the chemical system composing the living organism is nothing but a complicated network of rate processes from the chemical or physical point of view. Simulation of such a network may be done either by using an analogue computer or a digital differential analyzer, as has been done at the laboratory[...]. A digital system composed of some chemical reactions may play the role of a controlling computer of the living chemical plant, and the reactions having analogue behavior may be controlled by that computer as well as by analogue control systems which can be seen in primitive control" [98]. Schmitt [83] discussed the biological memory of macromolecular automata in his 1962 book, while Feynman [36] elaborated in his famous talk "There's plenty of room at the bottom" on the possibility to compute with atoms, molecules, and cells. Feynman was more focused on the miniaturization of the computer by storing and processing information at atomic levels, but his ideas were pursued by many others later (e. g. [14,21,86]), and ultimately led to what we know today as *nano-* and *molecular electronics* [24], but also greatly stimulated the earlier cellular computing research.

Michael Conrad, a pioneer in biologically-inspired, molecular, and cellular computing, as well as many others observed that living organisms process information in a very different way than digital computers [20]. Organisms typically exploit the inherent physical properties of the matter they are made of, while digital computers are based on multiple levels of abstractions, which are typically far away from the physical substrate and do not directly exploit any of its specific physical characteristics for the purpose of more efficient computation. Cellular computing, both in its natural and artificial dimension, tries to address these issues by drawing inspiration from real cells and by using them for the purpose of computation.

The principal reason for the increasing interest in cellular computing paradigms seem twofold. Firstly, Moore's Law [63] has dominated the progress of electronic circuits in the past decades, but fundamental limits of silicon-based technology now begin to become serious showstoppers [47]. The hope is that cellular computing will help to go beyond such limits by using alternative computing paradigms inspired by the cell or by using real cells, molecules, and atoms as information processing devices. Secondly, despite lots of progress, Turing's hope that "machines will eventually compete with men in all purely intellectual fields" [108] is far from accomplished. Biological systems outperform their silicon counterparts in various areas, such as for example with their ability to adapt, to recognize objects, or to self-repair. Again, by using cellular computing paradigms, the hope is that such bio-inspired

approaches will ultimately allow to go beyond existing limits.

Computation and Computability

Both from the perspective of computer science and biological organisms, a central question is what kind of operations a given system can perform, what kind of problems it can solve, and how efficiently it can do this. Despite their ubiquitousness and seemingly endless power, there are clear and fundamental limitations – that are too often ignored unfortunately – to what computers can do [43,55]. Two fields of computer science are relevant in this context: *computability theory* [27,61] deals with what problems can be solved, while *computational complexity theory* [68] deals with how efficient they can be solved.

Since Alan M. Turing's seminal paper on the abstract concept of the *Turing machine* (abbreviated, TM) [107], *computability* is well defined and we generally know what classes of problems can be solved by which classes of machines. The *Chomsky hierarchy* [94], for example, partitions the formal languages into classes of different expressive power. Each class can be associated a class of abstract automaton, which is able to generate and recognize the corresponding language. At the top of this hierarchy sits the Turing machine, which is able to generate the most expressive class of languages. Along with other abstract formalisms and machines, a Turing machine is able to carry out any effective computation, i. e., it can simulate any other machine capable of performing a well-defined computational procedure. In this context, "effective" is synonymous with "mechanical" or "algorithmic." In its original form, the *Church–Turing thesis* states that there is an equivalence between Church's λ-calculus and Turing machines. Since then, numerous more general and stronger forms of the thesis have emerged. For example, David Deutsch wrote: "Every finitely realizable physical system can be perfectly simulated by a universal model computing machine operating by finite means" (p. 99 in [30]). Despite claims from "hypercomputationalists," (e. g., see [28,104] for an overview of this debate) the fact is that no physically realizable device, which thus necessarily operates on finite resources and time, was ever able to compute functions (in the formal sense of the definition) that a Turing machine could not compute, which means that the Church–Turing thesis has not been disproved as of today.

It is obviously an interesting question to ask how and what biological systems "compute," what their limits are, and how efficient they can solve problems. Biological organisms naturally process some information in some way, but how – if at all – can these processes be described in an algorithmic manner? What kind of abstract automata are useful to describe the many computational mechanisms of a biological cell? Paton describes four general types of computations in [70] under which the information processing capabilities of a cell can be seen:

- *Sequential*. The most common form of machine, which usually has a central processor and a global memory. The instructions are carried out serially, such as for example by a Turing machine.
- *Parallel*. Several processors jointly execute a task. A wide range of parallel computing paradigms exist.
- *Distributed*. Similar to parallel computation, but the processors are usually located at larger distances.
- *Emergent*. The global behavior of such a computation emerges from the local interactions between the processors. This form is typical of self-organizing systems, cellular automata, and neural networks.

As he states, "[c]omputational models of the cell can utilize any of these views," however, most processes are best described as a parallel distributed computing paradigm with emergent properties. Paton also points out the risk of trying to force biological systems to "fit" into a certain computational model. Important details might be ignored and one is forced to think within the existing computing paradigms. We can conclude that abstract and formal models of computation have been very useful in describing the information processing capabilities of biological systems, however, not every formalism is well suited for such descriptions, even if his expressiveness matches that of the corresponding biological system. For example, modeling the highly parallel processes of a biological cell by a Turing machine is possible since parallel machines are not more expressive than sequential machines, however, the sequential paradigm does not offer a convenient and useful model in this context.

The Machine as a Cell

In 1999, Moshe Sipper, used the term "cellular computing" in a broad sense to introduce a general computing philosophy and framework. He hypothesized that "[t]his philosophy promises to provide new means for doing computation more efficiently – in terms of speed, cost, power dissipation, information storage, and solution quality. Simultaneously, cellular computing offers the potential to addressing much larger problem instances than previously possible, at least for some application domains" (p. 18 in [89]). Cellular computing consists of three essential and admittedly very general principles:

- Simplicity,

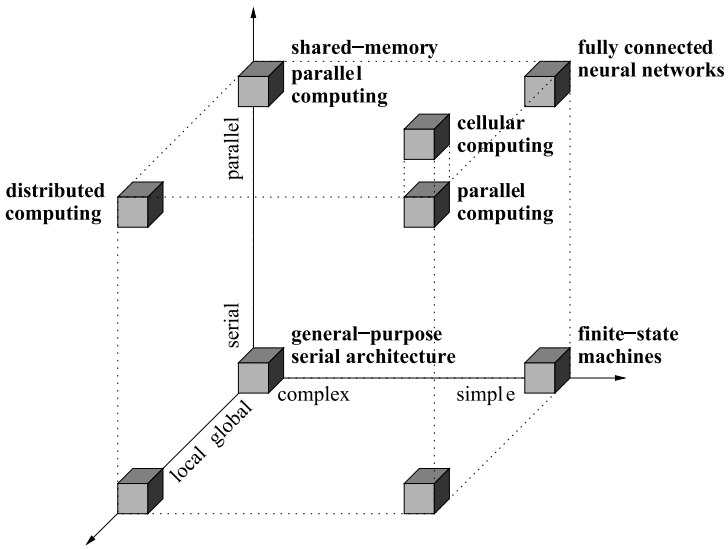

Cellular Computing, Figure 2
The "computing cube." Illustration of the cellular computing paradigm: simple + vastly parallel + local. Redrawn from Sipper [89] and extended. Cellular computing has been placed further along the parallelism axis to emphasize the "vastness" aspect

- Vast parallelism, and
- Locality.

Figure 2 illustrates these three principles. While most general-purpose processors are universal machines in the sense of a Turing machine, the concept of cellular computation expects the complexity of the basic unit, the cell, to be significantly lower, which is characterized by *simplicity*. While "simplicity" is not well or formally defined in this context, Sipper provides the example of an AND gate as being simple, while a Pentium processor is not. Most of today's computers are either based on a multi-core processor, which offers several cores that allow to process information in parallel. This trend – generally considered as an elegant way to keep up with Moore's law – is expected to continue in the next few years. As of November 2006, the Top500 supercomputing list [3] is headed by a BlueGene/L server with 131 072 processors. Cellular computing wants to be orders of magnitude above this level of complexity in terms of the number of cells (i. e., the processors) involved. Sipper uses the term *vast parallelism* to characterize systems where the number of cells is measured by the exponential notation 10^x. For example, an *Avogadro-scale system* would involve an Avogadro number of cells, i. e., 10^{23}. As we scale down electronic circuits to molecular and atomic dimensions and make use of self-assembling techniques, assemblies of this complexity become increasingly feasible. It is unlikely that systems of such complexity will ever be built on macroscopic levels. The last principle concerns the interactions among the cells: cellular computing is based on *local interactions* only, where no cell has a global view or control over the entire system. Obviously, the three principles are related and the cell's simplicity, for example, helps to achieve vast parallelism. Similarly, the vast parallelism makes it hard to implement non-local interactions. These interactions are illustrated in Fig. 2 by the "computing cube." Changing a single term in the "equation" simple + vastly parallel + local results in a different computing paradigm.

Compared to traditional parallel computing, which typically makes use of a rather small number of complex nodes, the cellular computing framework is based on a vast number of simple nodes. Also, parallel computers often have non-local interconnect topologies or even special nodes, which have some sort of global control over all other nodes, which cellular computing does not allow. Clearly, while it is already very challenging to find enough parallelism in applications to exploit the parallelism offered by traditional parallel computers, this problem will only be aggravated – and further moved to the programmer and the software – for fine-grained cellular computers. It is commonly accepted that such machines will likely only be more efficient for very specific problem domains which already offer plenty of inherent parallelism, such as for example image processing. Obviously, one can always implement some form of serial computing paradigm on

a parallel machine (e. g., implementing a Turing machine on the *Game of Life* [81]), but that completely defeats the purpose of such a machine.

This very general framework naturally allows for considerable flexibility in the models. Besides the three basic principles, a number of other important properties play a role, such as for example the detailed local interconnect topology, the cell's arrangement in space, the mobility, the uniformity, and the cell's behavior [89].

- *Arrangement.* While a majority of abstract cellular models do not consider explicit physical locations and spatial dimensions (e. g., the cell dimension of a cellular automata is not generally considered), more biologically and physically plausible models do. Regular grids in 2D (Fig. 3a) or 3D with rectangular or other geometries are most common, but irregular (Fig. 3c) or even non-rigid grids are possible. In addition, one needs to specify the boundary conditions of the grid (e. g., fixed or wrapped around) and whether the grid has a finite or infinite dimension (open boundary).
- *Interconnect topology.* Cellular computing is based on local interactions only, but there are many possibilities to interconnect neighboring cells in a non-global way. In case of a regular rectangular grid, purely local four- (von Neumann, see Fig. 3d) and eight-cell neighborhoods (Moore, see Fig. 3e) are most common [105]. However, interconnect topologies that are completely random or that have the small-world [118] or power-law [6] property are also possible since no cell has a global view over the whole system.
- *Mobility.* Cells can either be physically mobile (e. g., for ant colony optimization [32]) or immobile. Alternatively, their "program" can also move from one cell to another (e. g., for self-replicating loops [58]) without a physical relocation of the cell.
- *Uniformity.* Both the cell type (i. e, its state and behavior) and it's interconnections with the neighboring cells can be uniform (Fig. 3a) or non-uniform (Fig. 3b) to some degree. Most cellular paradigms use uniform cells and interconnection schemes.
- *State and Behavior.* Cells can have internal states that are either discrete or analog. Based on its internal state and the internal states read from the neighboring cells, the cell changes it own state according to some algorithm. This can be as simple as a look-up-table (abbreviated, LUT) or a mathematical function, or as complicated as a small program. The cells are usually updated in discrete time steps, either in a synchronous or asynchronous way. A continuous behavior is also possible, for example if the cell's behavior is guided by differential equations or analog electronics. Finally, both the behavior and the cell's updating can be deterministic or non-deterministic.

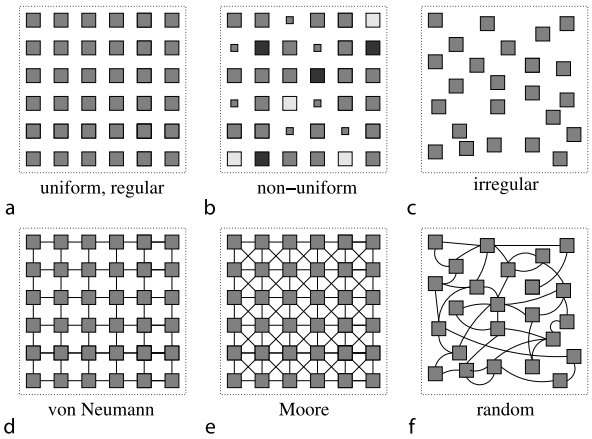

Cellular Computing, Figure 3
Illustration of different cell arrangements and interconnect topologies. a regular and uniform; b regular and non-uniform, different shadings and sizes indicate different cell programs; c irregular and uniform; d regular, uniform, von Neumann neighborhood; e regular, uniform, Moore neighborhood; f irregular, uniform, random neighborhood

Cellular computing does not claim to be a faithful model of biological cells, however, the framework is general and flexible enough to allow for biologically-plausible models as well. As Toffoli and Margulus point out in their book [105] with regards to cellular automata, the generality and flexibility of cellular approaches also comes with a cost. Instead of only a few variables as typical in traditional computing models, the cellular computing paradigm uses at least one variable (or "program") per cell. With the vast number of cells involved, the challenging task is thus to find the right rule(s) which allow to solve a given task. As already stated above, the cellular computing paradigm is ideally suited for tasks that already contain a high degree of parallelism and require local interactions only.

A large number of computing models have drawn inspiration from real cells. While a very large part of the work is concerned with modeling neural cells, a number of researchers were also interested in developmental and other aspects. For example, Astor and Adami [10] used a developmental model for the growth of neural networks. Each neuron is an autonomous entity, which behavior is only determined by its genetic information contained within the cell. An artificial chemistry is also used to model chemicals and substrates. Fleischer and Barr [37] intro-

duced a model of multicellular development that combines elements of the chemical, cell lineage, and mechanical models of morphogenesis pioneered by Turing [109], Lindenmayer [54], and Odell [65]. Cell migration is an important aspect of development, especially of the neural development. Compared to previous models, Fleischer's approach allows cells to move freely within the environment. Other pioneers of morphogenetic, developmental, and growing neural networks models, which necessarily involve cells in various forms, were for example Dellaert and Beer [29], Rust [82], Eggenberger [34], Gruau [42]. More abstract models that deal with the self-maintenance, the self-reproduction, and the self-replication came for example from Ono and Ikegami [66], Varela et al. [113], Langton [51], Mange et al. [58], and various others.

In the following, we will give three more detailed examples of cellular computing models.

Example: Cellular Automata

Cellular automata (abbreviated, CA) were originally conceived by Ulam and von Neumann in the 1940s to provide a formal framework for investigating the behavior of complex, extended systems [115]. CAs are dynamical systems in which space and time are discrete. In its basic version, a cellular automaton [105,120] consists of a regular grid of cells, each of which can be in one of a finite number of k possible states, updated synchronously (or asynchronously) in discrete time steps according to a local, usually identical (uniform, otherwise non-uniform) interaction rule. The state of a given cell is determined by the previous states of a specified neighborhood of cells. Figure 3d and e show two typical local CA neighborhoods, the von Neumann and the Moore neighborhood.

In CAs and many other cellular computing systems, there is a confound between the data and the computing devices, i. e., the structure, which process it. The abstract Turing machine and the von Neumann computer architecture make a clear distinction between these two. The Turing machine stores data on the tape, however, a CA cell can both store and process data and the machine can even extend or modify itself. Although it has been shown that elementary CAs are universal [22], this is more of theoretical than practical interest since most CAs that solve some real problem cannot – and do not need to – perform any universal computation.

One of the main challenges with cellular automata is to find the cell's rule(s) which allow to obtain a global behavior from the local interactions only. For example, given a specific task, say, detecting the contours of an object in an image, what are the individual cell rules (either uniform or non-uniform) that allow to successfully solve that task for a wide range of inputs? The large – or even vast – number of cells involved and the local interactions make this problem very challenging. Thus, most cellular automata are either programmed by hand – which becomes infeasible for larger systems and complex problems – or the rules are for example found by means of an evolutionary algorithm [18,87,88,92].

Example: Membrane Computing

Artificial chemistries [31] are man-made systems that are a very general formulation of abstract systems of *objects* that follow arbitrary *rules of interaction*. More formally speaking, an artificial chemistry essentially consists of a set of *molecules* S, a set of *rules* R, and a definition of the *reactor algorithm* A that describes how the set of rules is applied to the set of molecules. This very broad paradigm, inspired by bio-chemical systems, allows to describe many complex natural and artificial systems by means of simple rules of decentralized, local, and parallel interactions. The chemical paradigm thus also fits in the bigger framework of cellular computing.

In 1998, George Paun initiated *membrane computing* (also *P systems*) [72,73] as an abstract computational model afar inspired by biochemistry and by some of the basic features of biological membranes. Membrane computing makes use of a hierarchical membrane structure that is similar to the structure used in the chemical abstract machine as proposed by Berry and Boudol [15]. The evolution rules that transform the multisets are inspired by *Gamma systems* as proposed by Banâtre et al. [12].

A typical membrane system consists of cell-like membranes placed inside a unique "skin" membrane. Multisets of *objects* – usually strings of symbols – and a set of *evolution rules* (or *reaction rules*) are then placed inside the regions delimited by the membranes. As an example, a simple membrane system is depicted in Fig. 4. The evolution between system configurations is done nondeterministically but synchronously by applying the rules in parallel for all objects able to evolve. A sequence of transitions in a membrane system is called a *computation*. A computation *halts* when a halting configuration is reached, i. e., when no rule can be applied in any region. In classical membrane systems, a computation is considered successful if and only if it halts, but other interpretations of inputs and outputs are possible. A great feature of membrane systems is their inherent parallelism, which allows to trade space for time, for example to solve computationally hard problems. Using membrane division, it has been shown that NP-complete problems can be solved in poly-

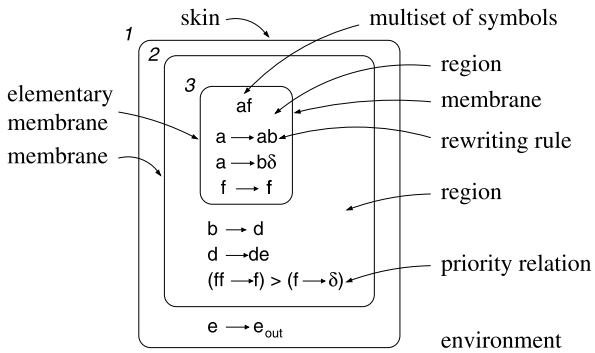

Cellular Computing, Figure 4
An example of a membrane system as given in [73]. The system generates n^2, $n \geq 1$, where n is the number of steps *before* the first application of the rule $a \rightarrow b\delta$. The rules are applied synchronously and in a maximum parallel manner, i. e., all rules that can be applied have to be applied. If multiple rules can be applied, one is picked nondeterministically

nomial and even linear time [73]. Most membrane systems are also Turing universal.

A wide variety of different membrane system flavors exist (see [2] for the latest publications), which have been applied to applications ranging from modeling spiking neural systems to economics.

Example: Amorphous Computing

In a 1996 white paper, Abelson et al. first described the *amorphous computing* (abbreviated, AC) framework [4]. Amorphous computing is the development of organizational principles and programming languages for obtaining coherent global behavior from the local cooperation of a myriad of unreliable parts that are interconnected in unknown, irregular, and time-varying ways. In biology, this question has been recognized as fundamental in the context of animals (such as ants and bees) that cooperate and form organizations. Amorphous computing asks this question in the field of computer science and engineering. Using the biological metaphor, the cells cooperate to form a multicellular organism (also called *programmable multitude*) under the direction of a genetic program shared by all members of the colony. Again, the cellular computing philosophy applies particularly well to the amorphous computing framework. The properties of an amorphous computer can be summarized under the following assumptions, where processor is used synonymous with cell:

- individual processors are identical and mass produced;
- processors possess no *a priori* knowledge of their location, orientation, or neighbors' identities;
- processors operate asynchronously although they have similar clock speeds;
- processors are distributed densely and randomly;
- processors are unreliable;
- processors communicate only locally and do not have a precision interconnect;
- the processors are arranged on a 2D surface or in 3D space;
- it is assumed that the number of particles is very large; and
- the communication model assumes that all processors have a circular broadcast of approximately the same fixed radius (large compared to the fixed size of a processor) and share a single channel.

The biggest challenge in amorphous computing – as with the majority of cellular computing paradigms – is how related to how to program the individual cells in order to obtain a consistent and robust global behavior. The amorphous computing research community came up with several automated approaches, which allow to compile a global system description into local interactions. Example are Nagpal's *origami shape language* [64], Coore's *growing point language* [23], and Butera's *process fragments* [16].

Cellular Computing Hardware

Cells have not only inspired abstract computing models but also real hardware, which is commonly called *biologically-inspired hardware* [93]. On the other hand, both specialized and non-specialized computers have been used for the acceleration and the faithful simulation of many biological processes related to the cell, particularly within the field of *bioinformatics* [11]. A biological cell is a massively parallel system, where thousands of highly complex processes run concurrently. Simulating these processes on a sequential von Neumann computer will obviously not be very efficient and will seriously limit the performance. Specialized hardware, which offers a high level of inherent parallelism, has thus attracted considerable attention in this field of cellular computing.

Cellular automata machines (abbreviated, CAMs) are specialized machines with the goal to very efficiently simulate cellular automata. For example, Hillis' connection machine [44] was the first milestone in that endeavor in the early 1980s. The processors were extremely simple and there was a strong emphasis on the interconnectivity, but not particularly on local topologies because the machines were targeted for supercomputer applications. Toffoli and Margolus' CAM machines [105] represent other examples of highly specialized machines. Nowadays, because

of their inherent fine-grained parallelism, reconfigurable circuits, such as *Field Programmable Gate Arrays* (abbreviated, FPGAs) [41,114], are ideal candidates to simulate cellular systems. For example, Petreska and Teuscher [77] proposed a very first hardware realization of membrane systems using reconfigurable circuits, which was based on a universal and minimal membrane hardware component that allowed to very efficiently evolve membrane systems in hardware. Another example of cellular hardware are *cellular neural network* (abbreviated, CNN) [19] chips, which are massively parallel analog processor arrays. The CNN approach is extremely powerful for specific applications, such as real-time image and video processing.

In the following, we will provide three more detailed examples of typical and biologically-inspired hardware, which is based on cellular paradigms and metaphors.

Example: Embryonics

Biological systems grow, live, adapt, and reproduce, characteristics that are not truly encompassed by any existing computing system. Sipper et al. [93] proposed the *POE model* of bio-inspired hardware systems, which stands for *phylogeny* (abbreviated, P), *ontogeny* (abbreviated, O), and *epigenesis* (abbreviated, E). This model allows to partition the hardware space along three axes, which, in very simplified terms, correspond to evolution (P), development (O), and learning (E). Each axis thus represents a different form of adaption and organization on a different time scale. The ultimate goal are machines that combine all three forms of adaption, so called *POEtic machines* [112].

The *embryonics project* [57,59] (embryonics stands for embryonic electronics), is an attempt to design highly-robust integrated circuits, endowed with the properties usually associated with the living world: self-repair (cicatrization) and self-replication. In this context, self-replication allows for a complete reconstruction of an organism by making a copy of itself. The approach draws inspiration from the basic processes of molecular biology and from the embryonic development of living beings, which is represented by the ontogenetic axis in the POE model. An embryonic circuit is based on a finite but arbitrarily large two-dimensional surface of silicon. This surface is divided into rows and columns, whose intersections define the cells. Since such cells (i. e., a small processor and its memory) have an identical physical structure (i. e., an identical set of logic operators and connections), the cellular array is homogeneous. Embryonics largely draws inspiration from the following biological features, which the majority of living beings (at the exception of unicellular organisms) share: (1) multicellular organization, (2) cellular division, and (3) cellular differentiation. Similar to biological cells, each embryonic cell is "universal" in the sense that it contains the whole of the organism's genetic material, the genome. This feature allows to implement self-repair in an elegant way since each cell is potentially able to replace every other cell within an embryonic organism. The final embryonics architecture makes use of four hierarchical levels of organization:

- *Molecule.* The basic primitive of the system, which essentially consists of a multiplexer associated with a programmable connection network. The multiplexer is duplicated to allow the detection of faults. The logic function of each molecule is defined by its molecular code, the *MOLCODE*.
- *Cell.* A finite set of molecules makes up a cell, which is essentially a processor with an associated memory.
- *Organism* A finite set of cells makes up an organism (synonymous with application), an application-specific multiprocessor system.
- *Population.* The organism can itself self-replicate, given rise to a population of identical organisms.

Figure 5 illustrates these four hierarchical levels. As a showcase of bio-inspired systems and hardware, a complete embryonic machine and application has been implemented in real hardware by using *Field Programmable Gate Arrays* (abbreviated, FPGAs) [41] to implement the basic molecule [102]. The main application, which used several thousand molecules, was a fault-tolerant electronic watch, the BioWatch [96]. The project illustrates how biological organisms, and cells in particular, can help to build alternative computing machines with properties not truly encompassed by any existing architecture.

Example: Field Programmable Gate Arrays

A *Field Programmable Gate Array* (abbreviated, FPGA) [41] is a reconfigurable device [114] which is based on a set of *configurable logic blocks* (abbreviated, CLBs) that are interconnected by a programmable interconnect fabric. The logic blocks of most FPGAs are rather simple, consisting typically of some combinational logic (e. g., a 4-input look-up-table) and some sequential logic (e. g., a flip-flop). Some CLBs offer additional memory too. The interconnect fabric is typically organized in a hierarchical manner with mostly local interconnections between CLBs, but with the possibility to establish limited long(er)-distance connections as well. Sophisticated design tools, compilers, and hardware description languages allow to map essentially any digital circuit onto such a reconfigurable chip, provided enough CLBs are available.

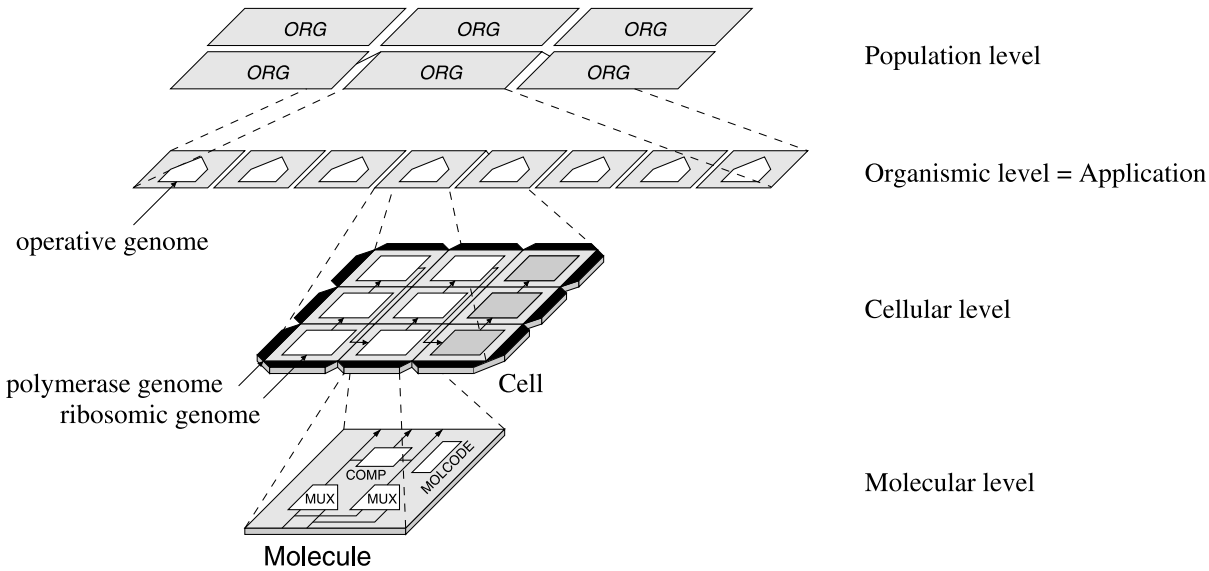

Cellular Computing, Figure 5
The four-level hierarchy of the embryonics landscape. Each level is configured by a part of the genome, which in parallel into all cells once their cell membranes have been constructed by a mechanism similar to cellular division. After this process, the cells differentiate according to their spatial location in the organism [57]

In accordance with the cellular computing framework, FPGAs are based on simple basic building blocks (the CLBs), are locally interconnected (in the sense that no cell controls the entire system), and today's high-end circuits contain hundreds of thousands of CLBs. FPGAs tend to be slower than full-custom VLSI circuits, but offer more flexibility at a lower price. The application of FPGA is obviously most interesting where their massive parallelism can be fully exploited, for example for image processing, neural networks [67], DNA sequence matching, or cellular automata. Sipper et al., for example, used FPGAs for their firefly CA machine [91], which were also used to implement on-line evolution algorithms to automatically find solutions to the synchronization task for CAs [25,88].

Example: The Cell Matrix

The *Cell Matrix* [1,33,56] is an extremely fine-grained and universal hardware architecture, not unlike the structure of an FPGA. The reconfigurable device consists of a homogeneous collection of cells that are interconnected in a nearest-neighbor scheme (e. g., von Neumann or other neighborhoods). Each cell performs very basic operations only that are implemented by a simple look-up table. Unlike traditional reconfigurable devices that are controlled by external systems, the Cell Matrix is self-configurable, i. e., each cell within the matrix can reconfigure other cells. In order to accomplish this, each cell can operate independently in one of two modes: (1) the D mode or (2) the C mode. In the D mode (processing mode), incoming data is processed by the cell's internal look-up table and used to generate the output. In the C mode (configuration mode), on the other hand, incoming data is used to re-write the cell's look-up table. A cell can therefore modify one of its neighbors by placing it in the C mode and by writing the data into the look-up table. Since a cell's look-up table governs the ability to control a neighboring cell's mode, a cell X, for example, can configure a neighboring cell Y in such a way that Y will subsequently configure a third cell Z. By combining cells to so-called *supercells*, hierarchical designs of almost any complexity can be realized. The small groups of cells then interact with nearby groups to achieve higher functionalities, which are used to perform more complex functions, and so on. The Cell Matrix architecture was extended in 2004 by Petraglio in his Ph D thesis [76].

The Cell as a Machine

"The brain computes!" This is accepted as a truism by the majority of neuroscientists engaged in discovering the principles employed in the design and operation of nervous systems. What is meant here is that any brain takes the incoming sensory data, encodes them into various biophysical variables, such as the membrane potential or neural firing rates, and subsequently performs a very

large number of ill-specified operations, frequently termed computations, on these variables to extract relevant features from the input" (p. 1 in [49]). Further on, Koch continues by elaborating that any physical process, which transforms variables, can in principle be thought of as computation, as long as it can be described by some mathematical formulation. As also noted by Conrad [20], living organisms process information in a very different way than digital computers and typically exploit the inherent physical properties of the matter they are made of. The direct, efficient, and sophisticated ways to process information in cells, molecules, and atoms are the results of billions of years of evolution.

In general, computation in living systems is more a question of what level is being analyzed and through which glasses the analyzer looks. A cell has thousands of information processing mechanisms, some are easy to map to existing computing paradigms, others are far away from any existing model and can therefore only be captured by new paradigms. If we want to use a biological cell as a computing device, we essentially have two possibilities:

- *Interpretation.* Interpret existing bio-chemical cellular mechanisms as computation.
- *Modification.* With the advances in bio-technology and bio-engineering in the last decade, bio-chemical processes at all organization levels of a cell can increasingly be modified for the specific purposes of computation.

For example, DNA can be described in computational terms rather easily and the DNA strand interpreted as data, which is being read by ribosomes is a fairly straightforward natural analog to the Turing machine [116]. But this undertaking becomes more interesting if the analogy can be reversed and the DNA is used to realize Boolean *in vivo* or *in vitro* logic gates in a well controlled manner.

As stated in the introduction, the concept of molecular automata and cells seen as information processing devices appeared around the 1960s and the field has seen impressive developments since then. The field is vast and we shall only give a very shallow taste here. The books by Amos [8], Calude and Paun [17], Paton et al. [69], and Sienko et al. [86] provide excellent overviews from the "cell as a machine" perspective on more recent work, while Paton [70], for example, deals with some of the earlier concepts.

Since Adelman's seminal work on *in vitro* DNA computation [5], this unconventional computing paradigm has played a very important role in cellular computation in both experiment and theory [74]. DNA computation clearly outperforms traditional silicon-based electronics in terms of speed, energy efficiency, and information density. However, while it is rather easy to explore huge search spaces, it is often non-trivial to find the correct answer within all the generated solutions, which has somehow limited the practical applications. Other work has focused on RNA editing as a computational process, which offers another interesting cellular paradigm for "biological software" [50]. Later, Winfree, Weiss, and various others, pioneered DNA-based and other *in vivo* logical circuits [13,84,85,119]. Based on deoxyribozymes, Stefanovic's group has developed molecular automata and logic [53,97]. Their circuits include an full adder, three-input logic gates, and an automaton that plays a version of the game tic-tac-toe. While such circuits serve as examples and proof-of-concepts, the group's long-term goal is to use engineered molecules to control autonomous therapeutic and diagnostic devices. These research areas are fast moving and lots of novel results are to be expected in the next few years.

Future Directions

"The simplest living cell is so complex that supercomputer models may never simulate its behavior perfectly. But even imperfect models could shake the foundations of biology" [40]. Gibbs further emphasizes that most attempts to create artificial life or to faithfully model biological systems suffered from a tremendous number of degrees of freedom. The high number of unconstrained system parameters can then lead to almost any desired behavior. Also, the models are often so complicated that they have a very limited ability to predict anything at all. Tomita believes that the study of the cell will never be complete unless its dynamic behavior on all levels is fully understood [106]. He suggests that the complex behavior can only be understood by means of sophisticated computer models that allow to undertake whole cell simulations. Endy and Brent [35] emphasize that "[p]ast efforts to model behavior of molecular and cellular systems over absolute time typically were qualitatively incomplete or oversimplified compared to available knowledge, and qualitatively incomplete in the sense that key numbers were unknown." It is estimated that complete E. coli simulations could run, perhaps, on a single processor system by 2020.

But the future not only lies in more faithful cellular models and in the ability to tweak cells for the purpose of computation. The construction of realistic living cells *in silico*, thought intractable for a long time, has now become within reach, and so does the possibility to build artificial living cells. Several efforts with this specific goal in mind are currently under way in both Europe and the United

States (see for example [78,79,101]). The question is, how and under which conditions simple life forms, which are able to self-maintain and self-replicate, can be synthesized artificially? Steen Rasmussen's team has taken one of the simpler and more artificial routes [79,80], but there is still a long way to go towards artificial cells that will be able to truly self-replicate and to hopefully perform some specific user-defined functions.

Cellular computing is still a very young field and lots of progress is to be expected in the next few decades, particularly in the areas of bio-technology, nano-technology, modeling, and computing. The main challenges in this interdisciplinary field are mostly related to the remaining wide-open gap between biological systems and machines [20]. On one hand, a much better understanding of cellular systems is required, on the other hand, we need to be able to apply this vast and growing body of knowledge to build better and more "lifelike" computing machines by either drawing inspiration from them or by using real cells and organisms for the purpose of computation.

The following research directions and challenges need to be addressed in the next few years:

1. Faithful qualitative and quantitative cellular models require a better understanding of the biological cell and its bio-chemical mechanism. For example, better non-invasive brain imaging techniques are required for a better understanding of neurons, populations of neurons, and brain regions.
2. Large-scale molecular simulations of cellular processes and large populations of cells. For example, by using specialized supercomputers, large-scale brain simulations only become possible now.
3. Novel computing paradigms are required to model and to understand biological systems. Boolean logic is appropriate for digital systems, where 0s and 1s can easily be represented by the presence and absence of electrical current, but alternative representations may be more appropriate for biological systems.
4. Novel programming paradigms for cellular systems made up from a vast number of cells. Traditional top-down design approaches do not typically scale up to such complexities. Nature-inspired bottom-up design approaches are required.

Bibliography

Primary Literature

1. Cell matrix corporation. http://www.cellmatrix.com. Accessed 12 Jul 2008
2. The P systems web page. http://ppage.psystems.eu. Accessed 12 Jul 2008
3. Top500 supercomputing sites. http://www.top500.org. Accessed 12 Jul 2008
4. Abelson H, Allen D, Coore D, Hanson C, Rauch E, Sussman GJ, Weiss R (2000) Amorphous computing. Commun ACM 43(5):74–82
5. Adelman LM (1994) Molecular computation of solutions to combinatorial problems. Science 266:1021–1024
6. Albert R, Barabasi AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
7. Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD (eds) (1994) Molecular Biology of the Cell, 3rd edn. Garland Publishing, New York
8. Amos M (ed) (2004) Cellular Computing. Oxford University Press, New York
9. Arbib MA (ed) (1995) The Handbook of Brain Theory and Neural Networks. MIT Press, Cambridge
10. Astor JC, Adami C (2001) A developmental model for the evolution of artificial neural networks. Artif life 6(3):189–218
11. Baldi P, Brunak S (2001) Bioinformatics: The Machine Learning Approach, 2nd edn. MIT Press, Cambridge
12. Banâtre JP, Coutant A, Le Metayer D (1988) A parallel machine for multiset transformation and its programming style. Future Gener Comput Syst 4:133–144
13. Benenson Y, Paz-Elizur T, Adar R, Keinan E, Leivneh Z, Shapiro E (2001) Programmable and autonomous computing machine made of biomolecules. Nature 414(6862):430–434
14. Bennett CH (1982) The thermodynamics of computation a review. Int J Theor Phys 21(12):905–940
15. Berry G, Boudol G (1992) The chemical abstract machine. Theor Comput Sci 96:217–248
16. Butera WJ (2002) Programming a Paintable Computer. Ph D thesis, MIT Media Lab, Cambridge
17. Calude CS, Paun G (2000) Computing with Cells and Atoms: An Introduction to Quantum and Membrane Computing. Taylor & Francis, New York
18. Capcarrere MS (2002) Cellular Automata and Other Cellular Systems: Design & Evolution. Ph D thesis, Swiss Federal Institute of Technology Lausanne
19. Chua LO, Roska T (2002) Cellular Neural Networks & Visual Computing. Cambridge University Press, Cambridge
20. Conrad M (1989) The brain-machine disanalogy. BioSystems 22(3):197–213
21. Conrad M, Liberman EA (1982) Molecular computing as a link between biological and physical theory. J Theor Biol 98(2):239–252
22. Cook M (2004) Universality in elementary cellular automata. Complex Syst 15(1):1–40
23. Coore D (1999) Botanical Computing: A Developmental Approach to Generating Interconnect Topologies on an Amorphous Computer. Ph D thesis, MIT Department of Electrical Engineering and Computer Science, Cambridge
24. Cuniberti G, Fagas G, Richter K (eds) (2005) Introducing Molecular Electronics. In: Lecture Notes in Physics, vol 680. Springer, Berlin
25. Das R, Crutchfield JP, Mitchell M, Hanson JE (1995) Evolving globally synchronized cellular automata. In: Eshelman LJ (ed) Proceedings of the Sixth International Conference on Genetic Algorithms. Morgan Kaufmann, San Francisco, pp 336–343
26. Davis BD (1961) The teleonomic significance of biosynthetic control mechanisms. Cold Spring Harbor Symp Quant Biol 26:1–10

27. Davis M (1958) Computability and Unsolvability. McGraw-Hill, New York
28. Davis M (2004) The myth of hypercomputation. In: Teuscher C (ed) Alan Turing: Life and Legacy of a Great Thinker. Springer, Berlin, pp 195–212 (reprinted with color images in 2005)
29. Dellaert F, Beer RD (1994) Toward an evolvable model of development for autonomous agent synthesis. In: Brooks RA, Maes P (eds) Artificial Life IV. Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems. A Bradford Book, MIT Press, Cambridge, pp 246–257
30. Deutsch D (1985) Quantum theory, the Church–Turing principle of the Universal Quantum Computer. Proc Royal Soc London A400:97–117
31. Dittrich P, Ziegler J, Banzhaf W (2001) Artificial chemistries a review. Artif Life 7(3):225–275
32. Dorigo M, Stützle T (2004) Ant Colony Optimization. MIT Press, Cambridge
33. Durbeck LJK, Macias NJ (2001) The Cell Matrix: An architecture for nanocomputing. Nanotechnology 12:217–230
34. Eggenberger P (1997) Creation of neural networks based on developmental and evolutionary principles. In: Gerstner W, Germond A, Hasler M, Nicod JD (eds) Proceedings of the International Conference on Artificial Neural Networks. ICANN'97, Lausanne, Switzerland. Lecture Notes in Computer Science, vol 1327. Springer, Berlin, pp 337–342
35. Endy D, Brent R (2001) Modelling cellular behavior. Nature 409(6818):391–395
36. Feynman RP (1960) There's plenty of room at the bottom: An invitation to enter a new field of physics. Caltech's Eng Sci (Feb 1960):22–36
37. Fleischer KW, Barr AH (1994) A simulation testbed for the study of multicellular development: The multiple mechanisms of morphogenesis. In: Langton CG (ed) Artificial Life III. SFI Studies in the Science of Complexity, vol XVII. Addison–Wesley, Redwood City, pp 389–416
38. Flemming W (1880) Beiträge zur Kenntnis der Zelle und ihrer Lebenserscheinungen. Arch Mikrosk Anat 18:151–289
39. Forbes N (2004) Imitation of Life: How Biology is Inspiring Computing. MIT Press, Cambridge
40. Gibbs WW (2001) Cybernetic cells. Sci Am 285(2):43–47
41. Gokhale M, Graham PS (2005) Reconfigurable Computing: Accelerating Computation with Field Programmable Gate Arrays. Springer, Berlin
42. Gruau F (1994) Neural Network Synthesis Using Cellular Encoding and the Genetic Algorithm. Ph D thesis, Ecole Normale Supérieure de Lyon
43. Harel D (2003) Computers Ltd.: What They Really Can't Do. Oxford University Press, New York
44. Hillis DW (1985) The Connection Machine. MIT Press, Cambridge
45. Hooke R (1665) Micrographia: Or, some physiological descriptions of minute bodies made by magnifying glasses. J. Martyn and J. Allestry, London
46. Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3:318–356
47. Kish LB (2002) End of Moore's law: Thermal (noise) death of integration in micro and nano electronics. Phys Lett A 305: 144–149
48. Kitano H (ed) (2001) Foundations of Systems Biology. MIT Press, Cambridge
49. Koch C (1999) Biophysics of Computation. Oxford University Press, New York
50. Landweber LF, Kari L (1999) The evolution of cellular computing: Nature's solution of a computational problem. BioSystems 52:3–13
51. Langton CG (1984) Self-reproduction in cellular automata. Physica D 10(1–2):135–144
52. Langton CG (ed) (1995) Artificial Life: An Overview. MIT Press, Cambridge
53. Lederman H, Macdonald J, Stefanovic D, Stojanovic MN (2006) Deoxyribozyme-based three-input logic gates and construction of a molecular full adder. Biochemistry 45(4): 1194–1199
54. Lindenmayer A (1968) Mathematical models for cellular interaction in development, parts I and II. J Theor Biol 18:280–315
55. Lloyd S (2000) Ultimate physical limits to computation. Nature 406:1047–1054
56. Macias NJ (1999) The PIG paradigm: the design and use of a massively parallel fine grained self-reconfigurable infinitely scalable architecture. In: Stoica A, Keymeulen D, Lohn J (eds) Proceedings of the First NASA/DOD Workshop on Evolvable Hardware. IEEE Computer Society, Los Alamitos, pp 175–180
57. Mange D, Sipper M, Stauffer A, Tempesti G 2000 Toward robust integrated circuits: The embryonics approach. Proc IEEE 88(4):516–540
58. Mange D, Stauffer A, Peparolo L, Tempesti G (2004) A macroscopic view of self-replication. Proc IEEE 92(12):1929–1945
59. Mange D, Tomassini M (eds) (1998) Bio-Inspired Computing Machines: Towards Novel Computational Architectures. Presses Polytechniques et Universitaires Romandes, Lausanne
60. McCulloch WS, Pitts WH (1943) A logical calculus of the ideas immanent in neural nets. Bull Math Biophys 5:115–133
61. Minsky ML 1967 Computation: Finite and Infinite Machines. Prentice-Hall, Englewood Cliffs
62. Monod J, Jacob F (1961) Teleonomic mechanisms in cellular metabolism, growth, and differentiation. Cold Spring Harbor Symp Quantit Biol 26:389–401
63. Moore GE (1965) Cramming more components onto integrated circuits. Electronics 38(8):114–117
64. Nagpal R (2001) Programmable Self-Assembly: Constructing Global Shape using Biologically-inspired Local Interactions and Origami Mathematics. Ph D thesis, MIT Department of Electrical Engineering and Computer Science, Cambridge
65. Odell GM, Oster G, Albrech P, Burnside B (1981) The mechanical basis of morphogenesis. In: Epithelial folding and invagination. Dev Biol 85:446–462
66. Ono N, Ikegami T (2000) Self-maintenance and self-reproduction in an abstract cell model. J Theor Biol 206(2):243–253
67. Ormondi AR, Rajapakse JC (eds) (2006) FPGA Implementations of Neural Networks. Springer, Dordrecht
68. Papadimitrou CH (1994) Computational Complexity. Addison–Wesley, Reading
69. Paton R, Bolouri H, Holcombe M, Parish JH, Tateson R (eds) (2004) Computation in Cells and Tissues. Springer, Berlin
70. Paton RC (1993) Some computational models at the cellular level. Biosystems 29:63–75
71. Pattee HH (1961) On the origin of macromolecular sequences. Biophys J 1(8):683–710
72. Paun G (2000) Computing with membranes. J Comput Syst

Sci 61(1):108–143 (first published in a TUCS Research Report, No 208, Nov 1998, http://www.tucs.fi)
73. Paun G (2002) Membrane Computing. Springer, Berlin
74. Paun G, Rozenberg G, Salomaa A (1998) DNA Computing: New Computing Paradigms. Springer, Heidelberg
75. Paweletz N (2001) Walter flemming: Pioneer of mitosis research. Nat Rev Mol Cell Biol 2:72–75
76. Petraglio E (2004) Fault tolerant self-replicating systems. Ph D thesis, Swiss Federal Institute of Technology (EPFL), Lausanne
77. Petreska B, Teuscher C (2004) A reconfigurable hardware membrane system. In: Martin-Vide C, Mauri G, Paun G, Rozenberg G, Salomaa A (eds) Membrane Computing. Lecture Notes in Computer Science, vol 2933. Springer, Berlin, pp 269–285
78. Pohorille A, Deamer D (2002) Artificial cells: Prospects for biotechnology. Trends Biotechnol 20(3):123–128
79. Rasmussen S, Chen L, Deamer D, Krakauer DC, Packard NH, Stadler PF, Bedau MA (2004) Transitions from nonliving to living matter. Science 303:963–965
80. Rasmussen S, Chen L, Nilsson M, Abe S (2003) Bridging nonliving and living matter. Artif Life 9(3):269–316
81. Rendell P (2002) Turing universality of the Game of Life. In: Adamatzky A (ed) Collision-Based Computing. Springer, London, pp 513–539
82. Rust AG (1998) Developmental Self-Organisation in Artificial Neural Networks. Ph D thesis, University of Hertfordshire
83. Schmitt FO (1962) Macromolecular Specificity and Biological Memory. MIT Press, Cambridge
84. Seelig G, Soloveichik D, Zhang DY, Winfree E (2006) Enzyme-free nucleic acid logic circuits. Science 314:1585–1588
85. Shapiro E, Gil B (2007) Logic goes in vitro. Nat Nanotechnol 2:84–85
86. Sienko T, Adamatzky A, Rambidi NG, Conrad M (eds) (2003) Molecular Computing. MIT Press, Cambridge
87. Sipper M (1996) Co-evolving non-uniform cellular automata to perform computations. Physica D 92:193–208
88. Sipper M (1997) Evolution of Parallel Cellular Machines: The Cellular Programming Approach. Springer, Heidelberg
89. Sipper M (1999) The emergence of cellular computing. IEEE Comput 32(7):18–26
90. Sipper M (2002) Machine Nature: The Coming Age of Bio-Inspired Computing. McGraw-Hill, New York
91. Sipper M, Goeke M, Mange D, Stauffer A, Sanchez E, Tomassini M (1997) The firefly machine: Online evolware. In: Proceedings of the 1997 IEEE International Conference on Evolutionary Computation (ICEC'97), Piscataway. IEEE Press, Piscataway, pp 181–186
92. Sipper M, Ruppin E (1997) Co-evolving architectures for cellular machines. Physica D 99:428–441
93. Sipper M, Sanchez E, Mange D, Tomassini M, Pérez-Uribe A, Stauffer A (1997) A phylogenetic, ontogenetic, and epigenetic view of bio-inspired hardware systems. IEEE Trans Evol Comput 1(1):83–97
94. Sipser M (2006) Introduction to the Theory of Computation, 2nd edn. Thomson, Boston
95. Stahl WR, Goheen HE (1963) Molecular algorithms. J Theor Biol 5(2):266–287
96. Stauffer A, Mange D, Tempesti G, Teuscher C (2001) Bio Watch: A giant electronic bio-inspired watch. In: Keymeulen D, Stoica A, Lohn J, Zebulum RS (eds) Proceedings of the Third NASA/DoD Workshop on Evolvable Hardware, EH-2001. IEEE Computer Society, Los Alamitos, pp 185–192
97. Stojanovic MN, Stefanovic D (2003) A deoxyribozyme-based molecular automaton. Nat Biotechnol 21:1069–1074
98. Sugita M (1961) Functional analysis of chemical systems in vivo using a logical circuit equivalent. J Theor Biol 1(2):415–430
99. Sugita M (1963) Functional analysis of chemical systems in vivo using a logical circuit equivalent. In: The idea of a molecular automaton. J Theor Biol 4(2):179–192
100. Sugita M, Fukuda N (1963) Functional analysis of chemical systems in vivo using a logical circuit equivalent. III: Analysis using a digital circuit combined with an analogue computer. J Theor Biol 5(3):412–425
101. Szostak JW, Bartel DP, Luisi PL (2001) Synthesizing life. Nature 409:387–390
102. Tempesti G, Mange D, Stauffer A, Teuscher C (2002) The Biowall: An electronic tissue for prototyping bio-inspired systems. In: Stoica A, Lohn J, Katz R, Keymeulen D, Zebulum RS (eds) Proceedings of the 2002 NASA/DoD Conference on Evolvable Hardware. IEEE Computer Society, Los Alamitos, pp 221–230
103. Teuscher C (2002) Turing's Connectionism. An Investigation of Neural Network Architectures. Springer, London
104. Teuscher C, Sipper M (2002) Hypercomputation: Hype or computation? Commun ACM 45(8):23–24
105. Toffoli T, Margolus N (1987) Cellular Automata Machines. MIT Press, Cambridge
106. Tomita M (2001) Whole-cell simulation: A grand challenge of the 21st century. Trends Biotechnol 19(6):205–210
107. Turing AM (1937) On computable numbers, with an application to the Entscheidungsproblem. Proc London Math Soc 42:230–265; Corrections. Proc London Math Soc 43:544–546
108. Turing AM (1950) Computing machinery and intelligence. Mind 59(236):433–460
109. Turing AM (1952) The chemical basis of morphogenesis. Philos Trans Royal Soc London B 237:37–72
110. Turing AM (1969) Intelligent machinery. In: Meltzer B, Michie D (eds) Machine Intelligence, vol 5. Edinburgh University Press, Edinburgh, pp 3–23
111. Turing AM (1992) Intelligent machinery. In: Ince DC (ed) Collected Works of A.M. Turing: Mechanical Intelligence. North-Holland, Amsterdam, pp 107–127
112. Tyrrell A, Sanchez E, Floreano D, Tempesti G, Mange D, Moreno JM, Rosenberg J, Alessandro Villa EP (2003) Poetic tissue: An integrated architecture for bio-inspired hardware. In: Tyrrell AM, Haddow PC, Torresen J (eds) Evolvable Systems: From Biology to Hardware. Proceedings of the 5th International Conference (ICES2003). Lecture Notes in Computer Science, vol 2606. Springer, Berlin, pp 129–140
113. Varela F, Maturana H, Uribe R (1974) Autopoiesis: The organization of living systems, its characterization and a model. BioSystems 5:187–196
114. Villasenor J, Mangione-Smith WH (1997) Configurable computing. Sci Am 276(6):54–59
115. von Neumann J (1966) Theory of Self-Reproducing Automata. University of Illinois Press, Urbana
116. Warner JR, Rich A, Hall CE (1962) Electron microscope studies of ribosomal clusters synthesizing hemoglobin. Science 138(3548):1299–1403

117. Watson JD, Crick FHC (1953) A structure for deoxyribose nucleic acid. Nature 171:737–738
118. Watts DJ, Strogatz SH (1998) Collective dynamics of small-world networks. Nature 393:440–442
119. Weiss R (2001) Cellular Computation and Communications using Engineered Genetic Regulatory Networks. Ph D thesis, MIT Department of Electrical Engineering and Computer Science, Cambridge
120. Wolfram S (1984) Cellular automata as models of complexity. Nature 311:419–424

Books and Reviews

Adamatzky A (ed) (2002) Collision-Based Computing. Springer, London
Adami C (1998) Introduction to Artificial Life. Springer, New York
Amos M (2005) Theoretical and Experimental DNA Computation. Springer, Berlin
Bentley BJ (2001) Digital Biology: Now Nature is Transforming our Technology. Headline Book Publishing, London
Ehrenfeucht A, Harju T, Petre I, Prescott DM, Rozenberg G (2004) Computation in Living Cells: Gene Assembly in Ciliates. Springer, Berlin
Hey AJG (ed) (1998) Feynman and Computation: Exploring the Limits of Computers. Westview, Boulder
Landweber LF, Winfree E (eds) (2002) Evolution as Computation. Springer, Berlin
Paun G, Rozenberg G (2002) A guide to membrane computing. J Theor Comput Sci 287(1):73–100
Petty MC, Bryce MR, Bloor D (eds) (1995) An Introduction to Molecular Electronics. Oxford University Press, New York
Pozrikidis C (ed) (2003) Modeling and Simulation of Capsules and Biological Cells. Chapman and Hall/CRC, Boca Raton
Sipper M (2002) Machine Nature: The Coming Age of Bio-Inspired Computing. McGraw-Hill, New York
Trimberger SM (1994) Field-Programmable Gate Array Technology. Kluwer, Boston
Wolfram S (2002) A New Kind of Science. Wolfram Media, Champaign
Zauner KP (2005) Molecular information technology. Crit Rev Solid State Mater Sci 30(1):33–69

Center Manifolds

George Osipenko
State Polytechnic University, St. Petersburg, Russia

Article Outline

Glossary
Definition of the Subject
Introduction
Center Manifold in Ordinary Differential Equations
Center Manifold in Discrete Dynamical Systems
Normally Hyperbolic Invariant Manifolds
Applications
Center Manifold in Infinite-Dimensional Space
Future Directions
Bibliography

Glossary

Bifurcation Bifurcation is a qualitative change of the phase portrait. The term "bifurcation" was introduced by H. Poincaré.

Continuous and discrete dynamical systems A dynamical system is a mapping $X(t,x)$, $t \in R$ or $t \in Z$, $x \in E$ which satisfies the group property $X(t+s,x) = X(t, X(s,x))$. The dynamical system is continuous or discrete when t takes real or integer values, respectively. The continuous system is generated by an autonomous system of ordinary differential equations

$$\dot{x} \equiv \frac{\mathrm{d}x}{\mathrm{d}t} = F(x) \tag{1}$$

as the solution $X(t,x)$ with the initial condition $X(0,x) = x$. The discrete system generated by a system of difference equations

$$x_{m+1} = G(x_m) \tag{2}$$

as $X(n,x) = G^n(x)$. The phase space E is the Euclidean or Banach.

Critical part of the spectrum Critical part of the spectrum for a differential equation $\dot{x} = Ax$ is $\sigma_c = \{$eigenvalues of A with zero real part$\}$. Critical part of the spectrum for a diffeomorphism $x \to Ax$ is $\sigma_c = \{$eigenvalues of A with modulus equal to 1$\}$.

Eigenvalue and spectrum If for a matrix (linear mapping) A the equality $Av = \lambda v$, $v \neq 0$ holds then v and λ are called eigenvector and eigenvalue of A. The set of the eigenvalues is the spectrum of A. If there exists k such that $(A - \lambda I)^k v = 0$, v is said to be generating vector.

Equivalence of dynamical systems Two dynamical systems f and g are topologically equivalent if there is a continuous one-to-one correspondence (homeomorphism) that maps trajectories (orbits) of f on trajectories of g. It should be emphasized that the homeomorphism need not be differentiable.

Invariant manifold In applications an invariant manifold arises as a surface such that the trajectories starting on the surface remain on it under the system evolution.

Local properties If $F(p) = 0$, the point p is an equilibrium of (1). If $G(q) = q$, q is a fixed point of (2). We study dynamics near an equilibrium or a fixed point of the system. Thus we consider the system in a neighborhood of the origin which is supposed to be an equi-

librium or a fixed point. In this connection we use terminology local invariant manifold or local topological equivalence.

Reduction principle In accordance with this principle a locally invariant (center) manifold corresponds to the critical part of the spectrum of the linearized system. The behavior of orbits on a center manifold determines the dynamics of the system in a neighborhood of its equilibrium or fixed point. The term "reduction principle" was introduced by V. Pliss [59].

Definition of the Subject

Let M be a subset of the Euclidean space R^n.

Definition 1 A set M is said to be smooth manifold of dimension $d \leq n$ if for each point $p \in M$ there exists a neighborhood $U \subset M$ and a smooth mapping $g\colon U \to U_0 \subset R^d$ such that there is the inverse mapping g^{-1} and its differential Dg^{-1} (matrix of partial derivatives) is an injection i. e. its maximal rang is d.

For sphere in R^3 which is a two-dimensional smooth manifold the introduction of the described neighborhoods is demonstrated in [77]. A manifold M is C^k-smooth, $k \geq 1$ if the mappings g and g^{-1} have k continuous derivatives. Moreover, chosing the mappings g and g^{-1} as C^∞-smooth or analytical, we obtain the manifold with the same smoothness.

Consider a continuous dynamical system

$$\dot{x} = F(x)\,, \quad x \in R^n\,, \tag{3}$$

where $F\colon R^n \to R^n$ is a C^k-smooth vector field, $k \geq 1$, i. e. the mapping F has k continuous derivatives. Let us denote the solution of (3) passing through the point $p \in R^n$ at $t = 0$ by $X(t,p)$. Suppose that the solution is determined for all $t \in R$. By the fundamental theorems of the differential equations theory, the hypotheses imposed on F guarantee the existence, uniqueness, and smoothness of $X(t,p)$ for all $p \in R^n$. In this case the mapping $X_t(p) = X(t,p)$ under fixed t is a diffeomorphism of R^n. Thus, F generates the smooth flow $X_t\colon R^n \to R^n$, $X_t(p) = X(t,p)$. The differential $DX_t(0)$ is a fundamental matrix of the linearized system $\dot{v} = DF(0)v$. A trajectory (an orbit) of the system (3) through x_0 is the set $T(x_0) = \{x = X(t,x_0),\ t \in R\}$.

Definition 2 A manifold M is said to be invariant under (3) if for any $p \in M$ the trajectory through p lies in M. A manifold M is said to be locally invariant under (3) if for every $p \in M$ there exists, depending on p, an interval $T_1 < 0 < T_2$, such that $X_t(p) \in M$ as $t \in (T_1,\ T_2)$. It means that the invariant manifold is formed by trajectories of the system and the locally invariant manifold consists of arcs of trajectories. An equilibrium is 0-dimensional invariant manifold and a periodic orbit is 1-dimensional one. The concept of invariant manifold is a useful tool for simplification of dynamical systems.

Definition 3 A manifold M is said to be invariant in a neighborhood U if for any $p \in M \bigcap U$ the moving point $X_t(p)$ remains on M as long as $X_t(p) \in U$. In this case the manifold M is locally invariant.

Near an equilibrium point O the system (3) can be rewritten in the form

$$\dot{x} = Ax + f(x)\,, \tag{4}$$

where $O = \{x = 0\}$, $A = DF(O)$, f is second-order at the origin, i. e. $f(0) = 0$ and $Df(0) = 0$. Let us show that the system (4) near O can be considered as a perturbation of the linearized system

$$\dot{x} = Ax\,. \tag{5}$$

For this we construct a C^k-smooth mapping g which coincides with f in a sufficiently small neighborhood of the origin and is C^1-close to zero.

To construct the mapping g one uses the C^∞-smooth cut-off function $\alpha\colon R^+ \to R^+$

$$\alpha(r) = \begin{cases} 1, & r < 1/2\,, \\ > 0, & 1/2 \leq r \leq 1\,, \\ 0, & r > 1\,. \end{cases}$$

Then set $g(x) = \alpha(|x|/\epsilon)f(x)$ where ε is a parameter. If $|x| > \epsilon$, $g(x) = 0$ and if $|x| < \epsilon/2$, $g(x) = f(x)$ and g is C^1-close to zero. Consider the system

$$\dot{x} = Ax + g(x)\,. \tag{6}$$

It is evident that the dynamics of (4) and (6) is the same on $U = \{|x| < \epsilon/2\}$. If the constructed system (6) has an invariant manifold M then $M \bigcap U$ is locally invariant manifold for the initial system (4) or, more precisely, the manifold M is invariant in U. It should be noted that for the case of infinite-dimensional phase space the described construction is more delicate (see details below).

Let us consider a C^k-smooth mapping ($k \geq 1$) $G\colon R^n \to R^n$ which has the inverse C^k-smooth mapping G^{-1}. The mapping G generates the discrete dynamical system of the form

$$x_{m+1} = G(x_m)\,, \quad m = \ldots, -2, -1, 0, 1, 2, \ldots\,. \tag{7}$$

Each continuous system $X(t,p)$ gives rise to the discrete system $G(x) = X(1,x)$ which is the shift operator on unit

time. Such a system preserves the orientation of the phase space,where as an arbitrary discrete system may change it.

Hence, the space of discrete systems is more rich than the space of continuous ones. Moreover, usually investigation of a discrete system is, as a rule, simpler than study of a differential equations one. In many instances the investigation of a differential equation may be reduced to study of a discrete system by Poincaré (first return) mapping. An orbit of (7) is the set $T = \{x = x_m,\ m \in \mathbb{Z}\}$, where x_m satisfies (7).

Near a fixed point O the mapping $x \to G(x)$ can be rewritten in the form

$$x \to Ax + f(x)\,, \tag{8}$$

where $O = \{x = 0\}$, $A = DG(O)$, f is second-order at the origin. It is shown above that near O (8) may be considered as a perturbation of the linearized system

$$x \to Ax\,. \tag{9}$$

Motivational Example

Consider the discrete dynamical system

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{pmatrix} X(x,y) = x + xy \\ Y(x,y) = \frac{1}{2}y + \frac{1}{2}x^2 + 2x^2y + y^3 \end{pmatrix}. \tag{10}$$

The origin $(0,0)$ is fixed point. Our task is to examine the stability of the fixed point. The linearized system at 0 is defined by the matrix

$$\begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix}$$

which has two eigenvalues 1 and 1/2. Hence, the first approximation system contracts along y-axis, whereas its action along x-axis is neutral. So the stability of the fixed point depends on nonlinear terms. Let us show that the curve $y = x^2$ is invariant for the mapping (10). It is enough to check that if a point (x,y) is on the curve then the image (X,Y) is on the curve, i. e. $Y = X^2$ as $y = x^2$. We have

$$\begin{aligned} Y|_{y=x^2} &= \tfrac{1}{2}y + \tfrac{1}{2}x^2 + 2x^2y + y^3|_{y=x^2} \\ &= x^2 + 2x^4 + x^6\,, \\ X^2|_{y=x^2} &= (x + xy)^2|_{y=x^2} \\ &= x^2 + 2x^4 + x^6\,. \end{aligned} \tag{11}$$

Thus, the curve $y = x^2$ is an invariant one-dimensional manifold which is a center manifold W^c for the system (10), see Fig. 1. The fixed point O is on W^c. The restriction of the system on the manifold is

$$x \to x + xy|_{y=x^2} = x(1 + x^2)\,. \tag{12}$$

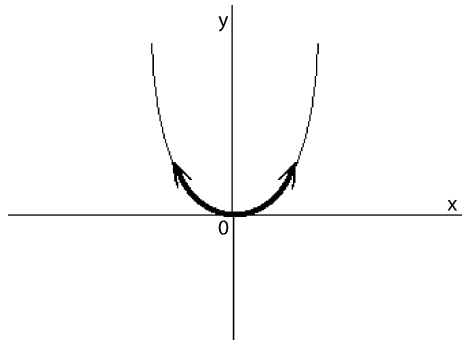

Center Manifolds, Figure 1
Dynamics on the invariant curve $y = x^2$

It follows that the fixed point 0 is unstable and $x_m \to \infty$ as $m \to \infty$, see Fig. 1. Hence, the origin is unstable fixed point for the discrete system (10). It turns out the system (10) near O is topologically equivalent to the system

$$\begin{aligned} x &\to x(1 + x^2)\,, \\ y &\to \tfrac{1}{2}y \end{aligned} \tag{13}$$

which is simpler than (10). The center manifold $y = x^2$ is tangent to the x-axis at the origin and W^c near 0 can be considered as a perturbed manifold of $\{y = 0\}$.

Introduction

In his famous dissertation "General Problem on Stability of Motion" [41] published at 1892 in Khar'kov (Ukraine) A.M. Lyapunov proved that the equilibrium O of the system (4) is stable if all eigenvalues of matrix A have negative real parts and O is unstable if there exists an eigenvalue with positive real part. He studied the case when some eigenvalues of A have negative real part and the rest of them have zero real one. Lyapunov proved that if the matrix A has a pair of pure imaginary eigenvalues and the other eigenvalues have negative real parts then there exists a two-dimensional invariant surface M through O, and the equilibrium O is stable if O is stable for the system restricted on M. Speaking in modern terms, Lyapunov proved the existence of "center manifold" and formulated "reduction principle" under the described conditions. Moreover, he found the stability condition by using power series expansions, which is an extension of his first method to evaluate stability of systems whose eigenvalues have zero real part. Now we use the same method to check stability [10,25,31,36,43,76]. The general reduction principle in stability theory was established by V. Pliss [59] in 1964, the term "center manifold" was introduced by A. Kelley at 1967 in the paper [38] where the existence of

a family of invariant manifolds through equilibrium was proved in general case.

As we see the center manifold can be considered as a perturbation of the center subspace of linearized system. H. Poincaré [61] was probably the first who perceived the importance of the perturbation problem and began to study conditions ensuring the preservation of the equilibriums and periodic orbits under a perturbation of differential equations. Hadamard [26] and Perron [57] proved the existence of stable and unstable invariant manifolds for a hyperbolic equilibrium point and, in fact, showed their preservation. The conditions necessary and sufficient for the preservation of locally invariant manifolds passing through an equilibrium have been obtained in [51]. Many results on the center manifold follow from theory of normal hyperbolicity [77] which studies dynamics of a system near a compact invariant manifold. They will be considered below. The books by Carr [10], Guckenheimer and Holmes [25], Marsden and McCracken [43], Iooss [36], Hassard, Kazarinoff and Wan [31], and Wiggins [76] are popular sources of the information about center manifolds.

Center Manifold in Ordinary Differential Equations

Consider a linear system of differential equations

$$\dot{v} = Lv\,, \tag{14}$$

where $v \in R^n$ and L is a matrix. Divide the eigenvalues of the matrix L into three parts: stable $\sigma_s = \{$eigenvalues with negative real part$\}$, unstable $\sigma_u = \{$eigenvalues with positive real part$\}$ and central (neutral, critical) $\sigma_c = \{$eigenvalues with zero real part$\}$. If the matrix does not have the eigenvalues with zero real part, the system is called hyperbolic. The matrix L has three eigenspaces: the stable subspace E^s, the unstable subspace E^u, and the central subspace E^c correspond to these parts of spectrum. The subspaces E^s, E^u, and E^c are invariant under the system (14), that is, the solution starting on the subspace stays on it [33]. The solutions on E^s have exponential tending to 0, the solutions on E^u have exponential growth. The subspaces E^s, E^c, E^u meet in pairs only at the origin and $E^s + E^c + E^u = R^n$, i. e. we have the invariant decomposition $R^n = E^s \oplus E^c \oplus E^u$. There exists a linear change of coordinates (see [33]) transforming (14) to the form

$$\begin{aligned} \dot{x} &= Ax\,, \\ \dot{y} &= By\,, \\ \dot{z} &= Cz\,, \end{aligned} \tag{15}$$

where the matrix A has eigenvalues with zero real part, the matrix B has eigenvalues with negative real part, and the matrix C has eigenvalues with positive real part. So, (14) decomposes into three independent systems of differential equations. It is known [1,23,24] that the system (15) is topologically equivalent to the system

$$\begin{aligned} \dot{x} &= Ax\,, \\ \dot{y} &= -y\,, \\ \dot{z} &= z\,. \end{aligned} \tag{16}$$

Thus, the dynamics of the system (14) is determined by the system $\dot{x} = Ax$ which is the restriction of (14) on the center subspace E^c. Our goal is to justify a similar "reduction principle" for nonlinear systems of differential equations.

Summarizing the results of V. Pliss [59,60], A. Kelley [38], N. Fenichel [19], and M. Hirsch, C. Pugh, M. Shub [34] we obtain the following theorem.

Theorem 1 (Existence Theorem) *Consider a C^k-smooth ($1 \le k < \infty$) system of differential equations $\dot{v} = F(v)$, $F(O) = 0$ and $L = DF(O)$. Let E^s, E^u and E^c be the stable, unstable, and central eigenspaces of L.*

Then near the equilibrium O there exist the following five C^k-smooth invariant manifolds (see Fig. 2):

- *The center manifold W^c, tangent to E^c at O;*
- *The stable manifold W^s, tangent to E^s at O;*
- *The unstable manifold W^u, tangent to E^u at O;*
- *The center-stable manifold W^{cs}, tangent to $E^c + E^s$ at O;*
- *The center-unstable manifold W^{cu}, tangent to $E^c + E^u$ at O;*

The stable and unstable manifolds are unique, but center, center-stable, and center-unstable manifolds are not necessarily unique.

Solutions on W^s have exponential tending to O as $t \to \infty$. Solutions on W^u have exponential tending to O as $t \to -\infty$.

Remark 1 The expression "near the equilibrium O" means that there exists a neighborhood $U(O)$, depending on F and k, where the statements of the Existence Theorem hold.

Remark 2 The linearized system $\dot{v} = DF(O)v$ has the invariant subspace $E^s + E^u$ whereas the complete system $\dot{v} = F(v)$ may not have any smooth invariant manifolds with the tangent space $E^s + E^u$ at O, see [5,30].

Representation of the Center Manifold

At first we suppose that our system does not have the unstable eigenspace and is transformed by the linear change

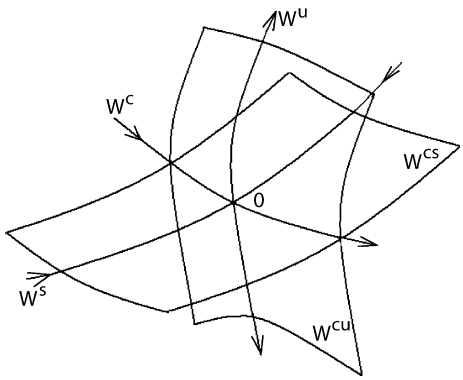

Center Manifolds, Figure 2
The classical five invariant manifolds

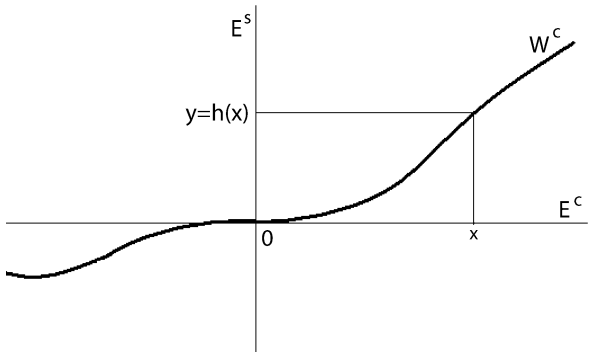

Center Manifolds, Figure 3
Representation of a center manifold

of coordinates mentioned above to the form

$$\begin{aligned} \dot{x} &= Ax + f(x, y)\,, \\ \dot{y} &= By + g(x, y)\,, \end{aligned} \tag{17}$$

where $E^c = \{(x, 0)\}$, $E^s = \{(0, y)\}$. Since the center manifold W^c is tangent to $E^c = \{(x, 0)\}$ at the origin $O = (0, 0)$, it can be represented near O in the form

$$W^c = \{(x, y) \,|\, |x| < \varepsilon,\ y = h(x)\}\,.$$

In other words the center manifold is represented as the graph of a smooth mapping $h\colon V \subset R^c \to R^s$, where c and s are dimensions of the center and stable subspaces, see Fig. 3. Since W^c goes through the origin and is tangent to E^s, we have the equalities

$$h(0) = 0\,, \quad Dh(0) = 0\,. \tag{18}$$

The invariance of W^c means that if an initial point (x, y) is in W^c, i. e. $y = h(x)$ then the solution $(X(t, x, y), Y(t, x, y))$ of (17) is in W^c, i. e. $Y(t, x, y)) = h(X(t, x, y))$. Thus, we get the invariance condition

$$Y(t, x, h(x))) = h(X(t, x, h(x)))\,. \tag{19}$$

Differentiating (19) with respect to t and putting $t = 0$ we get

$$Bh(x) + g(x, h(x)) = Dh(x)(Ax + f(x, h(x)))\,. \tag{20}$$

Thus, the mapping h have to be a solution of the partial differential equation (20) and satisfy (18).

Suppose that the system has the form

$$\begin{aligned} \dot{x} &= Ax + f(x, y, z)\,, \\ \dot{y} &= By + g_1(x, y, z)\,, \\ \dot{z} &= Cz + g_2(x, y, z)\,, \end{aligned} \tag{21}$$

where A has the eigenvalues with zero real part, B has the eigenvalues with negative real part, C has the eigenvalues with positive real part. Analogously to the previous case we can show that the center manifold is represented in the form

$$W^c = \{(x, y, z) \,|\, |x| < \epsilon,\ y = h_1(x),\ z = h_2(x)\}$$

where the mappings h_1 and h_2 satisfy the equations

$$\begin{aligned} &Bh_1(x) + g_1(x, h_1(x), h_2(x)) \\ &\qquad = Dh_1(x)(Ax + f(x, h_1(x), h_2(x))) \\ &Ch_2(x) + g_2(x, h_1(x), h_2(x)) \\ &\qquad = Dh_2(x)(Ax + f(x, h_1(x), h_2(x))) \end{aligned} \tag{22}$$

and the conditions

$$\begin{aligned} h_1(0) &= 0\,, \quad Dh_1(0) = 0 \\ h_2(0) &= 0\,, \quad Dh_2(0) = 0\,. \end{aligned} \tag{23}$$

Uniqueness and Nonuniqueness of the Center Manifold

The center manifold is nonunique in general.

Let us consider the illustrating example from [38]. The system has the form

$$\begin{aligned} \dot{x} &= x^2\,, \\ \dot{y} &= -y\,, \end{aligned} \tag{24}$$

where $(x, y) \in R^2$. It is obviously that the origin $(0, 0)$ is the equilibrium with single stable manifold $W^s = \{(0, y)\}$. There is a center manifold of the form $W^c = \{(x, 0)\}$. Moreover, there are the center manifolds that can be obtained when solving the equation

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{-y}{x^2}\,. \tag{25}$$

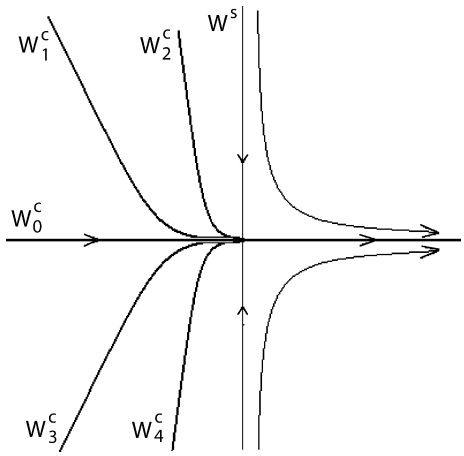

Center Manifolds, Figure 4
Nonuniqueness of the center manifold

The solution of (24) has the form

$$y(x) = a \exp\left(\frac{1}{x}\right) \tag{26}$$

for $x \neq 0$ and any constant $a \in R$. It follows that

$$W^c(a) = \{(x, y) | y = a \exp\left(\frac{1}{x}\right) \text{ as } x < 0 \text{ and } y = 0 \text{ as } x \geq 0\}$$

is center manifold for each a, being C^∞-smooth and pasting together the curve $\{y = a \exp(1/x)\}$ as $x < 0$ and the straight line $\{y = 0\}$ as $x \geq 0$., see Fig. 4.

However, the center manifolds possess the following weak uniqueness property. Let U be a neighborhood of the W^c. It turns out that the maximal invariant set I in U must be in W^c, that follows from the Reduction Theorem, see below. Consequently, all center manifolds contact on I. In this case the center manifold is called locally maximal. Thus, the center manifolds may differ on the trajectories leaving the neighborhood as $t \to \infty$ or $-\infty$.

For example, suppose that a center manifold is two-dimensional and a limit cycle is generated from the equilibrium through the Hopf bifurcation [31,35]. In this case the invariant set I is a disk bounded by the limit cycle and all center manifolds will contain this disk, see Fig. 5.

Smoothness

M. Hirsch, C. Pugh, M. Shub [34] proved that the center manifold is C^k-smooth if the system is C^k-smooth and $k < \infty$. Moreover if the kth derivative of the vector field is α-Hölder or Lipschitz mapping, the center manifold is the same.

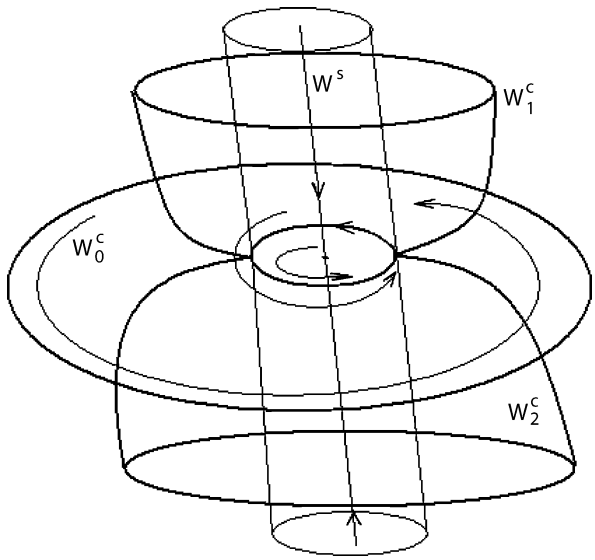

Center Manifolds, Figure 5
Weak uniqueness of the center manifold

However, if the system is C^∞ or analytic then the center manifold is not necessary the same. First consider the analytic case. It is clear that if the analytic center manifold exists then it is uniquely determined by applying the Taylor power series expansion. Consider the illustrating example [31]

$$\begin{aligned} \dot{x} &= -x^2 , \\ \dot{y} &= -y + x^2 \end{aligned} \tag{27}$$

which does not have an analytic center manifold. In fact, applying the Taylor power series expansions we obtain that the center manifold has to be given by the series $y = \sum_{n=2}^{\infty}(n-1)!x^n$ which diverges when $x \neq 0$.

If the system is C^∞-smooth then the Existence Theorem guaranties the existence of a C^k-smooth center manifold for any $k < \infty$. However, the center manifold may not be C^∞-smooth. Van Strien S.J. [71] showed that the system

$$\begin{aligned} \dot{x} &= -x^2 + \mu^2 , \\ \dot{y} &= -y - (x^2 - \mu^2) , \\ \dot{\mu} &= 0 . \end{aligned} \tag{28}$$

does not have a C^∞-smooth center manifold. In fact if the system is C^∞-smooth then for each k there is a neighborhood $U_k(O)$ of a C^k-smooth central manifold W^c. There are the systems for which the sequence $\{U_k\}$ can shrink to O as $k \to \infty$ [25,71]. The results of the papers [19,34]

show that the smoothness of the invariant manifold depends on the relation between Lyapunov exponents on the center manifold and on the normal subspace (stable plus unstable subspaces). This relation σ is included in the concept of the normal hyperbolicity. At an equilibrium $\sigma = \min(\mathrm{Re}\lambda_n/\mathrm{Re}\lambda_c)$, where $\mathrm{Re}\lambda_n/\mathrm{Re}\lambda_c > 0$, λ_c is the eigenvalue on the center subspace and λ_n is the eigenvalue on the normal subspace. The condition $\sigma > 1$ is necessary for a persistence of smooth invariant manifold. One can guarantee degree of smoothness k of the manifold provided $k < \sigma$. Moreover, there exists examples showing that the condition $k < \sigma$ is essential. It means that the system has to contract (expand) along $E^{\mathrm{s}}(E^{\mathrm{u}})k$-times stronger than along the center manifold.

If a center manifold to the equilibrium O exists then $\mathrm{Re}\,\lambda_c = 0$ and $\sigma = \infty$ at O, but near O may be other equilibrium (or other orbits) on the center manifold where $\sigma < \infty$, and as a consequence, the center manifold may not be C^∞-smooth. Let us consider the illustrating example [25]

$$\begin{aligned} \dot{x} &= \mu x - x^3 \,, \\ \dot{y} &= y + x^4 \,, \\ \dot{\mu} &= 0 \,. \end{aligned} \tag{29}$$

The point $O = (0, 0, 0)$ is equilibrium, the system linearized at O has the form

$$\begin{aligned} \dot{x} &= 0 \,, \\ \dot{y} &= y \,, \\ \dot{\mu} &= 0 \,. \end{aligned} \tag{30}$$

The system (30) has the following invariant subspace: $E^{\mathrm{s}} = \{0, 0, 0\}$, $E^{\mathrm{c}} = \{x, 0, \mu\}$, $E^{\mathrm{u}} = \{0, y, 0\}$. The μ-axis consists of the equilibriums $(0, 0, \mu_0)$ for the system (29). The system linearized at $(0, \mu_0, 0)$ is the following

$$\begin{aligned} \dot{x} &= \mu_0 x \,, \\ \dot{y} &= y \,, \\ \dot{\mu} &= 0 \,. \end{aligned} \tag{31}$$

The eigenvalues on the center subspace are 0 and μ_0, the eigenvalue on the normal (unstable) subspace is 1. Therefore the degree of smoothness is bounded by $\sigma = 1/\mu_0$. Detailed information and examples are given in [10,25,31,36,43,76].

Reduction Principle

As we saw above, a smooth system of differential equations near an equilibrium point O can be written in the form

$$\begin{aligned} \dot{x} &= Ax + f(x, y, z) \,, \\ \dot{y} &= By + g(x, y, z) \,, \\ \dot{z} &= Cz + q(x, y, z) \,, \end{aligned} \tag{32}$$

where $O = (0, 0, 0)$, A has eigenvalues with zero real part, B has eigenvalues with negative real part and C has eigenvalues with positive real part; $f(0, 0, 0) = g(0, 0, 0) = q(0, 0, 0) = 0$ and $Df(0, 0, 0) = Dg(0, 0, 0) = Dq(0, 0, 0) = 0$. In this case the invariant subspaces at O are $E^{\mathrm{c}} = \{(x, 0, 0)\}$, $E^{\mathrm{s}} = \{(0, y, 0)\}$, and $E^{\mathrm{u}} = \{(0, 0, z)\}$. The center manifold has the form

$$W^{\mathrm{c}} = \{(x, y, z) | x \in V \subset R^{\mathrm{c}},\ y = h_1(x),\ z = h_2(x)\} \,, \tag{33}$$

where c is the dimension of the center subspace E^{c}, the mappings $h_1(x)$ and $h_2(x)$ are C^k-smooth, $k \geq 1$, $h_1(0) = h_2(0) = 0$ and $Dh_1(0) = Dh_2(0) = 0$. The last equalities mean that the manifold W^{c} goes through O and is tangent to E^{c} at O.

Summarizing the results of V. Pliss [59,60], A. Shoshitaishvili [68,69], A. Reinfelds [63], K. Palmer [55], Pugh C.C., Shub M. [62], Carr J. [10], Grobman D.M. [24], and F. Hartman [29] we obtain the following theorem.

Theorem 2 (Reduction Theorem) *The system of differential Equations (32) near the origin is topologically equivalent to the system*

$$\begin{aligned} \dot{x} &= Ax + f(x, h_1(x), h_2(x)) \,, \\ \dot{y} &= -y \,, \\ \dot{z} &= z \,. \end{aligned} \tag{34}$$

The first equation is the restriction of the system on the center manifold. The theorem allows to reduce the investigation of the dynamics near a nonhyperbolic fixed point to the study of the system on the center manifold.

Construction of the Center Manifold

The center manifold may be calculated by using simple iterations [10]. However, this method does not have wide application and we consider a method based on the Taylor power series expansion. First it was proposed by A. Lyapunov in [41]. Suppose that the unstable subspace is 0, that is the last equation of the system (32) is absent. Such systems have a lot of applications in stability theory. So we consider a system

$$\begin{aligned} \dot{x} &= Ax + f(x, y) \,, \\ \dot{y} &= By + g(x, y) \,, \end{aligned} \tag{35}$$

The center manifold is the graph of the mapping $y = h(x)$ which satisfies the equation

$$Dh(x)(Ax + f(x, h(x))) - Bh(x) - g(x, h(x)) = 0 \quad (36)$$

and the conditions $h(0) = 0$ and $Dh(0) = 0$. Let us try to solve the equation by applying the Taylor power series expansion. Denote the left part of (36) by N(h(x)). J. Carr [10] and D. Henry [32] proved the following theorem.

Theorem 3 *Let $\phi\colon V(0) \subset R^c \to R^s$ be a smooth mapping, $\phi(0) = 0$ and $D\phi(0) = 0$ such that $N(\phi(x)) = o(|x|^m)$ for some $m > 0$ as $|x| \to 0$ then*

$$h(x) - \phi(x) = o(|x|^m) \text{ as } |x| \to 0$$

where $r(x) = o(|x|^m)$ means that $r(x)/|x|^m \to 0$ as $|x| \to 0$.

Thus, if we solve the Eq. (36) with a desired accuracy we construct h with the same accuracy. Theorem 3 substantiates the application of the described method. Let us consider a simple example from [76]

$$\begin{aligned} \dot{x} &= x^2 y - x^5 \,, \\ \dot{y} &= -y + x^2 \,, \end{aligned} \quad (37)$$

$(x, y) \in R^2$, the equilibrium is at the origin $(0, 0)$. The eigenvalues of the linearized system are 0 and -1. According to the Existence Theorem the center manifold is locally represented in the form

$$W^c = \{(x, y) \,|\, y = h(x),\ |x| < \delta,\ h(0) = 0, Dh(0) = 0)\} \,,$$

where δ is sufficiently small. It follows that h has the form

$$h = ax^2 + bx^3 + \cdots \,. \quad (38)$$

The equation for the center manifold is given by

$$Dh(x)(Ax + f(x, h(x))) - Bh(x) - g(x, h(x)) = 0 \,, \quad (39)$$

where $A = 0$, $B = -1$, $f(x, y) = x^2 y - x^5$, $g(x, y) = x^2$. Substituting (38) in (39) we obtain the equality

$$(2ax + 3bx^2 + \cdots)(ax^4 + bx^5 - x^5 + \cdots) + ax^2 + bx^3 - x^2 + \cdots = 0 \,. \quad (40)$$

Equating coefficients on each power of x to zero we obtain

$$\begin{aligned} x^2 &: \quad a - 1 = 0, \Rightarrow a = 1 \\ x^3 &: \quad b = 0, \\ \vdots & \quad \vdots \,. \end{aligned} \quad (41)$$

Therefore we have

$$h(x) = x^2 + 0x^3 + \cdots \,. \quad (42)$$

In this connection we have to decide how many terms (powers of x) have be computed? The solution depends on our goal. For example, suppose that we study the Lyapunov stability of the equilibrium $(0, 0)$ for (37). According to the Reduction Theorem the system (37) is topologically equivalent to the system

$$\begin{aligned} \dot{x} &= x^2 h(x) - x^5 \,, \\ \dot{y} &= -y \,, \end{aligned} \quad (43)$$

where $h(x) = x^2 + 0x^3 + \cdots$. Substituting h we obtain the equation

$$\dot{x} = x^4 + \ldots \,. \quad (44)$$

Hence the equilibrium is unstable. It should be noted that the calculation of bx^3 is unnecessary, since substituting $h(x) = x^2 + bx^3 + \cdots$ in the first equation of the system (43), we also obtain (44). Thus, it is enough to compute the first term ax^2 of the mapping h.

This example brings up the following question. The system (37) is topologically equivalent to the system

$$\begin{aligned} \dot{x} &= x^4 + \cdots \,, \\ \dot{y} &= -y \,, \end{aligned} \quad (45)$$

which has many center manifolds. From this it follows that the initial system has a lot of center manifolds. Which of center manifolds is actually being found when approximating the center manifold via power series expansion?

It turns out [10,70,74] that any two center manifolds differ by transcendentally small terms, i. e. the terms of their Taylor expansions coincide up to all existing orders. For example, the system (24) has the center manifolds of the form

$$W^c(a) = \left\{(x, y) | y = a \exp\left(\tfrac{1}{x}\right) \text{ as } x < 0 \text{ and } y = 0 \text{ as } x \geq 0\right\}$$

for any a. However, each manifold has null-expansion at 0.

Center Manifold in Discrete Dynamical Systems

Consider a dynamical system generated by the mapping

$$v \to Lv + G(v) \,, \quad (46)$$

where $v \in R^n$, L is a matrix and G is second-order at the origin. Without loss of generality we consider the system (46) as a perturbation of the linear system

$$x \to Lx \,. \quad (47)$$

Divide the eigenvalues of L into three parts:
stable $\sigma_s = \{$eigenvalues with modulus less than 1$\}$,
unstable $\sigma_u = \{$eigenvalues with modulus greater than 1$\}$,
critical $\sigma_c = \{$eigenvalues with modulus equal to 1$\}$.
The matrix L has three eigenspaces: the stable subspace E^s, the unstable subspace E^u, and the central subspace E^c, which correspond to the mentioned spectrum parts respectively, being $E^s \oplus E^u \oplus E^c = R^n$. The next theorem follows from the theorem on perturbation of ρ-hyperbolic endomorphism on a Banach space [34].

Theorem 4 (Existence Theorem) *Consider a C^k-smooth ($1 \le k < \infty$) discrete system*

$$v \to Lv + G(v) ,$$

$G(0) = 0$ and G is C^1-close to zero. Let E^s, E^u and E^c be the stable, unstable, and central eigenspaces of the linear mapping L.

Then near the fixed point O there exist the following C^k-smooth invariant manifolds:

- *The center manifold W^c tangent to E^c at O,*
- *The stable manifold W^s tangent to E^s at O,*
- *The unstable manifold W^u tangent to E^u at O,*
- *The center-stable manifold W^{cs} tangent to $E^c + E^s$ at O,*
- *The center-unstable manifold W^{cu} tangent to $E^c + E^u$ at O.*

The stable and unstable manifolds are unique, but center, center-stable, and center-unstable manifolds may not be unique.

The orbits on W^s have exponential tending to O as $m \to \infty$. The orbits on W^u have exponential tending to O as $m \to -\infty$.

There exists a linear change of coordinates transforming (46) to the form

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \to \begin{pmatrix} Ax + f(x,y,z) \\ By + g_1(x,y,z) \\ Cz + g_2(x,y,z) \end{pmatrix} , \tag{48}$$

where A has eigenvalues with modulus equal 1, B has eigenvalues with modulus less than 1, C has eigenvalues with modulus greater than 1. The system (48) at the origin has the following invariant eigenspaces: $E^c = \{(x,0,0)\}$, $E^s = \{(0,y,0)\}$, and $E^u = \{(0,0,z)\}$. Since the center manifold W^c is tangent to E^c at the origin $O = (0,0)$, it can be represented near O in the form

$$W^c = \{(x,y,z) \,|\, |x| < \epsilon,\ y = h_1(x), z = h_2(x)\} ,$$

where h_1 and h_2 are second-order at the origin, i. e. $h_{1,2}(0) = 0$, $Dh_{1,2}(0) = 0$. The invariance of W^c means that if a point $(x,y,z) \in W^c$, i. e. $y = h_1(x)$, $z = h_2(x)$, then $(Ax + f(x,y,z), By + g_1(x,y,z), Cz + g_2(x,y,z)) \in W^c$, i. e. $By + g_1(x,y,z) = h_1(Ax + f(x,y,z))$, $Cz + g_2(x,y,z) = h_2(Ax + f(x,y,z))$. Thus, we get the invariance property

$$\begin{aligned} Bh_1(x) + g_1(x, h_1(x), h_2(x)) \\ = h_1(Ax + f(x, h_1(x), h_2(x)) , \\ Ch_2(x) + g_2(x, h_1(x), h_2(x)) \\ = h_2(Ax + f(x, h_1(x), h_2(x)) . \end{aligned} \tag{49}$$

Results on smoothness and uniqueness of center manifold for discrete systems are the same as for continuous systems.

Theorem 5 (Reduction Theorem [55,62,63,64]) *The discrete system (48) near the origin is topologically equivalent to the system*

$$\begin{aligned} x &\to Ax + f(x, h_1(x), h_2(x)) , \\ y &\to By , \\ z &\to Cz . \end{aligned} \tag{50}$$

The first mapping is the restriction of the system on the center manifold. The theorem reduces the investigation of dynamics near a nonhyperbolic fixed point to the study of the system on the center manifold.

Normally Hyperbolic Invariant Manifolds

As it is indicated above, a center manifold can be considered as a perturbed invariant manifold for the center subspace of the linearized system. The normal hyperbolicity conception arises as a natural condition for the persistence of invariant manifold under a system perturbation. Informally speaking, an f-invariant manifold M, where f-diffeomorphism, is normally hyperbolic if the action of Df along the normal space of M is hyperbolic and dominates over its action along the tangent one. (the rigorous definition is below). Many of the results concerning center manifold follows from the theory of normal hyperbolic invariant manifolds. In particular, the results related to center eigenvalues of Df with $|\lambda| \neq 1$ may be obtained using this theory.

The problem of existence and preservation for invariant manifolds has a long history. Initial results on invariant (integral) manifolds of differential equations were obtaned by Hadamard [26] and Perron [57], Bogoliubov, Mitropolskii [8] and Lykova [48], Pliss [60], Hale [28], Diliberto [16] and other authors (for references see [39,77]). In the late 1960s and throughout 1970s the

theory of perturbations of invariant manifolds began to assume a very general and well-developed form. Sacker [66] and Neimark [50] proved (independently and by different methods) that a normally hyperbolic compact invariant manifold is preserved under C^1 perturbations. Final results on preservation of invariant manifolds were obtained by Fenichel [19,20] and Hirsch, Pugh, Shub [34]. The linearization theorem for a normally hyperbolic manifold was proved in [62].

Let $f: R^n \to R^n$ be a C^r-diffeomorphism, $1 \le r < \infty$, and a compact manifold M is f-invariant.

Definition 4 An invariant manifold M is called r-normally hyperbolic if there exists a continuous invariant decomposition

$$TR^n|_M = TM \oplus E^s \oplus E^u \tag{51}$$

of the tangent bundle TR^n into a direct sum of subbundles TM, E^s, E^u and constants a, $\lambda > 0$ such that for $0 \le k \le r$ and all $p \in M$,

$$\begin{aligned} |D^s f^n(p)|\ |(D^0 f^n(p))^{-1}|^k &\le a \exp(-\lambda n) \quad \text{for } n \ge 0\,, \\ |D^u f^n(p)|\ |(D^0 f^n(p))^{-1}|^k &\le a \exp(\lambda n) \quad \text{for } n \le 0\,. \end{aligned} \tag{52}$$

Here $D^0 f$, $D^s f$ and $D^u f$ are restrictions of Df on TM, E^s and E^u, respectively.

The invariance of the bundle E^* means that for every $p \in M_0$, the image of the fiber $E^*(p)$ at p under $Df(p)$ is the fiber $E^*(q)$ at $q = f(p)$. The bundles E^s and E^u are called stable and unstable, respectively. In other words, M is normally hyperbolic if the differential Df contracts (expands) along the normal (to M) direction and this contraction (expansion) is stronger in r times than a conceivable contraction (expansion) along M. Summarizing the results of [19,20,34,50,62,66] we obtain the following theorem.

Theorem 6 *Let a C^r diffeomorphism f be r-normally hyperbolic on a compact invariant manifold M with the decomposition $TR^n|_M = TM \oplus E^s \oplus E^u$. Then*

- *there exist invariant manifolds W^s and W^u near M, which are tangent at M to $TM \oplus E^s$ and $TM \oplus E^u$;*
- *the manifolds M, W^s and W^u are C^r-smooth;*
- *if g is another C^r-diffeomorphism C^r-close to f, then there exists the unique manifold M_g which is invariant and r-normally hyperbolic for g;*
- *near M f is topologically equivalent to $Df|_{E^s \oplus E^u}$, which in local coordinates has the form*

$$\begin{aligned} (x, y, z) \to (f(x), Df(x)y, Df(x)z)\,, \\ x \in M\,, \quad y \in E^s(x)\,, \quad z \in E^u(x)\,. \end{aligned} \tag{53}$$

Similarly result takes place for flows. Mañé [44] showed that a locally unique, preserved, compact invariant manifold is necessarily normally hyperbolic. Local uniqueness means that near M, an invariant set I of the perturbation g is in M_g. Conditions for the preservation of locally nonunique compact invariant manifolds and linearization were found by G. Osipenko [52,53].

Applications

Stability

As it was mentioned above, it was Lyapunov who pioneered in applying the concept of center manifold to establish the stability of equilibrium. V. Pliss [59] proved the general reduction principle in stability theory. According to this principle the equilibrium is stable if and only if it is stable on the center manifold. We have considered the motivating example where the center manifold concept was used. The system

$$\dot{x} = xy\,, \quad \dot{y} = -y + ax^2\,, \tag{54}$$

has the center manifold $y = h(x)$. Applying Theorem 3, we obtain

$$h(x) = ax^2 + \cdots\,.$$

The reduced system is of the form

$$\dot{x} = ax^3 + \ldots\,.$$

Thus, if $a < 0$, the origin is stable. The book [25] contains other examples of investigation of stability.

Bifurcations

Consider a parametrized system of differential equations

$$\dot{x} = A(\mu)x + f(x, \mu)\,, \tag{55}$$

where $\mu \in R^k$, f is C^1-small when μ is small. To study the bifurcation problem near $\mu = 0$ it is useful to deal with the extended system of the form

$$\begin{aligned} \dot{x} &= A(\mu)x + f(x, \mu)\,, \\ \dot{\mu} &= 0\,. \end{aligned} \tag{56}$$

Suppose that (55) has a n-dimensional center manifold for $\mu = 0$. Then the extended system (56) has a $n + k$-dimensional center manifold. The reduction principle guarantees that all bifurcations lie on the center manifold. Moreover, the center manifold has the form $y = h(x, \mu_0)$ for every μ_0 which is a solution of the equation $\dot{\mu} = 0$ and the mapping h may be presented as a power series in μ. Further interesting information regarding to recent advancements in the approximation and computation of center manifolds is in [37]. The book [31] deals with Hopf bifurcation and contains a lot of examples and applications. Invariant manifolds and foliations for nonautonomous systems are considered in [3,13]. Partial linearization for noninvertible mappings is studied in [4].

Center Manifold in Infinite-Dimensional Space

Center manifold theory is a standard tool to study the dynamics of discrete and continuous dynamical systems in infinite dimensional phase space. We start with the discrete dynamical system generated by a mapping of a Banach space. The existence of the center manifold of the mappings in Banach space was proved by Hirsch, Pugh and Shub [34]. Let $T\colon E \to E$ be a linear endomorphism of a Banach space E.

Definition 5 A linear endomorphism T is said to be ρ-hyperbolic, $\rho > 0$, if no eigenvalue of T has modulus ρ, i. e., its spectrum $\mathrm{Spect}(T)$ has no points on the circle of radius ρ in the complex plane C. A linear 1-hyperbolic isomorphism T is called hyperbolic.

For a ρ-hyperbolic linear endomorphism there exists a T-invariant splitting of $E = E_1 \oplus E_2$ such that the spectrum of the restriction $T_1 = T|_{E_1}$ lies outside of the disk of the radius ρ, where as the spectrum of $T_2 = T|_{E_2}$ lies inside it. So T_1 is automorphism and T_1^{-1} exists. It is known (for details see [34]) that one can define norms in E_1 and E_2 such that associated norms of T_1^{-1} and T_2 may be estimated in the following manner

$$|T_1^{-1}| < \tfrac{1}{\rho}\,, \quad |T_2| < \rho\,. \tag{57}$$

Conversely, if T admits an invariant splitting $T = T_1 \oplus T_2$ with $|T_1^{-1}| \equiv a$, $|T_2| \equiv b$, $ab < 1$ then $\mathrm{Spect}(T_1)$ lies outside the disk $\{\lambda \in C\colon |\lambda| < 1/a\}$, and $\mathrm{Spect}(T_2)$ lies in the disk $\{\lambda \in C\colon |\lambda| \le b\}$. Thus, T is ρ-hyperbolic with $b < \rho < 1/a$.

Theorem 7 *Let T be a ρ-hyperbolic linear endomorphism, $\rho \le 1$. Assume that $f\colon E \to E$ is C^r, $r \ge 1$, $f = T + g$, $f(0) = 0$ and there is $\delta > 0$ such that if g is a Lipschitz mapping and*

$$\mathrm{Lip}(g) = \mathrm{Lip}(f - T) \le \delta\,, \tag{58}$$

then there exists a manifold W_f, which is a graph of a C^1 map $\varphi\colon E_1 \to E_2$, i. e.

$$W_f = \{x + y \mid y = \varphi(x),\ x \in E_1,\ y \in E_2\}\,,$$

with the following properties:

- *W_f is f-invariant;*
- *if $\|T_1^{-1}\|^j \|T_2\| < 1$, $j = 1, \ldots, r$ then W_f is C^r and depends continuously on f in the C^r topology;*
- *$W_T = E_1$;*
- *for $x \in W_f$*

$$|f^{-1}(x)| \le (a + 2\varepsilon)\,|x|\,, \tag{59}$$

  where $a < 1/\rho$, and ε is small provided δ is small;
- *if $Df(0) = T$ then W_f is tangent to E_1 at 0.*

Suppose that the spectrum of T is contained in $A_1 \bigcup A_2$ where

$$A_1 = \{z \in C,\ |z| \ge 1\}\,, \quad A_2 = \{z \in C,\ |z| \le a < 1\}\,.$$

Corollary 1 *If $f\colon E \to E$ is C^r, $1 \le r < \infty$, $f(0) = 0$ and $\mathrm{Lip}(f - T) \le \epsilon$ is small, then there exists a center-unstable manifold $W^{\mathrm{cu}} = W_f$ which is the graph of a C^r function $E_1 \to E_2$. The center-unstable manifold is attractor, i. e. for any $x \in E$ the distance between $f^n(x)$ and W^{cu} tends to zero as $n \to \infty$. The manifold $W_f = W^{\mathrm{c}}$ is center manifold when $A_1 = \{z \in C,\ |z| = 1\}$.*

Apply this theorem to a mapping of the general form

$$u \to Au + f(u)\,, \tag{60}$$

where E is a Banach space, $u \in E$, $A\colon E \to E$ is a linear operator, f is second-order at the origin. The question arises of whether there exists a smooth mapping g such that g coincides with f in a sufficiently small neighborhood of the origin and g is C^1-close to zero? The answer is positive if E admits C^∞-norm, i. e. the mapping $u \to \|u\|$ is C^∞-smooth for $u \ne 0$. It is performed for Hilbert space. The desired mapping g may be constructed with the use of a cut-off function. To apply Theorem 7 to (60) we have to suppose *Hypothesis on C^∞-norm*: the Banach space E admits C^∞-norm.

Thus, Theorem 7 together with Hypothesis on C^∞-norm guarantee an existence of center manifold for the system (60) in a Banach space. A. Reinfelds [63,64] prove a reduction theorem for homeomorphism in a metric space from which an analog of Reduction Theorem 5 for Banach space follows.

Flows in Banach Space

Fundamentals of center manifold theory for flows and differential equations in Banach spaces can be found in [10,21,43,73]. In Euclidean space a flow is generated by a solution of a differential equation and the theory of ordinary differential equations guarantees the clear connection between flows and differential equations. In infinite dimensional (Banach) space this connection is more delicate and at first, we consider infinite dimensional flows and then (partial) differential equations.

Definition 6 A flow (semiflow) on a domain D is a continuous family of mappings $F_t\colon D \to D$ where $t \in R\,(t \geq 0)$ such that $F_0 = I$ and $F_{t+s} = F_t F_s$ for $t, s \in R\,(t, s \geq 0)$.

Theorem 8 (Center Manifold for Infinite Dimensional Flows [43]) *Let the Hypothesis on C^∞-norm be fulfilled and $F_t\colon E \to E$ be a semiflow defined near the origin for $0 \leq t \leq \tau$, $F_t(0) = 0$. Suppose that the mapping $F_t(x)$ is C^0 in t and C^{k+1} in x and the spectrum of the linear semigroup $DF_t(0)$ is of the form $e^{t(\sigma_1 \cup \sigma_2)}$, where $e^{t(\sigma_1)}$ is on the unit circle (i. e. σ_1 is on the imaginary axis) and $e^{t(\sigma_2)}$ is inside the unit circle with a positive distance from one as $t > 0$ (i. e. σ_2 is in the left half-plane). Let $E = E_1 \oplus E_2$ be the $DF_t(0)$-invariant decomposition corresponding to the decomposition of the spectrum $e^{t(\sigma_1 \cup \sigma_2)}$ and $dim E_1 = d < \infty$.*

Then there exists a neighborhood U of the origin and C^k-submanifold $W^c \subset U$ of the dimension d such that

- *W^c is invariant for F_t in U, $0 \in W^c$ and W^c is tangent to E_1 at 0;*
- *W^c is attractor in U, i. e. if $F_t^n(x) \in U$, $n = 0, 1, 2, \ldots$ then $F_t^n(x) \to W^c$ as $n \to \infty$.*

In infinite dimensional dynamics it is also popular notion of "inertial manifold" which is an invariant finite dimensional attracting manifold corresponding to W^c of the Theorem 8.

Partial Differential Equations

At the present the center manifold method takes root in theory of partial differential equations. Applying center manifold to partial differential equations leads to a number of new problems. Consider an evolution equation

$$\dot{u} = Au + f(u)\,, \tag{61}$$

where $u \in E$, E is a Banach space, A is a linear operator, f is second-order at the origin. Usually A is a differential operator defined on a domain $D \subset E$, $D \neq E$. To apply Theorem 8 we have to construct a semiflow F_t defined near the origin of E. As a rule, an infinite dimensional phase space E is a functional space where the topology may be chosen by the investigator. A proper choice of functional space and its topology may essentially facilitate the study. As an example we consider a nonlinear heat equation

$$u_t' = \Delta u + f(u)\,, \tag{62}$$

where u is defined on a bounded domain $\Omega \subset R^n$ with Dirichlet condition on the boundary $u|_{\partial\Omega} = 0$, $\Delta u = (\partial^2/\partial x_1^2 + \cdots + \partial^2/\partial x_n^2)u$ is the Laplace operator defined on the set

$$C_0^2 = \{u \in C^2(\overline{\Omega}),\ u = 0 \text{ on } \partial\Omega\}\,.$$

According to [49] we consider a Hilbert space $E = L^2(\Omega)$ with the inner product $< u, v >= \int_\Omega uvdx$. The operator Δ may be extended to a self-conjugate operator $A\colon D_A \to L^2(\Omega)$ with the domain D_A that is the closure of C_0^2 in $L^2(\Omega)$. Consider the linear heat equation $u_t' = \Delta u$ and its extension

$$u_t' = Au, \quad u \in D_A\,. \tag{63}$$

What do we mean by a flow generated by (63)?

Definition 7 A linear operator A is the infinitisemal generator of a continuous semigroup $U(t)$, $t \geq 0$ if:
$U(t)$ is a linear mapping for each t;
$U(t)$ is a semigroup and $\|U(t)u - u\| \to 0$ as $t \to +0$;

$$Au = \lim_{t \to +0} \frac{U(t)u - u}{t} \tag{64}$$

whenever the limit exists.

For Laplace operator we have $< \Delta u, u >\leq 0$ for $u \in C_0^2$, hence the operator A is dissipative. It guarantees [6] the existence of the semigroup $U(t)$ with the infinitesimal generator A. The mapping $U(t)$ is the semigroup $DF_t(0)$ in Theorem 8.

The next problem is the relation between the linearized system $u' = Au$ and full nonlinear Equation (61). To consider the mapping f in (62) as C^1-small perturbation, the Hypothesis on C^∞-norm has to be fulfilled. To prove the existence of the solution of the nonlinear equation with $u(0) = u_0$ one uses the Duhamel formula

$$u(t) = U_t u_0 + \int_0^t U_{t-s} f(u(s))\mathrm{d}s$$

and the iterative Picard method. By this way we construct the semiflow required to apply Theorem 8. The relation between the spectrum of a semigroup $U(t)$ and the spec-

trum of its infinitisemal generator gives rise to an essential problem. All known existence theorems are formulated in terms of the semigroup as in Theorem 8. The existence of the spectrum decompositions of a semigroup $U(t)$ and the corresponding estimations on the appropriate projections are supposed. This framework is inconvenient for many applications, especially in partial differential equations. In finite dimensions, a linear system $\dot{x} = Ax$ has a solution of the form $U(t)x = e^{At}x$ and λ is an eigenvalue of A if and only if $e^{\lambda t}$ is an eigenvalue of $U(t)$. We have the spectral equality

$$\mathrm{Spect}(U(t)) = e^{\mathrm{Spect}(A)t} .$$

In infinite dimensions, relating the spectrum of the infinitesimal generator A to that of the semigroup $U(t)$ is a spectral mapping problem which is often nontrivial. The spectral inclusion

$$\mathrm{Spect}(U(t)) \subset e^{\mathrm{Spect}(A)t}$$

always holds and the inverse inclusion is a problem that is solved by the spectral mapping theorems [12,18,22,65].

Often an application of center manifold theory needs to prove an appropriate version of the reduction theorem [7,11,12,14,15,18]. Pages 1–5 of the book [40] give an extensive list of the applications of center manifold theory to infinite dimensional problems. Here we consider a few of typical applications. Reaction-diffusion equations are typical models in chemical reaction (Belousov–Zhabotinsky reaction), biological systems, population dynamics and nuclear reaction physics. They have the form

$$u_t' = (K(\lambda) + D\Delta)u + f(u, \lambda) , \tag{65}$$

where K is a matrix depending on a parameter, D is a symmetric, positive semi-definite, often diagonal matrix, Δ is the Laplace operator, λ is a control parameter. The papers [2,27,31,67,78] applied center manifold theory to the study of the reaction-diffusion equation. Invariant manifolds for nonlinear Schrodinger equations are studied in [22,40]. In the elliptic quasilinear equations the spectrum of the linear operator is unbounded in unstable direction and in this case the solution does not generate even a semiflow. Nevertheless, center manifold technique is successfully applied in this case as well [45]. Henry [32] proved the persistence of normally hyperbolic closed linear subspace for semilinear parabolic equations. It follows the existence of the center manifold which is a perturbed manifold for the center subspace. I. Chueshov [14,15] considered a reduction principle for coupled nonlinear parabolic-hyperbolic partial differential equations that has applications in thermoelastisity. Mielke [47] has developed the center manifold theory for elliptic partial differential equations and has applied it to problems of elasticity and hydrodynamics. The results for non-autonomous system in Banach space can be found in [46]. The reduction principle for stochastic partial differential equations is considered in [9,17,75].

Future Directions

Now we should realize that principal problems in the center manifolds theory of finite dimensional dynamics have been solved and we may expect new results in applications. However the analogous theory for infinite dimensional dynamics is far from its completion. Here we have few principal problems such as the reduction principle (an analogue of Theorem 2 or 5) and the construction of the center manifold. The application of the center manifold theory to partial differential equations is one of promising and developing direction. In particular, it is very important to investigate the behavior of center manifolds when Galerkin method is applied due to a transition from finite dimension to infinite one. As it was mentioned above each application of center manifold methods to nonlinear infinite dimensional dynamics is nontrivial and gives rise to new directions of researches.

Bibliography

Primary Literature

1. Arnold VI (1973) Ordinary Differential Equations. MIT, Cambridge
2. Auchmuty J, Nicolis G (1976) Bifurcation analysis of reaction-diffusion equation (III). Chemical Oscilations. Bull Math Biol 38:325–350
3. Aulbach B (1982) A reduction principle for nonautonomous differential equations. Arch Math 39:217–232
4. Aulbach B, Garay B (1994) Partial linearization for noninvertible mappings. J Appl Math Phys (ZAMP) 45:505–542
5. Aulbach B, Colonius F (eds) (1996) Six Lectures on Dynamical Systems. World Scientific, New York
6. Balakrishnan AV (1976) Applied Functional Analysis. Springer, New York, Heidelberg
7. Bates P, Jones C (1989) Invariant manifolds for semilinear partial differential equations. In: Dynamics Reported 2. Wiley, Chichester, pp 1–38
8. Bogoliubov NN, Mitropolsky YUA (1963) The method of integral manifolds in non-linear mechanics. In: Contributions Differential Equations 2. Wiley, New York, pp 123–196
9. Caraballo T, Chueshov I, Landa J (2005) Existence of invariant manifolds for coupled parabolic and hyperbolic stochastic partial differential equations. Nonlinearity 18:747–767
10. Carr J (1981) Applications of Center Manifold Theory. In: Applied Mathematical Sciences, vol 35. Springer, New York

11. Carr J, Muncaster RG (1983) Applications of center manifold theory to amplitude expansions. J Diff Equ 59:260–288
12. Chicone C, Latushkin Yu (1999) Evolution Semigroups in Dynamical Systems and Differential Equations. Math Surv Monogr 70. Amer Math Soc, Providene
13. Chow SN, Lu K (1995) Invariant manifolds and foliations for quasiperiodic systems. J Diff Equ 117:1–27
14. Chueshov I (2007) Invariant manifolds and nonlinear master-slave synchronization in coupled systems. In: Applicable Analysis, vol 86, 3rd edn. Taylor and Francis, London, pp 269–286
15. Chueshov I (2004) A reduction principle for coupled nonlinesr parabolic-hyperbolic PDE. J Evol Equ 4:591–612
16. Diliberto S (1960) Perturbation theorem for periodic surfaces I, II. Rend Cir Mat Palermo 9:256–299; 10:111–161
17. Du A, Duan J (2006) Invariant manifold reduction for stochastic dynamical systems. http://arXiv:math.DS/0607366
18. Engel K, Nagel R (2000) One-parameter Semigroups for Linear Evolution Equations. Springer, New York
19. Fenichel N (1971) Persistence and smoothness of invariant manifolds for flows. Ind Univ Math 21:193–226
20. Fenichel N (1974) Asymptotic stability with rate conditions. Ind Univ Math 23:1109–1137
21. Gallay T (1993) A center-stable manifold theorem for differential equations in Banach space. Commun Math Phys 152:249–268
22. Gesztesy F, Jones C, Latushkin YU, Stanislavova M (2000) A spectral mapping theorem and invariant manifolds for nonlinear Schrodinger equations. Ind Univ Math 49(1):221–243
23. Grobman D (1959) Homeomorphism of system of differential equations. Dokl Akad Nauk SSSR 128:880 (in Russian)
24. Grobman D (1962) The topological classification of the vicinity of a singular point in *n*-dimensional space. Math USSR Sbornik 56:77–94; in Russian
25. Guckenheimer J, Holmes P (1993) Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vectors Fields. Springer, New York
26. Hadamard J (1901) Sur l'etaration et les solution asymptotiques des equations differentielles. Bull Soc Math France 29:224–228
27. Haken H (2004) Synergetics: Introduction and Advanced topics. Springer, Berlin
28. Hale J (1961) Integral manifolds of perturbed differential systems. Ann Math 73(2):496–531
29. Hartman P (1960) On local homeomorphisms of Euclidean spaces. Bol Soc Mat Mex 5:220
30. Hartman P (1964) Ordinary Differential Equations. Wiley, New York
31. Hassard BD, Kazarinoff ND, Wan YH (1981) Theory and Applications of Hopf Bifurcation. Cambridge University Press, Cambridge
32. Henry D (1981) Geometric theory of semilinear parabolic equations. Lect Notes Math 840:348
33. Hirsch M, Smale S (1974) Differential Equations, Dynamical Systems and Linear Algebra. Academic Press, Orlando
34. Hirsch M, Pugh C, Shub M (1977) Invariant manifolds. Lect Notes Math 583:149
35. Hopf E (1942) Abzweigung einer periodischen Lösung von einer stationären Lösung eines Differentialsystems. Ber Verh Sachs Akad Wiss Leipzig Math-Nat 94:3–22
36. Iooss G (1979) Bifurcation of Maps and Application. N-Holl Math Stud 36:105
37. Jolly MS, Rosa R (2005) Computation of non-smooth local centre manifolds. IMA J Numer Anal 25(4):698–725
38. Kelley A (1967) The stable, center stable, center, center unstable and unstable manifolds. J Diff Equ 3:546–570
39. Kirchgraber U, Palmer KJ (1990) Geometry in the Neighborhood of Invariant Manifolds of the Maps and Flows and Linearization. In: Pitman Research Notes in Math, vol 233. Wiley, New York
40. Li C, Wiggins S (1997) Invariant Manifolds and Fibrations for Perturbed Nonlinear Schrodinger Equations. Springer, New York
41. Lyapunov AM (1892) Problémé Générale de la Stabilité du Mouvement, original was published in Russia 1892, transtalted by Princeton Univ. Press, Princeton, 1947
42. Ma T, Wang S (2005) Dynamic bifurcation of nonlinear evolution equations and applications. Chin Ann Math 26(2): 185–206
43. Marsden J, McCracken M (1976) Hopf bifurcation and Its Applications. Appl Math Sci 19:410
44. Mañé R (1978) Persistent manifolds are normally hyperbolic. Trans Amer Math Soc 246:261–284
45. Mielke A (1988) Reduction of quasilinear elliptic equations in cylindrical domains with application. Math Meth App Sci 10:51–66
46. Mielke A (1991) Locally invariant manifolds for quasilinear parabolic equations. Rocky Mt Math 21:707–714
47. Mielke A (1996) Dynamics of nonlinear waves in dissipative systems: reduction, bifurcation and stability. In: Pitman Research Notes in Mathematics Series, vol 352. Longman, Harlow, pp 277
48. Mitropolskii YU, Lykova O (1973) Integral Manifolds in Nonlinear Mechanics. Nauka, Moscow
49. Mizohata S (1973) The Theory of Partial Differential Equations. Cambridge University Press, Cambridge
50. Neimark Y (1967) Integral manifolds of differential equations. Izv Vuzov, Radiophys 10:321–334 (in Russian)
51. Osipenko G, Ershov E (1993) The necessary conditions of the preservation of an invariant manifold of an autonomous system near an equilibrium point. J Appl Math Phys (ZAMP) 44:451–468
52. Osipenko G (1996) Indestructibility of invariant non-unique manifolds. Discret Contin Dyn Syst 2(2):203–219
53. Osipenko G (1997) Linearization near a locally non-unique invariant manifold. Discret Contin Dyn Syst 3(2):189–205
54. Palis J, Takens F (1977) Topological equivalence of normally hyperbolic dynamical systems. Topology 16(4):336–346
55. Palmer K (1975) Linearization near an integral manifold. Math Anal Appl 51:243–255
56. Palmer K (1987) On the stability of center manifold. J Appl Math Phys (ZAMP) 38:273–278
57. Perron O (1928) Über Stabilität und asymptotisches Verhalten der Integrale von Differentialgleichungssystem. Math Z 29:129–160
58. Pillet CA, Wayne CE (1997) Invariant manifolds for a class of dispersive. Hamiltonian, partial differential equations. J Diff Equ 141:310–326
59. Pliss VA (1964) The reduction principle in the theory of stability of motion. Izv Acad Nauk SSSR Ser Mat 28:1297–1324; translated (1964) In: Soviet Math 5:247–250
60. Pliss VA (1966) On the theory of invariant surfaces. In: Differential Equations, vol 2. Nauka, Moscow pp 1139–1150

61. Poincaré H (1885) Sur les courbes definies par une equation differentielle. J Math Pure Appl 4(1):167–244
62. Pugh C, Shub M (1970) Linearization of normally hyperbolic diffeomorphisms and flows. Invent Math 10:187–198
63. Reinfelds A (1974) A reduction theorem. J Diff Equ 10:645–649
64. Reinfelds A (1994) The reduction principle for discrete dynamical and semidynamical systems in metric spaces. J Appl Math Phys (ZAMP) 45:933–955
65. Renardy M (1994) On the linear stability of hyperbolic PDEs and viscoelastic flows. J Appl Math Phys (ZAMP) 45:854–865
66. Sacker RJ (1967) Hale J, LaSalle J (eds) A perturbation theorem for invariant Riemannian manifolds. Proc Symp Diff Equ Dyn Syst Univ Puerto Rico. Academic Press, New York, pp 43–54
67. Sandstede B, Scheel A, Wulff C (1999) Bifurcations and dynamics of spiral waves. J Nonlinear Sci 9(4):439–478
68. Shoshitaishvili AN (1972) Bifurcations of topological type at singular points of parameterized vector fields. Func Anal Appl 6:169–170
69. Shoshitaishvili AN (1975) Bifurcations of topological type of a vector field near a singular point. Trudy Petrovsky seminar, vol 1. Moscow University Press, Moscow, pp 279–309
70. Sijbrand J (1985) Properties of center manifolds. Trans Amer Math Soc 289:431–469
71. Van Strien SJ (1979) Center manifolds are not C^∞. Math Z 166:143–145
72. Vanderbauwhede A (1989) Center Manifolds, Normal Forms and Elementary Bifurcations. In: Dynamics Reported, vol 2. Springer, Berlin, pp 89–169
73. Vanderbauwhede A, Iooss G (1992) Center manifold theory in infinite dimensions. In: Dynamics Reported, vol 1. Springer, Berlin, pp 125–163
74. Wan YH (1977) On the uniqueness of invariant manifolds. J Diff Equ 24:268–273
75. Wang W, Duan J (2006) Invariant manifold reduction and bifurcation for stochastic partial differential equations. http://arXiv:math.DS/0607050
76. Wiggins S (1992) Introduction to Applied Nonlinear Dynamical Systems and Chaos. Springer, New York
77. Wiggins S (1994) Normally Hyperbolic Invariant Manifolds of Dynamical Systems. Springer, New York
78. Wulff C (2000) Translation from relative equilibria to relative periodic orbits. Doc Mat 5:227–274

Books and Reviews

Bates PW, Lu K, Zeng C (1998) Existence and persistence of invariant manifolds for semiflows in Banach spaces. Mem Amer Math Soc 135:129

Bates PW, Lu K, Zeng C (1999) Persistence of overflowing manifolds for semiflow. Comm Pure Appl Math 52(8):983–1046

Bates PW, Lu K, Zeng C (2000) Invariant filiations near normally hyperbolic invariant manifolds for semiflows. Trans Amer Math Soc 352:4641–4676

Babin AV, Vishik MI (1989) Attractors for Evolution Equations. Nauka. Moscow; English translation (1992). Elsevier Science, Amsterdam

Bylov VF, Vinograd RE, Grobman DM, Nemyskiy VV (1966) The Theory of Lyapunov Exponents. Nauka, Moscow (in Russian)

Chen X-Y, Hale J, Tan B (1997) Invariant foliations for C^1 semigroups in Banach spaces. J Diff Equ 139:283–318

Chepyzhov VV, Goritsky AYU, Vishik MI (2005) Integral manifolds and attractors with exponential rate for nonautonomous hyperbolic equations with dissipation. Russ J Math Phys 12(1): 17–39

Chicone C, Latushkin YU (1997) Center manifolds for infinite dimensional nonautonomous differential equations. J Diff Equ 141:356–399

Chow SN, Lu K (1988) Invariant manifolds for flows in Banach spaces. J Diff Equ 74:285–317

Chow SN, Lin XB, Lu K (1991) Smooth invariant foliations in infinite dimensional spaces. J Diff Equ 94:266–291

Chueshov I (1993) Global attractors for non-linear problems of mathematical physics. Uspekhi Mat Nauk 48(3):135–162; English translation in: Russ Math Surv 48:3

Chueshov I (1999) Introduction to the Theory of Infinite-Dimensional Dissipative Systems. Acta, Kharkov (in Russian); English translation (2002) http://www.emis.de/monographs/Chueshov/. Acta, Kharkov

Constantin P, Foias C, Nicolaenko B, Temam R (1989) Integral Manifolds and Inertial Manifolds for Dissipative Partial Differential Equations. Appl Math Sci, vol 70. Springer, New York

Gonçalves JB (1993) Invariant manifolds of a differentiable vector field. Port Math 50(4):497–505

Goritskii AYU, Chepyzhov VV (2005) Dichotomy property of solutions of quasilinear equations in problems on inertial manifolds. SB Math 196(4):485–511

Hassard B, Wan Y (1978) Bifurcation formulae derived from center manifold theory. J Math Anal Appl 63:297–312

Hsia C, Ma T, Wang S (2006) Attractor bifurcation of three-dimensional double-diffusive convection. http://arXiv:nlin.PS/0611024

Knobloch HW (1990) Construction of center manifolds. J Appl Math Phys (ZAMP) 70(7):215–233

Latushkin Y, Li Y, Stanislavova M (2004) The spectrum of a linearized 2D Euler operator. Stud Appl Math 112:259

Leen TK (1993) A coordinate independent center manifold reduction. Phys Lett A 174:89–93

Li Y (2005) Invariant manifolds and their zero-viscosity limits for Navier–Stokes equations. http://arXiv:math.AP/0505390

Osipenko G (1989) Examples of perturbations of invariant manifolds. Diff Equ 25:675–681

Osipenko G (1985, 1987, 1988) Perturbation of invariant manifolds I, II, III, IV. Diff Equ 21:406–412, 21:908–914, 23:556–561, 24:647–652

Podvigina OM (2006) The center manifold theorem for center eigenvalues with non-zero real parts. http://arXiv:physics/0601074

Sacker RJ, Sell GR (1974, 1976, 1978) Existence of dichotomies and invariant splitting for linear differential systems. J Diff Equ 15:429-458, 22:478–522, 27:106–137

Scarpellini B (1991) Center manifolds of infinite dimensional. Main results and applications. J Appl Math Phys (ZAMP) 43:1–32

Sell GR (1983) Vector fields on the vicinity of a compact invariant manifold. Lect Notes Math 1017:568–574

Swanson R (1983) The spectral characterization of normal hyperbolicity. Proc Am Math Soc 89(3):503–508

Temam R (1988) Infinite Dimensional Dynamical Systems in Mechanics and Physics. Springer, Berlin

Zhenquan Li, Roberts AJ (2000) A flexible error estimate for the application of center manifold theory. http://arXiv.org/abs/math.DS/0002138

Zhu H, Campbell SA, Wolkowicz (2002) Bifurcation analysis of a predator-prey system with nonmonotonic functional response. SIAM J Appl Math 63:636–682

Chaos and Complexity in Astrophysics, Introduction to

Steven N. Shore[1,2]
[1] Dipartmento di Fisica Enrico Fermi, Università di Pisa, Pisa, Italy
[2] Istituto Nazionale di Fiscia Nucleare, Sezione di Pisa, Pisa, Italy

We live in a hierarchically structured, complex Universe and the articles in this section demonstrate the breadth of vision required by astrophysicists addressing a vast spectrum of problems from dynamical systems to biological and galactic evolution to cosmology. Fundamentally, astronomical problems present a huge variety of scales of time, length, mass, and energy. For example, to understand the transport of energy in hot, opaque bodies without which we cannot understand the structure of stars requires kinetic theory of gases and atomic and molecular processes (the equation of state for stellar matter, and opacities), hydrodynamics, (especially turbulence for convective energy transport and circulation), magnetohydrodynamics and plasma physics (since stars generate magnetic fields by dynamo processes), nuclear and particle physics (for thermonuclear processing and energy loss), and even general relativity (for the late stages of evolution). But stars are formed and die within galaxies, self-gravitating ensembles of many billions of individual stars and cold clouds of millions of solar masses that are actually analogies of ecosystems. Thus the tools developed to address biological systems can be carried over into cosmic problems.

At the close of the 1920s, Arthur Eddington quipped that "someday, perhaps, we shall understand something as simple as a star". How wrong he was. Stars are self-gravitating nuclear engines and it is the mass of a gaseous body that determines whether it will or won't ignite its central thermonuclear energy source. Below a few percent of a solar mass, the stellar core never reaches the required temperature even if the densities are high enough to initiate nuclear reactions. Such a low mass object is called a brown dwarf if, during its initial contraction from an interstellar dark cloud, it can briefly burn deuterium. If even that source remains inaccessible, it is called a planet, regardless of whether it is bound in an orbit to another, more massive companion. Although various claims had been made throughout the 20th century for the presence of planetary systems, especially for the nearby very low mass object called Barnard's Star, the first direct detection of a planetary mass body orbiting another star came only in 1995 with the discovery of a massive Jupiter-like companion to the solar type star 51 Peg. Since then over 300 planets have been discovered orbiting stars other than the Sun, now including multiple systems and even several in other galaxies discovered by gravitational lensing. This sample is, however, currently still biased toward masses much greater than the Earth and systems that are still rather different from our solar system. But with the passage of years and the accumulation of velocity and photometric measurements, systems progressively more similar to ours have begun to emerge.

The first discoveries already, however, presented a paradox: It has generally been thought that the structure of our system is the product of thermal evaporation and restructuring the orbiting bodies by the early Sun, the evaporation of nearby gaseous masses and leaving behind only their rocky cores.

The detection of planetary mass bodies began, however, with the discovery of precisely such planets within a few hundred stellar radii of the central star. This has stimulated a renaissance in dynamical astronomy of few bodies and the re-discovery of effects previously thought to be of only minor importance in the present solar system. This too is reflected in the articles in this section. And because it is clear now that there must also be systems like our own, the study of extraterrestrial environments and the search for extrasolar Earths and the origin of life has emerged from the back rooms to take a central stage as a new sub-discipline: astrobiology (see ► Exobiology (theoretical), Complexity in, by Brandenburg and ► Exobiology and Complexity by Chaisson). It is particularly appropriate that this encyclopedia recognizes the immense promise of this field.

The interplanetary medium, the solar wind, is the topic of several articles. This is the supersonic mass outflow from the only star for which we have virtually all possible information. The wind, synonymous with the interplanetary medium, is the expanding coronal layer of the Sun. Heated by nonthemal mechanisms seated deep within he interior dynamo and convective envelope, magnetohydrodynamic waves and topological relaxation of the outer magnetic field (see ► Topological Magnetohydrodynamics and Astrophysics by Berger).

This is why several of the articles here deal with dynamo processes. This plasma is an ideal laboratory in which the Sun performs the experiments and we capture the results with in situ satellite measurements (see ► Self-

-Organization in Magnetohydrodynamic Turbulence by Veltri et al., and ► Space Plasmas, Dynamical Complexity in, by Chang).

Wave phenomena that are inaccessible under terrestrial laboratory conditions can be studied in detail with spacecraft, MHD turbulence can be observed on a vast range of length scales (from meters to astronomical units) and timescales (from milliseconds to days). We are in a new age of spacecraft measurement with SOHO and STEREO providing continuously monitored dynamical three dimensional imaging of the Sun.

And on the solar surface, in deep in its interior, we see the cyclic generation and decay of a dynamo whose superficial manifestations are the corona, magnetic loop structures of sizes comparable to the stellar radius, and dissipative processes that release high energy particles and locally heat the outer solar atmosphere. All of these topics show the complex interplay of many physical processes and are discussed in many of the articles in this section.

Cosmic rays were discovered at the start of the 20th century by Hess and others making balloon flights to high altitudes carrying devices for measuring atmospheric electricity. They discovered that the ionization increases with height, which led to the hypothesis of some form of extraterrestrial energetic particles impinging on the atmosphere. Further study showed these to be charged, high energy particles, protons, alpha particles, and electrons, whose source was a mystery. For some time, these were the only particles with which fundamental physics could be studied. Somehow, the universe is capable of producing relativistic ions and electrons that traverse the space between the galaxies, reaching energies well in excess of TeV. The Sun and solar type stars, during flares, are capable of accelerating electrons to GeV energies and beyond in milliseconds or shorter on scales of kilometers, but the process is still not understood. And on larger scales, the detection of TeV mission from supernova remnants, the supersonically ejected debris of stellar explosions, shows that these shocks can also accelerate particles on length scales of the size of the solar system throughout the life of the galaxy. Hyper-relativistic phenomena connected with stellar collapse and explosion, such as Gamma Ray Bursts, also demonstrate that complex acceleration processes happen on many different length and time scales in cosmic objects (see ► Acceleration Mechanisms by Melrose).

Historically, it was in problems arising within dynamical astronomy that chaos was first recognized – in the orbital dynamics of the solar system during the 19th century – and the richness of this field has extended far beyond the confines of the realm of the planets. It was the source of the analogies used by Boltzmann and Gibbs in the early stages of the development of statistical mechanics (see ► Astrophysics, Chaos and Complexity in, by Regev).

Beginning with the two body problem, for point masses the physical solution is almost trivial. The motion is conservative, the total energy is constant and, except for effect of general relativity for the most extreme systems, the problem can be solved in closed orbits knowing only the total energy and angular momentum. If one or both objects are extended, or if they are internally complicated, the problem ceases to be tractable in closed form. The orbits evolve in time, either periodically or secularly depending on whether or not there are internal dissipative processes in the constituent bodies. This feature, which complicates the dynamics of binary star systems and even planetary moons requires a sophisticated physical treatment of microphysical processes from gas laws to fluid mechanics.

Planetary ring systems are even more complex, showing a wealth of dynamical structuring that, from particle trajectories imaged in natural settings, produce images like those resulting from numerical models of nonlinear systems. At the next level, the apparently simple problem of the lunar orbit becomes almost impossible once tidal interactions and the effect of the Sun are included. This was the principal discovery of Poincare, Lyapunov, and the dynamicists of the late 19th century, that the perturbing influences on the motion of so small a body a the Moon, whose mass is negligible with respect to the two principals, has a complex temporal spectrum of resonances that produce, even without dissipation, extreme sensitivity of the predictions to the accuracy of the initial conditions for any dynamical calculation (see ► Orbital Dynamics, Chaos in, by Hadjidemetriou).

The extension of this complexity to N-body calculations (see ► Stellar Dynamics, *N*-body Methods for, by Makino) and the evolution of far more populous systems – clusters and galaxies (see ► Astrophysics: Dynamical Systems by Contopoulos) – are addressed in articles in this section.

Moving to the largest scale of all, the Universe as a whole, we reach the limits of current astronomical observations. In the last decade, gravitational wave observatories have opened a new type of astronomical observation, one connected with neither particles nor light and at the very limits of detection. During the formation of black holes and neutron stars, at the moment of collapse of tars either by intrinsic or extrinsic triggering, gravitational waves certainly carry away a significant portion of the binding energy of the initial system. Such emission is required by the theory of collapse and the detection of the secular evolution of binary pulsar orbits assures the

physical correctness of the fundamentally relativistic phenomenon. However, the Universe is a very big place and in all galaxies, whatever their evolutionary history and distance, there must be sources for such waves.

These produce an incoherent signal, unlike the strong, identifiable signature of a single collapse event, that should form a stochastic background from the integration of all such sources. It is also clear that at some stage in the earliest stages of the cosmic expansion there were fluctuations in the matter and fields that were strongly coupled with the spacetime structure. These were not dissipatively damped as would happen in a normal fluid.

Instead, their spectrum was systematically altered by the cosmic expansion to the point of visibility in the present universe. How this might be found and what it tells us about the first moments of the Big Bang and the subsequent evolution of cosmic sources during the Hubble time is also treated in this section (see ► Cosmic Gravitational Background, Stochastic by Ungarelli and ► Cosmic Strings by Achúcarro and Martins).

Finally, because of this huge range of timescales and multiplicity of processes involved in even some of the apparently simplest phenomena, astronomical observations present particularly challenging problems of data analysis (see ► Astronomical Time Series, Complexity in, by Scargle).

From dynamical systems to stochastic fluctuations of spacetime, the problem of finding periods and spectra in complex, not spatially resolved and partially sampled data takes us to the limit of current statistical methodology.

In fact, this is a fundamental feature of all astrophysical research, to stretch all of our human capacities, intellectual and technological, in an ever-widening panorama.

Chaos and Ergodic Theory

Jérôme Buzzi
C.N.R.S. and Université Paris-Sud, Orsay, France

Article Outline

Glossary
Definition of the Subject
Introduction
Picking an Invariant Probability Measure
Tractable Chaotic Dynamics
Statistical Properties
Orbit Complexity
Stability
Untreated Topics
Future Directions
Acknowledgments
Bibliography

Glossary

For simplicity, definitions are given for a continuous or smooth self-map or diffeomorphism T of a compact manifold M.

Entropy, measure-theoretic (or: metric entropy) For an ergodic invariant probability measure μ, it is the smallest exponential growth rates of the number of orbit segments of given length, with respect to that length, after restriction to a set of positive measure. We denote it by $h(T,\mu)$. See ► Entropy in Ergodic Theory and Subsect. "Local Complexity" below.

Entropy, topological It is the exponential growth rates of the number of orbit segments of given length, with respect to that length. We denote it by $h_{\text{top}}(f)$. See ► Entropy in Ergodic Theory and Subsect. "Local Complexity" below.

Ergodicity A measure is ergodic with respect to a map T if given any measurable subset S which is invariant, i. e., such that $T^{-1}S = S$, either S or its complement has zero measure.

Hyperbolicity A measure is hyperbolic in the sense of Pesin if at almost every point no Lyapunov exponent is zero. See ► Smooth Ergodic Theory.

Kolmogorov typicality A property is *typical in the sense of Kolmogorov* for a topological space $\mathcal{F}$ of parametrized families $f = (f_t)_{t\in U}$, U being an open subset of $\mathbb{R}^d$ for some $d \geq 1$, if it holds for f_t for Lebesgue almost every t and topologically generic $f \in \mathcal{F}$.

Lyapunov exponents
The Lyapunov exponents (► Smooth Ergodic Theory) are the limits, when they exist, $\lim_{n\to\infty} \frac{1}{n} \log \|(T^n)'(x).v\|$ where $x \in M$ and v is a non zero tangent vector to M at x. The Lyapunov exponents of an ergodic measure is the set of Lyapunov exponents obtained at almost every point with respect to that measure for all non-zero tangent vectors.

Markov shift (topological, countable state) It is the set of all infinite or bi-infinite paths on some countable directed graph endowed with the left-shift, which just translates these sequences.

Maximum entropy measure It is a measure μ which maximizes the measured entropy and, by the variational principle, realized the topological entropy.

Physical measure It is a measure μ whose basin, $\{x \in M\colon \forall \phi\colon M \to \mathbb{R}$ continuous $\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1}$

$\phi(f^k x) = \int \phi \, d\mu\}$ has nonzero volume.

Prevalence A property is *prevalent* in some complete metric, separable vector space X if it holds outside of a set N such that, for some Borel probability measure μ on X: $\mu(A + v) = 0$ for all $v \in X$. See [76,141,239].

Sensitivity on initial conditions T has *sensitivity to initial conditions* on $X' \subset X$ if there exists a constant $\rho > 0$ such that for every $x \in X'$, there exists $y \in X$, arbitrarily close to x, and $n \geq 0$ such that $d(T^n y, T^n x) > \rho$.

Sinai–Ruelle–Bowen measures It is an invariant probability measure which is absolutely continuous along the unstable foliation (defined using the unstable manifolds of almost every $x \in M$, which are the sets $W^u(x)$, of points y such that $\lim_{n\to\infty} \frac{1}{n} \log d(T^{-n} y, T^{-n} x) < 0$).

Statistical stability T is statistically stable if the physical measures of nearby deterministic systems are arbitrarily close to the convex hull of the physical measures of T.

Stochastic stability T is stochastically stable if the invariant measures of the Markov chains obtained from T by adding a suitable, smooth noise with size $\epsilon \to 0$ are arbitrarily close to the convex hull of the physical measures of T.

Structural stability T is structurally stable if any S close enough to T is topologically the same as T: there exists a homeomorphism $h: M \to M$ such that $h \circ T = S \circ h$ (orbits are sent to orbits).

Subshift of finite type It is a closed subset Σ_F of $\Sigma = \mathcal{A}^{\mathbb{Z}}$ or $\Sigma = \mathcal{A}^{\mathbb{N}}$ where $\mathcal{A}$ is a finite set satisfying: $\Sigma_F = \{x \in \Sigma : \forall k < \ell : x_k x_{k+1} \ldots x_\ell \notin F\}$ for some finite set F.

Topological genericity Let X be a Baire space, e. g., a complete metric space. A property is *(topologically) generic* in a space X (or holds for the (topologically) generic element of X) if it holds on a nonmeager set (or set of second Baire category), i. e., on a dense G_δ subset.

Definition of the Subject

Chaotic dynamical systems are those which present unpredictable and/or complex behaviors. The existence and importance of such systems has been known at least since Hadamard [126] and Poincaré [208], however it became well-known only in the sixties. We refer to [36,128,226,236] and [80,107,120,125,192,213] for the relevance of such dynamics in other fields, mathematical or not (see also ► Ergodic Theory: Interactions with Combinatorics and Number Theory, ► Ergodic Theory: Fractal Geometry). The numerical simulations of chaotic dynamics can be difficult to interpret and to plan, even misleading, and the tools and ideas of mathematical dynamical system theory are indispensable. Arguably the most powerful set of such tools is ergodic theory which provides a statistical description of the dynamics by attaching relevant probability measures. In opposition to single orbits, the statistical properties of chaotic systems often have good stability properties. In many cases, this allows an understanding of the complexity of the dynamical system and even precise and quantitative statistical predictions of its behavior. In fact, chaotic behavior of single orbits often yields global stability properties.

Introduction

The word *chaos*, from the ancient Greek $\chi\alpha o\sigma$, "shapeless void" [131] and "raw confused mass" [199], has been used [*$\chi\alpha o\sigma$ also inspired Van Helmont to create the word "gas" in the seventeenth century and this other thread leads to the molecular chaos of Boltzmann in the nineteenth century and therefore to ergodic theory itself*.] since a celebrated paper of Li and Yorke [169] to describe evolutions which however deterministic and defined by rather simple rules, exhibit unpredictable or complex behavior.

Attempts at Definition

We note that, like many ideas [237], this is not captured by a single mathematical definition, despite several attempts (see, e. g., [39,112,158,225] for some discussions as well as the monographs on chaotic dynamics [15,16,58,60,78,104,121,127,215,250,261]). Let us give some of the most well-known definitions, which have been given mostly from the topological point of view, i. e., in the setting of a self-map $T: X \to X$ on a compact metric space whose distance is denoted by d:

T has *sensitivity to initial conditions* on $X' \subset X$ if there exists a constant $\rho > 0$ such that for every $x \in X'$, there exists $y \in X$, arbitrarily close to x with a finite separating time:

$$\exists n \geq 0 \quad \text{such that} \quad d(T^n y, T^n x) > \rho \,.$$

In other words, any uncertainty on the exact value of the initial condition x makes $T^n(x)$ completely unknown for n large enough. If X is a manifold, then *sensitivity to initial conditions in the sense of Guckenheimer* [120] means that the previous phenomenon occurs for a set X' with nonzero volume.

T is *chaotic in the sense of Devaney* [94] if it admits a dense orbit and if the periodic points are dense in X. It implies sensitivity to initial conditions on X.

T is *chaotic in the sense of Li and Yorke* [169] if there exists an uncountable subset $X' \subset X$ of points, such that, for all $x \neq y \in X'$,

$$\liminf_{n\to\infty} d(T^n x, T^n y) = 0 \quad \text{and}$$
$$\limsup_{n\to\infty} d(T^n x, T^n y) > 0\,.$$

T has *generic chaos* in the sense of Lasota [206] if the set

$$\{(x,y) \in X \times X : \liminf_{n\to\infty} d(T^n x, T^n y) = 0 < \limsup_{n\to\infty} d(T^n x, T^n y)\}$$

is topologically generic (*see glossary.*) in $X \times X$.

Topological chaos is also sometimes characterized by *nonzero topological entropy* (▶ Entropy in Ergodic Theory): there exist exponentially many orbit segments of a given length. This implies chaos in the sense of Li and Yorke by [39].

As we shall see ergodic theory describes a number of chaotic properties, many of them implying some or all of the above topological ones. The main such property for a smooth dynamical system, say a $C^{1+\alpha}$-diffeomorphism of a compact manifold, is the existence of an invariant probability measure which is:

1. Ergodic (cannot be split) and aperiodic (not carried by a periodic orbit);
2. Hyperbolic (nearby orbits converge or diverge at a definite exponential rate);
3. Sinai–Ruelle–Bowen (as smooth as it is possible).

(For precise definitions we refer to ▶ Smooth Ergodic Theory or to the discussions below.) In particular such a situation implies nonzero entropy and sensitivity to initial condition of a set of nonzero Lebesgue measure (i. e., positive volume).

Before starting our survey in earnest, we shall describe an elementary and classical example, the full tent map, on which the basic phenomena can be analyzed in a very elementary way. Then, in Sect. "Picking an Invariant Probability Measure", we shall give some motivations for introducing probability theory in the description of chaotic but deterministic systems, in particular the unpredictability of their individual orbits. We define two of the most relevant classes of invariant measures: the physical measures and those maximizing entropy. It is unknown in which generality these measures exist and can be analyzed but we describe in Sect. "Tractable Chaotic Dynamics" the major classes of dynamics for which this has been done. In Sect. "Statistical Properties" we describe some of the finer statistical properties that have been obtained for such good chaotic systems: sums of observables along orbits are statistically undistinguishable from sums of independent and identically distributed random variables. Sect. "Orbit Complexity" is devoted to the other side of chaos: the complexity of these dynamics and how, again, this complexity can be analyzed, and sometimes classified, using ergodic theory. Sect "Stability" describes perhaps the most striking aspect of chaotic dynamics: the unstability of individual orbit is linked to various forms of stability of the global dynamics.

Finally we conclude by mentioning some of the most important topics that we could not address and we list some possible future directions.

Caveat. The subject-matter of this article is somewhat fuzzy and we have taken advantage of this to steer our path towards some of our favorite theorems and to avoid the parts we know less (some of which are listed below). We make no pretense at exhaustivity neither in the topics nor in the selected results and we hope that our colleagues will excuse our shortcomings.

Remark 1 In this article we only consider *compact, smooth and finite-dimensional* dynamical systems *in discrete time*, i. e., defined by self-maps. In particular, we have omitted the natural and important variants applying to flows, e. g., evolutions defined by ordinary differential equations but we refer to the textbooks (see, e. g.,[15,128,148]) for these.

Elementary Chaos: A Simple Example

We start with a toy model: the full tent map T of Fig. 1. Observe that for any point $x \in [0,1]$, $T^{-n}(x) = \{(\sigma(k,n)x + k)\cdot 2^{-n} : k = 0,1,\ldots,2^n-1\}$, where $\sigma(k,n) = \pm 1$. Hence $\bigcup_{n\geq 0} T^{-n}(x)$ is dense in $[0,1]$. It easily follows that T exhibits *sensitive dependence to initial conditions*. Even worse in this example, the qualitative asymptotic behavior can be completely changed by this arbitrarily small perturbation: x may have a dense orbit whereas y is eventually mapped to a fixed point! This is Devaney chaos [94].

This kind of unstability was first discovered by J. Hadamard [126] in his study of the geodesic flow (i. e., the frictionless movement of a point mass constrained to remain on a surface). At that time, such an unpredictability was considered a purely mathematical pathology, necessarily devoid of any physical meaning [*Duhem qualified Hadamard's result as "an example of a mathematical deduction which can never be used by physics" (see pp. 206–211 in* [103]*)!*].

Returning to out tent map, we can be more quantitative. At any point $x \in [0,1]$ whose orbit never visits 1/2, the *Lyapunov exponent* $\lim_{n\to\infty} \frac{1}{n} \log |(T^n)'(x)|$

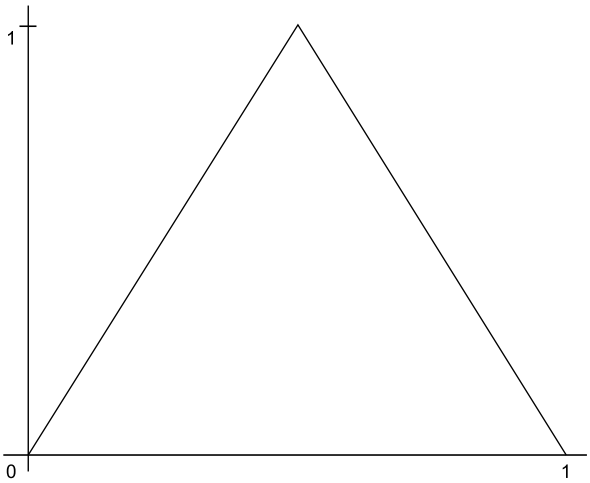

Chaos and Ergodic Theory, Figure 1
The graph of the full tent map $f(x) = 1 - |1 - 2x|$ over $[0, 1]$

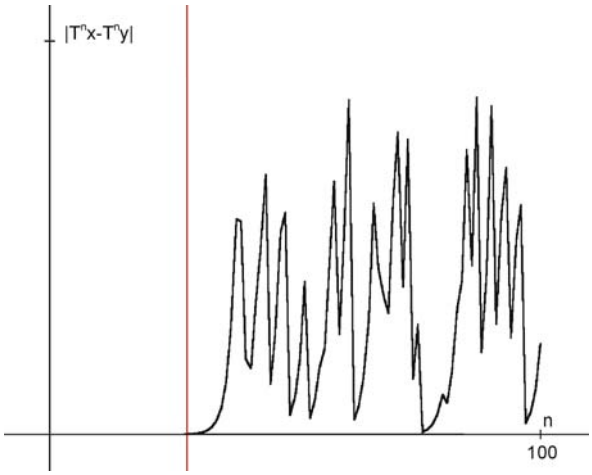

Chaos and Ergodic Theory, Figure 2
$|T^n(x) - T^n(y)|$ for $T(x) = 1 - |1 - 2x|$, $|x - y| = 10^{-12}$ and $0 \le n \le 100$. The vertical red line is at $n = 28$ and shows when $|T^n x - T^n y| \ge 0.5 \cdot 10^{-4}$ for the first time

is $\log 2$. (*See the glossary*.) Such a positive Lyapunov exponent corresponds to infinitesimally close orbits getting separated exponentially fast. This can be observed in Fig. 2. Note how this exponential speed creates a rather sharp transition.

It follows in particular that experimental or numerical errors can grow very quickly to size 1, [*For simple precision arithmetic the uncertainty is* 10^{-16} *which grows to size 1 in 38 iterations of* T.] i. e., the approximate orbit may contain after a while *no information about the true orbit*. This casts a doubt on the reliability of simulations. Indeed, a simulation of T on most computers will suggest that all orbits quickly converge to 0, which is completely false [*Such a collapse to 0 does really occurs but only for a countable subset of initial conditions in* $[0, 1]$ *whereas the points with dense orbit make a subset of* $[0, 1]$ *with full Lebesgue measure (see below). This artefact comes from the way numbers are represented – and approximated – on the computer: multiplication by even integers tends to "simplify" binary representations. Thus the computations involved in drawing Fig. 2 cannot be performed too naively.*]. Though somewhat atypical in its dramatic character, this failure illustrates the *unpredictability* and *unstability* of individual orbits in chaotic systems.

Does this mean that all quantitative predictions about orbits of T are to be forfeited? Not at all, *if we are ready to change our point of view* and look beyond a single orbit. This can be seen easily in this case. Let us start with such a global analysis from the topological point of view. Associate to $x \in [0, 1]$, a sequence $i(x) = i = i_0 i_1 i_2 \ldots$ of 0s and 1s according to $i_k = 0$ if $T^k x \le 1/2$, $i_k = 1$ otherwise. One can check that [*Up to a countable set of exceptions.*] $\{i(x) \colon x \in [0, 1]\}$ is the set $\Sigma_2 := \{0, 1\}^{\mathbb{N}}$ of *all* infinite sequences of 0s and 1s and that at most one $x \in [0, 1]$ can realize a given sequence as $i(x)$.

Notice how the transformation f becomes trivial in this representation:

$$i(f(x)) = i_1 i_2 i_3 \ldots \quad \text{if } i(x) = i_0 i_1 i_2 i_3 \ldots$$

Thus f is represented by the simple and universal "left-shift" on sequences, which is denoted by σ. This representation of a rather general dynamical system by the left-shift on a space of sequences is called *symbolic dynamics* (▶ Symbolic Dynamics), [171].

This can be a very powerful tool. Observe for instance how here it makes obvious that we have complete *combinatorial freedom* over the orbits of T: one can easily build orbits with various asymptotic behaviors: if a sequence of Σ_2 contains all the finite sequences of 0s and 1s, then the corresponding point has a dense orbit; if the sequence is periodic, then the corresponding point is itself periodic, to give two examples of the *richness* of the dynamics.

More quantitatively, the number of distinct subsequences of length n appearing in sequences $i(x)$, $x \in [0, 1]$, is 2^n. It follows that the *topological entropy* (▶ Entropy in Ergodic Theory) of T is $h_{\text{top}}(T) = \log 2$. [*For the coincidence of the entropy and the Lyapunov exponent see below.*] The positivity of the topological entropy can be considered as the signature of the complexity of the dynamics and considered as the definition, or at least the stamp, of a topologically chaotic dynamics.

Let us move on to a probabilistic point of view. Pick $x \in [0, 1]$ randomly according to, say, the uniform law in $[0, 1]$. It is then routine to check that $i(x)$ follows

the (1/2, 1/2)-Bernoulli law: the probability that, for any given k, $i_k(x) = 0$ is 1/2 and the i_ks are independent. Thus $i(x)$, seen as a sequence of random 0 and 1 when x is subject to the uniform law on $[0, 1]$, is *statistically undistinguishable from coin tossing!* This important remark leads to quantitative predictions. For instance, the *strong law of large numbers* implies that, for Lebesgue-almost every $x \in [0, 1]$ (i. e., for all $x \in [0, 1]$ except in a set of Lebesgue measure zero), the fraction of the time spent in any dyadic interval $I = [k \cdot 2^{-N}, \ell \cdot 2^{-N}] \subset [0, 1]$, $k, \ell, N \in \mathbb{N}$, by the orbit of x,

$$\lim_{n\to\infty} \frac{1}{n} \#\{0 \le k < n \colon T^k x \in I\} \tag{1}$$

exists and is equal to the length, 2^{-N}, of that interval. [*Eq.* (1) *in fact holds for any interval I. This implies that the orbit of almost every* $x \in [0, 1]$ *visits all subintervals of* $[0, 1]$, *i. e., the orbit is dense: in complete contradiction with the above mentioned numerical simulation!*] More generally, we shall see that, if $\phi \colon [0, 1] \to \mathbb{R}$ is any continuous function, then, for Lebesgue almost every-x,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \phi(T^k x) \quad \text{exists and is equal to} \quad \int \phi(x)\, \mathrm{d}x\,. \tag{2}$$

Using strong *mixing properties* (▶ Ergodicity and Mixing Properties) of the Lebesgue measure under T, one can prove further properties, e. g., sensitivity on initial conditions in the sense of Guckenheimer [*The Lebesgue measure is* weak-mixing*: Lebesgue-almost all couples of points* $(x, y) \in [0, 1]^2$ *get separated. Note that it is not true of every couple off the diagonal: Counter-examples can be found among couples* (x, y) *with* $T^n x = T^n y$ *arbitrarily close to 1.*] and study the fluctuations of the averages $\frac{1}{n} \sum_{k=0}^{n-1} \phi(Tx)$ by the way of limit theorems.

The above analysis relied on the very special structure of T but, as we shall explain, the ergodic theory of differentiable dynamical systems shows that all of the above (and much more) holds in some form for rather general classes of chaotic systems. The different chaotic properties are independent in general (e. g., one may have topological chaos whereas the asymptotic behavior of almost all orbits is periodic) and the proofs can become much more difficult. We shall nonetheless be rewarded for our efforts by the discovery of unexpected links between chaos and stability, complexity and simplicity as we shall see.

Picking an Invariant Probability Measure

One could think that dynamical systems, such as those defined by self-maps of manifolds, being *completely deterministic*, have nothing to do with probability theory. There are in fact several motivations for introducing various invariant probability measures.

Statistical Descriptions

An abstract goal might be to enrich the structure: a smooth self-map is a particular case of a Borel self-map, hence one can canonically attach to this map its set of all invariant Borel probability measures [*From now on all measures will be Borel probability measures except if it is explicitly stated otherwise.*], or just the set of *ergodic* [*A measure is* ergodic *if all measurable invariant subsets have measure 0 or 1. Note that arbitrary invariant measures are averages of ergodic ones, so many questions about invariant measures can be reduced to ergodic ones.*] ones. By the Krylov–Bogoliubov theorem (see, e. g., [148]), this set is non-empty for any continuous self-map of a compact space.

By the following fundamental theorem ▶ Ergodic Theorems, each such measure is the statistical description of some orbit:

Birkhoff Pointwise Ergodic Theorem *Let* $(X, \mathcal{F}, \mu)$ *be a space with a* σ*-field and a probability measure. Let* $f \colon X \to X$ *be a measure-preserving map, i. e.,* $f^{-1}(\mathcal{F}) \subset \mathcal{F}$ *and* $\mu \circ f^{-1} = \mu$. *Assume ergodicity of* μ *(See the glossary.) and (absolute) integrability of* $\phi \colon X \to \mathbb{R}$ *with respect to* μ. *Then for* μ*-almost every* $x \in X$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \phi(f^k x) \quad \textit{exists and is} \quad \int \phi\, \mathrm{d}\mu\,.$$

This theorem can be interpreted as saying that "time averages" coincide almost surely with "ensemble averages" (or "phase space average"), i. e., that Boltzmann's Ergodic Hypothesis of statistical mechanics [110] holds for dynamical systems that cannot be split in a measurable and non trivial way. [*This indecomposability is however often difficult to establish. For instance, for the hard ball model of a gas it is known only under some generic assumption (see* [238] *and the references therein).*] We refer to [161] for background.

Remark 2 One should observe that the existence of the above limit is not at all obvious. In fact it often fails from other points of view. One can show that for the full tent map $T(x) = 1 - |1 - 2x|$ analyzed above and many functions ϕ, the set of points for which it fails is large both *from the topological point of view* (it contains a dense G_δ set) and *from the dimension point of view* (it has Hausdorff dimension 1 [28]). This is an important point: the introduction of

invariant measures allows one to avoid some of the wilder pathologies.

To illustrate this let us consider the full tent map $T(x) = 1 - |1 - 2x|$ again and the two ergodic invariant measures: δ_0 (the Dirac measure concentrated at the fixed point 0) and the Lebesgue measure $\mathrm{d}x$. In the first case, we obtain a complex proof of the obvious fact that the time average at $x = 0$ (some set of full measure!) and the ensemble average with respect to δ_0 are both equal to $f(0)$. In the second case, we obtain a very general proof of the above Eq. (2).

Another type of example is provided by the contracting map $S\colon [0,1] \to [0,1]$, $S(x) = x/2$. S has a unique invariant probability measure, δ_0. For Birkhoff theorem the situation is the same as that of T and δ_0: it asserts only that the orbit of 0 is described by δ_0.

One can understand Birkhoff theorem as a (first and rather weak) *stability result*: the time averages are independent of the initial condition, *almost surely with respect to* μ.

Physical Measures

In the above silly example S, much more is true than the conclusion of Birkhoff Theorem: *all* points of $[0,1]$ are described by δ_0. This leads to the definition of the *basin* of a probability measure μ for a self-map f of a space M:

$$\mathcal{B}(\mu) := \Big\{ x \in M\colon \forall \phi\colon M \to \mathbb{R} \text{ continuous } \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \phi(f^k x) = \int \phi \, \mathrm{d}\mu \Big\} .$$

If M is a manifold, then there is a notion of volume and one can make the following definition. A *physical measure* is a probability measure whose basin has nonzero volume in M. Say that a dynamical system $f\colon M \to M$ on a manifold has a *finite statistical description* if there exists finitely many invariant probability measures $\mu_1, \ldots, \mu_n$ the union of whose basins is the whole of M, up to a set of zero Lebesgue measure.

Physical measures are among the main subject of interest as they are expected to be exactly those that are "experimentally visible". Indeed, if $x_0 \in \mathcal{B}(\mu)$ and $\epsilon_0 > 0$ is small enough, then, by Lebesgue density theorem, a point x picked according to, say, the uniform law in the ball $B(x_0, \epsilon_0)$ of center x_0 and radius ϵ_0, will be in $\mathcal{B}(\mu)$ with probability almost 1 and therefore its ergodic averages will be described by μ. Hence "experiments" can be expected to follow the physical measures and this is what is numerically observed in most of the situations (see however the caveat in the discussion of the full tent map).

The existence of a finite statistical description (or even of a physical measure) is, as we shall see, not automatic nor routine to prove. Attracting periodic points as in the above silly example provide a first type of physical measures. Birkhoff ergodic theorem asserts that absolutely continuous ergodic invariant measures, usually obtained from some expansion property, give another class of physical measures. These contracting and expanding types can be combined in the class of *Sinai–Ruelle–Bowen measures* [166] which are the invariant measures absolutely continuous "along expanding directions" (see for the precise but technical definition ▶ Smooth Ergodic Theory). Any ergodic Sinai–Ruelle–Bowen measure which is ergodic and without zero Lyapunov exponent [*That is, the set of points* $x \in M$ *such that* $\lim_{n\to\infty} \frac{1}{n} \log \|(f^n)'(x).v\| = 0$ *for some* $v \in T_x M$ *has zero measure.*] is a physical measure. Conversely, "most" physical measures [*For counter-examples see* [136].] are of this type [243,247].

Measures of Maximum Entropy

For all parameters $t \in [3.96, 4]$, the quadratic maps $Q_t(x) = tx(1-x)$, $Q_t\colon [0,1] \to [0,1]$, have nonzero topological entropy [91] and exponentially many periodic points [134]:

$$\lim_{n\to\infty} \frac{\#\{x \in [0,1]\colon Q_t^n(x) = x\}}{\mathrm{e}^{n h_{\mathrm{top}}(Q_t)}} = 1 .$$

On the other hand, by a deep theorem [117,178] there is an open and dense subset of $t \in [0,4]$, such that Q_t has a unique physical measure *concentrated on a periodic orbit*! Thus the physical measures can completely miss the topological complexity (and in particular the distribution of the periodic points). Hence one must look at other measures to get a statistical description of the complexity of such Q_t. Such a description is often provided by *measures of maximum entropy* μ_M whose measured entropy [*The usual phrases are "measure-theoretic entropy", "metric entropy".*] (▶ Entropy in Ergodic Theory) satisfies:

$$h(f, \mu_M) = \sup_{\mu \in M(f)} h(f, \mu) =^1 h_{\mathrm{top}}(f) .$$

$M(f)$ is the set of all invariant measures. [*One can restrict this to the ergodic invariant measures without changing the value of the supremum* (▶ Entropy in Ergodic Theory).] Equality 1 above is the *variational principle*: it holds for all continuous self-maps of compact metric spaces. One can say that the ergodic complexity (the complexity of f as seen by its *invariant measures*) captures the full topological complexity (defined by counting *all* orbits).

Remark 3 The variational principle implies the existence of "complicated invariant measures" as soon as the topological entropy is nonzero (see [47] for a setting in which this is of interest).

Maximum entropy measures do not always exist. However, if f is C^∞ smooth, then maximum entropy measures exist by a theorem of Newhouse [195] and they indeed describe the topological complexity in the following sense. Consider the probability measures:

$$\mu_{n,\epsilon} := \frac{1}{n}\sum_{k=0}^{n-1}\sum_{x\in E(n,\epsilon)} \delta_{f^k x}$$

where $E(n,\epsilon)$ is an arbitrary (ϵ,n)-separated subset [*See* ► Entropy in Ergodic Theory: $\forall x,y \in E(n,\epsilon)\, x \neq y \implies \exists 0 \leq k < n\ d(T^k x, T^k y) \geq \epsilon$.] of M with maximum cardinality. Then accumulation points for the weak star topology on the space of probability measures on M of $\mu_{n,\epsilon}$ when $n \to \infty$ and then $\epsilon \to 0$ are maximum entropy measures [185].

Let us quote two important additional properties, discovered by Margulis [179], that often hold for the maximum entropy measures:

- The *equidistribution of periodic points* with respect to some maximum entropy measure μ_M:

$$\mu_M = \lim_{n\to\infty} \frac{1}{\#\{x \in X : x = f^n x\}} \sum_{x=f^n x} \delta_x\,.$$

- The *holonomy invariance* which can be loosely interpreted by saying that the past and the future are independent conditionally on the present.

Other Points of View

Many other invariant measures are of interest in various contexts and we have made no attempt at completeness: for instance, invariant measures maximizing dimension [111,204], or pressure in the sense of the thermodynamical formalism [148,222], or some energy [9,81,146], or quasi-physical measures describing the dynamics around saddle-type invariant sets [104] or in systems with holes [75].

Tractable Chaotic Dynamics

The Palis Conjecture

There is, at this point, no general theory allowing the analysis of all dynamical systems or even of most of them despite many recent and exciting developments in the theory of generic C^1-diffeomorphisms [51,84]. In particular, the question of the generality in which physical measures exist remains open. One would like generic systems to have a *finite statistical description* (see Subsect. "Physical Measures"). This fails in some examples but these look exceptional and the following question is asked by Palis [200]:

> *Is it true that any dynamical system defined by a C^r-diffeomorphism on a compact manifold can be transformed by an arbitrarily small C^r-perturbation to another dynamical system having a finite statistical description?*

This is completely open though widely believed [*Observe, however, that such a statement is false for conservative diffeomorphisms with high order smoothness as KAM theory implies stable existence of invariant tori foliating a subset of positive volume.*]. Note that such a good description is not possible for all systems (see, e. g., [136,194]). Note that one would really like to ask about unperturbed "typical" [*The choice of the notion of typicality is a delicate issue. The Newhouse phenomenon shows that among C^2-diffeomorphisms of multidimensional compact manifolds, one cannot use topological genericity and get a positive answer. Popular notions are* prevalence *and* Kolmogorov genericity *– see the glossary.*] dynamical systems in a suitable sense, but of course this is even harder.

One is therefore led to make simplifying assumptions: typically of small dimension, uniform expansion/contraction or geometry.

Uniformly Expanding/Hyperbolic Systems

The most easily analyzed systems are those with uniform expansion and/or contraction, namely the uniformly expanding maps and uniformly hyperbolic diffeomorphisms, see ► Smooth Ergodic Theory. [*We require uniform hyperbolicity on the so-called chain recurrent set. This is equivalent to the usual Axiom-A and no-cycle condition.*] An important class of example is obtained as follows. Consider $A: \mathbb{R}^d \to \mathbb{R}^d$, a linear map preserving $\mathbb{Z}^d$ (i. e., A is a matrix with integer coefficients in the canonical basis) so that it defines a map $\bar{A}: \mathbb{T}^d \to \mathbb{T}^d$ on the torus. If there is a constant $\Lambda > 1$ such that for all $v \in \mathbb{R}^d$, $\|A.v\| \geq \Lambda\|v\|$ then $\bar{A}$ is a *uniformly expanding map*. If A has determinant ± 1 and no eigenvalue on the unit circle, then $\bar{A}$ is a *uniformly hyperbolic diffeomorphism* (► Smooth Ergodic Theory) (see also [60,148,215,233]). Moreover all C^1-perturbations of the previous examples are again uniformly expanding or uniformly hyperbolic. [*One can define uniform hyperbolicity for flows and an important class of examples is provided by the geodesic flow on compact*

manifolds with negative sectional curvature [148].] These uniform systems are sometimes called "strongly chaotic".

Remark 4 Mañé Stability Theorem (see below) shows that uniform hyperbolicity is a very natural notion. One can also understand on a more technical level uniform hyperbolicity as what is needed to apply an implicit function theorem in some functional space (see, e. g., [233]).

The existence of a finite statistical description for such systems has been proved since the 1970s by Bowen, Ruelle and Sinai [54,218,235] (the expanding case is much simpler [162]).

Theorem 1 *Let $f: M \to M$ be a $C^{1+\alpha}$ map of a compact manifold. Assume f to be (i) a uniformly expanding map on M or (ii) a uniformly hyperbolic diffeomorphism.*

- *f admits a finite statistical description by ergodic and hyperbolic Sinai–Ruelle–Bowen measures (absolutely continuous in case (i)).*
- *f has finitely many ergodic maximum entropy measures, each of which makes f isomorphic to a finite state Markov chain. The periodic points are uniformly distributed according to some canonical average of these ergodic maximum entropy measures.*
- *f is topologically conjugate [Up to some negligible subset.] to a subshift of finite type (See the glossary.)*

The construction of absolutely continuous invariant measures for a uniformly expanding map f can be done in a rather direct way by considering the pushed forward measures $\frac{1}{n}\sum_{k=0}^{n-1} f_*^k$ Leb and taking weak star limits while preventing the appearance of singularities, by, e. g., bounding some Hölder norm of the density using expansion and distortion of f.

The classical approach to the uniformly hyperbolic dynamics [52,222,233] is through symbolic dynamics and coding. Under the above hypothesis one can build a finite partition of M which is tailored to the dynamics (a *Markov partition*) so that the corresponding symbolic dynamics has a very simple structure: it is a full shift $\{1,\dots,d\}^{\mathbb{Z}}$, like in the example of the full tent map, or a subshifts of finite type. The above problems can then be solved using the thermodynamical formalism inspired from the statistical mechanics of one-dimensional ferromagnets [217]: ergodic properties are obtained through the spectral properties of a suitable transfer operator acting on some space of regular functions, e. g., the Hölder-continuous functions defined over the symbolic dynamics with respect to the distance $d(x,y) := \sum_{n\in\mathbb{Z}} 2^{-n} 1_{x_n \neq y_n}$ where $1_{s\neq t}$ is 1 if $s \neq t$, 0 otherwise.

A recent development [24,43,116] has been to find suitable Banach spaces to apply the transfer operator technique directly in the smooth setting, which not only avoids the complication of coding (or rather replace them with functional analytic preliminaries) but allows the use of the smoothness beyond Hölder-continuity which is important for finer ergodic properties.

Uniform expansion or hyperbolicity can easily be obstructed in a given system: a "bad" point (a critical point or a periodic point with multiplier with an eigenvalue on the unit circle) is enough. This leads to the study of other systems and has motivated many works devoted to relaxing the uniform hyperbolicity assumptions [51].

Pesin Theory

The most general such approach is *Pesin theory*. Let f be a $C^{1+\alpha}$-diffeomorphism [*It is an important open problem to determine to which extent Pesin theory could be generalized to the C^1 setting.*] f with an ergodic invariant measure μ. By Oseledets Theorem ▶ Smooth Ergodic Theory, for almost every x with respect to any invariant measure, the behavior of the differential $T_x f^n$ for n large is described by the Lyapunov exponents $\lambda_1,\dots,\lambda_d$, at x. Pesin is able to build charts around almost every orbit in which this asymptotic linear behavior describes that of f at the first iteration. That is, there are diffeomorphisms $\Phi_x: U_x \subset M \to V_x \subset \mathbb{R}^d$ with a "reasonable dependence on x" such that the differential of $\Phi_{f^n x} \circ f^n \circ \Phi_x^{-1}$ at any point where it is defined is close to a diagonal matrix with entries $(e^{(\lambda_1\pm\epsilon)n}, e^{(\lambda_2\pm\epsilon)n},\dots,e^{(\lambda_d\pm\epsilon)n})$.

In this full generality, one already obtains significant results:

- The entropy is bounded by the expansion: $h(f,\mu) \leq \sum_{i=1}^d \lambda_i^+(\mu)$ [219]
- At almost every point x, there are strong stable resp. unstable manifolds $W^{ss}(x)$, resp. $W^{uu}(x)$, coinciding with the sets of points y such that $d(T^n x, T^n y) \to 0$ exponentially fast when $n \to \infty$, resp. $n \to -\infty$. The corresponding holonomies are *absolutely continuous* (see, e. g., [59]) like in the uniform case. This allows Ledrappier's definition of Sinai–Ruelle–Bowen measures [166] in that setting.
- Equality in the above formula holds if and only if μ is a Sinai–Ruelle–Bowen measure [167]. More generally the entropy can be computed as $\sum_{i=1}^d \gamma_i(\mu)\lambda_i^+(\mu)$ where the γ_i are some fractal dimensions related to the exponents.

Under the only assumption of hyperbolicity (i. e., no zero Lyapunov exponent almost everywhere), one gets further properties:

- Existence of an hyperbolic measure which is not periodic forces $h_{\text{top}}(f) > 0$ [*However $h(f,\mu)$ can be zero.*] by [147].
- μ is *exact dimensional* [29,256]: the limit $\lim_{r\to 0} \log \mu(B(x,r))/\log r$ exists μ-almost everywhere and is equal to the Hausdorff dimension of μ (the infimum of the Hausdorff dimension of the sets with full μ-measure). This is deduced from a more technical "asymptotic product structure" property of any such measure.

For hyperbolic Sinai–Ruelle–Bowen measures μ, one can then prove, e. g.,:

- *local ergodicity* [202]: μ has at most countably many ergodic components and μ-almost every point has a neighborhood whose Lebesgue-almost every point are contained in the basin of an ergodic component of μ.
- *Bernoulli* [198]: Each ergodic component of μ is conjugate in a measure-preserving way, up to a period, to a Bernoulli shift, that is, a full shift $\{1,\dots,N\}^{\mathbb{Z}}$ equipped with a product measure. This in particular implies mixing and sensitivity on initial conditions for a set of positive Lebesgue measure.

However, establishing even such a weak form of hyperbolicity is rather difficult. The fragility of this condition can be illustrated by the result [44,45] that the topologically generic area-preserving surface C^1-diffeomorphism is either uniformly hyperbolic or has Lebesgue almost everywhere vanishing Lyapunov exponents, hence is never non-uniformly hyperbolic! (but this is believed to be very specific to the very weak C^1 topology). Moreover, such weak hyperbolicity is not enough, with the current techniques, to build Sinai–Ruelle–Bowen measures or analyze maximum entropy measures only assuming non-zero Lyapunov exponents. Let us though quote two conjectures. The first one is from [251] [*We slightly strengthened Viana's statement for expository reasons.*]:

Conjecture 1 *Let f be a $C^{1+\epsilon}$-diffeomorphism of a compact manifold. If Lebesgue-almost every point x has well-defined Lyapunov exponents in every direction and none of these exponents is zero, then there exists an absolutely continuous invariant σ-finite positive measure.*

The analogue of this conjecture has been proved for C^3 interval maps with unique critical point and negative Schwarzian derivative by Keller [150], but only partial results are available for diffeomorphisms [168].

We turn to measures of maximum entropy. As we said, C^∞ smoothness is enough to ensure their existence but this is through a functional-analytic argument (allowed by Yomdin theory [254]) which says nothing about their structure. Indeed, the following problem is open:

Conjecture 2 *Let f be a $C^{1+\epsilon}$-diffeomorphism of a compact surface. If the topological entropy of f is nonzero then f has at most countably many ergodic invariant measures maximizing entropy.*

The analogue of this conjecture has been proved for $C^{1+\epsilon}$ interval maps [64,66,70]. In the above setting a classical result of Katok shows the existence of uniformly hyperbolic compact invariant subsets with topological entropy arbitrarily close to that of f implying the existence of many periodic points:

$$\limsup_{n\to\infty} \frac{1}{n} \log \#\{x \in M\colon f^n(x) = x\} \geq h_{\text{top}}(f)\,.$$

The previous conjecture would follow from the following one:

Conjecture 3 *Let f be a $C^{1+\epsilon}$-diffeomorphism of a compact manifold. There exists an invariant subset $X \subset M$, carrying all ergodic measures with maximum entropy, such that the restriction $f|X$ is conjugate to a countable state topological Markov shift (See the glossary.).*

Systems with Discontinuities

We now consider stronger assumptions to be able to build the relevant measures.

The simplest step beyond uniformity is to allow discontinuities, considering piecewise expanding maps. The discontinuities break the rigidity of the uniformly expanding situation. For instance, their symbolic dynamics are usually no longer subshifts of finite type though they still retain some "simplicity" in good cases (see [68]).

To understand the problem in constructing the absolutely continuous invariant measures, it is instructive to consider the pushed forwards of a smooth measure. Expansion tends to keep the measure smooth whereas discontinuities may pile it up, creating non-absolute continuity in the limit. One thus has to check that expansion wins. In dimension 1, a simple fact resolves the argument: under a high enough iterate, one can make the expansion arbitrarily large everywhere, whereas a small interval can be chopped into at most two pieces.

Lasota and Yorke [165] found a suitable framework. They considered C^2 interval maps with $|f'(x)| \geq \text{const} > 1$ except at finitely many points. They used the Ruelle transfer operator directly on the interval. Namely they studied

$$(L\phi)(x) = \sum_{y\in T^{-1}x} \frac{\phi(y)}{|T'(y)|}$$

acting on functions $\phi\colon [0,1] \to \mathbb{R}$ with bounded variation and obtained the invariant density as the eigenfunction associated to the eigenvalue 1. One can then prove a *Lasota–Yorke inequality* (which might more accurately be called Doeblin–Fortet since it was introduced in the theory of Markov chains much earlier):

$$\|L\phi\|_{\mathrm{BV}} \le \alpha\|\phi\|_{\mathrm{BV}} + \beta\|\phi\|_1 \tag{3}$$

where $\|\cdot\|_{\mathrm{BV}}$, $\|\cdot\|_1$ are a strong and a weak norm, respectively and $\alpha < 1$ and $\beta < \infty$. One can then apply general theorems [143] or [196] (see [21] for a detailed presentation of this approach and its variants). Here α can essentially be taken as 2 (reflecting the locally simple discontinuities) divided by the minimum expansion: so $\alpha < 1$, perhaps after replacing T with an iterate. In particular, the existence of a finite statistical description then follows (see [61] for various generalizations and strengthenings of this result on the interval).

The situation in higher dimension is more complex for the reason explained above. One can obtain inequalities such as (3) on suitable if less simple functional spaces (see, e. g., [231]) but proving $\alpha < 1$ is another matter: discontinuities can get arbitrarily complex under iteration. [67,241] show that indeed, in dimension 2 and higher, piecewise uniform expansion (with a finite number of pieces) is not enough to ensure a finite statistical description if the pieces of the map have only finite smoothness. In dimension 2, resp. 3 or more, piecewise real-analytic, resp. piecewise affine, is enough to exclude such examples [65,240], resp. [242]. [82] has shown that, for any $r > 1$, an open and dense subset of piecewise C^r and expanding maps have a finite statistical description.

Piecewise hyperbolic diffeomorphisms are more difficult to analyze though several results (conditioned on technical assumptions that can be checked in many cases) are available [22,74,230,257].

Interval Maps with Critical Points

A more natural but also more difficult situation is a map for which the uniformity of the expansion fails because of the existence of critical points. [*Note that, by a theorem of Mañé a circle map without critical points or indifferent periodic point is either conjugate to a rotation or uniformly expanding* [181].]

A class which has been completely analyzed at the level of the above conjecture is that of *real-analytic families of maps of the interval* $f_t\colon [0,1] \to [0,1]$, $t \in I$, with a unique critical point, the main example being the quadratic family $Q_t(x) = tx(1-x)$ for $0 \le t \le 4$.

It is not very difficult to find quadratic maps with the following two types of behavior:

(stable) the orbit of Lebesgue-almost every $x \in [0,1]$ tends to an attracting periodic orbit;

(chaotic) there is an absolutely continuous invariant probability measure μ whose basin contains Lebesgue-almost every $x \in [0,1]$.

To realize the first it is enough to arrange the critical point to be periodic. One can easily prove that this stable behavior occurs on an open set of parameters –thus it is stable with respect to the parameter or the dynamical system. The second occurs for Q_4 with $\mu = \mathrm{d}x/\sqrt{\pi x(1-x)}$. It is much more difficult to show that this chaotic behavior occurs for a set of parameters of positive Lebesgue measure. This is a theorem of Jakobson [145] for the quadratic family (see for a recent variant [265]). Let us sketch two main ingredients of the various proofs of this theorem. The first is *inducing*: around Lebesgue-almost every point $x \in [0,1]$ one tries to find a time $\tau(x)$ and an interval $J(x)$ such that $f^{\tau(x)}\colon J(x) \to f^{\tau(x)}(J(x))$ is a map with good expansion and distortion properties. This powerful idea appears in many disguises in the non-uniform hyperbolic theory (see for instance [133,262]). The second ingredient is *parameter exclusion*: one removes the parameters at which a good inducing scheme cannot be built. More precisely one proceeds inductively, performing simultaneously the inducing and the exclusion, the good properties of the early stage of the inducing allowing one to control the measure of the parameters that need to be excluded to continue [30,145]. Indeed, the expansion established at a given stage allows to transfer estimates in the dynamical space to the parameter space.

Using methods from complex analysis and renormalization theory one can go much further and prove the following difficult theorems (actually the product of the work of many people, including Avila, Graczyk, Kozlovski, Lyubich, de Melo, Moreira, Shen, van Strien, Swiatek), which in particular solves Palis conjecture in this setting:

Theorem 2 ([117,160,178]) *Stable maps (that is, such that Lebesgue almost every orbit converges to one of finitely many periodic orbits) form an open and dense set among C^r interval maps, for any $r \ge 2$. [In fact this is even true for polynomials.].*

The picture has been completed in the *unimodal case* (that is, with a unique critical point):

Theorem 3 ([19,20,117,159,178]) *Let $f_t\colon [0,1] \to [0,1]$, $t \in [t_0, t_1]$, be a real-analytic family of unimodal maps. Assume that it is not degenerate [f_{t_0} and f_{t_1} are not conjugate]. Then:*

- *The set of t such that f_t is chaotic in the above sense has positive Lebesgue measure;*
- *The set of t such that f_t is stable is open and dense;*
- *The remaining set of parameters has zero Lebesgue measure. [This set of "strange parameters" of zero Lebesgue measure has however positive Hausdorff dimension according to work of Avila and Moreira. In particular each of the following situations is realized on a set of parameters t of positive Hausdorff dimension: non-existence of the Birkhoff limit at Lebesgue-almost every point, the physical measure is δ_p for p a repelling fixed point, the physical measure is non-ergodic.]*

We note that the theory underlying the above theorem yields much more results, including a very paradoxical rigidity of typical analytic families as above. See [19].

Non-uniform Expansion/Contraction

Beyond the dimension 1, only partial results are available. The most general of those assume uniform contraction or expansion along some direction, restricting the non-uniform behavior to an invariant sub-bundle often one-dimensional, or "one-dimensional-like".

A first, simpler situation is when there is a *dominated decomposition* with a uniformly expanding term: there is a continuous and invariant splitting of the tangent bundle, $T_\Lambda M = E^{\mathrm{uu}} \oplus E^{\mathrm{cs}}$, over some Λ an attracting set: for all unit vectors $v^u \in E^{\mathrm{uu}}$, $v^c \in E^{\mathrm{cs}}$,

$$\|(f^n)'(x).v^u\| \geq C\lambda^n \quad \text{and}$$
$$\|(f^n)'(x).v^c\| \leq C\mu^n \|(f^n)'(x).v^u\| \,.$$

Standard techniques (pushing the Riemannian volume of a piece of unstable leaf and taking limits) allow the construction of Gibbs u-states as introduced by [205].

Theorem 4 (Alves–Bonatti–Viana [8]) *[A slightly different result is obtained in [49].] Let $f\colon M \to M$ be a C^2 diffeomorphism with an invariant compact subset Λ. Assume that there is a dominated splitting $T_\Lambda M = E^u \oplus E^{cs}$ such that, for some $c > 0$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \prod_{k=0}^{n-1} \|f'(f^x)|E^{cs}\| \leq -c < 0$$

on a subset of Λ of positive Lebesgue measure. Then this subset is contained, up to a set of zero Lebesgue measure, in the union of the basins of finitely many ergodic and hyperbolic Sinai – Ruelle – Bowen measures.

The non-invertible, purely expansive version of the above theorem can be applied in particular to the following maps of the cylinder ($d \geq 16$, a is properly chosen close to 2 and α is small):

$$f(\theta, x) = (\mathrm{d}x \bmod 2\pi, a - x^2 + \epsilon \sin(\theta))$$

which are natural examples of maps with multidimensional expansion and critical lines considered by Viana [249]. A series of works have shown that the above maps fit in the above non-uniformly expanding setting with a proper control of the critical set and hence can be thoroughly analyzed through variants of the above theorem [4,249] and the references in [5]. For a, b properly chosen close to 2 and small ϵ, the following maps should be even more natural examples:

$$f(x, y) = (a - x^2 + \epsilon y, b - y^2 + \epsilon x)\,. \tag{4}$$

However the inexistence of a dominated splitting has prevented the analysis of their physical measures. See [66,70] for their maximum entropy measures.

Cowieson and Young have used completely different techniques (thermodynamical formalism, Ledrappier–Young formula and Yomdin theory on entropy and smoothness) to prove the following result (see [13] for related work):

Theorem 5 (Cowieson–Young [83]) *Let $f\colon M \to M$ be a C^∞ diffeomorphism of a compact manifold. Assume that f admits an attractor $\Lambda \subset M$ on which the tangent bundle has an invariant continuous decomposition $T_\Lambda M = E^+ \oplus E^-$ such that all vectors of $E^+ \setminus \{0\}$, resp. $E^- \setminus \{0\}$, have positive, resp. negative, Lyapunov exponents. Then any zero-noise limit measure μ of f is a Sinai–Ruelle–Bowen measure and therefore, if it is ergodic and hyperbolic, a physical measure.*

One can hope, that typically, the latter ergodicity and hyperbolicity assumptions are satisfied (see, e. g., [27]).

By pushing classical techniques and introducing new ideas for generic maps with one expanding and one weakly contracting direction, Tsujii has been able to prove the following generic result (which can be viewed as a 2-dimensional extension of some of the one-dimensional results above: one adds a uniformly expanding direction):

Theorem 6 (Tsujii [243]) *Let M be a compact surface. Consider the space of C^{20} self-maps $f\colon M \to M$ which admits directions that are uniformly expanded [More precisely, there exists a continuous, forward invariant cone field which is uniformly expanded under the differential of f.]*

Then existence of a finite statistical description is both topologically generic and prevalent in this space of maps.

Hénon-Like Maps and Rank One Attractors

In 1976, Hénon [130] observed that the diffeomorphism of the plane

$$H_{a,b}(x, y) = (1 - ax^2 + y, bx)$$

seemed to present a "strange attractor" for $a = 1.4$ and $b = 0.3$, that is, the points of a numerically simulated orbit seemed to draw a set looking locally like the product of a segment with a Cantor set. This attractor seems to be supported by the unstable manifold $W^u(P)$ of the hyperbolic fixed point P with positive absciss. On the other hand, the Plykin classification [207] excluded the existence of a uniformly hyperbolic attractor for a dissipative surface diffeomorphism.

For almost twenty years the question of the existence of such an attractor (by opposition to an attracting periodic orbit with a very long period) remained open. Indeed, one knew since Newhouse that, for many such maps, there exist infinitely many such periodic orbits which are very difficult to distinguish numerically. But in 1991 Benedicks and Carleson succeeded in proposing an argument refining (with considerable difficulties) their earlier proof of Jakobson one-dimensional theorem and established the first part of the following theorem:

Theorem 7 (Benedicks–Carleson [31]) *For any $\epsilon > 0$, for $|b|$ small enough, there is a set A with* $\mathrm{Leb}(A) > 0$ *satisfying: for all $a \in A$, there exists $z \in W^u(P)$ such that*

- *The orbit of z is dense in $W^u(P)$;*
- $\liminf_{n\to\infty} \frac{1}{n} \log \|(f^n)'(z)\| > 0$.

Further properties were then established, especially by Benedicks, Viana, Wang, Young [32,34,35,264]. Let us quote the following theorem of Wang and Young which includes the previous results:

Theorem 8 ([264]) *Let $T_{\mathrm{ab}}\colon S^1 \times [-1, 1] \to S^1 \times [-1, 1]$ be such that*

- $T_{a0}(S^1 \times [-1, 1]) \subset S^1 \times \{0\}$;
- *For $b > 0$, T_{ab} is a diffeomorphism on its image with*

$$c^{-1} b \le |\det T_{\mathrm{ab}}(x, y)| \le c \cdot b$$

for some $c > 1$ and all $(x, y) \in S^1 \times [-1, 1]$ and all (a, b).

Let $f_a\colon S^1 \to S^1$ be the restriction of T_{a0}. Assume that $f = f_0$ satisfies:

- *Non-degenerate critical points: $f'(c) = 0 \implies f''(c) \neq 0$;*
- *Negative Schwarzian derivative: for all $x \in S^1$ non-critical, $f'''(x)/f'(x) - 3/2(f''(x)/f'(x))^2 < 0$;*
- *No indifferent or attracting periodic point, i. e., x such that $f^n(x) = x$ and $|(f)'(x)| \le 1$;*
- *Misiurewicz condition: $d(f^n c, d) > c > 0$ for all $n \ge 1$ and all critical points c, d.*

Assume the following transversality condition on f at $a = 0$: for every critical point c, $(\mathrm{d}/da)(f_a(c_a) - p_a) \neq 0$ if c_a is the critical point of f_a near c and p_a is the point having the same itinerary under f_a as $f(c)$ under c. Assume the following non-degeneracy of T: $f_0'(c) = 0 \implies \partial T_{00}(c, 0)/\partial y \neq 0$.

- *T_{ab} restricted to a neighborhood of $S^1 \times \{0\}$ has a finite statistical description by a number of hyperbolic Sinai–Ruelle–Bowen measures bounded by the number of critical points of f;*
- *There is exponential decay of correlations and a Central Limit Theorem (see below) – except, in an obvious way, if there is a periodic interval with period > 1;*
- *There is a natural coding of the orbits that remains for ever close to $S^1 \times \{0\}$ by a closed invariant subset of a full shift.*

Very importantly, the above dynamical situation has been shown to occur near typical homoclinic tangencies: [190] proved that there is an open and dense subset of the set of all C^3 families of diffeomorphisms unfolding a first homoclinic tangency such that the above holds. However [201] shows that the set of parameters with a Henon-like attractor has zero Lebesgue density at the bifurcation itself, at least under an assumption on the so-called stable and unstable Hausdorff dimensions. [95] establishes positive density for another type of bifurcation. Furthermore [191] has related the Hausdorff dimensions to the abundance of uniformly hyperbolic dynamics near the tangency.

[248] is able to treat situations with more than one contracting direction. More recently [266] has proposed a rather general framework, with easily checkable assumptions in order to establish the existence of such dynamics in various applications. See also [122,252] for applications.

Statistical Properties

The ergodic theorem asserts that time averages of integrable functions converge to phase space averages for any ergodic system. The speed of convergence is quite arbitrary in that generality [161] (only upcrossing inequalities seem to be available [38,132]), however many results are available under very natural hypothesis as we are going to explain in this section. The underlying idea is that for sufficiently chaotic dynamics T and reasonably smooth ob-

servables ϕ, the time averages

$$A_n\phi(x) := \frac{1}{n}\sum_{k=0}^{n-1}\phi\circ T^k(x)$$

should behave as averages of independent and identically distributed random variables and therefore satisfy the classical limit theorems of probability theory.

The dynamical systems which are amenable to current technology are in a large part [*But other approaches are possible. Let us quote the work* [98] *on partially hyperbolic systems, for instance.*] those that can be reduced to the following type:

Definition 1 Let $T\colon X\to X$ be a nonsingular map on a probability, metric space $(X,\mathcal{B},\mu,d)$ with bounded diameter, preserving the probability measure μ. This map is said to be **Gibbs–Markov** if there exists a countable (measurable) partition α of X such that:

1. For all $a\in\alpha$, T is injective on a and $T(a)$ is a union of elements of α.
2. There exists $\lambda>1$ such that, for all $a\in\alpha$, for all points $x,y\in a$, $d(Tx,Ty)\geq\lambda d(x,y)$.
3. Let Jac be the *inverse* of the Jacobian of T. There exists $C>0$ such that, for all $a\in\alpha$, for all points $x,y\in a$, $|1-Jac(x)/Jac(y)|\leq Cd(Tx,Ty)$.
4. The map T has the "big image property": $\inf_{a\in\alpha}\mu(Ta)>0$.

Some piecewise expanding and C^2 maps are obviously Gibbs–Markov but the real point is that many dynamics can be reduced to that class by the use of inducing and tower constructions as in [262], in particular. This includes possibly piecewise uniformly hyperbolic diffeomorphisms, Collet–Eckmann maps of the interval [21] (typical chaotic maps in the quadratic family), billiards with convex scatterers [262], the stadium billiard [71], Hénon-like mappings [266].

We note that in many cases one is led to first analyze mixing properties through *decay of correlations*, i. e., to prove inequalities of the type [21]:

$$\left|\int_X\phi\cdot\psi\circ T^n\,\mathrm{d}\mu-\int_X\phi\,\mathrm{d}\mu\int_X\psi\,\mathrm{d}\mu\right|\leq\|\phi\|\cdot\|\psi\|_w\cdot a_n \tag{5}$$

where $(a_n)_{n\geq1}$ is some sequence converging to zero, e. g., $a_n=\mathrm{e}^{-\lambda n}$, $1/n^\alpha,\ldots$ and $\|\cdot\|$, $\|\cdot\|_w$ a strong and a weak norm (e. g., the variation norm and the L^1 norm). These rates of decay are often linked with return times statistics [263]. Rather general schemes have been developed to deduce various limit theorems such as those presented below from sufficiently quick decay of correlations (see notably [175] based on a dynamical variant of [113]).

Probabilistic Limit Theorems

The foremost limit property is the following:

Definition 2 A class C of functions $\phi\colon X\to\mathbb{R}$ is said to satisfy the Central Limit Theorem if the following holds: for all $\phi\in C$, there is a number $\sigma=\sigma(\phi)>0$ such that:

$$\lim_{n\to\infty}\mu\left(\left\{x\in X\colon\frac{A_n\phi(x)-\int\phi\,\mathrm{d}\mu}{\sigma n^{-1/2}}\leq t\right\}\right)=\int_{-\infty}^{t}\mathrm{e}^{-x^2/2\sigma^2}\frac{\mathrm{d}x}{\sqrt{2\pi}\sigma} \tag{6}$$

except for the degenerate case when $\phi(x)=\psi(Tx)-\phi(x)+\text{const}$.

The Central Limit Theorem can be seen in many cases as essentially a by-product of fast decay of correlations [175], i. e., if $\sum_{n\geq0}a_n<\infty$ in the notations of Eq. (5). It has been established for Hölder-continuous observables for many systems together with their natural invariant measures including: uniformly hyperbolic attractors, piecewise expanding maps of the interval [174], Collet–Eckmann unimodal maps on the interval [152,260], piecewise hyperbolic maps [74], billiards with convex scatterers [238], Hénon-like maps [35].

Remark 5 The classical Central Limit Theorem holds for square-integrable random variables [193]. For maps exhibiting *intermittency* (e.g, interval maps like $f(x)=x+x^{1+\alpha}$ mod 1 with an indifferent fixed point at 0) the invariant density has singularities and the integrability condition is non longer automatic for smooth functions. One can then observe convergence to stable laws, instead of the normal law [114].

A natural question is the speed of the convergence in (6). The *Berry–Esseen inequality*:

$$\left|\mu\left(\left\{\frac{A_n\phi(x)-\int\phi\,\mathrm{d}\mu}{\sigma n^{-1/2}}\leq t\right\}\right)-\int_{-\infty}^{t}\mathrm{e}^{-x^2/2\sigma^2}\frac{\mathrm{d}x}{\sqrt{2\pi}\sigma}\right|\leq\frac{C}{n^{\delta/2}}$$

for some $\delta>0$. It holds with $\delta=1$ in the classical, probabilistic setting.

The *Local Limit Theorem* looks at a finer scale, asserting that for any finite interval $[a,b]$, any $t\in\mathbb{R}$,

$$\lim_{n\to\infty} \sqrt{n}\mu\left(\left\{x \in X\colon nA_n\phi(x) \in [a,b] + \sqrt{n}t + n\int \phi \, \mathrm{d}\mu\right\}\right) = |b-a|\frac{\mathrm{e}^{-t^2/2\sigma^2}}{\sqrt{2\pi}\sigma}\,.$$

Both the Berry–Esseen inequality and the local limit theorem have been shown to hold for non-uniformly expanding maps [115] (also [62,216]).

Almost Sure Results

It is very natural to try and describe the statistical properties of the averages $A_n(x)$ for almost every x, instead of the weaker above statements in probability over x. An important such property is the *almost sure invariance principle*. It asks for the discrete random walk defined by the increments of $\phi \circ T^n(x)$ to converge, after a suitable renormalization, to a Brownian motion. This has been proved for systems with various degrees of hyperbolicity [92,98,135,183]. Another one is the almost sure Central Limit Theorem. In the independent case (e. g., if $X_1, X_2, \ldots$ are independent and identically distributed random variables in L^2 with zero average and unit variance), the almost sure Central Limit Theorem states that, almost surely:

$$\frac{1}{\log n}\sum_{k=1}^{n}\frac{1}{k}\delta_{\sum_{j=0}^{k-1} X_j/\sqrt{k}}$$

converges in law to the normal distribution. This implies, that, almost surely, for any $t \in \mathbb{R}$:

$$\lim_{n\to\infty}\frac{1}{\log n}\sum_{k=1}^{n}\frac{1}{k}\delta_{\left\{\sum_{j=0}^{k-1} X_j/\sqrt{k}\le t\right\}} = \int_{-\infty}^{t} \mathrm{e}^{-x^2/2\sigma^2}\frac{\mathrm{d}x}{\sqrt{2\pi}\sigma}$$

compare to (6).

A general approach is developed in [73], covering Gibbs–Markov maps and those that can be reduced to it. They show in particular that the dynamical properties needed for the classical Central Limit Theorem in fact suffice to prove the above almost invariance principle and even the almost sure version of the Central Limit Theorem (using general probabilistic results, see [37,255]).

Other Statistical Properties

Essentially all the statistical properties of sums of independent identically distributed random variables can be established for tractable systems. Thus one can also prove large deviations [156,172,259], iterated law of the logarithm, etc.

We note that the monograph [78] contains a nice introduction to the current work in this area.

Orbit Complexity

The orbit complexity of a dynamical system $f\colon M \to M$ is measured by its topological and measured entropies. We refer to ► Entropy in Ergodic Theory for detailed definitions.

The Variational Principle

Bowen–Dinaburg and Katok formulae can be interpreted as meaning that the topological entropy counts the number of arbitrary orbits whereas the measured entropy counts the number of orbits relevant for the given measure. In most situations, and in particular for continuous self-map of compact metric spaces, the following *variational principle* holds:

$$h_{\mathrm{top}}(f) = \sup_{\mu\in M(f)} h(f,\mu)$$

where $M(f)$ is the set of all invariant probability measures.

This is all the more striking in light of the fact that for many systems, the set of points which are typical from the point of view of ergodic theory [*for instance, those x such that* $\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\phi(T^k x)$ *exists for all continuous functions* ϕ.] is topologically negligible [*A subset of a countable union of closed sets with empty interior, that is, meager or first Baire category*.].

Strict Inequality For a fixed invariant measure, one can only assert that $h(f,\mu) \le h_{\mathrm{top}}(f)$. One should be aware that this inequality may be strict even for a measure with full support. For instance, it is not difficult to check that the full tent map with $h_{\mathrm{top}}(f) = \log 2$, admits ergodic invariant measures with full support and zero entropy.

There are also examples of dynamical systems preserving Lebesgue measure which have simultaneously positive topological entropy and zero entropy with respect to Lebesgue measure. That this occurs for C^1 surface diffeomorphisms preserving area is a simple consequence of a theorem of Bochi [44] according to which, generically in the C^1 topology, such a diffeomorphism is either uniformly hyperbolic or with Lyapunov exponents Lebesgue-almost everywhere zero. [*Indeed, it is easy to build such a diffeomorphism having both a uniformly hyperbolic compact invariant subset which will have robustly positive topological entropy and a non-degenerate elliptic fixed point which will prevent uniform hyperbolicity and therefore force all Lyapunov exponents to be zero. But Ruelle–Margulis*

inequality then implies that the entropy with respect to Lebesgue measure is zero.] Smooth examples also exist [46].

Remark 6 Algorithmic complexity [170] suggests another way to look at orbit complexity. One obtains in fact in this way another formula for the entropy. However this point of view becomes interesting in some settings, like extended systems defined by partial differential equations in unbounded space. Recently, [47] has used this approach to build interesting invariant measures.

Orbit Complexity on the Set of Measures

We have considered the entropies of each invariant measure separately, sharing only the common roof of topological entropy. One may ask how these different complexity sit together. A first answer is given by the following theorem in the symbolic and continuous setting. Let .

Theorem 9 (Downarowicz–Serafin [102]) *Let K be a Choquet simplex and $H\colon K \to \mathbb{R}$ be a convex function.*

Say that H is realized by a self-map $f\colon X \to X$ and its set $M(f)$ of f-invariant probability measures equipped with the weak star topology if the following holds. There exists an affine homeomorphism $\Psi\colon M(f) \to K$ such that, if $h\colon M(f) \to [0,\infty]$ is the entropy function, $H = h \circ \Psi$.

Then

- *H is realized by some continuous self-map of a compact space if and only if it is an increasing limit of upper semi-continuous and affine functions.*
- *H is realized by some subshift on a finite alphabet, i. e., by the left shift on a closed invariant subset Σ of $\{1,2,\dots,N\}^{\mathbb{Z}}$ for some $N < \infty$, if and only if it is upper semi-continuous*

Thus, in both the symbolic and continuous settings it is possible to have a unique invariant measure with any prescribed entropy. This stands in contrast to surface $C^{1+\epsilon}$-diffeomorphisms for which the set of the entropies of ergodic invariant measures is always the interval $[0, h_{\text{top}}(f)]$ as a consequence of [147].

Local Complexity

Recall that the topological entropy can be computed as: $h_{\text{top}}(f) = \lim_{\epsilon\to 0} h_{\text{top}}(f,\epsilon)$ where:

$$h_{\text{top}}(f,\epsilon) := \lim_{n\to\infty} \frac{1}{n} \log s(\delta, n, X)$$

where $s(\delta, n, E)$ is the maximum cardinality of a subset S of E such that:

$$x \neq y \implies \exists 0 \le k < n \; d(f^k x, f^k y) \ge \epsilon$$

(see Bowen's formula of the topological entropy ▶ Entropy in Ergodic Theory). Likewise, the measure-theoretic entropy $h(T,\mu)$ of an ergodic invariant probability measure μ is $\lim_{\epsilon\to 0} h(f,\mu,\epsilon)$ where:

$$h(T,\epsilon) := \lim_{n\to\infty} \frac{1}{n} \log r(\delta, n, \mu)$$

where $r(\delta, n, \mu)$ is the minimum cardinality of $C \subset X$ such that

$$\mu\Big(\big\{x \in X\colon \exists y \in C \quad \text{such that} \quad \forall 0 \le k < n \; d(f^k x, f^k y) < \epsilon\big\}\Big) > 1/2\,.$$

One can ask *at which scales does entropy arise for a given dynamical system?*, i. e., how the above quantities $h(T,\epsilon)$, $h(T,\mu,\epsilon)$ converge when $\epsilon \to 0$.

An answer is provided by the *local entropy*. [*This quantity was introduced by Misiurewicz* [186] *under the name* conditional topological entropy *and is called* tail entropy *by Downarowicz* [100].] For a continuous map f of a compact metric space X, it is defined as:

$$h_{\text{loc}}(f) := \lim_{\epsilon\to 0} h_{\text{loc}}(f,\epsilon) \quad \text{with}$$

$$h_{\text{loc}}(f,\epsilon) := \sup_{x\in X} h_{\text{loc}}(f,\epsilon,x) \quad \text{and}$$

$$h_{\text{loc}}(f,\epsilon,x) := \lim_{\delta\to 0} \limsup_{n\to\infty} \frac{1}{n} \log s\Big(\delta, n, \big\{y \in X\colon \forall k \ge 0 \; d(f^k y, f^k x) < \epsilon\big\}\Big)$$

Clearly from the above formulas:

$$h_{\text{top}}(f) \le h_{\text{top}}(f,\delta) + h_{\text{loc}}(f,\delta) \quad \text{and}$$
$$h(f,\mu) \le h(f,\mu,\delta) + h_{\text{loc}}(f,\delta)\,.$$

Thus the local entropy bounds the defect in uniformity with respect to the measure of the pointwise limit $h(f,\mu) = \lim_{\delta\to 0} h(f,\mu,\delta)$. An exercise in topology shows that the local entropy therefore also bounds the defect in upper semicontinuity of $\mu \mapsto h(f,\mu)$. In fact, by a result of Downarowicz [100] (extended by David Burguet to the non-invertible case), there is a *local variational principle*:

$$h_{\text{loc}}(f) = \lim_{\delta\to 0}\left(\sup_{\mu} h(f,\mu) - h(f,\mu,\delta)\right) = \sup_{\mu} \limsup_{\nu\to\mu} h(f,\nu) - h(f,\mu)$$

for any continuous self-map f of a compact metric space.

The local entropy is easily bounded for smooth maps using Yomdin's theory:

Theorem 10 ([64]) *For any C^r map f of a compact manifold, $h_{\mathrm{loc}}(f) \leq \frac{d}{r} \log \sup_x \|f'(x)\|$. In particular, $h_{\mathrm{loc}}(f) = 0$ if $r = \infty$.*

Thus C^∞ smoothness implies the existence of a maximum entropy measure (this was proved first by Newhouse) and the existence of *symbolic extension*: a subshift over a finite alphabet $\sigma : \Sigma \to \Sigma$ and a continuous and onto map $\pi : \Sigma \to M$ such that $\pi \circ \sigma = f \circ \sigma$. More precisely,

Theorem 11 (Boyle, Fiebig, Fiebig [56]) *Given a homeomorphism f of a compact metric space X, there exists a principal symbolic extension $\sigma : \Sigma \to \Sigma$, i. e., a symbolic extension such that, for every σ-invariant probability measure ν, $h(\sigma, \nu) = h(f, \nu \circ \pi^{-1})$, if and only if $h_{\mathrm{loc}}(f) = 0$.*

We refer to [55,101] for further results, including a realization theorem showing that the continuity properties of the measured entropy are responsible for the properties of symbolic extensions and also results in finite smoothness.

Global Simplicity

One can marvel at the power of mathematical analysis to analyze such complex evolutions. Of course another way to look at this is to remark that this analysis is possible *once this evolution has been fitted in a simple setting*: one had to move focus away from an individual, unpredictable orbit, of, say, the full tent map to the set of all the orbits of that map, which is essentially the set of all infinite sequences over two symbols: a very simple set indeed corresponding to full *combinatorial freedom* [[253] *describes a weakening of this which holds for all positive entropy symbolic dynamics.*]. The complete description of a given typical orbit requires an infinite amount of information, whereas the set of all orbits has a finite and very tractable definition. The complexity of the individual orbits is seen now as coming from purely random choices inside a simple structure.

The classical systems, namely uniformly expanding maps or hyperbolic diffeomorphisms of compact spaces, have a *simple* symbolic dynamics. It is not necessarily a full shift like for the tent map, but it is a subshift of finite type, i. e., a subshift obtained from a full shift by forbidding finitely many finite subwords. What happens outside of the uniform setting?

A fundamental example is provided by *piecewise monotone maps*, i. e., interval maps with finitely many critical points or discontinuities. The partition cut by these points defines a symbolic dynamics. This subshift is usually not of finite type. Indeed, the topological entropy taking arbitrary finite nonnegative values [*For instance, the topological entropy of the β-transformation, $x \mapsto \beta x \bmod 1$, is $\log \beta$ for $\beta \geq 1$.*], a representation that respects it has to use an uncountable class of models. In particular models defined by finite data, like the subshifts of finite type, cannot be generally adequate. However there are tractable "almost finite representations" in the following senses:

Most symbolic dynamics $\Sigma(T)$ of piecewise monotone maps T can be defined by finitely many infinite sequences, the *kneading invariants* of Milnor and Thurston: $\kappa_0^+, \kappa_1^-, \kappa_1^+, \ldots, \kappa_{d+1}^- \in \{0, \ldots, d\}^{\mathbb{N}}$ if d is the number of critical/discontinuity points. [*The kneading invariants are the suitably defined (left and right) itineraries of the critical/discontinuity points and endpoints.*] Namely,

$$\Sigma(T) = \{\alpha \in \{0, \ldots, d\}^{\mathbb{N}} : \forall n \geq 0 \quad \kappa_{\alpha_n}^+ \preceq \sigma^n \alpha \preceq \kappa_{\alpha_n+1}^-\}$$

where $\preceq$ is a total order on $\{0, \ldots, d\}^{\mathbb{N}}$ making the coding $x \mapsto \alpha$ non-decreasing. Observe how the kneading invariants determine $\Sigma(T)$ in an effective way: knowing their first n symbols is enough to know the sequences of length n which begin sequences of $\Sigma(T)$. We refer to [91] for the wealth of information that can be extracted from these kneading invariants following Milnor and Thurston [184].

This form of global simplicity can be extended to other classes of non-uniformly expanding maps, including those like Eq. (4) using the notions of subshifts and puzzles of *quasi-finite type* [68,69]. This leads to the notion and analysis of *entropy-expanding maps*, a new open class of non-uniformly expanding maps admitting critical hypersurfaces, defined purely in terms of entropies including the otherwise untractable examples of Eq. (4).

A generalization of the representation of uniform systems by subshifts of finite type is provided by *strongly positive recurrent countable state Markov shifts*, a subclass of Markov shifts (*see glossary.*) which shares many properties with the subshifts of finite type [57,123,124,224,229].

These "simple" systems admit a *classification result* which in particular identifies their measures with entropy close to the maximum [57]. Such a classification generalizes [164]. The "ideology" here is that complexity of individual orbits in a simple setting must come from randomness, but purely random systems are classified by their entropy according to Ornstein [197].

Stability

By definition, chaotic dynamical systems have orbits which are unstable and numerically unpredictable. It is all the more surprising that, once one accepts to consider

their dynamics globally, they exhibit very good stability properties.

Structural Stability

A simple form of stability is *structural stability*: a system $f: M \to M$ is structurally C^r-stable if any system g sufficiently C^r-close to f is topologically the same as f, formally: g is topologically conjugate, i. e., there is some homeomorphism [*If h were C^1, the conjugacy would imply, among other things, that for every p-periodic point:* $\det((f^p)'(x)) = \det((g^p)'(h(p)))$, *a much too strong requirement.*] $h: M \to M$ mapping the orbits of f to those of g, i. e., $g \circ h = h \circ f$.

Andronov and Pontryaguin argued in the 1930's that only such structurally stable systems are physically relevant. Their idea was that the model of a physical system is always known only to some degree of approximation, hence mathematical model whose structure depends on arbitrarily small changes should be irrelevant.

A first question is: *What are these structurally stable systems?* The answer is quite striking:

Theorem 12 (Mañé [182]) *Let $f: M \to M$ be a C^1 diffeomorphism of a compact manifold.*

f is structurally stable among C^1-diffeomorphisms of M if and only if f is uniformly hyperbolic on its chain recurrent set. [A point x is chain recurrent if, for all $\epsilon > 0$, there exists a finite sequence $x_0, x_1, \ldots, x_n$ such that $x_0 = x_n = x$ and $d(f(x_k), x_{k+1}) < \epsilon$. The chain recurrent set is the set of all chain recurrent points.]

A basic idea in the proof of the theorem is that failure of uniform hyperbolicity gives the opportunity to make an arbitrarily small perturbation contradicting the structural stability. In higher smoothness the required perturbation lemmas (e. g., the closing lemma [14,148,180]) are not available.

We note that uniform hyperbolicity without invertibility does not imply C^1-stability [210].

A second question is: *are these stable systems dense?* (So that one could offer structurally stable models for all physical situations). A deep discovery around 1970 is that this is not the case:

Theorem 13 (Abraham–Smale, Simon [3,234]) *For any $r \geq 1$ and any compact manifold M of dimension ≥ 3, the set of uniformly hyperbolic diffeomorphisms is not dense in the space of C^r diffeomorphisms of M. [They use the phenomenon called "heterodimensional homoclinic intersections".]*

Theorem 14 (Newhouse [194]) *For any $r \geq 2$, for any compact manifold M of dimension ≥ 2, the set of uniformly hyperbolic diffeomorphisms is not dense in the space of C^r diffeomorphisms of M. More precisely, there exists a non-empty open subset in this space which contains a dense G_δ subset of diffeomorphisms with infinitely many periodic sinks. [So these diffeomorphisms have no finite statistical description.]*

Observe that it is possible that uniform hyperbolicity could be dense among *surface C^1-diffeomorphisms* (this is the case for C^1 circle maps by a theorem of Jakobson [145]).

In light of Mañé C^1-stability theorem this implies that *structurally stable systems are not dense*, thus one can robustly see behaviors that are topologically modified by arbitrarily small perturbations (at least in the C^1-topology)! So one needs to look beyond these and face that topological properties of relevant dynamical system are not determined from "finite data". It is natural to ask whether the dynamics is almost determined by "sufficient data".

Continuity Properties of the Topological Dynamics

Structural stability asks the topological dynamics to remain *unchanged* by a small perturbation. It is probably at least as interesting to ask it to *change continuously*. This raises the delicate question of which topology should be put on the rather wild set of topological conjugacy classes. It is perhaps more natural to associate to the system a topological invariant taking value in a more manageable set and ask whether the resulting map is continuous.

A first possibility is Zeeman's *Tolerance Stability Conjecture*. He associated to each diffeomorphism the set of all the closures of all of its orbits and he asked whether the resulting map is continuous on a dense G_δ subset of the class of C^r-diffeomorphisms for any $r \geq 0$. This conjecture remains open, we refer to [85] for a discussion and related progress.

A simpler possibility is to consider our favorite topological invariant, the topological entropy, and thus ask whether the dynamical complexity as measured by the entropy is a stable phenomenon. $f \mapsto h_{\text{top}}(f)$ is lower semicontinuous for f among C^0 maps of the interval [*On the set of interval maps with a bounded number of critical points, the entropy is continuous* [188]. *Also $t \mapsto h_{\text{top}}(Q_t)$ is non-decreasing by complex arguments* [91], *though it is a non-smooth function.*] [187] and for f among $C^{1+\epsilon}$-diffeomorphisms of a compact surface [147]. [*It is an important open question whether this actually holds for C^1-diffeomorphisms. It fails for homeomorphisms* [214].] In both cases, one shows the existence of structurally stable invariant uniformly expanding or hyperbolic subsets with topological entropy close to that of the whole dynamics.

On the other hand $f \mapsto h_{\text{top}}(f)$ is upper semi-continuous for C^∞ maps [195,254].

Statistical Stability

Statistical stability is the property that *deterministic perturbations* of the dynamical system cause only small changes in the physical measures, usually with respect to the weak star topology on the space of measures. When the physical measure μ_g is uniquely defined for all systems g near f, statistic stability is the continuity of the map $g \mapsto \mu_g$ thus defined.

Statistical stability is known in the uniform setting and also in the piecewise uniform case, provided the expansion is strong enough [25,42,151] (otherwise there are counter-examples, even in the one-dimensional case).

It also holds in some non-uniform settings without critical behavior, in particular for maps with dominated splitting satisfying robustly a separation condition between the positive and negative Lyapunov exponents [247] (see also [7]).

However statistical unstability occurs in dynamics with critical behaviors [18]:

Theorem 15 *Consider the quadratic family $Q_t(x) = tx(1-x)$, $t \in [0,4]$. Let $I \subset [0,4]$ be the full measure subset of $[0,4]$ of good parameters t such that in particular Q_t admits a physical measure (necessarily unique). For $t \in I$, let μ_t be this physical measure.*

Lebesgue almost every $t \in I$ such that μ_t is not carried by a periodic orbit [At such parameters, statistical stability is easily proved.], is a discontinuity point of the map $t \mapsto \mu_t$ [$M([0,1])$ is equipped with the vague topology.]. However, for Lebesgue almost every t, there is a subset $I_t \subset I$ for which t is a Lebesgue density point [t is a Lebesgue density point if for all $r > 0$, the Lebesgue measure $m(I_t \cap [t-r, t+r]) > 0$ and $\lim_{\epsilon \to 0} m(I_t \cap [t-\epsilon, t+\epsilon])/2\epsilon = 1$.] and such that $\mu\colon I_t \to M([0,1])$ is continuous at t.

Stochastic Stability

A physically motivated and a technically easier approach is to study stability of the physical measure under *stochastic perturbations*. For simplicity let us consider a diffeomorphism of a compact subset of $\mathbb{R}^d$ allowing for a direct definition of *additive noise*. Let $\psi(x)\mathrm{d}x$ be an absolutely continuous probability law with compact support. [*Sometimes additional properties are required of the density, e. g., $\psi(x) = \phi(x)1_B$ where 1_B is the characteristic function of the unit ball and $C^{-1} \le \phi(x) \le C$ for some $C < \infty$.*] For $\epsilon > 0$, consider the Markov chain f_ϵ with state space M and transition probabilities:

$$P_\epsilon(x, A) = \int_A \psi\left(\frac{y - f(x)}{\epsilon}\right) \frac{\mathrm{d}y}{\epsilon^d} .$$

The evolution of measures is given by: $(f_\epsilon \mu)(A) = \int_M P_\epsilon(x, A)\, \mathrm{d}\mu$. Under rather weak irreducibility assumptions on f, f_ϵ has a *unique invariant measure* μ_ϵ (contrarily to f) and μ_ϵ is absolutely continuous. When f has a unique physical measure μ, it is said to be *stochastically stable* if $\lim_{\epsilon \to 0} \mu_\epsilon = \mu$ in the appropriate topology (the weak star topology unless otherwise specified).

It turns out that stochastic stability is a rather common property of Sinai–Ruelle–Bowen measures. It holds not only for uniformly expanding maps or hyperbolic diffeomorphisms [153,258], but also for most interval maps [25], for partially hyperbolic systems of the type $E^u \oplus E^{cs}$ or Hénon-like diffeomorphisms [6,13,33,40]. We refer to the monographs [21,41] for more background and results.

Untreated Topics

For reasons of space and time, many important topics have been left out of this article. Let us list some of them.

Other phenomena related to chaotic dynamics have been studied: entrance times [77,79], spectral properties and dynamical zeta functions [21], escape rates [104], dimension [204], differentiability of physical measures with respect to parameters [99,223], entropies and volume or homological growth rates [118,139,254].

As far as the structure of the setting is concerned, one can go beyond maps or diffeomorphisms or flows and study: more general group actions [128]; holomorphic and meromorphic structures [72] and the references therein; symplectic or volume-preserving [90,137] and in particular the Pugh–Shub program around stable ergodicity of partially hyperbolic systems [211]; random iterations [17,154,155].

A number of important problems have motivated the study of special forms of chaotic dynamics: equidistribution in number theory [105,109,138] and geometry [236]; quantum chaos [10,125]; chaotic control [227]; analysis of algorithms [245].

We have also omitted the important problem of *applying the above results*. Perhaps because of the lack of a general theory, this can often be a challenge (see for instance [244] for the already complex problem of verifying *uniform hyperbolicity* for a singular flow). Liverani has shown how theoretical results can lead to precise and efficient estimates for the toy model of piecewise expanding interval maps [176]. Ergodic theory implies that, in some settings at least, *adding noise may make some estimates more precise* (see [157]). We refer to [106] and the reference therein.

Future Directions

We conclude this article by a (very partial) selection of open problems.

General Theory

In dimension 1, we have seen that the analogue of the Palis conjecture (see above) is established (Theorem 2). However the description of the typical dynamics in Kolmogorov sense is only known in the unimodal, non-degenerate case by Theorem 3. Indeed, results like [63] suggest that the multimodal picture could be more complex.

In higher dimensions, our understanding is much more limited. As far as a general theory is concerned, a deep problem is the paucity of results on the generic dynamics in C^r smoothness with $r > 1$. The remarkable current progress in generic dynamics (culminating in the proof of the weak Palis conjecture, see [84], the references therein and [51] for background) seems restricted to the C^1 topology because of the lack of fundamental tools (e. g., closing lemmas) in higher smoothness. But Pesin theory requires higher smoothness at least technically. This is not only a hard technical issue but generic properties of physical measures, when they have been analyzed are often completely different between the C^1 case and higher smoothness [45].

Physical Measures

In higher dimensions, Benedicks and Carleson analysis of the Hénon map has given rise to a rather general theory of Hénon-like maps and more generally of the dynamical phenomena associated to homoclinic tangencies. However, the proofs are extremely technical. Could they be simplified? Current attempts like [266] center on the introduction of a simpler notion of critical points, possibly a non-inductive one [212].

Can this Hénon theory be extended to the weakly dissipative situation? to the conservative situation (for which the *standard map* is a well-known example defying analysis)? In the strongly dissipative setting, what are the typical phenomena on the complement of the Benedicks–Carleson set of parameters?

From a global perspective one of the main questions is the following:

> *Can infinitely many sinks coexist for a large set of parameters in a typical family or is Newhouse phenomenon atypical in the Kolmogorov or prevalent sense?*

This seems rather unlikely (see however [11]).

Away from such "critical dynamics", there are many results about systems with dominated splitting satisfying additional conditions. Can these conditions be weakened so they would be satisfied by typical systems satisfying some natural conditions (like robust transitivity)? For instance:

> *Could one analyze the physical measures of volume-hyperbolic systems?*

A more specific question is whether Tsujii's striking analysis of surface maps with one uniformly expanding direction can be extended to higher dimensions? can one weaken the uniformity of the expansion? The same questions for the corresponding invertible situation is considered in [83].

Maximum Entropy Measures and Topological Complexity

As we explained, C^∞ smoothness, by a Newhouse theorem, ensures the existence of maximum entropy measures, making the situation a little simpler than with respect to physical measures. This existence results allow in particular an easy formulation of the problem of the typicality of hyperbolicity:

> *Are maximum entropy ergodic measures of systems with positive entropy hyperbolic for most systems?*

A more difficult problem is that of the finite multiplicity of the maximum entropy measures. For instance:

> *Do typical systems possess finitely many maximum entropy ergodic measures?*

More specifically, can one prove intrinsic ergodicity (ie, uniqueness of the measure of maximum entropy) for an isolated homoclinic class of some diffeomorphisms (perhaps C^1-generic)? Can a generic C^1-diffeomorphism carry an infinite number of homoclinic classes, each with topological entropy bounded away from zero?

A perhaps more tractable question, given the recent progress in this area: Is a C^1-generic partially hyperbolic diffeomorphisms, perhaps with central dimension 1 or 2, intrinsically ergodic?

We have seen how uniform systems have simple symbolic dynamics, i. e., subshifts of finite type, and how interval maps and more generally entropy-expanding maps keep some of this simplicity, defining subshifts or puzzles of quasi-finite type [68,69]. [264] have defined symbolic dynamics for topological Hénon-like map which seems close to that of that of a one-dimensional system.

Can one describe Wang and Young symbolic dynamics of Hénon-like attractors and fit it in a class in which uniqueness of the maximum entropy measure could be proved?

More generally, can one define nice combinatorial descriptions, for surface diffeomorphisms? Can one formulate variants of the entropy-expansion condition [*For instance building on our "entropy-hyperbolicity"*.] of [66,70], that would be satisfied by a large subset of the diffeomorphisms?

Another possible approach is illustrated by the pruning front conjecture of [87] (see also [88,144]). It is an attempt to build a combinatorial description by trying to generalize the way that, for interval maps, kneading invariants determine the symbolic dynamics by considering the bifurcations from a trivial dynamics to an arbitrary one.

We hope that our reader has shared in our fascination with this subject, the many surprising and even paradoxical discoveries that have been made and the exciting current progress, despite the very real difficulties both in the analysis of such non-uniform systems as the Henon map and in the attemps to establish a general (and practical) ergodic theory of chaotic dynamical systems.

Acknowledgments

I am grateful for the advice and/or comments of the following colleagues: P. Collet, J.-R. Chazottes, and especially S. Ruette. I am also indebted to the anonymous referee.

Bibliography

Primary Literature

1. Aaronson J, Denker M (2001) Local limit theorems for partial sums of stationary sequences generated by Gibbs–Markov maps. Stoch Dyn 1:193–237
2. Abdenur F, Bonatti C, Crovisier S (2006) Global dominated splittings and the C^1 Newhouse phenomenon. Proc Amer Math Soc 134(8):2229–2237
3. Abraham R, Smale S (1970) Nongenericity of Ω-stability. In: Global Analysis, vol XIV. Proc Sympos Pure Math. Amer Math Soc
4. Alves JF (2000) SRB measures for non-hyperbolic systems with multidimensional expansion. Ann Sci Ecole Norm Sup 33(4):1–32
5. Alves JF (2006) A survey of recent results on some statistical features of non-uniformly expanding maps. Discret Contin Dyn Syst 15:1–20
6. Alves JF, Araujo V (2003) Random perturbations of nonuniformly expanding maps. Geometric methods in dynamics. I Astérisque 286:25–62
7. Alves JF, Viana M (2002) Statistical stability for robust classes of maps with non-uniform expansion. Ergod Theory Dynam Syst 22:1–32
8. Alves JF, Bonatti C, Viana M (2000) SRB measures for partially hyperbolic systems whose central direction is mostly expanding. Invent Math 140:351–398
9. Anantharaman N (2004) On the zero-temperature or vanishing viscosity limit for certain Markov processes arising from Lagrangian dynamics. J Eur Math Soc (JEMS) 6:207–276
10. Anantharaman N, Nonnenmacher S (2007) Half delocalization of the eigenfunctions for the Laplacian on an Anosov manifold. Ann Inst Fourier 57:2465–2523
11. Araujo V (2001) Infinitely many stochastically stable attractors. Nonlinearity 14:583–596
12. Araujo V, Pacifico MJ (2006) Large deviations for non-uniformly expanding maps. J Stat Phys 125:415–457
13. Araujo V, Tahzibi A (2005) Stochastic stability at the boundary of expanding maps. Nonlinearity 18:939–958
14. Arnaud MC (1998) Le "closing lemma" en topologie C^1. Mem Soc Math Fr (NS) 74
15. Arnol'd VI (1988) Geometrical methods in the theory of ordinary differential equations, 2nd edn. Grundlehren der Mathematischen Wissenschaften, 250. Springer, New York
16. Arnol'd VI, Avez A (1968) Ergodic problems of classical mechanics. W.A. Benjamin, New York
17. Arnold L (1998) Random dynamical systems. Springer Monographs in Mathematics. Springer, Berlin
18. Avila A (2007) personal communication
19. Avila A, Moreira CG (2005) Statistical properties of unimodal maps: the quadratic family. Ann of Math 161(2):831–881
20. Avila A, Lyubich M, de Melo W (2003) Regular or stochastic dynamics in real analytic families of unimodal maps. Invent Math 154:451–550
21. Baladi V (2000) Positive transfer operators and decay of correlations. In: Advanced Series in Nonlinear Dynamics, 16. World Scientific Publishing Co, Inc, River Edge, NJ
22. Baladi V, Gouëzel S (2008) Good Banach spaces for piecewise hyperbolic maps via interpolation. preprint arXiv:0711.1960 available from http://www.arxiv.org
23. Baladi V, Ruelle D (1996) Sharp determinants. Invent Math 123:553–574
24. Baladi V, Tsujii M (2007) Anisotropic Hölder and Sobolev spaces for hyperbolic diffeomorphisms. Ann Inst Fourier (Grenoble) 57:127–154
25. Baladi V, Young LS (1993) On the spectra of randomly perturbed expanding maps. Comm Math Phys 156:355–385. Erratum: Comm Math Phys 166(1):219–220, 1994
26. Baladi V, Kondah A, Schmitt B (1996) Random correlations for small perturbations of expanding maps, English summary. Random Comput Dynam 4:179–204
27. Baraviera AT, Bonatti C (2003) Removing zero Lyapunov exponents, English summary. Ergod Theory Dynam Syst 23:1655–1670
28. Barreira L, Schmeling J (2000) Sets of "non-typical" points have full topological entropy and full Hausdorff dimension. Isr J Math 116:29–70
29. Barreira L, Pesin Ya, Schmeling J (1999) Dimension and product structure of hyperbolic measures. Ann of Math 149(2):755–783
30. Benedicks M, Carleson L (1985) On iterations of $1 - ax^2$ on $(-1, 1)$. Ann of Math (2) 122(1):1–25
31. Benedicks M, Carleson L (1991) The dynamics of the Hénon map. Ann of Math 133(2):73–169

32. Benedicks M, Viana M (2001) Solution of the basin problem for Hénon-like attractors. Invent Math 143:375–434
33. Benedicks M, Viana M (2006) Random perturbations and statistical properties of Hénon-like maps. Ann Inst H Poincaré Anal Non Linéaire 23:713–752
34. Benedicks M, Young LS (1993) Sinaĭ–Bowen–Ruelle measures for certain Hénon maps. Invent Math 112:541–576
35. Benedicks M, Young LS (2000) Markov extensions and decay of correlations for certain Hénon maps. Géométrie complexe et systèmes dynamiques, Orsay, 1995. Astérisque 261: 13–56
36. Bergelson V (2006) Ergodic Ramsey theory: a dynamical approach to static theorems. In: International Congress of Mathematicians, vol II, Eur Math Soc, Zürich, pp 1655–1678
37. Berkes I, Csáki E (2001) A universal result in almost sure central limit theory. Stoch Process Appl 94(1):105–134
38. Bishop E (1967/1968) A constructive ergodic theorem. J Math Mech 17:631–639
39. Blanchard F, Glasner E, Kolyada S, Maass A (2002) On Li-Yorke pairs, English summary. J Reine Angew Math 547:51–68
40. Blank M (1989) Small perturbations of chaotic dynamical systems. (Russian) Uspekhi Mat Nauk 44:3–28, 203. Translation in: Russ Math Surv 44:1–33
41. Blank M (1997) Discreteness and continuity in problems of chaotic dynamics. In: Translations of Mathematical Monographs, 161. American Mathematical Society, Providence
42. Blank M, Keller G (1997) Stochastic stability versus localization in one-dimensional chaotic dynamical systems. Nonlinearity 10:81–107
43. Blank M, Keller G, Liverani C (2002) Ruelle–Perron–Frobenius spectrum for Anosov maps. Nonlinearity 15:1905–1973
44. Bochi J (2002) Genericity of zero Lyapunov exponents. Ergod Theory Dynam Syst 22:1667–1696
45. Bochi J, Viana M (2005) The Lyapunov exponents of generic volume-preserving and symplectic maps. Ann of Math 161(2):1423–1485
46. Bolsinov AV, Taimanov IA (2000) Integrable geodesic flows with positive topological entropy. Invent Math 140:639–650
47. Bonano C, Collet P (2006) Complexity for extended dynamical systems. arXiv:math/0609681
48. Bonatti C, Crovisier S (2004) Recurrence et genericite. Invent Math 158(1):33–104
49. Bonatti C, Viana M (2000) SRB measures for partially hyperbolic systems whose central direction is mostly contracting. Isr J Math 115:157–193
50. Bonatti C, Viana M (2004) Lyapunov exponents with multiplicity 1 for deterministic products of matrices. Ergod Theory Dynam Syst 24(5):1295–1330
51. Bonatti C, Diaz L, Viana M (2005) Dynamics beyond uniform hyperbolicity. A global geometric and probabilistic perspective. In: Mathematical Physics, III. Encyclopedia of Mathematical Sciences, 102. Springer, Berlin
52. Bowen R (1975) Equilibrium states and the ergodic theory of Anosov diffeomorphisms. In: Lecture Notes in Mathematics, vol 470. Springer, Berlin–New York
53. Bowen R (1979) Hausdorff dimension of quasi-circles. Inst Hautes Études Sci Publ Math 50:11–25
54. Bowen R, Ruelle D (1975) Ergodic theory of Axiom A flows. Invent Math 29:181–202
55. Boyle M, Downarowicz T (2004) The entropy theory of symbolic extensions. Invent Math 156:119–161
56. Boyle M, Fiebig D, Fiebig U (2002) Residual entropy, conditional entropy and subshift covers. Forum Math 14:713–757
57. Boyle M, Buzzi J, Gomez R (2006) Ricardo Almost isomorphism for countable state Markov shifts. J Reine Angew Math 592:23–47
58. Brain M, Berger A (2001) Chaos and chance. In: An introduction to stochastic aspects of dynamics. Walter de Gruyter, Berlin
59. Brin M (2001) Appendix A. In: Barreira L, Pesin Y (eds) Lectures on Lyapunov exponents and smooth ergodic theory. Proc Sympos Pure Math, 69, Smooth ergodic theory and its applications, Seattle, WA, 1999, pp 3–106. Amer Math Soc, Providence, RI
60. Brin M, Stuck G (2002) Introduction to dynamical systems. Cambridge University Press, Cambridge
61. Broise A (1996) Transformations dilatantes de l'intervalle et theoremes limites. Etudes spectrales d'operateurs de transfert et applications. Asterisque 238:1–109
62. Broise A (1996) Transformations dilatantes de l'intervalle et théorèmes limites. Astérisque 238:1–109
63. Bruin H, Keller G, Nowicki T, van Strien S (1996) Wild Cantor attractors exist. Ann of Math 143(2):97–130
64. Buzzi J (1997) Intrinsic ergodicity of smooth interval maps. Isr J Math 100:125–161
65. Buzzi J (2000) Absolutely continuous invariant probability measures for arbitrary expanding piecewise $\boldsymbol{R}$-analytic mappings of the plane. Ergod Theory Dynam Syst 20:697–708
66. Buzzi J (2000) On entropy-expanding maps. Unpublished
67. Buzzi J (2001) No or infinitely many a.c.i.p. for piecewise expanding C^r maps in higher dimensions. Comm Math Phys 222:495–501
68. Buzzi J (2005) Subshifts of quasi-finite type. Invent Math 159:369–406
69. Buzzi J (2006) Puzzles of Quasi-Finite Type, Zeta Functions and Symbolic Dynamics for Multi-Dimensional Maps. arXiv:math/0610911
70. Buzzi J () Hyperbolicity through Entropies. Saint-Flour Summer probability school lecture notes (to appear)
71. Bálint P, Gouëzel S (2006) Limit theorems in the stadium billiard. Comm Math Phys 263:461–512
72. Carleson L, Gamelin TW (1993) Complex dynamics. In: Universitext: Tracts in Mathematics. Springer, New York
73. Chazottes JR, Gouëzel S (2007) On almost-sure versions of classical limit theorems for dynamical systems. Probab Theory Relat Fields 138:195–234
74. Chernov N (1999) Statistical properties of piecewise smooth hyperbolic systems in high dimensions. Discret Contin Dynam Syst 5(2):425–448
75. Chernov N, Markarian R, Troubetzkoy S (2000) Invariant measures for Anosov maps with small holes. Ergod Theory Dynam Syst 20:1007–1044
76. Christensen JRR (1972) On sets of Haar measure zero in abelian Polish groups. Isr J Math 13:255–260
77. Collet P (1996) Some ergodic properties of maps of the interval. In: Dynamical systems, Temuco, 1991/1992. Travaux en Cours, 52. Hermann, Paris, pp 55–91
78. Collet P, Eckmann JP (2006) Concepts and results in chaotic dynamics: a short course. Theoretical and Mathematical Physics. Springer, Berlin
79. Collet P, Galves A (1995) Asymptotic distribution of entrance times for expanding maps of the interval. Dynamical systems

and applications, pp 139–152. World Sci Ser Appl Anal 4, World Sci Publ, River Edge, NJ
80. Collet P, Courbage M, Mertens S, Neishtadt A, Zaslavsky G (eds) (2005) Chaotic dynamics and transport in classical and quantum systems. Proceedings of the International Summer School of the NATO Advanced Study Institute held in Cargese, August 18–30, 2003. Edited by NATO Science Series II: Mathematics, Physics and Chemistry, 182. Kluwer, Dordrecht
81. Contreras G, Lopes AO, Thieullen Ph (2001) Lyapunov minimizing measures for expanding maps of the circle, English summary. Ergod Theory Dynam Syst 21:1379–1409
82. Cowieson WJ (2002) Absolutely continuous invariant measures for most piecewise smooth expanding maps. Ergod Theory Dynam Syst 22:1061–1078
83. Cowieson WJ, Young LS (2005) SRB measures as zero-noise limits. Ergod Theory Dynam Syst 25:1115–1138
84. Crovisier S (2006) Birth of homoclinic intersections: a model for the central dynamics of partially hyperbolic systems. http://www.arxiv.org/abs/math/0605387
85. Crovisier S (2006) Periodic orbits and chain-transitive sets of C^1-diffeomorphisms. Publ Math Inst Hautes Etudes Sci 104:87–141
86. Crovisier S (2006) Perturbation of C^1-diffeomorphisms and generic conservative dynamics on surfaces. Dynamique des diffeomorphismes conservatifs des surfaces: un point de vue topologique, 21, Panor Syntheses. Soc Math France, Paris, pp 1–33
87. Cvitanović P, Gunaratne GH, Procaccia I (1988) Topological and metric properties of Henon-type strange attractors. Phys Rev A 38(3):1503–1520
88. de Carvalho A, Hall T (2002) How to prune a horseshoe. Nonlinearity 15:R19-R68
89. de Castro A (2002) Backward inducing and exponential decay of correlations for partially hyperbolic attractors. Isr J Math 130:29–75
90. de la Llave R (2001) A tutorial on KAM theory. In: Smooth ergodic theory and its applications, Seattle, WA, 1999, pp 175–292. Proc Sympos Pure Math, 69. Amer Math Soc, Providence
91. de Melo W, van Strien S (1993) One-dimensional dynamics. In: Ergebnisse der Mathematik und ihrer Grenzgebiete (3), 25. Springer, Berlin
92. Denker M, Philipp W (1984) Approximation by Brownian motion for Gibbs measures and flows under a function. Ergod Theory Dyn Syst 4:541–552
93. Detmers MF, Liverani C (2007) Stability of Statistical Properties in Two-dimensional Piecewise Hyperbolic Maps. Trans Amer Math Soc 360:4777–4814
94. Devaney RL (1989) An introduction to chaotic dynamical systems, 2nd ed. Addison–Wesley, Redwood
95. Diaz L, Rocha J, Viana M (1996) Strange attractors in saddle-node cycles: prevalence and globality. Invent Math 125: 37–74
96. Dolgopyat D (2000) On dynamics of mostly contracting diffeomorphisms. Comm Math Phys 213:181–201
97. Dolgopyat D (2002) On mixing properties of compact group extensions of hyperbolic systems. Isr J Math 130:157–205
98. Dolgopyat D (2004) Limit theorems for partially hyperbolic systems. Trans Amer Math Soc 356:1637–1689
99. Dolgopyat D (2004) On differentiability of SRB states for partially hyperbolic systems. Invent Math 155:389–449
100. Downarowicz T (2005) Entropy structure. J Anal Math 96: 57–116
101. Downarowicz T, Newhouse S (2005) Symbolic extensions and smooth dynamical systems. Invent Math 160:453–499
102. Downarowicz T, Serafin J (2003) Possible entropy functions. Isr J Math 135:221–250
103. Duhem P (1906) La théorie physique, son objet et sa structure. Vrin, Paris 1981
104. Eckmann JP, Ruelle D (1985) Ergodic theory of chaos and strange attractors. Rev Mod Phys 57:617–656
105. Eskin A, McMullen C (1993) Mixing, counting, and equidistribution in Lie groups. Duke Math J 71:181–209
106. Fiedler B (ed) (2001) Ergodic theory, analysis, and efficient simulation of dynamical systems. Springer, Berlin
107. Fiedler B (ed) (2002) Handbook of dynamical systems, vol 2. North-Holland, Amsterdam
108. Fisher A, Lopes AO (2001) Exact bounds for the polynomial decay of correlation, $1/f$ noise and the CLT for the equilibrium state of a non-Hölder potential. Nonlinearity 14:1071–1104
109. Furstenberg H (1981) Recurrence in ergodic theory and combinatorial number theory. In: MB Porter Lectures. Princeton University Press, Princeton, NJ
110. Gallavotti G (1999) Statistical mechanics. In: A short treatise. Texts and Monographs in Physics. Springer, Berlin
111. Gatzouras D, Peres Y (1997) Invariant measures of full dimension for some expanding maps. Ergod Theory Dynam Syst 17:147–167
112. Glasner E, Weiss B (1993) Sensitive dependence on initial conditions. Nonlinearity 6:1067–1075
113. Gordin MI (1969) The central limit theorem for stationary processes. (Russian) Dokl Akad Nauk SSSR 188:739–741
114. Gouëzel S (2004) Central limit theorem and stable laws for intermittent maps. Probab Theory Relat Fields 128:82–122
115. Gouëzel S (2005) Berry–Esseen theorem and local limit theorem for non uniformly expanding maps. Ann Inst H Poincaré 41:997–1024
116. Gouëzel S, Liverani C (2006) Banach spaces adapted to Anosov systems. Ergod Theory Dyn Syst 26:189–217
117. Graczyk J, Świątek G (1997) Generic hyperbolicity in the logistic family. Ann of Math 146(2):1–52
118. Gromov M (1987) Entropy, homology and semialgebraic geometry. Séminaire Bourbaki, vol 1985/86. Astérisque No 145–146:225–240
119. Guivarc'h Y, Hardy J (1998) Théorèmes limites pour une classe de chaînes de Markov et applications aux difféomorphismes d'Anosov. Ann Inst H Poincaré 24:73:98
120. Guckenheimer J (1979) Sensitive dependence on initial conditions for one-dimensional maps. Commun Math Phys 70:133–160
121. Guckenheimer J, Holmes P (1990) Nonlinear oscillations, dynamical systems, and bifurcations of vector fields. Revised and corrected reprint of the 1983 original. In: Applied Mathematical Sciences, 42. Springer, New York
122. Guckenheimer J, Wechselberger M, Young LS (2006) Chaotic attractors of relaxation oscillators. Nonlinearity 19:701–720
123. Gurevic BM (1996) Stably recurrent nonnegative matrices. (Russian) Uspekhi Mat Nauk 51(3)(309):195–196; translation in: Russ Math Surv 51(3):551–552
124. Gurevic BM, Savchenko SV (1998) Thermodynamic formalism for symbolic Markov chains with a countable number of

states. (Russian) Uspekhi Mat Nauk 53(2)(320):3–106; translation in: Russ Math Surv 53(2):245–344
125. Gutzwiller M (1990) Chaos in classical and quantum mechanics. In: Interdisciplinary Applied Mathematics, 1. Springer, New York
126. Hadamard J (1898) Les surfaces à courbures opposees et leurs lignes geodesiques. J Math Pures Appl 4:27–73
127. Hasselblatt B, Katok A (2003) A first course in dynamics. In: With a panorama of recent developments. Cambridge University Press, New York
128. Hasselblatt B, Katok A (eds) (2002,2006) Handbook of dynamical systems, vol 1A and 1B. Elsevier, Amsterdam
129. Hennion H (1993) Sur un théorème spectral et son application aux noyaux lipchitziens. Proc Amer Math Soc 118:627–634
130. Hénon M (1976) A two-dimensional mapping with a strange attractor. Comm Math Phys 50:69–77
131. Hesiod (1987) Theogony. Focus/R. Pullins, Newburyport
132. Hochman M (2006) Upcrossing inequalities for stationary sequences and applications. arXiv:math/0608311
133. Hofbauer F (1979) On intrinsic ergodicity of piecewise monotonic transformations with positive entropy. Isr J Math 34(3):213–237
134. Hofbauer F (1985) Periodic points for piecewise monotonic transformations. Ergod Theory Dynam Syst 5:237–256
135. Hofbauer F, Keller G (1982) Ergodic properties of invariant measures for piecewise monotonic transformations. Math Z 180:119–140
136. Hofbauer F, Keller G (1990) Quadratic maps without asymptotic measure. Comm Math Phys 127:319–337
137. Hofer H, Zehnder E (1994) Symplectic invariants and Hamiltonian dynamics. Birkhäuser, Basel
138. Host B, Kra B (2005) Nonconventional ergodic averages and nilmanifolds. Ann of Math 161(2):397–488
139. Hua Y, Saghin R, Xia Z (2006) Topological Entropy and Partially Hyperbolic Diffeomorphisms. arXiv:math/0608720
140. Hunt FY (1998) Unique ergodicity and the approximation of attractors and their invariant measures using Ulam's method. (English summary) Nonlinearity 11:307–317
141. Hunt BR, Sauer T, Yorke JA (1992) Prevalence: a translation-invariant "almost every" on infinite-dimensional spaces. Bull Amer Math Soc (NS) 27:217–238
142. Ibragimov I, Linnik Yu, Kingman JFC (ed) (trans) (1971) Independent and stationary sequences of random variables. Wolters-Noordhoff, Groningen
143. Ionescu Tulcea CT, Marinescu G (1950) Théorie ergodique pour des classes d'opérations non complètement continues. (French) Ann of Math 52(2):140–147
144. Ishii Y (1997) Towards a kneading theory for Lozi mappings. I. A solution of the pruning front conjecture and the first tangency problem. Nonlinearity 10:731–747
145. Jakobson MV (1981) Absolutely continuous invariant measures for one-parameter families of one-dimensional maps. Comm Math Phys 81:39–88
146. Jenkinson O (2007) Optimization and majorization of invariant measures. Electron Res Announc Amer Math Soc 13:1–12
147. Katok A (1980) Lyapunov exponents, entropy and periodic orbits for diffeomorphisms. Inst Hautes Etudes Sci Publ Math No 51:137–173
148. Katok A, Hasselblatt B (1995) Introduction to the modern theory of dynamical systems. In: Encyclopedia of Mathematics and its Applications, 54. With a supplementary chapter by: Katok A, Mendoza L. Cambridge University Press, Cambridge
149. Keller G (1984) On the rate of convergence to equilibrium in one-dimensional systems. Comm Math Phys 96:181–193
150. Keller G (1990) Exponents, attractors and Hopf decompositions for interval maps. Ergod Theory Dynam Syst 10:717–744
151. Keller G, Liverani C (1999) Stability of the spectrum for transfer operators. Ann Scuola Norm Sup Pisa Cl Sci 28(4):141–152
152. Keller G, Nowicki T (1992) Spectral theory, zeta functions and the distribution of periodic points for Collet-Eckmann maps. Comm Math Phys 149:31–69
153. Kifer Yu (1977) Small random perturbations of hyperbolic limit sets. (Russian) Uspehi Mat Nauk 32(1)(193):193–194
154. Kifer Y (1986) Ergodic theory of random transformations. In: Progress in Probability and Statistics, 10. Birkhäuser Boston, Inc, Boston, MA
155. Kifer Y (1988) Random perturbations of dynamical systems. In: Progress in Probability and Statistics, 16. Birkhäuser Boston, Inc, Boston, MA
156. Kifer Yu (1990) Large deviations in dynamical systems and stochastic processes. Trans Amer Math Soc 321:505–524
157. Kifer Y (1997) Computations in dynamical systems via random perturbations. (English summary) Discret Contin Dynam Syst 3:457–476
158. Kolyada SF (2004) LI-Yorke sensitivity and other concepts of chaos. Ukr Math J 56:1242–1257
159. Kozlovski OS (2003) Axiom A maps are dense in the space of unimodal maps in the C^k topology. Ann of Math 157(2):1–43
160. Kozlovski O, Shen W, van Strien S (2007) Density of Axiom A in dimension one. Ann Math 166:145–182
161. Krengel U (1983) Ergodic Theorems. De Gruyter, Berlin
162. Krzyżewski K, Szlenk W (1969) On invariant measures for expanding differentiable mappings. Studia Math 33:83–92
163. Lacroix Y (2002) Possible limit laws for entrance times of an ergodic aperiodic dynamical system. Isr J Math 132:253–263
164. Ladler RL, Marcus B (1979) Topological entropy and equivalence of dynamical systems. Mem Amer Math Soc 20(219)
165. Lastoa A, Yorke J (1973) On the existence of invariant measures for piecewise monotonic transformations. Trans Amer Math Soc 186:481–488
166. Ledrappier F (1984) Proprietes ergodiques des mesures de Sinaï. Inst Hautes Etudes Sci Publ Math 59:163–188
167. Ledrappier F, Young LS (1985) The metric entropy of diffeomorphisms. I Characterization of measures satisfying Pesin's entropy formula. II Relations between entropy, exponents and dimension. Ann of Math 122(2):509–539, 540–574
168. Leplaideur R (2004) Existence of SRB-measures for some topologically hyperbolic diffeomorphisms. Ergod Theory Dynam Syst 24:1199–1225
169. Li TY, Yorke JA (1975) Period three implies chaos. Amer Math Monthly 82:985–992
170. Li M, Vitanyi P (1997) An introduction to Kolmogorov complexity and its applications. In: Graduate Texts in Computer Science, 2nd edn. Springer, New York
171. Lind D, Marcus B (1995) An introduction to symbolic dynamics and coding. Cambridge University Press, Cambridge
172. Liu PD, Qian M, Zhao Y (2003) Large deviations in Axiom A endomorphisms. Proc Roy Soc Edinb Sect A 133:1379–1388
173. Liverani C (1995) Decay of Correlations. Ann Math 142: 239–301

174. Liverani C (1995) Decay of Correlations in Piecewise Expanding maps. J Stat Phys 78:1111–1129
175. Liverani C (1996) Central limit theorem for deterministic systems. In: Ledrappier F, Lewowicz J, Newhouse S (eds) International conference on dynamical systems, Montevideo, 1995. Pitman Research Notes. In: Math 362:56–75
176. Liverani C (2001) Rigorous numerical investigation of the statistical properties of piecewise expanding maps – A feasibility study. Nonlinearity 14:463–490
177. Liverani C, Tsujii M (2006) Zeta functions and dynamical systems. (English summary) Nonlinearity 19:2467–2473
178. Lyubich M (1997) Dynamics of quadratic polynomials. I, II Acta Math 178:185–247, 247–297
179. Margulis G (2004) On some aspects of the theory of Anosov systems. In: With a survey by Richard Sharp: Periodic orbits of hyperbolic flows. Springer Monographs in Mathematics. Springer, Berlin
180. Mañé R (1982) An ergodic closing lemma. Ann of Math 116(2):503–540
181. Mañé R (1985) Hyperbolicity, sinks and measures in one-dimensional dynamics. Commun Math Phys 100:495–524
182. Mañé R (1988) A proof of the C^1 stability conjecture. Publ Math IHES 66:161–210
183. Melbourne I, Nicol M (2005) Almost sure invariance principle for nonuniformly hyperbolic systems. Comm Math Phys 260(1):131–146
184. Milnor J, Thurston W (1988) On iterated maps of the interval. Dynamical systems, College Park, MD, 1986–87, pp 465–563. Lecture Notes in Math, 1342. Springer, Berlin
185. Misiurewicz M (1976) A short proof of the variational principle for a Z_+^n action on a compact space. Bull Acad Polon Sci Sér Sci Math Astronom Phys 24(12)1069–1075
186. Misiurewicz M (1976) Topological conditional entropy. Studia Math 55:175–200
187. Misiurewicz M (1979) Horseshoes for mappings of the interval. Bull Acad Polon Sci Sér Sci Math 27:167–169
188. Misiurewicz M (1995) Continuity of entropy revisited. In: Dynamical systems and applications, pp 495–503. World Sci Ser Appl Anal 4. World Sci Publ, River Edge, NJ
189. Misiurewicz M, Smítal J (1988) Smooth chaotic maps with zero topological entropy. Ergod Theory Dynam Syst 8:421–424
190. Mora L, Viana M (1993) Abundance of strange attractors. Acta Math 171:1–71
191. Moreira CG, Palis J, Viana M (2001) Homoclinic tangencies and fractal invariants in arbitrary dimension. CRAS 333:475–480
192. Murray JD (2002,2003) Mathematical biology. I and II An introduction, 3rd edn. In: Interdisciplinary Applied Mathematics, 17 and 18. Springer, New York
193. Nagaev SV (1957) Some limit theorems for stationary Markov chains. Theor Probab Appl 2:378–406
194. Newhouse SE (1974) Diffeomorphisms with infinitely many sinks. Topology 13:9–18
195. Newhouse SE (1989) Continuity properties of entropy. Ann of Math 129(2):215–235; Erratum: Ann of Math 131(2):409–410
196. Nussbaum RD (1970) The radius of the essential spectrum. Duke Math J 37:473–478
197. Ornstein D (1970) Bernoulli shifts with the same entropy are isomorphic. Adv Math 4:337–352
198. Ornstein D, Weiss B (1988) On the Bernoulli nature of systems with some hyperbolic structure. Ergod Theory Dynam Syst 18:441–456
199. Ovid (2005) Metamorphosis. W.W. Norton, New York
200. Palis J (2000) A global view of dynamics and a conjecture on the denseness of tinitude of attractors. Asterisque 261: 335–347
201. Palis J, Yoccoz JC (2001) Fers a cheval non-uniformement hyperboliques engendres par une bifurcation homocline et densite nulle des attracteurs. CRAS 333:867–871
202. Pesin YB (1976) Families of invariant manifolds corresponding to non-zero characteristic exponents. Math USSR Izv 10:1261–1302
203. Pesin YB (1977) Characteristic exponents and smooth ergodic theory. Russ Math Surv 324:55–114
204. Pesin Ya (1997) Dimension theory in dynamical systems. In: Contemporary views and applications. Chicago Lectures in Mathematics. University of Chicago Press, Chicago
205. Pesin Ya, Sinaĭ Ya (1982) Gibbs measures for partially hyperbolic attractors. Ergod Theory Dynam Syst 2:417–438
206. Piorek J (1985) On the generic chaos in dynamical systems. Univ Iagel Acta Math 25:293–298
207. Plykin RV (2002) On the problem of the topological classification of strange attractors of dynamical systems. Uspekhi Mat Nauk 57:123–166. Translation in: Russ Math Surv 57: 1163–1205
208. Poincare H (1892) Les methodes nouvelles de la mecanique céleste. Paris, Gauthier–Villars
209. Pollicott M, Sharp R (2002) Invariance principles for interval maps with an indifferent fixed point. Comm Math Phys 229: 337–346
210. Przytycki F (1977) On *U*-stability and structural stability of endomorphisms satisfying. Axiom A Studia Math 60:61–77
211. Pugh C, Shub M (1999) Ergodic attractors. Trans Amer Math Soc 312:1–54
212. Pujals E, Rodriguez–Hertz F (2007) Critical points for surface diffeomorphisms. J Mod Dyn 1:615–648
213. Puu T (2000) Attractors, bifurcations, and chaos. In: Nonlinear phenomena in economics. Springer, Berlin
214. Rees M (1981) A minimal positive entropy homeomorphism of the 2-torus. J London Math Soc 23(2):537–550
215. Robinson RC (2004) An introduction to dynamical systems: continuous and discrete. Pearson Prentice Hall, Upper Saddle River
216. Rousseau–Egele J (1983) Un théorème de la limite locale pour une classe de transformations dilatantes et monotones par morceaux. Ann Probab 11:772–788
217. Ruelle D (1968) Statistical mechanics of a one-dimensional lattice gas. Comm Math Phys 9:267–278
218. Ruelle D (1976) A measure associated with axiom-A attractors. Amer J Math 98:619–654
219. Ruelle D (1978) An inequality for the entropy of differentiable maps. Bol Soc Brasil Mat 9:83–87
220. Ruelle D (1982) Repellers for real analytic maps. Ergod Theory Dyn Syst 2:99–107
221. Ruelle D (1989) The thermodynamic formalism for expanding maps. Comm Math Phys 125:239–262
222. Ruelle D (2004) Thermodynamic formalism. In: The mathematical structures of equilibrium statistical mechanics, 2nd edn. Cambridge Mathematical Library, Cambridge University Press, Cambridge
223. Ruelle D (2005) Differentiating the absolutely continuous in-

variant measure of an interval map f with respect to f. Comm Math Phys 258:445–453
224. Ruette S (2003) On the Vere-Jones classification and existence of maximal measures for countable topological Markov chains. Pacific J Math 209:366–380
225. Ruette S (2003) Chaos on the interval. http://www.math.u-psud.fr/~ruette
226. Saari DG (2005) Collisions, rings, and other Newtonian N-body problems. In: CBMS Regional Conference Series in Mathematics, 104. Published for the Conference Board of the Mathematical Sciences, Washington, DC. American Mathematical Society, Providence
227. Saperstone SH, Yorke JA (1971) Controllability of linear oscillatory systems using positive controls. SIAM J Control 9: 253–262
228. Sarig O (1999) Thermodynamic formalism for countable Markov shifts. Ergod Theory Dynam Syst 19:1565–1593
229. Sarig O (2001) Phase Transitions for Countable Topological Markov Shifts. Commun Math Phys 217:555–577
230. Sataev EA (1992) Invariant measures for hyperbolic mappings with singularities. Uspekhi Mat Nauk 47:147–202. Translation in: Russ Math Surv 47:191–251
231. Saussol B (2000) Absolutely continuous invariant measures for multidimensional expanding maps. Isr J Math 116:223–48
232. Saussol B (2006) Recurrence rate in rapidly mixing dynamical systems. Discret Contin Dyn Syst 15(1):259–267
233. Shub M (1987) Global stability of dynamical systems. Springer, New York
234. Simon R (1972) A 3-dimensional Abraham-Smale example. Proc Amer Math Soc 34:629–630
235. Sinai Ya (1972) Gibbs measures in ergodic theory. Uspehi Mat Nauk 27(166):21–64
236. Starkov AN (2000) Dynamical systems on homogeneous spaces. In: Translations of Mathematical Monographs, 190. American Mathematical Society, Providence, RI
237. Stewart P (1964) Jacobellis v Ohio. US Rep 378:184
238. Szász D (ed) (2000) Hard ball systems and the Lorentz gas. Encyclopedia of Mathematical Sciences, 101. In: Mathematical Physics, II. Springer, Berlin
239. Tsujii M (1992) A measure on the space of smooth mappings and dynamical system theory. J Math Soc Jpn 44:415–425
240. Tsujii M (2000) Absolutely continuous invariant measures for piecewise real-analytic expanding maps on the plane. Comm Math Phys 208:605–622
241. Tsujii M (2000) Piecewise expanding maps on the plane with singular ergodic properties. Ergod Theory Dynam Syst 20:1851–1857
242. Tsujii M (2001) Absolutely continuous invariant measures for expanding piecewise linear maps. Invent Math 143:349–373
243. Tsujii M (2005) Physical measures for partially hyperbolic surface endomorphisms. Acta Math 194:37–132
244. Tucker W (1999) The Lorenz attractor exists. C R Acad Sci Paris Ser I Math 328:1197–1202
245. Vallée B (2006) Euclidean dynamics. Discret Contin Dyn Syst 15:281–352
246. van Strien S, Vargas E (2004) Real bounds, ergodicity and negative Schwarzian for multimodal maps. J Amer Math Soc 17: 749–782
247. Vasquez CH (2007) Statistical stability for diffeomorphisms with dominated splitting. Ergod Theory Dynam Syst 27: 253–283
248. Viana M (1993) Strange attractors in higher dimensions. Bol Soc Bras Mat (NS) 24:13–62
249. Viana M (1997) Multidimensional nonhyperbolic attractors. Inst Hautes Etudes Sci Publ Math No 85:63–96
250. Viana M (1997) Stochastic dynamics of deterministic systems. In: Lecture Notes 21st Braz Math Colloq IMPA. Rio de Janeiro
251. Viana M (1998) Dynamics: a probabilistic and geometric perspective. In: Proceedings of the International Congress of Mathematicians, vol I, Berlin, 1998. Doc Math Extra I:557–578
252. Wang Q, Young LS (2003) Strange attractors in periodically-kicked limit cycles and Hopf bifurcations. Comm Math Phys 240:509–529
253. Weiss B (2002) Single orbit dynamics. Amer Math Soc, Providence
254. Yomdin Y (1987) Volume growth and entropy. Isr J Math 57:285–300
255. Yoshihara KI (2004) Weakly dependent stochastic sequences and their applications. In: Recent Topics on Weak and Strong Limit Theorems, vol XIV. Sanseido Co Ltd, Chiyoda
256. Young LS (1982) Dimension, entropy and Lyapunov exponents. Ergod Th Dynam Syst 6:311–319
257. Young LS (1985) Bowen–Ruelle measures for certain piecewise hyperbolic maps. Trans Amer Math Soc 287:41–48
258. Young LS (1986) Stochastic stability of hyperbolic attractors. Ergod Theory Dynam Syst 6:311–319
259. Young LS (1990) Large deviations in dynamical systems. Trans Amer Math Soc 318:525–543
260. Young LS (1992) Decay of correlations for certain quadratic maps. Comm Math Phys 146:123–138
261. Young LS (1995) Ergodic theory of differentiable dynamical systems. Real and complex dynamical systems. Hillerad, 1993, pp 293–336. NATO Adv Sci Inst Ser C Math Phys Sci 464. Kluwer, Dordrecht
262. Young LS (1998) Statistical properties of dynamical systems with some hyperbolicity. Ann Math 585–650
263. Young LS (1999) Recurrence times and rates of mixing. Isr J Math 110:153–188
264. Young LS, Wang D (2001) Strange attractors with one direction of instability. Comm Math Phys 218:1–97
265. Young LS, Wang D (2006) Nonuniformly expanding 1D maps. Commun Math Phys vol 264:255–282
266. Young LS, Wang D (2008) Toward a theory of rank one attractors. Ann Math 167:349–480
267. Zhang Y (1997) Dynamical upper bounds for Hausdorff dimension of invariant sets. Ergod Theory Dynam Syst 17: 739–756

Books and Reviews

Baladi V (2000) The magnet and the butterfly: thermodynamic formalism and the ergodic theory of chaotic dynamics. (English summary). In: Development of mathematics 1950–2000, Birkhäuser, Basel, pp 97–133

Bonatti C (2003) Dynamiques generiques: hyperbolicite et transitivite. In: Seminaire Bourbaki vol 2001/2002. Asterisque No 290:225–242

Kuksin SB (2006) Randomly forced nonlinear PDEs and statistical hydrodynamics in 2 space dimensions. In: Zurich Lectures in Advanced Mathematics. European Mathematical Society (EMS), Zürich

Liu PD, Qian M (1995) Smooth ergodic theory of random dynamical systems. In: Lecture Notes in Mathematics, 1606. Springer, Berlin
Ornstein D, Weiss B (1991) Statistical properties of chaotic systems. With an appendix by David Fried. Bull Amer Math Soc (NS) 24:11–116
Young LS (1999) Ergodic theory of chaotic dynamical systems. XIIth International Congress of Mathematical Physics, ICMP '97, Brisbane, Int Press, Cambridge, MA, pp 131–143

Chaotic Behavior of Cellular Automata

Julien Cervelle[1], Alberto Dennunzio[2], Enrico Formenti[3]
[1] Laboratoire d'Informatique de l'Institut Gaspard-Monge, Université Paris-Est, Marne la Vallée, France
[2] Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Milan, Italy
[3] Laboratoire I3S, Université de Nice-Sophia Antipolis, Sophia Antipolis, France

Article Outline

Glossary
Definition of the Subject
Introduction
Definitions
The Case of Cellular Automata
Results for Linear CA: Everything Is Detectable
Future Directions
Acknowledgments
Bibliography

Glossary

Equicontinuity All points are equicontinuity points (in compact settings).
Equicontinuity point A point for which the orbits of nearby points remain close.
Expansivity From two distinct points, orbits eventually separate.
Injectivity The next state function is injective.
Linear CA A CA with additive local rule.
Regularity The set of periodic points is dense.
Sensitivity to initial conditions For any point x there exist arbitrary close points whose orbits eventually separate from the orbit of x.
Strong transitivity There always exist points which eventually move from any arbitrary neighborhood to any point.
Surjectivity The next state function is surjective.
Topological mixing There always exist points which definitely move from any arbitrary neighborhood to any other.
Transitivity There always exist points which eventually move from any arbitrary neighborhood to any other.

Definition of the Subject

A discrete time dynamical system (DTDS) is a pair $\langle X, F\rangle$ where X is a set equipped with a distance d and $F\colon X \mapsto X$ is a mapping which is continuous on X with respect to the metric d. The set X and the function F are called the state space and the next state map. At the very beginning of the seventies, the notion of chaotic behavior for DTDS has been introduced in experimental physics [46]. Successively, mathematicians started investigating this new notion finding more and more complex examples. Although a general universally accepted theory of chaos has not emerged, at least some properties are recognized as basic components of possible chaotic behavior. Among them one can list: sensitivity to initial conditions, transitivity, mixing, expansively etc. [5,6,21,22,28,29,38,41,42,58,59,60].

In the eighties, S. Wolfram started studying some of these properties in the context of cellular automata (CA) [64]. These pioneering studies opened the way to a huge amount of successive paper which aimed to complete, precise and further develop the theory of chaos in the CA context (► Dynamics of Cellular Automata in Non–compact Spaces, ► Topological Dynamics of Cellular Automata, ► Ergodic Theory of Cellular Automata, [7,8,9,10,11,13,14,15,16,17,35,43,51]). This long quest has also been stimulated by the advent of more and more powerful computers which helped researchers in their investigations. However, remark that most of the properties involved in chaos definitions turned out to be undecidable [23,34,39,40,45]. Anyway, there are non-trivial classes of CA for which these properties are decidable. For this reason in the present work we focus on linear CA.

Introduction

Cellular automata are simple formal models for complex systems. They are used in many scientific fields ranging from biology to chemistry or from physics to computer science.

A CA is made of an infinite set of finite automata distributed over a regular lattice $\mathcal{L}$. All finite automata are identical. Each automaton assumes a value, chosen from a finite set A, called the *alphabet*. A *configuration* is a snapshot of the all automata values, i. e., a function from $\mathcal{L}$ to A.

In the present chapter, we consider the D dimensional lattice $\mathcal{L} = \mathbb{Z}^D$.

A *local rule* updates the value of an automaton on the basis of its current value and the ones of a fixed set of neighboring automata which are individuated by the *neighborhood frame* $N = \{\vec{u}_1, \ldots, \vec{u}_s\} \subset \mathbb{Z}^D$. Formally, the local rule is a function $f\colon A^s \to A$.

All the automata of the lattice are updated synchronously in discrete time steps. In other words, the local rule f induces a *global rule* $F\colon A^{\mathbb{Z}^D} \to A^{\mathbb{Z}^D}$, describing the evolution of the whole system from a generic time $t \in \mathbb{N}$ to $t+1$.

When one equips the configuration space with a metric, a CA can be viewed as a DTDS $\langle A^{\mathbb{Z}^D}, F\rangle$. The state space $A^{\mathbb{Z}^D}$ of a CA is also called the configuration space $A^{\mathbb{Z}^D}$. From now on, we identify a CA with the dynamical system induced by itself or with its global rule.

As we have already told, the studies on the chaotic behavior of CA have played a central role in the last twenty years. Unfortunately, most of the properties involved are undecidable. In this chapter, we illustrate these notions by means of a particular class of CA in which they turn out to be decidable constituting a valid source for examples and understanding. Indeed, we focus on *linear* CA, i. e., CA whose local rule is a linear function on a finite group G. For clarity's sake, we assume that G is the set $Z_m = \{0, 1, \ldots, m-1\}$ of integers modulo m and $+$ the cell-wise addition.

Despite their simplicity, linear CA exhibit many of the complex features of general CA ranging from trivial to the most complicated behaviors.

Definitions

Given a DTDS $\langle X, g\rangle$ the next state function induces deterministic dynamics by its iterated application starting from a given initial state. Formally, for a fixed state $x \in X$, the *dynamical evolution* or *orbit of initial state* x is the sequence $\{F^n(x)\}_{n\in\mathbb{N}}$. A state $p \in X$ is a periodic point if there exists an integer $n > 0$ such that $F^n(p) = p$.

Deterministic Chaos

We are particularly interested in the properties which can be considered as components of "chaotic" behavior. Among them, sensitivity to initial conditions is the most intriguing, at least at a level of popular divulgement. It captures the feature that small errors in experimental measurements lead to large scale divergence in the evolution.

Definition 1 (Sensitivity) A DTDS $\langle X, F\rangle$ is *sensitive to the initial conditions* (or simply *sensitive*) if there exists a constant $\varepsilon > 0$ such that for any state $x \in X$ and any $\delta > 0$ there is a state $y \in X$ such that $d(y, x) < \delta$ and $d(F^n(y), F^n(x)) > \varepsilon$ for some $n \in \mathbb{N}$.

Intuitively, a DTDS is sensitive if for any state x there exist points arbitrarily close to x which eventually separate from x under iteration of F. Sensitivity is a strong form of instability. In fact, if a system is sensitive to the initial conditions and we are not able to measure with infinite precision the initial state, we cannot predict its dynamical evolution. This means that experimental or casual errors can lead to wrong results.

The following is another form of unpredictability.

Definition 2 (Positive Expansivity) A DTDS $\langle X, F\rangle$ is *positively expansive* if there exists a constant $\varepsilon > 0$ such that for any pair of distinct states $x, y \in X$ we have $d(F^n(y), F^n(x)) \geq \varepsilon$ for some $n \in \mathbb{N}$.

Remark that in perfect spaces (i. e., spaces without isolated points), expansive DTDS are necessarily sensitive to initial conditions.

Sensitivity alone, notwithstanding its intuitive appeal, has the drawback that, once taken as the unique condition for characterizing chaos, it appears to be neither sufficient nor as intuitive as it seems at a first glance.

Indeed, for physicists, a chaotic system must be necessarily nonlinear (the term linear in its classical meaning refers to the linearity of the equation which governs the behavior of a system). However, it is easy to find linear systems (on the reals, for instance) which are sensitive to initial conditions. Hence, a chaotic system has to satisfy further properties other than sensitivity.

Definition 3 (Transitivity) A DTDS $\langle X, F\rangle$ is (topologically) *transitive* if for all non empty open subsets U and V of X there exists a natural number n such that $F^n(U) \cap V \neq \emptyset$.

Intuitively, a transitive DTDS has points which eventually move under iteration of F from one arbitrarily small neighborhood to any other. As a consequence, the dynamical system cannot be decomposed into two disjoint clopen sets which are invariant under the iterations of F. Indecomposability is an important feature since, roughly speaking, it guarantees that the system behaves in the same way in the whole state space. Finally, note that transitivity implies surjectivity in the case of compact spaces.

Some DTDS may exhibit stronger forms of transitivity.

Definition 4 (Mixing) A DTDS $\langle X, F\rangle$ is *topologically mixing* if for all non-empty open subsets U, V of X there exists a natural number m such that for every $n \geq m$ it holds that $F^n(U) \cap V \neq \emptyset$.

The previous notion is the topological version of the well-known mixing property of ergodic theory. It is particularly useful when studying product of systems. Indeed, the product of two transitive systems is not necessarily transitive; it is transitive if one of the two systems is mixing [27].

Moreover, remark that any non-trivial (i. e. with at least two points) mixing system is sensitive to initial conditions [44].

Definition 5 (Strong Transitivity) A DTDS $\langle X, F\rangle$ is *strongly transitive* if for any nonempty open set $U \subseteq X$ we have that $\bigcup_{n=0}^{+\infty} F^n(U) = X$.

A strongly transitive map F has points which eventually move under iteration of F from one arbitrarily small neighborhood to any other *point*. As a consequence, a strongly transitive map is necessarily surjective.

Finally, remark that many well-known classes of transitive DTDS (irrational rotations of the unit circle, the tent map, the shift map, etc.) exhibit the following form of transitivity.

Definition 6 (Total Transitivity) A DTDS $\langle X, F\rangle$ is total transitive if for all integers $n > 0$, the system $\langle X, F^n\rangle$ is transitive.

Proposition 1 ([27]) *Any mixing DTDS is totally transitive.*

At the beginning of the eighties, Auslander and Yorke introduced the following definition of chaos [5].

Definition 7 (AY-Chaos) A DTDS is AY-chaotic if it is transitive and it is sensitive to the initial conditions.

This definition involves two fundamental characteristics: the undecomposability of the system, due to transitivity and the unpredictability of the dynamical evolution, due to sensitivity.

We now introduce a notion which is often referred to as an element of *regularity* a chaotic dynamical system must exhibit.

Definition 8 (DPO) A DTDS $\langle X, F\rangle$ has the *denseness of periodic points* (or, it is *regular*) if the set of its periodic points is dense in X.

The following is a standard result for compact DTDS.

Proposition 2 *If a compact DTDS has DPO then it is surjective.*

In his famous book [22], Devaney modified the AY-chaos adding the denseness of periodic points.

Definition 9 (D-Chaos) A DTDS is said to be D-chaotic if it is sensitive, transitive and regular.

An interesting result states that sensitivity, despite its popular appeal, is redundant in the Devaney definition of chaos.

Proposition 3 ([6]) *Any transitive and regular DTDS with an infinite number of states is sensitive to initial conditions.*

Note that neither transitivity nor DPO are redundant in the Devaney definition of chaos [37]. Further notions of chaos can be obtained by replacing transitivity or sensitivity with stronger properties (like expansively, strong transitivity, etc.).

Stability

All the previous properties can be considered as components of a chaotic, and then unstable, behavior for a DTDS. We now illustrate some properties concerning conditions of stability for a system.

Definition 10 (Equicontinuous Point) A state $x \in X$ of a DTDS $\langle X, F\rangle$ is an *equicontinuous point* if for any $\varepsilon > 0$ there exists $\delta > 0$ such that for all $y \in X$, $d(y, x) < \delta$ implies that $\forall n \in \mathbb{N},\ d(F^n(y), F^n(x)) < \varepsilon$.

In other words, a point x is equicontinuous (or Lyapunov stable) if for any $\varepsilon > 0$, there exists a neighborhood of x whose states have orbits which stay close to the orbit of x with distance less than ε. This is a condition of local stability for the system.

Associated with this notion involving a single state, we have two notions of global stability based on the "size" of the set of the equicontinuous points.

Definition 11 (Equicontinuity) A DTDS $\langle X, F\rangle$ is said to be *equicontinuous* if for any $\varepsilon > 0$ there exists $\delta > 0$ such that for all $x, y \in X$, $d(y, x) < \delta$ implies that $\forall n \in \mathbb{N}$, $(F^n(y), F^n(x)) < \varepsilon$.

Given a DTDS, let E be its set of equicontinuity points. Remark that if a DTDS is equicontinuous then the set E of all its equicontinuity points is the whole X. The converse is also true in the compact settings. Furthermore, if a system is sensitive then $E = \emptyset$. In general, the converse is not true [43].

Definition 12 (Almost Equicontinuity) A DTDS is *almost equicontinuous* if the set of its equicontinuous points E is residual (i. e., it can be obtained by a infinite intersection of dense open subsets).

It is obvious that equicontinuous systems are almost equicontinuous. In the sequel, almost equicontinuous systems which are not equicontinuous will be called strictly

almost equicontinuous. An important result affirms that transitive systems on compact spaces are almost equicontinuous if and only if they are not sensitive [3].

Topological Entropy

Topological entropy is another interesting property which can be taken into account in order to study the degree of chaoticity of a system. It was introduced in [2] as an invariant of topological conjugacy. The notion of topological entropy is based on the complexity of the coverings of the systems. Recall that an open covering of a topological space X is a family of open sets whose union is X. The join of two open coverings $\mathcal{U}$ and $\mathcal{V}$ is $\mathcal{U} \vee \mathcal{V} = \{U \cap V \colon U \in \mathcal{U}, V \in \mathcal{V}\}$. The inverse image of an open covering $\mathcal{U}$ by a map $F \colon X \mapsto X$ is $F^{-1}(\mathcal{U}) = \{F^{-1}(U) \colon U \in \mathcal{U}\}$. On the basis of these previous notions, the entropy of a system $\langle X, F\rangle$ over an open covering $\mathcal{U}$ is defined as

$$H(X,F,\mathcal{U}) = \lim_{n\to\infty} \frac{\log |\mathcal{U} \vee F^{-1}(\mathcal{U}) \cdots \vee F^{-(n-1)}(\mathcal{U})|}{n} ,$$

where $|U|$ is the cardinality of U.

Definition 13 (Topological Entropy) The topological entropy of a DTDS $\langle X, F\rangle$ is

$$h(X,F) = \sup\{H(X,F,\mathcal{U}) \colon \mathcal{U} \text{ is an open covering of } X\} \tag{1}$$

Topological entropy represents the exponential growth of the number of orbit segments which can be distinguished with a certain good, but finite, accuracy. In other words, it measures the uncertainty of the system evolutions when a partial knowledge of the initial state is given.

There are close relationships between the entropy and the topological properties we have seen so far. For instance, we have the following.

Proposition 4 ([3,12,28]) *In compact DTDS, transitivity and positive entropy imply sensitivity.*

Cellular Automata

Consider the set of configurations C which consists of all functions from $\mathbb{Z}^D$ into A. The space C is usually equipped with the Thychonoff (or Cantor) metric d defined as

$$\forall a, b \in C, \quad d(a,b) = 2^{-n}, \quad \text{with} \quad n = \min_{\vec{v}\in\mathbb{Z}^D} \left\{ \|v\|_\infty \colon a(\vec{v}) \neq b(\vec{v}) \right\},$$

where $\|v\|_\infty$ denotes the maximum of the absolute value of the components of $\vec{v}$. The topology induced by d coincides with the product topology induced by the discrete topology on A. With this topology, C is a compact, perfect and totally disconnected space.

Let $N = \{\vec{u}_1, \ldots, \vec{u}_s\}$ be an ordered set of vectors of $\mathbb{Z}^D$ and $f \colon A^s \mapsto A$ be a function.

Definition 14 (CA) The D-dimensional CA based on the local rule f and the neighborhood frame N is the pair $\langle C\rangle, F$ where $F \colon C \mapsto C$ is the global transition rule defined as follows:

$$\forall c \in C, \quad \forall \vec{v} \in \mathbb{Z}^D, \quad F(c)(\vec{v}) = f(c(\vec{v} + \vec{u}_1), \ldots, c(\vec{v} + \vec{u}_s)) . \tag{2}$$

Note that the mapping F is (uniformly) continuous with respect to the Thychonoff metric. Hence, the pair $\langle C\rangle, F$ is a proper discrete time dynamical system.

Let $Z_m = \{0, 1, \ldots, m-1\}$ be the group of the integers modulo m. Denote by C_m the configuration space C for the special case $A = Z_m$. When C_m is equipped with the natural extensions of the sum and the product operations, it turns out to be a linear space. Therefore, one can exploit the properties of linear spaces to simplify the proofs and the overall presentation.

A function $f \colon Z_m^s \mapsto Z_m$ is said to be linear if there exist $\lambda_1, \ldots, \lambda_s \in Z_m$ such that it can be expressed as:

$$\forall (x_1, \ldots, x_s) \in Z_m^s, \quad f(x_1, \ldots, x_s) = \left[\sum_{i=1}^{s} \lambda_i x_i \right]_m$$

where $[x]_m$ is the integer x taken modulo m.

Definition 15 (Linear CA) A D-dimensional linear CA is a CA $\langle C\rangle_m, F$ whose local rule f is linear.

Note that for the linear CA Equ. (2) becomes:

$$\forall c \in C, \quad \forall \vec{v} \in \mathbb{Z}^D, \quad F(c)(\vec{v}) = \left[\sum_{i=1}^{s} \lambda_i c(\vec{v} + \vec{u}_i) \right]_m .$$

The Case of Cellular Automata

In this section the results seen so far are specialized to the CA setting, focusing on dimension one. The following result allows a first classification of one-dimensional CA according to their degree of chaoticity.

Theorem 1 ([43]) *A one-dimensional CA is sensitive if and only if it is not almost equicontinuous.*

In other words, for CA the dichotomy between sensitivity and almost equicontinuity is true and not only under

the transitivity condition. As a consequence, the family of all one-dimensional CA can be partitioned into four classes [43]:

(k1) equicontinuous CA;
(k2) strictly almost equicontinuous CA;
(k3) sensitive CA;
(k4) expansive CA.

This classification almost fits for higher dimensions. The problem is that there exist CA between the classes $K2$ and $K3$ (i. e., non sensitive CA without any equicontinuity point). Even relaxing $K2$ definition to "CA having some equicontinuity point", the gap persists (see, for instance [57]).

Unfortunately, much like most of the interesting properties of CA, the properties defining the above classification scheme are also affected by undecidability.

Theorem 2 ([23]) *For each $i = 1, 2, 3$, there is no algorithm to decide if a one-dimensional CA belongs to the class Ki.*

The following conjecture stresses the fact that nothing is known about the decidability for the membership in $K4$.

Conjecture 1 (Folklore) *Membership in class K4 is undecidable.*

Remark that the above conjecture is clearly false for dimensions greater than 1 since there do not exist expansive CA for dimension strictly greater than 1 [53].

Proposition 5 ([7,11]) *Expansive CA are strongly transitive and mixing.*

In CA settings, the notion of total transitivity reduces to the simple transitivity. Moreover, there is a strict relation between transitive and sensitive CA.

Theorem 3 ([49]) *If a CA is transitive then it is totally transitive.*

Theorem 4 ([28]) *Transitive CA are sensitive.*

As we have already seen, sensitivity is undecidable. Hence, in view of the combinatorial complexity of transitive CA, the following conjectures sound true.

Conjecture 2 *Transitivity is an undecidable property.*

Conjecture 3 [48] *Strongly transitive CA are (topologically) mixing.*

Chaos and Combinatorial Properties

In this section, when referring to a one-dimensional CA, we assume that $u_1 = \min N$ and $u_s = \max N$ (see also Sect. "Definitions"). Furthermore, we call *elementary* a one-dimensional CA with alphabet $A = \{0, 1\}$ and $N = \{-1, 0, 1\}$ (there exist 256 possible elementary CA which can be enumerated according to their local rule [64]).

In CA settings, most of the chaos components are related to some properties of combinatorial nature like injectivity, surjectivity and openness.

First of all, remark that injectivity and surjectivity are dimension sensitive properties in the sense of the following.

Theorem 5 ([4,39]) *Injectivity and surjectivity are decidable in dimension 1, while they are not decidable in dimension greater than 1.*

A one-dimensional CA is said to be a *right CA* (resp., *left CA*) if $u_1 > 0$ (resp., $u_s < 0$).

Theorem 6 ([1]) *Any surjective and right (or left) CA is topologically mixing.*

The previous result can be generalized in dimension greater than 1 in the following sense.

Theorem 7 *If for a given surjective D-dimensional CA there exists a $(D-1)$-dimensional hyperplane H (as a linear subspace of $\mathbb{Z}^D$) such that:*

1. *all the neighbor vectors stay on the same side of a H, and*
2. *no vectors lay on H,*

then the CA is topologically mixing.

Proof Choose two configurations c and d and a natural number r. Let U and V be the two distinct open balls of radius 2^{-r} and of center c and d, respectively (in a metric space (X, d) the open ball of radius $\delta > 0$ and center $x \in X$ is the set $B_\delta(x) = \{y \in X \mid d(y, x) < \delta\}$). For any integer $n > 1$, denote by N_n the neighbor frame of the CA F^n and with $d_n \in F^{-n}(d)$ any n-preimage of d. The values $c(\vec{x})$ for $\vec{x} \in O$ depend only on the values $d_n(\vec{x})$ for $\vec{x} \in O + N_n$, where $O = \{\vec{v} \mid \|v\|_\infty \le r\}$. By the hypothesis, there exists an integer $m > 0$ such that for any $n \ge m$ the sets O and $O + N_n$ are disjoint. Therefore, for any $n \ge m$ build a configuration $e_n \in C$ such that $e_n(\vec{x}) = d(\vec{x})$ for $\vec{x} \in O$, and $e_n(\vec{x}) = d_n(\vec{x})$ for $\vec{x} \in O + N_n$. Then, for any $n \ge m$, $e_n \in V$ and $F^n(e_n) \in U$. □

Injectivity prevents a CA from being strong transitive as stated in the following

Theorem 8 ([11]) *Any strongly transitive CA is not injective.*

Recall that a CA of global rule F is open if F is an open function. Equivalently, in the one-dimensional case, every configuration has the same numbers of predecessors [33].

Theorem 9 ([55]) *Openness is decidable in dimension one.*

Remark that mixing CA are not necessarily open (consider, for instance, the elementary rule 106, see [44]). The following conjecture is true when replacing strong transitivity by expansively [43].

Conjecture 4 *Strongly transitive CA are open.*

Recall that the shift map $\sigma : A^{\mathbb{Z}} \mapsto A^{\mathbb{Z}}$ is the one-dimensional linear CA defined by the neighborhood $N = \{+1\}$ and by the coefficient $\lambda_1 = 1$. A configuration of a one-dimensional CA is called jointly periodic if it is periodic both for the CA and the shift map (i. e., it is also spatially periodic). A CA is said to have the joint denseness of periodic orbits property (JDPO) if it admits a dense set of jointly periodic configurations. Obviously, JDPO is a stronger form of DPO.

Theorem 10 ([13]) *Open CA have JDPO.*

The common feeling is that (J)DPO is a property of a class wider than open CA. Indeed,

Conjecture 5 [8] *Every surjective CA has (J)DPO.*

If this conjecture were true then, as a consequence of Theorem 5 and Proposition 2, DPO would be decidable in dimension one (and undecidable in greater dimensions). Up to now, Conjecture 5 has been proved true for some restricted classes of one-dimensional surjective CA beside open CA.

Theorem 11 ([8]) *Almost equicontinuous surjective CA have JDPO.*

Consider for a while a CA whose alphabet is an algebraic group. A configuration is said to be *finite* if there exists an integer h such that for any i, $|i| > k$, $c(i) = 0$, where 0 is the null element of the group. Denote $s(c) = \sum_i c(i)$ the sum of the values of a finite configuration. A one-dimensional CA F is called *number conserving* if for any finite configuration c, $s(c) = s(F(c))$.

Theorem 12 ([26]) *Number-conserving surjective CA have DPO.*

If a CA F is number-conserving, then for any $h \in \mathbb{Z}$ the CA $\sigma^h \circ F$ is number-conserving. As a consequence we have that

Corollary 1 *Number-conserving surjective CA have JDPO.*

Proof Let F be a number-conserving CA. Choose $h \in \mathbb{Z}$ in such a way that the CA $\sigma^h \circ F$ is a (number-conserving) right CA. By a result in [1], both the CA $\sigma^h \circ F$ and F have JDPO. □

In a recent work it is proved that the problem of solving Conjecture 5 can be reduced to the study of mixing CA.

Theorem 13 ([1]) *If all mixing CA have DPO then every surjective CA has JDPO.*

As a consequence of Theorem 13, if all mixing CA have DPO then all transitive CA have DPO.

Permutivity is another easy-to-check combinatorial property strictly related to chaotic behavior.

Definition 16 (Permutive CA) A function $f : A^s \mapsto A$ is permutive in the variable a_i if for any $(a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_s) \in A^{s-1}$ the function $a \mapsto f(a_1, \ldots, a_{i-1}, a, a_{i+1}, \ldots, a_s)$ is a permutation.

In the one-dimensional case, a function f that is permutive in the leftmost variable a_1 (resp., rightmost a_s), it is called leftmost (resp. rightmost) permutive. CA with either leftmost or rightmost permutive local rule share most of the chaos components.

Theorem 14 ([18]) *Any one-dimensional CA based on a leftmost (or, rightmost) permutive local rule with $u_1 < 0$ (resp., $u_s > 0$) is topologically mixing.*

The previous result can be generalized to any dimension in the following sense.

Theorem 15 *Let f and N be the local rule and the neighborhood frame, respectively, of a given D-dimensional CA. If there exists i such that*

1. *f is permutive in the variable a_i, and*
2. *the neighbor vector $\vec{u}_i$ is such that $\|\vec{u}_i\|_2 = \max\{\|\vec{u}\|_2 \mid \vec{u} \in N\}$, and*
3. *all the coordinates of $\vec{u}_i$ have absolute value l, for some integer $l > 0$,*

then the given CA is topologically mixing.

Proof Without loss of generality, assume that $\vec{u}_i = (l, \ldots, l)$. Let U and V be two distinct open balls of equal radius 2^{-r}, where r is an arbitrary natural number. For any integer $n \geq 1$, denote with N_n the neighbor frame of the CA F_n, with f^n the corresponding local rule, and with $N_n(\vec{x})$ the set $\{\vec{x} + \vec{v} \mid \vec{v} \in N_n\}$ for a given $\vec{x} \in \mathbb{Z}^D$. Note that f^n is permutive in the variable corresponding to the neighbor vector $n\vec{u}_i \in N_n$. Choose two configurations $c \in U$ and $d \in V$. Let m be the smaller natural number such that $ml > 2r$. For any $n \geq m$, we are going to build a configuration $d_n \in U$ such that $F^n(d_n) \in V$. Set $d_n(\vec{z}) = c(\vec{z})$ for $\vec{z} \in O$ where $O = \{\vec{v} \mid \|\vec{v}\|_\infty \leq r\}$. In this way $d_n \in U$. In order to obtain $F^n(d_n) \in V$, it is required that $F^n(d_n)(\vec{x}) = d(\vec{x})$ for

each $\vec{x} \in O$. We complete the configuration d_n by starting with $\vec{x} = \vec{y}$, where $\vec{y} = (-r, \ldots, -r)$. Choose arbitrarily the values $d_n(\vec{z})$ for $\vec{z} \in (N_n(\vec{y}) \setminus O) \setminus \{\vec{y} + n\vec{u}_i\}$ (note that $O \subset N_n(\vec{y})$). By permutivity of f^n, there exists $a \in A$ such that if one set $d_n(\vec{y} + n\vec{u}_i) = a$, then $F^n(d_n)(\vec{y}) = d(\vec{y})$. Let now $\vec{x} = \vec{y} + \vec{e}_1$. Choose arbitrarily the values $d_n(\vec{z})$ for $\vec{z} \in (N_n(\vec{x}) \setminus N_n(\vec{y})) \setminus \{\vec{x} + n\vec{u}_i\}$. By the same argument as above, there exists $a \in A$ such that if one set $d_n(\vec{x} + n\vec{u}_i) = a$, then $F^n(d_n)(\vec{x}) = d(\vec{x})$. Proceeding in this way one can complete d_n in order to obtain $F^n(d_n) \in V$. □

Theorem 16 ([16]) *Any one-dimensional CA based on a leftmost (or, rightmost) permutive local rule with $u_1 < 0$ (resp., $u_s > 0$) has (J)DPO.*

Theorem 17 ([54]) *Any one-dimensional CA based on a leftmost and rightmost permutive local rule with $u_1 < 0$ and $u_s > 0$ is expansive.*

As a consequence of Proposition 5, we have the following result for which we give a direct proof in order to make clearer the result which follows immediately after.

Proposition 6 *Any one-dimensional CA based on a leftmost and rightmost permutive local rule with $u_1 < 0$ and $u_s > 0$ is strongly transitive.*

Proof Choose arbitrarily two configurations $c, o \in A^{\mathbb{Z}}$ and an integer $k > 0$. Let $n > 0$ be the first integer such that $nr > k$, where $r = \max\{-u_1, u_s\}$. We are going to construct a configuration $b \in A^{\mathbb{Z}}$ such that $d(b, c) < 2^{-k}$ and $F^n(b) = o$. Fix $b(x) = c(x)$ for each $x = nu_1, \ldots, nu_s - 1$. In this way $d(b, c) < 2^{-k}$. For each $i \in \mathbb{N}$ we are going to find suitable values $b(nu_s + i)$ in order to obtain $F^n(b)(i) = o(i)$. Let us start with $i = 0$. By the hypothesis, the local rule f^n of the CA F^n is permutive in the rightmost variable nu_s. Thus, there exists a value $a_0 \in A$ such that, if one sets $b(nu_s) = a_0$, we obtain $F^n(b)(0) = o(0)$. By the same reasons as above, there exists a value $a_1 \in A$ such that, if one set $b(nu_s + 1) = a_1$, we obtain $F^n(b)(1) = o(1)$. Proceeding in this way one can complete the configuration b for any position $nu_s + i$. Finally, since f^n is permutive also in the leftmost variable nu_1 one can use the same technique to complete the configuration b for the positions $nu_1 - 1, nu_1 - 2, \ldots$, in such a way that for any integer $i < 0$, $F^n(b)(i) = o(i)$. □

The previous result can be generalized as follows. Denote $\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_D$ the canonical basis of $\mathbb{R}^D$.

Theorem 18 *Let f and N be the local rule and the neighborhood frame, respectively, of a given D-dimensional CA. If there exists an integer $l > 0$ such that*

1. *f is permutive in all the 2^D variable corresponding to the neighbor vectors $(\pm l, \ldots, \pm l)$, and*
2. *for each vector $\vec{u} \in N$, we have $\|u\|_\infty \leq l$,*

then the CA F is strongly transitive.

Proof For the sake of simplicity, we only trait the case $D = 2$. For higher dimensions, the idea of the proof is the same. Let $\vec{u}_2 = (l, l)$, $\vec{u}_3 = (-l, l)$, $\vec{u}_4 = (-l, -l)$, $\vec{u}_5 = (l, -l)$. Choose arbitrarily two configurations $c, o \in A^{\mathbb{Z}^2}$ and an integer $k > 0$. Let $n > 0$ be the first integer such that $nl > k$. We are going to construct a configuration $b \in A^{\mathbb{Z}^2}$ such that $d(b, c) < 2^{-k}$ and $F^n(b) = o$. Fix $b(x) = c(x)$ for each $\vec{x} \neq n\vec{u}_2$ with $\|x\|_\infty \leq n$. In this way $d(b, c) < 2^{-k}$. For each $i \in \mathbb{Z}$ we are going to find suitable values for the configuration b in order to obtain $F^n(b)(i\vec{e}_1) = o(i\vec{e}_1)$. Let us start with $i = 0$. By the hypothesis, the local rule f^n of the CA F^n is permutive in the variable $n\vec{u}_2$. Thus, there exists a value $a_{(0,0)} \in A$ such that, if one set $b(n\vec{u}_1) = a_{(0,0)}$, we obtain $F^n(b)(\vec{0}) = o(\vec{0})$. Now, choose arbitrary values of b in the positions $(n+1)\vec{e}_1 + j\vec{e}_2$ for $j = -n, \ldots, n-1$. By the same reasons as above, there exists a value $a_{(0,1)} \in A$ such that, if one sets $b(nu_2 + 1\vec{e}_1) = a_{(0,1)}$, we obtain $F^n(b)(\vec{e}_1) = o(\vec{e}_1)$. Proceeding in this way, at each step i $(i > 1)$, one can complete the configuration b for all the positions $(n+i)\vec{e}_1 + j\vec{e}_2$ for $j = -n, \ldots, n$, obtaining $F^n(b)(i\vec{e}_1) = o(i\vec{e}_1)$. In a similar way, by using the fact that the local rule f^n of the CA F^n is permutive in the variable $n\vec{u}_3$, for any $i < 0$ one can complete the configuration b for all the positions $(-n+i)\vec{e}_1 + j\vec{e}_2$ for $j = -n, \ldots, n$, obtaining $F^n(b)(i\vec{e}_1) = o(i\vec{e}_1)$. Now, for each step $j = 1, 2, \ldots$, choose arbitrarily the values of b in the positions $i\vec{e}_1 + (n+j)\vec{e}_2$ and $i\vec{e}_1 + (n+j)\vec{e}_2$ with $i = -n, \ldots n-1$. The permutivity of f^n in the variables $n\vec{u}_2$, $n\vec{u}_3$, $n\vec{u}_5$, and $n\vec{u}_4$ permits one to complete the configuration b in the positions $(n+i)\vec{e}_1 + (n+j)\vec{e}_2$ for all integers $i \geq 0$, $(-n+i)\vec{e}_1 + (-n+j)\vec{e}_2$ for all integer $i < 0$, $(-n+i)\vec{e}_1 + (-n-j)\vec{e}_2$ for all integers $i \geq 0$, and $(n+i)\vec{e}_1 + (-n-j)\vec{e}_2$ for all integers $i < 0$, so that for each step j we obtain $\forall i \in \mathbb{Z}, F^n(b)(i\vec{e}_1 + j\vec{e}_2) = o(i\vec{e}_1 + j\vec{e}_2)$. □

CA, Entropy and Decidability

In [34], it is shown that, in the case of CA, the definition of topological entropy can be restated in a simpler form than (1).

The space-time diagram $S(c)$ generated by a configuration c of a D-dimensional CA is the $(D+1)$-dimensional infinite figure obtained by drawing in sequence the elements of the orbit of initial state c along the temporal axis.

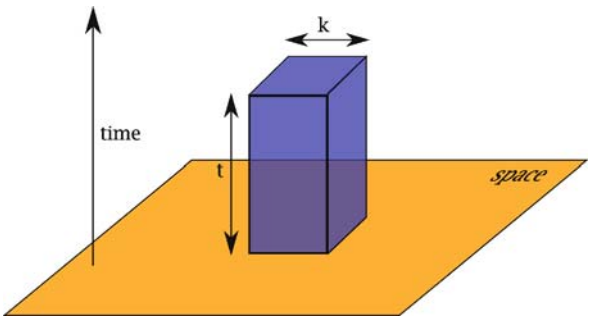

Chaotic Behavior of Cellular Automata, Figure 1
$\mathcal{N}(k, t)$ is the number of distinct *blue blocks* that can be obtained starting from any initial configuration (*orange plane*)

Formally, $S(c)$ is a function from $\mathbb{N} \times \mathbb{Z}^D$ in A defined as $\forall t \in \mathbb{N}, \forall \vec{v} \in \mathbb{Z}^D$, $S(c)(t, \vec{v}) = F^t(c)(\vec{v})$. For a given CA, fix a time t and a finite square region of side of length k in the lattice. In this way, a finite $(D+1)$-dimensional figure (hyper-rectangle) is identified in all space-time diagrams. Let $\mathcal{N}(k, t)$ be the number of distinct finite hyper-rectangles obtained by all possible space-time diagrams for the CA (i. e., $\mathcal{N}(k, t)$ is the number of the all space-time diagrams which are distinct in this finite region). The topological entropy of any given CA can be expressed as

$$h(C, F) = \lim_{k \to \infty} \lim_{t \to \infty} \mathcal{N}(k, t)$$

Despite the expression of the CA entropy is simpler than for a generic DTDS, the following result holds.

Theorem 19 ([34]) *The topological entropy of CA is uncomputable.*

Nevertheless there exist some classes of CA where it is computable [20,45]. Unfortunately, in most of these cases it is difficult to establish if a CA is a member of these classes.

Results for Linear CA: Everything Is Detectable

In the sequel, we assume that a linear CA on C_m is based on a neighborhood frame $N = \vec{u}_1, \ldots, \vec{u}_s$ whose corresponding coefficients of the local rule are $\lambda_1, \ldots, \lambda_s$. Moreover, without loss of generality, we suppose $\vec{u}_1 = \vec{0}$. In most formulas the coefficient λ_1 does not appear.

Decidability Results for Chaotic Properties

The next results state that all chaotic properties introduced in Section III are decidable. Yet, one can use the formulas to build samples of cellular automata that has the required properties.

Theorem 20 ([17,19,47])
Sensitivity *a linear CA is sensitive to the initial conditions if there exists a prime number p such that $p|m$ and $p \nmid \gcd\{\lambda_2, \ldots, \lambda_s\}$.*
Transitivity *a linear CA is topologically transitive if and only if $\gcd\{\lambda_2, \ldots, \lambda_s\} = 1$.*
Mixing *a linear CA is topologically mixing if and only if it is topologically transitive.*
Strong transitivity *a linear CA is strongly transitive if for each prime p dividing m, there exist at least two coefficients λ_i, λ_j such that $p \nmid \lambda_i$ and $p \nmid \lambda_j$.*
Regularity (DPO) *a linear CA has the denseness of periodic points if it is surjective.*

Concerning positive expansively, since in dimensions greater than one, there are no such CA, the following theorem characterizes expansively for linear CA in dimension one. For this situation we consider a local rule f with expression $f(x_{-r}, \ldots, x_r) = \left[\sum_{i=-r}^{r} a_i x_i\right]_m$.

Theorem 21 ([47]) *A linear one dimensional CA is positively expansive if and only if $\gcd\{m, a_{-r}, \ldots, a_{-1}\} = 1$ and $\gcd\{m, a_1, \ldots, a_r\} = 1$.*

Decidability Results for Other Properties

The next result was stated incompletely in [47] since the case of non sensitive CA without equicontinuity points is not treated, tough they exist [57].

Theorem 22 *Let F be a linear cellular automaton. Then the following properties are equivalent*

1. *F is equicontinuous*
2. *F has an equicontinuity point*
3. *F is not sensitive*
4. *for all prime p such that $p|m$, p divides $gcd(\lambda_2, \ldots, \lambda_s)$.*

Proof 1) $\Longrightarrow$ 2) and 2) $\Longrightarrow$ 3) are obvious. 3) $\Longrightarrow$ 4) is done by negating the formula for sensitive CA in Theorem 20.

Let us prove that 4) $\Longrightarrow$ 1). Suppose that F is a linear CA. We decompose $F = G + H$ by separating the term in λ_1 from the others:

$$H(x)(\vec{v}) = \lambda_1 x(\vec{v}) \quad G(x)(\vec{v}) = \left[\sum_{i=2}^{s} \lambda_i c(\vec{v} + \vec{u}_i)\right]_m .$$

Let $m = p_1^{\alpha_1} \cdots p_l^{\alpha_l}$ be the decomposition in prime factors and $a = \mathrm{lcm}\{\alpha_i\}$. The condition 4) gives that for all k, $\Pi_{i=1}^{l} p_i | \lambda_k$ and then m divides any product of a factors λ_i.

Let $\vec{v}$ be a vector such that for all i, $\vec{u}_i - \vec{v}$ has non-negative coordinates. Classically, we represent local rules

of linear CA by D-variable polynomials (this representation, together with the representation of configurations by formal power series allows to simplify the calculus of images through the iterates of the CA [36]). Let $X_1, \ldots, X_D$ be the variables. For $\vec{y} = (y_1, \ldots, y_D) \in \mathbb{Z}^D$, we note $X^{\vec{y}}$ the monomial $\Pi_{i=1}^{D} X_i^{y_i}$. We consider the polynomial P associated with G combined with a translation of vector $\vec{v}$, $P = \Pi_{i=2}^{s} \lambda_i X^{\vec{u}_i - \vec{v}}$. The coefficients of P^a are products of a factors λ_i hence $[P^a]_m = 0$. This means that the composition of G and the translation of vector $\vec{v}$ is nilpotent and then that G is nilpotent. As F is the sum of λ_1 times the identity and a nilpotent CA, we conclude that F is equicontinuous. □

The next theorem gives the formula for some combinatorial properties

Theorem 23 ([36]).
Surjectivity *a linear CA is surjective if and only if* $\gcd\{\lambda_1, \ldots, \lambda_s\} = 1$.
Injectivity *a linear CA is injective if and only if for each prime p decomposing m there exists an unique coefficient λ_i such that p does not divide λ_i.*

Computation of Entropy for Linear Cellular Automata

Let us start by considering the one-dimensional case.

Theorem 24 *Let us consider a one dimensional linear CA. Let $m = p_1^{k_1} \cdots p_h^{k_h}$ be the prime factor decomposition of m. The topological entropy of the CA is*

$$h(C, F) = \sum_{i=1}^{h} k_i (R_i - L_i) \log(p_i)$$

where $L_i = \min P_i$ and $R_i = \max P_i$, with $P_i = \{0\} \cup \{j\colon \gcd(a_j, p_i) = 1\}$.

In [50], it is proved that for dimensions greater than one, there are only two possible values for the topological entropy: zero or infinity.

Theorem 25 *A D-dimensional linear CA $\langle C, F\rangle$ with $D \geq 2$ is either sensitive and $h(C, F) = \infty$ or equicontinuous and $h(C, F) = 0$.*

By a combination of Theorem 25 and 20, it is possible to establish if a D dimensional linear CA with $D \geq 2$ has either zero or infinite entropy.

Linear CA, Fractal Dimension and Chaos

In this section we review the relations between strong transitivity and fractal dimension in the special case of linear CA. The idea is that when a system is chaotic then it produces evolutions which are complex even from a (topological) dimension point of view.

Any linear CA F, can be associated with its *W-limit set*, a subset of the $(D+1)$-dimensional Euclid space defined as follows. Let t_n be a sequence of integers (we call them times) which tends to infinity. A subset $S_F(t_n)$ of $(D+1)$-dimensional Euclid space represents a space-time pattern until time $t_n - 1$:

$$S_F(t_n) = \left\{(t, i) \text{s.t.} F^t(e_1)_i \neq 0,\ t < t_n\right\} .$$

A *W-limit set* for F is defined by $\lim_{n\to\infty} S_F(t_n)/t_n$ if the limit exists, where $S_F(t_n)/t_n$ is the contracted set of $S_F(t_n)$ by the rate $\frac{1}{t_n}$ i. e. $S_F(t_n)/t_n$ contains the point $(t/t_n, i/t_n)$ if and only if $S_F(t_n)$ contains the point (t, i). The limit $\lim_{n\to\infty} S_F(t_n)/t_n$ exists when $\liminf_{n\to\infty} S_F(t_n)/t_n$ and $\limsup_{n\to\infty} S_F(t_n)/t_n$ coincide, where

$$\liminf_{n\to\infty} \frac{S_F(t_n)}{t_n} = \left\{ x \in \mathbb{R}^{D+1} :\ \forall j, \exists x_j \in \frac{S_F(t_j)}{t_j}, x_j \to x \text{ when } j \to \infty \right\}$$

and

$$\limsup_{n\to\infty} \frac{S_F(t_n)}{t_n} = \left\{ x \in \mathbb{R}^{D+1} :\ \exists \left\{t_{n_j}\right\}, \forall j, \exists x_{n_j} \in \frac{S_F(t_{n_j})}{t_{n_j}}, x_{n_j} \to x \text{ when } j \to \infty \right\} ,$$

for a subsequence $\left\{t_{n_j}\right\}$ of $\{t_n\}$.

For the particular case of linear CA, the W-limit set always exists [30,31,32,56]. In the last ten years, the W-limit set of additive CA has been extensively studied [61,62,63]. It has been proved that for most of additive CA, it has interesting dimensional properties which completely characterize the set of quiescent configurations [56]. Here we link dimension properties of a W-limit set with chaotic properties. Correlating dimensional properties of invariant sets to dynamical properties has become during the years a fruitful source of new understanding [52].

Let X be a metric space. The *Hausdorff dimension* D_H of $V \subseteq X$ is defined as:

$$D_H(V) = \sup\left\{ h \in \mathbb{R} | \lim_{\epsilon\to 0} \left(\inf \sum |U_i|^h\right) = \infty \right\}$$

where the infimum is taken over all countable coverings U_i of V such that the diameter $|U_i|$of each U_i is less than ϵ (for more on Hausdorff dimension as well as other definitions of fractal dimension see [24]).

Given a CA F we denote $D_H(F)$ the Hausdorff dimension of its W-limit set.

Proposition 7 ([25,47]) *Consider a linear CA F over Z_{p^k} where p is a prime number. If $1 < D_H(F) < 2$ then F is strongly transitive.*

The converse relation is still an open problem. It would be also an interesting research direction to find out similar notions and results for general CA.

Conjecture 6 *Consider a linear CA F over Z_{p^k}, where p is a prime number. If F is strongly transitive then $1 < D_H(F) < 2$.*

Future Directions

In this chapter we reviewed the chaotic behavior of cellular automata. It is clear from the results seen so far that there are close similarities between the chaotic behavior of dynamical systems on the real interval and CA. To complete the picture, it remains only to prove (or disprove) Conjecture 5. Due to its apparent difficulty, this problem promises to keep researchers occupied for some years yet.

The study of the decidability of chaotic properties like expansively, transitivity, mixing etc. is another research direction which should be further addressed in the near future. It seems that new ideas are necessary since the proof techniques used up to now have been revealed as unsuccessful. The solution to these problems will be a source of new understanding and will certainly produce new results in connected fields.

Finally, remark that most of the results on the chaotic behavior of CA are concerned with dimension one. A lot of work should be done to verify what happens in higher dimensions.

Acknowledgments

This work has been supported by the Interlink/MIUR project "Cellular Automata: Topological Properties, Chaos and Associated Formal Languages", by the ANR Blanc Project "Sycomore" and by the PRIN/MIUR project "Formal Languages and Automata: Mathematical and Applicative Aspects".

Bibliography

Primary Literature

1. Acerbi L, Dennunzio A, Formenti E (2007) Shifting and lifting of cellular automata. In: Third Conference on Computability in Europe, CiE 2007, Siena, Italy, 18–23 June 2007. Lecture Notes in Computer Science, vol 4497. Springer, Berlin, pp 1–10
2. Adler R, Konheim A, McAndrew J (1965) Topological entropy. Trans Amer Math Soc 114:309–319
3. Akin E, Auslander E, Berg K (1996) When is a transitive map chaotic? In: Bergelson V, March P, Rosenblatt J (eds) Convergence in Ergodic Theory and Probability. de Gruyter, Berlin, pp 25–40
4. Amoroso S, Patt YN (1972) Decision procedures for surjectivity and injectivity of parallel maps for tesselation structures. J Comp Syst Sci 6:448–464
5. Auslander J, Yorke JA (1980) Interval maps, factors of maps and chaos. Tohoku Math J 32:177–188
6. Banks J, Brooks J, Cairns G, Davis G, Stacey P (1992) On Devaney's definition of chaos. Am Math Mon 99:332–334
7. Blanchard F, Maass A (1997) Dynamical properties of expansive one-sided cellular automata. Israel J Math 99:149–174
8. Blanchard F, Tisseur P (2000) Some properties of cellular automata with equicontinuity points. Ann Inst Henri Poincaré, Probabilité et Statistiques 36:569–582
9. Blanchard F, Kůrka P, Maass A (1997) Topological and measure-theoretic properties of one-dimensional cellular automata. Physica D 103:86–99
10. Blanchard F, Formenti E, Kůrka K (1998) Cellular automata in the Cantor, Besicovitch and Weyl topological spaces. Complex Syst 11:107–123
11. Blanchard F, Cervelle J, Formenti E (2005) Some results about chaotic behavior of cellular automata. Theor Comp Sci 349:318–336
12. Blanchard F, Glasner E, Kolyada S, Maass A (2002) On Li-Yorke pairs. J Reine Angewandte Math 547:51–68
13. Boyle M, Kitchens B (1999) Periodic points for cellular automata. Indag Math 10:483–493
14. Boyle M, Maass A (2000) Expansive invertible one-sided cellular automata. J Math Soc Jpn 54(4):725–740
15. Cattaneo G, Formenti E, Margara L, Mazoyer J (1997) A Shift-invariant Metric on S^Z Inducing a Non-trivial Topology. In: Mathematical Foundations of Computer Science 1997. Lecture Notes in Computer Science, vol 1295. Springer, Berlin, pp 179–188
16. Cattaneo G, Finelli M, Margara L (2000) Investigating topological chaos by elementary cellular automata dynamics. Theor Comp Sci 244:219–241
17. Cattaneo G, Formenti E, Manzini G, Margara L (2000) Ergodicity, transitivity, and regularity for linear cellular automata. Theor Comp Sci 233:147–164. A preliminary version of this paper has been presented to the Symposium of Theoretical Computer Science (STACS'97). LNCS, vol 1200
18. Cattaneo G, Dennunzio A, Margara L (2002) Chaotic subshifts and related languages applications to one-dimensional cellular automata. Fundam Inform 52:39–80
19. Cattaneo G, Dennunzio A, Margara L (2004) Solution of some conjectures about topological properties of linear cellular automata. Theor Comp Sci 325:249–271
20. D'Amico M, Manzini G, Margara L (2003) On computing the entropy of cellular automata. Theor Comp Sci 290:1629–1646
21. Denker M, Grillenberger C, Sigmund K (1976) Ergodic Theory on Compact Spaces.Lecture Notes in Mathematics, vol 527. Springer, Berlin
22. Devaney RL (1989) An Introduction to chaotic dynamical systems, 2nd edn. Addison-Wesley, Reading
23. Durand B, Formenti E, Varouchas G (2003) On undecidability of equicontinuity classification for cellular automata. Dis-

crete Mathematics and Theoretical Computer Science, vol AB. pp 117–128
24. Edgar GA (1990) Measure, topology and fractal geometry. Undergraduate texts in Mathematics. Springer, New York
25. Formenti E (2003) On the sensitivity of additive cellular automata in Besicovitch topologies. Theor Comp Sci 301(1–3):341–354
26. Formenti E, Grange A (2003) Number conserving cellular automata II: dynamics. Theor Comp Sci 304(1–3):269–290
27. Furstenberg H (1967) Disjointness in ergodic theory, minimal sets, and a problem in diophantine approximation. Math Syst Theor (now Theor Comp Syst) 1(1):1–49
28. Glasner E, Weiss B (1993) Sensitive dependence on initial condition. Nonlinearity 6:1067–1075
29. Guckenheimer J (1979) Sensitive dependence to initial condition for one-dimensional maps. Commun Math Phys 70:133–160
30. Haeseler FV, Peitgen HO, Skordev G (1992) Linear cellular automata, substitutions, hierarchical iterated system. In: Fractal geometry and Computer graphics. Springer, Berlin
31. Haeseler FV, Peitgen HO, Skordev G (1993) Multifractal decompositions of rescaled evolution sets of equivariant cellular automata: selected examples. Technical report, Institut für Dynamische Systeme, Universität Bremen
32. Haeseler FV, Peitgen HO, Skordev G (1995) Global analysis of self-similarity features of cellular automata: selected examples. Physica D 86:64–80
33. Hedlund GA (1969) Endomorphism and automorphism of the shift dynamical system. Math Sy Theor 3:320–375
34. Hurd LP, Kari J, Culik K (1992) The topological entropy of cellular automata is uncomputable. Ergodic. Th Dyn Sy 12:255–265
35. Hurley M (1990) Ergodic aspects of cellular automata. Ergod Theor Dyn Sy 10:671–685
36. Ito M, Osato N, Nasu M (1983) Linear cellular automata over z_m. J Comp Sy Sci 27:127–140
37. IV Assaf D, Gadbois S (1992) Definition of chaos. Am Math Mon 99:865
38. Kannan V, Nagar A (2002) Topological transitivity for discrete dynamical systems. In: Misra JC (ed) Applicable Mathematics in Golden Age. Narosa Pub, New Dehli
39. Kari J (1994) Reversibility and surjectivity problems of cellular automata. J Comp Sy Sci 48:149–182
40. Kari J (1994) Rice's theorem for the limit set of cellular automata. Theor Comp Sci 127(2):229–254
41. Knudsen C (1994) Chaos without nonperiodicity. Am Math Mon 101:563–565
42. Kolyada S, Snoha L (1997) Some aspect of topological transitivity – a survey. Grazer Mathematische Berichte 334:3–35
43. Kůrka P (1997) Languages, equicontinuity and attractors in cellular automata. Ergo Theor Dyn Sy 17:417–433
44. Kůrka P (2004) Topological and Symbolic Dynamics, vol 11 of Cours Spécialisés. Société Mathématique de France, Paris
45. Di Lena P (2006) Decidable properties for regular cellular automata. In: Navarro G, Bertolossi L, Koliayakawa Y (eds) Proceedings of Fourth IFIP International Conference on Theoretical Computer Science, pp 185–196. Springer, Santiago de Chile
46. Li TY, Yorke JA (1975) Period three implies chaos. Am Math Mon 82:985–992
47. Manzini G, Margara L (1999) A complete and efficiently computable topological classification of D-dimensional linear cellular automata over Z_m. Theor Comp Sci 221(1–2):157–177
48. Margara L (1999) On some topological properties of linear cellular automata. In: Kutylowski M, Pacholski L, Wierzbicki T (eds) Mathematical Foundations of Computer Science 1999 (MFCS99). Lecture Notes in Computer Science, vol 1672. Springer, Berlin, pp 209–219
49. Moothathu TKS (2005) Homogenity of surjective cellular automata. Discret Contin Dyn Syst 13:195–202
50. Morris G, Ward T (1998) Entropy bounds for endomorphisms commuting with k actions. Israel J Math 106:1–12
51. Nasu M (1995) Textile Systems for Endomorphisms and automorphisms of the shift, vol 114 of Memoires of the American Mathematical Society. American Mathematical Society, Providence
52. Pesin YK (1997) Dimension Theory in Dynamical Systems. Chicago Lectures in Mathematics. The University of Chicago Press, Chicago
53. Shereshevsky MA (1993) Expansiveness, entropy and polynomial growth for groups acting on subshifts by automorphisms. Indag Math 4:203–210
54. Shereshevsky MA, Afraimovich VS (1993) Bipermutative cellular automata are topologically conjugate to the one-sided Bernoulli shift. Random Comput Dynam 1(1):91–98
55. Sutner K (1999) Linear cellular automata and de Bruijn automata. In: Delorme M, Mazoyer J (eds) Cellular Automata, a Parallel Model, number 460 in Mathematics and Its Applications. Kluwer, Dordrecht
56. Takahashi S (1992) Self-similarity of linear cellular automata. J Comp Syst Sci 44:114–140
57. Theyssier G (2007) Personal communication
58. Vellekoop M, Berglund R (1994) On intervals, transitivity = chaos. Am Math Mon 101:353–355
59. Walters P (1982) An Introduction to Ergodic Theory. Springer, Berlin
60. Weiss B (1971) Topological transitivity and ergodic measures. Math Syst Theor 5:71–5
61. Willson S (1984) Growth rates and fractional dimensions in cellular automata. Physica D 10:69–74
62. Willson S (1987) Computing fractal dimensions for additive cellular automata. Physica D 24:190–206
63. Willson S (1987) The equality of fractional dimensions for certain cellular automata. Physica D 24, 179–189
64. Wolfram S (1986) Theory and Applications of Cellular Automata. World Scientific, Singapore

Books and Reviews

Akin E (1993) The general topology of dynamical systems. Graduate Stud. Math 1 Am Math Soc, Providence

Akin E, Kolyada S (2003) Li-Yorke sensitivity. Nonlinearity 16:1421–1433

Block LS, Coppel WA (1992) Dynamics in One Dymension. Springer, Berlin

Katok A, Hasselblatt B (1995) Introduction to the Modern Theory of Dynamical Systems. Cambridge University Press, Cambridge

Kitchens PB (1997) Symbolic dynamics: One-Sided, Two-Sided and Countable State Markov Shifts. Universitext. Springer, Berlin

Kolyada SF (2004) Li-yorke sensitivity and other concepts of chaos. Ukr Math J 56(8):1242–1257

Lind D, Marcus B (1995) An Introduction to Symbolic Dynamics and Coding. Cambidge University Press, Cambidge

Chaotic Dynamics in Nonequilibrium Statistical Mechanics

J. Robert Dorfman
Institute for Physical Science and Technology
and Department of Physics, University of Maryland,
College Park, USA

Article Outline

Glossary
Definition of the Subject
Introduction
The Roles of Ergodicity and Mixing
for the Approach to Thermodynamic Equilibrium
Integrable, Pseudo-chaotic, and Chaotic Systems
Anosov and Anosov-like Systems; the Chaotic Hypothesis
Applications of Dynamical Systems Theory
to Non-equilibrium Statistical Mechanics
Discussion
Future Directions
Acknowledgments
Bibliography

Glossary

Chaotic systems The time evolution of a deterministic mechanical system defines a trajectory in the phase space of all the generalized coordinates and generalized momenta. Consider two infinitesimally separated points that lie on two different trajectories in this phase space. If these two trajectories typically separate exponentially with time, the systems is called chaotic provided the set of all points with an exponentially separating partner is of positive measure.

Chaotic hypothesis The hypothesis that systems of large numbers of particles interacting with short ranged forces can be treated mathematically as if the system were chaotic with no pathologies in the mathematical description of the systems' trajectories in phase space.

Dynamical systems theory The mathematical theory of the time evolution in phase space, or closely related spaces, of a deterministic system, such as a mechanical system obeying Hamiltonian equations of motion.

Ergodic systems A mechanical system is called ergodic if a typical trajectory in a phase space of finite total measure spends a fraction of its time in a set which is equal to the ratio of the measure of the set to the total measure of the phase space.

Escape rate formula Consider a chaotic dynamical system with a phase space constructed in such a way that the phase space has some kind of an absorbing boundary. The set of points, $\mathcal{R}$, in the phase space such that trajectories through them never escape through the absorbing boundary either in the forward or the backward motion is called a repeller. One can define a set of Lyapunov exponents, $\lambda_i(\mathcal{R})$ and a Kolmogorov–Sinai entropy, $h_{\mathrm{KS}}(\mathcal{R})$ for motion on the repeller. Dynamical systems theory shows that the rate of escape, γ, of points, not on the repeller, through the boundary is given by

$$\gamma = \sum_i \lambda^+(\mathcal{R}) - h_{\mathrm{KS}}(\mathcal{R}) \,, \tag{1}$$

where the sum is over all of the positive Lyapunov exponents on the repeller.

Gaussian thermostats A dynamical friction acting on the particles in a mechanical system which keeps the total energy, or the total kinetic energy of the system at a fixed value. It was invented by Gauss as the simplest solution to the problem of finding the equations of motion for a mechanical system with a constraint of fixed energy.

Gelfand triplet An operator with right and left hand eigenfunctions, possibly defined in different function spaces, and an inner product of one function from the *right* space and one from the *left* space. Generally one of these spaces contains singular functions such as Schwartz distributions and the other contains sufficiently smooth functions so that the inner product is well defined. The term *rigged Hilbert space* is also used to denote a Gelfand triplet.

Hyperbolic dynamical system A chaotic system where the tangent space to almost all trajectories in its phase space can be separated into well-defined stable and unstable manifolds, that intersect each other transversally.

Kolmogorov–Sinai entropy per unit time A measure of the rate at which information about the initial point of a chaotic trajectory is produced in time. The exponential separation of trajectories in phase space, characteristic of chaotic motion, implies that trajectories starting at very close-by, essentially indistinguishable, initial points will eventually be distinguishable. Hence as time evolves we can specify more precisely the initial point of the trajectory. Pesin has proved that for closed, hyperbolic systems, the Kolmogorov–Sinai entropy is equal to the sum of the positive Lyapunov exponents. The Kolmogorov–Sinai entropy is often called the metric entropy.

Lyapunov exponents Lyapunov exponents, λ_i, are the rates at which infinitesimally close trajectories separate

or approach with time on the unstable and stable manifolds of a chaotic dynamical system. For closed phase spaces, that is, no absorbing boundaries present, Pesin theorem states that for hyperbolic dynamical system the Kolmogorov–Sinai entropy, h_{KS} is given by the sum of all the positive Lyapunov exponents.

$$h_{\mathrm{KS}} = \sum_i \lambda_i^+ \,. \tag{2}$$

Mixing systems Mixing systems are dynamical systems with stronger dynamical properties than ergodic systems in the sense that every mixing system is ergodic but the converse is not true. A system is mixing if the following statement about the time development of a set A_t is satisfied

$$\lim_{T\to\infty} \frac{\mu(A_T \cap B)}{\mu(B)} = \frac{\mu(A)}{\mu(\mathcal{E})} \,, \tag{3}$$

where B is any set of finite measure, and $\mu(\mathcal{E})$ is the measure of the full phase space. This statement is the mathematical expression of the fact that for a mixing system, every set of finite measure becomes uniformly distributed throughout the full phase space, with respect to the measure μ.

Normal variables Microscopic variables whose values are approximately constant on large regions of the constant energy surface in phase space.

Pseudo-chaotic systems A pseudo-chaotic system is a dynamical system where the separation of nearby trajectories is algebraic in time, rather than exponential. Pseudo-chaotic systems are weakly mixing as defined by the relation

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \mathrm{d}\tau \left[\mu(A_\tau \cap B) - \frac{\mu(A)\mu(B)}{\mu(\mathcal{E})} \right] = 0. \tag{4}$$

Sinai–Ruelle–Bowen (SRB) measure SRB measures for a chaotic system are invariant measures that are smooth on unstable manifolds and possibly singular on stable manifolds.

Stable manifold A stable manifold about a point P in phase space is the set of points that will approach P at time t approaches positive infinity, that is in the infinite future of the motion.

Transport coefficients Transport coefficients characterize the proportionality between the currents of particles, momentum, or energy in a fluid, and the gradients of density, fluid velocity or temperature in the fluid. The coefficients of diffusion, shear and bulk viscosity, and thermal conductivity are transport coefficients, and appear as coefficients of the second order gradients in the Navier–Stokes and similar equations.

Unstable manifold An unstable manifold about a point P in phase space is the set of points that will approach P as time approaches negative infinity, that is, as one follows the motion backwards in time to the infinitely remote past.

Definition of the Subject

For most of its history, non-equilibrium statistical mechanics has produced mathematical descriptions of irreversible processes by invoking one or another stochastic assumptions in order to obtain useful equations. Central to our understanding of transport in fluids, for example, are random walk processes, which typically are described by stochastic equations. These in turn lead to the Einstein relation for diffusion, and its generalizations to other transport processes. This relation, as formulated by Einstein, states that the mean square displacement of a diffusing particle grows linearly in time with a proportionality constant given by the coefficient of diffusion. If we assume that such a description applies to mechanical systems of many particles, we must explain the origins of irreversibility in deterministic – and time reversible – mechanical systems, and we must locate the source of stochasticity that is invoked to derive transport equations. For systems of particles that are chaotic, it is possible to make some progress on resolving these issues, and to gain some insights into the analogous properties of systems that are not chaotic or that have both chaotic and non-chaotic behaviors. This article summarizes the current status of the application of dynamical systems theory to non-equilibrium statistical mechanics, focuses on the behavior of chaotic systems, and presents some of the important open issues needing resolution.

Introduction

Statistical mechanics is devoted to the study of the collective properties of large numbers of particles, typically atoms, molecules, electrons, etc., but can also be stars, or even galaxies. The mechanical properties of the individual particles and their mutual interactions are presumed to be known, and the emphasis is on the statistical properties of collections of large numbers of them [1,2,3]. The study of statistical mechanics has always involved both an examination of its foundations and the devising of methods to compute thermodynamic, transport, and related quantities for systems of large numbers of particles. In recent years there has been a great deal of attention focused on the foundations of statistical mechanics prompted by developments in: (i) dynamical systems theory, particularly chaotic dynamics, (ii) the mathematics of Anosov

and hard-ball systems, and in (iii) the computational physics of systems of particles undergoing processes of various kinds [4,5,6,7,8,9,10,11,12,13,14]. Moreover, modern developments in quantum physics, particularly those of some relevance for the foundations of quantum mechanics, for cosmology, and for quantum computation have some bearing on the foundations of quantum statistical mechanics [15].

Non-equilibrium statistical mechanics is faced with the task of describing the wide range of non-thermodynamic, generally time dependent behaviors of large systems, such as the transport of particles, momentum, and/or energy from one region in space to another over some time interval. Despite a large number of, as yet, unsolved problems, non-equilibrium statistical mechanics has been able to explain and provide quantitative predictions for a wide range of transport phenomena. Here we will discuss some of these applications and describe the role played by the microscopic dynamics of the constituent particles, particularly when the microscopic dynamics is classical and chaotic [14]. Most of this article will be devoted to a study of the role of chaotic dynamics in non-equilibrium transport for classical, chaotic systems. We will also consider, to some extent, classical, non-chaotic systems as well as quantum systems, where the usual notions of chaotic dynamics do not apply but one can nevertheless understand some main features of the behavior of a quantum system by examining its classical counterpart, if there is one.

This article is devoted to the role played by chaotic dynamics in our understanding non-equilibrium statistical mechanics. It will focus on two main topics: (1) Basic issues in statistical mechanics, namely the approach of systems of particles to thermodynamic equilibrium. Here we give an updated view of the role of ergodic and mixing properties, proposed by Boltzmann and Gibbs, respectively, as the basis for understanding the approach to equilibrium [1,2,3,4]. (2) Applications of dynamical systems theory to non-equilibrium statistical mechanics and the theory of transport processes. Here we show that for chaotic systems, at least, on can find some very deep relationships between macroscopic transport and microscopic dynamics [5,6,7,8,9,10,11,12,13]. We also discuss, briefly, the closely related topic of fluctuation theorems for non-equilibrium stationary states which apply even in far-from equilibrium situations [16,17,18,19,20,21].

By way of introduction we begin with some observations about transport processes in fluids in order to set the stage for describing the role of chaotic dynamics in transport theory. As we will discuss in more detail below, normal transport processes in fluids are characterized by a linear growth in time of the mean square fluctuations of an appropriate time dependent microscopic variable of the system. Typical of such a fluctuation formula is the Einstein relation which states that the mean square displacement of a particle undergoing Brownian motion in a fluid grows linearly in time with a coefficient that is, apart from a numerical factor, the diffusion coefficient of the Brownian particle [22]. That is

$$\left\langle (\mathbf{r}(t) - \mathbf{r}(0))^2 \right\rangle = 2dDt\,. \tag{5}$$

Here $\mathbf{r}(t)$ is the spatial location of the Brownian particle at some time t, and d is the number of spatial dimensions of the system. D is the diffusion coefficient appearing in the linear relation, known as Fick's Law, that relates the probability current, $\mathbf{J}(\mathbf{r}, t)$ for the Brownian particle to its probability density, $P(\mathbf{r}, t)$,

$$\mathbf{J}(\mathbf{r}, t) = -D\nabla P(\mathbf{r}, t)\,. \tag{6}$$

The angular brackets denote an average over an ensemble of trajectories of time duration t, at least. At the heart of the Einstein formula is the fact that the Brownian particle undergoes a random walk on sufficiently long time scales. The distinguishing feature of a random walk is the linear growth in time of the mean square displacement. Thus, normal transport in general is a form of a random walk process. This was formalized by Helfand in 1960 [23], who generalized the Einstein formula to other transport processes such as viscous flow, and heat conduction. By normal transport we mean transport that can be described macroscopically by linear relations between the currents of globally conserved quantities such as mass, energy, and/or momentum, and the gradients of the densities of these quantities. The coefficients of proportionality are transport coefficients, such as the coefficients of shear and bulk viscosity, thermal conductivity, diffusion, and so on. These coefficients, for normal transport, do not depend on time, but may depend on the local densities of mass and the local temperature of the fluid.

Helfand was able to show that each transport process may be regarded as a random walk process characterized by a linear growth as a function of time of the mean square fluctuation of a microscopic quantity, $M_\sigma(\Gamma, t)$, called a Helfand moment. The Helfand moments depend on all of the phase space variables of the system, now denoted collectively by Γ, and time t. For normal transport, the generalized Einstein relation becomes

$$\left\langle (M_\sigma(\Gamma, t) - M_\sigma(\Gamma, t = 0))^2 \right\rangle = \sigma C_\sigma t\,. \tag{7}$$

Here σ is a transport coefficient appearing in the Navier–Stokes or diffusion equations, C_σ is a constant, and the

average is over an equilibrium ensemble. For diffusion of a Brownian particle, for example, the Helfand moment is simply the spatial location of the particle. This result implies that normal transport processes are essentially random walk processes with a generalized "displacement", namely the Helfand moment.

If we think of a system of particles, undergoing some kind of hydrodynamic flow as a deterministic dynamical system, with, typically but not exclusively, Hamiltonian dynamics, a deep question immediately presents itself: *Where does the randomness come from that is required for transport processes to be generalized random-walk processes?* While there are a variety of answers to this question, the main point of this article is to argue that for *chaotic systems* the randomness is an intrinsic property of the dynamics of the system [14], and then to illustrate some of the results that have been obtained recently which connect microscopic dynamics to macroscopic transport for such systems. We emphasize that many chaotic systems display the kind of randomness needed for good transport properties even though such systems are deterministic, and time reversible. However chaos is neither necessary nor sufficient for normal transport. There are examples of non-chaotic systems that exhibit normal diffusion The wind-tree model to be discussed below is an example [24,25]. There are also chaotic systems that exhibit abnormal transport. The two dimensional periodic Lorentz gas with circular scatterers is chaotic and exhibits both normal and abnormal diffusion, depending upon the lattice structure and the separation distance between neighboring scatterers [26,27].

Nevertheless, chaotic systems have sufficiently many nice properties that for them it is possible to obtain a number of rather, general connections between microscopic dynamics and macroscopic transport, and to understand the approach of systems to thermodynamic equilibrium from a dynamical standpoint. Such connections, should they exist, are not yet available for those non-chaotic systems which exhibit normal transport. The dynamical properties of most non-chaotic systems are not yet sufficiently well understood for the construction of a more or less general theory for generic connections between microscopic dynamics and macroscopic transport. Most realistic dynamical systems have a *mixed* phase space, where chaotic and non-chaotic regions, to be defined below, are intermingled [28,29]. It is generally supposed that for systems of large numbers of particles in macroscopic volumes, the non-chaotic regions occupy a very small part of phase space and can be ignored for all practical purposes. However this remains to be proved, and the effects of the non-chaotic regions require much more elucidation.

The plan of this article is as follows: We begin with a discussion of the foundations of statistical mechanics in Sect. "The Roles of Ergodicity and Mixing for the Approach to Equilibrium". There we discuss dynamical properties that are often considered to be important for statistical mechanics, such as ergodicity and mixing. There we argue that these properties, by themselves, are not sufficient to explain the approach to equilibrium of systems of large numbers of particles and that further notions are required. This will lead us in Sect. "Integrable, Pseudo-chaotic, and Chaotic Systems" to a discussion of some of the dynamical systems encountered in classical statistical mechanics, particularly integrable, pseudo-chaotic, and chaotic systems. In Sect. "Anosov and Anosov-like Systems; the Chaotic Hypothesis", we consider chaotic systems in more detail and show for a simple model of a chaotic system that the stochastic behavior needed for diffusion can be seen as a natural consequence of chaotic dynamics. Despite its simplicity and low dimensionality, this model will also allow us to provide a dynamical picture of the approach of a system to equilibrium. These models are simple examples of classes of dynamical systems called Anosov and Anosov-like systems. They have very useful mathematical properties for the description of non-equilibrium phenomena in simple systems, and it is convenient to assume, whenever possible, that the systems of interest to physicists have these good properties. These properties are then used for the applications of chaotic dynamics to non-equilibrium statistical mechanics in Sect. "Applications of Dynamical Systems Theory to Non-equilibrium". There we discuss a number of results that connect microscopic dynamical properties such as Lyapunov exponents and Kolmogorov–Sinai entropies [4,5,6,7,8,9,10,11,12,13,14] to macroscopic transport coefficients for chaotic systems. In Sect. "Discussion" we review the previous discussion and give brief comments about quantum systems, where the concepts of classical chaotic dynamics do not apply, but certain features of these systems can be understood in terms of the chaotic dynamics of their classical counterparts. Finally, we mention some future directions and open questions in Sect. "Future Directions".

The Roles of Ergodicity and Mixing for the Approach to Thermodynamic Equilibrium

Ergodic Systems

Boltzmann invented the notion of ergodicity in an attempt to reconcile the apparently irreversible behavior of large systems of particles, especially their approach to thermodynamic equilibrium, with the microscopic reversibility of

Newton's equations of motion [1,2,3]. Briefly put, Boltzmann argued as follows: Consider an isolated system of particles, and follow the system's trajectory on the appropriate constant energy surface in phase space. Suppose that the underlying microscopic dynamics of the system is such that the phase space trajectory of the system, over a long time interval, spends an amount of time in every region of phase space, A, that is proportional to the measure of that region, $\mu(A)$. Here the measure is the standard equilibrium phase space measure on a constant energy surface given by

$$\mu(A) = \int_A \frac{\mathrm{d}S}{|\nabla H|} , \tag{8}$$

where $\mathrm{d}S$ is the area of an infinitesimal region of the constant energy surface, and $|\nabla H|$ is the magnitude of the gradient of the Hamiltonian function for the system at the point of integration. Boltzmann's supposition above is called the *ergodic hypothesis*, and in mathematical terms it can be stated as [4]

$$\lim_{T\to\infty} \frac{\tau(A)}{T} = \frac{\mu(A)}{\mu(\mathcal{E})} . \tag{9}$$

Here $\tau(A)$ is the amount of time the system spends in region A during a time interval T, and $\mu(\mathcal{E})$ is the measure of the entire constant energy surface, which we assume to be finite. If one accepts the hypothesis, then one can easily show, by approximating integrals by sums, for example, that the *time average* of any well behaved dynamical quantity approaches its equilibrium, micro-canonical ensemble average as the time interval over which the average is taken approaches infinity. Boltzmann's hypothesis holds equally well for the time reversed motion of the system, and is consistent with the reversibility of the microscopic dynamics.

The kinds of elementary physical systems one studies in physics courses, such as coupled systems of harmonic oscillators, are generally not ergodic, although some simple systems, such as a single, one dimensional harmonic oscillator, can be shown to be ergodic. Another simple example is the discrete time motion of a point particle on the circumference of a circle, where the particle moves a fixed irrational fraction of the length of the circumference at successive time steps. In the long time limit, the circumference is uniformly covered by points visited by the particle [30]. A great deal of mathematical research over the past few decades, and longer, has been devoted to a study of more complicated ergodic systems. The first dramatic example of a system with ergodic properties, and one that has influenced most of the more recent efforts in this direction is the geodesic motion of a point on a surface of constant negative curvature. This was proved by E. Hopf [31], and the techniques employed remain useful today. Sinai and coworkers [32], as well as many other workers, particularly Simányi [33], have given careful mathematical proofs of ergodic behavior of various systems composed of hard disks or hard spheres, generally referred to as hard-ball systems. Turaev and Rom-Kedar, Donnay and others [28,29] have shown that by softening the hard sphere potential one may change an ergodic system into a non-ergodic one. Thus, the problem of proving that a system of interest to physicists is actually ergodic, or ergodic for practical purposes, is still far from a general solution.

Mixing Systems

Gibbs took a different approach to the problem of irreversibility. He used the analogy of an ink drop being stirred in a container of glycerin to introduce the stronger notion of a *mixing system* in his efforts to understand the approach of a non-equilibrium ensemble distribution function to an equilibrium distribution [4,8]. In considering a non-equilibrium phase space distribution one has to follow the trajectories of a set of points in phase space, not just a typical trajectory, as in the study of ergodic behavior. Gibbs suggested that the an initial set of points concentrated in a small region of phase space might spread out over the entire available phase space in the course of time, and become finely mixed throughout the phase space in much the same way as the ink drop will become mixed in the glycerin, in such a way that a coarse grained observation of the ink would lead to the conclusion that it is uniformly mixed, but a fine grained observation would reveal the individual threads of ink. Returning to phase space, Gibbs argued that although the measure of phase space occupied by the set of points should remain constant in time, in accordance with Liouville's theorem, the set eventually gets distributed over the constant energy surface such that a coarse grained observation of the phase space would lead to the conclusion that the set uniformly covers the energy surface, while a fine grained observation would reveal that the coverage consists of one long, thin set of total measure much less than that of the whole energy surface. Of course, mechanical reversibility ensures that one can recover the initial distribution of points by time reversing the motion but eventually mixing takes place for the time reversed motion, as well.

The mathematical definition of a mixing system [30] is given by looking at the time development of the initial set of phase points. We denote by A, the initial set of phase points on which the initial ensemble is concentrated, and

the set to which this initial set evolves after a time T, by A_T. Then, to examine the mixing of the set in phase space, we consider some arbitrary set of positive measure B and consider the intersection $B \cap A_T$. If the dynamics of the system is mixing, then eventually the fraction of the set B occupied by A_T should approach the fraction of A in the entire phase space, no matter what sets of positive measure A and B we consider. That is, a system is mixing if for any sets A and B of positive measure

$$\lim_{T\to\infty} \frac{\mu(A_T \cap B)}{\mu(B)} = \frac{\mu(A)}{\mu(\mathcal{E})} \,. \tag{10}$$

Here we used the conservation of phase space measure, namely that $\mu(A) = \mu(A_T)$. One can easily prove that a mixing system is also ergodic, but the reverse need not be true [30].

For a system of particles with the mixing property one can prove that a non-equilibrium phase space distribution will approach an equilibrium phase space distribution in a *weak*, long time limit [8]. That is, average quantities taken with respect to the non-equilibrium distribution approach equilibrium averages. The proof can easily be constructed by approximating integrals by sums in the usual way.

Mathematicians have considered proofs of the mixing property for various systems of interest to physicists. The most important of such systems are hard ball systems mentioned above, where the proofs of ergodicity and mixing are consequences of proving a stronger dynamical property, the Bernoulli property of the system, which implies that the system is mixing and therefore ergodic as well. We leave a proper definition of a Bernoulli system to the literature [34] and we will not consider it any further here. Needless to say, the class of systems which can be proved to be mixing is not yet large enough to encompass the typical systems studied in statistical mechanics or in kinetic theory, although considerable progress has been made in this direction over the past several years.

What is the Relevance of These Notions for Statistical Mechanics?

It would appear that with the proof that a system of a large number of particles in a reasonable container is mixing, the foundation for the applications of statistical mechanics to such a system would be secure. That this is not a complete explanation may be seen for the following reasons, among others:

1. We have assumed that our systems have fixed energies and that all the forces acting between the particles or on the system, due to the walls of the container, say, are conservative and known. Strictly speaking, this is not true of any laboratory system.
2. We have not examined how long it might take for the time average to be reasonably close to the ensemble average for an ergodic system or how long it would take a reasonable initial set to get uniformly mixed over the appropriate phase space, for a mixing system. One can argue that the appropriate times are very long, typically very much longer than the duration of an experiment. A partial but useful answer to this objection to the use of ergodicity and/or mixing properties is to consider Reduced distribution functions, particularly the single particle and two-particle distribution functions. These reduced distribution functions approach equilibrium, or more generally, local equilibrium forms, on much more realistic time scales.
The notion of local equilibrium arises in the context of the decay of a non-equilibrium state in a many particle system to total equilibrium, through hydrodynamic processes. In the usual picture, due to Chapman, Enskog, and Bogoliubov [35,36], the initial state becomes, on the time scale of the duration of a collision, one that can be described by reduced distribution functions. Then on the time scale of the time between collisions, set by the microscopic properties of the system such as the density and the size of the particles, the system becomes close to a state of local equilibrium with a local temperature, density and mean velocity. Finally, on a time scale set by the physical size of the container, the system relaxes to total equilibrium. The approach to local equilibrium is set by the dynamical interactions between the particles.
In addition, many non-equilibrium as well as equilibrium properties of a many-particle system can be formulated in terms of these reduced functions. This being the case, the questions are: What is the reason that reduced distribution functions approach local equilibrium forms, and what are the time scales involved in the approach to local equilibrium, and eventually to total equilibrium? For chaotic systems, at least, there is reason to believe that one can provide satisfactory answers to these questions.
Another approach to resolving the issue of time scales in applying these notions to laboratory systems, and one closely related to the use of reduced distribution functions is to focus attention on a set of microscopic variables called *normal variables* [1]. These variables are defined by the requirement that they not vary much over the constant energy surface. These variables then have the property that their value is almost the same

no matter where the phase point happens to be on the constant energy surface. In this picture an initial non-equilibrium state would be one where the values of the normal variables are far from their equilibrium or average values, but the time evolution of the system leads to regions where these variables have values closer to their equilibrium averages. In this picture the relevant time scale is the average time needed for the system to evolve to regions of phase space where the normal variables are close to their average values. Presumably this time will be much shorter than the time needed for the trajectory to cover the phase space. However, we need a detailed description of these variables and a description of their behavior as a system approaches local, then total equilibrium via kinetic and hydrodynamic processes, respectively. In the absence of such an understanding of normal variables, in general, the reduced distribution functions provide a clearer way to address the questions of time scales for the approach of distribution functions to their equilibrium values.

One can imagine, that both the reduced distribution resolution and the normal variable resolution of the time scale issue become easier to justify as the number of degrees of freedom, or the number of particles become large. In either case, although the phase space measure will also grow, and the time needed to cover it will increase, the relative fluctuations in the reduced distribution functions and normal variables will become smaller. It should be emphasized that the notions of ergodicity and mixing have an important role to play in statistical mechanics, despite the issues raised here concerning the time needed for covering the full phase space. Using ergodicity one can motivate the use of the micro-canonical ensemble and ultimately all of the Gibbs ensembles that provide methods for computing equilibrium properties of systems. Similarly, the mixing property can be used to show that time correlation functions decay in time, an important property needed for many applications of non-equilibrium statistical mechanics, particularly for the Green–Kubo formalism [2,3].

3. We have used classical mechanics as our description of dynamics, but nature is fundamentally quantum mechanical.

In the following sections we will address point (2) in some detail, and point (3), briefly. It is important to note, however, that the answers we shall provide, while encouraging, are very incomplete and much more work needs to be done to make these answers believable and secure. Here we consider point (1). As mentioned in the previous section, in order to produce a random walk of the Helfand moments needed for normal transport, some sort of stochastic or stochastic-like mechanism is required. Possible sources of stochastic behavior could be random external influences or some built-in structural randomness, however subtle and difficult to identify. Evidence for the effectiveness of built-in randomness is provided by a useful model of a statistical system, such as the Ehrenfest wind-tree model [24] where diamond shaped scatterers (trees) are placed at random in a plane, with their diagonals aligned parallel to the x- and y-axes. Then point particles (wind) move among the trees with velocities only in the $\pm x$ and $\pm y$ directions. The random placement of the trees provides a clear stochastic mechanism leading to normal diffusion and an approach to equilibrium and thereby simulate a mixing system. Such systems are not chaotic and some source of randomness beyond the dynamics is necessary for normal transport in such non-chaotic models. Moreover there are periodic versions of the wind-tree model that also exhibit normal diffusion [12,25], with more subtle sources of randomness, possibly connected with the splitting of two nearby trajectories at the corners of the scatterers. Below we turn our attention to the case of chaotic systems and argue that for such systems, the dynamics can be enough to allow normal transport and the approach of an ensemble of such systems to equilibrium, even when the system is spatially periodic.

Integrable, Pseudo-chaotic, and Chaotic Systems

The kind of systems one studies in courses on classical mechanics are typically *integrable system*. These are systems with N canonical coordinates, N canonical momenta, and N independent constants of the motion [2]. The canonical coordinates of such systems can be transformed to a set of action-angle variables, and when expressed in these variables, the systems become quasi-periodic in time, and the motion is confined to regions in phase space called invariant tori. Systems of coupled harmonic oscillators are good examples of integrable systems. There is another class of mechanical systems, of which the Ehrenfest wind-tree model described above is an example, called *pseudo-chaotic* systems [37]. These systems have the property that two infinitesimally close trajectories will separate algebraically with time, for sufficiently long times. For example, two distinct trajectories moving in initially parallel directions in a wind-tree system will eventually separate algebraically with time, no matter how close they are to each other initially, due to the random placement of scatterers and the fact that one of the two trajectories will eventually hit a scatterer that is "just missed" by

the other one. Pseudo-chaotic systems have a *weak mixing* property [30,37], namely that

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T d\tau \left[\mu(A_\tau \cap B) - \frac{\mu(A_\tau)\mu(B)}{\mu(\mathcal{E})}\right] = 0\,. \quad (11)$$

Chaotic systems are usually defined by the condition that two infinitesimally close trajectories will, in the appropriate long time limit, separate exponentially [4]. We will give a more careful discussion of chaotic systems in the next section.

Anosov and Anosov-like Systems; the Chaotic Hypothesis

Earlier in this discussion we argued that the notions of ergodicity and mixing are not, in themselves, fully sufficient for understanding the approach to thermodynamic equilibrium for systems of many particles due to problems of time scales. In addition we argued that in many cases reduced distribution functions may very well be better indicators of the approach to equilibrium than the full phase space distribution function. We also suggested that for chaotic systems at least, these observations can be verified to the extent that we can make some more detailed statements concerning the approach to equilibrium and transport properties. Here we will provide some justification for these comments.

It is helpful to consider simple model systems that display the kind of irreversible behavior that we would like to be able to describe for more realistic, many particle systems. Two model systems that have this feature are the baker's map [31,38] and the Arnold Cat Map, which itself is an example of more general models called hyperbolic toral automorphisms [14]. We will explain this terminology as we proceed.

The baker's map is a map of a two dimensional "unit square", $0 \le x, y \le 1$ onto itself, given by (see Fig. 1)

$$\begin{aligned} \begin{pmatrix} x' \\ y' \end{pmatrix} &= \begin{pmatrix} 2x \\ y/2 \end{pmatrix} \quad \text{for } 0 \le x \le 1/2\,; \\ \text{and} \quad &= \begin{pmatrix} 2x-1 \\ (y+1)/2 \end{pmatrix} \quad \text{for } 1/2 < x < 1\,. \end{aligned} \quad (12)$$

This map consists of a stretching of the square in the x-direction by a factor of 2, and squeezing in the y-direction by a factor of 1/2. The elongated image of the unit square is then cut in half and the right side is put on top of the left side so that a unit square is reconstructed, after each application of the map. Of course this is an area preserving map, and it has a discontinuity at $x = 1/2$.

Suppose the x, y-plane is our (two-dimensional) phase space and the motion of a phase point is confined to the unit square. We now apply some of the usual techniques of statistical mechanics to baker-map distribution functions defined on the unit square. The phase space distribution function changes at each time step according to

$$\begin{aligned} \rho_n(x, y,) &= \rho_{n-1}(x/2, 2y) \quad \text{for } 0 \le y \le 1/2\,; \\ \text{and} \quad &= \rho_{n-1}((x+1)/2, 2y-1) \quad \text{for } 1/2 < y < 1\,. \end{aligned} \quad (13)$$

For statistical mechanics, we often are interested in the reduced distribution functions of fewer variables. Here there are only two variables, so we consider the reduced distribution function, $W_n(x)$ obtained by integrating $\rho_n(x, y)$ over the y coordinate. Using Eq. (13), we obtain the equation [38]

$$W_n(x) = \frac{1}{2}\left[W_{n-1}\left(\frac{x}{2}\right) + W_{n-1}\left(\frac{x+1}{2}\right)\right]. \quad (14)$$

If one assumes that the initial value, $W_0(x)$, of the reduced distribution function is a reasonably smooth function of x, for example, if it can be represented by a convergent Fourier series, then $W_n(x)$ approaches a constant value as $n \to \infty$! The reason for this is that Eq. (14) says that the

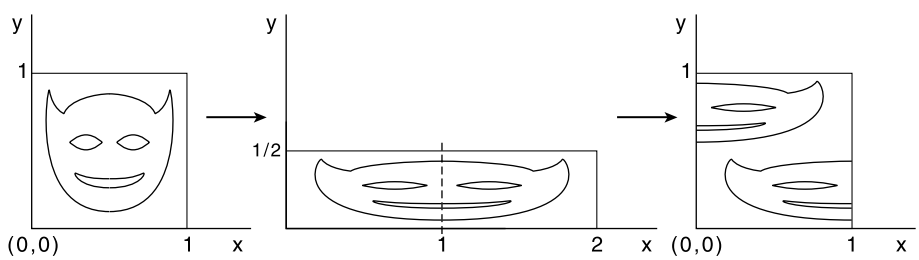

Chaotic Dynamics in Nonequilibrium Statistical Mechanics, Figure 1
The baker's map: The unit square is mapped onto itself by stretching by a factor *2* in the *x*-direction and by a compression by a factor of 1/2 in the *y*-direction. This is followed by cutting and re-arranging the resulting rectangle

reduced distribution at time n at a point x is the average value of the distribution at two points, $x/2$ and $(x+1)/2$ at the previous time, $n-1$. This averaging, if carried on long enough, produces a function which is constant in x. One can readily estimate the time it takes to reach this uniform state in the following way. Suppose that $l < 1$ is some length scale on which $W_0(x)$ varies with x. Then since the baker's map stretches sets of length l into sets of length $2l$ we can estimate the time to reach equilibrium as the time it takes a set of length l to be stretched to a set of length 1, which is $(-\ln l)/(\ln 2)$. This can be much shorter than the time it takes for the ergodic or mixing properties of the baker's map to make themselves manifest, which takes an additional time of order $\ln N/\ln 2$ to produce N horizontal strips of unit length, stacked in the y-direction. If we were to consider the reduced distribution function in the y variable, we would not get a nice equation with a distribution function that approaches equilibrium. Why that is so will be clear in a moment.

We have not introduced any stochastic features into our derivation of Eq. (14), but only integrated the Liouville equation over one of the variables. Thus we have been able to start from the Liouville equation, and by introducing nothing but an initial condition and integrating over the appropriate number of unmeasured variables, obtain an irreversible equation for a reduced distribution function which approaches equilibrium. Incidentally, the quantity $S = -\int_0^1 \mathrm{d}x\, W_n(x) \ln W_n(x)$ exhibits a monotonic increase with time n [8].

Our analysis of the approach to equilibrium of the function $W_n(x)$ made strong use of the stretching nature of the map in the x-direction. To get some further insight into the importance of the stretching of the phase space regions, we consider another model, the Arnold Cat Map [14] (see Fig. 2). Here the unit square, with opposite sides identified, represents a torus. This transformation maps the torus (that is, there is no cutting and moving of any section as there is in the baker's map) onto itself and may be described by a 2×2 matrix with unit determinant and integer elements. Such maps are called *toral automorphisms* [14], of which the Cat Map is a specific example. This map has an additional, *hyperbolic*, property, namely, that one of its eigenvalues be greater than unity. It follows from the fact that the determinant is unity, that the other eigenvalue is less than unity, and the product of the two is 1. The standard version of the map is given by the symmetric matrix

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \mathbf{T}\cdot\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\cdot\begin{pmatrix} x \\ y \end{pmatrix}, \quad \text{modulo } 1. \tag{15}$$

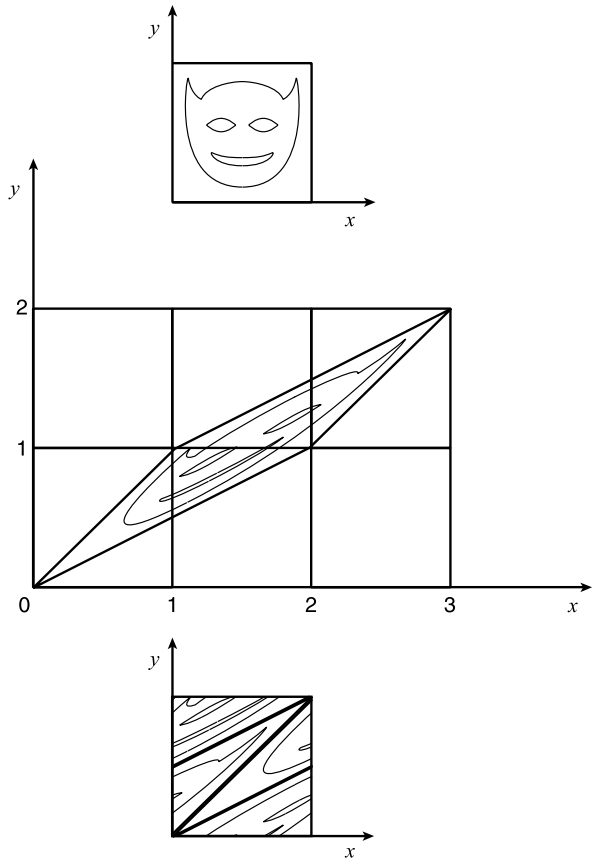

Chaotic Dynamics in Nonequilibrium Statistical Mechanics, Figure 2
The Arnold Cat Map of the unit torus onto itself

The eigenvalues of $\mathbf{T}$ are $(3 \pm \sqrt{5})/2$. The eigendirections are perpendicular to each other, and make non-zero angles with the x- and y-axes. There is a stretching direction which corresponds to the direction of the larger eigenvalue, and a contracting direction which corresponds to the smaller eigenvalue. While the equation for the projection of the "phase space" distribution function onto the x- or y-directions is not as simple as that for $W_n(x)$ given above, it is not difficult to write a computer program which shows the values of these projections after a few time steps, starting with some initial distribution on the unit torus. Figures 3 and 4 show the behavior of the projected distribution functions, $W_n(x)$ and $G_n(y)$, onto the x- and y-directions, respectively, for an initial set of points that is uniform over the region $0 \le x, y \le 0.1$ and zero everywhere else. Both these distribution functions become uniform after three or four time steps. In Figs. 4 and 5, we show the evolution of the phase space distribution function at some of the same times. Here we clearly see the difference on

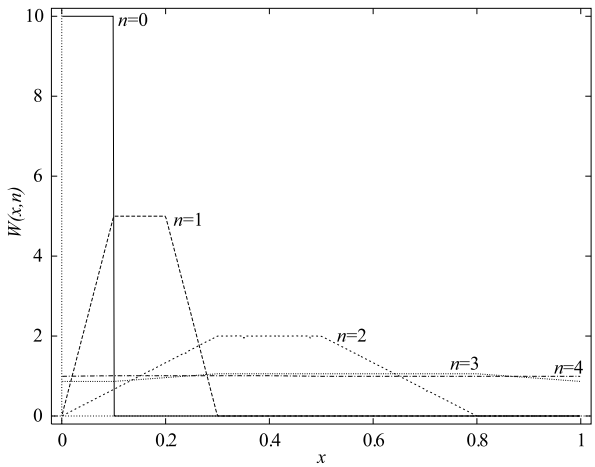

Chaotic Dynamics in Nonequilibrium Statistical Mechanics, Figure 3
The projection of the Cat Map distribution onto the x-axis, $W_n(x)$, at various times, n. By $n = 3, 4$ the distribution is essentially uniform

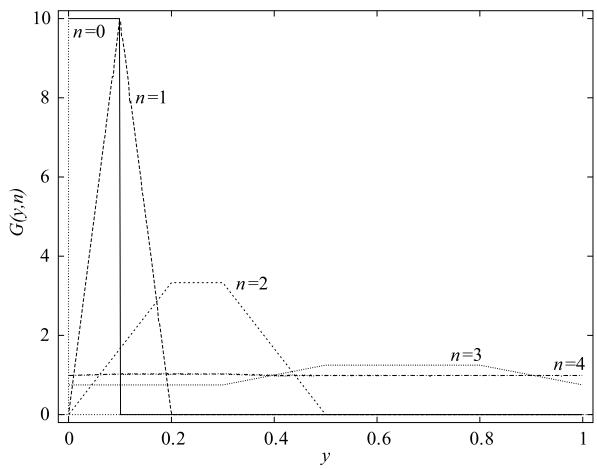

Chaotic Dynamics in Nonequilibrium Statistical Mechanics, Figure 4
The projection of the Cat Map distribution onto the y-axis, $G_n(y)$ at various times, n. By $n = 3, 4$ the distribution is essentially uniform

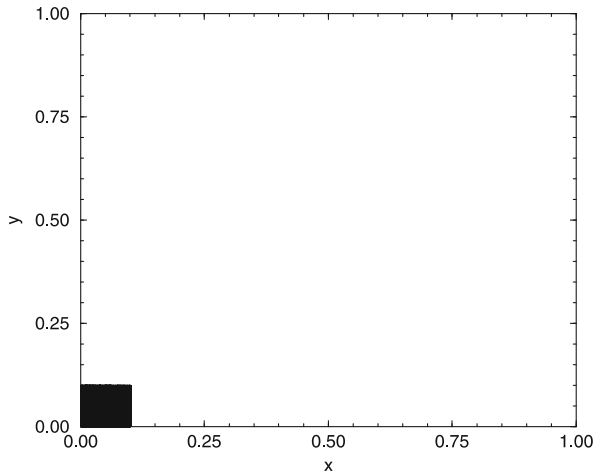

Chaotic Dynamics in Nonequilibrium Statistical Mechanics, Figure 5
The initial phase space distribution used to obtain Figs. 3 and 4

the rate of evolution of projected vs. full phase space distribution functions, and why, for simple models at least, the notions of ergodicity and mixing demand more than is physically required for an approach to equilibrium for the reduced distribution functions.

The important feature that both the baker's map and the cat map have in common is that they are both area preserving maps with a stretching direction, or *unstable* direction, and a contracting direction, called a *stable* direction [14]. These directions are associated with *stretching* and *contracting* factors that depend exponentially upon time. The logarithms of these factors per unit time are the *Lyapunov exponents* [4,6,14,39]. For the baker's map the two Lyapunov exponents are $\lambda_\pm = \pm \ln 2$, and for the cat map above, the two Lyapunov exponents are $\lambda_\pm = \ln[(3 \pm \sqrt{5})/2]$. In general, for higher dimensional systems, the stable and unstable directions span the so-called stable and unstable manifolds, respectively. For the baker's map the y-axis and the x-axis are stable and unstable manifolds, respectively, for the baker's map, as are the axes in the eigen-directions of the cat map. The importance of the unstable manifolds resides in the fact that on them non-uniform functions become smoothed with time, if they are not too singular, on a time scale that is determined by the positive Lyapunov exponents. In order to get a reduced distribution function that approaches an equilibrium distribution for long enough times, one clearly must project the Liouville distribution onto manifolds that are not orthogonal to all of the unstable directions. This is one reason to examine the Cat Map, namely, to show that projections onto either direction, x or y equally lead to a uniform distribution after a few time steps. This would be the case for the baker's map, if we were to project on any direction that makes a non-zero angle with the y-direction. Generally, the unstable directions in phase space are so complex that practically any projected distribution will not be orthogonal to them. Under time reversal the unstable and stable manifolds are interchanged, and the picture remains essentially the same. It is worth mention-

ing that if we were to project the distribution function onto a stable manifold we would not obtain an equilibrium distribution function, but rather a very complicated function since any irregularities in the initial distribution function become more irregular with time, along stable directions [6,8,40,41]. This is why a projection onto the y-axis would not lead to an equilibrium distribution for the baker's map, without some sort of coarse graining to "smooth out" an otherwise singular function.

Dynamical systems that have properties similar to the cat map or the baker's map are called *Anosov* or *Anosov-like* systems, respectively [6,14,18]. *Anosov* systems are continuous dynamical systems such that at every point in phase space there are stable and unstable directions, with positive and negative Lyapunov exponents, respectively. There may also be some *neutral* directions with zero Lyapunov exponents, but there must be at least one positive Lyapunov exponent. The presence of positive Lyapunov exponents is characteristic of a *hyperbolic* system. Furthermore there must be at least one trajectory in phase space that is dense, a requirement which is called *transitivity*. Systems of particles with some manageable singularities such as hard spheres are not strictly Anosov systems but are close enough to be considered *Anosov-like*. The Arnold Cat Map is an example of an Anosov system, while the baker's map is Anosov-like. Of course in a multidimensional Anosov system, there are a number of stable and unstable directions, and all stable manifolds intersect all unstable manifolds transversally. Systems with at least one positive Lyapunov exponents are commonly referred to as being *chaotic*. We have chose simple systems to illustrate the main ideas, but the reader should be aware that it is possible to define Lyapunov exponents for motion on invariant sets in phase space, even for sets of measure zero, such as periodic orbits, or as we discuss below, fractal repellers.

Returning to statistical mechanics, we see that the simple examples of the baker's map and the Cat Map give us some reason to believe that, for chaotic systems at least, reduced distribution functions will approach equilibrium values, or more precisely, local equilibrium values, on reasonable time scales. This conclusion depends, of course, on the assumption that the underlying microscopic dynamics is of the Anosov or Anosov-like variety. It is therefore useful to assume that systems of large numbers of particles with short ranged interactions, of general interest for statistical mechanics, are ergodic, Anosov-like systems as far as their dynamical properties are concerned. This assumption has been made a central feature of the dynamical systems approach to non-equilibrium statistical mechanics by Gallavotti and Cohen [18], who have called it *the chaotic hypothesis*. This hypothesis allows one to apply the results of chaotic dynamics, such as those discussed in the next section, to realistic systems. In actuality we know little a priori about the dynamics of such systems, such as their Lyapunov exponents, or the structure and properties of unstable and stable manifolds in phase space. Furthermore, due to the work of Turaev, Rom-Kedar, Donnay [28,29], and others, we know that if the intermolecular potential is smooth, there may exist regions in phase space where the dynamics is not chaotic. For these systems, the chaotic hypothesis implies the additional statement that the total volume of the non-chaotic regions in phase space is small enough so that for determining the important averages, they can be ignored.

Fractal Dimensions

An important concept used in applications of dynamical systems theory to statistical mechanics, among many other applications, is the notion of *fractal dimensions*. As we will mention in the next section, for chaotic systems many of the connections between macroscopic transport coefficients and microscopic dynamical quantities such as Lyapunov exponents can be related to the dimensions of fractal structures that form in the phase spaces of nonequilibrium systems. Generally, but not exclusively, fractals are defined by the statement that their dimension is not an integer. However, this definition is complicated by the fact that there are many ways to define the dimension of a set, and some fractals may have a dimension that is an integer by some definition and non-integer by others. Fractals that have a variety of different dimensions are referred to as *multifractals*. There are also other definitions of fractals based upon self-similarity at all scales, or upon differentiability and continuity properties. We refer to the literature for more details [42].

In order to present one definition of the fractal dimension of a set, imagine that the set is embedded an a d-dimensional space. Cover the set with cubes of length ϵ on a side, each labeled by an index i and suppose that one needs at least $N(\epsilon)$ of these cubes for a full coverage. Suppose further that we have a way to assign a measure μ_i to each cube. We can suppose that the total measure is normalized to some constant, say unity,

$$\sum_{1}^{N(\epsilon)} \mu_i = 1 \,. \tag{16}$$

The measure μ_i may be a normalized Lebesgue measure of the cube, or it may be the fraction of time that a typical trajectory on the fractal spends in cube i, for example. In

terms of this measure we can define a dimension D_q that depends on a continuous variable q, as

$$D_q = \frac{-1}{1-q} \lim_{\epsilon \to 0} \frac{I(q,\epsilon)}{\ln \epsilon} , \tag{17}$$

where $I(q,\epsilon)$ is given by

$$I(q,\epsilon) = \sum_{i=1}^{N(\epsilon)} \mu_i^q . \tag{18}$$

In this construction the quantity, D_0 is called the *box-counting dimension*, D_1 is called the *information dimension*, and D_2 is called the *correlation dimension*. The dimension D_q is a monotonic non-decreasing function of q. Other dimensions frequently employed to characterize fractals include the *Hausdorff dimension*, and we refer to the literature for a careful definition and discussion of this and other dimensions. As an example of a fractal, consider the *middle third Cantor Set*. This set is constructed by taking a unit interval, discarding the middle third of it, then take each of the two remaining pieces, discarding the middle third of each of them, and so on. It is easy to show that the dimensions, D_q, of the remaining set, are all equal to $\ln 2/ln3$.

Applications of Dynamical Systems Theory to Non-equilibrium Statistical Mechanics

The connections between normal transport, random walk processes, and the generalized Einstein formulae have allowed results from dynamical systems theory to be applied to non-equilibrium statistical mechanics in a direct way, at least for chaotic systems. However, one must study transport processes in a way that takes advantage of the fundamental properties of chaotic dynamics. Among the useful properties of chaotic systems are the formation of fractal structures in phase space. If the phase space is of sufficiently low dimensions then the fractal structures can be studied both by analytical methods and by computer simulated molecular dynamics. In addition to providing some insights into the approach to equilibrium for large systems, the dynamical systems approach gives us deep results of a more practical sort. Among these results are: (i) relations between transport coefficients and dynamical quantities, such as Lyapunov exponents and Kolmogorov–Sinai entropies, to be defined below [5,6,7,8,9,12,42,43,44,45,46]; (ii) a theory of entropy production in non-equilibrium steady states and in the approach of a system to equilibrium [5,6,7,8,9,10,16,17,18,20,40,49,50,51,52,53,54]; and (iii) a number of fluctuation theorems – Evans–Searles–Gallavotti–Cohen theorems – which describe fluctuations in entropy production in non-equilibrium steady states [16,17,18,19,20,21,52,55]. Here we will summarize these results, and we refer the reader to the literature for a more complete discussion of these and related topics.

Microscopic Dynamical Quantities and Macroscopic Transport Coefficients

Among the many reasons dynamical systems theory attracted the attention of workers in non-equilibrium statistical mechanics were the connections between macroscopic transport coefficients of a system of particles and quantities that characterize the system's chaotic behavior discovered by Gaspard and Nicolis [6,8,43,44,45,46,47], for Hamiltonian systems, and by Evans, Morriss, Hoover, Posch, and coworkers, for dissipative systems with Gaussian thermostats [5,7,9,12,48,49]. We discuss each case separately.

The Escape-Rate Formulae for Transport Coefficients

Previously we noted that the Helfand moments related to transport undergo a random walk motion in phase space. Gaspard and Nicolis pointed out that if one considers a random walk in a space with *absorbing* boundaries one can combine results from random walk theory and from dynamical systems theory to obtain very interesting previously unknown connections between transport coefficients and quantities that characterize the microscopic chaotic dynamics.

The random walk theory that one needs for this connection is based on a Fokker–Planck equation for the probability distribution of the Helfand moments, $P(M_\sigma, t)$, where M_σ is a Helfand moment. We assume that the transport process is normal, that is, that the mean square fluctuations of the Helfand moment grows linearly with time. In such a situation the Fokker–Planck equation takes the simple form of a diffusion equation:

$$\frac{\partial P(M_\sigma, t)}{\partial t} = \alpha \frac{\partial^2 P(M_\sigma, t)}{\partial M_\sigma^2} \tag{19}$$

Here α is a constant proportional to the transport coefficient σ. Suppose now that we solve this diffusion equation in M-space with boundary conditions that when M_σ reaches the values $\pm\chi/2$, $P(M_\sigma, t) = 0$. That is, the Helfand moments undergo a Brownian motion in a space with absorbing boundaries such that the probability distribution vanishes whenever $|M_\sigma|$ reaches a specific value for the first time. Under these conditions, the probability distribution will decay exponentially, with a decay rate γ

given by

$$\gamma = \alpha \left(\frac{\pi}{\chi}\right)^2 . \tag{20}$$

A remarkable result from dynamical systems theory is that for chaotic systems, another expression holds for the same escape rate in terms of Lyapunov exponents and a quantity called the Kolmogorov–Sinai entropy per unit time [6,14,56,57]. This microscopic escape-rate formula for the escape-rate, γ_{mic} is

$$\gamma_{\text{mic}} = \sum_i \lambda_i^+(\mathcal{R}) - h_{\text{KS}}(\mathcal{R}) . \tag{21}$$

All of the terms on the right hand side of Eq. (21) require some explanation. To get some insight into this formula, we now consider the microscopic dynamics of the system of particles and imagine that there is a set of initial conditions for all the particles for which the Helfand moment is in the region $-\chi/2 < M_\sigma < \chi/2$, and such that as the initial system evolves in time, either forward or backward, the Helfand moment will never reach the boundary. The set of all such initial conditions is called a *repeller* and denoted by $\mathcal{R}$. It is typically a fractal set of points in phase space, usually of measure zero in the set of all initial phase points, and a highly unstable set since an arbitrarily small displacement of the initial phase of the system from a point in $\mathcal{R}$ will lead to escape, unless the new initial point is also on the repeller. Now imagine that we consider an infinitesimally small set of points in $\mathcal{R}$, and observe how these points separate in phase space in the course of time. If the dynamics on $\mathcal{R}$ is Anosov-like, there will be a set of positive Lyapunov exponents which we have denoted by $\lambda_i^+(\mathcal{R})$ in Eq. (21). The positive Lyapunov exponents on the repeller describe the rate with which trajectories on the repeller move apart. apart. The other term in the escape-rate formula, $h_{\text{KS}}(\mathcal{R})$, is the Kolmogorov–Sinai entropy per unit time of the trajectories on $\mathcal{R}$, and can be understood in the following way. In general, dynamic entropies characterize the rate at which information about the exact trajectory of a system is gained or lost in the course of time. For example, if we know that the initial phase of a system is within some small region of phase space, then the *stretching* of the small region with time due to the dynamical instability allows us to locate the initial location of the trajectory ever more precisely as we follow the motion of the small initial set. If the system is closed, that is, there is no possibility of escape, then the amount of information about the initial location of the trajectory grows exponentially as the sum of the positive exponents. This result is known as Pesin's theorem. On the other hand, if the system is open and there is a possibility of escape, then the escaping trajectories lead to a loss of information [58]. Hence a better way to write Eq. (21) is

$$h_{\text{KS}}(\mathcal{R}) = \sum_i \lambda_i^+(\mathcal{R}) - \gamma_{\text{mic}} . \tag{22}$$

In any case we now have two expressions for the same escape-rate for a hyperbolic system, and by equating them, we obtain an expression for the transport coefficient in terms of microscopic dynamical quantities, as

$$\alpha = \lim_{\chi\to\infty} \frac{\chi^2}{\pi^2} \left[\sum_i \lambda_i^+(\mathcal{R}) - h_{\text{KS}}(\mathcal{R})\right] . \tag{23}$$

We have taken the large system limit to remove any possible dependence of the right hand side of Eq. (23) on the shape of the boundaries or on the size of the system. This result is due to Gaspard and Nicolis [43].

For two dimensional systems, one can reformulate the escape rate formula in terms of the information dimension, d_1, of the repeller along the unstable direction [6,59], which is given by $d_1 = h_{\text{KS}}/\lambda^+$, so that $\gamma = \lambda^+(1 - d_1)$.

The escape-rate formula has been used in molecular dynamics studies to obtain values for transport coefficients in a number of cases such as diffusion coefficients in periodic Lorentz gases, viscosities of simple systems and chemical reaction rates [27,47,60,61,62]. The results for the transport coefficients obtained by using the escape-rate methods agree with those obtained by other methods, often based on the Green–Kubo time correlation method, or, as discussed below, on Gaussian thermostat methods. There are as yet very few theoretical methods to compute the Kolmogorov–Sinai entropy for the repeller [6,27,47], while both the transport coefficients, and, as it turns out, the sum of the positive Lyapunov exponents on the repeller are sometimes amenable to treatment by the usual methods of statistical mechanics and kinetic theory [63,64,65].

Transport Coefficients and Phase-Space Contraction in Gaussian Thermostatted Systems An alternative method for relating macroscopic transport coefficients to Lyapunov exponents was developed by Evans, Hoover, Posch and co-workers in their work on developing computer algorithms to simulate non-equilibrium flows in many-particle systems [5,7,9,48,49,66,67,68,69,70,71,72,73,74,75]. A problem arose in these simulations because the systems tended to heat up considerably due to the presence of viscous friction or, for charged particle systems, ohmic heating. To counteract this heating, these authors introduced a fictitious thermostat which maintains a constant value for either the total energy or the total kinetic

energy of the system. The thermostat was introduced by a modification of the equations of motion by the addition of a frictional force in such a way that the Hamiltonian nature of the system, in the usual coordinate, momentum, and time variables, is lost. This kind of frictional force was first introduced by Gauss in his study of mechanical systems with various constraints.

For systems with Gaussian thermostats, Liouville's theorem, which states that the phase space density at a point moving under the equations of motion does not change with time, is no longer satisfied. It is replaced by a conservation equation that relates the time derivative of the distribution function to the parameters characterizing the frictional forces. As a consequence phase space volumes for thermostatted systems do not remain constant in time, but rather, on the average the phase space volume of a set decreases with time, the system approaches a *non-equilibrium* stationary state, and the system's distribution function approaches a distribution with fractal properties. The average rate of decrease of the phase space volume is given by a negative value of the sum of all of the Lyapunov exponents of the system. The non-equilibrium stationary state distribution is characterized by a special type of measure, called a Sinai–Ruelle–Bowen (SRB) measure, which describes the probability of finding a system in different regions of phase space. The SRB measure is characterized by the fact that it is smooth in the unstable directions in phase space but typically is fractal in the stable directions [14]. Simple examples of this type of measure can be found in the literature [41]. The resulting fractal is called an *attractor*.

It is important to note that the decrease of the phase space volume and the concomitant formation of an attractor with a non-trivial SRB measure, does not mean that the fractal has zero dimension. Indeed due to the decrease in phase space volume the *Lebesgue measure* of the phase space region where the trajectories are located approaches zero. However the dimension of the resulting fractal is not zero. The box counting dimension, D_0, of the fractal may even coincide with the dimension of phase space, and the information dimension, D_1, can often be expressed in terms of the Lyapunov exponents. For example for two-dimensional, thermostatted systems, with one positive Lyapunov exponent, λ^+ and one negative exponent, $-|\lambda^-|$, with $\lambda^+ - |\lambda^-| < 0$, the information dimension of the attractor is given by the *Kaplan–Yorke–Young formula* [4,42,69,71] for two dimensional ergodic systems, as $D_1 = 1 + \lambda^+/|\lambda|$.

The decrease of the phase space volume can also be related to the entropy production in the system+thermostat. This entropy production can also be related to the transport coefficients via the usual macroscopic laws of irreversible thermodynamics. Thus one has a way to relate transport coefficients to the sum of the Lyapunov exponents for these thermostatted systems.

To make this discussion more concrete, we consider the example of a hard-ball Lorentz gas [53,69,70,74,75]. We assume that the moving particles have a bounded free path between collisions with the fixed scatterers, and that the scatterers do not form traps without escape for the moving particles. We suppose the moving particles have mass. m, and carry a charge, q, (but still do not interact with each other) and are placed in a constant, external electric field, $\boldsymbol{E}$. Ordinarily, the field will accelerate the moving particles in such a way that their energy increases over time. To avoid this we add a frictional force to the equations of motion of the moving particle, and adjust the frictional force so as to keep the kinetic energy of the moving particle constant in time. The equations of motion for the position, $\boldsymbol{r}$ and velocity, $\boldsymbol{v}$, of the moving particle between collisions with scatterers are

$$\dot{\boldsymbol{r}} = \boldsymbol{v} \tag{24}$$

$$\dot{\boldsymbol{v}} = \frac{q\boldsymbol{E}}{m} - \alpha \boldsymbol{v}\,, \tag{25}$$

where α is a dynamical variable defined by the constant energy condition, $\dot{\boldsymbol{v}} \cdot \boldsymbol{v} = 0$. Thus we find that

$$\alpha = \frac{q\boldsymbol{v} \cdot \boldsymbol{E}}{m\boldsymbol{v}^2}\,. \tag{26}$$

The equations of motion given above for the moving particle must be supplemented by the equations describing the collisions of the moving particles with the scatterers. To simplify matters, we will suppose that the particles make instantaneous, elastic, and specular collisions with the scatterers.

If we define the Gibbs entropy for this system in terms of the phase space distribution function, $\rho(\boldsymbol{r}, \boldsymbol{v}, t)$, by

$$S_G(t) = -k_\mathrm{B} \int \mathrm{d}\boldsymbol{r} \int \mathrm{d}\boldsymbol{v} \rho(\boldsymbol{r}, \boldsymbol{v}, t)[\ln \rho(\boldsymbol{r}, \boldsymbol{v}, t) - 1]\,, \tag{27}$$

where k_B is Boltzmann's constant, one finds that

$$\frac{\mathrm{d}S_G}{\mathrm{d}t} = -k_\mathrm{B} \int \mathrm{d}\boldsymbol{r} \int \mathrm{d}\boldsymbol{v} \alpha \rho = -k_\mathrm{B} \langle \alpha \rangle\,. \tag{28}$$

The average value is taken with respect to the phase space distribution function for the moving particles, ρ. This entropy actually decreases with time, since on the average α is positive. This decrease in entropy must be matched, at

least, by an increase in entropy of the reservoir that is responsible for the additional frictional force. Thus

$$\frac{\mathrm{d}S_{\text{reservoir}}}{\mathrm{d}t} \geq k_{\mathrm{B}}\langle\alpha\rangle \geq 0\,. \tag{29}$$

We now take the entropy production in the reservoir to be given by the usual macroscopic laws, in particular,

$$\frac{\mathrm{d}S_{\text{reservoir}}}{\mathrm{d}t} = \frac{\boldsymbol{J}\cdot\boldsymbol{E}}{T} = \frac{\sigma\boldsymbol{E}^2}{T}\,, \tag{30}$$

where $\boldsymbol{J} = \sigma\boldsymbol{E}$ is the electrical current produced by the moving particles and σ is the coefficient of electrical conductivity. Then, by combining Eqs. (29) and (30), and assuming that we take the entropy production in the reservoir to be exactly $k_{\mathrm{B}}\langle\alpha\rangle$, we obtain and expression for σ as

$$\sigma = \frac{k_{\mathrm{B}}T\langle\alpha\rangle}{\boldsymbol{E}^2}\,. \tag{31}$$

The final step in this process is to relate the average friction coefficient, $\langle\alpha\rangle$, to the Lyapunov exponents for the trajectories of the moving particles. We note, first of all, that the volume, $\mathcal{V}$ of a small region in phase space will change exponentially in time with an exponent equal to the sum of all the Lyapunov exponents, as

$$\mathcal{V}(t) = \mathcal{V}(0)\exp\sum_i \lambda_i\,, \tag{32}$$

since the Lyapunov exponents describe the rates of stretching or of contracting of small distances in phase space. A simple argument shows that

$$\left\langle\frac{\mathrm{d}\ln\mathcal{V}}{\mathrm{d}t}\right\rangle = -\langle\alpha\rangle = \left\langle\sum_i \lambda_i\right\rangle\,. \tag{33}$$

In the non-equilibrium stationary state, all averages remain constant with time, so that we can now use Eq. (31) to obtain the desired connection between a transport coefficient, in this case the conductivity σ, and the average Lyapunov exponents,

$$\sigma = -\frac{k_{\mathrm{B}}T}{\boldsymbol{E}^2}\left\langle\sum_i \lambda_i\right\rangle\,. \tag{34}$$

This result has been used to determine the conductivity and the diffusion coefficient of the moving particles in the Lorentz gas by means of efficient computer algorithms for determining the Lyapunov exponents [70,74,75]. Similar methods have been used to determine the shear viscosity [48] and other transport properties of systems with short range, repulsive potentials.

One important result of the analysis of these thermostatted systems is the *conjugate pairing rule*. This is a generalization of the result for chaotic Hamiltonian systems that non-zero Lyapunov exponents come in pairs, with the same magnitude but with opposite signs so that members of each conjugate pair sum to zero [42]. This is called the *symplectic conjugate pairing rule*. For many chaotic systems with Gaussian thermostats it is possible to prove another conjugate pairing rule where the Lyapunov exponents form conjugate pairs which sum to a non-zero value, independent of which pair of exponents is chosen [48,76,77]. The conjugate pairing rule is very helpful for analyzing relations such as that given by Eq. (34), since the right hand side is completely determined by the sum of any one conjugate pair of exponents. The easiest pair to use is the pair formed by the positive and negative exponents with the largest magnitude.

Ruelle–Pollicott Resonances and Irreversible Processes

For a Hamiltonian N-particle system, the phase space distribution function, $\rho(\Gamma, t)$ evolves in time according to the relation given by the Liouville equation, in terms of a time evolution operator, $S(-t)$, as

$$\rho(\Gamma, t) = S(-t)\rho(\Gamma, 0) = \exp -t\mathcal{L}\rho(\Gamma, 0) = \rho(\Gamma(-t), 0), \tag{35}$$

where we indicate that the phase space variables Γ evolve as

$$\Gamma(t) = S(t)\Gamma = \exp t\mathcal{L}\Gamma\,. \tag{36}$$

Here $\mathcal{L}$ is the Liouville operator, when acting on differentiable functions, is given, for an isolated system of N particles, by

$$\mathcal{L} = \sum_{i=1}^{N}\boldsymbol{p}_i\cdot\frac{\partial}{\partial\boldsymbol{r}_i} + \sum_i^N \boldsymbol{F}_i\cdot\frac{\partial}{\partial\boldsymbol{p}_i}\,, \tag{37}$$

where $\boldsymbol{F}_i$ is the total force on particle i due to the other particles, and external sources such as walls of a container.

If one insists that the functions on which the time displacement operator acts be ordinary functions with well behaved derivatives, the exponential operators, S appearing in Eqs. (35), (36) are unitary with spectrum on the unit circle. In view of this observation it is difficult to see how a satisfactory theory of irreversible behavior, including decays in time, might be obtained from such an operator. The answer is to be found by enlarging the space of functions to include singular functions such as Schwartz

distributions [6]. The discussion in the previous subsection has indicated that distributions may evolve to fractals, which are typically non-differentiable functions, so it is very natural to include singular functions in the space of functions on which the time evolution operator may act. In this enlarged function space, the time evolution operator may have properties that do not obtain in a more restricted space of functions. Under these circumstances one studies the right and left eigenfunctions of the Liouville operator, $\mathcal{L}$ in a structure called a *Gelfand triplet* or *rigged Hilbert space* [78]. There are two sets of such triplets corresponding to forward and time reversed motion.

A Gelfand triplet is an operator, together with the two spaces spanned by its right and left eigenfunctions, and an inner product involving one function from each of the two spaces. If the right eigenfunctions are singular functions, the left eigenfunctions must be sufficiently well behaved so that the inner product of a function in the "singular" space with a function in the "smooth" space is well defined. The eigenvalues for the time displacement operator appear as two different sets of poles of the resolvent operator, $(z-\mathcal{L})^{-1}$, in the complex plane, one for forward and one for time-reversed motion. The poles are called *Ruelle–Pollicott* resonances [79,80,81,82], and they give the relaxation rates for various processes that can take place in the system. The time reversal symmetry of the motion provides the relation between the two sets of resonances in the complex plane. For simple models it is possible to construct the functions, singular and smooth, in a Gelfand triplet and to locate the resonances in the complex plane [83,84,85]. In addition to poles of the resolvent one may, in general, expect to find branch cuts, etc. in the complex plane [6]. The existence of these resonances provides an argument that the rates of relaxation to equilibrium found by using traditional methods of statistical mechanics, including kinetic theory, are not artifacts of the approximations made in arriving at these results, but are reflections of the existence of Ruelle–Pollicott resonances for the system.

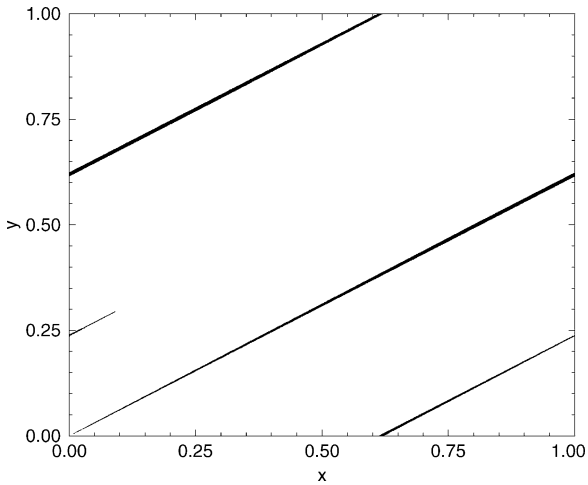

Chaotic Dynamics in Nonequilibrium Statistical Mechanics, Figure 6
The evolution of the phase space distribution of Figure 5 after three iterations of the Arnold Cat Map

Fractal Forms in Diffusion and a Dimension Formula

For chaotic systems, the relation between the mean square displacement of a diffusing particle and the diffusion coefficient conceals a fractal function. This can be understood on the basis of the following observation: Since the actual displacement of the diffusing particle depends on the initial phases of the particles in the system, even the slightest change in these phases will make a large change in the displacement of the Brownian particle. One way to try to capture some features of this variation of the displacement is to consider simple models of diffusion that are deterministic, chaotic, and diffusive [86,87]. The process of diffusion in these models is often referred to as deterministic diffusion, since no stochastic elements are introduced in the dynamics of the models. Here we illustrate this idea using the two dimensional periodic Lorentz gas with finite free paths of the moving particle between collisions, so that diffusion is well defined in this system [86,87,88].

It is very convenient to use the van Hove intermediate scattering function as a way of describing the diffusive process. This function, denoted by $F_{\boldsymbol{k}}(t)$, is defined as

$$F_{\boldsymbol{k}}(t) = \left\langle \mathrm{e}^{[i\boldsymbol{k}\cdot(\boldsymbol{r}(t)-\boldsymbol{r}(0))]} \right\rangle , \tag{38}$$

where the average is taken over an equilibrium ensemble. The van Hove function is related to the probability of finding the diffusing particle at point $\boldsymbol{r}$ at time t, $P(\boldsymbol{r}, t)$ by

$$P(\boldsymbol{r}, t) = \int \frac{\mathrm{d}\boldsymbol{k}}{(2\pi)^{\mathrm{d}}} P_{\boldsymbol{k}} F_{\boldsymbol{k}}(t) , \tag{39}$$

where d is the number of spatial dimensions of the system and $P_{\boldsymbol{k}}$ is the Fourier transform of the initial probability distribution for the diffusing particle. One can use a cumulant expansion in the exponent to show that $F_{\boldsymbol{k}}(t)$ is given by

$$F_{\boldsymbol{k}}(t) = \mathrm{e}^{s(k)t} , \tag{40}$$

where $s(k)$ is the wave-number dependent decay rate for diffusive motion. Thus, for wave numbers, k much smaller

than the inverse of a lattice spacing, in this case, unity, this function takes the form

$$s(k) = -Dk^2 + \tilde{D}k^4 + O(k^6)\,, \tag{41}$$

where D is the diffusion coefficient, $\tilde{D}$ is called the super-Burnett diffusion coefficient, etc. In order to see the physics that is obscured by taking the equilibrium average, consider the microscopic quantity that is averaged, namely, $\exp\left[i\boldsymbol{k}\cdot(\boldsymbol{r}(t)-\boldsymbol{r}(0))\right]$. If the motion of the particle is chaotic, the displacement over a time t is a very rapidly varying function of the initial point $\boldsymbol{r}(0)$. As a result, the exponential containing this displacement will be a rapidly oscillating function of the initial point. To capture these oscillations and to explore the fractal structure of the exponential function, we consider a partially averaged quantity, which we will call a normalized, incomplete van Hove function, $F_{\boldsymbol{k}}(\theta, t)$, defined by

$$F_{\boldsymbol{k}}(\theta, t) = \frac{\int_0^{\theta} \mathrm{d}\theta' \mathrm{e}^{[i\boldsymbol{k}\cdot(\boldsymbol{r}(\theta',t)-\boldsymbol{r}(\theta',0))]}}{\int_0^{2\pi} \mathrm{d}\theta' \mathrm{e}^{[i\boldsymbol{k}\cdot(\boldsymbol{r}(\theta',t)-\boldsymbol{r}(\theta',0))]}}\,. \tag{42}$$

Here we take initial points to be uniformly distributed just outside the surface of one of the scatterers in the periodic Lorentz gas, with initial velocity directed radially outward. The point on the surface is indicated by the angle θ taken with respect to some fixed direction. If one waits for a sufficient number of collisions to take place so that the motion is diffusive, and then plots $\mathrm{Im}\,F_{\boldsymbol{k}}(\theta, t)$ vs $\mathrm{Re}\,F_{\boldsymbol{k}}(\theta, t)$ one finds a fractal curve for small values of the wave number, $|\boldsymbol{k}|$. Computer results obtained by Claus et al. [89] for the hard disk Lorentz gas are plotted in Fig. 7 for various values of the wave number.

It is possible to prove that there exists a striking connection between the fractal Hausdorff dimension, $\mathcal{D}_H$, of the curve $(\mathrm{Re}\,F_{\boldsymbol{k}}(x), \mathrm{Im}\,F_{\boldsymbol{k}}(x))$, for the incomplete van Hove function [87,88], for small k, the diffusion coefficient, D, and the positive Lyapunov exponent, λ, characterizing the chaotic process underlying the diffusive motion of the particle. This connection is illustrated in Fig. 8 and given by the equation

$$\mathcal{D}_H(k) = 1 + \frac{D}{\lambda}k^2 + O(k^4)\,. \tag{43}$$

Also illustrated in Fig. 8 is the analogous curve obtained from the incomplete van Hove function when the hard disk potential is replaced by a repulsive Coulomb potential. For sufficiently high energy of the moving particle, the motion of the moving particle is chaotic and the van Hove function also shows fractal properties. Thus the fractal curves are not artifacts of the hard disk potential. Equation (43) illustrates the fact that for chaotic models such as the one discussed here, the incomplete van Hove function encodes in its structure both a macroscopic property of the system, the diffusion coefficient, and a microscopic property, the Lyapunov exponent. Such interesting connections are central to a deeper understanding of the microscopic foundations of transport processes.

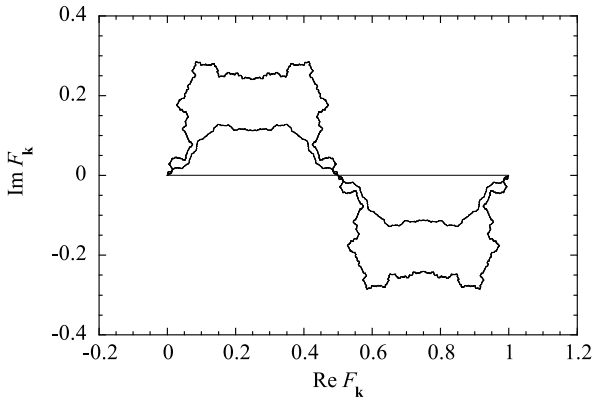

Chaotic Dynamics in Nonequilibrium Statistical Mechanics, Figure 7

The incomplete van Hove function for a periodic Lorentz gas where a point particle of unit mass and velocity undergoes elastic collisions with hard disks of unit radius forming a triangular lattice with interdisk distance $d = 2.3$: Curves of the cumulative functions for wavenumber $k_x = 0.0$, 0.5, and 0.9 with $k_y = 0$. Note that the fractality increases with the wavenumber

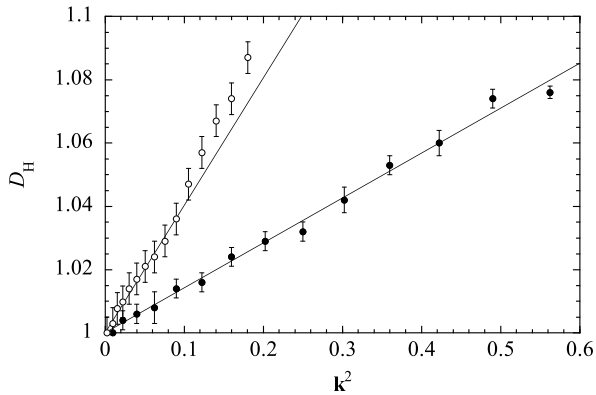

Chaotic Dynamics in Nonequilibrium Statistical Mechanics, Figure 8

Hausdorff dimension D_H of the incomplete van Hove function versus $k^2 = k_x^2$ ($k_y = 0$) for both periodic Lorentz gases with hard-disk scatterers (filled circles) and for the case where the hard disks are replaced by repulsive Coulomb scatterers on a square lattice, and the energy of the particle is sufficiently high for the motion to be chaotic (open circles). Both solid lines have slopes equal to $\mathcal{D}/\lambda$ for the respective diffusion coefficient $\mathcal{D}$ and Lyapunov exponent λ of the Lorentz gases

Entropy Production in Non-equilibrium, Hamiltonian Systems

One of the subjects of most active research in recent years has been the theory of entropy production in non-equilibrium systems. Dynamical systems theory has provided some new insights into this old problem. Briefly formulated, the problem is to explain the positive, irreversible production of entropy using the fundamental ideas in statistical mechanics, particularly the Liouville equation. The obstacle in this direction that must be overcome somehow is the fact that the Gibbs entropy

$$S_G(t) = -k_B \int \mathrm{d}\Gamma \rho(\Gamma, t)[\ln \rho(\Gamma, t) - 1] \,, \tag{44}$$

remains constant in time if $\rho(\Gamma, t)$ satisfies the Liouville equation for Hamiltonian systems

$$\frac{\mathrm{d}\rho(\Gamma, t)}{\mathrm{d}t} = 0 \,. \tag{45}$$

Here Γ represents all of the coordinate and momentum variables of the system. The usual way around this difficulty is to introduce a coarse grained entropy, obtained by either defining an entropy in terms of reduced distribution functions, as is done in Boltzmann's famous H-theorem, or by coarse graining the phase space itself, and defining an entropy in terms of the average phase space distribution in each of the coarse graining regions [9]. In either case, it is possible to show that the rate of production of the redefined entropy is positive, and for the Boltzmann case at least, is in agreement with the predictions of irreversible thermodynamics when the gas is close to a local equilibrium state.

Ideas from dynamical systems theory have not changed the fundamental need for coarse graining. Instead they have provided strong reasons for doing it, reasons that were lacking in the previous approaches, where the motivation for coarse graining seemed only to be that it was necessary to coarse grain in some way to get a positive entropy production. The central new idea in this area is that the phase space description of a non-equilibrium process requires, in the thermodynamic limit at least, the use of distribution functions that are defined on fractal sets. A simple example is provided by a system in which diffusion can take place in a region between two particle reservoirs, each reservoir being maintained at a different density of particles [6,40]. Then if the dynamics that leads to transport of the particles between the reservoirs is Anosov-like, one can argue, and in simple enough cases show explicitly, that the regions in phase space that correspond to regions of different density in the system get so tangled up and enmeshed that the distribution function is a wildly varying fractal function. As such, it is no longer differentiable and the steps that lead to the proof of the constancy of the Gibbs entropy can no longer be justified. The only way to treat this kind of fractal behavior of the distribution function is to smooth the function in one or another way, typically by defining cumulative distribution functions over small regions of phase space, leading to the construction of SRB measures. In simple cases, one can show that fractal forms appear in the phase space distribution function for systems in non-equilibrium stationary states produced by particle reservoirs, or in the relaxation of a system with particle diffusion to equilibrium, and that the rate of entropy production agrees with the predictions of irreversible thermodynamics [6,40,50,90,91]. However, much needs to be done to extend this work to other hydrodynamic processes and to understand why the results of macroscopic theory are obtained in this way. Recently, Gaspard has presented an expression for the rate of entropy production in chaotic as well as in stochastic systems as the difference in two dynamical entropies per unit time, one a form of the Kolmogorov–Sinai entropy per unit time, and the other, an entropy per unit time for the time reversed motion but as measured by the measure of the forward process. It is possible to prove that this rate of entropy production is positive, and for the cases studied so far, agrees with the results of irreversible thermodynamics [51,52].

Kinetic Theory Methods for Analytical Calculations of Lyapunov Spectra

As we demonstrated in the examples given above for the applications of chaos theory to non-equilibrium processes, an important role is played by Lyapunov exponents and Kolmogorov–Sinai entropies per unit time. Apart from computer simulations, it is not always clear how one might obtain expressions and values for these quantities. In some simple cases such as the baker's map and toral automorphisms, the calculation of the Lyapunov exponents is a simple matter. However these cases are very special and rarely are realistic models for physical systems. Thus the problem remains of finding general methods for an analytic determination of Lyapunov exponents and Kolmogorov–Sinai entropies. One such method is kinetic theory. Using the Boltzmann and related equations [8,35], van Beijeren and co workers have been able to obtain expressions for Lyapunov exponents, and in some cases, Kolmogorov–Sinai entropies, for dilute gases with short range potentials, such as hard ball systems, and for dilute Lorentz gases [8,63,64,65,92,93]. Results obtained this way are in

good agreement with the results of computer simulations. We refer to the literature for details.

Entropy Production in Systems with Gaussian Thermostats and the Cohen–Gallavotti Fluctuation Theorem

As a final example of the applications of dynamical systems theory to non-equilibrium statistical mechanics we briefly discuss one example of a number of closely related fluctuation theorems, first obtained by Evans, Searles, and by Gallavotti, and Cohen, and generalized by Spohn and Lebowitz, among others [16,17,18,19,20,21]. These fluctuation theorems are very closely related to a class of work-free energy fluctuation results due to Jarzynski [94], and Crooks [20], as well as to the expression for entropy production given by Gaspard [51,52], as mentioned above. By now the literature is quite extensive and the reader is directed there for more details. Here we discuss, briefly, the Gallavotti–Cohen version of a fluctuation theorem appropriate for systems with Gaussian thermostats [18]. In the discussion of systems with energy conserving thermostats, we have introduced another idea for entropy production, namely entropy production in a reservoir associated with phase space contraction of a thermostatted system. Here the central idea is that the thermostatted dynamics produces an attractor for the system's trajectories in phase space. The friction coefficient, which we have denoted by α, is a dynamical function taking on positive or negative values depending on the trajectory of the system in phase space. To produce an overall phase space contraction, the friction coefficient should be positive on average, but from time to time it may be negative. Loosely speaking, we might say that the entropy production in the system is positive when the value of α is positive at the time, and negative when α is negative. For example, in the Lorentz gas that we have been discussing, when the particle moves in the direction of the field, α is positive, and when it moves opposite to the direction of the field, α is negative, provided the charge q is positive. One might imagine some time interval τ, say, and ask for the probability that the time-averaged entropy production per unit time, ϵ_τ, as measured by the phase space contraction, has a value a over this time interval. This probability can be expressed in terms of the SRB measure on the attractor of the steady state system. The fluctuation theorem for this system is a result for the ratio of the probability that the time-average entropy production per unit time over an interval τ will be the value a, to the probability that this value will be $-a$. For a reversible, Anosov-like system with a Gaussian thermostat, the theorem is

$$\frac{\mathrm{Prob}(\epsilon_\tau = a)}{\mathrm{Prob}(\epsilon_\tau = -a)} = \mathrm{e}^{a\tau} \tag{46}$$

Note that for long times this ratio approaches zero or infinity depending upon the sign of a.

This result was first discovered by means of computer simulations by Evans, Cohen, and Morriss [16], and this observation was explained on the basis of Anosov-like dynamics by Gallavotti and Cohen [18]. A closely related fluctuation formula was derived by Evans and Searles [17]. By now this class of fluctuation theorems has been generalized considerably to include analogous results for stochastic and other kinds of systems. We refer to the literature mentioned above for further details.

Discussion

Here we described some aspects of the theory of irreversible processes, as seen from the point of view of dynamical systems theory. We described the notions of ergodic and mixing properties of a dynamical system and argued that they alone are insufficient for a full explanation of the approach to equilibrium as seen on laboratory time scales. However, we tried to argue that if the microscopic system has positive Lyapunov exponents connected to unstable manifolds in the phase space, then one can argue that an approach to equilibrium of reduced distribution functions can occur on much shorter times scales than those needed to establish the mixing of phase space regions throughout the entire phase space. Moreover the existence of a chaotic microscopic dynamics allows us to derive some very striking connections between macroscopic transport coefficients and quantities such as Lyapunov exponents, fractal dimensions, Kolmogorov–Sinai entropies, that characterize the underlying chaotic, microscopic dynamics of the system. Much, but not all, of our discussion was based on some very simple examples of classical chaotic systems with few degrees of freedom, the baker's map and the Arnold Cat Map. These are simple Anosov-like and Anosov systems where one can analyze many points in some detail. However, this analysis is still very far from an analysis of systems of real interest to statistical mechanics, where the number of degrees of freedom is generally much, much larger, even for systems of particles studied on computers, and where the dynamics is not always chaotic. Therefore, it is very much an open question to show that this picture persists when the systems studied are of macroscopic size, and the dynamics is treated in a more realistic way.

Finally we should say something about the role that quantum mechanics plays in our understanding of the approach of a system to equilibrium. "Quantum chaos"

is a subject that has developed an enormous literature [15,95,96,97], primarily associated with the quantum behavior of systems with few degrees of freedom which are chaotic in the classical limit of $\hbar \to 0$. This work is clearly of great importance for understanding the behavior of *mesoscopic* quantum devices such as quantum dots and related materials. However, there is no analog of dynamical chaos in quantum mechanics and no direct translation of the results described here to quantum systems. This is a result of the fact that the limit as $\hbar$ approaches zero *does not commute* with the limit of t approaching infinity. Therefore one cannot typically study the asymptotically long time behavior of a quantum system whose classical counterpart is chaotic, and then by taking the limit of $\hbar$ approaching zero, obtain the correct, chaotic behavior of the classical system. Instead one may look for some sign in the quantum motion that the classical system is chaotic. Typically a quantum system will exhibit some signs of the chaotic behavior of the classical system in the semi-classical region of small, but not zero $\hbar$, for some period of time known as the *Ehrenfest time*. The Ehrenfest time is essentially the time that it takes an initially small wave packet to expand to some characteristic length in the system. While the packet is small, the semi-classical behavior is essentially that of a classical system. The rate of the expansion of the wave packet is determined by the classical Lyapunov exponents. When the wave packet reaches a certain size, then the motion of the system is governed by interference and diffraction effects that have no classical counterpart. For example, while a classical particle moving among a random array of fixed scatterers exhibits normal diffusion, its quantum counterpart can be localized, or move diffusively, depending upon the spatial dimension of the system and the particle's energy.

The comments above suggest that chaotic dynamics has a large role to play in quantum systems when one looks at the semi-classical regime, namely the regime where $\hbar$, in proper dimensionless units, can be considered very small [95,96]. One important result, known as *Schnirelman's Theorem* [98]states that in the semi-classical limit, most of the wave functions for a quantum system, whose classical counterpart is chaotic and ergodic, become "equidistributed" on the constant energy surface. This means that the probability of finding of finding the system in some region becomes equal to the ratio of the measure of that region to the total measure of the region available to the system. Nevertheless, there are some wavefunctions for classically chaotic systems that exhibit *scars* in the semi-classical limit, where the wave function is concentrated on periodic orbits of the classical motion. The scarred wave functions then form a special class which do not satisfy Schnirelman's Theorem. In this connection we also mention *Berry's conjecture* [99] which states that the high energy wavefunctions for a classically chaotic, ergodic system can be represented as a Gaussian random function, such as a superposition of plane waves with random phases. For further details we refer to the literature listed above.

The understanding of the approach to equilibrium in macroscopic systems typically requires a treatment of quantum systems with a very large number of degrees of freedom. Such systems are not likely to be in one or in a superposition of a few energy eigenstates. Instead, such systems are likely to be in a state that is a superposition of a huge number of such quantum states, and destructive interference of the phase relations among the states is an important ingredient of the behavior of such systems [100]. In fact there is a large literature that deals with the phenomenon of the loss of coherence, or *decoherence*, of superpositions of large numbers of quantum states for a macroscopic system, and helps us understand a bit more clearly why classical mechanics is a good approximation for describing systems that we know to be intrinsically quantum mechanical.

Future Directions

We conclude with a list of open questions that will provide some indication of what directions might be fruitful further work in the future. This list is far from exhaustive, but focuses upon the particular issues addressed here.

1. Most of the results described here that connect microscopic dynamics and macroscopic transport, such as the escape rate formulae, Eq. (23), the dimension formula Eq. (43), and the Lyapunov exponents-transport coefficient relations for Gaussian thermostatted systems Eq. (34), have been proven for purely chaotic, Anosov or Anosov-like systems, only. The phase spaces of most realistic systems are not likely to be purely chaotic. Instead one expects realistic systems to have mixed phase spaces with both chaotic regions and non-chaotic regions. So far a satisfactory treatment of transport in mixed systems is in a very rudimentary shape, except for some models which have been studied extensively. It would be useful to have estimates of the size and importance of non-chaotic regions in many-particle systems and to determine what corrections, if any, are needed to generalize the results mentioned above to mixed systems.
2. Most of the results given here have been illustrated, and in some cases only derived, for low dimensional systems. How much of this discussion is relevant for

systems with many particles where the dimension of phase space is very large? For example, the formula connecting diffusion and Lyapunov exponents, given by Eq. (43), has only been derived for diffusion in two dimensional chaotic systems Lorentz gases. What is the generalization of this formula to higher dimensions and to general transport processes?

3. The fact that our detailed understanding of the role of chaos in non-equilibrium processes is for systems of low dimensionality suggests that applications of chaos to nanoscale systems might be very fruitful [52].
4. Pseudo-chaotic systems present a great challenge both to physicists and to mathematicians [101,102]. Are there general statements one can make about the motion of a particle in a collection of scatterers, such as hard squares or other scatterers where the motion is not chaotic at all, and the rate of separation of trajectories is at best algebraic? How do these properties depend on the geometrical structure and arrangements of the scatterers? Are there generalizations of the escape-rate formula, and the others mentioned here, to pseudo-chaotic systems?
5. There are at least a few logical gaps in the applications of dynamical systems theory to non-equilibrium statistical mechanics. One one hand we have argued that as far as the approach to equilibrium is concerned it makes sense to look at projected distributions since these distribution functions reach equilibrium forms long before the full phase space distribution function does. This statement itself needs a careful proof for realistic systems. Moreover, the derivations of the formulae that connect transport properties with microscopic dynamical quantities rely upon the use of the full phase space. This suggests that there is a fundamental issue of relevant time scales that needs to be resolved, in order to determine the time scales on which these results become valid.

Acknowledgments

I would like to thank Henk van Beijeren for reading a draft of this article and for his very helpful remarks. I would also like to thank Rainer Klages for his new book (Reference [12]), which was very helpful when preparing this article.

Bibliography

Primary Literature

1. Uhlenbeck GE, Ford GW (1963) Lectures in Statistical Mechanics, 2nd edn. Cambridge University Press, Cambridge
2. Toda M, Kubo R, Saito N (1992) Statistical Physics, vol I. Springer, Berlin
3. Kubo R, Toda M, Hashitsume (1992) Statistical Physics, vol II. Springer, Berlin
4. Eckmann JP, Ruelle D (1985) Ergodic theory of chaos and strange attractors. Rev Mod Phys 57:617
5. Evans DJ, Morriss GM (1990) Statistical Mechanics of Nonequilibrium Liquids, 2nd edn. Cambridge Univ Press, Cambridge
6. Gaspard P (1998) Chaos, Scattering, and Statistical Mechanics. Cambridge University Press, Cambridge
7. Hoover WG (1999) Time Reversibility, Computer Simulation, and Chaos. World Scientific Publishing Co, Singapore
8. Dorfman JR (1999) An Introduction to Chaos in Nonequilibrium Statistical Mechanics. Cambridge University Press, Cambridge
9. Ruelle D (1999) Smooth dynamics and new theoretical ideas in nonequilibrium statistical mechanics. J Stat Phys 95:393
10. Gallavotti G (1999) Statistical Mechanics – A Short Treatise. Springer, Berlin
11. Szasz D (ed) (2000) Hard-ball Systems and the Lorentz Gas. Encyclopedia of mathematical sciences, vol 101. Springer, Berlin
12. Klages R (2007) Microscopic Chaos, Fractals and Transport in Nonequilibrium Statistical Mechanics. World Scientific Publishing Co, Singapore
13. Klages R, van Beijeren H, Dorfman JR, Gaspard P (eds) (2004) Microscopic chaos and transport in many-particle systems. Special Issue of Physica D, vol 187:1–391
14. Katok A, Hasselblatt B (1995) Introduction to the Modern Theory of Dynamical Systems. Cambridge University Press, Cambridge
15. Srednicki M (1999) The approach to thermal equilibrium in quantized chaotic systems. J Phys A 32:1163
16. Evans DJ, Cohen EGD, Morriss GP (1993) Probability of second law violations in shearing steady flows. Phys Rev Lett 71:2401
17. Evans DJ, Searles DJ (2002) The fluctuation theorem. Adv Physics 51:1529
18. Gallavotti G, Cohen EGD (1995) Dynamical ensembles in stationary states. J Stat Phys 80:931
19. Lebowitz JL, Spohn H (1999) A Gallavotti–Cohen type symmetry in the large deviation functional for stochastic dynamics. J Stat Phys 95:333
20. Crooks GE (1999) Entropy fluctuation theorem and the nonequilibrium work relation for free energy differences. Phys Rev E 60:2721
21. Kurchan J (1998) Fluctuation theorem for stochastic dynamics. J Phys A 31:3719
22. Mazo RM (2002) Brownian Motion: Fluctuations, Dynamics, and Applications. Oxford University Press, Clarendon
23. Helfand E (1960) Transport coefficients from dissipation in a canonical ensemble. Phys Rev 119:1
24. Ehrenfest P, Ehrenfest T (1959) The Conceptual Foundations of the Statistical Approach in Mechanics. Cornell University Press, Ithaca
25. Dettmann CP, Cohen EGD (2000) Microscopic chaos and diffusion. J Stat Phys 101:775
26. Bunimovich L, Sinai YG (1981) Statistical properties of the Lorentz gas with periodic configuration of scatterers. Commun Math Phys 78:478
27. Gaspard, Baras F (1995) Chaotic scattering and diffusion in the Lorentz gas. Phys Rev E 51:5332

28. Turaev D, Rom-Kedar V (1998) Elliptic islands appearing in near-ergodic flows. Nonlinearity 11:575
29. Donnay VJ (1996) Elliptic islands in generalized Sinai billiards. Ergod Theory Dyn Syst 16:975
30. Walters P (1982) An Introduction to Ergodic Theory. Springer, Berlin
31. Hopf E (1937) Ergodentheorie. Springer, Berlin
32. Sinai YG (ed) (1991) Dynamical Systems, A Collection of Papers. World Scientific Publishing Co, Singapore
33. Simányi N (2004) Proof of the ergodic hypothesis for typical hard ball systems. Ann Henri Poincaré 5:203
34. Cornfeld IP, Fomin SV, Sinai YG (1982) Ergodic Theory. Springer, Berlin
35. Chapman S, Cowling TG (1970) The Mathematical Theory of Non-uniform Gases, 3rd edn. Cambridge University Press, Cambridge
36. Bogoliubov NN (1962) Problems of a dynamical theory in statistical physics. In: Studies in Statistical Mechanics, vol 1. North Holland Publishing Co, Amsterdam
37. Zaslavsky GM (2007) The Physics of Chaos in Hamiltonian Systems. Imperial College Press, London
38. Berry MV (1978) Regular and irregular motion. In: Jorna S (ed) Topics in Nonlinear Dynamics: A Tribute to Sir Edward Bullard, American Institute of Physics, New York
39. Tél T, Gruiz M (2006) Chaotic Dynamics: An Introduction Based on Classical Mechanics. Cambridge University Press, Cambridge
40. Gaspard P (1997) Entropy production in open vol preserving systems. J Stat Phys 88:1215
41. Tasaki S, Gilbert T, Dorfman JR (1998) An analytical construction of the SRB measures for baker-type maps. Chaos 8: 424
42. Ott E (2002) Chaos in Dynamical Systems. Cambridge University Press, Cambridge
43. Gaspard P, Nicolis G (1990) Transport properties, Lyapunov exponents and entropy per unit time. Phys Rev Lett 65:1693
44. Gaspard P (1992) Diffusion, effusion and chaotic scattering. J Stat Phys 68:673
45. Dorfman JR, Gaspard P (1995) Chaotic scattering theory of transport and reaction-rate coefficients. Phys Rev E 51:28
46. Gaspard P, Dorfman JR (1995) Chaotic scattering theory, thermodynamic formalism, and transport coefficients. Phys Rev E 52:3525
47. Viscardy S, Gaspard P (2003) Viscosity in the escape-rate formalism. Phys Rev E 68:041205
48. Evans DJ, Cohen EGD, Morriss GP (1990) Viscosity of a simple liquid from its maximal Lyapunov exponents. Phys Rev A 42:5990
49. Hoover WG, Posch HA (1994) Second-law irreversibility and phase space dimensionality loss from time-reversible nonequilibrium steady-state Lyapunov spectra. Phys Rev E 49:1913
50. Dorfman JR, Gaspard P, Gilbert T (2002) Entropy production of diffusion in spatially periodic deterministic systems. Phys Rev E 66:026110
51. Gaspard P (2004) Time reversed dynamical entropy and irreversibility in Markovian random processes. J Stat Phys 117:599
52. Gaspard P (2006) Hamiltonian dynamics, nanosystems, and nonequilibrium statistical mechanics. Phys A 369:201
53. Tél T, Vollmer J (2000) Entropy balance, multibaker maps, and the dynamics of the Lorentz gas. In: Szasz D (ed) Hard Ball Systems and the Lorentz Gas. Springer, Berlin
54. Vollmer J (2002) Chaos, spatial extension, transport, and nonequilibrium thermodynamics. Phys Rep 372:131
55. van Zon R, Cohen EGD (2004) Extended heat fluctuation theorems for a system with deterministic and stochastic forces. Phys Rev E 69:056121
56. Gaspard P, Rice SA (1989) Scattering from a classically chaotic repeller. J Chem Phys 90:2225
57. Gaspard P (1993) What is the role of chaotic scatttering in irreversible processes? Chaos 3:427
58. Dorfman JR, van Beijeren H (1997) Physica A 240:12
59. Tél T, Vollmer J, Breymann W (1996) Transient chaos: The origin of chaos in driven systems. Europhys Lett 35:659
60. Claus I, Gaspard P (2000) Microscopic chaos and reaction-diffusion processes in the periodic Lorentz gas. J Stat Phys 101:161
61. Claus I, Gaspard P, van Beijeren H (2004) Fractals and dynamical chaos in a random 2D Lorentz gas with sinks. Physica D 187:146
62. Bunimovich LA, Demers MF (2005) Deterministic models of the simplest chemical reactions. J Stat Phys 120:239
63. van Beijeren H, Dorfman JR (1995) Lyapunov exponents and Kolmogorov–Sinai entropy for the Lorentz gas at low densities. Phys Rev Lett 74:4412, erratum 77:1974
64. van Beijeren, Latz A, Dorfman JR (2001) Chaotic Properties of dilute, two and three dimensional random Lorentz gases II: open systems. Phys Rev E 63:016312
65. van Zon R, van Beijeren H, Dorfman JR (2000) Kinetiic theory estimates for the Kolmogorov–Sinai entropy and the largest Lyapunov exponents for dilute, hard ball gases and for dilute, random Lorentz gases. In: Szasz D (ed) Hard Ball Systems and the Lorentz Gas. Springer, Berlin
66. Evans DJ, Hoover WG, Failor BH, Moran B, Ladd AJC (1983) Nonequilibrium molecular dynamics via Gauss' principle of least constraint. Phys Rev A 28:1016
67. Posch HA, Hoover WG (1988) Lyapunov instability of dense Lennard–Jones fluids. Phys Rev A 38:473
68. Posch HA, Hoover WG (1989) Equilibrium and non-equilibrium Lyapunov spectra for dense fluids and solids. Phys Rev A 39:2175
69. Chernov NI, Eyink GL, Lebowitz JL, Sinai YG (1993) Steady state electrical conduction in the periodic Lorentz gas. Commun Math Phys 154:569
70. Baranyi A, Evans DJ, Cohen EGD (1993) Field-dependent conductivity and diffusion in a two-dimensional Lorentz gas. J Stat Phys 70:1085
71. Evans DJ, Cohen EGD, Searles DJ, Bonetto F (2000) Note on the Kaplan–Yorke dimension and linear transport coefficients. J Stat Phys 101:17
72. Dellago C, Glatz L, Posch H (1995) Lyapunov spectrum of the driven Lorentz gas. Phys Rev E 52:4817
73. Dellago C Posch HA, Hoover WG (1996) Lyapunov instability in a system of hard disks in equilibrium and nonequilibrium steady states. Phys Rev E 53:1485
74. Dettmann CP (2000) The Lorentz Gas: A paradigm for nonequilibrium steady states. In: Szasz D (ed) Hardball Systems and the Lorentz Gas. Springer, Berlin
75. Posch HA, Hirshl R (2000) Simulation of billiards and hard body fluids. In: Szasz D (ed) Hard Ball Systems and the Lorentz Gas. Springer, Berlin

76. Dettmann CP, Morriss GP (1996) Proof of Lyapunov exponent pairing for systems at constant kinetic energy. Phys Rev E 53:R5545
77. Wojtkowski M, Liverani C (1998) Conformally Symplectic Dynamics and the symmetry of the Lyapunov spectrum. Commun Math Phys 194:7
78. Bohm A, Gadella M (1990) Dirac Kets, Gamow Vectors and Gelfand Triplets: The Rigged Hilbert Space Formulation of Quantum Mechanics. Springer, Berlin
79. Pollicott M (1985) On the rate of mixing of Axiom-A flows. Inventiones Mathematicae 81:413
80. Pollicott M (1986) Meromorphic extensions of generalized zeta functions. Inventiones Mathematicae 85:147
81. Ruelle D (1986) Resonances of chaotic dynamical systems. Phys Rev Lett 56:405
82. Ruelle D (1986) Locating Resonances for Axiom-A dynamical systems. J Stat Phys 44:281
83. Dörfle M (1985) Spectrum and eigenfunctions of the Frobenius–Perron operator for the tent map. J Stat Phys 40:93
84. Gaspard P (1992) r-adic one dimensional maps and the Euler summation formula. J Phys A 25:L483
85. Gaspard P (1992) Diffusion in uniformly hyperbolic one dimensional maps and Appell polynomials. Phys Lett A 168:13
86. Fox RF (1997) Construction of the Jordan basis for the baker map. Chaos 7:254
87. Gaspard P (1996) Hydrodynamic modes as singular eigenstates of Liouvillian dynamics: Deterministic diffusion. Phys Rev E 53:4379
88. Gilbert T, Dorfman JR, Gaspard P (2001) Fractal dimension of the hydrodynamic modes of diffusion. Nonlinearity 14:339
89. Gaspard P, Claus I, Gilbert T, Dorfman JR (2001) Fractality of hydrodynamic modes of diffusion. Phys Rev Lett 86:1506
90. Fox RF (1998) Entropy evolution for the baker map. Chaos 8:462
91. Goldstein S, Lebowitz JL, Sinai YG (1998) Remark on the (Non)convergence of ensemble densities in dynamical systems. Chaos 8:393
92. van Zon R, van Beijeren H, Dellago C (1998) Largest Lyapunov exponent for many-particle systems at low densities. Phys Rev Lett 80:2035
93. de Wijn A, van Beijeren H (2004) Goldstone modes in Lyapunov spectra of hard sphere systems. Phys Rev E 70:016207
94. Jarzynski C (1997) Nonequilibrium equality for free energy differences. Phys Rev Lett 78:2960
95. Haake F (2001) Quantum Signatures of Chaos. Springer, Berlin
96. Stöckmann H-J (1999) Quantum Chaos: An Introduction. Cambridge University Press, Cambridge
97. Wojcik D (2006) Quantum maps with spatial extent: a paradigm for lattice quantum walks. Int J Mod Phys B 20:1969
98. Lazutkin VF (1993) KAM Theory and Semiclassical Approximations to Wave Functions. Springer, Berlin
99. Berry MV (1977) Regular and irregular wave functions. J Phys 10:2083
100. van Kampen N (1988) Ten theorems about quantum mechanical measurements. Physica A 153:97
101. Gutkin E (1996) Billiards in polygons: A survery of recent results. J Stat Phys 83:7
102. Tabachnikov S (2005) Billiards and Geometry. American Mathematical Society Press, Providence

Books and Reviews

Beck C, Schlögl F (1993) Thermodynamics of Chaotic Systems. Cambridge University Press, Cambridge
Tél T, Gaspard P, Nicolis G (eds) (1998) Focus Issue on Chaos and Irreversibility. Chaos 8(2):309–529
Rom-Kedar V, Zaslavsky G (eds) (2000) Focus Issue on Chaotic Kinetics and Transport. Chaos 10(1):1–288
Casati G, Chirikov B (eds) (1995) Quantum Chaos: Between Order and Disorder. Cambridge University Press, Cambridge
Dorfman JR (1998) Deterministic chaos and the foundation of the kinetic theory of gases. Phys Rep 301:151
Garbaczewski P, Olkiewicz R (eds) (2002) Dynamics of Dissipation. Lecture Notes in Physics, vol 597. Springer, Berlin

Charge-Based Solid-State Flying Qubits

Andrea Bertoni
CNR-INFM National Research Center
on NanoStructures and BioSystems at Surfaces (S3),
Modena, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Quantum Information Processing and Quantum Computation
Flying Qubits Based on Coherent Electron Transport in Coupled Quantum Wires
Entangled States and Estimation of Quantum Entanglement
Future Directions
Acknowledgment
Bibliography

Glossary

Bit Elementary unit of classical information represented by a binary digit.

Qubit (or quantum bit) Elementary unit of quantum information. The qubit refers also to the physical system whose state encodes the qubit of information.

Quantum gate Logical operation performed on one or a few qubits, that change their state according to a unitary transformation. Quantum gates are reversible by definition.

Quantum entanglement Property possessed by two or more quantum systems, when the state of the global system that includes all of them cannot be described by the simple composition of their states. Two entan-

gled systems show quantum correlations between their states that have no classical analogue.

Bell's inequality Relation between two sets of measurements performed on two quantum systems spatially separated. Bell's inequality can only be violated if the two systems are entangled.

Quantum wire Metallic or semiconductor wire with nanometric thickness. While the longitudinal current carrying states form a continuum in the energy spectrum, the transverse component of the carriers wave functions originates a discrete spectrum.

Two-dimensional electron gas (2DEG) Gas of electrons that are quantum-confined in one dimension and free to move in the remaining two. In the confinement direction the single-particle states have a quantized energy spectrum. 2DEGs are usually obtained through the modulation of the material conduction band in a semiconductor heterostructure.

Surface acoustic wave (SAW) Elastic acoustic wave that propagates on the surface of a material. In piezoelectric materials SAWs couples with electrons through the SAW-induced piezoelectric field.

Definition of the Subject

The physical implementation of the *qubit*, the dicotomic unit of quantum information, has been, and still is, the starting point of any proposal for a quantum information processing device. Its proper definition is a necessary, although not sufficient, condition for a practical system to be able to exploit the quite large number of quantum algorithms that have been proven to outperform the corresponding classical ones. The first period of quantum information processing research saw many proposals for qubits based on a large spectrum of systems and approaches, but later, the focus was put on solid-state devices, mainly due to their promises for scalability and better integrability with present-day semiconductor electronics. Furthermore, a deep understanding of the quantum-based physics of semiconductor devices is available, due to the long-standing research on the field, mainly driven by the huge commercial success of semiconductor microprocessors. In fact, the application of quantum mechanics to solid-state systems has resulted in a large number of semiconductor devices with novel functionalities, and the quantum mechanical aspects of the device properties start to dominate their behavior, as device dimensions get smaller. While this may become a problem for traditional electronic devices, it is a prerequisite for the implementation of quantum technology in general and quantum information processing in particular. These possibilities have inspired a number of proposals for creating qubits, quantum gates and quantum registers from semiconductor devices and using them for implementing quantum algorithms. Theoretical studies of the device properties have been undertaken, but the challenging technological hurdles have slowed down experimental progress.

Introduction

In what follows, we will describe a class of solid-state qubits, known as *charge flying qubits*, whose characteristic is the evolution with time of their spatial localization. We will first briefly introduce the concept of quantum information processing and, more specifically, of quantum computing. Then we will focus on the definition of the qubit and of quantum entanglement, stemming from the non-separability of the multi-particle state. The logic transformations that must be applied to the qubit in order to implement quantum algorithms, namely the *quantum gates*, will be described and the requirements for a practical realization of quantum computing devices, listed by DiVincenzo in his famous checklist [41], will be reviewed. Then we will enter into the details of a specific proposal for the realization of quantum bits and quantum gates, based on the charge state of electrons in semiconductor quantum wires.

In the last sections, we will return to the concept of qubit entanglement and will present specific calculations for the entanglement creation in a quantum-wire system and, more general, in a carrier-carrier scattering. In fact quantum entanglement can be considered both a resource and a detriment for quantum computing. It is a resource since any quantum algorithm is based on the entanglement of two or more qubit states, it is a detriment when a qubit is entangled with a degree of freedom that is outside the computational space, i. e. the environment, giving rise to the phenomenon of decoherence. It is the aim of the theoretical modeling to identify the proper system parameters for the controlled production of the first kind of entanglement, while avoiding, or at leas reducing, the second one.

The idea of a solid-state qubit, whose state is encoded by the localization of one electron traveling in two coupled semiconductor quantum wires, is here anticipated. The basic qubit structure consists of two parallel quantum wires, close to each other and separated by a potential barrier. The state of the qubit is represented by the location of a single electron traveling in one (state $|0\rangle$) or the other (state $|1\rangle$) of the two wires with a controlled velocity. If a coupling window is introduced between the wires through the lowering of the potential barrier that separates

them, the electronic wave function crossing the coupling region oscillates between the wires. When the electron reaches the end of the coupling window, the oscillation process terminates, and the wave function is eventually divided into two parts running along the left and right wires. The necessary superposition of quantum states is thus obtained. The introduction of a low potential barrier in one of the two wires, able to delay the propagation of the wave function with negligible reflection, delays the propagation of the part of the wave function in that particular wire and generates a phase shift between the two components of the qubit. Combinations of these two transformations allow for the implementation of a general single-qubit operation.

A two-qubit gate requires two pairs of wires. They are designed in such a way that only the electron running in the 1 component of the first qubit and the electron running in the 0 component of the second qubit feel their Coulomb interaction, thus generating the necessary entanglement. As a consequence, the electrons are slowed down, and the system undergoes a phase shift that originates from the spatial delay of the two interacting electrons. This last transformation and the two single-qubit transformations described above form a universal set of gates for quantum computing. Numerical simulation of the quantum dynamics of the proposed devices have been performed, in order to obtain the parameters for the experimental realization and to estimate the performance and decoherence times of these systems.

Prototypes of single-qubit devices have been realized experimentally, based on coupled electron wave guides obtained from two-dimensional electron gases (2DEGs) through split-gate or AFM nanolithography techniques. Their basic functioning is currently under investigation. The experiments and the different realization strategies will be mentioned.

Although the electrical properties of quantum wires are nowadays well established both theoretically and from the point of view of experimental realization, the feasibility of complex devices, with a network of coupled quantum wires, remains a challenge. Some considerations about the practical realizability of quantum computers based on solid-state flying qubits will be presented, and an example of a simple quantum-gate network will be given. A number of mechanisms of carrier injection into the QWRs, such as surface acoustic waves (SAW), electron pumps, resonant tunneling devices have been proposed in literature. We will describe, in particular, surface acoustic waves, that have demonstrated the ability to inject and drive single electrons in quantum wires. Numerical simulations of a simple model of such systems will be presented. In fact, surface acoustic waves are also expected to reduce decoherence effects due to the quantized energy spectrum induced on the embedded carriers. Concerning the detection of the final state of the charge qubit, it can be achieved by single-electron transistors placed at the end of each quantum wire. Different kinds of solid-state flying qubits have been proposed, as for example, the qubit encoding realized through the spin state of single electrons propagating in quantum wires [5]. We will not enter into the details of these proposals.

Quantum Information Processing and Quantum Computation

Over the few decades following the discovery of quantum mechanics, the tight link between information theory and physics has been recognized and, in 1961, Rolf Landauer pointed out that any irreversible process leading to erasure of information is necessarily a dissipative process. Landauer's statement [28] about the entropy cost of information, opened the search for understanding of how physics constrains the ability to use and manipulate information. Once recognized that "information is physical" [33], the next step is to consider that the universe is fundamentally quantum mechanical: the classical theories about information need to be revised according to quantum physics.

By taking into account the quantum nature of the information-storing devices it was clear that also the corresponding theory about information processing needed to include the distinctive feature of quantum physics: quantum entanglement. The first widespread proposal of a quantum information processing system was advanced by Richard Feynman, as a means of simulating a quantum system. In fact, an exact classical simulation of a quantum system needs to follow, separately, its possible quantum states, leading to a growth of the resources needed. On the other hand, a quantum simulation would follow a superposition state of the quantum system, with a corresponding superposition of the simulation tool.

However, quantum information processing burst onto the scene of active research in 1994, when Peter Shor proposed a quantum algorithm for the factorization of integer numbers that requires a computational time which is polynomial in $\log(N)$, with N the number to be factorized. Since no classical algorithm is known that can perform such a task more efficiently, the importance of the Shor algorithm was immediately realized. In fact, since many widely used public-key cryptographic methods are based on the practical impossibility of factorizing large integers, an "efficient" factorization algorithm could jeopardize their secrecy.

A number of branches of quantum information theory are nowadays fertile research topics, whose borders are, however, not sharp. We mention the fields of quantum algorithms, quantum cryptography, quantum communication, quantum teleportation and, finally, quantum computation. The starting point of all of the above tasks, is the *qubit*, i. e. the elementary particle of quantum information. A brief description of its basic properties, the formalism adopted to describe it, and the differences with classical bits, are given in the following. Then the specific case of a qubit based on coherent electron transport in coupled wires will be analyzed.

The Qubit

Classical computer systems store and process information in binary encoding. Any integer N is represented by an ordered sequence of n bits $a_n, a_{n-1}, \ldots, a_1, a_0$ according to

$$N = \sum_{i=0}^{n-1} a_i 2^i \tag{1}$$

with $0 \leq N \leq (2^n - 1)$ and $a_i \in \{0, 1\}$. In conventional computers, single bits are usually stored as the charge or magnetic state of an electronic device, and transistors are used to realize logic gates, i. e. the transformation of a multiple bit state according to an established rule. Any bit can be in one and only one of two possible states: either 0 or 1. In the quantum case, a bit of information is also represented by one of two possible states, namely $|0\rangle$ and $|1\rangle$, of a given system, but now, since the latter is described by the laws of quantum mechanics, the bit can be in any superposition of the two states. In fact, the states are associated with two specific eigenstates ($|\psi_0\rangle$ and $|\psi_1\rangle$ respectively) of a suitable observable and now the system can be in any state

$$|\psi\rangle = c_0|\psi_0\rangle + c_1|\psi_1\rangle \tag{2}$$

with c_0 and c_1 complex numbers and $|c_0|^2 + |c_1|^2 = 1$. This means that the quantum analog of a bit, the qubit [14], can be simultaneously in both states 0 and 1. If we measure the qubit we will find it with probability $|c_0|^2$ with the value 0 and with probability $|c_1|^2$ with the value 1. It is important to underline that Eq. (2) describes a coherent superposition rather than a generic mixture between 0 and 1. The essential point here is that there is always a base in which the value of the qubit system is well defined, while an incoherent mixture is a mixture whatever base we choose to describe it. As anticipated, the two states of a qubit are usually indicated with $|0\rangle$ and $|1\rangle$, following the Dirac ket notation used in quantum mechanics. Also, it must be noted that the space spanned by a single qubit is isomorphic to the spin space of a spin 1/2 particle. This will allow us to use the formalism for spin states and spin rotations when dealing with single qubits.

Let us suppose now to process some information encoded in an array of n bits. Classically, an input state must be chosen from the 2^n possibilities given by the array, then the information is elaborated, finally an output is produced. To have the output corresponding to a different input, another elaboration must be performed. Things are different if the information is encoded in an array of qubits and if the information is elaborated by means of some kind of "quantum machine" able to preserve the quantum coherence of states. In this case it is possible to create an input state that is a linear superposition of q classical inputs (2^n possibilities *each*), then the output of the elaboration will be the same linear superposition of the q corresponding classical outputs.

In this case, quantum computation exhibits a kind of "natural" massive parallelism, but this peculiarity of quantum mechanics can not be easily exploited. In fact, in order to know the output of the elaboration, a *measure*, in the quantum mechanical sense, must be performed, this producing a collapse of the quantum state into a single component of the linear combination. In our simple example this means that only one random result, among the results of the q classical computations, can be revealed. To get benefit from quantum processing of information, specific algorithms must be implemented, as mentioned in the previous section, that exploit the main peculiarity of quantum physics i. e. quantum entanglement.

Quantum Entanglement

The deep difference between classical information and quantum information becomes evident when a state of two or more qubits is taken. In this case the whole many-qubit system can be in a superposition of states in which none of the qubits has a definite value but there is a fixed correlation between their values. In other terms, an entangled state is any state that cannot be written as a direct product of single-qubit states.

For example, in the two-qubit entangled state

$$\frac{1}{\sqrt{2}} \left(|0\rangle_1|1\rangle_2 + |1\rangle_1|0\rangle_2\right) , \tag{3}$$

where $|0\rangle_n$ ($|1\rangle_n$) indicate the zero (one) state of the qubit n, the probability to obtain the value 0 as the outcome of a measure on either the first or the second qubit is 50%. However, once a measure has been performed, the

outcome of a measure on the remaining qubit is known with 100% confidence. The correlation between two systems in a joint quantum state is "stronger" than any classical correlation and leads to effects that are unknown in classical physics. This offers the possibility to encode information in a completely new way, in which the information is stored by the multi-qubit system as a whole and no information is carried by single qubits. In order to compact our notation, we will follow the usual convention of indicating the state of the qubits, lexicographically ordered, with a single ket vector and dropping their index. For example, $|01\rangle$ will be used to indicate the two-qubit state $|0\rangle_1|1\rangle_2$.

Entanglement is closely related to the issue of non-locality of quantum measurement. In fact it was evident since the early days of quantum physics that if two particles are in an entangled state, quantum theory predicts that a measurement process performed on a particle is able to change instantaneously the state of the other, even if they are spatially separated. It was indeed Schrödinger that introduced the word *entanglement* to indicate the superposition in a multi-particle system. Because of this "action at distance", quantum theory was strongly criticized by many physicists. Among them, Albert Einstein, Boris Podolsky and Nathan Rosen, in a famous paper [16] dated 1935, refused the non-locality implied by quantum mechanics and in the epistemological discussion that followed, addressed the new theory as incomplete. The discussion about incompleteness of quantum mechanics was considered merely philosophical until 1964. In that year John Bell showed [6] that, for entangled systems, measurements of correlated quantities should yield different results in the quantum mechanical case with respect to those expected if one assumes that the properties of the measured system are present prior to, and independent of, the observation (reality hypothesis).

In the last section, a proposal for an experimental setup able to violate Bell inequality, based on coherent electron transport in coupled quantum wires will be sketched and the generation of maximally entangled Bell's states by means of solid-state flying qubits will be analyzed. Some results of numerical simulations of such a system will be presented.

Bell's states, using the qubit formalism, are two-qubit states that do not carry any information on the state itself on the single qubits. In fact, the usual way is to encode a bit onto each system separately which gives the following four possibilities: $|00\rangle$, $|01\rangle$, $|10\rangle$, $|11\rangle$. By exploiting the features of entanglement it is possible to encode the information using four non-separable states of two qubits, the so-called Bell states:

$$|\Psi^+\rangle = \frac{(|01\rangle + |10\rangle)}{\sqrt{2}}\,, \qquad |\Psi^-\rangle = \frac{(|01\rangle - |10\rangle)}{\sqrt{2}}\,,$$
$$|\Phi^+\rangle = \frac{(|00\rangle + |11\rangle)}{\sqrt{2}}\,, \qquad |\Phi^-\rangle = \frac{(|00\rangle - |11\rangle)}{\sqrt{2}}\,. \tag{4}$$

Note that in each of the four states, any single qubit has 50% of probability to give 0 or 1 when measured. As a consequence a single-qubit reading does not provide even partial information on the state. Only a measure on the joint state of two qubits allows to read the stored information.

We finally note that the search for an experimental violation of Bell's inequality started straight after its discovery but it was only in the 1980s that the violation was obtained experimentally [3] using polarization entangled photons from a spontaneous parametric down conversion. On the other hand, the violation of Bell's inequality in semiconductor-based devices has not been, so far, revealed.

Quantum Gates and the Universal Set

In this section, we introduce some of the simplest quantum logic gates that are at the basis of more complex transformations presented in the following and introduce the concept of a universal set of quantum gates.

For a (classical) digital computer the simplest logic gate is the single-bit NOT gate. It simply changes the state of the input bit from 0 to 1 and vice versa. In the quantum case a NOT gate will perform the transformation

$$|\psi_{\text{in}}\rangle = c_0|0\rangle + c_1|1\rangle \xrightarrow{\text{NOT}} |\psi_{\text{out}}\rangle = c_0|1\rangle + c_1|0\rangle\,, \tag{5}$$

in which the complex coefficients of states $|0\rangle$ and $|1\rangle$ are exchanged.

Another useful single-qubit quantum gate that has no correspondence in classical information processing, is the Hadamard gate H, that transforms the two qubit states $|0\rangle$ and $|1\rangle$ according to:

$$|0\rangle \xrightarrow{\text{H}} \frac{|0\rangle + |1\rangle}{\sqrt{2}} \quad \text{and} \quad |1\rangle \xrightarrow{\text{H}} \frac{|0\rangle - |1\rangle}{\sqrt{2}}\,. \tag{6}$$

Although the idea of a quantum logic-gate operation is strictly related to that of classical logic operation, a crucial difference must be underlined. For many classical gates, like AND, OR, XOR, the number of input and output bits is different. In fact the basic logic gates used to implement digital computers are "many-to-one" operations and are not logically reversible. Once the input state is given, it is straightforward to compute the output state, but it is not possible to identify the input from the output. Thus, the logical transformations performed by these classical gates imply a loss of information. We will not enter here into the topic of reversible and non-reversible computation (see,

Charge-Based Solid-State Flying Qubits, Table 1
Truth table defining the controlled-NOT and Toffoli reversible gates

CNOT	
in	out
00	00
01	01
10	01
11	10

Toffoli			
in	out	in	out
000	000	100	000
001	001	101	101
010	010	110	111
011	010	111	110

for example, [7,33]), but simply report a fundamental result of quantum information theory: a quantum logic gate is a unitary transformation of the input state, therefore any quantum gate must be reversible. This means that it is not possible to start from the well-known classical logic gates and find quantum analogues (as for the NOT gate), but reversible transformations must be used instead, in order to identify a universal set of transformations.

A reversible classical gate of special interest, that will be described in the framework of solid-state flying qubit devices in the following section, is the controlled-NOT (CNOT). It is a two-bit to two-bit gate that has, as the first bit of output, a copy of the first bit of input, and as the second bit of output the XOR between the two input bits. The truth table for the controlled-NOT gate is given in Table 1.

It is easy to identify the quantum version of the CNOT gate. The operation performed on a generic two-qubit state $|\psi_{\text{in}}\rangle$ is:

$$c_a|00\rangle + c_b|01\rangle + c_c|10\rangle + c_d|11\rangle \xrightarrow{\text{CNOT}} c_a|00\rangle + c_b|01\rangle + c_c|11\rangle + c_d|10\rangle\,. \quad (7)$$

Let us introduce here a last reversible classical gate, the Toffoli gate (sometimes called control-control-NOT). It is a three-bit to three-bit gate whose operation is: the first two bits are unchanged, the third bit undergoes a NOT operation if, and only if, first and second bits are both 1. The truth table for the Toffoli gate is given in Table 1. Even in this case it is trivial to find the quantum version of the Toffoli gate. We will omit the explicit expression for the sake of brevity.

In strict analogy with the usual quantum mechanics formalism, an n-qubit state can be written both in ket formalism and as a column vector, once a basis is chosen.

Here and in the following the matrix representation of operators is expressed on the lexicographically ordered basis, i. e., $|0\ldots00\rangle, |0\ldots01\rangle, |0\ldots10\rangle, \ldots, |1\ldots11\rangle$. For example for a 2-qubit system, the basis will be $\{|00\rangle, |01\rangle, |10\rangle, |11\rangle\}$.

With this notation, vector representation of the 2-qubit state representing the entangled state $a|00\rangle + b|01\rangle + c|11\rangle$ results to be

$$\begin{pmatrix} a \\ b \\ 0 \\ c \end{pmatrix}. \quad (8)$$

This formalism allows a fast calculation of the effect of a quantum gate on a qubit state, once the transformation matrix is given. The matrix representation for three quantum gates introduced previously, (NOT, controlled-NOT, Hadamard) is:

$$\text{NOT} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

$$\text{CNOT} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

$$\text{H} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Let us now introduce a compact notation to represent the action performed by a transformation acting on a subset composed of n qubits of an m-qubit system (with $n < m$). The identity operation on the remaining $(m-n)$ qubits will be understood. Let us consider a generic basis ket of the m-qubit system $|b_1, b_2, \ldots, b_m\rangle$ and a generic basis ket of the n-qubit subsystem $|a_1, a_2, \ldots, a_n\rangle$ (with $a_i, b_i \in \{0, 1\}$). If $\mathbf{A}$ is a linear operator acting on $|a_1, a_2, \ldots, a_n\rangle$, we define the operator $\mathbf{A}^{(x_1, x_2, \ldots, x_n)}$ (with $x_i \le m$ and $x_i \ne x_j$ iff $i \ne j, \forall i, j \le n$) acting on the m-qubit state $|b_1, b_2, \ldots, b_m\rangle$ as the operator that transforms the n qubits state $|b_{x_1}, b_{x_2}, \ldots, b_{x_n}\rangle$, according the transformation $\mathbf{A}$ and leave the other qubits unchanged.

For example, the four-dimensional matrices $\mathbf{M}^{(1)}$ and $\mathbf{M}^{(2)}$ that represent the action of a generic single-qubit transformation

$$\mathbf{M} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

on a two-qubit system, and that will be used in the following sections, are:

$$\mathbf{M}^{(1)} = \begin{pmatrix} a & 0 & b & 0 \\ 0 & a & 0 & b \\ c & 0 & d & 0 \\ 0 & c & 0 & d \end{pmatrix},$$
$$\mathbf{M}^{(2)} = \begin{pmatrix} a & b & 0 & 0 \\ c & d & 0 & 0 \\ 0 & 0 & a & b \\ 0 & 0 & c & d \end{pmatrix}. \quad (9)$$

In fact, for a 2-qubit system the transformation $\mathbf{M}^{(1)}$ corresponds to the transformation $\mathbf{M}$ applied on the first qubit, while $\mathbf{M}^{(2)}$ corresponds to the transformation $\mathbf{M}$ applied on the second qubit.

To conclude this section, we introduce the concept of a universal set of gates, i. e. the set of simple transformations that can be combined in order to create any possible complex transformation of bits, (or qubits in the present case). We will show, in the following sections that, by using three types of coupled quantum wire devices, it is possible, at least in principle to implement such a universal set.

The concept is analogous to classical computation, where any given transformation can be obtained, for example, using AND and NOT gates. However, as mentioned above, the AND gate is not a reversible gate. This means that the transformations realized using the set {NOT, AND} are, in general, not reversible and we need a different choice for the quantum case.

In 1995 David DiVincenzo [40] showed that, for quantum computation, a universal set formed solely by one-qubit and two-qubit gates exists. A further simplification of the set of universal quantum gates was introduced by Barenco et al. [4]. They proved that a non-universal, classical two-bit gate can be found that, in conjunction with a generic quantum one-qubit gate, form a universal set for quantum computing. On the basis of this last result, we describe, in the following, the universal set implemented by the charge flying qubit.

Let us start by analyzing the one-qubit gate. The space of a two-component vector representing a single flying qubit is isomorphic to the spin space of a spin 1/2 particle. This means that a general transformation of one qubit is represented by a rotation matrix of the SU(2) group, i. e.

$$\mathbf{U}(\alpha,\beta,\theta) = \begin{pmatrix} e^{i(\alpha/2+\beta/2)}\cos\left(\frac{\theta}{2}\right) & e^{i(\alpha/2-\beta/2)}\sin\left(\frac{\theta}{2}\right) \\ -e^{i(-\alpha/2+\beta/2)}\sin\left(\frac{\theta}{2}\right) & e^{i(-\alpha/2-\beta/2)}\cos\left(\frac{\theta}{2}\right) \end{pmatrix}, \tag{10}$$

where α, β, θ are three real numbers representing three rotation angles around two orthogonal axes: α and β are rotations around Z, between them, a rotation around Y of γ is performed [39]. Any transformation $\mathbf{U}$ can be obtained using two of the three basic rotations:

$$\mathbf{R_x}(\theta) = \begin{pmatrix} \cos\left(\frac{\theta}{2}\right) & i\sin\left(\frac{\theta}{2}\right) \\ i\sin\left(\frac{\theta}{2}\right) & \cos\left(\frac{\theta}{2}\right) \end{pmatrix}, \tag{11}$$

around the X axis,

$$\mathbf{R_y}(\theta) = \begin{pmatrix} \cos\left(\frac{\theta}{2}\right) & \sin\left(\frac{\theta}{2}\right) \\ -\sin\left(\frac{\theta}{2}\right) & \cos\left(\frac{\theta}{2}\right) \end{pmatrix}, \tag{12}$$

around the Y axis, and

$$\mathbf{R_z}(\theta) = \begin{pmatrix} e^{i\frac{\theta}{2}} & 0 \\ 0 & e^{-i\frac{\theta}{2}} \end{pmatrix}, \tag{13}$$

around Z axis.

In particular, the generic transformation $\mathbf{U}$ in the form (10) can be obtained using only $\mathbf{R_y}$ and $\mathbf{R_z}$:

$$\mathbf{U}(\alpha,\beta,\theta) = \mathbf{R_z}(\alpha)\,\mathbf{R_y}(\theta)\,\mathbf{R_z}(\beta)\,. \tag{14}$$

Furthermore, one of the above three rotations can be obtained using the other two. In fact

$$\mathbf{R_x}(\theta) = \mathbf{R_z}(\pi/2)\,\mathbf{R_y}(\theta)\,\mathbf{R_z}(-\pi/2) \tag{15}$$

$$\mathbf{R_y}(\theta) = \mathbf{R_z}(-\pi/2)\,\mathbf{R_x}(\theta)\,\mathbf{R_z}(\pi/2) \tag{16}$$

$$\mathbf{R_z}(\theta) = \mathbf{R_x}(-\pi/2)\,\mathbf{R_y}(\theta)\,\mathbf{R_x}(\pi/2)\,. \tag{17}$$

This means that in order to obtain a generic single-qubit transformation (the first gate in our universal set) it is sufficient to have two of the rotations (11), (12), (13).

We now consider the two-qubit gate of the set. A typical choice for the latter is the CNOT introduced above. Another possibility is to use a two-qubit conditional phase shifter. This gate adds a phase factor to a given component of the qubit state. The usual form for the conditional phase shifter present in literature adds a phase $e^{i\gamma}$ to the state $|11\rangle$:

$$|\psi_{\text{in}}\rangle = c_a|00\rangle + c_b|01\rangle + c_c|10\rangle + c_d|11\rangle \longrightarrow$$
$$\longrightarrow |\psi_{\text{out}}\rangle = c_a|00\rangle + c_b|01\rangle + c_c|10\rangle + e^{i\gamma}c_d|11\rangle\,.$$

For any value of the angle γ, not equal to an integer multiple of π, we have a transformation that, together with the one-qubit gates $\mathbf{U}$ of Eq. (10), form a universal set.

In the following section, the conditional phase shifter used is slightly different from the one described above. In fact it adds a phase to the $|10\rangle$ component of the state. Its matrix representation is:

$$\mathbf{T}(\gamma) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & e^{i\gamma} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{18}$$

and forms, together with the one-qubit gates $\mathbf{U}$, a universal set.

The DiVincenzo Criteria

The concrete realization of quantum-computing capable hardware must face a number of general tasks and requirements that have been listed by David DiVincenzo [41] in the form of a checklist. Five criteria were listed, that must be met by physical systems that are to form the basis of a useful quantum computer.

For the sake of completeness, we report in the following the checklist, as given in [2]:

1. A scalable physical system of well-characterized qubits.
2. The ability to initialize the state of the qubits to a simple fiducial state.
3. Long relative decoherence times, much longer than the gate-operation time.
4. A universal set of quantum gates.
5. A qubit-specific measurement capability.

The above criteria are largely qualitative, but more quantitative bounds have been proposed [27,34]. A universally accepted criterion is the number of gate operations that can be executed before decoherence processes have degraded the quantum information significantly. This number is often estimated as the ratio of the decoherence time by the single gate operation time. However, additional mechanisms leading to the loss of quantum information must also be considered, such as the accuracy of the logic operation performed by the quantum gate. In fact, a problem that is often present in the design of quantum information devices is that optimization of one of the relevant parameters conflicts with the requirements for another operation. As an example, good readout sensitivity and fast gate operations require strong coupling to the classical apparatus that drives the device. However, strong interactions with the environment increases the speed of decoherence. Optimizing performance of the full system, i. e. from initialization through computation and readout, therefore forces one to make choices and to accept some additional decoherence if this results in improved readout sensitivity.

In addition to the above points, two additional points have been added in order to consider the issues related to the transfer of quantum information between different devices implementing the quantum operations:

6. The ability to interconvert stationary and flying qubits.
7. The ability to faithfully transmit flying qubits between specified locations.

The latter criteria, related to quantum computer networkability, are automatically met by proposals whose qubit definition is based on flying qubits. In fact, in order to perform a quantum computation the qubits must undergo a sequence of logical operations, represented by unitary transformations on the multi-qubit state. Two classes of proposals can be identified, according to the kind of qubit encoding [8]. The first one is characterized by qubits that are defined by the quantum state of given physical systems, fixed into space, with quantum gates implemented by changing in time the coupling between the qubits. Quantum computation based on charge and spin degrees of freedom of semiconductor quantum dots are an example of these implementations, where a logic operation is performed by a proper time-dependent tuning of interdot coupling. The second class is represented by the flying qubits, where, contrary to the previous case, quantum gates are fixed, while the position of the qubits changes in time. At each stage, the outputs of a quantum transformation, i. e. the state of a qubit register, must be moved to the inputs of the following quantum gate. This process, overlooked in the early times of quantum information processing research, results to be critical in viable practical realizations and strictly connected to the problem of quantum communication [41].

Flying Qubits Based on Coherent Electron Transport in Coupled Quantum Wires

In this section, a physical system able to perform the basic operations needed for quantum computation is proposed and investigated theoretically. As anticipated in the Introduction, the elementary qubit is defined as the state of an electron running along a couple of quantum wires. We will not enter into the details of the modeling of a realistic device, for which we refer to recent literature [1,29,32,38,42]. We will use hard-wall confining potentials and simple models in order to focus on the functioning principles of the quantum gates. However, we will use GaAs-AlGaAs heterostructure parameters in the simulations presented in the following. We will show that a proper design of the system, together with the action of Coulomb interaction between electrons, allows the implementation of basic one-qubit and two-qubit quantum logic gates.

Let us consider the quantum well formed by a modulation-doped heterostructure. If we assume low free carrier density, less than $10^{10}/\mathrm{cm}^2$, in the so-formed two-dimensional electron gas (2DEG), the confining potential profile along the direction orthogonal to the layers can be considered as a narrow triangular well (see, for example, [19]). Along this direction the wave function is quantized and has, for the lower, bound states, a discrete energy spectrum. The motion of the carriers in the plane parallel to the interfaces is free, so that, for any transverse energy, there exist a continuum of two-dimensional states called

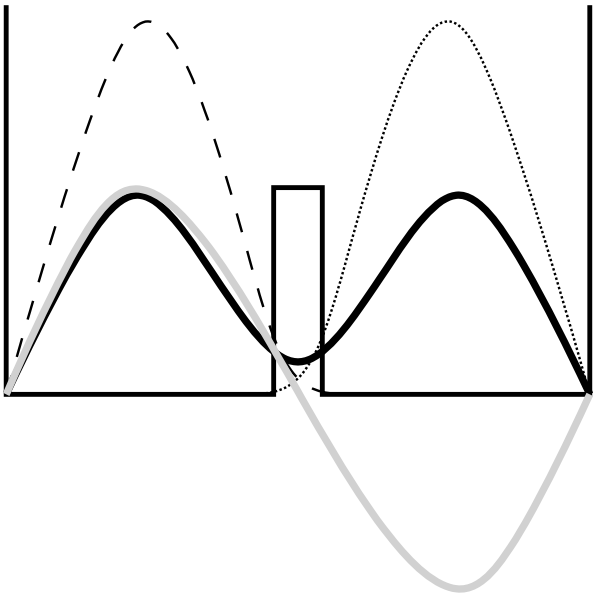

Charge-Based Solid-State Flying Qubits, Figure 1
Schematic transverse potential profile of two weakly coupled quantum wires (double well). The *dark* and *light solid lines* show the two lower eigenfunctions: $|\psi_e\rangle$ and $|\psi_o\rangle$, with even and odd parity, respectively. The *dashed* and *dotted lines* represent their sum ($|0\rangle$ state, localized in the left well) and difference ($|1\rangle$ state, localized in the right well), respectively

subbands. Such a 2DEG system is the basis from which quantum wire structures can be realized by lithographic process or electrostatic confinement. In the following, the confined direction, i. e. the material growth direction, is ignored since it is supposed that a single electron, injected in the system, will remain in the lowest subband.

The basic device consists of two parallel quantum wires [10]. It is supposed to operate at very low temperature in order to have a negligible number of electrons in the conduction band and to minimize the decoherence due to the interactions of the electrons with lattice vibrations. The transverse potential profile of the two-wire system is shown in Fig. 1 together with its ground eigenfunction $|\psi_e\rangle$, that has even parity, and the first excited state $|\psi_o\rangle$, that has odd parity.

Figure 1 shows also that the sum and difference of the two lower eigenfunctions are almost fully localized in the left and right well, respectively. The widths of the two wires are equal and constant along the whole device. This avoids any effect due to unalignment of subbands in the two quantum wires [24].

Let us consider a single electron injected into one of the two wires and assume that the injection process is able to keep the electron in the transverse ground state, and that, in the longitudinal direction along the wire, it is well described by a minimum uncertainty wave packet.

$$\psi(x,y) = \sqrt{\frac{2}{L}}\cos\left[\frac{\pi}{L}(x-x_0)\right] \frac{1}{\sqrt{\sigma\sqrt{2\pi}}} e^{-\left(\frac{y-y_0}{2\sigma}\right)^2} e^{ik_0 y}\,, \tag{19}$$

where, L is the width of the quantum wire, whose center is x_0, σ is the standard deviations of the Gaussian function and gives an estimate of the wave function extension in space along the wire direction, y_0 is the coordinate of the center of the wave function, k_0 is the wave number, representing the kinetic energy of the electron along the y direction. This is the initial condition for the following time-dependent numerical simulations.

The state of the qubit is defined as $|0\rangle$ if the electron is in the left wire, $|1\rangle$ if it is in the right wire. This means, in terms of transverse wave functions:

$$\begin{aligned}\langle x|0\rangle &= \frac{1}{\sqrt{2}}(\langle x|\psi_e\rangle + \langle x|\psi_o\rangle) = \sqrt{\frac{2}{L}}\cos\left(\frac{\pi}{L}(x-x_0)\right)\\ \langle x|1\rangle &= \frac{1}{\sqrt{2}}(\langle x|\psi_e\rangle - \langle x|\psi_o\rangle) = \sqrt{\frac{2}{L}}\cos\left(\frac{\pi}{L}(x-x_1)\right)\end{aligned} \tag{20}$$

where $x_0(x_1)$ is the central point of the left (right) well formed by the potential in the direction orthogonal to the wires.

The time evolution of the states defined by Eq. (20), as regards the x component, is simply given by a multiplying factor $e^{-i\omega t}$ with ω frequency of the two degenerate eigenstates.

One-Qubit Gates

If a coupling window is introduced between the wires, a significant lowering of the symmetric transverse state energy ω_e is produced. It can be easily shown that, with good approximation the energy of the anti-symmetric state is unchanged ($\omega_o = \omega$). It is well known (and it has been verified experimentally, see for example [18]) that the electronic wave function crossing this coupling region, oscillates between the wires with a period $\tau = 2\pi/(\omega - \omega_e)$. When the electron reaches the end of the coupling window, the oscillation process terminates and the wave function is eventually separated in two parts running along the left and the right wires, as shown in Fig. 2.

If T is the time that the electron spends in the coupling region, the transverse states in Eq. (20) evolve according

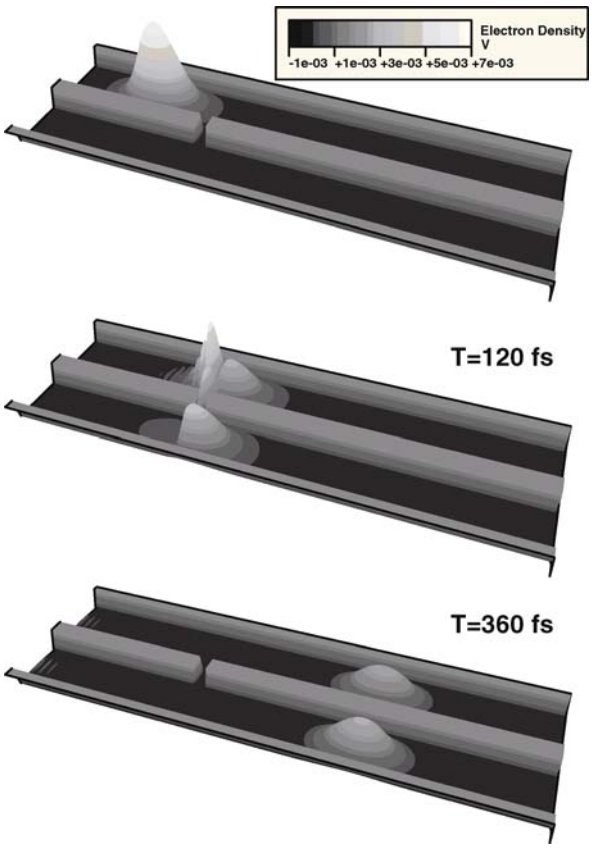

Charge-Based Solid-State Flying Qubits, Figure 2
Numerical solution of the time-dependent Schrödinger equation for one electron injected in a system of two coupled quantum wires [10]. The model potential profile and the square modulus of the wave function are shown at three different time steps. The width of the wires is $L = 6$ nm and the coupling window is 10 nm long. Note that different scales have been used for the two axes. The initial condition (in the upper image) is the wave function in Eq. (19) with $\sigma = 20$ nm and k_0 corresponding to an energy of 50 meV. The effective mass for GaAs has been used

to:

$$
\begin{aligned}
|0\rangle &\to \frac{e^{-i\omega T}}{\sqrt{2}}\left(e^{i(\omega-\omega_e)T}|\psi_e\rangle + |\psi_o\rangle\right) \\
&= e^{-i\omega T}\, e^{i\frac{\theta}{2}}\left(\cos\frac{\theta}{2}|0\rangle + i\sin\frac{\theta}{2}|1\rangle\right) \\
|1\rangle &\to \frac{e^{-i\omega T}}{\sqrt{2}}\left(e^{i(\omega-\omega_e)T}|\psi_e\rangle - |\psi_o\rangle\right) \\
&= e^{-i\omega T}\, e^{i\frac{\theta}{2}}\left(i\sin\frac{\theta}{2}|0\rangle + \cos\frac{\theta}{2}|1\rangle\right),
\end{aligned}
\tag{21}
$$

where $\theta = (\omega - \omega_e)T$. The length of the coupling window, the height of the barrier in it and the velocity of the electron, can be properly chosen to perform a given transfer of the electronic wave function between the wires. It is easy to see that, using the matrix formalism introduced in the previous section, the transformation $\mathbf{S}$ performed is represented by the rotation matrix $\mathbf{R_x}$ multiplied by a phase factor:

$$
\mathbf{S}(\theta) = e^{i\frac{\theta}{2}}\begin{pmatrix} \cos\left(\frac{\theta}{2}\right) & i\sin\left(\frac{\theta}{2}\right) \\ i\sin\left(\frac{\theta}{2}\right) & \cos\left(\frac{\theta}{2}\right)\end{pmatrix} = e^{i\frac{\theta}{2}}\mathbf{R_x}(\theta)\,. \tag{22}
$$

The global phase factor $e^{-i\omega T}$, due to the time evolution of the state, is ignored for simplicity: it is present in each of the transformations proposed and also in the free propagation of the electron, therefore it can be considered a global multiplying factor for the many-qubit system that gives no contribution.

Similarly, the exponential $e^{i\frac{\theta}{2}}$ of Eq. (22) results to be a global phase factor and can be neglected. Thus, in the following, the operation performed by the geometry described in this section (coupling window), will be indicated by either $\mathbf{S}(\theta)$ or $\mathbf{R_x}(\theta)$.

A second kind of transformation on a single qubit can be obtained with the introduction of a potential barrier able to delay the propagation of the wave function in one of the two wires (but low enough to avoid reflections [9]). This induces a phase shift between the two components of the qubit.

Let us suppose to have a qubit in a state $a|0\rangle + b|1\rangle$ (with a and b complex numbers and $|a^2 + b^2| = 1$). This means that the initial wave function $\psi_i(x, y)$: (a) in the longitudinal direction along the wires still has the form of a minimum uncertainty wave packet, (b) in the transverse x direction it is split in two parts, one in each wire, i. e., according to Eq. (20):

$$
\begin{aligned}
\psi_i(x, y) &= \left[a\cos\left(\frac{\pi}{L}(x - x_0)\right) + b\cos\left(\frac{\pi}{L}(x - x_1)\right)\right] \\
&\quad\times\sqrt{\frac{2}{L}}\frac{1}{\sqrt{\sigma\sqrt{2\pi}}}e^{-\left(\frac{y-y_0}{2\sigma}\right)^2}e^{ik_0y} \\
&= [a\langle x|0\rangle + b\langle x|1\rangle]\frac{1}{\sqrt{\sigma\sqrt{2\pi}}}e^{-\left(\frac{y-y_0}{2\sigma}\right)^2}e^{ik_0y}\,,
\end{aligned}
$$

or, as a column vector,

$$
\psi_i(x, y) = \frac{1}{\sqrt{\sigma\sqrt{2\pi}}}\begin{pmatrix} a\, e^{-(\frac{y-y_0}{2\sigma})^2}e^{ik_0y} \\ b\, e^{-(\frac{y-y_0}{2\sigma})^2}e^{ik_0y}\end{pmatrix}. \tag{23}
$$

It is known, from elementary quantum mechanics [30], that the effect of a potential barrier on a propagating wave function (like the one described in Eq. (19)) can be reduced, with good approximation, to a spatial delay ΔS,

provided either that (i) the mean kinetic energy of the wave packet is much higher than the barrier, or that (ii) $k_0 L$ is a multiple of π and the broadening in momentum space is small (σ is large).

If the delaying potential barrier is inserted on the 0 wire, only the $|0\rangle$ component of the wave function will undergo the spatial delay ΔS. The state of Eq. (23) will be transformed in

$$\psi_f(x,y) = \frac{1}{\sqrt{\sigma\sqrt{2\pi}}} \begin{pmatrix} a\, e^{-\left(\frac{y+\Delta S-y_0}{2\sigma}\right)^2} e^{ik_0(y+\Delta S)} \\ b\, e^{-\left(\frac{y-y_0}{2\sigma}\right)^2} e^{ik_0 y} \end{pmatrix} . \tag{24}$$

The delay ΔS is small compared to the parameter σ of the enveloping Gaussian function, thus the difference induced by the substitution $y \to y + \Delta y$ in the real exponential, is small compared to the variation of the imaginary exponential representing the enveloped plane wave. As a consequence it is possible to make the approximation

$$e^{-\left(\frac{y+\Delta S-y_0}{2\sigma}\right)^2} e^{ik_0(y+\Delta S)} \approx e^{-\left(\frac{y-y_0}{2\sigma}\right)^2} e^{ik_0(y+\Delta S)} \tag{25}$$

and to rewrite the state of Eq. (24) as

$$\psi_f(x,y) \approx \frac{1}{\sqrt{\sigma\sqrt{2\pi}}} e^{-\left(\frac{y-y_0}{2\sigma}\right)^2} e^{ik_0 y} \begin{pmatrix} a\, e^{ik_0\Delta S} \\ b \end{pmatrix} . \tag{26}$$

We introduce, now, an angle ϕ defined as

$$\phi = k_0 \Delta S . \tag{27}$$

It is easy to verify that the transformation $\psi_i \longrightarrow \psi_f$ on the qubit state is given by

$$\mathbf{R_0}(\phi) = \begin{pmatrix} e^{i\phi} & 0 \\ 0 & 1 \end{pmatrix} , \quad \text{or} \quad \mathbf{R_1}(\phi) = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\phi} \end{pmatrix} , \tag{28}$$

if the delay is realized on the wire representing the 0 or 1 component, respectively. The value of ϕ depends on the delay ΔS that depends, in turn, on the height and length of the potential barrier.

The transformations $\mathbf{R_0}$ and $\mathbf{R_1}$ can be combined to obtain a transformation $\mathbf{R_z}$, i. e. the rotation around the $\hat{z}$ axis defined in Eq. (13), that is rewritten here for convenience:

$$\mathbf{R_0}(\alpha/2)\mathbf{R_1}(-\alpha/2) = \mathbf{R_z}(\alpha) = \begin{pmatrix} e^{i\frac{\alpha}{2}} & 0 \\ 0 & e^{-i\frac{\alpha}{2}} \end{pmatrix} . \tag{29}$$

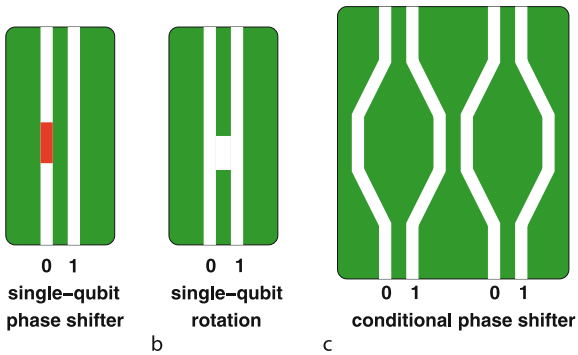

Charge-Based Solid-State Flying Qubits, Figure 3
Schematic representation of the three devices that constitute the universal set of quantum gates: a a delaying potential barrier (*red*) along the 0 wire realizes the R_0 transformation; b a coupling window between the wires realizes the R_x transformation; c a two-qubit Coulomb coupler is able to entangle the two qubit states and to realize a T transformation (see Eq. (18))

As described in the previous section, $\mathbf{R_z}$ together with $\mathbf{R_x}$, gives any rotation of the group SU(2). Thus, the two basic transformations proposed are able to realize the general rotation $\mathbf{U}$ of Eq. (10) that is the single-qubit gate of the universal set for the proposed qubit system.

A schematic representation of the two devices implementing the $\mathbf{R_0}$ and $\mathbf{R_x}$ transformations are reported in Fig. 3 (a) and (b), respectively.

It is straightforward to note that the transformations $\mathbf{R_x}$ and $\mathbf{R_z}$ for the electronic wave function are analogues to the transformations induced, in a two-path interferometer, by a photon beam splitter and a photon phase shifter. Furthermore, a simple calculation, that we omit for brevity, shows that the Hadamard gate of Eq. (6) can be obtained by using the above transformations.

Two-Qubit Gate

In order to implement the conditional phase shifter gate T, whose operation is described in Eq. (18), two qubits must be considered. Figure 3c represents the geometry of the two-qubit device. In the following we will call "first" the qubit represented by the left couple of wires, "second" the one represented by the right couple. In the two-qubit binary encoding of the state of the system, the second qubit is the less significant bit.

The device is designed as follows [10]. The four wires of the two qubits run parallel to each other and, in a small region, the wire 1 of the first qubit and the wire 0 of the second qubit get close to each other, enough to give rise to a significant Coulomb coupling between two electrons running along them. In fact, when the two electrons ap-

proach each other, part of their kinetic energy is transformed into a repulsive coupling potential and the velocity along the wires is reduced. Then, when the distance between them increases again to the original value, the potential energy is transformed back into kinetic energy and the initial velocity is restored. As a result of this process, the two electrons running along the right wire (state $|1\rangle$) of the first qubit and the left wire (state $|0\rangle$) of the second qubit, will suffer a delay in their propagation compared to the case in which either the first or the second qubit are not in one of the two central wires. This delay depends on the length of the coupling region and on the distance between the two central wires, and corresponds to a phase factor in front of the $|10\rangle$ component of the two-qubit system.

The transformation T induced by the described geometry (shown in Fig. 3c) can be also understood by analyzing the effect on the two electrons separately. If the first electron is in state $|1\rangle$ (the electron is in the wire near the second qubit), the slowing down caused on the second qubit is similar to the slowing down caused by a delaying potential barrier and the transformation on the second qubit is of the kind $\mathbf{R_0}$; similarly, if the second qubit is $|0\rangle$ the transformation on the first qubit is $\mathbf{R_1}$. As can be easily seen by direct inspection, if the whole two-qubits system is considered, the behavior described above gives rise to the two-qubits transformation T.

This proposal for the physical realization of a two-qubit gate is alternative to the one advanced in [25], where a CNOT gate was proposed (instead of the T gate proposed here) exploiting the difference in the oscillation period, induced by the Coulomb interaction between two electrons. It must be noted that, together with single qubit rotations, the conditional phase shifter is the basic transformation for the quantum discrete Fourier transform, needed to implement Shor's algorithm [17]. Starting from the CNOT gate, a quantum network is needed to implement a **T** transformation:

$$\mathbf{T}(\gamma) = \mathbf{R_1}^{(1)}\left(\frac{\gamma}{2}\right)\mathbf{R_1}^{(2)}\left(-\frac{\gamma}{2}\right)\text{CNOT}\ \ \mathbf{R_1}^{(2)}\left(\frac{\gamma}{2}\right)\text{CNOT}. \tag{30}$$

This network is simple, but the present proposal can avoid this step and directly realize the conditional phase shifter.

Electrons Driven by Surface Acoustic Waves

As an alternative to the ballistic propagation of electrons along the quantum wires, the use of surface acoustic waves (SAW) has been proposed. The SAW technology presents the drawback of some local heating near the SAW transducer. However, Barnes et al. [5] have successfully applied the SAW technique for the electron injection and transport along quantum wires. Within this approach a number of electrons are captured from a 2DEG and placed into the minima of a sinusoidal acoustic wave that propagates along the device. The 2DEG region is connected to the quantum-wire region in which a single electron must be injected. When the SAW minimum reaches the 1D channel, the trapped electrons undergo a further confinement due to the lateral potential that constitutes the wire, so that a moving quantum dot is formed. With a suitable choice of the SAW parameters it is possible to create a dot that carries a single electron. The electron moving inside the 1D channel is embedded in the minimum of the dot and, like in the free-propagation case, it experiences the cascade of

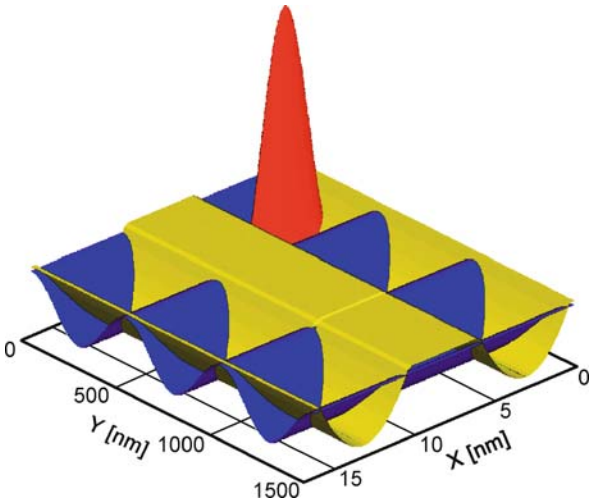

Charge-Based Solid-State Flying Qubits, Figure 4
Electron wave function (*red*) driven along a double wire device by the time-dependent sinusoidal potential of the SAW (*blue*) that are propagating from left to right

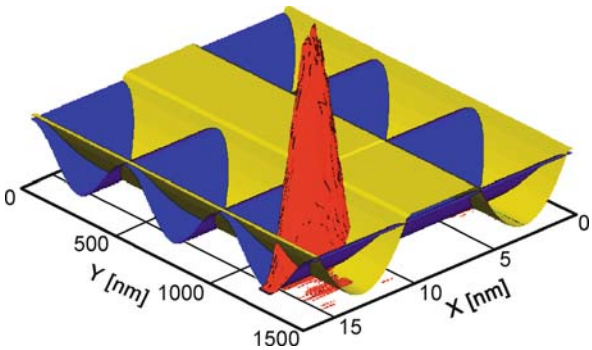

Charge-Based Solid-State Flying Qubits, Figure 5
Electron wave function (*red*) of Fig. 4 after the R_x transformation induced by the inter-wire coupling window at $y = 1000$ nm

quantum gates. In this way, the wave-packet could be more immune to decoherence effects, as it is confined along the three dimensions.

Figures 4 and 5 show the results, at initial and final time, respectively, of a time-dependent numerical simulation of a $\mathbf{R_x}$ gate device with SAW. Because of the long wave length of the SAWs (about 500 nm here), the spatial spreading of the electron wave function is large. Furthermore, the coupling window, located at $y = 1000$ nm, is small and hardly visible. However, the system parameters have been tuned in order to obtain a NOT operation and the electron wave function embedded in the SAW minimum is transferred from the wire 1 to the wire 0, thus confirming the proper functioning of the device.

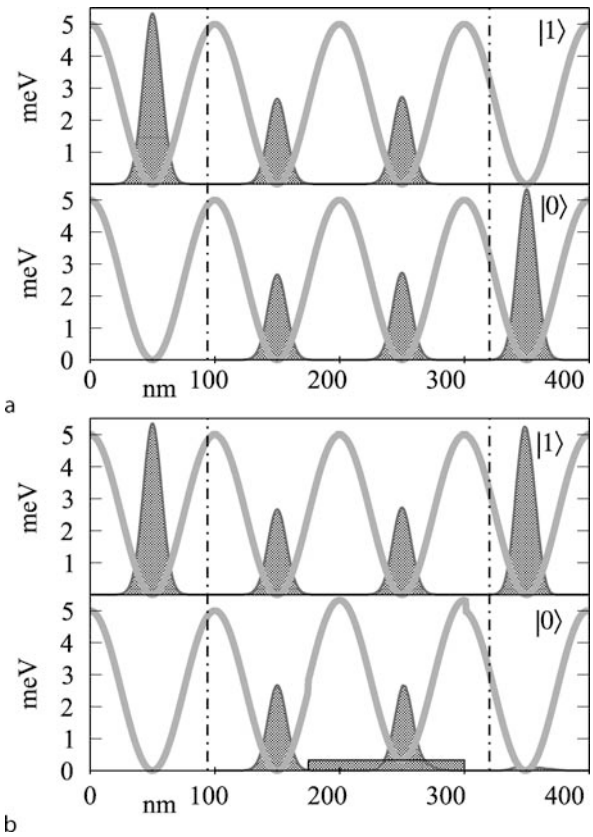

Charge-Based Solid-State Flying Qubits, Figure 6
One-dimensional time-dependent simulation of an electron wave function trapped inside a minimum of a surface acoustic wave (*sinusoidal solid line*), propagating from the left to the right. The square modulus of the wave function is shown at four different time steps (*shaded regions*). The coupled-wire device has two single-qubit rotations $R_x(\pi/2)$ (represented by *dashed-dotted lines*). In the *lower graph* a phase shifter $R_z(\pi)$ (small potential around 200–300 nm) is inserted in the wire 0 between the rotations, leading to a different final state of the qubit

While the action of the coupling window ($\mathbf{R_x}$ transformation) is not altered in the SAW approach, the functioning of the gates $\mathbf{R_z}$ and $\mathbf{T}$ need to be revised [12]. In fact, the confinement along the wire, due to the SAW, prevents the spatial delay of the wave packet, since the velocity of the electron is fixed by the SAW velocity. However, also in this case the $\mathbf{R_z}$ transformation can be realized by means of a potential barrier along the wire. Now the phase shift originates from the change of the energy levels of the moving dot, caused by the additional potential of the barrier. The $\mathbf{R_z}$ and $\mathbf{T}$ transformations in the frame of the SAW approach have been validated by numerical simulations and an example is shown in Fig. 6, where two 1D simulations for a SAW-driven electron are reported [8]. The wave function, shown at four different time steps, is initialized in the ground state of the sinusoidal SAW potential, inside the $|1\rangle$-state wire. Two quantum gate networks are simulated, namely $\mathbf{R_x}(\pi/2)\mathbf{R_x}(\pi/2)$ (with two coupling windows) and $\mathbf{R_x}(\pi/2)\mathbf{R_z}(\pi)\mathbf{R_x}(\pi/2)$ (with a coupling window, a potential barrier and a second coupling window) in the two uppermost and the two lowermost plots, respectively. In the second case the presence of a small potential barrier between the two splitting windows is able to change the final state of the qubit, as in the case of free propagation of the carrier. The use of SAW-induced electron interferometry in QWRs is being actively investigated both theoretically [13,23,37] and experimentally [15].

Entangled States and Estimation of Quantum Entanglement

The simulation presented in the previous section, is the result of a time-dependent Schrödinger solver that evolves the electrons in a single particle approximation. This is necessary due to the high computational effort that is needed for a ballistic evolution, with Coulomb interaction, in two dimensions. In the above approach, each electron wave-function $\psi(x, y)$ is represented, at a given time step, by a two-dimensional array. This means that no entanglement between electrons can be simulated. Nevertheless, the simulations performed validated the devices using a base of factorizable qubit states: the effect on entangled states can be easily deduced once the behavior on the given basis is known, thanks to the superposition principle. For example, the result of a transformation $\mathbf{R_x}^{(2)}(\pi/2)\mathbf{T}(\pi)\mathbf{R_x}^{(2)}(\pi/2)$ on a state $(|00\rangle + |10\rangle)$ is an entangled state:

$$\mathbf{R_x}^{(2)}(\pi/2)\mathbf{T}(\pi)\mathbf{R_x}^{(2)}(\pi/2)(|00\rangle + |10\rangle) = -(|01\rangle + i|10\rangle)\,. \quad (31)$$

The condition of separability imposed on the wave function will be dropped in this section. In fact, the possibility to simulate an entangled two-electron state allows the study of the physical system proposed in [26] for the test of Bell inequality using quantum-wire semiconductor devices.

Since the straightforward numerical simulation, in two dimensions, of the time evolution of $\psi(x_1, y_1, x_2, y_2, t)$ results to be too expensive from a computational point of view, a simpler semi-1D model can be used to study the proposed quantum wire system [11]. Let us consider, for the moment, a single qubit, i. e. an electron running along a couple of QWRs, with its wave function $\psi(x_1, y_1)$ eventually split between the wires. If the transverse component (orthogonal to the wires) of the electron wave function remains in the ground state of the transverse potential, i. e. the inter-subband scattering is negligible, then, along the device, the transverse dynamics can be neglected. This is obviously false in the regions of the $\mathbf{R_x}$ coupling windows, which will be considered in the following.

Two possible states of the qubit can be used to label the two possible forms for the transverse wave function, a cosine centered either in the wire 0 or 1:

$$\begin{aligned} x_1 = 0 &\longrightarrow \langle x|0\rangle = \sqrt{\frac{2}{L}}\cos\left(\frac{\pi}{L}(x - X_0)\right) \\ x_1 = 1 &\longrightarrow \langle x|1\rangle = \sqrt{\frac{2}{L}}\cos\left(\frac{\pi}{L}(x - X_1)\right) , \end{aligned} \tag{32}$$

where X_0 and X_1 are the central points of wires 0 and 1, respectively, and L is their width.

Concerning the $\mathbf{R_x}(\theta)$ transformation, within this semi-1D model, it is not possible to simulate directly the wave function splitting induced by the coupling window. In fact the wave function evolution that originates the transformation comes from the transverse dynamics whose detailed description is now lost. It is, however, possible to directly use the analytical expression of $\mathbf{R_x}$ in order to account for the coupling window, once it has been validated through numerical simulations [11]. The above considerations allow one to include the effect of an $\mathbf{R_x}(\theta)$ transformation, directly by the application of the proper transformation matrix. A similar approach has been used to include the two beam-splitters in the simulations of Fig. 6.

Let us consider now the third gate, namely the two-qubit conditional phase shifter sketched in Fig. 3c. As for the single-qubit case, it is possible to associate to each wire of the first or second qubit, a given value (either 0 or 1) of the x coordinate, considered now as a parameter. This is done for both the first and second qubit. With this approach, the model is strongly simplified: from a time-dependent Schrödinger equation for the five-variable wave function $\psi(x_1, y_1, x_2, y_2, t)$, to four equations:

$$\begin{aligned} i\hbar\frac{\partial}{\partial t}\psi_{x_1,x_2}(y_1, y_2, t) = &-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial y_1^2} + \frac{\partial^2}{\partial y_2^2}\right) \\ &\cdot \psi_{x_1,x_2}(y_1, y_2, t) \\ &+ V_{x_1,x_2}(y_1, y_2)\psi_{x_1,x_2}(y_1, y_2, t) \end{aligned} \tag{33}$$

with $x_1, x_2 \in \{0, 1\}$. During the time evolution, the four different components of the wave function

$$\psi_{x_1,x_2}(y_1, y_2, t) \quad \text{with} \quad x_1, x_2 \in \{0, 1\} , \tag{34}$$

are coupled by the $\mathbf{R_x}^{(1)}$ and $\mathbf{R_x}^{(2)}$ transformations, given by (see Eq. (9) for the notation):

$$\mathbf{R_x}^{(1)}(\theta) = \begin{pmatrix} \cos\left(\frac{\theta}{2}\right) & 0 & i\sin\left(\frac{\theta}{2}\right) & 0 \\ 0 & \cos\left(\frac{\theta}{2}\right) & 0 & i\sin\left(\frac{\theta}{2}\right) \\ i\sin\left(\frac{\theta}{2}\right) & 0 & \cos\left(\frac{\theta}{2}\right) & 0 \\ 0 & i\sin\left(\frac{\theta}{2}\right) & 0 & \cos\left(\frac{\theta}{2}\right) \end{pmatrix} , \tag{35}$$

$$\mathbf{R_x}^{(2)}(\theta) = \begin{pmatrix} \cos\left(\frac{\theta}{2}\right) & i\sin\left(\frac{\theta}{2}\right) & 0 & 0 \\ i\sin\left(\frac{\theta}{2}\right) & \cos\left(\frac{\theta}{2}\right) & 0 & 0 \\ 0 & 0 & \cos\left(\frac{\theta}{2}\right) & i\sin\left(\frac{\theta}{2}\right) \\ 0 & 0 & i\sin\left(\frac{\theta}{2}\right) & \cos\left(\frac{\theta}{2}\right) \end{pmatrix} . \tag{36}$$

The geometry of the system is contained in the two-particle potential $V_{x_1,x_2}(y_1, y_2)$. It consists of three terms: the two structure potentials along the wires 0 and 1 of each qubit, and the Coulomb interaction between the electrons:

$$V_{x_1,x_2}(y_1, y_2) = U_{x_1}(y_1) + V_{x_2}(y_2) + \frac{e^2}{\epsilon D_{x_1,x_2}(y_1, y_2)} . \tag{37}$$

where $D_{x_1,x_2}(y_1, y_2)$ represents the distance between point y_1 in x_1 wire of the first qubit and point y_2 in x_2 wire of the second qubit and is given explicitly by

$$D_{x_1,x_2}(y_1, y_2) = \sqrt{[p_{x_1}(y_1) - q_{x_2}(y_2)]^2 + [y_1 - y_2]^2} \tag{38}$$

where $p_{x_1}(y_1)$ is the x coordinate of a point that has y coordinate y_1 on the wire representing the x_1 state of the first qubit.

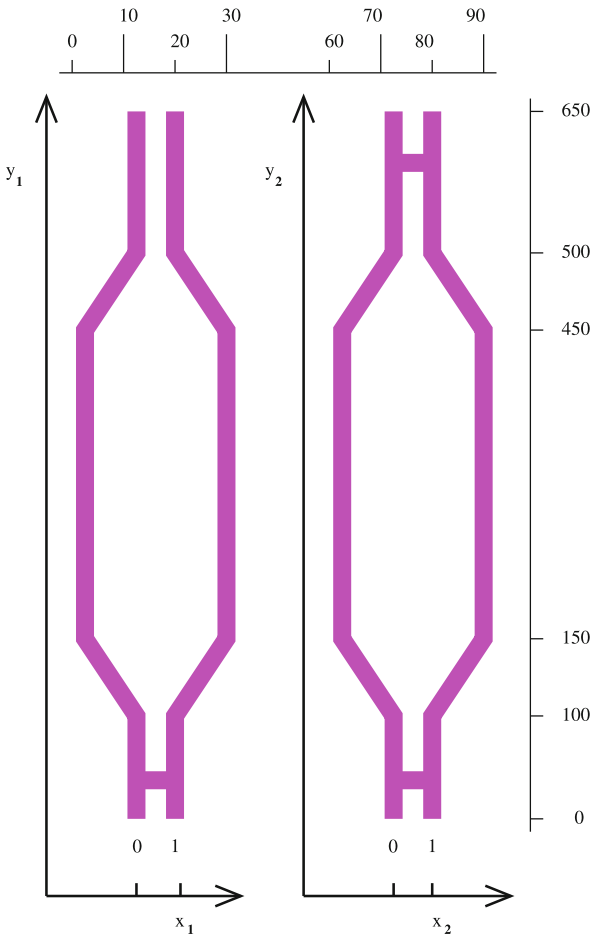

Charge-Based Solid-State Flying Qubits, Figure 7
QWRs network that realizes the transformation $R_x^{(2)}(\pi/2)T(\pi)R_x^{(2)}(\pi/2)R_x^{(1)}(\pi/2)$. The geometry shown is able to produce the four maximally entangled two-qubit Bell states [11]

Since the phase shift $\mathbf{R_0}$ of the first qubit is obtained with a delaying potential barrier inserted along a wire, its functioning is not altered with respect to the full-2D model. The potential barrier is accounted for in potential $U_0(y_1)$. Similarly, for the second qubit the potential is inserted in $V_0(y_1)$. We stress that, within the proposed universal set of gates, the conditional phase shifter is the only two-qubit gate and, as a consequence, the only one able to produce an entangled state.

In conclusion, let us present the result of a numerical simulation of a simple gate network able to produce a maximally entangled Bell's state. The transformation is

$$\mathbf{R_x}^{(1)}(\pi/2)\mathbf{R_x}^{(2)}(\pi/2)\mathbf{T}(\pi)\mathbf{R_x}^{(2)}(\pi/2)\,, \tag{39}$$

and the network obtained is schematized in Fig. 7. Two electrons are injected in the device. One electron is injected in the wire 1 of the first qubit and one in the wire 0 of the second qubit. The initial state is $|10\rangle$. It is easy to verify that the network described will perform the transformation:

$$\mathbf{R_x}^{(2)}(\pi/2)\mathbf{T}(\pi)\mathbf{R_x}^{(2)}(\pi/2)\mathbf{R_x}^{(1)}(\pi/2)|10\rangle = \frac{1-i}{2}(|01\rangle + |10\rangle)\,. \tag{40}$$

The numerical simulations performed, and shown in Fig. 8, confirm the good functioning of the device and the production of the entangled Bell's state

$$|\Psi^+\rangle = \frac{(|01\rangle + |10\rangle)}{\sqrt{2}}\,. \tag{41}$$

Future Directions

The devices presented in this work have, in principle, the full potentiality to create entangled states of electrons and to perform the logic operations of the universal quantum set of gates. For the modeling of realistic structures, we refer to the literature [1,29,32,38,42].

Concerning experimental realizations of prototypes of the single-qubit gates, we note that Pingue et al. [31] obtained the evidence of switching capabilities of the coupled-wire device, but the presence of occupied localized states in a quantum wire near the coupling window, also predicted in [29], and the Coulomb blockade regime of the experiment, made the device not suitable as a quantum gate.

Both vertical [20,22] and planar [35,36] double-channel structures are currently under investigation. In particular, Ramamoorty et al. [36] obtained switching characteristics whose temperature dependence show a clear signature of coherent behavior of a coupling-window device. It is also worth noting that Fischer et al. demonstrated that the control of the coupling between two modes of two 1D channels is feasible. The two modes were obtained exploiting the two minima of the conduction-band edge in the growth direction of a GaAs 2DEG [21] of two vertically coupled 2DEGs [22]: in both cases clear signatures of coherent coupling have been observed.

Acknowledgment

I would like to express my sincere thanks to the groups of Carlo Jacoboni and Massimo Rudan, that performed, together with the author, a large part of the work here reported. Furthermore, I thank, for most fruitful discussions, Marco Fanciulli and Peter Vogl.

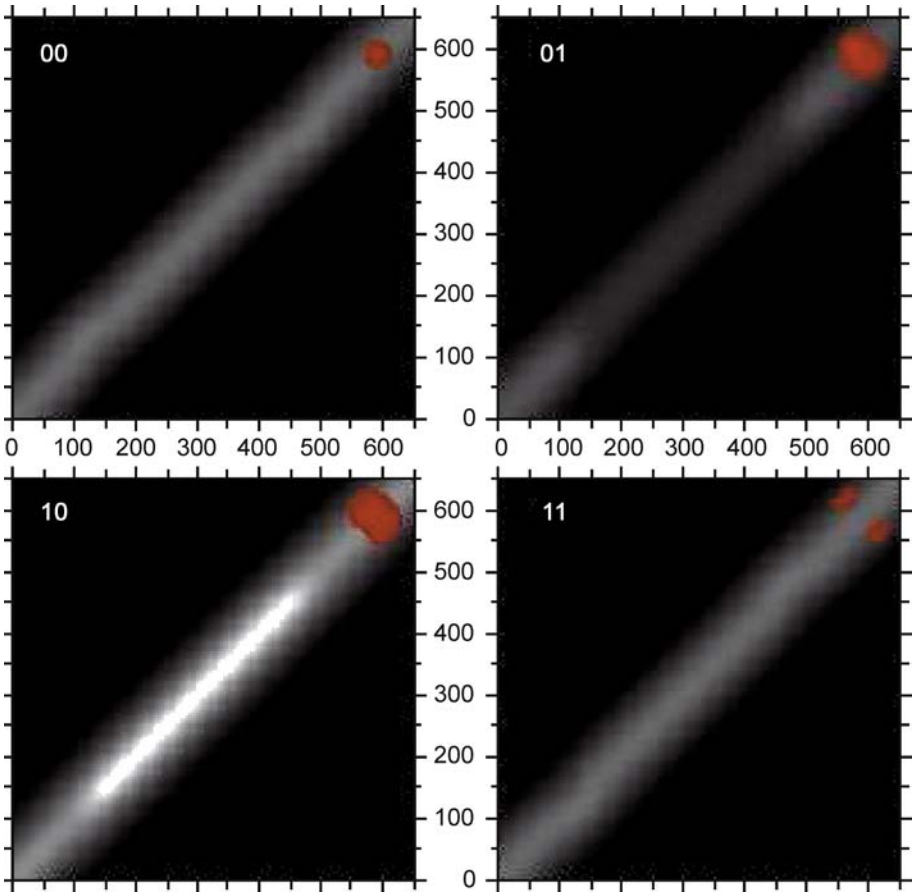

Charge-Based Solid-State Flying Qubits, Figure 8

Square modulus of the two-particle wave function (*red*) at the final time step of the simulation represented in Eq. (40). The four graphs represent the cases: $x_1 = 0, x_2 = 0; x_1 = 0, x_2 = 1; x_1 = 1, x_2 = 0; x_1 = 1, x_2 = 1$, as indicated in the *upper left corners* (see Eq. (33)). The *vertical and horizontal axes* of each graph represent the position y_1 of the first electron and y_2 of the second electron, respectively. The *white region along the diagonals* represent the electron–electron potential given by Eq. (37) [11]

Bibliography

Primary Literature

1. Akguc GB, Reichl LE, Shaji A, Snyder MG (2004) Bell states in a resonant quantum waveguide network. Phys Rev A 69:42303
2. ARDA (2004) Quantum Computation Roadmap. http://qist.lanl.gov. Accessed 1 June 2008
3. Aspect A, Grangier P, Roger G (1981) Experimental tests of realistic local theories via Bell's theorem. Phys Rev Lett 47:460
4. Barenco A, Bennett CH, Cleve R, DiVincenzo DP, Margolus N, Shor P, Sleator T, Smolin JA, Weinfurter H (1995) Elementary gates for quantum computation. Phys Rev A 52:3457
5. Barnes CHW, Shilton JM, Robinson AM (2000) Quantum computation using electrons trapped by surface accoustic waves. Phys Rev B 62:8410
6. Bell JS (1964) On Einstein Podolsky Rosen paradox Physics 1:195
7. Bennett CH (1982) The thermodynamics of computation – a review. Int J Theor Phys 21:905
8. Bertoni A (2007) Perspectives on solid-state flying qubits. J Comp Electr 6:67
9. Bertoni A, Bordone P, Brunetti R, Jacoboni C, Reggiani S (2000) Quantum logic gates based on coherent electron transport in quantum wires. Phys Rev Lett 84:5912
10. Bertoni A, Bordone P, Brunetti R, Jacoboni C, Reggiani S (2002) Numerical simulation of coherent transport in quantum wires for quantum computing. J Mod Opt 49:1219
11. Bertoni A, Ionicioiu R, Zanardi P, Rossi F, Jacoboni C (2002) Simulation of entangled electronic states in semiconductor quantum wires. Physica B 314:10
12. Bertoni A, Reggiani S (2004) Entanglement and quantum computing with ballistic electrons. Semicond Sci Technol 19:S113
13. Bordone P, Bertoni A, Rosini M, Reggiani S, Jacoboni C (2004) Coherent transport in coupled quantum wires assisted by surface acoustic waves. Semicond Sci Technol 19:412
14. Braunstein SL, Mann A, Revzen M (1992) Maximal violation of bell inequalities for mixed states. Phys Rev Lett 68:3259
15. Cunningham J, Pepper M, Talyanskii VI, Ritchie DA (2005)

Acoustic transport of electrons in parallel quantum wires. Acta Phys Polonica A 107:38
16. Einstein A, Podolsky B, Rosen N (1935) Can quantum-mechanical description of physical reality be considered complete? Phys Rev 47:777
17. Ekert A, Jozsa R (1996) Quantum computation and shor's factoring algorithm. Rev Mod Phys 68:733
18. Eugster CC, delAlamo JA, Rooks MJ, Melloch MR (1994) One-dimensional to onedimensional tunneling between electron waveguides. Appl Phys Lett 64:3157
19. Ferry DK (1997) Transport in Nanostructures. Cambridge University Press, Cambridge
20. Fischer SF (2007) Magnetotransport spectroscopy of dual one-dimensional electron systems. Int J Mod Phys B 21:1326
21. Fischer SF, Apetrii G, Kunze U, Schuh D, Abstreiter G (2005) Tunnel-coupled one-dimensional electron systems with large subband separations. Phys Rev B 74:115324
22. Fischer SF, Apetrii G, Kunze U, Schuh D, Abstreiter G (2006) Energy spectroscopy of controlled coupled quantum-wire states. Nature Phys 2:91
23. Furuta S, Barnes CH, Doran CJL(2004) Single-qubit gates and measurements in the surface acoustic wave quantum computer. Phys Rev B 70:205320
24. Governale M, Macucci M, Pellegrini P (2000) Shape of the tunneling conductance peaks for coupled electron waveguides. Phys Rev B 62:4557
25. Ionicioiu R, Amaratunga, Udrea F (2001) Quantum computation with balistic electrons. Int J Mod Phys B 15:125
26. Ionicioiu R, Zanardi P, Rossi F (2001) Testing bell's inequality with ballistic electrons in semiconductors. Phys Rev A 63: 50101
27. Knill E, Laflamme R, Martinez R, Tseng C-H (2000) An algorithmic benchmark for quantum information processing. Nature 404:368
28. Landauer R (1961) Irreversibility and heat generation in the computing process. IBM J Res Dev 5:183
29. Marchi A, Bertoni A, Reggiani S, Rudan M (2004) Investigation on single-electron dynamics in coupled gaas-algaas quantum wires. IEEE Trans Nanotech 3:129
30. Messiah A (1961) Quantum Mechanics. North-Holland, Amsterdam
31. Pingue P, Piazza V, Beltram F, Farrer I, Ritchie DA, Pepper M (2005) Coulomb blockade directional coupler. Appl Phys Lett 86:52102
32. Polizzi E, Ben Abdallah N (2002) Self-consistent three-dimensional models for quantum ballistic transport in open systems. Phys Rev B 66:245301
33. Preskill J (1998) Caltech lecture notes. http://www.theory.caltech.edu/~preskill/ph229. Accessed 1 June 2008
34. Preskill J (1998) Reliable quantum computers. Proc R Soc Lond A 454:3851
35. Ramamoorthy A, Bird JP, Reno JL (2006) Quantum asymmetry of switching in laterally coupled quantum wires with tunable coupling strength. Appl Phys Lett 89:153128
36. Ramamoorthy A, Bird JP, Reno JL (2006) Switching characteristics of coupled quantum wires with tunable coupling strength. Appl Phys Lett 89:13118
37. Rodriquez R, Oi DK, Kataoka M, Barnes CH, Ohshima T, Ekert AK (2005) Surface-acoustic-wave single-electron interferometry. Phys Rev B 72:85329
38. Sabathil M, Mamaluy D, Vogl P (2004) Prediction of a realistic quantum logic gate using the contact block reduction method. Semicond Sci Technol 19:S137
39. Tinkham M (1964) Group Theory and Quantum Mechanics. McGraw-Hill, New York
40. DiVincenzo DP (1995) Two-bit gates are universal for quantum computation. Phys Rev A 50:1015
41. DiVincenzo D (2000) The physical implementation of quantum computation. quant-ph/0002077
42. Zibold T, Vogl P, Bertoni A (2007) Theory of semiconductor quantum-wire based single- and two-qubit gates. Phys Rev B 76:195301

Books and Reviews

Benenti G, Casati G, Strini G (2004) Principles of Quantum Computation and Information, vol 1. World Scientific, Singapore. ISBN 9-812-38858-3

DiVincenzo DP (1995) Quantum Computation. Science 270:255

Harrison P (2005) Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures. Wiley-Interscience, Chichester. ISBN 0-470-01080-0

Hurt NE (2000) Mathematical Physics of Quantum Wires and Devices: From Spectral Resonances to Anderson Localization. Kluwer, Dordrecht. ISBN 0-792-36288-8

Nielsen M, Chuang I (2000) Quantum Computation and Quantum Information. Cambridge University Press, Cambridge. ISBN 0-521-63503-9

Chronological Calculus in Systems and Control Theory

MATTHIAS KAWSKI
Department of Mathematics, Arizona State University, Tempe, USA

Article Outline

Glossary
Definition of the Subject
Introduction, History, and Background
Fundamental Notions of the Chronological Calculus
Systems That Are Affine in the Control
Future Directions
Acknowledgments
Bibliography

Glossary

Controllability A control system is controllable if for every pair of points (*states*) p and q there exists an admissible control such that the corresponding solution curve that starts at p ends at q. Local controllability about a point means that all states in some open neighborhood can be reached.

Pontryagin Maximum Principle of optimal control Optimality of a control-trajectory pair geometrically is a property *dual* to local controllability in the sense that an optimal trajectory (endpoint) lies on the boundary of the reachable sets (after possibly augmenting the state of the system by the running cost). The maximum principle is a necessary condition for optimality. Geometrically it is based on analyzing the effect of families of control variations on the endpoint map. The chronological calculus much facilitates this analysis.

$\mathcal{E}(M) = C^\infty(M)$ The algebra of smooth functions on a finite dimensional manifold M, endowed with the topology of uniform convergence of derivatives of all orders on compact sets.

$\Gamma^\infty(M)$ The space of smooth vector fields on the manifold M.

Chronological calculus An approach to systems theory based on a functional analytic operator calculus, that replaces nonlinear objects such as smooth manifolds by infinite dimensional linear ones, by commutative algebras of smooth functions.

Chronological algebra A linear space with a bilinear product $\star$ that satisfies the identity $a \star (b \star c) - b \star (a \star c) = (a \star b) \star c - (b \star a) \star c$. This structure arises naturally via the product $(f \star g)_t = \int_0^t [f_s, g'_s] \mathrm{d}s$ of time-varying vector fields f and g in the chronological calculus. Here $[\ ,\ ]$ denotes the Lie bracket.

Zinbiel algebra A linear space with a bilinear product that satisfies the identity $a * (b * c) = (a * b) * c + (b * a) * c$. This structure arises naturally in the special case of affine control systems for the product $(U * V)(t) = \int_0^t U(s)V'(s)\mathrm{d}s$ of absolutely continuous scalar valued functions U and V.

The name Zinbiel is Leibniz read backwards, reflecting the duality with Leibniz algebras, a form of noncommutative Lie algebras. There has been some confusion in the literature with Zinbiel algebras incorrectly been called chronological algebras.

$\mathcal{II}\mathcal{F}(\mathcal{U}^Z)$ For a suitable space $\mathcal{U}$ of *time-varying scalars*, e. g. the space of locally absolutely continuous real-valued functions defined on a fixed time interval, and an indexing set Z, $\mathcal{II}\mathcal{F}(\mathcal{U}^Z)$ denotes the space of iterated integral functionals from the space of Z-tuples with values in $\mathcal{U}$ to the space $\mathcal{U}$.

Definition of the Subject

The chronological calculus is a functional analytic operator calculus tool for nonlinear systems theory. The central idea is to replace nonlinear objects by linear ones, in particular, smooth manifolds by commutative algebras of smooth functions. Aside from its elegance, its main virtue is to provide tools for problems that otherwise would effectively be untractable, and to provide new avenues to investigate the underlying geometry. Originally conceived to investigate problems in optimization and control, specifically for extending Pontryagin's Maximum Principle, the chronological calculus continues to establish itself as the preferred language of geometric control theory, and it is spawning new sets of problems, including its own set of algebraic structures that are now studied in their own right.

Introduction, History, and Background

This section starts with a brief historical survey of some landmarks that locate the chronological calculus at the interface of systems and control theory with functional analysis. It is understood that such brief survey cannot possibly do justice to the many contributors. Selected references given are meant to merely serve as starting points for the interested reader.

Many problems in modern systems and control theory are inherently nonlinear and e. g., due to conserved quantities or symmetries, naturally *live* on manifolds rather than on Euclidean spaces. A simple example is the problem of stabilizing the attitude of a satellite via feedback controls. In this case the natural state space is the tangent bundle $TSO(3)$ of a rotation group. The controlled dynamics are described by generally nonautonomous nonlinear differential equations. A key characteristic of their flows is their general lack of commutativity. Solutions of nonlinear differential equations generally do not admit closed form expressions in terms of the traditional sets of elementary functions and symbols. The chronological calculus circumvents such difficulties by reformulating systems and control problems in a different setting which is infinite dimensional, but linear. Building on well established tools and theories from functional analysis, it develops a new formalism and a precise language designed to facilitate studies of nonlinear systems and control theory.

The basic plan of replacing nonlinear objects by linear ones, in particular smooth manifolds by commutative algebras of smooth functions, has a long history. While its roots go back even further, arguably this approach gained much of its momentum with the path-breaking innovations by John von Neumann's work on the *"Mathematical Foundations of Quantum Mechanics"* [104] and Marshall Stone's seminal work on *"Linear Transformations in Hilbert Space"* [89], quickly followed by Israel Gelfand's dissertation on *"Abstract Functions and Linear Operators"* [34] (published in 1938). The fundamental concept of

maximal ideal justifies this approach of identifying manifolds with commutative normed rings (algebras), and vice versa. Gelfand's work is described as uniting previously uncoordinated facts and revealing the close connections between classical analysis and abstract functional analysis [102]. In the seventy years since, the study of Banach algebras, C^*-algebras and their brethren has continued to develop into a flourishing research area.

In the different arena of systems and control theory, major innovations at formalizing the subject were made in the 1950s. The Pontryagin Maximum Principle [8] of optimal control theory went far beyond the classical calculus of variations. At roughly the same time Kalman and his peers introduced the Kalman filter [51] for extracting signals from noisy observations, and pioneered state-space approaches in linear systems theory, developing the fundamental concepts of controllability and observability. Linear systems theory has bloomed and grown into a vast array of subdisciplines with ubiquitous applications. Via well-established transform techniques, linear systems lend themselves to be studied in the *frequency domain*. In such settings, systems are represented by linear operators on spaces of functions of a complex variable. Starting in the 1970s new efforts concentrated on rigorously extending linear systems and control theory to nonlinear settings. Two complementary major threads emerged that rely on differential geometric methods, and on operators represented by formal power series, respectively. The first is exemplified in the pioneering work of Brockett [10,11], Haynes and Hermes [43], Hermann [44], Hermann and Krener [45], Jurdjevic and Sussmann [50], Lobry [65], and many others, which focuses on state-space representations of nonlinear systems. These are defined by collections of vector fields on manifolds, and are analyzed using, in particular, Lie algebraic techniques. On the other side, the input-output approach is extended to nonlinear settings primarily through a formal power series approach as initiated by Fliess [31]. The interplay between these approaches has been the subject of many successful studies, in which a prominent role is played by Volterra series and the problem of *realizing* such input-output descriptions as a state-space system, see e. g. Brockett [11], Crouch [22], Gray and Wang [37], Jakubczyk [49], Krener and Lesiak [61], and Sontag and Wang [87].

In the late 1970s Agrachëv and Gamkrelidze introduced into nonlinear control theory the aforementioned abstract functional analytic approach, that is rooted in the work of Gelfand. Following traditions from the physics community, they adopted the name chronological calculus. Again, this abstract approach may be seen as much unifying what formerly were disparate and isolated pieces of knowledge and tools in nonlinear systems theory. While originally conceived as a tool for extending Pontryagin's Maximum Principle in optimal control theory [1], the chronological calculus continues to yield a stream of new results in optimal control and geometry, see e. g. Serres [85], Sigalotti [86], Zelenko [105] for very recent results utilizing the chronological calculus for studying the geometry.

The chronological calculus has led to a very different way of thinking about control systems, epitomized in e. g. the monograph on geometric control theory [5] based on the chronological calculus, or in such forceful advocacies for this approach for e. g. nonholonomic path finding by Sussmann [95]. Closely related are studies of nonlinear controllability, e. g. Agrachëv and Gamkrelidze [3,4], Tretyak [96,97,98], and Vakhrameev [99], including applications to controllability of the Navier–Stokes equation by Agrachëv and Sarychev [6]. The chronological calculus also lends itself most naturally to obtaining new results in averaging theory as in Sarychev [83], while Cortes and Martinez [20] used it in motion control of mechanical systems with symmetry and Bullo [13,68] for vibrational control of mechanical systems. Noteworthy applications include locomotion of robots by Burdick and Vela [100], and even robotic fish [71]. Instrumental is its interplay with series expansion as in Bullo [12,21] that utilize affine connections of mechanical systems. There are further applications to stability and stabilization Caiado and Sarychev [15,82], while Monaco et.al. [70] extended this approach to discrete-time dynamics and Komleva and Plotnikov used it in differential game theory [60].

Complementing such *applications* are new subjects of study such as the abstract algebraic structures that underlie the chronological calculus. The chronological algebra itself has been the subject of study as early as Agrachëv and Gamkrelidze [2]. The closely related structure of Zinbiel structures has recently found more attention, see work by Dzhumadil'daev [25,26], Kawski and Sussmann [57] and Kawski [54,55]. Zinbiel algebras arise in the special case when the dynamics ("*time-varying vector fields*") splits into a sum of products of time-varying coefficients and autonomous vector fields. There is an unfortunate confusion of terms in the literature as originally Zinbiel algebras had also been called chronological algebras. Recent usage disentangles these closely related, but distinct, structures and reflects the primacy of the latter term coined by Loday [66,67] who studies Leibniz algebras (which appear in cyclic homology). *Zinbiel* is simply *Leibniz* spelled backwards, a choice which reflect that Leibniz and Zinbiel are dual operands in the sense of Koszul duality as investigated by Ginzburg and Kapranov [36].

Fundamental Notions of the Chronological Calculus

From a Manifold to a Commutative Algebra

This and the next sections very closely follow the introductory exposition of Chapter 2 of Agrachëv and Sachkov [5], which is also recommended for the full technical and analytical details regarding the topology and convergence.

The objective is to develop the basic tools and formalism that facilitate the analysis of generally nonlinear systems that are defined on smooth manifolds. Rather than primarily considering points on a smooth manifold M, the key idea is to instead focus on the commutative algebra of $\mathcal{E}(M) = C^\infty(M, \mathbb{R})$ of real-valued smooth functions on M. Note that $\mathcal{E}(M)$ not only has the structure of a vector space over the field $\mathbb{R}$, but it also inherits the structure of a commutative ring under pointwise addition and multiplication from the codomain $\mathbb{R}$.

Every point $p \in M$ gives rise to a functional $\hat{p}\colon \mathcal{E}(M) \mapsto \mathbb{R}$ defined by $\hat{p}(\varphi) = \varphi(p)$. This functional is linear and multiplicative, and is a homomorphism of the algebras $\mathcal{E}(M)$ and $\mathbb{R}$: For every $p \in M$, for $\varphi, \psi \in \mathcal{E}(M)$ and $t \in \mathbb{R}$ the following hold

$$\hat{p}(\varphi + \psi) = \hat{p}\varphi + \hat{p}\varphi, \quad \hat{p}(\varphi \cdot \psi) = (\hat{p}\varphi) \cdot (\hat{p}\psi)\,,$$
$$\text{and} \quad \hat{p}(t\varphi) = t \cdot (\hat{p}\varphi)\,.$$

Much of the power of this approach derives from the fact that this correspondence is invertible: For a nontrivial multiplicative linear functional $\theta\colon \mathcal{E}(M) \mapsto \mathbb{R}$ consider its kernel $\ker\theta = \{\varphi \in \mathcal{E}(M)\colon \theta\varphi = 0\}$. A critical observation is that this is a maximal ideal, and that it must be of the form $\{\varphi \in \mathcal{E}(M)\colon \varphi(p) = 0\}$ for some, uniquely defined, $p \in M$. For the details of a proof see appendix A.1. of [5].

Proposition 1 *For every nontrivial multiplicative linear functional $\theta\colon \mathcal{E}(M) \mapsto \mathbb{R}$ there exists $p \in M$ such that $\theta = \hat{p}$.*

Note on the side, that there may be maximal ideals in the space of all multiplicative linear functionals on $\mathcal{E}(M)$ that do not correspond to any point on M – e. g. the ideal of all linear functionals that vanish on every function with compact support. But this does not contradict the stated proposition.

Not only can one *recover* the manifold M as a set from the commutative ring $\mathcal{E}(M)$, but using the weak topology on the space of linear functionals on $\mathcal{E}(M)$ one also recovers the topology on M

$$p_n \longrightarrow p \text{ if and only if } \forall f \in \mathcal{E}(M),\ \hat{p_n} f \longrightarrow \hat{p} f\,. \quad (1)$$

The smooth structure on M is recovered from $\mathcal{E}(M)$ in a trivial way: A function g on the space of multiplicative linear functionals $\hat{p}\colon p \in M$ is smooth if and only if there exists $f \in \mathcal{E}(M)$ such that for every $\hat{p}$, $g(\hat{p}) = \hat{p} f$.

In modern differential geometry it is routine to identify tangent vectors to a smooth manifold with either equivalence classes of smooth curves, or with first order partial differential operators. In this context, tangent vectors at a point $q \in M$ are derivations of $\mathcal{E}(M)$, that is, linear functionals $\hat{f}_q$ on $\mathcal{E}(M)$ that satisfy the Leibniz rule. Using $\hat{q}$, this means for every $\varphi, \psi \in \mathcal{E}(M)$,

$$\hat{X}_q(\varphi\psi) = (\hat{X}_q\varphi)(\hat{q}\psi) + (\hat{q}\varphi)(\hat{X}_q\psi)\,. \quad (2)$$

Smooth vector fields on M correspond to linear functionals $\hat{f}\colon \mathcal{E}(M) \mapsto \mathcal{E}(M)$ that satisfy for all $\varphi, \psi \in \mathcal{E}(M)$

$$\hat{f}(\varphi\psi) = (\hat{f}\varphi) \cdot \psi + \varphi \cdot (\hat{f}\psi)\,. \quad (3)$$

Again, the correspondence between tangent vectors and vector fields and the linear functionals as above is invertible. Write $\Gamma^\infty(M)$ for the space of smooth vector fields on M.

Finally, there is a one-to-one correspondence between smooth diffeomorphisms $\Phi\colon M \mapsto M$ and automorphisms of $\mathcal{E}(M)$. The map $\hat{\Phi}\colon \mathcal{E}(M) \mapsto \mathcal{E}(M)$ defined for $p \in M$ and $\varphi \in \mathcal{E}(M)$ by $\hat{\Phi}(\varphi)(p) = \varphi(\Phi(p))$ clearly has the desired properties. For the reverse direction, suppose $\Psi\colon \mathcal{E}(M) \mapsto \mathcal{E}(M)$ is an automorphism. Then for every $p \in M$ the map $\hat{p} \circ \Psi\colon \mathcal{E}(M) \mapsto \mathbb{R}$ is a nontrivial linear multiplicative functional, and hence equals $\hat{q}$ for some $q \in M$. It is easy to see that the map $\Phi\colon M \mapsto M$ is indeed a diffeomorphism, and $\Psi = \hat{\Phi}$.

In the sequel we shall omit the *hats*, and simply write, say, p for the linear functional $\hat{p}$. The role of each object is usually clear from the order. For example, for a point p, a smooth function φ, smooth vector fields f, g, h, with flow e^{th} and its tangent map $\left(e^{th}\right)_*$, what in traditional format might be expressed as $((e^{th})_* g)\varphi(e^{\tau f}(p))$ is simply written as $pe^{\tau f} g e^{th}\varphi$, not requiring any parentheses.

Frechet Space and Convergence

The usual topology on the space $\mathcal{E}(M)$ is the one of uniform convergence of all derivatives on compact sets, i. e., a sequence of functions $\{\varphi_k\}_{k=1}^\infty \subseteq \mathcal{E}(M)$ converges to $\phi \in \mathcal{E}(M)$ if for every finite sequence $f_1, f_2, \ldots, f_s$ of smooth vector fields on M and every compact set $K \subseteq M$ the sequence $\{f_s \ldots f_2 f_1 \varphi_k\}_{k=1}^\infty$ converges uniformly on K to $f_s \ldots f_2 f_1 \varphi$.

This topology is also obtained by a countable family of semi-norms $\|\cdot\|_{s,K}$ defined by

$$\|\varphi\|_{s,K} = \sup\{\, |p f_s \ldots f_2 f_1 \varphi| : \quad p \in K,\ f_i \in \Gamma^\infty(M),\ s \in \mathbb{Z}^+ \,\} \quad (4)$$

where K ranges over a countable collection of compact subsets whose union is all of M. In an analogous way define semi-norms of smooth vector fields $f \in \Gamma^\infty(M)$

$$\|f\|_{s,K} = \sup\{\ \|f\varphi\|_{s,K}\ :\ \|f\varphi\|_{s+1,K} = 1\ \}\,. \tag{5}$$

Finally, for every smooth diffeomorphism Φ of M, $s \in \mathbb{Z}^+$ and $K \subseteq M$ compact there exist $C_{s,K,\Phi} \in \mathbb{R}$ such that for all $\varphi \in \mathcal{E}(M)$

$$\|\Phi\varphi\|_{s,K}\ \le\ C_{s,K,\Phi}\,\|\varphi\|_{s,\varphi(K)}\,. \tag{6}$$

Regularity properties of one-parameter families of vector fields and diffeomorphisms (*"time-varying vector fields and diffeomorphisms"*) are understood in the weak sense. In particular, for a family of smooth vector fields $f_t \in \Gamma^\infty(M)$, its integral and derivative (if they exist) are defined as the operators that satisfy for every $\varphi \in \mathcal{E}(M)$

$$\left(\frac{\mathrm{d}}{\mathrm{d}t} f_t\right)\varphi = \frac{\mathrm{d}}{\mathrm{d}t}\left(f_t\varphi\right) \quad \text{and} \quad \left(\int_a^b f_t \mathrm{d}t\right)\varphi = \int_a^b \left(f_t\varphi\right)\mathrm{d}t\,. \tag{7}$$

The convergence of series expansions of vector fields and of diffeomorphisms encountered in the sequel are to be interpreted in the analogous weak sense.

The Chronological Exponential

This section continues to closely follow [5] which contains full details and complete proofs. On a manifold M consider generally time-varying differential equations of the form $\frac{\mathrm{d}}{\mathrm{d}t} q_t = f_t(q_t)$. To assure existence and uniqueness of solutions to initial value problems, make the typical regularity assumptions, namely that in every coordinate chart $U \subseteq M$, $x\colon U \mapsto \mathbb{R}^n$ the vector field $(x_* f_t)$ is (i) measurable and locally bounded with respect to t for every fixed x and (ii) smooth with locally bounded partial derivatives with respect to x for every fixed t. For the purposes of this article and for clarity of exposition, also assume that vector fields are complete. This means that solutions to initial value problems are defined for all times $t \in \mathbb{R}$. This is guaranteed if, for example, all vector fields considered vanish identically outside a common compact subset of M.

As in the previous sections, for each fixed t interpret $q_t\colon \mathcal{E}(M) \mapsto \mathbb{R}$ as a linear functional on $\mathcal{E}(M)$, and note that this family satisfies the time-varying, but linear differential equation, suggestively written as

$$\dot{q}_t = q_t \circ f_t\,. \tag{8}$$

on the space of linear functionals on $\mathcal{E}(M)$. It may be shown that under the above assumptions it has a unique solution, called the *right chronological exponential* of the vector field f_t as the corresponding flow. Formally, it satisfies for almost all $t \in \mathbb{R}$

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\overrightarrow{\exp}\int_0^t f_\tau\,\mathrm{d}\tau\right) = \left(\overrightarrow{\exp}\int_0^t f_\tau\,\mathrm{d}\tau\right)\circ f_t\,. \tag{9}$$

Analogously the *left chronological exponential* satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\overleftarrow{\exp}\int_0^t f_\tau\,\mathrm{d}\tau\right) = f_t \circ \left(\overleftarrow{\exp}\int_0^t f_\tau\,\mathrm{d}\tau\right)\,. \tag{10}$$

Formally, one obtains a series expansion for the chronological exponentials by rewriting the differential equation as an integral equation and solving it by iteration

$$q_t = q_0 + \int_0^t q(\tau)\circ f_\tau\,\mathrm{d}\tau \tag{11}$$

$$= q_0 + \int_0^t \left(q_0 + \int_0^\tau q(\sigma)\circ f_\sigma\,d\sigma\right)\circ f_\tau\,\mathrm{d}\tau \tag{12}$$

to eventually, formally, obtain the expansion

$$\overrightarrow{\exp}\int_0^t f_\tau\,\mathrm{d}\tau \sim \mathrm{Id} + \sum_{k=1}^\infty \int_0^t\int_0^{t_k}\cdots\int_0^{t_2} \cdot f_{\tau_k}\circ\cdots\circ f_{\tau_2}\circ f_{\tau_1}\,\mathrm{d}\tau_k\ldots\mathrm{d}\tau_2\,\mathrm{d}\tau_1 \tag{13}$$

and analogously for the left chronological exponential

$$\overleftarrow{\exp}\int_0^t f_\tau\,\mathrm{d}\tau \sim \mathrm{Id} + \sum_{k=1}^\infty \int_0^t\int_0^{t_k}\cdots\int_0^{t_2} \cdot f_{\tau_1}\circ f_{\tau_2}\circ\cdots\circ f_{\tau_k}\,\mathrm{d}\tau_k\ldots\mathrm{d}\tau_2\,\mathrm{d}\tau_1\,. \tag{14}$$

While this series never converges, not even in a weak sense, it nonetheless has an interpretation as an asymptotic expansion. In particular, for any fixed function $\phi \in \mathcal{E}(M)$ and any semi-norm as in the previous section, on any compact set one obtains an error estimate for the remainder after truncating the series at any order N which establishes that the truncation error is of order $O(t^N)$ as $t \to 0$. When one considers the restrictions to any f_t-invariant normed linear subspace $L \subseteq \mathcal{E}(M)$, i. e. for all t, $f_t(L) \subseteq L$, on which f_t is bounded, then the asymptotic series converges to the chronological exponential and it satisfies for every $\varphi \in L$

$$\left\|\left(\overrightarrow{\exp}\int_0^t f_\tau\,\mathrm{d}\tau\right)\varphi\right\| \le \mathrm{e}^{\int_0^t \|f_\tau\|\,\mathrm{d}\tau}\,\|\varphi\|\,. \tag{15}$$

In the case of analytic vector fields f_τ and analytic functions φ one obtains convergence of the series for sufficiently small t.

While in general there is no reason why the vector fields f_t should commute at different times, the chronological exponentials nonetheless still share some of the usual properties of exponential and flows, for example the composition of chronological exponentials satisfies for all $t_i \in \mathbb{R}$

$$\left(\overrightarrow{\exp}\int_{t_1}^{t_2} f_\tau \,\mathrm{d}\tau\right)\circ\left(\overrightarrow{\exp}\int_{t_2}^{t_3} f_\tau \,\mathrm{d}\tau\right) = \left(\overrightarrow{\exp}\int_{t_1}^{t_3} f_\tau \,\mathrm{d}\tau\right) . \tag{16}$$

Moreover, the left and right chronological exponentials are inverses of each other in the sense that for all $t_0, t_1 \in \mathbb{R}$

$$\begin{aligned}\left(\overrightarrow{\exp}\int_{t_0}^{t_1} f_\tau \,\mathrm{d}\tau\right)^{-1} &= \overrightarrow{\exp}\int_{t_1}^{t_0} f_\tau \,\mathrm{d}\tau \\ &= \overleftarrow{\exp}\int_{t_0}^{t_1} (-f_\tau)\,\mathrm{d}\tau .\end{aligned} \tag{17}$$

Variation of Parameters and the Chronological Logarithm

In control one rarely considers only a single vector field, and, instead, for example, is interested in the interaction of a perturbation or control vector field with a reference or drift vector field. In particular, in the nonlinear setting, this is where the chronological calculus very much facilitates the analysis. For simplicity consider a differential equation defined by two, generally time varying, vector fields (with the usual regularity assumptions)

$$\dot{q} = f_t(q) + g_t(q) . \tag{18}$$

The objective is to obtain a formula for the flow of the field $(f_t + g_t)$ as a *perturbation* of the flow of f_t. Writing

$$\Phi_t = \overrightarrow{\exp}\int_0^t f_\tau \,\mathrm{d}\tau \quad \text{and} \quad \Theta_t = \overrightarrow{\exp}\int_0^t (f_\tau + g_\tau)\,\mathrm{d}\tau, \tag{19}$$

this means one is looking for a family of operators Ψ_t such that for all t

$$\Theta_t = \Psi_t \circ \Phi_t . \tag{20}$$

Differentiating and using the invertibility of the flows one obtains a differential equation in the standard form

$$\begin{aligned}\dot{\Psi}(t) &= \Psi_t \circ \Phi_t \circ g_t \circ \Phi_t^{-1} = \Psi_t \circ (\mathrm{Ad}\Phi_t)\; g_t \\ &= \Psi_t \circ \left(\overrightarrow{\exp}\int_0^t \mathrm{ad} f_\tau \,\mathrm{d}\tau\right)\; g_t ,\end{aligned} \tag{21}$$

which has the unique solution

$$\Psi_t = \overrightarrow{\exp}\int_0^t \left(\overrightarrow{\exp}\int_0^\tau \mathrm{ad} f_\sigma \,\mathrm{d}\sigma\right)\; g_\tau \,\mathrm{d}\tau , \tag{22}$$

and consequently one has the *variations formula*

$$\begin{aligned}\overrightarrow{\exp}\int_0^t (f_\tau + g_\tau)\,\mathrm{d}\tau = \Bigg(\overrightarrow{\exp}\int_0^t \Bigg(\overrightarrow{\exp}\int_0^\tau \mathrm{ad} f_\sigma \,\mathrm{d}\sigma\Bigg) \\ \cdot\, g_\tau \,\mathrm{d}\tau\Bigg) \circ \left(\overrightarrow{\exp}\int_0^t f_\tau \,\mathrm{d}\tau\right) .\end{aligned} \tag{23}$$

This formula is of fundamental importance and used ubiquitously for analyzing controllability and optimality: One typically considers f_t as defining a reference dynamical system and then considers the effect of adding a control term g_t. A special application is in the theory of averaging where g_t is considered a perturbation of the field f_t, which in the most simple form may be assumed to be time-periodic. The monograph [81] by Sanders and Verhulst is the classical reference for the theory and applications of averaging using traditional language. The chronological calculus, in particular the above variations formula, have been used successfully by Sarychev [83] to investigate in particular higher order averaging, and by Bullo [13] for averaging in the context of mechanical systems and its interplay with mechanical connections. Instrumental to the work [83] is the notion of the chronological logarithm, which is motivated by rewriting the defining Eqs. (9) and (10) [2,80] as

$$\begin{aligned}f_t &= \left(\overrightarrow{\exp}\int_0^t f_\tau \,\mathrm{d}\tau\right)^{-1} \circ \frac{\mathrm{d}}{\mathrm{d}t}\left(\overrightarrow{\exp}\int_0^t f_\tau \,\mathrm{d}\tau\right) \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\overleftarrow{\exp}\int_0^t f_\tau \,\mathrm{d}\tau\right) \circ \left(\overleftarrow{\exp}\int_0^t f_\tau \,\mathrm{d}\tau\right)^{-1} .\end{aligned} \tag{24}$$

While in general the existence of the logarithm is a delicate question [62,83], under suitable hypotheses it is instrumental for the derivation of the nonlinear Floquet theorem in [83].

Theorem 1 (Sarychev [83]) *Suppose $f_t = f_{t+1}$ is a period-one smooth vector field generating the flow Φ_t and $\Lambda = \log \Phi_1$, then there exists a period-one family of diffeomorphisms $\Psi_t = \Psi_{t+1}$ such that for all t $\Phi_t = e^{t\Lambda} \circ \Psi_t$.*

The logarithm of the chronological exponential of a time-varying vector field εf_t may be expanded as a formal chronological series

$$\log \overrightarrow{\exp}\left(\int_0^1 \varepsilon f_\tau \,\mathrm{d}\tau\right) = \sum_{j=1}^{\infty} \varepsilon^j \Lambda^j . \tag{25}$$

The first three terms of which are given explicitly [83] as

$$\Lambda^1 = \int_0^1 f_\tau \, \mathrm{d}\tau \,, \quad \Lambda^2 = \frac{1}{2}\int_0^1 \left[\int_0^\tau f_\sigma \, \mathrm{d}\sigma \,,\, f_\tau\right] \mathrm{d}\tau,$$
$$\text{and} \quad \Lambda^3 = -\frac{1}{2}[\Lambda^1, \Lambda^2] + \frac{1}{3}\int_0^1 \mathrm{ad}^2\left(\int_0^\tau f_\sigma \, \mathrm{d}\sigma\right) f_\tau \, \mathrm{d}\tau \,. \quad (26)$$

Aside from such applications as averaging, the main use of the chronological calculus has been for studying the dual problems of optimality and controllability, in particular controllability of families of diffeomorphisms [4]. In particular, this approach circumvents the limitations encountered by more traditional approaches that use parametrized families of control variations that are based on a finite number of switching times. However, since [52] it is known that a general theory must allow also, e. g., for increasing numbers of switchings and other more general families of control variations. The chronological calculus is instrumental to developing such general theory [4].

Chronological Algebra

When using the chronological calculus for studying controllability, especially when differentiating with respect to a parameter, and also in the aforementioned averaging theory, the following *chronological product* appears almost everywhere: For two time-varying vector fields f_t and g_t that are absolutely continuous with respect to t, define their chronological product as

$$(f \star g)_t = \int_0^t [f_s, g_s'] \mathrm{d}s \quad (27)$$

where $[\cdot,\cdot]$ denotes the Lie bracket. It is easily verified that this product is generally neither associative nor satisfies the Jacobi identity, but instead it satisfies the chronological identity

$$x \star (y \star z) - y \star (x \star z) = (x \star y) \star z - (y \star x) \star z \,. \quad (28)$$

One may define an abstract chronological algebra (over a field $\mathbf{k}$) as a linear space that is equipped with a bilinear product that satisfies the chronological identity (28). More compactly, this property may be written in terms of the left translation λ that associates with any element $x \in \mathcal{A}$ of a not-necessarily associative algebra $\mathcal{A}$ the map $\lambda_x : y \mapsto xy$. Using this λ, the chronological identity (28) is simply

$$\lambda_{[x,y]} = [\lambda_x, \lambda_y] \,, \quad (29)$$

i. e., the requirement that the map $x \mapsto \lambda_x$ is a homomorphism of the algebra $\mathcal{A}$ with the commutator product $(x, y) \mapsto xy - yx$ into the Lie algebra of linear maps from $\mathcal{A}$ into $\mathcal{A}$ under its commutator product.

Basic general properties of chronological algebras have been studied in [2], including a discussion of free chronological algebras (over a fixed generating set S). Of particular interest are bases of such free chronological algebras as they allow one to minimize the number of terms in series expansions by avoiding redundant terms. Using the graded structure, it is shown in [2] that an ordered basis of a free chronological algebra over a set S may be obtained recursively from products of the form $\lambda_{b_1}\lambda_{b_2}\cdots\lambda_{b_k}s$ where $s \in S$ is a generator and $b_1 \preccurlyeq b_2 \preccurlyeq \cdots \preccurlyeq b_k$ are previously constructed basis elements. According to [2], the first elements of such a basis over the single-element generating set $S = \{s\}$ may be chosen as (using juxtaposition for multiplication)

$$b_1 = s \,, \quad b_2 = \lambda_{b_1}s = s^2 \,, \quad b_3 = \lambda_{b_2}s = s^2 s \,,$$
$$b_4 = \lambda_{b_1}\lambda_{b_1}s = ss^2 \,, \quad b_5 = \lambda_{b_3}s = (s^2 s)s \,,$$
$$b_6 = \lambda_{b_4}s = (ss^2)s \,, \quad b_7 = \lambda_{b_1}\lambda_{b_2}s = s(s^2 s) \,,$$
$$b_8 = \lambda_{b_2}\lambda_{b_1}s = s(ss^2) \,, \ldots \quad (30)$$

Using only this algebraic and combinatorial structure, one may define exponentials (and logarithms) and analyze the group of formal flows [2]. Of particular interest systems and control is an explicit formula that allows one to express products (compositions) of two or more exponentials (flows) as a single exponential. This is tantamount to a formula for the logarithm of a product of exponentials of two noncommuting indeterminates X and Y. The classical version is known as the Campbell–Baker–Hausdorff formula [16], the first few terms being

$$\mathrm{e}^X \cdot \mathrm{e}^Y = \mathrm{e}^{X+Y+\frac{1}{2}[X,Y]+\frac{1}{12}[X-Y,[X,Y]]-\frac{1}{24}[X,[Y,[X,Y]]]-\ldots} \,. \quad (31)$$

A well-known short-coming of the classical formula is that the higher-order iterated Lie brackets appearing in this formula are always linearly dependent and hence the coefficients are never well-defined. Hence one naturally looks for more compact, minimal formulas which avoid such redundancies. The chronological algebra provides one such alternative and has been developed and utilized in [2]. The interested reader is referred to the original reference for technical details.

Systems That Are Affine in the Control

Separating Time-Varying Control and Geometry

The chronological calculus was originally designed for work with nonstationary (time-varying) families of vec-

tor fields, and it arguably reaps the biggest benefits in that setting where there are few other powerful tools available. Nonetheless, the chronological calculus also much streamlines the analysis of autonomous vector fields. The best studied case is that of affine control systems in which the time-varying control can be separated from the geometry determined by autonomous vector fields. In classical notation such control systems are written in the form

$$\dot{x} = u_1(t)f_1(x) + \cdots + u_m(t)f_m(x)\,. \tag{32}$$

where f_i are vector fields on a manifold M and u is a measurable function of time taking values typically in a compact subset $U \subseteq \mathbb{R}^m$. This description allows one to also accommodate physical systems that have an *uncontrolled drift term* by simply fixing $u_1 \equiv 1$. For the sake of clarity, and for best illustrating the kind of results available, assume that the fields f_i are smooth and complete. Later we specialize further to real analytic vector fields. This set-up does not necessarily require an interpretation of the u_i as controls: they may equally well be disturbances, or it may simply be a dynamical system which splits in this convenient way.

As a starting point consider families of piecewise constant control functions $u\colon [0,T] \mapsto U \subseteq \mathbb{R}^m$. On each interval $[t_j, t_{j+1}]$ on which the control is constant $u_{[t_j,t_{j+1}]} = u^{(j)}$, the right hand side of (32) is a fixed vector field $g_j = u_1^{(j)} f_1 + \ldots u_m^{(j)} f_m$. The endpoint of the solution curve starting from $x(0) = p$ is computed as a directed product of ordinary exponentials (flows) of vector fields

$$p\mathrm{e}^{(t_1-t_0)g_1} \circ \mathrm{e}^{(t_2-t_1)g_2} \circ \cdots \mathrm{e}^{(T-t_{s-1})g_s}\,. \tag{33}$$

But even in this case of autonomous vector fields the chronological calculus brings substantial simplifications. Basically this means that rather than considering the expression (33) as a point on M, consider this product as a functional on the space $\mathcal{E}(M)$ of smooth functions on M. To reiterate the benefit of this approach [5,57] consider the simple example of a tangent vector to a smooth curve $\gamma\colon (-\varepsilon,\varepsilon) \mapsto M$ which one might want to define as

$$\dot{\gamma}(0) = \lim_{t\to 0} \frac{1}{t}(\gamma(t) - \gamma(0))\,. \tag{34}$$

Due to the lack of a linear structure on a general manifold, with a classical interpretation this does not make any sense at all. Nonetheless, when interpreting $\gamma(t)$, $\gamma(0)$, $\gamma'(0)$, and $\frac{1}{t}(\gamma(t)-\gamma(0))$ as linear functionals on the space $\mathcal{E}(M)$ of smooth functions on M this is perfectly meaningful. The meaning of the limit can be rigorously justified as indicated in the preceding sections, compare also [1].

A typical example to illustrate how this formalism almost trivializes important calculations involving Lie brackets is the following, compare [57]. Suppose f and g are smooth vector fields on the manifold M, $p \in M$, and $t \in R$ is sufficiently small in magnitude. Using that $\frac{\mathrm{d}}{\mathrm{d}t} p\mathrm{e}^{tf} = p\mathrm{e}^{tf} f$ and so on, one immediately calculates

$$\begin{aligned}&(\mathrm{d}/\mathrm{d}t)(p\mathrm{e}^{tf}\mathrm{e}^{tg}\mathrm{e}^{-tf}\mathrm{e}^{-tg})\\ &\quad = p\mathrm{e}^{tf} f\mathrm{e}^{tg}\mathrm{e}^{-tf}\mathrm{e}^{-tg} + p\mathrm{e}^{tf}\mathrm{e}^{tg} g\mathrm{e}^{-tf}\mathrm{e}^{-tg}\\ &\quad - p\mathrm{e}^{tf}\mathrm{e}^{tg}\mathrm{e}^{-tf} f\mathrm{e}^{-tg} - p\mathrm{e}^{tf}\mathrm{e}^{tg}\mathrm{e}^{-tf}\mathrm{e}^{-tg} g\,.\end{aligned} \tag{35}$$

In particular, at $t = 0$ this expression simplifies to $pf + pg - pf - pg = 0$. Analogously the second derivative is calculated as

$$\begin{aligned}&\frac{\mathrm{d}^2}{\mathrm{d}t^2}(p\mathrm{e}^{tf}\mathrm{e}^{tg}\mathrm{e}^{-tf}\mathrm{e}^{-tg}) = p\mathrm{e}^{tf} f^2\mathrm{e}^{tg}\mathrm{e}^{-tf}\mathrm{e}^{-tg} + p\mathrm{e}^{tf} f\\ &\cdot \mathrm{e}^{tg} g\mathrm{e}^{-tf}\mathrm{e}^{-tg} - p\mathrm{e}^{tf} f\mathrm{e}^{tg}\mathrm{e}^{-tf} f\mathrm{e}^{-tg} - p\mathrm{e}^{tf} f\mathrm{e}^{tg}\mathrm{e}^{-tf}\\ &\cdot \mathrm{e}^{-tg} g + p\mathrm{e}^{tf} f\mathrm{e}^{tg} g\mathrm{e}^{-tf}\mathrm{e}^{-tg} + p\mathrm{e}^{tf}\mathrm{e}^{tg} g^2\mathrm{e}^{-tf}\mathrm{e}^{-tg}\\ &- p\mathrm{e}^{tf}\mathrm{e}^{tg} g\mathrm{e}^{-tf} f\mathrm{e}^{-tg} - p\mathrm{e}^{tf}\mathrm{e}^{tg} g\mathrm{e}^{-tf}\mathrm{e}^{-tg} g - p\mathrm{e}^{tf} f\\ &\cdot \mathrm{e}^{tg}\mathrm{e}^{-tf} f\mathrm{e}^{-tg} - p\mathrm{e}^{tf}\mathrm{e}^{tg} g\mathrm{e}^{-tf} f\mathrm{e}^{-tg} + p\mathrm{e}^{tf}\mathrm{e}^{tg}\mathrm{e}^{-tf}\\ &\cdot f^2\mathrm{e}^{-tg} + p\mathrm{e}^{tf}\mathrm{e}^{tg}\mathrm{e}^{-tf} f\mathrm{e}^{-tg} g - p\mathrm{e}^{tf} f\mathrm{e}^{tg}\mathrm{e}^{-tf}\mathrm{e}^{-tg} g\\ &- p\mathrm{e}^{tf}\mathrm{e}^{tg} g\mathrm{e}^{-tf}\mathrm{e}^{-tg} g + p\mathrm{e}^{tf}\mathrm{e}^{tg}\mathrm{e}^{-tf} f\mathrm{e}^{-tg} g + p\mathrm{e}^{tf}\\ &\cdot \mathrm{e}^{tg}\mathrm{e}^{-tf}\mathrm{e}^{-tg} g^2\end{aligned} \tag{36}$$

which at $t = 0$ evaluates to

$$\begin{aligned}&pf^2 + pfg - pf^2 - pfg + pfg + pg^2 - pgf - pg^2\\ &- pf^2 - pgf + pf^2 + pfg - pfg - pg^2 + pfg + pg^2\\ &= 2pfg - 2pgf = 2p[f,g]\,.\end{aligned}$$

While this simple calculation may not make sense on a manifold with a classical interpretation in terms of points and vectors on a manifold, it does so in the context of the chronological calculus, and establishes the familiar formula for the Lie bracket as the limit of a composition of flows

$$p\mathrm{e}^{tf}\mathrm{e}^{tg}\mathrm{e}^{-tf}\mathrm{e}^{-tg} = p + t^2 p[f,g] + O(t^2) \;\text{ as }\; t \to 0. \tag{37}$$

Similar examples may be found in [5] (e. g. p. 36), as well as analogous simplifications for the variations formula for autonomous vector fields ([5] p. 43).

Asymptotic Expansions for Affine Systems

Piecewise constant controls are a starting point. But a general theory requires being able to take appropriate

limits and demands completeness. Typically, one considers measurable controls taking values in a compact subspace $U \subseteq \mathbb{R}^m$, and uses $L^1([0,t],U)$ as the space of admissible controls. Correspondingly the finite composition of exponentials as in (33) is replaced by continuous formulae. These shall still separate the effects of the time-varying controls from the underlying geometry determined by the autonomous vector fields. This objective is achieved by the Chen–Fliess series which may be interpreted in a number of different ways. Its origins go back to the 1950s when K. T. Chen [18], studying geometric invariants of curves in $\mathbb{R}^n$, associated to each smooth curve a formal noncommutative power series. In the early 1970s, Fliess [30,31] recognized the utility of this series for studying control systems. Using a set $X_1, X_2, \ldots, X_m$ of indeterminates, the Chen–Fliess series of a measurable control $u \in L^1([0,t],U)$ as above is the formal power series

$$S_{\mathrm{CF}}(T,u) = \sum_I \int_0^T u_{i_s}(t_s) \int_0^{t_{s-1}} \cdots \int_0^{t_3} u_{i_2}(t_2) \cdot \int_0^{t_2} u_{i_1}(t_1)\, \mathrm{d}t_1 \ldots \mathrm{d}t_s \cdot X_{i_1} \ldots X_{i_s} \quad (38)$$

where the sum ranges over all multi-indices $I = (i_1, \ldots, i_s)$, $s \geq 0$ with $i_j \in \{1, 2 \ldots m\}$. This series gives rise to an asymptotic series for solution curves of the system (32). For example, in the analytic setting, following [91], one has:

Theorem 2 *Suppose f_i are analytic vector fields on $\mathbb{R}^n$ $\phi\colon \mathbb{R}^n \mapsto \mathbb{R}$ is analytic and $U \subset \mathbb{R}^m$ is compact. Then for every compact set $K \subseteq \mathbb{R}^n$, there exists $T > 0$ such that the series*

$$S_{\mathrm{CF},f}(T,u,p)(\varphi) = \sum_I \int_0^T u_{i_s}(t_s) \int_0^{t_{s-1}} \cdots \int_0^{t_3} u_{i_2}(t_2) \cdot \int_0^{t_2} u_{i_1}(t_1)\, dt_1 \ldots dt_s \cdot p f_{i_1} \ldots f_{i_s} \varphi \quad (39)$$

converges uniformly to the solution of (32) for initial conditions $x(0) = p \in K$ and $u\colon [0,T] \mapsto U$ measurable.

Here the series $S_{\mathrm{CF},f}(T,u,p)$ is, for every fixed triple (f,u,p), interpreted as an element of the space of linear functionals on $\mathcal{E}(M)$. But one may abstract further. In particular, for each fixed multi-index I as above, the iterated integral coefficient is itself a functional that takes as input a control function $u \in \mathcal{U}^m$ and maps it to the corresponding iterated integral. It is convenient to work with the primitives $U_j\colon t \mapsto \int_0^t u_j(s)\, ds$ of the control functions u_j rather than the controls themselves. More specifically, if e. g. $\mathcal{U} = AC([0,T],[-1,1])$ then for every multi-index $I = (i_1, \ldots, i_s) \in \{1, \ldots, m\}^s$ as above one obtains the iterated integral functional $\Upsilon_I\colon \mathcal{U}^m \mapsto \mathcal{U}$ defined by

$$\Upsilon_I\colon U \mapsto \int_0^T U'_{i_s}(t_s) \int_0^{t_{s-1}} \cdots \int_0^{t_3} U'_{i_2}(t_2) \cdot \int_0^{t_2} U'_{i_1}(t_1)\, \mathrm{d}t_1 \mathrm{d}t_2 \ldots \mathrm{d}t_s\,. \quad (40)$$

Denoting by $\mathcal{II}\mathcal{F}(\mathcal{U}^Z)$ the linear space of iterated integral functionals spanned by the functionals of the above form, the map Υ maps multi-indices to $\mathcal{II}\mathcal{F}(\mathcal{U}^Z)$. The algebraic and combinatorial nature of these spaces and this map will be explored in the next section.

On the other side, there is the map $\mathcal{F}$ that maps a multi-index $(i_1, i_2, \ldots, i_s)$ to the partial differential operator $f_{i_1} \ldots f_{i_s}\colon \mathcal{E}(M) \mapsto \mathcal{E}(M)$, considered as a linear transformation of $\mathcal{E}(M)$. In the analytic case one may go further, since as a basic principle, the relations between the iterated Lie brackets of the vector fields f_j completely determine the geometry and dynamical behavior [90,91]. Instead of considering actions on $\mathcal{E}(M)$, it suffices to consider the action of these partial differential operators, and of the dynamical system (32) on a set of polynomial functions. In a certain technical sense, an associative algebra (of polynomial functions) is dual to the Lie algebra (generated by the analytic vector fields f_j). However, much further simplifications and deeper insights are gained by appropriately accounting for the intrinsic noncommutative structure of the flows of the system. Using a different product structure, the next section will *lift* the system (32) to a *universal* system that has no nontrivial relations between the iterated Lie brackets of the formal vector fields and which is modeled on an algebra of noncommutative polynomials (and noncommutative power series) rather than the space $\mathcal{E}(M)$ of all smooth functions on M. An important feature of this approach is that it does not rely on polynomial functions with respect to an a-priori chosen set of local coordinates, but rather uses polynomials which map to intrinsically geometric objects.

Zinbiel Algebra and Combinatorics

Just as the chronological calculus of time-varying vector fields is intimately connected to chronological algebras, the calculus of affine systems gives rise to its own algebraic and combinatorial structure. This, and the subsequent section, give a brief look into this algebra structure and demonstrate how, together with the chronological calculus, it leads to effective solution formulas for important control problems and beyond.

It is traditional and convenient to slightly change the notation and nomenclature. For further details see

[78] or [57]. Rather than using small integers $1, 2, \ldots, m$ to index the components of the control and the vector fields in (32), use an arbitrary index set Z whose elements will be called *letters* and denoted usually by $a, b, c, \ldots$ For the purposes of this note the *alphabet* Z will be assumed to be finite, but much of the theory can be developed for infinite alphabets as well. Multi-indices, i. e. finite sequences with values in Z are called *words*, and they have a natural associative product structure denoted by juxtaposition. For example, if $w = a_1 a_2 \ldots a_r$ and $z = b_1 b_2 \ldots b_s$ then $wz = a_1 a_2 \ldots a_r b_1 b_2 \ldots b_s$. The empty word is denoted e or 1. The set of all words of length n is Z^n, $Z^* = \cup_{n=0}^{\infty} Z^n$ and $Z^+ = \cup_{n=1}^{\infty} Z^n$ are the sets of all words and all nonempty words. Endow the associative algebra $A(Z)$, generated by the alphabet Z with coefficients in the field $\mathbf{k} = \mathbb{R}$, with an inner product $\langle \cdot, \cdot \rangle$ so that Z^* is an orthonormal basis. Write $\hat{A}(Z)$ for the completion of $A(Z)$ with respect to the uniform structure in which a fundamental system of basic neighborhoods of a polynomial $p \in A(Z)$ is the family of subsets $\{q \in A(Z) \colon \forall w \in W,\ \langle w, p - q \rangle = 0\}$ indexed by all finite subsets $W \subset Z^*$.

The smallest linear subspace of the associative algebra $A(Z)$ that contains the set Z and is closed under the commutator product $(w, z) \mapsto [w, z] = wz - zw$ is isomorphic to the free Lie algebra $L(Z)$ over the set Z. On the other side, one may equip $A^+(Z)$ with a Zinbiel algebra structure by defining the product $*\colon A^+(Z) \times A^+(Z \mapsto A^+(Z)$ for $a \in Z$ and $w, z \in Z^+$ by

$$w * a = wa \quad \text{and} \quad w * (za) = (w * z + z * w)a \tag{41}$$

and extending bilinearly to $A^+(Z) \times A^+(Z)$. One easily verifies that this product satisfies the Zinbiel identity for all $r, s, t \in A^+(Z)$

$$r * (s * t) = (r * s) * t + (s * r) * t\,. \tag{42}$$

The symmetrization of this product is the better known associative *shuffle product* $w \sqcup\!\sqcup z = w * z + z * w$ which may be defined on all of $A(Z) \times A(Z)$. Algebraically, the shuffle product is characterized as the transpose of the coproduct $\Delta\colon A(Z) \mapsto A(Z) \otimes A(Z)$ which is the *concatenation* product homomorphism that on letters $a \in Z$ is defined as

$$\Delta\colon a \mapsto a \otimes 1 + 1 \otimes a\,. \tag{43}$$

This relation between this coproduct and the shuffle is that for all $u, v, w \in A(Z)$

$$\langle \Delta(w), u \otimes v \rangle = \langle w, u \sqcup\!\sqcup v \rangle\,. \tag{44}$$

After this brief side-trip into algebraic and combinatorial objects we return to the control systems. The shuffle product $\sqcup\!\sqcup$ has been utilized for a long time in this and related contexts. But the Zinbiel product structure was only recently recognized as encoding important information. Its most immediate role is seen in the fact that the aforementioned map Υ from multi-indices to iterated integral functionals, now easily extends to a map (using the same name) $\Upsilon\colon A(Z) \mapsto \mathcal{II}\mathcal{F}(\mathcal{U}^Z)$ is a Zinbiel algebra homomorphism when the latter is equipped with a *pointwise* product induced by the product $*$ on $\mathcal{U} = AC([0, T], [-1, 1])$ defined by

$$(U * V)(t) = \int_0^t U(s)V'(s)\mathrm{d}s\,. \tag{45}$$

It is straightforward to verify that this product indeed equips $\mathcal{U}$ with a Zinbiel structure, and hence also $\mathcal{II}\mathcal{F}(\mathcal{U}^Z)$. On the side, note that the map Υ maps the shuffle product of words (noncommuting polynomials) to the associative pointwise product of scalar functions. The connection of the Zinbiel product with differential equations and the Chen–Fliess series is immediate. First *lift* the system (32) from the manifold M to the algebra $A(Z)$, roughly corresponding to the linear space of polynomial functions in an algebra of iterated integral functionals. Formally consider curves $S\colon [0, T] \mapsto A(Z)$ that satisfy

$$\dot{S} = S \sum_{a \in Z} a u_a(t)\,. \tag{46}$$

Using the Zinbiel product (45), and writing $U_a(t) = \int_0^t u_a(\tau)\mathrm{d}\tau$ for the primitives of the controls, the integrated form of the *universal control system* (46)

$$S(t) = 1 + \int_0^t S(\tau)F'(\tau)\mathrm{d}\tau \quad \text{with} \quad F = \sum_{a \in Z} U_a\,, \tag{47}$$

is most compactly written as

$$S = 1 + S * F\,. \tag{48}$$

Iteration yields the explicit series expansion

$$\begin{aligned}
S &= 1 + (1 + S * F) * F \\
&= 1 + F + ((1 + S * F) * F) * F \\
&= 1 + F + (F * F) + (((1 + S * F) * F) * F) * F \\
&= 1 + F + (F * F) + ((F * F) * F) \\
&\qquad + ((((1 + S * F) * F) * F) * F) * F \\
&\ \ \vdots \\
&= 1 + F + (F * F) + ((F * F) * F) \\
&\qquad + (((F * F) * F) * F) \ldots
\end{aligned}$$

Using intuitive notation for Zinbiel powers this solution formula in the form of an infinite series is compactly written as

$$S = \sum_{n=0}^{\infty} F^{*n} = 1 + F + F^{*2} + F^{*3} + F^{*4} + F^{*5} + F^{*6} + \cdots \tag{49}$$

The reader is encouraged to expand all of the terms and match each step to the usual computations involved in obtaining Volterra series expansions. Indeed this formula (49) is nothing more than the Chen–Fliess series (38) in very compact notation. For complete technical details and further discussion we refer the interested reader to the original articles [56,57] and the many references therein.

Application: Product Expansions and Normal Forms

While the Chen–Fliess series as above has been extremely useful to obtain precise estimates that led to most known necessary conditions and sufficient conditions for local controllability and optimality, compare e. g. [53,88,94], this series expansion has substantial shortcomings. For example, it is clear from Ree's theorem [77] that the series is an exponential Lie series. But this is not at all obvious from the series as presented (38). Most detrimental for practical purposes is that finite truncations of the series *never* correspond to any approximating systems. Much more convenient, in particular for path planning algorithms [47,48,63,64,75,95], but also in applications for numerical integration [17,46] are expansions as directed infinite products of exponentials or as an exponential of a Lie series.

In terms of the map $\mathcal{F}$ that substitutes for each letter $a \in Z$ the corresponding vector field f_a, one finds that the Chen–Fliess series (39) is simply the image of a natural object under the map $\Upsilon \otimes \mathcal{F}$. Indeed, under the usual identification of the space $\mathrm{Hom}(V, W)$ of linear maps between vector spaces V and W with the product $V^* \otimes W$, and noting that Z^* is an orthonormal basis for $A(Z)$, the identity map from $A(Z)$ to $A(Z)$ is identified with the series

$$\mathrm{Id}_{A(Z)} \equiv \sum_{w \in Z^*} w \otimes w \quad \in \quad A(Z) \otimes \hat{A}(Z)\,. \tag{50}$$

Thus, any rewriting of the combinatorial object on the right hand side will, via the map $\Upsilon \otimes \mathcal{F}$, give an alternative presentation of the Chen–Fliess series. In particular, from elementary consideration, using also the Hopf-algebraic structures of $A(Z)$, it is a-priori clear that there exist expansions in the forms

$$\sum_{w \in Z^*} w \otimes w = \exp\left(\sum_{h \in \mathcal{H}} \zeta_h \otimes [h]\right) = \overleftarrow{\prod_{h \in \mathcal{H}}} \exp\left(\xi_h \otimes [h]\right) \tag{51}$$

where $\mathcal{H}$ indexes an ordered basis $\{[h] \colon H \in \mathcal{H}\}$ of the free Lie algebra $L(Z)$ and for each $h \in \mathcal{H}$, ξ_h and ζ_h are polynomials in $A(Z)$ that are mapped by Υ to corresponding iterated integral functionals. The usefulness of such expression depends on the availability of simple formulas for $\mathcal{H}$, ξ_h and ζ_h.

Bases for free Lie algebras are well-known since Hall [42], and have been unified by Viennot [101]. Specifically, a Hall set over a set Z is any strictly ordered subset $\tilde{\mathcal{H}} \subseteq \mathcal{T}(\mathbb{Z})$ from the set $\mathcal{M}(Z)$ labeled binary trees (with leaves labeled by Z) that satisfies

(i) $Z \subseteq \tilde{\mathcal{H}}$
(ii) Suppose $a \in Z$. Then $(t, a) \in \tilde{\mathcal{H}}$ iff $t' \in \tilde{\mathcal{H}}$, $t' < a$ and $a < (t', a)$.
(iii) Suppose $u, v, w, (u, v) \in \tilde{\mathcal{H}}$. Then $(t', (t''', t'''')) \in \tilde{\mathcal{H}}$ iff $t''' \le t' \le (t''', t'''')$ and $t' < (t', (t''', t''''))$.

There are natural mappings $\theta \colon \mathcal{T}(Z) \mapsto L(Z) \subseteq A(Z)$ and $\vartheta \colon \mathcal{T}(Z) \mapsto A(Z)$ (the *foliage map*) that map labeled binary trees to Lie polynomials and to words, and which are defined for $a \in Z$, $(t', t'') \in \mathcal{T}(Z)$ by $\theta(a) = \vartheta(a) = a$ and recursively $\theta((t', t'')) = [\theta(t'), \theta(t'')]$ and $\vartheta((t', t'')) = \vartheta(t')\vartheta(t'')$. The image of a Hall set under the map θ is a basis for the free Lie algebra $L(Z)$. Moreover, the restriction of the map ϑ to a Hall-set is one-to-one which leads to a fundamental unique factorization property of words into products of Hall-words, and a unique way of recovering Hall trees from the Hall words that are their images under ϑ [101]. This latter property is very convenient as it allows one to use Hall words rather than Hall trees for indexing various objects.

The construction of Hall sets, as well as the proof that they give rise to bases of free Lie algebras, rely in an essential way on the process of *Lazard elimination* [9] which is intimately related to the solution of differential Eqs. (32) or (46) by a successive variation of parameters [93]. As a result, using Hall sets it is possible to obtain an extremely simple and elegant recursive formula for the coefficients ξ_h in (51), whereas formulas for the ζ_h are less immediate, although they may be computed by straightforward formulas in terms of natural maps of the Hopf algebra structure on $A(Z)$ [33]. Up to a normalization factor μ in terms of multi-factorials, the formula for the ξ_h is extremely simple

in terms of the Zinbiel product on $A(Z)$. For letters $a \in Z$ and Hall words $h, k, hk \in \mathcal{H}$ one has

$$\xi_a = a \quad \text{and} \quad \xi_{hk} = \mu_{hk} \cdot \xi_h * \xi_k \,. \tag{52}$$

Using completely different notation, this formula appeared originally in the work of Schützenberger [84], and was later rediscovered in various settings by Grayson and Grossman [38], Melancon and Reutenauer [69] and Sussmann [93]. In the context of control, it appeared first in [93].

To illustrate the simplicity of the recursive formula, which is based on the factorization of Hall words, consider the following *normal form* for a free nilpotent system (of rank $r = 5$) using a typical Hall set over the alphabet $Z = \{0, 1\}$. This is the closest nonlinear analogue to the Kalman controller normal form of a linear system, which is determined by Kronecker indices that classify the lengths of chains of integrators. For practical computations nilpotent systems are commonly used as approximations of general nonlinear systems – and every nilpotent system (that is, a system of form (32) whose vector fields f_a generate a nilpotent Lie algebra) is the image of a free nilpotent system as below under some projection map. For convenience, the example again uses the notation x_h instead of ξ_h.

$$\begin{aligned}
\dot{x}_0 &= u_0 \\
\dot{x}_1 &= u_1 \\
\dot{x}_{01} &= x_0 \cdot \dot{x}_1 = x_0\, u_1 \\
\dot{x}_{001} &= x_0 \cdot \dot{x}_{01} = x_0^2\, u_1 \\
&\quad \text{using } \vartheta^{-1}(001) = (0(01)) \\
\dot{x}_{101} &= x_1 \cdot \dot{x}_{01} = x_1 x_0\, u_1 \\
&\quad \text{using } \vartheta^{-1}(101) = (1(01)) \\
\dot{x}_{0001} &= x_0 \cdot \dot{x}_{001} = x_0^3\, u_1 \\
&\quad \text{using } \vartheta^{-1}(0001) = (0(0(01))) \\
\dot{x}_{1001} &= x_1 \cdot \dot{x}_{001} = x_1 x_0^2\, u_1 \\
&\quad \text{using } \vartheta^{-1}(1001) = (1(0(01))) \\
\dot{x}_{1101} &= x_1 \cdot \dot{x}_{101} = x_1^2 x_0\, u_1 \\
&\quad \text{using } \vartheta^{-1}(1101) = (1(1(01))) \\
\dot{x}_{00001} &= x_0 \cdot \dot{x}_{001} = x_0^4\, u_1 \\
&\quad \text{using } \vartheta^{-1}(00001) = (0(0(0(01)))) \\
\dot{x}_{10001} &= x_1 \cdot \dot{x}_{0001} = x_1 x_0^3\, u_1 \\
&\quad \text{using } \vartheta^{-1}(10001) = (1(0(0(01)))) \\
\dot{x}_{11001} &= x_1 \cdot \dot{x}_{1001} = x_1^2 x_0^2\, u_1 \\
&\quad \text{using } \vartheta^{-1}(11001) = (1(1(0(01)))) \\
\dot{x}_{01001} &= x_{01} \cdot \dot{x}_{001} = x_{01} x_0^3\, u_1 \\
&\quad \text{using } \vartheta^{-1}(01001) = ((01)(0(01))) \\
\dot{x}_{01101} &= x_{01} \cdot \dot{x}_{101} = x_{01} x_1^2 x_0\, u_1 \\
&\quad \text{using } \vartheta^{-1}(01101) = ((01)(1(01)))\,.
\end{aligned}$$

This example may be interpreted as a system of form (32) with the x_h being a set of coordinate functions on the manifold M, or as the truncation of the lifted system (46) with the x_h being a basis for a finite dimensional subspace of the linear space $\mathcal{E}(M)$. For more details and the technical background see [56,57].

Future Directions

The chronological calculus is still a young methodology. Its intrinsically infinite dimensional character demands a certain sophistication from its users, and thus it may take some time until it reaches its full potential. From a different point of view this also means that there are many opportunities to explore its advantages in ever new areas of applications. A relatively straightforward approach simply checks classical topics of systems and control for whether they are amenable to this approach, and whether this will be superior to classical techniques. The example of [70] which connects the chronological calculus with discrete time systems is a model example for such efforts. There are many further opportunities, some more speculative than others.

Control The chronological calculus appears ideally suited for the investigation of fully nonlinear systems – but in this area there is still much less known than in the case of nonlinear control systems that are affine in the control, or the case of linear systems. Specific conditions for the existence of solutions, controllability, stabilizability, and normal forms are just a few subjects that, while studies have been initiated [1,2,4], deserve further attention.

Computation In recent years much effort has been devoted to understanding the foundations of numerical integration algorithms in nonlinear settings, and to eventually design more efficient algorithms and eventually prove their superiority. The underlying mathematical structures, such as the compositions of noncommuting flows are very similar to those studies in control, as observed in e. g. [23], one of the earlier publications in this area. Some more recent work in this direction is [17,46,72,73,74]. This is also very closely related to the studies of the underlying combinatorial and algebraic structures – arguably [27,28,29,76] are some of the ones most closely related to the subject

of this article. There appears to be much potential for a two-way exchange of ideas and results.

Geometry Arguably one of the main thrusts will continue to be the use of the chronological calculus for studying the differential geometric underpinnings of systems and control theory. But it is conceivable that such studies may eventually abstract further – just like after the 1930s – to studies of generalizations of algebras that stem from systems on classical manifolds, much in the spirit of modern noncommutative geometry with [19] being the best-known proponent.

Algebra and combinatorics In general very little is known about possible ideal and subalgebra structures of chronological and Zinbiel algebras. This work was initiated in [2], but relatively little progress has been made since. (Non)existence of finite dimensional (nilpotent) Zinbiel algebras has been established, but only over a complex field [25,26]. Bases of free chronological algebras are discussed in [2], but many other aspects of this algebraic structure remain unexplored. This also intersects with the aforementioned efforts in foundations of computation – on one side there are Hopf-algebraic approaches to free Lie algebras [78] and nonlinear control [33,40,41], while the most recent breakthroughs involve dendriform algebras and related subjects [27,28,29]. This is an open field for uncovering the connections between combinatorial and algebraic objects on one side and geometric and systems objects on the other side. The chronological calculus may well serve as a vehicle to elucidate the correspondences.

Acknowledgments

This work was partially supported by the National Science Foundation through the grant DMS 05-09030

Bibliography

1. Agrachëv A, Gamkrelidze R (1978) Exponential representation of flows and chronological calculus. Math Sbornik USSR (Russian) 107(N4):487–532. Math USSR Sbornik (English translation) 35:727–786
2. Agrachëv A, Gamkrelidze R (1979) Chronological algebras and nonstationary vector fields. J Soviet Math 17:1650–1675
3. Agrachëv A, Gamkrelidze R, Sarychev A (1989) Local invariants of smooth control systems. Acta Appl Math 14:191–237
4. Agrachëv A, Sachkov YU (1993) Local controllability and semigroups of diffeomorphisms. Acta Appl Math 32:1–57
5. Agrachëv A, Sachkov YU (2004) Control Theory from a Geometric Viewpoint. Springer, Berlin
6. Agrachëv A, Sarychev A (2005) Navier Stokes equations: controllability by means of low modes forcing. J Math Fluid Mech 7:108–152
7. Agrachëv A, Vakhrameev S (1983) Chronological series and the Cauchy–Kowalevski theorem. J Math Sci 21:231–250
8. Boltyanski V, Gamkrelidze R, Pontryagin L (1956) On the theory of optimal processes (in Russian). Doklady Akad Nauk SSSR, vol.10, pp 7–10
9. Bourbaki N (1989) Lie Groups and Lie algebras. Springer, Berlin
10. Brockett R (1971) Differential geometric methods in system theory. In: Proc. 11th IEEE Conf. Dec. Cntrl., Berlin, pp 176–180
11. Brockett R (1976) Volterra series and geometric control theory. Autom 12:167–176
12. Bullo F (2001) Series expansions for the evolution of mechanical control systems. SIAM J Control Optim 40:166–190
13. Bullo F (2002) Averaging and vibrational control of mechanical systems. SIAM J Control Optim 41:542–562
14. Bullo F, Lewis A (2005) Geometric control of mechanical systems: Modeling, analysis, and design for simple mechanical control systems. Texts Appl Math 49 IEEE
15. Caiado MI, Sarychev AV () On stability and stabilization of elastic systems by time-variant feedback. ArXiv:math.AP/0507123
16. Campbell J (1897) Proc London Math Soc 28:381–390
17. Casas F, Iserles A (2006) Explicit Magnus expansions for nonlinear equations. J Phys A: Math General 39:5445–5461
18. Chen KT (1957) Integration of paths, geometric invariants and a generalized Baker–Hausdorff formula. Ann Math 65:163–178
19. Connes A (1994) Noncommutative geometry. Academic Press, San Diego
20. Cortés J, Martinez S (2003) Motion control algorithms for simple mechanical systems with symmetry. Acta Appl Math 76:221–264
21. Cortés J, Martinez S, Bullo F (2002) On nonlinear controllability and series expansions for Lagrangian systems with dissipative forces. Trans IEEE Aut Control 47:1401–1405
22. Crouch P (1981) Dynamical realizations of finite Volterra series. SIAM J Control Optim 19:177–202
23. Crouch P, Grossman R (1993) Numerical integration of ordinary differential equations on manifolds. J Nonlinear Sci 3:1–33
24. Crouch P, Lamnabhi-Lagarrigue F (1989) Algebraic and multiple integral identities. Acta Appl Math 15:235–274
25. Dzhumadil'daev A (2007) Zinbiel algebras over a q-commutator. J Math Sci 144:3909–3925
26. Dzhumadil'daev A, Tulenbaev K (2005) Nilpotency of Zinbiel algebras. J Dyn Control Syst 11:195–213
27. Ebrahimi-Fard K, Guo L (2007) Rota–Baxter algebras and dendriform algebras. J Pure Appl Algebra 212:320–339
28. Ebrahimi-Fard K, Manchon D, Patras F (2007) A Magnus- and Fer-type formula in dendriform algebras. J Found Comput Math (to appear) http://springerlink.com/content/106038/
29. Ebrahimi-Fard K, Manchon D, Patras F (2008) New identities in dendriform algebras. J Algebr 320:708–727
30. Fliess M (1978) Développements fonctionelles en indéterminées non commutatives des solutions d'équations différentielles non linéaires forcées. CR Acad Sci France Ser A 287:1133–1135
31. Fliess M (1981) Fonctionelles causales nonlinéaires et indeterminées noncommutatives. Bull Soc Math France 109:3–40
32. Gamkrelidze RV, Agrachëv AA, Vakhrameev SA (1991) Ordinary differential equations on vector bundles and chronological calculus. J Sov Math 55:1777–1848

33. Gehrig E (2007) Hopf algebras, projections, and coordinates of the first kind in control theory. Ph D Dissertation, Arizona State University
34. Gelfand I (1938) Abstract functions and linear operators. Math Sbornik NS 4:235–284
35. Gelfand I, Raikov D, Shilov G (1964) Commutative normed rings. (Chelsea) New York (translated from the Russian, with a supplementary chapter), Chelsea Publishing, New York
36. Ginzburg V, Kapranov M (1994) Koszul duality for operads. Duke Math J 76:203–272
37. Gray W, Wang Y (2006) Noncausal fliess operators and their shuffle algebra. In: Proc MTNS 2006 (Mathematical Theory of Networks and Systems). MTNS, Kyoto, pp 2805–2813
38. Grayson M, Grossman R (1990) Models for free nilpotent algebras. J Algebra 135:177–191
39. Grayson M, Larson R (1991) The realization of input-output maps using bialgebras. Forum Math 4:109–121
40. Grossman R, Larson R (1989) Hopf-algebraic structure of combinatorial objects and differential operators. Israel J Math 72:109–117
41. Grossman R, Larson R (1989) Hopf-algebraic structure of families of trees. J Algebra 126:184–210
42. Hall M (1950) A basis for free Lie rings and higher commutators in free groups. Proc Amer Math Soc 1:575–581
43. Haynes G, Hermes H (1970) Nonlinear controllability via Lie theory. SIAM J Control 8:450–460
44. Herman R (1963) On the accessibility problem in control theory. In: Int. Symp. Nonlinear Diff. Eqns. Nonlinear Mechanics. Academic Press, New York, pp 325–332
45. Hermann R, Krener A (1977) Nonlinear controllability and observability. IEEE Trans Aut Control 22:728–740
46. Iserles A, Munthe-Kaas H, Nrsett S, Zanna A (2000) Lie-group methods. Acta numerica 9:215–365
47. Jacob G (1991) Lyndon discretization and exact motion planning. In: Proc. Europ. Control Conf., pp 1507–1512, ECC, Grenoble
48. Jacob G (1992) Motion planning by piecewise constant or polynomial inputs. In: Proc. IFAC NOLCOS. Int Fed Aut, Pergamon Press, Oxford
49. Jakubczyk B (1986) Local realizations of nonlinear causal operators. SIAM J Control Opt 24:231–242
50. Jurdjevic V, Sussmann H (1972) Controllability of nonlinear systems. J Diff Eqns 12:95–116
51. Kalman R (1960) A new approach to linear filtering and prediction problems. Trans ASME – J Basic Eng 82:35–45
52. Kawski M (1988) Control variations with an increasing number of switchings. Bull Amer Math Soc 18:149–152
53. Kawski M (1990) High-order small-time local controllability. In: Sussmann H (ed) Nonlinear Controllability and Optimal Control. Dekker, pp 441–477, New York
54. Kawski M (2000) Calculating the logarithm of the Chen Fliess series. In: Proc. MTNS 2000, CDROM. Perpignan, France
55. Kawski M (2000) Chronological algebras: combinatorics and control. Itogi Nauki i Techniki 68:144–178 (translation in J Math Sci)
56. Kawski M (2002) The combinatorics of nonlinear controllability and noncommuting flows. In: Abdus Salam ICTP Lect Notes 8. pp 223–312, Trieste
57. Kawski M, Sussmann HJ (1997) Noncommutative power series and formal Lie-algebraic techniques in nonlinear control theory. In: Helmke U, Prätzel–Wolters D, Zerz E (eds) Operators, Systems, and Linear Algebra. Teubner, pp 111–128 , Stuttgart
58. Kirov N, Krastanov M (2004) Higher order approximations of affinely controlled nonlinear systems. Lect Notes Comp Sci 2907:230–237
59. Kirov N, Krastanov M (2005) Volterra series and numerical approximation of ODEs. Lect Notes Comp Sci 2907:337–344. In: Li Z, Vulkov L, Was'niewski J (eds) Numerical Analysis and Its Applications. Springer, pp 337–344, Berlin
60. Komleva T, Plotnikov A (2000) On the completion of pursuit for a nonautonomous two-person game. Russ Neliniini Kolyvannya 3:469–473
61. Krener A, Lesiak C (1978) The existence and uniqueness of Volterra series for nonlinear systems. IEEE Trans Aut Control 23:1090–1095
62. Kriegl A, Michor P (1997) The convenient setting of global analysis. Math Surv Monogr 53. Amer Math Soc, Providence
63. Lafferiere G, Sussmann H (1991) Motion planning for controllable systems without drift. In: IEEE Conf. Robotics and Automation. pp 1148–1153, IEEE Publications, New York
64. Lafferiere G, Sussmann H (1993) A differential geometric approach to motion planning. In: Li Z, Canny J (eds) Nonholonomic Motion Planning. Kluwer, Boston, pp 235–270
65. Lobry C (1970) Controllabilit'e des systèmes non linéares. SIAM J Control 8:573–605
66. Loday JL (1993) Une version non commutative des algèbres de Lie: les algèbres de Leibniz. Enseign Math 39:269–293
67. Loday JL, Pirashvili T (1996) Leibniz representations of Lie algebras. J Algebra 181:414–425
68. Martinez S, Cortes J, Bullo F (2003) Analysis and design of oscillatory control systems. IEEE Trans Aut Control 48:1164–1177
69. Melançon G, Reutenauer C (1989) Lyndon words, free algebras and shuffles. Canadian J Math XLI:577–591
70. Monaco S, Normand-Cyrot D, Califano C (2007) From chronological calculus to exponential representations of continuous and discrete-time dynamics: a lie-algebraic approach. IEEE Trans Aut Control 52:2227–2241
71. Morgansen K, Vela P, Burdick J (2002) Trajectory stabilization for a planar carangiform robot fish. In: Proc. IEEE Conf. Robotics and Automation. pp 756–762, New York
72. Munthe-Kaas H, Owren B (1999) Computations in a free Lie algebra. Royal Soc Lond Philos Trans Ser A 357:957–981
73. Munthe-Kaas H, Wright W (2007) On the Hopf algebraic structure of lie group integrators. J Found Comput Math 8(2):227–257
74. Munthe-Kaas H, Zanna A (1997) Iterated commutators, lie's reduction method and ordinary differential equations on matrix lie groups. In: Cucker F (ed) Found. Computational Math. Springer, Berlin, pp 434–441
75. Murray R, Sastry S (1993) Nonholonomic path planning: steering with sinusoids. IEEE T Autom Control 38:700–716
76. Murua A (2006) The Hopf algebra of rooted trees, free Lie algebras, and Lie series. J Found Comput Math 6:387–426
77. Ree R (1958) Lie elements and an algebra associated with shuffles. Annals Math 68:210–220
78. Reutenauer C (1991) Free Lie Algebras. Oxford University Press, New York
79. Rocha E (2003) On computation of the logarithm of the Chen–Fliess series for nonlinear systems. In: Zinober I, Owens D (eds)

Nonlinear and adaptive Control, Lect Notes Control Inf Sci 281:317–326, Sprtinger, Berlin
80. Rocha E (2004) An algebraic approach to nonlinear control theory. Ph D Dissertation, University of Aveiro, Portugal
81. Sanders J, Verhulst F (1985) Averaging methods in nonlinear dynamical systems. Appl Math Sci 59. Springer, New York
82. Sarychev A (2001) Lie- and chronologico-algebraic tools for studying stability of time-varying systems. Syst Control Lett 43:59–76
83. Sarychev A (2001) Stability criteria for time-periodic systems via high-order averaging techniques. In: Lect. Notes Control Inform. Sci. 259. Springer, London, pp 365–377
84. Schützenberger M (1958) Sur une propriété combinatoire des algèbres de Lie libres pouvant être utilisée dans un probléme de mathématiques appliquées. In: Dubreil S (ed) Algèbres et Théorie des Nombres. Faculté des Sciences de Paris vol 12 no 1 (1958-1959), Exposé no 1 pp 1–23
85. Serres U (2006) On the curvature of two-dimensional optimal control systems and zermelos navigation problem. J Math Sci 135:3224–3243
86. Sigalotti M (2005) Local regularity of optimal trajectories for control problems with general boundary conditions. J Dyn Control Syst 11:91–123
87. Sontag E, Wang Y (1992) Generating series and nonlinear systems: analytic aspects, local realizability and i/o representations. Forum Math 4:299–322
88. Stefani G (1985) Polynomial approximations to control systems and local controllability. In: Proc. 25th IEEE Conf. Dec. Cntrl., pp 33–38, New York
89. Stone M (1932) Linear Transformations in Hilbert Space. Amer Math Soc New York
90. Sussmann H (1974) An extension of a theorem of Nagano on transitive Lie algebras. Proc Amer Math Soc 45:349–356
91. Sussmann H (1983) Lie brackets and local controllability: a sufficient condition for scalar-input systems. SIAM J Cntrl Opt 21:686–713
92. Sussmann H (1983) Lie brackets, real analyticity, and geometric control. In: Brockett RW, Millman RS, Sussmann HJ (eds) Differential Geometric Control. pp 1–116, Birkhauser
93. Sussmann H (1986) A product expansion of the Chen series. In: Byrnes C, Lindquist A (eds) Theory and Applications of Nonlinear Control Systems. Elsevier, North-Holland, pp 323–335
94. Sussmann H (1987) A general theorem on local controllability. SIAM J Control Opt 25:158–194
95. Sussmann H (1992) New differential geometric methods in nonholonomic path finding. In: Isidori A, Tarn T (eds) Progr Systems Control Theory 12. Birkhäuser, Boston, pp 365–384
96. Tretyak A (1997) Sufficient conditions for local controllability and high-order necessary conditions for optimality. A differential-geometric approach. J Math Sci 85:1899–2001
97. Tretyak A (1998) Chronological calculus, high-order necessary conditions for optimality, and perturbation methods. J Dyn Control Syst 4:77–126
98. Tretyak A (1998) Higher-order local approximations of smooth control systems and pointwise higher-order optimality conditions. J Math Sci 90:2150–2191
99. Vakhrameev A (1997) A bang-bang theorem with a finite number of switchings for nonlinear smooth control systems. Dynamic systems 4. J Math Sci 85:2002–2016
100. Vela P, Burdick J (2003) Control of biomimetic locomotion via averaging theory. In: Proc. IEEE Conf. Robotics and Automation. pp 1482–1489, IEEE Publications, New York
101. Viennot G (1978) Algèbres de Lie Libres et Monoïdes Libres. Lecture Notes in Mathematics, vol 692. Springer, Berlin
102. Visik M, Kolmogorov A, Fomin S, Shilov G (1964) Israil Moiseevich Gelfand, On his fiftieth birthday. Russ Math Surv 19:163–180
103. Volterra V (1887) Sopra le funzioni che dipendono de altre funzioni. In: Rend. R Academia dei Lincei. pp 97–105, 141–146, 153–158
104. von Neumann J (1932) Mathematische Grundlagen der Quantenmechanik. Grundlehren Math. Wissenschaften 38. Springer, Berlin
105. Zelenko I (2006) On variational approach to differential invariants of rank two distributions. Diff Geom Appl 24:235–259

Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies

Michael Batty
Center for Advanced Spatial Analysis,
University College London, London, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Cities in Equilibrium
Urban Dynamics
Comprehensive System Models of Urban Structure
Future Directions
Bibliography

Glossary

Agent-based models Systems composed of individuals who act purposely in making locational/spatial decisions.

Bifurcation A process whereby divergent paths are generated in a trajectory of change in an urban system.

City size distribution A set of cities ordered by size, usually population, often in rank order.

Emergent patterns Land uses or economic activities which follow some spatial order.

Entropy maximizing The process of generating a spatial model by maximizing a measure of system complexity subject to constraints.

Equilibrium A state of the urban system which is balanced and unchanging.

Exponential growth The process whereby an activity changes through positive feedback on itself.

Fast dynamics A process of frequent movement between locations, often daily.

Feedback The process whereby a system variable influences another variable, either positively or negatively.

Fractal structure A pattern or arrangement of system elements that are self-similar at different spatial scales.

Land use transport model A model linking urban activities to transport interactions.

Life cycle effects Changes in spatial location which are motivated by aging of urban activities and populations.

Local neighborhood The space immediately around a zone or cell.

Logistic growth Exponential growth capacitated so that some density limit is not exceeded.

Lognormal distribution A distribution which has fat and long tails which is normal when examined on a logarithmic scale.

Microsimulation The process of generating synthetic populations from data which is collated from several sources.

Model validation The process of calibrating and testing a model against data so that its goodness of fit is optimized.

Multipliers Relationships which embody *n*th order effects of one variable on another.

Network scaling The in-degrees and out-degrees of a graph whose nodal link volumes follow a power law.

Population density profile A distribution of populations which typically follows an exponential profile when arrayed against distance from some nodal point.

Power laws Scaling laws that order a set of objects according to their size raised to some power.

Rank size rule A power law that rank orders a set of objects.

Reaction-diffusion The process of generating changes as a consequence of a reaction to an existing state and interactions between states.

Scale-free networks Networks whose nodal volumes follow a power law.

Segregation model A model which generates extreme global segregation from weak assumptions about local segregation.

Simulation The process of generating locational distributions according to a series of sub-model equations or rules.

Slow dynamics Changes in the urban system that take place over years or decades.

Social physics The application of classical physical principles involving distance, force and mass to social situations, particularly to cities and their transport.

Spatial interaction The movement of activities between different locations ranging from traffic distributions to migration patterns.

Trip distribution The pattern of movement relating to trips made by the population, usually from home to work but also to other activities such as shopping.

Urban hierarchy A set of entities physically or spatially scaled in terms of their size and areal extent.

Urban morphology Patterns of urban structure based on the way activities are ordered with respect to their locations.

Urban system A city represented as a set of interacting subsystems or their elements.

Definition of the Subject

Cities have been treated as systems for fifty years but only in the last two decades has the focus changed from aggregate equilibrium systems to more evolving systems whose structure emerges from the bottom up. We first outline the rudiments of the traditional approach focusing on equilibrium and then discuss how the paradigm has changed to one which treats cities as emergent phenomena generated through a combination of hierarchical levels of decision, driven in decentralized fashion. This is consistent with the complexity sciences which dominate the simulation of urban form and function. We begin however with a review of equilibrium models, particularly those based on spatial interaction, and we then explore how simple dynamic frameworks can be fashioned to generate more realistic models. In exploring dynamics, nonlinear systems which admit chaos and bifurcation have relevance but recently more pragmatic schemes of structuring urban models based on cellular automata and agent-based modeling principles have come to the fore. Most urban models deal with the city in terms of the location of its economic and demographic activities but there is also a move to link such models to urban morphologies which are clearly fractal in structure. Throughout this chapter, we show how key concepts in complexity such as scaling, self-similarity and far-from-equilibrium structures dominate our current treatment of cities, how we might simulate their functioning and how we might predict their futures. We conclude with the key problems that dominate the field and suggest how these might be tackled in future research.

Cities were first conceived as complex systems in the 1960s when architects and urban planners began to change their perceptions that cities be treated as 'works of art' to something much more akin to a functioning economic

system that required social engineering. Since the time of Newton, social scientists had speculated that social systems could be described using concepts from classical physics and the notion that interrelationships between the component parts of such systems might be articulated using concepts of mass, force and energy established a rudimentary framework that came to be known as social physics. Together with various macro-economic theories of how economies function based on Keynesian ideas of input, output and economic multipliers, cities were assumed to be equilibrium systems in which their interactions such as traffic, trade flows, and demographic migration could be modeled in analogy to gravitation. These flows were seen as complementary to the location of employment and population which reflected costs of travel and land rent, in turn the product of micro-economic theories of how agents resolved their demand for and the supply of space through the land market. Operational models for policy analysis were initially built on this basis and used for testing the impact of different plans for making cities more efficient, locationally and in terms of their movement/traffic patterns.

This early work did not emphasize the dynamics of urban change or the morphology of cities. Thus theories and models were limited in their ability to predict patterns of urban growth as reflected in sprawl and the regeneration of urban areas. The first wave of models which treated the city system in aggregate, static and top down fashion, fell into disrepute due to these limitations. To address them, the focus moved in the 1980s to more theoretical considerations in which static social physics types of model were embedded in nonlinear dynamic frameworks built around ideas in chaos, and bifurcation theory. A parallel development in simulating urban morphology was built around treating cities as fractals which evolved from the bottom up and operational models in which such morphologies were governed using cellular automata were developed. Developments in moving from aggregate to disaggregate or bottom-up individual-based models were spawned from these developments with agent-based modeling providing a new focus to the field. There is now a momentum for treating urban aggregates through new ideas about growth and form through scaling while new forms of representing cities through networks linking these to fractal morphologies are being developed through network science at different scales. These will ultimately lead to operational urban and transport models built from the bottom up which simulate spatial processes operating on networks and other morphologies. These have the potential to address key problems of urban growth and evolution, congestion, and inequality.

Introduction

Cities were first treated formally as systems when General System Theory and Cybernetics came to be applied to the softer social sciences in the 1950s. Ludwig von Bertalanffy [67] in biology and Norbert Wiener [73] in engineering gave enormous impetus to this emerging interdisciplinary field that thrust upon us the idea that phenomena of interest in many disciplines could be articulated in generic terms as 'systems'. Moreover the prospect that the systems approach could yield generic policy, control and management procedures applicable to many different areas, appeared enticing. The idea of a general systems theory was gradually fashioned from reflections on the way distinct entities which were clearly collections of lower order elements, organized into a coherent whole, displaying pattern and order which in the jargon of the mid-twentieth century was encapsulated in the phrase that "the whole is greater than the sum of the parts". The movement began in biology in the 1920s, gradually eclipsing parts of engineering in the 1950s and spreading to the management and social sciences, particularly sociology and political science in the 1960s. It was part of a wave of change in the social sciences which began in the late 19th century as these fields began to emulate the physical sciences, espousing positivist methods which had appeared so successful in building applicable and robust theory.

The focus then was on ways in which the elements comprising the system interacted with one another through structures that embodied feedbacks keeping the system sustainable within bounded limits. The notion that such systems have controllers to 'steer' them to meet certain goals or targets is central to this early paradigm and the science of "…control and communication in the animal and the machine" was the definition taken up by Norbert Wiener [73] in his exposition of the science of cybernetics. General system theory provided the generic logic for both the structure and behavior of such systems through various forms of feedback and hierarchical organization while cybernetics represented the 'science of steersmanship' which would enable such systems to move towards explicit goals or targets. Cities fit this characterization admirably and in the 1950s and 1960s, the traditional approach that articulated cities as structures that required physical and aesthetic organization, quickly gave way to deeper notions that cities needed to be understood as general systems. Their control and planning thus required much more subtle interventions than anything that had occurred hitherto in the name of urban planning.

Developments in several disciplines supported these early developments. Spatial analysis, as it is now called,

began to develop within quantitative geography, linked to the emerging field of regional science which represented a synthesis of urban and regional economics in which location theory was central. In this sense, the economic structure of cities and regions was consistent with classical macro and micro economics and the various techniques and models that were developed within these domains had immediate applicability. Applications of physical analogies to social and city systems, particularly ideas about gravitation and potential, had been explored since the mid 19th century under the banner of 'social physics' and as transportation planning formally began in the 1950s, these ideas were quickly adopted as a basis for transport modeling. Softer approaches in sociology and political science also provided support for the idea of cities as organizational systems while the notion of cybernetics as the basis for management, policy and control of cities was adopted as an important analogy in their planning [26,53].

The key ideas defined cities as sets of elements or components tied together through sets of interactions. The archetypal structure was fashioned around land use activities with economic and functional linkages between them represented initially in terms of physical movement, traffic. The key idea of feedback, which is the dynamic that holds a general system together, was largely represented in terms of the volume and pattern of these interactions, at a single point in time. Longer term evolution of urban structure was not central to these early conceptions for the focus was largely on how cities functioned as equilibrium structures. The prime imperative was improving how interactions between component land uses might be made more efficient while also meeting goals involving social and spatial equity. Transportation and housing were of central importance in adopting the argument that cities should be treated as examples of general systems and steered according to the principles of cybernetics.

Typical examples of such systemic principles in action involve transportation in large cities and these early ideas about systems theory hold as much sway in helping make sense of current patterns as they did when they were first mooted fifty or more years ago. Different types of land use with different economic foci interact spatially with respect to how employees are linked to their housing locations, how goods are shipped between different locations to service the production and consumption that define these activities, how consumers purchase these economic activities which are channeled through retail and commercial centers, how information flows tie all these economies together, and so on: the list of linkages is endless. These activities are capacitated by upper limits on density and capacity. In Greater London for example, the traffic has reached saturation limits in the central city and with few new roads being constructed over the last 40 years, the focus has shifted to improving public transport and to road pricing.

The essence of using a systems model of spatial interaction to test the impact of such changes on city structure is twofold: first such a model can show how people might shift mode of transport from road to rail and bus, even to walking and cycling, if differential pricing is applied to the road system. The congestion charge in central London imposed in 2003 led to a 30 percent reduction in the use of vehicles and this charge is set to increase massively for certain categories of polluting vehicles in the near future. Second the slightly longer term effects of reducing traffic are to increase densities of living, thus decreasing the length and cost of local work journeys, also enabling land use to respond by changing their locations to lower cost areas. All these effects ripple through the system with the city system models presented here designed to track and predict such *n*th order effects which are rarely obvious. Our focus in this chapter is to sketch the state-of-the-art in these complex systems models showing how new developments in the methods of the complexity sciences are building on a basis that was established half century ago.

Since early applications of general systems theory, the paradigm has changed fundamentally from a world where systems were viewed as being centrally organized, from the top down, and notions about hierarchy were predominant, to one where we now consider systems to be structured from the bottom up. The idea that one or the other – the centralized or the decentralized view – are mutually exclusive of each other is not entirely tenable of course but the balance has certainly changed. Theories have moved from structures and behaviors being organized according to some central control to theories about how systems retain their own integrity from the bottom up, endorsing what Adam Smith over 300 years ago, called "the hidden hand". This shift has brought onto the agenda the notion of equilibrium and dynamics which is now much more central to systems theory than it ever was hitherto. Systems such as cities are no longer considered to be equilibrium structures, notwithstanding that many systems models built around equilibrium are still eminently useful. The notion that city systems are more likely to be in disequilibrium, all the time, or even classed as far-from-equilibrium continually reinforcing the move away from equilibrium, is comparatively new but consistent with the speed of change and volatility in cities observed during the last fifty years.

The notion to that change is nowhere smooth but discontinuous, often chaotic, has become significant. Equi-

librium structures are renewed from within as unanticipated innovations, many technological but some social, change the way people make decisions about how they locate and move within cities. Historical change is important in that historical accidents often force the system onto a less than optimal path with such path dependence being crucial to an understanding of any current equilibria and the dynamic that is evolving. Part of this newly emerging paradigm is the idea that new structures and behaviors that emerge are often unanticipated and surprising. As we will show in this chapter, when we look at urban morphologies, they are messy but ordered, self-similar across many scales, but growing organically from the bottom up. Planned cities are always the exception rather than the rule and when directly planned, they only remain so for very short periods of time.

The new complexity sciences are rewriting the theory of general systems but they are still founded on the rudiments of structures composed of elements, now often called actors or agents, linked through interactions which determine the processes of behavior which keep the system in equilibrium and/or move it to new states. Feedback is still central but recently has been more strongly focused on how system elements react to one another through time. The notion of an unchanging equilibrium supported by such feedbacks is no longer central; feedback is now largely seen as the way in which these structures are evolved to new states. In short, system theory has shifted to consider such feedbacks in positive rather than negative terms although both are essential. Relationships between the system elements in terms of their interactions are being enriched using new ideas from networks and their dynamics [56]. Key notions of how the elements of systems scale relative to one another and relative to their system hierarchies have become useful in showing how local actions and interactions lead to global patterns which can only be predicted from the bottom up [54]. This new view is about how emergent patterns can be generated using models that grow the city from the bottom up [37], and we will discuss all these ideas in the catalog of models that we present below.

We begin by looking at models of cities in equilibrium where we illustrate how interactions between key system elements located in space follow certain scaling laws reflecting agglomeration economies and spatial competition. The network paradigm is closely linked to these ideas in structural terms. None of these models, still important for operational simulation modeling in a policy context, have an internal dynamic and thus we turn to examine dynamics in the next section. We then start with simple exponential growth, showing how it can be capacitated as logistic growth from which nonlinear behaviors can result as chaos and bifurcation. We show how these models might be linked to a faster dynamics built around equilibrium spatial interaction models but to progress these developments, we present much more disaggregate models based on agent simulation and cellular automata principles. These dynamics are then generalized as reaction-diffusion models.

Our third section deals with how we assemble more integrated models built from these various equilibrium and dynamic components or sub-models. We look at large-scale land use transport models which are equilibrium in focus. We then move to cellular automata models of land development, concluding our discussion with reference to the current development of fine scale agent-based models where each individual and trip maker in the city system is simulated. We sprinkle our presentation with various empirical applications, many based on data for Greater London showing how employment and population densities scale, how movement patterns are consistent with the underling infrastructure networks that support them, and how the city has grown through time. We show how the city can be modeled in terms of its structure and the way changes to it can be visualized. We then link these more abstract notions about how cities are structured in spatial-locational terms to their physical or fractal morphology which is a direct expression of their scaling and complexity. We conclude with future directions, focusing on how such models can be validated and used in practical policy-making.

Cities in Equilibrium

Arrangements of Urban Activities

Cities can usually be represented as a series of n locations, each identified by i, and ordered from $i = 1, 2, \ldots, n$. These locations might be points or areas where urban activity takes place, pertaining either to the inter-urban scale where locations are places not necessarily adjacent to one another or at the intra-urban scale where a city is exhaustively partitioned into a set of areas. We will use both representations here but begin with a generic formulation which does not depend on these differences per se.

It is useful to consider the distribution of locations as places where differing amounts of urban activity can take place, using a framework which shows how different arrangements of activity can be consistently derived. Different arrangements of course imply different physical forms of city. Assume there is N amount of activity to be distributed in n locations as $N_1, N_2, \ldots$ Beginning with N_1, there are $N!/[N_1!(N - N_1)!$ allocations of

N_1, $(N - N_1)!/[N_2!(N - N_1 - N_2)!$ allocations of N_2, $(N - N_1 - N_2)!/[N_3!(N - N_1 - N_2 - N_3)!$ of N_3 and so on. To find the total number of arrangements W, we multiply each of these quantities together where the product is

$$W = \frac{N!}{\prod_i N_i!} \,. \tag{1}$$

This might be considered a measure of complexity of the system in that it clearly varies systematically for different allocations. If all N activity were to be allocated to the first location, then $W = 1$ while if an equal amount of activity were to be allocated to each location, then W would vary according to the size of N and the number of locations n. It can be argued that the most likely arrangement of activities would be the one which would give the greatest possibility of distinct individual activities being allocated to locations and such an arrangement could be found by maximizing W (or the logarithm of W which leads to the same). Such maximizations however might be subject to different constraints on the arrangements which imply different conservation laws that the system must meet. This would enable different types of urban form to be examined under different conditions related to density, compactness, sprawl and so on, all of which might be formalized in this way.

To show how this is possible, consider the case where we now maximize the logarithm of W subject to meaningful constraints. The logarithm of Eq. (1) is

$$\ln W = \ln(N!) - \sum_i \ln(N_i!) \,, \tag{2}$$

which using Stirling's formula, simplifies to

$$\ln W \approx N + \ln(N!) - \sum_i N_i \ln N_i \,. \tag{3}$$

N_i which is the number of units of urban activity allocated to location i, is a frequency that can be normalized into a probability as $p_i = N_i/N$. Substituting for $N_i = Np_i$ in Eq. (3) and dropping the constant terms leads to

$$\ln W \propto -\sum_i p_i \ln p_i = H \,, \tag{4}$$

where it is now clear that the formula for the number of arrangements is proportional to Shannon's entropy H. Thus the process of maximizing $\ln W$ is the well-known process of maximizing entropy subject to relevant constraints and this leads to many standard probability distributions [66]. Analogies between city and other social systems with statistical thermodynamics and information theory were developed in the 1960s and represented one of the first formal approaches to the derivation of models for simulating the interaction between locations and the amount of activity attracted to different locations in city, regional and transport systems. As such, it has become a basis on which to build many different varieties of urban model [74].

Although information or entropy has been long regarded as a measure of system complexity, we will not take this any further here except to show how it is useful in deriving different probability distributions of urban activity. Readers are however referred to the mainstream literature for both philosophic and technical expositions of the relationship between entropy and complexity (for example see [42]). The measure H in Eq. (4) is at a maximum when the activity is distributed evenly across locations, that is when $p_i = 1/n$ and $H = \ln n$ while it is at a minimum when $p_i = 1$ and $p_j = 0, j = 1, 2, \ldots, n, i \neq j$, and $H = 0$. It is clear too that H varies with n; that is as the number of locations increases, the complexity or entropy of the system also increases. However what is of more import here is the kind of distribution that maximizing entropy generates when H is maximized subject to appropriate constraints. We demonstrate this as follows for a simple but relevant case where the key constraint is to ensure that the system reproduces the mean value of an attribute of interest. Let p_i be the probability of finding a place i which has P_i population residing there. Then we maximize the entropy

$$H = -\sum_i p_i \ln p_i \,, \tag{5}$$

subject to a normalization constraint on the probabilities

$$\sum_i p_i = 1 \,, \tag{6}$$

and a constraint on the mean population of places $\bar{P}$ in the system, that is

$$\sum_i p_i P_i = \bar{P} \,. \tag{7}$$

The standard method of maximizing Eq. (5) subject to constraint Eqs. (6) and (7) is to form a Langrangian L – a composite of the entropy and the constraints

$$L = -\sum_i p_i \ln p_i - \beta \left(\sum_i p_i - 1 \right) - \vartheta \left(\sum_i p_i P_i - \bar{P} \right) \tag{8}$$

where β and ϑ are multipliers designed to ensure that the constraints are met. Maximizing (8) with respect to p_i gives

$$\frac{\partial L}{\partial p_i} = \ln p_i - 1 - \beta - \vartheta P_i = 0 \,, \tag{9}$$

leading directly to a form for p_i which is

$$p_i = \exp(-\beta - 1)\exp(-\lambda P_i) = K\exp(-\vartheta P_i)\,. \quad (10)$$

K is the composite constant of proportionality which ensures that the probabilities sum to 1. Note also that the sign of the parameters is determined from data through the constraints. If we substitute the probability in Eq. (10) into the Shannon entropy, the measure of complexity of this system which is at a maximum for the given set of constraints, simplifies to $H = \beta + 1 + \vartheta\bar{P}$. There are various interpretations of this entropy with respect to dispersion of activities in the system although these represent a trade-off between the form of the distribution, in this case, the negative exponential, and the number of events or objects n which characterize the system.

Distributions and Densities of Population

The model we have derived can be regarded as an approximation to the distribution of population densities over a set of n spatial zones as long as each zone is the same size (area), that is, $A_i = A, \forall_i$ where nA is the total size (area) of the system. A more general form of entropy takes this area into account by maximizing the expected value of the logarithm of the density, not distribution, where the 'spatial' entropy is defined as

$$S = -\sum_i p_i \ln \frac{p_i}{A_i}\,, \quad (11)$$

with the probability density as p_i/A_i. Using this formula, the procedure simply generalizes the maximization to densities rather than distributions [10] and the model we have derived simply determines these densities with respect to an average population size $\bar{P}$. If we order populations over the zones of a city or even take their averages over many cities in a region or nation, then they are likely to be distributed in this fashion; that is, we would expect there to be many fewer zones or cities of high density than zones or cities of low density, due to competition through growth.

However the way this method of entropy-maximizing has been used to generate population densities in cities is to define rather more specific constraints that relate to space. Since the rise of the industrial city in the 19th century, we have known that population densities tend to decline monotonically with distance from the center of the city. More than 50 years ago, Clark [29] demonstrated quite clearly that population densities declined exponentially with distance from the center of large cities and in the 1960s with the application of micro-economic theory to urban location theory following von Thünen's [68] model, a range of urban attributes such as rents, land values, trip densities, and population densities were shown to be consistent with such negative exponential distributions [5]. Many of these models can also be generated using utility maximizing which under certain rather weak constraints can be seen as equivalent to entropy-maximizing [6]. However it is random utility theory that has been much more widely applied to generate spatial interaction models with a similar form to the models that we generate below using entropy-maximizing [20,46].

We will show how these typical micro-economic urban density distributions can be derived using entropy-maximizing in the following way. Maximizing S in Eq. (11) or H in Eq. (5) where we henceforth assume that the probability p_i is now the population density, we invoke the usual normalization constraint in Eq. (6) and a constraint on the average travel cost $\bar{C}$ incurred by the population given as $\sum_i p_i c_i = \bar{C}$ where c_i is the generalized travel cost/distance from the central business district (CBD) to a zone i. This maximization leads to

$$p_i = K\exp(-\mu c_i) \quad (12)$$

where μ is the parameter controlling the rate of decay of the exponential function, sometimes called the 'friction' of distance or travel cost.

Gravitational Models of Spatial Interaction

It is a simple matter to generalize this framework to generate arrangements of urban activities that deal with interaction patterns, that is movements or linkages between pairs of zones. This involves extending entropy to deal with two rather than one dimensional systems where the focus of interest is on the interaction between an origin zone called $i, i = 1, 2, \ldots, I$ and a destination zone $j, j = 1, 2, \ldots, J$ where there are now a total of IJ interactions in the system. These kinds of model can be used to simulate routine trips from home to work, for example, or to shop, longer term migrations in search of jobs, moves between residential locations in the housing market, as well as trade flows between countries and regions. The particular application depends on context as the generic framework is independent of scale.

Let us now define a two-dimensional entropy as

$$H = -\sum_i \sum_j p_{ij} \ln p_{ij}\,. \quad (13)$$

p_{ij} is the probability of interaction between origin i and destination j where the same distinctions between distribution and density noted above apply. Without loss of

generality, we will assume in the sequel that these variables p_{ij} covary with density in that the origin and destination zones all have the same area. The most constrained system is where we assume that all the interactions originating from any zone i must sum to the probability p_i of originating in that zone, and all interactions destined for zone j must sum to the probability p_j of being attracted to that destination zone. There is an implicit constraint that these origin and destination probabilities sum to 1, that is

$$\sum_i \sum_j p_{ij} = \sum_i p_i = \sum_j p_j = 1 \,, \tag{14}$$

but Eq. (14) is redundant with respect to the origin and destination normalization constraints which are stated explicitly as

$$\left.\begin{aligned} \sum_j p_{ij} &= p_i \\ \sum_i p_{ij} &= p_j \end{aligned}\right\} \,. \tag{15}$$

There is also a constraint on the average distance or cost traveled given as

$$\sum_i \sum_j p_{ij} c_{ij} = \bar{C} \,. \tag{16}$$

The model that is derived from the maximization of Eq. (13) subject to Eqs. (15) and (16) is

$$p_{ij} = K_i K_j p_i p_j \exp(-\gamma c_{ij}) \tag{17}$$

where K_i and K_j are normalization constants associated with Eq. (15), and γ is the parameter on the travel cost c_{ij} between zones i and j associated with Eq. (16). It is easy to compute K_i and K_j by substituting for p_{ij} from Eq. (17) in Eq. (15) respectively and simplifying. This yields

$$\left.\begin{aligned} K_i &= \frac{1}{\sum_j K_j p_j \exp(-\gamma c_{ij})} \\ K_j &= \frac{1}{\sum_i K_i p_i \exp(-\gamma c_{ij})} \end{aligned}\right\} \,, \tag{18}$$

equations that need to be solved iteratively.

These models can be scaled to deal with real trips or population simply by multiplying these probabilities by the total volumes involved, T for total trips in a transport system, P for total population in a city system, Y for total income in a trading system and so on. This system however forms the basis for a family of interaction models which can be generated by relaxing the normalization constraints; for example by omitting the destination constraint, $K_j = 1, \forall_j$, or by omitting the origin constraint, $K_i = 1, \forall_i$ or by omitting both where we need an explicit normalization constraint of the form $\sum_{ij} p_{ij} = 1$ in Eq. (14) to provide an overall constant K. Wilson [74] refers to this set of four models as: *doubly-constrained* – the model in Eqs. (17) and (18), the next two as *singly-constrained*, first when $K_i = 1, \forall_i$, the model is origin constrained, and second when $K_j = 1, \forall_j$, the model is destination constrained; and when we have no constraints on origins or destinations, we need to invoke the global constant K and the model is called *unconstrained*. It is worth noting that these models can also be generated in nearly equivalent form using random utility theory where they are articulated at the level of the individual rather than the aggregate trip-maker and are known as discrete choice models [20].

Let us examine one of these models, a singly-constrained model where there are origin constraints. This might be a model where we are predicting interactions from work to home given we know the distribution of work at the origin zones. Then noting that $K_j = 1, \forall_j$, the model is

$$p_{ij} = K_i p_i p_j \exp(-\gamma c_{ij}) = p_i \frac{p_j \exp(-\gamma c_{ij})}{\sum_j p_j \exp(-\gamma c_{ij})} \,. \tag{19}$$

The key issue with this sort of model is that not only are we predicting the interaction between zones i and j but we can predict the probability of locating in the destination zone p'_j, that is

$$p'_j = \sum_i p_{ij} = \sum_i p_i \frac{p_j \exp(-\gamma c_{ij})}{\sum_j p_j \exp(-\gamma c_{ij})} \,. \tag{20}$$

If we were to drop both origin and destination constraints, the model becomes one which is analogous to the traditional gravity model from which it was originally derived prior to the development of these optimization frameworks. However to generate the usual standard gravitational form of model in which the 'mass' of each origin and destination zone appears, given by P_i and P_j respectively, then we need to modify the entropy formula, thus maximizing

$$H = -\sum_i \sum_j p_{ij} \ln \frac{p_{ij}}{P_i P_j} \,, \tag{21}$$

subject to the normalization

$$\sum_i \sum_j p_{ij} = 1 \,, \tag{22}$$

and this time a constraint on the average 'logarithmic' travel cost $\overline{\ln C}$

$$\sum_i \sum_j p_{ij} \ln c_{ij} = \overline{\ln C} \,. \tag{23}$$

The model that is generated from this system can be written as

$$p_{ij} = K \frac{P_i P_j}{c_{ij}^{\eta}} \,, \tag{24}$$

where the effect of travel cost/distance is now in power law form with η the scaling parameter. Besides illustrating the fact that inverse power forms as well as negative exponential distributions can be generated in this way according to the form of the constraints, one is also able to predict both the probabilities of locating at the origins and the destinations from the traditional gravity model in Eq. (24).

Scaling, City Size, and Network Structure: Power Laws

Distance is a key organizing concept in city systems as we have already seen in the way various urban distributions have been generated. Distance is an attribute of nearness or proximity to the most accessible places and locations. Where there are the lowest distance or travel costs to other places, the more attractive or accessible are those locations. In this sense, distance or travel cost acts as an inferior good in that we wish to minimize the cost occurred in overcoming it. Spatial competition also suggests that the number of places that have the greatest accessibilities are few compared to the majority of places. If you consider that the most accessible place in a circular city is the center, then assuming each place is of similar size, as the number of places by accessibility increases, the lower the accessibility is. In short, there are many places with the same accessibility around the edge of the city compared to only one place in the center. The population density model in Eq. (12) implies such an ordering when we examine the frequency distribution of places according to their densities.

If we now forget distance for a moment, then it is likely that the distribution of places at whatever scale follows a distribution which declines in frequency with attributes based on size due to competition. If we look at all cities in a nation or even globally, there are far fewer big cities than small ones. Thus the entropy-maximizing framework that we have introduced to predict the probability (or frequency) of objects of a certain size occurring, is quite applicable in generating such distributions. We derived a negative exponential distribution in Eq. (10) but to generate a power law, all we need to do is to replace the constraint in Eq. (7) with its logarithmic equivalent, that is

$$\sum_i p_i \ln P_i = \overline{\ln P} \,, \tag{25}$$

and then maximize Eq. (5) subject to (6) and (25) to give

$$p_i = K P_i^{-\varphi} = \exp(-\beta - 1) \exp(-\varphi \ln P_i) \,, \tag{26}$$

where φ is the scaling parameter. Equation (26) gives the probability or frequency – the number of cities – for a zone (or city) with P_i population which is distributed according to an inverse power law. It is important to provide an interpretation of the constraint which generates this power law. Equation (25) implies that the system conserves the average of the logarithm of size which gives greater weight to smaller values of population than to larger, and as such, is recognition that the average size of the system is unbounded as a power function implies. With such distributions, it is unlikely that normality will prevail due to the way competition constrains the distribution in the long tail. Nevertheless in the last analysis, it is an empirical matter to determine the shape of such distributions from data, although early research on the empirical distributions of city sizes following Zipf's Law [81] by Curry [34] and Berry [21] introduced the entropy-maximizing framework to generate such size distributions.

The power law implied for the probability p_i of a certain size P_i of city or zone can be easily generalized to a two-dimensional equivalent which implies a network of interactions. We will maximize the two-dimensional entropy H in Eq. (13) subject to constraints on the mean logarithm of population sizes at origins and destinations which we now state as

$$\left.\begin{aligned} \sum_i \sum_j p_{ij} \ln P_i &= \sum_i p_i \ln P_i = \bar{P}_{\text{origins}} \\ \sum_i \sum_j p_{ij} \ln P_j &= \sum_j p_j \ln P_j = \bar{P}_{\text{destinations}} \end{aligned}\right\} \,, \tag{27}$$

where $p_i = \sum_j p_{ij}$ and $p_j = \sum_i p_{ij}$. Note however that there are no constraints on these origins and destination probabilities p_i and p_j per se but the global constraints in Eq. (14) must hold. This maximization leads to the model

$$p_{ij} = K P_i^{-\lambda_i} P_j^{-\lambda_j} = \frac{P_i^{-\lambda_i}}{\sum_i P_i^{-\lambda_i}} \frac{P_j^{-\lambda_j}}{\sum_j P_j^{-\lambda_j}} \,, \tag{28}$$

where it is clear that the total flows from any origin node or location i vary as

$$p'_i \propto P_i^{-\lambda_i} \,, \tag{29}$$

and the flows into any destination zone vary as

$$p'_j \propto P_j^{-\lambda_j} , \tag{30}$$

with the parameters λ_i and λ_j relating to the mean of the observed logarithmic populations associated with the constraint Eq. (27). Note that the probabilities for each origin and destination node or zone are independent from one another as there is no constraint tying them together as in the classic spatial interaction model where distance or travel cost is intrinsic to the specification.

These power laws can be related to recent explorations in network science which suggest that the number of in-degrees – the volume of links entering a destination in our terms – and the number of out-degrees – the volume emanating from an origin, both follow power laws [2]. These results have been widely observed in topological rather than planar networks where the focus is on the numbers of physical links associated with nodes rather than the volume of traffic on each link. Clearly the number of physical links in planar graphs is limited and the general finding from network science that the number of links scales as a power law cannot apply to systems that exist in two-dimensional Euclidean space [24]. However a popular way of transforming a planar graph into one which is non-planar is to invoke a rule that privileges some edges over others merging these into long links and then generating a topology which is based on the merged edges as constituting nodes and the links between the new edges as arcs. This is the method that is called space syntax [47] and it is clear that by introducing order into the network in this way, the in-degrees and out-degrees of the resulting topological graph can be scaling. Jiang [48] illustrates this quite clearly although there is some reticence to make such transformations and where planar graphs have been examined using new developments in network science based on small worlds and scale-free graph theory, the focus has been much more on deriving new network properties than on appealing to any scale-free structure [33].

However to consider the scale-free network properties of spatial interaction systems, each trip might be considered a physical link in and of itself, albeit that it represents an interaction on a physical network as a person making such an interaction is distinct in space and time. Thus the connections to network science are close. In fact the study of networks and their scaling properties has not followed the static formulations which dominate our study of cities in equilibrium for the main way in which such power laws are derived for topological networks is through a process of preferential attachment which grows networks from a small number of seed nodes [8]. Nevertheless, such dynamics appear quite consistent with the evolution of spatial interaction systems.

These models will be introduced a little later when urban dynamics are being dealt with. For the moment, let us note that there are various simple dynamics which can account not only for the distribution of network links following power laws, but also for the distribution of city sizes, incomes, and a variety of other social (and physical) phenomena from models that grow the number of objects according to simple proportionate growth consistent with the generation of lognormal distributions. Suffice it to say that although we have focused on urban densities as following either power laws or negative exponential functions in this section, it is entirely possible to use the entropy-maximizing framework to generate distributions which are log-normal, another alternative with a strong spatial logic. Most distributions which characterize urban structure and activities however are not likely to be normal and to conclude this section, we will review albeit very briefly, some empirical results that indicate the form and pattern of urban activities in western cities.

Empirical Applications: Rank-Size Representations of Urban Distributions

The model in Eq. (26) gives the probability of location in a zone i as an inverse power function of the population or size of that place which is also proportional to the frequency

$$f(p_i) \propto p_i = KP_i^{-\varphi} . \tag{31}$$

It is possible to estimate the scaling parameter φ in many different ways but a first test of whether or not a power law is likely to exist can be made by plotting the logarithms of the frequencies and population sizes and noting whether or not they fall onto a straight line. In fact a much more preferable plot which enables each individual observation to be represented is the cumulative function which is formed from the integral of Eq. (31) up to a given size; that is $F_i \propto P_i^{-\varphi+1}$. The counter-cumulative $F - F_i$ where F is the sum of all frequencies in the system – that is the number of events or cities – also varies as $P_i^{-\varphi+1}$ and is in fact the rank of the city in question. Assuming each population size is different, then the order of $\{i\}$ is the reverse of the rank, and we can now write the rank r of i as $r = F - F_i$. The equation for this rank-size distribution (which is the one that is usually used to fit the data) is thus

$$r = GP_r^{-\varphi+1} \tag{32}$$

where G is a scaling constant which in logarithmic form is $\ln r = G - (\varphi - 1) \ln P_r$. This is the equation that is im-

plicit in the rank-size plots presented below which reveal evidence of scaling.

First let us examine the scaling which is implicit in urban size distributions for the largest world city populations over 1 million in 2005, for cities over 100,000 in the USA in 2000, and for the 200 tallest buildings in the world in 2007. We could repeat such examples ad nauseum but these provide a good selection which we graph in rank-size logarithmic form in Fig. 1a, noting that we have normalized all the data by their means, that is by $<P_r>$ and $<r>$, as $P_r/<P_r>$ and $r/<r>$. We are only examining a very small number at the very top of the distribution and this is clearly not definitive evidence of scaling in the rest of the distribution but these plots do show the typical distributions of city size activities that have been observed in this field for over 50 years. As we will imply later, these signatures are evidence of self-organization and fractal structure which emerge through competition from the bottom up [15].

To illustrate densities in cities, we take employment and working population in small zones in Greater London, a city which has some 4.4 million workers. We rank-order the distribution in the same way we have done for world cities, and plot these, suitably normalized by their means, logarithmically in Fig. 1b. These distributions are in fact plotted as densities so that we remove aerial size effects. Employment densities $e_i = E_i/A_i$ can be interpreted as the number of work trips originating in employment zones $e_i = \sum_j T_{ij}$ – the volume of the out-degrees of each employment zone considered as nodes in the graph of all linkages between all places in the system, and population densities $h_j = P_j/A_j$ as the destination distributions $h_j = \sum_i T_{ij}$ – the in-degrees which measure all the trips destined for each residential zone from all employment zones. In short if there is linearity in the plots, this is evidence that the underlying interactions on the physical networks that link these zones are scaling. Figure 1b provides some evidence of scaling but the distributions are more similar to lognormal distributions than to power laws. This probably implies that the mechanisms for generating these distributions are considerably more complex than growth through preferential attachment which we will examine in more detail below [15].

Lastly, we can demonstrate that scaling in city systems also exists with respect to how trips, employment and population activities vary with respect to distance. In Fig. 1c, we have again plotted the employment densities $e_i = E_i/A_i$ at origin locations and population densities $h_j = P_j/A_j$ at destination locations but this time against distances $d_{\mathrm{CBD}\to i}$ and $d_{\mathrm{CBD}\to j}$ from the center of London's CBD in logarithmic terms. It is clear that there is significant correlation but also a very wide spread of values around the log-linear regression lines due to the fact that the city is multi-centric. Nevertheless the relationships appears to be scaling with these estimated as $e_i = 0.042 d_{\mathrm{CBD}\to i}^{-0.98}$, ($r^2 = -0.30$), and $h_j = 0.029 d_{\mathrm{CBD}\to j}^{-0.53}$,

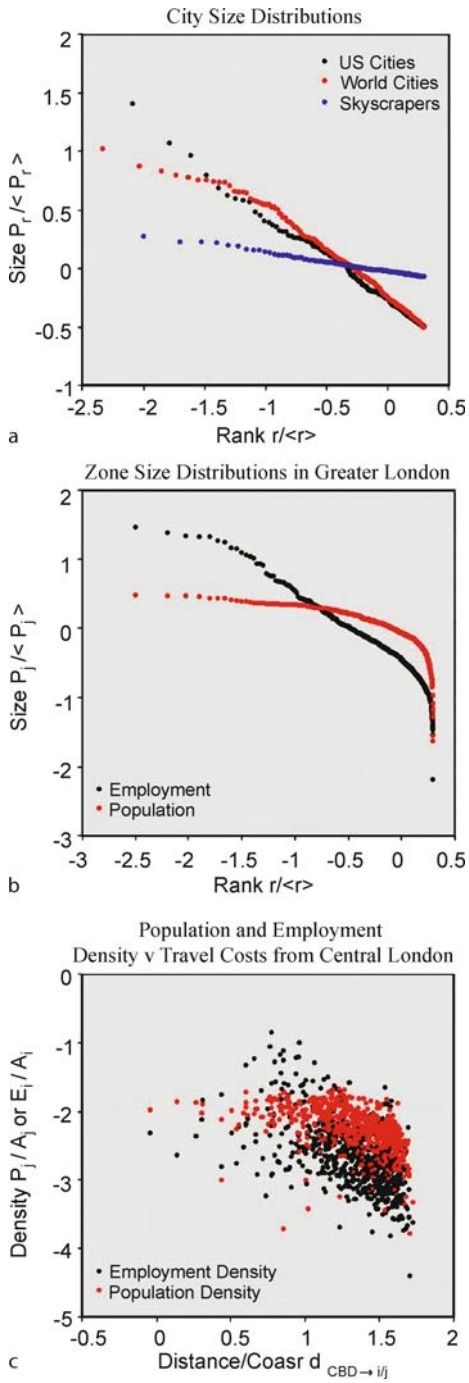

Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 1
Scaling distributions in world cities and in Greater London

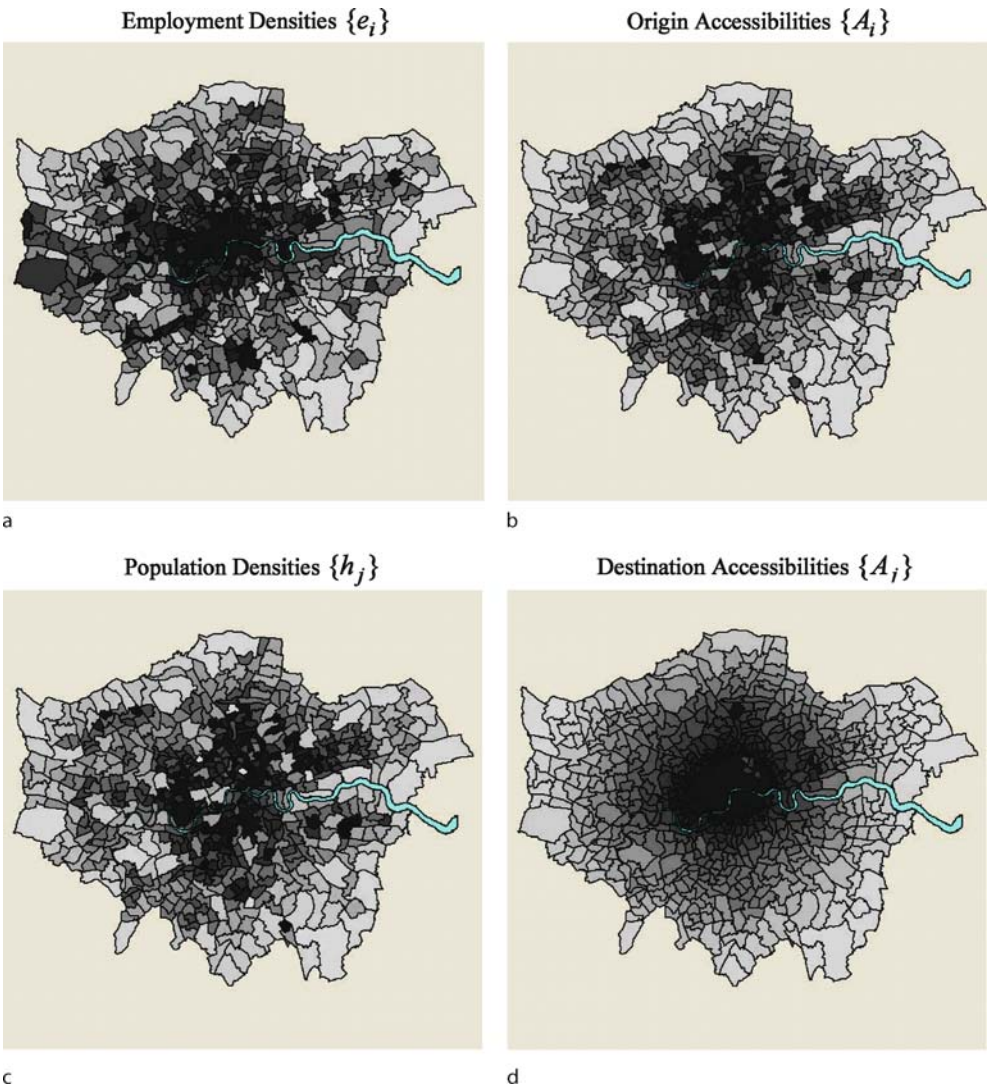

Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 2
Employment, population and accessibilities in Greater London (greatest extent is 55kms east to west; 45kms north to south)

($r^2 = -0.23$). However more structured spatial relationships can be measured by accessibilities which provide indices of overall proximity to origins or destinations, thus taking account of the fact that there are several competing centers. Accessibility can be measured in many different ways but here we use a traditional definition of potential based on employment accessibility A_i to populations at destinations, and population accessibility A_j to employment at origins defined as

$$\left.\begin{aligned} A_i &\propto \sum_j \frac{h_j}{c_{ij}} \\ A_j &\propto \sum_i \frac{e_i}{c_{ij}} \end{aligned}\right\}, \tag{33}$$

where c_{ij} is, as before, the generalized cost of travel from employment origin i to population destination j. In Fig. 2a, b, we compare the distribution of employment densities e_i with accessibility origins A_i and in Fig. 2c, d, population densities h_j with accessibility destinations A_j. Each set of maps is clearly correlated with higher associations than in Fig. 1c which take account of only the single CBD. Regressing $\{\ln e_i\}$ on $\{\ln A_i\}$ and $\{\ln p_j\}$ on $\{\ln A_j\}$ gives an approximate scaling with 31% of the variance accounted for in terms of origin accessibility and 41% for destination accessibility. These relations appear linear but there is still considerable noise in the data which undoubtedly reflects the relative simplicity of the models and

the fact that accessibility is being measured using current transport costs without any reference to the historical evolution of the city's structure. It is, however, building blocks such as these that constitute the basis for operational land use transport models that have developed for comparative static and quasi-dynamic forecasting that we will discuss below.

Urban Dynamics

Aggregate Development

Models of city systems have largely been treated as static for at first sight, urban structure in terms of its form and to some extent its function appears stable and long-lasting. During the industrial era, cities appeared to have a well-defined structure where land uses were arranged in concentric rings according to their productivity and wealth around a central focus, usually the central business district (CBD), the point where most cities were originally located and exchange took place. Moreover data on how cities had evolved were largely absent and this reinforced the focus on statics and equilibria. Where the need to examine urban change was urgent, models were largely fashioned in terms of the simplest growth dynamics possible and we will begin with these here.

The growth of human populations in their aggregate appears to follow an exponential law where the rate of change σ is proportional to the size of the population itself $P(t)$, that is

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = \sigma P(t) \,. \tag{34}$$

It is easy to show that starting from an initial population $P(0)$, the growth is exponential, that is

$$P(t) = P(0)\exp(\sigma t) \,, \tag{35}$$

which is the continuous form of model. When formulated discretely, at time steps $t = 1, 2, \ldots, T$, Eq. (34) can be written as $P(t) - P(t-1) = \beta P(t-1)$ which leads to

$$P(t) = (1+\beta)P(t-1) \,. \tag{36}$$

Through time from the initial condition $P(0)$, the trajectory is

$$P(t) = (1+\beta)^t P(0) \,. \tag{37}$$

$1+\beta$ is the growth rate. If $\beta > 0$, Eq. (37) shows exponential growth, if $\beta < 0$, exponential decline, and if $\beta = 0$, the population is in the steady state and simply reproduces itself.

This simple growth model leads to *smooth change*, and any discontinuities or breaks in the trajectories of growth or decline must come about through an external change in the rate from the outside environment. If we assume the growth rate fluctuates around a mean of one with β varying randomly, above -1, then it is not possible to predict the trajectory of the growth path. However if we have a large number of objects which we will assume to be cities whose growth rates are chosen randomly, then we can write the growth equation for each city as

$$P_i(t) = [1+\beta_i(t)]P_i(t-1) \,, \tag{38}$$

which from an initial condition $P_i(0)$ gives

$$P_i(t) = \prod_{\tau=1}^{t} [1+\beta_i(\tau)]P_i(0) \,. \tag{39}$$

This is growth by proportionate effect; that is, each city grows in proportion to its current size but the growth rate in each time period is random. In a large system of cities, the ultimate distribution of these population sizes will be lognormal. This is easy to demonstrate for the logarithm of Eq. (39) can be approximated by

$$\ln P_i(t) = \ln P_i(0) + \sum_{\tau=1}^{t} \beta_i(\tau) \,, \tag{40}$$

where the sum of the random components is an approximation to the log of the product term in Eq. (39) using Taylor's expansion. This converges to the lognormal from the law of large numbers. It was first demonstrated by Gibrat [43] for social systems but is of considerable interest here in that the fat tail of the lognormal can be approximated by an inverse power law. This has become the default dynamic model which underpins an explanation of the rank-size rule for city populations first popularized by Zipf [81] and more recently confirmed by Gabaix [41] and Blank and Solomon [23] among others. We demonstrated this in Fig. 1a for the world city populations greater than 1 million and for US city populations greater than 100,000. As such, it is the null hypothesis for the distribution of urban populations in individual cities as well as population locations within cities.

Although Gibrat's model does not take account of interactions between the cities, it does introduce diversity into the picture, simulating a system that in the aggregate is non-smooth but nevertheless displays regularity. These links to aggregate dynamics focus on introducing slightly more realistic constraints and one that is of wide relevance is the introduction of capacity constraints or limits on the

level to which a population might grow. Such capacitated growth is usually referred to as logistic growth. Retaining the exponential growth model, we can limit this by moderating the growth rate σ according to an upper limit on population $P_{\max}$ which changes the model in Eq. (34) and the growth rate σ to

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = \left[\sigma\left(1 - \frac{P(t)}{P_{\max}}\right)\right] P(t) \,. \tag{41}$$

It is clear that when $P(t) = P_{\max}$, the overall rate of change is zero and no further change occurs. The continuous version of this logistic is

$$P(t) = \frac{P_{\max}}{1 + \left(\frac{P_{\max}}{P(0)} - 1\right)\exp(-\sigma t)} \,, \tag{42}$$

where it is easy to see that as $t \to \infty$, $P(t) \to P_{\max}$.

The discrete equivalent of this model in Eq. (41) follows directly from $P(t) - P(t-1) = \beta[1 - (P(t-1)/P_{\max})]P(t-1)$ as

$$P(t) = \left[1 + \beta\left(1 - \frac{P(t-1)}{P_{\max}}\right)\right] P(t-1) \,, \tag{43}$$

where the long term dynamics is too intricate to write out as a series. Equation (43) however shows that the growth component β is successively influenced by the growth of the population so far, thus preserving the capacity limit through the simple expedient of adjusting the growth rate downwards. As in all exponential models, it is based on proportionate growth. As we noted above, we can make each city subject to a random growth component $\beta_i(t)$ while still keeping the proportionate effect.

$$P(t) = \left[1 + \beta_i(t)\left(1 - \frac{P(t-1)}{P_{\max}}\right)\right] P(t-1) \,. \tag{44}$$

This model has not been tested in any detail but if $\beta_i(t)$ is selected randomly, the model is a likely to generate a lognormal-like distribution of cities but with upper limits being invoked for some of these. In fact, this stochastic equivalent also requires a lower integer bound on the size of cities so that cities do not become too small [14]. Within these limits as long as the upper limits are not too tight, the sorts of distributions of cities that we observe in the real world are predictable.

In the case of the logistic model, remarkable and unusual discontinuous nonlinear behavior can result from its simple dynamics. When the β component of the growth rate is $\beta < 2$, the predicted growth trajectory is the typical logistic which increases at an increasing rate until an inflection point after which the growth begins to slow, eventually converging to the upper capacity limit of $P_{\max}$. However when $\beta \cong 2$, the population oscillates around this limit, bifurcating between two values. As the value of the growth rate increases towards 2.57, these oscillations get greater, the bifurcations doubling in a regular but rapidly increasing manner. At the point where $\beta \cong 2.57$, the oscillations and bifurcations become infinite, apparently random, and this regime persists until $\beta \cong 3$ during which the predictions look entirely chaotic. In fact, this is the regime of 'chaos' but chaos in a controlled manner from a deterministic model which is not governed by externally induced or observed randomness or noise.

These findings were found independently by May [52], Feigenbaum [38], Mandelbot [51] among others. They relate strongly to bifurcation and chaos theory and to fractal geometry but they still tend to be of theoretical importance only. Growth rates of this magnitude are rare in human systems although there is some suggestion that they might occur in more complex coupled biological systems of predator-prey relations. In fact one of the key issues in simulating urban systems using this kind of dynamics is that although these models are important theoretical constructs in defining the scope of the dynamics that define city systems, much of these dynamic behaviors are simplistic. In so far as they do characterize urban systems, it is at the highly aggregate scale as we demonstrate a little later. The use of these ideas in fact is much more applicable to extending the static equilibrium models of the last section and to demonstrate these, we will now illustrate how these models might be enriched by putting together logistic behaviors with spatial movement and interaction.

One way of articulating urban dynamics at the intra-urban level is to identify different speeds of change. In particular we can define a fast dynamics that relates to how people might move around the city on daily basis, for example, in terms of the journey to work, and a slower dynamics that relates to more gradual change that relates to the size of different locations affected by residential migrations. We can model the fast dynamics using a singly-constrained spatial interaction which distributes workers to residential locations which we define using previous notation where all variables are now time scripted by (t): $T_{ij}(t)$ trips between zones i and j, employment $E_i(t)$ at origin zone i, population $P_j(t)$ at destination zone j, the friction of distance parameter $\gamma(t)$, and the travel cost $c_{ij}(t)$ between zones i and j. The model is defined as

$$T_{ij}(t) = E_i(t)\frac{P_j(t)\exp[-\gamma(t)c_{ij}(t)]}{\sum_j P_j(t)\exp[-\gamma(t)c_{ij}(t)]} \,, \tag{45}$$

from which we can predict residential population $P'_j(t)$ as

$$P'_j(t) = \sum_i T_{ij}(t) = \sum_i E_i(t) \frac{P_j(t)\exp[-\gamma(t)c_{ij}(t)]}{\sum_j P_j(t)\exp[-\gamma(t)c_{ij}(t)]} . \quad (46)$$

This is the fast dynamics but each zone is capacitated by an unchanging upper limit on population where the zonal population changes slowly in proportion to its existing size through internal migration and in response to the upper limit $P_{j\max}$. The change in terms of this slower dynamic from t to $t+1$ is modeled as

$$\Delta P_j(t+1) = \beta[P_{j\max} - P'_j(t)]P'_j(t) \quad (47)$$

with the long term trajectory thus given as

$$P_j(t+1) = \left(1 + \beta[P_{j\max} - P'_j(t)]\right) P'_j(t) . \quad (48)$$

Clearly $P_j(t)$ will converge to $P_{j\max}$ as long as $P'_j(t)$ is increasing while the fast dynamics is also updated in each successive time period from

$$P'_j(t+1) = \sum_i T_{ij}(t+1) = \sum_i E_i(t+1) \frac{P_j(t+1)\exp[-\gamma(t+1)c_{ij}(t+1)]}{\sum_j P_j(t+1)\exp[-\gamma(t+1)c_{ij}(t+1)]} . \quad (49)$$

We may have an even slower dynamics relating to technological or other social change which changes $P_{j\max}$ while various other models may be used to predict employment for example, which itself may be a function of another fast dynamics relating to industrial and commercial interactions. The time subscripted variables travel $c_{ij}(t+1)$ and the friction of distance parameter $\gamma(t+1)$ might be changes that reflect other time scales. We might even have lagged variables independently introduced reflecting stocks or flows at previous time periods $t-1$, $t-2$ etc. Wilson [75,76] has explored links between these spatial interaction entropy-maximizing models and logistic growth and has shown that in a system of cities or zones within a city, unusual bifurcating behavior in terms of the emergence of different zonal centers can occur when parameter values, particularly the travel cost parameter $\gamma(t+1)$, cross certain thresholds.

There have been many proposals involving dynamic models of city systems which build on the style of nonlinear dynamics introduced here and these all have the potential to generate discontinuous behavior. Although Wilson [75] pioneered embedding dynamic logistic change into spatial interaction models, there have been important extensions to urban predator-prey models by Dendrinos and Mullaly [36] and to bifurcating urban systems by Allen [3,4], all set within a wider dynamics linking macro to micro through master equation approaches [45]. A good summary is given by Nijkamp and Reggiani [57] but most of these have not really led to extensive empirical applications for it has been difficult to find the necessary rich dynamics in the sparse temporal data sets available for cities and city systems; at the macro-level, a lot of this dynamics tends to be smoothed away in any case. In fact, more practical approaches to urban dynamics have emerged at finer scale levels where the agents and activities are more disaggregated and where there is a stronger relationship to spatial behavior. We will turn to these now.

Dynamic Disaggregation: Agents and Cells

Static models of the spatial interaction variety have been assembled into linked sets of sub-models, disaggregated into detailed types of activity, and structured so that they simulate changes in activities through time. However, the dynamics that is implied in such models is simplistic in that the focus has still been very much on location in space with time added as an afterthought. Temporal processes are rarely to the forefront in such models and it is not surprising that a more flexible dynamics is emerging from entirely different considerations. In fact, the models of this section come from dealing with objects and individuals at much lower/finer spatial scales and simulating processes which engage them in decisions affecting their spatial behavior. The fact that such decisions take place through time (and space) makes them temporal and dynamic rather through the imposition of any predetermined dynamic structures such as those used in the aggregate dynamic models above. The models here deal with individuals as agents, rooted in cells which define the space they occupy and in this sense, are highly disaggregate as well as dynamic. These models generate development in cities from the bottom up and have the capability of producing patterns which are emergent. Unlike the dynamic models of the last section, their long term spatial behavior can be surprising and often hard to anticipate.

It is possible however to use the established notation for equilibrium models in developing this framework based on the generic dynamic $P_i(t) = P_i(t-1) + \Delta P_i(t)$ where the change in population $\Delta P_i(t)$ can be divided into two components. The first is the usual proportionate effect, the positive feedback induced by population on itself which is defined as the *reactive* element of change $\omega P_i(t-1)$. The second is the *interactive* element, change

that is generated from some action-at-a-distance which is often regarded as a diffusion of population from other locations in the system. We can model this in the simplest way using the traditional gravity model in Eq. (24) but noting that we must sum the effects of the diffusion over the destinations from where it is generated as a kind of accessibility or potential. The second component of change is $\phi P_i(t-1)K\sum_j P_j(t-1)/c_{ij}^{\eta}$ from which the total change between t and $t-1$ is

$$\Delta P_i(t) = \omega P_i(t-1) + \phi P_i(t-1) K \sum_j \frac{P_j(t-1)}{c_{ij}^{\eta}} + \varepsilon_i(t-1). \tag{50}$$

We have also added a random component $\varepsilon_i(t-1)$ in the spirit of our previous discussion concerning growth rates. We can now write the basic reaction-diffusion equation, as it is sometimes called, as

$$\begin{aligned} &P_i(t) \\ &= P_i(t-1) + \Delta P_i(t) \\ &= P_i(t-1)\left(1 + \omega + \phi K \sum_j \frac{P_j(t-1)}{c_{ij}^{\eta}} + \varepsilon_i(t-1)\right). \end{aligned} \tag{51}$$

This equation looks as though it applies to a zonal system but we can consider each index i or j simply a marker of location, and each population activity can take on any value; for single individuals it can be 0 or 1 while it might represent proportions of an aggregate population or total numbers for the framework is entirely generic. As such, it is more likely to mirror a slow dynamics of development rather than a fast dynamics of movement although movement is implicit through the diffusive accessibility term.

We will therefore assume that the cells are small enough, space-wise, to contain single activities – a single household or land use which is the cell state – with the cellular tessellation usually forming a grid associated with the pixel map used to visualize data input and model output. In terms of our notation, population in any cell i must be $P_i(t) = 1$ or 0, representing a cell which is occupied or empty with the change being $\Delta P_i(t) = -1$ or 0 if $P_i(t-1) = 1$ and $\Delta P_i(t) = 1$ or 0 if $P_i(t-1) = 0$. These switches of state are not computed by Eq. (51) for the way these cellular variants are operationalized is through a series of rules, constraints and thresholds. Although consistent with the generic model equations, these are applied in more ad hoc terms. Thus these models are often referred to as automata and in this case, as cellular automata (CA).

The next simplification which determines whether or not a CA follows a strict formalism, relates to the space over which the diffusion takes place. In the fast dynamic equilibrium models of the last section and the slower ones of this, interaction is usually possible across the entire space but in strict CA, diffusion is over a local neighborhood of cells around i, Ω_i, where the cells are adjacent. For symmetric neighborhoods, the simplest is composed of cells which are north, south, east and west of the cell in question, that is $\Omega_i = n, s, e, w$ – the so-called von Neumann neighborhood, while if the diagonal nearest neighbors are included, then the number of adjacent cells rises to 8 forming the so-called Moore neighborhood. These highly localized neighborhoods are essential to processes that grow from the bottom up but generate global patterns that show emergence. Rules for diffusion are based on switching a cell's state on or off, dependent upon what is happening in the neighborhood, with such rules being based on counts of cells, cell attributes, constraints on what can happen in a cell, and so on.

The simplest way of showing how diffusion in localized neighborhoods takes place can be demonstrated by simplifying the diffusion term in Eq. (50) as follows. Then $\phi K\sum_j P_j(t-1) = \phi K\sum_j P_j(t-1)c_{ij}^{-\eta}$ as $c_{ij} = 1$ when $\Omega_i = n, s, e, w$. The cost is set as a constant value as each cell is assumed to be small enough to incur the same (or no) cost of transport between adjacent cells. Thus the diffusion is a count of cells in the neighborhood i. The overall growth rate is scaled by the size of the activity in i but this activity is always either $P_i(t-1) = 1$ or 0, presence or absence. In fact this scaling is inappropriate in models that work by switching cells on and off for it is only relevant when one is dealing with aggregates. This arises from the way the generic equation in (51) has been derived and in CA models, it is assumed to be neutral. Thus the change Eq. (50) becomes

$$\Delta P_i(t) = \omega + \phi K \sum_j P_j(t-1) + \varepsilon_i(t)\,, \tag{52}$$

where this can now be used to determine a threshold $Z_{\max}$ over which the cell state is switched. A typical rule might be

$$P_i(t) = \begin{cases} 1, & \text{if } [\omega + \phi K \sum_j P_j(t-1) + \varepsilon_i(t)] > Z_{\max} \\ 0, & \text{otherwise}\,. \end{cases} \tag{53}$$

It is entirely possible to separate the reaction from the diffusion and consider different combinations of these effects

sparking off a state change. As we have implied, different combinations of attributes in cells and constraints within neighborhoods can be used to effect a switch, much depending on the precise specification of the model.

In many growth models based on CA, the strict limits posed by a local neighborhood are relaxed. In short, the diffusion field is no longer local but is an information or potential field consistent with its use in social physics where action-at-distance is assumed to be all important. In the case of strict CA, it is assumed that there is no action-at-a-distance in that diffusion only takes place to physically adjacent cells. Over time, activity can reach all parts of the system but it cannot hop over the basic cell unit. In cities, this is clearly quite unrealistic as the feasibility of deciding what and where to locate does not depend on physical adjacency. In terms of applications, there are few if any urban growth models based on strict CA although this does rather beg the question as to why CA is being used in the first place. In fact it is more appropriate to call such models cell-space or CS models as Couclelis [32] has suggested. In another sense, this framework can be considered as one for agent-based modeling where the cells are not agents and where there is no assumption of a regular underlying grid of cells [12,13]. There may be such a grid but the framework simply supposes that the indices *i* and *j* refer to locations that may form a regular tessellation but alternatively may be mobile and changing. In such cases, it is often necessary to extend the notation to deal with specific relations between the underlying space and the location of each agent.

Empirical Dynamics: Population Change and City Size

We will now briefly illustrate examples of the models introduced in this section before we then examine the construction of more comprehensive models of city systems. Simple exponential growth models apply to rapidly growing populations which are nowhere near capacity limits such as entire countries or the world. In Fig. 3, we show the growth of world population from 2000 BCE to date where it is clear that the rate of growth may be faster than the exponential model implies, although probably not as fast as double exponential. In fact world population is likely to slow rapidly over the next century probably mirroring global resource limits to an extent which are clearly illustrated in the growth of the largest western cities. In Figs. 4a, b, we show the growth in population of New York City (the five boroughs) and Greater London from 1750 to date and it is clear that in both cases, as the cities developed, population grew exponentially only to slow as the upper density limits of each city were reached.

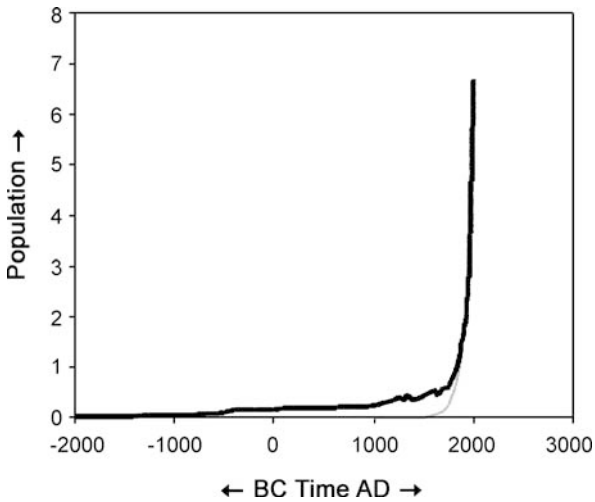

Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 3
Exponential world population growth. The *fitted exponential curve* is shown in *grey* where for the most part it is coincident with the observed growth, except for the very long period before the Industrial Revolution (before 1750)

Subsequent population loss and then a recent return of population to the inner and central city now dominate these two urban cores, which is reminiscent of the sorts of urban dynamic simulated by Forrester [39] where various leads and lags in the flow of populations mean that the capacity limit is often overshot, setting up a series of oscillations which damp in the limit. Forrester's model was the one of the first to grapple with the many interconnections between stocks and flows in the urban economy although these relationships were predicated hypothetically in simple proportionate feedback terms. Together they generated a rich dynamics but dominated by growth which was capacitated, thus producing logistic-like profiles with the leads and lags giving damped oscillations which we illustrate from his work in Fig. 5 [11]. We will see that the same phenomena can be generated from the bottom up as indeed Forrester's model implies, using cellular automata within a bounded spatial system.

Dynamics which arise from bottom-up urban processes can be illustrated for a typical CA/CS model, DUEM (Dynamic Urban Evolutionary Model) originally developed by Xie [78]. In the version of the model here, there are five distinct land uses – housing, manufacturing/primary industry, commerce and services, transport in the form of the street/road network, and vacant land. In principle, at each time period, each land use can generate quantities and locations of any other land use although in practice only industry, commerce and housing can gener-

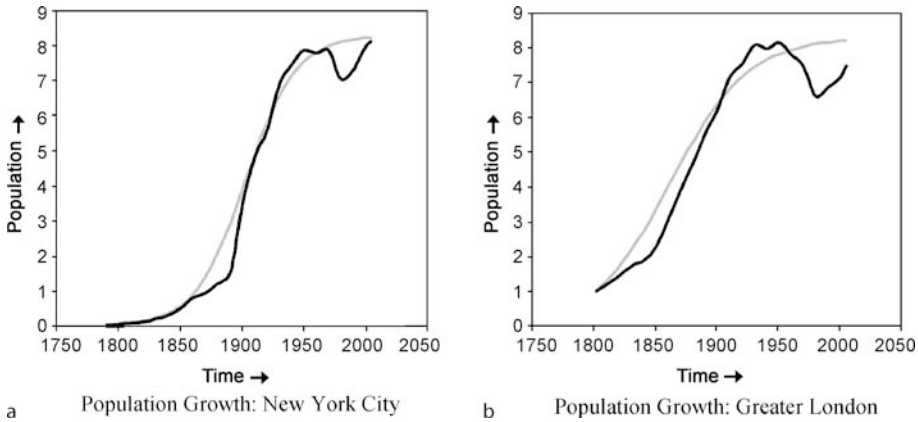

Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 4
Logistic population growth

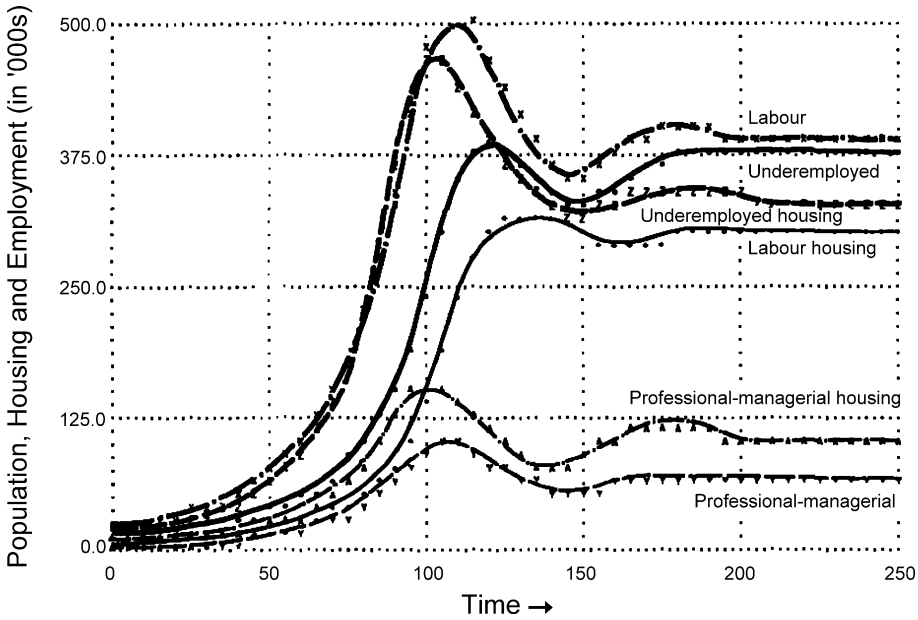

Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 5
Oscillating capacitated growth in a version of the Forrester *Urban Dynamics* model (from [11])

ate one another as well as generating streets. Streets do not generate land uses other than streets themselves. Vacant land is regarded as a residual available for development which can result from a state change (decline) in land use. The way the generation of land uses takes place is through a rule-based implementation of the generic Eq. (51) which enables a land use k, $P_i^k(t)$, to be generated from any other land use ℓ, $P_j^\ell(t-1)$. Land uses are also organized across a life cycle from initiating through mature to declining. Only initiating land uses which reflect their relative newness can spawn new land use. Mature remain passive in these terms but still influence new location while declining land uses disappear, thus reflecting completion of the life cycle of built form.

We are not able to present the fine details of the model here (see Batty, Xie and Sun [19], and Xie and Batty [79]) but we can provide a broad sketch. The way initiating land uses spawn new ones is structured according to rule-based equations akin to the thresholding implied in Eq. (53). In fact, there are three spatial scales at which these thresholds are applied ranging from the most local neighborhood through the district to the region itself. The neighborhood

exercises a trigger for new growth or decline based on the existence or otherwise of the street network, the district uses the densities of related land uses and distance of the new land use from the initiating use to effect a change, while the region is used to implement hard and fast constraints on what cells are available or not for development. Typically an initiating land use will spawn a new land use in a district only if the cells in question are vacant and if they are not affected by some regional constraint on development with these rules being implemented first. The probability of this land use occurring in a cell in this district is then fixed according to its distance from the initiating location. This probability is then modified according to the density of different land uses that exist around each of these potential locations – using compatibility constraints – and then in the local neighborhood, the density of the street network is examined. If this density is not sufficient to support a new use, the probability is set equal to zero and the cell in question does not survive this process of allocation. At this point, the cell state is switched from 'empty' or 'vacant' to 'developed' if the random number drawn is consistent with the development probability determined through this process.

Declines in land use which are simply switches from developed to vacant in terms of cell state are produced through the life cycling of activities. When a mature land use in a cell reaches a certain age, it moves into a one period declining state and then disappears at the end of this time period, the cell becoming vacant. Cells remain vacant for one time period before entering the pool of eligible locations for new development. In the model as currently constituted, there is no internal migration of activities or indeed any mutation of uses but these processes are intrinsic to the model structure and have simply not been invoked. The software for this model has been written from scratch in *Visual C++* with the loosest coupling possible to GIS through the import of raster files in different proprietary formats. The interface we have developed, shown below, enables the user to plant various land use seeds into a virgin landscape or an already developed system which is arranged on a suitably registered pixel grid which can be up to 3K x 3K or 9 million pixels in size. A map of this region forms the main window but there are also three related windows which show the various trajectories of how different land uses change through time with the map and trajectories successively updated in each run.

A feature which is largely due to the fact that the model can be run quickly through many time periods, is that the system soon grows to its upper limits with exponential growth at first which then becomes logistic or capacitated. In Fig. 6, we show how this occurs from planting a random selection of land use seeds in the region and then letting these evolve until the system fills. Because there are lags in the redevelopment of land uses in the model due to the life cycle effects, as the system fills, land is vacated. This increases the space available for new development leading to oscillations of the kind reflected in Forrester's model shown in Fig. 5 and more controversially in the real systems shown for New York City and Greater London in Fig. 4. In this sense, a CA model has a dynamics which is equivalent to that of the more top-down dynamics where growth is modeled by exponential or logistic functions. CA models however generate this as an emergent phenomena from the bottom up.

Our last demonstration of CA really does generate emergent phenomena. This is a model of residential movement that leads to extreme segregation of a population classified into two distinct groups which we will call red R and green G. Let us array the population on a square grid of dimension 51 x 51 where we place an R person next to a G person in alternate fashion, arranging them in checker board style as in Fig. 7a. The rule for being satisfied with one's locational position viz a viz one's relationship to other individuals is as follows: persons of a different group will live quite happily, side by side with each other, as long as there are as many persons of the same persuasion in their local neighborhood. The neighborhood in this instance is the eight cells that surround a person on the checkerboard in the *n, s, e, w,* and *nw, se, sw,* and *ne* positions. If however a person finds that the persons of the opposing group outnumber those of their own group, and this would occur if there were more than 4 persons of the opposite persuasion, then the person in question would change their allegiance. In other words, they would switch their support to restore their own equilibrium which ensures that they are surrounded by at least the same number of their own group. There is a version of this model that is a little more realistic in which a person would seek another location – move – if this condition were not satisfied rather than change their support, but this is clearly not possible in the completely filled system that we have assumed; we will return to this slightly more realistic model below.

In Fig. 7a, the alternative positioning shown in the checker board pattern meets this rule and the locational pattern is in 'equilibrium': that is, no one wants to change their support to another group. However let us suppose that just six persons out of a total of 2601 (51 x 51 agents sitting on the checker board) who compose about 0.01 percent of the two populations, change their allegiance. These six changes are easy to see in Fig. 7a where we assume that four R persons of the red group, change in their allegiance to support the green group, and two Gs change the op-

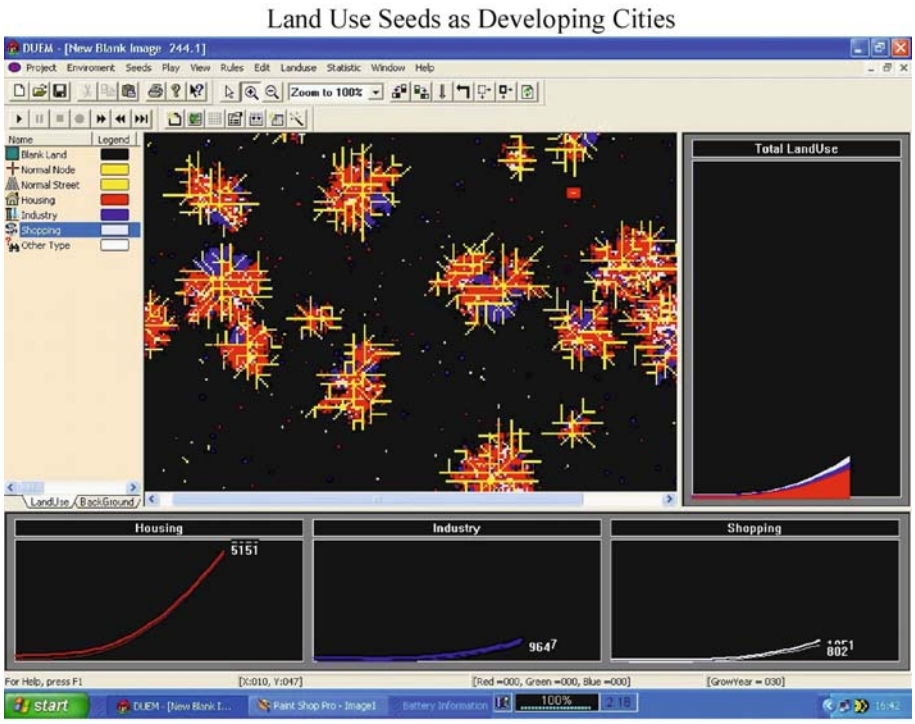

a

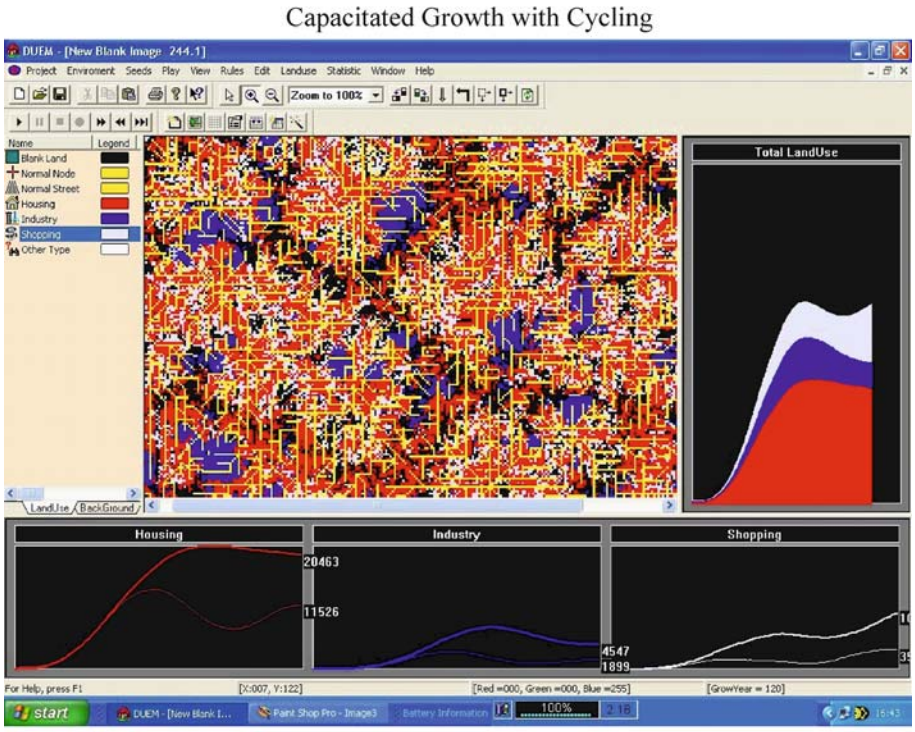

b

Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 6
Cellular growth using the *DUEM* model

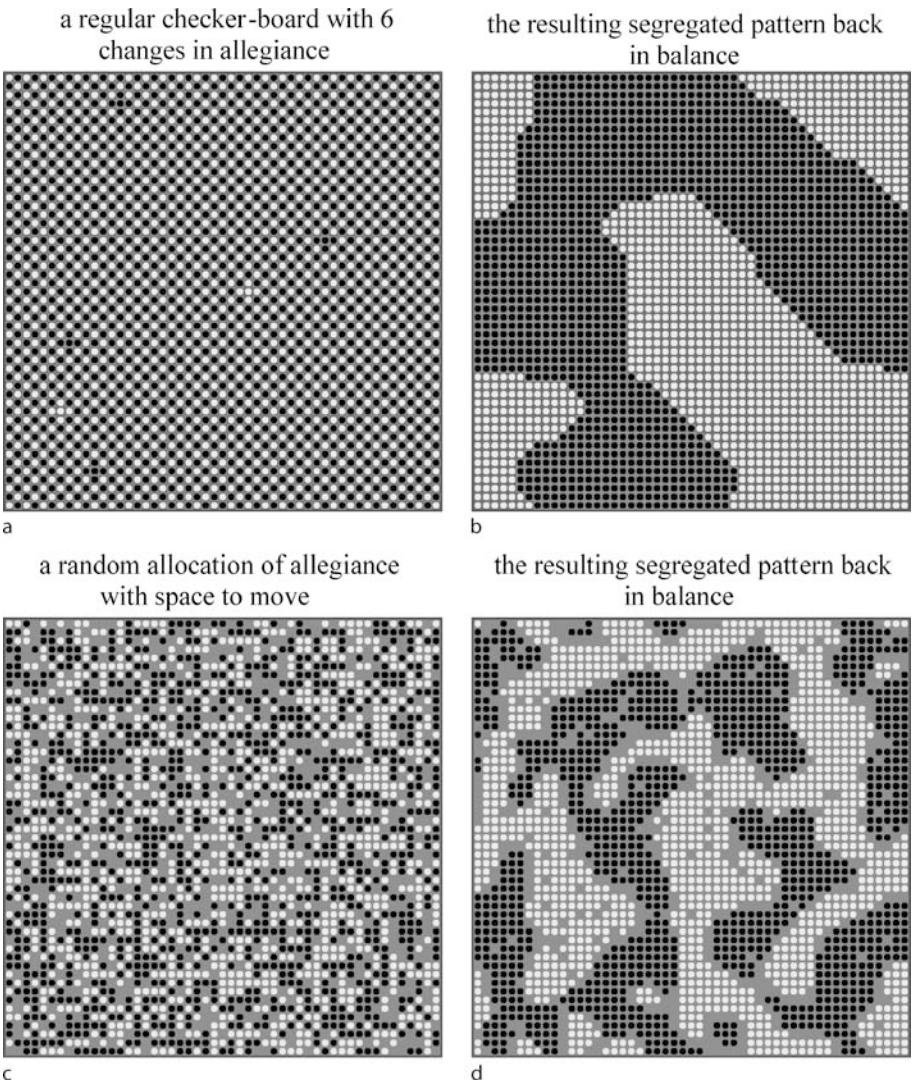

Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 7
Emergent segregation: A fragile equality (a) gives way to segregation (b); A random mix with available space (c) gives way to segregation (d)

posite way. What then happens is the equilibrium is upset in these locations but instead of being quickly restored by local changes, this sets off a mighty unraveling which quickly changes the locational complexion of the system to one where the Rs are completely and utterly segregated from the Gs. We show this in Fig. 7b. From a situation where everyone was satisfied and mixed completely, we get dramatic segregation which is a most unusual consequence. At first sight, one would never imagine that with so mild a balance of preferences, such segregation would take place. The ultimate pattern implies that Rs will live nowhere near Gs unless they really have to and there is nowhere else to live and vice versa. If an R or a G could not tolerate more than one person of a different kind living near them, then such segregation would be understandable but this is not the case: Rs are quite content to live in harmony with Gs as long as the harmony is equality.

This model was first proposed more than 30 years by Schelling [61,62]. In fact we can make this a little more realistic if we provide some free space within the system. In this case, we assume that 1/3 of the lattice is empty of persons of any kind, 1/3 composed of Rs, and 1/3 of Gs, and we mix these randomly as we show in Fig. 7c. Now the rule is slightly different in that if there are more opposition per-

sons around a person of one persuasion, then that person will try to move his or her location to a more preferential position. This sets up a process of shuffling around the checker board but as we show in Fig. 7d, quite dramatic shifts take place in location which leads to the segregation shown. This is the kind of effect that takes place in residential areas in large cities where people wish to surround themselves with neighbors of their own kind. What is surprising about the phenomena which makes it 'emergent' is that for very mild preferential bias, dramatic segregation can take place. Of course if the preferences for like neighbors are very strong anyway, then segregation will take place. But in reality, such preferences are usually mild rather than strong, yet extreme segregation takes place anyway. The conclusion is that cities often look more segregated around racial and social lines than the attitudes of their residents might suggest.

Comprehensive System Models of Urban Structure

Integrated Land Use Transport Models

The various components used to model cities in equilibrium were quickly assembled into structures that attempted to simulate urban structure and growth from the 1960s onwards. These models were referred to as land use transport models in that their aim was to simulate the locations of different land uses and their consequent patterns of traffic generation, usually according to spatial interaction principles based on gravitational assumptions. But they usually represented cities as demographic and economic activities – population, households, employment and so on – rather than as residential, commercial or industrial land use. In short the city system was seen to operate at the level of the location of activities which then consumed space through land use from which traffic was generated, and once urban activities and their interactions were predicted, appropriate translations were made into land use. As we shall see, this is not as unproblematic as was originally thought.

The integration of urban activities and their interactions – land use and transport – can be accomplished using a variety of economic frameworks built around economic relationships between activities. Traditionally these have been represented as input-output models where one activity is linked to another and it is possible to predict the chain of linkages between all the activities using multipliers. We will illustrate this for two activities: we assume that employment E is divided into an unpredictable component, sometimes considered as employment that is *basic* B and export orientated in the economy, and employment that is *non-basic* S where $E = B + S$. Non-basic employment services the population P from which it is derived as $S = bP$. If we then consider that population can be generated by applying an activity rate a to employment as $P = aE$, we have the rudiments of a generative sequence that forms a structure for predicting activities and their locations which are highly interdependent. Simple manipulation of these relationships shows that $E = B(1 - ba)^{-1}$ where $(1 - ba)^{-1}$ is the multiplier central to traditional macro-economic theory.

If we now consider that employment and population are related spatially through their interactions, we model the relationship between employment as population using a singly-constrained sub-model

$$P_j = a \sum_i T_{ij} = a \sum_i E_i \frac{F_j c_{ij}^{-\psi}}{\sum_k F_k c_{ik}^{-\psi}} , \tag{54}$$

where T_{ij} are work trips between i and j, F_j is some measure of attraction at residential location j, and ψ is the friction of distance/travel cost parameter. Employment is modeled in reverse direction as

$$E_i = b \sum_j S_{ji} = b \sum_j P_j \frac{F_i c_{ji}^{-\varpi}}{\sum_k F_k c_{ki}^{-\varpi}} , \tag{55}$$

where S_{ji} are employment demands in j from i, F_i is some measure of attraction at residential location i, and ϖ is the friction of distance/travel cost parameter. These two equations for the two sectors are not usually solved simultaneously but the chain is broken in that we start with basic employment B in Eq. (54), predicting basic population, then using this basic population in Eq. (55) to produce an increment of non-basic employment which in turn is used to predict the next increment of non-basic population in Eq. (54). This iteration converges to the multiplier relationships $E = B(1 - ba)^{-1}$ and $P = bB(1 - ba)^{-1}$.

This kind of sequence can be disaggregated indefinitely with respect to population and employment types and linked demands to other sectors. Education, leisure and so on can be added to the framework making the model ever more comprehensive. This was the model first developed by Lowry [50]. It is still the most widely applied of all operational urban models and has been elaborated in various ways, some of them dealing with partial dynamics [11]. Their theoretical pedigree is rooted largely in regional economics, location theory and the new urban economics which represent the spatial equivalents of classical macro and micro economics. The most coherent recent statement in this vein is based on applications of trade theory to the urban economy as reflected in the work of Fujita,

Krugman and Venables [40] but there is a long heritage of empirical models in the Lowry [50] tradition which continue to be built [71].

These models now incorporate the four-stage transportation modeling process of trip generation, distribution, modal split and assignment explicitly and they are consistent with discrete choice methods based on utility maximizing in their simulation of trip-making [20]. They have been slowly adapted to simulate dynamic change although they still tend to generate the entire activity pattern of the city in one go, and they remain parsimonious in that the assumption is that all the outcomes from the model can be tested in terms of their goodness of fit. They have also become more disaggregate and there are now links to physical land use although they still remain at the level of activity allocation despite their nomenclature as land use transport models. In short, this class of models is the most operational in that newer styles tend to be less comprehensive in their treatment of urban activities and transportation. Probably the most highly developed of these models currently is the UrbanSim model [69] although the MEPLAN, TRANUS and IRPUD models, whose most recent versions were developed in the EU Propolis [60] project, also represent the state-of-the-art.

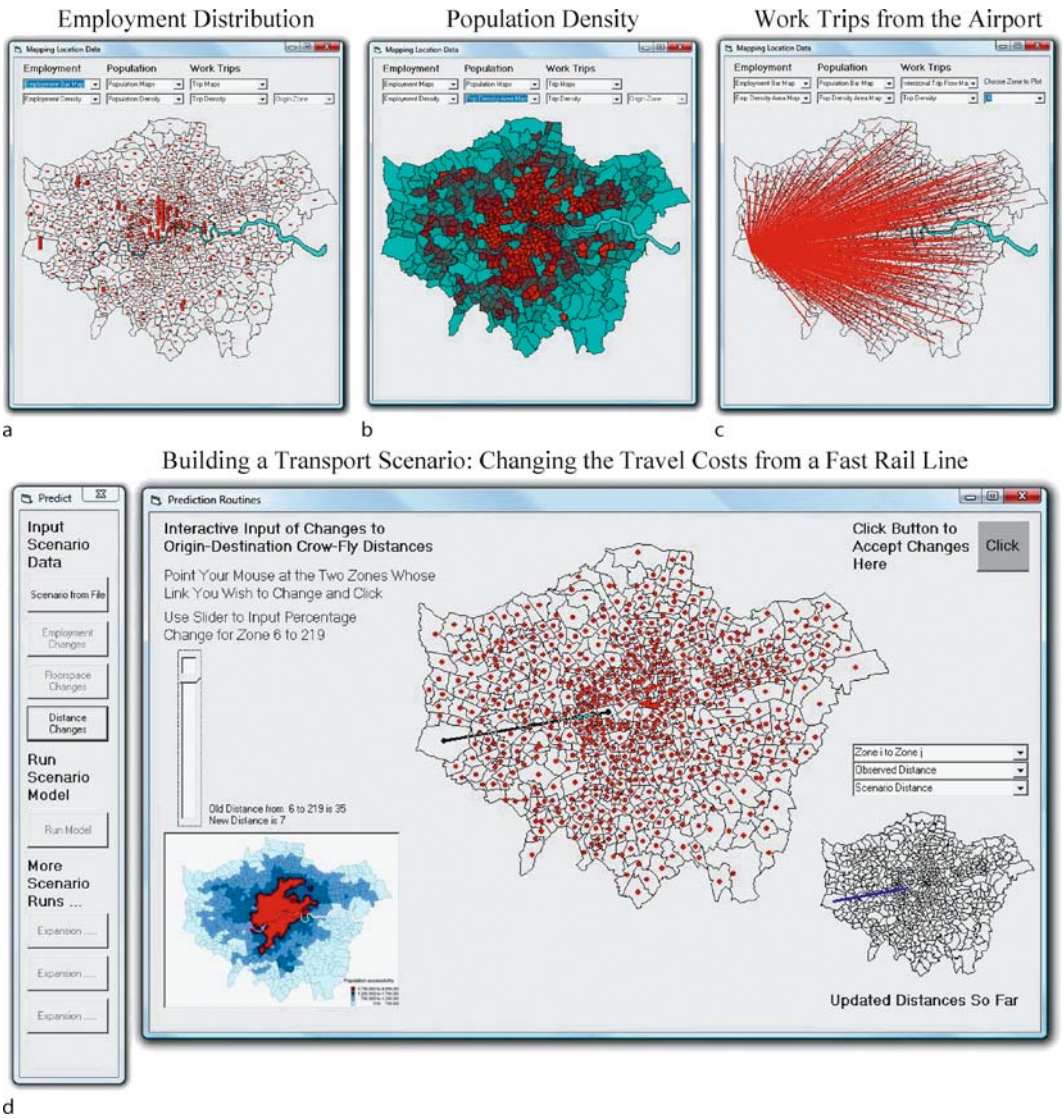

Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 8
Visualization of outputs from a Greater London land use transport model

To conclude this section, it is worth showing a visualization from one of these land use transport models which we have recently built for the London region as part of an integrated assessment of climate change in the metropolis. The component we show is a residential location model which predicts the flow of workers from employment locations to residential areas using four different modes of transport and disaggregated into five employment and five household types. In Fig. 8a, we show some outputs from the model – the observed employment distribution, the pattern of population density, and total work trips from the airport (Heathrow) zone in the base year simulation 2005. This kind of model assumes that employment and the travel cost network are exogenously determined and thus 'what-if' style questions can be thrown at the model to be evaluated in terms of the impact of changes in the transport network and employment volumes on the location of population. We illustrate such a scenario builder for changes in the transport routes and costs in Fig. 8b which provides some sense of how such complexity can be visualized. These are key issues in planning policy for the future growth of London, particularly with respect to flooding in the Thames Estuary which is likely to be affected by climate change. These kinds of models are hardly routine but they are being developed now in many places.

Agent-Based and Cellular Automata Models of Land Development

The first bottom-up CA models applicable to urban structure and growth can be traced back to the 1960s. Chapin and Weiss [27] used cell-space (CS) simulation whose locational attractions were based on linear regression, in their models of urban growth in Greensboro, North Carolina. Lathrop and Hamburg [49] used gravitational models to effect the same in simulating growth in the Buffalo-Niagara region while from a rather different perspective, Tobler [65] used CA-like simulation to generate a movie of growth in the Detroit region. All these applications were on the edge of the mainstream which 30 years ago was based not on formal dynamics but on cross-sectional equilibrium models of the variety presented above. In the intervening years, CA insofar as it was considered a simulation tool, was regarded as important mainly for its pedagogic and analytical value [32]. It was not until the early 1990s that models began to emerge which were considered to be close enough to actual urban growth patterns to form the basis for simulation and prediction. In fact, there still exists a recurrent debate about whether or not CA models are more important for their pedagogic value rather than for their abilities to simulate real systems. These require gross simplifications of model processes and spatial units, sometimes rendering them further from reality than the static cross-sectional models that came before.

The three earliest attempts at such modeling were geared to simulating rapid urban growth for metropolitan regions, medium-sized towns, and suburban areas. Batty and Xie [18] developed simulations of suburban residential sprawl in Amherst, New York, where a detailed space-time series of development was used to tune the model. Clarke and Gaydos [31] embarked on a series of simulations of large-scale metropolitan urban growth in the Bay Area and went on to model a series of cities in the US in the Gigalopolis project. White and Engelen [72] developed a CA model for Cincinnati from rather crude temporal land use data and in all these cases, the focus was on land development, suburbanization, and sprawl. Since then, several other groups have developed similar models focusing on suburbanization in Australian cities [70], 'desakota' – rapid urban growth in rural areas in China – specifically in the Pearl River Delta [80], diffused urban growth in Northern Italy [22], and rapid urbanization in Latin American cities [35]. Other attempts at modeling and predicting sprawl have been made by Papini et al. [58] for Rome and Cheng [28] for Wuhan, while Engelen's group at RIKS in the Netherlands has been responsible for many applications of their model system to various European cities [9].

There are at least four applications which do not focus on urban growth per se. Wu and Webster [77] have been intent on adding spatial economic processes and market clearing to such models, while Portugali and Benenson [59] in Tel-Aviv have focused their efforts on intra-urban change, particularly segregation and ghettoization. Semboloni [64] has worked on adding more classical mechanisms to his CA models reflecting scale and hierarchy as well as extending his simulations to the third dimension, while there have been several attempts by physicists to evolve a more general CA framework for urban development which links to new ideas in complexity such as self-organized criticality and power law scaling [7,63].

It is worth showing some graphics from such CA models as they are being applied to real cities. In Fig. 9, we show how the DUEM model can be used to simulate the pattern of development change in the Detroit region of South East Michigan. In a sense because we live in world dominated by a somewhat unhealthy interest in growth, it might be assumed that all the models we have presented here are only geared to simulating new development. In fact, each of these models can simulate decline or reproduce the steady state because CA models can solely deal with transitions and change in the existing fabric as we il-

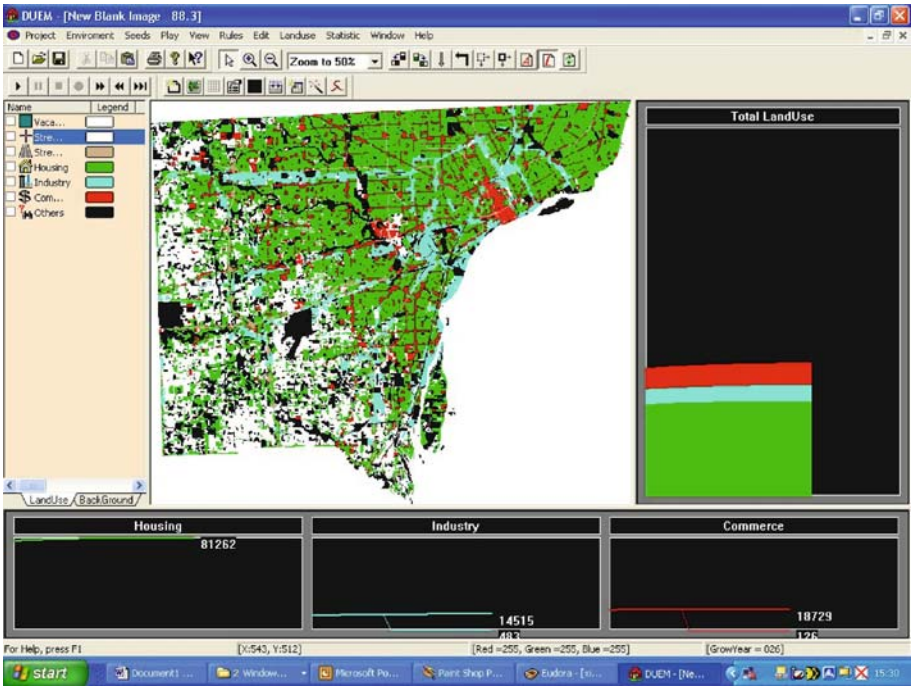

Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 9
Simulating very slow growth and rapid decline in the Detroit region using the CA *DUEM* model

lustrated earlier in the Schelling segregation model. This is the case in Detroit where the population has rapidly adjusted and segregated its locations in the last 50 years but in a context where the overall growth has been extremely modest with many areas growing very fast in the suburbs but the central areas declining at similar rates. The profile in Fig. 9 is akin to a steady state rather than the overall exponential growth or decline shown in previous examples.

There are some agent-based models at the land use or activities level which enable predictions of future urban patterns but the main focus is at the very micro-level where local movements in terms of traffic are being simulated [25]. Several models that approach the agent-based ideal originate from other areas. TRANSIMS is a hybrid in that its roots are in agent-based simulation of vehicles but it has been scaled to embrace urban activities [55] and even UrbanSim can be interpreted through the agent paradigm. A parallel but significant approach to individualistic modeling is based on micro-simulation which essentially samples individual behavior from more aggregate distributions and constructs synthetic agent-based models linked to spatial location [30]. This is a rapidly changing field at the present time with no agreement about terminology. The term agent is being used to describe many different types of models with some focusing on unique objects ranging from cells or points in space where activities or individuals exist to models of institutions and groups with only implicit spatial positioning [44].

Models of Urban Morphology

The models introduced above do not capture many of the physical features of cities and regions in terms of their morphology. Cities are highly organized with respect to their form, displaying as we have already seen in terms of city size, clusters of activity on all scales, in short, fractals [16]. Insofar as static equilibrium models are able to reproduce this form and to an extent they are able to do so, this is largely because some of the structure of the city is input into these models through existing employment and population distributions which have already captured elements of the morphology. There are competitive effects in these models too that are intrinsic to these simulations with the dynamic models based on cellular automata closest to reflecting these processes in urban form. This is because the process of development is generated from the bottom up and agglomeration is a key feature of the processes of development that are simulated as in some of the

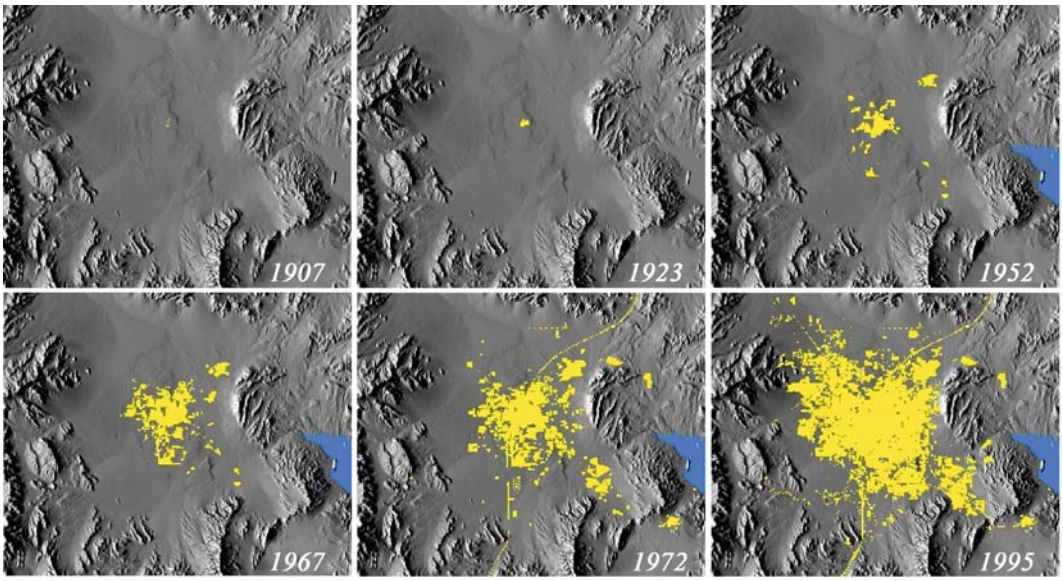

Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 10
The growth of Las Vegas from 1907 to 1995 (from [1])

models discussed in the last section. Here we will simply illustrate some of the evolving forms that various combinations of the models already discussed are able to simulate. This shows how various processes of land development and travel behavior can come together to generate structures that are close to what we observe in the real world.

A good example of the urban growth which has been rapid over the last 50 years is Las Vegas, the fastest growing metropolitan area in the United States which is illustrated in Fig. 10 [1]. The sprawl does not look very different from time period to time period although it is clear that growth is clustered and these clusters tend to merge as the city grows. In this sense, the pattern always looks like more of the same from time period to time period but inside the city, things have changed rather more dramatically as the place has moved from desert oasis and staging post prior to 1950 to the entertainment and gambling capital of the US. Exponential growth of population, employment and tourism is implied by this volume of urban development mirroring the simplest 'un-capacitated' growth model in Eqs. (36) and (37). The fact that the city has grown in some directions rather than others is largely due to a combination of physical and accidental historical factors and does not imply any differences in the way growth has occurred from one time period to the next.

Cellular automata models can generate such growth where entirely local development rules are operated uniformly across the space to grow a city from a single seed. This can lead to fractal patterns, patterns that are self-similar in form with respect to scale, of the kind observed in real cities. In Fig. 11a, we show how the operation of deterministic rules where a cell is developed if there is one and only one cell already developed in its immediate neighborhood, leads to a growing structure. This is a typical example of a modular principle that preserves a certain level of density and space when development occurs but when operated routinely and exhaustively leads to cellular growth that is regular and self-similar across scales, hence fractal. In Fig. 11b, the shape of the structure generated is now circular in that development eventually occurs everywhere. The city fills up completely but the order in which this takes place is a result of development taking place at each time period with random probability. This is the effect of introducing 'noise' or 'diversity' into the model used to generate the sequence in Fig. 11a.

If urban growth is modular and scales in the simplistic way that is portrayed in these models of fractal growth, then it is not surprising that there is a tendency to explain such patterns generically, without regard to growth per se; to study these as if they represent systems with an equilibrium pattern that simply scales through time. But this is a trap that must be avoided. Dig below the surface, and examine the processes of growth and the activities that occupy these forms, disaggregate the scale and change the time interval, and this image of an implied stability changes quite radically. During the era pictured in

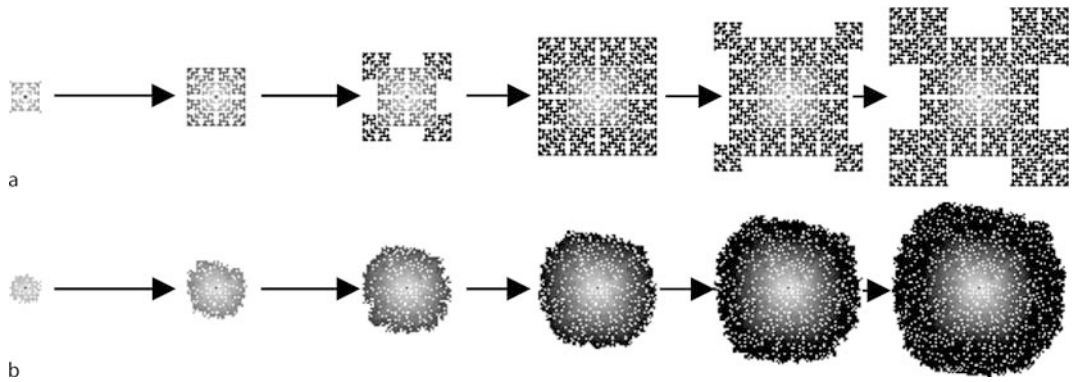

Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 11
Growth from the bottom up. a deterministic growth based on developing cells if one and only one cell is already developed in their 8 cell adjacent neighborhood, and b stochastic growth based on developing cell if any cell is developed in the adjacent neighborhood according to a random probability

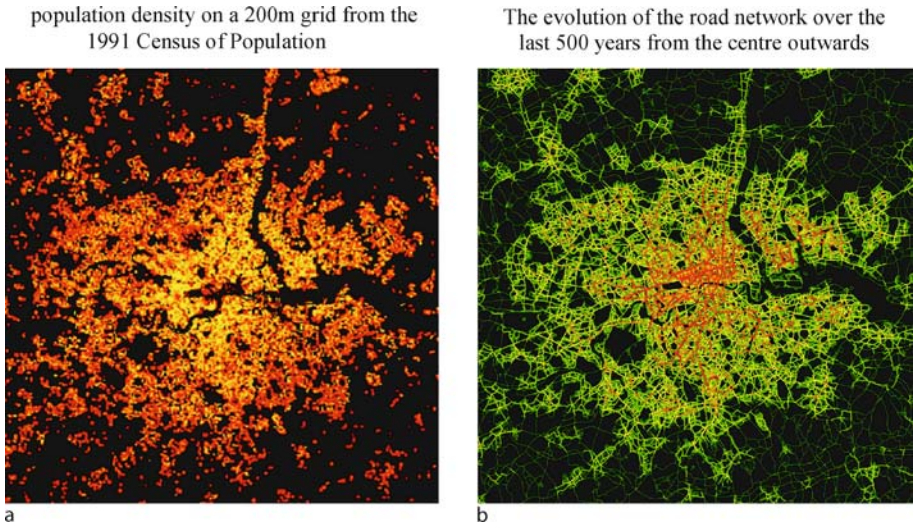

Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies, Figure 12
Greater London: self-similar clusters and the connectivity network within the sprawl

Fig. 10, technology has changed dramatically. Las Vegas did not acquire its gambling functions until the 1950s but by then it was already growing fast and the subsequent injection of cash into its local economy, the largest per capita in the western world for those who reside there, did little to change the pattern of explosive growth that followed. The manner in which people moved in the early Las Vegas was by horse and wagon but the city could only grow with the car, the plane and air-conditioning, not to say the incredible information technologies that now dictate how one gambles, wins, and loses.

Our six frame 'movie' of the growth of Las Vegas does reveal that the established pattern of adding to the periphery is not entirely the complete story for small blobs of development seem to attach themselves and then are absorbed back into the growing mass as growth catches them up. In this case, this is simply housing being constructed a little beyond the edge due to the mechanics of the development process. In older, more established settlement patterns such as those in Western Europe for example, this might be the absorption of older villages and freestanding towns into the growing sprawl. Consider the picture of population density in London recorded in 1991 and illustrated in Fig. 12a. Here there are many towns and villages that existed long before London grew to embrace them. If we define the metropolis as the connected network of set-

tlement that fills an entire space where everyone can connect to everybody else either directly or indirectly, the picture is similar as we show in Fig. 12b.

One could envisage London being connected in this way with a much sparser network of links while at the other extreme the entire space could be filled. In fact, it would seem that the level of connectivity which has evolved with respect to the density of the space filled is just enough for the city to function as a whole. It is this morphology and degree of connectivity that marks the fact that the city has reached a level of self-organization which is regarded as critical. If connectivity were greater, more space would be filled and many more connections put in place but the structure would contain a certain degree of redundancy making it inefficient. Below this, the system would not be connected at all and it would not function as a metropolis. In fact there are strong relationships to this characterization of urban settlement as a porous media in which a phase transition might take place as the system fills up which in network terms, is like a percolation threshold [13]. The models that we have sketched above all provide ways of generating these kinds of morphology, albeit through somewhat different mechanisms than the obvious way in which growth in physical systems takes place. The forms generated constitute an essential check on the adequacy or otherwise of these system models.

Future Directions

The biggest problems facing the development of complex systems models in general and those applied to cities in particular involve validation. The move from articulating systems as organized entities structured from the top down based on some sort of centralized control mechanisms to systems that grow in an uncoordinated way from the bottom up have also shifted our perspective from developing systems model in a parsimonious way to developing much richer models requiring more detailed data. In short, complexity theory has changed the basis for theory and model selection from an insistence that all models must be testable against data to an acceptance that if there is a strong reason why some non-testable propositions should be included in a model (as models with very rich behaviors and processes imply), then these should be included even if they cannot be tested. This is consistent with the shift from aggregate to disaggregate modeling, from the focus on equilibrium to dynamics, and on processes and behaviors rather than simply outcomes.

This changes the entire basis of validation and combined with the difficulties of articulating processes which are clearly relevant but often unobservable, the way in which models might be useful in policy making in complex systems is changing too. Modeling is now much more contingent on context and circumstance than at any time in the past. The use of multiple models, counter modeling and the synthesis of different and often contradictory model structures is now taken for granted in systems where we consider there may be no optimal solutions and where there will always be dissent from what is regarded as acceptable. Many newer models such as those based on cellular and agent-based structures and those which postulate a dynamics that involves bifurcations that are often of only theoretical interest until one such dynamic is observed, are unlikely to meet the canons of parsimony in which unambiguous tests can be made against data. These limits to validation begin to suggest that complex system models need to be classified on a continuum of ways in which they can be tested and used in practice which will depend on the type of model, the context, and the users involved [17].

In terms of more substantive developments, the question of dynamics is still of burning importance in developing better models of cities. There is an intrinsic problem of articulating urban processes of change from sparsely populated data bases which often contain only the aggregate outcomes of multiple processes. The way in which our commonsense observations of decision making in cities can be linked to more considered outcomes represented in data has barely been broached in developing good models of urban spatial behavior. In agent-based modeling, the role of cognition is important while the question of defining agents at appropriate levels is a major research focus, particularly when it comes to aggregates which are of a more abstract nature, such as groups and institutions. However what is of clear importance is the fact that as our focus becomes finer and as we disaggregate to ever more detailed levels, we then begin to represent policy processes into which these models might be nested in more detailed ways, implying that policy making and planning itself might be simply one other feature of these system models.

In short in our quest for more detail and for embracing a wider environment, city models have come to encapsulate the control mechanisms themselves as intrinsic to their functioning. It is at this point that we need much better ways of showing how such models can be used in practice. To an extent, this implies that we need to link these system models to their wider context of use and application, showing how other conceptions, other systems models, might be related to them in less formal ways than in terms of the science we have presented here. This has always been a challenge for the application of complex-

ity theory to human and social systems, and it will remain the cutting edge of this field whose rationale is the prediction and design of more efficient, equitable, and sustainable cities.

Bibliography

Primary Literature

1. Acevedo W, Gaydos L, Tilley J, Mladinich C, Buchanan J, Blauer S, Kruger K, Schubert J (1997) Urban land use change in the Las Vegas valley. US Geological Survey, Washington
2. Albert R, Jeong H, Barabási A-L (1999) Diameter of the world wide web. Nature 401:130–131
3. Allen PM (1982) Evolution, modelling, and design in a complex world. Environ Plan B 9:95–111
4. Allen PM (1998) Cities and regions as self-organizing systems: models of complexity. Taylor and Francis, London
5. Alonso W (1964) Location and land use. Harvard University Press, Cambridge
6. Anas A (1983) Discrete choice theory, information theory and the multinomial logit and gravity models. Transp Res B 17B:13–23
7. Andersson C, Rasmussen S, White R (2002) Urban settlements transition. Environ Plan B 29:841–865
8. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512
9. Barredo JI, Kasanko M, McCormick N, Lavalle C (2003) Modeling dynamic spatial processes: simulation of urban future scenarios through cellular automata. Landscape Urban Plan 64:145–160
10. Batty M (1974) Spatial entropy. Geograph Anal 6:1–31
11. Batty M (1976) Urban modelling: algorithms, calibration, predictions. Cambridge University Press, Cambridge
12. Batty M (2005) Agents, cells, and cities: new representational models for simulating multiscale urban dynamics. Environ Plan A 37(8):1373–1394
13. Batty M (2005) Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals. MIT Press, Cambridge
14. Batty M (2006) Rank clocks. Nature 444:592–596
15. Batty M (2008) The size, scale, and shape of cities. Science 319(5864):769–771
16. Batty M, Longley PA (1994) Fractal cities: a geometry of form and function. Academic Press, San Diego
17. Batty M, Torrens PM (2005) Modelling and prediction in a complex world. Futures 37(7):745–766
18. Batty M, Xie Y (1994) From cells to cities. Environ Plan B 21:s31–s48
19. Batty M, Xie Y, Sun Z (1999) Modeling urban dynamics through GIS-based cellular automata. Comput Environ Urban Syst 23:205–233
20. Ben Akiva M, Lerman S (1985) Discrete choice analysis. MIT Press, Cambridge
21. Berry BJL (1964) Cities as systems within systems of cities. Papers Proc Region Sci Assoc 13:147–164
22. Besussi E, Cecchini A, Rinaldi E (1998) The diffused city of the italian north-east: identification of urban dynamics using cellular automata urban models. Comp Environ Urban Syst 22:497–523
23. Blank A, Solomon S (2000) Power laws in cities population, financial markets and internet sites: scaling and systems with a variable number of components. Physica A 287:279–288
24. Cardillo A, Scellato S, Latora V, Porta S (2006) Structural properties of planar graphs of urban street patterns. Phys Rev E 73:066107-1–8
25. Castle CJE, Crooks AT (2006) Principles and concepts of agent-based modelling for developing geospatial simulations. Working Paper 110. Centre for Advanced Spatial Analysis, University College London, London
26. Chadwick GF (1971) A systems view of planning. Pergamon Press, Oxford
27. Chapin FS, Weiss SF (1968) A probabilistic model for residential growth. Transp Res 2:375–390
28. Cheng J (2003) Modelling spatial and temporal urban growth. Ph D Thesis, ITC Dissertation 99, ITC, Enschede, Netherlands
29. Clark C (1951) Urban population densities. J Royal Stat Soc Ser A 114:490–496
30. Clarke G (ed) (1996) Microsimulation for urban and regional policy analysis. Pion Press, London
31. Clarke KC, Gaydos LJ (1998) Loose coupling a cellular automaton model and GIS: long-term growth prediction for San Francisco and Washington/Baltimore. Int J Geograph Inform Sci 12:699–714
32. Couclelis H (1985) Cellular worlds: a framework for modeling micro-macro dynamics. Environ Plan A 17:585–596
33. Crucitti P, Latora V, Porta S (2006) Centrality measures in spatial networks of urban streets. Phys Rev E 73:036125-1-5
34. Curry L (1964) The random spatial economy: an exploration in settlement theory. Ann Assoc Amer Geograph 54:138–146
35. de Almeida CM, Batty M, Câmara G, Cerqueira GC, Monteiro AMV, Pennachin CP, Soares-Filho BS (2003) Stochastic cellular automata modeling of urban land use dynamics: empirical development and estimation. Comput Environ Urban Syst 27:481–509
36. Dendrinos DS, Mullally H (1985) Urban evolution: studies in the mathematical ecology of cities. Oxford University Press, Oxford
37. Epstein JM, Axtell RL (1996) Growing artificial societies: social science from the bottom up. MIT Press, Cambridge
38. Feigenbaum MJ (1980) The metric universal properties of period doubling bifurcations and the spectrum for a route to turbulence. Ann New York Acad Sci 357:330–336
39. Forrester JW (1969) Urban dynamics. MIT Press, Cambridge
40. Fujita M, Krugman P, Venables AJ (1999) The spatial economy: cities, regions, and international trade. MIT Press, Cambridge
41. Gabaix X (1999) Zipf's law for cities: an explanation. Quart J Econom 114:739–767
42. Gell-Man M (1994) The quark and the jaguar: adventures in the simple and the complex. Freeman and Company, New York
43. Gibrat R (1931) Les inégalités économiques. Librarie du Recueil Sirey, Paris
44. Gilbert N (2007) Agent-based models. Sage Inc., Thousand Oaks
45. Haag G (1989) Dynamic decision theory: applications to urban and regional topics. Kluwer, Dordrecht
46. Helbing D, Nagel K (2004) The physics of traffic and regional development. Contemp Phys 45:405–426
47. Hillier B (1996) Space is the machine. Cambridge University Press, Cambridge
48. Jiang B (2007) A topological pattern of urban street networks: universality and peculiarity. Physica A 384:647–655

49. Lathrop GT, Hamburg JR (1965) An opportunity-accessibility model for allocating regional growth. J Amer Inst Plan 31:95–103
50. Lowry IS (1964) Model of metropolis. Memorandum RM-4035-RC. Rand Corporation, Santa Monica
51. Mandelbot BB (1983) The fractal geometry of nature. Freeman, New York
52. May RM (1976) Simple mathematical models with very complicated dynamics. Nature 261:459–467
53. McLoughlin JB (1969) Urban and regional planning: a systems approach. Faber and Faber, London
54. Miller JH, Page SE (2007) Complex adaptive systems: an introduction to computational models of social life. Princeton University Press, Princeton
55. Nagel K, Beckman RJ, Barrett CL (1999) TRANSIMS for urban planning. LA-UR 984389. Los Alamos National Laboratory, Los Alamos
56. Newman M, Barabási A-L, Watts DJ (2006) The structure and dynamics of networks. Princeton University Press, Princeton
57. Nijkamp P, Reggiani A (1992) Interaction, evolution and chaos in space. Springer, Berlin
58. Papini L, Rabino GA, Colonna A, Di Stefano V, Lombardo S (1998) Learning cellular automata in a real world: the case study of the rome metropolitan area. In: Bandini S, Serra R, Suggi Liverani F (eds) Cellular automata: research towards industry: ACRI'96. Proc of the 3rd Conference on cellular automata for research and industry. Springer, London, pp 165–183
59. Portugali J, Benenson I (1996) Human agents between local and global forces in a self-organizing city. In: Schweitzer F (ed) Self-organization of complex structures: from individual to collective dynamics. Gordon and Breach, London, pp 537–545
60. Propolis (2004) PROPOLIS (policies and research of policies for land use and transport for increasing urban sustainability). Final report for the Commission of the European Communities. LT Consultants Ltd, Helsinki
61. Schelling TC (1969) Models of segregation. Amer Econom Rev Papers Proc 58(2):488–493
62. Schelling TC (1978) Micromotives and macrobehavior. Norton and Company, New York
63. Schweitzer F, Steinbrink J (1997) urban cluster growth: analysis and computer simulation of urban aggregations. In: Schweitzer F (ed) Self-organization of complex structures: from individual to collective dynamics. Gordon and Breach, London, pp 501–518
64. Semboloni F (2000) The growth of an urban cluster into a dynamic self-modifying spatial pattern. Environ Plan B 27:549–564
65. Tobler WR (1970) A computer movie simulating population growth in the detroit region. Econom Geograph 42:234–240
66. Tribus M (1969) Rational, descriptions, decisions and designs. Pergamon Press, New York
67. von Bertalanffy L (1969) General system theory: foundations, development, applications. George Braziller, New York. Revised Edition (1976)
68. von Thünen JH (1826) Von Thunen's isolated state. Pergamon, Oxford. (1966 translation from the 1826 German Edition Der isolierte Staat in Beziehung auf Landwirtschaft und Nationaloekonomie by P. G. Hall)
69. Waddell P (2002) UrbanSim: modeling urban development for land use, transportation and environmental planning. J Amer Plan Assoc 68:297–314
70. Ward DP, Murray AT, Phinn SR (2000) A stochastically constrained cellular model of urban growth. Comput Environ Urban Syst 24:539–558
71. Wegener M (2005) Urban land-use transportation models. In: Maguire DJ, Batty M, Goodchild MF (eds) GIS, spatial analysis, and modeling. ESRI Press, Redlands, pp 203–220
72. White RW, Engelen G (1993) Cellular automata and fractal urban form: a cellular modelling approach to the evolution of urban land use patterns. Environ Plan A 25:1175–1193
73. Wiener N (1965) Cybernetics: or the control and communication in the animal and the machine, 2nd edn. MIT Press, Cambridge
74. Wilson AG (1970) Entropy in urban and regional modelling. Pion Press, London
75. Wilson AG (1981) Catastrophe theory and bifurcation; applications to urban and regional systems. University of California Press, Berkeley
76. Wilson AG (2007) Boltzmann, Lotka, and Volterra and spatial structural evolution: a integrated methodology for some dynamical systems. J Royal Soc Interface 1–7. doi:10.1098/rsif.2007.1288
77. Wu F, Webster CJ (1998) Simulation of land development through the integration of cellular automata and multicriteria evaluation. Environ Plan B 25:103–126
78. Xie Y (1994) Analytical models and algorithms for cellular urban dynamics. Unpublished Ph D dissertation, State University of New York at Buffalo, Buffalo
79. Xie Y, Batty M (2005) Integrated urban evolutionary modeling. In: Atkinson PM, Foody GM, Darby SE, Wu F (eds) Geodynamics. CRC Press, Boca Raton, pp 273–293
80. Yeh A G-O, Li X (2000) A 'Grey-Cell' constrained ca model for the simulation of urban forms and developments in the planning of sustainable cities using GIS. Centre of Urban Planning and Environmental Management. University of Hong Kong, Pokfulam
81. Zipf GK (1949) Human behavior and the principle of least effort. Addison-Wesley, Cambridge

Books and Reviews

Barabási A (2002) Linked: the new science of networks. Perseus Publishing, New York

Batty M, Couclelis H, Eichen M (1997) Urban systems as cellular automata. Environ Plan B 24:159–164

Benenson I, Torrens PM (2004) Geosimulation: automata-based modeling of urban phenomena. Wiley, London

Clarke M, Wilson AG (1993) Dynamics of urban spatial structure: progress and problems. J Region Sci 21:1–18

Couclelis H (1997) From cellular automata models to urban models: new principles for model development and implementation. Environ Plan B 24:165–174

Dendrinos DS, Sonis M (1990) Chaos and socio-spatial dynamics. Springer, New York

Haggett P, Chorley R (1969) Network analysis in geography. Edward Arnold, London

Helbing D, Molnar P Farkas IJ, Bolay K (2001) Self-organizing pedestrian movement. Environ Plan B 28:361–383

Holland JH (1995) Hidden order: how adaptation builds complexity. Addison-Wesley, Reading

Isard W (1956) Location and space-economy: a general theory relating to industrial location, market areas, land use, trade and urban structure. MIT Press, Cambridge
Jacobs J (1961) The death and life of great american cities. Vintage Books, Random House
Krugman PR (1996) The self-organizing economy. Blackwell, Cambridge
Portugali J (2000) Self-organization and the city. Springer, Berlin
Resnick M (1994) Termites, turtles and traffic jams: explorations in massively parallel micro-worlds. MIT Press, Cambridge
Sanders L, Pumain D, Mathian H, Guerin-Pace F, Bura S (1997) SIMPOP: a multiagent system for the study of urbanism. Environ Plan B 24:287–305
Simon HA (1969, 1996) Sciences of the artificial. MIT Press, Cambridge
Stewart JQ, Warntz W (1958) Physics of population distribution. J Region Sci 1:99–123
White RW (1998) Cities and cellular automata. Discrete Dyn Nature Soc 2:111–125
Willumsen LG, de Ortuzar JD (1990) Modelling transport. Wiley, Chichester
Wilson AG (2000) Complex spatial systems: the modelling foundations of urban and regional analysis. Pearson Education, Harlow
Wilson AG (1974) Urban and regional models in geography and planning. Wiley, Chichester

Climate Change and Agriculture

Cynthia Rosenzweig
NASA/Goddard Institute for Space Studies,
Columbia University, New York, USA

Article Outline

Glossary
Introduction
Conclusions
Future Directions
Bibliography

Glossary

Anthropogenic emissions Greenhouse gas emissions that are produced as a result of humans through such developments as industry or agriculture.

Greenhouse gases The gases of the atmosphere that create the greenhouse effect, which keeps much of the heat from the sun from radiating back into outer space. Greenhouse gases include, in order of relative abundance: Water vapor, carbon dioxide, methane, nitrous oxide, ozone, and CFCs. Greenhouse gases come from natural sources and human activity; present CO_2 levels are ~380 ppmv, approximately 100 ppmv higher than they were in pre-industrial times.

Soybean cyst nematode *Heterodera glycines*, a plant-parasite that infects the roots of soybean, with the female becoming a cyst. Infection causes various symptoms, including a serious loss of yield.

El Niño-southern oscillation (ENSO) A phenomenon in the equatorial Pacific Ocean characterized by a positive sea-surface temperature departure from normal (for the 1971–2000 base period) in the Niño 3.4 region greater than or equal in magnitude to 0.5°C, averaged over three consecutive months.

North atlantic Oscillation (NAO) A hemispheric, meridional oscillation in atmospheric mass with centers of action near Iceland and over the subtropical Atlantic.

Vegetative index A simple numerical indicator used to analyze remote sensing measurements, often from space satellites, to determine how much photosynthesis is occurring in an area.

Soil organic carbon All the organic compounds within the soil without living roots and animals.

Introduction

The term climate change refers to an overall shift of mean climate conditions in a given region. The warming trend associated with anthropogenic emissions of greenhouse gases and the enhanced greenhouse effect of the atmosphere can and should be regarded as a "climate change" when viewed on the time scale of decades or a few centuries.

Climate change exacerbates concerns about agricultural production and food security worldwide. At global and regional scales, food security is prominent among the human concerns and ecosystem services under threat from dangerous anthropogenic interference in the earth's climate [17,29,50]. At the national scale, decision-makers are concerned about potential damages that may arise in coming decades from climate change impacts, since these are likely to affect domestic and international policies, trading patterns, resource use, regional planning, and human welfare.

While agro-climatic conditions, land resources and their management are key components of food production, both supply and demand are also critically affected by distinct socio-economic pressures, including current and projected trends in population and income growth and distribution, as well as availability and access to technology and development. In the last three decades, for instance, average daily per capita intake has risen globally from 2,400 to 2,800 calories, spurred by economic growth, improved production systems, international trade, and glob-

alization of food markets [12]. Feedbacks of such growth patterns on cultures, personal tastes, and lifestyles have in turn led to major dietary changes – mainly in developing countries – where shares of meat, fat, and sugar in total food intake have increased significantly [13]. Thus, the consequences of climate change on world food demand and supply will depend on many interactive dynamic processes.

Agriculture plays two fundamental roles in human-driven climate change. It is one of the key human sectors that will be affected by climate change over the coming decades, thus requiring adaptation measures. Agriculture is also a major source of greenhouse gases to the atmosphere, including carbon dioxide (CO_2) due to land-use change and farm operations; methane (CH_4) from rice production and livestock husbandry, and nitrous oxide (N_2O) from application of nitrogen fertilizer. As climate changes as well as socio-economic pressures shape future demands for food, fiber and energy, synergies can be identified between adaptation and mitigation strategies, so that robust options that meet both climate and societal challenges can be developed. Ultimately, farmers and others in the agricultural sector will be faced with the dual task of contributing to global reductions of carbon dioxide and other greenhouse gas emissions, while coping with an already-changing climate.

A changing climate due to increasing anthropogenic emissions of greenhouse gases will affect both the productivity and geographic distribution of crop and pasture species. The major climate factors contributing to these responses include increasing atmospheric carbon dioxide, rising temperature, and increasing extreme events, especially droughts and floods. These factors in turn will affect water resources for agriculture, grazing lands, livestock, and associated agricultural pests. Effects will vary, depending on the degree of change in temperature and precipitation and on the particular management system and its location. Several studies have suggested that recent warming trends in some regions may have already had discernible effects on some agricultural systems [17].

Climate change projections are uncertain in regard to both the rate and magnitude of temperature and precipitation variation in the coming decades. This uncertainty arises from a lack of precise knowledge of how climate system processes will change and of how population growth, economic and technological developments, and land-use patterns will evolve in the coming century [16,17]. Despite these uncertainties, the ultimate significance of the climate change issue is related to its global reach, affecting agricultural regions throughout the world in complex ways. After approximately two decades of research, ten major conclusions may be drawn in regard to climate change and agriculture.

Effects on Agricultural Systems Will Be Heterogeneous

Global studies on projected climate change effects on agriculture show that negative and positive effects will occur both within countries and across the world. In large countries such as the United States, Russia, Brazil, and Australia, agricultural regions will likely be affected quite differently. Some regions will experience increases in production and some declines (see, e. g., [34]). At the international level, this implies possible shifts in comparative advantage for production of export crops. This also implies that adaptive responses to climate change will necessarily be complex and varied. Due to differences in global climate model projections and decadal variability, it is impossible to project exact effects in any one location for any given decade.

Developing Countries Are More Vulnerable

Despite general uncertainties about the rate and magnitude of climate change and especially about consequent hydrological changes, regional and global studies have consistently shown that agricultural production systems in the mid and high latitudes are more likely to benefit in the near term (to mid-century), while production systems in the low-latitudes are more likely to decline (Fig. 1) [17]. In biophysical terms, rising temperatures will likely push many crops beyond their limits of optimal growth and yield. Higher temperatures will intensify the evaporative demand of the atmosphere, leading to greater water stress, especially in semi-arid regions. Since most developing countries are located in lower-latitude regions (some which are indeed semi-arid) while most developed countries are located in the more humid mid- to high latitudes, this finding suggests a divergence in vulnerability between these groups of nations, with far-reaching implications for future world food security [31,37].

Furthermore, developing countries often have fewer resources with which to devise appropriate adaptation measures to meet changing agricultural conditions. The combination of potentially greater climate stresses and lower adaptive capacity in developing countries creates different degrees of vulnerability between rich and poor nations as they confront global warming. This difference is due in part to the potentially greater detrimental impacts of a changing climate in areas that are already warm (particularly if such areas are also dry), and in part to the generally lower levels of adaptive capacity in developing countries.

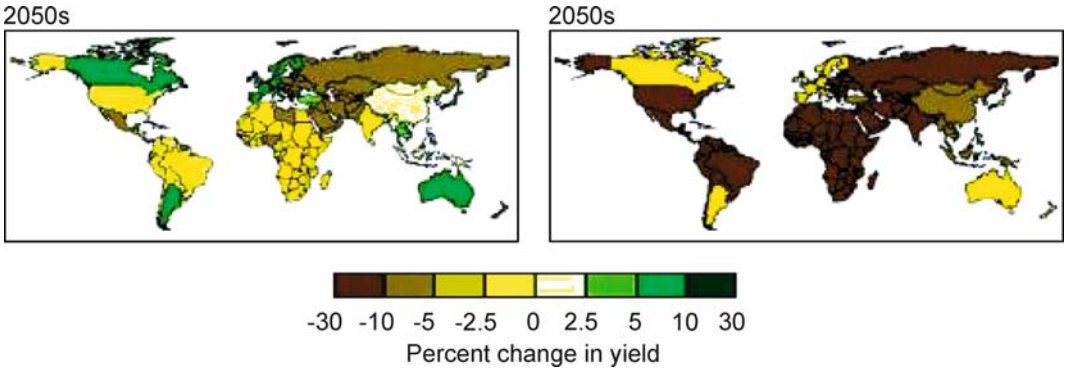

Climate Change and Agriculture, Figure 1
Potential changes (%) in national cereal yields for the 2050s (compared with 1990) under the HadCM3 SRES A1FI with (*left*) and without (*right*) CO_2 effects [31]

Development Path Matters

Since climate is not the only driving force on agriculture, researchers now conduct scenario analysis that include linked sets of population projections, economic growth rates, energy technology improvements, land-use changes, and associated emissions of greenhouse gases [31]. Regional patterns related to economic development and adaptive capacity contribute to differing levels of climate change impacts [17]. Scenarios with higher economic growth rates and less attention to environmental issues lead to high temperatures and reduced adaptive capacity, which in turn lead to pronounced decreases in yields both regionally and globally. Scenarios with lower greenhouse gas emissions and greater attention to environmental issues lead to lower amounts of temperature rise and crop production declines.

Long-Term Effects Are Negative for Both Developed and Developing Countries

If the effects of climate change are not abated, production in the mid- and high-latitudes is likely to decline in the longer term (i. e., above ~3°C warming) (Fig. 2) [16]. These results are consistent over a range of temperature, precipitation, and direct CO_2 effects tested, and are due primarily to the detrimental effects of heat and water stress as temperatures rise. While the beneficial effects of CO_2 may eventually level out, the detrimental effects of warmer temperatures and greater water stress are more likely to be progressive in all regions. Although the precise levels of CO_2 effects on crops and their contribution to global crop production are still active areas of research [47], global impacts are likely to turn negative in all regions sometime during the second half of the century, and perhaps before then for some major crops. For instance, by 2050 climate change is projected to have a downward pressure on yields of almonds, walnuts, avocados, and table grapes in California. Opportunities for expansion into cooler regions have been identified, but this adaptation would require substantial investments and may be limited by non-climatic constraints [24].

Water Resources Are Key

Recent flooding and heavy precipitation events in the US and worldwide have caused great damage to crop production. If the frequency of these weather extremes were to increase in the near future, as recent trends for the US indicate and as projected by global climate models [17,48], the cost of crop losses in the coming decades could rise dramatically. US corn production losses due to excess soil moisture, already significant under current climate, may double during the next thirty years, causing additional damages totaling an estimated $3 billion per year (Fig. 3) [41]. These costs may either be borne directly by those farmers affected or may need to be transferred to private or governmental insurance and disaster relief programs. There is also concern for tractability in the spring and water-logging in the summer in mid and high latitudes where precipitation is projected to increase.

Changes in crop water demand and water availability will affect the reliability of irrigation, which competes for growing municipal and industrial demands [42]. Studies link climate change scenarios with hydrologic, agricultural, and planning models to estimate water availability for agriculture under changing climate conditions, to explore changes in ecosystem services, and to evaluate adap-

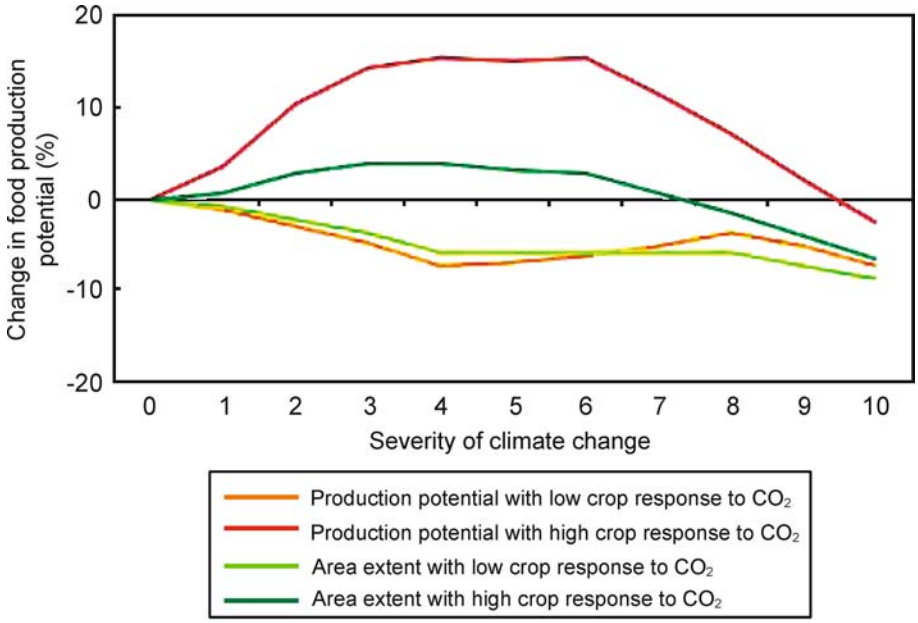

Climate Change and Agriculture, Figure 2
Change in food production potential in relation to severity of climate change [13]

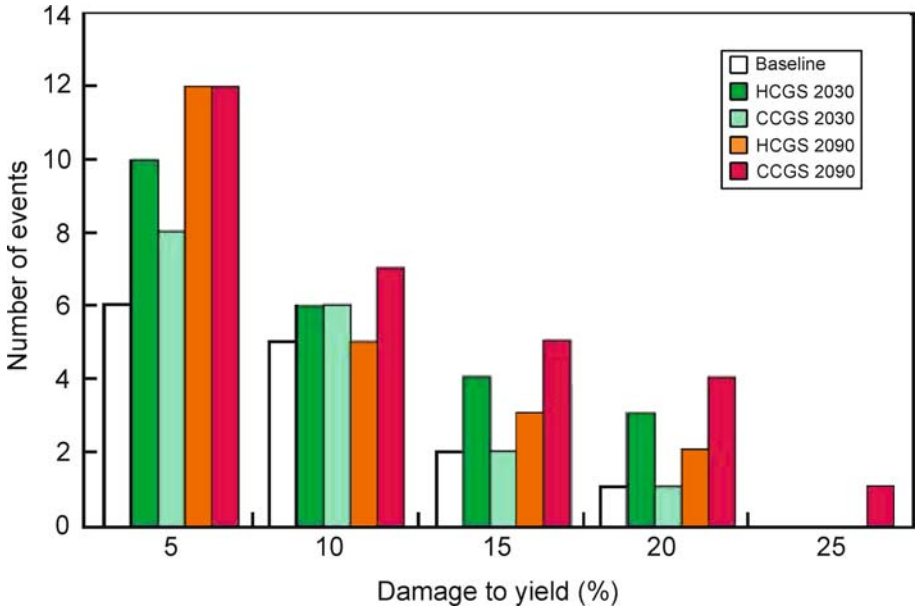

Climate Change and Agriculture, Figure 3
Number of events causing damage to maize yields due to excess soil moisture conditions, averaged over all study sites, under current baseline (1951–1998) and climate change conditions. The Hadley Center (HC) and Canadian Center (CC) scenarios with greenhouse gas and sulfate aerosols (GS) were used. Events causing a 20% simulated yield damage are comparable to the 1993 US Midwest floods [41]

tation strategies for the water resources and agriculture sectors. Major irrigated agricultural regions are very likely to be affected by changing supplies of and demands for water due a changing climate especially under conditions of expansion of irrigated lands [42].

Cultivars are available for agricultural adaptation to the projected changes, but their demand for water may be higher than currently adapted varieties. Thus, even in relatively water-rich areas, changes in water demand due to climate change effects on agriculture and increased demand from urban growth will require timely improvements in crop cultivars, irrigation and drainage technology, and water management. In tropical regions, the use of agroforestry may be an economically feasible way to protect crop plants from extremes in microclimates and soil moisture [22].

Agricultural Pests and Diseases May Spread

Increased pest damage arises from changes in production systems, enhanced resistance of some pests to pesticides, and the production of crops in warmer and more humid climatic regions where crops are more susceptible to pests. Changes in crop management techniques, particularly the intensification of cropping, reduction in crop rotations, and increase in monocultures, have increased the activity of pests. The expansion of worldwide trade in food and plant products has also increased the impact of weeds, insects, and diseases on crops. The geographical ranges of several important insects, weeds, and pathogens in the US have recently expanded, including soybean cyst nematode (*Heterodera glycines*) and sudden death syndrome (*Fusarium solani* f. sp. *glycines*) (Fig. 4) [15,39,43].

Current climate trends and extreme weather events may be directly and indirectly contributing to the increased pest damage [17,39,52]. Downy mildew (*Plasmopara viticola*) epidemics on grape, the most serious grapevine disease in northern Italy, may increase under climate change, even though reduced precipitation may have a counterbalancing effect on disease pressure [44].

Such changes need to be put in the context of the global increases in pest-induced losses of crops in all regions since the 1940s [30,33] and the more than 33-fold increase in both the amount and toxicity of pesticide used over the same period [33]. Climate change thus may exacerbate environmental and public health issues related to agricultural chemicals [39], since increased applications of agricultural chemicals are likely to be needed in response to increasing disease pressure. Improved knowledge of the effects of climate on host–pathogen interactions will contribute to the adaptive capacity of agro-ecosystems.

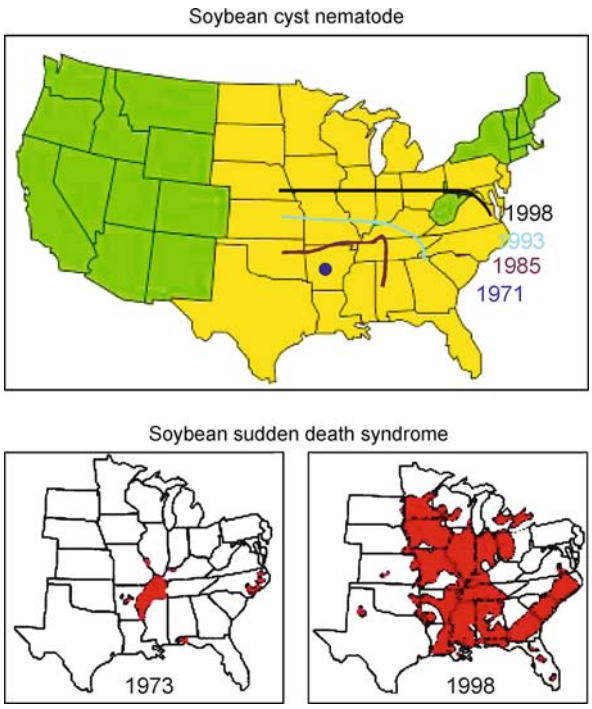

Climate Change and Agriculture, Figure 4
Range expansion of soybean cyst nematode (*Heterodera glycines*) from 1971 to 1989 (*top*) and soybean sudden death syndrome (*Fusarium solani f. sp. Glycines*) from 1973 to 1998 (*bottom*) in North America [40]

Current Climate Stress Is a Key Entry Point for Climate Change

There is an important interplay between current and future climate stresses. Since farmers have dealt with climatic fluctuations since the advent of agriculture, improving strategies for dealing with present climate extremes – such as droughts, floods, and heatwaves – is an important way to prepare for climate change. Many agricultural regions are affected by the major climate variability systems, including the processes known as the El Niño-Southern Oscillation (ENSO) and the North Atlantic Oscillation (NAO) [36]. The El Niño phase of the ENSO cycle tends to bring rainfall to Uruguay, while La Niña brings drought, as shown in Fig. 5 for 1998, an El Niño year, and 2000, the following La Niña.

In terms of prediction tools, ENSO models provide the opportunity for testing and validation of climate prediction and assessment on shorter seasonal-to-interannual time-scales. Skill in predicting climate changes on shorter time-scales, particularly the ENSO periods of the last twenty years when good observations exist, may lend credence to projections of global warming over the longer-term. As global climate models are further developed with

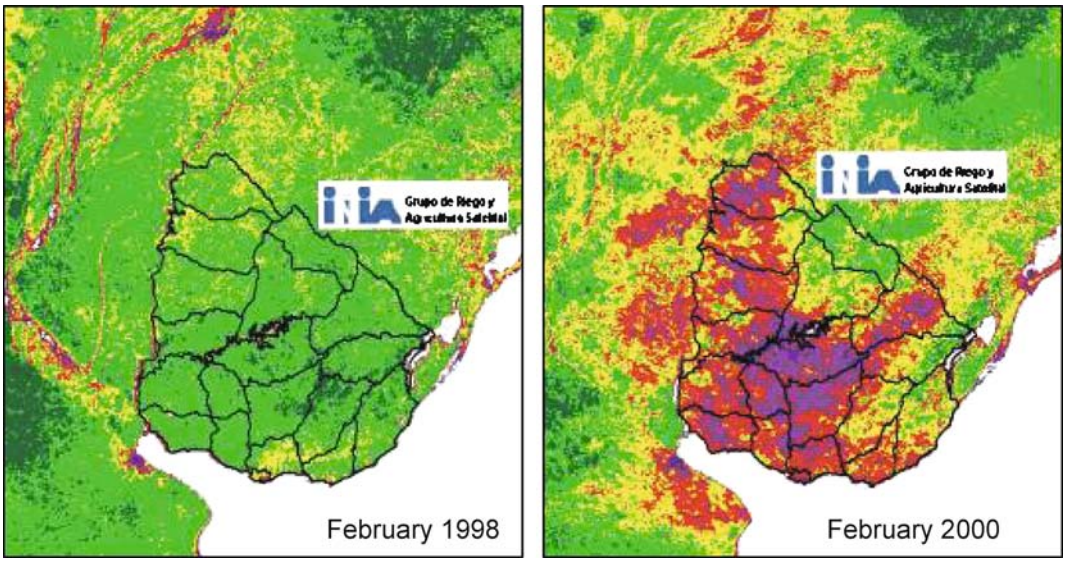

Climate Change and Agriculture, Figure 5
Vegetative Index (NDVI) for El Niño (1998) and La Niña (2000) years in Uruguay. *Green* = adequate water conditions; *red/purple* = drought conditions [2]

improved parametrizations and higher spatial resolution, they are likely to improve simulations of ENSO and other large-scale variability processes. The interaction of these systems with underlying anthropogenic trends caused by increasing greenhouse gas concentrations in the atmosphere is an active area of contemporary climate science. For regions directly affected by ENSO and other systems, such changes, if they do indeed occur, may become important manifestations of global warming.

Adaptation Is Necessary

'Coping range' is a useful paradigm for improving responses to climate stresses of today and preparing for the climate changes of tomorrow [20]. An agricultural system may currently exist within a 'coping range' of climate variability that may be exceeded as incidence of extreme events increases under changing climate conditions (Fig. 6). The goal is to increase the coping range over which an agricultural system may thrive under such changes through the process of adaptation.

Adaptation can help farmers to minimize negative impacts of climate on human activities and ecosystems and to take advantage of potential beneficial changes. *Adaptation* to climate change can be defined as the range of actions taken in response to changes in local and regional climatic conditions [45]. Adaptation responses include both *autonomous adaptation* actions (i. e., those taken spontaneously and independently by individual farmers), and *planned adaptation* actions (i. e., those facilitated by climate-specific regulations and incentives put in place by regional, national and international policies) (see Table 1) [17]. In terms of the multiple factors impinging on agriculture, however, system responses to socio-economic, institutional, political or cultural pressures may outweigh response to climate change alone in driving the evolution of agricultural systems. The *adaptive capacity* of a system, in the context of climate change, can be viewed as the full set of system skills – i. e., technical solutions available to farmers in order to respond to climate stresses – as determined by the socio-economic and cultural settings, plus institutional and policy contexts, prevalent in the region of interest [17].

While current agronomic research confirms that at the field level crops would respond positively to elevated CO_2 in the absence of climate change (e. g., [1,17,19,21], the associated impacts of high temperatures, altered patterns of precipitation, and possibly increased frequency of extreme events (such as drought and floods) are likely to require a range of adaptation responses, some of which are listed in Table 2.

Mitigation Reduces Long-Term Risk

Agriculture has an important role to play in mitigation of climate change. *Mitigation* is defined as intervention aimed at reducing the severity of climate change by reducing the atmospheric concentration of greenhouse gases, ei-

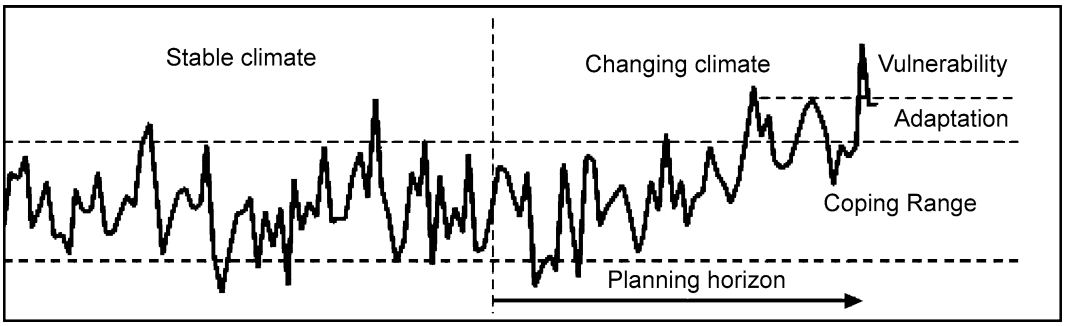

Climate Change and Agriculture, Figure 6
Coping range of climate variability (adapted from [20])

Climate Change and Agriculture, Table 1
Adaptation approaches to climate impacts on agriculture

Approach	Definition	Operation
Autonomous	Adaptation that does not constitute a conscious response to climatic stimuli but is triggered by ecological changes in natural systems and by market or welfare changes in human systems. Also referred to as spontaneous adaptation.	Crop calendar shifts (planting, input schedules, harvesting) Cultivar changes Crop-mix changes
Planned	Adaptation that is the result of a deliberate policy decision, based on an awareness that conditions have changed or are about to change and that action is required to return to, maintain, or achieve a desired state.	Land-use incentives Irrigation infrastructure Water pricing Germplasm development programs

Climate Change and Agriculture, Table 2
Key agronomic impacts and adaptation responses

Agricultural impacts	Adaptation response
Biomass increase under elevated CO_2	Cultivar selection and breeding to maximize yield
Acceleration of maturity due to higher temperature	Cultivar selection and breeding of slower maturing types
Heat stress during flowering and reproduction	Early planting of spring crops
Crop losses due to increased droughts and floods	Changes in crop mixtures and rotations; warning systems; insurance
Increased pest damage	Improved management; increased pesticide use; biotechnology

ther by reducing emissions or by enhancing sinks. There are several major ways that the agricultural sector can contribute to climate change mitigation.

Soil Carbon Sequestration Of the approximately 150 GT of carbon that were lost in the last century due to land conversion to agriculture and subsequent production, about two thirds were lost due to deforestation and one-third, roughly 50 GT, due to cultivation of current agricultural soils and exports as food products [50]. The latter figure represents the maximum theoretical amount of carbon that could be restored in agricultural soils. In practice, as long as 40–50% of total above-ground grain or fruit production is exported as food to non-agricultural areas, the actual carbon amount that can be restored in agricultural soils is much lower.

Efforts to improve soil quality and raise soil organic carbon (SOC) levels include crop management and conservation tillage techniques. These practices have evolved as means to enhance sustainability and resilience of agricultural systems, rather than with soil carbon sequestration as primary motivation. They include so-called "best practice" agricultural techniques, such as use of cover crops and/or nitrogen fixers in rotation cycles, judicious use of fertilizers and organic amendments, soil water management improvements to irrigation and drainage, and improved varieties with high biomass production.

Conventional tillage is defined [27] as the mechanical manipulation of the topsoil that leaves no more than 15% of the ground surface covered with crop residues. In contrast, no-till management is defined as the avoidance of mechanical manipulation of the topsoil so as to leave it

undisturbed and covered with surface residues from harvesting the prior crop to planting the new crop.

Best agricultural practices can result in a net augmentation of soil carbon and in enhanced productivity due to better soil structure and soil moisture management. The relevant practices include precise and timely applications and spatial allocation of fertilizers, use of slow-release fertilizers, prevention of erosion, shortening or elimination of fallow periods, use of high-residue cover crops and green-manure crops, and minimized mechanical disturbance of soil (e. g., zero tillage). Altogether, such practices may lead to partial or even complete restoration of the soil's organic carbon content where it had been depleted. In some cases, it might even be possible to store more carbon than had originally been present in the "virgin" soil. Where the soils had been severely degraded and their agricultural productivity greatly impaired, they may be converted to grassland or afforested so as to serve as carbon sinks.

The overall potential for carbon storage depends in each case on such factors as climate, type of vegetation, topography, depth and texture of the soil, past use (or abuse), and current management.

Along with sequestering carbon, these practices have the potential to improve soils in developing countries. In areas such as West Africa, soil fertility depletion has been described as the single most important constraint to food security [4]. Studies in smallholder agricultural farms in Africa have already illustrated significant increases in system carbon and productivity through organic-inorganic resources management (Roose and Barthes [35])

Biofuels Agriculture may help to mitigate anthropogenic greenhouse emissions through the production of biofuels. As has been demonstrated by ethanol based on corn production, issues involved with biofuel production include potential competition with food production, increased pollution from fertilizers and pesticides, and loss of biodiversity. Biofuels derived from low-input high-diversity mixtures of native grassland perennials can provide more usable energy, greater greenhouse gas reduction, and less agrichemical pollution per hectare than corn grain ethanol or soybean biodiesel [46]. The higher net energy results arise because perennial grasses require lower energy inputs and produce higher bioenergy yield. Furthermore, all aboveground biomass of the grasses can be converted to energy, rather than just the seed of either corn or soybean. These perennial grasses also sequester carbon at significant rates [46].

Other Greenhouse Gases Because of the greater global warming potential (GWP) of methane (21) and nitrous oxide (310) compared to carbon dioxide (1), reductions of non-CO_2 greenhouse gas emissions from agriculture can be quite significant and achieved via the development of more efficient rice (for methane) and livestock production systems (for both methane and nitrous oxide). In intensive agricultural systems with crops and livestock production, direct CO_2 emissions are predominantly connected to field crop production and are typically in the range of 150–200 kg C ha^{-1} yr^{-1} [14,51]. Recent full greenhouse gas analysis of different farm systems in Europe showed that such CO_2 emissions represent only 10–15% of the farm total, with emissions of CH_4 contributing 25–30% and emissions of N_2O accounting for as much as 60% of total CO_2-equivalent greenhouse gas emissions from farm activities. The N_2O contribution arises from substantial nitrogen volatilization from fertilized fields and animal waste, but it is also a consequence of its very high GWP.

In Europe, methane emissions are mostly linked to cattle digestive pathways; its contribution also dominates that of CO_2, due in part to methane's high GWP. Mitigation measures for methane production in livestock include improved feed and nutrition regimes, as well as recovery of bio-gas for on-farm energy production. Effective reduction of N_2O emissions is more difficult, given the largely heterogeneous nature of emissions in space and time and thus the difficulty of timing fertilizer applications and/or manure management. Large uncertainties in emission factors also complicate the assessment of efficient N_2O-reduction strategies. Current techniques focus on reduction of absolute amounts of fertilizer nitrogen applied to fields, as well as on livestock feeding regimes that reduce animal excreta.

Climate Change Effects on Agriculture Are Occurring Already

Agricultural effects of observed climate changes are being documented in many regions of the world (Fig. 7). Changes in crop phenology provide important evidence of responses to recent regional climate change. Such changes are apparent in perennial crops, such as fruit trees and wine-making varieties of grapes, which are less dependent on yearly management decisions by farmers than annual crops and are also often easier to observe.

Phenological changes are often observed in tandem with changes in management practices by farmers. Between 1951 and 2004 in Europe, agricultural crops have advanced 2.1 days/decade in response to recent warming in spring [28]. In Sahelian countries, increasing temperature in combination with rainfall reduction has led to

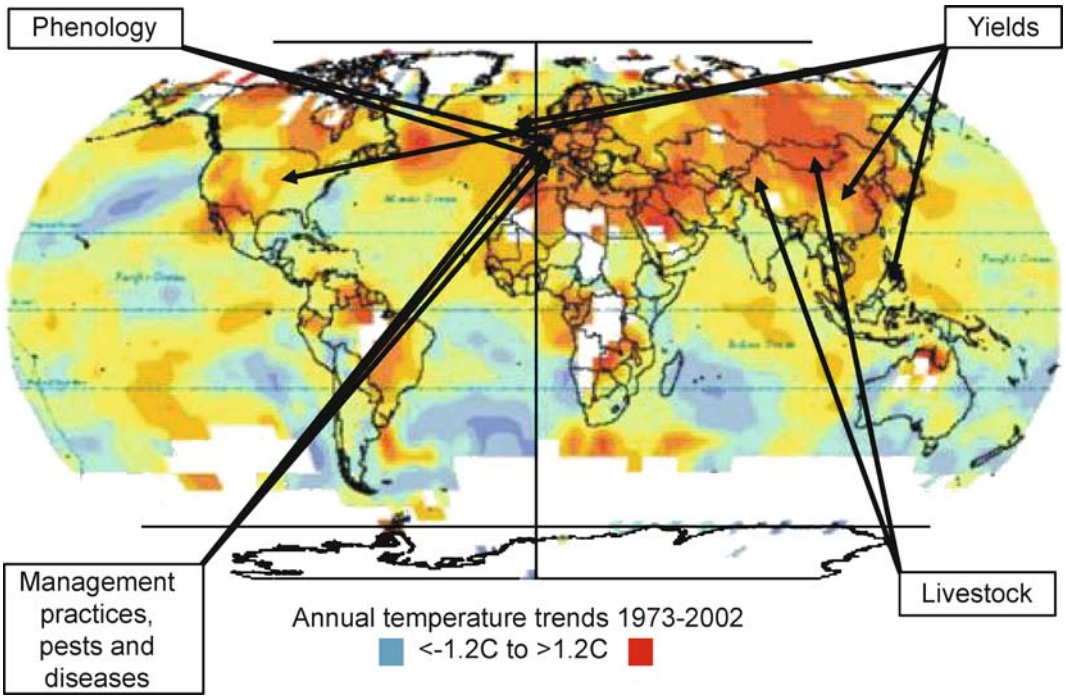

Climate Change and Agriculture, Figure 7
Locations of observed changes in agriculture in response to climate changes

a reduced length of vegetative period, no longer allowing present varieties to complete their cycle [5].

A negative effect of warming for local rice production has been observed by the International Rice Research Institute (IRRI) in the Philippines (yield loss of 15% for 1°C increase of growing-season minimum temperature in the dry season) [32]; a similar effect has been noted on hay yield in the UK (1°C increase in July–August led to a 0.33 t/ha loss) [6]. At the county level, US maize and soybean yields are demonstrating a positive effect of cooler and wetter years in the Midwest and hotter and drier years in the North-west plains [23]. In the case of the Sahel region of Africa, warmer and drier conditions have served as a catalyst for a number of other factors that have accelerated a decline in groundnut production [49]. For livestock, one study in Tibet reports a significant relationship of improved performance with warming in high mountainous conditions [9], while pasture biomass in Mongolia has been negatively affected by the warmer and drier climate, as observed at a local station [3] or at the regional scale by remote sensing [10].

Conclusions

Climate change brings both challenges and opportunities to agriculture. Farmers and researchers are being called on to simultaneously adapt to and mitigate climate change through a myriad of activities involving management practices, crop breeding, and new production systems. Some of these can be mutually re-enforcing, especially in view of the projected increased climate variability under climate change. This is because, most mitigation techniques currently considered in agriculture, including reduced tillage, were originally designed as "best practice" management strategies, aimed at enhancing the long-term stability and resilience of cropping systems in the face of climate variability or of increased cultivation intensity. By increasing the ability of soils to hold soil moisture and to better withstand erosion, and by enriching ecosystem biodiversity through the establishment of more diversified cropping systems, mitigation techniques implemented locally for soil carbon sequestration may also help cropping systems to better withstand droughts and/or floods, both of which are projected to increase in frequency and severity in future warmer climates. As climate change progresses, agriculture will continue to play a leading role in responding to a dynamic environment.

Future Directions

The time has come to incorporate climate as an essential factor in development planning and implementation. In

the past, responses to climate variability were often too narrowly focused and lacking in institutional fit. Development programs are now beginning to include recommendations to mainstream responses to climate variability and change.

Magalhães [25] gave as an example of the need for a broad focus when considering climate in planning the early drought policies in Northern Brazil. Responses had focused on improving the water-supply infrastructure (e. g. building dams and digging wells) rather than on redressing the social and economic vulnerabilities and the need to build human capital by means of education, institutions, and market incentives for sustainability. Adaptive policies should be broadly conceived so as to increase and secure household entitlements, to change land-use patterns that lead to degradation, and to develop means of support for inhabitants that are less sensitive to the vagaries of climate.

At what levels or scales of organization (national, regional, household, and individual) can the variability of climate and the sustainability of agricultural production be addressed effectively? Because many nations encompass several and often most numerous climatic zones, the challenge faced by national agricultural managers and policy make is to foster sustainability at the regional level while building a foundation from the bottom-up at the individual as well as at the household and community levels. Dilley [8] believes that greater benefits of food security could be realized if knowledge were made more readily available at the household level, thereby improving the ability of more people to make even small adjustments based on anticipated climatic conditions. Multiple-scale efforts are clearly needed, with pathways of communication among the various levels and sectors of society.

Regions do not exist in isolation, as evidenced by the effects that extreme climate events occurring along the Atlantic coast of South America have on the sustainability of the inland Amazon rainforest. At least part of the pressure to deforest the Amazon region arises from the westward migration of farmers from Northern Brazil who suffer from ENSO-related droughts there (Magalhães [25]). Thus, policies related to sustainability and climate issues need to take regional interactions and their direct and indirect linkages into account.

Beyond interconnections among regions within a nation, there are the larger national concerns of economics, finance, and international relations that may affect the range of climate-adaptation policy choices. Engaging with the international community on issues of climate variability and change can lead to capacity building in both developed and developing countries, as climate and societal processes are studded and as improved understanding is incorporated into policies. This can be accomplished through interactions with international bodies dealing with climate variability and change, such as the World Meterological Organization and the United Nations Framework Convention on Climate Change. There is a growing realization that climate plays an important role in sustainable development: It is a component of natural capital, an occasional trigger to socio-economic crises caused by extreme events, and a long-term component of global environmental change.

Bibliography

1. Ainsworth EA, Long SP (2005) What have we learned from 15 years of free-air CO_2 enrichment (FACE)? A meta-analysis of the response of photosynthesis, canopy properties and plant production to rising CO_2. New Phytologist 165:351–275
2. Baethgen WE, Giménez A (2002) Seasonal climate forecasts and the agricultural sector of Uruguay. In: Examples of ENSO-Society Interactions. The International Research Institute for Climate and Socity, New York. http://iri.columbia.edu/climate/ENSO/societal/resource/example/Baethgen.html
3. Batimaa P (2005) The potential impact of climate change and vulnerability and adaptation assessment for the livestock sector of Mongolia. Assessments of Impacts and Adaptations to Climate Change. AIACC, Washington DC, 20 pp
4. Bationo A, Kihara J, Vanlauwe B, Waswa B, Kimetu J (2007) Soil organic carbon dynamics, functions and management in West African agro-ecosystems. Agric Syst 94:13–25
5. Ben Mohamed A, Duivenbooden NV, Abdoussallam S (2002) Impact of climate change on agricultural production in the Sahel, Part 1. Methodological approach and case study for millet in Niger. Clim Chang 54:327–348
6. Cannell MGR, Palutikof JP, Sparks TH (1999) Indicators of climate change in the UK. DETR, London, 87 pp
7. Chmielewski FM, Muller A, Bruns E (2004) Climate changes and trends in phenology of fruit trees and field crops in Germany, 1961–2000. Agric Forest Meteorol 121(1-2):69–78
8. Dilley M (2003) Regional responses to climate variability in Southern Africa. In: O'Brian K, Vogel C (eds) Coping with climate variability: The use of seasonal climate forecasts in Southern Africa. Ashgate, Hampshire, pp 35–47
9. Du MY, Kawashima S, Yonemura S, Zhang XY, Chen SB (2004) Mutual influence between human activities and climate change in the Tibetan Plateau during recent years. Global Planet. Change 41:241–249
10. Erdenetuya M (2004) Application of remote sensing in climate change study: Vulnerability and adaptation assessment for grassland ecosystem and livestock sector in Mongolia project. AIACC Annual Report, Washington DC
11. Fischer G, Shah M, Velthuizen H, Nachtergael FO (2001) Global agro-ecological assessment for agriculture in the 21st century. International Institute for Applied Systems Analysis. IIASA, Laxenburg
12. Fischer G, Shah M, van Velthuizen H (2002) Climate change and agricultural vulnerability, special report to the UN World Summit on Sustainable Development, Johannesburg 2002. IIASA, Laxenburg

13. Fischer, Shah GM, Tubiello FN, van Velhuizen H (2005) Socio-economic and climate change impacts on agriculture: An itegrated assessment, 1990–2080. Philos Trans R Soc B-Biol Sci 360(1463):2067–2083
14. Flessa H, Ruser R, Dörsch P, Kamp T, Jimenez MA, Munch JC, Beese F (2002) Integrated evaluation of greenhouse gas emissions (CO_2, CH_4, N_2O) from two farming systems in southern Germany. Agric Ecosyst Environ 91:175–189
15. Hartman GL, Noel GR, Gray LE (1995) Occurrence of soybean sudden death syndrome in east-central Illinois and associated yield losses. Plant Disease 79:314–318
16. IPCC (2007) Climate change 2001: The scientific basis. Contributions of Working Group I to the fourth assessment report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge
17. IPCC (2007) Climate change 2007: Impacts, adaptation, and vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge
18. IPCC (2007) Climate Change 2001: Mitigation. Contribution of Working Group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, 1076 pp
19. Jablonski LM, Wang X, Curtis PS (2002) Plant reproduction under elevated CO_2 conditions: A meta-analysis of reports on 79 crop and wild species. New Phytologist 156(1):9–26
20. Jones PD, Mann ME (2004) Climate over past millennium. Rev Geophys 42:RG2002
21. Kimball BA, Kobayashi K, Bindi M (2002) Response of agriculture crops to free-air CO_2 enrichment. Adv Agron 77:293–368
22. Lin BB (2007) Agroforestry management as an adaptive strategy against potential microclimate extremes in coffee agriculture. Agric Forest Meteorol 144(1-2):85–94
23. Lobell DB, Asner GP (2003) Climate and management contributions to recent trends in US agricultural yields. Science 299:1032
24. Lobell DB, Field CB, Cahill KN, Bonfils C (2006) Impacts of future climate change on California perennial crop yields: Model projections with climate and crop uncertainties. Agricult Forest Meteorol 141:208–218
25. Magalhães AR (2000) Sustainable development: Climate and policy linkages. In: Proceedings of the International Forum on Climate Prediction, Agriculture and Development, 26–-28 April. International Research Institute for Climate Prediction, Palisades, New York, pp 3–10
26. Magalhães AR, Glantz MH (eds) (1992) Socioeconomic impacts of climate variations and policy response in Brazil. Esquel Brazil Foundation, Brasilia
27. Marland G, West TO, Schlamadinger B, Canella L (2003) Managing soil organic carbon in agriculture: The net effect on greenhouse gas emissions. Tellus 55B:613–621
28. Menzel A, von Vopelius J, Estrella N, Schleip C, Dose V (2006) Farmers' annual activities are not tracking speed of climate change. Climate Res 32:201–207
29. Millennium Ecosystem Assessment (2005) Ecosystem and human well-being: Synthesis. Island, Washington DC
30. Oerke EC, Dehne HW, Schohnbeck F, Weber A (1995) Crop production and crop protection: Estimated losses in major food and cash crops. Elsevier, Amsterdam, 830 pp
31. Parry ML, Rosenzweig C, Iglesias A, Livermore M, Fischer G (2004) Effects of climate change on global food production under SRES emissions and socio-economic scenarios. Glob Environ Chang 14:53–67
32. Peng SB, Huang JL, Sheehy JE, Laza RC, Visperas RM, Zhong XH, Centeno GS, Khush GS, Cassman KG (2004) Rice yields decline with higher night temperature from global warming. Proc Natl Acad Sci USA 101(27):9971–9975
33. Pimentel D (1997) Pest management in agriculture, In: Pimentel D (ed) Techniques for reducing pesticide use: Environmental and economic benefits. Wiley, Chichester, pp 1–12
34. Reilly J, Tubiello F, McCarl B, Abler D, Darwin R, Fuglie K, Hollinger S, Izaurralde C, Jagtap S, Jones J, Mearns L, Ojima D, Paul E, Paustian K, Riha S, Rosenberg N, Rosenzweig C (2003) US agriculture and climate change: New results. Clim Chang 57:43–69
35. Roose E, Barthes B (2001) Organic matter management for soil conservation and productivity restoration in Africa: a contribution from Francophone research. Nutrient Cycling in Agroecosystems 61(1–2):159–170
36. Rosenzweig C, Hillel D (2008) Climate variability and the global harvest. Oxford University Press, Oxford
37. Rosenzweig C, Parry ML (1994) Potential impacts of climate change on world food supply. Nature 367:133–138
38. Rosenzweig C, Tubiello F (2007) The interactions of adaptions and mitigation strategies in agriculture. Mit Adapt Strategies Glob Change 12(5):855–873
39. Rosenzweig C, Iglesias A, Yang XB, Epstein PR, Chivian E (2000) Implications of climate change for US agriculture: Extreme weather events, plant diseases, and pests. Center for Health and the Global Environment, Harvard Medical School. Cambridge, 56 pp
40. Rosenzweig C, Iglesias A, Yang XB, Epstein PR, Chivian E (2000) Climate change and extreme weather events: Implications for food production, plant diseases, and pests. Global Change and Human Health 2(2)
41. Rosenzweig C, Tubiello FN, Goldberg R, Mills E, Bloomfield J (2002) Increased crop damage in the US from excess precipitation under climate change. Glob Environ Chang 12:197–202
42. Rosenzweig C, Strzepek KM, Major DC, Iglesias A, Yates DN, McClusky A, Hillel D (2004) Water resources for agriculture in a changing climate: International case studies. Glob Environ Chang 14:345–360
43. Roy KW, Rupe JC, Hershman DE, Abney TS (1997) Sudden death syndrome. Plant Dis 81:1100–1111
44. Salinari F, Giosue S, Tubiello FN, Rettori A, Rossi V, Spanna F, Rosenzweig C, Gullino ML (2006) Downy mildew (Plasmopara viticola) epidemics on grapevine under climate change. Glob Chang Biol 12(7):1299–1307
45. Smit B, Burton I, Klein RJT et al (2000) An anatomy of adaptation to climate change and variability. Climate Chang 45(1):223–251
46. Tillman D, Hill J, Lehman C (2006) Carbon-negative biofuels from low-input high-diversity grassland biomass. Science 314:1598–1600
47. Tubiello FN, Amthor JS, Boote KJ, Donatelli M, Easterling W, Fischer G, Gifford RM, Howden M, Reilly J, Rosenzweig C (2006) Crop response to elevated CO2 and world food supply – A comment on "Food for Thought …" by Long et al. Science 312:1918–1921:2006. Eur J Agron 26(3):215–223 APR 2007
48. National Assessment Synthesis Team (2001) Climatic Change Impacts on the US: The potential consequences of climate

variability and change. US Global Change Research Program, Washington DC
49. Van Duivenbooden N, Abdoussalam S, Mohamed AB (2002) Impact of climate change on agricultural production in the Sahel. Part 2. Case study for groundnut and cowpea in Niger. Clim Chang 54:349–368
50. Watson RT, Noble IR, Bolin B, Ravindranath NH, Verando DJ, Dokken DJ (2000) IPCC special reports. Land use, land-use change, and forestry. Cambridge Univ Press, Cambridge, 324
51. West TO, Marland G (2002) A synthesis of carbon sequestration, carbon emissions, and net carbon flux in agriculture: Comparing tillage practices in the United States. Agr Ecosyst Environ 91:217–232
52. Yang XB, Scherm H (1997) El Niño and infectious disease. Science 275:739

Climate Change, Economic Costs of

Richard S. J. Tol[1,2,3,4]
[1] Economic and Social Research Institute, Dublin, Ireland
[2] Institute for Environmental Studies, Vrije Universiteit, Amsterdam, The Netherlands
[3] Department of Spatial Economics, Vrije Universiteit, Amsterdam, The Netherlands
[4] Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Issues
Total Costs
Marginal Costs
Policy Implications
Future Directions
Acknowledgments
Bibliography

Glossary

Direct costs The direct cost equals quantity times price.

Discount factor The discount factor t years into the future equals one over one plus the discount rate, raised to the power t.

Discount rate The discount rate is the annual rate of decline of the value of consumption. It is roughly equal to the rate of interest, or the opportunity cost of capital. The discount rate consists of two components: the rate at which consumers get better off, and the rate of pure time preference (or impatience). The discount rate is not related to the rate of inflation, which is the annual rate of decline of the value of money.

Economics Economics is the social science which studies human behavior with regard to the relationship between ends and scarce means which have alternative uses.

Equity weights Equity weights are applied to aggregate national impacts to a global total. Equity weights are often set to unity, but sometimes equity weights equal the ratio of nationally average per capita income over global average per capita income.

Indirect costs The indirect costs equal all costs that are not direct costs. This includes the price change induced by the change in quantity (partial equilibrium effects), the changes in other markets (general equilibrium effects), and the changes at later times (dynamic effects).

Marginal costs The marginal cost of greenhouse gas emissions equals the first partial derivative of the net present value of the total costs of climate change to emissions.

Monetary valuation Monetary valuation is a set of techniques and their application that attempts to express in monetary terms the value to humans of changes in environmental goods and services. Negative impacts are typically expressed as an income loss that would give an equivalent loss in welfare.

Neo-classical economics Although historians would refer to neo-classical economics as the dominant form of economic research between 1860 and 1910, common usage has neo-classical economics as a synonym for mainstream or orthodox economics. In that sense, neo-classical economics is a style of research, characterized by empirical rigour, mathematical rigour, and micro-founded macro-relationships.

Net present value The net present value is the sum of all future costs and benefits, weighted by the discount factor.

Total costs The total cost of climate change equals the direct and the indirect costs of climate change, that is, the difference in welfare between a scenario with climate and a scenario without.

Definition of the Subject

The economic costs of climate change include all positive and negative impacts of the enhanced greenhouse effect and the resulting changes in the atmosphere and ocean on all human consumers and producers. Total costs refer to the difference in human welfare between a scenario with climate change and a scenario without climate change.

Marginal costs refer to the difference in human welfare between two scenarios with a slightly different climate, normalized by the amount of greenhouse gas emissions that would induce that difference. Estimates of the economic costs of climate change are important to assess the size of the climate problem relative to other problems, and to compare the costs of climate change to the costs of greenhouse gas emission reduction.

Introduction

Calls for greenhouse gas emission reduction are often phrased as a moral imperative. While tempting, this is wrong. Firstly, there is no moral agreement. Emission reduction could save polar bears but it would cost coal miner's jobs and raise the price of food for the malnourished. Moral imperatives are easy if a policy has only benefits. As soon as a policy has both costs and benefits, one has to make trade-offs and choose the lesser evil. Secondly, there is no avoiding dangerous interference with the climate system. Emission reduction would slow down the melting of the Greenland ice cap, and reduce the probability of a collapse of the West-Antarctic Ice Sheet – but it would not stop the melting or bring the change to zero. Thirdly, we have no obligations to future generations or poor people. Such duties are self-imposed. And even if we choose to help others, there are many ways to do this. Would our grandchildren prefer a richer but warmer world, or a poorer but colder one? Would the grandchildren of the Bangladeshis like us to reduce greenhouse gas emissions, help them to adapt to climate change, or help their grandparents grow rich?

This chapter looks into such questions. It is written from the thoroughly relativistic perspective of a neo-classical economist. The basic principles of the economic theory of climate change are quite simple. Greenhouse gas emissions are an externality. Externalities are unintended – we burn coal to make electricity, not to emit carbon dioxide – and uncompensated – carbon dioxide is freely dumped in the atmosphere – consequences of economic activity. Externalities should be internalized, that is, emitters should pay for their emissions. The price of emissions should equal the damage done by the emissions. That is all.

There is now a vast literature on the economics of climate change – started by Nordhaus [73,74]. A large part of that literature is about the deviation between the simple policy prescription of economic theory and the complexities of actual policy. Another large part of the literature is about the costs of greenhouse gas emission reduction. The literature on the economic costs of climate change is only a small one. It is reviewed here.

Section "Issues" discusses the methodological, conceptual, and moral issues one has to confront when estimating the economic impacts of climate change. Section "Total Costs" reviews estimates of the total economic impact. Section "Marginal Costs" surveys estimates of the marginal impacts. Section "Policy Implications" concludes by assessing the policy implications.

Issues

Scenarios

A scenario is a set of assumptions on future conditions that is coherent, internally consistent, and not implausible [70]. Climate scenarios are usually derived from modeling experiments with General Circulation Models (GCM). Climate scenarios include simple statistics such as the global mean surface air temperature, and complex results such as spatial patterns of rainfall extremes. Climate scenarios may also include low-probability events, such as a disruption of the thermohaline circulation in the Atlantic Ocean, or the collapse of the West Antarctic Ice Sheet.

Scenarios also include population, economic activity, greenhouse gas emissions, and land use. Besides driving the climate models, these components are also important as they determine the vulnerability of social and economic systems to climate change over time. Although poorer societies are generally believed to be more vulnerable to climate change, this is by no means a simple relationship. Some impacts tend to fall with economic growth. The impacts of climate change on infectious diseases are a prime example. Malaria does not kill middle income people, because they can afford prevention and (if necessary) cure. Some impacts tend to rise with economic growth. The impact of climate change on biodiversity and species loss is one example. People tend to care more about these matters as their income grows, and further developed economies put more pressure on nature. Other impacts may rise first and then fall with economic growth. Urban air quality is one example. The very poor have nothing to foul the air with, and the very rich do not like foul air and have the wherewithal to prevent it. Climate change is likely to increase malaria, reduce biodiversity and worsen air quality. In some case, climate change and economic growth work together to increase the impacts, while in other cases they pull in opposite directions.

Valuation Approach

There are various techniques for the monetary valuation of climate change impacts. Some values of impacts are directly based on observed prices. Agriculture and dike building are examples. Other values can be indirectly mea-

sured on the basis of observed market prices for surrogate products or services. One example is human health, which is not traded directly but indirectly through safety measures, labor markets, and health care. The challenge in these instances is to model future market prices that are consistent with the underlying socioeconomic scenario. For yet other impacts, no market values exist, and hypothetical prices are needs. Notable impacts are on non-commercial ecosystems and biodiversity.

Because it is practically impossible to estimate each exposure-response relationship or value at the respective geographical location of a climate change impact, data from previous studies focusing on different locations and different policy contexts are inevitable. Furthermore, most climate change impacts will take place in the future, for which by definition no data are available. Therefore it is important to know when data from other studies can be used and under what conditions, and how to extrapolate values from today to tomorrow.

The majority of recent studies still adopt benefit transfer methods for the evaluation of climate impacts. However, benefit transfer is not very reliable [91]. For this reason, more attention should be given to original valuation research in the context of climate change.

An example of such a study is Li et al. [54] who analyze the willingness-to-pay (WTP) of American citizens for climate policy by means of the contingent valuation method. They find that the median American citizen is willing to pay about $ 15/tC. Berrens et al. [7] find a willingness-to-pay between $ 200 and $ 1760 per US household per year (0.2–2.3% of income) for US ratification of the Kyoto Protocol. (Manne and Richels [60] estimate that the costs of US ratification would be 0.75% of GDP in 2010.) Hersch and Viscusi [46] find that Europeans are willing to pay up to 3.7% more for petrol if that helps combat climate change. Viscusi and Zeckhauser [117] find that Harvard students are willing to pay $ 0.50/gallon (a 25% price increase) or 3% of their expected annual income for greenhouse gas emission reduction. (This study also showed that these students underestimate projected warming in Boston by about 50%, while the authors made them believe that carbon dioxide emission reduction would be effective for slowing climate change in the next 30 years.) There is scope for similar applications of WTP techniques, mainly to account for spatial and socio-economic differences in individuals' preferences.

Direct and Higher Order Impacts

Most studies to date have estimated the direct costs of climate change. Direct costs equal the physical change (e. g., the dikes to be reinforced) times their price (in this example, the costs per dike length and dike height). Direct costs are easy to compute (but see Subsect. "Valuation Approach"), but probably underestimate the real economic costs.

The higher order impacts come in three kinds. Firstly, climate change may impact the market under consideration. For example, if dikes are being reinforced everywhere, then the costs of dike building is likely to go up as materials, machinery, and skilled labour is difficult to get. Secondly, the impact of climate change on one market may spill over into other markets. For example, dike building may increase the costs of construction, as the same materials and skills are used. Dike building is capital intensive and may drive up the interest rate. Thirdly, the impact of climate change may affect economic growth. For example, money invested in dike building is not invested elsewhere.

A number of recent studies have examined the economy-wide implications of sea level rise [10], tourism [8], and health [9]. While it is perhaps too early to draw firm conclusions from this body of research, the studies suggest that the indirect effects of climate change impacts can both enlarge and diminish the direct economic impacts of climate change. The distribution of gains and losses is another difference between direct costs and general equilibrium effects. Whereas direct costs are limited to those directly affected, markets would spread the impact to their suppliers, clients, and competitors as well as to financial markets.

Fankhauser and Tol [35] show that the economic growth impact of climate change is as large as the direct impact of climate change. Acemoglu et al. [1] and Masters and McMillan [62] show that differences in climate explain part of observed differences in economic development. Easterly and Levine [28] show that the link is at most weak and indirect, and it is not clear whether the mechanisms that may have been active in the past, still hold for present and future.

Adaptation

One cannot study the costs of climate change impact without also studying, or at least making assumptions about the costs of adaptation [109]. Studies focusing on costs of the impacts make widely differing assumptions about the amount of adaptation that will take place. While some studies completely ignore adaptation, other studies consider arbitrary levels of adaptation, or assume optimal adaptation. No studies use realistic models of adaptation [109]. There is little research that shows how adaptation costs compare to the potential damages of not adapt-

ing. The impacts of climate change and the capacity to adapt would be affected with the level of development and flexibility of the economy [123]. Hence, the future success and nature of adaptation depends on the assumed socio-economic scenario.

Aggregation: Temporal

Climate change is a slow process. Today's emissions will affect the climate for decades to centuries, and sea level for centuries to millennia. As cause and effect are separated in time, so are costs (of emission reduction) and benefits (of avoided climate change). The procedure to make commensurate costs and benefits at different points in time is called discounting. Discounting is as common as it is controversial. See [3] for an excellent discussion.

Individuals discount future gains or losses because of two reasons. (People may also discount the future because it is more uncertain than the present, but in this case discounting is used as a shortcut for an uncertainty analysis.) First, money earns interest. Second, people are impatient. The first reason is widely accepted. Davidson [26] is one of the few exceptions. On the second reason, there is virtual consensus too. All ethical arguments show that people should not discount (e. g., [12]). All empirical evidence shows that people do nonetheless (e. g., [79,80]).

Climate change is a large-scale problem. Therefore, the discount rate of society is more relevant than the individual discount rate. The appropriate measure of the growth rate of money is the average growth rate of per capita consumption. Again, there is little dispute on this. But should the social rate of discount also include a measure of impatience? Again, philosophers agree: Impatience is immoral. However, this implies that a government would deviate from the will of the people. This may be defended with the argument that the government is the guardian of future, yet unborn people. However, the empirical evidence is clear in this case too: Governments are impatient [31].

Discounting is more profound over long periods than over short ones. Discounting implies that climate change damages that occur in a century or so are largely irrelevant. This realization has led people to rethink the fundamental principles of discounting, particularly

(a) the notion that the procedure of discounting results from the intertemporal allocation of resources of an individual agent; and
(b) the assumption that discounting is exponential.

To start with the individual perspective, Lind [56] and Lind and Schuler [55] argue that *earmarked* investment is a crucial assumption in discounting. The discount factor measures the trade-off between consumption now and consumption later, where consumption later is contingent on a specific investment plan. As the current generation cannot commit near-future generations to maintain their investments for the benefit of far-future generations, discounting breaks down between generations. Schelling [95] agrees. The alternative is to decide explicitly on the resource allocation between generations. Chichilnisky [19] shows that discounting coincides with a dictatorship of the present generation over future generations. Gerlagh and Keyzer [38] show that discounting is equivalent to the present generation owning all future resources. This is objectionable from a moral standpoint, but it is reality. This line of research has not led to practical alternatives to discounting.

Conventional discounting is exponential: The discount factor is $(1+r)^{-t}$, where r is the discount rate and t is time. Some people argue that the functional specification of conventional discounting is wrong. The first component is empirical. Conventional exponential discounting has that the relative difference between two years is always equal, regardless of their distance from the present. That is, the difference between year 10 and 11 is the same as the distance between year 100 and 101. However, many people would in fact argue that the difference between year 10 and 11 is equal to the difference between year 100 and 110. Such hyperbolic discounting [22] is very similar to exponential discounting for short periods, but the difference is substantial for long periods. The similarity between exponential and hyperbolic discounting in the short run is important, because a switch to hyperbolic discounting would imply a drastic overhaul of long-term decisions only.

There are two further arguments for hyperbolic discounting cf. Dasgupta and Maskin [25]. The first is due to Weitzman [118]. He shows that, if one is uncertain what discount rate to use, then the lowest discount rate becomes increasingly dominant over time. The certainty-equivalent discount rate falls with time, and the difference between years shrinks in the more distant future. Consider the following example. After one year, the average of a 1% and a 10% discount rate is

$$1-\left(\frac{\left(1.01^{-1}+1.10^{-1}\right)}{2}\right)^{1/1}=5.0\,\%$$

(and not 5.5%). After 100 years,

$$1-\left(\frac{\left(1.01^{-100}+1.10^{-100}\right)}{2}\right)^{1/100}=1.7\,\%\,.$$

That is, the average approaches the minimum as time pro-

gresses. One may criticize this as a short cut for a full uncertainty analysis. However, Gollier [39,40] shows that the same is true if a government somehow aggregates the individual discount rates of its citizens. In the long run, the preferences of the person with the lowest discount rate become increasingly important, and the discount rate declines over time.

Guo et al. [41] and Newell and Pizer [71] show that hyperbolic discounting leads to higher estimates of the social cost of carbon. However, the quantitative effect is limited by the fact that hyperbolic discount rates are high for the first decades.

Aggregation: Spatial

Climate change is a global problem. Carbon dioxide and other greenhouse gases mix uniformly in the atmosphere. This implies most of the impacts of one country's emissions fall on other countries. The same is true for the benefits of emission reduction. The impacts on different countries need to be aggregated somehow.

Two methods dominate the literature. In the first and oldest method, regional impacts are quantified in local currencies, converted to dollars, say, and added up [33,103]. This is simple, but the disadvantage is that similar impacts are treated differently. Most disturbingly, climate-change-induced deaths in rich countries receive a greater weight than climate-change-induced deaths in poor countries. The second method, known as equity weighing, corrects for this [6,36]. Rather than simply adding regional estimates, the regional utility-equivalents are added and then converted back to money according to an assumed global welfare function. A big disadvantage of this method is that climate-change-induced deaths are treated differently than deaths by other, national causes. The reason for this discrepancy is that equity weighing, as practiced in the literature, explicitly assumes a global decision maker.

In the meta-analysis of Tol [107], the median estimate of the marginal damage costs of carbon dioxide is \$ 10/tC without equity weights, and \$ 54/tC with equity weights. So, equity weighing is obviously important. The reason is simple. Poor countries are more vulnerable to climate change. Poor countries have little economic weight. Equity weights correct for this.

Morally, this may be the right thing to do. However, national governments also have a certain obligation to defend the interests of their citizens. A narrow interpretation of self-interest would suggest that impacts abroad be ignored (unless they spill over, e. g., through international migration). Then, climate change policy would be very limited, as most impacts will be abroad. However, the principle of good neighborhood is well established, both morally and legally. This entails that one should avoid doing harm to others; and should pay compensation if harm is done nonetheless (e. g., [113]).

A rational actor would avoid doing harm if that is cheaper than the compensation paid. From a national perspective, the relevant damages are then the impacts on the own country plus the compensation paid to other countries. Schelling [93] forcefully argues that compensation should equal the welfare loss of the victim rather than the welfare loss that the culprit would have experienced had she been the victim. This argues for aggregation of monetized impact estimates without equity weighing.

However, compensation would need to be paid only once. Furthermore, a country would also reasonably expect to be compensated itself. This implies that the damage to a country equals the global damage times its share in causing the problem. Defining the latter is a thorny issue, as the cause-effect chain is long, complex, and uncertain. One would need to make arbitrary decisions on cause, effect and their connection.

Uncertainty

Climate change is plagued by uncertainty [16]. Partly, this is because our understanding of climate change and its impacts is incomplete. For the larger part, however, this is because climate change will take place in the future, partly driven by future emissions, and impacting a future world. Future research and observations may reduce the uncertainty, although surprises may increase the uncertainty just as well, but uncertainty will never disappear. Learning and irreversibility play a crucial role in how to deal with uncertainty. Events that may or may not occur in some distant future, but whose consequences can be alleviated once it becomes clear if they would occur, should not worry us too much. On the other hand, if an effect is irreversible (e. g., species extinction), we may want to prevent it regardless of how uncertain it is and regardless of what future research will show (according to the "precautionary principle"). Another crucial part of dealing with uncertainty is risk aversion. Essentially, this determines how much weight we place on negative surprises. A risk neutral decision maker would cancel negative surprises against positive ones, but a risk adverse decision maker would not. Recent work has shown that the marginal damage costs of carbon dioxide are indeed very sensitive to the assumed degree of risk aversion. Although uncertainty and risk are often emphasized – often in a casual way – only few studies seek to quantify its implications (e. g., [51]).

In a recent paper, Weitzman [120] shows that, under a wide range of standard assumptions, the uncertainty about climate change is so large that the expected value of the social costs of climate change is infinite. Earlier, Tol [106] showed this for a specific model. This implies that uncertainty should take central stage in the analysis of climate policy. The Weitzman result throws up a number of methodological issues that will need to be resolved before the policy implications of this work become clear.

Completeness

The impacts of climate change that have been quantified and monetized include the impacts on agriculture and forestry, water resources, coastal zones, energy consumption, air quality, and human health. Obviously, this list is incomplete. Also within each impact category, the assessment is incomplete. Studies of the impacts of sea level rise on coastal zones, for instance, typically omit saltwater intrusion in groundwater [72]. Furthermore, studies typically compare the situations before and after climate change, but ignore that there will be a substantial period during which adaptation is suboptimal – the costs of this are not known.

Some of the missing impacts are most likely negative. Diarrhoea impacts have been quantified recently [57]. Like malaria, diarrhoea is a disease that is driven by poverty but sensitive to climate. Including diarrhoea tightens the link between development and climate policy. Increasing water temperatures would increase the costs of cooling power plants [101]. Redesigning urban water management systems, be it for more or less water, would be costly [4], as would implementing the safeguards against the increased uncertainty about future circumstances. Roads and bridges would suffer from weather conditions for which they were not designed; this would imply either disruption of traffic or expensive retrofits. Extratropical storms may well increase, leading to greater damage and higher building standards [27]. Expenditures on these things are relatively small. Even if climate change would double or triple the cost, the impact would be small. Ocean acidification would reduce marine biodiversity, and may well harm fisheries [52]. Ocean fisheries are only a small, and declining fraction of GDP, while there are ready substitutes for wild fish protein (notably fish farming). The value of biodiversity is unclear (see below).

Other missing impacts are probably positive. Higher wind speeds in the mid-latitudes would decrease the costs of wind and wave energy [11,44]. Less sea ice would improve the accessibility of arctic harbours, would reduce the costs of exploitation of oil and minerals in the Arctic, and may even open up new transport routes between Europe and East Asia [121]. Warmer weather would reduce expenditures on clothing and food, and traffic disruptions due to snow and ice [15]. Also in these cases, the impact of climate change is likely to be small relative to the economy.

Some missing impacts are positive in some places, and negative in others. Tourism is an example. Climate change may well drive summer tourists towards the poles and up the mountains [42,43]. People, however, are unlikely to change the time and money spent on holiday making. The effect is a redistribution of tourist revenue [8]. The global impact is close to zero, but regional impacts are measured in tens of billions of dollars – positive in temperate, rich countries, and negative in tropical, poor countries. This exacerbates the already skewed distribution of climate impacts. Some ski resorts may go out of business, and others would need expensive snowmaking equipment [29,97]. Other ski resorts would profit from the reduced competition. Although regional impacts may be substantial, at the global scale positives and negatives cancel.

Other impacts are simply not known. Some rivers may see an increase in flooding, and others a decrease [53]. At the moment, only a limited number of rivers have been studied in detail, and it is unclear how to extrapolate to other rivers. It is clear though, that land use and water management may greatly increase or reduce impacts. Although river floods wreak substantial havoc and damages of a single event can reach substantial numbers, average flood damage is in fact small relative to the economy [112]. Tropical storms do more damage, although a substantial share of the impact is due to bad planning rather than bad weather [14]. Nonetheless, tropical storms may prevent capital accumulation and the plantation of lucrative crops such as banana [30,69]. Unfortunately, it is not known how climate change would alter the frequency, intensity, and spread of tropical storms [63,89].

The missing impacts discussed above are probably small. There are also bigger gaps in the coverage of climate change impact studies. Climate change is likely to have a profound impact on biodiversity, but quantitative predictions are rare [13]. Although the economic impact of a small change in biodiversity is known to be small [88], the value of large biodiversity changes is unknown but could well be substantial [18]. There is a small but unknown chance that climate change will be more dramatic than is typically assumed in the impacts literature. This may be because of shutdown of the thermohaline circulation [61], a collapse of the Greenland or West-Antarctic Ice Sheet [84], or a release of large amounts of methane [45]. The economic analysis of such scenarios has only just begun [57]. It may be that climate change

would lead to large-scale migration [64] and violent conflict, although there is only weak empirical support for this [49,124]. Finally, climate change impact studies stop at the end of the 21st century. In 2100, impacts are negative, and getting more negative at an accelerating pace. It is not known how rapidly things would get worse in the 22nd century without emission abatement.

Although the sign of the aggregate unknown impacts is not known, risk aversion would lead one to conclude that greenhouse gas emission reduction should be more stringent than suggested by a cost-benefit analysis based on the quantified impacts only. However, the size of the bias is unknown too – so the main policy implication is that more research is needed.

Total Costs

The first studies of the welfare impacts of climate change were done for the USA [21,74,98,102]. Although Nordhaus [74] (see also Ayres and Walter [5]) extrapolated his US estimate to the world, the credit for the first serious study of the global welfare impacts goes to Fankhauser [32,33], although Hohmeyer and Gaertner [48] earlier published some low quality estimates. Other global estimates include those by Nordhaus [76,77], Tol [103], Nordhaus and Yang [82], Plambeck and Hope [90], Nordhaus and Boyer [81], Mendelsohn et al. [66,68], Tol [105], Maddison [59], Hope [50], Rehdanz and Maddison [92] and Nordhaus [78]. Note that Stern et al. [100] is based on Hope [50].

This is a rather short list of studies, and an even shorter list of authors. This problem is worse if one considers that Nordhaus and Mendelsohn are colleagues; that Fankhauser, Maddison and Tol are students of Pearce; and that Rehdanz is a student of Maddison and Tol; while Hope's (and Stern's) estimates are averages of Fankhauser's and Tol's. Although most fields are dominated by a few people, dominance is here for want of challengers. The effect of this is hard to gauge. The reasons are lack of funding (this work is too applied for academic sources, while applied agencies do not like the typical results and pre-empt this by not funding it), lack of daring (this research requires making many assumptions, and taking on well-entrenched incumbents), and lack of reward (the economics profession frowns on the required interdisciplinarity). In addition, many people, including many economists, would argue that climate change is beyond cost-benefit analysis and that monetary valuation is unethical.

Table 1 shows some characteristics of these studies. A few insights emerge. First, the welfare impact of a doubling of the atmospheric concentration of carbon dioxide on the current economy is relatively small. Although the estimates differ, impacts are not more than a few percent of GDP. The estimates of Hope [50], Mendelsohn et al. [66,68] and Tol [105] even point to initial benefits of climate change. (Studies published after 1995 all have regions with net gains and net losses due to global warming, whereas earlier studies only find net losses.) With such estimates, it is no surprise that cost-benefit analyses of climate change recommend only limited greenhouse gas emission reduction – for instance, Nordhaus [75] argues that the optimal rate of emission reduction is 10–15%, one of the more contentious findings of the climate economics literature.

Second, although the impact is relatively small, it is not negligible. A few per cent of GDP in annual damage is a real concern.

Third, climate change may initially have positive impacts. This is partly because the higher ambient concentration of carbon dioxide would reduce water stress in plants and may make them grow faster – although this effect is now believed to be weaker [58]. Another reason is that the global economy is concentrated in the temperate zone, where a bit of warming may well be welcomed because of reductions in heating costs and cold-related health problems. At the same time, the world population is concentrated in the tropics, where the impacts of initial climate change are probably negative. Even though initial *economic* impacts are positive, it does not necessarily follow that greenhouse gas emissions should be subsidized. The climate responds rather slowly to changes in emissions, so the initial impacts cannot be avoided. Impacts start falling – that is, additional climate change reduces global welfare – roughly at the same time as climate change can be influenced by present and future emission reduction [47].

The fourth insight is that relative impacts are higher in poorer countries (see also Yohe and Schlesinger [122]). This is because poorer countries have a lower adaptive capacity [2], particularly in health [108], and have a greater exposure to climate change, particularly in agriculture and water resources. Furthermore, poorer countries tend to be hotter and therefore closer to temperature limits and short on spatial analogues should it get warmer still. At the same time, there are fewer studies on the impacts of climate change on developing countries than on developed countries. Although research is scarce [83], there is little reason to assume that climate change impacts would be homogeneous within countries; certainly, certain economic sectors (e. g., agriculture), regions (e. g., the coastal zone) and age groups (e. g., the elderly) are more heavily affected than

Climate Change, Economic Costs of, Table 1
Econonic impact estimates of climate change; numbers in brackets are either standard deviations or confidence intervals

Study	Warming	Impact	Minimum	Region	Maximum	Region
Nordhaus [76]	3.0	−1.3				
Nordhaus [77]	3.0	−4.8 (−30.0 to 0.0)				
Fankhauser [33]	2.5	−1.4	−4.7	China	−0.7	Eastern Europe and the former Soviet Union
Tol [103]	2.5	−1.9	−8.7	Africa	−0.3	Eastern Europe and the former Soviet Union
Nordhaus and Yang [82][a]	2.5	−1.7	−2.1	Developing countries	0.9	Former Soviet Union
Plambeck and Hope [60][a]	2.5	−2.5 (−0.5 to −11.4)	−8.6 (−0.6 to −39.5)	Asia (w/o China)	0.0 (−0.2 to 1.5)	Eastern Europe and the former Soviet Union
Mendelsohn et al. [66][a,b,c]	2.5	0.0 0.1	−3.6 −0.5	Africa	4.0 1.7	Eastern Europe and the former Soviet Union
Nordhaus and Boyer [81]	2.5	−1.5	−3.9	Africa	0.7	Russia
Tol [105]	1.0	2.3 (1.0)	−4.1 (2.2)	Africa	3.7 (2.2)	Western Europe
Maddison [59][a,d,e]	2.5	−0.1	−14.6	South America	2.5	Western Europe
Rehdanz and Maddison [92][a,c]	1.0	−0.4	−23.5	Sub-Saharan Africa	12.9	South Asia
Hope [50] [a]	2.5	0.9 (−0.2 to 2.7)	−2.6 (−0.4 to 10.0)	Asia (w/o China)	0.3 (−2.5 to 0.5)	Eastern Europe and the former Soviet Union
Nordhaus [78]	2.5	−0.9 (0.1)				

[a]Note that the global results were aggregated by the current author.
[b]The top estimate is for the "experimental" model, the bottom estimate for the "cross-sectional" model.
[c]Note that Mendelsohn et al. only include market impacts.
[d]Note that the national results were aggregated to regions by the current author for reasons of comparability.
[e]Note that Maddison only considers market impacts on households.

others. This has two policy implications. Firstly, recall that greenhouse gas mix uniformly in the atmosphere. It does not matter where they are emitted or by whom, the effect on climate change is the same. Therefore, any justification of stringent emission abatement is an appeal to consider the plight of the poor and the impacts imposed on them by the rich [94,95]. While this makes for wonderful rhetoric and fascinating research (e. g., [104]), reality shows little compassion for the poor by the rich. Secondly, if poverty is the root cause for vulnerability to climate change, one may wonder whether stimulating economic growth or emission abatement is the better way to reduce impacts. Indeed, Tol and Yohe [115] argue that the economic growth foregone by stringent abatement more than offsets the avoided impacts of climate change, at least for malaria, while Tol [108] shows that development is a cheaper way of reducing climate-change-induced malaria than is emission reduction. Moreover, richer countries may find it easier and cheaper to compensate poorer countries for the climate change damages caused, than to reduce greenhouse gas emissions. Such compensation may be explicit and financial, but would more likely take the shape of technical and financial assistance with adaptation (cf. [85]).

The agreement between the studies is remarkable if one considers the diversity in methods. The studies of Fankhauser, Hope, Nordhaus, and Tol all use the enumerative method: 'physical' impact estimates are obtained one by one, from 'natural science' papers based on 'process-based' models or 'laboratory experiments'. These physical impacts are multiplied with their respective prices, and added up. The 'prices' are obtained by benefit transfer. In contrast, Mendelsohn's work is based on direct, empirical estimates of the welfare impacts, using observed variations in prices and expenditures to discern the effect of climate (e. g., [67]). Mendelsohn estimates are done per sector and then added up, but physical modelling and benefit transfer are avoided. Nordhaus [78] uses empirical estimates of the *aggregate* climate impact on income, while Maddison [59] looks at patterns of *aggregate* household consumption. Like Mendelsohn, Nordhaus and Maddison rely exclusively on observations, but they assume that all climate effects are aggregated by the economy into

incomes and expenditures. Rehdanz and Maddison [92] also empirically estimate the aggregate impact, but use self-reported happiness as an indicator; their approach is similar to that of Nordhaus and Maddison, but the indicator is subjective rather than objective. The enumerative studies of Fankhauser etc rely on controlled experiments (albeit with detailed, process-based models in most cases). This has the advantages of ease of interpretation and physical realism, but the main disadvantage is that certain things are kept constant that would change in reality; adaptation is probably the key element. The statistical studies of Mendelsohn etc rely on uncontrolled experiments. This has the advantage that everything varies as in reality, but the disadvantages are that the assessment is limited to observed variations (which may be small compared to projected changes, particularly in the case of carbon dioxide concentration) and that effects may be spuriously attributed to climate. Therefore, the variety of methods enhances confidence, not in the individual estimates, but in the average.

The shortcomings of the estimates are at least as interesting. Welfare losses are approximated with direct costs, ignoring general equilibrium and even partial equilibrium effects (see below). In the enumerative studies, impacts are assessed independently of one another, even if there is an obvious overlap as between water resources and agriculture. Estimates are often based on extrapolation from a few detailed case studies, and extrapolation is to climate and levels of development that are very different from the original case study. Valuation is based on benefit transfer, driven only by difference in per capita income. Realistic modelling of adaptation is problematic, and studies either assume no adaptation or perfect adaptation. Many impacts are unquantified, and some of these may be large (see below). The uncertainties are unknown – only 4 of the 14 estimates in Table 1 have some estimate of uncertainty. These problems are gradually solved, but progress is slow. Indeed, the above list of caveats is similar to that in Fankhauser and Tol [34].

Marginal Costs

Although the number of studies of the *total* costs of climate change is small, a larger number of studies estimate the *marginal* costs. The marginal damage cost of carbon dioxide is defined as the net present value of the incremental damage due to an infinitesimally small increase in carbon dioxide emissions. If this is computed along the optimal trajectory of emissions, the marginal damage cost equals the Pigou tax. Marginal damage cost estimates derive from total cost estimates – the fact that there are more estimates available, does not imply that we know more about the marginal costs than we do about the total costs. In fact, some of the total cost estimates [59,66,68,78,92] have yet to be used for marginal cost estimation, so that the empirical basis is actually smaller.

Tol [110] gathers 211 estimates of the SCC from 47 studies. The studies were grouped in those that were peer-reviewed and those that were not. Some studies are based on original estimates of the total costs of climate change, while other studies borrow total costs estimates from other studies. Most studies use incremental or marginal calculus to estimate the SCC, as they should, while a few others use average impacts or an unspecified method. Some studies assume that climate changes but society does not, while other studies include a dynamic model of vulnerability. A few studies use entirely arbitrary assumptions about future climate change, while most studies are based on internally consistent scenarios. These classifications are used as quality indicators. More recent studies were given a higher weight. Many studies report multiple estimates. Most of the estimates are sensitivity analyses around a central estimate, and some estimates are only included to (approximately) reproduce an earlier study. Tol [110] introduces additional weights to account for this.

Tol [110] adjusts a Fisher–Tippett kernel density estimator to 211 data points, weighted as describe above. The 211 estimates provide the modes. Only a few of the studies provide an estimate of the uncertainty. Therefore, the standard deviation is set equal to the sample standard deviation.

Table 2 shows selected characteristics of the kernel distribution for the whole sample and selected sub-samples.

Splitting the sample by discount rate used has the expected effect: A higher discount rate implies a lower estimate of the SCC and a thinner tail. Table 2 also shows that estimates in the peer reviewed literature are lower and less uncertain than estimates in the gray literature.

Splitting the sample by publication date, shows that the estimates of the SCC published before AR2 [87] were larger than the estimates published between AR2 and AR3 [99], which in turn were larger than the estimates published since. Note that these differences are not statistically significant if one considers the means and standard deviation. However, the kernel distribution clearly shifts to the left. Therefore, AR4 [96] were incorrect to conclude that the economic estimates of the impact of climate change have *increased* since 2001. In their words (pp. 781): "There is some evidence that initial new market benefits from climate change will peak at a lower magnitude and sooner than was assumed for the TAR, and it is likely that there will be higher damages for larger magni-

Climate Change, Economic Costs of, Table 2

Selected characteristics (mode, mean, standard deviation, median, 90-percentile, 95-percentile, 99-percentile, percentile of the Stern estimate) of the joint probability density of the social cost of carbon for the whole sample (all) and selected subsamples (pure rate of time preference, review process, and publication date)

	All	PRTP			Review		Publication date		
		0%	1%	3%	peer	gray	<1996	96–01	>2001
Mode	35	129	56	14	20	53	36	37	27
Mean	127	317	80	24	71	196	190	120	88
St.Dev.	243	301	70	21	98	345	392	179	121
Median	74	265	72	21	48	106	88	75	62
90%	267	722	171	51	170	470	397	274	196
95%	453	856	204	61	231	820	1555	482	263
99%	1655	1152	276	82	524	1771	1826	867	627
Stern	0.92	0.56	1.00	1.00	0.97	0.84	0.86	0.92	0.96

tudes of global mean temperature increases than was estimated in the TAR." It is unclear how Schneider et al. [96] reached this conclusion, but it is not supported by the data presented here.

The SCC estimate by Stern et al. [100] is almost an outlier in the entire sample (excluding, of course, the Stern estimate itself). Depending on the kernel density, the Stern estimate lies between the 90th and the 94th percentile. It fits in better with estimates that use a low discount rate and were not peer-reviewed – characteristics of the Stern Review – but even in comparison to those studies, Stern et al. [100] are on the high side. The Stern estimate also fits in better with the older studies. This is no surprise, as the PAGE model (e. g., [50]) is calibrated to [87] and [99]. Other criticism of the Stern Review can be found in [24,65,79,80,115,119,120].

Policy Implications

The policy implications of the above findings are several, and not necessarily in line with the conceived wisdom of climate policy. First and foremost, the economic impacts literature points out that climate change is a problem. Initial climate change may be beneficial, but it cannot be avoided. This is a sunk benefit. The avoidable part of climate change is in all likelihood negative. This justifies greenhouse gas emission reduction.

Second, the estimates of the marginal damage costs justify some emission abatement, but not too much. For instance, the future price of carbon dioxide emission permits in the European Trading System is around $ 100/tC. Using the market rate of discount, the expected social cost of carbon is only $ 24/tC. This climate policy has a benefit-cost ratio of 0.24. EU climate policy is therefore too stringent. Of course, European climate policy does pass the cost-benefit test according to CEC [17], but this study does not meet conventional standards of academic quality [111]. Earlier, Pearce [89] similarly concluded that the UK cost-benefit analysis [20] is deficient, while also the *Stern Review* [100] has been criticized in the academic literature (e. g., [80,119]).

Third, climate policy is about ethics rather than about economics [37,116]. The judgment what to do about greenhouse gas emissions rests on the values one attaches to far-flung countries and distant futures. The ethics are not straightforward, however. If one places a lot of weight on the future, one should make a trade-off between increasing investment in capital goods, education, emission reduction, or technology. If one places a lot of weight on people in poor countries, one should make a trade-off between adaptation, development, emission reduction, and trade reform.

Fourth, the uncertainties about the economic impact of climate change are profound. Partly, this is because the subject is complex. A large share of the uncertainty can be explained, however, by the dearth of research funding. Although climate change is often said to be the largest (environmental) challenge of our times, very few researchers are funded to substantiate or refute that claim.

Future Directions

Further research is therefore needed. Several problems with past and present research are identified above. Firstly, research into the economic impact of climate change is rightly classified as "applied research". This implies, however, that research funding comes from bodies with a stake in the result, and that quality and independence are not necessarily overriding concerns. The *Stern Review* is the most prominent example in the recent past of a study that started with the conclusions and worked back to identify the required assumptions. The new *Centre for Climate*

Change Economics and Policy at the *London School of Economics* may fall into the same trap. Because the stakes in climate policy are large, academic quality of research must be guaranteed.

Secondly, there are only two groups of independent academics who study the economic impact of climate change. These groups are not sufficiently funded. More importantly, these groups are rarely challenged. Combined with the first problem, it is therefore important to establish a third group of independent, academic economists to study the impact of climate change.

Thirdly, research on the impact of climate change, economic and otherwise, has been lamp-posting. After the groundbreaking work in the early 1990s, researchers have refined previous estimates. Little attention has been paid to those impacts for which no previous estimates exist. While this is the normal procedure of gradual progress in scientific research, the study of the impact of climate change is still in its formative stages. Not just *more*, but particularly *different* research is needed – into the economic effects of climate change on biodiversity, on violent conflict, on ice shelves and ocean current, and on economic development in the long term.

Acknowledgments

My thoughts on this subject were shaped by discussions with Hadi Dowlatabadi, Tom Downing, Sam Fankhauser, David Maddison, Rob Mendelsohn, Bill Nordhaus, David Pearce, Steve Schneider, Joel Smith, and Gary Yohe.

Bibliography

1. Acemoglu D, Johnson S, Robinson JA (2001) The colonial origins of comparative development: an empirical investigation. Am Econ Rev 91:1369–1401
2. Adger WN (2006) Vulnerability. Glob Environ Chang 16:268–281
3. Arrow KJ, Cline WR, Maeler KG, Munasinghe M, Squitieri R, Stiglitz JE (1996) Intertemporal equity, discounting, and economic efficiency. In: Bruce JP, Lee H, Haites EF (eds) Climate change 1995: economic and social dimensions – contribution of working group iii to the second assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge, pp 125–144
4. Ashley RM, Balmfort DJ, Saul AJ, Blanskby JD (2005) Flooding in the future – predicting climate change, risks and responses in urban areas. Water Sci Technol 52(5):265–273
5. Ayres RU, Walter J (1991) The greenhouse effect: damages, costs and abatement. Environ Resour Econ 1:237–270
6. Azar C, Sterner T (1996) Discounting and distributional considerations in the context of global warming. Ecol Econ 19:169–184
7. Berrens RP, Bohara AK, Jenkins-Smith HC, Silva CL, Weimer DL (2004) Information and effort in contingent valuation surveys: application to global climate change using national internet samples. J Environ Econ Manag 47:331–363
8. Berrittella M, Bigano A, Roson R, Tol RSJ (2006) A general equilibrium analysis of climate change impacts on tourism. Tour Manag 27(5):913–924
9. Bosello F, Roson R, Tol RSJ (2006) Economy-wide estimates of the implications of climate change: human health. Ecol Econ 58:579–591
10. Bosello F, Roson R, Tol RSJ (2007) Economy-wide estimates of the implications of climate change: sea level rise. Environ Resour Econ 37:549–571
11. Breslow PB, Sailor DJ (2002) Vulnerability of wind power resources to climate change in the continental united states. Renew Energy 27(4):585–598
12. Broome J (1992) Counting the cost of global warming. White Horse Press, Cambridge
13. Burkett VR, Wilcox DA, Stottlemyer R, Barrow W, Fagre D, Baron J, Price J, Nielson JL, Allen CD, Peterson DL, Ruggerone G, Doyle T (2005) Nonlinear dynamics in ecosystem response to climate change: case studies and policy implications. Ecol Complex 2(4):357–394
14. Burton I, Kates RW, White GF (1993) The environment as hazard, 2nd edn. The Guilford Press, New York
15. Carmichael CG, Gallus Jr WA, Temeyer BR, Bryden MK (2004) A winter weather index for estimating winter road maintenance costs in the midwest. J Appl Meteorol 43(11):1783–90
16. CBO (2005) Uncertainty in analyzing climate change: policy implications. congress of the united states. Congressional Budget Office, Washington
17. CEC (2005) Winning the battle agains global climate change – background paper. Commission of the European Communities, Brussels
18. Champ PA, Boyle KJ, Brown TC (eds) (2003) A primer on nonmarket valuation. Kluwer, Dordrecht
19. Chichilnisky G (1996) An axiomatic approach to sustainable development. Soc Choice Welf 13(2):219–248
20. Clarkson R, Deyes K (2002) Estimating the social cost of carbon emissions. Working Paper 140. The Public Enquiry Unit – HM Treasury, London
21. Cline WR (1992) The economics of global warming. Institute for International Economics, Washington
22. Cropper ML, Aydede SK, Portney PR (1992) Rates of time preference for saving lives. Am Econ Rev 82(2):469–472
23. Darwin RF (1999) A FARMer's view of the ricardian approach to measuring agricultural effects of climatic change. Clim Chang 41(3–4):371–411
24. Dasgupta P (2007) Commentary: The stern review's economics of climate change. Natl Inst Econ Rev 199:4–7
25. Dasgupta P, Maskin E (2005) Uncertainty and hyperbolic discounting. Am Econ Rev 95(4):1290–1299
26. Davidson MD (2006) A social discount rate for climate damage to future generations based on regulatory law. Clim Chang 76:55–72
27. Dorland C, Tol RSJ, Palutikof JP (1999) Vulnerability of the netherlands and northwest europe to storm damage under climate change. Clim Chang 43:513–535
28. Easterly W, Levine R (2003) Tropics, germs, and crops: how endowments influence economic development. J Monet Econ 50:3–39
29. Elsasser H, Buerki R (2002) Climate change as a threat to tourism in the alps. Clim Res 20(3):253–257

30. Ennos AR (1997) Wind as an ecological factor. Trends Ecol Evol 12(3):108–111
31. Evans DJ, Sezer H (2004) Social discount rates for six major countries. Appl Econ Lett 11:557–560
32. Fankhauser S (1994) The economic costs of global warming damage: a survey. Glob Environ Chang 4(4):301–309
33. Fankhauser S (1995) Valuing climate change – the economics of the greenhouse, 1st edn. EarthScan, London
34. Fankhauser S, Tol RSJ (1996) Recent advancements in the economic assessment of climate change costs. Energy Policy 24(7):665–673
35. Fankhauser S, Tol RSJ (2005) On climate change and economic growth. Resour Energy Econ 27:1–17
36. Fankhauser S, Tol RSJ, Pearce DW (1997) The aggregation of climate change damages: a welfare theoretic approach. Environ Resour Econ 10:249–266
37. Gardiner SM (2006) A perfect moral storm: climate change, intergenerational ethics and the problem of moral corruption. Environ Values 15:397–413
38. Gerlagh R, Keyzer MA (2001) Sustainability and the intergenerational distribution of natural resource entitlements. J Public Econ 79:315–341
39. Gollier C (2002) Discounting an uncertain future. J Public Econ 85:149–166
40. Gollier C (2002) Time horizon and the discount rate. J Econ Theor 107:463–473
41. Guo JK, Hepburn C, Tol RSJ, Anthoff D (2006) Discounting and the social cost of carbon: a closer look at uncertainty. Environ Sci Policy 9(5):203–216
42. Hamilton JM, Maddison DJ, Tol RSJ (2005) Climate change and international tourism: a simulation study. Glob Environ Chang 15(3):253–266
43. Hamilton JM, Maddison DJ, Tol RSJ (2005) The effects of climate change on international tourism. Clim Res 29:255–268
44. Harrison GP, Wallace AR (2005) Sensitivity of wave energy to climate change. IEEE Trans Energy Convers 20(4):870–877
45. Harvey D, Huang Z (1995) Evaluation of the potential impact of methane clathrate destabilization on future global warming. J Geophys Res 100:2905–2926
46. Hersch J, Viscusi WK (2006) The generational divide in support for environmental policies: european evidence. Clim Chang 77:121–136
47. Hitz S, Smith JB (2004) Estimating global impacts from climate change. Glob Environ Chang 14:201–218
48. Hohmeyer O, Gaertner M (1992) The costs of climate change – a rough estimate of orders of magnitude. Fraunhofer-Institut fur Systemtechnik und Innovationsforschung, Karlsruhe
49. Homer-Dixon TF (1994) Environmental scarcities and violent conflict: evidence from cases. Int Secur 19(1):5–40
50. Hope CW (2006) The marginal impact of co2 from PAGE2002: an integrated assessment model incorporating the IPCC's five reasons for concern. Integr Assess J 6(1):19–56
51. Hope CW, Maul P (1996) Valuing the impact of CO_2 emissions. Energy Policy 24(3):211–219
52. Kikkawa T, Kita J, Ishumatsu A (2004) Comparison of the lethal effect of CO_2 and acidification on red sea bream (pagrus major) during the early developmental stages. Marine Pollut Bull 48(1–2):108–110
53. Kundzewicz ZW, Gracyk D, Maurer T, Pinskwar I, Radziejewski M, Svensson C, Szwed M (2005) Trend detection in river flow series: 1, annual maximum flow. Hydrol Sci J 50(5):797–810
54. Li H Berrens RP, Bohara AK, Jenkins-Smith HC, Silva CL, Weimer DL (2004) Exploring the beta model using proportional budget information in a contingent valuation study. Ecol Econ 28:329–343
55. Lind RC, Schuler RE (1998) Equity and discounting in climate change decisions. In: Nordhaus WD (ed) Economics and policy issues in climate change. Resources for the Future, Washington, pp 59–96
56. Lind RC (1995) Intergenerational equity, discounting, and the role of cost-benefit analysis in evaluating global climate policy. Energy Policy 23(4/5):379–389
57. Link PM, Tol RSJ (2004) Possible economic impacts of a shutdown of the thermohaline circulation: an application of FUND. Port Econ J 3:99–114
58. Long SP, Ainsworth EA, Leakey ADB, Noesberger J, Ort DR (2006) Food for thought: lower-than-expected crop yield stimulation with rising CO2 concentrations. Science 312: 1918–1921
59. Maddison DJ (2003) The amenity value of the climate: the household production function approach. Resour Energy Econ 25:155–175
60. Manne AS, Richels RG (2004) US rejection of the kyoto protocol: the impact on compliance costs and CO_2 emissions. Energy Policy 32:447–454
61. Marotzke J (2000) Abrupt climate change and thermohaline circulation: mechanisms and predictability. Proc Natl Acad Sci 97:1347–1350
62. Masters WA, McMillan MS (2001) Climate and scale in economic growth. J Econ Growth 6:167–186
63. McDonald RE, Bleaken DG, Cresswell DR, Pope VD, Senior CA (2005) Tropical storms: representation and diagnosis in climate models and the impacts of climate change. Clim Dyn 25(1):19–36
64. McLeman R, Smit B (2006) Migration as an adaptation to climate change. Clim Chang 76:31–53
65. Mendelsohn RO (2006) A critique of the stern report. Regulation (Winter 2006–2007):42–46
66. Mendelsohn RO, Morrison W, Schlesinger ME, Andronova NG (2000) Country-specific market impacts of climate change. Clim Chang 45:553–569
67. Mendelsohn RO, Nordhaus WD, Shaw D (1994) The impact of climate on agriculture: a ricardian analysis. Am Econ Rev 84(4):753–771
68. Mendelsohn RO, Schlesinger ME, Williams LJ (2000) Comparing impacts across climate models. Int Assess 1:37–48
69. Mulcahy M (2004) Weathering the storms: hurricanes and risk in the british greater caribbean. Bus Hist Rev 78(4):635–663
70. Nakicenovic N, Swart RJ (eds) (2001) IPCC special report on emissions scenarios. Cambrigde University Press, Cambridge
71. Newell RG, Pizer WA (2003) Discounting the distant future: how much do uncertain rates increase valuations? J Environ Econ Manag 46:52–71
72. Nicholls RJ, Tol RSJ (2006) Impacts and responses to sea-level rise: a global analysis of the SRES scenarios over the 21st Century. Phil Trans Royal Soc A Math Phys Eng Sci 361(1841):1073–1095
73. Nordhaus WD (1982) How fast should we graze the global commons? Am Econ Rev 72(2):242–246
74. Nordhaus WD (1991) To Slow or Not to Slow: The economics of the greenhouse effect. Econ J 101:920–937
75. Nordhaus WD (1993) Rolling the 'DICE': An optimal transition

path for controlling greenhouse gases. Resour Energy Econ 15:27–50
76. Nordhaus WD (1994) Managing the global commons: the economics of climate change. The MIT Press, Cambridge
77. Nordhaus WD (1994) Expert opinion on climate change. Am Sci 82(1):45–51
78. Nordhaus WD (2006) Geography and macroeconomics: new data and new findings. Proc Natl Acad Sci 103(10):3510–3517. www.pnas.org/cgi/doi/10.1073/pnas.0509842103
79. Nordhaus WD (2007) Critical assumptions in the stern review on climate change. Science 317:201–202
80. Nordhaus WD (2007) A review of the stern review on the economics of climate change. J Econ Lit 45(3):686–702
81. Nordhaus WD, Boyer JG (2000) Warming the world: economic models of global warming. The MIT Press, Cambridge
82. Nordhaus WD, Yang Z (1996) RICE: A regional dynamic general equilibrium model of optimal climate-change policy. Am Econ Rev 86(4):741–765
83. O'Brien KL, Sygna L, Haugen JE (2004) Vulnerable or resilient? A multi-scale assessment of climate impacts and vulnerability in norway. Clim Chang 64:193–225
84. Oppenheimer M, Alley RB (2005) Ice sheets, global warming, and article 2 of the UNFCCC. Clim Chang 68:257–267
85. Paavola J, Adger WN (2006) Fair adaptation to climate change. Ecol Econ 56:594–609
86. Pearce DW (2003) The social cost of carbon and its policy implications. Oxford Rev Econ Policy 19(3):1–32
87. Pearce DW, Cline WR, Achanta AN, Fankhauser S, Pachauri RK, Tol RSJ, Vellinga P (1996) The social costs of climate change: greenhouse damage and the benefits of control. In: Bruce JP, Lee H, Haites EF (eds) Climate Change 1995: Economic and Social Dimensions – Contribution of Working Group III to the Second Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, pp 179–224
88. Pearce DW, Moran D (1994) The economic value of biodiversity. EarthScan, London
89. Pielke Jr RA, Landsea C, Mayfield M, Laver J, Pasch R (2005) Hurricanes and global warming. Bull Am Meteorol Soc 86(11):1571–1575
90. Plambeck EL, Hope CW (1996) PAGE95 – An updated valuation of the impacts of global warming. Energy Policy 24(9):783–793
91. Ready R, Navrud S, Day B, Dubourg R, Machado F, Mourato S, Spaninks F, Rodriguez MXV (2004) Benefit transfer in europe: How reliable are transfers between countries? Environ Resour Econ 29(1):67–82
92. Rehdanz K, Maddison DJ (2005) Climate and happiness. Ecol Econ 52:111–125
93. Schelling TC (1984) Choice and consequence. Harvard University Press, Cambridge
94. Schelling TC (1992) Some economics of global warming. Am Econ Rev 82:1–14
95. Schelling TC (1995) Intergenerational discounting. Energy Policy 23(4/5):395–401
96. Schneider SH, Semenov S, Patwardhan A, Burton I, Magadya CHD, Oppenheimer M, Pittock AB, Rahman A, Smith JB, Suarez A, Yamin F (2007) Assessing key vulnerability and the risk from climate change. In: Parry ML et al (eds) Climate Change 2007: Impacts, Adaptation and Vulnerability – Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, pp 779–810
97. Scott D, McBoyle G, Mills B (2003) Climate change and the skiing industry in southern ontario (Canada): exploring the importance of snowmaking as a technical adaptation. Clim Res 23:171–181
98. Smith JB (1996) Standardized estimates of climate change damages for the united states. Clim Chang 32(3):313–326
99. Smith JB, Schellnhuber HJ, Mirza MMQ, Fankhauser S, Leemans R, Lin E, Ogallo L, Pittock B, Richels RG, Rosenzweig C, Tol RSJ, Weyant JP, Yohe GW (2001) Vulnerability to climate change and reasons for concern: a synthesis. In: Mccarthy JJ, Canziani OF, Leary NA, Dokken DJ, White KS (eds) Climate change 2001: Impacts, adaptation, and vulnerability, vol 19. Cambridge University Press, Cambridge, pp 913–967
100. Stern NH, Peters S, Bakhshi V, Bowen A, Cameron C, Catovsky S, Crane D, Cruickshank S, Dietz S, Edmonson N, Garbett SL, Hamid L, Hoffman G, Ingram D, Jones B, Patmore N, Radcliffe H, Sathiyarajah R, Stock M, Taylor C, Vernon T, Wanjie H, Zenghelis D (2006) Stern review: The economics of climate change. Cambridge University Press, Cambridge
101. Szolnoky C, Buzas K, Clement A (1997) Impacts of the climate change on the operation of a freshwater cooled electric power plant. Periodica Polytecnica: Civil Eng 41(2):71–94
102. Titus JG (1992) The costs of climate change to the united states. In: Majumdar SK et al (eds) Global climate change: implications, challenges and mitigation measures. Pennsylvania Academy of Science, Easton, pp 384–409
103. Tol RSJ (1995) The damage costs of climate change – towards more comprehensive calculations. Environ Resour Econ 5:353–374
104. Tol RSJ (2001) Equitable cost-benefit analysis of climate change. Ecol Econ 36(1):71–85
105. Tol RSJ (2002) New estimates of the damage costs of climate change, Part I: Benchmark Estimates. Environ Resour Econ 21(1):47–73
106. Tol RSJ (2003) Is the uncertainty about climate change too large for expected cost-benefit analysis? Clim Chang 56(3):265–289
107. Tol RSJ (2005) The marginal damage costs of carbon dioxide emissions: an assessment of the uncertainties. Energy Policy 33(16):2064–2074
108. Tol RSJ (2005) Emission abatement versus development as strategies to reduce vulnerability to climate change: an application of FUND. Environ Dev Econ 10:615–629
109. Tol RSJ (2005) Adaptation and mitigation: trade-offs in substance and methods. Environ Sci Policy 8:572–578
110. Tol RSJ (2007) The social cost of carbon: trends, outliers and catastrophes, research unit sustainability and global change FNU-144. Hamburg University and Centre for Marine and Atmospheric Science, Hamburg
111. Tol RSJ (2007) Europe's long-term climate target: A critical evaluation. Energy Policy 35:424–432
112. Tol RSJ, van der Grijp NM, Olsthoorn AA, van der Werff PE (2003) Adapting to climate change: A case study of riverine flood risks in the netherlands. Risk Analysis 23(3):575–583
113. Tol RSJ, Verheyen R (2004) State responsibility and compensation for climate change damages – a legal and economic assessment. Energy Policy 32:1109–1130
114. Tol RSJ, Yohe GW (2006) Of dangerous climate change and dangerous emission reduction. In: Schellnhuber HJ, Cramer

W, Nakicenovic N, Wigley T, Yohe G (eds) Avoiding dangerous climate change. Cambridge University Press, Cambridge, pp 291–298
115. Tol RSJ, Yohe GW (2006) A review of the stern review. World Econ 7(4):233–250
116. Toman M (2006) Values in the economics of climate change. Environ Values 15:365–379
117. Viscusi WK, Zeckhauser RJ (2006) The perception and valuation of the risks of climate change: A rational and behavioral blend. Clim Chang 77:151–177
118. Weitzman ML (2001) Gamma discounting. Am Econ Rev 91(1):260–271
119. Weitzman ML (2007) A review of the stern review on the economics of climate change. J Econ Lit 45(3):703–724
120. Weitzman ML (2008) On modeling and interpreting the economics of catastrophic climate change. Rev Econ Stat
121. Wilson KJ, Falkingham J, Melling H, de Abreu R (2004) Shipping in the canadian arctic: Other possible climate change scenarios. Int Geosci Remote Sens Symp 3:1853–1856
122. Yohe GW, Schlesinger ME (2002) The economic geography of the impacts of climate change. J Econ Geogr 2:311–341
123. Yohe GW, Tol RSJ (2002) Indicators for social and economic coping capacity – moving towards a working definition of adaptive capacity. Glob Environ Chang 12(1):25–40
124. Zhang DD, Jim CY, Lin GCS, He YQ, Wang JJ, Lee HF (2006) Climatic change, wars and dynastic cycles in China over the last millennium. Clim Chang 76:459–477

Climate Change and Human Health

Hartmut Grassl
Max Planck Institute for Meteorology,
Hamburg, Germany

Article Outline

Glossary
Introduction
Climate Change Impact on Human Health
Climate Change Impacts on Plants
with Consequences for Human Health
Concluding Remarks
Bibliography

Glossary

Health As defined by the World Health Organization (WHO) health is the state of complete physical, mental and social well-being and not merely the absence of disease or infirmity.

Human bioclimate The fundamental issue in human biometeorology is the assessment of the direct health effects of the atmospheric environment from heat exchange, to solar radiation and air pollution.

Climate change related direct health effects Climate change always impacts on human bioclimate, presently it leads to increased heat stress and heat stress frequency, higher ultraviolet radiation doses especially in summer, longer allergic pollen seasons and new allergens as well as intensified photo-smog.

Thermal stress and mortality Summer heat waves in mid-latitudes and elsewhere increase without doubt mortality; hence also highlight lack of correct adaptive measures, i. e. heat waves impact most strongly in societies with lack of social cohesion.

Global expansion of tropical diseases The observed recent global warming has increased the incidence and enlarged the distribution of some tropical diseases due to the expansion of suitable conditions for both vectors and pathogens. A northward spread has been observed for West Nile fever, Leishmaniasis and Chikungunya fever and a climate-driven spread has in parts also been recorded for malaria, dengue fever and other vector-borne infectious diseases.

Vector-borne diseases In epidemiology a vector is an organism transmitting a pathogen from one of its reservoirs (e. g. ruminants, birds) to another one (e. g. human) without falling ill. Such vectors for tropical diseases are: mosquitoes, biting flies, bugs, lice, flea's and mites. Typical vector-borne diseases are malaria, yellow fever, dengue fever, West Nile fever, Leishmaniasis, Chikungunya fever. For some of these diseases global warming is the cause of the observed expansion or intensification. The complex web of reservoir organism, pathogens, vectors and infected organisms with different dependence on climate parameters often hinders a full understanding. Hence, surprises are and will be common.

Arbo viruses Arbo viruses are transmitted by arthropods (**ar**thropod-**bo**rne) to vertebrates and hence in parts also to humans. Besides yellow fever, tick borne encephalitis and dengue fever about 150 other diseases are due to virus infections by insects and spiders (arthropods). In very complex transmission cycles climatic conditions play a central role. The occurrence of unusual arbo virus infections is often related to changes in climatic conditions. Therefore, the partly dramatic global increase of some arbo virus infections is also driven at least in part by the ongoing global anthropogenic climate change.

Arbo viruses transmitted by Aedes mosquitoes Aedes mosquitoes and Aedes-transmitted arbo viruses such as the dengue and yellow fever viruses are a growing global threat. The primary vector of these diseases, *Aedes aegypti*, has re-emerged throughout the tropics,

but also *Aedes albopictus* has emerged as one of the worst invasive species taking the role of *Aedes aegypti*. Direct human activities like global trade are mainly responsible for the spread of these vectors and global warming – indirectly anthropogenic as well – cannot be ruled out as a contributor. With further warming temperate regions like Central Europe could also become areas for *Aedes albopictus*.

Malaria and global warming Although the Anopheles mosquitoes transmitting the protozoae (e. g. *Plasmodium falciparum*) causing malaria are strongly dependent on temperature and suitable small water reservoirs for the larvae, the spread of malaria in recent years is more a consequence of deficiencies in public health systems of many countries rather than due to the observed global warming and concomitant precipitation changes.

Blue-tongue disease in Europe Since August 2006 the blue-tongue-disease of cattle and sheep (serotype 8 from South Africa) spread within months from the Netherlands to Belgium, Luxembourg, France and Germany alone at the end of 2006. The vector carrying the virus is the ceratopogonid, biting midge, *Culicoides obsoletus*. The new disease for ruminants in Western and Central Europe is primarily a consequence of globalization but the extremely warm winter 2006/2007 in Western and Central Europe supported further spreading.

Carbon dioxide fertilization and quality of food From field experiments at elevated carbon dioxide concentrations (close to a doubling of preindustrial values) it is known that agricultural yield increases for C3 plants by 10 to 30 percent; however, frequently also reduced nitrogen content in plant tissue including seeds is observed. Hence, food quality may be lowered.

Changes in the pollen season and new pollen The onset of flowering of plants in mid and high latitudes is mainly triggered by temperature. Therefore, warming in recent decades has caused an earlier start of pollen in the air, often also leading to a longer pollen season and for some pollen also to higher abundance stimulated by higher carbon dioxide concentration. Hence, susceptible individuals will suffer from pollen allergy longer and even perennial allergic symptoms may become possible.

Introduction

The basis of our life is energy from the sun, water from the skies and biomass production by plants on land and in the oceans. If we ask for the key climate parameters – given the size of the planet and its mean distance to the sun – we get a very similar answer: energy flux density of the sun, precipitation and land surface parameters, mostly determined by vegetation. Hence, climate is the key natural resource. If this resource changes rapidly, as it does now, life in all its forms is affected as well. Therefore, it is a must for decision makers to deal with climate change. Here only one facet of climate change is discussed: the health of humans. It is rather astonishing that it has not already been studied intensively since the beginning of the climate change debate, as other environmental policy decisions were nearly totally driven by health consequences for humans [12].

The consequences of climate change for the health of humans, animals and plants are complex with direct and indirect relations between causes and impacts. As any organism on land is in permanent struggle with the local weather a rather strong capacity for adaptation exists; however, the organisms are trained only with existing weather and climate variability during the life-time of an organism. For new weather extremes accompanying any climate change this adaptation capability is no longer given. Hence, a changing climate will often become a threat to health. As several health related factors may change simultaneously, "multiplying" impacts, which – if alone – would not have gone beyond existing adaptive capability, may do so. An example is heat stress during heat waves accompanied not only by high ultraviolet radiation levels but also by high near surface concentrations of the strong oxidant ozone, exacerbating the heat stress.

According to the World Health Organization (WHO) "Health" is defined as "the state of complete physical, mental and social well being and not merely the absence of disease or infirmity". This comprehensive and also ambitious understanding of the term health can also be applied to animals and plants as well as ecosystems (the terms ecosystem health and environmental health have been used very often in recent literature). Therefore, consequences of climate change on health are not confined to diseases but also include reduced well-being or reduced strength of an organism as well as weakened functions in an ecosystem or a socio-economic system.

This contribution to the Encyclopedia of Complexity does not focus exclusively on climate change and human health (see Sect. "Climate Change Impact on Human Health") because climate change impacts on animals and plants (see Sect. "Climate Change Impacts on Plants with Consequences for Human Health") also have consequences for human well-being and human health. An example of the complexity is the changed composition of nutrients in grains as a consequence of higher carbon dioxide levels in the atmosphere, in addition mod-

ulated by changed climate parameters that may in turn enhance infestations of plant diseases with consequences for the composition of our food. The section on Climate Change Impacts on Plants with Consequences for Human Health tries to collect known knowledge, reports on potential threats and chances for political reactions as well as pointing out major open research questions.

Climate Change Impact on Human Health

Whenever climate changes all ecosystems have to react, i. e. they try to adapt to changed climate that must also include new weather extremes. Typical reactions are shift of biomes, altered ecosystem composition, new geographical patterns of animal and plant diseases, new relations between predator and prey. As major so-called natural climate changes have occurred also in the recent few million years the first major question in the present anthropogenic climate change era is: Can earlier so-called abrupt climate changes be used as an example for projections into the future? The answer is no, because the mean global temperature change rate to occur in the 21st century exceeds by far even the most rapid natural ones. The largest and most rapid *global* mean temperature change rates in the recent few million years have been the collapses of the large northern hemisphere ice sheets that have led to a mean global warming of up to 5°C and a sea level rise of about 120 m in roughly 10,000 years. Projections for the 21st century without stringent climate policy [4] range between 1.5°C and about 4°C, i. e. an acceleration of at least a factor 30. Hence, the past cannot be used as an analogue. In other words: The key goal of the United Nations Framework Convention on Climate Change (UNFCCC) that speaks of avoidance of a dangerous interference with the climate system, cannot be met without globally coordinated climate policy. It requests, in addition, a climate policy helping to stabilize greenhouse gas concentrations within a time frame that firstly natural ecosystems are able to adapt to climate change, secondly food production is not generally threatened and a sustainable economic development remains possible.

In this situation with ill-adapted forests, rapidly changed patterns of infectious diseases for plants, animals and humans looking back through climatic history does also not help directly. Besides intensified research on changed disease patterns in the very recent past a stringent globally coordinated climate policy under UNFCCC is the best insurance against massively altered disease patterns.

This paper will concentrate on climate change impacts on human health but will not exclude totally impacts on food production. Major points will be "thermal stress" before "vector-borne diseases" and also prolonged allergen seasons and new allergens are discussed.

Changed Thermal Stress

Heat or cold stress forces our body to adapt to keep the core body temperature within a narrow temperature interval of about five degrees centigrade (35 to a maximum of slightly above 40°C). While extreme cold stress events have diminished and will further diminish as a consequence of ongoing global warming, extreme heat stress will increase dramatically (see Fig. 1), when the frequency distributions of temperature at a certain location are shifted by only a few degrees centigrade and may be broadened. Up to now only very few places on the Earth's surface exist where survival of a human being is nearly impossible. This would certainly happen if the wet bulb temperature (roughly equivalent to a ventilated sweating naked body) surmounts about 35°C. Under present climate conditions such areas do not exist, but coastal areas of the Red Sea come closest to it during on-shore winds after a sunny day that was heating surface waters to about 35°C. Hence, heat strokes can in principle be avoided by adequate behavior, if buildings are well insulated and properly ventilated and if an individual behaves. Therefore, the huge death toll caused by major heat waves, namely about 55,000 people that died during the heat wave in summer 2003 in Europe [5] is rather an indication for "social freeze" and ill-adapted buildings in our industrialized or developing societies than for really intolerable thermal conditions. Many have died because of lack of care. New weather extremes always demask weaknesses in our security-related infrastructure. A famous example is the large difference in the number of people dying during a summer heat wave in 1995 in two US cities (Chicago and Philadelphia) just because of intensified public care in Philadelphia where the weak and poor citizens where brought by the city administration to the cooled malls during daytime.

As Fig. 2 clearly demonstrates the mortality anomaly caused by heat waves is a fact observed over several decades (here in a developed country). Lowering the anomaly means not only investment in better warning systems but also enhanced social care in general.

Both heat and cold stress increase mortality as underlined by Fig. 3 for European countries. In the future heat stress category 3 will occur much more often and the lowering of the increased mortality will also be a sign of an improved public health system.

A further point to be made with respect to enhanced thermal stress as a consequence of climate change is the rapidly mounting heat stress in the inner tropics, where

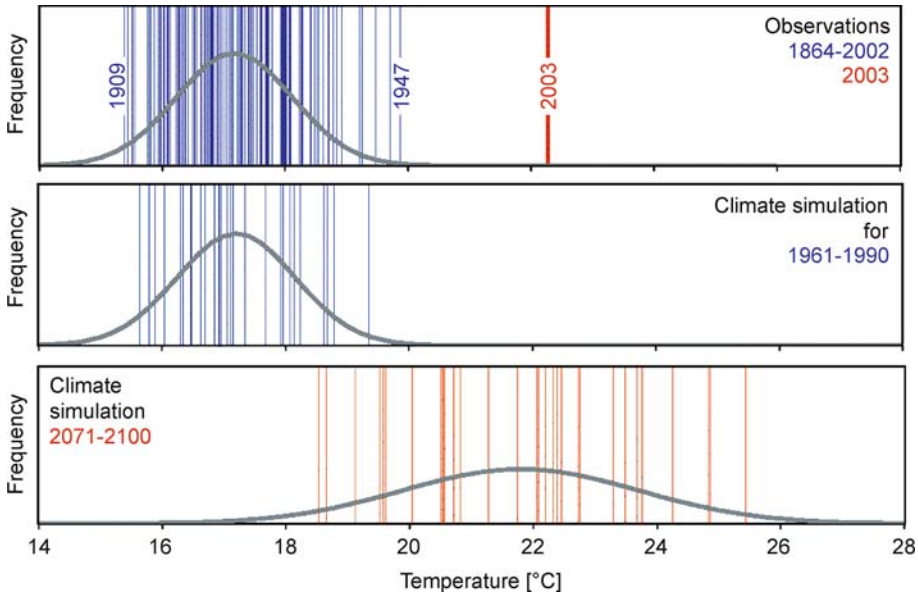

Climate Change and Human Health, Figure 1
Observed (*upper panel*) and modeled summer mean temperatures for the Swiss Plateau both for present climate (1961 to 1990) (*central panel*) and the last three decades of the 21st century for scenario A2. Please note that the exceptional summer 2003 would occur every second year, if globally coordinated climate policy would not exist. From [13]

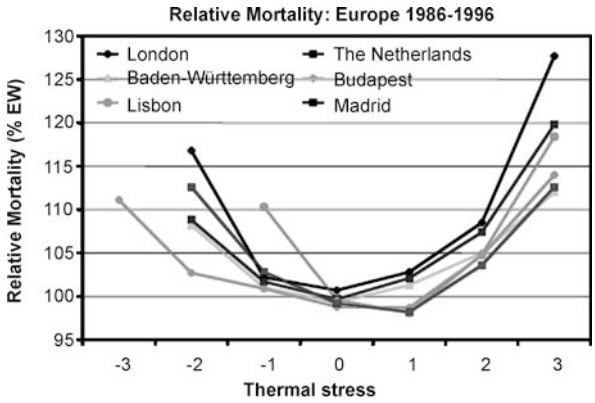

Climate Change and Human Health, Figure 2
Mean relative mortality in percent for different thermal stress categories observed in Europe during 1986 to 1996. Please note that the mortality increases by more than a factor 2 when heat stress category 2 is replaced by category 3. From [5]

dew points of about 25°C will more often occur, if global warming also continues there. The ability to work with high efficiency is shrinking there rapidly with rising temperatures and dew points. As is well known economic development of developing countries needs a cooled or well-ventilated work place.

Impact on Photochemical Smog

Photochemical smog is formed if solar radiation stimulates chemical reactions in a polluted atmosphere. Emissions of non-methane hydrocarbons and nitrogen oxides ($NO + NO_2$) lead to the formation of ozone and other oxidizing toxic trace gases as well as aerosol particles. In mid-latitudes photochemical smog is typically strongest in late spring and summer and it has been the reason for some environmental policy making. Heat waves with intense solar radiation lead to major photochemical smog episodes. The higher frequency and longer duration of heat waves during recent and foreseen global warming will intensify health problems already existing during heat waves. As Tables 1 and 2 demonstrate the hot summer in Germany in 2003 has increased the number of days for both ozone and aerosol load (PM_{10}) where limit values have been surmounted in Germany dramatically [14,15], e. g. by a factor of 2 or more both for PM_{10} in rural areas and ozone in all areas.

Fighting against photochemical smog will become even more demanding in a warming world.

Changed Ultraviolet Radiation

Ozone protects us to a large extent from the dangerous part of ultraviolet solar radiation through absorp-

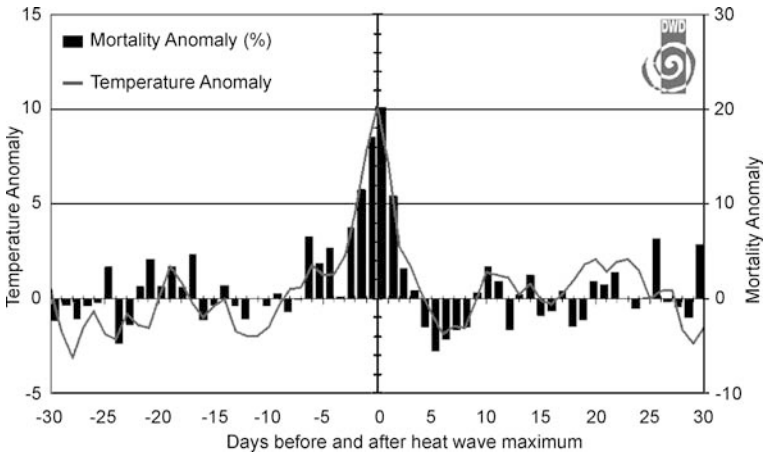

Climate Change and Human Health, Figure 3
Mean mortality anomaly during a heat wave (30 days before and after its maximum) resulting from nine observed heat waves in the state of Baden Württemberg (Germany) for the period 1968 to 1997. Source: [5]

Climate Change and Human Health, Table 1
Days with 8-hour mean ozone concentration above 120 µg m^{-3} since 1990 for all stations in the German network

1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
22	22	28	23	32	29	20	22	19	21	19	21	19	51	19

Climate Change and Human Health, Table 2
Days with PM_{10} values above 50 µg m^{-3} for the years 2001 to 2004

2001	2002	2003	2004
Stations with highest values			
117	103	132	73
Stations in cities (average)			
65	75	83	55
Remote stations in cities (average)			
22	30	38	16
Rural stations (average)			
7	12	17	5

tion in the ultraviolet (UV-B) range from 0.28 to 0.32 µm wavelength. The observed latitude-dependent reduction of stratospheric ozone (from 8 to 50 km height in high latitudes), which constitutes about 90% of the total ozone column content and which is caused by chlorine compounds stemming from decay products of chlorofluorocarbons (CFCs) and other halocarbons, has increased UV-B radiation especially in spring of both hemispheres at higher latitudes. This stratospheric ozone decrease is strongest in Antarctic spring in the so-called ozone hole. This ozone depletion has certainly increased and still will increase the skin cancer and cataract incidence that has dramatically grown in recent decades. However, the ozone depletion contribution is largely buried in the variability and change of exposure of our body to UV radiation, which is due to changed behavior, especially in developed countries. The key question for the forthcoming decades is: How fast will ozone column content recovery be after the banning of CFCs and other chlorine and bromine containing compounds? Present knowledge says: Full recovery after several decades with a chance of a super-recovery caused by a further increase of the greenhouse effect of the atmosphere, which lowers stratospheric temperatures.

Health Effects Caused by Other Changed Climate Parameters

Climate change shifts and reshapes frequency distributions of meteorological and hydrological parameters, thereby multiplying the occurrence of known extremes and leading to new ones (see Fig. 1, where this is demonstrated for temperature). Therefore, the health of many more millions is affected by intensified flooding, higher storm surges, and many other weather-related disasters.

While the highest death toll of weather-related disasters was – for thousands of years – due to droughts, with up to 10 million people dying per decade around 1930, the highest death toll is now caused by flooding and wind damage due to tropical cyclones and hurricanes (Red Cross, …). The main reason for this change is – besides

more dwellings in flood-prone areas, often already lying slightly below mean sea level – the international aid bringing food and seeds into the drought-affected areas; as long as civil war does not prevent this help.

Hence, climate change also calls – irrespective of globally coordinated climate change mitigation policies – for a coordinated climate change adaptation policy in the coming decades, because we have to adapt to the already unavoidable climate change. Mitigation measures meant to avoid the un-tolerable climate change will only become effective in decades due to the inertia of the climate system caused by the slow reaction of oceans and ice sheets. In other words: If flooding is prevented by strengthened dikes, anticipating shifted frequency distributions of precipitation, cholera epidemics will not occur. Diking has to become an international activity, as the emitters in industrialized countries are causing more flooding and sea level rise on a global scale, co-financed by the already existing but strongly to be increased adaptation fund under the UNFCCC, its Kyoto Protocol and the follow-up protocol envisaged to be signed in 2009.

As an aside I will here report on reactions of our body to high carbon dioxide levels in the atmosphere. Very often if many people are gathered in the same closed room and breathe the same air some will ask for fresh air (i. e. they require more oxygen). The need for fresh air is because of too high carbon dioxide levels. Regulations in some German states concerning ventilation in classrooms provide ventilation rules to avoid carbon dioxide concentrations above 1,500 ppm, a level after which concentration diminishes and some students may even develop signs of a beginning headache. The oxygen content of air has fallen by a bit more than a tenth of a percent only to about 20.84 percent at which point ventilation by opening windows becomes a must. Hence, ventilation of our living rooms means pushing out carbon dioxide.

Transmission of Infectious Diseases from Birds to Humans

In recent years migratory birds have been named as a cause for the transmission of bird flu (avian influenza) to humans because they can in principle transmit the influenza virus to chickens, geese, turkeys and ducks in our farms from where the virus infects humans that come into close contact with the fowl or their products. However, very often the cause for the long-range transmission is global trade and tourism on the one hand plus industrialized animal husbandry in developed and emerging countries on the other. The latter has been found as the principal cause for the comparably rapid mutation of slightly pathogenic bird flu viruses to highly pathogenic ones (Bairlein and Metzger, 2008). These in turn can reach wild bird species which then can rapidly transmit them to new areas, if the viruses are only slightly pathogenic for them. But also pathogens, vectors or reservoir species can inadvertently be introduced by globalization to regions where no longer-term co-evolution could have taken place. Further global warming will intensify the shift of pathogens with the shift of (migratory) birds, but will only add to the further rapid distribution of pathogens caused by globalization (Smith et al., 2007).

Allergies Caused by Pollen

In most countries allergies have recently increased dramatically. In Germany, for example, 20 to 30% of the population suffers from allergies. Most abundant is the allergic rhino conjunctivitis (hay fever), often turning into asthma bronchiale, caused by allergic pollen in air. Climate change has led to longer pollen seasons, in parts to more pollen, changed pollen spectrum and also new pollen [9].

From studies in Europe, North America and Japan an earlier flowering of 1 to 3 days per decade has been reported during the recent decades [11]. For some species, especially the late flowering ones a prolonged pollen season has been found [2,3], in parts caused also by long-range transport of the pollen. Consequently, for many people in mid-latitudes suffering from several pollen the pollen season became longer, sometimes already starting in December and ending only in October after the flowering of the neophyte ragweed (*Ambrosia artemisiifolia*) for Europe.

From differences between cities and rural areas as well as from laboratory studies it became known that pollen abundance increases with carbon dioxide concentration for some species.

In combination with air pollution allergic reactions have been shown to intensify [3].

Invasion of Allergenic Neophytes With ongoing global warming two processes combine for the spread of neophytes: Firstly, global trade and tourism transmitting plants and their seeds within days and weeks around the globe to all inhabited places and into ocean basins and secondly, increased temperatures allowing more and more often establishment of exotic species in new areas. A famous example where both mentioned processes work in combination is the invasion of the strongly allergenic ragweed (*Ambrosia artemisiifolia*) from North America to Europe in the 19th century with a large spread after the Second World War. Ragweed flowers from late August to Septem-

ber and its pollen also undergo long-range transport [1]. In recent decades it spread strongly in Central Europe facilitated by agricultural practice and by inadvertent transport with bird food as well as higher temperatures.

Indirect Health Effects Caused by Climate Change

Many infectious diseases are transmitted by vectors, i. e. by animals (very often insects) transferring a pathogen from another animal and/or human without suffering from the pathogen themselves. The complicated web of hosts, pathogens, reservoirs and vectors is dependent on many factors, among them climate variables, foremost temperature and precipitation. Therefore, distribution patterns and incidence of infectious diseases will also be modified by climate change; however, disentangling the cause and effect relationships is often extremely difficult. Since tropical infectious diseases are the most common of these diseases and especially temperature sensitive they will be a focus in this section. The World Health Organization (WHO) points to a highly probable increase of morbidity and mortality due to higher prevalence and a pole-ward shift of tropical infectious diseases due to further global warming.

Increase of Vector-Borne Tropical Infectious Diseases

As always in scientific investigations it is especially difficult to derive long-term trends of certain variables influenced by many parameters like changes in land use, socio-economic conditions and climate or weather patterns. This is also true for tropical infectious diseases. Hence, only few trend analyses exist, e. g. for malaria reaching higher elevations in Africa.

The main vectors for infectious diseases are mosquitoes, biting flies, bugs, lice, fleas, and mites. Of about more than a million insect species roughly 17,000 have adapted to a blood-sucking mode of life. A small minority of these are vectors of pathogens. The pathogens are viruses, bacteria, protozoa or filarioses. Pathogens are either multiplied within the vectors without change of form (Arboviruses, Rickettsiae, Bacteria), multiplied with change of form (Protozoa) or changed without multiplication (Filaria). From uptake of a pathogen during a blood meal until the infective stage temperature is *the* factor determining duration, hence, higher temperatures can lead to enhanced spread of vector-borne diseases as for example observed for dengue fever.

According to number of people infected by blood-sucking insects *malaria* comes first with 300 to 500 million cases per year of which 1 to 2 million die, especially children. About 70 mosquito species of the genera Anopheles transmit four different protozoa (*Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium ovale*, *Plasmodium malariae*). The mosquito *Anopheles gambiae* in Africa is the most important vector. Because malaria has been eradicated in many developed countries in Europe an enhanced potential for malaria must not lead to a re-introduction there. However, in countries with a weak public health system higher prevalence and new malaria infected areas are highly probable.

About 120 million people in Asia, Africa, South and Central America are infected by *lymphatic filariasis* (elephantiasis) transmitted by different genera of mosquitoes. The larvae of the worm develop (from 0.4 to about 1.5 mm) in the vector and can be transmitted during the next blood meal to humans. The development period is temperature-dependent. Hence, a potential for a further spread exists, but the Global Program to Eliminate Lymphatic Filariasis (WHO, 2006) may reduce incidence strongly.

Dengue fever affects about 50 million persons per year and it is a clearly growing threat to human health in about 100 tropical and subtropical countries. The virus belonging to the flaviviruses is transmitted mainly by the *Aedes aegypti* mosquito. Vector and virus show strong temperature dependence in their development. But the disappearance of *Aedes aegypti* from Europe around 1950 points to the probably strong application of the insecticide DDT. As often, measures taken against infectious diseases can easily off-set climate influences.

Onchocerciasis (in severe cases leading to river blindness) affects about 37 million persons, mainly in Africa, the microfilaria of the worm "*Onchocerca volvulus*" are taken up from the human skin by the vector, a simuliidae species, during a first blood meal, they develop through two skinnings within the vector and are transmitted back to a human being during a further blood meal after about 7 to 10 days. Whether the vector, developing in running water and sucking blood during daylight, is already reacting to climate change is not known although a temperature and precipitation dependence clearly exists.

Many other tropical vector-borne infectious diseases exist, like loiasis affecting about 13 million people in tropical Africa, "Schlafkrankheit" with about 60,000 new cases per year or West Nile fever that has reached the USA. In all cases it is a complex mix of influencing factors with positive and negative feedbacks, which inhibits a clear separation of a climate contribution to changed patterns and severity. Hence, there will be many surprises often based both on transport of vectors by growing trade and tourism and better survival conditions in higher latitudes due to higher temperatures and/or changed moisture conditions.

The best means to cope with changed disease distribution patterns is a strong public health system and links between the systems in different countries.

Climate Change Impacts on Plants with Consequences for Human Health

Our food is produced by plants on land and in the ocean, even if we eat meat or fish because animals also feed ultimately on plants. Hence, a climate change impact on plants may have strong indirect effects on human health. However, even if the climate change impact were small on certain plants, their way of reacting to enhanced CO_2 concentration could still have consequences for food production. In reality climate change and elevated CO_2 concentration act together and both have largely different impacts on plant types and species. The temperature dependence of photosynthesis rate as presented in Fig. 4 for very different plants, shows the rather steep decline of this rate for maize, a C4-plant, at high temperatures above about 40°C. Adding to this finding that there is nearly no positive feedback to higher CO_2 levels for C4-plants, yield in tropical areas would be reduced if plants have to assimilate at leave temperatures of about 40°C. Therefore, Working Group II of IPCC concluded [4]: "Crop productivity is projected to increase slightly at mid- to high latitudes for local mean temperature increases of up to 1–3°C depending on the crop, and then decrease beyond that in some regions. At lower latitudes, especially seasonally dry and tropical latitudes, crop productivity is projected to decrease for even small local temperature increases (1–2°C), which would increase the risk of hunger."

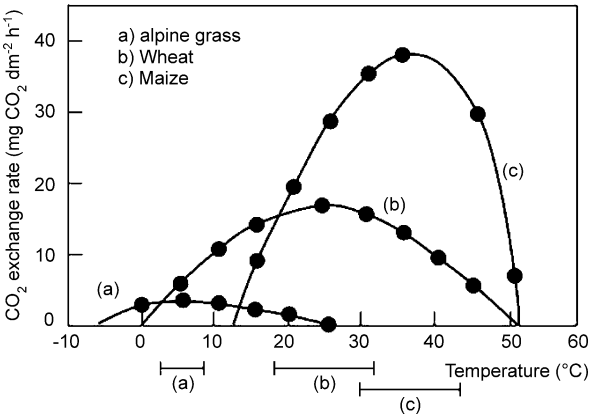

Climate Change and Human Health, Figure 4
Temperature dependence of photosynthesis rate for an alpine grass, wheat and maize. Please note the strongly differing optimal temperature ranges. From [10]

If C3-plants (e. g. wheat, rice, potato, sugar beet) live at higher CO_2-concentrations their photosynthesis rate increases rather linearly with CO_2-concentration, if water stress and nutrient scarcity are not limiting their photosynthesis rate. Consequently, the delay of decades between CO_2-concentration rise and full expression of the warming, to which we are already committed, is a "window of opportunity" for high crop yields of C3-plants. This may in the long-term also have global consequences, because the CO_2-concentration could rise additionally if the (high latitude) forests (C3-plants), acting as a sink for anthropogenic CO_2 presently, would lose this capacity under higher climate change stress. Whether and when this will occur is not yet known.

Changes in Food Composition for Main Crops at Elevated CO_2-Concentration

An important consequence of elevated CO_2-concentration would be changed composition of plant tissue, and especially of seeds, as it could have immediate health consequences for animals and humans eating them. The sparse body of published studies is nearly unanimously pointing to a loss of nitrogen content in leaves and stems as well as in seeds (see Fig. 5 and [6] for an overview). The results of so-called **F**ree **A**ir **C**arbon Dioxide **E**nrichment (FACE) studies at roughly doubled CO_2-concentrations ($\sim$ 550 ppm) are all reporting higher nitrogen content for plant tissue and a bit less for seeds. This negative consequence for our food is not yet fully acknowledged in the public, because of the very different impacts for different parts of society. For a grain-producing farmer the CO_2 fertilization effect leads to higher yields of C3-plants, a dampening of photo-smog yield reductions, higher water use efficiency for both C3- and C4-plants and thus less drought impact while for the baker the wheat quality for baking bread declines, and the cattle as well as the consumer get less healthy food.

Shift of Biomes and Migration

As already observed, precipitation is redistributed due to rather different warming patterns. And projections of precipitation changes in the 21st century, as published in the Fourth Assessment Report of [4], can be summarized in the following sentence: Rather humid areas will on average get more precipitation, especially in high latitudes, while semi-arid areas will get on average less (with strongly lowered soil moisture at higher temperatures). This is bad news for many millions of people in semi-arid zones of the tropics and subtropics. The main impact on plants is water stress, hence lower crop yields. But also less food for cat-

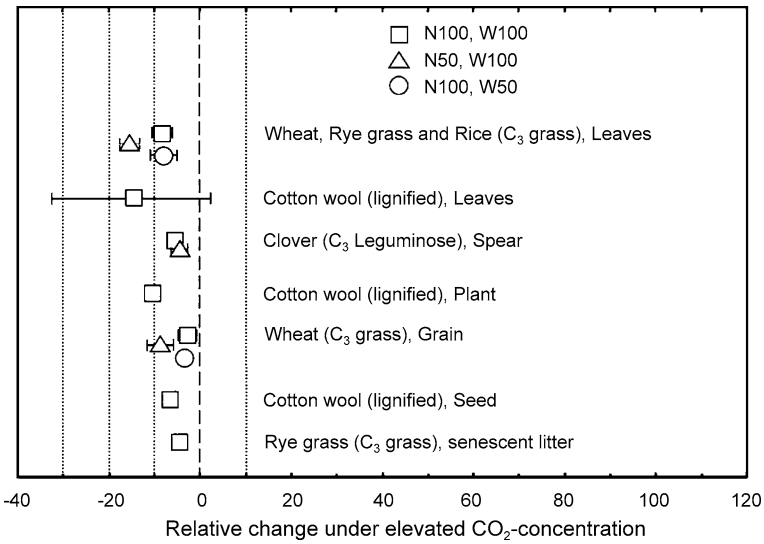

Climate Change and Human Health, Figure 5
Changes in nitrogen content in plant tissue under elevated CO_2-concentration (~ 550 ppm) depending on nitrogen fertilizer and water availability (50 or 100%). From [16]

tle in pasture lands, continued or enhanced malnutrition for the poor, aggravation of existing deficiencies in public health care systems in these areas will be the negative consequences of climate change impact on the biomes of these areas. In other words: Undermined public health systems, loss of livelihoods and finally migration will be the dire consequences of anthropogenic climate change. The socio-economic and political reactions to this threat cannot be foreseen in detail. However, the international economic cooperation between developing and industrialized countries has to take as a priority adaptation to unavoidable climate change as the best means to lower impacts on public health.

Concluding Remarks

Although threat to human health was often the cause for environmental policy making, e. g. in the case of desulfurization of power plant exhaust, the manifold threats to human health as a consequence of global anthropogenic climate change have rarely been named as a key reason for climate change policy measures. What else is the partly still growing mal-nutrition of subsistence farming communities in the desertification-prone semi-arid tropics and subtropics than a threat to human health? On the other hand most new threats to our health caused by the spread of vector-borne infectious diseases due to higher temperatures or those caused by new weather extremes can be strongly reduced by proper health system up-grading and pre-cautionary measures that strengthen security-related infrastructure. However, this will probably not be the case in developing countries already suffering from (very) weak public health systems; unless the preliminary decisions of the 13th Conference of the Parties to the UNFCCC lead to a new international and binding protocol in 2009 as a follow-on to the Kyoto Protocol that then stipulates that a fixed portion of the revenues of international greenhouse gas emission trading should be used for adaptation measures in developing countries. A large share has to be invested in public health systems in poor developing countries in order to help the poorer parts of societies typically suffering most from epidemics and weather-related catastrophes.

If earlier tropical vector-borne infectious diseases, like the West Nile fever, reach developed countries, research to get proper vaccines will be stimulated within large pharmaceutical companies. If the threat remains confined to the poor South, this research effort will often not exist, because developing countries' normal population cannot afford the expensive new drugs or vaccines that remain property right protected for years. It is high time that political summits deal with this problem and WHO helps to circumvent this barrier as pointed out recently by the Nobel Prize laureate for economics (Stiglitz, 2006).

Bibliography

1. Bohren C, Memillod G, Delabays N (2006) Common ragweed in Switzerland: development of a nation-wide concerted action. J Plant Dis Prot, Special Issue XX, Ulmer, Stuttgart, pp 497–503

2. Confalioneri U, Menne B, Akhtar R, Ebi KL, Hanengue M, Kovats RS, Revich B, Woodward A (2007) Human health. In: Parry MO, Lanziani OF, Palutikof JP, van der Linden PJ, Hanson CE (eds) Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of IPCC. Cambridge University Press, Cambridge, pp 391–431
3. D'Amato G, Cecci L, Bonini S, Nunes C, Annesi-Maesano I, Behrendt H, Liccardi G, Popov T, van Canwenberge P (2007) Allergenic pollen and pollen allergy in Europe. Allergy 62:976–990
4. Intergovernmental Panel on Climate Change (IPCC) (2007) Science of Climate Change, Contribution of Working Group I to the Fourth Assessment Report. Cambridge University Press, Cambridge
5. Jendritzky G, Koppe C (2008) Die Auswirkungen von thermischen Belastungen auf die Mortalität. In: Lozán JL, Grassl H, Jendritzky G, Karbe L, Reise K (eds) Warnsignale Klima: Gesundheitsrisiken. Wissenschaftliche Auswertungen, Hamburg, ISBN 978-39809668-4-9
6. Kimball BA (2004): Global environmental change: implications for agricultural productivity. Crop Env Bioinform 1:251–263
7. Laschewski G, Jendritzky G (2002) Effects of the thermal environment on human health: an investigation on 30 years of daily mortality data from SW Germany. Clim Res 21:91–103
8. McMichael A, Campbell-Lendrum DH, Corvalán CF, Ebi KL, Githeko AK, Schraga ID, Woodward A (2003) Climate Change and Human Health: Risks and Responses. WHO, Geneva
9. Menzel A et al. (2006) European phonological response to climate change matches the warming pattern. Glob Chang Biol 12:1969–1976
10. Rosenzweig C, Hillel D (1998) Carbon Dioxide, Climate Change and Crop Yields. In: Rosenzweig D, Hillel D (eds) Climate Change and the Global Harvest. Potential Impacts of the Greenhouse Effect on Agriculture. Oxford University Press, Oxford, pp 70–100
11. Rosenzweig C, Cassassa G, Karoly DJ, Imeson A, Liu C, Menzel A, Rawlins S, Toot TL, Seguin B, Tryjanowski P (2007) Assessment of observed changes and responses innatural and managed systems. In: Parry MO, Lanziani OF, Palutikof JP, van der Linden PJ, Hanson CE (eds) Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of IPCC. Cambridge University Press, Cambridge, pp 79–131
12. Sauerborn R (2007) Climate change: an agenda for research and teaching in public health. Scand J Public Health 1–3
13. Schär C, Vidale PL, Lüthi D, Frei C, Häberli C, Liniger MA, Appenzeller C (2004) The role of increasing temperature variability for European summer heat waves. Nature 427:332–336
14. UBA (2007) www.env-it.de/luftdaten/download/public/docs/pollutant/03/Jahr/Ozberi06.pdf
15. UBA (2007) www.env-it.de/luftdaten/download/public/docs/pollutant/PM10_gesamt_2001-2006.pdf
16. Weigel H-J, Manderscheid R, Fangmeier A, Högy P (2008) Mehr Kohlendioxid in der Atmosphäre: Fluch oder Segen für die Landwirtschaft. In: Lozán JL, Grassl H, Jendritzky G, Karbe L und Reise K (eds) Warnsignale Klima: Gesundheitsrisiken. Wissenschaftliche Auswertungen, Hamburg, ISBN 978-39809668-4-9

Climate Modeling, Global Warming and Weather Prediction, Introduction to

Hartmut Grassl
Max Planck Institute for Meteorology,
Hamburg, Germany

All systems operating away from thermodynamic equilibrium develop structures. The planet Earth will always be far away from thermodynamic equilibrium because of the strongly differing solar radiation input as a function of latitude and season. Hence, the differential heating of the surface and also the atmosphere must lead to temperature gradients, in turn causing pressure gradients that create currents in the ocean and the wind in the atmosphere. On a rotating sphere (in reality the geoid, which is close to a rotational ellipsoid) these flows form low and high pressure systems both in the ocean and the atmosphere which are able to reduce latitudinal gradients but never come close to thermodynamic equilibrium because of continuing differential heating. The average temperatures and flow fields, as well as their strong spatial and temporal variability, are a function of land/sea distribution and atmospheric composition, especially depending on water and ice in clouds. The strongly climatically relevant gases in the atmosphere are, to a large extent, a consequence of life on Earth.

Ranking all radiatively active gases in the atmosphere according to their influence on weather and climate shows the exceptional composition of the atmosphere: Water vapor (H_2O) in all three phases but largely as a gas, carbon dioxide (CO_2), ozone (O_3), nitrous oxide (N_2O) and methane (CH_4) constitute only three thousandths of the atmospheric mass, yet they largely determine how much solar radiation reaches the surface, e. g., through clouds, and how much thermal or terrestrial radiation leaves from there to space, again a strong function of clouds and the above-mentioned gases. The average surface temperature is thus strongly depending on the concentration of the gases mentioned, which are all greenhouse gases. They do not strongly absorb solar radiation but do absorb terrestrial radiation, thereby forcing the surface and the lower atmosphere to warm in order to reach nearly equilibrium between absorbed and emitted energy. Any growth or reduction of greenhouse gas concentrations increases or decreases average surface temperature and thus changes climate.

The climate system, i. e., its interacting components atmosphere, ocean, land, vegetation, soils and crust, shows

both a remarkable stability and high sensitivity. Over many million years the greenhouse effect of about 30 K has varied only by about +/ − 5 K with respect to present interglacial temperatures, thus has been stable in terms of temperature varying only from 283 to 293 K. +5 K however meant melting of all inland ice sheets and −5 K a new strong glaciation with major ice sheets reaching 40 to 50° N (see ► Cryosphere Models).

The observed strong increases of all long-lived naturally occurring greenhouse gases (+35% for CO_2, +120% for methane, and +10% for N_2O since 1750) have stimulated a mean global warming which now has emerged from strong climate variability. In 2007, the Fourth Assessment Report of the Intergovernmental Panel on Climate Change concluded: "The understanding of anthropogenic warming and cooling influences has improved … leading to very high confidence that the global average net effect of human activities since 1750 has been one of warming". On the other hand, the high sensitivity of the climate system is demonstrated in so called abrupt climate change events (see ► Abrupt Climate Change Modeling).

The strongest global one of these – besides the impact of celestial bodies – is deglaciation after an intense glaciation in about 5000 to 10 000 years with a concomitant temperature increase of 4 to 5 K, caused by the slow latitudinal redistribution of solar radiation due to Earth orbit parameter changes as a consequence of slowly changing gravitational forcing by the neighboring planets, mainly Venus, Jupiter and Saturn. Because climate is a key natural resource for plants, animals and humans, any rapid climate change threatens life on Earth. Therefore agriculture (see ► Climate Change and Agriculture), forestry and all economic activities (see ► Climate Change, Economic Costs of) will be impacted by anthropogenic climate change in the 21st century, leading to strong consequences in societal behavior vis-a-vis this challenge, e. g., the one caused by growing inequity between those societies causing climate change, the industrialized countries, and those suffering first or more strongly such as subsistence farmers in semi-arid tropical areas. Climate Models developed so far include many physical processes (see ► Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface as an example), parts of atmospheric chemistry, and vegetated land surface atmosphere interactions, but still lack reaction of ocean biomass production to enhanced CO_2 levels. Finally, it is apparent that global warming will have an effect on human health and that this will include effects on food crops and animals (see ► Climate Change and Human Health).

The low horizontal resolution of global climate models has stimulated nested regional climate models (see ► Regional Climate Models: Linking Global Climate Change to Local Impacts) delivering enhanced output in areas with strong topography or sea/land contrasts. The slowly emerging Earth System Models are no longer driven by changed atmospheric composition alone but by emissions, i. e., they can calculate resulting greenhouse gas concentrations. However they are still not advanced enough to answer the question: When will growing climate change stress turn the present uptake of anthropogenic CO_2 into forests through the CO_2-fertilization into an additional CO_2 source for the atmosphere due to a generally weakened vegetation? It is common knowledge that weather, for example the passage of a coldfront at a certain location, can be forecast only for up to two weeks because of the intrinsically chaotic behavior of atmospheric flow. The accuracy of present day weather forecast models, which are very similar to the atmospheric component of climate models, has advanced strongly, driven by higher spatial resolutions and better parametrizations of sub-grid scale processes in the models, assimilation of more (and especially satellite data) into the models and ensemble forecasting. This has recently led to the same forecast accuracy up to about 10 days in the southern hemisphere, where an in situ observing system for the starting fields of the model is largely lacking.

The increased attention the topic climate change has finally attracted in the political arena will certainly accelerate progress in this field, despite the complex nature of the functioning of the Earth system.

There are two additional articles on climate change which were recruited for other sections. These articles are: ► Dynamic Games with an Application to Climate Change Models and ► System Dynamics Models of Environment, Energy and Climate Change.

Cognitive Robotics

Kazuhiko Kawamura[1], Will Browne[2]
[1] Center for Intelligent Systems, Vanderbilt University, Nashville, USA
[2] Cybernetics, The University of Reading, Reading, UK

Article Outline

Glossary
Definition of the Subject
Introduction
What Is Cognition?
Cognitive Architectures
Cognitive Robotics
Case Study
Future Directions
Bibliography

Glossary

Amygdala Amygdala consists of almond-shaped groups of neurons located within the limbic lobe in the brain. The amygdale performs primary roles in the formation and storage of memories associated with emotions and is said to have a substantial role in mental states.

Basal ganglia Basal Ganglia are a collection of subcortical neuronal group and have a significant role in the control of movement.

Central executive agent (CEA) CEA is a cognitive or compound agent responsible for high-level executive control, such as reasoning, task switching and realization of internal rehearsal, e. g. within the ISAC cognitive architecture.

Chinese room argument John Searle developed a thought experiment called the "Chinese Room" argument against what he calls "strong AI". Searle describes a scenario in which a person who knows no Chinese is locked in a room full of boxes of Chinese symbols together with a book of instructions for manipulating symbols. This person receives questions in Chinese from under the door. If the person in the room is able to pass out Chinese symbols using the instruction book to produce correct answers to the questions, he passed the Turing Test for intelligence in Chinese, but he does not understand a word of Chinese.

Connectionist models Connectionist models of cognition are structured on the concept of neural networks. Connectionist networks provide an account for the complex behavior in a way parallel distributed processing (PDP) does. There is no way to distinguish between simple and complex representations in connectionist models. In this sense, they are considered to be subsymbolic.

Cortex Cortex (or cerebral cortex) is a surface structure in the brain responsible for many brain functions including attention, sensory processing, motor functions, awareness, language processing and arguably consciousness. The human cortex is 2–4 mm thick and consists of large sheets of mostly layered neurons.

First-order cybernetics First-order cybernetics considers control and communication in the animal and machine, where the agent receives feedback, including utility of its actions, from the environment.

Cartesian theater A centered locus in the brain called *Cartesian materialism*, because it is the view one arrives at when one discards Descartes' dualism, but fails to discard the associated imagery of a central (but material) theater where it all comes together.

Global workspace Multiple parallel *specialist* processes compete and co-operate for access to a *global workspace*. If granted access to the global workspace, the information a process has to offer is *broadcast* back to the entire set of specialists.

Humanoid robots There is no universally accepted definition for a humanoid robot today. However, it is widely accepted that a humanoid robot must have a body somewhat resembled to a human body, exhibit human-like behavior, and be able to interact with humans using human-level intelligence. As of today, no existing humanoid robots satisfy all these requirements.

ISAC ISAC stands for Intelligent Soft Arm Control. The name arises from the fact that the arm is highly compliant and safe for working with and around people. In its multiagent architecture called the Intelligent Machine Architecture (IMA), human and many modules within the humanoid are represented as distinct agents within a common computational framework.

Minimum robust representationalism (MRR) MRR is a notion, rather than a formal definition, put forward by Clark and Grush that addresses the problem of internal representation when addressing cognitive phenomena. The emphasis on emulators differs from the classical ideas of cognitivism and representationalism. Transparent (i. e. analytically traceable) emulator circuitry is the minimal needed to usefully consider representations of external states.

Multiagent systems (MAS) A multiagent system (MAS) is a software system composed of multiple agents and collectively capable of reaching goals that are difficult

to achieve by an individual agent. An agent within MAS can be autonomous in the sense that it has own decision-making capability or non-autonomous like a simple input-output device. MAS agents can include human agents like the case study in this section.

Neural networks Neural networks, or artificial neural networks, are a class of networks of simple processing units which can exhibit complex behavior. They were inspired by the way biological nerve systems, such as the brain, process information. Simple neural networks consist of three layers, input, hidden and output.

Production systems Production systems are symbolic artificial intelligence systems, i. e. they manipulate symbols, instead of numbers. Production systems are composed three parts: a global database, production rules and a control structure. Production rules (or productions) are called if-then rules.

Second-order cybernetics Second-order cybernetics recognizes that the agent has an important effect back on the environment and the two systems affect each other.

Definition of the Subject

Cognitive Robotics is an emerging field of robotics, which will continue to evolve for the years to come. The field of cognitive robotics generally comprises the design and use of robots with human-like intelligence in perception, motor control and high-level cognition. To realize cognitive robots many overlapping disciplines are needed, e. g. robotics, artificial intelligence, cognitive science, neuroscience, biology, philosophy, psychology, and cybernetics. Thus attempting to tightly define the subject is not constructive as often its nature is amorphous, growing and a strict definition could exclude future relevant work.

Work by Clark and Grush [1] towards a cognitive robot definition is well respected. We quote some important considerations below:

- We hold that fluent, coupled real-world action-taking is a necessary component of cognition.
- Cognition, we want to say, requires both fluent real-world coupling and the capacity to improve such engagements by the use of de-coupled, off-line reasoning.
- Cognizers, on our account, must display the capacity for environmentally decoupled thought and the contemplation of options. The cognizer is thus a being who can think or reason about its world without directly engaging those aspects of the world that its thoughts concern.

In 2006 international researchers representing many of the above disciplines attended the Cognitive Robotics, Intelligence and Control Workshop (COGRIC) [2], which discussed the future of cognitive robotics. Important concepts in cognitive robotics were held to be the ability to form internal states or models for reasoning and decision making (consequently planning), learning from experience, self reflection, embodiment and situatedness, perception to measure world, external behavior/internal representation and importing ideas from human cognition including perception and learning (the entire system) whilst understanding brains and minds. From discussions, it was concluded that it was not worth debating boundaries between consciousness, cognition and intelligence. Future opportunities included lifelong adaptability and developmental learning.

Other researchers have considered this topic, with different list of key features that a cognitive robot should/would possess. Key features that participants identified in the workshop are:

- Ability to perceive the world in a similar way to humans (or better) (e. g., "active perception", Dana Ballard [3], "ecological approach to perception", JJ Gibson [4])
- Ability to communicate with humans using natural language and mental models (robust human-robot interaction, such as overcoming the frame of reference problem, Alan Schultz [5])
- Ability to develop cognition through sensory-motor coordination (e. g., "morphological computation", Rolf Pfeifer [6])
- Ability to have a sense of self awareness (internal states, machine consciousness: Igor Alexander [7], Owen Holland [8] vs. Kevin O'Reagan [9])
- Ability to use attention and emotion (not in a sense of social robots) to control behaviors (K. Kawamura [10])

Robots may be justified in terms of ameliorating a labor shortage, working in hazardous environments, and undertaking repetitive tasks. However, the importance of cognitive robotics is in enabling robots to work autonomously in real-world environments. Currently, robots are not intelligent enough to operate without supervision in real-world situations, e. g. due to difficulties in applying existing knowledge to new situations, selecting between competing goals, coping with multiple sensory input and completing multiple tasks each with subtle variations. Deliberative robots may be able to weld a car, but higher order cognition (Fig. 1) is needed to care for the elderly as the environment is much more unstructured, dynamic, sensory rich, noisy, multitask, varied input/communication, emotive and interactive.

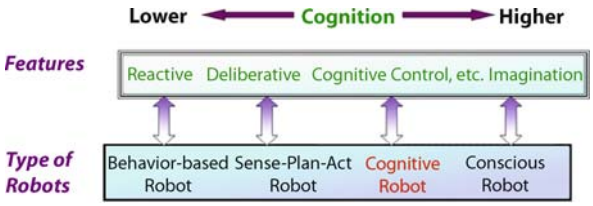

Cognitive Robotics, Figure 1
A spectrum of cognition in robotics. Adapted from [11]

Finally we quote Pentti Haikonen, Nokia Research Center:

"The next big push in robotics may well be machine cognition. Cognitive robotics, that are able to perceive and move as well as we humans do, are the focus of several national and international research programs in EU, Japan and USA. The emerging cognitive technology (as opposed to cognitive science) tries to emulate human information processing and in doing so utilizes experience, affordance (in terms of meaning), attention and even emotion" [2].

Introduction

Humans have sought to create human-like artificial creatures for menial tasks, for coworkers and for the intellectual challenge. Historical stories of Golems raised by religious rituals from mud have lead to the science-fiction of androids assisting or terrorizing the human race. Fiction has now been replaced by increasingly serious attempts by industry and academics, which is treated here as it relates to complexity.

Complexity covers systems where the interaction of multiple components leads to behaviors unachievable by any individual component alone. Cognitive robots require multiple components to function, often in parallel and in harmony, with the goal to produce intelligent human-like behavior and skills.

The origin of modern cognitive robotics comes from the field of cybernetics, the study of control and communication in living organisms, machines and organizations. The term cybernetics was popularized by Norbert Wiener in his 1948 book [12]. Cybernetics had a crucial influence on many important concepts, such as goal-directed behavior generation, self organization and situated nature of intelligence, which are now commonly used in the intelligent robotics community.

In the 1960s, researchers in artificial intelligence (AI) pursued the concept of intelligence from a more deterministic point of view. AI, or commonly known as strong AI, was predicated on the presumption that intelligence or knowledge can be represented as production systems and stored inside of a machine, i. e. computer. However, researchers in cybernetics claim that intelligence is an attribute of an interaction with the environment rather than a commodity stored in a computer and must be actively constructed by a machine itself (e. g., [13]).

The concept of strong AI is that a machine's processors can become a mind, exhibiting many aspects of intelligent behavior, especially sapience and insight into its inner workings. Weak AI, on the other hand, maintains that machines can only reproduce human attributes through the interaction with the environment, which sets constraints on cognitive robots. The best known example of a counter argument to claims made by the strong AI approach is the Chinese Room argument made by John Searle [14]. The debate Searle caused continues today.

In 1999, Clark and Grush [1] pointed out the problem with the notion of internal knowledge or world representation (a part of strong AI argument) and offered a solution called the Minimal Robust Representationalism (MRR). According to MRR, "Cognition, we want to say, requires both fluent real-world coupling and the capacity to improve such engagements by the use of de-coupled, off-line reasoning" ([1], p. 13). We will revisit this notion later in Sect. "Case Study".

As the opening paragraph stated, building a humanoid robot with which to embody human-level intelligence has been a dream of many AI and robotic researchers. In 1993, Rodney Brooks began to work on a first generation cognitive robot [15], which was an upper-torso humanoid and used to generate human-like behaviors. His effort led to a number of government and commercial humanoid development projects in Japan, such as Honda's ASIMO [16] and Toyota's personal assistant humanoid robot [17] during the 1990s and early 2000s. One of the strong arguments for humanoid robot development was that there is a strong desire to replicate human behavior within embedded artificial agents and humanoid robots are the best embedded system to do so. Other researchers consider the field too new to draw this conclusion, thus the jury is still out, e. g. no one knows whether an elderly person prefers a human-like robotic companion to a robotic wheelchair that provides specific needs for the person.

Recently, a number of international projects aimed to develop humanoid robot capabilities that can be used to test developmental theories of cognitive development and language acquisition were initiated (e. g. RoboGroup [18], RobotCub [19], Synergistic Intelligence [20]). The overall objectives of these projects are to study the perceptual, representational, reasoning and learning capabilities of embodied systems in human-centered environments. In particular, these projects propose to use humanoid robots

to study mental development processes through robots' interaction with the environment similar to children's mental development processes. For example, RobotCub project, one of the most ambitious among them, aims to develop a set of fully functional humanoid cub robots the size of a 2-year old child.

The aim of this chapter is to provide an overview of the emerging field of cognitive robotics that will evolve over the years into a truly interdisciplinary field. The organization of the rest of the chapter is: Sect. "What Is Cognition?" covers the concept and definition of cognition. Section "Cognitive Architectures" covers major approaches to cognition: symbolic, connectionist and hybrid. Section "Cognitive Robotics" gives an overview of various cognitive robots in terms of important aspects and formative examples. Section "Case Study" summarizes Kawamura's attempt to realize a humanoid robot emphasizing the cognitive task of internal rehearsal. Finally Sect. "Future Directions" provides future directions for the field of cognitive robotics.

What Is Cognition?

This section explores cognition in its influence on constructing advanced robots, rather than debating philosophical insights into its nature. By investigating the brain's functionality, we can gain insight into how to form an artificial cognitive mind.

A goal of advanced/cognitive robots is to endow a level of intelligence and cognitive skills typically associated with people or animals. Although the goal may be based on nature, the methods do not have to be naturally inspired. Much of the early classical symbolic AI-type cognition work (details in Sect. "Cognitive Architectures") were based on computer science techniques, such as production rules, finite state machines and functional programming. Due to issues with scalability, generality, robustness, disturbance rejection and coping with environmental properties (such as uncertainty, noise, nonlinearities, non-determinism and epistasis), biologically inspired methods, such as connectionism became popular in the 1980s.

Although robotic cognition often takes inspiration from biological evidence, it does not guarantee that the outcomes will be equivalent due to discrepancies in evidence and implementation. Investigative techniques, from cell straining to functional magnetic resonance imaging or fMRIs, have been used to 'map' the brain. However, there is much debate on the accuracy and interpretation of the results, with the caveat that the brain's structure/functionality is "not as simple as it appears" often stated. It is argued that a boxology (where functional boxes are connected by lines), which may be easily represented in a computer program, could not replicate the biological evidence.

Mapping brain activity in awake, functioning animals is considered more valuable for cognitive research than when the animal is sleeping, e. g. sensory–motor coordination requires interaction with an environment. Insight can also be gained into self organizing and self reflection during sub-conscious processes using modern neuroscience methods.

Tests for cognition have a tradition in animal as well as human studies, e. g. maze exploration (path planning), T-maze latent learning, mirror test, and relational studies. These are being adapted for robots, e. g. recognition of own mirror image by a small robot [21].

The above tests identify key features and attributes for cognition, but individually do not represent a definition for recognizing cognition. Cognition may be considered as ill-defined due to many alternative definitions. A simple definition came from the foresight initiative by the United Kingdom Department of Trade and Industry (UKDTI) that states:

"Cognitive systems are natural or artificial information processing systems, including those responsible for perception, learning, reasoning and decision-making and for communication and action" [22].

Applying definitions is also controversial, e. g. at what point does something stop being cognitive? (e. g. are ants cognitive as an individual and/or do they display collective cognition?).

Cognitive Architectures

The task of developing a general cognitive architecture and subsequent computational models is greatly complicated by the lack of a clear definition of cognition as discussed in Sect. "What is Cognition?". Nevertheless, cognitive architectures are designed to propose structural properties of the modeled cognitive systems such as humans. Cognitive architectures are classified as symbolic, connectionist, or hybrid (i. e. symbolic-connectionist). As John Anderson, one of the pioneers of the unified theory of cognition, observed, "it is difficult to answer a question regarding which (architecture) might be the correct one or the most correct one since these architectures are often quite removed from empirical phenomena that they are supposed to account for" [23]. Nonetheless, cognitive architectures generally implement cognition as a whole as opposed to cognitive models (the topic of Sect. "Cognitive Robotics"), which focus on particular problems or applications. The aim of this section is to provide the reader a short summary and the

important aspects of symbolic, connectionist and hybrid approaches using representative cognitive architectures in the fields of cognitive science, artificial intelligence, cognitive neuroscience, and machine cognition.

Symbolic Approach to Cognition

The symbolic approach to cognition can be best seen in Newell and Simon's physical-symbol system hypothesis [24] which states that "a physical symbol system has the necessary and sufficient means of general intelligent action". It means that any system that manipulates symbols is sufficient for producing intelligent behavior, and further that all intelligent systems are necessarily implementations of physical-symbol systems. The symbolic approach to cognition dominated the filed of cognitive science and artificial intelligence during the 1970s and 80s. Well-known examples of architectures that fall within this paradigm include ACT-R, Soar and EPIC.

ACT-R (Adaptive Character of Thought-Rational) ACT-R is a cognitive architecture for analyzing human cognitive performance which was developed by John R. Anderson and his group at Carnegie Mellon University [25]. ACT-R has been inspired by the work of Allen Newell and also inspired by the progresses of cognitive neuroscience. It is not a computational model, rather it is a framework for modeling specific human cognitive abilities whose predictions can be compared with human performance. That is, models of assumptions in cognitive processes can be created using ACT-R, allowing results to be easily observed, visualized, and compared with human performance data.

ACT-R makes use of symbolic processing and a sub-symbolic network. Symbolic processing is done through a set of production rules of the classic IF … THEN form. The sub-symbolic level runs concurrently with the symbolic level and consists of parallel processes that influence the performance of the symbolic system, such as conflict resolution, data retrieval, and rule execution. This allows the designer to have extra control over various conditions that affect the system performance without having to create explicit production rules. The central feature of the architecture is that all processing depends upon the current goal of the system. Functionally, ACT-R consists of three basic components: modules, buffers, and a pattern matcher as shown in Fig. 2.

There are two types of modules: perceptual-motor, and memory modules. Perceptual-motor modules (i. e., Visual and Motor Modules in Fig. 2 Declarative Memory) interface with the real world by perceiving environment and

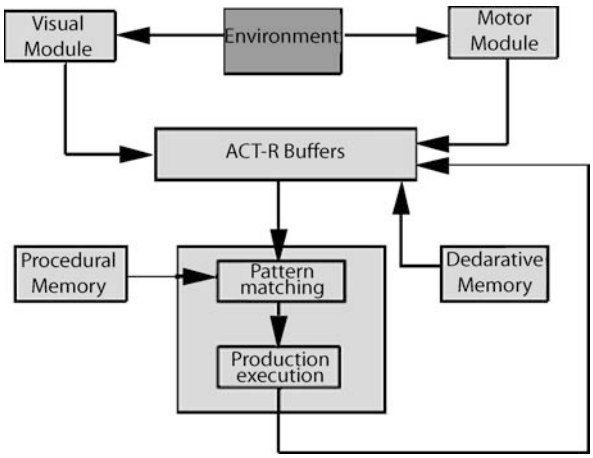

Cognitive Robotics, Figure 2
Functional structure of ACT-R. Adapted from [25]

generating actions. Memory modules represent ACT-R's most important assumption, i. e. human knowledge can be divided into two kinds of representations: declarative and procedural. Declarative memory consists of facts which can be used to represent goals. Procedural memory describes tasks in the form of production rules. Each production rule consists of a condition in which it should be executed, and an instructional component describing what to do.

Buffers hold declarative memory modules to be used by ACT-R and are somewhat analogous to the human working memory system [26]. Declarative information that is relevant to that goal is activated and considered to be in the focus of attention. The contents of the buffers are analyzed and a production rule is matched by the Pattern Matching module. A production rule is selected based on how well the conditions of execution associated with the production rule are met. The reader is referred to the ACT-R web site [25] for further information on the architecture and useful links.

SOAR (State, Operator And Result) Soar is another example of a production system-based cognitive architecture that implements goal-directed human behavior. It was created by John Laird, Allen Newell, and Paul Rosenbloom at Carnegie Mellon University [27]. Soar has been widely used by AI researchers to model different aspects of human behavior. Initially, Soar had been presented as a problem-solving architecture for artificial intelligence. Since then, Soar has been applied both within artificial intelligence and psychology for modeling human cognition.

Similar to ACT-R, underlying the Soar architecture is that knowledge representations can be made in forms of

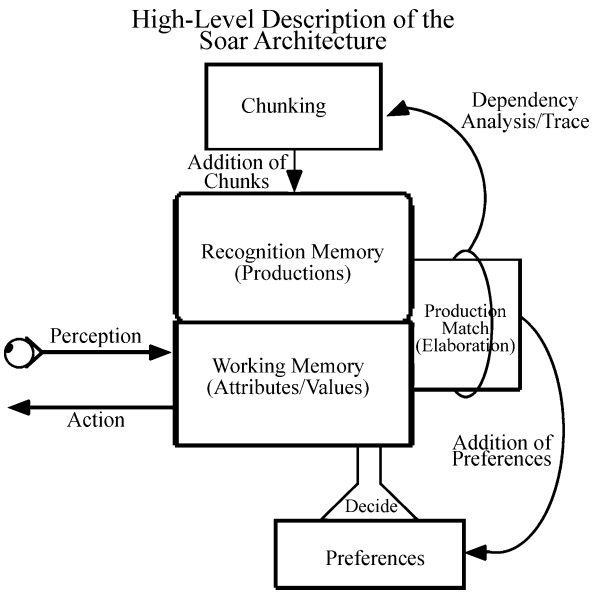

Cognitive Robotics, Figure 3
Soar cognitive architecture. Adapted from [28]

procedural, declarative and episodic. Also Soar tries to address a collection of mechanisms of mind. Figure 3 illustrates a high-level Soar architecture showing major mechanisms.

Soar represents all tasks as collections of problem spaces, following Newell's problem space principle [29]. Problem spaces are made up of a set of states and operators that manipulate the states. Soar begins work on a task by choosing a problem space, then an initial state in the space. Soar represents the goal of the task as some final state in the problem state. There is a single current state that represents the information about the problem or task being processed, including information about the current situation, the current state of the model, the goal toward which the model is working, and information relevant to that goal. This information is brought into the Soar's dynamic memory (SDM). All productions that match the current contents of dynamic memory then fire in parallel. The results usually cause changes in the dynamic memory, so other productions may now match. Those productions are then allowed to fire causing further matching. This process continues until no new productions fire. At this time, there may be a set of possible actions available as results from the productions firing. Action decision is processed by the *preference system*. When an action is performed, this usually results in a new state, and thus the operation can be repeated.

All learning in Soar occurs through the "chunking" mechanism, which is a form of explanation-based learning and a by-product of impasse resolution [30]. When an impasse is resolved and results in a change in the contents of dynamic memory, information about the resolution is stored as a new production rule. Impasses occur frequently during problem solving in Soar therefore, so does learning. The reader is referred to the Soar web site [28] for complete documentations and useful inks.

EPIC (Executive-Process/Interactive-Control) EPIC is a cognitive architecture developed by David E. Kieras and David E. Meyer at the University of Michigan [31]. While ACT-R and Soar focus primarily on the internal information processing of humans, EPIC extends to include the perceptual and motor aspects of human cognition. A schematic overview of the EPIC architecture is shown in Fig. 4. It consists of a cognitive processor in the form of a production system, a set of perceptual and motor processors and a simulated task environment allowing interaction with the surroundings. EPIC has three types of simulated sensory organs: visual, auditory, and tactile. For each sensory organ, there is a perceptual processor that works in parallel to each other. These processors receive information from the sensory organs, convert it into a symbolic form, and send the symbolic representation into working memory.

There are motor processors for each of the three sensory processors: ocular, vocal, and manual. The inputs to the motor processors are abstract symbolic representations of responses, which are then translated into explicit responses and passed to the motor processors. Information from the motor processors is passed to the working memory as a feedback.

Overall control of this architecture is attained through the use of task-specific control processes. There is no central control structure. Instead, production rules are used that are of the same form, but functionally separate from those used to encode information. Limitations in the EPIC architecture are postulated to be in the capacity of perceptual-motor processes.

The most important issue being pursued with EPIC is said to be the nature of human multiple-task performance. Major applications are user interface design to telephone operator workstations and cockpit systems, in which operation speed is critical and multiple perceptual and motor modalities are involved. The reader is referred to the EPIC web site [33] for complete documentations and useful inks.

Connectionist Approach

The connectionist approach to cognition provides an alternative theory of mind to the symbolic approach. It be-

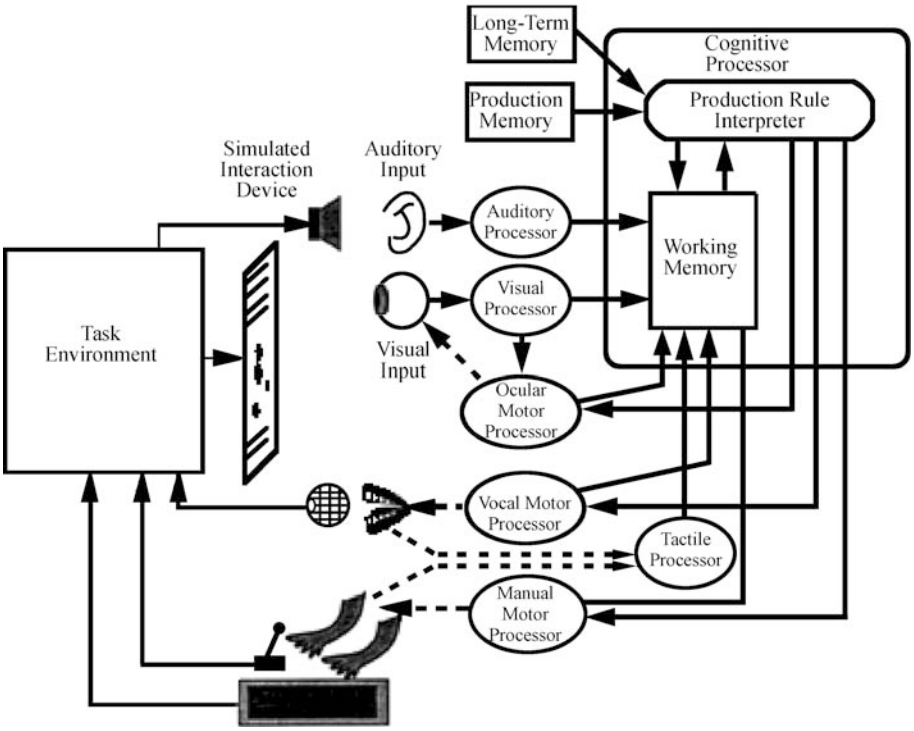

Cognitive Robotics, Figure 4
EPIC architecture. Adapted from [31,32] with kind permission from Professor David Meyer

came popular in the 1980s and its popularity has not been diminished. The connectionist approach differs from the symbolic approach in almost all major dimensions. Connectionist (or parallel distributed processing) models offer many attractive features when compared with standard symbolic models. They include a level of biological plausibility, parallel distributed representations, pattern generalization performance, and adaptive learning.

Connectionist models represent information throughout a network of simple, but highly interconnected units or nodes. In some connectionist networks, each unit has a particular meaning. In others, such as neural networks, the nodes are individually meaningless and information is represented as a function of the simultaneous activation of multiple nodes. Neural networks [34] are one of the most popular connectionist models used in many fields, including artificial intelligence and robotics. The number of different connectionist architectures available today is large; to discuss them all is beyond the scope of this section. Instead, this section will focus on two unique architectures, which the authors believe generic enough to give the reader a good sense how to design connectionist-based cognitive robots. For more general discussion of connectionist approach, the reader is referred to an excellent review [35]. For more technical details, the reader is referred to, e. g. the Handbook on Parallel and Distributed Processing [36].

Connectionist Cognitive Machine Architecture Pentti Haikonen of Nokia Research Center takes a connectionist approach to build an embodied cognitive architecture. Instead of creating specialized models that reproduce cognitive processes directly at higher representational level like ACT-R, Haikonen proposes a cognitive architecture based on a distributed signal representation as a building block as shown in Fig. 5. In Fig. 5, "the preprocessed sensory information in the form of distributed signal representation is forwarded to feedback neurons, which also accept feed-

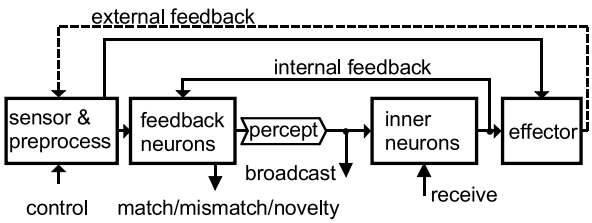

Cognitive Robotics, Figure 5
Perception/response loop module. Reprinted from [37] with kind permission from Dr. Pentti Haikonen

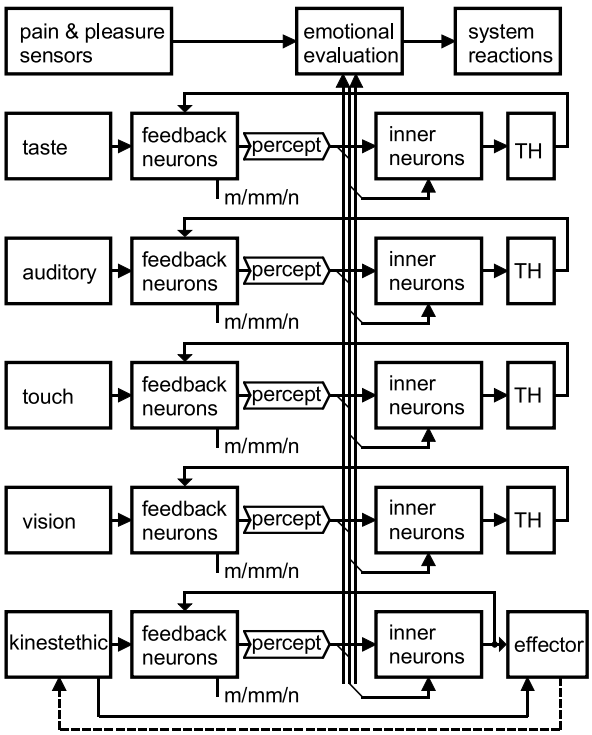

Cognitive Robotics, Figure 6
Haikonen's cognitive machine architecture. Reprinted from [37] with kind permission from Dr. Pentti Haikonen

back signals from the system". The feedback and the sensory information are compared with each other and the eventual match, mismatch and novelty conditions are indicated by respective signals. The internal feedback indicates also the intended act that is to be executed subsequently by the effector and will therefore be expected as the corresponding sensory percept. In that case, "the feedback signal is comparable to the so-called corollary discharge which is forwarded from the motor area to sensory areas in the brain" [37].

Figure 6 illustrates a schematic diagram of Haikonen's cognitive architecture. The cognitive architecture determines the cross-connections between the sensory perception/response loop modules of the Fig. 5. The architecture allows the association of complex additional meanings with percepts and also allows the generation of emotional significance, the "emotional soundtrack" to be associated with memorized entities and episodes. This emotional significance is then used to control attention. Each module works on its own and produces streams of percepts about environment and the module's own internal states. This output is then fed back to the feedback neurons and via this route becomes a percept.

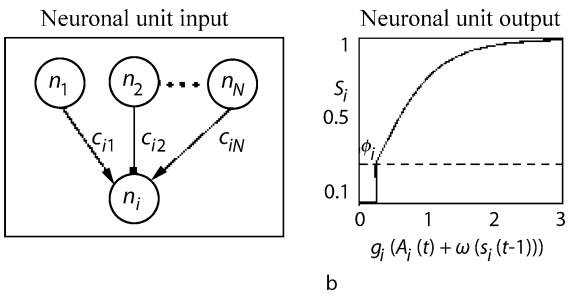

Cognitive Robotics, Figure 7
Typical neuronal unit model used in SNMs. Reprinted from [40] with kind permission from Dr. Jeffrey Krichmar

Synthetic Neural Modeling The synthetic neural modeling approach is a general method of testing theories of brain function at the system level [38]. It was developed by a group of researchers at the Neurosciences Institute over the last two decades. These models are designed to demonstrate the theory of neuronal group selection (TNGS) proposed by Edelman [39]. That is, it is based on the premise that neuronal circuits in the brain, formed with functional properties, the value of which is not knowable a priori, compete with one other to participate in the determination of behavior via strengthening of their connections with other brain circuits. Basic unit of selection in the TNGS are neuronal group, collections of dozens, or hundreds of neurons that are strongly connected to each other via synapses. A synthetic neural model (SNM) comprises a device, an environment and a neural system simulation [40]. Figure 7 shows a typical SNM unit (a) and its output (b). In (a), neuronal unit i receives input from N neuronal units via synaptic connections labeled c_{ij}, $j = 1, 2, \ldots, N$. Each connection has a relative strength that can be either excitatory (e. g. c_{i1} and c_{iN} are positive) or inhibitory (e. g. c_{i2} is negative). In (b), the output of a neuron is subject to a function based on its current activity and the input from other neurons. Below a specific threshold φ_i, the output is 0. SNMs have been embedded in a series of mobile robots called Darwin [38]. Section "Cognitive Robotics" includes a short description of such a Darwin model.

Hybrid Approach

The hybrid architecture is intended to capitalize on the complementary strengths of production-systems and connectionist architectures to implement a human dual processing theory of controlled and automatic processing.

CAP2 (Controlled and Automatic Processing 2) CAP2 is a hybrid cognitive architecture, incorporating both sym-

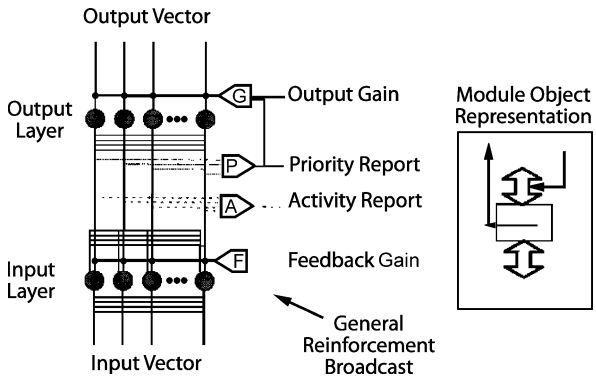

Cognitive Robotics, Figure 8
Microstructure of a CAP2 module. Reprinted from [41] with kind permission from Professor Walter Schneider

bolic (e. g., ACT-R) and connectionist (e. g., PDP) elements. It was developed by Walter Schneider and others at the University of Pittsburgh [41]. It is based on a human dual processing theory of automatic and controlled processing [42]. The CAP2 architecture is implemented with entirely connectionist components, but has networks that operate as sequential control structures to behave like a production system. The basic module of the CAP2 architecture is shown in Fig. 8. It involves two layers of units and five control signals. It is modeled after the cortical columns in the human cortex. An input vector of activation enters the module from below, passes through a connection matrix that reflects prior learning and evokes a new activation pattern in the input layer. The input layer then activates the output layer through a second connection matrix and sends a vector of activation to other modules.

A schematic overview of the CAP2 architecture, which implements a hierarchy of modules or modular processors is shown in the Fig. 9. At the top of this hierarchy is an executive control network, which is responsible for the control of the flow of information between modules. This control of information flow is conducted according to the current goal of the system, thus making it conceptually similar to the ARC-R architecture.

Cognitive Robotics

Now that the relevant aspects of cognition have been identified and the formative architectures in the field have been outlined, this section explores cognitive robotics. It does not attempt to provide necessary and sufficient measures for determining if a device is a cognitive robot as the field is still maturing and such definitions are likely to prove limiting when exploring the still somewhat unknown nature of cognition. Nor will this section present a history of cognitive robots as it is more insightful to detail important aspects of the field illustrating concepts with features of individual robots.

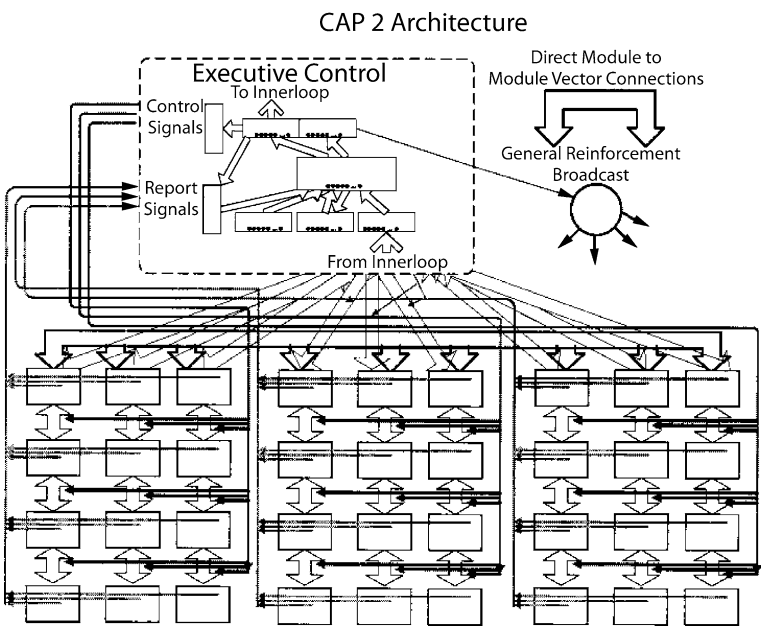

Cognitive Robotics, Figure 9
The CAP2 architecture showing three regions with four levels in each region. Reprinted from [41] with kind permission from Professor Walter Schneider

The first aspect is that a cognitive robot is embodied, which is commensurate with the situated nature of intelligence. It is noted that most robots have self-contained processing relating to a single environment often with a single objective. Increased wireless data transfer rates have resulted in the possibility of the processing being remote from the sensing and action. Furthermore, a single processor could now be embodied in multiple locations simultaneously, whilst communicating with other embodied and remotely embodied processes. Note: a disembodied processor has no online environmental interaction, which is employed occasionally where an offline agent is used to quickly develop promising strategies based on the model of the real environment built or learnt from embodied interactions.

Arguably the first robots exploring cognition were built by the cyberneticist Grey Walter in the late 1940s. The idea of cognitive robotics had not been developed in 1948 when the first tortoise was designed. The legacy of Grey Walter [43] is such that some of his ideas have influenced biologically inspired robotics and are still relevant today. The tortoises were based on analogue electronics as digital electronics barely existed at the time. They relied on relays – threshold devices – to implement behavioral transitions. Phototaxis behaviors were obtained as well as touch-controlled behavior and results were reported on the mirror test [44].

Walter was also an eminent neuroscientist, a field that continues to have much input and importance into cognitive robotics. This highlights the equal support of all the associated fields. This support is a two-way process, e. g. the intellectual challenge of creating cognitive robotics includes testing insights into natural cognition. One unique robot, or a design study, along this line is CRONOS robot (Fig. 10) being developed by Owen Holland [45]. Holland calls his approach *anthropomimetric robotics*.

Cybernetics arose from the realization that many seemingly diverse fields share similar properties; namely the importance of feedback from the environment to a system. This feedback is central to communication and control in order to achieve efficacy of action. The theoretical insight from cybernetics into dynamical systems is very relevant to cognitive robots.

First-order cybernetics considers the effect of the environment through feedback on the robot, whilst second-order cybernetics acknowledges that the robot can also affect the environment and thus the meta-system must be taken into account. Consequently, the environment can determine the level of cognition that any robot can display. An example is where an exploratory robot has to learn only not to collide with obstacles, so the robot learns to stay away from obstacles by staying still. Braitenberg, a cybernetician and a neuroanatomist, originated the Braitenberg Vehicles as a thought experiment to "illuminate the key issues of what we may call cybernetics or artificial intelligence or cognitive science" [46] based on adjusting the morphology of a vehicle in order to respond differently to similar environmental stimuli [47].

Cognitive Robotics, Figure 10
CRONOS robot. Reprinted with kind permission from Professor Owen Holland

Morphological computation is an important modern research field where the robot is designed to complement the environment. Whereas evolution has had millennia to adapt creatures to their environment in order to exhibit cognition, robots must be designed to fit their environment. An example is the robot Puppy designed by Iida and Pfeifer [48], where stable gates are obtained through efficient mechanical design and simple processing rather than convoluted cognitive learning. A morphological design guide, adapted from Pfeifer and Scheier [49], is shown in Fig. 11.

Perception interfaces with the environment, so the sensor set-up of the robot is core to cognition. Human eyes are tuned to the wavelengths of primary colors so will have a stronger reaction to these colors (e. g. red) than a robot with a CCD camera that treats visible frequencies equally. Humans have multiple types of sensors (more than the anecdotal five), whilst robots may have many more (such as laser range finding). Thus robot cognition of an environment will differ, subtly or significantly, from human cognition.

Even if a human and a robot could be instantaneously given the same perceptions there is the *explanatory gap*, where the quality of the experience is important. Smelling

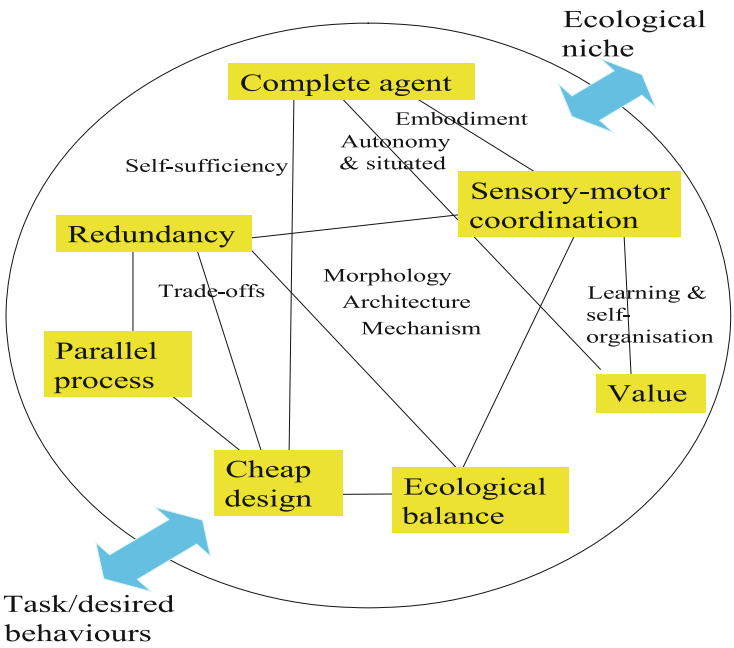

Cognitive Robotics, Figure 11
Morphological design guide. Adapted from [49]

a red rose will bring a deeper quality of experience to a human due to past associations. Humans also utilize *active perception*, where the perception of an object is deliberately manipulated in order to experience it to a greater extent. Thus a cognitive robot will need to perceive by exercising a skill, rather than passively receiving information.

The perceptions need to be represented internally within a robot in order to manipulate them to determine the appropriate action. Either connectionist or symbolic representations may be utilized. Connectionist approaches reputedly mimic the representation in the human brain, which is difficult to understand. Symbols have proved useful when transparency for human readability and understanding of cognitive processing is required. The *symbolic grounding problem* [50] addresses how symbols get their meaning and in a deeper sense, what it is to experience the symbol. The word 'red' is very emotive to humans, but storing the letter sequence 'r e d' in a memory address does not have the same connotations. The more symbols are experienced, the greater the grounding.

A connectionist architecture based on the biological evidence has been successfully used by Krichmar [40] to replicate the behavior of rats in a Morris Water Maze (a simple water-based maze where a rat is required to use memory and path planning in order to locate a safe platform) (Fig. 12). Through much post processing of the learned artificial neuronal links an additional artificial pathway was deemed necessary, which was subsequently found to exist in nature. This is an example where cognitive studies can be enhanced through cognitive robotics.

Although it is possible to perceive and represent a multitude of external information it would be too slow and take too much memory if every detail was processed and stored. Thus a cognitive robot must have *attention* on the salient features of the domain, whilst ignoring others. This reasoning continues as the actions are available to the agent (*affordances*) within a domain must be determined for analysis/selection. Classical search is too slow so key feature selection is required. Once an action has been selected the *frame problem* (Frame Problem in Artificial Intelligence) occurs as the cognitive robot must determine what to update dependent on the results of effecting this action.

It has been argued that the mind has a central executive that acts as a librarian for memories [51]. Although this may provide a practical solution, it is undetermined what tells the central executive how to act. As an alternative approach, Noelle proposes that:

"Since working memory is employed extensively during cognitive processing in humans and animals, it is reasonable to conjecture that robots could benefit from the capabilities provided by a working memory system. Benefits could include the focusing of attention on the most relevant features of the current tasks" [52].

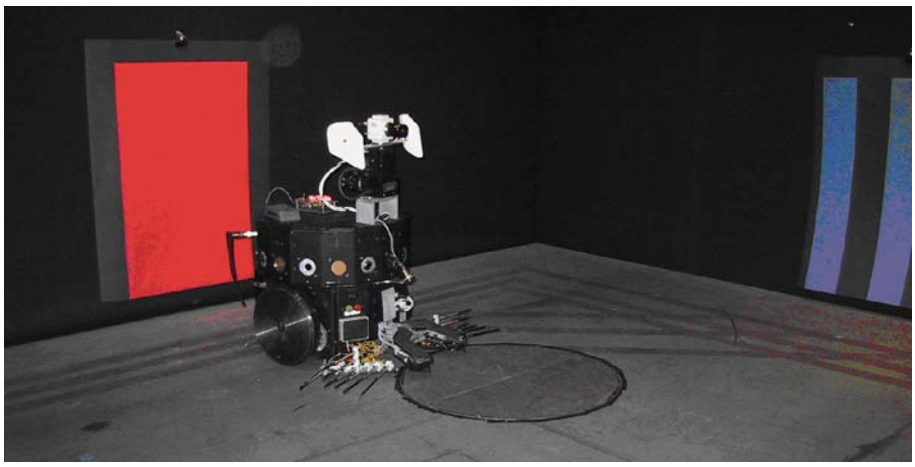

Cognitive Robotics, Figure 12
Darling robot exploring dry version of Morris Water Maze. Reprinted with kind permission from Dr. Jeffrey Krichmar

It is worth noting here that memory is not considered uniform in humans, with types such as episodic, procedural and semantic being identified [53]. Episodic memory encodes individual events, whilst semantic memory expresses the essence of memories. It is argued that a mechanism, such as abstraction, is required to transfer between episodic and semantic memory. Memory is also temporal in nature, which has caused much debate into its categorization. Long-term, short-term and working memory are hypothesized, but it is the ability to recall appropriate past events/lessons in a timely manner and compare these with prescient events that is important to cognitive robotics.

Learning from feedback encompasses one of the most active areas of AI research, so it will not be detailed here. Instead, a few important aspects for cognitive robotics are highlighted. To map the vast world into memory requires compactness as well as the recall of salient features. States, actions and feedback can be *imprinted*, which although reliable is memory intensive. The ability to *generalize*, i. e. remove irrelevant information, improves performance but more steps are necessary. *Abstraction* is where patterns, both spatial and temporal, are identified independent of their associations, e. g. the concept of a tree without recalling specific instances. The ability to scale learning is a main challenge in the field of AI.

Many advanced robotic platforms exist or are emerging, such as ASIMO [16] and iCub [19]. Unfortunately, many modern actuators require much power, space and advanced control. Lack of compliance, which is inherent in most animals and other material properties, can limit the behavior in a cognitive robot. CRONOS [45] is an example of a platform with natural actions to mimic human cognition. Thus a new 'breed' of robots will require new control methods as their nonlinear nature results in many existing control methods being invalid.

Now that we have considered an embodied robot with the rudiments of cognition, there are aspects of cognition, behavior and social interaction that need to be considered.

A cognitive robot will need to plan a series of actions over time. This will require a series of state-action links, an internal world model, or a method that utilizes the world itself as a model. Cartesian theater, global workspace [54] and other architectures have been proposed for such a model.

Despite the ability to memorize and learn from situations, a robot will need to react to unknown/unfamiliar situations. Interpolation and extrapolation are popular techniques, but are often hard to compute and cannot be guaranteed to be effective. Instead, artificial emotions are being considered as action modifiers, goal setters and decision makers. Emotions are also useful to aid human robotic interaction, but their cognitive uses are much wider [55].

Most previous robots were designed to achieve a single or limited number of similar tasks and goals. Increasing the range of competencies of a robot will require it to select between goals, so a value system will be needed on which to base decision-making. Science fiction may suggest hard-coded laws for robotics [56], but this could lead to conflicts and contradictions. Debates have begun on the ethics and legal aspects of giving robots autonomy in goal setting, especially if this enables new goals to be determined.

Robots operating in a domestic or social setting will need to assess the intentions of others to aid collaboration and prevent destructions/accidents. Humans have inherited neural circuitry to recognize faces and use body lan-

guage/gestures as important communication tools. Much work is required in this field and it is unproven on how humans will react to robots in a social setting [5].

Work in the 1970s by Masahiro Mori on the Uncanny Valley hypothesizes that a robot must appear and behave very humanlike to be acceptable [57]. Alternatively it is better to make the robot mechanical in appearance, so it would be acceptable as a mechanoid.

After considering what future cognitive robots may need it is worth noting that their processing is unlikely to be serial, silicon-based digital systems. Parallel systems are more timely when processing large quantities of data. Analog systems could encompass non-linear and possibly nondeterministic behavior. Finally, biological substrates, which exhibit both the above properties, are being developed for use in robot control [58].

Case Study

Introduction

This section presents a case study on the humanoid robot ISAC [59] that one of the authors (Kawamura) is developing. Specifically, the task of reaching for a named object involving internal rehearsal is explored as sense, memory and planning are all required and the implementation of an 'inner sense' where sensory experiences and consequences of different behaviors may be anticipated is developed.

The first lesson is that humanoid robots are indeed complex systems due to the number of components that must interact. Figure 13 shows ISAC that is an upper bodied humanoid robot incorporating stereo vision, audio and touch sensors with pneumatic actuation for multiple arm and hand movement. Sensing inputs are mapped to non-linear and flexible actuators through parallel, distributed processing that represent perception, reasoning and learning.

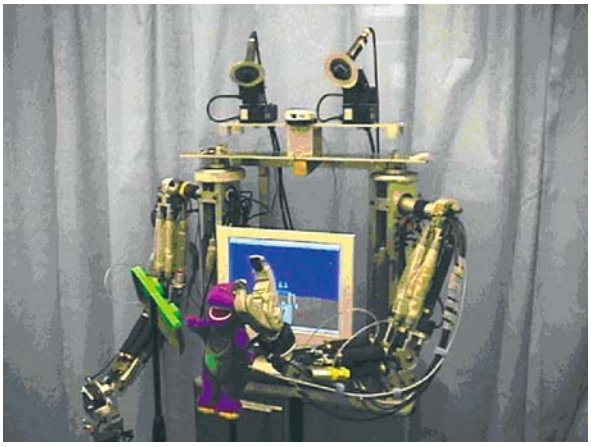

Cognitive Robotics, Figure 13
ISAC reaching to Barney with left arm

Architecture

During earlier development it was realized that enhancement and maintenance of such robotic software systems could benefit from domain-specific guidelines that promote code reuse and integration through software agents. This led his group to develop a multiagent-based robot control architecture based on the Intelligent Machine Architecture (IMA) [60,61]. IMA allows for modular design and the development of subsystems from perception modeling to behavior control through the collections of software agents and associated memories. Figure 14 was configured using IMA agents and associated memory structures. It consists of a number of IMA agents and a set of memory structures. Within the cognitive architecture, the Self Agent represents the robot's self [62] and is responsible for multiple aspects of cognition including internal rehearsal.

In the ISAC architecture, the Self Agent handles the dual sensory-motor loops as shown in Fig. 15. The First–Order Response Agent (FRA) is responsible for the reactive and routine responses of the system while the Central Executive Agent (CEA) is responsible for the cognitive response. The Internal Rehearsal System (IRS) takes the working memory chunks as the motor commands, the current situation as the external state and sends a rehearsed result to CEA. If IRS produces a poor prediction, CEA will suppress the Activator Agent, replace the working memory chunks, and tell the Activator Agent to switch action.

It is known that humans are able to have sensory experiences in the absences of external stimuli as illustrated by experimental results of, e. g. Lee and Thompson [63]. It thus seemed reasonable to assume the existence of an 'inner sense' where sensory experiences and consequences of different behaviors may be anticipated. The idea of the existence of such an inner sense (or model) does not necessary go against the theory of embedded intelligence advocated by a number of researchers e. g., Brooks [64], Clancy [65], Clark [66], Pfeifer [49] who de-emphasize the role of internal world models and instead emphasize the situated and embodied nature of intelligence. An alternative to internal world models is the 'simulation hypothesis' by Hesslow [67] which accounts for the 'inner world' in terms of internal simulation of perception and behavior. Kawamura's approach may be termed as a "grounded

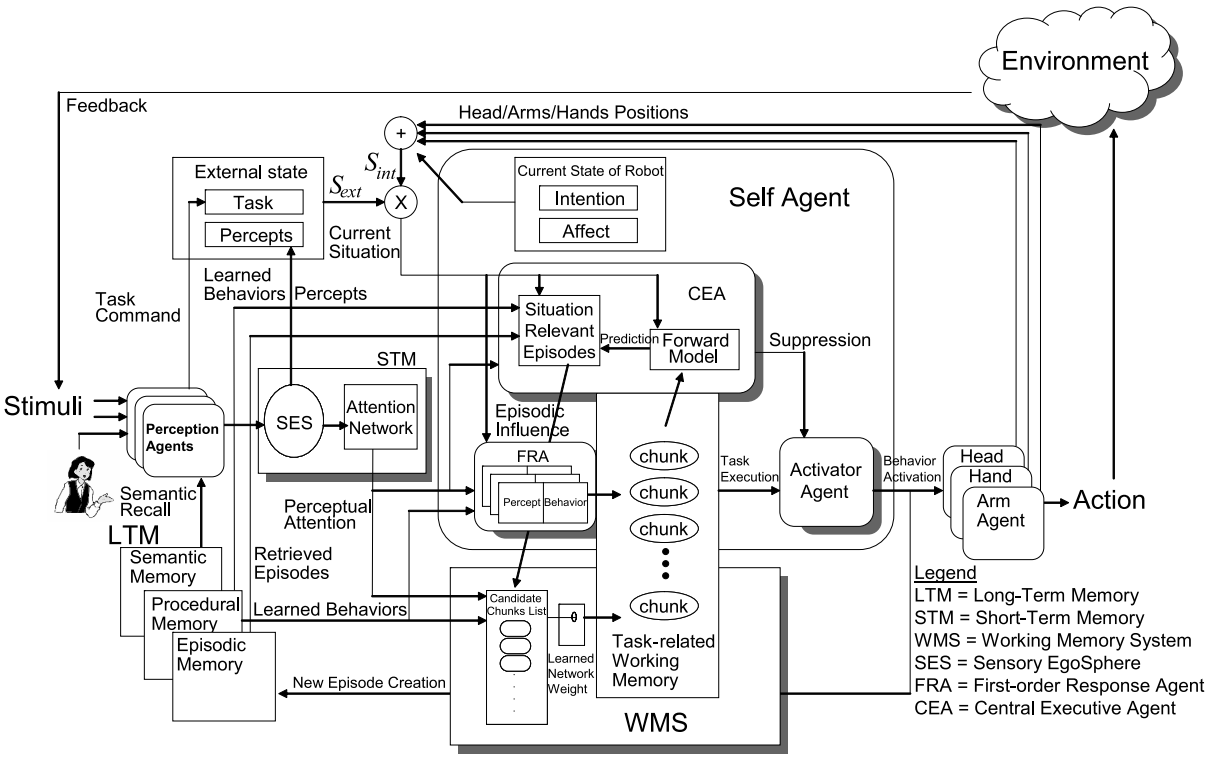

Cognitive Robotics, Figure 14
IMA-based cognitive robot architecture

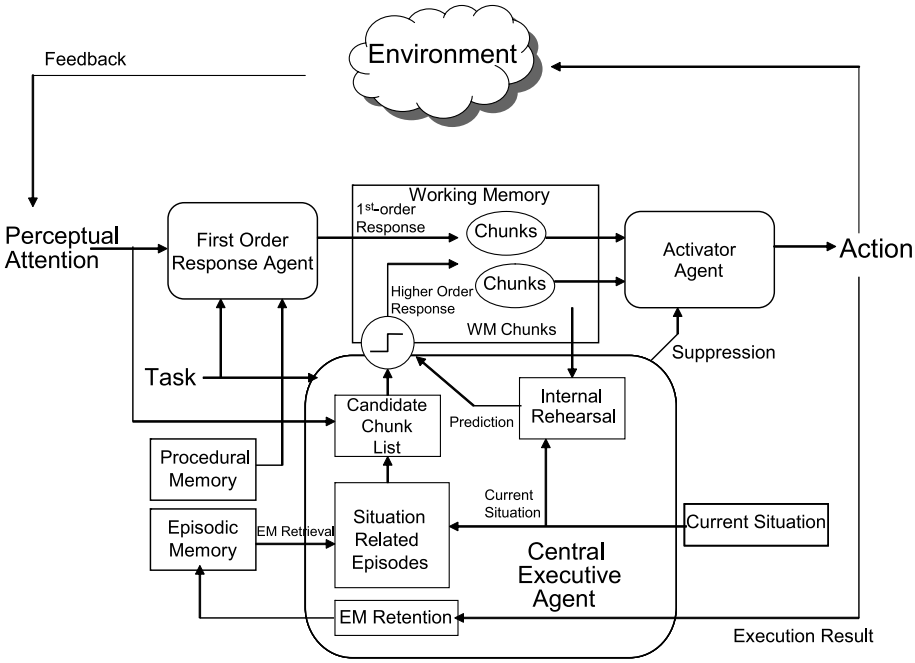

Cognitive Robotics, Figure 15
ISAC self agent cognitive cycle

internal simulation" utilizing one type of internal representation of perception and behavior.

Design of the Internal Rehearsal System (IRS)

Brain-inspired internal simulation research has now moved into the robotics field. For example, Shanahan [68] proposed a cognitive architecture for a mobile robot, shown in Fig. 16, involving two sensory-motor loops: the reactive or first-order loop and the cognitive or higher-order loop. The first-order loop involves the sensory cortex (SC), motor cortex (MC), and basal ganglia (BG). This loop directly maps sensory input to motor actuation. The higher-order loop internally rehearses the decision from the first-order loop and changes the output of the system based on the observation of this rehearsal through the Amygdala (Am) or the emotion system.

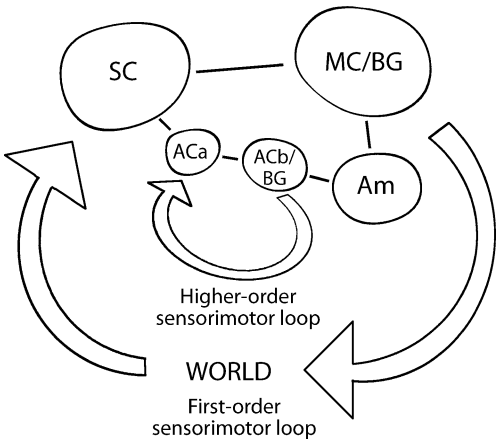

Cognitive Robotics, Figure 16
A top-level schematic diagram of two interacting sensory-motor loops in the brain. Adapted from [68]

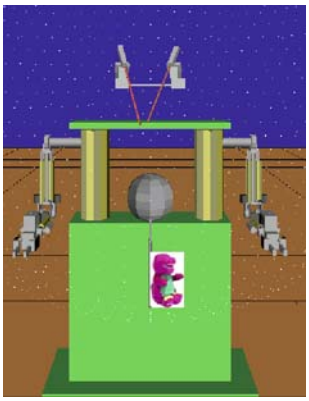

Cognitive Robotics, Figure 17
ISAC simulator displaying a Barney doll as a sphere

When IRS is invoked by CEA, it takes the current behavior chunk as the motor command and the current environment ISAC is in as the current state. After CEA selects a behavior to perform the skill described by the task, IRS internally rehearses the behavior with the percept corresponding to the current. If a collision occurs with the percept during the rehearsal, IRS returns the percept, the step in interpolation where the collision occurred, and the total number of joint steps in the interpolated motion to CEA.

The following experiment is designed to evaluate how FRA, CEA, and IRS work together. The experiment involves two percepts: Barney (target) and a Lego toy (obstacle) (Fig. 17).

1. A task to *reach-to-Barney* is given to ISAC. FRA immediately places ReachRight and Barney into the working memory (WM) as chunks.
2. Using the chunks, IRS will try to reach to the Barney with the right arm, but predicts a collision with the Lego toy.
3. CEA will suppress the Activator Agent based on this prediction from IRS.
4. CEA will use the episodic retrieval technique and replace the chunk ReachRight to ReachLeft.
5. IRS will reach to the Barney with the virtual left arm. This reach will be successful.
6. CEA will let the Activator Agent proceed to reach to the Barney using the left arm.

Performance

When ISAC was given a command to reach to the Barney, FRA placed two chunks "ReachRight" and "Barney" into the working memory (Note: ISAC normally begins with its right arm). Both the Activator Agent and IRS began to process these chunks. IRS completed the computation within 3.2 s and sent its results to CEA. At the same time, the Activator Agent sent a motion command to the Right Arm Agent to perform the reaching motion. The Right Arm Agent would take 11 s to perform this type of reach if no obstacle exists. When IRS finished, the following output was sent to CEA: [15 68 lego_toy]. This means that during the simulation, IRS determined a collision with the Lego toy in the fifteenth step of the interpolated *reach* behavior out of the total of 68 interpolated steps. Figure 18 shows the trajectories of the right arm collision points during the rehearsal. CEA took this result and determined that it did not reach to the Barney. CEA then suppressed the Activator Agent and prevented the right arm from further action.

CEA then decided to use the left arm and replaces the working memory chunks with "ReachLeft" and "Barney".

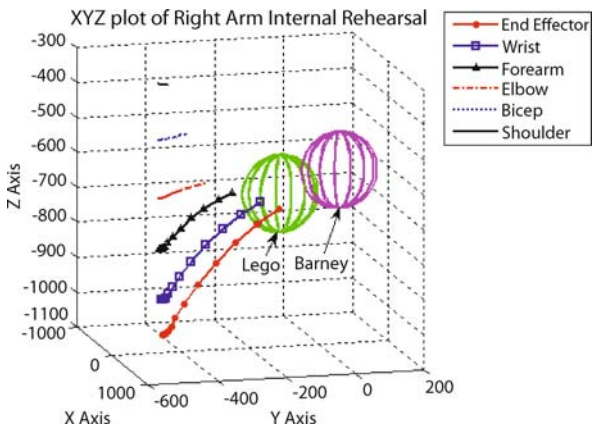

Cognitive Robotics, Figure 18
Right arm internal rehearsal

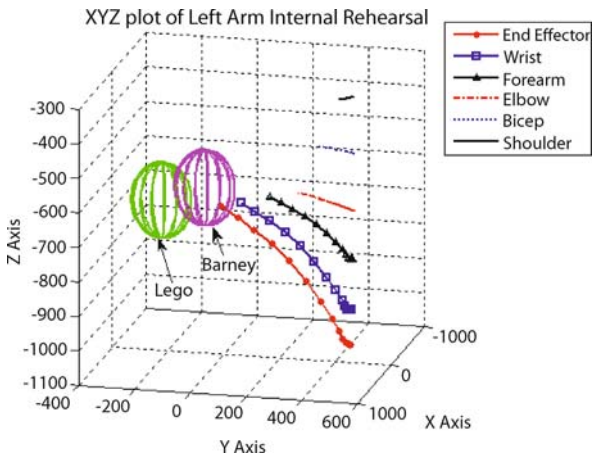

Cognitive Robotics, Figure 19
Left arm internal rehearsal

IRS and the Activator Agent were once again initiated, and IRS internally rehearsed the reach skill and determined no collision with the Lego toy (Fig. 19). Both the wrist and end effector points entered the Barney percept sphere on the sixteenth step out of the total of 69 interpolation steps. CEA determines this as a success and did not impede the Activator Agent thus allowing ISAC to reach to Barney using the left arm (Fig. 13).

Summary

This section illustrated how dual mechanisms of coupled real world of action-taking loop and off-line inner reasoning loop can work together to improve cognitive skills for a robot using internal rehearsal.

Future Directions

Cognitive robotics is still an evolving field with many possible and exciting future directions. To date, much progress has been made in architecture, perception, mobility, reasoning, learning from environmental embodiment, and advanced actuators. However, it will take years before we will see truly integrated cognitive robots in everyday life.

A few of the challenges to cognitive robots are listed below:

- Developing flexible systems, with ability to cope with multiple tasks, environments and disturbances
- Design strategies for learning by taking into account morphological and material constraints
- Cognitive skills development through social interaction
- Robots with mental states and emotions
- Modeling consciousness and its interaction with cognition
- Built in sensation and communication.

As robotic technology continues to penetrate every aspect of human society, the importance of social acceptance, such as trustworthiness and ethics, will become important. The field of cognitive robotics is expected to play a leading role in this area in the future. It is our hope that within decades, we will see true cognitive robots that will be accepted by the general public.

Bibliography

Primary Literature

1. Clark A, Grush R (1999) Towards a cognitive robotics. Adapt Behav 7(1):5–16
2. COGRIC (2006) Cognitive Robotics, Intelligence and Control. http://www.cogric.reading.uk/. Accessed 16–18 Aug 2006
3. Ballard DH (1991) Animal Vision. Artif Intell 48(1):1–27
4. Gibson JJ (1979) The Ecological Approach to Visual Perception. Houghton Mifflin, Boston
5. Trafton JG, Schultz AC, Bugajska M, Mintz F (2005) Perspective-taking with Robots: Experiments and models. In: IEEE International Workshop on Robots and Human Interactive Communication, Nashville, pp 580–584
6. Pfeifer R, Bongard J (2007) How the Body Shapes the Way We Think: A New View of Intelligence. MIT Press, MA
7. Alexander I (2005) The World in My Mind, My Mind in the World: Key Mechanisms of Consciousness in People, Animals and Machines. Imprint Academic, UK
8. Holland O (ed) (2003) Machine Consciousness. Imprint Academic, UK
9. O'Reagan JK http://nivea.psycho.univ-paris5.fr/. Accessed 28 Aug 2008
10. Kawamura K et al (2006) From Intelligent Control to Cognitive Control. In: 11th International Symposium on Robotics and Applications (ISORA), Budapest

11. Kawamura K, Browne WN (2006) Tutorial on Cognitive Robots. IEEE RO-MAN, Hertfordshire
12. Wiener N (1948) Cybernetics, or Control and Communication in the Animal and the Machine. MIT Press, MA
13. Heylighen F, Joslyn C (2001) Cybernetics and Second-Order Cybernetics. In: Meyers RA (ed) Encyclopedia of Physical Science and Technology, 3rd edn. Academic Press, New York
14. Searle J (1999) The China Room. In: Wilson RA, Kei F (eds) The MIT Encyclopedia of the Cognitive Science. MIT Press, MA
15. COG www.ai.mit/edu/projects/humanoid-robotics-group/cog/
16. Honda http://world.honda.com/ASIMO/. Accessed 28 Aug 2008
17. Toyota http://www.toyota.co.jp/en/special/robot/. Accessed 28 Aug 2008
18. RoboGroup http://www.isc.cnrs.fr/dom/RoboGroup/htm/. Accessed 28 Aug 2008
19. RobotCub http://www.robotcub.org/. Accessed 28 Aug 2008
20. Synergistic Inteligence http://www.jeap.org/humanoids/pdf/MAsada.pdf. Accessed 28 Aug 2008
21. Takeno J et al (2005) Experiments and examination of mirror image cognition using a small robot. In: IEEE International Symposium on Computational Intelligence in Robotics and Automation. Espo, Finland, pp 493–498
22. Foresight http://www.foresight.gov.uk/index.asp. Accessed 28 Aug 2008
23. Anderson JR (1983) The Architecture of Cognition. Harvard University Press, MA
24. Newell A, Simon HA (1972) Human Problem Solving. Prentice-Hall, NJ
25. ACT-R http://act-r.psy.cmu.edu/about/. Accessed 28 Aug 2008
26. Anderson JR, Bothell D, Byme MD, Douglass S, Lebiere C, Qin Y (2004) An integrated theory of the mind. Psychol Rev 111(4):1036–1060
27. Laird J, Newell A, Rosenbloom P (1987) Soar – An architecture for general intelligence. Artif Intell 33:1–64
28. Soar http://sitemaker.umich.edu/soar/home. Accessed 28 Aug 2008
29. Newell A (1990) Unified Theories of Cognition. Harvard University Press, MA
30. Young RM, Lewis RL (1999) The Soar cognitive architecture and human working memory. In: Miyake A, Shah P (eds) Models of Working Memory: Mechanisms of active maintenance and executive control. Cambridge University Press, NY, pp 224–256
31. Kieras DE, Meyer DE (1997) An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. Hum-Comput Interact 12:391–438
32. Meyer DE, Kieras DE (1997) A computational theory of executive cognitive processes and multiple-task performance: Part 1. Basic Mechanisms. Psychol Rev 104:3–65
33. EPIC http://www.umich.edu/~bcalab/epic.html. Accessed 28 Aug 2008
34. Abdi H, Valentin D, Edelman BE (1999) Neural Networks. Sage Publications, CA
35. Connectionism http://plato.stanford.edu/entries/connectionism/. Accessed 28 Aug 2008
36. Blazewicz J, Ecker K, Plateau B, Trystram D (eds) (2000) Handbook on Parallel and Distributed Processing. Springer, Germany
37. Haikonen PO (2006) Towards the times of miracles and wonder; a model for a conscious machine. Brain Inspired Cognitive Systems (BICS), Athens
38. Krichmar JL, Edelman GM (2003) Brain-Based Devices: Intelligent systems based on principles of the nervous system. In: IEEE/RSJ Int Conf on Intelligent Robotics and Systems. Las Vegas, pp 940–945
39. Edelman GM (1987) Neural Darwinism: The Theory of Neuronal Group Selection. Basic Books, NY
40. Krichmar JL, Reeke GN (2005) The Darwin Brain-Based Automata: Synthetic Neural Models and Real-World Devices. In: Reeke GN, Poznanski RR, Lindsay KA, Rosenberg JR, Sporns O (eds) Modeling in the Neurosciences: From Biological Systems to Neuromimetic Robotics. Taylor and Francis F, London
41. Schneider W (2000) Working Memory in a Multilevel Hybrid Connectionist Control Architecture (CAP2). In: Miyake A, Shah P (eds) Models of Working Memory: Mechanisms of active maintenance and executive control. Cambridge University Press, NY, pp 340–374
42. Schneider W, Shffrin RM (1977) Controlled and automatic human information processing: Detection, search, and attention. Psychol Rev 84:1–66
43. Holland O (2003) Exploration and high adventure: the legacy of Grey Walter. Phil Trans R Soc Lond A 361:2085–2121
44. Gallup GG Jr (1970) Chimpanzees: Self recognition. Sci 167: 86–7
45. CRONOS http://cswww.essex.ac.uk/staff/owen/machine/cronos.html. Accessed 28 Aug 2008
46. Arbib MA (1984) Forward. In: Braitenberg V (ed) Vehicles: Experiments in Systhetic Psychology. MIT Press, NY
47. Braitenberg V (1965) Taxis, kinesis, and decussation. Prog Brain Res 17:210–222
48. Iida F, Pfeifer R (2004) Self-Stabilization and Behavioural Diversity of Embodied Adaptive Locomotion. In: Iida F, Pfeifer R, Steels L, Kuniyoshi Y (eds) Embodied artificial intelligence. Springer, Berlin, pp 119–129
49. Pfeifer R, Scheier C (1999) Understanding Intelligence. MIT Press, MA
50. Mayo MJ (2003) Symbol Grounding and its Implications for Artificial Intelligence. In: Proc. 26th Australas Comput Sci Conf, vol 16. Adelaide, pp 55–60
51. Rubinstein JS, Meyer DE, Evans JE (2001) Executive control of cognitive processes in task switching. J Exp Psychol Hum Percept Perform 27(4):763–797
52. Phillips JL, Noelle DC (2005) A biologically inspired working memory framework for robots. In: IEEE International Workshop on Robots and Human Interactive Communication, Nashville, TN, pp 599–604
53. Gazzaniga MS (ed) (2004) The Cognitive Neurosciences III. MIT Press, Boston
54. Baars BJ (1997) In the Theater of Consciousness: The Workspace of the Mind. Oxford University Press, Oxford
55. Browne WN, Tingley C (2006) Developing an Emotion-Based Architecture for Autonomous Agents. In: Third International Conference on Autonomous Robots and Agents (ICARA 2006), Palmerston North, pp 225–230
56. Asimov I (1950) I, Robot. Fawcett Crest, NY
57. Gee FC, Browne WN, Kawamura K (2005) Uncanny Valley revisited. In: Proc IEEE Robot and Human Interactive Communication, Nashville, pp 151–157

58. Potter SM (2001) Distributed Processing in Cultured Neuronal Networks. Prog Brain Res 130:1–14
59. CIS http://eecs.vanderbilt.edu/CIS/CRL. Accessed 28 Aug 2008
60. Pack T, Wilkes DM, Kawamura K (1997) A software architecture for integrated service robot development. In: IEEE Trans on Systems, Man and Cybernetics October, pp 3774–3779
61. IMA http://eecs.vuse.vanderbilt.edu/cis/concepts/ima.shtml. Accessed 28 Aug 2008
62. Kawamura K, Dodd W, Ratanaswasd P, Gutierrez R (2005) Development of a robot with a sense of self. In: 6th IEEE Int Symposium on Computational Intelligence in Robotics and Automation, Espoo
63. Lee DN, Thompson AI (1982) Vision in Action: The Control of Locomotion. In: Ingle D, Gooddale MA, Mansfield RJW (eds) Analysis of Visual Behavior. MIT Press, Cambridge, pp 411–433
64. Brooks RA (1986) A robust layered control system for a mobile robot. IEEE J Robot Autom RA-2(1):14–23
65. Clancey WJ (1997) Situated Cognition. Cambridge University Press, NY
66. Clark A (1997) Being There: Putting Brains, Body, and World Together Again. MIT Press, Cambridge
67. Hesslow G (2002) Conscious thought as simulation of behavior and perception. Trends Cogn Sci 6(6):242–247
68. Shanahan MP (2006) A cognitive architecture that combines internal simulation with a global workspace. Conscious Cogn 15:433–449

Books and Reviews

Bekey G (2005) Autonomous Robots: From Biological Inspiration to Implementation and Control. MIT Press, MA
Brooks RA (1999) Cambrian Intelligence: The Early History of the New AI. MIT Press, MA
Dennett DC (1991) Consciousness Explained. Little, MA
Gabriel M, Moore J (eds) (1990) Learning and Computational Neuroscience: Foundations of Adaptive Networks. MIT Press, MA
Haikonen PO (2003) The Cognitive Approach to Conscious Machines. Imprint Academic, UK
Hauser MD (2000) Wild Minds. Henry Holt and Company, NY
Meystel AM, Albus JS (2002) Intelligent Systems: Architecture, Design, and Control. Wiley, NY
Nolfi S, Floreano D (2004) Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-organizing Machines. Bradford Book, MIT Press, MA
Picard RW (1997) Affective Computing. MIT Press, MA
Siegwart R, Nourbakhsh IR (2004) Introduction to Autonomous Mobile Robots. Bradford Book, MIT Press, MA

Collective Transport and Depinning

Lei-Han Tang
Department of Physics, Hong Kong Baptist University, Kowloon Tong, Hong Kong SAR, China

Article Outline

Glossary
Definition of the Subject
Introduction
Elastic Manifolds and Impurity Pinning
Driven Depinning, Critical Properties, and Scaling
Analytical Treatments of the Depinning Transition and Universality
Self-Organized Criticality
Interface Depinning in Anisotropic Medium
Future Directions
Bibliography

Glossary

Elastic manifold An elastic manifold is a spatially extended object whose energy is given by a quadratic function of the gradients of its transverse displacements. In statistical physics, it is used as a coarse-grained description of the low-energy modes in an ordered structure.

Pinning Pinning is a common phenomenon in condensed systems where the relevant degrees of freedom are trapped to an energy minimum and hence respond dynamically only when the external driving exceeds a certain threshold. Pinning is often caused by defects or impurities in the system but it may also be due to intrinsic properties such as the existence of a periodic lattice that breaks the translational symmetry of space.

Scaling Scaling describes power-law relationships among two or more physical quantities. Physical systems at continuous phase transitions are often found to exhibit scale invariance, i. e., structural and dynamic properties on different length and time scales can be mapped onto each other through suitable scale transformations. Such properties are characterized by a set of scaling exponents. The renormalization group theory, through its flow equations under a scale transformation, provides a systematic method to compute these exponents.

Universality Universality is a key concept in the classification of systems which exhibit scale invariance. Models in the same universality class have identical scaling properties and are described by the same set of scaling exponents. Therefore the identification of possible universality classes is one of the key issues in the study of, e. g., critical phenomenon at continuous phase transitions. The origin of universality is best illustrated by the fixed-point structure of the renormalization group flow equations. Studies have shown that dimensionality, symmetry, and conservation laws are the key factors that determine a particular universality class. There is, however, a great deal of theoretical interest to identify new principles that determine universality classes, or exceptions to the above rule, particularly in driven nonequilibrium systems.

Definition of the Subject

Collective transport takes place in systems which exhibit highly correlated response to external driving. This is in contrast to, e. g., electrical conduction in a normal metal, where free electrons drift independently under an applied electric field, leading to Ohm's law. Collective motion of microscopic degrees of freedom, on the other hand, often yields a nonlinear response or even threshold behavior, where steady-state transport sets in only when the driving exceeds a certain critical strength. In the subcritical regime, the collective modes are pinned by defects or impurities in the system. An increasing degree of correlation and corporation takes place as the threshold is approached. The subject plays an important role in the study of a broad class of solid state phenomena, including charge and spin density wave transport, hysteresis in dirty magnets, and nonlinear current-voltage characteristics in type-II superconductors. It has also found interesting applications outside physics, such as in crack propagation and earthquake modeling. Important theoretical developments that took place in the early 1990s culminated in the formulation of a functional renormalization group theory for the nonequilibrium depinning transition. The analytical framework enables systematic computation of the critical properties and, perhaps more importantly, elucidation of universality through its fixed point structures. Subsequent work by many research groups have established a close link between driven depinning and the sandpile models of self-organized criticality (SOC). This connection has been fruitfully explored to gain a better understanding of the long-range spatial correlations and intermittent temporal activities in the two classes of problems.

Introduction

The motion of a water drop down a glass window under gravity illustrates many salient features of collective transport and depinning, although the phenomenon itself is surprisingly rich in physics and chemistry when examined in microscopic detail [18,51,73]. The size of the drop is governed by the so called capillary length (typically of the order of a few millimeters) from elementary physics: On this scale the surface tension that makes a water drop spherical competes with the gravity which acts to deform the droplet. The actual descent of the droplet, which usually follows a zigzag path with an ever changing speed, is a result of many factors, not least the random force set by the local wettability of the glass surface. With a microscope, one may observe the hysteric advancement of the contact line separating wet and dry regions on the substrate. The interplay between the surface tension, the pinning force on the contact line, and the driving force provided by gravity gives rise to a complex and intermittent dynamical behavior encompassing a wide range of length and time scales.

Similar types of collective and intermittent transport exist in solids. In certain low-dimensional materials such as $NbSe_3$, periodic modulation of the electronic density, known as the charge-density-wave (CDW) [33], develops spontaneously at sufficiently low temperatures through the Peierls instability. An applied electric field in the direction of the charge modulation exerts a body force on the CDW much the same way as gravity does on the water droplet. However, in the presence of impurity atoms or crystal defects, the CDW does not move unless the electric field exceeds a certain threshold value, as seen in experiments [33,85]. The vortex lattice in type-II superconductors is another example of modulated electronic structures within a solid [44]. Pinning of the vortex lattice by intrinsic or artificial defects is essential for achieving a high critical supercurrent in these materials.

Yet another class of collective pinning phenomena in solids involves the dynamics of topological defects such as crystal dislocations [31] or magnetic domain walls [10]. These objects have internal dimensions lower than that of the embedding medium, and hence can explore inhomogeneities in the surrounding environment. The driven motion of these objects has a great effect on the physical properties of their host system, e. g. plastic deformation of a solid due to glide and climb motion of dislocations, and hysteresis effect [78,80,88] related to the pinning of domain walls separating regions of opposite magnetization.

The development of quantitative theories for impurity pinning and the driven depinning transition began in the 1970s after fundamental breakthroughs in many-body physics and equilibrium critical phenomena [29,44,45,50]. Research in this area has traditionally followed two separate approaches: The microscopic approach that attempts to explain the observed behavior starting from the fundamental laws of physics, and the phenomenological approach that focuses on the large-scale properties using the simplest models possible. The second approach, which is popular in the field of statistical physics, has the advantage of mathematical simplicity and clarity. It facilitates identification of the underlying symmetries of the original problem, and the establishment of universality classes through which model systems can be classified and compared. However, real physical systems often contain complications that prevent a direct comparison between model predictions and experimental observations (see, e. g. [85]). This is where the microscopic approach, popular among

condensed matter theorists, comes to aid. In the best studied cases, a microscopic theory allows one to derive or estimate system-specific parameters and properties, and to suggest correct phenomenological models and possible improvements when discrepancies are found.

The present article focuses on the recent theoretical developments in the statistical physics of this class of problems. This is partly motivated by the significant progress that has been achieved in a quantitative characterization of the depinning transition over the past two decades. Similar to the critical phenomena in thermal equilibrium systems, static and dynamic fluctuations exhibit scaling properties with exponents that fall into well-defined universality classes determined by the symmetry and dimensionality of the problem. These concepts allow one to establish precise relationships among models proposed under different physical contexts. The successful application of the renormalization group methods to the depinning transition has led to a deeper analytical understanding of the observed critical phenomena. Our choice is also motivated by the close connection between depinning and the subject of self-organized criticality [5], which enjoys broad interest in the complexity community.

We shall start with a brief review of systems, mostly from solid state physics, where collective modes and depinning play an important role in the interpretation of the observed transport phenomenon. A generic mathematical formulation of this class of problems, known as the elastic manifold in a disordered potential, is introduced, along with a discussion of the complex energy landscape underlying its equilibrium and dynamic properties. This is followed by a description of the critical properties at the driven depinning transition and numerical results. The basic analytic structure of the threshold solution can be seen in a mean-field theory. We then summarize the main results of functional renormalization group calculations which provide a systematic and quantitative characterization of the depinning transition. The relationship between driven depinning and self-organized criticality is explained. The influence of medium anisotropy on the depinning transition is briefly discussed. Finally, we mention a few open problems in the field for future work.

Elastic Manifolds and Impurity Pinning

Elasticity, Order, and Symmetry Breaking

Although long-range spatial correlations that underlie collective transport can be generated dynamically under certain conditions, we shall focus on systems for which such correlations are already present before external driving is applied. In particular, we shall assume that the system is in a low temperature ordered state and responds elastically to external perturbations. This assumption allows us to concentrate on the large scale properties of the system in the depinning process, while leaving the microscopic, system-specific behavior to a separate discussion.

Elasticity is a familiar concept in macroscopic physics. An object is said to be *elastic* if it deforms under an applied force in a continuous and reversible manner. Microscopically, elasticity is intimately related to the existence of an ordered state that breaks a continuous symmetry [4]. This point can be appreciated with the example of a crystalline solid. Atoms in the solid form a periodic structure in space with energetically preferred unit cells and lattice constants. This state is said to break the translational symmetry of the particle system as the density of atoms is no longer uniform in space. A uniform translation of the structure does not lead to a change in the system energy. On the other hand, relative displacement of atoms distorts the unit cells and leads to an increase in energy which grows quadratically with the displacement. Thus, the system has acquired a form of rigidity by breaking a continuous symmetry to gain order. Elasticity is a manifestation of this rigidity when the ordered state is perturbed by external forces or internal impurities and defects.

Both charge-density waves and vortex lattices are periodic structures in space, and hence naturally their behavior is similar to a crystalline solid. There are, however, differences in the number of components needed to describe a generic deformation in each case. A CDW with a single modulation wave vector $\mathbf{Q}$ corresponds to an electron density $\rho(\mathbf{r}) = \rho_0 + \rho_1 \cos(\mathbf{Q}\cdot\mathbf{r} + \varphi)$, where ρ_0 is the average density and ρ_1 the modulation amplitude due to charge ordering. A uniform phase shift $\varphi \to \varphi + c$ moves the CDW uniformly against the underlying lattice. Weak deformation of the CDW is described by a spatially varying phase $\varphi(\mathbf{r})$. The elastic energy of a CDW depends quadratically on the phase gradient $\nabla\varphi$, with different elastic constants along and perpendicular to the modulation vector $\mathbf{Q}$ [29]. In the case of the vortex lattice, a two-component vector field $\mathbf{u}(\mathbf{r})$ perpendicular to the vortex lines is needed to describe a general distortion of the ideal structure. The construction of the elastic energy parallels that of a crystalline solid. In an isotropic medium, three elastic constants are needed to describe the energetics of a vortex line array [44,45].

Topological defects in an ordered structure also behave elastically when deformed from their ideal positions under a great variety of circumstances, though this form of elasticity has a different microscopic origin. A dislocation in a solid, for example, has a core where the atomic arrangement differs from elsewhere in the crystal. This gives rise

to an excess amount of energy proportional to the length of the line. The interface between two bulk phases has an excess free energy (commonly known as the surface tension) proportional to the surface area for the same reason. This applies also to domain walls in a magnet.

Elastic Manifolds in a Disordered Potential

The energy of a topological defect is affected by the presence of impurities or point defects in the system. For example, an impurity atom sitting in the core of a dislocation has a different atomic environment and hence a different energy than when it resides in a normal region. This effect gives rise to a short-range and often attractive force between the impurity and the dislocation core. The tendency for the dislocation to distort itself in order to adapt to the impurity configuration is countered by the elastic energy cost. In general, the two competing forces lead to a large number of metastable conformations of the dislocation core, separated by energy barriers. The effect of impurities on interfaces and magnetic domain walls can be considered in a similar way.

A unified statistical mechanical description of impurity pinning can be formulated in terms of a D-dimensional elastic manifold embedded in d spatial dimensions [28,34,49]. Let $\mathbf{r} \in R^D$ be the internal coordinates of the manifold and $\mathbf{u}(\mathbf{r}) \in R^{d-D}$ be the transverse displacement of the manifold at point $\mathbf{r}$ with respect to a flat configuration. When the deformation is small, the excess volume of the manifold is a quadratic function of the gradient of $\mathbf{u}(\mathbf{r})$. The following energy function includes effects of both elasticity and impurities,

$$E(\{\mathbf{u}\}) = \int \mathrm{d}^D r \left[\frac{\gamma}{2} |\nabla \mathbf{u}|^2 + V_R(\mathbf{r}, \mathbf{u}) \right] . \tag{1}$$

Here γ is the stiffness constant of the manifold, and $V_R(\mathbf{r}, \mathbf{u})$ is a random potential arising from the interaction between the manifold and the impurities in the medium. In most theoretical treatments, the set of random variables $\{V_R(\mathbf{r}, \mathbf{u})\}$, which take particular values in a given sample, is assumed to be Gaussian distributed with the mean $\langle V_R(\mathbf{r}, \mathbf{u}) \rangle = 0$ and the second moment $\langle V_R(\mathbf{r}_1, \mathbf{u}_1) V_R(\mathbf{r}_2, \mathbf{u}_2) \rangle = R(\mathbf{u}_1 - \mathbf{u}_2) \delta(\mathbf{r}_1 - \mathbf{r}_2)$. Here $\delta(\mathbf{r})$ is a short-ranged function which vanishes beyond a coarse-graining length $a_\parallel$.

The form of the disorder correlator $R(\mathbf{u})$ in the transverse directions is dictated by the symmetries of the original physical problem [28]. For contact interactions and randomly distributed impurities in the embedding space, $R(\mathbf{u})$ is short-ranged with a characteristic decay length $a_\perp$. A different situation is encountered when the manifold represents an interface that separates two bulk phases, each affected differently by the impurities which serve as a "random field" (i. e., the impurity atom has different chemical potential in the two bulk phases). The potential $V_R(\mathbf{r}, u)$ represents the cumulative effect of impurities swept by the interface when it is displaced to $u(\mathbf{r})$ from a reference configuration. Consequently $R(u) \sim |u|$ for large u.

A CDW in the presence of impurities can also be described by Eq. (1), where $u(\mathbf{r}) = \varphi(\mathbf{r})$ is a scalar field and $\mathrm{d} = D$. Since the impurity interacts with the local charge density which is a periodic function of the local phase, the potential $V_R(\mathbf{r}, \varphi)$ is also periodic in φ. Consequently, $R(\varphi)$ is periodic in φ as well.

Application of Eq. (1) to the vortex lattice requires special care. On scales smaller than the spacing between vortex lines, each vortex line behaves as an independent object. However, this description fails on large scales, where the periodicity of the lattice changes the nature of the problem [30,68].

Rugged Energy Landscape, Critical Dimension, and the Pinning Length

The two terms in Eq. (1) represent competing effects on the manifold: Elasticity favors a flat manifold with a constant $\mathbf{u}$, while the attractive force from impurities gives rise to a spatially varying $\mathbf{u}(\mathbf{r})$. The configuration $\mathbf{u}(\mathbf{r})$ which minimizes Eq. (1) satisfies the following force-equilibrium condition:

$$\gamma \nabla^2 \mathbf{u} - \nabla_{\mathbf{u}} V_R(\mathbf{r}, \mathbf{u}) = 0 . \tag{2}$$

Complications arise when Eq. (2) has many solutions. Each solution corresponds to a local minimum of the energy function (1), separated by energy barriers from other local energy minima. The resulting energy surface (or landscape) is known as *rugged*.

Proper characterization of the energy landscape defined by Eq. (1) has been a long-standing problem in the statistical mechanics of disordered systems. For $D < 4$ and weak disorder, the ruggedness appears when the system size is greater than a characteristic length L_c along the manifold, known as the *pinning length*. This important observation was due to Larkin [44] for flux lines, Fukuyama, Lee and Rice [29,50] for CDWs, and Imry and Ma [37] for magnetic domain walls. On scales $L < L_c$, elasticity limits the transverse displacement $\mathbf{u}$ (known as roughness) to be within the respective correlation length $a_\perp$ of the impurity potential $V_R(\mathbf{r}, \mathbf{u})$, so that the manifold lies within a single minimum of the energy surface. On the scale L_c, the typical

strength of the random force $\nabla_{\mathbf{u}} V_R$, averaged over a volume L_c^D, is estimated to be $\left|\overline{\nabla_{\mathbf{u}} V_R}\right| \simeq \frac{R^{1/2}(0)}{a_\perp} L_c^{-D/2}$. Balancing it with the elastic force $\gamma a_\perp / L_c^2$, one obtains,

$$L_c = \left(\frac{\gamma^2 a_\perp^4}{R(0)}\right)^{1/(4-D)} . \tag{3}$$

For $L > L_c$, Eq. (2) has an exponentially increasing number of solutions.

For $D > 4$, the elastic term in (2) dominates over the random force term on large length scales. Consequently, the rugged energy landscape is a small scale phenomenon which occurs when L_c is larger than the correlation length $a_\parallel$ of the disorder potential $V_R(\mathbf{r}, \mathbf{u})$ parallel to the manifold. From Eq. (3) we see that this condition requires the disorder strength to be sufficiently strong. Weak disorder is not able to produce pinning for manifolds with internal dimension greater than 4.

The qualitative analysis above shows that the energetics of the manifold problem are qualitatively different below and above four dimensions. The existence of a *critical dimension* $D_c = 4$ suggests systematic renormalization group approach to the problem. However, earlier attempts based on a perturbative treatment of the disorder potential failed to produce a renormalizable theory [1,22,32]. This was first achieved successfully by Daniel Fisher [28] in 1986 in the equilibrium case. Extension of the scheme to the driven depinning transition is discussed in Sect. "Analytical Treatments of the Depinning Transition and Universality".

Driven Depinning, Critical Properties, and Scaling

The Driven Depinning Transition

An applied force $\mathbf{F}$ coupled linearly to the displacement field $\mathbf{u}(\mathbf{r})$ tilts the equilibrium energy landscape as defined by Eq. (1) towards a particular direction. For small $\mathbf{F}$, the manifold makes small adjustments locally to reach a new stationary state where force equilibrium is re-established, and this happens for the majority of solutions to Eq. (2). The number of such solutions, however, continues to decrease as $\mathbf{F}$ increases. When the magnitude of $\mathbf{F}$ exceeds a certain critical value, all stationary states disappear, and the manifold enters a running state. The transition from stationary to running states with increasing $\mathbf{F}$ is known as the *driven depinning transition*.

Figure 1 illustrates the dependence of the steady-state velocity v of the manifold against $F = |\mathbf{F}|$. In the absence of thermal fluctuations, there is a well-defined threshold F_c that separates the pinned from moving regimes. At finite temperatures, the transition from a pinned to a moving manifold is smeared out by thermally activated creep motion. At low temperatures, the rounding effect is weak except in a very small region around F_c.

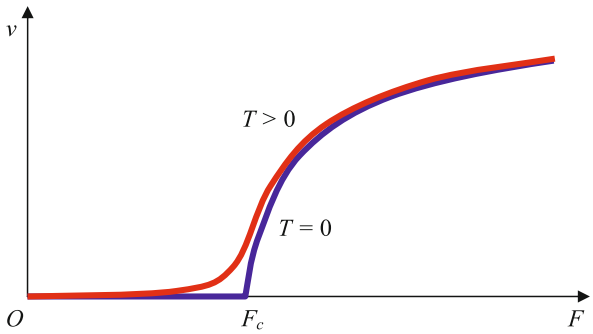

Collective Transport and Depinning, Figure 1
Schematic plot showing the manifold velocity *v* against the driving force *F* at zero (*blue*) and finite (*red*) temperatures. Below the threshold F_c, the manifold is pinned by impurities in the medium, but thermal activation may generate a small but finite velocity

The discussion in Subsect. "Rugged Energy Landscape, Critical Dimension, and the Pinning Length" on the pinning length can be used to estimate the critical force needed to depin the manifold. For $D < 4$, the maximum pinning effect is seen on the scale L_c, where the typical strength of the first two terms in (5) is given by [10,26,50]

$$F_c \simeq \gamma a_\perp / L_c^2 = \frac{R^{2/(4-D)}(0)}{\gamma^{D/(4-D)} a_\perp^{(4+D)/(4-D)}} . \tag{4}$$

This is also the force needed to depin the manifold. For $D > 4$, pinning is possible only if $L_c > a_\parallel$ or $R^{1/2}(0) a_\perp^{-1} a_\parallel^{-D/4} > \gamma a_\perp / a_\parallel^2$, i. e., the strength of the pinning force is stronger than that of the elastic force on the minimal scale $a_\parallel$.

Continuum Model of the Manifold Dynamics

A dynamical model for the manifold can be constructed by assuming the motion to be completely overdamped. This is quite reasonable for the applications mentioned in Sect. "Introduction", where the displacement $\mathbf{u}(\mathbf{r}, t)$ represents a course-grained variable which changes on a time scale much longer than the relaxation time of underlying microscopic processes. With this assumption, the equation of motion for the manifold takes the form,

$$\mu^{-1} \frac{\partial \mathbf{u}}{\partial t} = \gamma \nabla^2 \mathbf{u} + \boldsymbol{\eta}(\mathbf{r}, \mathbf{u}) + \mathbf{F} .$$

Here μ is known as the mobility of the manifold, and $\boldsymbol{\eta}(\mathbf{r}, \mathbf{u}) = -\nabla_{\mathbf{u}} V_R(\mathbf{r}, \mathbf{u})$ is the random pinning force.

An interface is described by a height function $u(\mathbf{r}, t)$, so that the dynamical equation, which defines the so-called *linear interface model* (LIM), takes a scalar form [10,26],

$$\mu^{-1}\frac{\partial u}{\partial t} = \gamma\nabla^2 u + \eta(\mathbf{r}, u) + F\,. \tag{5}$$

The same equation applies to a CDW, but with $\eta(\mathbf{r}, u + 2\pi) = \eta(\mathbf{r}, u)$ periodic in the phase variable [27]. Using the example of a single vortex line in three dimensions, Ertas and Kardar [23] have shown that the extra transverse dimensions do not change the main features of the depinning process. Hence Eq. (5) can be considered as the generic description of the driven manifold problem, where $u(\mathbf{r}, t)$ stands for the component of the transverse displacement in the driven direction.

In Layman's terms, Eq. (5) may be viewed as describing the advance of a military front where the attacking side has more ammunition and manpower F but the defending side is able to exploit the hilly ground positions $\eta(\mathbf{r}, u)$ to pose an effective resistance. In addition, stretching the front line into a convoluted form results in a decrease in attacking power and is hence discouraged!

For the analysis of Eq. (5), it is convenient to describe pinning effects in terms of the random force $\eta(\mathbf{r}, u)$ instead of the random potential $V_R(\mathbf{r}, u)$. Without loss of generality, one may assume the mean value of $\eta(\mathbf{r}, u)$ to be zero. For Gaussian distributed random forces, it is suffice to specify the statistics of $\eta(\mathbf{r}, u)$ with the correlator,

$$\langle \eta(\mathbf{r}_1, u_1)\,\eta(\mathbf{r}_2, u_2)\rangle = \Delta(u_1 - u_2)\,\delta(\mathbf{r}_1 - \mathbf{r}_2)\,. \tag{6}$$

On the "bare" scale, $\Delta(u) = -\partial^2 R/\partial u^2$ but as shown in [66] and [48], this relation breaks down in the driven case upon coarse-graining of the original degrees of freedom.

Critical Properties and Scaling Laws

Let us first consider the interface depinning problem as described by Eq. (5). This model has been studied extensively both analytically [14,66,69] and numerically [53,76] in recent years, and the main characteristics of the solution have been well-understood. These findings are summarized below.

The form of Eq. (5) suggests a "non-crossing condition" as first noted by Middleton [58]. Consider two interface configurations $u_1(\mathbf{r}, t_0)$ and $u_2(\mathbf{r}, t_0)$ at some initial time t_0. If $u_1(\mathbf{r}, t_0) < u_2(\mathbf{r}, t_0)$ for all $\mathbf{r}$ on the interface, one can easily show that $u_1(\mathbf{r}, t) < u_2(\mathbf{r}, t)$ at any later time $t > t_0$, i. e., the two solutions never touch if initially one is completely behind the other. This property in particular implies that the interface velocity $v = \overline{\mathrm{d}u/\mathrm{d}t}$ (averaged over all sites $\mathbf{r}$) in the steady state is a unique and continuous function of F, ruling out first order phase transition in this class of models.

On the moving side but close to the depinning threshold, advancement of the interface can be described with the help of Fig. 2. The thick line illustrates the interface position at a given time t_0. Pinning yields a roughness which grows as a power-law of the distance between two points on the interface, i. e.,

$$\left\langle \left[u(\mathbf{r}, t_0) - u(\mathbf{r}', t_0)\right]^2 \right\rangle \sim |\mathbf{r} - \mathbf{r}'|^{2\zeta}\,, \tag{7}$$

where ζ is known as the *roughness exponent*. This behavior holds on a range of length scales from the pinning length L_c to a correlation length $\xi_\parallel$ along the interface. The transverse displacement of the interface on scale $\xi_\parallel$ is given by $\xi_\perp \sim \xi_\parallel^\zeta$.

Motion of the interface within a correlation time τ also obeys scaling. On time scales shorter than τ, the interface advances through a sequence of rapid, localized movements (known as avalanches) of varying size up to the scale $\xi_\parallel$. The average time for a given site to move by a distance Δu grows as a power law: $\Delta t \sim \Delta u^{1/\beta}$. During this time, activities within a distance $l \sim \Delta t^{1/z}$ along the interface are correlated. The exponent $z = \zeta/\beta$ is known as the *dynamical exponent*. This scaling terminates when l reaches the correlation length $\xi_\parallel$, or $\Delta t = \tau \sim \xi_\parallel^z$, where the interface as a whole advances to a new disorder environment with a different distribution of pinning sites and pinning forces.

As F decreases towards F_c, the size of each correlated domain grows to infinity in a power-law fashion as well, e. g., $\xi_\parallel \sim |F - F_c|^{-\nu}$, where ν is known as the *correlation length exponent*. The three exponents ζ, z and ν together characterize the critical properties of the interface at the depinning transition. Through dimensional arguments, one may determine the critical behavior of quantities other than those discussed above. For example, the interface velocity near the transition can be estimated from $v \simeq \xi_\perp/\tau \sim \xi_\parallel^\zeta/\xi_\parallel^z \sim |F - F_c|^{\nu(z-\zeta)}$, hence the corresponding *velocity exponent* is given by,

$$\theta = \nu(z - \zeta) \tag{8}$$

The region between two interface configurations separated by time τ, as shown in Fig. 2, can be described as a set of "bubbles", each representing a correlated volume of base area $\xi_\parallel^D$ and height $\xi_\perp$. The average strength of the pinning forces within each bubble has a variation of the order $\xi_\parallel^{-D/2}\xi_\perp^{-1/2}$ from the system-wide average. Variation of this quantity among bubbles should be smaller than the excess driving force $F - F_c$ so that depinning can take place

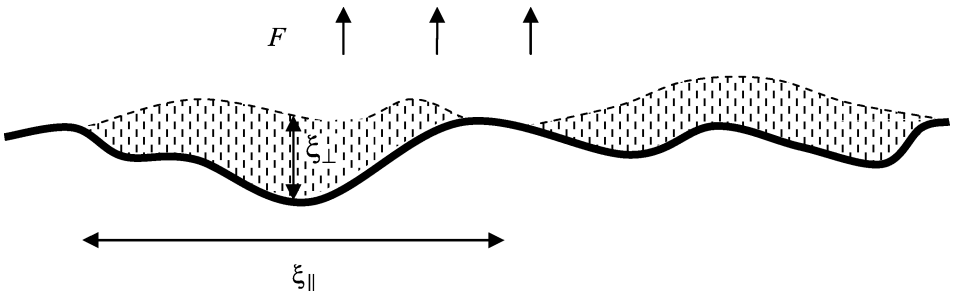

Collective Transport and Depinning, Figure 2
An interface above but close to the depinning threshold. Depinning events within a correlation length $\xi_{\parallel}$ and correlation time τ are highly correlated and obey scaling at the transition. Shaded area represents the volume swept by the interface over the time interval τ

Collective Transport and Depinning, Table 1
Scaling exponents at the depinning transition. Numbers in parentheses indicate uncertainties in the last digit

	Interface[a]			CDW			MF $D \geq 4$
	$D = 1$	$D = 2$	$D = 3$	$D = 1$	$D = 2$	$D = 3$	
ζ (roughness)	1.25(1)	0.75(2)	0.35(1)	0			0
z (dynamic)	1.42(4)	1.56(6)	1.75(15)	1	1.32(4)[b]	1.65(6)[b]	2
ν, ν_f (correlation length)	1.33(2)	0.80(1)	0.606(4)	0.4(1)[c] 2.01(2)	0.5(1)[c] 0.98(3)[b]	0.5(1)[c] 0.68(4)[b]	$1/2$
θ (velocity)	0.25(3)	0.64(2)	0.84(2)	0.45(5)[c]	0.64(3)[b,c]	0.81(3)[b,c]	1

[a]Leschhorn H [53]; [b]Middleton AA, Fisher DS [61]; Narayan O, Middleton AA [67]; [c]Mayer CR, Sethna JP [63].

uniformly across the system. This condition is fulfilled if the correlation length exponent satisfies the following inequality [35,66,69],

$$\nu \geq \frac{2}{D + \zeta} \tag{9}$$

The above description of interface depinning applies also to the CDW with one important caveat. Due to the non-crossing condition, steady-state motion of the CDW at $F > F_c$ is periodic in time, i. e., $\varphi(\mathbf{r}, t + \tau) = \varphi(\mathbf{r}, t) + 2\pi$ for all $\mathbf{r}$, with τ being the period of the attractor [58]. The dynamic phase advance at different sites has thus bounded variations described by $\zeta = 0$. Above but close to F_c, the activity at a given site, as measured by the phase velocity $\dot{\varphi}(\mathbf{r}, t)$, is typically concentrated in time windows much shorter than T, but acquires long-ranged spatial correlations up to a correlation length $\xi_{\parallel} \sim (F - F_c)^{-\nu}$. Both analytical [65] and numerical calculations [63] indicate that $\nu = 1/2$ in all dimensions, therefore violating Eq. (9). Although the origin of this behavior has not been settled completely [67], a plausible explanation is that fluctuations of the pinning force in a given region of the system are compensated by a static phase distortion $\varphi_0(\mathbf{r}) = \varphi(\mathbf{r}, t = 0)$, so that the system behaves much more homogeneous than the naïve estimate used to obtain Eq. (9). The sample-to-sample fluctuations of F_c, on the other hand, have been shown recently [7,25] to obey a Gaussian distribution with a width proportional to $L^{-D/2}$as predicted, where L stands for the linear system size.

Numerical Results for the Critical Exponents

The depinning transition of the elastic manifolds and the CDW has been studied extensively using various lattice models. Middleton and Fisher [60,61], Myers and Sethna [63], and Narayan and Middleton [67] simulated a discretized version of Eq. (5) for CDW depinning in one to three dimensions. The random force is given by $\eta(\mathbf{r}, \varphi) = V \sin(\varphi - \beta(\mathbf{r}))$, where V is the strength of the pinning force, and $\beta(\mathbf{r})$ is the preferred phase at site $\mathbf{r}$, chosen randomly from site to site. Values of critical exponents as determined in their numerical work are given in Table 1, which show good agreement with the analytic results in Sect. "Analytical Treatments of the Depinning Transition and Universality". The exponent ν_f (second row) is determined from quantities related to the avalanche propagation below the depinning threshold.

Leschorn [53] carried out large-scale simulations of a discretized version of the LIM. Exponents he obtained are also given in Table 1. These values are in good agreement with more recent studies [76]. Note that the rough-

ness exponent ζ for a one-dimensional interface at the depinning threshold is greater than one. This gives rise to a subtlety in using Eq. (7) to measure ζ as discussed in [56].

Analytical Treatments of the Depinning Transition and Universality

The depinning transition differs from the usual critical phenomena in thermal equilibrium systems in terms of the vast separation of fast and slow time scales. Therefore the successful development of a renormalization group theory in the early 1990s to effectively capture this unique feature was an important milestone in the analytical studies of nonequilibrium phase transitions. The work acquired broader significance due to the later discovered correspondence between depinning and sand pile models of self-organized criticality (SOC), which we discuss in Sect. "Self-Organized Criticality".

Mean-Field Theory

Fisher [27] and others [43,52,55,65] considered a mean-field approximation to Eq. (5) which can be treated analytically. It is instructive to examine the calculation in some detail here as the solution reveals several important features of the depinning transition which extend to lower dimensions, while the mathematical manipulations can be kept at an elementary level.

In sufficiently high dimensions, one may approximate the Laplacian in Eq. (5) by a spring force, whose equilibrium point is set by the mean interface position $\bar{u}(t)$. The resulting dynamical equation for a given site on the interface takes the form,

$$\mathrm{d}u/\mathrm{d}t = \gamma(\bar{u} - u) + \eta(u) + F\,. \tag{10}$$

For simplicity we have dropped μ in (5) through a redefinition of t. In the steady-state, $\bar{u} = vt$, where v is the interface velocity to be determined self-consistently from the solution of Eq. (10). Since the disorder $\eta(u)$ is uncorrelated along the interface, the system-wide average $\bar{u}$ can be replaced by an average over the distribution of $\eta(u)$.

Following the discussion in [55], we adopt a moving frame and define $x = u - vt - (F - v)/\gamma$ to be the displacement away from the equilibrium position when the random force $\eta(u)$ is absent. In terms of x, Eq. (10) reads (after a rigid translation of $\eta(u)$),

$$\mathrm{d}x/\mathrm{d}t = -\gamma x + \eta(vt + x)\,. \tag{11}$$

The self-consistency condition becomes,

$$\gamma\,\langle x\rangle = -F + v\,. \tag{12}$$

The driving force F now disappears from the equation of motion (11) and can be computed from (12) as a function of v, once a solution to (11) is found.

As illustrated in Fig. 3, Eq. (11) describes the dynamics of an overdamped particle in a moving potential,

$$W(x,t) = \frac{\gamma}{2}x^2 + U(vt + x)\,. \tag{13}$$

The random part $U(vt + x) = -\int_{x_0}^{vt+x}\eta(u)\mathrm{d}u$ travels at a constant velocity v to the left or to the right depending

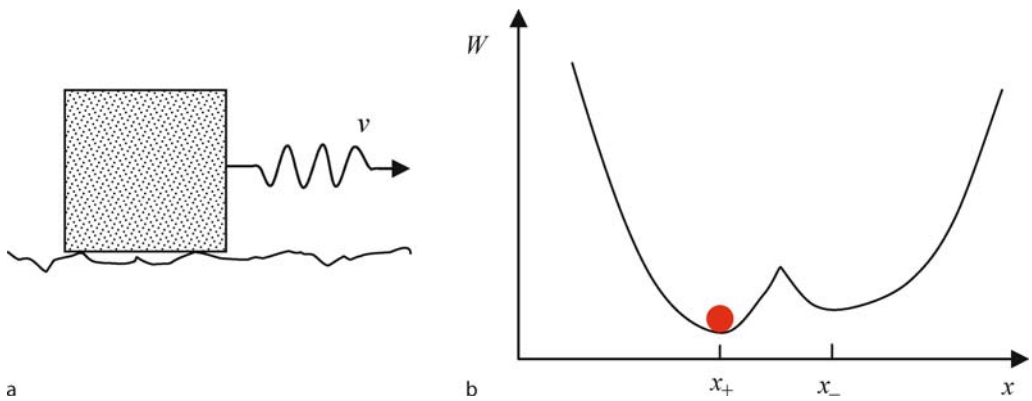

Collective Transport and Depinning, Figure 3
The mean-field model of depinning. a Block (representing the interface) dragged by a spring whose right end moves at a slow but constant velocity *v*. The block is sitting on a rough surface and immersed in a very viscous fluid, so that inertia effects can be neglected. b The effective potential *W* on the block at a particular instant when viewed in a co-moving frame. *W* is a sum of two parts: A quadratic spring potential plus a random potential *U* which travels slowly to the left. In the absence of thermal motion, the block (represented by the red ball in the figure) traces the left most minima x_+ of *W* until it disappears, followed by an "avalanche" into the next available minimum

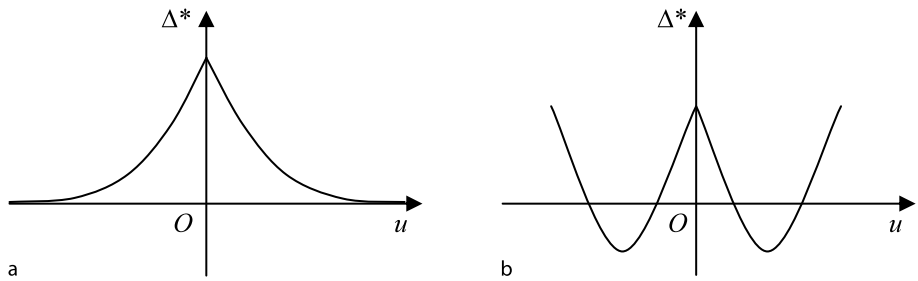

Collective Transport and Depinning, Figure 4
Fixed point solutions to Eq. (14) for a the linear interface models and b the charge-density wave depinning

on the sign of v. In the quasi-static limit $v \to 0$, the particle stays in a local minimum of the potential W for the whole time, and is interrupted occasionally by jumps when the minimum it follows disappears. When the potential travels to the left ($v > 0$), the particle traces the leftmost minimum x_+ of W. The opposite occurs when the potential travels to the right.

With the above picture in mind, we may use Eq. (12) to calculate the thresholds $F_c^+ = -\gamma \lim_{v\to 0+}\langle x\rangle$ and $F_c^- = -\lim_{v\to 0-}\langle x\rangle$ for forward and backward depinning, respectively. For a continuous, slow-varying function $\eta(u)$ with a sufficiently small amplitude, the potential W has a unique minimum all the time, and hence $F_c^+ = F_c^- = 0$, i. e., the system is never pinned. However, for a discontinuous function $\eta(u)$ (or for sufficiently strong disorder), the potential W has, from time to time, more than one minimum, and hence $F_c^+ = -F_c^- > 0$. Note that only upward jumps in η (i. e. a sudden weakening of the pinning force) give rise to upward cusps in the potential which generate double minima in W.

When the potential $W(x,t)$ travels at a small but finite velocity v, the ball shown in Fig. 3b is displaced to the leftmost minimum of W by an amount proportional to v due to viscosity. From Eq. (12) and the value of F_c determined above, one finds $v = a(F - F_c)$, where a is a numerical constant. This behavior is confirmed by exact solution of a "two-state" model for the pinning force [52]. The mean-field velocity exponent is thus given by $\theta_{\mathrm{MF}} = 1$.

Functional Renormalization Group

Equation (5) has been studied analytically since the 1970s. Earlier attempts by Efetov and Larkin [22] and others [26,43] based on a perturbative expansion of the noise term around a flat reference state ran into divergences for $D < D_c = 4$, which can not be cured with the usual renormalization group (RG) procedure. Breakthroughs were achieved in 1992 by Narayan and Fisher [65] for the CDW depinning and by Nattermann et al. [69] for the driven interface. In the field theory jargon, the model has an infinite number of relevant operators that correspond to the coefficients in a power-law expansion of the random force correlator $\Delta(u)$. Upon momentum shell integration on a running scale L, the one-loop corrections to these coefficients can be summarized in terms of a functional RG flow for $\Delta(u)$,

$$\frac{\mathrm{d}\Delta(u)}{\mathrm{d}\ln L} = -cL^{\varepsilon}\frac{\mathrm{d}^2}{\mathrm{d}u^2}\left[\frac{1}{2}\Delta^2(u) - \Delta(u)\Delta(0)\right] . \qquad (14)$$

Here $\varepsilon = 4 - D$ and c is a constant. Dependence on L can be absorbed with the scaling transformation $u \to L^{\zeta}u$, $\Delta \to L^{2\zeta-\varepsilon}\Delta$. The resulting flow equation has an infinite number of fixed points, which are distinguished by the number of nodes of the function Δ. Figure 4 shows two such solutions, the first with no node at $\zeta = \varepsilon/3$, and the second with infinitely many nodes (periodic) at $\zeta = 0$. The periodic fixed point is associated with the CDW depinning transition. The one with no node describes the interface depinning transition. The most prominent feature of the fixed point function Δ^* is the cusp singularity at $u = 0$, which is responsible for the failure of previous RG treatments.

Equation (14) is supplemented with a RG equation for the mobility μ that determines the dynamical response on scale L,

$$\frac{\mathrm{d}\ln\mu}{\mathrm{d}\ln L} = -c\Delta''(0+)L^{\varepsilon} . \qquad (15)$$

The elastic constant γ, on the other hand, is not renormalized. The latter property yields the scaling relation

$$\nu(2-\zeta) = 1 . \qquad (16)$$

The functional RG analysis has been carried out to two-loop order by Le Doussal and collaborators [14,48]. This study is particularly useful as it helps to clear up several

subtleties which can not be easily resolved in the one-loop calculation. For interface depinning, exponents up to the second order in ε are given by,

$$\begin{aligned}
\zeta &= \frac{1}{3}\varepsilon + 0.047771\varepsilon^2 \\
z &= 2 - \frac{2}{9}\varepsilon - 0.0432087\varepsilon^2 \\
\nu &= \frac{1}{2-\zeta} = \frac{1}{2} + \frac{1}{12}\varepsilon + 0.0258316\varepsilon^2 \\
\theta &= \nu(z-\zeta) = 1 - \frac{1}{9}\varepsilon + 0.040123\varepsilon^2 \,.
\end{aligned} \tag{17}$$

Good agreement is seen with the numerical values given in Table 1.

For CDW depinning, the two-loop analysis yields the following exponents,

$$\begin{aligned}
\zeta &= 0 \\
z &= 2 - \frac{1}{3}\varepsilon - \frac{1}{9}\varepsilon^2 \\
\nu &= \frac{1}{2-\zeta} = \frac{1}{2} \\
\theta &= \nu(z-\zeta) = 1 - \frac{1}{6}\varepsilon - \frac{1}{18}\varepsilon^2 \,.
\end{aligned} \tag{18}$$

Interestingly, here the expression for the dynamical exponent z truncated at first order in ε appears to be in better agreement with the simulation results given in Table 1.

Very recently, Rosso, Le Doussal and Wiese [77] proposed and implemented a scheme to measure numerically the effective pinning force and the renormalized correlator Δ on large distances. The idea is similar to the mean-field model discussed above, with the particle replaced by the center of mass of the manifold in a weak confining potential that introduces a running cut-off to spatial correlations. Their numerical results provide direct confirmation of the cusp singularity in Δ, which has become the hallmark of the RG theories for the manifold problem.

Contact Line Depinning

The above analysis has been extended by Ertas and Kardar [24] to the contact line depinning problem. The increase in surface free energy (or surface area) due to a deformation of the contact line from an ideal straight configuration has been worked out previously by Joanny and de Gennes [39]. Result of this analysis, generalized to a "contact manifold" with D internal dimensions, is that the Laplace term in Eq. (5) is replaced by a nonlocal term, $\gamma\nabla^2 u \to \int \mathrm{d}^D\mathbf{r}' K_D(\mathbf{r}-\mathbf{r}')[u(\mathbf{r}',t) - \bar{u}(t)]$. Here $\bar{u}(t)$ denotes the mean displacement. The function $K_D(\mathbf{r}) \sim |\mathbf{r}|^{-D-1}$ describes the decay of the long-ranged elastic coupling along the contact manifold, assuming the modes in the "bulk" have relaxed. This form of elastic coupling changes the critical dimension to $D_c = 2$. In terms of $\varepsilon = 2 - D$, the critical exponents of the depinning transition from a two-loop renormalization group calculation by Le Doussal et al. [48] are given by,

$$\begin{aligned}
\zeta &= \frac{1}{3}\varepsilon + 0.13245\varepsilon^2 \\
z &= 1 - \frac{2}{9}\varepsilon - 0.1132997\varepsilon^2 \\
\nu &= \frac{1}{2-\zeta} = 1 + \frac{1}{3}\varepsilon + 0.24356\varepsilon^2 \\
\theta &= \nu(z-\zeta) = 1 - \frac{2}{9}\varepsilon - 0.1873737\varepsilon^2 \,.
\end{aligned} \tag{19}$$

Interestingly, the same long-range coupling is found to describe tensile crack propagation [75]. Numerical simulations [21] of a discretized long-range model in $D = 1$ ($\varepsilon = 1$) yield $\zeta = 0.385(5)$, $z = 0.770(5)$, $\nu = 1.625(10)$, $\theta = 0.625(5)$, in good agreement with the theory.

Self-Organized Criticality

Close to the depinning threshold, motion of the system is intermittent. The growth activities are highly localized and exhibit strong correlations over time. The natural separation of slow and fast events at the onset of depinning invites comparison to models of self-organized criticality (SOC) and of earthquakes, which were made popular through the seminal work of Bak, Tang and Wissenfeld (BTW) [5,6] on the "sandpile" model. This observation prompted a number of detailed investigations of the microscopic processes leading to the build-up of spatiotemporal correlations as the depinning threshold is approached from below, as summarized in [72]. Subsequent work by several groups [67,71] established a direct link between these two classes of models. This development has also contributed to a better understanding of SOC in automaton models. It opens the door to a systematic exploration of their scaling properties which were hitherto hindered by the lack of a suitable continuum representation.

The mapping between depinning and SOC automaton models is most easily understood by considering $u(\mathbf{r}, t)$ in Eq. (5) as the accumulated activity of a given lattice site $\mathbf{r}$ after time t [71]. Take for example the original sandpile model introduced by BTW on a square lattice. The system is specified by an integer height $z(x, y)$ on each lattice site (x, y). Starting from a random set of values $z_0(x, y)$, "sand" is gradually added through a sequence of

events $z(x, y) \to z(x, y) + 1$, each time only a single site is chosen randomly from all available sites in the system for updating. If the operation leads to an overflow, i. e., $z(x, y) > K$, then toppling takes place according to the rule,

$$\begin{aligned} z(x, y) &\to z(x, y) - 4 \\ z(x \pm 1, y) &\to z(x \pm 1, y) + 1 \\ z(x, y \pm 1) &\to z(x, y \pm 1) + 1 \,. \end{aligned} \tag{20}$$

This process is continued until no site has a height exceeding K. Let $u(x, y; t)$ be the total number of toppling events at site (x, y) from the beginning of the process. Counting the number of sand grains received and given away by the site yields immediately

$$\begin{aligned} z(x, y; t) &= z_0(x, y) - 4u(x, y; t) + u(x + 1, y; t) \\ &\quad + u(x - 1, y; t) + u(x, y + 1; t) \\ &\quad + u(x, y - 1; t) \\ &= z_0(x, y) + \nabla^2 u \,, \end{aligned} \tag{21}$$

where ∇^2 is the Laplace operator on the lattice. The condition $z(x, y; t) > K$ for toppling $u \to u + 1$ is thus equivalent to

$$\nabla^2 u + z_0(x, y) - K > 0 \,. \tag{22}$$

The above growth rule is almost identical to a lattice automaton model introduced by Narayan and Middleton [67] for CDW depinning. Comparing (22) to Eq. (5), we see that here (i) u is restricted to take only integer values plus a random, site dependent shift $\beta(\mathbf{r})$, which can be considered as the preferred phase set by the local disorder; (ii) The dynamical rules obey a global symmetry $u \to u + 1$, as in CDW depinning. The usual wisdom in critical phenomena suggests that the two models are in the same universality class due to property (ii), while property (i) is a microscopic detail which does not affect the scaling properties at the depinning transition. Indeed, various scaling exponents determined through analytical and numerical studies of the two models are shown to be consistent with each other [67,82]. Subtle differences exist, however, in the way the system is driven to criticality [67,72]. The effect of boundary conditions also need to be treated with care [67].

To eliminate inertia effects in real sand pile avalanches [64] which prevented direct comparison with the BTW model, the Oslo group [15] designed a two-dimensional rice pile experiment that displays power-law scaling and self-organized criticality. The group also proposed an automaton model, known as the Oslo model, which yields critical exponents in good agreement with the experiment. The Oslo model differs from the BTW model only in the choice of the threshold value K for toppling: Instead of taking a constant value, K is site-dependent. Its value is updated after each toppling event at the site. Consequently, condition (22) changes to,

$$\nabla^2 u + z_0(\mathbf{r}) - K(\mathbf{r}, u) > 0 \,. \tag{23}$$

The symmetry $u \to u + 1$ is no longer present. A similar discussion as above links the Oslo rice-pile model to the LIM of depinning [8,71]. Other SOC models, such as the Manna model [57] and the Zaitsev model [87], have similar updating rules as the Oslo model and hence belong to the LIM universality class as well. The mechanism of self-organization and the development of spatiotemporal correlations in the SOC and depinning models have been discussed in detail by Paczuski, Maslov and Bak [72], who have also identified a large number of scaling relations that give all other exponents in terms of two basic ones.

As an example, let us consider the power-law distribution of the avalanche size which can be measured, e. g., by the total number of toppling events s when a new grain is added. For a system of linear size L, the distribution satisfies the finite-size scaling,

$$P(s, L) = s^{-\tau} F(s/L^{\delta}) \,. \tag{24}$$

The mapping discussed above identifies $\delta = D + \zeta$, i. e., the dimension of the total volume swept by the interface in a system-wide avalanche. On the other hand, a sum rule yields the following expression,

$$\tau = 2 - \frac{\zeta + \nu^{-1}}{\zeta + D} \tag{25}$$

For example, the exponents $\zeta = 0$ and $\nu = 1/2$ for the CDW depinning predicts $\tau = 2(D - 1)/D$, in excellent agreement with simulation results on the sandpile model in two and three dimensions [82]. Care must be taken, however, in applying Eq. (25) to the $D = 1$ rice pile model in the boundary driven case, as noted by Paczuski and Boettcher [71].

Interface Depinning in Anisotropic Medium

Easy and Hard Directions of Depinning

The linear interface model defined by Eq. (5) together with Eq. (6) for the random force is statistically invariant under the *tilt transformation* $u \to u + \mathbf{s} \cdot \mathbf{r}$, where $\mathbf{s}$ is the slope of the interface. Hence the depinning threshold F_c, as well

as other critical properties, are independent of the orientation of the interface. This property holds for an interface moving in an isotropic medium.

When the medium is anisotropic, one or more of the model parameters that enter the estimate Eq. (4) for F_c may change with $\mathbf{s}$. Consequently, the depinning threshold $F_c(\mathbf{s})$ becomes orientation-dependent [83]. An interface oriented in the "easy" direction enters the moving phase at the weakest driving. However, a moving interface in this direction is generically unstable against faceting towards directions of slower growth, particularly near the depinning threshold. To illustrate this point, we may consider a part of the interface that is slowed down by an exceptionally strong pin. As the rest of the interface moves forward at a higher velocity, a local tilt develops that moves the interface away from the easy direction. This further strengthens the pinning effect, producing a cone-like structure. Indeed, the depinning transition in the easy direction has been shown to be of first order with a velocity jump [38].

The story is different for an interface oriented in the "hard" direction. Both numerical [2,3,11,84] and experimental [2,11,12] studies have shown that the depinning transition here is continuous, but the critical exponents take different values from those of the LIM. Yet another set of critical exponents are encountered when the interface is forced to tilt away from the hard direction by, say boundary conditions [83]. The resulting depinning problem is related to *directed percolation*, which we consider in some detail below.

A Lattice Automaton

Tang and Leschhorn [84] introduced a lattice automaton for imbibition in a two-dimensional porous medium. A very similar model was studied by Buldyrev et al. [11] around the same time and reported together with experimental results. The automaton model is defined as follows. On a square lattice, each site (i, j) is assigned a random pinning force $\eta(i, j)$, uniformly distributed in the interval [0,1). At $t = 0$, the interface h_i is completely flat. In each time step, growth $h_i \to h_i + 1$ is performed in parallel when either of the following two conditions is satisfied: (i) the random force $\eta(i, h_i)$ on the interface at column i is less than a pre-specified driving force f; or (ii) $h_i < h_{i-1} - 1$ or $h_i < h_{i+1} - 1$. Thus to obtain a completely pinned interface, we must have $|h_i - h_{i-1}| \leq 1$ and $\eta(i, h_i) > f$ (called a blocking site) for all i. This condition requires the existence of a *directed percolating path* through blocking sites. Such paths exist if the density p of blocking sites exceeds a critical value $p_c \simeq 0.539$. Consequently, the depinning threshold is given by $f_c = 1 - p_c \simeq 0.461$. It is possible to show that, for $f < f_c$, the interface stops at the first directed percolating path that lies fully above its initial position.

The roughness of the incipient directed percolating path at $p = p_c$ is $\zeta = \nu_\perp / \nu_\parallel \simeq 0.63$, where $\nu_\parallel = 1.733$ and $\nu_\perp = 1.097$ are exponents that characterize the divergence of parallel and perpendicular correlation lengths, $\xi_\parallel \sim |p - p_c|^{-\nu_\parallel}, \xi_\perp \sim |p - p_c|^{-\nu_\perp}$, respectively [41]. This is also the roughness exponent assumed by the interface at the depinning threshold $f = f_c$. Simulations [84] have shown that the dynamic exponent $z = 1$ at the depinning transition of the above model, while the crossover length exponent $\nu = \nu_\parallel$. From Eq. (8) one obtains $\theta = \nu_\parallel - \nu_\perp \simeq 0.63$.

The orientational dependence of the depinning threshold in the automaton model has been studied by applying the helical boundary condition $h_{i+L} = h_i + sL$ which forces a global tilt [83]. The threshold force is indeed found to be lower. This behavior matches well with the well-known property of directed percolation: For $p > p_c$, percolating paths fall within a cone of opening angle $\varphi(p)$ around the symmetry direction [41]. If the average tilt s of the interface exceeds this angle, no spanning percolating path exists and the interface is free to move. The roughness exponent of the cone boundary is $\zeta = 1/2$, which belongs to yet another universality class of interface depinning away from a symmetry direction [83].

The above automaton model has been generalized by Buldyrev et al. [12] to higher dimensions. The directed percolating strings are replaced by directed percolating surfaces which have been considered in the context of resistor-diode percolation [20]. The roughness exponent of such a surface at the percolation threshold in $D + 1$ dimensions is given in Table 2. Havlin et al. [36] investigated in detail the temporal sequence of growth activities at the depinning transition, which exhibits an interesting super-diffusive behavior. They argued that the dynamic exponent z that describes the growth in size of the affected region after an initial depinning event is related to the scaling of minimal path length with Euclidean distance on the critical (isotropic) percolation cluster in D dimensions. The values of z from their work are also given in Table 2.

As in the $1 + 1$ dimensional case, a tilted interface has a lower depinning threshold. Just above the depinning threshold, the interface decomposes into stripes of inactive regions perpendicular to the tilt, separated by active fronts that propagate towards the "easy" direction at a finite velocity. Based on this phenomenology, Tang, Kardar and Dhar [83] determined exponents characterizing the depinning transition of a tilted interface. In particular, the velocity exponent $\theta = 1$ in all dimensions.

Collective Transport and Depinning, Table 2
Summary of the exponents for directed percolation depinning along the hard direction. After [2]

	$D = 1$	$D = 2$	$D = 3$	$D = 4$	$D = 5$	$D = 6$	MF
ζ (roughness)	0.63(1)	0.48(3)	0.35(1)	0.27(5)	0.25(5)	0.2(2)	0
z (dynamic)	1	1.15(5)	1.36 (5)	1.58(5)	1.7(1)	1.8(2)	2
ν (correlation length)	1.73(2)	1.16(5)	0.95(10)	0.66(10)	0.6(1)	0.5(1)	$1/2$
θ (velocity)	0.58(7)	0.8(2)	1.0(2)	1.0(2)			1

Quenched KPZ Equation

The slope-dependent depinning threshold can be modeled explicitly by adding the Kardar–Parisi–Zhang (KPZ) term to Eq. (5). The resulting equation of motion takes the form,

$$\mu^{-1}\frac{\partial u}{\partial t} = \gamma\nabla^2 u + \frac{1}{2}\lambda(\nabla u)^2 + \eta(\mathbf{r}, u) + F\,. \qquad (26)$$

In the original proposal [40], the λ-term is due to a kinematic effect known as lateral growth, and hence its strength is proportional to the interface velocity v. At the depinning threshold, however, this term is present only when the medium is anisotropic. A positive λ describes depinning along a hard direction, while a negative λ corresponds to growth along an easy direction. Directed numerical integrations [17,54] of Eq. (26) in $(1 + 1)$ dimensions yielded exponents consistent with the directed percolation models. The relevance of the KPZ nonlinearity in modifying the critical behavior of the LIM has also been confirmed by renormalization group calculations [47,81].

Future Directions

Through the intensive work by many groups in the past fifteen years, the depinning transition of elastic manifolds in a disordered medium has emerged as one of the best understood nonequilibrium critical phenomena with nontrivial scaling properties. The discovery of a renormalizable continuum theory provided the much needed theoretical foundation for the identification of universality classes and symmetry principles. Owing to this remarkable development, models and experimental systems that differ in microscopic details can be compared and classified with regard to their threshold behavior. More recent refinements of the theory by the Paris group (e. g., [46]) have yielded deeper insights on the role of the cusp singularity of the random force correlator in capturing the multiple-minima energy landscape of the disordered manifold problem.

Within the class of models describe by Eq. (5), there are still a few unresolved issues. For example, the roughness exponent ζ of the $(1 + 1)$-dimensional LIM has been found numerically to be indistinguishable to $5/4$, suggesting the possibility of an exact derivation. The effect of a thermal noise term added to the deterministic model has not been well-understood [13,59,86]. There now exist very good numerical estimates of the exponent ψ that describes the vanishing of the interface velocity $v \sim T^{\psi}$ with temperature T at the depinning threshold, but its value does not agree with any of the existing theoretical proposals [13]. Another puzzle is the closeness between the numerical estimates for the dynamical exponent z given in Table 1 and the one-loop RG prediction for the CDW depinning.

The mapping between the CDW depinning and the sandpile models offers a very powerful tool for the development of RG theories for systems that exhibit SOC. To our knowledge, this connection has not been fully explored so far. It would be interesting to see if SOC models other than the sandpile and the rice-pile, such as the Olami–Feder–Christensen model [70] for earthquakes, can also be mapped to some type of depinning model. Conversely, it would be nice to find out whether the lessons learned from the detailed characterization of avalanche dynamics [19] can contribute to a better understanding of critical correlations in the LIM at the depinning threshold from small to large scales.

On the experimental side, perhaps the best studied system is the CDW depinning, yet the CDW velocity generally grows faster than linear above the depinning threshold [85], while the elastic depinning theory predicts $\theta < 1$. It has been suggested that the apparent discrepancy could be due to a combination of factors: Nonuniform driving, thermal activation across energy barriers, and plasticity when dislocations in the phase field is present. Simultaneous presence of strong pinning sites and weak collective pinning in a finite system may also give rise to a complex I–V curve in experimental systems. While some of these effects have been investigated theoretically [79] (see also the review by Brazovskii and Nattermann), a clear picture is yet to emerge.

Moulinet et al. [62] designed an experimental system to study the roughness and dynamics of the contact line of a viscous fluid. The roughness exponent ζ obtained from experimental measurements is 0.5, much bigger than the value 0.385 obtained from numerical simulations [21] of the model considered by Ertas and Kardar [24]. Simi-

lar discrepancies have been reported in the experimental studies [9,74] of the roughness of crack fronts. Modifications of the LIM to include nonlinear couplings and wave-like dynamics [9] for avalanche propagation have been suggested. However, many unresolved issues remain.

It is perhaps not surprising to see that real physical systems almost always contain complications that invalidate direct comparison with the linear elasticity theory of depinning. The pinned state just below the transition is highly metastable and hence is susceptible to various relaxational processes that could possibly invalidate our simple assumptions (see, however [42] for a discussion of correlation lengths below the elastic depinning threshold). Inhomogeneities inside the sample, e. g., macroscopic variations in the concentration of impurity atoms, may also have a dramatic effect on the critical properties associated with the depinning transition. Incorporating the relevant microscopic processes into the generic description discussed here is expected to yield a more complete theory of depinning. It is hoped that future work in this direction will unleash the full impact of the theoretical progress that have been made in the past two decades, which in itself has been intellectually extremely stimulating.

Bibliography

Primary Literature

1. Aharony A, Imry Y, Ma SK (1976) Lowering of dimensionality in phase transitions with random fields. Phys Rev Lett 37:1364–1367
3. Amaral LAN, Barabasi AL, Stanley HE (1994) Universality classes for interface growth with quenched disorder. Phys Rev Lett 73:62–65
2. Amaral LAN, Barabasi AL, Buldyrev SV, Harrington ST, Havlin S, Lahijany RS, Stanley HE (1995) Avalanches and the directed percolation depinning model: Experiments, simulations, and theory. Phys Rev E 51:4655–4673
4. Anderson PW (1984) Basic notions of condensed matter physics. Benjamin/Cummings, Menlo Park
5. Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality: An explanation for $1/f$ noise. Phys Rev Lett 59:381–384
6. Bak P, Tang C, Wiesenfeld K (1988) Self-organized criticality. Phys Rev A 38:364–374
7. Bolech CJ, Rosso A (2004) Universal statistics of the critical depinning force of elastic systems in random media. Phys Rev Lett 93:125701
8. Bonachela JA, Chaté H, Dornic I, Munoz MA (2007) Absorbing states and elastic interfaces in random media: Two equivalent descriptions of self-organized criticality. Phys Rev Lett 98:155702
9. Bouchaud E, Bouchaud JP, Fisher DS, Ramanathan S, Rice JR (2002) Can crack front waves explain the roughness of cracks? J Mech Phys Solids 50:1703–1725
10. Bruinsma R, Aeppli G (1984) Interface motion and nonequilibrium properties of the random-field ising model. Phys Rev Lett 52:1547–1550
11. Buldyrev SV, Barabasi AL, Caserta F, Havlin S, Stanley HE, Vicsek T (1992) Anomalous interface roughening in porous media: Experiment and model. Phys Rev A 45:R8313–R8316
12. Buldyrev SV, Barabasi AL, Havlin S, Kertesz J, Stanley HE, Xenias HS (1992) Anomalous interface roughening in 3D porous media: Experiment and model. Physica A 191:220–226
13. Bustingorry S, Kolton AB, Giamarchi T (2008) Thermal rounding of the depinning transition. Europhys Lett 81:26005
14. Chauve P, Le Doussal P, Wiese KJ (2001) Renormalization of pinned elastic systems: How does it work beyond one loop? Phys Rev Lett 86:1785–1788
15. Christensen K, Corral A, Frette V, Feder J, Jossang T (1996) Tracer dispersion in a self-organized critical system. Phys Rev Lett 77:107–110
16. Coppersmith SN, Millis AJ (1991) Diverging strains in the phase-deformation model of sliding charge-density waves. Phys Rev B 44:7799–7807
17. Csahók Z, Honda K, Vicsek T (1993) Dynamics of surface roughening in disordered media. J Phys A 26:L171–L178
18. de Gennes PG (1985) Wetting: Statics and dynamics. Rev Mod Phys 57:827–863
19. Dhar D (2006) Theoretical studies of self-organized criticality. Physica A 369:29–70
20. Dhar D, Barma M, Phani MK (1981) Duality transformations for two-dimensional directed percolation and resistance problems. Phys Rev Lett 47:1238–1241
21. Duemmer O, Krauth W (2007) Depinning exponents of the driven long-range elastic string. J Stat Mech 2007:P01019
22. Efetov KB, Larkin AI (1977) Charge-density wave in a random potential. Sov Phys JETP 45:1236–1241
23. Ertas D, Kardar M (1994) Anisotropic scaling in depinning of a flux line. Phys Rev Lett 73:1703–1706
24. Ertas D, Kardar M (1994) Critical dynamics of contact line depinning. Phys Rev E 49:R2532–R2535
25. Fedorenko AA, Doussal PL, Wiese KJ (2006) Universal distribution of threshold forces at the depinning transition. Phys Rev B 74:041110
26. Feigel'man MV (1983) Propagation of a plane front in an inhomogeneous medium. Sov Phys JETP 58:1076–1077
27. Fisher DS (1985) Sliding charge-density waves as a dynamic critical phenomenon. Phys Rev B 31:1396–1427
28. Fisher DS (1986) Interface fluctuations in disordered systems: $5-\varepsilon$ expansion and failure of dimensional reduction. Phys Rev Lett 56:1964–1967
29. Fukuyama H, Lee PA (1978) Dynamics of the charge-density wave. I. Impurity pinning in a single chain. Phys Rev B 17:535–541
30. Giamarchi T, Le Doussal P (1994) Elastic theory of pinned flux lattices. Phys Rev Lett 72:1530–1533
31. Granato A, Lüke K (1956) Theory of mechanical damping due to dislocations. J Appl Phys 27:583–593
32. Grinstein G (1976) Ferromagnetic phase transitions in random fields: The breakdown of scaling laws. Phys Rev Lett 37:944–947
33. Grüner G (1988) The dynamics of charge-density waves. Rev Mod Phys 60:1129–1181
34. Halpin-Healy T, Zhang YC (1995) Kinetic roughening phenomena, stochastic growth, directed polymers and all that. Aspects of multidisciplinary statistical mechanics. Phys Rep 254:215–414

35. Harris AB (1974) Effect of random defects on the critical behaviour of Ising models. J Phys C 7:1671–1692
36. Havlin S, Amaral LAN, Buldyrev SV, Harrington ST, Stanley HE (1995) Dynamics of surface roughening with quenched disorder. Phys Rev Lett 74:4205–4208
37. Imry Y, Ma SK (1975) Random-field instability of the ordered state of continuous symmetry. Phys Rev Lett 35:1399–1401
38. Jeong H, Kahng B, Kim D (1996) Anisotropic surface growth model in disordered media. Phys Rev Lett 77:5094–5097
39. Joanny JF, de Gennes PG (1984) A model for contact angle hysteresis. J Chem Phys 81:552–562
40. Kardar M, Parisi G, Zhang YC (1986) Dynamic scaling of growing interfaces. Phys Rev Lett 56:889–892
41. Kinzel W (1982) Directed Percolation. In: Deutscher G, Zallen R, Adler J (eds) Percolation structures and processes. Annals of the Israel Physical Society, vol 5. Hilger, Bristol, p 425
42. Kolton AB, Rosso A, Giamarchi T, Krauth W (2006) Dynamics below the depinning threshold in disordered elastic systems. Phys Rev Lett 97:057001
43. Koplik J, Levine H (1985) Interface moving through a random background. Phys Rev B 32:280–292
44. Larkin AI (1970) Effect of inhomogeneities on the structure of the mixed state of superconductors. Sov Phys JETP 31:784–786
45. Larkin AI, Ovchinnikov YN (1979) Pinning in type-II superconductors. J Low Temp Phys 34:409–428
47. Le Doussal P, Wiese KJ (2003) Functional renormalization group for anisotropic depinning and relation to branching processes. Phys Rev E 67:016121
48. Le Doussal P, Wiese KJ, Chauve P (2002) Two-loop functional renormalization group theory of the depinning transition. Phys Rev B 66:174201
49. Le Doussal P, Wiese KJ, Chauve P (2004) Functional renormalization group and the field theory of disordered elastic systems. Phys Rev E 69:026112
46. Le Doussal P, Muller M, Wiese KJ (2008) Cusps and shocks in the renormalized potential of glassy random manifolds: How functional renormalization group and replica symmetry breaking fit together. Phys Rev B 77:064203
50. Lee PA, Rice TM (1979) Electric field depinning of charge density waves. Phys Rev B 19:3970–3980
51. Leger L, Joanny JF (1992) Liquid spreading. Rep Prog Phys 55:431–486
52. Leschhorn H (1992) Interface motion in a random medium: Mean field theory. J Phys A 25:L555–L560
53. Leschhorn H (1993) Interface depinning in a disordered medium? Numerical results. Physica A 195:324–335
54. Leschhorn H (1996) Anisotropic interface depinning: Numerical results. Phys Rev E 54:1313–1320
56. Leschhorn H, Tang LH (1993) Comment on "Elastic string in a random potential". Phys Rev Lett 70:2973
55. Leschhorn H, Nattermann T, Stepanow S, Tang LH (1997) Driven interface depinning in a disordered medium. Ann Phys (Leipzig) 6:1–34
57. Manna S (1991) Two-state model of self-organized criticality. J Phys A 24:L363–L369
58. Middleton AA (1992) Asymptotic uniqueness of the sliding state for charge-density waves. Phys Rev Lett 68:670–673
59. Middleton AA (1992) Thermal rounding of the charge-density-wave depinning transition. Phys Rev B 45:9465–9468
60. Middleton AA, Fisher DS (1991) Critical behavior of pinned charge-density waves below the threshold for sliding. Phys Rev Lett 66:92–95
61. Middleton AA, Fisher DS (1993) Critical behavior of charge-density waves below threshold: Numerical and scaling analysis. Phys Rev B 47:3530–3552
62. Moulinet S, Guthmann C, Rolley E (2002) Roughness and dynamics of a contact line of a viscous fluid on a disordered substrate. Eur Phys J E 8:437–443
63. Myers CR, Sethna JP (1993) Collective dynamics in a model of sliding charge-density waves. I. Critical behavior. Phys Rev B 47:11171–11192
64. Nagel SR (1992) Instabilities in a sandpile. Rev Mod Phys 64:321–325
65. Narayan O, Fisher DS (1992) Critical behavior of sliding charge-density waves in $4-\varepsilon$ dimensions. Phys Rev B 46:11520–11549
66. Narayan O, Fisher DS (1993) Threshold critical dynamics of driven interfaces in random media. Phys Rev B 48:7030–7042
67. Narayan O, Middleton AA (1994) Avalanches and the renormalization group for pinned charge-density waves. Phys Rev B 49:244–256
68. Nattermann T (1990) Scaling approach to pinning: Charge density waves and giant flux creep in superconductors. Phys Rev Lett 64:2454–2457
69. Nattermann T, Stepanow S, Tang LH, Leschhorn H (1992) Dynamics of interface depinning in a disordered medium. J Phys II (Paris) 2:1483–1488
70. Olami Z, Feder HJS, Christensen K (1992) Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes. Phys Rev Lett 68:1244–1247
71. Paczuski M, Boettcher S (1996) Universality in sandpiles, interface depinning, and earthquake models. Phys Rev Lett 77:111–114
72. Paczuski M, Maslov S, Bak P (1996) Avalanche dynamics in evolution, growth, and depinning models. Phys Rev E 53:414–443
73. Podgorski T, Flesselles JM, Limat L (2001) Corners, cusps, and pearls in running drops. Phys Rev Lett 87:036102
74. Ponson L, Bonamy D, Bouchaud E (2006) Two-dimensional scaling properties of experimental fracture surfaces. Phys Rev Lett 96:035506
75. Ramanathan S, Fisher DS (1998) Onset of propagation of planar cracks in heterogeneous media. Phys Rev B 58:6026–6046
76. Rosso A, Hartmann AK, Krauth W (2003) Depinning of elastic manifolds. Phys Rev E 67:021602
77. Rosso A, Le Doussal P, Wiese KJ (2007) Numerical calculation of the funcitonal renormalization group fixed-point functions at the depinning transition. Phys Rev B 75:220201(R)
78. Ryu KS, Akinaga H, Shin SC (2007) Tunable scaling behaviour observed in Barkhausen criticality of a ferromagnetic film. Nature Physics 3:547–550
79. Saunders K, Schwarz JM, Marchetti MC, Middleton AA (2004) Mean-field theory of collective transport with phase slips. Phys Rev B 70:024205
80. Sethna JP, Dahmen K, Kartha S, Krumhansl JA, Roberts BW, Shore JD (1993) Hysteresis and hierarchies: Dynamics of disorder-driven first-order phase transformations. Phys Rev Lett 70:3347–3350
81. Stepanow S (1995) Dynamics of growing interfaces in disordered medium: The effect of lateral growth. J Phys II (France) 5:11–18

82. Tang C, Bak P (1988) Critical exponents and scaling relations for self-organized criticality phenomena. Phys Rev Lett 60:2347–2350
84. Tang LH, Leschhorn H (1992) Pinning by directed percolation. Phys Rev A 45:R8309–R8312
83. Tang LH, Kardar M, Dhar D (1995) Driven depinning in anisotropic media. Phys Rev Lett 74:920–923
85. Thorne RE (2005) A history of the I–V characteristic of CDW conductors. J Phys IV (France) 131:89–94
86. Vandembroucq D, Skoe R, Roux S (2004) Universal depinning force fluctuations of an elastic line: Application to finite temperature behavior. Phys Rev E 70:051101
87. Zaitsev SI (1992) Robin Hood as self-organized criticality. Physica A 189:411–416
88. Zapperi S, Cizeau P, Durin G, Stanley HE (1998) Dynamics of a ferromagnetic domain wall: Avalanches, depinning transition, and the Barkhausen effect. Phys Rev B 58:6353–6366

Books and Reviews

Alava M, Dubé M, Rost M (2004) Imbibition in disordered media. Adv Phys 53:83–175
Blatter G, Feigelman MV, Geshkenbein VB, Larkin AI, Vinokur VM (1994) Vortices in high-temperature superconductors. Rev Mod Phys 66:1125–1388
Brazovskii S, Nattermann T (2004) Pinning and sliding of driven elastic systems: From domain walls to charge density waves. Adv Phys 53:177–252
de Gennes PG, Brochard-Wyart F, Quéré D (2003) Capillarity and Wetting Phenomena: Drops, Bubbles, Pearls, Waves. Springer, New York
Fisher DS (1998) Collective transport in random media: From superconductors to earthquakes. Phys Rep 301:113–150
Kardar M (1998) Nonequilibrium dynamics of interfaces and lines. Phys Rep 301:85–112
Quéré D (2005) Non-sticking drops. Rep Prog Phys 68:2495–2532
Sethna JP, Dahmen KA, Myers CR (2001) Crackling noise. Nature 410:242–250
Turcotte DL (1999) Self-organized criticality. Rep Prog Phys 62:1377–1429

Community Structure in Graphs

Santo Fortunato[1], Claudio Castellano[2]
[1] Complex Networks Lagrange Laboratory (CNLL), ISI Foundation, Torino, Italy
[2] SMC, INFM-CNR and Dipartimento di Fisica, "Sapienza" Università di Roma, Roma, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Elements of Community Detection
Computer Science: Graph Partitioning
Social Science: Hierarchical and k-Means Clustering
New Methods
Testing Methods
The Mesoscopic Description of a Graph
Future Directions
Bibliography

Glossary

Graph A graph is a set of elements, called *vertices* or *nodes*, where pairs of vertices are connected by relational links, or *edges*. A graph can be considered as the simplest representation of a complex system, where the vertices are the elementary units of the system and the edges represent their mutual interactions.

Community A community is a group of graph vertices that "belong together" according to some precisely defined criteria which can be measured. Many definitions have been proposed. A common approach is to define a community as a group of vertices such that the density of edges between vertices of the group is higher than the average edge density in the graph. In the text also the terms *module* or *cluster* are used when referring to a community.

Partition A partition is a split of a graph in subsets with each vertex assigned to only one of them. This last condition may be relaxed to include the case of overlapping communities, imposing that each vertex is assigned to at least one subset.

Dendrogram A dendrogram, or hierarchical tree, is a branching diagram representing successive divisions of a graph into communities. Dendrograms are frequently used in social network analysis and computational biology, especially in biological taxonomy.

Scalability Scalability expresses the computational complexity of an algorithm. If the running time of a community detection algorithm, working on a graph with n vertices and m edges, is proportional to the product $n^\alpha m^\beta$, one says that the algorithm scales as $O(n^\alpha m^\beta)$. Knowing the scalability allows to estimate the range of applicability of an algorithm.

Definition of the Subject

Graph vertices are often organized into groups that seem to live fairly independently of the rest of the graph, with which they share but a few edges, whereas the relationships between group members are stronger, as shown by the large number of mutual connections. Such groups of vertices, or communities, can be considered as independent compartments of a graph. Detecting communities is of great importance in sociology, biology and com-

puter science, disciplines where systems are often represented as graphs. The task is very hard, though, both conceptually, due to the ambiguity in the definition of community and in the discrimination of different partitions and practically, because algorithms must find "good" partitions among an exponentially large number of them. Other complications are represented by the possible occurrence of hierarchies, i. e. communities which are nested inside larger communities, and by the existence of overlaps between communities, due to the presence of nodes belonging to more groups. All these aspects are dealt with in some detail and many methods are described, from traditional approaches used in computer science and sociology to recent techniques developed mostly within statistical physics.

Introduction

The origin of graph theory dates back to Euler's solution [1] of the puzzle of Königsberg's bridges in 1736. Since then a lot has been learned about graphs and their mathematical properties [2]. In the 20th century they have also become extremely useful as representation of a wide variety of systems in different areas. Biological, social, technological, and information networks can be studied as graphs, and graph analysis has become crucial to understand the features of these systems. For instance, social network analysis started in the 1930s and has become one of the most important topics in sociology [3,4]. In recent times, the computer revolution has provided scholars with a huge amount of data and computational resources to process and analyze these data. The size of real networks one can potentially handle has also grown considerably, reaching millions or even billions of vertices. The need to deal with such a large number of units has produced a deep change in the way graphs are approached [5,6,7,8,9].

Real networks are not random graphs. The random graph, introduced by P. Erdös and A. Rényi [10], is the paradigm of a disordered graph: in it, the probability of having an edge between a pair of vertices is equal for all possible pairs. In a random graph, the distribution of edges among the vertices is highly homogeneous. For instance, the distribution of the number of neighbors of a vertex, or *degree*, is binomial, so most vertices have equal or similar degree. In many real networks, instead, there are big inhomogeneities, revealing a high level of order and organization. The degree distribution is broad, with a tail that often follows a power law: therefore, many vertices with low degree coexist with some vertices with large degree. Furthermore, the distribution of edges is not only globally, but also locally inhomogeneous, with high concentrations of edges within special groups of nodes, and low concentrations between these groups. This feature of real networks is called *community structure* and is the topic of this chapter. In Fig. 1 a schematic example of a graph with community structure is shown.

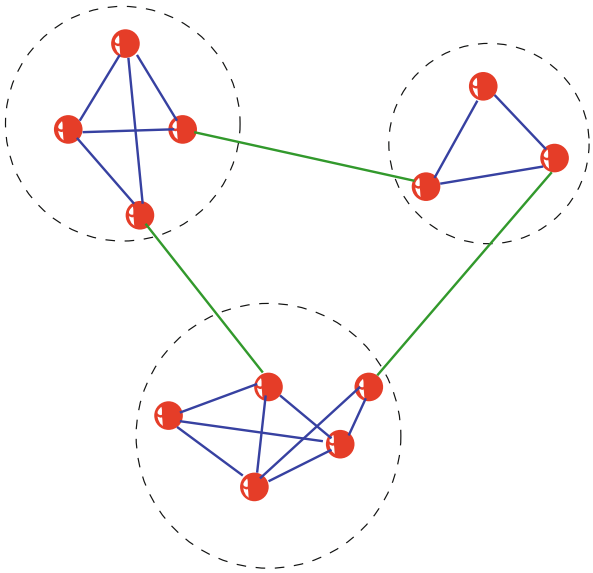

Community Structure in Graphs, Figure 1
A simple graph with three communities, highlighted by the *dashed circles*

Communities are groups of vertices which probably share common properties and/or play similar roles within the graph. So, communities may correspond to groups of pages of the World Wide Web dealing with related topics [11], to functional modules such as cycles and pathways in metabolic networks [12,13], to groups of related individuals in social networks [14,15], to compartments in food webs [16,17], and so on.

Community detection is important for other reasons, too. Identifying modules and their boundaries allows for a classification of vertices, according to their topological position in the modules. So, vertices with a central position in their clusters, i. e. sharing a large number of edges with the other group partners, may have an important function of control and stability within the group; vertices lying at the boundaries between modules play an important role of mediation and lead the relationships and exchanges between different communities. Such classification seems to be meaningful in social [18,19,20] and metabolic networks [12]. Finally, one can study the graph where vertices are the communities and edges are set between modules if there are connections between some of their vertices in the original graph and/or if the modules overlap. In this

way one attains a coarse-grained description of the original graph, which unveils the relationships between modules. Recent studies indicate that networks of communities have a different degree distribution with respect to the full graphs [13]; however, the origin of their structures can be explained by the same mechanism [21].

The aim of community detection in graphs is to identify the modules only based on the topology. The problem has a long tradition and it has appeared in various forms in several disciplines. For instance, in parallel computing it is crucial to know what is the best way to allocate tasks to processors so as to minimize the communications between them and enable a rapid performance of the calculation. This can be accomplished by splitting the computer cluster into groups with roughly the same number of processors, such that the number of physical connections between processors of different groups is minimal. The mathematical formalization of this problem is called *graph partitioning*. The first algorithms for graph partitioning were proposed in the early 1970s. Clustering analysis is also an important aspect in the study of social networks. The most popular techniques are hierarchical clustering and k-means clustering, where vertices are joined into groups according to their mutual similarity.

In a seminal paper, Girvan and Newman proposed a new algorithm, aiming at the identification of edges lying between communities and their successive removal, a procedure that after a few iterations leads to the isolation of modules [14]. The intercommunity edges are detected according to the values of a centrality measure, the edge betweenness, that expresses the importance of the role of the edges in processes where signals are transmitted across the graph following paths of minimal length. The paper triggered a big activity in the field, and many new methods have been proposed in the last years. In particular, physicists entered the game, bringing in their tools and techniques: spin models, optimization, percolation, random walks, synchronization, etc., became ingredients of new original algorithms. Earlier reviews of the topic can be found in [22,23].

Section "Elements of Community Detection" is about the basic elements of community detection, starting from the definition of community. The classical problem of graph partitioning and the methods for clustering analysis in sociology are presented in Sect. "Computer Science: Graph Partitioning" and "Social Science: Hierarchical and k-Means Clustering", respectively. Section "New Methods" is devoted to a description of the new methods. In Sect. "Testing Methods" the problem of testing algorithms is discussed. Section "The Mesoscopic Description of a Graph" introduces the description of graphs at the level of communities. Finally, Sect. "Future Directions" highlights the perspectives of the field and sorts out promising research directions for the future.

This chapter makes use of some basic concepts of graph theory, that can be found in any introductory textbook, like [2]. Some of them are briefly explained in the text.

Elements of Community Detection

The problem of community detection is, at first sight, intuitively clear. However, when one needs to formalize it in detail things are not so well defined. In the intuitive concept some ambiguities are hidden and there are often many equally legitimate ways of resolving them. Hence the term "Community Detection" actually indicates several rather different problems.

First of all, there is no unique way of translating into a precise prescription the intuitive idea of community. Many possibilities exist, as discussed below. Some of these possible definitions allow for vertices to belong to more than one community. It is then possible to look for overlapping or nonoverlapping communities. Another ambiguity has to do with the concept of community structure. It may be intended as a single partition of the graph or as a hierarchy of partitions, at different levels of coarse-graining. There is then a problem of comparison. Which one is the best partition (or the best hierarchy)? If one could, in principle, analyze all possible partitions of a graph, one would need a sensible way of measuring their "quality" to single out the partitions with the strongest community structure. It may even occur that one graph has no community structure and one should be able to realize it. Finding a good method for comparing partitions is not a trivial task and different choices are possible. Last but not least, the number of possible partitions grows faster than exponentially with the graph size, so that, in practice, it is not possible to analyze them all. Therefore one has to devise smart methods to find 'good' partitions in a reasonable time. Again, a very hard problem.

Before introducing the basic concepts and discussing the relevant questions it is important to stress that the identification of topological clusters is possible only if the graphs are *sparse*, i. e. if the number of edges m is of the order of the number of nodes n of the graph. If $m \gg n$, the distribution of edges among the nodes is too homogeneous for communities to make sense.

Definition of Community

The first and foremost problem is how to define precisely what a community is. The intuitive notion presented in

the Introduction is related to the comparison of the number of edges joining vertices within a module ("intracommunity edges") with the number of edges joining vertices of different modules ("intercommunity edges"). A module is characterized by a larger density of links "inside" than "outside". This notion can be however formalized in many ways. Social network analysts have devised many definitions of subgroups with various degrees of internal cohesion among vertices [3,4]. Many other definitions have been introduced by computer scientists and physicists. In general, the definitions can be classified in three main categories.

- *Local definitions*. Here the attention is focused on the vertices of the subgraph under investigation and on its immediate neighborhood, disregarding the rest of the graph. These prescriptions come mostly from social network analysis and can be further subdivided in *self-referring*, when one considers the subgraph alone, and *comparative*, when the mutual cohesion of the vertices of the subgraph is compared with their cohesion with the external neighbors. Self-referring definitions identify classes of subgraphs like *cliques*, *n-cliques*, *k-plexes*, etc. They are *maximal subgraphs*, which cannot be enlarged with the addition of new vertices and edges without losing the property which defines them. The concept of clique is very important and often recurring when one studies graphs. A clique is a maximal subgraph where each vertex is adjacent to all the others. In the literature it is common to call cliques also non-maximal subgraphs. Triangles are the simplest cliques, and are frequent in real networks. Larger cliques are rare, so they are not good models of communities. Besides, finding cliques is computationally very demanding: the Bron–Kerbosch method [24] runs in a time growing exponentially with the size of the graph. The definition of clique is very strict. A softer constraint is represented by the concept of n-clique, which is a maximal subgraph such that the distance of each pair of its vertices is not larger than n. A k-plex is a maximal subgraph such that each vertex is adjacent to all the others except at most k of them. In contrast, a *k-core* is a maximal subgraph where each vertex is adjacent to at least k vertices within the subgraph. Comparative definitions include that of *LS set*, or *strong community*, and that of *weak community*. An LS set is a subgraph where each node has more neighbors inside than outside the subgraph. Instead, in a weak community, the total degree of the nodes inside the community exceeds the external total degree, i. e. the number of links lying between the community and the rest of the graph. LS sets are also weak communities, but the inverse is not true, in general. The notion of weak community was introduced by Radicchi et al. [25].
- *Global definitions*. Communities are structural units of the graph, so it is reasonable to think that their distinctive features can be recognized if one analyzes a subgraph with respect to the graph as a whole. Global definitions usually start from a *null model*, i. e. a graph which matches the original in some of its topological features, but which does not display community structure. After that, the linking properties of subgraphs of the initial graph are compared with those of the corresponding subgraphs in the null model. The simplest way to design a null model is to introduce randomness in the distribution of edges among the vertices. A random graph à la Erdös–Rényi, for instance, has no community structure, as any two vertices have the same probability to be adjacent, so there is no preferential linking involving special groups of vertices. The most popular null model is that proposed by Newman and Girvan and consists of a randomized version of the original graph, where edges are rewired at random, under the constraint that each vertex keeps its degree [26]. This null model is the basic concept behind the definition of *modularity*, a function which evaluates the goodness of partitions of a graph into modules (see Sect. "Evaluating Partitions: Quality Functions"). Here a subset of vertices is a community if the number of edges inside the subset exceeds the expected number of internal edges that the subset would have in the null model. A more general definition, where one counts small connected subgraphs (*motifs*), and not necessarily edges, can be found in [27]. A general class of null models, including that of modularity, has been designed by Reichardt and Bornholdt [28].
- *Definitions based on vertex similarity*. In this last category, communities are groups of vertices which are similar to each other. A quantitative criterion is chosen to evaluate the similarity between each pair of vertices, connected or not. The criterion may be local or global: for instance one can estimate the distance between a pair of vertices. Similarities can be also extracted from eigenvector components of special matrices, which are usually close in value for vertices belonging to the same community. Similarity measures are at the basis of the method of hierarchical clustering, to be discussed in Sect. "Social Science: Hierarchical and k-Means Clustering". The main problem in this case is the need to introduce an additional criterion to select meaningful partitions.

It is worth remarking that, in spite of the wide variety of definitions, in many detection algorithms communities are not defined at all, but are a byproduct of the procedure. This is the case of the divisive algorithms described in Sect. "Divisive Algorithms" and of the dynamic algorithms of Sect. "Dynamic Algorithms".

Evaluating Partitions: Quality Functions

Strictly speaking, a partition of a graph in communities is a split of the graph in clusters, with each vertex assigned to only one cluster. The latter condition may be relaxed, as shown in Sect. "Overlapping Communities". Whatever the definition of community is, there is usually a large number of possible partitions. It is then necessary to establish which partitions exhibit a real community structure. For that, one needs a *quality function*, i. e. a quantitative criterion to evaluate how good a partition is. The most popular quality function is the modularity of Newman and Girvan [26]. It can be written in several ways, as

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j) \,, \tag{1}$$

where the sum runs over all pairs of vertices, A is the adjacency matrix, k_i the degree of vertex i and m the total number of edges of the graph. The element A_{ij} of the adjacency matrix is 1 if vertices i and j are connected, otherwise it is 0. The δ-function yields one if vertices i and j are in the same community, zero otherwise. Because of that, the only contributions to the sum come from vertex pairs belonging to the same cluster: by grouping them together the sum over the vertex pairs can be rewritten as a sum over the modules

$$Q = \sum_{s=1}^{n_m} \left[\frac{l_s}{m} - \left(\frac{d_s}{2m} \right)^2 \right] . \tag{2}$$

Here, n_m is the number of modules, l_s the total number of edges joining vertices of module s and d_s the sum of the degrees of the vertices of s. In Eq. (2), the first term of each summand is the fraction of edges of the graph inside the module, whereas the second term represents the expected fraction of edges that would be there if the graph were a random graph with the same degree for each vertex. In such a case, a vertex could be attached to any other vertex of the graph, and the probability of a connection between two vertices is proportional to the product of their degrees. So, for a vertex pair, the comparison between real and expected edges is expressed by the corresponding summand of Eq. (1).

Equation (2) embeds an implicit definition of community: a subgraph is a module if the number of edges inside it is larger than the expected number in modularity's null model. If this is the case, the vertices of the subgraph are more tightly connected than expected. Basically, if each summand in Eq. (2) is non-negative, the corresponding subgraph is a module. Besides, the larger the difference between real and expected edges, the more "modular" the subgraph. So, large positive values of Q are expected to indicate good partitions. The modularity of the whole graph, taken as a single community, is zero, as the two terms of the only summand in this case are equal and opposite. Modularity is always smaller than one, and can be negative as well. For instance, the partition in which each vertex is a community is always negative. This is a nice feature of the measure, implying that, if there are no partitions with positive modularity, the graph has no community structure. On the contrary, the existence of partitions with large negative modularity values may hint to the existence of subgroups with very few internal edges and many edges lying between them (*multipartite structure*).

Modularity has been employed as a quality function in many algorithms, like some of the divisive algorithms of Sect. "Divisive Algorithms". In addition, modularity optimization is itself a popular method for community detection (see Sect. "Modularity Optimization"). Modularity also allows to assess the stability of partitions [29] and to transform a graph into a smaller one by preserving its community structure [30].

However, there are some caveats on the use of the measure. The most important concerns the value of modularity for a partition. For which values one can say that there is a clear community structure in a graph? The question is tricky: if two graphs have the same type of modular structure, but different sizes, modularity will be larger for the larger graph. So, modularity values cannot be compared for different graphs. Moreover, one would expect that partitions of random graphs will have modularity values close to zero, as no community structure is expected there. Instead, it has been shown that partitions of random graphs may attain fairly large modularity values, as the probability that the distribution of edges on the vertices is locally inhomogeneous in specific realizations is not negligible [31]. Finally, a recent analysis has proved that modularity increases if subgraphs smaller than a characteristic size are merged [32]. This fact represents a serious bias when one looks for communities via modularity optimization and is discussed in more detail in Sect. "Modularity Optimization".

Hierarchies

Graph vertices can have various levels of organization. Modules can display an internal community structure, i. e.

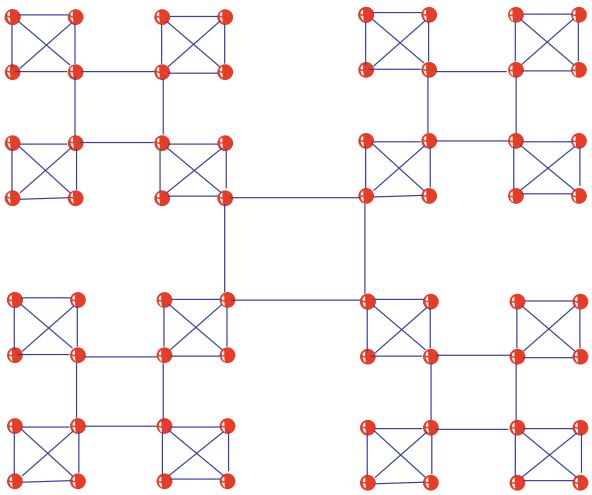

Community Structure in Graphs, Figure 2
Schematic example of a hierarchical graph. Sixteen modules with four vertices each are clearly organized in groups of four

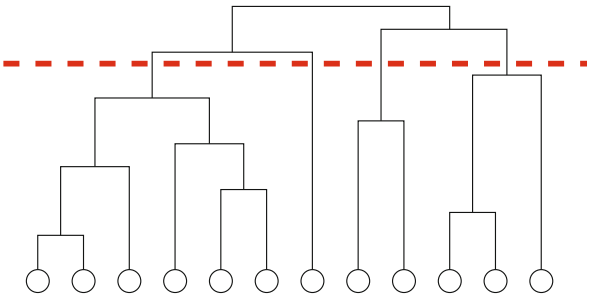

Community Structure in Graphs, Figure 3
A dendrogram, or hierarchical tree. *Horizontal cuts* correspond to partitions of the graph in communities. Reprinted figure with permission from [26]

they can contain smaller modules, which can in turn include other modules, and so on. In this case one says that the graph is hierarchical (see Fig. 2). For a clear classification of the vertices and their roles inside a graph, it is important to find all modules of the graph as well as their hierarchy.

A natural way to represent the hierarchical structure of a graph is to draw a *dendrogram*, like the one illustrated in Fig. 3. Here, partitions of a graph with twelve vertices are shown. At the bottom, each vertex is its own module. By moving upwards, groups of vertices are successively aggregated. Merges of communities are represented by horizontal lines. The uppermost level represents the whole graph as a single community. Cutting the diagram horizontally at some height, as shown in the figure (dashed line), displays one level of organization of the graph vertices. The diagram is hierarchical by construction: each community belonging to a level is fully included in a community at a higher level. Dendrograms are regularly used in sociology and biology. The technique of hierarchical clustering, described in Sect. "Social Science: Hierarchical and *k*-Means Clustering", lends itself naturally to this kind of representation.

Overlapping Communities

As stated in Sect. "Evaluating Partitions: Quality Functions", in a partition each vertex is generally attributed only to one module. However, vertices lying at the boundary between modules may be difficult to assign to one module or another, based on their connections with the other vertices. In this case, it makes sense to consider such intermediate vertices as belonging to more groups, which are then called *overlapping communities* (Fig. 4). Many real networks are characterized by a modular structure with sizeable overlaps between different clusters. In social networks, people usually belong to more communities, according to their personal life and interests: for instance a person may have tight relationships both with the people of its working environment and with other individuals involved in common free time activities.

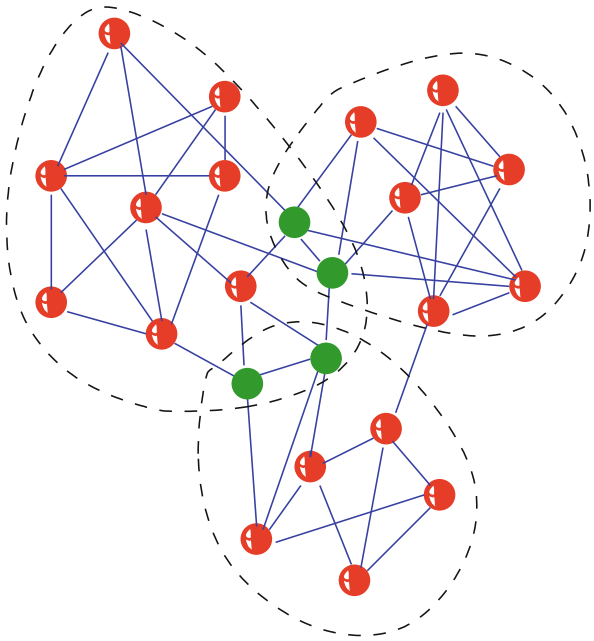

Community Structure in Graphs, Figure 4
Overlapping communities. In the partition highlighted by the *dashed* contours, some vertices are shared between more groups

Accounting for overlaps is also a way to better exploit the information that one can derive from topology. Ideally, one could estimate the degree of participation of a vertex in different communities, which corresponds to the likelihood that the vertex belongs to the various groups. Community detection algorithms, instead, often disagree in the classification of peripheral vertices of modules, because they are forced to put them in a single cluster, which may be the wrong one.

The problem of community detection is so hard that very few algorithms consider the possibility of having overlapping communities. An interesting method has been recently proposed by G. Palla et al. [13] and is described in Sect. "Clique Percolation". For standard algorithms, the problem of identifying overlapping vertices could be addressed by checking for the stability of partitions against slight variations in the structure of the graph, as described in [33].

Computer Science: Graph Partitioning

The problem of graph partitioning consists in dividing the vertices in g groups of predefined size, such that the number of edges lying between the groups is minimal. The number of edges running between modules is called *cut size*. Figure 5 presents the solution of the problem for a graph with fourteen vertices, for $g = 2$ and clusters of equal size.

The specification of the number of modules of the partition is necessary. If one simply imposed a partition with the minimal cut size, and left the number of modules free, the solution would be trivial, corresponding to all vertices ending up in the same module, as this would yield a vanishing cut size.

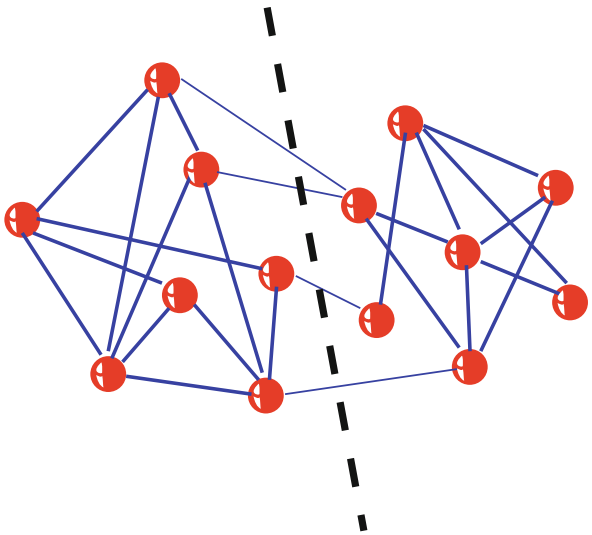

Community Structure in Graphs, Figure 5
Graph partitioning. The *cut* shows the partition in two groups of equal size

Graph partitioning is a fundamental issue in parallel computing, circuit partitioning and layout, and in the design of many serial algorithms, including techniques to solve partial differential equations and sparse linear systems of equations. Most variants of the graph partitioning problem are NP-hard, i. e. it is unlikely that the solution can be computed in a time growing as a power of the graph size. There are however several algorithms that can do a good job, even if their solutions are not necessarily optimal [34]. Most algorithms perform a bisection of the graph, which is already a complex task. Partitions into more than two modules are usually attained by iterative bisectioning.

The *Kernighan–Lin algorithm* [35] is one of the earliest methods proposed and is still frequently used, often in combination with other techniques. The authors were motivated by the problem of partitioning electronic circuits onto boards: the nodes contained in different boards need to be linked to each other with the least number of connections. The procedure is an optimization of a benefit function Q, which represents the difference between the number of edges inside the modules and the number of edges lying between them. The starting point is an initial partition of the graph in two clusters of the predefined size: such initial partition can be random or suggested by some information on the graph structure. Then, subsets consisting of equal numbers of vertices are swapped between the two groups, so that Q has the maximal increase. To reduce the risk to be trapped in local maxima of Q, the procedure includes some swaps that decrease the function Q. After a series of swaps with positive and negative gains, the partition with the largest value of Q is selected and used as starting point of a new series of iterations. The Kernighan–Lin algorithm is quite fast, scaling as $O(n^2)$ in worst-case time, n being the number of vertices. The partitions found by the procedure are strongly dependent on the initial configuration and other algorithms can do better. However, the method is used to improve on the partitions found through other techniques, by using them as starting configurations for the algorithm.

Another popular technique is the *spectral bisection method*, which is based on the properties of the Laplacian matrix. The Laplacian matrix (or simply Laplacian) of a graph is obtained from the adjacency matrix A by placing on the diagonal the degrees of the vertices and by changing the signs of the other elements. The Laplacian has all non-negative eigenvalues and at least one zero eigenvalue,

as the sum of the elements of each row and column of the matrix is zero. If a graph is divided into g connected components, the Laplacian would have g degenerate eigenvectors with eigenvalue zero and can be written in block-diagonal form, i. e. the vertices can be ordered in such a way that the Laplacian displays g square blocks along the diagonal, with entries different from zero, whereas all other elements vanish. Each block is the Laplacian of the corresponding subgraph, so it has the trivial eigenvector with components $(1, 1, 1, \ldots, 1, 1)$. Therefore, there are g degenerate eigenvectors with equal non-vanishing components in correspondence of the vertices of a block, whereas all other components are zero. In this way, from the components of the eigenvectors one can identify the connected components of the graph.

If the graph is connected, but consists of g subgraphs which are weakly linked to each other, the spectrum will have one zero eigenvalue and $g-1$ eigenvalues which are close to zero. If the groups are two, the second lowest eigenvalue will be close to zero and the corresponding eigenvector, also called *Fiedler vector*, can be used to identify the two clusters as shown below.

Every partition of a graph with n vertices in two groups can be represented by an index vector $\mathbf{s}$, whose component $\mathbf{s}_i$ is $+1$ if vertex i is in one group and -1 if it is in the other group. The cut size R of the partition of the graph in the two groups can be written as

$$R = \frac{1}{4}\mathbf{s}^{\mathrm{T}}\mathbf{L}\mathbf{s}\,, \tag{3}$$

where $\mathbf{L}$ is the Laplacian matrix and $\mathbf{s}^{\mathrm{T}}$ the transpose of vector $\mathbf{s}$. Vector $\mathbf{s}$ can be written as $\mathbf{s} = \sum_i a_i \mathbf{v}_i$, where $\mathbf{v}_i$, $i = 1, \ldots, n$ are the eigenvectors of the Laplacian. If $\mathbf{s}$ is properly normalized, then

$$R = \sum_i a_i^2 \lambda_i\,, \tag{4}$$

where λ_i is the Laplacian eigenvalue corresponding to eigenvector $\mathbf{v}_i$. It is worth remarking that the sum contains at most $n-1$ terms, as the Laplacian has at least one zero eigenvalue. Minimizing R equals to the minimization of the sum on the right-hand side of Eq. (4). This task is still very hard. However, if the second lowest eigenvector λ_2 is close enough to zero, a good approximation of the minimum can be attained by choosing $\mathbf{s}$ parallel to the Fiedler vector $\mathbf{v}_2$: this would reduce the sum to λ_2, which is a small number. But the index vector cannot be perfectly parallel to $\mathbf{v}_2$ by construction, because all its components are equal in modulus, whereas the components of $\mathbf{v}_2$ are not. The best one can do is to match the signs of the components. So, one can set $\mathbf{s}_i = +1(-1)$ if $\mathbf{v}_2^i > 0$ (<0). It may happen that the sizes of the two corresponding groups do not match the predefined sizes one wishes to have. In this case, if one aims at a split in n_1 and $n_2 = n - n_1$ vertices, the best strategy is to order the components of the Fiedler vector from the lowest to the largest values and to put in one group the vertices corresponding to the first n_1 components from the top or the bottom, and the remaining vertices in the second group. If there is a discrepancy between n_1 and the number of positive or negative components of $\mathbf{v}_2$, this procedure yields two partitions: the better solution is the one that gives the smallest cut size.

The spectral bisection method is quite fast. The first eigenvectors of the Laplacian can be computed by using the Lanczos method [36], that scales as $m/(\lambda_3 - \lambda_2)$, where m is the number of edges of the graph. If the eigenvalues λ_2 and λ_3 are well separated, the running time of the algorithm is much shorter than the time required to calculate the complete set of eigenvectors, which scales as $O(n^3)$. The method gives in general good partitions, that can be further improved by applying the Kernighan–Lin algorithm.

Other methods for graph partitioning include level-structure partitioning, the geometric algorithm, multilevel algorithms, etc. A good description of these algorithms can be found in [34].

Graph partitioning algorithms are not good for community detection, because it is necessary to provide as input both the number of groups and their size, about which in principle one knows nothing. Instead, one would like an algorithm capable to produce this information in its output. Besides, using iterative bisectioning to split the graph in more pieces is not a reliable procedure.

Social Science: Hierarchical and *k*-Means Clustering

In social network analysis, one partitions actors/vertices in clusters such that actors in the same cluster are more similar between themselves than actors of different clusters. The two most used techniques to perform clustering analysis in sociology are *hierarchical clustering* and *k-means clustering*.

The starting point of hierarchical clustering is the definition of a similarity measure between vertices. After a measure is chosen, one computes the similarity for each pair of vertices, no matter if they are connected or not. At the end of this process, one is left with a new $n \times n$ matrix X, the similarity matrix. Initially, there are n groups, each containing one of the vertices. At each step, the two most similar groups are merged; the procedure continues until all vertices are in the same group.

There are different ways to define the similarity between groups out of the matrix X. In *single linkage clustering*, the similarity between two groups is the minimum element x_{ij}, with i in one group and j in the other. On the contrary, the maximum element x_{ij} for vertices of different groups is used in the procedure of *complete linkage clustering*. In *average linkage clustering* one has to compute the average of the x_{ij}.

The procedure can be better illustrated by means of dendrograms, like the one in Fig. 3. One should note that hierarchical clustering does not deliver a single partition, but a set of partitions.

There are many possible ways to define a similarity measure for the vertices based on the topology of the network. A possibility is to define a distance between vertices, like

$$x_{ij} = \sqrt{\sum_{k \neq i,j} (A_{ik} - A_{jk})^2} \,. \tag{5}$$

This is a dissimilarity measure, based on the concept of structural equivalence. Two vertices are structurally equivalent if they have the same neighbors, even if they are not adjacent themselves. If i and j are structurally equivalent, $x_{ij} = 0$. Vertices with large degree and different neighbors are considered very "far" from each other. Another measure related to structural equivalence is the Pearson correlation between columns or rows of the adjacency matrix,

$$x_{ij} = \frac{\sum_k (A_{ik} - \mu_i)(A_{jk} - \mu_j)}{n\sigma_i\sigma_j} \,, \tag{6}$$

where the averages $\mu_i = (\sum_j A_{ij})/n$ and the variances $\sigma_i = \sum_j (A_{ij} - \mu_i)^2/n$.

An alternative measure is the number of edge- (or vertex-) independent paths between two vertices. Independent paths do not share any edge (vertex), and their number is related to the maximum flow that can be conveyed between the two vertices under the constraint that each edge can carry only one unit of flow (max-flow/min-cut theorem). Similarly, one could consider all paths running between two vertices. In this case, there is the problem that the total number of paths is infinite, but this can be avoided if one performs a weighted sum of the number of paths, where paths of length l are weighted by the factor α^l, with $\alpha < 1$. So, the weights of long paths are exponentially suppressed and the sum converges.

Hierarchical clustering has the advantage that it does not require a preliminary knowledge on the number and size of the clusters. However, it does not provide a way to discriminate between the many partitions obtained by the procedure, and to choose that or those that better represent the community structure of the graph. Moreover, the results of the method depend on the specific similarity measure adopted. Finally, it does not correctly classify all vertices of a community, and in many cases some vertices are missed even if they have a central role in their clusters [22].

Another popular clustering technique in sociology is k-means clustering [37]. Here, the number of clusters is preassigned, say k. The vertices of the graph are embedded in a metric space, so that each vertex is a point and a distance measure is defined between pairs of points in the space. The distance is a measure of dissimilarity between vertices. The aim of the algorithm is to identify k points in this space, or *centroids*, so that each vertex is associated to one centroid and the sum of the distances of all vertices from their respective centroids is minimal. To achieve this, one starts from an initial distribution of centroids such that they are as far as possible from each other. In the first iteration, each vertex is assigned to the nearest centroid. Next, the centers of mass of the k clusters are estimated and become a new set of centroids, which allows for a new classification of the vertices, and so on. After a sufficient number of iterations, the positions of the centroids are stable, and the clusters do not change any more. The solution found is not necessarily optimal, as it strongly depends on the initial choice of the centroids. The result can be improved by performing more runs starting from different initial conditions.

The limitation of k-means clustering is the same as that of the graph partitioning algorithms: the number of clusters must be specified at the beginning, the method is not able to derive it. In addition, the embedding in a metric space can be natural for some graphs, but rather artificial for others.

New Methods

From the previous two sections it is clear that traditional approaches to derive graph partitions have serious limits. The most important problem is the need to provide the algorithms with information that one would like to derive from the algorithms themselves, like the number of clusters and their size. Even when these inputs are not necessary, like in hierarchical clustering, there is the question of estimating the goodness of the partitions, so that one can pick the best one. For these reasons, there has been a major effort in the last years to devise algorithms capable of extracting a complete information about the community structure of graphs. These methods can be grouped into different categories.

Divisive Algorithms

A simple way to identify communities in a graph is to detect the edges that connect vertices of different communities and remove them, so that the clusters get disconnected from each other. This is the philosophy of divisive algorithms. The crucial point is to find a property of intercommunity edges that could allow for their identification. Any divisive method delivers many partitions, which are by construction hierarchical, so that they can be represented with dendrograms.

Algorithm of Girvan and Newman. The most popular algorithm is that proposed by Girvan and Newman [14]. The method is also historically important, because it marked the beginning of a new era in the field of community detection. Here edges are selected according to the values of measures of *edge centrality*, estimating the importance of edges according to some property or process running on the graph. The steps of the algorithm are:

1. Computation of the centrality for all edges;
2. Removal of edge with largest centrality;
3. Recalculation of centralities on the running graph;
4. Iteration of the cycle from step 2.

Girvan and Newman focused on the concept of *betweenness*, which is a variable expressing the frequency of the participation of edges to a process. They considered three alternative definitions: edge betweenness, current-flow betweenness and random walk betweenness.

Edge betweenness is the number of shortest paths between all vertex pairs that run along the edge. It is an extension to edges of the concept of site betweenness, introduced by Freeman in 1977 [20]. It is intuitive that intercommunity edges have a large value of the edge betweenness, because many shortest paths connecting vertices of different communities will pass through them (Fig. 6). The betweenness of all edges of the graph can be calculated in a time that scales as $O(mn)$, with techniques based on breadth-first-search [26,38].

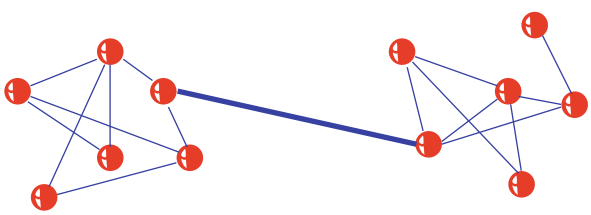

Community Structure in Graphs, Figure 6
Edge betweenness is highest for edges connecting communities. In the figure, the *thick* edge in the middle has a much higher betweenness than all other edges, because all shortest paths connecting vertices of the two communities run through it

Current-flow betweenness is defined by considering the graph a resistor network, with edges having unit resistance. If a voltage difference is applied between any two vertices, each edge carries some amount of current, that can be calculated by solving Kirchoff's equations. The procedure is repeated for all possible vertex pairs: the current-flow betweenness of an edge is the average value of the current carried by the edge. Calculation of current-flow betweenness requires the inversion of an $n \times n$ matrix (once), followed by obtaining and averaging the current for all pairs of nodes. Each of these two tasks takes a time $O(n^3)$ for a sparse matrix.

The random-walk betweenness of an edge says how frequently a random walker running on the graph goes across the edge. We remind that a random walker moving from a vertex follows each edge with equal probability. A pair of vertices is chosen at random, s and t. The walker starts at s and keeps moving until it hits t, where it stops. One computes the probability that each edge was crossed by the walker, and averages over all possible choices for the vertices s and t. The complete calculation requires a time $O(n^3)$ on a sparse graph. It is possible to show that this measure is equivalent to current-flow betweenness [39].

Calculating edge betweenness is much faster than current-flow or random walk betweenness ($O(n^2)$ versus $O(n^3)$ on sparse graphs). In addition, in practical applications the Girvan–Newman algorithm with edge betweenness gives better results than adopting the other centrality measures. Numerical studies show that the recalculation step 3 of Girvan–Newman algorithm is essential to detect meaningful communities. This introduces an additional factor m in the running time of the algorithm: consequently, the edge betweenness version scales as $O(m^2 n)$, or $O(n^3)$ on a sparse graph. Because of that, the algorithm is quite slow, and applicable to graphs with up to $n \sim 10\,000$ vertices, with current computational resources. In the original version of Girvan–Newman's algorithm [14], the authors had to deal with the whole hierarchy of partitions, as they had no procedure to say which partition is the best. In a successive refinement [26], they selected the partition with the largest value of modularity (see Sect. "Evaluating Partitions: Quality Functions"), a criterion that has been frequently used ever since. There have been countless applications of the Girvan–Newman method: the algorithm is now integrated in well known libraries of network analysis programs.

Algorithm of Tyler et al. Tyler, Wilkinson and Huberman proposed a modification of the Girvan–Newman algorithm, to improve the speed of the calculation [40].

The modification consists in calculating the contribution to edge betweenness only from a limited number of vertex pairs, chosen at random, deriving a sort of Monte Carlo estimate. The procedure induces statistical errors in the values of the edge betweenness. As a consequence, the partitions are in general different for different choices of the sampling pairs of vertices. However, the authors showed that, by repeating the calculation many times, the method gives good results, with a substantial gain of computer time. In practical examples, only vertices lying at the boundary between communities may not be clearly classified, and be assigned sometimes to a group, sometimes to another. The method has been applied to a network of people corresponding through email [40] and to networks of gene co-occurrences [41].

Algorithm of Fortunato et al. An alternative measure of centrality for edges is information centrality. It is based on the concept of efficiency [42], which estimates how easily information travels on a graph according to the length of shortest paths between vertices. The information centrality of an edge is the variation of the efficiency of the graph if the edge is removed. In the algorithm by Fortunato, Latora and Marchiori [43], edges are removed according to decreasing values of information centrality. The method is analogous to that of Girvan and Newman, but slower, as it scales as $O(n^4)$ on a sparse graph. On the other hand, it gives a better classification of vertices when communities are fuzzy, i. e. with a high degree of interconnectedness.

Algorithm of Radicchi et al. Because of the high density of edges within communities, it is easy to find loops in them, i. e. closed non-intersecting paths. On the contrary, edges lying between communities will hardly be part of short loops. Based on this intuitive idea, Radicchi et al. proposed a new measure, the edge clustering coefficient, such that low values of the measure are likely to correspond to intercommunity edges [25]. The edge clustering coefficient generalizes to edges the notion of clustering coefficient introduced by Watts and Strogatz for vertices [44]. The latter is the number of triangles including a vertex divided by the number of possible triangles that can be formed. The edge clustering coefficient is the number of loops of length g including the edge divided by the number of possible cycles. Usually, loops of length $g = 3$ or 4 are considered. At each iteration, the edge with smallest clustering coefficient is removed, the measure is recalculated again, and so on. The procedure stops when all clusters obtained are LS-sets or "weak" communities (see Sect. "Definition of Community"). Since the edge clustering coefficient is a local measure, involving at most an extended neighborhood of the edge, it can be calculated very quickly. The running time of the algorithm to completion is $O(m^4/n^2)$, or $O(n^2)$ on a sparse graph, so it is much shorter than the running time of the Girvan–Newman method. On the other hand, the method may give poor results when the graph has few loops, as it happens in several non-social networks. In this case, in fact, the edge clustering coefficient is small and fairly similar for all edges, and the algorithm may fail to identify the bridges between communities.

Modularity Optimization

If Newman–Girvan modularity Q (Sect. "Evaluating Partitions: Quality Functions") is a good indicator of the quality of partitions, the partition corresponding to its maximum value on a given graph should be the best, or at least a very good one. This is the main motivation for modularity maximization, perhaps the most popular class of methods to detect communities in graphs. An exhaustive optimization of Q is impossible, due to the huge number of ways in which it is possible to partition a graph, even when the latter is small. Besides, the true maximum is out of reach, as it has been recently proved that modularity optimization is an NP-hard problem [45], so it is probably impossible to find the solution in a time growing polynomially with the size of the graph. However, there are currently several algorithms able to find fairly good approximations of the modularity maximum in a reasonable time.

Greedy techniques. The first algorithm devised to maximize modularity was a greedy method of Newman [46]. It is an agglomerative method, where groups of vertices are successively joined to form larger communities such that modularity increases after the merging. One starts from n clusters, each containing a single vertex. Edges are not initially present, they are added one by one during the procedure. However, modularity is always calculated from the full topology of the graph, since one wants to find its partitions. Adding a first edge to the set of disconnected vertices reduces the number of groups from n to $n - 1$, so it delivers a new partition of the graph. The edge is chosen such that this partition gives the maximum increase of modularity with respect to the previous configuration. All other edges are added based on the same principle. If the insertion of an edge does not change the partition, i. e. the clusters are the same, modularity stays the same. The number of partitions found during the procedure is n, each with a different number of clusters, from n to 1. The largest value of modularity in this subset of partitions is the approximation of the modularity maximum given by the algorithm. The update of the modularity value at each iteration step can be performed in a time $O(n + m)$, so the algorithm runs to completion in a time $O((m + n)n)$,

or $O(n^2)$ on a sparse graph, which is fast. In a later paper by Clauset et al. [47], it was shown that the calculation of modularity during the procedure can be performed much more quickly by use of max-heaps, special data structures created using a binary tree. By doing that, the algorithm scales as $O(md \log n)$, where d is the depth of the dendrogram describing the successive partitions found during the execution of the algorithm, which grows as $\log n$ for graphs with a strong hierarchical structure. For those graphs, the running time of the method is then $O(n \log^2 n)$, which allows to analyze the community structure of very large graphs, up to 10^7 vertices. The greedy algorithm is currently the only algorithm that can be used to estimate the modularity maximum on such large graphs. On the other hand, the approximation it finds is not that good, as compared with other techniques. The accuracy of the algorithm can be considerably improved if one accounts for the size of the groups to be merged [48], or if the hierarchical agglomeration is started from some good intermediate configuration, rather than from the individual vertices [49].

Simulated annealing. Simulated annealing [50] is a probabilistic procedure for global optimization used in different fields and problems. It consists in performing an exploration of the space of possible states, looking for the global optimum of a function F, say its maximum. Transitions from one state to another occur with probability 1 if F increases after the change, otherwise with a probability $\exp(\beta \Delta F)$, where ΔF is the decrease of the function and β is an index of stochastic noise, a sort of inverse temperature, which increases after each iteration. The noise reduces the risk that the system gets trapped in local optima. At some stage, the system converges to a stable state, which can be an arbitrarily good approximation of the maximum of F, depending on how many states were explored and how slowly β is varied. Simulated annealing was first employed for modularity optimization by R. Guimerá et al. [31]. Its standard implementation combines two types of "moves": local moves, where a single vertex is shifted from one cluster to another, taken at random; global moves, consisting of merges and splits of communities. In practical applications, one typically combines n^2 local moves with n global ones in one iteration. The method can potentially come very close to the true modularity maximum, but it is slow. Therefore, it can be used for small graphs, with up to about 10^4 vertices. Applications include studies of potential energy landscapes [51] and of metabolic networks [12].

Extremal optimization. Extremal optimization is a heuristic search procedure proposed by Boettcher and Percus [52], in order to achieve an accuracy comparable with simulated annealing, but with a substantial gain in computer time. It is based on the optimization of local variables, expressing the contribution of each unit of the system to the global function at study. This technique was used for modularity optimization by Duch and Arenas [53]. Modularity can be indeed written as a sum over the vertices: the local modularity of a vertex is the value of the corresponding term in this sum. A fitness measure for each vertex is obtained by dividing the local modularity of the vertex by its degree. One starts from a random partition of the graph in two groups. At each iteration, the vertex with the lowest fitness is shifted to the other cluster. The move changes the partition, so the local fitnesses need to be recalculated. The process continues until the global modularity Q cannot be improved any more by the procedure. At this stage, each cluster is considered as a graph on its own and the procedure is repeated, as long as Q increases for the partitions found. The algorithm finds an excellent approximation of the modularity maximum in a time $O(n^2 \log n)$, so it represents a good tradeoff between accuracy and speed.

Spectral optimization. Modularity can be optimized using the eigenvalues and eigenvectors of a special matrix, the modularity matrix B, whose elements are

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m} , \tag{7}$$

where the notation is the same used in Eq. (1). The method [54,55] is analogous to spectral bisection, described in Sect. "Computer Science: Graph Partitioning". The difference is that here the Laplacian matrix is replaced by the modularity matrix. Between Q and B there is the same relation as between R and L in Eq. (3), so modularity can be written as a weighted sum of the eigenvalues of B, just like Eq. (4). Here one has to look for the eigenvector of B with largest eigenvalue, $\mathbf{u}_1$, and group the vertices according to the signs of the components of $\mathbf{u}_1$, just like in Sect. "Computer Science: Graph Partitioning". The Kernighan–Lin algorithm can then be used to improve the result. The procedure is repeated for each of the clusters separately, and the number of communities increases as long as modularity does. The advantage over spectral bisection is that it is not necessary to specify the size of the two groups, because it is determined by taking the partition with largest modularity. The drawback is similar as for spectral bisection, i. e. the algorithm gives the best results for bisections, whereas it is less accurate when the number of communities is larger than two. The situation could be improved by using the other eigenvectors with positive eigenvalues of the modularity matrix. In addition, the eigenvectors with the most negative eigenvalues

are important to detect a possible multipartite structure of the graph, as they give the most relevant contribution to the modularity minimum. The algorithm typically runs in a time $O(n^2 \log n)$ for a sparse graph, when one computes only the first eigenvector, so it is faster than extremal optimization, and slightly more accurate, especially for large graphs.

Finally, some general remarks on modularity optimization and its reliability. A large value for the modularity maximum does not necessarily mean that a graph has a community structure. Random graphs can also have partitions with large modularity values, even though clusters are not explicitly built in [31,56]. Therefore, the modularity maximum of a graph reveals its community structure only if it is appreciably larger than the modularity maximum of random graphs of the same size [57].

In addition, one assumes that the modularity maximum delivers the "best" partition of the network in communities. However, this is not always true [32]. In the definition of modularity (Eq. (2)) the graph is compared with a random version of it, that keeps the degrees of its vertices. If groups of vertices in the graphs are more tightly connected than they would be in the randomized graph, modularity optimization would consider them as parts of the same module. But if the groups have less than $\sqrt{m}$ internal edges, the expected number of edges running between them in modularity's null model is less than one, and a single interconnecting edge would cause the merging of the two groups in the optimal partition. This holds for every density of edges inside the groups, even in the limit case in which all vertices of each group are connected to each other, i. e. if the groups are cliques. In Fig. 7 a graph is made out of n_c identical cliques, with l vertices each, connected by single edges. It is intuitive to think that the modules of the best partition are the single cliques: instead, if n_c is larger than about l^2, modularity would be higher for the partition in which pairs of consecutive cliques are parts of the same module (indicated by the dashed lines in the figure). The problem holds for a wide class of possible null models [58]. Attempts have been made to solve it within the modularity framework [59,60,61].

Modifications of the measure have also been suggested. Massen and Doye proposed a slight variation of modularity's null model [51]: it is still a graph with the same degree sequence as the original, and with edges rewired at random among the vertices, but one imposes the additional constraint that there can be neither multiple edges between a pair of vertices nor edges joining a vertex with itself (self-edges). Muff, Rao and Caflisch remarked that modularity's null model implicitly assumes that each vertex could be attached to any other, whether in real cases a cluster is usually connected to few other clusters [62]. Therefore, they proposed a local version of modularity, in which the expected number of edges within a module is not calculated with respect to the full graph, but considering just a portion of it, namely the subgraph including the module and its neighboring modules.

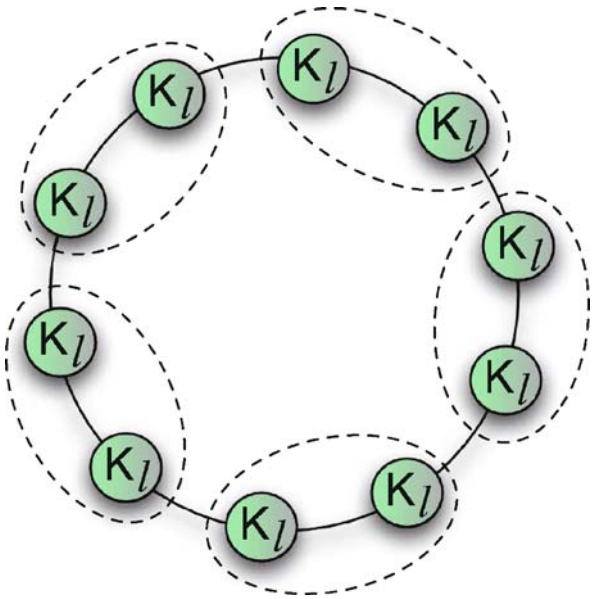

Community Structure in Graphs, Figure 7
Resolution limit of modularity optimization. The natural community structure of the graph, represented by the individual cliques (*circles*), is not recognized by optimizing modularity, if the cliques are smaller than a scale depending on the size of the graph. Reprinted figure with permission from [32]

Spectral Algorithms

As discussed above, spectral properties of graph matrices are frequently used to find partitions. Traditional methods are in general unable to predict the number and size of the clusters, which instead must be fed into the procedure. Recent algorithms, reviewed below, are more powerful.

Algorithm of Donetti and Muñoz. An elegant method based on the eigenvectors of the Laplacian matrix has been devised by Donetti and Muñoz [63]. The idea is simple: the values of the eigenvector components are close for vertices in the same community, so one can use them as coordinates to represent vertices as points in a metric space. So, if one uses M eigenvectors, one can embed the vertices in an M-dimensional space. Communities appear as groups of points well separated from each other, as illustrated in Fig. 8. The separation is the more visible, the larger the number of dimensions/eigenvectors M. The space points

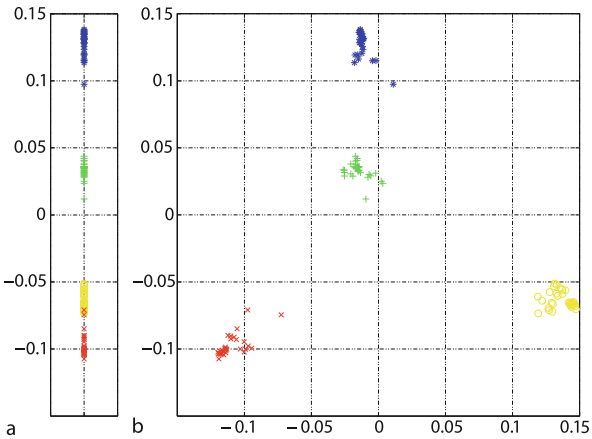

Community Structure in Graphs, Figure 8
Spectral algorithm by Donetti and Muñoz. Vertex *i* is represented by the values of the *i*th components of Laplacian eigenvectors. In this example, the graph has an adhoc division in four communities, indicated by different symbols. The communities are better separated in two dimensions (b) than in one (a). Reprinted figure with permission from [63]

are grouped in communities by hierarchical clustering (see Sect. "Social Science: Hierarchical and *k*-Means Clustering"). The final partition is the one with largest modularity. For the similarity measure between vertices, Donetti and Muñoz used both the Euclidean distance and the angle distance. The angle distance between two points is the angle between the vectors going from the origin of the M-dimensional space to either point. Applications show that the best results are obtained with complete-linkage clustering. The algorithm runs to completion in a time $O(n^3)$, which is not fast. Moreover, the number M of eigenvectors that are needed to have a clean separation of the clusters is not known a priori.

Algorithm of Capocci et al. Similarly to Donetti and Muñoz, Capocci et al. used eigenvector components to identify communities [64]. In this case the eigenvectors are those of the *normal matrix*, that is derived from the adjacency matrix by dividing each row by the sum of its elements. The eigenvectors can be quickly calculated by performing a constrained optimization of a suitable cost function. A similarity matrix is built by calculating the correlation between eigenvector components: the similarity between vertices i and j is the Pearson correlation coefficient between their corresponding eigenvector components, where the averages are taken over the set of eigenvectors used. The method can be extended to directed graphs. It is useful to estimate vertex similarities, however it does not provide a well-defined partition of the graph.

Algorithm of Wu and Huberman. A fast algorithm by Wu and Huberman identifies communities based on the properties of resistor networks [65]. It is essentially a method for bisectioning graph, similar to spectral bisection, although partitions in an arbitrary number of communities can be obtained by iterative applications. The graph is transformed into a resistor network where each edge has unit resistance. A unit potential difference is set between two randomly chosen vertices. The idea is that, if there is a clear division in two communities of the graph, there will be a visible gap between voltage values for vertices at the borders between the clusters. The voltages are calculated by solving Kirchoff's equations: an exact resolution would be too time consuming, but it is possible to find a reasonably good approximation in a linear time for a sparse graph with a clear community structure, so the more time consuming part of the algorithm is the sorting of the voltage values, which takes time $O(n \log n)$. Any possible vertex pair can be chosen to set the initial potential difference, so the procedure should be repeated for all possible vertex pairs. The authors showed that this is not necessary, and that a limited number of sampling pairs is sufficient to get good results, so the algorithm scales as $O(n \log n)$ and is very fast. An interesting feature of the method is that it can quickly find the natural community of any vertex, without determining the complete partition of the graph. For that, one uses the vertex as source voltage and places the sink at an arbitrary vertex. The same feature is present in an older algorithm by Flake et al. [11], where one uses max-flow instead of current flow.

Previous works have shown that also the eigenvectors of the *transfer matrix* $\mathbf{T}$ can be used to extract useful information on community structure [66,67]. The element T_{ij} of the transfer matrix is $1/k_j$ if i and j are neighbors, where k_j is the degree of j, otherwise it is zero. The transfer matrix rules the process of diffusion on graphs.

Dynamic Algorithms

This section describes methods employing processes running on the graph, focusing on spin-spin interactions, random walk and synchronization.

Q-state Potts model. The Potts model is among the most popular models in statistical mechanics [68]. It describes a system of spins that can be in q different states. The interaction is ferromagnetic, i. e. it favors spin alignment, so at zero temperature all spins are in the same state. If antiferromagnetic interactions are also present, the ground state of the system may not be the one where all spins are aligned, but a state where different spin values coexist, in homogeneous clusters. If Potts spin variables are assigned to the vertices of a graph with com-

munity structure, and the interactions are between neighboring spins, it is likely that the topological clusters could be recovered from like-valued spin clusters of the system, as there are many more interactions inside communities than outside. Based on this idea, inspired by an earlier paper by Blatt, Wiseman and Domany [69], Reichardt and Bornholdt proposed a method to detect communities that maps the graph onto a q-Potts model with nearest-neighbors interactions [70]. The Hamiltonian of the model, i. e. its energy, is the sum of two competing terms, one favoring spin alignment, one antialignment. The relative weight of these two terms is expressed by a parameter γ, which is usually set to the value of the density of edges of the graph. The goal is to find the ground state of the system, i. e. to minimize the energy. This can be done with simulated annealing [50], starting from a configuration where spins are randomly assigned to the vertices and the number of states q is very high. The procedure is quite fast and the results do not depend on q. The method also allows to identify vertices shared between communities, from the comparison of partitions corresponding to global and local energy minima. More recently, Reichardt and Bornholdt derived a general framework [28], in which detecting community structure is equivalent to finding the ground state of a q-Potts model spin glass [71]. Their previous method and modularity optimization are recovered as special cases. Overlapping communities can be discovered by comparing partitions with the same (minimal) energy, and hierarchical structure can be investigated by tuning a parameter acting on the density of edges of a reference graph without community structure.

Random walk. Using random walks to find communities comes from the idea that a random walker spends a long time inside a community due to the high density of edges and consequent number of paths that could be followed. Zhou used random walks to define a distance between pairs of vertices [72]: the distance d_{ij} between i and j is the average number of edges that a random walker has to cross to reach j starting from i. Close vertices are likely to belong to the same community. Zhou defines the "global attractor" of a vertex i to be the closest vertex to i (smallest d_{ij}), whereas the "local attractor" of i is its closest neighbor. Two types of communities are defined, according to local or global attractors: a vertex i has to be put in the same community of its attractor and of all other vertices for which i is an attractor. Communities must be minimal subgraphs, i. e. they cannot include smaller subgraphs which are communities according to the chosen criterion. Applications to real and artificial networks show that the method can find meaningful partitions. In a successive paper [73], Zhou introduced a measure of dissimilarity between vertices based on the distance defined above. The measure resembles the definition of distance based on structural equivalence of Eq. (5), where the elements of the adjacency matrix are replaced by the corresponding distances. Graph partitions are obtained with a divisive procedure that, starting from the graph as a single community, performs successive splits based on the criterion that vertices in the same cluster must be less dissimilar than a running threshold, which is decreased during the process. The hierarchy of partitions derived by the method is representative of actual community structures for several real and artificial graphs. In another work [74], Zhou and Lipowsky defined distances with biased random walkers, where the bias is due to the fact that walkers move preferentially towards vertices sharing a large number of neighbors with the starting vertex. A different distance measure between vertices based on random walks was introduced by Latapy and Pons [75]. The distance is calculated from the probabilities that the random walker moves from a vertex to another in a fixed number of steps. Vertices are then grouped into communities through hierarchical clustering. The method is quite fast, running to completion in a time $O(n^2 \log n)$ on a sparse graph.

Synchronization. Synchronization is another promising dynamic process to reveal communities in graphs. If oscillators are placed at the vertices, with initial random phases, and have nearest-neighbor interactions, oscillators in the same community synchronize first, whereas a full synchronization requires a longer time. So, if one follows the time evolution of the process, states with synchronized clusters of vertices can be quite stable and long-lived, so they can be easily recognized. This was first shown by Arenas, Díaz–Guilera and Pérez–Vicente [76]. They used Kuramoto oscillators [77], which are coupled two-dimensional vectors endowed with a proper frequency of oscillations. If the interaction coupling exceeds a threshold, the dynamics leads to synchronization. Arenas et al. showed that the time evolution of the system reveals some intermediate time scales, corresponding to topological scales of the graph, i. e. to different levels of organization of the vertices. Hierarchical community structure can be revealed in this way. Based on the same principle, Boccaletti et al. designed a community detection method based on synchronization [79]. The synchronization dynamics is a variation of Kuramoto's model, the opinion changing rate (OCR) model [80]. The evolution equations of the model are solved for decreasing values of a parameter that tunes the strength of the interaction coupling between neighboring vertices. In this way, different partitions are recovered: the partition with the largest value of modularity is chosen. The algorithm scales in a time $O(mn)$, or $O(n^2)$ on

sparse graphs, and gives good results on practical examples. However, synchronization-based algorithms may not be reliable when communities are very different in size.

Clique Percolation

In most of the approaches examined so far, communities have been characterized and discovered, directly or indirectly, by some global property of the graph, like betweenness, modularity, etc., or by some process that involves the graph as a whole, like random walks, synchronization, etc. But communities can be also interpreted as a form of local organization of the graph, so they could be defined from some property of the groups of vertices themselves, regardless of the rest of the graph. Moreover, very few of the algorithms presented so far are able to deal with the problem of overlapping communities (Sect. "Overlapping Communities"). A method that accounts both for the locality of the community definition and for the possibility of having overlapping communities is the Clique Percolation Method (CPM) by Palla et al. [13]. It is based on the concept that the internal edges of community are likely to form cliques due to their high density. On the other hand, it is unlikely that intercommunity edges form cliques: this idea was already used in the divisive method of Radicchi et al. (see Sect. "Divisive Algorithms"). Palla et al. define a k-clique as a complete graph with k vertices. Notice that this definition is different from the definition of n-clique (see Sect. "Definition of Community") used in social science. If it were possible for a clique to move on a graph, in some way, it would probably get trapped inside its original community, as it could not cross the bottleneck formed by the intercommunity edges. Palla et al. introduced a number of concepts to implement this idea. Two k-cliques are *adjacent* if they share $k-1$ vertices. The union of adjacent k-cliques is called *k-clique chain*. Two k-cliques are connected if they are part of a k-clique chain. Finally, a *k-clique community* is the largest connected subgraph obtained by the union of a k-clique and of all k-cliques which are connected to it. Examples of k-clique communities are shown in Fig. 9. One could say that a k-clique community is identified by making a k-clique "roll" over adjacent k-cliques, where rolling means rotating a k-clique about the $k-1$ vertices it shares with any adjacent k-clique. By construction, k-clique communities can share vertices, so they can be overlapping. There may be vertices belonging to non-adjacent k-cliques, which could be reached by different paths and end up in different clusters. In order to find k-clique communities, one searches first for maximal cliques, a task that is known to require a running time that grows exponentially with the size of the graph. However, the authors found that, for the real networks they analyzed, the procedure is quite fast, allowing to analyze graphs with up to 10^5 vertices in a reasonably short time. The actual scalability of the algorithm depends on many factors, and cannot be expressed in closed form. The algorithm has been extended to the analysis of weighted [81] and directed [82] graphs. It was recently used to study the evolution of community structure in social networks [83]. A special software, called *CFinder*, based on the CPM, has been designed by Palla and coworkers and is freely available. The CPM has the same limit as the algorithm of Radicchi et al.: It assumes that the graph has a large number of cliques, so it may fail to give meaningful partitions for graphs with just a few cliques, like technological networks.

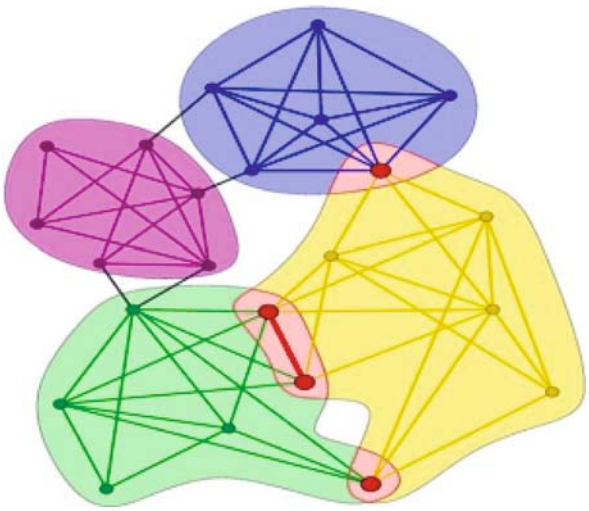

Community Structure in Graphs, Figure 9
Clique Percolation Method. The example shows communities spanned by adjacent 3-cliques (*triangles*). Overlapping vertices are shown by the *bigger* dots. Reprinted figure with permission from [13]

Other Techniques

This section describes some algorithms that do not fit in the previous categories, although some overlap is possible.

Markov Cluster Algorithm (MCL). This method, invented by van Dongen [84], simulates a peculiar process of flow diffusion in a graph. One starts from the *stochastic matrix* of the graph, which is obtained from the adjacency matrix by dividing each element A_{ij} by the degree of i. The element S_{ij} of the stochastic matrix gives the probability that a random walker, sitting at vertex i, moves to j. The sum of the elements of each column of S is one. Each it-

eration of the algorithm consists of two steps. In the first step, called expansion, the stochastic matrix of the graph is raised to an integer power p (usually $p = 2$). The entry M_{ij} of the resulting matrix gives the probability that a random walker, starting from vertex i, reaches j in p steps (diffusion flow). The second step, which has no physical counterpart, consists in raising each single entry of the matrix M to some power α, where α is now real-valued. This operation, called inflation, enhances the weights between pairs of vertices with large values of the diffusion flow, which are likely to be in the same community. Next, the elements of each row must be divided by their sum, such that the sum of the elements of the row equals one and a new stochastic matrix is recovered. After some iterations, the process delivers a stable matrix, with some remarkable properties. Its elements are either zero or one, so it is a sort of adjacency matrix. Most importantly, the graph described by the matrix is disconnected, and its connected components are the communities of the original graph. The method is really simple to implement, which is the main reason of its success: as of now, the MCL is one of the most used clustering algorithms in bioinformatics. Due to the matrix multiplication of the expansion step, the algorithm should scale as $O(n^3)$, even if the graph is sparse, as the running matrix becomes quickly dense after a few steps of the algorithm. However, while computing the matrix multiplication, MCL keeps only a maximum number k of non-zero elements per column, where k is usually much smaller than n. So, the actual worst-case running time of the algorithm is $O(nk^2)$ on a sparse graph. A problem of the method is the fact that the final partition is sensitive to the parameter α used in the inflation step. Therefore several partitions can be obtained, and it is not clear which are the most meaningful or representative.

Maximum likelihood. Newman and Leicht have recently proposed an algorithm based on traditional tools and techniques of statistical inference [85]. The method consists in deducing the group structure of the graph by checking which possible partition better "fits" the graph topology. The goodness of the fit is measured by the likelihood that the observed graph structure was generated by the particular set of relationships between vertices that define a partition. The latter is described by two sets of model parameters, expressing the size of the clusters and the connection preferences among the vertices, i. e. the probabilities that vertices of one cluster are linked to any vertex. The partition corresponding to the maximum likelihood is obtained by iterating a set of coupled equations for the variables, starting from a suitable set of initial conditions. Convergence is fast, so the algorithm could be applied to fairly large graphs, with up to about 10^6 vertices. A nice feature of the method is that it discovers more general types of vertex classes than communities. For instance, multipartite structure could be uncovered, or mixed patterns where multipartite subgraphs coexist with communities, etc. In this respect, it is more powerful than most methods of community detection, which are bound to focus only on proper communities, i. e. subgraphs with more internal than external edges. In addition, since partitions are defined by assigning probability values to the vertices, expressing the extent of their membership in a group, it is possible that some vertices are not clearly assigned to a group, but to more groups, so the method is able to deal with overlapping communities. The main drawback of the algorithm is the fact that one needs to specify the number of groups at the beginning of the calculation, a number that is often unknown for real networks. It is possible to derive this information self-consistently by maximizing the probability that the data are reproduced by partitions with a given number of clusters. But this procedure involves some degree of approximation, and the results are often not good.

L-shell method. This is an agglomerative method designed by Bagrow and Bollt [86]. The algorithm finds the community of any vertex, although the authors also presented a more general procedure to identify the full community structure of the graph. Communities are defined locally, based on a simple criterion involving the number of edges inside and outside a group of vertices. One starts from a vertex-origin and keeps adding vertices lying on successive shells, where a shell is defined as a set of vertices at a fixed geodesic distance from the origin. The first shell includes the nearest neighbors of the origin, the second the next-to-nearest neighbors, and so on. At each iteration, one calculates the number of edges connecting vertices of the new layer to vertices inside and outside the running cluster. If the ratio of these two numbers ("emerging degree") exceeds some predefined threshold, the vertices of the new shell are added to the cluster, otherwise the process stops. Because of the local nature of the process, the algorithm is very fast and can identify communities very quickly. By repeating the process starting from every vertex, one could derive a *membership matrix* M: the element M_{ij} is one if vertex j belongs to the community of vertex i, otherwise it is zero. The membership matrix can be rewritten by suitably permutating rows and columns based on their mutual distances. The distance between two rows (or columns) is defined as the number of entries whose elements differ. If the graph has a clear community structure, the membership matrix takes a block-diagonal form, where the blocks identify the communities. Unfortunately, the rearrangement of the matrix requires a time $O(n^3)$, so

it is quite slow. In a different algorithm, local communities are discovered through greedy maximization of a local modularity measure [87].

Algorithm of Eckmann and Moses. This is another method where communities are defined based on a local criterion [88]. The idea is to use the clustering coefficient [44] of a vertex as a quantity to distinguish tightly connected groups of vertices. Many edges mean many loops inside a community, so the vertices of a community are likely to have a large clustering coefficient. The latter can be related to the average distance between pairs of neighbors of the vertex. The possible values of the distance are 1 (if neighbors are connected) or 2 (if they are not), so the average distance lies between 1 and 2. The more triangles there are in the subgraph, the shorter the average distance. Since each vertex has always distance 1 from its neighbors, the fact that the average distance between its neighbors is different from 1 reminds what happens when one measures segments on a curved surface. Endowed with a metric, represented by the geodesic distance between vertices/points, and a curvature, the graph can be embedded in a geometric space. Communities appear as portions of the graph with a large curvature. The algorithm was applied to the graph representation of the World Wide Web, where vertices are Web pages and edges are the hyperlinks that take users from a page to the other. The authors found that communities correspond to Web pages dealing with the same topic.

Algorithm of Sales–Pardo et al. This is an algorithm designed to detect hierarchical community structure (see Sect. "Hierarchies"), a realistic feature of many natural, social and technological networks, that most algorithms usually neglect. The authors [89] introduce first a similarity measure between pairs of vertices based on Newman–Girvan modularity: basically the similarity between two vertices is the frequency with which they coexist in the same community in partitions corresponding to local optima of modularity. The latter are configurations for which modularity is stable, i. e. it cannot increase if one shifts one vertex from one cluster to another or by merging or splitting clusters. Next, the similarity matrix is put in block-diagonal form, by minimizing a cost function expressing the average distance of connected vertices from the diagonal. The blocks correspond to the communities and the recovered partition represents the largest scale organization level. To determine levels at lower scales, one iterates the procedure for each subgraph identified at the previous level, which is considered as an independent graph. The method yields then a hierarchy by construction, as communities at each level are nested within communities at higher levels. The algorithm is not fast, as both the search of local optima for modularity and the rearrangement of the similarity matrix are performed with simulated annealing, but delivers good results for computer generated networks, and meaningful partitions for some social, technological and biological networks.

Algorithm by Rosvall and Bergstrom. The modular structure can be considered as a reduced description of a graph to approximate the whole information contained in its adjacency matrix. Based on this idea, Rosvall and Bergstrom [90] envisioned a communication process in which a partition of a network in communities represents a synthesis Y of the full structure that a signaler sends to a receiver, who tries to infer the original graph topology X from it. The best partition corresponds to the signal Y that contains the most information about X. This can be quantitatively assessed by the maximization of the mutual information $I(X;Y)$ [91]. The method is better than modularity optimization, especially when communities are of different size. The optimization of the mutual information is performed by simulated annealing, so the method is rather slow and can be applied to graphs with up to about 10^4 vertices.

Testing Methods

When a community detection algorithm is designed, it is necessary to test its performance, and compare it with other methods. Ideally, one would like to have graphs with known community structure and check whether the algorithm is able to find it, or how closely can come to it. In any case, one needs to compare partitions found by the method with "real" partitions. How can different partitions of the same graph be compared? Danon et al. [92] used a measure borrowed from information theory, the *normalized mutual information*. One builds a *confusion matrix* N, whose element N_{ij} is the number of vertices of the real community i that are also in the detected community j. Since the partitions to be compared may have different numbers of clusters, N is usually not a square matrix. The similarity of two partitions A and B is given by the following expression

$$I(A,B) = \frac{-2\sum_{i=1}^{c_A}\sum_{j=1}^{c_B} N_{ij}\log(N_{ij}N/N_{i.}N_{.j})}{\sum_{i=1}^{c_A} N_{i.}\log(N_{i.}/N) + \sum_{j=1}^{c_B} N_{.j}\log(N_{.j}/N)}, \tag{8}$$

where $c_B(c_A)$ is the number of communities in partition $A(B)$, $N_{i.}$ is the sum of the elements of N on row i and $N_{.j}$ is the sum of the elements of N on column j. Another useful measure of similarity between partitions is the *Jaccard index*, which is regularly used in scientometric research.

Given two partitions A and B, the Jaccard index is defined as

$$I_J(A,B) = \frac{n_{11}}{n_{11} + n_{01} + n_{10}}\,, \tag{9}$$

where n_{11} is the number of pairs of vertices which are in the same community in both partitions and $n_{01}(n_{10})$ denotes the number of pairs of elements which are put in the same community in $A(B)$ and in different communities in $B(A)$. A nice presentation of criteria to compare partitions can be found in [93].

In the literature on community detection, algorithms have been generally tested on two types of graphs: computer generated graphs and real networks. The most famous computer generated benchmark is a class of graphs designed by Girvan and Newman [14]. Each graph consists of 128 vertices, arranged in four groups with 32 vertices each: 1–32, 33–64, 65–96 and 97–128. The average degree of each vertex is set to 16. The density of edges inside the groups is tuned by a parameter z_{in}, expressing the average number of edges shared by each vertex of a group with the other members (internal degree). Naturally, when z_{in} is close to 16, there is a clear community structure (see Fig. 10a), as most edges will join vertices of the same community, whereas when $z_{\text{in}} \leq 8$ there are more edges connecting vertices of different communities and the graph looks fuzzy (see Fig. 10c). In this way, one can realize different degrees of mixing between the groups. In this case the test consists in calculating the similarity between the partitions determined by the method at study and the natural partition of the graph in the four equal-sized groups. The similarity can be calculated by using the measure of Eq. (8), but in the literature one used a different quantity, i. e. the fraction of correctly classified vertices. A vertex is correctly classified if it is in the same cluster with at least 16 of its "natural" partners. If the model partition has clusters given by the merging of two or more natural groups, all vertices of the cluster are considered incorrectly classified. The number of correctly classified vertices is then divided by the total size of the graph, to yield a number between 0 and 1. One usually builds many realizations of the graph for a particular value of z_{in} and computes the average fraction of correctly classified vertices, which is a measure of the sensitivity of the method. The procedure is then iterated for different values of z_{in}. Many different algorithms have been compared with each other according to the diagram where the fraction of correctly classified vertices is plotted against z_{in}. Most algorithms usually do a good job for large z_{in} and start to fail when z_{in} approaches 8. The recipe to label vertices as correctly or incorrectly classified is somewhat arbitrary, though, and measures like those of Eq. (8) and (9) are probably more objective. There is also a subtle problem concerning the reliability of the test. Because of the randomness involved in the process of distributing edges among the vertices, it may well be that, in specific realizations of the graph, some vertices share more edges with members of another group than of their own. In this case, it is inappropriate to consider the initial partition in four groups as the real partition of the graph.

Tests on real networks usually focus on a limited number of examples, for which one has precise information about the vertices and their properties.

The most popular real network with a known community structure is the social network of Zachary's karate club (see Fig. 11). This is a social network representing the personal relationships between members of a karate club at an American university. During two years, the sociologist Wayne Zachary observed the ties between members, both inside and outside the club [94]. At some point, a conflict arose between the club's administrator (vertex 1) and one of the teachers (vertex 33), which led to the split of the club

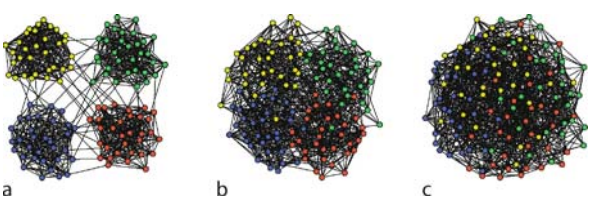

Community Structure in Graphs, Figure 10
Benchmark of Girvan and Newman. The three pictures correspond to $z_{\text{in}} = 15$ (a), $z_{\text{in}} = 11$ (b) and $z_{\text{in}} = 8$ (c). In c the four groups are basically invisible. Reprinted figure with permission from [12]

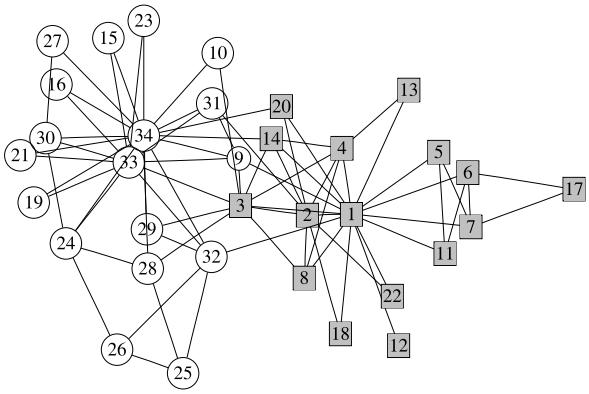

Community Structure in Graphs, Figure 11
Zachary's karate club network, an example of graph with known community structure. Reprinted figure with permission from [26]

in two smaller clubs, with some members staying with the administrator and the others following the instructor. Vertices of the two groups are highlighted by squares and circles in Fig. 11. The question is whether the actual social split could be predicted from the network topology. Several algorithms are actually able to identify the two classes, apart from a few intermediate vertices, which may be misclassified (e. g. vertices 3, 10). Other methods are less successful: for instance, the maximum of Newman–Girvan modularity corresponds to a split of the network in four groups [53,63]. It is fundamental however to stress that the comparison of community structures detected by the various methods with the split of Zachary's karate club is based on a very strong assumption: that the split actually reproduced the separation of the social network in two communities. There is no real argument, beyond common wisdom, supporting this assumption.

Two other networks have frequently been used to test community detection algorithms: the network of American college football teams derived by Girvan and Newman [14] and the social network of bottlenose dolphins constructed by Lusseau [95]. Also for these networks the caveat applies: Nothing guarantees that "reasonable" communities, defined on the basis of non-topological information, must coincide with those detected by methods based only on topology.

The Mesoscopic Description of a Graph

Community detection algorithms have been applied to a huge variety of real systems, including social, biological and technological networks. The partitions found for each system are usually similar, as the algorithms, in spite of their specific implementations, are all inspired by close intuitive notions of community. What are the general properties of these partitions? The analysis of partitions and their properties delivers a *mesoscopic description* of the graph, where the communities, and not the vertices, are the elementary units of the topology. The term mesoscopic is used because the relevant scale here lies between the scale of the vertices and that of the full graph. A simple question is whether the communities of a graph are usually about of the same size or whether the community sizes have some special distribution. It turns out that the distribution of community sizes is skewed, with a tail that obeys a power law with exponents in the range between 1 and 3 [13,22,23,47]. So, there seems to be no characteristic size for a community: small communities usually coexist with large ones.

As an example, Fig. 12 shows the cumulative distribution of community sizes for a recommendation network

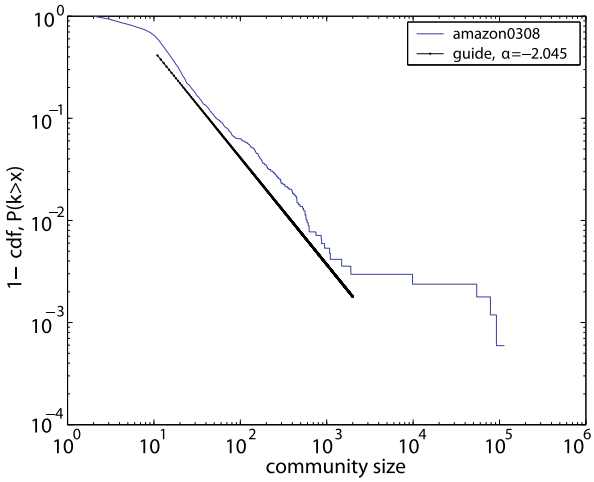

Community Structure in Graphs, Figure 12
Cumulative distribution of community sizes for the Amazon purchasing network. The partition is derived by greedy modularity optimization. Reprinted figure with permission from [47]

of the online vendor Amazon.com. Vertices are products and there is a connection between item A and B if B was frequently purchased by buyers of A. We remind that the cumulative distribution is the integral of the probability distribution: if the cumulative distribution is a power law with exponent α, the probability distribution is also a power law with exponent $\alpha + 1$.

If communities are overlapping, one could derive a network, where the communities are the vertices and pairs of vertices are connected if their corresponding communities overlap [13]. Such networks seem to have some special properties. For instance, the degree distribution is a particular function, with an initial exponential decay followed by a slower power law decay. A recent analysis has shown that such distribution can be reproduced by assuming that the graph grows according to a simple preferential attachment mechanism, where communities with large degree have an enhanced chance to interact/overlap with new communities [21].

Finally, by knowing the community structure of a graph, it is possible to classify vertices according to their roles within their community, which may allow to infer individual properties of the vertices. A nice classification has been proposed by Guimerá and Amaral [12,96]. The role of a vertex depends on the values of two indices, the z-score and the participation ratio, that determine the position of the vertex within its own module and with respect to the other modules. The z-score compares the internal degree of the vertex in its module with the average internal degree of the vertices in the module. The par-

ticipation ratio says how the edges of the vertex are distributed among the modules. Based on these two indices, Guimerá and Amaral distinguish seven roles for a vertex. These roles seem to be correlated to functions of vertices: in metabolic networks, for instance, vertices sharing many edges with vertices of other modules ("connectors") are often metabolites which are more conserved across species than other metabolites, i. e. they have an evolutionary advantage [12].

Future Directions

The problem of community detection is truly interdisciplinary. It involves scientists of different disciplines both in the design of algorithms and in their applications. The past years have witnessed huge progresses and novelties in this topic. Many methods have been developed, based on various principles. Their scalability has improved by at least one power in the graph size in just a couple of years. Currently partitions in graphs with up to millions of vertices can be found. From this point of view, the limit is close, and future improvements in this sense are unlikely. Algorithms running in linear time are very quick, but their results are often not very good.

The major breakthrough introduced by the new methods is the possibility of extracting graph partitions with no preliminary knowledge or inputs about the community structure of the graph. Most new algorithms do not need to know how many communities there are, a major drawback of computer science approaches: they derive this information from the graph topology itself. Similarly, algorithms of new generation are able to select one or a few meaningful partitions, whereas social science approaches usually produce a whole hierarchy of partitions, which they are unable to discriminate. Especially in the last two years, the quality of the output produced by some algorithms has considerably improved. Realistic aspects of community structure, like overlapping and hierarchical communities, are now often taken into account.

The main question is: is there at present a good method to detect communities in graphs? The answer depends on what is meant by "good". Several algorithms give satisfactory results when they are tested as described in Sect. "Testing Methods": in this respect, they can be considered good. However, if examined in more detail, some methods disclose serious limits and biases. For instance, the most popular method used nowadays, modularity optimization, is likely to give problems in the analysis of large graphs. Most algorithms are likely to fail in some limit, still one can derive useful indications from them: from the comparison of partitions derived by different methods one could extract the cores of real communities. The ideal method is one that delivers meaningful partitions and handles overlapping communities and hierarchy, possibly in a short time. No such method exists yet.

Finding a good method for community detection is a crucial endeavor in biology, sociology and computer science. In particular, biologists often rely on the application of clustering techniques to classify their data. Due to the bioinformatics revolution, gene regulatory networks, protein–protein interaction networks, metabolic networks, etc., are now much better known that they used to be in the past and finally susceptible to solid quantitative investigations. Uncovering their modular structure is an open challenge and a necessary step to discover properties of elementary biological constituents and to understand how biological systems work.

Bibliography

Primary Literature

1. Euler L (1736) Solutio problematis ad geometriam situs pertinentis. Commentarii Academiae Petropolitanae 8:128–140
2. Bollobás B (1998) Modern Graph Theory. Springer, New York
3. Wasserman S, Faust K (1994) Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge
4. Scott JP (2000) Social Network Analysis. Sage Publications Ltd, London
5. Barabási AL, Albert R (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
6. Dorogovtsev SN, Mendes JFF (2003) Evolution of Networks: from biological nets to the Internet and WWW. Oxford University Press, Oxford
7. Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45:167–256
8. Pastor-Satorras R, Vespignani A (2004) Evolution and structure of the Internet: A statistical physics approach. Cambridge University Press, Cambridge
9. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex Networks: Structure and Dynamics. Phys Rep 424:175–308
10. Erdös P, Rényi A (1959) On Random Graphs. Publicationes Mathematicae Debrecen 6:290–297
11. Flake GW, Lawrence S, Lee Giles C, Coetzee FM (2002) Self-Organization and Identification of Web Communities. IEEE Comput 35(3):66–71
12. Guimerà R, Amaral LAN (2005) Functional cartography of complex metabolic networks. Nature 433:895–900
13. Palla G, Derényi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435:814–818
14. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Nat Acad Sci USA 99(12): 7821–7826
15. Lusseau D, Newman MEJ (2004) Identifying the role that animals play in their social networks. Proc R Soc Lond B 271: S477–S481

16. Pimm SL (1979) The structure of food webs. Theor Popul Biol 16:144–158
17. Krause AE, Frank KA, Mason DM, Ulanowicz RE, Taylor WW (2003) Compartments exposed in food-web structure. Nature 426:282–285
18. Granovetter M (1973) The Strength of Weak Ties. Am J Sociol 78:1360–1380
19. Burt RS (1976) Positions in Networks. Soc Force 55(1):93–122
20. Freeman LC (1977) A Set of Measures of Centrality Based on Betweenness. Sociometry 40(1):35–41
21. Pollner P, Palla G, Vicsek T (2006) Preferential attachment of communities: The same principle, but a higher level. Europhys Lett 73(3):478–484
22. Newman MEJ (2004) Detecting community structure in networks. Eur Phys J B 38:321–330
23. Danon L, Duch J, Arenas A, Díaz-Guilera A (2007) Community structure identification. In: Caldarelli G, Vespignani A (eds) Large Scale Structure and Dynamics of Complex Networks: From Information Technology to Finance and Natural Science. World Scientific, Singapore, pp 93–114
24. Bron C, Kerbosch J (1973) Finding all cliques on an undirected graph. Commun ACM 16:575–577
25. Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D (2004) Defining and identifying communities in networks. Proc Nat Acad Sci USA 101(9):2658–2663
26. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69:026113
27. Arenas A, Fernández A, Fortunato S, Gómez S (2007) Motif-based communities in complex networks. arXiv:0710.0059
28. Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. Phys Rev E 74:016110
29. Massen CP, Doye JPK (2006) Thermodynamics of community structure. arXiv:cond-mat/0610077
30. Arenas A, Duch J, Fernándes A, Gómez S (2007) Size reduction of complex networks preserving modularity. New J Phys 9(6):176–180
31. Guimerà R, Sales-Pardo M, Amaral LAN (2004) Modularity from fluctuations in random graphs and complex networks. Phys Rev E 70:025101(R)
32. Fortunato S, Barthélemy M (2007) Resolution limit in community detection. Proc Nat Acad Sci USA 104(1):36–41
33. Gfeller D, Chappelier J-C, De Los Rios P (2005) Finding instabilities in the community structure of complex networks. Phys Rev E 72:056135
34. Pothen A (1997) Graph partitioning algorithms with applications to scientific computing. In: Keyes DE, Sameh A, Venkatakrishnan V (eds) Parallel Numerical Algorithms. Kluwer Academic Press, Boston, pp 323–368
35. Kernighan BW, Lin S (1970) An efficient heuristic procedure for partitioning graphs. Bell Syst Tech J 49:291–307
36. Golub GH, Van Loan CF (1989) Matrix computations. John Hopkins University Press, Baltimore
37. JB MacQueen (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley. University of California Press, pp 281–297
38. Brandes U (2001) A faster algorithm for betweenness centrality. J Math Sociol 25(2):163–177
39. Newman MEJ (2005) A measure of betweenness centrality based on random walks. Soc Netw 27:39–54
40. Tyler JR, Wilkinson DM, Huberman BA (2003) Email as spectroscopy: automated discovery of community structure within organizations. In: Huysman M, Wenger E, Wulf V (eds) Proceeding of the First International Conference on Communities and Technologies. Kluwer Academic Press, Amsterdam
41. Wilkinson DM, Huberman BA (2004) A method for finding communities of related genes. Proc Nat Acad Sci USA 101(1): 5241–5248
42. Latora V, Marchiori M (2001) Efficient behavior of small-world networks. Phys Rev Lett 87:198701
43. Fortunato S, Latora V, Marchiori M (2004) A method to find community structures based on information centrality. Phys Rev E 70:056104
44. Watts D, Strogatz SH (1998) Collective dynamics of "small-world" networks. Nature 393:440–442
45. Brandes U, Delling D, Gaertler M, Görke R, Hoefer M, Nikoloski Z, Wagner D (2007) On finding graph clusterings with maximum modularity. In: Proceedings of the 33rd International Workshop on Graph-Theoretical Concepts in Computer Science (WG'07). Springer, Berlin
46. Newman MEJ (2004) Fast algorithm for detecting community structure in networks. Phys Rev E 69:066133
47. Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70:066111
48. Danon L, Díaz-Guilera A, Arenas A (2006) The effect of size heterogeneity on community identification in complex networks. J Stat Mech Theory Exp 11:P11010
49. Pujol JM, Béjar J, Delgado J (2006) Clustering algorithm for determining community structure in large networks. Phys Rev E 74:016107
50. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220(4598):671–680
51. Massen CP, Doye JPK (2005) Identifying communities within energy landscapes. Phys Rev E 71:046101
52. Boettcher S, Percus AG (2001) Optimization with extremal dynamics. Phys Rev Lett 86:5211–5214
53. Duch J, Arenas A (2005) Community detection in complex networks using extremal optimization. Phys Rev E 72:027104
54. Newman MEJ (2006) Modularity and community structure in networks. Proc Nat Acad Sci USA 103 (23):8577–8582
55. Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74:036104
56. Reichardt J, Bornholdt S (2007) Partitioning and modularity of graphs with arbitrary degree distribution. Phys Rev E 76:015102(R)
57. Reichardt J, Bornholdt S (2006) When are networks truly modular? Phys D 224:20–26
58. Kumpula JM, Saramäki J, Kaski K, Kertész J (2007) Limited resolution in complex network community detection with Potts model approach. Eur Phys J B 56:41–45
59. Arenas A, Fernándes A, Gómez S (2007) Multiple resolution of the modular structure of complex networks. arXiv:physics/0703218
60. Ruan J, Zhang W (2007) Identifying network communities with high resolution. arXiv:0704.3759
61. Kumpula JM, Saramäki J, Kaski K, Kertész J (2007) Limited resolution and multiresolution methods in complex network community detection. In: Kertész J, Bornholdt S, Mantegna RN (eds) Noise and Stochastics in Complex Systems and Finance. Proc SPIE 6601:660116
62. Muff S, Rao F, Caflisch A (2005) Local modularity measure for network clusterizations. Phys Rev E 72:056107

63. Donetti L, Muñoz MA (2004) Detecting network communities: a new systematic and efficient algorithm. J Stat Mech Theory Exp P10012
64. Capocci A, Servedio VDP, Caldarelli G, Colaiori F (2004) Detecting communities in large networks. Phys A 352(2–4):669–676
65. Wu F, Huberman BA (2004) Finding communities in linear time: a physics approach. Eur Phys J B 38:331–338
66. Eriksen KA, Simonsen I, Maslov S, Sneppen K (2003) Modularity and extreme edges of the Internet. Phys Rev Lett 90(14):148701
67. Simonsen I, Eriksen KA, Maslov S, Sneppen K (2004) Diffusion on complex networks: a way to probe their large-scale topological structure. Physica A 336:163–173
68. Wu FY (1982) The Potts model. Rev Mod Phys 54:235–268
69. Blatt M, Wiseman S, Domany E (1996) Superparamagnetic clustering of data. Phys Rev Lett 76(18):3251–3254
70. Reichardt J, Bornholdt S (2004) Detecting fuzzy community structure in complex networks. Phys Rev Lett 93(21):218701
71. Mezard M, Parisi G, Virasoro M (1987) Spin glass theory and beyond. World Scientific Publishing Company, Singapore
72. Zhou H (2003) Network landscape from a Brownian particle's perspective. Phys Rev E 67:041908
73. Zhou H (2003) Distance, dissimilarity index, and network community structure. Phys Rev E 67:061901
74. Zhou H, Lipowsky R (2004) Network Brownian motion: A new method to measure vertex-vertex proximity and to identify communities and subcommunities. Lect Notes Comput Sci 3038:1062–1069
75. Latapy M, Pons P 92005) Computing communities in large networks using random walks. Lect Notes Comput Sci 3733: 284–293
76. Arenas A, Díaz-Guilera A, Pérez-Vicente CJ (2006) Synchronization reveals topological scales in complex networks. Phys Rev Lett 96:114102
77. Kuramoto Y (1984) Chemical Oscillations, Waves and Turbulence. Springer, Berlin
78. Arenas A, Díaz-Guilera A (2007) Synchronization and modularity in complex networks. Eur Phys J ST 143:19–25
79. Boccaletti S, Ivanchenko M, Latora V, Pluchino A, Rapisarda A (2007) Detecting complex network modularity by dynamical clustering. Phys Rev E 76:045102(R)
80. Pluchino A, Latora V, Rapisarda A (2005) Changing opinions in a changing world: a new perspective in sociophysics. Int J Mod Phys C 16(4):505–522
81. Farkas I, Ábel D, Palla G, Vicsek T (2007) Weighted network modules. New J Phys 9:180
82. Palla G, Farkas IJ, Pollner P, Derényi I, Vicsek T (2007) Directed network modules. New J Phys 9:186
83. Palla G, Barabási A-L, Vicsek T (2007) Quantifying social groups evolution. Nature 446:664–667
84. van Dongen S (2000) Graph Clustering by Flow Simulation. Ph D thesis, University of Utrecht, The Netherlands
85. Newman MEJ, Leicht E (2007) Mixture models and exploratory analysis in networks. Proc Nat Acad Sci USA 104(23):9564–9569
86. Bagrow JP, Bollt EM (2005) Local method for detecting communities. Phys Rev E 72:046108
87. Clauset A (2005) Finding local community structure in networks. Phys Rev E 72:026132
88. Eckmann J-P, Moses E (2002) Curvature of co-links uncovers hidden thematic layers in the World Wide Web. Proc Nat Acad Sci USA 99(9):5825–5829
89. Sales-Pardo M, Guimerá R, Amaral LAN (2007) Extracting the hierarchical organization of complex systems. arXiv:0705.1679
90. Rosvall M, Bergstrom CT (2007) An information-theoretic framework for resolving community structure in complex networks. Proc Nat Acad Sci USA 104(18):7327–7331
91. Shannon CE, Weaver V (1949) The Mathematical Theory of Communication. University of Illinois Press, Champaign
92. Danon L, Díaz-Guilera A, Duch J, Arenas A (2005) Comparing community structure identification. J Stat Mech Theory Exp P09008
93. Gustafsson M, Hörnquist M, Lombardi A (2006) Comparison and validation of community structures in complex networks. Physica A 367:559–576
94. Zachary WW (1977) An information flow model for conflict and fission in small groups. J Anthr Res 33:452–473
95. Lusseau D (2003) The emergent properties of a dolphin social network. Proc R Soc Lond B 270(2):S186–188
96. Guimerá R, Amaral LAN (2005) Cartography of complex networks: modules and universal roles. J Stat Mech Theory Exp P02001

Books and Reviews

Bollobás B (2001) Random Graphs. Cambridge University Press, Cambridge

Chung FRK (1997) Spectral Graph Theory. CBMS Regional Conference Series in Mathematics 92. American Mathematical Society, Providence

Dorogovtsev SN, Mendes JFF (2003) Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford University Press, Oxford

Elsner U (1997) Graph Partitioning: a Survey. Technical Report 97-27, Technische Universität Chemnitz, Chemnitz

Comparison of Discrete and Continuous Wavelet Transforms

Palle E. T. Jorgensen[1], Myung-Sin Song[2]
[1] Department of Mathematics, The University of Iowa, Iowa City, USA
[2] Department of Mathematics and Statistics, Southern Illinois University, Edwardsville, USA

Article Outline

Glossary
Definition of the Subject
Introduction
The Discrete vs. Continuous Wavelet Algorithms
List of Names and Discoveries
History
Tools from Mathematics
A Transfer Operator
Future Directions
Literature
Acknowledgments
Bibliography

This *glossary* consists of a list of terms used inside the paper in mathematics, in probability, in engineering, and, on occasion, in physics. To clarify the seemingly confusing use of up to four different names for the same idea or concept, we have further added informal explanations spelling out the reasons behind the differences in current terminology from neighboring fields.

DISCLAIMER: This glossary has the structure of four areas. A number of terms are listed line by line, and each line is followed by explanation. Some "terms" have up to four separate (yet commonly accepted) names.

Glossary

MATHEMATICS: function (measurable), PROBABILITY: random variable, ENGINEERING: signal, PHYSICS: state Mathematically, functions may map between any two sets, say, from X to Y; but if X is a probability space (typically called Ω), it comes with a σ-algebra $\mathcal{B}$ of measurable sets, and probability measure P. Elements E in $\mathcal{B}$ are called events, and $\mathrm{P}(E)$ the probability of E. Corresponding measurable functions with values in a vector space are called random variables, a terminology which suggests a stochastic viewpoint. The function values of a random variable may represent the outcomes of an experiment, for example "throwing of a die".

Yet, function theory is widely used in engineering where functions are typically thought of as signal. In this case, X may be the real line for time, or $\mathbb{R}^d$. Engineers visualize functions as signals. A particular signal may have a stochastic component, and this feature simply introduces an extra stochastic variable into the "signal", for example noise.

Turning to physics, in our present application, the physical functions will be typically be in some L^2-space, and L^2-functions with unit norm represent quantum mechanical "states".

MATHEMATICS: sequence (incl. vector-valued), PROBABILITY: random walk, ENGINEERING: time-series, PHYSICS: measurement Mathematically, a sequence is a function defined on the integers $\mathbb{Z}$ or on subsets of $\mathbb{Z}$, for example the natural numbers $\mathbb{N}$. Hence, if time is discrete, this to the engineer represents a time series, such as a speech signal, or any measurement which depends on time. But we will also allow functions on lattices such as $\mathbb{Z}^d$.

In the case $d = 2$, we may be considering the grayscale numbers which represent exposure in a digital camera. In this case, the function (grayscale) is defined on a subset of $\mathbb{Z}^2$, and is then simply a matrix.

A random walk on $\mathbb{Z}^d$ is an assignment of a sequential and random motion as a function of time. The randomness presupposes assigned probabilities. But we will use the term "random walk" also in connection with random walks on combinatorial trees.

MATHEMATICS: nested subspaces, PROBABILITY: refinement, ENGINEERING: multiresolution, PHYSICS: scales of visual resolutions While finite or infinite families of nested subspaces are ubiquitous in mathematics, and have been popular in Hilbert space theory for generations (at least since the 1930s), this idea was revived in a different guise in 1986 by Stéphane Mallat, then an engineering graduate student. In its adaptation to wavelets, the idea is now referred to as the multiresolution method.

What made the idea especially popular in the wavelet community was that it offered a skeleton on which various discrete algorithms in applied mathematics could be attached and turned into wavelet constructions in harmonic analysis. In fact what we now call multiresolutions have come to signify a crucial link between the world of discrete wavelet algorithms, which are popular in computational mathematics and in engineering (signal/image processing, data mining, etc.) on the one side, and on the other side continuous wavelet bases in function spaces, especially in $L^2(\mathbb{R}^d)$. Further, the multiresolution idea closely mimics how fractals are analyzed with the use of finite function systems.

But in mathematics, or more precisely in operator theory, the underlying idea dates back to work of John von Neumann, Norbert Wiener, and Herman Wold, where nested and closed subspaces in Hilbert space were used extensively in an axiomatic approach to stationary processes, especially for time series. Wold proved that any (stationary) time series can be decomposed into two different parts: The first (deterministic) part can be exactly described by a linear combination of its own past, while the second part is the opposite extreme; it is *unitary*, in the language of von Neumann.

Von Neumann's version of the same theorem is a pillar in operator theory. It states that every isometry in a Hilbert space $\mathcal{H}$ is the unique sum of a shift isometry and a unitary operator, i. e., the initial Hilbert space $\mathcal{H}$ splits canonically as an orthogonal sum of two subspaces $\mathcal{H}_s$ and $\mathcal{H}_u$ in $\mathcal{H}$, one which carries the shift operator, and the other $\mathcal{H}_u$ the unitary part. The shift isometry is defined from a nested scale of closed spaces V_n, such that the intersection of these spaces is $\mathcal{H}_u$. Specifically,

$$\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \cdots \subset V_n \subset V_{n+1} \subset \cdots$$

$$\bigwedge_n V_n = \mathcal{H}_u\,, \quad \text{and} \quad \bigvee_n V_n = \mathcal{H}\,.$$

However, Stéphane Mallat was motivated instead by

the notion of scales of resolutions in the sense of optics. This, in turn, is based on a certain "artificial-intelligence" approach to vision and optics, developed earlier by David Marr at MIT, an approach which imitates the mechanism of vision in the human eye.

The connection from these developments in the 1980s back to von Neumann is this: Each of the closed subspaces V_n corresponds to a level of resolution in such a way that a larger subspace represents a finer resolution. Resolutions are relative, not absolute! In this view, the relative complement of the smaller (or coarser) subspace in larger space then represents the visual detail which is added in passing from a blurred image to a finer one, i. e., to a finer visual resolution.

This view became an instant hit in the wavelet community, as it offered a repository for the fundamental father and the mother functions, also called the scaling function φ, and the wavelet function ψ. Via a system of translation and scaling operators, these functions then generate nested subspaces, and we recover the scaling identities which initialize the appropriate algorithms. What results is now called the family of pyramid algorithms in wavelet analysis. The approach itself is called the multiresolution approach (MRA) to wavelets. And in the meantime various generalizations (GMRAs) have emerged.

In all of this, there was a second "accident" at play: As it turned out, pyramid algorithms in wavelet analysis now lend themselves via multiresolutions, or nested scales of closed subspaces, to an analysis based on frequency bands. Here we refer to bands of frequencies as they have already been used for a long time in signal processing.

One reason for the success in varied disciplines of the same geometric idea is perhaps that it is closely modeled on how we historically have represented numbers in the positional number system. Analogies to the Euclidean algorithm seem especially compelling.

MATHEMATICS: operator, PROBABILITY: process, ENGINEERING: black box, PHYSICS: observable (if self-adjoint) In linear algebra students are familiar with the distinctions between (linear) transformations T (here called "operators") and matrices. For a fixed operator $T\colon V \to W$, there is a variety of matrices, one for each choice of basis in V and in W. In many engineering applications, the transformations are not restricted to be linear, but instead represent some experiment ("black box", in Norbert Wiener's terminology), one with an input and an output, usually functions of time. The input could be an external voltage function, the black box an electric circuit, and the output the resulting voltage in the circuit. (The output is a solution to a differential equation.)

This context is somewhat different from that of quantum mechanical (QM) operators $T\colon V \to V$ where V is a Hilbert space. In QM, selfadjoint operators represent observables such as position Q and momentum P, or time and energy.

MATHEMATICS: Fourier dual pair, PROBABILITY: generating function, ENGINEERING: time/frequency, PHYSICS: *P*/*Q* The following dual pairs position Q/momentum P, and time/energy may be computed with the use of Fourier series or Fourier transforms; and in this sense they are examples of Fourier dual pairs. If for example time is discrete, then frequency may be represented by numbers in the interval $[0, 2\pi)$; or in $[0, 1)$ if we enter the number 2π into the Fourier exponential. Functions of the frequency are then periodic, so the two endpoints are identified. In the case of the interval $[0, 1)$, 0 on the left is identified with 1 on the right. So a low frequency band is an interval centered at 0, while a high frequency band is an interval centered at 1/2. Let a function W on $[0, 1)$ represent a probability assignment. Such functions W are thought of as "filters" in signal processing. We say that W is low-pass if it is 1 at 0, or if it is near 1 for frequencies near 0.

Low-pass filters pass signals with low frequencies, and block the others.

If instead some filter W is 1 at 1/2, or takes values near 1 for frequencies near 1/2, then we say that W is high-pass; it passes signals with high frequency.

MATHEMATICS: convolution, PROBABILITY: —, ENGINEERING: filter, PHYSICS: smearing Pointwise multiplication of functions of frequencies corresponds in the Fourier dual time-domain to the operation of convolution (or of Cauchy product if the time-scale is discrete.) The process of modifying a signal with a fixed convolution is called a linear filter in signal processing. The corresponding Fourier dual frequency function is then referred to as "frequency response" or the "frequency response function".

More generally, in the continuous case, since convolution tends to improve smoothness of functions, physicists call it "smearing."

MATHEMATICS: decomposition (e. g., Fourier coefficients in a Fourier expansion), PROBABILITY: —, ENGINEERING: analysis, PHYSICS: frequency components Calculating the Fourier coefficients is "analysis," and adding up the pure frequencies (i. e., summing the Fourier series) is called synthesis. But this view carries over more generally to engineering where there

are more operations involved on the two sides, e. g., breaking up a signal into its frequency bands, transforming further, and then adding up the "banded" functions in the end. If the signal out is the same as the signal in, we say that the analysis/synthesis yields perfect reconstruction.

MATHEMATICS: **integrate (e. g., inverse Fourier transform),** PROBABILITY: **reconstruct,** ENGINEERING: **synthesis,** PHYSICS: **superposition** Here the terms related to "synthesis" refer to the second half of the kind of signal-processing design outlined in the previous paragraph.

MATHEMATICS: **subspace,** PROBABILITY: —, ENGINEERING: **resolution,** PHYSICS: **(signals in a) frequency band** For a space of functions (signals), the selection of certain frequencies serves as a way of selecting special signals. When the process of scaling is introduced into optics of a digital camera, we note that a nested family of subspaces corresponds to a grading of visual resolutions.

MATHEMATICS: **Cuntz relations,** PROBABILITY: —, ENGINEERING: **perfect reconstruction from subbands,** PHYSICS: **subband decomposition**

$$\sum_{i=0}^{N-1} S_i S_i^* = \mathbf{1}\,, \quad \text{and} \quad S_i^* S_j = \delta_{i,j}\mathbf{1}\,.$$

MATHEMATICS: **inner product,** PROBABILITY: **correlation,** ENGINEERING: **transition probability,** PHYSICS: **probability of transition from one state to another** In many applications, a vector space with inner product captures perfectly the geometric and probabilistic features of the situation. This can be axiomatized in the language of Hilbert space; and the inner product is the most crucial ingredient in the familiar axiom system for Hilbert space.

MATHEMATICS: $\boldsymbol{f_{\text{out}} = Tf_{\text{in}}}$, PROBABILITY: —, ENGINEERING: **input/output,** PHYSICS: **transformation of states** Systems theory language for operators $T\colon V \to W$. Then vectors in V are input, and in the range of T output.

MATHEMATICS: **fractal,** PROBABILITY: —, ENGINEERING: —, PHYSICS: — Intuitively, think of a fractal as reflecting similarity of scales such as is seen in fern-like images that look "roughly" the same at small and at large scales. Fractals are produced from an infinite iteration of a finite set of maps, and this algorithm is perfectly suited to the kind of subdivision which is a cornerstone of the discrete wavelet algorithm. Self-similarity could refer alternately to space, and to time. And further versatility is added, in that flexibility is allowed into the definition of "similar".

MATHEMATICS: —, PROBABILITY: —, ENGINEERING: **data mining,** PHYSICS: — The problem of how to handle and make use of large volumes of data is a corollary of the digital revolution. As a result, the subject of data mining itself changes rapidly. Digitized information (data) is now easy to capture automatically and to store electronically. In science, commerce, and industry, data represent collected observations and information: In business, there are data on markets, competitors, and customers. In manufacturing, there are data for optimizing production opportunities, and for improving processes. A tremendous potential for data mining exists in medicine, genetics, and energy. But raw data are not always directly usable, as is evident by inspection. A key to advances is our ability to *extract information and knowledge* from the data (hence "data mining"), and to understand the phenomena governing data sources. Data mining is now taught in a variety of forms in engineering departments, as well as in statistics and computer science departments.

One of the structures often hidden in data sets is some degree of *scale*. The goal is to detect and identify one or more natural global and local scales in the data. Once this is done, it is often possible to detect associated similarities of scale, much like the familiar scale-similarity from multidimensional wavelets, and from fractals. Indeed, various adaptations of wavelet-like algorithms have been shown to be useful. These algorithms themselves are useful in *detecting* scale-similarities, and are applicable to other types of pattern recognition. Hence, in this context, generalized multiresolutions offer another tool for discovering structures in large data sets, such as those stored in the resources of the Internet. Because of the sheer volume of data involved, a strictly manual analysis is out of the question. Instead, sophisticated query processors based on statistical and mathematical techniques are used in generating insights and extracting conclusions from data sets.

Multiresolutions Haar's work in 1909–1910 had implicitly the key idea which got wavelet mathematics started on a roll 75 years later with Yves Meyer, Ingrid Daubechies, Stéphane Mallat, and others—namely the idea of a multiresolution. In that respect Haar was ahead of his time. See Figs. 1 and 2 for details.

$$\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots\,, \quad V_0 + W_0 = V_1$$

The word "multiresolution" suggests a connection to optics from physics. So that should have been a hint to

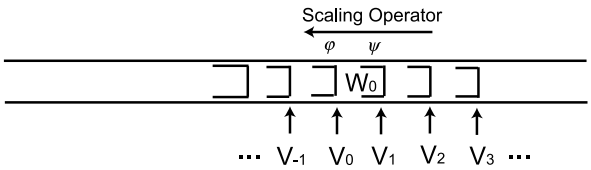

Comparison of Discrete and Continuous Wavelet Transforms, Figure 1
Multiresolution. $L^2(\mathbb{R}^d)$-version (continuous); $\varphi \in V_0$, $\psi \in W_0$

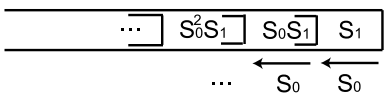

Comparison of Discrete and Continuous Wavelet Transforms, Figure 2
Multiresolution. $l^2(\mathbb{Z})$-version (discrete); $\varphi \in V_0$, $\psi \in W_0$

mathematicians to take a closer look at trends in signal and image processing! Moreover, even staying within mathematics, it turns out that as a general notion this same idea of a "multiresolution" has long roots in mathematics, even in such modern and pure areas as operator theory and Hilbert-space geometry. Looking even closer at these interconnections, we can now recognize scales of subspaces (so-called multiresolutions) in classical algorithmic construction of orthogonal bases in inner-product spaces, now taught in lots of mathematics courses under the name of the Gram–Schmidt algorithm. Indeed, a closer look at good old Gram–Schmidt reveals that it is a matrix algorithm, Hence new mathematical tools involving non-commutativity!

If the signal to be analyzed is an image, then why not select a fixed but suitable *resolution* (or a subspace of signals corresponding to a selected resolution), and then do the computations there? The selection of a fixed "resolution" is dictated by practical concerns. That idea was key in turning computation of wavelet coefficients into iterated matrix algorithms. As the matrix operations get large, the computation is carried out in a variety of paths arising from big matrix products. The dichotomy, continuous vs. discrete, is quite familiar to engineers. The industrial engineers typically work with huge volumes of numbers.

Numbers! – So why wavelets? Well, what matters to the industrial engineer is not really the wavelets, but the fact that special wavelet functions serve as an efficient way to encode large data sets – I mean encode for computations. And the wavelet algorithms are computational. They work on numbers. Encoding numbers into pictures, images, or graphs of functions comes later, perhaps at the very end of the computation. But without the graphics, I doubt that we would understand any of this half as well as we do now. The same can be said for the many issues that relate to the crucial mathematical concept of self-similarity, as we know it from fractals, and more generally from recursive algorithms.

Definition of the Subject

In this paper we outline several points of view on the interplay between discrete and continuous wavelet transforms; stressing both pure and applied aspects of both. We outline some new links between the two transform technologies based on the theory of representations of generators and relations. By this we mean a finite system of generators which are represented by operators in Hilbert space. We further outline how these representations yield sub-band filter banks for signal and image processing algorithms.

The word "wavelet transform" (WT) means different things to different people: Pure and applied mathematicians typically give different answers the questions "What is the WT?" And engineers in turn have their own preferred quite different approach to WTs. Still there are two main trends in how WTs are used, the *continuous* WT on one side, and the *discrete* WT on the other. Here we offer a user friendly outline of both, but with a slant toward geometric methods from the theory of operators in Hilbert space.

Our paper is organized as follows: For the benefit of diverse reader groups, we begin with Glossary (Sect. "Glossary"). This is a substantial part of our account, and it reflects the multiplicity of how the subject is used.

The concept of multiresolutions or multiresolution analysis (MRA) serves as a link between the discrete and continuous theory.

In Sect. "List of Names and Discoveries", we summarize how different mathematicians and scientists have contributed to and shaped the subject over the years.

The next two sections then offer a technical overview of both discrete and the continuous WTs. This includes basic tools from Fourier analysis and from operators in Hilbert space. In Sect. "Tools from Mathematics" and Sect. "A Transfer Operator", we outline the connections between the separate parts of mathematics and their applications to WTs.

Introduction

While applied problems such as time series, signals and processing of digital images come from engineering and

from the sciences, they have in the past two decades taken a life of their own as an exciting new area of applied mathematics. While searches in Google on these keywords typically yield sites numbered in the millions, the diversity of applications is wide, and it seems reasonable here to narrow our focus to some of the approaches that are both more mathematical and more recent. For references, see for example [1,6,23,31]. In addition, our own interests (e. g., [20,21,27,28]) have colored the presentation below. Each of the two areas, the discrete side, and the continuous theory is huge as measured by recent journal publications. A leading theme in our article is the independent interest in a multitude of interconnections between the discrete algorithm and their uses in the more mathematical analysis of function spaces (continuous wavelet transforms). The mathematics involved in the study and the applications of this interaction we feel is of benefit to both mathematicians and to engineers. See also [20]. An early paper [9] by Daubechies and Lagarias was especially influential in connecting the two worlds, discrete and continuous.

The Discrete vs. Continuous Wavelet Algorithms

The Discrete Wavelet Transform

If one stays with function spaces, it is then popular to pick the d-dimensional Lebesgue measure on $\mathbb{R}^d$, $d = 1, 2, \ldots$, and pass to the Hilbert space $L^2(\mathbb{R}^d)$ of all square integrable functions on $\mathbb{R}^d$, referring to d-dimensional Lebesgue measure. A wavelet basis refers to a family of basis functions for $L^2(\mathbb{R}^d)$ generated from a finite set of normalized functions ψ_i, the index i chosen from a fixed and finite index set I, and from two operations, one called scaling, and the other translation. The scaling is typically specified by a d by d matrix over the integers $\mathbb{Z}$ such that all the eigenvalues in modulus are bigger than one, lie outside the closed unit disk in the complex plane. The d-lattice is denoted $\mathbb{Z}^d$, and the translations will be by vectors selected from $\mathbb{Z}^d$. We say that we have a wavelet basis if the triple indexed family $\psi_{i,j,k}(x) := |\det A|^{j/2}\psi(A^j x + k)$ forms an orthonormal basis (ONB) for $L^2(\mathbb{R}^d)$ as i varies in I, $j \in \mathbb{Z}$, and $k \in \mathbb{R}^d$. The word "orthonormal" for a family F of vectors in a Hilbert space $\mathcal{H}$ refers to the norm and the inner product in $\mathcal{H}$: The vectors in an orthonormal family F are assumed to have norm one, and to be mutually orthogonal. If the family is also total (i. e., the vectors in F span a subspace which is dense in $\mathcal{H}$), we say that F is an orthonormal basis (ONB.)

While there are other popular wavelet bases, for example frame bases, and dual bases (see e. g., [2,18] and the papers cited there), the ONBs are the most agreeable at least from the mathematical point of view.

That there are bases of this kind is not at all clear, and the subject of wavelets in this continuous context has gained much from its connections to the discrete world of signal- and image-processing.

Here we shall outline some of these connections with an emphasis on the mathematical context. So we will be stressing the theory of Hilbert space, and bounded linear operators acting in Hilbert space $\mathcal{H}$, both individual operators, and families of operators which form algebras.

As was noticed recently the operators which specify particular subband algorithms from the discrete world of signal-processing turn out to satisfy relations that were found (or rediscovered independently) in the theory of operator algebras, and which go under the name of Cuntz algebras, denoted $\mathcal{O}_N$ if n is the number of bands. For additional details, see e. g., [21].

In symbols the C^*-algebra has generators $(S_i)_{i=0}^{N-1}$, and the relations are

$$\sum_{i=0}^{N-1} S_i S_i^* = \mathbf{1} \tag{1}$$

(where $\mathbf{1}$ is the identity element in $\mathcal{O}_N$) and

$$\sum_{i=0}^{N-1} S_i S_i^* = \mathbf{1}\,, \quad \text{and} \quad S_i^* S_j = \delta_{i,j}\mathbf{1}\,. \tag{2}$$

In a representation on a Hilbert space, say $\mathcal{H}$, the symbols S_i turn into bounded operators, also denoted S_i, and the identity element $\mathbf{1}$ turns into the identity operator I in $\mathcal{H}$, i. e., the operator $I\colon h \to h$, for $h \in \mathcal{H}$. In operator language, the two formulas (1) and (2) state that each S_i is an isometry in $\mathcal{H}$, and that the respective ranges $S_i\mathcal{H}$ are mutually orthogonal, i. e., $S_i\mathcal{H} \perp S_j\mathcal{H}$ for $i \neq j$. Introducing the projections $P_i = S_i S_i^*$, we get $P_i P_j = \delta_{i,j} P_i$, and

$$\sum_{i=0}^{N-1} P_i = I\,.$$

In the engineering literature this takes the form of programming diagrams: Fig. 3. If the process of Fig. 3 is repeated, we arrive at the discrete wavelet transform (Fig. 4) or stated in the form of images ($n = 5$) (Fig. 5).

Selecting a resolution subspace $V_0 =$ closure span $\{\varphi(\cdot - k) | k \in \mathbb{Z}\}$, we arrive at a wavelet subdivision $\{\psi_{j,k} | j \geq 0, k \in \mathbb{Z}\}$, where $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$, and the continuous expansion $f = \sum_{j,k}\langle\psi_{j,k}|f\rangle\psi_{j,k}$ or the discrete analogue derived from the isometries, $i = 1, 2, \ldots, N-1$, $S_0^k S_i$ for $k = 0, 1, 2, \ldots$; called the discrete wavelet transform.

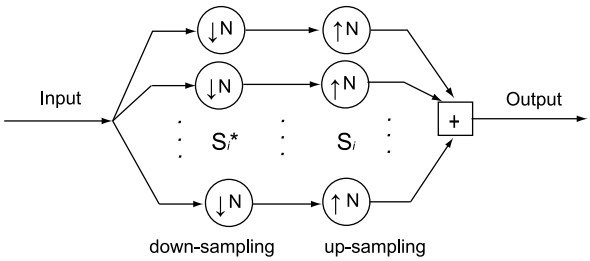

Comparison of Discrete and Continuous Wavelet Transforms, Figure 3
Perfect reconstruction in a subband filtering as used in signal- and image-processing

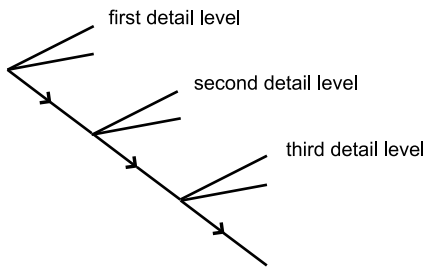

Comparison of Discrete and Continuous Wavelet Transforms, Figure 4
Binary decision tree for a sequential selection of filters.

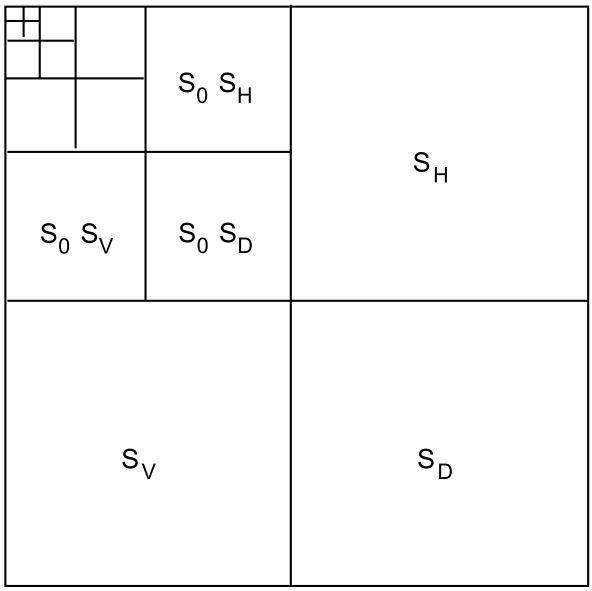

Comparison of Discrete and Continuous Wavelet Transforms, Figure 5
The subdivided squares represent the use of the pyramid subdivision algorithm to image processing, as it is used on pixel squares. At each subdivision step the *top left-hand square* represents averages of nearby pixel numbers, averages taken with respect to the chosen low-pass filter; while the three directions, horizontal, vertical, and diagonal represent detail differences, with the three represented by separate bands and filters. So in this model, there are four bands, and they may be realized by a tensor product construction applied to dyadic filters in the separate *x*- and the *y*-directions in the plane. For the discrete WT used in image-processing, we use iteration of four isometries S_0, S_H, S_V, and S_D with mutually orthogonal ranges, and satisfying the following sum-rule $S_0S_0^* + S_HS_H^* + S_VS_V^* + S_DS_D^* = I$, with I denoting the identity operator in an appropriate l^2-space

Notational Convention In algorithms, the letter N is popular, and often used for counting more than one thing.

In the present contest of the discrete wavelet algorithm (DWA) or DWT, we count two things, "the number of times a picture is decomposed via subdivision". We have used n for this. The other related but different number N is the number of subbands, $N = 2$ for the dyadic DWT, and $N = 4$ for the image DWT. The image-processing WT in our present context is the tensor product of the 1-D dyadic WT, so $2 \times 2 = 4$. Caution: Not all DWAs arise as tensor products of $N = 2$ models. The wavelets coming from tensor products are called separable. When a particular image-processing scheme is used for generating continuous wavelets it is not transparent if we are looking at a separable or inseparable wavelet!

To clarify the distinction, it is helpful to look at the representations of the Cuntz relations by operators in Hilbert space. We are dealing with representations of the two distinct algebras $\mathcal{O}_2$, and $\mathcal{O}_4$; two frequency subbands vs. four subbands. Note that the Cuntz $\mathcal{O}_2$, and $\mathcal{O}_4$ are given axiomatic, or purely symbolically. It is only when subband filters are chosen that we get representations. This also means that the choice of N is made initially; and the same N is used in different runs of the programs. In contrast, the number of times a picture is decomposed varies from one experiment to the next!

Summary: $N = 2$ for the dyadic DWT: The operators in the representation are S_0, S_1. One average operator, and one detail operator. The detail operator S_1 "counts" local detail variations.

Image-processing. Then $N = 4$ is fixed as we run different images in the DWT: The operators are now: S_0, S_H, S_V, S_D. One average operator, and three detail operator for local detail variations in the three directions in the plane.

The Continuous Wavelet Transform

Consider functions f on the real line $\mathbb{R}$. We select the Hilbert space of functions to be $L^2(\mathbb{R})$ To start a continuous WT, we must select a function $\psi \in L^2(\mathbb{R})$ and $r, s \in \mathbb{R}$

Comparison of Discrete and Continuous Wavelet Transforms, Figure 6
***n* = 2 Jorgensen. The selection of filters (Fig 4) is represented by the use of one of the operators S_i in Fig 3. A planar version of this principle is illustrated in Fig 6. For a more detailed discussion, see e. g., [3]**

such that the following family of functions

$$\psi_{r,s}(x) = r^{-1/2}\psi\left(\frac{x-s}{r}\right)$$

creates an over-complete basis for $L^2(\mathbb{R})$. An over-complete family of vectors in a Hilbert space is often called a coherent decomposition. This terminology comes from quantum optics. What is needed for a continuous WT in the simplest case is the following representation valid for all $f \in L^2(\mathbb{R})$:

$$f(x) = C_\psi^{-1} \iint_{\mathbb{R}^2} \langle \psi_{r,s} | f \rangle \psi_{r,s}(x) \frac{\mathrm{d}r\mathrm{d}s}{r^2}$$

where $C_\psi := \int_{\mathbb{R}} |\hat{\psi}(\omega)|^2 \frac{\mathrm{d}\omega}{\omega}$ and where $\langle \psi_{r,s} | f \rangle = \int_{\mathbb{R}} \overline{\psi_{r,s}(y)} f(y) \mathrm{d}y$. The refinements and implications of this are spelled out in tables in Sect. "Connections to Group Theory".

Some Background on Hilbert Space

Wavelet theory is the art of finding a special kind of basis in Hilbert space. Let $\mathcal{H}$ be a Hilbert space over $\mathbb{C}$ and denote the inner product $\langle \cdot \mid \cdot \rangle$. For us, it is assumed linear in the second variable. If $\mathcal{H} = L^2(\mathbb{R})$, then

$$\langle f \mid g \rangle := \int_{\mathbb{R}} \overline{f(x)}\, g(x)\, \mathrm{d}x \,.$$

If $\mathcal{H} = \ell^2(\mathbb{Z})$, then

$$\langle \xi \mid \eta \rangle := \sum_{n \in \mathbb{Z}} \bar{\xi}_n \eta_n \,.$$

Let $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$. If $\mathcal{H} = L^2(\mathbb{T})$, then

$$\langle f \mid g \rangle := \frac{1}{2\pi} \int_{-\pi}^{\pi} \overline{f(\theta)}\, g(\theta)\, \mathrm{d}\theta \,.$$

Functions $f \in L^2(\mathbb{T})$ have Fourier series: Setting $e_n(\theta) = \mathrm{e}^{in\theta}$,

$$\hat{f}(n) := \langle e_n \mid f \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{e}^{-in\theta} f(\theta)\, \mathrm{d}\theta \,,$$

and

$$\|f\|^2_{L^2(\mathbb{T})} = \sum_{n \in \mathbb{Z}} \left|\hat{f}(n)\right|^2 .$$

Similarly if $f \in L^2(\mathbb{R})$, then

$$\hat{f}(t) := \int_{\mathbb{R}} \mathrm{e}^{-ixt} f(x)\, \mathrm{d}x \,,$$

and

$$\|f\|^2_{L^2(\mathbb{R})} = \frac{1}{2\pi} \int_{\mathbb{R}} \left|\hat{f}(t)\right|^2 \mathrm{d}t \,.$$

Let J be an index set. We shall only need to consider the case when J is countable. Let $\{\psi_\alpha\}_{\alpha \in J}$ be a family of nonzero vectors in a Hilbert space $\mathcal{H}$. We say it is an *orthonormal basis* (ONB) if

$$\langle \psi_\alpha \mid \psi_\beta \rangle = \delta_{\alpha,\beta} \qquad \text{(Kronecker delta)} \tag{3}$$

and if

$$\sum_{\alpha \in J} |\langle \psi_\alpha \mid f \rangle|^2 = \|f\|^2 \quad \text{holds for all} f \in \mathcal{H} \,. \tag{4}$$

If only Eq. (4) is assumed, but not Eq. (3), we say that $\{\psi_\alpha\}_{\alpha \in J}$ is a (normalized) *tight frame*. We say that it is a *frame* with *frame constants* $0 < A \le B < \infty$ if

$$A\,\|f\|^2 \le \sum_{\alpha \in J} |\langle \psi_\alpha \mid f \rangle|^2 \le B\,\|f\|^2$$

holds for all $f \in \mathcal{H}$.

Introducing the rank-one operators $Q_\alpha := |\psi_\alpha\rangle\langle\psi_\alpha|$ of Dirac's terminology, see [3], we see that $\{\psi_\alpha\}_{\alpha\in J}$ is an ONB if and only if the Q_α's are projections, and

$$\sum_{\alpha\in J} Q_\alpha = I \qquad (= \text{the identity operator in } \mathcal{H})\,. \tag{5}$$

It is a (normalized) tight frame if and only if Eq. (5) holds but with no further restriction on the rank-one operators Q_α. It is a frame with frame constants A and B if the operator

$$S := \sum_{\alpha\in J} Q_\alpha$$

satisfies

$$AI \le S \le BI$$

in the order of Hermitian operators. (We say that operators $H_i = H_i^*$, $i = 1, 2$, satisfy $H_1 \le H_2$ if $\langle f \mid H_1 f\rangle \le \langle f \mid H_2 f\rangle$ holds for all $f \in \mathcal{H}$). If h, k are vectors in a Hilbert space $\mathcal{H}$, then the operator $A = |h\rangle\langle k|$ is defined by the identity $\langle u \mid Av\rangle = \langle u \mid h\rangle\langle k \mid v\rangle$ for all $u, v \in \mathcal{H}$.

Wavelets in $L^2(\mathbb{R})$ are generated by simple operations on one or more functions ψ in $L^2(\mathbb{R})$, the operations come in pairs, say scaling and translation, or phase-modulation and translations. If $N \in \{2, 3, \dots\}$ we set

$$\psi_{j,k}(x) := N^{j/2}\psi\left(N^j x - k\right) \quad \text{for } j, k \in \mathbb{Z}\,.$$

Increasing the Dimension In wavelet theory, [7] there is a tradition for reserving φ for the father function and ψ for the mother function. A 1-level wavelet transform of an $N \times M$ image can be represented as

$$\mathbf{f} \mapsto \begin{pmatrix} \mathbf{a}^1 & | & \mathbf{h}^1 \\ - & & - \\ \mathbf{v}^1 & | & \mathbf{d}^1 \end{pmatrix} \tag{6}$$

where the subimages $\mathbf{h}^1, \mathbf{d}^1, \mathbf{a}^1$ and $\mathbf{v}^1$ each have the dimension of $N/2$ by $M/2$.

$$\begin{aligned}
\mathbf{a}^1 &= V_m^1 \otimes V_n^1 : \varphi^A(x, y) = \varphi(x)\varphi(y) \\
&= \sum_i \sum_j h_i h_j \varphi(2x - i)\varphi(2y - j) \\
\mathbf{h}^1 &= V_m^1 \otimes W_n^1 : \psi^H(x, y) = \psi(x)\varphi(y) \\
&= \sum_i \sum_j g_i h_j \varphi(2x - i)\varphi(2y - j) \\
\mathbf{v}^1 &= W_m^1 \otimes V_n^1 : \psi^V(x, y) = \varphi(x)\psi(y) \\
&= \sum_i \sum_j h_i g_j \varphi(2x - i)\varphi(2y - j) \\
\mathbf{d}^1 &= W_m^1 \otimes W_n^1 : \psi^D(x, y) = \psi(x)\psi(y) \\
&= \sum_i \sum_j g_i g_j \varphi(2x - i)\varphi(2y - j)
\end{aligned} \tag{7}$$

where φ is the father function and ψ is the mother function in sense of wavelet, V space denotes the average space and the W spaces are the difference space from multiresolution analysis (MRA) [7].

In the formulas, we have the following two indexed number systems $\mathbf{a} := (h_i)$ and $\mathbf{d} := (g_i)$, $\mathbf{a}$ is for averages, and $\mathbf{d}$ is for local differences. They are really the input for the DWT. But they also are the key link between the two transforms, the discrete and continuous. The link is made up of the following scaling identities:

$$\begin{aligned}
\varphi(x) &= 2\sum_{i\in\mathbb{Z}} h_i \varphi(2x - i)\,; \\
\psi(x) &= 2\sum_{i\in\mathbb{Z}} g_i \varphi(2x - i)\,;
\end{aligned}$$

and (low-pass normalization) $\sum_{i\in\mathbb{Z}} h_i = 1$. The scalars (h_i) may be real or complex; they may be finite or infinite in number. If there are four of them, it is called the "four tap", etc. The finite case is best for computations since it corresponds to compactly supported functions. This means that the two functions φ and ψ will vanish outside some finite interval on a real line.

The two number systems are further subjected to orthogonality relations, of which

$$\sum_{i\in\mathbb{Z}} \bar{h}_i h_{i+2k} = \frac{1}{2}\delta_{0,k} \tag{8}$$

is the best known.

The systems h and g are both low-pass and high-pass filter coefficients. In Eq. (7), $\mathbf{a}^1$ denotes the first averaged image, which consists of average intensity values of the original image. Note that only φ function, V space and h coefficients are used here. Similarly, $\mathbf{h}^1$ denotes the first detail image of horizontal components, which consists of intensity difference along the vertical axis of the original image. Note that φ function is used on y and ψ function on x, W space for x values and V space for y values; and both h and g coefficients are used accordingly. The data $\mathbf{v}^1$ denote the first detail image of vertical components, which consists of intensity difference along the horizontal axis of the original image. Note that φ function is used on x and ψ function on y, W space for y values and V space

for x values; and both h and g coefficients are used accordingly. Finally, $\mathbf{d}^1$ denotes the first detail image of diagonal components, which consists of intensity difference along the diagonal axis of the original image. The original image is reconstructed from the decomposed image by taking the sum of the averaged image and the detail images and scaling by a scaling factor. It could be noted that only ψ function, W space and g coefficients are used here. See [28,33].

This decomposition not only limits to one step but it can be done again and again on the averaged detail depending on the size of the image. Once it stops at certain level, quantization (see [26,32]) is done on the image. This quantization step may be lossy or lossless. Then the lossless entropy encoding is done on the decomposed and quantized image.

The relevance of the system of identities Eq. (8) may be summarized as follows. Set

$$m_0(z) := \frac{1}{2}\sum_{k\in\mathbb{Z}} h_k z^k \quad \text{for all } z \in \mathbb{T}\,;$$

$$g_k := (-1)^k \bar{h}_{1-k} \quad \text{for all } k \in \mathbb{Z}\,;$$

$$m_1(z) := \frac{1}{2}\sum_{k\in\mathbb{Z}} g_k z^k; \quad \text{and}$$

$$(S_j f)(z) = \sqrt{2} m_j(z) f(z^2)\,,$$
$$\text{for } j = 0,1\,,\quad f \in L^2(\mathbb{T})\,,\quad z \in \mathbb{T}\,.$$

Then the following conditions are equivalent:

(a) The system of Eq. (8) is satisfied.
(b) The operators S_0 and S_1 satisfy the Cuntz relations.
(c) We have perfect reconstruction in the subband system of Fig. 4.

Note that the two operators S_0 and S_1 have equivalent matrix representations. Recalling that by Parseval's formula, we have $L^2(\mathbb{T}) \simeq l^2(\mathbb{Z})$. So representing S_0 instead as an $\infty \times \infty$ matrix acting on column vectors $x = (x_j)_{j\in\mathbb{Z}}$ we get

$$(S_0 x)_i = \sqrt{2}\sum_{j\in\mathbb{Z}} h_{i-2j} x_j$$

and for the adjoint operator $F_0 := S_0^*$, we get the matrix representation

$$(F_0 x)_i = \frac{1}{\sqrt{2}}\sum_{j\in\mathbb{Z}} \bar{h}_{j-2i} x_j$$

with the overbar signifying complex conjugation. This is computational significance to the two matrix representations, both the matrix for S_0, and for $F_0 := S_0^*$, is slanted.

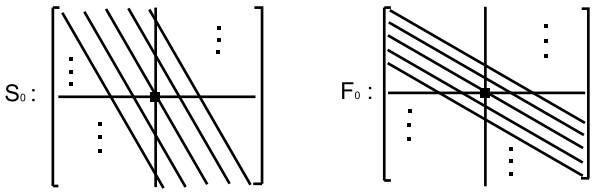

Comparison of Discrete and Continuous Wavelet Transforms, Figure 7
Matrix representation of filters operations.

However, the slanting of one is the mirror-image of the other, i. e., Fig. 7.

Significance of Slanting The slanted matrix representations refers to the corresponding operators in L^2. In general operators in Hilbert function spaces have many matrix representations, one for each orthonormal basis (ONB), but here we are concerned with the ONB consisting of the Fourier frequencies z^j, $j \in \mathbb{Z}$. So in our matrix representations for the S operators and their adjoints we will be acting on column vectors, each infinite column representing a vector in the sequence space l^2. A vector in l^2 is said to be of finite size if it has only a finite set of non-zero entries.

It is the matrix F_0 that is effective for iterated matrix computation. Reason: When a column vector x of a fixed size, say 2s is multiplied, or acted on by F_0, the result is a vector y of half the size, i. e., of size s. So $y = F_0 x$. If we use F_0 and F_1 together on x, then we get two vectors, each of size s, the other one $z = F_1 x$, and we can form the combined column vector of y and z; stacking y on top of z. In our application, y represents averages, while z represents local differences: Hence the wavelet algorithm.

$$\begin{bmatrix} \vdots \\ y_{-1} \\ y_0 \\ y_1 \\ \vdots \\ - \\ \vdots \\ z_{-1} \\ z_0 \\ z_1 \\ \vdots \end{bmatrix} = \begin{bmatrix} F_0 \\ - \\ F_1 \end{bmatrix} \begin{bmatrix} \vdots \\ x_{-2} \\ x_{-1} \\ x_0 \\ x_1 \\ x_2 \\ \vdots \end{bmatrix}$$

$$y = F_0 x$$
$$z = F_1 x$$

Comparison of Discrete and Continuous Wavelet Transforms, Table 1
Splitting of a signal into filtered components: average and a scale of detail. Several instances: Continuous vs. discrete Operator representation

$N = 2$	Overcomplete Basis	Dual Bases
Continuous resolution	$C_\psi^{-1} \iint_{\mathbb{R}^2} \frac{\mathrm{d}r\mathrm{d}s}{r^2} \lvert\psi_{r,s}\rangle\langle\psi_{r,s}\rvert = \mathbf{1}$	$C_{\psi,\tilde\psi}^{-1} \iint_{\mathbb{R}^2} \frac{\mathrm{d}r\mathrm{d}s}{r^2} \lvert\psi_{r,s}\rangle\langle\tilde\psi_{r,s}\rvert = \mathbf{1}$
Discrete resolution	$\sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}} \lvert\psi_{j,k}\rangle\langle\psi_{j,k}\rvert = \mathbf{1}$, $\psi_{j,k}$ Corresponding to $r = 2^{-j}, s = k2^{-j}$	$\sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}} \lvert\psi_{j,k}\rangle\langle\tilde\psi_{j,k}\rvert = \mathbf{1}$

$N \geq 2$	Isometries in ℓ^2	Dual Operator System in ℓ^2
Sequence spaces	$\sum_{i=0}^{N-1} S_i S_i^* = \mathbf{1}$, where $S_0, \ldots, S_{N-1}$ are adjoints to the quadrature mirror filter operators F_i, i. e., $S_i = F_i^*$	$\sum_{i=0}^{N-1} S_i \tilde S_i^* = \mathbf{1}$, for a dual operator system $S_0, \ldots, S_{N-1}$, $\tilde S_0, \ldots, \tilde S_{N-1}$

Comparison of Discrete and Continuous Wavelet Transforms, Table 2
Application of the operator representation to specific signals, contiuous and discrete

$C_\psi^{-1} \iint_{\mathbb{R}^2} \frac{\mathrm{d}r\mathrm{d}s}{r^2} \lvert\langle \psi_{r,s} \mid f\rangle\rvert^2 = \lVert f\rVert_{L^2}^2 \quad \forall f \in L^2(\mathbb{R})$	$C_{\psi,\tilde\psi}^{-1} \iint_{\mathbb{R}^2} \frac{\mathrm{d}r\mathrm{d}s}{r^2} \langle f \mid \psi_{r,s}\rangle\langle \tilde\psi_{r,s} \mid g\rangle = \langle f \mid g\rangle \quad \forall f, g \in L^2(\mathbb{R})$
$\sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}} \lvert\langle \psi_{j,k} \mid f\rangle\rvert^2 = \lVert f\rVert_{L^2}^2 \quad \forall f \in L^2(\mathbb{R})$	$\sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}} \langle f \mid \psi_{j,k}\rangle\langle \tilde\psi_{j,k} \mid g\rangle = \langle f \mid g\rangle \quad \forall f, g \in L^2(\mathbb{R})$
$\sum_{i=0}^{N-1} \lVert S_i^* c\rVert^2 = \lVert c\rVert^2 \quad \forall c \in \ell^2$	$\sum_{i=0}^{N-1} \langle S_i^* c \mid \tilde S_i^* d\rangle = \langle c \mid d\rangle \quad \forall c, d \in \ell^2$

Connections to Group Theory

The first line in the two tables below is the continuous wavelet transform. It comes from what in physics is called *coherent vector decompositions*. Both transforms applies to vectors in Hilbert space $\mathcal{H}$, and $\mathcal{H}$ may vary from case to case. Common to all transforms is vector input and output. If the input agrees with output we say that the combined process yields the identity operator image. $\mathbf{1}\colon \mathcal{H} \to \mathcal{H}$ or written $\mathbf{1}_{\mathcal{H}}$. So for example if $(S_i)_{i=0}^{N-1}$ is a finite operator system, and input/output operator example may take the form

$$\sum_{i=0}^{N-1} S_i S_i^* = \mathbf{1}_{\mathcal{H}} .$$

The Summary of and variations on the resolution of the identity operator $\mathbf{1}$ in L^2 or in ℓ^2, for ψ and $\tilde\psi$ where $\psi_{r,s}(x) = r^{-\frac{1}{2}}\psi\left(\frac{x-s}{r}\right)$,

$$C_\psi = \int_{\mathbb{R}} \frac{\mathrm{d}\omega}{\lvert\omega\rvert} \lvert\hat\psi(\omega)\rvert^2 < \infty ,$$

similarly for $\tilde\psi$ and $C_{\psi,\tilde\psi} = \int_{\mathbb{R}} \frac{\mathrm{d}\omega}{\lvert\omega\rvert} \overline{\hat\psi(\omega)}\, \hat{\tilde\psi}(\omega)$ is given in Table 1.

Then the assertions in Table 1 amount to the equations in Table 2.

A function ψ satisfying the resolution identity is called a *coherent vector* in mathematical physics. The representation theory for the $(ax+b)$-group, i. e., the matrix group $G = \left\{ \left(\begin{smallmatrix} a & b \\ 0 & 1 \end{smallmatrix}\right) \mid a \in \mathbb{R}_+,\ b \in \mathbb{R} \right\}$, serves as its underpinning. Then the tables above illustrate how the $\{\psi_{j,k}\}$ wavelet system arises from a discretization of the following unitary representation of G:

$$\left(U_{\left(\begin{smallmatrix} a & b \\ 0 & 1 \end{smallmatrix}\right)} f\right)(x) = a^{-\frac{1}{2}} f\left(\frac{x-b}{a}\right)$$

acting on $L^2(\mathbb{R})$. This unitary representation also explains the discretization step in passing from the first line to the second in the tables above. The functions $\{\psi_{j,k} \mid j, k \in \mathbb{Z}\}$ which make up a wavelet system result from the choice of a suitable coherent vector $\psi \in L^2(\mathbb{R})$, and then setting

$$\psi_{j,k}(x) = \left(U_{\left(\begin{smallmatrix} 2^{-j} & k\cdot 2^{-j} \\ 0 & 1 \end{smallmatrix}\right)} \psi\right)(x) = 2^{\frac{j}{2}} \psi\left(2^j x - k\right) .$$

Even though this representation lies at the historical origin of the subject of wavelets, the $(ax + b)$-group seems to be now largely forgotten in the next generation of the wavelet community. But Chaps. 1–3 of [7] still serve as a beautiful presentation of this (now much ignored) side of the subject. It also serves as a link to mathematical physics and to classical analysis.

List of Names and Discoveries

Many of the main discoveries summarized below are now lore.

1807 *Jean Baptiste Joseph Fourier mathematics, physics (heat conduction)*
Expressing functions as sums of sine and cosine waves of frequencies in arithmetic progression (now called Fourier series).

1909 *Alfred Haar mathematics*
Discovered, while a student of David Hilbert, an orthonormal basis consisting of step functions, applicable both to functions on an interval, and functions on the whole real line. While it was not realized at the time, Haar's construction was a precursor of what is now known as the Mallat subdivision, and multiresolution method, as well as the subdivision wavelet algorithms.

1946 *Denes Gabor (Nobel Prize): physics (optics, holography)*
Discovered basis expansions for what might now be called time-frequency wavelets, as opposed to time-scale wavelets.

1948 *Claude Elwood Shannon mathematics, engineering (information theory)*
A rigorous formula used by the phone company for sampling speech signals. Quantizing information, entropy, founder of what is now called the mathematical theory of communication.

1976 *Claude Garland, Daniel Esteban (both) signal processing*
Discovered subband coding of digital transmission of speech signals over the telephone.

1981 *Jean Morlet petroleum engineer*
Suggested the term "ondelettes." J.M. decomposed reflected seismic signals into sums of "wavelets (Fr.: ondelettes) of constant shape," i. e., a decomposition of signals into wavelet shapes, selected from a library of such shapes (now called wavelet series). Received somewhat late recognition for his work. Due to contributions by A. Grossman and Y. Meyer, Morlet's discoveries have now come to play a central role in the theory.

1985 *Yves Meyer mathematics, applications*
Mentor for A. Cohen, S. Mallat, and other of the wavelet pioneers, Y.M. discovered infinitely often differentiable wavelets.

1989 *Albert Cohen mathematics (orthogonality relations), numerical analysis*
Discovered the use of wavelet filters in the analysis of wavelets — the so-called Cohen condition for orthogonality.

1986 *Stéphane Mallat mathematics, signal and image processing*
Discovered what is now known as the subdivision, and multiresolution method, as well as the subdivision wavelet algorithms. This allowed the effective use of operators in the Hilbert space $L^2(\mathbb{R})$, and of the parallel computational use of recursive matrix algorithms.

1987 *Ingrid Daubechies mathematics, physics, and communications*
Discovered differentiable wavelets, with the number of derivatives roughly half the length of the support interval. Further found polynomial algorithmic for their construction (with coauthor Jeff Lagarias; joint spectral radius formulas).

1991 *Wayne Lawton mathematics (the wavelet transfer operator)*
Discovered the use of a transfer operator in the analysis of wavelets: orthogonality and smoothness.

1992 *The FBI using wavelet algorithms in digitizing and compressing fingerprints*
C. Brislawn and his group at Los Alamos created the theory and the codes which allowed the compression of the enormous FBI fingerprint file, creating A/D, a new database of fingerprints.

2000 *The International Standards Organization*
A wavelet-based picture compression standard, called JPEG 2000, for digital encoding of images.

1994 *David Donoho statistics, mathematics*
Pioneered the use of wavelet bases and tools from statistics to "denoise" images and signals.

History

While wavelets as they have appeared in the mathematics literature (e. g., [7]) for a long time, starting with Haar in 1909, involve function spaces, the connections to a host of discrete problems from engineering is more subtle. Moreover the deeper connections between the discrete algorithms and the function spaces of mathematical analysis are of a more recent vintage, see e. g., [31] and [21].

Here we begin with the function spaces. This part of wavelet theory refers to continuous wavelet transforms

(details below). It dominated the wavelet literature in the 1980s, and is beautifully treated in the first four chapters in [7] and in [8]. The word "continuous" refers to the continuum of the real line $\mathbb{R}$. Here we consider spaces of functions in one or more real dimensions, i. e., functions on the line $\mathbb{R}$ (signals), the plane $\mathbb{R}^2$ (images), or in higher dimensions $\mathbb{R}^d$, functions of d real variables.

Tools from Mathematics

In our presentation, we will rely on tools from at least three separate areas of mathematics, and we will outline how they interact to form a coherent theory, and how they come together to form a link between what is now called the discrete and the continuous wavelet transform. It is the discrete case that is popular with engineers ([1,23,29,30]), while the continuous case has come to play a central role in the part of mathematics referred to as harmonic analysis, [8]. The three areas are, operator algebras, dynamical systems, and basis constructions:

a. Operator algebras. The theory of operator algebras in turn breaks up in two parts: One the study of "the algebras themselves" as they emerge from the axioms of von Neumann (von Neumann algebras), and Gelfand, Kadison and Segal (C^*-algebras.) The other has a more applied slant: It involves "the representations" of the algebras. By this we refer to the following: The algebras will typically be specified by generators and by relations, and by a certain norm-completion, in any case by a system of axioms. This holds both for the norm-closed algebras, the so called C^*-algebras, and for the weakly closed algebras, the von Neumann algebras. In fact there is a close connection between the two parts of the theory: For example, representations of C^*-algebras generate von Neumann algebras.
To talk about representations of a fixed algebra say A we must specify a Hilbert space, and a homomorphism ρ from A into the algebra $\mathcal{B}(H)$ of all bounded operators on $\mathcal{H}$. We require that ρ sends the identity element in A into the identity operator acting on $\mathcal{H}$, and that $\rho(a^*) = (\rho(a))^*$ where the last star now refers to the adjoint operator.
It was realized in the last ten years (see for example [3,21,22] that a family of representations that wavelets which are basis constructions in harmonic analysis, in signal/image analysis, and in computational mathematics may be built up from representations of an especially important family of simple C^*-algebras, the Cuntz algebras. The Cuntz algebras are denoted $\mathcal{O}_2, \mathcal{O}_3, \ldots$, including $\mathcal{O}_\infty$.

b. Dynamical systems. The connection between the Cuntz algebras $\mathcal{O}_N$ for $N = 2, 3, \ldots$ are relevant to the kind of dynamical systems which are built on branching-laws, the case of $\mathcal{O}_N$ representing N-fold branching. The reason for this is that if N is fixed, $\mathcal{O}_N$ includes in its definition an iterated subdivision, but within the context of Hilbert space. For more details, see e. g., [12,13,14,15,16,17,22].

c. Analysis of bases in function spaces. The connection to basis constructions using wavelets is this: The context for wavelets is a Hilbert space $\mathcal{H}$, where $\mathcal{H}$ may be $L^2(\mathbb{R}^d)$ where d is a dimension, $d = 1$ for the line (signals), $d = 2$ for the plane (images), etc. The more successful bases in Hilbert space are the orthonormal bases ONBs, but until the mid 1980s, there were no ONBs in $L^2(\mathbb{R}^d)$ which were entirely algorithmic and effective for computations. One reason for this is that the tools that had been used for 200 years since Fourier involved basis functions (Fourier wave functions) which were not localized. Moreover these existing Fourier tools were not friendly to algorithmic computations.

A Transfer Operator

A popular tool for deciding if a candidate for a wavelet basis is in fact an ONB uses a certain transfer operator. Variants of this operator is used in diverse areas of applied mathematics. It is an operator which involves a weighted average over a finite set of possibilities. Hence it is natural for understanding random walk algorithms. As remarked in for example [12,20,21,22], it was also studied in physics, for example by David Ruelle, who used to prove results on phase transition for infinite spin systems in quantum statistical mechanics. In fact the transfer operator has many incarnations (many of them known as Ruelle operators), and all of them based on N-fold branching laws.

In our wavelet application, the Ruelle operator weights in input over the N branch possibilities, and the weighting is assigned by a chosen scalar function W. the and the W-Ruelle operator is denoted R_W. In the wavelet setting there is in addition a low-pass filter function m_0 which in its frequency response formulation is a function on the d-torus $\mathbf{T}^d = \mathbb{R}^d/\mathbb{Z}^d$.

Since the scaling matrix A has integer entries A passes to the quotient $\mathbb{R}^d/\mathbb{Z}^d$, and the induced transformation $r_A : \mathbb{T}^d \to \mathbb{T}^d$ is an N-fold cover, where $N = |\det A|$, i. e., for every x in $\mathbb{T}^d$ there are N distinct points y in $\mathbb{T}^d$ solving $r_A(y) = x$.

In the wavelet case, the weight function W is $W = |m_0|^2$. Then with this choice of W, the ONB problem for a candidate for a wavelet basis in the Hilbert space $L^2(\mathbb{R}^d)$

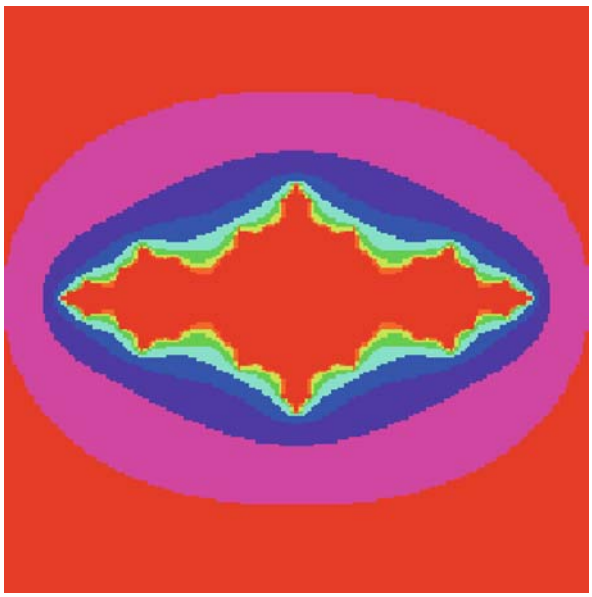

Comparison of Discrete and Continuous Wavelet Transforms, Figure 8
Julia set with $c = -1$ These images are generated by Mathematica by authors for different c values for $\varphi_c(z) = z^(2) + c$.

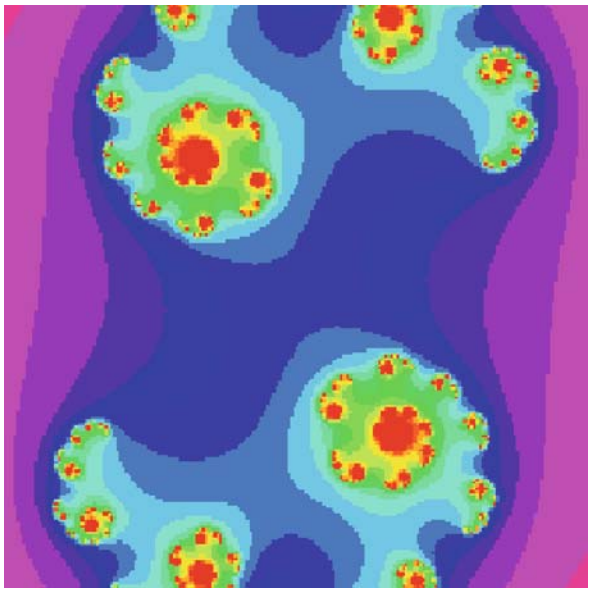

Comparison of Discrete and Continuous Wavelet Transforms, Figure 9
Julia set with $c = 0.45 - 0.1428i$ These images are generated by Mathematica by authors for different c values for $\varphi_c(z) = z^(2) + c$.

as it turns out may be decided by the dimension of a distinguished eigenspace for R_W, by the so called Perron–Frobenius problem.

This has worked well for years for the wavelets which have an especially simple algorithm, the wavelets that are initialized by a single function, called the scaling function. These are called the multiresolution analysis (MRA) wavelets, or for short the MRA-wavelets. But there are instances, for example if a problem must be localized in frequency domain, when the MRA-wavelets do not suffice, where it will by necessity include more than one scaling function. And we are then back to trying to decide if the output from the discrete algorithm, and the $\mathcal{O}_N$ representation is an ONB, or if it has some stability property which will serve the same purpose, in case where asking for an ONB is not feasible.

Future Directions

The idea of a scientific analysis by subdividing a fixed picture or object into its finer parts is not unique to wavelets. It works best for structures with an inherent self-similarity; this self-similarity can arise from numerical scaling of distances. But there are more subtle non-linear self-similarities. The Julia sets in the complex plane are a case in point [4,5,10,11,24,25]. The simplest Julia set come from a one parameter family of quadratic polynomials $\varphi_c(z) = z^2 + c$, where z is a complex variable and where c is a fixed parameter. The corresponding Julia sets J_c have a surprisingly rich structure. A simple way to understand them is the following: Consider the two branches of the inverse $\beta_\pm = z \mapsto \pm\sqrt{z - c}$. Then J_c is the unique minimal non-empty compact subset of $\mathbb{C}$, which is invariant under $\{\beta_\pm\}$. (There are alternative ways of presenting J_c but this one fits our purpose. The Julia set J of a holomorphic function, in this case $z \mapsto z^2 + c$, informally consists of those points whose long-time behavior under repeated iteration, or rather iteration of substitutions, can change drastically under arbitrarily small perturbations.) Here "long-time" refers to large n, where $\varphi^{(n+1)}(z) = \varphi(\varphi^{(n)}(z))$, $n = 0, 1, \ldots$, and $\varphi^{(0)}(z) = z$.

It would be interesting to adapt and modify the Haar wavelet, and the other wavelet algorithms to the Julia sets. The two papers [13,14] initiate such a development.

Literature

As evidenced by a simple Google check, the mathematical wavelet literature is gigantic in size, and the manifold applications spread over a vast number of engineering journals. While we cannot do justice to this volume st literature, we instead offer a collection of the classics [19] edited recently by C. Heil et al.

Acknowledgments

We thank Professors Dorin Dutkay and Judy Packer for helpful discussions. Work supported in part by the U.S. National Science Foundation.

Bibliography

1. Aubert G, Kornprobst P (2006) Mathematical problems in image processing. Springer, New York
2. Baggett L, Jorgensen P, Merrill K, Packer J (2005) A non-MRA C^r frame wavelet with rapid decay. Acta Appl Math 1–3:251–270
3. Bratelli O, Jorgensen P (2002) Wavelets through a looking glass: the world of the spectrum. Birkhäuser, Birkhäuser, Boston
4. Braverman M (2006) Parabolic julia sets are polynomial time computable. Nonlinearity 19(6):1383–1401
5. Braverman M, Yampolsky M (2006) Non-computable julia sets. J Amer Math Soc 19(3):551–578 (electronic)
6. Bredies K, Lorenz DA, Maass P (2006) An optimal control problem in medical image processing Springer, New York, pp 249–259
7. Daubechies I (1992) Ten lectures on wavelets. CBMS-NSF Regional Conference Series in Applied Mathematics, vol 61, SIAM, Philadelphia
8. Daubechies I (1993) Wavelet transforms and orthonormal wavelet bases. Proc Sympos Appl Math, Amer Math Soc 47:1–33, Providence
9. Daubechies I, Lagarias JC (1992) Two-scale difference equations. II. Local regularity, infinite products of matrices and fractals. SIAM J Math Anal 23(4):1031–1079
10. Devaney RL, Look DM (2006) A criterion for sierpinski curve julia sets. Topology Proc 30(1):163–179. Spring Topology and Dynamical Systems Conference
11. Devaney RL, Rocha MM, Siegmund S (2007) Rational maps with generalized sierpinski gasket julia sets. Topol Appl 154(1):11–27
12. Dutkay DE (2004) The spectrum of the wavelet galerkin operator. Integral Equations Operator Theory 4:477–487
13. Dutkay DE, Jorgensen PET (2005) Wavelet constructions in non-linear dynamics. Electron Res Announc Amer Math Soc 11:21–33
14. Dutkay DE, Jorgensen PET (2006) Hilbert spaces built on a similarity and on dynamical renormalization. J Math Phys 47(5):20
15. Dutkay DE, Jorgensen PET (2006) Iterated function systems, ruelle operators, and invariant projective measures. Math Comp 75(256):1931–1970
16. Dutkay DE, Jorgensen PET (2006) Wavelets on fractals. Rev Mat Iberoam 22(1):131–180
17. Dutkay DE, Roysland K (2007) The algebra of harmonic functions for a matrix-valued transfer operator. arXiv:math/0611539
18. Dutkay DE, Roysland K (2007) Covariant representations for matrix-valued transfer operators. arXiv:math/0701453
19. Heil C, Walnut DF (eds) (2006) Fundamental papers in wavelet theory. Princeton University Press, Princeton, NJ
20. Jorgensen PET (2003) Matrix factorizations, algorithms, wavelets. Notices Amer Math Soc 50(8):880–894
21. Jorgensen PET (2006) Analysis and probability: wavelets, signals, fractals. grad texts math, vol 234. Springer, New York
22. Jorgensen PET (2006) Certain representations of the cuntz relations, and a question on wavelets decompositions. In: Operator theory, operator algebras, and applications. Contemp Math 414:165–188 Amer Math Soc, Providence
23. Liu F (2006) Diffusion filtering in image processing based on wavelet transform. Sci China Ser F 49(4):494–503
24. Milnor J (2004) Pasting together julia sets: a worked out example of mating. Exp Math 13(1):55–92
25. Petersen CL, Zakeri S (2004) On the julia set of a typical quadratic polynomial with a siegel disk. Ann Math (2) 159(1):1–52
26. Skodras A, Christopoulos C, Ebrahimi T (2001) JPEG 2000 still image compression standard. IEEE Signal Process Mag 18:36–58
27. Song MS (2006) Wavelet image compression. Ph.D. thesis, University of Iowa
28. Song MS (2006) Wavelet image compression. In: Operator theory, operator algebras, and applications. Contemp. Math., vol 414. Amer. Math. Soc., Providence, RI, pp 41–73
29. Strang G (1997) Wavelets from filter banks. Springer, Singapore, pp 59–110
30. Strang G (2000) Signal processing for everyone. Computional mathematics driven by industrial problems (Martina F 1999), pp 365–412. Lect Notes Math, vol 1739. Springer, Berlin
31. Strang G, Nguyen T (1996) Wavelets and filter banks. Wellesley-Cambridge Press, Wellesley
32. Usevitch BE (Sept. 2001) A tutorial on modern lossy wavelet image compression: foundations of jpeg 2000. IEEE Signal Process Mag 18:22–35
33. Walker JS (1999) A primer on wavelets and their scientific applications. Chapman & Hall, CRC

Complex Dynamics of Traffic Management, Introduction to

Boris S. Kerner
GR/ETI, HPC: G021, Daimler AG, Sindelfingen, Germany

Traffic management is an extremely complex dynamic process associated with the spatiotemporal behavior of many-particle non-linear systems. In this Encyclopedia of Complexity and Systems Science, there are several review articles devoted to various aspects of the complex dynamics of vehicular traffic management, air traffic management, and pedestrian traffic management.

The complexity of vehicular traffic management is associated with non-linear interactions between the following three main dynamic processes:

(i) travel *decision* behavior,
(ii) traffic *assignment* in a traffic network, and
(iii) traffic flow behavior, in particular, *traffic congestion* occurrence within the network.

Travel decision behavior determines travel demand. Traffic assignment in the network is associated with travel supply. However, traffic congestion occurring within the traffic network restricts free flow travel. This influences both travel decision behavior (e. g., because of traffic congestion, a person decides to stay at home or travel by train rather than by car) and traffic assignment in the network (e. g., because of traffic congestion on a route from an origin to a destination usually used, a person changes their route of travel). The articles presented in this Encyclopedia cover all these main scientific fields of traffic management.

The review article of Goulias (see ► Travel Behavior and Demand Analysis and Prediction) is devoted to a detailed consideration of the theoretical and modeling approaches to travel behavior analysis. As shown in Goulias's article, traveler values and attitudes are referred to motivational, cognitive, situational, and disposition factors determining human behavior. Goulias presents a sequential chronological order merging technological innovations and theoretical innovations to the modeling of travel behavior. There are a diverse variety of approaches to the modeling of travel behavior, which are based on mathematical methods like multi-agent microsimulation widely used in many other fields of systems science. Inputs to these models are the typical regional model data about social, economic, and demographic information of potential travelers and land use information to create schedules followed by people in their everyday life. The output are detailed lists of activities pursued, times spent in each activity, and travel information from activity to activity. The in-depth understanding of transportation-related human behavior is essential to all kinds of traffic management. This is because travel behavior refers primarily to the modeling and analysis of travel demand.

Traffic assignment and control as well as methods for traffic prediction in various traffic networks are reviewed in the articles of Rakha and Tawfik (see ► Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment), Gartner and Stamatiadis (see ► Traffic Networks, Optimization and Control of Urban), Hegyi, Bellemans and De Schutter (see ► Freeway Traffic Management and Control), Rehborn and Klenov (see ► Traffic Prediction of Congested Patterns). In particular, Rakha and Tawfik review methods for modeling of dynamic traffic assignment in traffic networks. Gartner and Stamatiadis consider methods for traffic modeling and light signal control in city and urban traffic networks. Traffic control methods developed for traffic on freeway networks are discussed in the article of Hegyi, Bellemans, and De Schutter. Rehborn and Klenov discuss methods for traffic prediction required for dynamic traffic assignment in urban and freeway traffic networks.

In an urban network with short enough network links, dynamic traffic phenomena are determined mostly by traffic signals and other traffic regulations at link intersections. The dynamic traffic phenomena in the urban network can be nevertheless very complex, because traffic regulations at one of the link intersections can have a great influence on the probability of traffic congestion occurring on other links of the network. In their article, Gartner and Stamatiadis (► Traffic Networks, Optimization and Control of Urban) present a historical model development beginning from a consideration of traffic control methods at an isolated link intersection to a discussion of very complex approaches to dynamic traffic signal control and optimization in a complex city network.

Driver interactions, which are always essential for traffic dynamics on highways, lead to very complex spatiotemporal phenomena in vehicular traffic.

Hegyi, Bellemans and De Schutter (► Freeway Traffic Management and Control) present an introduction to control and managements methods currently used in field trials on freeways as well as to some recent control and managements approaches. These methods include on-ramp feedback metering, congestion pricing, speed limitation, collective and individual route guidance systems. Future traffic management methods will include also diverse driver assistance systems, in particular, with the use of vehicle ad-hoc networks and vehicle-to-infrastructure communication.

A dynamic traffic management in a traffic network requires a dynamic traffic assignment model. Rakha and Tawfik (► Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment) provide a detailed discussion of the problem of dynamic traffic assignment. A dynamic traffic assignment model should find the link inflows for the network. The model usually includes a traffic flow model (traffic modeler), which makes a prognosis of traffic in the network, and a traffic routing model (traffic router) associated with the problem of traffic optimization. The traffic router computes the sequence of roadways that minimize some utility objective function, i. e., minimize travel "costs" of the traffic network. Examples of the travel costs are travel time, fuel consumption, or HC and CO_2 emissions. In a real traffic network, traffic routing can be organized through the use for example of individual and collective driver guidance systems as well as congestion pricing. The traffic modeler and router are connected by a feedback loop. As a result, traffic congestion in the network predicted by the traffic modeler changes results of dynamic traffic assignment considerably. For this rea-

son, the traffic modeler should model traffic congestion as close as possible to real traffic congestion found in empirical observations.

In the dynamic traffic management models, a reliable prognosis of traffic is also required. As shown in the article of Rehborn and Klenov (► Traffic Prediction of Congested Patterns), there are two main approaches to traffic prognosis: (a) "physics-of-traffic", i. e., traffic prognosis models and methods based on the understanding of the nature of empirical spatiotemporal features of traffic patterns; (b) methods of mathematical statistics or artificial intelligence, which are also called "data mining" methods, in which the understanding of traffic patterns is not necessarily needed.

Data mining models for traffic prognosis do not usually require the understanding of traffic data, which these methods use for traffic prediction. Using a huge number of historical traffic data, such a method learns predictable features of traffic data without the pretension of their understanding. The data mining methods used for traffic prognosis include statistical methods based on regression, a wavelet approach, or filtering models as well as neural networks. It should be noted that there are many articles in this Encyclopedia, in which these and other methods of artificial intelligence are reviewed.

The physics-of-traffic approach to traffic prognosis is based on the understanding of measurements of traffic variables made in space and time as well as on simulations of the traffic variables associated with empirical data. The key element for a successful prediction concept lies in the correct and reliable understanding of the empirical features of the traffic process and their reproducible characteristics. Furthermore, this understanding of real traffic should be incorporated in a traffic flow model, which should explain and predict the empirical traffic features. Therefore, in the physics-of-traffic approach a traffic flow prediction model is based on empirical features of phase transitions or/and resulting congested patterns; current and historical measured traffic data are used in the model, which reconstructs traffic phases and makes the tracking and prediction of the propagation of traffic congestion in a traffic network.

Thus we see that the understanding of empirical traffic congestion is the key for effective traffic management, control, organization, and all other traffic applications.

Empirical congested patterns exhibit a complex spatiotemporal behavior that was studied during the last 75 years by several generations of scientists. It was found that traffic congestion in the traffic network results from traffic breakdown in an initially free flow: vehicle speeds decrease abruptly to lower speeds in congested traffic. Traffic breakdown is observed mostly at highway bottlenecks. A bottleneck can be a result of road works, on- and off-ramps, a decrease in the number of freeway lanes, road curves and road gradients, bad weather conditions, accidents, etc. In congested traffic a "stop-and-go" mode, i. e., a sequence of moving traffic jams is very often observed.

However, the puzzle of empirical spatiotemporal features of traffic congestion has been solved only recently. Consequently, an alternative traffic flow theory that is called three-phase traffic theory and the associated three-phase traffic flow models, which can predict and explain the empirical spatiotemporal features of traffic congestion, have had to be introduced. The development of various traffic flow theories and modeling approaches to traffic congestion is discussed in (► Traffic Congestion, Modeling Approaches to). These modeling approaches include microscopic, mesoscopic and macroscopic traffic flow models, in particular based on cellular automata and gas kinetic models widely used in many other fields of systems science. A probabilistic theory of traffic breakdown is considered in the review article of Kerner and Klenov (see ► Traffic Breakdown, Probabilistic Theory of). Spatiotemporal features of traffic congested patterns resulting from traffic breakdown are considered in the review article (► Traffic Congestion, Spatiotemporal Features of). Empirical observations of traffic congestion discussed in these articles show that phenomena of traffic breakdown and the subsequent evolution of resulting congested patterns are associated with diverse phase transitions and complex spatiotemporal self-organization effects in vehicular traffic.

On the one hand, the phase transitions and self-organization traffic phenomena, which can occur randomly over time and space within a traffic network, depend considerably on the spatiotemporal distribution of traffic demand in the network. On the other hand, the occurrence of traffic congestion changes travel decision behavior and traffic assignment considerably. In turn, travel decision behavior and traffic assignment determine the spatiotemporal distribution of traffic demand in this network. Thus the recent understanding of empirical features of the phase transitions and spatiotemporal effects in vehicular traffic (see ► Traffic Congestion, Modeling Approaches to and ► Traffic Congestion, Spatiotemporal Features of) should be taken into account in future models and theories of dynamic traffic assignment as well as in the subsequent development of traffic management and control strategies. This will be one of the most important and difficult challenges for transportation research in the next future.

Thus the complexity of traffic management is associated with diverse phase transitions in vehicular traffic re-

sulting in various complex self-organizing spatiotemporal congested patterns propagating within a traffic network as well as with the necessity in the optimization of the traffic congested patterns. This optimization should ensure either the dissolution of traffic congestion or, if this is not possible to achieve, the minimization of the influence of traffic congestion on travel costs like travel time, fuel consumption, or HC and CO_2 emissions.

Modeling approaches to vehicular traffic management discussed in the mentioned above review articles of the Encyclopedia are also very important for studies of air traffic and pedestrian traffic.

Air traffic continues to grow at a steady pace in the world. For this reason, the discussion of air traffic dynamics as well as approaches to air traffic modeling made in the review article of Sridhar and Sheth (see ► Air Traffic Control, Complex Dynamics of) is an important contribution to the fields of traffic management.

Currently, we can observe a continuous growth of publications devoted to modeling approaches to pedestrian traffic and crowd dynamics. In particular, the interest to pedestrian traffic is associated with necessity in the development of effective evacuation strategies required for the relocation of people to safely escape hazardous disaster impacts. During the evacuation, destinations are chosen to escape the hazardous disaster impacts as quickly as possible. This is a peculiarity of the evacuation process in comparison with vehicle traffic in which every vehicle has usually a defined destination. This very interesting and important field of traffic management is considered in the review articles of Helbing and Johansson, Schadschneider et al., and Goudie.

Helbing and Johansson (see ► Pedestrian, Crowd and Evacuation Dynamics) and Schadschneider et al. (see ► Evacuation Dynamics: Empirical Results, Modeling and Applications) consider various aspects of the modeling of pedestrian interactions during evacuation as well as a comparison of the associated theoretical and empirical results. Based on an example of Australian fire zone research, Goudie (see ► Evacuation as a Communication and Social Phenomenon) considers evacuation as a communication and social phenomenon. Goudie's article describes the influence of communication and social aspects on possible evacuation scenarios.

Acknowledgments

I would like to thank Konstadinos Goulias, Sergey Klenov, Hesham Rakha, Hubert Rehborn, Andreas Schadschneider, and Bart De Schutter for useful suggestions.

Complex Gene Regulatory Networks – from Structure to Biological Observables: Cell Fate Determination

Sui Huang, Stuart A. Kauffman
Institute for Biocomplexity and Informatics,
Department of Biological Sciences, University of Calgary,
Calgary, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
Overview: Studies of Networks in Systems Biology
Network Architecture
Network Dynamics
Cell Fates, Cell types: Terminology and Concepts
History of Explaning Cell Types
Boolean Networks as Model for Complex GRNs
Three Regimes of Behaviors for Boolean Networks
Experimental Evidence from Systems Biology
Future Directions and Questions
Bibliography

Glossary

Transcription factor A protein that binds to the regulatory region of a target gene (its promoter or enhancer regions) and thereby, controls its expression (*transcription* of the target gene into a mRNA which is ultimately *translated* into the *protein* encoded by the target gene). Transcription factors account for temporal and contextual specificity of the expression of genes; for instance, a developmentally regulated gene is expressed only during a particular phase in development and in particular tissues.

Gene regulatory network (GRN) Transcription factors regulate the expression of other transcription factor genes as well as other 'non-regulatory' genes which encode proteins, such as metabolic enzymes or structural proteins. A regulatory relationship between two genes thus is formalized as: "transcription factor A is the regulator of target gene B" or: A → B. The entirety of such regulatory interactions forms a network = the gene regulatory network (GRN). Synonyms: genetic network or gene network, transcriptional network.

Gene network architecture and topology The GRN can be represented as a "*directed graph*". The latter consists of a set of nodes (= vertices) representing the genes connected by arrows (directed *links* or *edges*)

representing the regulatory interactions pointing from the regulator to the regulated target gene. [In contrast, in an *undirected* graph, the links are simple lines without arrowheads. The protein-interaction network can be represented as undirected graph]. The topology of a network is the structure of this graph and is an abstract notation, without physicality, of all the potential regulatory interactions between the genes. Topology usually is used to denote the simple interactions captured by the directed graph. For defining the network dynamics, however, additional aspects of interactions need to be specified, including: modalities or "sign" of an arrow (inhibitory vs. activating regulation), the 'transfer functions' (relationship between magnitude of input to that of the output = target gene) and the logical function (notably, in Boolean network, defining how multiple inputs are related to each other and are integrated to shaping the output). In this article when all this additional information is implied to be included, the term gene network *architecture* is used. Thus, the graph topology is a subset of network architecture.

Gene network dynamics The collective change of the gene expression levels of the genes in a network, essentially, the change over time of the network state *S*.

State space Phase space = the abstract space that contains all possible states *S* of a dynamical system. For (autonomous) gene regulatory networks, each state *S* is specified by the configuration of the expression levels of each of the *N* genes of the network; thus a system state S is one point in the *N*-dimensional state space. As the system changes its state over time, S moves along trajectories in the state space.

Transcriptome Gene expression pattern across the entire (or large portion of) the genome, measured at the level of mRNA levels. Used as synonym to "gene expression profile". The transcriptome can in a first approximation be considered a snapshot of the network state *S* of the GRN in gene network dynamics.

Cell type A distinct, whole-cell phenotype characteristic of a mature cell specialized to exert an organ-specific function. Example of cell types are: liver cell, red blood cell, skin fibroblast, heart muscle cell, fat cell, etc. Cell types are characterized by their distinct cell morphology and their gene expression pattern.

Cell fate A potential developmental outcome of a (stem or progenitor) cell. A cell fate of a stem cell can be the development into a particular mature cell type.

Multipotency The ability of a cell to generate multiple cell types; a hallmark of stem cells. Stem cells are said to be multipotent (see also under Stem cells).

Stem cell A multi-potent cell capable of "self-renewal" (division in which both daughter cells have the same degree of multi-potency as the mother cell) and can give rise to multiple cell types. There is a hierarchy of multipotency: a *toti-potent embryonic stem cell* can generate all possible cell types in the body, including extra-embryonic tissues, such as placenta. A *pluripotent embryonic stem cell* can generate tissues of three germ layers, i. e., it can produce all cell types of the foetus and the adult. A *multipotent (sensu stritiore) stem cell* of a tissue (e. g., blood) can give rise to all cell types of that tissue (e. g., a hematopoietic stem cell can produce all the blood cells). A multipotent *progenitor* cell can give rise to more than one cell types within a tissue (e. g. the granulocyte-monocyte progenitor cell).

Cell lineage Developmental trajectory of a multipotent cell towards one of multiple cell types, e. g., the "myeloid lineage" among blood cells, comprising the white blood cells granulocytes, monocytes, etc. Thus, a cell fate decision is a decision between multiple lineages accessible to a stem or progenitor cell.

Differentiation The process of cell fate decision in a stem or progenitor cell and the subsequent maturation into a mature cell type.

Definition of the Subject

Current studies of complex gene regulatory networks (GRN) in which thousands of genes regulate each others' expression have revealed interesting features of the network structure using graph theory methods. But how does the particular network architecture translate into biology? Since individual genes alter their expression level as a consequence of the network interactions, the genome-wide gene expression pattern (*transcriptome*), which manifests the dynamics of the entire network, changes as a whole in a highly constrained manner. The transcriptome in turn determines the cell phenotype. Hence, the constraints in the global dynamics of the GRN directly map into the most elementary "biological observable": the existence of distinct cell types in the metazoan body and their development from pluripotent stem cells.

In this article a historical overview of the various levels at which GRNs are studied, starting from network architecture analysis to the dynamics are first presented. An introduction is given to continuous and discrete value models of GRN commonly used to understand the dynamics of small genetic circuits or of large genome wide networks, respectively. This will allow us to explain how the intuitive metaphor of the "epigenetic landscape", a key idea that was proposed by Waddington in the 1940s to explain

the generation of discrete cell fates, formally arises from gene network dynamics. This central idea appears in its modern form in formal and molecular terms as the concept that cell types represent *attractor states* of GRNs, first proposed by Kauffman in 1969. This raises two fundamental questions currently addressed by experimental biologists in the era of "systems biology": (1) Are cell types in the metazoan body indeed high-dimensional attractors of GRNs and (2) is the dynamics of GRNs in the "critical regime" – poised between order and chaos? A summary of recent experimental findings on these questions is given, and the broader implications of network concepts for cell fate commitment of stem cells are also briefly discussed. The idea of the epigenetic landscape is key to our understanding of how genes and gene regulatory networks give rise to the observable cell behavior, and thus to a formal and integrated view of molecular causation in biology.

Introduction

With the rise of "systems biology" over the past decade, molecular biology is moving away from the paradigm of linear genetic pathways which has long served as linear chains of causation (e. g., Gene A → Gene B→ Gene C→ phenotype) in explaining cell behaviors. It has begun to embrace the idea of molecular networks as an integrated information processing system of the cell [93,138]. The departure from the gene-centered, mechanistic 'arrow-arrow' schemes that embody 'proximal causation' [189] towards an integrative view will also entail a change in our paradigm of what an "explanation" means in biology: How do we map the collective behavior of thousands of interacting genes, obtained from molecular dissection, to the "***biological observable***"? The latter term, borrowed from the statistical physics idea of the "macroscopic observable", is most prosaically epitomized in whole-cell behavior in metazoan organisms: its capacity to adopt a large variety of easily recognizable, discretely distinct phenotypes, such as a liver cell vs. a red blood cell, or different functional states, such as the proliferative, quiescent or the apoptotic state.

All these morphologically and functionally distinct phenotypes are produced by the very same set of genes of the genome as a result of the joint action of the genes. This is achieved by the differential expression ("turning ON or OFF") of individual genes. Thus, in a first approximation, each cell phenotype is determined by a specific configuration of the status of expression of all individual genes across the genome. This genome-wide ***gene expression pattern or profile***, or ***transcriptome***, is the direct output of the GRN and maps almost uniquely into a cell phenotype (see overview in Fig. 1).

A ***network*** is the most elementary conceptualization of a complex system which is composed of interacting elements and whose behavior as an entity we wish to understand: it formalizes how a set of distinct elements (nodes) influence each other as predetermined by a fixed scheme of their interactions (links between the nodes) and the modality of interactions (rules associated with each node). In this article we focus on the ***gene regulatory network*** (GRN), the network formed by interactions through which genes regulate each other's expression (Fig. 1a), and we ask how they control the global ***behavior*** of the network, and thereby, govern development of cells into the thousands of cell types found in the metazoan body.

Because in a network a node exerts influence onto others, we can further formalize networks as ***directed graphs***, that is, the links are arrows pointing from one node to another (Fig. 1b). The information captured in an undirected or directed graph is the ***network topology*** (see Glossary). In addition, the arrows have a ***modality*** ("sign"), namely, activating or inhibiting their target gene. But since each node can receive several inputs ("upstream regulators"), it is more appropriate to combine modality of interaction together with the way the target gene integrates the various inputs to change its expression behavior (= output). Thus, each node can be assigned a ***function*** that maps all its inputs in a specific way to the output (Fig. 1, top). For instance, "promoter logics" [209] which may dictate that two stimulating inputs, the transcription factors, act synergistically, or that one inhibitory input may, when present, override all other activating inputs, is one way to represent such an input integrating function.

Here we use ***network architecture*** as a term that encompasses both network topology as well as the interaction modalities or the functions. The latter add the ingredients to the topology information that are necessary to describe the ***dynamics*** (***behavior***) of the network.

The Core Gene Regulatory Network (GRN) in Mammals

How complex is the effective GRN of higher, multicellular organisms, such as mammals? Virtually every cell type in mammals contains the ∼25,000 genes in the genome [44, 158,178] which could potentially interact with each other. However, in a first approximation, we do not need to deal with all the 25,000 genes but only with those intrinsic regulators that have direct influence on other genes. A subset of roughly 5–10% of the genes in the genome encode ***transcription factors*** (TF) [194] a class of DNA binding pro-

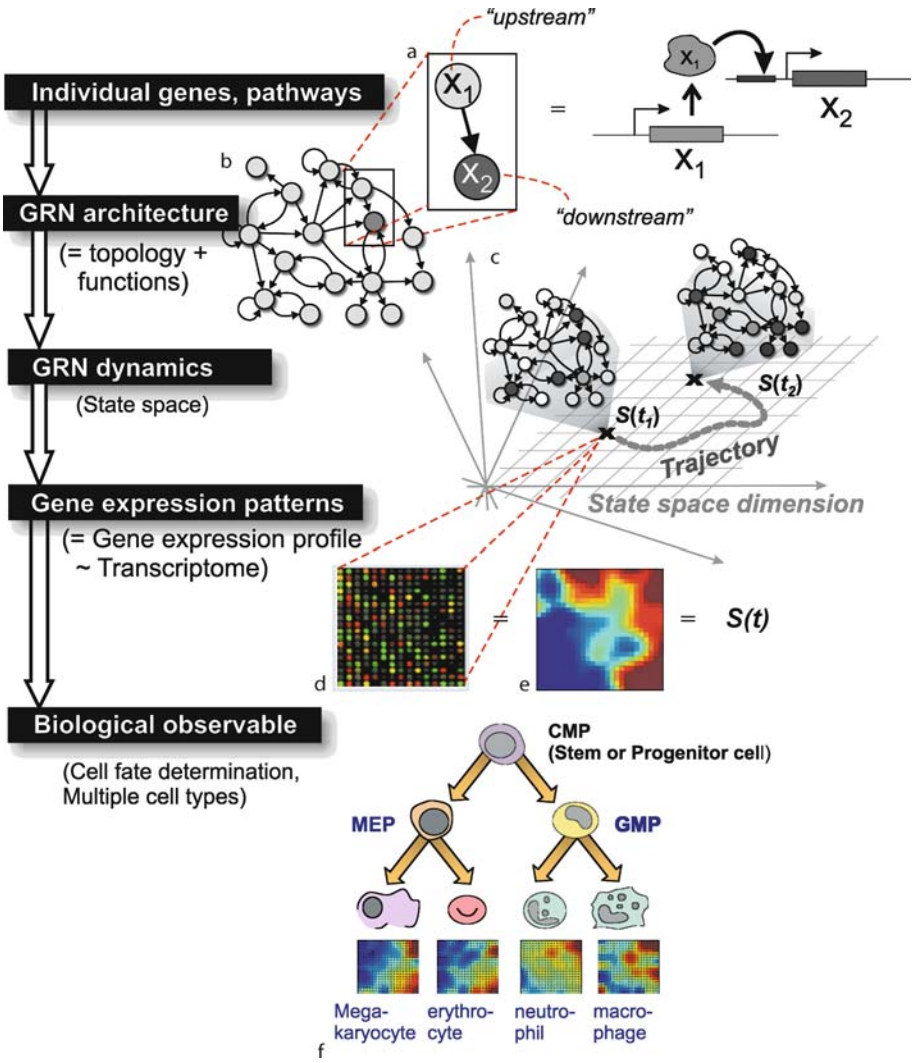

Complex Gene Regulatory Networks – from Structure to Biological Observables: Cell Fate Determination, Figure 1

Overview: Concepts at different levels for understanding emergence of biological behavior from genes and networks. a Elementary interaction: regulator gene X_1 interacts with target (regulated) gene X_2, simplified as one link of a graph and the corresponding standard 'cartoon' it represents. b Notation of gene regulatory network topology as *directed graph*. c Schematic of a higher dimensional gene expression state space, showing that a network state $S(t)$ maps onto one point (denoted as *cross*) in the state space. *Dashed curve/arrow* represents state space trajectory. d Spotted DNA microarray for measuring gene expression (mRNA) profiles, representing a network state $S(t)$ and e the associated "GEDI map" visualization of gene expression pattern. The map is obtained using the program GEDI [57,83]. This program places genes that behave similarly (with respect to their expression in a set of microarray measurements) onto the same pixel (= minicluster of genes) of the map. Similar miniclusters are arranged in nearby pixels in the two-dimensional picture of an $n \times m$ array of pixels. The assignment of genes to a pixel is achieved by a self-organizing map (SOM) algorithm. The *color of each pixel* represents the centroid gene expression level (mRNA abundance) of the genes in the minicluster. For a stack of GEDI maps, all genes are forced to be assigned to the same pixels in the different maps, hence the global coherent patterns of each GEDI map allows for a one-glance 'holistic' comparison of gene expression profiles of different conditions $S(t)$ (tissue types, time points in a trajectory, etc). f Scheme of branching development of a subset of four different blood cells with their distinct GEDI maps, starting from the multipotent *CMP* = common myeloid progenitor cell [150]. *MEP* = Megakaryocte-Erythorid progenitor cell; *GMP* = Granulocyte-Monocyte progenitor

teins that regulate the expression of genes by binding to the promoter region of the 'target genes' (Fig. 1a). Another few hundred loci in the genome encode microRNAs (miRNA), which are transcribed to RNA but do not encode proteins. miRNAs regulate the expression of genes by interfering with the mRNA of their target genes based on sequence complementarity (for reviews see [41,89,141]). Thus, in this discussion we can assume a regulatory network of around 3000 regulator genes rather than the 25,000 genes of the genome. The genes that do not encode TFs are the "effector genes", encoding the work horses of the cell, including metabolic enzymes and structural genes. In our approximation, we also assume that the effector genes do not directly control other genes (although they may have global effects such as change of pH that may affect the expression of some genes). Further we do not count genes encoding proteins of the signal transduction machinery since they act to mediate external signals (such as hormones) that can be viewed as perturbations to the network and can in a first approximation be left out when discussing cell-intrinsic processes. Thus, the directed graph describing the genome-wide GRN network has a "medusa" structure [112], with a core set of regulators (medusa head) and a periphery of regulated genes (arms) which in a first approximation do not feedback to the core.

The Core GRN as a Graph That Governs Cell Phenotype

The next question is: do the 3000 core regulatory genes form a *connected* graph or rather independent (detached) "modules"? The idea of modularity would have justified the classical paradigm of independent causative pathways and has in fact actively been promoted in an attempt to mitigate the discouragement in view of the unfathomable complexity of the genome [85]. While a systematic survey that would provide a precise number is still not available, we can, based on patchy knowledge from the study of individual TFs, safely assume that a substantial fraction of TFs control the expression of more than one other TF. Many of them also control entire batteries [50] of effector genes, while perhaps a third subset may be specialized in regulating solely the effector genes. In any case, the core regulatory network of ∼3000 nodes controls directly or indirectly the entire gene expression profile of 25,000 genes, and hence, the cell phenotype.

Then, assuming that on a average each TF controls at least two (typically more) other TFs [128,188], and considering statistical properties of random evolved networks (graphs) [36], we can safely assume that the core transcriptional network is a connected graph or at least its giant component (largest connected subgraph) covers the vast majority of its nodes. This appears to be the case in GRNs of simple organisms for which more data are available for parts of the network, as in yeast [128], *C. elegans* [54]; sea urchin [50] or, for even more limited subnetworks in mammalian systems [32,183] although studies focused on selected subnetworks may be subjected to investigation bias. However, recent analysis of DNA binding by 'master transcription factors' in mammalian cells using systemic (hence in principle unbiased) chromatin immunoprecipitation techniques [177] show that they typically bind to hundreds if not thousands of target genes [48,70,105,155], strongly suggesting a global interconnectedness of the GRN. This however, does not exclude the possibility that the genome-wide network may exhibit some "modularity" in terms of weakly connected modules that are locally densely connected [166].

Structure of This Article

The goal of this article is to present both to biologists and physicists a set of basic concepts for understanding how the maps of thousands of interacting genes that systems biology researchers are currently assembling, ultimately control cell phenotypes. In Sect. "Overview: Studies of Networks in Systems Biology" we present an overview to experimental biologists on the history of the analysis of networks in systems biology. In Sect. "Network Architecture" we briefly discuss core issues in studies of network *topology*, before explaining basic ideas of network *dynamics* in Sect. "Network Dynamics" based on two-gene circuits. The central concepts of multi-stability will be explained to biologists, assuming a basic calculus background. Waddington's *epigenetic landscape* will also be addressed in this conjunction. In Sect. "Cell Fates, Cell Types: Terminology and Concepts" we discuss the actual "*biological observable*" that the gene regulatory network controls by introducing to non-biologists central concepts of cell fate regulation. Section "History of Explaning Cell Types" offers a historical overview of various explanations for metazoan cell type diversity, including dynamical, network-based concepts, as well as more 'reductionist' explanations that still prevail in current mainstream biology. Here the formal link between network concepts and Waddington's landscape metaphor will be presented. We then turn from small gene circuits to large, complex networks, and in Sect. "Boolean Networks as Model for Complex GRNs" we introduce the model of random Boolean networks which, despite their simplicity, have provided a useful conceptual framework and paved the way to learning what an integrative understanding

of global network dynamics would look like, leading to the first central hypothesis: *Cell types may be high-dimensional attractors of the complex gene regulatory network*. In Sect. "Three Regimes of Behaviors for Boolean Networks" the more fundamental dynamical properties of ordered, critical and chaotic behavior are discussed, leading to the second hypothesis: *Networks that control living cells may be in the critical regime*. In Sect. "Experimental Evidence from Systems Biology" we summarize current experimental findings that lend initial support to these ideas, and in Sect. "Future Directions and Questions" we conclude with an outlook on how these general concepts may impact future biology.

Overview: Studies of Networks in Systems Biology

Cellular networks of biological molecules currently studied by "systems biologists" encompass three large classes: metabolic, protein–protein interaction and gene regulatory (transcriptional) networks (GRN). The formalization of these systems into networks is often taken for granted but is an important issue. It is noteworthy that metabolic network diagrams [193] represent physical networks in that there is a *flow* of information or energy in the links, and thus, the often used metaphors inspired by man-made transport or communication networks ('bottleneck', 'hubs' etc.) are more appropriate than in the other two classes. Moreover, metabolic networks are subjected to the constraint of mass preservation at each node – in compliance with Lavoisier's mass conservation in chemistry. In contrast to such flow networks, protein–protein interaction networks and GRN are abstract notations of "influence networks" in which the nodes influence the behavior of other nodes directly or indirectly – as represented by the links of the graph. There is no actual flow of matter in the links (although of course information is exchanged) and there is no obvious physical law that constrains the architecture, thus allowing for much richer variations of the network structure. The links are hence abstract entities, representing *potential* interactions assembled from independent observations that may not coexist in a particular situation. Thus, in this case the network is rather a convenient graphical representation of a collection of potential interactions.

Studies of such biomolecular influence networks can fundamentally be divided between two levels (Table 1): (A) network architecture and (B) network dynamics (system behavior).

The study of network architecture in molecular biology can further be divided into (A1) efforts to *determine* the actual graph structure that represent the specific interactions of the genes for particular instance (a species) and (A2) the more recent *analysis* of its structure, e. g., for interesting topological features [2]. The determination of the graph structure of the genome-wide GRNs in turn is either achieved (*i*) by direct *experimental demonstration* of the physical interactions, which has in the past decade greatly benefited from novel massively parallel technologies, such as chromatin IP, promoter one-hybrid or protein binding DNA microarrays [35,54,177], or (*ii*) via theoretical *inference* based on observed correlations in gene expression behavior from genome-wide gene expression profiling experiments. Such correlations are the consequence of mutual dependencies of expression in the influence network.

Complex Gene Regulatory Networks – from Structure to Biological Observables: Cell Fate Determination, Table 1
Overview of typical directions and levels in the study of gene regulatory networks

A. Network Architecture
A1. Determination of network architecture
► Experimental
► Theoretical = ***Network inference***
A2. Analysis of network architecture
Identification/characterization of "interesting" structural features
B. Network Dynamics
For small networks
B1. Modeling: theoretical prediction of behavior
of a real circuit based on (partially known, assumed) architecture + experimental verification
For complex networks (full architecture not known)
B2. Theoretical
Study of generic dynamics of simulated ensembles of networks of particular architecture classes
B3. Experimental
Measurement and analysis of high-dimensional state space trajectories of a real network

The next section provides a concise overview of the study of the network architecture, which due to limitations of available data is mostly an exercise in studying graph topology, and briefly discusses associated problems.

Network Architecture

Network Inference

An emerging challenge is the determination of entire networks, including the connection graph (topology), the direction of interactions ('arrows") and, ideally, the interaction modality and logical function, using systematic inference from genome-wide patterns of gene expression ("inverse problem"). Gene expression profiles (transcriptomes) represent a snap-shot of the mRNA levels in cell or tissue samples. The arrival of DNA microarrays for efficient measuring gene expression profiles for almost all the genes in the genome has stimulated the development of algorithms that address the daunting inverse problem, and the number of proposed algorithms has recently exploded. Most approaches concentrate on inference based on the more readily available static gene expression profiles, although time course microarray experiments, where time evolution of transcriptomes are monitored at close intervals would be advantageous, especially to infer the arrow direction of network links (directed graph) and the interaction functions.

However, because of uncertainties and the paucity of experimental data, systematic network inference faces formidable technical and formal challenges, and most theoretical work has been developed and tested based on simulated networks. A fundamental concern is that microarray based expression levels reflect the average expression of a population of millions of cells that – as recently demonstrated – exhibit vastly diverse behaviors even when they are clonal. Thus, the actual quantity used for inference is not a direct but a convoluted manifestation of network regulation (this issue is discussed in Sect. "Are Cell Types Attractors?"). Moreover, while mRNA levels reflect relatively well the activity status of the corresponding gene promoter, i. e., revealing the *regulated* activity, they are poor indicator of the *regulating* activity of a gene, because of the loose relationship between mRNA level and effective activity of the transcription factor that it encodes. Since the true network architecture ('gold standard') is not known, validation of the theoretical approaches remains unsatisfactory. Nevertheless, the recent availability of large numbers of gene expression profiles and the increasing (although not complete) coverage of gene regulation maps for single-cell organisms (notably E. coli and yeast) open the opportunity to directly study the mapping between gene expression profile and network structure [39,42,61,88,135,190].

Here we refer to [2,131,139] for a survey of inference methods and instead briefly discuss the study of network topology before we move on to network dynamics.

Analyzing Network Structures

Once the network topology is known, even if direction, modality and logics of links are not specified to offer the complete system architecture, it can be analyzed using graph theory tools for the presence of global or local structures (subgraphs) that are "interesting". A potentially interesting feature can be defined as a one which cannot be explained by chance occurrence in some randomized version of the graph (null model), i. e., which departs from what one would expect in a "random" network (see below). Most of the graph theoretical studies have been stimulated by the *protein–protein interaction networks* which have been available for some years, of which the best characterized is that of the yeast *S. cerevisiae* [29]. Such networks represent non-directed graphs, since the links between the nodes (proteins) have been determined by the identification of physical protein interactions (heterodimer or higher complex formation). Here we provide only a cursory overview of this exploding field, while focusing on conceptual issues.

A large array of structural network features, many of them inspired by the study of ecological and social networks [152,160,199] have been found in biomolecular networks. These features include global as well as local features, such as, to mention a few, the scale-free or broad-scale distribution of the connectivity k_i of the nodes i [3,10,20], betweenness of node i [107,208], hierarchical organization [164], modularity [102,121,133,166,206], assortativity [140], and enrichment for specific local topology motifs [147], etc. (for a review of this still expanding field see [2,21,29,152]).

The global property of a scale-free distribution of connectivity k_i, which has attracted most attention early on and quickly entered the vocabulary of the biologist, means that the probability $P(k_i)$ for an arbitrary node (gene) i in the network to have the connectivity k_i has the form $P(k_i) \sim k_i^{\gamma}$ where the characteristic constant γ is the power-law exponent – the slope of the line in a $P(k_i)$ vs. k_i double-logarithmic plot. This distribution implies that there is no characteristic scale, i. e., no stable average value of k: sampling of larger number of nodes N will lead to larger "average" k values. In other words, there is an "unexpectedly" high fraction of nodes which are highly connected ("hubs") while the majority exhibits low connec-

tivity. This property has attracted as much interest as it has stirred controversy because of the connotation of "universality" of scale-freeness on the one hand, and several methodological concerns on the other hand [21,24,73,76,175,181,193]

The Problem of Choosing the Null Model In addition to well-articulated caveats due to incompleteness, bias and noise of the data [29,52], an important general methodological issue in the identification of structural features of interest, especially if conclusion on functionality and evolution is drawn, is the choice of the appropriate "null model" – or "negative control" in the lingo of experimentalists. A total random graph obviously is not an ideal null model, for on its background, any bias due to obvious constraints imposed by the physical reality of the molecular network will appear non-random and hence, be falsely labeled "interesting" even if one is not interested in the physical constraints but in evidence for functional organization [14,93]. For instance, the fact that gene duplication, a general mechanism of genome growth during evolution, will promote the generation of the network motif in which one regulator regulates two target genes, needs to be considered before a "purposeful" enrichment for such a motif due to increased fitness is assumed [93]. Similarly, the rewiring of artificial regulatory networks through reshuffling of *cis* and *trans* regions or the construction of networks based on promoter-sequence information content reveal constraints that lead to bias for particular structures in the absence of selection pressure [18,46]

The problem amounts to the practical question of which structural property of the network should be preserved when randomizing an observed graph to generate a null model. Arbitrarily constrained randomization based on preservation of some a priori graph properties [18,140] thus may not suffice. However, the question of which feature to keep cannot be satisfactorily answered given the lack of detailed knowledge of all physical processes underlying the genesis and evolution of the network. The more one knows of the latter, the more appropriate constraints can be applied to the randomization in order to expose novel biological feature that cannot be (trivially) explained by the neglected elementary construction constrains.

Evolution of Network Structure: Natural Selection or Natural Constraint?

This question goes beyond the technicality of choosing a null model for it reaches into the realm of a deeper questions of evolutionary biology: which features are inevitably linked to the very physical processes through which networks have grown in evolution and which arose due to natural selection because they confer a survival advantage [15,77,78,93,203]? Often this question is taken for granted and an all-mighty selection process is assumed that can produce any structure as long as it contributes sufficiently to fitness. This however, would require that during natural selection the random, mutation-driven reshuffling of nodes and connections has no constrains and that Darwinian evolution explores the entire space of possible architectures. Clearly this is not the case: physical constraints in the rearrangement of DNA (insertions, deletions, duplications, conversion, etc) [16,170,187] as well as graph theoretical considerations channel the possibilities for how one graph can transform into another [93,203]. For instance, growth of the network due to the increase of genome-size (gene number) by gene duplication can, without the invisible hand of selection, give rise to the widely-found scale-free structure [23,186] although it remains to be seen whether this mechanism (and not some more fundamental statistical process) accounts for the ubiquitous (near) scale-free topology. The fact that the scale-free (or at least, broad-scale [10]) architecture has dynamical consequences (see Sect. "Architectural Features of Large Networks and Their Dynamics") raises the question whether properties such as robustness may be inherent to some structure that is "self-organized", rather than sculpted by the invisible hand of natural selection. Thus, one certainly cannot simply argue that the scale-free structure has evolved "because of some functional advantage". Instead, it is reminded here that natural selection can benefit from spontaneous, self-organized structures [116]. This structuralist view [200], in which the physically most likely and not the biologically most functional is realized, need to be considered when analyzing the anatomy of networks [93].

In summary, the choice of the null model has to be made carefully and requires knowledge of the biochemical and physicochemical process that underlies genome evolution. This methodological caveat has its counterpart in the identification of "interesting" nucleotide sequence motifs when analyzing genome sequences [167].

Gene-Specific Information: More Functional Analysis Based on Topology

Beyond pure graph theoretical analysis, there have been attempts to link the topology with functional biological significance. For instance, one question is how the global graph structure changes (such as the size of the giant component) when nodes or links are randomly or selectively (e. g., hubs) removed. It should be noted that the term

"robustness" used in such structural studies [5] refer to networks with the aforementioned connotation of transport or communication function with flow in the links and thus, differs fundamentally from robustness in a *dynamical* sense in the influence networks that we will discuss below.

Another approach towards connecting network topology with biological functionality is to employ bioinformatics and consider the *biological identity* of the genes or proteins represented by the nodes. Then one can ask whether some graph-related node properties (e. g., degree of connectivity, betweenness/centrality, contribution to network entropy, etc.) are correlated with known biological properties of the genes, of which the most prosaic is the "essentiality" of the protein, derived from genetic studies [84,137,207,208]. To mention just a few of the earlier studies of this expanding field, it has been suggested that proteins with large connectivity ("hubs") appear to be enriched for "essential proteins" [104], that hubs evolve slower [68], and that they tend to be older in evolution [58] – in accordance with the model of preferential attachment that generates the scale-free distribution of the connectivity $\underline{k}_i$. However, many of these findings have been contested for statistical or other reasons [27,28,67,106,107]. The conclusion of such functional bioinformatics analysis need to be re-examined when more reliable and complete data become available.

Network Dynamics

While the graph theoretical studies, including the ramifications into core questions of evolution outlined above, are interesting eo ipso, life is breathed into these networks only with the 'dynamics'. Understanding the system-level dynamics of networks is an essential step in our desire to map the static network diagrams which are merely "*anatomical* observables", into the *functional* observable of cell behavior. Dynamics is introduced by considering that a given gene (node) has a time-dependent expression value $x_i(t)$ which in a first approximation represents the activity state of gene i (an active gene is expressed and post-translationally activated). Instead of seeing dynamics as a sequence of gene activation, epitomizing the chain of causation of the gene-centric view (as outlined in Sect. "Introduction") a goal in the study of complex systems is to understand the integrated, "emergent" behavior of systems as a holistic entity. We thus define a system (network) state $S(t)$ as $S(t) = [x_1(t), x_2(t), \ldots, x_N(t)]$ which is determined by the expression values x_i of all the N genes of a network which change over time. It is obvious that not all theoretically possible state configurations S can be realized since the individual variables x_i cannot change independently. This global constraint on S imposed by the gene regulatory interactions ultimately determines the "system behavior" which is equivalent to whole-cell phenotype changes, such as switching between cell phenotypes. Thus, one key question is: given a particular network architecture, what is the dynamics of S, and does it reproduce typical cell behaviors, or even predict the particular behavior of a specific cell? Before plunging into complex GRNs, let us introduce basic concepts of modeling dynamics using small circuits (Fig. 2, B1 in Table 1.).

Small Gene Circuits

Basic Formalism The dynamics of the network can be written as a system of ordinary differential equations (ODE) that describe the rate of change of x_i as a function of the state of the all the x_j:

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{F}(\mathbf{x}) \,, \tag{1}$$

where $\mathbf{x}$ is the state vector $\mathbf{x}(t) = [x_1, x_2, \ldots, x_N]$ for a network of N genes, and $\mathbf{F}$ describes the interactions (including the interaction matrix), defining how the components influence each other. Concretely, for individual genes in small circuits, e. g. of two genes, $N = 2$:

$$\begin{aligned} \frac{\mathrm{d}x_1}{\mathrm{d}t} &= f_1(x_1, x_2) \\ \frac{\mathrm{d}x_2}{\mathrm{d}t} &= f_2(x_1, x_2) \,. \end{aligned} \tag{1a}$$

Here the functions f_i are part of the network "architecture" that determine how the inputs onto gene i jointly determine its dynamical behavior – as defined in Sect. "Introduction". An example of f is shown in Fig. 2b. Its form is further specified by *system parameters* that are constants for the time period of observation. The solution of these system equations is the movement of the state vector $\mathbf{x}$ in the N-dimensional ***gene expression state space*** spanned by x_i.

$$\mathbf{x}(t) = S(t) = [x_1, x_2, \ldots, x_N] \,. \tag{2}$$

The Network State For a larger number of genes, it is convenient to describe the dynamics of the network with the vector $\mathbf{x}$ which roughly represents the system state S, Eq. (2), that biologists measure using DNA microarrays and is known as the "*gene expression profile*" (see Sect. "Definition of the Subject", Fig. 1). We will thus refer to S instead of the vector x in discussing biological sys-

tem states. Since Eq. (1) is a first order differential equation, S specifies a system state at time t which for $t = 0$ represents an ***initial condition***. In other words, unlike in macroscopic mechanical systems with inertia, in which the velocity dx_i/dt of the "particles" is also important to specify an initial state, in cells a photographic snapshot of all the positions of x_i, that is $S(t)$ itself, specifies the system state. (This will become important for experimental monitoring of network dynamics).

Role of Dynamical Models in Biology Small networks, or more precisely, "*circuits*" of a handful of interacting proteins and genes, have long been the object of modeling efforts in cell biology that use ODEs of the type of (1a) to model the behavior of the circuit [109,192] (Table 1, B1). In contrast, the objects of interest in the study of complex system sciences are large, i. e., "complex" networks of thousands of nodes, such as the 3000-node core GRN, mentioned earlier, and has largely been driven by

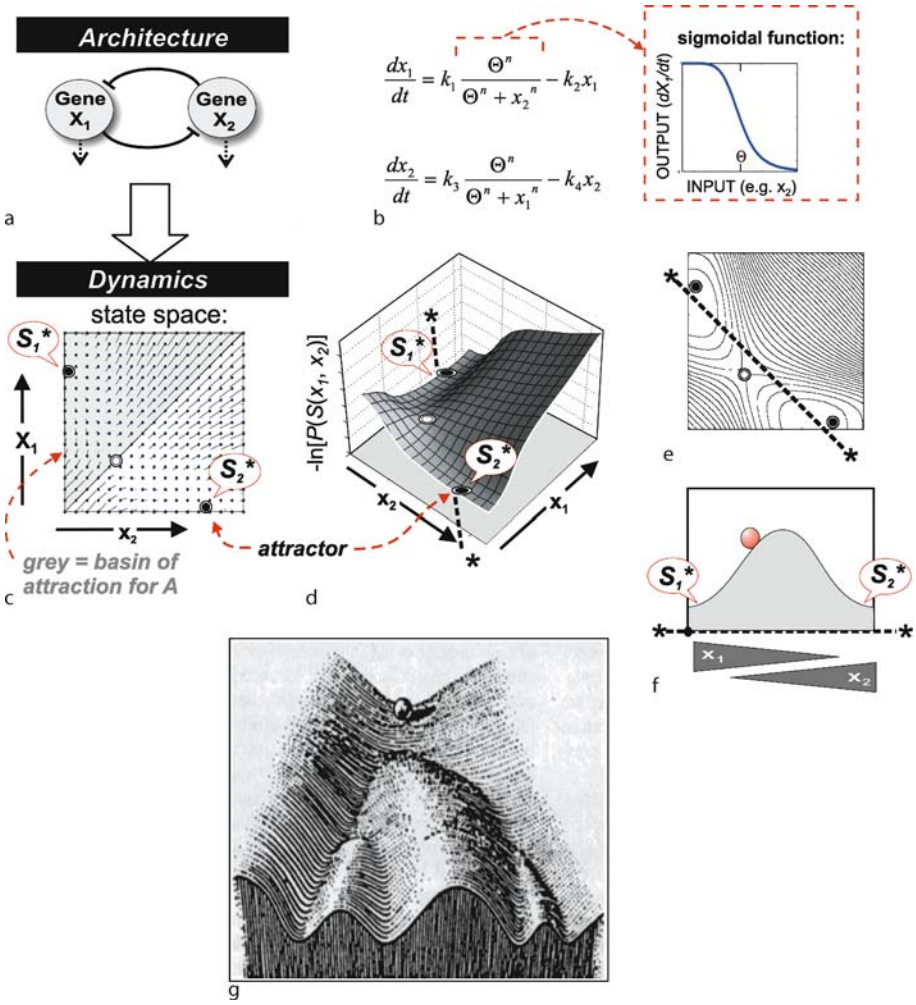

Complex Gene Regulatory Networks – from Structure to Biological Observables: Cell Fate Determination, Figure 2
From gene circuit architecture to dynamics, exemplified on a bistable two-gene circuit. a Circuit architecture: two mutually inhibitory genes X_1 and X_2 which are expressed at constant rate and inactivated with first-order kinetics ("bistable toggle switch"). b Typical ODE system equations for the circuit in a (see [71,109]). c State space (x_1–x_2-phase-plane). *Each dot* is an example initial state $S_o = [x_1(t = 0), x_2(t = 0)]$ with the emanating tick revealing direction and extent that the state S_o would travel in the next time unit Δt. *Solid circles*, S_1^* and S_2^*, denote stable fixed-points ("attractors"). *Empty circle* denotes unstable fixed-point (saddle-node). d, e, f Various schematic representations of the probability $P(S)$ (for a noisy circuit) to find the system in state $S = [x_1, x_2]$. The "elevation" (z-axis over the x_1–x_2 plane) is calculated as $-\ln(P)$; thus, the most probable = stable states are the lowest in the emerging landscape. g Waddington's metaphoric epigenetic landscape, in the 1957 version [198]

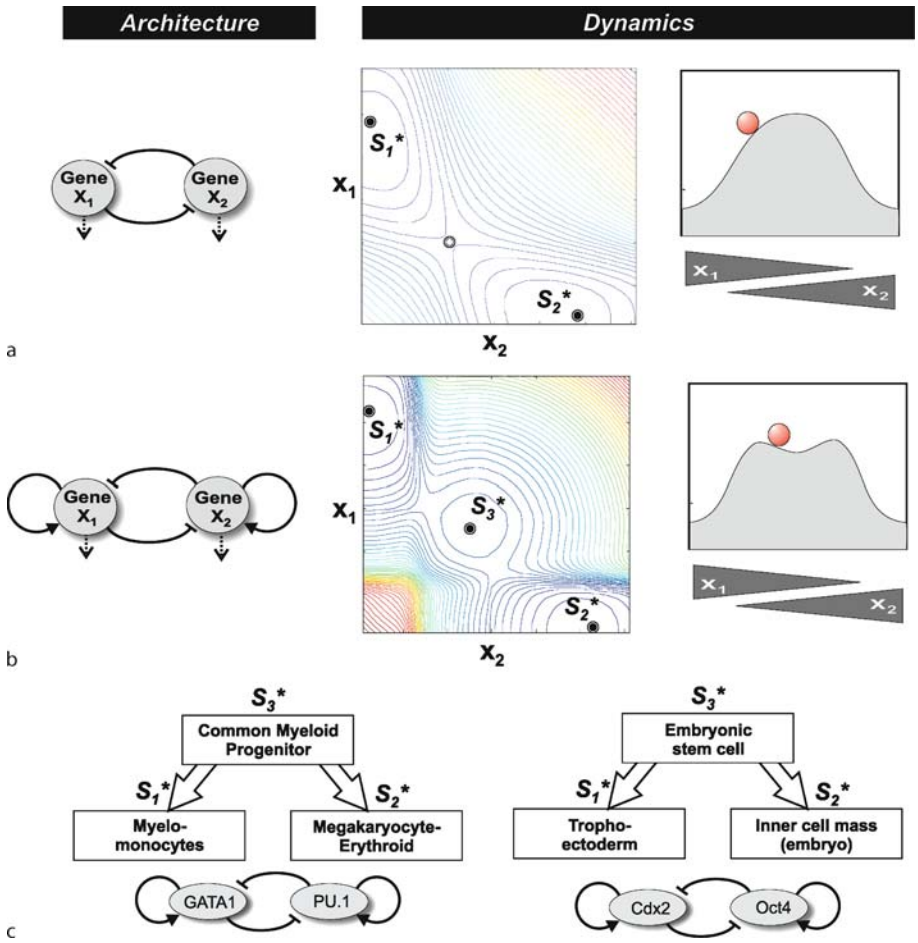

Complex Gene Regulatory Networks – from Structure to Biological Observables: Cell Fate Determination, Figure 3

Expansion of the bistable two-gene circuit by self-stimulation – creating a central metastable attractor. a The bistable circuit of Fig. 2 and associated epigenetic landscape, now as *contour graph*. b If the two genes exert self-activation, then, for a large range of parameters and additive effect of the two inputs that each gene receives, the epigenetic landscape is altered such that the central unstable fixed-point in a. becomes locally stable, S_3^*, giving rise to robust "tristable" dynamics [99]. c Two specific examples of observed gene circuits that represent the circuit of *C*. Such network motifs are typically found to regulate the branch points where multipotent progenitor cells make binary cell fate decisions [99,162]. The metastable central attractor S_3^* can be modeled as representing the metastable bi-potent progenitor cell which is poised in a state of locally stable indeterminacy between the two prospective fate attractors S_1^* and S_2^*

experimental monitoring of $S(t)$ using microarrays since the function $\boldsymbol{F}$ that represents the architecture of the network is not known (Table 1, B3).

What is the actual aim of modeling a biological network? In mathematical modeling of small gene circuits or of signal transduction pathways, one typically predicts the temporal evolution (time course) of the concentration of the modeled variables $x_1(t)$, or $x_2(t)$, and characterizes critical points, such as stable steady states or oscillations in the low-dimensional state space, e. g. in the $x_1 - x_2$ phase plane, after solving the system equations, as exemplified in Figs. 2 and 3. Unknown parameter values have to be estimated, either based on previous reports or obtained by fitting the modeled $x_i(t)$ to the observed time course. Successful prediction of behaviors also serves to *validate* the assumed circuit architecture. From the ***stability analysis*** of the resulting behavior [151], generalization as to the ***dynamical robustness*** (low sensitivity to noise that cause $x(t)$ to randomly fluctuate) and ***structural robustness*** (preservation of similar dynamical behavior even if network architecture is slightly rewired by mutations) can be made [7,34]. Both types of robustness of the system shall not be confounded with the robustness which is sometimes encountered in the analysis of network topol-

ogy where preservation of some graph properties (such as global connectedness) upon deletion of nodes or links is examined ("error tolerance") (as discussed in Sect. "Network Architecture").

Although in reality most genes and proteins are embedded in the larger, genome-wide network, small, idealized circuit models which implicitly accept the absence of many unknown links as well as inputs from the nodes in the global network outside the considered circuit often surprisingly well predict the observed kinetics of the variables of the circuit. This not only points to intrinsic structural robustness of the class of natural circuits, in that circuit architectures slightly different from that of the real system can generate the observed dynamics. But it also suggests that obviously, for some not well-understood reasons, local circuits are to some extent dynamically insulated from the larger network in which they are embedded although they are topologically not detached.

Such "functional modularity" may be a property of the particular architecture of the evolved complex GRNs. In fact, in models of complex networks (discussed in the next section), evidence for just such "functional modularity" has been found with respect to network dynamics [116]. In brief preview (see Sect. "Three Regimes of Behaviors for Boolean Networks"), work on random Boolean network models of GRN has revealed three "regimes" of dynamical behaviors: ordered, critical, and chaotic, as described in Sect. "Ordered and Chaotic Regimes of Network Behavior". In the ordered and critical regimes, many or most genes become "frozen" into "active" or "inactive" states, leaving behind *functionally isolated "islands" of genes* which, despite being connected to other islands through the interactions of the network, are free to vary their activities without impacting the behaviors of other functionally isolated islands of genes.

Complex Networks

Despite a recent spate of publications in which the temporal change of expression of individual genes have been predicted based on small circuit diagrams, such predictions do not provide understanding of the integrated cell behavior, such as the change of cell phenotype that may involve thousands genes in the core GRN. Thus, analysis of genome-wide network state is needed to understand the biological observable. In a first approximation, it is plausible that the state of a cell, such as the particular cell (pheno-)type in a multi-cellular organism, is defined by its genome-wide gene expression profile, or *transcriptome* $T = [x_1, x_2, \ldots, x_N]$ with $N = 25,000$. In fact, microarray analysis of various cells and tissues reveals globally distinct, tissue specific patterns of gene expression profiles that can easily be discerned as shown in Fig. 1f. As mentioned above (Sect. "Definition of the Subject"), the gene expression profile across the genome is governed by the core regulatory network of transcription factors (TFs) which enslave the rest of the genome. Thus, in our approximation the network state S of the core transcriptional network of 3000 or so genes essentially controls the entire (genome-wide) gene expression profile.

For clarity of formalization, it is important to note that one genome in principle encodes exactly one fixed network, since the network connections are defined by the specific molecular interactions between the protein structure of TFs and the DNA sequence motif of the cis-regulatory promoter elements they recognize. Both are encoded by the genomic sequence. The often encountered notion that "networks change during development" and that "every cell type has its own network" is in this strict formalism incorrect – the genes absent in one cell type must directly or indirectly have been repressed (and sometimes, continuously kept repressed) by other genes that are expressed. Thus, the genome (of one species) directly maps into one (time-invariant) network architecture which in turn can generate a variety of states $S(t)$. It is the state S that changes in time and is distinct in different cell types or in different functional states within one cell type. Only genetic mutations in the genome will "rewire" the network and change its architecture.

If the network state $S(t)$ of the core GRN directly maps into a state of a cell (the biological observable), then one question is: What is the nature of the integrated dynamics of S in complex, irregularly wired GRN and is it compatible with observable cell behavior?

The study of the dynamics of complex networks (Table 1, B) may at first glance appear to be impeded by our almost complete ignorance about the architecture and interaction modalities between the genes. However, we cannot simply extrapolate from the mindset of studying small circuits with rate equations to the analysis of large networks. This is not only impossible due to the lack of information about the detailed structure of the entire network – the function **F** in Eq. 1 is unknown – but it may also be numerically hard to do. Yet, despite our ignorance about the network architecture, much can be learned if we reset our focus plane to the larger picture of the network. In this regard, the study of the dynamics of complex network can have two distinct goals (see Table 1).

One line of research (Table 1, B2) overcomes the lack of information about the network architecture by taking an ensemble approach [112] and asks: *What is the typical behavior of a statistical ensemble of networks that is*

characterized by an architectural feature (e. g. average connectivity k_i, power-law exponent γ)? This computationally intense approach typically entails the use of discrete-valued gene networks (Sect. "Boolean Networks as Model for Complex GRNs"). As will be discussed below, such analysis has led to the definition of three broad classes of behaviors: chaotic, critical and ordered.

The second approach (Table 1, B3) to the dynamics of complex GRN is closer to experimental biology and exploits the availability of gene expression profile measurements using DNA microarray technology. Such measurements provide snapshots of the state of the network $S(t)$ over N genes, covering almost the entire transcriptome, and thus, reveals the direct output of the GRN as a distributed control system. Monitoring $S(t)$ and its change in time during biological processes at the whole-cell level will reveal the constraint imposed on the dynamics of $S(t)$ by network interactions and can, in addition to providing data for the inference problem (Sect. "Network Inference"), expose particular dynamical properties of the transcriptome that can be correlated with the biological observable.

Cell Fates, Cell types: Terminology and Concepts

In order to appreciate the meaning of the network state S and how it maps to the biological observable, we will now present (to non-biologists) in more detail the most prosaic biological observable of gene network dynamics: cell fate determination during development.

Stem Cells and Progenitor Cells in Multi-cellular Organism

A hall-mark of multi-cellular organisms is the differentiation of the omnipotent zygote (fertilized egg) via *totipotent* and *pluripotent embryonic stem cells* and *multipotent tissue stem cells* into functionally distinct mature "cell types" of the adult body, such as red blood cell, skin cells, nerve cells, liver cells, etc. This is achieved through a branching (tree-like) scheme of successive specialization into lineages. If the fertilized egg is represented by the main trunk of the "tree of development", then think of cells at the branching points of developmental paths as the stem cells. One example of a multipotent stem cell is the hematopoietic stem cells (HSC) which is capable of differentiating into the entire palette of blood cells, such as red blood cells, platelets and the variety of white blood cells. The last branch points represent *progenitor* cells which have a lesser developmental potential but still can chose between a few cell types (e. g., the common granulocyte-macrophage progenitor (GMP) cell, Fig. 1f). Finally, the outmost branches of the tree represent the mature, terminally differentiated cell types of the body.

A cell that can branch into various lineages is said to be "multipotent". It is a "*stem cell*" when it has the potential to self-renew by cell division, maintaining its differentiation potential, and to create the large family of cells of an entire tissue (e. g., the hematopoietic stem cell). Thus, progenitor cells, which proliferate but cannot infinitively self-renew, are strictly not stem cells. The commitment to a particular cell phenotype (a next generation branch of the tree) is also referred to as a "cell fate" since the cell at the proximal branching point is "fated" to commit to one of its prospective cell types.

Development and Differentiation

The diversification of the embryonic stem cell to yield the spectrum of the thousands [94] or so cell types in the body occurs in a process of successive branching events, at which multipotent cells commit to one fate and which appear to be binary in nature. Thus, multipotent cells make an either-or decision between typically two lineages – although more complex schemes have been proposed [65]. Moreover, it is generally assumed that during natural development there is only ***diversification*** of developmental paths but no *confluence* from different lineages, although recently exceptions to this rule have been reported for the hematopoietic system [1].

As cells develop towards the outer branches of the "tree of development", they become more and more specialized and progressively lose their competence to proliferate and diversify ("potency"). They also develop the phenotype features of a mature cell type; for instance, in the case of red blood cells, they adopt the flat, donut-like shape and synthesize haemoglobin. This process is called ***differentiation***. Most cells then also loose their capacity to divide, that is, to enter into the proliferative state. Thus, mature, terminally differentiated cells are typically quiescent or "*post mitotic*".

The branching scheme of cell types imposes another fundamental property: cell types are *discrete* entities. They are distinct from each other, i. e., they are well separated in the "phenotype space" and are stable [171]. There is thus no continuum of phenotype. As Waddington, a prominent embryologist of the last century, recognized in the 1940s: Cell types are "well-recognisable" entities and "intermediates are rare" [198]. In addition to the (quasi-) discreteness between branches of the same level, discreteness between the stages within one developmental path is also apparent: a multipotent stem cell at a branching point is in a discrete stage and can be identified based on molecular markers,

isolated and cultured as such. Hence, a "stem cell" is not just a snapshot of an intermediate stage within a continuous process of development, but a discrete metastable entity.

The flow down the developmental paths, from a stem cell to the terminally differentiated state is, despite the pauses at the various metastable stem and progenitor cell levels, essentially unidirectional and, with a few exceptions, irreversible. In some tissues, the mature cells, as is the case with liver, pancreas or endothelial cells, can upon injury revert to a phenotype similar to that of the last immature (progenitor) stage and resume proliferation to restore the lost cell population, upon which they return to the differentiated, quiescent state.

Biologists often speak in somewhat loose manner of cells "***switching***" their phenotype. This may refer to switching from a progenitor state to a terminally differentiated state (differentiation) or within a progenitor state from the quiescent to other functional states, such as the proliferative or the apoptotic state (apoptosis = programmed cell death). In any case, such intra-lineage switching between different functional states also represents discontinuous, quasi-discrete transitions of whole-cell behaviors. The balance between division, differentiation and death in the progenitor or stem cell compartment of a tissue thus consists of state transitions that entail all-or-none decisions. This balance is at the core of organismal development and tissue homeostasis.

Now we can come back to the network formalism: if the network state S maps directly into the biological observable, what are the properties of the network architecture that confer its ability to produce the properties of the biological observable outlined in this section: discreteness and robustness of cell types, discontinuity of transitions, successive binary diversification and directionality of these processes? Addressing these questions is the long-term goal of a theory of the multicellular organism. In the following we describe the status of research toward this goal.

History of Explaning Cell Types

Waddington's Epigenetic Landscape and Bistable Genetic Circuits

One of the earliest conceptualization of the existence of discreteness of cell types was the work of C. Waddington, who proposed "epigenetic regulation" in the 1940s, an idea that culminated in the famous figure of the "epigenetic landscape" (Fig. 2g). This metaphor, devoid of any formal basis, let alone relationship to gene regulation, captures the basic properties of discrete entities and the instability of intermediates. The term 'epigenetic' was coined by Waddington to describe distinct biological traits that arise from the interplay of the same set of genes and does not require the assumption of a distinct, "causal" gene to be explained [196]. The 1957 version of the epigenetic landscape [198] (Fig. 2g) also implies that multipotent cells are destined to make either-or decisions between two prospective lineages, as embodied in the "watershed" regions (Fig. 2g).

Almost at the same time as Waddington, in 1948 Max Delbrück proposed a generic concept of differentiation into two discrete states in a biochemical system that can be described by equations of the form (3), consisting of two mutual inhibiting metabolites, $x_1(t)$ and $x_2(t)$ that exhibits bistability [53] (For a detailed qualitative explanation, see Fig. 2). The dynamics of such a system is graphically represented in the two dimensional state space spanned by x_1 and x_2 (Fig. 2c-f). ***Bistable dynamics*** implies that there are two steady states S_1^* and S_2^* that satisfy $\mathrm{d}x_1/\mathrm{d}t = \mathrm{d}x_2/\mathrm{d}t = 0$ and are *stable* fixed-points of the system. For the system equations of Fig. 2b, this behavior is observed for a large range of parameters. In a nutshell, the mutual inhibition renders the balanced steady state $S_3^*(x_1 = x_2)$ unstable so that the system settles down in either the steady state $S_1^*(x_1 \gg x_2)$ or $S_2^*(x_1 \ll x_2)$ when kicked out of this unstable fixed point. These two stable steady states and their associated gene activity patterns are discretely separated in the $x_1 - x_2$ state space and have been postulated to represent the differentiated state of cells, thus, corresponding to Waddington's valleys. This was the first conceptualization of a cellular differentiated state as a stable fixed-point of a non-linear dynamical system. Soon after Monod and Jacob discovered the principle of gene regulation they also proposed a circuit of the same architecture as Delbrück [148], but consisting of two mutually suppressing genes instead of metabolites, to explain differentiation in bacteria as a bistable system.

Bistability, Tristability and Multistability

The central idea of bistability is that the very same system, in this case, a gene circuit composed of two genes x_1 and x_2, can under some conditions produce two distinct stable states separated by an unstable stationary-state, and hence, it can switch in almost discontinuous manner from one state (S_1^*) to another (S_2^*) ("toggle switch") (Fig. 2). Bistability is a special (the simplest) case of ***multi-stability***, an elementary (but not necessary) property of systems with non-linear interactions, such as those underlying the gene regulatory influences as detailed in Fig. 2b. Which stable state (in this case, S_1^* or S_2^*) a network occupies depends

on the ***initial condition***, the position of $S(t = 0)$ in the $x_1 - x_2$ state space (Fig. 2c). The two stable steady-states, S_1^* or S_2^*, are called "***attractors***" since the system in such states will return to the characteristic gene expression pattern in response to small ***perturbations*** to the system by enforced change of the values of x_1 or x_2. Similarly, after an external influence that places the network at an unstable initial state $S'(t = 0)$ within the ***basin of attraction*** of attractor S_1^* (gray area in Fig. 2c), and hence cause the network to settle down in the stable state S_1^*, the network will stay there even after the causative influence that initially put it in the state S' has disappeared. Thus, attractor states confer memory of the network state. In contrast, larger perturbations beyond a certain threshold will trigger a transition from one attractor to another, thus explaining the observed discontinuous "switch-like" transitions between two stable states in a continuous system.

After Delbrück, Monod and Jacob, numerous theoretical [74] and, with the rise of systems biology, experimental works have further explored similar simple circuitries that produce bistable behavior (see ref. in [38]). Such circuits and the predicted switch-like behavior have been found in various biological systems, including in gene regulatory circuits in Escherichia coli [157], mammalian cell differentiation [43,99,127,168] as well as in protein signal transduction modules [12,64,205]. Artificial gene regulatory circuits have been constructed using simple recombinant DNA technology to verify model predictions [22,71,123].

Circuit analyzes have been expanded to cover more complex circuits in mammalian development. It appears that there is a common theme in the circuits that govern cell differentiation (Sect. "Cell Fates, Cell Types: Terminology and Concepts"): interconnected pairs of mutually regulating genes that are often also self-regulatory, as shown in Fig. 3b [43,99,168]. Such circuit diagrams may be crucial in controlling the binary diversification at developmental branch points (Fig. 1f) [99]. In the case where the two mutually inhibiting genes are also are self-stimulatory, as summarized in Fig. 3b, an interesting modification of the bistable dynamics can be obtained. Assuming independent (additive) influence of the self-stimulatory and cross-inhibitory inputs, this circuit will convert the central unstable state (saddle) S_3^* (Fig. 2, 3a) that separate the two stable steady-states into a stable steady state, thus generating ***tristable*** behavior [99]. The third, central stable fixed point has, in symmetrical cases, the gene expression configuration $S_3^*[x_1 \sim x_2]$ (Fig. 3b).

The *promiscuous* expression of intermediate-low levels of x_1 *and* x_2 *in the locally stable state* S_3^* has been associated with a stem or progenitor cell that can differentiate into the cell represented by the attractors $S_1^*(x_1 \gg x_2)$ or $S_2^*(x_1 \ll x_2)$. In fact, the common progenitor cells (Fig. 1f) have been shown to express "promiscuous" expression of genes thought to be specific for either of the lineages in which it will have to commit to, so-to speak providing a "preview" or "multi-lineage priming" of the gene expression of its prospective cell fates [47,59,92]. For instance, the *common myeloid progenitor* (CMP, Fig. 1) which can commit to either the *erythroid* or the *myeloid* lineage, expresses both the erythroid-specific transcription factor GATA1 (= x_1) and the myeloid specific transcription factor PU.1 (= x_2) at intermediate levels (Fig. 3c). The metastable [GATA1∼PU.1] configuration generates a state of *indeterminacy* or "suspense" [146] awaiting the signal, be it a biological instruction or stochastic perturbation, that resolves it to become either one of the more stable states [GATA1 ≫ PU.1] or [GATA1 ≪ PU.1] when cells differentiate into the erythroid or myeloid lineage, respectively. Thus, multi-potency and indeterminacy of a progenitor cell can be defined purely dynamically and does not require a much sought after "stemness" gene (again, a concept that arose from the gene-centered view) [195]. The metastable state also captures the notion of a "higher potential energy" [79] that drives development and hence, may account for the arrow of time in ontogenesis.

A Formalism for Waddington's Epigenetic Landscape Metaphor

Obviously, the valleys in Waddington's epigenetic landscape represent the stable steady states (attractors) while the hill-tops embody the unstable fixed-points, as shown in Fig. 2. How can Waddington's metaphor formally be linked to gene network dynamics and the state space structure? For a one dimensional system $\mathrm{d}x/\mathrm{d}t = f(x)$, this is easily shown with the following analogy: An "energy" landscape represents the cumulative "work" performed/received when "walking" in the state space against / along the vector field of the "force" $f(x)$ (Fig. 2c). Thus, the "potential energy" is obtained by integrating $f(x)$ (the right-hand side of the system equation, Eq. (1)) over the state space variables x.

$$V(x) = -\int f(x)\mathrm{d}x \,. \tag{3}$$

Here the state space dimension x is given the meaning of a physical space, as that pertaining to a landscape, and the integral $V(x)$ is the sum of the "forces" experienced by the network states $S(t) = x(t)$ over a path in x that drive $S(t)$ to the stable states (Fig. 2c). The neg-

ative sign establishes the notion of a potential in that the system loses energy as it moves towards the stable steady states which are in the valleys ("lowest energy"). Higher ($N > 1$) dimensional systems (Eq. 1) are in general non-integrable (unless there exists a continuously differentiable (potential) function $V(x_1, x_2, \ldots, x_N)$ for which $f_1 \mathrm{d}x_1 + f_2 \mathrm{d}x_2 + \cdots + f_N \mathrm{d}x_N \cdots = 0$ is the *exact total differential*, so that $-grad(V)=\mathbf{F}(\mathbf{x})$ with $\mathbf{F}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_N(\mathbf{x})]^{\mathrm{T}}$). Thus, there is in general no potential function that could represent a "landscape"; however, the idea of some sort of "potential landscape" is widely (and loosely) used, especially in the case of two-dimensional systems where the third dimension can be depicted as a cartographic elevation $V(x_1, x_2)$. An elevation function $V(x_1, x_2)$ can be obtained in stochastic systems where $\mathbf{x}(t)$ is subjected to random fluctuations due to gene expression noise [87]. Then $V(S)$ is related to the probability $P(S)$ to find the system in state $S = (x_1, x_2, \ldots)$, e. g., $V(S) = -\ln(P(S)$ [118,177]. It should however be kept in mind that the "quasi-potential" V is not a true (conservative) potential, since the vector field is not a conservative field.

The above treatment of the dynamics of the gene regulatory circuit explains the valleys and hills of Waddington's landscape but still lacks the "directionality" of the overall flow on the landscape, depicted by Waddington as the slope from the back to the front of his landscape. (This *arrow or time of development* will briefly be discussed in the outlook Sect. "Future Directions and Questions").

In summary, the epigenetic landscape that Waddington proposed based on his careful observation of cell behavior and that he reshaped over decades [103,179,196,197,198], can now be given both a molecular biology correlate (the gene regulatory networks) and a formal framework (probability landscape of network states S). The landscape idea lies at the heart of the connection between molecular network topology and biological observable.

The Molecular Biology View: "Epigenetic" Marks of Chromatin Modification

Although it is intuitively plausible that stable steady states of the circuits of Delbrück and of Monod and Jacob may represent the stable differentiated state, this explanation of a biological observable in terms of a dynamical system was not popular in the community of experimental molecular biologists and was soon sidelined as molecular biology, and with it the gene-centered view, came to dominate biology. The success in identifying novel genes (and their mutated alleles) and the often straightforward explanation of a phenotype offered by the mere discovery of a (mutant) gene triggered a hunt for such "explanatory genes" to account for biological observables. Gene circuits had to give place to the one-gene-one trait concept, leading to an era in which a new gene-centered epistemological habit, sometimes referred to as "*genetic determinism*", prevailed [180]. Genetic determinism, a particular form of reductionism in biology, was to last for another fifty years after Delbrück's proposal of bistability. Only as the "low hanging fruits" of simple genotype-phenotype relationships seemed to have all been picked, and genome-wide gene expression measurements became possible was the path cleared for the rise of "system biology" and biocomplexity that we witness today.

In genetic determinism, macroscopic observables are reduced in a qualitative manner to the existence of genes or regulatory interactions which at best "form" a linear chain of events. Such "pathways", most lucidly manifest in the arrow-arrow schemes of modern cell biology papers, serve as a mechanistic explanation and satisfied the intellectual longing for causation. It is in light of this thinking that the molecular biologists' explanation of the phenomenon of cell type determination has to be seen. Absent a theory of cell fate diversification and in view of the stunning stability of cell types, a dogma thus came into existence according to which the type identity of cells, once committed to, is irreversibly carved in stone [161]. Rare transdifferentiation events (switch of cell lineages) were regarded as curiosities. The observation that cell types express type-specific genes was explained by the presence of cell-type specific transcription factors. For instance, red blood cells express haemoglobin because of the presence of GATA1 – a lineage specific transcription factor that not only promotes commitment to the erythroid lineage (as discussed above, Sect. "Bistability, Tristability and Multistability") but also controls haemoglobin expression [149]. Conversely, the absence of gene activity, for instance, non-expression of liver-specific genes in erythrocytes, was explained by the silencing of the not needed genes by covalent DNA methylation and histone modifications (methylation, acetylation, etc.) [13,65,69,117,122] which modify chromatin structure, and thereby, control the access of the transcription machinery to the regulatory sites on the DNA. But who controls the controller?

Chromatin modifications [117,122] are thought to confer discrete alterations of gene expression state that are stable and essentially irreversible. This idea of permanent marks on DNA represents the conceptual cousin of mutations (but without altering DNA sequence) and was thus in line with the spirit of genetic determinism. Accordingly, they were readily adopted as an explanation of the apparently irreversible cell type-specific gene inactivation

and were given the attribute "epigenetic" to contrast them from the genetic changes that involve alteration of DNA sequences. But enzymes responsible for covalent DNA and chromatin modification are not gene-locus specific, leaving open how the cell type specific gene expression *pattern* is orchestrated.

It is important to mention here a disparity in the current usage of the term "epigenetics" [103]. In modern molecular biology, "epigenetics" is almost exclusively used to refer to DNA methylation and covalent histone modifications; this meaning is taken for granted even among authors who comment on the very usage of this term [13,26,103], and are unaware that memory effects can arise purely dynamically without a distinct material substrate, as discussed in Sect. "Bistability, Tristability and Multistability". In contrast, biological physicists use "epigenetic" to describe precisely phenomena, such as multi-stability (Sect. "Bistability, Tristability and Multistability") that are found in non-linear systems – a usage that comes closer to Waddington's original metaphor for illustrating the discreteness and stability of cell types [197,198]. The use of Waddington's "epigenetic landscape" has recently seen a revival [75,165] in the modern literature in the context of chromatin modifications but remains loosely metaphoric and without a formal basis.

Rethinking the Histone Code

The idea of methylation and histone modifications as a second code beyond the DNA sequence ("histone code") that cells use to "freeze" their type-specific gene expression pattern relied on the belief that these covalent modifications act like permanent marks (only erased in the germline when an oocyte is formed). This picture is now beginning to change.

First, recent biochemical analysis suggest that the notion of a static, irreversible "histone code" is oversimplified, casting doubt on the view that histone modification is the molecular substrate of "epigenetic memory" [122,126,144,191]. With the accumulating characterization of chromatin modifying enzymes, notably those controlling histone lysine (de)methylation [122,126,144,191], it is increasingly recognized that the covalent "epigenetic" modifications are bidirectional (reversible) and highly dynamic. Second, cell fate plasticity, most lucidly evident in the long-known but rarely observed transdifferentiation events, or in the artificial reprogramming of cells into embryonic stem cells either by nuclear transfer-mediated cloning [91] or genetic manipulation [143,156,184,201], confirm that the "epigenetic" marks are reversible – given that the appropriate biochemical context is provided. If what was thought of as permanent molecular marks is actually dynamical and reversible – what then maintains lineage-specific gene expression patterns in an inheritable fashion across cell divisions?

In addition, as mentioned above, there is another, more fundamental question from a conceptual point of view: chromatin-based marking of gene expression status is a generic, not locus – specific process – the same enzymes can apply or remove the covalent marks on virtually any gene in the genome. They are dumb. So what smart system orchestrates the DNA methylation and histone modification machinery at the tens of thousands of loci in the genome so that the appropriate set of genes is (in)activated to generate the cell type-specific patterns of gene expression?

A system-level view avoids the conundrum caused by the mechanistic, proximal explanation [189] of the gene-centered view. A complex systems approach led to the idea that the genome-wide network of transcriptional regulation can under some conditions spontaneously orchestrate, thanks to self-organization, the establishment of lineage- specific gene expression profiles [114], as will be discussed below. In fact, the picture of chromatin modification as primum movens that operates "upstream" of the transcription factors (TFs), controlling their access to the regulatory elements in promoter regions, must be revised in light of a series of new observations. Evidence is accumulating that the controller itself is controlled – namely, by the TFs they are thought to control: TFs may actually take the initiative by recruiting the generic chromatin-modifying enzymes to their target loci [51,80,125,144,145,182,185]. It is even possible that a mutual, cooperative dynamical interdependence between TFs and chromatin-modifying enzymes may establish locus-specific, switch-like behavior that commands "chromatin status" changes [56,132]. In fact, an equivalent of the indeterminacy state where both opposing lineage specific TFs balance each other (Sect. "Bistability, Tristability and Multistability", Fig. 3b) is found at the level of chromatin modification, in that some promoters exhibit "bivalent" histone modification in which activating and suppressing histone methylations coexist [25]. Such states in fact are associated with TFs expressed at low level – in agreement with the central attractor S_3^* of the tristable model (Fig. 3b). Thus, chromatin modification is at least in part "downstream" of TFs and may thus act to add additional stability to the dynamical states that arise from the network of transcriptional regulation. If correct, such a relationship will allow us to pass primary responsibility for establishing the observable gene expression patterns back to transcription factors. With its genome-wide regulatory connections the

GRN is predestined for the task of coordination and distributed information processing.

But under what conditions can a complex, apparently randomly wired network of 3000 regulators create, with such stunning reliability and accuracy, the thousands of distinct, stable, gene expression profiles that are associated with a meaningful biological state, such as a cell type? Studies of Boolean networks as toy models in the past 40 years have provided important insights.

Boolean Networks as Model for Complex GRNs

General Comment on Simplifying Models

The small-circuit model discussed in Sect. "Small Gene Circuits" represents arbitrarily cut-out fragments of the genome-wide regulatory network. A cell phenotype, however, is determined by the gene expression profile over thousands of genes. How can we study the entire network of 25,000 genes, or at least the core GRN with 3000 transcription factors even if most of the details of interactions remain elusive?

The use of random Boolean networks as a generic network model, independent of a specific GRN architecture of a particular species, was proposed by Kauffman in 1969 – at the time around which the very idea of small gene regulatory circuits were presented by Monod and Jacob [148]. In random Boolean networks the dynamics is implemented by assuming discrete-valued genes that are either ON (expressing the encoded protein) or OFF (silenced). The interaction function (Sect. "Introduction", and **F** in Sect. "Small Gene Circuits") that determines how the ON/OFF status of multiple inputs of a target gene map into its behavior (output status) is a logical (Boolean) function B, and the network topology is randomly-wired, with the exception of some deliberately fixed features. Thus, the work on random Boolean networks allowed the study of the generic behavior of large networks of thousands of genes even before molecular biology could deliver the actual connections and interaction logics of the GRN of living systems (for review, see [8]).

The lack of detailed knowledge about specific genes, their interaction functions and the formidable computational cost for modeling genome-wide networks has warranted a coarse-graining epitomized by the Boolean network approach. But more specifically, in the broader picture of system-wide dynamics, the discretization of gene expression level is also justified because (*i*) the above discussed steep sigmoidal shape of "transfer functions" (Fig. 2b) that describe the influence of one gene onto another's expression rate can be approximated by a step-function and/or (*ii*) the local dynamics produced by such small gene circuit modules is in fact characterized by discontinuous transitions between discrete states as shown in Sect. "Network Dynamics".

In addition, the Boolean network approach offers several advantages over the exhaustive, maximally detailed models favored by engineers who seek to understand a *particular instance* of a system rather than *typical* properties of a *class* of systems. The simplification opens a new vista onto fundamental principles that may have been obscured by the details since, as philosophers and physicists have repeatedly articulated, there is no understanding without simplification, and "less can be more" [11,30,159]. An important practical advantage of the simplification in the Boolean network approach is the possibility to study *statistical ensembles* of ten thousands of network instances, i. e., entire classes of network architectures, and to address the question of how a particular architecture type maps into a particular type of dynamic behavior. Some of the results obtained in fact are valid for both discrete and continuous behavior of the network nodes [17].

Model Formalism for Boolean Networks

In the Kauffman model of random Boolean networks gene activity values are binary (1 = ON, and 0 = OFF) [111,115,116] and time is also discretized. Thus, a Boolean network is a generalized form of cellular automata but without the aspect of physical space and the particular neighborhood relations. Then, in analogy to continuous systems, a network of N elements i ($i = 1, 2, \ldots, N$), defines a network state S at any given discrete time step t: $\boldsymbol{S}(t) = [x_1(t), x_2(t), \ldots, x_N(t)]$ where x_i is the activity status that now only takes the values 1 or 0. The principles are summarized in Fig. 4. The state space that represents the entire dynamics of the network is finite and contains 2^N states. However, again, not all states are equally likely to be realized and observed, since genes do not behave independently but influence each other's expression status.

Regulation of gene i by its incoming network connections is modeled by the Boolean function B_i that is associated with each gene i and maps the configuration of the activity status (1 or 0) of its input genes (upstream regulators of gene i) into the new value of x_i for the next time point. Thus, the argument of the Boolean function B_i is the input vector $\boldsymbol{R}_i(t) = [x_1(t), x_2(t), \ldots, x_{ki}(t)]$, where k_i is the number of inputs that the gene i receives. At each time step, the value of each gene is updated: $x_i(t+1) = B_i[\boldsymbol{R}_i(t)]$. The logical function B_i can be formulated as a "truth table" which is convenient for large k's (Fig. 4a). In the widely studied case where each gene has exactly $k = 2$ inputs, the Boolean function can be one of

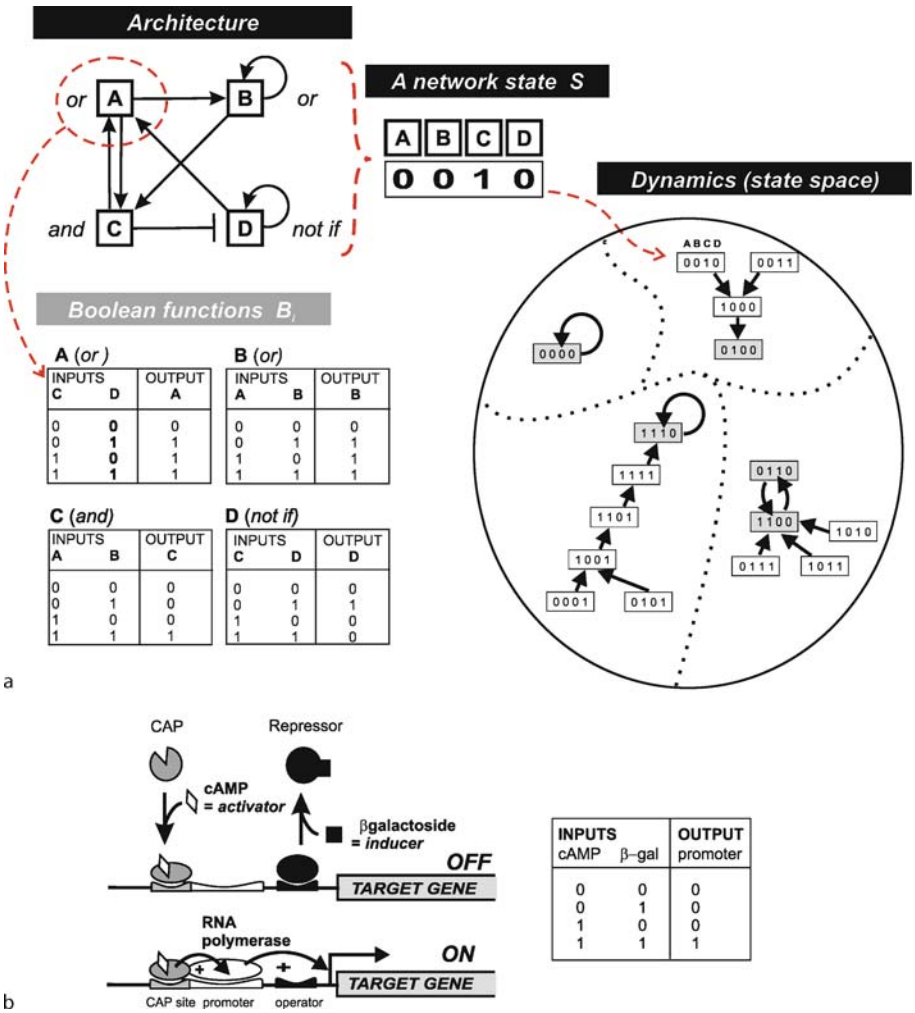

Complex Gene Regulatory Networks – from Structure to Biological Observables: Cell Fate Determination, Figure 4

Principles of Boolean network models of GRN, exemplified on a $N = 4$ gene network. a From *architecture* to dynamics. The four genes i: A, B, C, D interact as indicated by the directed graph (*top left panel*), and each of the genes is assigned a Boolean function B_i as indicated, with the corresponding "truth table" shown below. A network state is a string of $N = 4$ binary variable, thus there are $2^N = 2^4 = 16$ possible *network states S*. They collectively establish the entire state space and can be arranged in a *state transition map* according to the state transitions imposed by the Boolean functions (*right panel*). The attractor states are colored *gray*. In the example, there are three point attractors and one limit cycle attractor (of period $T = 2$). The dotted lines in the state space denote the attractor boundaries. b Example for capturing the regulation at the promoter of the lactose operon as an "AND" Boolean function. Note that there are many ways to define what constitute an input – in the case shown, the allosteric ligands cAMP ("activator") and a β-galactoside ("inducer"), such as allolactose, give rise to a two-input Boolean function 'AND'

the set of $2^{(2 \wedge k)} = 16$ classical Boolean operators, such as AND, OR, NOTIF, etc. [109,116]. Figure 4b shows the example of the well-studied lactose operon and how its regulatory characteristics can be captured as an AND Boolean function for the two inputs.

In the simplest model, all genes are updated in every time step. This synchrony is artificial for it assumes a central clock in the cell which is not likely to exist, although some gating processes from oscillations in the redox potential has been reported [120]. The idealization of synchrony, however, facilitates the study of large Boolean networks, and many of the results that have been found with synchronous Boolean networks carry over to networks with asynchrony of updating that has been implemented in various ways [40,72,81,119]. For synchronous networks the entire network state $\boldsymbol{S}$ can also be viewed as being updated by the updating function U: $\boldsymbol{S}(t+1) = U[\boldsymbol{S}(t)]$, where U summarizes all the N Boolean functions B_i. This

facilitates some treatments and is represented in the ***state transition table*** that lists all possible network states in one column and its successor in a second column. This leads to a higher-level, directed graph that represents the entire dynamics of a network as illustrated below.

Dynamics in Boolean Networks

The state transition table captures the entire dynamical behavior of the network and can conveniently be depicted as a ***state transition map***, a directed graph in which a node represents one of the 2^N possible state $S(t)$ (a box with a string or 1, 0s in Fig. 4a. representing the gene expression pattern). Such diagrams are particularly illustrative only for N up to $\sim$10, since they display all possible states S of the finite state space [204]. The states are connected by the arrows which represent individual state transitions and collectively depict the ***trajectories*** in the state space (Fig. 4a, right panel).

Since the Boolean functions are *deterministic*, a state $S(t)$ unambiguously transitions into *one* successor state $S(t+1)$. In contrast, a state can have multiple predecessors, since two or more different states can be updated into the same state. Hence, trajectories can converge but not diverge, i. e., there is no branching of trajectories emanating from one state. This property of "losing information about the ancestry" is essential to the robustness of the dynamics of networks.

In updating the network states over time, $S(t)$ can be represented as a walk on the directed graph of the state transition map. Because of the finiteness of the state space in discrete networks, $S(t)$ will eventually encounter an already visited state, and because of the absence of divergence, the set of successive states will be the same as in the previous encounter. Thus, no matter what the initial state is, the network will eventually settle down in either a set of cycling states (which form a *limit cycle*) or in a single stable state that updates onto itself. Accordingly, these states to which all the trajectories are "attracted" to are the ***attractors*** of the network. They are equivalent to stable oscillators or stable fixed-points, respectively, in the continuous description of gene circuits (Sect. "Bistability, Tristability and Multistability"). In other words, because of the regulatory interactions between the genes, the system cannot freely choose any gene expression configuration S. Again, most network states S in the state space are thus unstable and transition to other states to comply with the Boolean rules until the network finds an attractor. And as with the small, continuous systems (Sect. "Bistability, Tristability and Multistability") the set of states S that "drain" to an attractor state constitute its ***basin of attraction***.

However, unlike continuous systems, Boolean network dynamics do not produce unstable steady-states that can represent the indeterminacy of undecided cell states that correspond to stem cells about to make an either-or decision between two lineages (see Sect. "Bistability, Tristability and Multistability"). Instead, basins of attraction are "disjoint" areas of state space.

Attractors as Cell Types Kauffman proposed that the high-dimensional attractor states represent cell types in metazoan organisms – thus expanding the early notion of steady state in small circuits to networks of thousands of genes [111,116] (Sect. "Bistability, Tristability and Multistability"). This provides a natural explanation for the stability of the genome-wide expression profile that is characteristic of and defines a cell type as well as for the stable coordination of genome-wide gene expression oscillations in proliferating cells that undergo the cell division cycle [202]. The correspondence of attractors in large networks with the cell type specific transcriptome is a central hypothesis that links the theoretical treatment of the dynamics of complex networks with experimental cell biology.

Use of Boolean Networks to Model Real Network Dynamics Owing to their simple structure, Boolean networks have been applied in place of differential equations to model real-world networks for which a rudimentary picture of the topology with only few details about the interaction functions is known. Hereby, individual Boolean functions are assigned to the network nodes either based on best guesses, informed by qualitative descriptions from the experimental literature, or randomly. This approach has yielded surprisingly adequate recapitulation of biological behaviors, indicating that the topology itself accounts for a great deal of the dynamical constraints [4,49,60,62,95,130]. Such studies also have been used to evaluate the dynamical regime (discussed in the next section) of real biological networks [19].

Three Regimes of Behaviors for Boolean Networks

The simplicity and tractability of the Boolean network formalism has stimulated a broad stream of investigations that has led to important insights in the fundamental properties of the generic dynamics that large networks can generate [116] even before progress in genomics could even offer a first glimpse on the actual architecture of a real gene regulatory network [112].

Using the ensemble approach, the architectural parameters that influence the global, long-term behavior of

complex networks with N up to 100,000 has been determined [114,116]. As mentioned in the introduction, based on the latest count of the genome size and the idea that the core transcriptional network essentially governs the global dynamics of gene expression profiles, networks of $N \sim 3000$ would have sufficed. Nevertheless, a striking result from the ensemble studies was that for a broad class of architectures, even a complex, irregular (randomly wired) network can produce 'ordered' dynamics with stable patterns of gene expression, thus potentially delivering the "biological observable". In general, the global behavior of ensembles of Boolean networks can be divided in three broad regimes [116]: an *ordered* and a *chaotic* regime, and a regime that represents behavior at their common border, the *critical* regime.

Ordered and Chaotic Regimes of Network Behavior

In networks in the ***ordered*** regime, two randomly picked initial states $S_1(t=0)$ and $S_2(t=0)$ that are closed to one another in state space as measured by the Hamming distance $H[H(t) = |S_1(t) - S_2(t)| \equiv$ the number of genes whose activities differ in the two states $S_1(t)$ and $S_2(t)]$ will exhibit trajectories that on average will quickly converge (that is, the Hamming distance between the two trajectories will on average decrease with time) [55]. The two trajectories will settle down in one fixed-point attractor or a limit cycle attractor that will have a small period T compared to N and thus, produce very stable system behaviors. Such networks in general have a small number of attractors with typically have small periods T and drain large basins of attraction [116]. Numerical analyzes of large ensembles suggest that the average period length scales with $\sqrt{N}$. The state transition map, as shown for $N = 4$ in Fig. 4a, will show that trajectories converge onto attractor states from many different directions and are in general rather short [Maliackal, unpublished] so that attractor basins appear as compact, with high rotational symmetry and hence, "bushy".

It is important to stress here that if a cell type is an attractor, then different cell types are different attractors, and, in the absence of the unstable steady states present in continuous dynamical systems (see Sect. "Bistability, Tristability and Multistability") differentiation consists in perturbations that move a system state from one attractor into the basin of another attractor from which it flows to the new attractor state that encodes the gene expression profile of the new cell type. Examination of the bushy basins in the ordered regime makes it clear, as do numerical investigations and experiments (Sect. "Are Cell Types Attractors?"), that multiple pathways can lead from one attractor to another attractor – a property that meets resistance in the community of pathway-centered biologists.

In contrast, in networks in the ***chaotic*** regime, two randomly placed initial states $S_1(t=0)$ and $S_2(t=0)$ that are initially close to one another (in terms of Hamming distance) will generate trajectories that on average will *diverge* and either end with high likelihood in two different attractors, or they may appear to "wander" aimlessly in state space. This happens because the attractor is a limit cycle attractor with very long period T – on the scale of 2^N so that in the worst case a trajectory may visit most if not all 2^N possible network configurations S. For a small network of just $N = 200$, this is a limit cycle in the order of length $2^{100} \approx 10^{30}$ time steps. As a point of comparison, the universe is some 10^{17} seconds old. Given the hyperastronomic size of this number, this "limit cycle" will appear as an aperiodic and endless stream of uncorrelated state transitions, as if the system is on a "permanent transient". Thus, networks in the chaotic regime are not stable, trajectories tend to diverge (at least initially), and their behavior is sensitive to the initial state. In the state transition map, the small attractors typically receive trajectories with long transients that arrive from a few state space directions and hence, in contrast to the bushy attractors of the ordered regime, the basins have long thin branches and appear "tree-like".

The definition of "chaos" for discrete networks given here is distinct from that of (deterministic) chaos in continuous systems, where time evolution of infinitesimally closed initial states can be monitored and shown to diverge. Nevertheless, the degree of chaos in discrete networks, as qualitatively outlined above, is well-defined and can be quantified based on the slope of the curve in the so-called Derrida plot which assesses how a large number of random pairs of initial states evolves in one time step [55].

More recently, it was shown that it is possible to determine the behavior class from the architecture, without simulating the state transitions and determining the Derrida plot, simply by calculating the expected *average sensitivity* from all the N Boolean functions B_i [173]. In addition, a novel distance measure that uses normalized compression distances (NCD) – which captures the complexity in the difference between two states $S(t)$ better than the Hamming distance used in the Derrida plot – has been proposed to determine the regime of networks [153].

Critical Networks: Life at Edge of Chaos?

Critical networks are those which exhibit a dynamical behavior just at the edge between that of the ordered and the chaotic regime, and have been postulated to be opti-

mally poised to govern the dynamics of living cells [114, 116]. *Ordered* behavior would be too rigid, in that most perturbations of a stable attractor would be absorbed and the network would return to the same attractor state, minimizing the possibility for a network to change its internal equilibrium state in response to external signals (which are modeled as flipping individual genes from ON to OFF or vice versa). *Chaotic* behavior, on the other hand, would be too sensitive to such perturbations because trajectories diverge – so that the network would wander off in state space and fail to exhibit robust behavior. *Critical* networks may represent the optimal mix between stability and responsiveness, hence conferring robustness to random perturbations (noise) and adaptability in response to specific signals. Critical network have several remarkable features:

First, consider the need of cells to plausibly make the maximum number of reliable discriminations, and to act on them in the presence of noise with maximum reliability. Then 'deep' in the ordered regime, convergence of trajectories in state space is high, hence, as explained above (Sect. "Dynamics in Boolean Networks") information is constantly discarded. In this "lossy" regime, information about past discriminations is thus easily lost. Conversely, in the chaotic regime and with only a small amount of noise, the system will diverge and hence cannot respond reliably to external signals (perturbations). It seems plausible that optimal capacity to categorize and act reliably is found in critical networks, or networks slightly in the ordered regime.

Second, it has recently been shown that a measure of the correlation of pairs of genes with respect to their altering activities, called "***mutual information***" (MI) is maximized for critical networks [154]. The MI measures the mutual dependence of two variables (vectors), such as two genes based on their expression in a set of states $S(t)$. Consider two genes in a synchronous Boolean network, A and B. The mutual information between x_A and x_B is defined as $\mathrm{MI(A,B)} = H(x_A) + H(x_B) - H(x_A, x_B)$. Here, $H(x)$ is the entropy of the variable x, and $H(x, y)$ is the joint entropy of x and y. Mutual information is 0 if either gene A or B is unchanging, or if A and B are changing in time in an uncorrelated way. Mutual information is greater than 0 and bounded by 1.0, if A and B are fluctuating in a correlated way. Thus, *critical networks maximize the correlated changing behavior of variables in the genetic network*. This new result strongly suggests, at least in the ensemble of random Boolean networks, that critical networks can coordinate the most complex organized behavior.

Third, the "basin entropy" of a Boolean network, which characterizes the way the state space is partitioned in to the disjoint basins of attraction of various sizes (Fig. 4a) also exposes a particular property of critical networks [124]. If the size or "weight", W_i, of a basin of attraction i is the fraction of all the 2^N states that flow to that attractor, then the basin entropy is defined as $\Sigma\, W_i \log(W_i)$. The remarkable result is that *only* critical networks have the property that this basin entropy continues to increase as the size of the network increases. By contrast, ordered and chaotic networks have basin entropies that first increase, but then stop increasing with network size [124]. If one thinks of basins of attraction and attractors not only as cell fates, or cell types, but as distinct specific cellular programs encoded by the network, then only critical networks appear to have the capacity to expand the diversity of what "a cell can do" with increasing network size. Again, this strongly suggests that critical networks can carry out the most complex coordinated behaviors. Thus, GRN may have evolved under natural selection (or otherwise – see Sect. "Evolution of Network Structure: Natural Selection or Natural Constraint?") to be critical.

Finally, it is noteworthy that while Boolean networks were invented to model GRNs, the variables can equally be interpreted as any kind of two-valued states of components in a cell, and the Boolean network becomes a *causal network* concerning events in a cell, including GRN as a subset of such events. This suggests that not only GRN but the entire network of processes in cells, including signal transduction and metabolic processes, that is, information, mass and energy flow, may optimally be coordinated if the network is critical.

If life is poised to be in the critical regime, then the questions follow: which architecture produces ordered, critical and chaotic behaviors, and are living cells in fact the critical regime?

Architectural Features of Large Networks and Their Dynamics

As outlined in Sect. "Network Architecture", the recent availability of data on gene regulation in real networks, although far from complete, has triggered the study of complex, irregular network topologies as static graphs. This line of investigation is now beginning to merge with the study of the dynamics of generic Boolean networks. Below we summarize some of the interesting architecture features and their significance for global dynamics in term of the three regimes. First, studies of generic dynamics in ensembles of Boolean networks have established the following major structure-dynamics relationships:

(1) *The average input connectivity per node, k.* Initial studies on Boolean networks by Kauffman assumed a homogenous distribution of inputs k. It was found

that $k = 2$ networks are in the ordered/critical regime (given other parameters, see below) [116]. Above a critical k_c value (which depends on other parameters, see below) networks behave chaotically. Analysis of continuous, linearized models also suggest that in general, sparsity of connections is more likely to promote ordered dynamics – or stability [142].

(2) The *distribution of the connectivity (degree)* over the individual network genes. As mentioned in Sect. "Network Architecture", the global topology of many complex molecular networks appears to have a connectivity distribution that approximates a power-law. For directed graphs such as the GRN, one needs to distinguish between the distribution of the input and output connectivities. Analyses of the dynamics of random Boolean networks with either scale-free input [66] or output connectivity distribution suggest that this property favors the ordered regime for a given value of the parameter p ("internal homogeneity", see below) [6]; Specifically, if the slope of the scale free distribution (power-law exponent γ) is greater than 2.5, the corresponding Boolean network is ordered, regardless of the value of the parameter p (see below). For values of p approaching 1.0, $\gamma > 2.0$ suffices to assure ordered dynamics.

(3) The nature of the Boolean functions is an important aspect of the network architecture that also influences the global dynamics. In the early studies, Kauffman characterized Boolean functions with respect to these two important features [116]:

(a) "*internal homogeneity p*". The parameter p ($0.5 < p < 1$) is the proportion of either 1s or 0s in the output column of the truth table of the Boolean function (Fig. 4). Thus, a function with $p = 0.5$ has equal numbers of 1s and 0s for the output of all input configurations. Boolean functions with p-values close to 1 or 0 are said to exhibit high internal homogeneity.

(b) "*Canalizing function*". A Boolean function B_i of target gene i is said to be canalizing if at least one of its inputs has one value (either 1 or 0) that determines the output of gene i (1 or 0), independently of the values of the other components of the input vector. If the two values of the input j determines *both* output values of gene i [a "fully canalizing" function, e. g., if $x_j(t) = 1$ (or 0, respectively), then $x_i(t+1) = 1$ (or 0, respectively)], then the other inputs have no influence on the output at all and the "*effective* input connectivity" of i, $k_{i\text{eff}}$ is smaller than the "*nominal*" k_i. For instance, for Boolean functions with $k = 2$, only two of the 16 possible functions, XOR and XNOR, are not canalizing. Four functions are "fully canalizing", i. e., are effectively $k = 1$ functions (TRANSFER, COMPLEMENT).

From ensemble studies it was found that both a high internal homogeneity p and a high proportion of canalizing functions contribute to ordered behavior [116].

Do Real Networks Have Architecture Features That Suggest They Avoid Chaos?

Only scant data is available for transcriptional networks, and it must be interpreted with due caution since new data may, given sampling bias and artefacts, especially for nearly scale-free distributions, affect present statistics. In any case, existing data indicate that in fact the average input connectivity of GRNs is rather low, and far from $k \sim N$ which would lead to chaotic behavior. Specifically, analysis of available (partial) transcriptional networks suggest that the *input* degrees approximates an *exponential* distribution while the *output* degree distribution seems to be *scale-free* although the number of nodes are rather small to reliably identify a power-law distribution [54,82].

GRN data from *E. coli*, for which the most complete, hand-curated maps of genome-wide gene regulation exists [169] and from partial gene interaction networks from yeast, obtained mostly by chromatin-precipitation/microarray (ChIP-chip) [128], indicate that the average input connectivity (which is exponentially distributed) is below 4 [82,134,188]. A recent work on the worm *C. elegans* using the Yeast-one-hybrid system on a limited set of 72 promoters found an average of 4 DNA-binding proteins per promoter [54]. Thus, while such analyzes await correction as coverage increase, clearly, real GRNs for microbial and lower metazoan are sparse, and hence more likely to be in or near the ordered regime.

In contrast to the input, the output connectivity appears to be power-law distributed for yeast and bacteria, and perhaps also for *C. elegans*, if the low coverage of the data available so far can be trusted [54,82]. The paucity of data points precludes reliable estimates of the power-law exponents. However, the scale-free property, if confirmed for the entire GRN, may well also contribute to avoiding chaotic and increasing the ordered or critical regime [6]. Nevertheless, It is reminded here that the deeper meaning of the scale-freeness per se and its genesis (natural selection due to functionality or not) are not clear – it may be an inevitable manifestation of fundamental statistics rather than evolved under natural selection (Sect. "Evolution of Network Structure: Natural Selection or Natural Constraint?").

As for the use of Boolean functions, analysis of a set of 150 experimentally verified regulatory mechanism of well-studied promoters revealed an enrichment for canalizing functions, again, in accordance with the architecture criteria associated with ordered dynamics [86]. Similarly, when canalizing functions were randomly imposed onto the published yeast protein–protein interaction network topology to create a architecture whose dynamics was then simulated, ordered behavior was observed [113].

In this conjunction it is interesting to mention ***microRNAs*** (miRNAs), a recently discovered class of non-coding RNA which act by inhibiting gene expression at the post-transcriptional level through sequence complementarity to mRNA [41,89]. Hence, they suppress a gene independent of the TF constellation at the promoter. Thus, miRNAS epitomize the most simple and powerful molecular realization of a canalizing function. Their existence may therefore shift the network behavior from the chaotic towards the critical or ordered regime. Interestingly, RNA-based gene regulation is believed to have appeared before the metazoan radiation 1000 million years ago, and microRNAs are thought to have evolved in ancestors of Bilateria [141]. In fact, many miRNA play key roles in cell fate determination during tissue specification of development of vertebrates [41,172], and a composite feedback loop circuit involving both microRNA and TFs has been described in neutrophil differentiation [63]. The shift of network behavior from the chaotic towards the critical or ordered regime may indeed enable the coordination of complex gene expression patterns during ontogeny of multicellular systems which requires maximal information processing capacity to ensure the coexistence of stability and diversity.

There are, as mentioned in Sect. "Analyzing Network Structures", many more global and local topological features that have been found in biomolecular networks that appear to be interesting, as defined in the sense in Sect. "Network Architecture" (enriched above some null model graph). It would be interesting to test how they contribute to producing chaotic, critical or ordered behavior. The impact of most of these topology features on the global dynamics, notably the three regimes of behaviors, remains unknown, since most functional interpretation of network motifs have focused on local dynamics [9,136]. The low quality and availability of experimental data of GRN architectures opens at the moment only a minimal window into the dynamical regimes of biological networks. The improvement with respect to coverage and quality of real GRNs, to be expected in the coming years, is mostly driven by a reductionist agenda whose intrinsic aim is to exhaustively enumerate all the "pathways" that may serve "predictive modeling" of gene behaviors, as detailed in Sect. "Small Gene Circuits". However, with the concepts introduced here, a framework now exists that warrants a deeper, genome-wide analysis of the relationship between structure and biological function. Such analysis should also address the fundamental dualism between inevitable self-organization (due to intrinsic constraints from physical laws) and natural selection (of random mutants for functional advantages) [46,93,203] to ask whether criticality is self-organized (see Sect. "Evolution of Network Structure: Natural Selection or Natural Constraint?").

Experimental Evidence from Systems Biology

The experimental validation of the central concepts that were erected based on theoretical analysis of network dynamics, notably the ensemble approach of Boolean networks, amounts to addressing the following two questions:

1. Are cell types attractors?
2. Is the dynamical behavior of the genomic GRN in the critical regime?

As experimental systems biology begins to reach beyond the systematic characterization of genes and their interactions it has already been possible to design experiments to obtain evidence to address these questions.

Are Cell Types Attractors?

Obviously, the observable gene expression profiles S^*, as shown in Fig. 1e, are stationary (steady) states and characteristic of a cell type. But "steady state" ($d\mathbf{x}/dt = 0$ in Eq. (1)) does not necessary imply a stable (self-stabilizing) attractor state that attracts nearby states. In the absence of knowledge of the architecture of the underlying GRN (we do not know the function $\mathbf{F}$ in Eq. (1)) we cannot perform the standard formal analysis, such as linear stability analysis around S^*, to determine whether a stationary state is stable or not; however, use of microarray-based expression profiling not to identify individual genes as in the gene-centered view, but in a novel integrated manner (Table 1, B3) for the analysis of $S(t)$ provides a way to address the question. The qualitative properties of an attractor offer a handle, in that a high-dimensional attractor state S^*, be it a fixed-point or a small limit cycle or even a strange attractor in the state space, requires that the volume of the states around it contracts to the attractor, i. e., $\mathrm{div}(\mathbf{F}) < 0$. [90].

Convergence of Trajectories Thus, one consequence is that trajectories emanating from states around (and near) S^* converge towards it from most (ideally, all) directions

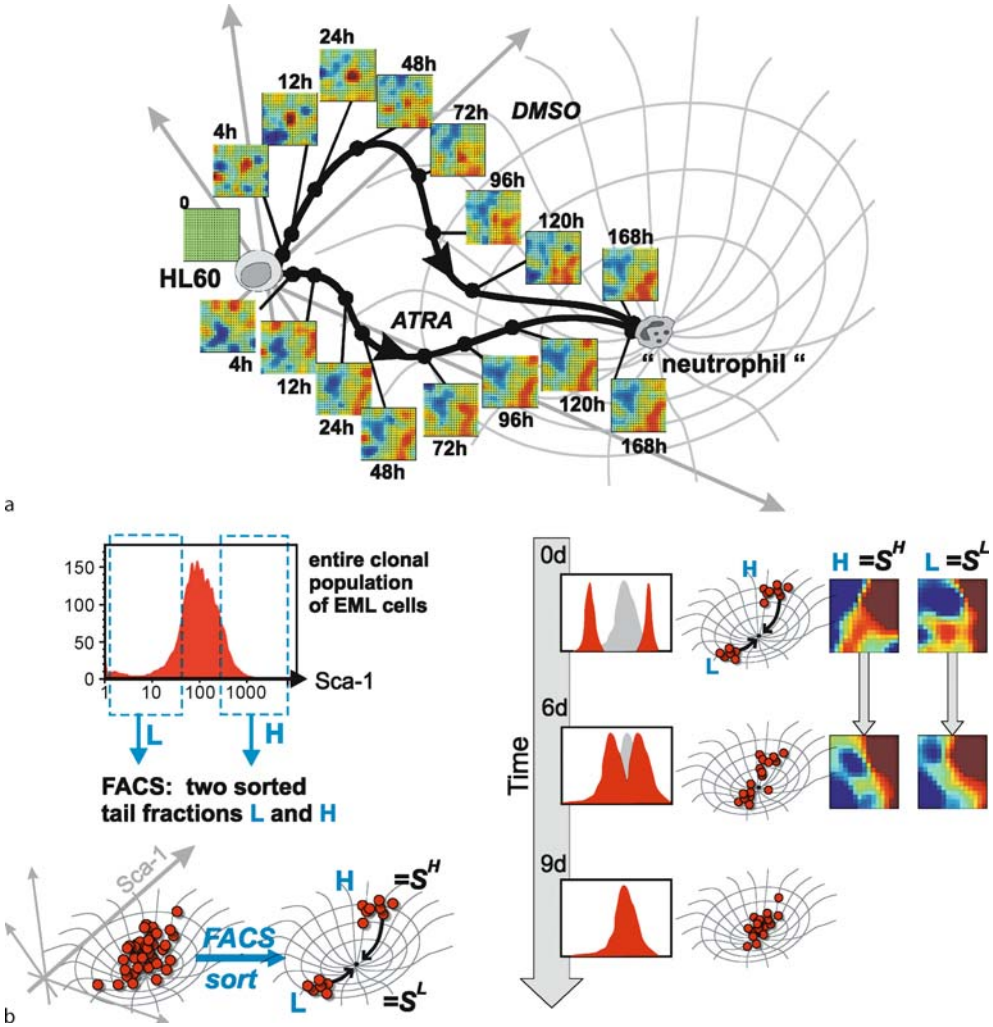

Complex Gene Regulatory Networks – from Structure to Biological Observables: Cell Fate Determination, Figure 5

Experimental evidence that a cell fate (differentiated state, cell type) is a high-dimensional stable attractor in gene expression state space. Evidence is based on convergence (a) from different directions of high-dimensional trajectories to the gene expression pattern of the differentiated state, or on the relaxation of states placed near the border of the basin of attraction back to the "center" of the attractor state (b). a Gene expression profile dynamics of HL60 cells treated with ATRA (all-trans retinoid acid) or DMSO (dimethylsulfoxide) at 0h to differentiate them into neutrophil-like cells. Selected gene expression profile snapshots along the trajectories shown as GEDI maps (as in Fig. 1), schematically placed along the state space trajectories. GEDI maps show the convergence of $N \sim 2800$ genes towards very similar patterns at 168h [98]. b Heterogeneity of a population of clonal (genetically identical) cells is exploited to demonstrate the relaxation towards the attractor state. The histogram (*top, left panel*) from flow cytrometric measurement shows the inherent heterogeneity (spread) of cells with respect to stem cell marker Sca-1 expression in EML cells which cannot be attributed solely to measurement or gene expression noise but reflect metastable cell-individuality [37]. Two spontaneous "outlier" subfractions, expressing either low (L) or high (H) levels of the Sca-1, were sorted using FACS (fluorescence-activated cell sorting) and cultured independently. They represent "small perturbations" of the attractor state. Each sorted subpopulation will over a period of 5–9 days restore the parental distribution (*right panel, schematic*). Gene expression profiling (shown as as GEDI maps) reveal that the H and L subpopulations are distinct with respect not only to Sca-1 expression levels but also to that of multiple other genes. Thus, the two outlier cell fractions are at distinct states S^H and S^L in the high-dimensional state space of which Sca-1 is only one distinguishing dimension. The restoration of the parental distribution of Sca-1 is accompanied by the approaching of the two distinct gene expression profiles S^H and S^L of the spontaneously perturbed cells to become similar to each other (and to that of the parental population), indicating a high-dimensional attractor state

and in all dimensions of the state space. It is technically challenging to sample multiple high-dimensional initial states near the attractor states and demonstrate that they all converge towards S^* over time. However, the fact that the promyeloid precursor cells HL60 (a leukemic cell line) can be triggered to differentiate into mature neutrophil-like cells by an array of distinct chemical agents can be exploited [45]. This historical observation itself, without analysis of $S(t)$ by gene expression profiling, already suggests that the neutrophil state is an attractor, because it reflects robustness and means that the detailed history of how it is reached doesn't matter.

But the advent of microarray-based gene expression profiling technologies has recently opened up the possibility to show that (at least two) distinct trajectories converge towards gene expression profile of the neutrophil state S_{Neutr} if S_{Neutr} is an attractor state [98]. Thus, HL60 cells were treated with either one of two differentiation-inducing reagents, all-trans-retinoid acid (ATRA) and dimethyl sulfoxide (DMSO), two chemically unrelated compounds, and the changes of the transcriptome over time in response to either treatment were measured at multiple time intervals to monitor the two trajectories, $S^{\text{ATRA}}(t)$ and $S^{\text{DMSO}}(t)$, respectively (see Fig. 5a for details). In fact the two trajectories first diverged to an extent that the two state vectors lost most of their correlation (i. e. the two gene expression profiles $S^{\text{ATRA}}(24\,\text{h})$ and $S^{\text{DMSO}}(24\,\text{h})$ were maximally different at $t = 24$ hours after stimulation). But subsequently, they converged to very similar $S(t)$ values when the cells reached the differentiated neutrophil state for both treatments (Fig. 5a). The convergence was not complete, but quite dramatic relative to the maximally divergent state at 24 h, and was contributed by around 2800 of the 3800 genes monitored (for details see [98]). Thus, it appears that at least the artificial drug-induced differentiated neutrophil state is so stable that it can apparently orchestrate the expression of thousands of genes to produce the appropriate cell-type defining expression pattern S from quite distinct perturbed cellular states. Although only two trajectories have been measured rather than an entire state space volume, the convergence with respect to 2800 state space dimensions is strongly indicative of a high-dimensional attractor state.

Relaxation After Small Perturbations and the Problem of Cell Heterogeneity A second way to expose qualitative properties of a high-dimensional attractor S^* is to perform a weak perturbation of S^*, into a state S' near the edge of (but within) the basin of attraction (where S' should differ from S^* with respect to as many genes x_i as possible), and to observe its return of the network to S^*. This more intuitive property of an attractor state was difficult to measure because of a phenomenon often neglected by theorists and experimentalists alike: cell population heterogeneity due to "gene expression noise". Microarray measurements, as is the case with many biochemical analysis methods, require the material of millions of cells, thus the measured $S(t)$ is actually a population average. This can be problematic because the population is heterogeneous [33]: gene expression level of gene i can typically differ by as much as 100 or more fold between two individual cells within a clonal (genetically identical) population [37]. Thus, virtually all genes i exhibit a broad (log-normal) histogram (Fig. 5b, inset) when the expression level of an individual gene x_i is measured at the single-cell level across a population [38,101,129]. Such cell-to-cell variability is often explained by "gene expression noise" caused by random temporal fluctuations due to low copy numbers of specific molecules in the cell [108]. However, there may be other possibly deterministic sources that generate metastable variant cells [33,176]. Such non-genetic, enduring individuality of cells translates into the picture of a cell population forming a cloud of points in state space around the attractor S^*, in which each cell represents a single point. Application of a low dose perturbation intended to allow a system to relax back to the attractor state then will be interpreted by individual cells differently: Those positioned at the border of the cloud may be kicked out of the basin of attraction, and move to another attractor, thus masking the trajectory of relaxation. In fact, single-cell resolution measurements of the response to low dose stimulation in cell populations confirmed this picture of heterogeneity, in that a differentiation inducer given at low dose to trigger a partial response (weak perturbation) produced a bimodal distribution of gene expression of differentiation markers: Some cells differentiated, other did not [38].

However, the spontaneous heterogeneity eo ipso, consisting of transient but persistent variant cells within a population [37], allows us to demonstrate the relaxation to the attractor when single cell-level manipulation and analysis are performed: Physical isolation of the population fractions at two opposite the edge of the cloud (basin of attraction), based on one single state space dimension x_k can substitute for the weak perturbation that places cells to the border of a basin (see Fig. 5b for details). Indeed such "outlier" population fractions exhibited not only distinct levels of x_k expression but also globally distinct gene expression profiles (despite being members of the same clonal population). Cells of both outlier fractions eventually "flowed back" to populate the

state space region around the attractor state and restored the original distribution (shape of the cloud) [37]. The time scale of this relaxation (>5 days) was similar to that for the HL60 cells to converge to the attractor of the differentiated neutrophils. This result is summarized in Fig. 5b. Again, the spontaneous regeneration of the distinct gene expression profile of a macroscopically observable cell phenotype, consisting of thousands of genes, supports the notion of a high-dimensional attractor in gene expression space that maps into a distinct, observable cell type.

Are Gene Regulatory Networks in Living Cells Critical?

The second central question we ask in this article is whether GRNs produce a global dynamics that is in the critical regime (Sect. "Three Regimes of Behaviors for Boolean Networks"). This question is not as straightforward to address experimentally. First, the notion of order, chaos and criticality are defined in the models as properties of network ensembles. Second, it is cumbersome to sample a large number of pairs of initial states and monitor their high-dimensional trajectories to determine whether they on average converge, diverge or stay "parallel".

As mentioned earlier (Sect. "Dynamics in Boolean Networks"), one approach for obtaining a first glimpse of an idea as to whether real networks are critical or not was recently proposed by Aldana and coworkers [19] and refs. herein): They "imposed" a dynamics onto real biological networks, for which only the topology but not the interaction functions are known, by treating them as Boolean networks, whereby the Boolean functions were guessed or randomly assigned according to some rules. Such studies suggest that these networks, given their assumed topologies, are in the critical regime.

To more directly characterize the observed dynamics of natural systems in terms of the three regimes, several indirect approaches have been taken. These strategies are based on novel measurable quantities computed from several schemes of microarray experiments that are now available but were not originally generated with an intention to answer this question. By first determining in simulated networks whether that quantity is associated also with criticality in silico, inference is then made as to what regime of system behavior the observed system resembles. Three such pieces of evidence provide a first hint that perhaps, living cells may indeed be in the critical regime:

(i) Gene expression profile changes during Hela cell cycle progression was compared with the detailed temporal structure of the updating of network states in Boolean networks. Tithe discretized real gene expression data of cells progressing in the cell cycle were compared with that of simulated state cycles of random Boolean networks in the three regimes in terms of the Lempev-Zip complexity of time series. This led to the conclusion that the dynamical behavior of thousands of genes was most consistent with either ordered or critical behavior, but not chaotic behavior [174].

(ii) A striking property predicted from analysis of simulated critical Boolean networks is that if a randomly selected gene is deleted and the number of other genes that change their expression as a consequence of that single-gene deletion ("avalanche size") is measured, and such single-gene deletion experiments is repeated many times, the avalanche-sizes will exhibit a power law distribution with a slope of $-\gamma = 1.5$. This specific behavior is only seen in critical networks. Analysis of just such data for over 200 single deletion mutants in yeast [163] from the experiments reported by Hughes et al. [100]was recently performed. It was found that the distribution of the avalanche sizes not only approached a power law, but that the slope was also -1.5 [163]. This result was insensitive to altering the criterion for defining a "change in gene activity" from two-fold to five fold in calculating the avalanche size.

(iii) A more direct determination of the regime of network dynamics was recently reported, in which an analogous analysis as the Derrida plot (see Sect. "Ordered and Chaotic Regimes of Network Behavior") was performed [153]. Macrophage gene expression profiles were measured at various time points in response to various perturbations (= stimulation of Toll-like receptors with distinct agents), offering a way to measure the time evolution of a large number of similar initial states. Here, instead of the Hamming distance the normalized compression distance NCD (as mentioned in Sect. "Ordered and Chaotic Regimes of Network Behavior") was used to circumvent the problems associated with Derrida curves. The results were consistent with critical dynamics, in that on average, for many pairs of initial states, their distance at time t and $t + \Delta t$ was on average equal, thus neither trajectory convergence nor divergence took place.

Future Directions and Questions

The ideas of a state space representing the dynamics of the network, and of its particular structure that stems from the

constraints imposed by the regulatory interactions and can be epitomized as an 'epigenetic landscape', serve as a conceptual bridge linking network architecture with the observable biological behavior of a cell. In this framework, the attractors (valleys) are disjoint regions in the state space landscape and represent cell fates and cell types. One profound and bold hypothesis is that for networks to have a landscape with attractors that optimally convey cell lineage identity and robustness of their gene expression profile, yet allow enough flexibility for cell phenotype switches during development, the networks must be poised at the boundary between the chaotic and the ordered regime. Such critical networks may be a universal property of networks that have maximal information processing capacity. But what can we naturally do with this conceptual framework presented here? And what are the next questions to ask for the near future?

Clearly, new functional genomics analysis techniques will soon advance the experimental elucidation of the architecture of the GRN of various species, including that of metazoan organisms. This will drastically expand the opportunities for theoretical analysis of the architecture of networks and finally afford a much closer look at the dynamics without resorting to simulated network ensembles.

However, beyond network analysis in terms of mathematical formalisms, the concepts of integrated network dynamics presented here should also pave the way for a formal rather than descriptive understanding of tissue homeostasis and development as well as diseases, such as cancer. One corollary of the idea that cell types are attractors is that cancer cells are also attractors – which are lurking somewhere in state space (near embryonic or stem cell attractors) and are normally avoided by the physiological developmental trajectories. They become accessible in pathological conditions and trap cells in them, preventing terminal differentiation (this idea is discussed in detail in [33,96]).

Such biological interpretation of the concepts presented here will require that these concepts be expanded to embrace the following aspects of dynamical biological systems that are currently not well understood but can already be framed:

Attractor Transitions and Development

If cell fates are attractors, then development is a flow in state space trajectories which then represent the developmental trajectories. But how do cells, e. g., a progenitor cell committing to a particular cell fate, move from one attractor to another? Currently, two models are being envisioned for the "flow between attractors" that constitutes developmental paths in the generation of the multi-cellular organisms:

(i) In one model, stimulated by the studies in Boolean networks, the cell "jumps" from one attractor to another in response to a distinct perturbation (e. g., developmental signal) which is represented as the imposed alteration of the expression status of a set of genes ("bit-flipping" in binary Boolean networks). This corresponds to the displacement of the network state $S(t)$ by an external influence to a new position in the basin of another attractor.

(ii) The second model posits that the landscape changes, and has its roots in the classical modeling of nonlinear dynamical systems: For instance, the attractor of the progenitor cell (valley) may be converted into an unstable steady-state point (hill top) at which point the network (cell) will spontaneously be attracted by either of the two attractors on each side of the newly formed hill. This is exemplified in a model for fate decision in bipotent progenitor cells in hematopoiesis [99]. In this model, a change of the landscape structure ensues from a change in the network architecture caused by the slow alteration of the value of a system parameter that controls the interaction strength in the system equations (see Sect. "Small Gene Circuits"). Thus, the external signal that triggers the differentiation exerts its effect by affecting the system parameters. Increasing the decay rates of x_1 and x_2 in the example of Fig. 3b will lead to the disappearance of the central attractor and convert the tristable landscape of Fig. 3b to the bistable one of Fig. 3a [99]. Such qualitative changes that occur during the slow increase or decrease of a system parameter are referred to as a ***bifurcation***. Note that this second model takes a narrower view, assuming that the network under study is not the global, universal and fixed network of the entire genome, as discussed in Sect. "Complex Networks", but rather a subnetwork that can change its architecture based on the presence or absence of gene products of genes that are outside of the subnetwork.

Noisy Systems

In a bistable switching system, it is immediately obvious that random fluctuations in x_i due to "gene expression noise" may trigger a transition between the attractors and thus, explain the stochastic phenotype transitions, as has been recently shown for several micro-organismal systems and discussed for mammalian cell differentiation [87,97,

101,108]. The observed heterogeneity of cells in a nominally identical cell population,caused either by "gene expression noise" or other diversifying processes, such as the random partitioning of molecules at cell divisions, implies that cell states cannot be viewed as deterministic points and trajectories in the state space, but rather as moving "clouds" – held together by the attractors.

More recently, on the basis of such bistable switches, it has even been proposed and shown in synthetic networks that environmental bias in fluctuation magnitude may explicitly control the switch to the physiologically favorable attractor state – because gene expression noise may be higher (relative to the deterministically controlled expression levels) when cells are in the attractor that dictates a gene expression pattern that is incompatible with a given environment [110]. Such biased gene expression noise thus may drive the flow between the attractors, and hence, (on average) guide the system through attractor transitions to produce various cell fates in a manner that is commensurate with development of the tissue. On the other hand, it may also explain the local stochasticity of cell fate decisions observed for many stem and progenitor cells.

Directionality

The concepts of cell fates and cell types as attractors, as well as any mechanisms that explains attractor transitions during development and differentiation, do not explain the overall directionality of development, or the "*arrow of time*" in ontogenesis: Why is development essentially a one-way process, i. e., irreversible in time, given that the underlying regulatory events, the switching ON or OFF of genes, are fully reversible? Where does the overall slope (from back to front) in Waddington's epigenetic landscape (Fig. 2g) come from? One idea is that "gene expression noise" may play the role of thermal noise (heat) in thermodynamics in explaining the irreversibility of processes. Alternatively, the network could be specifically wired, perhaps through natural selection so that attractor transitions are highly biased for one direction.

Beyond Cell Autonomous Processes

To understand development we need of course to open our view beyond the cell autonomous dynamics of GRNs. Some of the genes expressed in particular states $S(t)$ encode secreted proteins that affect the gene expression hence, state $S(t)$ of neighboring cells. Such inter-cell communication establishes a network at a higher level, with its own constrained dynamics that need to be incorporated in the models of developmental trajectories in gene expression state space.

Evolution

As discussed several times in this article, when networks are studied in the light of evolution, a central question arises: how do mutations, which essentially rewire the network by altering the nature of regulatory interactions, give rise to the particular architecture of the GRN that we find today? What are the relative roles in shaping particular network architectures, of (*i*) constraints due to physical (graph-theoretical) laws and self-organization vs. that of (*ii*) natural selection for functionality? If selection plays a major role, can it select for such features as the global landscape structure, or even for criticality? Or can the latter even "self-organize" without adaptive selection [31]? Such questions go far beyond the current analysis of Darwinian mechanisms of robustness and evolvability of networks. They are at the heart of the quest in biocomplexity research for fundamental principles of life, of which the process of natural selection itself is a just subset.

Regulatory networks offer a accessible and formalizable object of study to begin to ask these questions.

Bibliography

Primary Literature

1. Adolfsson J, Mansson R, Buza-Vidas N, Hultquist A, Liuba K, Jensen CT, Bryder D, Yang L, Borge OJ, Thoren LA, Anderson K, Sitnicka E, Sasaki Y, Sigvardsson M, Jacobsen SE (2005) Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage commitment. Cell 121:295–306
2. Aittokallio T, Schwikowski B (2006) Graph-based methods for analysing networks in cell biology. Brief Bioinform 7:243–55
3. Albert R (2005) Scale-free networks in cell biology. J Cell Sci 118:4947–57
4. Albert R, Othmer HG (2003) The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in Drosophila melanogaster. J Theor Biol 223:1–18
5. Albert R, Jeong H, Barabasi AL (2000) Error and attack tolerance of complex networks. Nature 406:378–82
6. Aldana M, Cluzel P (2003) A natural class of robust networks. Proc Natl Acad Sci USA 100:8710–4
7. Aldana M, Balleza E, Kauffman S, Resendiz O (2007) Robustness and evolvability in genetic regulatory networks. J Theor Biol 245:433–48
8. Aldana M, Coppersmith S, Kadanoff LP (2003) Boolean dynamics with random couplings. In: Kaplan E, Marsden JE, Sreenivasan KR (eds) Perspectives and problems in nonlinear science. A celebratory volume in honor of Lawrence Sirovich. Springer, New York
9. Alon U (2003) Biological networks: the tinkerer as an engineer. Science 301:1866–7

10. Amaral LA, Scala A, Barthelemy M, Stanley HE (2000) Classes of small-world networks. Proc Natl Acad Sci USA 97:11149–52
11. Anderson PW (1972) More is different. Science 177:393–396
12. Angeli D, Ferrell JE Jr., Sontag ED (2004) Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems. Proc Natl Acad Sci USA 101: 1822–1827
13. Arney KL, Fisher AG (2004) Epigenetic aspects of differentiation. J Cell Sci 117:4355–63
14. Artzy-Randrup Y, Fleishman SJ, Ben-Tal N, Stone L (2004) Comment on Network motifs: simple building blocks of complex networks and Superfamilies of evolved and designed networks. Science 305:1107; author reply 1107
15. Autumn K, Ryan MJ, Wake DB (2002) Integrating historical and mechanistic biology enhances the study of adaptation. Q Rev Biol 77:383–408
16. Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA (2004) Structure and evolution of transcriptional regulatory networks. Curr Opin Struct Biol 14:283–91
17. Bagley RJ, Glass L (1996) Counting and classifying attractors in high dimensional dynamical systems. J Theor Biol 183:269–84
18. Balcan D, Kabakcioglu A, Mungan M, Erzan A (2007) The information coded in the yeast response elements accounts for most of the topological properties of its transcriptional regulation network. PLoS ONE 2:e501
19. Balleza E, Alvarez-Buylla ER, Chaos A, Kauffman A, Shmulevich I, Aldana M (2008) Critical dynamics in genetic regulatory networks: examples from four kingdoms. PLoS One 3:e2456
20. Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286:509–12
21. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5: 101–113
22. Becskei A, Seraphin B, Serrano L (2001) Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. EMBO J 20:2528–2535
23. Berg J, Lassig M, Wagner A (2004) Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. BMC Evol Biol 4:51
24. Bergmann S, Ihmels J, Barkai N (2004) Similarities and differences in genome-wide expression data of six organisms. PLoS Biol 2:E9
25. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125:315–26
26. Bird A (2007) Perceptions of epigenetics. Nature 447:396–8
27. Bloom JD, Adami C (2003) Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein–protein interactions data sets. BMC Evol Biol 3:21
28. Bloom JD, Adami C (2004) Evolutionary rate depends on number of protein–protein interactions independently of gene expression level: response. BMC Evol Biol 4:14
29. Bork P, Jensen LJ, von Mering C, Ramani AK, Lee I, Marcotte EM (2004) Protein interaction networks from yeast to human. Curr Opin Struct Biol 14:292–9
30. Bornholdt S (2005) Systems biology. Less is more in modeling large genetic networks. Science 310:449–51
31. Bornholdt S, Rohlf T (2000) Topological evolution of dynamical networks: global criticality from local dynamics. Phys Rev Lett 84:6114–7
32. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, Gifford DK, Melton DA, Jaenisch R, Young RA (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122:947–56
33. Brock A, Chang H, Huang SH Non-genetic cell heterogeneity and mutation-less tumor progression. Manuscript submitted
34. Brown KS, Hill CC, Calero GA, Myers CR, Lee KH, Sethna JP, Cerione RA (2004) The statistical mechanics of complex signaling networks: nerve growth factor signaling. Phys Biol 1: 184–195
35. Bulyk ML (2006) DNA microarray technologies for measuring protein-DNA interactions. Curr Opin Biotechnol 17:422–30
36. Callaway DS, Hopcroft JE, Kleinberg JM, Newman ME, Strogatz SH (2001) Are randomly grown graphs really random? Phys Rev E Stat Nonlin Soft Matter Phys 64:041902
37. Chang HH, Hemberg M, Barahona M, Ingber DE, Huang S (2008) Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. Nature 453:544–547
38. Chang HH, Oh PY, Ingber DE, Huang S (2006) Multistable and multistep dynamics in neutrophil differentiation. BMC Cell Biol 7:11
39. Chang WC, Li CW, Chen BS (2005) Quantitative inference of dynamic regulatory pathways via microarray data. BMC Bioinformatics 6:44
40. Chaves M, Sontag ED, Albert R (2006) Methods of robustness analysis for Boolean models of gene control networks. Syst Biol (Stevenage) 153:154–67
41. Chen K, Rajewsky N (2007) The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet 8: 93–103
42. Chen KC, Wang TY, Tseng HH, Huang CY, Kao CY (2005) A stochastic differential equation model for quantifying transcriptional regulatory network in Saccharomyces cerevisiae. Bioinformatics 21:2883–90
43. Chickarmane V, Troein C, Nuber UA, Sauro HM, Peterson C (2006) Transcriptional dynamics of the embryonic stem cell switch. PLoS Comput Biol 2:e123
44. Claverie JM (2001) Gene number. What if there are only 30,000 human genes? Science 291:1255–7
45. Collins SJ (1987) The HL-60 promyelocytic leukemia cell line: proliferation, differentiation, and cellular oncogene expression. Blood 70:1233–1244
46. Cordero OX, Hogeweg P (2006) Feed-forward loop circuits as a side effect of genome evolution. Mol Biol Evol 23:1931–6
47. Cross MA, Enver T (1997) The lineage commitment of haemopoietic progenitor cells. Curr Opin Genet Dev 7: 609–613
48. Dang CV, O'Donnell KA, Zeller KI, Nguyen T, Osthus RC, Li F (2006) The c-Myc target gene network. Semin Cancer Biol 16:253–64
49. Davidich MI, Bornholdt S (2008) Boolean network model predicts cell cycle sequence of fission yeast. PLoS ONE 3:e1672
50. Davidson EH, Erwin DH (2006) Gene regulatory networks and the evolution of animal body plans. Science 311:796–800
51. de la Serna IL, Ohkawa Y, Berkes CA, Bergstrom DA, Dacwag CS, Tapscott SJ, Imbalzano AN (2005) MyoD targets chromatin

remodeling complexes to the myogenin locus prior to forming a stable DNA-bound complex. Mol Cell Biol 25:3997–4009
52. Deane CM, Salwinski L, Xenarios I, Eisenberg D (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol Cell Proteomics 1:349–56
53. Delbrück M (1949) Discussion. In: Unités biologiques douées de continuité génétique Colloques Internationaux du Centre National de la Recherche Scientifique. CNRS, Paris
54. Deplancke B, Mukhopadhyay A, Ao W, Elewa AM, Grove CA, Martinez NJ, Sequerra R, Doucette-Stamm L, Reece-Hoyes JS, Hope IA, Tissenbaum HA, Mango SE, Walhout AJ (2006) A gene-centered C. elegans protein-DNA interaction network. Cell 125:1193–205
55. Derrida B, Pomeau Y (1986) Random networks of automata: a simple annealed approximation. Europhys Lett 1:45–49
56. Dodd IB, Micheelsen MA, Sneppen K, Thon G (2007) Theoretical analysis of epigenetic cell memory by nucleosome modification. Cell 129:813–22
57. Eichler GS, Huang S, Ingber DE (2003) Gene Expression Dynamics Inspector (GEDI): for integrative analysis of expression profiles. Bioinformatics 19:2321–2322
58. Eisenberg E, Levanon EY (2003) Preferential attachment in the protein network evolution. Phys Rev Lett 91:138701
59. Enver T, Heyworth CM, Dexter TM (1998) Do stem cells play dice? Blood 92:348–51; discussion 352
60. Espinosa-Soto C, Padilla-Longoria P, Alvarez-Buylla ER (2004) A gene regulatory network model for cell-fate determination during Arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles. Plant Cell 16:2923–39
61. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol 5:e8
62. Faure A, Naldi A, Chaouiya C, Thieffry D (2006) Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle. Bioinformatics 22:e124–31
63. Fazi F, Rosa A, Fatica A, Gelmetti V, De Marchis ML, Nervi C, Bozzoni I (2005) A minicircuitry comprised of microRNA-223 and transcription factors NFI-A and C/EBPalpha regulates human granulopoiesis. Cell 123:819–31
64. Ferrell JE Jr., Machleder EM (1998) The biochemical basis of an all-or-none cell fate switch in Xenopus oocytes. Science 280:895–8
65. Fisher AG (2002) Cellular identity and lineage choice. Nat Rev Immunol 2:977–82
66. Fox JJ, Hill CC (2001) From topology to dynamics in biochemical networks. Chaos 11:809–815
67. Fraser HB, Hirsh AE (2004) Evolutionary rate depends on number of protein–protein interactions independently of gene expression level. BMC Evol Biol 4:13
68. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary rate in the protein interaction network. Science 296:750–2
69. Fuks F (2005) DNA methylation and histone modifications: teaming up to silence genes. Curr Opin Genet Dev 15: 490–495
70. Gao H, Falt S, Sandelin A, Gustafsson JA, Dahlman-Wright K (2007) Genome-wide identification of estrogen receptor α binding sites in mouse liver. Mol Endocrinol 22:10–22
71. Gardner TS, Cantor CR, Collins JJ (2000) Construction of a genetic toggle switch in Escherichia coli. Nature 403:339–342
72. Gershenson C (2002) Classification of random Boolean networks. In: Standish RK, Bedau MA, Abbass HA (eds) Artificial life, vol 8. MIT Press, Cambridge, pp 1–8
73. Gisiger T (2001) Scale invariance in biology: coincidence or footprint of a universal mechanism? Biol Rev Camb Philos Soc 76:161–209
74. Glass L, Kauffman SA (1972) Co-operative components, spatial localization and oscillatory cellular dynamics. J Theor Biol 34:219–37
75. Goldberg AD, Allis CD, Bernstein E (2007) Epigenetics: a landscape takes shape. Cell 128:635–8
76. Goldstein ML, Morris SA, Yen GG (2004) Problems with fitting to the power-law distribution. Eur Phys J B 41:255–258
77. Goodwin BC, Webster GC (1999) Rethinking the origin of species by natural selection. Riv Biol 92:464–7
78. Gould SJ, Lewontin RC (1979) The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proc R Soc Lond B Biol Sci 205:581–98
79. Graf T (2002) Differentiation plasticity of hematopoietic cells. Blood 99:3089–101
80. Grass JA, Boyer ME, Pal S, Wu J, Weiss MJ, Bresnick EH (2003) GATA-1-dependent transcriptional repression of GATA-2 via disruption of positive autoregulation and domain-wide chromatin remodeling. Proc Natl Acad Sci USA 100:8811–6
81. Greil F, Drossel B, Sattler J (2007) Critical Kauffman networks under deterministic asynchronous update. New J Phys 9:373
82. Guelzim N, Bottani S, Bourgine P, Kepes F (2002) Topological and causal structure of the yeast transcriptional regulatory network. Nat Genet 31:60–3
83. Guo Y, Eichler GS, Feng Y, Ingber DE, Huang S (2006) Towards a holistic, yet gene-centered analysis of gene expression profiles: a case study of human lung cancers. J Biomed Biotechnol 2006:69141
84. Hahn MW, Kern AD (2005) Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol 22:803–6
85. Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402:C47–52
86. Harris SE, Sawhill BK, Wuensche A, Kauffman SA (2002) A model of transcriptional regulatory networks based on biases in the observed regulation rules. Complexity 7:23–40
87. Hasty J, Pradines J, Dolnik M, Collins JJ (2000) Noise-based switches and amplifiers for gene expression. Proc Natl Acad Sci USA 97:2075–80
88. Haverty PM, Hansen U, Weng Z (2004) Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification. Nucleic Acids Res 32:179–88
89. He L, Hannon GJ (2004) MicroRNAs: small RNAs with a big role in gene regulation. Nat Rev Genet 5:522–31
90. Hilborn R (1994) Chaos and nonlinear dynamics: An introduction for scientists and engineers, 2 edn. Oxford University Press, New York
91. Hochedlinger K, Jaenisch R (2006) Nuclear reprogramming and pluripotency. Nature 441:1061–7
92. Hu M, Krause D, Greaves M, Sharkis S, Dexter M, Heyworth C, Enver T (1997) Multilineage gene expression precedes commitment in the hemopoietic system. Genes Dev 11:774–85

93. Huang S (2004) Back to the biology in systems biology: what can we learn from biomolecular networks. Brief Funct Genomics Proteomics 2:279–297
94. Huang S (2007) Cell fates as attractors – stability and flexibility of cellular phenotype. In: Endothelial biomedicine, 1st edn, Cambridge University Press, New York, pp 1761–1779
95. Huang S, Ingber DE (2000) Shape-dependent control of cell growth, differentiation, and apoptosis: switching between attractors in cell regulatory networks. Exp Cell Res 261:91–103
96. Huang S, Ingber DE (2006) A non-genetic basis for cancer progression and metastasis: self-organizing attractors in cell regulatory networks. Breast Dis 26:27–54
97. Huang S, Wikswo J (2006) Dimensions of systems biology. Rev Physiol Biochem Pharmacol 157:81–104
98. Huang S, Eichler G, Bar-Yam Y, Ingber DE (2005) Cell fates as high-dimensional attractor states of a complex gene regulatory network. Phys Rev Lett 94:128701
99. Huang S, Guo YP, May G, Enver T (2007) Bifurcation dynamics of cell fate decision in bipotent progenitor cells. Dev Biol 305:695–713
100. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH (2000) Functional discovery via a compendium of expression profiles. Cell 102:109–26
101. Hume DA (2000) Probability in transcriptional regulation and its implications for leukocyte differentiation and inducible gene expression. Blood 96:2323–8
102. Ihmels J, Bergmann S, Barkai N (2004) Defining transcription modules using large-scale gene expression data. Bioinformatics 20:1993–2003
103. Jablonka E, Lamb MJ (2002) The changing concept of epigenetics. Ann N Y Acad Sci 981:82–96
104. Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411:41–42
105. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316:1497–502
106. Jordan IK, Wolf YI, Koonin EV (2003) No simple dependence between protein evolution rate and the number of protein–protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol Biol 3:1
107. Joy MP, Brock A, Ingber DE, Huang S (2005) High-betweenness proteins in the yeast protein interaction network. J Biomed Biotechnol 2005:96–103
108. Kaern M, Elston TC, Blake WJ, Collins JJ (2005) Stochasticity in gene expression: from theories to phenotypes. Nat Rev Genet 6:451–64
109. Kaplan D, Glass L (1995) Understanding Nonlinear Dynamics, 1st edn. Springer, New York
110. Kashiwagi K, Urabe I, Kancko K, Yomo T (2006) Adaptive response of a gene network to environmental changes by fitness-induced attractor selection. PLoS One, 1:e49
111. Kauffman S (1969) Homeostasis and differentiation in random genetic control networks. Nature 224:177–8
112. Kauffman S (2004) A proposal for using the ensemble approach to understand genetic regulatory networks. J Theor Biol 230:581–90
113. Kauffman S, Peterson C, Samuelsson B, Troein C (2003) Random Boolean network models and the yeast transcriptional network. Proc Natl Acad Sci USA 100:14796–9
114. Kauffman SA (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor Biol 22:437–467
115. Kauffman SA (1991) Antichaos and adaptation. Sci Am 265:78–84
116. Kauffman SA (1993) The origins of order. Oxford University Press, New York
117. Khorasanizadeh S (2004) The nucleosome: from genomic organization to genomic regulation. Cell 116:259–72
118. Kim KY, Wang J (2007) Potential energy landscape and robustness of a gene regulatory network: toggle switch. PLoS Comput Biol 3:e60
119. Klemm K, Bornholdt S (2005) Stable and unstable attractors in Boolean networks. Phys Rev E Stat Nonlin Soft Matter Phys 72:055101
120. Klevecz RR, Bolen J, Forrest G, Murray DB (2004) A genomewide oscillation in transcription gates DNA replication and cell cycle. Proc Natl Acad Sci USA 101:1200–5
121. Kloster M, Tang C, Wingreen NS (2005) Finding regulatory modules through large-scale gene-expression data analysis. Bioinformatics 21:1172–9
122. Kouzarides T (2007) Chromatin modifications and their function. Cell 128:693–705
123. Kramer BP, Fussenegger M (2005) Hysteresis in a synthetic mammalian gene network. Proc Natl Acad Sci USA 102: 9517–9522
124. Krawitz P, Shmulevich I (2007) Basin entropy in Boolean network ensembles. Phys Rev Lett 98:158701
125. Krysinska H, Hoogenkamp M, Ingram R, Wilson N, Tagoh H, Laslo P, Singh H, Bonifer C (2007) A two-step, PU.1-dependent mechanism for developmentally regulated chromatin remodeling and transcription of the c-fms gene. Mol Cell Biol 27:878–87
126. Kubicek S, Jenuwein T (2004) A crack in histone lysine methylation. Cell 119:903–6
127. Laslo P, Spooner CJ, Warmflash A, Lancki DW, Lee HJ, Sciammas R, Gantner BN, Dinner AR, Singh H (2006) Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. Cell 126:755–66
128. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 298:799–804
129. Levsky JM, Singer RH (2003) Gene expression and the myth of the average cell. Trends Cell Biol 13:4–6
130. Li F, Long T, Lu Y, Ouyang Q, Tang C (2004) The yeast cell-cycle network is robustly designed. Proc Natl Acad Sci USA 101:4781–6
131. Li H, Xuan J, Wang Y, Zhan M (2008) Inferring regulatory networks. Front Biosci 13:263–75
132. Lim HN, van Oudenaarden A (2007) A multistep epigenetic switch enables the stable inheritance of DNA methylation states. Nat Genet 39:269–75
133. Luo F, Yang Y, Chen CF, Chang R, Zhou J, Scheuermann RH (2007) Modular organization of protein interaction networks. Bioinformatics 23:207–14

134. Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. Nature 431:308–12
135. MacCarthy T, Pomiankowski A, Seymour R (2005) Using large-scale perturbations in gene network reconstruction. BMC Bioinformatics 6:11
136. Mangan S, Alon U (2003) Structure and function of the feed-forward loop network motif. Proc Natl Acad Sci USA 100:11980–5
137. Manke T, Demetrius L, Vingron M (2006) An entropic characterization of protein interaction networks and cellular robustness. JR Soc Interface 3:843–50
138. Marcotte EM (2001) The path not taken. Nat Biotechnol 19:626–627
139. Margolin AA, Califano A (2007) Theory and limitations of genetic network inference from microarray data. Ann N Y Acad Sci 1115:51–72
140. Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. Science 296:910–3
141. Mattick JS (2007) A new paradigm for developmental biology. J Exp Biol 210:1526–47
142. May RM (1972) Will a large complex system be stable? Nature 238:413–414
143. Meissner A, Wernig M, Jaenisch R (2007) Direct reprogramming of genetically unmodified fibroblasts into pluripotent stem cells. Nat Biotechnol 25:1177–1181
144. Mellor J (2006) Dynamic nucleosomes and gene transcription. Trends Genet 22:320–9
145. Metzger E, Wissmann M, Schule R (2006) Histone demethylation and androgen-dependent transcription. Curr Opin Genet Dev 16:513–7
146. Mikkers H, Frisen J (2005) Deconstructing stemness. Embo J 24:2715–9
147. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298:824–7
148. Monod J, Jacob F (1961) Teleonomic mechanisms in cellular metabolism, growth, and differentiation. Cold Spring Harb Symp Quant Biol 26:389–401
149. Morceau F, Schnekenburger M, Dicato M, Diederich M (2004) GATA-1: friends, brothers, and coworkers. Ann N Y Acad Sci 1030:537–54
150. Morrison SJ, Uchida N, Weissman IL (1995) The biology of hematopoietic stem cells. Annu Rev Cell Dev Biol 11:35–71
151. Murray JD (1989) Mathematical biology, 2nd edn (1993). Springer, Berlin
152. Newman MEJ (2003) The structure and function of complex networks. SIAM Review 45:167–256
153. Nykter M, Price ND, Aldana M, Ramsey SA, Kauffman SA, Hood L, Yli-Harja O, Shmulevich I (2008) Gene expression dynamics in the macrophage exhibit criticality. Proc Natl Acad Sci USA 105:1897–900
154. Nykter M, Price ND, Larjo A, Aho T, Kauffman SA, Yli-Harja O, Shmulevich I (2008) Critical networks exhibit maximal information diversity in structure-dynamics relationships. Phys Rev Lett 100:058702
155. Odom DT, Zizlsperger N, Gordon DB, Bell GW, Rinaldi NJ, Murray HL, Volkert TL, Schreiber J, Rolfe PA, Gifford DK, Fraenkel E, Bell GI, Young RA (2004) Control of pancreas and liver gene expression by HNF transcription factors. Science 303:1378–81
156. Okita K, Ichisaka T, Yamanaka S (2007) Generation of germline-competent induced pluripotent stem cells. Nature 448:313–7
157. Ozbudak EM, Thattai M, Lim HN, Shraiman BI, Van Oudenaarden A (2004) Multistability in the lactose utilization network of Escherichia coli. Nature 427:737–740
158. Pennisi E (2003) Human genome. A low number wins the GeneSweep Pool. Science 300:1484
159. Picht P (1969) Mut zur utopie. Piper, München
160. Proulx SR, Promislow DE, Phillips PC (2005) Network thinking in ecology and evolution. Trends Ecol Evol 20:345–53
161. Raff M (2003) Adult stem cell plasticity: fact or artifact? Annu Rev Cell Dev Biol 19:1–22
162. Ralston A and Rossant J (2005) Genetic regulation of stem cell origins in the mouse embryo. Clin Genet 68:106–12
163. Ramo P, Kesseli J, Yli-Harja O (2006) Perturbation avalanches and criticality in gene regulatory networks. J Theor Biol 242:164–70
164. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. Science 297:1551–5
165. Reik W, Dean W (2002) Back to the beginning. Nature 420:127
166. Resendis-Antonio O, Freyre-Gonzalez JA, Menchaca-Mendez R, Gutierrez-Rios RM, Martinez-Antonio A, Avila-Sanchez C, Collado-Vides J (2005) Modular analysis of the transcriptional regulatory network of E. coli. Trends Genet 21:16–20
167. Robins H, Krasnitz M, Barak H, Levine AJ (2005) A relative-entropy algorithm for genomic fingerprinting captures host-phage similarities. J Bacteriol 187:8370–4
168. Roeder I, Glauche I (2006) Towards an understanding of lineage specification in hematopoietic stem cells: a mathematical model for the interaction of transcription factors GATA-1 and PU.1. J Theor Biol 241:852–65
169. Salgado H, Santos-Zavaleta A, Gama-Castro S, Peralta-Gil M, Penaloza-Spinola MI, Martinez-Antonio A, Karp PD, Collado-Vides J (2006) The comprehensive updated regulatory network of Escherichia coli K-12. BMC Bioinformatics 7:5
170. Samonte RV, Eichler EE (2002) Segmental duplications and the evolution of the primate genome. Nat Rev Genet 3:65–72
171. Sandberg R, Ernberg I (2005) Assessment of tumor characteristic gene expression in cell lines using a tissue similarity index (TSI). Proc Natl Acad Sci USA 102:2052–7
172. Shivdasani RA (2006) MicroRNAs: regulators of gene expression and cell differentiation. Blood 108:3646–53
173. Shmulevich I, Kauffman SA (2004) Activities and sensitivities in boolean network models. Phys Rev Lett 93:048701
174. Shmulevich I, Kauffman SA, Aldana M (2005) Eukaryotic cells are dynamically ordered or critical but not chaotic. Proc Natl Acad Sci USA 102:13439–44
175. Siegal ML, Promislow DE, Bergman A (2007) Functional and evolutionary inference in gene networks: does topology matter? Genetica 129:83–103
176. Smith MC, Sumner ER, Avery SV (2007) Glutathione and Gts1p drive beneficial variability in the cadmium resistances of individual yeast cells. Mol Microbiol 66:699–712
177. Southall TD, Brand AH (2007) Chromatin profiling in model organisms. Brief Funct Genomic Proteomic 6:133–40
178. Southan C (2004) Has the yo-yo stopped? An assessment of human protein-coding gene number. Proteomics 4:1712–26

179. Stern CD (2000) Conrad H. Waddington's contributions to avian and mammalian development, 1930–1940. Int J Dev Biol 44:15–22
180. Strohman R (1994) Epigenesis: the missing beat in biotechnology? Biotechnology (N Y) 12:156–64
181. Stumpf MP, Wiuf C, May RM (2005) Subnets of scale-free networks are not scale-free: sampling properties of networks. Proc Natl Acad Sci USA 102:4221–4
182. Suzuki M, Yamada T, Kihara-Negishi F, Sakurai T, Hara E, Tenen DG, Hozumi N, Oikawa T (2006) Site-specific DNA methylation by a complex of PU.1 and Dnmt3a/b. Oncogene 25:2477–88
183. Swiers G, Patient R, Loose M (2006) Genetic regulatory networks programming hematopoietic stem cells and erythroid lineage specification. Dev Biol 294:525–40
184. Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126:663–76
185. Tapscott SJ (2005) The circuitry of a master switch: Myod and the regulation of skeletal muscle gene transcription. Development 132:2685–95
186. Taylor JS, Raes J (2004) Duplication and divergence: The evolution of new genes and old ideas. Annu Rev Genet 38: 615–643
187. Teichmann SA, Babu MM (2004) Gene regulatory network growth by duplication. Nat Genet 36:492–6
188. Thieffry D, Huerta AM, Perez-Rueda E, Collado-Vides J (1998) From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. Bioessays 20:433–40
189. Tinbergen N (1952) Derived activities; their causation, biological significance, origin, and emancipation during evolution. Q Rev Biol 27:1–32
190. Toh H, Horimoto K (2002) Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. Bioinformatics 18:287–97
191. Trojer P, Reinberg D (2006) Histone lysine demethylases and their impact on epigenetics. Cell 125:213–7
192. Tyson JJ, Chen KC, Novak B (2003) Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. Curr Opin Cell Biol 15:221–231
193. van Helden J, Wernisch L, Gilbert D, Wodak SJ (2002) Graph-based analysis of metabolic networks. Ernst Schering Res Found Workshop:245–74
194. van Nimwegen E (2003) Scaling laws in the functional content of genomes. Trends Genet 19:479–84
195. Vogel G (2003) Stem cells. 'Stemness' genes still elusive. Science 302:371
196. Waddington CH (1940) Organisers and genes. Cambridge University Press, Cambridge
197. Waddington CH (1956) Principles of embryology. Allen and Unwin Ltd, London
198. Waddington CH (1957) The strategy of the genes. Allen and Unwin, London
199. Watts DJ (2004) The "new" science of networks. Ann Rev Sociol 20:243–270
200. Webster G, Goodwin BC (1999) A structuralist approach to morphology. Riv Biol 92:495–8
201. Wernig M, Meissner A, Foreman R, Brambrink T, Ku M, Hochedlinger K, Bernstein BE, Jaenisch R (2007) In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. Nature 448:318–24
202. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, Brown PO, Botstein D (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol Biol Cell 13:1977–2000
203. Wilkins AS (2007) Colloquium Papers: Between "design" and "bricolage": Genetic networks, levels of selection, and adaptive evolution. Proc Natl Acad Sci USA 104 Suppl 1:8590–6
204. Wuensche A (1998) Genomic regulation modeled as a network with basins of attraction. Pac Symp Biocomput:89–102
205. Xiong W, Ferrell JE Jr. (2003) A positive-feedback-based bistable 'memory module' that governs a cell fate decision. Nature 426:460–465
206. Xu X, Wang L, Ding D (2004) Learning module networks from genome-wide location and expression data. FEBS Lett 578:297–304
207. Yu H, Greenbaum D, Xin Lu H, Zhu X, Gerstein M (2004) Genomic analysis of essentiality within protein networks. Trends Genet 20:227–31
208. Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS Comput Biol 3:e59
209. Yuh CH, Bolouri H, Davidson EH (2001) Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. Development 128:617–29

Books and Reviews

Huang S (2004) Back to the biology in systems biology: what can we learn from biomolecular networks. Brief Funct Genomics Proteomics 2:279–297

Huang S (2007) Cell fates as attractors – stability and flexibility of cellular phenotype. In: Endothelial biomedicine, 1st edn. Cambridge University Press, New York, pp 1761–1779

Huang S, Ingber DE (2006) A non-genetic basis for cancer progression and metastasis: self-organizing attractors in cell regulatory networks. Breast Dis 26:27–54

Kaneko K (2006) Life: An introduction to complex systems biology, 1edn. Springer, Berlin

Kauffman SA (1991) Antichaos and adaptation. Sci Am 265:78–84

Kauffman SA (1993) The origins of order. Oxford University Press, New York

Kauffman SA (1996) At home in the universe: the search for the laws of self-organization and complexity. Oxford University Press, New York

Laurent M, Kellershohn N (1999) Multistability: a major means of differentiation and evolution in biological systems. Trends Biochem Sci 24:418–422

Wilkins AS (2007) Colloquium papers: Between "design" and "bricolage": Genetic networks, levels of selection, and adaptive evolution. Proc Natl Acad Sci USA 104 Suppl 1:8590–6

Complexity in Earthquakes, Tsunamis, and Volcanoes, and Forecast, Introduction to

William H. K. Lee
US Geological Survey (Retired), Menlo Park, USA

Article Outline

Introduction
Earthquakes
Tsunamis
Volcanoes
Discussions

This Introduction is intended to serve as a 'road map' for readers to navigate through the 42 Encyclopedia articles on earthquakes, tsunamis, and volcanoes. Selecting the topics and authors was somewhat subjective, as it is not possible to cover the vast existing literature with only 42 articles. They are, however, representative of the wide range of problems investigated in connection with these natural phenomena. I will introduce these articles by grouping them into sections and then into subsections. However, some articles belong to more than one section or one subsection, reflecting the inter-related nature of earthquakes, tsunamis and volcanoes. For the benefit of the readers, I will point to certain issues discussed in some of the articles which, in my view, have not been settled completely. I have also taken the liberty of quoting or paraphrasing sentences from many of these articles when introducing them, but I do not claim to be accurate. It is best for these articles to speak for themselves.

I wish to thank Bernard Chouet for helping me in planning and reviewing the manuscripts of the volcanoes section. I am grateful to Bernard Chouet, Edo Nyland, Jose Pujol, Chris Stephens, and Ta-liang Teng for their helpful comments that greatly improved this manuscript.

Introduction

Earthquakes, tsunamis, and volcanic eruptions are complex and often inter-related natural phenomena with disastrous impact to society rivaling those caused by the worst floods or storms. The 1556 Huaxian earthquake in the Shansi province of China claimed over 830,000 lives. The total economic loss of the 1995 Kobe earthquake in Japan was estimated at US $200 billion. The 2004 Indian Ocean tsunami (triggered by the Sumatra–Andaman earthquake of December 26) bought devastation thousands of miles away with fatalities exceeding 280,000. The 79 AD eruption of Mount Vesuvius near Naples, Italy buried the towns of Pompeii and Herculaneum. The 1902 eruption of Mount Pelée, Martinique, totally destroyed the town of St. Pierre.

Insurance companies classify major natural catastrophes as storms, floods, or earthquakes (including tsunamis, and volcanic eruptions). Since 1950, about 2.5 million people have died due to these catastrophes and overall economic losses have totaled about US $2 trillion in current dollar values. Earthquakes, tsunamis, and volcanic eruptions have accounted for about half of the fatalities and more than one third of the total economic losses. Geoscientists have attempted to predict these events, but with limited success. There are many reasons for such slow progress: (1) systematic monitoring of earthquakes, tsunamis and volcanoes requires large capital investment for instruments and very long-term support for operation and maintenance; (2) catastrophic earthquakes, tsunamis and volcanic eruptions occur rarely, and (3) politicians and citizens are quick to forget these hazards in the face of other more frequent and pressing issues. But with continuing rapid population growth and urbanization, the loss potential from these natural hazards in the world is quickly escalating.

With advances in nonlinear dynamics and complexity studies, geoscientists have applied modern nonlinear techniques and concepts such as chaos, fractal, critical phenomena, and self-organized criticality to the study of earthquakes, tsunamis and volcanoes. Here we sample these efforts, mainly in seismicity modeling for earthquake prediction and forecast, along with articles that review recent progress in studying earthquakes, tsunamis, and volcanoes. Although predictability is desirable, it is also possible to reduce these natural hazards with more practical approaches, such as early warning systems, hazard analysis, engineering considerations, and other mitigation efforts. Several articles in this Encyclopedia discuss these practical solutions.

Earthquakes

When a sudden rupture occurs in the Earth, seismic waves are generated. When these waves reach the Earth's surface, we may feel them as a series of vibrations, which we call an earthquake. Instrumental recordings of earthquakes have been made since the latter part of the 19th century by seismographic stations and networks from local to global scales. The observed data have been used, for example, (1) to compute the source parameters of earthquakes, (2) to determine the physical properties of the Earth's interior, (3) to test the theory of plate tectonics, (4) to map active faults, (5) to infer the nature of damaging ground shaking, (6) to carry out seismic hazard analyzes, and (7) to predict and forecast earthquakes. A satisfactory theory of the complex earthquake process has not yet been achieved, and realistic equations for modeling earthquakes do not exist at present. There is, however, good progress towards a physical foundation for the earth-

quake source process, partly as a result of research directed toward earthquake prediction.

Earthquake Monitoring, and Probing the Earth's Interior

Earthquakes are complex natural phenomena, and their monitoring requires an interdisciplinary approach, including using tools from other scientific disciplines and engineering. In ► Earthquake Monitoring and Early Warning Systems, W.H.K. Lee and Y.M. Wu presented a summary of earthquake monitoring, a description of the products derived from the analysis of seismograms, and a discussion of the limitations of these products. The basic results of earthquake monitoring are summarized in earthquake catalogs, which are lists of origin time, hypocenter location, and magnitude of earthquakes, as well as other source parameters. Lee and Wu describe the traditional earthquake location method formulated as an inverse problem. In ► Earthquake Location, Direct, Global-Search Methods, Lomax et al. review a different approach using direct-search over a space of possible locations, and discuss other related algorithms. Direct-search earthquake location is important because, relative to the traditional linearized method, it is both easier to apply to more realistic Earth models and is computational more stable. Although it has not been widely applied because of its computational demand, it shows great promise for the future as computer power is advancing rapidly.

The most frequently determined parameter after 'location' is 'magnitude', which is used to characterize the 'size' of an earthquake. A brief introduction to the quantification of earthquake size, including magnitude and seismic moment, is given in ► Earthquake Monitoring and Early Warning Systems by Lee and Wu. Despite its various limitations, magnitude provides important information concerning the earthquake source. Magnitude values have an immense practical value for realistic long-term disaster preparedness and risk mitigation efforts. A detailed review, including current practices for magnitude determinations, appears in ► Earthquake Magnitude by P. Bormann and J. Saul.

Besides computing earthquake source parameters, earthquake monitoring also provides data that can be used to probe the Earth's interior. In ► Tomography, Seismic, J. Pujol reviews a number of techniques designed to investigate the interior of the Earth using arrival times and/or waveforms from natural and artificial sources. The most common product of a tomographic study is a seismic velocity model, although other parameters, such as attenuation and anisotropy, can also be estimated. Seismic tomography generally has higher resolution than that provided by other geophysical methods, such as gravity and magnetics, and furnishes information (1) about fundamental problems concerning the internal structure of the Earth on a global scale, and (2) for tectonic and seismic hazard studies on a local scale.

In ► Seismic Wave Propagation in Media with Complex Geometries, Simulation of, H. Igel et al. present the state-of-the-art in computational wave propagation. They point to future developments, particularly in connection with the search for efficient generation of computational grids for models with complex topography and faults, as well as for the combined simulation of soil-structure interactions. In addition to imaging subsurface structure and earthquake sources, 3-D wave simulations can forecast strong ground motions from large earthquakes. In the absence of deterministic prediction of earthquakes, the calculation of earthquake scenarios in regions with sufficiently well-known crustal structures and faults will play an important role in assessing and mitigating potential damage, particularly those due to local site effects.

In addition to the classical parametrization of the Earth as a layered structure with smooth velocity perturbation, a new approach using scattered waves that reflect Earth's heterogeneity is introduced by H. Sato in his article on ► Seismic Waves in Heterogeneous Earth, Scattering of. For high-frequency seismograms, envelope characteristics such as the excitation level and the decay gradient of coda envelopes and the envelope broadening of the direct wavelet are useful for the study of small-scale inhomogeneities within the Earth. The radiative transfer theory with scattering coefficients calculated from the Born approximation and the Markov approximation for the parabolic wave equation are powerful mathematical tools for these analyzes. Studies of the scattering of high-frequency seismic waves in the heterogeneous Earth are important for understanding the physical structure and the geodynamic processes that reflect the evolution of the solid Earth.

Earthquake Prediction and Forecasting

A fundamental question in earthquake science is whether earthquake prediction is possible. Debate on this question has been going on for decades without clear resolution. Are pure observational methods without specific physical understanding sufficient? Earthquakes have been instrumentally monitored continuously for about 100 years (although not uniformly over the Earth), but reliable and detailed earthquake catalogs cover only about 50 years. Consequently, it seems questionable that earthquakes can be

predicted solely on the basis of observed seismicity patterns, given that large earthquakes in a given region have recurrence intervals ranging from decades to centuries or longer. Despite progress made in earthquake physics, we are still not able to write down all the governing equations for these events and lack sufficient information about the Earth's properties. Nevertheless, many attempts have been and are being made to predict and forecast earthquakes. In this section, several articles based on empirical and physics-based approaches will be briefly introduced.

In ► Geo-complexity and Earthquake Prediction, V. Keilis-Borok et al. present an algorithmic prediction method for individual extreme events having low probability but large societal impact. They show that the earthquake prediction problem is necessarily intertwined with problems of disaster preparedness, the dynamics of the solid Earth, and the modeling of extreme events in hierarchical complex systems. The algorithms considered by Keilis-Borok et al. are based on premonitory seismicity patterns and provide alarms lasting months to years. Since the 1990s, these alarms have been posted for use in testing such algorithms against newly occurred large earthquakes. Some success has been achieved, and although the areas for the predicted earthquakes are very large and the predicted time windows are very long, such predictions can be helpful for the officials and the public to undertake appropriate preparedness.

Stochastic models are a practical way of bridging the gap between the detailed modeling of a complex system and the need to fit models to limited data. In ► Earthquake Occurrence and Mechanisms, Stochastic Models for, D. Vere-Jones presents a brief account of the role and development of stochastic models of seismicity, from the first empirical studies to current models used in earthquake probability forecasting. The author combines a model of the physical processes generating the observable data (earthquake catalogs) with a model for the errors, or uncertainties, in our ability to predict those observables.

D.A. Yuen et al. propose the use of statistical approaches and data-assimilation techniques to earthquake forecasting in their article on ► Earthquake Clusters over Multi-dimensional Space, Visualization of. The nature of the spatial-temporal evolution of earthquakes may be assessed from the observed seismicity and geodetic measurements by recognizing nonlinear patterns hidden in the vast amount of seemingly unrelated data. The authors endeavor to bring across the basic concept of clustering and its role in earthquake forecasting, and conclude that the clustering of seismic activity reflects both the similarity between clusters and their correlation properties.

In ► Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, G. Zoeller et al. present a combined approach to understanding seismicity and the emergence of patterns in the occurrence of earthquakes based on numerical modeling and data analysis. The discussion and interpretation of seismicity in terms of statistical physics leads to the concept of 'critical states', i. e. states in the seismic cycle with an increased probability for abrupt changes involving large earthquakes. They demonstrate that numerical fault models are valuable for understanding the underlying mechanisms of observed seismicity patterns, as well as for practical estimates of future seismic hazard.

D. Sornette and M.J. Werner in ► Seismicity, Statistical Physics Approaches to stress that the term 'statistical' in 'statistical physics' has a different meaning than as used in 'statistical seismology'. Statistical seismology has been developed as a marriage between probability theory, statistics, and earthquake occurrences without considerations of earthquake physics. In statistical physics approaches to seismicity, researchers strive to derive statistical models from microscopic laws of friction, damage, rupture, etc. Sornette and Werner summarize some of the concepts and tools that have been developed, including the leading theoretical physical models of the space-time organization of earthquakes. They then present several examples of the new metrics proposed by statistical physicists, underlining their strengths and weaknesses. They conclude that a holistic approach emphasizing the interactions between earthquakes and faults is promising, and that statistical seismology needs to evolve into a genuine physically-based statistical physics of earthquakes.

In ► Earthquake Networks, Complex, S. Abe and N. Suzuki discuss the construction of a complex earthquake network obtained by mapping seismic data to a growing stochastic graph. This graph, or network, turns out to exhibit a number of remarkable physical and mathematical behaviors that share common traits with many other complex systems. The scale-free and small-world natures are typical examples in complex earthquake networks.

Electromagnetic phenomena associated with earthquakes, such as earthquake light have been reported throughout almost all human history. Until rather recently, however, most such observations were unreliable and best described as folklore. In ► Earthquakes, Electromagnetic Signals of, S. Uyeda et al. summarize the scientific search for electromagnetic precursors for earthquake prediction. The presumption is that since earthquakes occur when slowly increasing tectonic stress in the Earth's crust reaches a critical level; the same stress may give rise to some electromagnetic phenomena. Research on possi-

ble relationships was initiated in several countries around the world in the 1980s. Two main approaches are (1) the monitoring of possible emissions from focal regions in a wide range of frequency from DC to VHF, and (2) the monitoring of possible anomalies in the transmission of man-made electromagnetic waves of various frequencies over focal regions. Despite much circumstantial evidence, earthquake-related electromagnetic signals, in particular those at a pre-seismic stage are not yet widely accepted to be associated with earthquakes.

Observational programs focused on searching for reliable precursory phenomena in seismicity, seismic velocities, tilt and strain, electromagnetic signals, chemical emissions and animal behavior, claim some successes but no systematic precursors have been identified. In ► Earthquake Forecasting and Verification, J.R. Holliday et al. stress that reliable earthquake forecasting will require systematic verification. They point out that although earthquakes are complex phenomena, systematic scaling laws such as the Gutenberg–Richter frequency-magnitude relation have been recognized. The Gutenberg–Richter relation is given by: $\log N(M) = a - bM$, where M is the earthquake magnitude, $N(M)$ is the number of earthquakes with magnitude greater than or equal to M, and a and b are constants. Since $b \approx 1$, this means that the number of earthquakes increase tenfold for each decrease of one magnitude unit. This suggests that *large* earthquakes occur in regions where there are large numbers of *small* earthquakes. On this basis, the regions where large earthquakes will occur can be forecast with considerable accuracy, but the Gutenberg–Richter relation provides no information about the precise occurrence times.

Earthquake Engineering Considerations and Early Warning Systems

Since seismic hazards exist in many regions of the world, three major strategies are introduced to reduce their societal impacts: (1) to avoid building in high seismic-risk areas, (2) to build structures that can withstand the effects of earthquakes, and (3) to plan for earthquake emergencies. The first strategy is not very practical because, with rapid population growth, many economically productive activities are increasingly located in high seismic-risk areas. However, by mapping active faults and by studying past earthquakes, we may estimate the risk potential from earthquakes and plan our land use accordingly. The second strategy depends on the skills of engineers, and also requires seismologists to provide realistic estimates of the ground motions resulting from expected earthquakes. The third strategy includes attempting to predict earthquakes reliably well in advance to minimize damage and casualties, and also requires the cooperation of the entire society. Although we are far from being able to predict earthquakes reliably, earthquake early warning systems can provide critical information to reduce damage and causalities, as well as to aid rescuing and recovery efforts.

Accurate prediction of the level and variability of near-source strong-ground motions in future earthquakes is one of the key challenges facing seismologists and earthquake engineers. The increasing number of near-source recordings collected by dense strong-motion networks exemplifies the inherent complexity of near-field ground shaking, which is governed by a number of interacting physical processes. Characterizing, quantifying, and modeling ground-motion complexity requires a joint investigation of (1) the physics of earthquake rupture, (2) wave-propagation in heterogeneous media, and (3) the effects of local site conditions. In ► Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, P.M. Mai discusses briefly the beginnings of strong-motion seismology and the recognition of ground-motion complexity. Using two well-recorded recent earthquakes, the author introduces the observational aspects of near-field ground shaking and describes the basic mathematical tools used in the computation of ground motion. The key elements for characterizing and modeling ground-motion complexity are also explained, supplemented by a concise overview of the underlying physical processes.

With increasing urbanization worldwide, earthquake hazards pose ever greater threats to lives, property, and livelihoods in populated areas near major active faults on land or near offshore subduction zones. Earthquake early-warning (EEW) systems can be useful tools for reducing the impact of earthquakes, provided that the populated areas are favorably located with respect to earthquake sources and their citizens are properly trained to respond to the warning messages. Under favorable conditions, an EEW system can forewarn an urban population of impending strong shaking with lead times that range from a few seconds to a few tens of seconds. A lead time is the time interval between issuing a warning and the arrival of the *S*- and surface waves, which are the most destructive due to their large amplitudes. Even a few seconds of advance warning is useful for pre-programmed emergency measures at various critical facilities, such as the deceleration of rapid-transit vehicles and high-speed trains, the orderly shutoff of gas pipelines, the controlled shutdown of some high-technological manufacturing operations, the safe-guarding of computer facilities, and bringing elevators to a stop at the nearest floor.

Recent advances in early warning methodologies are summarized by W.H.K Lee and Y.M. Wu in the second part of their article, ► Earthquake Monitoring and Early Warning Systems. In ► Earthquake Early Warning System in Southern Italy, A. Zollo et al. analyze and illustrate the main scientific and technological issues related to the implementation and management of the earthquake early warning system under development in the Campania region of southern Italy. The system is designed to issue alerts to distant coastal targets using data from a dense seismic network deployed in the Apennine belt region. The authors note that earthquake early warning systems can also help mitigate the effects of earthquake-induced disasters such as fires, explosions, landslides, and tsunamis. Earthquake early warning systems can be installed at relatively low cost in developing countries, where even moderate-size earthquakes can cause damage comparable to that caused by much larger earthquakes in developed countries.

Nonlinear problems in structural earthquake engineering deal with the dynamic response of meta-stable, man-made buildings subjected to strong earthquake shaking. During earthquakes, structures constructed on soft sediments and soils deform together with the underlying soil. Strong shaking forces the soil-structure systems to evolve through different levels of nonlinear response, with continuously changing properties that depend upon the time history of excitation and on the progression and degree of damage. In ► Earthquake Engineering, Non-linear Problems in, M.D. Trifunac first briefly discuss the literature on complex and chaotic dynamics of simple mechanical oscillators, and then introduces the dynamic characteristics and governing equations of the meta-stable structural dynamics in earthquake engineering. He describes the nature of the solutions of the governing equations in terms of both the vibrational and the wave representations. The author also addresses the dynamic instability, material and geometric nonlinearities, and complexities of the governing equations associated with nonlinear soil-structure interaction.

Structural health monitoring and structural damage detection refers to the processes of determining and tracking the structural integrity and assessing the nature of damage in a structure. An important and challenging problem is being able to detect the principal components of damage in structures (as they occur during or soon after the earthquake) before physical inspection. In the article, ► Earthquake Damage: Detection and Early Warning in Man-Made Structures, M.I. Todorovska focuses on global methods and intermediate-scale methods, which can point to the parts of the structure that have been damaged. Recently, structural identification and health monitoring of buildings based on detecting changes in wave travel time through the structure has received renewed attention and has proven to be very promising.

Earthquake Physics

Brittle deformation, which is the primary mode of deformation of the Earth's crust in response to tectonic stress, is manifested by faulting at the long timescale and by earthquakes at the short timescale. It is one of the best-known examples of a system exhibiting self-organized criticality. A full understanding of this system is essential for evaluating earthquake hazards, but our current understanding is sketchy. In ► Brittle Tectonics: A Non-linear Dynamical System, C.H. Scholz shows that an earthquake dynamic system has two characteristic length scales, W^* and W^{**}. An earthquake nucleates within the seismogenic zone and initially propagates in all directions along its perimeter, acting as a 3D crack. When its dimension exceeds W^*, the rupture is restricted to propagating in the horizontal direction, acting as a 2D crack. Thus a symmetry breakage occurs at the dimension W^*. *Small* earthquakes, with dimensions smaller than W^*, are not self-similar with *large* earthquakes, those with lengths larger than W^*. The same occurs for suprafaults at the dimension W^{**} (a suprafault is the shear relaxation structure that includes a fault and its associated ductile shear zone).

Earthquake prediction is desirable for reducing seismic hazards, but we lack an understanding of how and why earthquakes begin and grow larger or stop. Theoretical and laboratory studies show that a quasi-static rupture growth precedes dynamic rupture. Thus, detecting the quasi-static rupture growth may lead to forecasting the subsequent dynamic rupture. In ► Earthquake Nucleation Process, Y. Iio reviews studies that analyze the early portions of observed waveforms, and summarizes what we presently understand about earthquake nucleation process. An earthquake initiates over a small patch of a fault, and then their rupture fronts expand outward until they stop. Some large earthquakes have a rupture extent greater than 1000 km, while fault lengths of small microearthquakes range over only a few meters. Surprisingly, the concept that earthquakes are self-similar is widely accepted despite fault length ranging over 6 orders of magnitude. One example of such similarity is the proportionality of average fault slip to fault length, which implies a constant static stress drop, independent of earthquake size.

The self-similarity law raises a fundamental question, namely what is the difference between large and small earthquakes? One end-member model represents earth-

quakes as ruptures that grow randomly and then terminate at an earlier stage for smaller earthquakes, but continue longer for larger earthquakes. This type of model has been proposed mainly to explain the frequency–magnitude distribution of earthquakes (the Gutenberg–Richter relation), but it implies that it is impossible to forecast the final size of an earthquake at the time the rupture initiates. However, the other end-member model predicts that larger earthquakes have a larger 'seed' than smaller earthquakes, and that large and small earthquakes are different even at their beginnings.

Geoscientists have long sought an understanding of how earthquakes interact. Can earthquakes trigger other earthquakes? The answer is clearly yes over short time and distance scales, as in the case of mainshock–aftershock sequences. Over increasing time and distance scales, however, this question is more difficult to answer. In ► Earthquakes, Dynamic Triggering of, S.G. Prejean and D.P. Hill explore the most distant regions over which earthquakes can trigger other earthquakes. This subject has been the focus of extensive research over the past twenty five years, and offers a potentially important key to improving our understanding of earthquake nucleation. In this review, the authors discuss physical models and give a description of documented patterns of remote dynamic triggering.

Models of the earthquake source have been successfully used in predicting many of the general properties of seismic waves radiated from earthquakes. These general properties can be derived from a simple omega-squared spectral shape. In ► Earthquake Scaling Laws, R. Madariaga derives general expressions for energy, moment and stress in terms of measured spectral parameters, and shows that earthquake sources can be reduced to a single family with the three parameters of moment, corner frequency and radiated energy. He suggests that most of the properties of the seismic spectrum and slip distribution can be explained by a simple crack model. Whether an earthquake is modeled as a simple circular crack or as a complex distribution of such cracks, the result is the same.

In ► Earthquake Source: Asymmetry and Rotation Effects, R. Teisseyre presents a consistent theory describing an elastic continuum subjected to complex internal processes, considers all the possible kinds of the point-related motions and deformations, and defines a complex rotation field including spin and twist. Also included in the discussion is a new description of the source processes, including the role of rotation in source dynamics, an explanation of co-action of the slip and rotation motions, and a theory of seismic rotation waves. Rotational seismology is an emerging field, and a progress report is provided in the Appendix in ► Earthquake Monitoring and Early Warning Systems by W.H.K. Lee and Y.M. Wu.

Some New Tools to Study Earthquakes

The Global Positioning System (GPS) is a space-based Global Navigation Satellite System. Using signals transmitted by a constellation of GPS satellites, the positions of ground-based receivers can be calculated to high precision, making it possible to track relative movements of points on the Earth's surface over time. Unlike older geodetic surveying methods (which involved periodically but infrequent measuring angles, distances, or elevations between points), GPS can provide precise 3-D positions over a range of sampling rates and on a global scale. GPS equipment is easy to use and can be set up to collect data continuously. Since its early geophysical applications in the mid-1980s, this versatile tool, which can be used to track displacements over time periods of seconds to decades, has become indispensable for crustal deformation studies, leading to many important insights and some surprising discoveries. In ► GPS: Applications in Crustal Deformation Monitoring, J. Murray-Moraleda focuses on applications of GPS data to the studies of tectonic, seismic, and volcanic processes. The author presents an overview of how GPS works and how it is used to collect data for geophysical studies. The article also describes a variety of ways in which GPS data have been used to measure crustal deformation and investigate the underlying processes.

The concept of a seismic cycle involves processes associated with the accumulation and release of stress on seismogenic faults, and is commonly divided into three intervals: (1) the coseismic interval for events occurring during an earthquake, (2) the postseismic interval immediately following an earthquake, and (3) the interseismic period in between large earthquakes. In ► Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, R. Lohman explores how we can draw conclusions about fault zone slip at depths far greater than are directly accessible to us, based on how the Earth's surface deforms during, before, and after earthquakes.

Atmospheric sound can be radiated by the displacement or rupture of the Earth's surface induced by earthquakes, tsunamis, and volcanoes, and by the flow and excitation of fluids during volcanic eruptions. These complex and potentially cataclysmic phenomena share some common physics, yet represent different ways of converting energy into atmospheric sound. In ► Infrasound from Earthquakes, Tsunamis and Volcanoes, M. Garces and A. LePichon discuss some of the signal features unique to earthquakes, tsunamis, and volcanoes captured by the

present generation of infrasound arrays. They also discuss contemporary methods for the analysis, interpretation, and modeling of these diverse signals, and consider some of the associated geophysical problems that remain unsolved.

Tsunamis

Tsunamis are oceanic gravity waves generated by seafloor deformation due to earthquakes, volcanic eruptions, landslides, or asteroid impacts. Earthquake tsunamis, such as the 2004 Indian Ocean tsunami (caused by the Sumatra–Andaman earthquake of December 26), are the most frequent type of tsunamis. However, large volcanic eruptions, such as the 1883 Krakatau eruption (in the Sunda strait between the islands of Java and Sumatra) also cause ocean-wide tsunamis. Landslides (which are often triggered by earthquakes) cause large tsunamis locally, but their effects are usually limited to the immediate vicinity of the source.

Modeling: Forward and Inverse Approaches

Forward-modeling of a tsunami starts from given initial conditions, computes its propagation in the ocean, and calculates the tsunami arrival times and/or water run-up heights along the coasts. Once the initial conditions are provided, the propagation and coastal behavior can be numerically computed for an actual bathymetry. These calculations are useful for early tsunami warning and for detailed hazard estimations. However, the initial conditions associated with tsunami generation processes are still poorly known, because large tsunamis are rare and the tsunami generation in the open ocean is not directly observable. Currently, the tsunami source is estimated indirectly, mostly on the basis of seismological analysis, but a more direct estimation of the tsunami source is essential to better understand the tsunami generation process and to more accurately forecast the effects of a tsunami along the coasts.

In ► Tsunamis, Inverse Problem of, K. Satake reviews inverse methods used in the quantification of tsunami sources from the observations. The author describes the tsunami generation by earthquakes, with an emphasis on the fault parameters and their effects on tsunami propagation, including shallow water theory and numerical computation. The author then summarizes the tsunami observations, including instrumental sea-level data and run-up height estimates for modern, historical and prehistoric tsunamis. He also describes methods for modeling and quantifying a tsunami source, and for analyzing tsunami travel times, amplitudes and waveforms. He concludes with an estimation of earthquake fault parameters derived from waveform inversion of tsunami data, and a discussion of heterogeneous fault motion and its application for tsunami warning.

Tsunami inundation is the one of the final stages of tsunami evolution, when the wave encroaches upon and floods dry land. It is during this stage that a tsunami is most destructive and takes the vast majority of its victims. To gauge the near-shore impact of tsunami inundation, engineers and scientists rely primarily on three different methods: (1) field survey of past events, (2) physical experimentation in a laboratory, and (3) numerical modeling. In ► Tsunami Inundation, Modeling of, P.J. Lynett focuses on numerical simulations. He reviews tsunami generation and open ocean propagation, and discusses the physics of near-shore tsunami evolution, hydrodynamic modeling of tsunami evolution, moving shoreline algorithms, and effect of topographical features on inundation.

Tsunami Forecasting and Warning

The original definition of 'tsunami earthquake' was given by H. Kanamori (Phys Earth Planet Inter 6:346–359, 1972) as "an earthquake that produces a large-size tsunami relative to the value of its surface wave magnitude (M_S)". The true damage potential that a tsunami earthquake represents may not be recognized by conventional near real-time seismic analysis methods that utilize measurements of relatively high-frequency signals, and thus the threat may only become apparent upon the arrival of the tsunami waves on the local shores. Although tsunami earthquakes occur relatively infrequently, the effect on the local population can be devastating, as was most recently illustrated by the July 2006 Java tsunami earthquake, which was quickly followed by tsunami waves two to seven meters high, traveling as far as two kilometers inland and killing at least 668 people.

It is important to note that the definition of 'tsunami earthquake' is distinct from that of 'tsunamigenic earthquake'. A tsunamigenic earthquake is any earthquake that excites a tsunami. Tsunami earthquakes are a specific subset of tsunamigenic earthquakes. In ► Tsunami Earthquakes, J. Polet and H. Kanamori describe the characteristics of tsunami earthquakes and the possible factors involved in the anomalously strong excitation of tsunamis by these events. They also discuss a possible model for these infrequent, but potentially very damaging events.

Tsunamis are among nature's most destructive hazards. Typically generated by large, underwater shallow earthquakes, tsunamis can cross an ocean basin in a matter of hours. Although difficult to detect, and not dangerous

while propagating in open ocean, tsunamis can unleash awesome destructive power when they reach coastal areas. With advance warning, populations dwelling in coastal areas can be alerted to evacuate to higher ground and away from the coast, thus saving many lives.

Tsunami travels at about the same speed of a commercial airliner, however, seismic waves can travel at speeds more than 40 times greater. Because of this large disparity in speed, scientists rely on seismic methods to detect the possibility of tsunami generation and to warn coastal populations of an approaching tsunami well in advance of its arrival. The seismic P-wave for example, travels from Alaska to Hawaii in about 7 min, whereas a tsunami will take about 5.5 hours to travel the same distance. Although over 200 sea-level stations reporting in near-real time are operating in the Pacific Ocean, it may take an hour or more, depending on the location of the epicenter, before the existence (or not) of an actual tsunami generation is confirmed. In other ocean basins where the density of sea-level instruments reporting data in near real-time is less, the delay in tsunami detection is correspondingly longer. However, global, regional, and local seismic networks, and the infrastructure needed to process the large amounts of seismic data that they record, are well in place around the world. For these reasons, tsunami warning centers provide initial tsunami warnings to coastal populations based entirely on the occurrence of a large shallow offshore earthquake. It is well-known, however, that large shallow offshore earthquakes may or may not be tsunamigenic.

In ► Tsunami Forecasting and Warning, O. Kamigaichi discusses the complexity problem in tsunami forecasting for large local events, and describes the Tsunami Early Warning System in Japan. Tsunami disaster mitigation can be achieved effectively by the appropriate combination of software and hardware countermeasures. Important issues for disaster mitigation includes: (1) improving people's awareness of the tsunami hazards, (2) imparting the necessity of spontaneous evacuation when people notice an imminent threat of tsunami on their own (feeling strong shaking near the coast, seeing abnormal sea level change, etc), (3) giving clear directions on how to respond to the tsunami forecast, and (4) conducting tsunami evacuation drills. The author notes that in tsunami forecasting, a trade-off exists between promptness and accuracy/reliability.

In ► Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, B. Hirshorn and S. Weinstein describe the basic method used by the Pacific Tsunami Warning Center (PTWC) mainly for large teleseismic events. Software running at the PTWC processes in real time seismic signals from over 150 seismic stations worldwide provided by various seismic networks. Automatic seismic event detection algorithms page the duty scientists for any earthquake occurring worldwide over about Magnitude 5.5. Other automatic software locates these events, and provides a first estimate of their magnitude and other source parameters in near real time. Duty scientists then refine the software's automated source parameter estimates and issue a warning if necessary. The authors also describe their ongoing efforts to improve estimates of earthquake source parameters.

Wedge Mechanics, Submarine Landslides and Slow Earthquakes

A study of the mechanics of wedge-shaped geological bodies, such as accretionary prisms in subduction zones and fold-and-thrust belts in collision zones, is interesting because they enable us to use the observed morphology and deformation of these bodies to constrain properties of the thrust faults underlying them. The fundamental process described in wedge mechanics is how gravitational force, in the presence of a sloping surface, is balanced by basal stress and internal stress. The internal state of stress depends on the rheology of the wedge. The most commonly assumed wedge rheology for geological problems is perfect Coulomb plasticity, and the model based on this rheology is referred to as the Coulomb wedge model.

The connection between wedge mechanics and great earthquakes and tsunamis at subduction zones is an emerging new field of study. In their article, ► Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Wang et al. cover the topics of stable and critical Coulomb wedges, dynamic Coulomb wedge, stress drop and increase in a subduction earthquake, and tsunamigenic coseismic seafloor deformation. Better constraints are needed to quantify how stresses along different down-dip segments of the subduction fault evolve with time throughout an earthquake cycle and how the evolution impacts wedge and seafloor deformation. Submarine monitoring in conjunction with land-based monitoring at subduction zones that are currently in different phases of earthquake cycles will allow us to better understand the evolution of fault and wedge stresses during the interseismic period. In this regard, cabled seafloor monitoring networks including borehole observatories being designed or implemented at different subduction zones will surely yield valuable data in the near future.

The term 'submarine landslide' encompasses a multitude of gravitational mass failure features at areal scales from square meters to thousands of square kilometers. The

term 'slow earthquake' describes a discrete slip event that produces millimeter to meter-scale displacements identical to those produced during earthquakes but without the associated seismic shaking. Recently, a GPS network on the south flank of Kilauea volcano, Hawaii, recorded multiple slow earthquakes on the subaerial portion of a large landslide system that extends primarily into the submarine environment. Since catastrophic failure of submarine landslides can cause a tsunami they represent significant hazards to coastal zones. Because submarine landslide systems are among the most active as well as spatially confined deforming areas on Earth, they are excellent targets for understanding the general fault failure process. In ► Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, B.A. Brooks et al. present a review of this interdisciplinary topic of interest in geodesy, seismology, tsunamis, and volcanology.

Volcanoes

About 1,500 volcanoes have erupted one or more times during the past 10,000 years, and since A.D. 1600, volcanic disasters have killed about 300,000 people and resulted in property damage and economic loss exceeding hundreds of millions of dollar. Articles in this section are intended to summarize recent research in: (1) volcano seismology, (2) physical processes involved in volcanoes, and (3) modeling volcanic eruptions and hazards warning.

Volcano Seismology

Magma transport in a volcano is episodic due to the inherent instability of magmatic systems at all time scales. This episodicity is reflected in seismic activity, which originates in dynamic interactions between gas, liquid and solid along magma transport paths that involve complex geometries. The description of the flow processes is governed by the nonlinear equations of fluid dynamics. In volcanic fluids, further complexity arises from the strong nonlinear dependence of magma rheology on temperature, pressure, and water and crystal content, and nonlinear characteristics of associated processes underlying the physico-chemical evolution of liquid-gas mixtures constituting magma.

In ► Volcanoes, Non-linear Processes in, B. Chouet presents a brief review of volcano seismology and addresses basic issues in the quantitative interpretation of processes in active volcanic systems. Starting with an introduction of the seismic methodology used to quantify the source of volcano seismicity, the author then focuses on sources originating in the dynamics of volcanic fluids. A review of some of the representative source mechanisms of Long-Period (LP) and Very Long-Period (VLP) signals is followed by a description of a mesoscale computational approach for simulating two-phase flows of complex magmatic fluids. Refined understanding of magma and hydrothermal transport dynamics therefore requires multidisciplinary research involving detailed field measurements, laboratory experiments, and numerical modeling. Such research is fundamental to monitoring and interpreting the subsurface migration of magma that often leads to eruptions, and thus would enhance our ability to forecast hazardous volcanic activity.

Volcano seismicity produces a wide variety of seismic signals that provide glimpses of the internal dynamics of volcanic systems. Quantitative approaches to analysis and interpret volcano-seismic signals have been developed since the late 1970s. The availability of seismic equipments with wide frequency and dynamic ranges since the early 1990s has revealed a variety of volcano-seismic signals over a wide range of periods. Quantification of the sources of volcano-seismic signals is crucial to achieving a better understanding of the physical states and dynamics of magmatic and hydrothermal systems. In ► Volcano Seismic Signals, Source Quantification of, H. Kumagai provides the theoretical basis for a quantification of the sources of volcano-seismic signals. The author focuses on the phenomenological representation of seismic sources, waveform inversion to estimate source mechanisms, spectral analysis based on an autoregressive model, and physical properties of fluid-solid coupled waves.

Among various eruptive styles, Strombolian activity is easier to study because of its repetitive behavior. Since Strombolian activity offers numerous interesting seismic signals, a growing attention has been devoted to the application of waveform inversion for imaging conduit geometry and retrieving eruption dynamics from seismological recordings. Quantitative models fitting seismological observations are a powerful tool for interpreting seismic recordings from active volcanoes. In ► Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, L. D'Auria and M. Martini discuss the mechanism of generation of Very-Long Period (VLP) signals accompanying Strombolian explosions. This eruptive style, occurring at many basaltic volcanoes worldwide, is characterized by the ascent and the bursting of large gas slugs. The mechanism of formation, ascent and explosion of bubbles and slugs and their relation with eruptive activity has been studied theoretically and by analogue simulations. The authors report results from numerical simulations, focusing on the seismic signals generated by pressure variations applied to the conduit walls.

Physical Processes in Volcanoes

The dynamics of solid-liquid composite systems are relevant to many problems, including how melts or aqueous fluids migrate through the mantle and crust toward the surface, how deformation and fracture in these regions are influenced by the existence of fluids, and also how these fluids can be observed in seismic tomographic images. In ► Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in, Y. Takei introduces a general continuum mechanical theory for macroscopic dynamics of solid-liquid composite systems, and emphasizes on how such interactions with pore geometry can be studied. The author then discusses the determinability of porosity and pore geometry from seismic tomographic images, and presents a practical method to assess porosity and pore geometry from tomographic V_P and V_S images.

A volcano consists of solids, liquids, gases, and intermediate materials of any two of these phases. Mechanical and thermodynamical interactions between these phases are essential in the generating a variety of volcanic activities. In particular, the gas phase is mechanically distinct from the other phases and plays an important role in dynamic phenomena in volcanoes. In ► Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, M. Ichihara and T. Nishimura discuss several bubble dynamics phenomena from the viewpoint that a bubbly fluid acts as an impulse generator of observable signals, such as earthquakes, ground deformations, airwaves, and an eruption itself. The authors focus on the notion that the impulse is excited by non-linear coupling between internal processes in a bubbly fluid and an external perturbation. The importance of these processes has recently become noticed as a possible triggering mechanism of eruptions, earthquakes, and inflation of a volcano.

Our capability to mitigate volcano hazards relies in large part on forecasting explosive events, a process which requires a high degree of understanding of the physicochemical factors operating during explosive volcanism. The approaches taken to gain an understanding of explosive volcanism have relied on a combination of field observations, theoretical models and laboratory models of materials and mechanisms. In ► Volcanic Eruptions, Explosive: Experimental Insights, S.J. Lane and M.R. James first review aspects of the volcanic materials literature, with the aim of illustrating the nature of molten rock, the complexity of which underpins most explosive volcanic processes. Experimental modeling of these processes can then build on the materials understanding. Such experiments involve investigation of the behavior of natural volcanic products at laboratory time and length scales, including the response of magma samples to rapid changes in pressure and temperature, the fall behavior of silicate particles in the atmosphere, and the generation and separation of electrostatic charge during explosive eruptions.

In ► Volcanic Eruptions: Cyclicity During Lava Dome Growth, O. Melnik et al. consider the process of slow extrusion of very viscous magma that forms lava domes. Dome-building eruptions are commonly associated with hazardous phenomena, including pyroclastic flows generated by dome collapses, explosive eruptions, and volcanic blasts. These eruptions commonly display fairly regular alternations between periods of high and low or no activity with time scales from hours to years. Usually hazardous phenomena are associated with periods of high magma discharge rate. Hence, understanding the causes of pulse activity during extrusive eruptions is an important step towards forecasting volcanic behavior, and especially the transition to explosive activity when magma discharge rate increases by a few orders of magnitude. In recent years the risks have escalated because the population density in the vicinity of many active volcanoes has increased.

Modeling Volcanic Eruptions and Hazards Warning

While a wide range of complex deterministic models exists to model various volcanic processes, these provide little in the way of information about future activity. Being the (partially) observed realization of a complex system, volcanological data are inherently stochastic in nature, and need to be modeled using statistical models. In ► Volcanic Eruptions: Stochastic Models of Occurrence Patterns, M.S. Bebbington considers models of eruption occurrence, omitting techniques for forecasting the nature and effect of the eruption. As the track record of a potentially active volcano provides the best method of assessing its future long-term hazards, the author first briefly reviews the provenance and characteristics of the data available, and then discusses various taxonomies for stochastic models. The examples of Mount Etna and Yucca Mountain are selected for more detailed examination partly because many, somewhat contradictory, results exist. Different models make different assumptions, and vary in how much information they can extract from data. In addition, the data used often varies from study to study, and the sensitivity of models to data is important, but too often ignored.

In ► Volcanic Hazards and Early Warning, R.I. Tilling highlights the range in possible outcomes of volcano unrest and reviews some recent examples of the actual outcomes documented for several well-monitored volcanoes. The author also discusses the challenge for emergency-

management authorities, as well as challenges in achieving refined predictive capability. To respond effectively to a developing volcanic crisis, timely and reliable early warnings are absolutely essential; they can be achieved only by a greatly improved capability for eruption prediction. This in turn depends on the quantity and quality of volcano-monitoring data and the diagnostic interpretation of such information.

Discussions

The terms 'Prediction' and 'forecasting' are often used interchangeably. However, the commonly accepted definition of an 'earthquake prediction' is a concise statement, in advance of the event, of the time, location, and magnitude of a future earthquake. To be practically useful to the society, the time window must be short (in days or months), the location extent small (within tens of kilometers), and the magnitude precise ($\pm$ 0.5 unit). A 'forecast', on the other hand, is more loosely defined as the probability of an occurrence of a large earthquake in a given region (e. g., southern California) during the coming decades or centuries. The broader time intervals associated with forecasts allows society to consider and implement mitigation efforts over a large region.

As an observational seismologist and on a personal note, I am skeptical that earthquakes can be reliably predicted before (1) we have collected accurate data of their occurrences over a sufficiently long period of time, and (2) we have a good understanding of the physical processes that create them. Although earthquakes have been known from antiquity, accurate earthquake catalogs exist only since about the 1960s. Since the recurrence of a damaging earthquake in a given area is often more than 100 years (some even thousands of years), it is obvious that we lack the necessary observed data. C. Lanczos (*Linear Differential Operators*, Van Nostrand-Reinhold, 1961) said it well in general: "a lack of information cannot be remedied by any mathematical trickery". Nevertheless, we must apply new concepts and tools to extract as much useful information as possible from the existing data. Indeed, we must thank many pioneers for enlightening us with many interesting and tentative results about earthquakes, tsunamis and volcanoes that they managed to extract from inadequate and insufficient data.

Because most tsunamis are generated by earthquakes, successes in predicting tsunamis depend on predicting earthquakes and recognizing them as tsunamigenic. Predicting a volcanic eruption is a little easier, as the location is known and there are often some observable phenomena preceding it. However, the exact time and the intensity and extent of an eruption are difficult to predict because the volcanic processes are very complex involving gas, liquid and solid phases.

Fortunately, earthquake, tsunami and volcano hazards can be reduced by employing sound engineering practices, early warning systems, and hazard analysis, using many of the tools and concepts that were developed for prediction. Since fatalities and economic loss from a single catastrophic event can reach 100,000 or more, and $100 billion or more, respectively, it is imperative that governments should support long-term monitoring with modern instruments and research, including complexity studies of earthquakes, tsunamis and volcanoes.

Complexity and Non-linearity in Autonomous Robotics, Introduction to

Warren E. Dixon
Department of Mechanical and Aerospace Engineering, Gainesville, USA

Historically, robotic systems have played a key role in manufacturing, hazardous material handling, exploration and surveillance, search and rescue, and military applications. For most of these applications, single robot systems with traditional sensing and actuation are able to execute predetermined tasks in a well defined environment. The ability to develop a robotic system capable of executing tasks in an environment that is not well defined or is dynamic has been a daunting task. The development of such a capability has been limited by the complexity required in sensing, cognitive decision making, new actuation capabilities, and robustness and responsiveness of the control system. The complexity of these challenges increases when groups of robots are required to interact with each other or with people in a social or physical manner. This volume of work is a collection of recent and emerging efforts at the dawn of a new millennium to develop enabling technologies for new frontiers where robots are able to execute tasks in dynamic and unstructured environments.

Sensing is a fundamental requirement for an autonomous system to efficiently interact with its environment. Vision is arguably the primary environmental sensor used by human beings and many other animals to understand how to interact with dynamic and unstructured environments. While traditional sensor modalities will continue to be an integral part of emerging robotic systems, new research efforts will be driven by the desire

for human-like interpretation and reaction to image information. However, the use of image-based feedback adds complexity and new challenges. One challenge is how an autonomous system can find and track objects of interest in an image of dense, cluttered environments (see ► Motion Prediction for Continued Autonomy). Once a target(s) of interest is determined, how can a robot use a set of two dimensional images to interpret its relative position and orientation with respect to different objects (see ► Image Based State Estimation). Another challenge is how to use reconstructed and estimated information from an image to develop a stable closed-loop error system (see ► Adaptive Visual Servo Control).

Given feedback from vision sensors, traditional sensors (e. g., sonar, radar, lidar, encoders), or a fusion of these sensors, another complexity is how to best exploit sensor and other information discovered during the execution of a plan, including autonomous selection of which behavior(s) to invoke, in what sequence, and by what method. The ability to organize and apply situational knowledge to macro-level planning and decision-making can be enabled through a scaffolding provided by a computer architecture that provides interfaces for different sources of information (see ► Software Architectures for Autonomy). As robotic systems operate in increasingly more complicated environments, more intelligence is required by the autonomous system. The ultimate goal for an autonomous system is to enable a level of cognition with human-like intelligence in perception, motor control, and high-level cognition. Beyond sensing and reaction, cognition enables a robotic system to reason about an environment without direct interaction, enabling the capability to plan for selecting between competing goals, coping with multiple sensory inputs, and completing multiple tasks (see ► Cognitive Robotics). Computer architectures that provide a backbone for organizing and interfacing with different sensing and actuation components and the ability for autonomous systems to exhibit increasing levels of intelligence are especially important as robotics move from behind chained off work cells in an assembly line to direct emotional and physical interaction with people (see ► Human Robot Interaction).

In addition to new frontiers in sensing, computer architectures, and autonomous reasoning, advances are also required in the construction and actuation of an autonomous robotic system. For example, since people spend a significant portion of their time inside buildings, the building itself could be designed as a morphing robot surface that is articulated, programmable, and embedded with integrated digital technologies (see ► Continuum Robots). Or, as robots are forecasted to take an increasing role for in-home care, robots will need new methods of physically interacting with people and the environment. For example, many biological systems have shown that significant articulation can be achieved through tongues, trunks, and tentacles. Robots inspired by these systems could also morph their shape to their environment to access difficult-to-reach areas, and to perform adaptive grasping using whole arm manipulation (see also ► Continuum Robots).

New actuation and manipulation methods place new challenges on control systems. One challenge includes the lack of precise mathematical models of the environment and its interaction with the robot through actuators and sensors. As autonomous robots interact with people, the robot may be required to interpret events that a person describes using linguistic terms. These challenges motivate the need for new control designs that may be inspired by the way that humans tend to work with vague or imprecise concepts (see ► Neuro-fuzzy Control of Autonomous Robotics).

Given the complexity that arises from developing a single autonomous system capable of interacting with uncertain and dynamic environments, motivation exists to develop technologies for groups of robotic systems. Distributed robotics hold the promise to enable groups of robots to collaborate to solve complex tasks such as monitoring a vast environment, manipulating large objects, or building advanced structures. One challenge to enable groups of robots to collaborate efficiently is how to develop path planning and motion coordination methods so that teams of autonomous mobile robots can share the same workspace while avoiding interference with each other, and/or while achieving group motion objectives (see ► Multiple Mobile Robot Teams, Path Planning and Motion Coordination in). Obtaining an optimal solution for the coordination of multiple robot vehicles is a challenge because an integrated approach for designing communication, sensing, and control systems must all be considered as constraints on the system performance (see ► Distributed Controls of Multiple Robotic Systems, An Optimization Approach). The design of a single robot requires the balancing of many factors (e. g., cost, sensing, processing capabilities), many of which are conflicting. In addition to these factors, distributed robotics adds the complexity of multiple robots. For example, there are different advantages and disadvantages to designing a homogeneous versus a heterogeneous team (see ► Distributed Robotic Teams: A Framework for Simulated and Real–World Modeling).

Despite different design tradeoffs for teams of robots, some advanced teams hold the potential for unprece-

dented versatility and robustness at a low cost. For example, modular self-reconfigurable (MSR) robot teams are composed of a large number of repeated modules that can rearrange their connectedness to form a large variety of structures to suit the task (see ► Modular Self-Reconfigurable Robots). In contrast to self-reconfigurable robotics, self-replication utilizes an original unit to actively assemble an exact copy of itself from passive components. This has the potential to result in an exponential growth in the number of robots available to perform a job, thus drastically shortening the original unit's task time (see ► Self-replicating Robotic Systems). To enable self replication, a robot must exploit resources in the environment such as raw materials and energy. Insights to address this challenge can be provided from nature, where biological systems must forage for resources. Foraging robots are mobile robots capable of searching for and, when found, transporting objects to one or more collection points. Foraging robots may be single robots operating individually, or multiple robots operating collectively (see ► Foraging Robots).

New technologies are emerging in sensing, actuation, reasoning, and control. As these technologies mature, new levels of autonomy are made possible. As robotic systems are able to operate in increasingly more dynamic and uncertain environments, an exciting new era for robotics is evident.

Complexity in Systems Level Biology and Genetics: Statistical Perspectives

David A. Stephens
Department of Mathematics and Statistics,
McGill University, Montreal, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
Mathematical Representations of the Organizational Hierarchy
Transcriptomics and Functional Genomics
Metabolomics
Future Directions
Bibliography

Glossary

Systems biology The holistic study of biological structure, function and organization.

Probabilistic graphical model A probabilistic model defining the relationships between variables in a model by means of a graph, used to represent the relationships in a biological network or pathway.

MCMC Markov chain Monte Carlo – a computational method for approximating high-dimensional integrals using Markov chains to sample from probability distributions, commonly used in Bayesian inference.

Microarray A high-throughput experimental platform for collecting functional gene expression and other genomic data.

Cluster analysis A statistical method for discovering subgroups in data.

Metabolomics The study of the metabolic content of tissues.

Definition of the Subject

This chapter identifies the challenges posed to biologists, geneticists and other scientists by advances in technology that have made the observation and study of biological systems increasingly possible. High-throughput platforms have made routine the collection vast amounts of structural and functional data, and have provided insights into the working cell, and helped to explain the role of genetics in common diseases. Associated with the improvements in technology is the need for statistical procedures that extract the biological information from the available data in a coherent fashion, and perhaps more importantly, can quantify the certainty with which conclusions can be made. This chapter outlines a biological hierarchy of structures, functions and interactions that can now be observed, and detail the statistical procedures that are necessary for analyzing the resulting data. The chapter has four main sections. The first section details the historical connection between statistics and the analysis of biological and genetic data, and summarizes fundamental concepts in biology and genetics. The second section outlines specific mathematical and statistical methods that are useful in the modeling of data arising in bioinformatics. In sections three and four, two particular issues are discussed in detail: functional genomic via microarray analysis, and metabolomics. Section five identifies some future directions for biological research in which statisticians will play a vital role.

Introduction

The observation of biological systems, their processes and inter-reactions, is one of the most important activities in modern science. It has the capacity to provide direct insight into fundamental aspects of biology, genetics, evo-

lution, and indirectly will inform many aspects of public health. Recent advances in technology – high-throughput measurement platforms, imaging – have brought a new era of increasingly precise methods of investigation. In parallel to this, there is an increasingly important focus on statistical methods that allow the information gathered to be processed and synthesized. This chapter outlines key statistical techniques that allow the information gathered to be used in an optimal fashion.

Although its origin is dated rather earlier, the term *Systems Biology* (see, for example, [1,2,3]) has, since 2000, been used to describe the study of the operation of biological systems, using tools from mathematics, statistics and computer science, supplanting *computational biology* and *bioinformatics* as an all-encompassing term for quantitative investigation in molecular biology. Most biological systems are hugely complex, involving chemical and mechanical processes operating at different scales. It is important therefore that information gathered is processed coherently, according to self-consistent rules and practices, in the presence of the uncertainty induced by imperfect observation of the underlying system. The most natural framework for coherent processing of information is that of probabilistic modeling.

Statistical Versus Mathematical Modeling

There is a great tradition of mathematical and probabilistic modeling of biology and genetics; see [4] for a thorough review. The mathematization of biology, evolution and heredity began at the end of the nineteenth century, and continued for the first half of the twentieth century, by far pre-dating the era of molecular biology and genetics that culminated at the turn of the last millennium with the human genome project. Consequently, the mathematical models of, say, evolutionary processes that were developed by Yule [5] and Fisher and Wright [6,7,8], and classical models of heredity, could only be experimentally verified and developed many years after their conception. It could also be convincingly argued that through the work of F. Galton, K. S. Pearson and R. A. Fisher, modern statistics has its foundation in biology and genetics.

In parallel to statistical and *stochastic* formulation of models for biological systems, there has been a more recent focus on the construction of *deterministic* models to describe observed biological phenomena. Such models fall under the broad description *Mathematical Biology*, and have their roots in applied mathematics and dynamical systems; see, for example, [8,9] for a comprehensive treatment. The distinction between stochastic and deterministic models is important to make, as the objectives and tools used often differ considerably. This chapter will restrict attention to stochastic models, and the processing of observed data, and thus is perhaps more closely tied to the immediate interests of the scientist, although some of the models utilized will be inspired by mathematical models of the phenomena being observed.

Fundamental Concepts in Biology and Genetics

To facilitate the discussion of statistical methods applied to systems biology, it is necessary to introduce fundamental concepts from molecular biology and genetics; see the classic text [10] for full details. Attention is restricted to eukaryotic, organisms whose cells are constructed to contain a nucleus which coding information is encapsulated.

- The cell nucleus is a complex architecture containing several nuclear domains [11] whose organization is not completely understood, but the fundamental activity that occurs within the nucleus is the production and distribution of proteins.
- Dioxyribonucleic acid (DNA) is a long string of nucleotides that encodes biological information, and that is copied or *transcribed* into ribonucleic acid (RNA), which in turn enables the formation of proteins. Specific segments of the DNA, genes, encode the proteins, although non-coding regions of DNA – for example, promoter regions, transcription factor binding sites – also have important roles. Genetic variation at the nucleotide level, even involving a single nucleotide, can disrupt cellular activity. In humans and most other complex organisms, DNA is arranged into chromosomes, which are duplicated in the process of mitosis. The entire content of the DNA of an organism is termed the *genome*.
- Proteins are macromolecules formed by the *translation* of RNA, comprising amino acids arranged (in primary structure) in a linear fashion, comprising domains with different roles, and physically configured in three dimensional space. Proteins are responsible for all biological activities that take place in the cell, although proteins may have different roles in different tissues at different times, due to the *regulation* of transcription.
- Proteins interact with each other in different ways in different contexts in interaction networks that may be dynamically organized. Genes are also regarded as having indirect interactions through gene regulatory networks.
- Genetic variation amongst individuals in a population is due to mutation and selection, which can be regarded as stochastic mechanisms. Genetic information in the form of DNA passes from parent to offspring, which

promulgates genetic variation. Individuals in a population are typically related in evolutionary history. Similarly, proteins can also thought to be related through evolutionary history.
- Genetic disorders are the result of genetic variation, but the nature of the genetic variation can be large- or small-scale; at the smallest scale, variation in single nucleotides (*single nucleotide polymorphisms* (SNPs)) can contribute to the variation in observed traits.

Broadly, attention is focused on the study of *structure* and *function* of DNA, genes and proteins, and the nature of their *interactions*. It is useful, if simplistic, to view biological activities in terms of an organizational hierarchy of inter-related chemical reactions at the level of DNA, protein, nucleus, network and cellular levels. A holistic view of mathematical modeling and statistical inference requires the experimenter to model simultaneously actions and interactions of all the component features, whilst recognizing that the component features cannot observed directly, and can only be studied through separate experiments on often widely different platforms. It is the role of the bioinformatician or systems biologist to synthesize the data available from separate experiments in an optimal fashion.

Mathematical Representations of the Organizational Hierarchy

A mathematical representation of a biological system is required that recognizes, first, the complexity of the system, secondly, its potentially temporally changing nature, and thirdly the inherent uncertainties that are present. It is the last feature that necessitates the use of probabilistic or stochastic modeling.

An aphorism commonly ascribed to D.V. Lindley states that "*Probability is the language of uncertainty*"; probability provides a coherent framework for processing information in the presence of imperfect knowledge, and through the paradigm of Bayesian theory [12] provides the mathematical template for statistical inference and prediction. In the modeling of complex systems, three sorts of uncertainty are typically present:

- *Uncertainty of Structure:* Imperfect knowledge of the connections between the interacting components is typically present. For example, in a gene regulatory network, it may be possible via the measurement of gene co-expression to establish which genes interact within the network, but it may not be apparent precisely how the organization of regulation operates, that is which genes regulate the expression of other genes.
- *Uncertainty concerning Model Components:* In any mathematical or probabilistic model of a biological system, there are model components (differential equations, probability distributions, parameter settings) that must be chosen to facilitate implementation of the model. These components reflect, but are not determined by, structural considerations.
- *Uncertainty of Observation:* Any experimental procedure carries with it uncertainty induced by the measurement of underlying system, that is typically subject to random measurement error, or noise. For example, many biological systems rely on imaging technology, and the extraction of the level of signal of a fluorescent probe, for a representation of the amount of biological material present. In microarray studies (see Sect. "Microarrays"), comparative hybridization of messenger RNA (mRNA) to a medium is a technique for measuring gene expression that is noisy due to several factors (imaging noise, variation in hybridization) not attributable to a biological cause.

The framework to be built must handle these types of uncertainty, and permit inference about structure and model components.

Models Derived from Differential Equations

A deterministic model reflecting the dynamic relationships often present in biological systems may be based on the system of ordinary differential equations (ODEs)

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{g}(\boldsymbol{x}(t)) \tag{1}$$

where $\boldsymbol{x}(t) = (x_1(t), \ldots, x_d(t))^{\mathsf{T}}$ represent the levels of d quantities being observed $\dot{\boldsymbol{x}}(t)$ represents time derivative, and $\boldsymbol{g}$ is some potentially non-linear system of equations, that may be suggested by biological prior knowledge or prior experimentation. The model in Eq. (1) is a classical "Mathematical Biology" model, that has been successful in representing forms of organization in many biological systems (see, for example [13] for general applications). Suppressed in the notation is a dependence on system parameters, $\boldsymbol{\theta}$, a k-dimensional vector that may be presumed fixed, and "tuned" to replicate observed behavior, or estimated from observed data. When data representing a partial observation of the system are available, inferences about $\boldsymbol{\theta}$ can be made, and models defined by ODE systems are of growing interest to statisticians; see, for example, [15,16,17].

Equation (1) can be readily extended to a *stochastic differential equation* (SDE) system

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{g}(\boldsymbol{x}(t)) + \boldsymbol{dz}(t) \tag{2}$$

where $\boldsymbol{z}(t)$ is some stochastic process that renders the solution to Eq. (2) a stochastic process (see, for example, [18] for a comprehensive recent summary of modeling approaches and inference procedures, and a specific application in [19]). The final term $\boldsymbol{dz}(t)$ represents the infinitesimal stochastic increment in $\boldsymbol{z}(t)$. Such models, although particularly useful for modeling activity at the molecular level, often rely on simplifying assumptions (linearity of $\boldsymbol{g}$, Gaussianity of $\boldsymbol{z}$) and the fact that the relationship structure captured by $\boldsymbol{g}$ is known. Inference for the parameters of the system can be made, but in general require advanced computational methods (Monte Carlo (MC) and Markov chain Monte Carlo (MCMC)).

Probabilistic Graphical Models

A simple and often directly implementable approach is based on a *probabilistic graphical model*, comprising a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, described by a series of nodes $\mathcal{N}$, edges $\mathcal{E}$, and a collection of random variables $\boldsymbol{X} = (X_1, \ldots, X_d)^{\mathsf{T}}$ placed at the nodes, all of which may be dynamically changing. See, for example [20] for a recent summary, [14] for mathematical details and [22] for a biological application.

The objective of constructing such a model is to identify the joint probability structure of $\boldsymbol{X}$ given the graph $\mathcal{G}$, which possibly is parametrized by parameters $\boldsymbol{\phi}$, $f_X(\boldsymbol{x}|\boldsymbol{\phi}, \mathcal{G})$. In many applications, $\boldsymbol{X}$ is not directly observed, but is instead inferred from observed data, $\boldsymbol{Y}$, arising as noisy observations derived from $\boldsymbol{X}$. Again, a k-dimensional parameter vector $\boldsymbol{\theta}$ helps to characterize the stochastic dependence of $\boldsymbol{Y}$ on $\boldsymbol{X}$ by parametrize the conditional probability density $f_{Y|X}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$. The joint probability model encapsulating the probabilistic structure of the model is

$$f_{X,Y}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}) = f_X(\boldsymbol{x}|\boldsymbol{\phi}, \mathcal{G}) f_{Y|X}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) \tag{3}$$

The objectives of inference are to learn about $\mathcal{G}$ (the uncertain structural component) and parameters $(\boldsymbol{\theta}, \boldsymbol{\phi})$ (the uncertain model parameters and observation components).

The graph structure $\mathcal{G}$ is described by $\mathcal{N}$ and $\mathcal{E}$. In holistic models, $\mathcal{G}$ represents the interconnections between interacting modules (genomic modules, transcription modules, regulatory modules, proteomic modules, metabolic modules etc.) and also the interconnections within modules in the form of sub graphs. The nodes $\mathcal{N}$ (and hence $\boldsymbol{X}$) represent influential variables in the model structure, and the edges $\mathcal{E}$ represent dependencies. The edge connecting two nodes, if present, may be *directed* or *undirected* according to the nature of the influence; a directed edge indicates the direction of *causation*, an undirected edge indicates a *dependence*.

Causality is a concept distinct from dependence (association, co-variation or correlation), and represents the influence of one node on one or more other nodes (see, for example, [23] for a recent discussion of the distinction with examples, and [24,25] for early influential papers discussing how functional dependence may be learned from real data). A simple causal relationship between three variables X_1, X_2, X_3 can be represented

which encodes a *conditional independence relationship* between X_1 and X_3 given X_2, and a factorization of the joint distribution

$$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2).$$

Similarly, the equivalent graphs

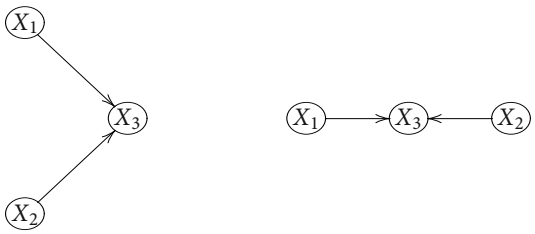

encode $\quad p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3|x_1, x_2)$

whereas the graphs for conditional independence of X_2 and X_3 given X_1 are

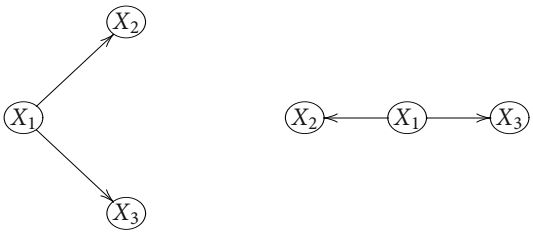

encoding $\quad p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_1). \quad (4)$

Such simple model assumptions are the building blocks for the construction of highly complex graphical representations of biological systems. There is an important difference between analysis based purely on simultaneous observation of all components of the system, which can typically on yield inference on *dependencies* (say, covariances measured in the joint probability model $p(\boldsymbol{x})$ – see, for

example [26,27,28] and analysis based on *interventions*-genomic knock–out experiments, chemical or biological challenges, transcriptional/translational perturbation such as RNA interference (RNAi) – that may yield information on casual links; see, for example [29,30].

Bayesian Statistical Inference

Given a statistical model for observed data such as Eq. (3), inference for the parameters $(\boldsymbol{\theta}, \boldsymbol{\phi})$ and the graph structure $\mathcal{G}$ is required. The optimal coherent framework is that of *Bayesian statistical inference* (see for example [31]), that requires computation of the *posterior distribution* for the unknown (or *unobservable*) quantities given by

$$\pi(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}|\boldsymbol{x}, \boldsymbol{y}) \propto f_{X,Y}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}) p(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}) = L(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}|\boldsymbol{x}, \boldsymbol{y}) p(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}) \quad (5)$$

a probability distribution from which can be computed parameter estimates with associated uncertainties, and predictions from the model. The terms $L(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}|\boldsymbol{x}, \boldsymbol{y})$ and $p(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G})$ are termed *likelihood* and *prior probability distribution* respectively. The likelihood reflects the observed data, and the prior distribution encapsulates biological prior knowledge about the system under study. If the graph structure is known in advance, the prior distribution for that component can be set to be degenerate. If, as in many cases of probabilistic graphical models, the $\boldsymbol{x}$ are unobserved, then the posterior distribution incorporates them also,

$$\pi(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}, \boldsymbol{x}|, \boldsymbol{y}) \propto f_{Y|X}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) f_X(\boldsymbol{x}|\boldsymbol{\phi}, \mathcal{G}) p(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}) \quad (6)$$

yielding a *latent* or *state-space* model, otherwise interpreted as a *missing data* model.

The likelihood and prior can often be formulated in a hierarchical fashion to reflect believed causal or conditional independence structures. If a graph $\mathcal{G}$ is separable into two sub graphs $\mathcal{G}_1, \mathcal{G}_2$ conditional on a connecting node η, similar to the graph in Eq. (4), then the probability model also factorizes into a similar fashion; for example, X_1 might represent the amount of expressed mRNA of a gene that regulates two separate functional modules, and X_2 and X_3 might be the levels of expression of collections of related proteins. The hierarchical specification also extends to parameters in probability models; a standard formulation of a Bayesian hierarchical model involves specification of conditional independence structures at multiple levels of within a graph. The following three-level hierarchical model relates data $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_p)$ at level 1, to a population of parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)^{\mathsf{T}}$ at level 2, to hyper parameters $\boldsymbol{\psi}$ at level 3

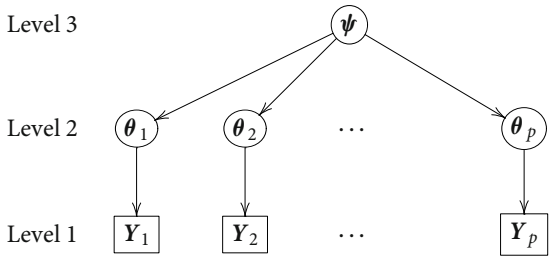

(7)

yielding the factorization of the Bayesian full joint distribution as

$$f_{X,Y,\boldsymbol{\psi},\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\psi}) = p(\boldsymbol{\psi}) \left\{ \prod_{i=1}^{p} p(\boldsymbol{\theta}_i|\boldsymbol{\psi}) \right\} \left\{ \prod_{i=1}^{p} p(\boldsymbol{Y}_i|\boldsymbol{\phi}_i) \right\} .$$

Bayesian Computation

The posterior distribution is, potentially, a high-dimensional multivariate function on a complicated parameter space. The proportionality constant in Eq. (5) takes the form

$$f_{X,Y}(\boldsymbol{x}, \boldsymbol{y}) = \int f_{X,Y}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}) p(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}) \, \mathrm{d}\boldsymbol{\theta} \, \mathrm{d}\boldsymbol{\phi} \, \mathrm{d}\mathcal{G} \quad (8)$$

and in Eq. (6) takes the form

$$f_Y(\boldsymbol{y}) = \int f_{X,Y}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}) p(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{G}) \, \mathrm{d}\boldsymbol{\theta} \, \mathrm{d}\boldsymbol{\phi} \, \mathrm{d}\mathcal{G} \, \mathrm{d}\boldsymbol{x} \quad (9)$$

and is termed the *marginal likelihood* or *prior predictive distribution* for the *observable* quantities $\boldsymbol{x}$ and $\boldsymbol{y}$. In formal Bayesian theory, it is the representation of the distribution of the observable quantities through the paradigm of *exchangeability* that justifies the decomposition in Eq. (8) into likelihood and prior, and justifies, via asymptotic arguments, the use of the posterior distribution for inference (see Chaps. 1–4 in [13] for full details). It is evident from these equations that exact computation of the posterior distribution necessitates high-dimensional integration, and in many cases this cannot be carried out analytically.

Numerical Integration Approaches Classical numerical integration methods, or analytic approximation methods are suitable only in low dimensions. Stochastic numerical integration, for example Monte Carlo integration, approximates expectations by using empirical averages of

functional of samples obtained from the target distribution; for probability distribution $\pi(\boldsymbol{x})$, the approximation of $\mathrm{E}_\pi[g(\boldsymbol{X})]$,

$$\mathrm{E}_\pi[g(\boldsymbol{X})] = \int g(\boldsymbol{x})\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} < \infty$$

is achieved by randomly sampling $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ (N large) from $\pi()$, and using the estimate

$$\widehat{\mathrm{E}}_\pi[g(\boldsymbol{X})] = \frac{1}{N}\sum_{i=1}^{n} g(\boldsymbol{x}_i).$$

An adaptation of the Monte Carlo method can be used if the functions g and π are not "similar" (in the sense that g is large in magnitude where π is not, and *vice versa*); *importance sampling* uses the representation

$$\mathrm{E}_\pi[g(\boldsymbol{X})] = \int g(\boldsymbol{x})\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int \frac{g(\boldsymbol{x})\pi(\boldsymbol{x})}{p(\boldsymbol{x})} p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

for some pdf $p()$ having common support with π, and constructs an estimate from a sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ from $p()$ of the form

$$\widehat{\mathrm{E}}_\pi[g(\boldsymbol{X})] = \frac{1}{N}\sum_{i=1}^{n} \frac{g(\boldsymbol{x}_i)\pi(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i)}.$$

Under standard regularity conditions, the corresponding estimators converge to the required expectation. Further extensions are also useful:

- *Sequential Monte Carlo:* Sequential Monte Carlo (SMC) is an adaptive procedure that constructs a sequence of improving importance sampling distributions. SMC is a technique that is especially useful for inference problems where data are collected sequentially in time, but is also used in standard Monte Carlo problems (see [32]).
- *Quasi Monte Carlo:* Quasi Monte Carlo (QMC) utilizes uniform **but not random** samples to approximate the required expectations. It can be shown that QMC can produce estimators with lower variance than standard Monte Carlo.

Markov Chain Monte Carlo Markov chain Monte Carlo (MCMC) is a stochastic Monte Carlo method for sampling from a high-dimensional probability distribution $\pi(\boldsymbol{x})$, and using the samples to approximate expectations with respect to that distribution. An ergodic, discrete-time Markov chain is defined on the support of π in such a way that the stationary distribution of the chain exists, and is equal to π. Dependent samples from π are obtained by collecting realized values of the chain after it has reached its stationary phase, and then used as the basis of a Monte Carlo strategy.

The most common MCMC algorithm is known as the *Metropolis–Hastings* algorithm which proceeds as follows. If the state of the d-dimensional chain $\{\boldsymbol{X}_t\}$ at iteration t is given by $\boldsymbol{X}_t = \boldsymbol{u}$, then a candidate state $\boldsymbol{v}$ is generated from conditional density $q(\boldsymbol{u}, \boldsymbol{v}) = q(\boldsymbol{v}|\boldsymbol{u})$, and accepted as the new state of the chain (that is, $\boldsymbol{X}_{t+1} \stackrel{\text{def}}{=} \boldsymbol{v}$) with probability $\alpha(\boldsymbol{u}, \boldsymbol{v})$ given by

$$\alpha(\boldsymbol{u}, \boldsymbol{v}) = \min\left\{1, \frac{\pi(\boldsymbol{v})q(\boldsymbol{v}, \boldsymbol{u})}{\pi(\boldsymbol{u})q(\boldsymbol{u}, \boldsymbol{v})}\right\}.$$

A common MCMC approach involves using a *Gibbs sampler strategy* that performs iterative sampling with updating from the collection of *full conditional* distributions

$$\begin{aligned}\pi(\boldsymbol{x}_j|\boldsymbol{x}_{(j)}) &= \pi(\boldsymbol{x}_j|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, \boldsymbol{x}_d) \\ &= \frac{\pi(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d)}{\pi(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, \boldsymbol{x}_d)}, \quad j = 1, \ldots, d\end{aligned}$$

rather than updating the components of $\boldsymbol{x}$ simultaneously. There is a vast literature on MCMC theory and applications; see [33,34] for comprehensive treatments.

MCMC re-focuses inferential interest from computing posterior analytic functional forms to producing posterior samples. It is an extremely flexible framework for computational inference that carries with it certain well-documented problems, most important amongst them being the assessment of *convergence*. It is not always straightforward to assess when the Markov chain has reached its stationary phase, so certain monitoring steps are usually carried out.

Bayesian Modeling: Examples

Three models that are especially useful in the modeling of systems biological data are *regression models*, *mixture models, and state-space models*. Brief details of each type of model follow.

Regression Models Linear regression models relate an observed response variable Y to a collection of predictor variables $X_1, X_2, \ldots, X_d$ via the model for the ith response

$$Y_i = \beta_0 + \sum_{j=1}^{d} \beta_j X_{ij} + \epsilon_i = \boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\beta} + \epsilon_i$$

say, or in vector form, for $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathsf{T}}$,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)^{\mathsf{T}}$ is a vector of real-valued parameters, and ϵ is a vector random variable with zero-mean and variance-covariance matrix Σ. The objective in the analysis is to make inference about $\boldsymbol{\beta}$, to understand the influence of the predictors on the response, and to perform prediction for Y. The linear regression model (or General Linear Model) is extremely flexible: the design matrix $\boldsymbol{X}$ can be formed from arbitrary, possibly non-linear *basis* functions of the predictor variables. By introducing a covariance structure into Σ, it is possible to allow for dependence amongst the components of $\boldsymbol{Y}$, and allows for the possibility of modeling repeated measures, longitudinal or time-series data that might arise from multiple observation of the same experimental units.

An extension that is often also useful is to *random effect* or *mixed* models that take into account any repeated measures aspect to the recorded data. If data on an individual (person, sample, gene etc) is $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{id})^{\mathsf{T}}$, then

$$\boldsymbol{Y}_i = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{U}_i + \epsilon_i \tag{10}$$

where $\boldsymbol{Z}$ is a $d \times p$ constant design matrix, and $\boldsymbol{U}_i$ is a $p \times 1$ vector of *random effects* specific to individual i. Typically the random effect vectors are assumed to be drawn from a common population. Similar formulations can be used to construct semi-parametric models that are useful for flexible modeling in regression.

Mixture Models A mixture model presumes that the probability distribution of variable $\boldsymbol{Y}$ can be written

$$f_{Y|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \omega_k f_k(\boldsymbol{y}|\theta_k) \tag{11}$$

where $f_1, f_2, \ldots, f_K$ are distinct component densities indexed by parameters $\theta_1, \ldots, \theta_K$, and for all k, $0 < \omega_k < 1$, with

$$\sum_{k=1}^{K} \omega_k = 1.$$

The model can be interpreted as one that specifies that with probability ω_k, $\boldsymbol{Y}$ is drawn from density f_k, for $k = 1, \ldots, K$. Hence the model is suitable for modeling in cluster analysis problems.

This model can be extended to an *infinite mixture model*, which has close links with *Bayesian non-parametric* modeling. A simple infinite mixture/Bayesian non-parametric model is the *mixture of Dirichlet processes* (MDP) model [35,36]: for parameter $\alpha > 0$ and distribution function F_0, an MDP model can be specified using the following hierarchical specification: for a sample of size n, we have

$$\begin{aligned} Y_i|\theta_i &\sim f_{Y|\theta}(y|\theta_i) \quad i = 1, \ldots, n \\ \theta_1, \ldots, \theta_n &\sim DP(\alpha, F_0) \end{aligned}$$

where $DP(\alpha, F_0)$ denotes a *Dirichlet process*. The $DP(\alpha, F_0)$ model may be sampled to produce $\theta_1, \theta_2, \ldots, \theta_n$ using the *Polya-Urn* scheme

$$\begin{aligned} \theta_1 &\sim F_0 \\ \theta_k|\theta_1, \ldots, \theta_{k-1} &\sim \frac{\alpha}{\alpha + k - 1} F_0 + \frac{1}{\alpha + k - 1} \sum_{j=1}^{k-1} \delta_{\theta_j} \end{aligned}$$

where δ_x is a point mass at x. For θ_k, conditional on $\theta_1, \ldots, \theta_{k-1}$, the Polya-Urn scheme either samples θ_k from F_0 (with probability $\alpha/(\alpha + k - 1)$), or samples $\theta_k = \theta_j$ for some $j = 1, \ldots, k-1$ (with probability $1/(\alpha + k - 1)$). This model therefore induces *clustering* amongst the θ values, and hence has a structure similar to the finite mixture model – the distinct values of $\theta_1, \ldots, \theta_n$ are identified as the cluster "centers" that index the component densities in the mixture model in Eq. (11). The degree of clustering is determined by α; high values of α encourage large numbers of clusters.

The MDP model is a flexible model for statistical inference, and is used in a wide range of applications such as density estimation, cluster analysis, functional data analysis and survival analysis. The component densities can be univariate or multivariate, and the model itself can be used to represent the variability in observed data or as a prior density. Inference for such models is typically carried out using MCMC or SMC methods [32,33]. For applications in bioinformatics and functional genomic, see [37,38].

State-Space Models A state-space model is specified through a pair of equations that relate a collection of *states*, $\boldsymbol{X}_t$, to *observations* $\boldsymbol{Y}_t$ that represent a system and how that system develops over time. For example, the relationship could be model led as

$$\begin{aligned} \boldsymbol{Y}_t &= f(\boldsymbol{X}_t, \boldsymbol{U}_t) \\ \boldsymbol{X}_{t+1} &= g(\boldsymbol{X}_t, \boldsymbol{V}_t) \end{aligned}$$

where f and g are vector-valued functions, and $(\boldsymbol{U}_t, \boldsymbol{V}_t)$ are random error terms. A *linear state-space model* takes the form

$$\begin{aligned} \boldsymbol{Y}_t &= A_t \boldsymbol{X}_t + \boldsymbol{c}_t + \boldsymbol{U}_t \\ \boldsymbol{X}_{t+1} &= B_t \boldsymbol{X}_t + \boldsymbol{d}_t + \boldsymbol{V}_t \end{aligned}$$

for deterministic matrices A_t and B_t and vectors $\boldsymbol{c}_t$ and $\boldsymbol{d}_t$. The $\boldsymbol{X}_t$ represent the values of unobserved states, and the

second equation represents the evolution of these states through time (see [39]).

State-space models can be used as models for scalar, vector and matrix-valued quantities. One application is evolution of a covariance structure, for example, representing dependencies in a biological network. If the network is dynamically changing through time, a model similar to those above is required but where $\boldsymbol{X}_t$ is a square, positive-definite matrix. For such a network, therefore, a probabilistic model for positive-definite matrices can be constructed from the Wishart/Inverse Wishart distributions [40]. For example, we may have for $t = 1, 2, \ldots,$

$$\boldsymbol{Y}_t \sim \text{Normal}(0, \boldsymbol{X}_t)$$

$$\boldsymbol{X}_{t+1} \sim \text{InverseWishart}(\nu_t, \boldsymbol{X}_t)$$

where degrees of freedom parameter ν_t is chosen to induce desirable properties (stationary, constant expectation etc.) in the sequence of $\boldsymbol{X}_t$ matrices.

Transcriptomics and Functional Genomics

A key objective in the study of biological organization is to understand the mechanisms of the transcription of genomic DNA into mRNA that initiates the production of proteins and hence lies at the center of the functioning of the nuclear engine. In a cell in a particular tissue at a particular time, the nucleus contains the entire mRNA profile (*transcriptome*) which, if it could be measured, would provide direct insight into the functioning of the cell. If this profile could be measured in a dynamic fashion, then the patterns of gene regulation for one, several or many genes could be studied. Broadly, if a gene is "active" at any time point, it is producing mRNA transcripts, sometimes at a high rate, sometimes at a lower rate, and understanding the relationships between patterns of up- and down-regulation lies at the heart of uncovering pathways, or networks of interacting genes. *Transcriptomics* is the study of the entirety of recorded transcripts for a given genome in a given condition. *Functional genomics*, broadly, is the study of gene function via measured expression levels and how it relates to genome structure and protein expression.

Microarrays

A common biological problem is to detect *differential expression* levels of a gene in two or more tissue or cell types, as any differences may contribute to the understanding of the cellular organization (pathways, regulatory networks), or may provide a mechanism for discrimination between future unlabeled samples. An important tool for the analysis of these aspects of gene function is the *microarray*, a medium onto which DNA fragments (or *probes*) are placed or etched. Test sample mRNA fragments are tagged with a fluorescent marker, and then allowed to bond or *hybridize* with the matching DNA probes specific to that nucleotide sequence, according to the usual biochemical bonding process. The microarray thus produces a measurement of the mRNA content of the test sample for each of the large number of DNA sequences bound to the microarray as probes. Microarrays typically now contain tens of thousands of probes for simultaneous investigation of gene expression in whole chromosomes, or even whole genomes for simple organisms. The hybridization experiments are carried out under strict protocols, and every effort is made to regularize the production procedures, from the preparation stage through to imaging. Typically, replicate experiments are carried out.

Microarray experiments have made the study of gene expression routine; instantaneous measurements of mRNA levels for large numbers of different genes can be obtained for different tissue or cell types in a matter of hours. The most important aspects of a statistical analysis of gene expression data are, therefore, twofold; the analysis should be readily implementable for large data sets (large numbers of genes, and/or large numbers of samples), and should give representative, robust and reliable results over a wide range of experiments.

Since their initial use as experimental platforms, microarrays have become increasingly sophisticated, allowing measurement of different important functional aspects. Arrays containing whole genomes of organisms can be used for investigation of function, copy-number variation, SNP variation, deletion/insertion sites and other forms of DNA sequence variation (see [41] for a recent summary). High-throughput technologies similar in the form of printed arrays are now at the center of transcriptome investigation in several different organisms, and also widely used for genome-wide investigation of common diseases in humans [42,43]. The statistical analysis of such data represents a major computational challenge. In the list below, a description of details of *first* and *second* generation microarrays is given.

- **First Generation Microarray Studies**
  From the mid 1990s, comparative hybridization experiments using microarrays or gene-chips began to be widely used for the investigation of gene expression. The two principal types of array used were cDNA arrays and oligonucleotide arrays:
  - **cDNA microarrays**: In cDNA microarray competitive hybridization experiments, the mRNA levels of a genes in a target sample are compared to the

mRNA level of a control sample by attaching fluorescent tags (usually red and green respectively for the two samples) and measuring the relative fluorescence in the two channels. Thus, in a test sample (containing equal amounts of target and control material), differential expression **relative** to the control is either in terms of *up-regulation* or *down-regulation* of the genes in the target sample. Any genes that are up-regulated in the target compared to the control and hence that have larger amounts of the relevant mRNA, will fluoresce as predominantly red, and any that are down-regulated will fluoresce green. Absence of differences in regulation will give equal amounts of red and green, giving a yellow fluor. Relative expression is measured on the log scale

$$y = \log \frac{x_{\text{TARGET}}}{x_{\text{CONTROL}}} = \log \frac{x_{\text{R}}}{x_{\text{G}}} \tag{12}$$

where x_{R} and x_{G} are the fluorescence levels in the RED and GREEN channels respectively.

- **Oligonucleotide arrays:** The basic concept oligonucleotide arrays is that the array is produced to interrogate specific target mRNAs or genes by means of a number of oligo probes usually of length no longer than 25 bases; typically 10-15 probes are used to hybridize to a specific mRNA, with each oligo probe designed to target a specific segment of the mRNA sequence. Hybridization occurs between oligos and test DNA in the usual way. The novel aspect of the oligonucleotide array is the means by which the **absolute** level of the target mRNA is determined; each *perfect match* (PM) probe is paired with a *mismatch* (MM) probe that is identical to the prefect match probe **except** for the nucleotide in the center of the probe, for which a mismatch nucleotide is substituted, as indicated in the diagram below.

  PM: ATGTATACTATT[A]TGCCTAGAGTAC

  MM: ATGTATACTATT[C]TGCCTAGAGTAC

  The logic is that the target mRNA, which has been fluorescently tagged, will bind perfectly to the *PM* oligo, and not bind at all to the *MM* oligo, and hence the absolute amount of the target mRNA present can be obtained as the difference $x_{PM} - x_{MM}$ where x_{PM} and x_{MM} are the measurements of for the *PM* and *MM* oligos respectively.

- **Second Generation Microarrays**
  In the current decade, the number of array platforms has increased greatly. The principle of of hybridization of transcripts to probes on a printed array is often still the fundamental biological component, but the design of the new arrays is often radically different. Some of the new types of array are described below (see [44] for a summary).
  - **ChiP-Chip:** ChIP-chip (*chromatin immunoprecipitation* chip) arrays are *tiling* array with genomic probes systematically covering whole genomes or chromosomes that is used to relate protein expression to DNA sequence by mapping the binding sites of transcription factor and other DNA-binding proteins. See [45] for an application and details of statistical issues.
  - **ArrayCGH:** Array comparative genome hybridization (ArrayCGH) is another form of tiling array that is used to detect *copy number variation* (the variation in the numbers of repeated DNA segments) in subgroups of individuals with the aim of detecting important variations related to common diseases. See [46,47].
  - **SAGE:** Serial Analysis of Gene Expression (SAGE) is a platform for monitoring the patterns of expression of many thousands of transcripts in one sample, which relies on the sequencing of short cDNA tags that correspond to a sequence near one end of every transcript in a tissue sample. See [48,49,50].
  - **Single Molecule Arrays:** Single Molecule Arrays rely on the binding of single mRNA transcripts to the spots on the array surface, and thus allows for extremely precise measurement of transcript levels: see [51]. Similar technology is used for precise protein measurement and antibody detection. See [52].

Statistical Analysis of Microarray Data

In a microarray experiment, the experimenter has access to expression/expression profile data, possibly for a number of replicate experiments, for each of a (usually large) number of genes. Conventional statistical analysis techniques and principles (hypothesis testing, significance testing, estimation, simulation methods/Monte Carlo procedures) are used in the analysis of microarray data. The principal biological objectives of a typical microarray analysis are:

- **Detection of differential expression:** up- or down-regulation of genes in particular experimental contexts, or in particular tissue samples, or cell lines at a given time instant.
- **Understanding of temporal aspects of gene regulation:** the representation and modeling of patterns of changes in gene regulation over time.

- **Discovery of gene clusters:** the partitioning of large sets of genes into smaller sets that have common patterns of regulation.
- **Inference for gene networks/biological pathways:** the analysis of co-regulation of genes, and inference about the biological processes involving many genes concurrently.

There are typically several key issues and models that arise in the analysis of microarray data: such methods are described in detail in [53,54,55,56]. For a Bayesian modeling perspective, see [57].

- **Array normalization:** Arrays are often imaged under slightly different experimental conditions, and therefore the data are often very different even from replicate to replicate. This is a systematic experimental effect, and therefore needs to be adjusted for in the analysis of differential expression. A misdiagnosis of differential expression may be made purely due to this systematic experimental effect.
- **Measurement error**: The reported (relative) gene expression levels models are only in fact proxies for the true level gene expression in the sample. This requires a further level of variability to be incorporated into the model.
- **Random effects modeling**: It may be necessary to use *mixed* regression models, where gene specific *random-effects* terms are incorporated into the model.
- **Multivariate analysis**: The covariability of response measurements, in time course experiments, or between PM and MM measurements for an oligonucleotide array experiment, is best handled using multivariate modeling.
- **Testing:** One- and two-sample hypothesis testing techniques, based on parametric and non-parametric testing procedures can be used in the assessment of the presence of differential expression. For detecting more complex (patterns of) differential expression, in more general structured models, the tools of analysis of variance (ANOVA) can be used to identify the chief sources of variability.
- **Multiple testing/False discovery**: In microarray analysis, a classical statistical analysis using significance testing needs to take into account the fact that a very large number of tests are carried out. Hence significance levels of tests must be chosen to maintain a required *family-wise error rate*, and to control the *false discovery rate*.
- **Classification:** The genetic information contained in a gene expression profile derived from microarray experiments for, say, an individual tissue or tumor type may be sufficient to enable the construction of a *classification rule* that will enable subsequent classification of new tissue or tumor samples.
- **Cluster analysis:** Discovery of subsets of sets of genes that have common patterns of regulation can be achieved using the statistical techniques of *cluster analysis* (see Sect. "Clustering").
- **Computer-intensive inference**: For many testing and estimation procedures needed for microarray data analysis, simulation-based methods (bootstrap estimation, Monte Carlo and permutation tests, Monte Carlo and MCMC) are often necessary, especially when complex Bayesian models are used.
- **Data compression/feature extraction:** The methods of principal components analysis and extended linear modeling via *basis functions* can be used to extract the most pertinent features of the large microarray data sets.
- **Experimental design**: Statistical experimental design can assist in determining the number of replicates, the number of samples, the choice of time points at which the array data are collected and many other aspects of microarray experiments. In addition, power and sample size assessments can inform the experimenter as to the statistical worth of the microarray experiments that have been carried out.

Typically, data derived from both types of microarrays are highly noise and artefact corrupted. The statistical analysis of such data is therefore quite a challenging process. In many cases, the replicate experiments are very variable. The other main difficulty that arises in the statistical analysis of microarray data is the dimensionality; a vast number of gene expression measurements are available, usually only on a relatively small number of individual observations or samples, and thus it is hard to establish any general distributional models for the expression of a single gene.

Clustering

Cluster analysis is an unsupervised statistical procedure that aims to establish the presence of identifiable subgroups (or *clusters*) in the data, so that objects belonging to the same cluster resemble each other more closely than objects in different clusters; see [58,59] for comprehensive summaries.

In two or three dimensions, clusters can be visualized by plotting the raw data. With more than three dimensions, or in the case of dissimilarity data (see below), analytical assistance is needed. Broadly, clustering algorithms fall into two categories:

- **Partitioning Algorithms:** A partitioning algorithm divides the data set into K clusters, where and the algorithm is run for a range of K-values. Partitioning methods are based on specifying an initial number of groups, and iteratively reallocating observations between groups until some equilibrium is attained. The most famous algorithm is the *K-Means* algorithm in which the observations are iteratively classified as belonging to one of K groups, with group membership is determined by calculating the centroid for each group (the multidimensional version of the mean) and assigning each observation to the group with the closest centroid. The K-means algorithm alternates between calculating the centroids based on the current group memberships, and reassigning observations to groups based on the new centroids. A more robust method uses mediods rather than centroids (that is, medians rather than means in each dimension, and more generally, any distance-based allocation algorithm could be used.
- **Hierarchical Algorithms:** A hierarchical algorithm yields an entire hierarchy of clustering for the given data set. *Agglomerative methods* start with each object in the data set in its own cluster, and then successively merges clusters until only one large cluster remains. *Divisive methods* start by considering the whole data set as one cluster, and then splits up clusters until each object is separated. Hierarchical algorithms are discussed in detail in Sect. "Hierarchical Clustering".

Data sets for clustering of N observations can either take the form of an $N \times p$ data matrix, where rows contain the different observations, and columns contain the different variables, or an $N \times N$ dissimilarity matrix, whose (i,j)th element is d_{ij}, the distance or dissimilarity between observations i and j that obeys the usual properties of a metric. Typical data distance measures between two data points i and j with measurement vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ are the L_1 and L_2 Euclidean distances, and the grid-based *Manhattan distance* for discrete variables, or the *Hamming distance* for binary variables. For *ordinal* (ordered categorical) or *nominal* (label) data, other dissimilarities can be defined.

Hierarchical Clustering Agglomerative hierarchical clustering initially places each of the N items in its own cluster. At the first level, two objects are to be clustered together, and the pair is selected such that the potential function increases by the largest amount, leaving $N-1$ clusters, one with two members, the remaining $N-2$ each with one. At the next level, the optimal configuration of $N-2$ clusters is found, by joining two of the existing clusters. This process continuous until a single cluster remains containing all N items. At each level of the hierarchy, the merger chosen is the one that leads to the smallest increase in some objective function.

Classical versions of the hierarchical agglomeration algorithm are typically used with *average*, *single* or *complete* linkage methods, depending on the nature of the merging mechanism. Such criteria are inherently heuristic, and more formal *model-based* criteria can also be used. Model-based clustering is based on the assumption that the data are generated by a mixture of underlying probability distributions. Specifically, it is assumed that the population of interest consists of K different sub populations, and that the density of an observation from the the sub population is for some unknown vector of parameters. Model-based clustering is described in more detail in Sect. "Model–Based Hierarchical Clustering".

The principal display plot for a clustering analysis is the *dendrogram* which plots all of the individual data objects linked by means of a binary "tree". The dendrogram represents the structure inferred from a hierarchical clustering procedure which can be used to partition the data into subgroups as required if it is cut at a certain "height" up the tree structure. As with many of the aspects of the clustering procedures described above, it is more of a heuristic graphical representation rather than a formal inferential summary. However, the dendrogram is readily interpretable, and favored by biologists.

Model-Based Hierarchical Clustering Another approach to hierarchical clustering is *model-based clustering* (see for example [60,61]), which is based on the assumption that the data are generated by a mixture of K underlying probability distributions as in Eq. (11). Given data matrix $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)^{\mathsf{T}}$, let $\gamma = (\gamma_1, \ldots, \gamma_N)$ denote the cluster labels, where $\gamma_i = k$ if the ith data point comes from the kth sub-population. In the classification procedure, the maximum likelihood procedure is used to choose the parameters in the model.

Commonly, the assumption is made that the data in the different sub-populations follow multivariate normal distributions, with mean u_k and covariance matrix Σ_k for cluster k, so that

$$\begin{aligned} f_{Y|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta}) &= \sum_{k=1}^{K} \omega_k f_k(\boldsymbol{y}|u_k, \Sigma_k) \\ &= \sum_{k=1}^{K} \omega_k \frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma_k|^{1/2}} \\ &\qquad \exp\left\{-\frac{1}{2}(\boldsymbol{y}-u_k)^{\mathsf{T}} \Sigma_k^{-1} (\boldsymbol{y}-u_k)\right\} \end{aligned}$$

where $\Pr[\gamma_i = k] = \omega_k$. If $\Sigma_k = \sigma^2 \boldsymbol{I}_p$ is a $p \times p$ matrix, then maximizing the likelihood is the same as minimizing the sum of within-group sums of squares and corresponds to the case of hyper-spherical clusters with the same variance. Other forms of Σ_k yield clustering methods that are appropriate in different situations. The key to specifying this is the singular value or eigen decomposition of Σ_k, given by eigenvalues $\lambda_1, \ldots, \lambda_p$ and eigen vectors $\mathbf{v}_1, \ldots, \mathbf{v}_p$, as in Principal Components Analysis [62]. The eigen vectors of Σ_k, specify the orientation of the kth cluster, the largest eigenvalue λ_1 specifies its variance or size, and the ratios of the other eigenvalues to the largest one specify its shape. Further, if $\Sigma_k = \sigma_k^2 \boldsymbol{I}_p$, the criterion corresponds to hyper spherical clusters of different sizes, and by fixing the eigenvalue ratios $\alpha_j = \lambda_j/\lambda_1$ for $j = 2, 3, \ldots, p$ across clusters, other cluster shapes are encouraged.

Model-Based Analysis of Gene Expression Profiles

The clustering problem for vector-valued observations can be formulated using models used to represent the gene expression patterns via the *extended linear model*, that is, a linear model in non-linear basis functions; see, for example, [63,64] for details.

Generically, the aim of the statistical model is to capture the behavior of the gene expression ratio y_t as a function of time t. The basis of the modeling strategy would be to use models that capture the characteristic behavior of expression profiles likely to be observed due to different forms of regulation. A regression framework and model can be adopted. Suppose that Y_t is model led using a linear model

$$Y_t = \boldsymbol{X}_t \boldsymbol{\beta} + \varepsilon_t$$

where $\boldsymbol{X}_t$ is (in general) a $1 \times p$ vector of specified functions of t, and $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector. In vector representation, the gene expression profile over times $t_1, \ldots, t_T$ can be written $\boldsymbol{Y} = (Y_1, \ldots, Y_T)$,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon\,. \tag{13}$$

The precise form of design matrix $\boldsymbol{X}$ will be specified to model the time-variation in signal. Typically the random error terms $\{\varepsilon_t\}$ are taken as independent and identically distributed Normal random variables with variance σ^2, implying that the conditional distribution of the responses $\boldsymbol{Y}$ is multivariate normal

$$\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 \sim N\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_T\right) \tag{14}$$

where now $\boldsymbol{X}$ is $T \times p$ where $\boldsymbol{I}_T$ is the $T \times T$ identity matrix.

In order to characterize the underlying gene expression profile, the parameter vector $\boldsymbol{\beta}$ must be estimated. For this model, the maximum likelihood/ordinary least squares estimates of $\boldsymbol{\beta}$ and σ^2 are

$$\widehat{\boldsymbol{\beta}}_{\mathrm{ML}} = \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} \qquad \widehat{\sigma}^2 = \frac{1}{T-p}\left(\boldsymbol{y} - \widehat{\boldsymbol{y}}\right)^{\mathsf{T}}\left(\boldsymbol{y} - \widehat{\boldsymbol{y}}\right)$$

for fitted values $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\mathrm{ML}} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}$.

Bayesian Analysis in Model–Based Clustering In a Bayesian analysis of the model in (13) a joint prior distribution $\pi\left(\boldsymbol{\beta}, \sigma^2\right)$ is specified for $\left(\boldsymbol{\beta}, \sigma^2\right)$, and a posterior distribution conditional on the observed data is computed for the parameters. The calculation proceeds using Eq. (5) (essentially with $\mathcal{G}$ fixed).

$$\pi\left(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, \boldsymbol{X}\right) = \frac{L\left(\boldsymbol{y}; \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2\right)\pi\left(\boldsymbol{\beta}, \sigma^2\right)}{\int L\left(\boldsymbol{y}; \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2\right)\pi\left(\boldsymbol{\beta}, \sigma^2\right)\,\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\sigma^2}$$

where $L\left(\boldsymbol{y}; \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2\right)$ is the likelihood function. In the linear model context, a conjugate prior specification is used where

$$\begin{aligned} \pi\left(\boldsymbol{\beta}|\sigma^2\right) &\equiv \mathrm{Normal}\left(\boldsymbol{v}, \sigma^2 \boldsymbol{V}\right) \\ \pi\left(\sigma^2\right) &\equiv \mathrm{IGamma}\left(\frac{\alpha}{2}, \frac{\gamma}{2}\right) \end{aligned} \tag{15}$$

($\boldsymbol{v}$ is $p \times 1$, $\boldsymbol{V}$ is $p \times p$ positive-definite and symmetric, all other parameters are scalars) and IGamma denotes the inverse Gamma distribution. Using this prior, standard Bayesian calculations show that conditional on the data

$$\begin{aligned} \pi\left(\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2\right) &\equiv \mathrm{Normal}\left(\boldsymbol{v}^*, \sigma^2 \boldsymbol{V}^*\right) \\ \pi\left(\sigma^2|\boldsymbol{y}\right) &\equiv \mathrm{IGamma}\left(\frac{T+\alpha}{2}, \frac{c+\gamma}{2}\right) \end{aligned} \tag{16}$$

where

$$\begin{aligned} \boldsymbol{V}^* &= \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \boldsymbol{V}^{-1}\right)^{-1} \\ \boldsymbol{v}^* &= \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \boldsymbol{V}^{-1}\right)^{-1}\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} + \boldsymbol{V}^{-1}\boldsymbol{v}\right) \\ c &= \boldsymbol{y}^{\mathsf{T}}\boldsymbol{y} + \boldsymbol{v}^{\mathsf{T}}\boldsymbol{V}^{-1}\boldsymbol{v} - \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} + \boldsymbol{V}^{-1}\boldsymbol{v}\right)^{\mathsf{T}} \\ &\qquad \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \boldsymbol{V}^{-1}\right)^{-1}\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} + \boldsymbol{V}^{-1}\boldsymbol{v}\right) \end{aligned} \tag{17}$$

In regression modeling, it is usual to consider a centered parametrization for $\boldsymbol{\beta}$ so that $\boldsymbol{v} = 0$, giving

$$\begin{aligned} \boldsymbol{v}^* &= \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \boldsymbol{V}^{-1}\right)^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} \\ c &= \boldsymbol{y}^{\mathsf{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathsf{T}}\boldsymbol{X}^{\mathsf{T}}\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \boldsymbol{V}^{-1}\right)^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} \\ &= \boldsymbol{y}^{\mathsf{T}}\left(\boldsymbol{I}_T - \boldsymbol{X}\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \boldsymbol{V}^{-1}\right)^{-1}\boldsymbol{X}^{\mathsf{T}}\right)\boldsymbol{y} \end{aligned}$$

A critical quantity in a Bayesian clustering procedure is the marginal likelihood, as in Eq. (8), for the data in light of the model:

$$f_Y(\boldsymbol{y}) = \int f_{Y|\boldsymbol{\beta},\sigma^2}\left(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2\right)\pi\left(\boldsymbol{\beta}|\sigma^2\right)\pi\left(\sigma^2\right)\,\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\sigma^2\,. \tag{18}$$

Combining terms above gives that

$$f_Y(\boldsymbol{y}) = \left(\frac{1}{\pi}\right)^{T/2}\frac{\gamma^{\alpha/2}\Gamma\left(\frac{T+\alpha}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)}\frac{|\boldsymbol{V}^*|^{1/2}}{|\boldsymbol{V}|^{1/2}}\frac{1}{\{c+\gamma\}^{(T+\alpha)/2}} \tag{19}$$

This expression is the marginal likelihood for a single gene expression profile. For a collection of profiles belonging to a single cluster, $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N$, Eq. (19) can again be evaluated and used as the basis of a dissimilarity measure as an input into a hierarchical clustering procedure. The marginal likelihood in Eq. (19) can easily be re-expressed for clustered data. The basis of the hierarchical clustering method outlined in [64] proceeds by agglomeration of clusters from N to 1, with the two clusters that lead to the **greatest increase** marginal likelihood score at each stage of the hierarchy. This method works for profiles of arbitrary length, potentially with different observation time points, however it is computationally most efficient when the time points are the same for each profile.

The design matrix $\boldsymbol{X}$ is typically expressed via nonlinear basis functions, for example truncated polynomial splines, Fourier bases or wavelets. For T large, it is usually necessary to use a projection through a lower number of bases; for example, for a single profile, $\boldsymbol{X}$ becomes $T \times p$ and $\boldsymbol{\beta}$ becomes $p \times 1$, for $T > p$. Using different designs, many flexible models for the expression profiles can be fitted. In some cases, the linear mixed effect formulation in Eq. (10) can be used to construct the spline-based models; in such models, some of the β parameters are themselves assumed to be random effects (see [65]).

For example, in *harmonic regression*, regression in the Fourier bases is carried out. Consider the extended linear model

$$Y_t = \sum_{j=0}^{p}\beta_j g_j(t) + \varepsilon_t$$

where $g_0(t) = 1$ and

$$g_j(t) = \begin{cases}\cos(\phi_j t) & j \text{ odd}\\ \sin(\phi_j t) & j \text{ even}\end{cases}$$

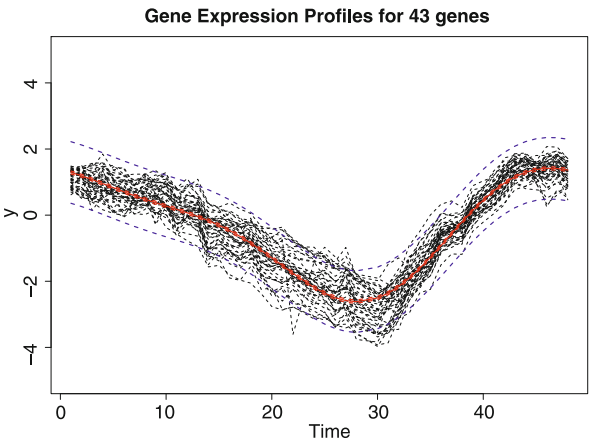

Complexity in Systems Level Biology and Genetics: Statistical Perspectives, Figure 1
Cluster of gene expression profiles obtained using Bayesian hierarchical model-based clustering: data from the intraerythrocytic developmental cycle of protozoa *Plasmodium falciparum*. Clustering achieved using harmonic regression model with $k = 2$. Solid red line is posterior mean for this cluster, dotted red lines are point wise 95% credible intervals for the cluster mean profile, and dotted blue lines are point wise 95% credible intervals for the observations

where p is an even number, $p = 2k$ say, and $\phi_j, j = 1, 2, \ldots, k$ are constants with $\phi_1 < \phi_2 < \cdots < \phi_k$. For fixed t, $\cos(\phi_j t)$ and $\sin(\phi_j x)$ are also fixed and this model is a linear model in parameters

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathsf{T}}\,.$$

This model can be readily fitted to time-course expression profiles. The plot below is a fit of the model with $k = 2$ to a cluster of profiles extracted using the method described in [64] from the malaria protozoa *Plasmodium falciparum* data set described in [66].

One major advantage of the Bayesian inferential approach is that any biological prior knowledge that is available can be incorporated in a coherent fashion. For example, the data in Figure 1 illustrate periodic behavior related to the cyclical nature of cellular organization, and thus the choice of the Fourier bases is a natural one.

Choosing the Number of Clusters: Bayesian Information Criterion A hierarchical clustering procedure gives the sequence by which the clusters are merged (in agglomerative clustering) or split (in divisive clustering) according the model or distance measure used, but does not give an indication for the number of clusters that are present in the data (under the model specification). This is obviously an important consideration. One advantage of the model-

based approach to clustering is that it allows the use of statistical model assessment procedures to assist in the choice of the number of clusters. A common method is to use approximate *Bayes factors* to compare models of different orders (i. e. models with different numbers of clusters), and gives a systematic means of selecting the parametrization of the model, the clustering method, and also the number of clusters (see [67]).

The Bayes factor is the posterior odds for one model against the other assuming neither is favored a priori. A reliable approximation to twice the log Bayes factor called the *Bayesian Information Criterion* (BIC), which, for model M fitted to n data points is given by

$$\mathrm{BIC}_M = -2 \log L_M(\widehat{\boldsymbol{\theta}}) + d_M \log n$$

where L_M is the Bayesian marginal likelihood from Eq. (18), $L_M(\widehat{\boldsymbol{\theta}})$ is the maximized log likelihood of the data for the model M, and d_M is the number of parameters estimated in the model. The number of clusters is not considered a parameter for the purposes of computing the BIC. The smaller (more negative) the value of the BIC, the stronger the evidence for the model.

Classification via Model–Based Clustering Any clustering procedure can be used as the first step in the construction of *classification* rules. Suppose that it, on the basis of an appropriate decision procedure, it is known that there are C clusters, and that a set of existing expression profiles $y_1, \ldots, y_N$ have been allocated in turn to the clusters. Let $z_1, \ldots, z_N$ be the cluster allocation labels for the profiles. Now, suppose further that the C clusters can be decomposed further into two subsets of sizes C_0 and C_1, where the subsets represent perhaps clusters having some common, known biological function or genomic origin. For example, in a cDNA microarray, it might be known that the clones are distinguishable in terms of the organism from which they were derived. A new objective could be to allocate a novel gene and expression profile to one of the subsets, and one of the clusters within that subset.

Let y_{ijk}, for $i = 0, 1$, $j = 1, 2, \ldots, C_i$, $k = 1, 2, \ldots, N_{ij}$ denote the kth profile in cluster j in subset i. Let y^* denote a new profile to be classified, and ξ^* be the binary classification-to-subset, and z^* the classification-to-cluster variable for y^*. Then, by Bayes Rule, for $i = 1, 2$,

$$P\left[\xi^* = i | y^*, y, z\right] \propto p\left(y^* | \xi^* = i, y, z\right) P\left[\xi^* = i | y, z\right] \tag{20}$$

The two terms in Eq. (20) can be determined on the basis of the clustering output.

Metabolomics

The term *metabolome* refers to the total metabolite content of an organic sample (tissue, blood, urine etc) obtained from a living organism which represents the products of a higher level of biological interaction than that which occurs within the cell. *Metabolomics* and *metabonomics* are the fields in biomedical investigation that combines the application of nuclear magnetic resonance (NMR) *spectroscopy* with multivariate statistical analysis in studies of the composition of the samples. Metabonomics is often used in reference to the static chemical content of the sample, whereas metabolomics is used to refer to the dynamic evolution of the metabolome. Both involve the measurement of the metabolic response to interventions – see for example [68] – and applications of metabolomics include several in public health and medicine [69,70].

Statistical Methods for Spectral Data

The two principal spectroscopic measurement platforms, NMR and Mass Spectrometry (MS) yield alternative representations of the metabolic spectrum. They produce spectra (or profiles) that consist of several thousands of individual measurements at different resonances or masses. There are several phases of processing of such data; preprocessing using smoothing, alignment and de-noising, peak separation, registration and signal extraction. For an extensive discussion, see [62].

An NMR spectrum consists of measurements of the intensity or frequency of different biochemical compounds (metabolites) represented by a set of *resonances* dependent upon the chemical structure, and can be regarded as a linear combination of peaks (nominally of various widths) that correspond to singletons or multiple peaks according to the neighboring chemical environment. A typical spectrum extracted from rat urine is depicted in Fig. 2 see [71]. Two dominant sharp peaks are visible.

Features of the spectra that require specific statistical modeling include multiple peaks for a single compound, variation in peak shape, and chemical shifts induced by variation in experimental pH Signals from different metabolites can be highly overlapped and subject to peak position variation due primarily to pH variations in the samples, and there are many small scale features (see Fig. 3). Statistical methods of pre-processing NMR spectra for statistical analysis which address the problems outlined above, using, for example, dynamic time warping to achieve alignment of resonance peaks across replicate spectra as a form or spectral registration form part of the necessary holistic Bayesian framework.

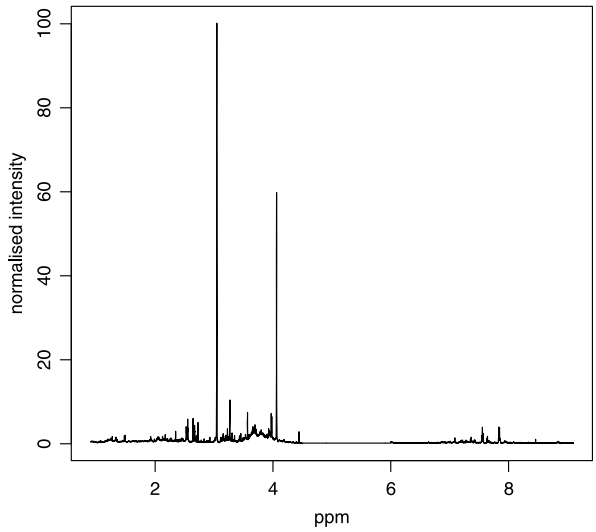

Complexity in Systems Level Biology and Genetics: Statistical Perspectives, Figure 2
A normalized rat urine spectrum. The ordinate is parts per million, the abscissa is intensity after standardization

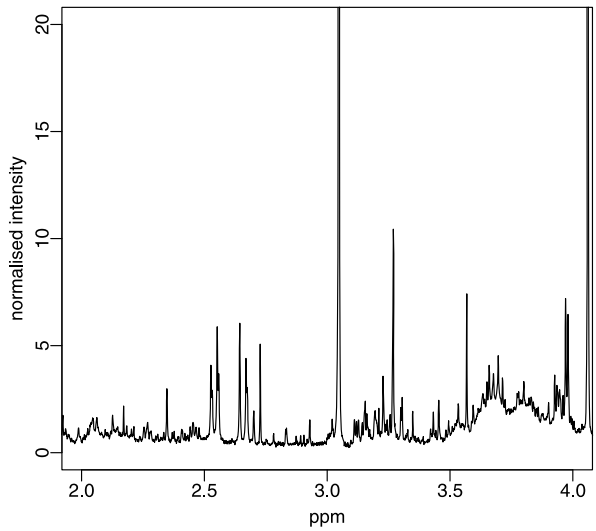

Complexity in Systems Level Biology and Genetics: Statistical Perspectives, Figure 3
Magnified portion of the spectrum showing small scale features

Classical statistical methods for metabolic spectra include the following:

- **Principal Components Analysis (PCA) and Regression:** a linear data projection method for dimension reduction, feature extraction, and classification of samples in an unsupervised fashion, that is, without reference to labeled cases.
- **Partial Least Squares (PLS):** a non-linear projection method similar to PCA, but implemented in a supervised setting for sample discrimination.
- **Clustering:** Clusters of spectra, or peaks within spectra, can be discovered using similar techniques to those described in Sect. "Clustering".
- **Neural Networks:** Flexible non-linear regression models constructed from simple mathematical functions that are learned from the observation of cases, that are ideal models for classification. The formulation of an neural netweork involves three levels of interlinked variables; *outputs*, *inputs*, and *hidden variables*, interpreted as a collection of unobserved random variables that form the hidden link between inputs and outputs.

Bayesian Approaches

The Bayesian framework is a natural one for incorporating genuine biological prior knowledge into the signal reconstruction, and typically useful prior information (about fluid composition, peak location, peak multiplicity) is available. In addition, a hierarchical Bayesian model structure naturally allows construction of plausible models for the spectra across experiments or individuals.

- **Flexible Bayesian Models:** The NMR spectrum can be represented as a noisy signal derived from some underlying and biologically important mechanism. Basis-function approaches (specifically, wavelets) have been much used to represent non-stationary time-varying signals [65,71,72,73]. The sparse representation of the NMR spectrum in terms of wavelet coefficients makes them an excellent tool in data compression, yet these coefficients can still be easily transformed back to the spectral domain to give a natural interpretation in terms of the underlying metabolites.
  Figure 4 depicts the reconstruction of the rat urine spectrum in the region between 2.5 and 2.8 ppm using wavelet methods; see [71].
- **Bayesian Time Series Models for Complex Non-stationary Signals:** See for example [74]. The duality between semi-parametric modeling of functions and latent time series models allows a view of the analysis of the underlying NMR spectrum not as a set of point wise evaluations of a function, but rather as a (time-ordered) series of correlated observations with some identifiable latent structure. Time series models, computed using dynamic calculation (filtering), provide a method for representing the NMR spectra parsimoniously.
- **Bayesian Mixture Models:** A reasonable generative model for the spectra is one that constructs the

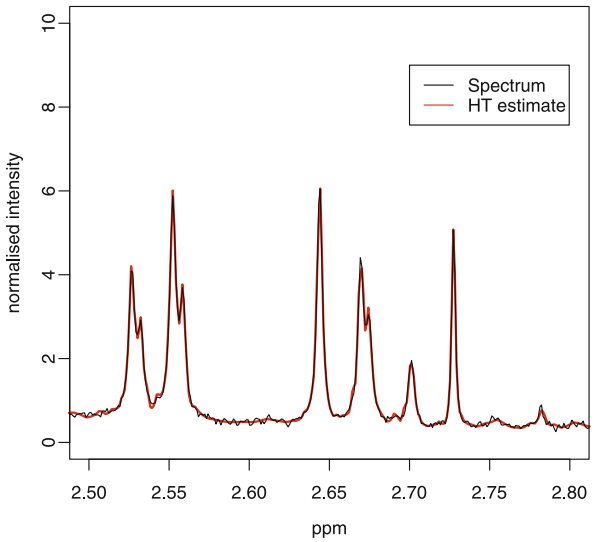

Complexity in Systems Level Biology and Genetics: Statistical Perspectives, Figure 4
Wavelet reconstruction of a region of the spectrum results under the "Least Asymmetric wavelet" with four vanishing moments using the hard thresholding (HT)

spectra from a large number of symmetric peaks of varying size, corresponding to the contributions of different biochemical compounds. This can be approximated using a finite mixture model, where the number, magnitudes and locations, of the spectral contributions are unknown. Much recent research has focused on the implementation of computational strategies for Bayesian mixtures, in particular Markov chain Monte Carlo (MCMC) and Sequential Monte Carlo (SMC) have proved vital. The reconstruction of NMR spectra is a considerably more challenging area than those for which mixture modeling is conventionally used, as many more individual components are required. Flexible semi-parametric mixture models have been utilized in [75,76], whilst fully non-parametric mixture models similar to those described in Sect. "Mixture Models" can also be used [73].

A major advantage of using the fully Bayesian framework is that, once again, all relevant information (the spectral data itself, knowledge of the measurement processes for different experimental platforms, the mechanisms via which multiple peaks and shifts are introduced) can be integrated in a coherent fashion. In addition, prior knowledge about the chemical composition of the samples can be integrated via a prior distribution constructed by inspection of the profiles for training samples. At a higher level of synthesis, the Bayesian paradigm offers a method for integrating metabolomic data with other functional or structural data, such as gene expression or protein expression data. Finally, the metabolic content of tissue changes temporally, so dynamic modeling of the spectra could also be attempted.

Future Directions

Biological data relating to structure and function of genes, proteins and other biological substances are now available from a wide variety of platforms. Researchers are beginning to develop methods for coherent combination of data from different experimental processes to get an entire picture of biological cause and effect. For example, the effective combination of gene expression and metabonomic data will be of tremendous utility. A principal challenge is therefore the fusion of expression data derived from different experimental platforms, and seeking links with sequence and ontological information available. Such fusion will be critical in the future of statistical analysis of large scale systems biology and bioinformatics data sets.

In terms of public health impact of systems biology and statistical genomics, perhaps the most prominent is the study of common diseases through high-throughput genotyping of single nucleotide polymorphisms (SNPs). In genome wide association studies, SNP locations that correlate with disease status or quantitative trait value are sought. In such studies, the key statistical step involves the selection of informative predictors (SNP or genomic loci) from a large collection of candidates. Many such genome wide studies have been completed or are ongoing (see [42,43,77,78]). Such studies represent huge challenges for statisticians and mathematical modelers, as the data contain many subtle structures but also as the amount of information is much greater than that available for typical statistical analysis.

Another major challenge to the quantitative analysis of biological data comes in the form of image analysis and extraction. Many high throughput technologies rely on the extraction of information from images, either in static form, or dynamically from a series of images. For example it is now possible to track the expression level of mRNA transcripts in real-time ([79,80,81]), and to observe mRNA transcripts moving from transcription sites to translation sites (see for example [82]). Imaging techniques can also offer insights into aspects of the dynamic organization of nuclear function by studying the positioning of nuclear compartments and how those compartments reposition themselves in relation to each other through time. The challenges for the statistician are to develop real-time analysis methods for tracking and quantifying the nature and

content of such images, and tools from spatial modeling and time series analysis will be required.

Finally, *flow cytometry* can measure characteristics of millions of cells simultaneously, and is a technology that offers many promises for insights into biological organization and public health implications. However, quantitative measurement and analysis methods are only yet in the early stages of development, but offer much promise (see [83,84]).

Bibliography

1. Kitano H (ed) (2001) Foundations of Systems Biology. MIT Press, Cambridge
2. Kitano H (2002) Computational systems biology. Nature 420(6912):206–210
3. Alon U (2006) An Introduction to Systems Biology. Chapman and Hall, Boca Raton
4. Edwards AWF (2000) Foundations of mathematical genetics, 2nd edn. Cambridge University Press, Cambridge
5. Yule GU (1924) A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis. Philos Trans R Soc Lond Ser B 213:21–87
6. Fisher RA (1922) On the dominance ratio. Proc R Soc Edinburgh 42:321–341
7. Fisher RA (1930) The genetical theory of natural selection. Clarendon Press, Oxford
8. Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159
9. Murray JD (2002) Mathematical Biology: I An Introduction. Springer, New York
10. Murray JD (2003) Mathematical Biology: II Spatial Models and Biomedical Applications. Springer, New York
11. Lewin B (2007) Genes, 9th edn. Jones & Bartlett Publishers, Boston
12. Spector DL (2001) Nuclear domains. J Cell Sci 114(16):2891–3
13. Bernardo JM, Smith AFM (1994) Bayesian Theory. Wiley, New York
14. Haefner JW (ed) (2005) Modeling Biological Systems: Principles and Applications, 2nd edn. Springer, New York
15. Ramsay JO, Hooker G, Campbell D, Cao J (2007) Parameter estimation for differential equations: a generalized smoothing approach. J Royal Stat Soc: Series B (Methodology) 69(5):741–796
16. Donnet S, Samson A (2007) Estimation of parameters in incomplete data models defined by dynamical systems. J Stat Plan Inference 137(9):2815–2831
17. Rogers S, Khanin R, Girolami M (2007) Bayesian model-based inference of transcription factor activity. BMC Bioinformatics 8(Suppl 2) doi:10.1186/1471-2105-8-S2-S2
18. Wilkinson DJ (2006) Stochastic Modelling for Systems Biology. Chapman & Hall (CRC), Boca Raton
19. Heron EA, Finkenstädt B, Rand DA (2007) Bayesian inference for dynamic transcriptional regulation; the hes1 system as a case study. Bioinformatics 23(19):2596–2603
20. Airoldi EM (2007) Getting started in probabilistic graphical models. PLoS Comput Biol 3(12):e252
21. Husmeier D, Dybowski R, Roberts S (eds) (2005) Probabilistic Modelling in Bioinformatics and Medical Informatics. Springer, Ney York
22. Friedman N (2004) Inferring cellular networks using probabilistic graphical models. Science 303:799–805
23. Opgen-Rhein R, Strimmer K (2007) From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. BMC Syst Biol 1:37:1–10
24. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc Natl Acad Sci USA 97:12182–12186
25. Friedman N, Linial M, Nachman I, Pe'er D (2000) Using bayesian networks to analyze expression data. J Comput Biol 7:601–620
26. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA Jr, Marks JR, Nevins JR (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci USA 98(11):462–467
27. Dobra A, Hans C, Jones B, Nevins J, Yao G, West M (2004) Sparse graphical models for exploring gene expression data. J Multivar Anal 90:196–212
28. Jones B, Carvalho C, Dobra A, Hans C, Carter C, West M (2005) Experiments in stochastic computation for high dimensional graphical models. Stat Sci 20:388–400
29. Markowetz F, Bloch J, Spang R (2005) Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. Bioinformatics 21:4026–4032
30. Eaton D, Murphy KP (2007) Exact Bayesian structure learning from uncertain interventions. Artificial Intelligence & Statistics 2:107–114
31. Robert CP (2007) The Bayesian Choice: From Decision–Theoretic Foundations to Computational Implementation. Texts in Statistics, 2nd edn. Springer, New York
32. Doucet A, de Freitas N, Gordon NJ (eds) (2001) Sequential Monte Carlo Methods in Practice, Statistics for Engineering and Information Science. Springer, New York
33. Robert CP, Casella G (2005) Monte Carlo Statistical Methods. Texts in Statistics, 2nd edn. Springer, New York
34. Gamerman D, Lopes HF (2006) Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. Texts in Statistical Science, 2nd edn. Chapman and Hall (CRC), Boca Raton
35. Antoniak CE (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann Stat 2:1152–1174
36. Escobar MD, West M (1995) Bayesian density estimation and inference using mixtures. J Am Stat Assoc 90(430):577–588
37. Dahl DB (2006) Model-based clustering for expression data via a Dirichlet process mixture model. In: Do KA, Müller P, Vannucci M (eds) Bayesian Inference for Gene Expression and Proteomics. University Press, Cambridge, Chap 10
38. Kim S, Tadesse MG, Vannucci M (2006) Variable selection in clustering via Dirichlet process mixture models. Biometrika 93(4):877–893
39. West M, Harrison J (1999) Bayesian Forecasting and Dynamic models, 2nd edn. Springer, New York
40. Philipov A, Glickman ME (2006) Multivariate stochastic volatility via Wishart processes. J Bus Econ Stat 24(3):313–328
41. Gresham D, Dunham MJ, Botstein D (2008) Comparing whole genomes using DNA microarrays. Nat Rev Genet 9:291–302
42. The Wellcome Trust Case Control Consortium (2007) Association scan of 14,500 nonsynonymous snps in four diseases identifies autoimmunity variants. Nat Genet 39:1329–1337

43. The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678
44. Liu XS (2007) Getting started in tiling microarray analysis. PloS Comput Biol 3(10):1842–1844
45. Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, Brown M, Liu XS (2006) Model-based analysis of tiling-arrays for ChIP-chip. Proc Natl Acad Sci USA 103(33):12457–62 (2006)
46. Freeman JL et al (2006) Copy number variation: New insights in genome diversity. Genome Res 16:949–961
47. Urban AE et al (2006) High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. Proc Natl Acad Sci USA 103(12):4534–4539
48. Saha S et al (2002) Using the transcriptome to annotate the genome. Nat Biotech 20:508–512
49. Shadeo A et al (2007) Comprehensive serial analysis of gene expression of the cervical transcriptome. BMC Genomics 8:142
50. Robinson SJ, Guenther JD, Lewis CT, Links MG, Parkin IA (2007) Reaping the benefits of SAGE. Methods Mol Biol 406:365–386
51. Lu C, Tej SS, Luo S, Haudenschild CD, Meyers BC, Green PJ (2005) Elucidation of the Small RNA Component of the Transcriptome. Science 309(5740):1567–1569
52. Weiner H, Glökler J, Hultschig C, Büssow K, Walter G (2006) Protein, antibody and small molecule microarrays. In: Müller UR, Nicolau DV (eds) Microarray Technology and Its Applications. Biological and Medical Physics. Biomedical Engineering. Springer, Berlin, pp 279–295
53. Speed TP (ed) (2003) Statistical Analysis of Gene Expression Microarray Data. Chapman & Hall/CRC, Bacon Raton
54. Parmigiani G, Garett ES, Irizarry RA, Zeger SL (eds) (2003) The Analysis of Gene Expression Data. Statistics for Biology and Health. Springer, New York
55. Wit E, McClure J (2004) Statistics for Microarrays: Design, Analysis and Inference. Wiley, New York
56. Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S (eds) (2005) Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Statistics for Biology and Health. Springer, New York
57. Do KA, Müller P, Vannucci M (2006) Bayesian Inference for Gene Expression and Proteomics. Cambridge University Press, Cambridge
58. Everitt BS, Landau S, Leese M (2001) Cluster Analysis, 4th edn. Hodder Arnold, London
59. Kaufman L, Rousseeuw PJ (2005) Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Series in Probability and Statistics, 2nd edn. Wiley, Ney York
60. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL (2001) Model-based clustering and data transformation for gene expression data. Bioinformatics 17:977–987
61. McLachlan GJ, Bean RW, Peel D (2002) A mixture model-based approach to the clustering of microarray expression data. Bioinformatics 18:413–422
62. De Iorio M, Ebbels TMD, Stephens DA (2007) Statistical techniques in metabolic profiling. In: Balding DJ, Bishop M, Cannings C (eds) Handbook of Statistical Genetics, 3rd edn. Wiley, Chichester, Chap 11
63. Heard NA, Holmes CC, Stephens DA, Hand DJ, Dimopoulos G (2005) Bayesian coclustering of Anopheles gene expression time series: Study of immune defense response to multiple experimental challenges. Proc Natl Acad Sci USA 102(47):16939–16944
64. Heard NA, Holmes CC, Stephens DA (2006) A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves. J Am Stat Assoc 101(473):18–29
65. Morris JS, Brown PJ, Baggerly KA, Coombes KR (2006) Analysis of mass spectrometry data using Bayesian wavelet-based functional mixed models. In: Do KA, Müller P, Vannucci M (eds) Bayesian Inference for Gene Expression and Proteomics. Cambridge University Press, Cambridge, pp 269–292
66. Bozdech Z, Llinás M, Pulliam BL, Wong ED, Zhu J, DeRisi JL (2003) The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. PLoS Biol 1(1):E5
67. Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90:773–795
68. Nicholson JK, Connelly J, Lindon JC, Holmes E (2002) Metabonomics: a platform for studying drug toxicity and gene function. Nat Rev Drug Discov 1:153–161
69. Lindon JC, Nicholson JK, Holmes E, Antti H, Bollard ME, Keun H, Beckonert O, Ebbels TM, Reily MD, Robertson D (2003) Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project. Toxic Appl Pharmacol 187:137
70. Brindle JT, Annti H, Holmes E, Tranter G, Nicholson JK, Bethell HWL, Clarke S, Schofield SM, McKilligin E, Mosedale DE, Graingerand DJ (2002) Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics. Nat Med 8:143
71. Yen TJ, Ebbels TMD, De Iorio M, Stephens DA, Richardson S (2008) Analysing real urine spectra with wavelet methods. (in preparation)
72. Brown PJ, Fearn T, Vannucci M (2001) Bayesian wavelet regression on curves with applications to a spectroscopic calibration problem. J Am Stat Soc 96:398–408
73. Clyde MA, House LL, Wolpert RL (2006) Nonparametric models for proteomic peak identification and quantification. In: Do KA, Müller P, Vannucci M (eds) Bayesian Inference for Gene Expression and Proteomics. Cambridge University Press, Cambridge, pp 293–308
74. West M, Prado R, Krystal A (1999) Evaluation and comparison of EEG traces: Latent structure in non-stationary time series. J Am Stat Assoc 94:1083–1095
75. Ghosh S, Grant DF, Dey DK, Hill DW (2008) A semiparametric modeling approach for the development of metabonomic profile and bio-marker discovery. BMC Bioinformatics 9:38
76. Ghosh S, Dey DK (2008) A unified modeling framework for metabonomic profile development and covariate selection for acute trauma subjects. Stat Med 30;27(29):3776–88
77. Duerr RH et al (2006) A Genome–Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene. Science 314(5804):1461–1463
78. Sladek R et al (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445:881–885
79. Longo D, Hasty J (2006) Imaging gene expression: tiny signals make a big noise. Nat Chem Biol 2:181–182
80. Longo D, Hasty J (2006) Dynamics of single-cell gene expression. Mol Syst Biol 2:64
81. Wells AL, Condeelis JS, Singer RH, Zenklusen D (2007) Imaging real-time gene expression in living systems with single-tran-

script resolution: Image analysis of single mRNA transcripts. CSH Protocols, Cold Springer Habor
82. Rodriguez AJ, Condeelis JS, Singer RH, Dictenberg JB (2007) Imaging mRNA movement from transcription sites to translation sites. Semin Cell Dev Biol 18(2):202–208
83. Lizard G (2007) Flow cytometry analyses and bioinformatics: Interest in new softwares to optimize novel technologies and to favor the emergence of innovative concepts in cell research. Cytom A 71A:646–647
84. Lo K, Brinkman RR, Gottardo R (2008) Automated gating of flow cytometry data via robust model-based clustering. Cytom Part A 73A(4):321–332

Complex Networks and Graph Theory

Geoffrey Canright
Telenor R&I, Fornebu, Norway

Article Outline

Glossary
Definition of the Subject
Introduction
Graphs, Networks, and Complex Networks
Structure of Networks
Dynamical Network Structures
Dynamical Processes on Networks
Graph Visualization
Future Directions
Bibliography

Glossary

Directed/Undirected graph A set of vertices connected by directed or undirected edges. A directed edge is one-way ($A \to B$), while an undirected edge is two-way or symmetric: $A - B$.

Network For our purposes, a network is defined identically to a graph: it is an abstract object composed of vertices (nodes) joined by (directed or undirected) edges (links). Hence we will use the terms 'graph' and 'network' interchangeably.

Graph topology The list of nodes i and edges (i, j) or $(i \to j)$ defines the topology of the graph.

Graph structure There is no single agreed definition for what constitutes the "structure" of a graph. To the contrary: this question has been the object of a great deal of research—research which is still ongoing.

Node degree distribution One crude measure of a graph's structure. If n_k is the number of nodes having degree k in a graph with N nodes, then the set of n_k is the node degree distribution—which is also often expressed in terms of the frequencies $p_k = n_k/N$.

Small-worlds graph A "small-worlds graph" has two properties: it has short path lengths (as is typical of random graphs)—so that the "world" of the network is truly "small", in that every node is within a few (or not too many) hops of every other; and secondly, it has (like real social networks, and *unlike* random graphs) a significant degree of *clustering*—meaning that two neighbors of a node have a higher-than-random probability of also being linked to one another.

Graph visualization The problem of displaying a graph's topology (or part of it) in a 2D image, so as to give the viewer insight into the structure of the graph. We see that this is a hard problem, as it involves both the unsolved problem of what we mean by the structure of the graph, and also the combined technological/psychological problem of conveying useful information about a (possibly large) graph via a 2D (or quasi-3D) layout. Clearly, the notion of a good graph visualization is dependent on the use to which the visualization is to be put—in other words, on the information which is to be conveyed.

Section Here, a 'bookkeeping' definition. This article introduces the reader to all of the other articles in the Section of the Encyclopedia which is titled "Complex Networks and Graph Theory". Therefore, whenever the word 'Section' (with a large 'S') is used in this 'roadmap' article, the word refers to that Section of the Encyclopedia. To avoid confusion, the various subdivisions of this roadmap article will be called 'parts'.

Definition of the Subject

The basic network concept is very simple: objects, connected by relationships. Because of this simplicity, the concept turns up almost everywhere one looks. The study of networks (or equivalently, graphs), both theoretically and empirically, has taken off over the last ten years, and shows no sign of slowing down. The field is highly interdisciplinary, having important applications to the Internet and the World Wide Web, to social networks, to epidemiology, to biology, and in many other areas. This introductory article serves as a reader's guide to the 13 articles in the Section of the Encyclopedia which is titled "Complex Networks and Graph Theory". These articles will be discussed in the context of three broad themes: network structure; dynamics of network structure; and dynamical processes running over networks.

Introduction

In the past ten years or so, the study of graphs has exploded, leaving forever the peaceful sanctum of pure

mathematics to become a fundamental concept in a vigorous, ill-defined, interdisciplinary, and important field of study. The most common descriptive term for this field is the "study of complex networks". This term is distinguished from the older, more mathematically-bound term "graph theory" in two ways. First – and perhaps most important – this new field is not just theoretical; to the contrary, the curious researcher finds that there is an enormous variety of empirically obtained graphs/networks (we use the terms interchangeably here) available as datasets on the Web. That is, one studies real, measured graphs; and not surprisingly, this empirical connection gives the endeavor much more of an applied flavor also.

The second distinction is perhaps not well motivated by the words used; but the "study of complex networks" typically means studying networks whose structure deviates in important ways from the "classical random graphs" of Erdős and Rényi [10,11]. We note that these two points are related: as one turned to studying real networks [18], one found that they were not described by classical random graphs [24,25]; hence one was forced to look at other kinds of structures.

This article is a reader's guide to the other articles in the Section entitled "Complex Networks and Graph Theory". The inclusion of both terms was very deliberate: graph theory gives the mathematical foundation for the more messy endeavor called "complex networks"; and the two fields have a strong and fruitful interaction. In this article I will describe the 13 other articles of this Section.

These 13 articles amply document the interdisciplinary nature of this exciting field: we find represented mathematics, biology, the Web and the Internet, software, epidemiology, and social networks (with the latter field also having its own Section in the Encyclopedia). For this reason, I will not define the parts of this article by field of study, but rather by general *themes* which run through essentially all studies of networks (at least in principle, and often in fact). In each part of this article I will point out those articles which represent that part's theme to a significant degree. Hence these themes are meant to tie together all of the articles into a simple framework.

In the next part "Graphs, Networks, and Complex Networks", I will concisely present the basic terminology which is in use. Then in entitled part the "Structure of Networks" I discuss the knotty question of the *structure* of a graph; this problem is the first theme, and it is very much unfinished business. In part "Dynamical Network Structures" we look at network topologies (and structures) which are *dynamic* rather than static. This is clearly an important theme, since (i) real empirical networks are necessarily dynamic (on some – often short – time scale), and (ii) the study of how networks grow and evolve can be highly useful for understanding how they have the structure we observe today. Then in part "Dynamical Processes on Networks" we look at the very large question of dynamical processes which take place *over* networks. Examples which illustrate the importance of this topic are epidemic spreading over social and computer networks, and the activation/inhibition processes going on over gene (or neural) networks. Clearly the progress of such dynamical processes may be strongly dependent on the underlying network structure; hence we see that all of these themes (parts "Structure of Networks"–"Dynamical Processes on Networks") are tightly related.

In part "Graph Visualization" we look briefly at an important but somewhat 'orthogonal' theme, namely, the problem of *graph visualization:* given a graph's topology, how can one best present this information visually to a human viewer? Finally, part "Future Directions" offers a very brief, modest, and personal take on the very large question of "future directions" in research on complex networks.

Graphs, Networks, and Complex Networks

Graphs

One of the earliest applications of graph theory may be found in Euler's solution, in 1736, of the problem called 'The seven bridges of Königsberg'. Euler considered the problem of finding a continuous path crossing each of the seven bridges of Königsberg exactly once, and solved the problem by representing each connected piece of land as a vertex (node) of a graph, and each bridge as an undirected (two-way) link. (For a nice presentation of this problem, see http://mathforum.org/isaac/problems/bridges2.html.) This little problem is an excellent example of the power of mathematics to extract understanding via abstraction: the lay person may stare at the bridges, islands etc, and try various ideas—but reducing the entire problem to an abstract graph, composed only of nodes and links, aids the application of pure reason, leading to a final and utterly convincing solution.

To study graphs is to study discrete objects and the relationships between them. Hence graph theory may be regarded as a branch of combinatorics. Erdős and Rényi [10,11] founded the study of "classical random graphs". These graphs are specified by the node number N (which typically is assumed to grow large), and by links laid down at random between these nodes, subject to various constraints. One such constraint is that every node have exactly k links—giving a ($k-$) *regular random graph.* A more relaxed constraint is simply to specify m links in total (so that the average node degree is $\langle k \rangle = m/N$). Fur-

ther relaxing of this constraint gives that every possible link, out of the set of

$$\begin{bmatrix} N \\ 2 \end{bmatrix} = N(N-1)/2$$

possible links, is included with probability p. This gives the average node degree (now averaged over many graphs) as $\langle k \rangle = \langle m \rangle / N = p(N-1)/2$. While all of these types of classical random graphs are similar "in spirit", those with the fewest constraints are those for which it is easiest to prove things.

We are fortunate to have in our Section the article ► Random Graphs, A Whirlwind Tour of by Fan Chung. In this article, the reader is given a good introduction to classical random graphs, along with a thorough presentation of the more modern theory of random graphs with a new, and more realistic, type of constraint—namely that they should, on average, have a given *node degree distribution*. This new theory makes the study of random graphs extremely relevant for today's empirically-anchored research: many empirical graphs are characterized by their node degree distribution, and many of these in fact have a *power-law degree distribution*, such that the number n_k of nodes having degree k varies with k by a power law: $n_k \sim k^{-\beta}$. This work is useful, precisely because a random graph with a given node degree distribution is the "most typical" graph of the set of graphs with that degree distribution. Hence, statements about such random graphs are statements about typical graphs with the same degree distribution—unless and until we know more about the empirical graphs.

Networks

As noted earlier in this introduction, we consider the terms 'network' and 'graph' to be interchangeable. Nevertheless there is a bit more to be said about the term. The 'network' concept motivates and infuses a vigorous and lively research activity that has more or less exploded since (roughly) the work of Watts and Strogatz [24,25]. Much of their work was motivated by the 'small worlds problem'. The latter dates back to the work of Milgram [18] (and even earlier). Milgram posed the question: how far is it from Kansas (or Nebraska) to Boston, Massachusetts—when the distance is measured in 'hops' between people who know one another? Modern language would rephrase this question as follows: is the US acquaintanceship network a 'small world'? Milgram's answer was 'yes': after disregarding the chains (of letters—that was the mechanism for a 'hop') that never reached the target, the average path length was roughly 5–6 hops—a 'small world'.

The explosion of interest in the last 10 years is well documented in the set of references in "Books and Reviews" (below).

We also include in this part an introductory discussion of directed graphs. Directed graphs have directed links: it no longer suffices to say "i and j are linked", because there is a *directionality* in the linking: $(i \to j)$ or $(j \to i)$ (or both). Some of the mathematical background for understanding directed graphs is provided in the article in this Section by Berman and Shaked-Monderer (► Non-negative Matrices and Digraphs). One quickly finds, upon coming in contact with directed graphs, that one is in a rather different world—the mathematics has changed, the structures are different, and one's intuition often fails.

Directed graphs are however here to stay, and well worth study. We cite two classic examples to demonstrate the extreme relevance of directed graphs. First, there is early and pioneering work by Stuart Kauffman [14,15] on genetic regulatory networks. These form directed graphs, because the links express the fact that gene $G1$ regulates gene $G2$ $(G1 \to G2)$—a relationship which is by no means symmetric. Understanding gene regulation and expression is a fundamental problem in biology—and we see that the problem may be usefully expressed as one of understanding *dynamics on a directed graph*. The article in this Section by Huang and Kauffman (► Complex Gene Regulatory Networks – from Structure to Biological Observables: Cell Fate Determination) brings us up to date on this exciting problem.

A more well known directed graph, which plays a role for many of us in our daily lives, is the *Web graph* [1,2,8]. We navigate among Web pages using hyperlinks—one-way pointers taking us (e. g.) from page A to page B $(A \to B)$. The utility of the Web as a source of information—and as a platform for interaction—is enormous. Hence the Web is well worth intense study, both intellectually and practically.

Hyperlinks are useful, not only for navigation, but also to aid in *ranking* Web pages. This utility was clearly pointed out in two early papers on Web link analysis [7,17]. "Web link analysis" may be viewed most simply as a process which takes (all or part of) the Web graph as input, and which gives 'importance scores' for each Web page as output. The PageRank approach to link analysis of Brin and Page [7] has almost certainly played a significant role in the meteoric rise of the Web search company Google. Hence there is both practical and commercial value in understanding the Web graph. We have two complementary papers in this Section which treat this important area: that by Adamic on ► World Wide Web,

Graph Structure, and that by Bjelland et al. on ► Link Analysis and Web Search.

Complex Networks

Now we come to the last substantial word in the title of this Section: 'complex'. This is a word that (as yet) lacks a precise scientific meaning. There are in fact too many definitions; in other words, there is no generally-agreed precise definition. For a good overview of work on this thorny problem we refer the reader to [4].

Fortunately, for the purposes of this Section, we need only a very simple definition: 'complex networks' are those which are *not* well modeled by classical random graphs (described above, and in the article by Fan Chung). The use of the term 'complex networks' is very widespread—while examples which give a clear definition are less common. We note that the definition given here is also cited by Dorogovtsev in his article in this Section (► Growth Models for Networks).

Structure of Networks

We have noted already that there is no single, simple answer to the question "what is the structure of this graph"? With this caveat in mind, we offer the reader a guide to the articles in this Section which address the structure of networks.

Undirected Graphs

We have already noted the article by Fan Chung (► Random Graphs, A Whirlwind Tour of), giving an up-to-date overview of the properties of random graphs with a given node degree distribution—with the well-studied case being power-law graphs. She cites a number of experimental results which indicate that the experimental exponents β (taken from the power-law degree distributions) fall in one range ($2 < \beta < 3$) for social and technological networks, and another, rather distinct range ($\beta < 2$) for biological networks. This may be regarded as a (quantitative) structural difference between these two types of network; Chung offers an explanation based on qualitatively distinct growth mechanisms (see the next part of this article).

Fortunato and Castellano (► Community Structure in Graphs) offer an excellent overview of another broad approach to network structure. Here the idea is to understand structure in terms of substructure (and sub-substructure, etc). That is: here is a network. Can we identify subgraphs—possibly overlapping, possibly disjoint—of this network that in some sense "belong together"? In other words: how can one identify the *community structure* of a graph? The review of Fortunato and Castellano is on the one hand very thorough—and on the other hand makes clear that there is no one agreed answer to this question. There are indeed almost as many answers as there are theoretical approaches; and this problem has received a lot of attention. Fortunato and Castellano note that there is currently a "favorite" approach, defined by finding subgraphs with high *modularity* [21]. Roughly speaking, a subgraph with high modularity has a higher density of internal links than that found for the same subgraph in a randomized 'null model' for the same graph. Fortunato and Castellano give a careful discussion of the strengths and weaknesses of this approach to community detection, as well as of many others.

Since the work of Watts and Strogatz [24,25], the definition of a 'small-world graph' has included two criteria. First, one must have the short average path length which Milgram found in his experiments, and which gives rise to the term 'small worlds'. However, one also insists that a 'true' small-world graph should locally resemble a real social network, in that there is a significant degree of *clustering*. Roughly speaking, this means that the probability that two of my acquaintances know each other is higher than random. More mathematically, to say that a graph G has high clustering means that the incidence of *closed triangles* of links in G is higher than expected in a randomized 'null model' for G.

A closed triangle is a small, simply defined subgraph, and in studying clustering one studies the statistics of the occurrence of this small subgraph. A more general term for this type of small subgraph is a *motif* [19]. Much as with clustering and triangles, one defines a set of motifs, and then generates a *significance profile* for any given network, comparing the frequency of each motif in the profile to that of a corresponding random graph. Valverde and Solé (► Motifs in Graphs) offer a stimulating overview of the study of motifs in networks of many types—both directed and undirected. They point out a remarkable consistency of motif significance profiles across networks with very different origins—for example, a software graph and the gene network of a bacterium—and argue that this consistency is best understood in terms of historical accident, rather than in terms of functionality.

The article by Liljeros (► Human Sexual Networks) looks at empirical human sexual networks. He addresses the evidence for and against the notion that such networks are power-law (also known as "scale free"). This question is important for understanding epidemic spreading over networks—especially in the light of the results of Pastor-Satorras and Vespignani [23], which showed that epidemic spreading on power-law networks is more difficult to stop than was predicted by earlier models using

the well-mixed (all-to-all) approximation. Liljeros examines carefully what is known about the structure of human sexual networks, noting the great difficulty inherent in gathering extensive and/or reliable data.

The article by He, Siganos, and Faloutsos (▶ Internet Topology) looks at a very different empirical network, namely the physical Internet, composed (at the lowest level) of routers and cables. The naïve newcomer might assume that the Internet, being an engineered system, is fully mapped out, and hence its topology should be readily "understood". The article by He et al. presents the reality of the Internet: it is largely self-organized (especially at the higher level of organization, the 'Autonomous System' or AS level); it is far from trivial to experimentally map out this network—even at the AS level; and there is not even agreement on whether or not the AS-graph is a power-law graph—which is after all a rather crude measure of structure of the network. He et al. describe recent work which offers a neat resolution of the conflicting and ambiguous data on this question. They then go on to describe more imaginative models for the structure of the Internet, going beyond simply the degree distribution, and having names like 'Jellyfish' and 'Medusa'.

Directed Graphs

A generic directed graph is immediately distinguished from its undirected counterparts in that a natural unit of substructure is obvious, and virtually always present. That is the 'strongly connected component' or SCC (termed 'class' by Berman and Shaked-Monderer). That is, even when a directed graph is connected, there are node pairs which cannot reach one another by following directed paths. An SCC C is then a maximal set of nodes satisfying the constraint that every node in C is reachable, via a directed path, from every other node in C. The SCCs then form disjoint sets (equivalence classes), and every node is in one SCC. In short, the very notion of 'reachability' is more problematic in directed graphs: all nodes in the same SCC can reach one another, but otherwise, all bets are off!

Lada Adamic (▶ World Wide Web, Graph Structure) gives a good overview of what is known empirically about the structure of the *Web graph*—that abstract and yet very real object in which a node is a Web page, and a link is a (one-way) hyperlink. The Web graph is highly dynamic, in at least two ways: pages have a finite lifetime, with new ones appearing while old ones disappear; also, many Web pages are dynamic, in that they generate new content when accessed—and can in principle represent an infinite amount of content. Also, of course, the Web is huge. Lada Adamic presents the problems associated with '*crawling*' the Web to map out its topology. The reader will perhaps not be surprised to learn that the Web graph obeys a power law—both for the indegree distribution and for the outdegree distribution. Adamic discusses several other measures for the structure of the Web graph, including its gross SCC structure (the 'bow tie'), its diameter, reciprocity, clustering, and motifs. The problem of path lengths and diameter is less straightforward for directed graphs, since the unreachability problem produces infinite or undefined path length for many pairs.

A nice bonus in the article by Adamic is the discussion of some popular and well studied *subgraphs* of the Web graph—query connection graphs, Weblogs, and Wikipedia. Query connection graphs are subgraphs built from a hit list, and can give useful information about the query itself, and the likelihood that the hit list will be satisfactory. Weblogs and Wikipedia are well known to most Web-literate people; and it is fascinating to see these daily phenomena subjected to careful scientific analysis.

The article by Bjelland et al. (▶ Link Analysis and Web Search) has perhaps the closest ties to the mathematical presentation of Berman and Shaked-Monderer (▶ Non-negative Matrices and Digraphs). This is because Web link analysis tends to focus on the *principal eigenvector* of the graph's adjacency matrix (or of some modification of the adjacency matrix), while Berman and Shaked-Monderer discuss in some detail the spectral properties of this matrix, and give some results about the principal eigenvector. The latter yields *importance* or *authority scores* for each page. These scores are the principal output of Web link analysis; and in fact Berman and Shaked-Monderer cite PageRank as an outstanding example of an application of the theory [22]. Bjelland et al. explain the logic leading to the choice of the principal eigenvector, as a way of 'harvesting' information from the huge set of 'collective recommendations' that the hyperlinks constitute. They also present the principal approaches to link analysis (the 'big three'), and place them in a simple logical framework which is completed by the arrival of a new, fourth approach. In addition, Bjelland et al. discuss a number of technical issues related to Web link analysis—which is, after all, a practical and commercial field of endeavor, as well as an object for research and understanding.

Jennifer Dunne (▶ Food Webs) presents a rather special type of directed graph taken from biology: the food web. She offers a very concise definition of this concept in her glossary, which we reproduce here: "the network of feeding interactions among diverse co-occurring species in a particular habitat". We note that most feeding interactions are one-way: bird B eats insect I, but insect I does not eat bird B. However, food webs are rather special among

empirical directed graphs, in that they have a lower incidence of loops than that found in typical directed graphs. Early work indicated (or even assumed) that a food web is loop-free; but this has been shown not to be strictly true. (A simple, but real example of a loop is cannibalism: A eats A; but much longer loops also exist.) The field faces (as many do) a real problem in getting good empirical data, and empirically obtained food webs tend to be small. For example, Table 1 of Dunne presents data for 16 empirical food webs, ranging in size from 25 nodes to 172.

Our second biological application of directed graphs is presented by Huang and Kauffman (► Complex Gene Regulatory Networks – from Structure to Biological Observables: Cell Fate Determination). This article presents gene regulatory networks (GRN). The directed link ($G1 \rightarrow G2$) in a GRN expresses the fact that gene $G1$ regulates (by inhibition or activation) the expression of gene $G2$, via intermediate proteins (typically transcription factors). The links are thus inherently one way, although reciprocal links ($G1 \leftrightarrow G2$) do occur. The article of Huang and Kauffman is very comprehensive: the structure of a GRN is only their starting point, as they seek to understand and model the *dynamical development process*, in which the set of on/off states of the cell genome moves towards an *attractor* (a steady or quasi-steady state), which represents a mature and stable cell type.

Returning to structure (of GRNs), we again face a severe data-extraction problem, which is compounded by the fact that the 'interaction modality' (activation or inhibition, plus or minus) of each link must also be known before one can study dynamical development of gene expression over the GRN. In short: one needs to know more than just the presence and direction of the links; one needs their type. Huang and Kauffman give a thorough discussion of these problems, and argue for studies using an "ensemble approach". This is much like the random graph approach, in that it takes a set of statistical constraints which are empirically determined, and then studies (often via simulation) a randomly generated ensemble of graphs which satisfy the statistical constraints. In short, the ensemble approach takes the structure as determined by these constraints (node degree distribution, etc), and then studies random graphs with this structure. Note that both the graph topology and the interaction modalities of the links are randomized in this ensemble approach.

Dynamical Network Structures

We have already had a lot to say about the structure of networks—and yet we have left out one important dimension, namely *time*. Empirical networks change over time, so that most measurements that map out such a network are taking a snapshot. Now we will look explicitly at studies addressing the dynamical evolution of networks.

One classic study is the paper by Barabasi and Albert [5]. Here the preferential-attachment (or "rich get richer") model was introduced as an explanation for the ubiquitous power-law degree distributions. In other words, a *growth* (developmental) model was used to explain properties of *snapshots* of "mature" networks. In the preferential-attachment model, new nodes which join a network link to existing nodes with a biased probability distribution—so that the probability of linking to an existing node of degree k is proportional to k.

This simple model indeed gives (after long time) a power-law distribution; and the ideas in [5] stimulated a great interest in various growth models for networks. We are fortunate to have, in this Section of the Encyclopedia, the article by Sergey Dorogovtsev entitled ► Growth Models for Networks. This article is of course very short compared to the volume [9]—but it offers a good overview of the field, and a thorough updating of that volume, covering a broad range of questions and ideas, including the simple linear preferential attachment model, and numerous variations of it. Also, a distinctly different class of growth models, termed "optimization based models", is presented, and compared to the class of models involving forms of preferential attachment. In optimization based models, new nodes place links so as to optimize some function of the resulting network properties.

We also mention, in the context of growth models, the article by Fan Chung (► Random Graphs, A Whirlwind Tour of). As noted above, she has pointed out a tendency for biological networks to have significantly smaller exponents than technological networks; and she includes a good discussion of both preferential attachment (which tends to give the larger, technological, exponents) and duplication models. The latter involve new nodes "copying" or duplicating (probabilistically) the links of existing nodes. Chung observes that such duplication mechanisms do exist in biology, and also shows that they can give exponents in the appropriate range.

Huang and Kauffman (► Complex Gene Regulatory Networks – from Structure to Biological Observables: Cell Fate Determination) also offer a limited discussion of evolution of the genome—not so much as a topic in itself, but because (again) understanding genome evolution can help in understanding—and even determining—the structure of today's genome. They mention both duplication and preferential attachment. Also, it is quite interesting to note the parallels between their discussion and that of Valverde and Solé (► Motifs in Graphs)—in that Huang and Kauff-

man also argue that many observed structures may be due, not to selection pressure or enhanced functionality, but simply to physical constraints on the evolutionary process, and/or to historical accident. The duplication mechanism in fact turns up again in the article by Valverde and Solé, who argue that this growth mechanism can largely account for the high frequency of occurrence of some motifs.

Finally, we note that growth is not the only dynamical process operative in networks. Just as the mature body is constantly shedding and regenerating cells, many mature networks are subject to constant small topology changes. One interesting class of non-growth dynamics is the *attack*. That is: how robust is a given network against a systematic attack which deletes nodes and/or links? A classic study in this direction is [3], which studied the attack tolerance of scale-free networks. There are many reasons to believe that such networks are very well connected, and this study added more: attacking random nodes had little effect on the functionality of the network (as crudely measured by the network diameter, and by the size of the largest surviving component). Thus such networks may be termed "robust"—but only with regard to this kind of "uninformed" attack. The same study showed that a "smart" attack, removing the nodes in descending order of node degree (i. e., highest degree first), caused the network to break down much more rapidly—thus highlighting once again the crucial role played by the 'hubs' in power-law networks.

Jennifer Dunne (► Food Webs) reports some studies of the robustness of food webs to attack—where here the 'attack' is extinction of a node (species). The dynamics is slightly different from that in the previous paragraph, however, because of the phenomenon of secondary extinction: removal of one species can cause another (or several others) to go extinct as well, if they are dependent on the first for their food supply. Also, food webs are not well modeled by power-law degree distributions. Nevertheless, the cited studies indicate that removing high-degree species again causes considerably more damage (as measured by secondary extinctions and web fragmentation) than removing low-degree species. Dunne also reports studies seeking to simulate the effects of "ecologically plausible" extinction scenarios; here we find that the studied food webs are in fact very robust to such extinction scenarios—a result which (perhaps) confirms our prejudice, that today's ecosystems are here because they are robust.

Dynamical Processes on Networks

Our next theme is also about dynamics—not *of* the network topology, however, but *over* it. That is, it is often of interest to study *processes* which occur on the nodes (changing their state), and which transmit something (information) over the links, so that a change in one node's state induces a change in other nodes' states.

We can render these ideas less abstract by considering the concrete example of epidemic spreading. The nodes can be individuals (or computers, or mobile phones), and the network is then a network of social contacts (or a computer or phone network). The elementary point that the disease is spread via contact is readily captured by the network model. A classic study in this regard, which strongly underscored the value of network models, was that (mentioned earlier) of Pastor-Satorras and Vespignani [23]. Here it was shown that classical threshold thinking about epidemic spreading fails when the network is scale-free: the effective threshold is *zero*. This study, in yet another way, revealed that such networks are extremely well connected—and it stimulated a great deal of thought about prevention strategies.

We have already mentioned the article by Liljeros (► Human Sexual Networks) in this Section, with its evidence for power-law, or nearly power-law, degree distributions in human sexual networks. Liljeros offers a sober and careful discussion of the implications of the theoretical result just cited, for understanding and preventing the spread of sexually transmitted diseases on finite networks. One consequence of a power-law node degree distribution—that a few individuals will have an *extremely* high number of contacts—may seem at first glance implausible or even impossible; and yet, as Liljeros points out, careful empirical studies have tended to support this prediction.

The human sexual network is of course not static; people change partners, and in fact people with many partners tend to change more often. Hence, the dynamics of the network topology must be considered, including how it affects the dynamics of the epidemic spreading process going on over the topology. A term which captures this interplay is "concurrence"—which tells, not how many partners you have, but how many partners you have "simultaneously"—i. e., within a time window which allows the passing of the disease from one contact to another. Considering concurrence leads to another type of graph, termed a 'line graph'—a structure in which contact relationships become nodes, which are linked when they are concurrent.

The reader will not be surprised to hear that epidemic spreading over directed graphs is qualitatively different from the same process over undirected graphs. The topic has real practical interest, because—as pointed out by Kephart and White [16] and by Newman, Forrest and Balthrop [20]—the effective network over which computer

viruses propagate is typically directed. We invite the interested reader to consult these sources—and to note the striking similarities between the picture of the 'email address graph' in [20] and the gross structure of the Web graph [8] (and Figure 1 of Adamic).

Another classic study of dynamical processes on networks is that by Duncan Watts [24,25] on synchronization phenomena over networks. Our Section includes a thorough and up-to-date survey of this problem, primarily from the theoretical side, by Chen et al. (► Synchronization Phenomena on Networks). The field is large, as there exists a wide variety of dynamical processes on nodes, and types of inter-node coupling, which may (or may not) lead to synchronization over the network. Chen et al. focus on three main themes which neatly summarize the field. First, there is the synchronization process itself, and the theory which allows one to predict when synchronization will occur. Next comes the question of how the network structure affects the tendency to synchronization. The third theme is a logical followup to the second: if we know something about how structure affects synchronization, can we not find design methods which enhance the tendency to synchronize? This latter theme includes a number of ingenious methods for 'rewiring' the network in order to enhance its synchronizability. For example: it is found that a very strong community structure (recall the article by Castellano and Fortunato) will inhibit synchronization; and so one has studied networks called *'entangled networks'*, which are systematically rewired so as to have essentially no community structure, and so are optimally synchronizable.

Dunne (► Food Webs) discusses dynamics of species number over food webs. That is, the links telling "who eats who" also mediate transfers of biomass—a simple dynamical process. And yet—as we know from the dynamics of the simple two-species Lotka–Volterra equations [13]—simple nonlinear rules can give rise to complex behavior. Dunne reports that the behavior does not get more simple as one studies larger networks; one finds the same types of asymptotic behavior—equilibrium, limit cycle, and chaotic dynamics. It is a nontrivial task to study nonlinear dynamical models over tens or hundreds of nodes, and at the same time anchor the theory and simulation in reality. One approach to doing this has been to insist that the resulting model satisfy certain stability criteria (hence conforming to reality!); these criteria are framed in terms of 'species persistence' (not too many extinctions) and/or 'population stability' (limited fluctuations in species mass for all species).

A dynamical process also plays a central role in the discussion of Huang and Kauffman (► Complex Gene Regulatory Networks – from Structure to Biological Observables: Cell Fate Determination). In a simplified but highly nontrivial model, the state of the gene network is modeled by a vector $\boldsymbol{S}(t)$, which takes binary values (0 or 1, i. e., 'off' or 'on') at each node in the network, at each time t. Also, the interactions between genes are modeled by Boolean truth tables—i. e., a given gene's (binary) output state is some Boolean function of all of its input states. The resulting 'Boolean network model' is at once very simple (discrete time, discrete binary states), and yet very complex, in the sense that it is impossible to predict the behavior of such models without simulating them. Huang and Kauffman describe the three regimes of dynamical behavior found: an 'ordered' regime, a 'chaotic' regime, and an intermediate 'critical' regime. While the ordered regime gives stable behavior, Huang and Kauffman argue that biology favors the critical regime. One gets the flavor of their argument if one recalls that the same genome must be able to converge to many different cell types—so its dynamics cannot be *too* stable—and yet those same cell types must be 'stable enough'. The discussion of the latter point includes a remarkable recent experimental study which graphically shows the return of two cell populations to the same 'preferred' stable attractor, after two distinct perturbations, via two distinct paths in state space.

Valverde and Solé (► Motifs in Graphs) also discuss dynamical processes, in terms of motifs. They give a clear picture for the simple test case of the three-node motif called 'Feed Forward Loop' or FFL. Here again—as for the GRNs of Huang and Kauffman—the modality (activation/inhibition) of the three links must be considered, and the resulting possibilities are either 'coherent' (non-conflicting) or 'incoherent'. It is interesting to note that the FFL motif has been studied both via simulation, and experimentally, in real gene networks. Valverde and Solé also make contact with the article of Chen et al. (► Synchronization Phenomena on Networks) by briefly discussing the connection between network synchronizability and the distribution of motif types.

Graph Visualization

The problem of graph visualization is a marvelous mixture of art and science. To produce a good graph visualization is to translate the science—what is known factually and analytically about the graph—into a 2D (or quasi-3D) image that the human brain can appreciate. Here the word 'appreciate' can include both understanding and the experience of beauty. Both of these aspects are present in great quantities in the many figures of this Section. We invite the reader interested in visualization to visit and compare

the following figures: Figure 2 in Fan Chung's article; Figures 1 and 3 in Dunne; Figures 2 and 5 in Liljeros; Figure 17 in Chen et al.; and Figures 1, 2, and 6b in Valverde and Solé.

Graph visualization is included in the articles of this Section because it is a vital tool in the study of networks, and also an unfinished research challenge. The article by Vladimir Batagelj (► Complex Networks, Visualization of) offers a fine overview of this large field. This article in fact has a very broad scope, touching upon many visualization problems (molecules, Google maps) which may be called "information visualization". The connection to networks (no pun intended) is that information is readily understood in terms of connections between units—i. e., as a network.

We note that one natural way of 'understanding' a graph is in terms of its *communities*; and since community substructure offers a way (perhaps even hierarchical) of 'coarse-graining' the network, methods for defining communities often lead naturally to methods for graph visualization. One example of this may be found in Figure 10 of Fortunato and Castellano (► Community Structure in Graphs); for other examples, see [12] and [6]. Batagelj describes a wide variety of graph visualization methods; but a careful reading of his article shows that many methods which are useful for large graphs are dependent on finding and exploiting some kind of community structure. The terms used for approaches in this vein include 'multilevel algorithms', 'clustering', 'block modeling', coarse graining, partitions, and hierarchies.

The lessons learned from Fortunato and Castellano are brought home again in the article by Batagelj: there is no 'magic bullet' that gives a universally satisfying answer to the problem of visualizing large networks. To quote Batagelj in regard to an early graph visualization: "Nice as a piece of art, but with an important message: there are big problems with visualization of dense and/or large networks." A more recent example is a 2008 visualization of the Internet: here there is a 'natural' unit of coarse graining, namely the autonomous system or AS level (recall the article by He et al. on Internet Topology), so that the network of almost 5 million nodes reduces to 'only' about 18 000 ASes. Yet the resulting visualization (Figure 4 of Batagelj) clearly reveals that 18 000 nodes is still 'large' relative to human visual processing capacity.

For other beautiful and mystifying examples of this same point, I recommend figures 5 and 7 in the same article. The difficulty of visualizing large networks is perhaps most succinctly captured in the outline of Batagelj's stimulating article. Besides the normal introductory parts, we find only two others: 'Attempts' and 'Perspectives'. Yet the article is by no means discouraging—it is rather fascinating and inspiring, and I encourage the reader to read and enjoy it.

Future Directions

The many articles related to "Complex Networks and Graph Theory" have each offered their own view of 'Future directions' for the corresponding field of study. Therefore it would be both redundant and presumptuous for me to attempt the same task for all of these fields. Instead I will offer a very short and entirely personal assessment of the 'Future' for the broad field of complex networks and graph theory.

I view the field somewhat as a living organism: it is highly dynamic, and still growing vigorously. New ideas continue to pop up—many of which could not be covered in this Section, simply due to practical limitations. Also, there is a fairly free flow of ideas across disciplines. The reader has perhaps already gotten a feeling for this cross-boundary flow, by seeing the same basic ideas crop up in many articles, on problems coming for distinctly different traditional disciplines. In short, I feel that the interdisciplinarity of this field is real, it is vigorous and healthy, and it is exciting.

Another aspect contributing to the excellent health of the field is its strong connection to empiricism. Network studies thrive on getting access to real, empirically-obtained graphs—we have seen this over and over again in discussion of the articles in this Section. From the Web graph with its tens of billions of nodes, to food webs with perhaps a hundred nodes, the science of networks is stimulated, challenged, and enriched by a steady influx of new data.

Finally, the study of complex networks is eminently practical. The path to direct application can be very short. Again we cite the Web graph and Google's PageRank algorithm as an example of this. The study of gene networks is somewhat farther from immediate application; but the possible benefits from a real understanding of cell and organism development can be enormous. The same holds for the problem of epidemic spreading. These examples are only picked out to illustrate the point; all of the articles, and topics represented by them, are not far removed from practical application.

In short: the field is exciting, vigorous, and interdisciplinary, and offers great practical benefits to society. The study of graphs and networks shows no signs of becoming moribund. Hence I will hazard a guess about the future: that the field will continue to grow and inspire excitement for many years to come. It is my hope that many readers

of this Section will be infected by this excitement, and will choose to join in the fun.

Bibliography

Primary Literature

1. Adamic LA (1999) The Small World Web. In: Proc 3rd European Conf Research and Advanced Technology for Digital Libraries, ECDL, London, pp 443–452
2. Albert R, Jeong H, Barabasi AL (1999) Diameter of World-Wide Web. Nature 410:130–131
3. Albert R, Jeong H, Barabasi AL (2000) Error and attack tolerance of complex networks. Nature 406:378–382
4. Badii R, Politi A (1997) Complexity: Hierarchical Structures and Scaling in Physics. Cambridge Nonlinear Science Series, vol 6. Cambridge University Press, Cambridge
5. Barabasi AL, Albert R (1999) Emergence of Scaling in Random Networks. Science 286:509–512
6. Bjelland J, Canright G, Engø-Monsen K, Remple VP (2008) Topographic Spreading Analysis of an Empirical Sex Workers' Network. In: Ganguly N, Mukherjee A, Deutsch A (eds) Dynamics on and of Complex Networks. Birkhauser, Basel,also in http://delis.upb.de/paper/DELIS-TR-0634.pdf
7. Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine. In: Proceedings of the seventh international conference on World Wide Web, pp 107–117
8. Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata S, Tomkins A, Wiener J (2000) Graph structure in the web. Comput Netw 33:309–320
9. Dorogovtsev SN, Mendes JFF (2003) Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford University Press, Oxford
10. Erdős P (1947) Some remarks on the theory of graphs. Bull Amer Math Soc 53:292–294
11. Erdős P, Rényi A (1959) On random graphs, I. Publ Math Debrecen 6:290–297
12. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA 99:8271–8276
13. Kaplan D, Glass L (1995) Understanding Nonlinear Dynamics. Springer, New York
14. Kauffman S (1969) Homeostasis and differentiation in random genetic control networks. Nature 224:177–8
15. Kauffman SA (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor Biol 22:437–467
16. Kephart JO, White SR (1991) Directed-graph epidemiological models of computer viruses. In: Proceedings of the 1991 IEEE Computer Society Symposium on Research in Security and Privacy, pp 343–359
17. Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5):604–632
18. Milgram S (1967) The Small World Problem. Psychol Today 2:60–67
19. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network Motifs: Simple Building Blocks of Complex Networks. Science 298:824–827
20. Newman MEJ, Forrest S, Balthrop J (2002) Email networks and the spread of computer viruses. Phys Rev E 66:35–101
21. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69:026113
22. Page L, Brin S, Motwani R, Winograd T (1998) The PageRank citation ranking: Bringing order to the web. Technical report. Stanford University, Stanford
23. Pastor-Satorras R, Vespignani A (2001) Epidemic spreading in scale-free networks. Phys Rev Lett 86:3200–3203
24. Watts DJ (1999) Small Worlds: The Dynamics of Networks between Order and Randomness. Princeton Studies in Complexity. Princeton University Press, Princeton
25. Watts D, Strogatz S (1998) Collective dynamics of 'small world' networks. Nature 393:440–442

Books and Reviews

Albert R, Barabasi AL (2002) Statistical Mechanics of Complex Networks. Rev Mod Phys 74:47–97

Bornholdt S, Schuster HG (2003) Handbook of Graphs and Networks: From the Genome to the Internet. Wiley-VCH, Berlin

Caldarelli G, Vespignani A (2007) Large Scale Structure and Dynamics of Complex Networks: From Information Technology to Finance and Natural Science. Cambridge University Press, Cambridge

Chung F, Lu L (2006) Complex Graphs and Networks, CBMS Regional Conference. Series in Mathematics, vol 107. AMS, Providence

da F Costa L, Osvaldo N, Oliveira J, Travierso G, Rodrigues FA, Paulino R, Boas V, Antiqueira L, Viana MP, da Rocha LEC Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications. Working Paper: http://arxiv.org/abs/0711.3199

Kauffman SA (1993) The origins of order. Oxford University Press, Oxford

Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45:167–256

Newman M, Barabasi A, Watts DJ (2006) The Structure and Dynamics of Networks. Princeton Studies in Complexity. Princeton University Press, Princeton

Complex Networks, Visualization of

Vladimir Batagelj
University of Ljubljana, Ljubljana, Slovenia

Article Outline

Glossary
Definition of the Subject
Introduction
Attempts
Perspectives
Bibliography

Glossary

For basic notions on graphs and networks, see the articles by Wouter de Nooy: ► Social Network Analysis, Graph Theoretical Approaches to and by Vladimir

Batagelj: ► Social Network Analysis, Large-Scale in the Social Networks Section. For complementary information on graph drawing in social network analysis, see the article by Linton Freeman: ► Social Network Visualization, Methods of.

***k*-core** A set of vertices in a graph is a *k*-core if each vertex from the set has an internal (restricted to the set) degree of at least *k* and the set is maximal – no such vertex can be added to it.

Network A network consists of vertices linked by lines and additional data about vertices and/or lines. A network is large if it has at least some hundreds of vertices. Large networks can be stored in computer memory.

Partition A partition of a set is a family of its nonempty subsets such that each element of the set belongs to exactly one of the subsets. The subsets are also called classes or groups.

Spring embedder is another name for the energy minimization graph drawing method. The vertices are considered as particles with repulsive force between them, and lines as springs that attract or repel the vertices if they are too far or too close, respectively. The algorithm is a means of determining an embedding of vertices in two or three dimensional space that minimizes the 'energy' of the system.

Definition of the Subject

The earliest pictures containing graphs were magic figures, connections between different concepts (for example the Sephirot in Jewish Kabbalah), game boards (nine men's morris, pachisi, patolli, go, xiangqi, and others) road maps (for example Roman roads in Tabula Peutingeriana), and genealogical trees of important families [33].

The notion of the graph was introduced by Euler. In the eighteenth and nineteenth centuries, graphs were used mainly for solving recreational problems (Knight's tour, Eulerian and Hamiltonian problems, map coloring). At the end of the nineteenth century, some applications of graphs to real life problems appeared (electric circuits, Kirchhoff; molecular graphs, Kekulé). In the twentieth century, graph theory evolved into its own field of discrete mathematics with applications to transportation networks (road and railway systems, metro lines, bus lines), project diagrams, flowcharts of computer programs, electronic circuits, molecular graphs, etc.

In social science the use of graphs was introduced by Jacob Moreno around 1930 as a basis of his sociometric approach. In his book *Who shall survive?* [36], a relatively large network *Sociometric geography of community – map III* (435 individuals, 4350 lines) is presented. Linton Freeman wrote a detailed account of the development of social network analysis [20] and the visualization of social networks [19]. The networks studied in social network analysis until the 1990s were mostly small – some tens of vertices.

Introduction

Through the 1980s, the development of information technology (IT) laid the groundwork for the emerging field of computer graphics. During this time, the first algorithms for graph drawing appeared:

- Trees: Wetherell and Shannon [45].
- Acyclic graphs: Sugiyama [42].
- Energy minimization methods (spring embedders) for general graphs: Eades [17], Kamada and Kawai [30], Fruchterman and Reingold [21].

In energy minimization methods, vertices are considered as particles with repulsive force between them and lines as springs that attract or repel the vertices if they are too far or too close, respectively. The algorithms provide a means of determining an embedding of vertices in two or three dimensional space that minimizes the 'energy' of the system.

As early as 1963, William Tutte proposed an algorithm for drawing planar graphs [43] and Donald E. Knuth put forth an algorithm for drawing flowcharts [31].

A well known example of an early graph visualization was produced by Alden Klovdahl using his program View_Net – see Fig. 1. As nice a piece of art as it was, it held an important message: there are big problems with the visualization of dense, large graphs.

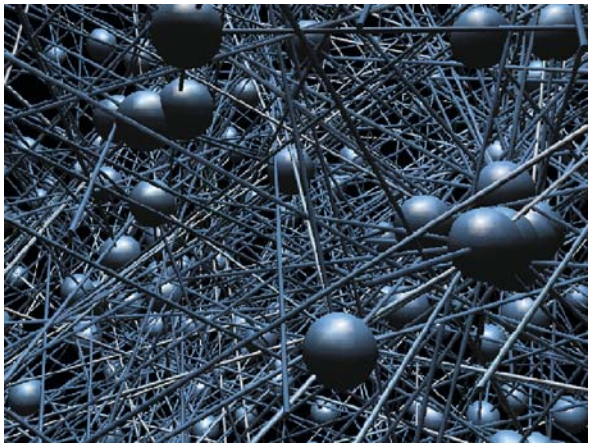

Complex Networks, Visualization of, Figure 1
Klovdahl: Social links in Canberra, Australia

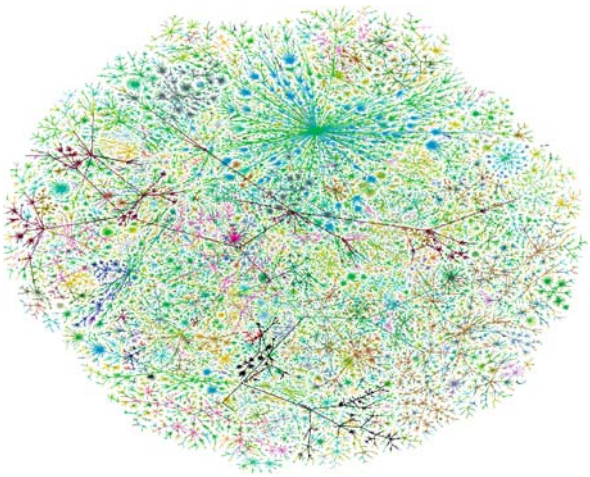

Complex Networks, Visualization of, Figure 2
Network of traceroute paths for 29 June 1999

These developments led to the emergence of a new field: graph drawing. In 1992 a group of computer scientists and mathematicians (Giuseppe Di Battista, Peter Eades, Michael Kaufmann, Pierre Rosenstiehl, Kozo Sugiyama, Roberto Tamassia, Ioannis Tollis, and others) started the conference *International Symposium on Graph Drawing* which takes place each year. The proceedings of the conference are published in the *Lecture Notes in Computer Science* series by Springer [24]. To stimulate new approaches to graph drawing, a *graph drawing contest* accompanies each conference. Many papers on graph drawing are published in the *Journal of Graph Algorithms and Applications* [79].

Most of the efforts of the graph drawing community were spent on problems of drawing special types of graphs (trees, acyclic graphs, planar graphs) or using special styles (straight lines, orthogonal, grid-based, circular, hierarchical) and deriving bounds on required space (area) for selected types of drawing.

In the 1990s, further development of IT (GUI, multimedia, WWW) made large graph analysis a reality. For example, see the studies of large organic molecules (PDB [51]), the Internet (Caida [52]), and genealogies (White and Jorion [46], FamilySearch [60]). In chemistry, several tools for dynamic three-dimensional visualization and inspection of molecules were developed (Kinemage [48], Rasmol [87], MDL Chime [82]).

One of the earliest systems for large networks was SemNet (see Fairchild, Poltrock, Furnas [18]), used to explore knowledge bases represented as directed graphs.

In 1991 Tom Sawyer Software [91] was founded – the premier provider of high performance graph visualization,

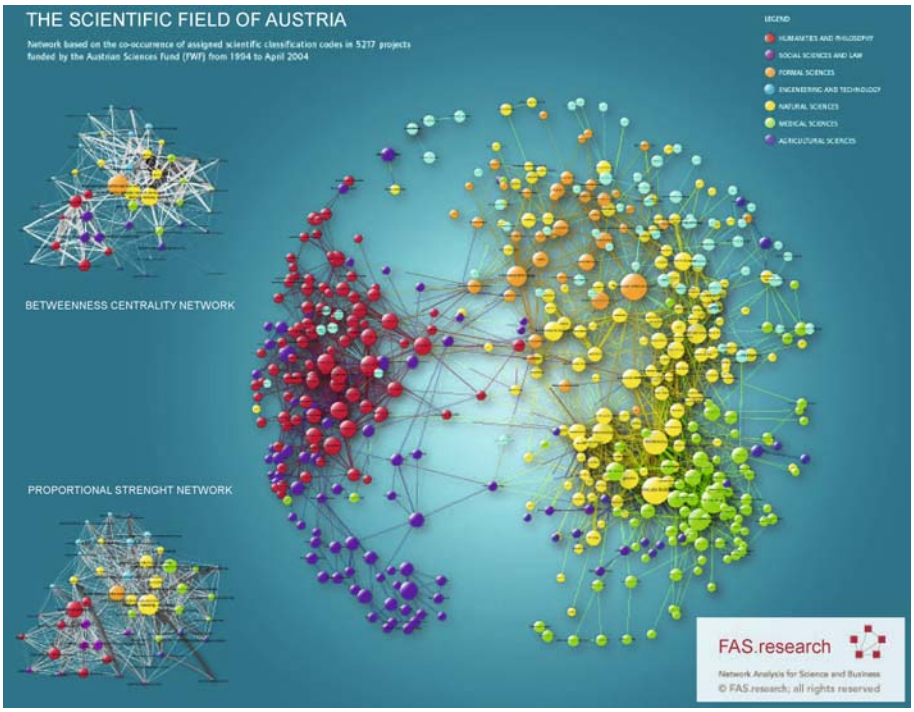

Complex Networks, Visualization of, Figure 3
FAS: The scientific field of Austria

layout, and analysis systems that enable the user to see and interpret complex information to make better decisions.

In 1993 at AT&T, the development of GraphViz tools for graph visualization began (dot, neato, dotty, tcldot, libgraph) [69]. Becker, Eick and Wilks developed a SeeXYZ family of network visualization programs [8].

In 1996 Vladimir Batagelj and Andrej Mrvar started the development of *Pajek* – a program for large network analysis and visualization [5].

In 1997 at La Sepienza, Rome, the development of GDToolkit [63] started as an extension of LEDA (Library of Efficient Data types and Algorithms) to provide implementations of several classical graph-drawing algorithms. The new version, GDT 4.0 (2007), produced in collaboration with University of Perugia, is LEDA independent.

Graham Wills, a principal investigator at Bell Labs (1992–2001), built the Nicheworks system for the visual analysis of very large weighted network graphs (up to a million vertices).

In the summer of 1998 Bill Cheswick and Hal Burch started work on the *Internet Mapping Project* at Bell Labs [53]. Its goal was to acquire and save Internet topological data over a long period of time. This data has been used in the study of routing problems and changes, distributed denial of service (DDoS) attacks, and graph theory. In the fall of 2000 Cheswick and Burch moved to a spin-off from Lucent/Bell Labs named Lumeta Corporation. Bill Cheswick is now back at AT&T Labs. Figure 2 shows a network obtained from traceroute paths for 29 June 1999 with nearly 100 000 vertices.

In the years 1997–2004, Martin Dodge maintained his *Cybergeography Research* web pages [58]. The results were published in the book *The Atlas of Cyberspace* [14]. A newer, very rich, site on information visualization is *Visual complexity* [95], where many interesting ideas on network visualizations can be found. These examples, and many others, can also be accessed from the Infovis *1100+ examples of information visualization* site [77]. An interesting collection of graph/network visualizations can be found also on the CDs of Gerhard Dirmoser. The collection also contains many artistic examples and other pictures not produced by computers.

In 1997 Harald Katzmair founded FAS research in Vienna, Austria [61], a company providing network analysis services. FAS emphasizes the importance of nice-looking final products (pictures) for customers by using graphical tools to enhance the visual quality of results obtained from network analysis tools. In Fig. 3, a network of Austrian research projects is presented. A similar company, Aguidel [49], was founded in France by Andrei Mogoutov, author of the program Réseau-Lu.

Every year (from 2002) at the INSNA Sunbelt conference [78], the *Viszards group* has a special session in which they present their solutions – analysis and visualizations of selected networks or types of networks. Most of the selected networks are large (KEDS, Internet Movie Data Base, Wikipedia, Web of Science).

Attempts

The new millennium has seen several attempts to develop programs for drawing large graphs and networks. Most of the following descriptions are taken verbatim from the programs' web pages.

Stephen Kobourov, with his collaborators from the University of Arizona, developed two graph drawing systems: GRIP (2000) [22,70] and Graphael (2003) [66]. GRIP – *Graph dRawing with Intelligent Placement* was designed for drawing large graphs and uses a multi-dimensional force-directed method together with fast energy function minimization. It employs a simple recursive coarsening scheme – rather than being placed at random, vertices are placed intelligently, several at a time, at locations close to their final positions.

The Cooperative Association for Internet Data Analysis (CAIDA) [52], co-founded in 1998 by kc claffy, is an independent research group dedicated to investigating both the practical and theoretical aspects of the Internet to promote the engineering and maintenance of a robust, scalable, global Internet infrastructure. They have been focusing primarily on understanding how the Internet is evolving, and on developing a state-of-the-art infrastructure for data measurement that can be shared with the entire research community.

Figure 4 represents a macroscopic snapshot of the Internet for two weeks: 1–17 January 2008. The graph reflects 4 853 991 observed IPv4 addresses and 5 682 419 IP links. The network is aggregated into a topology of Autonomous Systems (ASes). The abstracted graph consists of 17 791 ASes (vertices) and 50 333 peering sessions (lines).

Walrus is a tool for interactively visualizing large directed graphs in three-dimensional space. It is best suited to visualizing moderately sized graphs (a few hundred thousand vertices) that are nearly trees. Walrus uses three-dimensional hyperbolic geometry to display graphs under a fisheye-like magnifier. By bringing different parts of a graph to the magnified central region, the user can examine every part of the graph in detail. Walrus was developed by Young Hyun at CAIDA based on research by Tamara Munzner. Figure 5 presents two examples of visualizations produced with Walrus.

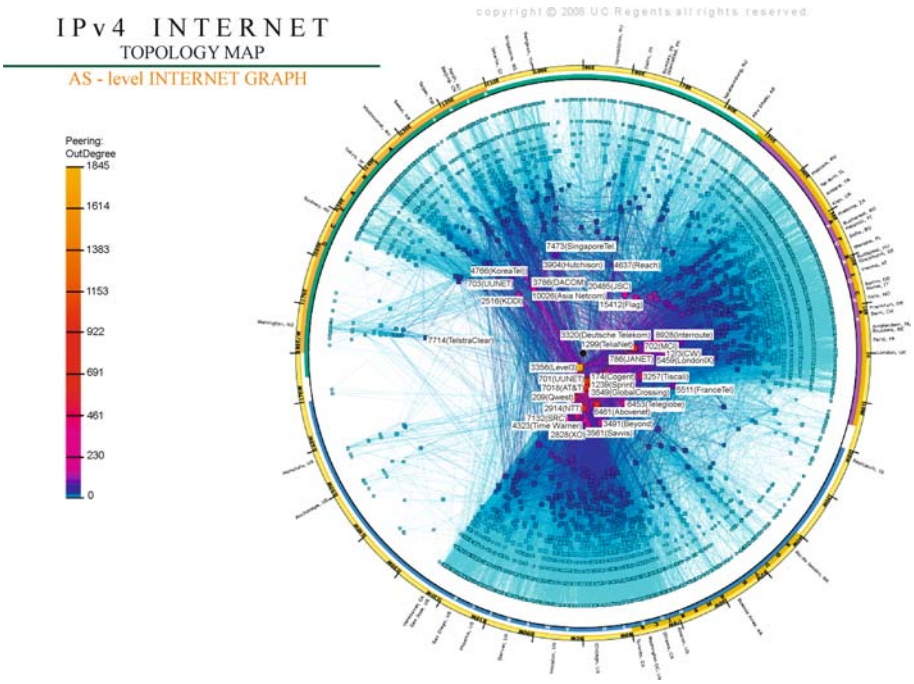

Complex Networks, Visualization of, Figure 4
CAIDA: AS core 2008

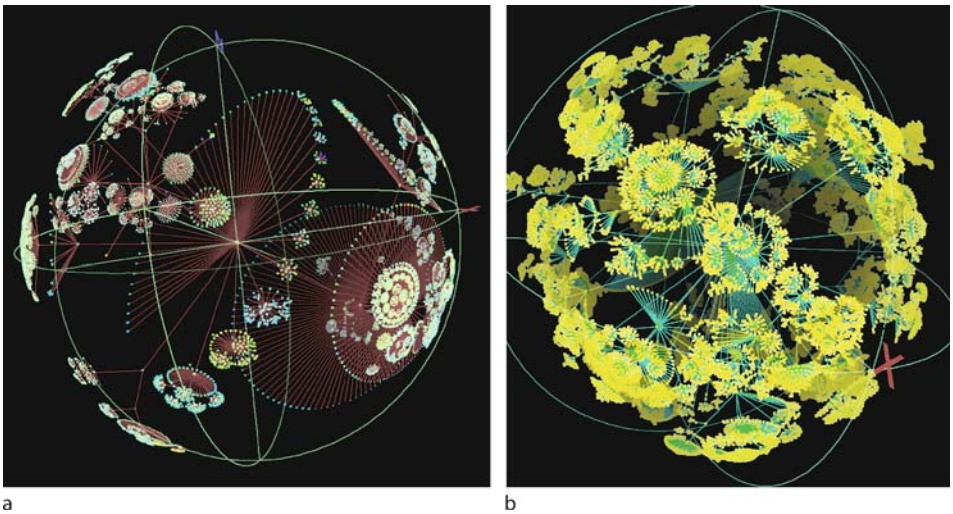

Complex Networks, Visualization of, Figure 5
Walrus

Some promising algorithms for drawing large graphs have been proposed by Ulrik Brandes, Tim Dwyer, Emden Gansner, Stefan Hachul, David Harel, Michael Jünger, Yehuda Koren, Andreas Noack, Stephen North, Christian Pich, and Chris Walshaw [11,16,23,25,26,32,39,44]. They are based either on a multilevel energy minimization approach or on an algebraic or spectral approach that reduces to some application of eigenvectors.

The multilevel approach speeds-up the algorithms. *Multilevel algorithms* are based on two phases: a *coarsening phase*, in which a sequence of coarse graphs with decreasing sizes is computed, and a *refinement phase*, in which successively finer drawings of graphs are computed, using the drawings of the next coarser graphs and a variant of a suitable force-directed single-level algorithm [25]. The fastest algorithms combine the multilevel approach

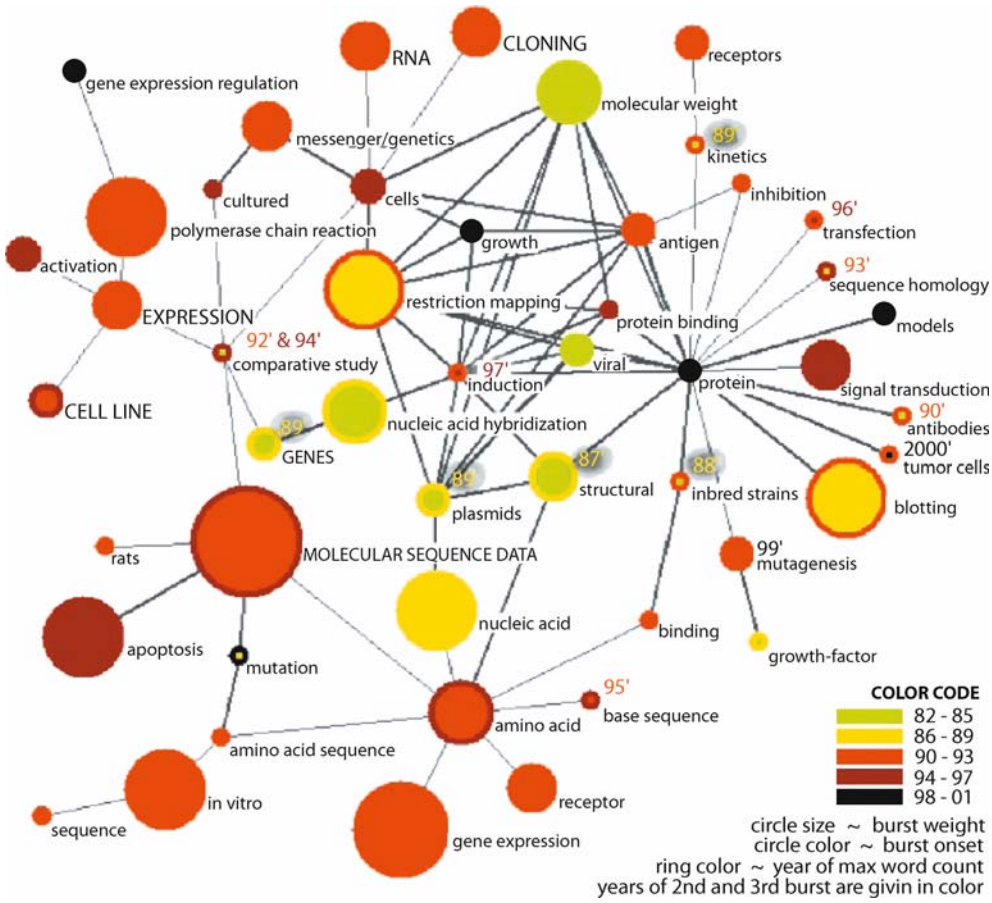

Complex Networks, Visualization of, Figure 6
Katy Börner: Text analysis

with fast approximation of long range repulsive force using nested data structures, such as quadtree or kd-tree.

Katy Börner from Indiana University, with her collaborators, produced several visualizations of scientometric networks such as Backbone of Science [9] and Wikipedia [73]. They use different visual cues to produce information-rich visualizations – see Fig. 6. She also commissioned the *Map of Science* based on data (800 000 published papers) from Thomson ISI and produced by Kevin Boyack, Richard Klavans and Bradford Paley [9].

Yifan Hu from AT&T Labs Information Visualization Group developed a multilevel graph drawing algorithm for visualization of large graphs [29]. The algorithm was first implemented in 2004 in Mathematica and released in 2005. For demonstration he applied it to the *University of Florida Sparse Matrix collection* [56] that contains over 1500 square matrices. The results are available in the *Gallery of Large Graphs* [74]. The largest graph (van-Heukelum/cage15) has 5 154 859 vertices and 47 022 346 edges. In Fig. 7, selected pictures from the gallery are presented.

From the examples that we have given, we can see that, in some cases, graph drawing algorithms can reveal symmetries in a given graph and also a 'structure' ((sub)trees, clusters, planarity, etc.). Challenges remain in devising ways to represent graphs with dense parts.

For dense parts, a better approach is to display them using matrix representation. This representation was used in 1999 by Vladimir Batagelj, Andrej Mrvar and Matjaž Zaveršnik in their partitioning approach to visualization of large graphs [7] and is a basis of systems such as *Matrix Zoom* by James Abello and Frank van Ham, 2004 [1,2], and *MatrixExplorer* by Nathalie Henry and Jean-Daniel Fekete, 2006 [27]. A matrix representation is determined by an ordering of vertices. Several algorithms exist that can produce such orderings. A comparative study of them was published by Chris Mueller [37,84]. In Fig. 8 three orderings of the same matrix are presented. Most ordering al-

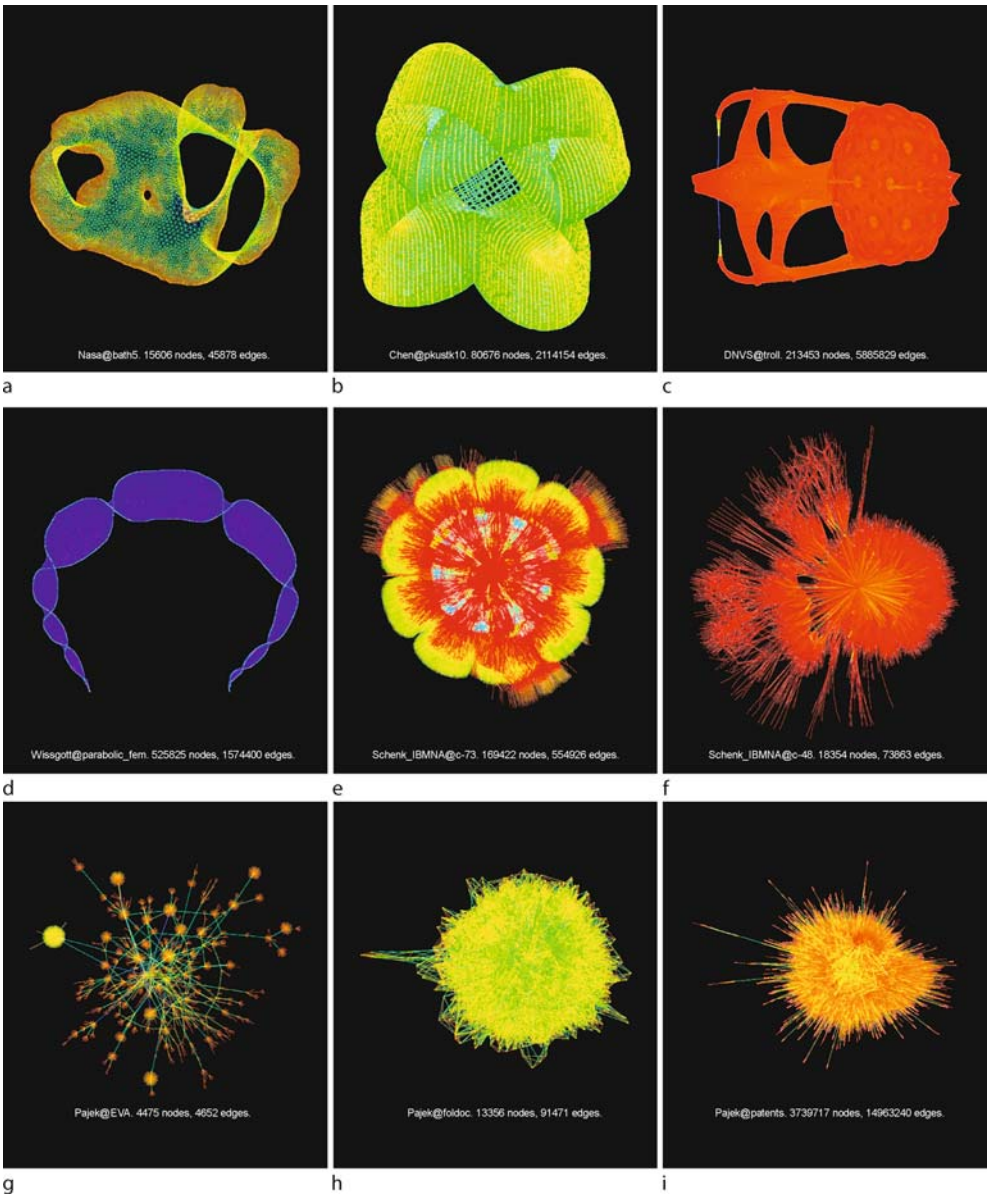

Complex Networks, Visualization of, Figure 7
Examples from the Gallery of Large Graphs

gorithms were originally designed for applications in numerical, rather than data, analysis. The orderings can also be determined using clustering or blockmodeling methods [15].

An important type of networks are *temporal networks*, where the presence of vertices and lines changes through time. Visualization of such networks requires special approaches (Sonia [88], SVGanim [90], TecFlow [75]). An interesting approach to visualization of temporal networks was developed by Ulrik Brandes and his group [12].

Perspectives

In this section, we present a collection of ideas on how to approach visualization of large networks. These ideas are only partially implemented in different visualization solutions.

While the technical problems of graph drawing strive for a single 'best' picture, network analysis is also a part of data analysis. Its goal is to gain insight not only into the structure and characteristics of a given network, but also

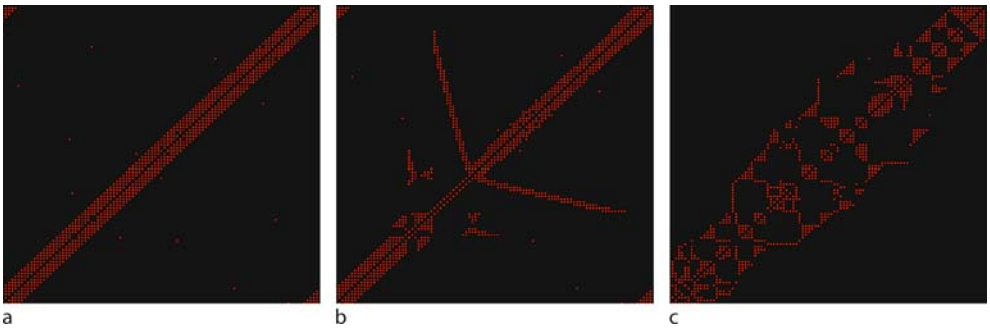

Complex Networks, Visualization of, Figure 8
Matrix representations

Complex Networks, Visualization of, Figure 9
Big picture, V. Batagelj, AE'04

into how this structure influences processes going on over the network. We usually need several pictures to present the obtained results.

Small graphs can be presented in their totality and in detail within a single view. In a comprehensive view of large graphs, details become lost – conversely a detailed view can encompass only a part of a large graph.

The literature on graph drawing is dominated by the 'sheet of paper' paradigm – the solutions and techniques are mainly based on the assumption that the final result is a static picture on a sheet of paper. In this model, to present a large data set we need a large 'sheet of paper' – but this has a limit. Figure 9 presents a visualization of a symmetrized subnetwork of 5952 words and 18 008 associations from the *Edinburgh Associative Thesaurus* [59] prepared by Vladimir Batagelj on a 3 m × 5 m 'sheet of paper' for Ars Electronica, Linz 2004 *Language of networks* exhibition.

The main tool for dealing with large objects is abstraction. In graphs, abstraction is usually realized using a hierarchy of partitions. By shrinking selected classes of a partition we obtain a smaller reduced graph. The main operations related to abstraction are:

- Cut-out: Display of only selected parts (classes of partition) of a graph;

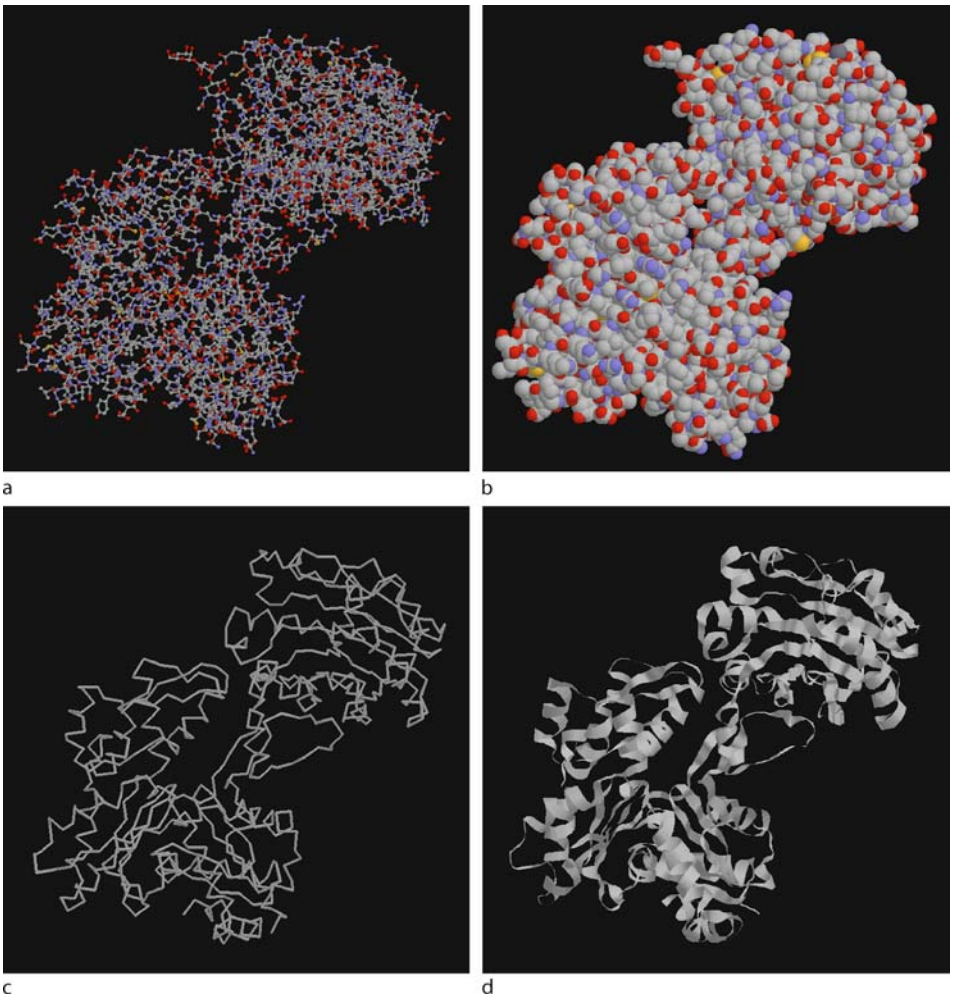

Complex Networks, Visualization of, Figure 10
Glasses: Rasmol displays – BallStick, SpaceFill, Backbone, Ribbons

- Context: Display details of selected parts (classes) of a graph and display the rest of the graph in some reduced form;
- Model: Display the reduced graph with respect to a given partition;
- Hierarchy: Display the tree representing the nesting of graph partitions.

In larger, denser networks there is often too much information to be presented at once. A possible answer is an interactive layout on a computer screen where the user controls what (s)he wants to see.

The computer screen is a medium which offers many new possibilities: parallel views (global and local); brushing and linking; zooming and panning; temporary elements (additional information about the selected elements, labels, legends, markers, etc.); highlighted selections; and others. These features can and should be maximally leveraged to support data analytic tasks; or repeating Shneiderman's mantra: overview first, zoom and filter, then details on-demand (extended with: Relate, history and extract) [40].

When interactively inspecting very large graphs, a serious problem appears: how does one avoid the "lost within the forest" effect? There are several solutions that can help the user maintain orientation:

- Restart option: Returns the user to the starting point;
- Introduction of additional orientation elements: Allows elements to be switched on and off.
- Multiview: Presents at least two views (windows):

- Map view: Shows an overall global view which contains the current position and allows 'long' moves (jumps). For very large graphs, a map view can be combined with zooming or fish-eye views.
- Local view: Displays a selected portion of the graph.

Additional support can be achieved by implementing trace, backtrack, and replay mechanisms and guided tours.

An interactive dynamic visualization of a graph on the computer screen need not be displayed in its totality. Inspecting a visualization, the user can select which parts and elements will be displayed and in what way. See, for example, TouchGraph [92].

Closely related to the multiview concept are the associated concepts of glasses, lenses and zooming. Glasses affect the entire window, while lenses affect only selected region or elements.

By selecting different glasses, we can obtain different views on the same data supporting different visualization aims. For example, in Fig. 10 four different glasses (ball and stick, space-fill, backbone, ribbons) were applied in the program Rasmol to the molecule 1atn.pdb (deoxyribonuclease I complex with actin).

Another example of glasses is presented in Fig. 11. The two pictures were produced by James Moody [35]. The graph pictured was obtained by applying spring em-

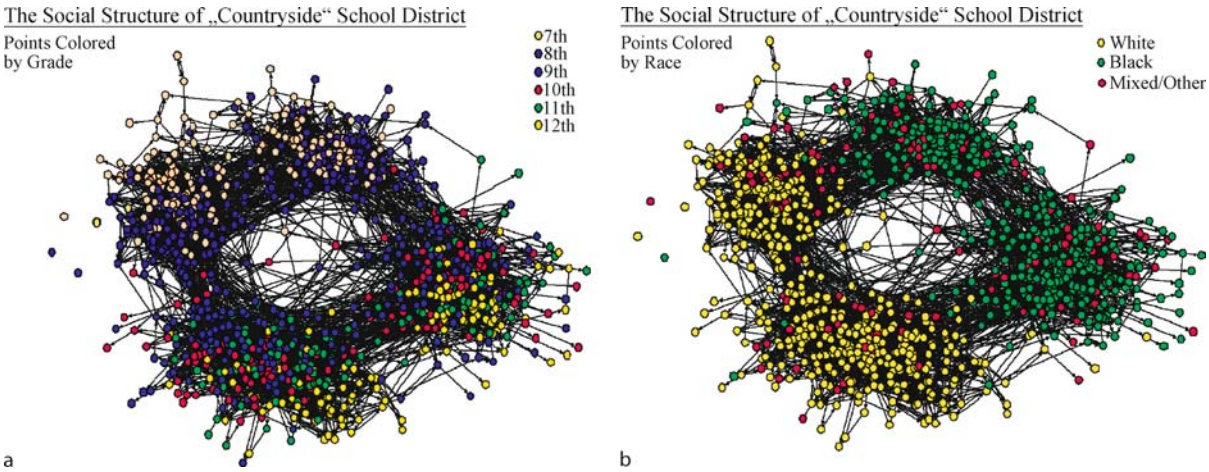

Complex Networks, Visualization of, Figure 11
Glasses: Display of properties – school

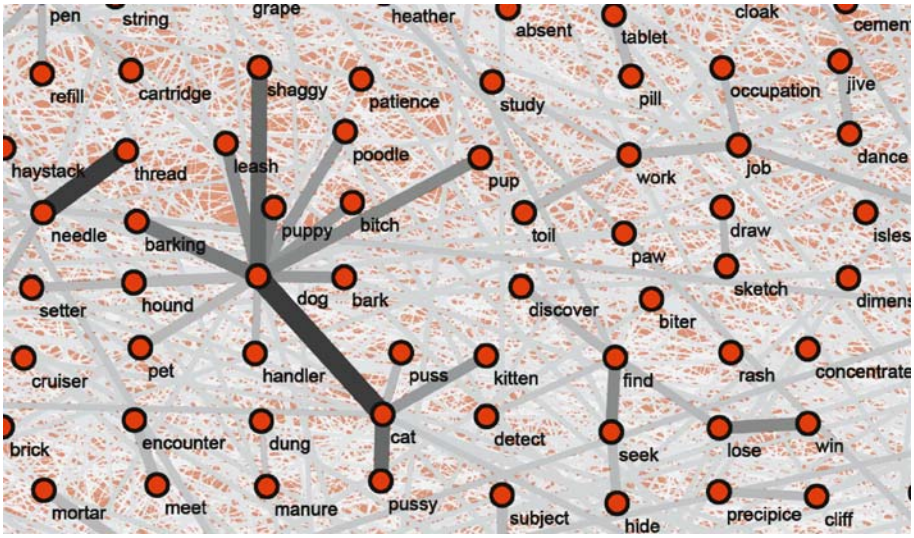

Complex Networks, Visualization of, Figure 12
Part of the big picture

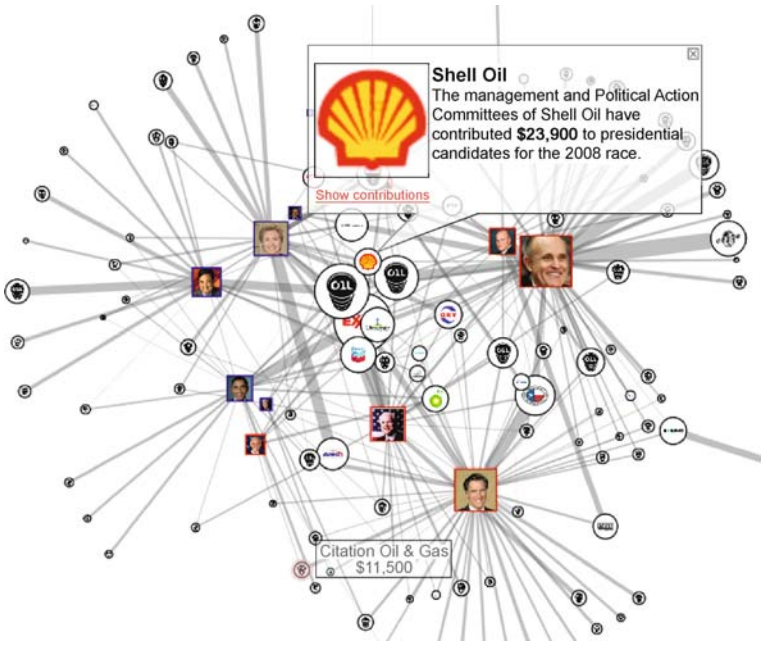

Complex Networks, Visualization of, Figure 13
Lenses: Temporary info about the selected vertex

Complex Networks, Visualization of, Figure 14
Zoom, glasses, lenses, navigation: Google Maps

bedders. It represents the friendships among students in a school. The glasses are the coloring of its vertices by different partitions: an age partition (left picture) and a race partition (right picture). This gives us an explanation of the four groups in the resulting graph picture, characterized by younger/older and white/black students.

Figure 12 shows a part of the big picture presented in Fig. 9. The glasses in this case are based on ordering the edges in increasing order of their values and drawing them in this order – stronger edges cover the weaker. The picture emphasizes the strongest substructures; the remaining elements form a background.

There are many kinds of glasses in representation of graphs, for example, fish-eye views, matrix representation, using application field conventions (genealogies, molecules, electric circuits, SBGN), displaying vertices only, selecting the type of labels (long/short name, value), displaying only the important vertices and/or lines, size of vertices determined by core number or "betweennes".

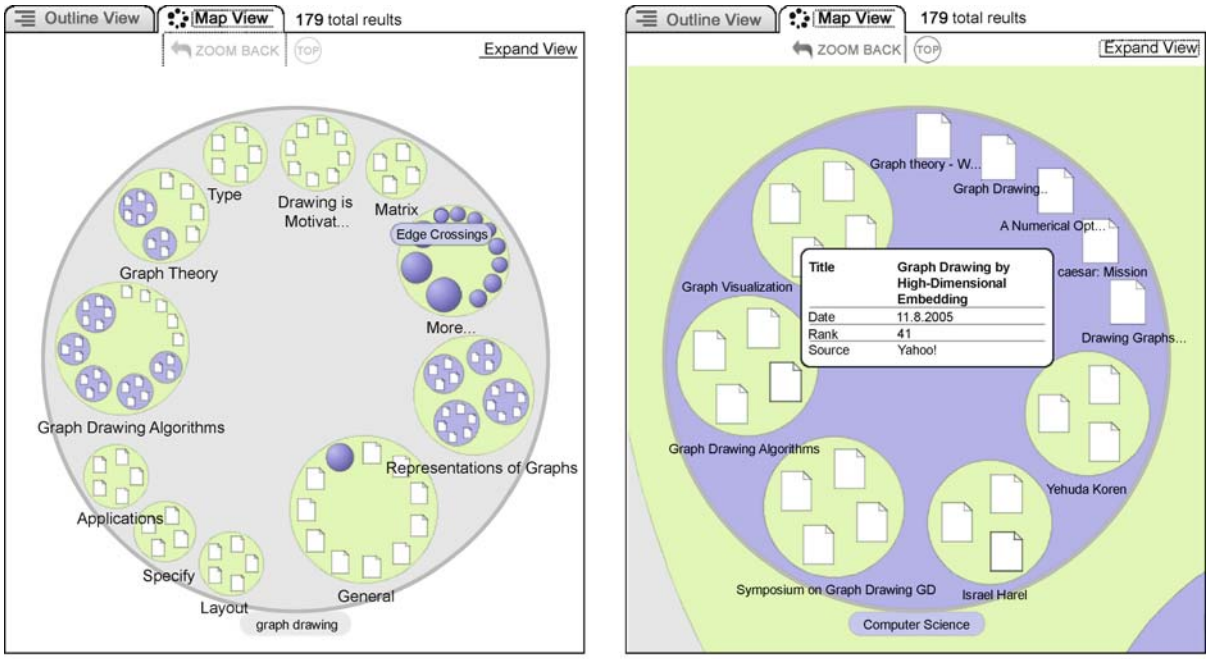

a b

Complex Networks, Visualization of, Figure 15
Zoom, glasses, lenses, navigation: Grokker

An example of lens is presented in Fig. 13 – contributions of companies to various presidential candidates from *Follow the Oil Money* by Greg Michalec and Skye Bender-deMoll [62]. When a vertex is selected, information about that vertex is displayed. Another possible use of a lens would be to temporarily enhance the display of neighbors of a selected vertex [94] or to display their labels. The "shaking" option used in *Pajek* to visually identify all vertices from a selected cluster is also a kind of lens; so are the matrix representations of selected clusters in NodeTrix [72].

Additional enhancement of a presentation can be achieved by the use of support elements such as labels, grids, legends, and various forms of help facilities.

An important concept connected with zooming is the level of detail, or LOD – subobjects are displayed differently depending on the zoom depth.

A nice example of a combination of these techniques is the Google Maps service [65] – see Fig. 14. It combines zooming, glasses (Map, Satellite, Terrain), navigation (left, right, up, down) and lenses (info about points). The maps at different zoom levels provide information at different levels of detail and in different forms.

A similar approach could be used for inspection of a large graph or network by examining selected hierarchical clusterings of its vertices. To produce higher level 'maps,' different methods can be used: *k*-core representation [4], density contours [83], generalized blockmodeling [15], clustering [71] (Fig. 15), preserving only important vertices and lines, etc. In visualizing 'maps,' new graphical elements (many of them still to be invented) can be used (see [13,80], p. 223 in [15]) to preserve or to indicate information about structures at lower levels.

The *k*-core representation [4,81] is based on *k*-core decomposition of a network [6,7] and was developed by Alessandro Vespignani and his collaborators [54]. Fig-

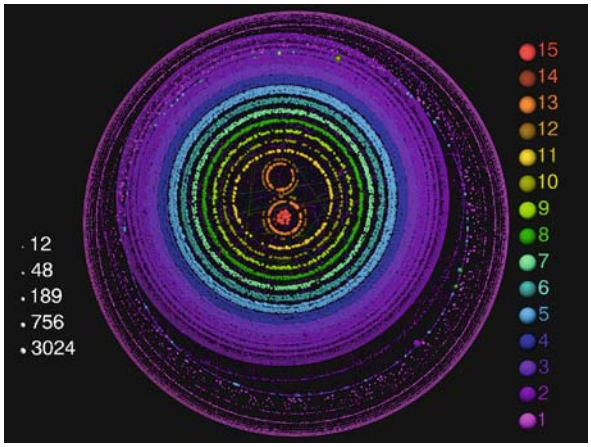

Complex Networks, Visualization of, Figure 16
***k*-core structure of a portion of the web at the .fr domain**

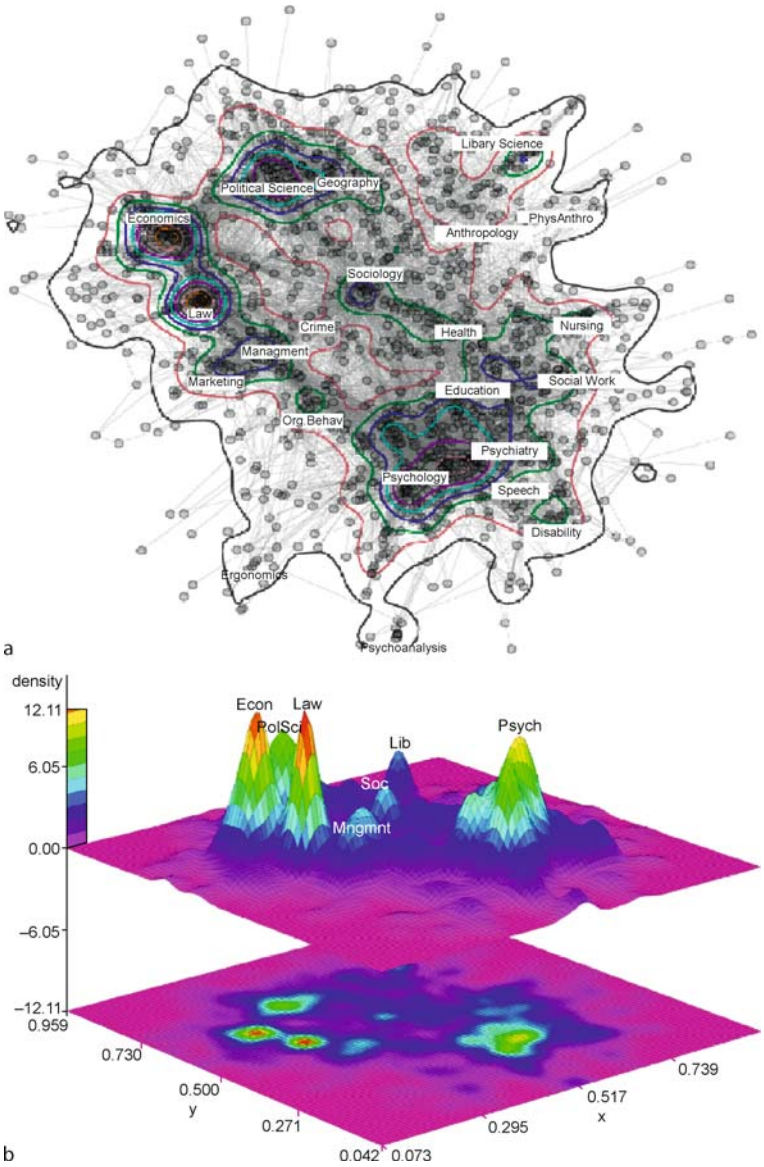

Complex Networks, Visualization of, Figure 17
Density structure

ure 16 shows a portion of the web at the .fr domain with 1 million pages. Each node represents a web page and each edge is a hyperlink between two pages.

Density contours were introduced by James Moody in 2006. First, a spring embedder layout of a (valued) network is determined. Next, vertices and lines are removed and replaced by density contours. Figure 17 shows this process applied to the case of a social science co-citation network. The left side shows the network layout and the right bottom part presents the corresponding density contours.

The basic steps in graph/network visualization are:

$$\text{graph/network} \rightarrow \boxed{\text{analysis}} \rightarrow \text{layouts} \rightarrow \boxed{\text{viewer}} \rightarrow \text{pictures}\,.$$

Development of different tools can be based on this scheme, depending on the kind of users (simple, advanced) and the tasks they address (reporting, learning, monitoring, exploration, analysis). In some cases, a simple viewer will be sufficient (for example SVG viewer, X3D viewer, or a special graph layout viewer), in others

a complete network analysis system is needed (such as Geomi [3,64], ILOG [76], Pajek [50], Tulip [93], yFiles [96]).

To visualize a network, layouts are obtained by augmenting network data with results of analysis and users' decisions. In *Pajek*'s input format, there are several layout elements from *Pajek*'s predecessors (see *Pajek*'s manual, pp. 69–73 in [50]). As in typesetting

$$\text{text} + \text{formatting} = \text{formatted text}$$

so in network visualization

$$\text{network} + \text{layout} = \text{picture}\,.$$

It would be useful to define a common layout format (an extension of GraphML [68]?) so that independent viewer modules can be developed and combined with different layout algorithms. Some useful ideas can be found in the nViZn ("envision") system [89]. To specify layouts we can borrow from typesetting the notion of *style*.

Bibliography

Primary Literature

1. Abello J, van Ham F (2004) Matrix zoom: A visual interface to semi-external graphs. IEEE Symposium on Information Visualization, October 10–12 2004, Austin, Texas, USA, pp 183–190
2. Abello J, van Ham F, Krishnan N (2006) ASK-GraphView: A large scale graph visualization system. IEEE Trans Vis Comput Graph 12(5):669–676
3. Ahmed A, Dwyer T, Forster M, Fu X, Ho J, Hong S, Koschützki D, Murray C, Nikolov N, Taib R, Tarassov A, Xu K (2006) GEOMI: GEOmetry for maximum insight. In: Healy P, Eades P (eds) Proc 13th Int Symp Graph Drawing (GD2005). Lecture Notes in Computer Science, vol 3843. Springer, Berlin, pp 468–479
4. Alvarez-Hamelin JI, DallAsta L, Barrat A, Vespignani A (2005) Large scale networks fingerprinting and visualization using the k-core decomposition. In: Advances in neural information processing systems 18, Neural Information Processing Systems, NIPS 2005, December 5–8, 2005, Vancouver, British Columbia, Canada
5. Batagelj V, Mrvar A (2003) Pajek – analysis and visualization of large networks. In: Jünger M, Mutzel P (eds) Graph drawing software. Springer, Berlin, pp 77–103
6. Batagelj V, Zaveršnik M (2002) Generalized cores. arxiv cs.DS/0202039
7. Batagelj V, Mrvar A, Zaveršnik M (1999) Partitioning approach to visualization of large graphs. In: Kratochvíl J (ed) Lecture notes in computer science, vol 1731. Springer, Berlin, pp 90–97
8. Becker RA, Eick SG, Wilks AR (1995) Visualizing network data. IEEE Trans Vis Comput Graph 1(1):16–28
9. Boyack KW, Klavans R, Börner K (2005) Mapping the backbone of science. Scientometrics 64(3):351–374
10. Boyack KW, Klavans R, Paley WB (2006) Map of science. Nature 444:985
11. Brandes U, Pich C (2007) Eigensolver methods for progressive multidimensional scaling of large data. In: Proc 14th Intl Symp Graph Drawing (GD '06). Lecture notes in computer science, vol 4372. Springer, Berlin, pp 42–53
12. Brandes U, Fleischer D, Lerner J (2006) Summarizing dynamic bipolar conflict structures. IEEE Trans Vis Comput Graph (special issue on Visual Analytics) 12(6):1486–1499
13. Dickerson M, Eppstein D, Goodrich MT, Meng J (2005) Confluent drawings: Visualizing non-planar diagrams in a planar way. J Graph Algorithms Appl (special issue for GD'03) 9(1):31–52
14. Dodge M, Kitchin R (2001) The atlas of cyberspace. Pearson Education, Addison Wesley, New York
15. Doreian P, Batagelj V, Ferligoj A (2005) Generalized blockmodeling. Cambridge University Press, Cambridge
16. Dwyer T, Koren Y (2005) DIG-COLA: Directed graph layout through constrained energy minimization. INFOVIS 2005:9
17. Eades P (1984) A heuristic for graph drawing. Congressus Numerantium 42:149–160
18. Fairchild KM, Poltrock SE, Furnas GW (1988) SemNet: Three-dimensional representations of large knowledge bases. In: Guindon R (ed) Cognitive science and its applications for human-computer interaction. Lawrence Erlbaum, Hillsdale, pp 201–233
19. Freeman LC (2000) Visualizing social networks. J Soc Struct 1(1). http://wwww.cmu.edu/joss/content/articles/volume1/Freeman/
20. Freeman LC (2004) The development of social network analysis: A study in the sociology of science. Empirical, Vancouver
21. Fruchterman T, Reingold E (1991) Graph drawing by force directed placement. Softw Pract Exp 21(11):1129–1164
22. Gajer P, Kobourov S (2001) GRIP: Graph drawing with intelligent placement. Graph Drawing 2000 LNCS, vol 1984:222–228
23. Gansner ER, Koren Y, North SC (2005) Topological fisheye views for visualizing large graphs. IEEE Trans Vis Comput Graph 11(4):457–468
24. Graph Drawing. Lecture Notes in Computer Science, vol 894 (1994), 1027 (1995), 1190 (1996), 1353 (1997), 1547 (1998), 1731 (1999), 1984 (2000), 2265 (2001), 2528 (2002), 2912 (2003), 3383 (2004), 3843 (2005), 4372 (2006), 4875 (2007). Springer, Berlin
25. Hachul S, Jünger M (2007) Large-graph layout algorithms at work: An experimental study. JGAA 11(2):345–369
26. Harel D, Koren Y (2004) Graph drawing by high-dimensional embedding. J Graph Algorithms Appl 8(2):195–214
27. Henry N, Fekete J-D (2006) MatrixExplorer: A dual-representation system to explore social networks. IEEE Trans Vis Comput Graph 12(5):677–684
28. Herman I, Melancon G, Marshall MS (2000) Graph visualization and navigation in information visualization: A survey. IEEE Trans Vis Comput Graph 6(1):24–43
29. Hu YF (2005) Efficient and high quality force-directed graph drawing. Math J 10:37–71
30. Kamada T, Kawai S (1988) An algorithm for drawing general undirected graphs. Inf Proc Lett 31:7–15
31. Knuth DE (1963) Computer-drawn flowcharts. Commun ACM 6(9):555–563
32. Koren Y (2003) On spectral graph drawing. COCOON 2003: 496–508
33. Kruja E, Marks J, Blair A, Waters R (2001) A short note on the history of graph drawing. In: Proc Graph Drawing 2001. Lecture notes in computer science, vol 2265. Springer, Berlin, pp 272–286
34. Lamping J, Rao R, Pirolli P (1995) A focus+context technique

based on hyperbolic geometry for visualizing large hierarchies. CHI 95:401–408
35. Moody J (2001) Race, school integration, and friendship segregation in America. Am J Soc 107(3):679–716
36. Moreno JL (1953) Who shall survive? Beacon, New York
37. Mueller C, Martin B, Lumsdaine A (2007) A comparison of vertex ordering algorithms for large graph visualization. APVIS 2007, pp 141–148
38. Munzner T (1997) H3: Laying out large directed graphs in 3D hyperbolic space. In: Proceedings of the 1997 IEEE Symposium on Information Visualization, 20–21 October 1997, Phoenix, AZ, pp 2–10
39. Noack A (2007) Energy models for graph clustering. J Graph Algorithms Appl 11(2):453–480
40. Shneiderman B (1996) The eyes have it: A task by data type taxonomy for information visualization. In: IEEE Conference on Visual Languages (VL'96). IEEE CS Press, Boulder
41. Shneiderman B, Aris A (2006) Network visualization by semantic substrates. IEEE Trans Vis Comput Graph 12(5):733–740
42. Sugiyama K, Tagawa S, Toda M (1981) Methods for visual understanding of hierarchical systems. IEEE Trans Syst, Man, Cybern 11(2):109–125
43. Tutte WT (1963) How to draw a graph. Proc London Math Soc s3-13(1):743–767
44. Walshaw C (2003) A multilevel algorithm for force-directed graph-drawing. J Graph Algorithms Appl 7(3):253–285
45. Wetherell C, Shannon A (1979) Tidy drawing of trees. IEEE Trans Softw Engin 5:514–520
46. White DR, Jorion P (1992) Representing and computing kinship: A new approach. Curr Anthr 33(4):454–463
47. Wills GJ (1999) NicheWorks-interactive visualization of very large graphs. J Comput Graph Stat 8(2):190–212

Web Resources

48. 3D Macromolecule analysis and Kinemage home page: http://kinemage.biochem.duke.edu/. Accessed March 2008
49. Aguidel: http://www.aguidel.com/en/. Accessed March 2008
50. Batagelj V, Mrvar A (1996) Pajek – program for analysis and visualization of large network: http://pajek.imfm.si. Accessed March 2008. Data sets: http://vlado.fmf.uni-lj.si/pub/networks/data/. Accessed March 2008
51. Brookhaven Protein Data Bank: http://www.rcsb.org/pdb/. Accessed March 2008
52. Caida: http://www.caida.org/home/. Accessed March 2008. Walrus gallery: http://www.caida.org/tools/visualization/walrus/gallery1/. Accessed March 2008
53. Cheswick B: Internet mapping project – map gallery: http://www.cheswick.com/ches/map/gallery/. Accessed March 2008
54. Complex Networks Collaboratory: http://cxnets.googlepages.com/. Accessed March 2008
55. Cruz I, Tamassia R (1994) Tutorial on graph drawing. http://graphdrawing.org/literature/gd-constraints.pdf. Accessed March 2008
56. Davis T: University of Florida Sparse Matrix Collection: http://www.cise.ufl.edu/research/sparse/matrices. Accessed March 2008
57. Di Battista G, Eades P, Tamassia R, Tollis IG (1994) Algorithms for drawing graphs: An annotated bibliography. Comput Geom: Theory Appl 4:235–282. http://graphdrawing.org/literature/gdbiblio.pdf. Accessed March 2008
58. Dodge M: Cyber-Geography Research: http://personalpages.manchester.ac.uk/staff/m.dodge/cybergeography/. Accessed March 2008
59. Edinburgh Associative Thesaurus (EAT): http://www.eat.rl.ac.uk/. Accessed March 2008
60. FamilySearch: http://www.familysearch.org/. Accessed March 2008
61. FASresearch, Vienna, Austria: http://www.fas.at/. Accessed March 2008
62. Follow the Oil Money: http://oilmoney.priceofoil.org/. Accessed March 2008
63. GDToolkit – Graph Drawing Toolkit: http://www.dia.uniroma3.it/~gdt/gdt4/index.php. Accessed March 2008
64. GEOMI (Geometry for Maximum Insight): http://www.cs.usyd.edu.au/~visual/valacon/geomi/. Accessed March 2008
65. Google Maps: http://maps.google.com/. Accessed March 2008
66. Graphael: http://graphael.cs.arizona.edu/. Accessed March 2008
67. Graphdrawing home page: http://graphdrawing.org/. Accessed March 2008
68. GraphML File Format: http://graphml.graphdrawing.org/. Accessed March 2008
69. Graphviz: http://graphviz.org/. Accessed March 2008
70. GRIP: http://www.cs.arizona.edu/~kobourov/GRIP/. Accessed March 2008
71. Grokker – Enterprise Search Management and Content Integration: http://www.grokker.com/. Accessed March 2008
72. Henry N, Fekete J-D, Mcguffin M (2007) NodeTrix: Hybrid representation for analyzing social networks: https://hal.inria.fr/inria-00144496. Accessed March 2008
73. Herr BW, Holloway T, Börner K (2007) Emergent mosaic of wikipedian activity: http://www.scimaps.org/dev/big_thumb.php?map_id=158. Accessed March 2008
74. Hu YF: Gallery of Large Graphs: http://www.research.att.com/~yifanhu/GALLERY/GRAPHS/index1.html. Accessed March 2008
75. iCKN: TeCFlow – a temporal communication flow visualizer for social network analysis: http://www.ickn.org/. Accessed March 2008
76. ILOG Diagrams: http://www.ilog.com/. Accessed March 2008
77. Infovis – 1100+ examples of information visualization: http://www.infovis.info/index.php?cmd=search&words=graph&mode=normal. Accessed March 2008
78. INSNA – International Network for Social Network Analysis: http://www.insna.org/. Accessed March 2008
79. Journal of Graph Algorithms and Applications: http://jgaa.info/. Accessed March 2008
80. KartOO visual meta search engine: http://www.kartoo.com/. Accessed March 2008
81. LaNet-vi – Large Network visualization tool: http://xavier.informatics.indiana.edu/lanet-vi/. Accessed March 2008
82. MDL Chime: http://www.mdli.com/. Accessed March 2008
83. Moody J (2007) The network structure of sociological production II: http://www.soc.duke.edu/~jmoody77/presentations/soc_Struc_II.ppt. Accessed March 2008
84. Mueller C: Matrix visualizations: http://www.osl.iu.edu/~chemuell/data/ordering/sparse.html. Accessed March 2008
85. OLIVE, On-line library of information visualization environments: http://otal.umd.edu/Olive/. Accessed March 2008

86. Pad++: Zoomable user interfaces: Portal filtering and 'magic lenses': http://www.cs.umd.edu/projects/hcil/pad++/tour/lenses.html. Accessed March 2008
87. RasMol Home Page: http://www.umass.edu/microbio/rasmol/index2.htm. Accessed March 2008
88. Sonia – Social Network Image Animator: http://www.stanford.edu/group/sonia/. Accessed March 2008
89. SPSS nViZn: http://www.spss.com/research/wilkinson/nViZn/nvizn.html. Accessed March 2008
90. SVGanim: http://vlado.fmf.uni-lj.si/pub/networks/pajek/SVGanim. Accessed March 2008
91. Tom Sawyer Software: http://www.tomsawyer.com/home/index.php. Accessed March 2008
92. TouchGraph: http://www.touchgraph.com/. Accessed March 2008
93. Tulip: http://www.labri.fr/perso/auber/projects/tulip/. Accessed March 2008
94. Viégas FB, Wattenberg M (2007) Many Eyes: http://services.alphaworks.ibm.com/manyeyes/page/Network_Diagram.html. Accessed March 2008
95. Visual complexity: http://www.visualcomplexity.com/vc/. Accessed March 2008
96. yWorks/yFiles: http://www.yworks.com/en/products_yfiles_about.htm. Accessed March 2008

Books and Reviews

Bertin J (1967) Sémiologie graphique. Les diagrammes, les réseaux, les cartes. Mouton/Gauthier-Villars, Paris/La Haye
Brandes U, Erlebach T (eds) (2005) Network analysis: Methodological foundations. LNCS. Springer, Berlin
Carrington PJ, Scott J, Wasserman S (eds) (2005) Models and methods in social network analysis. Cambridge University Press, Cambridge
de Nooy W, Mrvar A, Batagelj V (2005) Exploratory social network analysis with Pajek. Cambridge University Press, Cambridge
di Battista G, Eades P, Tamassia R, Tollis IG (1999) Graph drawing: Algorithms for the visualization of graphs. Prentice Hall, Englewood Cliffs
Jünger M, Mutzel P (eds) (2003) Graph drawing software. Springer, Berlin
Kaufmann M, Wagner D (2001) Drawing graphs, methods and models. Springer, Berlin
Tufte ER (1983) The visual display of quantitative information. Graphics, Cheshire
Wasserman S, Faust K (1994) Social network analysis: Methods and applications. Cambridge University Press, Cambridge
Wilkinson L (2000) The grammar of graphics. Statistics and Computing. Springer, Berlin

Complex Systems and Emergent Phenomena

Henrik Jeldtoft Jensen
Institute for Mathematical Sciences and Department of Mathematics, Imperial College London, London, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Equilibrium Averages
The Two-Dimensional XY-Model
Lack of Ordering in Two Dimensions
Vortex Unbinding
The Vortex Unbinding Transition in Other Systems
Non-Equilibrium Systems
1/f Fluctuations – a Langevin Approach
Self-organized Structure Formation – Ant Trails
Summary and Future Directions
Acknowledgments
Bibliography

Glossary

Correlations is the degree, to which events at different positions and at different times depend on or influence each other, is measured by correlation functions. If two events are statistically independent, the correlation between them is zero. The opposite is not necessarily the case, but one will often expect that if correlations decay, the mutual dependence does likewise.

Correlation function describes correlations between two quantities and depends on their separation in time and space.

Complex systems consist of a large number of interacting components. The interactions give rise to emergent hierarchical structures. The components of the system and properties at systems level typically change with time. A complex system is inherently open and its boundaries often a matter of convention.

Equilibrium In statistical mechanics the prototype equilibrium system consists of a "small" system in thermal contact with another system, the latter being big enough to act as a heat bath. A heat bath is defined as a system so big that when it exchanges energy with the small system the temperature of the heat bath remains the same. The statistical properties of equilibrium systems are independent of time.

Generalized rigidity is a term introduced by P.W. Anderson [1] to describe the situation, when a many component system acts as a globally connected unit, in the sense that if one apply a force at one point, the effect can be transmitted across the system. Ice has rigidity, if we push at one point, the entire piece of ice will start moving. If the ice, on the other hand melts to water, a force applied locally will only have an effect locally.

Hamiltonian expresses the energy of a system as a function of the degrees of freedom, in terms of which the system is defined at the considered level of description. Emergence in physical systems can sometimes be understood in terms of lumping degrees of freedom, in the Hamiltonian, together in sets of effective degrees of freedom, e. g. the center of mass of a solid body.

Non-equilibrium systems is a term used to describe any system that is not in equilibrium. Needless to say this is a characterization of limited value, since there are many very different types of systems included in this category.

Order parameter is a quantity that allows one to discriminate between two phases of a physical system. The order parameter changes from zero to non-zero as one passes from one phase to the other. To identify the relevant order parameter is often non-trivial and, is in itself, a first important step.

Renormalization group analysis is a systematic mathematical procedure that enables a derivation of the emergent behavior at the macroscopic systems level. The behavior at long length and time scales is obtained from the underlying microskopic short length scales and fast dynamics.

Statistical mechanics seeks to understand how properties at systems level emerge from the level of the system-components and their interactions. This often involves the application of probability theory, and a number of mathematical techniques. Throughout, we draw a distinction between statistical mechanics and statistical physics. The latter is mainly concerned with the microscopic foundation of thermodynamics and, e. g., phenomena such as phase transitions and superconductivity.

Definition of the Subject

Matter in the universe is organized in a hierarchical structure. At the bottom (if there is one) we have elementary particles, atoms and molecules from which we get macro molecules like proteins and DNA, these are the building blocks of organelles, which together form the cells. From cells we get organs, which put together form organisms: animals and plants of a great variety of species. One level of structure emerges from the level below. Is it possible to scientifically describe, let alone, predict emergence. Sometimes emergence is described as a phenomena beyond analysis. The perplexity with which this concept is sometimes met, is well illustrated by the following quote from a recent call for participation in a meeting, held by the British research council EPSRC, to look at ways to explore emergence in complex systems. Emergence is described in the following words: "For the first time since the enlightenment in the western tradition we have started to understand that there are non-causal systems in which some things just "are". The concept of emergence by whisch patterns of possibility arise through interactions of agents over time, accepts that even with the same starting conditions the same pattern would not necessarily repeat." Another attitude is represented by Lord Roberts May's statement in a recent lecture that "when people say something is an emergent property, it just means they don't understand the phenomena".

In this brief article we will argue that emergence is neither an empty concept nor a mysterious non-causal enigma. On the contrary, emergence is central to scientific enquiry. Emergence occurs when many components interact and combine to form an identifiable system. In philosophy this is the observation that quantitative changes accumulate and give rise to new qualitative changes. A proposition that can be traced from the ancient Greek philosophers through Hegel to Dialectical Materialism. In physics Phil Anderson famously summed the fact that new levels of organization need new types of description, up in the phrase "More is Different" [2]. By following the tradition of statistical mechanics it is sometimes possible to reduce significantly the confusion surrounding what emergence is, and how it can be investigated and described. Kenneth Wilson got the 1982 Nobel prize for his Renormalization Group theory, which is a particular beautiful method for extracting certain emergent properties with great mathematical detail and precision [3].

Introduction

Statistical mechanics is concerned with the interaction of many components. From interactions between the components at one given level the aim is to understand the collective coherent behavior, which emerges as many components are put together. It is through the interactions of the components at one level that the next level emerges. We consider a collection of *interacting* atoms and the outcome is, say, a transistor. In some cases the microscopic details of the properties of the *individual* building blocks are not so crucial. It happens that the collective behavior is controlled by general properties of the *interactions* between the components more than by the intrinsic properties of the components. A number of methods have been developed to bridge the gap between the individual components comprising a system and the collective whole. This often involves predictions of the asymptotic behavior at long distances and at long times. In particular the philosophy and

technique of the Renormalization Group have been successful in a number of cases in doing this. Many other approaches exist, some in which a coarse grained description is sought similar to the one used in fluid dynamics, other times the collective systems level behavior is generated by use of Individual Based Agent models typically simulated on computers. The latter method goes beyond traditional statistical mechanics, nevertheless it is reasonable to consider simulations of agent based models as part of statistical mechanics since they are very similar in spirit to traditional statistical mechanics simulations starting from the microscopic degrees of freedom, as e. g. extensively done for the celebrated Ising model used in studies of magnetic systems, melting and many other phenomena.

In this article we will discuss a number of examples of how statistical mechanics is able to deal with emergent phenomena. Our first example is from the theory of equilibrium properties of magnetic materials. In the model the microscopic magnetic moments – or spin – combine to form macroscopic coherent vortices. The vortices are bound in pairs at low temperature. When the temperature is increased, the biggest pairs are able to unbind, or fall apart. This is a subtle collective effect caused by the smaller pairs weakening the binding force between the vortices in the biggest pairs. We use this case to describe in some detail how the statistical mechanics formalism of Boltzmann and Gibbs allows us to identify the macroscopic excitations. These macroscopic excitations are what constitute the macroscopic "components" of the system, once they are identified we can calculate the unbinding of the vortex pairs. This discussion focuses on spatial aspects. The mathematical description developed from this model has applications to a range of very different phenomena, such as melting, superconductivity, superfluidity, electrical charges in two-dimensional space and crystal growth. This is a good typical example of how the mathematical formalism of statistical mechanics is able to deliver understanding, and a description, transcending the particular and unify apparently disconnected phenomena.

Many-component systems also often exhibit emergent temporal behavior that is caused by the interactions and ensuing collective motion of the components. For example, one typically see very persistent correlations, or long time memory, in the macro-dynamics of many-component systems. A particular version of this phenomena is called $1/f$ fluctuations, and we explain below what this is, and how it is related to long time correlations or long memory effects. To make the discussion concrete we will present the details of a very simple model of diffusing particles. The model might be related to motorway traffic.

To go beyond models taken from physics we will finish by a discussion of models inspired by the observed collective behavior of social insects. Aspects of trail formation and mound building of ants and termites have been reproduced in computer simulations. Often the models consider "agents" with a tendency to perform random walks and picking up and laying down material. The agents deviate from random walking when they come across traces of smell – pheromones – laid down by other ants. This indirect interaction can lead to the formation of surprisingly intricate structures of trails and mounds.

Since essentially any activity within statistical mechanics is concerned with a description of emerging phenomena a very large literature exists and we list here only a few books of particular relevance to the view point of the present article [1,4,5,6,7,8,9,10,11,12,13] and some more specialized papers as we go along.

Equilibrium Averages

To describe how statistical mechanics is able to identify structures emerging at the macroscopic level we briefly recall how macroscopic (or systems level) quantities are obtained through averaging procedures. The reason equilibrium systems can be analyzed in particular detail is that the situation, where the systems of interest can be considered as in thermal equilibrium with a heat bath, allows for the determination of the probability weights of the individual micro-states. One starts out with the following fundamental hypothesis concerning isolated or closed systems:

- **Micro Canonical Ensemble**. For a closed system it is assumed that *all* micro-states, consistent with the macroscopic constraints, occur with *equal* probability.

The macroscopic constraints can, for example, be the total energy E (which is constant for a closed system) and the volume V. Denote by $\Omega(E, V)$ the total number of micro-states possible under these constrains. Meaning the components or particles of the system have to be located within the given volume V, and that when we add all the energies of the particles the sum must equal E. The probability $p(s)$ that the system is in a particular state s is then

$$p(s) = \frac{1}{\Omega(E, V)} \,. \tag{1}$$

Closed systems are not very interesting in the sense that one is unable to interact with them. A much more interesting situation is when the system **S** under consideration is brought in contact with a heat bath **B** or heat reservoir. The heat bath is a system so big that even when it exchanges energy with the small system of experimental interest, the heat bath remains unchanged. Say a cup of tea

in contact with the Pacific Ocean. The heat bath is characterized by its temperature T. We can now use the fundamental hypothesis above to determine the probabilistic weights for the states of **S**. Since the combined system $\mathbf{B}+\mathbf{S}$ is closed the weights for the combined system is given by the Micro Canonical Ensemble, i. e. all micro-states of the combined system are equally likely. The number of micro-states for the combined system of total energy $E_{\mathrm{Tot}} = E_{\mathbf{B}} + E_{\mathbf{S}}$ will be a product

$$\Omega_{\mathrm{Tot}}(E_{\mathrm{Tot}}) = \Omega_{\mathbf{B}}(E_{\mathbf{B}})\Omega_{\mathbf{S}}(E_{\mathbf{S}})\,. \tag{2}$$

Here one neglects the interactions between the heat bath and the system. Now focus on one particular micro-state s of **S** of energy $E_{\mathbf{s}}$. Since we have a particular state, s, in mind we have $\Omega_{\mathbf{S}}(s) = 1$. This state can be combined in many ways with states of the bath **B** as long as those fulfill the constraint $E_{\mathrm{Tot}} = E_{\mathbf{B}} + E_s$. So the probability, $p(s)$, for finding the system **S** in s, when **S** is in equilibrium with the bath, is proportional to $\Omega_{\mathbf{B}}(E_{\mathrm{Tot}} - E_s)$. Namely

$$p(s) = \frac{\Omega_{\mathbf{B}}(E_{\mathrm{Tot}} - E_s)}{\sum_{\mathrm{state}} \Omega_{\mathbf{B}}(E_{\mathrm{Tot}} - E_{\mathrm{state}})}, \tag{3}$$

the denominator ensures normalization. In order to introduce the temperature into the mathematical formalism it turns out that we should consider the logarithm of $p(s)$. We have

$$\ln[p(s)] = \text{constant} + \ln[\Omega_{\mathbf{B}}(E_{\mathrm{Tot}} - E_s)] \tag{4}$$

$$= \text{constant} + \ln[\Omega_{\mathbf{B}}(E_{\mathrm{Tot}})] - \frac{\partial \ln[\Omega_{\mathbf{B}}(E_{\mathrm{Tot}})]}{\partial E_{\mathrm{Tot}}} E_s \tag{5}$$

$$= \text{constant} - \frac{1}{k_{\mathrm{B}} T} E_s\,. \tag{6}$$

Here we Taylor expanded to linear order to obtain the first equality. The second equality follows, because it can be shown by use of the first and second law of thermodynamics that the temperature is given by

$$\frac{1}{k_{\mathrm{B}} T} = \frac{\partial \ln[\Omega_{\mathbf{B}}(E_{\mathrm{Tot}})]}{\partial E_{\mathrm{Tot}}}\,. \tag{7}$$

This is obtained in the following way. The first and second law of thermodynamics lead to the following thermodynamic identity $\mathrm{d}E = T\mathrm{d}S - p\mathrm{d}V$ where the entropy $S = k_{\mathrm{B}} \ln[\Omega(E)]$. Since the thermodynamic identity takes the form of an exact differential we conclude that $\partial E/\partial S = T$ from which Eq. (7) follows. We now conclude

$$p(s) = \frac{\mathrm{e}^{-\frac{E_s}{k_{\mathrm{B}} T}}}{Z}\,, \tag{8}$$

where the constant Z is obtained from the normalization condition

$$\sum_{\mathrm{states}} p(s) = 1\,, \tag{9}$$

to be given by

$$Z = \sum_{\mathrm{states}} = \mathrm{e}^{-\frac{E_s}{k_{\mathrm{B}} T}}\,. \tag{10}$$

We conclude that the probabilistic weights, needed to calculate the average macroscopic behavior of a system in contact with a heat bath at temperature T, is given by the (Boltzmann) weights in Eq. (8). And we mention that a large number of average quantities can be calculated from the sum in Eq. (10). This important sum is called the *partition function* or partition sum. Some states, or configurations of the microscopic degrees of freedom, will contribute more to the partition sum than others, such configurations can sometimes be identified as macroscopic collective excitations. These may possess a degree of robustness and stability and can in such cases be identified as macroscopic emergent objects with specific properties that can be considered essential building blocks. Perhaps it is instructive to have the following picture in mind. Think of a pool table. To describe the motion of the balls we can either follow the trajectories of all the individual molecules making up 15 colored balls or we can notice that some of the molecules move together in a coordinated way and thereby form each of the 15 balls. We can therefore instead simply follow the trajectories of the center of mass (COM) of each of the balls. Obviously we lose a lot of information since we can't go from the COM of the balls to the motion of all the molecules; whereas we can drive the COM motion if we know the motion of all the molecules. Hence we note that emergence involves a loss of information. We will discuss in detail an important and illustrative example in the next section.

The Two-Dimensional *XY*-Model

Here we describe how the averaging procedure described in the previous section can be structured in a way that allows the introduction of new effective collective degrees of freedom. These describe macroscopic excitations created by the coherent motion of a huge number of microscopic variables. When the collective degrees have been identified information concerning the detailed motion of the microscopic variables can be neglected, and one is in this way able to reduce the computational effort needed and at the same time identify the essential emergent structures. A particular clear example of this procedure consists

of the physics of systems modeled by the so-called two-dimensional XY-model.

We start our discussion by considering the formation of vortices in the sea of two-dimensional magnetic moments. The individual microscopic magnetic moments sit on the sites of a two-dimensional square lattice and the direction of the moments are confined to two dimensions, see Fig. 1 (A beautiful online interactive simulation can be found on Hans Weber's web page at http://www.mt.luth.se/~weber/). Each magnetic moment can be thought of as a magnetic needle, or an arrow, confined to two dimensions and pointing in a specific direction described by the angle θ. The magnetic moment number i is given by the vector $\mathbf{S}_i = (\cos(\theta_i), \sin(\theta_i))$.

We will use it as our reference model. We think of the model as consisting of planar rotors of unit length arranged on a two-dimensional square lattice. The Hamiltonian of the system is given by

$$H = -J \sum_{\langle i,j \rangle} \mathbf{S}_i \cdot \mathbf{S}_j = -J \sum_{\langle i,j \rangle} \cos(\theta_i - \theta_j) \,. \tag{11}$$

Here J is the coupling constant between the magnetic moments, $\langle i, j \rangle$ denotes summation over all nearest neighbor sites in the lattice, and θ_i denotes the angle of the rotor on site i with respect to some (arbitrary) polar direction in the two-dimensional vector space containing the rotors.

We shall see below how the components, the rotors, work together to form certain collective coherent structures: topological defects or topological charges. In Fig. 1 these excitations are depicted. Each consists of a whirl or vortex in the configuration of the rotors. There are vortices of opposite sign. As one moves around in a positive direction along a contour encircling the center of a vortex (black circle) the rotors perform a full rotation in the positive direction as well. When we move around one of the anti-vortices (white circle) in a similar way, the rotors undergo a full rotation in the negative direction. Although these charges are here seen as arising from rotors or magnetic moments, the impressive fact is that these topological charges also represent Coulomb charges in two dimensions, or dislocations in two-dimensional crystals, or vortices in two-dimensional superconductors or a large number of other collective excitations. The interaction between

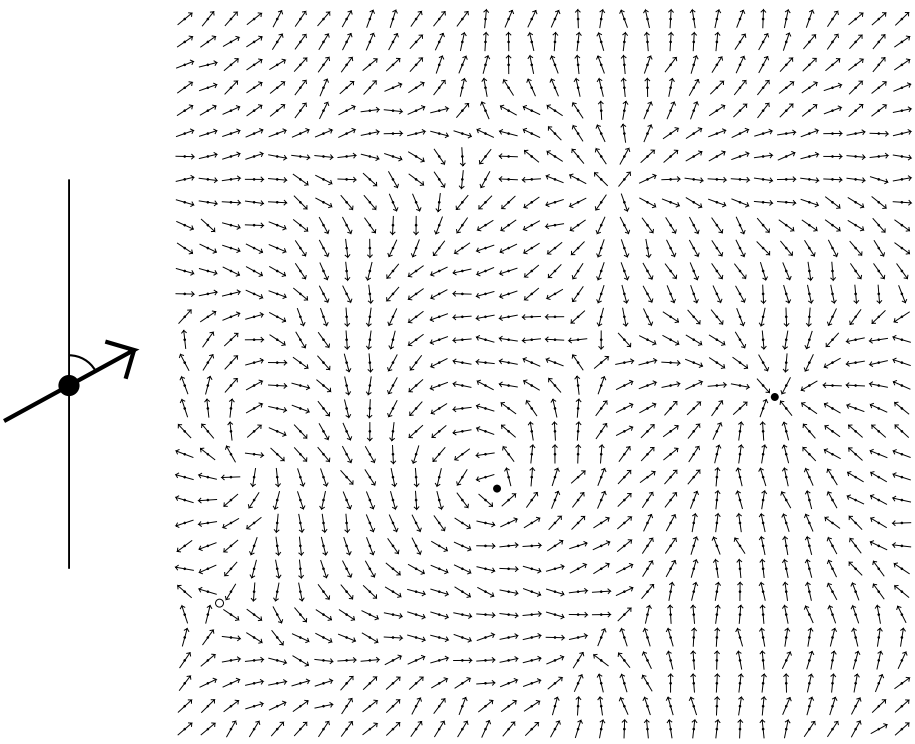

Complex Systems and Emergent Phenomena, Figure 1

Rotor configuration of the *XY*-model. Vortices (*black circle*) and anti-vortices (*white circle*) are clearly seen. The configuration is neutral, i. e. there is an equal number of vortices and anti-vortices; but only one white circle is indicated. The reader may find it amusing to try to locate the missing anti-vortex

the topological charges depends in all cases logarithmically on the spatial separation and this leads to some very general collective behavior, most spectacular the logarithmic dependence on separation causes a certain type of phase transition: the Kosterlitz–Thouless transition [14].

If we assume that the direction of the rotors varies smoothly from site to site, we can approximate $\cos(\theta_i - \theta_j)$ by the first two terms $1 - 1/2(\theta_i - \theta_j)^2$ in the Taylor expansion of cos. The sum over the nearest neighbors corresponds to the discrete Laplace operator, which we can express in terms of partial derivatives through $\theta_i - \theta_j = \partial_x \theta$ for two sites i and j which differs by one lattice spacing in the x-direction. This leads to the continuum Hamiltonian

$$H = E_0 + \frac{J}{2} \int \mathrm{d}\mathbf{r} (\nabla\theta)^2 \,. \tag{12}$$

Here $E_0 = 2JN$ is the energy of the completely aligned ground state of N rotors.

The thermodynamics of the system is obtained from the partition function

$$Z = \mathrm{e}^{-\beta E_0} \int D[\theta] \exp\left\{ -\beta \frac{J}{2} \int \mathrm{d}\mathbf{r} (\nabla\theta)^2 \right\} \,, \tag{13}$$

a functional integral over all possible configurations of the director field $\theta(\mathbf{r})$. Not all configurations will be of the same importance. By focusing on the terms in the sum that contribute most, we can identify the configurations that may be used as building blocks at the next level of description. Since the energy appears in the exponential with a negative sign in front, the most significant contributions will be those with the smaller energy – thus we have to pick out the local minima. We therefore divide the integral over $\theta(\mathbf{r})$ into a sum over the local minima θ_{vor} of $H[\theta]$ plus fluctuations θ_{sw} around the minima

$$Z = \mathrm{e}^{-\beta E_0} \sum_{\theta_{\mathrm{vor}}} \int D[\theta_{\mathrm{sw}}] \exp\Big\{ -\beta (H[\theta_{\mathrm{vor}}] + \frac{1}{2} \int \mathrm{d}\mathbf{r}_1 \int \mathrm{d}\mathbf{r}_2 \theta_{\mathrm{sw}}(\mathbf{r}_1) \frac{\delta^2 H}{\delta\theta(\mathbf{r}_1)\delta\theta(\mathbf{r}_2)} \theta_{\mathrm{sw}}(\mathbf{r}_2)) \Big\} \,. \tag{14}$$

The field configurations corresponding to local minima of H are solutions to the extremal condition

$$\frac{\delta H}{\delta\theta(\mathbf{r})} = 0 \;\Rightarrow\; \nabla^2\theta(\mathbf{r}) = 0 \,. \tag{15}$$

There are two types of solutions to this equation. The first consists of the ground state $\theta(\mathbf{r}) =$ constant. The second type of solutions consist of vortices in the director field (see Fig. 1) and are obtained by imposing the following set of boundary conditions on the circulation integral of $\theta(\mathbf{r})$:

1) For all closed curves encircling the position $\mathbf{r}_0$ of the center of the vortex

$$\oint \nabla\theta(\mathbf{r}) \cdot \mathrm{d}\mathbf{l} = 2\pi n \,, \quad n = 1, 2, \ldots \,. \tag{16}$$

2) For all paths that don't encircle the vortex position $\mathbf{r}_0$

$$\oint \nabla\theta(\mathbf{r}) \cdot \mathrm{d}\mathbf{l} = 0 \,. \tag{17}$$

Condition 1) imposes a singularity in the director field. Note the circulation integral *must* be equal to an integer times 2π since we circle a closed path and therefore $\theta(\mathbf{r})$ has to point in the same direction after traversing the path as it did when we started.

We can estimate the energy of a vortex in the following way. The problem is spherical symmetric, hence the vortex field θ_{vor} must be of the form $\theta(\mathbf{r}) = \theta(r)$. The dependence on r can be found from Eq. (16). We calculate the circulation integral along a circle of radius r centered at the position $\mathbf{r}_0$ of the vortex

$$2\pi n = \oint \nabla\theta(\mathbf{r}) \cdot \mathrm{d}\mathbf{l} = 2\pi r |\nabla\theta| \,. \tag{18}$$

We solve and obtain $|\nabla\theta(r)| = n/r$. Substitute this result into the Hamiltonian Eq. (12)

$$E_{\mathrm{vor}} - E_0 = \frac{J}{2} \int \mathrm{d}\mathbf{r} [\nabla\theta(\mathbf{r})]^2 \tag{19}$$

$$= \frac{Jn^2}{2} \int_0^{2\pi} \mathrm{d}\phi \int_a^L r\,\mathrm{d}r \frac{1}{r^2} \tag{20}$$

$$= \pi n^2 J \ln \frac{L}{a} \,. \tag{21}$$

Here a denotes the lattice constant and L is the linear size of the considered lattice. The circulation condition Eq. (16) creates a distortion in the phase field $\theta(\mathbf{r})$ that persists infinitely far from the center of the vortex. $|\nabla\theta|$ decays only as $1/r$ leading to a logarithmic divergence of the energy. Hence we need to take into account that the integral over r in Eq. (20) is cut-off for large r-values by the finite system size L and for small r-values by the lattice spacing a. We recall that our continuum Hamiltonian is an approximation to the lattice Hamiltonian in Eq. (11). A vortex with the factor n in Eq. (16) larger than one is called multiple charged. We notice that the energy of the vortex is quadratic in the charge. In a macroscopically large system even the energy of a single charge vortex will be large, and therefore we do not expect individual vortices to be thermally induced.

Consider now a pair consisting of a single charged vortex and a single charged anti-vortex. When we encircle the vortex, we pick up $\oint \mathrm{d}\mathbf{l} \cdot \nabla\theta = 2\pi$ and when

we encircle the anti-vortex, we pick up $\oint \mathrm{d}\mathbf{l} \cdot \nabla\theta = -2\pi$. Hence, if we choose a path large enough to enclose both vortices, we pick up a circulation of the phase equal to $2\pi + (-2\pi) = 0$. I. e. the distortion of the phase field $\theta(\mathbf{r})$ from the vortex-anti-vortex pair is able to cancel out at distances from the center of the two vortices large compared to the separation R between the vortex and the anti-vortex, see Fig. 1. This explains why the energy of the vortex pair is of the form [15,16,17]

$$E_{2\mathrm{vor}}(R) = 2E_c + E_1 \ln(R/a) \,. \tag{22}$$

Where E_c is the energy of a vortex core and E_1 is proportional to J. In detail, the phase field $\theta_{2\mathrm{vor}}(\mathbf{r})$ of a vortex located at $\mathbf{r} = (-a, 0)$ and an anti-vortex located at $\mathbf{r} = (a, 0)$ is given by [15]

$$\theta_{2\mathrm{vor}}(\mathbf{r}) = \mathrm{arctg}\left(\frac{2ay}{a^2 - r^2}\right) \,. \tag{23}$$

Significant aspects of the macroscopic behavior of the XY-model can be understood by treating the vortices as particles characterized by their position and their charge and ignoring the underlying sea of rotors. Indication of this follows from the expression for the energy of a pair of vortices in Eq. 22. This energy is given in terms of the relative position of the two vortices, no reference is needed to the microscopic rotor field given in Eq. 23. The pairs of vortices have dramatic effects on the macroscopic behavior of the XY-model. At low temperature the vortices are organized in fairly small bound pairs, as the temperature is increased and more thermal energy is available the separation between paired up vortices grow and at a certain temperature the pairs break apart with the effect that the individual vortices now can move freely around as they are no longer kept in check by their partner of the opposite charge. The result is the Kosterlitz–Thouless transition which manifests itself in various ways in different realizations of the XY-model. Before we discuss this transition we will look at the average ordering of the rotors. This quantity – the magnetization – is usually able to monitor if a dramatic change in the macroscopic behavior occurs as a function of temperature. But not so in the 2d XY-model. To understand this makes it clearer how important it is to identify correctly the emergent excitations of a many component system.

Lack of Ordering in Two Dimensions

In order to highlight the peculiarity of two dimensions we consider the d-dimensional XY-model. We imagine a d-dimensional cubic lattice. Each lattice site contains a planar rotor or a phase. In the continuum limit the Hamiltonian is still given by Eq. (12) except the integral over $\mathbf{r}$ is now a d-dimensional integral and therefore the factor J is replaced by Ja^{2-d}. The average size of the projection of the rotors along, say, the x-direction in $\mathbf{S}$ space, i. e. the magnetization, is

$$\langle S_x \rangle = \langle \cos\theta(\mathbf{r}) \rangle \tag{24}$$

$$= \langle \cos\theta(0) \rangle \,. \tag{25}$$

Note that we might as well have chosen the y-direction. The model is isotropic and the x and the y directions are equivalent. When $\langle S_x \rangle \neq 0$ a preferred direction is singled out in the sense that on average $\mathbf{S}$ points in the direction given by $\langle S_x \rangle$. In this case we say that the rotor field possesses order. In contrast if $\langle S_x \rangle = 0$ we also have $\langle S_y \rangle = 0$, since the model is isotropic. The zero projection comes about because the rotors circulate around and on average point equally much in all directions. So we say that the rotor field is disordered or does not possess any ordering.

First we neglect the singular vortex contributions (which is perfectly safe at low temperature) and Fourier transform the phase field

$$\theta(\mathbf{r}) = \int \frac{\mathrm{d}\mathbf{k}}{(2\pi)^d} \hat{\theta}(\mathbf{k}) \mathrm{e}^{-i\mathbf{k}\cdot\mathbf{r}} \tag{26}$$

$$\theta(0) = \int \frac{\mathrm{d}\mathbf{k}}{(2\pi)^d} \hat{\theta}(\mathbf{k}) \tag{27}$$

$$\int \mathrm{d}\mathbf{r} (\nabla\theta)^2 = \int \frac{\mathrm{d}\mathbf{k}}{(2\pi)^d} k^2 \theta(\hat{\mathbf{k}}) \hat{\theta}(-\mathbf{k}) \,. \tag{28}$$

These equations are substituted into the expression

$$\begin{aligned} \langle S_x \rangle &= \frac{\int D[\theta] \cos(\theta(0)) \mathrm{e}^{-\beta H}}{\int D[\theta] \mathrm{e}^{-\beta H}} \\ &= Re\left(\frac{\int D[\theta] \mathrm{e}^{-\beta H + i\theta(0)}}{Z}\right) \,. \end{aligned} \tag{29}$$

After some algebra one obtains the following expression

$$\langle S_x \rangle = \exp\left(-\frac{T}{2Ja^{2-d}} S_d \int_{\pi/L}^{\pi/a} \mathrm{d}k k^{d-3}\right) \,. \tag{30}$$

The behavior of $\langle S_x \rangle$ is controlled by the integral

$$I(L) = \int_{\frac{\pi}{L}}^{\frac{\pi}{a}} \mathrm{d}k k^{d-3} \,. \tag{31}$$

The behavior of $I(L)$ strongly depends on the dimension d. For $d < 2$ we have $I(L) \sim L^{2-d} \to \infty$ as $L \to \infty$. Hence,

$\langle S_x \rangle = 0$ in the limit of large systems for dimensions less than 2. For $d > 2$ we have that

$$I(L) \to A = \frac{1}{d-2}\left(\frac{\pi}{a}\right)^{d-2} \tag{32}$$

and therefore

$$\langle S_x \rangle = \exp\left(-\frac{S_d}{2Ja^{2-d}}AT\right) > 0\,. \tag{33}$$

Finally for $d = 2$ the integral $I(L)$ is logarithmically divergent $I(L) = \ln(L/a)$ which is sufficient to force $\langle S_x \rangle$ to zero for any non-zero temperature.

We conclude that there is no ordered phase according to the behavior of $\langle S_x \rangle$ at low temperature for $d \leq 2$. For $d < 2$ this means that there is no phase transition. The same was for a while thought to be the case for $d = 2$. Since the thermal motion included in the calculation of $\langle S_x \rangle$ is able to prevent a preferred direction and hence ensure $\langle S_x \rangle = 0$ for $T > 0$ including other types of excitations, such as vortices, can surely not make $\langle S_x \rangle$ different from zero. So it is safe to conclude that the rotors are unable to order along a common direction for any non-zero temperature and it was accordingly expected that a phase transition in the 2d YX-model was excluded. This conclusion was reached since in magnetic systems the average of the local magnetic moment, i. e. $\langle S_x \rangle$ is the order parameter and the phase transition takes place at the temperature where the order parameter changes from zero to a non-zero value. It turned out that by identifying the vortices as emergent collective excitations and by understanding their physical effects, a phase transition of a new kind was discovered in the 2d XY-model.

Vortex Unbinding

An indication of the importance of vortices as the temperature is increased can be obtained from the following simple and heuristic argument. We estimate the free energy of a single vortex. The Helmholtz free energy is given by the difference between the energy and the entropy multiplied by the temperature $F = E - TS$. The energy is given by Eq. (21). We estimate the entropy from the number of places where we can position the vortex center, namely on each of the $(L/a)^2$ plaquette of the square lattice, i. e., $S = k_B \ln(L^2/a^2)$. Accordingly, the free energy is given by

$$F = E_0 + (\pi J - 2k_B T)\ln(L/a)\,. \tag{34}$$

For $T < \pi J/2k_B$ the free energy will diverge to plus infinity as $L \to \infty$. At temperatures $T > \pi J/2k_B$ the system can lower its free energy by producing vortices: $F \to -\infty$ as $L \to \infty$. This simple heuristic argument points to the fact that the logarithmic dependence on system size of the energy of the vortex combines with the logarithmic dependence of the entropy to produce the subtleties of the vortex unbinding transition. Assume a different dependence of the energy on system size and one will either have thermal activation of vortices at all temperatures (in case $E_{\text{vor}} \to \text{const.} < \infty$) or vortices will not be activated at any temperature (in case $E_{\text{vor}} \sim (L/a)^b$ with $b > 0$). It is the logarithmic size dependence of the 2d vortex energy that allows the outcome of the competition between the entropy and the energy to change qualitatively at a certain finite temperature T_{KT}.

In reality it is not single vortices of the same sign that proliferate at a certain temperature. What happens is that the larger vortex pairs which are bound together for temperatures below T_{KT} unbind at T_{KT}. This is a collective effect that can be treated quantitatively by use of a special Renormalization Group method design by Kosterlitz [18]. The vortex pairs induced as one approaches T_{KT} disturb the phase field so much that the effective value of the vortex binding term E_1 in the vortex pair free energy, that is Eq. (22) generalized to non-zero temperature, is driven to zero for large vortex separations. In the next section we shall see in detail how this happens.

The Spin Wave Stiffness

As our concern in this article is with emergent entities, we will now briefly discuss how a focus on, and an understanding of, the vortex degrees of freedom makes it possible to identify and describe the previously "hidden" phase transition in the 2d XY-system.

The effect of the thermally activated vortex pairs is describe by the temperature dependent *spin wave stiffness* ρ_s^R. This is an example of what Philip W. Anderson calls a generalized rigidity [1]. The spin wave stiffness describes how much free energy it costs to apply a twist, or gradient, to the rotors (also called spins):

$$\theta(\mathbf{r}) = \theta_0(\mathbf{r}) + \mathbf{v}_{\text{ex}} \cdot \mathbf{r}\,, \tag{35}$$

here $\theta_0(\mathbf{r})$ is allowed to vary according to the canonical ensemble. The increase in the free energy is given by

$$F(\mathbf{v}_{\text{ex}}) - F(0) = \frac{1}{2}V\rho_s^R v_{\text{ex}}^2\,. \tag{36}$$

A number of comments concerning the notation are illuminating. The notation $\mathbf{v}_{\text{ex}}$ for the gradient applied to the phase field $\theta(\mathbf{r})$ has its origin in the fact that the same physics, as we describe here, applies to superfluid films and superconducting films. In these cases the field $\theta(\mathbf{r})$ is the phase of the complex order parameter, the wave function

of the super-fluid. Being the phase of a quantum mechanical wave function the gradient of $\theta(\mathbf{r})$ is related to a probability current and thereby to the velocity field of the superfluid. The notation ρ_s^R is meant to remind one that this phase rigidity, is determined by the **density** of superfluid in the case of a superfluid or a superconductor. The superscript R in ρ_s^R indicates that thermal excitations renormalize the quantity. It follows immediately from the Hamiltonian in Eq. (12) that at zero temperature $\rho_s^R = J = \rho_s$. The spin wave stiffness is similar to the shear constant of a material. The shear constant determines how the (free) energy increase when a shear deformation is imposed. As temperature is increased the shear constant decreases and drops abruptly to zero when the solid melts into a liquid.

To obtain ρ_s^R one calculates the left hand side of Eq. (36). Details can be found in the wonderful book by Chaikin and Lubensky [9]. The phase field is split into two parts

$$\theta_0(\mathbf{r}) = \theta_s(\mathbf{r}) + \theta_v(\mathbf{r}) , \tag{37}$$

where the first term describes smooth spin waves and the second term contains the singular vortex contribution. The free energy is obtained from $F = k_B T \ln Z$ and the partition function is given by Eq. (13). To calculate Z introduce Fourier transforms of the phase field. After quite a bit of algebra one arrives at the following simple expression

$$\rho_s^R = \rho_s - \frac{1}{2}\frac{\rho_s^2}{T} \lim_{k\to 0} \frac{\langle \hat{n}(\mathbf{k})\hat{n}(-\mathbf{k})\rangle_0}{k^2} , \tag{38}$$

which expresses the renormalized stiffness in terms of the correlation function of the Fourier transform of the vortex density function

$$n(\mathbf{r}) = \sum_\alpha n_\alpha \delta(\mathbf{r} - \mathbf{r}_\alpha) , \tag{39}$$

for a collection of vortices of charge n_α (see Eq. (16)) with centers located at positions $\mathbf{r}_\alpha$. The vortices are now described entirely by their position exactly like if they were ordinary particles. So what started out as a complex configuration in the field of rotors is now possible to treat as point particles. The effect of the extended disturbance of the rotor field is taken care of by the interaction energy between two vortices. The thermodynamic average in Eq. (38) is over the canonical ensemble with no twist imposed, hence the subscript 0. Eq. (38) can be used to determine how the spin wave stiffness behaves at large distances as a function of temperature. We will discuss how in the next section.

The KT Transition

Let us first summarize the phenomenology of the Kosterlitz–Thouless transition. As the temperature is increased more and more vortex pairs are thermally activated. This makes ρ_s^R decrease, see Eq. (38). This corresponds to a decrease in the increment of the free energy induced by a certain twist $\mathbf{v}_{ex}$. We can understand the effect from the fact that the phase field $\theta(\mathbf{r})$ becomes more and more distorted as the temperature is increased, hence the extra perturbation caused by $\mathbf{v}_{ex}$ becomes relatively less important. Quantitatively one finds

$$\rho_s^R = \begin{cases} \rho_s^R(T_{KT}^-)[1 + \text{const.}(T_{KT} - T)^{1/2}] & \text{for } T < T_{KT} \\ 0 & \text{for } T > T_{KT} . \end{cases} \tag{40}$$

Here, T_{KT} is the Kosterlitz–Thouless temperature at which vortex pairs unbind. The value of T_{KT} differs from one system to another. In the 2d XY-model $T_{KT}/J \simeq 0.893 \pm 0.002$ [19]. The remarkable thing is, that the ratio

$$\rho_s^R(T_{KT}^-)/T_{KT} = 2/\pi \tag{41}$$

is universal for all systems that undergoes a KT-transition. Since $\rho_s^R(T_{KT}^+) = 0$ Eq. (41) is referred to as the universal jump. The correlation length $\xi(T)$ behaves in a very unusual way as one approaches T_{KT} from above. We are used to a relatively slow algebraic divergence of the correlation length as the critical temperature is approached. For the KT-transition the divergence is, however, much faster

$$\xi(T) \sim \exp\left(\frac{\text{const.}}{(T - T_{KT})^{1/2}}\right) \quad \text{for} \quad T > T_{KT} . \tag{42}$$

Can we in a simple way understand this exponential divergence? Yes, we can. The phase field is significantly distorted by unbound vortices, since these vortices are not screened by a nearby anti-vortex. I. e. the phases $\theta(\mathbf{r})$ can remain correlated over distances shorter than the typical distance $D = 1/\sqrt{n_{ub}}$ between unbound vortices of density n_{ub} [20]. Or in other words, we expect the correlation length $\xi \sim D$. The vortices are thermally induced and therefore their density is expected to depend on the temperature through a Boltzmann factor $\exp(-E_{vor}/T)$. The situation described here is exactly what happens in the one-dimensional so-called ϕ^4 model. This model supports thermally activated solitons. The correlation length is set by the inverse of the soliton density and diverges exponentially as the temperature goes to zero [21]. The same thing, in a slightly simpler version, also happens in the

one-dimensional Ising model. This argument can indicate the cause of the exponential dependence of ξ. But it is no more than an indication since the exponential dependence in Eq. (42) is significantly different from a simple Boltzmann factor. This difference is due to corrective renormalization effects.

Continuous phase transitions are accompanied by divergences in thermodynamic quantities caused by the divergence of the correlation length as the critical temperature T_c is approached. The singular part of the free energy density f can be estimated as the amount of thermal energy T_c within a correlated volume ξ^d, which gives $f \sim T_c/\xi^d$. The specific heat c_V is given by the second derivative of the free energy $c_V = -T\partial^2 f/\partial T^2 \sim \partial^2 \xi^{-d} \partial T^2$. For the KT-transition the exponential divergence of $\xi(T)$ in Eq. (42) is so rapid and occur over such a narrow temperature range that the divergence in c_V cannot be resolved in simulations or in experiment. This is another reason why the vortex unbinding transition remained unnoticed for so long. It doesn't leave any dramatic signature in the thermodynamic quantities. However, as mentioned above, the macroscopic rigidity clearly changes at the transition.

The Vortex Unbinding Transition in Other Systems

We have above mentioned that not only the XY-model exhibits the Kosterlitz–Thouless vortex unbinding transition [22]. Any two-dimensional system that supports thermally induced "charges" or topological defects that interact logarithmically will undergo this transition. The $U(1)$ symmetry of the phase field $\theta(\mathbf{r})$ of the XY-model is also present in the Ginzburg–Landau free energy of superfluids and of superconductors. The topological excitations in the case of a superfluid consist of vortices in the flow of the superfluid. Vortices like those observed when one empties a bath tub. In thin superfluid helium film such vortices destruct the superfluid phase with increasing temperature according to the scenario of the KT-transition [23,24].

The situation is slightly more complicated in superconductors. Because the superfluid in this case is charged (the superconducting pairs of electrons), screening effects play a role [9,10,22]. However, for thin superconducting films of thickness δ the effective screening length is given by $\lambda_{\text{eff}} = \lambda^2/\delta$, which can easily become a macroscopic length. In this case the loss of superconductivity is caused by the unbinding of vortex pairs according to the KT-transition. The broken pairs can move freely when they respond to an applied electric current. As they move they cause phase slips in the superconducting order parameter. These phase slips induce a voltage drop according to the Josephson relation. The superconductor is now unable to support an electric current without a voltage drop, i. e. it is not a superconductor any longer [25,26,27,28].

Dislocations in two-dimensional crystals interact through the strain field. Two edge dislocations of opposite sign correspond to an extra row of atoms inserted along the line connecting the location of the two dislocation cores. The extra line of atoms produces strain and leads to an increase in the energy which is logarithmic in the separation between the two dislocations. Thus, the situation is very similar to the one encountered in the XY-model. When the dislocations unbind, free dislocations are produced. A shear applied to the system can now be accommodated by the mobile dislocations without an increase in the (free) energy. I. e., the shear constant has dropped to zero and the system is melted. The 2d melting theory of Kosterlitz–Thouless–Halperin–Nelson–Young predicts that melting occurs in two stages. At the first stage dislocations unbind and make the shear constant drop to zero. The dislocations are topological defects, their effect on the order of the lattice are, however, not very dramatic. Before the unbinding of dislocations, the translational and the orientational order of the lattice are both described by correlation functions that depend algebraically on distance. When the dislocations unbind the translational correlation function becomes exponential but the orientational correlations remain algebraic. At a somewhat higher temperature topological defects called disclinations unbind with the effect that the orientational order becomes exponential. Details can be found in Chaikin and Lubensky [9].

There are many other cases where the logarithmic vortex interaction and the KT-transition play a role. For instance, the shape of surfaces in three dimensions may undergo a transition from smooth to rough [29]. Assume that the surface energy of the two-dimensional surface is proportional to the area of the surface. And assume that the surface is defined in terms of its height $h(x, y)$ above the xy-plane, i. e. no over hangs. In other words the points on the surface have the coordinates $(x, y, h(x, y))$. The Hamiltonian for the surface is then

$$H = \sigma \int \mathrm{d}x \int \mathrm{d}y \sqrt{1 + (\nabla h)^2} \,. \tag{43}$$

Here σ is a measure of the surface tension. If the height only varies slowly as a function of (x, y) we can assume $|\nabla h| \ll 1$ and then expand the square root. In this approximation the Hamiltonian in Eq. (43) can be written as

$$H = \sigma L^2 + \frac{1}{2} \int \mathrm{d}x \int \mathrm{d}y (\nabla h)^2 \,, \tag{44}$$

(L is the linear size of the system in the xy-plane) which is equivalent to Eq. (19), and we expect the same physical

phenomenology to apply to the surface as we found for the XY-model.

Non-Equilibrium Systems

The Boltzmann probabilities cannot be used to calculate the macroscopic properties when we deal with situations that have no equivalence among systems in contact with a heat bath. At present there is no general procedure for the determination of the probability weights of the individual microstates. This doesn't mean that statical mechanics can't describe how macroscopic properties emerge in systems out of equilibrium. Indeed approaches of broad interest exist. Here we will briefly introduce the use of Langevin equations and algorithmic models through two concrete simplistic models. The first is inspired by the very long memory or correlation times often observed in complex systems. We will discuss slow flowing motorway traffic. The other is concerned with the emergence of self-organized structures among interacting living organisms. Specifically we will think of the formation of ant trails.

1/f Fluctuations – a Langevin Approach

In this section we describe an example of emergence in time. We will assume that interaction between the components of our system forces these to diffuse around, rather than to move around in a ballistic manner. The result is a time signal that contains very strong correlations and is characterized by what is denoted a $1/f$ power spectrum [30].

We want to study correlations in a time signal $f(t)$. We measure the signal again and again at two times separated by T time units. To make life simple we will *neglect* the normalization factor in the empirical averages, i. e. we don't divide by the number of terms in the sum in Eq. (45) below. A justification for this is that we are interested in the functional dependence of the correlations on the time interval T and not so much interested in the actual specific value of the correlation coefficient. Since the correlation coefficient will depend on T we talk about the correlation function. Moreover, since we are correlating the signal with itself we talk about the autocorrelation function given by:

$$\begin{aligned} C(T) &= \sum_t [f(t) - \langle f(t)\rangle][f(t+T) - \langle f(t+T)\rangle] \\ &= \int \mathrm{d}t [f(t) - \langle f(t)\rangle][f(t+T) - \langle f(t+T)\rangle] \\ &= \int \mathrm{d}t [f(t) - \langle f(t)\rangle][f(t+T) - \langle f(t)\rangle] \,. \end{aligned} \tag{45}$$

In the last equality we made use of the fact that the average of value $f(t)$ and $f(t+T)$ are identical.

The autocorrelation function is an important object for the study of memory effects or causality effects in a signal. The correlation function is equivalent to the *power spectrum*. The power spectrum of the signal $f(t)$ is defined as

$$S_f(\omega) = |\hat{f}(\omega)|^2 \,. \tag{46}$$

That is the absolute value squared of the Fourier transform of the signal and the power spectrum is related to the Fourier transform of the autocorrelation function:

$$S_f(\omega) = \hat{C}(\omega) \,. \tag{47}$$

This relation explains why power spectra that approximately depend inversely proportional on the frequency

$$S_f(\omega) \propto 1/\omega^\beta \,, \tag{48}$$

with $\beta \simeq 1$, are of special interest [31,32,33]. Namely, at a somewhat heuristic level, we can substitute Eq. (48) into Eq. (47) and then into

$$C(T) = \int_{-\infty}^{\infty} \mathrm{d}\omega \hat{C}(\omega) \mathrm{e}^{-i\omega t} \,, \tag{49}$$

to obtain

$$C(T) = \int_{-\infty}^{\infty} \mathrm{d}\omega \omega^{-\beta} \mathrm{e}^{-i\omega T} \tag{50}$$

$$= T^{1-\beta} \int_{-\infty}^{\infty} \mathrm{d}u u^{-\beta} \mathrm{e}^{u} \,. \tag{51}$$

We made the substitution $u = \omega T$ and note that the integral in the above equation now is independent of T. So when $\beta \simeq 1$, the correlation function $C(T)$ depends very weakly on T meaning very slow decay of correlations. This indicates the particular interest in power spectra that approximately decays as one over the frequency – called $1/f$ noise. The way we carried the argument through is slightly dangerous due to possible divergent integrals; the conclusion is, however, sound.

Transport by Diffusion

For concreteness imagine a piece of motorway stretching from $x = -\infty$ to $x = \infty$. At $x = 0$ vehicles can enter or leave at an intersection. We will develop a model for the time evolution of the density of cars $n(x, t)$ at position x at time t. Since the cars – particles – only can leave or enter our system at $x = 0$, at all other positions, $x \neq 0$, changes

during a brief time interval δ in the number of particles in a small interval $[x, x+\delta x]$ about x

$$\delta n(x,t) = n(x,t+\delta t)\delta x - n(x,t)\delta x \tag{52}$$

will be caused by a difference during the time δt between the number of particles leaving the section $[x, x+\delta x]$ at $x+\delta x$ and the number of particles entering the section at x. Let $J(x,t)$ denote the particle current (number of particles crossing the position x at time t per time unit). We can then write

$$\delta n(x,t) = J(x+\delta x,t)\delta t - J(x,t)\delta t\,. \tag{53}$$

Substituting Eq. (53) into Eq. (52) we obtain

$$n(x,t+\delta t)\delta x - n(x,t)\delta x = -J(x+\delta x,t)\delta t + J(x,t)\delta t$$
$$\Downarrow \tag{54}$$

$$\frac{\partial n(x,t)}{\partial t} = -\frac{\partial J(x,t)}{\partial x}\,. \tag{55}$$

The last equation follows in the limit $\delta x \to 0$ and $\delta t \to 0$.

This equation is exact and only assumes conservation of the particles. To obtain a closed equation for $n(x,\ t)$ we need to relate $J(x,\ t)$ to $n(x,\ t)$. And to do so we need to make assumptions concerning the nature of how the particles, or cars, move along the line. Let us imagine that congestion makes it impossible for cars to move freely. On the contrary assume that the drivers are forced to effectively perform random walks; i. e. diffuse along the motorway.

We emphasize that it is through this assumption that the cars, or particles, are made to interact. Here we assume that the diffusion is a result of over-crowding and interaction amongst the cars. Obviously particles might perform diffusive motion as a result of other interactions. Pollen in water diffuses because it is bombarded by large numbers of water molecules. In any case some sort of complex interaction is always responsible for diffusive motion since the particles otherwise would move around according to Newton's laws. The model might in fact be more relevant to small particles (pollen, say) suspended in a long narrow strip of water – or something else.

To model the jamming and resulting diffusive motion, we will assume that the net particle current (at coarse grained level) is from high particle density to low particle density, and linear in the density difference. We express this as

$$J(x,t) = -\gamma\frac{\partial n(x,t)}{\partial x}\,. \tag{56}$$

We combine Eq. (55) and Eq. (56) to obtain a closed form of dynamical equation for $n(x,t)$:

$$\frac{\partial n(x,t)}{\partial t} = \gamma\frac{\partial^2 n(x,t)}{\partial x^2}\,. \tag{57}$$

This is the well-known diffusion equation. It describes how inhomogeneities in the density $n(x,\ t)$ relaxes by diffusion. The equation describes a closed system. Next we include the particles that might be added or removed at a certain rate $g(x,\ t)$ at position x at time t. According to the description above we have in particular in mind that $g(x,\ t)$ must describe cars leaving and entering at $x=0$. We will later return to how this particular requirement can be imposed on $g(x,\ t)$. We now have our final equation of motion for $n(x,\ t)$

$$\frac{\partial n(x,t)}{\partial t} = \gamma\frac{\partial^2 n(x,t)}{\partial x^2} + g(x,t)\,. \tag{58}$$

This is an inhomogeneous partial differential equation and we solve it easily by Fourier transformation

$$n(x,t) = \int_{-\infty}^{\infty}\frac{\mathrm{d}k}{2\pi}\int_{-\infty}^{\infty}\frac{\mathrm{d}\omega}{2\pi}\hat{n}(k,\omega)\mathrm{e}^{i(kx+\omega t)}\,. \tag{59}$$

Substitute into Eq. (58) to obtain an expression for $\hat{n}(k,\omega)$ in terms of the Fourier transform of the drive $\hat{g}(k,\omega)$

$$\hat{n}(k,\omega) = \frac{\hat{g}(k,\omega)}{i\omega+\gamma k^2}\,. \tag{60}$$

Now substitute Eq. (60) into Eq. (59):

$$n(x,t) = \int_{-\infty}^{\infty}\frac{\mathrm{d}k}{2\pi}\int_{-\infty}^{\infty}\frac{\mathrm{d}\omega}{2\pi}\frac{\hat{g}(k,\omega)}{i\omega+\gamma k^2}\mathrm{e}^{i(kx+\omega t)}\,. \tag{61}$$

Next we want to focus on the density fluctuations at a specific position $x_0 > 0$. Therefore we define

$$N(t) \equiv n(x_0,t) - \langle n(x_0,t)\rangle_t\,, \tag{62}$$

where we have subtracted the temporal averaged density. We will determine the power spectrum of $N(t)$ and for this purpose need the Fourier transform

$$\hat{N}(\omega) = \int_{-\infty}^{\infty}\mathrm{d}t N(t)\mathrm{e}^{-i\omega t} \tag{63}$$

$$= \int_{-\infty}^{\infty}\mathrm{d}t\ n(x_0,t)\mathrm{e}^{-i\omega t} - \langle n(x_0,t)\rangle_t \int_{-\infty}^{\infty}\mathrm{d}t\ \mathrm{e}^{-i\omega t} \tag{64}$$

$$= \int_{-\infty}^{\infty}\frac{\mathrm{d}k}{2\pi}\frac{\hat{g}(k,\omega)}{i\omega+\gamma k^2}\mathrm{e}^{ikx_0} - \langle n(x_0,t)\rangle_t\delta(\omega)\,. \tag{65}$$

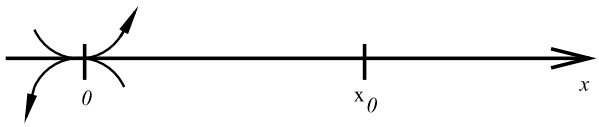

Complex Systems and Emergent Phenomena, Figure 2
Cars/particles diffusing up and down a motorway stretching from $x = -\infty$ to $x = \infty$. At $x = 0$ an intersection allows the vehicles to enter or leave the motor way. At $x = x_0$ a traffic warden is monitoring the number of vehicles, $N(t)$, in front of him

That is how far we can go without further assumptions concerning the nature of the drive $g(x, t)$. Since this source term is meant to represent vehicles entering and leaving at position $x = 0$ we will now use

$$g(x, t) = \delta(x)\chi(t) \Rightarrow \hat{g}(\omega) = \hat{\chi}(\omega) \,. \tag{66}$$

We then have that for $x \neq 0$ the source $g(x, t) = 0$ and at $x = 0$ the temporal variation in the flow onto and away from the "motorway" is given by $\chi(t)$. From Eq. (65) we get

$$\hat{N}(\omega) = \hat{\chi}(\omega) \int_{-\infty}^{\infty} \mathrm{d}k \frac{\mathrm{e}^{ikx_0}}{i\omega + \gamma k^2} \,. \tag{67}$$

The power spectrum is finally calculated as the absolute value square of $\hat{N}(\omega)$

$$|\hat{N}(\omega)|^2 = \frac{|\hat{\chi}(\omega)|^2}{4\gamma\omega} \mathrm{e}^{-\sqrt{\frac{2\omega}{\gamma}}x_0} \,. \tag{68}$$

The power spectrum of the density fluctuations is clearly influenced by the power spectrum of $\chi(t)$. Let us assume that vehicles enter and leave at the intersection in a totally uncorrelated manner (perhaps not a totally realistic assumption) which translates into $|\hat{\chi}(\omega)|^2 =$ constant. In this case

$$|\hat{N}(\omega)|^2 \propto \frac{1}{\omega} \mathrm{e}^{-\sqrt{\frac{2\omega}{\gamma}}x_0} \,. \tag{69}$$

For frequencies so small that $\sqrt{(2\omega)/\gamma}x_0 < 1$ we have $\exp(-\sqrt{(2\omega)/\gamma}x_0) \simeq 1$ and therefore

$$|\hat{N}(\omega)|^2 \propto \frac{1}{\omega} \quad \text{for} \quad \omega < \frac{\gamma}{2x_o^2} \equiv \frac{1}{2T_{\text{diff}}} \,. \tag{70}$$

Where we introduced the time scale $T_{\text{diff}} = x_o^2/\gamma$. This is the characteristic time it takes for particles, under going diffusion with a diffusion constant γ, to move from $x = 0$ to $x = x_0$.

Very long temporal correlations, as indicated by the $1/f$ behavior of the power spectrum, is observed in very many and diverse situations: the light intensity from quasars, the ocean current, the pitch or pressure fluctuations in speech and music, the flow of traffic, the fluctuations in the resistivity of a conductor, and many more [31, 32]. Is the model we have sketched above able to explain the observed $1/f$ correlations in all these many different systems? No, probably not. Surface driven diffusion doesn't seem to be central to all these situations. The question whether a general explanation for $1/f$ exists is still an open one.

We have considered $1/f$ fluctuations here for mainly two reasons. It is a fascinating problem which is often encountered in complex systems and our discussion illustrates how one can use stochastic differential equations to go beyond equilibrium statistical mechanics to analyze temporal emergent behavior.

Self-organized Structure Formation – Ant Trails

As our next example of non-equilibrium statistical mechanics and emergence in complex systems we will briefly consider an algorithmic model inspired by ant trail formation. The model is schematic, simplistic and its relevance to the actual mechanism involved when real ants form trails is not known in detail. Nevertheless, the model is an interesting example of how emergent structures can appear from a dynamical algorithm.

The model describes how the path selected by ants gradually converge towards the *shortest* path between the nest and a location of food, as a result of ants' tendency to walk around at random combined with a preference for following the smell (the pheromone trail) left behind by preceding ants [12,13]. The model discussed here is presented by Danilo Benzatti at http://ai-depot.com/CollectiveIntelligence/Ant.html; but a rich literature on individual based models of formation of patterns by social insects exists [12,13,34,35,36].

We imagine a network of possible paths around the ant nest (see Fig. 3).

The nest is located at one node. Around the nest many edges connect a system of nodes. The ants can only move along the edges and in this way travel from node to node. At some of the nodes food is present. Often more than one

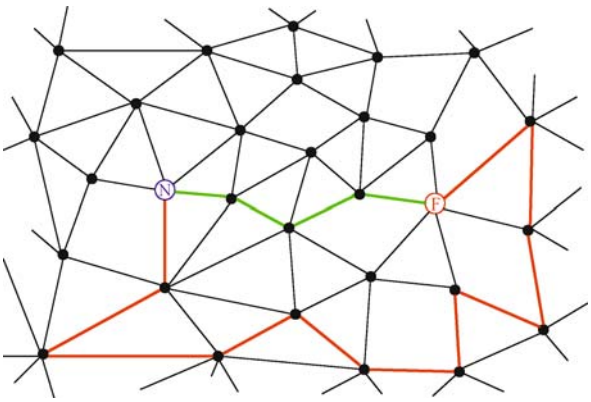

Complex Systems and Emergent Phenomena, Figure 3
A network of possible paths the ants can choose between as they travel from the nest at *N* to the food at *F*. The *green* short path will maintain a high level of pheromones, and will continue to attract ants, while the longer *red* path will gradually sustain a high level of pheromones and will therefore cease to attract ants

possible route will lead from the nest to the food. How do ants, with no global overview and no sophisticated means of measuring traveled distance, identify the shortest route from the nest to a food position? The assumption is that all what an individual ant is able to do is:

(1) **Ant Motion:** Perform random walks.
(2) **Effect of ant on path:** Lay down a unit of pheromone when traversing an edge.
(3) **Feedback:** Feel attracted to pheromones deposited on the edges. The more pheromones the more an ant is attracted to an edge; and the more likely it is that the ant chooses to walk along that edge.
(4) **Return to nest:** When food is found the ant follows its own pheromone trail back to the nest.

The reason this algorithm can converge towards the shortest path, is that the shortest path between the nest and the location of the food will have more ants traveling along its edges per time, and therefore will have a higher pheromone concentration than the longer paths.

The details of the algorithm is as follows. Let the nodes be enumerated by $i = 1, 2, \ldots, N$. Let edges between node number i and node number j be labeled E_{ij}. Each edge carries a time-dependent pheromone weight $\phi(E_{ij}, t)$. When an ant traverses an edge it deposits one unit of pheromone leading to

$$\phi(E_{ij}, t) \mapsto \phi(E_{ij}, t+1) = \phi(E_{ij}, t) + 1 \,. \tag{71}$$

At each node the ants chose probabilistically between the edges sprouting from the node. The probability $p(E_{ij}, t)$ that an ant chooses a certain edge E_{ij} among all the edges connected to a specific node number i is given by

$$p(E_{ij}, t) = \frac{\phi(E_{ij}, t)}{\sum_l \phi(E_{il}, t)} \,. \tag{72}$$

At the start of the simulation all ants are in the nest. Next they begin their random exploration of the network. When an ant locates the food, it picks up a unit of food and returns to the nest along the path it followed on its way out. On the return journey pheromones are also laid down, reinforcing this specific path. Finally it is assumed that the pheromones evaporate at a constant rate, say,

$$\phi(E_{ij}, t) \mapsto \phi(E_{ij}, t+1) = \phi(E_{ij}, t) - \nu\Theta(E_{ij}, t) \,. \tag{73}$$

(Note the Θ-function [defined as $\Theta(x) = 1$ for $x > 0$ and $\Theta(x) = 0$ for $x \leq 0$] ensures that the evaporation stops when there are no more pheromones on an edge, i. e. when $\phi(E_{ij} = 0.$) Thus, paths rarely used will lose their pheromone signature and will not be attractive to the ant, whereas the pheromone level will be maintained along the more often used paths. Thus, the shorter paths with a more frequent ant-traffic will prevail over the longer paths, since the ants on the latter paths visit the individual edges less frequently.

The collective effect of this algorithm is to find the shortest path between nest and food, although the individual ant doesn't need to know that this is what is going on. For more detail see [37] and for modeling of emergent collective intelligence and pattern formation amongst social insects see [12,13,34,35,36]. It is still an open question how to represent algorithmic models, like the one described in this section, in a precise way.

Summary and Future Directions

We have described a specific example where one can follow in great detail the steps from one level of structure to the next in the hierarchical order of matter. Topological defects arise as coherent structures of the "atoms" at one level and can be considered as (composite) particles at the next level. Their interaction can be derived from the behavior of the constituent "atoms". Many different systems may support composite particles that interact in the same way. We looked in particular at vortex physics, where the Kosterlitz–Thouless transition is caused by the logarithmic interaction between the topological defects. The one most important fact in determining the KT-transition is that both energy and entropy depend logarithmically on length scale for the two-dimensional topological charges. This example is hoped to make clear that within equilib-

rium statistical mechanics emergent phenomena can be described and analyzed in great quantitative detail.

When we move to systems out of equilibrium, which of course by far constitute the majority, no universally applicable formalism exists so far. We illustrated, however, by two very different examples that emergent collective behavior produced by interactions between components can also in non-equilibrium situations be modeled either by use of various mathematical techniques or through the application of computer simulations.

I will finish by suggesting the following *conclusion*. There is nothing mysterious about emergent phenomena. They are a wonderful thing – but they are not of new character, something science never has seen or dealt with before. On the contrary, I will claim that understanding emergent phenomena is exactly what all science is aimed at. Often this is not so explicitly clear as in the examples discussed above where the emphasis is explicitly on the macroscopic, or systems level, effects of the interactions between the components constituting the system. Nevertheless, even when one studies, say, atomic physics (as in contrast to statistical physics), one is dealing with the effect of interacting components. An atom consists after all of interacting protons, neutrons and electrons and the properties of the atom are the emergent result of the interactions between these particles.

Notwithstanding, the focus of the statistical mechanics approach is towards generality. As illustrated by our discussion of the XY-model and the Kosterlitz–Thouless transition, the same phenomenology can be observed in many very different systems with very different types of components (magnetic moments, atoms in a lattice, superfluids etc.), if the interactions between the components possess equivalence at a mathematical level. It is this generality that leads people to suggest that even simple model studies may sometimes be of relevance to seemingly much more complicated situations. In the future we are bound to see the statistical mechanics approach to emergent phenomena being applied to a much broader range of problems than was traditionally the case. We see attempts in fields like biology and economics, but also in linguistics, to develop pertinent statistical mechanics models. Examples include phenomena ranging from gene regulation to the social behavior of insect colonies and from stock market fluctuations to management of logistics. So far statistical mechanics has mainly been developed under the influence of physics and methods like the Renormalization Group arose. It will be very interesting to follow how statistical mechanics broadens its arsenal of tools as emergent phenomena in other fields are approached from the view point of statistical mechanics.

Acknowledgments

I am deeply grateful to many people with whom I have collaborated and learnt from during many years concerning matters relevant to the topic of the present article. In particular it is a pleasure to mention Hans Fogedby, Peter Minnhagen, and Hans Weber.

Bibliography

1. Anderson PW (1984) Basiv notions of condensed matter physics. Benjamin/Cummings, Menlo Park
2. Anderson PW (1972) More is different. Science 177:393–396
3. Wilson K (1982) The renormalization group and critical phenomena. Available via DIALOG. http://nobelprize.org/nobel_prizes/physics/laureats/1982/wilson-lecture.html. Accessed 10 Jul 2008
4. Reif F (1965) Fundamentals of statistical and thermal physics. McGraw-Hill, New York
5. Shang-Keng M (1985) Statistical mechanics. World Scientific, Singapore
6. Kadanoff LP (2000) Statistical Physics. Statics, Dynamics and Renormalization. World Scientific, Singapore
7. Binney JJ, Dowrick NJ, Fisher AJ, Newman MEJ (1992) The theory of critical pnenomena. An introduction to the renormaliztion group. Oxford University Press, Oxford
8. Nelson DR (2002) Defects and geometry in condensed matter physics. Cambridge University Press, Cambridge
9. Chaikin PM, Lubensky TC (1995) Principles of condensed matter physics. Cambridge University Press, Cambridge
10. Tinkham M (2004) Introduction to superconductity, 2nd edn. Dover Publications, New York
11. Sornette D (2004) Critical phenomena in natural sciences. Chaos, fractals, selforganization and disorder: concepts and tools. Springer, Berlin
12. Camazine S, Deneubourg J-L, Franks NR, Sneyd J, Theraulaz G, Bonabeau E (2001) Self-organization in biological systems. Princeton University Press, Princeton
13. Solé R, Goodwin B (2000) Signs of life, how complexity pervades biology. Basic Books, New York
14. Kosterlitz JM, Thouless DJ (1973) Ordering, metastability and phase transitions in two-dimensional systems. J Phys C Solid State Phys 6:1181–1203
15. Cataudella V, Minnhagen P (1990) Simple estimate for vortex fluctuations in connection with high-t_c superconductors. Phys C 166:442–450
16. Weber H, Jensen HJ (1991) Crossover from three- to two-dimensional behavior of the vortex energies in layered *xy*-models for high-t_c superconductors. Phys Rev B 44:454–457
17. Minnhagen P, Olsson P (1991) Monte carlo calculation of the vortex interaction for high-t_c superconductors. Phys Rev B 44:4503–4511
18. Kosterlitz JM (1974) The critical properties of the two-dimensional xy model. J Phys C Solid State Phys 7:1046–1060
19. Olsson P, Minnhagen P (1991) On the helicity modulus, the critical temperature and monte carlo simulations for the two-dimensional *xy*-model. Phys Scr 43:203–209
20. Jensen HJ, Weber H (1992) A phenomenological study of vortices in a two dimensional *xy*-model in a magnetic field. Phys Rev B 45:10468–10472

21. Jensen HJ, Fogedby HC (1985) Phonon-kink interference in the φ^4 model. Phys Scr 31:210–214
22. Minnhagen P (1987) The two-dimensional coulomb gas, vortex unbinding, and superfluid-superconducting films. Rev Mod Phys 59:1001–1066
23. Ambegaokar V, Halperin BI, Nelson DR, Siggia ED (1978) Dissipation in two-dimensional superfluids. Phys Rev Lett 40:783–786
24. Ambegaokar V, Halperin BI, Nelson DR, Siggia ED (1980) Dynamics of superfluid films. Phys Rev B 21:1806–1826
25. Kadin AM, Epstein K, Gildman AM (1983) Renormalization and the Kosterlitz–Thouless transition in a two-dimensional superconductor. Phys Rev B 27:6691–6702
26. Huberman BA, Myerson RJ, Doniach S (1978) Dissipation near the critical point of a two-dimensional superfluid. Phys Rev Lett 40:780–782
27. Doniach S, Huberman BA (1979) Topological excitations in two-dimensional superconductors. Phys Rev Lett 42:1169–1172
28. Beasley MR, Mooij JE, Orlandon TP (1979) Possibility of vortex-antivortex pair dissociation in two-dimensional superconductors. Phys Rev Lett 42:1165–1168
29. Barabási A-L, Stanley HE (1995) Fractal concepts in surface growth. Cambridge University Press, Cambridge
30. Milotti E 1/*f* noise: a pedagogical review. Available via DIALOG. http://arxiv.org/pdf/physics/0204033. Accessed 10 Jul 2008
31. Press WH (1978) Flicker noises in astronomy and elsewhere. Comments Mod Phys C 7:103–119
32. Weissman MB (1988) 1/*f* noise and other slow, nonexponential kinetics in condensed matter. Rev Mod Phys 60:537–571
33. Grinstein G, Hwa T, Jensen HJ (1992) $1/f^{\alpha}$ noise in dissipative transport. Phys Rev A 45:R559–R562
34. Schweitzer F, Lao K, Family F (1997) Active random walkers simulate trunk trail formation by ants. Biosystems 41:153–166
35. Bonabeauc E, Theraulaz G, Deneubourg J-L, Franks NR, Rafelsberger O, Joly J-L, Blanco S (1998) A model for the emergence of pillars, walls and royal chambers in termite nests. Phil Trans Royal Soc B 353:1561–1576
36. Feltell D, Bai L, Jensen HJ (2006) An individual approach to modeling emergent structure in termite swarm systems. Int J Model, Identif Control 1:43–54
37. Benzatti D http://ai-depot.com/CollectiveIntelligence/Ant.html

Composites, Multifunctional

Anette M. Karlsson[1], Mosobalaje O. Adeoye[2]
[1] Dept. of Mechanical Engineering, University of Delaware, Newark, USA
[2] Dept. of Materials Science & Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria

Article Outline

Glossary
Definition of the Subject
Introduction
Functionalized Carbon Reinforced Polymer Matrix Composites
Functionalized Composites with Carbon Nanotubes
Composites as Actuators and Sensors
Biomineral Composites
Self-healing Composites
Multifunctional Coatings
Other Multifunctional Composites
Future Directions
Bibliography

Glossary

Biocomposite Biological composite materials made through biological processes usually in living organisms. Biocomposites formed through biomineralization processes are referred to as biomineral composites

Biomimetics Also known as biomimicry, is the study of biological systems with the aim of applying the methods and processes in these systems to the design of engineering materials and systems to produce engineering devices and structures with superior or comparable functionalities.

Biomineralization Mineralization carried out through biological processes to convert organic materials to inorganic materials to form biominerals. Biomineral composites are composed of inorganic minerals formed through biomineralization processes by living organisms in organic matrices of proteins and polysaccharides.

Composite A combination of two or more monophase materials arranged into a material (or "material structure") to produce one or more particular properties that are superior to the same properties exhibited by the individual component materials.

Multifunctionality This is the ability of a material or device to perform two or more functions simultaneously or consecutively.

Self healing The ability of a structure to repair damage without external intervention. For example, a small cut on a human skin will be repaired automatically by the body. Man-made structures with such ability are under early development.

Definition of the Subject

A composite material is a combination of two or more monophase materials arranged into a single entity material (or "material structure") to produce one or more particular properties that are superior to the same properties exhibited by the individual component materials. Until recently, most composites were designed with a single

function in mind, most commonly focusing on improving strength or durability of the material. Many high performance composites have been developed and are performing excellently in this respect. For example, carbon fiber reinforced epoxy is now commonly used in many engineering applications. Carbon fibers have high theoretical strength and high stiffness, but are brittle and therefore readily break if a very small flaw is initiated. By embedding the fiber in epoxy, which is relatively ductile (less flaw sensitive) compared to the carbon fiber, the composite structure combines the strength and stiffness of the carbon fibers with the ductility of the epoxy. A natural extension of this concept is to develop materials that are optimized for engineering applications in conditions that demand more than one function being performed by a single material. For example, a material acting as a structural support may as well function as a thermal sensor at the same time.

A multifunctional composite, therefore, is a composite material that is capable of performing two or more functions simultaneously or consecutively. The need for improved performance in current engineering application and in entirely new areas of application has been the major driving force behind the design and development of multifunctional composites. Multifunctionality is achieved in a composite by deliberately and purposefully engineering the microsructures of the component materials and the composite as a whole usually at the micro- and nano-range levels (that is at the length scales of μm and nm). When the reinforcement scale is in the nano-region the resulting composites are commonly referred to as *nanocomposites*, which is a new frontier in materials science and engineering. Today, much effort is being geared towards research and development of multifunctional composites at various materials research departments and institutions across the world.

Introduction

Multifunctional composites are designed, through microstructural modification, to enhance or introduce new material properties in order to improve or increase the functionalities of a structure, with respect to a broad range of properties. This can include various combinations of magnetic, electronic, electrical, optical, chemical, thermal and mechanical properties. In other words, a multifunctional composite provides two or more functionalities simultaneously or sequentially with improved performance, with less complexity, cost and weight compared to a structure where these functions are provided by individual components. In many cases, a multifunctional composite is the only means by which the combination of the desired functions can be achieved.

Requirements for high performance, durability, conservation of natural resources, low cost, and in many applications miniaturization have led to research and development of multifunctional composites with high specific properties, e. g., high strength per unit weight. These have especially been the driving forces for applications in space exploration, aerospace, information technology and energy production and transmission. Thus, materials science is a critical tool to bring together different materials as composites that can perform more than one function. Such composites are designed and engineered at various scales, ranging from the atomic level and up. For example, load-bearing composites have been developed to simultaneously act as thermal sensors by utilizing carbon nanotube reinforced polymers [109].

A very important multifunctional composite group that has been researched over the past two decades is the group of carbon nanotube composites because of the remarkable mechanical, electrical, thermal and structural properties exhibited by the fiber-like carbon nanotubes reinforcing component of the composites [35]. Carbon nanotubes are added to polymeric matrices to form carbon nanotube polymer composites with excellent mechanical and electrical properties [105]. With the nanotubes the scale of reinforcement is now in the nano region hence these composites are termed nanocomposites, and will be discussed in Sect. "Functionalized Carbon Reinforced Polymer Matrix Composites".

Living organisms are made up of multifunctional materials. An example is the human skin that functions as a container and protector for all internal organs and the human structure. It also serves as a heat sensor, touch sensor, and an outlet for sweat and oil. The human skin, made up of various cells and layers of cells consisting of blood vessels, sensory receptors, glands, and hair follicles, is therefore a natural multifunctional composite. Moreover, human skin has the amazing ability of self healing. Consequently, nature serves as an inspiration for the development of multifunctional materials. A second class of multifunctional natural composites is the biomineralized materials found in living organisms. Biomineral composites are natural materials, a group of bioceramic-biopolymer composites, produced through cell-mediated processes [40]. They are composed of inorganic nano- or micro-scale amorphous or crystalline minerals formed through biologically induced or biologically controlled mineralization processes by living organisms in organic matrices of proteins and polysaccharides [42,116]. Their functions include structural support, mechanical protection, move-

ment, grinding, and gravity or magnetic field sensing. This class of multifunctional composites is of importance in biomimetics and is of interest to both biologists and materials scientists. They are tough materials combining high hardness with high fracture resistance. Examples are bone, dentine, enamel, shells, scales, eggshells and sponge silica skeletons. There are many other biocomposites that are not organic-inorganic such as skin, wood and leaf – these are organic and are also multifunctional.

Many other types of composite materials are now being developed aiming towards multifunctional composites as demands are increasing for smaller, lightweight but smarter products. Reinforcing components could be particles (0-dimensional), fibers (1-dimensional) or plates (2-dimensional). The orientation of these reinforcing components in a matrix could be random, unidirectional or bidirectional, and could be laminar. They are in the form of distinct phases in the matrix with sizes varying from macrophases to nanophases. When the size of the reinforcement is in the nano-range, the composite is referred to as a nanocomposite. When the nanoreinforcements are composed of a functional nanophase the composite is then a multifunctional nanocomposite in which the nanophase provides an advanced functional behavior through enhanced properties such as mechanical, chemical, biological, electrical, magnetic, optical properties.

A Short Note on Carbon

Carbon is a critical, enabling material and consequently an important component of many multifunctional composites. We will, therefore, outline some of the key properties for this element.

Carbon is one on the most common elements found on earth. It appears in three major forms: diamond, graphite, and as various forms of "fullerenes" and "carbon nanotubes." The two former versions occur naturally, whereas the latter is typically artificially produced. Diamond, with a cubic crystal structure where the atoms are arranged in a tetrahedral configuration, is the hardest material found in nature. It is an electric insulator. In graphite, the atoms are arranged in hexagonal planes that are stacked parallel to each other. This layered nature of the structure makes graphite anisotropic and relatively soft, since fracture easily occurs along these layers. In this form, carbon is electrically conductive.

A third form of carbon was categorized in 1985 [54] (even though earlier work indicated their existence) and will play a key role in functionalizing composite materials. This third form is a variant of graphite, where one layer of graphite is wrapped to form a sphere or cylinder. The initial form described was a 60-carbon atom sphere, where the carbon atoms are arranged in 20 hexagons and 12 pentagons on the surface of the sphere. Soon after this discovery, variants of the 60-atom sphere were found, where, in particular, the tubular form, carbon nanotube (CNT) have provided an exciting new research field [50]. There are endless variants of how CNTs can be assembled, but the major categories can be summarized as single- and multi-wall tubes along with their major division by the atomic structure (armchair or zigzag; if arbitrary direction "chiral" see Fig. 1). CNTs are one of the most promising materials for functionalizing composites, due to their unique properties. For example, the armchair styled CNT are metallic, but the other versions are semiconductors. In particular, single CNTs are excellent conductor, and can therefore serve as nano-sized electrical wires. CNTs also have excellent mechanical properties, for example as high as 1 TPa elastic modulus has been reported [48,60]. Multi-walled CNTs consist of either single walled tubes layered

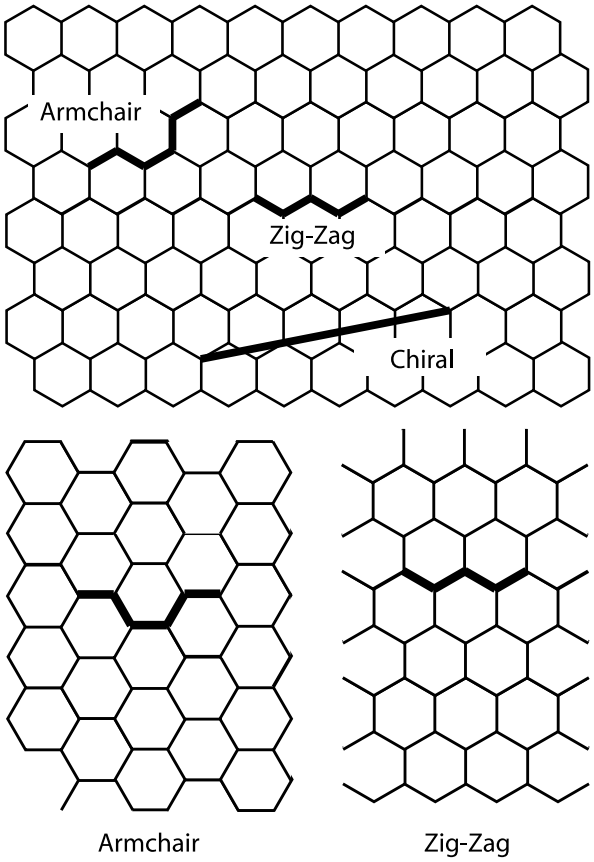

Composites, Multifunctional, Figure 1
Illustration of the various atomic configurations a carbon nanotube can assume. A carbon atom is located in each of the corners of the hexagons

over each other ("Russian doll-style"), or as a continuous, rolled sheet forming a wrapped layered sheet (parchment-style). There are several interesting review papers discussing CNT tubes, for example by Thostenson et al. [106] focusing on mechanical properties, by Gooding [43] focusing on electrochemistry and by Bandaru [5], discussing electrical properties of CNTs. Two recent review papers on theoretical aspects of the thermomechanical properties of carbon nanotubes can be found in [20,21].

Functionalized Carbon Reinforced Polymer Matrix Composites

Carbon reinforced, polymer matrix composite are now commonly used in engineering applications, due to their relatively high strength and low weight. The carbon fibers give the composite their strength and stiffness, for which the structures traditionally have been optimized. However, recent developments make it possible to utilize the ability of carbon fibers to be electrically conductive. This opens up a broad range of opportunities for functionalizing carbon reinforced polymer matrix composites.

Electromagnetic Shielding

Electromagnetic interference (EMI) is caused by electromagnetic radiations emitted from sources that carry electrical current. A common source for EMI is an electronic device, but natural sources such as the Northern Lights and the Sun can also cause EMI. EMI results in anything from annoying "sound sparkles" on the television or cell phone to a malfunction of a device or a system (e. g., an aircraft). In warfare, EMI can be used to disrupt communications by the hostile side. Thus, it is important to shield aircrafts and communication devices from EMI. Carbon reinforced polymer matrix composites can readily be designed to shield EMI. Since carbon is electrically conducting, the carbon fibers can be used to reflect the radiations. By adding conductive fillers into the polymer matrix, the composite material can work efficiently as an EMI shield. It is now commonly used in many consumer products to shield them from EMI, both for protecting the machine and from spreading EMI the machine is generating. For example Bagwell et al. [4] added short copper fibers to increase EMI. A review of EMI shielding can be found in reference [27] and a review of conducting polymer composites that are primarily optimized for EMI shield is found in [101].

Electrified Carbon Fiber Polymer Matrix Composites

Interestingly, research from the groups of Sierakowski and Zhupanska have shown that the material properties in a carbon fiber polymer matrix composite may change when subjected to an electromechanical field [97,98,121]. Early work [98] indicated that the strength, in particular the resistance to fracture and delamination, increases when the composite is subjected to an electric current. Several factors contributes to this, including that (i) the mechanical and electromagnetic fields are coupled when mechanical and electromagnetic loads are imposed simultaneously; (ii) the heat generated in the conducting carbon fibers are transferred to the polymer matrix; and possibly that (iii) the failure mechanisms change when the structure is subjected to an electromagnetic field. Recent work [97,121] where the impact resistance was investigated, show that the gains are short-term. The impact resistance initially can increase as much as 30%, but for a structure subjected to long-term exposure to an electromagnetic field, the gain is reduced back to the initial properties. The losses appear to be caused by the increasing temperature of the polymer due to the heating of the carbon fibers. Nevertheless, a structure can temporarily be strengthened by imposing an electromagnetic field. Moreover, the heat generated in the carbon fibers could potentially be used to activate self healing mechanisms (self healing mechanisms are discussed in Sect. "Self-healing Composites")

Functionalized Composites with Carbon Nanotubes

As discussed in the introductory section of carbon, carbon nanotubes are probably one of the single most promising materials to functionalize composites. The possibilities appear to be endless and a few limited examples will be discussed here.

Single-walled carbon nanotubes (SWNTs) have great promise for functionalizing composite materials. They are light weight and have high mechanical strength, high thermal and electric conductivity and unique optoelectronic properties. They are also light weight, with a small diameter and high aspect ratio. However, these properties may be compromised when incorporated into a polymer matrix. This is primarily caused by the SWNT not being compatible with the polymer matrixes. This results in the SWNTs tending to agglomerate into clusters. When the SWNT are not bonded properly to the polymer matrix and/or appear in clusters, their unique properties may not be transferred to the composite materials. Therefore, significant efforts are being aimed towards improving the dispersion and bonding.

Chen et al. [19] suggest that molecular engineering is a viable approach to achieve good mechanical strength and retain the electric conductivity of CNT. They pointed out

that the problem with dispersion and bonding is due to the smooth surface of the SWNT. In a mix of SWNT and a polymer matrix, the mechanical load can be transferred though mechanical interactions between the SWNT and the polymer matrix, via van der Waals' interactions or covalent bonds, or via special non-covalent bonds, such as hydrogen bonding. The covalent bonding is in general the strongest type of bonding. However, when this is implemented, the electrical and thermal properties are often seriously challenged since these bonds tend to interfere with the SWNT structure [19].

CNTs can be used for energy absorption, which have been shown by Chen and co-workers [22,45,86]. In their work, they developed a solid-liquid composite which combines a non-wetting liquid with a hydrophobic nanoporous solid. The basic premise is that a liquid is absorbed into nanopores (such as the inside of a CNT) when a pressure is applied on the system. They showed that this infiltration absorbs and converts mechanical work into solid-liquid interface energy, with high energy absorption (10–100 J/g). Due to the ultra-high specific surface area of the nanopores, this is several orders of magnitude higher than conventional energy absorption materials. Moreover, by varying the interface energy, the energy absorption performance may be adjusted in a wide range, suitable for damping protections, vibration proof, or blast resistance. The interfacial energy can be changed by using chemical admixtures, or using viscous liquid. If the load rate can be controlled, this can also change the interfacial energy. In addition, by using functional liquids (such as electrolytes), the ion density at the nanopore-liquid interface may be perturbed by external mechanical or thermal fields. Thus, the multifunctional solid-liquid nanocomposite may harvest thermal and mechanical energies into electricity [46,87].

Polyaniline (PANi) is formed by polymerizing aniline (phenylamine, aminobenzene) which is an aromatic amine with the formula $C_6H_5NH_2$. PANi is a conductive polymer, and consequently has great potential to be a useful material component in multifunctional composites. When aniline is polymerized with the presence of multi-walled carbon nanotubes (MWNT), to form a polyaniline-carbon nanotube composite, the MWNT are coated with PANi and form a three-dimensional network within a matrix of PANi. This results in a composite with excellent electro-optical properties [90]. Preliminary work has shown that PANi can also work as a biological sensor [61]. Here the SWNT was wrapped with a single-stranded DNA and mixed with a self-doped polyaniline, poly(anilineboronic acid). The composite is able to detect nanomolar concentration of dopamine (a naturally occurring hormone). The sensitivity for detecting dopamine was increased with a factor of four by adding the SWNT compared to the self-doped polyaniline [61].

Polypyrrole (PPy) is formed from synthesized (connected) pyrrole, where pyrrole is a heterocyclic aromatic organic compound, C_4H_4NH. PPy have been used for corrosion protection of metals, discussed in Sect. "Multifunctional Coatings". In a similar manner as PANi, PPy is conducting. An interesting application is to use PPy with MWNT, where supercapacitive properties have been measured [49]. To achieve this, the MWNTs must be aligned and then coated by an appropriate layer of PPy. Alignment of the MWNTs can be obtained by growing them on a quartz glass under appropriate conditions, described for example by Hughes et al. [49]. Measurement of the charge storage capacity of the aligned-MWNT-PPy composite film show several times charge storage than either PPy or MWNT alone (e. g., 2.55 F/cm^2 for the composite film compared to 0.62 F/cm^2 for pure PPy film) [49]. Thus, aligned MWNTs coated with PPy have potential applications for supercapacitors and batteries, as well as sensors.

Composites as Actuators and Sensors

Ionic polymers have received significant attention during the last decade, based on their ability to work as electromechanical transducer and potentials to work as sensors and/or actuators, and also as electrolytes in polymer fuel cells. Today, the preferred polymer materials are perfluorosulfonic acid (PFSA) polymers, a class of fluoropolymers consisting of a hydrophobic polytetrafluoroethylene (PTFE) backbone attached to hydrophilic sulfonic acid groups (SO_3^-) or carboxylate groups via the fluorocarbon polymer side chains. Upon water uptake, the hydrophobic groups ionize and attach to the hydronium ions (H_3O^+) to provide a conductive path for proton transport, while the polymer network maintains the overall structure of the membrane. Commercial material products include Nafion®[1] membranes and Flemion®[2] membranes. These membranes are also used as the proton exchange membranes (PEM) for fuel cell applications. A PEM functions as a "filter" (an electrolyte), letting protons through the polymer membrane, whereas the electrons are forced to take a path outside the PEM. From the path of the electrons, electric energy can be harvested.

Actuators and sensors can be made by an ionic polymer metal composite (IPMC) [3,8,9,56,57,74,75,76,89],

[1] Nafion® is a registered trademark of E.I. DuPont De Nemours & Co.

[2] Flemion® is a registered trademark of Asahi Glass Group in Japan.

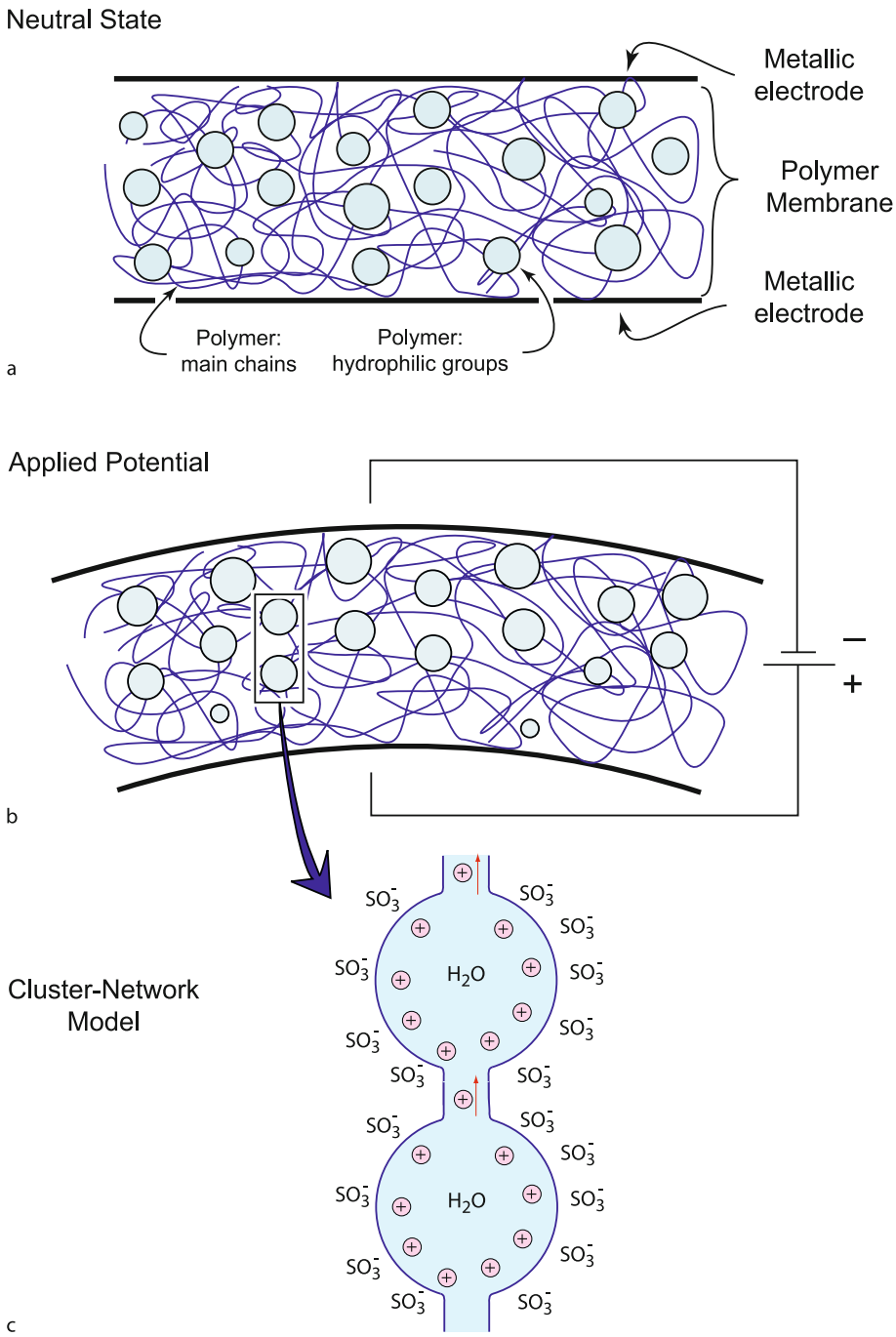

Composites, Multifunctional, Figure 2

A schematic of the principle of an IPMC actuator. a Geometry before a voltage is applied highlighting the morphology of the polymer chains (hydrophobic) and the hydrophilic groups forming a cluster. b The structure deforms when an external voltage is applied. c The cluster-network model for Nafion® membranes, illustrating the cation migration. Adapted from [47,67,76,95]

with a potential application as artificial muscles [3,8,9,10,56,57,67,76,94,95,112,114]. In this case, the ionic polymer is sandwiched between two metallic electrodes. The ionic polymer is hydrated, typically with water or ionic liquids such as salts [3], to achieve its functionality. The metal electrodes are flexible layers, resulting in a soft and flexible actuator which can perform large dynamic deformation if placed in an alternating electric field, see Fig. 2.

Currently, platinum (Pt) and gold (Au) are the preferred electrode materials, which defuses into the ionic polymer membrane, resulting in a material gradient over the thickness of the IPMC. The actuation is governed by the mobile cations (positively charged) moving towards the fixed cathode (negatively charged), resulting in a biased morphology and consequently a bending of the membrane as indicated in Fig. 2. When the current is switched, the location of the cathode is reversed and the cations will consequently move towards the other side, causing the actuator to move in the reverse. Alternatively, the IPMC can work as a sensor, where it generates a voltage if it is suddenly bent.

Nafion® polymer based IPMCs relaxes (reduces its deflection), whereas Flemion® polymer based IPMCs slowly increases its deflection under constant voltage. This is attributed to the mobile cations initially repelling the sulfonic acid groups in Nafion® polymer (which gives a fast actuation), but when the polarization is held constant, the cations relocates slowly, relaxing the IPMC. The carboxylate in Flemion® polymers are weaker in polarizing the structure, and therefore the relaxation is not observed [75,76].

These actuators can strain up to 3% for voltages less than 7 V [67]. This induces a significant bending, where stresses up to 30 MPa are reported [67]. Future research efforts are focused on reducing the negative effect of dehydration (an ionizing liquid is needed for the function), as well as addressing the reduced efficiency over time. Current applications range from fins to robotic fish to artificial eyes [67].

Biomineral Composites

Nature has presented us with a variety of biological composites (biocomposites) such as skin, bone, shell, leaf and wood, which are all multifunctional in living organisms. Some are organic in nature, e. g. leaf and wood, whereas some are organic-inorganic, e. g. bone. Attempting to mimic nature, the design of these biocomposites are now intensely studied, including the structure and microstructure and the physical, chemical, electrical, magnetic and even optical properties. For example, using a biomimetics process M. C. Chang and co-workers have been able to develop a synthetic bone [17,18].

Biocomposites formed through a biomineralization process are referred to as biomineral composites. Examples of biomineral composites include bone, dentine, enamel, mollusk shells, crustacean exoskeletons, eggshells, sponge silica skeletons, and a variety of transition metal minerals produced by different bacteria (see references in references [42] and [33]). The functions of the biomineral composites include structural support, mechanical protection and movement, anchoring (to another body or to ocean floor) grinding, filtering, gravity or magnetic field sensing, optical and piezoelectric [33]. Biomineral composites are composed of an organic matrix of proteins, lipids and polysaccharides. The structure consist of a nano- or micro-scale amorphous or crystalline minerals formed by a biologically induced or controlled mineralization processes, through complex chemical interactions between organic and inorganic matrices [2,42,91,116]. The structure is usually complex with the organic and the mineral components tightly interwoven at the nanoscale level, highly ordered and hierarchical to give high strength, rigidity along with mechanical and chemical stability, that are superior to synthetic materials made from the same materials. Biomineral composites are normally designed to function under a narrow range of environmental conditions, such as narrow temperature regimes and restricted mechanical loads. However, the compositions of biocomposites can be altered over time when a gradual change in environmental conditions occurs, to achieve necessary properties for the survival of the organisms [63]. Mollusks and sponges are known to make use of sophisticated biomineralization mechanisms to obtain structures that exhibit attractive combinations of strength, stiffness, resilience, and energy absorbing capabilities [63]. Even though the mechanisms of biomineralization are yet to be fully understood [69], biomineral composites are of much interest to warrant in-depth studies by biologists and chemists as well as material scientists. The structures of a few biomineral composites that are multifunctional are briefly described below.

The crustacean exoskeleton is a layered structure made up of the epicuticle layer, which is the topmost layer, the exocuticle layer and the endocuticle layer, which is the innermost [25]. The epicuticule is a relatively thin layer of about 2–4 µm. It is waxy, acting as a diffusion barrier [77]. The exocuticle is about 150–180 µm in *Homarus americanus* (American lobster) claw with the endocuticle 3–4 times this thickness. As a comparison, for the *Callinected sapidus* (Atlantic blue crab) claw, the exocuticle is about 40–50 µm and the endocuticle 6–8 times thicker [25]. The exocuticle and the endocuticle are the major load bearing structures of the exoskeleton and are made up of multiple fibrous layers arranged parallel to the surface. The fibrous layer consists of chitin-protein (chitin is a biological polysaccharide with the generic formula $(C_8H_{13}O_5N)_n$) fibrils bonded by a matrix of minerals and other proteins. Each of these fibrous layers is rotated by a small angle relative to the next layer in parallel, building up to a band of

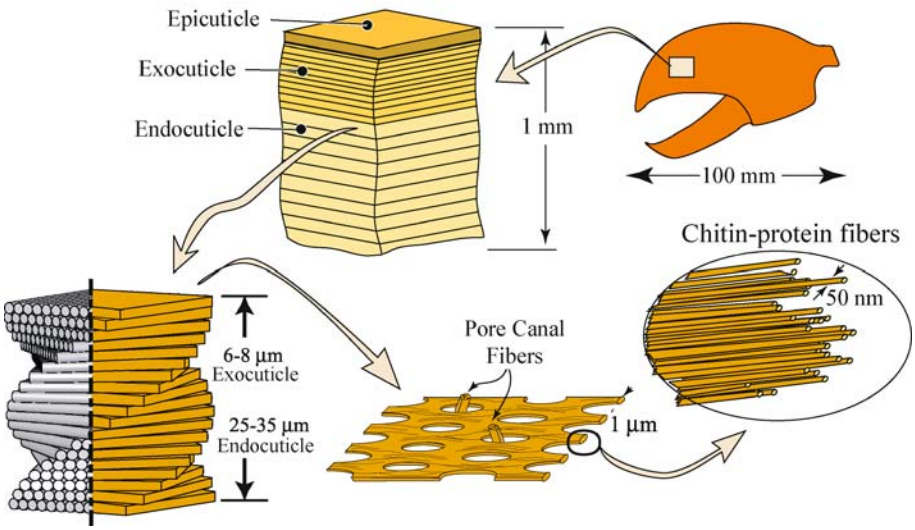

Composites, Multifunctional, Figure 3

The cross-section (*top*) of the exoskeleton of a Homarus americanus (American lobster) taken from the claw (*top right*). The outer layer (epicuticle) acts as a diffusion barrier while the exocuticle and endocuticle layers are load bearing structures built of mineralized fibrous chitin protein. The helicoidal nature of the arrangement of the fibrous layers of the exocuticle and the endocuticle layers is shown on the *left*. The pore canals and the pore canal fibers in a layer are shown (*bottom*). Each layer is composed of chitin protein fibers (*bottom right*). Note that the exocuticle layer is denser than the endocuticle layer

layers that is twisted by 180° to form a helicoidal architecture (Fig. 3). The exoskeleton has through-the-thickness holes, *pore canals*, through which chitin-protein macrofibrils fibers, *pore canal fibers* run perpendicularly to the layers. Even though some disagreement exists in the literature, the pore canal fibers appear to run from the bottom of the endocuticle to the top of the exocuticle [25,28]. The pore canals and the pore canal fibers fill important functions in building the exoskeleton after molting. Moreover, Cheng et al. [25] showed that the pore canal fibers are important for strength of the exoskeleton. In all, the multiscaled structure of the exoskeleton, a biomineral composite, provide the crustacean with a strong structural support, an impervious defense covering for the body of the crustacean, and also serve as a carrying, holding and tearing tool in case of attack or feeding.

Nacre is another natural biomineral nanocomposite; it is also known as mother-of-pearl. It is the iridescent lining on the inside of the shells of many sea-going bivalves and gastropods such as oysters, mussels and abalones. Like many other biomineral composites, nacre has a hierarchical structure. It is composed of about 95% inorganic hexagonal platelets of aragonite (a crystallographic form of $CaCO_3$) 5 to 8 µm wide and 0.2 to 0.5 µm thick [6], arranged in a continuous parallel lamina in 5% organic matrix composed of elastic biopolymers (such as chitin, lustrin and silk-like proteins and polysaccharides). The organic biopolymer is typically 5 to 20 nm thick. Nacre has received significant attention in recent years due to its high ductility, enhanced toughness and fracture strength, along with its low weight, resulting in excellent specific properties. Its fracture resistance is about 1000 to 3000 times greater than that of its component aragonite crystals [30,31]. Its high toughness is as a result of the ductility of the organic matrix in connection with the repeated unfolding of the protein molecules. The nanostructure resembles a brickwork arrangement with a significant overlap of the platelets and the organic matrix serving as the mortal [15]. This architecture is a critical factor that is responsible for the high fracture strength observed in nacre [6,41]. In addition, Li et al. [59] showed that the rotation of the nano-sized grain during loading is a key contributor to the high ductility. Many studies have been carried out on nacre using various experimental and modeling techniques to study its formation, its structure and morphology, and its deformation and properties, especially the mechanical properties [6,16,52,55,62,63,65,73,85].

Mimicking this material is of interest in the design of high performance materials such as impact resistance armor [7,55,92]. Using layer by layer assembly technique, Podsiadlo et al. [84] have been able to prepare a nanostructured analogue of nacre from nanometer sized sheets of montmorillonite clay and a polyelectrolyte. Artificial nacre

has also been synthesized [103]. Mimicking nacre in creating one nano-layer of material at a time, Nicholas Kotov and his team have evolved a process that allows the creation of materials one nano-layer at a time. They use this process to produce a new material from clay nanosheets and a water-soluble polymer that shares chemistry with white glue. The material is transparent, very strong, yet lighter in weight [64].

Animal skeletons are made up of bones which are hard and rigid tissues. Bone is self-healing and can continuously regenerate itself. Moreover, bone is relatively stiff and tough, and can withstand and adapt over time to local stresses. These properties make bone a reliable biological structural material. Bone is considered as a nanocomposite of minerals and proteins [18,26]: It is composed of a matrix impregnated with calcium carbonate, calcium sulphate and small amounts of sodium and magnesium. The matrix, consist of collagen fibers impregnated with crystals of hydroxyapatite, $Ca_5(PO_4)_3(OH)$, and water. Among the functions of the bone is support, locomotion, protection for soft and delicate organs (like the skull protecting the brain), manufacturing of blood cells and homeostasis. An example of bone is the femur; it has a covering of a tough, strong membrane, called periosteum which is richly supplied with blood vessels. Next to the periosteum is a layer of compact bone and bone forming cells (osteoblasts), arranged in concentric layers (lamellae) with round small interconnecting canals (the Haversian canals) that contain blood vessels, nerves and lymph vessels. Embedded in this hard bone matrix are osteocytes, which are associated with bone deposition and bone remodeling. Inside the compact bone is a thin soft membrane known as the endosteum that encloses the marrow cavity that contains soft tissues, the yellow marrow [88,108]. The growth and strengthening of bone is stimulated by mechanical stresses through strain detection. This function is carried out by specialized cells within the bone that are sensitive to and respond to strains. In the absence of mechanical stresses bone becomes weak and less developed. Thus, the more the bone performs its intended functions the stronger it becomes. Mimicking this ability will produce excellent synthetic engineering materials for structural or load-bearing applications.

In a study of the sponge *Euplectella sp.*, Aizenberg et al. [2] described the structure of its skeleton as being hierarchical in nature. It has a layered arrangement that gives it a high resistance to crack propagation. The skeleton is thus composed of a layered biocomposite material. The structure is described in Fig. 4, after Aizenberg et al. [2]. The microstructure is made up of consolidated hydrated silica nanoparticles forming sets of concentric rings glued together by an organic matrix to form spicules. This layer approach provides toughness and resistance to crack propagation. These spicules are then assembled in parallel into bundles within a silica matrix to form struts. The struts are arranged to make the cylindrical cage with the ability to resist tensile and shearing stresses and a significant capacity for recovery after deformation of the skeleton.

Aizenberg et al. [1] showed that the tunic spicules of the ascidian *P. pachydermatina* are a biocomposite material with well-defined domains of both amorphous and crystalline calcium carbonates, separated by an insoluble organic layer. The crystalline calcium carbonate is polycrystalline calcite, and it forms around the amorphous calcium carbonate which serves as the core. The calcium carbonate layers contain magnesium and proteins with a higher content in the amorphous than in the crystalline. The amino acid compositions of macromolecules associated with the two mineral phases are also quite different.

Wood is a natural biocomposite material with a multifunctional capability. It differs from the above mentioned bicomposites in that it is *not* a biomineral composite. Wood is a naturally hard and tough biocomposite material that forms the trunk or stem of trees. The material consists essentially of elongated hollow cells that carry nutrients from the roots to the leaves. The cells make up the cellulose fibers that are arranged generally in the grain direction, parallel to the surface of the trunk. Cellulose, with a generic chemical formular $(C_6H_{10}O_5)_n$, is one of the major chemical constituents of wood, others are hemicellulose and lignin. It is a linear polymer with thousands of mers in a single molecule and it constitutes about 40 to 50% of wood. The cross section of wood is composed of several layers: the outer bark, the inner bark, the cambium, the sapwood, the heartwood, and the pitch. The thickness of each layer depends on the age of the tree, the species to which it belongs, and on the particular tree. The cambium layer is microscopically thin and it grows by cell division to increase the diameter of the trunk. The tree trunk increases in diameter by addition of new peripheral growth layers that constitute the growth rings. The sapwood layer conducts moisture, minerals, oxygen, and nitrogen. As the stem (or trunk) grows in diameter, the sapwood progressively forms the heartwood. The heartwood is the thickest of the layers and it is the one that provides the structural strength. It is usually darker in color because of the mineral deposits, gums and resins that are present in it. Cutting across these layers horizontally are tissues called wood rays radiating out from the center outward. They help in storing and transferring nutrients. Botanically, woods are classified as softwoods and hardwoods depending on their basic cellular structure and on how moisture moves within

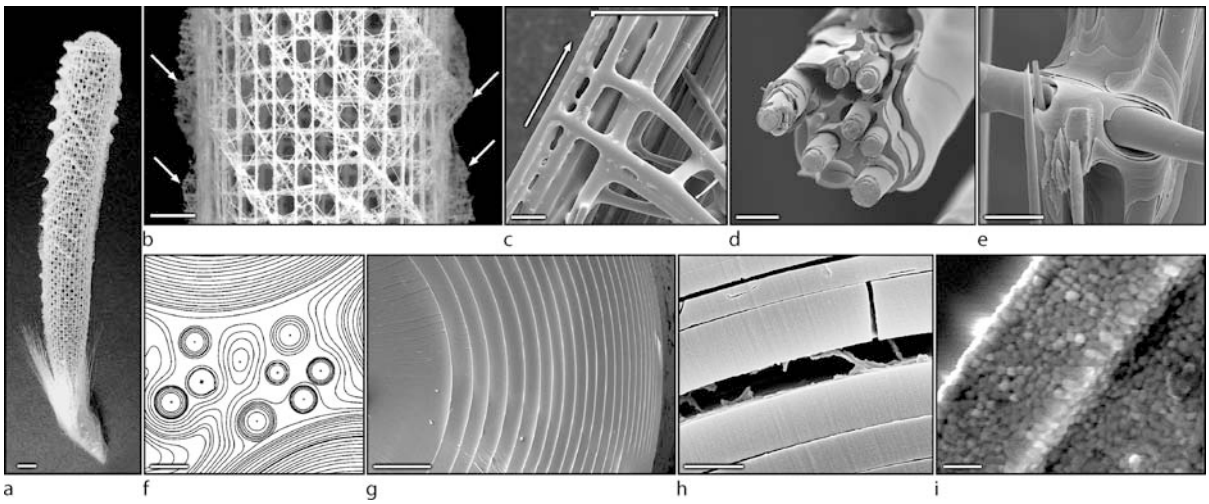

Composites, Multifunctional, Figure 4

Structural analysis of the mineralized skeletal system of Euplectella sp. a Photograph of the entire skeleton, showing cylindrical glass cage. Scale bar: 1 cm. b Fragment of the cage structure showing the square-grid lattice of vertical and horizontal struts with diagonal elements arranged in a chessboard manner. Orthogonal ridges on the cylinder surface are indicated by *arrows*. Scale bar: 5 mm. c Scanning electron micrograph (SEM) showing that each strut (enclosed by a *bracket*) is composed of bundled multiple spicules (the *arrow* indicates the long axis of the skeletal lattice). Scale bar: 100 µm. d SEM of a fractured and partially HF-etched single beam revealing its ceramic fiber-composite structure. Scale bar: 20 µm. e SEM of the HF-etched junction area showing that the lattice is cemented with laminated silica layers. Scale bar: 25 µm. f Contrast-enhanced SEM image of a cross section through one of the spicular struts, revealing that they are composed of a wide range of different-sized spicules surrounded by a laminated silica matrix. Scale bar: 10 µm. g SEM of a cross section through a typical spicule in a strut, showing its characteristic laminated architecture. Scale bar: 5 µm. h SEM of a fractured spicule, revealing an organic interlayer. Scale bar: 1 µm. i Bleaching of biosilica surface revealing its consolidated nanoparticulate nature. Scale bar: 500 nm. (Taken from reference [2] with permission from AAAS)

the living tree. Softwoods are mainly made up of long cells of between 3 and 5 mm called tracheids. Hardwoods, on the other hand, are mainly made up of two kinds of cells, wood fibers (0.7 to 3 mm long) and vessel elements (with wide ranging lengths).

The important physical properties of wood are moisture content, permeability, shrinkage, density. These give it the multifunctional capability such as serving as a super-structure, acting as a nutrient storage and transport medium, the ability to withstand harsh weather, and self-healing. The properties, however, vary greatly across species and also depend on factors such as the age of the tree, stem form, type of soil and climate. Wood is anisotropic with the mechanical properties varying across the growth rings and along the height up the tree. The mechanical, electrical and thermal characteristics of wood make it a popular excellent engineering material over ages.

Leaf is an organic biocomposite that is flat, broad and thin. It is a plant organ in which photosynthesis is carried out. The upper surface of the leaf is waxy for the purpose of water-proofing. It performs functions such as converting sunlight to chemical energy in the mesophyll, transporting glucose, water and minerals through out the plant by the vascular bundle; it is water-proof and provides shade for the tree. The cross-section is made up of different layers in this order from the top: upper cuticle, upper epidermis, palisade mesophyll, spongy mesophyll, lower epidermis and lower cuticle. Embedded in the mesophyll layers is the vascular bundle (phloem and xylem) and air spaces for the supply of air (carbon dioxide) and moisture that comes in through the stomata that dotted the lower epidermis through the lower cuticle. The broadness of leaf allows it to gather as much sunlight as possible as a supply of the energy needed for photosynthesis. Leaf provides a system that could be mimicked in designing materials for energy conversion and at the same time distributes the product.

Biocomposites materials produced by nature have properties that could be beneficial when reproduced in synthetic materials. The design, manufacture from simple raw materials, economical use of raw materials and energy, multifunctionality and degradability of biocomposites are inspirational to biomimetics or biomimicry in the design and manufacture of synthetic engineering materials. Another inspiration from nature is the building from bottom up, from atomic or molecular level to the macro structural level. This provides for efficient use of raw ma-

terials. A thorough study of nature's biocomposite materials could, therefore, yield viable procedures and techniques for the design and manufacture of synthetic engineering materials with excellent combination of properties that will provide for multifunctionality in them.

Self-healing Composites

Biological systems have an outstanding ability in self-healing; that is, automatically detecting and repairing damaged tissue. The repair is made by a material similar to the original (causing a scar tissue), or identical tissue (leaving the damage area undetectable after repair). For humans, the latter can for example be observed in bone, whereas the former on skin. Moreover, biological systems can adapt to new conditions. Humans build more muscles and bones if we increase our daily exercise regime and a tree grows branches to find the most sunlight.

Man-made structures traditionally do not have these features. If damage occurs, damage will continue to accumulate until it is detected via human intervention or until the structures fails. Today, there are many sophisticated means of detecting failures. Even though tremendous advances have been made towards detecting and repairing damage in structures, *self-healing materials* would significantly improve the reliability of structures. During the last decade, significant advances have been made towards developing self-healing composites. The current approaches results in a "scar tissue" in the sense that the original material is not exactly reproduced, but that the structure will function satisfactory. Two materials systems will be discussed here; (i) ceramic matrix composites and (ii) polymer matrix composites.

Ceramic materials typically have high strength and stiffness, and retain their strength even at high temperatures. Since they are brittle and therefore are considered unreliable (tend to break without warning signs that metals exhibit, such as plastic deformations, especially under tensile stresses) their uses have been limited. To negotiate the brittle response, ceramics can be reinforced. Most commonly, ceramics are reinforced with a second ceramic, forming a ceramic matrix composite (CMC). The bonding between the reinforcement and the matrix is a key parameter that governs the toughness of the CMC. By controlling the interface material, referred to as "interphase", the interfacial bonding can be optimized, allowing for the load transfer between the matrix and the fibers, deflects matrix cracking and serve as a barrier towards diffusion. When considering the strength of the interface material, it must be optimized to be "just right," not too strong and not too weak.

Self-healing of CMCs can be achived through high temperature oxidation in silicon carbide based CMCs. Silicon carbide (SiC) reinforced with SiC fiber (SiC/SiC composites) or carbon fibers (SiC/C composites) is mechanically durable up to 1500°C and is therefore considered a promising composite for high temperature applications such as engines and gas turbines [70]. Even though the fibers are brittle in a similar manner as the matrix, the fibers and matrix work synergistically to create a ductile material. Micro-cracks develop during loading, but these micro-cracks are stopped by the microstructural features, simulating a non-linear response paralleling yielding in metals. When silicon is subjected to elevated temperatures, it quickly forms a protective coating of silica (SiO_2). Unfortunately, the micro-cracks that are associated with the CMC become pathways for oxygen to penetrate into the structure, causing internal oxidation. This can weaken the structure. By coating the fibers with carefully selected materials, the oxidation can be controlled and result in self-healing of the micro-cracks. Graphitic carbons ("pyrocarbons," "PyC") and boron nitride (BN) have emerged as the most prominent interphase [39,44,70,71,72]. When oxygen diffuses through the micro-cracks, a fluid oxide is formed due to the oxidation, filling the cracks, Fig. 5. These glassy oxides that form can be optimized through the interphase. For example, when a borosilica glass is used as an additional coating on the fibers, no loss in composite strength was observed after 200 h at elevated temperatures [39]. Thus, a self-healing mechanism in silicon carbide (SiC) matrix composite reinforced with SiC or carbon fibers has been observed, which is caused by oxidation at high temperatures. The oxidation occurs at temperatures above 800°C. The self-healing can continue until the reducing material has been consumed.

Polymer composites might be the most promising systems where self-healing mechanisms can be developed. There are several reasons for this. Polymer based systems are in general less expensive than ceramic based systems, and tends to be easier to work with. All self-healing ceramic systems are based on activating the healing process through subjecting the material structure to heat. Even though some self-healing polymer systems are based on heating, the temperature regime for healing polymer is significantly lower than for ceramics, thus simplifying the process. Furthermore, many self-healing approaches to polymer systems are not dependent on heating. Lastly, several different approaches for self-healing of polymer have been developed so far, thus inviting alternative approaches for self-healing.

One simple concept of healing a damaged structure is to subject the material to elevated temperatures as was

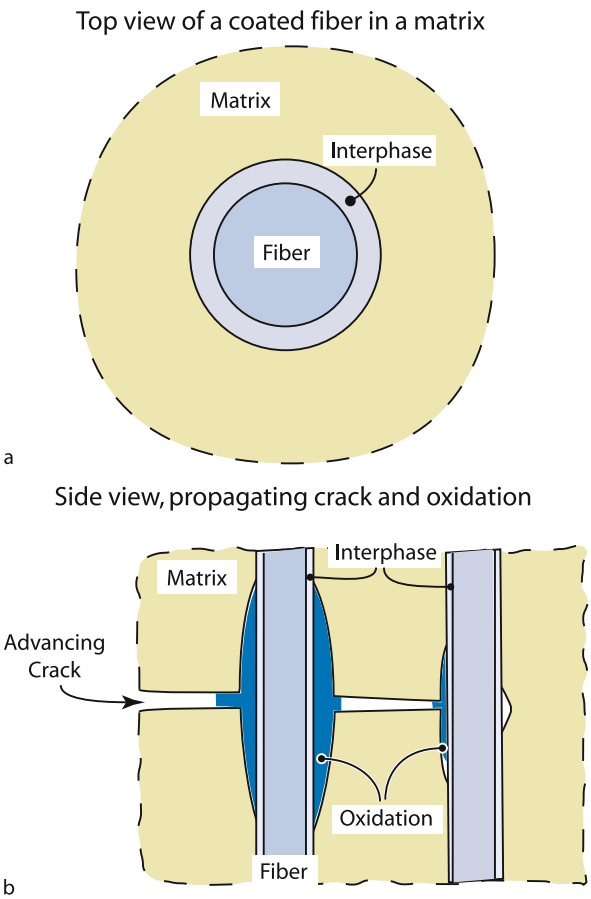

Composites, Multifunctional, Figure 5
Examples of interphases used in SiC matrix composites a schematic top view of a fiber with a coating (interphase) in a matrix b schematic side view of an advancing crack exposing the interphase that oxidizes, resulting in self-healing

seen for ceramics. This idea is based on that the elevated temperature activates a chemical process that results in healing. For example, by utilizing a particular thermally reversible reaction (e. g., a selected Diels–Alder reaction) for cross linking linear polymer chains, a fractured material can be healed, as demonstrated by Chen and co-workers [23,24]. For one material system investigated, after fracture and subsequent heat treatment, the original strength was regained. In this case, the polymer (a furan-based monomer synthesized with 1.8-bis(maleimido)-3.6-dioxaoctane) was subjected to mechanical testing leading to fracture followed by heat treatment at 130°C for 30 min. In this case, there was no loss in the original strength when the structure was re-tested [24]. This approach to self-healing has the clear advantage that no additional material is needed to be added: the material is intrinsically self-healing. However, only a very limited set of materials that can self-heal by cross-linking the polymer chains upon reheating. An alternative approach for healing via heating utilizes an additional material phase incorporated into the original material. When subjected to sufficient heating, this additional material is activated, and can mend the damage structure. For example, a composite made of glass fiber reinforced epoxy retained its stiffness after the heated repair [120]. By adding a heat-activated material to the composite structure, a design engineer would not be strictly limited to a narrow set of materials. However, a major disadvantage with both of these heat activated healing methods is that the healing is not automatic; rather, the structure needs to be treated in a separate process. (This may be differentiated from the ceramic self-healing that was discussed above. The ceramic is operating at the temperature at which oxidation occurs, but this polymer operates at temperature lower than where self-healing appears.) Nevertheless, depending on the application, this approach can be quite useful.

A more convenient approach to self-healing of a structure compared to the heat activated systems described above is a system that heals itself without active interference. Most approaches aiming to achieve this are based on introducing one or more phases into the composite material. These additional material phases are automatically activated when damage occurs. A successful approach have been developed by White, Sottos and co-workers, where spherical microcapsules containing a "healing agent" and a second phase containing a catalyst are embedded in a polymer matrix composite [12,13,14,53,115]. When a crack propagates, the micro capsules in the crack path burst and release their healing agent into the crack, Fig. 6. As the healing agent fills the crack, it will eventually contact the catalysts. When this occurs, the healing agent will polymerize, filling the crack and effectively healing the crack. In the work by White, Sottos and co-workers, the agent was a dicyclopentadiene (DCPD) monomer and the catalyst a bis(tricyclohexylphoshine)benzylidine rethenum (IV) dichloride (a Grubbs' catalyst) [12,13,14,53,115]. This results in a ring-opening metathesis polymerization (ROMP) of the DCPD, resulting in a highly cross-linked polymer. When stabilized with 100–200 ppm p-tert-butylcatechol, the DCPD has a long life and healing can be achieved even for aging structures. The micro capsules containing the DCPD ranged from 40–240 µm and were made with poly-ureaformaldehyde. In early work, curing for 48 h was required to retain 45% of the initial strength (if the curing occurred at 80°C, up to 80% of initial strength was achieved) [12,115]. Subsequent studies showed that 10 h were sufficient to achieve full polymer-

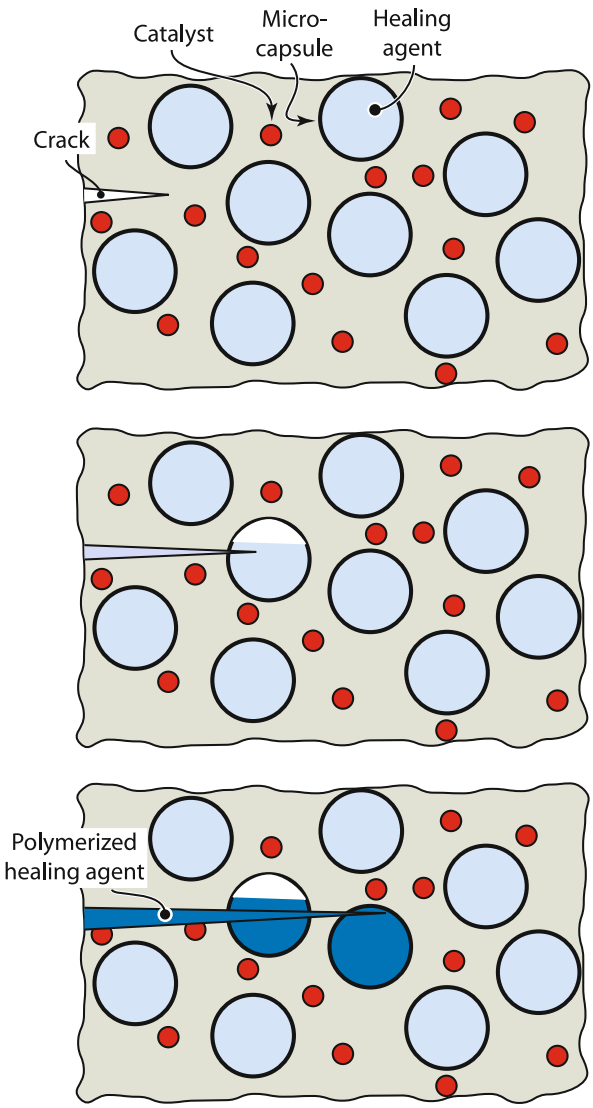

Composites, Multifunctional, Figure 6
Schematic of a propagating crack in a polymer with micro-capsules filled with a healing agent

ization (full healing or full strength), and that the fatigue life can be increased with over 200% if the structure is allowed to rest sufficiently for the healing agent to polymerize [13,14]. Most engineering structures are allowed to "rest" between operations. For example, cars are normally used for commuting to work and get sufficient time to "rest," both day and night, whereas airplanes are scheduled for regular maintenance that keeps them grounded for many hours that may be sufficient for the polymerization to take place. Thus, this system is a promising approach towards extending the lifetime of polymer matrix composites.

The micro-sphere approach has the clear advantage of being possible to be incorporated into a range of materials, and that no particular treatment is needed to activate the healing processes. There are, however, some drawbacks. These include that the shell of the micro-capsules have to be designed so that it breaks when a crack has developed in the bulk material, and that the healing agent comes across the catalyst. Moreover, the up to 200 μm diameter spheres can possibly interfere with the reinforcement of the polymer, including introducing an unwanted waviness of the fiber reinforcement. The latter drawback can be addressed by replacing the micro-spheres with hollow micro-cylinders [11,36,68,82,83]. The current state-of-the-art for hollow micro-cylinders focuses on using commercial hollow glass fibers embedded in composite materials [11,82,83]. In a similar manner as to the case of micro-spheres, the hollow cylinders are filled with a "healing agent" that is activated once the fiber breaks. A two-phase epoxy system is used, where the epoxy resin is stored in one set of cylinders and the hardening agent is stored in a second set of cylinders. In a layered composite material, the hollow glass fibers are aligned with the reinforcement fibers, for example the fibers with the epoxy resin are aligned with the 0°-ply and the fibers with the hardening agent with the 90°-ply. When cracks develop and propagate, the glass tubes break, allowing the epoxy and the hardening-agent to fill the damaged zone. The materials are selected so that the epoxy cures at ambient conditions. The major challenge with this approach relates to the difficulty of finding suitable glass tubes. Ideally, the properties of the hollow glass tubes should match that of the original reinforcement, so they can replace or enhance the composite structure. To address this, Pang and Bond [83] purchased commercial borosilicate glass tubing and using in-house facilities drew the fibers to external diameter of 60 μm and inner diameter of approximately 42 μm. The fibers were filled with a commercial epoxy repair agent (MY750 Ciba–Geigy) and the corresponding hardening agent respectively [83]. In this case, about 90% of the strength of the original strength is retained after repair, but the strength degrades with time.

With the exception of the reversible cross linking polymers, the repair schemes discussed so far for polymer composites are all based on one-time repairs; once a micro-capsule or micro-fiber breaks the healing agent is consumed, and no further healing will occur if another crack should develop again at the same point. In contrast, self-healing in biological systems can occur multiple times for repeated injuries, assuming a reasonable frequency of injuries. In animals, this is possible by the continuous flow of an intelligent mixture of biochemicals in the vascular network, which is related to the circulatory system. Some at-

tempts are made to mimic vascular network for self-healing [104,107], where a constant supply of healing agent could potentially be provided. There are several manufacturing issues involved here, and opens up many potential research avenues.

Multifunctional Coatings

Applying coating on a structure is many time a cost effective way of obtaining a multifunctional composite material. There are many examples that illustrate this, for example environmental barrier coatings, coatings for increased wear resistance, and thermal barrier coatings. Most coatings combine several functions by having multiple layers where each layer contributes a particular function.

Coatings for corrosion resistance (sometimes referred to as environmental coatings) are probably one of the most common classes of coatings, for example covering steel with zinc to obtain a galvanic protection, (the zinc is sacrificed to protect the underlying steel). Even though functional, these cannot be categorized as multifunctional and will not be discussed here. More advanced coatings are now being developed to protect steel and other metals. Some of these coatings have multifunctional capacities. Some particular interesting coating materials are electropolymerized polymer composites including polyaniline (PANi) [51,78,93,102] and polypyrrole (PPy) [51,102]. Polyaniline (PANi) is formed by polymerizing aniline (phenylamine, aminobenzene), $C_6H_5NH_2$ and polypyrrole (PPy) is synthesized pyrrole, C_4H_4NH. The polymers are typically deposited through an electrochemical synthesis in which the thickness can be controlled. The corrosion resistance depends on the deposition parameters including applied potential and the feeding rate of the monomers. By producing a composite of PPy and PANi, the corrosion rate can be reduced with more than two order of magnitudes compared to unprotected materials [51]. To achieve this improved rate, it is crucial to ensure a proper bonding of the structure of the polymer composites deposited on the metal, which must be controlled through the processing parameters [102]. This class of coatings protects in sulfuric acid (H_2SO_4), not so well in hydrochloric acid (HCl), and not at all in a Sodium chloride (NaCl) solution [93]. Thus, care must be taken when using this type of coating for corrosion protection, but evidently, it can be quite useful for a range of applications.

A second class of important coatings is coatings used to ensure low friction and increased wear resistance of the underlying structure, tribological coatings. These coatings are critical for a range of applications, including moving contacts (e. g., bearings), materials processing (e. g., drilling), and applications where addition of lubricants or materials debris from wearing is unacceptable (e. g., food processing, medical implants). Also, by reducing friction in moving parts in vehicles, the fuel efficiency of the vehicle can be significantly increased. When optimizing a coating for wear resistance, the goal is to reach as high hardness as possible [110]. When combining wear resistance with low friction, many other aspects much be considered. There are now several systems used as solid lubricants which allows for both low friction and wear resistance. These include diamond and diamond-like carbon, graphite, molybdenum disulfide, hexagonal boron nitride, boric acid as well as soft metals [37]. An interesting example of a low friction wear-resistant coating consists of a composite coating made from a titanium nitride matrix, TiN, with molybdenum sulphides, MoS_x, dispersed as a second phase. Up to 8% (by weight) addition of MoS_x does not effect the hardness of the coating (thus promoting wear resistance), but decreased the coefficient of friction with more than a factor of two, and consequently increasing the life up to 500 times compared to the TiN coating alone [29].

Of high interest is to combine the corrosion resistance coating with the tribological coatings. This multifunctional coating would then resist both corrosion, wear and provide a low coefficient of friction [113,117]. This would increase the lifetime of many engineering applications, and for example, increase the fuel efficiency of vehicles since it reduces energy losses due to friction. Tribological coatings under stress and at the same time in aggressive environments degrade at a significantly faster rate compared to if they were subjected to wear alone or to the aggressive environments alone, since the two conditions aid each other to aggravate the deterioration of the coatings, as illustrated in Fig. 7. In a recent review by Wood [117] it is clear that this is a research area that requires significant attention in the future.

The last class of coatings we will discuss is coatings used for high temperature protection: thermal barrier coatings (TBCs), used to protect the underlying metallic substrate. TBCs are commonly used for protecting superalloys in gas turbines (both stationary for energy production and mobile for propulsion of airplanes). These systems are a material system with multiple layers, in which each layer is optimized for a particular function. It is made up of a *bond coat* being deposited on a metallic base, after which a ceramic *top coat* is deposited, see Fig. 8. In gas turbine engines, a thermal gradient over the top coat of up to 150°C is achieved from active cooling of the superalloy and by selecting top coat materials with relatively low thermal conductivities [38,66,79,99,100,118]. The metal-

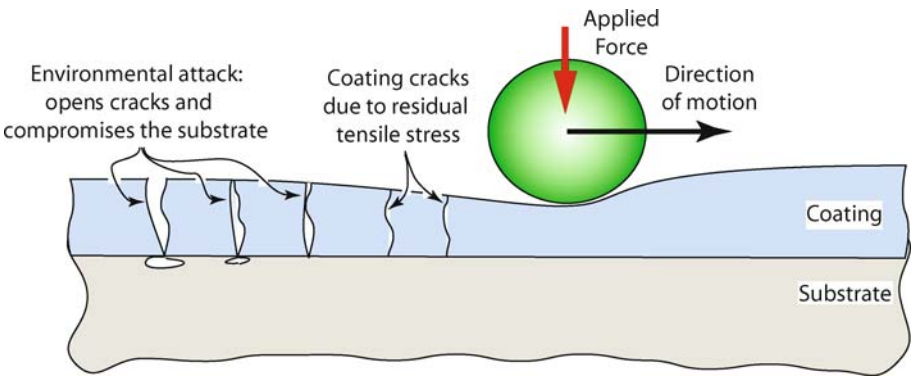

Composites, Multifunctional, Figure 7
Illustration of a degradation caused by combined wear and corrosion of a coated structure. Wear (illustrated by a sphere moving on the surface with an applied normal force) causes cracks in the coating. As the cracks develop, the environment can reach the substrate and deteriorate the substrate quickly

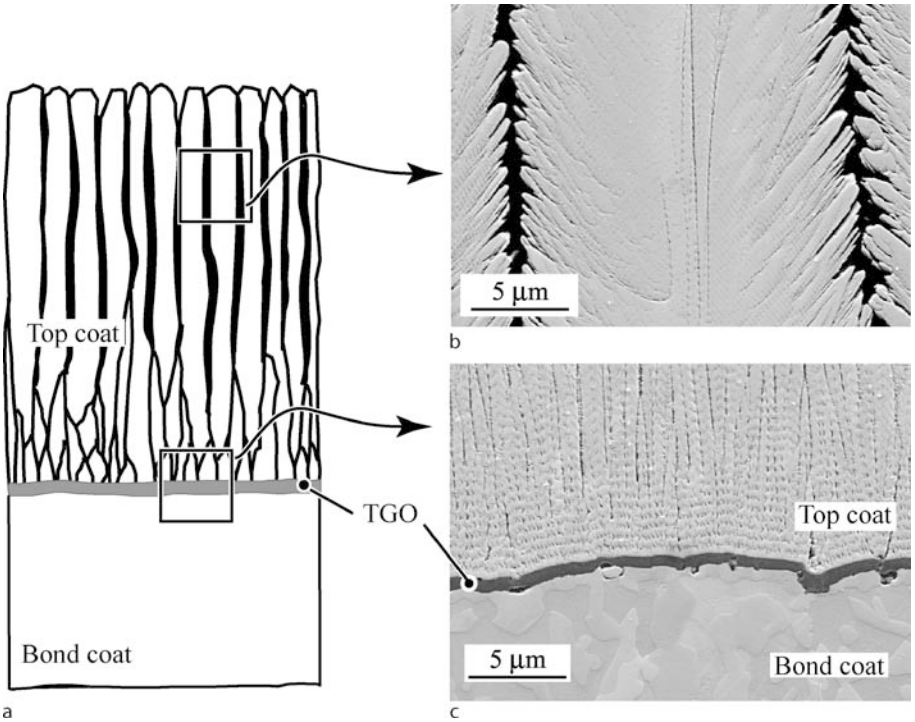

Composites, Multifunctional, Figure 8
An example of a thermal barrier coating produced by electron-beam physical vapor deposition (EB-PVD). a Sketch of the system; b scanning electron microscopy (SEM) image of the cross-section of the top coat, close to the surface; c SEM image of the interface between the top coat, thermally grown oxide (TGO) and the bond coat. Note that b and c are of the same scale. The top coat (yttria stabilized zirconia) is intentionally made porous to allow for strain tolerance during thermal cycling

lic bond coat provides oxidation protection to the superalloy by sacrificing itself by supplying aluminum to form an alpha-alumina scale (α-Al_2O_3) between the bond coat and the top coat (Fig. 8). As the alumina scale grows, the aluminum content decreases in the bond coat, ultimately changing the bond coat properties [32,96]. Thus, it is important to control the chemical content of the materials since even small amounts of critical trace elements can enhance or reduce the interfacial fracture toughness of the structure. Even though TBCs have been used for more

than a decade, improvements of these multifunctional materials are still critical. By improving their reliability and durability, gas turbine powerplants and aircraft engines can become significantly more efficient, reducing their fuel consumption and reducing the pollution associated with energy production and propulsion.

Other Multifunctional Composites

Polyaniline (PANi) as discussed earlier has also been used as nanoparticles (rather than the matrix) to achieve multifunctional composites. For example Dispenza et al. [34] used PANi particles in a hydrogel matrix obtaining a biocompatible nanocomposite with properties suitable for the development of optoelectronic devices. The composite was obtained by a multistep process, starting with water dispersion polymerization of aniline, followed by γ-irradiation. The γ-irradiation cross links the PANi to the steric stabilizers (either poly-vinyl-pyrrolidone, PVP or poly-vinyl-alcohol, PVA). Depending on the processing conditions, various properties can be obtained, but the PANi particles remain spherical [34]. The composites can undergo two optical transitions and the fluorescence signals can vary in wave-length as a function of pH-value.

Composites with 3-dimensional reinforcements have been developed in order to eliminate a number of shortcomings like low shear and transverse stiffness and strength exhibited by laminated composites, with application in areas such as the aerospace industry. 3-dimensional braided reinforcements give reinforcing support in all the three directions thereby suppressing delamination in the composite and giving a higher damage tolerance [80,81,119].

Another group of composites that are multifunctional are the hybrid composite materials. A hybrid composite is made by combing two or more types of fibers in a single matrix material or in two or more types of matrix materials. This gives a greater possibility of achieving multifunctions by changing the combinations of fibers and/or matrix materials [111]. However, the interactions of the constituent components of a hybrid composite and large number of design variables involved make the design of such a composite complex. Examples are carbon-aramid reinforced epoxy, glass-carbon reinforced epoxy, and carbon-Kevlar reinforced epoxy. Lee [58] classified hybrid composites as (1) interply or tow-by-tow, in which tows of the two or more constituent types of fiber are mixed in a regular or random manner; (2) sandwich hybrids, also known as core-shell, in which one material is sandwiched between two layers of another; (3) interply or laminated, where alternate layers of two (or more) materials are stacked in a regular manner; (4) intimately mixed hybrids, where the constituent fibers are made to mix as randomly as possible so that no over-concentration of any one type is present in the material; (5) other kinds, such as those reinforced with ribs, pultruded wires, thin veils of fiber or combinations of the above [58].

Future Directions

Biomimetics is seeking to mimic nature to design and produce materials comparable or better than the ones nature has produced. The goal is to be able to synthesize organs for replacement in the body. Such organs should be able to perform and grow just like the natural one being replaced. This will eliminate looking for or waiting for donors wanting to donate such a needed organ which, in some cases, the body of the patient may reject in the end.

For non-biological applications, mimicking multifunctional biocomposites should be directed at designing synthetic composite materials that can simultaneously perform more than one function. Self-healing is another aspect being targeted to be achieved in material design. The future goal is not only to achieve multifunctionality in composite materials but also such materials should be self-healing. One great lesson from nature is that nature does not waste materials in building its structures yet with incredible relevant properties. This is because nature builds from bottom up, from molecular level to macro level. Ability to control the design and structure of the material on the molecular level will allow production and fabrication of components, devices and structures with incredible properties and functionalities without excessive use of material or energy inputs. This is a great advantage in conservation of material and energy.

Bibliography

Primary Literature

1. Aizenberg J, Lambert G, Weiner S, Addadi L (2002) Factors involved in the formation of amorphous and crystalline calcium carbonate: A study of an ascidian skeleton. J Am Chem Soc 124:32–39
2. Aizenberg J, Weaver JC, Thanawala MS, Sundar VC, Morse DE, Fratzl P (2005) Skeleton of Euplectella sp.: Structural hierarchy from the nanoscale to the macroscale. Science 309:275–278
3. Akle BJ, Bennett MD, Leo DJ (2006) High-strain ionomeric-ionic liquid electroactive actuators. Sens Actuators A-Phys 126:173–181
4. Bagwell RM, McManaman JM, Wetherhold RC (2006) Short shaped copper fibers in an epoxy matrix: Their role in a multifunctional composite. Compos Sci Technol 66:522–530
5. Bandaru PR (2007) Electrical properties and applications of carbon nanotube structures. J Nanosci Nanotechnol 7:1239–1267

6. Barthelat F (2007) Biomimetics for next generation materials. Philos Trans R Soc A-Math Phys Eng Sci 365:2907–2919
7. Barthelat F, Tang H, Zavattieri PD, Li CM, Espinosa HD (2007) On the mechanics of mother-of-pearl: A key feature in the material hierarchical structure. J Mech Phys Solids 55:306–337
8. Bennett MD, Leo DJ (2004) Ionic liquids as stable solvents for ionic polymer transducers. Sens Actuators: A-Physical 115:79–90
9. Bennett MD, Leo DJ, Wilkes GL, Beyer FL, Pechar TW (2006) A model of charge transport and electromechanical transduction in ionic liquid-swollen Nafion membranes. Polymer 47:6782–6796
10. Biddiss E, Chau T (2006) Electroactive polymeric sensors in hand prostheses: Bending response of an ionic polymer metal composite. Med Eng Phys 28:568–578
11. Bleay SM, Loader CB, Hawyes VJ, Humberstone L, Curtis PT (2001) A smart repair system for polymer matrix composites. Compos Pt A-Appl Sci Manuf 32:1767–1776
12. Brown EN, White SR, Sottos NR (2004) Microcapsule induced toughening in a self-healing polymer composite. J Mater Sci 39:1703–1710
13. Brown EN, White SR, Sottos NR (2005) Retardation and repair of fatigue cracks in a microcapsule toughened epoxy composite – part 1: Manual infiltration. Compos Sci Technol 65:2466–2473
14. Brown EN, White SR, Sottos NR (2005) Retardation and repair of fatigue cracks in a microcapsule toughened epoxy composite – Part II: In situ self-healing. Compos Sci Technol 65:2474–2480
15. Bruet BJF, Qi HJ, Boyce MC, Panas R, Tai K, Frick L, Ortiz C (2005) Nanoscale morphology and indentation of individual nacre tablets from the gastropod mollusc Trochus niloticus. J Mater Res 20:2400–2419
16. Bruet BJF, Qi HJ, Boyce MC, Panas R, Tai K, Frick L, Ortiz C (2005) Nanoscale morphology and indentation of individual nacre tablets from the gastropod mollusc Trochus niloticus (vol 20, Pg 2400, 2005). J Mater Res 20:3157–3157
17. Chang MC, Ikoma T, Kikuchi M, Tanaka J (2001) Preparation of a porous hydroxyapatite/collagen nanocomposite using glutaraldehyde as a cross-linkage agent. J Mater Sci Lett 20:1199–1201
18. Chang MC, Ko CC, Douglas WH (2003) Conformational change of hydroxyapatite/gelatin nanocomposite by glutaraldehyde. Biomaterials 24:3087–3094
19. Chen J, Ramasubramaniam R, Xue C, Liu H (2006) A versatile, molecular engineering approach to simultaneously enhanced, multifunctional carbon-nanotube-polymer composites. Adv Functi Mater 16:114–119
20. Chen X, Cao G (2007) Review: Atomistic studies of mechanical properties of carbon nanotubes. J Theor Comput Nanosci 4:823–839
21. Chen X, Huang Y (2008) Nanomechanics modeling and simulation of carbon nanotubes. J Eng Mech 134:211–216
22. Chen X, Surani FB, Kong X, Punyamurtula VK, Qiao Y (2006) Energy absorption performance of a steel tube enhanced by a nanoporous material functionalized liquid. Appl Phys Lett 89:241918
23. Chen XX, Dam MA, Ono K, Mal A, Shen HB, Nutt SR, Sheran K, Wudl F (2002) A thermally re-mendable cross-linked polymeric material. Science 295:1698–1702
24. Chen XX, Wudl F, Mal AK, Shen HB, Nutt SR (2003) New thermally remendable highly cross-linked polymeric materials. Macromolecules 36:1802–1807
25. Cheng L, Wang L, Karlsson AM (2008) Image analyses of two crustacean exoskeletons and implications of the exoskeletal microstructure on the mechanial behavior. J Mater Res (in press)
26. Choi H, Sofranko AC, Dionysiou DD (2006) Nanocrystalline TiO2 photocatalytic membranes with a hierarchical mesoporous multilayer structure: Synthesis, characterization, and multifunction. Adv Funct Mater 16:1067–1074
27. Chung DDL (2001) Electromagnetic interference shielding effectiveness of carbon materials. Carbon 39:279–285
28. Compere P, Goffinet G (1987) Ultrastructural shape and 3-dimensional organization of the intracuticular canal systems in the mineralized cuticle of the green crab Carcinus–Maenas. Tissue Cell 19:839–857
29. Cosemans P, Zhu X, Celis JP, Van Stappen M (2003) Development of low friction wear-resistant coatings. Surf Coat Technol 174:416–420
30. Currey JD (1977) Mechanical-properties of mother of pearl in tension. Proc Royal Soc London Ser B-Biol Sci 196:443–463
31. Currey JD (2005) Hierarchies in Biomineral Structures. Science 309:253–254
32. Darzens S, Karlsson AM (2004) On the microstructural development in platinum-modified nickel-aluminide bond coats. Surf Coat Technol 177:108–112
33. De Stasio G, Schmitt MA, Gellman SH (2005) Spectromicroscopy at the organic-inorganic interface in biominerals. Am J Sci 305:673–686
34. Dispenza C, Leone M, Lo Presti C, Librizzi F, Spadaro G, Vetri V (2006) Optical properties of biocompatible polyaniline nanocomposites. J Non-Cryst Solids 352:3835–3840
35. Dresselhaus MS, Dresselhaus G, Eklund PC (1996) Science of fullerenes and carbon nanotubes. Academic Press, New York
36. Dry C (1996) Procedures developed for self-repair of polymer matrix composite materials. Compos Struct 35:263–269
37. Erdemir A (2005) Review of engineered tribological interfaces for improved boundary lubrication. Tribol Int 38:249–256
38. Evans AG, Mumm DR, Hutchinson JW, Meier GH, Petit FS (2001) Mechanisms controlling the durability of thermal barrier coatings. Prog Mater Sci 46:505–553
39. Ferraris M, Montorsi M, Salvo M (2000) Class coating for Sicf-Sic composites for high-temperature application. Acta Mater 48:4721–4724
40. Fincham AG (2007) Structural Biology of the Enamel Proteins. http://www.usc.edu/hsc/dental/Info/Research/9.html. Accessed 19 Sep 2007
41. Ghosh P, Katti D, Katti K, Mohanty B, Verma D (2005) Mechanical properties of biological nanocomposite nacre: Multiscale modeling and experiments on nacre from Red Abalone. Mater Res Soc Symp Proc 898:7–12
42. Gilbert PUPA, Abrecht M, Frazer BH (2005) The organic-mineral interface in biominerals. Rev Miner Geochem 59:157–185
43. Gooding JJ (2005) Nanostructuring electrodes with carbon nanotubes: A review on electrochemistry and applications for sensing. Electrochimica Acta 50:3049–3060
44. Guo QG, Song JR, Liu L, Zhang BJ (1999) Relationship between oxidation resistance and structure of B4c-Sic/C composites with self-healing properties. Carbon 37:33–40
45. Han A, Chen X, Surani FB, Qiao Y (2008) Rate-dependent in-

filtration of a viscous liquid in nanopores. Appl Phys Lett in press
46. Han A, Qiao YT (2007) Hermal effects on infiltration of a solubility- sensitive volume memory liquid. Phil Mag Lett 87:25–31
47. Heitner-Wirguin C (1996) Recent advances in perfluorinated ionomer membranes: Structure, properties and applications. J Membr Sci 120:1–33
48. Hernandez E, Goze C, Bernier P, Rubio A (1998) Elastic properties of C and Bxcynz composite nanotubes. Phys Rev Lett 80:4502–4505
49. Hughes M, Shaffer MSP, Renouf AC, Singh C, Chen GZ, Fray J, Windle AH (2002) Electrochemical capacitance of nanocomposite films formed by coating aligned arrays of carbon nanotubes with polypyrrole. Adv Mater 14:382–385
50. Iijima S (1991) Helical microtubules of graphitic carbon. Nature 354:56–58
51. Iroh JO, Zhu YR, Shah K, Levine K, Rajagopalan RR, Uyar T, Donley M, Mantz R, Johnson J, Voevodin NN, Balbyshev VN, Khramovb AN (2003) Electrochemical synthesis: A novel technique for processing multi-functional coatings. Prog Org Coat 47:365–375
52. Katti K, Katti DR, Tang J, Pradhan S, Sarikaya M (2005) Modeling mechanical responses in a laminated biocomposite – Part II: Nonlinear responses and nuances of nanostructure. J Mater Sci 40:1749–1755
53. Kessler MR, Sottos NR, White SR (2003) Self-healing structural composite materials. Compos Pt A-Appl Sci Manuf 34:743–753
54. Kroto HW, Heath JR, Obrien SC, Curl RF, Smalley RE (1985) C-60 – Buckminsterfullerene. Nature 318:162–163
55. Kumar P, Nukala VV, Šimunović S (2005) Statistical physics models for nacre fracture simulation. Phys Rev E 72:041919
56. Lee MJ, Jung SH, Kim GS, Moon I, Lee S, Mun MS (2007) Actuation of the artificial muscle based on ionic polymer metal composite by electromyography (Emg) signals. J Intell Mater Syst Struct 18:165–170
57. Lee S, Kim KJ, Park IS (2007) Modeling and experiment of a muscle-like linear actuator using an ionic polymer metal composite and its actuation characteristics. Smart Mater Struct 16:583–588
58. Lee SM (1989) Dictionary of composite materials technology. Technomic Publishing Company Inc, Lancaster
59. Li XD, Xu ZH, Wang RZ (2006) In situ observation of nanograin rotation and deformation in nacre. Nano Lett 6:2301–2304
60. Lu JP (1997) Elastic properties of carbon nanotubes and nanoropes. Phys Rev Lett 79:1297–1300
61. Ma YF, Ali SR, Dodoo AS, He HX (2006) Enhanced sensitivity for biosensors: Multiple functions of DNA-wrapped single-walled carbon nanotubes in self-doped polyaniline nanocomposites. J Phys Chem B 110:16359–16365
62. Ma ZJ, Huang J, Sun J, Wang GN, Li CZ, Xie LP, Zhang RQ (2007) A novel extrapallial fluid protein controls the morphology of nacre lamellae in the pearl oyster, Pinctada Fucata. J Biol Chem 282:23253–23263
63. Mayer G (2005) Rigid biological systems as models for synthetic composites. Science 310:1144–1147
64. McGee T (2007) Biomimicry: Nacre inspires transparent strong as steel plastic. Science & Technology (biopolymer) http://www.treehugger.com/files/2007/10/plastic_steel.php. Accessed 1 Nov 2007
65. Metzler RA, Abrecht M, Olabisi RM, Ariosa D, Johnson CJ, Frazer BH, Coppersmith SN, Gilbert PUPA (2007) Architecture of columnar nacre, and implications for its formation mechanism. Phys Rev Lett 98:268102
66. Miller RA (1984) Oxidation-based model for thermal barrier coating life. J Am Ceram Soc 67:517–521
67. Mirfakhrai T, Madden JDW, Baughman RH (2007) Polymer artificial muscles. Mater Today 10:30–38
68. Motuku M, Vaidya UK, Janowski GM (1999) Parametric studies on self-repairing approaches for resin infused composites subjected to low velocity impact. Smart Mater Struct 8:623–638
69. Mount AS, Wheeler AP, Paradkar RP, Snider D (2004) Hemocyte-mediated shell mineralization in the eastern oyster. Science 304:297–300
70. Naslain R (2004) Design, preparation and properties of non-oxide CMCS for application in engines and nuclear reactors: An overview. Compos Sci Technol 64:155–170
71. Naslain R, Guette A, Rebillat F, Pailler R, Langlais F, Bourrat X (2004) Boron-bearing species in ceramic matrix composites for long-term aerospace applications. J Solid State Chem 177:449–456
72. Naslain RR, Pailler R, Bourrat X, Bertrand S, Heurtevent F, Dupel P, Lamouroux F (2001) Synthesis of highly tailored ceramic matrix composites by pressure-pulsed CVI. Solid State Ion 141:541–548
73. Nassif N, Pinna N, Gehrke N, Antonietti M, Jager C, Colfen H (2005) Amorphous layer around aragonite platelets in nacre. Proc Natl Acad Sci USA 102:12653–12655
74. Nemat-Nasser S, Li JY (2000) Electromechanical response of ionic polymer-metal composites. J Appl Phys 87:3321–3331
75. Nemat-Nasser S, Wu YX (2003) Comparative experimental study of ionic polymer-metal composites with different backbone ionomers and in various cation forms. J Appl Phys 93:5255–5267
76. Nemat-Nasser S, Wu YX (2006) Tailoring the actuation of ionic polymer-metal composites. Smart Mater Struct 15:909–923
77. Neville AC (1975) Biology of the arthropod cuticle. Springer, Berlin
78. Ozyilmaz AT, Erbil A, Yazici B (2004) Investigation of corrosion behaviour of stainless steel coated with polyaniline via electrochemical impedance spectroscopy. Prog Org Coat 51:47–54
79. Padture NP, Gell M, Jordan EH (2002) Materials science – Thermal barrier coatings for gas-turbine engine applications. Science 296:280–284
80. Pandey R, Hahn HT (1996) Designing with 4-step braided fabric composites. Compos Sci Technol 56:623–634
81. Pandey R, Hahn HT (1996) Visualization of 4-step braided fabric composites. Compos Sci Technol 56:161–170
82. Pang JWC, Bond IP (2005) 'Bleeding Composites' – Damage detection and self-repair using a biomimetic approach. Compos Pt A-Appl Sci Manuf 36:183–188
83. Pang JWC, Bond IP (2005) A hollow fibre reinforced polymer composite encompassing self-healing and enhanced damage visibility. Compos Sci Technol 65:1791–1799
84. Podsiadlo P, Liu ZQ, Paterson D, Messersmith PB, Kotov NA (2007) Fusion of seashell nacre and marine bioadhesive analogs: High-strength nanocompoisite by layer-by-layer assembly of clay and L-3,4-dihydroxyphenylaianine polymer. Adv Mater 19:949–955

85. Qi HJ, Bruet BJF, Palmer JS, Ortiz C, Boyce MC (2006) Micromechanics and macromechanics of the tensile deformation of nacre. In: Holzapfel GA, Ogden RW (eds) Mechanics of biological tissue. Springer, Berlin, pp 189–203
86. Qiao Y, Cao G, Chen X (2007) Effect of gas molecules on nanofluidic behaviors. J Am Chem Soc 129:2355–2359
87. Qiao Y, Punyamurtula VK, Han A (2007) Mechanoelectricity of a nanoporous monel – Electrolyte solution system. J Power Sources 164:931
88. Royal Society of Chemistry (RSC) (2006) Glass bones, education in chemistry. http://www.rsc.org/Education/EiC/issues/2006Nov/GlassBones.asp. Accessed 25 Sep 2007
89. Sadeghipour K, Salomon R, Neogi S (1992) Development of a novel electrochimically active membrane and "smart" material based vibration sensor/damper. Smart Mater Struct 1:172–179
90. Sainz R, Benito AM, Martinez MT, Galindo JF, Sotres J, Baro AM, Corraze B, Chauvet O, Dalton AB, Baughman RH, Maser WK (2005) A soluble and highly functional polyaniline-carbon nanotube composite. Nanotechnology 16:S150–S154
91. Sarikaya M, Fong H, Sunderland N, Flinn BD, Mayer G, Mescher A, Gaino E (2001) Biomimetic model of a sponge-spicular optical fiber – Mechanical properties and structure. J Mater Res 6:1420–1428
92. Sarikaya M, Liu J, Aksay IA (1995) Nacre: Properties, crystallography, morphology and formation. In: Sarikaya M, Aksay IA (eds) Biomimetics design and processing of materials. American Institute of Physics, New York pp 34–90
93. Sathiyanarayanan S, Devi S, Venkatachari G (2006) Corrosion protection of stainless steel by electropolymerised pani coating. Prog Org Coat 56:114–119
94. Shahinpoor M, Kim KJ (2000) The effect of surface-electrode resistance on the performance of ionic polymer-metal composite (IPMIC) artificial muscles. Smart Mater Struct 9:543–551
95. Shahinpoor M, Kim KJ (2001) Ionic polymer-metal composites: I. fundamentals. Smart Mater Struct 10:819–833
96. Shi J, Darzens S, Karlsson AM (2005) Aspects of the morphological evolution in thermal barrier coatings and the intrinsic thermal mismatch therein. Mater Sci Eng A-Struct Mater Prop Microstruct Processing 392:301–312
97. Sierakowski RL, Telitchev IY, Zhupanska OI (2008) On the impact response of electrified carbon fiber polymer matrix composites: Effects of electric current intensity and duration. Compos Sci Technol 68:639–649
98. Snyder DR, Sierakowski RL, Chenette ER, Aus JW (2001) Preliminary assessment of electro-termo-magnetically loaded composite panel impact resistance/crack propagation with high speed digital laser photography. 24th international congress on high-speed photography and photonics. Proc SPIE 4183:488–513
99. Stiger M, Yanar N, Topping M, Pettit F, Meier G (1999) Thermal barrier coatings for the 21st century. Z Metallkunde 90:1069–1078
100. Strangman TE (1985) Thermal barrier coatings for turbine airfoils. Thin Solid Films 127:93–105
101. Strumpler R, Glatz-Reichenbach J (1999) Conducting polymer composites. J Electroceram 3:329–346
102. Tan CK, Blackwood DJ (2003) Corrosion protection by multilayered conducting polymer coatings. Corros Sci 45:545–557
103. Tang ZY, Kotov NA, Magonov S, Ozturk B (2003) Nanostructured artificial nacre. Nat Mater 2:413–418
104. Therriault D, White SR, Lewis JA (2003) Chaotic mixing in three-dimensional microvascular networks fabricated by direct-write assembly. Nat Mater 2:265–271
105. Thostenson ET, Chou TW (2006) Processing-structure-multifunctional property relationship in carbon nanotube/epoxy composites. Carbon 44:3022–3029
106. Thostenson ET, Ren ZF, Chou TW (2001) Advances in the science and technology of carbon nanotubes and their composites: A review. Compos Sci Technol 61:1899–1912
107. Toohey KS, Sottos NR, Lewis JA, Moore JS, White SR (2007) Self-healing materials with microvascular networks. Nat Mater 6:581–585
108. University of the Western Cape UWC (2007) Bone, internet BioEd project. http://www.botany.uwc.ac.za/sci_ed/grade10/mammal/bone.htm. Accessed 26 Sep 2007
109. Veedu VP, Cao AY, Li XS, Ma KG, Soldano C, Kar S, Ajayan PM, Ghasemi-Nejhad MN (2006) Multifunctional composites using reinforced laminae with carbon-nanotube forests. Nat Mater 5:457–462
110. Veprek S, Veprek-Heijman MGJ, Karvankova P, Prochazka J (2005) Different approaches to superhard coatings and nanocomposites. Thin Solid Films 476:1–29
111. Wan YZ, Lian JJ, Huang Y, He F, Wang YL, Jiang HJ, Xin JY (2007) Preparation and characterization of three-dimensional braided carbon/kevlar/epoxy hybrid composites. J Mater Sci 42:1343–1350
112. Wang J, Xu CY, Taya M, Kuga Y (2006) Mechanical stability optimization of flemion-based composite artificial muscles by use of proper solvent. J Mater Res 21:2018–2022
113. Wang LP, Zhang JY, Zeng ZX, Lin YM, Hu LT, Xue QJ (2006) Fabrication of a nanocrystalline Ni-Co/Coo functionally graded layer with excellent electrochemical corrosion and tribological performance. Nanotechnology 17:4614–4623
114. Weiland LM, Leo DJ (2005) Computational analysis of ionic polymer cluster energetics. J Appl Phys 97:10
115. White SR, Sottos NR, Geubelle PH, Moore JS, Kessler MR, Sriram SR, Brown EN, Viswanathan S (2001) Autonomic healing of polymer composites. Nature 409:794–797
116. Wilt FH (2005) Developmental biology meets materials science: Morphogenesis of biomineralized structures. Dev Biol 280:15–25
117. Wood RJK (2007) Tribo-corrosion of coatings: A review. J Phys D-Appl Phys 40:5502–5521
118. Wright P (1998) Influence of cyclic strain on life of a PVD TBC. Mater Sci Eng A 245:191–200
119. Wu XQ, Li JL, Shenoi RA (2006) Measurement of braided preform permeability. Compos Sci Technol 66:3064–3069
120. Zako M, Takano N (1999) Intelligent material systems using epoxy particles to repair microcracks and delamination damage in GFRP. J Intell Mater Syst Struct 10:836–841
121. Zhupanska OI, Sierakowski RL (2007) Effects of an electromagnetic field on the mechanical response of composites. J Composite Mater 41:633–652

Books and Reviews

Bar-Cohen Y (2006) Biomimetics: Biologically inspired technologies. CRC Press Taylor and Francis Group LLC, New York

Currey J (1984) The mechanical adaptations of bones. Princeton University Press, Princeton

Forest Products Laboratory (1999) Wood handbook – Wood as an engineering material. Gen Tech Rep FPL–GTR–113. US Department of Agriculture, Forest Service, Forest Products Laboratory, Madison
Lowenstam HA, Weiner S (1989) On biomineralization. Oxford University Press, New York
Mann S (2001) Biomineralization. Oxford University Press, New York
Wainwright SA, Biggs WD, Currey JD, Gosline JM (1976) Mechanical design in organisms. Wiley, New York

Computational Chemistry, Introduction to Complexity in

Danail Bonchev
Department of Mathematics and Applied Mathematics and Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, USA

Defining complexity one remembers the words of St. Augustine: "I know very well what time is, until I have to explain it to somebody else." Intuitively so clear, complexity is so difficult to define rigorously. Complexity is sometimes mistaken with size, large systems are presumably complex. Are they always? Is a big piece of rock more complex than a tiny living cell? What makes a system complex is not the number of its components, but the number and kind of their interactions. As a result of these interactions the system formed has properties and behavior different from those of its components. The dynamics of a complex system's interactions is described by more sophisticated mathematical models (nonlinear differential equations). But how much complexity is needed to observe such emergent properties and dynamics? Given the number of components how many interactions would produce a new behavior?

We may not be able to specify an ideal definition of complexity or a detailed reply to these questions. However, what we can and must do is to quantify complexity, to measure it, as we do with electric charge without knowing what exactly it is. In the realm of computational chemistry, this has been done during the last 30 years defining a variety of measures of *molecular complexity* – compositional, structural, and dynamic. While molecular compositional complexity is straightforward for calculation from the number of atoms of different elements, the structural complexity of molecules can be quantitatively assessed by a variety of descriptors (termed also topological indices), based on graph theory and information theory. Presenting molecular structure as a graph gives rise to a multitude of quantitative descriptors. However, only few of them satisfy the set of criteria formulated in this field. A major requirement for a complexity descriptor is to match in a series of molecules certain patterns of increasing complexity, based on the number of complexifying elements – branches, cycles, central organization, and others (see ► Topological Complexity of Molecules). New methods consider as effective complexity measures the overall subgraph count, the sum of vertex degrees of all subgraphs ("overall connectivity"), and the count of all random walks. Shannon's information theory has also been broadly used in quantifying complexity of chemical systems. Shannon's function was found to be very useful in characterizing different distributions of atoms in molecules. It reflects the increase in complexity with the increase in the diversity of atoms and bonds in molecules, the diversity being caused not only by the elemental composition of atoms, but also by the vertex degrees and vertex total distances distributions in molecular graphs. Moreover, Shannon's information indices calculated from electron distributions in atoms and nucleon distributions in atomic nuclei were shown to be an important tool for predicting the properties of superheavy chemical elements, and their isotopes (see ► Information Theoretic Complexity Measures). The universality of the graph theoretical and information theoretical methods developed in the field of computational chemistry makes them also of interest in network complexity analysis in biology, social sciences, computer sciences, and other branches of science and technology.

A higher-level theoretical framework of molecular similarity and complexity is the quantum mechanical, density-functional formalism (DFT). This quantitative approach enables comparing and ordering of molecular structures, which finds application in a quantum version of the linear and nonlinear QSPR (see ► Quantum Similarity and Quantum Quantitative Structure-Properties Relationships (QQSPR)). Tuning the relative reaction rates of the different steps of a complex chemical reaction would enable the optimal design of many industrially important reactions. The DFT technique was recently used to address such a fine tuning procedure (see ► Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis). Such an approach promises to provide in the near future tools for effective screening of potential new materials with tailored catalytic properties.

Drug discovery is a major field of interest for applied computational chemistry. The lengthy process of design of a new drug starts with a search for the best computational model that relates molecular structure, composition, and properties to certain biological activity that has the desired therapeutic effect. This vast area of research, which spans over the last 50 years and is known as QSAR (Quantitative Structure-Activity Relationships) modeling,

reduces strongly the potential drug candidates selected for synthesis and testing. While composition and properties are experimentally determined quantities, the translation of molecular structure into numbers is not unique, and includes a multitude of structural parameters termed molecular descriptors (see ► Drug Design, Molecular Descriptors in). Different classes of descriptors capture different aspects of molecular structure. The manner in which atoms are connected to each other is characterized by graph theory-based structural invariants called topological or two-dimensional (2D) indices. Three-dimensional (3D) descriptors are used to account for the metric and geometry of molecules, while conformational (or 4D) parameters complete the picture by accounting for the presence of a variety of 3D structures in molecules having free rotating atomic groups. Distinct classes include quantum chemical descriptors, such as atomic charges, bond orders, superdelocalizabilities and dipole moments, as well as physicochemical properties like solubility and binding constants. No theory exists to predict which of these descriptor classes would work the best for a given class of biologically active compounds, thus making all of them useful tools in drug design.

The search for new promising drug candidates experienced a quantum leap, when the intelligent guesses based on analogies with naturally occurring compounds were replaced by automated searches in huge databases of chemical compounds. The use of such combinatorial libraries has led to an explosive increase in the number of identified therapeutic targets, and enabled producing tailor-made drug molecules. All this increased considerably the complexity of the *in silico* methods used in drug discovery. The QSAR modeling reached a stage of maturity, with rigorously defined criteria for selection of molecular descriptors, as well as selection of methods for model validation and determining the applicability domains of the models derived (see ► QSAR Modeling and QSAR Based Virtual Screening, Complexity and Challenges of Modern). It was realized that no particular QSAR modeling method has an advantage over the others, as well as that each specific technique and each specific class of descriptors have their unique contribution to the relationship with the examined biological activity. This resulted in the creation of combinatorial QSAR methodology, which produces a multitude of models, satisfying a set of acceptability criteria.

In parallel with the optimization of traditional QSAR methods and their adaptation to the complex datasets available, the field of drug design benefited tremendously by using the methods of artificial intelligence. Most of these methods have been inspired by the manner in which the living things function and evolve (see ► Drug Design with Artificial Intelligence Methods). Genetic and evolutionary algorithms are stochastic methods that solve optimization problems by evolving solutions based on concepts of DNA genetics and Darwinian evolution. The ant colony optimization (ACO) algorithms mimic the behavior of some ant species when searching for the shortest path to a food source. The particle swarm optimization (PSO) algorithm mirrors the social behavior of large groups of individuals, such as bird flocking, fish schooling, and animal herding. Swarm intelligence algorithms are used in drug design for diverse application, including selection of structural descriptors for QSAR models, enzyme-inhibitor docking, and gene expression. Another class of algorithms, termed Artificial Immune Systems (AIS), makes use of learning and memory principles used in the human immune system. AIS models found application in diverse biological and medical problems, such as prediction of protein structure, cancer diagnosis, classification of gene expression data, and recognition of ECG arrhythmia. Machine learning (decision trees, support vector machines, k-nearest neighbors, etc.) is an important field of artificial intelligence methods. It extracts information from experimental data by computational and probabilistic methods, and by using a set of rules it predicts the properties of objects not included in the learning set (see ► Drug Design with Machine Learning). Machine-learning techniques, such as artificial neural networks (ANNs), have been widely applied to QSAR data (see ► Drug Design with Artificial Neural Networks). Its success story began with the creation of algorithms for the training of multilayer feed-forward (MLF) artificial neural networks. MLF was widely used to model biological and toxicological properties of chemical compounds. Other networks used in drug design are self-organizing maps, counter-propagation networks, and probabilistic neural networks.

Biochemistry became for biology what physics has been for chemistry – the fundament to build on and explain. The essence of life, and even its origin, is nowadays in a process of redefining, proceeding from the self-organization of biomolecules (DNA, RNA, proteins, metabolites) in biochemical networks. All aspects of biological development, such as cell differentiation, tissue multilayering, segmentation and left-right asymmetry, can be related to the physicochemical processes of self-organizing. The complex dynamics of self-organizing is characterized by oscillations, pattern formation, and the emergence of multiple steady states. Origin of life is the domain where chemical self-organization and biological evolution meet (see ► Biological Development and Evolution, Complexity and Self-organization in). Networks of different kinds

and sizes appear at different stages of molecular evolution. (see ► Molecular Evolution, Networks in). The evolution of networks is not yet studied in detail, yet it is already established that gene mutations cause addition and deletion of nodes in protein–protein interaction networks, transcriptional regulatory networks, and metabolic networks, whereas rewiring of nodes is an evolutionary mechanism observed only in the first two networks.

Complexity of biological systems is multifaceted. Traditionally, the genome sequence complexity is viewed as a universal method for comparing species complexity, despite the controversy with some of the complexity measures examined. The Kolmogorov's information, Shannon's information, and Jensen–Shannon divergence are among the most frequently used measures (see ► Biological Complexity and Biochemical Information). Other types of biocomplexity, based on the structure and functionality of biological systems, have also been analyzed. Recently, the advance in the study of biological networks has offered new options for the assessment of biological complexity, including modeling the evolution of complexity in artificial life systems. Essential for each specific quantitative measure of network complexity is its divergence from the random network having the same size and the same average connectivity. Modular and motif (subgraph) information content of biological networks were shown to reflect well the complex network organization. Complexity of the dynamics of biochemical reactions is closely related to the cell fractal structure, and has important consequences for the dynamic behavior of cells. The processes of organizing the intracellular macromolecular systems obey chaotic dynamics. Non-linear interactions in and between spatial and temporal domains and over wide ranges of scales underlie the emergent properties of complex biological systems (see ► Biochemistry, Chaotic Dynamics, Noise, and Fractal Space in). With their nonlinearity, fractal dimensions, and attractors-controlled dynamics, the processes of the self-organizing chaos are of major importance for complexity theory. The high complexity of biological systems is a challenge for the traditional modeling of dynamics based on continuous models and ordinary differential equations (ODE). The cellular automata method (see ► Cellular Automata Modeling of Complex Biochemical Systems) appears as a viable alternative for modeling of processes within biological systems, including their metabolic, protein, and gene regulatory networks, as well as for identifying dynamic patterns, and devising of strategies for pathway control.

Chemistry and chemical engineering have entered into the 21st century with a broad spectrum of new classes of sophisticated and even "smart" materials of high complexity. Nanotechnology offers structures of unusual properties based on atomic clusters, nanotubes, and DNA. Composite materials and special polymers extend essentially the areas of application of classical materials. Many of the new materials are created with the technology of molecular self-assembly, adopted from the living nature. The high complexity of nanoscale materials (see ► Nanoscale Atomic Clusters, Complexity of) stems from the wealth of possible structures, and the very strong size-dependence of their properties. Clusters have great potential as a means of optimization of existing technologies, developing principally new technological processes, materials, and devices.

DNA provides basic building blocks for constructing nanostructures with specific functional features like molecular recognition, self-assembly, and predictable structure (see ► DNA-Templated Self-Assembly of Protein Arrays and Highly Conductive Nanowires). Periodic protein arrays can be constructed by templated self-assembly onto DNA nanogrids. This enables producing target materials with predictable 3D structure like highly conductive nanowires. The principles of biomolecular self-assembly can be extended to the modern materials synthesis, leading to a broad range of new materials and processes with significant technological impact (see ► Self-assembled Materials). The complex nature of self-assembling processes makes their modeling a nontrivial task. The methods used cover from atomic scale (molecular dynamics and Monte Carlo simulations) to mesoscale (kinetic Monte Carlo and phase field modeling) to lattice methods, to Random Sequential Adsorption (RSA) and cellular automata techniques. The development of complex, multi-scale modeling approaches is under way and shows great promise.

The mass-production of industrial polymers was dominated for a long time by polymers with linear macromolecular structure like polyethylene and polypropylene,. The appearance of nonlinear polymers with a variety of branched-tree topologies offered new classes of materials with properties unseen in linear polymers (see ► Polymers, Non-linearity in). The entanglement of branches in the star- and comb-shaped, and hyper-branched (known also as dendrimers) polymers, results in complicated conformational statistics and dynamics. The latter is modeled surprisingly well using graph theory, providing reliable prediction of the materials' properties in melt and solid states. Even more unusual properties are offered by composites. Composite materials combine the best features of different materials and are ideally suited to achieve multifunctionality and to form materials that have a broad spectrum of desired properties (see ► Composites, Multifunc-

tional). Carbon fiber and carbon nanotube-based composites can be formed with materials as diverse as polymers, cement, and ceramics. Perhaps the most unusual property demonstrated by some complex composites is the self-healing. Inspired from the technological secrets of the living things, such dream materials not requiring continuous support will become one of the most exciting novelties of 21st century.

Computational and Theoretical Nanoscience, Introduction to

Yong S. Joe
Center for Computational Nanoscience,
Department of Physics and Astronomy,
Ball State University, Muncie, USA

Nanoscience and nanotechnology change the nature of almost every human-made object in this century. Advances in the field of nanoscience empower us with new tools for proving electronic devices with ever-decreasing scale. Many people have projected that nanometer-scale devices will continue this trend, bringing control of matter to unprecedented scales. This includes scale reduction not only in microelectronics, but also in fields such as quantum-switch-based computing in the shorter term. These advances have the potential to change the way we engineer our environment, construct and control systems, and interact in society.

Computational science, which has emerged as a third way of doing research, one that complements theory and experiment, plays a key role in developing our understanding of materials at the nanometer scale and in the development "by-design" of new nanoscale materials and devices. Hence, modeling and simulation are now integral components of scientific research.

It is essential to have a detailed understanding of quantum effects in electronic transport to design devices effectively at the nanoscale, sustain the miniaturization trends of integrated circuits, and create new engineered nanostructures [see ► Quantum Phenomena in Semiconductor Nanostructures]. The resonance phenomena has a special attention in the electronic transport of non-interacting electrons through the infinite rectilinear quantum wires with impurities and one-dimensional rings with impurities connected to current leads [see ► Resonances in Electronic Transport Through Quantum Wires and Rings]. The Fano resonance, which is a manifestation of the interference between a localized state and the continuum states, is investigated in an Aharonov–Bohm ring and in an open three-terminal interferometer with a quantum dot [see ► Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring]. It is also important to understand quantum transport and Kondo physics in coupled quantum dots with Anderson impurity [see ► Quantum Impurity Physics in Coupled Quantum Dots]. The complex quantum dots, treated as artificial molecules, are an excellent model system for studying dynamical symmetries and Kondo effect. The simplest of such complex objects is the double quantum dot which consists of two islands with confined electrons [see ► Tunneling Through Quantum Dots with Discrete Symmetries]. Since the quantum interference of the system depends not only on the dynamical phase but also on the geometric phase, the investigation of the geometric phase and its effect on nanostructures in a dynamical process of the environment is of particular interest [see ► Geometric Phase and Related Phenomena in Quantum Nanosystems]. In studying complex systems we have to deal with coupled phenomena and processes at a multitude of different spatial and temporal scales. Understanding interactions in the low dimensional semiconductor nanostructures and its response at multiple scales is a fundamental quest of modern science. [See ► Nanoscale Processes, Modeling Coupled and Transport Phenomena in Nanotechnology].

The spin phenomena in mesoscopic transport have moved only recently into the focus of attention, as one branch of the field of spintronics. The interplay between quantum coherence with confinement- or interaction-effects gives rise to a variety of unexpected spin phenomena in mesoscopic conductors and allows moreover to control and engineer the spin of the charge carriers. The spin interference is often the basis for spin-valves, spin-filters and detectors, and spin-switches. Their underlying mechanisms may gain relevance on the way to future semiconductor-based spin electronics. A quantitative and theoretical understanding of spin-dependent mesoscopic transport calls for developing efficient and flexible numerical algorithms.

The electronic and transport properties of quantum dot spin transistors are studied with emphasis on single-electron tunneling and shell structure using comprehensive modeling approach. For example, self-consistent calculation of the Poisson and Schrödinger equations within the spin-density-functional theory and the exact diagonalization of the many-body Schrödinger equation can be employed to describe spin transport properties of quantum dots [see ► Quantum Dot Spin Transistors, Self–consistent Simulation of]. With rising interest in spin-dependent transport, the interplay of the electron spin and

charge degree of freedom has been exploited in a variety of spin interference devices. Charge and spin transport through phase-coherent conductors of mesoscopic scales carry imprints of wave interference as predominant and characteristic features [see ► Spin-Polarized Quantum Transport in Mesoscopic Conductors: Computational Concepts and Physical Phenomena]. Novel physical phenomena that may lead to improved memory devices and advances in quantum information processing are closely related to spin-orbit interactions. Hence, it is important to investigate a semiclassical wave-packet description of spin transport in the presence of electric fields by explaining how the microscopic motion of carriers gives rise to a spin current [see ► Semiclassical Spin Transport in Spin-Orbit Coupled Systems].

In the meantime, the theory of exchange-correlation energy is of great interest in modern density-functional many-body approaches. Hence, the spin dependent exchange-correlation energy in an interacting system such as the two-dimensional electron layers is of critical importance in the proper design of modern nano-structure devices and quantum-well lasers from a technological point of view [see ► Spin Dependent Exchange and Correlation in Two-Dimensional Electron Layers]. In addition, the time revolution of correlation functions in disordered spin systems is also worth to investigate [see ► Spin Dynamics in Disordered Solids].

Carbon nanotubes are referred to as the fabric of nanotechnology, and will play a central role in the future development of this technology. Understanding the properties of nanotubes, via computational simulation studies, has been one of the most intensive areas of research in physical sciences. Carbon nanotubes form the fourth allotrope of crystalline carbon after graphite, diamond, and a variety of caged-like fullerene molecules. Their mechanical properties make them stronger than steel, and their thermal conductivity is faster than copper. They have very exotic electronic-conduction properties by changing their geometry or by introducing topological defects into their structure and therefore, their electronic conductance can change from metals to semi-conductors. In general, carbon nanotubes show very different mechanical, thermal, electronic and optical properties.

The investigation into nanotube properties has prompted an intensive experimental and theoretical/computational research. One of the most active areas has involved the use of predictive computational modeling of their mechanical, thermal and mass transport properties [see ► Carbon Nanotubes, Thermo-mechanical and Transport Properties of]. One of the computational methods includes atomistic simulations using tight-binding molecular dynamics to study the nucleation and formation of carbon fullerenes and single-walled carbon nanotubes [see ► Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation]. Furthermore, the non-equilibrium Green's Function approach can be employed to study nanowire field effect transistors (FETs) such as a silicon nanowire FETs and carbon nanotube FETs [see ► Quantum Simulations of Ballistic Nanowire Field Effect Transistors]. Therefore, carbon nanotubes are viewed as highly relevant nanostructures with extensive potential applications.

The physical implementation of the quantum bit (or qubit) has been, and still is, the starting point of any proposal for quantum information processing device. The research of quantum computing focuses on qubits and quantum entanglement based on both a large spectrum of systems and solid-state devices. Hence, it is interesting to consider the practical realizability of quantum computers based on solid-state flying qubits and a simple quantum-gate network [see ► Charge-Based Solid-State Flying Qubits].

For device sizes in the range of tens of nanometers, the atomistic granularity of constituent materials cannot be neglected. Effects of atomistic strain, surface roughness, unintentional doping, the underlying crystal symmetries, or distortions of the crystal lattice can have a dramatic impact on the device operation and performance. In an atomistic simulation, one takes into account both the atomistic/granular and quantum properties of the underlying nanostructure. The variety of geometries, materials, and doping configurations in semiconductor devices at the nanoscale suggests that a general Nanoelectronic Modeling tool (NEMO 3-D) is needed [see ► Multimillion Atom Simulations with Nemo3D].

There is a persistent need of miniaturization of machines and energy conversion devices for various engineering applications. In the field of robotics, there is a need to develop nano-scale actuators, motors and other machine components [see ► Viral Protein Nano-Actuators, Computational Studies of Bio-nanomachines]. These nano-scale robots, machines, and sensors can be used to deliver drugs to specific locations in the body, or detect individual cancer cells, or be used as molecular filters to separate minute particles from the environment. On the other hand, the field computation, which is a model of computation that information is represented primarily in fields, is to provide a mathematical language for describing information processing in the brain and in other natural and artificial systems [see ► Field Computation in Natural and Artificial Intelligence]. This field computation can be expected to provide an increasingly important analyt-

ical and intuitive framework for understanding massively parallel analog computation in natural and artificial intelligence.

In conclusion, there is no doubt that computational and theoretical nanoscience research such as quantum electronic and spintronic nanodevices will yield new applications in the near future for sensing, information processing, and quantum computation.

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis

Anton Kokalj
Department of Physical and Organic Chemistry, Jožef Stefan Institute, Ljubljana, Slovenia

Article Outline

Glossary
Definition of the Subject
Introduction
Ab Initio Electronic Structure Calculations
Description of Activated Processes
Description of the Overall Reaction
Theory and Trends of Chemisorption and Reactivity on Transition Metals
Examples
Future Directions
Bibliography

Glossary

Surface site is a position on the surface. High-symmetry positions are usually named as top, bridge, and hollow sites and designate positions over a surface atom, over two neighboring surface atoms, and over the void in between neighboring surface atoms, respectively.

Surface reaction center is a group of atoms on the surface where an elementary reaction occurs. The term reaction site will be used occasionally.

Adsorption is a process where molecules from the gas-phase or from liquid-solution bind to a solid or liquid surface. In this article, adsorption from the gas-phase to a solid surface is considered exclusively.

Chemisorption is a type of adsorption where a molecule binds to the surface through a direct chemical bond, whereas in physisorption the binding is due to van der Waals interaction.

Catalyst's activity is related to the rate of catalytic reaction, and can be expressed as amount of product made per unit time per active site (i.e. turnover frequency).

Catalyst's selectivity is a ratio between the amount of desired product obtained per amount of consumed reactants.

DFT Density Functional Theory
GGA Generalized Gradient Approximation
PES Potential Energy Surface
MEP Minimum Energy Path
NEB Nudged Elastic Band
fcc Face-Centered Cubic
IS Initial State
TS Transition State
FS Final State
BEP Brønsted–Evans–Polanyi
MD Molecular Dynamics
TM Transition Metal; term "early TM" designates transition metals with less than half-filled *d*-states, whereas "late TM" are those with more than half filled *d*-states.
DOS Density of States
STM Scanning Tunneling Microscopy
OMC Oxametallacycle
EO Ethylene epoxide
Ac Acetaldehyde

Definition of the Subject

First-principles (in Latin *ab initio*) quantum-mechanical-based computer simulations can deliver an atomic-level insight into elementary constituents of phenomena such as chemical reactions. The knowledge and understanding thus obtained may be used to propose new reaction-centers with on-demand tailored catalytic properties. The purpose of this article is to demonstrate with several examples how to gain such an insight and on this basis how to help design new *heterogeneous* reaction centers. In particular, surfaces of transition metals are considered, which are traditionally known as good heterogeneous catalysts. However, *ab initio* calculations of the properties of individual reaction centers present only the first part of the endeavor, because dynamical properties (e. g. overall reaction rates) are, in general, determined by the interplay between many elementary atomic-scale processes. The recipe for calculating the properties of such a dynamical system may therefore be accomplished by a hierarchy approach, where different computational methods are utilized for different time and lengths scales. For example, the information provided by the atomic-scale quantum-mechanical calculations can be utilized by statistical mechanics methods like kinetic Monte Carlo to follow the time evo-

lution of the system. Such (statistical) methods are beyond the subject of this article and are just shortly outlined.

Introduction

A chemical reaction is a process of conversion of substances into other substances. Consider for example a simple chemical reaction, where molecules A and B react so as to form molecules C and D. The molecules A and B that enter into reaction are called reactants, and the molecules C and D are the reaction products. To understand the mechanism of reaction we need to envisage the atomic structure of the involved molecules and the trajectory during their inter-conversion. In general, during chemical reaction some bonds are broken and new bonds are formed. Perhaps the simplest picture of a chemical bond—a bond linking two atoms together—is that of a mechanical spring governed by Hook's law. This means that atoms "feel" a harmonic potential around equilibrium positions. Imagine for example that during chemical reaction an *a*–*b* bond breaks and *b*–*c* bond forms. This means that the *a*–*b* bond will be stretched first (the potential energy will increase) and eventually broken (point of maximum potential energy), and then a new *b*–*c* bond will start to form (the potential energy will decrease). This process is schematically shown in Fig. 1. Already this simple picture shows that chemical reaction is an activated process, that is, the system has to cross an energy barrier when passing from reactants to products. In reality the majority of chemical reactions are activated. Moreover, in many cases, the energy barrier of the reaction is so large that the reaction does not take place: more precisely the rate of reaction is too slow to be observed. Namely, the rate of chemical reaction depends exponentially on the energy barrier: the larger the barrier, the smaller is the rate. To accelerate a chemical reaction, a "third" substance, called a catalyst, is used. A catalyst is a substance that accelerates the rate of a chemical reaction, but is itself not consumed by the reaction. In terms of the above discussion of activated process, this means that the catalyst reduces the height of the activation barrier (see Fig. 1), for example, by means of providing a new reaction mechanism.

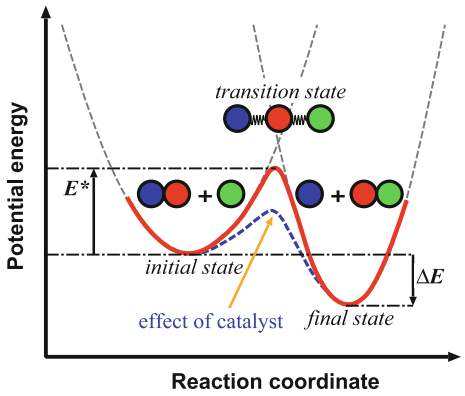

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 1
Reaction energy profile of a simple reaction (*red curve*): $ab + c \rightarrow a + bc$. Around equilibrium positions (local minima) atoms feel harmonic potential (indicated by *two dashed gray parabolas*). As a consequence chemical reaction (bond-breaking and bond-making) will involve overcoming an energy barrier, E^*. The effect of a catalyst on the chemical reaction is to reduce the E^* (*blue curve*) and correspondingly to enhance the rate of reaction

A catalyst can be either homogeneous or heterogeneous. A homogeneous catalyst is in the same phase (solid, liquid, and gas), whereas heterogeneous catalyst is in a different phase with respect to reactants. A typical example of heterogeneous catalysis is where a solid catalyst is used to accelerate the gas-phase reaction. At present, about 90% of all chemicals are produced as heterogeneously catalyzed processes [62], where catalysts are solids, and provide a range of products such as fuels, fertilizers and plastics. Moreover heterogeneous catalysis is used to clean poisonous emissions from power plants, cars, and industrial production. Because of the technological and economical importance of catalysis, it makes it worthwhile to improve the catalysts even if the gain is as minor as only a few percentage points or less in performance. The efficiency of the production process strongly depends on the specific properties of the catalyst. A deeper knowledge of the interactions between the reactants and the catalyst and the understanding of the reaction mechanism would allow for the optimization of the production of many chemicals.

To understand the reaction mechanism of a given chemical reaction in detail, it is clear that the above-mentioned spring-picture of the chemical bonds holding the atoms within the molecule together will not suffice. Moreover, when a catalyst is used to accelerate the rate of reaction, the complexity of the system rises, because the number of degrees of freedom increases substantially. The progress attained in the last few decades in the field of computer simulations of matter at the atomic scale have opened new perspectives toward an atomic-scale understanding of the catalysts. In particular, electronic structure calculations, which for the description of solids are mainly based on Density Functional Theory (DFT), can provide information that are otherwise hard to obtain from experiment, such as the details of the reaction mechanisms and identification of the nature of the transition states. Therefore they help elucidate the factors that determine the ac-

tivity and selectivity of catalysts, and consequently provide a way for their improvement.

Ab Initio Electronic Structure Calculations

The dynamics of a chemical reaction could, in principle, be treated by solving the time-dependent Schrödinger equation. In practice, this is not feasible (except maybe for a few trivial cases), and therefore several approximations are made. For example, one could proceed with a time-independent Schrödinger equation that is simplified with Born–Oppenheimer approximation. The time-independence stems from the fact, that, for example, the Hamiltonian (the energy operator) of an isolated molecule or crystal (in the absence of any external time-varying "*influence*") does not depend on time, whereas the Born–Oppenheimer approximation allows one to decouple the electronic motion from the motion of nuclei. This is made possible, because the mass of the nucleus is about three orders of magnitude heavier than the mass of the electrons and therefore the time scale of ionic dynamics is much larger than the electronic one. This means that one picks a given fixed configuration of nuclei (kinetic energy of nuclei is thus neglected) and then solves the electronic problem (hence the name *electronic structure theory*). Formally, this can be written as:

$$\widehat{H}_{\{\mathbf{R}\}}\Psi_{\{\mathbf{R}\}} = E(\{\mathbf{R}\})\Psi_{\{\mathbf{R}\}} \,, \tag{1}$$

where $\widehat{H}_{\{\mathbf{R}\}}$ and $\Psi_{\{\mathbf{R}\}}$ are the Hamiltonian and wave-function of n interacting electrons moving in the field of N fixed nuclei with coordinates $\{\mathbf{R}\}$. $E(\{\mathbf{R}\})$ is the energy of the Hamiltonian $\widehat{H}_{\{\mathbf{R}\}}$, and can be seen as the potential energy that the nuclei "feel" at configuration $\{\mathbf{R}\}$. As for the analysis of chemical reactions, it is often only the ground-state $E(\{\mathbf{R}\})$ that is of interest. The $E(\{\mathbf{R}\})$ is therefore a function of the coordinates of nuclei. Solving Eq. (1) for many different configurations of the nuclei, would allow for the construction of the potential-energy-surface (PES), that can be used to analyze the motion of nuclei. Once the PES is constructed the forces acting on atoms can be calculated as $\mathbf{F}_I = -\frac{\partial E(\{\mathbf{R}\})}{\partial \mathbf{R}_I}$ (I designates a given atom), and the dynamics (e. g. motion of atoms) can be studied, for example, by integrating the classical equation of motions—a method known as *ab initio* molecular dynamics (MD). However, the construction of PES is only feasible for very simple systems, for others *ab initio* MD consists of solving Eq. (1) at each snapshot during the *discretized* time evolution (t, $t+\delta t$, $t+2\delta t, \ldots$) of the system. At each snapshot the forces are calculated by Hellman–Feynman theorem as $\mathbf{F}_I = -\frac{\partial E(\{\mathbf{R}\})}{\partial \mathbf{R}_I} = -\langle \Psi_{\{\mathbf{R}\}} | \frac{\partial H_{\{\mathbf{R}\}}}{\partial \mathbf{R}_I} | \Psi_{\{\mathbf{R}\}} \rangle$, and the system is propagated by δt accordingly. Note that in such MD treatment the quantum effects of nuclei (e. g. tunneling) are neglected.

For practical purposes, even solving Eq. (1) is too complicated (analytical solution does not exist, except for a few trivial cases), and the n-electron equation is further simplified into a set of one-electron equations with the following form:

$$\left[-\tfrac{1}{2}\nabla^2 + \upsilon_{\text{eff}}(\mathbf{r})\right]\phi_i = \varepsilon_i \phi_i, \quad i = 1, n \,, \tag{2}$$

where $\upsilon_{\text{eff}}(\mathbf{r})$ is an effective one-electron potential (for notational simplicity, nonmagnetic insulator is considered). One possible scheme is the Hartree–Fock (HF) method, where the many-electron wave-function Ψ is approximated by an anti-symmetrized product (*Slater determinant*) of one-electron orbitals, ϕ_i, thus satisfying the Pauli exclusion principle according to which the wave-function has to be anti-symmetric with respect to exchange of two electrons. In HF method the effective potential $\upsilon_{\text{eff}}(\mathbf{r})$ is given by:

$$\widehat{\upsilon}^{\text{HF}}_{\text{eff}}(\mathbf{r}) = \upsilon(\mathbf{r}) + \int_{\mathbf{r}'} \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}' - \widehat{K}(\mathbf{r}) \,. \tag{3}$$

The $\upsilon(\mathbf{r})$ is the *external* potential due to nuclei charges, $\rho(\mathbf{r}')$ is the electron density (the integral accounts for the electron-electron Coulomb interaction), and $\widehat{K}(\mathbf{r})$ is the exchange operator that arises due to "anti-symmetry" of the wave-function. The HF method, although computationally appealing, gives too poor results due to neglect of the correlation effects. The procedure for treating correlations is well established, but the corresponding wave-function-based methods are computationally very demanding [104].

Another computationally appealing method is provided by Density Functional Theory, where the electronic structure problem is simplified by realizing that the ground-state energy is a functional of the electron ground-state density [35]. The practical DFT approach also consists of solving a set of one-electron equations (*Kohn–Sham* equations [46]), similar to those of HF, but the DFT formalism—although formally exact—allows one to treat the correlation effects, in practice, in an approximate way. In DFT the effective potential $\upsilon_{\text{eff}}(\mathbf{r})$ is given by:

$$\upsilon^{\text{KS}}_{\text{eff}}(\mathbf{r}) = \upsilon(\mathbf{r}) + \int_{\mathbf{r}'} \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}' + \upsilon_{\text{xc}}(\mathbf{r}) \,, \tag{4}$$

where the $\upsilon_{\text{xc}}(\mathbf{r})$ is the exchange-correlation potential, which is given in terms of exchange-correlation (XC) energy functional $E_{\text{xc}}[\rho]$ as $\upsilon_{\text{xc}}(\mathbf{r}) = \delta E_{\text{xc}}[\rho]/\delta\rho(\mathbf{r})$. There are two types of approximations in DFT, depending on how

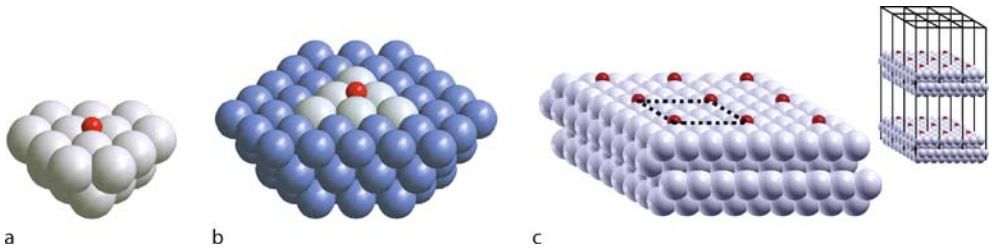

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 2
Three types of surface models: a cluster, b embedded cluster, and c slab model (surface unit cell is indicated by *dashed line*). The multi-slab flavor of the slab model is shown in the *inset* of c

the $E_{xc}[\rho]$ is approximated [79]. In the local-density-approximation (LDA) the XC functional depends solely on electron density (i. e. the space is partitioned into infinitesimally small cubes, which are treated as uniform electron gas), whereas the generalized-gradient-approximation (GGA) incorporates also the density gradient terms into the XC functional. The LDA usually gives poor estimates of activation energies (poorly describes the bond-breaking and bond-making), because it favors homogeneous systems. Because a more bonded system is more homogeneous, the LDA overestimates the binding energies, and correspondingly severely underestimates the activation energy (E^*) of bond-making processes (and vice versa for bond-breaking reactions). On the other hand, the GGA gives more accurate binding and activation energies. There are many flavors of GGA functionals, for the description of solids the most popular are those of Perdew [80,81].

Modeling the Surface

In order to describe a surface (or a catalyst) with a computer, one needs to construct an appropriate model. Such models consists of a limited number of atoms, because it is not possible to treat a macroscopic number of (inequivalent) atoms. Three types of surface models are used (see Fig. 2):

Cluster model: surface is represented by a cluster of atoms that is obtained by a "finite cut" from an ideal lattice. This model builds on the assumption that adsorption is a phenomenon of strong local character. The cluster models of metallic surfaces are prone to cluster-size effects [34,86,109], in particular, the adsorption energy is known to undergo large variations as a function of the cluster size and shape [34,86].

Embedding scheme: embedding schemes divide the adsorbate/substrate system into at least two zones: (i) the cluster region and (ii) an outer region. The essence of embedding is to correctly connect the cluster to the outer part of the model so as to diminish the size effects of the cluster model. The most sophisticated embedding schemes are based on Green function formalism [6,24].

Periodic slab model: describes the surface by an infinite periodic slab in the *xy*-direction with finite thickness in the *z*-direction. The structure (and description) of such a slab is entirely given by a surface unit cell. There are two flavors: a single slab and a multi-slab model. The latter is actually an infinite 3D model, where individual slabs are separated by vacuum layers that are thick enough to prevent adjacent slab–slab interaction.

A given surface can be cut from the bulk crystal with many possible orientations. Miller indices are used to designate the surface plane by three integers as (*hkl*). These indices denote a plane that intercepts the three lattice vectors at $1/h$, $1/k$, and $1/l$, respectively. Figure 3 shows the three low-Miller index surfaces of an fcc (face-centered-cubic) crystal, which are (100), (110), and (111).

Description of Activated Processes

As mentioned in the Introduction, a chemical reaction is an activated process, and reactants have to overcome an energy barrier (E^*) during their conversion to products. The transition over the energy barrier occurs on average once per $t_{slow} \approx t_{fast} e^{E^*/kT}$, where t_{fast} is the time-scale characteristic to microscopic dynamics, and kT is the thermal energy. Usually $E^* \gg kT$ and the two time-scales differ by many orders of magnitude (the transition itself is very fast but it occurs rarely). Hence activated processes cannot be studied directly by molecular-dynamics (MD) simulations, as this would require an unfeasible number of MD integration steps before a single transition would occur (the time step used in MD must be smaller than the t_{fast}).

In general, there are several energy barriers involved in a chemical reaction. In this case the description of the overall chemical reaction is decomposed into several elementary steps, where each elementary step is associated

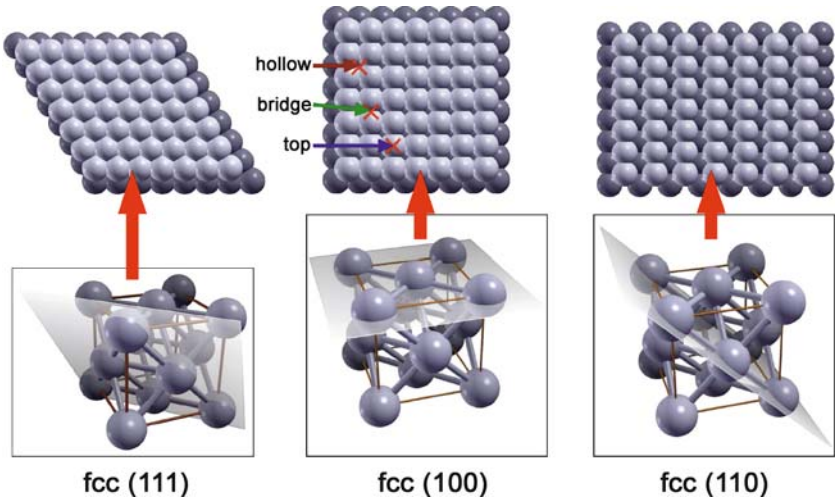

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 3
Three low Miller index surfaces of an fcc (face-centered-cubic) crystal: (111), (100), and (110). In the *bottom row* of plots the direction of the plane used to cut the corresponding surfaces is indicated. The positions of high-symmetry surface sites (*hollow*, *bridge*, and *top*) are also shown on the (100) surface

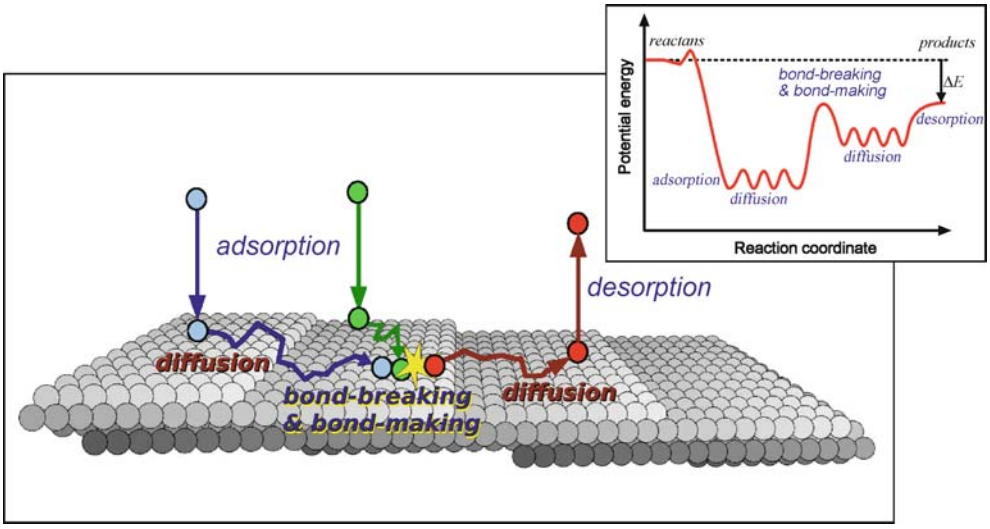

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 4
Chemical reactions consist of several elementary steps. The Figure shows a typical mechanism of surface catalyzed reaction (i. e. Langmuir–Hinshelwood mechanism), which consists of the following reaction steps: adsorption of reactants, their diffusion on the surface, one or more elementary steps where some bonds are broken and new ones are formed so as to form the product molecules, which eventually desorb from the surface. The *inset* shows the associated reaction energy profile

with (at most) one energy barrier. Figure 4 shows elementary steps involved in a typical surface-catalyzed reaction. Therefore, one needs to consider each elementary step in order to understand the overall chemical reaction. The hierarchical approach to the problem would consist of first analyzing each elementary step individually by electronic structure calculations, and then to assemble them into an overall picture by other methods, such as, for example, kinetic Monte Carlo (Sect. "Putting It All Together").

Let us consider an activated elementary step, such as depicted in Fig. 1. A stable molecular configuration on the left side of the barrier is called initial-state (IS) and that on the right side the final-state (FS) of an elementary step, whereas the point of the highest energy along the path

from IS to FS is called a "transition state" (TS), and its energy difference with respect to IS, is the activation energy, E^*:

$$E^* = E_{\mathrm{TS}} - E_{\mathrm{IS}} \,. \tag{5}$$

But for an isolated diatomic molecule, the potential-energy-surface is multidimensional, hence there is an infinite number of trajectories going from IS to FS, and among there is also the minimum-energy-path (MEP), which is characterized by vanishing force components orthogonal to the path. The MEP's tangent is therefore parallel to the force vector and the configuration of the highest energy along the MEP is the TS—a first-order stationary point (see Fig. 5).

The calculation of PES is, due to its multidimensional nature, computationally too demanding to be sampled accurately but for a few simple cases. However, in many instances it turns out that the knowledge of the stationary points on the PES and its shape around them is sufficient

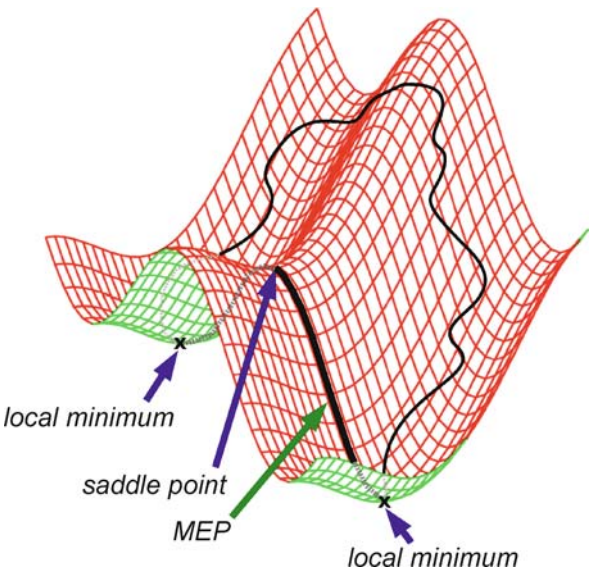

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 5

Two-dimensional potential-energy-surface (PES). Two trajectories are shown linking the two local minima: one of them is the minimum-energy-path (MEP) that is characterized by the vanishing force components orthogonal to the path and passes through the saddle point. In an elementary chemical reaction, reactant (initial-state, IS) and product (final-state, FS) are associated with local minima on the PES, and the reactant is transformed into the product by proceeding along a given trajectory on the PES that links the two minima. Statistically, the most probable trajectory displays the smallest barrier, hence it passes through the saddle point (first-order stationary point). The configuration at the saddle point is called the transition state (TS)

to calculate reaction rates. Accurate estimates of transition rates, can be obtained, for example, by using a statistical approach known as *transition-state theory* (TST) [16,18]. According to TST the (forward) rate constant depends exponentially on the activation energy, E^*, and is given by a van't Hoff–Arrhenius type expression:

$$k_{\mathrm{f}} = \nu \mathrm{e}^{-E^*/kT} \,, \tag{6}$$

where k is a Boltzmann constant (please note that k and k_{f} label two distinct quantities), T is temperature, and ν is a prefactor, which depends on the shape of the PES around the stationary points. The letter "f" in the k_{f} indicates that this is the rate-constant for the forward direction of the elementary step, i. e. from IS to FS. There will also be backward jumps from FS to IS, and the net rate, r, of the elementary step will be given by the difference between the forward and backward rates, $r = r_{\mathrm{f}} - r_{\mathrm{b}}$. The TST also provides an insight into the physical origin of prefactor ν. In the harmonic TST approximation, where the PES around the IS and TS is expanded in normal modes, the prefactor ν can be described in terms of the vibration frequencies ν_i of IS and TS, and Eq. (6) can be written as:

$$k_{\mathrm{f}} = \frac{\prod_i^{3N} \nu_i^{\mathrm{IS}}}{\prod_i^{3N-1} \nu_i^{\mathrm{TS}}} \mathrm{e}^{-E^*/kT} \,. \tag{7}$$

Note that the product in the denominator is only over $3N - 1$ frequencies (N being the number of atoms), because the TS's normal mode tangent to reaction coordinate cancels out in the derivation of TST. On the basis of (harmonic) TST it is therefore clear why the knowledge of the stationary points on the PES and its shape around them suffices (in many cases) to calculate reaction rate. In the one-dimensional case, Eq. (7) would reduce to $k_{\mathrm{f}} = \nu^{\mathrm{IS}} \mathrm{e}^{-E^*/kT}$, where ν^{IS} is the oscillator frequency around the IS. Bond-stretching frequencies are on the order of 10^{12} to $10^{13}\mathrm{s}^{-1}$, and this gives an idea about the typical value of the prefactor in Eq. (6).

Identification of Transition-States

The dependence of the rate constant on the E^* is exponential and only linear on the prefactor ν, Eq. (6). It is precisely for this reason why the identification of the transition state is so crucial: knowing its structure and understanding its chemistry would allow for the design of a reaction site that would lower the energy of TS and consequently increase the rate of the elementary step. The premise behind this is that by "chemically" acting on the reaction-center, the prefactor would not change substantially, and the exponential dependence on E^* would prevail in determining

the rate. Moreover, it is much more intuitive to act on the energetic stability of the TS than on the "shape" of the PES around the IS and TS. As for the overall reaction rate, one would consequently need to lower the energy barrier for all elementary steps to increase the overall rate, but it turns out that in catalytic reactions, some barriers are anti-correlated, that is, lowering a barrier for one elementary step increases the barrier for some other elementary step.

At this point one could ask: why bother with computer atomistic simulations to identify transition states involved in given chemical reactions? The reason is that transition states are very hard to identify experimentally, because of their too short lifetimes. Moreover, the identification of transition states is quite demanding even by computer simulations: usually at least one order of magnitude heavier than the identification of equilibrium configurations (local minima on PES). The reason is that the methods based solely on the first derivatives of PES—the forces—do not converge to saddle points. Among the most popular methods for the identification of saddle points is the nudged-elastic-band method (NEB) [32,33], where instead of searching solely for TS, one considers a path connecting the IS and FS. This path is then discretized into several structures called images. The images are connected by "springs" to prevent them from sliding down the PES into the nearest local minimum during the optimization procedure (the spring forces are allowed to act parallel to the path and the true forces orthogonal to it). The NEB method therefore converges the path to the MEP. The highest point along a stable MEP—which by definition is a saddle point (a first-order stationary point)—is then found by allowing one of the images to "climb up" until the forces vanish. Actually the NEB will converge to MEP's whenever these are stable (i. e. when the curvature of the PES on hyperplanes orthogonal to the MEP itself is everywhere positive). When the MEP's are not stable different techniques based on the use of collective variables should be preferred [110].

Brønsted–Evans–Polanyi Relationship

Because the PES is such an illusive object, chemists have long attempted to develop some concepts that would surpass the explicit knowledge of the PES and even of the transition states. Although these are mainly empirical, they are useful and can be used as guides or in explanatory purposes when searching for improved reaction sites. Various relationships have been proposed, and they are derived from some assumptions of the shape of the potential-energy-surface. Perhaps the most widely used, at least in the field of heterogeneous catalysis, is the *Brønsted–Evans–Polanyi* (BEP) relationship [7,17,62], according to which the activation energy for a set of related elementary reactions correlates linearly with the reaction energy, ΔE (ΔE is defined graphically in Figs. 1 and 7):

$$E^* = a + b\Delta E \,. \tag{8}$$

This relation tells us that the smaller ΔE is, the smaller will be the activation energy. Small here means negative, that is, exothermic or thermodynamically favored. In Eq. (8) there are two parameters, a and b, which give the slope and the intercept of the BEP line. These two parameters were given various names, e. g. the slope is called the reaction coefficient, whereas the intercept is the intrinsic barrier of reaction [62]. It has been shown that the BEP relation holds well for dissociation reactions at surfaces [3,71], because the transition state is *late*, product-like, and variations in the stability of dissociated products are likely to reflect the stability of the transition state. Analogously, for an *early*, reactant-like, transition state the variation in the stability of the reactant will correspondingly affect the stability of the transition state, but the barrier will be little affected. This is schematically shown in Fig. 6. For this reason the BEP slope is large for reactions with *late* transition state, and small for *early* transition state.

Description of the Overall Reaction

Decomposition into Elementary Steps

A heterogeneous catalytic reaction involves several elementary steps, such as the adsorption of the gas-phase reactants on the solid surface, their diffusion on the surface, one or more elementary steps involving breaking some of the reactant chemical bonds and making new ones so as to form the product molecules, which eventually desorb from the surface (see Fig. 4). Consider, for example, ammonia (NH_3) synthesis, which is among the largest industrial catalytic processes (Haber–Bosch process) and one of the most studied [94]. The overall reaction is:

$$N_2 + 3H_2 \rightarrow 2NH_3 \,, \tag{9}$$

and it consists of the following elementary steps [94] (omitting diffusion of adsorbed species):

$$\begin{array}{lr} 1. & N_{2(g)} \rightarrow N_{2(ad)} \\ 2. & N_{2(ad)} \rightarrow 2N_{(ad)} \\ 3. & H_{2(g)} \rightarrow 2H_{(ad)} \\ 4. & N_{(ad)} + H_{(ad)} \rightarrow NH_{(ad)} \\ 5. & NH_{(ad)} + H_{(ad)} \rightarrow NH_{2(ad)} \\ 6. & NH_{2(ad)} + H_{(ad)} \rightarrow NH_{3(ad)} \\ 7. & NH_{3(ad)} \rightarrow NH_{3(g)} \,, \end{array} \tag{10}$$

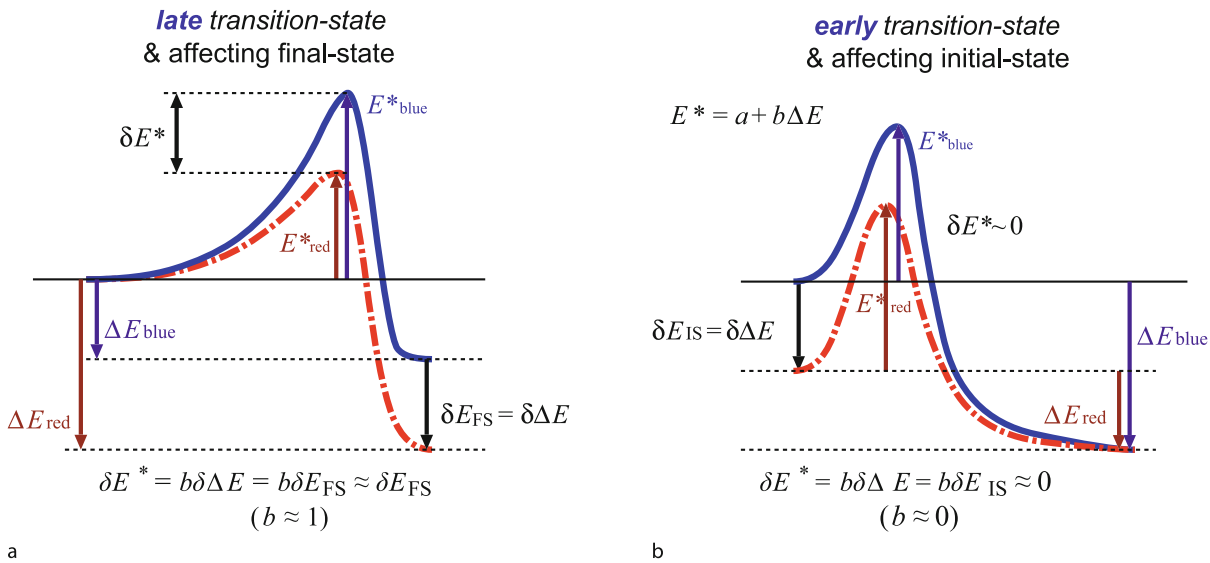

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 6
Effect on the variation of relative stability of final (FS) and initial state (IS) on the activation energy for *late*- and *early*-type transition state (TS), respectively, according to the BEP relation. a For *late* TS the stabilization of FS stabilizes the TS leading to a reduction of the activation energy, *E. b For *early* TS the stabilization of IS stabilizes the TS, but this affects the *E** very little**

where labels (g) and (ad) designate the gas-phase and adsorbed species, respectively. This set of elementary steps can be classified into three groups:

(i) dissociative adsorption of nitrogen (N_2) and hydrogen (H_2), Eqs. (10.1)–(10.3);
(ii) reactions of the dissociated N and H species to form the product NH_3 species: in this step, the adsorbed N is successively hydrogenated, Eqs. (10.4)–(10.6);
(iii) desorption of the product ammonia molecules, Eq. (10.7).

Sabatier Principle

As for the above NH_3 synthesis, a good catalyst must facilitate both the dissociation of reactant N_2 and H_2 molecules, as well as the formation of the product molecules. Because the reactants are the molecules in the gas phase, the reaction energy of dissociation steps is given by the dissociative chemisorption energies. According to the BEP relation, Eq. (8), the more negative (exothermic) is the chemisorption energy the smaller will be the activation energy, and the easier (faster) the formation of intermediates. This indicates that a very reactive surface should be used that forms strong bonds with the dissociated species. However, there is an opposite trend for the formation of product molecules, see Fig. 7 (it is worth mentioning that the BEP slope for association reactions tends to be much smaller than for dissociation reactions [71]). In addition, if the surface is too reactive intermediates will fully cover it thus inhibiting the reaction. Therefore, weakly chemisorbed intermediates would be required to form the product molecules. The best catalyst will be therefore a compromise with an intermediate chemisorption energy. This is the *principle of Sabatier*, which results in the so-called volcano activity plots (see inset of Fig. 7).

Considerations presented above demonstrate why the surface acts as a catalyst: the main effect is to stabilize the intermediates, which reduces the activation energies accordingly. But the stabilization should not be overdone, because then the reaction rate would decrease again.

Putting It All Together

From the preceding section it is rather obvious that the dynamics of the overall reaction may develop pattern(s), that may not be guesses merely by considering elementary steps individually. Namely, macroscopic dynamical properties are, in general, determined by the interplay between many elementary atomic-scale processes. For this reason, after the individual elementary steps have been quantified, they need to be put into a broader context that will link them together so as to construct a coherent picture of the overall reaction.

Mean-Field Description: Rate Equations One could setup rate equations for all elementary steps. The procedure for doing that is well established [15,61,102]. Consider for example, an elementary step where adsorbed

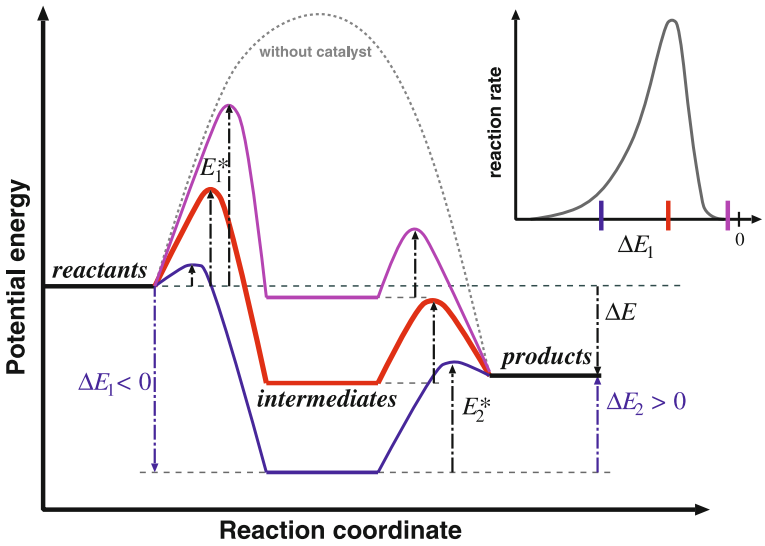

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 7

Surface acts as the catalyst, because it stabilizes the intermediates, and therefore decreases the activation energy. The more they are stabilized (the more negative ΔE_1) the smaller is the activation energy for their formation (*left barrier*, E_1^*), but the barrier for their conversion into products will display the opposite trend (*right barrier*, E_2^*). The optimum is therefore a compromise between the two anti-correlated effects, hence the catalyst should bind intermediates moderately (*Sabatier principle*). *Inset*: the rate of reaction displays a *volcano* dependence on the intermediate chemisorption energy (ΔE_1)

species A and B react so as to form adsorbed species AB:

$$A_{(ad)} + B_{(ad)} \rightleftharpoons AB_{(ad)} \,. \tag{11}$$

Naively speaking, the forward rate, r_f, of this step will be given by the "*probability of* A *and* B *to meet at the same site*" times the "*probability to overcome the barrier*" per unit time. If we define the coverage of species x as $\theta_x = N_x/N$, where N_x is the number of adsorbed x molecules and N is the number of all surface sites, then the first probability is given by the product of the two coverages, $\theta_A\theta_B$, whereas the second probability is given by the forward rate constant, k_f. In this case the rate is normalized to the number of "jumps" per site per unit time. The reverse rate, r_b, will be analogously given by $\theta_{AB} \times k_b$, where k_b is the rate constant for the backward direction (activation energy in the backward direction is $E^*_{bck} = E^* - \Delta E$, e. g. see Fig. 1). The net rate is therefore given by:

$$\begin{aligned} r = r_f - r_b &= \theta_A\theta_B k_f - \theta_{AB} k_b \\ &= \theta_A\theta_B\left[\nu_f e^{-E^*/kT}\right] - \theta_{AB}\left[\nu_b e^{-(E^*-\Delta E)/kT}\right] , \end{aligned} \tag{12}$$

where ν_f and ν_b designate prefactors for the forward and backward direction, respectively. Note, however, that if the *probability* is expressed as $\theta_A\theta_B$ then A and B are treated as independent. Therefore, in this scheme the lateral interactions between the adsorbed species can be treated only by mean-field description through dependence of the activation energies and prefactors on the coverage. After rate equations are constructed for all elementary steps the so-obtained set of equations needs to be solved. The resulting equations may be highly nonlinear and very interesting patterns such as kinetic oscillations and chaos may develop [116]. In order to simplify the treatment several approximations may be applied, such as steady-state (SSA) and quasi-equilibrium approximation (QEA) [61].

Kinetic Monte Carlo A more detailed description of the overall reaction could be obtained with a *kinetic Monte Carlo* (kMC) simulation [19,115], but it requires larger computational effort. In kMC the surface is usually mapped onto a two-dimensional lattice and the adsorbate positions are associated with lattice sites (adsorption for example is simulated by random "arrivals" of molecules to the lattice sites). At each simulation step the rates of all considered elementary processes are evaluated. A given process is then executed by a random selection with appropriately weighted probability (i. e. the rate of jumps is proportional to the associated barriers). In kMC the time evolution is coarse-grained to time scale appropriate for the rare activated events. At each step a detailed atomic configuration is known, and the activation energies can be calculated taking into account proper local environment.

In kMC the number of particles as well as the number of steps is very large, and to calculate all the interactions and activation energies on the fly with ab-initio methods is far from feasible. Therefore, a model Hamiltonian (e.q. *lattice gas* Hamiltonian) is constructed that allows for fast on the fly calculation of required energies. A model Hamiltonian may be built by expanding it into various interaction terms that are determined from a number of *ab initio* pre-calculated structures [85].

Rate-Limiting Step An analysis of the rates of the involved elementary steps reveals that not all the elementary steps are equally significant in determining the rate of the overall chemical reaction. The step that most significantly determines the rate of the overall chemical reaction is called the rate-determining step. This may imply that the rate-determining step is the "slowest" among all elementary steps, however under steady state conditions, by definition, the rates of all the steps will be equal. For example, in the ammonia synthesis mentioned above, the dissociation of N_2 is rate-determining. There may be more than one kinetically relevant steps in a given chemical reaction. From Eq. (12), we see that the rate depends exponentially on activation energy and temperature, whereas the dependence on the other parameters is weaker. Therefore, if a given elementary step has much larger E^* than the others, it is likely to be rate-determining. Note also that under different experimental conditions (T, P) different steps may be rate-determining. For example, under large enough T the E^*/kT ratio will become small enough, and other factors, such as, the availability of sites and partial pressures of reactants may prevail.

Optimum Chemisorption Energy For a class of heterogeneous reactions, where dissociative adsorption is the rate-determining step, like the ammonia synthesis considered above, an analysis of the reaction rates gives a universal result that the optimum chemisorption energy should be, broadly speaking, in the range from -1.0 to -2.0 eV per molecule [3,76]. This universality is a consequence of the BEP dependence of the activation energy on the chemisorption energy. This implies that the stronger the bond in the diatomic molecule, the more reactive will be the "optimum" catalyst. For example, the bond of N_2 is stronger than that of oxygen, O_2, therefore a good catalyst for reactions involving dissociation of N_2 is more reactive than the catalyst for reactions involving the O_2 dissociation. The catalysts for ammonia synthesis are metals such as ruthenium (Ru) and iron (Fe), whereas oxidation reactions typically involve less reactive silver (Ag), palladium (Pd) and platinum (Pt) catalysts.

The general result presented in the paragraph above calls for an in-depth understanding of the factors that govern the dissociative chemisorption energies and these factors will be underlined in Sect. "Theory and Trends of Chemisorption and Reactivity on Transition Metals". Theory and computer simulations have played an important role here. They helped to provide an understanding of chemisorption phenomenon and moreover computer simulations based on DFT have provided systematic databases of dissociative chemisorption energies.

Although the range of optimum chemisorption energies suggested above is relatively broad, it turns out that for a given diatomic molecule this range typically involves only a few elemental metals, and none of them may even lie close to the volcano curve maximum [76]. Hence a fine-tuning of the chemisorption energy would be required so as to move toward the volcano maximum. A simple rational approach would be to form a bimetallic alloy by simple interpolation [39], i. e., by mixing a given metal that lies on the left side of the volcano peak (too strong chemisorption energy, E_{chem}) with the one on the right (too weak E_{chem}). As for the ammonia synthesis discussed above, mixing Mo—which binds N too strongly—with Co —which binds N too weakly—results in a catalyst that is closer to the volcano maximum and therefore more active than the best elemental metals Fe, Ru, and Os [39] (see Fig. 8). This is precisely what was found experimentally [38,47]. Note that both Mo and Co display lower activity than Fe, Ru, and

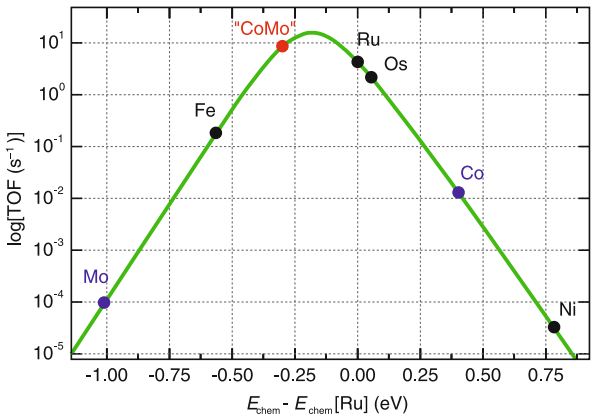

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 8

Calculated volcano plot [39]: catalyst's activity (i. e. turnover frequency, TOF) for ammonia synthesis as a function of chemisorption energy of nitrogen with respect to Ru, $E_{chem} - E_{chem}$[Ru] (eV per N_2 molecule). Curve is calculated by microkinetic model assuming linear BEP relation between nitrogen chemisorption energy and the activation energy for N_2 dissociation. On the basis of results from [39]

Os, and without the knowledge of the principal factor that governs the activity, this would only be guessed by error-and-trial.

For the mixed catalyst to have the interpolated properties, several requirements have to be fulfilled. In particular the constituent metals should not segregate. More issues about alloys will be discussed in Sect. "Theory and Trends of Chemisorption and Reactivity on Transition Metals".

Selectivity Issues

Tuning solely the reactivity of a catalyst—for example according to the above-mentioned Sabatier volcano plot principle—in order to optimize its activity is in many cases not sufficient. Chemical reactions are accompanied with the formation of (unwanted) byproducts. In some cases the selectivity issues may be crucial (selectivity can be defined as a ratio between the amount of desired product obtained per amount of consumed reactants), because some of the reaction byproducts may even poison the catalyst, such as, for example, coke formation during the syngas production (e. g. see Sect. "Improvement of Steam-Reforming Catalyst"). Therefore, it is clear that improving the catalyst's selectivity is of great importance.

DFT simulations can help improve the catalyst's selectivity by disentangling the mechanisms that govern the reactivity of catalysts toward competitive reaction pathways. Consider the formation of two competitive species from a given adsorbed intermediate:

$$\begin{aligned} A_{(ad)} &\rightarrow B \\ A_{(ad)} &\rightarrow C\,. \end{aligned} \tag{13}$$

If B and C are adsorbed species, then the BEP relation suggests that in order to favor the formation of, for example, B species the reaction center should be designed such as to relatively enhance the magnitude of adsorption energy of B with respect to C. However, if the species B and C are gas-phase molecules, or if the reactions do not obey the BEP relation, the design of the proper reaction center that will favor the formation of wanted species will be more difficult. A possible strategy is to act on stabilities of involved transition states by designing such a reaction center that will stabilize the TS of the desired reaction channel with respect to those of undesired pathways. To do so the transition states have to be identified: knowing their structure and understanding their chemistry would allow one to act on their stability. In this respect, the understanding of the factors that determine the adsorbate–substrate bonding should be most helpful. For this reason a simple chemisorption model and its applicability under different circumstances will be presented below. Then it will be demonstrated with a few examples how quantum mechanical computer simulations can be used to help in designing novel reaction-centers with tailored properties.

Theory and Trends of Chemisorption and Reactivity on Transition Metals

Ab initio electronic structure methods, such as DFT, allow one to calculate the properties needed for the quantitative description of chemical reactions, such as equilibrium and transition state structures, their energies, and vibrational properties. On the other hand, one needs a simple model, which would describe in terms of a few parameters the trends of these properties, and therefore provide understanding of the most important factors that govern them.

Below a short description of the Hammer–Nørskov chemisorption model [28,30] is given, which is based on the Newns–Anderson chemisorption theory [75]. This model provides a simple description of the chemisorption trends on transition metal (TM) surfaces. Consider, for example, that a single one-electron state of an adsorbed atom interacts with a TM surface. This state will interact with all the valence electron states of the surface metal atoms, which can be described with *s*- and *d*-band (i. e. valence electrons of a given (isolated) TM atom are described by discrete *s*- and *d*-states, but in a solid these states form *s*- and *d*-band). The former is very broad and approximately half filled for all TM (containing one electron/TM-atom). The *d*-band is instead much narrower, as shown in Fig. 9a. The delocalized *s*-states interact weakly with the adsorbate state, resulting in its broadening, whereas the localized *d*-states (which can be approximately considered as molecular state) interact strongly forming the adsorbate–surface bonding- and antibonding-states. This is schematically shown in Fig. 9b. Because the *s*-band and its interaction with the adsorbate is similar for all TM, the variation in the adsorbate–surface bonding is described merely by considering the *d*-band in the Hammer–Nørskov model. The variation of the adsorbate–surface bonding is determined mainly by two effects: (i) the amount of empty antibonding adsorbate–surface states; note that anti-bonding interaction is repulsive, so the more these states are empty the stronger is the adsorbate–surface interaction, and (ii) Pauli type repulsion arising from the orthogonalization of the overlapping adsorbate and metal states. This term is proportional to the square of coupling matrix elements between adsorbate state and metal *d*-states.

As for the effect (i), the amount of the empty antibonding states is related to the position of the *d*-band with respect to the Fermi level, ε_F (Fermi level is the energy of the highest occupied state in the metal). In the Hammer–

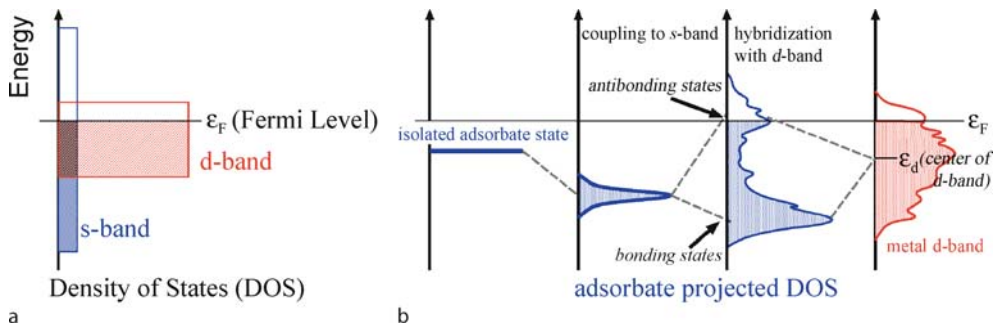

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 9
a Electronic structure of transition metal (TM); b schematic illustration of adsorbate interaction with TM surface

Nørskov model this position is described by the center of the *d*-band, ε_d. The higher is the ε_d with respect to ε_F the more empty are the antibonding states and the stronger is the adsorbate–surface bond. This explains why the reactivity of TM decreases from left to right in the Periodic Table: the more the *d*-band is occupied the more below the Fermi level it is (lower ε_d). Figure 10 shows the trend of the DFT calculated chemisorption energies of oxygen on close packed surfaces of 4*d* transition metals. It is due to this trend and the Sabatier principle that mainly *late* transition metals (i. e. those with more than a half filled *d*-band) are catalytically interesting, because *early* transition metals (i. e. those with less than a half filled *d*-band) bind adsorbates too strongly. The reactivity also reduces when going down from the 3*d* to 5*d* row in the Periodic Table, because the latter states are more extended and therefore the effect (ii) is more repulsive.

For a given adsorbate, its interaction to the surface can be affected not only by changing the metal (as shown in Fig. 10), but also by other means that are more subtle, and can be used to fine-tune the adsorbate–surface interaction. Consider what happens to the reactivity of a given *late* transition metal if the width of the *d*-band is reduced by some means. DFT calculations reveal that in such cases no charging or decharging of *d*-states occurs [30], and in order to keep the number of *d*-electrons fixed the *d*-band has to move up causing an up-shift of its center thus increasing the adsorbate–surface interaction. This is schematically presented in Fig. 11. In general, the less *effectively* a given TM atom is bonded the smaller is the width of its local *d*-band (the density-of-states projected to *d*-states of the considered metal atom), resulting in increased reactivity for late TM. The variation of the width of the *d*-band can be achieved, for example, by varying the coordination of surface atoms or by strain. Consider for example (111) and (100) low Miller index surfaces of an fcc metal (these two surfaces are shown in Fig. 3): the former has nine-fold coordinated surface atoms, whereas the latter has eight-

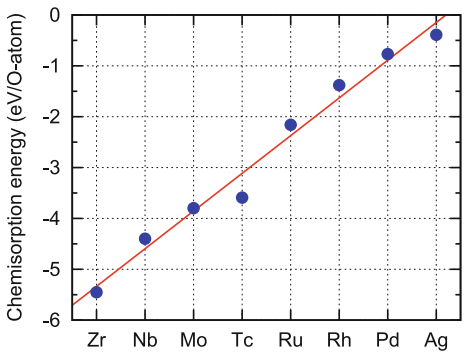

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 10
DFT calculated dissociative chemisorption energies of oxygen with respect to gas-phase O_2 molecule (in eV per O-atom) on closed-packed surfaces of 4*d* transition metals. On the basis of results from Refs. [13,54]. The *line* is drawn to guide the eye

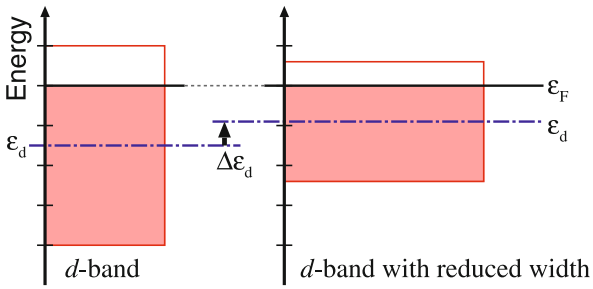

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 11
Effect of the reduced band width on the center of the *d*-band for *late* (more than a half filled *d*-band) transition metal: in order to keep the number of electrons fixed (number of electrons is given by the colored area below ε_F), the *d*-band has to shift up, resulting in the $\Delta\varepsilon_d$ shift of the center of the *d*-band, ε_d. The effect would be the opposite for *early* (less than a half filled *d*-band) transition metal: the ε_d would be down-shifted

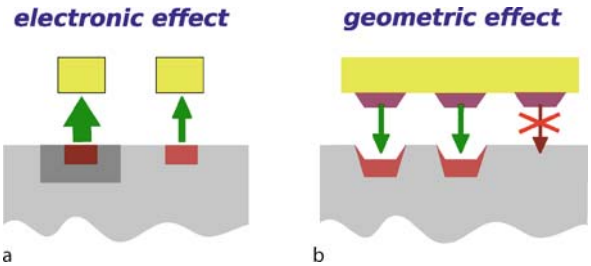

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 12
Schematic representation of electronic and geometric effects. a The adsorbate–surface interaction can be tuned by changing the electronic properties of the binding site by acting on its environment (electronic effect). b The adsorbate–surface interaction is hindered if the ensemble of the active binding site on the surface does not match the "geometric" requirements of the adsorbate (geometric effect)

fold coordinated surface atoms. Surface atoms of the (111) facet are therefore effectively more bonded than those of the (100) facet, and therefore we may anticipate that the former binds simple adsorbates less than the latter.

Considerations presented above were made for an atom (with a single state) interacting with the TM surface, but the arguments can be generalized to molecules by considering several adsorbate states interacting with the surface. The above *electronic* effects are still valid (e. g. see Fig. 13 for the correlation between the *d*-band center and molecular adsorption and activation energies), but in addition to that there is another effect, which is related to the geometry of the molecule and surface. Indeed, the two types of effects are termed *electronic* (or ligand) and *geometric* (or ensemble) effects, and provide a basic framework for tuning the reactivity of surface reaction centers. The way the two effects work is presented schematically in Fig. 12, and they may be described as:

- Electronic (ligand) effect is due to modifications of catalytic properties due to the interaction of a surface atom that the molecule binds to with its neighbors (e. g. shift of the d-band center).
- Geometric (ensemble) effect is related to the number of active surface atoms required for a given molecule to bind.

Above we implicitly assume that the larger the adsorbate–surface bond strength the larger the reactivity of the surface (i. e. the two terms were used casually; here surface reactivity toward dissociation reactions is meant). This can by justified by the BEP relation and also by the fact that the interaction of TS with the surface can be treated like a molecule–surface interaction, hence the stronger is this interaction the lower is the energy barrier.

Special Active Sites at Surfaces

What we have not mentioned explicitly so far is that not all sites on any real surface are alike. For example, lattice imperfections are always present on surfaces. This is particularly the case for industrial catalysts which consist of small dispersed particles on a ceramic support and display a large variety and concentration of defects [99]. It may be anticipated on the basis of the *d*-band model that such structural variations should influence the reactivity. Indeed, nowadays a large amount of experimental evidence indicates that the binding of adsorbates at surfaces is strengthened by the presence of low-coordinated

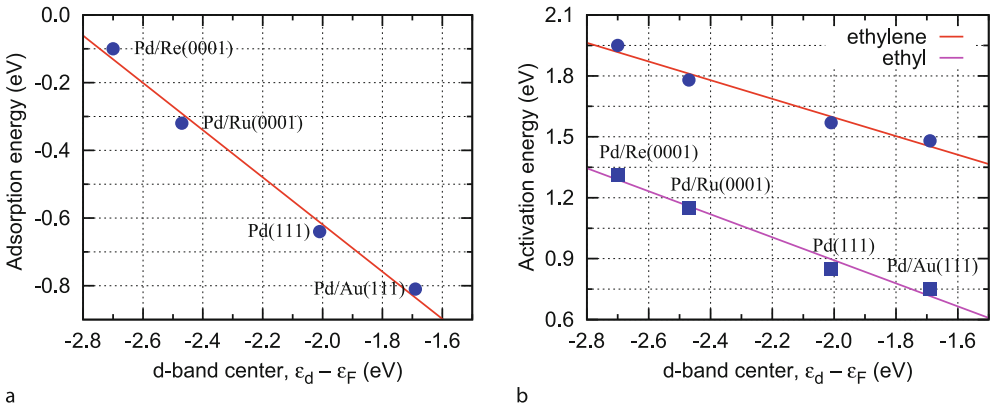

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 13
a DFT calculated ethylene adsorption energy as a function of the *d*-band center (ε_d) of the Pd(111) surface and commensurate Pd overlayer over close packed Re(0001), Ru(0001), and Au(111) surfaces. b DFT calculated C–H bond activation energies for ethylene (C_2H_4, ●) and ethyl (C_2H_5, ■) dehydrogenation versus Pd *d*-band center for the same systems as in a. On the basis of results from [77]

defects [83,93,98,112]. Below we shortly survey how the reactivity can be tuned by strain, defects and alloying.

Strain With computer simulations the effect of strain can be very easily modeled by merely increasing/decreasing the lattice spacings. Although such calculations are fictitious, they are useful to discern the effects of strain on the reactivity, and show that surface reactivity increases with tensile strain and vice-versa [65]. As anticipated on the basis of the *d*-band model, this is due to the fact that a reduced overlap between the *d*-states at neighboring atoms reduces the width of the *d*-band. Experimentally the strained sites can be made, for example, by growing atomically thin commensurate overlayers on support with different lattice spacings [9,87,95]. This way both compressive and tensile strained overlayers can be achieved (depending on the relative lattice spacings of support with respect to metal constituting the overlayers). An example is shown in Fig. 13, which shows the adsorption energy of ethylene (C_2H_4) and activation energies for dehydrogenation of ethylene and ethyl (C_2H_5) on palladium overlayers over rhenium (Re), ruthenium (Ru), and gold (Au) [77]. Correlation between these energies and the *d*-band center is rather good. The Pd overlayer on Au is more reactive than pure Pd, because lattice spacings of Au are larger than that of Pd and vice-versa for Ru. So Pd overlayer experiences tensile and compressive strain on Au and Ru, respectively. Rhenium is an exception, because of very similar lattice spacings to that of Pd, yet the reactivity trend is captured by the *d*-band center, which is apparently affected by the electronic interaction between Pd and Re.

Surface Defects The importance of surface defects in determining the reactivity of heterogeneous catalysts can hardly be overestimated [30]. Not only do defects usually bind adsorbates more strongly than perfect facets do, but in certain cases the reactivity of the step defects is so much larger than that of the terraces as to dominate the dissociation reaction rate at concentrations as low as 1% of step atoms [14]. This can be attributed both to electronic and geometric effects. As for the former, the width of the *d*-band of atoms at defects is reduced due to their decreased coordination (up-shift of ε_d), as shown in Fig. 14. The reactivity is then further increased by geometric effect, because of the availability of new sites. The extent of the two effects is apparent from Fig. 15, which displays the TS for NO dissociation under three different circumstances [26]: (a) shows the TS structure over the perfect facet of Ru(0001), and in (b) the same TS structure (no geometric effect) is shown near the step-edge defect. The reduction of E^* by 0.5 eV as going from (a) to (b) is therefore due to increased reactivity of step atoms (*electronic effect*). The barrier then further decreases by 0.7 eV, by taking advantage of the geometry of the step (*geometric effect*), because the TS structure shown in (c) benefits from being highly coordinated. Note that the TS's for the dis-

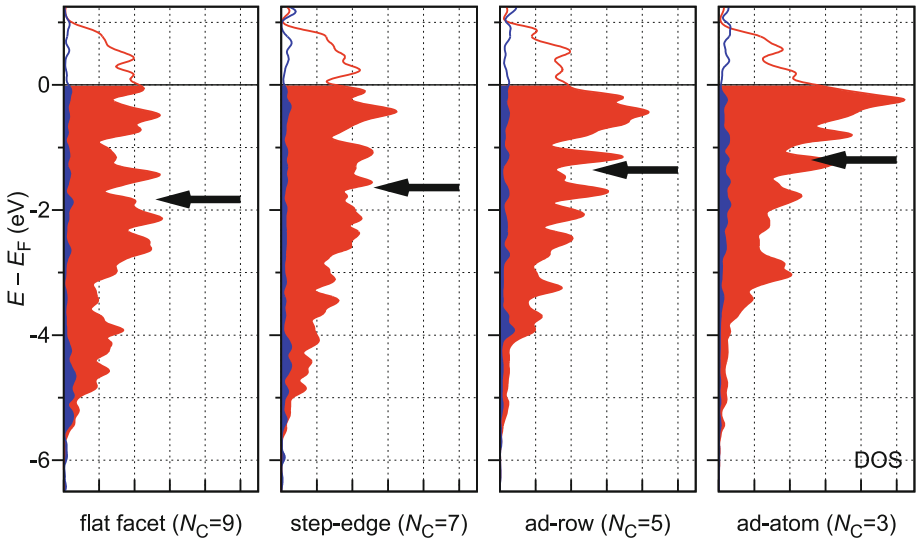

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 14
Density of states (DOS) projected to progressively less coordinated metal atom at the Rh surface ($N_C \equiv$ coordination number). From left to right: perfect (111) facet and step-edge, ad-row, and ad-atom defects thereon. The *d*- and *s*-band are colored *red* and *blue*, respectively. *Arrows* indicate the position of the *d*-band centers (ε_d). On the basis of results from [48]

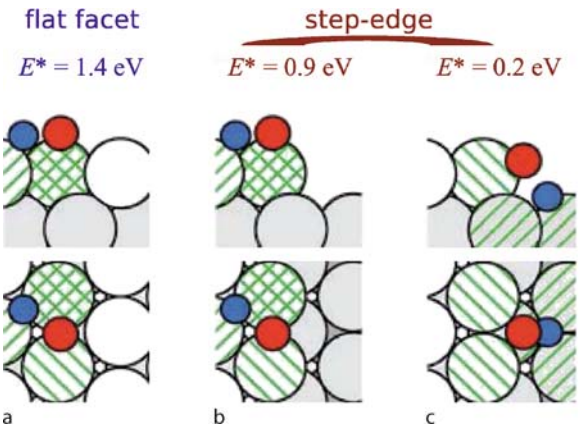

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 15

Activation energies and transition state structures for NO dissociation over flat Ru(0001) facet a, and step-edge defect thereon b,c [26,30]. In b the TS structure is analogous to that shown in a, whereas in c the advantage of the step-edge geometry is taken into account. Note that the reaction center at flat facet a involves three surface Ru atoms, where that of step-edge c involves five surface Rh atoms. From [26,27]. With kind permission from Springer Science and Business Media

sociation of diatomic molecules such as N_2, NO, O_2, and CO are late [71,76], atomic-like, and isolated atoms are "unhappy" (reactive) and they want to be as highly coordinated as possible. So on one hand, the coordination of metal atoms at the step-edge defect is reduced, but on the other hand the step-edge defects provide the surface sites with high coordination. Therefore, at step-edges the electronic and geometric effects are cooperative resulting in their large reactivity toward dissociation reactions.

Indeed, the surface defects are so much more reactive than flat terraces, that even gold—the noblest among metals [29]—under certain circumstances becomes reactive enough to display very interesting catalytic properties [37]. For example, gold is inert toward the oxidation of carbon monoxide ($CO+\frac{1}{2}O_2 \rightarrow CO_2$), because of its inability to dissociate oxygen molecules. However, small gold nanoparticles supported on oxides (see Fig. 16) have been found to be active [31,107]. This behavior may be attributed to several effects: (i) the abundance of low coordinated reactive sites, which increases with decreasing the particle size (these sites appear at the intersections of low Miller-index planes and include edge, kink and even adatom defects) [66,84]; (ii) strain induced due to oxide-support and small particle size [66], and (iii) the appearance of new special sites at the perimeter of the oxide–metal interface (e. g. due to oxide-induced charge transfer to adsorbate) [60,72,73,74].

Alloying Alloying presents a very interesting way of changing the reactivity of the surface. On one hand addition of another metal (impurity) into a given host metal changes its electronic structure, but on the other hand it opens a way to tailor a selectivity by a geometric effect. The electronic effect on the reactivity of alloys is well captured by the shift of the *d*-band center, and these shifts have been calculated by DFT and tabulated for many catalytically interesting bimetallic alloys [88], thus providing a guideline for an on-demand tuning of the electronic structure of the surface. An interesting and counterintuitive example of alloying is that in some cases the addition of an inert IB metal enhances the adsorption ener-

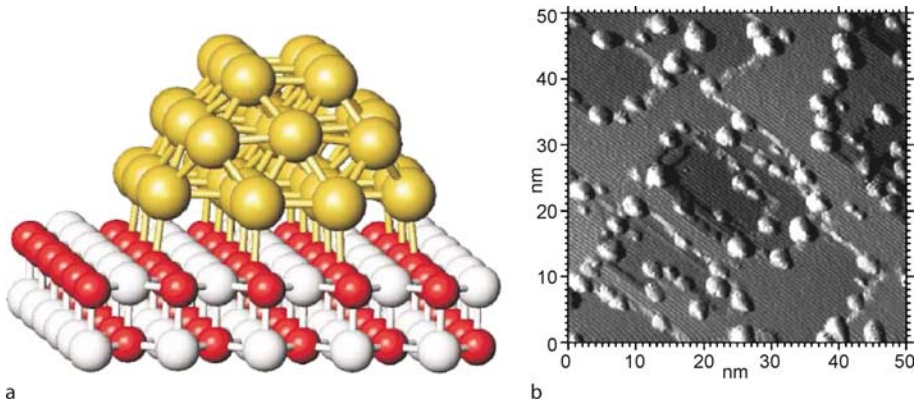

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 16

Low coordinated surface defects are so reactive that even gold—the noblest among metals [29]—becomes catalytically active in the limit of very small nanoparticles. a A model of gold nanoparticles on a MgO(100) support. Reprinted figure with permission from [72]. b An STM (scanning-tunneling-microscope) image of gold nanoparticles on TiO_2(110). From [107], reprinted with permission from AAAS

gies on more reactive VIII group metals [28,77]. This is mainly due to strain [88], i. e., if a "small" metal atom is added into a "larger" host, it will experience a tensile strain and vice-versa (and the *d*-band center will shift accordingly).

As for the geometric effect, it has long been known from experiments that the addition of a group-IB metal (Cu, Ag, Au) to group-VIII metals (Fe, Co, Ni; Ru, Rh, Pd; Os, Ir, Pt) can lead to an increased selectivity of the catalyst [97]. In particular, it is known that such catalysts hinder hydrogenolysis reactions which involve C–C bond breaking, but do not significantly affect dehydrogenation reactions which involve C–H bond breaking [97]. For example, the hydrogenolysis reaction require sites with a larger ensemble of active atoms than dehydrogenation reactions. Already the addition of a small amount of IB metal will substantially decrease the number of sites composed of a large ensemble of active metal atoms, thus disfavoring reactions that require large ensembles of active atoms (as shown schematically by Fig. 12b) and making the dehydrogenation reactions—which require a small ensemble of active atoms—more selective [97].

Tuning the surface composition of an alloy represents an exciting way of designing new reaction centers with tailored catalytic properties. However, the actual composition cannot be varied at random. Some combinations of metals may lead to segregation, while for others the composition at the surface differs from that in the bulk (the amount of a given metal at the surface is either enriched or depleted). An extensive database of surface segregation energies have been compiled by DFT calculations [89], and the segregation tendency may be explained by surface energy and crystal structure differences between the host and impurity. Even if the two metals do not form a bulk alloy (i. e. they are immiscible), they may form an alloy on the surface. Such alloys have been shown to have interesting catalytic properties [2,22,23,44,97]. An example of a AuNi surface alloy used to catalyze steam reforming will be presented in Sect. "Improvement of Steam-Reforming Catalyst".

An exiting novel class of near surface alloys have been considered recently where the impurity metal atoms are located primarily in the subsurface layer [22,44] (see Fig. 17). The DFT calculations indicate that such alloys yield superior catalytic behavior for hydrogen related reaction, because they bind atomic hydrogen weakly, yet they are able to dissociate the H_2 molecule efficiently [22]. As for the water–gas shift reaction ($CO + H_2O \rightarrow CO_2 + H_2$), such a type of subsurface alloy was also suggested to be promising due to the ability to bind CO weakly and to dissociate water easily [44].

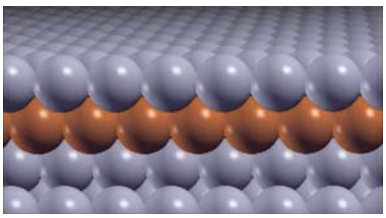

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 17
An idealized model of a near-surface alloy (NSA) with impurity metal in the subsurface layer. NSA's have interesting catalytic properties [2,22,23,44,97]

Examples

As described above the reactivity of special sites on the surface can be tuned in several ways, which are relatively easy to achieve with computer DFT simulations, because there we are assembling a given structure "atom by atom", and hence we have total control over it. Computer simulations allow one to consider and investigate structures under various circumstances that are otherwise unfeasible in reality with the aim to gain new understanding. One could envisage many different surface reaction centers and some of them may have very appealing catalytic properties. Below a few examples are presented with the aim to demonstrate how new reaction centers may be proposed on the basis of insight gained by careful analysis of results of quantum mechanical computer simulations.

Tuning the Relative Rates of Methane Dehydrogenation

One of the problems in the conversion of methane (CH_4)—the principal component of natural gas—to more useful liquid-phase chemicals, such as methanol, is that, although many catalysts are able to cleave the C–H bonds, they do so in a nonselective manner. The efficiency of catalysts is thus limited by the tendency of dehydrogenation to proceed until graphite is eventually formed on the surface, thus poisoning the catalyst [20]. Although a symbolic reaction equation may look as simple as $CH_4 + ? \rightarrow CH_3OH + \cdots$ it appears in reality so difficult that—despite its importance —a direct conversion of methane to methanol (or other liquid chemicals) is not yet feasible. For this reason let us ask a simpler question: is it possible to design a reactive center that will cleave the C–H bonds of methane selectively?

By careful analysis of the results of DFT computer simulations, it has been shown how to design such a reaction center that effectively activates the first dehydrogenation

step while hindering the second one [48,49,114]:

$$\begin{aligned} &1.\quad CH_4 \rightleftharpoons CH_3 + H\,, \\ &2.\quad CH_3 \rightleftharpoons CH_2 + H\,, \end{aligned} \tag{14}$$

The activation energies can be used as a measure of reaction rate constants (e. g. Eq. (6), $k_f = \nu e^{-E^*/kT}$), therefore a target reaction center has to reverse the natural order of the two barriers' heights. This means that the barrier for the first step has to be lower than that for the second.

Local Geometry of the Reaction Site Kokalj et al. [48] have investigated the effect of local geometry of the reaction center on the two energy barriers. These authors considered the Rh(111) surface because Rh is known for its reactivity toward the C–H bond cleavage of CH_4 [103]. The local geometry of the reaction center was varied by "creating" progressively less coordinated defective sites on the surface, passing from perfect (111) facet with nine-fold coordinated surface atoms (coordination number, $N_C = 9$), to step-edge ($N_C = 7$), added row ($N_C = 5$), and finally to an ad-atom defect ($N_C = 3$). These surface structures are shown in Fig. 18, and the corresponding *d*-band projected density of states is shown in Fig. 14. For each one of them, the DFT calculated optimal adsorption energies of the involved CH_4, methyl (CH_3), ethylene (CH_2), and hydrogen (H) species are reported in Fig. 19a. The adsorption of CH_4, CH_3, and CH_2 is rather sensitive to the coordination of the metal atoms at the binding site, whereas the adsorption of hydrogen is not so much so.

As for CH_3 and CH_2, we may see that they are stabilized at low coordinated defects, as expected on the basis of the Hammer–Nørskov *d*-band model, except for the ad-atom defect, where the trend is reversed. Moreover, the ad-atom behaves rather unexpectedly also with respect to C–H activation energies, as shown in Fig. 19b: although the barrier for the first dehydrogenation step decreases as expected with the coordination of the reaction center, the dehydrogenation of methyl ($CH_3 \rightarrow CH_2 + H$) is hindered at an ad-atom defect, where the first dehydrogenation step is instead most favored. Therefore, the behavior of the ad-atom is so peculiar that neither can the adsorption of CH_3 and CH_2 be explained solely by the Hammer–Nørskov *d*-band model (*electronic* effect) nor can the dehydrogenation barriers be explained by the BEP relation. As will be shown below this is related to structural features of the ad-atom (*geometric* effect). We proceed by analyzing the involved structures, in particular, adsorbed CH_3, CH_2,

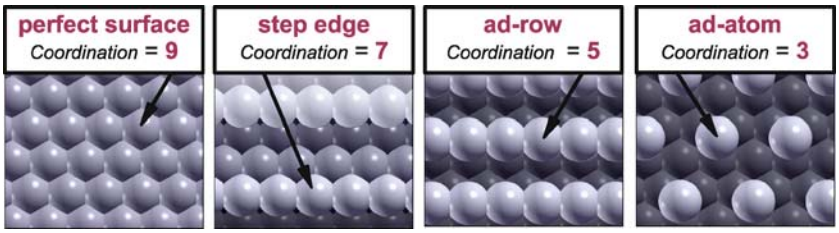

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 18
Different local geometries of the reaction centers on Rh(111). From *left to right*: perfect surface, step-edge, ad-row, and ad-atom. Reprinted with permission from [48]

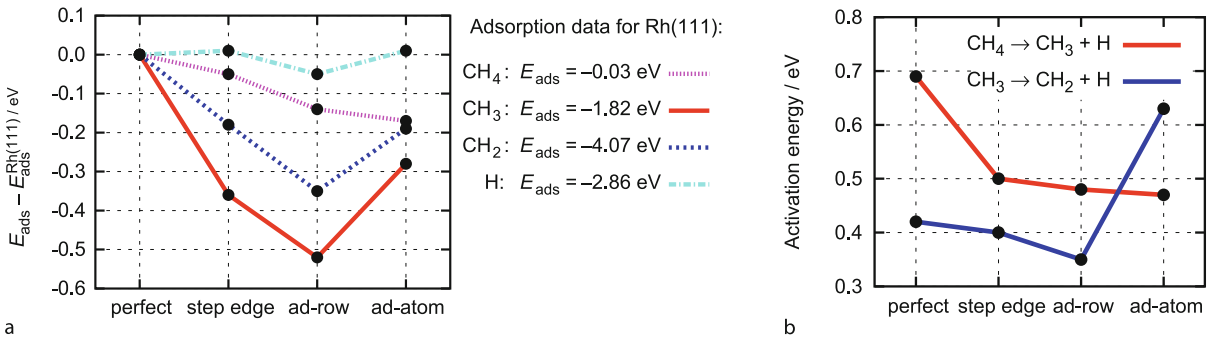

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 19
a Adsorption energies of methane (CH_4), methyl (CH_3), methylene (CH_2) and atomic hydrogen (H) on a step-edge ($N_C = 7$), ad-row ($N_C = 5$), and ad-atom ($N_C = 3$) with respect to the Rh(111) facet ($N_C = 9$). b Activation energies for the first two steps of methane dehydrogenation as a function of local geometry of the reaction site. On the basis of results from [48]

and the corresponding TS structures for the two dehydrogenation steps.

Analysis of CH_3 Adsorption Methyl is a radical whose unsaturated C atom strongly binds to the metal surface. But in addition to this C–metal bond, methyl also displays a peculiar three-center C–H–metal bond, which is in part responsible for the peculiar behavior mentioned above. The three-center C–H–metal bond is usually referred to as *agostic bond* in organometallic chemistry [12,25,111]. The *agostic* bond results essentially from the hybridization of 1*e* C–H bonding orbitals with the *d* states of the metal surface [68,78]. *Agostic* bonds occur when the adsorption geometry allows for a small H–metal distance. Consider for example a methyl radical adsorbed at an fcc hollow site with H atoms pointing toward either the nearest hollow sites (no agostic bonds, Fig. 20a) or toward the nearest metal atoms (thus giving rise to agostic bonds, Fig. 20b). The difference between the two adsorption energies is rather large (≈ 0.4 eV); also note that in the latter case, the C–H distance is slightly larger than in the former. This effect is enhanced when going from the (111) facet to the step edge and ad-row: the C–H distance increases, Fig. 20, and more charge is delocalized from the C–H bond toward the Rh atom, Fig. 21. On the other hand, at an ad-atom defect the H–metal distance is quite large, indicating a smaller *agostic* interaction, and consequently explaining the reversed trend of methyl adsorption energy at the ad-atom.

Agostic bonding also stretches and correspondingly weakens the C–H bond [69], thus helping break it, and this may partially explain why the ad-atom reaction center is not so efficient for methyl dehydrogenation.

Analysis of CH_2 Adsorption To explain the reversed trend of methylene adsorption energy at the ad-atom, note that to couple the two unpaired electrons of the methylene diradical, CH_2 has to form (at least) two bonds with the substrate. The local structure of an ad-atom defect, however, is such that this requirement is not easily fulfilled. As passing from a (111) facet to a step edge and an added row, CH_2 binds to progressively less coordinated atoms, however at an ad-atom defect it bridges the ad-atom ($N_C = 3$) to a surface atom underneath ($N_C = 10$). The average coordination number of the two bridged atoms is thus $N_C = 6.5$, close to the value $N_C = 7$ of the step edge. As a result, the adsorption energy of methylene at an ad-atom and at a step edge are very similar: -4.26 and -4.25 eV, respectively (see Fig. 19a).

Structural Analysis of Transition States An analysis of a large number of identified TS structures, not only those on currently considered reaction centers, but also on Ru(0001) [11], Ni(111) [52,70,108], Pd(100) [113], and Pt(110)(1 × 2) [82] reveals that the TS's display some universal features [49,114], which are evident from Fig. 22, and can be described as:

(i) For both reactions the TS is *late*, i. e., its structure is close to that of the final state: the C–H distance for the detaching H atom at the TS is in the range 1.6–1.7 Å, to be compared with an equilibrium bond length of 1.1 Å.
(ii) The TS of the first dehydrogenation step involves only one metal atom: the CH_3 fragment is located at the top site.
(iii) On the other hand, the TS for the second dehydrogenation step involves two metal atoms: the CH_2 (methylene) fragment is located at the bridge site, while the dissociating H atom is at the top site.

We further notice that although the adsorption energy of H is rather insensitive to the coordination of the metal reaction center (see Fig. 19), this is so only for the best adsorption sites corresponding to each reaction center (hollow or bridge). However, at the TS of the $CH_3 \rightarrow CH_2 + H$, the H atom is located instead close to a top site (see Fig. 22b). The hydrogen adsorption energies on top of a metal atom are also rather similar for the (111) facet, at step edge, and at ad-row (≈ -2.45 eV). However, on top of the ad-atom the adsorption is about 0.2 eV less stable, and this further contributes to the large value of the E^* at an ad-atom defect.

Requirements for Reaction Center The insight obtained on the basis of the analysis presented above, may be used to design new reaction centers, which would enhance the first reaction while hindering the second, by not only acting on the local geometry, but also on the local chemical composition. On the basis of the above analysis the following three requirements should be met [49]:

(i) One of the CH_2–metal bonds should be substantially weakened;
(ii) The strength of the H–metal bond at the top site should be reduced, and
(iii) *Agostic* C–H–metal bonding of CH_3 should be prevented.

Note that *agostic* interactions are small, hence the effect of *agostic* bonding on the reaction barrier can only be small. On the other hand, the CH_2–surface and H–surface bonds are strong, and a large reduction of these bond strengths can affect the reaction barrier substantially. It turns out

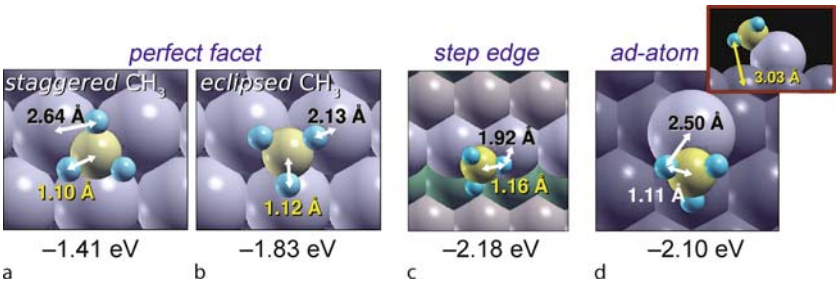

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 20

a,b Staggered and eclipsed orientation of methyl at an fcc site of perfect Rh(111). In **a** the orientation of CH_3 is such that H atoms are staggered with the nearest metal atoms, whereas in **b** they are eclipsed thus forming agostic bonds. **c** CH_3 adsorbed at the step edge with one of the H atoms pointing toward the step metal atom thus forming the agostic bond. **d** CH_3 at the ad-atom: agostic bonding is hindered due to large H-metals distances. Adapted from [49]

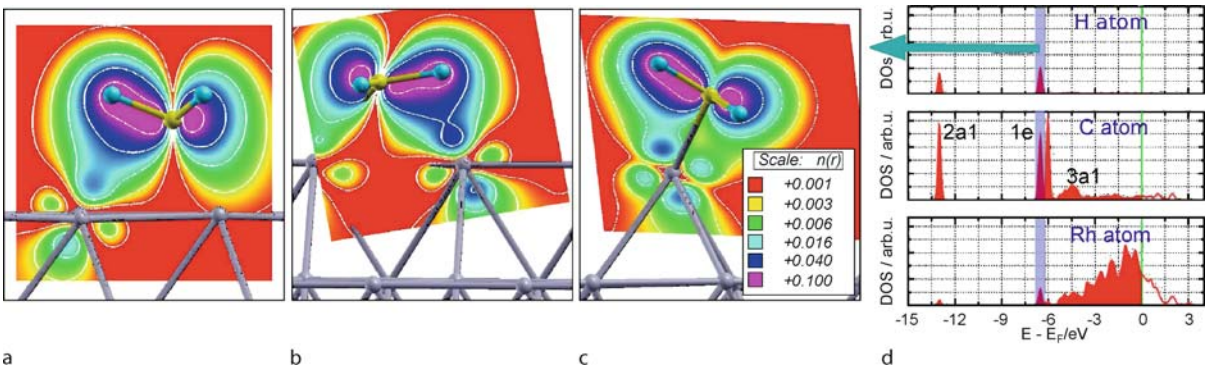

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 21

a, b, c Electron charge density, integrated over the energy window marked by a *blue stripe* in **d**, illustrating the extent of three center C–H–metal *agostic* bonding of methyl adsorbed on **a** a Rh(111) facet, **b** a step edge, and **c** an ad-atom. The magnitude of ILDOS increases from *red to violet* following a *rainbow* scale. Five contours are drawn in logarithmic scale from 10^{-1} to $10^{-3} e/a_0^3$. **d** Density of states projected to H, C, and Rh atom. Adapted from [49]

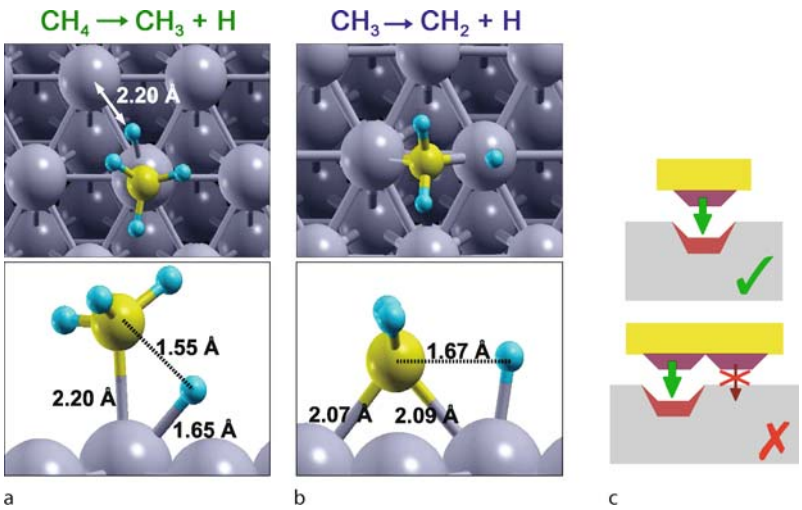

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 22

a, b Main features of the transition state structures for the first two dehydrogenation reactions on a flat Rh(111) surface. **c** A reaction center composed of a single reactive site is not compatible with the structural characteristics of TS shown in **b**, and would therefore destabilize it. Adapted from [49]

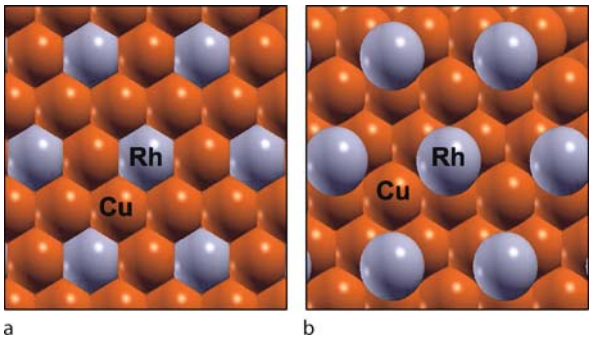

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 23
a Rh atom substitutionally embedded into Cu(111) and b Rh ad-atom on Cu(111). Reprinted with permission from [49]

that the above requirements can be fulfilled more or less simultaneously.

Local Chemical Composition of Reaction Site The requirements just outlined, suggest that a proper reaction center should be composed of solely one reactive atom, embedded, for example, in an inert environment (shown schematically in Fig. 22c). Consider, for example, an isolated reactive Rh atom substitutionally embedded into a more inert Cu(111) surface, such as shown in Fig. 23a [49]. This can be seen as a model for a Cu-rich phase of a RhCu alloy. Bulk RhCu alloy shows two phases: a Rh-rich phase (with Cu concentration, $X_{Cu} \lesssim 0.1$) and a Cu-rich phase (with $X_{Cu} \gtrsim 0.8$) [63]. However, on the surface the amount of Cu is enriched, e. g., already at a small amount of Cu ($X_{Cu} > 0.05$) the Cu-rich phase will form on the surface [10].

At this reaction center (Fig. 23a), the CH_2 forms one strong bond with the substitutional Rh atom, while the other bond with a Cu atom is much weaker. In addition, agostic bonding is also prevented. Therefore, it performs much better than the Rh(111): the calculated E^* for the first dehydrogenation step is 0.70 eV, whereas that for the second is as large as 0.84 eV (to be compared with 0.69 and 0.42 eV at the Rh(111), respectively) [49]. Snapshots of the two reactions are shown in Fig. 24. Given that the two reaction barriers on clean Cu(111) are predicted to be 1.7 and 1.5 eV [49], dehydrogenation would selectively occur near the Rh atom.

Tuning the chemical composition of the reaction center therefore substantially increases the barrier for the second dehydrogenation step, but hardly affects the barrier for the first step in the current case. However, by combining the structural effects with the chemical effects just outlined, it is possible to selectively increase the barrier of the second dehydrogenation step, while reducing that

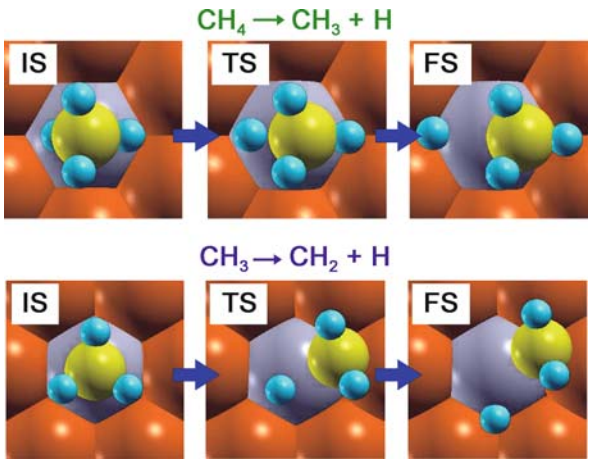

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 24
Initial- (IS), transition- (TS), and final-state (FS) structures for the first two steps of methane dehydrogenation over the Rh atom substitutionally embedded into Cu(111). Reprinted with permission from [49]

of the first. Consider for example a Rh ad-atom adsorbed onto a Cu(111) surface, as shown in Fig. 23b [49]. In this case, the reaction barrier for the first reaction is very small, $E^* = 0.35$ eV, while the barrier for the second reaction is quite large, $E^* = 0.89$ eV [49]. In order to illustrate the effects of the local atomic structure and local chemical composition on the two dehydrogenation steps, the reaction energy profiles at the considered reaction centers are shown in Fig. 25.

The reaction center based on metal ad-atoms, such as just considered (Fig. 23b), is merely an academic model system, because the naked metal ad-atoms would either cluster or diffuse into the bulk. Nevertheless, the kind of arguments and analysis presented and utilized above are instrumental to the understanding of the mechanisms responsible for the reactivity of reaction centers on real catalysts and to the design and realization of new materials with tailored catalytic properties. As for the ad-atoms, the problem is of course open on how to stabilize them. For example, Zhang and Hu [114] considered isolated Pt ad-atoms on an oxide MoO_3(010) surface, and calculated that the barrier for methane dehydrogenation is significantly lower than that on a Pt(111), while the further dehydrogenation of methyl is blocked.

Improvement of Steam-Reforming Catalyst

Today, the conversion of methane to liquid chemicals is utilized via an in-direct route: in a first step the methane is transformed into a syngas (a mixture of H_2 and CO). Then in the second step the liquid chemicals such as, for exam-

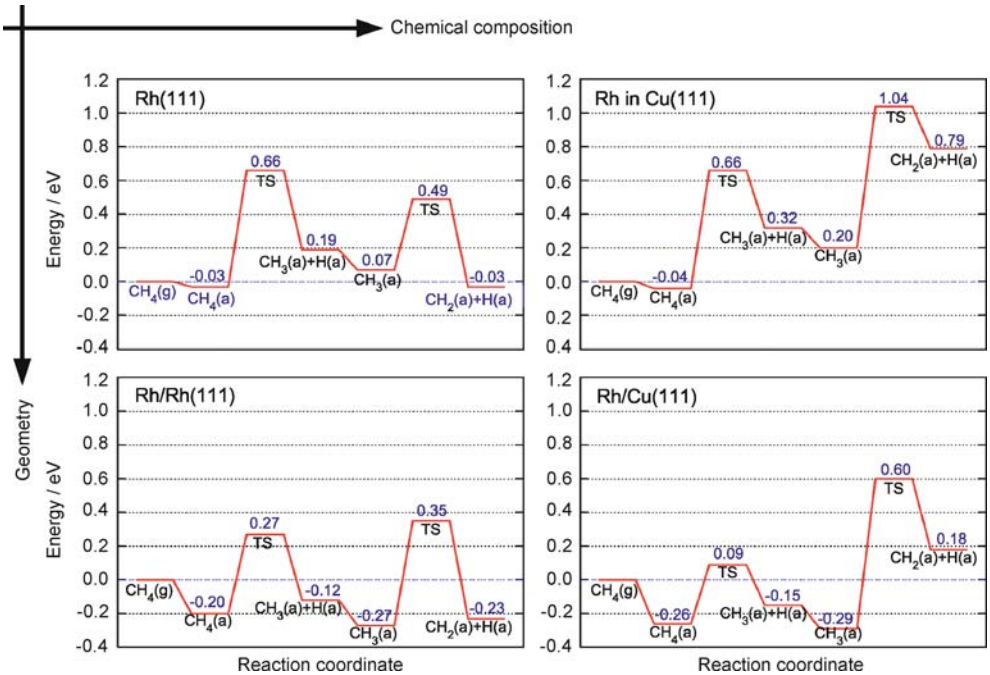

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 25

Reaction profiles for the two dehydrogenation reactions. *Arrows* indicate the improved parameter of the reaction center. The zero level is the energy of the gas phase methane. Difference between the labels "CH_3(a)+H(a)" and "CH_3(a)" is that for the latter the H(a) diffused to an equivalent site far from CH_3(a). Labels Rh in Cu(111), Rh/Rh(111), and Rh/Cu(111) stand for Rh atom substitutionally embedded into Cu(111) (Fig. 23a), Rh ad-atom on Rh(111), and Rh ad-atom on Cu(111) (Fig. 23b), respectively. Reprinted with permission from [49]

ple, methanol are synthesized from syngas. Note, however, that a direct route to methanol synthesis would surpass a very expensive syngas production.

The syngas may be produced via several processes, one of them being the steam reforming:

$$CH_4 + H_2O \rightleftharpoons CO + 3H_2 \quad (\Delta H^0_{298K} = 2.14\,\text{eV})\,. \quad (15)$$

The elementary reaction steps can be classified into two groups, and can be written as (omitting adsorption, diffusion, and desorption steps):

(i) dissociation of reactants:

$$\begin{aligned} &1. \quad CH_4 \rightarrow CH_{3(ad)} + H_{(ad)} \\ &2. \quad CH_{3(ad)} \rightarrow CH_{2(ad)} + H_{(ad)} \\ &3. \quad CH_{2(ad)} \rightarrow CH_{(ad)} + H_{(ad)} \\ &4. \quad CH_{(ad)} \rightarrow C_{(ad)} + H_{(ad)} \\ &5. \quad H_2O \rightarrow OH_{(ad)} + H_{(ad)} \\ &6. \quad OH_{(ad)} \rightarrow O_{(ad)} + H_{(ad)} \end{aligned} \quad (16)$$

(ii) formation of products:

$$\begin{aligned} &1. \quad H_{(ad)} + H_{(ad)} \rightarrow H_2 \\ &2. \quad O_{(ad)} + C_{(ad)} \rightarrow CO \end{aligned} \quad (17)$$

On the basis of detailed investigation of elementary processes by means of atomistic DFT computer-simulations and surface-science experiments, Besenbacher et al. [2] designed an improved high surface area alloy catalyst for steam-reforming process. The process is obviously limited by the ability of a catalyst to activate hydrocarbon molecules, and its tendency to form an undesired graphite, $C_{(ad)} \rightarrow C_{(s)}$, which poisons the catalysts. As already stated, catalysts are able to break C–H bonds, but in a nonselective manner, which is precisely what is done in steam-reforming, where methane is stripped off all the H atoms, Eqs. (16.1)–(16.4). At this point, two issues are crucial: (i) the surface should prevent naked $C_{(ad)}$ atoms from clustering so as to avoid the graphite growth, and (ii) the $C_{(ad)}$ should not bind too strongly to the surface: recall that the less the $C_{(ad)}$ binds to the surface the easier it will react with chemisorbed $O_{(ad)}$.

A traditional catalyst in steam-reforming is based on Ni as the active element [45]. Kratzer and Besenbacher et al. [2,52] considered adding a small amount of Au into the Ni. The idea is to prevent the $C_{(ad)}$ atoms from clustering. Although the two metals do not form a bulk alloy—they are immiscible in the bulk—they do form an alloy in the

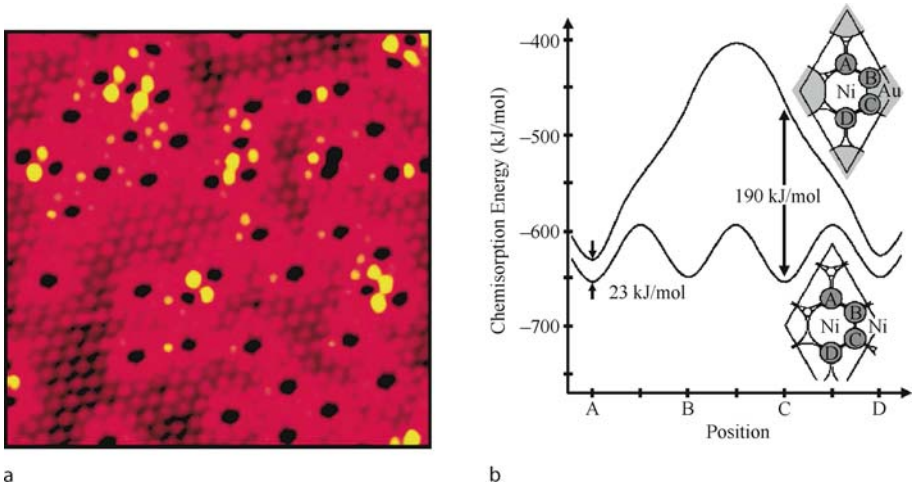

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 26
a STM image of a surface NiAu alloy with 7% of a surface Au atoms. Au atoms are black and the neighboring Ni atoms are brighter than other Ni atoms indicating the modification in electronic structure. b DFT calculated adsorption energy of a C atom as a function of its position on the surface; *bottom* curve shows the diffusion profile of the C atom on clean Ni(111) along A–B–C–D positions shown in the *bottom inset*. *Top* curve shows the same diffusion profile for a situation where one Ni atom was substituted by an Au atom as shown by *top inset*. From [2], reprinted with permission from AAAS

outermost surface layer, as shown by an STM image in Fig. 26a.

As already discussed in Sect. "Tuning the Relative Rates of Methane Dehydrogenation", the transition states for dehydrogenations are similar on various transition metals, so in the current case they will be similar to those displayed in Fig. 22 for the first two dehydrogenation steps. DFT calculations by Kratzer et al. [52] predicted that the addition of Au into Ni(111) decreases the reactivity toward the C–H bond breaking, in accord with experimental observation [36]. The activation barrier for the first dehydrogenation step increases by 0.17 eV at the reaction center involving one Au atom, and more than twice that for the center with two Au atoms. This is at variance with what was predicted on RhCu alloy [49] (Sect. "Tuning the Relative Rates of Methane Dehydrogenation"), where the barrier for CH_4 dehydrogenation was almost not affected by the presence of Cu. This can be explained by the fact that Ni and Au atoms differ substantially in size, whereas Cu and Rh do not. Addition of "larger" Au impurity onto the Ni layer therefore down-shifts the *d*-band center of Ni thus decreasing its reactivity [52]. Nevertheless, the effect of addition of Au on dehydrogenation is relatively modest compared to the effect it has on naked $C_{(ad)}$ atoms and on formation of graphite. In particular, a Au atom makes all the nearest neighboring hollow sites completely unstable for adsorption of carbon atoms (by almost 2 eV), and also destabilizes by 0.24 eV the second nearest neighbor sites, as shown in Fig. 26b. This has a tremendous consequence: a single Au atom excludes six surface Ni metal atoms from binding with $C_{(ad)}$ (corresponding to six nearest neighbor hollow sites), and additionally makes the next neighboring sites less stable and therefore somewhat more susceptible to CO formation. Therefore, even a small amount of Au prevents the $C_{(ad)}$ atoms from clustering and therefore from graphite being formed.

That actually all works as suggested by DFT calculations, the authors have performed several experiments: (i) they first synthesized a Au/Ni catalyst in a high surface area form on a $MgAl_2O_4$ support, then (ii) they performed a surface structure characterization by means of EXAFS to confirm the Ni and Au actually formed surface alloy, and finally (iii) they measured the steam-reforming activity of a catalyst. They showed that, while the Ni catalyst deactivates rapidly, the activity of the Au/Ni was almost constant under the measured time interval (see Fig. 27) [2].

It is instructive to compare the RhCu model catalyst discussed in Sect. "Tuning the Relative Rates of Methane Dehydrogenation", designed to suppress the dehydrogenation of methyl, to the current NiAu catalyst: the AuNi catalyst consists of individual inert atoms in a reactive substrate, whereas the RhCu model catalyst consists of isolated reactive atoms embedded in an inert (or less reactive) substrate. This is entirely due to the geometric (ensemble) effect: The C atom requires sites composed of a large ensemble of active surface atoms, whereas for the dehydro-

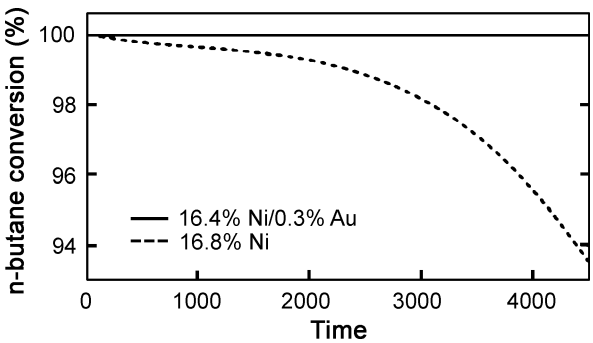

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 27
Conversion of *n*-butane on Ni (*dashed line*) and NiAu surface-alloy (*solid line*) as a function of time during steam reforming. From [2], reprinted with permission from AAAS

genation of methyl already two reactive neighboring atoms are sufficient.

Ethylene Epoxidation

To further illustrate how quantum mechanical simulations help disentangle the mechanisms responsible for the reactivity and selectivity of a catalyst, we consider one more example, the epoxidation of ethylene, which is one of the most important selective oxidation processes based on heterogeneous metallic catalysis [1,53,92,96]. Ethylene epoxide (oxirane) is an important intermediate in the fabrication of glycols and polyols, and silver is a remarkably good catalyst for this reaction with a selectivity at high temperature and high pressure up to 80% in the presence of promoters such as Cl and Cs [53,92,96]. Although a symbolic reaction equation may look as simple as:

$$C_2H_4 + \tfrac{1}{2}O_2 \rightarrow C_2H_4O\,, \tag{18}$$

the elucidation of the reaction mechanism has been a subject of long debate. Its understanding has emerged only recently and was made possible by the synergy between the surface-science experiments and DFT computer simulations. Early studies have mainly focused on the identity of active oxygen species in the reaction and tried to resolve the longstanding question whether molecular or atomic species are the active oxidant [8,21,91]. It has been established only recently that the reaction proceeds via the surface oxametallacycle (OMC) intermediate, which then transforms either to ethylene epoxide (EO) or to acetaldehyde (Ac) [5,42,43,55,56,57,59,67,100,101], the latter reaction leading to undesired total combustion. The overall reaction can be described by the following minimal-sequence of steps:

$$\begin{array}{lr} 1. & O_{2(g)} \rightarrow O_{2(ad)} \\ 2. & O_{2(ad)} \rightarrow 2O_{(ad)} \\ 3. & C_2H_{4(g)} \rightarrow C_2H_{4(ad)} \\ 4. & C_2H_{4(ad)} + O_{(ad)} \rightarrow OMC_{(ad)} \\ 5a. & OMC_{(ad)} \rightarrow EO \\ 5b. & OMC_{(ad)} \rightarrow Ac\,. \end{array} \tag{19}$$

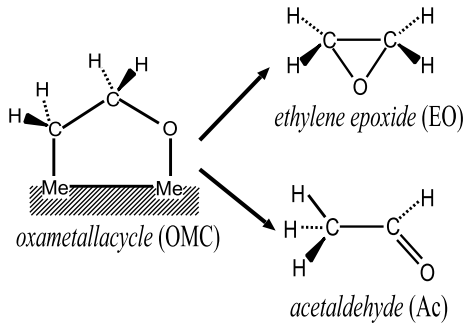

For a more elaborate list of elementary steps see, for example, the work of Stegelmann [101]. For ethylene epoxidation to occur, the catalyst must be reactive enough to dissociate the oxygen molecule [53], but it should be also mild enough so as to form and to stabilize an OMC intermediate, rather than to break C–H bonds, either in the ethylene or in the OMC intermediate [51]. Silver fulfills all these requirements quite well: it is among the most appropriate metals for reactions involving dissociation of oxygen (Sect. "Optimum Chemisorption Energy"), does not activate the C–H bonds [51,92,96] and also binds the OMC weakly to the surface [50,56,64] so that its transformation to EO is facile (as will be shown below). As for dissociation of oxygen, for example, palladium would be also acceptable (Sect. "Optimum Chemisorption Energy"), but at variance with Ag is able to break the C–H bonds [90] (see also Fig. 13b).

Provided that the C–H bond activation is not an issue, the main reaction byproduct involves the formation of Ac [51,56,57]. The Ac molecule is thermodynamically more stable than the EO by about 1 eV. The reason that the EO can be synthesized is because a proper catalyst makes the barrier for its formation smaller or at least comparable to the barrier for Ac formation. The epoxidation of ethylene is therefore an example of a kinetically favored reaction. Linic et al. [56,57] showed that the selectivity of a catalyst is determined by the branching reaction of OMC to form EO rather than the competitive Ac. In particular, at given temperature it would mainly depend on

the difference between the Ac and EO activation energies, $\Delta E^* = E^*_{\mathrm{Ac}} - E^*_{\mathrm{EO}}$ (assuming the two corresponding prefactors are similar). The selectivity, S, can be approximately expressed as:

$$S = \frac{r_{\mathrm{EO}}}{r_{\mathrm{EO}} + r_{\mathrm{Ac}}} \approx \frac{\mathrm{e}^{-E^*_{\mathrm{EO}}/kT}}{\mathrm{e}^{-E^*_{\mathrm{EO}}/kT} + \mathrm{e}^{-E^*\mathrm{Ac}/kT}} = \frac{1}{1 + \mathrm{e}^{-\Delta E^*/kT}}, \quad (20)$$

where r_{EO} and r_{Ac} are the rates of EO and Ac formation, respectively. Therefore, the larger is the ΔE^* the larger is the selectivity. The ability to stabilize the transition state for EO formation with respect to that for Ac formation would therefore lead to increased selectivity.

Bocquet et al. [4] realized on the basis of DFT calculations that the selectivity is linked to the conformational arrangement of the OMC intermediate. By considering the trajectories of atoms during the EO and Ac formation and the structure of the two transition states, they suggested that staggered OMC conformation favors the formation of Ac, whereas eclipsed OMC conformation favors the EO formation. This is due to the fact that the Ac formation involves the shift of the H-atom from the first C atom to the second C atom, which is assisted if the two CH_2 groups of the OMC intermediate are in staggered conformation. This is schematically shown in Fig. 28. Hence, if on a given surface both staggered and eclipsed forms exist, the difference between the two activation energies would be related to the relative energy stability of the two conformational OMC forms.

Kokalj et al. [51] provided on the basis of DFT calculations an alternative explanation as to what determines the selectivity. These authors applied the BEP relationship for the EO and Ac formation on a number of different circumstances, including (111) and (100) surfaces of Rh, Cu, Ag, and Au. This analysis revealed that the formation of Ac follows remarkably well the BEP relation, whereas the formation of EO does so to a lesser extent. By knowing that the EO and Ac bind weakly to the surface [51,64], the main message from this BEP analysis was that the stronger is the OMC–surface interaction the larger are the two barriers [51], which demonstrates that for a facile OMC to EO and Ac transformation a less reactive catalyst is required, such as, for example, silver.

Although an approximate magnitude of the activation barriers for the EO and Ac formation can be determined from the BEP principle, the BEP principle alone is not accurate enough to estimate the selectivity of a catalyst toward the EO formation (see Fig. 30b). The identification of other important factors that determine the catalyst's selectivity toward the formation of EO with respect to Ac formation was made possible by the structural analysis of the TS's for both reactions. Figure 29 schematically depicts the TS's for the EO and Ac formation from OMC on Ag(100). The main difference between the two is that, in TS for EO formation the C–surface bond is fully broken, whereas in TS for Ac formation both C– and O–surface bonds are only partially broken [51]. This may suggest that the two bonds contribute differently to the two activation energies. In particular the stronger is the OMC's O–surface bond with respect to the C–surface bond, the more selective the substrate will be toward EO formation. This argument was quantified by decomposing the OMC–surface interaction into C– and O–surface contributions, which were approximated by the interaction of methyl (CH_3·) and methoxy

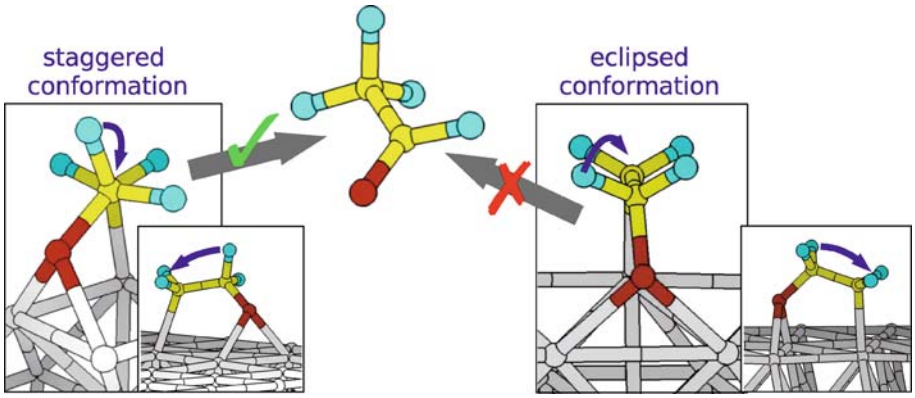

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 28
The formation of acetaldehyde (Ac) proceeds through the staggered conformation of the two CH_2 groups in oxametallacycle (OMC). Bocquet et al. [4] realized on the basis of DFT calculations that the selectivity in the OMC branching reaction toward either EO or Ac is related to the relative stabilities of eclipsed and staggered conformations of the OMC

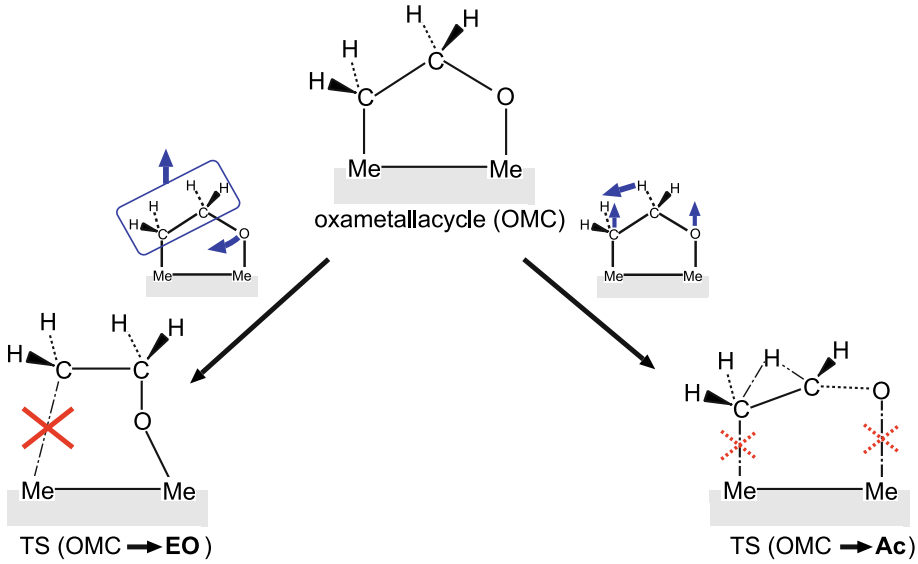

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 29
Schematic presentation of an OMC intermediate and transition states for EO (*left*) and Ac (*right*) formation. In the TS leading to EO the C–surface bond is broken, whereas in the TS leading to Ac both C– and O–surface bonds are elongated. During the EO formation the O–atom moves beneath the ethylene fragment, which is concomitantly shifted upward (indicated by *blue arrows*) and as a consequence the C–metal bond is ruptured, whereas in the formation of Ac the whole molecule is upshifted during the 1,2 hydrogen shift, and consequently the C– and O–surface bonds are partially broken. Based on [51]

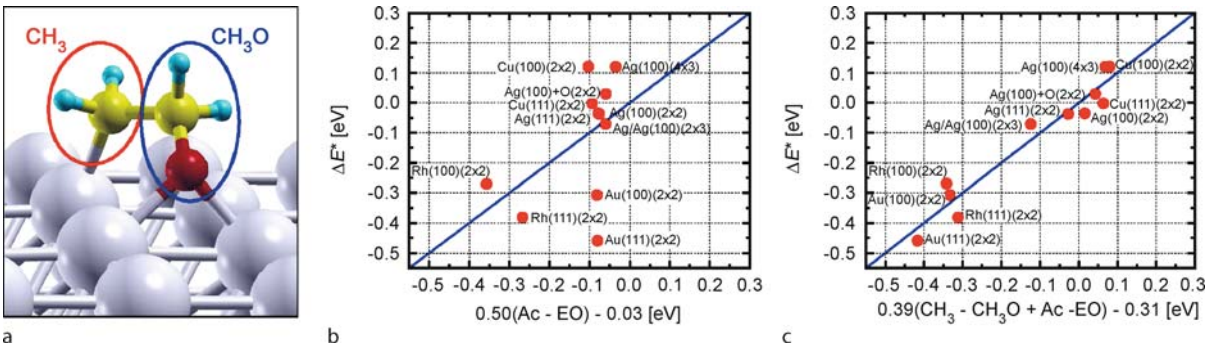

Computer-Aided Design of the Reaction Site in Heterogeneous Catalysis, Figure 30
a The OMC–surface interaction can be decomposed into C– and O–surface terms, which can be approximated by the CH_3– and CH_3O–surface interaction, respectively. This leads to an indicator embodied in Eq. (21) that predicts the $\Delta E^* = E^*_{Ac} - E^*_{EO}$ remarkably better than the standard BEP relation. b Standard BEP and c indicator of Eq. (21) in predicting the ΔE^*. The rms and the maximum errors are 0.19 and 0.38 eV, respectively. Reprinted from [51]

($CH_3O\cdot$) radicals with the surface, respectively, as shown in Fig. 30a. By combining this observation with the BEP relation and after some trivial algebra, an equation for the ΔE^* was derived. This equation was than fitted to calculated data, and that resulted in the following expression [51]:

$$\Delta E^* \simeq 0.39\left[E_{ads}^{CH_3} - E_{ads}^{CH_3O} + E_{ads}^{Ac} - E_{ads}^{EO}\right] - 0.31\,, \quad (21)$$

where E_{ads} stands for adsorption energy of corresponding species. The quality of the fit given by Eq. (21) is compared to that obtained from the simple BEP relation in Figs. 30b and 30c. The indicator as embodied in Eq. (21) is able to estimate ΔE^* with an accuracy better than 0.1 eV (the rms and maximum errors are 0.05 and 0.07 eV, respectively), and it suggests that ΔE^* is mainly determined by two contributions: (i) *differential bonding affinity* of the catalyst toward the O- and C-atoms of the OMC, and (ii) the difference between the adsorption energies of the two final states, Ac and EO.

Recently, Linic et al. [58] showed on the basis of DFT computational screening, that Cu/Ag alloy should display a greater selectivity toward EO compared to pure

Ag, which was later confirmed by surface science experiments [40,41]. However, no explanation was given of why this is so. This finding can be straightforwardly explained by the model presented above [51]. The reasons are that Cu shows larger bonding affinity toward the O-atom with respect to the C-atom than silver (this also explains why Cu is intrinsically more selective than Ag for OMC transformation into EO [105,106]). Because Cu is also more reactive than Ag, this implies that on diluted $Ag_{1-x}Cu_x$ alloys ($x \ll 1$, assuming no Cu-Cu nearest neighbors), the most stable OMC binds its O-atom to Cu and its C-atom to Ag (for instance, such OMC orientation is preferred by 0.16 eV on Cu/Ag(100)), enhancing the O–metal bond strength and making the C–metal bond "relatively" weaker. According to the *differential bonding affinity* term of Eq. (21), this makes the formation of EO more selective. In particular, on Cu/Ag(100) alloy ΔE^* is larger than on Ag by 0.17 eV, where 0.04 eV is due to BEP contribution, and 0.10 eV is due to *differential bonding affinity* [51].

Future Directions

Electronic structure calculations, such as DFT, have reached the point, where they can provide understanding of the atomic-scale details of the elementary steps and the mechanisms underlying the reactivity of a catalyst toward specific chemical reactions. This knowledge can then be exploited in the search for better catalysts, and surpasses the *traditional* error-and-trial procedure. In particular, applications are emerging, where new materials are screened computationally. Ever increasing computer power will make it possible to screen large numbers of new materials in the future, and this has great potential in the search for new materials with tailored catalytic properties.

Bibliography

1. Barteau MA (2006) Surface science and the advancement of direct olefin epoxidation. Surf Sci 600:5021
2. Besenbacher F, Chorkendorff I, Clausen BS, Hammer B, Molenbroek AM, Nørskov JK, Stensgaard I (1998) Design of a surface alloy catalyst for steam reforming. Science 279:1913
3. Bligaard T, Nørskov JK, Dahl S, Matthiesen J, Christensen CH, Sehested J (2004) The Brønsted–Evans-Polanyi relation and the volcano curve in heterogeneous catalysis. J Catal 224:206
4. Bocquet M-L, Loffreda D (2005) Ethene epoxidation selectivity inhibited by twisted oxametallacycle: A DFT study on Ag surface-oxide. J Am Chem Soc 127:17207–17215
5. Brainard RL, Madix RJ (1989) Oxidation of tert-butyl alcohol to isobutylene oxide on a silver(110) surface: the role of unactivated carbon-hydrogen bonds in product selectivity. J Am Chem Soc 111:3826
6. Brivio GP, Trioni MI (1999) The adiabatic molecule–metal surface interaction: Theoretical approaches. Rev Mod Phys 71(1):231–265
7. Brönsted JN (1928) Acid and basic catalysis. Chem Rev 5:231–338
8. Campbell CT, Paffett MT (1984) The role of chlorine promoters in catalytic ethylene epoxidation over the Ag(110) surface. Appl Surf Sci 19:28–42
9. Campbell RA, Rodriguez JA, Goodman DW (1992) Chemical and electronic properties of ultrathin metal films: The Pd/Re(0001) and Pd/Ru(0001) systems. Phys Rev B 46(11):7077–7087
10. Chou S-C, Yeh C-T, Chang T-H (1997) Adsorption of hydrogen on dispersed copper-rhodium bimetallic crystallites. J Phys Chem B 101:5828
11. Ciobica IM, Frechard F, van Santen RA, Kleyn AW, Hafner J (2000) A DFT study of transition states for C–H activation on the Ru(0001) surface. J Phys Chem B 104:3364
12. Crabtree RH (1995) Aspects of methane chemistry. Chem Rev 95:897
13. Crawford P, Hu P (2007) Trends in C–O and C–N bond formations over transition metal surfaces: An insight into kinetic sensitivity in catalytic reactions. J Chem Phys 126:194706
14. Dahl S, Logadottir A, Egeberg RC, Larsen JH, Chorkendorff I, Törnqvist E, Nørskov JK (1999) Role of steps in N_2 activation on Ru(0001). Phys Rev Lett 83:1814–1817
15. Dumesic JA (1999) Analyses of reaction schemes using De Donder relations. J Catal 185:496–505
16. Evans MG, Polanyi M (1935) Some applications of the transition state method to the calculation of reaction velocities, especially in solution. Trans Faraday Soc 31:875
17. Evans MG, Polanyi M (1936) Further considerations on the thermodynamics of chemical equilibria and reaction rates. Trans Faraday Soc 32:1333–1360
18. Eyring H (1935) The activated complex in chemical reactions. J Chem Phys 3(2):107–115
19. Fichthorn KA, Weinberg WH (1991) Theoretical foundations of dynamical Monte Carlo simulations. J Chem Phys 95(2):1090–1096
20. Gesser HD, Hunter NR (1998) A review of C-1 conversion chemistry. Catal Today 42:183
21. Grant RB, Lambert RM (1985) A single crystal study of the silver-catalysed selective oxidation and total oxidation of ethylene. J Catal 92:364
22. Greeley J, Mavrikakis M (2004) Alloy catalysts designed from first principles. Nature Matter 3:810
23. Greeley J, Thomas Jaramillo F, Bonde J, Chorkendorff I, Nørskov JK (2006) Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. Nature Matter 5:909
24. Grimley TB, Pisani C (1974) Chemisorption theory in the Hartree–Fock approximation. J Phys C: Solid State Phys 7:2831–2848
25. Hall C, Perutz RN (1996) Transition metal alkane complexes. Chem Rev 96:3125
26. Hammer B (1999) Bond activation at monatomic steps: NO dissociation at corrugated Ru(0001). Phys Rev Lett 83(18):3681–3684
27. Hammer B (2006) Special sites at noble and late transition metal catalysts. Top Catal 37:3
28. Hammer B, Nørskov JK (1995) Electronic factors determining the reactivity of metal surfaces. Surf Sci 343:211
29. Hammer B, Nørskov JK (1995) Why gold is the noblest of all the metals. Nature 376:238

30. Hammer B, Nørskov JK (2000) Theoretical surface science and catalysis - calculations and concepts. Adv Catal 45:71–129
31. Haruta M, Yamada N, Kobayashi T, Iijima S (1989) Gold catalysts prepared by coprecipitation for low-temperature oxidation of hydrogen and of carbon monoxide. J Catal 115:301
32. Henkelman G, Jonsson H (2000) Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. J Chem Phys 113:9978
33. Henkelman G, Uberuaga BP, Jonsson H (2000) A climbing image nudged elastic band method for finding saddle points and minimum energy paths. J Chem Phys 113:9901
34. Hermann K, Bagus PS, Nelin CJ (1987) Size dependence of surface cluster models: CO adsorbed on Cu(100). Phys Rev B 35:9467
35. Hohenberg P, Kohn W (1964) Inhomogeneous electron gas. Phys Rev 136(3B):B864–B871
36. Holmblad PM, Larsen HJ, Chorkendorff I (1996) Modication of Ni(111) reactivity toward CH_4, CO, and D_2 by two-dimensional alloying. J Chem Phys 104:7289
37. Hvolbæk B, Janssens TVW, Clausen BS, Falsig H, Christensen CH, Nørskov JK (2007) Catalytic activity of Au nanoparticles. Nano Today 2:14
38. Jacobsen CJH (2000) Novel class of ammonia synthesis catalysts. Chem Commun 12:1057
39. Jacobsen CJH, Dahl S, Clausen BS, Bahn S, Logadottir A, Nørskov JK (2001) Catalyst design by interpolation in the periodic table: Bimetallic ammonia synthesis catalyst. J Am Chem Soc 123:8404
40. Jankowiak JT, Barteau AM (2005) Ethylene epoxidation over silver and copper-silver bimetallic catalysts: II. Cs and Cl promotion. J Catal 236:379
41. Jankowiak JT, Barteau MA (2005) Ethylene epoxidation over silver and copper-silver bimetallic catalysts: I. kinetics and selectivity. J Catal 236:366
42. Jones GS, Mavrikakis M, Mark Barteau A, John Vohs M (1998) First synthesis, experimental and theoretical vibrational spectra of an oxametallacycle on a metal surface. J Am Chem Soc 120:2196–3204
43. Klust A, Madix RJ (2006) Partial oxidation of higher olefins on Ag(111): Conversion of styrene to styrene oxide, benzene, and benzoic acid. Surf Sci 600:2025
44. Knudsen J, Nilekar AU, Vang RT, Schnadt J, Kunkes EL, Dumesic JA, Mavrikakis M, Besenbacher F (2007) A Cu/Pt near-surface alloy for water-gas shift catalysis. J Am Chem Soc 129:6485–6490
45. Kochloefl K (1997) Steam reforming. In: Ertl G, Knözinger H, Weitkamp J (eds) Handbook of Heterogeneous Catalysis, vol 4, chapter 3.2. Wiley, Weinheim, p 1819–1831
46. Kohn W, Sham LJ (1965) Self-consistent equations including exchange and correlation effects. Phys Rev 140(4A):A1133–A1138
47. Kojima R, Aika K-I (2000) Cobalt molybdenum bimetallic nitride catalysts for ammonia synthesis. Chem Lett 29:514
48. Kokalj A, Bonini N, Sbraccia C, de Gironcoli S, Baroni S (2004) Engineering the reactivity of metal catalysts: a model study of methane dehydrogenation on Rh(111). J Am Chem Soc 126:16732
49. Kokalj A, Bonini N, de Gironcoli S, Sbraccia C, Fratesi G, Baroni S (2006) Methane dehydrogenation on Rh@Cu(111): A first-principles study of a model catalyst. J Am Chem Soc 128:12448
50. Kokalj A, Gava P, de Gironcoli S, Baroni S (2008) Activated adsorption of ethylene on atomic-oxygen covered Ag(100) and Ag(210): formation of an oxametallacycle. J Phys Chem C 112:1019–1027. doi:10.1021/jp0747961
51. Kokalj A, Gava P, de Gironcoli S, Baroni S (2008) What determines the catalyst's selectivity in the ethylene epoxidation reaction. J Catal 254:304–309
52. Kratzer P, Hammer B, Nørskov JK (1996) A theoretical study of CH_4 dissociation on pure and gold-alloyed Ni(111) surface. J Chem Phys 105:5595
53. Lambert RM, Williams FJ, Cropley RL, Palermo A (2005) Heterogeneous alkene epoxidation: past, present and future. J Mol Catal A 228:27
54. Li W-X, Stampfl C, Scheffler M (2002) Oxygen adsorption on Ag(111): A density-functional investigation. Phys Rev B 65:075407
55. Linic S, Barteau MA (2002) Formation of a stable surface oxametallacycle that produces ethylene oxide. J Am Chem Soc 124:310
56. Linic S, Barteau MA (2003) Construction of a reaction coordinate and a microkinetic model for ethylene epoxidation on silver from DFT calculations and surface science experiments. J Catal 214:200–212
57. Linic S, Barteau MA (2003) Control of ethylene epoxidation selectivity by surface oxametallacycles. J Am Chem Soc 125:4034
58. Linic S, Jankowiak J, Barteau MA (2004) Selectivity driven design of bimetallic ethylene epoxidation catalysts from first principles. J Catal 224:489–493
59. Linic S, Piao H, Adib K, Barteau MA (2004) Ethylene epoxidation on Ag: Identification of the crucial surface intermediate by experimental and theoretical investigation of its electronic structure. Angew Chem Int Ed 43:2918
60. Liu Z-P, Gong X-Q, Kohanoff J, Sanchez C, Hu P (2003) Catalytic role of metal oxides in gold-based catalysts: A first principles study of CO oxidation on TiO_2 supported Au. Phys Rev Lett 91(26):266102
61. Lynggaard H, Andreasen A, Stegelmann C, Stoltze P (2004) Analysis of simple kinetic models in heterogeneous catalysis. Prog Surf Sci 77:71–137
62. Masel RI (1996) Principles of adsorption and reaction on solid surfaces. Wiley, New York
63. Massalski TB, Okamoto H, Subramanian PR, Kacprezak L (1990) Binary Alloy Phase Diagrams, vol 2E. ASM, Materials Parks, OH
64. Mavrikakis M, Doren DJ, Barteau MA (1998) Density functional theory calculations for simple oxametallacycles: Trends across the periodic table. J Phys Chem B 102:394–399
65. Mavrikakis M, Hammer B, Nørskov JK (1998) Effect of strain on the reactivity of metal surfaces. Phys Rev Lett 81(13):2819–2822 Sep
66. Mavrikakis M, Stoltze P, Nørskov JK (2000) Making gold less noble. Catal Lett 64:101
67. Medlin JW, Barteau MA, Vohs JM (2000) Oxametallacycle formation via ring-opening of 1-epoxy-3-butene on Ag(110): a combined experimental/theoretical approach. J Mol Catal A 163:129
68. Michaelides A, Hu P (1999) Methyl chemisorption on Ni(111) and C–H–M multicentre bonding: a density functional theory study. Surf Sci 437:362

69. Michaelides A, Hu P (2001) Softened C–H modes of adsorbed methyl and their implications for dehydrogenation: An ab initio study. J Chem Phys 114:2523
70. Michaelides A, Hu P (2000) A first principles study of CH_3 dehydrogenation on Ni(111). J Chem Phys 112:8120
71. Michaelides A, Liu Z-P, Zhang CJ, Alavi A, King DA, Hu P (2003) Identification of general linear relationships between activation energies and enthalpy changes for dissociation reactions at surfaces. J Am Chem Soc 125:3704
72. Molina LM, Hammer B (2003) Active role of oxide support during CO oxidation at Au/MgO. Phys Rev Lett 90(20):206102
73. Molina LM, Hammer B (2004) Theoretical study of CO oxidation on Au nanoparticles supported by MgO(100). Physical Review B (Condensed Matter and Materials Physics) 69(15):155424
74. Molina LM, Rasmussen MD, Hammer B (2004) Adsorption of O_2 and oxidation of CO at Au nanoparticles supported by TiO_2(110). J Chem Phys 120:7673
75. Newns DM (1969) Self-consistent model of hydrogen chemisorption. Phys Rev 178(3):1123–1135
76. Nørskov JK, Bligaard T, Logadottir A, Bahn S, Hansen LB, Bollinger M, Bengaard H, Hammer B, Sljivancanin Z, Mavrikakis M, Xu Y, Dahl S, Jacobsen CJH (2002) Universality in heterogeneous catalysis. J Catal 209:275
77. Pallassana V, Neurock M (2000) Electronic factors governing ethylene hydrogenation and dehydrogenation activity of pseudomorphic Pd_{ML}/Re(0001), Pd_{ML}/Ru(0001), Pd(111) and Pd_{ML}/Au(111) surfaces. J Catal 191:301–317
78. Papoian G, Nørskov JK, Hoffmann R (2000) A comparative theoretical study of hydrogen, methyl, and ethyl chemisorption on the Pt(111) surface. J Am Chem Soc 122:4129
79. Parr RG, Yang W (1989) Density-functional theory of atoms and molecules. Oxford University Press, New York, USA
80. Perdew JP, Burke K, Ernzerhof M (1996) Generalized gradient approximation made simple. Phys Rev Lett 77(18):3865–3868
81. Perdew JP, Wang Y (1992) Accurate and simple analytic representation of the electron-gas correlation energy. Phys Rev B 45:13244–13249
82. Petersen MA, Jenkins SJ, King DA (2004) Theory of methane dehydrogenation on Pt{110}(1 × 2). Part II: Microscopic reaction pathways for $CH_x \rightarrow CH_{x-1}$ (x = 1–3). J Phys Chem B 108:5920
83. Ramsier RD, Gao Q, Waltenburg NH, JT Yates Jr. (1994) Thermal dissociation of NO on Pd surfaces: The influence of step sites. J Chem Phys 100:6837–6845
84. Remediaki IN, Lopez N, Nørskov JK (2005) CO oxidation on rutile-supported Au nanoparticles. Angew Chem Int Ed 44:1824
85. Reuter K, Scheffler M (2006) First-principles kinetic Monte Carlo simulations for heterogeneous catalysis: Application to the CO oxidation at RuO_2(110). Phys Rev B 73(4):045433
86. Ricart JM, Clotet A, Illas F, Rubio J (1994) The analysis of the chemisorption bond from uncorrelated and correlated cluster model wave functions. J Chem Phys 100:1988
87. Rodriguez JA, Goodman DW (1992) The nature of metal-metal bond in bimetallic surfaces. Science 257:897
88. Ruban A, Hammer B, Stoltze P, Skriver HL, Nørskov JK (1997) Surface electronic structure and reactivity of transition and noble metals. J Mol Catal A 115:421
89. Ruban AV, Skriver HL, Nørskov JK (1999) Surface segregation energies in transition-metal alloys. Phys Rev B 59(24):15990–16000
90. van Santen RA (1997) Epoxidation catalysis using heterogeneous catalysts. In: Ertl G, Knözinger H, Weitkamp J (eds) Handbook of Heterogeneous Catalysis, vol 5, chap 4. Wiley, Weinheim, pp 2244–2252
91. van Santen RA, de Groot CPM (1986) The mechanism of ethylene epoxidation. J Catal 98:530–539
92. van Santen RA (1997) Epoxidation catalysis using heterogeneous catalysts In: Handbook of Heterogeneous Catalysis, vol 5, chap. 4.6.1. Wiley, Weinheim, pp 2244–2252
93. Savio L, Vattuone L, Rocca M (2001) The role of steps and terrace width in gas-surface interaction: O_2/Ag(410). Phys Rev Lett 87:276101
94. Schlögl R (1997) Ammonia synthesis. In: Ertl G, Knözinger H, Weitkamp J (eds) Handbook of Heterogeneous Catalysis, vol 4, chap 2.1. Wiley, Weinheim, pp 1697–1748
95. Sellidj A, Koel BE (1994) Electronic and CO chemisorption properties of ultrathin Pd films vapor deposited on Au(111). Phys Rev B 49(12):8367–8376 Mar
96. Serafin JG, Liu AC, Seyedmonir SR (1998) Surface science and the silver-catalyzed epoxidation of ethylene: an industrial perspective. J Mol Catal A 131:157
97. Sinfelt JH (1983) Bimetallic Catalysts. Willey, New York
98. Somorjai GA (1981) Chemistry in two dimensions: Surfaces. Cornell University Press, Ithaca
99. Somorjai GA (1994) Introduction to surface chemistry and catalysis. Willey, Inc., New York
100. Stacchiola D, Wu G, Kaltchev M, Tysoe WT (2000) Spectroscopic studies of ethylene adsorption on oxygen-modified Ag(111) at high pressures. Surf Sci 463:81
101. Stegelmann C, Schiødt NC, Campbell CT, Stoltze P (2004) Microkinetic modeling of ethylene oxidation over silver. J Catal 221:630
102. Stoltze P (2000) Microkinetic simulation of catalytic reactions. Prog Surf Sci 65:65–150
103. Svensson M, Blomberg MRA, Siegbahn PEM (1991) Reaction of second-row transition-metal atoms with methane. J Am Chem Soc 113:7076
104. Szabo A, Ostlund NS (1996) Modern quantum chemistry: introduction to advanced electronic structure theory. Dover publications, Inc., Mineola
105. Torres D, Lopez N, Illas F (2006) A theoretical study of coverage effects for ethylene epoxidation on Cu(111) under low oxygen pressure. J Catal 243:404
106. Torres D, Lopez N, Illas F, Lambert RM (2005) Why copper is intrinsically more selective than silver in alkene epoxidation: Ethylene oxidation on Cu(111) versus Ag(111). J Am Chem Soc 127:10774
107. Valden M, Lai X, Goodman DW (1998) Onset of catalytic activity of gold clusters on titania with the appearance of nonmetallic properties. Science 281:1647
108. Watwe RM, Bengaard HS, Rostrup-Nielsen JR, Dumesic JA, Nørskov JK (2000) Theoretical studies of stability and reactivity of CH_x species on Ni(111). J Catal 189:16
109. Whitten JL, Yang H (1996) Theory of chemisorption and reactions on metal surfaces. Surf Sci Rep 24:55
110. Wu Y, Schmitt JD, Car R (2004) Mapping potential energy surfaces. J Chem Phys 121:1193–1200
111. Zaera F (1995) An organometallic guide to the chemistry of hydrocarbons moieties on transition metal surfaces. Chem Rev 95:2651

112. Zambelli T, Wintterlin J, Trost J, Ertl G (1996) Identification of the active sites of a surface-catalyzed reaction. Science 273:1688–1690
113. Zhang CJ, Hu P (2002) Methane transformation to carbon and hydrogen on Pd(100): Pathways and energetics from density functional theory calculations. J Chem Phys 116:322
114. Zhang CJ, Hu P (2002) The possibility of single C–H bond activation in CH_4 on MoO_3-supported Pt catalyst: A density functional theory study. J Chem Phys 116:4281
115. Zhdanov VP (2002) Monte Carlo simulations of oscillations, chaos and pattern formation in heterogeneous catalytic reactions. Surf Sci Rep 45:231–326
116. Zhdanov VP, Kasemo B (1994) Kinetic phase transitions in simple reactions on solid surfaces. Surf Sci Rep 20:113–189

Computer Graphics and Games, Agent Based Modeling in

Brian Mac Namee
School of Computing, Dublin Institute of Technology, Dublin, Ireland

Article Outline

Glossary
Definition of the Subject
Introduction
Agent-Based Modelling in Computer Graphics
Agent-Based Modelling in CGI for Movies
Agent-Based Modelling in Games
Future Directions
Bibliography

Glossary

Computer generated imagery (CGI) The use of computer generated images for special effects purposes in film production.

Intelligent agent A hardware or (more usually) software-based computer system that enjoys the properties autonomy, social ability, reactivity and pro-activeness.

Non-player character (NPC) A computer controlled character in a computer game – as opposed to a player controlled character.

Virtual character A computer generated character that populates a virtual world.

Virtual world A computer generated world in which places, objects and people are represented as graphical (typically three dimensional) models.

Definition of the Subject

As the graphics technology used to create virtual worlds has improved in recent years, more and more importance has been placed on the behavior of virtual characters in applications such as games, movies and simulations set in these virtual worlds simulations. The behavior of these virtual characters should be believable in order to create the illusion that virtual worlds are populated with living characters. This has led to the application of agent-based modeling to the control of virtual characters. There are a number of advantages of using agent-based modeling techniques which include the fact that they remove the requirement for hand controlling all agents in a virtual environment, and allow agents in games to respond to unexpected actions by players or users.

Introduction

Advances in computer graphics technology in recent years have allowed the creation of realistic and believable virtual worlds. However, as such virtual worlds have been developed for applications spanning games, education and movies it has become apparent that in order to achieve real believability, virtual worlds must be populated with life-like virtual characters. This is where the application of agent-based modeling has found a niche in the areas of computer graphics and, in a huge way, computer games. Agent-based modeling is a perfect solution to the problem of controlling the behaviors of the virtual characters that populate a virtual world. In fact, because virtual characters are embodied and autonomous these applications require an even stronger notion of agency than many other areas in which agent-based modeling is employed.

Before proceeding any further, and because there are so many competing alternatives, it is worth explicitly stating the definition of an intelligent agent that will inform the remainder of this article. Taken from [83] an intelligent agent is defined as "… *a hardware or (more usually) software-based computer system that enjoys the following properties:*

- ***autonomy:*** *agents operate without the direct intervention of humans or others, and have some kind of control over their actions and internal state;*
- ***social ability:*** *agents interact with other agents (and possibly humans) via some kind of agent-communication language;*
- ***reactivity:*** *agents perceive their environment, (which may be the physical world, a user via a graphical user interface, a collection of other agents, the INTERNET, or*

perhaps all of these combined), and respond in a timely fashion to changes that occur in it;

- ***pro-activeness:*** *agents do not simply act in response to their environment, they are able to exhibit goal-directed behavior by taking the initiative."*

Virtual characters implemented using agent-based modeling techniques satisfy all of these properties. The characters that populate virtual worlds should be fully autonomous and drive their own behaviors (albeit sometimes following the orders of a director or player). Virtual characters should be able to interact believably with other characters and human participants. This property is particularly strong in the case of virtual characters used in games which by their nature are particularly interactive. It is also imperative that virtual characters appear to perceive their environments and react to events that occur in that environment, especially the actions of other characters or human participants. Finally virtual characters should be pro-active in their behaviors and not always require prompting from a human participant in order to take action.

The remainder of this article will proceed as follows. Firstly, a broad overview of the use of agent-based modeling in computer graphics will be given, focusing in particular on the genesis of the field. Following on from this, the focus will switch to the use of agent-based modeling techniques in two particular application areas: computer generated imagery (CGI) for movies, and computer games. CGI has been used to astounding effect in movies for decades, and in recent times has become heavily reliant on agent-based modeling techniques in order to generate CGI scenes containing large numbers of computer generated extras. Computer games developers have also been using agent-based modeling techniques effectively for some time now for the control of non-player characters (NPCs) in games. There is a particularly fine match between the requirements of computer games and agent-based modeling due to the high levels of interactivity required.

Finally, the article will conclude with some suggestions for the future directions in which agent-based modeling technology in computer graphics and games is expected to move.

Agent-Based Modelling in Computer Graphics

The serious use of agent-based modeling in computer graphics first arose in the creation of autonomous groups and crowds – for example, crowds of people in a town square or hotel foyer, or flocks of birds in an outdoor scene. While initially this work was driven by visually unappealing simulation applications such as fire safety testing for buildings [75], focus soon turned to the creation of visually realistic and believable crowds for applications such as movies, games and architectural walkthroughs. Computer graphics researchers realized that creating scenes featuring large virtual crowds by hand (a task that was becoming important for the applications already mentioned) was laborious and time-consuming and that agent-based modeling techniques could remove some of the animator's burden. Rather than requiring that animators hand-craft all of the movements of a crowd, agent-based systems could be created in which each character in a crowd (or flock, or swarm) could drive its own behavior. In this way the behavior of a crowd would emerge from the individual actions of the members of that crowd.

Two of the earliest, and seminal, examples of such systems are Craig Reynolds' Boids system [64] and Tu & Terzopoulos' animations of virtual fish [76]. The Boids system simulates the flocking behaviors exhibited in nature by schools of fish, or flocks of birds. The system was first presented at the prestigious SIGGRAPH conference (www.siggraph.org) in 1987 and was accompanied by the short movie "Stanley and Stella in: Breaking the Ice". Taking influence from the area of artificial life (or aLife) [52], Reynolds postulated that the individual members of a flock would not be capable of complex reasoning, and so flocking behavior must emerge from simple decisions made by individual flock members. This notion of emergent behavior is one of the key characteristics of aLife systems.

In the original Boids system, each virtual agent (represented as a simple particle and known as a *boid*) used just three rules to control its movement. These were separation, alignment and cohesion and are illustrated in Fig. 1. Based on just these three simple rules extremely realistic flocking behaviors emerged. This freed animators from the laborious task of hand-scripting the behavior of each creature within the flock and perfectly demonstrates the advantage offered by agent-based modeling techniques for this kind of application.

The system created by Tu and Terzopoulos took a more complex approach in that they created complex models of biological fish. Their models took into account fish physiology, with a complex model of fish muscular structure, along with a perceptual model of fish vision. Using these they created sophisticated simulations in which properties such as schooling and predator avoidance were displayed. The advantage of this approach was that it was possible to create unique, unscripted, realistic simulations without the intervention of human animators. Terzopoulos has since gone on to apply similar techniques to the control of virtual humans [68].

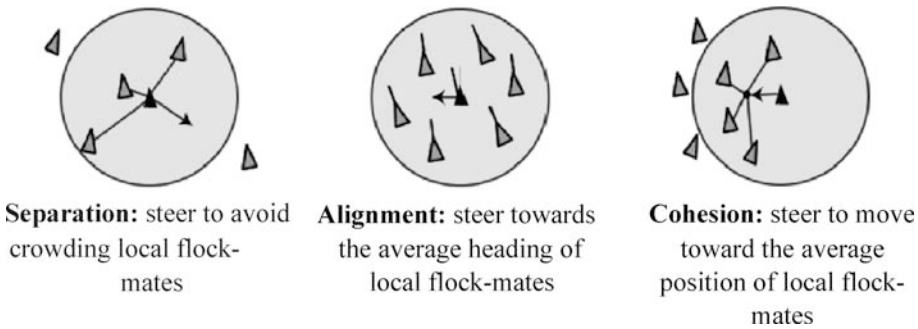

Computer Graphics and Games, Agent Based Modeling in, Figure 1
The three rules used by Reynolds' original Boids system to simulate flocking behaviors

Moving from animals to crowds of virtual humans, the Virtual Reality Lab at the Ecole Polytechnique Fédérale de Lausanne in Switzerland (vrlab.epfl.ch) led by Daniel Thalmann has been at the forefront of this work for many years. They group currently has a highly evolved system, VICrowd, for the animation of virtual crowds [62] which they model as a hierarchy which moves from individuals to groups to crowds. This hierarchy is used to avoid some of the complications which arise from trying to model large crowds in real time – one of the key gaols of ViCrowd.

Each of the levels in the ViCrowd hierarchy can be modeled as an agent and this is done based on *beliefs*, *desires* and *intentions*. The beliefs of an agent represent the information that the agent possesses about the world, including information about places, objects and other agents. An agent's desires represent the motivations of the agent regarding objectives it would like to achieve. Finally, the intentions of an agent represent the actions that an agent has chosen to pursue. The belief-desire-intention (BDI) model of agency was proposed by Rao and Georgeff [61] and has been used in many other application areas of agent-based modeling.

ViCrowd has been used in ambitious applications including the simulation of a virtual city comprised of, amongst other things, a train station a park and a theater [22]. In all of these environments the system was capable of driving the believable behaviors of large groups of characters in real-time.

It should be apparent to readers from the examples given thus far that the use of agent-based modeling techniques to control virtual characters gives rise to a range of unique requirements when compared to the use of agent based modeling in other application areas. The key to understanding these is to realize that the goal in designing agents for the control of virtual characters is typically not to design the most efficient or effective agent, but rather to design the most interesting or believable character. Outside of very practical applications such as evacuation simulations, when creating virtual characters, designers are concerned with maintaining what Disney, experts in this field, refer to as the *illusion of life* [36].

This refers to the fact that the user of a system must believe that virtual characters are living, breathing creatures with goals, beliefs, desires, and, essentially, lives of their own. Thus, it is not so important for a virtual human to always choose the most efficient or cost effective option available to it, but rather to always choose reasonable actions and respond realistically to the success or failure of these actions. With this in mind, and following a similar discussion given in [32], some of the foremost researchers in virtual character research have the following to say about the requirements of agents as virtual characters.

Loyall writes [46] that "*Believable agents are personality-rich autonomous agents with the powerful properties of characters from the arts.*" Coming from a dramatic background it is not surprising that Loyall's requirements reflect this. Agents should have strong personality and be capable of showing emotion and engaging in meaningful social relationships.

According to Blumberg [11], "… *an autonomous animated creature is an animated object capable of goal-directed and time-varying behavior*". The work of Blumberg and his group is very much concerned with virtual creatures, rather than humans in particular, and his requirements reflect this. Creatures must appear to make choices which improve their situation and display sophisticated and individualistic movements.

Hayes–Roth and Doyle focus on the differences between "*animate characters*" and traditional agents [27]. With this in mind they indicate that agents' behaviors must be "*variable rather than reliable*", "*idiosyncratic instead of predictable*", "*appropriate rather than correct*", "*effective instead of complete*", "*interesting rather than effi-*

cient", and "*distinctively individual as opposed to optimal*".

Perlin and Goldberg [59] concern themselves with building believable characters "*that respond to users and to each other in real-time, with consistent personalities, properly changing moods and without mechanical repetition, while always maintaining an author's goals and intentions*".

Finally, in characterizing believable agents, Bates [7] is quite forgiving requiring "*only that they not be clearly stupid or unreal*". Such broad, shallow agents must "*exhibit some signs of internal goals, reactivity, emotion, natural language ability, and knowledge of agents … as well as of the … micro-world*".

Considering these definitions, Isbister and Doyle [32] identify the fact that the consistent themes which run through all of the requirements given above match the general goals of agency – virtual humans must display autonomy, reactivity, goal driven behavior and social ability – and again support the use of agent-based modeling to drive the behavior of virtual characters.

The Spectrum of Agents

The differences between the systems mentioned in the previous discussion are captured particularly well on the *spectrum of agents* presented by Aylett and Luck [5]. This positions agent systems on a spectrum based on their capabilities, and serves as a useful tool in differentiating between the various systems available. One end of this spectrum focuses on *physical agents* which are mainly concerned with simulation of believable physical behavior, (including sophisticated physiological models of muscle and skeleton systems), and of sensory systems. Interesting work at this end of the spectrum includes Terzopoulos' highly realistic simulation of fish [76] and his virtual stuntman project [21] which creates virtual actors capable of realistically synthesizing a broad repertoire of lifelike motor skills.

Cognitive agents inhabit the other end of the agent spectrum and are mainly concerned with issues such as reasoning, decision making, planning and learning. Systems at this end of the spectrum include Funge's cognitive modeling approach [26] which uses the situation calculus to control the behavior of virtual characters, and Nareyek's work on planning agents for simulation [55], both of which will be described later in this article.

While the systems mentioned so far sit comfortably at either end of the agent spectrum, many of the most effective inhabit the middle ground. Amongst these are *c4* [13], used to great effect to simulate a virtual sheep dog with the ability to learn new behaviors, *Improv* [59] which augments sophisticated physical human animation with scripted behaviors and the *ViCrowd* system [62] which sits on top of a realistic virtual human animation system and uses planning to control agents' behavior.

Virtual Fidelity

The fact that so many different agent-based modeling systems, for the control of virtual humans exist gives rise to the question why? The answer to this lies in the notion of *virtual fidelity*, as described by Badler [6]. Virtual fidelity refers to the fact that virtual reality systems need only remain true to actual reality in so much as this is required by, and improves, the system.

In [47] the point is illustrated extremely effectively. The article explains that when game designers are architecting the environments in which games are set, the scale to which these environments are created is not kept true to reality. Rather, to ease players' movement in these worlds, areas are designed to a much larger scale, compared to character sizes, than in the real world. However, game players do not notice this digression from reality, and in fact have a negative response to environments that are designed to be more true to life finding them cramped. This is a perfect example of how, although designers stay true to reality for many aspects of environment design, the particular blend of virtual fidelity required by an application can dictate certain real world restrictions can be ignored in virtual worlds.

With regard to virtual characters, virtual fidelity dictates that the set of capabilities which these characters should display is determined by the application which they are to inhabit. So, the requirements of an agent-based modeling system for CGI in movies would be very different to those of a agent-based modeling system for controlling the behaviors of game characters.

Agent-Based Modelling in CGI for Movies

With the success of agent-based modeling techniques in graphics firmly established there was something of a search for application areas to which they could be applied. Fortunately, the success of agent-based modeling techniques in computer graphics was paralleled with an increase in the use of CGI in the movie industry, which offered the perfect opportunity. In many cases CGI techniques were being used to replace traditional methods for creating expensive, or difficult to film scenes. In particular, scenes involving large numbers of people or animals were deemed no longer financially viable when set in the real world. Creating these scenes using CGI involved painstak-

ing hand animation of each character within a scene, which again was not financially viable.

The solution that agent-based modeling offers is to make each character within a scene an intelligent agent that drives its own behavior. In this way, as long as the initial situation is set up correctly scenes will play out without the intervention of animators. The facts that animating for movies does not need to be performed in real-time, and is in no way interactive (there are no human users involved in the scene), make the use of agent-based modeling a particularly fine match for this application area.

Craig Reynolds' Boids system [64] which simulates the flocking behaviors exhibited in nature by schools of fish, or flocks of birds and was discussed previously is one of the seminal examples of agent-based modeling techniques being used in movie CGI. Reynold's approach was first used for CGI in the 1999 film "Batman Returns" [14] to simulate colonies of bats. Reynold's technologies have been used in "The Lion King" [4] and "From Dusk 'Till Dawn" [65] amongst other films. Reynolds' approach was so successful, in fact, that he was awarded an Academy Award for his work in 1998.

Similar techniques to those utilized in the Boids system have been used in many other films to animate such diverse characters as ants, people and stampeding wildebeest. Two productions which were released in the same year, "Antz" [17] by Dreamworks and "A Bug's Life" [44] by Pixar took great steps in using CGI effects to animate large crowds for. For "Antz" systems were developed which allowed animators easily create scenes containing large numbers of virtual characters modeling each as an intelligent agent capable of obstacle avoidance, flocking and other behaviors. Similarly, the creators of "A Bug's Life" created tools which allowed animators easily combine pre-defined motions (known as alibis) to create behaviors which could easily be applied to individual agents in scenes composed of hundreds of virtual characters.

However, the largest jump in the use of agent-based modeling in movie CGI was made in the recent Lord of the Rings trilogy [33,34,35]. In these films the bar was raised markedly in terms of the sophistication of the virtual characters displayed and the sheer number of characters populating each scene. To achieve the special effects shots required by the makers of these films, the Massive software system was developed by Massive Software (www.massivesoftware.com). This system [2,39] uses agent-based modeling techniques, again inspired by aLife, to create virtual extras that control their own behaviors. This system was put to particularly good use in the large scale battle sequences that feature in all three of the Lord of the Rings films. Some of the sequences in the final film of the trilogy, the Return of the King, contain over 200,000 digital characters.

In order to create a large battle scene using the Massive software, each virtual extra is represented as an intelligent agent, making its own decisions about which actions it will perform based on its perceptions of the world around it. Agent control is achieved through the use of fuzzy logic based controllers in which the state of an agent's brain is represented as a series of motivations, and knowledge it has about the world – such as the state of the terrain it finds itself on, what kinds of other agents are around it and what these other agents are doing. This knowledge about the world is perceived through simple simulated visual, auditory and tactile senses. Based on the information they perceive agents decide on a best course of action. Designing the brains of these agents is made easier that it might seem at first by the fact that agents are developed for short sequences, and so a small range of possible tasks. So for example, separate agent models would be used for a fighting scene and a celebration scene.

In order to create a large crowd scene using Massive animators initially set up an environment populating it with an appropriate cast of virtual characters where the brains of each character are slight variations (based on physical and personality attributes) of a small number of archetypes. The scene will then play itself out with each character making it's own decisions. Therefore there is no need for any hand animation of virtual characters. However, directors can view the created scenes and by tweaking the parameters of the brains of the virtual characters have a scene play out in the exact way that they require.

Since being used to such impressive effect in the Lord of the Rings trilogy (the developers of the Massive system were awarded an academy award for their work), the Massive software system has been used in numerous other films such as "I, Robot" [60], "The Chronicles of Narnia: The Lion, the Witch and the Wardrobe" [1] and "Ratatouille" [10] along with numerous television commercials and music videos.

While the achievements of using agent-based modeling for movie CGI are extremely impressive, it is worth noting that none of these systems run in real-time. Rather, scenes are rendered by banks of high powered computers, a process that can take hours for relatively simple scenes. For example, the famous Prologue battle sequence in the "Lord of the Rings: The Fellowship of the Ring" took a week to render. When agent-based modeling is applied to the real-time world of computer games, things are very different.

Agent-Based Modelling in Games

Even more so than in movies, agent-based modeling techniques have been used to drive the behaviors of virtual characters in computer games. As games have become graphically more realistic (and in recent years they have become extremely so) game-players have come to expect that games are set in hugely realistic and believable virtual worlds. This is particularly evident in the widespread use of realistic physics modeling which is now commonplace in games [67]. In games that make strong use of physics modeling, objects in the game world topple over when pushed, float realistically when dropped in water and generally respond as one would expect them to. Players expect the same to be true of the virtual characters that populate virtual game worlds. This can be best achieved by modeling virtual characters as embodied virtual agents. However, there are a number of constraints which have a major influence on the use of agent-based modeling techniques in games.

The first of these constraints stems from the fact that modern games are so highly interactive. Players expect to be able to interact with all of the characters they encounter within a game world. These interactions can be as simple as having something to shoot at or having someone to race against; or involve much more sophisticated interactions in which a player is expected to converse with a virtual character to find out specific information or to cooperate with a virtual character in order to accomplish some task that is key to the plot of a game. Interactivity raises a massive challenge for practitioners as there is very little restriction in terms of what the player might do. Virtual characters should respond in a believable way at all times regardless of how bizarre and unexpected the actions of the player might be.

The second challenge comes from the fact that the vast majority of video games should run in real time. This means that the computational complexity must be kept to a reasonable level as there are only a finite number of processor cycles available for AI processing. This problem is magnified by the fact that an enormous amount of CPU power it usually dedicated to graphics processing. When compared to the techniques that can be used for controlling virtual characters in films some of the techniques used in games are rudimentary due to this real-time constraint.

Finally, modern games resemble films in the fact that creators go to great lengths to include intricate storylines and control the building of tension in much the way that film script writers do. This means that games are tested heavily in order to ensure that the game proceeds smoothly and that the level of difficulty is finely tuned so as to always hold the interest of a player. In fact, this testing of games has become something of a science in itself [77]. Using autonomous agents gives game characters the ability to do things that are unexpected by the game designers and so upset their well laid plans. This can often be a barrier to the use of sophisticated techniques such as learning.

Unfortunately there is also a barrier to the discussion of agent-based modeling techniques used in commercial games. Because of the very competitive nature of the games industry, game development houses often consider the details of how their games work as valuable trade secrets to be kept well guarded. This can make it difficult to uncover the details of how particularly interesting features of a game are implemented. While this situation is improving – more commercial game developers are speaking at games conferences about how their games are developed and the release of game systems development kits for the development of game modifications (or *mods*) allows researchers to plumb the depths of game code – it is still often impossible to find out the implementation details of very new games.

Game Genres

Before discussing the use of agent-based modeling in games any further, it is worth making a short clarification on the kinds of computer games that this article refers to. When discussing modern computer games, or video games, this article does not refer to computer implementations of traditional games such as chess, backgammon or card games such as solitaire. Although these games are of considerable research interest (chess in particular has been the subject of extremely successful research [23]) they are typically not approached using agent-based modeling techniques. Typically, artificial intelligence approaches to games such as these rely largely on sophisticated searching techniques which allow the computer player to search through a multitude of possible future situations dictated by the moves it will make and the moves it expects its opponent to make in response. Based on this search, and some clever heuristics that indicate what constitutes a good game position for the computer player, the best sequence of moves can be chosen. This searching technique relies on the fact that there are usually a relatively small number of moves that a player can make at any one time in a game. However, the fact that the ancient Chinese game of Go-Moku has not, to date, been mastered by computer players [80] illustrates the restrictions of such techniques.

The common thread linking together the kinds of games that this article focuses on is that they all contain computer controlled virtual characters that possess

Computer Graphics and Games, Agent Based Modeling in, Figure 2
A screenshot of the upcoming action game Rogue Warrior from Bethesda Softworks (image courtesy of Bethesda Softworks)

a strong notion of agency. Efforts are often made to separate the many different kinds of modern video games that are the focus of this article into a small set of descriptive genres. Unfortunately, much like in music, film and literature, no categorization can hope to perfectly capture the nuances of all of the available titles. However, a brief mention of some of the more important game genres is worth while (a more detailed description of game genres, and artificial intelligence requirements of each is given in [41]).

The most popular game genre is without doubt the *action game* in which the player must defeat waves of demented foes, typically (for increasingly bizarre motivations) bent upon global destruction. Illustrative examples of the genre include *Half-Life 2* (www.half-life2.com) and the *Halo* series (www.halo3.com). A screenshot of the upcoming action game *Rogue Warrior* (www.bethsoft.com) is shown in Fig. 2.

Strategy games allow players to control large armies in battle with other people, or computer controlled opponents. Players do not have direct control over their armies, but rather issue orders which are carried out by agent-based artificial soldiers. Well regarded examples of the genre include the *Age of Empires* (www.ageofempires.com) and *Command & Conquer* (www.commandandconquer.com) series.

Role playing games (such as the *Elder Scrolls* (www.elderscrolls.com) series) place game players in expansive virtual worlds across which they must embark on fantastical quests which typically involve a mixture of solving puzzles, fighting opponents and interacting with non-player characters in order to gain information. Figure 3 shows a screenshot of the aforementioned role-playing game *The Elder Scrolls IV: Oblivion*.

Almost every sport imaginable has at this stage been turned into a computer based *sports game*. The challenges in developing these games are creating computer controlled opponents and team mates that play the games at a level suitable to the human player. Interesting examples include *FIFA Soccer 08* (www.fifa08.ea.com) and *Forza Motorsport 2* (www.forzamotorsport.net).

Finally, many people expected that the rise of massively multi-player online games (MMOGs), in which hundreds of human players can play together in an online world, would sound the death knell for the use of virtual non-player characters in games. Examples of MMOGs include *World of Warcraft* (www.worldofwarcraft.com) and *Battlefield 2142* (www.battlefield.ea.com). However, this has not turned out to be the case as there are still large numbers of single player games being produced and even MMOGs need computer controlled characters for roles that players do not wish to play.

Of course there are many games that simply do not fit into any of these categorizations, but that are still relevant for a discussion of the use of agent-based techniques – for example *The Sims* (www.thesims.ea.com) and the *Microsoft Flight Simulator* series (www.microsoft.com/games/flightsimulatorx). However the categorization still serves to introduce those unfamiliar with the subject to the kinds of games up for discussion.

Computer Graphics and Games, Agent Based Modeling in, Figure 3
A screenshot from Bethesda Softwork's role playing game The Elder Scrolls IV: Oblivion (image courtesy of Bethesda Softworks)

Implementing Agent-Based Modelling Techniques in Games

One of the earliest examples of using agent-based modeling techniques in video games was its application to path planning. The ability of non-player characters (NPCs) to manoeuvre around a game world is one of the most basic competencies required in games. While in very early games it was sufficient to have NPCs move along prescripted paths, this soon become unacceptable. Games programmers soon began to turn to AI techniques which might be applied to solve some of the problems that were arising. The A* path planning algorithm [74] was the first example of such a technique to find wide-spread use in the games industry. Using the A* algorithm NPCs can be given the ability to find their own way around an environment. This was put to particularly fine effect early on in real-time strategy games where the units controlled by players are semi-autonomous and are given orders rather than directly controlled. In order to use the A* algorithm a game world must be divided into a series of cells each of which is given a rating in terms of the effort that must be expended to cross it. The A* algorithm then performs a search across these cells in order to find the shortest path that will take a game agent from a start position to a goal.

Since becoming widely understood amongst the game development community many interesting additions have been made to the basic A* algorithm. It was not long before three dimensional versions of the algorithm became commonly used [71]. The basic notion of storing the energy required to cross a cell within a game world has also been extended to augment cells with a wide range of other useful information (such as the level of danger in crossing a cell) that can be used in the search process [63].

The next advance in the kind of techniques being used to achieve agent-based modeling in games was the finite state machine (FSM) [30]. An FSM is a simple system in which a finite number of *states* are connected in a directed graph by *transitions* between these states. When used for the control of NPCs, the nodes of an FSM indicate the possible actions within a game world that an agent can perform. Transitions indicate how changes in the state of the game world or the character's own attributes (such as health, tiredness etc) can move the agent from one state to another.

Figure 4 shows a sample FSM for the control of an NPC in a typical action game. In this example the behaviors of the character are determined by just four states – CHASE, ATTACK, FLEE and EXPLORE. Each of these states provides an action that the agent should take. For exam-

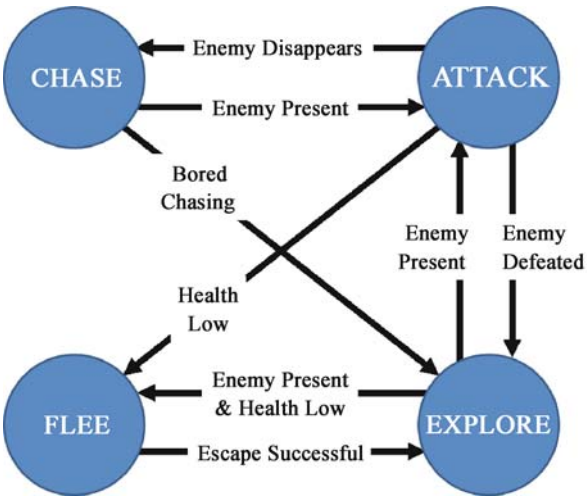

Computer Graphics and Games, Agent Based Modeling in, Figure 4
A simple finite state machine for a soldier NPC in an action game

ple, when in the EXPLORE state the character should wander randomly around the world, or while in the FLEE state the character should determine a direction to move in that will take it away from its current enemy and move in that direction. The links between the states show how the behaviors of the character should move between the various available states. So, for example, if while in the ATTACK state the agent's health measure becomes low, they will move to the FLEE state and run away from their enemy.

FSMs are widely used because they are so simple, well understood and extremely efficient both in terms of processing cycles required and memory usage. There have also been a number of highly successful augmentations to the basic state machine model to make them more effective, such as the introduction of layers of parallel state machines [3], the use of fuzzy logic in finite state machines [19] and the implementation of cooperative group behaviors through state machines [72].

The action game *Halo 2* is recognized as having a particularly good implementation of state machine based NPC control [79]. At any time an agent could be in any one of the four states *Idle*, *Guard/Patrol*, *Attack/Defend*, and *Retreat*. Within each of these states a set of rules was used in order to select from a small set of appropriate actions for that state – for example a number of different ways to attack the player. The decisions made by NPCs were influenced by a number of character attributes including strength, speed and cowardliness. Transition between states was triggered by perceptions made by characters simulated senses of vision and hearing and by internal attributes such as health. The system implemented also allowed for group behaviors allowing NPCs to hold conversations and cooperate to drive vehicles.

However, FSMs are not without their drawbacks. When designing FSMs developers must envisage every possible situation that might confront an NPC over the course of a game. While this is quite possible for many games, if NPCs are required to move between many different situations this task can become overwhelming. Similarly, as more and more states are added to an FSM designing the links between these states can become a mammoth undertaking.

From [31] the definition of rule based systems states that they are "*… comprised of a database of associated rules. Rules are conditional program statements with consequent actions that are performed if the specified conditions are satisfied*". Rule based systems have been applied extensively to control NPCs in games [16], in particular for the control of NPCs in role-playing games. NPCs behaviors are scripted using a set of rules which typically indicate how an NPC should respond to particular events within the game world. Borrowed from [82], the listing below shows a snippet of the rules used to control a warrior character in the RPG *Baldur's Gate* (www.bioware.com).

```
IF
  // If my nearest enemy is not within 3
  !Range(NearestEnemyOf(Myself),3)
  // and is within 8
  Range(NearestEnemyOf(Myself),8)
THEN
  // 1/3 of the time
  RESPONSE #40
    // Equip my best melee weapon
    EquipMostDamagingMelee()
    // and attack my nearest enemy, checking every 60
    // ticks to make sure he is still the nearest
    AttackReevalutate(NearestEnemyOf (Myself),60)
  // 2/3 of the time
  RESPONSE #60
    // Equip a ranged weapon
    EquipRanged()
    // and attack my nearest enemy, checking every 30
    // ticks to make sure he is still the nearest
    AttackReevalutate(NearestEnemyOf (Myself), 30)
```

The implementation of an NPC using a rule-based system would consist of a large set of such rules, a small set of which would fire based on the conditions in the world at any given time. Rule based systems are favored by game developers as they are relatively simple to use and can be

exhaustively tested. Rule based systems also have the advantage that rule sets can be written using simple proprietary scripting systems [9], rather than full programming languages, making them easy to implement. Development companies have also gone so far as to make these scripting languages available to the general public, enabling them to author there own rule sets.

Rule based systems, however, are not without their drawbacks. Authoring extensive rule sets is not a trivial task, and they are usually restricted to simple situations. Also, rule based systems can be restrictive in that they don't allow sophisticated interplay between NPCs motivations, and require that rule set authors foresee every situation that the NPC might find itself in.

Some of the disadvantages of simple rule based systems can be alleviated by using more sophisticated inference engines. One example uses Dempster Schafer theory [43] which allows rules to be evaluated by combining multiple sources of (often incomplete) evidence to determine actions. This goes some way towards supporting the use of rule based systems in situations where complete knowledge is not available.

ALife techniques have also been applied extensively in the control of game NPCs, as much as a philosophy as any particular techniques. The outstanding example of this is *The Sims* (thesims.ea.com) a surprise hit of 2000 which has gone on to become the best selling PC game of all time. Created by games guru Will Wright the Sims puts the player in control of the lives of a virtual family in their virtual home. Inspired by aLife, the characters in the game have a set of motivations, such as hunger, fatigue and boredom and seek out items within the game world that can satisfy these desires. Virtual characters also develop sophisticated social relationships with each other based on common interest, attraction and the amount of time spent together. The original system in the Sims has gone on to be improved in the sequel *The Sims 2* and a series of expansion packs.

Some of the more interesting work in developing techniques for the control of game characters (particularly in action games) has been focused on developing interesting sensing and memory models for game characters. Players expect when playing action games that computer controlled opponents should suffer from the same problems that players do when perceiving the world. So, for example, computer controlled characters should not be able to see through walls or from one floor to the next. Similarly, though, players expect computer controlled characters to be capable of perceiving events that occur in a world and so NPCs should respond appropriately to sound events or on seeing the player.

One particularly fine example of a sensing model was in the game *Thief: The Dark Project* where players are required to sneak around an environment without alerting guards to their presence [45]. The developers produced a relatively sophisticated sensing model that was used by non-player characters which modeled visual effects such as not being able to see the player if they were in shadows, and moving some way towards modeling acoustics so that non-player characters could respond reasonably to sound events.

2004's *Fable* (fable.lionhead.com) took the idea of adding memory to a game to new heights. In this adventure game the player took on the role of a hero from boyhood to manhood. However, every action the player took had an impact on the way in which the game world's population would react to him or her as they would remember every action the next time they met the player. This notion of long-term consequences added an extra layer of believability to the game-playing experience.

Serious Games & Academia

It will probably have become apparent to most readers of the previous section that much of the work done in implementing agent-based techniques for the control of NPCs in commercial games is relatively simplistic when compared to the application of these techniques in other areas of more academic focus, such as robotics [54]. The reasons for this have been discussed already and briefly relate to the lack of available processing resources and the requirements of commercial quality control. However, a large amount of very interesting work is taking place in the application of agent-based technologies in academic research, and in particular the field of serious games. This section will begin by introducing the area of serious games and then go on to discuss interesting academic projects looking at agent-based technologies in games.

The term serious games [53] refers to games designed to do more than just entertain. Rather, serious games, while having many features in common with conventional games, have ulterior motives such as teaching, training, and marketing. Although games have been used for ends apart from entertainment, in particular education, for a long time, the modern serious games movement is set apart from these by the level of sophistication of the games it creates. The current generation of serious games is comparable with main-stream games in terms of the quality of production and sophistication of their design. Serious games offer particularly interesting opportunities for the use of agent-based modeling techniques due to the facts that they often do not

have to live up to the rigorous testing of commercial games, can have the requirement of specialized hardware rather than being restricted to commercial games hardware and often, by the nature of their application domains, require more in-depth interactions between players and NPCs.

The modern serious games movement can be said to have begun with the release of *America's Army* (www.americasarmy.com) in 2002 [57]. Inspired by the realism of commercial games such as the *Rainbow 6* series (www.rainbow6.com), the United States military developed America's Army and released it free of charge in order to give potential recruits a flavor of army life. The game was hugely successful and is still being used today as both a recruitment tool and as an internal army training tool.

Spurred on by the success of America's Army the serious games movement began to grow, particularly within academia. A number of conferences sprung up and notably the Serious Games Summit became a part of the influential Game Developer's Conference (www.gdconf.com) in 2004.

Some other notable offerings in the serious games field include *Food Force* (www.food-force.com) [18], a game developed by the United Nations World Food Programme in order to promote awareness of the issues surrounding emergency food aid; *Hazmat Hotzone* [15], a game developed by the Entertainment Technology Centre at Carnegie Mellon University to train fire-fighters to deal with chemical and hazardous materials emergencies; *Yourself!Fitness* (www.yourselffitness.com) [53] an interactive virtual personal trainer developed for modern games consoles; and *Serious Gordon* (www.seriousgames.ie) [50] a game developed to aid in teaching food safety in kitchens. A screen shot of Serious Gordon is shown in Fig. 5.

Over the past decade, interest in academic research that is directly focused on artificial intelligence, and in particular agent-based modelling techniques and their application to games (as opposed to the general virtual character/computer graphics work discussed previously) has grown dramatically. One of the first major academic research projects into the area of Game-AI was led by John Laird at the University of Michigan, in the United States. The SOAR architecture was developed in the early nineteen eighties in an attempt to *"develop and apply a unified theory of human and artificial intelligence"* [66]. SOAR is essentially a rule based inference system which takes the current state of a problem and matches this to production rules which lead to actions.

Computer Graphics and Games, Agent Based Modeling in, Figure 5
A screenshot of Serious Gordon, a serious game developed to aid in the teaching of food safety in kitchens

After initial applications into the kind of simple puzzle worlds which characterized early AI research [42], the SOAR architecture was applied to the task of controlling computer generated forces [37]. This work lead to an obvious transfer to the new research area of game-AI [40].

Initially the work of Laird's group focused on applying the SOAR architecture to the task of controlling NPC opponents in the action game *Quake* (www.idsoftware.com) [40]. This proved quite successful leading to opponents which could successfully play against human players, and even begin to plan based on anticipation of what the player was about to do. More recently Laird's group have focused on the development of a game which requires more involved interactions between the player and the NPCs. Named *Haunt 2*, this game casts the player in the role of a ghost that must attempt to influence the actions of a group of computer controlled characters inhabiting the ghost's haunted house [51]. The main issue that arises with the use the SOAR architecture is that it is enormously resource hungry, with the NPC controllers running on a separate machine to the actual game.

At Trinity College in Dublin in Ireland, the author of this article worked on an intelligent agent architecture, the Proactive Persistent Agent (PPA) architecture, for the control of background characters (or support characters) in character-centric games (games that focus on character interactions rather than action, e. g. role-playing games) [48,49]. The key contributions of this work were that it made possible the creation of NPCs that were capable of behaving believably in a wide range of situations and allowed for the creation of game environments which it appeared had an existence beyond their interactions with players. Agent behaviors in this work were based on models of personality, emotion, relationships to other characters and behavioral models that changed according to the current role of an agent. This system was used to develop a stand alone game and as part of a simulation of areas within Trinity College. A screenshot of this second application is shown in Fig. 6.

At Northwestern University in Chicago the Interactive Entertainment group has also applied approaches from more traditional research areas to the problems facing game-AI. Ian Horswill has led a team that are attempting to use architectures traditionally associated with robotics for the control of NPCs. In [29] Horswill and Zubek consider how perfectly matched the behavior based architectures often used in robotics are with the requirements of NPC control architectures. The group have demonstrated some of their ideas in a test-bed environment built on top

Computer Graphics and Games, Agent Based Modeling in, Figure 6
Screenshots of the PPA system simulating parts of a college

of the game Half-Life [38]. The group also looks at issues around character interaction [85] and the many psychological issues associated with creating virtual characters asking how we can create virtual game agents that display all of the foibles that make us relate to characters in human stories [28].

Within the same research group a team led by Ken Forbus have extended research previously undertaken in conjunction with the military [24] and applied it to the problem of terrain analysis in computer strategy games [25]. Their goal is to create strategic opponents which are capable of performing sophisticated reasoning about the terrain in a game world and using this knowledge to identify complex features such as ambush points. This kind of high level reasoning would allow AI opponents play a much more realistic game, and even surprise human players from time to time, something that is sorely missing from current strategy games.

As well as this work which has spring-boarded from existing applications, a number of projects began expressly to tackle problems in game-AI. Two which particularly stand out are the Excalibur Project, led by Alexander Nareyek [55] and work by John Funge [26]. Both of these projects have attempted to applying sophisticated planning techniques to the control of game characters.

Nareyek uses constraint based planning to allow game agents reason about their world. By using techniques such as local search Nareyek has attempted to allow these sophisticated agents perform resource intensive planning within the constraints of a typical computer game environment. Following on from this work, the term *anytime agent* was coined to describe the process by which agents actively refine original plans based on changing world conditions. In [56] Narayek describes the directions in which he intends to take this work in future.

Funge uses the situational calculus to allow agents reason about their world. Similarly to Nareyek he has addressed the problems of a dynamic, ever changing world, plan refining and incomplete information. Funge's work uses an extension to the situational calculus which allows the expression of uncertainty. Since completing this work Funge has gone on to be one of the founders of AiLive (www.ailive.net), a middleware company specializing in AI for games.

While the approaches of both of these projects have shown promise within the constrained environments to which they have been applied during research, (and work continues on them) it remains to be seen whether such techniques can be successfully applied to a commercial game environment and all of the resource constraints that such an environment entails.

One of the most interesting recent examples of agent-based work in the field of serious games is that undertaken by Barry Silverman and his group at the University of Pennsylvania in the United States [69,70]. Silverman models the protagonists in military simulations for use in training programmes and has taken a very interesting approach in that his agent models are based on established cognitive science and behavioral science research. While Silverman admits that many of the models described in the cognitive science and behavioral science literature are not well quantified enough to be directly implemented, he has adapted a number of well respected models for his purposes. Silverman's work is an excellent example of the capabilities that can be explored in a serious games setting rather than a commercial game setting, and as such merits an in depth discussion. A high-level schematic diagram of Silverman's approach is shown in Fig. 7 and shows the agent architecture used by Silverman's system, PMFserv.

The first important component of the PMFserv system is the biology module which controls biological needs using a metaphor based on the flow of water through a system. Biological concepts such as hunger and fatigue are simulated using a series of reservoirs, tanks and valves which model the way in which resources are consumed by the system. This biological model is used in part to model stress which has an important impact on the way in which agents make decisions. To model the way in which agent performance changes under pressure Silverman uses *performance moderator functions* (PMFs). An example of one of the earliest PMFs used is the Yerkes–Dodson "inverted-u" curve [84] which illustrates that as mental arousal is increased performance initially improves, peaks and then trails off again. In PMFserv a range of PMFs are used to model the way in which behavior should change depending on stress levels and biological conditions.

The second important module of PMFserv attempts to model how personality, culture and emotion affect the behavior of an agent. In keeping with the rest of their system PMFserv uses models inspired by cognitive science to model emotions. In this case the well known OCC model [58], which has been used in agent-based applications before [8], is used. The OCC model provides for 11 pairs of opposite emotions such as pride and shame, and hope and fear. The emotional state of an agent with regard to past, current and future actions heavily influences the decisions that the agent makes.

The second portion of the Personality, Culture, Emotion module uses a value tree in order to capture the values of an agent. These values are divided into a Preference Tree which captures long term desired states for the world, a Standards Tree which relates to the actions that an agent

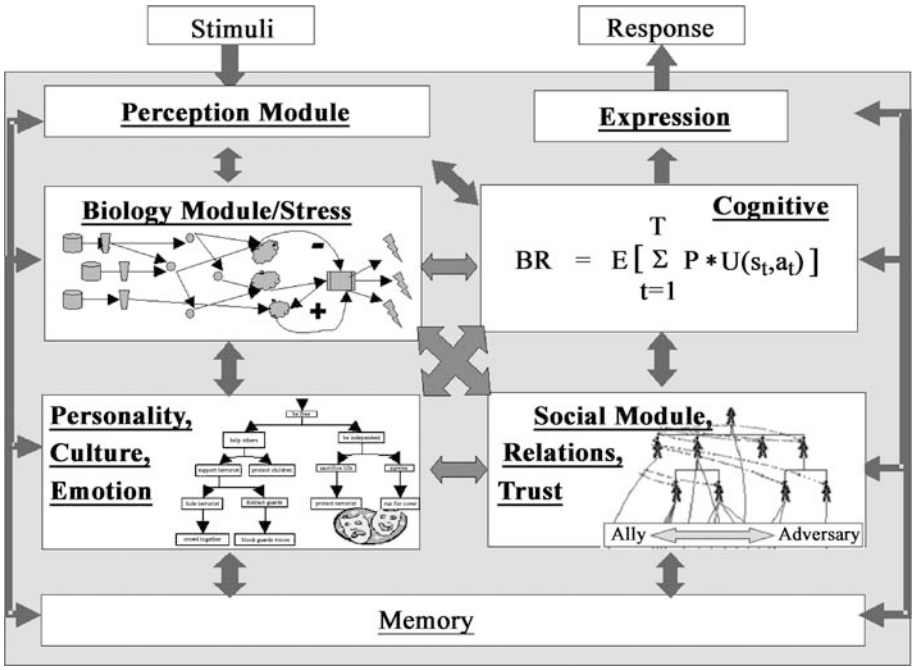

Computer Graphics and Games, Agent Based Modeling in, Figure 7
A schematic diagram of the main components of the PMFserv system (with kind permission of Barry Silverman)

believes it can or cannot follow in order to achieve these desired states and a Goal Tree which captures short term goals.

PMFserv also models the relationships between agents (Social Model, Relations, Trust in Fig. 7). The relationship of one agent to another is modeled in terms of three axes. The first is the degree to which the other agent is thought of as a human rather than an inanimate object – locals tend to view American soldiers as objects rather than people. The second axis is the cognitive grouping (ally, foe etc) to which the other agent belongs and whether this is also a group to which the first agent has an affinity. Finally, the valence, or strength, of the relationship is stored. Relationships continually change based on actions that occur within the game world. Like the other modules of the system this model is also based on psychological research [58].

The final important module of the PMFserv architecture is the Cognitive module which is used to decide on particular actions that agents will undertake. This module uses inputs from all of the other modules to make these decisions and so the behavior of PMFserv agents is driven by their stress levels, relationships to other agents and objects within the game world, personality, culture and emotions. The details of the PMFserv cognitive process are beyond the scope of this article, so it will suffice to say that action selection is based on a calculation of the utility of a particular action to an agent, with this calculation modified by the factors listed above.

The most highly developed example using the PMFserv model is a simulation of the 1993 event in Mogadishu, Somalia in which a United States military Black Hawk helicopter crashed, as made famous by the book and film "Black Hawk Down" [12]. In this example, which was developed as a military training aid as part of a larger project looking at agent implementations within such systems [78,81] the player took on the role of a US army ranger on a mission to secure the helicopter wreck in a modification (or "mod") of the game *Unreal Tournament* (www.unreal.com). A screenshot of this simulation is shown in Fig. 8.

The PMFserv system was used to control the behaviors of characters within the game world such as Somali militia, and Somali civilians. These characters were imbued with physical attributes, a value system and relationships with other characters and objects within the game environment. The sophistication of PMFserv was apparent in many of the behaviors of the simulations NPCs. One particularly good example was the fact that Somali women would offer themselves as human shields for militia fighters. This behavior was never directly programmed into the agents make-up, but rather emerged as a result of their values and assessment of their situation. PMFserv remains one of the most sophisticated current agent implementa-

Computer Graphics and Games, Agent Based Modeling in, Figure 8
A screenshot of the PMFserv system being used to simulate the Black Hawk Down scenario (with kind permission of Barry Silverman)

tions and shows the possibilities when the shackles of commercial game constraints are thrown off.

Future Directions

There is no doubt that with the increase in the amount of work being focused on the use of agent-based modeling in computer graphics and games there will be major developments in the near future. This final section will attempt to predict what some of these might be.

The main development that might be expected in all of the areas that have been discussed in this article is an increase in the depth of simulation. The primary driver of this increase in depth will be the development of more sophisticated agent models which can be used to drive ever more sophisticated agent behavior. The PMFserv system described earlier is one example of the kinds of deeper systems that are currently being developed. In general computer graphics applications this will allow for the creation of more interesting simulations including previously prohibitive features such as automatic realistic facial expressions and other physical expressions of agents' internal states. This would be particularly use in CGI for movies in which, although agent based modeling techniques are commonly used for crowd scenes and background characters, main characters are still animated almost entirely by hand.

In the area of computer games it can be expected that many of the techniques being used in movie CGI will filter over to real-time game applications as the processing power of game hardware increases – this is a pattern that has been evident for the past number of years. In terms of depth that might be added to the control of game characters one feature that has mainly been conspicuous by its absence in modern games is genuine learning by game agents. 2000's *Black & White* and its sequel *Black & White 2* (www.lionhead.com) featured some learning by one of the game's main characters that the player could teach in a reinforcement manner [20]. While this was particularly successful in the game, such techniques have not been more widely applied. One interesting academic project in this area is the NERO project (www.nerogame.org) which allows a player to train an evolving army of soldiers and have them battle the armies of other players [73]. It is expected that these kinds of capabilities will become more and more common in commercial games.

One new feature of the field of virtual character control in games is the emergence of specialized middleware. Middleware has had a massive impact in other areas of game development including character mod-

eling (for example Maya available from www.autodesk.com) and physics modeling (for example Havok available from www.havok.com). AI focused middleware for games is now becoming more common with notable offerings including AI-Implant (www.ai-implant.com) and Kynogon (www.kynogon.com) which perform path finding and state machine based control of characters. It is expected that more sophisticated techniques will over time find their way into such software.

To conclude the great hope for the future is that more and more sophisticated agent-based modeling techniques from other application areas and other branches of AI will find their way into the control of virtual characters.

Bibliography

Primary Literature

1. Adamson A (Director) (2005) The Chronicles of Narnia: The Lion, the Witch and the Wardrobe. Motion Picture. http://adisney.go.com/disneypictures/narnia/lb_main.html
2. Aitken M, Butler G, Lemmon D, Saindon E, Peters D, Williams G (2004) The Lord of the Rings: the visual effects that brought middle earth to the screen. International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), Course Notes
3. Alexander T (2003) Parallel-State Machines for Believable Characters. In: Massively Multiplayer Game Development. Charles River Media
4. Allers R, Minkoff R (Directors) (1994) The Lion King. Motion Picture. http://disney.go.com/disneyvideos/animatedfilms/lionking/
5. Aylett R, Luck M (2000) Applying Artificial Intelligence to Virtual Reality: Intelligent Virtual Environments. Appl Artif Intell 14(1):3–32
6. Badler N, Bindiganavale R, Bourne J, Allbeck J, Shi J, Palmer M (1999) Real Time Virtual Humans. In: Proceedings of the International conference on Digital Media Futures.
7. Bates J (1992) The Nature of Characters in Interactive Worlds and the Oz Project. Technical Report CMU-CS-92–200. School of Computer Science, Carnegie Melon University
8. Bates J (1992) Virtual reality, art, and entertainment. Presence: J Teleoper Virtual Environ 1(1):133–138
9. Berger L (2002) Scripting: Overview and Code Generation. In: Rabin S (ed) AI Game Programming wisdom. Charles River Media
10. Bird B, Pinkava J (Directors) (2007) Ratatouille. Motion Picture. http://disney.go.com/disneyvideos/animatedfilms/ratatouille/
11. Blumberg B (1996) Old Tricks, New Dogs: Ethology and Interactive Creatures. Ph D Thesis, Media Lab, Massachusetts Institute of Technology
12. Bowden M (2000) Black Hawk Down. Corgi Adult
13. Burke R, Isla D, Downie M, Ivanov Y, Blumberg B (2002) Creature Smarts: The Art and Architecture of a Virtual Brain. In: Proceedings of Game-On 2002: the 3rd International Conference on Intelligent Games and Simulation, pp 89–93
14. Burton T (Director) (1992) Batman Returns. Motion Picture. http://www.warnervideo.com/batmanmoviesondvd/
15. Carless S (2005) Postcard From SGS 2005: Hazmat: Hotzone – First-Person First Responder Gaming. Retrieved October 2007, from Gamasutra: www.gamasutra.com/features/20051102/carless_01b.shtml
16. Christian M (2002) A Simple Inference Engine for a Rule Based Architecture. In: Rabin S (ed) AI Game Programming Wisdom. Charles River Media
17. Darnell E, Johnson T (Directors) (1998) Antz. Motion Picture. http://www.dreamworksanimation.com/
18. DeMaria R (2005) Postcard from the Serious Games Summit: How the United Nations Fights Hunger with Food Force. Retrieved October 2007, from Gamasutra: www.gamasutra.com/features/20051104/demaria_01.shtml
19. Dybsand E (2001) A Generic Fuzzy State Machine in C++. In: Rabin S (ed) Game Programming Gems 2. Charles River Media
20. Evans R (2002) Varieties of Learning. In: Rabin S (ed) AI Game Programming Wisdom. Charles River Media
21. Faloutsos P, van de Panne M, Terzopoulos D (2001) The Virtual Stuntman: Dynamic Characters with a Repetoire of Autonomous Motor Skills. Comput Graph 25(6):933–953
22. Farenc N, Musse S, Schweiss E, Kallmann M, Aune O, Boulic R et al (2000) A Paradigm for Controlling Virtual Humans in Urban Environment Simulations. Appl Artif Intell J Special Issue Intell Virtual Environ 14(1):69–91
23. Feng-Hsiung H (2002) Behind Deep Blue: Building the Computer that Defeated the World Chess Champion. Princeton University Press
24. Forbus K, Nielsen P, Faltings B (1991) Qualitative Spatial Reasoning: The CLOCK Project. Artif Intell 51:1–3
25. Forbus K, Mahoney J, Dill K (2001) How Qualitative Spatial Reasoning Can Improve Strategy Game AIs. In: Proceedings of the AAAI Spring Symposium on AI and Interactive Entertainment
26. Funge J (1999) AI for Games and Animation: A Cognitive Modeling Approach. A.K. Peters
27. Hayes-Roth B, Doyle P (1998) Animate Characters. Auton Agents Multi-Agent Syst 1(2):195–230
28. Horswill I (2007) Psychopathology, narrative, and cognitive architecture (or: why NPCs should be just as screwed up as we are). In: Proceedings of AAAI Fall Symposium on Intelligent Narrative Technologies
29. Horswill I, Zubek R (1999) Robot Architectures for Believable Game Agents. In: Proceedings of the 1999 AAAI Spring Symposium on Artificial Intelligence and Computer Games
30. Houlette R, Fu D (2003) The Ultimate Guide to FSMs in Games. In: Rabin S (ed) AI Game Programming Wisdom 2. Charles River Media
31. IGDA (2003) Working Group on Rule-Based Systems Report. International Games Development Association
32. Isbister K, Doyle P (2002) Design and Evaluation of Embodied Conversational Agents: A Proposed Taxonomy. In: Proceedings of the AA- MAS02 Workshop on Embodied Conversational Agents: Lets Specify and Compare Them! Bologna, Italy
33. Jackson P (Director) (2001) The Lord of the Rings: The Fellowship of the Ring. Motion Picture. http://www.lordoftherings.net/
34. Jackson P (Director) (2002) The Lord of the Rings: The Two Towers. Motion Picture. http://www.lordoftherings.net/
35. Jackson P (Director) (2003) The Lord of the Rings: The Return of the King. Motion Picture. http://www.lordoftherings.net/

36. Johnston O, Thomas F (1995) The Illusion of Life: Disney Animation. Disney Editions
37. Jones R, Laird J, Neilsen P, Coulter K, Kenny P, Koss F (1999) Automated Intelligent Pilots for Combat Flight Simulation. AI Mag 20(1):27–42
38. Khoo A, Zubek R (2002) Applying Inexpensive AI Techniques to Computer Games. IEE Intell Syst Spec Issue Interact Entertain 17(4):48–53
39. Koeppel D (2002) Massive Attack. http://www.popsci.com/popsci/science/d726359b9fa84010vgnvcm1000004eecbccdrcrd.html. Accessed Oct 2007
40. Laird J (2000) An Exploration into Computer Games and Computer Generated Forces. The 8th Conference on Computer Generated Forces and Behavior Representation
41. Laird J, van Lent M (2000) Human-Level AI's Killer Application: Interactive Computer Games. In: Proceedings of the 17th National Conference on Artificial Intelligence
42. Laird J, Rosenbloom P, Newell A (1984) Towards Chunking as a General Learning Mechanism. The 1984 National Conference on Artificial Intelligence (AAAI), pp 188–192
43. Laramée F (2002) A Rule Based Architecture Using Dempster-Schafer theory. In: Rabin S (ed) AI Game Programming Wisdom. Charles River Media
44. Lasseter J, Stanton A (Directors) (1998) A Bug's Life; Motion Picture. http://www.pixar.com/featurefilms/abl/
45. Leonard T (2003) Building an AI Sensory System: Examining the Deign of Thief: The Dark Project. In: Proceedings of the 2003 Game Developers' Conference, San Jose
46. Loyall B (1997) Believable Agents: Building Interactive Personalities. Ph D Thesis, Carnegie Melon University
47. Määta A (2002) Realistic Level Design for Max Payne. In: Proceedings of the 2002 Game Developer's conference, GDC 2002
48. Mac Namee B, Cunningham P (2003) Creating Socially Interactive Non Player Characters: The μ-SIC System. Int J Intell Games Simul 2(1)
49. Mac Namee B, Dobbyn S, Cunningham P, O'Sullivan C (2003) Simulating Virtual Humans Across Diverse Situations. In: Proceedings of Intelligent Virtual Agents '03, pp 159–163
50. Mac Namee B, Rooney P, Lindstrom P, Ritchie A, Boylan F, Burke G (2006) Serious Gordon: Using Serious Games to Teach Food Safety in the Kitchen. The 9th International Conference on Computer Games: AI, Animation, Mobile, Educational & Serious Games CGAMES06, Dublin
51. Magerko B, Laird JE, Assanie M, Kerfoot A, Stokes D (2004) AI Characters and Directors for Interactive Computer Games. The 2004 Innovative Applications of Artificial Intelligence Conference. AAAI Press, San Jose
52. Thalmann MN, Thalmann D (1994) Artificial Life and Virtual Reality. Wiley
53. Michael D, Chen S (2005) Serious Games: Games That Educate, Train, and Inform. Course Technology PTR
54. Muller J (1996) The Design of Intelligent Agents: A Layered Approach. Springer
55. Nareyek A (2001) Constraint Based Agents. Springer
56. Nareyek A (2007) Game AI is Dead. Long Live Game AI! IEEE Intell Syst 22(1):9–11
57. Nieborg D (2004) America's Army: More Than a Game. Bridging the Gap;Transforming Knowledge into Action through Gaming and Simulation. Proceedings of the 35th Conference of the International Simulation and Gaming Association (ISAGA), Munich
58. Ortony A, Clore GL, Collins A (1988) The cognitive structure of emotions. Cambridge University Press, Cambridge
59. Perlin K, Goldberg A (1996) Improv: A System for Scripting Interactive Actors in Virtual Worlds. In: Proceedings of the ACM Computer Graphics Annual Conference, pp 205–216
60. Proyas A (Director) (2004) I, Robot. Motion Picture. http://www.irobotmovie.com
61. Rao AS, Georgeff MP (1991) Modeling rational agents within a BDI-architecture. In: Proceedings of Knowledge Representation and Reasoning (KR&R-91). Morgan Kaufmann, pp 473–484
62. Musse RS, Thalmann D (2001) A Behavioral Model for Real Time Simulation of Virtual Human Crowds. IEEE Trans Vis Comput Graph 7(2):152–164
63. Reed C, Geisler B (2003) Jumping, Climbing, and Tactical Reasoning: How to Get More Out of a Navigation System. In: Rabin S (ed) AI Game Programming Wisdom 2. Charles River Media
64. Reynolds C (1987) Flocks, Herds and Schools: A Distributed Behavioral Model. Comput Graph 21(4):25–34
65. Rodriguez R (Director) (1996) From Dusk 'Till Dawn. Motion Picture
66. Rosenbloom P, Laird J, Newell A (1993) The SOAR Papers: Readings on Integrated Intelligence. MIT Press
67. Sánchez-Crespo D (2006) GDC: Physical Gameplay in Half-Life 2. Retrieved October 2007, from gamasutra.com: http://www.gamasutra.com/features/20060329/sanchez_01.shtml
68. Shao W, Terzopoulos D (2005) Autonomous Pedestrians. In: Proceedings of SIGGRAPH/EG Symposium on Computer Animation, SCA'05, pp 19–28
69. Silverman BG, Bharathy G, O'Brien K, Cornwell J (2006) Human Behavior Models for Agents in Simulators and Games: Part II: Gamebot Engineering with PMFserv. Presence Teleoper Virtual Worlds 15(2):163–185
70. Silverman BG, Johns M, Cornwell J, O'Brien K (2006) Human Behavior Models for Agents in Simulators and Games: Part I: Enabling Science with PMFserv. Presence Teleoper Virtual Environ 15(2):139–162
71. Smith P (2002) Polygon Soup for the Programmer's Soul: 3D Path Finding. In: Proceedings of the Game Developer's Conference 2002, GDC2002
72. Snavely P (2002) Agent Cooperation in FSMs for Baseball. In: Rabin S (ed) AI Game Programming Wisdom. Charles River Media
73. Stanley KO, Bryant BD, Karpov I, Miikkulainen R (2006) Real-Time Evolution of Neural Networks in the NERO Video Game. In: Proceedings of the Twenty-First National Conference on Artificial Intelligence, AAAI-2006. AAAI Press, pp 1671–1674
74. Stout B (1996) Smart Moves: Intelligent Path-Finding. Game Dev Mag Oct
75. Takahashi TS (1992) Behavior Simulation by Network Model. Memoirs of Kougakuin University 73, pp 213–220
76. Terzopoulos D, Tu X, Grzeszczuk R (1994) Artificial Fishes with Autonomous Locomotion, Perception, Behavior and Learning, in a Physical World. In: Proceedings of the Artificial Life IV Workshop. MIT Press
77. Thompson C (2007) Halo 3: How Microsoft Labs Invented a New Science of Play. Retrieved October 2007, from wired.com: http://www.wired.com/gaming/virtualworlds/magazine/15-09/ff_halo
78. Toth J, Graham N, van Lent M (2003) Leveraging gaming in DOD modelling and simulation: Integrating performance and behavior moderator functions into a general cognitive archi-

tecture of playing and non-playing characters. Twelfth Conference on Behavior Representation in Modeling and Simulation (BRIMS, formerly CGF), Scotsdale, Arizona
79. Valdes R (2004) In the Mind of the Enemy: The Artificial Intelligence of Halo 2. Retrieved October 2007, from HowStuffWorks.com: http://entertainment.howstuffworks.com/halo2-ai.htm
80. van der Werf E, Uiterwijk J, van den Herik J (2002) Programming a Computer to Play and Solve Ponnuki-Go. In: Proceedings of Game-On 2002: The 3rd International Conference on Intelligent Games and Simulation, pp 173–177
81. van Lent M, McAlinden R, Brobst P (2004) Enhancing the behavioral fidelity of synthetic entities with human behavior models. Thirteenth Conference on Behavior Representation in Modeling and Simulation (BRIMS)
82. Woodcock S (2000) AI Roundtable Moderator's Report. In: Proceedings of the Game Developer's Conference 2000 (GDC2000)
83. Wooldridge M, Jennings N (1995) Intelligent Agents: Theory and Practice. Know Eng Rev 10(2):115–152
84. Yerkes RW, Dodson JD (1908) The relation of strength of stimulus to rapidity of habit formation. J Comp Neurol Psychol 18:459–482
85. Zubek R, Horswill I (2005) Hierarchical Parallel Markov Models of Interaction. In: Proceedings of the Artificial Intelligence and Interactive Digital Entertainment Conference, AIIDE 2005

Books and Reviews

DeLoura M (ed) (2000) Game Programming Gems. Charles River Media
DeLoura M (ed) (2001) Game Programming Gems 2. Charles River Media
Dickheiser M (ed) (2006) Game Programming Gems 6. Charles River Media
Kirmse A (ed) (2004) Game Programming Gems 4. Charles River Media
Pallister K (ed) (2005) Game Programming Gems 5. Charles River Media
Rabin S (ed) (2002) Game AI Wisdom. Charles River Media
Rabin S (ed) (2003) Game AI Wisdom 2. Charles River Media
Rabin S (ed) (2006) Game AI Wisdom 3. Charles River Media
Russell S, Norvig P (2002) Artificial Intelligence: A Modern Approach. Prentice Hall
Treglia D (ed) (2002) Game Programming Gems 3. Charles River Media

Computing in Geometrical Constrained Excitable Chemical Systems

Jerzy Gorecki[1,2], Joanna Natalia Gorecka[3]
[1] Institute of Physical Chemistry, Polish Academy of Science, Warsaw, Poland
[2] Faculty of Mathematics and Natural Sciences, Cardinal Stefan Wyszynski University, Warsaw, Poland
[3] Institute of Physics, Polish Academy of Science, Warsaw, Poland

Article Outline

Glossary
Definition of the Subject
Introduction
Logic Gates, Coincidence Detectors and Signal Filters
Chemical Sensors Built with Structured Excitable Media
The Ring Memory and Its Applications
Artificial Chemical Neurons with Excitable Medium
Perspectives and Conclusions
Acknowledgments
Bibliography

Glossary

Some of the terms used in our article are the same as in the article of Adamatzky in this volume, and they are explained in the glossary of ► Reaction-Diffusion Computing.

Activator A substance that increases the rate of reaction.

Excitability Here we call a dynamical system excitable if it has a single stable state (the rest state) with the following properties: if the rest state is slightly perturbed then the perturbation uniformly decreases as the system evolves towards it. However, if the perturbation is sufficiently strong it may grow by orders of magnitude before the system approaches the rest state. The increase in variables characterizing the system (usually rapid if compared with the time necessary to reach the rest state) is called an excitation. A forest is a classical example of excitable medium and a wildfire that burns it is an excitation. A dynamical system is non-excitable if applied perturbations do not grow up and finally decay.

Excitability level The measure of how strong a perturbation has to be applied to excite the system. For example, for the Ru-catalyzed Belousov–Zhabotinsky reaction, increasing illumination makes the medium less excitable. The decrease in the excitability level can be observed in reduced amplitudes of spikes and in decreased velocities of autowaves.

Firing number The ratio between the number of generated excitations to the number of applied external stimulations. In most of the cases we define the firing number as the ratio between the number of output spikes to the number of arriving pulses.

Inhibitor A substance that decreases the rate of reaction or even prevents it.

Medium In the article we consider a chemical medium, fully characterized by local concentrations of reagents and external conditions like temperature or illumination level. The time evolution of concentrations is gov-

erned by a set of reaction-diffusion equations, where the reaction term is an algebraic function of variables characterizing the system and the non-local coupling is described by the diffusion operator. We are mainly concerned with a two dimensional medium (i. e., a membrane with a solution of reagents used in experiments), but the presented ideas can be also applied to one-dimensional or three-dimensional media.

Refractory period A period of time during which an excitable system is incapable of repeating its response to an applied, strong perturbation. After the refractory period the excitable medium is ready to produce an excitation as an answer to the stimulus.

Spike, autowave In a spatially distributed excitable medium, a local excitation can spread around as a pulse of excitation. Usually a propagating pulse of excitation converges to a stationary shape characteristic for the medium, which is not dependent on initialization and propagates with a constant velocity – thus it is called an autowave.

Subexcitability A system is called subexcitable if the amplitude and size of the initiated pulse of excitation decreases in time. However, if the decay time is comparable with the characteristic time for the system, defined as the ratio between system size and pulse velocity, then pulses in a subexcitable system travel for a sufficiently long distance to carry information. Subexcitable media can be used to control the amplitude of excitations. Subexcitability is usually related to system dynamics, but it may also appear as the result of geometrical constraints. For example, a narrow excitable channel surrounded by a non-excitable medium may behave as a subexcitable system because a propagating pulse dies out due to the diffusion of the activator in the neighborhood.

Definition of the Subject

It has been shown in the article of Adamatzky ► Reaction-Diffusion Computing that an excitable system can be used as an information processing medium. In such a medium, information is coded in pulses of excitation; the presence of a single excitation or of a group of excitations forms a message. Information processing discussed by Adamatzky is based on a homogeneous excitable medium and the interaction between pulses in such medium. Here we focus our attention on a quite specific type of excitable medium that has an intentionally introduced structure of regions characterized by different excitability levels. As the simplest case we consider a medium composed of excitable regions where autowaves can propagate and non-excitable ones where excitations rapidly die. Using such two types of medium one can, for example, construct signal channels: stripes of excitable medium where pulses can propagate surrounded by non-excitable areas thick enough to cancel potential interactions with pulses propagating in neighboring channels. Therefore, the propagation of a pulse along a selected line in an excitable system can be realized in two ways. In a homogeneous excitable medium, it can be done by a continuous control of pulse propagation and the local feedback with activating and inhibiting factors (► Reaction-Diffusion Computing). In a structured excitable medium, the same result can be achieved by creating a proper pattern of excitable and non-excitable regions. The first method gives more flexibility, the second is just simpler and does not require a permanent watch. As we show in this article, a number of devices that perform simple information processing operations, including the basic logic functions, can be easily constructed with structured excitable medium. Combining these devices as building blocks we can perform complex signal processing operations. Such an approach seems similar to the development of electronic computing where early computers were built of simple integrated circuits.

The research on information processing with structured excitable media has been motivated by a few important problems. First, we would like to investigate how the properties of a medium can be efficiently used to construct devices performing given functions, and what tasks are the most suitable for chemical computing. There is also a question of generic designs valid for any excitable medium and specific ones that use unique features of a particular system (for example, a one-dimensional generator of excitation pulses that can be built with an excitable surface reaction [23]). In information processing with a structured excitable medium, the geometry of the medium is as important as its dynamics, and it seems interesting to know what type of structures are related to specific functions. In the article we present a number of such structures characteristic for particular information processing operations.

Another important motivation for research comes from biology. Even the simplest biological organisms can process information and make decisions important for their life without CPU-s, clocks or sequences of commands as it is in the standard von Neumann computer architecture [15]. In biological organisms, even at a very basic cellular level, excitable chemical reactions are responsible for information processing. The cell body considered as an information processing medium is highly structured. We believe that analogs of geometrical structures used for certain types of information processing op-

erations in structured excitable media will be recognized in biological systems, so we will better understand their role in living organisms. At a higher level, the analogies between information processing with chemical media and signal processing in the brain seems to be even closer because the excitable dynamics of calcium in neural tissue is responsible for signal propagation in nerve system [30]. Excitable chemical channels that transmit signals between processing elements look similar to dendrites and axons. As we show in the article, the biological neuron has its chemical analog, and this allows for the construction of artificial neural networks using chemical processes. Having in mind that neural networks are less vulnerable to random errors than classical algorithms one can go back from biology to man-made computing and adopt the concepts in a fast excitable medium, for example especially prepared semiconductors (► Unconventional Computing, Novel Hardware for).

The article is organized in the following way. In the next section we discuss the basic properties of a structured chemical medium that seem useful for information processing. Next we consider the binary information coded in propagating pulses of concentration and demonstrate how logic gates can be built. In the following chapter we show that a structured excitable medium can acquire information about distances and directions of incoming stimuli. Next we present a simple realization of read-write memory cell, discuss its applications in chemical counting devices, and show its importance for programming with pulses of excitation. In the following section we present a chemical realization of artificial neurons that perform multiargument operation on sets of input pulses. Finally we discuss the perspectives of the field, in particular more efficient methods of information coding and some ideas of self-organization that can produce structured media capable of information processing.

Introduction

Excitability is the wide spread behavior of far-from-equilibrium systems [35,39] observed, for example, in chemical reactions (Bielousov–Zhabotynsky BZ-reaction [45], CO oxidation on Pt [37], combustion of gases [26]) as well as in many other physical (laser action) and biochemical (signaling in neural systems, contraction of cardiovascular tissues) processes [51]. All types of excitable systems share a common property that they have a stable stationary state (the rest state) they reside in when they are not perturbed. A small perturbation of the rest state results only in a small-amplitude linear response of the system that uniformly decays in time. However, if a perturbation is sufficiently large then the system can evolve far away from the rest state before finally returning to it. This response is strongly nonlinear and it is accompanied by a large excursion of the variables through phase space, which corresponds to an excitation peak (a spike). The system is refractory after a spike, which means that it takes a certain recovery time before another excitation can take place. The excitability is closely related with relaxation oscillations and the phenomena differ by one bifurcation only [56].

Properties of excitable systems have an important impact on their ability to process information. If an excitable medium is spatially distributed then an excitation at one point of the medium (usually seen as an area with a high concentration of a certain reagent), may introduce a sufficiently large perturbation to excite the neighboring points as the result of diffusion or energy transport. Therefore, an excitation can propagate in space in the form of a pulse. Unlike mechanical waves that dissipate the initial energy and finally decay, traveling spikes use the energy of the medium to propagate and dissipate it. In a typical excitable medium, after a sufficiently long time, an excitation pulse converges to the stationary shape, independent of the initial condition what justifies to call it an autowave. Undamped propagation of signals is especially important if the distances between the emitting and receiving devices are large. The medium's energy comes from the nonequilibrium conditions at which the system is kept. In the case of a batch reactor, the energy of initial composition of reagents allows for the propagation of pulses even for days [41], but a typical time of an experiment in such conditions is less than an hour. In a continuously fed reactor pulses can run as long as the reactants are delivered [44].

If the refractory period of the medium is sufficiently long then the region behind a pulse cannot be excited again for a long time. As a consequence, colliding pulses annihilate. This type of behavior is quite common in excitable systems. Another important feature is the dispersion relation for a train of excitations. Typically, the first pulse is the fastest, and the subsequent ones are slower, which is related to the fact that the medium behind the first pulse has not relaxed completely [63]. For example, this phenomenon is responsible for stabilization of positions in a train of pulses rotating on a ring-shaped excitable area. However, in some systems [43] the anomalous dispersion relation is observed and there are selected stable distances between subsequent spikes. The excitable systems characterized by the anomalous dispersion can play an important role in information processing, because packages of pulses are stable and thus the information coded in such packages can propagate without dispersion.

The mathematical description of excitable chemical media is based on differential equations of reaction-diffusion type, sometimes supplemented by additional equations that describe the evolution of the other important properties of the medium, for example the orientation of the surface in the case of CO oxidation on a Pt surface [12,37]. Numerical simulations of pulse propagation in an excitable medium which are presented in many papers [48,49] use the FitzHugh–Nagumo model describing the time evolution of electrical potentials in nerve channels [17,18,54]. The models for systems with the Belousov–Zhabotinsky (BZ) reaction, for example the Rovinsky and Zhabotinsky model [61,62] for the ferroin catalyzed BZ reaction with the immobilized catalyst, can be derived from "realistic" reaction schemes [16] via different techniques of variable reduction. Experiments with the Ru-catalyzed, photosensitive BZ reaction have become standard in experiments with structured excitable media because the level of excitation can be easily controlled by illumination; see for example [7,27,33]. Light catalyzes the production of bromine that inhibits the reaction, so non illuminated regions are excitable and those strongly illuminated are not. The pattern of excitable (dark) and non-excitable (transparent) fields is just projected on a membrane filled with the reagents. For example, the labyrinth shown in Fig. 1 has been obtained by illuminating a membrane through a proper mask. The presence of a membrane is important because it stops convection in the solution and reduces the speed and size of spikes, so studied systems can be smaller. The other methods of forming excitable channels based on immobilizing a catalyst by imprinting it on a membrane [68] or attaching it by lithography [21,71,72] have been also used, but they seem to be more difficult.

Numerical simulations of the Ru-catalyzed BZ reaction can be performed with different variants of the Oregonator model [9,19,20,38]. For example, the three-variable model uses the following equations:

$$\varepsilon_1 \frac{\partial u}{\partial t} = u(1-u) - w(u-q) + D_u \nabla^2 u \quad (1)$$

$$\frac{\partial v}{\partial t} = u - v \quad (2)$$

$$\varepsilon_2 \frac{\partial w}{\partial t} = \phi + fv - w(u+q) + D_w \nabla^2 w \quad (3)$$

where u, v and w denote dimensionless concentrations of the following reagents: $HBrO_2$, $Ru(4,4'\text{-dm-bpy})_3^{3+}$, and Br^-, respectively. In the considered system of equations, u is an activator and v is an inhibitor. The set of Oregonator equations given above reduces to the two-variable model cited in the article of Adamatzky ► Reaction-Diffusion Computing if the processes responsible for bromide production are very fast if compared to the other reactions ($\varepsilon_2 \ll \varepsilon_1 \ll 1$). In such cases, the local value of w can be calculated assuming that it corresponds to the stationary solution of the third equation. If such a w is substituted into the first equation, one obtains the two-variable Oregonator model. In the equations given above, the units of space and time are dimensionless and they have been chosen to scale the reaction rates. Here we also neglected the diffusion of ruthenium catalytic complex because usually it is much smaller than those of the other reagents. The reaction-diffusion equations describing the time evolution of the system can be solved with the standard numerical techniques [57].

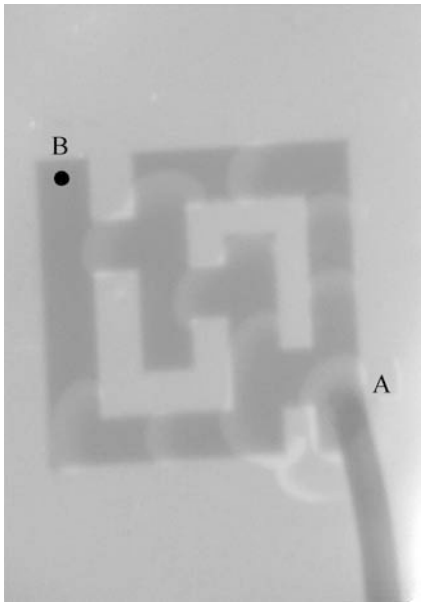

Computing in Geometrical Constrained Excitable Chemical Systems, Figure 1
Pulses of excitation propagating in a labyrinth observed in an experiment with a Ru-catalyzed BZ-reaction. The excitable areas are *dark*, the non-excitable *light*. The source of a train of pulses (a tip of a silver wire) is placed at the point A

The parameter ϕ represents the rate of bromide production caused by illumination and it is proportional to the applied light intensity. Therefore, by adjusting the local illumination (or choosing the proper ϕ as a function of space variables in simulations) we create regions with the required level of excitability, like for example excitable stripes insulated by a non-excitable neighborhood. Of course, the reagents can freely diffuse between the regions characterized by different illuminations.

The structure of the equations describing the system's evolution in time and space gives the name "reaction diffusion computing" [4] to information processing with an

excitable medium. Simulations play an important role in tests of potential information processing devices, because, unlike in experiment, the conditions and parameters of studied systems can be easily adjusted with the required precision and kept forever. Most of the information processing devices discussed below were first tested in simulations and next verified experimentally. One of the problems is related to the short time of a typical experiment in a batch condition (a membrane filled with reagents) that does not exceed one hour. In such systems the period in which the conditions can be regarded as stable is usually shorter than 30 minutes and within this time an experimentalist should prepare the medium, introduce the required illumination and perform observations. The problem can be solved when one uses a continuously fed reactor [44], but experimental setups are more complex. On the other hand, simulations often indicate that the range of parameters in which a given phenomenon appears is very narrow. Fortunately experiments seem to be more robust than simulations and the expected effects can be observed despite of inevitable randomness in reagent preparation. Having in mind the relatively short time of experiments in batch conditions and the low stability of the medium, we believe that applications of a liquid chemical excitable medium like the Ru-catalyzed BZ reaction as information processors (wetware) are rather academic and they are mainly oriented on the verification of ideas. Practical applications of reaction-diffusion computers will probably be based on an other type of medium, like structured semiconductors (► Unconventional Computing, Novel Hardware for).

In an excitable reaction-diffusion medium, spikes propagate along the minimum time path. Historically, one of the first applications of a structured chemical medium used in information processing was the solution of the problem of finding the shortest path in a labyrinth [67]. The idea is illustrated in Fig. 1. The labyrinth is build of excitable channels (dark) separated by non-excitable medium (light) that does not allow for interactions between pulses propagating in different channels. Let us assume that we are interested in the distance between the right bottom corner (point A) and the left upper corner (point B) of the labyrinth shown. The algorithm that scans all the possible paths between these points and selects the shortest one is automatically executed if the paths in labyrinth are build of an excitable chemical medium. To see how it works, let us excite the medium at the point A. The excitation spreads out through the labyrinth, separates at the junctions and spikes enter all possible paths. During the time evolution, pulses of excitation can collide and annihilate, but the one that propagates along the shortest path has always unexcited medium in front. Knowing the time difference between the moment when the pulse is initiated at the point A and the moment when it arrives at the point B and assuming that the speed of a pulse is constant, we can estimate the length of shortest path linking both points. The algorithm described above is called the "prairie fire" algorithm and it is automatically executed by an excitable medium. It finds the shortest path in a highly parallel manner scanning all possible routes at the same time. It is quite remarkable that the time required for finding the shortest path does not depend on the the complexity of labyrinth structure, but only on the distance between the considered points.

Although the estimation of the minimum distance separating two points in a labyrinth is relatively easy (within the assumption that corners do not significantly change pulse speed) it is more difficult to tell what is the shortest path. To do it one can trace the pulse that arrived first and plot a line tangential to its velocity. This idea was discussed in [6] in the context of finding the length of the shortest path in a nonhomogeneous chemical medium with obstacles. The method, although relatively easy, requires the presence of an external observer who follows the propagation of pulses. An alternative technique of extracting the shortest path based on the image processing was demonstrated in [59]. One can also locate the shortest path connecting two points in a labyrinth in a purely chemical way using the coincidence of excitations generated at the endpoints of the path. Such a method, described in [29], allows one to find the midpoint of a trajectory, and thus locate the shortest path point by point.

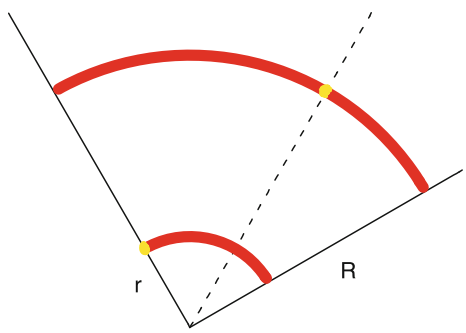

Computing in Geometrical Constrained Excitable Chemical Systems, Figure 2

The idea of angle trisection. Angle arms are linked with *red* arc shaped excitable channels with the radius ratio $r : R = 1 : 3$. The channels have been excited at the same time at the right arm of the angle. At the moment when the excitation pulse (*a yellow dot*) on a smaller arch reaches the second arm the excitation on the other arch shows a point on a line trisecting the angle (*the dashed line*)

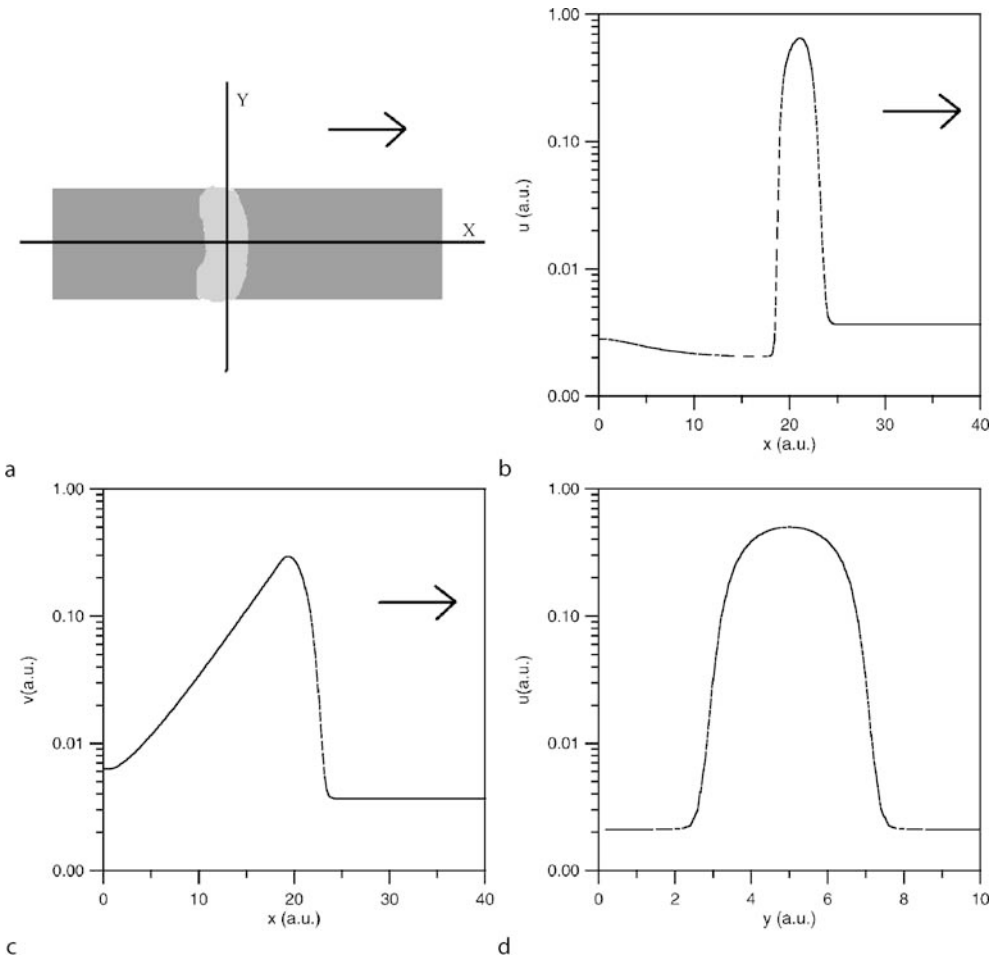

Computing in Geometrical Constrained Excitable Chemical Systems, Figure 3

The shape of the excitation pulse in a stripe of excitable medium. The *arrow* indicates the direction of propagation. Calculations were done for the Oregonator model (Eqs. (1)–(3)) and the following values of parameters were applied: $f = 1.12$, $q = 0.002$, $\varepsilon_1 = 0.08$, $\varepsilon_2 = 0.00097$, $\phi_{\text{excitable}} = 0.007$, $\phi_{\text{non-excitable}} = 0.075$. a The position of a spike on the stripe, b, c concentration of activator and inhibitor along the *x*-axis, d concentration of activator along the *y*-axis

The approximately constant speed of excitation pulses in a reaction-diffusion medium allows one to solve some geometrically oriented problems. For example one can "measure" the number π by comparing the time of pulse propagation around a non-excitable circle of the radius d with the time of propagation around a square with the same side [69]. Similarly, a constant speed of propagating pulses can be used to obtain a given fraction of an angle. For example, a trisection of an angle can be done if the angle arms are linked with two arc-shaped excitable channels as shown in Fig. 2. The ratio of channel radii should be equal to 3. Both channels are excited at the same time at points on the same arm. The position of excitation at the larger arch at the moment when the excitation propagating on the shorter arch reaches the other arm belongs to a line that trisects the angle.

The examples given above are based on two properties of structured excitable medium: the fact that the non-excitability of the neighborhood can restrict the motion of an excitation pulse to a channel, and that the speed of propagation depends on the excitability level of the medium but not on channel shape. This is true when channels are wide and the curvature is small. There are also other properties of an excitable medium useful for information processing. A typical shape of a pulse propagating in a stripe of excitable medium is illustrated in Fig. 3. The pulse moves from the left to the right as the arrow indicates. The profiles on Fig. 3b,c show cross sections of the concentrations

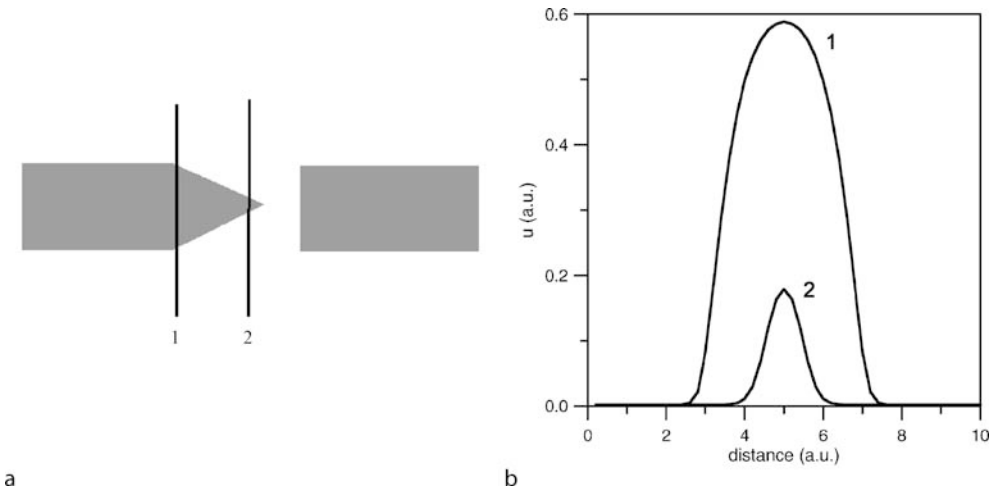

Computing in Geometrical Constrained Excitable Chemical Systems, Figure 4
The shape of activator concentration in an excitable medium of a triangular shape. a The structure of excitable (*dark*) and non-excitable (*light*) regions, b concentration of activator along the lines marked in a. The profiles correspond to times when the activator reaches its maximum at a given line

of the activator $u(x, y)$ and inhibitor $v(x, y)$ along the horizontal axis of the stripe at a selected moment of time. The peak of the inhibitor follows activator maximum and is responsible for the refractory character of the region behind the pulse. Figure 3d illustrates the profile of an activator along the line perpendicular to the stripe axis. The concentration of the activator reaches its maximum in the stripe center and rapidly decreases at the boundary between the excitable and non-excitable areas. Therefore, the width of the excitable channel can be used as a parameter that controls the maximum concentration of an activator in a propagating pulse.

Figure 4 shows profiles of a concentration in an excitable channel with a triangular shape. The two curves on Fig. 4b illustrate the profile of u along the lines 1 and 2, respectively measured at the time when the concentration of u on a given line reaches its maximum. It can be seen that the maximum concentration of the activator decreases when a pulse propagates towards the tip of the triangle. This effect can be used to build a chemical signal diode. Let us consider two pieces of excitable medium, one of triangular shape and another rectangular, separated by a non-excitable gap as shown on Fig. 4a. It is expected that a perturbation of the rectangular area by a pulse propagating towards the tip of the triangular one is much smaller than the perturbation of the triangular area by a pulse propagating towards the end of rectangular channel. Using this effect, a chemical signal diode that transmits pulses only in one direction can be constructed just by selecting the right width of non-excitable medium separating two pieces of excitable medium: a triangular one and a rectangular one [5,40].

The idea of a chemical signal diode presented in Fig. 4a was, in some sense, generalized by Davydov et al. [46], who considered pulses of excitation propagating on a 2-dimensional surface in 3-dimensional space. It has been shown that the propagation of spikes on surfaces with rapidly changing curvature can be unidirectional. For example, such an effect occurs when an excitation propagates on the surface of a tube with a variable diameter. In such a case, a spike moving from a segment characterized by a small diameter towards a larger one is stopped.

For both of the chemical signal diodes mentioned above, the excitability of the medium is a non trivial function of some space variables. However, a signal diode can be also constructed when the excitability of the medium changes in one direction only, so in a properly selected coordinate system it is a function of a single space variable. For example, the diode behavior in a system where the excitability level is a triangular function of a single space variable has been confirmed in numerical simulations based on the Oregonator model of a photosensitive, Ru-catalyzed BZ reaction (Eqs. (1)–(3)) [76]. If the catalyst is immobilized, then pulses that enter the region of inhomogeneous illumination from the strongly illuminated site are not transmitted, whereas the pulses propagating in the other direction can pass through (see Fig. 5a). A similar diode-like behavior resulting from a triangular profile of medium excitability can be expected for oxidation of CO on a Pt surface. Calculations demonstrate that a triangular

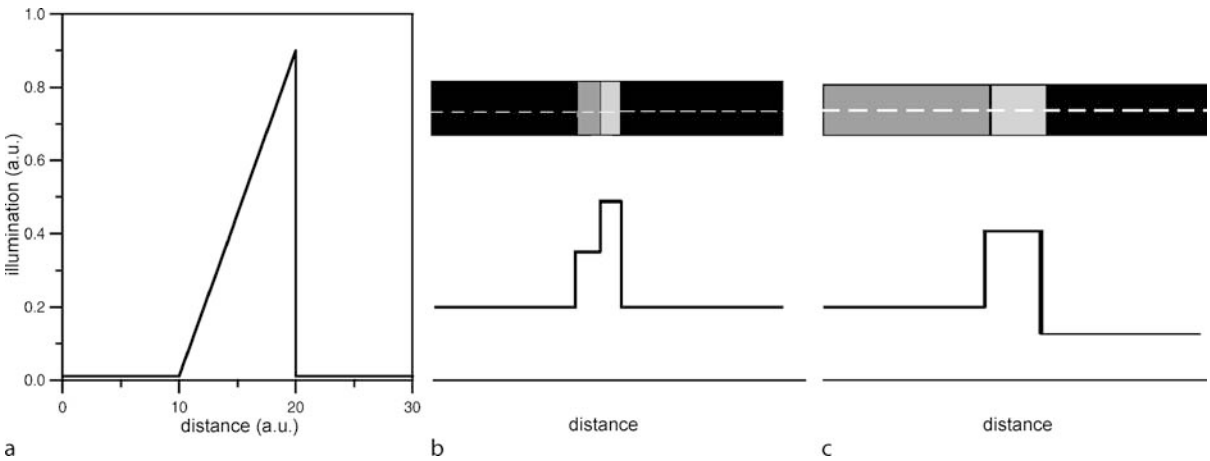

Computing in Geometrical Constrained Excitable Chemical Systems, Figure 5

The excitability as a function of a space variable in different one-dimensional realizations of a signal diode with BZ- reaction inhibited by light: a a triangular profile of illumination, b the illumination profile in a diode composed of two non-excitable barriers, c the illumination profile in a single barrier diode with nonsymmetrical excitable inputs. The *upper graphs* in b and c illustrate illumination on a membrane used in experiments [25]

profile of temperature (in this case, reduced with respect to that which characterizes the excitable medium) allows for the unidirectional transmission of spikes characterized by a high surface oxygen concentration [23]. However, a realization of the chemical signal diode can be simplified. If the properties of excitable channels on both sites of a diode are the same, then the diode can be constructed with just two stripes of non-excitable medium characterized by different excitabilities as illustrated in Fig. 5b. If the symmetry is broken at the level of input channels then the construction of a signal diode can be yet simpler, and it reduces to a single narrow non-excitable gap with a much lower excitability than that of the neighboring channels (cf. Fig. 5c). In both cases the numerical simulations based on the Oregonator model have shown that a diode works. The predictions of simulations have been qualitatively confirmed by experimental results [25]. Different realizations of a signal diode show that even for very simple signal processing devices the corresponding geometrical structure of excitable and non-excitable regions is not unique. Alternative constructions of chemical information processing devices seem important because they tell us on the minimum conditions necessary to build a device that performs a given function. In this respect a diode built with a single non-excitable region looks interesting, because such situation may occur at a cellular level, where the conditions inside the cell are different from those around. The diode behavior in the geometry shown on Fig. 5c indicates that the unidirectional propagation of spikes can be forced by a channel in a membrane, transparent to molecules or ions responsible for signal transmission.

Wave-number-dependent transmission through a non-excitable barrier is another feature of a chemical excitable medium important for information processing [48]. Let us consider two excitable areas separated by a stripe of non-excitable medium and a pulse of excitation propagating in one of those areas. The perturbation of the area behind the stripe introduced by an arriving pulse depends on the direction of its propagation. If the pulse wavevector is parallel to the stripe then the perturbation is smaller than in the case when it arrives perpendicularly [48]. Therefore, the width of the stripe can be selected such that pulses propagating perpendicularly to the stripe can cross it, whereas a pulse propagating along the stripe do not excite the area on the other side. This feature is frequently used to arrange the geometry of excitable channels such that pulses arriving from one channel do not excite the other.

Non-excitable barriers in a structured medium can play a more complex role than that described above. The problem of barrier crossing by a periodic train of pulses can be seen as an excitation via a periodic perturbation of the medium. It has been studied in detail in [13]. The answer of the medium is quite characteristic in the form of a devil-staircase-like firing number as a function of perturbation strength. In the case of barrier crossing, the strength of the excitation behind a barrier generated by an arriving pulse depends on the character of the non-excitable medium, the barrier width, and the frequency of the incoming signal (usually, due to an uncompleted relaxation of the medium, the amplitude of spikes decreases with frequency). A typical complex frequency transformation

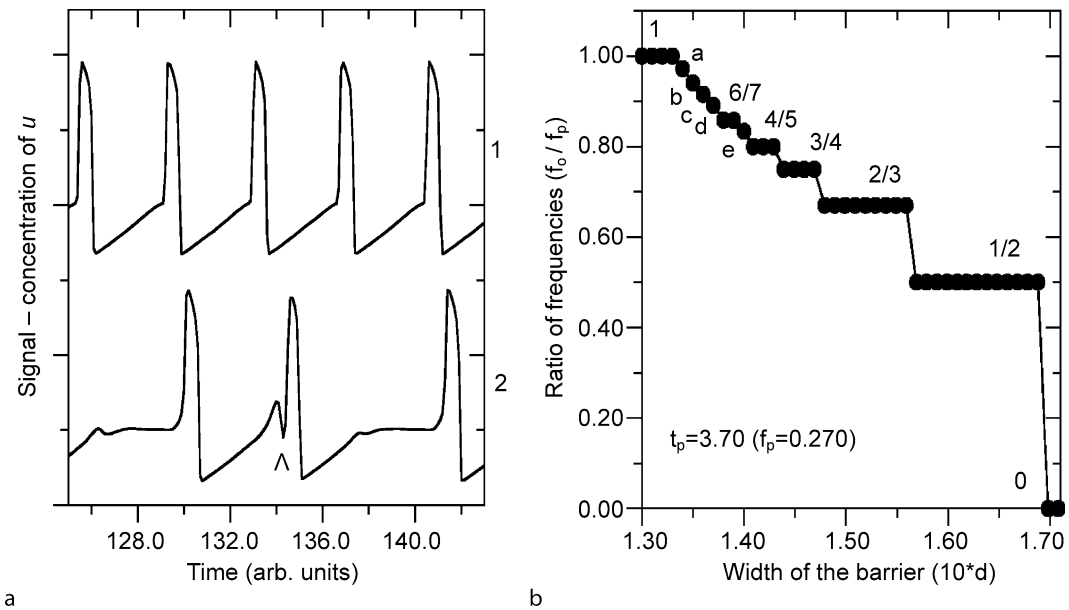

Computing in Geometrical Constrained Excitable Chemical Systems, Figure 6

Frequency transformation on a barrier – the results for the FitzHugh–Nagumo model [64]. a The comparison of an arriving train of pulses (1) and the transmitted signal (2). b A typical dependence of the firing number as a function of barrier width. The plateaus are labeled with corresponding values of firing number (1, 6/7, 4/5, 3/4, 2/3, 1/2 and 0). At points labeled a–e the following values have been observed: a – 35/36; b – 14/15; c – 10/11; d – 8/9 and e – 5/6

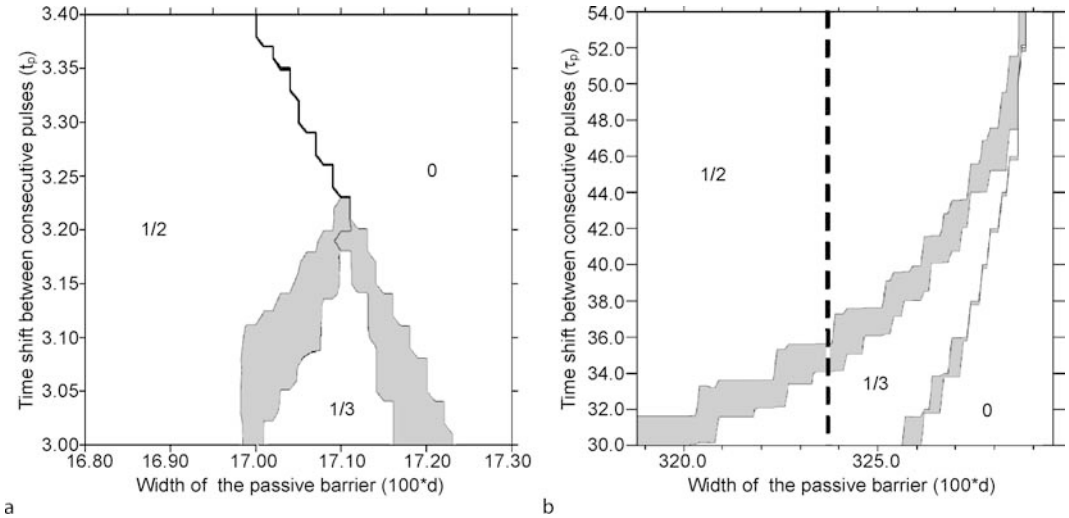

Computing in Geometrical Constrained Excitable Chemical Systems, Figure 7

The firing number as a function of the barrier width *d* and the interval of time between consecutive pulses (t_p). Labels in the *white* areas give the firing number, and the *gray* color marks values of parameters where more complicated transformations of frequency occur. a Results calculated for the FitzHugh–Nagumo model, b the Rovinsky–Zhabotinsky model. Both space and time are in dimensionless units

after barrier crossing is illustrated in Fig. 6. Experimental and numerical studies on firing number of a transmitted signal have been published [10,64,72,74]. It is interesting that the shape of regions characterized by the same firing number in the space of two parameters, barrier width and signal frequency, is not generic and depends on the type of the excitable medium. Figure 7 compares the firing numbers obtained for FitzHugh–Nagumo and Rovinsky–

Zhabotinsky models. In the first case, trains of pulses with small periods can cross wider barriers than trains characterized by low frequency; for the second model the dependence is reversed.

Logic Gates, Coincidence Detectors and Signal Filters

The simplest application of excitable media in information processing is based on the assumption that the logical FALSE and TRUE variables are represented by the rest state and by the presence of an excitation pulse at a given point of the system respectively. Within such interpretation a pulse represents a bit of information propagating in space. When the system remains in its rest state, no information is recorded or processed, which looks plausible for biological interpretation. Information coded in excitation pulses is processed in regions of space where pulses interact (via collision and subsequent annihilation or via transient local change in the properties of the medium). In this section we demonstrate that the geometry of excitable channels and non-excitable gaps can be tortured (the authors are grateful to prof. S. Stepney for this expression) to the level at which the system starts to perform the simplest logic operations on pulses. The binary chemical logic gates can be used as building blocks for devices performing more complex signal processing operations. Information processing with structured excitable media is "unconventional" because it is performed without a clock that sequences the operations, as it is in the standard von Neumann type computer architecture. On the other hand, in the signal processing devices described below, the proper timing of signals is important and this is achieved by selecting the right length and geometry of channels. In some cases, for example for the operations on trains of pulses, the presence of a reference signal, which plays a role similar to a clock, would significantly help to process information [70]. Historically, the logical gates were the first devices that have been realized with a structured chemical medium [1,2,14,65,73,75].

The setup of channels that execute the logic sum (OR) operation is illustrated in Fig. 8 [48]. The gate is composed of three excitable stripes (marked gray) surrounded by a non-excitable medium. The gaps separating the stripes have been selected such that a pulse that arrives perpendicularly to the gap can excite the area on the other side and a pulse that propagates parallel to the gap does not generate sufficient perturbation to excite the medium behind the gap. If there is no input pulse, the OR gate remains in the rest state and does not produce any output. A pulse in any of the input channels I1 and I2 can cross the gap separating these channels from the output O and an output spike appears. The excitation of the output channel generated by a pulse from one of the input channels propagates parallel to the gap separating the other input channel so it does not interfere with the other input as seen in Fig. 8b and c. The frequency at which the described OR gate operates is limited by the refractory period of the output medium. If the signals from both input channels arrive, but the time difference between pulses is smaller than the refractory time, then only the first pulse will produce the output spike.

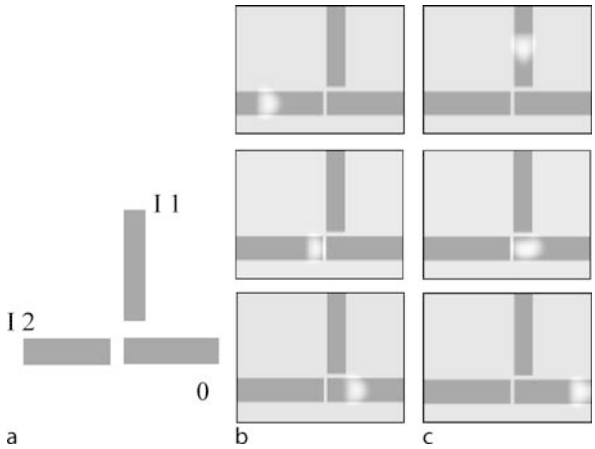

Computing in Geometrical Constrained Excitable Chemical Systems, Figure 8
The distribution of excitable channels (*dark*) that form the OR gate. b and c illustrate the time evolution of a single pulse arriving from inputs I1 and I2, respectively

The gate that returns the logic product (AND) of input signals is illustrated in Fig. 9. The width of the gap separating the output channel O from the input one (I1,I2) is selected such that a single excitation propagating in the input channel does not excite the output (Fig. 9b). However, if two counterpropagating pulses meet, then the resulting perturbation is sufficiently strong to generate an excitation in the output channel (Fig. 9c). Therefore, the output signal appears only when both sites of the input channel have been excited and pulses of excitation collided in front of the output channel. The width of the output channel defines the time difference between input pulses treated as simultaneous. Therefore, the structure shown in Fig. 9 can also be used as a detector of time coincidence between pulses.

The design of the negation gate is shown in Fig. 10. The gaps separating the input channels, the source channel and the output channel can be crossed by a pulse that propagates perpendicularly, but they are non-penetrable for pulses propagating parallel to the gaps. The NOT gate

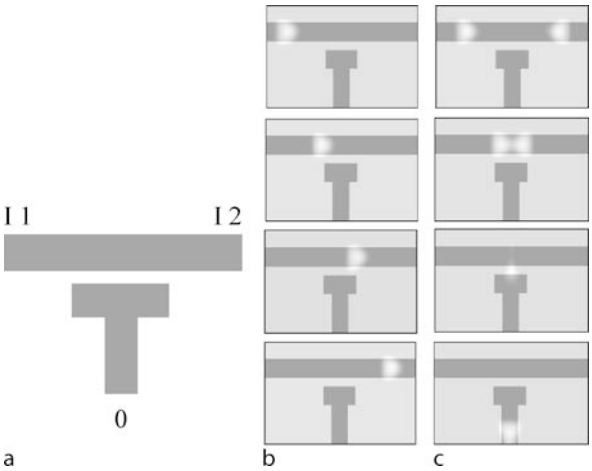

Computing in Geometrical Constrained Excitable Chemical Systems, Figure 9
The distribution of excitable channels (*dark*) that form the AND gate. b and c illustrate the response of the gate to a single pulse arriving from input I1 and to a pair of pulses, respectively

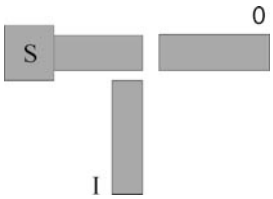

Computing in Geometrical Constrained Excitable Chemical Systems, Figure 10
The distribution of excitable channels (*dark*) that form the NOT gate

should deliver an output signal if the input is in the rest state. Therefore, it should contain a source of excitation pulses (marked S). If the input is in the rest state then pulses from the source propagate unperturbed and enter the output channel. If there is an excitation pulse in the input channel then it enters the channel linking the source with the output and annihilates with one of pulses generated by the source. As a result no output pulse appears. At the first look, the described NOT gate works fine, but if we assume that a single pulse is used in information coding than the input pulse should arrive at the right time to block the source. Therefore, additional synchronization of the source is required. If information is coded in trains of pulses then the frequency of the source should match with the one used for coding.

The structure of excitable channels for the exclusive OR (XOR) gate is illustrated in Fig. 11c [34]. Two input channels bring signals to the central area C. The output

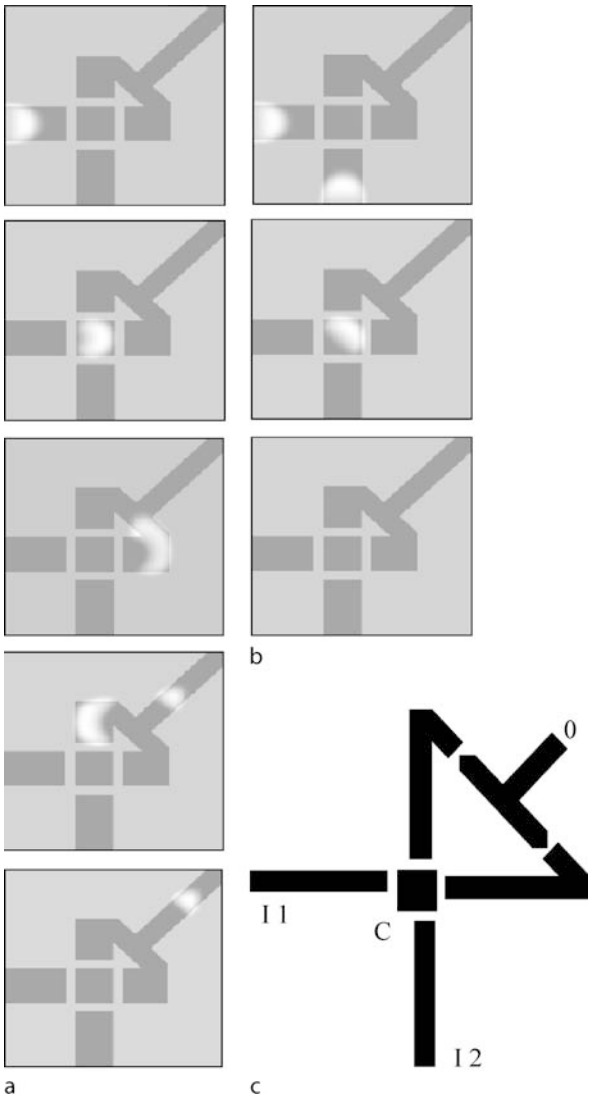

Computing in Geometrical Constrained Excitable Chemical Systems, Figure 11
c the distribution of excitable channels (*dark*) that form the XOR gate. a and b illustrate the response of the gate to a single pulse arriving from input I1 and to a pair of pulses, respectively

channels are linked together via diodes (Fig. 4) that stop possible backward propagation. As in the previous cases only pulses perpendicular to the gaps can pass through non-excitable gaps between the central area and both input and output channels. The shape of the central area has been designed such that an excitation generated by a single input pulse propagates parallel to one of the outputs and is perpendicular to another. As the result one of the output channels is excited (see Fig. 11a). However, if pulses from both input channels arrive at the same time then the

wavevector of the excitation in the central part is always parallel to the boundaries (Fig. 11b). Therefore, no output signal appears. Of course, there is no output signal if none of the inputs are excited. It is worth noticing that for some geometries of the XOR gate the diodes in output channels are not necessary because the backward propagation does not produce a pulse with a wavevector perpendicular to a gap as seen in the fourth frame of Fig. 11a.

Another interesting example of behavior resulting from the interaction of pulses has been observed in a cross-shaped structure built of excitable regions, separated by gaps penetrable for perpendicular pulses [65,66] shown in Fig. 12. The answer of cross-shaped junction to a pair pulses arriving from two perpendicular directions has been studied as a function of the time difference between pulses. Of course, if the time difference is large, pulses propagate independently along their channels. If the time difference is small the cross-junction acts like the AND gate and the output excitation appears in one of the corner areas. However, for a certain time difference the first arriving pulse is able to redirect the other and force it to follow. The effect is related to uncompleted relaxation of the central area of the junction at the moment when the second pulse arrives. Pulse redirection seems to be an interesting effect from the point of programming with excitation pulses, but in practice it requires a high precision in selecting the right time difference.

The logic gates described above can also be applied to transform signals composed of many spikes. For example, two trains of pulses can be added together if they pass through an OR gate. The AND gate creates a signal composed of coinciding pulses from both trains. It is also easy to use a structured excitable medium and generate a signal that does not contain coinciding pulses [50]. The structure of the corresponding device is illustrated in Fig. 13. All non-excitable gaps are penetrable for perpendicular pulses. If a pulse of excitation arrives from one of the input channels then it excites the segment A and propagates towards the other end of it. If there is no spike in the other input channel then this excitation propagates unperturbed and activates the output channel. However, if a pulse from the other channel arrives it annihilates with the original

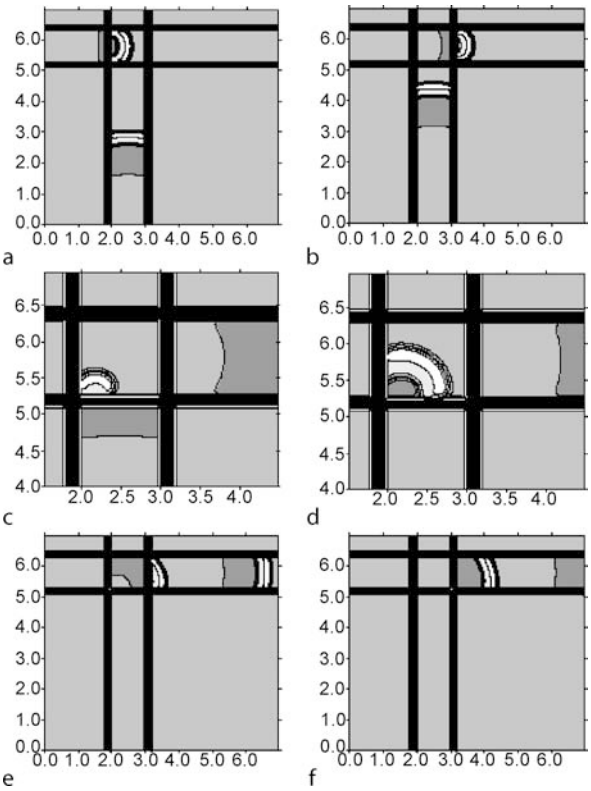

Computing in Geometrical Constrained Excitable Chemical Systems, Figure 12

The distribution of excitable and non-excitable regions in a cross-shaped junction. Here the excitable regions are *gray* and the non-excitable *black*. Consecutive figures illustrate an interesting type of time evolution caused by interaction of pulses. Two central figures are enlarged in order to show how uncompleted relaxation influences the shape of the next pulse

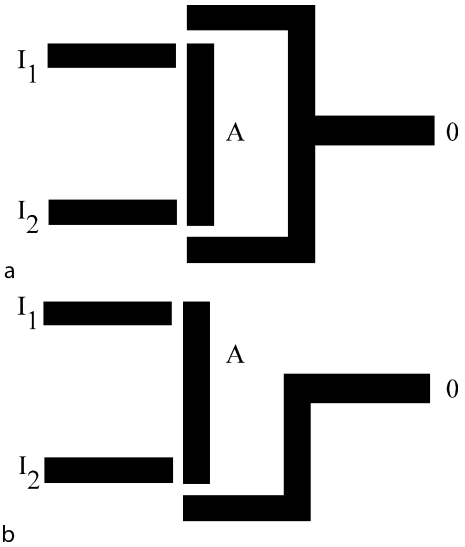

Computing in Geometrical Constrained Excitable Chemical Systems, Figure 13

The distribution of excitable channels in devices that compare trains of pulses. The device shown in a applies the XOR operation to a pair of signals and makes a signal composed of spikes that do not coincide with the other signal. b generates a signal composed of spikes that arrive through I1 and do not coincide with excitation pulses coming from I2

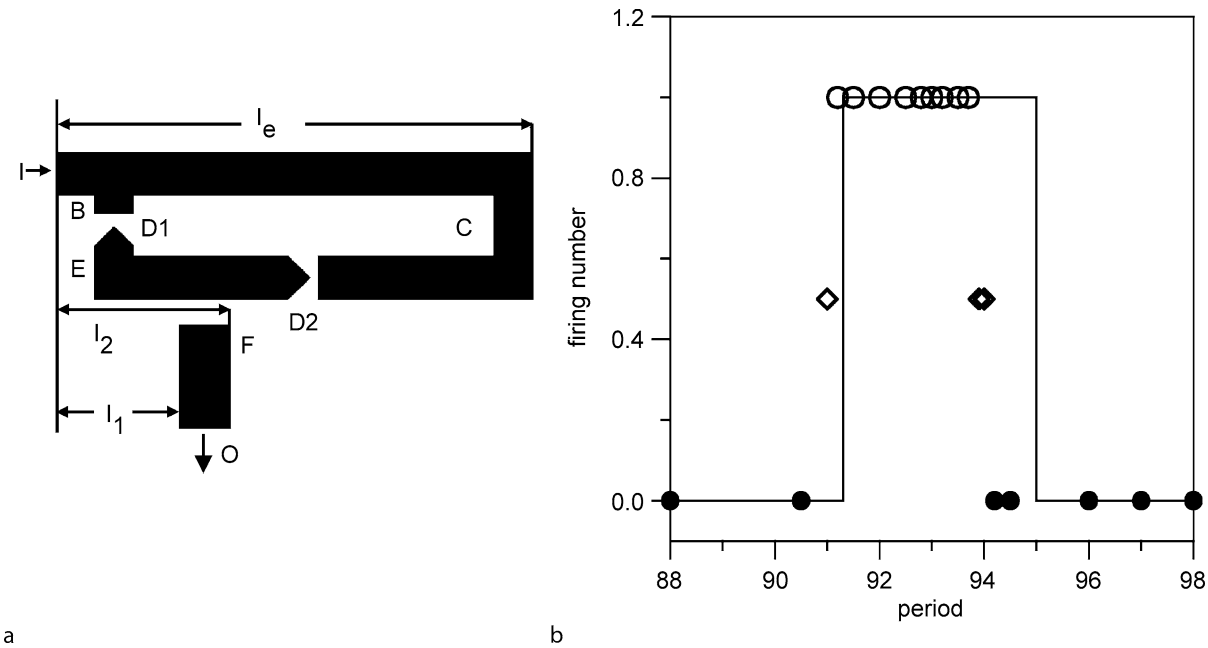

Computing in Geometrical Constrained Excitable Chemical Systems, Figure 14
a The distribution of excitable and non-excitable regions in a band filter. b Typical characteristics of the band filter [22]. The *black dots* mark periods for which the signal is not transmitted, the *empty* ones indicate full transmission, the *empty diamonds* mark the periods of the arriving signal for which every second spike is transmitted

pulse and no output signal is generated. The structure illustrated on Fig. 13a can be easily transformed into a device that compares two trains of pulses such that the resulting signal is composed of spikes of the first signal that do not coincide with the pulses of the second train. Such a device can be constructed just by neglecting one of the output channels. Figure 13b illustrates the geometry of the device that produces the output signal composed of pulses in arriving at the input I1 that have no corresponding spikes in the signal arriving from I2.

The coincidence detector can be used as a frequency filter that transmits periodic signals within a certain frequency interval [22,50]. The idea of such a filter is shown in Fig. 14. The device tests if the time between subsequent spikes of the train remains in the assumed range. As illustrated the signal arriving from the input I separates and enters the segment E through diodes D1 and D2. The segments E and F form a coincidence detector (or an AND gate, c. f. Fig. 9). The excitation of the output appears when an excitation coming to E via the segment C and D2 is in the coincidence with the subsequent spike of the train that arrived directly via D1. The time shift for which coincidences are tested is decided by the difference in lengths of both paths. The time resolution depends on the width of the F channel (here $l_2 - l_1$). For periodic signals the presented structure works as a frequency filter and transmits signals within the frequency range $f_{\pm} = v/(\Delta r \pm (\Delta w)/2)$, where Δr is the difference of distances traveled by spikes calculated to the point in E placed above the center of F channels ($\sim 2 * l_e$), Δw is the width of the output channel, and v is the velocity of a spike in the medium. Typical characteristics of the filter are illustrated on Fig. 14b. The points represent results of numerical simulations and the line shows the filter characteristics calculated from the equation given above. It can be noticed that at the ends of the transmitted frequency band a change in output signal frequency is observed. This unwelcome effect can be easily avoided if a sequence of two identical filters is used. A filter tuned to a certain frequency will also pass any of its harmonics because the output signal can be generated by the coincidence of every second, third, etc. pulses in the train.

Chemical Sensors Built with Structured Excitable Media

Even the simplest organisms without specialized nerve systems or brains are able to search for the optimum living conditions and resources of food. We should not be astonished by this fact because even chemical systems can receive and process information arriving from their neighborhood. In this section we describe how simple struc-

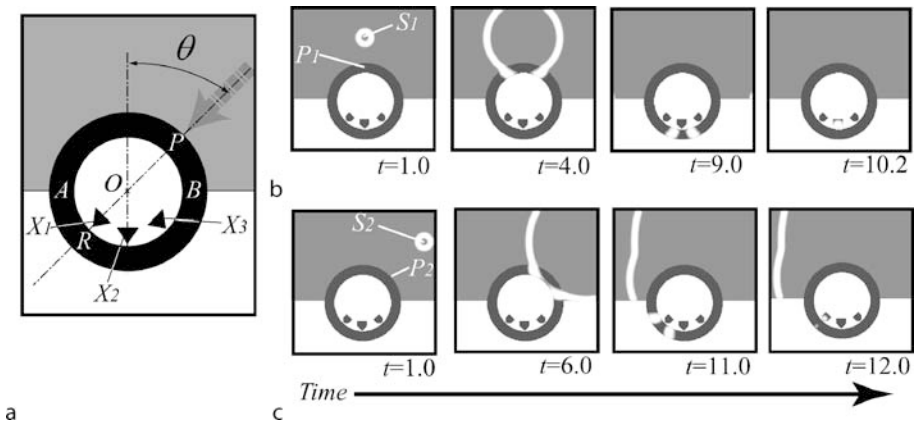

Computing in Geometrical Constrained Excitable Chemical Systems, Figure 15
The distribution of excitable areas (*gray* and *black*) and non-excitable regions (*white*) in the ring-shaped direction detector. *Triangles* marked X_1, X_2 and X_3 show the output channels. b and c illustrate time evolution of pulses generated from sources at different locations. The excited output channel depends on the direction of the source (copied from [53] with the permission of the authors)

tures built of excitable media can be used for direction and distance sensing. In order to simplify the conversion between an external stimulus and information processed by a sensor, we assume that the environment is represented by the same excitable medium as the sensor itself, so a stimulus can directly enter the sensor and be processed.

One possible strategy of sensing is based on a one-to-one relationship between the measured variable and the activated sensor channel [53]. An example of a sensor of that type is illustrated on Fig. 15. It is constructed with a highly excitable, black ring surrounded by a number of coincidence detectors, denoted as X_1, X_2 and X_3. The excitation of a detector appears if a pair of pulses collide on the ring in front of it. Let us consider a homogeneous excitable environment and a spherical pulse of excitation with the source at the point S_1. High excitability of the ring means that excitations on the ring propagate faster than those in the surrounding environment. At a certain moment the pulse arrives at the point P_1, which lies on a line connecting the S_1 and the center of the ring (see Fig. 15b). The arriving pulse creates an excitation on the ring originating from the point P_1. This excitation splits into a pair of pulses rotating in opposite directions and after propagating around the ring, they collide at the point, which is symmetric to P with respect to the center of the ring O. The point of collision can be located by an array of coincidence detectors and thus we have information on the wave vector of the arriving pulse. In this method the resolution depends on the number of detectors used because each of them corresponds to a certain range of the measured wave vectors. The fact that a pulse has appeared in a given output channel implies that no other channel of the sensor gets excited. It is interesting that the output information is reversed in space if compared with the input one; the left sensor channels are excited when an excitation arrives from the right and vice versa.

The geometry of the direction sensor that sends information as an excitation pulse in one of its detector channels can be yet simplified. Two such realizations are illustrated in Fig. 16. The excitable areas that form the sensor are marked black, the excitable surrounding medium where stimuli propagate is gray, and the non-excitable areas are white. Let us consider an excitation generated in the medium M by a source S. It is obvious that if the source is located just above the sensor then pulses of excitation are generated at the ends of the sensor channel D at the same time and they finally annihilate above the central coincidence channel and generate an output pulse in it. If the source is located at a certain angle with respect to the vertical line then the annihilation point is shifted off the center of D. We have performed a series of numerical simulations to find the relation between the position of the source and the position of annihilation point in D. In Fig. 16c and 16d the position of the annihilation point is plotted as a function of the angle between the source position and the vertical line. Although the constructions of sensors seem to be very similar, the working range of angles is larger for the sensor shown on Fig. 16b whereas the sensor shown in Fig. 16a offers slightly better resolution. Using information from two detectors of direction, the position of the source of excitation can be easily located because it is placed on the intersection of the lines representing the detected directions [77].

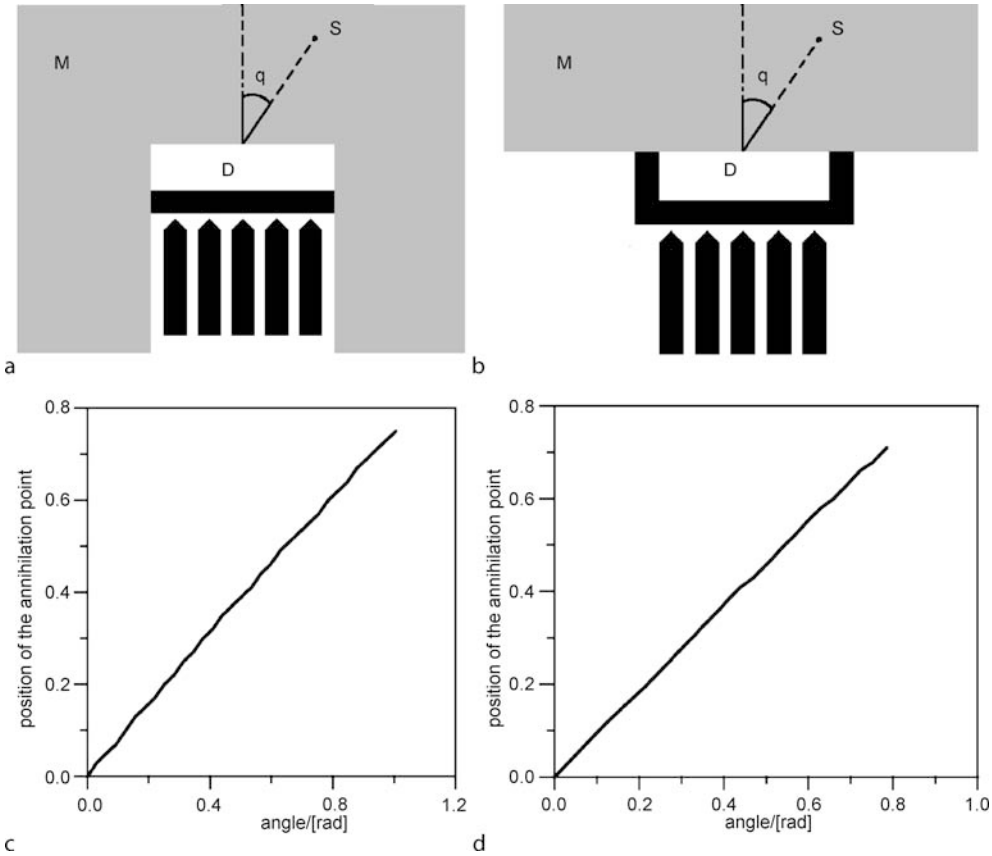

Computing in Geometrical Constrained Excitable Chemical Systems, Figure 16

The geometry of excitable (*black*) and diffusive (*white*) areas in two realizations of simplified distance sensors. The *gray* color marks the excitable medium around the sensor, *S* marks the position of excitation source. c and d show the position of the annihilation point in the channel D as a function of the angle *q* between the source position and the vertical line. The half-length of D is used as the scale unit

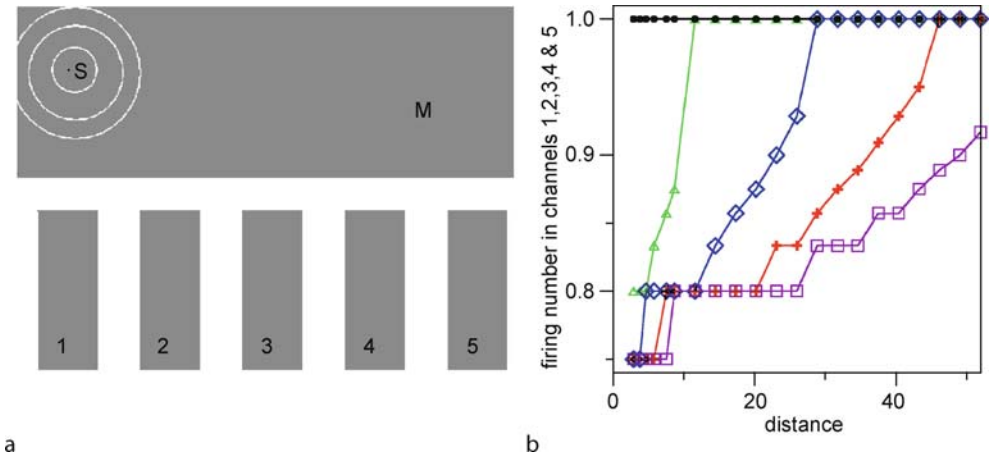

Computing in Geometrical Constrained Excitable Chemical Systems, Figure 17

The distance sensor based on frequency transformation. a The distribution of excitable channels, b the firing numbers in different channels as a function of distance; *black, green, blue red and violet curves* correspond to signals in channels 1–5, respectively

Another strategy of sensing is based on the observation that frequencies of pulses excited in a set of sensor channels by an external periodic perturbation contain information on the location of a stimulus. Within this strategy there is no direct relationship between the number of sensor channels and sensor resolution, and, as we show below, a sensor with a relatively small number of channels can quite precisely estimate the distance separating it from the excitation source. As we have mentioned in the Introduction, the frequency of a chemical signal can change after propagating through a barrier made of non-excitable medium. For a given frequency of arriving spikes the frequency of excitations behind the barrier depends on the barrier width and on the angle between the normal angle to the barrier and the wave vector of arriving pulses. This effect can be used for sensing. The geometrical arrangement of excitable and non-excitable areas in a distance sensor is shown in Fig. 17. The excitable signal channels (in Fig. 17a they are numbered 1–5) are wide enough to ensure stable propagation of spikes. They are separated from one another by parallel non-excitable gaps that do not allow for interference between pulses propagating in the neighboring channels. The sensor channels are separated from the excitable medium M by the non-excitable sensor gap G. The width of this gap is very important. If the gap is too wide then no excitation of the medium M can generate a pulse in the sensor channels. If the gap is narrow then any excitation in front of the sensor can pass G and create a spike in each sensor channel so the signals sent out by the sensor channels are identical. However, there is a range of gap widths such that the firing number depends on the wave vector characterizing a pulse at the gap in front of the channel. If the source S is close to the array of sensor channels, then the wave vectors characterizing excitations in front of various channels are significantly different. Thus the frequencies of excitations in various channels should differ too. On the other hand, if the source of excitations is far away from the gap G then the wave vectors in front of different channels should be almost identical and the frequencies of excitations should be the same. Therefore, the system illustrated in Fig. 17a can sense the distance separating it from the source of excitations. If this distance is small then the firing numbers in neighboring sensor channels are different and these differences decrease when the source of excitations moves away. A typical distance dependence of firing numbers observed in different channels is illustrated in Fig. 17b. This result has been obtained in numerical simulations based on the Oregonator model.

The range of sensed distances depends on the number of sensor channels. A similar sensor, but with 4 sensor channels, was studied in [28]. Comparing the results, we observe that the presence of 5th channel significantly improves the range of distances for which the sensor operates. On the other hand, the additional channel has almost no effect on sensor resolution at short distances. The sensor accuracy seems to be a complex function related to the width of the sensor gap and the properties of channels. The firing number of a single channel as a function of the distance between the sensor and the source has a devil-staircase-like form with long intervals where the function is constant corresponding to simple fractions as illustrated in Fig. 5b. In some range of distances the steps in firing numbers of different channels can coincide, so the resolution in this range of distances is poor. For the other regions, the firing numbers change rapidly and even small changes in distance can be easily detected.

The signal transformation on a barrier depends on the frequency of incoming pulses (c. f. Fig. 6) so the distance sensor that works well for one frequency may not work for another. In order to function properly for different stimuli the sensor has to be adapted to the conditions it operates. In practice such adaptation can be realized by the comparison of frequencies of signals in detector channels with the frequency in a control channel. For example the control channel can be separated from the medium M by a barrier so narrow that every excitation of the medium generates a spike. The comparison between the frequency of exci-

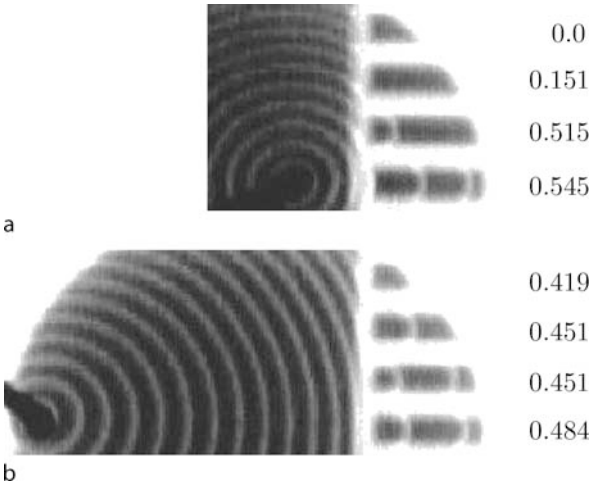

Computing in Geometrical Constrained Excitable Chemical Systems, Figure 18

Two snapshots from the experimental realization of the distance sensor with four channels. In the *upper* figure the source (1 mm thick silver wire) is placed 2 mm away from the sensor; in the *bottom* one the source is 12 mm away. The firing numbers are given next to the corresponding channels

tations in the control channel and in the sensor channels can be used to adjust the sensor gap G. If the frequency of excitations in the sensor channels is the same as in the control channels then the width of the gap should be increased or its excitability level decreased. On the other hand if the frequency in the sensor channel is much smaller than in the control channel (or null) then the gap should be more narrow or more excitable. Such an adaptation mechanism allows one to adjust the distance detector to any frequency of arriving excitations.

The fact that the distance detector described above actually works has been confirmed in experiments with a photosensitive Ru-catalyzed BZ reaction. Typical snapshots from two experiments performed for the source placed 2 and 12 mm away from the sensors are shown in Fig. 18. The firing numbers observed in different sensor channels confirm qualitatively the predictions of numerical simulations. If the source of excitations is close to the sensor gap, then the differences between firing numbers observed in neighboring channels are large. On the other hand, when the source of excitations is far away from the sensor, the frequencies in different channels become similar. The range of distances at which the sensor works is measured in centimeters so it is of the same order as the sensor size.

The Ring Memory and Its Applications

The devices discussed in the previous section can be classified as instant machines [58] capable of performing just the task they have been designed for. A memory where information coded in excitation pulses can be written-in, kept, read-out and, if necessary, erased, significantly increases the information processing potential of structured excitable media. Moreover, due to the fact that the state of memory can be changed by a spike, the memory allows for programming with excitation pulses. One possible realization of a chemical memory is based on the observation that a pulse of excitation can rotate on a ring-shaped excitable area as long as the reactants are supplied and the products removed [41,52,55]. Therefore, a ring with a number of spikes rotating on it can be regarded as a loaded memory cell. Such memory can be erased by counterpropagating pulses. The idea of memory with loading pulses rotating in one direction and erasing pulses in another has been discussed in [49]. If the ring is big then it can be used to memorize a large amount of information because it has many states corresponding to different numbers of rotating pulses. However, in such cases, loading the memory with subsequent pulses may not be reliable because the input can be blocked by the refractory tail left by one of al-

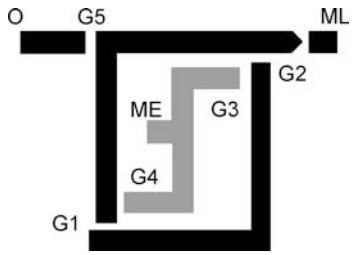

Computing in Geometrical Constrained Excitable Chemical Systems, Figure 19
The distribution of excitable channels (*dark*) that form a memory cell. The cell is composed of the loading channel ML, the memory ring, the erasing channel ME (marked *gray*) and the output channel O

ready rotating pulses. The same effect can block the erasing pulses. Therefore, the memory capable of storing just a single bit seems to be more reliable and we consider it in this section. Such memory has two states: if there is a rotating pulse the ring is in the logical TRUE state (we call it loaded); if there is no pulse the state of memory corresponds to the logical FALSE and we call such memory erased.

Let us consider the memory illustrated in Fig. 19. The black areas define the memory ring, the output channel O, and the loading channel ML. The memory ring is formed by two L-shaped excitable areas. The areas are separated by gaps and, as we show below, with a special choice of the gaps the symmetry of the ring is broken and unidirectional rotation ensured. The Z-shaped excitable area composed of gray segments inside the ring forms the erasing channel. The widths of all non-excitable gaps separating excitable areas are selected such that a pulse of excitation propagating perpendicularly to the gap excites the active area on the other site of the gap, but the gap is impenetrable for pulses propagating parallel to the gap. The memory cell can be loaded by a spike arriving from the ML channel. Such a spike crosses the gap and generates an excitation on the ring rotating counterclockwise. The information about loaded memory is periodically sent out as a series of spikes through the output channel O. The rotating pulse does not affect the erasing channel because it always propagates parallel to it. The erasing excitation is generated in the center of the Z-shaped area and it splits into two erasing pulses. These pulses can cross the gaps separating the erasing channel from the memory ring and create a pair of pulses rotating clockwise. A spike that propagates clockwise on the memory ring is not stable because it is not able to cross any of the gaps and dies. It also does not produce any output signal. Therefore, if the memory has not been loaded then an erasing excitation does not load it. On the

other hand, if the memory is loaded then clockwise rotating pulses resulting from the excitation of the erasing channel annihilate with the loading pulse and the memory is erased. The idea of using two places where erasing pulses can enter the memory ring is used to ensure that at least one of those places is fully relaxed and so one of erasing pulses can always enter the ring.

In order to verify that such memory works, we have performed a number of simulations using the Rovinsky–Zhabotinsky model of a BZ reaction and have done the experiments. We considered a loaded memory and at a random time the excitation was generated in the middle of the erasing channel. In all cases, such an excitation erased the memory ring. A typical experimental result is illustrated in Fig. 20. Here the channels are 1.5 mm thick and the gaps are 0.1 mm wide. Figure 20a shows the loaded memory and initiated pair of pulses in the erasing channel. In Fig. 20b one of the pulses from the erasing channel enters the memory ring. The memory ring in front of the left part of erasing channel is still in the refractory state so the left erasing pulse has not produced an excitation on the ring. Finally in Fig. 20c we observe the annihilation of the loading pulse with one of erasing pulses. Therefore, the state of memory changed from loaded to unloaded. The experiment was repeated a few times and the results were in a qualitative agreement with the simulations: the loaded memory cell kept its information for a few minutes and it was erased after every excitation of the erasing channel.

Figure 21 illustrates two simple, yet interesting, applications of a memory cell. Figure 21a shows a switchable unidirectional channel that can be opened or closed depending on the state of memory [29]. The channel is constructed with three excitable segments A, B and C separated by signals diodes D1 and D2 (c. f. Fig. 4). The mid segment B is also linked with the output of the memory ring M. Here the erasing channels of M are placed outside the memory ring, but their function is exactly the same as in the memory illustrated in Fig. 19. The idea of switchable channel is similar to the construction of the NOT gate (Fig. 10). If the memory is not loaded then the spike propagation from input I to output O is unperturbed. However, if the memory is loaded then pulses of excitation periodically enter the segment B and annihilate with the transmitted signal. As a result, the channel is either open or blocked depending on the state of the memory. The excitations generated by the memory ring do not spread outside the B segment. On one end their propagation is stopped by the diode D1, on the other by the geometry of the junction between the B channel and the memory output. The state of memory is controlled by the excitation pulses coming from loading and erasing channels, so switchable channels can be used in devices that are programmable with excitation pulses. Figure 21b illustrates a self-erasing memory cell that changes its state from loaded to erased after a certain time. In such a memory cell, the output channel is connected with the erasing one. When the memory is loaded the output signal appears. After some time decided by the length of connecting channels an output pulse returns as an erasing pulse and switches the memory to

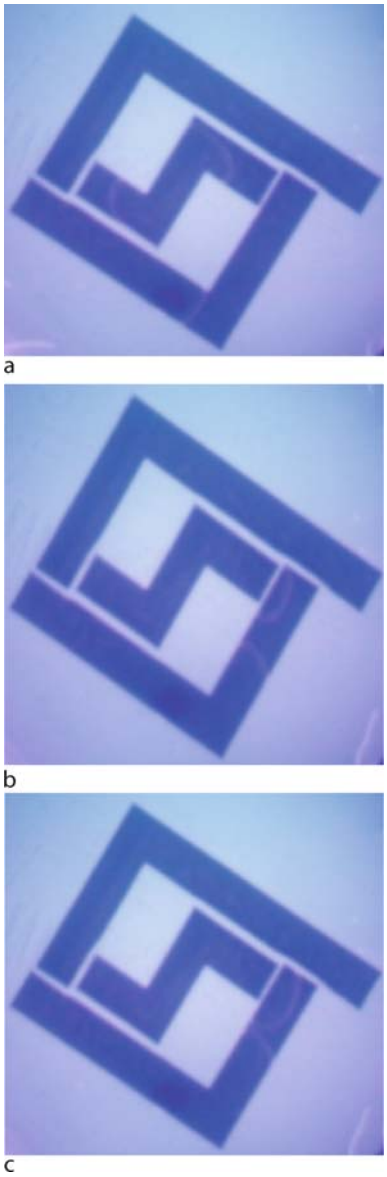

Computing in Geometrical Constrained Excitable Chemical Systems, Figure 20
Three snapshots from an experiment with memory erasing. The memory ring is formed by two L-shaped excitable channels, the Z-shaped erasing channel is inside the ring

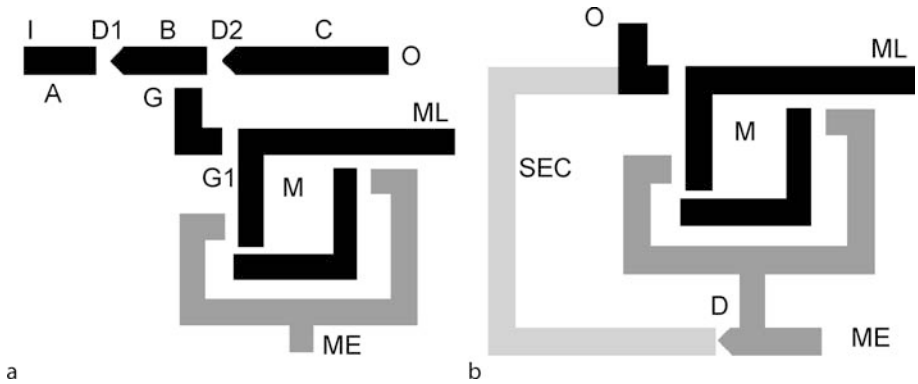

Computing in Geometrical Constrained Excitable Chemical Systems, Figure 21
Two simple applications of a memory cell. a A switchable unidirectional channel that stops or transmits signals depending on the state of memory. b A self-erasing memory cell that changes its state after a certain time. SEC marks the connection between the memory output and the erasing channel

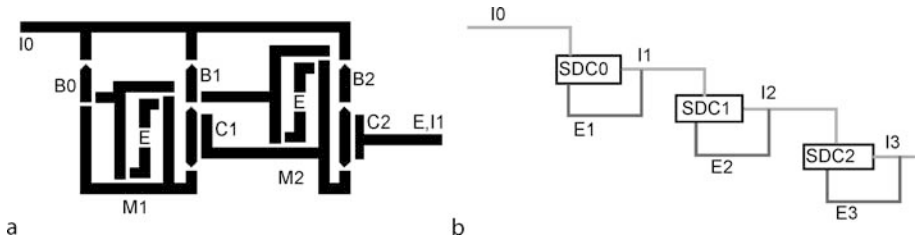

Computing in Geometrical Constrained Excitable Chemical Systems, Figure 22
The counter of excitation pulses that arrive at the input I0. a shows the geometry of excitable channels (*black*) in a single digit counter for the positional representation with the base 3. b is a schematic illustration of the cascade of single digit counters that provides a positional representation. The feedback signals from E1, E2 and E3 channels erase the memory of the single digit counters

unloaded state. This behavior is an example of a simple feedback process common in self regulating information processing systems.

Using memory cells, signal diodes, and coincidence detectors, one can construct devices which perform more complex signal processing operations. As an example, we present a simple chemical realization of a device that counts arriving spikes and returns their number in any chosen positional representation [27]. Such a counter can be assembled from single digit counters. The construction of a single digit counter depends on the representation used. Here, as an example, we consider the positional representation with the base 3. The geometry of a single digit counter is schematically shown in Fig. 22. Its main elements are two memory cells M_1 and M_2 and two coincidence detectors C_1 and C_2. At the beginning let us assume that none of the memory cells are loaded. When the first pulse arrives through the input channel I_0, it splits at all junctions and excitations enter segments B_0, B_1 and B_2. The pulse that has propagated through B_0 loads the memory cell M_1. The pulses that have propagated through B_1 and B_2 die at the bottom diodes of segments C_1 and C_2 respectively. Thus, the first input pulse loads the memory M_1 and does not change the state of M_2. When M_1 is loaded, pulses of excitation are periodically sent to segments B_0 and C_1 via the bottom channel. Now let us consider what happen when the second pulse arrives. It does not pass through B_0 because it annihilates with the pulses arriving from the memory M_1. The excitations generated by the second pulse can enter B_1 and B_2. The excitation that propagated through B_2 dies at the bottom diode of the segment C_2. The pulse that has propagated through B_1 enters C_1, annihilates with a pulse from memory M_1 and activates the coincidence detector. The output pulse from the coincidence detector loads the memory M_2. Therefore, after the second input pulse both memories M_1 and M_2 are loaded. If the third pulse arrives the segments B_0 and B_1 are blocked by spikes sent from the memory rings. The generated excitation can enter channel B_2 and its collision with a pulse coming from the memory cell M_2 activates the output channel of C_2. The output signal is directed to the counter of responsible

for the digit at next position (I1) and it is also used to erase all memory cells. Thus after the third pulse both memory cells M1 and M2 are erased. The counter shown in Fig. 22 returns a digit in a representation with the base 3: here 0 is represented by the $(M_1, M_2) = (0, 0)$, 1 by $(1, 0)$, 2 by $(1, 1)$ and the next pulse changes the state of memory cell into $(M_1, M_2) = (0, 0)$. Of course, using $n - 1$ memory cells in a single digit counter we can represent digits of the system with base n. A cascade of single digit counters (see Fig. 22b) gives a positional representation of the number of arriving pulses.

Artificial Chemical Neurons with Excitable Medium

In this section we discuss a simple realization of an artificial neuron with structured excitable medium and show how neuron networks can be used in programmable devices. We consider a chemical analogy of the McCulloch–Pitts neuron, i. e., a device that produces the output signal if the combined activation exceeds a critical value [32]. The geometry of a chemical neuron is inspired by a the structure of biological neuron [30]. One of its realizations with an excitable medium is illustrated on Fig. 23. Another geometry of a neuron has been discussed in [24]. In an artificial chemical neuron, like in real neurons, dendrites (input channels 1–4) transmit weak signals which are added together through the processes of spatial and temporal integration inside the cell body (part C). If the aggregate excitation is larger than the threshold value the cell body gets excited. This excitation is transmitted as an output signal down the axon (the output channel) and the amplitude of the output signal does not depend on the value of integrated inputs but only on the properties of the medium that makes the output channel. In Fig. 23a the axon is not shown and we assume that it is formed by an excitable channel located perpendicularly above the cell body. In the construction discussed in [24] both dendrites and the axon were on a single plane. We have studied the neuron using numerical simulations based on the Oregonator model and the reaction-diffusion equations have been solved on a square grid. The square shape of the neuron body and the input channels shown on Fig. 23a allows for precise definition of the boundary between the excitable and non-excitable parts. The idea behind the chemical neuron is similar to that of the AND gate: perturbations introduced by multiple inputs combine and generate a stronger excitation of the cell body than this resulting from a single input pulse. Therefore, it is intuitively clear that if we are able to adjust the amplitudes of excitations coming from individual input channels, then the output channel becomes excited only when the required number of excitations arrive from the inputs.

In the studied neuron, the amplitudes of spikes in input channels have been adjusted by sub-excitability of

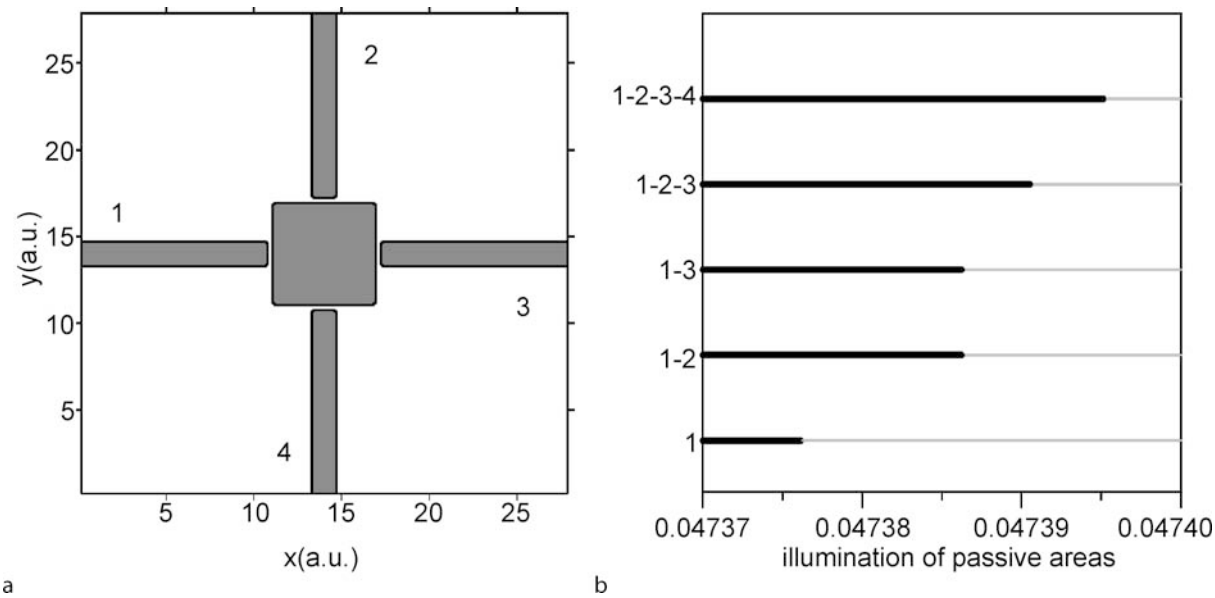

Computing in Geometrical Constrained Excitable Chemical Systems, Figure 23
Artificial chemical neuron. a The geometry of excitable and non-excitable areas; b The response of the neuron to different types of excitations as a function of the illumination of non-excitable regions. The numbers given on the *left list* the excited channels. The values of ϕ_p for which the neuron body gets excited are marked by a *thick line*

these channels which can be controlled by the channel width or by the illumination of surrounding non-excitable medium. In our simulations we considered different values of ϕ_p for non-excitable areas, whereas $\phi_a = 0.007$ characterizing excitable regions has been fixed for dendrites and the neuron body. Simulation results shown on Fig. 23b indicate that the properties of chemical neurons are very sensitive with respect to changes in ϕ_p. The thick line marks the values of ϕ_p for which the output signal appears. For the parameters used when ϕ_p is slightly below 0.047377 any single excitation produces an output signal. On the other hand if $\phi_p > 0.047397$ then even the combined excitation of all inputs is not sufficient to excite the neuron. In between those two limiting values we observe all other thresholds; i. e., an output excitation as the result of two or three combined inputs. Therefore, by applying the proper illumination level the structure shown in Fig. 23a can work as a four input McCulloch–Pitts neuron with the required threshold. A similar high sensitivity of the neuron properties on ϕ_a is also observed. It means that neuron properties can be controlled with tiny changes in system parameters.

The chemical neuron illustrated in Fig. 23a can be used to program signal processing with pulses of excitation. If we set the illuminations such that any two input pulses produce the output and use one of the inputs to control the device, then if the control pulse is present the device performs the OR operation on the other inputs. If there is no control pulse, it calculates an alternative on conjunctions of all pairs of channels.

The geometry of a network constructed with neurons can be easily controlled if the switchable channels illustrated on Fig. 21b are used to establish or cut connections between processing elements. The state of the memory that controls the channel may depend on the state of the network through a feedback mechanism what allows for network training. In the chemical programming described above pulses of concentration of the same reagent are used to store and process information, and to program the medium. Therefore, the output signal may be directly used to change the network geometry. External programming factors like illumination or a temperature field [3] are difficult for applications in three dimensional structures. The pulse based programming seems to be easier if the proper geometry of switchable channels and of the feedbacks is introduced.

At the first look it seems that a practical realization of a network built of chemical neurons may be difficult. However, the geometry of a considered neuron looks quite similar to structures of phases formed in a multicomponent system. For example, the diamond structure in an oil-water-surfactant system, which spontaneously appears at certain thermodynamic conditions [11], has a form of centers linked with the four nearest neighbors. If the reactants corresponding for excitability are soluble in water, but not in oil then the water rich phase forms the structure of excitable channels and processing elements just as the result of thermodynamic conditions. Within a certain range of parameters, such a structure is thermodynamically stable. This means that the network has an auto-repair ability and robustness against unexpected destruction. The subexcitability of a channel is related to its diameter, so the required value can be obtained by selecting the right composition of the mixture and the conditions at which the phase transition occurs. Moreover, the structure is three dimensional, which allows for a higher density of processing elements than that obtained with the classical two-dimensional techniques, for example lithography.

Perspectives and Conclusions

Perspectives

In this article we have described a number of simple devices constructed with structured excitable chemical media which process information coded in excitation pulses. All the considered systems process information in an unconventional (non-von Neumann) way; i. e., without an external clock or synchronizing signal that controls the sequence of operations. On the other hand, in many cases the right timing of performed operation is hidden in the geometrical distribution and sizes of excitable regions. The described devices can be used as building blocks for more complex systems that process signals formed of excitation pulses. Some of the discussed devices can be controlled with spikes. Therefore, there is room for programming and learning. However, the further development of applications of structured excitable medium for information processing along this line seem to follow the evolution of classical electronic computing. In our opinion it would be more interesting to step away from this path and learn more about the potential offered by excitable media.

It would be interesting to study new types of excitable media suitable for information processing. Electrical analogs of reaction-diffusion systems (► Unconventional Computing, Novel Hardware for) seem promising, because they are more robust than the wetware based on liquid BZ media. In such media, spikes propagate much faster and the spatial scale can be much reduced if compared with typical chemical systems. They seem to be promising candidates for the hardware in geometrically oriented problems and direct image process-

ing [4]. However, two interesting properties of chemical reaction-diffusion systems are lost in their electrical analogs. First, chemical information processing systems, unlike the electronic ones, integrate two functions: the chemical reactivity and the ability to process information. Having in mind the excitable oxidation of CO on a Pt surface and the potential application of this medium for information processing [23], we can think of catalysts that are able to monitor their activity and report it. Second, the media described in ► Unconventional Computing, Novel Hardware for are two-dimensional. Most of the systems discussed in this article can be realized in three dimensions. The use of two dimensional media in chemical experiments is mainly related with significant difficulties in observations of three dimensional chemical excitations [42]. Potential application of phase transitions for channel structure generation described in the previous section should increase the interest in information processing with three dimensional medium.

Studies on more effective methods of information coding are important for the further development of reaction-diffusion computing. The most simple translation of chemistry into the language of information science based on the equivalence between the presence of a spike and the TRUE logic value is certainly not the most efficient. It can be expected that information interpreted within multi-valued logic systems is more suitable for transformations with the use of structured media [47]. As an example, let us consider three-valued logic encoded by excitation pulses in the following way: the lack of a pulse represents a logical FALSE (F), two pulses separated by time δt correspond at the logical TRUE (T), and one pulse is interpreted as a "nonsense" (★). Let us assume that two signals coded as described above are directed to the inputs of the AND gate illustrated on Fig. 9 and that they are fully synchronized at the input. The gap between the input and output channels is adjusted such that an activator diffusing from a single traveling pulse is not sufficient to generate a new pulse in the output channel, but the excitation resulting from the collision of facing pulses at the junction point exceeds the threshold and the output pulse is generated. Moreover, let us assume that the system is described by the dynamics for which signals are transformed as illustrated in Fig. 7a (for example FitzHugh–Nagumo dynamics) and that the time difference between spikes δt is selected such that the second spike can pass the gap. The output signal appears when spikes from both input channels are in coincidence or when the second spike from the same input arrives. As a result the device performs the following function within 3-valued logic [47]:

$\boxdot_{\mathcal{F}_1}$	T	F	★
T	T	★	★
F	★	F	F
★	★	F	★

The same operation, if performed on a classical computer would require a procedure that measures time between spikes. The structured excitable medium performs it naturally provided that the time δt is adjusted with the properties of the medium and the geometry of the gap. Of course, similar devices that process variables of n-valued logic can be also constructed with structured excitable media.

Conclusions

In the article we have presented a number of examples that should convince the reader that structured excitable media can be used for information processing. The future research will verify if this branch of computer science is fruitful. It would be important to find new algorithms that can be efficiently executed using a structured excitable medium. However, the ultimate test for the usefulness of ideas should come from biology. It is commonly accepted that excitable behavior is responsible for information processing and coding in living organisms [8,31,36,60]. We believe that studies on chemical information processing will help us to better understand these problems. And, although at the moment computing with a homogeneous excitable medium seems to offer more applications than that with the structured one, we believe the proportions will be reversed in the future. After all, our brains are not made of a single piece of a homogeneous excitable medium.

Acknowledgments

The research on information processing with structured excitable medium has been supported by the Polish State Committee for Scientific Research project 1 P03B 035 27.

Bibliography

Primary Literature

1. Adamatzky A, De Lacy Costello B (2002) Experimental logical gates in a reaction-diffusion medium: The XOR gate and beyond. Phys Rev E 66:046112
2. Adamatzky A (2004) Collision-based computing in Belousov–Zhabotinsky medium. Chaos Soliton Fractal 21(5):1259–1264
3. Adamatzky A (2005) Programming Reaction-Diffusion Processors. In: Banatre J-P, Fradet P, Giavitto J-L, Michel O (eds) LNCS, vol 3566. Springer, pp 47–55

4. Adamatzky A, De Lacy Costello B, Asai T (2005) Reaction-Diffusion Computers. Elsevier, UK
5. Agladze K, Aliev RR, Yamaguchi T, Yoshikawa K (1996) Chemical diode. J Phys Chem 100:13895–13897
6. Agladze K, Magome N, Aliev R, Yamaguchi T, Yoshikawa K (1997) Finding the optimal path with the aid of chemical wave. Phys D 106:247–254
7. Agladze K, Tóth Á, Ichino T, Yoshikawa K (2000) Propagation of Chemical Pulses at the Boundary of Excitable and Inhibitory Fields. J Phys Chem A 104:6677–6680
8. Agmon-Snir H, Carr CE, Rinzel J (1998) The role of dendrites in auditory coincidence detection. Nature 393:268–272
9. Amemiya T, Ohmori T, Yamaguchi T (2000) An Oregonator-class model for photoinduced Behavior in the $Ru(bpy)_3^{2+}$–Catalyzed Belousov–Zhabotinsky reaction. J Phys Chem A 104:336–344
10. Armstrong GR, Taylor AF, Scott SK, Gaspar V (2004) Modelling wave propagation across a series of gaps. Phys Chem Chem Phys 6:4677–4681
11. Babin V, Ciach A (2003) Response of the bicontinuous cubic D phase in amphiphilic systems to compression or expansion. J Chem Phys 119:6217–6231
12. Bertram M, Mikhailov AS (2003) Pattern formation on the edge of chaos: Mathematical modeling of CO oxidation on a Pt(110) surface under global delayed feedback. Phys Rev E 67:036207
13. Dolnik M, Finkeova I, Schreiber I, Marek M (1989) Dynamics of forced excitable and oscillatory chemical-reaction systems. J Phys Chem 93:2764–2774; Finkeova I, Dolnik M, Hrudka B, Marek M (1990) Excitable chemical reaction systems in a continuous stirred tank reactor. J Phys Chem 94:4110–4115; Dolnik M, Marek M (1991) Phase excitation curves in the model of forced excitable reaction system. J Phys Chem 95:7267–7272; Dolnik M, Marek M, Epstein IR (1992) Resonances in periodically forced excitable systems. J Phys Chem 96:3218–3224
14. Epstein IR, Showalter K (1996) Nonlinear Chemical Dynamics: Oscillations, Patterns, and Chaos. J Phys Chem 100:13132–13147
15. Feynman RP, Allen RW, Heywould T (2000) Feynman Lectures on Computation. Perseus Books, New York
16. Field RJ, Körös E, Noyes RM (1972) Oscillations in chemical systems. II. Thorough analysis of temporal oscillation in the bromate-cerium-malonic acid system. J Am Chem Soc 94:8649–8664
17. FitzHugh R (1960) Thresholds and plateaus in the Hodgkin-Huxley nerve equations. J Gen Physiol 43:867–896
18. FitzHugh R (1961) Impulses and physiological states in theoretical models of nerve membrane. Biophys J 1:445–466
19. Field RJ, Noyes RM (1974) Oscillations in chemical systems. IV. Limit cycle behavior in a model of a real chemical reaction. J Chem Phys 60:1877–1884
20. Gaspar V, Bazsa G, Beck MT (1983) The influence of visible light on the Belousov–Zhabotinskii oscillating reactions applying different catalysts. Z Phys Chem (Leipzig) 264:43–48
21. Ginn BT, Steinbock B, Kahveci M, Steinbock O (2004) Microfluidic Systems for the Belousov–Zhabotinsky Reaction. J Phys Chem A 108:1325–1332
22. Gorecka J, Gorecki J (2003) T-shaped coincidence detector as a band filter of chemical signal frequency. Phys Rev E 67:067203
23. Gorecka J, Gorecki J (2005) On one dimensional chemical diode and frequency generator constructed with an excitable surface reaction. Phys Chem Chem Phys 7:2915–2920
24. Gorecka J, Gorecki J (2006) Multiargument logical operations performed with excitable chemical medium. J Chem Phys 124:084101 (1–5)
25. Gorecka J, Gorecki J, Igarashi Y (2007) One dimensional chemical signal diode constructed with two nonexcitable barriers. J Phys Chem A 111:885–889
26. Gorecki J, Kawczynski AL (1996) Molecular dynamics simulations of a thermochemical system in bistable and excitable regimes. J Phys Chem 100:19371–19379
27. Gorecki J, Yoshikawa K, Igarashi Y (2003) On chemical reactors that can count. J Phys Chem A 107:1664–1669
28. Gorecki J, Gorecka JN, Yoshikawa K, Igarashi Y, Nagahara H (2005) Sensing the distance to a source of periodic oscillations in a nonlinear chemical medium with the output information coded in frequency of excitation pulses. Phys Rev E 72:046201 (1–7)
29. Gorecki J, Gorecka JN (2006) Information processing with chemical excitations – from instant machines to an artificial chemical brain. Int J Unconv Comput 2:321–336
30. Haken H (2002) Brain Dynamics. In: Springer Series in Synergetics. Springer, Berlin
31. Häusser M, Spruston N, Stuart GJ (2000) Diversity and Dynamics of Dendritic Signaling. Science 290:739–744
32. Hertz J, Krogh A, Palmer RG (1991) Introduction to the theory of neural computation. Addison-Wesley, Redwood City
33. Ichino T, Igarashi Y, Motoike IN, Yoshikawa K (2003) Different operations on a single circuit: Field computation on an Excitable Chemical System. J Chem Phys 118:8185–8190
34. Igarashi Y, Gorecki J, Gorecka JN (2006) Chemical information processing devices constructed using a nonlinear medium with controlled excitability. Lect Note Comput Science 4135:130–138
35. Kapral R, Showalter K (1995) Chemical Waves and Patterns. Kluwer, Dordrecht
36. Kindzelskii AL, Petty HR (2003) Intracellular calcium waves accompany neutrophil polarization, formylmethionyl-leucylphenylalanine stimulation, and phagocytosis: a high speed microscopy study. J Immunol 170:64–72
37. Krischer K, Eiswirth M, Ertl GJ (1992) Oscillatory CO oxidation on Pt(110): modelling of temporal self-organization. J Chem Phys 96:9161–9172
38. Krug HJ, Pohlmann L, Kuhnert L (1990) Analysis of the modified complete Oregonator accounting for oxygen sensitivity and photosensitivity of Belousov–Zhabotinskii systems. J Phys Chem 94:4862–4866
39. Kuramoto Y (1984) Chemical oscillations, waves, and turbulence. Springer, Berlin
40. Kusumi T, Yamaguchi T, Aliev RR, Amemiya T, Ohmori T, Hashimoto H, Yoshikawa K (1997) Numerical study on time delay for chemical wave transmission via an inactive gap. Chem Phys Lett 271:355–360
41. Lázár A, Noszticzius Z, Försterling H-D, Nagy-Ungvárai Z (1995) Chemical pulses in modified membranes I. Developing the technique. Physica D 84:112–119; Volford A, Simon PL, Farkas H, Noszticzius Z (1999) Rotating chemical waves: theory and experiments. Physica A 274:30–49
42. Luengviriya C, Storb U, Hauser MJB, Müller SC (2006) An elegant method to study an isolated spiral wave in a thin layer

of a batch Belousov–Zhabotinsky reaction under oxygen-free conditions. Phys Chem Chem Phys 8:1425–1429
43. Manz N, Müller SC, Steinbock O (2000) Anomalous dispersion of chemical waves in a homogeneously catalyzed reaction system. J Phys Chem A 104:5895–5897; Steinbock O (2002) Excitable Front Geometry in Reaction-Diffusion Systems with Anomalous Dispersion. Phys Rev Lett 88:228302
44. Maselko J, Reckley JS, Showalter K (1989) Regular and irregular spatial patterns in an immobilized-catalyst Belousov–Zhabotinsky reaction. J Phys Chem 93:2774–2780
45. Mikhailov AS, Showalter K (2006) Control of waves, patterns and turbulence in chemical systems. Phys Rep 425:79–194
46. Morozov VG, Davydov NV, Davydov VA (1999) Propagation of Curved Activation Fronts in Anisotropic Excitable Media. J Biol Phys 25:87–100
47. Motoike IN, Adamatzky A (2005) Three-valued logic gates in reaction–diffusion excitable media. Chaos, Solitons & Fractals 24:107–114
48. Motoike I, Yoshikawa K (1999) Information Operations with an Excitable Field. Phys Rev E 59:5354–5360
49. Motoike IN, Yoshikawa K, Iguchi Y, Nakata S (2001) Real–Time Memory on an Excitable Field. Phys Rev E 63:036220 (1–4)
50. Motoike IN, Yoshikawa K (2003) Information operations with multiple pulses on an excitable field. Chaos, Solitons & Fractals 17:455–461
51. Murray JD (1989) Mathematical Biology. Springer, Berlin
52. Nagai Y, Gonzalez H, Shrier A, Glass L (2000) Paroxysmal Starting and Stopping of Circulatong Pulses in Excitable Media. Phys Rev Lett 84:4248–4251
53. Nagahara H, Ichino T, Yoshikawa K (2004) Direction detector on an excitable field: Field computation with coincidence detection. Phys Rev E 70:036221 (1–5)
54. Nagumo J, Arimoto S, Yoshizawa S (1962) An Active Pulse Transmission Line Simulating Nerve Axon. Proc IRE 50: 2061–2070
55. Noszticzuis Z, Horsthemke W, McCormick WD, Swinney HL, Tam WY (1987) Sustained chemical pulses in an annular gel reactor: a chemical pinwheel. Nature 329:619–620
56. Plaza F, Velarde MG, Arecchi FT, Boccaletti S, Ciofini M, Meucci R (1997) Excitability following an avalanche-collapse process. Europhys Lett 38:85–90
57. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2007) Numerical Recipes, 3rd edn. The Art of Scientific Computing. Available via http://www.nr.com.
58. Rambidi NG, Maximychev AV (1997) Towards a Biomolecular Computer. Information Processing Capabilities of Biomolecular Nonlinear Dynamic Media. BioSyst 41:195–211
59. Rambidi NG, Yakovenchuk D (1999) Finding paths in a labyrinth based on reaction-diffusion media. BioSyst 51:67–72; Rambidi NG, Yakovenchuk D (2001) Chemical reaction-diffusion implementation of finding the shortest paths in a labyrinth. Phys Rev E 63:026607
60. Rambidi NG (2005) Biologically Inspired Information Processing Technologies: Reaction-Diffusion Paradigm. Int J Unconv Comp 1:101–121
61. Rovinsky AB, Zhabotinsky AM (1984) Mechanism and mathematical model of the oscillating bromate-ferroin-bromomalonic acid reaction. J Phys Chem 88:6081–6084
62. Rovinsky AB (1986) Spiral waves in a model of the ferroin catalyzed Belousov–Zhabotinskii reaction. J Phys Chem 90:217–219
63. Sielewiesiuk J, Gorecki J (2002) Chemical Waves in an Excitable Medium: Their Features and Possible Applications in Information Processing. In: Klonowski W (ed) Attractors, Signals and Synergetics. 1st European Interdisciplinary School on Nonlinear Dynamics for System and Signal Analysis Euroattractor 2000, Warsaw, 6–15 June 2000. Pabst, Lengerich, pp 448–460
64. Sielewiesiuk J, Gorecki J (2002) On complex transformations of chemical signals passing through a passive barrier. Phys Rev E 66:016212; Sielewiesiuk J, Gorecki J (2002) Passive barrier as a transformer of chemical signal frequency. J Phys Chem A 106:4068–4076
65. Sielewiesiuk J, Gorecki J (2001) Chemical impulses in the perpendicular junction of two channels. Acta Phys Pol B 32:1589–1603
66. Sielewiesiuk J, Gorecki J (2001) Logical functions of a cross junction of excitable chemical media. J Phys Chem A 105:8189–8195
67. Steinbock O, Toth A, Showalter K (1995) Navigating complex labyrinths – optimal paths from chemical waves. Science 267:868–871
68. Steinbock O, Kettunen P, Showalter K (1995) Anisotropy and Spiral Organizing Centers in Patterned Excitable Media. Science 269:1857–1860
69. Steinbock O, Kettunen P (1996) Chemical clocks on the basis of rotating pulses. Measuring irrational numbers from period ratios. Chem Phys Lett 251:305–308
70. Steinbock O, Kettunen P, Showalter K (1996) Chemical Pulse Logic Gates. J Phys Chem 100:18970–18975
71. Suzuki K, Yoshinobu T, Iwasaki H (1999) Anisotropic waves propagating on two–dimensional arrays of Belousov–Zhabotinsky oscillators. Jpn J Appl Phys 38:L345–L348
72. Suzuki K, Yoshinobu T, Iwasaki H (2000) Unidirectional propagation of chemical waves through microgaps between zones with different excitability. J Phys Chem A 104:6602–6608
73. Suzuki R (1967) Mathematical analysis and application of iron-wire neuron model. IEEE Trans Biomed Eng 14:114–124
74. Taylor AF, Armstrong GR, Goodchild N, Scott SK (2003) Propagation of chemical waves across inexcitable gaps. Phys Chem Chem Phys 5:3928–3932
75. Toth A, Showalter K (1995) Logic gates in excitable media. J Chem Phys 103:2058–2066
76. Toth A, Horvath D, Yoshikawa K (2001) Unidirectional wave propagation in one spatial dimension. Chem Phys Lett 345: 471–474
77. Yoshikawa K, Nagahara H, Ichino T, Gorecki J, Gorecka JN, Igarashi Y (2009) On Chemical Methods of Direction and Distance Sensing. Int J Unconv Comput 5:53–65

Books and Reviews

Hjelmfelt A, Ross J (1994) Pattern recognition, chaos, and multiplicity in neural networks of excitable systems. Proc Natl Acad Sci USA 91:63–67

Nakata S (2003) Chemical Analysis Based on Nonlinearity. Nova, New York

Rambidi NG (1998) Neural network devices based on reaction–diffusion media: an Approach to artificial retina. Supramol Sci 5:765–767

Storb U, Müller SC (2004) Scroll waves. In: Scott A (ed) Encyclopedia of nonlinear sciences. Routledge, Taylor and Francis Group, New York, pp 825–827

Computing with Solitons

Darren Rand[1], Ken Steiglitz[2]
[1] Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, USA
[2] Computer Science Department, Princeton University, Princeton, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Manakov Solitons
Manakov Soliton Computing
Multistable Soliton Collision Cycles
Experiments
Future Directions
Bibliography

Glossary

Integrable This term is generally used in more than one way and in different contexts. For the purposes of this article, a partial differential equation or system of partial differential equations is *integrable* if it can be solved explicitly to yield *solitons* (qv).

Manakov system A system of two cubic Schrödinger equations where the self- and cross-phase modulation terms have equal weight.

Nonlinear Schrödinger equation A partial differential equation that has the same form as the Schrödinger equation of quantum mechanics, with a term nonlinear in the dependent variable, and for the purposes of this article, interpreted classically.

Self- and cross-phase modulation Any terms in a nonlinear Schrödinger equation that involve nonlinear functions of the dependent variable of the equation, or nonlinear functions of a dependent variable of another (coupled) equation, respectively.

Solitary wave A solitary wave is a wave characterized by undistorted propagation. Solitary waves do not in general maintain their shape under perturbations or collisions.

Soliton A soliton is a solitary wave which is also robust under perturbations and collisions.

Turing equivalent Capable of simulating any Turing Machine, and hence by Turing's Thesis capable of performing any computation that can be carried out by a sequence of effective instructions on a finite amount of data. A machine that is Turing equivalent is therefore as powerful as any digital computer. Sometimes a device that is Turing equivalent is called "universal."

Definition of the Subject

Solitons are localized, shape-preserving waves characterized by robust collisions. First observed as water waves by John Scott Russell [29] in the Union Canal near Edinburgh and subsequently recreated in the laboratory, solitons arise in a variety of physical systems, as both temporal pulses which counteract dispersion and spatial beams which counteract diffraction.

Solitons with two components, vector solitons, are computationally universal due to their remarkable collision properties. In this article, we describe in detail the characteristics of Manakov solitons, a specific type of vector soliton, and their applications in computing.

Introduction

In this section, we review the basic principles of soliton theory and spotlight relevant experimental results. Interestingly, the phenomena of soliton propagation and collision occur in many physical systems despite the diversity of mechanisms that bring about their existence. For this reason, the discussion in this article will treat temporal and spatial solitons interchangeably, unless otherwise noted.

Scalar Solitons

A pulse in optical fiber undergoes dispersion, or temporal spreading, during propagation. This effect arises because the refractive index of the silica glass is not constant, but is rather a function of frequency. The pulse can be decomposed into a frequency range—the shorter the pulse, the broader its spectral width. The frequency dependence of the refractive index will cause the different frequencies of the pulse to propagate at different velocities, giving rise to dispersion. As a result, the pulse develops a chirp, meaning that the individual frequency components are not evenly distributed throughout the pulse. There are two types of dispersion: normal and anomalous. If the longer wavelengths travel faster, the medium is said to have *normal*dispersion. If the opposite is true, the medium has *anomalous* dispersion.

The response of a dielectric such as optical fiber is nonlinear. Most of the nonlinear effects in fiber originate from nonlinear refraction, where the refractive index n depends on the intensity of the propagating field according to the relation

$$n = n_0 + n_2|E|^2, \tag{1}$$

where n_0 is the linear part of the refractive index, $|E^2|$ is the optical intensity, and n_2 is the coefficient of nonlinear contribution to the refractive index. Because the material responds almost instantaneously, on the order of femtoseconds, and because the phase shift $\Delta\phi$ is proportional to n, each component of an intense optical pulse sees a phase shift proportional to its intensity. Since the frequency shift $\delta\omega = -(\partial\Delta\phi)/(\partial t)$, the leading edge of the pulse is red-shifted ($\delta\omega < 0$), while the trailing edge is blue-shifted ($\delta\omega > 0$), an effect known as self-phase modulation (SPM). As a result, if the medium exhibits normal dispersion, the pulse is broadened; for anomalous dispersion, the pulse is compressed. Under the proper conditions, this pulse compression can exactly cancel the linear, dispersion-induced broadening, resulting in distortionless soliton propagation. For more details, see the book by Agrawal [3].

The idealized mathematical model for this pulse propagation is the nonlinear Schrödinger equation (NLSE):

$$i\frac{\partial u}{\partial z} \pm \frac{1}{2}\frac{\partial^2 u}{\partial x^2} + |u|^2 u = 0, \tag{2}$$

where $u(z, x)$ is the complex-valued field envelope, z is a normalized propagation distance and x is normalized time propagating with the group velocity of the pulse. The second and third terms describe dispersion and the intensity-dependent Kerr nonlinearity, respectively. The coefficient of the dispersion term is positive for anomalous dispersion and negative for normal dispersion. Equation (2), known as the *scalar* NLSE, is integrable—that is, it can be solved analytically, and collisions between solitons are 'elastic,' in that no change in amplitude or velocity occurs as a result of a collision. Zakharov and Shabat [38] first solved this equation analytically using the inverse scattering method. It describes, for example, the propagation of picosecond or longer pulses propagating in lossless optical fiber.

Two solitons at different wavelengths will collide in an optical fiber due to dispersion-induced velocity differences. A schematic of such a collision is depicted in Fig. 1. The scalar soliton collision is characterized by two phenomena— a position and phase shift—both of which can be understood in the same intuitive way. During collision, there will be a local increase in intensity, causing a local increase in the fiber's refractive index, according to Eq. (1). As a result, both the soliton velocity and phase will be affected during the collision.

From an all-optical signal processing perspective, the phase and position shifts in a soliton collision are not useful. This is because these effects are independent of any soliton properties that are changed by collision; that is, the result of one collision will not affect the result of subsequent collisions. Scalar solitons are therefore not useful for complex logic or computing, which depend on multiple, cascaded interactions.

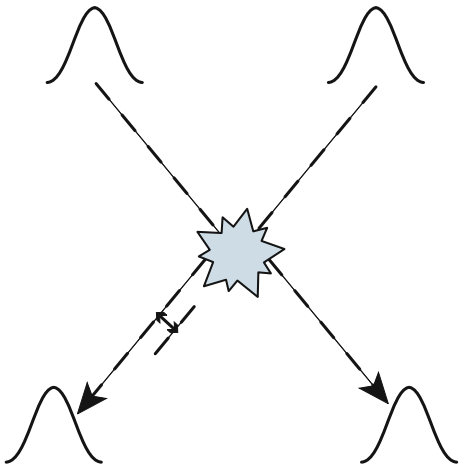

Computing with Solitons, Figure 1
Schematic of a scalar soliton collision, in which amplitude and velocities are unchanged. The two soliton collision effects are a position shift (depicted through the translational shift in the soliton path) and phase shift (not pictured)

Despite this setback, it was discovered later that a system similar to the scalar NLSE, the Manakov system [19], possesses very rich collisional properties [26] and is integrable as well. Manakov solitons are a specific instance of two-component vector solitons, and it has been shown that collisions of Manakov solitons are capable of transferring information via changes in a complex-valued polarization state [16].

Vector Solitons

When several field components, distinguished by polarization and/or frequency, propagate in a nonlinear medium, the nonlinear interaction between them must be considered as well. This interaction between field components results in intensity-dependent nonlinear coupling terms analogous to the self-phase modulation term in the scalar case. Such a situation gives rise to a set of coupled nonlinear Schrödinger equations, and may allow for propagation of *vector* solitons. For the case of two components propagating in an ideal sl medium with no higher-order effects and only intensity-dependent nonlinear coupling, the equations become:

$$\begin{aligned} i\frac{\partial u_1}{\partial z} + \frac{\partial^2 u_1}{\partial x^2} + 2\mu(|u_1|^2 + \alpha|u_2|^2)u_1 = 0\,, \\ i\frac{\partial u_2}{\partial z} + \frac{\partial^2 u_2}{\partial x^2} + 2\mu(|u_2|^2 + \alpha|u_1|^2)u_2 = 0\,, \end{aligned} \tag{3}$$

where $u_1(z,x)$ and $u_2(z,x)$ are the complex-valued pulse envelopes for each component, μ is a nonlinearity parameter, and α describes the ratio between self- and cross-phase modulation contributions to the overall nonlinearity. Only for the special case of $\alpha = 1$ are Eqs. (3) integrable. First solved using the method of inverse scattering by Manakov [19], Eqs. (3) admit solutions known as Manakov solitons. For nonintegrable cases ($\alpha \neq 1$), some analytical solitary-wave solutions are known for specific cases, although in general a numerical approach is required [36]. The specific case of $\alpha = 2/3$, for example, corresponds to linearly birefringent polarization maintaining fiber, and will be considered in more detail in Sect. "Experiments".

Due to their multicomponent structure, vector solitons have far richer collision dynamics than their scalar, one-component counterparts. Recall that scalar collisions are characterized by phase and position shifts only. Vector soliton collisions also exhibit these effects, with the added feature of possible intensity redistributions between the component fields [19,26]. This process is shown schematically in Fig. 2. In the collision, two conservation relations are satisfied: (i) the energy in each soliton is conserved and (ii) the energy in each component is conserved. It can be seen that when the amplitude of one component in a soliton increases as a result of the collision, the other component decreases, with the opposite exchange in the second soliton. The experimental observation of this effect will be discussed in Sect. "Experiments". In addition to fundamental interest in such solitons, collisions of vector solitons make possible unique applications, including collision-based logic and universal computation [16,27,34,35], as discussed in Sect. "Manakov Soliton Computing".

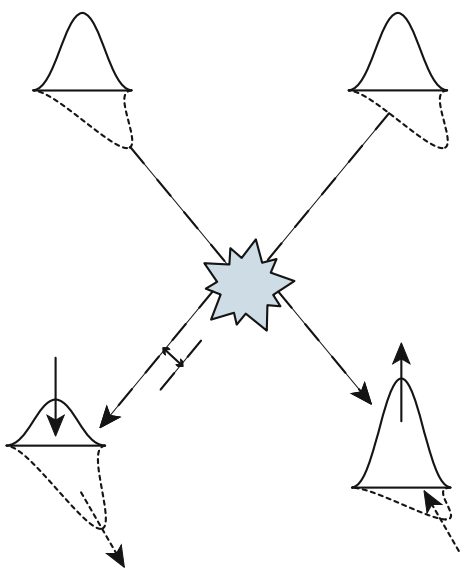

Computing with Solitons, Figure 2
Schematic of a vector soliton collision, which exhibits a position shift and phase shift (not pictured), similar to the scalar soliton collision (cf. Fig. 1). Vector soliton collisions also display an energy redistribution among the component fields, shown here as two orthogonal polarizations. Arrows indicate direction of energy redistribution

Manakov Solitons

As mentioned in Sect. "Introduction", computation is possible using vector solitons because of an energy redistribution that occurs in a collision. In this section, we provide the mathematic background of Manakov soliton theory, in order to understand soliton computing and a remarkable way to achieve bistability using soliton collisions as described in Sects. "Manakov Soliton Computing" and "Multistable Soliton Collision Cycles", respectively.

The Manakov system consists of two coupled NLSEs [19]:

$$\begin{aligned} i\frac{\partial q_1}{\partial z} + \frac{\partial^2 q_1}{\partial x^2} + 2\mu(|q_1|^2 + |q_2|^2)q_1 = 0\,, \\ i\frac{\partial q_2}{\partial z} + \frac{\partial^2 q_2}{\partial x^2} + 2\mu(|q_1|^2 + |q_2|^2)q_2 = 0\,, \end{aligned} \tag{4}$$

where $q_1(x,z)$ and $q_2(x,z)$ are two interacting optical components, μ is a positive parameter representing the strength of the nonlinearity, and x and z are normalized space and propagation distance, respectively. As mentioned in Sect. "Vector Solitons", the Manakov system is a special case of Eqs. (3) with $\alpha = 1$. The two components can be thought of as components in two polarizations, or, as in the case of a photorefractive crystal, two uncorrelated beams [11].

Manakov first solved Eqs. (4) by the method of inverse scattering [19]. The system admits single-soliton, two-component solutions that can be characterized by a complex number $k \equiv k_R + ik_I$, where k_R represents the energy of the soliton and k_I the velocity, all in normalized units. The additional soliton parameter is the complex-valued polarization state $\rho \equiv q_1/q_2$, defined as the (z- and x-independent) ratio between the q_1 and q_2 components.

Figure 3 shows the schematic for a general two-soliton collision, with initial parameters ρ_1, k_1 and ρ_L, k_2, corresponding to the right-moving and left-moving solitons, respectively. The values of k_1 and k_2 remain constant during collision, but in general the polarization state changes. Let ρ_1 and ρ_L denote the respective soliton states before impact, and suppose the collision transforms ρ_1 into ρ_R, and ρ_L into ρ_2. It turns out that the state change undergone

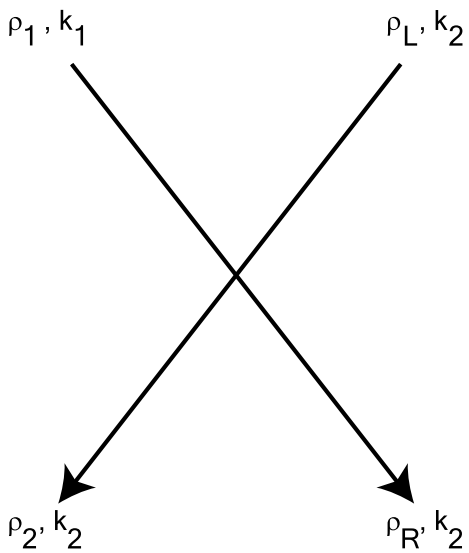

Computing with Solitons, Figure 3
Schematic of a general two-soliton collision. Each soliton is characterized by a complex-valued polarization state ρ and complex parameter k. Reprinted with permission from [34]. Copyright by the American Physical Society

by each colliding soliton takes on the very simple form of a linear fractional transformation (also called a bilinear or Möbius transformation). Explicitly, the state of the emerging left-moving soliton is given by [16]:

$$\rho_2 = \frac{[(1-g)/\rho_1^* + \rho_1]\rho_L + g\rho_1/\rho_1^*}{g\rho_L + (1-g)\rho_1 + 1/\rho_1^*} , \tag{5}$$

where

$$g \equiv \frac{k_1 + k_1^*}{k_2 + k_1^*} . \tag{6}$$

The state of the right-moving soliton is obtained similarly, and is

$$\rho_R = \frac{[(1-h^*)/\rho_L^* + \rho_L]\rho_1 + h^*\rho_L/\rho_L^*}{h^*\rho_1 + (1-h^*)\rho_L + 1/\rho_L^*} , \tag{7}$$

where

$$h \equiv \frac{k_2 + k_2^*}{k_1 + k_2^*} . \tag{8}$$

We assume here, without loss of generality, that k_{1R}, $k_{2R} > 0$.

Several properties of the linear fractional transformations in Eqs. (5) and (7) are derived in [16], including the characterization of inverse operators, fixed points, and implicit forms. In particular, when viewed as an operator every soliton has an *inverse*, which will undo the effect of the operator on the state. Note that this requires that the inverse operator have the same k parameter as the original, a condition that will hold in our application of computing in the next section.

These state transformations were first used by Jakubowski et al. [16] to describe logical operations such as NOT. Later, Steiglitz [34] established that arbitrary computation was possible through time gating of Manakov (1 + 1)-dimensional spatial solitons. We will describe this in Sect. "Manakov Soliton Computing".

There exist several candidates for the physical realization of Manakov solitons, including photorefractive crystals [4,5,9,11,30], semiconductor waveguides [17], quadratic media [33], and optical fiber [23,28]. In Sect. "Experiments", we discuss in detail an experiment with vector solitons in linearly birefringent optical fiber.

Manakov Soliton Computing

We described in the previous section how collisions of Manakov solitons can be described by transformations of a complex-valued state which is the ratio between the two Manakov components. We show in this section that general computation is possible if we use (1+1)-dimensional spatial solitons that are governed by the Manakov equations and if we are allowed to time-gate the beams input to the medium. The result is a dynamic computer without spatially fixed gates or wires, which is unlike most present-day conceptions of a computer that involve integrated circuits, in which information travels between logical elements that are fixed spatially through fabrication on a silicon wafer. We can call such a scheme "nonlithographic," in the sense that there is no architecture imprinted on the medium.

The requirements for computation include cascadability, fanout, and Boolean completeness. The first, cascadability, requires that the output of one device can serve as input to another. Since any useful computation consists of many stages of logic, this condition is essential. The second, fanout, refers to the ability of a logic gate to drive at least two similar gates. Finally, Boolean completeness makes it possible to perform arbitrary computation.

We should emphasize that although the model we use is meant to reflect known physical phenomena, at least in the limit of ideal behavior, the result is a mathematical one. Practical considerations of size and speed are not considered here, nor are questions of error propagation. In this sense the program of this article is analogous to Fredkin and Toffoli [13] for ideal billiard balls, and Shor [31] for quantum mechanics. There are however several can-

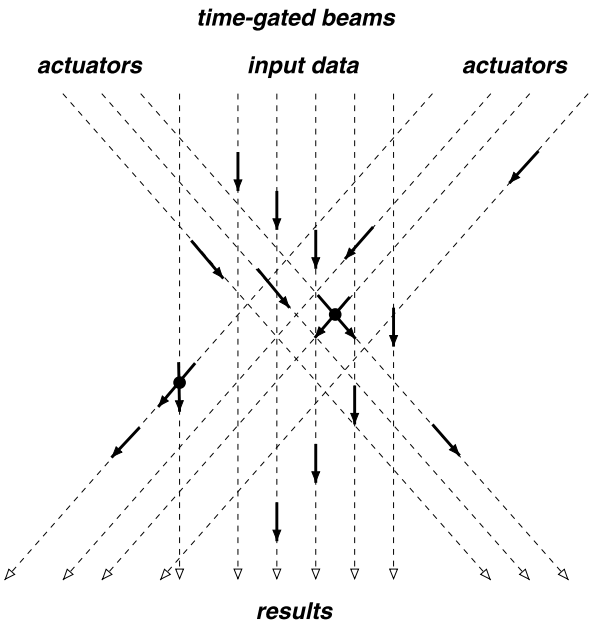

Computing with Solitons, Figure 4
The general physical arrangement considered in this paper. Time-gated beams of spatial Manakov solitons enter at the top of the medium, and their collisions result in state changes that reflect computation. Each solid arrow represents a beam segment in a particular state. Reprinted with permission from [34]. Copyright by the American Physical Society

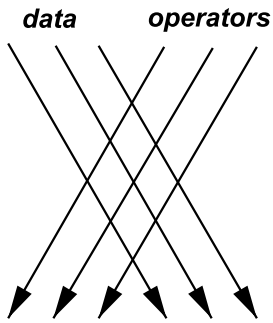

Computing with Solitons, Figure 5
Colliding spatial solitons. Reprinted with permission from [34]. Copyright by the American Physical Society

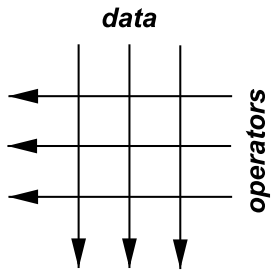

Computing with Solitons, Figure 6
Convenient representation of colliding spatial solitons. Reprinted with permission from [34]. Copyright by the American Physical Society

didates for physical instantiation of the basic ideas in this paper, as noted in the previous section.

Although we are describing computation embedded in a homogeneous medium, and not interconnected *gates* in the usual sense of the word, we will nevertheless use the term *gates* to describe prearranged sequences of soliton collisions that effect logical operations. We will in fact adopt other computer terms to our purpose, such as *wiring* to represent the means of moving information from one place to another, and *memory* to store it in certain ways for future use.

We will proceed in the construction of what amounts to a complete computer in the following stages: First we will describe a basic gate that can be used for FANOUT. Then we will show how the same basic configuration can be used for NOT, and finally, NAND. Then we will describe ways to use time gating of the input beams to interconnect signals. The NAND gate, FANOUT, and interconnect are sufficient to implement any computer, and we conclude with a layout scheme for a general-purpose, and hence Turing-equivalent computer. The general picture of the physical arrangement is shown in Fig. 4.

Figure 5 shows the usual picture of colliding solitons, which can work interchangeably for the case of temporal or spatial solitons. It is convenient for visualization purposes to turn the picture and adjust the scale so the axes are horizontal and vertical, as in Fig. 6. We will use binary logic, with two distinguished, distinct complex numbers representing TRUE and FALSE, called 1 and 0, respectively. In fact, it turns out to be possible to use complex 1 and 0 for these two state values, and we will do that throughout this paper, but this is a convenience and not at all a necessity. We will thus use complex polarization states 1 and 0 and logical 1 and 0 interchangeably.

FANOUT

We construct the FANOUT gate by starting with a COPY gate, implemented with collisions between three down-moving, vertical solitons and one left-moving horizontal soliton. Figure 7 shows the arrangement. The soliton state labeled *in* will carry a logical value, and so be in one of the two states 0 or 1. The left-moving soliton labeled *actuator* will be in the fixed state 0, as will be the case throughout this paper. The plan is to adjust the (so far) arbitrary states z and y so that $out = in$, justifying the name COPY. It is reasonable to expect that this might be possible, be-

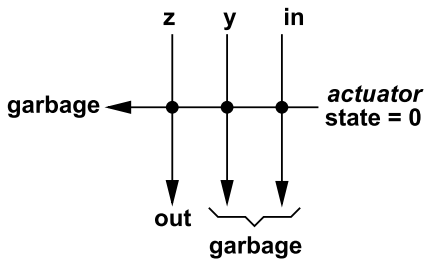

Computing with Solitons, Figure 7
COPY gate. Reprinted with permission from [34]. Copyright by the American Physical Society

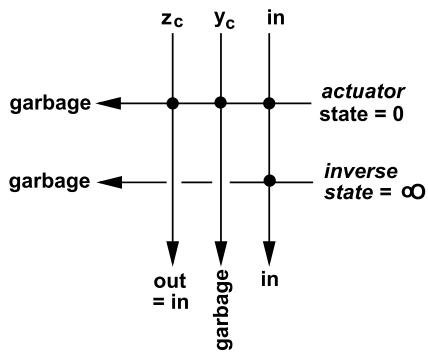

Computing with Solitons, Figure 8
FANOUT gate. Reprinted with permission from [34]. Copyright by the American Physical Society

cause there are four degrees of freedom in the two complex numbers z and y, and two complex equations to satisfy: that *out* be 1 and 0 when *in* is 1 and 0, respectively. Values that satisfy these four equations in four unknowns were obtained numerically. We will call them z_c and y_c. It is not always possible to solve these equations; Ablowitz et al. [1] showed that a unique solution is guaranteed to exist in certain parameter regimes. However, explicit solutions have been found for all the cases used in this section, and are given in Table 1.

To be more specific about the design problem, write Eq. (5) as the left-moving product $\rho_2 = L(\rho_1, \rho_L)$, and similarly write Eq. (7) as $\rho_R = R(\rho_1, \rho_L)$. The successive left-moving products in Fig. 7 are $L(in, 0)$ and $L(y, L(in, 0))$. The *out* state is then $R(z, L(y, L(in, 0)))$. The stipulation that 0 maps to 0 and 1 maps to 1 is expressed by the following two simultaneous complex equations in two complex unknowns

$$R(z, L(y, L(0, 0))) = 0 \,,$$
$$R(z, L(y, L(1, 0))) = 1 \,.$$

It is possible to solve for z as a function of y and then eliminate z from the equations, yielding one complex equation in the one complex unknown y. This is then solved numerically by grid search and successive refinement. There is no need for efficiency here, since we will require solutions in only a small number of cases.

To make a FANOUT gate, we need to recover the input, which we can do using a collision with a soliton in the state which is the inverse of 0, namely ∞ [16]. Figure 8 shows the complete FANOUT gate. Notice that we indicate collisions with a dot at the intersection of paths, and require that the continuation of the inverse soliton not intersect the continuation of z that it meets. We indicate that by a broken line, and postpone the explanation of how this "wire crossing" is accomplished. It is immaterial whether the continuation of the inverse operator hits the continuation of y, because it is not used later. We call such solitons *garbage* solitons.

NOT and ONE Gates

In the same way we designed the complex pair of states (z_c, y_c) to produce a COPY and FANOUT gate, we can find a pair (z_n, y_n) to get a NOT gate, mapping 0 to 1 and 1 to 0; and a pair (z_1, y_1) to get a ONE gate, mapping both 0 and 1 to 1. These (z, y) values are given in Table 1.

We should point out that the ONE gate in itself, considered as a one-input, one-output gate, is not invertible, and could never be achieved by using the continuation of one particular soliton through one, or even many collisions. This is because such transformations are always nonsin-

Computing with Solitons, Table 1
Parameters for gates when soliton speeds are 1

gate	z	y
COPY	$-0.24896731 - 0.62158212 \cdot I$	$2.28774210 + 0.01318152 \cdot I$
NOT	$-0.17620885 + 0.38170630 \cdot I$	$0.07888703 - 1.26450654 \cdot I$
ONE	$-0.45501471 - 1.37634227 \cdot I$	$1.43987094 + 0.64061349 \cdot I$
Z-CONV	$0.31838068 - 0.43078735 \cdot I$	$-0.04232340 + 2.17536612 \cdot I$
Y-CONV	$1.37286955 + 0.88495501 \cdot I$	$-0.58835758 - 0.18026939 \cdot I$

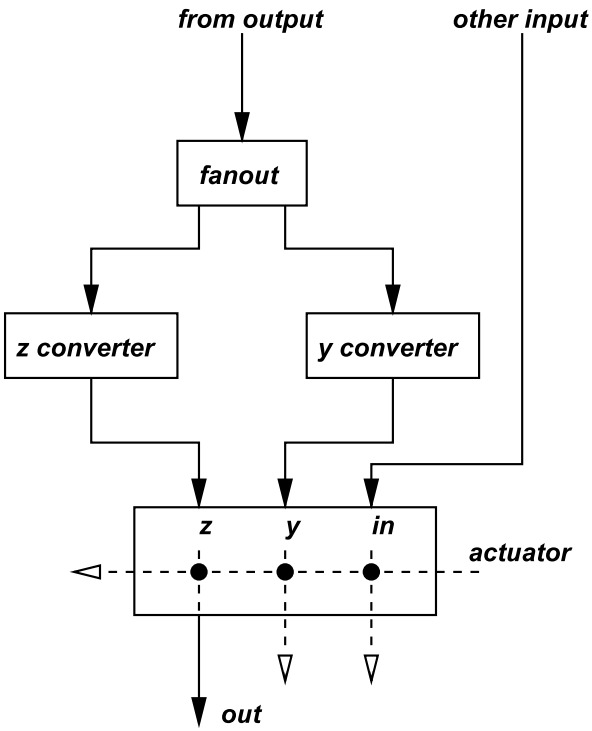

Computing with Solitons, Figure 9
A NAND gate, using converter gates to couple copies of one of its inputs to its *z* and *y* parameters. Reprinted with permission from [34]. Copyright by the American Physical Society

gular linear fractional transformations, which are invertible [16]. The transformation of state from the input to the continuation of z is, however, much more complicated and provides the flexibility we need to get the ONE gate. It turns out that this ONE gate will give us a row in the truth table of a NAND, and is critical for realizing general logic.

Output/Input Converters, Two-Input Gates, and NAND

To perform logic of any generality we must of course be able to use the output of one operation as the input to another. To do this we need to convert logic (0/1) values to some predetermined z and y values, the choice depending on the type of gate we want. This results in a two-input, one-output gate.

As an important example, here's how a NAND gate can be constructed. We design a z-converter that converts 0/1 values to appropriate values of z, using the basic three-collision arrangement shown in Fig. 7. For a NAND gate, we map 0 to z_1, the z value for the ONE gate, and map 1 to z_n, the z value for the NOT gate. Similarly, we construct a y-converter that maps 0 to y_1 and 1 to y_n. These z- and y-converters are used on the fanout of one of the inputs, and the resulting two-input gate is shown in Fig. 9. Of course these z- and y-converters require z and y values themselves, which are again determined by numerical search (see Table 1).

The net effect is that when the left input is 0, the other input is mapped by a ONE gate, and when it is 1 the other input is mapped by a NOT gate. The only way the output can be 0 is if both inputs are 1, thus showing that this is a NAND gate. Another way of looking at this construction is that the 2×2 truth table of (left input)×(right input) has as its 0 row a ONE gate of the columns (1 1), and as its 1 row a NOT gate of the columns (1 0).

The importance of the NAND gate is that it is *universal* [20]. That is, it can be used with interconnects and fanouts to construct any other logical function. Thus we have shown that with the ability to "wire" we can implement any logic using the Manakov model.

We note that other choices of input converters result in direct realizations of other gates. Using input converters that convert 0 and 1 to (z_c, y_c) and (z_n, y_n), respectively, results in a truth table with first row (0 1) and second row (1 0), an XOR gate. Converting 0 and 1 to (z_c, y_c) and (z_1, y_1), respectively, results in an OR gate, and so on.

Time Gating

We next take up the question of interconnecting the gates described above, and begin by showing how the continuation of the input in the COPY gate can be restored without affecting the other signals. In other words, we show how a simple "wire crossing" can be accomplished in this case.

For spatial solitons, the key flexibility in the model is provided by assuming that input beams can be time-gated; that is, turned on and off. When a beam is thus gated, a finite segment of light is created that travels through the medium. We can think of these finite segments as finite light pulses, and we will call them simply *pulses* in the remainder of this article.

Figure 10a shows the basic three-collision gate implemented with pulses. Assuming that the actuator and data pulses are appropriately timed, the actuator pulse hits all three data pulses, as indicated in the projection below the space-space diagram. The problem is that if we want a later actuator pulse to hit the rightmost data pulse (to invert the state, for example, as in the FANOUT gate), it will also hit the remaining two data pulses because of the way they must be spaced for the earlier three collisions.

We can overcome this difficulty by sending the actuator pulse from the left instead of the right. Timing it appropriately early it can be made to miss the first two data

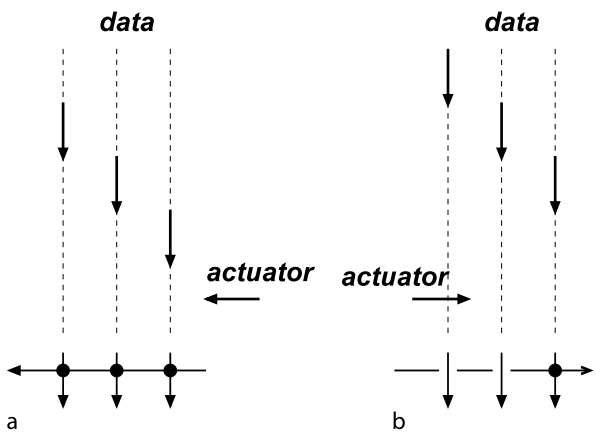

Computing with Solitons, Figure 10
a When entered from the right and properly timed, the actuator pulse hits all three data pulses, as indicated in the projection at the bottom; b When entered from the left and properly timed, the actuator pulse misses two data pulses and hits only the rightmost data pulse, as indicated in the projection at the bottom. Reprinted with permission from [34]. Copyright by the American Physical Society

pulses, and hit the third, as shown in Fig. 10b. It is easy to check that if the velocity of the right-moving actuator solitons is algebraically above that of the data solitons by the same amount that the velocity of the data solitons is algebraically above that of the left-moving actuator solitons, the same state transformations will result. For example, if we choose the velocities of the data and left-moving actuator solitons to be $+1$ and -1, we should choose the velocity of the right-moving actuator solitons to be $+3$. This is really a consequence of the fact that the g and h parameters of Eqs. (6) and (8) in the linear fractional transformation depend only on the difference in the velocities of the colliding solitons.

Wiring

Having shown that we can perform FANOUT and NAND, it remains only to show that we can "wire" gates so that any outputs can be fed to any inputs. The basic method for doing this is illustrated in Fig. 11. We think of data as stored in the down-moving pulses in a column, which we can think of as "memory". The observer moves with this frame, so the data appears stationary.

Pulses that are horizontal in the three-collision gates shown in previous figures will then appear to the observer to move upward at inclined angles. It is important to notice that these upward diagonally moving pulses are evanescent in our picture (and hence their paths are shown dashed in the figure). That is, once they are used, they do

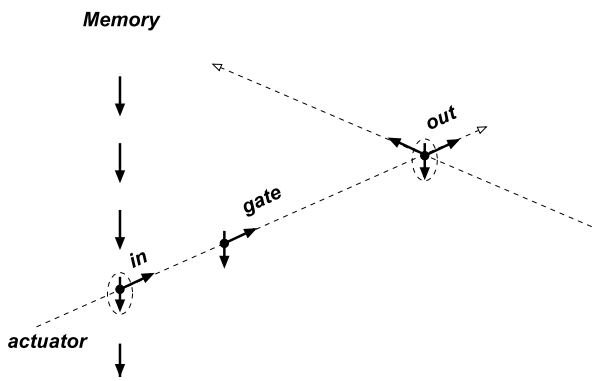

Computing with Solitons, Figure 11
The frame of this figure is moving down with the data pulses on the left. A data pulse in memory is operated on with a three-collision gate actuated from the left, and the result deposited to the upper right. Reprinted with permission from [34]. Copyright by the American Physical Society

not remain in the picture with a moving frame and hence cannot interfere with later computations. However, all vertically moving pulses remain stationary in this picture.

Once a diagonal trajectory is used for a three-collision gate, reusing it will in general corrupt the states of all the stationary pulses along that diagonal. However, the original data pulse (gate input) can be restored with a pulse in the state inverse to the actuator, either along the same diagonal as the actuator, provided we allow enough time for the result (the gate output, a stationary z pulse) to be used, or along the other diagonal.

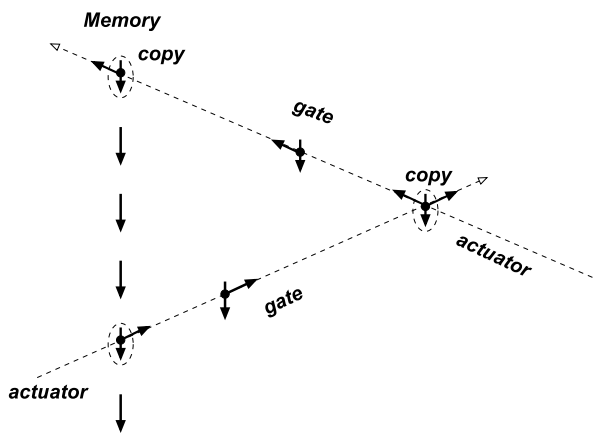

Computing with Solitons, Figure 12
A data pulse is copied to the upper right, this copy is copied to the upper left, and the result put at the top of memory. The original data pulse can then be restored with an inverse pulse and copied to the left in the same way. Reprinted with permission from [34]. Copyright by the American Physical Society

Suppose we want to start with a given data pulse in the memory column and create two copies above it in the memory column. Figure 12 shows a data pulse at the lower left being copied to the upper right with a three-collision COPY gate, initiated with an actuator pulse from the left. This copy is then copied again to the upper left, back to a waiting z pulse in the memory column. After the first copy is used, an inverse pulse can be used along the lower left to upper right diagonal to restore the original data pulse. The restored data pulse can then be copied to the left in the same way, to a height above the first copy, say, and thus two copies can be created and deposited in memory above the original.

A Second Speed and Final FANOUT and NAND

There is one problem still remaining with a true FANOUT: When an original data pulse in memory is used in a COPY operation for FANOUT, two diagonals are available, one from the lower left to the upper right, and the other from the lower right to the upper left. Thus, two copies can be made, as was just illustrated. However, when a data pulse is deposited in the memory column as a result of a logic operation, the logical operation itself uses at least one diagonal, which leaves at most one free. This makes a FANOUT of the *output* of a gate impossible with the current scheme.

A simple solution to this problem is to introduce another speed, using velocities ± 0.5, say, in addition to ± 1. This effectively provides four rather than two directions in which a pulse can be operated on, and allows true FANOUT and general interconnections. Figure 13 shows such a FANOUT; the data pulse at the lower left is copied to a position above it using one speed, and to another position, above that, using another.

Finally, a complete NAND gate is shown in Fig. 14. The gate can be thought of as composed of the following steps:

- input 2 is copied to the upper left, and that copy transformed by a z-converter to the upper right, placing the z pulse for the NAND gate at the top of the figure;
- after the copy of input 2 is used, input 2 is restored with an inverse pulse to the upper left;
- input 2 is then transformed to the upper right by a y-converter;
- input 1 is copied to the upper right, to a position collinear with the z- and y-converted versions of the other input;
- a final actuator pulse converts the z pulse at the top to the output of the NAND gate.

Note that the output of the NAND has used two diagonals, which again shows why a second speed is needed if we are to use the NAND output as an input to subsequent logical operations. The y operator pulses, middle components in the three-collision COPY and converter gates, are not shown in the figure, but room can always be made for them to avoid accidental collisions by adding only a constant amount of space.

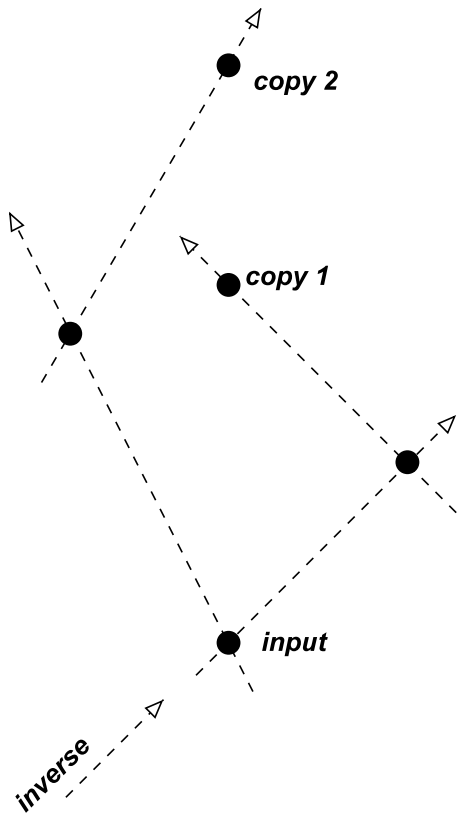

Computing with Solitons, Figure 13
The introduction of a second speed makes true FANOUT possible. For simplicity, in this and the next figure, data and operator pulses are indicated by solid dots, and the y operator pulses are not shown. The paths of actuator pulses are indicated by dashed lines. Reprinted with permission from [34]. Copyright by the American Physical Society

Universality

It should be clear now that any sequence of three-collision gates can be implemented in this way, copying data out of the memory column to the upper left or right, and performing NAND operations on any two at a time in the way shown in the previous section. The computation can proceed in a breadth-first manner, with the results of each successive stage being stored above the earlier results. Each additional gate can add only a constant amount of height

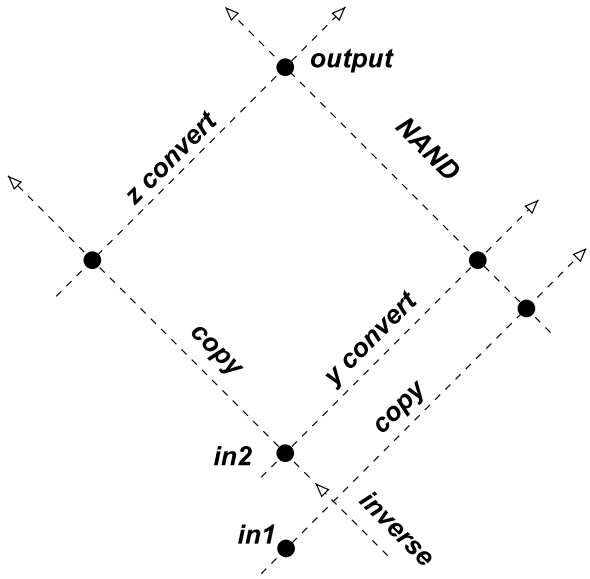

Computing with Solitons, Figure 14
Implementation of a NAND gate. A second speed will be necessary to use the output. Reprinted with permission from [34]. Copyright by the American Physical Society

and width to the medium, so the total area required is no more than proportional to the square of the number of gates.

The "program" consists of down-moving y and z operator pulses, entering at the top with the down-moving data, and actuator pulses that enter from the left or right at two different speeds. In the frame moving with the data, the data and operator pulses are stationary and new results are deposited at the top of the memory column. In the laboratory frame the data pulses leave the medium downward, and new results appear in the medium at positions above the old data, at the positions of newly entering z pulses.

Discussion

We have shown that in principle any computation can be performed by shining time-gated lasers into a completely homogeneous nonlinear optical medium. This result should be viewed as mathematical, and whether the physics of vector soliton collisions can lead to practical computational devices is a subject for future study. With regard to the economy of the model, the question of whether time gating is necessary, or even whether two speeds are necessary, is open.

We note that the result described here differs from the universality results for the ideal billiard ball model [13], the Game of Life [7], and Lattice Gasses [32], for example, in that no internal mirrors or structures of any kind are used inside the medium. To the author's knowledge, whether internal structure is necessary in these other cases is open.

Finally, we remark that the model used is reversible and dissipationless. The fact that some of the gate operations realized are not in themselves reversible is not a contradiction, since extra, "garbage" solitons [13] are produced that save enough state to run the computation backwards.

Multistable Soliton Collision Cycles

Bistable and multistable optical systems, besides being of some theoretical interest, are of practical importance in offering a natural "flip-flop" for noise immune storage and logic. We show in this section that simple cycles of collisions of solitons governed by the Manakov equations can have more than one distinct stable set of polarization states, and therefore these distinct equilibria can, in theory, be used to store and process information. The multistability occurs in the polarization states of the beams; the solitons themselves do not change shape and remain the usual sech-shaped solutions of the Manakov equations. This phenomenon is dependent only on simple soliton collisions in a completely homogeneous medium.

The basic configuration considered requires only that the beams form a closed cycle, and can thus be realized in any nonlinear optical medium that supports spatial Manakov solitons. The possibility of using multistable systems of beam collisions broadens the possibilities for practical application of the surprisingly strong interactions that Manakov solitons can exhibit, a phenomenon originally described in [26]. We show here by example that a cycle of three collisions can have two distinct foci surrounded by basins of attractions, and that a cycle of four collisions can have three.

The Basic Three-Cycle and Computational Experiments

Figure 15 shows the simplest example of the basic scheme, a cycle of three beams, entering in states A, B, and C, with intermediate beams a, b, and c. For convenience, we will refer to the beams themselves, as well as their states, as A, B, C, etc. Suppose we start with beam C initially turned off, so that $A = a$. Beam a then hits B, thereby transforming it to state b. If beam C is then turned on, it will hit A, closing the cycle. Beam a is then changed, changing b, etc., and the cycle of state changes propagates clockwise. The question we ask is whether this cycle converges, and if so, whether

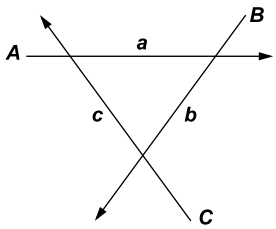

Computing with Solitons, Figure 15
The basic cycle of three collisions. Reprinted with permission from [35]. Copyright by the American Physical Society

it will converge with any particular choice of complex parameters to exactly zero, one, two, or more foci. We answer the question with numerical simulations of this cycle.

A typical computational experiment was designed by fixing the input beams A, B, C, and the parameters k_1 and k_2, and then choosing points a randomly and independently with real and imaginary coordinates uniformly distributed in squares of a given size in the complex plane. The cycle described above was then carried out until convergence in the complex numbers a, b, and c was obtained to within 10^{-12} in norm. Distinct foci of convergence were stored and the initial starting points a were categorized by which focus they converged to, thus generating the usual picture of basins of attraction for the parameter a. Typically this was done for 50,000 random initial values of a, effectively filling in the square, for a variety of parameter choices A, B, and C. The following results were observed:

- In cases with one or two clear foci, convergence was obtained in every iteration, almost always within one or two hundred iterations.
- Each experiment yielded exactly one or two foci.
- The bistable cases (two foci) are somewhat less common than the cases with a unique focus, and are characterized by values of k_R between about 3 and 5 when the velocity difference Δ was fixed at 2.

Figure 16 shows a bistable example, with the two foci and their corresponding basins of attraction. The parameter k is fixed in this and all subsequent examples at $4 \pm i$ for the right- and left-moving beams of any given collision, respectively. The second example, shown in Fig. 17, shows that the basins are not always simply connected; a sizable island that maps to the upper focus appears within the basin of the lower focus.

A Tristable Example Using a Four-Cycle

Collision cycles of length four seem to exhibit more complex behavior than those of length three, although it is dif-

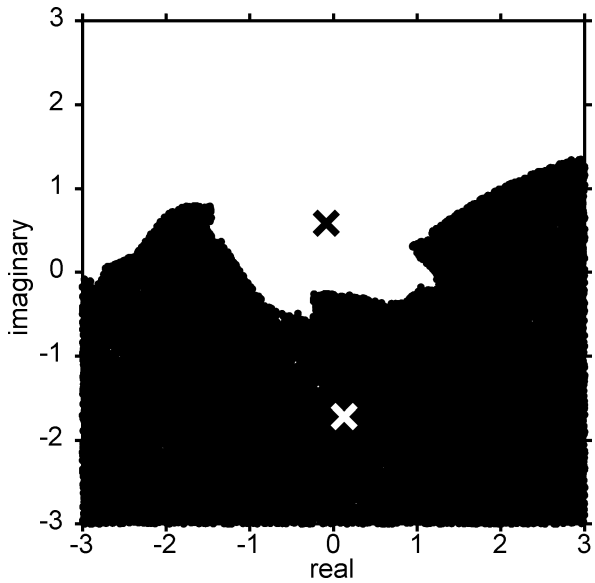

Computing with Solitons, Figure 16
The two foci and their corresponding basins of attraction in the first example, which uses a cycle of three collisions. The states of the input beams are $A = -0.8 - i \cdot 0.13$, $B = 0.4 - i \cdot 0.13$, $C = 0.5 + i \cdot 1.6$; and $k = 4 \pm i$. Reprinted with permission from [35]. Copyright by the American Physical Society

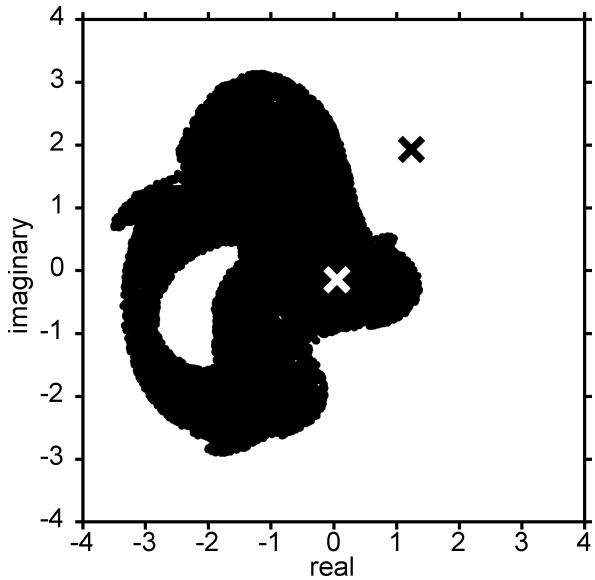

Computing with Solitons, Figure 17
A second example using a cycle of three collisions, showing that the basins need not be simply connected. The states of the input beams are $A = 0.7 - i \cdot 0.3$, $B = -1.1 - i \cdot 0.5$, $C = 0.4 + i \cdot 0.81$; and $k = 4 \pm i$. Reprinted with permission from [35]. Copyright by the American Physical Society

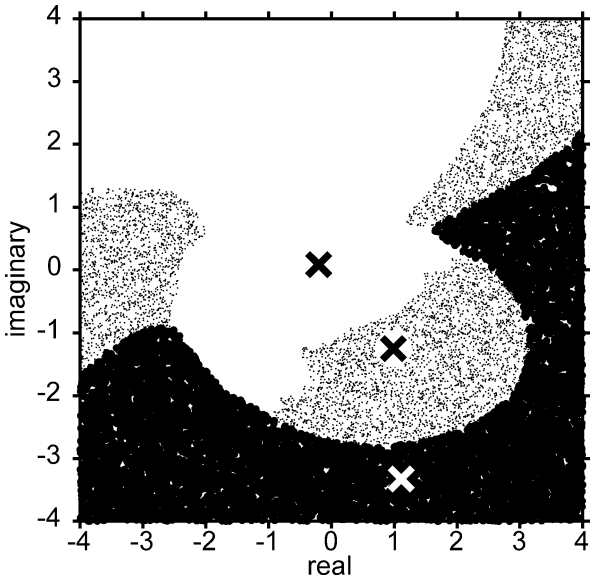

Computing with Solitons, Figure 18
A case with three stable foci, for a collision cycle of length four. The states of the input beams are $A = -0.39 - i \cdot 0.45$, $B = 0.22 - i \cdot 0.25$, $C = 0.0 + i \cdot 0.25$, $D = -0.51 + i \cdot 0.48$; and $k = 4 \pm i$. Reprinted with permission from [35]. Copyright by the American Physical Society

ficult to draw any definite conclusions because the parameter spaces are too large to be explored exhaustively, and there is at present no theory to predict such highly nonlinear behavior. If one real degree of freedom is varied as a control parameter, we can move from bistable to tristable solutions, with a regime between in which one basin of attraction disintegrates into many small separated fragments. Clearly, this model is complex enough to exhibit many of the well-known features of nonlinear systems.

Fortunately, it is not difficult to find choices of parameters that result in very well behaved multistable solutions. For example, Fig. 18 shows such a tristable case. The smallest distance from a focus to a neighboring basin is on the order of 25% of the interfocus distance, indicating that these equilibria will be stable under reasonable noise perturbations.

Discussion

The general phenomenon discussed in this section raises many questions, both of a theoretical and practical nature. The fact that there are simple polarization-multistable cycles of collisions in a Manakov system suggests that similar situations occur in other vector systems, such as photorefractive crystals or birefringent fiber. Any vector system with the possibility of a closed cycle of soliton collisions becomes a candidate for multistability, and there is at this point really no compelling reason to restrict attention to the Manakov case, except for the fact that the explicit state-change relations make numerical study much easier.

The simplified picture we used of information traveling clockwise after we begin with a given beam a gives us stable polarization states when it converges, plus an idea of the size of their basins of attractions. It is remarkable that in all cases in our computational experience, except for borderline transitional cases in going from two to three foci in a four-cycle, this circular process converges consistently and quickly. But understanding the actual dynamics and convergence characteristics in a real material requires careful physical modeling. This modeling will depend on the nature of the medium used to approximate the Manakov system, and is left for future work. The implementation of a practical way to switch from one stable state to another is likewise critically dependent on the dynamics of soliton formation and perturbation in the particular material at hand, and must be studied with reference to a particular physical realization.

We remark also that no iron-clad conclusions can be drawn from computational experiments about the numbers of foci in any particular case, or the number possible for a given size cycle—despite the fact that we regularly used 50,000 random starting points. On the other hand, the clear cases that have been found, such as those used as examples, are very characteristic of universal behavior in other nonlinear iterated maps, and are sufficient to establish that bi- and tristability, and perhaps higher-mode multistability, is a genuine mathematical characteristic, and possibly also physically realizable. It strongly suggests experimental exploration.

We restricted discussion in this section to the simplest possible structure of a single closed cycle, with three or four collisions. The stable solutions of more complicated configurations are the subject of continuing study. A general theory that predicts this behavior is lacking, and it seems at this point unlikely to be forthcoming. This forces us to rely on numerical studies, from which, as we point out above, only certain kinds of conclusions can be drawn. We are fortunate, however, in being able to find cases that look familiar and which are potentially useful, like the bistable three-cycles with well separated foci and simply connected basins of attraction.

It is not clear however, just what algorithms might be used to find equilibria in collision topologies with more than one cycle. It is also intriguing to speculate about how collision configurations with particular characteristics can be designed, how they can be made to interact, and how they might be controlled by pulsed beams. There is

promise that when the ramifications of complexes of vector soliton collisions are more fully understood they might be useful for real computation in certain situations.

Application to Noise-Immune Soliton Computing

Any physical instantiation of a computing technology must be designed to be immune from the effects of noise buildup from logic stage to logic stage. In the familiar computers of today, built with solid-state transistors, the noise-immunity is provided by physical state restoration, so that voltage levels representing logical "0" and "1" are restored by bistable circuit mechanisms at successive logic stages. This is state restoration at the physical level.

For another example, proposed schemes for quantum computing would be impractical without some means of protecting information stored in qubits from inevitable corruption by the rest of the world. The most common method proposed for accomplishing this is error correction at the software level, state restoration at the logical level.

In the collision-based scheme for computing with Manakov solitons described in Sect. "Manakov Soliton Computing", there is no protection against buildup of error from stage to stage, and some sort of logical state-restoration would be necessary in a practical realization. The bistable collision cycles of Manakov solitons described in this section, however, offer a natural computational building block for soliton computation with physical state restoration. This idea is explored in [27]. Figure 19 illustrates the approach with a schematic diagram of a NAND gate, implemented with bistable cycles to represent bits. The input bits are stored in the collision cycles (1) and (2), which have output beams that can be made to collide with input beam A of cycle (3), which represents the output bit of the gate. These inputs to the gate, shown as dashed lines, change the state of beam A of the ordinarily bistable cycle (3) so that it becomes *monostable*. The state of cycle (3) is then steered to a known state. When the input beams are turned off, cycle (3) returns to its normal bistable condition, but with a known input state. Its state then evolves to one of two bits, and the whole system of three collision cycles can be engineered so that the final state of cycle (3) is the NAND of the two bits represented by input cycles (1) and (2). (See [27] for details.)

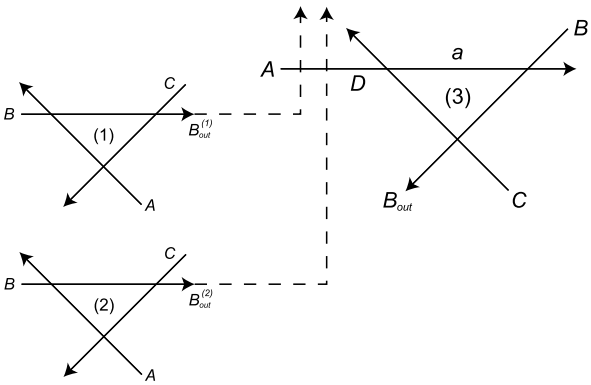

Computing with Solitons, Figure 19
Schematic of NAND gate using bistable collision cycles. Reprinted with permission from [27]. Copyright by Old City Publishing

A computer based on such bistable collision cycles is closer in spirit to present-day ordinary transistor-based computers, with a natural noise-immunity and state-restoration based on physical bistability. As mentioned in the previous subsection, however, the basic bistable cycle phenomenon awaits laboratory verification, and much remains to be learned about the dynamics, and eventual speed and reliability of such systems.

Experiments

The computation schemes described in the previous sections obviously rely on the correct mathematical modeling of the physics proposed for realization. We next describe experiments that verify some of the required soliton phenomenology in optical fibers. Specifically, we highlight the experimental observation of temporal vector soliton propagation and collision in a birefringent optical fiber [28]. This is both the first demonstration of temporal vector solitons with two mutually-incoherent component fields, and of vector soliton collisions in a Kerr nonlinear medium.

Temporal soliton pulses in optical fiber were first predicted by Hasegawa and Tappert [14], followed by the first experimental observation by Mollenauer et al. [24]. In subsequent work, Menyuk accounted for the birefringence in polarization maintaining fiber (PMF) and predicted that vector solitons, in which two orthogonally polarized components trap each other, are stable under the proper operating conditions [21,22]. For birefringent fibers, self-trapping of two orthogonally polarized pulses can occur when XPM-induced nonlinearity compensates the birefringence-induced group velocity difference, causing the pulse in the fiber's fast axis to slow down and the pulse in the slow axis to speed up. The first demonstration of temporal soliton trapping was performed in the sub picosecond regime [15], in which additional ultrashort pulse effects such as Raman scattering are present. In particular, this effect results in a red-shift that is linearly proportional to the propagation distance, as observed in a later temporal

soliton trapping experiment [25]. Recently, soliton trapping in the picosecond regime was observed with equal amplitude pulses [18]. However, vector soliton propagation could not be shown, because the pulses propagated for less than 1.5 dispersion lengths. In other work, phase-locked vector solitons in a weakly birefringent fiber laser cavity with nonlinear coherent coupling between components were observed [12].

The theoretical model for linearly birefringent fiber is the following coupled nonlinear Schrödinger equation (CNLSE):

$$i\left(\frac{\partial A_x}{\partial z}+\beta_{1x}\frac{\partial A_x}{\partial t}\right)-\frac{\beta_2}{2}\frac{\partial^2 A_x}{\partial t^2}+\gamma\left(|A_x|^2+\frac{2}{3}|A_y|^2\right)A_x = 0\,,$$
$$i\left(\frac{\partial A_y}{\partial z}+\beta_{1y}\frac{\partial A_y}{\partial t}\right)-\frac{\beta_2}{2}\frac{\partial^2 A_y}{\partial t^2}+\gamma\left(|A_y|^2+\frac{2}{3}|A_x|^2\right)A_y = 0\,, \quad (9)$$

where t is the local time of the pulse, z is propagation distance along the fiber, and $A_{x,y}$ is the slowly varying pulse envelope for each polarization component. The parameter $\beta_{1x,y}$ is the group velocity associated with each fiber axis, and β_2 represents the group velocity dispersion, assumed equal for both polarizations. In addition, we neglect higher order dispersion and assume a loss less medium with an instantaneous electronic response, valid for picosecond pulses propagating in optical fiber.

The last two terms of Eqs. (9) account for the nonlinearity due to SPM and XPM, respectively. In linearly birefringent optical fiber, a ratio of 2/3 exists between these two terms. When this ratio equals unity, the CNLSE becomes the integrable Manakov system of Eqs. (4). On the other hand, solutions of Eqs. (9) are, strictly speaking, solitary waves, not solitons. However, it was found in [36] that the family of symmetric, single-humped (fundamental or first-order) solutions, to which the current investigation in this section belongs, are all stable. Higher-order solitons, characterized by multiple humps, are unstable. Furthermore, it was shown in [37] that collisions of solitary waves in Eqs. (9) can be described by application of perturbation theory to the integrable Manakov equations, indicating the similarities between the characteristics of these two systems.

Experimental Setup and Design

The experimental setup is shown in Fig. 20. We synchronized two actively mode-locked erbium-doped fiber lasers (EDFLs)—EDFL1 at 1.25 GHz repetition rate, and EDFL2 at 5 GHz. EDFL2 was modulated to match with the lower

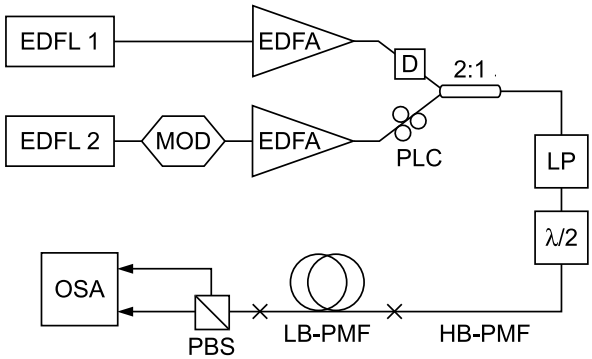

Computing with Solitons, Figure 20
Experimental setup. EDFL: Erbium-doped fiber laser; EDFA: Erbium-doped fiber amplifier; MOD: modulator; D: tunable delay line; PLC: polarization loop controller; 2:1: fiber coupler; LP: linear polarizer; λ/2: half-wave plate; HB-PMF and LB-PMF: high and low birefringence polarization maintaining fiber; PBS: polarization beam splitter; OSA: optical spectrum analyzer. Reprinted with permission from [28]. Copyright by the American Physical Society

repetition rate of EDFL1. Each pulse train, consisting of 2 ps pulses, was amplified in an erbium-doped fiber amplifier (EDFA) and combined in a fiber coupler. To align polarizations, a polarization loop controller (PLC) was used in one arm, and a tunable delay line (TDL) was employed to temporally align the pulses for collision. Once combined, both pulse trains passed through a linear polarizer (LP) and a half-wave plate to control the input polarization to the PMF. Approximately 2 m of high birefringence (HB) PMF preceded the specially designed 500 m of low birefringence (LB) PMF used to propagate vector solitons. Although this short length of HB-PMF will introduce some pulse splitting (on the order of 2–3 ps), the birefringent axes of the HB- and LB-PMF were swapped in order to counteract this effect. At the output, each component of the vector soliton was then split at a polarization beam splitter, followed by an optical spectrum analyzer (OSA) for measurement.

The design of the LB-PMF required careful control over three characteristic length scales: the (polarization) beat length, dispersion length L_d, and nonlinear length L_{nl}. A beat length $L_b = \lambda/\Delta n = 50$ cm was chosen at a wavelength of 1550 nm, where Δn is the fiber birefringence. According to the approximate stability criterion of [8], this choice allows stable propagation of picosecond vector solitons. By avoiding the sub picosecond regime, ultrashort pulse effects such as intrapulse Raman scattering will not be present. The dispersion $D = 2\pi c\beta_2/\lambda^2 = 16$ ps/km nm and $L_d = 2T_0^2/|\beta_2| \approx 70$ m, where $T_0 = T_{FWHM}/1.763$ is a characteristic pulse width related to the

full width at half maximum (FWHM) pulse width. Since $L_d \gg L_b$, degenerate four-wave mixing due to coherent coupling between the two polarization components can be neglected [23]. Furthermore, the total propagation distance is greater than 7 dispersion lengths.

Polarization instability, in which the fast axis component is unstable, occurs when $L_{nl} = (\gamma P)^{-1}$ is of the same order of magnitude or smaller than L_b, as observed in [6]. The nonlinearity parameter $\gamma = 2\pi n_2/\lambda A_{eff} = 1.3\ (\text{km W})^{-1}$, with Kerr nonlinearity coefficient $n_2 = 2.6 \times 10^{-20}\ \text{m}^2/\text{W}$ and measured effective mode area $A_{eff} = 83\ \mu\text{m}^2$. In the LB-PMF, the fundamental vector soliton power $P \approx 14$ W, thus $L_{nl} = 55\ \text{m} \gg L_b$, mitigating the effect of polarization instability.

Vector Soliton Propagation

We first studied propagation of vector solitons using both lasers independently. The wavelength shift for each component is shown in Fig. 21a as a function of the input polarization angle ϕ, controlled through the half-wave plate. Due to the anomalous dispersion of the fiber at this wavelength, the component in the slow (fast) axis will shift to shorter (longer) wavelengths to compensate the birefringence. The total amount of wavelength shift between components $\Delta\lambda_{xy} = \Delta\beta_1/D = 0.64$ nm, where $\Delta\beta_1 = |\beta_{1x} - \beta_{1y}| = 10.3$ ps/km is the birefringence-induced group velocity difference and dispersion $D = 2\pi c\beta_2/\lambda^2 = 16$ ps/km nm.

As ϕ approaches 0° (90°), the vector soliton approaches the scalar soliton limit, and the fast (slow) axis does not shift in wavelength, as expected. At $\phi = 45°$, a symmetric shift results. For unequal amplitude solitons, the smaller component shifts more in wavelength than the larger component, because the former experiences more XPM. Numerical simulations of Eqs. (9), given by the dashed lines of Fig. 21a, agree very well with the experimental results. Also shown in Fig. 21 are two cases, $\phi = 45°$ and 37°, as well as the numerical prediction. The experimental spectra show some oscillatory features at 5 GHz, which are a modulation of the EDFL2 repetition rate on the optical spectrum. A sample input pulse spectrum from EDFL1 is shown in the inset of Fig. 21, which shows no modulation due to the limited resolution of the OSA. Vector solitons from both lasers produced similar results. In this and all subsequent plots in this section, the slow and fast axis components are depicted by solid and dashed lines, respectively.

As the two component amplitudes become more unequal, satellite peaks become more pronounced in the smaller component. These features are also present in the simulations, but are not as dominant (cf. Fig. 21d and e). We attribute this to the input pulse, which is calibrated for the $\phi = 45°$ case, because the power threshold for vector soliton formation in this case is largest due to the 2/3 factor between SPM and XPM nonlinear terms in the CNLSE. As the input is rotated towards unequal components, there will be extra power in the input pulse, which will radiate in the form of dispersive waves as the vector soliton forms. Due to the nature of this system, these dispersive waves can be nonlinearly trapped, giving rise to the satellite features in the optical spectra. This effect is not as prevalent in the simulations because the threshold was numerically determined at each input angle ϕ.

Vector Soliton Collision

To prepare the experiment for a collision, we operated both lasers simultaneously, detuned in wavelength to allow for dispersion-induced walkoff, and adjusted the delay line in such a way that the collision occurred halfway down the fiber. We define a collision length $L_{coll} = 2T_{FWHM}/D\Delta\lambda$, where $\Delta\lambda$ is the wavelength separation between the two vector solitons. For our setup, $\Delta\lambda = 3$ nm, and $L_{coll} = 83.3$ m. An asymptotic theory of soliton collisions, in which a full collision takes place, requires at least 5 collision lengths. The total fiber length in this experiment is equal to 6 collision lengths, long enough to ensure sufficient separation of solitons before and after collision. In this way, results of our experiments can be compared to the asymptotic theory, even though full numerical simulations will be shown for comparison. To quantify our results, we introduce a quantity $R \equiv \tan^2\phi$, defined as the amplitude ratio between the slow and fast components.

Recall that in Sect. "Manakov Solitons", we introduced the Manakov equations (Eqs. (4)), and described collision-induced transformations of the polarization state of the soliton, which come about due to the asymptotic analysis of the soliton collision. The polarization state is the ratio between the two components $\rho \equiv A_x/A_y = \cot\phi \exp(i\Delta\theta)$, and is therefore a function of the polarization angle ϕ and the relative phase $\Delta\theta$ between the two components. In the context of the experiments described in this section, these state transformations (Eqs. (5) and (7)) predict that the resulting energy exchange will be a function of amplitude ratios $R_{1,2}$, wavelength separation $\Delta\lambda$, and the relative phase $\Delta\theta_{1,2}$ between the two components of each soliton, where soliton 1 (2) is the shorter (longer) wavelength soliton.

A word of caution is in order at this point. An interesting consequence of the 2/3 ratio between SPM and XPM, which sets the birefringent fiber model apart from

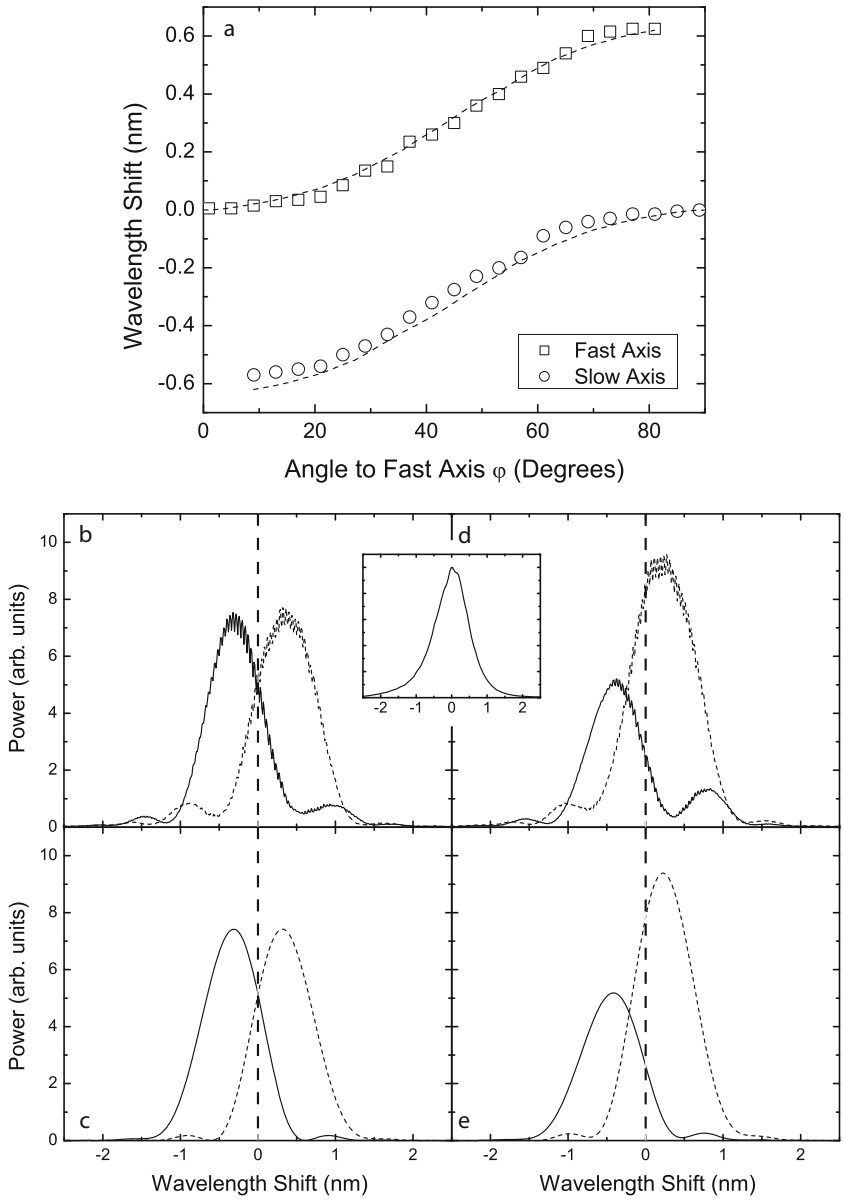

Computing with Solitons, Figure 21

Arbitrary-amplitude vector soliton propagation. a Wavelength shift vs. angle to fast axis ϕ, numerical curves given by dashed lines; b and d experimental results for $\phi = 45°$ and 37° with EDFL2, respectively. *Inset:* input spectrum for EDFL1; c and e corresponding numerical simulations of $\phi = 45°$ and 37°, respectively. The slow and fast axis components are depicted by solid and dashed lines, respectively. Reprinted with permission from [28]. Copyright by the American Physical Society

the Manakov model, is the relative phase between the two components. For the Manakov soliton, each component 'feels' the same amount of total nonlinearity, because the strengths of both SPM and XPM are equal. Therefore, regardless of the polarization angle, the amount of total nonlinear phase shift for each component is the same (even though the contributions of SPM and XPM phase shifts are in general not equal). As a result, the relative phase between the two components stays constant during propagation, as does the polarization state. This is *not* the case for vector solitons in birefringent fiber. For the case of equal amplitudes, each component does experience the same amount of nonlinear phase shift, and therefore the polarization state is constant as a function of propagation

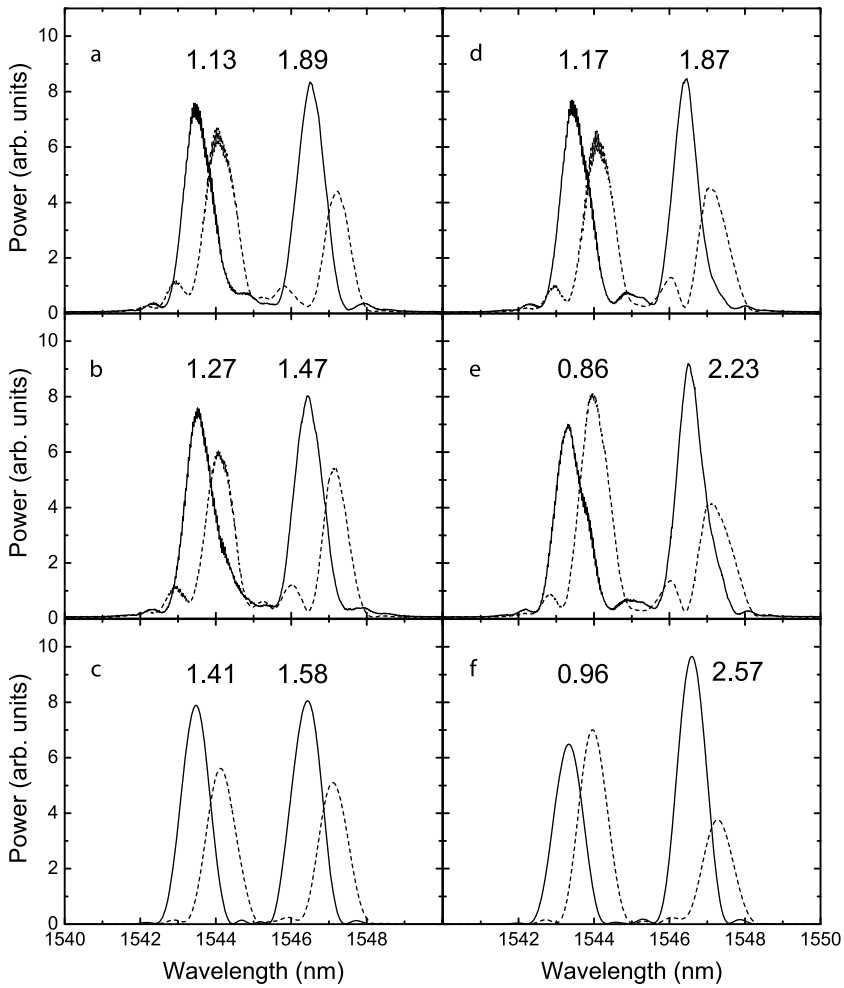

Computing with Solitons, Figure 22

Demonstration of phase-dependent energy-exchanging collisions. a–c Short HB-PMF; d–f long HB-PMF; a,d experiment, without collision; b, e experiment, with collision; c,f simulated collision result with c $\Delta\theta_2 = 90°$ and f $\Delta\theta_2 = 50°$. Values of slow-fast amplitude ratio *R* are given above each soliton. The slow and fast axis components are depicted by solid and dashed lines, respectively. Reprinted with permission from [28]. Copyright by the American Physical Society

distance. However, for arbitrary (unequal) amplitudes, the total phase shift for each component will be different. Consequently, the relative phase will change *linearly* as a function of propagation distance, and the polarization state will not be constant. As a result, the collision-induced change in polarization state, while being a function of the amplitude ratios $R_{1,2}$ and wavelength separation $\Delta\lambda$, will also depend upon the collision position due to the propagation dependence of the relative phase $\Delta\theta_{1,2}(z)$. To bypass this complication, we ensure that all collisions occur at the same spatial point in the fiber.

Because only one half-wave plate is used in our experiment (see Fig. 20), it was not possible to prepare each vector soliton individually with an arbitrary R. In addition, due to the wavelength dependence of the half-wave plate, it was not possible to adjust $\Delta\lambda$ without affecting R.

First, we investigated the phase dependence of the collision. This was done by changing the length of the HB-PMF entering the LB-PMF, while keeping R and $\Delta\lambda$ constant. As a result, we could change $\Delta\theta_{1,2}$ due to the birefringence of the HB-PMF. Approximately 0.5 m of HB-PMF was added to ensure that the total amount of temporal pulse splitting did not affect the vector soliton formation. The results are shown in Fig. 22, where Fig. 22a–c and d–f correspond to the short and long HB-PMFs, respectively. Figure 22a and d show the two vector solitons, which propagate independently when no collision occurs; as expected, the two results are similar because the OSA

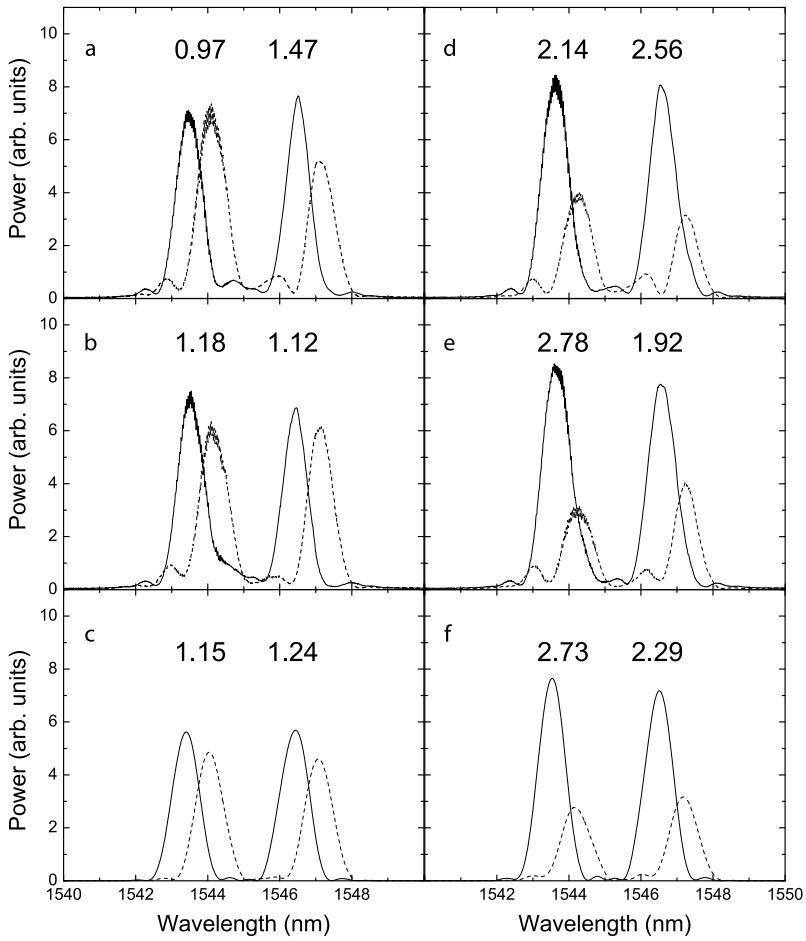

Computing with Solitons, Figure 23

Additional energy-exchanging collisions. a,d Experiment, without collision; b,e experiment, with collision; c,f simulated collision result, using $\Delta\theta_2 = 90°$ inferred from the experiment of Fig. 22. Values of slow-fast amplitude ratio *R* are given above each soliton. The slow and fast axis components are depicted by solid and dashed lines, respectively. Reprinted with permission from [28]. Copyright by the American Physical Society

measurement does not depend on $\Delta\theta_{1,2}$. The result of the collision is depicted in Fig. 22b and e, along with the corresponding simulation results in Fig. 22c and f.

In both of these collisions, an energy exchange between components occurs, and two important relations are satisfied: the total energy in each soliton and in each component is conserved. It can be seen that when one component in a soliton increases as a result of the collision, the other component decreases, with the opposite exchange in the second soliton. The difference between these two collisions is dramatic, in that the energy redistributes in opposite directions. For the simulations, idealized sech pulses for each component were used as initial conditions, and propagation was modeled without accounting for losses. The experimental amplitude ratio was used, and (without loss of generality [16,19,26]) $\Delta\theta_2$ was varied while $\Delta\theta_1 = 0$. Best fits gave $\Delta\theta_2 = 90°$ (Fig. 22c) and 50° (Fig. 22f). Despite the model approximations, experimental and numerical results all agree to within 15%.

In the second set of results (Fig. 23), we changed R while keeping all other parameters constant. More specifically, we used the short HB-PMF, with initial phase difference $\Delta\theta_2 = 90°$, and changed the amplitude ratio. In agreement with theoretical predictions, the same direction of energy exchange is observed as in Fig. 22a–c.

Spatial Soliton Collisions

We mention here analogous experiments with spatial solitons in photorefractive media by Anastassiou et al. In [4], it is shown that energy is transferred in a collision of vector spatial solitons in a way consistent with the predictions

for the Manakov system (although the medium is a saturable one, and only approximates the Kerr nonlinearity). The experiment in [5] goes one step farther, showing that one soliton can be used as an intermediary to transfer energy from a second soliton to a third. We thus are now at a point where the ability of both temporal and spatial vector solitons to process information for computation has been demonstrated.

Future Directions

This article discussed computing with solitons, and attempted to address the subject from basic physical principles to applications. Although the nonlinearity of fibers is very weak, the ultralow loss and tight modal confinement make them technologically attractive. By no means, however, are they the only potential material for soliton-based information processing. Others include photorefractive crystals, semiconductor waveguides, quadratic media, and Bose–Einstein condensates, while future materials research may provide new candidate systems.

From a computing perspective, scalar soliton collisions are insufficient. Although measurable phase and position shifts do occur, these phenomena cannot be cascaded to affect future soliton collisions and therefore cannot transfer information from one collision to the next. Meaningful computation using soliton collisions requires a new degree of freedom; that is, a new component. Collisions of vector solitons display interesting energy-exchanging effects between components, which can be exploited for arbitrary computation and bistability.

The vector soliton experiments of Sect. "Experiments" were proof-of-principle ones. The first follow-up experiments with temporal vector solitons in birefringent fiber can be directed towards a full characterization of the collision process. This can be done fairly simply using the experimental setup of Fig. 20 updated in such a way as to allow independent control of two vector soliton inputs. This would involve separate polarizers and half-waveplates, followed by a polarization preserving fiber coupler.

Cascaded collisions of temporal solitons also await experimental study. As demonstrated in photorefractive crystals with a saturable nonlinearity [5], one can show that information can be passed from one collision to the next. Beyond a first demonstration of two collisions is the prospect of setting up a multi-collision feedback cycle. Discussed in Sect. "Multistable Soliton Collision Cycles", these collision cycles can be bistable and lead to interesting applications in computation.

Furthermore, the recent work of Ablowitz et al. [2] shows theoretically that the useful energy-redistribution properties of vector soliton collisions extend perfectly to the semi-discrete case: that is, to the case where space is discretized, but time remains continuous. This models, for example, propagation in an array of coupled nonlinear waveguides [10]. The work suggests alternative physical implementations for soliton switching or computing, and also hints that the phenomenon of soliton information processing is a very general one.

Bibliography

1. Ablowitz MJ, Prinari B, Trubatch AD (2004) Soliton interactions in the vector nls equation. Inverse Problems 20(4):1217–1237
2. Ablowitz MJ, Prinari B, Trubatch AD (2006) Discrete vector solitons: Composite solitons, Yang–Baxter maps and computation. Studies in Appl Math 116:97–133
3. Agrawal GP (2001) Nonlinear Fiber Optics, 3rd edn. Academic Press, San Diego
4. Anastassiou C, Segev M, Steiglitz K, Giordmaine JA, Mitchell M, Shih MF, Lan S, Martin J (1999) Energy-exchange interactions between colliding vector solitons. Phys Rev Lett 83(12):2332–2335
5. Anastassiou C, Fleischer JW, Carmon T, Segev M, Steiglitz K (2001) Information transfer via cascaded collisions of vector solitons. Opt Lett 26(19):1498–1500
6. Barad Y, Silberberg Y (1997) Phys Rev Lett 78:3290
7. Berlekamp ER, Conway JH, Guy RK (1982) Winning ways for your mathematical plays, Vol. 2. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], London
8. Cao XD, McKinstrie CJ (1993) J Opt Soc Am B 10:1202
9. Chen ZG, Segev M, Coskun TH, Christodoulides DN (1996) Observation of incoherently coupled photorefractive spatial soliton pairs. Opt Lett 21(18):1436–1438
10. Christodoulides DN, Joseph RI (1988) Discrete self-focusing in nonlinear arrays of coupled waveguides. Opt Lett 13:794–796
11. Christodoulides DN, Singh SR, Carvalho MI, Segev M (1996) Incoherently coupled soliton pairs in biased photorefractive crystals. Appl Phys Lett 68(13):1763–1765
12. Cundiff ST, Collings BC, Akhmediev NN, Soto-Crespo JM, Bergman K, Knox WH (1999) Phys Rev Lett 82:3988
13. Fredkin E, Toffoli T (1982) Conservative logic. Int J Theor Phys 21(3/4):219–253
14. Hasegawa A, Tappert F (1973) Transmission of stationary nonlinear optical pulses in dispersive dielectric fibers 1: Anomalous dispersion. Appl Phys Lett 23(3):142–144
15. Islam MN, Poole CD, Gordon JP (1989) Opt Lett 14:1011
16. Jakubowski MH, Steiglitz K, Squier R (1998) State transformations of colliding optical solitons and possible application to computation in bulk media. Phys Rev E 58(5):6752–6758
17. Kang JU, Stegeman GI, Aitchison JS, Akhmediev N (1996) Observation of Manakov spatial solitons in AlGaAs planar waveguides. Phys Rev Lett 76(20):3699–3702
18. Korolev AE, Nazarov VN, Nolan DA, Truesdale CM (2005) Opt Lett 14:132
19. Manakov SV (1973) On the theory of two-dimensional stationary self-focusing of electromagnetic waves. Zh Eksp Teor Fiz 65(2):505–516, [Sov. Phys. JETP 38, 248 (1974)]
20. Mano MM (1972) Computer Logic Design. Prentice-Hall, Englewood Cliffs

21. Menyuk CR (1987) Opt Lett 12:614
22. Menyuk CR (1988) J Opt Soc Am B 5:392
23. Menyuk CR (1989) Pulse propagation in an elliptically birefringent Kerr medium. IEEE J Quant Elect 25(12):2674–2682
24. Mollenauer LF, Stolen RH, Gordon JP (1980) Experimental observation of picosecond pulse narrowing and solitons in optical fibers. Phys Rev Lett 45(13):1095–1098
25. Nishizawa N, Goto T (2002) Opt Express 10:1151-1160
26. Radhakrishnan R, Lakshmanan M, Hietarinta J (1997) Inelastic collision and switching of coupled bright solitons in optical fibers. Phys Rev E 56(2):2213–2216
27. Rand D, Steiglitz K, Prucnal PR (2005) Signal standardization in collision-based soliton computing. Int J of Unconv Comp 1:31–45
28. Rand D, Glesk I, Brès CS, Nolan DA, Chen X, Koh J, Fleischer JW, Steiglitz K, Prucnal PR (2007) Observation of temporal vector soliton propagation and collision in birefringent fiber. Phys Rev Lett 98(5):053902
29. Russell JS (1844) Report on waves. In: Report of the 14th meeting of the British Association for the Advancement of Science, Taylor and Francis, London, pp 331–390
30. Shih MF, Segev M (1996) Incoherent collisions between two-dimensional bright steady-state photorefractive spatial screening solitons. Opt Lett 21(19):1538–1540
31. Shor PW (1994) Algorithms for quantum computation: Discrete logarithms and factoring. In: 35th Annual Symposium on Foundations of Computer Science, IEEE Press, Piscataway, pp 124–134
32. Squier RK, Steiglitz K (1993) 2-d FHP lattice gasses are computation universal. Complex Systems 7:297–307
33. Steblina VV, Buryak AV, Sammut RA, Zhou DY, Segev M, Prucnal P (2000) Stable self-guided propagation of two optical harmonics coupled by a microwave or a terahertz wave. J Opt Soc Am B 17(12):2026–2031
34. Steiglitz K (2000) Time-gated Manakov spatial solitons are computationally universal. Phys Rev E 63(1):016608
35. Steiglitz K (2001) Multistable collision cycles of Manakov spatial solitons. Phys Rev E 63(4):046607
36. Yang J (1997) Physica D 108:92–112
37. Yang J (1999) Multisoliton perturbation theory for the Manakov equations and its applications to nonlinear optics. Phys Rev E 59(2):2393–2405
38. Zakharov VE, Shabat AB (1971) Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media. Zh Eksp Teor Fiz 61(1):118–134, [Sov. Phys. JETP 34, 62 (1972)]

Conduction and Diffusion in Percolating Systems

Barry D. Hughes
Department of Mathematics and Statistics,
University of Melbourne, Melbourne, Australia

Article Outline

Glossary
Definition of the Subject
Introduction
Continuum Models: Steady-State Phenomena
Lattice Models: Steady-State Phenomena
Random Motion in a Random Environment
Future Directions
Bibliography

Glossary

Backbone In a lattice percolation problem above the percolation threshold, that fraction of the infinite cluster with two disjoint connections to infinity.

Correlation length (ξ) Length scale in a randomly structured system over which the system cannot be regarded as homogeneous.

Critical exponent Exponent characterizing the dominant behavior of an observable quantity near a percolation threshold; e. g. $\xi \sim \text{constant} \times (p_c - p)^{-\nu}$ as $p \uparrow p_c$.

Effective medium approximation Approximate description of an inhomogeneous system obtained by matching averaged local fluctuations in properties in a self-consistent manner.

Flux Vector-valued function (in a continuum) or signed scalar (in a discrete system) quantifying transport or conduction rates.

Lattice Discrete structure (network/graph) of sites (nodes/vertices) connected by bonds (links/edges), including periodic lattices, random lattices, and tree-like or self-similar pseudolattices.

Percolation theory Idealized model of a random medium. In the classical discrete case, the bonds of a lattice are independently open with probability p (Bernoulli bond percolation) or the sites of a lattice are independently occupied with probability p (Bernoulli site percolation). There are various continuum analogues.

Percolation threshold (p_c) Dividing point in parameter space separating cases where long-range connectivity is precluded (infinite connected sets exist with probability 0) from those where long-range connectivity occurs (infinite connected sets exist with positive probability).

Percolative system Random two-phase continuous or discrete system in which one phase is deemed void or non-conducting; usually a percolation threshold exists in such systems.

Potential (V) Function distributed over space or over the sites of a network from which steady-state transport or conduction may be determined.

Pseudolattice A non-periodic discrete structure of sites (nodes/vertices) and bonds (links/edges), most commonly either topologically tree-like or geometrically self-similar.

Random walk Model of random motion, especially on lattices, consisting of a sequence of steps separated by constant or random time intervals.

Recurrent random walk Random walk process on a lattice for which the walker returns to the starting site with probability 1.

Renormalization Converting a system, either exactly or approximately, to a related system with a different characteristic length scale.

Transient random walk Random walk process on a lattice for which the walker has probability less than 1 of returning to the starting site.

Definition of the Subject

The problem of determining the macroscopic structural and transport properties of microscopically non-uniform materials has a long history and is of such central importance that it finds applications in an astonishingly wide range of areas of science and technology, from developmental biology to xerography [109,140,141]. Consider systems that consist of two phases, each of which is homogeneous in its properties, mixed in some way to create an inhomogeneous material. Let one phase occupy a fraction ϕ of the volume. If $\phi \ll 1$, so that the inhomogeneous structure is in some sense dilute, the possibility arises of determining the effective properties of the material as an expansion in powers of ϕ. Simple examples of this, when the dilute phase consists of identical spheres distributed in some reasonable manner, are the prediction variously associated with the names of Maxwell, Clausius, Mossotti, Lorenz, Lorentz and others that spheres of dielectric permittivity ϵ_1 embedded in a matrix of permittivity ϵ_0 produce an effective permittivity ϵ^* given by

$$\frac{\epsilon^* - \epsilon_0}{\epsilon^* + 2\epsilon_0} \approx \frac{\epsilon_1 - \epsilon_0}{\epsilon_1 + 2\epsilon_0}\phi\,; \tag{1}$$

and the prediction of Einstein [43,44] that the effective viscosity η^* of a random suspension of rigid, spheres occupying a volume fraction ϕ in a liquid of viscosity η is given by $\eta^* = \eta\big[1 + \frac{5}{2}\phi + o(\phi)\big]$ as $\phi \to 0$. Analogous calculations can be performed for many other "dilute" systems, leading to predictions at low volume fraction for effective electrical conductivity and other attributes. Simple examples demonstrate that as the volume fraction ϕ increases, quantitative information on exactly how the phases are distributed becomes essential for accurate prediction of, or even construction of decent bounds for, effective properties—knowing the volume fractions of the phases is not enough.

We address inhomogeneous systems and especially two-phase systems—both continua and lattices—that are randomly structured, with particular emphasis on systems in which only one phase (with relative abundance ϕ) sustains transport. The simplest such models arise by defining steady-state transport processes or random motions on the occupied phase in the lattice percolation model (► Percolation Phase Transition) of Broadbent and Hammersley [25,62,63,64] or its continuum analogues. (► Continuum Percolation). We call all such models *percolative models.* They exhibit threshold behavior at $\phi = \phi_c$, separating globally nonconducting states when the transport-sustaining phase is sufficiently rare ($0 \le \phi < \phi_c$) from conducting states when the transport-sustaining phase is sufficiently abundant ($\phi_c < \phi \le 1$). When the transport-sustaining phase only sparsely spans a large region, that is, just above the percolation threshold ϕ_c, asymptotic power-law dependence of the transport properties on $\phi - \phi_c$ is observed, but many results now deemed well-known still evade rigorous proof.

Introduction

We address problems of transport and conduction, both in continua and on lattices, where there are randomly distributed local transport properties, especially the cases where there are two phases present (for continua) or there are bonds or sites of two types present (for lattices). Although we later pay most attention to the lattice case, for the present and in Sect. "Continuum Models: Steady-State Phenomena" we discuss primarily continua in which two material phases with different properties are present. Frequently, one of these phases is empty space (void space). Within regions comprising only one phase, all properties of the system are uniform. The properties of regions containing both phases depend not only on the relative proportions of each phase present, but also on the way the phases are distributed. The fundamental question to be answered is this: if we pretend that the system is homogeneous, as it might indeed appear if we viewed a very large piece of it from a distance, what are the effective transport properties of the system?

A complete solution of the problem is not to be expected if the microstructure is elaborate or especially subtle, but sufficient progress has been made since the 1950s that for a number of conceptually simple models, some exact results and many approximate results of decent quality are available. The two cases amenable to treatment are periodic microstructure and random microstructure. The former case, which is not our present concern, presents few conceptual challenges and is increasingly amenable

to numerical calculation, since the determination of the global properties of the system can usually be reduced to a study of a single finite region (the fundamental repeat unit) and this is usually computationally tractable.

Interest in modeling systems in which small (usually identical) pieces of one phase are randomly distributed in some way and the remaining phase occupies the rest of space goes back at least to James Clerk Maxwell (1831–1879), Ludvig Valentin Lorenz (1829–1891) and Hendrik Antoon Lorentz (1853–1928). The work of these authors in the latter half of the nineteenth century and some antecedents, culminating in results such as Eq. (1), is reviewed by Landauer [94], Markov [102] and Milton [109]. In the first two decades of the twentieth century the problem was addressed by Einstein [43,44] and others, but four major conceptual developments were needed to bring the subject beyond its infancy.

(a) In 1935 Bruggeman [27], building on ideas of earlier authors but arguing with much greater clarity, developed an *effective medium approximation*, in which an unknown overall property of the composite system, such as its effective conductivity σ_{eff}, is computed in an approximate but self-consistent manner that takes account of fluctuations in the corresponding local property. In the extreme case in which one phase is nonconducting, Bruggeman's approach predicts that for sufficiently small volume fraction ϕ of the conducting phase, there is no conductivity, but there is a critical volume fraction ϕ_c of the conducting phase above which there is conduction, with effective conductivity $\sigma_{\text{eff}} \sim \text{constant} \times (\phi - \phi_c)$ as $\phi \downarrow \phi_c$. Bruggeman's original treatment pays no attention to the manner in which the phases are randomly or periodically distributed.

(b) In 1965, Beran [11,12] established the foundations of a statistical study of effective properties of random media, in which the random placement of phases is statistically quantified. In particular, Beran brought out the importance of considering individual realizations of a random system equipped with a proper probability structure, and drew attention to the important issue of ergodicity, which in this context concerns the relation between volume averages (averages of properties over large volumes in a single realization) and ensemble averages (averages over many realizations).

(c) In 1957, Broadbent and Hammersley [25,62,63,64] considered a lattice model of random media, motivated by a problem of gas mask design for British coal miners. Although quantified transport was not part of their original model—they only addressed connectivity—this work is at the conceptual heart of modern discussions. In their model, now known as, the *percolation theory* lattice is randomly stripped so that either individual bonds remain with probability p (bond percolation) or individual sites remain with probability p (site percolation); the parameter p is the analogue of volume fraction ϕ in a two-phase random continuum, with what remains of the lattice viewed as a conducting phase in the original system. There is a well-defined relative abundance p of the conducting phase, called the *percolation threshold* p_c, above which there is long-distance connectivity and transport is possible. Although the percolation threshold plays for the lattice model a role analogous to the critical volume fraction predicted by Bruggeman's effective medium approximation (and indeed effective medium approximations can be set up for lattice systems), the existence of a precise percolation threshold p_c was rigorously established in 1957, and its value is known exactly for some lattices and approximately but to very high precision for other lattices. The probability that a given site belongs to an infinite connected cluster of the active phase when that phase has relative abundance p is the percolation probability $P_\infty(p)$, and it is accepted that $P_\infty(p) \sim \text{constant} \times (p - p_c)^\beta$ as $p \downarrow p_c$, with $\beta = 5/36$ in simple two-dimensional lattice systems and $\beta \leq 1$ more generally. A number of the techniques and concepts that have developed over 50 years for lattice-based percolation theory have now found their way into natural continuous-medium analogues [104]. Both the lattice and continuum percolation models have many applications [109,138,140,166].

(d) That percolation ideas would be significant in physics was realized almost immediately [5] following the 1957 paper of Broadbent and Hammersely. The final conceptual step needed was first articulated in 1968 by Ziman [174]: the stochastic geometry underlying percolation theory, though essential to the description of transport in random media, does not completely characterize the transport properties: there is also the specific contribution of the transport mechanism. The natural attempt to connect a transport coefficient such as effective conductivity $\sigma_{\text{eff}}(p)$ to the percolation probability $P_\infty(p)$ by writing $\sigma_{\text{eff}}(p) \propto P_\infty(p)$ [42] was refuted by experiments in 1971 of Last and Thouless [95] on the conductivity of a sheet of colloidal graphite with holes randomly punched in it in a manner appropriate to simulate site percolation on the square lattice. The experiments suggested that $\sigma_{\text{eff}}(p) \sim \text{constant} \times (p - p_c)^t$ as $p \downarrow p_c$, with $t > 1 \gg \beta$. The subtlety of the problem arises from the fact that just above the percolation threshold, the sample-spanning structure is tenuous and tortuous, with significant implications on observable quantities that characterize the transport process.

The developments stemming from the approximate work of Bruggeman and the rigorous work of Beran [(a) and (b) above] are well covered in the expansive texts of Milton [109] and Torquato [166]. The bulk of the present article is devoted to random medium problems—both continuous and discrete, though favoring the latter—with a percolative aspect [(c) and (d) above]. Steady-state continuum problems are discussed first in Sect. "Continuum Models: Steady-State Phenomena", with the physical contexts addressed in Sect. "Contexts". These contexts are also claimed by advocates of the analogous lattice models discussed in Sect. "Lattice Models: Steady-State Phenomena", to which we pay more attention. The model on which we primarily focus is the *random resistor network* (introduced in Sect. "The Random Resistor Problem"), which reveals most of the key features of random media, especially the existence of percolation thresholds and the subtlety of the active paths in the system close to the percolation threshold. Although many important and elegant results are now available for geometrical and topological aspects of the underlying percolation process, as discussed in the books by Bollobas and Riordan [21], Grimmett [59] and Hughes [75], and ► Percolation Phase Transition, considerable work is still needed to resolve major questions concerning the random resistor network and related systems.

We do not consider the mechanical properties (such as elastic modulii or fracture resistance) of random media, although many of the ideas and results presented here have been extended to mechanical properties in the literature. Less restricted surveys of the properties of random media can be found in the major texts of Milton [109], Sahimi [140,141] and Torquato [166]. Time-evolving problems in randomly structured media are of great interest, but our discussion in Sect. "Random Motion in a Random Environment" strongly emphasizes the model of random walk on a random lattice derived from a percolation process, poetically described as the *ant in the labyrinth* [24,33,75].

Continuum Models: Steady-State Phenomena

Despite the enormous efforts expended on discrete inhomogeneous and percolative systems, most scientific applications that motivate these studies arise in materials perhaps most naturally treated as continua. We briefly summarize several contexts in which percolative systems arise and review selected findings for continua that expose major concepts later discussed more extensively for the more tractable discrete analogues. Examples considered

(i) are scalar, so that we have scalar transport coefficients rather than tensors;
(ii) involve real potentials, so we consider only steady direct current electrical conduction, rather than frequency-dependent alternating current conduction;
(iii) have no explicit time dependence, and
(iv) have linear constitutive response.

The reader may refer to the expansive texts of Milton [109] and Sahimi [140,141] for discussions without the restrictions (i)–(iv).

Contexts

Problems Equivalent to Electrical Conduction In a variety of physical contexts the following mathematical problem arises. Let Ω be a domain (a connected spatial region), let $\sigma(\mathbf{r})$ (which we shall call a *transport coefficient*) be a prescribed function of position $\mathbf{r}$ in Ω, and let ∇ denote the usual gradient operator. We are to find a *potential* $V(\mathbf{r})$ that satisfies the equation

$$\nabla \cdot (\sigma \nabla V) = 0, \quad \mathbf{r} \in \Omega \,, \tag{2}$$

subject to prescribed conditions on the boundary $\partial\Omega$ of Ω, usually either the Dirichlet boundary condition [$V(\mathbf{r})$ prescribed], the Neumann [$\mathbf{n} \cdot \nabla V$ prescribed, where $\mathbf{n}$ is a unit vector normal to $\partial\Omega$ and, for consistency, the surface integral of $\mathbf{n} \cdot \nabla V$ over $\partial\Omega$ is zero], or Dirichlet and Neumann conditions applied to disjoint components of $\partial\Omega$. The potential V, which is uniquely defined for the Dirichlet boundary-value problem and is unique up to an additive constant for the pure Neumann boundary-value problem, is associated with a vector field

$$\mathbf{E} = -\nabla V \,, \tag{3}$$

and the field $\mathbf{E}$ and the transport coefficient σ determine a flux vector

$$\mathbf{J} = \sigma \mathbf{E} \,. \tag{4}$$

Equation (2) can be interpreted as a statement of a conservation law under steady-state conditions for a substance carried by the flux vector $\mathbf{J}$. Where the transport coefficient has surfaces of discontinuity (as is the case in two-phase media), one interprets the partial differential equations as holding in the distributional sense and the boundary conditions of continuity of V and of $\mathbf{n} \cdot \sigma \nabla V$ (where $\mathbf{n}$ is normal to the phase interface) follow.

The canonical example of the problem embodied in Eqs. (2)–(4) is steady state (direct current) electrical conduction, with V the electrostatic potential, $\mathbf{E}$ the electric field, $\mathbf{J}$ the electric current, and σ the electrical conductivity. Six different interpretations of the same mathematical

Conduction and Diffusion in Percolating Systems, Table 1
Mathematically equivalent potential theory problems

Transport coefficient	Flux law	Potential–field relation
Electrical conductivity σ	$\mathbf{J} = \sigma\mathbf{E}$	$\mathbf{E} = -\nabla V$ from Maxwell's $\nabla \times \mathbf{E} = \mathbf{0}$
Dielectric permittivity ϵ	$\mathbf{D} = \epsilon\mathbf{E}$	$\mathbf{E} = -\nabla V$ from Maxwell's $\nabla \times \mathbf{E} = \mathbf{0}$
Magnetic permeability μ	$\mathbf{B} = \mu\mathbf{H}$	$\mathbf{H} = -\nabla V$ from Maxwell's $\nabla \times \mathbf{H} = \mathbf{0}$

Transport coefficient	Flux–potential relation	
Thermal conductivity κ	Fourier's Law (temperature T)	$\mathbf{Q} = -\kappa\nabla T$
Diffusivity D	Fick's Law (pressure P, viscosity η)	$\mathbf{q} = -(k/\eta)\nabla P$
Permeability k	Darcy's Law (concentration c)	$\mathbf{J} = -D\nabla c$

problem are given in Table 1. For the first three interpretations, the potential V is the consequence of a fundamental physical equation of the form $\nabla \times \mathbf{E} = \mathbf{0}$. In the remaining examples in Table 1), the flux law $\mathbf{J} = \sigma\mathbf{E}$ and the potential–field relation $\mathbf{E} = -\nabla V$ do not have individual fundamental physical interpretations.

Caveats to the Electrical Interpretation In our discussion we shall use the terminology appropriate to electrical conductivity, however in interpreting the results in other contexts several caveats are needed. For time-dependent diffusive processes, we really wish to solve the equation

$$\frac{\partial c}{\partial t} = \nabla \cdot (D\nabla c)\,; \tag{5}$$

the electrical conduction analogue applies only to the steady state. Similar issues arise for heat conduction. For diffusion, problems in which the tracer is injected at an isolated point (a point source) are of great interest. In the extreme case in which one phase sustains diffusion and the other does not, the effect of placing the source at an arbitrary point in the phase that sustains diffusion depends critically on whether the source falls in a conducting region of finite extent or of infinite extent and these issues are not covered by a discussion confined to the electrical conduction analogue.

Although we have included porous medium permeability in Table 1, this exact equivalence to the electrical conduction problem only applies when we work at length scales large compared to pore sizes, so that the notion of permeability—whether constant or variable—is meaningful. A problem of great interest arises when one considers fluid flow through the void space of a medium consisting of an impermeable solid phase and a void phase [139]. In this case, there is no local partial differential equation at all for the solid phase, while in the void space the fluid is subject to the low Reynolds number limit of the Navier–Stokes equations.

The Idea of Homogenization

If the local conductivity σ is bounded, but fluctuates in some manner, one may seek a "homogenized" description of the medium, replacing the variable conductivity problem $\nabla \cdot (\sigma\nabla V) = 0$ by a uniform conductivity problem $\nabla \cdot (\sigma_{\text{eff}}\nabla V) = 0$ that matches it in some sense that needs to be made precise, with the constant *effective conductivity* σ_{eff} to be determined. A porous medium, viewed as a two-phase void/solid system as mentioned above, homogenizes in a different manner, since the equations for the desired uniform system, embodying Darcy's Law, are structurally different from the equations that govern the flow in the pore space [139]. We do not discuss this case here.

There are two basic rigorous approaches. One, on which we focus in the present article, is introduced in Sect. "Effective Conductivities". The other, which is generally described as *homogenization theory*, may be briefly summarized as follows. There are three length scales in the problem: a microscale, on which the structure fluctuates (perhaps strongly), a mesoscale over which unstructured continuum behavior emerges, and a macroscale over which parameters defined at the mesoscale may vary slowly. One may define volume averages $\langle\mathbf{E}\rangle$ and $\langle\mathbf{J}\rangle$ of the electric field over some representative domain large compared to the length scale on which microstructure fluctuates (that is, a mesoscale domain) and introduce a homogenized conductivity $\widehat{\sigma}$ defined by

$$\langle\mathbf{J}\rangle = \widehat{\sigma}\langle\mathbf{E}\rangle\,. \tag{6}$$

Homogenization is successful if, in an appropriate limit as the microscale dimension is sent to zero, a well-defined $\widehat{\sigma}$ emerges. The theory of homogenization has been well worked out for periodic microstructures [4,10], which are not our concern here, but is more challenging for the random microstructure case, although considerable progress has also been made there [56,79]; surveys of recent work will be found in the texts of Milton [109] and

Torquato [166]. The first strong results were derived for uniformly elliptic partial differential operators, which in the present context means that for some constants α and β we have $0 < \alpha \leq \sigma \leq \beta < \infty$, precluding the analysis of percolative systems, but extensions to percolative systems are discussed in the book of Jikov, Kozlov and Oleinik [79]. Percolative systems are subtle because the length scale on which the inhomogeneity is important—the correlation length—diverges at the percolation threshold; and the detailed geometrical arrangement of the conducting and non-conducting phases is very important except perhaps for very low or very high volume fractions of the conducting phase. It should be emphasized that in homogenizing random media, the determination of exact values of $\widehat{\sigma}$, either by formula or in terms of an algorithm for easy, highly precise computation, largely remains a distant goal. Establishing the existence of a well-defined $\widehat{\sigma}$ and determining any of its nontrivial qualitative properties is already a major achievement.

Effective Conductivities

As an alternative to the homogenization approach outlined above, one may simply define effective properties of an inhomogeneous system by a black box approach. In the electrical analogy, and considering three dimensions for definiteness, take a finite rectangular prism $-M < x < M$, $-M < y < M$, $0 \leq z \leq L$, occupied by the inhomogeneous conductor, with all boundaries save for $z = 0$ and $z = L$ being insulated ($\mathbf{n} \cdot \sigma \nabla V = 0$). If one specifies the potential to be $V = V_0$ at $z = 0$ and $V = 0$ at $z = L$, and the z-component of the current through the surface $x = 0$ is $J(x, y)$ per unit area, then a reasonable definition of the effective conductivity σ_{eff} is

$$\sigma_{\text{eff}} = \lim_{L\to\infty} \lim_{M\to\infty} \frac{1}{V_0/L} \int_{-M}^{M} \int_{-M}^{M} \frac{J(x, y)\mathrm{d}x\mathrm{d}y}{(2M)^2} , \quad (7)$$

where L/M is held constant as L, $M \to \infty$. This is the most natural definition for physical experiments or computer simulation and the one we adopt below in our more detailed discussion of discrete models. Actually, for a randomly-structured two-phase system, unless one establishes that for the particular statistical model of the microstructure, the limit is well-defined and takes the same value for almost all realizations of the microstructure (that is, the limit exists and takes a unique value with probability 1), the integral has to be averaged over all realizations of the microstructure before the limit $L, M \to \infty$ is taken.

An alternative definition replaces the insulated boundary conditions on all faces other than $z = 0$ or $z = L$ by $V = V_0(L - z)/L$. One could alternatively set up a definition of effective conductivity by prescribing a uniform injection of current per unit area across the face $z = 0$ and studying the induced potential difference.

Golden and Papanicolaou [56] have examined the relation between the black box definition of effective conductivity for a finite region and the definition from homogenization, and shown their equivalence under the conditions of uniform ellipticity. Stronger results applicable to percolative systems will be found in Jikov et al. [79].

While results on the existence of a well-defined effective conductivity σ_{eff} are of course of significant interest, one really wants to determine its value.

Exact Results Few nontrivial exact results on the effective conductivity are available. Consider a one-dimensional conductor of length L, comprising N independent random conducting elements of length Δ, with the kth element having conductivity σ_k and sustaining a potential drop V_k. Then the current flowing is $J = \sigma_k V_k/\Delta$, the same in each element, and the field is

$$\frac{1}{L}\sum_{k=1}^{N} V_k = \frac{1}{N\Delta}\sum_{k=1}^{N} \frac{J\Delta}{\sigma_k} \to J\langle\sigma^{-1}\rangle , \quad (8)$$

where we have taken the limit $N \to \infty$ and used the Strong Law of Large Numbers [47]. Hence we have the exact result that

$$\sigma_{\text{eff}} = \frac{1}{\langle\sigma^{-1}\rangle} . \quad (9)$$

For two-dimensional systems there are a number of "phase interchange" or "duality" results, due to Keller [82] and others, discussed carefully by Milton [109]. In particular, an infinite chessboard, with the white squares having conductivity σ_{w} and the black squares having conductivity σ_{b} produces an exact effective conductivity

$$\sigma_{\text{eff}} = \sqrt{\sigma_{\text{w}}\sigma_{\text{b}}} . \quad (10)$$

Bounds The electrical energy dissipation rate associated with a potential distribution V and corresponding electric field $\mathbf{E}$ and current $\mathbf{J}$ in the finite $2M \times 2M \times L$ rectangular prism of Sect. "Effective Conductivities" is

$$\begin{aligned} \mathcal{E}\{V\} &= \int_{-M}^{M}\int_{-M}^{M}\int_{0}^{L} \mathbf{J}\cdot\mathbf{E}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z \\ &= \int_{-M}^{M}\int_{-M}^{M}\int_{0}^{L} \sigma|\nabla V|^2\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z . \end{aligned} \quad (11)$$

Consider a potential V that satisfies the partial differential equation $\nabla \cdot (\sigma \nabla V) = 0$ in the interior and the boundary conditions $v = V_0$ on $z = 0$, $V = 0$ on $z = L$ and $\mathbf{n} \cdot \sigma \nabla V = 0$ on other surfaces: we call this an *admissible potential*. Let W, a *trial potential*, be any other function that satisfies the same boundary conditions but may not satisfy the partial differential equation. If we write $\eta = W - V$, expand $\mathcal{E}\{W\} = \mathcal{E}\{V + \eta\}$ and use the Divergence Theorem and the boundary conditions, we easily show that $\mathcal{E}\{W\} \geq \mathcal{E}\{V\}$, so that admissible potentials minimize energy dissipation. That is, we have a variational principle. It can be shown under reasonable conditions that

$$\sigma_{\text{eff}} \leq \frac{1}{(V_0/L)^2 (2M)^2 L} \int_{-M}^{M} \int_{-M}^{M} \int_{0}^{L} \sigma |\nabla W|^2 \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z \,, \tag{12}$$

so that trial potentials can be used to construct rigorous upper bounds on the effective conductivity. A complementary variational principle that leads to lower bounds on the effective conductivity can be derived by analyzing trial current distributions. The variational approach has its origin in work in 1955 of Brown [26] and has been developed in diverse contexts by many authors (Brown, Beran, Prager, Hashin and Shtrikman, Davies, ...) with varying degrees of rigor [109,166]. Bergman [15] introduced an alternative approach to variational methods for deriving bounds on effective conductivities, based on complex variable techniques. Bergman's work and many important consequences are reviewed by Milton [109].

The most easily derived bounds hold without restriction on the distribution of phases in a multiphase system, and also apply for continuously varying local conductivities:

$$\frac{1}{\langle \sigma \rangle} \leq \sigma_{\text{eff}} \leq \langle \sigma \rangle \,, \tag{13}$$

where angle brackets denote the volume average. These bounds, first identified in 1912 by Wiener [168], are the best possible general bounds, since they become equalities in the case of slabs of homogeneous material parallel to the face $z = 0$ (conductors in series—$\sigma_{\text{eff}} = 1/\langle \sigma^{-1} \rangle$) or no variation in conductivity with z (conductors in parallel—$\sigma_{\text{eff}} = \langle \sigma \rangle$). If no additional assumptions are made except for statistical isotropy, the 1962 bounds of Hashin and Shtrikman [67] become the best possible. For high and low conductivity phases with volume fractions ϕ_{high} and ϕ_{low} and conductivities $\sigma_{\text{high}} \geq \sigma_{\text{low}}$, the Hashin–Shtrikman bounds in d dimensions ($d = 2$ or 3) are

$$\sigma_{\text{low}} + \frac{d\phi_{\text{high}}\sigma_{\text{low}}(\sigma_{\text{high}} - \sigma_{\text{low}})}{d\sigma_{\text{low}} + \phi_{\text{low}}(\sigma_{\text{high}} - \sigma_{\text{low}})} \leq \sigma_{\text{eff}} \leq \sigma_{\text{high}} - \frac{d\phi_{\text{low}}\sigma_{\text{high}}(\sigma_{\text{high}} - \sigma_{\text{low}})}{d\sigma_{\text{high}} - \phi_{\text{high}}(\sigma_{\text{high}} - \sigma_{\text{low}})} \,. \tag{14}$$

If we consider the case of a percolative system, where we have $\sigma_{\text{low}} = 0$, $\sigma_{\text{high}} = \sigma_1$ and for brevity we write $\phi_{\text{high}} = \phi$ (that is, ϕ is the volume fraction for the conducting phase) the lower bounds in (13) and (14) become trivial ($\sigma_{\text{eff}} \geq 0$) and neither upper bound is very strong, as shown by Fig. 1. For the case of spheres with independently located centers (technically, Poisson points) and either constant radii, or random radii with a reasonable distribution, it can be rigorously established [104] that for sufficiently small ϕ, the probability that there is an infinite connected region of the conducting phase is zero and consequently the effective conductivity will be zero below a percolation threshold ϕ_c. Numerical simulations with Poisson-centered conducting spheres of constant radius show that $\phi_c \approx 0.2895 \pm 0.0005$ [135]: the gray rectan-

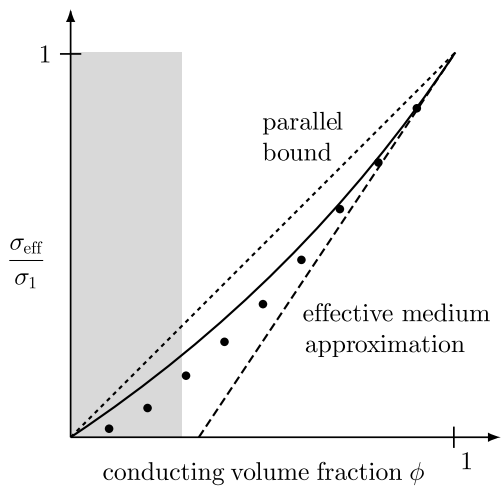

Conduction and Diffusion in Percolating Systems, Figure 1
Bounds on the effective conductivity σ_{eff} for a conducting phase of conductivity σ_1 randomly and homogeneously distributed at volume fraction ϕ, with remaining space occupied by a nonconducting phase. The *gray* area lies below the percolation threshold estimate $\phi_c \approx 0.2895 \pm 0.0005$ of Rintoul and Torquato [135] for overlapping conducting spheres in a nonconducting matrix. The solid discs are numerical estimates of Kim and Torquato [87] for overlapping non-conducting spheres in a conducting matrix, where $\phi_c \approx 0.03$. *Upper broken line*: elementary ("parallel") bound $\sigma_{\text{eff}}/\sigma_1 \leq \phi$. Solid curve: Hashin–Shtrikman bound $\sigma_{\text{eff}}/\sigma_1 \leq 2\phi/(3 - \phi)$ [67]. Lower broken line: effective medium approximation $\sigma_{\text{eff}}/\sigma_1 \approx (3/2)(\phi - 1/3)$ for $\phi \geq 1/3$ from work of Bruggeman [27]

gle in Fig. 1 spans the interval $0 \le \phi \le \phi_c$ for this system, and parts of the upper bound curves (13) and (14) that intersect the Gray rectangle miss the threshold behavior entirely and are misleadingly uninformative for this specific system.

In Fig. 1 we also show as discs numerical estimates [87] of the conductivity in the case where the randomly placed spheres are non-conducting, the rest of space is conducting and the percolation threshold is much smaller ($\phi_c \approx 0.03$ [87]); physicists call this the *Swiss cheese model.* The Hashin–Shtrikman upper bound (14) works better because the percolation threshold is so low, but the Hashin–Shtrikman lower bound remains useless. Much better bounds than those shown are available for non-percolative problems, especially if additional information about correlations between phase locations is available [109,166].

Approximations Clarifying the work of earlier authors, in 1935 Bruggeman [27] derived an *effective medium approximation* for the effective conductivity. This approach represents an uncontrolled approximation—it does not give rigorous bounds and a priori estimation of its accuracy is not possible. One considers a spherical inclusion of unspecified constant conductivity within a uniform medium whose conductivity is taken to be the unknown effective conductivity σ_{eff}, which will be estimated "self-consistently". For a prescribed constant field $\mathbf{E}$ at infinity, the field within the inclusion is calculated. The requirement that, when averaging over material properties within the inclusion, the average field in the inclusion does not differ from the field at infinity yields an approximate equation for σ_{eff},

$$\left\langle \frac{3\sigma_{\text{eff}}}{\sigma + 2\sigma_{\text{eff}}} \right\rangle = 1 \,. \tag{15}$$

In the case in which $\sigma = \sigma_1$ with probability ϕ, corresponding to a volume fraction ϕ of the conducting phase, this approximation predicts that for a percolative system

$$\sigma_{\text{eff}} = \frac{3}{2}\left(\phi - \frac{1}{3}\right)\sigma_1 \,. \tag{16}$$

The physical requirement that $\sigma_{\text{eff}} \ge 0$ leads one to interpret this as predicting that $\sigma_{\text{eff}} = 0$ for $0 \le \phi \le 1/3$, so that there is a threshold for the conductivity at $\phi = 1/3$.

As shown in Fig. 1, this approximation closely conforms to the Hashin–Shtrikman upper bound near $\phi = 1$. Although it has the desirable feature of predicting a conductivity threshold, the actual predicted threshold (1/3) is not particularly close to numerical estimate for randomly placed spheres with independent centers (0.2895) and is very far from the estimated threshold for the Swiss cheese model (0.03).

Differences Between Continuum and Discrete Models

The considerable emphasis in the literature on discrete (lattice) models is partly motivated by the folklore of *universality*, under which, apart from some reasonable conditions that exclude direct, long-range connections and other oddities, qualitative features and associated critical exponents of analogous problems depend only on dimensionality. While this appears to be broadly true, there are some significant exceptions, two of which we mention here.

(a) A parameter value where properties of the infinite system which are elsewhere analytic lose analyticity is called a *critical point*. In the percolation model of Broadbent and Hammersley [25,62,63,64] on periodic lattices, there is only one critical point associated with global connectedness [106], namely the percolation threshold, and this coincides with the only critical point for the effective conductivity. Following on work of Kozlov [93], Jikov et al. [79] have discussed carefully a random chessboard, where the plane is divided into equal squares, colored independently. Squares are black (conductivity σ_b) with probability ϕ and white (conductivity σ_w) with probability $1-\phi$. For $\phi = 1/2$, Eq. (10) holds, but more importantly, in the random chessboard model there are effectively two thresholds for a given color, one for long-distance connectivity using corner connections and one for long-distance connectivity through adjoining edges of the squares. If we let $p_c \approx 0.59$ denote the site percolation threshold of the square lattice (▶ Percolation Thresholds, Exact), then the following results hold as $\sigma_b \to 0$ with σ_w held fixed. For $0 \le \phi \le 1 - p_c$, the white phase has large components connected via edges of squares and

$$\lim_{\sigma_b \to 0} \sigma_{\text{eff}} = \sigma_w f(\phi) > 0 \,.$$

For $1 - p_c < \phi < p_c$ neither color has large components connected via edges of squares, but both colors have large components connected via corners, and

$$c_1(\phi)\sqrt{\sigma_w \sigma_b} \le \sigma_{\text{eff}} \le c_2(\phi)\sqrt{\sigma_w \sigma_b} \,.$$

For $p_c < \phi < 1$, we have $\sigma_b \le \sigma_{\text{eff}} \le c_3(\phi)\sigma_b$. Berlyand and Golden [16] have established the stronger result that for $1 - p_c < \phi < p_c$, $\sigma_{\text{eff}} = \sqrt{\sigma_w \sigma_b} + O(\sigma_b)$ as $\sigma_b \to 0$ with σ_w held fixed. Torquato et al. [167] have reported numerical studies.

(b) The universality concept presupposes that all analogous systems in the same dimension have the same crit-

ical exponents for a given system property. Thus, for example, in a percolative continuous system where ϕ is the volume fraction of the conducting phase, we expect to see

$$\sigma_{\text{eff}} \sim \text{constant} \times (\phi - \phi_c)^t \quad \text{as } \phi \downarrow \phi_c \,. \tag{17}$$

where the conductivity exponent t (the general existence of which remains to be established, even for specific two and three dimensional systems) is independent of microstructure, and the same for lattices and continua in a given dimension. This expectation turns out to be wrong, with hopes being dashed by computational studies of Feng et al. [48] on the Swiss cheese model introduced in Sect. "Bounds": in two dimensions the lattice and Swiss cheese continuum model exponents are close together and possibly coincident; in three dimensions the exponent t for the Swiss cheese continuum is about 0.5 larger than the exponent t for the simple cubic lattice.

Lattice Models: Steady-State Phenomena

Notwithstanding the caveats exposed in Sect. "Differences Between Continuum and Discrete Models", we turn to a detailed discussion of discrete systems. We consider *lattices* (also known as networks or graphs) of *sites* (also known as nodes or vertices) connected by *bonds* (also called links or edges). The number of bonds attached to a site is its *coordination number* (also called degree or valence), and two sites that are attached to the same bond are called *nearest-neighbour sites.* The most commonly studied lattices are periodic in structure, the simplest example being the *hypercubic lattice* $\mathbb{Z}^d$ of sites with integer coordinates, connected by bonds of unit length, so that each site has coordination number $2d$. Other important infinite networks are sometimes called *pseudolattices* to distinguish them from periodic structures. The most important examples of pseudolattices are self-similar or fractal structures [100] (▶ Scaling Properties, Fractals, and the Renormalization Group Approach to Percolation) and tree-like structures, especially the *Cayley tree* or *Bethe lattice* [165] of coordination number z, in which there are no closed loops.

Key Results from Percolation Theory

In percolation theory for lattices sites, are declared *occupied* or *vacant* in some random way and bonds are declared *open* or *closed* in some random way. A site is deemed to belong to a cluster of size $n \geq 1$ (where n may be finite or infinite) if it is present and connected via open bonds to $n-1$ other sites; a vacant site is a cluster of size 0. In the (Bernoulli) *bond problem*, all sites are occupied, and bonds are independently declared open with probability p and closed with probability $1-p$. In the (Bernoulli) *site problem*, sites are occupied with probability p and vacant with probability $1-p$, and a bond is declared open if and only if the two sites it joins are both occupied. Every bond percolation problem is equivalent to a site percolation problem on a related lattice, so that it suffices to state general results for site percolation, although in many examples modeling transport and conduction, the language of bond percolation is more appropriate.

Viewed at sufficiently low resolution, a realization of bond or site percolation other than at the percolation threshold appears homogeneous. A measure of the length scale over which inhomogeneous structure is important is furnished by the *correlation length* $\xi(p)$. One way of defining this precisely on $\mathbb{Z}^d$ is to let τ_m denote the probability that the sites at $\mathbf{0}$ and $(m, 0, 0, \ldots, 0)$ belong to a common finite cluster. Then $\xi(p) = -\lim_{m\to\infty} m^{-1} \ln \tau_m$.

Rigorous General Results The following results are rigorously proven [21,59,75] for Bernoulli site percolation, provided that the lattice or pseudolattice is homogeneous in the sense that all sites are equivalent (as is the case for $\mathbb{Z}^d$ and the Bethe lattice), and the number of sites at most n bonds distant from a given site grows no faster than $c\exp(an^\kappa)$, where $a > 0$, $c > 0$ and $\kappa < 1$.

(i) There exists a percolation threshold $p_c > 0$ such that the probability $P_\infty(p)$ of a chosen site belonging to an infinite cluster is zero for $p < p_c$ and strictly greater than zero for $p > p_c$.

(ii) The probability $P_n(p)$ that a chosen site belongs to a cluster of size n decays exponentially rapidly with increasing n when $p < p_c$.

The exponential decay of $P_n(p)$ ensures the finiteness of the mean cluster size $\chi(p) = \sum_{n=0}^{\infty} nP_n(p)$ for $p < p_c$. The sum remains meaningful for $p > p_c$ if we exclude the infinite cluster $n = \infty$ and $\chi(p)/(1 - P_\infty(p))$ becomes the mean cluster size, conditioned on the cluster being finite. A result related to (ii) establishes that on $\mathbb{Z}^d$, the correlation length is well-defined and finite for $p < p_c$.

(iii) The mean cluster size $\chi(p)$ is finite for $p < p_c$ but $\chi(p) \geq pp_c(p_c - p)^{-1}$ when $p < p_c$, so that $\chi(p) \to \infty$ as $p \uparrow p_c$.

(iv) For a given value of p, there is a number k_0 (which may take no values other than 0, 1 or ∞) such that the number of distinct infinite clusters of occupied sites on the lattice takes the value k_0 with probability 1. When $P_\infty(p) = 0$, $k_0 = 0$.

For some pseudolattices, such as the Bethe lattice, if $P_\infty(p) > 0$ then $k_0 = \infty$; however, for a class of lattices which includes normal periodic lattices, if $0 < P_\infty(p) < 1$

then $k_0 = 1$. That is, when an infinite cluster exists it is unique.

(v) For $p > p_c$ a site is part of the backbone if it lies on the infinite cluster, and within that cluster there are two independent paths to infinity. If $B(p)$ is the probability that a given site belongs to the backbone then $B(p) \le p^{-1} P_\infty(p)^2$.

For a discussion of the determination of and actual values of site and percolation thresholds for particular lattices, see ► Percolation Thresholds, Exact.

Heuristics A heuristic scaling theory for percolation theory (► Scaling Properties, Fractals, and the Renormalization Group Approach to Percolation introduced in 1979 by Stauffer [153] and analogies between percolation and statistical mechanics first exhibited in 1972 by Fortuin and Kasteleyn [51,52] lead to the following folklore [75]. There are *critical exponents* that describe the non-analytic behavior of important attributes of the percolation model at the percolation threshold. If we use the proportionality symbol $\propto$ as a short notation, so that $f(p) \propto g(p)$ as $p \to p_c$ means that $f(p)/g(p) \to$ nonzero constant, then the critical exponents are defined as follows:

cluster size distribution at p_c

$$P_n(p_c) \propto n^{-1-1/\delta} \quad \text{as } n \to \infty\,; \tag{18}$$

mean cluster size below p_c

$$\chi(p) \propto (p_c - p)^{-\gamma} \quad \text{as } p \uparrow p_c\,; \tag{19}$$

mean cluster size above p_c

$$\chi(p) \propto (p_c - p)^{-\gamma'} \quad \text{as } p \downarrow p_c\,; \tag{20}$$

correlation length below p_c

$$\xi(p) \propto (p_c - p)^{-\nu} \quad \text{as } p \uparrow p_c\,; \tag{21}$$

correlation length above p_c

$$\xi(p) \propto (p_c - p)^{-\nu'} \quad \text{as } p \downarrow p_c\,; \tag{22}$$

percolation probability

$$P_\infty(p) \propto (p - p_c)^{\beta} \quad \text{as } p \downarrow p_c\,; \tag{23}$$

backbone probability

$$B(p) \propto (p - p_c)^{\beta_{bb}} \quad \text{as } p \downarrow p_c\,. \tag{24}$$

For the Bethe lattice, it is established that $\delta = 2, \gamma = 1$, $\beta = 1$ and $\beta_{bb} = 2$ [75], but for standard lattices even the existence of critical exponents in the sense explained above is not rigorously established, although as noted below, a few weaker results consistent with this picture, such as $\lim_{p\downarrow p_c} \ln[P_\infty(p)]/\ln(p - p_c) = \frac{5}{36}$ for site percolation on the triangular lattice, are now proven.

Folklore asserts that exponents are the same for the all site and bond problems in a given dimension (universality), vary significantly with the dimension d in low dimensions, and are independent of dimension for $d \ge 6$, where they take the so-called *mean-field values* $\beta = \gamma = 1$ and $\nu = \frac{1}{2}$. (These values are rigorously established by work of Hara and Slade [66] for standard percolation for $d \ge 19$, and for a spread-out model of percolation for $d \ge 6$.) At $d = 6$, the prefactors in the dominant asymptotic forms that define the critical exponents may be modified by the inclusion of a power of $\ln|p - p_c|$. It is generally believed that $\gamma = \gamma'$ and $\nu = \nu'$. On modest hypotheses on lattice structure it is rigorously established [75] from work of various authors that, provided the exponents exist, $\beta \le 1$, $\gamma \ge 1$, $\gamma/d \le \nu \le \gamma$, and $\nu \ge 2/d$, but two much more informative scaling relations are accepted:

$$\gamma + 2\beta = \beta(\delta + 1) \tag{25}$$

in all cases, and for $d \le 6$,

$$d\nu = 2\beta + \gamma\,. \tag{26}$$

All credible numerical evidence supports the existence of critical exponents consistent with these scaling laws, they are rigorously established for site percolation on the triangular lattice [148], and natural probabilistic conditions equivalent to (26) for bond percolation on $\mathbb{Z}^d$ are known [22].

Rigorous Results for Two Dimensions For several two-dimensional lattices a number of exact results are known: in particular $p_c = 1/2$ for site percolation on the triangular lattice (coordination number 6) and bond percolation on the square lattice (coordination number 4). In these particular problems it has also been rigorously established that the percolation probability $P_\infty(p)$ is continuous for $0 \le p \le 1$, so that there is no infinite cluster present right at the percolation threshold. This result is believed to be true, but remains unproven, in three dimensions.

Major innovations in probability since 2000, primarily associated with the work of Lawler, Schramm, Smirnov and Werner, have led to the following results being rigorously established for the case of site percolation on the two-dimensional triangular lattice [148]:

$$P_\infty(p) = (p - \tfrac{1}{2})^{5/36 + o(1)} \quad \text{as } p \downarrow \tfrac{1}{2}\,; \tag{27}$$

$$\chi(p) = |p - \tfrac{1}{2}|^{-43/18 + o(1)} \quad \text{as } p \to \tfrac{1}{2}\,; \tag{28}$$

$$\xi(p) = |p - \tfrac{1}{2}|^{-4/3 + o(1)} \quad \text{as } p \to \tfrac{1}{2}\,. \tag{29}$$

This is almost (but not quite) a proof of existence of critical exponents in the sense of physicists, and moreover shows

that the scaling relations (25) and (26) hold in this one particular case and the accepted folklore for two dimensions that

$$\beta = \tfrac{5}{36}, \qquad \gamma = \gamma' = \tfrac{43}{18}\,, \qquad \nu = \nu' = \tfrac{4}{3}\,, \tag{30}$$

arrived at from scaling arguments, statistical mechanical arguments, heuristic arguments and numerical evidence, is indeed correct in this one particular case. The exact backbone exponent β_{bb} remains unknown, but it has been shown to be determined by the leading eigenvalue of a differential operator [96].

The Random Resistor Problem

The canonical discrete model of transport and conduction is the *random resistor network*, in which the bonds of a lattice or some less regular network are resistors of random resistance. A more extended discussion of this model, including proofs of many results stated here, will be found in Hughes [75]; for the Dirichlet, Thomson and Rayleigh Principles discussed below see also Doyle and Snell [40]. The first study of random resistor networks may fairly be ascribed to 1971 work of Kirkpatrick [88,89], although an equivalent system had already been studied as early as 1956 by Fatt [45] as a model for porous media, with hydraulic conductivities of bonds playing the role of resistance and pressure playing the role of voltage.

Problem Definition If we use Greek subscripts to index bonds of the lattice (each bond being assigned a nominal direction of orientation), then the (direct) current I and the potential difference V across a particular bond α are related by Ohm's Law $I = G_\alpha V$, where G_α is the conductance of bond α (the reciprocal of the resistance). Current is a signed quantity. If the potential at site $\mathbf{s}'$ is less than that at site $\mathbf{s}$ then the current flowing from $\mathbf{s}$ to $\mathbf{s}'$ is positive.

It is customary to assume that the conductances of different bonds are independent, identically distributed random variables. Note that one might consider more generally the case in which $I = \mathcal{G}_\alpha(V)$ where $\mathcal{G}_\alpha$ is a nonlinear function with several random parameters, but this problem is both subtle [29] and difficult [19,20,61,83,105,161] and will not be discussed here; see Sahimi [141].

If the bonds of a random resistor network have independent, identically distributed conductances governed by the probability density function

$$f(g) = (1-p)\delta_+(g) + ph(g), \quad 0 < p < 1\,, \tag{31}$$

where $h(g)$ is the conditional probability of the conductance, given that the conductance is nonzero, we call the network the *general percolation conduction problem*. We use the notation $\delta_+(g)$ to distinguish a delta function that acts on the right from the usual symmetric delta function $\delta(x)$. The specific case in which $h(g) = \delta(g - g_0)$, that is,

$$f(g) = (1-p)\delta_+(g) + p\delta(g - g_0)), \quad 0 < p < 1\,, \tag{32}$$

will be called the *standard* percolation conduction problem. The definitions (31) and (32) are directly associated with bond percolation. A standard percolation problem for site percolation arises on assigning each bond that links two occupied sites the conductance g_0, and making all other bonds non-conducting.

In general, we have a lattice with a potential (voltage) $V_{\mathbf{s}}$ at each site $\mathbf{s}$ and a current $i_{\mathbf{rs}}$ flowing from site $\mathbf{r}$ to site $\mathbf{s}$ along a bond of conductance $G_{\mathbf{rs}} = G_{\mathbf{sr}}$ (or, equivalently, resistance $R_{\mathbf{rs}} = 1/G_{\mathbf{rs}}$). From Ohm's Law

$$i_{\mathbf{rs}} = G_{\mathbf{rs}}(V_{\mathbf{r}} - V_{\mathbf{s}}) \tag{33}$$

and Kirchhoff's Law of conservation of current $\sum_{\mathbf{s}} i_{\mathbf{rs}} = I_{\mathbf{r}}$, we have

$$\sum_{\mathbf{s}} G_{\mathbf{rs}}(V_{\mathbf{r}} - V_{\mathbf{s}}) = I_{\mathbf{r}}\,, \tag{34}$$

where $I_{\mathbf{r}}$ is the current supply into site $\mathbf{r}$, positive if current is injected and negative if it is withdrawn. At most sites $\mathbf{r}$, we have $I_{\mathbf{r}} = 0$, but we impose the additional physical requirement that $\sum_{\mathbf{r}} I_{\mathbf{r}} = 0$, the sum being taken over all sites where current is injected or withdrawn. When appropriate boundary conditions are imposed on finite pieces of periodic lattice, or on finite networks more generally, the problem is completely specified and numerical determination of the potential distribution over the sites and the associated currents is possible; from this estimates of an effective conductivity σ_{eff} that we define in Sect. "Lattices of Dimension $d \geq 2$" below emerge.

Bounds for General Resistor Networks Before addressing the specific case of periodic lattices, we consider arbitrary finite networks. The *effective conductance* $C^{\mathrm{eff}}_{\mathbf{ab}}$ *between two sites* $\mathbf{a}$ and $\mathbf{b}$ and its reciprocal, the *effective resistance* $R^{\mathrm{eff}}_{\mathbf{ab}}$ *between the sites*, are defined by

$$C^{\mathrm{eff}}_{\mathbf{ab}} = \frac{1}{R^{\mathrm{eff}}_{\mathbf{ab}}} = \frac{I}{V_{\mathbf{a}} - V_{\mathbf{b}}}\,, \tag{35}$$

if a current I injected at site $\mathbf{a}$ and withdrawn at site $\mathbf{b}$ is associated with a potential difference $V_{\mathbf{a}} - V_{\mathbf{b}}$ between the sites. (Effective conductivities arise by scaling effective conductances with appropriate measures of network size.)

Variational principles enable bounds on the effective conductance to be deduced.

DIRICHLET'S PRINCIPLE. Consider a finite resistor network with bond conductances $G_{\mathbf{rs}}$ $(0 \le G_{\mathbf{rs}} < \infty)$ and with a specified potential difference V between sites $\mathbf{a}$ and $\mathbf{b}$. Let $\{W_{\mathbf{s}}\}$ be a *trial potential distribution* , that is, any set of real numbers subject to the requirement that $W_{\mathbf{a}} - W_{\mathbf{b}} = V$. Let $\{V_{\mathbf{s}}\}$ be an admissible potential distribution, that is, a set of real numbers satisfying the Kirchhoff–Ohm Law (34) for $\mathbf{s} \neq \mathbf{a}$ or $\mathbf{b}$, as well as the constraint $V_{\mathbf{a}} - V_{\mathbf{b}} = V$. Then the effective conductance between sites $\mathbf{a}$ and $\mathbf{b}$ is given by

$$C^{\text{eff}}_{\mathbf{ab}} = \frac{1}{2V^2} \sum_{\mathbf{r},\mathbf{s}} G_{\mathbf{rs}}(V_{\mathbf{r}} - V_{\mathbf{s}})^2 \le \frac{1}{2V^2} \sum_{\mathbf{r},\mathbf{s}} G_{\mathbf{rs}}(W_{\mathbf{r}} - W_{\mathbf{s}})^2 . \tag{36}$$

THOMSON'S PRINCIPLE. Consider a finite resistor network with bond resistances $R_{\mathbf{rs}}$ $(0 \le R_{\mathbf{rs}} < \infty)$ and a specified current $I_{\mathbf{a}}$ injected at site $\mathbf{a}$ and withdrawn at site $\mathbf{b}$. Let $\{j_{\mathbf{sr}}\}$ be a *trial current distribution*, that is: $j_{\mathbf{rs}} = -j_{\mathbf{sr}}$; $\sum_{\mathbf{s}} j_{\mathbf{rs}} = 0$ if $\mathbf{r} \neq \mathbf{a}$ and $\mathbf{r} \neq \mathbf{b}$; and $\sum_{\mathbf{s}} j_{\mathbf{as}} = I_{\mathbf{a}} = -\sum_{\mathbf{s}} j_{\mathbf{bs}}$. Let $\{i_{\mathbf{sr}}\}$ be an *admissible current distribution*, that is, a trial current distribution for which Ohm's Law $V_{\mathbf{r}} - V_{\mathbf{s}} = i_{\mathbf{rs}} R_{\mathbf{rs}}$ holds. Then the effective resistance between sites $\mathbf{a}$ and $\mathbf{b}$ is given by

$$R^{\text{eff}}_{\mathbf{ab}} = \frac{1}{2I_{\mathbf{a}}^2} \sum_{\mathbf{r},\mathbf{s}} i_{\mathbf{rs}}^2 R_{\mathbf{rs}} \le \frac{1}{2I_{\mathbf{a}}^2} \sum_{\mathbf{r},\mathbf{s}} j_{\mathbf{rs}}^2 R_{\mathbf{rs}} . \tag{37}$$

The sums in the inequalities in Dirichlet's and Thomson's Principles may be interpreted as the rate of energy dissipation corresponding to the potential distribution or the current distribution respectively. These two principles assert that admissible potential and current distributions minimize the rate of energy dissipation. From these two principles a number of important results follow, including the following.

RAYLEIGH'S MONOTONICITY PRINCIPLE. For finite resistor networks, the resistance between two sites does not decrease if the resistances of some or all of the bonds are increased.

Exactly Solvable Models In one dimension, N random resistors in series yield an effective conductivity (here the overall conductance per unit length) $\sigma_{\text{eff}} = \langle G^{-1} \rangle^{-1}$; the argument is the same as that used above for a random one-dimensional continuum. A slightly less trivial problem for which exact results are available is the standard percolation conduction problem on a Bethe lattice of coordination number z [70,75,156,157]. Pick an arbitrary site $\mathbf{s}_0$ of the lattice, set that site to have potential V_0 and prescribe that the potential is zero at the periphery of the tree (that is, the potential decays asymptotically to zero as we move outwards from $\mathbf{s}_0$ along any infinite path of open bonds that may be present). If a current I_0 flows from site $\mathbf{s}_0$ into the tree, then the tree conductance may be defined as $T = I_0/V_0$. Where the average is taken over all realizations of the system, it can be shown rigorously that $\langle T \rangle$ is zero for $p < p_c = (z-1)^{-1}$ and

$$\langle T \rangle \sim \frac{2zg_0 c}{z-2} [p(z-1) - 1]^2 \quad \text{as } p \downarrow p_c , \tag{38}$$

where $c \approx 0.761$ is a constant. Since $p_c = 1/(z-1)$ for the Bethe lattice, this result is consistent with the hypothesis that long-distance conduction is possible in the standard percolation conduction problem (32) if and only if $p > p_c$. It is sometimes argued that the Bethe lattice correctly represents the critical behavior of periodic lattices of sufficiently high dimensionality and that would imply that effective conductivity σ_{eff} defined in Sect. "Lattices of Dimension $d \ge 2$" satisfies $\sigma_{\text{eff}} \propto (p - p_c)^2$ as $p \downarrow p_c$. Straley [159] has identified the flaw in this argument—in a Bethe lattice of finite size, a significant fraction of all sites are at the boundary—and has given a definition of the effective conductivity σ_{eff} of a Bethe lattice that gives bonds in all regions of the lattice comparable significance and leads to $\sigma_{\text{eff}} \propto (p - p_c)^3$ as $p \downarrow p_c$.

For some remarks concerning self-similar pseudolattices for which exact results are available, see Sect. "Conduction on Fractals".

Lattices of Dimension $d \ge 2$

For random resistor problems more appropriately related to transport and conduction problems in d-dimensional space, we consider periodic lattices. There have been numerical investigations of the standard percolation conduction on topologically random lattices created by the Voronoï algorithm [77,78], but we shall not pursue these extensions of the theory.

Defining the Conductivity Consider the specific case of a finite hypercube Λ_L cut from from the simple hypercubic lattice $\mathbb{Z}^d$, with $(L+1)^d$ sites at locations $\mathbf{s} = (s_1, s_2, \ldots, s_d)$ with integer coordinates subject to the inequalities $0 \le s_j \le L$. The faces $s_1 = 0$ and $s_1 = L$ are subjected to a potential difference V, so that there is a potential gradient V/L. In three dimensions, this corresponds to a cubic array of conducting elements sandwiched between two perfectly conducting plates. Boundary conditions on the other faces should be unimportant when L is

large, but in numerical work periodic boundary conditions are often used (the faces $s_j = 0$ and $s_j = L$ are joined for $j = 2, \ldots d$). If a total current I flows in response to the imposed potential difference, then the current per unit area (that is, per site) of the face $s_1 = 0$ is $I/(L+1)^{d-1}$. We therefore define the (specific) conductivity of the lattice as

$$\lim_{L\to\infty} \frac{I/(L+1)^{d-1}}{V/L} = \lim_{L\to\infty} \frac{L^{2-d} I}{V} . \tag{39}$$

For the random resistor network, one needs to discuss the existence of this limit for each realization ω of the lattice and one is especially interested in the mean value of the limit, which we take as defining the *effective conductivity* of the system:

$$\sigma_{\text{eff}} = \lim_{L\to\infty} \langle L^{2-d} I/V \rangle = \lim_{L\to\infty} \left\langle L^{2-d} C_L \right\rangle , \tag{40}$$

where $C_L = I/V$ is the conductance between the faces of the hypercube Λ_L.

For a considerable period the existence of a well-defined effective conductivity remained unproven, with the case of percolative problems providing an especially severe challenge, but this hurdle was eventually overcome [79]. In particular, for percolative problems it is now established there is a well-defined $\sigma_{\text{eff}}(p)$. The averages over realizations in Eq. (40) can be dropped, with the limit existing with probability 1, and being 0 for $p < p_c$ and strictly positive for $p > p_c$

Bounding the Conductivity The Dirichlet and Thompson Principles and Rayleigh's Monotonicity Law are stated for the case in which current is injected at one site and withdrawn at another, but by introducing appropriate links of zero resistance, the laws apply equally well to the case in which a potential difference V is maintained across two parallel faces of the hypercube Λ_L used in the definition of the effective conductivity, and a current I flows between the parallel faces.

THE SERIES AND PARALLEL BOUNDS. For a random resistor network on the simple hypercubic lattice $\mathbb{Z}^d$ with independent, identically distributed bond conductances, the conductance C_L between two parallel faces of the hypercube Λ_L satisfies the inequalities

$$L^{-1}(L+1)^{d-1}\langle G^{-1}\rangle^{-1} \le \langle C_L \rangle \le L^{-1}(L+1)^{d-1}\langle G \rangle, \tag{41}$$

where the random variable G corresponds to the conductance of any one bond of the lattice.

Hence $\langle L^{2-d} C_L \rangle$ is bounded above as $L \to \infty$ if the individual bond conductances G have finite mean, and is also bounded away from zero if the individual bond resistances $(1/G)$ have finite mean. Under these conditions, if σ_{eff} exists, then from Eq. (40), we have

$$\langle G^{-1} \rangle^{-1} \le \sigma_{\text{eff}} \le \langle G \rangle . \tag{42}$$

Hammersley [65] has used Dirichlet's Principle to derive some stronger results than the series and parallel bounds (42) for independent conductances. Chayes and Chayes [30] used variational principles to derive a not directly useful lower bound on the conductivity for the standard percolation conduction problem, and the following more encouraging upper bound.

UPPER BOUND OF CHAYES AND CHAYES. For the standard bond percolation conduction problem on the d-dimensional hypercubic lattice, with open bonds having conductance g_0, the effective conductivity $\sigma_{\text{eff}}(p)$ satisfies the inequality

$$\sigma_{\text{eff}}(p) \le g_0 \Pr\{\text{a given bond is part of the backbone}\} . \tag{43}$$

The *conductivity exponent* for the percolation conduction problem is defined by writing

$$\sigma_{\text{eff}}(p) \propto (p - p_c)^t \quad \text{as } p \downarrow p_c . \tag{44}$$

There is no rigorous proof, even in two dimensions, of the existence of the critical exponent, even in the weak sense that $t = \lim_{p\downarrow p_c} \ln[\sigma_{\text{eff}}(p)]/\ln(p - p_c)$. However, from the result (v) in Sect. "Rigorous General Results" and the upper bound of Chayes and Chayes we know that the conductivity is identically zero below the percolation threshold, as one's intuition demands, and if the accepted critical exponents do exist then we have the inequality

$$t \ge \beta_{\text{bb}} \ge 2\beta , \tag{45}$$

where β_{bb} is the backbone exponent and β is the exponent for the percolation probability $P_\infty(p)$. From numerical estimates $\beta_{\text{bb}} \approx 0.52$ in two dimensions, and we know that $2\beta = \frac{5}{18} \approx 0.28$, while $t > 1$ (see Sect. "Numerical Estimates of the Conductivity Critical Exponent t"), so neither equality is very sharp.

Scaling Theory for the Conductivity Straley [158] developed a heuristic scaling argument to describe the conductivity of a random resistor network for which the probability density function for the individual bond conductances is

$$f(g) = (1-p)\delta(g-a) + p\delta(g-b) , \tag{46}$$

where $a \ll b$. The effective conductivity σ_{eff} changes in the obvious manner when the units of conductivity or

length change. A prefactor μ in the conductivity can be regarded as accounting for this. Straley proposed that σ_{eff} is a homogeneous function near the percolation threshold:

$$\sigma_{\text{eff}}(p) \approx \mu S([p-p_c]\lambda^{-1}, a\mu^{-1}\lambda^{-A}, \mu b^{-1}\lambda^{-B}) , \quad (47)$$

with the approximation assumed valid when each argument of S is less than unity. The function S is singular when any of its arguments vanishes. The parameter λ embodies the interrelations via scaling of the parameters a, b and $p-p_c$. Setting $a=0$, $\lambda=|p-p_c|$ and $\mu=\lambda^B b$ gives

$$\sigma_{\text{eff}}(p) \approx |p-p_c|^B bS(\text{sgn}[p-p_c], 0, 1) . \quad (48)$$

This case describes the conductivity of the classical percolation conduction problem, so we are led to the conclusion that $S(-1,0,1)=0$ and $S(1,0,1)>0$, and we identify B with the conductivity exponent t. If instead we set $1/b=0$, $\lambda=|p-p_c|$ and $\mu=\lambda^{-A}a$, we obtain

$$\sigma_{\text{eff}}(p) \approx |p-p_c|^{-A} aS(\text{sgn}[p-p_c], 1, 0) . \quad (49)$$

This case describes the superconductivity problem in which a fraction $1-p$ of the bonds are normal conductors, while the remaining bonds have infinite conductance, so we conclude that $S(-1,1,0)<\infty$ and $S(1,1,0)=\infty$ and we identify A with the Superconductivity exponent s. Thus

$$\sigma_{\text{eff}}(p) \approx \mu S([p-p_c]\lambda^{-1}, a\mu^{-1}\lambda^{-s}, \mu b^{-1}\lambda^{-t}) . \quad (50)$$

If we now take $p=p_c$ and $0<a<b<\infty$, then setting $\lambda^{t+s}=a/b$ and $\mu=b\lambda^t$, we obtain

$$\sigma_{\text{eff}}(p_c) \approx a^u b^{1-u} S(0,1,1) , \quad (51)$$

where the exponent u is given by

$$u = t/(s+t) . \quad (52)$$

Special Results for two Dimensions Each two-dimensional lattice L with well-defined faces (polygons with the bonds as sides, not crossed by any bonds) is associated with a *dual lattice* L^{D}, obtained by placing a site of L^{D} in every face of L and joining sites of L^{D} by bonds if the original faces of L share a common bond in L. The square lattice is its own dual lattice; the triangular lattice and the hexagonal lattice form a dual pair. Duality is central to the exact determination of percolation thresholds in two dimensions ► Percolation Thresholds, Exact, but also has useful implications for the random resistor problem [17,101,160], which parallel the phase interchange relations for two-dimensional continua (Sect. "Effective Conductivities"). For the square lattice, if the probability density function for bond conductances is

$$f(g) = (1-p)\delta(g-a) + p\delta(g-b) , \quad (53)$$

then the effective conductivity $\sigma_{\text{eff}}(p)$ satisfies the equation

$$\sigma_{\text{eff}}(p)\sigma_{\text{eff}}(1-p) = ab , \quad (54)$$

and in particular $\sigma_{\text{eff}}(\frac{1}{2}) = \sqrt{ab}$. In the notation of Straley's scaling theory (Sect. "Scaling Theory for the Conductivity"), this implies that $u=\frac{1}{2}$ for the square lattice, and so $s=t$ for the square lattice. Although this does not prove rigorously that the conductivity exponent t and superconductivity exponent s coincide for the square lattice, it is taken as evidence for the claim that $s=t$ for all reasonable two-dimensional lattices. Numerical evidence is strongly against this result remaining true in higher dimensions.

It can also be shown from duality that for the square lattice, if

$$f(g) = \frac{1}{gD\sqrt{2\pi}} \exp\left[-\frac{(\log g - \log g_0)^2}{2D^2}\right] , \quad (55)$$

then $\sigma_{\text{eff}} = g_0$.

Effective Medium Approximations The effective medium ideas of Sect. "Approximations" were adapted to the estimation of the conductivity of random resistor networks on periodic lattices by Kirkpatrick [88,89]. A much fuller modern account than given here of the implementation of effective medium ideas for lattice systems and the application of the ideas to more subtle problems and in different contexts, will be found in Sahimi [140].

The random network is replaced by a network that is uniform with unknown bond conductance g_*, except for one special bond, which has conductance G. Let the lattice have coordination number z. It can be shown that the average fluctuation in current or potential difference across the special bond due to its differing in conductance from g_* vanishes provided that

$$\left\langle \frac{G-g_*}{g_* + (2/z)(G-g_*)} \right\rangle = 0 , \quad (56)$$

and this gives a self-consistent approximate determination of g_*. For the square and simple cubic lattices, g_* is an approximation for σ_{eff}. For other lattices, σ_{eff} is a constant multiple of g_*. In the case of a percolative conductance distribution, the approximation produces an estimate of the percolation threshold.

For the standard percolation conduction problem (32), where a fraction p of the bonds has nonzero conductance g_0, the effective medium approximation produces $g_*/g_0 \approx (p-2/z)/(1-2/z)$ and as bond conductances must be non-negative this prediction is interpreted as

meaning that

$$\frac{g_*}{g_0} \approx \begin{cases} 0, & p \le 2/z, \\ \dfrac{p - 2/z}{1 - 2/z}, & p > 2/z\,. \end{cases} \tag{57}$$

That is, the effective medium approximation predicts that the percolation threshold is $2/z$. This is fortuitously exact for the square lattice bond problem and reasonable for the triangular bond and honeycomb bond problems (► Percolation Thresholds, Exact), but works less well in three dimensions. The construction of effective medium approximations for site percolation problems is more delicate [75].

The major qualitative deficiency of the effective medium approximation is the conductivity exponent prediction: it asserts that $t = 1$ for all dimensions. This is already a poor approximation in two dimensions ($t \approx 1.3$—see Table 2) and is grossly misleading in three dimensions ($t \approx 2$—see Table 3). The effective medium approximation also predicts that the critical exponent is unchanged if the standard percolation conduction problem is replaced by the more general model (31), provided that the mean resistance of conducting bonds is finite, that is,

$$\int_0^\infty g^{-1} h(g) \mathrm{d}g < \infty\,. \tag{58}$$

When the conditional probability density function for the conductance of bonds of nonzero conductance has the asymptotic form $h(g) \propto g^{-\alpha}$ as $g \downarrow 0$ with $0 < \alpha < 1$, so that the condition (58) is violated, the effective medium approximation predicts that $t = 1/(1-\alpha)$ [75,92].

The effective medium approximation works well in non-percolative systems, and in percolative systems has some use when p is close to 1.

Renormalization By analogy with statistical mechanics applications, any process by which a lattice system is replaced by a similar lattice with bonds of different lengths is called (*real-space* or *position-space*) *renormalization*. Such a process can be performed exactly on some self-similar pseudolattices, but not on standard periodic lattices.

Numerical estimates of the effective conductivity based on simulation of finite lattice fragments often use an algorithm in which a finite subnetwork is replaced by a single effective bond for which the conductance may be calculated exactly [32,57,97]. This is especially effective near the percolation threshold, and is sometimes loosely described as an exact renormalization procedure [32]. The name is a little misleading, since the new random resistor network cannot generally be embedded in a natural way within the original underlying lattice structure.

There was considerable interest in real-space renormalization in the 1970s and 1980s, as it gave approximate predictions for critical exponents that differ from the so-called mean-field (high dimension) values of the geometrical exponents of percolation theory, and from the effective medium predictions of the conductivity exponent. Like the effective medium approximation, real-space renormalization is an uncontrolled approximation for which a priori estimates of the quality of the approximation are not available. For detailed accounts of real-space renormalization ideas applied to percolation and conduction, see Sahimi [140] and ► Scaling Properties, Fractals, and the Renormalization Group Approach to Percolation.

The effect of renormalization is to reduce the correlation length in the system, moving the system away from the percolation threshold. Sahimi et al. [144] have given an improved, though still uncontrolled approximation, in which approximate renormalization is used to map the system to parameter values where the effective medium approximation is more accurate. This renormalized effective medium approximation effectively increases the interval in which the effective medium approximation works, but like all effective medium treatments cannot accurately

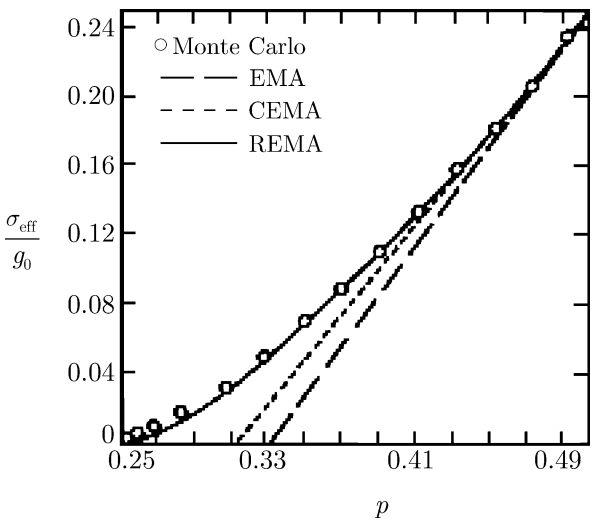

Conduction and Diffusion in Percolating Systems, Figure 2
Effective conductivity of the simple cubic lattice bond problem for the standard percolation conduction model (32). Monte Carlo simulation data (*circles* [89] give clear evidence that the conductivity exponent *t* exceeds 1. The curves represent uncontrolled approximations: EMA, simple effective medium approximation [89]; CEMA, cluster effective medium approximation [1]; REMA, hybrid theory (renormalized effective medium approximation) [144]. Figure reproduced from [144]

portray the ultimate asymptotic behavior as $p \downarrow p_c$. An example of the resulting approximation is given in Fig. 2, together with the simple effective medium approximation [89], an alternative improved effective medium approximation [1] and simulation data [89]. The reader should note from the simulation data, which the approximate curve of Sahimi et al. matches well except very close to the threshold, the clear nonlinear behavior of the conductivity, in striking contrast to the global linear behavior predicted by the simple effective medium theory and the ultimate linear behavior also predicted by the improved approximations near their predicted approximate percolation thresholds.

Numerical Estimates of the Conductivity Exponent *t*

Although it is possible to estimate the conductivity exponent t from an approximate determination of $\sigma_{\text{eff}}(p)$ on an interval including p_c, greater computational efficiency comes from the extension of the finite scaling ideas of Fisher [49] to the percolation conduction problem [32,57,97,112,143].

We explain the argument for the hypercubic lattice $\mathbb{Z}^d$, but the argument should apply for all periodic lattices. As the percolation threshold is approached, the correlation length $\xi(p)$ in the infinite lattice diverges:

$$\xi(p) \propto |p - p_c|^{-\nu} . \tag{59}$$

Although in principle the critical exponents for the correlation length for $p > p_c$ (ν') and for $p < p_c$ (ν) could differ, they have been proved rigorously to coincide for site percolation on the triangular lattice and there is no evidence to support their being different for other lattices. The argument is therefore written out in terms of ν.

Consider the finite hypercube $\Lambda_L \subset \mathbb{Z}^d$ introduced in Sect. "Defining the Conductivity". The finite lattice Λ_L will represent the infinite lattice to a good approximation provided that $\xi(p) \ll L$. We assume that the conductivity $\sigma_L(p)$ of the finite lattice depends on $|p - p_c|$ only through the ratio $\xi(p)/L \propto (L|p - p_c|^\nu)^{-1}$, and we write

$$\sigma_L(p) \approx L^{-\kappa} f(L|p - p_c|^\nu) . \tag{60}$$

Since the conductivity is nonzero at $p = p_c$ for some realizations of the finite lattice, $f(0) \neq 0$. We assume that $f(u) \propto u^\lambda$ as $u \to \infty$ so that $\sigma_L(p) \propto L^{-\kappa}[L|p - p_c|^\nu]^\lambda$. To recover the appropriate asymptotic law (44) for the conductivity of the infinite lattice, that is, $\sigma_{\text{eff}} \propto (p - p_c)^t$ as $p \to p_c^+$, we require $\lambda = \kappa$ (to cancel out the L-dependence), and to obtain the correct dependence on $p - p_c$, we need $t = \nu\lambda$. Hence

$$\sigma_L(p) \approx L^{-t/\nu} f(L|p - p_c|^\nu) \tag{61}$$

and in particular

$$\sigma_L(p_c) \approx L^{-t/\nu} f(0) . \tag{62}$$

The best available estimates of t come from the use of this asymptotic relation to determine t/ν and then using the exact value $\nu = \frac{4}{3}$ for $d = 2$ or independent numerical estimates of ν for $d \geq 3$.

In comparison with estimating the purely geometric or topological parameters of a percolation model, conductivity estimates are comparatively expensive in terms of computer time. There is a trade-off between having high precision in the values of the average conductivity of the finite lattice and having large enough L values to use the asymptotic relation (62). The most precise estimates of t/ν presently available come from Monte Carlo simulations, in which a random subset of the possible realizations of the lattice is generated for a sequence of values of L. Although the difference equations for the potential can be solved by standard techniques of numerical linear algebra, it is more efficient to perform a sequence of transformations that replace locally complicated small sections of the conducting backbone of the system by equivalent single bonds [32,57,97]; this is especially effective in two dimensions. An alternative technique for two-dimensional systems can be devised using transfer matrices, with the conductivity of one realization of an infinite strip of width L being estimated for a sequence of values of L [38,39], and a generalization of this method that works in higher dimensions is also available [28,72]. Other efficient methods of solving the relevant linear equations are also known [91].

Finite-size scaling arguments also apply to the superconductivity problem in which a fraction p of the bonds have infinite conductance, and one finds that

$$\sigma_L(p_c) \propto L^{s/\nu} \quad \text{as } L \to \infty . \tag{63}$$

A selection of the better estimates for t/ν and s/ν and corresponding predictions for t and s are given in Table 2 ($d = 2$) and Table 3 ($d = 3$), together with a few estimates not derived via finite-size scaling. A number of plausible conjectures on relations between various exponents, based on heuristic arguments or phenomenological relations observed from early numerical estimates, are now comprehensively refuted [57,75]. Grassberger [57] has carefully assessed the relative merits of different schemes for estimating t/ν in two dimensions, and notes major issues with the quality of random number generators used for some simulations, and problems arising from inappropriate assumptions concerning correction to scaling terms for the

Conduction and Diffusion in Percolating Systems, Table 2

Estimates of the conductivity (t) and superconductivity (s) exponents in two dimensions. Where values of t/ν or s/ν are given, these values were obtained first by finite-size scaling or related ideas, and the values of t or s were subsequently deduced. In two dimensions it is known that $\nu = \frac{4}{3}$ exactly (asterisked entries use this value to compute t or s from t/ν or s/ν), and also that $s = t$ exactly, but no such results are available for $d = 3$

t/ν	t	Source
0.95 ± 0.05		Real metal-insulator mixtures [123]
0.95 ± 0.01	1.28 ± 0.03	Transfer matrix [38]
0.968 ± 0.005	$\approx 1.291^*$	transfer matrix (bond) [170]
0.970 ± 0.009	$\approx 1.293^*$	Enumerate random walks on backbone [73]
≈ 0.972	$\approx 1.296^*$	Simulation of random walks [131]
$0.973^{+0.005}_{-0.003}$	$1.297^{+0.007}_{-0.004}$	Finite-size scaling [98]
0.975 ± 0.005	$\approx 1.300^*$	transfer matrix (site)[170]
0.979 ± 0.006	$\approx 1.305^*$	finite-size scaling [132]
	1.31 ± 0.04	Monte Carlo [50]
0.9825 ± 0.0008	$1.3100 \pm 0.0011^*$	Finite-size scaling (bond and site) [57]
s/ν	s	Source
0.9745 ± 0.0015	1.299 ± 0.002	Special purpose computer [118]
0.977 ± 0.010	$\approx 1.303^*$	Transfer matrix [72]

Conduction and Diffusion in Percolating Systems, Table 3

Estimates of the conductivity (t) and superconductivity (s) exponents in three dimensions. Where values of t/ν or s/ν are given, these values were obtained first by finite-size scaling or related ideas, and the values of t or s were subsequently deduced. Neither ν nor p_c are known exactly and estimates are sensitive to choices made by the cited authors. All estimates are for the simple cubic lattice, except for the random walk estimate [136] for the simple cubic lattice site problem

t/ν	ν used	t	Source
2.095 ± 0.016	0.89 ± 0.01	1.867 ± 0.035	Finite-size scaling [143]
2.26 ± 0.04			Special purpose computer [120]
2.276 ± 0.012	0.88 ± 0.02	2.003 ± 0.047	Finite-size scaling [55]
2.282 ± 005			Current distribution moments [8]
2.305 ± 0.015	≈ 0.88	≈ 2.0	Finite-size scaling [32]
		2.02 ± 0.02	Monte Carlo as $p \to p_c$ [32]
≈ 2.315			Generalized transfer matrix [28]
2.48 ± 0.07			Random walks [136]
s/ν	ν used	s	Source
0.782 ± 0.019			Finite-size scaling [137]
0.85 ± 0.04	≈ 0.88	≈ 0.75	Tranfer matrix [72]
0.835 ± 0.005			Special purpose computer [119]

square lattice bond problem. The square lattice site problem, and the bond problems on the triangular and honeycomb lattices, are better behaved [57,133]. Early estimates of t and s by finite-size scaling for lattices with unknown percolation thresholds are partly compromised by inaccuracy in numerical estimates of thresholds, but in some cases numerical estimates of thresholds are now available to extravagant precision (e. g., square lattice site problem, $p_c = 0.592\,746\,5 \pm 0.000\,000\,4$; simple cubic lattice site problem, $p_c = 0.311\,607\,7 \pm 0.000\,000\,4$ [37]).

A weak inequality relating the critical exponents t and ν can be deduced on the basis of finite-size scaling. For finite fragments of the square lattice, a duality argument shows that there is probability $\frac{1}{2}$ that the fragment is spanned by a conducting path at $p = \frac{1}{2}$. Each such realization of the system gives a conductance between the connected sides of the lattice which is no smaller than aL^{-2}, this worst case arising when there is a single path passing through all lattice sites. It follows that for the square lattice, $\sigma_L(p_c) \geq (a/2)ML^{-2}$ and so $t \leq 2\nu$ in two dimensions.

Conduction on Fractals Some progress has been made for percolation conduction problems on self-similar fractal structures [100], which have been proposed as possible models for the backbone of large clusters at the percolation threshold [53]. The plane Sierpinski lattice (triangular gasket) has percolation threshold $p_c = 1$, but unlike the linear chain which also has $p_c = 1$, the Sierpinski lattice has nontrivial finite-size scaling behavior, and the analogue of t/ν is $\log(5/3)/\log(2) \approx 0.73$ [53]. The analogue of s/ν has been evaluated numerically for this system as 0.27 ± 0.03 [162]. For a further discussion of fractal structures in the context of percolation see ► Scaling Properties, Fractals, and the Renormalization Group Approach to Percolation.

Structural Speculations

Despite recent spectacular advances in the understanding of two-dimensional percolation mentioned in Sect. "Rigorous Results for two Dimensions", an adequate theory for the conductivity exponent t and superconductivity exponent s remains elusive, even in two dimensions. A number of attempts have been made to model the structure of the backbone just above the percolation, and to model what physicists have called the "incipient infinite cluster" at the percolation threshold p_c, in the hope of predicting the values of t, s and related "dynamical" exponents of percolating systems.

The Backbone The backbone in a percolating system includes all sites with two independent connections to infinity. Only the backbone can carry current in the random resistor problem, although because of local symmetries, not all bonds of the backbone will carry nonzero current. On periodic lattices, the existence of a well-defined nonzero backbone probability $B(p)$ for $p > p_c$ ensures (as a consequence of ergodic theory [18]) a well-defined density of backbone sites, and so the number $B_L(p)$ of sites with two disjoint connections to the boundary in a hypercube of side L is asymptotically a multiple of $L^d B(p)$ as $L \to \infty$. Natural finite-size scaling analysis [155] (cf. Sect. "Numerical Estimates of the Conductivity Critical Exponent t" above) based on the functional form $B_L(p) = L^\kappa b(L|p - p_c|^\nu)$ with $b(z) \propto z^\lambda$ as $z \to \infty$ and the requirement that $B(p) \propto (p - p_c)^{\beta_{bb}}$ as $p \downarrow p_c$ leads to the conclusion that $d = \kappa + \lambda$ and $\beta_{bb} = \lambda\nu$. Thus, in particular, $B_L(p_c) \propto L^{d_{bb}}$, where $d_{bb} = d - \beta_{bb}/\nu$ is a formula believed to hold by physicists for $d \le 6$, interpreted by them as the fractal dimension [100] of the backbone at the percolation threshold, and estimated by simulation. Grassberger [57] estimates that $d_{bb} = 1.6432 \pm 0.0008$ from a careful study of the square lattice site and bond percolation problems. The occurrence of this noninteger exponent suggests delicate structure of the backbone just above the percolation threshold.

The first attempts to model the backbone in detail and thereby make predictions about the conductivity exponent t date from 1975 and 1976 and are due to de Gennes [34] and Skal and Shklovskii [147]. In their view, the backbone consisted of a "superlattice" of nodes connected by macrobonds or links. If the node spacing is estimated as $\xi(p)$, the correlation length, and the conductivity of a macrobond is given by $g(p) \propto (p - p_c)^\zeta$ for some exponent ζ, then the effective conductivity is predicted to scale as

$$\sigma_{\text{eff}}(p) \approx \xi(p)^{2-d} g(p) \propto (p - p_c)^{(d-2)\nu+\zeta} , \qquad (64)$$

the factor of $\xi(p)^{2-d}$ being motivated by Eq. (39). This prediction implies that the conductivity exponent is $t = \nu(d-2) + \zeta$, and it was suggested [34] that $\zeta = 1$ in all dimensions. However, Chayes and Chayes [31] proved rigorously that $\zeta > \nu$ in all dimensions, implying that $t > \nu(d-1)$. For $d = 2$, this gives $t > \frac{4}{3}$, a result which is no longer credible given the precision recently attained in numerical estimates and the simple superlattice model is unviable.

Problems with the superlattice model were already noted by Kirkpatrick [90] in 1978. He proposed instead a self-similar fractal model, and this idea was further pursued by Gefen et al. [53]. An implication of the self-similar fractal model that $t \le (d-1)\nu$ is consistent with numerical data in low dimensions, but fails for $d = 6$, where it is accepted that $\nu = \frac{1}{2}$ and $t = 3$.

The most plausible picture of the backbone structure just above the percolation threshold is inelegantly described as the "nodes–links–blobs" model [151,152], with approximately self-similar structures over smallest scales linked together in some way by tortuous, quasi-one-dimensional links but the structure is far from completely characterized.

The Incipient Infinite Cluster Since the percolation probability $P_\infty(p)$ is proven rigorously to be zero at the percolation threshold for several two-dimensional percolation problems, there is with probability 1 no infinite cluster present at the percolation threshold. The *incipient infinite cluster* at the percolation threshold in the sense originally used by physicists is therefore a dubious construct. Kesten has given a rigorous discussion of two mathematically respectable candidates for the incipient infinite cluster [59,86], using appropriately defined conditional probabilities.

(i) Take $p > p_c$ and work with conditional probabilities, conditioning on the event that the origin is part of an infinite cluster. Then take the limit $p \downarrow p_c$.
(ii) Take $p = p_c$ and work with conditional probabilities, conditioning on the event that the origin belongs to a connected cluster that reaches the boundary of a box of side length $2n$ centered on the origin. Then take the limit as $n \to \infty$.

For bond percolation on the square lattice, these two apparently different definitions have been proved consistent [86]. Moreover, for this specific problem, it can be shown that the expected number of sites of the incipient infinite cluster in a box of side $2n$ is (to within a slowly-varying prefactor) $n^{2-\eta}$, where the exponent η governs the decay of the probability that the origin is connected to a box of side $2n$. It has been proved that for the triangular lattice bond problem, $\eta = \frac{5}{24}$ [148], and this should be generally true for two dimensions. In physicists' terminology, this shows that the fractal dimension of the incipient infinite cluster in two dimensions is $d_{\mathrm{iic}} = \frac{43}{24} \approx 1.792$, which is outside the 1983 estimates $d_{\mathrm{iic}} = 1.90 \pm 0.1$ [130] and $d_{\mathrm{iic}} = 1.900 \pm 0.009$ [81].

The theory of the incipient infinite cluster and its relation to finite-size scaling considerations are completely resolved in two dimensions [23,76]. Some illuminating if less complete conclusions are available in higher dimensions [23]. Physicists' scaling theories suggest that $d_{\mathrm{iic}} = d - \beta/\nu$ for $2 \le d \le 6$ [155]. However, it is not the incipient infinite cluster but rather the backbone that is core to the determination of the conductivity exponent, so that these results do not go far enough for the present purposes.

Random Motion in a Random Environment

One may replace the steady-state transport and conduction processes discussed in Sect. "Continuum Models: Steady-State Phenomena" and "Lattice Models: Steady-State Phenomena" for continua and lattices respectively by time-dependent analogues. Our discussion of the continuum version is very brief.

Continuum Problems

As noted in Sect. "Caveats to the Electrical Interpretation", problems of diffusion in random media are easily defined. An account of the homogenization of the diffusion Eq. (5), including references to earlier work, will be found in Jikov et al. [79]. In particular for random media subject a uniform ellipticity condition (equivalent to bounding the local diffusivity away from zero) and with a spatially stationary distribution of the local diffusivity, the system can be homogenized for almost all realizations of the random diffusivity distribution, and the corresponding homogenized diffusivity is independent of the realization. Percolative problems, where the diffusivity is zero at a given arbitrary point with nonzero probability are more difficult to analyze, but Jikov et al. [79] give a proof that for the random chessboard (squares with diffusivities 0 or 1 with probabilities $1 - p$ and p respectively), the scaled displacement $\mathbf{X}(t)$ of a diffusion process commencing in the conducting phase converges in distribution to a zero mean, isotropic bivariate normal random variable. If D_0 denotes the homogenized diffusivity for the steady-state problem treated like a steady conduction problem, then the diffusivity governing the time-dependent problem is $D_0/P_\infty(p)$, where $P_\infty(p)$ is the density of the infinite conducting cluster (and indeed equal to the percolation probability for site percolation on the square lattice).

For a discussion of many aspects and applications of diffusion in condensed matter physics, an area that has sustained interest in the properties of heterogeneous and randomly microstructured materials for a long time, see the collection edited by Heitjans and Kärger [71]. One interesting diffusion problem that we do not address, since it does not manifest percolation properties, arises when one phase acts as an irreversible trap. The survival probability diffusing objects in the presence of dilute random traps decays as $\exp(-kt^{3/5})$ at large times t in three dimensions rather than the naively expected $\exp(-kt)$, due to the unexpected statistical importance of rare, large trap-free regions [58,74,80].

Lattice Problems

The lattice-based models for stochastic transport processes in random environments that we address may loosely be described as *random walks*, since they derive conceptually from the random walk problems first posed by in 1905 by Pearson [126] and in 1919 by Pólya [128,129], but the precise meaning of the phrase *random walk* varies widely in the literature. The classic account of the traditional theory of discrete time, translationally invariant random walks on $\mathbb{Z}^d$ is Spitzer [150]; substantial extensions to the theory needed for more general problems will be found in Doyle and Snell [40], Hughes [74,75], Telcs [163] and Woess [169]. The most recent comprehensive rigorous reviews of random walks in random environments are the papers by Zeitouni [172,173]. For other discussions of random walks in random environments see the books by Hughes [75] and Révész [134].

Nearest-Neighbour Stepping Discrete-Time Random Walk We consider the case in which the probability that a walker presently at site $\mathbf{s}'$ next steps to site $\mathbf{s}$ is $p(\mathbf{s}|\mathbf{s}')$, nonzero only when $\mathbf{s}$ is a nearest neighbor of $\mathbf{s}'$. The walk is called *recurrent* if there is probability 1 that the walker eventually returns to the starting site and *transient* otherwise. A recurrent walker is certain to visit all sites accessible to it, and visits its own starting site infinitely often (with probability 1). A transient walker revisits the starting site at most finitely many times.

The walk is of Pólya type if $p(\mathbf{s}|\mathbf{s}') = 1/z(\mathbf{s}')$, where $z(\mathbf{s}')$ is the coordination number of $\mathbf{s}'$, that is, all nearest-neighbor steps are equally likely. Pólya's most famous result is that for Pólya walk on $\mathbb{Z}^d$, the walk is recurrent for $d = 1$ and $d = 2$, but transient for $d \geq 3$. Translationally invariant walks are well-suited to Fourier analysis. This is how much of the classical theory is derived, and the discrete time evolution is simplified by generating function methods [74,150]. In particular, for Pólya walks on $\mathbb{Z}^d$, the position distribution of the random walker is asymptotically Gaussian (normal) at large times and the probability $p_n(\mathbf{s}_0)$ that the walker will be found at the starting site $\mathbf{s}_0$ after n steps decays in proportion to $n^{-d/2}$.

The number S_n of distinct sites visited in the first n steps is of interest in a number of contexts [74]; in the probability literature this quantity is sometimes misleadingly called the range of the random walk. For classical Pólya walks on $\mathbb{Z}^d$, the mean number of distinct sites visited has the asymptotic behavior [115]

$$\langle S_n \rangle \sim \begin{cases} (8n/\pi)^{1/2}, & d = 1\,, \\ \pi n/\ln n, & d = 2\,, \\ (1-R)n, & d \geq 3\,, \end{cases} \tag{65}$$

where R is the probability of eventual return to the starting site. The following more general results hold [74] for translationally invariant walks on periodic lattices. For transient walks, $\langle S_n \rangle \sim (1-R)n$. For recurrent walks, if $p_n(\mathbf{s}_0) \sim \text{constant} \times n^{-H/2}$ with $0 < H < 2$, or equivalently,

$$\sum_{n=0}^{\infty} p_n(\mathbf{0})\xi^n \sim \text{constant} \times (1-\xi)^{H/2-1} \quad \text{as } \xi \uparrow 1\,, \tag{66}$$

then

$$\langle S_n \rangle \sim \text{constant} \times n^{H/2}\,. \tag{67}$$

The case $H = 2$ induces logarithmic prefactors. The exponent H is variously described as the *harmonic dimension* [74], *spectral dimension* or *fraction dimension* [2], and $H = d$ for the simplest problems on $\mathbb{Z}$ or $\mathbb{Z}^2$.

For random walk problems where the generating function and Fourier analysis techniques used to derive the preceding results are not available, the important questions to addressed in other ways are as follows.

(i) Is the walk recurrent or transient?
(ii) Is the walk diffusive (large-n Gaussian or normal limiting behavior) and if so, what is the value of diffusion constant?
(iii) In non-diffusive cases, how does the mean-square displacement grow with n?
(iv) What is the value of the harmonic dimension?

Canonical Models for Random Environments Random environments for random walkers to experience may be constructed in various ways. One simple way is to assign to each bond of the lattice $\mathbb{Z}^d$ (or some other periodic lattice) an independent nonzero weight, which might be interpreted as the conductance of a random resistor. If $G_{\mathbf{r},\mathbf{s}} = G_{\mathbf{r},\mathbf{s}}$ denotes the conductance of a bond joining nearest-neighbor sites $\mathbf{r}$ and $\mathbf{s}$, then the transition probability may be defined to be

$$p(\mathbf{s}|\mathbf{s}') = \frac{G_{\mathbf{s},\mathbf{s}'}}{\sum_{\mathbf{s}''} G_{\mathbf{s}'',\mathbf{s}'}}\,, \tag{68}$$

the sum over $\mathbf{s}''$ being restricted to nearest neighbors of $\mathbf{s}'$. If all bonds emanating from a given site have equal conductance, the walk is of Pólya type.

Consider the following two problems on the infinite conducting lattice just introduced:

- a discrete-time random walk, governed by (68) starting from site $\mathbf{s}_0$;
- electrical conduction between site $\mathbf{s}_0$ and infinity.

Then it can be shown [40] that the walk is transient if and only if the total resistance between $\mathbf{s}_0$ and infinity is finite. This result, and various other relations between random walks and electric networks [40,75,117], represent one of the most important avenues for progress in the study of random walks in random environments.

Pemantle and Peres [127] have shown that for independent, identically distributed conductances, the question of recurrence or transience of this model can be settled (for all realizations of the environment save for a set of zero probability) by an analysis of Pólya random walk on the random lattice arising from the bond percolation on the same lattice. The original walk is transient if and only if for some p with $p_c < p < 1$ the Pólya walk on the infinite cluster in the bond percolation problem is transient.

Master Equations and Generalized Master Equations Differential-difference equations of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}c(\mathbf{s},t) = \sum_{\mathbf{s}'} W(\mathbf{s},\mathbf{s}')c(\mathbf{s}',t) - M(\mathbf{s})c(\mathbf{s},t) \tag{69}$$

describe a process in which the concentration of a mobile substance at site $\mathbf{s}$ evolves by transfer of substance from other sites (gaining substance at rate $W(\mathbf{s},\mathbf{s}')c(\mathbf{s}',t)$ from site $\mathbf{s}'$) and loss of substance to other sites (total rate of loss $M(\mathbf{s})c(\mathbf{s},t)$ to all other sites). If the substance is to be globally conserved, one needs

$$M(\mathbf{s}) = \sum_{\mathbf{s}'} W(\mathbf{s}',\mathbf{s}) \tag{70}$$

and if

$$p(\mathbf{s},t) = \frac{c(\mathbf{s},t)}{\sum_{\mathbf{s}'} c(\mathbf{s}',t)} \tag{71}$$

denotes the normalized concentration then what physicists call the *master equation* [122],

$$\frac{\mathrm{d}}{\mathrm{d}t}p(\mathbf{s},t) = \sum_{\mathbf{s}'}[W(\mathbf{s},\mathbf{s}')p(\mathbf{s}',t) - W(\mathbf{s}',\mathbf{s})p(\mathbf{s},t)]\,, \tag{72}$$

results. This equation is interpreted as the law governing the probability of the position of some type of random walk process in continuous time, and indeed is a standard equation from the theory of Markov processes.

To obtain a richer class of possible behaviors, physicists have also considered *generalized master equations* [84], where the transition rates $W(\mathbf{s},\mathbf{s}')$ are replaced by memory kernels $W(\mathbf{s},\mathbf{s}',t)$:

$$\frac{\mathrm{d}}{\mathrm{d}t}p(\mathbf{s},t) = \int_0^t \sum_{\mathbf{s}'}[W(\mathbf{s},\mathbf{s}',t-t')p(\mathbf{s}',t') - W(\mathbf{s}',\mathbf{s},t-t')p(\mathbf{s},t')]\mathrm{d}t'\,. \tag{73}$$

For Eqs. (72) and(73) the transition rates or memory kernels are often nonzero only for nearest-neighbor sites. Simple models of random media arise by imposing symmetry,

$$W(\mathbf{s},\mathbf{s}') = W(\mathbf{s}',\mathbf{s}) \quad \text{or} \quad W(\mathbf{s},\mathbf{s}',t) = W(\mathbf{s}',\mathbf{s}',t),$$

and making the rate coefficients $W(\mathbf{s},\mathbf{s}')$ or parameters in the memory kernels $W(\mathbf{s},\mathbf{s}',t)$ independent random variables. If we introduce Laplace transforms, writing

$$\widehat{f}(u) = \int_0^\infty e^{-ut}f(t)\mathrm{d}t\,, \tag{74}$$

then the transformed master and generalized master equations for a process started at the origin site $\mathbf{0}$ become

$$u\widehat{p}(\mathbf{s},u) - \delta_{\mathbf{s},\mathbf{0}} = \sum_{\mathbf{s}'} W(\mathbf{s},\mathbf{s}')[\widehat{p}(\mathbf{s}',u) - \widehat{p}(\mathbf{s},u)], \tag{75}$$

$$u\widehat{p}(\mathbf{s},u) - \delta_{\mathbf{s},\mathbf{0}} = \sum_{\mathbf{s}'} \widehat{W}(\mathbf{s},\mathbf{s}',u)[\widehat{p}(\mathbf{s}',u) - \widehat{p}(\mathbf{s},u)]. \tag{76}$$

In both cases, the Laplace transform equations are able to be interpreted as the equations governing a random resistor network, with all sites having an additional connection to zero potential (an earth connection) of conductance u. If we attempt to match Eq. (75), with random rates $W(\mathbf{s},\mathbf{s}')$, to an equivalent uniform system with rates W_*, then it should be anticipated that W_* is a function of the variable u, since the relative importance of the earth connection at $\mathbf{s}$ compared to its connections to other sites will fluctuate over the lattice. Consequently, as discussed in Sect. "Exactly and Approximately Solved Continuous-Time Problems", the appropriate real-time description of the system may be expected to correspond to a uniform memory kernel generalized master equation, rather than a simple, uniform transition rate master equation.

Exactly Solved Discrete-Time Problems Elegant exact results are scarce for random processes in random environments, with the strongest and most informative results limited to one dimension (that is, on $\mathbb{Z}$), where percolative problems become essentially trival: for $p < 1$ a Pólya walker is confined to a finite interval and essentially its position has a limiting distribution (to be more precise, the positions after even or odd numbers of steps have limiting distributions, which may be different). Averaging over all realizations of the environment we find that the mean-square displacement after n steps is $o(n)$. This unusual behavior is an artifact of averaging: as n increases, in more and more realizations the walker's displacements have saturated and only the effects of increasingly rare large intervals contribute to the continued growth of the mean-square displacement, and various properties of the system can be deduced by appropriate averaging over the positions of the left and right boundary sites [121].

A less trivial one-dimensional model introduced in 1972 by Temkin [164] brings out the subtle possibilities for random walks in random environments. Consider the stepping law

$$p(l|l') = A_{l'}\delta_{l,l'+1} + (1 - A_{l'})\delta_{l,l'-1}\,, \tag{77}$$

where $\{A_l\}$ is a set of independent, identically distributed random variables. Let an overbar denote an average over all realizations of the set $\{A_l\}$. Solomon [149] has shown

that: a walker initially stepping right is certain to return if and only if $\overline{\log[(1-A)/A]} \geq 0$; a walker initially stepping left is certain to return if and only if $\overline{\log[A/(1-A)]} \geq 0$, and the walk is recurrent if and only if $\overline{\log[A/(1-A)]} = 0$. In 1982, Sinai [146] produced the astonishing result that in the recurrent case, on the additional assumption that A_l is bounded away from 0 and 1, the mean-square displacement is $O(\ln^4 n)$, in place of the standard $O(n)$ for one-dimensional Pólya walkers.

Exactly and Approximately Solved Continuous-Time Problems Alexander et al. [3] have discussed exactly solvable one-dimensional master equations with independent, identically distributed random transition rates. If the probability density function for the rates is denoted by $f(w)$, they show that the mean-square displacement grows as $2W_0 t$, and the random system is equivalent in its long-time properties to a uniform system in which all transition rates are W_0, where

$$\frac{1}{W_0} = \int_0^\infty \frac{f(w)\mathrm{d}w}{w}\,, \tag{78}$$

so long as the integral on the right is finite. When this is not the case, the random motion is *sub-diffusive* : if $f(w) \to f(0) > 0$ as $w \to 0$ then the mean-square displacement grows as $t/\ln t$, while if $f(w) \propto w^{-\alpha}$ as $w \to 0$ the mean-square displacement grows as $t^{(2-2\alpha)/(2-\alpha)}$ (distribution-induced nonuniversality). There are several of other approaches to exactly solvable one-dimensional master equation problems, including asymmetric rate problems [75].

A number of authors have extended the idea of an effective medium approximation from the random resistor problem (Sect. "Effective Medium Approximations") or other contexts to the problem of random master equations. For historical details and a full account of the analysis see Sahimi et al. [142] or Hughes [74]. As noted in Sect. "Master Equations and Generalized Master Equations", in the Laplace transform domain, the master equation with random coefficients is equivalent to a random resistor network, and the uniform transition rate produced by the matching procedure is a function of the transform variable, so that the approximately equivalent uniform system is governed by a generalized master equation. In the percolative case in which the rate coefficient associated with a bond is nonzero only with probability $p < 1$, the approximately equivalent uniform system does not support motion if $p < 2/z$ (the predicted percolation threshold, where z is the coordination number of the underlying lattice on which the percolation process is realized). For $p \downarrow 2/z$, the effective diffusion constant $D(p)$ (a constant multiple of the time derivative of the mean-square displacement) is predicted to vanish linearly with $p - 2/z$ for the standard case where all nonzero rate coefficients are equal, but the exponent is predicted to change in more general cases if the average of W^{-1} is infinite, where W is the random rate associated with an arbitrary bond, conditional on the rate being nonzero.

In one dimension, the effective medium approximation reproduces a number of the exact results of Alexander et al. [3]. Its predictions for higher-dimensional systems, while not of great accuracy, do have one interesting aspect. Generalized master equations with memory kernels $W(\mathbf{s}, \mathbf{s}', t) = \phi(t)p(\mathbf{s}|\mathbf{s}')$ are naturally associated with continuous-time random walk processes in which the walker waits for a random time with probability density function $\psi(t)$ between successive steps, and then moves with the stepping law $p(\mathbf{s}|\mathbf{s}')$. The classic exponential waiting time density $\psi(t) = \kappa \exp(-\kappa t)$ arises if and only $\phi(t) = \kappa\delta_+(t)$ and the generalized master equation reduces to the ordinary master equation in this case. In general, the connection between ϕ and ψ is most simply expressed in terms of the Laplace transform: $\widehat{\psi}(u) = \widehat{\phi}(u)/[u + \widehat{\phi}(u)]$ [74,85]. Any form of nondiffusive behavior that may arise is naturally associated with a non-exponential waiting time density, and for $p < 2/z \approx p_c$ the associated waiting time is predicted to be defective, that is,

$$\int_0^\infty \psi(t)\mathrm{d}t < 1 \tag{79}$$

and the equivalent continuous time random walker takes only finitely many steps. An alternative approach to modeling nondiffusive behavior of possible relevance to percolative systems can be developed using fractional calculus ideas [107,108].

The Ant in the Labyrinth

We conclude our survey of transport and conduction in random environments, with a discussion for discrete-random walks of the problem of the ant in the labyrinth, introduced in 1975 by Brandt [24] and proposed by de Gennes [33] as a probe of the geometry of percolation. The ant, a random walker, moves through a random labyrinth generated by applying site or bond percolation to a periodic lattice [24,33,75,110,111,113]. Because the analogue of a strong ellipticity condition is not satisfied for these percolative problems, many of the more general results from rigorous analyzes of random walks in random environments [172,173] do not apply.

Models Let the original lattice from which the labyrinth is made have coordination number z. Then at a site $\mathbf{s}$ of the labyrinth, the coordination number will be $Z_\omega(\mathbf{s}) \le z$, where $Z_\omega(\mathbf{s})$ is determined by the specific realization ω of site or bond percolation used in constructing the labyrinth. There are two canonical choices for the ant's mode of stepping through the labyrinth, the distinction between them having been clearly drawn by Mitescu and Roussenq [111] and Majid et al. [99].

myopic ant (Pólya case) —On arrival at site $\mathbf{s}$, the ant looks at the $Z_\omega(\mathbf{s})$ adjacent sites onto which it is permitted to step, assigns each of them probability $1/Z_\omega(\mathbf{s})$ and chooses one of them at random.

blind ant —On arrival at site $\mathbf{s}$, the ant attempts to move to one of the z adjacent sites on the original lattice. If this move is not allowed for the labyrinth, the ant remains at site $\mathbf{s}$.

In a given time interval, a blind ant visits fewer distinct sites than a myopic ant.

It is generally believed that the qualitative properties of myopic (Pólya) ants and blind ants are similar, but there is an important difference when the ant is introduced on a finite cluster [111]: the "equilibrium" probability distributions of the ant, approached asymptotically as the number of steps grows without bound, differ for the two models. The existence of equilibrium probability distributions on finite clusters can be discussed from the point of view of Markov chains [46], where such distributions are called "invariant measures".

Simulations suggest [41,113,125,145] that when the mean-square displacement $\langle R_n^2 \rangle$ is averaged over environments, as $n \to \infty$ we have

$$\overline{\langle R_n^2(p) \rangle} = \begin{cases} R_\infty^2 - A(p)\exp\{[-n/\theta(p)]^w\} + \cdots, & p < p_c, \\ Bn^{2k} + \cdots, & p = p_c, \\ CD(p)n + \cdots, & p > p_c, \end{cases} \tag{80}$$

where any dependence of a quantity on p or n is explicitly indicated. Here $D(p)$ is an effective diffusion constant and the exponent k, which would be 1/2 for classical diffusion, is dimension-dependent, with $k < 1/2$ in low-dimensional systems. In particular, estimates of k for the simple cubic lattice are 0.20 ± 0.01 [125] and 0.200 ± 0.002 [41]. The value of the exponent w was initially believed close to 1 [113] in three dimensions, but later evidence [125] suggests that $w \approx 0.4$ in three dimensions.

In the mathematical literature, careful distinctions are drawn between results that apply to the *annealed* case (corresponding to averaging over all realizations of the disordered system) and the *quenched* case (involving statements about realizations ω of the disorder. For $p \le p_c$ the ant commences with probability 1 on a cluster of finite size, and in any individual realization ω of the system, the mean-square displacement will converge to a finite value (blind case) or oscillate between two finite values (myopic case for some lattices, where oscillations can occur if the cluster divides into sites accessible only after an even number of steps and sites accessible only after an odd number of steps). For $p = p_c$, at a given value of n the displacement will have saturated (that is, be close to its asymptotic value) in some environments but not in others. This partitioning of the realizations of the environments, and the tortuous structure of large clusters that are not yet fully explored by the ant, give rise to the nonclassical exponent $k < 1/2$.

Scaling Theory for the Ant in the Labyrinth Heuristic work [9,69,75,124,155] raises attractive possible connections between exponents characterizing the asymptotic behavior of the ant in the labyrinth and the geometrical exponents of percolation theory. A number of rigorous scaling relations involving various exponents that characterize random walk and electrical conduction on lattices have been derived by Telcs [163], but these are not sufficiently informative to establish rigorously the results now briefly described or to enable exponent values to be predicted.

We shall denote averages for walks, conditioned on the walks taking place on clusters of m sites, by $\langle \cdot \rangle_m$. Consider first the expected number of distinct sites visited. At the percolation threshold, if there is some kind of effective harmonic dimension for large clusters (cf. Sect. "Lattice Problems"), which one might loosely call the harmonic dimension of the incipient cluster (cf. Sect. "The Incipient Infinite Cluster") and denote by H_{iic}, then the expected number of distinct sites visited by a walker on a cluster of m sites should evolve as $\langle S_n \rangle_m \propto n^{H_{\mathrm{iic}}/2}$ for $1 \ll \langle S_n \rangle_m \ll m$. The scaling hypothesis [124] $\langle S_n \rangle_m \approx n^{H_{\mathrm{iic}}/2}\phi(m/n^{H_{\mathrm{iic}}/2})$ and the asymptotic law $P_m(p) \propto m^{-1-1/\delta}$ for the cluster size distribution gives a simple prediction [6,124] for the mean number of distinct sites visited, averaged both over cluster sizes and over walks on individual clusters:

$$\overline{\langle S_n \rangle} \approx \sum_{m=1}^{\infty} m^{-1-1/\delta} n^{H_{\mathrm{iic}}/2}\phi(m/n^{H_{\mathrm{iic}}/2}) \approx n^{(1-1/\delta)H_{\mathrm{iic}}/2}; \tag{81}$$

the last result arises from approximating the sum by an integral. Three-dimensional simulations give $\overline{\langle S_n \rangle} \propto n^{0.54\pm0.02}$ [124].

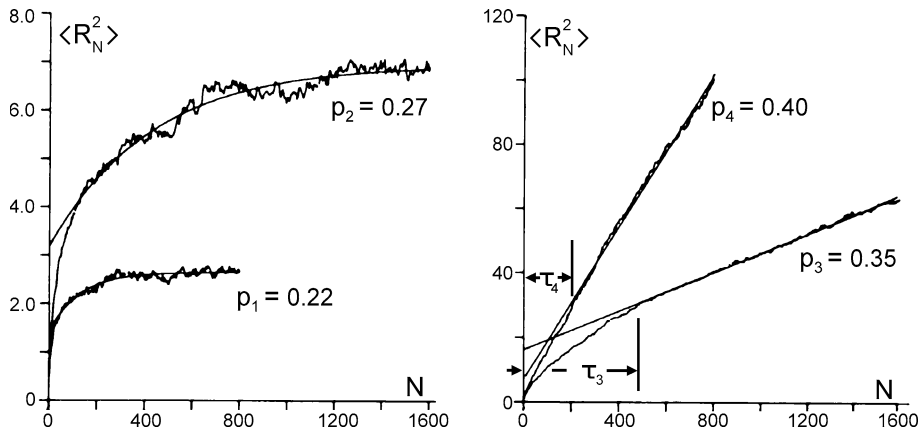

Conduction and Diffusion in Percolating Systems, Figure 3

Simulations of the ant in the labyrinth [113]: myopic ants (Pólya walkers) on the simple cubic lattice under site percolation ($p_c \approx 0.31$) for four values of p. Simulation data (*erratic curves*) shows the mean-square displacement after N steps, averaged over realizations of the environment when a fraction p of all sites are occupied: a shows $p < p_c$ ($p_1 = 0.22$, $p_2 = 0.27$), with the smooth curves least-squares fits to $R^2_\infty - A(p)\exp[-n/\theta(p)]$; b shows $p > p_c$ ($p_1 = 0.35$, $p_4 = 0.40$), with the straight-line asymptotes corresponding to effectively diffusive behaviour and τ_k corresponding to the number of steps needed for diffusive behaviour to be manifest for $p = p_k$ (Figure reproduced with permission from [114])

A similar analysis may be performed for the mean-square displacement at the percolation threshold [54,75]. If the fractal dimension of the incipient infinite cluster is denoted by d_{iic} (cf. Sect. "The Incipient Infinite Cluster") one may propose that the mean-square displacement on clusters of size m is $\langle R_n^2 \rangle_m \propto n^{2\nu_{\mathrm{iic}}}$ for $1 \ll \langle R_n^2 \rangle_m^{1/2} \ll m^{1/d_{\mathrm{iic}}}$. The naturally associated scaling assumption $\langle R_n^2 \rangle_m \approx n^{2\nu_{\mathrm{iic}}} \psi(m^{1/d_{\mathrm{iic}}}/n^{2\nu_{\mathrm{iic}}})$ predicts the mean-square displacement averaged both over cluster sizes and over walks on individual clusters to be

$$\begin{aligned} \overline{\langle R_n^2 \rangle} &\approx \sum_{m=1}^{\infty} m^{-1-1/\delta} n^{2\nu_{\mathrm{iic}}} \psi(m^{1/d_{\mathrm{iic}}}/n^{2\nu_{\mathrm{iic}}}) \\ &\approx n^{2\nu_{\mathrm{iic}}[1-d_{\mathrm{iic}}/(2\delta)]} \,. \end{aligned} \tag{82}$$

The accepted scaling relations $d_{\mathrm{iic}} = (\beta + \gamma)/\nu$ and $\gamma = \beta(\delta - 1)$ give the alternative form $\overline{\langle R_n^2 \rangle} \propto n^{2\nu_{\mathrm{iic}}[1-\beta/(2\nu)]}$ and simulation data is consistent with this picture [68].

A conjecture of Alexander and Orbach [2] published in 1982 suggesting that $H_{\mathrm{iic}} = 4/3$ for all dimensions d briefly raised the possibility that the conductivity exponent t could be simply related to geometrical exponents in percolation theory, and that other dynamical exponents related to random walk processes could be determined. Given the known geometrical exponents in two dimensions, the Alexander–Orbach conjecture predicted that $t = 91/72 \approx 1.2639$ in two dimensions. It is now generally accepted that the Alexander–Orbach conjecture, while an excellent approximation, is not precisely correct [9,116].

A number of other observable quantities can be estimated by scaling arguments [9,75,155]. For example, just below the percolation threshold, it is predicted [154] that the mean-square displacement (averaged over all walks and realizations) has the limiting value $R^2_\infty \propto (p_c - p)^{2\nu-\beta}$.

Rigorous Results for Discrete Time The most important rigorous result is probably the following theorem, which shows that Pólya's simple criterion for transience $d \geq 3$ is not destroyed by the percolation process, provided the system is above the percolation threshold, and the ant starts on the infinite cluster [60].

Theorem 1 (Theorem of Grimmett, Kesten and Zhang) *For myopic ants (Pólya walkers) walking on the infinite cluster generated by bond percolation on the simple hypercubic lattice at $p > p_c$, the walk is transient with probability 1 if $d \geq 3$.*

As with most rigorous results in this area, the proof is of antisocial length. However, Rayleigh's Monotonicity Law and the correspondence between random walks and electric circuits makes it relatively straightforward to show that the recurrence of the ordinary Pólya walk on the square lattice $\mathbb{Z}^2$ ensures the recurrence of myopic ant motion on every realization of an infinite cluster in bond or site percolation on the square lattice.

For walks on the incipient infinite cluster, a nebulous concept made precise by either of the equivalent ap-

proaches of Kesten discussed in Sect. "The Incipient Infinite Cluster", it is known [86] that the general behavior is subdiffusive ($|\mathbf{X}_n| = O(n^{1/2-\epsilon})$ for some $\epsilon > 0$) for the square lattice bond problem, but the exact exponent characterizing the growth of $|\mathbf{X}_n|$ is unknown.

Rigorous Results for Continuous Time For the natural master-equation analogue of the myopic ant (corresponding to a myopic ant executing a continuous-time random walk with an exponential waiting time density of mean waiting time 1 at all sites) several exact results that mirror known results for Pólya walks on periodic lattices are available. In stating these results, we assume that our labyrinth is generated by bond percolation on $\mathbb{Z}^d$ with $d \geq 2$ and $p > p_c$, we write C_∞^ω to denote the (unique with probability 1) infinite cluster in the realization ω of the random environment, and we assign coordinates such that the the origin $\mathbf{0}$ is a site of the infinite cluster. We write $\Pr^\omega$ to denote probabilities associated with random walks in the fixed realization ω of the environment. The position at time t of a random walker starting from the origin at time 0 is $\mathbf{X}_t$. We say that a result holds for almost all environments if it holds with probability 1, where probability is measured with respect to the realizations of bond percolation for the given value of p.

The following results, published in 2004 [7,103], go further than results previously established only for $d = 2$ [35,36] and show that there is no anomalous behavior above the percolation threshold for the myopic ant problem for all $d \geq 2$: in the long time limit, the process strongly resembles classical diffusion.

Mathieu and Remy [103] have proved that there exists $c_1(p, d)$ (independent of time or the realization ω of the environment) such that for almost all environments, the quenched bound

$$\sup_{\mathbf{y} \in C_\infty^\omega} \Pr^\omega\{\mathbf{X}_t = \mathbf{y}\} \leq \frac{c_1(p, d)}{t^{d/2}} \tag{83}$$

holds. This result confirms the transience of the walk for $d \geq 2$, a result already known from the theorem of Grimmett, Kesten and Zhang [60] for the discrete-time walk, stated in Sect. "Rigorous Results for Discrete Time". Mathieu and Remy also establish the analogous result to (83) for site percolation on the two-dimensional square lattice, and an annealed lower bound proportional to $t^{-d/2}$ for bond percolation on $\mathbb{Z}^d$.

Barlow [7] has constructed upper and lower bounds for the quenched transition density

$$q_t^\omega(\mathbf{x}, \mathbf{y}) = \frac{1}{Z^\omega(\mathbf{y})} \Pr^\omega\{X_{t+\tau} = \mathbf{y} \,|\, X_\tau = \mathbf{x}\} \tag{84}$$

that hold with probability 1 for sufficiently large time t:

$$\begin{aligned} &\frac{c_1(d, p)}{t^{d/2}} \exp\Big[-\frac{c_2(d, p)}{t}|\mathbf{x} - \mathbf{y}|_1^2\Big] \leq q_t^\omega(\mathbf{x}, \mathbf{y}) \\ &\leq \frac{c_3(d, p)}{t^{d/2}} \exp\Big[-\frac{c_4(d, p)}{t}|\mathbf{x} - \mathbf{y}|_1^2\Big] . \end{aligned} \tag{85}$$

Here if $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_d)$, one defines

$$|\mathbf{x} - \mathbf{y}|_1 = \sum_{k=1}^{d} |x_k - y_k| . \tag{86}$$

Sufficiently large t is quantified by $t \geq \sup\{|\mathbf{x} - \mathbf{y}|_1, S_\mathbf{x}^\omega\}$, where the finite but realization-dependent quantities $S_\mathbf{x}^\omega$ deal with problems associated with an initial period of potentially anomalous behavior. Barlow's results establish the essentially diffusive asymptotic behavior of the walk. For further results in the quenched case, see Berger and Biskup [13].

The above results pertain only to walks of Polyá type on the infinite cluster for $p > p_c$. For a discussion of biased random walks on the infinite cluster in two-dimensional percolation, see Berger et al. [14].

Future Directions

Problems of transport and conduction in random systems subject to the kinds of uniform ellipticity restrictions that preclude percolation phenomena have become increasingly well understood, though the address by Zeitouni [171] at the 2002 International Congress of Mathematicians summarizes a number of areas in which progress on the problem of random walk in a random environment remains to be made. Moreover, an adequate theoretical understanding does not entirely remove the tedious problem of accurately estimating overall system properties.

The situation concerning truly percolative problems remains less satisfactory. The generally classical behavior of such systems above the percolation threshold means that, for modeling purposes in science and technology, one often may reasonably replace the random system by a uniform analogue, in much the same way as one can do this for systems with ellipticity restrictions, but with the important caveat that close to the percolation threshold, the length or time scales at which the uniform treatment is accurate may in practice be unacceptably large. A fuller understanding therefore of processes at or near the percolation threshold remains the greatest challenge.

There has been spectacular progress since the late 1970s with the understanding (both heuristic and rigor-

ous) of the geometrical side of percolation theory, especially for lattices. We have elegant results on the existence of classical or mean-field behavior for large dimension d, and the knowledge (though gaps in rigor remain) that mean-field behavior applies for $d \geq 7$ and in a slightly weaker sense for $d = 6$ but not for smaller d. More importantly, through rigorous work on conformal invariance and on Schramm–Loewner evolution processes, exact two-dimensional values of geometrical critical exponents and exact two-dimensional scaling relations are now properly established. An analogous rigorous account of dynamical exponents characterizing transport in two-dimensional percolating systems is perhaps not too much to ask for, though the way is not yet clear. It would be surprising indeed if there were significant progress in the short term on the rigorous analysis of three-dimensional systems at or near the percolation threshold.

Bibliography

1. Ahmed G, Blackman JA (1979) On theories of transport in disordered media. J Phys C 12:837–853
2. Alexander S, Orbach R (1982) Density of states on fractals: 'fractons'. J Phys Lett 43:L625–L631
3. Alexander S, Bernasconi J, Schneider WR, Orbach R (1981) Excitation dynamics in random one-dimensional systems. Rev Mod Phys 53:175–198
4. Allaire G (1992) Homogenization and two-scale convergence. SIAM J Math Anal 23:1482–1518
5. Anderson PW (1958) Absence of diffusion in certain random lattices. Phys Rev 109:1492–1505
6. Angles d'Auriac JC, Rammal R (1983) Scaling analysis for random walk properties on percolation clusters. J Phys C 16:L825–L830
7. Barlow MT (2004) Random walks on supercritical percolation clusters. Ann Probab 32:3024–3084
8. Batrouni GG, Hansen A, Larson B (1996) Current distribution in the three-dimensional random resistor network at the percolation threshold. Phys Rev E 53:2292–2297
9. Ben-Avraham D, Havlin S (2000) Diffusion and Reactions in Fractals and Disordered Systems. Cambridge University Press, Cambridge
10. Bensoussan A, Lions J-L, Papanicolaou G (1978) Asymptotic Analysis for Periodic Structures. North–Holland, Amsterdam
11. Beran MJ (1965) Statistical Continuum Theories. Trans Soc Rheol 9:339–355
12. Beran MJ (1968) Statistical Continuum Theories. Wiley, New York
13. Berger N, Biskup M (2007) Quenched invariance principle for simple random walk on percolation clusters. Probab Theory Relat Fields 137:83–120
14. Berger N, Gantert N, Peres Y (2007) The speed of biased random walk on percolation clusters. arXiv:math/0211303v3; revised version of (2003) Probab Theory Relat Fields 126:221–242
15. Bergman DJ (1978) Analytical properties of the complex effective dielectric constant of a composite medium with applications to the derivation of rigorous bounds and to percolation problems. In: Garland JC, Tanner DB (eds) Electrical Transport and Optical Properties of Inhomogeneous Media, AIP Conference Proceedings, No 40. American Institute of Physics, New York, pp 46–61
16. Berlyand L, Golden K (1994) Exact result for the effective conductivity of a continuum percolation model. Phys Rev B 50:2114–2117
17. Bernasconi J, Schneider WR, Weismann HJ (1977) Some rigorous results for random planar conductance networks. Phys Rev B 16:5250–5255
18. Billingsley P (1965) Ergodic Theory and Information. Wiley, New York
19. Blumenfeld R, Meir Y, Aharony A, Harris AB (1987) Resistance fluctuations in randomly diluted networks. Phys Rev B 35:3524–3535
20. Blumenfeld R, Meir Y, Harris AB, Aharony A (1986) Infinite set of exponents describing physics on fractal networks. J Phys A 19:L791–L796
21. Bollobas B, Riordan O (2006) Percolation. Cambridge University Press, Cambridge
22. Borgs C, Chayes JT, Kesten H, Spencer J (1999) Uniform boundedness of critical crossing probabilities implies hyperscaling. Random Struct Algorithms 15:368–413
23. Borgs C, Chayes JT, Kesten H, Spencer J (2001) The birth of the infinite cluster: finite-size scaling in percolation. Commun Math Phys 224:153–204
24. Brandt WW (1975) Use of percolation theory to estimate effective diffusion coefficients of particles migrating on various ordered lattices and in a random network structure. J Chem Phys 63:5162–5167
25. Broadbent SR, Hammersley JM (1957) Percolation processes. I Crystals and mazes. Proc Camb Philos Soc 53:629–641
26. Brown WF (1955) Solid mixture permittivities. J Chem Phys 23:1514–1517
27. Bruggeman DAG (1935) Berechnung verschiedener physikalischer Konstanten von heterogen Substanzen. I Dielektrizitätskonstanten und Leitfähigkeiten der Mischkörper aus isotropen Substanzen. Annalen Phys (Leipzig) 24:636–679
28. Byshkin MS, Turkin AA (2005) A new method for the calculation of the conductivity of inhomogeneous systems. J Phys A 38:5057–5067
29. Calvert B, Keady G (1993) Braess's paradox and power-law nonlinearities in networks. J Aust Math Soc B 35:1–22
30. Chayes JT, Chayes L (1986) Bulk transport properties and exponent inequalities for random resistor and flow networks. Commun Math Phys 105:133–152
31. Chayes JT, Chayes L (1987) On the upper critical dimension of Bernoulli percolation. Commun Math Phys 113:27–48
32. Clerc JP, Podolskiy VA, Sarychev AK (2000) Precise determination of the conductivity exponent of 3D percolation using exact numerical renormalization. Eur Phys J B 15:507–516
33. de Gennes PG (1976) La percolation: un concept unificateur. Rech 7:919–927
34. de Gennes PG (1976) On a relation between percolation theory and the elasticity of gels. J Phys Lett 37:L1–L2
35. de Masi A, Ferrari PA, Goldstein S, Wick WD (1985) Invariance principle for reversible Markov processes with application to diffusion in the percolation regime. In: Durrett R (ed) Particle Systems, Random Media and Large Deviations, Contem-

porary Mathematics, Vol 41. American Mathematical Society, Providence, pp 71–85
36. de Masi A, Ferrari PA, Goldstein S, Wick WD (1989) An invariance principle for reversible Markov processes. Applications to random motions in random environments. J Stat Phys 55:787–855
37. Deng Y, Blöte HWJ (2005) Monte Carlo study of the site-percolation model in two and three dimensions. Phys Rev B 72:016126
38. Derrida B, Vannimenus J (1982) A transfer matrix approach to random resistor networks. J Phys A 15:L557–L564
39. Derrida B, Zabolitzky JG, Vannimenus J, Stauffer D (1984) A transfer matrix program to calculate the conductivity of random resistor networks. J Stat Phys 36:31–42
40. Doyle PG, Snell JL (1984) Random Walks and Electric Networks, Carus Mathematical Monograph No. 22. Mathematical Association of America, Washington
41. Duering E, Roman HE (1991) Corrections to scaling for diffusion exponents on three-dimensional percolation systems at criticality. J Stat Phys 64:851–858
42. Eggarter TP, Cohen MH (1970) Simple model for density of states and mobility of an electron in a gas of hard-core scatterers. Phys Rev Lett 25:807–810
43. Einstein A (1906) Eine neue Bestimmung der Moleküldimensionen. Annalen Phys 19:289–306
44. Einstein A (1911) Berichtigung zu meiner Arbeit: 'Eine neue Bestimmung der Moleküldimensionen'. Annalen Phys 34:591–592
45. Fatt I (1956) The network model of porous media [in 3 parts]: I—Capillary pressure characteristics; II—Dynamic properties of a single size tube network; III—Dynamic properties of networks with tube radius distribution. Trans Am Inst Min Metall Petroleum Eng, Petroleum Branch 207:144–159, 160–163, 164–177
46. Feller W (1970) An Introduction to Probability Theory and its Applications, vol 1, 3rd edn. Wiley, New York
47. Feller W (1971) An Introduction to Probability Theory and its Applications, vol 2, 2nd edition. Wiley, New York
48. Feng S, Halperin BI, Sen PN (1987) Transport properties of continuum systems near the percolation threshold. Phys Rev B 35:197–214
49. Fisher ME (1971) The theory of critical point singularities. In: Green MS (ed) Critical Phenomena: Enrico Fermi Summer School. Academic Press, New York, pp 1–99
50. Fogelholm R (1980) The conductivity of large percolation network samples. J Phys C 13:L571–L574
51. Fortuin CM (1972) On the random cluster model. II The percolation model. Physica 58:393–418
52. Fortuin CM, Kasteleyn PW (1972) On the random cluster model. I Introduction and relation to other models. Physica 57:536–564
53. Gefen Y, Aharony A, Mandelbrot BB, Kirkpatrick S (1981) Solvable fractal family and its possible relation to the backbone at percolation. Phys Rev Lett 47:1771–1774
54. Gefen Y, Aharony A, Alexander S (1983) Anomalous diffusion on percolating clusters. Phys Rev Lett 50:77–80
55. Gingold DB, Lobb CJ (1990) Percolative conduction in three dimensions. Phys Rev B 42:8220–8224
56. Golden K, Papanicolaou G (1983) Bounds for effective parameters of heterogeneous media by analytic continuation. Commun Math Phys 90:473–491
57. Grassberger P (1999) Conductivity exponent and backbone dimension in 2-d percolation. Physica A 262:251–263
58. Grassberger P, Procaccia I (1982) The long-time properties of diffusion in a medium with static traps. J Chem Phys 77:6281–6284
59. Grimmett G (1999) Percolation, 2nd edition. Springer, Berlin
60. Grimmett GR, Kesten H, Zhang Y (1993) Random walk on the infinite cluster of the percolation model. Probab Theory Relat Fields 96:33–44
61. Gu GQ, Yu KW (1992) Effective conductivity of nonlinear composites. Phys Rev B 46:4502–4507
62. Hammersley JM (1957) Percolation processes. II. The connective constant. Proc Camb Philos Soc 53:642–645
63. Hammersley JM (1957) Percolation processes. Lower bounds for the critical probability. Ann Math Stat 28:791–795
64. Hammersley JM (1961) Comparison of atom and bond percolation processes. J Math Phys 2:728–733
65. Hammersley JM (1988) Mesoadditive processes and the specific conductivity of lattices. J Appl Probab, Special vol 25A, edited by Gani J, 347–358
66. Hara T, Slade G (1994) Mean-field behaviour and the lace expansion. In: Grimmett G (ed) Probability and Phase Transition. Kluwer, Dordrecht, pp 87–122
67. Hashin Z, Shtrikman S (1962) A variational approach to the theory of the effective magnetic permeability of multiphase materials. J Appl Phys 33:3125–3131
68. Havlin S and Ben-Avraham D (1983) Diffusion and fracton dimensionality on fractals and on percolation clusters. J Phys A 16:L483–L487
69. Havlin S, Ben-Avraham D, Sompolinsky H (1983) Scaling behavior of diffusion on percolation clusters. Phys Rev A 27:1730–1733
70. Heinrichs J, Kumar N (1975) Simple exact treatment of conductance in a random Bethe lattice. J Phys C 8:L510–L516
71. Heitjans P, Kärger J (eds)(2005) Diffusion in Condensed Matter: Methods, Material, Models. Springer, Berlin
72. Herrmann HJ, Derrida B, Vannimenus J (1984) Superconductivity exponents in two- and three-dimensional percolation. Phys Rev B 30:4080–4082
73. Hong DC, Havlin S, Herrmann HJ, Stanley HE (1984) Breakdown of the Alexander–Orbach conjecture for percolation: exact enumeration of random walks on percolation backbones. Phys Rev B 30:4083–4086
74. Hughes BD (1995) Random Walks and Random Environments, vol 1: Random Walks. Clarendon Press, Oxford
75. Hughes BD (1996) Random Walks and Random Environments, vol 2: Random Environments. Clarendon Press, Oxford
76. Járai AA (2003) Incipient infinite percolation clusters in 2D. Ann Probab 31:444–485
77. Jerauld GR, Hatfield JC, Scriven LE, Davis HT (1984) Percolation and conduction on Voronoi and triangular networks: a case study in topological disorder. J Phys C 17:1519–1529
78. Jerauld GR, Scriven LE, Davis HT (1984) Percolation and conduction on the 3D Voronoi and regular networks: a second case study in topological disorder. J Phys C 17:3429–3439
79. Jikov VV, Kozlov SM, Oleinik OA (1994) Homogenization of Differential Operators and Integral Functionals. Springer, Berlin
80. Kayser RF, Hubbard JB (1983) Diffusion in a medium with a random distribution of static traps. Phys Rev Lett 51:79–82

81. Kapitulnik A, Aharony A, Deutscher G, Stauffer D (1983) Self-similarity and correlations in percolation theory. J Phys A 16:L269–L274
82. Keller JB (1964) A theorem on the conductivity of a composite medium. J Math Phys 5:548–549
83. Kenkel SW, Straley JP (1982) Percolation theory of nonlinear circuit elements. Phys Rev Lett 49:767–770
84. Kenkre VM (1982) The master equation approach: coherence, energy transfer, annihilation, and relaxation. In: Kenkre VM, Reineker P, Exciton Dynamics in Molecular Crystals and Aggregates. Springer, Berlin, pp 1–109
85. Kenkre VM, Montroll EW, Shlesinger MF (1973) Generalized master equations for continuous-time random walks. J Stat Phys 9:45–50
86. Kesten H (1986) The incipient infinite cluster in two-dimensional percolation. Probab Theory Relat Fields 73:369–394
87. Kim IC, Torquato S (1992) Effective conductivity of suspensions of overlapping spheres. J Appl Phys 71:2727–2735
88. Kirkpatrick S (1971) Classical transport in disordered media: scaling and effective-medium theories. Phys Rev Lett 27:1722–1725
89. Kirkpatrick S (1973) Percolation and conduction. Rev Mod Phys 45:574–588
90. Kirkpatrick S (1978) The geometry of the percolation threshold. In: Garland JC, Tanner DB (eds) Electrical Transport and Optical Properties of Inhomogeneous Media, AIP Conference Proceedings, No 40. American Institute of Physics, New York, pp 99–116
91. Knudsen HA, Fazekas S (2006) Robust algorithm for random resistor networks using hierarchical domain structure. J Comput Phys 211:700–718
92. Kogut PM, Straley JP (1979) Distribution-induced non-universality of the percolation conductivity exponents. J Phys C 12:2151–2159
93. Kozlov SM (1989) Geometric aspects of averaging. Russ Math Surv 44(2):91–144
94. Landauer R (1978) Electrical conductivity in inhomogeneous media. In: Garland JC, Tanner DB (eds) Electrical Transport and Optical Properties of Inhomogeneous Media, AIP Conference Proceedings, No 40. American Institute of Physics, New York, pp 2–43
95. Last BJ, Thouless DJ (1971) Percolation theory and electrical conductivity. Phys Rev Lett 27:1719–1721
96. Lawler GF, Schramm O, Werner W (2002) One-arm exponent for critical 2D percolation. Electronic Journal of Probability, vol 7, paper 2 (http://.www.math.washington.edu/~ejpecp/EjpVol7/paper2.html)
97. Lobb CJ, Frank DJ (1979) Large-cell renormalization group calculation of the percolation conductivity critical exponent. J Phys C 12:L827–L830
98. Lobb CJ, Frank DJ (1984) Percolative conduction and the Alexander–Orbach conjecture in two dimensions. Phys Rev B 30:4090–4092
99. Majid I, Ben-Avraham D, Havlin S, Stanley HE (1984) Exact-enumeration approach to random walks on percolation clusters in two dimensions. Phys Rev B 30:1626–1628
100. Mandelbrot BB (1982) The Fractal Geometry of Nature. W.H. Freeman, San Francisco
101. Marchant J, Gabillard B (1975) Sur le calcul d'un réseau résistif aléatoire. Comptes Rendus Acad Sci (Paris) B 281:261–264
102. Markov KZ (2000) Elementary micromechanics of heterogeneous media. In: Markov K, Preziosi L (eds) Heterogeneous Media: Micromechanics, Modeling, Methods and Simulations. Birkhäuser, Boston, pp 1–62
103. Mathieu P, Remy E (2004) Isoperimetry and heat kernel decay on percolation clusters. Ann Probab 32:100–128
104. Meester R, Roy R (1996) Continuum Percolation. Cambridge University Press, Cambridge
105. Meir Y, Blumenfeld R, Aharony A, Harris AB (1986) Series analysis of randomly diluted nonlinear resistor networks. Phys Rev B 34:3424–3428
106. Men'shikov MV (1986) Coincidence of critical points in percolation problems. Sov Math Dokl 33:856–859
107. Metzler R, Klafter J (2000) The random walker's guide to anomalous diffusion: a fractional dynamics approach. Phys Rep 339:1–77
108. Metzler R, Klafter J (2004) The restaurant at the end of the random walk: recent developments in the description of anomalous transport by fractional dynamics. J Phys A 37:R161-R208
109. Milton GW (2002) The Theory of Composites. Cambridge University Press, Cambridge
110. Mitescu CD, Roussenq J (1976) Une fourmi dans un labyrinthe: diffusion dans un système de percolation Comptes Rendus de l'Académie des Sciences (Paris) A 283:999–1001
111. Mitescu CD, Roussenq J (1983) Diffusion on percolation clusters. In: Deutscher G, Zallen R, Adler J (eds) Percolation Processes and Structures. Annals of the Israel Physical Society, vol 5. Adam Hilger, Bristol, pp 81–100
112. Mitescu CD, Allain M, Guyon E, Clerc JP (1982) Electrical conductivity of finite-size percolation networks. J Phys A 15:2523–2531
113. Mitescu CD, Ottavi H, Roussenq J (1978) Diffusion on percolation lattices: the labyrinthine ant. In: Garland JC, Tanner DB (eds) Electrical Transport and Optical Properties of Inhomogeneous Media, AIP Conference Proceedings, No 40. American Institute of Physics, New York, pp 377–381
114. Mitescu CD, Ottavi H, Roussenq J (1978). In: AIP Conference Proceedings, No 40. American Institute of Physics, New York, pp 377–381
115. Montroll EW, Weiss GH (1965) Random walks on lattices. II J Math Phys 6:167–181
116. Nakayama T, Yakubo K, Orbach RL (1994) Dynamical properties of fractal networks: scaling, numerical simulations, and physical realizations. Rev Mod Phys 66:381–443
117. Nash-Williams CStJA (1959) Random walks and electric currents in networks. Proc Camb Philos Soc 18:931–958
118. Normand J-M, Herrmann HJ, Hajjar M (1988) Precise calculation of the dynamical exponent of two-dimensional percolation. J Stat Phys 52:441–446
119. Normand J-M, Herrmann HJ (1990) Precise numerical determination of the superconducting exponent of percolation in three dimensions. Intern J Mod Phys C 1:207–214
120. Normand J-M, Herrmann HJ (1995) Precise determination of the conductivity exponent of 3D percolation using "Percola". Intern J Mod Phys C 6:207–214
121. Odagaki T, Lax M (1980) ac hopping conductivity of a one-dimensional bond percolation model. Phys Rev Lett 45:847–850
122. Oppenheim I, Shuler KE, Weiss GH (1977) Stochastic Processes in Chemical Physics: The Master Equation. MIT Press, Cambridge, Massachusetts

123. Palevski A, Deutscher G (1984) Conductivity measurements on a percolation fractal. J Phys A 17:L895–L898
124. Pandey RB, Stauffer D (1983) Fractal dimensionality and number of sites visited of the ant in the labyrinth. J Phys A 16:L511–L513
125. Pandey RB, Stauffer D, Margolina A, Zabolitzky JG (1984) Diffusion on random systems above, below and at their percolation threshold in two and three dimensions. J Stat Phys 34:427–450
126. Pearson K (1905) The problem of the random walk. Nature 72:294
127. Pemantle R, Peres Y (1996) On which graphs are all random walks in random environments transient? In: Aldous D, Pemantle R (eds) Random Discrete Structures, IMA Volumes in Mathematics and Its Applications, No. 76. Springer, New York, pp 207–211
128. Pólya G (1919) Quelques problèmes de probabilité se rapportant à la 'promenade au hasard'. Enseign Math 20:444–445
129. Pólya G (1921) über eine Aufgabe der Wahrscheinlichkeitsrechnung betreffend die Irrfahrt im Straßennetz. Mathematische Annalen 83:149–160
130. Puech L, Rammal R (1983) Fractal geometry and anomalous diffusion in the backbone of percolation clusters. J Phys C 16:L1179–L1202
131. Rammal R, Angles d'Auriac JC, Benoit A (1984) Universality of the spectral dimension of percolation clusters. Phys Rev B 30:4087–4089
132. Rammal R, Lemieux MA, Tremblay AMS (1985) Comment on 'ε-expansion for the conductivity of a random resistor network'. Phys Rev Lett 54:1087
133. Reš, I (2001) Corrections to scaling for percolative conduction: anomalous behavior at small *L*. Phys Rev B 64:224304
134. Révész P (2005) Random Walk in Random and Non-random Environments. World Scientific, Singapore
135. Rintoul MD, Torquato S (1997) Precise determination of the critical threshold and exponents in a three-dimensional continuum percolation model. J Phys A 30:L585–L592
136. Roman HE (1990) Diffusion in three-dimensional random systems at their percolation threshold. J Stat Phys 58:375–382
137. Sahimi M (1984) Finite-size scaling calculation of conductivity of three-dimensional conductor–superconductor networks at the percolation threshold. J Phys C 17:L355–L358
138. Sahimi M (1994) Applications of Percolation Theory. Taylor and Francis, London
139. Sahimi M (1995) Flow and Transport in Porous Media and Fractured Rock. VCH Verlagsgesellschaft, Weinheim
140. Sahimi M (2003) Heterogeneous Materials, vol 1: Linear Transport and Optical Properties. Springer, New York
141. Sahimi M (2003) Heterogeneous Materials, vol 2: Nonlinear and Breakdown Properties and Atomistic Modeling. Springer, New York
142. Sahimi M, Hughes BD, Scriven LE, Davis HT (1983) Stochastic transport in disordered systems. J Chem Phys 78:6849–6864
143. Sahimi M, Hughes BD, Scriven LE, Davis HT (1983) Critical exponent of percolation conductivity by finite-size scaling. J Phys C 16:L521–L527
144. Sahimi M, Hughes BD, Scriven LE, Davis HT (1983) Real-space renormalization and effective-medium approximation to the percolation conduction problem. Phys Rev B 28:307–311
145. Seifert E, Suessenbach M (1984) Tests of universality for percolative diffusion. J Phys A 17:L703–L708
146. Sinai YG (1982) The limiting behavior of a one-dimensional random walk in a random environment. Theory Probab Appl 27:256–268
147. Skal AS, Shklovskii BI (1975) Topology of an infinite cluster in the theory of percolation and its relationship to the theory of hopping conduction. Sov Phys Semicond 8:1029–1032
148. Smirnov S, Werner W (2001) Critical exponents for two-dimensional percolation. Math Res Lett 8:729–744
149. Solomon F (1975) Random walks in a random environment. Ann Probab 3:1–31
150. Spitzer F (1976) Principles of Random Walk, 2nd edition. Springer, Berlin
151. Stanley HE (1977) Cluster shapes at the percolation threshold: an effective cluster dimensionality and its connection with critical point exponents. J Phys A 10:L211–L220
152. Stanley HE, Coniglio A (1983) Fractal structure of the incipient infinite cluster in percolation. In: Deutscher G, Zallen R, Adler J (eds) Percolation Processes and Structures, Annals of the Israel Physcal Society, vol 5. Adam Hilger, Bristol, pp 101–120
153. Stauffer D (1979) Scaling theory of percolation clusters. Phys Rep 54:1–74
154. Stauffer D (1985) Introduction to Percolation Theory. Taylor and Francis, London
155. Stauffer D, Aharony A (1994) Introduction to Percolation Theory, corrected 2nd edition. Taylor and Francis, London
156. Stinchcombe RB (1973) The branching model for percolation theory and electrical conductivity. J Phys C 6:L1–L5
157. Stinchcombe RB (1974) Conductivity and spin-wave stiffness in disordered systems: an exactly soluble model. J Phys C 7:197–203
158. Straley JP (1976) Critical phenomena in resistor networks. J Phys C 9:783–795
159. Straley JP (1977) Random resistor tree in an applied field. J Phys C 10:3009–3013
160. Straley JP (1977) Critical exponents for the conductivity of random resistor lattices. Phys Rev B 15:5733–5737
161. Straley JP, Kenkel SW (1984) Percolation theory for nonlinear conductors. Phys Rev B 29:6299–6305
162. Taitelbaum H, Havlin S (1988) Superconductivity exponent for the Sierpinski gasket. J Phys A 21:2265–2271
163. Telcs A (2006) The Art of Random Walks. Lecture Notes in Mathematics, vol 1885. Springer, Berlin
164. Temkin DE (1972) One-dimensional random walks in a two-component chain. Sov Math Dokl 13:1172–1176
165. Thorpe MF (1982) Bethe lattices. In: Thorpe MF (ed) Excitations in Disordered Systems. Plenum, New York, pp 85–107
166. Torquato S (2002) Random Heterogeneous Materials: Microstructure and Macroscopic Properties. Springer, New York
167. Torquato S, Kim IC, Kule D (1999) Effective conductivity, dielectric constant, and diffusion coefficient of digitized composite media via first-passage-time equations. J Appl Phys 85:1560–1571
168. Wiener O (1912) Die Theorie des Mischkörpers für das Feld des stationären Strömung. Abh Mathematisch-Phys Kl K Sächs Ges Wiss 32:509–604
169. Woess W (2000) Random Walks on Infinite Graphs and Groups. Cambridge University Press, Cambridge
170. Zabolitzky JG (1984) Monte Carlo evidence against the Alexander–Orbach conjecture for percolation conductivity. Phys Rev B 30:4076–4079

171. Zeitouni O (2002) Random walks in random environments. In: Daquien LI (ed) Proceedings of the International Congress of Mathematicians. vol 3. Higher Education Press, Beijing, pp 117–127
172. Zeitouni O (2004) Random walks in random environment. In: Tavaré S, Zeitouni O (eds) Lectures on Probability and Statistics (Ecole d'Eté de Probabilités de Saint-Flour XXXI), Lecture Notes in Mathematics, vol 1837. Springer, Berlin, pp 1–188
173. Zeitouni O (2006) Random walks in random environments. J Phys A: Math Gen 39:R433–R464
174. Ziman J (1968) The localization of electrons in ordered and disordered systems. I Percolation of classical particles. J Phys C 1:1532–1538

Consciousness and Complexity

Anil K. Seth[1], Gerald M. Edelman[2]
[1] Department of Informatics, University of Sussex, Brighton, UK
[2] The Neurosciences Institute, San Diego, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Consciousness and Complexity
Neural Complexity
Information Integration
Causal Density
Empirical Evidence
Related Theoretical Proposals
Outlook
Bibliography

Glossary

Thalamocortical system The network of highly interconnected cortical areas and thalamic nuclei that comprises a large part of the mammalian brain. The cortex is the wrinkled surface of the brain, the thalamus is a small walnut-sized structure at its center. An intact thalamocortical system is essential for normal conscious experience.

Theory of neuronal group selection (TNGS) A large-scale selectionist theory of brain development and function with roots in evolutionary theory and immunology. According to this theory, brain dynamics shape and are shaped by selection among highly variant neuronal populations guided by value or salience.

Neural correlate of consciousness Patterns of activity in brain regions or groups of neurons that have privileged status in the generation of conscious experience. *Explanatory* correlates are neural correlates that in addition account for key properties of consciousness.

Dynamic core A distributed and continually shifting coalesence of patterns of activity among neuronal groups within the thalamocortical system. According to the TNGS, neural dynamics within the core are of high neural complexity by virtue of which they give rise to conscious discriminations.

Neural complexity A measure of simultaneous functional segregation and functional integration based on information theory. A system will have high neural complexity if each of its components can take on many different states and if these states make a difference to the rest of the system.

Small-world networks Networks in which most nodes are not neighbors of one another, but most nodes can be reached from every other by a small number of hops or steps. Small-world networks combine high clustering with short path lengths. They can be readily identified in neuroanatomical data, and they are well suited to generating dynamics of high neural complexity.

Metastability Dynamics that are characterized by segregating and integrating influences in the temporal domain; metastable systems are neither totally stable nor totally unstable.

Definition of the Subject

How do conscious experiences, subjectivity, and apparent free will arise from their biological substrates? In the mid 1600s Descartes formulated this question in a form that has persisted ever since [1]. According to Cartesian dualism consciousness exists in a non-physical mode, raising the difficult question of its relation to physical interactions in the brain, body and environment. Even in the late twentieth century, consciousness was considered by many to be outside the reach of natural science [2], to require strange new physics [3], or even to be beyond human analysis altogether [4]. Over the last decade however, there has been heightened interest in attacking the problem of consciousness through scientific investigation [5,6,7,8,9]. Succeeding in this inquiry stands as a key challenge for twenty-first century science.

Conventional approaches to the neurobiology of consciousness have emphasized the search for so-called 'neural correlates': Activity within brain regions or groups of neurons that has privileged status in the generation of conscious experience [10]. An important outcome of this line of research has been that consciousness is closely tied to neural activity in the thalamocortical system, a network of

cortical areas and subcortical nuclei that forms a large part of the vertebrate brain [11,12]. Yet correlations by themselves cannot supply explanations, they can only constrain them. A promising avenue toward explanation is to focus on key properties of conscious experience and to identify neural processes that can account for these properties; we can call these processes *explanatory correlates*. This article clarifies some of the issues surrounding this approach and describes ways of characterizing quantitatively the *complexity* of neural dynamics as a candidate explanatory correlate.

Complexity is a central concept within many branches of systems science and more generally across physics, statistics, and biology; many quantitative measures have been proposed and new candidates appear frequently [13,14,15]. The complexity measures described in this article are distinguished by focusing on the extent to which a system's dynamics are *differentiated* while at the same time *integrated*. This conception of complexity accounts for a fundamental feature of consciousness, namely that every conscious experience is composed of many different distinguishable components (differentiation) and that every conscious experience is a unified whole (integration). According to the theoretical perspective described here, the combination of these features endows consciousness with a discriminatory capability unmatched by any other natural or artificial mechanism.

While the present focus is on consciousness and its underlying mechanisms, it is likely that the measures of complexity we describe will find application not only in neuroscience but also in a wide variety of natural and artificial systems.

Introduction

Consciousness

Consciousness is that which is lost when we fall into a dreamless sleep and returns when we wake up again. As William James emphasized, consciousness is a *process* and not a 'thing' [5]. Conscious experiences have content such as colors, shapes, smells, thoughts, emotions, inner speech, and the like, and are commonly accompanied by a sense of self and a subjective perspective on the world (the 'I'). The phenomenal aspects of conscious content (the 'redness' of red, the 'warmth' of heat, etc.) are in philosophical terminology called *qualia* [16].

It is important to distinguish between conscious *level*, which is a position on a scale from brain-death and coma at one end to vivid wakefulness at the other, and conscious *content*, which refers to composition of a conscious scene at a given (non-zero) conscious level. Obviously,

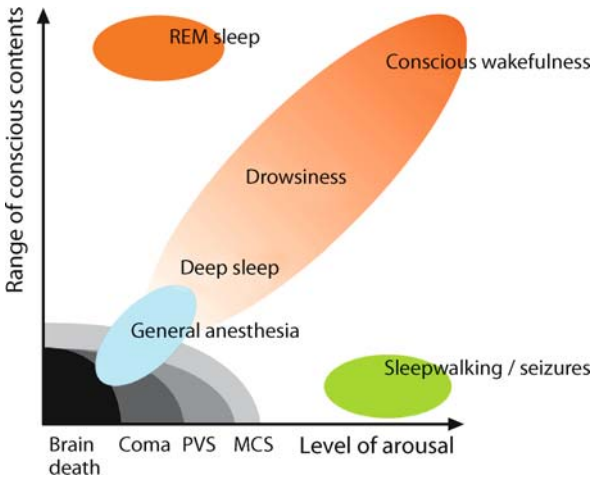

Consciousness and Complexity, Figure 1
Conscious level is correlated with the range of possible conscious contents. PVS = persistent vegetative state, MCS = minimally conscious state. Adapted from [21]

conscious level and conscious content are related inasmuch as the range of possible conscious contents increases with conscious level (see Fig. 1). It is also possible to differentiate *primary* (sensory) consciousness from *higher-order* (meta) consciousness [6,17]. Primary consciousness refers to the presence of perceptual conscious content (colors, shapes, odors, etc.). Higher-order consciousness (HOC) refers to the fact that we are usually conscious of being conscious; that is, human conscious contents can refer to ongoing primary conscious experiences. HOC is usually associated with language and an explicit sense of selfhood [18] and good arguments can be made that primary consciousness can exist in principle in the absence of HOC, and that in many animals it probably does [19,20].

Consciousness as Discrimination

There are many aspects of consciousness that require explanation (see Table 1). However, one especially salient aspect that has been too often overlooked is that every conscious scene is both *integrated* and *differentiated* [22]. That is, every conscious scene is experienced 'all of a piece', as unified, yet every conscious scene is also composed of many different parts and is therefore one among a vast repertoire of possible experiences: When you have a particular experience, you are distinguishing it from an enormous number of alternative possibilities. On this view, conscious scenes reflect informative discriminations in a very high dimensional space where the dimensions reflect all the various modalities that comprise a conscious

Consciousness and Complexity, Table 1

Thirteen features of consciousness that require theoretical explanation. Items 1–6 are to some degree open to quantitative measurement whereas items 7–13 are more readily understood through logical and qualitative analysis. This list is drawn from [23] and a related list appears in [17]

1	Consciousness is accompanied by irregular, low-amplitude, fast (12–70 Hz) electrical brain activity.
2	Consciousness is associated with activity within the thalamocortical complex, modulated by activity in subcortical areas.
3	Consciousness involves distributed cortical activity related to conscious contents.
4	Conscious scenes are unitary.
5	Conscious scenes occur serially – only one conscious scene is experienced at a time.
6	Conscious scenes are metastable and reflect rapidly adaptive discriminations in perception and memory.
7	Conscious scenes comprise a wide multimodal range of contents and involve multimodal sensory binding.
8	Conscious scenes have a focus/fringe structure; focal conscious contents are modulated by attention.
9	Consciousness is subjective and private, and is often attributed to an experiencing 'self'.
10	Conscious experience is reportable by humans, verbally and non-verbally.
11	Consciousness accompanies various forms of learning. Even implicit learning initially requires consciousness of stimuli from which regularities are unconsciously extracted.
12	Conscious scenes have an allocentric character. They show intentionality, yet are shaped by egocentric frameworks.
13	Consciousness is a necessary aspect of decision making and adaptive planning.

experience: sounds, smells, body signals, thoughts, emotions, and so forth (Fig. 2).

Because the above point is fundamental, it is useful to work through a simple example (adapted from [22,24]). Consider a blank rectangle that is alternately light and dark (Fig. 3a,b). Imagine that this rectangle is all there is, that you are seated in front of it, and that you have been instructed to say "light" and "dark" as appropriate. A simple light-sensitive diode is also in front of the screen and beeps whenever the screen is light. Both you and the diode can perform the task easily, therefore both you and the diode can discriminate these two states: lightness and darkness. But each time the diode beeps, it is entering into one of a total of two possible states. It is minimally differentiated. However, when you say "light" or "dark" you are reporting one out of an enormous number of possible experiences. This point is emphasized by considering a detailed image such as a photograph (Fig. 3c). A conscious person will readily see this image as distinct both from the blank rectangle and from a scrambled version of the same image (Fig. 3d). The diode, however, would classify both images and the rectangle as "light" (depending on its threshold), because it is insufficiently differentiated to capture the differences between the three.

Consider now an idealized digital camera. The electronics inside such a camera will enter a different state for the scrambled image than for the non-scrambled image; indeed, there will be a distinct state for any particular image. A digital camera is capable of much greater differentiation than the diode, but it is still not capable of discrimination because it is minimally *integrated*. In idealized form it is a collection of many independent light-sensitive diodes that must, to a good approximation, remain functionally independent from each other. From the perspective of this camera the image and the scrambled image are equivalent. *We* (as conscious organisms) can tell the difference between the two is because we integrate the many different parts of the image to form a coherent whole. We perceive each part of the image in relation to all the other parts, and we perceive each image in relation to all other possible images and possible conscious experiences that we may have. Successful discrimination therefore requires *both* integration *and* differentiation, and it can be hypothesized that it is this balance that yields the unity and diversity central to conscious experience.

Experimental evidence as well as intuition testifies to the fundamental nature of integration and differentiation in consciousness. A striking example is provided by so-called 'split brain' patients whose cortical hemispheres have been surgically separated. When presented with two independent visuospatial memory tasks, one to each hemisphere, they perform both very well [25]. In contrast, normal subjects cannot avoid integrating the independent signals into a single conscious scene which yields a much harder problem, and performance is correspondingly worse. In general, normal subjects are unable to perform multiple tasks simultaneously if they both require conscious input and they cannot make more than one conscious decision within the so-called 'psychological refractory period', a short interval of a few hundred milliseconds [26].

A loss of differentiation can be associated with the impoverishment of conscious contents following brain trauma. In 'minimally conscious' or 'persistent vegetative' states the dynamical repertoire of the thalamocortical system is reduced to the extent that adaptive behavioral responses are excluded [21]. In less dramatic cases focal cortical lesions can delete specific conscious contents; for example, damage to cortical region V4 can remove color

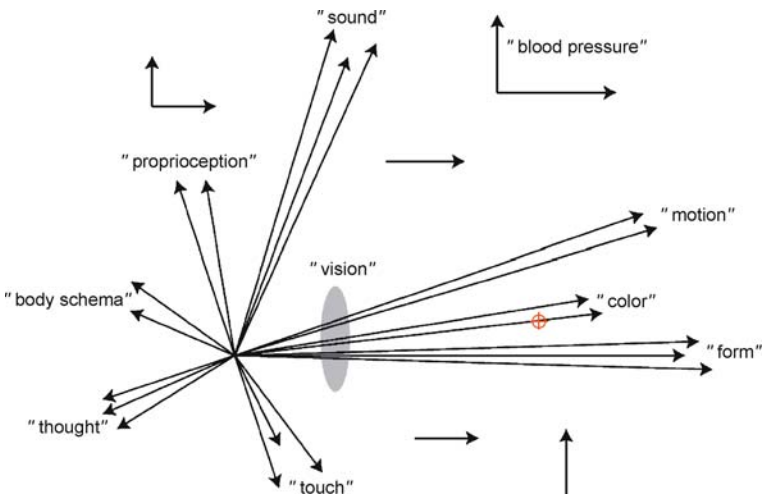

Consciousness and Complexity, Figure 2
The figure shows an *N*-dimensional neural space corresponding to the dynamic core (see Sect. "Consciousness and Complexity"). *N* is the number of neuronal groups that, at any time, are part of the core, where *N* is normally very large (much larger than is plotted). The appropriate neural reference space for the conscious experience of 'pure red' would correspond to a discriminable point in the space (marked by the *red cross*). Focal cortical damage can delete specific dimensions from this space

dimensions from the space of possible experiences (cerebral achromatopsia [27]; c.f., Fig. 2). Reportable conscious experience is also eliminated during generalized epileptic seizures and slow-wave sleep. Neural activity in these states is again poorly differentiated, showing hypersynchrony (epilepsy) or a characteristic synchronous 'burst pause' pattern (sleep) [22].

Consciousness and Complexity

The Dynamic Core Hypothesis

The notion that consciousness arises from neural dynamics that are simultaneously differentiated and integrated is expressed by the *dynamic core hypothesis* (DCH). This hypothesis has two parts [22,28]:

- A group of neurons can contribute directly to conscious experience only if it is part of a distributed functional cluster (the dynamic core) that, through reentrant interactions in the thalamocortical system, achieves high integration in hundreds of milliseconds.
- To sustain conscious experience, it is essential that this functional cluster be highly differentiated, as indicated by high values of complexity.

The concept of a *functional cluster* refers to a subset of a neural system with dynamics that displays high statistical dependence internally and comparatively low statistical dependence with elements outside the subset: A functional cluster 'speaks mainly to itself' [29]. Conceiving of the dynamic core as a functional cluster implies that the boundaries of the neural substrates of consciousness are continually shifting, with neuronal groups exiting and entering the core according to the flow of conscious contents and the corresponding discriminations being made. *Reentry* refers to the recursive exchange of signals among neural areas across massively parallel reciprocal connections and which in the context of the DCH serve to bind the core together. It is important to distinguish reentry from 'feedback' which refers to the recycling of an error signal from an output to an input [30,31]. The interpretation of *complexity* in the context of the DCH is the subject of Sect. "Neural Complexity"; for now we remark that it provides a quantitative measure of neural dynamics that is maximized by simultaneous high differentiation and high integration.

The Theory of Neuronal Group Selection

The DCH emerged from the theoretical framework provided by the 'theory of neural group selection' (TNGS), otherwise known as 'neural Darwinism' [18,32,33]. This section summarizes some of this essential background.

The TNGS is a biological perspective on brain processes with roots in evolutionary theory and immunology. It suggests that brain development and dynamics are *selectionist* in nature, and not instructionist, in contrast to computers which carry out explicit symbolic instructions. Four aspects of selectionist processes are emphasized: di-

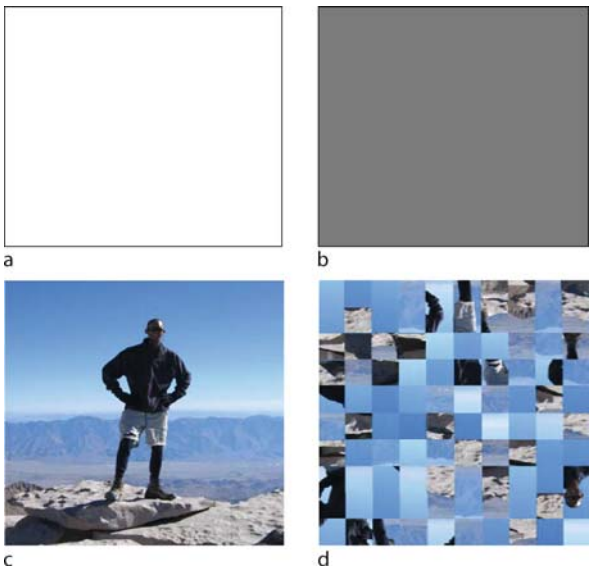

Consciousness and Complexity, Figure 3
a A light-colored rectangle. b A dark-colored rectangle. c A detailed image (the summit of Mount Whitney, California). d A scrambled version of the same image. A simple light-sensitive diode would be able to discriminate a from b, but not among a, c, and d, since all these images would appear as 'light'. An idealized digital camera would enter a different state for each image a, b, c, and d, but would not discriminate between c and d because the camera does not integrate the various parts of each image to form a coherent whole. We can discriminate among all images because (i) our brain is capable of sufficient differentiation to enter a distinct state for each image, and (ii) our brain is capable of integrating the various parts of each image to form a coherent whole

versity, amplification/reproduction, selection, and degeneracy. *Diversity* in the brain is reflected in highly variant populations of neuronal groups where each group consists of hundreds to thousands of neurons of various types. This variation arises as a result of developmental and epigenetic processes such as cell division, migration, and axonal growth; subsequent strengthening and weakening of connections among cells (synapses) via experience and behavior generates further diversity. *Amplification* and *selection* in the brain are constrained by *value*, which reflects the salience of an event and which can be positive or negative as determined by evolution and learning. Value is mediated by diffuse ascending neural pathways originating, for example, in dopaminergic, catecholaminergic, and cholinergic brainstem nuclei [34]. As a result of value-dependent synaptic plasticity, connections among neuronal groups that support adaptive outcomes are strengthened, and those that do not are weakened. Finally, *degeneracy* emphasizes that in adaptive neural systems many structurally different combinations can perform the same function and yield the same output. Degeneracy is a key feature of many biological systems that endows them with adaptive flexibility [35,36]. It is conspicuously absent in artificial systems that are correspondingly fragile (some artificial systems make use of 'redundancy' which differs from degeneracy in that specific functional units are explicitly duplicated; redundancy provides the robustness but not the flexibility of degeneracy).

According to the TNGS, primary consciousness arises when brain areas involved in ongoing perception are linked via reentry to brain areas responsible for a value-based memory of previous perceptual categorizations. On this view, primary consciousness manifests as a 'remembered present' (akin to William James' 'specious present') by which an animal is able to exploit adaptive links between immediate or imagined circumstances and that animal's previous history of value-driven behavior (Fig. 4).

The TNGS and the dynamic core hypothesis are closely related [17,18]. They share the general claim that the neural mechanisms underlying consciousness arose in evolution for their ability to support multimodal discriminations in a high-dimensional space. In addition, the reentrant interactions linking immediate perception to value-category memory are precisely those that are suggested to bind together the dynamic core. Finally, the vast diversity of neural groups is central both to the original TNGS in providing a substrate for selection and to the DCH, in providing an essential component of neural complexity.

Consciousness and the Dynamic Core

We can now summarize the DCH and its origin in the TNGS. Consciousness is entailed by extensive reentrant interactions among neuronal populations in the thalamocortical system, the so-called dynamic core. These interactions, which support high-dimensional discriminations among states of the dynamic core, confer selective advantages on the organisms possessing them by linking current perceptual categorizations to value-dependent memory. The high dimensionality of these discriminations is proposed to be a direct consequence of the rich complexity of the participating neural repertoires. Just as conscious scenes are both differentiated and integrated at the phenomenal level to yield high-dimensional discriminations, so too are the reentrant dynamics of their underlying neural mechanisms differentiated and integrated. Critically according to the TNGS, conscious qualia *are* the high-dimensional discriminations entailed by this balance of differentiation and integration as reflected in high complexity.

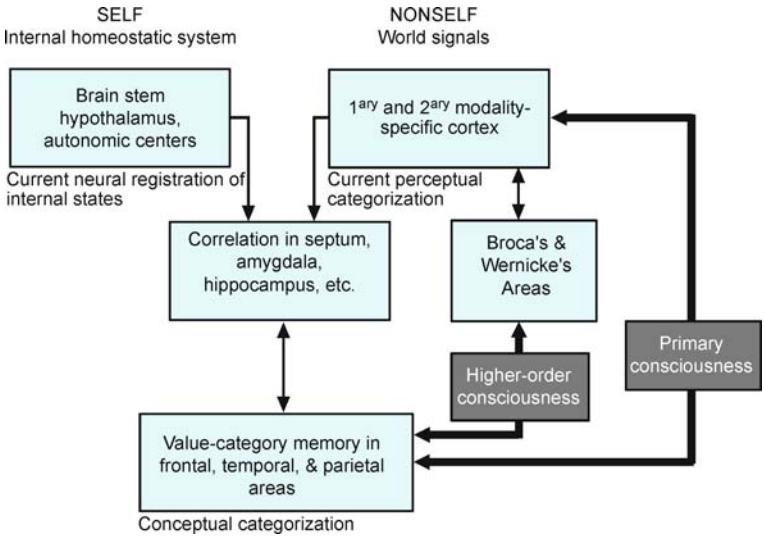

Consciousness and Complexity, Figure 4

Primary consciousness and HOC in the TNGS. Signals related to value and signals from the world are correlated and produce value-category memories. These memories are linked by reentry to current perceptual categorization, resulting in primary consciousness. Higher-order consciousness depends on further reentry between value-category memory and current categorization via areas involved in language production and comprehension. Reprinted from [17]

Any theory of consciousness must confront the question of whether conscious experiences have causal effects in the physical world [37]. Responding positively reflects common sense but it seems contrary to science to suggest non-physical causes for physical events. Responding negatively respects the causal closure of the physical world but appears to suggest that conscious experiences are 'epiphenomenal' and could in principle be done without (an implication that may be particularly troubling for experiences of 'free will' [38]). The TNGS addresses this quandary via the notion of *entailment*. According to the TNGS, dynamic core processes entail particular conscious experiences in the same way that the molecular structure of hemoglobin entails its particular spectroscopic properties: it simply could not be otherwise [18]. Therefore, although consciousness does not cause physical events, there exist particular physical causal chains (the neural mechanisms underlying consciousness) that by necessity entail corresponding conscious experiences: The conscious experience cannot be 'done without'.

Measuring Consciousness and Complexity

Having covered basic elements of the DCH and its origin in the TNGS, we turn now to the issue of measuring complexity in neural dynamics. To be useful in this context, candidate measures should satisfy several constraints. We have already mentioned that a suitable measure should reflect the fact that consciousness is a dynamic process [5], not a thing or a capacity. This point is particularly important in light of the observation that conscious scenes arise ultimately from transactions between organisms and environments, and these transactions are fundamentally processes [39]. (This characterization does not, however, exclude 'off-line' conscious scenes, for example those experienced during dreaming, reverie, abstract thought, planning, or imagery). A suitable measure should also take account of causal interactions within a neural system, and between a neural system and its surroundings – i. e., bodies and environments. Finally, to be of practical use, a suitable measure should also be computable for systems composed of large numbers of neuronal elements.

Obviously, the quantitative characterization of complexity can constitute only one aspect of a scientific theory of consciousness. This is true at both the neural level and at the level of phenomenal experience. At the neural level, no single measure could adequately describe the complexity of the underlying brain system (this would be akin, for example, to claiming that the complex state of the economy could be described by the gross domestic product alone). At the phenomenal level, conscious scenes have many diverse features [18,19], several of which do not appear to be readily quantifiable (see Table 1). These include subjectivity, the attribution of conscious experience to a self, and intentionality, which reflects the observation that consciousness is largely about events and objects. A critical is-

sue nevertheless remains: how can measurable aspects of the neural underpinnings of consciousness be characterized [32,132]?

Neural Complexity

A fundamental intuition about complexity is that a complex system is neither fully ordered (e. g., a crystal) nor fully disordered (e. g., an ideal gas). This intuition is compatible with the central theme of the DCH, namely that the neural dynamics within the dynamic core should be both integrated and differentiated. The following definition of *neural complexity* (C_N), first proposed in 1994 [40], satisfies these intuitions and provides a practical means for assessing the complexity of neural and other systems.

Mathematical Definition

Consider a neural system X composed of N elements (these may be neurons, neuronal groups, brain regions, etc.). A useful description of the dynamical connectivity of X is given by the joint probability distribution of the activities of its elements. Assuming that this function is Gaussian, this is equivalent to the covariance matrix of the system's dynamics COV(X). Importantly, COV(X) captures the total effect of all (structural) connections within a system upon deviation from statistical independence of the activities of a pair of elements, and not just the effect of any direct anatomical connection linking them [41]. Given COV(X) and assuming that the dynamics of X are covariance stationary (i. e., having unchanging mean and variance over time) the entropy of the system $H(\mathrm{X})$ is given by:

$$H(\mathrm{X}) = \tfrac{1}{2} \ln \left((2\pi \mathrm{e})^N \, |\mathrm{COV}(\mathrm{X})| \right)$$

where $|.|$ denotes the matrix determinant [42]. $H(\mathrm{X})$ measures the overall degree of statistical independence exhibited by the system; i. e., its degree of differentiation. Knowing the entropy of a system allows calculation of the mutual information (MI) between two systems, or between two subsets of a single system. The MI between systems (or subsets) A and B measures the uncertainty about A that is accounted for by the state of B and is defined as $\mathrm{MI}(\mathrm{A};\mathrm{B}) = H(\mathrm{A}) + H(\mathrm{B}) - H(\mathrm{AB})$ [43].

The integration of X, $I(\mathrm{X})$, measures the system's overall deviation from statistical independence. All elements in a highly integrated system are tightly coupled in their activity. With x_i denoting the ith element of X, $I(\mathrm{X})$ can be calculated as:

$$I(\mathrm{X}) = \sum_{i=1}^{N} H(x_i) - H(\mathrm{X}) \, .$$

$I(\mathrm{X})$ is equivalent to the measure 'multi-information' which was introduced several decades ago [44]. Having ex-

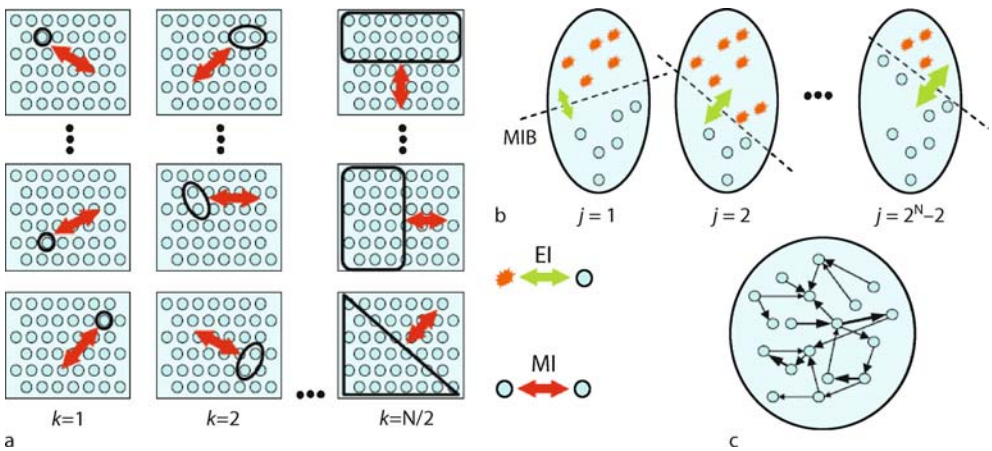

Consciousness and Complexity, Figure 5

Measuring integration and differentiation in neural dynamics. a Neural complexity C_N is calculated as the ensemble average mutual information (MI) between subsets of a given size and their complement, summed over all subset sizes (k) (adapted from Fig. 2 in [46]). *Small circles* represent neuronal elements and *red arrows* indicate MI between subsets and the remainder of the system. b Information integration Φ is calculated as the effective information (EI) across the 'minimum information bipartition' (MIB). To calculate EI for a given bipartition (j), one subset is injected with maximally entropic activity (*orange stars*) and MI across the partition is measured. c Causal density c_d is calculated as the fraction of interactions that are causally significant according to a multivariate Granger casuality analysis. A weighted (and unbounded) version of causal density (c_{dw}) can be calculated as the summed magnitudes of all significant causal interactions (depicted schematically by *arrow width*). Reprinted with permission from [132]

pressions for MI, $H(\mathrm{X})$, and $I(\mathrm{X})$ allows $C_\mathrm{N}(\mathrm{X})$ to be expressed in two equivalent ways. First, $C_\mathrm{N}(\mathrm{X})$ can be calculated by summing the average MI between subsets of various sizes, for all possible bipartitions of the system:

$$C_\mathrm{N}(\mathrm{X}) = \sum_k \left\langle MI\left(\mathrm{X}_j^k; \mathrm{X} - \mathrm{X}_j^k\right)\right\rangle , \quad (1)$$

where X_j^k is the jth bipartition of size k, and $\langle . \rangle$ is the average across index j (Fig. 5a). $C_\mathrm{N}(\mathrm{X})$ can also be expressed in terms of integration:

$$C_\mathrm{N}(\mathrm{X}) = \sum_k \left(\frac{k}{n} I(\mathrm{X}) - \left\langle I\left(\mathrm{X}_j^k\right)\right\rangle\right) , \quad (2)$$

where $\langle I(\mathrm{X}_j^k)\rangle$ is the average integration of all subsets of size k. $C_\mathrm{N}(\mathrm{X})$ will be high if small subsets of the system show high statistical independence, but large subsets show low statistical independence. In other words, $C_\mathrm{N}(\mathrm{X})$ will be high if each of its subsets can take on many different states and if these states make a difference to the rest of the system.

Because the full $C_\mathrm{N}(\mathrm{X})$ can be computationally expensive to calculate for large systems, it is useful to have an approximation that considers only bipartitions consisting of a single element and the rest of the system. There are three mathematically equivalent ways of expressing this approximation, which is denoted $C(\mathrm{X})$:

$$\begin{aligned} C(\mathrm{X}) &= H(\mathrm{X}) - \sum_{k=1}^{N} H\left(x_i | \mathrm{X} - x_i\right) \\ &= \sum_i MI(x_i; \mathrm{X} - x_i) - I(\mathrm{X}) \\ &= (n-1)\, I(\mathrm{X}) - n \left\langle I(\mathrm{X} - x_i)\right\rangle , \end{aligned} \quad (3)$$

where $H(x_i|\mathrm{X} - x_i)$ denotes the conditional entropy of each element x_i given the entropy of the rest of the system $\mathrm{X} - \mathrm{X}_i$. These three expressions are equivalent for all X, whether they are linear or non-linear, and neither $C_\mathrm{N}(\mathrm{X})$ nor $C(\mathrm{X})$ can adopt negative values.

Recently, De Lucia et al. [45] have developed a different approximation to $C_\mathrm{N}(\mathrm{X})$ that is calculated directly from topological network properties (i. e., without needing covariance information). Their measure of 'topological $C_\mathrm{N}(\mathrm{X})$' is based on the eigenvalue spectrum of the connectivity matrix of a network. While topological $C_\mathrm{N}(\mathrm{X})$ offers substantial savings in computational expense it carries the assumption that the network is activated by independent Gaussian noise and therefore cannot be used to measure neural complexity in conditions in which a network is coupled to inputs and outputs (see Subsect. "Complexity and Behavior" below).

Connectivity and Complexity

There is a growing consensus that features of neuroanatomical organization impose important constraints on the functional dynamics underlying cognition [47,48]. Accordingly, several studies have addressed the relationship between structural connectivity and neural complexity [49,50,51,52,53].

One useful approach employs evolutionary search procedures (genetic algorithms [54]) to specify the connection structure of simple networks under various fitness (cost) functions. A population of networks $\mathrm{X}_1 \ldots \mathrm{X}_\mathrm{N}$ is initialized ('generation zero') with each member having random connectivity. Each network X_i is then evaluated according to a fitness function [for example, maximize $C(\mathrm{X})$] and those that score highly, as compared to the other networks in the population, are subjected to a small amount of random 'mutation' (i. e., small random changes in connectivity) and proceed to the next 'generation'. This procedure is repeated for many generations until the population contains networks that score near-optimally on the fitness function, or until the experimenter is satisfied that no further improvement is likely.

Sporns and colleagues applied a version of evolutionary search to find distinctive structural motifs associated with $H(\mathrm{X})$, $I(\mathrm{X})$, and $C(\mathrm{X})$ [49]. In this study, the initial population consisted of ten networks each with $N = 32$ nodes and $K = 256$ connections and with fixed identical positive weights w_{ij}. The fitness function was determined by the value of $H(\mathrm{X})$, $I(\mathrm{X})$, or $C(\mathrm{X})$ calculated from the covariance matrix of each network, assuming activation by covariance-stationary Gaussian noise. In each case they found that the resulting networks had distinctive structural features, as revealed both by simple visual inspection and by analysis using a variety of graph-theoretic measures. Networks optimized for $H(\mathrm{X})$ contained mostly reciprocal connections without any apparent local clustering. Networks optimized for $I(\mathrm{X})$ were highly clustered (i. e., neighboring nodes connect mainly to each other [55]) and had a long characteristic path length (i. e., a high mean separation between any two nodes in terms of number of intervening nodes). Finally, networks optimized for $C(\mathrm{X})$ had high clustering (and high reciprocal connectivity) coupled with a short characteristic path length. Strikingly, these networks were very similar to the so-called 'small world' class of network in which dense groups of nodes are connected by a relatively small number of reciprocal 'bridges' [55]. These networks also had a high proportion of 'cycles' (routes through the network that return to their starting point) and very low wiring lengths [49].

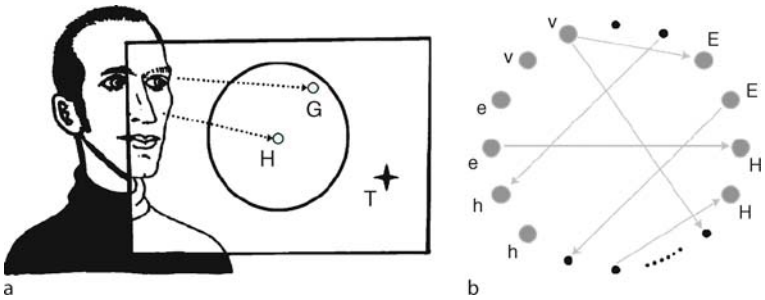

Consciousness and Complexity, Figure 6

Target fixation model. a The agent controls head-direction (H) and eye-direction (not shown) in order to move a gaze point (G) towards a target (T). b Neural network controller. The six input neurons are shown on the *left* and the four output neurons on the *right*. Each pair of inputs (v,e,h) responds to *x*, *y* displacements: 'v' neurons to displacements of G from T, 'h' neurons to displacements of H from an arbitrary origin ('straight ahead'), and 'e' neurons to displacements of H from the eye-direction. The four output neurons control head direction (H) and eye-direction relative to the head (H). For clarity only four of the 22 interneurons are shown. *Thin gray lines* show synaptic connections. Only a subset of the 256 connections are shown. Adapted from [59]

Sporns et al. extended the above findings by calculating $C(X)$ for networks reflecting the known cortical connectivity of both the macaque visual cortex and the entire cat cortex. In both cases covariance matrices were obtained by assuming linear dynamics, equal connection strengths, and activation by covariance-stationary Gaussian noise. They found that both networks gave rise to high $C(X)$ as compared to random networks with equivalent distributions of nodes and connections. Indeed, the networks seemed to be near-optimal for $C(X)$ because random rewiring of connections led in almost all cases to a reduction in $C(X)$ [49].

In a separate study using a non-linear neuronal network model including excitatory and inhibitory units, Sporns showed that regimes of high $C(X)$ coincided with 'mixed' connection patterns consisting of both local and long-range connections [56]. This result lines up with the previous study [49] in suggesting an association between small-world properties and complex dynamics. In addition, Sporns and Kötter found that networks optimized for the number of functional 'motifs' (small repeating patterns) had high $C(X)$ but those optimized for structural motifs did not [57] suggesting that high complexity reflects the presence of large functional repertoires. Finally, $C(X)$ seems to associate with fractal patterning, but not in a simple sense that fractal networks are optimal for complexity [53,58]. Rather, fractality seems to be one among several structural attributes that contribute to the emergence of small-world features and complex dynamics. Together, these results indicate that only certain classes of network are able to support dynamics that combine functional integration with functional segregation and that these networks resemble in several ways those found in neuroanatomical systems.

Complexity and Behavior

An important claim within the DCH is that complex dynamics provide adaptive advantages during behavior. To test this claim, Seth and Edelman examined the relationship between behavior and neural complexity in a simple agent-based computational model [59]. They evolved networks similar to those in [49] ($N = 32$, $K = 256$) by selecting for their ability to guide target fixation behavior in a simulation model requiring coordination of 'head' and 'eye' movements (Fig. 6). Networks were evolved in both 'simple' and 'complex' environments where environmental complexity was reflected by unpredictable target movement and by variation in parameters affecting head and eye movement. Consistent with the DCH, networks supporting target fixation in rich environments showed higher $C(X)$ than their counterparts adapted to simple environments. This was true both for dynamics exhibited during behavior in the corresponding environments ('interactive' complexity), and for dynamics evoked with Gaussian noise ('intrinsic' complexity).

Sporns and Lungarella explored the relationship between $C(X)$ and behavior in a different way [60]. As in [59], networks acted as neural controllers during performance of a task (in this case, control of a simulated arm to reach for a target). However, instead of evolving for successful behavior, networks were evolved directly for high $C(X)$. Strikingly, selecting for high $C(X)$ led to networks that were able to perform the task, even though performance on the task had not been explicitly selected for. Finally, Lungarella and Sporns asked how $C(X)$ depends on sensorimotor coupling by comparing neural dynamics of a robotic sensory array in two conditions: (i) unperturbed foveation behavior, and (ii) decoupling of sensory input

and motor output via 'playing back' previously recorded motor activity [61]. They found significantly higher $C(X)$ when sensorimotor coupling was maintained.

Taken together, the above results suggest a strong link between high neural complexity and flexible, adaptive behavior. Of course, in none of these studies is any claim made that the corresponding networks are in any sense conscious.

Extensions and Limitations

The concept of neural complexity has been extended to characterize the selectional responses of neural systems to inputs in terms of 'matching' complexity C_M [62]. C_M measures how well the intrinsic correlations within a neural system fit the statistical structure of a sensory stimulus. Simulations show that C_M is high when intrinsic connectivity is modified so as to differentially amplify those intrinsic correlations that are enhanced by sensory input, possibly reflecting the capacity of a neurally complex system to 'go beyond the information given' in a stimulus [62]. Despite this possibility C_M has not been investigated as thoroughly as has C_N.

C_N has several limitations. In its full form it is computationally prohibitive to calculate for large networks, but in approximation it is less satisfying as a measure. Also, C_N does not reflect complexity in the temporal domain since functional connections are analyzed at zero-lag [23]. Finally, C_N does not take into account *directed* (causal) dynamical interactions for the simple reason that MI is a symmetric measure. This last point is addressed by the alternative measures described below.

Information Integration

The most prominent alternative to C_N is 'information integration' (Φ) [24,63]. Unlike C_N, Φ reflects causal interactions because it is based on 'effective information' (EI), a directed version of MI that relies on the replacement of the outputs of different subsets of the studied system with maximum entropy signals.

Mathematical Definition

Φ is defined as the effective information across the informational 'weakest-link' of a system, the so-called *minimum information bipartition* (MIB; Fig. 5b). It is calculated by the following procedure [63].

Given a system of N elements, identify all possible bipartitions of the system. For each bipartition $A|B$, replace the outputs from A by uncorrelated noise (i. e., maximally entropic activity), and measure how differentiated are the responses of its complement (B). This is the effective information (EI) between A and B:

$$\mathrm{EI}(A \to B) = \mathrm{MI}(A_{H\max}; B) ,$$

where $\mathrm{MI}(A_{H\max}; B)$ is the mutual information between A and B when the outputs from A have maximal entropy. $\mathrm{EI}(A \to B)$ measures the capacity for causal influence of partition A on its complement B (i. e., all possible effects of A on B). Given that $\mathrm{EI}(A \to B)$ and $\mathrm{EI}(B \to A)$ are not necessarily equal, one can define:

$$\mathrm{EI}(A \leftrightarrow B) = \mathrm{EI}(A \to B) + \mathrm{EI}(B \to A) .$$

The minimum information bipartition (MIB) is the bipartition for which the normalized $\mathrm{EI}(A \leftrightarrow B)$ is lowest. Normalization is accomplished by dividing $\mathrm{EI}(A \leftrightarrow B)$ by $\min\{H_{\max}(A); H_{\max}(B)\}$, so that effective information is bounded by the maximum entropy available. The resulting MIB corresponds to the informational weakest link of the system, and the Φ value of the system is the non-normalized $\mathrm{EI}(A \leftrightarrow B)$ across the MIB.

A further stage of analysis has been described [63] in which a system can be decomposed into 'complexes' by calculating Φ for different subsets of elements; a complex is a subset having $\Phi > 0$ that is not included in a larger subset with higher Φ. For a given system, the complex with the maximum value of Φ is called the 'main complex'.

Information Integration, Connectivity, and Complexity

As with neural complexity it is useful to explore what kinds of network structure lead to high values of Φ. Because of computational constraints only comparatively small networks have been investigated for their ability to generate high Φ (i. e., $N = 8$, $K = 16$ as opposed to $N = 32$, $K = 256$ as in [49]). In an initial study, networks optimized for Φ had highly heterogeneous connectivity patterns with no two elements having the same sets of inputs and outputs [63]. At the same time, all nodes tended to emit and receive the same number of connections. These two properties arguably subserve functional segregation and integration, respectively [63].

Although both Φ and C_N depend on a combination of functional integration and segregation, they are sensitive to different aspects of network dynamics. C_N reflects an average measure of integration that, unlike Φ, does not require heterogeneous connectivity. On the other hand, unlike C_N, Φ is determined by the value of an informational measure (EI) across only a single bipartition (the MIB) and is not modified by dynamical transactions across

the remainder of the network. Finally, as mentioned above, Φ but not C_N is sensitive to causality.

Limitations and Extensions

As with C_N, Φ does not measure complexity in the temporal domain [23]. There are also substantial limitations attending measurement of Φ for non-trivial systems. First, it is not possible in general to replace the outputs of arbitrary subsets of neural systems with uncorrelated noise. An alternative version of Φ can be envisaged in which 'transfer entropy' (TE) [64], a directed version of MI, is substituted for EI. TE can be calculated from the dynamics generated by a neural system during behavior and therefore does not require arbitrary perturbation of a system; it measures the *actual* causal influence across partitions whereas EI measures the *capacity* for causal influence. However, a version of Φ based on TE does not in general find the informational 'weakest link' (MIB) of a system since the MIB depends on capacity and not on transient dynamics.

Second, unlike C_N there is presently no well-defined approximation for Φ that removes the need to examine all possible bipartitions of a system. However, it may be possible to make use of some informal heuristics. For example, bipartitions for which the normalized value of EI will be at a minimum will be most often those that cut the system in two halves, i. e., midpartitions [63]. Similarly, a representative rather than exhaustive number of perturbations may be sufficient to obtain at least an estimated value of Φ [63].

The Information Integration Theory of Consciousness

Φ occupies a central place in the 'information integration theory of consciousness' (IITC, [24]). According to this theory, consciousness *is* information integration as measured by Φ. The nature of the conscious content in a system with high Φ is determined by the particular informational relationships within the main complex (the complex with the highest Φ). While there are many similarities between the DCH and the IITC, most obviously that both make strong appeal to a measure of complexity, there are also important differences of which we emphasize two:

(i) Because Φ measures the capacity for information integration, it does not depend on neural activity per se. The IITC predicts that a brain where no neurons were active, but in which they were potentially able to react, would be conscious (perhaps of nothing). Similarly, a brain in which each neuron were stimulated to fire as an exact replica of your brain, but in which synaptic interactions had been blocked, would *not* be conscious [24]. The DCH has neither of these implications.

(ii) On the IITC Φ is an adequate measure of the 'quantity' of consciousness, therefore any system (biological or artificial) with sufficiently high Φ would necessarily be conscious. According to the DCH, high C_N is necessary but not sufficient for consciousness.

Point (ii) is particularly important in view of the finding that an arbitrarily high Φ can be obtained by a system as simple as a Hopfield network, which is a fully connected network with simple binary neuronal elements [23]. By choosing the synaptic strengths according to an exponential rule it can be shown that the corresponding Φ value scales linearly with network size, such that $\Phi(\mathrm{X}) = N$ bits for a network X of size N nodes. On the IITC this result leads to the counterintuitive conclusion that a sufficiently large Hopfield network will be conscious. Another challenge for the IITC in this context is the fact that the probability distributions determining entropy values (and therefore by extension Φ values) depend on subjective decisions regarding the spatial and temporal granularity with which the variables in a system are measured ([23,65] but see [24]).

Causal Density

A balance between dynamical integration and differentiation is likely to involve dense networks of causal interactions among neuronal elements. Causal density (c_d) is a measure of causal interactivity that captures both differentiated and integrated aspects of these interactions [23,66]. It differs from C_N by detecting causal interactions, differs from Φ by being sensitive to dynamical interactions across the whole network, and differs from both by being based not on information theory but instead on multivariate autoregressive modeling.

Mathematical Definition

Causal density (c_d) measures the fraction of interactions among neuronal elements in a network that are causally significant (Fig. 5c). It can be calculated by applying 'Granger causality' [67,68], a statistical concept of causality that is based on prediction: If a signal x_1 causes a signal x_2, then past values of x_1 should contain information that helps predict x_2 above and beyond the information contained in past values of x_2 alone [67]. In practice, Granger causality can be tested using multivariate regression modeling [69]. For example, suppose that the temporal dynamics of two time series, $x_1(t)$ and $x_2(t)$ (both of length T),

can be described by a bivariate autoregressive model:

$$x_1(t) = \sum_{j=1}^{p} A_{11,j} x_1(t-j) + \sum_{j=1}^{p} A_{12,j} x_2(t-j) + \xi_1(t)$$
$$x_2(t) = \sum_{j=1}^{p} A_{21,j} x_1(t-j) + \sum_{j=1}^{p} A_{22,j} x_2(t-j) + \xi_2(t) \tag{4}$$

where p is the maximum number of lagged observations included in the model (the model order, $p < T$), A contains the coefficients of the model, and ξ_1, ξ_2 are the residuals (prediction errors) for each time series. If the variance of ξ_1 (or ξ_2) is reduced by the inclusion of the x_2 (or x_1) terms in the first (or second) equation, then it is said that x_2 (or x_1) *G-causes* x_1 (or x_2). In other words, x_2 G-causes x_1 if the coefficients in A_{12} are jointly significantly different from zero. This relationship can be tested by performing an F-test on the null hypothesis that $A_{12,j} = 0$, given assumptions of covariance stationarity on x_1 and x_2. The magnitude of a significant interaction can be measured either by the logarithm of the F-statistic [70] or, more simply, by the log ratio of the prediction error variances for the restricted (R) and unrestricted (U) models:

$$gc_{2\to 1} = \begin{cases} \log \dfrac{\mathrm{var}(\xi_{1\mathrm{R}(12)})}{\mathrm{var}(\xi_{1\mathrm{U}})}, & \text{if } gc_{2\to 1} \text{ is significant} \\ 0, & \text{otherwise,} \end{cases}$$

where $\xi_{1\mathrm{R}(12)}$ is derived from the model omitting the $A_{12,j}$ (for all j) coefficients in Eq. (4) and $\xi_{1\mathrm{U}}$ is derived from the full model, where $gc_{2\to 1}$ refers to G-causality from x_2 to x_1.

Importantly, G-causality is easy to generalize to the multivariate case in which the G-causality of x_1 is tested in the context of multiple variables $x_2 \ldots x_N$. In this case, x_2 G-causes x_1 if knowing x_2 reduces the variance in x_1's prediction error when the activities of all other variables $x_3 \ldots x_n$ are also included in the regression model. Both bivariate and multivariate G-causality have been usefully applied to characterizing causal interactions in simulated [71,72] and biological [73] neural systems.

Following a Granger causality analysis, both normalized (c_d) and non-normalized (c_dw) versions of causal density of a network X with N nodes can be calculated as:

$$c_\mathrm{d}(\mathrm{X}) = \frac{\alpha}{N(N-1)},$$
$$c_\mathrm{dw}(\mathrm{X}) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} gc_{j\to i}$$

where α is the total number of significant causal interactions and $N(N-1)$ is the total number of pairwise interactions among elements. While normalized causal density is bounded to the range [0,1] the non-normalized version is unbounded.

High causal density indicates that elements within a system are both globally coordinated in their activity (in order to be useful for predicting each others activity) and at the same time dynamically distinct (reflecting the fact that different elements contribute in different ways to these predictions). Therefore, as with both C_N and Φ, c_d reflects both functional integration and functional segregation in network dynamics.

Conditions Leading to High Causal Density

In terms of connectivity, computational models show that both fully connected networks (having near-identical dynamics at each node) and a fully disconnected networks (having independent dynamics at each node) have low c_d and c_dw; by contrast, randomly connected networks have much higher values [71]. More detailed connectivity studies remain to be conducted.

An initial attempt to analyze behavioral conditions leading to high causal density was made by [66] revisiting the model of target fixation described previously [59]. To recapitulate, in this model networks were evolved in both 'simple' and 'complex' environments where environmental complexity was reflected by unpredictable target movement and by variation in parameters affecting head and eye movement (Fig. 6). Causal density in this model was calculated from first-order differenced time series of the ten sensorimotor neurons and it was found that highest values of causal density occurred for networks evolved and tested in the complex environments. These results mirrored those obtained with C_N, indicating an association between a high value of a complexity measure and adaptive behavior in a richly structured environment.

Extensions and Limitations of Causal Density

A practical problem for calculating causal density is that multivariate regression models become difficult to estimate accurately as the number of variables (i. e., network elements) increases. For a network of N elements, the total number of parameters in the corresponding multivariate model grows as pN^2, and the number of parameters to be estimated for any single time series grows linearly (as pN), where p is the model order (Eq. (4)). We note that these dependencies are much lower than the factorial dependency associated with Φ and C_N, and may therefore be more readily circumvented. One possible approach may

involve the use of Bayesian methods for limiting the number of model parameters via the introduction of prior constraints on significant interactions [74]. In neural systems, such prior constraints may be derived, for example, on the basis of known neuroanatomy or by anatomically based clustering procedures.

Several other extensions to causal density are suggested by enhancements to the statistical implementation of Granger causality:

- Non-linear G-causality methods based, for example, on radial-basis-function kernels allow causal density to detect both linear and non-linear causal interactions [75,76].
- Partial G-causality (based on partial coherence) enhances robustness to common input from unobserved variables, supporting more accurate estimates of causal density in systems that cannot be fully observed [77].
- G-causality has a frequency-dependent interpretation [70,73] allowing causal density to be assessed in specific frequency bands.

Causal Density and Consciousness

Although causal density is not attached to any particular theory of consciousness, it aligns closely with the DCH because it is inherently a measure of process rather than capacity. Causal density cannot be inferred from network anatomy alone, but must be calculated on the basis of explicit time series representing the dynamic activities of network elements during behavior. It also depends on all causal interactions within the system, and not just on those interactions across a single bipartition, as is the case for Φ. Finally, causal density incorporates the temporal dimension more naturally than is the case for either C_N or Φ; while the latter measures functional interactions at zero-lag only, causal density incorporates multiple time lags as determined by the order parameter p (Eq. (4)).

The foregoing descriptions make clear that although existing formal measures may have heuristic value in identifying functional integration and functional segregation in neural dynamics, they remain inadequate in varying degrees. C_N can reflect process, can be computed for large systems in approximation, but does not capture causal interactions. Φ captures causal interactions, is infeasible to compute for large neural systems, and can be shown to grow without bound even for certain simple networks. Also, Φ is a measure of capacity rather than process but this is a deliberate feature of the IITC. c_d reflects all causal interactions within a system and is explicitly a measure of process, but it also is difficult to compute for large systems. An additional and important practical limitation of C_N, Φ, and c_d is that they apply only to statistically stationary dynamics.

Empirical Evidence

We turn now to empirical evidence relevant to the DCH. Much of this evidence comes from patients with focal brain lesions and neuroimaging of healthy subjects using functional magnetic resonance imaging (fMRI), electroencephelography (EEG) and magnetoencephalography (MEG). Although current experimental methods are not sufficiently powerful to confirm or refute the DCH, their application, separately and in combination, yields much useful information. A detailed review appears in [17]; below we select some pertinent features.

Involvement of the Thalamocortical System

A wealth of experimental evidence attests to thalamocortical involvement in consciousness, as demonstrated by both clinical studies and by experiments using normal subjects. Clinical studies show that damage to non-specific (intralaminar) thalamic nuclei can abolish consciousness in toto [78], whereas damage to cortical regions often deletes specific conscious features such as color vision, visual motion, conscious experiences of objects and faces, and the like [79]. No other brain structures show these distinctive effects when damaged.

Conscious functions in normal subjects are usefully studied by comparison with closely matched controls who perform the function unconsciously, an approach known as 'contrastive analysis' [80,81,82]. An emerging consensus among contrastive studies is that conscious contents correspond to widespread thalamocortical activation as compared to unconscious controls [81]. For example, Dehaene and colleagues have shown widespread fMRI activation peaks in parietal, prefrontal, and other cortical regions for conscious perception of visual words, as compared to unconscious inputs which activated mainly primary visual cortex [83]. Along similar lines, a recent fMRI study of motor sequence learning showed a shift from widespread cortical involvement during early learning (when conscious attention is required) to predominantly subcortical involvement during later learning phases (when skill production is comparatively 'automatic') [84].

Dynamical Correlates: Binocular Rivalry

A classical example of contrastive analysis makes use of the phenomenon of binocular rivalry, in which different

images are projected to each eye [85]. Despite the integrative nature of consciousness these images, if sufficiently different, are not combined into a single composite; rather, conscious experience alternates between them. Srinivasan and colleagues used magnetoencephalography to measure brain responses to flickering visual stimuli under rivalrous conditions [86,87]. A vertical grating flickering at one frequency was presented to one eye and a horizontal grating flickering at another frequency was presented to the other; these different frequencies allowed stimulus-specific brain responses to be isolated in the neuromagnetic signal, a technique known as 'frequency tagging' [88]. As expected for such stimuli, subjects perceived only one grating at any given time. It was found that the power of the frequency-tag of a stimulus was higher by 30–60% across much of the cortical surface when that stimulus was perceptually dominant compared to when it was perceptually suppressed. Moreover, there was a large increase in coherence among distant brain regions, consistent with the idea that conscious perception is associated with widespread integration of neural dynamics mediated by reentry. Although this coherence increase is not a direct test of the DCH it is consistent with the theory and underscores the value of looking for *dynamical* correlates of consciousness.

Cosmelli and colleagues have used a similar paradigm to show that development of a perceptual dominance period arises in neural terms as an extended dynamical process involving the propagation of activity throughout a distributed brain network beginning in occipital regions and extending into more frontal regions [89]. Such a 'wave of consciousness' might reflect underlying integrative processes that lead to the formation of a dynamic core [90]. Chen and colleagues modified the rivalry paradigm so that subjects saw both gratings with both eyes but had to differentially *pay attention* to one or the other [91]. Power increases but not coherence increases were found for the attended stimulus, suggesting that attention may not involve the same global integrative processes implicated in consciousness. Finally, Srinivasan has shown that coherence increases during dominance are due partly to increased phase locking to the external stimulus and partly to increased synchrony among intrinsic network elements, again in line with the idea that consciousness involves coalescence of a distributed functional cluster within the brain [92].

Sleeping, Waking, and Anesthesia

Binocular rivalry involves constant conscious level and changing conscious *content*. Experimental evidence relevant to conscious *level* comes from studies involving transitions between sleeping and waking, anesthesia, epileptic absence seizures and the like. Many studies have tracked changes in endogenous activity across these various transitions but direct assessments of specific complexity measures are mostly lacking. Nonetheless, current findings are broadly consistent with the DCH. As noted in Subsect. "Consciousness as Discrimination", absence seizures and slow-wave sleep (but not rapid-eye-movement sleep) are characterized by hypersynchronous neural activity that may correspond to reduced functional segregation [22]. Anesthesia has particular promise for further experimental study because global anesthetic states can be induced via a wide variety of pharmacological agents having diverse physiological effects. Moreover, proposed unifying frameworks, such as Mashour's 'cognitive unbinding' theory [93], share with the DCH the idea that loss of consciousness can arise from diminished functional integration. In line with Mashour's proposal, John and colleagues have observed at anesthetic loss-of-consciousness (i) functional disconnection along the rostrocaudal (front-to-back) axis and across hemispheres (measured by coherence changes), and (ii) domination of the EEG power spectrum by strongly anteriorized low frequencies [94].

The development of transcranial magnetic stimulation (TMS) has opened the possibility of studying effective (causal) connectivity in the brain. TMS non-invasively disrupts specific cortical regions by localized electromagnetic induction. Massimini and colleagues combined high-density EEG with TMS to test whether effective connectivity among distant brain regions is diminished during sleep [95]. Applying a TMS pulse to premotor cortex during quiet wakefulness led to a sequence of waves propagating throughout the cortex, but this widespread propagation was mostly absent during non-rapid eye movement (slow wave) sleep. While these results do not allow calculation of any specific complexity measure they are consistent with both the IITC and the DCH.

In summary, it is clear that enhanced experimental and analytical methods are needed in order to test adequately whether C_N (or other specific measures) are modulated as predicted by the DCH. Initial attempts to calculate C_N directly from neural dynamics have not been successful (see [96] for a review) although a link to complexity is suggested by the discovery of small-world networks in functional brain dynamics [96,97].

Related Theoretical Proposals

Dynamical Systems Theory and Metastability

We have already mentioned that the measures of complexity discussed in this article apply only to statistically

stationary dynamics (Subsect. "Causal Density and Consciousness"). This restriction contrasts sharply with an alternative tradition in theoretical neuroscience that focuses on non-stationary brain dynamics and which emphasizes the tools of dynamical systems theory. This alternative tradition can be traced back to early suggestions of Turing [98] and Ashby [99] and was concisely expressed by Katchalsky in 1974: "… waves, oscillations, macrostates emerging out of cooperative processes, sudden transitions, patterning, etc., seem made to order to assist in the understanding of integrative processes in the nervous system" [100]. More recently the dynamical systems approach has been championed in neuroscience by, among others, Haken [101] under the rubric 'coordination dynamics' and Freeman who has produced a steady stream of papers exploring dynamical principles in brain activity [102,103]. Valuable reviews of work in this tradition can be found in [104,105,106].

A key concept in the dynamical systems approach is 'metastability' which describes dynamics that are "distinguished by a balanced interplay of integrating and segregating influences" (see p. 26 in [107]). While this definition is obviously similar to the intuition driving neural complexity, metastability has been fleshed out, not in the concepts of information theory or time-series analysis, but instead in the language of attractor dynamics. A dynamical system inhabits a metastable regime when there are no stable fixed points but only partial attraction to certain phase relationships among the system variables. At the level of neural dynamics metastability may reflect the ongoing creation and dissolution of neuronal assemblies across distributed brain regions [105,107,108]. A now classical experimental example of metastability comes from a study in which subjects were asked to flex a finger in response to a periodic tone, initially in a syncopated manner [107]. As the frequency of the tone increases, the syncopated response becomes harder to maintain until a critical point is reached at which the subject switches to a synchronous mode of response. Strikingly, this behavioral phase transition is accompanied by a corresponding transition in the patterning of neuromagnetic cortical signals. At the critical point, where there is partial attraction to both syncopated and synchronous response modes, both behavioral and neural dynamics are dominated by metastability. For other evidence of metastability in the brain see [109].

Metastability characterizes an important aspect of conscious experience, namely that conscious events are rapidly adaptive and fleeting [17]. Consciousness is remarkable for its present-centeredness [5,6]. Immediate experience of the sensory world may last at most a few seconds and our fleeting cognitive present is surely less than half a minute in duration. This present-centeredness has adaptive value for an organism by allowing time enough to recruit a broad network of task-related neural resources while permitting neural dynamics to evolve responses to subsequent events. Thus, conscious experience can be described by an interplay of segregating and integrating influences in both the temporal (metastability) and spatial (complexity) domains. A key theoretical challenge is to work out in greater detail the relationship between these two concepts.

Global Workspace Theory

Beginning in 1988 [80] Baars developed a cognitive theory of consciousness under the rubric 'global workspace (GW) theory' [80,81,110]. The cornerstone of GW theory is the idea that consciousness involves a central resource (the GW) which enables distribution of signals among numerous otherwise informationally encapsulated and functionally independent specialized processors. GW theory states that mental content becomes conscious mental content when it gains access to the GW such that it can influence a large part of the brain and a correspondingly wide range of behaviors. A key aspect of GW theory is that conscious contents unfold in an integrated, serial manner but are the product of massively parallel processing among the specialized processors. The integrated states of the GW follow each other in a meaningful but complex progression that depends on multiple separate processes, each of which might have something of value to add to the ongoing constitution of the GW. Although these notions are compatible with the DCH, they do not by themselves specify dynamical properties to the same level of detail.

A dynamical approach to GW theory has been pursued by Wallace [112] and separately by Dehaene, Changeux and colleagues [111,113,114]. Wallace adopts a graph-theoretic perspective proposing that the GW emerges as a 'giant component' among transient collections of otherwise unconscious processors. The formation of a giant component in graph theory denotes a phase transition at which multiple sub-networks coalesce to form a large network including the majority of network nodes [115]. Dehaene and colleagues have built a series of computational models inspired by GW theory, to account for a variety of psychological phenomena including 'attentional blink' [111], 'inattentional blindness' [114], and effortful cognitive tasks [113]. These models are based on the concept of a 'neuronal global workspace' in which sensory stimuli mobilize excitatory neurons with long-range cortico-cortical axons, leading to the genesis of global activ-

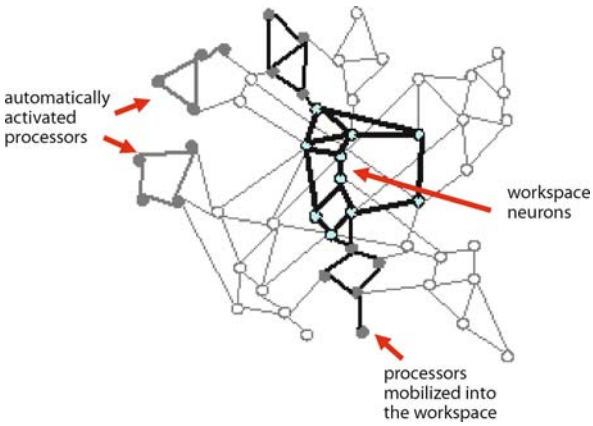

Consciousness and Complexity, Figure 7
A schematic of the neuronal global workspace. A central global workspace, constituted by long-range cortico-cortical connections, assimilates other processes according to their salience. Other automatically activated processors do not enter the global workspace. Adapted from [111]

ity patterns among so-called 'workspace neurons' (Fig. 7). This model, and that of Wallace both, predict that consciousness is 'all or nothing' – i. e., a gradual increase in stimulus visibility should be accompanied by a sudden transition (ignition) of the neuronal GW into a corresponding activity pattern. As with Wallace, although some dynamic properties of the neuronal GW have been worked out and are compatible with the DCH, a rigorous account of how the model relates to neural complexity has not been attempted.

Neuronal Synchrony and Neuronal Coalitions

The association of neural synchrony with consciousness arose from its proposed role as a mechanism for solving the so-called 'binding problem', which in general terms refers to the problem of coordinating functionally segregated brain regions. The binding problem is most salient in visual perception for which the functional and anatomical segregation of visual cortex contrasts sharply with the unity of a visual scene. Since the 1980s a succession of authors have proposed that the binding problem is solved via neuronal synchronization [116,117] and both experimental evidence [118,119] and computational models have borne out the plausibility of this mechanism [31]. In the 1990s, starting with an early paper by Crick and Koch [7], this proposal grew into the hypothesis that consciousness itself is generated by transient synchronization among widely distributed neuronal assemblies, with particular emphasis on oscillations in the gamma band ($\sim$ 40 Hz) [39,120]. In support of this idea we have already seen that conscious perception correlates with increased synchrony of a (non-gamma-band) visual 'frequency tag' [87], and several studies have reported associations between gamma-band synchrony and consciousness [121,122,123]. However, synchrony-based theories of binding (and by extension consciousness) remain controversial [124] and there is not yet evidence that disruptions of gamma-band synchrony lead to disruptions of conscious contents [12].

From the perspective of the DCH a deeper concern with the synchrony hypothesis is that it accounts only for integration, and not for the combination of integration and differentiation that yields the discriminatory power of consciousness. In a recent position paper [125], Crick and Koch reversed their previous support for gamma-band synchrony as a sufficient mechanism for consciousness, favoring instead the notion of competition among 'coalitions' of neurons in which winning coalitions determine the contents of consciousness at a given time. Such neuronal coalitions bear similarities to the decades-old notion of Hebbian assemblies [126] on a very large and dynamic scale. They also suggest that unconscious processing may consist largely of feed-forward processing whereas consciousness may involve standing waves created by bidirectional signal propagation, a proposal advanced as well by Lamme [127]. Crick and Koch note that the 'coalition' concept is similar to the dynamic core concept [125] although lacking in the detailed formal specification of the latter.

A possible role for gamma-band synchrony in both the DCH and in Crick and Koch's framework is that it may facilitate the formation but not the ongoing activity of the core (or a coalition) [125]. In this light it is suggestive that correlations between gamma-band synchrony and consciousness tend to occur at early stages of conscious perception [121,122].

Outlook

Scientific accounts of consciousness continue to confront the so-called 'hard problem' of how subjective, phenomenal experiences can arise from 'mere' physical interactions in brains, bodies, and environments [128,129]. It is possible that new concepts will be required to overcome this apparent conceptual gap [130]. It is equally likely that increasing knowledge of the mechanisms underlying consciousness will lead these philosophical conundrums to fade away, unless they have empirical consequences [81,125]. In short, to expect a scientific resolution to the hard problem as it is presently conceived may be to misunderstand the role of science in explaining na-

ture. A scientific theory cannot presume to replicate the experience that it describes or explains; a theory of a hurricane is not a hurricane [18]. If the phenomenal aspect of experience is irreducible, so is the fact that physics has not explained why there is something rather than nothing, and this ontological limit has not prevented physics from laying bare many mysteries of the universe.

The approach described in this article is one of developing *explanatory correlates* of consciousness, namely properties of neural dynamics that are experimentally testable and that account for key properties of conscious experience. Thanks to accelerating progress in experimental techniques and increased attention to theory, the outlook for this approach is healthy. We close by suggesting some key areas for further study:

Development of a large-scale model of a dynamic core Although progress in large scale neural network modeling has been rapid [131], we currently lack a sufficiently detailed model of environmentally coupled thalamocortical interactions needed to test the mechanistic plausibility of the DCH. Having such a model should also allow substantive connections to be drawn between the DCH and GW theory.

Development of new experimental methods New methods are needed to track neuronal responses at sufficient spatio-temporal resolutions to support accurate estimation of C_N and other complexity measures during different conscious and unconscious conditions. Among current methods fMRI has poor time resolution and measures neural activity indirectly, while MEG/EEG lacks spatial acuity and is unable to record details of thalamic responses.

Complexity and metastability New theory is needed to relate the class of complexity measures described in this article to metastability, which analyzes functional segregation and integration in the temporal domain.

Emergence and 'downward causality' New theory is also needed to better understand how global dynamical states arise from their basal interactions and how these global states can constrain, enslave, or otherwise affect properties at the basal level [39]. Applied to consciousness and to cognitive states in general, such downward causality can suggest functional roles and may even help reconcile the phenomenology of free-will with physiological fact.

Bibliography

Primary Literature

1. Haldane E, Ross G (1985) The philosophical work of Descartes. Cambridge University Press, Cambridge
2. Popper K, Eccles JF (1977) The self and its brain. Springer, New York
3. Penrose R (1994) Shadows of the mind: A search for the missing science of consciousness. Oxford University Press, Oxford
4. McGinn C (1991) The problem of consciousness. Blackwell, Oxford
5. James W (1904) Does consciousness exist? Philos J Psychol Sci Methods 1:477–491
6. Edelman GM (1989) The remembered present. Basic Books, New York
7. Crick F, Koch C (1990) Towards a neurobiological theory of consciousness. Semin Neurosci 2:263–275
8. Dalton TC, Baars BJ (2003) Consciousness regained: The scientific restoration of mind and brain. In: Dalton TC, Evans RB (eds) The lifecycle of psychological ideas: Understanding the prominence and the dynamics of intellectual change. Springer, Berlin, pp 203–247
9. Koch C (2004) The quest for consciousness: A neurobiological approach. Roberts, Greenwood Village
10. Metzinger T (2000) Neural correlates of consciousness: Empirical and conceptual questions. Press MIT, Cambridge
11. Llinás R, Ribary U, Contreras D, Pedroarena C (1998) The neuronal basis for consciousness. Philos Trans Soc R Lond Biol B Sci 353:1841–1849
12. Rees G, Kreiman G, Koch C (2002) Neural correlates of consciousness in humans. Nat Rev Neurosci 3(4):261–70
13. Crutchfield JP, Young K (1989) Inferring statistical complexity. Phys Rev Lett 63:105–108
14. Zurek WH (1990) Complexity, entropy, and the physics of information. Addison-Wesley, Redwood City
15. Adami C (2002) What is complexity? Bioessays 24:1085–1094
16. Tye M (2007) Qualia. In: Zalta E (ed) The Stanford Encyclopedia of Philosophy (Summer 2008 Edition). http://plato.stanford.edu/archives/sum2008/entries/qualia/
17. Seth AK, Baars BJ (2005) Neural Darwinism and consciousness. Conscious Cogn 14:140–168
18. Edelman GM (2003) Naturalizing consciousness: A theoretical framework. Proc Natl Acad Sci USA 100(9):5520–5524
19. Seth AK, Baars BJ, Edelman DB (2005) Criteria for consciousness in humans and other mammals. Conscious Cogn 14(1):119–139
20. Edelman DB, Baars BJ, Seth AK (2005) Identifying the hallmarks of consciousness in non-mammalian species. Conscious Cogn 14(1):169–187
21. Laureys S (2005) The neural correlate of (un)awareness: Lessons from the vegetative state. Trends Cogn Sci 9:556–559
22. Tononi G, Edelman GM (1998) Consciousness and complexity. Science 282:1846–1851
23. Seth AK, Izhikevich E, Reeke GN, Edelman GM (2006) Theories and measures of consciousness: An extended framework. Proc Natl Acad Sci USA 103(28):10799–10804
24. Tononi G (2004) An information integration theory of consciousness. Neuroscience BMC 5(1):42
25. Gazzaniga MS (2005) Forty-five years of split-brain research and still going strong. Nat Rev Neurosci 6:653–659
26. Pashler H (1994) Dual-task interference in simple tasks: data and theory. Psychol Bull 116:220–244
27. Zeki S (1990) A century of cerebral achromatopsia. Brain 113(6):1721–1777
28. Edelman GM, Tononi G (2000) A universe of consciousness: How matter becomes imagination. Basic Books, New York

29. Tononi G, McIntosh AR, Russell DP, Edelman GM (1998) Functional clustering: identifying strongly interactive brain regions in neuroimaging data. Neuroimage 7:133–149
30. Tononi G, Sporns O, Edelman GM (1992) Reentry and the problem of integrating multiple cortical areas: Simulation of dynamic integration in the visual system. Cerebral Cortex 2(4):31–35
31. Seth AK, McKinstry JL, Edelman GM, Krichmar JL (2004) Visual binding through reentrant connectivity and dynamic synchronization in a brain-based device. Cerebral Cortex 14:1185–99
32. Edelman GM (1987) Neural Darwinism. Basic Books, New York
33. Edelman GM (1993) Selection and reentrant signaling in higher brain function. Neuron 10:115–125
34. Friston KJ, Tononi G, Reeke GN, Sporns O, Edelman GM (1994) Value-dependent selection in the brain: Simulation in a synthetic neural model. Neuroscience 59:229–243
35. Tononi G, Sporns O, Edelman GM (1999) Measures of degeneracy and redundancy in biological networks. Proc Natl Acad Sci USA 96:3257–3262
36. Edelman GM, Gally J (2001) Degeneracy and complexity in biological systems. Proc Natl Acad Sci USA 98(24):13763–13768
37. Kim J (1998) Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation. Press MIT/Bradford Books, Cambridge
38. Wegner D (2003) The illusion of conscious will. Press MIT, Cambridge
39. Thompson E, Varela F (2001) Radical embodiment: Neural dynamics and consciousness. Trends Cogn Sci 5:418–425
40. Tononi G, Sporns O, Edelman GM (1994) A measure for brain complexity: Relating functional segregation and integration in the nervous system. Proc Natl Acad Sci USA 91:5033–5037
41. Vanduffel W, Payne BR, Lomber SG, Orban GA (1997) Functional impact of cerebral connections. Proc Natl Acad Sci USA 94:7617–7620
42. Papoulis A, Pillai SU (2002) Probability, random variables, and stochastic processes, 4th edn. McGraw-Hill, New York
43. Jones DS (1979) Elementary information theory. Clarendon Press, Oxford
44. McGill WJ (1954) Multivariate information transmission. Trans IEEE Inform Theory 4:93–111
45. de Lucia M, Bottaccio M, Montuori M, Pietronero L (2004) A topological approach to neural complexity. Phys Rev E 71:016114
46. Tononi G, Edelman GM, Sporns O (1998) Complexity and coherency: Integrating information in the brain. Trends Cogn Sci 2:474–484
47. Bressler SL (1995) Large-scale cortical networks and cognition. Brain Res Brain Res Rev 20:288–304
48. Friston K (2002) Beyond phrenology: what can neuroimaging tell us about distributed circuitry? Annu Rev Neurosci 25: 221–250
49. Sporns O, Tononi G, Edelman GM (2000) Theoretical neuroanatomy: Relating anatomical and functional connectivity in graphs and cortical connection matrices. Cerebral Cortex 10:127–141
50. Sporns O, Tononi G (2002) Classes of network connectivity and dynamics. Complexity 7(1):28–38
51. Seth AK, Edelman GM (2004) Theoretical neuroanatomy: Analyzing the structure, dynamics and function of neuronal networks. In: Ben Naim E, Fraunfelder H, Toroczkai Z (eds) Complex networks. Lecture Notes in Physics. Springer, Berlin, pp 487–518
52. Sporns O, Chialvo D, Kaiser M, Hilgetag C (2004) Organization, development and function of complex brain networks. Trends Cogn Sci 8:418–425
53. Buckley CL, Bullock S (2007) Spatial embedding and complexity: The small-world is not enough. In: Almeida e Costa F (ed) Proceedings of the Ninth European Conference on Artificial Life. Springer, Berlin, pp 986–995
54. Mitchell M (1997) An introduction to genetic algorithms. Press MIT, Cambridge
55. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small world' networks. Nature 393:440–442
56. Sporns O (2004) Complex neural dynamics. In: Jirsa VK, Kelso JAS (eds) Coordination dynamics: Issues and trends. Springer, Berlin, pp 197–215
57. Sporns O, Kötter R (2004) Motifs in brain networks. PLoBiol S 2:e369–e369
58. Sporns O (2006) Small-world connectivity, motif composition, and complexity of fractal neuronal connections. Biosystems 85:55–64
59. Seth AK, Edelman GM (2004) Environment and behavior influence the complexity of evolved neural networks. Adapt Behav 12:5–21
60. Sporns O, Lungarella M (2006) Evolving coordinated behavior by maximizing information structure. In: Rocha L, Yaeger L, Bedau MS, Floreano D, Goldstone RL, Vespignani A (eds) Proceedings of the 10th European Conference on Artificial Life. Press MIT, Cambridge, pp 323–330
61. Lungarella M, Sporns O (2006) Mapping information flow in sensorimotor networks. PLoComput S Biol 2:e144–e144
62. Tononi G, Sporns O, Edelman GM (1996) A complexity measure for selective matching of signals by the brain. Proc Natl Acad Sci USA 93:3422–3427
63. Tononi G, Sporns O (2003) Measuring information integration. Neuroscience BMC 4(1):31
64. Schreiber T (2000) Measuring information transfer. Phys Rev Lett 85(2):461–4
65. Werner G (2007) Perspectives on the neuroscience of cognition and consciousness. Biosystems 87:82–95
66. Seth AK (2005) Causal connectivity of evolved neural networks during behavior. Netw Comput Neural Syst 16:35–54
67. Granger CWJ (1969) Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37:424–438
68. Seth AK (2007) Granger causality. Scholarpedia, p 15501
69. Hamilton JD (1994) Time series analysis. Princeton University Press, Princeton
70. Geweke J (1982) Measurement of linear dependence and feedback between multiple time series. J Amer Stat Assoc 77:304–13
71. Seth AK (2008) Causal networks in simulated neural systems. Cogn Neurodyn 2:49–64
72. Seth AK, Edelman GM (2007) Distinguishing causal interactions in neural populations. Neural Comput 19:910–933
73. Ding M, Chen Y, Bressler S (2006) Granger causality: Basic theory and application to neuroscience. In: Schelter S, Winterhalder M, Timmer J (eds) Handbook of Time Series Analysis. Wiley, Wienheim, pp 438–460

74. Zellner A (1971) An introduction to Bayesian inference in econometrics. Wiley, New York
75. Ancona N, Marinazzo D, Stramaglia S (2004) Radial basis function approaches to nonlinear granger causality of time series. Phys Rev E 70:056221
76. Chen Y, Rangarajan G, Feng J, Ding M (2004) Analyzing multiple nonlinear time series with extended Granger causality. Phys Lett A 324:26–35
77. Gao S, Seth AK, Kendrick K, Feng J. Partial granger causality: Eliminating exogenous input. J Neurosci Methods 172(1)79–93
78. Bogen JE (1995) On the neurophysiology of consciousness: I An overview. Conscious Cogn 4:52–62
79. Kolb B, Whishaw IQ (1996) Fundamentals of human neuropsychology, 4th edn. Freeman WH, New York
80. Baars BJ (1988) A cognitive theory of consciousness. Cambridge University Press, New York
81. Baars BJ (2002) The conscious access hypothesis: Origins and recent evidence. Trends Cogn Sci 6:47–52
82. Lau HC, Passingham RE (2006) Relative blindsight in normal observers and the neural correlate of visual consciousness. Proc Natl Acad Sci USA 103:18763–18768
83. Dehaene S, Naccache L, Cohen L, Bihan DL, Mangin JF, Poline JB, Rivière D (2001) Cerebral mechanisms of word masking and unconscious repetition priming. Nat Neurosci 4:752–758
84. Floyer-Lea A, Matthews PM (2004) Changing brain networks for visuomotor control with increased movement automaticity. Neurophysiol J 92:2405–2412
85. Blake R, Logothetis N (2002) Visual competition. Nat Rev Neurosci 3:13–21
86. Tononi G, Srinivasan R, Russell DP, Edelman GM (1998) Investigating neural correlates of conscious perception by frequency-tagged neuromagnetic responses. Proc Natl Acad Sci USA 95:3198–3203
87. Srinivasan R, Russell DP, Edelman GM, Tononi G (1999) Increased synchronization of magnetic responses during conscious perception. J Neurosci 19:5435–5448
88. Silberstein RB, Schier MA, Pipingas A, Ciorciari J, Wood SR, Simpson DG (1990) Steady-state visually evoked potential topography associated with a visual vigilance task. Brain Topogr 3:337–347
89. Cosmelli D, David O, Lachaux J-P, Martinerie J, Garnero L, Renault B, Varela F (2004) Waves of consciousness: Ongoing cortical patterns during binocular rivalry. Neuroimage 23:128–140
90. Nunez P, Srinivasan R (2006) A theoretical basis for standing and traveling brain waves measured with human EEG with implications for an integrated consciousness. Clin Neurophysiol 117:2424–2435
91. Chen Y, Seth AK, Gally JS, Edelman GM (2003) The power of human brain magnetoencephalographic signals can be modulated up or down by changes in an attentive visual task. Proc Natl Acad Sci USA 100:3501–3506
92. Srinivasan R (2004) Internal and external neural synchronization during conscious perception. Int Bifurcat J Chaos 14:825–842
93. Mashour GA (2004) Consciousness unbound: Toward a paradigm of general anesthesia. Anesthesiology 100:428–433
94. John ER, Prichep LS, Kox W, Valdés-Sosa P, Bosch-Bayard J, Aubert E, Tom M, di Michele F, Gugino LD (2001) Invariant reversible QEEG effects of anesthetics. Conscious Cogn 10:165–183
95. Massimini M, Ferrarelli F, Huber R, Esser SK, Singh H, Tononi G (2005) Breakdown of cortical effective connectivity during sleep. Science 309:2228–2232
96. Stam CJ, Jones BF, Nolte G, Breakspear M, Scheltens P (2007) Small-world networks and functional connectivity in Alzheimer's disease. Cereb Cortex 17:92–99
97. Stephan KE, Hilgetag CC, Burns GA, O'Neill MA, Young MP, Kötter R (2000) Computational analysis of functional connectivity between areas of primate cerebral cortex. Philos Trans Soc R Lond Biol B Sci 355:111–126
98. Turing A (1950) Computing machinery and intelligence. Mind 59:433–460
99. Ashby WR (1952) Design for a brain: The origin of adaptive behaviour. Chapman Hall, London
100. Katchalsky A, Rowland V, Hubermann B (1974) Neurosci Res Prog Bull 12
101. Haken H (2002) Brain dynamics. Springer, New York
102. Freeman WJ (1975) Mass action in the nervous system. Academic Press, New York
103. Freeman WJ (2005) A field-theoretic approach to understanding scale-free neocortical dynamics. Biol Cybern 92(6):350–359
104. Kelso JAS (1995) Dynamic patterns: The self-organisation of brain and behavior. Press MIT, Cambridge
105. Werner G (2007) Metastability, criticality and phase transitions in brain and its models. Biosystems 90:496–508
106. Izhikevich EM (2006) Dynamical systems in neuroscience: The geometry and excitability of bursting. Press MIT, Cambridge
107. Bressler S, Kelso J (2001) Cortical coordination dynamics and cognition. Trends Cogn Sci 5:26–36
108. Friston KJ (1997) Transients, metastability, and neuronal dynamics. Neuroimage 5:164–171
109. Fingelkurts A (2004) Making complexity simpler: Multivariability and metastability in the brain. Int Neurosci J 114:843–862
110. Baars BJ (2005) Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. Prog Brain Res 150:45–53
111. Dehaene S, Sergent C, Changeux J-P (2003) A neuronal network model linking subjective reports and objective physiological data during conscious perception. Proc Natl Acad Sci USA 100:8520–8525
112. Wallace R (2005) Consciousness: A mathematical treatment of the neuronal global workspace model. Springer, New York
113. Dehaene S, Kerszberg M, Changeux J-P (1998) A neuronal model of a global workspace in effortful cognitive tasks. Proc Natl Acad Sci USA 95:14529–14534
114. Dehaene S, Changeux J-P (2005) Ongoing spontaneous activity controls access to consciousness: a neuronal model for inattentional blindness. PLoBiol S 3:e141–e141
115. Bollobás B (1985) Random graphs. Academic Press, London
116. von der Malsburg C (1995) Binding in models of perception and brain function. Curr Opin Neurobiol 5:520–526
117. Singer W, Gray CM (1995) Visual feature integration and the temporal correlation hypothesis. Annu Rev Neurosci 18:555–586

118. Fries P, Reynolds JH, Rorie AE, Desimone R (2001) Modulation of oscillatory neuronal synchronization by selective visual attention. Science 291:1560–1563
119. Steinmetz PN, Roy A, Fitzgerald PJ, Hsiao SS, Johnson KO, Niebur E (2000) Attention modulates synchronized neuronal firing in primate somatosensory cortex. Nature 404:187–190
120. Engel AK, Fries P, König P, Brecht M, Singer W (1999) Temporal binding, binocular rivalry, and consciousness. Conscious Cogn 8:128–151
121. Melloni L, Molina C, Pena M, Torres D, Singer W, Rodriguez E (2007) Synchronization of neural activity across cortical areas correlates with conscious perception. Neurosci J 27: 2858–2865
122. Palva S, Linkenkaer-Hansen K, Näätänen R, Palva J (2005) Early neural correlates of conscious somatosensory perception. Neurosci J 25:5248–5258
123. Meador KJ, Ray PG, Echauz JR, Loring DW, Vachtsevanos GJ (2002) Gamma coherence and conscious perception. Neurology 59:847–854
124. Shadlen MN, Movshon JA (1999) Synchrony unbound: a critical evaluation of the temporal binding hypothesis. Neuron 24:67–77 111
125. Crick F, Koch C (2003) A framework for consciousness. Nat Neurosci 6:119–126
126. Hebb DO (1949) The organization of behavior. Wiley, New York
127. Lamme VAF (2006) Towards a true neural stance on consciousness. Trends Cogn Sci 10:494–501
128. Chalmers DJ (1996) The conscious mind: In search of a fundamental theory. Oxford University Press, Oxford
129. Bennett MR, Hacker PMS (2003) Philosophical foundations of neuroscience. Blackwell, Oxford
130. Thompson E (2004) Life and mind: A tribute to Francisco Varela. Phenomenol Cogn Sci 3:381–98
131. Izhikevich E, Gally JS, Edelman GM (2004) Spike-timing dynamics of neuronal groups. Cerebral Cortex 14:933–944
132. Seth AK, Dienes Z, Cleeremans A, Overgaard M, Pessoa L (2008) Measuring consciousness: relating behavioural and neurophysiological approaches. Trends Cogn Sci 12:314–21

Books and Reviews

Baars DJ, Banks WP, Newman JB (2003) Essential sources in the scientific study of consciousness. Press MIT, Cambridge
Dorogovtsev SN, Mendes JFF (2003) Evolution of Networks: from biological networks to the Internet and WWW. Oxford University Press, Oxford
Edelman GM (2004) Wider than the sky: The phenomenal gift of consciousness. Yale University Press, New Haven
Edelman GM, Tononi G (2000) A universe of consciousness: How matter becomes imagination. Basic Books, New York
Koch C (2004) The Quest for Consciousness: A Neurobiological Approach. Roberts, Greenwood Village
Metzinger T (2000) Neural correlates of consciousness. Press MIT, Cambridge
Tononi G (2004) An information integration theory of consciousness. Neuroscience BMC 5(1):42
Seth AK, Izhikevich E, Reeke GN, Edelman GM (2006) Theories and measures of consciousness: An extended framework. Proc Natl Acad Sci USA 103(28):10799–10804

Continuum Percolation

Isaac Balberg
The Racah Institute of Physics, The Hebrew University, Jerusalem, Israel

Article Outline

Glossary
Definition of the Subject
Introduction
Lattice Percolation
Principal Issues in Continuum Percolation
Percolation Thresholds in the Continuum
The Critical Behavior of the Dynamical Properties in the Continuum
Future Directions
Acknowledgments
Bibliography

Glossary

Site percolation Assuming a lattice in which the sites can be occupied, with a probability p^{s}, two sites are assumed locally connected if they are nearest neighbors and occupied. A group of sites in which each of its' sites is occupied and locally connected to at least one other site in the group is called a connected cluster. If there is a cluster that connects the edges of an "infinite" lattice, the cluster is called the percolation cluster. The lowest p^{s} for which such a cluster is found is called the percolation threshold, $p_{\mathrm{c}}^{\mathrm{s}}$.

Bond percolation Assuming a lattice, the segment between nearest neighbor sites is called a bond. A bond can be occupied with a probability p^{b}. Two bonds are locally connected if they are both occupied and have a common site. A group of bonds in which each is connected to at least one other bond is called a connected cluster of bonds. If the span of this cluster is infinite we have a "percolation cluster" and the lowest p^{b} for which such a cluster is formed is called the bonds percolation threshold, $p_{\mathrm{c}}^{\mathrm{b}}$.

Critical behavior Percolation can be modeled as a phase transition and thus various properties have a power-low dependence on $|p - p_{\mathrm{c}}|$ for small values of $|p - p_{\mathrm{c}}|/p_{\mathrm{c}}$. The exponent that describes this dependence is called the critical exponent.

Universal and non universal behavior If the critical exponent depends only on the dimensionality of the system, we say that the critical behavior is universal. If the

exponent depends on other parameters of the system, we say that the behavior is non universal.

Continuum percolation In a system of objects one can define a local connectivity criterion such as the overlap of pores in porous media. The objects are assumed to be randomly distributed in space, they may have various and variable shapes and sizes and they may interact with each other. A group of objects such that each object is locally connected to at least one other object in the group is a connected cluster. The cluster that has an infinite span is the "percolating" cluster in the continuum and the lowest concentration of objects that yield such a cluster defines the percolation threshold.

Critical fractional volume The content of the objects at the onset of global connectivity, i. e. the percolation threshold, is usually characterized in the continuum by the measurable fractional volume content of the objects of interest in the system. This volume at the percolation threshold is the critical fractional volume.

Excluded volume In the case where the local connectivity is determined by a partial overlap of the volumes of two equal objects, the volume in space in which the two centers of the two objects can be, and must be, in order to have such an overlap is defined as the excluded volume of the object. If the objects are not equal one has to define a corresponding average.

Average bonds per object The average number of objects that are locally connected to a given object in the system. This quantity, at the percolation threshold, B_c, is the quantity that characterizes topologically the onset of global connectivity i. e. the onset of percolation.

"Pointedness" The parameter that can be defined qualitatively as the degree of deviation from sphericity of a given object. The manifestation of this property for different objects, that have the same excluded volume, is that with its increase the value of B_c decreases.

Critical behavior of dynamical properties The critical behavior of a property that has to do with flow in the system, such as electrical conductivity and fluid permeability. In the continuum, it is related to the distribution of the local values of the dynamical parameter associated with a given bond. If the average of this parameter effects the global critical behavior of the system, its contribution is added to the "universal" critical behavior that is found in lattices.

The random void and the inverted random void systems A system of pores or topologically similar systems which are reminiscent of the "Swiss Cheese". In the random void (RV) case one is concerned with the network excluding the pores. The mirror image of this system, i. e. when one is concerned with the network that consists of the pores (or particles that can coalesce) is called the inverted random void (IRV) system. In these systems the "neck" formed by the separation (RV) or the overlap (IRV) of two adjacent pores determines the local dynamical property.

Tunneling percolation The conduction process in a system where the "local connectivity" is not determined by a geometrical contact but is determined electronically. This is in particular by inter-object tunneling. The corresponding connected system exhibits a percolation-like critical behavior of the dynamical properties.

"Physically controlled percolation" When an externally applied quantity, such as an electric field or mechanical pressure, affects the parameters that characterize the percolation-like behavior, one may call the corresponding phenomena "physically controlled percolation".

Definition of the Subject

In this review we will describe the two main features that are very different in lattice percolation and in continuum percolation and which bring about new concepts that are not encountered in the "more traditional" and well developed theory of percolation in lattices [121]. The origin of the difference is that in continuum percolation we are dealing with "real" objects that have sizes and shapes, and that are randomly distributed in space, while in lattices we are dealing with abstract mathematical objects such as dots and line segments in a priori defined locations. The result is that in the continuum we have to use different quantities to describe the percolation threshold and we obtain a much richer variability of the corresponding behaviors of the dynamical properties, such as the electrical transport. To equip the reader with the basic concepts needed for the discussion of continuum percolation we also include a short introduction to the relevant principles of lattice percolation and then we discuss in some detail the very basic concepts and achievements in continuum percolation. Finally, we discuss open questions and possible future developments associated with this field.

Percolation theory is concerned with the effects of local and global connectivity on the geometrical and dynamical properties of systems. Within the limited scope of this review we cannot discuss the numerous systems and many properties associated with continuum percolation and we deal only with the principles of this field. However, in order to illustrate the great variety of systems and properties that constitute this area of research, we mention now single, typical, examples of the most conspicuous of those.

Continuum percolation encompasses all the many body systems from the smallest, the elementary particles systems [104], to the largest, the galactic-cosmological systems [109]. Correspondingly continuum percolation was applied to discuss and describe structures and phenomena in physics, chemistry, earth sciences and biology, as well as communication and traffic networks. As some typical examples, let us mention, the electrical properties of numerous composites [128], porous media [125], microemulsions [36,69], disordered semiconductors [137], disordered superconductors [91], molecular [75] and macromolecular [39] liquids, nano-tubes in composites [45,64] and suspensions [64], quantum dot composites [27], thin metal films [118], layered materials [47], quasi-crystals [98], chemical networks [92] and transport in them [96], biological networks [134] and flow in them [117], bioinformatics [99], and ecological systems [132]. Then there are the properties associated with the flow of charge carriers, such as electrical conductivity [10], the flow of liquids such as the permeability [88], as well as corresponding processes [77] such as displacements [133], drainage [6], dispersion [107], hydrological flows [31] and diffusion [130]. Very closely related is the class of rheological properties [57] such as the viscosity [84]. The possible variation in the local ("bond strength") parameters makes the concepts of continuum percolation also applicable to information and traffic management [135]. Finally, there are the related areas of the elastic [81], the dielectric [67] and the magnetic [122] properties, and their relation to electrical properties [28,85,93].

In this review we will limit ourselves to a discussion of the percolation threshold and electrical conductivity, as they are the most studied properties in "real systems", with the hope that the principles that are understood to govern this property will be sufficient for approaching the understanding of the various systems and properties such as those mentioned above. In particular this review is expected to be also helpful for the analysis of related behaviors such as the electrical noise [127], piezoresistivity [71,128], magneto-transport [29], alternating current conductivity [43], and the effects of interplay between electrical and elasto-thermal properties [8]. An example of a generalization of the concepts used in the discussion of the electrical conductivity to other properties is the extension of the theory [127] and the experimental analysis [102] of the electrical conductivity to the electrical noise.

Turning to the historical highlights of the development of lattice percolation theory and continuum percolation theory we realize that by the very nature of this limited and subjective review there will be many important works that will not be mentioned here and need to be summarized in a more expanded review. We point out however that our previous reviews [11,17,26,32] cover the basic understanding of the present subject for the corresponding years.

The history of percolation theory is usually considered to begin with the works of Flory in 1941 [60] and Stockmayer in 1943 [123] on gelation, and, by definition, as a flow problem in the work of Broadbent and Hammersley in 1957 [33], who also coined the name of the corresponding processes as percolation phenomena [136]. These works were followed by very many works that consolidated the concepts associated with the theory of percolation on lattices. Excellent general reviews on lattice percolation were given by Zallen in 1983 [136], Stauffer and Aharony in 1992 [121], and Sahimi in 1994 [105] and some aspects of it were considered in more specialized reviews such as by Kirkpatrick in 1973 [80], Shklovskii and Efros in 1984 [114], Isichenko in 1992 [77] and Sahimi in 1998 [106].

The principal concepts, that were developed for lattices were the site and bond occupation probabilities, p, and the resulting clusters, that are groups of occupied sites or bonds that are connected by having pairs of occupied sites (or bonds) when the latter are nearest neighbors. The corresponding theory was concerned with the properties of the clusters that enable this connectivity, and in particular those of the infinite-spanning cluster, when it exists. The onset of this cluster at a given occupation probability, p_c, is associated with the onset of global connectivity as well as with the possible onsets of "flows" through the systems. The major developments of percolation theory were then along two lines that will be emphasized in this review: the determination of the onset of global connectivity, and the variation of geometrical and dynamical properties of such systems as the percolation threshold is approached. The corresponding developments in the 1960's culminated in the 1969 work of Kasteleyn and Fortuin (that is detailed in Fortuin and Kasteleyn [62]). They were able to show that the p transition through p_c can be described as a second order phase transition [119], thus justifying a posteriori the previous findings that the above various properties behave as power laws of $|p - p_c|$, with powers that are only dimensionally dependent. The latter behavior as $p \to p_c$ is known as the critical behavior, and the independence of the exponents on other system parameters is referred to as the universal critical behavior of the quantities in percolation systems. While the initial motivation to consider percolation problems was to account for the behavior of "real" systems, the considerations of such systems and the realization that they cannot be simply dis-

cussed using the terms of lattice systems was postponed to the 1970's. Since "real" systems are made of objects and these are randomly dispersed in the continuum, the corresponding research field of "real" systems became synonymous with the later coined concept of "continuum" percolation.

The first major step in trying to account generally for more realistic systems can be considered the finding of Scher and Zallen (hereafter S&Z) in 1970 [108]. They found that for properly placed hard core spheres in lattice points, the onset of global connectivity will take place at a volume fraction of the objects in the system, τ_c, that is nearly a constant for all possible lattices in a given dimension D. This is in contrast with the relatively wide range of p_c values for the various lattices in the same D, thus suggesting that at τ_c there is an onset of a fundamental topological connectivity. Similar invariants were found by Powell in 1979 [97] who essentially used a model in which "conducting" and "non-conducting" spheres fill a box under "gravitation". Another approach, which relates to problems of flow via porous-like media or impurity band conduction-like systems, was taken by Holcomb and Rehr in 1969 [76], and Pike and Seager in 1974 [95]. They considered essentially the fractional occupied volume of spherical pores, when the local connectivity is defined by partial overlap of the pores, at the onset of global connectivity. They also found invariants in the case for systems for which no volume can be attached to the object such as widthless sticks in a two dimensional system [25]. A very significant insight into the problem of percolation thresholds of systems in the continuum was provided by Shante and Kirkpatrick (hereafter S&K) in 1971 [112]. They confirmed that the underlying topology of the connected objects, i. e. the average number of connected permeable spheres, per given sphere, at the onset of percolation, is the same as that of the "fine grid" lattice. It took then more than a decade until Balberg and co-workers in 1983 [14,20] and 1984 [21,22] provided the generalization of the above concepts to zero-volume and "soft-core" non-parallel objects. They systematically established the concepts of the average excluded volume, $\langle V_{ex}\rangle$, and the number of bonds per object, B_c. Consequently the quantity $B_c = \rho_c\langle V_{ex}\rangle$, where ρ_c is the number of objects per unit volume at the onset of percolation, became the guide for *the phenomenological theory* of the percolation threshold in the continuum [131] and its application [30]. For example, these concepts of $\langle V_{ex}\rangle$ and B_c have enabled the explanation of the nearly zero volume of the conducting phase needed for the onset of percolation in many systems [9] and the analysis of the various cases where the objects are not entirely permeable [17].

The first step towards a rigorous derivation of results in continuum percolation is probably the suggestion of Coniglio, DeAngelis and Forlani who developed, in 1977 [44], a series expansion in the objects density, ρ, to describe the average size of the finite clusters in fluid systems. This utilization of liquid theory [74] was translated to the prediction of the ρ_c values for partially permeable spheres by DeSimone, Stratt and Demoulini in 1986 [49], and to an analytic prediction of trends, in the B_c dependence on the type of the objects, by Bug, Safran and Webman in 1985 [34]. Extension of the latter approach and the application of extrapolation methods (by fitting the results to the known average cluster size dependence on $|\rho - \rho_c|$) finally yielded very accurate predictions for B_c [3,53]. In particular the work of Drory et al. in 1997 [56] has lent a firm basis for a rigorous theory for percolation thresholds in the continuum, which followed Drory's works in 1996 [50,51] and 1997 [52] that proved that continuum percolation is also a second order-like phase transition. An insight concerning the variation of B_c (that is much less straightforward than the variation of ρ_c) from a system of one type of objects to a system of another type of objects was suggested by Alon, Balberg and Drory who introduced the concept of "*pointedness*" in 1991 [4]. Recently, some works on permeable systems in high dimensions have shown that applying intuition for the prediction of various specific behaviors is not always justified [63,68]. The case at hand was the finding of Wagner, Balberg and Klein, in 2006 [131], that for $D \to \infty$ the value of B_c can become smaller than the intuitively expected unity [136].

Turning to the critical behavior of percolation systems in the continuum, we note that the study of the dynamical properties, and in particular the study of the electrical resistivity, became an intensive area of research in the 1970's. The first work to test experimentally the predicted properties on "real" systems can be considered to be the study of Abeles, Pinch and Gittleman in 1975 [2] regarding the critical behavior of the electrical resistivity of granular metals [1] at the onset of metallic conduction. They found that the conductivity exponent that characterizes the critical behavior, t, was 1.9 in accordance with the universal values predicted for lattices. This was followed by numerous works that yielded the same conclusion on other electrical properties, such as the electrical noise [40]. In fact various experiments on "real" systems have also confirmed the lattice-like geometrical-structural exponents, such as the correlation length exponent, ν [18,79] and the fractal dimension of the percolation cluster, D_f [113]. In parallel, the developments in computers and consequent computer simulations in the 1980's have enabled quite accurate simulations for the determination of the critical exponents in

continuum systems, yielding again very good agreement with the universal predictions [20]. On the other hand, the many experimental results that did not conform with the universal predictions [7,94] were largely ignored by the mainstream of the research. Even the "toy" model of non-universal behavior of the electrical resistivity, proposed by Kogut and Straley (hereafter K&S) in 1979 [82], was bypassed for a few years, until it became the cornerstone of the understanding of non universal behavior of the dynamical properties. This overlooking of the theory of K&S was apparently for a good reason, since all that time there was no known link to connect the features of the continuum system with the parameters of their "toy" model. This "missing" link was provided by the work of Halperin, Feng and Sen in 1985 [72]. They showed that non universality can be a fundamental property of dynamical properties in continuum systems. By finding the distribution function of the pore sizes, in sedimentary rock-like systems, they were able to show its immediate consequence regarding the resistors value distribution, and thus the appearance of non universality in accordance with the model of K&S. Their finding opened the way then to the understanding of the many dynamical properties in the various continuum systems that were mentioned above. In particular, soon after these findings, in 1987, Balberg showed [10] that tunneling conduction in random systems can also be mapped onto a K&S-like resistor distribution. This realization has enabled, at least, the qualitative interpretation of the numerous experimental results in the broadest group of studied continuum systems, i. e. composite materials, and the broadest group of studied properties, i. e. the various electrical transport properties. The very simple theoretical predication that was derived for the latter case yielded, however, critical exponents that were much larger than the many experimentally observed values that were reported in the literature [26]. The other difficulty with the corresponding physical picture was that it did not explain how is it that systems in which all the particles are electrically connected on the one hand, and in which there is no geometrical connectivity, between any pair of them, on the other hand, yield behaviors that are identical to those obtained in bona fide percolation systems [13]. The two problems have been resolved in principle only very recently, the first by Grimaldi and Balberg in 2006 [70], and the other by Toker et al. in 2003 [124]. These developments will be described in more detail below.

Introduction

Percolation phenomena are associated with the consequences of the connectivity in systems composed of elementary members that are distributed in some random manner, after a criterion for the connectivity between two such elements is defined. The corresponding theory that deals with those phenomena tries to describe geometrical structures of connected parts in those systems as well as their dynamical properties. The most interesting and consequently the most studied manifestations of percolation phenomena are those taking place in the close proximity to the onset of global connectivity in the system. This onset is known as the percolation threshold, and the behavior of the geometrical and dynamical properties in the vicinity of this threshold is known as the critical behavior of the percolation system. It was shown that this critical behavior can be described as a second-order phase transition and thus, close enough to the percolation threshold, the various properties are given by power laws [121] that are reminiscent of the dependencies found in the most well-known phase transitions [119] such as the simplest magnetic transitions [116].

There are thousands of papers that have been written on the theory and the applications of percolation in the continuum. These applications are found in very many fields of physics, chemistry, biology, geophysics and communications, and there are very many relevant properties that are discussed in terms of the concepts of this theory. Since all these cannot be summarized within the framework of the present review we will concentrate here on the theory and the understanding of the principal concepts and on the major experimental manifestations of the theory and its consequences. In particular we will focus on the central essential concept of the *percolation threshold* in the continuum and on the dynamical property that was studied more than any other property in corresponding systems, i. e., the *electrical conductivity*.

Since the fundamental concepts that are used for the description of percolation in the continuum are based on the great advancements made in the theory of lattice percolation, we will start this review by briefly outlining the principles of the latter theory. Following the fact that there are excellent textbooks [106,121,136] and many review articles on lattice percolation, this background will be given very briefly in Sect. "Lattice Percolation". It is expected that this brief background will be sufficient for the understanding of the basic ideas of continuum percolation that will be discussed then in more detail here, in Sects. "Principal Issues in Continuum Percolation", "Percolation Thresholds in the Continuum", and "The Critical Behavior of the Dynamical Properties in the Continuum". In those sections we will put emphasis on the properties of continuum percolation that cannot be extrapolated simply from their lattice counterparts. For example,

the almost zero-percolation threshold as well as the non universal behavior of various physical properties do not follow the scenarios of, and expectations from, lattice percolation. Finally, in Sect. "Future Directions" we consider the major issues in continuum percolation the solution of which seem to be important in order to deepen our understanding of this and related fields of research. In Sect. "Bibliography" we provide a rather very limited list of papers relevant to the subjects considered here with the hope that this will be sufficient as a guide for the application of continuum percolation theory in various and many related research areas.

Lattice Percolation

To introduce the above mentioned basic concepts we start by considering the very simple model of the two dimensional square lattice that is illustrated in Fig. 1. Let us assume a probability, p, that a lattice site is occupied, and denote the occupied sites by full circles. The step that is very important here, and will be very crucial when we discuss continuum percolation, is the definition of the local connectivity between members in a given system. In the present lattice these members are the occupied sites and two such sites are defined to be connected if they are first nearest neighbors. In the particular case shown in Fig. 1 the number of nearest neighbors of a given site (the coordination number of the lattice), z, is 4. We can now define a (connected) cluster size as the number of sites such that each of them is connected to at least one (same or other) occupied site in the cluster. In Fig. 1 we see then clusters of size 1, size 2, size 4 and a larger cluster that connects opposite edges of the system. The latter is called the spanning cluster. In percolation theory we are concerned mainly with "infinite" systems, i. e. with systems where the details (e. g. the sites themselves and the intersite distances) are much smaller than the size of the system. Correspondingly, in an "infinite" system we call the spanning cluster, the percolation, or the infinite, cluster. The illustration given in Fig. 1 reflects then a finite (small) portion of the infinite system. Such small portions are used below for the description and definitions of local and global quantities of the system, but it has to be remembered that quantitative characterizations of the various properties are considered meaningful only in the (statistically sufficient) "infinite" system limit.

From Fig. 1 it is easy to appreciate that the number of clusters and their sizes will grow as the site occupation, p, increases, as can be derived quantitatively by various models and simulations [121,136]. The most important consequence, however, is that there will be a value of

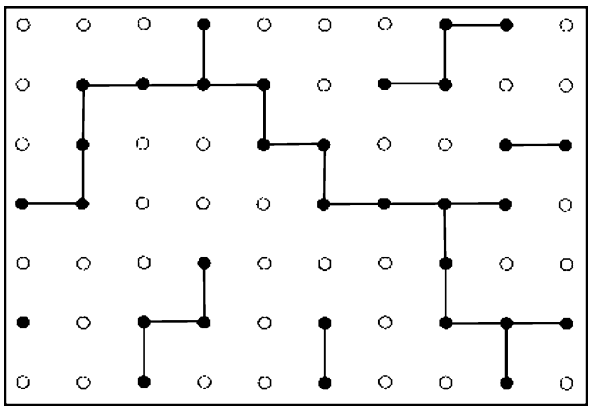

Continuum Percolation, Figure 1
A two dimensional illustration of a portion of an infinite lattice. The open circles represent lattice sites and the full circles represent occupied lattice sites. The segments connecting two nearest neighbor occupied sites represent a local connection or a bond. In this illustration there are finite clusters of connected sites (of size 1, 2 and 4) and a "spanning" cluster that connects two opposite edges of the system. In a system of infinite size the latter is known as the "infinite" or the "percolation" cluster. (From [32])

p, that is known in the literature as p_c, such that for $p < p_c$ there will be no percolation cluster, while for $p \geq p_c$ there will be such a cluster. This p_c is known as the *percolation threshold* of the lattice. The behavior of the various geometrical or physical properties for p values in the close proximity of p_c is known as the *critical behavior*. The above mentioned considerations of the percolation as a phase transition yield then that these properties will be well described by power laws of the proximity to the threshold $|p - p_c|/p_c$. For example, the average cluster size, S, is defined usually [121] as the average number of sites in a cluster, per an occupied site in the lattice. The average that is thus found for the finite clusters in the lattice, behaves, for both, $p > p_c$ and $p < p_c$, as

$$S \propto |p - p_c|^{-\gamma} \,. \tag{1}$$

In particular, the phase transition-like behavior yields that exponents, such as γ, depend only on the lattice dimension D, and are thus independent of the details (e. g. the lattice type) of the system. This property is known as the *universality* and the corresponding dependence on $|p - p_c|$ as the *universal behavior*. The universal values of γ are, for example, 2.4 for $D = 2$ and 1.8 for $D = 3$. The other important property of the clusters is their geometrical extent, i. e. their effective (e. g. gyration) radius (say, in units of the lattice spacing). The corresponding average for the finite clusters in the lattice (e. g., per sites connected to a given site in the cluster; Stauffer and Aharony [121]) de-

fines then the "connectivity" length in the system, ξ. This length, as to be expected from the above discussion, behaves as

$$\xi \propto |p - p_c|^{-\upsilon} , \tag{2}$$

where ν is a "universal" exponent. The quantity ξ, known also as the *correlation length*, is the most important parameter of the system since it characterizes its connectivity and provides the basic length scale for dealing with the behavior of the properties of interest in the system. The well known values of υ are 1.33 for $D = 2$ and 0.88 for $D = 3$.

We note in passing that in Fig. 1 we may also consider the occupancy of the bonds (i. e. the segments between the sites) for which the local bond connectivity criterion is the occupation of two bonds that share a site [129]. It is obvious that for a given lattice there is a simple numerical factor that relates the site occupancy, p^s, and the bond occupancy, p^b [136]. Below, we will use then the more general p concept to describe both quantities except when the difference between the two types of occupancy needs to be emphasized.

Turning to the dynamic properties associated with the connectivity of the system in lattices we consider now the global conductance of a system, G. This is the dynamic property that we will discuss in Sect. "The Critical Behavior of the Dynamical Properties in the Continuum". for continuum systems. Viewing then the system shown in Fig. 1 as consisting of "bonds" (between two occupied sites), let us assume that each bond is a resistor bar of a given resistance value. In such a system there will be electrical conduction from right to left or from top to bottom, only if there is a spanning cluster. Examining the system illustrated in Fig. 1 from this point of view, we see that only a fraction of the 18 resistor-bonds in this cluster (15 from left to right or 11 from top to bottom) will participate in the conduction in the respective direction. On the other hand, there will be a few resistance-bonds (3 from left to right, or 7 from top to bottom) that do not participate in the conduction process. The bonds of the first type belong to the *backbone* of the percolation cluster while the others are known as the *"dead ends"* of the percolation cluster.

Let us consider now the structure of the backbone net. We can imagine the backbone as (say, a cubic) net of "links" and "nodes". These "links" intersect then at the "nodes" that define the corresponding net. The average length of the links is easily appreciated to be of the order of ξ, since the "holes" in the backbone net can encapsulate only finite clusters, and these, as we saw above, have their diameter distributed around ξ. Turning to the structure of the links, we can look at it from a geometrical point of view or an electrical point of view. In the first picture let us consider those bonds in the link that connect two sites in the link such that there is no other "indirect" connection between them within the distance ξ (i. e. when only the members of the link) are included. We call the corresponding bond a "*singly connected bond*". From the electrical point of view the latter bonds are simply those that carry the same current as the link (say, through a plane that intersects the link, but only through members of the link). The parts of the backbone that are confined between two adjacent "singly connected bonds" are known as the "blobs" of the backbone. Hence, since the system will conduct only for $p > p_c$, the backbone can be envisioned as consisting of, say, a square or a cubic, network of "links" that are made of series connections of "singly connected bonds" and "blobs". The model described here is known as the *links-nodes-blobs* (the LNB) model [32,114,120,121].

Considering the expected resistance of a cubic (or hypercubic) LNB network let us assume that the resistance of a link is R_ξ. If the sample size is L we have then, on the average, L/ξ links that connect one edge of the sample with its counterpart. Correspondingly, there will be $(L/\xi)^{D-1}$ parallel links between these "opposite-edges", where D is the dimensionality of the system. Since the link's resistance is R_ξ, the resistance of the whole network $\mathbf{R}_L$ will be given by

$$\mathbf{R}_L = R_\xi (\xi/L)^{D-2} . \tag{3}$$

The crucial step here is then "only" to estimate the value of R_ξ [32,114]. The corresponding easiest approach (that can be termed also the lattice approach) is to assume that all the bond-resistors have the same resistance value $\mathbf{r}_0$ [121,122]. In passing we note that when the resistors' values come from a distribution, even if a lattice structure is assumed as the platform of the resistors network, the problem is classified as a problem in continuum percolation. This follows of course the fact that in "real" (or continuum) systems a distribution in the resistance values of the local resistors is usually expected (see Sect. "The Local Resistors and Their Distributions"). The deviation from the simple $\mathbf{r}_0$ assumption that is used in lattice models will be shown in Sect. "The Critical Behavior of the Dynamical Properties in the Continuum" to lead to the very different and rich behaviors that are encountered in continuum percolation.

However, with the above assumption of all resistors having the same value $\mathbf{r}_0$, we can estimate the value of R_ξ in the LNB model as follows. We saw that the link consists of singly connected bonds and blobs. Let us determine first how many singly connected bonds we expect to have in a link of length ξ. Assuming a lattice with $p > p_c$ and assuming that there are L_1 such bonds in this link [114],

let us cut out $(p - p_c)/p$ of the bonds in the whole lattice. The number of singly connected bonds that will be cut in a link will be then $L_1(p - p_c)/p$. If, upon the increase of L_1 as $p \to p_c$, this quantity reaches unity, there will be on the average one singly connected bond missing in a link and "all" the links will not conduct. In the limit of $\xi \to \infty$ (i. e. $(p - p_c)/p_c \ll 1$) we approach then the onset of "no percolation" with $L_1 \propto (p - p_c)^{-1}$. If we neglect the blobs (that, following the above, are aggregates of parallel resistors and thus have lower resistance than the same length chain of singly connected resistors), the resistance of the link can be estimated by its lower bound of $R_\xi = \mathbf{r}_0 L_1 \propto (p - p_c)^{-1}$. Since the blobs do contribute to the resistance of the link, one expects that $R_\xi > \mathbf{r}_0 L_1$ and that R_ξ will behave as $(p - p_c)^{-\zeta}$ where ζ is an exponent that is dimensional-dependent. While the calculation of the value of ζ is not straightforward [122], it can be well approximated by fractal models [48]. That the value of ζ is larger, but not much larger, than unity is to be expected, since, as in the "dilution" of the network of links (when $p \to p_c$), there is also a dilution in the local structure of the blobs, and the links will approach the limit of a chain of singly connected bonds. In other words as $p \to p_c$ the resistance $\mathbf{R}_L$ will be affected by both the dilution of the links in the network Eq. (3) and the dilution of the resistors in the blobs. The latter effect means a "stronger" (in comparison with the one that would be encountered in a link that is made only of "singly connected bonds") divergence of the sample resistance as $p \to p_c$ and thus it is manifested by ζ that are, for $D < 6$, larger than 1. Indeed, it was found that $\zeta \approx 1.3$ for $D = 2$ and $\zeta \approx 1.1$ for $D = 3$ [121]. All the above results can be summarized now by presenting the critical behavior of the global resistance, $\mathbf{R}_L$, in a percolation system of size L, by

$$\mathbf{R}_L (\equiv 1/G) \propto (p - p_c)^{-t} , \tag{4}$$

where t is the critical exponent of the global conductance G, that is given, in view of the above, by

$$t = (D - 2)\nu + \zeta . \tag{5}$$

The numerical t values that were derived by various corresponding analytical approximations, or Monte Carlo simulations, are known as the universal values of the conductivity exponent, t_{un}. These values are, in particular [121,136], 1.3 for $D = 2$ and ≈ 2 for $D = 3$. The deviation from these "lattice" values that represent the behavior of the "global network" in lattices will be discussed in Sect. "The Critical Behavior of the Dynamical Properties in the Continuum".

Principal Issues in Continuum Percolation

In the previous section we have briefly reviewed the principal concepts of lattice percolation where the systems have sites (or bonds) on lattices and these can be empty or occupied with a probability $p^{\mathrm{s}}(p^{\mathrm{b}})$. This yields a percolation threshold p_c that depends on the particular lattice of interest. The physical properties on the other hand, all of which are associated with the existence of the corresponding bonds in the system, are all characterized by the same single-valued physical parameters. In continuum percolation the system is composed of objects (or structures) that are *randomly placed in space*, that may be of *various sizes and various shapes* and, if non-spherical, may have a *distribution of their orientations*. Correspondingly, the physical parameters that characterize the bonds may vary from one bond to another in a manner that may or may not be determined solely by the local geometry and environment of the bond. The two most significant issues that are of concern in continuum percolation are then the connectivity of the system, as reflected by the percolation threshold, in systems of objects, and the possibility of a different critical behavior of the geometrical and physical properties in comparison with that of lattices.

To make this general discussion more specific, we start by an illustration of a well known prototype of a two dimensional continuum percolation system which consists of randomly distributed circles, as shown in Fig. 2. The circles in this model can be described as permeable, overlapping, interpenetrating, or of "soft core". Such a system can represent a metal sheet in which holes have been drilled at random. In this case the electrical conduction between opposite edges of the sample is carried by the "background" (the sample without the holes) of the sheet. We call this "interpretation" of the system, the random void (RV), or the "Swiss Cheese" model. The system shown in Fig. 2 can be also considered as representing a mirror image of the RV system, i. e. to be taken to represent pores (filled, say, with a conducting liquid) in a porous medium, or a collection of isolated and fused metallic grains embedded in an insulating matrix as in granular metals. In these cases, the conducting parts are the pores or the grains, and their network provide the global conductivity of the system. The transport in the latter system can be of course of a liquid or an electrical charge. Correspondingly, this view of the system is known as the inverted random void (IRV) model.

In the latter model the criterion for local connectivity (that is equivalent to the simple bond connection in lattice percolation) is rather simple; two circles are connected if they partially overlap. While (as we will see later) the local

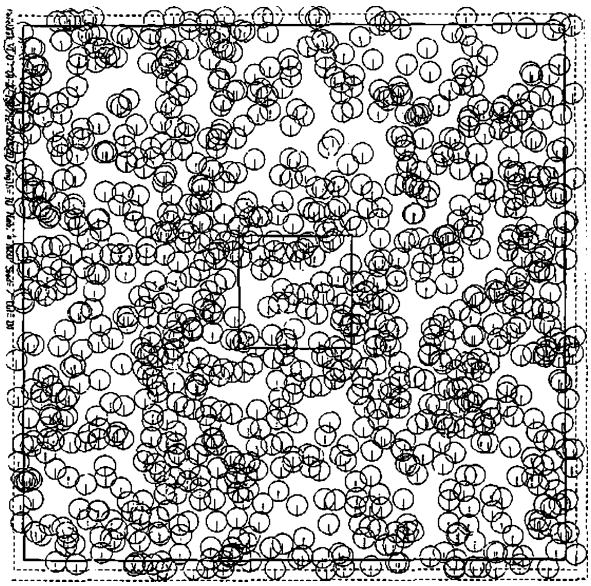

Continuum Percolation, Figure 2
A two dimensional illustration of a continuum system. The circles are distributed randomly and two of them are considered connected if they partially overlap. There are finite clusters here as well as an "infinite" cluster of connected circles. A small finite portion of the system is enclosed in the square shown at the center. (From [130])

connectivity criterion is not always so simple in continuum systems, this one is easy to understand intuitively and thus it is the most abundant model in continuum percolation. In what follows, and in particular, for the discussion of the "connecting criteria" and their interpretation in the continuum, we will usually illustrate our arguments by figures that show small portions of the infinite system, such as the area enclosed in the square that confines the center of Fig. 2.

It is already the first examination of the system shown in Fig. 2 that makes one appreciate the fundamental questions of continuum percolation: how to define quantitatively the percolation threshold in such a system (there is no apparent meaning to p), and how the local degree of objects overlap determines the local and global geometrical and dynamical properties of the system. We will discuss these two issues in Sects. "Percolation Thresholds in the Continuum" and "The Critical Behavior of the Dynamical Properties in the Continuum" of our review. In contrast, we will not describe here the numerous natural and artificial systems for which the consequences of these issues are of great significance, leaving these for more specific reviews to come. We provided, however, a small representative list of such systems in Sect. "Definition of the Subject".

Percolation Thresholds in the Continuum

The Basic Concepts

The first attempt to characterize the percolation threshold in the continuum, which is usually credited as the beginning of the field of continuum percolation, was the 1970 [108] work of Scher and Zallen (S&Z), who based their argument on the "behavior" of a regular lattice, such as the one shown in Fig. 3. There, the red (say, the conducting) spheres occupy lattice sites, with a probability p^s, as in the lattice models described in Sect. "Lattice Percolation". The green (insulating) spheres correspond to empty sites in the lattice. The local connectivity criterion here is simply the red spheres occupation of nearest neighbor sites, as in the lattice, but here, if we consider the spheres as objects, the connectivity between two adjacent spheres, can be looked upon, as defined by a *single point contact* between two such spheres. Hence, if the diameter of the hard core spheres is equal to the lattice spacing the connectivity of the network stays the same as in the corresponding lattice. We can consider then the fractional volume of the conducting (red) phase which is simply given by $\phi = p^s f$, where f is the filling factor of the system (the maximum fraction of the volume in a given lattice that can be filled by hard spheres of equal volume). Correspondingly, one can define the percolation threshold by the total occupied volume fraction of the conducting spheres, i. e. $\phi_c = p_c^s f$. It is obvious then that $\phi - \phi_c \propto p^s - p_c^s$, and that the critical behavior of a given property Ξ in this case will be determined by

$$\Xi \propto (\phi - \phi_c)^{\chi} , \tag{6}$$

where χ is the corresponding-lattice critical exponent. S&Z noticed that ϕ_c, unlike p_c^s (or p_c^b), is nearly a dimensional invariant, yielding values very close to 45 (area %) for all 2D lattices and very close to 16 (vol.%) for all 3D lattices. This brought them to the important realization that if we "shake" the system shown in Fig. 3 to yield a disordered system of touching spheres, its topology will not be significantly modified, and following the above close invariance, both ϕ and ϕ_c will be as good parameters for the characterization of the disordered arrangement of spheres, as p^s and p_c^s are for lattices. The latter system can be also envisioned as a result of randomly dropping red and green spheres "gravitationally" into a container so that the "gravitation" secures the (single point) contacts between adjacent spheres [97]. Following the above concluded topology, it was of no surprise then that the latter approach confirmed the ideas and the results of S&Z.

We note of course that the above systems is by no means general or a prototype of a continuum system.

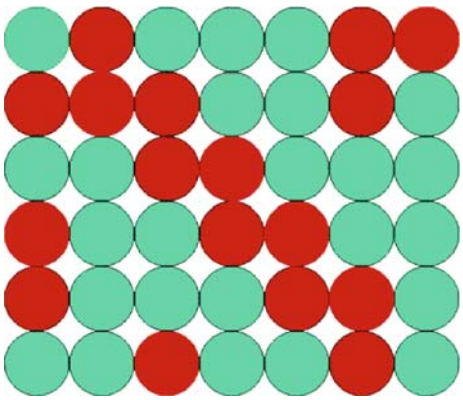

Continuum Percolation, Figure 3
A small portion of a square lattice that is comprised by "conducting" (red) and "insulating" (green) circles. The local connectivity is defined here by the "single point" contact of two red circles, and the global connectivity is determined by the spanning cluster of touching red circles

Rather, it represent an example of simple mapping of a continuum problem onto a lattice problem. While *the use of the fractional volume as a parameter*, which is equivalent in a way to p^s in lattices, was an important step in launching a theory for continuum percolation, their ϕ_c values were misused later, and to this date, for the characterization of continuum systems that are very different than the ones for which these, very particular values (secured single-point contacts), were calculated. For example, consider a system of red spheres that are embedded randomly in a continuous insulating matrix. Such a system represents many common composites (see Sect. "The Critical Behavior of the Dynamical Properties in the Continuum"). It can be easily appreciated that $\phi_c = 16$ vol.% in such a system of hard spheres yields a very sparse composite [1] and that the metallic (touching grain) conduction will be set at a much higher vol.% (of the order of 50 vol.% in the granular metals that constitute in fact an IRV system; Fonseca and Balberg [61]). Moreover, in practice there is no random system in which the "contact" or the "bond" between non permeable elements is "automatically" provided (if at all, it is approached only at the very high close packing limit of $\phi > 64$ vol.% [136,138]. Hence, there is no resemblance between the abundant common composite (with a continuous insulating matrix) and the models of S&Z and Powell, and thus the value of $\phi_c = 16$ vol.% is by no means a general approximate-invariant. This problem was discussed in more detail by Balberg and Binenbaum in 1987 [17].

In 1971 [112] Shante and Kirkpatrick (S&K) approached the percolation problem from the opposite end, i. e., considering permeable objects as in the case shown in Fig. 2. Like S&Z, it was obvious to S&K that the most promising approach to continuum systems was to try and map them onto a lattice problem, for which the theory was quite developed in the 1970's. They took what may be considered the most important single step in correlating lattice percolation and continuum percolation, proving that, at least in a specific systems, the topology of the two can be mapped onto each other. Their approach took into account an earlier empirical observation of Dalton, Domb and Sykes in 1964 [46] that in lattices, where z (the coordination number of the sites) approaches infinity, the value of zp_c^s extrapolates to 4.5 in 2D and to 2.8 in 3D. The quantity zp_c^s is, however, the average number of near-neighbors that are connected to a given site and it can be considered then as the number of occupied "bonds" per site. Hence, the quantity zp_c^s can be interpreted simply as the number of the occupied neighbor sites of a given site at the onset of percolation in the $z \to \infty$ limit. On the other hand, if one considers a system of permeable objects one can "implant" them anywhere in space. In particular for permeable spheres, the density of the possible centers of spheres which overlap a given sphere is also unlimited.

Following this observation, S&K conjectured that the topology of the $z \to \infty$ lattice and the continuum of permeable spheres is the same at the onset of percolation, i. e., that the number of "occupied" bonds per site, or the number of overlapping spheres for a given sphere, will be the same. Topologically, then, these two systems are expected to have the same onset of global connectivity. To test this idea, S&K considered the fact that the probability of randomly choosing a geometrical point that will be out of a given sphere of a volume v, in a unit size system, is $(1 - v)$. The probability that this point will be out of all the ρ_c spheres in this unit volume system, at the percolation threshold, is $(1 - v)^{\rho_c}$. In a large system that is statistically sufficient (i. e., $v \ll 1$ but $\rho_c \to \infty$) we can write the latter quantity as $e^{-v\rho_c}$. On the other hand, the probability of a point to be in any of the spheres (at the onset of percolation) is simply ϕ_c, the fractional volume that is covered by the ρ_c spheres. Hence, $\phi_c = 1 - e^{-v\rho_c}$. For a permeable sphere of radius a, in order not to overlap "another" sphere, its center has to be at a distance of $2a$ from the center of the "first" sphere. Correspondingly, the volume in which the center of the second sphere is not "allowed" to be (if no overlap is permitted) is not v but, rather, $8v$, or for hyperspheres in a D dimensional system it is, $v2^D$. On the other hand, the average number of centers of spheres that overlap a given sphere is simply $\rho_c v2^D$. We have then that the quantity $\rho_c v2^D$, is the critical number of bonds per sphere (or center, or site), B_c. Applying now their above

conjecture, i. e. that $B_c = zp_c^s(z \to \infty)$ one finds that

$$\phi_{Dc} = 1 - \exp(-B_c/2^D) , \quad (7)$$

for a system of hyperspheres of dimension D.

Using the above mentioned "lattice" values of B_c for $D = 2$ and $D = 3$, with the prediction of Eq. (7), S&K found an excellent agreement with the Monte Carlo findings of $\phi_{2c} = 68$ area % and $\phi_{3c} = 29$ vol.%. In fact the B_c values were confirmed directly for permeable spheres later (in 1987 [16]) by Balberg and Binenbaum. The most important step of S&K was then the recognition of the fact that the number of bonds per site is the proper topological parameter to describe the connectivity in continuum as well as in lattice systems. In passing we note that following the above argument for any system of parallel aligned "regular objects" (such as boxes or cylinders) the relation $B_c = \rho_c v 2^D$ is expected to be obeyed [21,115].

In 1984 [22] Balberg et al. realized that the ideas of S&K can be generalized to non spherical objects, and even to objects that have no "volumes" of their own. Such is the case of two dimensional "widthless sticks" [20,25], where the local connectivity criterion is the intersection of two sticks, and the onset of percolation does not require any area of the sticks. This is in contrast with the systems mentioned so far. The way this generalization was conducted [9] followed the procedure of S&K with the topological expectation that for any system (regardless of the particular objects) B_c is the meaningful general topological parameter, and as such, it is expected to be a finite invariant. Following the B_c values in lattices of various dimensions [136], their values are intuitively expected to be between 1 and 5. In particular, B_c is always the number of the "other" objects centers (or sites) within the "excluded volume" V_{ex} of a given object, where the latter is the volume in which the centers of two objects must be in order for them to overlap. As such, the quantity V_{ex} is well defined for any permeable (or partially permeable object, see below). This yields that, in general, $B_c = V_{ex}\rho_c$. Naturally, the simple relation we had above for spheres Eq. (7) can be generalized now to all dimensions and for all types of permeable objects of a finite volume. This is done by noting that we can write ϕ_c by using, instead of $v\rho_c$ as above, the relation, $v\rho_c V_{ex}/V_{ex} = vB_c/V_{ex}$ for an object of hypervolume v and a hyper-excluded volume, V_{ex}, thus yielding that

$$\phi_c = 1 - e^{-B_c(v/V_{ex})} . \quad (8)$$

In 1986 [9] Balberg noted that Eq. (8) has far-reaching consequences, since it says that we can have a connected system with very minute content of the "percolating" or "conducting" phase ($v\rho_c$ or ϕ_c) involved. This follows from the fact, seen in Eq. (8), that ϕ_c can be as small as desired provided that v/V_{ex} is very as small as requireds. Indeed, this result was able to explain the $\phi_c \to 0$ values in porous systems, such as sedimentary rocks [7,110] and geological cracks [58], as well as in cermets such as cellular composites [94,128] in which the conducting phase is arranged along lines or planes [101]. Moreover, the fact that for lattices we have always finite p_c^s (or p_c^b) values that are larger than 0.3 for 2D and larger than 0.1 for 3D, made this result, of ϕ_c values as small as desired, quite surprising, and demonstrated that the mapping of the continuum onto lattices is not as simple as might have been conjectured from the pioneering works of S&Z and S&K (who were concerned only with systems of spheres).

The above generalization is also very useful for systems with distributions of object sizes, shapes and orientations, when a proper determination of the average excluded volume of the system, $\langle V_{ex}\rangle$, is carried out [22,53]. Hence, the above $B_c = \rho_c V_{ex}$ relation can be generalized to:

$$B_c = \rho_c \langle V_{ex}\rangle . \quad (9)$$

This result is very significant and important from the point of view of applications since if B_c is nearly an invariant (as to be expected topologically in view of the above, i. e. its value changes only in a very limited range, see also Sect. "The Origin of the Different $\mathbf{B}_c$ Values in Systems of Different Objects" below) for various objects, one can predict the variation of ρ_c with the variation of the $\langle V_{ex}\rangle$ values in a group of objects. Hence, by assuming a constant B_c, trends in the behavior of percolation thresholds, ρ_c, as a function of the object's parameters, can be evaluated readily [9,37,83,103]. Moreover, one can predict the actual vol.% of the conducting phase (ϕ_c or $\rho_c v$) by just knowing the values of B_c and $\langle V_{ex}\rangle$ that can be well approximated by models (see below).

It turns out that the application of the concepts exhibited by Eqs. (8) and (9), beyond the very simple solution of $V_{ex} = 2^D v$ suggested for spheres (and for parallel convex shaped objects by Pike and Seager [95], and by Skal and Shklovskii, in 1974 [115]), has enabled the application of the above considerations to very many continuum systems. This was done in particular by the use and/or modification of the most celebrated model of non trivial systems, i. e., that of the three dimensional continuum system of capped cylinders [22]. These capped cylinders (of length L, radius W/2 and two capped spheres of radius W/2) span objects from spheres to cylinders with a single "close to an invariant" (see Sect. "The Origin of the Different B_c Values in Systems of Different Objects"), value of B_c. The basis of this model has been applied recently by Berham

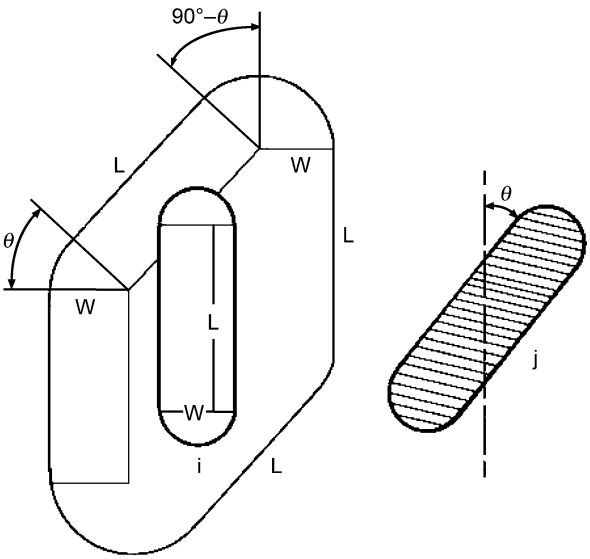

Continuum Percolation, Figure 4
The volume and the excluded volume of two adjacent capped cylinders. The latter is for two cylinders (*i* and *j*) the angle between which is θ. (From [22])

and Sastry in 2007 [30] even to rather non-regular "wavy" shaped objects. The system of capped cylinders became then the prototype of continuum percolation systems of "non trivial" (non spherical) permeable and non-permeable objects. The volume of this capped cylinder is simply $v = (4\pi/3)(W/2)^3 + \pi(W/2)^2 L$, but its average excluded volume, shown in Fig. 4, is given by

$$\langle V_{\mathrm{ex}} \rangle = (32\pi/3)(W/2)^3 + 8(W/2)^2 L + 4(W/2)L^2 \langle \sin\theta \rangle \,, \tag{10}$$

where θ is the angle between the two of possible intersecting cylinders, and the average is over the particular distribution of the θ values [22,89]. The "non trivial" aspect is manifested by the fact that $\langle V_{\mathrm{ex}} \rangle$ is not proportional to v, and that in the large aspect ratio case (i. e. $L \gg W$) we have that $v/\langle V_{\mathrm{ex}} \rangle \propto W/L$, while for the small aspect ratio case we recover the above mentioned, trivial $v/V_{\mathrm{ex}} = 1/8$ result for spheres. The system described here can be shown to represent very many natural and artificial systems where the objects are elongated [38] or have a negligible volume [101]. We note also that another system that is very helpful for the description of various systems is that of disks with a radius a, and a thickness t. These disks have a volume $v = \pi a^2 t$, while their excluded volume, in the isotropic orientation case, is given by $V_{\mathrm{ex}} = \pi^2 a^3$ [37]. It is apparent that in the thin disk limit ($t \ll a$), we get that $v/V_{\mathrm{ex}} \propto t/a$, explaining the very small ϕ_c values observed, as mentioned above, in sedimentary rocks, in systems of geological cracks and in cellular composites. In passing we note that the low dimensionality of the conducting phase inclusions in these systems explains the correlation between low percolation thresholds and the high critical conductivity-exponents (that are associated with low dimensional tunneling systems, see Sect. "The Average Resistance in the Case of Tunneling-Percolation").

Following the above, let us try and guess now the possible variations of B_c as one goes from systems that consist of permeable (soft-core) spheres to quasi-permeable spheres with a hard core of radius b and a wrapping soft shell with a thickness of $a - b$. We mentioned already that B_c is 2.8 in the 3D soft core limit. We also know then, as found directly [17], that the value of B_c decreases, in the transition with the increase b, towards a value of 1.5. What we have available empirically are, those limits, the fact that what influences the behavior are volume quantities, and the above expectation that B_c will decrease between the two limits as $b/a \to 0$. Hence, we can simply expect empirically that

$$B_c = 1.5 + 1.3\left[1 - (b/a)^3\right] \,. \tag{11}$$

Considering the above mentioned topological meaning of B_c we recall that for lattices we realized already that in the $z \to \infty$ limit, $B_c = 2.8$, and we note that for the other limit we can apply the conjecture of Ziman in 1979 [138] (following the fact that $z p_c^b$ is an invariant in lattices such that $z p_c^b = 1.5$ in 3D) that the smallest z (existing or imaginary) needed for percolation in 3D lattices, is 1.5. In this case we expect then that the approximation:

$$B_c = 1.5 + 1.3\left[1 - (1.5/z)\right] \,, \tag{12}$$

will be reasonable for lattices. If the connecting topology in the continuum and in lattices is the same (as found to be the case in the $b/a \to 0$ limit by S&K), B_c should scale equally well with $(b/a)^3$ or $1.5/z$. Indeed, as is clearly seen in Fig. 5, Monte Carlo simulations on a series of systems that represent the "soft core" to the "hard core" transition [17] have confirmed this mapping of the continuum system onto the lattice system. This establishes that, indeed, the underlying connectivity is the same in lattices and in continuum systems of partially permeable spheres (i. e., beyond the soft core S&K limit). A reproduction of the corresponding B_c values themselves can be achieved and explained phenomenologically as we will see in the following section. We noted already that from the practical point of view, one can generally predict the percolation threshold ρ_c once one knows the value of B_c (or an approximate of it) and the value of V_{ex}. In the present case we have the value of B_c (Eq. (11)) and

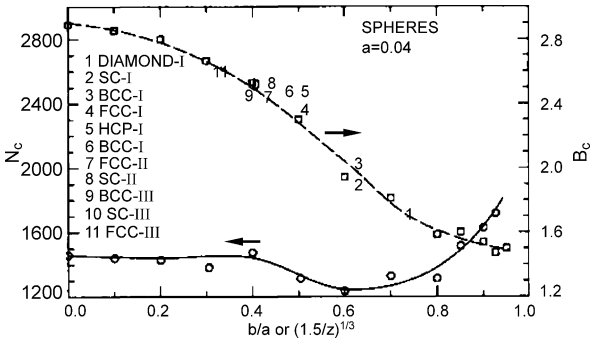

Continuum Percolation, Figure 5

Monte Carlo results for the dependence of the critical number of spheres in the system, N_c (denoted by circles, solid curve) and B_c (denoted by squares, dashed curves) on the spheres' b/a ratio in a continuum system of semi-permeable spheres. Also shown is the dependence of B_c on $(1.5/z)^{1.3}$ for three dimensional lattice systems. (From [16])

$V_{ex} = (32\pi/3)(a^3 - b^3)$. Hence, as the hard core limit is approached, and B_c gets the value of 1.5, we expect that ρ_c ($= N_c$ for a system of a unit volume) will diverge (of course up to the close packing limit that is possible). This is well exhibited by the Monte Carlo determination of N_c that is shown in Fig. 5. We note that the above derived phenomenological behavior does not tell us why the value of B_c itself decreases in the $b/a \to 1$ and $z \to 1.5$ limits (see Sect. "The Origin of the Different B_c Values in Systems of Different Objects").

A somewhat different view of the hard core-soft shell problem leads us to examine continuum percolation properties that can be considered to be even more remote from those of lattice systems. A conspicuous case of such a property is the effect of objects interaction due to a potential associated with the objects [41,90]. To introduce this scenario, we can consider the above hard core-soft shell sphere as being composed of a region of infinite repulsive interaction at distances r (from its center) smaller than $2b$, and a shell of zero potential for $r > 2b$. We may further look at the problem as a one in which the density of the objects is not uniform around an object (here e. g. $\rho = 0$ for $r < 2b$ and $\rho = \rho_s$ for $r > 2b$). If we use then the corresponding radial distribution function $g_R(r)$ of the particles (or objects' centers), the average number of bonds per site will be given by $B = \langle\rho\rangle \int_0^{2a} g_R(r) \mathrm{d}^3 r$, where $\langle\rho\rangle$ is the average density of objects in the whole system and a is the radius of the soft shell (that determines the local connectivity of the system). If we have a more complicated potential (i. e. when $g_R(r)$ has values different from 0 or 1 as above), B, and thus B_c, will depend on the parameters of this distribution, such as on the interaction potential that determines the $\rho(r)$ dependence [4]. A case that was considered in detail and is relevant, for example, to microemulsions [69], is that of an attractive potential shell that "wraps" the hard core ($b \le r \le (1+\lambda)b$) and has a constant, non-zero value, in the interval of λb. In that case, it has been shown first by Bug et al. [35] (and later by Drory, Balberg and Berkowitz [54,55]) that interactions can raise or lower the percolation thresholds, i. e., the value of the global ρ_c value, according to the details of the magnitude of the potential and the ratios between b, λ and a.

Physically, this competition is not too surprising, since the attractive potential will cause the increase of the density of objects in the attractive potential region, but for a globally given ρ, it will cause a reduction of the density outside this region. The two effects will be manifested by an increased probability for the formation of the connected clusters, on the one hand, but by the reduction of the connectivity between different clusters, on the other hand. The interesting observations are then the non-trivial results of the interplay between these two effects that is determined by the system parameters, i. e. by the dependence of the values of B_c and ρ_c on them.

From the point of view of continuum percolation theory, the important conclusion is that even in this explicit interaction case we can characterize the system by its B_c value, correlate it with system parameters, and apparently generalize it to systems where the mapping onto lattice models is, conceptually and/or computationally, not obvious. On the other hand, it is apparent that such derivations can account for general connectivity problems where ρ or other "bonding strength" parameters are not uniform in space or in the network.

The Generalized Thresholds and the Critical Behavior in the Continuum

Let us turn now to see how the above mentioned quantities play a role in the determination of the most important parameter in the study of the critical behavior, i. e. the proximity to the percolation threshold. Following the fact that $B_c = \rho_c \langle V_{ex} \rangle$, we simply have the relation $B - B_c \propto \rho - \rho_c$. Hence, for B very close to B_c and ϕ very close to ϕ_c (which is the interesting "critical" range), we can further generalize this relation (by using Eq. (8)) to

$$\phi - \phi_c \propto B - B_c \propto \rho - \rho_c \,. \tag{13}$$

This relation is a very important one from the practical point of view, since ϕ is the only parameter that one can determine readily in experimental studies. However, the requirement for small $\phi - \phi_c$ values must be kept then in

mind when the critical behavior is studied in continuum systems. In fact, another very important precaution has to be taken into account when studying the critical behavior as a function of $\phi - \phi_c$. We saw above that in many (e. g. sedimentary rocks and cellular composites) systems, ϕ_c can be very small. Hence, a small $\phi - \phi_c$ (with respect to unity) does not actually tell us about the proximity to the threshold. The parameter that one should consider is a normalized proximity to the threshold, such as $(\phi - \phi_c)/\phi_c$. We note that the latter problem is not severe in lattices of the "practical" ($D = 2$ or $D = 3$) dimensions because p_c^s and p_c^b for all these lattices are larger than 0.12 [136]. On the other hand, as noted by Grimaldi and Balberg in 2006 [70], the overlooking of the above two precautions is one of the reasons that when considering experimental results, improper comparisons with lattice-like critical behavior have been suggested. Following the above discussion on the soft core-hard core transition we also note that in the small $B - B_c$ limit, Eq. (13) covers all systems, including those where ϕ represents te hard core fractional volume in partially permeable (hard-core/soft-shell) objects. This is easy to appreciate since at the close vicinity of the percolation threshold the ratio of the hard core volume to the entire (hard core and soft shell) volume can be considered a constant and since we know that the relation given by Eq. (13) holds in limit as shown by the S&Z case. Indeed, many computer simulations on many continuum systems [15,65,66], where the systems are well defined (and ϕ_c can be approached much more closely than in the experimental studies on "real" systems), have revealed the critical behavior of lattices. This is quite an important realization when one tries to explain the critical behavior in composites, considering the fact that the rigorous development that we described above was given for permeable objects.

From our above discussion we can expect then *the same critical behavior* for all types of objects, and may further conjecture that this is the case when other local connectivity criteria are applied. In fact, thus far all Monte Carlo simulations have confirmed these expectations. We note, however, that a high precision study of ϕ_c and the critical exponents of a system of permeable spheres, was claimed to indicate a small variations with respect to the corresponding lattice values of the same exponents [100]. This conclusion can be contrasted by the fact that rigorous exact mapping of the permeable circles system onto the fine grid lattice system was demonstrated by McCarthy already in 1987 [86]. As of now, however, further work will be needed in order to determine whether the above small deviation is a matter of accuracy or proximity to ϕ_c (see the discussion following Eq. (13) above), or if there is still something more fundamental here that has not been satisfactorily studied thus far.

The Origin of the Different B_c Values in Systems of Different Objects

So far we have taken B_c to be an empirical parameter, as are the p_c^s and p_c^b quantities in lattice percolation, and the application of the relation given by Eq. (9) was proven extremely useful for the determination of trends in the variation of measurable percolation thresholds (i. e., ρ_c) when a constant (system independent, for a class of materials, e. g. boxes with a different aspect ratio) B_c value was assumed. However, the various values of B_c, while being confined to the 1–5 range (see above) are not the same for systems of objects of different shapes. These differences are important to understand since, as we discussed above, this is the only parameter that accounts for the global topological connectivity of percolation systems in the continuum. We note in passing that while the B_c values can be derived rigorously (see Sect. "Rigorous Determination of the Percolation Threshold"), their relation to the particular objects in the system under consideration is not a priori transparent from the corresponding derivations. At this point we concentrate then on the reasons for the different B_c values for different "types" of objects. For example, B_c *is 2.8 for permeable spheres and is only 2.6 for parallel permeable cubes*. While this difference seems minute (in comparison with the larger differences to be mentioned below), it is a convenient case for the illustration of the concept of "pointedness" that enables to explain the observed differences in the B_c values.

The reason for the above mentioned variations in the B_c values had hardly been considered prior to the work of Alon, Balberg, and Drory in 1991 [4] that yielded an insight into the apparent "different topology" associated with different objects. They derived this understanding using a heuristic or "semi-rigorous" argument, which will be illustrated here by the different B_c values for spheres and cubes, as well as by the variation of B_c in the soft core to hard core transition.

Let us consider then a system of permeable spheres of radius r_0 and a system of parallel permeable cubes with an edge a_0. Choosing their size such that they have the same excluded volume $\langle V_{ex} \rangle$ (in this case, also the same volume), we note that $\langle V_{ex} \rangle = (4\pi/3)(2r_0)^3 = (2a_0)^3$. Now let us consider the maximum distance between the centers of two spheres that can overlap. This distance is $2r_0$, and thus the possible distances, l, between the centers of two partially overlapping spheres is in the range $0 \le l \le 2r_0$. In contrast, for three dimensional cubes that are also con-

nected by a partial overlap (and have the same $\langle V_{ex} \rangle$), the corresponding range is $0 \leq l \leq \sqrt{3}a_0$, i. e., $0 \leq l \leq 2.7r_0$. Hence, since the density of the centers of the objects is uniform in both cases, on the average, two partially overlapping cubes can "on the average span" a larger distance than that of partially overlapping spheres. Correspondingly, in order to span an "infinite" cluster, we will need a smaller number of cubes than spheres, i. e., the value of ρ_c will be smaller in the case of cubes compared with the case of spheres. Since we have chosen V_{ex} to be the same in the above two systems, the corresponding value of ρ_c and thus the value of $B_c (= \rho_c V_{ex})$ will be lower in the case of the cubes. Of course, the exact ratio between the B_c values of the two systems cannot be derived from the qualitative-simple illustration given here, but a more refined argument has been shown to heuristically yield such ratios exactly [4].

Since the apparent difference between the cube and the sphere is that the former has corners ("points"), one can say that the "pointedness" (that was defined by Alon, Balberg and Drory accurately) brings about the lower value of B_c. We can correspondingly conclude that since the average "covered" l "per excluded volume" will be larger with a higher aspect ratio of elongated objects, such as boxes that are isotropically distributed in space, we can further expect that for elongated boxes, of the same $\langle V_{ex} \rangle$, the value of ρ_c will decrease with this ratio. For these, as mentioned above, the corresponding trend of the decrease of B_c with the aspect ratio was confirmed both rigorously and computationally [53,131] to decrease from about 2.7 to 1.2 for aspect ratios between 1 and 500. The latter trend was confirmed also for other types of permeable elongated objects [30,103].

A similar argument can apply for spheres of a hard core of radius b and a soft shell with a wrapping thickness of $a - b$. The "local span" of the two objects' in the $b = 0$ case is in the $0 \leq l \leq 2a$ range, while for the hard-core sphere with a soft shell it is $2b \leq l \leq 2a$, so that the average l value is larger in the latter case. Correspondingly, for the same average ρ in the system, the "span" of larger-hard core spheres will be larger, or the value of ρ_c will be smaller in the hard core-soft shell case, suggesting the decrease of B_c from 2.8 to 1.5, as we saw already in Fig. 5. The above result is of great conceptual importance, since it shows that while the phenomenological theory of the excluded volume theory was developed for permeable objects, *all consequences regarding the trends* in the values of the percolation thresholds (such as the effect of the aspect ratio) *apply also upon the approach to the hard core limit*. This can be illustrated simply by replacing W by $W - 2b$, where b is the hard core radius, in Eq. (10), and noting that as the aspect ratio increases the last term will eventually become the dominant term for any $2b/W$ ratio. In fact from the above arguments concerning the aspect ratio and the soft core to hard core transition, we conclude that the two effects can be compounded. This is also important from the practical point of view, since one can be guided, for example, for "real" composites that are composed of hard core objects [19], by the many works and the numerous simulations [21] that were concerned with the corresponding (easier to study) systems of permeable objects. A case in point is the decrease of the percolation threshold in carbon-black-polymer composites as the conducting particles become elongated (or, as better known, of "higher structure" [10,11,102]).

Rigorous Determination of the Percolation Threshold

Thus far we have considered the derivation of the concepts and the empirical trends in the behavior of the percolation thresholds as a function of the correlation between, and the properties of, the individual objects, as well as the corresponding global averages such as the anisotropy of a system. However, as we saw above, the particular fundamental values of the thresholds were either derived from lattice quasi-invariants (notably $p_c^s z$ for $z \to \infty$), heuristic arguments or computer simulations. Noting that a "complete" theory with a firm basis requires a rigorous derivation of the fundamental quantities from first principles, it is always a challenge to achieve such a derivation and the theoretical justification for the experimental and computational observations and their trends. This challenge was recognized also for the problem at hand, i. e., the rigorous finding of the "fundamental-topological" percolation threshold B_c (e. g., $B_c = 4.5$ for circles and $B_c = 2.8$ for spheres), which we encountered in the previous sections.

The above challenge was appreciated in 1977 [73] when Haan and Zwanzig considered the existence of bonds, defined by the overlap of permeable objects, and applied a graph theory for the analogous lattice-bond percolation problem. This attempt, which involved a lengthy cluster enumeration and high-order overlap integrals, has not been followed however by others, probably because of the latter reasons. The breakthrough in trying to find a rigorous route which is physically more transparent can be attributed then to the work of Coniglio, DeAngelis and Forlani in that year#160;[44]. They applied for the above purpose the formalism of the connectedness functions that is well known in the theory of liquids [74]. For the sake of brevity we will describe here their approach by the following simple "intuitive" arguments.

If we define a local (direct) connectedness criterion (e. g., overlap between permeable spheres) and a global connectedness (e. g., being on the same cluster), we can derive a direct connectedness function $C(\boldsymbol{x}, \boldsymbol{y})$, such that if $\boldsymbol{r}$ is the vector $\boldsymbol{x} - \boldsymbol{y}$ connecting "directly" the two objects at $\boldsymbol{x}$ and at $\boldsymbol{y}$, $\rho(\boldsymbol{r})C(\boldsymbol{r})\mathrm{d}^3 r$ is the probability that given a particle at $\boldsymbol{r} = \boldsymbol{0}$, there will be a particle, in the volume $\mathrm{d}^3 r$ around $\boldsymbol{r}$, that is locally connected to the particle at $\boldsymbol{r} = \boldsymbol{0}$. Here, $\rho(\boldsymbol{r})$ is the particles' (or objects') density in the system. The quantity $C(\boldsymbol{x}, \boldsymbol{y})$ accounts, in addition to the direct connection, for all possible additional "indirect" connections between the "directly connected" these two particles, and thus it can be presented by corresponding "closed diagrams" [56]. Similarly, one can define a "total" pair connectedness function $g(\boldsymbol{z}, \boldsymbol{y})$ between two objects (say, at $\boldsymbol{z}$ and at $\boldsymbol{y}$) which involves all the connecting (direct and indirect) routes between them. The function $g(\boldsymbol{r})$ (where $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{z}$) can be defined then by the probability, $\rho(\boldsymbol{r})g(\boldsymbol{r})\mathrm{d}^3 r$, of finding a particle in the volume element $\mathrm{d}^3 r$ around $\boldsymbol{r}$, which is connected (i. e., belonging to the same cluster) to the particle that is given to exist at $\boldsymbol{r} = \boldsymbol{0}$. Since $g(\boldsymbol{x}, \boldsymbol{y})$ is composed of "direct" connectivity steps such as $C(\boldsymbol{x}, \boldsymbol{z})$ and $C(\boldsymbol{z}, \boldsymbol{y})$, the "total" $g(\boldsymbol{x}, \boldsymbol{y})$ function can be expressed by the corresponding (Ornstein-Zernike) "chain" rule

$$g(\mathbf{x}, \mathbf{y}) = C(\mathbf{x}, \mathbf{y}) + \int \rho(\mathbf{z})C(\mathbf{x}, \mathbf{z})g(\mathbf{z}, \mathbf{y})\mathrm{d}^3 z \,, \tag{14}$$

where $\rho(\boldsymbol{z})$ is the density of particles (or centers of objects) in a volume of $\mathrm{d}^3 z$ around $\boldsymbol{z}$. For simplicity we will assume in the following that $\rho(\mathbf{z}) = \rho$, where ρ is a constant-uniform density in our problem. The crucial step that connects relation (14) with the present continuum percolation problem, i. e. with the average cluster size, is the consideration of its Fourier transform. For a "wave vector" $\boldsymbol{k}$ this relation yields that for the transformed functions $C(\boldsymbol{k})$ and $g(\boldsymbol{k})$ we will have that $g(\mathbf{k}) = C(\mathbf{k})/(1 - \rho C(\mathbf{k}))$. If we are interested in the integral $\int g(\boldsymbol{r})\mathrm{d}^3 r$, i. e. in the Fourier transform $g(\boldsymbol{0}) \equiv g(\boldsymbol{k} = \boldsymbol{0})$, we get in particular that

$$g(\boldsymbol{0}) = C(\boldsymbol{0})/(1 - \rho C(\boldsymbol{0})) \,. \tag{15}$$

The new important step made in the work of Coniglio, DeAngelis and Forlani in 1977 [44] was that, as can be appreciated intuitively, the average cluster size S in the system is simply given by

$$S = 1 + \int \rho g(\boldsymbol{r})\mathrm{d}^3 r \,, \tag{16}$$

where the unity here represents the given initial object (with a center at, say, $\boldsymbol{r} = \boldsymbol{0}$), $\rho\mathrm{d}^3 r$ is the probability of finding an object center in the space element $\mathrm{d}^3 r$, and $g(\boldsymbol{r})$ is essentially the percolation correlation function [121], i. e., the probability that, if there are objects centers at $\boldsymbol{r} = \boldsymbol{0}$ and $\boldsymbol{r}$, they will be on the same cluster. From Eq. (16) we have that $\rho g(\boldsymbol{0}) = S - 1$ and thus using Eq. (15), we obtain that

$$S = 1/(1 - \rho C(\boldsymbol{0})) \,. \tag{17}$$

Hence, knowing that at the threshold $S \to \infty$, we get that $\rho_c = 1/C(\boldsymbol{0})$, and thus, in principle, "all that is left" is to calculate $C(\boldsymbol{0})$.

There are basically two approaches for the calculation of $C(\boldsymbol{0})$. The first one is immediately called for by Eq. (14), from which it is apparent that if another relation between $g(\boldsymbol{x}, \boldsymbol{y})$ and $C(\boldsymbol{x}, \boldsymbol{y})$ is known, one can solve both functions and then find the value of S. Such a relation can be derived from corresponding approximations. One of these approximations, known as the Percus–Yevick closure [74]), was used for the solution of the problem at hand. Since the application of this closure was of a limited success [49]), and since it would require a deviation from the simple outline of our main argument regarding the rigorous derivation of ρ_c, it will not be detailed here.

The conceptually simpler and physically more transparent approach to the problem can be attributed to the suggestion of Bug, Safran and Webman in 1985 [34]. They were able to determine the initial terms in a (diagrammatic) series expansion of the form $C(\boldsymbol{0}) = \sum c_n \rho^n$, and to relate the initial coefficients to the concept of the excluded volume. This already enabled them to derive the trends in the behavior of the thresholds. However, this use, of what one may call an order-by-order (or series) approach, did not yield (in that work) actual values for the thresholds. The full utilization of this order-by-order approach came in the work of Drory and coworkers in 1990 [3], 1991 [53] and 1997 [56]. In passing we note that Drory also got in 1996 [51] and 1997 [52] a firm basis for the conjecture that continuum percolation is a phase transition, as proven previously [62] to be the case of lattice percolation.

Turning to the utilization of the latter approach we noticed already that in principle, $C(\boldsymbol{0}) = 1/\rho_c$. However, in practice, this does not provide directly an accurate enough value for ρ_c for the rather short series that were obtained thus far. On the other hand a more successful, albeit biased, approximation has yielded, as mentioned below, very accurate results. For the description of the latter we start then from Eq. (17) using the well understood relation that we had considered before, i. e., replacing ρ by $B = \rho V_{\mathrm{ex}}$. Hence, the aim of the procedure is to determine the, by now, well understood and well-characterized parameter B_c. Following the above, we can write S in terms

of a power series of the form $S = \sum a_n B^n$, which is similar to the form used for the derivation of S by a power series of p in lattice percolation [121] In the present case, having the quantity $C(\mathbf{0})$ expressed as a power series of the form $C(\mathbf{0}) = \sum b_n B^n$ enables the comparison of the coefficients on both sides of Eq. (17), yielding then relations between the coefficients a_n and b_n. Here, for example, $a_0 = 1$, $b_0 = 0$, $a_1 = b_1$, $a_2 = b_2 + b_1 a_1$, $a_3 = b_3 + b_2 a_1 + b_1 a_2$, and $a_4 = b_4 + b_3 a_1 + b_2 a_2 + b_1 a_3$. Having these relations and implementing the fact (see Eq. (1)) that S has a critical behavior of the form

$$S \propto |B - B_c|^{-\gamma} , \tag{18}$$

where γ is the well-established critical exponent for lattices and, as expected from universality, also for the continuum (see above), one can fit the series of S to the behavior described by Eq. (18), in order to find the value of B_c. Of course the "only" input needed for the implementation of this procedure is then the determination of, as many b_n coefficients as possible and a good extrapolation of S. However, the derivation of the coefficients b_n is not trivial and becomes more and more complicated with increasing n [53]. So far these values of b_n have been determined up to $n = 5$ for permeable hyperspheres, yielding, however, very accurate values for B_c that are only a few % off the Monte Carlo estimates for corresponding $D = 2$ to $D = 6$ systems [3]. Similar successes for systems of elongated boxes [53] and hard core-soft shell cubes [56] have been already obtained. In spite of those achievements, further basic and computational developments in the study of this subject are called for. These will be outlined in Sect. "Future Directions".

The Critical Behavior of the Dynamical Properties in the Continuum

As we pointed out above, within the limited framework of this review, it will not be possible to consider all the many various dynamical properties (mentioned in Sect. "Definition of the Subject") that are determined by the added locally variable features in continuum systems and compare their behaviors with those derived for lattices. Correspondingly, we will confine our discussion to the conceptually most conspicuous representative, which is also the most studied, dynamical property, i. e., the electrical conductivity. It can be shown, as done by Rubin et al. in 1999 [102] for the electrical noise, that the basic arguments associated with the critical behaviors of the other properties are similar in principle, and that the specific details of those behaviors can be derived along the same lines that are used here for the electrical conductivity.

In the previous sections, we saw that one property that makes continuum percolation an area of a much wider diversity and application, in comparison with what one finds in lattice percolation, is the percolation threshold. The global properties of a percolation system were own there to depend on the details of the systems, where these may vary in many respects from one system to another. On the other hand, as far as the critical behavior is concerned, we have seen that the global behaviors of the geometrical-statistical properties are expected to be exactly the same as those found in lattices, provided that one uses the topological concept of bonds per site. In those cases a simple mapping, which is intuitively straightforward, of the "percolating phase" content onto the lattice occupation probability, can be made. Following the fact that for a given lattice there is a constant ratio between p^b and p^s [136], we will simply use here p and p_c as a generalized "conducting phase" content and describe the critical behavior of the dynamical properties in the continuum using these parameters. When needed however, the relation between these and the commonly measurable quantities, such as the corresponding partial volume contents of the conducting phase, τ and τ_c, will be emphasized.

Does the above behavior of the geometrical-statistical properties give us a hint as to the critical behavior of the dynamic (or "physical") properties of the system? To answer this question, let us consider the two possible features that are fixed in lattices, and may be random (or follow another distribution) in the continuum. This is easy to do by reexamining the system shown in Fig. 2 as a system of permeable spheres. In this figure we can see that the continuum randomness is manifested by the positions of the centers of the spheres and the random-like degrees of overlap between two adjacent spheres. If these spheres are envisioned as pores filled with a conducting fluid, the latter feature determines the local resistance between a pair of them. This will yield a distribution in the resistance values of the actual resistors that connect pairs of pores in the system. The simplest approach to evaluate the contribution of the latter effect (that has no counterpart in lattice percolation) to the global conductance of the system is to assign a resistor, with a resistance value that is taken from the corresponding distribution, to the occupied bonds on a given lattice [111]. A more direct simulation of the system that involves also the random "implantation" of the spheres has also been shown [23] to yield the predicted results that will be given below. The purpose of the following discussion is then to show how corresponding distributions of the local resistance values can guide us to the determination of the critical behavior in "real" continuum systems.

The simplest system to realize then is the lattice system of Fig. 1, when one assumes that the "bond" or the conducting "bar" between two occupied nearest neighbor sites corresponds to a resistor, the value of which is taken from a given distribution. A priori, one does not foresee too much of a surprise here, and a "universal" (lattice-like) behavior of the conductivity is expected, except that the individual local resistance value, $\mathbf{r}_0$ (associated with R_ξ; see Eq. (3)), will be replaced now by an average value of the local resistors, $\langle \mathbf{r} \rangle$. In fact, the finding of a universal value for t in many "real" systems [2,19,25] suggested that this distribution-effect will not yield a critical behavior that is different from the behavior in lattices, as appears to be the case for the geometrical-statistical properties. Considering this problem Kogut and Straley (K&S) developed in 1979 [82] a simple "toy" model that suggested that for some distributions, the above universality will hold, while for others, it will not. It took six years, however, until the missing link of the "toy" distribution and "real", or continuum, systems was made by Halperin, Feng and Sen in 1985 [72], and attention was given finally to the results of K&S. The work of K&S, which can be considered as the cornerstone of the field of "non-universal" behavior of dynamical properties in the continuum, will be reviewed below, after a modified-generalized account of the basic physics that is involved in their model will be given.

What apparently took the above-mentioned six years in order to understand, even in principle, the emerging numerous experimental observations of non-universal t values in the 1970's and 1980's (i. e., $t \neq t_{\mathrm{un}}$; [10,11,94]), and to appreciate the theory of K&S, was *the unsatisfied need to find a distribution function* of the resistors values in a given "real", natural or artificial, system. This breakthrough came then with the work of Halperin, Feng and Sen in 1985 [72], who were able to show that in sedimentary rocks the values of the local resistors are essentially distributed according to the "toy model" function of K&S. In fact, it will be emphasized all along this part of the review that the determination of the resistors distribution function is the crucial step for the understanding of the critical behavior of the electrical conductivity (or other dynamical properties) in a given continuum system. The basic physics of this key issue, which has not been reviewed before, will be described below in some detail.

The Basic Physics of the Non-universal Behavior of the Conductivity

The physics of the critical behavior of the global resistance in continuum systems can be described as follows. Suppose we have a distribution $f(g)$ of the values of the individual conductors (the resistors that occupy the bonds) in our infinite lattice. Let us assume that the system has a fractional bond occupation p and that the percolation threshold is p_c such that $p > p_c$.

We have seen that in lattice percolation (Sect. "Lattice Percolation") the resistance of the system can be written as

$$\mathbf{R}_L = \mathbf{r}_0 A_R L_1^{\zeta} (L/\xi)^{2-D} \,, \tag{19}$$

where A_R is some constant (i. e., ξ, or $p - p_c$, independent) parameter that accounts for the local and global geometry of the system and $\mathbf{r}_0$ is the resistance value of the individual occupied bond in the system. The question of interest here is then the behavior of $\mathbf{R}_L$, when we have a distribution of the individual resistance values (as expected in continuum systems) rather than a single $\mathbf{r}_0$ value. Let us denote the conductance of the individual resistor by g, so that the local value of the resistance is given by g^{-1}. Now suppose that the g values are taken from a distribution $f(g)$ that is confined to the range $g_1 \leq g \leq g_2$. Let us also assume that in our system there is a largest possible g_2 value (e. g., a zero distance tunneling or a complete overlap of IRV spheres; see Sect. "The Local Resistors and Their Distributions", below). The most important assumption in the following consideration is however that there is no correlation between the location of the resistor in space and its value, and thus one would expect that if there is such a distribution, $\mathbf{r}_0$ in Eq. (19) should be simply replaced by $\langle \mathbf{r} \rangle$ where

$$\langle \mathbf{r} \rangle = \langle g^{-1} \rangle = \int_{g1}^{g2} g^{-1} f(g) \mathrm{d}g \,. \tag{20}$$

We note in passing that such an average is meaningful if it is taken on $1/g$ (rather than on g), since it is the series connection of resistors that determines the resistance of the link (see Eq. (3)).

In principle, g_1 can be any value in the interval $0 < g_1 \leq g_2$ (a $g_1 = 0$ value corresponds to an unoccupied bond), and it is apparent that for a well defined finite-constant $g_1 > 0$ value, the critical behavior described by Eq. (19), with $\langle \mathbf{r} \rangle$ replacing $\mathbf{r}_0$, will be the same as in Eqs. (3)–(5), i. e., the universal (lattice-like) critical behavior will be obtained. What if $g_1 \to 0$ (but $g_1 \neq 0$; see above)? In this case it is apparent that the result will depend on $f(g)$. If $f(g)$ decreases as $g \to 0$, the value of the average $\langle \mathbf{r} \rangle$, as defined by Eq. (20), will be finite, as above. On the other hand, if $f(g)$ is a constant or it increases towards $g = 0$ (i. e. $f(g)$ diverges at $g = 0$), the value of $\langle \mathbf{r} \rangle$ will be ∞. In other words, if there are "enough" $g \to 0$

conductors in the system, the average will be determined by them and we will also get that $\mathbf{R}_L \to \infty$.

To consider however the behavior of $\mathbf{R}_L$ in more detail (i. e. the way in which $\langle \mathbf{r} \rangle$ diverges) for a non-decreasing $f(g)$ distribution as $g \to 0$, we use the following estimate of $\langle \mathbf{r} \rangle$ for the $g_1 \to 0$ case as follows. Let us list the values of the local conductances in the system in a descending order. Now, electrical conductance in the system is only possible if $p > p_c$, and thus we can choose the fraction (or subset), p_c of the p local conductors in the system, that contains the "top" g, p_c-values from that list. Following our basic assumption that there is no correlation between the location of the bond and its attached g value, we can use the fact that any randomly selected subset of p_c conductors is, by definition, a percolating cluster, and conclude that the above chosen p_c subset of the top value conductors constitute a percolation cluster. Of course, the largest conductance value in the system, g_2, is also the largest possible g value in the so chosen subset. For convenience let us choose now the value of g_2, by the relation $F(g_2) = 1$, where $F(g)$ is the indefinite integral of the function $f(g)$. This is done without loss of generality since all other g values in the system can be normalized accordingly. Then, the lowest g value of the conductors in the above subset p_c of p has the corresponding normalized value g_c. Mathematically, this g_c value is given then by

$$p \left[\int_{gc}^{g2} f(g) \mathrm{d}g \right] = p_c \,. \tag{21}$$

This yields that $F(g_c) = (p - p_c)/p$, and thus that $g_c = F^{-1}[(p - p_c)/p]$, where F^{-1} is the inverse function of F. We have then a very significant result that connects the conductors' values in the system with the proximity to the percolation threshold. In particular, it is the nature of the F^{-1} function that determines the "pace" at which $g_c \to 0$, as p is made to become closer to p_c. The very important observation here is that for $p = p_c$ we have to exhaust all the resistors in the system and thus $g_c = 0$. The question that we would like to resolve is, following the variation of g_c as p approaches p_c, how will $\langle \mathbf{r} \rangle$ change as a consequence of this variation.

As above, we can distinguish here between three cases, but now we can also get some physical insight as to the corresponding critical behaviors. If $f(g)$ is a constant (in the $0 \le g \le 1$ interval) g_c will approach 0 at exactly the same "pace" as $p - p_c$. If $f(g)$ increases as $g \to 0$ (i. e. $f(g)$ diverges as $g \to 0$), g_c will approach $g = 0$ "faster" than $p - p_c$, and if $f(g)$ decreases toward $g = 0$, g_c will approach $g = 0$ at a "slower pace" than p approaches p_c. In other words, $1/g_c$ which is a measure of the resistors that determine the value of $\langle \mathbf{r} \rangle$ (see below), will not diverge in the latter case when $p \to p_c$. On the other hand, if $1/g_c$ diverges as $p \to p_c$ there is a possibility that a dynamic property that depends on g will also diverge then. Correspondingly then the above $p - p_c$ dependence of $1/g_c$ on $p - p_c$ dependence of $\langle \mathbf{r} \rangle$ may contribute to a non universal behavior of $\mathbf{R}_L$.

In order to derive a quantitative determination of the critical behavior of $\mathbf{R}_L$ let us try now to evaluate the behavior of $\langle \mathbf{r} \rangle$. Above, we have selected a subset of the conductors in the system (i. e. a subset of the resistors that participate in the conduction process). The average resistance of this subset is

$$\langle \mathbf{r_c} \rangle = \int_{gc}^{g2} g^{-1} f(g) \mathrm{d}g \,. \tag{22}$$

Any other subset of p_c resistors will have an $\langle \mathbf{r} \rangle$ which is larger than $\langle \mathbf{r_c} \rangle$ (or equal to it), and thus $\langle \mathbf{r_c} \rangle$ yields the lowest estimate of $\langle \mathbf{r} \rangle$ that can be inserted in Eq. (19) instead of the $\mathbf{r}_0$ value in Eq. (19). Considering the fact that we are concerned with $p \to p_c$ and that $g_c \to 0$, the contribution of the very low g-values (with values of $g < g_c$) to the "measured" network resistance becomes negligible as $g_c \to 0$. This is in particular so if the values of g are changing by orders of magnitude as $g \to 0$ [87]. Hence, this $\langle \mathbf{r_c} \rangle$ of the "bypassing" network becomes a "more and more" accurate estimate for the value of $\langle \mathbf{r} \rangle$ in the $g_c \to 0$ limit, and consequently a reasonable replacement for $\mathbf{r}_0$ in Eq. (19).

Let us try and get a physical feeling for the correspondingly expected critical behavior, i. e., for the dependence of $\mathbf{R}_L$ on $(p - p_c)$. For that purpose we follow the dependence of $\langle \mathbf{r_c} \rangle$ on $(p - p_c)$ by considering the behavior of $1/g_c$ when $p \to p_c$. The above mentioned "list" of conductors, as p is made to decrease (i. e., some of these occupied bonds are eliminated as p_c is approached), we have to go "down the list" of the conductors in the system, in order to choose p_c bonds from the new p, so that the smallest g_c involved is smaller (or at most equal) in comparison with the g_c value of the larger (i. e. "former") p. Of course, it is not enough that g_c will just become smaller as $p \to p_c$, since if the availability of these low-value g_c conductors becomes scarce as $p \to p_c$, they will not contribute significantly to the network (or to the average $\langle \mathbf{r_c} \rangle$) and there will be an effective constant $\langle \mathbf{r_c} \rangle$ value (which is independent of $p - p_c$) that will determine the resistance $\langle \mathbf{r} \rangle$ and thus the global resistance of the system (see examples below, and in the following sections). If this is not the case, the effect of $\langle \mathbf{r} \rangle$ on $\mathbf{R}_L$ is clear then; not only changes of the

network connectivity (Eq. (19)) take place as $p \to p_c$, but also, the average resistor value that contributes to the conduction can diverge, yielding an additional $p - p_c$ divergence to the "universal" contribution of Eq. (19). Hence, a compounded, non-universal (or non lattice-like) behavior can be obtained.

For a more quantitative illustration of the above discussion let us consider the simple case of $f(g) = 1$ in the interval $0 \leftarrow g \leq 1$. In this case we have from (Eq. (22)) that $\langle \mathbf{r_c} \rangle \propto \log(1/g_c)$ and from (Eq. (21)) that $g_c \propto p - p_c$. Hence, substituting $\langle \mathbf{r_c} \rangle$ instead of $\mathbf{r}_0$ in Eq. (19) will provide a logarithmic divergence to $\mathbf{R}_L$, as $p \to p_c$, in addition to the power law divergence of the global or the "lattice-like" contribution. It is also apparent then that it is enough that when $f(g)$ increases towards $g = 0$ (i. e., $f(g)$ diverges as $g \to 0$), a stronger (say, power law) divergence of $\langle \mathbf{r_c} \rangle$ with $p - p_c$ will be obtained. On the other hand, there will be no divergence of $\langle \mathbf{r_c} \rangle$ if $f(g)$ decreases (in this case $f(g)$ diminishes) towards $g = 0$. In the latter case we have that the approach of p to p_c is associated with the "disappearance" of the large $(1/g)$ resistors, while in the former, this approach is associated with the "appreciable" availability of these resistors ("at the expense" of the smaller $(1/g)$ resistors). From the above analysis we see that the crossover between the diverging and non diverging $\langle \mathbf{r_c} \rangle$, as $p \to p_c$, occurs when $f(g)$ "overcomes" the increase in the "summation" of the resistors, as reflected by the $(\mathrm{d}g/g)$ term in Eq. (22). This is easy to comprehend in the LNB picture as a competition between the possible introduction of large $(1/g)$ resistors in the links (the $f(g)$ effect), and the rate of increase in the "weighted" values of the "needed" individual resistors in the link (the $\mathrm{d}g/g$ effect) as the links become longer (i. e. as $p \to p_c$). This scenario of the "need to supply enough high-resistance resistors" in order to get the divergence of $\langle \mathbf{r_c} \rangle$ as $p \to p_c$, will be important as we turn to consider "real" continuum systems. Physically, we found here an important principle of the $\langle \mathbf{r_c} \rangle$ behavior, the appearance of non universal behavior will be determined by the competition between the $f(g)$ and the $\mathrm{d}g/g$ effects. One notes is passing that for other dynamical properties (that are described by functions of g other that $1/g$) the "$\mathrm{d}g/g$" effect will be changed accordingly [42].

In "real" systems, however, the $f(g)$ distributions may be more complicated than described above, yielding behaviors that were not accounted for in the literature until very recently [70,71]. To illustrate such a distribution and to summarize the above discussion, let us consider an $f(g)$ function that is peaked at some value g_m and diminishes towards both, $g = 1$ and $g = 0$. Such a distribution from $g_2 = 0.4$ to $g_1 = 0$ is illustrated in Fig. 6. As p is de-

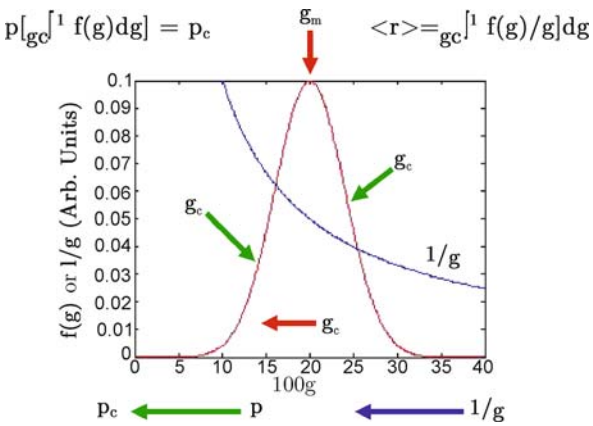

Continuum Percolation, Figure 6

An illustration of a peaked distribution of the local conductances in a continuum system, as well as the variation of the participating g-values as $g \to 0$. The relation between the approach of g_c to the $g \to 0$ limit and the corresponding direction of the $p - p_c$ variations is indicated by the arrows in the figure. The principal equations that determine g_c and the approximated average resistor in the sample $\langle r \rangle$ are also given

creased towards p_c, the g_c values "move first" along the $f(g)$ part where $f(g)$ increases faster than $1/g$. Over this range, i. e. $0.4 \leq g_c \leq g_m$, we have (see above) the conditions that yield a diverging-like behavior of $\langle \mathbf{r_c} \rangle$. Hence, as the decrease of p is associated with the decrease of g_c, we will get an apparent non universal behavior. Continuing with the decrease of p to the p-range for which $f_(g)$ decreases with $1/g$ (i. e. the $0 \leftarrow g \leq g_m$ range), the apparent behavior of $\langle \mathbf{r_c} \rangle$ will become of the type that does not have a divergence and will thus yield a universal-like behavior of the system in the corresponding p-range. We note that while, strictly speaking, the critical behavior (i. e., at the $p \to p_c$ limit) is universal for this $f(g)$, in "practice", i. e., when the p values are somewhat removed from p_c, the behavior observed (i. e. for $g > g_m$) may be characterized as non universal. In fact, considering the increase in the "pace" in which $f(g)$ varies with $1/g$ in the $g > g_m$ regime, one would expect for the large g (or large p) values, a behavior that is close to the universal behavior, and then that this behavior will turn to a more pronounced non universal-like behavior as g and p are decreasing (but still for $g > g_m$). This pronounced behavior will be weakened however, as g_m will be approached, until upon the decrease of $f(g)$ for $g < g_m$ a universal behavior is obtained. Hence, a peak in the "local" t value is expected for some p (i. e. at the inflection point of $f(g)$ in the $g_m < g < 0.4$ interval) and a transition from a non universal-like to a universal behavior will be obtained at some lower p. Examining the various experimental data in the literature [128] that

suggested a non universal behavior for cases where $f(g)$ is expected to be of the form shown in Fig. 6, one must conclude that the p value that corresponds to the g_m peak lies below the p values-limit available in the acquisition of the experimental data. In fact, the range of the g values between g_2 and g_m usually consists of a few orders of magnitude in the variation of g [70] and thus the experimental data that is fitted to a range of a few orders of magnitude in $\mathbf{R}_L$, while appearing to represent a single "non universal" t value, does not disclose the variation in the value of t as $p \to p_c$, due to the relatively limited range of the latter (see also Sect. "The Average Resistance in the Case of Tunneling-Percolation" below).

The Model of Kogut and Straley and Beyond

As we have shown above, the crucial factor in the determination of the critical behavior of an average dynamical property is the variation of its extreme value as $p - p_c \to 0$. K&S have proposed a particular distribution of the local conductances in a system which can exhibit the three basic behaviors that we mentioned above. This distribution (which as we show below turned out to be very useful for the description of best known real systems) yielded then analytic expressions for the diverging or non diverging $\langle \mathbf{r_c} \rangle$ values as $p \to p_c$. The distribution that they suggested was

$$f(g) = 1/(1-\alpha) g^{-\alpha} , \tag{23}$$

which is normalizable for $\alpha < 1$, in the range $1 > g > 0$. For this distribution one gets from Eq. (21) that

$$g_c = \left[(p - p_c)/p\right]^{1/(1-\alpha)} \tag{24}$$

and from Eq. (22), that

$$\langle \mathbf{r_c} \rangle = \left[(1-\alpha)/\alpha\right]\left[g_c^{-\alpha} - 1\right] . \tag{25}$$

We see here the three behaviors described above as follows. If $\alpha = 0$ we have that $f(g) = 1$ as discussed above. If $\alpha < 0$, $\langle \mathbf{r_c} \rangle$ is not diverging as $p \to p_c$ and it is the constant $(1-\alpha)/(-\alpha)$, i. e., the system behaves in the $g \to 0$ limit as if it is made of local resistors all having the latter value. If $\alpha > 0$, we get that in the interesting regime of $p \to p_c$ (i. e., where $g_c \to 0$),

$$\langle \mathbf{r_c} \rangle \approx \left[(1-\alpha)/\alpha\right] g_c^{-\alpha} \propto \left[(p - p_c)\right]^{-\alpha/(1-\alpha)} . \tag{26}$$

We recall that the resistance of the sample is approximated by $\mathbf{R}_L \propto \langle \mathbf{r_c} \rangle (p - p_c)^{-t_{un}}$, where t_{un} describes the effect of the connectivity of the global network of the system (Eqs. (3)–(5)). Correspondingly, the critical behavior of the electrical conductivity of the system, which determines the critical exponent t (see Eq. (4)), will be given now by

$$t = t_{un} + \alpha/(1-\alpha) . \tag{27}$$

In what follows we will show that the mapping of the most conspicuous systems in the continuum onto the K&S distribution is possible and that it is associated with the corresponding local geometrical parameters. We will also note the very important case where this mapping is not simple and other approaches must be taken in order to determine $\langle \mathbf{r} \rangle$ and thus the $\mathbf{R}_L$ dependence on $p - p_c$, as the percolation threshold is approached. For example, the application of the Effective Medium Approximation (EMA) by Grimaldi and Balberg in 2006 [70] yielded the confirmation of the rather more complicated behavior that was discussed above in relation to Fig. 6, on the one hand, and which explained the moderate $t - t_{un}$ (that are in the 1–10 range) values that were obtained in numerous composites [128], on the other hand.

The Local Resistors and Their Distributions

As pointed out above, the conductors' distribution, as given by K&S was considered to be quite abstract until this resistors value distribution was suggested for some real systems by Halperin, Feng and Sen in 1985 [72]. Once this distribution is given, its mapping onto the K&S distribution is, as shown below [12], quite straight forward. Another way to derive the corresponding critical behavior was based on the use of a physically more transparent and more rigorous approach [59]. The corresponding findings were the very crucial steps that have finally explained the universal and non universal behaviors of "real" systems. In passing, we note that a posteriori, the critical behavior estimated by the above consideration of $\langle \mathbf{r_c} \rangle$, is also confirmed by the latter approach, as will be shown below.

Since both approaches have been given in detail in the literature, we will only outline here the principal steps in their utilization for the determination of the corresponding t values. Starting then with the first approach, let us assume that the resistance of a local resistor in the system, $1/g$, is determined by a single local geometrical parameter, ε, and that the distribution of the ε values in the system, $h(\varepsilon)$, is known. The distribution function of the local conductor values, $f(g)$ is simply given then by [10]

$$f(g) = h(\varepsilon)(\mathrm{d}\varepsilon/\mathrm{d}g) . \tag{28}$$

"All one has to find" then, from the local configuration that occur in the system, are the $h(\varepsilon)$ function (which is usually not easy to do; see below) and the function $\varepsilon(g)$ (which is usually easy to do).

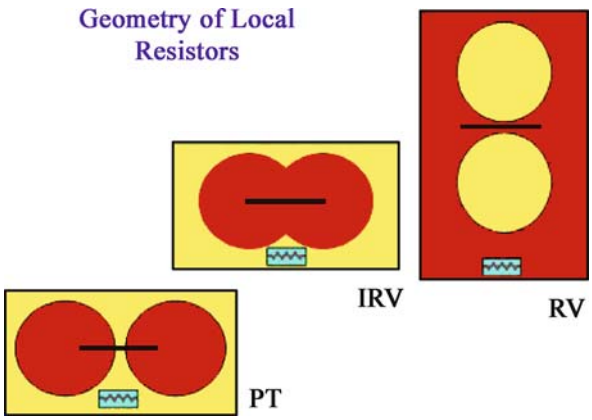

Continuum Percolation, Figure 7
The three major configurations of "real" systems in the continuum. The red color represents the conducting phase, and the yellow color represents the insulating phase in the systems. The bars and the resistors represent the "necks" and the corresponding dominant local resistors in the system. The configurations represent, from right to left, the random void, the inverted random void and the percolation tunneling models

There are three principal local configurations that were studied in detail. Let us discuss these configurations and consider the single geometrical parameter, ε, that describes them. These three configurations are illustrated in Fig. 7, where the "conducting" phase is colored in red and the "insulating" phase is colored in yellow. Starting from the random void (RV, sedimentary rock-like system, or a metallic bulk in which holes were punched out), that is also known as the Swiss cheese-like system, we can assume that the distance between the surfaces of the spherical pores is ε, and there is a "neck" of volume $(2\sqrt{b\varepsilon})C_D\varepsilon^{D-1}$ between them. Here, b is the radius of the pore and C_D is a constant (which is 1 in $D = 2$ and $\pi/4$ in $D = 3$). This neck determines the resistance of the neck that is associated with the two adjacent spheres [72] since beyond this "neck" the local resistance in the system is relatively small. The value of the local resistor in the system is well approximated then by the resistance of the "neck" which is given by $\rho_0(2\sqrt{b\varepsilon})/C_D\varepsilon^{D-1}$, where ρ_0 is the resistivity of the conducting medium. Hence, the resistance of the local resistor is varied with ε as $1/g \propto \varepsilon^{3/2-D}$ (i. e., the local conductance can be expressed by $g \propto \varepsilon^{D-3/2}$).

In the second configuration, known as the inverted random void (IRV) system, such as a granular metal above the onset of metallic conductivity [61], or a system of pores filled with a conducting liquid [32,59] the overlap between two spheres can be characterized by the parameter $\varepsilon = 2b - r$, where r is the distance between the centers of the adjacent spheres. Correspondingly, the neck for conduction here is approximated to be $2\sqrt{\varepsilon b}$ long, and it has a cross section of $C_D(b\varepsilon)^{(D-1)/2}$ ($C_D = 2$ in $D = 2$ and $C_D = \pi$ in $D = 3$), yielding that $g \propto \varepsilon^{D/2-1}$. For the above two cases we have then that $g \propto \varepsilon^m$, where $m = D - 3/2$ in the RV model and $m = D/2 - 1$ in the IRV model.

The third case is that of conduction between two spheres that do not overlap for which the distance between their centers is r and the distance between their surfaces is $r - 2b$. In that case, the conduction between the two spheres is by tunneling and the charge transfer probability between the surfaces is expected to decrease as $\exp[-(r - 2b)/2d]$, where $2d$ is the effective (say, the WKB) tunneling decay coefficient. This percolation-tunneling (PT) case will be considered in Sects. "The Tunneling Percolation Problem", "The Average Resistance in the Case of Tunneling-Percolation".

Let us turn now to the corresponding $h(\varepsilon)$ and $f(g)$ distributions in the first two cases. Noting that the derivation, theoretically or experimentally, of these distributions is the most difficult part (see above and below) in our problem, one can consider the first suggestion of a "real" $h(\varepsilon)$ dependence, and its use in the present context, to be the breakthrough that lead to the understanding of the non universal behavior of the dynamic properties in "real" systems. This was achieved, as pointed out already, in the 1985 [72] work of Halperin Feng and Sen, who showed that in porous media $h(\varepsilon)$ is a constant, h_0, for $\varepsilon \to 0$. The corresponding distribution is, however, normalizable since $h(\varepsilon)$ tapers off with increasing ε, above a certain value of ε (as is obvious from the finite size of the system, see below). Being interested in the smallest ε values, i. e., the largest resistors in the system, one can easily apply the "recipe" given by Eq. (28). Using then the general $g \propto \varepsilon^m$ relation, where m is the relevant exponent, we have that $f(g) \propto \mathrm{d}g/\mathrm{d}\varepsilon \propto h_0 g^{(1-m)/m}$ and thus, that the α value of K&S is $1 - 1/m$. In the above RV case, for which $m = D - 3/2$, we have then that $t - t_{\mathrm{un}} = \alpha/(1 - \alpha) = m - 1 = D - 5/2$. Hence, from the consideration of K&S, it is easy to conclude that for $D = 2$ a universal behavior will be found, while for $D = 3$, a value of $t - t_{\mathrm{un}} = 1/2$ is predicted, as confirmed indeed by the corresponding simulations of Sen, Roberts and Halperin in 1985 [111].

The above procedure was generalized by Balberg in 1998 [12] for any power law distribution of the $h(\varepsilon) \propto \varepsilon^{-\omega}$ type which, following the above, yields that $f(g) \propto g^{(1-m-\omega)/m}$ and correspondingly (see Eq. (27)) that $t - t_{\mathrm{un}} = (m + \omega - 1)/(1 - \omega)$. Of course we note that we must have that $\omega < 1$, so that the distribution is normalizable. For the three dimensional RV case we will generally get then that $t - t_{\mathrm{un}} = (D - 5/2 + \omega)/(1 - \omega)$. It is impor-

tant to note that if ω varies from one ε range to another, the behavior obtained will be more complicated. In particular, transitions between universal-like and non universal-like behaviors may be obtained as $p \to p_c$. Such a variation, with the consequences of a none monotonic $h(\varepsilon)$, and thus a non monotonic $f(g)$, has not been considered in the literature for this case. However, following the above discussion of Fig. 6, one can account for now such transitions. In fact, a similar situation in principle, was evaluated very recently for the percolation tunneling problem, that will be described in Sect. "The Average Resistance in the Case of Tunneling-Percolation". A mapping similar to that of the RV model was obtained for the IRV model [12,23] yielding that $\alpha \leq 0$ for $D \leq 4$, and thus the divergence of the conductance is obtained (for the $\omega = 0$ case) only for $D \geq 5$.

Let us derive now these results in a physically more transparent manner [59,72]. We start by assuming the LNB model and try to find the relation between the parameter ε, that is relevant to the value of the resistors, and the proximity parameter $p - p_c$, so that a relation such as that of the K&S model can be found between the average local resistance $\langle \mathbf{r_c} \rangle$ and $p - p_c$. For the $\varepsilon \to 0$ limit, which corresponds to the larger $1/g$ resistors (that must be included in the network when $p \to p_c$), we assume, as above, that $h(\varepsilon) = h_0$ and that $g \propto \varepsilon^{-m}$. The corresponding $\langle \mathbf{r_c} \rangle$ is determined then by

$$\langle \mathbf{r_c} \rangle \propto \int_{\delta}^{\Delta} h_0 \varepsilon^{-m} \, \mathrm{d}\varepsilon \,, \tag{29}$$

where Δ is the ε value above which $h(\varepsilon)$ tapers off with increasing ε, and δ is the smallest ε value (which is associated with the smallest g value), in a typical link, that is made of L_1 singly connected bonds. The probability that ε will be larger than a given δ for a bond in this link is simply $1 - \int_0^{\delta} h_0 \, \mathrm{d}\varepsilon$, and thus for all L_1 bonds in the link it is $(1 - \int_0^{\delta} h_0 \, \mathrm{d}\varepsilon)^{L1} = \exp(-\delta L_1 h_0)$. This means that at $\delta \approx 1/L_1 h_0$ this probability starts to become significantly smaller than unity, and thus this relation provides a good estimate of δ, i. e. that $\delta \propto 1/L_1$. Having this value and considering that $\delta \ll \Delta$ and that $m > 1$ (see above), we get from Eq. (29) that

$$\langle \mathbf{r_c} \rangle \propto h_0 \delta^{-m+1} \approx h_0^m L_1^{m-1} \propto (p_- p_c)^{1-m} \,. \tag{30}$$

This result yields, as above, (see Eqs. (26) and (27)) that $t = t_{\mathrm{un}} + (m - 1)$. As we saw for the RV model, $m = D - 3/2$, and thus for the corresponding three dimensional system we get $t = t_{\mathrm{un}} + 1/2$, in accordance with the above result that was obtained from the K&S model. Similarly, for the IRV model we got that $m = D/2 - 1$, and thus for $D = 3$, we have that $m < 1$ and $\alpha < 0$ yielding that $t = t_{\mathrm{un}}$. We can further attach the distribution term $\varepsilon^{-\omega}$ as above and get that instead of the relation of $\delta L_1 h_0 \approx 1$ we will have that $\delta^{(1-\omega)} L_1 h_0/(1-\omega) \approx 1$ and thus, that in the more general case, $\langle \mathbf{r_c} \rangle \propto h_0 \delta^{-m+1-\omega} \approx h_0^m L_1^{(\omega+m-1)/(1-\omega)} \propto (p_- p_c)^{-(\omega+m-1)/(1-\omega)}$ or $t = t_{\mathrm{un}} + (m + \omega - 1)/(1 - \omega)$, where ω is as above [12].

The Tunneling Percolation Problem

Before turning to the problem of the average value of the local resistor in the case of tunneling percolation, let us examine the unique relation between tunneling and percolation [13]. Tunneling is essentially the principal mechanism of electrical conduction in most classes of composite materials. This is to be distinguished from the cases that involve a coalescence of particles encounters one that in granular metals [2], porous media (the IRV model; see above) and microemulsions [69], where the onset of a continuous geometrical network is associated with the onset of a global "metallic" conductivity. In most other cases, such as in granular metals in the dielectric regime [1] and in various semiconductor composites [26,27], the conductivity between the particles (or objects) is generally controlled by tunneling.

As one considers the tunneling in the percolation-systems context, one notices that it challenges our above understanding of continuum percolation as follows. In lattices (S&Z-like), or in the above RV and IRV systems, the geometrical connectivity and the electrical connectivity are very clear and both are associated with the geometrical continuity of the conducting phase. This is also the case when we can define a soft shell that wraps the hard-core so that the overlaps of the soft shells provides a continuous conducting phase, as can be envisioned to be the case in microemulsions [35]. When the conductivity is provided only by tunneling, there is no geometrical connectivity (or continuity) but there is electrical connectivity. The latter connectivity, however, is problematic, since all the conducting objects are connected electrically to each other (albeit with a different "strength" that is determined by the exponential decay of the tunneling probability). Hence, the system has a "zero span" of the geometrical connectivity and an "infinite span" of the electrical connectivity. It appears then to be quite surprising [13] that systems corresponding to this "counter" percolation configuration, such as many types of composites, exhibit computational [24,78] or experimental [128] well-defined percolation thresholds with a universal or a non universal critical behavior. This problem was essentially clarified

only recently as to be described below. We note then in passing that until quite recently the very many experimental or computational studies on numerous composite systems that were discussed in terms of a bona-fide percolation critical behavior have done so with no justification.

The first explanation for the above paradox was given only very recently by Toker et al. in 2003 [124]. They concluded, by examining the fractal dimension of the "electrical" percolation network (that was derived from local probe microscopy on real composites), that if the tunneling decay parameter is very small compared with the size of the objects (say, spheres), in practice, only the network of adjacent neighbors (see above) determines the global measurable conductivity of the system. In that case the percolation threshold is simply associated with by the concentration of "near neighbor" objects, such as in the case of a system of objects of a "large" hard-core and a thin soft shell. In this case the percolation tunneling problem can be mapped onto such a model. The dynamical properties are determined then by the local resistors distribution (see also Sect. "The Average Resistance in the Case of Tunneling-Percolation") as in the above given discussion.

When this is not the case, i. e. when the tunneling distance is of the order of the size of the conducting objects, such as the case in granular metals [1,26,124] or quantum dot semiconductors [27] in the dielectric regime, the observation of a percolation threshold and a critical behavior is less obvious. It turns out that the solution to this problem can be based on the same model as in the above case if we take into account the accumulated information that we have on the radial distribution (RDF), $G(r)$, of a collection of hard spheres. This function that is defined [138] as the probability $G(r)\mathrm{d}r$ to find a particle center within the interval r and $r + \mathrm{d}r$ from a given particle, shows that there is an evolution of the peak of "nearly touching" particles as we go from the dilute, or small b, limit to the (hard core) dense spheres limit, i. e. as the particles size and/or density increase. In particular, the common conclusion of all the corresponding models is that the larger the particles phase content (i. e. the particles density and/or their size) the narrower the "tail" of the near neighbors separations. This is equivalent in the present context to the statement that more particles will overlap their neighbors within the narrow shell, d, that is of the order of the tunneling distance. In other words, the way the hard spheres arrange themselves yields a preferred conduction network that can be looked upon as made of an IRV system with the soft shell, the thickness of which is of the order of the tunneling decay constant. Hence, the model is much like the one proposed above for the $b >> d$ case. It is of course appreciated that as the system becomes very dilute it also becomes reminiscent of the near neighbor hopping model [114]. Indeed, a close examination of the dependence of the conductivity on the conducting phase content in granular metals [26] reveals a density regime in which both, percolation and hopping, can simultaneously account for the experimental data, while only a percolation model can account for the data obtained on the "large spheres" system of Carbon black-Polymer composites [102].

The Average Resistance in the Case of Tunneling-Percolation

The determination of the average local resistance in the tunneling-percolation problem was solved thus far accurately only for the 1D (or quasi 1D; see below) systems [10,11]. At higher dimensions, the lack of analytic expressions for the $f(g)$ distribution does not enable a simple evaluation of the corresponding critical behavior. However, with the information already available [5,126] it is possible, as described below, to obtain a good physical understanding and derive semi-quantitative results for the critical behavior [78].

We start then with the simple case of a linear chain of geometrically, but not electrically, isolated "metallic" spheres [71], the hard core radius of which is b and the average distance between which is $2a$ ($= 1/N_1$, where N_1 is the average number of spheres per unit length). The random distribution of the distances of the centers of the nearest neighbor spheres from a given sphere center in the corresponding 1D system is the well-known 1D Hertz distribution [10,11,126] that can be written as [26]

$$h_1(r) = \left[1/(2a-2b)\right]\exp\left[-(r-2b)/(2a-2b)\right], \quad (31)$$

where r is the distance from the center of the reference sphere. On the other hand, the local tunneling conductance between two such spheres is simply given by

$$g = g_0\exp\left[-(r-2b)/2d\right], \quad (32)$$

where g_0 is a corresponding geometrical-physical constant and $2d$ is the tunneling decay constant. For the calculation of the average value of the local resistance we will neglect the resistance between "non-adjacent" spheres, i. e., we assume that $b \gg d$. Applying Eqs. (28), (31) and (32), we get that for the nearest neighbor connections (or bonds; [10,11,26])

$$f_1(g) \propto \exp\left[(r-2b)/2d\right]\left[1-d/(a-b)\right] \propto g^{-\alpha}, \quad (33)$$

where,

$$\alpha = 1 - d/(a-b). \quad (34)$$

Following the above K&S prediction (Eq. (27)) for the exponent t we get then that

$$t - t_{\mathrm{un}} = (a - b)/d - 1\,. \tag{35}$$

Considering the discussion in Sect. "The Basic Physics of the Non-universal Behavior of the Conductivity", regarding Eq. (22), we see that if d is larger than $(a - b)$, there is not "enough" supply of distant neighbors, or large resistors ("the $f(g)$ effect"), to overcome the needed $1/g$ resistors ("the $\mathrm{d}g/g$ effect") and thus to yield the diverging $\langle \mathbf{r_c} \rangle$. In contrast, if $d < a - b$, even though the "supply" (Eq. (31)) decreases with increasing r (i. e. with decreasing g values), this "supply" is enough for the "fastly" increasing $1/g$ values, to yield an increase in $f(g)$ as $g \to 0$. This is an important observation that results from the fact that we have two terms in Eq. (28) so that the behavior of $f(g)$ is not necessarily the behavior expected by just examining $h_1(r)$. Physically then, as above, in the $d < a - b$ case the decrease in the available small g values (or the large r values) is slower than the increase in the value of the large resistors, $1/g$, as $r \to \infty$ or (Eq. (32)). Correspondingly, one obtains the divergence of $f(g)$ (Eq. (33)) as $g \to 0$ i. e., when p approaches p_c. This description is manifested quantitatively of course by the above mapping of the tunneling resistor problem onto the K&S model, with an α value that determines the transition between a non-diverging to a diverging distribution, i. e. $\alpha = 0$, which in the present case corresponds to $d/(a - b) = 1$ (Eq. (34)).

For completeness, let us present the physically more transparent procedure for the determination of $t - t_{\mathrm{un}}$, such as the one of Halperin, Feng and Sen in 1985 [72], which we outlined in Sect. "The Local Resistors and Their Distributions", but this time for the tunneling-percolation problem. This approach [102] is similar then to the one considered for the RV and IRV models, but for the present problem we assume that the largest distance between two adjacent spheres in the link is s (rather than the smallest distance, δ, in the RV and IRV problems). We are above interested then in the probability that the largest r in the link of L_1 singly connected resistors is smaller than some prechosen s. This probability (following Eq. (31) and the procedure given in Sect. "The Local Resistors and Their Distributions") is simply $\{1 - \exp[-(s - 2b)/(2a - 2b)]\}^{L1} = \exp\{-L_1 \exp[(s - 2b)/(2a - 2b)]\}$. Correspondingly, this probability starts to be small at a value of about $\exp(-1)$, and thus the condition for s being the largest "bond-length" in the link is that $L_1 \approx 1/\exp[(s - 2b)/(2a - 2b)]$. On the other hand, the average resistance in the link, $\langle \mathbf{r_c} \rangle$, is now proportional to $\int_{2b}^{s} \exp\{[(r - 2b)/(2a - 2b)] - [(r - 2b)/2d]\}\mathrm{d}r$. The dominant term in the integral is then $\exp\{[(s - 2b)/(2a - 2b)][1 - (a - b)/d]\}$, or in view of the above, the dominant term of $\langle \mathbf{r_c} \rangle$ is $L_1^{-[1-(a-b)/d]} \propto (p - p_c)^{1-[(a-b)/d]}$. Correspondingly, we get (as we already found above from our mapping of the problem onto the K&S distribution) that in this relatively simple 1D case, $t - t_{\mathrm{un}} = (a - b)/d - 1$.

This finding has been adopted not only for the relativle simple 3D cases where the above $h_1(r)$ distribution is still appropriate [71] but also, very recently, to general 3D random systems. The simpler cases include cellular composites in which the system can be assumed to be made of insulating (say, cubic) blocks, and of conducting channels with a diameter just wider than $2b$ which include metallic spheres of diameter $2b$. The channels are further assumed to coincide with the edges of the insulating cubes [71,78]. In practice, however, the comparison of the prediction of Eq. (35) with experimental observations on "real" cellular composites [128] shows that the above predictions yield t values that are very exaggerated. If one tries to provide then a "more realistic" model of such a composite, one must also assume that the chains are not infinite and that thier number per unit length, N_1, (and thus a) may vary from one chain to the other, according to the make of the composite. This is even to the degrees that $(a - b)/d$ may be larger or smaller than unity in different chains. The corresponding distribution of N_1 in the chains may affect then the critical behavior by yielding an effective "global" α. For this case that was considered in detail by Grimaldi et al. in 2003 [71], it is obvious that the chains' $N_1 (= 1/2a)$ values will contribute to a universal behavior (high N_1) or to a non universal (low N_1) behavior of the system. The global behavior will be determined then finally by the distribution of the various N_i's values in the system.

A priori one may think that this model of Grimaldi et al. should apply to systems of higher dimensions as one approaches the percolation threshold, since in the Links, Nodes, Blobs (LNB) model we consider a conducting network of such large chains that are, at first sight, one dimensional in nature. At second sight one notices two problems with that assumption. First, the distribution function of the nearest neighbor sites in the $D > 1$ case, $h_D(r)$, is different from the one given in Eq. (31) and, second, and more severe, the LNB link is made of singly connected bonds that are determined by the connectivity of the $D > 1$ network and not (necessarily) by the first nearest neighbors, as in the simple 1D case.

Before giving the recent "exact" solution to the problem [78] let us see the basic underlying physics of it. The global conductance as $p \to p_c$ can be concluded semi-quantitavely by noting that for the very large $(r - 2b)$ values (that are relevant to the critical behavior), the $h_3(r - 2b)$ distribution function is expected to decrease

as $\exp[-(r/2a)^3]$ for all neighbors [5]. Here, a is the average (see below) distance between neighboring sphere centers in the above 3D system. Hence, "finally" (i. e., for the largest r's or smallest g's) the "supply" of the g's (i. e. the function $h_3(r-2b)$) will not meet the "demand" of Eq. (32). This situation is different then from the situation described above (Eq. (33)) for the 1D case of $d < a - b$, for any value of $d/(a-b)$. Correspondingly, as demonstrated below, for these large r's, $f(g)$ will be a decreasing function of $1/g$. Strictly speaking, for the general 3D system the "ultimate" (i. e. at the $p \to p_c$ limit) the critical behavior will be universal as discussed in Sect. "The Basic Physics of the Non-universal Behavior of the Conductivity". However, the latter limit is rather experiamentally [70] and so far computationallay [78] a non accessible (see below) and thus there is a great interest in the critical behavior of the system on route to this limit. To consider the latter behavior let us use the 1990 [126] result of Torquato, Lu and Robinstein for the 3D distribution function of the inter-particle distances of the first nearest neighbors, $h_3(r)$. The latter result can guide us in predicting the behavior of the more general $D > 1$ systems.

The relevant dominant term that as implied by their $h_3(r)$ is

$$H_3(r-2b) \propto A(\rho)(r-2b)^2 \exp\left\{-\left[8A(\rho)/b^3\right]\left[r^3-(2b)^3\right]\right\} , \quad (36)$$

where, $A(\rho) = [\rho v(1+\rho v)/(1-\rho v)^3]$, v is the hard core spheres volume $(4\pi/3)b^3$ in a system of a unit volume and ρ is their concentration (i. e., per unit volume of the sample). This concentration is related to the above $2a(> 2b)$ parameter (i. e. the average distance between the sphere-centers of two nearest neighbors) that, we simply define it here by $\int_{2b}^{\infty} h_3(r)rdr$ [70]. A very rough, but geometrically more transparent, estimate of a (which is enough for the present discussion of the basic physics of the problem) can be given by assuming that the spheres are implanted randomly (as in the $a \gg b$ case). In this case a is associated with the average "territory" of a sphere, and is thus given by the relation

$$(4\pi/3)a^3\rho = 1 . \quad (37)$$

Returning to Eq. (36), we can follow the behavior of $f(g)$ by applying Eqs. (28), (32) and (36) in a way that yielded Eq. (33). The first feature that we encounter in $H_3(r-2b)$, and which has no counter part in $h_1(r-2b)$, is a peak at some r_m that is of the order of $2(a-b)$. For $r < r_m$ both $H_3(r-2b)$ and the resistance term $1/g \propto \exp[(r-2b)/2d]$ increase with increasing r and thus $f(g)$ increases with $1/g$. The rate of this increase may be very large for the small $r-2b$ values due to the dominance of the $(r-2b)^2$ term in Eq. (36) that can be translated to a $g^{-\alpha}$ behavior with $\alpha > 1$ for a limited g range, in some cases [24]. Of course, this strong dependence will be weakened for larger r's (smaller g's) as follows. Since $H_3(r-2b)$ stops increasing at r_m it is obvious that the maximum rate of the increase of $f(g)$ with $1/g$ will be obtained at some intermediate value r_p that is smaller than r_m. This yields then a corresponding g_p value (Eq. 32) such that $g_p > g_m$. However, the increase of $f(g)$ with $1/g$ does not necessarily stop at g_m. To see that let us now consider the $r > r_m$ regime. The dominant term of $H_3(r-2b)$ there is the exponential decrease of this function and thus, using the procedure suggested by Eqs. (33) and (36), we can express $f(g)$ now as $f(g) \propto g^{-\alpha(r)}$ where now we have a "local α" that is given by:

$$\alpha(r) = 1 - \left\{2d\left[8A(\rho)/b^3\right]\left[r^2+br+(2b)^2\right]\right\} . \quad (38)$$

Noting that r is a measure of $1/g$ we have here then, unlike the constant α of the 1D case, a "locally" varying $\alpha(r)$. Hence, for the $r > r_m$ regime we will have that for small (in comparison with $2a-2b$), $r-2b$ values, $\alpha(r) > 0$, i. e. $f(g)$ will still be increasing with increasing $1/g$. For larger r's however, $\alpha(r)$ will "eventually" become negative and thus $f(g)$ will decrease with $1/g$. Denoting the r value, at which $\alpha(r) = 0$, by r_t, we can summarize the behavior of $f(g)$ as follows. For small $1/g$ values $f(g)$ increases with g, obtaining a maximum rate of this increase at some g_p which is larger than g_m. Then, at some g_t (such that $g_t < g_m$) $f(g)$ will decrease as a function of $1/g$. Following now the premise of our model (as described in Sect. "The Basic Physics of the Non-universal Behavior of the Conductivity") i. e. the increase of $1/g_c$ with the decrease of $p-p_c$) we can now predict the critical behavior of the global conductance of the system and in particular the corresponding critical exponent as p approaches p_c. For p far above p_c, $\langle \mathbf{r_c} \rangle$ will have a diverging-like behavior with $p-p_c$ and thus a non universal value will be obtained for t. The value of this non universal t will reach its maximum at a p value, p_p, and then, for still smaller p values, it will decrease until at a low enough p value, p_t, it will become the universal value t_{un}. Hence, as we saw above, strictly speaking, the critical (i. e. in the $p \to p_c$ limit) behavior of the general $D > 1$ systems will always be the universal behavior.

Since the value of $1/g_c$ and thus the values of $\langle \mathbf{r_c} \rangle$ and $\langle \mathbf{r} \rangle$ diverge as $p - p_c \to 0$, minute variations in the value of p cause very strong (orders of magnitude) variations in the former quantities. This implies that the variations that we described above for $1/g$, and thus for $f(g)$, actually take place over a narrow range of p when it is close to p_c. The

first implication of that is that the use of a constant a (say, the one that applies to p_c) in the above equations (e. g. Eq. (37)) is a very good approximation. The second implication is that the "locally" (r range, g range or p range, see also Eq. (38)) determined value of t should be defined differentially as

$$t = \mathrm{d}[\ln G]/\,\mathrm{d}[\ln(p - p_c)]\,, \tag{39}$$

where G is the global conductance of the system. Using this definition and the g_c dependence on $p - p_c$, the above behavior of $f(g)$ can be translated to a p (or a $p - p_c$) dependence of t. We concluded above that far from p_c, t will increase with decreasing $p - p_c$ until it obtains its maximum, and from there on (i. e. for smaller p values) it will decrease until it reaches the corresponding universal value. The questions that arise are, of course, at what p, and for what value of $t - t_{un}$ will the peak t-value and the t-value transition happen, and how does this behavior depend on the parameters of the system [78]. The mathematical expression for the above described behaior has been developed very recently by Johner et al. [78] using the application of the Effective Medium Approximation (EMA). In that case, one essentially usses a lattice model, but with a distribution of bond or g-values taken by using the $f_3(g)$ function that is obtained by applying Eqs. (28), (32) and (36) to the basic EMA equation. They were able to show then that:

$$t - t_{un} = \left[(a-b)/d\right]\left\{\ln\left[yp/\left(p-p_c\right)\right]\right\}^{-2/3}\,, \tag{40}$$

where, $y = \exp[2b/(2a - 2b)]^3$. It is easy to appreciate that this expression follows the expected behavior as otlined above.

Following our above discussion and the result given in Eq. (40), it is apparent that we cannot have a simple-complete mapping of the $D > 1$ systems (Eq. (38)) onto the K&S model as we had for the 1D model (Eq. (33)). On the other hand, the constant parameter α, of Eqs. (33) and (34), can be applied for the characterization of the above general $D > 1$ systems, since the three relevant length scales for all ($D \geq 1$) tunneling models are still the same i. e., a (see above), b and d. Using then $\alpha = 1 - d/(a - b)$ as the system characterization parameter, we show in Fig. 8 the above differential $t(p - p_c)$ dependence, as obtained by the application of the Effective Medium Approximation (EMA) to the 3D percolation tunneling system [70]. In this approximation, one notes that $t_{un} = 1$, but this (EMA value) is not important in the present context, since we are interested here in $t - t_{un}$. The results shown in Fig. 8, confirm then the above mentioned expectations and in particular that (as in the 1D case) the

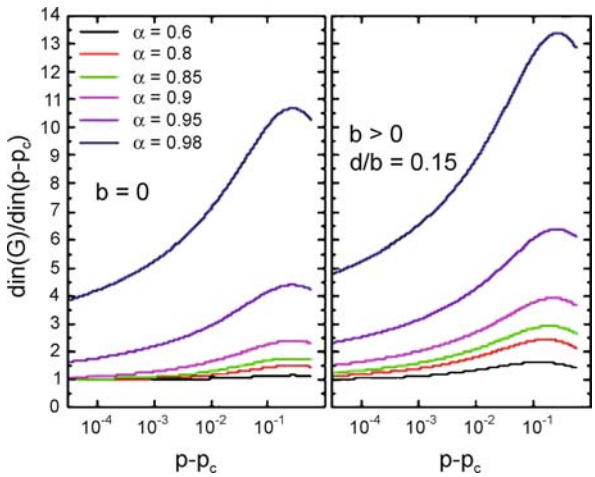

Continuum Percolation, Figure 8
The dependence of the (differential) conductivity exponent as calculated using the Effective Medium Approximation for a system of spheres, where 2*b* is their diameter and 2*d* is the tunneling decay constant. The results are presented for various values of the 1D parameter *α*, such that the higher the values of *α*, the smaller the local conductance values that are involved in the global conductance. (From [70])

larger the value of α (in the $0 \leq \alpha \leq 1$ interval), i. e. the smaller the tunneling decay range in comparison with the inter-particle distances r (or, the smaller the corresponding g-values that are involved in the conduction), the more pronounced the non universal-like behavior (the larger the t values) that accompanies the conduction process. However, in contrast with the $D = 1$ case, the $t > t_{un}$ values are maintained only over a limited $p - p_c$ range, and "eventually" (as $p \to p_c$) a decrease of t towards the value of t_{un} takes place. In addition to this main result, i. e., the confirmation of the peak in the t dependence on p and its decrease to t_{un}, as p approaches p_c, we can derive further conclusions regarding the α dependence of the curves. First, for very small values of α, a non universal behavior will hardly be observed, and second, even though the shift of the $t(p)$ peak with increasing α is very small, it is obvious from our comparison of Eqs. (32) and (36) that this increase shifts the peak towards lower $p - p_c$ values. As to the meaning of this shift in practice, let us note that for a given type of a system (i. e., b and d fixed), the larger the value of a (the smaller the value of p), the more dilute the system. Correspondingly, as apparent from the smaller g values involved in the conduction process then, the decay of $H_3(r)$ in Eq. (36) will be "slower", and the non universal behavior will be also emphasized by the extension of the non universal behavior values into the lower $p - p_c$ regime. A similar argument can be derived from Eq. (38)

since the smaller the value of d/b the larger the value of $\alpha(r)$ i. e. the value of t for a given r. In passing, we also note that the parameters that determine the critical behavior can be controlled externally. For example, a voltage application effect on d, or a pressure application and temperature change on a, will change the value of α and thus the value of t. In that case, we expect to get a new class of phenomena that may be called "physically controlled percolation".

Considering the corresponding numerous data in the literature [128], one finds that there is no convincing direct experimental or Monte Carlo evidence, on a genuine continuum system, for the variation of $t(p)$ as predicted in Fig. 8, but there is some indirect evidence for that behavior (see below). The main reason for not observing the peaked $t(p)$ dependence in the experimental and computational works seems to be the accuracy of these data, or more importantly, the limited $p - p_c$ range that is actually accessible in the corresponding studies. For the experimental data, the latter effect can be appreciated as follows [70]. The proximity to the percolation threshold is usually given, as in Fig. 8, by the parameter $p - p_c$. However, this is not a good parameter for that purpose of the description of the proximity to the threshold in continuum systems. If both p and p_c are small, a small $p - p_c$ does not mean that the system is in close proximity to the critical regime. Rather, a quantity such as $(p - p_c)/p_c$ is a better indication of the proximity to the threshold. The problem is not too severe in lattices when one considers the resistors distribution imposed on lattices [88,111] since the p_c values are larger than 0.1 for two and three dimensions. This problem may be very severe, however, when one interprets experimental data in the continuum, where the proximity parameter is given in terms of the fractional occupied volume τ, and the corresponding critical fractional volume τ_c, i. e. by $(\tau - \tau_c)$. Here, while both, τ and τ_c can be very small, $(\tau - \tau_c)/\tau_c$ can be quite large, indicating that the system is far from the critical region. In this case, single, high, non universal-like, t values, are not too surprising in view of the above discussion, and the limited $(\tau - \tau_c)/\tau_c$ range hardly enables one to detect deviations from a single t value [70]. Computer simulations, on the other hand, "suffer" from finite size effects, and while a variable $t(p)$ behavior has been found [24,78], it has not yet been established to be beyond the influence of such effects.

However, more convincing, though indirect, confirmation of the t value variation with $p - p_c$ seems to be indicated by the experimental data on other electrical properties such as the resistance noise. For this property, the variation of the corresponding exponent, as a function of $\tau - \tau_c$, was found to indicate the decrease of t as τ approaches τ_c [40,42,85]. Hence, in view of the above, the transitions from a non universal behavior to a universal behavior seem to be well understood and well accounted for semiquantitatively, but further work is needed in order to examine this effect more quantitatively.

Future Directions

The relevance of the connectivity to the many areas of science mentioned in Sect. "Definition of the Subject", makes the achievements in the theory of continuum percolation of great importance to numerous natural and artificial systems. In particular, the understanding and conclusions derived from this theory in the last forty years serve as guides to the analysis and understanding of the various properties of these systems. Naturally then, future studies of continuum percolation are expected to proceed along two directions. The first is the further development of the theory itself and issues of principle and depth of understanding, such as the ones emphasized in the present review. The other direction is the application of the general conclusions and principles derived from the first direction for the understanding of the systems and properties such as those mentioned in Sect. "Definition of the Subject". The latter direction deserves quite a few reviews that will center on particular types of systems (say, solid composites and porous media) or particular properties (such as electrical noise and heterogeneous-fluid viscosity). Considering the wide scope of these many issues, we will mention here, as examples, only a couple of them that are relevant to systems and properties that became of wide interest only very recently.

On the very basic-scientific level, the establishment of continuum percolation as a phase transition that can be mapped onto the lattice percolation phase transition is desired, in order to determine rigorously and conclusively that the critical behavior of the geometrical-statistical properties is the same. Another fundamental theoretical problem is the derivation of the percolation thresholds and critical exponents rigorously. The first question that arises in that context is whether this can be done by another approach than the common one used thus far, i. e., the application of the theory of liquids, and whether such an approach can yield better results than that of the latter application. The second question is whether within the rather developed application of the theory of liquids, one can find a way to facilitate its implementation to the degree that will enable unbiased determination of percolation thresholds and critical exponents, as well as its application for "less trivial" objects than those considered thus

far. On the more fundamental level of the latter problem, it is very interesting to know why this application is so successful for the description of the percolation behavior, even though the theory of liquids was not designed for the study of phase transitions.

Considering the empirical approach to the theory we note that the usefulness and the generalization of the concept of "pointedness" have not been studied extensively, in spite of its impressive predications concerning trends in the behavior of percolation thresholds in the continuum. The transparent physical-geometrical meaning of this concept calls then for its further utilization on the one hand, and for trying to understand the reasons for its success on the other hand. In particular we note that at present, the latter approach can be used much more readily and for many more systems than the above mentioned rigorous approach that is based on the theory of liquids. A specific problem that needs attention in order to derive a more complete understanding of the percolation threshold in continuum percolation is the value of B_c at the hard-core limit

Turning to the behavior of the dynamical properties, it appears that most of the basic principles of the corresponding theory are well understood by now. However, the problem of variable "bond strengths" is not well accounted for, in general, and beyond the "strong-interaction" limit (e. g., tunneling only to nearest neighbors) in particular. The meaning of the percolation threshold in corresponding systems and the predictions of the corresponding critical behaviors, are still open. In particular, the evaluation of the "residual" conductivity in comparison with the dominant percolation-like subnetwork has not been given thus far. For this, there is the need for the other corresponding parameters that characterize the charge, the mass, or energy, transfer between the objects in the system. It further appears however, that the most important information that one needs, when these principles are applied for the understanding of particular continuum systems, is the distribution function of the sizes and shapes of the objects in the system, as well as their separations. As emphasized in this review, this information is usually scarce, thus limiting the possible applications of the general understanding and principles outlined in Sect. "The Critical Behavior of the Dynamical Properties in the Continuum" to specific systems. Here, the utilization of very modern experimental characterization techniques, such as local probe microscopies, should provide the link needed for the above application. On the theoretical-computational end, the distribution functions associated with the structure, and thus with the dynamical physical properties, need to be derived, in order to appreciate trends in the critical behavior (i. e., the values of the critical exponents) in various systems. These trends will become amenable for examination, however, only with the ability to produce or find systems where the percolation threshold can be approached much more closely than at present, in the experiments, and to create much larger systems (in order to avoid finite size effects) in the computations. Such developments are expected to yield the (generally missing so far) information on the dependence of the critical exponents on the proximity to the percolation threshold. It appears then that the meaning of the percolation threshold, when farther than closest neighbors are involved (such as in the case of tunneling-percolation systems where the "interaction" range is not much smaller than the size of the objects), as well as the critical behavior of such systems, beyond the intuitive picture suggested in Sect. "The Critical Behavior of the Dynamical Properties in the Continuum", is very much called for.

Let us now turn to outline a few typical (of the numerous) questions associated with specific systems, in order to illustrate the issues and systems whose understanding rests on the concepts mentioned above. Such are the effects of charging and quantum confinement within corresponding particles that are embedded in a percolating system. The understanding of the interplay between these effects and the effects of the neighboring network is still in its infancy. This interplay has been shown recently to have pronounced effects not only on the transport, but also on the phototransport in such systems. Also, the effects of special features of the network, such as its fractal dimension, on the dynamical behavior are still a matter of controversy. Again, local probe microscopy, on the experimental end, and corresponding theoretical-computational work, on the other end, may resolve the corresponding problems. Another issue of great interest is the consideration of objects that are more complicated in their shape than the convex objects dealt with in the present review. Such are the systems that are of great present interest, i. e., carbon nanotube composites, where the objects can be described as having a "wavy" shape. One question that arises there, and that was dealt with only very recently, is the effect of this "wavyness" on the percolation threshold. A broader issue that is relevant to many systems, but for which no systematic discussion has been given thus far, is the relation between the trends in the critical behavior of the electrical, mechanical and rheological properties, in corresponding solids or molten composites. These are of great importance in the fields of chemical-material and electrical-material engineering.

Finally, a subject that has not been discussed in the context of continuum percolation, and, as we saw above,

is mainly concerned with global properties of systems, is the statistics of finite clusters where there is, or there is no, interaction between the objects. This is of great importance for systems where the finite clusters as such determine the properties of interest, such as in cases where there is a charge transfer by a delocalization process. This is of importance, for example, in the cases of illumination-generated charge carriers, since there, the local connectivity of the system around the particle, in which the carriers were optically generated, determines the probability of their separation and thus the "deconfinement" and corresponding elimination of the radiative recombination.

Acknowledgments

The present review could not have been written without the stimulation and the intensive collaboration that I had with the many colleagues and students, whose papers that were co-authored with me, are cited in this review. In particular, I would like to thank Dr. C. Grimaldi and Dr. N. Wagner for the on going collaboration and for the critical reading of the present manuscript. This work was supported by the Israel Science Foundation (ISF). The author holds the Enrique Berman chair in Solar Energy Research at the Hebrew University.

Bibliography

Primary Literature

1. Abeles B (1976) Granular metal films. Appl Solid State Sci 6:1–117
2. Abeles B, Pinch HL, Gittleman JI (1975) Percolation conductivity in W-Al_2O_3 granular metal films. Phys Rev Lett 36:257–260
3. Alon U, Drory A, Balberg I (1990) Systematic derivation of percolation thresholds in continuum systems. Phys Rev A 42:4634–4638
4. Alon U, Balberg I, Drory A (1991) New, heuristic, percolation criterion for continuum systems. Phys Rev Lett 66:2879–2882
5. Ambegaokar V, Halperin BI, Langer JS (1971) Hopping conductivity in disordered systems. Phys Rev B 4:2612–2620
6. Andrade JS et al (2000) Flow between two sites on a percolation cluster. Phys Rev E 62:8270–8281
7. Archie GE (1942) The electrical resistivity log as an aid in determining some reservoir characteristics. Trans Am Inst Min Metall Pet Eng 146:54–62
8. Azulay D et al (2003) Electrical-themal switching in carbon black-polymer composites as a local effect. Phys Rev Lett 90:236601
9. Balberg I (1986) Excluded-volume explanation of Archie's law. Phys Rev B 33:3618–3620
10. Balberg I (1987) Tunneling and nonuniversal conductivity in composite materials. Phys Rev Lett 59:1305–1308
11. Balberg I (1987) Recent developments in continuum percolation. Philos Mag B 56:991–1002
12. Balberg I (1998) New limits on the continuum-percolation transport exponents. Phys Rev B 57:13351–13354
13. Balberg I (2002) A comprehensive picture of the electrical transport phenomena in carbon black-polymer composites. Carbon 40:139–143
14. Balberg I, Binenbaum N (1983) A Computer study of the percolation threshold in a two-dimensional anisotropic system of conducting sticks. Phys Rev B 28:3799–3812
15. Balberg I, Binenbaum N (1985) Cluster structure and conductivity of three-dimensional continuum Systems. Phys Rev A 31:1222–1225
16. Balberg I, Binenbaum N (1987) Invariant properties of the percolation thresholds in the soft core-hard core transition. Phys Rev A 35:5174–5177
17. Balberg I, Binenbaum N (1987) Scher and Zallen criterion: applicability to composite systems. Phys Rev B 35:8749–8752
18. Balberg I, Blanc J (1985) Capacitive noise spectra of a disordered material. Phys Rev B 31:8295–8297
19. Balberg I, Bozowski S (1982) Percolation in composites of random stick-like conducting particles. Solid State Commun 44:551–554
20. Balberg I, Binenbaum N, Anderson CH (1983) Critical behavior of the two-dimensional sticks system. Phys Rev Lett 51:1605–1609
21. Balberg I, Binenbaum N, Wagner N (1984) Percolation thresholds in the three-dimensional sticks system. Phys Rev Lett 52:1465–1609
22. Balberg I et al (1984) Excluded volume and its relation to the onset of percolation. Phys Rev B 30:3933–3943
23. Balberg I et al (1988) Critical behavior of the electrical resistance and its noise in inverted random-void systems. Phys Rev Lett 60:1887–1890
24. Balberg I et al (1990) Tunneling and percolation behavior in granular metals. Mater Res Soc Symp Proc 195:233–238
25. Balberg I, Berkowitz B, Drachsler GE (1991) Application of a percolation model to flow in fractured hard rocks. J Geophys Rev 96:10015–10021
26. Balberg I et al (2004) Percolation and tunneling in composite materials. Int J Mod Phys B 18:2091–2121
27. Balberg I et al (2007) Fundamental transport processes in ensembles of silicon quantum dots. Phys Rev B 75:225–329
28. Bergman DJ (2003) Exact relation between critical exponents for elastic stiffness and electrical conductivity of percolation systems. Phys B Condens Matter 338:240–246
29. Bergman DJ et al (1983) Critical behavior of the low-field Hall conductivity at a percolation threshold. Phys Rev Lett 50:1512–1515
30. Berhan L, Sastry AM (2007) Modeling percolation in high-aspect-ration fiber systems. Phys Rev E 75:041–121
31. Berkowitz B, Balberg I (1992) Percolation approach to the problem of hydraulic conductivity in porous media. Transp Porous Media 9:275–286
32. Berkowitz B, Balberg I (1993) Percolation theory and its application to groundwater hydrology. J Water Resour Res 29:775–794
33. Broadbent SR, Hammersley JM (1957) Percolation processes, crystals and mazes. Proc Camb Philos Soc 53:629–641
34. Bug ALR, Safran SA, Webman I (1985) Continuum percolation of rods. Phys Rev Lett 54:1412–1415
35. Bug ALR et al (1985) Do interactions raise or lower a percolation threshold? Phys Rev Lett 55:1896–1899

36. Cametti C et al (1990) Theory and experiment of electrical conductivity and percolation locus in water. Phys Rev Lett 64:1461–1464
37. Charlaix E, Guyon E, Rivier N (1984) A criterion for percolation threshold in a random array of plates. Solid State Commun 50:999–1002
38. Charlaix E, Guyon E, Roux S (1987) Permeability of a random array of fractures of widely varying apertures. Transp Porous Media 2:31–43
39. Chatterjee AP (2000) Continuum percolation in macromolecular fluids. J Chem Phys 113:9310–9317
40. Chen CC, Chou YC (1985) Electrical-conductivity fluctuations near the percolation threshold. Phys Rev Lett 54:2529–2532
41. Chiew YC, Glandt ED (1983) Percolation behavior of permeable and of adhesive spheres. J Phys A 16:2599–2608
42. Chiteme C, McLachlan DS, Balberg I (2003) 1/f or flicker noise in cellular percolation systems. Phys Rev B 67:024207
43. Clerc JP et al (1990) The ac electrical conductivity of binary-disordered systems, percolation clusters, fractals and related models. Adv Phys 39:191–308
44. Coniglio A, DeAngelis U, Forlani A (1977) Pair connectedness and cluster size. J Phys A Math Gen 10:1123–1139
45. Dalmas F et al (2006) Carbon nanotube-filled polymer composites. Numerical simulations of electrical conductivity in three-dimensional entangled fibrous networks. Acta Mater 54:2923–2931
46. Dalton NW, Domb C, Sykes MF (1964) Dependence of the critical concentration of dilute ferromagnet on the range of interaction. Proc Phys Soc 83:496–498
47. Day AR et al (2003) Spectral representation of the electrical properties of layered materials. Phys B Condens Matter 338:24–30
48. de Arcangelis L, Redner S, Coniglio A (1985) Anomalous voltage distribution of random resistor networks and a new model for the backbone at the percolation threshold. Phys Rev B 31:4725–4727
49. DeSimone T, Stratt RM, Demoulini S (1986) Continuum percolation in an interacting system: Exact solution of the Percus-Yevick equation for connectivity in liquids. Phys Rev Lett 56:1140–1143
50. Drory A (1996) Theory of continuum percolation. I. General formalism. Phys Rev E 54:5992–6002
51. Drory A (1996) Theory of continuum percolation. II. Mean field theory. Phys Rev E 54:6003–6013
52. Drory A (1997) Exact solution of a one-dimensional continuum percolation model. Phys Rev E 55:3878–3885
53. Drory A et al (1991) Analytic derivation of percolation thresholds in anisotropic systems of permeable objects. Phys Rev A 43:6604–6612
54. Drory A, Balberg I, Berkowitz B (1994) Random-adding determination of percolation thresholds in interacting systems. Phys Rev E 49:R949–952
55. Drory A, Balberg I, Berkowitz B (1995) Application of the central-particle potential approximation for percolation in interacting systems. Phys Rev E 52:4482–4494
56. Drory A et al (1997) Theory of continuum percolation. III. Low-density expansion. Phys Rev E 56:1379–1395
57. Du F et al (2004) Nanotube networks in polymer nanocomposites: Rheology and electrical conductivity. Macromolecules 37:9048–9055
58. Englman R, Gur Y, Jaeger Z (1983) Fluid flow through a crack network in rocks. J Appl Mech 50:707–711
59. Feng S, Halperin BI, Sen PN (1987) Transport properties of continuum systems near the percolation threshold. Phys Rev B 35:197–214
60. Flory PM (1941) Molecular size distribution in three dimensional polymers. J Am Chem Soc 63:3083–3100
61. Fonseca LF, Balberg I (1993) Resistivity and electrical noise in granular metal composites. Phys Rev B 48:14915–14924
62. Fortuin CM, Kasteleyn PW (1972) Random-cluster model. 1. Introduction and relation to other models. Phys 57:536–564
63. Fortunato S, Stauffer D, Coniglio A (2004) Percolation in high dimension is not understood. Phys A 334:307–311
64. Foygel M et al (2004) Theoretical computational studies of carbon nanotube composites and suspensions: electrical and thermal conductivity. Phys Rev B 71:164201
65. Gawlinski ET, Redner S (1983) Monte Carlo renormalization group for continuum percolation with excluded-volume interactions. J Phys A Math Gen 16:1063–1071
66. Gawlinski ET, Stanley HE (1981) Continuum percolation in two dimensions: Monte carlo tests of scaling and universality for non-interacting discs. J Phys A Math Gen 14:L291–L299
67. Grannan DM, Garland JC, Tanner DB (1981) Critical behavior of the dielectric constant of a random composite near the percolation threshold. Phys Rev Lett 46:375–378
68. Grassberger P (2003) Critical percolation in high dimensions. Phys Rev E 67:036101
69. Grest GS et al (1986) Dynamic percolation in microemulsions. Phys Rev A 33:2842–2845
70. Grimaldi C, Balberg I (2006) Tunneling and non-universality in continuum percolation systems. Phys Rev Lett 96:066602
71. Grimaldi C et al (2003) Segregated tunneling-percolation model for transport nonuniversality. Phys Rev B 68:024207
72. Halperin BI, Feng S, Sen PN (1985) Differences between lattice and continuum percolation transport exponents. Phys Rev Lett 54:2391–2394
73. Haan SW, Zwanzig R (1977) Series expansions in a continuum percolation problem. J Phys A Math Gen 10:1547–1555
74. Hansen PH, McDonald RM (1986) Theory of Simple Liquids. Academic, London
75. Heyes DM, Melrose JR (1998) Percolation thresholds of simple liquids. J Phys A Math Gen 21:4075–4081
76. Holcomb DF, Rehr JJ (1969) Percolation in heavily doped semiconductors. Phys Rev 183:773–776
77. Isichenko MB (1992) Percolation, statistical topography, and transport in random media. Rev Mod Phys 64:961–1043
78. Johner N et al (2008) Transport exponent in a three-dimensional continuum tunneling-percolation model. Phys Rev B 77:174–204
79. Kapitulnik A, Deutscher G (1982) Percolation characteristics in discontinuous thin films of Pb. Phys Rev Lett 43:1444–1448
80. Kirkpatrick S (1973) Percolation and conduction. Rev Mod Phys 45:574–588
81. Knite M et al (2002) Electric and elastic properties of conductive poylmer nanocomposites on macro- and nanoscales. Mater Sci Eng C 19:15–19
82. Kogut PM, Straley J (1979) Distribution–induced non-universality of the percolation conductivity exponents. J Phys C Solid State Phys 12:2151–2159

83. Laria D, Vericat F (1989) Percolation behavior of long permeable objects: A reference interaction-site-model study. Phys Rev B 40:353–360
84. Lin C-R, Chen W-J (1999) The links–nodes-blobs model for shear thinning-yield stress fluids. Colloid Polym Sci 277:1019–1025
85. Mandal P et al (1997) Temperature and magnetic field dependence of the resistivity of carbon-black composites. Phys Rev B 55:452–456
86. McCarthy JF (1987) Continuum percolation of disks and the random lattice. Phys Rev Lett 58:2242–2244
87. Miller A, Abrahams E (1960) Impurity conduction in low concentrations. Phys Rev 120:745–755
88. Murat M, Mariner S, Bergman DJ (1986) A transfer matrix study of the conductivity and permeability exponents in continuum percolation. J Phys A Math Gen 19:L275–L279
89. Neda Z, Florian R, Brechet Y (1999) Reconsideration of continuum percolation of isotropically oriented sticks in three dimensions. Phys Rev E 59:3717–3719
90. Netemeyer SC, Glandt ED (1986) Percolation behavior of the square-well fluid. J Chem Phys 85:6054–6059
91. Octavio M et al (1988) Nonuniversal critical behavior in the critical current of superconducting composites. Phys Rev B 37:9292–9297
92. Pagnotta SE et al (2005) Glassy behavior of a percolative water-protein system. Phys Rev E 71:031506
93. Park S et al (2004) Percolation conduction in the half-metallic ferromagnetic and ferroelectric mixture of (La,Lu,Sr)MnO_3. Phys Rev Lett 92:167206
94. Pike GE (1978) Conductivity of thick films (cermet) resistors as a function of metallic particle volume fraction. In: Garland JC, Tanner DB (eds) Electrical Transport and Optical Properties of Inhomogeneous Media (AIP Conf Proc 40). AIP, New York, pp 366–371
95. Pike GE, Seager CH (1974) Percolation and conductivity: a computer study I. Phys Rev B 10:1421–1434
96. Planes J et al (1998) Transport properties of polyanilline-cellulose-acetate blends. Phys Rev B 58:7774–7785
97. Powell MJ (1979) Site percolation in randomly packed spheres. Phys Rev B 20:4194–4198
98. Rapp O, Shinivas V, Poon SJ (2005) Critical exponents at the metal-insulator transition in AlPdRe quasicrystals. Phys Rev B 71:012202
99. Re A et al (2006) Correlated fragile site expression allows the identification of candidate fragile genes involved in immunity and associated with carcinogenesis. BMC Bioinformatics 7:413
100. Rintoul MD, Torquato S (1997) Precise determination of the critical threshold and exponents in three-dimensional continuum percolation model. J Phys A Math Gen 30:L585–L592
101. Robinson PC (1983) Numerical calculations of critical densities for lines and planes. J Phys A Math Gen 17:2823–2830
102. Rubin Z et al (1999) Critical behavior of the electrical transport properties in a tunneling-percolation system. Phys Rev B 49:12196–12199
103. Saar MO, Manga M (2002) Continuum percolation of randomly oriented soft-core prisms. Phys Rev E 65:056131
104. Satz H, Fortunato S (2001) Percolation and confinement in SU(2) gauge theory. Nucl Phys A 681:466C–471C
105. Sahimi M (1994) Application of Percolation Theory. Taylor, London
106. Sahimi M (1998) Non-linear and non-local transport in heterogeneous media: from long-range correlated percolation to fracture and materials breakdown. Phys Rep Rev Sect Phys Lett 306:214–295
107. Sahimi M, Imdakm AO (1988) The effect of morphological disorder on hydrodynamic dispersion in flow through porous media. Phys A Math Gen 21:3833–3870
108. Scher H, Zallen R (1970) Critical density in percolation processes. J Chem Phys 53:3759–3761
109. Schrijver CJ et al (1992) Patterns in the photosphereic magnetic-field and percolation theory. Astron Astrophys 253: L1–L4
110. Sen PN, Scala C, Cohen MH (1981) A self-similar model for sedimentary rocks with application to the dielectric constant of fused glass beads. Geophysics 46:781–795
111. Sen PN, Roberts JN, Halperin BI (1985) Nonuniversal critical exponents for transport in percolating systems with a distribution of bond strengths. Phys Rev B 32:3306–3308
112. Shante VKS, Kirkpatrick S (1971) An introduction to percolation theory. Adv Phys 20:325–357
113. Shimoni N et al (2002) tomographic-like reconstruction of the percolation cluster as a phase transition. Phys Rev B 66:020102R
114. Shklovskii BI, Efros AL (1984) Electronic properties of doped semiconductors. Springer, New York
115. Skal AS, Shklovskii BI (1974) Influence of impurity concentration on the hopping conduction in semiconductors. Sov Phys Semicond 7:1058–1059
116. Smart JS (1968) Effective field theories of Magnetism. Sanders, Philadelphia
117. Sokolowska D, Krol-Otwinowska A, Moscicki JK (2004) Water network percolation transition in hydrated yeast. Phys Rev E 70:052901
118. Song Y, Lee S-I, Gaines JR (1992) Ac conductivity and 1/*f* noise in a Cr-film lattice-percolation system. Phys Rev B 46:14–20
119. Stanley HE (1971) Introduction to phase transitions and critical phenomena. Clarendon, Oxford
120. Stanley HE (1977) Cluster shapes at the percolation threshold: an effective cluster dimensionality and its connection with critical-point exponents. J Phys A Math Gen 10:L211–L220
121. Stauffer D, Aharony A (1992) Introduction to percolation theory. Taylor, London
122. Stinchcombe RB (1976) Conductivity and spin-wave stiffness in disordered systems-an exactly soluble model. J Phys C Solid State Phys 7:179–203
123. Stockmayer WH (1943) Theory of molecular size distribution and gel formation in branched-chin polymers. J Chem Phys 11:45–55
124. Toker D et al (2003) Tunneling and percolation in metal-insulator composite materials. Phys Rev B 68:041403
125. Thompson AH, Katz AJ, Krohn CE (1987) The microgeometry and transport properties of sedimentary rocks. Adv Phys 365:625–694
126. Torquato S, Lu B, Rubinstein J (1990) Nearest neighbor distribution function in many-body systems. Phys Rev A 41:2059–2075
127. Tremblay A-MS, Fourcade B, Breton P (1989) Multifractals and noise in metal-insulator composites. Phys A 157:89–100
128. Vionnet-Menot S et al (2005) Tunneling-percolation origin of nonuniversality: theory and experiments. Phys Rev B 76:064201

129. Vyssotsky VA et al (1961) Critical percolation probabilities (bond problem). Phys Rev 123:1566 1567
130. Wagner N, Balberg I (1987) Anomalous diffusion and continuum percolation. J Stat Phys 59:369–382
131. Wagner N, Balberg I, Klein D (2006) Monte carlo results for continuum percolation in low and high dimensions. Phys Rev E 74:021127
132. Williams JC, Snyder SA (2005) Restoring habitat corridors in fragmented landscapes using optimization and percolation models. Environ Model Assess 10:239–250
133. Wilkinson D (1986) Percolation effects in immiscible displacement. Phys Rev A 34:1380–1391
134. Wille JJ, Elson EL, Okamoto RJ (2006) Cellular and matrix mechanics of bioartificial tissues during continuous cyclic stretch. Ann Biomed Eng 34:1678–1690
135. Wu ZH et al (2006) Optimal paths in complex networks with correlated weights: the worldwide airport network. Phys Rev E 74:056104
136. Zallen R (1983) The Physics of Amorphous Solids. Wiley, New York
137. Zallen R, Scher H (1971) Percolation on a continuum and the localization delocalization transition in amorphous semiconductors. Phys Rev B 4:4471–4479
138. Ziman JM (1979) Models of Disorder. Cambridge University Press, New York

Books and Reviews

Balberg I (1987) Recent developments in continuum percolation. Philos Mag B 56:991–1002
Balberg I et al (2004) Percolation and tunneling in composite materials. Int J Mod Phys B 18:2091–2121
Berkowitz B, Balberg I (1993) Percolation theory and its application to groundwater hydrology. J Water Resour Res 29:775–794
Isichenko MB (1992) Percolation, statistical topography, and transport in random media. Rev Mod Phys 64:961–1043
Kirkpatrick S (1973) Percolation and conduction. Rev Mod Phys 45:574–588
Sahimi M (1994) Applications of Percolation Theory. Taylor, London
Sahimi M (1998) Non-linear and non-local transport in heterogeneous media: from long-range correlated percolation to fracture and materials breakdown. Phys Rep Rev Sect Phys Lett 306:214–295
Shklovskii BI, Efros AL (1984) Electronic properties of doped semiconductors. Springer, New York
Stauffer D, Aharony A (1992) Introduction to percolation theory. Taylor, London
Zallen R (1983) The Physics of Amorphous Solids. Wiley, New York

Continuum Robots

Ian D. Walker[1], Keith E. Green[2]
[1] Department of Electrical and Computer Engineering, Clemson University, Clemson, USA
[2] School of Architecture, Clemson University, Clemson, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Case Study 1: Trunk Continuum Robots – OctArm
Case Study 2: Continuum Robot Surfaces – AWE
Summary
Future Directions
Acknowledgments
Bibliography

Glossary

Degree of freedom (Sub)set of a physically reconfigurable system whose configuration can always be specified by a single variable.

Kinematics Geometric (and differential geometric) models of mechanical systems, not including effects of dynamics (masses, forces, inertias, etc.).

Robot A physically reconfigurable system that has multiple degrees of freedom and is (at least partially) computer-controlled.

Actuators Devices used to transfer mass (herein portions of a robot system) from one place to another within the system's environment.

Sensors Devices used to infer information, usually either about the internal state of a system, or about external environment surrounding that system.

Control Use of sensors and actuators to reconfigure a system (herein robots) according to a desired plan.

Continuum Continuous in nature; herein usually the backbone/core of a robot.

OctArm Continuum robot manipulator inspired by octopus arms.

AWE ("Animated Work Environment") Reconfigurable environment featuring continuum robot components, focused on work environments featuring computing.

Definition of the Subject

In this article, we discuss some of the key issues involved in the design and implementation of the emerging class of "invertebrate-like" continuum robots. Using two case studies of continuum robots developed recently at Clemson University, we overview the issues involved in realizing continuum robots and their deployment. The potential of these types of robots for enhanced productivity in novel applications is discussed.

In the first case study, we describe the design of the "OctArm" continuum manipulator robot hardware, and discuss the results of field testing of these novel "trunk-

like" robots. OctArm robots are able to adapt their shape to their environment, to access difficult-to-reach areas, and to perform adaptive grasping using their entire arms. Lessons learned and implications for future robot manipulators in the field are discussed.

In the second case study, we describe the new "Animated Work Environment" (AWE) concept. AWE is an articulated, programmable, interior environment and chassis, embedded with integrated digital technologies. Featuring a continuum, morphing robot surface controlled by a user-friendly interface, AWE is characterized as adaptive and robust when applied to a range of work activities and working populations currently not accommodated by robotics.

The OctArm and AWE were selected as case studies to represent the potential broad range of application of continuum robots: the former suggests how continuum robots might operate in environments too harsh and hazardous for routine human activity, and the latter offers the promise of continuum robots in everyday interior environments supporting human collaboration and productivity.

Introduction

Industrial robot manipulators typically feature a small finite number (typically 4–6) of serially-connected rigid links. This is well-matched to factory "pick-and-place" applications, but such manipulators have not been generally successful in tackling more complex tasks (such as interacting with humans) in less structured environments. A frequently considered question in robot manipulator research is: "what happens as the number of joints becomes much larger, and ultimately increases towards infinity?" Examples of serial rigid-link systems with many joints exist in nature (snakes, mammalian and fish backbones, for example), and these biological "existence proofs" have provided motivation and insight for robot analysts and hardware designers for many years. This interest has resulted in the development of several special sub-classes of manipulators, collectively known as *redundant manipulators* [86].

A manipulator is said to be (kinematically) *redundant* if its configuration (joint) space degrees of freedom exceed its task space degrees of freedom. Hence, since spatial (end effector) task motions are generally six-dimensional, seven or eight degree of freedom spatial rigid-link manipulators are usually considered to be redundant. A manipulator is generally considered to be *hyperredundant* if its configuration (joint) space degrees of freedom greatly exceed its task space degrees of freedom. Hyperredundant manipulators have enhanced potential to use their extra joints for whole arm grasping/manipulation and maneuvering within tight obstacle fields. Their anticipated applications therefore include operations in congested environments (disaster relief, medical applications, etc.).

The field of hyperredundant manipulators has evolved in two major directions: "vertebrate-like" rigid-link designs, and "invertebrate-like" continuum manipulators [86]. Here we describe recent developments in continuum robots. Continuum robots are most typically defined by their "continuous backbone structures" – structures without the skeletal design of traditional rigid-link, robot manipulators [66,78]. Defined by this unusual structure, continuum robots may be described as "invertebrate-like" as opposed to the "vertebrate-like" nature of traditional robots [33]. Much research in the area takes inspiration from biology, and has often been strongly influenced by the design and functionality of invertebrate structures such as octopus arms [46,50,79,81] and elephant trunks [19,30,46,76,85]. As continuum designs can be scaled down to very small sizes [69,72,73], potential applications include medicine [4,12,68,75,82].

The main feature of continuum structures is their inherent smoothness. Instead of bending at discrete points (joints, or "elbows") along the "backbone", they can smoothly bend anywhere along their structure. Several designs, including the OctArm robots discussed in this article, additionally feature the ability to extend (elongate) along the length of the backbone. Almost all continuum robot designs exhibit significant compliance, i. e. they inherently present a "soft" rather than "stiff" interface to the environment.

Two functions in particular define the potential advantages of continuum robots: operation in congested environments and whole-arm grasping. In congested environments, the lack of "elbows" and the ability to bend at many locations allows continuum structures to "snake" in and through very tight obstacle fields. (Snakes are vertebrates, but their "links" are small and many, and their general movements match that of continuum structures.) There are numerous important arenas in which key applications of traditional robotics are currently impractical (many surgical applications, search and rescue operations in collapsed buildings, etc.) and where the effective deployment of continuum structures exhibiting snake-like behaviors would make a significant contribution to Society.

In whole-arm grasping, continuum robots again prove their potential advantage: their smooth and compliant nature allows them to gently interact with the world by adapting their shape to that of environmental objects (i. e.

contact along a continuum of the robot) [67]. Note that this is the form of adaptive manipulation used by biological continuum structures such as elephant trunks and octopus arms.

In the following case studies, we outline the design and development of two very different types of continuum robot design, the "OctArm" series of manipulator, and the "AWE" continuum surface. In both cases, the ability of the continuum robot structures to bend smoothly is key to enabling the innovative application scenarios.

Case Study 1: Trunk Continuum Robots – OctArm

Introduction to Octarm Robots

Continuum robot hardware designs can be sub-divided (via their means of actuation) into the categories of "extrinsically actuated" and "intrinsically actuated". Extrinsically actuated continuum robots have their actuators located outside the continuum structure itself, and typically transmit power to the continuum structure though tendons. Intrinsically actuated continuum robots have the actuators embedded in the continuum structure.

At Clemson, we have developed continuum robot hardware actuated both extrinsically and intrinsically. In this section, we discuss the OctArm series of intrinsically actuated continuum arms developed over the past several years under funding from DARPA. The OctArm design was significantly influenced by the underlying design of octopus arms [46,50,70]. Key OctArm design goals were to keep the arm as "soft" and compliant as possible, while incorporating the actuators in the body of the arm.

The resulting robots feature a serially arranged set of "sections". Each section corresponds roughly to a link in a conventional rigid link robot [71]. The variables determining changes in shape of each section correspond conceptually to "joint angles" in conventional robot arms. However, the shape of each section in the OctArm design is determined by two or more variables in a coupled nonlinear relationship. A series of kinematic models have been developed in order to command the OctArms into desired shapes. These models are easily generalized to a wide variety of previously proposed and constructed continuum robots.

Overview of Octarm Robots

The OctArm series of continuum robots feature three (or four) sections, each section having three degrees of freedom, for a total of nine (or twelve) degrees of freedom per arm (see Fig. 1, showing a three-section version) [64]. Each section has two degrees of bending freedom, and one degree of extension. The bending and extension in a section are achieved via variation of pressure in three "air muscles" (McKibben actuators) built into the system (in fact, the air muscles – tubes in the figure – form the majority of the arm structure).

Continuum Robots, Figure 1
Three-section Octarm V continuum robot

The arms are approximately one meter long, with a roughly circular cross-sectional profile, and are tapered from base to tip. The air pressure in the actuators is controlled via commercial pressure regulators, which form the inner loop of the feedback control system. At a higher level, commands to the pressure regulators are provided via a real-time kernel on a PC-104 system, with wireless interface to an operator control PC/joystick. The control code is written in C+. Algorithm development and user interface code utilize Matlab and Simulink modules [20].

The shape of the arms is measured via cables integrated into the arm. The cables change length as the arm moves; changes in cable length are then measured by encoders at the base; and the shape of the arm is inferred by changes in the cable length, following a geometric model. This approach gives a good approximate shape of the arm, which is displayed to the operator via a graphical interface, driven by the sensor information in real-time.

Related Work

Continuum robot hardware designs (usually arms/trunks or fingers/whiskers, though some robot snake/worm [33] designs also fit the classification) have appeared in the literature since the early 1960's [2]. Although there have been many proposed continuum robot designs [51], and numerous hardware realizations [3,10,15,29,35,36,37,47,53,61,62,84], until recently they have largely been "laboratory curiosities", used for a few "proof-of-concept" demonstrations and then abandoned. The main reason for

this has been a lack of underlying theory, corresponding to traditional robot kinematics and dynamics, to enable the coordination of their degrees of freedom. This deficiency meant that even "simple" tasks such as moving the tip of the robots in a straight line were difficult if not impossible to achieve. This in turn has made it very difficult to apply continuum robots to real tasks.

However, in the last ten years, a complete and general body of theory (at least for kinematics) has been established [38,39,40,58,59]. Our group at Clemson University has been a major contributor to this theory [24,25,26,30,31,32,41,43,44,45,74,80] building on existing pioneer efforts [13,14,16,17,83]. This emerging body of understanding now enables the consistent coordination of the degrees of freedom within general continuum structures. Programming of continuum robots can now theoretically be done in terms of "task space" and "configuration space" in an analogous way to that for conventional robots.

At Clemson, the first author's group has demonstrated the ability for the "OctArm" continuum robot hardware to successfully maneuver through complex obstacle fields and grasp and manipulate objects of many different shapes over a wide range (orders of magnitude) of size and weight, and with widely different physical characteristics (rigid, soft, flexible ...) [8,9,18]. The following section describes the deployment of the OctArm hardware in realistic field trials.

Example Octarm Field Trials

The OctArm system underwent extensive field trials in the spring of 2005 and 2006 at the Southwest Research Institute (SwRI) in San Antonio, Texas, USA [42,52]. The SwRI testbed is specially designed for the systematic evaluation of robotic systems in the field. The test environment includes rubble piles and a dry riverbed which can be flooded to create turbulent conditions. The main goal for the trials was to evaluate the ability of the OctArm system to stably and adaptively grasp a wide range of objects under a variety of conditions. For the field tests, OctArm V was mounted on a Foster–Miller TALON platform (shown in Fig. 2).

The OctArm base was attached to the second link of the TALON robot arm. The control valves and two air tanks provided nine channels of controlled pneumatic pressure. The control computer was mounted on the back of the TALON. The TALON and arm were controlled via wireless connection. The system was operated under joystick control via the wireless link and in view of the operator. More details of the user interface used are described in [20].

Continuum Robots, Figure 2
OctArm V mounted on Foster–Miller TALON system

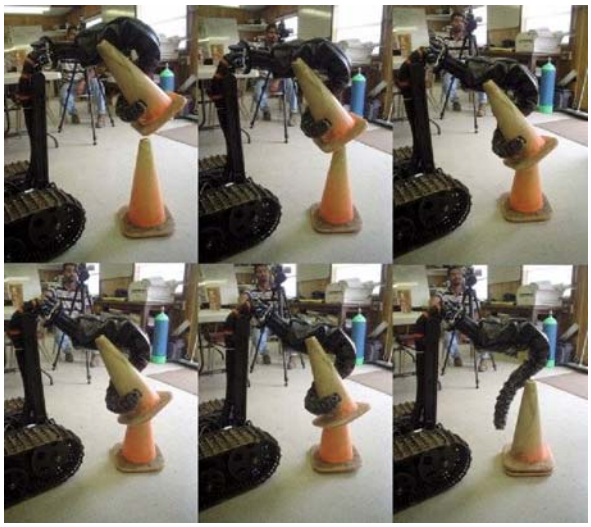

Continuum Robots, Figure 3
Cone stacking task

Tasks included stacking and unstacking traffic cones (see Fig. 3). The ability of the system to grasp objects, such as spheres and cylinders, over a wide range of scales was recorded. These tasks are inherently problematic for traditional parallel jaw-end effectors. The operations were timed, videotaped, and recorded in detail to provide a baseline performance measure for continuum robots under the above conditions.

The system was also operated in water. Submerged in water (shown in Fig. 4), the OctArm attempted to grasp various payloads and to maintain grasps under turbulent flow. This tested the potential of the system for robust grasping under disturbances. The system was operated as well alongside rubble piles. This provided an initial benchmark for remote operation in congested environments.

Continuum Robots, Figure 4
Grasp in turbulent water

During the trials, the OctArm grasped and manipulated various sized cylinders and spheres. The system was able to successfully conform to the varying shapes of these payloads, demonstrating an ability to adapt to environmental conditions not found in traditional manipulators. The continuum system was able to successfully grasp both spherical (balls) and elongated objects (pieces of wood) within the underwater environment, and to maintain the grasps despite significant flow.

The results of the trials clearly demonstrated that continuum robot arms exhibit more adaptive and versatile operation than do conventional robot manipulators. Note that the continuum arm in these trials featured no shape sensing; the shapes displayed to the user was inferred only from the pressures at the actuator input valves. This proved a highly inaccurate estimate, as external effects such as sag due to gravity or external loading were not accounted for in any way. Shortly after the trials described in this section were performed, we integrated shape sensors into the design in order to reduce this inaccuracy.

Case Study 2: Continuum Robot Surfaces – AWE

Introduction to AWE

Our second case study features a novel robotic system where the continuum "trunk" is expanded into a continuum "surface". The surface forms a key part of our "Animated Work Environment" (AWE), aimed at work activities and working populations currently not accommodated by robotics [27,28].

A recent trend in the nature, place and organization of working life is the growing complexity of work, the emergence of new working populations (older workers, unskilled workers, telecommuters and flexible shift workers), and the increasing likelihood that workers are working, at least part-time, at home. As the home becomes more an office, the office is becoming more a home, where "hot desks," lounges, sofas, WI-FI and internal networks have replaced office cubicles and hard-wired, isolated workstations. These dramatic transformations in the nature of work, combined with unprecedented new technologies associated with working life, suggest a re-evaluation of the relations between workers, their technologies and their work environments, and a redesign of the work environment itself as a socially and technologically responsive system occupying both home and office.

While investigators continue to realize promising components of the "intelligent" work environment – projectors, screens, tablets, sensors, actuators and other digital devices – our collaborative team is focused upon a novel but wholly compatible aspect of the "intelligent" work environment: the physical work space itself. The AWE concept challenges knowledge and understanding in both Architecture and Engineering by defining the "robot as a room" and the "room as a robot." This is a redefinition of what constitutes Architecture and Robotics and is not only a conceptual leap in the respective disciplines, but a fully appropriate, even necessary response to a condition in working life that is both social and technological.

The need to "program the room" both stimulates and is enabled by existing and ongoing efforts in Information Technology and "intelligent environments." More broadly, AWE challenges present understanding in both Architecture and Computer and Information Science and Engineering by recognizing computer software, networks and devices not as isolated aspects of a digital society but as constituents of an integral environmental system, far more productive and more accessible than any of its parts. In this way, the AWE concept aims to expand the vision of researchers, developers and manufacturers of information technologies to recognize the physical environment as an integral and necessary part of the dynamic interaction between people and the digital realm.

Overview of AWE

The ongoing dramatic transformation in working life, including the introduction of ever new digital technologies, presents problems and opportunities to all workers, particularly to segments of the working population that are emerging and neglected: telecommuters, flexible shift workers, single parents, elders, recent immigrants, the obese, the handicapped and, other individuals requiring special accommodations.

Continuum Robots, Figure 5
AWE concept – SLEEPING

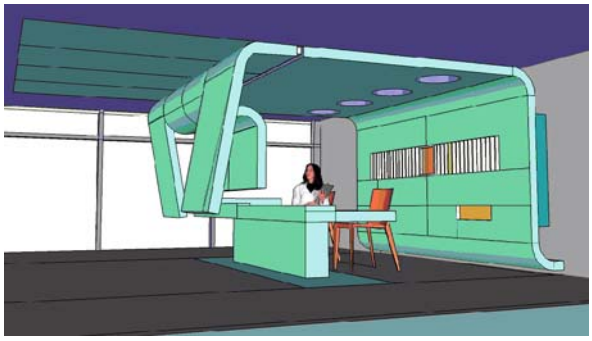

Continuum Robots, Figure 6
AWE concept – COMPOSING

This dramatic shift in the nature, place and organization of working life motivates our research which, in the simplest of terms, involves the designing, prototyping, demonstrating and evaluating of a prototypical "robot-room" with embedded Information Technologies that we call an "Animated Work Environment" (AWE) (Figs. 5 and 6).

The strength of AWE is made clearer by recognizing what it isn't: it isn't a building, or a room, or a "stand-alone" device, or a software application, or a piece of furniture. Instead, AWE is a user-friendly, programmable environment, both digital and analog, high-tech and low-tech, fitted to home and office, that users adjust along a continuum, providing the sense of being more "at home" or more "at work," more leisurely or more productive, more efficient or more innovative, while facilitating multiple activities.

In concept, AWE is envisioned as an information-rich environment featuring the ability to continuously "morph" its continuum surfaces to accommodate a wide range of user needs. At the core of this environment (though not exclusively comprising it) are smooth, continuously deformable "smart" continuum surfaces whose configuration, and hence functionality, are user-controllable. In addition to this novel aspect, AWE embodies a range of "off-the-shelf" Information Technology (IT) components: embedded commercially-available sensors that, when suitably exploited, make AWE user-friendly and intelligent; radio-frequency identification (RFID) tags that allow AWE to associate printed and digital materials; and integrated display screens, scanners, projectors, keyboards and audio speakers that make AWE useful as a total work environment programmable to suit a range of work needs and situations.

Related Work

Few precedents for AWE are found in the work environments being researched and developed by IT and the architectural design industries. The most promising of these efforts are perhaps IBM's "Blue Space" [34] and IDEO's "Q" [65], both important steps towards integrating IT and Design. Compared to our vision of AWE, however, "Blue Space" contains fewer and more timid "smart" components and a narrower range of embedded IT peripherals; and "Q," accommodates only one user who must be seated, supports a far more limited range of work activities, and cannot be reconfigured in the way AWE promises. A more enticing integration of design and IT is found in numerous efforts offered by Phillips as "Ambient Intelligence" 0; but as single products rather than complex environments, these too fall short of AWE's promise.

Additionally, while numerous research efforts in ubiquitous computing [6,7,22,23,60] are viewed by the AWE team as significant, compatible, and parallel efforts to our own research, the difference between the AWE project and these efforts is made clear by a compelling research project from University of Massachusetts titled "NeTS: Animated Spaces for the Digital Society" [60]. While this project's title and broad objective to "bridge the gap between the physical and digital worlds" sounds strikingly similar to our own, "NeTS" focuses on employing Radio Frequency Identification to link objects with their placement in a room, rather than creating, as envisioned for AWE, the physical room itself, rendered intelligent and embedded with a range of IT technologies. Our AWE research effort focuses less on developing and interfacing individual IT elements that might be embedded in AWE, and more on cultivating the dynamic, reinforcing and reconfigurable relationships between the digital and physical (i. e. environmental) realms responsive to workers and working life in a digital society.

Precedents that most significantly impact our development of AWE, a robot-room, are those few where Architectural and CISE researchers collaborate to realize intelligent physical environments. In broad theoretical terms, the AWE team is inspired by two such convergences

drawn from William Mitchell's trilogy of books on IT and the built environment [55,56,57]. The first is Mitchell's vision of the building as a computer: "The building of the near future will function more and more like large computers" [56]. The second is Mitchell's vision of the building as a robot: "Our buildings will become … robots for living in" [56].

It is worth noting that Mitchell moves easily between MIT's School of Architecture and its Media Lab, and that his vision of the building-as-robot is not altogether radical when we recognize that the first known treatise of architecture, by Vitruvius, laid claims to machine design as the proper domain of architecture [49,77]. Indeed, the "smart house" projects by Mitchell's group [48] and others, and, more significantly, the programmable, robotic wall by dECOi [21] and robotic pavilions by Kas Oosterhuis [63], hint at the promise of converging architectural design, interface design and robotic design. AWE draws from these precedents the open source, chassis/plug-in strategy of MIT's *House* [48], the seamless integration of IT and architecture suggested by the MIT/FPC *Media House* [54], and the real-time configurability of the programmable *WEB* pavilion of Oosterhuis [63].

However as "houses," "walls," and "pavilions," these significant precedents for our research, by definition, are not explicitly designed for specified human activities and so, suffer from being too diffuse; whereas AWE, operating at the smaller scale of the room, employing novel morphing surfaces, and devoted to facilitating working life, makes more of its association of new technologies, the physical environment and very particular social conditions.

Example AWE Scenario

The following scenario begins to illustrate the operation and promise of AWE.

Situation Laura, a single-mother/biologist, presents her research proposal before a committee in two weeks. Laura's proposal draws significantly from documents in digital format, notes from a recent conference, a video she made, and several books. She is working on her proposal mostly at home but also at her faculty office. Meanwhile Laura's nephew Roberto, a college student, is visiting from Latin America.

Before AWE Laura's documents are spread over the desk, computer, floor, and filing cabinet of her home office. Books are opened, face-down, to critical pages. One of these books and two of the digital papers are most important to her proposal. She's periodically disrupted by her four-year old child, Eric, who pulls materials off Laura's desk when he wants her attention. Laura senses that her nephew Roberto has a real interest in science, but she doesn't know how to sustain this interest over his visit, and, anyhow, he's on vacation – easily distracted and easily bored.

Working with AWE for the first time AWE is installed and ready to operate in Laura's living room (Fig. 5). It looks to her more like a dining room table than an office desk, a cubicle, or a work station. Laura quickly recognizes six prominent icons on AWE's work surface that read: COMPOSING, PRESENTING, COLLABORATING, MEETING, VIEWING, PLAYING.

As Laura urgently needs to work on her research proposal, she imagines COMPOSING best defines her needs and engages that icon. Quickly and steadily, four ribbons descend from the ceiling, each embedded with a computer screen arrayed at eye level when seated at the work surface (Fig. 6). The four screens light-up and a keyboard and mouse pop up from the work surface. In the short time that AWE sets itself up for COMPOSING, Laura discovers a toggle option marked SMART / PROGRAM and decides on toggling SMART mode. Nothing happens; that is, until Laura has organized her printed materials and coffee cup on AWE's work surface and sits before its array of four lit screens.

Laura takes notice of a slowly moving curtain encircling one end of AWE's work surface to provide privacy and sound dampening for COMPOSING. Laura's scientific mind correctly infers that AWE, in its simple intelligence, has recognized that she is presently seated, COMPOSING, and so is likely to require this kind of gentle enclosure to facilitate her work. She also senses that AWE has adjusted its overhead lighting to facilitate her focused work. But the desk is too low for Laura's height. Laura moves her hand towards the work surface where she sees the outline of a hand, marked RISE, at which point the surface rises to a height at which Laura is comfortable. The surface stops rising when Laura removes her hand. Laura now presses and holds the COMPOSING button to save this particular setting and then types "Laura – Research Presentation" to define this personal setting. At a later occasion, Laura can simply press COMPOSING to return AWE to this configuration.

Laura now assigns four digital documents to the four digital screens: two screens display the two most important articles, a third displays the video, and the fourth screen displays the titles of some non-digital documents Laura is using to compose the presentation – her handwritten conference notes, nine printed articles, and four books. Laura

clicks on the title of a particular printed article she's soon to need, and the screen reports to her that this article (with attached RFID tag), is located in slot-24 of AWE's physical filing system. The lit door over file slot-24 helps Laura quickly locate and retrieve the article.

Laura now touches, one at a time, the two screens displaying the two most important articles for her proposal and these, in AWE's simple intelligence, follow Laura's hand, extending and slightly twisting to a more accessible position for her. In a like manner, Laura sends the two screens displaying the video and the non-digital documents a short distance away and above her so that she is aware of, but not distracted by them.

Forty minutes into her work, Laura becomes aware of her son Eric, growing restless. But Eric, seeing his mother intently working through AWE's translucent privacy curtain, entertains himself for another five minutes. Then, the inevitable: Eric, intent on getting his mother's attention, lunges for the printed article Laura has just retrieved. At that moment, Laura toggles the icon on AWE's work surface labeled COVER/UNCOVER to COVER. A thin but sturdy surface originates from below, covering Laura's materials that had been distributed methodically by her across the work surface. The new surface provides space for Laura and Eric to color some pictures for ten minutes – their pact.

After ten minutes of coloring, Eric seems content to entertain himself for a little while longer, and Laura toggles UNCOVER to reveal the organization of her work surface again. Laura decides it's time to rehearse the first segment of her presentation. Simply engaging the PRESENTING icon establishes the new configuration (Fig. 7). (Alternatively, Laura could select INTENT INFERENCING in AWE's set-up menu, so that AWE begins configuring the PRESENTING mode when it recognizes Laura standing in a location where one would give a presentation.) In PRESENTING mode, three of AWE's display screens retract into the ceiling and the fourth screen, displaying Laura's developing *Powerpoint* presentation, spins 180 degrees to face her. Three projectors also are turned on: one projects the *Powerpoint* presentation on the wall surface next to Laura while the other two projectors display, on two adjacent walls, the visual sensation of a small lecture hall where one might make a presentation. AWE's lighting and acoustics are likewise adjusted to simulate the atmosphere of the hall.

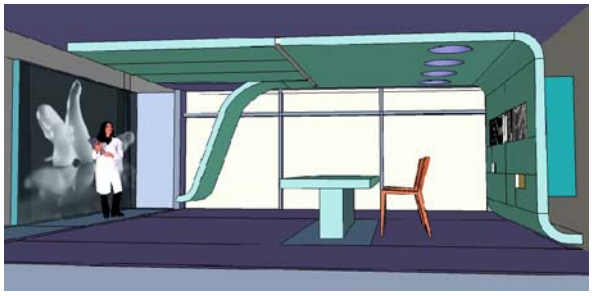

Continuum Robots, Figure 7
AWE concept – Presenting

After 30 minutes of rehearsing, Laura remembers its time she drop Eric at the day care and then, to return to her faculty office for a scheduled appointment. Before leaving the house, Laura withdraws, from a USB port in AWE, a portable storage device containing the digital information held in her home-based AWE system.

Stopped at a red traffic light on her way to the University, Laura has an epiphany about her developing proposal. Upon entering her faculty lab where AWE is also installed, Laura inserts the portable storage device into AWE to quickly match the AWE configuration she left at home. Without having to reach for a notepad and without losing the time of transcription, Laura is easily integrating her "epiphany" into the developing presentation. Laura also adjusts AWE's on-screen ATMOSPHERE settings to better approximate, for her next rehearsal, the sense of the conference room where she will finally present her proposal.

At home, Roberto employs AWE to view materials that Laura has given him access to. Among these is a recording of Laura's earlier conversation with Roberto about AWE and her developing work, presented by AWE as a video for Roberto to review and examine. Roberto, a college student, prefers lounging in an upholstered chair to sitting at a desk. Using AWE's simple intelligence, Roberto adjusts AWE's work surface from the default VIEWING configuration to a slightly tilted and lower configuration that better suits his lounging. As well, Roberto elects to bring the single screen displaying Laura's video more proximate to him, guiding the screen with his hand. Roberto saves this VIEWING figuration as, simply, "Roberto" for future retrieval.

When Laura returns from the University, Roberto excitedly announces to his aunt that using AWE, he's discovered a curious link between something she's considering in her research and research activities occurring in his native country, which he discovered on the internet.

When Roberto grows tired from all his explorations facilitated by AWE, Laura, with a simple activation of the COMPOSING option, recovers the physical organization of her materials as before, and continues developing her presentation.

In sum, AWE affords responsiveness to needs – particular and dynamic, cognitive and kinesthetic – of individual users in fluid social organizations and work patterns.

Summary

The two case studies discussed in this article demonstrate the potential for continuum structures to transform the nature of robot applications. The OctArm continuum robots described herein have successfully grasped and manipulated objects over a wide range of sizes and scales in the laboratory, demonstrating the ability to adapt their shape to that of a wide variety of payloads. Additionally, the field tests demonstrated the ability of the continuum robots to operate successfully in realistic application environments, both in air and in water, and to maintain grasp-stability under dynamic disturbances.

AWE, by recognizing the physical environment as an integral aspect of the dynamic interaction between people and the digital realm, constitutes an early, socially significant initiative featuring continuum robots by a team of "collaborative environment designers," defined by Mark Burry as "architects" working "along with their new collaborating experts … in computer science … and engineering" [11].

Future Directions

The AWE environment, featuring continuum robot surfaces, is a "look at the future" for continuum robotics. AWE is a direct extension of the activities within Clemson University's *Animated Architecture Lab*, a research and teaching body founded by the second author in which engineering, architecture and the social sciences converge. The AWE project is currently being realized, as a result of these efforts, as a working prototype at full-scale.

Our ongoing efforts with the OctArm robots focus on enhanced designs incorporating significantly higher arm strength, and on the integration of sensors to guide remote operation in congested obstacle fields. We anticipate "multi-arm" versions of the hardware being developed in the near future.

Acknowledgments

The authors acknowledge support from the US National Science Foundation under grant number IIS-0534423, and from the Defense Advanced Research Projects Agency (DARPA) Defense Sciences Office through the Space and Naval Warfare Systems Center, San Diego, Contract Number N66001-03-C-8043.

Bibliography

1. Aarts E, Marzano S (2003) The New Everyday: Views on Ambient Intelligence. 010 Publishers, Rotterdam
2. Anderson VC, Horn RC (1967) Tensor arm manipulator design. Trans ASME, vol. 67-DE-57. American Society of Mechanical Engineers, New York, pp 1–12
3. Aoki T, Ochiai A, Hirose S (2004) Study on slime robot: development of the mobile robot prototype model using bridle bellows. In: Proc. IEEE International Conference on Robotics and Automation, New Orleans, Louisiana, pp 2808–2813
4. Bailly Y, Amirat Y (2005) Modeling and Control of a Hybrid Continuum Active Catheter for Aortic Aneurysm Treatment. In: Proceedings of the IEEE International Conference on Robotics and Automation, Barcelona, Spain, pp 936–941
5. Beyer H, Holtzblatt K (1999) Contextual Design: Defining Customer-Centered Systems. Morgan Kaufman Publishers, San Francisco
6. Binder T, De Michelis G, Gervautz M, Jacucci G, Matkovic K, Psik T, Wagner I (2004) Supporting configurability in a mixed-media environment for design students. In: Thomas P (ed) Personal Ubiquitous Computing, vol. 8, Issue 5. Springer, London, pp 310–325
7. Bondarenko O, Janssen R (2005) Documents at Hand: Learning from Paper to Improve Digital Technologies. In: Proceedings of the Computer-Human-Interaction (CHI) 2005 Conference, Association for Computing Machinery. ACM, New York
8. Braganza D, Nath N, Walker ID, Dawson D (2006) Neural Network Grasping Controller for Continuum Robots. Proc. IEEE Conference on Decision and Control, San Diego, CA, pp 6445–6449
9. Braganza D, McIntyre M, Walker ID, Dawson D (2006) Whole Arm Grasping Control for Redundant Robot Manipulators. In: Proc. American Control Conference, Minneapolis, MN, pp 3194–3199
10. Buckingham R (2002) Snake arm robots, Industrial Robot. Int J 29(3):242–245
11. Burry M (2004) Convergent Design. In: Redmond J, Durling D, de Bono A (eds) Futureground Design Research Society International Conference vol 2 Proc. Monash University Press, Victoria
12. Chen G, Tu PM, Herve TR, Prelle C (2005) Design and modeling of a micro-robotic manipulator for colonoscopy. In: Proc. 5th Int Workshop on Research and Education in Mechatronics, Annecy, France, pp 109–114
13. Chrikjian GS (1992) Theory and Applications of Hyper-redundant Robotic Mechanisms. Ph D Thesis, Department of Applied Mechanics, California Institute of Technology. Pasadena
14. Chirikjian GS, Burdick JW (1991) Selected Applications and Intrinsic Kinematics of Hyper-Redundant Robots. In: Proc. 4th American Nuclear Society Topical Meeting on Robotics and Remote Systems, pp 61–70
15. Chirikjian GS, Burdick JW (1993) Design and Experiments with a 30 DOF Robot. In: Proc. IEEE International Conference on Robotics and Automation, Atlanta, pp 113–119
16. Chirikjian GS (1995) Hyper-redundant manipulator dynamics: a continuum approximation. Adv Robot 9(3):217–243
17. Chirikjian GS, Burdick JW (1994) A modal approach to hyper-redundant manipulator kinematics. IEEE Trans Robot Autom 10(3):343–354

18. Chitrakaran V, Behal A, Dawson D, Walker ID (2007) Setpoint Regulation of Continuum Robots Using a Fixed Camera. Robotica 25(5):581–586
19. Cieslak R, Morecki A (1999) Elephant trunk type elastic manipulator – a tool for bulk and liquid type materials transportation. Robotica 17:11–16
20. Csencsits M, Jones BA, McMahan W, Walker ID (2005) User interfaces for continuum robot arms. In: Proc. IEEE/RSJ Int Conf on Intelligent Robots and Systems, Edmonton, Canada, pp 3011–3018
21. deCOI (2004) Aegis Hypersurface. In: Mark Galthorpe/dECOi, "Precise Indeterminacy", Praxis: J Writ Build, Issue 6:28–45
22. Ganz A (2006) Principal Investigator, NeTS: Animated Spaces for the Digital Society: The ASPEN Architecture, NSF Award Abstract – #0434985
23. Garlan D, Siewiorek D, Smailagic A, Steenkiste P (2002) Project Aura: Toward Distraction-Free Pervasive Computing. In: Satyanarayanan M (ed) IEEE Pervasive Computing. IEEE, New York, pp 22–30
24. Gravagne IA, Walker ID (2000) Kinematic Transformations for Remotely Actuated Planar Continuum Robots. In: Proc. IEEE International Conference on Robotics and Automation, San Francisco, CA. IEEE, New York, pp 19–26
25. Gravagne IA, Walker ID (2002) Manipulability, Force and Compliance Analysis for Planar Continuum Manipulators. IEEE Trans Robot Autom 18(3):263–273
26. Gravagne IA, Rahn CD, Walker ID (2003) Large deflection dynamics and control for planar continuum robots. IEEE/ASME Trans Mechatron 8(2):299–307
27. Green KE, Gugerty LJ, Walker ID, Witte JC (2005) AWE (Animated Work Environment): Ambient Intelligence in Working Life. In: Proc. Conference on Intelligent Ambience and Well–Being (Ambience 05), Tampere, Finland, September 2005, pp 1–7 (CD-ROM proceedings)
28. Green KE, Walker ID, Gugerty LJ, Witte JC (2006) Three Robot–Rooms/The AWE Project. In: Proc. CHI 2006, Montreal, Canada, April 2006, pp 809–814
29. Greypilgrim Company http://www.greypilgrim.com. Accessed 2 Aug 1999
30. Hannan MW, Walker ID (2001) Analysis and Experiments with an Elephant's Trunk Robot. Adv Robot 15(8):847–858
31. Hannan MW, Walker ID (2003) Kinematics and the Implementation of an elephant's trunk manipulator and other continuum style robots. J Robot Syst 20(2):45–63
32. Hannan MW, Walker ID (2005) Real-Time Shape Estimation for Continuum Robots Using Vision. Robotica 23(5):645–651
33. Hirose S (1993) Biologically inspired robots. Oxford University Press, New York
34. IBM Blue Space Research (2005) http://www.research.ibm.com/bluespace
35. Ikuta K, Ichikawa H, Suzuki K, Yajima D, Multi-degree of Freedom Hydraulic Pressure Driven Safety Active Catheter. In: Proc. IEEE International Conference on Robotics and Automation, pp 4161–4166
36. Immega G (1992) Tentacle-like manipulators with adjustable tension lines. US Patent #5,317,952
37. Immega G, Antonelli K (1995) The KSI tentacle manipulator. In: Proc. IEEE Intl. Conf. Robotics and Automation, Nagoya, Japan, pp 3149–3154
38. Ivanescu M, Stoian V (1995) A variable structure controller for a tentacle manipulator. In: Proc. IEEE Intl. Conf. Robotics and Automation, Nagoya, Japan, pp 3155–3160
39. Ivanescu M, Bizdoaca N, Pana D (2003) Dynamic control for a tentacle manipulator with SMA actuators. In: Proc. IEEE Intl. Conf. Robotics and Automation, Taipei, Taiwan, pp 2079–2084
40. Ivanescu M, Popescu N, Popescu D (2005) A Variable Length Tentacle Manipulator Control System. In: Proc. IEEE International Conference on Robotics and Automation, Barcelona, Spain, pp 3274–3279
41. Jones BA, McMahan W, Walker ID (2004) Design and analysis of a novel pneumatic manipulator. In: Proc. 3rd IFAC Symposium on Mechatronic Systems, Sydney, Australia, pp 745–750
42. Jones BA, Csencsits M, McMahan W, Chitrakaran V, Grissom M, Pritts M, Rahn CD, Walker ID (2006) Grasping, manipulation, and exploration tasks with the OctArm continuum manipulator, video in Proceedings of the International Conference on Robotics and Automation, Orlando, FL
43. Jones BA, Walker ID (2006a) Three-Dimensional Modeling and Display of Continuum Robots. In: Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Beijing, China, pp 5872–5877
44. Jones BA, Walker ID (2006b) Kinematics for Multi-Section Continuum Robots. IEEE Trans Robot 22(1):43–55
45. Jones BA, Walker ID (2006c) Practical Kinematics for Real-Time Implementation of Continuum Robots. IEEE Trans Robot 22(6):1087–1099
46. Kier WM, Smith KK (1985) Tongues, tentacles and trunks: The biomechanics of movement in muscular-hydrostats. Zool. J Linn Soc 83:307–324
47. Lane DM, Davies JBC, Robinson G, O'Brien DJ, Sneddon J, Seaton E, Elfstrom E (1999) The AMADEUS dextrous subsea hand: design, modeling, and sensor processing. IEEE J Ocean Eng 24(1):96–111
48. Larson K (2004) House_n Current Projects http://architecture.mit.edu/~kll/Project%20List%20Sept-2004.pdf. Accessed 2 Jun 2008
49. McCarter R (ed) (1987) Building; Machines. Pamphlet Architecture/Princeton Architectural Press, New York
50. McMahan W, Jones BA, Walker ID, Chitrakaran V, Seshadri A, Dawson D (2004) Robotic manipulators inspired by cephalopod limbs. In: Proc. of the CDEN Design Conf., Montreal, Canada, pp 1–10
51. McMahan W, Jones BA, Walker ID (2005) Design and implementation of a multi-section continuum robot: Air-Octor. In: Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems, Edmonton, Canada, pp 3345–3352
52. McMahan W, Jones BA, Chitrakaran V, Csencsits M, Grissom M, Pritts M, Rahn CD, Walker ID (2006), Field trials and testing of the OctArm continuum manipulator. In: Proceedings of the International Conference on Robotics and Automation, Orlando, FL, USA, pp 2336–2341
53. Mehling JS, Diftler MA, Chu M, Valvo M (2006) A Minimally Invasive Tendril Robot for In-Space Inspection. In: Proc. BioRob 2006 Conference, pp 690–695
54. MIT Media Lab/FPC (2002), Media House Project: BCN, 26.09–06.10, Barcelona: FPC
55. Mitchell WJ (2000) City of Bits. MIT, Cambridge
56. Mitchell WJ (2000) e-topia. MIT, Cambridge
57. Mitchell WJ (2003) ME++. MIT, Cambridge
58. Mochiyama H, Shimemura E, Kobayashi H (1998) Shape Correspondence between a Spatial Curve and a Manipulator with

Hyper Degrees of Freedom. Proc. IEEE International Conference on Robotics and Automation, pp 161–166
59. Mochiyama H, Suzuki T (2003) Kinematics and dynamics of a cable-like hyper-flexible manipulator. ICRA, Taipei, Taiwan, pp 3672–3677
60. Niranjan RK, Ganz A (2004) Animated Space Architecture for Multimedia Experience – ASkME. In: Proceedings, Broadnets 2004, San Jose, pp 1–6
61. O. C. Robotics http://www.ocrobotics.com/mediagallery/images.htm. Accessed 30 May 2008
62. Ohno H, Hirose S (2001) Design of slim slime robot and its gait of locomotion. In: Proc. IEEE/RSJ Int Conf on Intelligent Robots and Systems, Maui, Hawaii, pp 707–715
63. Oosterhuis K (2003) Hyperbodies: Towards an E-motive architecture. Birkäuser, Basel
64. Pritts MB, Rahn CD (2004) Design of an artificial muscle continuum robot. In: Proc. IEEE Intl. Conf. Robotics and Automation, New Orleans, Louisiana, pp 4742–4746
65. 'Q' Mobile workstation concept for Steelcase (2005) http://www.ideo.com/portfolio/re.asp?x=12378. Accessed 2 Jun 2008
66. Robinson G, Davies JBC (1999) Continuum robots – a state of the art. In: Proc. IEEE Intl. Conf. Robotics and Automation, Detroit, Michigan, pp 2849–2854
67. Salisbury K, Townsend W, Ebrman B, DiPietro D (1988) Preliminary design of a whole-arm manipulation system (WAMS). In: Proc. IEEE Intl. Conf. Robotics and Automation, Philadelphia, Pennsylvania, pp 254–260
68. Sears P, Dupont P (2006) A Steerable Needle Technology Using Curved Concentric Tubes. In: Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems, Beijing, China, pp 2850–2856
69. Simaan N (2005) Snake-Like Units Using Flexible Backbones and Actuation Redundancy for Enhanced Miniaturization. In: Proc. IEEE International Conference on Robotics and Automation, Barcelona, Spain, pp 3023–3028
70. Smith KK, Kier WM (1989) Trunks, tongues, and tentacles: Moving with skeletons of muscle. Am Sci 77:28–35
71. Spong MW, Hutchinson S, Vidyasagar M (2006) Robot modeling and control. John Wiley & Sons, New York
72. Suzumori K, Iikura S, Tanaka H (1991) Development of flexible microactuator and its applications to robotic mechanisms. In: Proc. IEEE Intl. Conf. Robotics and Automation, Sacramento, California, pp 1622–1627
73. Suzumori K, Iikura S, Tanaka H (1992), Applying a flexible microactuator to robotic mechanisms. Control Syst Mag IEEE 12:21–27
74. Tatlicioglu E, Walker ID, Dawson D (2007) Dynamic Modelling for Planar Extensible Continuum Robot Manipulators. In: Proc. IEEE International Conference on Robotics and Automation, Rome, Italy, pp 1357–1362
75. Thomann F, Betemps M, Redarce T (2003) The development of a bendable colonoscopic tip. In: International Conference on Robotics and Automation, Taipei, Taiwan, pp 658–663
76. Tsukagoshi H, Kitagawa A, Segawa M (2001) Active Hose: an artificial elephant's nose with maneuverability for rescue operation. In: Proc. IEEE Intl. Conf. Robotics and Automation, Seoul, Korea, pp 2454–2459
77. Vitruvius (1985) De Architectura [On Architecture]. trans. F. Granger. Harvard University Press, Cambridge
78. Walker ID (2000) Issues in Creating 'Invertebrate' Robots. In: Proc. Conference on Adaptive Manipulation with Animals and Machines, Montreal, Canada, pp 1–6
79. Walker ID, Dawson D, Flash T, Grasso F, Hanlon R, Hochner B, Kier WM, Pagano C, Rahn CD, Zhang Q (2005) Continuum Robot Arms Inspired by Cephalopods. In: Proc. SPIE Conference on Unmanned Ground Vehicle Technology VII, Orlando, FL, pp 303–314
80. Walker ID, Carreras C, McDonnell R, Grimes G (2006) Extension Versus Bending for Continuum Robots. Int J Adv Robot Syst 3(2):171–178
81. Wilson JF, Li D, Chen Z, George RT (1993) Flexible Robot Manipulators and Grippers: Relatives of elephant Trunks and Squid Tentacles. In: Dario P (ed) Robots and Biological Systems: Towards a New Bionics? Nato Asi Series. Springer, New York, pp 474–494
82. Xu K, Simaan N (2006) Actuation Compensation for Flexible Surgical Snake-like Robots with Redundant Remote Actuation. In: Proc. IEEE International Conference on Robotics and Automation, pp 4148–4154
83. Yamada H, Hirose S (2006) Study on the 3D Shape of Active Cord Mechanism. In: Proc. IEEE International Conference on Robotics and Automation, pp 2890–2895
84. Yang J, Potratz J, Abdel-Malek K (2006) A Hyper-Redundant Continuous Robot. In: Proc. IEEE International Conference on Robotics and Automation. pp 1854–1859
85. Yang J, Pena Pitarch E, Potratz J, Beck S, Abdel-Malek K (2006) Synthesis and analysis of a flexible elephant trunk robot. Adv Robot 20(6): 631–659
86. Siciliano B, Khatib O (2008) Springer Handbook of Robotics. Chapter 11, Kinematically Redundant Manipulators. in press

Control of Non-linear Partial Differential Equations

Fatiha Alabau-Boussouira[1],
Piermarco Cannarsa[2]
[1] L.M.A.M., Université de Metz, Metz, France
[2] Dipartimento di Matematica, Università di Roma "Tor Vergata", Rome, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Controllability
Stabilization
Optimal Control
Future Directions
Bibliography

Glossary

$\mathbb{R}$ denotes the **real line**, $\mathbb{R}^n$ the n-dimensional Euclidean space, $x \cdot y$ stands for the Euclidean scalar product of $x, y \in \mathbb{R}^n$, and $|x|$ for the norm of x.

State variables quantities describing the state of a system; in this note they will be denoted by u; in the present setting, u will be either a function defined on a subset of $\mathbb{R} \times \mathbb{R}^n$, or a function of time taking its values in an Hilbert space H.

Space domain the subset of $\mathbb{R}^n$ on which state variables are defined.

Partial differential equation a differential equation containing the unknown function as well as its partial derivatives.

State equation a differential equation describing the evolution of the system of interest.

Control function an external action on the state equation aimed at achieving a specific purpose; in this note, control functions they will be denoted by f; f will be used to denote either a function defined on a subset of $\mathbb{R} \times \mathbb{R}^n$, or a function of time taking its values in an Hilbert space F. If the state equation is a partial differential equation of evolution, then a control function can be:

1. *distributed* if it acts on the whole space domain;
2. *locally distributed* if it acts on a subset of the space domain;
3. *boundary* if it acts on the boundary of the space domain;
4. *optimal* if it minimizes (together with the corresponding trajectory) a given cost;
5. *feedback* if it depends, in turn, on the state of the system.

Trajectory the solution of the state equation u_f that corresponds to a given control function f.

Distributed parameter system a system modeled by an evolution equation on an infinite dimensional space, such as a partial differential equation or a partial integro-differential equation, or a delay equation; unlike systems described by finitely many state variables, such as the ones modeled by ordinary differential equations, the information concerning these systems is "distributed" among infinitely many parameters.

$\mathbb{1}_A$ denotes the **characteristic function** of a set $A \subset \mathbb{R}^n$, that is,

$$\mathbb{1}_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \in \mathbb{R}^n \setminus A \end{cases}$$

∂_t, ∂_{x_i} denote **partial derivatives** with respect to t and x_i, respectively.

$L^2(\Omega)$ denotes the **Lebesgue space** of all real-valued square integrable functions, where functions that differ on sets of zero Lebesgue measure are identified.

$H_0^1(\Omega)$ denotes the **Sobolev space** of all real-valued functions which are square integrable together with their *first order* partial derivatives in the sense of distributions in Ω, and vanish on the boundary of Ω; similarly $H^2(\Omega)$ denotes the space of all functions which are square integrable together with their *second order* partial derivatives.

$H^{-1}(\Omega)$ denotes the dual of $H_0^1(\Omega)$.

$\mathcal{H}^{n-1}$ denotes the $(n-1)$-dimensional **Hausdorff measure**.

H denotes a **normed space**s over $\mathbb{R}$ with norm $\|\cdot\|$, as well as an **Hilbert space** with the scalar product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$.

$L^2(0, T; H)$ is the space of all square integrable functions $f\colon [0, T] \to H$; $C([0, T]; H)$ (continuous functions) and $H^1(0, T; H)$ (Sobolev functions) are similarly defined .

Given Hilbert spaces F and H, $\mathcal{L}(F, H)$ denotes the (Banach) space of all bounded linear operators $\Lambda\colon F \to H$ with norm $\|\Lambda\| = \sup_{\|x\|=1} \|\Lambda x\|$ (when $F = H$, we use the abbreviated notation $\mathcal{L}(H)$); $\Lambda^*\colon H \to F$ denotes the adjoint of Λ given by $\langle \Lambda^* u, \phi \rangle = \langle u, \Lambda \phi \rangle$ for all $u \in H$, $\phi \in F$.

Definition of the Subject

Control theory (abbreviated, CT) is concerned with several ways of influencing the evolution of a given system by an external action. As such, it originated in the nineteenth century, when people started to use mathematics to analyze the perfomance of mechanical systems, even though its roots can be traced back to the calculus of variation, a discipline that is certainly much older. Since the second half of the twentieth century its study was pursued intensively to address problems in aerospace engineering, and then economics and life sciences. At the beginning, CT was applied to systems modeled by ordinary differential equations (abbreviated, ODE). It was a couple of decades after the birth of CT—in the late sixties, early seventies—that the first attempts to control models described by a partial differential equation (abbreviated, PDE) were made. The need for such a passage was unquestionable: too many interesting applications, from diffusion phenomena to elasticity models, from fluid dynamics to traffic flows on networks and systems biology, can be modeled by a PDE.

Because of its peculiar nature, control of PDE's is a rather deep and technical subject: it requires a good knowledge of PDE theory, a field of enormous interest in its own right, as well as familiarity with the basic aspects of CT for ODE's. On the other hand, the effort put into this research direction has been really intensive. Mathemati-

cians and engineers have worked together in the construction of this theory: the results—from the stabilization of flexible structures to the control of turbulent flows—have been absolutely spectacular.

Among those who developed this subject are A. V. Balakrishnan, H. Fattorini, J. L. Lions, and D. L. Russell, but many more have given fundamental contributions.

Introduction

The basic examples of controlled partial differential equations are essentially two: the heat equation and the and the wave equation. In a bounded open domain $\Omega \subset \mathbb{R}^n$ with sufficiently smooth boundary Γ the *heat equation*

$$\partial_t u = \Delta u + f \quad \text{in } Q_T \doteq (0,T) \times \Omega \tag{1}$$

describes the evolution in time of the temperature $u(t,x)$ at any point x of the body Ω. The term $\Delta u = \partial^2_{x_1} u + \cdots + \partial^2_{x_n} u$, called the Laplacian of u, accounts for heat diffusion in Ω, whereas the additive term f represents a heat source. In order to solve the above equation uniquely one needs to add further data, such as the initial distribution u_0 and the temperature of the boundary surface Γ of Ω. The fact that, for any given data $u_0 \in L^2(\Omega)$ and $f \in L^2(Q_T)$ Eq. (1) admits a unique weak solution u_f satisfying the boundary condition

$$u = 0 \quad \text{on } \Sigma_T \doteq (0,T) \times \Gamma \tag{2}$$

and the initial condition

$$u(0,x) = u_0(x) \quad \forall x = (x_1, \ldots, x_n) \in \Omega \tag{3}$$

is well-known. So is the maximal regularity result ensuring that

$$u_f \in H^1\left(0,T; L^2(\Omega)\right) \cap C\left([0,T]; H^1_0(\Omega)\right) \cap L^2\left(0,T; H^2(\Omega)\right) \tag{4}$$

whenever $u_0 \in H^1_0(\Omega)$. If problem (1)–(3) possesses a unique solution which depends continuously on data, then we say that the problem is *well-posed*.

Similarly, the *wave equation*

$$\partial^2_t u = \Delta u + f \quad \text{in} \quad Q_T \tag{5}$$

describes the vibration of an elastic membrane (when $n = 2$) subject to a force f. Here, $u(t,x)$ denotes the displacement of the membrane at time t in x. The initial condition now concerns both initial displacement and velocity:

$$\forall x \in \Omega \quad \begin{cases} u(0,x) = u_0(x) \\ \partial_t u(0,x) = u_1(x)\,. \end{cases} \tag{6}$$

It is useful to treat the above problems as a first order *evolution equation* in a Hilbert space H

$$u'(t) = Au(t) + Bf(t) \quad t \in (0,T)\,, \tag{7}$$

where $f(t)$ takes its valued in another Hilbert space F, and $B \in \mathcal{L}(F,H)$. In this abstract set-up, the fact that (7) is related to a PDE translates into that the closed linear operator A is not defined on the whole space but only on a (dense) subspace $D(A) \subset H$, called the *domain* of A; such a property is often referred to as the *unboundedness* of A.

For instance, in the case of the heat equation (1), $H = L^2(\Omega) = F$, and A is defined as

$$\begin{cases} D(A) = H^2(\Omega) \cap H^1_0(\Omega) \\ Au = \Delta u\,, \quad \forall u \in D(A)\,, \end{cases} \tag{8}$$

whereas $B = I$.

As for the wave equation, since it is a second order differential equation with respect to t, the Hilbert space H should be given by the product $H^1_0(\Omega) \times L^2(\Omega)$. Then, problem (5) is turned into the first order equation

$$U'(t) = \mathcal{A}U(t) + Bf(t) \quad t \in (0,T)\,,$$

where

$$U = \begin{pmatrix} u \\ v \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ I \end{pmatrix}, \quad F = L^2(\Omega)\,.$$

Accordingly, $\mathcal{A}\colon D(\mathcal{A}) \subset H \to H$ is given by

$$\begin{cases} D(\mathcal{A}) = \left(H^2(\Omega) \cap H^1_0(\Omega)\right) \times H^1_0(\Omega) \\ \mathcal{A}U = \begin{pmatrix} 0 & I \\ A & 0 \end{pmatrix} U = \begin{pmatrix} v \\ Au \end{pmatrix} \quad \forall U \in D(\mathcal{A})\,, \end{cases}$$

where A is taken as in (8).

Another advantage of the abstract formulation (7) is the possibility of considering locally distributed or boundary source terms. For instance, one can reduce to the same set-up the equation

$$\partial_t u = \Delta u + \mathbb{1}_\omega f \quad \text{in } Q_T\,, \tag{9}$$

where $\mathbb{1}_\omega$ denotes the characteristic function of an open set $\omega \subset \Omega$, or the nonhomogeneus boundary condition of Dirichlet type

$$u = f \quad \text{on } \Sigma_T\,, \tag{10}$$

or Neumann type

$$\frac{\partial u}{\partial \nu} = f \quad \text{on } \Sigma_T\,, \tag{11}$$

where ν is the outward unit normal to Γ. For Eq. (9), B reduces to multiplication by $\mathbb{1}_\omega$—a bounded operator on $L^2(\Omega)$; conditions (10) and (11) can also be associated to suitable linear operators B—which, in this case, turn out to be unbounded. Similar considerations can be adapted to the wave equation (5) and to more general problems.

Having an efficient way to represent a source term is essential in control theory, where such a term is regarded as an external action, the *control function*, exercised on the *state variable* u for a purpose, of which there are two main kinds:

- *positional*: $u(t)$ is to approach a given target in X, or attain it exactly at a given time $t > 0$;
- *optimal*: the pair (u, f) is to minimize a given functional.

The first criterion leads to *approximate* or *exact controllability* problems in time t, as well as to *stabilization* problems as $t \to \infty$. Here, the main tools will be provided by certain estimates for partial differential operators that allow to study the states that can be attained by the solution of a given controlled equation. These issues will be addressed in Sects. "Controllability" and "Stabilization" for linear evolution equations. Applications to the heat and wave equations will be discussed in the same sections.

On the other hand, *optimal control problems* require analyzing the typical issues of optimizations: existence results, necessary conditions for optimality, sufficient conditions, robustness. Here, the typical problem that has been successfully studied is the Linear Quadratic Regulator that will be discussed in Sect. "Linear Quadratic Optimal Control".

Control problems for nonlinear partial differential equations are extremely interesting but harder to deal with, so the literature is less rich in results and techniques. Nevertheless, among the problems that received great attention are those of fluid dynamics, specifically the *Euler equations*

$$\partial_t u + (u \cdot \nabla)u + \nabla p = 0$$

and the *Navier–Stokes equations*

$$\partial_t u - \mu \Delta u + (u \cdot \nabla)u + \nabla p = 0$$

subject to a *boundary control* and to the incompressibility condition $\operatorname{div} u = 0$.

Controllability

We now proceed to introduce the main notions of controllability for the evolution equation (7). Later on in this section we will give interpretations for the heat and wave equations.

In a given Hilbert space H, with scalar product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$, let

$$A\colon D(A) \subset H \to H$$

be the *infinitesimal generator* of a *strongly continuous semigroup* e^{tA}, $t \geq 0$, of bounded linear operators on X. Intuitively, this amounts to saying that $u(t) \doteq \mathrm{e}^{tA} u_0$ is the unique solution of the Cauchy problem

$$\begin{cases} u'(t) = Au(t) & t \geq 0 \\ u(0) = u_0\,, \end{cases}$$

in the classical sense for $u_0 \in D(A)$, and in a suitable generalized sense for all $u_0 \in H$. Necessary and sufficient conditions in order for an unbounded operator A to be the infinitesimal generator of a strongly continuous semigroup are given by the celebrated Hille–Yosida Theorem, see, e. g. [99] and [55].

Abstract Evolution Equations

Let F be another Hilbert space (with scalar product and norm denoted by the same symbols as for H), the so-called *control space*, and let $B\colon F \to H$ be a linear operator, that we will assume to be bounded for the time being. Then, given $T > 0$ and $u_0 \in H$, for all $f \in L^2(0, T; F)$ the Cauchy problem

$$\begin{cases} u'(t) = Au(t) + Bf(t) & t \geq 0 \\ u(0) = u_0 \end{cases} \tag{12}$$

has a unique *mild solution* $u_f \in C([0, T]; H)$ given by

$$u_f(t) = \mathrm{e}^{tA} u_0 + \int_0^t \mathrm{e}^{(t-s)A} Bf(s) \quad \forall t \geq 0 \tag{13}$$

Note 1 Boundary control problems can be reduced to the same abstract form as above. In this case, however, B in (12) turns out to be an unbounded operator related to suitable fractional powers of $-A$, see, e. g., [22].

For any $t \geq 0$ let us denote by $\Lambda_t\colon L^2(0, t; F) \to H$ the bounded linear operator

$$\Lambda_t f = \int_0^t \mathrm{e}^{(t-s)A} Bf(s)\,\mathrm{d}s \quad \forall f \in L^2(0, t; F)\,. \tag{14}$$

The *attainable* (or *reachable*) set from u_0 at time t, $\mathcal{A}(u_0, t)$ is the set of all points in H of the form $u_f(t)$ for some control function f, that is

$$\mathcal{A}(u_0, t) \doteq \mathrm{e}^{tA} u_0 + \Lambda_t L^2(0, t; F)\,.$$

We introduce below the main notions of controllability for (7). Let $T > 0$.

Definition 1 System (7) is said to be:

- **exactly controllable** in time T if $\mathcal{A}(u_0, T) = H$ for all $u_0 \in H$, that is, if for all $u_0, u_1 \in H$ there is a control function $f \in L^2(0, T; F)$ such that $u_f(T) = u_1$;
- **null controllable** in time T if $0 \in \mathcal{A}(u_0, T)$ for all $u_0 \in H$, that is, if for all $u_0 \in H$ there is a control function $f \in L^2(0, T; F)$ such that $u_f(T) = 0$;
- **approximately controllable** in time T if $\mathcal{A}(u_0, T)$ is dense in H for all $u_0 \in H$, that is, if for all $u_0, u_1 \in H$ and for any $\varepsilon > 0$ there is a control function $f \in L^2(0, T; F)$ such that $\|u_f(T) - u_1\| < \varepsilon$.

Clearly, if a system is exactly controllable in time T, then it is also null and approximately controllable in time T. Although these last two notions of controllability are strictly weaker than strong controllability, for specific problems—like when A generates a strongly continuous group—some of them may coincide.

Since controllability properties concern, ultimately, the range of the linear operator Λ_T defined in (14), it is not surprising that they can be characterized in terms of the adjoint operator $\Lambda_T^*\colon H \to L^2(0, T; F)$, which is defined by

$$\int_0^T \langle \Lambda_T^* u(s), f(s)\rangle \mathrm{d}s = \langle u_0, \Lambda_T f\rangle \qquad \forall u \in H\,,\ \forall f \in L^2(0, T; F)\,.$$

Such a characterization is the object of the following theorem. Notice that the above identity and (14) yield

$$\Lambda_T^* u(s) = B^* \mathrm{e}^{(T-s)A^*} u \quad \forall s \in [0, T]\,.$$

Theorem 1 *System (7) is:*

- ***exactly controllable*** *in time T if and only if there is a constant $C > 0$ such that*

$$\int_0^T \left\| B^* e^{tA^*} u \right\|^2 dt \geq C\|u\|^2 \quad \forall u \in H\,; \tag{15}$$

- ***null controllable*** *in time T if and only if there is a constant $C > 0$ such that*

$$\int_0^T \left\| B^* e^{tA^*} u \right\|^2 dt \geq C\left\| e^{TA^*} u \right\|^2 \quad \forall u \in H\,; \tag{16}$$

- ***approximately controllable*** *in time T if and only if, for every $u \in H$,*

$$B^* e^{tA^*} u = 0 \quad t \in [0, T] \ \ a.e. \implies u = 0\,. \tag{17}$$

To benefit the reader who is more familiar with optimization theory than abstract functional analysis, let us explain, by a variational argument, why estimate (16) implies null controllability. Consider, for every $\varepsilon > 0$, the penalized problem

$$\min \left\{ J_\varepsilon(f) \,:\, f \in L^2(0, T; H) \right\},$$

where

$$J_\varepsilon(f) = \frac{1}{2}\int_0^T \|f(t)\|^2 \,\mathrm{d}t + \frac{1}{2\varepsilon}\|u_f(T)\|^2 \qquad \forall f \in L^2(0, T; H)\,.$$

Since J_ε is strictly convex, it admits a unique minimum point f_ε. Set $u_\varepsilon = u_{f_\varepsilon}$. Recalling (13) we have, By Fermat's rule,

$$0 = J_\varepsilon'(f_\varepsilon)g = \int_0^T \langle f_\varepsilon(t), g(t)\rangle \,\mathrm{d}t + \frac{1}{\varepsilon}\langle u_\varepsilon(T), \Lambda_T g\rangle \quad \forall g \in L^2(0, T; H)\,. \tag{18}$$

Therefore, passing to the adjoint of Λ_T,

$$\int_0^T \left\langle f_\varepsilon(t) + \frac{1}{\varepsilon}\big(\Lambda_T^* u_\varepsilon(T)\big)(t), g(t)\right\rangle \mathrm{d}t = 0 \qquad \forall g \in L^2(0, T; H)\,,$$

whence, owing to (14),

$$f_\varepsilon(t) = -\frac{1}{\varepsilon}\big(\Lambda_T^* u_\varepsilon(T)\big)(t) = -B^* v_\varepsilon(t) \qquad \forall t \in [0, T]\,, \tag{19}$$

where $v_\varepsilon(t) \doteq \frac{1}{\varepsilon}\mathrm{e}^{(T-t)A^*} u_\varepsilon(T)$ is the solution of the *dual problem*

$$\begin{cases} v' + A^* v = 0 & t \in [0, T] \\ v(T) = \frac{1}{\varepsilon} u_\varepsilon(T)\,. \end{cases}$$

It turns out that

$$\frac{1}{2}\int_0^T \|f_\varepsilon(t)\|^2 \,\mathrm{d}t + \frac{1}{\varepsilon}\|u_\varepsilon(T)\|^2 \leq C\|u_0\|^2 \qquad \forall \varepsilon > 0 \tag{20}$$

for some positive constant C. Indeed, observe that, in view of (19),

$$\begin{cases} \left\langle u_\varepsilon' - Au_\varepsilon + BB^* v_\varepsilon\,,\, v_\varepsilon \right\rangle = 0\,, & u_\varepsilon(0) = u_0 \\ \left\langle v_\varepsilon' + A^* v_\varepsilon\,,\, u_\varepsilon \right\rangle = 0\,, & v_\varepsilon(T) = \frac{1}{\varepsilon} u_\varepsilon(T)\,. \end{cases}$$

So,

$$\int_0^T \left[\frac{\mathrm{d}}{\mathrm{d}t} \langle u_\varepsilon \,,\, v_\varepsilon \rangle + \left\| B^* v_\varepsilon \right\|^2 \right] dt = 0 \,,$$

hence

$$\frac{1}{\varepsilon} \| u_\varepsilon(T) \|^2 + \int_0^T \left\| B^* v_\varepsilon \right\|^2 \mathrm{d}t \;=\; \langle u_0 \,,\, v_\varepsilon(0) \rangle \,. \tag{21}$$

Now, apply estimate (16) with $u = \frac{u_\varepsilon(T)}{\varepsilon}$ and note that $v_\varepsilon(T-t) = \mathrm{e}^{tA^*} \frac{u_\varepsilon(T)}{\varepsilon}$ to obtain

$$\int_0^T \left\| B^* v_\varepsilon(t) \right\|^2 \mathrm{d}t \geq C \left\| v_\varepsilon(0) \right\|^2$$

for some positive constant C. Hence, (20) follows from (21) and (19).

Finally, from (20) one deduces the existence of a weakly convergent subsequence f_{ε_j} in $L^2(0, T; F)$. Then, called f_0 the weak limit of f_{ε_j}, $u_{\varepsilon_j}(t) \to u_{f_0}(t)$ for all $t \in [0, T]$. So, owing to (20), $u_{f_0}(T) = 0$.

Heat Equation

It is not hard to see that the heat equation (9) with Dirichlet boundary conditions (2) fails to be exactly controllable. On the other hand, one can show that it is null controllable in any time $T > 0$, hence approximately controllable. Let ω be an open subset of Ω such that $\overline{\omega} \subset \Omega$.

Taking

$$H = L^2(\Omega) = F \,, \quad Bf = \mathbb{1}_\omega f \quad \forall f \in L^2(\Omega)$$

and A as in (8), one obtains that, for any $u_0 \in L^2(\Omega)$ and $f \in L^2(Q_T)$, the initial-boundary value problem

$$\begin{cases} \partial_t u = \Delta u + \mathbb{1}_\omega f & \text{in } Q_T \\ u = 0 & \text{on } \Sigma_T \\ u(0, x) = u_0(x) & x \in \Omega \end{cases} \tag{22}$$

has a unique mild solution $u_f \in C([0, T]; L^2(\Omega))$. Moreover, multiplying both sides of equation (9) by u and integrating by parts, it is easy to see that

$$\partial_{x_i} u \in L^2(Q_T) \quad \forall i = 1, \ldots, n \,. \tag{23}$$

Notice that the above property already suffices to explain why the heat equation cannot be exactly controllable: it is impossible to attain a state $u_1 \in L^2(\Omega)$ which is not compatible with (23).

On the other hand, null controllability holds true in any positive time.

Theorem 2 *Let $T > 0$ and let ω be an open subset of Ω such that $\overline{\omega} \subset \Omega$. Then the heat equation (9) with homogeneous Dirichlet boundary conditions is null controllable in time T, i. e., for every initial condition $u_0 \in L^2(\Omega)$ there is a control function $f \in L^2(Q_T)$ such that the solution u_f of (22) satisfies $u_f(T, \cdot) \equiv 0$. Moreover,*

$$\iint_{Q_T} |f|^2 dx dt \leq C_T \int_\Omega |u_0|^2 dx$$

for some positive constant C_T.

The above property is a consequence of the abstract result in Theorem 1 and of concrete estimates for solutions of parabolic equations. Indeed, in order to apply Theorem 1 one has to translate (16) into an estimate for the heat operator. Now, observing that both A and B are self-adjoint, one promptly realizes that (16) reduces to

$$\int_0^T \int_\omega |v(t, x)|^2 \mathrm{d}x \mathrm{d}t \geq C \int_\Omega |v(T, x)|^2 \mathrm{d}x \tag{24}$$

for every solution v of the problem

$$\begin{cases} \partial_t v = \Delta v & \text{in } Q_T \\ v = 0 & \text{on } \Sigma_T \,. \end{cases} \tag{25}$$

Estimate (24) is called an *observability inequality* for the heat operator for obvious reasons: problem (25) is not well-posed since the initial condition is missing. Nevertheless, if, "observing" a solution v of such a problem on the "small" cylinder $(0, T) \times \omega$, you find that it vanishes, then you can conclude that $v(T, \cdot) \equiv 0$ in the whole domain Ω. Thus, $v(0, \cdot) \equiv 0$ by backward uniqueness.

In conclusion, as elegant as the abstract approach to null controllability may be, one is confronted by the difficult task of proving observability estimates. In fact, for the heat operator there are several ways to prove inequality (24). One of the most powerful, basically due to Fursikov and Imanuvilov [65], relies on global *Carleman estimates*. Compared to other methods that can be used to derive observability, such a technique has the advantage of applying to second order parabolic operators with variable coefficients, as well as to more general operators.

Global Carleman estimates are a priori estimates in weighted norms for solutions of the problem

$$\begin{cases} \partial_t v = \Delta v + f & \text{in } Q_T \\ v = 0 & \text{on } \Sigma_T \,. \end{cases} \tag{26}$$

regardless of initial conditions. The weight function is usually of the form

$$\psi_r(t, x) \doteq \theta(t) \left(\mathrm{e}^{2r \|\phi\|_{\infty, \Omega}} - \mathrm{e}^{r\phi(x)} \right) \quad (t, x) \in Q_T \,, \tag{27}$$

where r is a positive constant, ϕ is a given function in $C^2(\overline{\Omega})$ such that

$$\nabla\phi(x) \neq 0 \quad \forall x \in \overline{\Omega}\,, \tag{28}$$

and

$$\theta(t) \doteq \frac{1}{t(T-t)} \quad 0 < t < T\,.$$

Note that

$$\theta > 0\,, \quad \theta(t) \to \infty \quad t \to 0\,, T$$

$$\psi_r > 0\,, \quad \psi_r(t,x) \to \infty \quad t \downarrow 0\,, t \uparrow T\,.$$

Using the above notations, a typical global Carleman estimate for the heat operator is the following result obtained in [65]. Let us denote by $\nu(x)$ the outword unit normal to Γ at a point $x \in \Gamma$, and by

$$\frac{\partial\phi}{\partial\nu}(x) = \nabla\phi(x)\cdot\nu(x)$$

the normal derivative of ϕ at x.

Theorem 3 *Let Ω be a bounded domain of $\mathbb{R}^n$ with boundary of class C^2, let $f \in L^2(Q_T)$, and let ϕ be a function satisfying (28). Let v be a solution of (26). Then there are positive constants r, s_0 and C such that, for any $s > s_0$,*

$$\begin{aligned} s^3 \iint_{Q_T} \theta^3(t)|v(t,x)|^2 e^{-2s\psi_r} dxdt \\ \leq C \iint_{Q_T} |f(t,x)|^2 e^{-2s\psi_r} dxdt \\ + Cs \int_0^T \theta(t)dt \\ \times \int_\Gamma \frac{\partial\phi}{\partial\nu}(x) \left|\frac{\partial v}{\partial\nu}(t,x)\right|^2 e^{-2s\psi_r}\, d\mathcal{H}^{n-1}(x) \end{aligned} \tag{29}$$

It is worth underlying that, thanks to the singular behavior of θ near 0 and T, the above result is independent of the initial value of v. Therefore, it can be applied, indifferently, to any solution of (26) as well as to any solution of the *backward problem*

$$\begin{cases} \partial_t v + \Delta v = f & \text{in } Q_T \\ v = 0 & \text{on } \Sigma_T\,. \end{cases}$$

Moreover, inequality (29) can be completed adding first and second order terms to its left-hand side, each with its own adapted power of s and θ.

Instead of trying to sketch the proof of Theorem 3, which would go beyond the scopes of this note, it is interesting to explain how it can be used to recover the observability inequality (24), which is what is needed to show that the heat equation is null controllable. The reasoning—not completely straightforward—is based on the following topological lemma, proved in [65].

Lemma 1 *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with boundary Γ of class C^k, for some $k \geq 2$, and let $\omega \subset \Omega$ be an open set such that $\overline{\omega} \subset \Omega$.*

Then there is function $\phi \in C^k(\overline{\Omega})$ such that

$$\begin{cases} (i) \quad \phi(x) = 0 \quad \text{and} \quad \frac{\partial\phi}{\partial\nu}(x) < 0 \quad \forall x \in \Gamma \\ (ii) \quad \{x \in \Omega | \nabla\phi(x) = 0\} \subset \omega\,. \end{cases} \tag{30}$$

Now, given a solution v of (25) and an open set ω such that $\overline{\omega} \subset \Omega$, let $\omega' \subset\subset \omega'' \subset\subset \omega$ be subdomains with smooth boundary. Then the above lemma ensures the existence of a function ϕ such that

$$\{x \in \Omega | \nabla\phi(x) = 0\} \subset \omega'\,.$$

"Localizing" problem (25) onto $\Omega' \doteq \Omega \setminus \omega'$ by a cutoff function $\eta \in C^\infty(\mathbb{R}^n)$ such that

$$0 \leq \eta \leq 1\,, \quad \eta \equiv 1 \quad \text{on } \mathbb{R}^n \setminus \omega''\,, \quad \eta \equiv 0 \quad \text{on } \omega'\,,$$

that is, taking $w = \eta v$, gives

$$\begin{cases} \partial_t w = \Delta w + h & \text{in } Q'_T \doteq (0,T)\times\Omega' \\ w(t,\cdot) = 0 & \text{on } \partial\Omega' = \partial\Omega \cup \partial\omega'\,, \end{cases} \tag{31}$$

with $h := -v\Delta\eta + 2\nabla\eta\cdot\nabla u$. Since $\nabla\phi \neq 0$ on Ω', Theorem 3 can be applied to w on Q'_T to obtain

$$\begin{aligned} s^3 \iint_{Q'_T} \theta^3 |w|^2 \mathrm{e}^{-2s\psi_r} \mathrm{d}x\mathrm{d}t \\ \leq C \iint_{Q'_T} |h|^2 \mathrm{e}^{-2s\psi_r} \mathrm{d}x\mathrm{d}t \\ + Cs \int_0^T \theta \mathrm{d}t \int_\Gamma \frac{\partial\phi}{\partial\nu} \left|\frac{\partial w}{\partial\nu}\right|^2 \mathrm{e}^{-2s\psi_r} d\mathcal{H}^{n-1} \\ + Cs \int_0^T \theta \mathrm{d}t \int_{\partial\omega'} \frac{\partial\phi}{\partial\nu} \left|\frac{\partial w}{\partial\nu}\right|^2 \mathrm{e}^{-2s\psi_r} d\mathcal{H}^{n-1} \\ \leq C \iint_{Q'_T} |h|^2 \mathrm{e}^{-2s\psi_r} \mathrm{d}x\mathrm{d}t \end{aligned}$$

for s sufficiently large. On the other hand, for any $0 < T_0 < T_1 < T$,

$$s^3 \iint_{Q'_T} \theta^3 |w|^2 e^{-2s\psi_r} dx dt$$
$$\geq s^3 \int_{T_0}^{T_1} dt \int_{\Omega\setminus\omega} \theta^3 |w|^2 e^{-2s\psi_r} dx dt$$
$$\geq \int_{T_0}^{T_1} dt \int_{\Omega\setminus\omega} |v|^2 dx$$

Therefore, recalling the definition of h,

$$\int_{T_0}^{T_1} dt \int_{\Omega\setminus\omega} |v|^2 dx \leq C \iint_{Q'_T} |h|^2 e^{-2s\psi_r} dx dt$$
$$\leq C \int_0^T dt \int_{\omega''\setminus\omega'} \left[|\nabla^2\eta|^2 v^2 + |\nabla\eta|^2 |\nabla v|^2\right] e^{-2s\psi_r} dx$$
$$\leq C \int_0^T dt \int_\omega |v|^2 dx + C \int_0^T dt \times \int_{\omega''\setminus\omega'} |\nabla v|^2 e^{-2s\psi_r} dx\,.$$

Now, fix $T_0 = T/3$, $T_1 = 2T/3$ and use Caccioppoli's inequality (a well-known estimate for solution of elliptic and parabolic PDE's)

$$\int_0^T dt \int_{\omega''\setminus\omega'} |\nabla v|^2 e^{-2s\psi_r}\, dx \leq C \int_0^T dt \int_\omega |v|^2 e^{-2s\psi_r}\, dx\,,$$

to conclude that

$$\int_{T/3}^{2T/3} dt \int_{\Omega\setminus\omega} |v|^2\, dx \leq C \int_0^T dt \int_\omega |v|^2\, dx$$

or

$$\int_{T/3}^{2T/3} dt \int_\Omega |v|^2\, dx \leq (1+C) \int_0^T dt \int_\omega |v|^2\, dx$$

for some constant C. Then, the dissipativity of the heat operator (that is, the fact that $\int_\Omega |v(t,x)|^2 dx$ is decreasing with respect to t) implies that

$$\int_\Omega v^2(T,x)\, dx \leq \frac{3}{T} \int_{T/3}^{2T/3} dt \int_\Omega v^2(t,x)\, dx \leq (1+C)\frac{3}{T} \int_0^T dt \int_\omega v^2(t,x)\, dx\,,$$

which is exactly (24).

Wave Equation

Compared to the heat equation, the wave equation (5) exhibits a quite different behavior from the point of view of exact controllability. Indeed, on the one hand, there is no obstruction to exact controllability since no regularizing effect is connected with wave propagation. On the other hand, due to the finite speed of propagation, exact controllability cannot be expected to hold true in arbitrary time, as null controllability does for the heat equation.

In fact, a typical result that holds true for the wave equation is the following, where a boundary control of Dirichlet type acts on a part $\Gamma_1 \subset \Gamma$, while homogeneous boundary conditions are imposed on $\Gamma_0 = \Gamma \setminus \Gamma_1$:

$$\begin{cases} \partial_t^2 u = \Delta u & \text{in } Q_T \\ u = f \mathbb{1}_{\Gamma_1} & \text{on } \Sigma_T \\ u(0,x) = u_0(x)\,, \partial_t u(0,x) = u_1(x) & x \in \Omega \end{cases} \tag{32}$$

Observe that problem (32) is well-posed taking

$$u_0 \in L^2(\Omega)\,, \quad u_1 \in H^{-1}(\Omega)$$
$$f \in L^2(0,T;L^2(\Gamma))$$
$$u \in C([0,T];L^2(\Omega)) \cap C^1([0,T];H^{-1}(\Omega))\,.$$

Theorem 4 *Let Ω be a bounded domain of $\mathbb{R}^n$ with boundary of class C^2 and suppose that, for some point $x_0 \in \mathbb{R}^n$,*

$$\begin{cases} (x-x_0)\cdot\nu(x) > 0 & \forall x \in \Gamma_1 \\ (x-x_0)\cdot\nu(x) \leq 0 & \forall x \in \Gamma_0\,. \end{cases}$$

Let

$$R = \sup_{x\in\Omega} |x - x_0|\,.$$

If $T > 2R$, then, for all $(u_0,u_1), (v_0,v_1) \in L^2(\Omega) \times H^{-1}(\Omega)$ there is a control function $f \in L^2(0,T;L^2(\Gamma))$ such that the solution u_f of (32) satisfies

$$u_f(T,x) = v_0(x)\,, \quad \partial_t u_f(T,x) = v_1(x)\,.$$

As we saw for abstract evolution equations, the above exact controllability property is proved to be equivalent to an observability estimate for the dual homogeneous problem using, for instance, the Hilbert Uniqueness Method (HUM) by J.-L. Lions [86].

Bibliographical Comments

The literature on controllability of parabolic equations and related topics is so huge, that no attempt to provide a com-

prehensive account of it would fit within the scopes of this note. So, the following comments have to be taken as a first hint for the interested reader to pursue further bibliographical research.

The theory of exact controllability for parabolic equations was initiated by the seminal paper [58] by Fattorini and Russell. Since then, it has experienced an enormous development. Similarly, the multiplier method to obtain observability inequalities for the wave equation was developed in [17,73,74,77,86]. Some fundamental early contributions were surveyed by Russell [108]. The next essential progress was made in the work by Lebeau and Robbiano [83] and then by Fursikov and Imanuvilov in a series of papers. In [65] one can find an introduction to global Carleman estimates, as well as applications to the controllability of several ODE's. In particular, the presentation of this paper as for observability inequalities and Carleman estimates for the heat operator is inspired by the last monograph. General perspectives for the understanding of global Carleman estimates and their applications to unique continuation and control problems for PDE's can be found in the works by Tataru [113,114,115,116]. Usually, the above approach requires coefficients to be sufficiently smooth. Recently, however, interesting adaptations of Carleman estimates to parabolic operators with discontinuous coefficients have been obtained in [21,82].

More recently, interest has focussed on control problems for nonlinear parabolic equations. Different approaches to controllability problems have been proposed in [57] and [44]. Then, null and approximate controllability results have been improved by Fernandez–Cara and Zuazua [61,62]. Techniques to produce insensitizing controls have been developed in [117]. These techniques have been successfully applied to the study of Navier–Stokes equations by several authors, see e. g. [63].

Fortunately, several excellent monographs are now available to help introduce the reader to this subject. For instance, the monograph by Zabczyk [121] could serve as a clean introduction to control and stabilization for finite- and infinite-dimensional systems. Moreover, [22,50,51], as well as [80,81] develop all the basic concepts of control and system theory for distributed parameter systems with special emphasis on abstract formulation. Specific references for the controllability of the wave equation by HUM can be found in [86] and [74]. More recent results related to series expansion and Ingham type methods can be found in [75]. For the control of Navier–Stokes equations the reader is referred to [64], as well as to the book by Coron [43], which contains an extremely rich collection of classical results and modern developments.

Stabilization

Stabilization of flexible structures such as beams, plates, up to antennas of satellites, or of fluids as, for instance, in aeronautics, is an important part of CT. In this approach, one wants either to derive feedback laws that will allow the system to autoregulate once they are implemented, or study the asymptotic behavior of the stabilized system i. e. determine whether convergence toward equilibrium states as times goes to infinity holds, determine its speed of convergence if necessary or study how many feedback controls are required in case of coupled systems.

Different mathematical tools have been introduced to handle such questions in the context of ODE's and then of PDE's. Stabilization of ODE's goes back to the work of Lyapunov and Lasalle. The important property is that trajectories decay along Lyapunov functions. If trajectories are relatively compact in appropriate spaces and the system is autonomous, then one can prove that trajectories converge to equilibria asymptotically. However, the construction of Lyapunov functions is not easy, in general.

This section will be concerned with some aspects of the stabilization of second order hyperbolic equations, our model problem being represented by the wave equation with distributed damping

$$\begin{cases} \partial_{tt}u - \Delta u + a(x)u_t = 0 & \text{in } \Omega \times \mathbb{R}\,, \\ u = 0 & \text{on } \Sigma = (0,\infty) \times \Gamma \\ (u, \partial_t u)(0) = (u^0, u^1) & \text{on } \Omega\,, \end{cases} \tag{33}$$

in a bounded domain $\Omega \subset \mathbb{R}^n$ with a smooth boundary Γ. For $n = 2$, $u(t,x)$ represents the displacement of point x of the membrane at time t. Therefore, equation (33) describes an elastic system. The energy of such a system is given by

$$E(t) = \frac{1}{2}\int_\Omega \left[|u_t(t,x)|^2 + |\nabla u(t,x)|^2\right] \mathrm{d}x\,.$$

When $a \geq 0$, the *feedback* term $a(x)u_t$ models friction: it produces a loss of energy through a dissipation phenomenon. More precisely, multiplying the equation in (33) by u_t and integrating by parts on Ω, it follows that

$$E'(t) = -\int_\Omega a(x)|u_t|^2\,\mathrm{d}x \leq 0\,, \quad \forall t \geq 0\,. \tag{34}$$

On the other hand, if $a \equiv 0$, then the system is *conservative*, i. e., $E(t) = E(0)$ for all $t \geq 0$.

Another well-investigated stabilization problem for the wave equation is when the feedback is localized on a part Γ_0 of the boundary Γ, that is,

$$\begin{cases} \partial_{tt}u - \Delta u = 0 & \text{in } \Omega \times \mathbb{R} \\ \frac{\partial u}{\partial \nu} + u_t = 0 & \text{on } \Sigma_0 = (0,\infty) \times \Gamma_0 \\ u = 0 & \text{on } \Sigma_1 = (0,\infty) \times (\Gamma \setminus \Gamma_0) \\ (u, \partial_t u)(0) = (u^0, u^1) \end{cases} \tag{35}$$

In this case, the dissipation relation (34) takes the form

$$E'(t) = -\int_{\Gamma_0} |u_t|^2 \, \mathrm{d}\mathcal{H}^{n-1} \le 0\,, \quad \forall t \ge 0\,.$$

In many a situation—such as to improve the quality of an acoustic hall—one seeks to reduce vibrations to a minimum: this is why stabilization is an important issue in CT. We note that the above system has a unique stationary solution—or, *equilibrium*—given by $u \equiv 0$. Stabilization theory studies all questions related to the convergence of solutions to such an equilibrium: existence of the limit, rate of convergence, different effects of nonlinearities in both displacement and velocity, effects of geometry, coupled systems, damping effects due to memory in viscoelastic materials, and so on.

System (33) is said to be:

- **strongly stable** if $E(t) \to 0$ as $t \to \infty$;
- **(uniformly) exponentially stable** if $E(t) \le Ce^{-\alpha t}E(0)$ for all $t \ge 0$ and some constants $\alpha > 0$ and $C \ge 0$, independent of u^0, u^1.

This note will focus on some of the above issues, such as geometrical aspects, nonlinear damping, indirect damping for coupled systems and memory damping.

Geometrical Aspects

A well-known property of the wave equation is the so-called *finite speed of propagation*, which means that, if the initial conditions u^0, u^1 have compact support, then the support of $u(t,\cdot)$ evolves in time at a finite speed. This explains why, for the wave equation, the geometry of Ω plays an essential role in all the issues related to control and stabilization.

The size and localization of the region in which the feedback is active is of great importance. In this paper such a region, denoted by ω, is taken as a subset of Ω of positive Lebesgue measure. More precisely, a is assumed to be continuous on $\overline{\Omega}$ and such that

$$a \ge 0 \quad \text{on} \quad \Omega \quad \text{and} \quad a \ge a_0 \quad \text{on} \quad \omega\,, \tag{36}$$

for some constant $a_0 > 0$. In this case, the feedback is said to be *distributed*. Moreover, it is said to be *globally* distributed if $\omega = \Omega$ and *locally* distributed if $\Omega \setminus \omega$ has positive Lebesgue measure.

Two main methods have been used or developed to study stabilization, namely the *multiplier method* and *microlocal analysis*. The one that gives the sharpest results is based on microlocal analysis. It goes back to the work of Bardos, Lebeau and Rauch [17], giving geodesics sufficient conditions on the region of active control for exact controllability to hold. These conditions say that each ray of geometric optics should meet the control region. Burq and Gérard [25] showed that these results hold under weaker regularity assumptions on the domain and coefficients of the operators (see also [26,27]). These geodesics conditions are not explicit, in general, but they allow to get decay estimates of the energy under very general hypotheses.

The multiplier method is an explicit method, based on energy estimates, to derive decay rates (as well as observability and exact controllability results). For boundary control and stabilization problems it was developed in the works of several authors, such as Ho [38,73], J.-L. Lions [86], Lasiecka–Triggiani, Komornik–Zuazua [76], and many others. Zuazua [123] gave an explicit geometric condition on ω for a semilinear wave equation subject to a locally distributed damping. Such a condition was then relaxed K. Liu [87] (see also [93]) who introduced the so-called piecewise multiplier method. Lasiecka and Triggiani [80,81] introduced a sharp trace regularity method which allows to estimate boundary terms in energy estimates. There also exist intermediate results between the geodesics conditions of Bardos–Lebeau–Rauch and the multiplier method, obtained by Miller [95] using differentiable escape functions.

Zuazua's multiplier geometric condition can be described as follows. If a subset O of $\overline{\Omega}$ is given, one can define an ε-neighborhood of O in $\overline{\Omega}$ as the subset of points of Ω which are at distance at most ε of O. Zuazua proved that if the set ω is such that there exists a point $x_0 \in \mathbb{R}^n$—an *observation* point—for which ω contains an ε-neighborhood of $\Gamma(x^0) = \{x \in \partial\Omega\,,\ (x - x^0)\cdot\nu(x) \ge 0\}$, then the energy decays exponentially. In this note, we refer to this condition as (MGC).

If a vanishes for instance in a neighborhood of the two poles of a ball Ω in $\mathbb{R}^n$, one cannot find an observation point x_0 such that (MGC) holds. K. Liu [87] (see also [93]) introduced a piecewise multiplier method which allows to

choose several observation points, and therefore to handle the above case.

Introduce disjoint lipschitzian domains Ω_j of Ω, $j = 1, \ldots, J$, and observation points $x^j \in \mathbb{R}^N$, $j = 1, \ldots, J$ and define

$$\gamma_j(x^j) = \{x \in \partial\Omega_j\,,\ (x - x^j) \cdot \nu_j(x) \geq 0\}$$

Here ν_j stands for the unit outward normal vector to the boundary of Ω_j. Then the piecewise multiplier geometrical condition for ω is:

$$\omega \supset \mathcal{N}_\varepsilon \left(\cup_{j=1}^J \gamma_j(x^j) \cup \left(\Omega \setminus \cup_{j=1}^J \Omega_j\right)\right) \quad \text{(PWMGC)}$$

It will be denoted by (PWMGC) condition in the sequel.

Assume now that a vanishes in a neighborhood of the two poles of a ball in $\mathbb{R}^n$. Then, one can choose two subsets Ω_1 and Ω_2 containing, respectively, the two regions where a vanishes and apply the piecewise multiplier method with $J = 2$ and with the appropriate choices of two observation points and ε. The multiplier method consists of integrating by parts expressions of the form

$$\int_t^T \int_\Omega \left(\partial_t^2 u - \Delta u + a(x) u_t\right) Mu \,\mathrm{d}x\,\mathrm{d}t = 0 \qquad \forall 0 \leq t \leq T\,,$$

where u stands for a (strong) solution of (33), with an appropriate choice of Mu. Multipliers have generally the form

$$Mu = (m(x) \cdot \nabla u + c\,u)\,\psi(x)\,,$$

where m depends on the observation points and ψ is a cut-off function. Other multipliers of the form $Mu = \Delta^{-1}(\beta u)$, where β is a cut-off function and Δ^{-1} is the inverse of the Laplacian operator with homogeneous Dirichlet boundary conditions, have also used.

The geometric conditions (MGC) or (PWMGC) serve to bound above by zero terms which cannot be controlled otherwise. One can then prove that the energy satisfies an estimate of the form

$$\int_t^T E(s)\,\mathrm{d}s \leq cE(t) + \int_t^T \left(\int_\Omega a(x)|u_t|^2 + \int_\omega |u_t|^2 \right) \mathrm{d}s \qquad \forall t \geq 0\,. \quad (37)$$

Once this estimate is proved, one can use the dissipation relation to prove that the energy satisfies integral inequalities of Gronwall type. This is the subject of the next section.

Decay Rates, Integral Inequalities and Lyapunov Techniques

The Linear Feedback Case Using the dissipation relation (34), one has

$$\int_t^T \int_\Omega a|u_t|^2\,\mathrm{d}x\,\mathrm{d}s \leq \int_t^T -E'(s)\,\mathrm{d}s \leq E(t) \qquad \forall\ 0 \leq t \leq T\,.$$

On the other hand, thanks to assumption (36) on a

$$\int_t^T \int_\omega u_t^2\,\mathrm{d}x\,\mathrm{d}s \leq \frac{1}{a_0} \int_t^T \int_\Omega a|u_t|^2\,\mathrm{d}x\,\mathrm{d}s \leq \frac{1}{a_0} E(t) \quad \forall 0 \leq t \leq T\,.$$

By the above inequalities and (37), E satisfies

$$\int_t^T E(s)\,\mathrm{d}s \leq cE(t)\,, \quad \forall 0 \leq t \leq T\,. \quad (38)$$

Since E is a nonincreasing function and thanks to this integral inequality, Haraux [71] (see also Komornik [74]) proved that E decays exponentially at infinity, that is

$$E(t) \leq E(0) \exp\left(1 - t/c\right), \quad \forall t \geq c\,. \quad (39)$$

This proof is as follows. Define

$$\phi(t) = \exp(t/c) \int_t^\infty E(s)\,\mathrm{d}s \quad \forall t \geq 0\,.$$

Thanks to (38) ϕ is nonincreasing on $[0, \infty)$, so that

$$\phi(t) \leq \phi(0) = \int_0^\infty E(s)\,\mathrm{d}s\,.$$

Using once again (38) with $t = 0$ in this last inequality and the definition of ϕ, one has

$$\int_t^\infty E(s)\,\mathrm{d}s \leq cE(0)\exp(-t/c) \quad \forall t \geq 0\,.$$

Since E is a nonnegative and nonincreasing function

$$cE(t) \leq \int_{t-c}^t E(s)\,\mathrm{d}s \leq \int_{t-c}^\infty E(s)\,\mathrm{d}s \leq cE(0)\exp(-(t-c)/c)\,,$$

so that (39) is proved.

An alternative method is to introduce a modified (or perturbed) energy E_ε which is equivalent to the natural one for small values of the parameter ε as in Komornik and

Zuazua [76]. Then one shows that this modified energy satisfies a differential Gronwall inequality so that it decays exponentially at infinity. The exponential decay of the natural energy follows then at once. In this case, the modified energy is indeed a Lyapunov function for the PDE. The natural energy cannot be in general such a Lyapunov function due to the finite speed of propagation (consider initial data which have compact support compactly embedded in $\Omega\backslash\omega$).

There are also very interesting approaches using the frequency domain approach, or spectral analysis such as developed by K. Liu [87] Z. Liu and S. Zheng [88]. In the sequel, we concentrate on the integral inequality method. This method has been generalized in several directions and we present in this note some results concerning extensions to

- nonlinear feedback
- indirect or single feedback for coupled system
- memory type feedbacks

Generalizations to Nonlinear Feedbacks Assume now that the feedback term $a(x)u_t$ in (33) is replaced by a nonlinear feedback $a(x)\rho(u_t)$ where ρ is a smooth, increasing function satisfying $v\rho(v) \geq 0$ for $v \in \mathbb{R}$, linear at ∞ and with polynomial growth close to zero, that is: $\rho(v) = |v|^p$ for $|v| \leq 1$ where $p \in (1, \infty)$.

Assume moreover that ω satisfies Zuazua's multiplier geometric condition (MGC) or Liu's piecewise multiplier method (PWMGC). Then using multipliers of the space and time variables defined as $E(s)^{(p-1)/2}\, Mu(x)$ where $Mu(x)$ are multipliers of the form described in section 5.1 and integrating by parts expressions of the form

$$\int_t^T E(s)^{(p-1)/2} \times \int_\Omega \big(\partial_t^2 u - \Delta u + a(x)\rho(u_t)\big) Mu(x)\,\mathrm{d}x\,\mathrm{d}s = 0\,,$$

one can prove that the energy E of solutions satisfies the following inequality for all $0 \leq t \leq T$

$$\int_t^T E^{(p+1)/2}(s)\,\mathrm{d}t \leq cE^{(p+1)/2}(t) + c\int_t^T E^{(p-1)/2}(s) \times \left(\int_\Omega \rho(u_t)^2 + \int_\omega |u_t|^2\right).$$

One can remark than an additional multiplicative weight in time depending on the energy has to be taken. This weight is $E^{(p-1)/2}$. Then as in the linear case, but in a more involved way, thanks to the dissipation relation

$$E'(t) = -\int_\Omega a(x)u_t\rho(u_t)\,, \tag{40}$$

one can prove that E satisfies the following nonlinear integral inequality

$$\int_t^T E^{(p+1)/2}(s)\,\mathrm{d}s \leq cE(t)\,, \quad \forall 0 \leq t \leq T\,.$$

Thanks to the fact that E is nonincreasing, a well-known result by Komornik [74] shows that E is polynomially decaying, as $t^{-2/(p-1)}$ at infinity. The above type results have been obtained by many authors under weaker form (see also [40,41,71,98,122]).

Extensions to nonlinear feedbacks without growth conditions close to zero have been studied by Lasiecka and Tataru [78], Martinez [93], W. Liu and Zuazua [89], Eller Lagnese and Nicaise [56] and Alabau–Boussouira [5]. We present the results obtained in this last reference since they provide *optimal* decay rates.

The method is as follows. Define respectively the linear and nonlinear kinetic energies

$$\begin{cases} \int_\omega |u_t|^2\,\mathrm{d}x\,, \\ \int_\Omega |a(x)\rho(u_t)|^2\,\mathrm{d}x\,, \end{cases}$$

and use a weight function in time $f(E(s))$ which is to be determined later on in an optimal way. Integrating by parts expressions of the form

$$\int_t^T f(E(s)) \int_\Omega \big(\partial_t^2 u - \Delta u + a(x)\rho(u_t)\big) Mu(x)\,\mathrm{d}x\,\mathrm{d}s = 0\,,$$

one can prove that the energy E of solutions satisfies the following inequality for all $0 \leq t \leq T$

$$\int_t^T E(s)f(E(s))\,\mathrm{d}s \leq cf(E(t)) + c\int_t^T f(E(s)) \times \left(\int_\Omega |a(x)\rho(u_t)|^2 + \int_\omega |u_t|^2\right). \tag{41}$$

The difficulty is to determine the optimal weight under general growth conditions on the feedback close to 0, in particular for cases for which the feedback decays to 0 faster than polynomials.

Assume now that the feedback satisfies

$$g(|v|) \leq |\rho(v)| \leq Cg^{-1}(|v|)\,, \quad \forall |v| \leq 1\,, \tag{42}$$

where g is continuously differentiable on $\mathbb{R}$ strictly increasing with $g(0) = 0$ and

$$\begin{cases} g \in C^2([0, r_0]) \,, r_0 \text{ sufficiently small} \,, \\ H(.) = \sqrt{.}g(\sqrt{.}) \text{ is strictly convex on } \left[0, r_0^2\right] \,, \\ g \text{ is odd} \,. \end{cases}$$

Moreover, ρ is assumed to have a linear growth at infinity. We define the *optimal* weight function f as follows.

We first extend H to a function $\hat{H}$ define on all $\mathbb{R}$

$$\hat{H}(x) = \begin{cases} H(x) \text{ if } x \in \left[0, r_0^2\right] \,, \\ +\infty \text{ otherwise} \,, \end{cases}$$

then, define a function F as follows:

$$F(y) = \begin{cases} \dfrac{\hat{H}^*(y)}{y} & \text{if } \; y \in (0, +\infty) \,, \\ 0 & \text{if } \; y = 0 \,, \end{cases}$$

where $\hat{H}^*$ stands for the convex conjugate of $\hat{H}$, that is

$$\hat{H}^*(y) = \sup_{x \in \mathbb{R}} \{x\, y - \hat{H}(x)\} \,.$$

Then the *optimal* weight function f is determined in the following way

$$f(s) = F^{-1}(s/2\beta) \quad s \in \left[0, 2\beta r_0^2\right) \,,$$

where β is of the form $\max(\eta_1, \eta_2 E(0))$, η_1 and η_2 being explicit positive constants.

One can prove that the above formulas make sense, and in particular that F is invertible and smooth. More precisely, F is twice continuously differentiable strictly increasing, one-to-one function from $[0, +\infty)$ onto $[0, r_0^2)$. Note that since the feedback is supposed to be linear at infinity, if one wants to obtain results for general growth types of the feedback, one can assume convexity of H only in a neighborhood of 0.

One can prove from (41) that there exists an (explicit) $T_0 > 0$ such that for all initial data, E satisfies the following nonlinear integral inequality

$$\int_t^T E(s) f(E(s))\, \mathrm{d}s \le T_0 E(t) \quad \forall 0 \le t \le T \,. \tag{43}$$

This inequality is proved thanks to convexity arguments as follows. Thanks to the convexity of $\hat{H}$, one can use Jensen's inequality and (42), so that

$$\int_{\Omega_t} |a(x)\rho(u_t)|^2\, \mathrm{d}x \le \gamma_1(t) \hat{H}^{-1} \times \left(\frac{1}{\gamma_1(t)} \int_\Omega a(x) u_t \rho(u_t)\, \mathrm{d}x \right)$$

In a similar way, one proves that

$$\int_{\omega_t} |u_t|^2\, \mathrm{d}x \le \gamma_2(t) \hat{H}^{-1} \left(\frac{1}{\gamma_2(t)} \int_\Omega a(x) u_t \rho(u_t)\, \mathrm{d}x \right)$$

where Ω_t and ω_t are time-dependent sets of respective Lebesgue measures $\gamma_1(t)$ and $\gamma_2(t)$ on which the velocity $u_t(t, x)$ is sufficiently small. Using the above two estimates, together with the linear growth of ρ at infinity, one proves

$$\int_t^T f(E(s)) \left(\int_\Omega |a(x)\rho(u_t)|^2 + \int_\omega |u_t|^2 \right) \le \int_t^T f(E(s)) \hat{H}^{-1} \left(\frac{1}{c} \int_\Omega a(x) u_t \rho(u_t)\, \mathrm{d}x \right)$$

Using then Young's inequality, together with the dissipation relation (40) in the above inequality, one obtains

$$\int_t^T f(E(s)) \left(\int_\Omega |a(x)\rho(u_t)|^2 + \int_\omega |u_t|^2 \right) \le C_1 \int_t^T \hat{H}^\star \left(f(E(s)) \right) \mathrm{d}s + C_2 E(t) \,, \tag{44}$$

where $C_i > 0$ $i = 1, 2$ is a constant independent of the initial data. Using the dissipation relation (40) in the above inequality, this gives for all $0 \le t \le T$

Combining this last inequality with (41) gives

$$\int_t^T E(s) f(E(s))\, \mathrm{d}s \le \beta \int_t^T (\hat{H})^\star \left(f(E(s)) \right) \mathrm{d}s + C_2 E(t)$$

where β is chosen of the form $\max(\eta_1, \eta_2 E(0))$, η_1 and η_2 being explicit positive constants to guarantee that the argument E of f stays in the domain of definition of f. Thus (43) is proved, thanks to the fact that the weight function has been chosen so that

$$\beta \hat{H}^\star(f(E(s)) = \frac{1}{2} E(s) f(E(s)) \quad \forall\, 0 \le s \,.$$

Therefore E satisfies a nonlinear integral inequality with a weight function $f(E)$ which is defined in a semi-explicit way in general cases of feedback growth.

The last step is to prove that a nonincreasing and nonnegative absolutely continuous function E satisfying a nonlinear integral inequality of the form (43) is decaying at infinity, and to establish at which rate this holds. For this, one proceeds as in [5].

Let $\eta > 0$ and $T_0 > 0$ be fixed given real numbers and F be a strictly increasing function from $[0, +\infty)$ on $[0, \eta)$, with $F(0) = 0$ and $\lim_{y \to +\infty} F(y) = \eta$.

For any $r \in (0, \eta)$, we define a function K_r from $(0, r]$ on $[0, +\infty)$ by

$$K_r(\tau) = \int_\tau^r \frac{\mathrm{d}y}{yF^{-1}(y)}\,, \tag{45}$$

and a function ψ_r which is a strictly increasing function from $[\frac{1}{F^{-1}(r)}, +\infty)$ onto $[\frac{1}{F^{-1}(r)}, +\infty)$ by

$$\psi_r(z) = z + K_r(F(\frac{1}{z})) \geq z, \quad \forall z \geq \frac{1}{F^{-1}(r)}\,, \tag{46}$$

One can prove that if E is a nonincreasing, absolutely continuous function from $[0, +\infty)$ on $[0, +\infty)$, satisfying $0 < E(0) < \eta$ and the inequality

$$\int_t^T E(s)F^{-1}(E(s))\,\mathrm{d}s \leq T_0 E(S)\,, \quad \forall\, 0 \leq t \leq T\,, \tag{47}$$

then E satisfies the following estimate:

$$E(t) \leq F\left(\frac{1}{\psi_r^{-1}(\frac{t}{T_0})}\right), \quad \forall t \geq \frac{T_0}{F^{-1}(r)}\,, \tag{48}$$

where r is any real such that

$$\frac{1}{T_0}\int_0^{+\infty} E(\tau)F^{-1}(E(\tau))\,\mathrm{d}\tau \leq r \leq \eta\,.$$

Thus, one can apply the above result to E with $\eta = r_0^2$ and show that $\lim t \to +\infty E(t) = 0$, the decay rate being given by estimate (48).

If g is polynomial close to zero, one gets back that the energy $E(t)$ decays as $t^{\frac{-2}{p-1}}$ at infinity. If $g(v)$ behaves as $\exp(-1/|v|)$ close to zero, then $E(t)$ decays as $1/(\ln(t))^2$ at infinity.

The usefulness of convexity arguments has been first pointed out by Lasiecka and Tataru [78] using Jensen's inequality and then in different ways by Martinez [93] (the weight function does not depend on the energy) and W. Liu and Zuazua [89] and Eller Lagnese and Nicaise [56]. Optimal decay rates have been obtained by Alabau–Boussouira [5,6] using a weight function determined through the theory of convex conjugate functions and Young's (named also as Fenchel–Moreau's) inequality. This argument was also used by W. Liu and Zuazua [89] in a slightly different way and combined to a Lyapunov technique. Optimality of estimates in [5] is proved in one-dimensional situation and for boundary dampings applying optimality results of Vancostenoble [119] (see also Martinez and Vancostenoble [118]).

Indirect Damping for Coupled Systems

Many complex phenomena are modeled through coupled systems. In stabilizing (or controlling) energies of the vector state, one has very often access only to some components of this vector either due to physical constraints or to cost considerations. In this case, the situation is to stabilize a full system of coupled equation through a reduced number of feedbacks. This is called indirect damping. This notion has been introduced by Russell [109] in 1993.

As an example, we consider the following system:

$$\begin{cases} \partial_t^2 u - \Delta u + \partial_t u + \alpha v = 0 \\ \partial_t^2 v - \Delta v + \alpha u = 0 \end{cases} \quad \text{in } \Omega \times \mathbb{R}\,, \quad u = 0 = v \quad \text{on } \partial\Omega \times \mathbb{R}\,. \tag{49}$$

Here, the first equation is damped through a linear distributed feedback, while no feedback is applied to the second equation. The question is to determine if this coupled system inherits any kind of stability for nonzero values of the coupling parameter α from the stabilization of the first equation only.

In the finite dimensional case, stabilization (or control) of coupled ODE's can be analyzed thanks to a powerful rank type condition named Kalman's condition. The situation is much more involved in the case of coupled PDE's.

One can show first show that the above system fails to be exponentially stable (see also [66] for related results). More generally, one can study the stability of the system

$$\begin{cases} u'' + A_1 u + Bu' + \alpha v = 0 \\ v'' + A_2 v + \alpha u = 0 \end{cases} \tag{50}$$

in a separable Hilbert space H with norm $|\cdot|$, where A_1, A_2 and B are self-adjoint positive linear operators in H. Moreover, B is assumed to be a bounded operator. So, our analysis applies to systems with internal damping supported in the whole domain Ω such as (49); the reader is referred to [1,2] for related results concerning boundary stabilization problems (see also Beyrath [23,24] for localized indirect dampings).

In light of the above observations, system (50) fails to be exponentially stable, at least when H is infinite dimensional and A_1 has a compact resolvent as in (49). Indeed it is shown in Alabau, Cannarsa and Komornik [8] that

the total energy of sufficiently smooth solutions of (50) decays polynomially at infinity whenever $|\alpha|$ is small enough but nonzero. From this result we can also deduce that any solution of (50) is strongly stable regardless of its smoothness: this fact follows by a standard density argument since the semigroup associated with (50) is a contraction semigroup.

A brief description of the key ideas of the approach developed in [2,8] is as follows. Essentially, one uses a finite iteration scheme and suitable multipliers to obtain an estimate of the form

$$\int_0^T E(u(t), v(t))\mathrm{d}t \le c \sum_{k=0}^{j} E(u^{(k)}(0), v^{(k)}(0)) \qquad \forall\, T \ge 0\,, \quad (51)$$

where j is a positive integer and E denotes the total energy of the system

$$E(u,v) = \frac{1}{2}\left(|A_1^{1/2}u|^2 + |u'|^2\right) + \frac{1}{2}\left(|A_2^{1/2}v|^2 + |v'|^2\right) + \alpha\langle u, v\rangle\,.$$

Once (51) is proved, an abstract lemma due to Alabau [1,2] shows that $E(u(t), v(t))$ decays polynomially at ∞. This abstract lemma can be stated as follows.

Let A be the infinitesimal generator of a continuous semi-group $\exp(tA)$ on an Hilbert space $\mathcal{H}$, and $D(A)$ its domain. For U^0 in $\mathcal{H}$ we set in all the sequel $U(t) = \exp(tA)U^0$ and assume that there exists a functional E defined on $C([0,+\infty), \mathcal{H})$ such that for every U^0 in $\mathcal{H}$, $E(\exp(.A))$ is a non-increasing, locally absolutely continuous function from $[0,+\infty)$ on $[0,+\infty)$. Assume moreover that there exist an integer $k \in \mathbb{N}^\star$ and nonnegative constants c_p for $p = 0, \dots k$ such that

$$\int_S^T E(U(t))\mathrm{d}t \le \sum_{p=0}^{k} c_p E(U^{(p)}(S)) \qquad \forall 0 \le S \le T\,, \forall U^0 \in D(A^k)\,. \quad (52)$$

Then the following inequalities hold for every U^0 in $D(A^{kn})$ and all $0 \le S \le T$ where n is any positive integer:

$$\int_S^T E(U(\tau))\frac{(\tau - S)^{n-1}}{(n-1)!}\mathrm{d}\tau \le c \sum_{p=0}^{kn} E(U^{(p)}(S))\,, \quad (53)$$

and

$$E(U(t)) \le c \sum_{p=0}^{kn} E(U^{(p)}(0))t^{-n} \qquad \forall t > 0\,, \quad \forall U^0 \in D(A^{kn})\,,$$

where c is a constant which depends on n.

First (53) is proved by induction on n. For $n = 1$, it reduces to the hypothesis (52). Assume now that (53) holds for n and let U^0 be given in $D(A^{k(n+1)})$. Then we have

$$\int_S^T \int_t^T E(U(\tau))\frac{(\tau - t)^{n-1}}{(n-1)!}\mathrm{d}\tau\mathrm{d}t \le c \sum_{p=0}^{kn} \int_S^T E(U^{(p)}(t))\mathrm{d}t \qquad \forall\, 0 \le S \le T\,, \forall\, U^0 \in D(A^{kn}).$$

Since U^0 is in $D(A^{k(n+1)})$ we deduce that $U^{(p)}(0) = A^p U^0$ is in $D(A^k)$ for $p \in \{0, \dots kn\}$. Hence we can apply the assumption (52) to the initial data $U^{(p)}(0)$. This together with Fubini's Theorem applied on the left hand side of the above inequality give (53) for $n + 1$. Using the property that $E(U(t))$ is non increasing in (53) we easily obtain the last desired inequality.

Applications to wave-wave, wave-Petrowsky equations and various concrete examples hold.

The above results have been studied later on by Batkai, Engel, Prüss and Schnaubelt [18] using very interesting resolvent and spectral criteria for polynomial stability of abstract semigroups. The above abstract lemma in [2] has also been generalized using interpolation theory. One should note that this integral inequality involving higher order energies of solutions is not of differential nature, unlike Haraux's and Komornik's integral inequalities. Another approach based on decoupling techniques and for slightly different abstract systems have been introduced by Ammar Khodja Bader and Ben Abdallah [12].

Spectral conditions have also been studied by Z. Liu [88] and later on by Z. Liu and Rao [90], Loreti and Rao [92] for peculiar abstract systems and in general for coupled equations only of the *same* nature (wave-wave for instance), so that a dispersion relation for the eigenvalues of the coupled system can be derived. Also these last results are given for internal stabilization only. Because of the above limitations, Z. Liu–Rao and Loreti–Rao's results are less powerful in generality than the ones given by Alabau, Cannarsa and Komornik [8] and Alabau [2]. Moreover results through energy type estimates and integral inequalities can be generalized to include nonlinear indirect dampings as shown in [7]. On the other side spectral

methods are very useful to obtain optimal decay rates provided that one can determine at which speed the eigenvalues approach the imaginary axis for high frequencies.

Memory Dampings

We consider the following model problem

$$\begin{cases} u_{tt}(t,x) - \Delta u(t,x) + \int_0^t \beta(t-s)\,\Delta u(s,x)\,\mathrm{d}s = \\ \qquad\qquad\qquad\qquad |u(t,x)|^\gamma u(t,x) \\ u(t,\cdot)_{|\partial\Omega} = 0 \\ (u(0,\cdot), u_t(0,\cdot)) = (u_0, u_1) \end{cases} \tag{54}$$

where $0 < \gamma \le \frac{2}{N-2}$ holds. The second member is a source term. The damping

$$\int_0^t \beta(t-s)\Delta u(s,x)\,\mathrm{d}s$$

is of memory type.

The energy is defined by

$$\begin{aligned} E_u(t) = {} & \frac{1}{2}\|u_t(t)\|^2_{L^2(\Omega)}\,\mathrm{d}x \\ & + \frac{1}{2}\left(1 - \int_0^t \beta(s)\,\mathrm{d}s\right)\|\nabla u(t)\|^2_{L^2(\Omega)} \\ & - \frac{1}{\gamma+2}\|u(t)\|^{\gamma+2}_{L^{\gamma+2}(\Omega)} \\ & + \frac{1}{2}\int_0^t \beta(t-s)\|\nabla u(t) - \nabla u(s)\|^2_{L^2(\Omega)}\,\mathrm{d}s \end{aligned}$$

The damping term produces dissipation of the energy, that is (for strong solutions)

$$\begin{aligned} E_u'(t) = {} & -\frac{1}{2}\beta(t)\|\nabla u(t)\|^2 \\ & + \frac{1}{2}\int_0^t \beta'(t)\|\nabla u(s) - \nabla u(t)\|^2\,\mathrm{d}s \le 0 \end{aligned}$$

One can consider more general abstract equations of the form

$$u''(t) + Au(t) - \int_0^t \beta(t-s)Au(s)\,\mathrm{d}s = \nabla F(u(t)) \qquad t \in (0,\infty) \tag{55}$$

in a Hilbert space X, where $A\colon D(A) \subset X \to X$ is an accretive self-adjoint linear operator with dense domain, and ∇F denotes the gradient of a Gâteaux differentiable functional $F\colon D(A^{1/2}) \to \mathbb{R}$. In particular, equation (54) fits into this framework as well as several other classical equations of mathematical physics such as the linear elasticity system.

We consider the following assumptions.

Assumptions (H1)

1. A is a self-adjoint linear operator on X with dense domain $D(A)$, satisfying
$$\langle Ax, x\rangle \ge M\|x\|^2 \quad \forall x \in D(A) \tag{56}$$
for some $M > 0$.
2. $\beta\colon [0,\infty) \to [0,\infty)$ is a locally absolutely continuous function such that
$$\int_0^\infty \beta(t)\mathrm{d}t < 1\ \beta(0) > 0 \quad \beta'(t) \le 0 \qquad \text{for a.e. } t \ge 0\,.$$
3. $F\colon D(A^{1/2}) \to \mathbb{R}$ is a functional such that
   1. F is Gâteaux differentiable at any point $x \in D(A^{1/2})$;
   2. for any $x \in D(A^{1/2})$ there exists a constant $c(x) > 0$ such that
   $$|DF(x)(y)| \le c(x)\|y\|, \qquad \text{for any } y \in D(A^{1/2}),$$
   where $DF(x)$ denotes the Gâteaux derivative of F in x; consequently, $DF(x)$ can be extended to the whole space X (and we will denote by $\nabla F(x)$ the unique vector representing $DF(x)$ in the Riesz isomorphism, that is, $\langle \nabla F(x), y\rangle = DF(x)(y)$, for any $y \in X$);
   3. for any $R > 0$ there exists a constant $C_R > 0$ such that
   $$\|\nabla F(x) - \nabla F(y)\| \le C_R\|A^{1/2}x - A^{1/2}y\|$$
   for all $x, y \in D(A^{1/2})$ satisfying $\|A^{1/2}x\|\,,\ \|A^{1/2}y\| \le R$.

Assumptions (H2)

1. There exist $p \in (2,\infty]$ and $k > 0$ such that
$$\beta'(t) \le -k\beta^{1+\frac{1}{p}}(t) \quad \text{for a.e. } t \ge 0$$
(here we have set $\frac{1}{p} = 0$ for $p = \infty$).
2. $F(0) = 0$, $\nabla F(0) = 0$, and there is a strictly increasing continuous function $\psi\colon [0,\infty) \to [0,\infty)$ such that $\psi(0) = 0$ and
$$|\langle \nabla F(x), x\rangle| \le \psi(\|A^{1/2}x\|)\|A^{1/2}x\|^2 \quad \forall x \in D(A^{1/2})\,.$$

Under these assumptions, global existence for sufficiently small (resp. all) initial data in the energy space can be proved for nonvanishing (resp. vanishing) source terms.

It turns out that the above energy methods based on multiplier techniques combined with linear and nonlinear integral inequalities can be extended to handle memory dampings and applied to various concrete examples such as wave, linear elastodynamic and Petrowsky equations for instance. This allows to show in [10] that exponential as well as polynomial decay of the energy holds if the kernel decays respectively exponentially or polynomially at infinity.

The method is as follows. One evaluates expressions of the form

$$\int_t^T \langle u''(s) + Au(s) - \int_0^t \beta * Au(s) - \nabla F(u(s), Mu\rangle \, \mathrm{d}s$$

where the multipliers Mu are of the form $\phi(s)(c_1(\beta * u)(s) + c_2(s)u)$ with ϕ which is a differentiable, nonincreasing and nonnegative function, and c_1 being a suitable constant, whereas c_2 may be chosen dependent on β.

Integrating by parts the resulting relations and performing some involved estimates, one can prove that for all $t_0 > 0$ and all $T \geq t \geq t_0$

$$\int_t^T \phi(s)E(s)\,\mathrm{d}s \leq C\phi(0)E(t) + \int_t^T \phi(s) \times \int_0^s \beta(s-\tau)\left\|A^{1/2}u(s) - A^{1/2}u(\tau)\right\|^2 \,\mathrm{d}\tau\,\mathrm{d}s\,,$$

If $p = \infty$, that is if the kernel β decays exponentially, one can easily bound the last term of the above estimate by $cE(t)$ thanks to the dissipation relation.

If $p \in (2, \infty)$, one has to proceed differently since the term

$$\int_t^T \phi(s) \int_0^s \beta(s-\tau)\left\|A^{1/2}u(s) - A^{1/2}u(\tau)\right\|^2 \,d\tau\,\mathrm{d}s$$

cannot be directly estimated thanks to the dissipation relation. To bound this last term, one can generalize an argument of Cavalcanti and Oquendo [37] as follows. Define, for any $m \geq 1$,

$$\varphi_m(t) := \int_0^t \beta^{1-\frac{1}{m}}(t-s)\|A^{1/2}u(s) - A^{1/2}u(t)\|^2\mathrm{d}s\,, \qquad t \geq 0\,. \quad (57)$$

Then, we have for any $T \geq S \geq 0$

$$\int_S^T E_u^{\frac{m}{p}}(t)\int_0^t \beta(t-s)\|A^{1/2}u(s) - A^{1/2}u(t)\|^2\mathrm{d}s\mathrm{d}t \leq CE_u^{\frac{p}{p+m}}(S)\left(\int_S^T E_u^{1+\frac{m}{p}}(t)\varphi_m(t)\mathrm{d}t\right)^{\frac{m}{p+m}} \quad (58)$$

for some constant $C > 0$. Suppose that, for some $m \geq 1$, the function φ_m defined in (57) is bounded. Then, for any $S_0 > 0$ there is a positive constant C such that

$$\int_S^\infty E_u^{1+\frac{m}{p}}(t)\mathrm{d}t \leq C\Big(E_u^{\frac{m}{p}}(0) + \|\varphi_m\|_\infty^{\frac{m}{p}}\Big)E_u(S) \qquad \forall\, S \geq S_0\,. \quad (59)$$

One uses this last result first with $m = 2$ noticing that φ_2 is bounded and $\phi = E^{2/p}$. This gives a first energy decay rate as $(t+1)^{-p/2}$. This estimate shows that φ_1 is bounded. Then one applies once again the last result with $m = 1$ and $\phi = E^{1/p}$. One deduces then that E decays as $(t+1)^{-p}$ which is the *optimal* decay rate expected.

Bibliographical Comments

For an introduction to the multiplier method, we refer the interested reader to the books of J.-L. Lions [86], Komornik [74] and the references therein. The celebrated result of Bardos Lebeau and Rauch is presented in [86]. A general abstract presentation of control problems for hyperbolic and parabolic equations can be found in the book of Lasiecka and Triggiani [80,81]. Results on spectral methods and the frequency domain approach can be found in the book of Z. Liu [88]. There also exists an interesting approach developed for bounded feedback operators by Haraux and extended to the case of unbounded feedbacks by Ammari and Tucsnak [11]. In this approach, the polynomial (or exponential) stability of the damped system is proved thanks to the corresponding observability for the undamped (conservative) system. Such observability results for weakly coupled undamped systems have been obtained for instance in [3].

Many other very interesting issues have been studied in connection to semilinear wave equations, see [34,123] and the references therein, and damped wave equations with nonlinear source terms [39].

Well-posedness and asymptotic properties for PDE's with memory terms have first been studied by Dafermos [53,54] for convolution kernels with past history (convolution up to $t = -\infty$), by Prüss [103] and Prüss and Propst [102] in which the efficiency of different models of dampings are compared to experiments (see also Londen Petzeltova and Prüss [91]). Decay estimates for the energy of solutions using multiplier methods combined with Lyapunov type estimates for an equivalent energy are proved in Munoz Rivera [97], Munoz Rivera and Salvatierra [96], Cavalcanti and Oquendo [37] and Giorgi Naso and Pata [67] and many other papers.

Optimal Control

As for positional control, also for optimal control problems it is convenient to adopt the abstract formulation introduced in Sect. "Abstract Evolution Equations". Let the state space be represented by the Hilbert space H, and the state equation be given in the form (12), that is

$$\begin{cases} u'(t) = Au(t) + Bf(t) & t \in [0, T] \\ u(0) = u_0 \,. \end{cases} \tag{60}$$

Recall that A is the infinitesimal of a strongly continuous semigroup, e^{tA}, in H, B is a (bounded) linear operator from F (the control space) to H, and u_f stands for the unique (mild) solution of (60) for a given control function $f \in L^2(0, T; H)$.

A typical optimal control problem of interest for PDE's is the *Bolza problem* which consists in

$$\begin{cases} \text{minimizing the } \textit{cost functional} \\ \quad J(f) \doteq \int_0^T L(t, u_f(t), f(t))\mathrm{d}t + \ell\left(u_f(T)\right) \\ \text{over all controls } \; f \in L^2(0, T; F)\,. \end{cases} \tag{61}$$

Here, T is a positive number, called the *horizon*, whereas L and ℓ are given functions, called the *running cost* and *final cost*, respectively. Such functions are usually assumed to be *bounded below*.

A control function $f_* \in L^2(0, T; F)$ at which the above minimum is attained is called an *optimal control* for problem (61) and the corresponding solution u_{f_*} of (60) is said to be an *optimal trajectory*. Alltogether, $\{u_{f_*}, f_*\}$ is called an *optimal (trajectory/control) pair*.

For problem (61) the following issues will be addressed in the sections below:

- *the existence* of controls minimizing functional J;
- *necessary conditions* that a candidate solution must satisfy;
- *sufficient conditions for optimality* provided by the dynamic programming method.

Other problems of particular interest to CT for PDE's are problems with an *infinite horizon* ($T = \infty$), problems with a *free horizon* T and a final *target*, and problems with constraints on both control variables and state variables. Moreover, the study of nonlinear variants of (60), including semilinear problems of the form

$$\begin{cases} u'(t) = Au(t) + h(t, u(t), f(t)) & t \in [0, T] \\ u(0) = u_0\,, \end{cases} \tag{62}$$

is strongly motivated by applications. The discussion of all these variants, however, will not be here pursued in detail.

Traditionally, in optimal control theory, state variables are denoted by the letters $x, y, \ldots$, whereas $u, v, \ldots$ are reserved for control variables. For notational consistency, in this section $u(\cdot)$ will still denote the state of a given system and $f(\cdot)$ a control function, while ϕ will stand for a fixed element of control space F.

Existence of Optimal Controls

From the study of finite dimensional optimization it is a familiar fact that the two essential ingredients to guarantee the existence of minima are compactness and lower semicontinuity. Therefore, it is clear that, in order to obtain a solution of the optimal control problem (60)–(61), one has to make assumptions that allow to recover such properties. The typical hypotheses that are made for this purpose are the following:

- *coercivity*: there exist constants $c_0 > 0$ and $c_1 \in \mathbb{R}$ such that

$$\ell(\phi) \geq c_1 \quad \text{and} \quad L(t, u, \phi) \geq c_0\|\phi\|^2 + c_1 \qquad \forall (t, u, \phi) \in [0, T] \times H \times F \tag{63}$$

- *convexity*: for every $(t, u) \in [0, T] \times H$

$$\phi \mapsto L(t, u, \phi) \quad \text{is convex on} \quad F\,. \tag{64}$$

Under the above hypotheses, assuming lower semicontinuity of ℓ and of the map $L(t, \cdot, \phi)$, it is not hard to show that problem (60)–(61) has at least one solution. Indeed, assumption (63) allows to show that any minimizing sequence of controls $\{f_k\}$ is bounded in $L^2(0, T; H)$. So, it admits a subsequence, still denoted by $\{f_k\}$ which converges weakly in $L^2(0, T; H)$ to some function f. Then, by linearity, $u_{f_k}(t)$ converges to $u_f(t)$ for every $t \in [0, T]$. So, using assumption (64), it follows that f is a solution of (60)–(61).

The problem becomes more delicate when the Tonelli type coercivity condition (63) is relaxed, or the state equation is nonlinear as in (62). Indeed, the convergence of $u_{f_k}(t)$ is no longer ensured, in general. So, in order to recover compactness, one has to make further assumptions, such as the compactness of e^{tA}, or structural properties of L and h. For further reading, one may consult the monographs [22,85], and [79], for problems where the running and final costs are given by quadratic forms (the so-called Linear Quadratic problem), or [84] and [59] for more general optimal control problems.

Necessary Conditions

Once the existence of a solution to problem (60)–(61) has been established, the next important step is to provide

conditions to detect a candidate solution, possibly showing that it is, in fact, optimal. By and large the optimality conditions of most common use are the ones known as Pontryagin's Maximum Principle named after the Russian mathematician L.S. Pontryagin who greatly contributed to the development of control theory, see [100,101].

So, suppose $\{u_*, f_*\}$, where $u_* = u_{f_*}$ is a candidate optimal pair and consider the so-called adjoint system

$$\begin{cases} -p'(t) = A^* p(t) + \partial_u L(t, u_*(t), f_*(t)) = 0 \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad t \in [0, T] \ \text{ a.e.} \\ p(T) = \partial \ell(u_*(T)) \,, \end{cases}$$

where $\partial_u L(t, u, \phi)$ and $\partial \ell(u)$ denote the Fréchet gradients of the maps $L(t, \cdot, \phi)$ and ℓ at u, respectively. Observe that the above is a backward linear Cauchy problem with terminal condition, which can obviously be reduced to a forward one by the change of variable $t \to T - t$. So, it admits a unique mild solution, labeled p_*, which is called the *adjoint state* associated with $\{u_*, f_*\}$.

Pontryagin's Maximum Principle states that, if $\{u_*, f_*\}$ is optimal, then

$$\langle p_*(t), Bf_*(t)\rangle + L(t, u_*(t), f_*(t)) = \min_{\phi \in F} \left[\langle p_*(t), B\phi\rangle + L(t, u_*(t), \phi)\right] \qquad t \in [0, T] \ \text{ a.e.} \tag{65}$$

The name Maximum Principle rather than Minimum Principle, as it would be more appropriate, is due to the fact that, traditionally, attention was focussed on the *maximization*—instead of minimization—of the functional in (61). Even today, in most models from economics, one is interested in maximizing payoffs, such as revenues, utility, capital and so on. In that case, (65) would still be true, with a "max" instead of a "min".

At first glance, it might be hard to understand the revelance of (65) to problem (61). To explain this, introduce the function, called the *Hamiltonian*,

$$\mathcal{H}(t, u, p) = \min_{\phi \in F} \left[\langle p, B\phi\rangle + L(t, u, \phi)\right] \qquad (t, u, p) \in [0, T] \times H \times H \,. \tag{66}$$

Then, Fermat's rule yields $B^* p + \partial_\phi L(t, u, \phi) = 0$ at every $\phi \in F$ at which the minimum in (66) is attained. Therefore, from (65) it follows that

$$B^* p_*(t) + \partial_\phi L(t, u_*(t), f_*(t)) = 0 \quad t \in [0, T] \ \text{ a.e.} \tag{67}$$

which provides a much-easier-to-use optimality condition.

There is a vast literature on necessary conditions for optimality for distributed parameter systems. The set-up that was considered above can be generalized in several ways: one can consider nonlinear state equations as in (62), nonsmooth running and finals costs, constraints on both state and control, problems with infinite horizon or exit times. Further reading and useful references on most of these extensions can be found in the aforementioned monographs [22,79,84,85], and in [59] which is mainly concerned with time optimal control problems.

Dynamic Programming

Though useful as it may be, Pontryagin's Maximum Principle remains a necessary condition. So, without further information, it does not suffice to prove the optimality of a give trajectory/control pair. Moreover, even when the map $\phi \mapsto \partial_\phi L(t, u, \phi)$ turns out to be invertible, the best result identity (67) can provide, is a representation of $f_*(t)$ in terms of $u_*(t)$ and $p_*(t)$: not enough to determine $f_*(t)$, in general.

This is why other methods to construct optimal controls have been proposed over the years. One of the most interesting ones is the so-called *dynamic programming* method (abbreviated, DP), initiated by the work of R. Bellman [20]. Such a method will be briefly described below in the set-up of distributed parameter systems.

Fix $T > 0$, s such that $0 \le s \le T$, and consider the optimal control problem

to minimize

$$J^{s,v}(f) = \int_s^T L\left(t, u_f^{s,v}(t), f(t)\right) \mathrm{d}t + \ell\left(u_f^{s,v}(T)\right) \tag{68}$$

over all control functions $f \in L^2(s, T; F)$, where $u_f^{s,v}(t)$ is the solution of the controlled system

$$\begin{cases} u'(t) = Au(t) + Bf(t) \quad t \in [s, T] \\ u(s) = v \,. \end{cases} \tag{69}$$

The *value function* U associated to (68)-(69) is the real-valued function defined by

$$U(s, v) = \inf_{f \in L^2(s,T;F)} J^{s,v}(f) \quad \forall (s, v) \in [0, T] \times H \,. \tag{70}$$

A fundamental step of DP is the following result, known as Bellman's *optimality principle*.

Theorem 5 *For any* $(s, v) \in [0, T] \times H$ *and any* $f \in L^2(s, T; F)$

$$U(s, v) \leq \int_s^r L\left(t, u_f^{s,v}(t), f(t)\right) dt + U\left(r, u_f^{s,v}(r)\right) \qquad \forall r \in [s, T].$$

Moreover, $f^*(\cdot)$ *is optimal if and only if*

$$U(s, v) = \int_s^r L\left(t, u_f^{s,v}(t), f(t)\right) dt + U\left(r, u_f^{s,v}(r)\right) \qquad \forall r \in [s, T].$$

The connection between DP and optimal control is based on the properties of the value function. Indeed, applying Bellman's optimality principle, one can show that, if U is Fréchet differentiable, then

$$\begin{cases} \partial_s U(s, v) + \langle Av, \partial_v U(s, v)\rangle + \mathcal{H}(s, v, \partial_v U(s, v)) = 0 \\ \qquad\qquad\qquad\qquad\qquad\qquad (s, v) \in (0, T) \times D(A) \\ U(T, v) = \ell(v) \quad v \in H \end{cases}$$

where $\mathcal{H}$ is the Hamiltonian defined in (66). The above equation is the celebrated *Hamilton–Jacobi equation* of DP. To illustrate its connections with the original optimal control problem, a useful formal argument—that can, however, be made rigorous—is the following. Consider a sufficiently smooth solution W of the above problem and let $(s, v) \in (0, T) \times D(A)$. Then, for any trajectory/control pair $\{u, f\}$,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} W(t, u(t)) &= \partial_s W(t, u(t)) + \langle \partial_v W(t, u(t)), Au(t) \\ &\quad + Bf(t)\rangle \\ &= \langle \partial_v W(t, u(t)), Bf(t)\rangle \\ &\quad - \mathcal{H}(t, u(t), \partial_v W(t, u(t))) \\ &\geq -L(t, u(t), f(t)) \end{aligned} \tag{71}$$

by the definition of $\mathcal{H}$. Therefore, integrating from s to T,

$$\ell(u(T)) - W(s, v) \geq -\int_s^T L(t, u(t), f(t))\mathrm{d}t,$$

whence $J^{s,v}(f) \geq W(s, v)$. Thus, taking the infimum over all $f \in L^2(s, T; F)$,

$$W(s, v) \leq U(s, v) \quad \forall (s, v) \in (0, T) \times D(A). \tag{72}$$

Now, suppose there is a control function $f_* \in L^2(s, T; F)$ such that, for all $t \in [s, T]$,

$$\begin{aligned} &\langle \partial_v W(t, u_*(t)), Bf_*(t)\rangle + L(t, u_*(t), f_*(t)) \\ &\qquad = \mathcal{H}(t, u_*(t), \partial_v W(t, u_*(t))), \end{aligned} \tag{73}$$

where $u_*(\cdot) = u_{f_*}^{s,v}(\cdot)$. Then, from (71) and (73) it follows that

$$\frac{\mathrm{d}}{\mathrm{d}t} W(t, u_*(t)) = -L(t, u(t), f(t)),$$

whence

$$W(s, v) = J^{s,v}(f_*) \geq U(s, v).$$

From the above inequality and (72) it follows that $W(s, v) = U(s, v)$ for all $(s, v) \in (0, T) \times D(A)$, hence for all $(s, v) \in (0, T) \times H$ since $D(A)$ is dense in H. So, f_* is an **optimal control**.

Note 2 The above considerations lead to the following procedure to obtain optimal an optimal trajectory:

- find a smooth solution of the Hamilton–Jacobi equation;
- for every $(t, v) \in (0, T) \times D(A)$ provide a feedback $f(t, v)$ such that

$$\begin{aligned} &\langle \partial_v W(t, v), Bf(t, v)\rangle + L(t, v, f(t, v)) \\ &\qquad = \mathcal{H}(t, v, \partial_v W(t, v)) \end{aligned}$$

- solve the so-called *closed loop equation*

$$\begin{cases} u'(t) = Au(t) + Bf(t, u(t)) \quad t \in [s, T] \\ u(s) = v \end{cases}$$

Notice that not only is trajectory u optimal, but the corresponding control f is given in feedback form as well.

Linear Quadratic Optimal Control

One of the most successful applications of DP is the so-called Linear Quadratic optimal control problem. Consider problem (68)–(69) with costs L and ℓ given by

$$L(t, u, \phi) = \langle M(t)u, u\rangle + \langle N(t)\phi, \phi\rangle \qquad \forall (T, u, \phi) \in [0, T] \times H \times F$$

and

$$\ell(u) = \langle Du, u\rangle \quad \forall u \in H,$$

where

- $M\colon [0, T] \to \mathcal{L}(H)$ is continuous, $M(t)$ is symmetric and $\langle M(t)u, u\rangle \geq 0$ for every $(t, u) \in [0, T] \times H$;
- $N\colon [0, T] \to F$ is continuous, $N(t)$ is symmetric and $\langle N(t)\phi, \phi\rangle \geq c_0|\phi|^2$ for every $(t, \phi) \in [0, T] \times F$ and some constant $c_0 > 0$;

- $D \in \mathcal{L}(H)$ is symmetric and $\langle Du, u\rangle \geq 0$ for every $u \in H$.

Then, assumptions (63) and (64) are satisfied. So, a solution to (68)–(69) does exist. Moreover, it is unique because of the strict convexity of functional $J^{s,v}$.

In order to apply DP, one computes the Hamiltonian

$$\begin{aligned}\mathcal{H}(t,u,p) &= \min_{\phi\in F}\Big[\langle p, B\phi\rangle + \langle M(t)u,u\rangle + \langle N(t)\phi,\phi\rangle\Big]\\ &= \langle M(t)u,u\rangle - \frac{1}{4}\langle BN^{-1}(t)B^*p,p\rangle\,,\end{aligned}$$

where the above minimum is attained at

$$\phi_*(t,p) = -\frac{1}{2}N^{-1}(t)B^*p\,. \tag{74}$$

Therefore, the Hamilton–Jacobi equation associated to the problem is

$$\begin{cases}\partial_s W(s,v) + \langle Av, \partial_v W(s,v)\rangle + \langle M(s)v,v\rangle\\ \quad -\frac{1}{4}\langle BN^{-1}(s)B^*\partial_v W(s,v), \partial_v W(s,v)\rangle = 0\\ \qquad\qquad\qquad\qquad \forall (s,v)\in(0,T)\times D(A)\\ w(T,v) = \langle Dv,v\rangle \quad \forall v\in H\end{cases}$$

It is quite natural to search a solution of the above problem in the form

$$W(s,v) = \langle P(s)v,v\rangle \quad \forall (s,v)\in[0,T]\times H\,,$$

with $P\colon [0,T]\to\mathcal{L}(H)$ continuous, symmetric and such that $\langle P(t)u,u\rangle \geq 0$. Substituting into the Hamilton–Jacobi equation yields

$$\begin{cases}\langle P'(s)v,v\rangle + \langle [A^*P(s)+P(s)A]v,v\rangle + \langle M(s)v,v\rangle\\ \quad -\langle BN^{-1}(s)B^*P(s)v, P(s)v\rangle = 0\\ \qquad\qquad\qquad\qquad \forall (s,v)\in(0,T)\times D(A)\\ \langle P(T)v,v\rangle = \langle Dv,v\rangle \quad \forall v\in H\end{cases}$$

Therefore, P must be a solution of the so-called *Riccati equation*

$$\begin{cases}P'(s) + A^*P(s) + P(s)A + M(s)\\ \quad -P(s)BN^{-1}(s)B^*P(s) = 0 \qquad \forall s\in(0,T)\\ P(T) = D\end{cases}$$

Once a solution $P(\cdot)$ of the Riccati equation is known, the procedure described in Note 2 can be applied. Indeed, recalling (74) and the fact that $\partial_v W(t,v) = 2P(t)v$, one concludes that $f(t,v) = -N^{-1}(t)B^*P(t)v$ is a feedback law. So, solving the closed loop equation

$$\begin{cases}u'(t) = [A - BN^{-1}(t)B^*P(t)]u(t) \quad t\in(s,T)\\ u(s) = v\end{cases}$$

one obtains the unique optimal trajectory of problem (68)–(69).

In sum, by DP one reduces the original Linear Quadratic optimal control problem to the problem of finding the solution of the Riccati equation, which is easier to solve than the Hamilton–Jacobi equation.

Bibliographical Comments

Different variants of the Riccati equation have been successfully studied by several authors in connection with different state equations and cost functionals, including boundary control problems and problems for other functional equations, see [22,79] and the references therein. Sometimes, the solution of the Riccati equation related to a linearized model provides feedback stabilization for nonlinear problems as in [104].

Unfortunately, the DP method is hard to implement for general optimal control problems, because of several obstructions: nonsmoothness of solutions to Hamilton–Jacobi equations, selection problems that introduce discontinuities, unboundedness of the coefficients, numerical complexity. Besides the Linear Quadratic case, the so-called Linear Convex case is the other example that can be studied by DP under fairly general conditions, see [14]. For nonlinear optimal control problems some of the above difficulties have been overcome extending the notion of viscosity solutions to infinite dimensional spaces, see [45,46,47,48,49], see also [28,29,30,31,32,33] and [112]. Nevertheless, finding additional ideas to make a generalized use of DP for distributed parameter systems possible, remains a challenging problem for the next generations.

Future Directions

In addition to all considerations spread all over this article on promising developments of recent—as well as established—research lines, a few additional topics deserve to be mentioned.

The one subject that has received the highest attention, recently, is that of numerical approximation of control problems, from the point of view of both controllability and optimal control. Here the problem is that, due to high frequency spurious numerical solutions, stable algorithms for solving initial-boundary value problems do not necessarily yield convergent algorithms for computing controls. This difficulty is closely related to the existence of concentrated numerical solutions that escape the observation mechanisms. Nevertheless, some interesting results have been obtained so far, see, e. g., [124,125].

Several interesting results for nonlinear control problems have been obtained by the *return method*, developed initially by Coron [42] for a stabilization problem. This and other techniques have then been applied to fluid models ([68,69]), the Korteweg–de Vries equation ([105,106,107]), and Schrödinger type equations ([19]), see also [43] and the references therein. It seems likely that these ideas, possibly combined with other techniques like Carleman estimates as in [70], will lead to new exiting results in the years to come.

A final comment on null controllability for *degenerate parabolic equations* is in order. Indeed, many problems that are relevant for applications are described by parabolic equation equations in divergence form

$$\partial_t u = \nabla\cdot(A(x)\nabla u)+b(x)\cdot\nabla u+c(t,x)u+f \quad \text{in} \quad Q_T\,,$$

or in the general form

$$\partial_t u = \mathrm{Tr}\,[A(x)\nabla^2 u]+b(x)\cdot\nabla u+c(t,x)u+f \quad \text{in} \quad Q_T\,,$$

where $A(x)$ is a symmetric matrix, positive definite in Ω but possibly singular on Γ. For instance, degenerate parabolic equations arise in fluid dynamics as suitable transformations of the Prandtl equations, see, e. g., [94]. They can also be obtained as Kolmogorov equations of diffusions processes on domains that are invariant for stochastic flows, see, e. g., [52]. The latter interpretation explains why they have been applied to biological problems, such as gene frequency models for population genetics (see, e. g., the Wright–Fischer model studied in [111]).

So far, null controllability properties of degenerate parabolic equations have been fully understood only in dimension one: for some kind of degeneracy, null controllability holds true (see [36] and [9]), but, in general, one can only expect regional null controllability (see [35]). Since very little is known on null controllability for degenerate parabolic equations in higher space dimensions, it is conceivable that such a topic will provide interesting problems for future developments.

Bibliography

1. Alabau F (1999) Stabilisation frontière indirecte de systèmes faiblement couplés. C R Acad Sci Paris Sér I 328:1015–1020
2. Alabau F (2002) Indirect boundary stabilization of weakly coupled systems. SIAM J Control Optim 41(2):511–541
3. Alabau-Boussouira F (2003) A two-level energy method for indirect boundary observability and controllability of weakly coupled hyperbolic systems. SIAM J Control Optim 42(3):871–906
4. Alabau-Boussouira F (2004) A general formula for decay rates of nonlinear dissipative systems. C R Math Acad Sci Paris 338:35–40
5. Alabau-Boussouira F (2005) Convexity and weighted integral inequalities for energy decay rates of nonlinear dissipative hyperbolic systems. Appl Math Optim 51(1):61–105
6. Alabau-Boussouira F (2006) Piecewise multiplier method and nonlinear integral inequalities for Petrowsky equations with nonlinear dissipation. J Evol Equ 6(1):95–112
7. Alabau-Boussouira F (2007) Asymptotic behavior for Timoshenko beams subject to a single nonlinear feedback control. NoDEA 14(5–6):643–669
8. Alabau F, Cannarsa P, Komornik V (2002) Indirect internal damping of coupled systems. J Evol Equ 2:127–150
9. Alabau-Boussouira F, Cannarsa P, Fragnelli G (2006) Carleman estimates for degenerate parabolic operators with applications to null controllability. J Evol Equ 6:161–204
10. Alabau-Boussouira F, Cannarsa P, Sforza D (2008) Decay estimates for second order evolution equations with memory. J Funct Anal 254(5):1342–1372
11. Ammari K, Tucsnak M (2001) Stabilization of second order evolution equations by a class of unbounded feedbacks. ESAIM Control Optim Calc Var 6:361–386
12. Ammar-Khodja F, Bader A, Benabdallah A (1999) Dynamic stabilization of systems via decoupling techniques. ESAIM Control Optim Calc Var 4:577–593
13. Barbu V (2003) Feedback stabilization of Navier–Stokes equations. ESAIM Control Optim Calc Var 9:197–206
14. Barbu V Da Prato G (1983) Hamilton Jacobi equations in Hilbert spaces. Pitman, London
15. Barbu V, Lasiecka I, Triggiani R (2006) Tangential boundary stabilization of Navier–Stokes equations. Mem Amer Math Soc 181(852):128
16. Barbu V, Triggiani R (2004) Internal stabilization of Navier–Stokes equations with finite-dimensional controllers. Indiana Univ Math J 53(5):1443–1494
17. Bardos C, Lebeau G, Rauch R (1992) Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary. SIAM J Control Optim 30:1024–1065
18. Bátkai A, Engel KJ, Prüss J, Schnaubelt R (2006) Polynomial stability of operator semigroups. Math Nachr 279(13–14):1425–1440
19. Beauchard K (2005) Local controllability of a 1-D Schödinger equation. J Math Pures Appl 84(7):851–956
20. Bellman R (1957) Dynamic Programming. Princeton University Press, Princeton
21. Benabdallah A, Dermenjian Y, Le Rousseau J (2007) Carleman estimates for the one-dimensional heat equation with a discontinuous coefficient and applications to controllability and an inverse problem. J Math Anal Appl 336(2):865–887
22. Bensoussan A, Da Prato G, Delfour MC, Mitter SK (1993) Representation and Control of Infinite Dimensional Systems. Systems and Control: Foundations and applications, Birkhäuser, Boston
23. Beyrath A (2001) Stabilisation indirecte localement distribué de systèmes faiblement couplés. C R Acad Sci Paris Sér I Math 333(5):451–456
24. Beyrath A (2004) Indirect linear locally distributed damping of coupled systems. Bol Soc Parana Mat 22(2):17–34
25. Burq N, Gérard P (1997) Condition nécessaire et suffisante pour la contrôlabilité exacte des ondes. C R Acad Sci Paris Sér I Math 325(7):749–752

26. Burq N, Hitrik M (2007) Energy decay for damped wave equations on partially rectangular domains. Math Res Lett 14(1):35–47
27. Burq N, Lebeau G (2001) Mesures de défaut de compacité, application au système de Lamé. Ann Sci École Norm Sup (4) 34(6):817–870
28. Cannarsa P (1989) Regularity properties of solutions to Hamilton–Jacobi equations in infinite dimensions and nonlinear optimal control. Diff Integral Equ 2:479–493
29. Cannarsa P, Da Prato G (1990) Some results on non-linear optimal control problems and Hamilton–Jacobi equations in infinite dimensions. J Funct Anal 90:27–47
30. Cannarsa P, Gozzi F, Soner HM (1991) A boundary value problem for Hamilton–Jacobi equations in Hilbert spaces. Applied Math Optim 24:197–220
31. Cannarsa P, Gozzi F, Soner HM (1993) A dynamic programming approach to nonlinear boundary control problems of parabolic type. J Funct Anal 117:25–61
32. Cannarsa P, Di Blasio G (1995) A direct approach to infinite dimensional Hamilton–Jacobi equations and applications to convex control with state constraints. Differ Integral Equ 8:225–246
33. Cannarsa P, Tessitore ME (1996) Infinite dimensional Hamilton–Jacobi equations and Dirichlet boundary control problems of parabolic type. SIAM J Control Optim 34:1831–1847
34. Cannarsa P, Komornik V, Loreti P (2002) One-sided and internal controllability of semilinear wave equations with infinitely iterated logarithms. Discret Contin Dyn Syst 8(3):745–756
35. Cannarsa P, Martinez P, Vancostenoble J (2004) Persistent regional null controllability for a class of degenerate parabolic equations. Commun Pure Appl Anal 3(4):607–635
36. Cannarsa P, Martinez P, Vancostenoble (2005) Null Controllability of degenerate heat equations. Adv Differ Equ 10(2):153–190
37. Cavalcanti MM, Oquendo HP (2003) Frictional versus viscoelastic damping in a semilinear wave equation. SIAM J Control Optim 42:1310–1324
38. Chen G (1981) A note on boundary stabilization of the wave equation. SIAM J Control Optim 19:106–113
39. Chueshov I, Lasiecka I, Toundykov D (2008) Long-term dynamics of semilinear wave equation with nonlinear localized interior damping and a source term of critical exponent. Discret Contin Dyn Syst 20(3):459–509
40. Conrad F, Rao B (1993) Decay of solutions of the wave equation in a star-shaped domain with nonlinear boundary feedback. Asymptot Anal 7:159–177
41. Conrad F, Pierre (1994) Stabilization of second order evolution equations by unbounded nonlinear feedback. Ann Inst H Poincaré Anal Non Linéaire 11(5):485–515, Asymptotic Anal 7:159–177
42. Coron JM (1992) Global asymptotic stabilization for controllable systems without drift. Math Control Signals Syst 5(3):295–312
43. Coron JM (2007) Control and nonlinearity. Mathematical surveys and monographs vol 136, Providence, RI: xiv+426
44. Coron JM, Trélat E (2004) Global steady-state controllability of one-dimensional semilinear heat equations. SIAM J Control Optim 43(2):549–569
45. Crandall MG, Lions PL (1985) Hamilton Jacobi equation in infinite dimensions I: Uniqueness of viscosity solutions. J Funct Anal 62:379–396
46. Crandall MG, Lions PL (1986) Hamilton Jacobi equation in infinite dimensions II: Existence of viscosity solutions. J Funct Anal 65:368–425
47. Crandall MG, Lions PL (1986) Hamilton Jacobi equation in infinite dimensions III. J Funct Anal 68:214–247
48. Crandall MG, Lions PL (1990) Hamilton Jacobi equation in infinite dimensions IV: Hamiltonians with unbounded linear terms. J Funct Anal 90:237–283
49. Crandall MG, Lions PL (1991) Hamilton Jacobi equation in infinite dimensions V: Unbounded linear terms and *B*-continuous solutions. J Funct Anal 97:417–465
50. Curtain RF, Weiss G (1989) Well posedness of triples of operators (in the sense of linear systems theory). Control and Estimation of Distributed Parameter Systems (Vorau, 1988). Internat. Ser. Numer. Math., vol. 91, Birkhäuser, Basel
51. Curtain RF, Zwart H (1995) An introduction to infinite-dimensional linear systems theory. Texts in Applied Mathematics, vol 21. Springer, New York
52. Da Prato G, Frankowska H (2007) Stochastic viability of convex sets. J Math Anal Appl 333(1):151–163
53. Dafermos CM(1970) Asymptotic stability inviscoelasticity. Arch Ration Mech Anal 37:297–308
54. Dafermos CM (1970) An abstract Volterra equation with applications to linear viscoelasticity. J Differ Equ 7:554–569
55. Engel KJ, R Nagel R (2000) One-parameter semigroups for linear evolution equations. Springer, New York
56. Eller M, Lagnese JE, Nicaise S (2002) Decay rates for solutions of a Maxwell system with nonlinear boundary damping. Comp Appl Math 21:135–165
57. Fabre C, Puel J-P, Zuazua E (1995) Approxiamte controllability of the semilinear heat equation. Proc Roy Soc Edinburgh Sect A 125(1):31–61
58. Fattorini HO, Russell DL (1971) Exact controllability theorems for linear parabolic equations in one space dimension. Arch Rat Mech Anal 4:272–292
59. Fattorini HO (1998) Infinite Dimensional Optimization and Control theory. Encyclopedia of Mathematics and its Applications, vol 62. Cambridge University Press, Cambridge
60. Fattorini HO (2005) Infinite dimensional linear control systems. North-Holland Mathematics Studies, vol 201. Elsevier Science B V, Amsterdam
61. Fernández-Cara E, Zuazua E (2000) Null and approximate controllability for weakly blowing up semilinear heat equations. Ann Inst H Poincaré Anal Non Linéaire 17(5):583–616
62. Fernández-Cara E, Zuazua E (2000) The cost approximate controllability for heat equations: The linear case. Adv Differ Equ 5:465–514
63. Fernández-Cara E, Guerrero S, Imanuvilov OY, Puel J-P (2004) Local exact controllability of the Navier–Stokes system. J Math Pures Appl 83(9–12):1501–1542
64. Fursikov A (2000) Optimal control of distributed systems. Theory and applications. Translation of Mathematical Monographs, vol 187. American Mathematical Society, Providence
65. Fursikov A, Imanuvilov OY (1996) Controllability of evolution equations. Lecture Notes, Research Institute of Mathematics, Seoul National University, Seoul
66. Gibson JS (1980) A note on stabilization of infinite dimensional linear oscillators by compact linear feedback. SIAM J Control Optim 18:311–316

67. Giorgi C, Naso MG, Pata V (2005) Energy decay of electromagnetic systems with memory. Math Models Methods Appl Sci 15(10):1489–1502
68. Glass O (2000) Exact boundary controllability of 3-D Euler equation. ESAIM Control Optim Calc Var 5:1–44
69. Glass O (2007) On the controllability of the 1-D isentropic Euler equation. J Eur Math Soc (JEMS) 9(3):427–486
70. Guerrero S (2007) Local controllability to the trajectories of the Navier–Stokes system with nonlinear Navier-slip boundary conditions. ESAIM Control Optim Calc Var 12(3):484–544
71. Haraux A (1978) Semi-groupes linéaires et équations d'évolution linéaires périodiques. Publication du Laboratoire d'Analyse Numérique no 78011. Université Pierre et Marie Curie, Paris
72. Haraux A (1989) Une remarque sur la stabilisation de certains systèmes du deuxième ordre en temps. Portugal Math 46(3):245–258
73. Ho LF (1986) Observabilité frontière de l'équation des ondes. C R Acad Sci Paris Sér I Math 302(12):443–446
74. Komornik V (1994) Exact Controllability and Stabilization. The Multiplier Method. Collection RMA, vol 36. Masson–John Wiley, Paris–Chicester
75. Komornik V, Loreti P (2005) Fourier series in control theory. Springer, New York
76. Komornik V, Zuazua E (1990) A direct method for the boundary stabilization of the wave equation. J Math Pures Appl 69:33–54
77. Lasiecka I, Lions J-L, Triggiani R (1986) Nonhomogeneous boundary value problems for second order hyperbolic operators. J Math Pures Appl 65(2):149–192
78. Lasiecka I, Tataru D (1993) Uniform boundary stabilization of semilinear wave equation with nonlinear boundary damping. Differ Integral Equ 8:507–533
79. Lasiecka I, Triggiani R (1991) Differential and algebraic Riccati equations with application to boundary/point control problems: continuous theory and approximation theory. Lecture Notes in Control & Inform Sci, vol 164. Springer, Berlin
80. Lasiecka I, Triggiani R (2000) Control theory for partial differential equations: continuous and approximation theories. I. Encyclopedia of Mathematics and its Applications, vol 74. Cambridge University Press, Cambridge
81. Lasiecka I, Triggiani R (2000) Control theory for partial differential equations: continuous and approximation theories. II. Encyclopedia of Mathematics and its Applications, vol 75. Cambridge University Press, Cambridge
82. Le Rousseau J (2007) Carleman estimates and controllability results for the one-dimensional heat equation with BV coefficients. J Differ Equ 233(2):417–447
83. Lebeau G, Robbiano L (1995) Exact control of the heat equation. Comm Partial Differ Equ 20(1–2):335–356
84. Li X, Yong J (1995) Optimal control of infinite dimensional systems. Systems & Control: Foundations & Applications. Birkhäuser, Boston
85. Lions J-L (1971) Optimal control of systems governed by partial differential equations. Springer, New-York
86. Lions J-L (1988) Contrôlabilité exacte et stabilisation de systèmes distribués I-II. Masson, Paris
87. Liu K (1997) Locally distributed control and damping for the conservative systems. SIAM J Control Optim 35:1574–1590
88. Liu Z, Zheng S (1999) Semigroups associated with dissipative systems. Chapman Hall CRC Research Notes in Mathematics, vol 398. Chapman Hall/CRC, Boca Raton
89. Liu WJ, Zuazua E (1999) Decay rates for dissipative wave equations. Ric Mat 48:61–75
90. Liu Z, Rao R (2007) Frequency domain approach for the polynomial stability of a system of partially damped wave equations. J Math Anal Appl 335(2):860–881
91. Londen SO, Petzeltová H, Prüss J (2003) Global well-posedness and stability of a partial integro-differential equation with applications to viscoelasticity. J Evol Equ 3(2):169–201
92. Loreti P, Rao B (2006) Optimal energy decay rate for partially damped systems by spectral compensation. SIAM J Control Optim 45(5):1612–1632
93. Martinez P (1999) A new method to obtain decay rate estimates for dissipative systems with localized damping. Rev Mat Complut 12:251–283
94. Martinez P, Raymond J-P, Vancostenoble J (2003) Regional null controllability for a linearized Crocco type equation. SIAM J Control Optim 42(2):709–728
95. Miller L (2002) Escape function conditions for the observation, control, and stabilization of the wave equation. SIAM J Control Optim 41(5):1554–1566
96. Muñoz Rivera JE, Peres Salvatierra A (2001) Asymptotic behaviour of the energy in partially viscoelastic materials. Quart Appl Math 59:557–578
97. Muñoz Rivera JE (1994) Asymptotic behaviour in linear viscoelasticity. Quart Appl Math 52:628–648
98. Nakao M (1996) Decay of solutions of the wave equation with a local nonlinear dissipation. Math Ann 305:403–417
99. Pazy A (1968) Semigroups of linear operators and applications to partial differential equations. Springer Berlin
100. Pontryagin LS (1959) Optimal regulation processes. Uspehi Mat Nauk 14(1):3–20
101. Pontryagin LS, Boltyanskii VG, Gamkrelidze RV, Mishchenko EF (1962) The mathematical theory of optimal processes. Interscience Publishers John Wiley & Sons, Inc., New York–London, translated from the Russian by Trirogoff KN, edited by Neustadt LW
102. Propst G, Prüss J (1996) On wave equation with boundary dissipation of memory type. J Integral Equ Appl 8:99–123
103. Prüss J (1993) Evolutionary integral equations and applications. Monographs in Mathematics, vol 87. Birkhäuser Verlag, Basel
104. Raymond J-P (2006) Feedback boundary stabilization of the two-dimensional Navier–Stokes equations. SIAM J Control Optim 45(3):790–828
105. Rosier L (1997) Exact boundary controllability for the Korteweg–de Vries equation on a bounded domain. ESAIM Control Optim Calc Var 2:33–55
106. Rosier L (2000) Exact boundary controllability for the linear Korteweg–de Vries equation on the half-line. SIAM J Control Optim 39(2):331–351
107. Rosier L, Zhang BY (2006) Global stabilization of the generalized Korteweg–de Vries equation posed on a finite domain. SIAM J Control Optim 45(3):927–956
108. Russell DL (1978) Controllability and stabilizability theorems for linear partial differential equations: recent progress and open questions. SIAM Rev 20(4):639–739
109. Russell DL (1993) A general framework for the study of indirect damping mechanisms in elastic systems. J Math Anal Appl 173(2):339–358

110. Salamon D (1987) Infinite-dimensional linear systems with unbounded control and observation: a functional analytic approach. Trans Am Math Soc 300(2):383–431
111. Shimakura N (1992) Partial differential operators of elliptic type. Translations of Mathematical Monographs, vol 99-American Mathematical Society, Providence
112. Tataru D (1992) Viscosity solutions for the dynamic programming equations. Appl Math Optim 25:109–126
113. Tataru D (1994) A-priori estimates of Carleman's type in domains with boundary. J Math Pures Appl 75:355–387
114. Tataru D (1995) Boundary controllability for conservative P.D.E. Appl Math Optim 31:257–295
115. Tataru D (1996) Carleman estimates and unique continuation near the boundary for P.D.E.'s. J Math Pures Appl 75:367–408
116. Tataru D (1997) Carleman estimates, unique continuation and controllability for anizotropic PDE's. Optimization methods in partial differential equations, South Hadley MA 1996. Contemp Math vol 209. Am Math Soc pp 267–279, Providence
117. de Teresa L (2000) Insensitizing controls for a semilinear heat equation. Comm Partial Differ Equ 25:39–72
118. Vancostenoble J, Martinez P (2000) Optimality of energy estimates for the wave equation with nonlinear boundary velocity feedbacks. SIAM J Control Optim 39:776–797
119. Vancostenoble J (1999) Optimalité d'estimation d'énergie pour une équation des ondes amortie. C R Acad Sci Paris, 328, série I, pp 777–782
120. Vitillaro E (2002) Global existence for the wave equation with nonlinear boundary damping and source terms. J Differ Equ 186(1):259–298
121. Zabczyk J (1992) Mathematical control theory: an introduction. Birkhäuser, Boston
122. Zuazua E (1989) Uniform stabilization of the wave equation by nonlinear feedbacks. SIAM J Control Optim 28:265–268
123. Zuazua E (1990) Exponential decay for the semilinear wave equation with locally distributed damping. Comm Partial Differ Equ 15:205–235
124. Zuazua E (2006) Control and numerical approximation of the heat and wave equations. In: Sanz-Solé M, Soria J, Juan Luis V, Verdera J (eds) Proceedings of the International Congress of Mathematicians, vol I, II, III, European Mathematical Society, Madrid, pp 1389–1417
125. Zuazua E (2006) Controllability and observability of partial differential equations: Some results and open problems. In: Dafermos CM, Feireisl E (eds) Handbook of differential equations: evolutionary differential equations, vol 3. Elsevier/North-Holland, Amsterdam, pp 527–621

Cooperative Games

Roberto Serrano[1,2]
[1] Department of Economics, Brown University, Providence, USA
[2] IMDEA-Social Sciences, Madrid, Spain

Article Outline

Glossary
Definition of the Subject
Introduction
Cooperative Games
The Core
The Shapley Value
Future Directions
Bibliography

Glossary

Game theory Discipline that studies strategic situations.

Cooperative game Strategic situation involving coalitions, whose formation assumes the existence of binding agreements among players.

Characteristic or coalitional function The most usual way to represent a cooperative game.

Solution concept Mapping that assigns predictions to each game.

Core Solution concept that assigns the set of payoffs that cannot be improved upon by any coalition.

Shapley value Solution concept that assigns the average of marginal contributions to coalitions.

Definition of the Subject

Cooperative game theory It is one of the two counterparts of game theory. It studies the interactions among coalitions of players. Its main question is this: Given the sets of feasible payoffs for each coalition, what payoff will be awarded to each player? One can take a positive or normative approach to answering this question, and different solution concepts in the theory lean towards one or the other.

Core It is a solution concept that assigns to each cooperative game the set of payoffs that no coalition can improve upon or block. In a context in which there is unfettered coalitional interaction, the core arises as a good positive answer to the question posed in cooperative game theory. In other words, if a payoff does not belong to the core, one should not expect to see it as the prediction of the theory if there is full cooperation.

Shapley value It is a solution that prescribes a single payoff for each player, which is the average of all marginal contributions of that player to each coalition he or she is a member of. It is usually viewed as a good normative answer to the question posed in cooperative game theory. That is, those who contribute more to the groups that include them should be paid more.

Although there were some earlier contributions, the official date of birth of game theory is usually taken to be 1944, year of publication of the first edition of the *Theory*

of Games and Economic Behavior, by John von Neumann and Oskar Morgenstern [42]. The core was first proposed by Francis Ysidro Edgeworth in 1881 [13], and later reinvented and defined in game theoretic terms in [14]. The Shapley value was proposed by Lloyd Shapley in his 1953 Ph D dissertation [37]. Both the core and the Shapley value have been applied widely, to shed light on problems in different disciplines, including economics and political science.

Introduction

Game theory is the study of games, also called strategic situations. These are decision problems with multiple decision makers, whose decisions impact one another. It is divided into two branches: non-cooperative game theory and cooperative game theory. The actors in non-cooperative game theory are individual players, who may reach agreements only if they are self-enforcing. The non-cooperative approach provides a rich language and develops useful tools to analyze games. One clear advantage of the approach is that it is able to model how specific details of the interaction among individual players may impact the final outcome. One limitation, however, is that its predictions may be highly sensitive to those details. For this reason it is worth also analyzing more abstract approaches that attempt to obtain conclusions that are independent of such details. The cooperative approach is one such attempt, and it is the subject of this article.

The actors in cooperative game theory are coalitions, that is, groups of players. For the most part, two facts, that coalitions can form and that each coalition has a feasible set of payoffs available to its members, are taken as given. Given the coalitions and their sets of feasible payoffs as primitives, the question tackled is the identification of final payoffs awarded to each player. That is, given a collection of feasible sets of payoffs, one for each coalition, can one predict or recommend a payoff (or set of payoffs) to be awarded to each player? Such predictions or recommendations are embodied in different solution concepts.

Indeed, one can take several approaches to answering the question just posed. From a positive or descriptive point of view, one may want to get a prediction of the likely outcome of the interaction among the players, and hence, the resulting payoff be understood as the natural consequence of the forces at work in the system. Alternatively, one can take a normative or prescriptive approach, set up a number of normative goals, typically embodied in axioms, and try to derive their logical implications. Although authors sometimes disagree on the classification of the different solution concepts according to these two criteria – as we shall see, the understanding of each solution concept is enhanced if one can view it from very distinct approaches –, in this article we shall exemplify the positive approach with the core and the normative approach with the Shapley value. While this may oversimplify the issues, it should be helpful to a reader new to the subject.

The rest of the article is organized as follows. Sect. "Cooperative Games" introduces the basic model of a cooperative game, and discusses its assumptions as well as the notion of solution concepts. Sect. "The Core" is devoted to the core, and Sect. "The Shapley Value" to the Shapley value. In each case, some of the main results for each of the two are described, and examples are provided. Sect. "Future Directions" discusses some directions for future research.

Cooperative Games

Representations of Games. The Characteristic Function

Let us begin by presenting the different ways to describe a game. The first two are the usual ways employed in non-cooperative game theory.

The most informative way to describe a game is called its *extensive form*. It consists of a game tree, specifying the timing of moves for each player and the information available to each of them at the time of making a move. At the end of each path of moves, a final outcome is reached and a payoff vector is specified. For each player, one can define a *strategy*, i. e., a complete contingent plan of action to play the game. That is, a strategy is a function that specifies a feasible move each time a player is called upon to make a move in the game.

One can abstract from details of the interaction (such as timing of moves and information available at each move), and focus on the concept of strategies. That is, one can list down the set of strategies available to each player, and arrive at the *strategic* or *normal form* of the game. For two players, for example, the normal form is represented in a bimatrix table. One player controls the rows, and the other the columns. Each cell of the bimatrix is occupied with an ordered pair, specifying the payoff to each player if each of them chooses the strategy corresponding to that cell.

One can further abstract from the notion of strategies, which will lead to the *characteristic function form* of representing a game. From the strategic form, one makes assumptions about the strategies used by the complement of a coalition of players to determine the feasible payoffs for the coalition (see, for example, the derivations in [7,42]). This is the representation most often used in cooperative game theory.

Thus, here are the primitives of the basic model in cooperative game theory. Let $N = \{1, \ldots, n\}$ be a finite set of players. Each non-empty subset of N is called a *coalition*. The set N is referred to as the *grand coalition*. For each coalition S, we shall specify a set $V(S) \subset \mathbb{R}^{|S|}$ containing $|S|$-dimensional payoff vectors that are feasible for coalition S. This is called the characteristic function, and the pair (N, V) is called a *cooperative game*. Note how a reduced form approach is taken because one does not explain what strategic choices are behind each of the payoff vectors in $V(S)$. In addition, in this formulation, it is implicitly assumed that the actions taken by the complement coalition (those players in $N \setminus S$) cannot prevent S from achieving each of the payoff vectors in $V(S)$. There are more general models in which these sorts of externalities across coalitions are considered, but we shall ignore them in this article.

Assumptions on the Characteristic Function

Some of the most common technical assumptions made on the characteristic function are the following:

(1) For each $S \subseteq N$, $V(S)$ is closed. Denote by $\partial V(S)$ the boundary of $V(S)$. Hence, $\partial V(S) \subseteq V(S)$.
(2) For each $S \subseteq N$, $V(S)$ is comprehensive, i. e., for each $x \in V(S)$, $\{x\} - \mathbb{R}^{|S|}_{+} \subseteq V(S)$.
(3) For each $x \in \mathbb{R}^{|S|}$,

$$\partial V(S) \cap \left(\{x\} + \mathbb{R}^{|S|}_{+}\right)$$

is bounded.
(4) For each $S \subseteq N$, there exists a continuously differentiable representation of $V(S)$, i. e., a continuously differentiable function $g_S : \mathbb{R}^{|S|} \to \mathbb{R}$ such that

$$V(S) = \{x \in \mathbb{R}^{|S|} | g_S(x) \leq 0\} .$$

(5) For each $S \subseteq N$, $V(S)$ is non-leveled, i. e., for every $x \in \partial V(S)$, the gradient of g_S at x is positive in all its coordinates.

With the assumptions made, $\partial V(S)$ is its Pareto frontier, i. e., the set of vectors $x_S \in V(S)$ such that there does not exist $y_S \in V(S)$ satisfying that $y_i \geq x_i$ for all $i \in S$ with at least one strict inequality.

Other assumptions usually made relate the possibilities available to different coalitions. Among them, a very important one is *balancedness*, which we shall define next:

A collection $\mathcal{T}$ of coalitions is balanced if there exists a set of weights $w(S) \in [0, 1]$ for each $S \in \mathcal{T}$ such that for every $i \in N$, $\sum_{S \in \mathcal{T}, S \supseteq \{i\}} w(S) = 1$. One can think of these weights as the fraction of time that each player devotes to each coalition he is a member of, with a given coalition representing the same fraction of time for each player. The game (N, V) is balanced if $x_N \in V(N)$ whenever $(x_S) \in V(S)$ for every S in a balanced collection $\mathcal{T}$. That is, the grand coalition can always implement any "time-sharing arrangement" that the different subcoalitions may come up with.

The characteristic function defined so far is often referred to as a *non-transferable utility (NTU)* game. A particular case is the *transferable utility (TU)* game case, in which for each coalition $S \subseteq N$, there exists a real number $v(S)$ such that

$$V(S) = \left\{x \in \mathbb{R}^{|S|} : \sum_{i \in S} x_i \leq v(S)\right\} .$$

Abusing notation slightly, we shall denote a TU game by (N, v). In the TU case there is an underlying nummeraire – money – that can transfer utility or payoff at a one-to-one rate from one player to any other. Technically, the theory of NTU games is far more complex: it uses convex analysis and fixed point theorems, whereas the TU theory is based on linear inequalities and combinatorics.

Solution Concepts

Given a characteristic function, i. e., a collection of sets $V(S)$, one for each S, the theory formulates its predictions on the basis of different *solution concepts*. We shall concentrate on the case in which the grand coalition forms, that is, cooperation is totally successful. Of course, solution concepts can be adapted to take care of the case in which this does not happen.

A *solution* is a mapping that assigns a set of payoff vectors in $V(N)$ to each characteristic function game (N, V). Thus, a solution in general prescribes a set, which can be empty, or a singleton (when it assigns a unique payoff vector as a function of the fundamentals of the problem). The leading set-valued cooperative solution concept is the core, while one of the most used single-valued ones is the Shapley value for TU games.

There are several criteria to evaluate the reasonableness or appeal of a cooperative solution. As outlined above, in a normative approach, one can propose axioms, abstract principles that one would like the solution to satisfy, and the next step is to pursue their logical consequences. Historically, this was the first argument to justify the Shapley value. Alternatively, one could start by defending a solution on the basis of its definition alone. In the case of the core, this will be especially natural: in a context in which

players can freely get together in groups, the prediction should be payoff vectors that cannot be improved upon by any coalition. One can further enhance one's positive understanding of the solution concept by proposing games in extensive form or in normal form played non-cooperatively by players whose self-enforcing agreements lead to a given solution. This is simply to provide non-cooperative foundations or non-cooperative implementation to the cooperative solution in question, and it is an important research agenda initiated by John Nash in [25], referred to as the Nash program (see [34] for a recent survey). Today, there are interesting results of these different kinds for many solution concepts, which include axiomatic characterizations and non-cooperative foundations. Thus, one can evaluate the appeal of the axioms and the non-cooperative procedures behind each solution to defend a more normative or positive interpretation in each case.

The Core

The idea of agreements that are immune to coalitional deviations was first introduced to economic theory by Edgeworth in [13], which defined the set of coalitionally stable allocations of an economy under the name "final settlements." Edgeworth envisioned this concept as an alternative to *competitive equilibrium* [43], of central importance in economic theory, and was also the first to investigate the connections between the two concepts. Edgeworth's notion, which today we refer to as the *core*, was rediscovered and introduced to game theory in [14]. The origins of the core were not axiomatic. Rather, its simple and appealing definition appropriately describes stable outcomes in a context of unfettered coalitional interaction.

The core of the game (N, V) is the set of payoff vectors

$$C(N, V) = \{x \in V(N): \ \nexists S \subseteq N, x_S \in V(S) \setminus \partial V(S)\}.$$

In words, it is the set of feasible payoff vectors for the grand coalition that no coalition can upset. If such a coalition S exists, we shall say that S can improve upon or block x, and x is deemed unstable. That is, in a context where any coalition can get together, when S has a blocking move, coalition S will form and abandon the grand coalition and its payoffs x_S in order to get to a better payoff for each of the members of the coalition, a plan that is feasible for them.

Non-Emptiness

The core can prescribe the empty set in some games. A game with an empty core is to be understood as a situation of strong instability, as any payoffs proposed to the grand coalition are vulnerable to coalitional blocking.

Example Consider the following simple majority 3-player TU game, in which the votes of at least two players makes the coalition winning. That is, we represent the situation by the following characteristic function: $v(S) = 1$ for any S containing at least two members, $v(\{i\}) = 0$ for all $i \in N$. Clearly, $C(N, v) = \emptyset$. Any feasible payoff agreement proposed to the grand coalition will be blocked by at least one coalition.

An important sufficient condition for the non-emptiness of the core of NTU games is balancedness, as shown in [32]:

Theorem 1 (Scarf [32]) *Let the game (N, V) be balanced. Then $C(N, V) \neq \emptyset$.*

For the TU case, balancedness is not only sufficient, but it becomes also necessary for the non-emptiness of the core:

Theorem 2 (Bondareva [9]; Shapley [39]) *Let (N, v) be a TU game. Then, (N, v) is balanced if and only if $C(N, v) \neq \emptyset$.*

The Connections with Competitive Equilibrium

In economics, the institution of markets and the notion of prices are essential to the understanding of the allocation of goods and the distribution of wealth among individuals. For simplicity in the presentation, we shall concentrate on exchange economies, and disregard production aspects. That is, we shall assume that the goods in question have already been produced in some fixed amounts, and now they are to be allocated to individuals to satisfy their consumption needs.

An *exchange economy* is a system in which each agent i in the set N has a consumption set $Z_i \subseteq \mathbb{R}^l_+$ of commodity bundles, as well as a preference relation over Z_i and an initial endowment $\omega_i \in Z_i$ of the commodities. A feasible *allocation* of goods in the economy is a list of bundles $(z_i)_{i \in N}$ such that $z_i \in Z_i$ and $\sum_{i \in N} z_i \leq \sum_{i \in N} \omega_i$. An allocation is *competitive* if it is supported by a *competitive equilibrium*. A competitive equilibrium is a price-allocation pair $(p, (z_i)_{i \in N})$, where $p \in \mathbb{R}^l \setminus \{0\}$ is such that

- for every $i \in N$, z_i is top-ranked for agent i among all bundles z satisfying that $pz \leq p\omega_i$,
- and $\sum_{i \in N} z_i = \sum_{i \in N} \omega_i$.

In words, this is what the concept expresses. First, at the equilibrium prices, each agent demands z_i, i. e., wishes to purchase this bundle among the set of affordable bundles, the budget set. And second, these demands are such that all markets clear, i. e., total demand equals total supply.

Note how the notion of a competitive equilibrium relies on the principle of private ownership (each individual owns his or her endowment, which allows him or her to access markets and purchase things). Moreover, each agent is a price-taker in all markets. That is, no single individual can affect the market prices with his or her actions; prices are fixed parameters in each individual's consumption decision. The usual justification for the price-taking assumption is that each individual is "very small" with respect to the size of the economy, and hence, has no market power.

One difficulty with the competitive equilibrium concept is that it does not explain where prices come from. There is no single agent in the model responsible for coming up with them. Walras in [43] told the story of an auctioneer calling out prices until demand and supply coincide, but in many real-world markets there is no auctioneer. More generally, economists attribute the equilibrium prices to the workings of the forces of demand and supply, but this appears to be simply repeating the definition. So, is there a different way one can explain competitive equilibrium prices?

As it turns out, there is a very robust result that answers this question. We refer to it as the *equivalence principle* (see, e. g., [6]), by which, under certain regularity conditions, the predictions provided by different game-theoretic solution concepts, when applied to an economy with a large enough set of agents, tend to converge to the set of competitive equilibrium allocations. One of the first results in this tradition was provided by Edgeworth in 1881 for the core. Note how the core of the economy can be defined in the space of allocations, using the same definition as above. Namely, a feasible allocation is in the core if it cannot be blocked by any coalition of agents when making use of the coalition's endowments.

Edgeworth's result was generalized later by Debreu and Scarf in [11] for the case in which an exchange economy is replicated an arbitrary number of times (Anderson studies in [1] the more general case of arbitrary sequences of economies, not necessarily replicas). An informal statement of the Debreu–Scarf theorem follows:

Theorem 3 (Debreu and Scarf [11]) *Consider an exchange economy. Then,*

(i) The set of competitive equilibrium allocations is contained in the core.
(ii) For each non-competitive core allocation of the original economy, there exists a sufficiently large replica of the economy for which the replica of the allocation is blocked.

The first part states a very appealing property of competitive allocations, i. e., their coalitional stability. The second part, known as the core convergence theorem, states that the core "shrinks" to the set of competitive allocations as the economy grows large.

In [3], Aumann models the economy as an atomless measure space, and demonstrates the following core equivalence theorem:

Theorem 4 (Aumann [3]) *Let the economy consists of an atomless continuum of agents. Then, the core coincides with the set of competitive allocations.*

For readers who wish to pursue the topic further, [2] provides a recent survey.

Axiomatic Characterizations

The axiomatic foundations of the core were provided much later than the concept was proposed. These characterizations are all inspired by Peleg's work. They include [26,27], and [36] – the latter paper also provides an axiomatization of competitive allocations in which core convergence insights are exploited.

In all these characterizations, the key axiom is that of *consistency*, also referred to as the reduced game property. Consistency means that the outcomes prescribed by a solution should be "invariant" to the number of players in the game. More formally, let (N, V) be a game, and let σ be a solution. Let $x \in \sigma(N, V)$. Then, the solution is consistent if for every $S \subseteq N$, $x_S \in \sigma(S, V_{xS})$, where (S, V_{xS}) is the reduced game for S given payoffs x, defined as follows. The feasible set for S in this reduced game is the projection of $V(N)$ at $x_{N\setminus S}$, i. e., what remains after paying those outside of S:

$$V_{xS}(S) = \{y_S : (y_S, x_{N\setminus S}) \in V(N)\} .$$

However, the feasible set of $T \subset S$, $T \neq S$, allows T to make deals with any coalition outside of S, provided that those services are paid at the rate prescribed by $x_{N\setminus S}$:

$$V_{xS}(T) = \{y_T \in \cup_{Q\subseteq N\setminus S}(y_T, x_Q) \in V(T \cup Q)\} .$$

It can be shown that the core satisfies consistency with respect to this reduced game. Moreover, consistency is the central axiom in the characterization of the core, which, depending on the version one looks at, uses a host of other axioms; see [26,27,36].

Non-cooperative Implementation

To obtain a non-cooperative implementation of the core, the procedure must embody some feature of anonymity,

since the core is usually a large set and it contains payoffs where different players are treated very differently. For instance, if the procedure always had a fixed set of moves, typically the prediction would favor the first mover, making it impossible to obtain an implementation of the entire set of payoffs.

The model in [30] builds in this anonymity by assuming that negotiations take place in continuous time, so that anyone can speak at the beginning of the game, and at any point in time, instead of having a fixed order. The player that gets to speak first makes a proposal consisting of naming a coalition that contains him and a feasible payoff for that coalition. Next, the players in that coalition get to respond. If they all accept the proposal, the coalition leaves and the game continues among the other players. Otherwise, a new proposal may come from any player in N. It is shown that, if the TU game has a non-empty core (as well as any of its subgames), a class of stationary self-enforcing predictions of this procedure coincide with the core. If a core payoff is proposed to the grand coalition, there are no incentives for individual players to reject it. Conversely, a non-core payoff cannot be sustained because any player in a blocking coalition has an incentive to make a proposal to that coalition, who will accept it (knowing that the alternative, given stationarity, would be to go back to the non-core status quo). [24] offers a discrete-time version of the mechanism: in this work, the anonymity required is imposed on the solution concept, by looking at the order-independent equilibria of the procedure.

The model in [33] sets up a market to implement the core. The anonymity of the procedure stems from the random choice of broker. The broker announces a vector $(x_1, \ldots, x_n)$, where the components add up to $v(N)$. One can interpret x_i as the price for the productive asset held by player i. Following an arbitrary order, the remaining players either accept or reject these prices. If player i accepts, he sells his asset to the broker for the price x_i and leaves the game. Those who reject get to buy from the broker, at the called out prices, the portfolio of assets of their choice if the broker still has them. If a player rejects, but does not get to buy the portfolio of assets he would like because someone else took them before, he can always leave the market with his own asset. The broker's payoff is the worth of the final portfolio of assets that he holds, plus the net monetary transfers that he has received. It is shown in [33] that the prices announced by the broker will always be his top-ranked vectors in the core. If the TU game is such that gains from cooperation increase with the size of coalitions, a beautiful theorem of Shapley in [41] is used to prove that the set of all equilibrium payoffs of this procedure will coincide with the core. Core payoffs are here understood as those price vectors where all arbitrage opportunities in the market have been wiped out. Also, procedures in [35] implement the core, but do not rely on the TU assumption, and they use a procedure in which the order of moves can be endogenously changed by players. Finally, yet another way to build anonymity in the procedure is by allowing the proposal to be made by brokers outside of the set N, as done in [28].

An Application

Consider majority games within a parliament. Suppose there are 100 seats, and decisions are made by simple majority so that 51 votes are required to pass a piece of legislation.

In the first specification, suppose there is a very large party – player 1 –, who has 90 seats. There are five small parties, with 2 seats each. Given the simple majority rules, this problem can be represented by the following TU characteristic function: $v(S) = 1$ if S contains player 1, and $v(S) = 0$ otherwise. The interpretation is that each winning coalition can get the entire surplus – pass the desired proposal. Here, a coalition is winning if and only if player 1 is in it. For this problem, the core is a singleton: the entire unit of surplus is allocated to player 1, who has all the power. Any split of the unit surplus of the grand coalition ($v(N) = 1$) that gives some positive fraction of surplus to any of the small parties can be blocked by the coalition of player 1 alone.

Consider now a second problem, in which player 1, who continues to be the large party, has 35 seats, and each of the other five parties has 13 seats. Now, the characteristic function is as follows: $v(S) = 1$ if and only if S either contains player 1 and two small parties, or it contains four of the small parties; $v(S) = 0$ otherwise. It is easy to see that now the core is empty: any split of the unit surplus will be blocked by at least one coalition. For example, the entire unit going to player 1 is blocked by the coalition of all five small parties, which can award 0.2 to each of them. But this arrangement, in which each small party gets 0.2 and player 1 nothing, is blocked as well, because player 1 can bribe two of the small parties (say, players 2 and 3) and promise them 1/3 each, keeping the other third for itself, and so on. The emptiness of the core is a way to describe the fragility of any agreement, due to the inherent instability of this coalition formation game.

The Shapley Value

Now consider a transferable utility or TU game in characteristic function form. The number $v(S)$ is referred to as the worth of S, and it expresses S's initial position (e. g.,

the maximum total amount of surplus in nummeraire – money, or power – that S initially has at its disposal.

Axiomatics

Shapley in [37] is interested in solving in a fair and unique way the problem of distribution of surplus among the players, when taking into account the worth of each coalition. To do this, he restricts attention to single-valued solutions and resorts to the axiomatic method. He proposes the following axioms on a single-valued solution:

(i) Efficiency: The payoffs must add up to $v(N)$, which means that all the grand coalition surplus is allocated.
(ii) Symmetry: If two players are substitutes because they contribute the same to each coalition, the solution should treat them equally.
(iii) Additivity: The solution to the sum of two TU games must be the sum of what it awards to each of the two games.
(iv) Dummy player: If a player contributes nothing to every coalition, the solution should pay him nothing.

(To be precise, the name of the first axiom should be different. In an economic sense, the statement does imply efficiency in superadditive games, i. e., when for every pair of disjoint coalitions S and T, $v(S) + v(T) \leq v(S \cup T)$. In the absence of superadditivity, though, forming the grand coalition is not necessarily efficient, because a higher aggregate payoff can be obtained from a different coalition structure.)

The surprising result in [37] is this:

Theorem 5 (Shapley [37]) *There is a unique single-valued solution to TU games satisfying efficiency, symmetry, additivity and dummy. It is what today we call the Shapley value, the function that assigns to each player i the payoff*

$$Sh_i(N, v) = \sum_{S, i \in S} \frac{(|S| - 1)!(|N| - |S|)!}{|N|!} [v(S) - v(S \setminus \{i\})].$$

That is, the Shapley value awards to each player the average of his marginal contributions to each coalition. In taking this average, all orders of the players are considered to be equally likely. Let us assume, also without loss of generality, that $v(\{i\}) = 0$ for each player i.

What is especially surprising in Shapley's result is that nothing in the axioms (with the possible exception of the dummy axiom) hints at the idea of marginal contributions, so marginality in general is the outcome of all the axioms, including additivity or linearity. Among the axioms utilized by Shapley, additivity is the one with a lower normative content: it is simply a mathematical property to justify simplicity in the computation of the solution. Young in [45] provides a beautiful counterpart to Shapley's theorem. He drops additivity (as well as the dummy player axiom), and instead, uses an axiom of marginality. Marginality means that the solution should pay the same to a pleyar in two games if his or her marginal contributions to coalitions is the same in both games. Marginality is an idea with a strong tradition in economic theory. Young's result is "dual" to Shapley's, in the sense that marginality is assumed and additivity derived as the result:

Theorem 6 (Young [45]) *There exists a unique single-valued solution to TU games satisfying efficiency, symmetry and marginality. It is the Shapley value.*

Apart from these two, [19] provides further axiomatizations of the Shapley value using the idea of potential and the concept of consistency, as described in the previous section.

There is no single way to extend the Shapley value to the class of NTU games. There are three main extensions that have been proposed: the Shapley λ-transfer value [40], the Harsanyi value [16], and the Maschler–Owen consistent value [23]. They were axiomatized in [5,10,17], respectively.

The Connections with Competitive Equilibrium

As was the case for the core, there is a value equivalence theorem. The result holds for the TU domain (see [4,8,38]). It can be shown that the Shapley value payoffs can be supported by competitive prices. Furthermore, in large enough economies, the set of competitive payoffs "shrinks" to approximate the Shapley value. However, the result cannot be easily extended to the NTU domain. While it holds for the λ-transfer value, it need not obtain for the other extensions. For further details, the interested reader is referred to [18] and the references therein.

Non-cooperative Implementation

Reference [15] was the first to propose a procedure that provided some non-cooperative foundations of the Shapley value. Later, other authors have provided alternative procedures and techniques to the same end, including [20,21,29,44].

We shall concentrate on the description of the procedure proposed by Hart and Mas-Colell in [20]. Generalizing an idea found in [22], which studies the case of $\delta = 0$ – see below –, Hart and Mas-Colell propose the following non-cooperative procedure. With equal probabil-

ity, each player $i \in N$ is chosen to publicly make a feasible proposal to the others: $(x_1, \ldots, x_n)$ is such that the sum of its components cannot exceed $v(N)$. The other players get to respond to it in sequence, following a prespecified order. If all accept, the proposal is implemented; otherwise, a random device is triggered. With probability $0 \leq \delta < 1$, the same game continues being played among the same n players (and thus, a new proposer will be chosen again at random among them), but with probability $1 - \delta$, the proposer leaves the game. He is paid 0 and his resources are removed, so that in the next period, proposals to the remaining $n - 1$ players cannot add up to more than $v(N \setminus \{i\})$. A new proposer is chosen at random among the set $N \setminus \{i\}$, and so on.

As shown in [20], there exists a unique stationary self-enforcing prediction of this procedure, and it actually coincides with the Shapley value payoffs for any value of δ. (Stationarity means that strategies cannot be history dependent). As $\delta \to 1$, the Shapley value payoffs are also obtained not only in expectation, but with independence of who is the proposer. One way to understand this result, as done in [20], is to check that the rules of the procedure and stationary behavior in it are in agreement with Shapley's axioms. That is, the equilibrium relies on immediate acceptances of proposals, stationary strategies treat substitute players similarly, the equations describing the equilibrium have an additive structure, and dummy players will have to receive 0 because no resources are destroyed if they are asked to leave. It is also worth stressing the important role in the procedure of players' marginal contributions to coalitions: following a rejection, a proposer incurs the risk of being thrown out and the others of losing his resources, which seem to suggest a "price" for them.

In [21], the authors study the conditions under which stationarity can be removed to obtain the result. Also, [29] uses a variant of the Hart and Mas-Colell procedure, by replacing the random choice of proposers with a bidding stage, in which players bid to obtain the right to make proposals.

An Application

Consider again the class of majority problems in a parliament consisting of 100 seats. As we shall see, the Shapley value is a good way to understand the power that each party has in the legislature.

Let us begin by considering again the problem in which player 1 has 90 seats, while each of the five small parties has 2 seats. It is easy to see that the Shapley value, like the core in this case, awards the entire unit of surplus to player 1: effectively, each of the small parties is a dummy player, and hence, the Shapley value awards zero to each of them.

Consider a second problem, in which player 1 is a big party with 35 seats, and there are 5 small parties, with 13 seats each. The Shapley value awards 1/3 to the large party, and, by symmetry, 2/15 to each of the small parties. To see this, we need to see when the marginal contributions of player 1 to any coalition are positive. Recall that there are 6! possible orders of players. Note how, if player 1 arrives first or second in the room in which the coalition is forming, his marginal contribution is zero: the coalition was losing before he arrived and continues to be a losing coalition after his arrival. Similarly, his marginal contribution is also zero if he arrives fifth or sixth to the coalition; indeed, in this case, before he arrives the coalition is already winning, so he adds nothing to it. Thus, only when he arrives third or fourth, which happens a third of the times, does he change the nature of the coalition, from losing to winning. This explains his Shapley value share of 1/3. In this game, the Shapley value payoffs roughly correspond to the proportion of seats that each party has.

Next, consider a third problem in which there are two large parties, while the other four parties are very small. For example, let each of the large parties have 48 seats (say, players 1 and 2), while each of the four small parties has only one seat. Now, the Shapley value payoffs are 0.3 to each of the two large parties, and 0.1 to each of the small ones. To see this, note that the marginal contribution of a small party is only positive when he comes fourth in line, and out of the preceding three parties in the coalition, exactly one of them is a large party, i. e., 72 orders out of the 5! orders in which he is fourth. That is, $(72/5!) \times (1/6) = 1/10$. In this case, the competition between the large parties for the votes of the small parties increases the power of the latter quite significantly, with respect to the proportion of seats that each of them holds.

Finally, consider a fourth problem with two large parties (players 1 and 2) with 46 seats each, one mid-size party (player 3) with 5 seats, and three small parties, each with one seat. First, note that each of the three small parties has become a dummy player: no winning coalition where he belongs becomes losing if he leaves the coalition, and so players 4, 5 and 6 are paid zero by the Shapley value. Now, note that, despite the substantial difference of seats between each large party and the mid-size party, each of them is identical in terms of marginal contributions to a winning coalition. Indeed, for $i = 1, 2, 3$, player i's marginal contribution to a coalition is positive only if he arrives second or third or fourth or fifth (and out of the preceding players in the coalition, exactly one is one of the non-dummy players). Note how the Shap-

ley value captures nicely the changes in the allocation of power due to each different political scenario. In this case, the fierce competition between the large parties for the votes of player 3, the swinging party to form a majority, explains the equal share of power among the three.

Future Directions

This article has been a first approach to cooperative game theory, and has emphasized two of its most important solution concepts. The literature on these topics is vast, and the interested reader is encouraged to consult the general references listed below. For the future, one should expect to see progress of the theory into areas that have been less explored, including games with asymmetric information and games with coalitional externalities. In both cases, the characteristic function model must be enriched to take care of the added complexities.

Relevant to this encyclopedia are issues of complexity. The complexity of cooperative solution concepts has been studied (see, for instance, [12]). In terms of computational complexity, the Shapley value seems to be easy to compute, while the core is harder, although some classes of games have been identified in which this task is also simple.

Finally, one should insist on the importance of novel and fruitful applications of the theory to shed new light on concrete problems. In the case of the core, for example, the insights of core stability in matching markets have been successfully applied by Alvin Roth and his collaborators to the design of matching markets in the "real world" (e. g., the job market for medical interns and hospitals, the allocation of organs from doners to patients, and so on) – see [31].

Bibliography

Primary Literature

1. Anderson RM (1978) An elementary core equivalence theorem. Econometrica 46:1483–1487
2. Anderson RM (2008) Core convergence. In: Durlauff S, Blume L (eds) The New Palgrave Dictionary of Economics, 2nd edn. McMillan, London
3. Aumann RJ (1964) Markets with a continuum of traders. Econometrica 32:39–50
4. Aumann RJ (1975) Values of markets with a continuum of traders. Econometrica 43:611–646
5. Aumann RJ (1985) An axiomatization of the non-transferable utility value. Econometrica 53:599–612
6. Aumann RJ (1987) Game theory. In: Eatwell J, Milgate M, Newman P (eds) The New Palgrave Dictionary of Economics, Norton, New York
7. Aumann RJ, Peleg B (1960) Von Neumann–Morgenstern solutions to cooperative games without side payments. Bull Am Math Soc 66:173–179
8. Aumann RJ, Shapley LS (1974) Values of Non-Atomic Games. Princeton University Press, Princeton
9. Bondareva ON (1963) Some applications of linear programming methods to the theory of cooperative games (in Russian). Problemy Kibernetiki 10:119 139
10. de Clippel G, Peters H, Zank H (2004) Axiomatizing the Harsanyi solution, the symmetric egalitarian solution and the consistent solution for NTU-games. I J Game Theory 33:145–158
11. Debreu G, Scarf H (1963) A limit theorem on the core of an economy. Int Econ Rev 4:235–246
12. Deng X, Papadimitriou CH (1994) On the complexity of cooperative solution concepts. Math Oper Res 19:257–266
13. Edgeworth FY (1881) Mathematical Psychics. Kegan Paul Publishers, London. reprinted in 2003) Newman P (ed) F. Y. Edgeworth's Mathematical Psychics and Further Papers on Political Economy. Oxford University Press, Oxford
14. Gillies DB (1959) Solutions to general non-zero-sum games. In: Tucker AW, Luce RD (eds) Contributions to the Theory of Games IV. Princeton University Press, Princeton, pp 47–85
15. Gul F (1989) Bargaining foundations of Shapley value. Econometrica 57:81–95
16. Harsanyi JC (1963) A simplified bargaining model for the n-person cooperative game. Int Econ Rev 4:194–220
17. Hart S (1985) An axiomatization of Harsanyi's non-transferable utility solution. Econometrica 53:1295–1314
18. Hart S (2008) Shapley value. In: Durlauff S, Blume L (eds) The New Palgrave Dictionary of Economics, 2nd edn. McMillan, London
19. Hart S, Mas-Colell A (1989) Potencial, value and consistency. Econometrica 57:589–614
20. Hart S, Mas-Colell A (1996) Bargaining and value. Econometrica 64:357–380
21. Krishna V, Serrano R (1995) Perfect equilibria of a model of n-person non-cooperative bargaining. I J Game Theory 24:259–272
22. Mas-Colell A (1988) Algunos comentarios sobre la teoria cooperativa de los juegos. Cuadernos Economicos 40:143–161
23. Maschler M, Owen G (1992) The consistent Shapley value for games without side payments. In: Selten R (ed) Rational Interaction: Essays in Honor of John Harsanyi. Springer, New York
24. Moldovanu B, Winter E (1995) Order independent equilibria. Games Econ Behav 9:21–34
25. Nash JF (1953) Two person cooperative games. Econometrica 21:128–140
26. Peleg B (1985) An axiomatization of the core of cooperative games without side payments. J Math Econ 14:203–214
27. Peleg B (1986) On the reduced game property and its converse. I J Game Theory 15:187–200
28. Pérez-Castrillo D (1994) Cooperative outcomes through non-cooperative games. Games Econ Behav 7:428–440
29. Pérez-Castrillo D, Wettstein D (2001) Bidding for the surplus: a non-cooperative approach to the Shapley value. J Econ Theory 100:274–294
30. Perry M, Reny P (1994) A non-cooperative view of coalition formation and the core. Econometrica 62:795–817
31. Roth AE (2002) The economist as engineer: game theory, experimentation and computation as tools for design economics. Econometrica 70:1341–1378
32. Scarf H (1967) The core of an *N* person game. Econometrica 38:50–69

33. Serrano R (1995) A market to implement the core. J Econ Theory 67:285–294
34. Serrano R (2005) Fifty years of the Nash program, 1953–2003. Investigaciones Económicas 29:219–258
35. Serrano R, Vohra R (1997) Non-cooperative implementation of the core. Soc Choice Welf 14:513–525
36. Serrano R, Volij O (1998) Axiomatizations of neoclassical concepts for economies. J Math Econ 30:87–108
37. Shapley LS (1953) A value for *n*-person games. In: Tucker AW, Luce RD (eds) Contributions to the Theory of Games II. Princeton University Press, Princeton, pp 307–317
38. Shapley LS (1964) Values of large games VII: a general exchange economy with money. Research Memorandum 4248-PR. RAND Corporation, Santa Monica
39. Shapley LS (1967) On balanced sets and cores. Nav Res Logist Q 14:453–460
40. Shapley LS (1969) Utility comparison and the theory of games. In La Décision: Agrégation et Dynamique des Ordres de Préférence. CNRS, Paris
41. Shapley LS (1971) Cores of convex games. I J Game Theory 1:11–26
42. von Neumann J, Morgenstern O (1944) Theory of Games and Economic Behavior. Princeton University Press, Princeton
43. Walras L (1874) Elements of Pure Economics, or the Theory of Social Wealth. English edition: Jaffé W (ed) Reprinted in 1984 by Orion Editions, Philadelphia
44. Winter E (1994) The demand commitment bargaining and snowballing of cooperation. Econ Theory 4:255–273
45. Young HP (1985) Monotonic solutions of cooperative games. I J Game Theory 14:65–72

Books and Reviews

Myerson RB (1991) Game Theory: An Analysis of Conflict. Harvard University Press, Cambridge
Osborne MJ, Rubinstein A (1994) A Course in Game Theory. MIT Press, Cambridge
Peleg B, Sudholter P (2003) Introduction to the Theory of Cooperative Games. Kluwer, Amsterdam. 2nd edn. Springer, Berlin
Roth AE, Sotomayor M (1990) Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis. Cambridge University Press, Cambridge

Cooperative Games (Von Neumann–Morgenstern Stable Sets)

Jun Wako[1], Shigeo Muto[2]
[1] Department of Economics, Gakushuin University, Tokyo, Japan
[2] Graduate School of Decision Science and Technology, Tokyo, Institute of Technology, Tokyo, Japan

Article Outline

Glossary
Definition of the Subject
Introduction
Stable Sets in Abstract Games
Stable Set and Core
Stable Sets in Characteristic Function form Games
Applications of Stable Sets in Abstract and Characteristic Function Form Games
Stable Sets and Farsighted Stable Sets in Strategic Form Games
Applications of Farsighted Stable Sets in Strategic Form Games
Future Directions
Bibliography

Glossary

Characteristic function form game A characteristic function form game consists of a set of players and a characteristic function that gives each group of players, called a coalition, a value or a set of payoff vectors that they can gain by themselves. It is a typical representation of cooperative games. For characteristic function form games, several solution concepts are defined such as von Neumann – Morgenstern stable set, core, bargaining set, kernel, nucleolus, and Shapley value.

Abstract game An abstract game consists of a set of outcomes and a binary relation, called domination, on the outcomes. Von Neumann and Morgenstern presented this game form for general applications of stable sets.

Strategic form game A strategic form game consists of a player set, each player's strategy set, and each player's payoff function. It is usually used to represent non-cooperative games.

Imputation An imputation is a payoff vector in a characteristic function form game that satisfies group rationality and individual rationality. The former means that the players divide the amount that the grand coalition of all players can gain, and the latter says that each player is assigned at least the amount that he/she can gain by him/herself.

Domination Domination is a binary relation defined on the set of imputations, outcomes, or strategy combinations, depending on the form of a given game. In characteristic function form games, an imputation is said to dominate another imputation if there is a coalition of players such that they can realize their payoffs in the former by themselves, and make each of them better off than in the latter. Domination given a priori in abstract games can be also interpreted in the same way. In strategic form games, domination is defined on the basis of commonly beneficial changes of strategies by coalitions.

Internal stability A set of imputations (outcomes, strategy combinations) satisfies internal stability if there is no domination between any two imputations in the set.

External stability A set of imputations (outcomes, strategy combinations) satisfies external stability if any imputation outside the set is dominated by some imputation inside the set.

von Neumann–Morgenstern stable set A set of imputations (outcomes, strategy combinations) is a von Neumann–Morgenstern stable set if it satisfies both internal and external stability.

Farsighted stable set A farsighted stable set is a more sophisticated stable set concept mainly defined for strategic form games. Given two strategy combinations x and y, we say that x indirectly dominates y if there exist a sequence of coalitions $S^1, \ldots, S^p$ and a sequence of strategy combinations $y = x^0, x^1, \ldots, x^p = x$ such that each coalition S^j can induce strategy combination x^j by a joint move from x^{j-1}, and all members of S^j end up with better payoffs at $x^p = x$ compared to the payoffs at x^{j-1}. A farsighted stable set is a set of strategy combinations that is stable both internally and externally with respect to the indirect domination. A farsighted stable set can be defined in abstract games and characteristic function form games.

Definition of the Subject

The von Neumann–Morgenstern stable set for solution (hereafter stable set) is the first solution concept in cooperative game theory defined by J. von Neumann and O. Morgenstern. Though it was defined cooperative games in characteristic function form, von Neumann and Morgenstern gave a more general definition of a stable set in abstract games. Later, J. Greenberg and M. Chwe cleared a way to apply the stable set concept to the analysis of non-cooperative games in strategic and extensive forms. Though general existence of stable sets in characteristic function form games was denied by a 10-person game presented by W.F. Lucas, stable sets exist in many important games. In voting games, for example, stable sets exist, and they indicate what coalitions can be formed in detail. The core, on the other hand, can be empty in voting games, though it is one of the best known solution concept in cooperative game theory. The analysis of stable sets is not necessarily straightforward, since it can reveal a variety of possibilities. However, stable sets give us deep insights into players' behavior in economic, political and social situations such as coalition formation among players.

Introduction

For studies of economic or social situations where players can take cooperative behavior, the stable set was defined by von Neumann and Morgenstern [31] as a solution concept for characteristic function form cooperative games. They also defined the stable set in abstract games so that one can apply the concept to more general games including non-cooperative situations. Greenberg [9] and Chwe [3] cleared a way to apply the stable set concept to the analysis of non-cooperative games in strategic and extensive forms.

The stable set is a set of outcomes satisfying two stability conditions: internal and external stability. The internal stability means that between any two outcomes in the set, there is no group of players such that all of its members prefer one to the other and they can realize the preferred outcome. The external stability means that for any outcome outside the set, there is a group of players such that all of its members have a commonly preferred outcome in the set and they can realize it. Though general existence was denied by Lucas [18] and Lucas and Rabie [21], the stable set has revealed many interesting behavior of players in economic, political, and social systems.

Von Neumann and Morgenstern (also Greenberg) assumed only a single move by a group of players. Harsanyi [13] first pointed out that stable sets in characteristic function form games may fail to cover "farsighted" behavior of players. Harsanyi's work inspired Chwe's [3] contribution to the formal study of foresight in social environments.

Chwe paid attention to a possible chain of moves such that a move of a group of players will bring about a new move of another group of players, which will further cause a third group of players to move, and so on. Then the group of players moving first should take into account a sequence of moves that may follow, and evaluate their profits obtained at the end. By incorporating such a sequence of moves, Chwe [3] defined a more sophisticated stable set, which we call a farsighted stable set in what follows. Recent work by Suzuki and Muto [51,52] showed that the farsighted stable set provides more reasonable outcomes than the original (myopic) stable set in important classes of strategic form games.

The rest of the chapter is organized as follows. Section "Stable Sets in Abstract Games" presents the definition of a stable set in abstract games. Section "Stable Set and Core" shows basic relations between the two solution concepts of stable set and core. Section "Stable Sets in Characteristic Function form Games" gives the definition of a stable set in characteristic function form games. Section "Applications of Stable Sets in Abstract and Char-

acteristic Function Form Games" first discusses general properties of stable sets in characteristic function form games, and then presents applications of stable sets in abstract and characteristic function games to political and economic systems. Section "Stable Sets and Farsighted Stable Sets in Strategic Form Games" gives the definitions of a stable set and a farsighted stable set in strategic form games. Section "Applications of Farsighted Stable Sets in Strategic Form Games" discusses properties of farsighted stable sets and their applications to social and economic situations. Section "Future Directions" ends the chapter with remarks. Section "Bibliography" offers a list of references.

Stable Sets in Abstract Games

An *abstract game* is a pair $(W, \succ)$ of a set of outcomes W and an irreflexive binary relation $\succ$ on W, where irreflexivity means that $x \succ x$ holds for no element $x \in W$. The relation $\succ$ is interpreted as follows: if $x \succ y$ holds, then there must exist a set of players such that they can induce x from y by themselves and all of them are better off in x.

A subset K of W is called a *stable set* of abstract game $(W, \succ)$ if the following two conditions are satisfied:

1. Internal stability: For any two elements $x, y \in K$, $x \succ y$ never holds.
2. External stability: For any element $z \notin K$, there must exist $x \in K$ such that $x \succ z$.

We explain more in detail what the external and internal stability conditions imply in the definition of a stable set. Suppose players have common understanding that each outcome inside a stable set is "stable" and that each outcome outside the set is "unstable". Here the "stability" means that no group of players has an incentive to deviate from it, and the "instability" means that there is at least one group of players that has an incentive to deviate from it. Then the internal and external stability conditions guarantee that the common understanding is never disproved, and thus continues to prevail.

In fact, suppose the set is both internally and externally stable, and pick any outcome in the set. Then by internal stability, no group of players can be better off by deviating from it and inducing an outcome inside the set. Thus no group of players reaches an agreement to deviate, which makes each outcome inside the set remain stable. Deviating players may be better off by inducing an outcome outside the set; but outcomes outside the set are commonly considered to be unstable. Thus deviating players can never expect that such an outcome will continue. Next pick any outcome outside the set. Then by external stability, there exists at least one group of players who can become better off by deviating from it and inducing an outcome inside the set. The induced outcome is considered to be stable since it is in the set. Hence the group of players will deviate. Hence each outcome outside the set remains unstable.

Stable Set and Core

Another solution concept that is widely known is a core. For a given abstract game $G = (W, \succ)$, a subset C of W is called the *core* of G if $C = \{x \in W |$ there is no $y \in W$ with $y \succ x\}$. From the definition, the core satisfies internal stability. Thus the core C of G is contained in any stable set of G if the latter exists. To see this, suppose that $C \not\subset K$ for a stable set K, and C is non-empty, i. e., $C \neq \emptyset$. (If $C = \emptyset$, then clearly $C \subset K$.) Pick any element $x \in C \backslash K$. Since $x \notin K$, by external stability there exists $y \in K$ with $y \succ x$, which contradicts $x \in C$.

When the core of a game satisfies external stability, it has very strong stability. It is called the *stable core*. The stable core is the unique stable set of the game.

Stable Sets in Characteristic Function form Games

An n-person *game in characteristic function form* with transferable utility is a pair (N, v) of a player set $N = \{1, 2, \ldots, n\}$ and a characteristic function v on the set 2^N of all subsets of N such that $v(\emptyset) = 0$. Each subset of N is called a *coalition*. The game (N, v) is often called a TU-game. A characteristic function form game without transferable utility is called an NTU-game: its characteristic function gives each coalition a set of payoff vectors to players. For NTU-games and their stable sets, refer to Aumann and Peleg [1] and Peleg [35]. In this section, hereafter, we deal with only TU characteristic function form games, and refer to them simply as characteristic function form games.

Let (N, v) be a characteristic function form game. The characteristic function v assigns a real number $v(S)$ to each coalition $S \subseteq N$. The value $v(S)$ indicates the worth that coalition S can achieve by itself.

An n-dimensional vector $x = (x_1, x_2, \ldots, x_n)$ is called a payoff vector. A payoff vector x is called an *imputation* if the following two conditions are satisfied:

1. Group rationality: $\sum_{i=1}^{n} x_i = v(N)$,
2. Individual rationality: $x_i \geq v(\{i\})$ for each $i \in N$.

The first condition says that all players cooperate and share the worth $v(N)$ that they can produce. The second condition says that each player must receive at least the amount

that he/she can gain by him/herself. Let A be the set of all imputations.

Let x, y be any imputations and S be any coalition. We say that x *dominates* y *via* S and write this as $x \operatorname{dom}_S y$ if the following two conditions are satisfied:

1. Coalitional rationality: $x_i > y_i$ for each $i \in S$,
2. Effectivity: $\sum_{i \in S} x_i \leq v(S)$.

The first condition says that every member of coalition S strictly prefers x to y. The second condition says that coalition S can guarantee the payoff x_i for each member $i \in S$ by themselves. We say that x *dominates* y (denoted by $x \operatorname{dom} y$) if there exists at least one coalition S such that $x \operatorname{dom}_S y$.

It should be noted that a pair (A, dom) is an abstract game defined in Sect. "Stable Sets in Abstract Games". It is easily seen that "dom" is an irreflexive binary relation on A. A stable set and the core of game (N, v) are defined to be a stable set and the core of the associated abstract game (A, dom), respectively.

Since von Neumann and Morgenstern defined the stable set, its general existence had been one of the most important problems in game theory. The problem was eventually solved in a negative way. Lucas [18] found the following 10-person characteristic function form game in which no stable set exists.

A game with no stable set: Consider the following 10-person game:

$$\begin{aligned}
N &= \{1, 2, \ldots, 10\}\,, \\
v(N) &= 5\,, \quad v(\{1,3,5,7,9\}) = 4\,, \\
v(\{3,5,7,9\}) &= v(\{1,5,7,9\}) = v(\{1,3,7,9\}) = 3\,, \\
v(\{1,2\}) &= v(\{3,4\}) = v(\{5,6\}) = v(\{7,8\}) \\
&= v(\{9,10\}) = 1\,, \\
v(\{3,5,7\}) &= v(\{1,5,7\}) = v(\{1,3,7\}) = v(\{3,5,9\}) \\
&= v(\{1,5,9\}) = v(\{1,3,9\}) = 2\,, \\
v(\{1,4,7,9\}) &= v(\{3,6,7,9\}) = v(\{5,2,7,9\}) = 2 \ \text{ and} \\
v(S) &= 0 \ \text{ for all other } \ S \subset N\,.
\end{aligned}$$

Though this game has no stable set, it has a nonempty core. A game with no stable set and an empty core was also found by Lucas and Rabie [21].

We remark on a class of games in which a stable core exists. As mentioned before, if a stable set exists, it always contains the core, which is of course true also in characteristic function form games. Furthermore, in characteristic function form games, there is an interesting class, called convex games, in which the core satisfies external stability. That is, the core is a stable core. A characteristic function form game (N, v) is a *convex game* if for any $S, T \subseteq N$ with $S \subset T$ and for any $i \notin T$, $v(S \cup \{i\}) - v(S) \leq v(T \cup \{i\}) - v(T)$, i. e., the bigger coalition a player joins, the larger the player's contribution becomes. In convex games, the core is large and satisfies external stability. For the details, refer to Shapley [45].

Though general existence is denied, the stable set provides us with very useful insights into many economic, political, and social issues. In the following, we will present some stable set analyzes applied to those issues.

Applications of Stable Sets in Abstract and Characteristic Function Form Games

Symmetric Voting Games

This section deals with applications of stable sets to voting situations. Let us start with a simple example.

Example 1 Suppose there is a committee consisting of three players 1, 2 and 3. Each player has one vote. Decisions are done according to a simple majority rule. That is, to pass a bill, at least two votes are necessary.

Before analyzing players' behavior, we first formulate the situation as a characteristic function form game. Let the player set be $N = \{1, 2, 3\}$. Since a coalition of a simple majority of players can pass any bill, we give value 1 to such coalitions. Other coalitions can pass no bill. We thus give them value 0. Hence the characteristic function is given by

$$v(S) = \begin{cases} 1 & \text{if } |S| \geq 2\,, \\ 0 & \text{if } |S| \leq 1\,, \end{cases}$$

where $|S|$ denotes the number of players in coalition S. The set of imputations is given by

$$A = \{x = (x_1, x_2, x_3) | x_1 + x_2 + x_3 = 1, x_1, x_2, x_3 \geq 0\}\,.$$

One stable set of this game is given by the set K consisting of three imputations, $(1/2, 1/2, 0), (1/2, 0, 1/2), (0, 1/2, 1/2)$. A brief proof is the following. Since each of the three imputations has only two numbers 1/2 and 0, internal stability is trivial. To show external stability, take any imputation $x = (x_1, x_2, x_3)$ from outside K. Suppose first $x_1 < 1/2$. Since $x \notin K$, at least one of x_2 and x_3 is less than 1/2. We assume $x_2 < 1/2$. Then $(1/2, 1/2, 0)$ dominates x via coalition $\{1, 2\}$. Next suppose $x_1 = 1/2$. Since $x \notin K, 0 < x_2, x_3 < 1/2$. Thus $(0, 1/2, 1/2)$ dominates x via coalition $\{2, 3\}$. Finally suppose $x_1 > 1/2$. Then $x_2, x_3 < 1/2$, and thus $(0, 1/2, 1/2)$ dominates x via coalition $\{2, 3\}$. Thus the proof of external stability is complete. This three-point stable set indicates that a two-person coalition is formed, and that players in the coalition share equally the outcome obtained by passing a bill.

This game has another three types of stable sets. First, any set $K_c^1 = \{x \in A | x_1 = c\}$ with $0 \le c < 1/2$ is a stable set. The internal stability of each K_c^1 is trivial. To show external stability, take any imputation $x = (x_1, x_2, x_3) \notin K_c^1$. Suppose $x_1 > c$. Define by $y_1 = c, y_2 = x_2 + (x_1 - c)/2, y_3 = x_3 + (x_1 - c)/2$. Then $y \in K_c^1$ and $y \operatorname{dom}_{\{2,3\}} x$. Next suppose $x_1 < c$. Notice that at least one of x_2 and x_3 is less than $1-c$ since $c < 1/2$. Suppose without loss of generality $x_2 < 1-c$. Since $c < 1/2$, we have $(c, 1-c, 0) \in K_c^1$ and $(c, 1-c, 0) \operatorname{dom}_{\{1,2\}} x$. Thus external stability holds. This stable set indicates that player 1 gets a fixed amount c and players 2 and 3 negotiate for how to allocate the rest $1-c$. Similarly, any sets $K_c^2 = \{x \in A | x_2 = c\}$ and $K_c^3 = \{x \in A | x_3 = c\}$ with $0 \le c < 1/2$ are stable sets.

The three-person game of Example 1 has no other stable set. See von Neumann and Morgenstern [31]. The former stable set is called a *symmetric* (or *objective*) stable set, while the latter types are called *discriminatory* stable sets.

As a generalization of the above result, symmetric stable sets are found in general n-person simple majority voting games. An n-person characteristic function form game (N, v) with $N = \{1, 2, \ldots, n\}$ is called a *simple majority voting game* if

$$v(S) = \begin{cases} 1 & \text{if } |S| > n/2\,, \\ 0 & \text{if } |S| \le n/2\,. \end{cases}$$

A coalition S with $v(S) = 1$, i. e., with $|S| > n/2$, is called a *winning* coalition. A winning coalition including no smaller winning coalitions is called a *minimal winning* coalition. In simple majority voting games, a minimal winning coalition means a coalition of $(n+1)/2$ players if n is odd, or $(n+2)/2$ players if n is even. The following theorem holds. See Bott [2] for the proof.

Theorem 1 *Let (N, v) be a simple majority voting game. Then the following hold.*

(1) If n is odd, then the set

$$K = \langle \underbrace{2/(n+1), \ldots, 2/(n+1)}_{\frac{n+1}{2}}, \underbrace{0, \ldots, 0}_{\frac{n-1}{2}} \rangle$$

is a stable set where the symbol $\langle x \rangle$ denotes the set of all imputations obtained from x through permutations of its components.

(2) If n is even, the set

$$K = \langle \{x \in A | \underbrace{x_1 = \ldots = x_{n/2}}_{\frac{n}{2}} \ge \underbrace{x_{(n/2)+1} = \ldots = x_n}_{\frac{n}{2}}\} \rangle$$

is a stable set, where

$$A = \left\{ x = (x_1, \ldots, x_n) \middle| \sum_{i=1}^{n} x_i = 1, x_1, \ldots, x_n \ge 0 \right\}$$

and $\langle Y \rangle = \bigcup_{x \in Y} \langle x \rangle$.

It should be noted from (1) of Theorem 1 that when the number of players is odd, a minimal winning coalition is formed. The members of the coalition share equally the total profit. On the other hand, when the number of players is even, (2) of Theorem 1 shows that every player may gain a positive profit. This implies that the grand coalition of all players is formed. In negotiating for how to share the profit, two coalitions, each with $n/2$ players, are formed and profits are shared equally within each coalition. Since at least $n/2 + 1$ players are necessary to win when n is even, an $n/2$-player coalition is the smallest coalition that can prevent its complement from winning. Such a coalition is called a *minimal blocking* coalition. When n is odd, an $(n+1)/2$-player minimal winning coalition is also a minimal blocking coalition.

General Voting Games

In this section, we present properties of stable sets and cores in general (not necessarily symmetric) voting games. A characteristic function from game (N, v) is called a *simple game* if $v(S) = 1$ or 0 for each nonempty coalition $S \subseteq N$. A coalition S with $v(S) = 1$ (resp. $v(S) = 0$) is a winning coalition (resp. losing coalition). A simple game is called a *voting game* if it satisfies (1) $v(N) = 1$, (2) if $S \subseteq T$, then $v(S) \le v(T)$, and (3) if S is winning, then $N - S$ is losing. The first condition implies that the grand coalition N is always winning. The second condition says that a superset of a winning coalition is also winning. The third condition says that there are no two disjoint winning coalitions. It is easily shown that the simple majority voting game studied in the previous section satisfies these conditions. A player has a veto if he/she belongs to every winning coalition. As for cores of voting games, the following theorem holds.

Theorem 2 *Let (N, v) be a voting game. Then the core of (N, v) is nonempty if and only if there exists a player with veto.*

Thus the core is not a useful tool for analyzing voting situations with no veto player. In simple majority voting games, no player has a veto, and thus the core is empty. The following theorem shows that stable sets always exist.

Theorem 3 *Let (N, v) be a voting game. Let S be a minimal winning coalition and define a set K by*

$$K = \left\{ x \in A \middle| \sum_{i \in S} x_i = 1, x_i = 0 \forall i \notin S \right\} .$$

Then K is a stable set.

Thus in voting games, a minimal winning coalition is always formed, and they gain all the profit. For the proofs of these theorems, see Owen [34]. Further results on stable sets in voting games are found in Bott [2], Griesmer [12], Heijmanns [16], Lucas et al. [20], Muto [26,28], Owen [32], Rosenmüller [36], Shapley [43,44].

Production Market Games

Let us start with a simple example.

Example 2 There are four players, each having one unit of a raw material. Two units of the raw material are necessary for producing one unit of an indivisible commodity. One unit of the commodity is sold at p dollars.

The situation is formulated as the following characteristic function form game. The player set is $N = \{1, 2, 3, 4\}$. Since two units of the raw material are necessary to produce one unit of the commodity, the characteristic function v is given by

$$v(S) = 2p \text{ if } |S| = 4 , \quad v(S) = p \text{ if } |S| = 3, 2 , \quad v(S) = 0 \text{ if } |S| = 1, 0 .$$

The set of imputations is

$$A = \{x = (x_1, x_2, x_3, x_4) | x_1 + x_2 + x_3 + x_4 = 2p, \; x_1, x_2, x_3, x_4 \geq 0\} .$$

The following set K is one of the stable sets of the game:

$$K = \langle \{x = (x_1, x_2, x_3, x_4) \in A | x_1 = x_2 = x_3 \geq x_4\} \rangle .$$

To show internal stability, take two imputations $x = (x_1, x_2, x_3, x_4)$ with $x_1 = x_2 = x_3 \geq x_4$ and $y = (y_1, y_2, y_3, y_4)$ in K. Suppose x dominates y. Since $x_1 = x_2 = x_3 \geq p/2 \geq x_4$, the domination must hold via coalition $\{i, 4\}$ with $i = 1, 2, 3$. Then we have a contradiction $2p = \sum_{i=1}^{4} x_i > \sum_{i=1}^{4} y_i = 2p$, since $y \in K$ implies that the largest three elements of y are equal. To show external stability, take $z = (z_1, z_2, z_3, z_4) \notin K$. Suppose $z_1 \geq z_2 \geq z_3 \geq z_4$. Then $z_1 > z_3$. Define $y = (y_1, y_2, y_3, y_4)$ by

$$y_i = \begin{cases} z_3 + \dfrac{z_1 + z_2 - 2z_3}{4} & \text{for } i = 1, 2, 3 , \\ z_4 + \dfrac{z_1 + z_2 - 2z_3}{4} & \text{for } i = 4 . \end{cases}$$

Then $y \in K$ and $y \operatorname{dom}_{\{3,4\}} z$, since $y_3 > z_3$, $y_4 > z_4$ and $y_3 + y_4 \leq p = v(\{3, 4\})$.

This stable set shows that in negotiating for how to share the profit of $2p$ dollars, three players form a coalition and share equally the gain obtained through collaboration. At least two players are necessary to produce the commodity. Thus a three-player coalition is the smallest coalition that can prevent its complement from producing the commodity, i. e., a minimal blocking coalition. We would claim that in the market a minimal blocking coalition is formed and that profits are shared equally within the coalition.

An extension of the model was given by Hart [14] and Muto [27]. Hart considered the following production market with n players, each holding one unit of a raw material. To produce one unit of an indivisible commodity, k units of raw materials are necessary. The associated production market game is defined by the player set $N = \{1, 2, \ldots, n\}$ and the characteristic function v given by

$$v(S) = \begin{cases} 0 & \text{if } 0 \leq |S| < k , \\ p & \text{if } k \leq |S| < 2k , \\ \vdots & \\ jp & \text{if } jk \leq |S| < (j+1)k , \\ \vdots & \\ hp & \text{if } hk \leq |S| < n , \end{cases} ,$$

where $n = hk + r$ and h, r are integers such that $h \geq 1$ and $0 \leq r \leq k - 1$. When $h = 1$,

$$v(S) = \begin{cases} 0 & \text{if } |S| < k , \\ p & \text{if } |S| \geq k . \end{cases}$$

The following theorem holds.

Theorem 4 *Suppose $h = 1$. Let $t = n - k + 1$ and $n = tu + w$ where u, w are integers such that $u \geq 1$ and $0 \leq w \leq t - 1$. Then the following set K is a stable set.*

$$K = \langle \{x = (x_1, \ldots, x_n) \in A | \; x_1 = \ldots = x_t \geq x_{t+1} = \ldots = x_{2t} \geq \ldots \geq x_{tu+1} = \ldots = x_n = 0\} \rangle ,$$

where

$$A = \left\{ x = (x_1, \ldots, x_n) \middle| \sum_{i=1}^{n} x_i = p, x_1, \ldots, x_n \geq 0 \right\}$$

is the set of imputations.

The theorem shows that in negotiating for how to share the profit, minimal blocking coalitions, i. e., coalitions of

$n-k+1$ players, are formed and within each coalition, profits are shared equally. Players failing to form a coalition gain nothing. When $h \geq 2$, the following theorem holds.

Theorem 5 *Suppose $h \geq 2$. Let*

$$K = \left\langle \left\{ x = (x_1, \ldots, x_n) \in A \middle| \right. \right.$$
$$\left. \left. x_1 = \cdots = x_{n-k+1} \geq \frac{p}{k} \geq x_{n-k+2} = \cdots = x_n \right\} \right\rangle ,$$

where

$$A = \left\{ x = (x_1, \ldots, x_n) \middle| \sum_{i=1}^{n} x_i = hp, x_1, \ldots, x_n \geq 0 \right\} .$$

Then K is a stable set if and only if

$$n \geq (h+1)(k-1) .$$

Therefore if n is large or k is small, then a minimal blocking coalition is formed and the rest of the players also form a coalition. Within each coalition, profits are shared equally.

The next example deals with the case in which more than one raw materials are necessary to produce a commodity.

Example 3 Two types of raw materials P and Q are needed, one unit each, to produce one unit of an indivisible commodity, which is sold at p dollars. Player 1 holds one unit of raw material p, and each of players 2 and 3 holds one unit of raw material Q.

This situation is formulated as the following characteristic function form game. The player set is $N = \{1, 2, 3\}$. Since one unit of raw materials P and Q are necessary to produce the commodity, the characteristic function ν is given by

$$\nu(N) = p ,$$
$$\nu(\{1,2\}) = \nu(\{1,3\}) = p , \quad \nu(\{2,3\}) = 0 ,$$
$$\nu(\{1\}) = \nu(\{2\}) = \nu(\{3\}) = 0 , \quad \nu(\emptyset) = 0 .$$

The set of imputations is

$$A = \{x = (x_1, x_2, x_3) | x_1 + x_2 + x_3 = p, \ x_1, x_2, x_3 \geq 0\} .$$

The following set K is one of the stable sets in this game:

$$K = \{x = (x_1, x_2, x_3) \in A | x_2 = x_3\} .$$

To show internal stability, take two imputations $x = (x_1, x_2, x_3)$ and $y = (y_1, y_2, y_3)$ in K and suppose x dominates y. Then the domination must hold via coalitions $\{1, 2\}, \{1, 3\}$ since values of other coalitions (except $\{1, 2, 3\}$) are 0. If $x \operatorname{dom}_{\{1,2\}} y$, then $x_1 > y_1$ and $x_2 > y_2$ hold. Thus we have a contradiction $p = \sum_{i=1}^{3} x_i > \sum_{i=1}^{3} y_i = p$. The domination via $\{1, 3\}$ leads to the same contradiction. To show external stability, take any imputation $z = (z_1, z_2, z_3) \notin K$. Then $z_2 \neq z_3$. Without loss of generality, let $z_2 < z_3$. Define $y = (y_1, y_2, y_3)$ by

$$y_i = \begin{cases} z_1 + \dfrac{z_3 - z_2}{3} & \text{for } i = 1 , \\ z_2 + \dfrac{z_3 - z_2}{3} & \text{for } i = 2 , \\ z_2 + \dfrac{z_3 - z_2}{3} & \text{for } i = 3 . \end{cases}$$

Then $y \in K$ and $y \operatorname{dom}_{\{1,2\}} z$, since $y_1 > z_1, y_2 > z_2$ and $y_1 + y_2 < \nu(\{1, 2\})$. This stable set shows that in negotiating for how to share the profit p dollars, players 2 and 3 form a coalition against player 1 and share equally the gain obtained through collaboration.

There exist other stable sets in which players 2 and 3 collaborate but they do not share equally the profit. More precisely, the following set

$$K = \{x = (x_1, x_2, x_3) \in A |$$
$$x_2 \text{ and } x_3 \text{ move towards the same direction}\}$$

is a stable set, where "move towards the same direction" means that if x_2 increases then x_3 increases, and if x_2 decreases then x_3 decreases.

A generalization of the results above is given by the following theorem due to Shapley [42]. Shapley's original theorem is more complicated and holds in more general markets.

Theorem 6 *Suppose there are m players, $1, \ldots, m$, each holding one unit of raw material P, and n players, $m+1, \ldots, m+n$, each holding one unit of raw material Q. To produce one unit of an indivisible commodity, one unit of each of raw materials Pand Q is necessary. One unit of commodity is sold at p dollars. In this market, the following set K is a stable set.*

$$K = \{x = (x_1, x_2, \ldots, x_{m+n}) \in A |$$
$$x_1 = \cdots = x_m, x_{m+1} = \cdots = x_{m+n}\} .$$

where

$$A = \left\{ x = (x_1, \ldots, x_m, x_{m+1}, \ldots, x_{m+n}) \middle| \right.$$
$$\left. \sum_{i=1}^{m+n} x_i = p \times \min(m, n), \ x_1, \ldots, x_{m+n} \geq 0 \right\} ,$$

is the set of imputations of this game.

This theorem shows that players holding the same raw material form a coalition and share equally the profit gained through collaboration.

For further results on stable sets in production market games, refer to Hart [14], Muto [27], Owen [34]. Refer also to Lucas [19], Owen [33], Shapley [41], for further general studies on stable sets.

Assignment Games

The following two sections deal with markets in which indivisible commodities are traded between sellers and buyers, or bartered among agents. The first market is the assignment market originally introduced by Shapley and Shubik [47].

An assignment market consists of a set of $n(\geq 1)$ buyers $B = \{1, \ldots, n\}$ and a set of n sellers $F = \{1', \ldots, n'\}$. Each seller $k' \in F$ is endowed with one indivisible commodity to sell, which is called object k'. Thus F also denotes the set of n objects in the market. The objects are differentiated. Each buyer $i \in B$ wants to buy at most one of the objects, and places a nonnegative monetary valuation $u_{ik'}(\geq 0)$ for each object $k' \in F$. The matrix $U = (u_{ik'})_{(i,k')\in B\times F}$ is called the valuation matrix. The sellers place no valuation for any objects. An *assignment market* is denoted by $M(B, F, U)$. We remark that an assignment market with $|B| \neq |F|$ can be transformed into the market with $|B| = |F|$ by adding dummy buyers resp. sellers, and zero rows resp. columns correspondingly to valuation matrix U.

For each coalition $S \subseteq B \cup F$ with $S \cap B \neq \emptyset$ and $S \cap F \neq \emptyset$, we define assignment problem $P(S)$ as follows:

$$\begin{aligned} P(S)\colon \overline{m}(S) = \max_x & \sum_{(i,k')\in(S\cap B)\times(S\cap F)} u_{ik'}x_{ik'} \\ s.t. & \sum_{k'\in S\cap F} x_{ik'} \leq 1 \quad \text{for all } \ i \in S \cap B\,, \\ & \sum_{i\in S\cap B} x_{ik'} \leq 1 \quad \text{for all } \ k' \in S \cap F\,, \\ & x_{ik'} \geq 0 \quad \text{for all } \ (i, k') \in (S \cap B) \times (S \cap F)\,. \end{aligned}$$

Assignment problem $P(S)$ has at least one optimal integer solution (see Simonnard [49]), which gives an optimal matching between sellers and buyers in S that yields the highest possible surplus in S. Without loss of generality, we assume that the rows and columns of valuation matrix U are arranged so that the diagonal assignment x^* with $x^*_{ii'} = 1, i = 1, \ldots, n$, is one of the optimal solutions to $P(B \cup F)$.

For a given assignment market $M(B, F, U)$, we define the associated *assignment game* G to be the characteristic function form game $(B \cup F, v)$. The player set of G is $B \cup F$. The characteristic function v is defined as follows: $v(S) = \overline{m}(S)$ for each $S \subseteq F \cup B$ with $S \cap B \neq \emptyset$ and $S \cap F \neq \emptyset$. For coalitions only of sellers or buyers, they cannot produce surplus from trade. Thus $v(S) = 0$ for each S with $S \subseteq B$, $S \subseteq F$, or $S = \emptyset$. The imputation set of G is

$$A = \left\{(w, p) \in \Re^B_+ \times \Re^F_+ \,\middle|\, \sum_{i\in B} w_i + \sum_{k'\in F} p_{k'} = v(B\cup F)\right\}.$$

Shapley and Shubik [47] proved that for any assignment game G, the core C is given by the set of optimal solutions to the dual problem of assignment problem $P(B \cup F)$, i. e.,

$$\begin{aligned} C = \{&(w, p) \in A | w_i + p_{k'} \geq u_{ik'} = v(\{i, k'\}) \\ &\text{for each } \ (i, k') \in B \times F\}\,, \end{aligned}$$

and thus the core is nonempty. They also showed that sellers' core payoff vector p gives market prices of the respective objects at which the demand and supply equilibrates for each object.

The general existence of stable sets in assignment games is still unsolved. However, as mentioned in Sect. "Stable Set and Core", if a game has the stable core, i. e., the core with external stability, then it is the unique stable set of the game. Thus we consider when an assignment game has the stable core.

Given an assignment market $M(B, F, U)$, we say that valuation matrix U satisfies the *dominant diagonal* condition if all of its diagonal entries are row and column maximums, i. e.,

$$\begin{aligned} u_{ii'} &= \max\{u_{ik'} | k' \in F\} \\ &= \max\{u_{ji'} | j \in B\} \quad \text{for each } \ i = 1, \ldots, n\,. \end{aligned}$$

The dominant diagonal condition implies that each buyer i can yield the maximum surplus by purchasing the object of seller i'. Thus the players do not have to compete for partners. They come to be more concerned with the bargaining with his/her best matched partner. Then it is proved that valuation matrix U satisfies the dominant diagonal condition if and only if the core of G includes the imputations $(\underline{w}, \overline{p})$ and $(\overline{w}, \underline{p})$, where $\underline{w}_i = 0, \overline{p}_{i'} = u_{ii'}, \overline{w}_i = u_{ii'}$, and $\underline{p}_{i'} = 0$ for each $i \in N$. Furthermore, the following theorem holds.

Theorem 7 *Let $M(B, F, U)$ be any assignment market with $|B| = |F|$. The associated assignment game G has the stable core if and only if valuation matrix U satisfies the dominant diagonal condition.*

It is also proved that an assignment game G is convex if and only if valuation matrix U satisfies that $u_{ik'} = 0$ for each $(i, k') \in B \times F$ with $i \neq k$. This implies that the core has the von Neumann–Morgenstern stability in a larger class of assignment games including convex assignment games. For more details, refer to Solymosi and Raghavan [50].

House Barter Games

In this section, we consider a market in which only indivisible commodities are bartered. This market was originally considered by Shapley and Scarf [46].

The market we consider has $n(\geq 2)$ players, each endowed with one indivisible commodity, e. g., a house. Let $N = \{1, \ldots, n\}$ be the set of players. The n indivisible commodities are differentiated. The commodity initially owned by player i is called *house i*. Thus N also denotes the set of houses in the market. We assume that each player wants to own exactly one house, and no player disposes any house. Each player i has a complete, reflexive, and transitive preference relation R_i on N. Here, jR_ih denotes that player i prefers house j at least as well as house h. Let jP_ih denote that player i strictly prefers j to h, and jI_ih denote that player i is indifferent between j and h. Defining players' preferences this way, we assume that each player strictly prefers owning a house to owning no house. The bundle $R = (R_i)_{i\in N}$ of players' preference relations is called a *preference profile*.

There is no divisible good such as money in the market. The players only exchange their houses to make a mutually beneficial trade. Thus an *allocation* of the market is defined to be a bijection x from N onto N, where $x(i)$ denotes the house assigned to player i in x. An allocation can be regarded as a permutation of N. An allocation x is also indicated by the vector $x = (x_1, \ldots, x_n)$ with $x_i = x(i)$ for each $i \in N$. Let A be the set of allocations. The market above is referred to as house barter market $M(N, R)$, or briefly market M.

Let x, y be any allocations of market M. For each coalition $S \subseteq N$, let $x(S)$ be the set of houses assigned to the members of S in x, i. e.,

$$x(S) = \{j \in N | j = x(i) \quad \text{for some} \quad i \in S\} .$$

We say that x *weakly dominates* y (denoted by x wdom y) if there exists a coalition S satisfying the following conditions:

1. $x(i)R_iy(i)$ for each $i \in S$ with P_i holding for at least one $i \in S$,
2. $x(S) = S$.

The second condition is the effectivity condition, which requires that each player i in S can obtain house $x(i)$ by exchanging their own endowments. We say that x *strongly dominates* y (denoted by x sdom y) if $x(i)P_iy(i)$ for *each* $i \in S$, and $x(S) = S$. We use the notations x wdom$_S$ y and x sdom$_S$ y to indicate the associated coalition S.

An allocation x is said to be *individually rational* if $x(i)R_ii$ for each player $i \in N$. An allocation x is *Pareto efficient* if there exists no allocation $y \in A$ with y wdom$_N$ x. If there is no $y \in A$ with y sdom$_N$ x, then x is *weakly Pareto efficient*. The three sets of individually rational, Pareto efficient, and weakly Pareto efficient allocations are denoted by IR, PA, and WPA, respectively.

We define cores and stable sets of market M by cores and stable sets of the associated *house barter games* (A, wdom) and (A, sdom), which are the abstract games with the outcome sets given by the allocation set A, and the binary relations on A given by the weak and strong dominations, respectively.

A nonempty subset of A is referred to as a *wdom stable set* of market M if it is a stable set of abstract game (A, wdom). A nonempty subset of A is referred to as a *sdom stable set stable set* of market M if it is a stable set of abstract game (A, sdom). The wdom and sdom stable sets are the stable sets defined by the weak and strong dominations, respectively. A subset of A, which may be empty, is called the *strict core* of market M if it is the core of abstract game (A, wdom). The *core* of market M is the core of abstract game (A, sdom). The strict core and the core of market M are the cores defined by the weak and strong dominations, respectively.

From the definitions above, a wdom stable set is a subset of PA, and an sdom stable set is a subset of WPA. However, both wdom and sdom stable sets may not be subsets of IR. The strict core is a subset of $PA \cap IR$. The core is a subset of $WPA \cap IR$.

Shapley and Scarf [46] proved that the core is nonempty for any house barter market $M(N, R)$. However, since external stability is not imposed on the core, the core does not necessarily coincide with an sdom stable set. In fact, the following example shows that there is a house barter market with no sdom stable set.

Example 4 Let M_1 be the market with the player set $N = \{1, 2, 3\}$ and the following preference profile:

1) $2 \; P_1 \; 3 \; P_1 \; 1$,
2) $3 \; P_2 \; 1 \; P_2 \; 2$,
3) $1 \; P_3 \; 2 \; P_3 \; 3$,

Market M_1 has six allocations: $x^1 = (2, 3, 1)$, $x^2 = (2, 1, 3)$, $x^3 = (1, 3, 2)$, $x^4 = (3, 2, 1)$, $x^5 = (3, 1, 2)$, $x^6 =$

$(1,2,3)$. From the preference profile, x^1 is clearly a core allocation. Allocation x^1 strongly dominates x^5 and x^6. Let $X = \{x^2, x^3, x^4\}$. We note that x^1 does not strongly dominate any $x^k \in X$. Here, suppose market M_1 has an sdom stable set K. Then the external stability of K implies $x^1 \in K$ and $K \cap X \neq \emptyset$. Note that $x^2 \operatorname{sdom}_{\{1,2\}} x^4$, $x^4 \operatorname{sdom}_{\{1,3\}} x^3$, and $x^3 \operatorname{sdom}_{\{2,3\}} x^2$. This together with the internal stability of K implies that K can contain only one allocation $x^k \in X$. However, the allocation x^k strongly dominates only one allocation in $X \backslash \{x^k\}$. This means that K does not have external stability, which is a contradiction. Thus, there exists no sdom stable set in market M_1.

Market M_1 however has a nice feature: the singleton $\{x^1\}$ is the strict core, and x^1 weakly dominates all the other allocations $x^2, \ldots, x^6$. In addition, every player shows only strict preferences. Noticing these facts, Roth and Postlewaite [38] proved that for any house barter market $M(N,R)$, if each player has a strict preference relation, then the strict core is a singleton, and it is the unique wdom stable set, i. e., the wdom stable core. Wako [54] proved that this property is extended as follows:

Theorem 8 *For any house barter market $M(N,R)$, if the strict core SC is nonempty, then it is the unique wdom stable set. Furthermore, for any $x, y \in SC$, $x(i) I_i y(i)$ for each $i \in N$.*

Even if some players do not have strict preference relations, the strict core is characterized as the wdom stable core as far as it is nonempty. The wdom stable core SC of market M has nice properties. First, each allocation $x \in SC$ is individually rational, Pareto efficient, and not weakly dominated by any other allocations since x is a strict core allocation. Secondly, any allocation outside SC is weakly dominated by some allocation in SC, since SC is a wdom stable set. Thirdly, even if SC contains different allocations, we may choose any one of them, since they are indifferent for each player.

However, the strict core can be empty when indifferences are allowed in preference relations. Shapley and Scarf [46] already discussed this point by the following example.

Example 5 Let M_2 be the market with the player set $N = \{1,2,3\}$ and the following preference profile:

1) $2 \; P_1 \; 3 \; I_1 \; 1$,
2) $1 \; I_2 \; 3 \; P_2 \; 2$,
3) $2 \; P_3 \; 1 \; I_3 \; 3$,

It can be verified that the strict core of M_2 is empty, and that the sets $V^1 = \{(2,3,1),(2,1,3)\}$ and $V^2 = \{(1,3,2),(3,1,2)\}$ are both wdom stable sets of M_2. Thus Theorem 8 does not carry over to the cases with the strict core being empty.

Quint and Wako [40] then gave a necessary and sufficient condition for the strict core to be nonempty. For each player $i \in N$ and each nonempty coalition$S \subseteq N$, let $B_i(S)$ be the set of player i's most-preferred house in S, i. e., $B_i(S) = \{h \in S | h R_i j \text{ for each } j \in S\}$. We call a partition $T = \{T_1, \ldots, T_m\}$ of N a *partition by minimal self-mapped sets (PMSS)* if each $T_k \in T$ satisfies the following conditions:

$$T_k = \bigcup_{i \in T_k} B_i\big(N \backslash \cup_{l=1}^{k-1} T_l\big) \text{ and there is no } S \subset T_k \text{ with } S = \bigcup_{i \in S} B_i\big(N \backslash \cup_{l=1}^{k-1} T_l\big).$$

There exists at least one PMSS for any market M. When there are more than one PMSSs, each PMSS consists of the same sets with only the order of some sets being different. We say that group $T_k \in T$ is a lower (resp. higher) group of $T_l \in T$ if $k > l$ (resp. $k < l$). The fact that $T = \{T_1, \ldots, T_m\}$ is a PMSS means that for each player i of any given group $T_k \in T$, the houses most preferred by i among the endowments of groups T_k and lower are all owned in his/her group T_k. The following theorem holds.

Theorem 9 *Let $T = \{T_1, \ldots, T_m\}$ be a PMSS of a given house barter market $M(N,R)$. Then the strict core is nonempty if and only if there exists an allocation $x \in A$ such that*

$$x(T_k) = T_k \text{ and } x(i) \in B_i\big(N \backslash \cup_{l=1}^{k-1} T_l\big) \text{ for each } i \in T_k \text{ and each } T_k \in T.$$

The necessary and sufficient condition above requires that in any group $T_k \in T$, each player $i \in T_k$ can obtain his/her most-preferred houses (among those owned in groups T_k and lower) through a feasible exchange within his/her own group T_k. We refer to this condition as *segmentability*. Suppose a house barter market has segmentability, and that some players in a group T_k of PMSS T have more preferable houses in higher groups. However, such houses can be exchanged within the higher groups in a mutually beneficial way, and by the definition of a PMSS, no player in the higher groups has an incentive to trade with lower groups. In this sense, the market is segmented into distinct groups. It follows from Theorems 8 and 9 that in house barter markets, segmentability is necessary and sufficient for the existence of the wdom stable core. Quint

and Wako [40] gave a polynomial-time algorithm to examine segmentability of a given house barter market. The following example shows a house barter market with segmentability.

Example 6 Let M_3 be the market with the player set $N = \{1, 2, 3, 4, 5, 6\}$ and the following preference profile:

1) $2\ P_1\ 3\ P_1\ 5\ P_1\ 4\ P_1\ 1\ P_1\ 6$,
2) $1\ I_2\ 3\ P_2\ 4\ P_2\ 6\ P_2\ 5\ P_2\ 2$,
3) $1\ P_3\ 2\ P_3\ 3\ P_3\ 4\ P_3\ 5\ P_3\ 6$,
4) $2\ P_4\ 5\ P_4\ 6\ P_4\ 3\ P_4\ 4\ P_4\ 1$,
5) $1\ I_5\ 4\ P_5\ 5\ P_5\ 3\ P_5\ 6\ P_5\ 2$,
6) $3\ P_6\ 6\ P_6\ 1\ P_6\ 2\ P_6\ 4\ P_6\ 5$.

Although market M_3 has two PMSSs, $T = \{T_1 = \{1, 2, 3\}, T_2 = \{4, 5\}, T_3 = \{6\}\}$ and $T' = \{T'_1 = \{1, 2, 3\}, T'_2 = \{6\}, T'_3 = \{4, 5\}\}$, the differences are only the order of the groups. The wdom stable core of M_3 is the set $K = \{(2, 3, 1, 5, 4, 6)\}$.

The house barter market was also discussed by Moulin [25] from a wide perspective of cooperative microeconomics and game theory. Ehlers [5] initiated a study on stable sets of two-sided matching games, which were originally studied by Gale and Shapley [8]. For a comprehensive study on two-sided matching games, refer to Roth and Sotomayor [39].

Stable Sets and Farsighted Stable Sets in Strategic Form Games

We first defined a stable set in an abstract game in Sect. "Stable Sets in Abstract Games". This means that we can also define a stable set of a strategic form game. In this section we introduce a more sophisticated stable set concept: a *farsighted stable set* of a strategic form game.

Let $G = (N, \{X_i\}_{i \in N}, \{u_i\}_{i \in N})$ be an *n-person strategic form game*, where $N = \{1, 2, \ldots, n\}$ is the set of players, X_i is the set of *strategies* of player i, and u_i is player i's *payoff function*, $u_i : X = X_1 \times X_2 \times \cdots \times X_n \to \Re$ (the set of real numbers).

For any two strategy combinations $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n) \in X$, we say that x *is induced from* y *via coalition* $S \subseteq N$ if $y_i = x_i$ for each $i \in N \backslash S$, that is, the combination x is reached from y by moves only of players in S. It is easily seen from the definition that if x is induced from y via coalition S, y is induced from x via coalition S. Thus we write this inducement relation as $x \leftrightarrow_S y$. We say that x *indirectly dominates* y (denoted by $x \succ y$) if there exist a sequence of strategy combinations $y = x^0, x^1, \ldots, x^{p-1}, x^p = x$ and a sequence of coalitions $S^1, \ldots, S^p$ such that for each $j = 1, \ldots, p$, $x^{j-1} \leftrightarrow_{S^j} x^j$ and $u_i(x) > u_i(x^{j-1})$ for each $i \in S^j$. We sometimes say that x *indirectly dominates* y *starting with* S^1 (denoted by $x \succ_{S^1} y$) to specify the set of players which deviates first from y.

We hereupon remark that in the definition of indirect domination we implicitly assume that joint moves by groups of players are neither once-and-for-all nor binding, i. e., some players in a deviating group may later make another move with players within or even outside the group. It should be noted that the indirect domination defined above is borrowed from Chwe [3]. Though Harsanyi [13] first proposed the notion of indirect domination, his definition was given in characteristic function form games.

When $p = 1$ in the definition of indirect domination, we simply say that x *directly dominates* y, which is denoted by $x \succ^d y$. When we want to specify a deviating coalition, we say that x directly dominates y via coalition S, which is denoted by $x \succ^d_S y$. This direct domination in strategic form games was defined by Greenberg [9].

Let pairs $(X, \succ)$ and $(X, \succ^d)$ be the abstract games associated with game G. A *farsighted stable set* of G is defined to be a stable set of abstract game $(X, \succ)$ with indirect domination. A stable set of abstract game $(X, \succ^d)$ with direct domination is simply called a *stable set* of G.

Applications of Farsighted Stable Sets in Strategic Form Games

Existence of farsighted stable sets in strategic form games remains unsolved. Nevertheless, it has turned out through applications that farsighted stable sets give much sharper insights into players' behavior in economic, political, and social situations than (myopic) stable sets with direct dominations. In what follows, we show the analyses of farsighted stable sets in the prisoner's dilemma and two types of duopoly market games in strategic form.

Prisoner's Dilemma

To make discussion as clear as possible, we will focus on the following particular version of the prisoner' s dilemma shown below. Similar results hold in general prisoner' s dilemma games.

Prisoner's Dilemma:

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	4, 4	0, 5
	Defect	5, 0	1, 1

$$\begin{tikzcd}
\text{(Cooperate, Cooperate)} \arrow[r, "2"] \arrow[d, "1"'] & \text{(Cooperate, Defect)} \arrow[d, "1"'] \\
\text{(Defect, Cooperate)} \arrow[r, "2"] & \text{(Defect, Defect)} \arrow[ul, "12"']
\end{tikzcd}$$

Cooperative Games (Von Neumann–Morgenstern Stable Sets), Figure 1

Here $\rightarrow$ denotes a direct domination. For example, "(Cooperate, Cooperate)$\xrightarrow{2}$(Cooperate,Defect) means (Cooperate, Defect) $\succ^d_{\{2\}}$ (Cooperate,Cooperate). "(Cooperate, Cooperate) $\xleftarrow{12}$ (Defect, Defect)" means (Cooperate, Cooperate) $\succ^d_{\{1,2\}}$ (Defect, Defect)

We first present a farsighted stable set derived when two players use only pure strategies. In this case, the set of strategy combinations is $X = \{$(Cooperate, Cooperate), (Cooperate, Defect), (Defect, Cooperate), (Defect, Defect)$\}$, where in each combination, the former (resp. the latter) is player 1's (resp. 2's) strategy.

The direct domination relation of this game is summarized as in Fig. 1.

From Fig. 1, no stable set (with direct domination) exists in the prisoner's dilemma. However, since (Cooperate, Cooperate) $\succ$ (Cooperate, Defect), (Defect,Cooperate) and there is no other indirect domination, the singleton $\{$(Cooperate, Cooperate)$\}$ is the unique farsighted stable set with respect to $\succ$. Hence if the two players are farsighted and make a joint but not binding move, the farsighted stable set succeeds in showing that cooperation of the players results in the unique stable outcome.

We now study stable outcomes in the mixed extension of the prisoner's dilemma, i. e., the prisoner's dilemma with mixed strategies played. Let $X_1 = X_2 = [0, 1]$ be the sets of mixed strategies of players 1 and 2, respectively, and let $t_1 \in X_1$ (resp. $t_2 \in X_2$) denote the probability that player 1 (resp. 2) plays "Cooperate". It is easily seen that the minimax payoffs to players 1 and 2 are both 1 in this game. We call a strategy combination that gives both players at least (resp. more than) their minimax payoffs, an *individually rational* (resp. a *strictly individually rational*) strategy combination. We then have the following theorem.

Theorem 10 *Let*

$$T = \{(t_1, t_2) | 1/4 < t_1 \leq 1\,,\, t_2 = 1\} \cup \{(t_1, t_2) | t_1 = 1, 1/4 < t_2 \leq 1\}\,,$$

and define the singleton $K^1(t_1, t_2) = \{(t_1, t_2)\}$ for each $(t_1, t_2) \in T$. Let $K^2 = \{(0, 0), (1, 1/4)\}$ and $K^3 = \{(0, 0), (1/4, 1)\}$. Then the sets K^2, K^3, and any $K^1(t_1, t_2)$ with $(t_1, t_2) \in T$ are farsighted stable sets of the mixed extension of the prisoner's dilemma. There are no other types of farsighted stable sets in the mixed extension of the prisoner's dilemma.

This theorem shows that if the two players are farsighted and make a joint but not binding move in the prisoner's dilemma, then essentially a single Pareto efficient and strictly individually rational strategy combination results as a stable outcome. i. e., $K^1(t_1, t_2)$. We, however, have two exceptional cases K^2, K^3 that (Defect, Defect) could be stable together with one Pareto efficient point at which one player gains the same payoff as in (Defect, Defect).

n-Person Prisoner's Dilemma

We consider an n-person prisoner's dilemma. Let $N = \{1, \ldots, n\}$ be the player set. Each player i has two strategies: C (Cooperate) and D (Defect). Let $X_i = \{C, D\}$ for each $i \in N$. Hereafter we refer to a strategy combination as a *state*. The set of states is $X = \prod_{i \in N} X_i$. For each coalition $S \subset N$, let $X_S = \prod_{i \in S} X_i$ and $X_{-S} = \prod_{i \in S^c} X_i$, where S^c denotes the complement of S with respect to N. Let x_s and x_{-s} denote generic elements of X_S and X_{-S}, respectively. Player i's payoff depends not only on his/her strategy but also on the number of other cooperators. Player i's payoff function $u_i : X \rightarrow \Re$ is given by $u_i(x) = f_i(x_i, h)$, where $x \in X, x_i \in X_i$ (player i's choice in x), and h is the number of players other than i playing C. We call the strategic form game thus defined an *n-person prisoner's dilemma game*.

For simplifying discussion, we assume that all players are homogeneous and each player has an identical payoff function. That is, f_i's are identical, and simply written as f unless any confusion arises. We assume the following properties on the function f

Assumption 1

(1) $f(D, h) > f(C, h)$ for all $h = 0, 1, \ldots, n-1$
(2) $f(C, n-1) > f(D, 0)$
(3) $f(C, h)$ and $f(D, h)$ are strictly increasing in h.

Property (1) means that every player prefers playing D to playing C regardless of which strategies other players play. Property (2) means that if all players play C, then each of them gains a payoff higher than the one in $(D, \ldots, D)$.

Property (3) means that if the number of cooperators increases, every player becomes better off regardless of which strategy he/she plays.

It holds from Property (1) that $(D, \ldots, D)$ is the unique Nash equilibrium of the game. Here for $x, y \in X$, we say that *y is Pareto superior to x* if $u_i(y) \geq u_i(x)$ for all $i \in N$ and $u_i(y) > u_i(x)$ for some $i \in N$. The state $x \in X$ is said to be *Pareto efficient* if there is no $y \in X$ that is Pareto superior to x. By Property (2), $(C, \ldots, C)$ is Pareto superior to $(D, \ldots, D)$. Together with Property (3), $(C, \ldots, C)$ is Pareto efficient.

Given a state x, we say that x is *individually rational* if for all $i \in N$, $u_i(x) \geq \min_{y_{-i} \in X_{-i}} \max_{y_i \in X_i} u_i(y)$. If a strict inequality holds, we say that x is *strictly individually rational*. It holds from (1), (3) of Assumption 8.1 that $\min_{y_{-i} \in X_{-i}} \max_{y_i \in X_i} u_i(y) = f(D, 0)$.

The following theorem shows that any state that is strictly individually rational and Pareto efficient is itself a singleton farsighted stable set. That is, any strictly individually rational and Pareto efficient outcome is stable if the players are farsighted. Refer to Suzuki and Muto [51,52] for the details.

Theorem 11 *For n-person prisoner's dilemma game, if x is a strictly individually rational and Pareto efficient state, then $\{x\}$ is a farsighted stable set.*

Duopoly Market Games

We consider two types of duopoly: Cournot quantity-setting duopoly and Bertrand price-setting duopoly. For simplifying discussion, we will consider a simple duopoly model in which firms' cost functions and a market demand function are both linear. Similar results, however, hold in more general duopoly models.

There are two firms 1,2, each producing a homogeneous good with the same marginal cost $c > 0$. No fixed cost is assumed.

(1) **Cournot duopoly:** Firms' strategic variables are their production levels. Let x_1 and x_2 be production levels of firms 1 and 2, respectively. The market price $p(x_1, x_2)$ for x_1 and x_2 is given by

$$p(x_1, x_2) = \max(a - (x_1 + x_2), 0) ,$$

where $a > c$. We restrict the domain of production of both firms to $0 \leq x_i \leq a - c$, $i = 1, 2$. This is reasonable since a firm would not overproduce to make a nonpositive profit. When x_1 and x_2 are produced, firm i's profit is given by

$$\pi_i(x_1, x_2) = (p(x_1, x_2) - c)x_i .$$

Thus Cournot duopoly is formulated as the following strategic form game

$$G^C = (N, \{X_i\}_{i=1,2} , \{\pi_i\}_{i=1,2}) ,$$

where the player set is $N = \{1, 2\}$, each player's strategy set is a closed interval between 0 and $a - c$, i.e., $X_1 = X_2 = [0, a-c]$, and their payoff functions are π_i, $i = 1, 2$. Let $X = X_1 \times X_2$. The joint profit of two firms is maximized when $x_1 + x_2 = (a - c)/2$.

(2) **Bertrand duopoly:** Firms' strategic variables are their price levels. Let

$$D(p) = \max (a - p, 0)$$

be the market demand at price p. Then the total profit at p is

$$\prod(p) = (p - c)D(p) .$$

We restrict the domain of price level p of both firms to $c \leq p \leq a$. This assumption is also reasonable since a firm would avoid a negative profit. The total profit $\prod(p)$ is maximized at $p = (a + c)/2$, which is called a *monopoly price*.

Let p_1 and p_2 be prices of firms 1 and 2, respectively. We assume that if firms' prices are equal, then they share equally the total profit, otherwise all sales go to the lower pricing firm of the two. Thus firm i's profit is given by

$$\rho_i(p_i, p_j) = \begin{cases} \prod(p_i) & \text{if } p_i < p_j \\ \prod(p_i)/2 & \text{if } p_i = p_j \\ 0 & \text{if } p_i > p_j \end{cases}$$
$$\text{for } i, j = 1, 2, i \neq j .$$

Hence Bertrand duopoly is formulated as the strategic form game

$$G^B = \left(N, \{Y_i\}_{i=1,2}, \{\rho_i\}_{i=1,2}\right) ,$$

where $N = \{1, 2\}$, $Y_1 = Y_2 = [c, a]$, and $\rho_i (i = 1, 2)$ is i's payoff function. Let $Y = Y_1 \times Y_2$.

It is well-known that a Nash equilibrium is uniquely determined in either market: $x_1 = x_2 = (a - c)/3$ in the Cournot market, and $p_1 = p_2 = c$ in the Bertrand market.

The following theorem holds for the farsighted stable sets in Cournot duopoly.

Theorem 12 *Let $(x_1, x_2) \in X$ be any strategy pair with $x_1 + x_2 = (a - c)/2$. Then the singleton $\{(x_1, x_2)\}$ is a farsighted stable set. Furthermore, every farsighted stable set is of the form $\{(x_1, x_2)\}$ with $x_1 + x_2 = (a - c)/2$ and $x_1, x_2 \geq 0$.*

As mentioned before, any strategy pair (x_1, x_2) with $x_1 + x_2 = (a - c)/2$ and $x_1, x_2 \geq 0$ maximizes two firms' joint profit. This suggests that the von Neumann–Morgenstern stability together with firms' farsighted behavior produce joint profit maximization even if firms' collaboration is not binding.

As for Bertrand duopoly, we have the following theorem, which claims that the monopoly price pair is itself a farsighted stable set, and no other farsighted stable set exists. Therefore the von Neumann–Morgenstern stability together with firms' farsighted behavior attain efficiency (from the standpoint of firms) also in Bertrand duopoly. Refer to Suzuki and Muto [53] for the details.

Theorem 13 *Let $p = (p_1, p_2)$ be the pair of monopoly prices, i. e., $p_1 = p_2 = (a + c)/2$. Then the singleton $\{p\}$ is the unique farsighted stable set.*

For studies of stable sets with direct domination in duopoly market games, refer to Muto and Okada [29,30]. Properties of stable sets and Harsanyi's original farsighted stable sets in pure exchange economies are investigated by Greenberg et al. [10]. For further studies on stable sets and farsighted stable sets in strategic form games, refer to Kaneko [17], Mariotti [24], Xue [55,56], Diamantoudi and Xue [4].

Future Directions

In this paper, we have reviewed applications of von Neumann–Morgenstern stable sets in abstract games, characteristic function form games, and strategic form games to economic, political and social systems.

Stable sets give us insights into coalition formation among players in the systems in question. Farsighted stable sets, especially applied to some economic systems, show that players' farsighted behavior leads to Pareto efficient outcomes even though their collaboration is not binding. The stable set analysis is also applicable to games with infinitely many players. Those analyses show us new approaches to large economic and social systems with infinitely many players. For the details, refer to Hart [15], Einy et al. [7], Einy and Shitovitz [6], Greenberg et al. [11], Shitovitz and Weber [48], and Rosenmüller and Shitovitz [37]. There is also a study on the linkage between common knowledge of Bayesian rationality and achievement of stable sets in generalized abstract games. Refer to Luo [22,23] for the details.

Analyses of social systems by applying the concepts of farsighted stable sets as well as stable sets must further advance theoretical studies on games in which players inherently take both cooperative and non-cooperative behavior. Those studies will in turn have impacts on developments of economics, politics, sociology, and many applied social sciences.

Bibliography

Primary Literature

1. Aumann R, Peleg B (1960) Von Neumann–Morgenstern solutions to cooperative games without side payments. Bull Am Math Soc 66:173–179
2. Bott R (1953) Symmetric solutions to majority games. In: Kuhn HW, Tucker AW (eds) Contribution to the theory of games, vol II. Annals of Mathematics Studies, vol 28. Princeton University Press, Princeton, pp 319–323
3. Chwe MS-Y (1994) Farsighted coalitional stability. J Econ Theory 63:299–325
4. Diamantoudi E, Xue L (2003) Farsighted stability in hedonic games. Soc Choice Welf 21:39–61
5. Ehlers L (2007) Von Neumann–Morgenstern stable sets in matching problems. J Econ Theory 134:537–547
6. Einy E, Shitovitz B (1996) Convex games and stable sets. Games Econ Behav 16:192–201
7. Einy E, Holzman R, Monderer D, Shitovitz B (1996) Core and stable sets of large games arising in economics. J Econ Theory 68:200–211
8. Gale D, Shapley LS (1962) College admissions and the stability of marriage. Am Math Mon 69:9–15
9. Greenberg J (1990) The theory of social situations: an alternative game theoretic approach. Cambridge University Press, Cambridge
10. Greenberg J, Luo X, Oladi R, Shitovitz B (2002) (Sophisticated) stable sets in exchange economies. Games Econ Behav 39:54–70
11. Greenberg J, Monderer D, Shitovitz B (1996) Multistage situations. Econometrica 64:1415–1437
12. Griesmer JH (1959) Extreme games with three values. In: Tucker AW, Luce RD (eds) Contribution to the theory of games, vol IV. Annals of Mathematics Studies, vol 40. Princeton University Press, Princeton, pp 189–212
13. Harsanyi J (1974) An equilibrium-point interpretation of stable sets and a proposed alternative definition. Manag Sci 20:1472–1495
14. Hart S (1973) Symmetric solutions of some production economies. Int J Game Theory 2:53–62
15. Hart S (1974) Formation of cartels in large markets. J Econ Theory 7:453–466
16. Heijmans J (1991) Discriminatory von Neumann–Morgenstern solutions. Games Econ Behav 3:438–452
17. Kaneko M (1987) The conventionally stable sets in noncooperative games with limited observations I: Definition and introductory argument. Math Soc Sci 13:93–128
18. Lucas WF (1968) A game with no solution. Bull Am Math Soc 74:237–239
19. Lucas WF (1990) Developments in stable set theory. In: Ichiishi T et al (eds) Game Theory and Applications, Academic Press, New York, pp 300–316
20. Lucas WF, Michaelis K, Muto S, Rabie M (1982) A new family of finite solutions. Int J Game Theory 11:117–127

21. Lucas WF, Rabie M (1982) Games with no solutions and empty cores. Math Oper Res 7:491–500
22. Luo X (2001) General systems and φ-stable sets: a formal analysis of socioeconomic environments. J Math Econ 36:95–109
23. Luo X (2006) On the foundation of stability. Academia Sinica, Mimeo, available at http://www.sinica.edu.tw/~xluo/pa14.pdf
24. Mariotti M (1997) A model of agreements in strategic form games. J Econ Theory 74:196–217
25. Moulin H (1995) Cooperative Microeconomics: A Game-Theoretic Introduction. Princeton University Press, Princeton
26. Muto S (1979) Symmetric solutions for symmetric constant-sum extreme games with four values. Int J Game Theory 8:115–123
27. Muto S (1982) On Hart production games. Math Oper Res 7:319–333
28. Muto S (1982) Symmetric solutions for (n,k) games. Int J Game Theory 11:195–201
29. Muto S, Okada D (1996) Von Neumann–Morgenstern stable sets in a price-setting duopoly. Econ Econ 81:1–14
30. Muto S, Okada D (1998) Von Neumann–Morgenstern stable sets in Cournot competition. Econ Econ 85:37–57
31. von Neumann J, Morgenstern O (1953) Theory of Games and Economic Behavior, 3rd ed. Princeton University Press, Princeton
32. Owen G (1965) A class of discriminatory solutions to simple n-person games. Duke Math J 32:545–553
33. Owen G (1968) n-Person games with only 1, n-1, and n-person coalitions. Proc Am Math Soc 19:1258–1261
34. Owen G (1995) Game theory, 3rd ed. Academic Press, New York
35. Peleg B (1986) A proof that the core of an ordinal convex game is a von Neumann–Morgenstern solution. Math Soc Sci 11:83–87
36. Rosenmüller J (1977) Extreme games and their solutions. In: Lecture Notes in Economics and Mathematical Systems, vol 145. Springer, Berlin
37. Rosenmüller J, Shitovitz B (2000) A characterization of vNM-stable sets for linear production games. Int J Game Theory 29:39–61
38. Roth A, Postlewaite A (1977) Weak versus strong domination in a market with indivisible goods. J Math Econ 4:131–137
39. Roth A, Sotomayor M (1990) Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis. Cambridge University Press, Cambridge
40. Quint T, Wako J (2004) On houseswapping, the strict core, segmentation, and linear programming, Math Oper Res 29:861–877
41. Shapley LS (1953) Quota solutions of n-person games. In: Kuhn HW, Tucker TW (eds) Contribution to the theory of games, vol II. Annals of Mathematics Studies, vol 28. Princeton University Press, Princeton, pp 343–359
42. Shapley LS (1959) The solutions of a symmetric market game. In: Tucker AW, Luce RD (eds) Contribution to the theory of games, vol IV. Annals of Mathematics Studies, vol 40. Princeton University Press, Princeton, pp 145–162
43. Shapley LS (1962) Simple games: An outline of the descriptive theory. Behav Sci 7:59–66
44. Shapley LS (1964) Solutions of compound simple games. In Tucker AW et al (eds) Advances in Game Theory. Annals of Mathematics Studies, vol 52. Princeton University Press, Princeton, pp 267–305
45. Shapley LS (1971) Cores of convex games. Int J Game Theory 1:11–26
46. Shapley LS, Scarf H (1974) On cores and indivisibilities. J Math Econ 1:23–37
47. Shapley LS, Shubik M (1972) The assignment game I: The core. Int J Game Theory 1:111–130
48. Shitovitz B, Weber S (1997) The graph of Lindahl correspondence as the unique von Neumann–Morgenstern abstract stable set. J Math Econ 27:375–387
49. Simonnard M (1966) Linear programming. Prentice-Hall, New Jersey
50. Solymosi T, Raghavan TES (2001) Assignment games with stable core. Int J Game Theory 30:177–185
51. Suzuki A, Muto S (2000) Farsighted stability in prisoner's dilemma. J Oper Res Soc Japan 43:249–265
52. Suzuki A, Muto S (2005) Farsighted stability in n-person prisoner's dilemma. Int J GameTheory 33:431–445
53. Suzuki A, Muto S (2006) Farsighted behavior leads to efficiency in duopoly markets. In: Haurie A, et al (eds) Advances in Dynamic Games. Birkhauser, Boston, pp 379–395
54. Wako J (1991) Some properties of weak domination in an exchange market with indivisible goods. Jpn Econ Rev 42:303–314
55. Xue L (1997) Nonemptiness of the largest consistent set. J Econ Theory 73:453–459
56. Xue L (1998) Coalitional stability under perfect foresight. Econ Theory 11:603–627

Books and Reviews

Lucas WF (1992) Von Neumann–Morgenstern stable sets. In: Aumann RJ, Hart S (eds) Handbook of Game Theory with Economic Applications, vol 1. North-Holland, Amsterdam, pp 543–590

Shubik M (1982) Game theory in the social sciences: Concepts and solutions. MIT Press, Boston

Cooperative Multi-hierarchical Query Answering Systems

Zbigniew W. Ras[1,2], Agnieszka Dardzinska[3]

[1] Department of Computer Science, University of North Carolina, Charlotte, USA

[2] Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

[3] Department of Computer Science, Białystok Technical University, Białystok, Poland

Article Outline

Glossary
Definition of the Subject
Introduction
Multi-hierarchical Decision System
Cooperative Query Answering
Future Directions
Bibliography

Glossary

Autonomous information system An autonomous information system is an information system existing as an independent entity.

Intelligent query answering Intelligent query answering is an enhancement of query-answering into a sort of intelligent system (capable or being adapted or molded). Such systems should be able to interpret incorrectly posed questions and compose an answer not necessarily reflecting precisely what is directly referred to by the question, but rather reflecting what the intermediary understands to be the intention linked with the question.

Knowledge base Knowledge base is a collection of rules defined as expressions written in predicate calculus. These rules have a form of associations between conjuncts of values of attributes.

Ontology Ontology is an explicit formal specification of how to represent objects, concepts and other entities that are assumed to exist in some area of interest and relationships holding among them. Systems that share the same ontology are able to communicate about the domain of discourse without necessarily operating on a globally shared theory. A system commits to ontology if its observable actions are consistent with the definitions in the ontology.

Semantics The meaning of expressions written in some language as opposed to their syntax which describes how symbols may be combined independently of their meaning.

Definition of the Subject

One way to make a Query Answering System (QAS) intelligent is to assume the hierarchical structure of their attributes. Such systems have been investigated by Cuppens and Demolombe [3], Gal and Minker [4], and Gaasterland et al. [6], and they are called cooperative. Queries submitted to them are built, in a classical way, from values of attributes describing objects in an information system S and from two-argument functors "and", "or". Instead of "or", we use the symbol "+". Instead of "and", we use the symbol "*". Let us assume that QAS is associated with an information system S. Now, if query q submitted to QAS fails, then any attribute value listed in q can be generalized and the number of objects supporting q in S may increase. In cooperative systems, these generalizations are controlled either by users [4], or by methods based on knowledge discovery [12]. Conceptually, a similar approach has been proposed by Lin [11]. He defines a neighborhood of an attribute value which we can interpret as its generalization (or its parent in the corresponding hierarchical attribute structure). When query fails, then the query answering system is trying to replace values in a query by new values from their corresponding neighborhoods. QAS for S can also collaborate and exchange knowledge with other information systems. In all such cases, it is called intelligent. In papers [14,15] the query answering strategy was based on a guided process of knowledge (rules) extraction and knowledge exchange among systems. Knowledge extracted from information systems collaborating with S was used to construct new attributes in S and/or impute null or hidden values of attributes in S. This way we do not only enlarge the set of queries which QAS can successfully answer but also increase the overall number of retrieved objects and their confidence. Some attributes in S can be distinguished. We usually call them decision attributes. Their values represent concepts which can be defined in terms of the remaining attributes in S, called classification attributes. Query languages for such information systems are built only from values of decision attributes and from two-argument functors "+", "*" [16]. The semantics of queries is defined in terms of semantics of values of classification attributes. Precision and recall of QAS is strictly dependent on the support and confidence of the classifiers used to define queries.

Introduction

Responses by QAS to submitted queries do not always contain the information desired and although they may be logically correct, can sometimes be misleading. Research in the area of intelligent query answering rectifies these problems. The classical approach is based on a cooperative method called relaxation for expanding an information system and related to it queries [3,4]. The relaxation method expands the scope of a query by relaxing the constraints implicit in the query. This allows QAS to return answers related to the original query as well as the literal answers which may be of interest to the user.

This paper concentrates on multi-hierarchical decision systems which are defined as information systems with several hierarchical distinguished attributes called decision attributes. Their values are used to build queries. We give the theoretical framework for modeling such systems and its corresponding query languages. Standard interpretation and the classifier-based interpretation of queries are introduced and used to model the quality (precision, recall) of QAS.

Multi-hierarchical Decision System

In this section we introduce the notion of a multi-hierarchical decision system S and the query language built from atomic expressions containing only values of the decision attributes in S. Classifier-based semantics and the standard semantics of queries in S are proposed. The set of objects X in S is defined as the interpretation domain for both semantics. Standard semantics identifies all objects in X which should be retrieved by a query. Classifier-based semantics gives weighted sets of objects which are retrieved by queries. The notion of precision and recall of QAS in the proposed setting is introduced. We use only rule-based classifiers to define the classifier-based semantics. By improving the confidence and support of the classifiers we improve the precision and recall of QAS.

Definition 1 By a multi-hierarchical decision system we mean a triple $S = (X, A \cup D, V)$, where X is a nonempty, finite set of objects, $D = \{d[i] : 1 \leq i \leq k\}$ is a set of hierarchical decision attributes, A is a nonempty finite set of classification attributes, and $V = \bigcup\{V_a : a \in A \cup D\}$ is a set of their values.

We assume that: V_a, V_b are disjoint for any $a, b \in A \cup D$, such that $a \neq b$, $a : X \to V_a$ is a partial function for every $a \in A \cup D$.

Definition 2 By a set of decision queries (d-queries) for S we mean a least set T_D such that:

- $0, 1 \in T_D$,
- if $w \in \bigcup\{V_a : a \in D\}$, then $w, \sim w \in T_D$,
- if $t_1, t_2 \in T_D$, then $(t_1 + t_2), (t_1 * t_2) \in T_D$.

Definition 3 Decision query t is called simple if $t = t_1 * t_2 * \ldots * t_n$ and

$$(\forall j \in \{1, 2, \ldots, n\}) \left[\left(t_j \in \bigcup\{V_a : a \in D\}\right) \vee \left(t_j = \sim w \wedge w \in \bigcup\{V_a : a \in D\}\right)\right].$$

Definition 4 By a set of classification terms (c-terms) for S we mean a least set T_C such that:

- $0, 1 \in T_C$,
- if $w \in \bigcup\{V_a : a \in A\}$, then $w, \sim w \in T_C$,
- $t_1, t_2 \in T_C$, then $(t_1 + t_2), (t_1 * t_2) \in T_C$.

Definition 5 Classification term t is called simple if $t = t_1 * t_2 * \ldots * t_n$ and

$$(\forall j \in \{1, 2, \ldots, n\}) \left[\left(t_j \in \bigcup\{V_a : a \in A\}\right) \vee \left(t_j = \sim w \wedge w \in \bigcup\{V_a : a \in A\}\right)\right].$$

Definition 6 By a classification rule we mean any expression of the form $[t_1 \to t_2]$, where t_1 is a simple classification term and t_2 is a simple decision query.

Definition 7 Semantics M_S of c-terms in $S = (X, A \cup D, V)$ is defined in a standard way as follows:

- $M_S(0) = 0$, $M_S(1) = X$,
- $M_S(w) = \{x \in X : w = a(x)\}$ for any $w \in V_a$, $a \in A$,
- $M_S(\sim w) = \{x \in X : (\exists v \in V_a)[v = a(x) \,\&\, v \neq w]\}$ for any $w \in V_a$, $a \in A$,
- if t_1, t_2 are terms, then

$$M_S(t_1 + t_2) = M_S(t_1) \cup M_S(t_2),$$
$$M_S(t_1 * t_2) = M_S(t_1) \cap M_S(t_2).$$

Now, we introduce the notation used for values of decision attributes. Assume that the term $d[i]$ also denotes the first granularity level of a hierarchical decision attribute $d[i]$. The set $\{d[i, 1], d[i, 2], d[i, 3], \ldots\}$ represents the values of attribute $d[i]$ at its second granularity level. The set $\{d[i, 1, 1], d[i, 1, 2], \ldots, d[i, 1, n_i]\}$ represents the values of attribute d at its third granularity level, right below the node $d[i, 1]$. We assume here that the value $d[i, 1]$ can be refined to any value from $\{d[i, 1, 1], d[i, 1, 2], \ldots, d[i, 1, n_i]\}$, if necessary. Similarly, the set $\{d[i, 3, 1, 3, 1], d[i, 3, 1, 3, 2], d[i, 3, 1, 3, 3], d[i, 3, 1, 3, 4]\}$ represents the values of attribute d at its fourth granularity level which are finer than the value $d[i, 3, 1, 3]$.

Now, let us assume that a rule-based classifier is used to extract rules describing simple decision queries in S. We denote that classifier by **RC**. The definition of semantics N_S of c-terms is **RC** independent whereas the definition of semantics M_S of d-queries is **RC** dependent.

Definition 8 Classifier-based semantics M_S of d-queries in $S = (X, A \cup D, V)$ is defined as follows:

If t is a simple d-query in S and $\{r_j = [t_j \to t] : j \in J_t\}$ is a set of all rules defining t which are extracted from S by classifier **RC**, then $M_S(t) = \{(x, p_x) : (\exists j \in J_t)(x \in M_S(t_j)[p_x = \Sigma\{\mathrm{conf}(j) \cdot \mathrm{sup}(j) : x \in M_S(t_j) \,\&\, j \in J_t\} / \Sigma\{\mathrm{sup}(j) : x \in M_S(t_j) \,\&\, j \in J_t\}]$, where $\mathrm{conf}(j), \mathrm{sup}(j)$ denote the confidence and the support of the rule $[t_j \to t]$, correspondingly.

Definition 9 Attribute value $d[j_1, j_2, \ldots, j_n]$ in $S = (X, A \cup D, V)$ is dependent on $d[i_1, i_2, \ldots, i_k]$ in S, if one of the following conditions hold:

1) $n \leq k \,\&\, (\forall m \leq n)\,[i_m = j_m]$,
2) $n > k \,\&\, (\forall m \leq k)\,[i_m = j_m]$.

Otherwise, $d[j_1, j_2, \ldots, j_n]$ is called independent from $d[i_1, i_2, \ldots, i_k]$ in S.

Example 1 The attribute value $d[\mathbf{2}, \mathbf{3}, \mathbf{1}, \mathbf{2}]$ is dependent on the attribute value $d[\mathbf{2}, \mathbf{3}, \mathbf{1}, \mathbf{2}, 5, 3]$. Also, $d[\mathbf{2}, \mathbf{3}, \mathbf{1}, \mathbf{2}, \mathbf{5}, \mathbf{3}, 2, 4]$ is dependent on $d[\mathbf{2}, \mathbf{3}, \mathbf{1}, \mathbf{2}, \mathbf{5}, \mathbf{3}]$.

Definition 10 Let $S = \big(X, A \cup \{d[1], d[2], \ldots, d[k]\}, V\big)$, $w \in V_{d[i]}$, and $IV_{d[i]}$ be the set of all attribute values in $V_{d[i]}$ which are independent from w.

Standard semantics N_S of d-queries in S is defined as follows:

- $N_S(0) = 0$, $N_S(1) = X$,
- if $w \in V_{d[i]}$, then $N_S(w) = \{x \in X\colon d[i](x) = w\}$, for any $1 \le i \le k$
- if $w \in V_{d[i]}$, then $N_S(\sim w) = \{x \in X\colon (\exists v \in IV_{d[i]})\ [d[i](x) = v]\}$, for any $1 \le i \le k$
- if t_1, t_2 are terms, then

$$N_S(t_1 + t_2) = N_S(t_1) \cup N_S(t_2)\,,$$
$$N_S(t_1 * t_2) = N_S(t_1) \cap N_S(t_2)\,.$$

Definition 11 Let $S = (X, A \cup D, V)$, t is a d-query in S, $N_S(t)$ is its meaning under standard semantics, and $M_S(t)$ is its meaning under classifier-based semantics. Assume that $N_S(t) = X_1 \cup Y_1$, where $X_1 = \{x_i, i \in I_1\}$, $Y_1 = \{y_i, i \in I_2\}$. Assume also that $M_S(t) = \{(x_i, p_i)\colon i \in I_1\} \cup \{(z_i, q_i)\colon i \in I_3\}$ and $\{y_i, i \in I_2\} \cap \{z_i, i \in I_3\} = \emptyset$.

By precision of a classifier-based semantics M_S on a d-query t, we mean

$$\mathrm{Prec}(M_S, t) = [\Sigma\{p_i\colon i \in I_1\} + \Sigma\{(1 - q_i)\colon i \in I_3\}] \\ /[\mathrm{card}(I_1) + \mathrm{card}(I_3)]\,.$$

By recall of a classifier-based semantics M_S on a d-query t, we mean

$$\mathrm{Rec}(M_S, t) = [\Sigma\{p_i\colon i \in I_1\}]/[\mathrm{card}(I_1) + \mathrm{card}(I_2)]\,.$$

Example 2 Assume that $N_S(t) = \{x_1, x_2, x_3, x_4\}$, $M_S(t) = \{(x_1, p_1), (x_2, p_2), (x_5, p_5), (x_6, p_6)\}$. Then:

$$\mathrm{Prec}(M_S, t) = [p_1 + p_2 + (1 - p_5) + (1 - p_6)]/4\,,$$
$$\mathrm{Rec}(M_S, t) = [p_1 + p_2]/4\,.$$

Cooperative Query Answering

There are cases when classical Query Answering Systems fail to return any answer to a d-query q but still a satisfactory answer can be found. For instance, let us assume that in a multi-hierarchical decision system $S = (X, A \cup D, V)$, where $D = \{d[1], d[2], \ldots, d[k]\}$, there is no single object whose description matches the query q. Assuming that a distance measure between objects in S is defined, then by generalizing q, we may identify objects in S whose descriptions are closest to the description of q. This problem is similar to the problem when the granularity of an attribute value used in a query q is finer than the granularity of the corresponding attribute used in S. By replacing such attribute values in q by more general values used in S, we may retrieve objects from S which satisfy q.

Definition 12 The distance δ_S between two attribute values $d[j_1, j_2, \ldots j_n]$, $d[i_1, i_2, \ldots, i_m]$ in $S = (X, A \cup D, V)$, where $j_1 = i_1$, $p \ge 1$, is defined as follows:

1) if $[j_1, j_2, \ldots, j_p] = [i_1, i_2, \ldots, i_p]$ and $j_{p+1} \ne i_{p+1}$, then $\delta_S\big[d[j_1, j_2, \ldots, j_n], d[i_1, i_2, \ldots, i_m]\big] = 1/2^{p-1}$
2) if $n \le m$ and $[j_1, j_2, \ldots, j_n] = [i_1, i_2, \ldots, i_n]$, then $\delta_S\big[d[j_1, j_2, \ldots, j_n], d[i_1, i_2, \ldots, i_m]\big] = 1/2^n$

The second condition, in the above definition, represents the average case between the best and the worth case.

Example 3 Following the above definition of the distance measure, we get:

1. $\delta_S\big[d[2, 3, 2, 4], d[2, 3, 2, 5, 1]\big] = 1/4$
2. $\delta_S\big[d[2, 3, 2, 4], d[2, 3, 2]\big] = 1/8$

Let us assume that $q = q(a[3, 1, 3, 2], b[1], c[2])$ is a d-query submitted to S. The notation $q(a[3, 1, 3, 2], b[1], c[2])$ means that q is built from $a[3, 1, 3, 2], b[1], c[2]$ which are the atomic attribute values in S. Additionally, we assume that attribute a is not only hierarchical but also it is ordered. It basically means that the difference between the values $a[3, 1, 3, 2]$ and $a[3, 1, 3, 3]$ is smaller than between the values $a[3, 1, 3, 2]$ and $a[3, 1, 3, 4]$. Also, the difference between any two elements in $\{a[3, 1, 3, 1], a[3, 1, 3, 2], a[3, 1, 3, 3], a[3, 1, 3, 4]\}$ is smaller than between $a[3, 1, 3]$ and $a[3, 1, 2]$.

Now, we outline a possible strategy which QAS can follow to solve q. Clearly, the best solution for answering q is to identify objects in S which precisely match the d-query submitted by the user. If it fails, we try to identify objects which match d-query $q(a[3, 1, 3], b[1], c[2])$. If we succeed, then we try d-queries $q(a[3, 1, 3, 1], b[1], c[2])$ and $q(a[3, 1, 3, 3], b[1], c[2])$. If we fail, then we should succeed with $q(a[3, 1, 3, 4], b[1], c[2])$. If we fail with $q(a[3, 1, 3], b[1], c[2])$, then we try $q(a[3, 1], b[1], c[2])$ and so on.

Cooperative Multi-hierarchical Query Answering Systems, Table 1
Multi-hierarchical decision system *S*

X	e	f	g	…	…	a	b	c	d
x_1	$e[1$	$f[1]$	…	…	…	$a[1]$	$b[2]$	$c[1,1]$	$d[3]$
x_2	$e[2]$	$f[1]$	…	…	…	$a[1,1]$	$b[2,1]$	$c[1,1,1]$	$d[3,1,2]$
x_3	$e[2]$	$f[1]$	…	…	…	$a[1,1,1]$	$b[2,2,1]$	$c[2,2]$	$d[1]$
x_4	$e[1]$	$f[2]$	…	…	…	$a[2]$	$b[2,2]$	$c[1,1]$	$d[1,1]$

To present this cooperative strategy in a more precise way, we use an example and start with a very simple dataset. Namely, we assume that S has four decision attributes which belong to the set $\{a, b, c, d\}$. System S contains only four objects listed in Table 1.

Now, we assume that d-query $q = a[1,2] * b[2] * c[1,1] * d[3,1,1]$ is submitted to the multi-hierarchical decision system S (see Table 1). Clearly, q fails in S.

Jointly with q, also a threshold value for a minimum support can be supplied as a part of a d-query. This threshold gives the minimal number of objects that need to be returned as an answer to q. When QAS fails to answer q, the nearest objects satisfying q have to be identified.

The algorithm for finding these objects is based on the following steps:

If QAS fails to identify a sufficient number of objects satisfying q in S, then the generalization process starts. We can generalize either attribute a or d. Since the value $d[3,1,2]$ has a lower granularity level than $a[1,1]$, then we generalize $d[3,1,2]$ getting a new query $q_1 = a[1,2] * b[2] * c[1,1] * d[3,1]$. But q_1 still fails in S. Now, we generalize $a[1,1]$ getting a new query $q_2 = a[1] * b[2] * c[1,1] * d[3,1]$. Objects x_1, x_2 are the only objects in S which support q_2.

If the user is only interested in one object satisfying the query q, then we need to identify which object in $\{x_1, x_2\}$ has a distance closer to q.

Clearly,

$$\begin{aligned}\delta_S[q, x_1] &= \delta_S\big[[a[1,2], b[2], c[1,1], d[3,1,1]], \\ &\qquad [a[1], b[2], c[1,1], d[3]]\big] \\ &= 1/4 + 0 + 0 + 1/4 = 1/2\,, \\ \delta_S[q, x_2] &= \delta_S\big[[a[1,2], b[2], c[1,1], d[3,1,1]], \\ &\qquad [a[1,1], b[2,1], c[1,1,1], d[3,1,2]]\big] \\ &= 1/4 + 1/4 + 1/8 + 1/8 = 3/4\,,\end{aligned}$$

which means x_1 is the winning object.

Note that the cooperative strategy only identifies objects satisfying d-queries and it identifies objects to be returned by QAS to the user. The confidence assigned to these objects depends on the classifier **RC**.

Future Directions

We have introduced the notion of system-based semantics and user-based semantics of queries. User-based semantics are associated with the indexing of objects by a user which is time consuming and unrealistic for very large sets of data. System-based semantics are associated with automatic indexing of objects in X which strictly depends on the support and confidence of classifiers and depends on the precision and recall of a query answering system. The quality of classifiers can be improved by a proper enlargement of the set X and the set of features describing them which differentiate the real-life objects from the same semantic domain as X in a better way. An example, for instance, is given in [16]. The quality of a query answering system can be improved by its cooperativeness. Both precision and recall of QAS is increased if no-answer queries are replaced by generalized queries which are answered by QAS on a higher granularity level than the initial level of queries submitted by users. Assuming that the system is distributed, the quality of QAS for multi-hierarchical decision system S can also be improved through collaboration among sites [14,15].

The key concept of intelligent QAS based on collaboration among sites is to generate global knowledge through knowledge sharing. Each site develops knowledge independently which is used jointly to produce global knowledge. Assume that two sites S_1 and S_2 accept the same ontology of their attributes and share their knowledge in order to solve a user query successfully. Also, assume that one of the attributes at site S_1 is confidential. The confidential data in S_1 can be hidden by replacing them with null values. However, users at S_1 may treat them as missing data and reconstruct them with the knowledge extracted from S_2 [10]. The vulnerability illustrated in this example shows that a security-aware data management is an essential component for any intelligent QAS to ensure data confidentiality.

Bibliography

1. Chmielewski MR, Grzymala-Busse JW, Peterson NW (1993) The rule induction system LERS – a version for personal computers. Found Comput Decis Sci 18(3-4):181–212
2. Chu W, Yang H, Chiang K, Minock M, Chow G, Larson C (1996) Cobase: A scalable and extensible cooperative information system. J Intell Inf Syst 6(2/3):223–259
3. Cuppens F, Demolombe R (1988) Cooperative answering: a methodolgy to provide intelligent access to databases. Pro-

ceeding of the Second International Conference on Expert Database Systems, pp 333–353
4. Gal A, Minker J (1988) Informative and cooperative answers in databases using integrity constraints. Natural Language Understanding and Logic Programming, North Holland, pp 277–300
5. Gaasterland T (1997) Cooperative answering through controlled query relaxation. IEEE Expert 12(5):48–59
6. Gaasterland T, Godfrey P, Minker J (1992) Relaxation as a platform for cooperative answering. J Intell Inf Syst 1(3):293–321
7. Giannotti F, Manco G (2002) Integrating data mining with intelligent query answering. Logics in Artificial Intelligence. Lecture Notes in Computer Science, vol 2424. Springer, Berlin, pp 517–520
8. Godfrey P (1993) Minimization in cooperative response to failing database queries. Int J Coop Inf Syst 6(2):95–149
9. Guarino N (1998) Formal ontology in information systems. IOS Press, Amsterdam
10. Im S, Ras ZW (2007) Protection of sensitive data based on reducts in a distributed knowledge discovery system. Proceedings of the International Conference on Multimedia and Ubiquitous Engineering (MUE 2007), Seoul. IEEE Computer Society, pp 762–766
11. Lin TY (1989) Neighborhood systems and approximation in relational databases and knowledge bases. Proceedings of the Fourth International Symposium on Methodologies for Intelligent Systems. Poster Session Program, Oak Ridge National Laboratory, ORNL/DSRD-24, pp 75–86
12. Muslea I (2004) Machine Learning for Online Query Relaxation. Proceedings of KDD-2004, Seattle. ACM, pp 246–255
13. Pawlak Z (1981) Information systems – theoretical foundations. Inf Syst J 6:205–218
14. Ras ZW, Dardzinska A (2004) Ontology based distributed autonomous knowledge systems. Inf Syst Int J 29 (1):47–58
15. Ras ZW, Dardzinska A (2006) Solving Failing Queries through Cooperation and Collaboration, Special Issue on Web Resources Access. World Wide Web J 9(2):173–186
16. Ras ZW, Zhang X, Lewis R (2007) MIRAI: Multi-hierarchical, FS-tree based Music Information Retrieval System. In: Kryszkiewicz M et al (eds) Proceedings of RSEISP 2007. LNAI, vol 4585. Springer, Berlin, pp 80–89

Coordination Dynamics

James A. S. Kelso
Human Brain and Behavior Loboratory,
Center for Complex Systems and Brain Sciences,
Florida Atlantic University, Boca Raton, USA

Article Outline

Glossary
Definition of the Subject
Introduction
History of Coordination Dynamics: Synergy and Rhythmic Order
Conceptual Foundations of Coordination Dynamics: Self-organizing Dynamical Systems
Finding Dynamical Laws of Coordination
Empirical Foundations of Coordination Dynamics: Pattern Generation, Stability and Phase Transitions
Three Deceptively Simple Experiments
The Theoretical Modeling Strategy of Coordination Dynamics: Symmetry and Bifurcations
Metastable Coordination Dynamics
Coordination of Multiple Components: From Quadrupeds to Brains
"Collective Minds"
Modifying Coordination: Meaningful Information
A Brief Survey of Applications and Elaborations of Coordination Dynamics
Future Directions and Conclusions: The Complementary Nature of Coordination Dynamics
Acknowledgments
Bibliography

Glossary

Coordination dynamics Coordination Dynamics, defined broadly as the science of coordination, describes, explains and predicts how patterns of coordination form, adapt, persist and change in living things. In coordination dynamics the parts communicate via mutual information exchange and information is meaningful and specific to the forms coordination takes. Coordination dynamics embraces both spontaneous self-organizing tendencies and the need to guide or direct them in specific ways in a single conceptual framework. Life, brain, mind and behavior are hypothesized to be linked by virtue of sharing a common underlying coordination dynamics.

Synergies Synergies (*aka* coordinative structures) are functional groupings of structural elements (e. g. neurons, muscles, joints) that are temporarily constrained to act as a single coherent unit. They arise in many contexts on many levels of biological organization, from the genetic to the social. Synergies are the key to understanding biological coordination and as such are the significant units of coordination dynamics. The synergy hypothesis is an hypothesis about how Nature handles biological complexity.

Self-organization The 'self' in the word self-organization refers to the ability of an open system that exchanges matter, energy and information with the environment, to organize itself. Spontaneous patterns arise solely as a result of the dynamics of the system with no specific ordering influence imposed from the outside and no

homunculus-like agent inside. Nonequilibrium phase transitions are the hallmark of self-organization in living things.

Collective variables Collective variables (*aka order parameters* in physics or *coordination variables* in coordination dynamics) are relational quantities that are created by the cooperation among the individual parts of a system. Yet they, in turn, govern the behavior of the individual parts. This is sometimes referred to as *circular* or *reciprocal causality*. In coordination dynamics, the identification of coordination variables depends on the level of description. What is "macro" at one level may be "meso" or "micro" at another.

Control parameters Control parameters refer to naturally occurring environmental conditions or intrinsic, endogenous factors that move the system through its repertoire of patterns and cause them to change. Experimentally, you only know for certain you have identified a control parameter if, when varied, it causes the system's behavior to change qualitatively or discontinuously, i. e., to change its functional state.

Metastability Metastability arises due to broken symmetry in the coordination dynamics where the unstable and stable fixed points (phase- and frequency-locked states) have disappeared due to tangent or saddle-node bifurcations leaving behind only remnants of the fixed points. Metastability is the simultaneous realization of two competing tendencies: the tendency of the components to couple together and the tendency of the components to express their intrinsic independent behavior. Metastability has been hailed as a new principle of organization in complex living systems, including the brain, reconciling apparent contraries such as individual and collective, part and whole, competition and cooperation, integration and segregation, and so forth.

Definition of the Subject

Even before man speculated about the nature and sources of his own experiences, he was probably curious about the agencies by which animal motion was affected. Life and motion are almost synonymous terms. Franklin Fearing [48]

In his preface to the *Principia*, Isaac Newton speculated that not just the motions of the planets, the Moon and the tides could be explained by the forces of attraction and repulsion, but all other natural phenomena as well. Despite the hubris, "self-motion", Newton recognized, "was beyond our understanding" [64]. Three and a half centuries later, the problem remains: the goal-directed coordinated movements of animals are not mere mechanical consequences of the laws of physics, at least as we know them. Despite the many remarkable applications of physics to biology and entire fields devoted to them (e. g. biomechanics, biophysics, nanophysics, etc.) and despite the successes of modern molecular biology, the great unresolved problem of biology remains: *how complex living things are coordinated in space and time*. To recognize that coordination is often purposeful and goal-directed does not at all mean a return to vitalism. But it does pose the scientific challenge of extending physics to understand coordination in living things. Coordination dynamics is a response to this challenge: it is a conceptual framework and research program that deals fundamentally with *animate* (and animated) self-organizing dynamical systems (see also [181,182]). That is, it deals with animate organisms anchored to and engaged in their surrounding worlds. Table 1 compares some of the features of classical mechanics and coordination dynamics. The Table is not intended to be inclusive or to convey the idea that there have been no candidate "paradigms" between classical mechanics and coordination dynamics. For present purposes, "the complex systems" paradigm may represent the most recent break from or extension of classical mechanics. In the complex systems paradigm, self-organization would replace organization, open systems would replace closed systems, change, disorder and process would replace stasis, order and equilibrium, etc. [163]. Coordination dynamics goes a step further. In coordination dynamics it is not organization versus self-organization, order versus disorder, closed versus open systems, reductionism versus emergentism, etc., but rather both aspects that are necessary for an exhaustive account of phenomena and a deeper understanding of coordination in living things. Hence, rather than view these features in opposition, they are better viewed as complementary [100,104,110].

Coordination represents one of the most striking, most taken for granted, yet least understood features of all living things. Imagine a living system whose component parts and processes, on any level of description one choose to examine, did not interact with each other or with their surrounds. Such a collective "cell," "organ," "organism" or "society" would possess neither structure nor function. Coordination can be seen almost everywhere we look, whether in the regulatory interactions among genes that affect how an organism develops and how some diseases like cancer occur, the tumbling and twisting of the bacterial flagellum, the coordinated responses of organisms to constantly varying environmental stimuli, the coordination among nerve cells that produce basic forms of locomotion in vertebrates and invertebrates, the coordination among cell assemblies of the brain that underlies our

Coordination Dynamics, Table 1
Some complementary features of classical mechanics and coordination dynamics

Classical Mechanics	Coordination Dynamics
machines; organized	"organisms "; self-organized
inanimate	animate
motion	coordination (function); animation
matter	materially instantiated; organization
forces	information (semantic) and information exchange; couplings
fundamental dimensions (M, L, T)	collective or coordination variables
space, time	spatiotemporal
linear; smooth dynamics	essentially nonlinear; bifurcations; multi- and metastable
deterministic	fluctuations/variability play key role
decomposable;	"emergent "; synergistic;
motion of whole = sum of motion of the parts	motion of whole > sum of motion of the parts
homogeneous parts, elements, components	heterogeneous units and connectivity
micro versus macro level distinction	level independent strategy; 'one level down '
fixed laws	extensible laws; regularities
context-free	context-dependent
machine/artifactual perspective on order	organic/natural perspective on order

awareness, ability to think, remember, decide and act, the miraculous coordination between the lungs, larynx, lips and tongue that belies a child's first word, the learned coordination among fingers and brain that allows the skilled pianist to play a concerto, the congruence of motion and emotion in dance, drama and everyday life, the coordination between people – like rowers in a racing eight – working together to achieve a common goal. Everything is coordinated.

What do we mean by the word coordination? Coordination is not only spatial and temporal order. Rather, it refers to different kinds and degrees of *functional* order among interacting parts and processes in space and time. Newtonian mechanics may define limits on how biological systems are coordinated, but it says nothing about their functional organization, per se. *How are complex living things coordinated in space and time? What is the nature of the basic interactions that give rise to patterns of coordinated behavior? Why do they take the form they do*? These questions lie at the heart of understanding coordination. Given the ubiquity of coordinated behavior in living things, one might have expected its lawful basis to have been uncovered many years ago. However, it is only in the last 25 years or so, and under quite peculiar circumstances, that basic laws for a quantitative description of coordination have been found.

Introduction

A centipede was happy quite,
Until a frog in fun said:
"Pray tell which leg comes after which?"
This raised her mind to such a pitch,
She lay distracted in the ditch,
Considering how to run. Anon

Coordination Dynamics – the science of coordination – refers to the concepts, methods and tools used to describe, explain and predict how patterns of coordination form, adapt, persist, and change in living things. It is about identifying coordinated patterns in the behavior of living things and expressing how these patterns evolve and change in terms of dynamical laws or rules. The dynamics here refers to equations of motion for key *coordination variables* or *order parameters* [73] that characterize coordinated patterns of behavior on multiple levels of description. As the name implies, the dynamics deals with coordination, not (or not only) with matter in motion: *coordination dynamics* (see Table 1). Through an intimate relationship between theory, experiment, analysis and modeling, Coordination Dynamics seeks to identify the laws, principles and mechanisms underlying coordinated behavior within and between different levels of organization, explicitly addressing the connection between levels. Thus, a goal of Coordination Dynamics is to identify the nature of the functional and context-dependent coordination *within* a part of a system (e. g., the firing of cells in the heart or neurons in a part of the brain), *between* different parts of a system (e. g., parts of the brain, parts of the body, members of an audience) and *between* different kinds of system (e. g., stimuli and responses, organisms and environments, humans and robots, etc.). In

coordination dynamics, the coupling between things is realized by many mechanisms, but is fundamentally informational in nature. That is, Coordination dynamics deals specifically with *meaningfully* coupled, self-organizing systems: the parts communicate via mutual information exchange and information is meaningful and specific to the forms coordination takes. As a source of biological order and pattern, self-organization has received much less attention than the teleonomic, directed nature of living things captured by terms like "program", "blueprint", "template" and so forth. Instead of treating them as opposing theories, Coordination Dynamics unites the spontaneous, self-organizing nature of coordination and the obviously directed, agent-like properties characteristic of animate nature into a single framework [100,104,110]. It does this by studying how functionally meaningful information arises from spontaneous self-organizing processes and how it in turn modifies, guides and directs them.

Coordination Dynamics is both multi- and interdisciplinary, engaging relevant aspects and subfields of psychology, philosophy, biology, neuroscience, computer science, engineering, mathematics and physics. For Coordination Dynamics, a complete understanding of coordination phenomena on *any* given level of description requires: i) specifying the individual coordinating elements and their properties; ii) identifying key parameters, boundary and task conditions that constrain coordination; and (iii) showing how interactions among coordinating elements produce or generate patterns of coordination. By demonstrating in specific cases how the nonlinear coupling among the parts produces coordinated behavior Coordination Dynamics demystifies the popular term "emergence". Even more subtly, a certain régime of Coordination Dynamics called *metastability* resolves the longstanding dichotomy between the whole and the parts by explicitly showing how the individual parts of the system may retain a certain degree of autonomy while still coordinating as a whole. To the extent that they transcend the particular mechanisms through which coordination is realized, the principles of Coordination Dynamics may be said to be "universal" and hence have the potential to describe and explain coordinated behavior in a number of fields ([75]; see also ► Movement Coordination and ► Social Coordination, from the Perspective of Coordination Dynamics). In the case of movement coordination, for example, Coordination Dynamics provides the basic laws for a quantitative description of phenomena that are observed when humans interact in a certain way with themselves, with other humans and with their environment [59,100].

History of Coordination Dynamics: Synergy and Rhythmic Order

Coordination Dynamics arose as a response to the fundamental problem of control and coordination in complex, biological systems: the problem of degrees of freedom. Consider an ordinary movement of the human body. The body itself consists of over 790 muscles and 100 joints that have co-evolved in a complex environment with a nervous system that contains $\sim 10^{12}$ neurons and neuronal connections. On the sensory side, billions of receptor elements embedded in skin, joints and muscles inform the mover about his movement. Clearly, any ordinary human activity requires the cooperation among very many structurally diverse elements – a miracle that we take for granted (e. g. [164]). How does nature compress the very high dimensional state space of such a complex system into something lower dimensional and controllable? An attractive hypothesis proposed by the Russian physiologist Bernstein (1896–1966 [17]) is that in complex living systems, the individual elements are not controlled directly but are rather organized into collectives called *synergies*. Synergies are functional groupings of structural elements (e. g. neurons, muscles, joints) that are temporarily constrained to act as a single coherent unit. Just as new states of matter form under certain conditions when a group of atoms behaves as a single particle (e. g., the Bose–Einstein condensate) so a new state of biological *function* emerges when large ensembles of different elements form a synergy. The synergy hypothesis is therefore an hypothesis about how Nature handles biological complexity. Synergies may appear in many contexts on many levels of biological organization, from the genetic to the social. Depending on context, synergies can accomplish different functions using some of the same anatomical components (e. g., those used for speaking and chewing) and the same function using different components (e. g. writing one's name with a pen attached to the big toe). Once assembled, the degrees of freedom composing a synergy take care of themselves in a relatively autonomous organization. The assembling and disassembling of synergies may be said to be "soft" demanding little energy: synergies are ready to become something else at an instant. They are the "atoms", the significant units of biological function [105,107].

The hallmark of a synergy is that the individual elements adjust to mutual fluctuations and to fluctuations in the external force field (and more generally, the synergy's environment) in order to sustain integrity of function. As a consequence, natural variations (which from the scientist's view may be seen as "errors") that occur in the individual elements of the synergy are compensated by adjust-

ments ("covariations") in other members of the synergy in such a way as to maintain a given function stable. Retaining stability is, for a synergy, the retaining of functional integrity [105,204]. Stability, therefore, plays a key role in coordination dynamics, where the great challenge is to discover what the stability is of. Since the key variables of coordination are not known a priori in living things, they must be identified through empirical research. This, as we shall see, follows a particular strategy.

In the late 1970s and early 80s technological developments and sophisticated computer methods for analyzing complex, multidegree of freedom movements enabled stringent experimental tests of the synergy hypothesis to be carried out [112,120]. Invariably, the experiments, which ranged from postural control to speech production and complex finger and limb movements ([105,107,136,201,204] for reviews), showed: a) that a perturbation to any part of the putative synergy is rapidly compensated for by remotely linked elements in such a way as to preserve system function; b) that the same elements are used in different functions in different ways; c) that different elements may accomplish the same function; and d) that the adjustments observed could in all cases be said to be meaningful, task- and context-specific. All this evidence for the existence of synergies attests to the tremendous redundancy or degeneracy of biological systems [46].

All scientific journeys begin with a single step. The identification of synergies as significant structural–functional units of biological coordination was an important one for the development of Coordination Dynamics. Synergies simplify control by reducing the number of variables that must be independently specified: as constraints, they make control and coordination of complex, multivariable systems possible. But understanding goes far beyond identification. How are synergies formed? What principles govern their assembly? And how does one synergy change spontaneously to another as internal or external conditions change? How can distinct synergies co-exist among the same set of components? And how are individual components of the synergy engaged and disengaged as circumstances change?

Insights into these questions come from the work of a largely unheralded genius called Erich von Holst [215], a behavioral physiologist who spent his life studying coordination in a wide variety of creatures – from worms to man. Von Holst's research will not give us answers to all the questions about synergies but it will provide key insights into the essence of coordination and a stepping stone to finding the underlying principles. Using an experimental model system that allowed him to measure an elementary synergy – the to and fro motions of the fins of the swimming dogfish *Labrus* under carefully controlled water flow conditions – von Holst identified at least three basic types of coordination: *absolute* coordination, in which component parts are locked together in time (like the synchronized flashing of fireflies, a couple making love or phase synchrony between parts of the brain); partial or *relative* coordination, in which the component parts 'lock in' only transiently and then break apart as circumstances change (like a little boy walking hand in hand with his father on the beach; dad must slow down and/or son add a step so that they can stay together); and no coordination at all, in which the component parts behave quite independently (as occurs in the locomotion of millipedes and centipedes when the same little boy chops off their middle legs, or perhaps after persistent, long term practice in playing the piano or the violin). Various blends, mixes and transitions between these coordinated behaviors were also observed – always matching the exigencies of the internal and external environment.

Why might some kind of common principle exist for such diverse phenomena? The reason is that the same basic coordination phenomena seem to cut across a wide range of levels, creatures and functions. Among those observed are: patterns of coordination remain stable in time despite continuous, and often unexpected perturbations; the ease with which component parts and processes are flexibly engaged and disengaged as functional demands or environmental conditions change; the existence of multiple coordination patterns – so-called multifunctionality – that effectively satisfy the same set of circumstances; the selection of particular coordination patterns that are exquisitely tailored to suit the current needs of the organism; adaptation of coordination to changing internal and external contingencies; smooth and abrupt transitions from one coordinated pattern to another; transitions from partially to fully coordinated patterns and vice-versa; persistence of a coordinated pattern even when conditions that led to the establishment of the pattern have changed (a kind of memory), and so forth. Such phenomena appear so commonly and so consistently as to suggest the existence of an underlying lawfulness or regularity that transcends the differences between systems. Nature, as the saying goes, operates with ancient themes. Or maybe nature just is what it is.

Conceptual Foundations of Coordination Dynamics: Self-organizing Dynamical Systems

Given we accept the empirical facts about synergies and rhythmic order in the nervous system and the movements of living things, what concepts, methods and tools do we

use to understand them? Were synergies simply rigid mechanical entities built by an engineer or an intelligent designer, control theory with its programs, reference levels, comparators, feedforward and feedback error correcting mechanisms and so forth, might have seemed an obvious place to look for explanatory tools. A program instructs the parts of a system what to do and when to do it. Feedback may then be used to correct errors in the outcome. But now what? The system receives an error signal: How does it know which of its many parts to correct? In a complex system composed of very many components it may take a very long time to come up with a solution, a problem computer scientists refer to as an NP-complete problem, where NP means "non-deterministic polynomial time complete". Biology with its degeneracy and redundancy has no such problem. "Error" signals from one part of a synergy are rapidly compensated by other members. So if anything, the *machine perspective* on order and regulation (Table 1) seems to compound the problem rather than solve it.

Coordination Dynamics takes its inspiration from *a natural, organic perspective*, i. e., how nature handles complexity (Table 1). It is well-known that pattern formation in open, nonequilibrium physical and chemical systems such as fluids, lasers and chemical reactions can emerge spontaneously. These patterns arise solely as a result of the dynamics of the system with no specific ordering influence imposed from the outside environment and no homunculus-like agent inside. Such "self-organized" pattern formation is a cooperative phenomenon that results from the interaction of large numbers of interacting subsystems [73,152]. It should be stressed here that there is no "self" inside the system responsible for prescribing or coding the emergent pattern. The 'self' in self-organization comes from the fact that given the ability to exchange matter, energy and information with the environment, the system organizes itself. That the organism is an open system is one of two essential criteria for life postulated by Francis Crick in *Of Molecules and Men* [33], yet it has received much less attention in biology than Crick's other criterion, the need for organisms to reproduce and pass on 'copies' of themselves to their descendants. Here already we see a dichotomy between a complex system's natural ordering tendencies and the need (at least in living systems) to guide that order in specific ways. Coordination dynamics (Table 1) reconciles this dichotomy by viewing these two fundamental aspects as complementary ([100,104,110]).

In his general theory of nonequilibrium phase transitions called "synergetics" Haken [73] showed that close to critical points where a so-called *control parameter* crosses a threshold, very complex, high-dimensional systems can be completely described by a much lower dimensional dynamics specified in terms of only a few collective variables or *order parameters*. What do these terms mean? *Control Parameters* refer to naturally occurring environmental conditions or intrinsic, endogenous factors that on first blush appear analogous to what an experimental scientist might call an independent variable. But the concept is entirely different, and the implications for experimental design far reaching [96]. The role of control parameters is to move the system through its repertoire of patterns and cause them to change. In fact, you only know for certain you have identified a control parameter if, when varied, it causes the system's behavior to change qualitatively or discontinuously, i. e., to change state. In a dynamical system, when a parameter changes smoothly, the attractor in general also changes smoothly. Sizeable changes in the input have little or no effect on the resulting output. However, when the control parameter passes through a critical point or threshold in an intrinsically nonlinear dynamical system an abrupt change in the attractor can occur. This sensitive dependence on parameters is called a *bifurcation* in mathematics, or a *nonequilibrium phase transition* in physical theories of pattern formation [73]. Indeed, control parameters are often referred to in mathematics as *bifurcation parameters*. Qualitative change does not mean that quantification is impossible. To the contrary, qualitative change is at the heart of pattern formation and, provided care is taken to evaluate system timescales (e. g., how quickly the control parameter is changed relative to the typical time of the system to react to perturbations; see [121]) quantitative predictions ensue that can be tested experimentally (see Sect. "The Theoretical Modeling Strategy of Coordination Dynamics: Symmetry and Bifurcations").

Collective variables are relational quantities that are created by the cooperation among the individual parts of a system. Yet they, in turn, govern the behavior of the individual parts. This is sometimes referred to as *circular* or *reciprocal causality*. In self-organizing systems the stranglehold of linear causality is broken. At best, simple cause-effect relations are the exception, not the rule. Depending on where the system lives in the space of its parameters, many causes can produce the same effect or the same cause can have multiple effects. One can intuit why the concept of collective or coordination variable is central to a science of coordination. The reason is that interactions in such systems are so complicated that understanding may only be possible in terms of system-specific collective or coordination variables. The latter are not necessarily "macroscopic quantities". In coordination dynamics, the identification of coordination variables depends on the level of

description. What is "macro" at one level may be "meso" or "micro" at another. This strategy of folding together all aspects within the dynamics of collective or coordination variables embraces the full complexity of living things on a given level of description without proliferating arbitrary divisions (for a nice discussion see [203]).

In nonequilibrium systems, the enormous compression of degrees of freedom near critical points arises because events occur on different timescales: the faster individual elements in the system become "enslaved" to the slower, "emergent" collective variables [73]. Alternatively, one may conceive of a hierarchy of timescales for various processes involved in coordination. On a given level of the hierarchy are coordination variables subject to constraints (e. g., of the task) that act as boundary conditions on the coordination dynamics. At the next level down are component processes and events that typically operate on faster timescales. Notice for the 'tripartite scheme' of Coordination Dynamics (see pp. 66–67 in [100]) the key is to choose a level of description and understand the relation between adjacent levels, not reduce to some "fundamental" lower level (Table 1). In coordination dynamics, no level is any more or less fundamental than any other. A complete description of a phenomenon always requires three adjacent tiers: The boundary conditions and control parameters that establish the context for particular coordination phenomena to occur; the collective level and its dynamics; the component level and its dynamics including the nonlinear coupling between components.

Dynamic instability is the generic mechanism underlying self-organized pattern formation and change in all (open) systems coupled to their internal or external environments [153]. Near instability the individual elements, in order to accommodate to current conditions, must order themselves in new or different ways. The patterns that emerge at nonequilibrium phase transitions may be defined as attractive states of the collective variable dynamics. That is, the collective variable may converge in time to a certain limit set or attractor solution, a nonequilibrium steady state. Attractors can be *fixed points*, in which all initial conditions converge to some stable rest state. Attractors can also be *periodic*, exhibiting preferred rhythms or orbits on which the system settles regardless of where it starts. Or, there can be so-called *strange* attractors; strange because they exhibit *deterministic chaos*, a type of irregular behavior resembling random noise, yet often containing pockets of quite ordered behavior. Stable fixed point, limit cycle and chaotic solutions as well as a wide variety of other transient and irregular behaviors are possible in the *same* system, depending on the values of control parameters (and their time dependence). Moreover, fluctuations are always present, constantly testing whether a given pattern is stable. Fluctuations are not just noise; rather, by probing the stability of existing states they allow the system to discover new, more adaptive patterns that suit the prevailing circumstances (boundary conditions, control parameters; Table 1).

How might these conceptual tools aid our understanding of biological coordination? On first blush, it might seem a gigantic leap from the physics and mathematics of pattern formation in nonequilibrium systems to the problem of coordination in living things. Yet in science, analogy often plays a major role in bringing about conceptual breakthroughs. Although initially the analogy may seem far-fetched, great science often starts with a vague idea which, when followed by crucial experiments and mathematical theory renders the vague idea exact. A key aspect to appreciate is that cooperative phenomena in physical systems are typically independent of the particular molecular machinery or material substrate that instantiates them. This is because the elementary components are the same, i. e. homogeneous. On the other hand, in living, evolved things the component elements are often quite different. Thus any theory of coordination of living things will have to take into account the heterogeneity of its component elements. Perhaps as a consequence of inherent heterogeneity (and no doubt the advancement of technology) the tendency in biology is to focus more and more on specific processes at ever smaller and smaller scales. As a result, building huge data bases may sometimes appear to take precedence over finding scientific laws [55].

Finding Dynamical Laws of Coordination

What if biological coordination were shown to be a self-organized phenomenon? Might that be a springboard to finding laws of coordination? In the sense of T.S. Kuhn [131] such questions appear to call out for a new paradigm, special entry points where irrelevant details may be pruned away exposing the essential aspects one is trying to understand. Inspired by synergetics (and paradoxical though it may seem) the key to determining if coordination as a self-organized phenomenon is to focus on *qualitative change*, places where abrupt switches or bifurcations in coordination occur. Qualitative change is crucial because it affords a clear distinction between one coordination pattern and another, thereby enabling one to identify the key *collective variables or order parameters* that define coordination states and their coordination dynamics. If a complex system is changing smoothly and linearly it is hard to distinguish the variables that matter, so-called state variables, from the ones that don't. Qualita-

tive change may also be used to infer relevant quantities in more naturalistic settings. In situations where many variables may be changing in uncontrolled ways, the one(s) that change(s) abruptly are likely to be the most meaningful, both for the phenomena themselves and our understanding of them [96]. Likewise, any parameter that induces qualitative behavioral change qualifies as a *control parameter*. This is the reason why stability is so important. As a control parameter crosses a critical value the previously stable pattern becomes unstable and the system switches to a different pattern that is now stable beyond the critical point. The quite general predictions of nonequilibrium phase transition theory are a strong enhancement of fluctuations (critical fluctuations) and a strong increase in the pattern's relaxation time (critical slowing down) as the transition is approached. Obviously, if nonequilibrium phase transitions are a basic mechanism of self-organization and if, as hypothesized, the forces of evolution and self-organization form the core of biological order and coordination, it should be possible to discover nonequilibrium phase transitions and their signature features in real experiments. If not, the theory that coordination in living things is due fundamentally to self-organization must go the way of all beautiful theories that are negated by the facts.

Empirical Foundations of Coordination Dynamics: Pattern Generation, Stability and Phase Transitions

In coordination dynamics, the payoff of knowing collective variables and control parameters is high: they enable one to obtain the dynamical rules of behavior on a chosen level of description. By adopting the same strategy "one level down", the individual components and their dynamics may be studied and identified. It is the nonlinear interaction between the parts that creates coordinative patterns of the whole thereby building a bridge across levels of description (Table 1). This ability to derive phenomena from lower levels of description is at the core of what scientists usually mean by the word "understanding". In general, in complex living systems it is difficult to isolate the components and study their dynamics. The reason is that the individual components seldom exist outside the context of the functioning whole, and have to be studied as such.

If phase transitions hold the key to finding laws of coordination, where should we look for them? A central criterion for a law-based approach to coordination is reproducibility of the phenomenon in question. Although not everything is rhythmic, rhythms represent a wide variety of coordinated behaviors in a very large number of different biological systems at very many levels [24,70,71,97,119,174,212,215] and seem like an obvious entry point. One only has to look at the extensive field of so-called "central pattern generators" (CPGs) in invertebrate and vertebrate neurobiology to find remarkable similarities in the patterns that living creatures produce. Terms such as "swimming" CPG, "flight" CPG, and "locomotor" CPG reflect the reproducibility of patterns and their functional significance. Synchronization and desynchronization, frequency- and phase-locking are ubiquitous features of such patterns reflecting a high degree of neural and behavioral coordination (see Sects. "History of Coordination Dynamics: Synergy and Rhythmic Order" and "Coordination of Multiple Components: From Quadrupeds to Brains").

What then of phase transitions? And what connection exists, if any, between rhythms and phase transitions? It is well-known that quadrupeds and indeed many creatures including birds and fish exhibit characteristic gaits and may switch flexibly between them depending on circumstances. In the neurobiology literature, a key question is always "where are the switches in this thing"? [1]. Rather than assume the existence of switches, a priori, the scientific approach of coordination dynamics is to investigate the necessary and sufficient conditions that give rise to switching. Inspired by theories of self-organization in nature, coordination dynamics asks if switching may take the form of a nonequilibrium phase transition. The idea is not so far fetched as it seems. Many years ago, order-order transitions were hypothesized by Erwin Schrödinger [178] to be a crucial principle of biological organization and hypothesized to be the "new laws in the organism" [109].

Three Deceptively Simple Experiments

To investigate order-order transitions experimentally, consider an experimental paradigm introduced some years ago in which human beings are asked to move their two index fingers back and forth rhythmically [94,95]. In one condition (call it parallel, Fig. 1) they are told to alternate finger movements at a comfortable rate, one finger flexing in time as the other extends. In another separate condition (call it mirror) they are told to flex both fingers together and extend both fingers together at the same time. The key part of the experiment is that participants are instructed to increase the speed at which they perform these movements. For better experimental control a pacing metronome whose frequency can be systematically increased (say every 10 cycles called a plateau) may be used for subjects to follow. The main results are shown in Fig. 1 and described in the figure's caption.

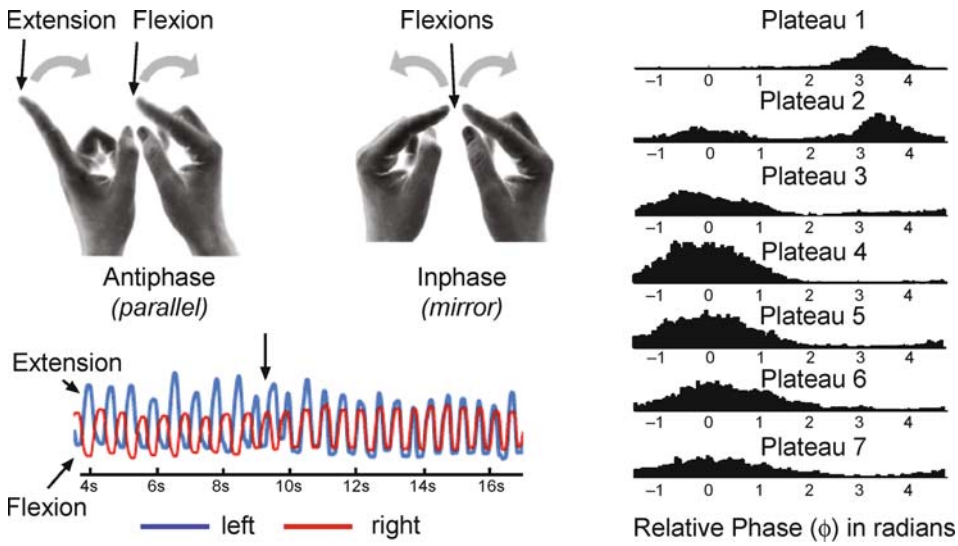

Coordination Dynamics, Figure 1

Phase transitions in bimanual movements. *Left side*. On the top are the two experimental conditions (parallel, antiphase and mirror, in phase) in the Kelso paradigm. The lower plot shows the time series of the finger movements in a representative run. As rate increases, trials initiated in the antiphase pattern spontaneously switch to the in-phase mirror pattern. The critical frequency is identified with an arrow. In contrast, trials initiated in the in-phase pattern do not switch as frequency increases (not shown). *Right side*. Distributions of relative phase between finger movements for plateaus of increasing frequency of movement. Initially the relative phase is concentrated at π radians, indicating antiphase is a stable pattern. On *plateau 2*, relative phase is concentrated around 0 and π radians, showing the bistability of antiphase and inphase coordination. For higher frequency plateaus the relative phase is concentrated at 0 radians, indicating that inphase is the only stable pattern of coordination (adapted from [7] with permission)

Experimental studies of bimanual rhythmic movement demonstrate that humans in the absence of learning procedures are able to produce two patterns of coordination at low frequency values, but only one – the symmetrical, in-phase mode – as frequency is scaled beyond a critical value. This is a remarkable result given the complexity of the nervous system, the body and their interaction with the world.

Consider another example, this time involving a human coordinating with an environmental signal [115,218]. In this experimental setup a single limb or finger is moved such that peak flexion occurs in between the beats of a pacing metronome, i. e. in a syncopated fashion. When the metronome frequency is increased, once again a critical value is reached where participants switch spontaneously to coordinating peak flexion on the beat, i. e. in a pattern of synchronization with the metronome. No such switching occurs when subjects begin in the synchronized mode of coordination. We may refer to these effects as a very basic example of *coordination between an organism and its environment*.

Now, consider the case of two people interacting with each other, an elementary form of interpersonal or *social coordination*. In this situation, each individual is instructed to oscillate a limb (the lower leg in this case) in the same or opposite direction to that of the other person [170]. In order to do the task, there must be a medium of interaction (vision, sound, touch, smell…) through which humans can couple. In this case, the two people watch each other (for details of this and other work see ► Social Coordination, from the Perspective of Coordination Dynamics). Then, either by an instruction from the experimenter or by following a metronome whose frequency is systematically increased, the social dyad speeds up their movements. When moving their legs up and down in the same direction, the two members of the dyad remain synchronized across a broad range of speeds. However, when moving their legs in the *opposite* direction (one person's leg extending at the knee while the other's is flexing), such is not the case. Once again at a certain critical rate participants change their behavior spontaneously so that their legs now move in the *same* direction.

The 'nonlinear' paradigm of coordination dynamics as illustrated in these three simple experiments has led to a wide range of investigations in many fields and a surprising variety of contexts (see Sect. "Collective Minds" for brief summary; also Books and Reviews) including detailed studies of underlying brain mechanisms using the full armamentarium of imaging technologies. Although

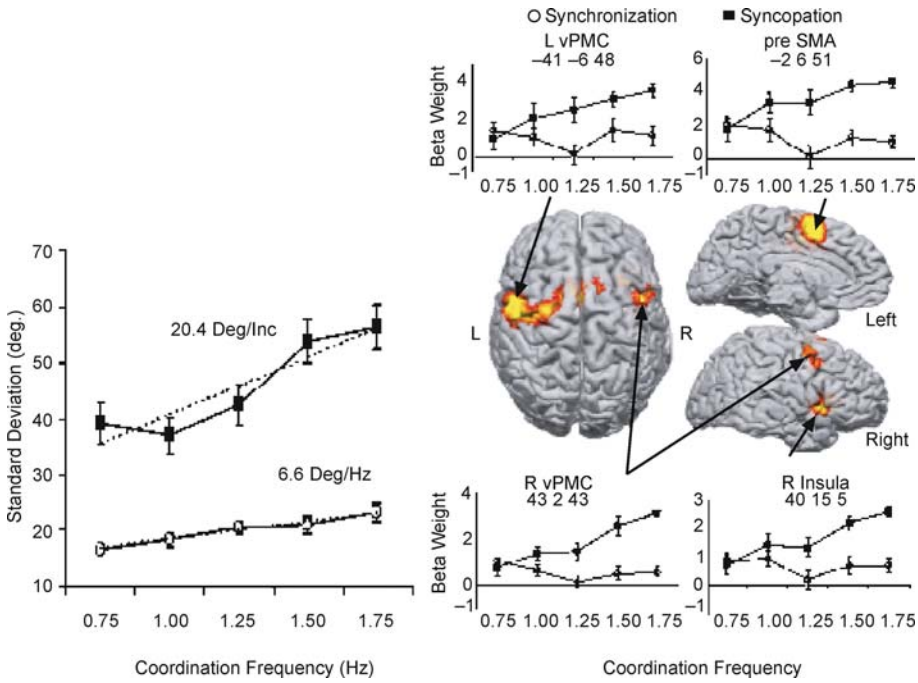

Coordination Dynamics, Figure 2

A pattern of generic results from combined brain and behavioral studies of coordination (adapted from [81]). On the *left* are behavioral results showing how relative phase variability is initially higher and increases more rapidly with the control parameter of frequency for syncopated antiphase (*filled squares*) than synchronized inphase coordination (*open circles*) with an external event. On the *right* are corresponding brain activations that comprise a network that is exquisitely sensitive to the stability of antiphase coordination. Areas depicted are left and right ventral premotor cortex (vPMC), pre supplementary motor area (pre-SMA) and right insula. The X Y Z Talairach coordinate of the peak voxel for each region is provided. Notice how the behavioral and brain data track each other

the findings would take us too far afield for present purposes ([19,20,191]; for reviews see [59,81]) two particular results are worth noting. First is that the basic paradigm has led to the first direct evidence of phase transitions in the human brain seen using both large scale electrode EEG [143,216] and SQuID arrays [35,61,62,108,123,124]. Second, and even more telling, is that fMRI evidence indicates that regardless of whether one is coordinating the two hands or coordinating with an external signal, a common network concerned with the *stability* of coordination is involved (see [81,156] for reviews).

The pattern of experimental findings described in this section illustrates an important conceptual distinction between coordination dynamics and other theories of self-organization [163]. In the latter, at bifurcation points or phase transitions, the system switches to a new, higher level of organization called a *dissipative structure*. Dissipative structures are so named because, compared with the "simpler" structures or patterns they replace, they require more energy to sustain them (ibid., p.xv). Not so in the order-order transitions of coordination dynamics. In coordination dynamics, the new organization that appears at bifurcation points is 'simpler' than the one it replaces and requires less energy. For example, brain electrical activity actually drops across the antiphase to inphase transition even though the system is being driven faster [124]. In Fig. 2, blood oxygen level dependent (BOLD) activity in certain brain regions is shown to increase as the stability of the antiphase pattern decreases. It is obvious that the increasing metabolic energy demands of the brain for antiphase relative to inphase coordination will diminish once the transition to in phase occurs. Thus, the key principle behind the 'simpler' self-organizing structures that emerge in coordination dynamics are based, not (or not only) upon energy per se as in the theory of dissipative structures, but on the system's *information processing demands*. Intuitively, the antiphase pattern is more difficult to coordinate as rate or frequency is increased causing the system to switch to a pattern that is easier to perform under the current conditions. Importantly, coordination dynamics replaces vague terms such as 'task difficulty' and 'task complexity' by quantitative behavioral measures of

stability and quantitative brain measures of BOLD and neuroelectric activity.

The Theoretical Modeling Strategy of Coordination Dynamics: Symmetry and Bifurcations

The three pieces of experimental evidence described above cut across entirely different kinds of things and events (auditory, visual and proprioceptive sensations, finger and leg movements, people and brains, etc.). The common denominator is that all these things and processes are meaningfully coupled together in time under particular boundary conditions (task instructions, environmental context, manipulated parameters, etc.). The phenomena observed hint at an aspect that any basic law should exhibit, namely that although the patterns of coordination observed are realized by different physical structures and physicochemical processes, laws and regularities are abstract and relational.

How then do we go about identifying the actual underlying laws? More specifically, how do we explain the coordinative phenomena observed experimentally? As stressed above, in contrast to certain physical systems like the laser, in biological coordination the path from the microscopic level to collective order parameters is not known and cannot (yet?) be derived from first principles like conservation laws. In coordination dynamics we have to: a) identify the order parameters or coordination variables and their low-dimensional dynamics empirically; b) determine the key control parameters that move the system though its coordinative states; and c) relate different levels though a study of the individual subsystems and their nonlinear interaction.

Determining the dynamics of coordination variables is non-trivial. In all three experimental situations, the relative phase ϕ or phase relation between the component elements appears to qualify as a suitable order parameter or coordination variable. The reasons are as follows: ϕ characterizes the patterns of spatiotemporal order observed, in phase and anti-phase; ϕ changes far more slowly than the variables that describe the individual coordinating components (e. g., position, velocity, acceleration, electromyographic activity of contracting muscles, neuronal ensemble activity in particular brain regions, etc.); ϕ changes abruptly at the transition and is only weakly dependent on parameters outside the transition; and ϕ appears to obey a dynamics in which the patterns may be characterized as *attractors* or *attractive states* of some underlying dynamical system. Since in all cases the frequency or rate clearly drives the system through different coordination patterns without actually prescribing them, frequency qualifies as a control parameter.

Determining the coordination dynamics means mapping observed, reproducibly stable patterns onto attractors of the dynamics. A general strategy is to assume sufficiently higher order dynamics and expand the vector field of these dynamics in a Fourier series:

$$\dot{\phi} = f(\phi) = a_0 + a_1 \sin(\phi) + a_2 \sin(2\phi) + \ldots + b_1 \cos(\phi) + b_2 \cos(2\phi) + \ldots . \quad (1)$$

Symmetry may be used to classify patterns and restrict the functional form of the coordination dynamics. Symmetry means "no change as a result of change": pattern symmetry means a given pattern is symmetric under a group of transformations. A transformation is an operation that maps one pattern onto another, e. g. in the first experimental case, left-right transformation exchanges homologous limbs within a bimanual pattern. If all relative phases are equivalent after the transformation, then the pattern is considered invariant under this operation. Symmetry serves two purposes. First it serves as a pattern classification tool allowing for the identification of basic coordination patterns that can be captured theoretically. Given a symmetry group, one can determine all invariant patterns. Second, imposing symmetry restrictions on the dynamics itself limits possible solutions and allows one to arrive at a coordination dynamics that contains the patterns as *different* stationary states of the *same* nonlinear dynamical system. In other words basic coordination patterns correspond to attractors of the relative phase for adequate parameter values. For example, left-right symmetry of homologous limbs leads to invariance under the transformation $\phi \to \phi$ so that the simplest dynamical system that accommodates the experimental observations is:

$$\dot{\phi} = f(\phi) = -a \sin(\phi) - 2b \sin(2\phi) \quad (2)$$

where ϕ is the relative phase between the movements of the two individuals, $\dot{\phi}$ is the derivative of ϕ with respect to time, and the ratio b/a is a control parameter corresponding to the movement rate in the experiment. An equivalent formulation of Eq. (1) is

$$\dot{\phi} = -\partial V(\phi)/\partial\phi \quad \text{with } V(\phi) = -a \cos\phi - b \cos 2\phi . \quad (3)$$

In the literature, Eqs. (2) and (3) are the equations at the collective level of the HKB model of coordinated behavior, after Haken, Kelso and Bunz [76,106,139]. Figure 3 (top) allows one to develop an intuitive understanding of the behavior of Eqs. (2),(3), as well as to connect the key concepts

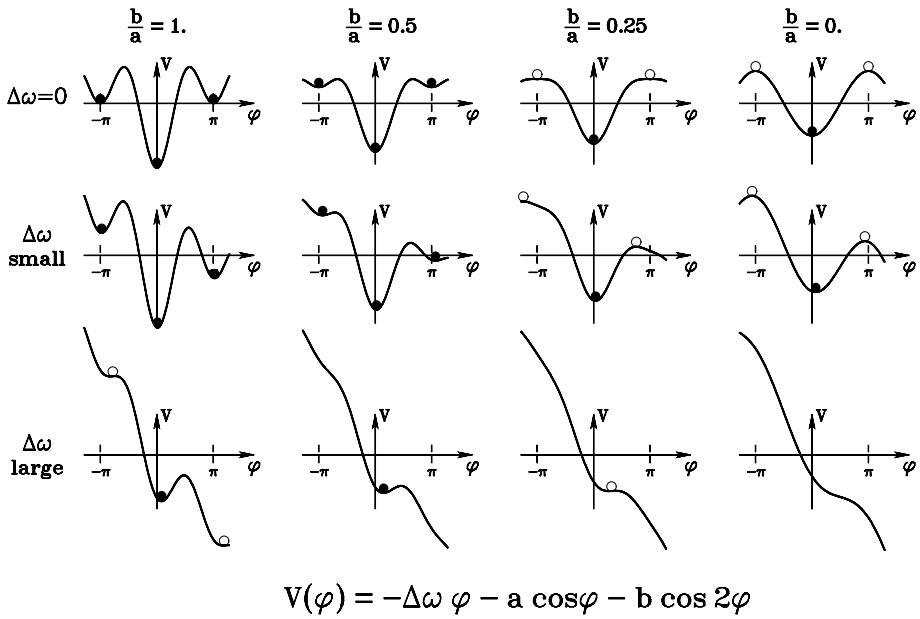

Coordination Dynamics, Figure 3
The potential, $V(\phi)$ of Eq. (3) (with $\Delta\omega = 0$) and Eq. (5) (with $\Delta\omega \neq 0$). *Black balls* symbolize stable coordinated behaviors and *white balls* correspond to unstable behavioral states (see text for details)

of stability and instability in self-organized dynamical systems to the observed experimental facts.

The dynamics can be visualized as a particle moving in a potential function $V(\phi)$. The minima of the potential are points of vanishing force, giving rise to stable solutions of the elementary coordination dynamics. As long as the speed parameter (b/a) is slow, meaning the cycle period is long, Eq. (3) has two stable fixed point attractors, coordinative states at $\phi = 0$ and $\phi = \pm\pi$ rad. Thus, two coordinated behavioral patterns coexist for exactly the same parameter values, the essentially nonlinear feature of *bistability* (Table 1). Such bi- and in general multi-stability is the dynamical signature of *multifunctionality* which can be seen at many levels in living things. As the ratio b/a is decreased, meaning that the cycle period gets shorter as the system speeds up, the formerly stable fixed point at $\phi = \pm\pi$ rad becomes unstable, and turns into a *repellor* (open circles). Any small perturbation will now kick the system into the *basin of attraction* of the stable fixed point corresponding to an in-phase coordination pattern at $\phi = 0$. Notice also that once there, the system's behavior will stay in the in-phase attractor, even if the direction of the control parameter is reversed. This is called *hysteresis*, a basic form of memory in nonlinear dynamical systems.

What about the individual components? Research has established that these take the form of self-sustaining oscillators, archetypal of all time-dependent behavior [12,14,47,92,93]. The particular functional form of the oscillator need not occupy the reader here (see ► Movement Coordination which uses empirical facts and symmetry arguments to restrict and thereby identify the component dynamics). More important is the nature of the nonlinear coupling that produces emergent coordination. The simplest, perhaps fundamental biophysical coupling that guarantees all the observed emergent properties of coordination: multistability, flexible switching among coordination states and primitive memory, is:

$$K_{12} = \left(\dot{X}_1 - \dot{X}_2\right)\left\{\alpha + \beta(X_1 - X_2)^2\right\} , \qquad (4)$$

where X_1 and X_2 are the individual components and α and β are coupling parameters. Notice that the 'next level up', the level of coordinated behavioral patterns and the dynamical rule that governs them (Eqs. (2) and (3)), can be *derived* from the level below, the individual components and their nonlinear interaction. One may call this *constructive reductionism*: by focusing on adjacent levels, under the boundary constraints of the task, the individual parts can be put together to create the behavior of the whole.

The basic self-organized dynamics, Eqs. (2) and (3) have been extended in numerous ways, only a few of which are mentioned here.

- *Critical slowing down* and *enhancement of fluctuations*. Introducing stochastic forces into Eqs. (2) and (3) ([175,210,211] see Chap. 11 in [74] and [101,121] for

a thorough discussion) allows key predictions of coordination dynamics to be tested and quantitatively evaluated [113,121,172]. Critical slowing is easy to understand from Fig. 3 (top). As the minima at $\phi = \pm\pi$ become shallower and shallower, the time it takes to adjust to a small perturbation takes longer and longer. Thus, the local relaxation time is predicted to increase as the instability is approached because the restoring force (given as the gradient of the potential) becomes smaller. Likewise, the variability of ϕ is predicted to increase due to the flattening of the potential near the transition point. Both predictions have been confirmed in a wide variety of experimental systems, including recordings of the human brain ([81,100,174] for review).

- *Symmetry breaking*. Notice that Eqs. (2) and (3) are *symmetric*: the dynamical system is 2π periodic and is identical under left-right reflection (ϕ is the same as $-\phi$). This assumes that the individual components are identical, which as remarked upon earlier, is seldom the case in living things where symmetries are broken all the time.

In terms of the development of the theory, an important experimental example of symmetry breaking is the case of coordinating movement with a visual stimulus: visual stimuli and limb movement are obviously not equivalent. Thus $\phi \to \phi$ symmetry cannot be assumed. This means that symmetry partners of coordination patterns with systematic phase leads or lags do not coexist at the same parameter values. To accommodate this fact, a term $\Delta\omega$ is incorporated into the dynamics [115]:

$$\begin{aligned} \dot{\phi} &= \Delta\omega - a\sin\phi - 2b\sin 2\phi\,, \quad \text{and} \\ V(\phi) &= -\Delta\omega\phi - a\cos\phi - b\cos 2\phi \end{aligned} \tag{5}$$

for the equation of motion and the potential respectively.

Note that Eq. (5) falls out naturally from an analysis of the oscillators, ω_1 and ω_2, viz.

$$\dot{\phi} = \frac{\omega_1^2 - \omega_2^2}{2\Omega} + (\alpha + 2\beta R^2)\sin\phi - \beta R^2 \sin 2\phi \tag{6}$$

for

$$\Delta\omega = \frac{\omega_1^2 - \omega_2^2}{2\Omega} \approx \omega_1 - \omega_2 \tag{7}$$

with

$$\begin{aligned} a &= -(\alpha + 2\beta R^2) \\ b &= \tfrac{1}{2}\beta R^2\,. \end{aligned} \tag{8}$$

Small values of $\Delta\omega$ shift the attractive fixed points (Fig. 3 middle) in an adaptive manner. For larger values of $\Delta\omega$ the attractors disappear entirely (Fig. 3 bottom) causing the relative phase to drift: no coordination between the components appears to be possible. Note, however, that the dynamics still retain some curvature (Fig. 3 bottom right): even though there are no attractors there is still attraction to where the attractors used to be. The reason is that the difference ($\Delta\omega$) between the individual components is sufficiently large that they do their own thing, while still retaining a *tendency* to cooperate. The introduction of the symmetry breaking term $\Delta\omega$ in Eq. (5) changes the entire coordination dynamics (layout of the fixed points, bifurcation structure) of the original HKB system in which $\Delta\omega = 0$. This is important to realize because it is the subtle interplay between the coupling ($k = b/a$) and the symmetry breaking term $\Delta\omega$ in Eq. (5) that gives rise to *metastability*.

Sometimes in the literature Eq. (5) is referred to collectively as the Haken–Kelso–Bunz equation. Though convenient, this is technically incorrect and fails to recognize both the intellectual contributions to its extension and the conceptual consequences thereof. For reasons of symmetry and simplicity, the original HKB equation did not contain the symmetry breaking term, $\Delta\omega$ [115] nor did it treat fluctuations explicitly [175] both of which are crucial for capturing the broad range of phenomena observed and testing further predictions. In particular, without $\Delta\omega$ there is: a) no fixed point shift, a sign of adaptation to changing circumstances, see Fig. 3 and 4; b) the bifurcation is a saddle node not, as in the original HKB equation, a pitchfork. These are different normal forms [98]; and c) most important of all, the original HKB equation does not and cannot exhibit metastability which is the key to understanding the complementary relationship between the synergic tendency of the elements to couple (integration) and at the same time to express their individual differences (segregation). The oscillators in the original HKB formulation were identical thereby excluding metastability. For these reasons, it seems wise to refer to Eq. (5) (with its stochastic aspect included) as the extended HKB equation.

Equation (5) is a bit strange. Even though it is an order parameter equation of motion that describes *coordinative* behavior (in words, phi dot is a function of phi), it includes also a parameter ($\Delta\omega$) that arises as a result of differences among the *individual components*. Equation (5) is thus a strange mixture of the whole and the parts, the global and the local, the cooperative and the competitive, the collective and the individual. Were the components identical, $\Delta\omega$ would be zero and we would not see component differences affecting the behavior of the whole (Fig. 3 top row). Equation (5) would simply reflect the behavior of the collective untarnished by component properties, a purely

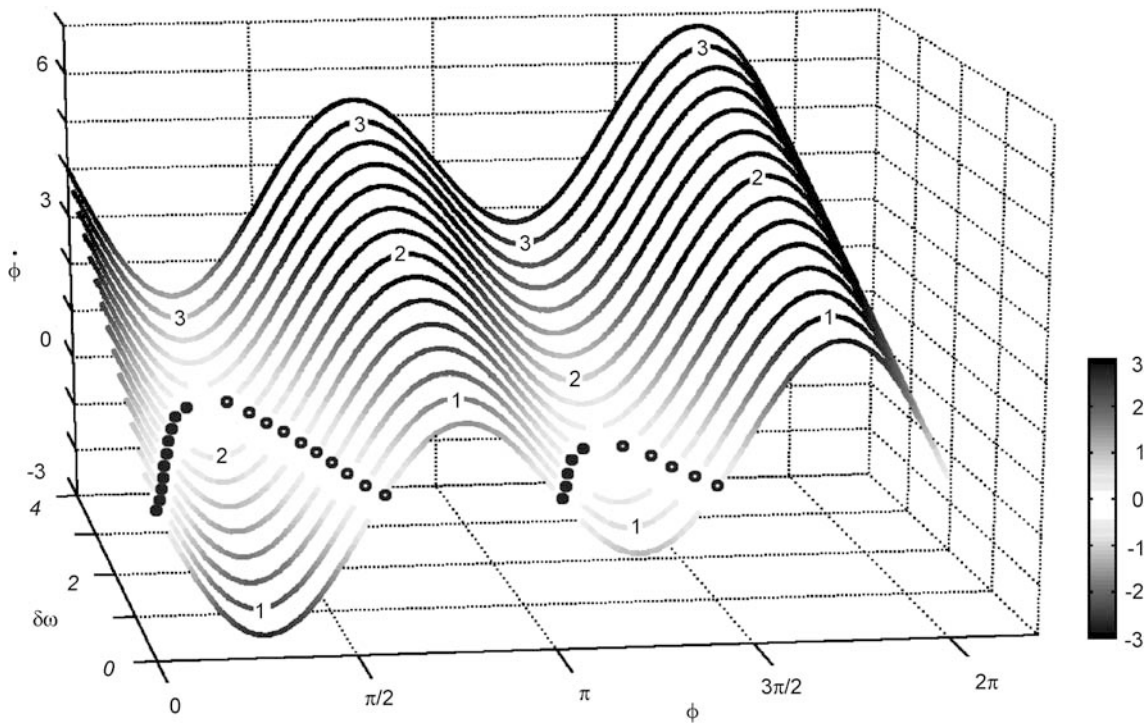

Coordination Dynamics, Figure 4

Elementary coordination law (Eq. (5)). Surface formed by a family of flows of the coordination variable ϕ (in radians) as a function of its time derivative $\dot{\phi}$ for increasing values of $\Delta\omega$. For this example, the coupling is fixed: $a = 1$ and $b = 1$. When $\dot{\phi}$ reaches zero (*flow line becoming white*), the system ceases to change and fixed point behavior is observed. Note that the fixed points here refer to *emergent coordination states* produced by nonlinearly coupled elements. Stable and unstable fixed points at the intersection of the flow lines with the isoplane $\dot{\phi} = 0$ are represented as filled and open circles respectively. Three representative *lines labeled 1 to 3* illustrate the different régimes of the coordination dynamics. Following the *flow line 1* from left to right, two stable fixed points (*filled circles*) and two unstable fixed points (*open circles*) exist. This flow belongs to the multistable (here bistable) régime of dynamics. Following *line 2* from left to right, one pair of stable and unstable fixed points is met on the left, but notice the complete disappearance of fixed point behavior on the right side of the figure. That is, a qualitative change (bifurcation; phase transition) has occurred due to the loss of stability of the coordination state near antiphase, π rad. The flow now belongs to the monostable régime. Following *line 3* from left to right, no stable or unstable fixed points exist yet a subtle form of coordination – neither completely ordered (synchronized) nor completely disordered (desynchronized) – still remains. This is the metastable régime

emergent interaction – the HKB equation. It is the fact that *both* the components *and* their (nonlinear) interaction appear at the same level of description that gives rise to the array of coexisting tendencies characteristic of metastability. The history of coordination (Sect. "History of Coordination Dynamics: Synergy and Rhythmic Order") may now be seen in a new light: Eq. (5) is a basic representation of a *synergy*, a low dimensional dynamic of a metastable organization in which the tendency of the parts to act together coexists with a tendency of the parts to do their own thing (see Chap. 4 in [100]). If indeed the synergy is a unit of life and mind as proposed in [105,107] then it is metastability that endows the synergy with robustness and flexibility, enabling the same parts to participate in multiple functions.

Metastable Coordination Dynamics

From States to Tendencies

Etymologically, 'metastability', comes from the latin '*meta*' (beyond) and '*stabilis*' (able to stand). In coordination dynamics, metastability corresponds to a régime near a saddle-node or tangent bifurcation in which stable coordination states no longer exist (e. g., in-phase synchronization where the relative phase between coordinating components lingers at zero), but attraction remains to where those fixed points used to be (see Fig. 3, bottom row). This gives rise to a dynamical flow consisting of *both* phase trapping and phase scattering.

To best visualize the emergence of metastability, Fig. 4 shows the flow of the dynamics for the elementary coor-

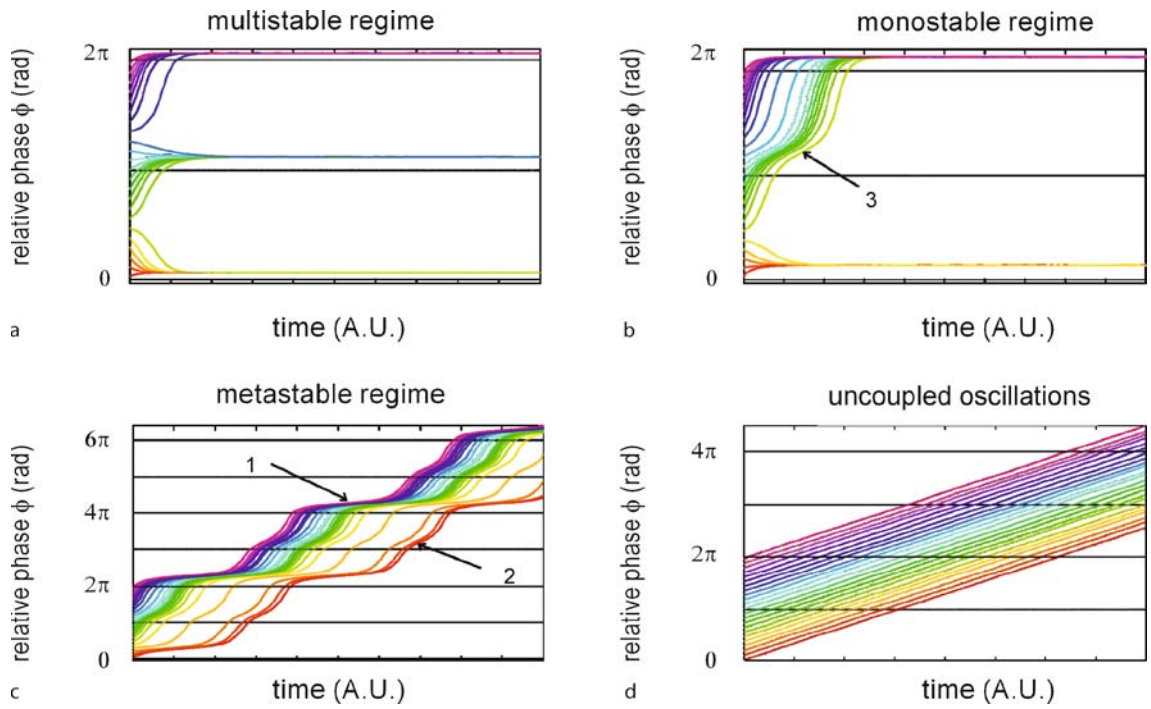

Coordination Dynamics, Figure 5

How the key coordination variable or order parameter of the elementary coordination law (Eq. (5)) behaves over time. Shown is a family of trajectories of the relative phase ϕ over time (in Arbitrary Units) arising from a range of initial conditions sampled between 0 and 2π radians, in the multistable (a), monostable (b) and metastable régimes (c) of Eq. (5). For the uncoupled case (d) the trajectories never converge indicating that the oscillations are completely independent of each other. Trajectories in the multistable régime (a) converge either to an attractor located slightly above $\phi = 0$ rad modulo 2π or to another attractor located slightly above $\phi = \pi$ rad modulo 2π. In the monostable régime (b), trajectories converge to an attractor located slightly above $\phi = 0$ rad modulo 2π. In the trajectories of relative phase for the metastable régime (c unwrapped to convey continuity), there is no longer any persisting convergence to the attractors, but rather a succession of periods of rapid drift (*escapes*) interspersed with periods inflecting toward, but not remaining on the horizontal (*dwells*). Note dwells near $\phi = 0$ rad modulo 2π in the metastable régime (e. g. dwell at about 4π rad annotated 1 in c) and nearby $\phi = \pi$ rad modulo 2π (dwell at about 3π rad annotated 2 in c) are reminiscent of the transient obtained for certain initial conditions in the monostable régime (Fig. 5b, annotation 3). The key point is that in the metastable régime the system's behavior is a blend of coupled and independent behavior

dination law (Eq. (5)) across a range of $\Delta\omega$ values with the coupling parameter, $k = b/a = 1$ fixed. Stable fixed points (attractors) are presented as filled circles and unstable fixed points (repellors) as open circles. Here, fixed points of the coordination dynamics correspond to phase- and frequency relationships between oscillatory processes.

The surface shown in Fig. 4 defines three regions under the influence of the symmetry breaking term $\Delta\omega$. In the first region present in the lower part of the surface, the system is multistable: two stable attracting fixed points (filled circles) represent possible alternative states. Which one the system settles in depends on initial conditions and the size of the basin of attraction. In an intermediate region, following the line labeled 2 from left to right, the weakest attractor near anti-phase (right side) disappears after it collides with its associated repellor somewhere near $\Delta\omega = 1.3$, but the strongest attractor (left side) is still present as well as its repellor partner. Finally in the third region in the upper part of the surface, the dynamics become metastable. Following the line labeled 3 from left to right, no fixed points exist anymore: this part of the surface no longer intersects the isoplane $\dot{\phi} = 0$ where the fixed points are located. Strictly speaking coordination *states* qua frequency- and phase-synchrony no longer exist in the metastable régime of the coordination dynamics. Metastability is thus a subtle blend of coupling and intrinsic tendencies in which behavior is neither completely ordered (synchronized) nor completely disordered (desynchronized). Both tendencies coexist.

How do individual and coordination behavior evolve in time in the metastable régime? A unique flow now exists in which the dynamics may be characterized by

places where the trajectory of the coordination variable converges and pauses around the horizontal and places where the trajectory drifts or diverges from the horizontal. Let us define the former as a *dwell time*, and the latter as an *escape time*. In Fig. 5c we show two locations for the dwell times: one that lingers a long time before escaping (e. g. Fig. 5c, annotation 1) slightly above the more stable in-phase pattern near 0 rad (modulo 2π), and the other that lingers only briefly (e. g. Fig. 5c, annotation 2) slightly above π (modulo 2π). These inflections recur over and over again as long as the system self-organizes in the metastable régime, i. e. as long as it does not undergo a phase transition to a locked or unlocked state. Despite the complete absence of phase-locked attractors, the coordinating elements in the metastable régime do not behave totally independently. Rather, their interdependence takes the form of dwellings (phase gathering tendencies) nearby the remnants of the fixed points (cf. Fig. 3 bottom; Figs. 4, 5c) and may be nicely expressed by concentrations in the histogram of the relative phase (see Chap. 4 in [100]).

Recently metastability has been hailed as a "new principle" of coordination in the brain and has been embraced by a number of neuroscientists as playing a role in various cognitive functions, even consciousness itself (e. g. [44,45,57,58,111,127,187,212]). According to a recent review [49]:

> *Metastability is an entirely new conception of brain functioning where the individual parts of the brain exhibit tendencies to function autonomously at the same time as they exhibit tendencies for coordinated activity* [19,20,97,100].

For Coordination Dynamics, metastability's significance lies not in the word itself but in what it means for understanding coordination in living things. In coordination dynamics, as shown in its most elementary form (Eq. (5)), metastability is not a concept or an idea, but a direct result of the broken symmetry of a system of (nonlinearly) coupled (nonlinear) oscillators. Such a design principle for the brain seems highly plausible given that rhythms in the brain are ubiquitous, operate over a broad range of frequencies and are strongly associated with various sensory, motor and cognitive processes [10,24,100].

The Creation of Information

There is another reason for proposing metastable coordination dynamics as the essential way the brain and perhaps all complex organizations work. It concerns an analogy to how physicists understand how we know the universe we live in. According to Quantum Mechanics, out of a universe in which quantum indeterminacy rules – the wave function is spread out over all of space – nature selects an alternative. Information is thereby created. The way this is done in practice is that a device is built in which an interactive material is placed in a physically, electrically or chemically metastable state. According to the late quantum measurement theorist, H.S. Greene [69]:

> *It is the observable transition between this metastable state and a more stable state that conveys the essential information concerning a sub-microscopic event that would otherwise go undetected … The functional material of the detector must be macroscopic and in a metastable state which allows the quantal interaction to become manifest at the macroscopic level.* (see p. 173 in [69])

This is how some physicists view the creation of information: bit from it, as it were (in contrast to John Archibald Wheeler's 'it from bit'). Quantum Mechanics thus implies the creation of new information in the process of measurement and observation. Likewise, we have seen in the human brain that information (as a marginally coupled, phase-locked state) is created and destroyed in the metastable régime of the coordination dynamics, where tendencies for apartness and togetherness, individual and collective, segregation and integration, phase synchrony and phase scattering *coexist*. New information is created because the system operates in a special régime where the slightest nudge will put it into a new coordinated state. In this way, the (essentially nonlinear) coordination dynamics creates new, informationally meaningful coordination states that can be stabilized over time. The *stability of information over time* is guaranteed by the coupling between component parts and processes and may constitute a dynamic kind of (non-hereditary) memory. It does not seem a big step then to say that once created, this information can then guide, modify and direct the system's dynamics. As we shall see in Sect. "Modifying Coordination: Meaningful Information" studies of intentional change, environmental change, learning and so forth have demonstrated both empirically and theoretically that an intentional goal – as memorized information – acts in the same information space as the coordination dynamics ([114,173]; see also [141]).

Coordination of Multiple Components: From Quadrupeds to Brains

Phase- and frequency synchronization, the coupling among oscillatory processes, are an example, par excel-

lence, of self-organization in natural systems [76,119,132,161,219]. Think of the famous clocks on the wall, no two exactly the same in frequency, but somehow ticking perfectly in time with each other. The weakest of coupling, whether through vibration in the wall or displacement of the air, enables the clocks to be mutually coordinated without any coordinator at all. Similar phenomena have been seen in the brain and have been hypothesized to play a key role in "binding": oscillations in the brain appear to be coupled or "bound" together into a coherent network when people attend to a stimulus, perceive, think and act (e. g., [34,43,67,184,214]). For example, synchronization in the so-called gamma band (approx. 30–50 Hz) has been proposed as a neural correlate of consciousness. The Journal *Science* announced the synchronization effects observed in monkey cortex as "The Mind Revealed" (see also [34]).

The brain, it seems, has latched on to phase synchrony as a principle of self-organization. Though the connection is seldom made, phase- and frequency- synchronization is typical of central pattern generators (CPGs), neural circuits in vertebrates and invertebrates that generate timing sequences without feedback from the periphery or the help of reflexes (see Sect. "Empirical Foundations of Coordination Dynamics: Pattern Generation, Stability and Phase Transitions"; [70,71]). Indeed, it is the temporal order observed that allows us to talk about 'pattern generators' in the first place. Though the specific mechanisms are obviously different between the visual cortex of the monkey and the stomatogastric ganglion of the lobster, the dynamic patterns are the same, hinting at the source of an underlying principle [97,100,174]. But what form might the coordination dynamics of the brain take?

Obviously when it comes to the brain there are, in principle, very many regions to coordinate. In practice, however, only a restricted set of regions appear to be functionally connected during particular tasks (see, e. g. Fig. 2). The idea, then, is that one could use the Central Pattern Generator (CPG) design for quadrupedal locomotion [32,66,176] as a basic model of interaction among cortical pattern generators. This is not as far fetched as it seems. It is likelier than not that the precursors to the structure and function of the cerebral cortex are self-contained circuits in the spinal cord and brainstem that generate intrinsic patterns of rhythmic activity [70,71,220]. Such CPGs typically work by transforming tonic driving inputs into detailed spatiotemporal patterns of (usually oscillatory) activity. Several of the properties of CPGs are conserved throughout evolution rendering them a likely candidate for the basic building blocks of the brain [72]. The hypothesis proposed here is that cortical pattern generators may underlie the coordination that is needed for everything the brain is purported to do – think, feel, remember, act, socialize, etc.

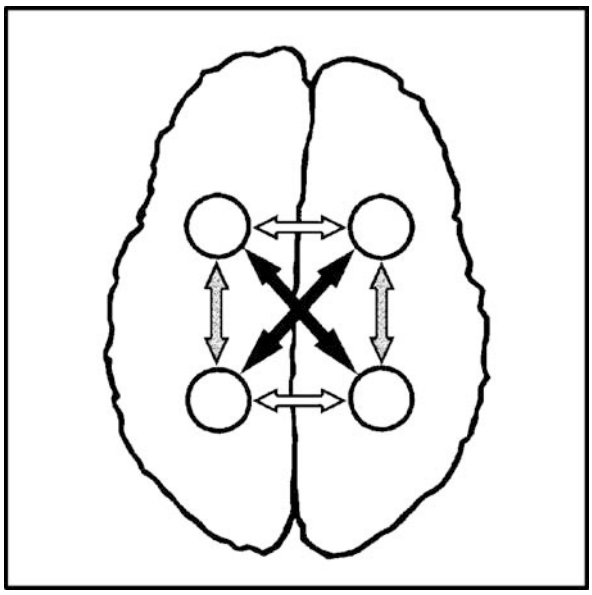

Coordination Dynamics, Figure 6
A schematic of brain coordination dynamics among four brain regions. Each *circle* represents an area of the brain capable of intrinsic oscillation and the *arrows* correspond to connections among brain areas giving rise to cortical pattern generation (see text)

Following the footsteps of basic coordination dynamics, in the quadruped analogy each "limb" corresponds to a neural region capable of intrinsic oscillatory activity and the patterns emerge from (broken) symmetries and changes in coupling between neural regions. For example, Fig. 6 shows a cartoon of the neuroanatomical connections underlying the anterior-posterior coordination of the hemispheres of the brain.

The relevant variables are four phase variables, ϕ_{ij} (with $i \in$ {right, left hemisphere}, $j \in$ {anterior, posterior} characterizing the oscillatory behavior of each brain area with respect to its timing. Much research on coordination dynamics shows that the *relative phase* is a key coordination variable or order parameter although it is quite possible that amplitudes and frequencies are important variables too [4,57,111]. For the sake of simplicity, we stick to the case of interareal cortical coordination between 4 brain regions, where a set of 3 relative phases suffices to characterize any pattern uniquely.

As already illustrated, a key notion is to use symmetry to classify patterns and restrict the functional form of the coordination dynamics. Here, pattern symmetry means

a given cortical pattern is symmetric under a group of transformations. As we have noted, a transformation is an operation that maps one pattern onto another, e. g. the anterior-posterior (a-p) transformation exchanges anterior and posterior regions within a cortical pattern. If all relative phases are equivalent after the transformation, then the pattern is considered invariant under this operation.

Symmetry serves as a pattern classification tool allowing for the identification of basic cortical patterns that can be captured theoretically. Given a symmetry group, one can determine all invariant patterns. For example, certain idealized cortical patterns are invariant under the symmetry group generated by the following operations: exchange of anterior and posterior, exchange of left and right, and inversion of all phases (inversion of time). A good way to illustrate these patterns is with phase pictograms.

Imposing symmetry restrictions on the dynamics itself limits possible solutions and allows one to arrive at a coordination dynamics that contains the patterns as *different* stationary states of the *same* nonlinear dynamical system. In other words basic cortical patterns correspond to attractive states of the relative phase for adequate parameter values:

$$\begin{aligned}\dot{\phi}_{ij} = \sum_{n=1}^{\infty}\{&\underbrace{A_n \sin(n(\phi_{ij} - \phi_{\hat{i}j}))}_{\text{Homologous contralateral coupling (white arrows)}} \\ &+ \underbrace{C_n \sin(n(\phi_{ij} - \phi_{i\hat{j}}))}_{\text{Ipsilateral coupling (gray arrows)}} \\ &+ \underbrace{E_n \sin(n(\phi_{ij} - \phi_{\hat{i}\hat{j}}))}_{\text{Nonhomologous contralateral coupling (black arrows)}}\}\end{aligned} \tag{9}$$

where A_n, C_n and E_n are parameters and a hat over an index means that the opposite value is taken, e. g. if $i =$ right, then $\hat{i} =$ left.

To analyze the solutions to the phase dynamics, for the sake of simplicity the coordination dynamics may be restricted to second order. Higher orders generate parameter régimes where many patterns may coexist; first order removes the possibility that some patterns may coexist. Also, diagonal coupling (black arrows in Fig. 6) may be neglected by setting $E_n = 0$. Notice that an effective diagonal coupling still exists because two couplings are sufficient to stabilize any pattern of activity among 4 cortical areas. Moreover, it is easier to generalize a system with contralateral and ipsilateral couplings to systems in which more areas are involved.

In sum, following exactly the basic theoretical modeling strategy (Eq. (1)) the dynamical system takes the following form:

$$\begin{aligned}\dot{\phi}_{\mathrm{rp}} &= A_1[\sin(\phi_{\mathrm{rp}} - \phi_{\mathrm{lp}}) + \sin(\phi_{\mathrm{la}})] \\ &\quad + A_2[\sin(2(\phi_{\mathrm{rp}} - \phi_{\mathrm{lp}})) + \sin(2\phi_{\mathrm{la}})] \\ &\quad + 2C_1 \sin(\phi_{\mathrm{rp}}) + 2C_2 \sin(2\phi_{\mathrm{lp}}) \\ \dot{\phi}_{\mathrm{lp}} &= A_1[\sin(\phi_{\mathrm{lp}} - \phi_{\mathrm{rp}}) + \sin(\phi_{\mathrm{la}})] \\ &\quad + A_2[\sin(2(\phi_{\mathrm{lp}} - \phi_{\mathrm{rp}})) + \sin(2\phi_{\mathrm{la}})] \\ &\quad + C_1[\sin(\phi_{\mathrm{lp}} - \phi_{\mathrm{la}}) + \sin(\phi_{\mathrm{rp}})] \\ &\quad + C_2[\sin(2\phi_{\mathrm{lp}} - \phi_{\mathrm{la}})) + \sin(2\phi_{\mathrm{rp}})] \\ \dot{\phi}_{\mathrm{la}} &= 2A_1 \sin(\phi_{\mathrm{la}}) - \phi_{\mathrm{lp}}) + 2A_2 \sin(2\phi_{\mathrm{la}}) \\ &\quad + C_1[\sin(\phi_{\mathrm{la}} - \phi_{\mathrm{lp}}) + \sin(\phi_{\mathrm{rp}})] \\ &\quad + C_2[\sin(2(\phi_{\mathrm{la}} - \phi_{\mathrm{lp}})) + \sin(2\phi_{\mathrm{rp}})]\,.\end{aligned} \tag{10}$$

Solving $\dot{\phi}_{ij} = 0$ yields stationary solutions that correspond to idealized cortical "gaits". Trot, pace, gallop and jump patterns may be identified as multistable or monostable solutions in various parameter régimes. Patterns of lower symmetry can also be captured. Obviously the foregoing analysis is intended for illustrative purposes only. The examples provided in Fig. 7 and 8 are only a few of very many possible cortical patterns that can be obtained by further symmetry groups. It is important to empha-

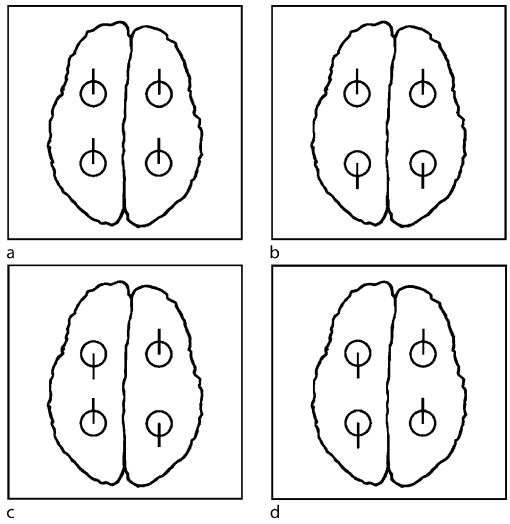

Coordination Dynamics, Figure 7

Brain phase pictograms. Each brain area is represented as a *circle*. The spatial arrangement of the circles viewed from looking down on the top of the head represents the brain's hypothesized anterior-posterior and left-right functional organization. Phase is represented by the angle the stick makes on each circle, with the reference phase being zero for the right frontal region. If all phases are rotated by the same amount in the same direction, the cortical coordination pattern remains the same. The patterns in a, b, c and d are those idealized cortical patterns that remain invariant under anterior-posterior, left-right and time inversion operations. Notice that the relative phase between any two brain regions is either inphase or antiphase

size that not all patterns in a given symmetry group are observable. Which ones are actually observed is dictated by the coordination dynamics, in particular a given pattern's *stability* which, as we have seen, can be measured (cf. Sect. "Empirical Foundations of Coordination Dynamics: Pattern Generation, Stability and Phase Transitions"). As in the simpler cases of coordination dynamics described above, *switching* among cortical patterns is due, not to switches per se but to instabilities – phase transitions or bifurcations in the phase dynamics. Neuromodulators are candidate control parameters capable of sculpting cortical patterns by leading the system through instabilities [130]. Moreover, when the oscillatory frequencies in the anterior and posterior regions of the brain are slightly different, a kind of partial coordination among cortical regions may occur. This is exactly the *metastable* coordination dynamics of the brain as described in the previous section. Here again, the key point is that the rules of the game appear to be run by principles of coordination dynamics and symmetry. As always, experiments are now needed to test this hypothesis. EEG measures of cross-frequency phase synchrony of the human brain may reflect a start in this direction (e. g., [84,158]). More direct attempts are underway in our laboratory [199].

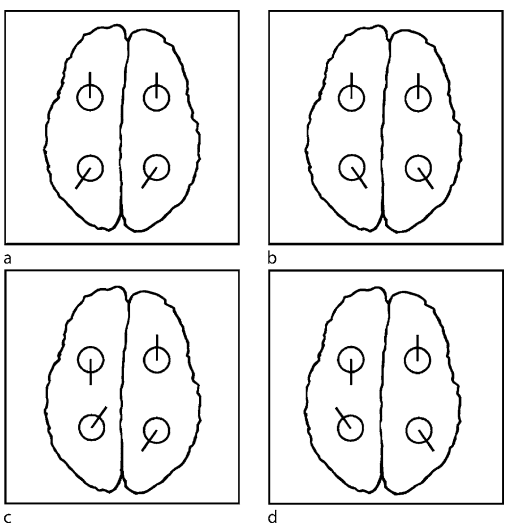

Coordination Dynamics, Figure 8
Representative brain phase pictograms corresponding to cortical patterns of lower symmetry. Here the anterior-posterior symmetry is dropped and the cortical patterns that remain form two one parameter families. One family consists of in-phase ordering *within* anterior and posterior areas and any fixed phase relation *between* anterior and posterior regions (a,b). The other family (c,d) consists of anti-phase relations within frontal and anterior regions and any fixed phase relations between them

"Collective Minds"

The basic coordination dynamics for two and four nonlinearly interacting components (Eqs. (2)–(10)) can readily be elaborated as a model of emergent coordinated behavior or "group cohesion" among very many anatomically different components (see, e. g. [5]). Self-organized behavioral patterns such as singing in a group or making a "wave" during a football game are common, yet unstudied examples. By virtue of information exchange nearest neighbors adjust their motions to each other generating, and being influenced by, their social environment. Recently, Néda and colleagues [148,149] have examined a simpler group activity: applause in theater and opera audiences in Romania and Hungary. After an exceptional performance, initially thunderous incoherent clapping gives way to slower, synchronized clapping. Measurements indicate that the clapping period suddenly doubles at the onset of the synchronized phase, and slowly decreases as synchronization is lost. This pattern is a cultural phenomenon in many parts of Europe: a collective request for an encore. Increasing frequency (decreasing period) is a measure of the urgency of the message, and culminates in the transition back to noise when the performers reappear. These results are readily explained by a model of a group of globally coupled nonlinear oscillators [132]:

$$\frac{\mathrm{d}\phi_k}{\mathrm{d}t} = \omega_k + \frac{K}{N}\sum_{j=1}^{N}\sin(\phi_j - \phi_k) \tag{11}$$

in which a critical coupling parameter, K_c determines the different modes of clapping behavior. K is a function of the dispersion (D) of clapping frequencies:

$$K_c = \sqrt{\frac{2}{\pi^3}}\mathrm{D}\,. \tag{12}$$

During fast clapping, synchronization is not possible due to the large dispersion of clapping frequencies. Slower, synchronized clapping at double the period arises when small dispersion appears. Period doubling rhythmic applause tends not to occur in big open-air concerts where the informational coupling among the audience is small. K can also be societally imposed. In Eastern European communities during communist times, synchronization was seldom destroyed because enthusiasm was often low for the "great leader's" speech. For people in the West, the cultural information content of different clapping patterns may be quite different. Regardless, the mathematical descriptions for coordinated behavior – of social dyads and the psychology of large groups – are remarkably similar.

Modifying Coordination: Meaningful Information

Unlike the behavior of inanimate things, the self-organizing dynamics of animate behavior is based on information (Table 1), though not in the standard sense of data communicated across a channel [180]. In coordination dynamics, *collective or coordination variables are context-dependent and intrinsically meaningful*. Context-dependence does not imply subjectivity and lack of reproducibility. Nor does it mean that every new context requires a new collective variable or order parameter. As we have seen already, for example, within- and between-person coordinated behaviors are described by the same self-organizing coordination dynamics. One of the consequences of identifying the latter is that in order to *modify or change* the system's behavior, new information (in the form say, of an environmental input, a task to be learned, or an intention to change behavior) is expressed in terms of parameters acting on system-relevant collective dynamics. On the one hand, the benefit of identifying collective variables is that they embrace the full complexity of the system and hence provide the relevant information about *what* to modify. On the other, the collective variable dynamics – prior to the introduction of any new information – influences how that information can be used. The upshot is that information is not lying out there as mere data: information is meaningful to the extent that it modifies, and is modified by, the collective variable dynamics.

A minimum mathematical form for the full coordination dynamics which encompasses both spontaneous self-organizing tendencies and specific parametric influences is

$$\dot{\phi} = f(\phi) + f\,\mathrm{inf}(\phi) \tag{13}$$

where the first term is the typical so-called "intrinsic dynamics" e. g., of Eq. (5) or Eq. (10) and the second term represents 'informational forcing', i. e., a perturbation of the vector field of the dynamics attracting the system toward a required coordination pattern. It is important to emphasize that the plus sign in Eq. (13) is for operational purposes only, affording the measurement of the complementary contributions to the coordination dynamics of both spontaneous and directed (parametric) influences. The conceptual advantage of Eq. (13) is that information acts in the same space as the collective variables that define the intrinsic coordination patterns, i. e., those patterns that characterize spontaneous coordination tendencies. Thus, information is not arbitrary with respect to the dynamics. A corollary of this formulation is that information has no meaning outside its influence on the intrinsic dynamics. They are cut, as Sheets-Johnstone [182] remarks, from the same dynamic cloth.

Intentional Dynamics

Self-organizing processes, in the manner of Haken's synergetics, provide a theoretical foundation for all forms of coordination. However, we do not want to throw the baby out with the bathwater. Coordinated behavior often has a goal-directedness to it as well. We humans, for example, have no doubt whatsoever that it is us, and us alone, that direct the motions of our own bodies. Where do agency and directedness come from? A clue comes from considering the elementary spontaneous movements we are born with which consist of a large repertoire of spontaneous (thus self-organized) movements – making a fist, kicking, sucking, etc. etc. Only at some point does the child realize – through his own movements and the kinaesthetic sensations they give rise to – that these movements are his own. If one attaches the string of a mobile to his foot, he comes to realize that it is *his* kicking movements that are causing the mobile to move in ways that *he* likes. The pre-existing repertoire enables activities to happen before we make them happen. Evolutionarily constrained self-organizing coordination tendencies ('intrinsic dynamics') thus appear to lie at the origins of conscious agency. They are, in the words of the philosopher Maxine Sheets-Johnstone, "the mother of all cognition", presaging every conscious mind that ever said "I". From spontaneous self-organized behavior emerges the self – "I am" "I do" and from there a huge range of potentialities ('I can do'). "I-ness" arises from spontaneity, and it is this "I" that directs human action. As Sheets-Johnstone [182] cogently remarks, we literally discover ourselves in movement. In our spontaneity of movement, we discover arms that extend, mouths that open, knees that flex and so forth. We make sense of ourselves as living things.

Following these insights, consider briefly how Coordination Dynamics addresses the role of intentional information in bringing about behavioral change. How is the process of intentionally switching among patterns of coordination to be understood? According to the theory, the relative stability of the intrinsic patterns plays a role in determining how easily the system can switch in and out of coordination states. As defined previously, "intrinsic dynamics" expresses the fact that the system (which may include the brain) – prior to any specific input – already possesses a repertoire of behavioral patterns that are unique to each individual. Theoretically, intention parametrizes the intrinsic dynamics in two ways: (i) by destabilizing an ongoing pattern and stabilizing a target pattern [114,137,171]; and (ii) by stabilizing an intrinsically unstable pattern that under the current circumstances might otherwise become unstable and switch (see

Chap. 5 in [100], [114]). Measurement of switching time shows that intention both acts upon and is constrained by the intrinsic dynamics of coordination. First, the system switches far faster from less stable to more stable patterns (as measured by variability) than vice-versa. Second, the data show that it is possible to intentionally stabilize intrinsically unstable patterns under conditions in which they would normally switch. Both results are in excellent agreement with theory [114,171,173].

The neural basis of the interaction between intrinsic brain dynamic activity and *intentional* pattern selection and switching is just beginning to be studied and the results look very promising [39]. *Spontaneous* switching between patterns is known to be associated with increased activity in prefrontal, premotor and parietal regions [3,37,146], a network that is compatible with the stability dependent circuits described in Fig. 2 (see also Jantzen & Kelso, 2007). Increased activity reported in specific brain regions appears to reflect the loss of pattern stability that precedes spontaneous pattern switching. New results from our laboratory show that there is greater activity in the basal ganglia (BG) – a region known to be crucial for starting and controlling voluntary movements [68] – when moving from a more to a less stable pattern [39]. The heightened level of activity in BG may be related to the stability of the original pattern, the stabilization of the selected pattern switched into or both. Regardless, this intriguing result suggests that the basal ganglia play a key role in parametrizing the coordination dynamics.

Stimulus (Parametric) Stabilization and Change

Not only internally generated information is able to stabilize and destabilize coordination states under suitable circumstances: coupling sound, vision or touch conditions to specific aspects of an individual movement have been shown, not only to modify the movement but to *globally* stabilize coordination [25,50,122,133]. Thus the role of 'stimuli' in Coordination Dynamics is much more than to trigger preset motor commands or provide feedback to the motor system. To account for these kinds of effects, Jirsa and colleagues [89] introduced the notion of *parametric stabilization*: coupling specific sensory input parametrically to a set of limit cycle oscillators (see also [4,91]):

$$\begin{aligned} \ddot{x}_1 + f(x_1, \dot{x}_1)\dot{x}_1 + \omega^2 x_1 &= g(x_1, \dot{x}_1, x_2, \dot{x}_2) + \varepsilon(t)x_1 \\ \ddot{x}_2 + f(x_2, \dot{x}_2)\dot{x}_2 + \omega^2 x_2 &= g(x_2, \dot{x}_2, x_1, \dot{x}_1) + \varepsilon(t)x_2 \end{aligned} \tag{14}$$

where f is a nonlinear oscillator function, g represents the HKB coupling (Eq. (4)), ω is the eigenfrequency of the oscillator and $\varepsilon(t)$ represents sensory information. Here again in Eq. (14) we see a key aspect of coordination dynamics, namely that perception and action, sensory information and the dynamics of movement are inextricably linked. Notice the linkage in this case is of a parametric, multiplicative nature which is necessary to account for both the local changes to component trajectories produced by sensory information (called 'anchoring') and the global stabilization effects on the coordination dynamics. Fink et al. [50] for example, were able to show that such localized and specific sensory information was capable of shifting (and thereby delaying) the critical point at which phase transitions occurred.

A Brief Survey of Applications and Elaborations of Coordination Dynamics

The foregoing discussion pertains to just two of the many kinds of adaptive modification of coordination dynamics that have been investigated in the literature. Here only a flavor can be provided. The sample includes, but is by no means limited to: the processes underlying the ability of biological systems to stabilize intrinsically unstable systems [26,54,199]; the initiation (including 'false starts') and coordination of discrete, discontinuous behaviors [52,80,86,112,188,189] including neurally-based comparisons with those of a continuous, rhythmic nature [166,185]; the spontaneous recruitment and annihilation of biomechanical degrees of freedom to accomplish task and environmental conditions [21,23,51,118]; the coordination dynamics of trajectory formation [22,38] and cursive handwriting [6]; the important role that perception [145] and attention [2,27,147,190,193,194] play in modulating coordinative stability; how practice and learning alter the entire coordination repertoire by reshaping the landscape of the coordination dynamics using competitive and cooperative mechanisms [53,140,151,177,222,223]; the stabilization and consolidation of new memorized states of coordination and the dynamics of the forgetting process [128,129]; how handedness amplifies asymmetries in the coordination dynamics [2,201], and so forth. The same concepts and methods have been applied to problems ranging from maintaining posture and stabilizing postural sway [8,9,42,83] to understanding how concurrent cognitive tasks modulate coordination dynamics [159,183].

Theoretical research at the neural level has progressed from phenomenological modeling at behavioral (e. g. [60,76,89,115,118,175,201]) and brain levels [88,205] to neurobiologically-grounded accounts of both unimanual [56,63,125] and bimanual coordination [87] that are

based on known cellular and neural ensemble properties of the cerebral cortex. Recent work [85] has extended this neural theory to include the heterogeneous connectivity between neural ensembles in the cortex. Once general laws of coordination at behavioral and brain levels have been identified, it has proved possible to derive them from a deeper theory founded on neuroanatomical and neurophysiological facts, thereby causally connecting different levels of description [117] for review). The neural theory, in turn, poses a number of challenges to experiment, such as how synaptic and cellular properties are influenced by learning, arousal and attention [103].

Remarkable applications of coordination dynamics have occurred in expected directions (though none the less remarkable for all that) including many physical activities and sports such as the relation between respiration and locomotion [36], juggling [13,79], gymnastics [142], running [41], tennis [157], swimming [179], boxing [135], skiing [154], golf [125] and even riding horses [134] to name only a few, as well as in entirely unexpected directions, such as modeling coordination of infant breathing as a way to understand the effects of premature birth [65], studies of coordination dynamics in children with Developmental Coordination Disorder [213] and the introduction of coordination dynamics therapy to treat a wide variety of CNS disorders and diseases (e. g. [167,168,169]; see also [207]). Principles of coordination dynamics have been shown to apply to perceptual grouping as nicely illustrated by the classic bistable properties of reversible figures such as the Necker cube (e. g. [100]), pattern recognition [77], the visual perception of spatiotemporal inphase and antiphase moving stimuli [18,40,78,221] and speech categorization [28,202]. In many cases the foregoing research findings have expanded, if not overturned, conventional explanations of phenomena that have seldom considered dynamics.

Increase in research activity using the concepts and methods of coordination dynamics has been such that the term has taken on a life of its own in different fields. Thus, it is commonplace in the literature to hear the words 'cognitive', 'brain', 'neural', 'social', 'behavioral', 'developmental' 'multimodal', 'postural', etc., qualify and precede the words *coordination dynamics*. The dynamical approach is currently center stage in a number of fields, for example, dynamical neuroscience (e. g., [90,160]), dynamical cognitive science (e. g., [16,162,186,209]), behavioral [217] and task [165] dynamics, dynamical social psychology (e. g., [15,155,206]), dynamical systems accounts of development (e. g., [135,149,191,192,204]; see also [195]) and its implications are under careful consideration in philosophical circles (e. g. [11,29,30,31,198]).

Future Directions and Conclusions: The Complementary Nature of Coordination Dynamics

General laws and principles of biological coordination – to the extent they exist – are, by definition, abstract and mathematical. Yet, these laws are always conditioned by and realized by specific mechanisms and contexts. Over the last twenty-five years, often using the field of animate movement as an entry point it has been shown that the same coordination dynamics applies to functional coordination in a wide variety of situations. Although the basic laws for a quantitative description of the phenomena observed when human beings (and human brains) move, interact with the environment and with each other are the same, the anatomical, mechanical and physiological mechanisms realizing these dynamics are obviously not. Laws and mechanisms are complementary aspects of coordination dynamics.

Current research and theory views coordination as arising from the mutual interplay of constraints on multiple levels of description – ranging from the intrinsic properties and modes of interaction among cells and cellular ensembles in brain circuitry to biomechanical influences at the behavioral level all the way to cognitive and task constraints. Coordination dynamics is not only a theoretical framework, but also a research program that explicitly attempts to incorporate and connect known constraints at multiple levels of description. For instance, coordination dynamics successfully identified and later quantified the form of the nonlinear coupling among interacting components. In showing that the stability and change of coordination is due to nonlinear interactions among individual components coordination dynamics removes some of the mysticism behind the contemporary terms "emergence" and "self-organization". At the same time, coordination dynamics expands and modifies the concept of self-organization in non-living systems by introducing new concepts to account for the fact that coordination is not only characterized by self-organization but also by directed or supervised forms of coordination. The two origins or cornerstones of coordination dynamics may be reconciled by showing how meaningful information originates from self-organizing processes and may in turn modify them.

In studies of coordinated movement, the field to which coordination dynamics owes its origins, it has proven useful to try to isolate the role of various constraints and how they are mediated by the central nervous system. On the one hand, this strategy has helped identify different factors that serve to stabilize coordination under conditions in which it may otherwise become unstable and susceptible to change. On the other hand, a focus on iso-

lating particular constraints can lead, albeit unwittingly, to dichotomies (e. g., coordination principles versus neuromuscular-skeletal mechanisms of implementation) that may not be so useful. In reality it seems safe to conclude that a coalition of constraints – acting on multiple levels – impinges upon the stability of coordination depending on task and environmental context and the mover's intent. For example, the multilevel theory offered by Kelso [100] connects task goals (Level 1) to constraints on nonlinear oscillators (Level 3), the interactions among which determine the coordinative patterns observed (Level 2). Thus, rather than pose "abstract laws of coordination dynamics" against "neuromuscular-skeletal determinants of coordination", more important is to understand how the balance between identified constraints plays itself out in the course of any coordinated activity. Situations in which constraints are placed in competition with each other often prove to be highly revealing [122].

Throughout this article, every effort has been made to articulate the key notions of coordination dynamics, both conceptual and technical, and to present them in close proximity in order to help both the novice and the expert reader. The behavioral simplicity of the basic coordination patterns studied in the laboratory is deceptive; their understanding, however, requires recent advances in physics and mathematics. The theoretical concepts and methods of coordination dynamics are likely to play an ever greater role in the social, behavioral, economic, cognitive and neurosciences, especially as the interactions among disciplines continues to grow. Up to now, the use of nonlinear dynamics is still quite restricted, and often metaphorical. One reason is that the tools are difficult to learn, and require a degree of mathematical sophistication. Their implementation in real systems is nontrivial, requiring a different approach to experimentation and observation. Another reason is that the dynamical perspective is often cast in opposition to more conventional theoretical approaches, instead of as an aid or complement to understanding. The former tends to emphasize decentralization, collective decision-making and cooperative behavior among many interacting elements. The latter tends to focus on individual psychological processes such as intention, perception, attention, memory and so forth. Yet there is increasing evidence that intending, perceiving, attending, deciding, emoting and remembering have a dynamics as well. The language of dynamics serves to bridge individual and group processes. In each case, dynamics must be filled with content, with key variables and parameters obtained for the systems under study. A beauty about coordination dynamics is that the coordination variables or order parameters are semantic, relational quantities that "enfold" different aspects together thereby reducing often arbitrary divisions. Every system is different, but what we learn about one may aid in understanding another. What may be most important of all is to see animated living things in the light of a theory – coordination dynamics – that embraces both spontaneous self-organizing and directed processes, the complementary nature.

Acknowledgments

The writing of this article was supported by NIMH Innovations Award MH 42900, NINDS Grant NS48220, the US Office of Naval Research and the Pierre de Fermet chair of the Midi-Pyréncés Région for which the author is grateful.

Bibliography

Primary Literature

1. Abbott LF (2006) Where are the switches on this thing? In: van Hemmen JL, Sejnowski TJ (eds) 23 Problems in Systems Neuroscience. Oxford University Press, Oxford, pp 423–431
2. Amazeen EL, Amazeen PG, Treffner PJ, Turvey MT (1997) Attention and handedness in bimanual coordination dynamics. J Exp Psychol Hum Perc Perf 23:1552–1560
3. Aramaki Y, Honda M, Okada T, Sadato N (2006) Neural correlates of the spontaneous phase transition during bimanual coordination. Cereb Cortex 16:1338–1348
4. Assisi CG, Jirsa VK, Kelso JAS (2005) Dynamics of multifrequency coordination using parametric driving: Theory and Experiment. Biol Cybern 93:6–21
5. Assisi CG, Jirsa VK, Kelso JAS (2005) Synchrony and clustering in heterogeneous networks with global coupling and parameter dispersion. Phys Rev Lett 94:018106
6. Athènes S, Sallagoïty I, Zanone PG, Albaret JM (2004) Evaluating the coordination dynamics of handwriting. Hum Mov Sci 23:621–641
7. Banerjee A (2007) Neural information processing underlying rhythmic bimanual coordination: theory, method and experiment. Ph D Thesis, Complex Systems and Brain Sciences, Florida Atlantic University
8. Bardy BG (2004) Postural coordination dynamics in standing humans. In: Jirsa VK, Kelso JAS (eds) Coordination Dynamics: Issues and Trends. Springer, Berlin
9. Bardy BG, Oullier O, Bootsma RJ, Stoffregen TA (2002) The dynamics of human postural transitions. J Exp Psychol Hum Perc Perf 28:499–514
10. Başar E (2004) Memory and Brain Dynamics: Oscillations Integrating Attention, Perception, Learning, and Memory. CRC Press, Boca Raton
11. Bechtel W (1998) Representations and cognitive explanations: Assessing the dynamicist challenge in cognitive science. Cogn Sci 22:295–318
12. Beek PJ, Beek WJ (1988) Tools for constructing dynamical models of rhythmic movement. Hum Mov Sci 7:301–342
13. Beek PJ, Turvey MT (1992) Temporal Patterning in Cascade Juggling. J Exp Psychol Hum Perc Perf 18:934–947

14. Beek PJ, Rikkert WEI, van Wieringen PCW (1996) Limit cycle properties of rhythmic forearm movements. J Exp Psychol Hum Perc Perf 22:1077–1093
15. Beek PJ, Verschoor F, Kelso JAS (1997) Requirements for the emergence of a dynamical social psychology. Psychol Inq 8:100–104
16. Beer RD (1999) Dynamical approaches to cognitive science. Trends Cogn Sci 4:91–99
17. Bernstein N (1967) The coordination and regulation of movement. Pergammon, Oxford
18. Bingham GP, Schmidt RC, Zaal FT (1999) Visual perception of the relative phasing of human limb movements. Perc Psychophys 61:246–258
19. Bressler SL (2003) Cortical coordination dynamics and the disorganization syndrome I schizophrenia. Neuropsychopharmacology 28:S35-S39
20. Bressler SL, Kelso JAS (2001) Cortical coordination dynamics and cognition. Trends Cogn Sci 5:26–36
21. Buchanan JJ, Kelso JAS (1999) To switch or not to switch: Recruitment of degrees of freedom stabilizes biological coordination. J Motor Behav 31:126–144
22. Buchanan JJ, Kelso JAS, de Guzman GC (1997) The self-organization of trajectory formation: I Experimental evidence. Biol Cybern 76:257–273
23. Buchanan JJ, Kelso JAS, de Guzman GC, Ding M (1997) The spontaneous recruitment and annihilation of degrees of freedom in rhythmic hand movements. Hum Mov Sci 16:1–32
24. Buzsáki G (2006) Rhythms of the Brain. Oxford University Press, Oxford
25. Byblow WD, Carson RG, Goodman D (1994) Expressions of asymmetries and anchoring in bimanual coordination. Hum Mov Sci 13:3–28
26. Cabrera JL, Milton JG (2004) Stick balancing: On-off intermittency and survival times. Nonlinear Sci 11:305–317
27. Carson RG, Chua R, Byblow WD, Poon P, Smethurst CS (1999) Changes in posture alter the attentional demands of voluntary movement. Proc R Soc Lond B 266:853–857
28. Case P, Tuller B, Ding M, Kelso JAS (1995) Evaluation of a dynamical model of speech perception. Perc Psychophys 57:977–988
29. Chemero A (2001) Dynamical explanation and mental representation. Trends Cogn Sci 5:140–141
30. Chemero A, Silberstein M (2008) After the philosophy of mind. Philos Sci (in press)
31. Clark A (1997) Being there. MIT Press, Cambridge
32. Collins JJ, Stewart IN (1993) Coupled nonlinear oscillators and the symmetries of animal gaits. J Nonlinear Sci 3:349–392
33. Crick FHC (1966) Of molecules and men. University of Washington Press, Seattle
34. Crick FHC (1994) The astonishing hypothesis. Scribner, New York
35. Daffertshofer A, Peper CE, Beek PJ (2000) Spectral analysis of event-related encephalographic signals. Phys Lett A 266:290–302
36. Daffertshofer A, Huys R, Beek PJ (2004) Dynamical coupling between locomotion and respiration. Biol Cybern 90: 157–164
37. Debaere F, Swinnen SP, Beatse E, Sunaert S, van HP, Duysens J (2001) Brain areas involved in interlimb coordination: a distributed network. Neuroimage 14:947–958
38. DeGuzman GC, Kelso JAS, Buchanan JJ (1997) The self-organization of trajectory formation: II Theoretical model. Biol Cybern 76:275–284
39. DeLuca C, Jantzen KJ, Bertollo M, Comani S, Kelso JAS (2008) The role of basal ganglia in the intentional switching between coordination patterns of different stability. Paper presented at 18th Annual Meeting of the Neural Control of Movement, Naples, Florida 29 April-4 May 2008
40. De Rugy RA, Oullier O, Temprado JJ (2008) Stability of rhythmic visuo-motor tracking does not depend on relative velocity. Exp Brain Res 184:269–273
41. Diedrich FJ, Warren WH Jr (1995) Why change gaits? Dynamics of the walk-run transition. J Exp Psychol Hum Perc Perf 21:183–202
42. Dijkstra TMH, Schoner G, Giese MA, Gielen CCAM (1994) Frequency dependence of the action-perception cycle for postural control in a moving visual environment: Relative phase dynamics. Biol Cybern 71:489–501
43. Eckhorn R, Bauer R, Jordan W, Borsch M, Kruse W, Munk M, Reitboeck HJ (1988) Coherent oscillations: a mechanism of feature linking in the visual cortex. Multiple electrode correlation analyses in the cat. Biol Cybern 60:121–130
44. Edelman GM (2004) Naturalizing consciousness: a theoretical framework. Proc Nat Acad Sci USA 100:520–524
45. Edelman G, Tononi G (2000) A Universe of Consciousness. Basic Books, New York
46. Edelman GE, Gally JA (2001) Degeneracy and complexity in biological systems. Proc Nat Acad Sci USA 98:13763–13768
47. Eisenhammer T, Hubler A, Packard N, Kelso JAS (1991) Modeling experimental time series with ordinary differential equations. Biol Cybern 65(2):107–112
48. Fearing F (1930/1970) Reflex Action: a study in the history of physiological psychology. MIT Press, Cambridge
49. Fingelkurts AnA, Fingelkurts AlA (2004) Making complexity simpler: multivariability and metastability in the brain. Int J Neurosci 114:843–862
50. Fink PW, Kelso JAS, Jirsa VK, Foo P (2000) Local and global stabilization of coordination by sensory information. Exp Brain Res 134:9–20
51. Fink PW, Kelso JAS, DeGuzman GC (2000) Recruitment of degrees of freedom stabilizes coordination. J Exp Psychol Hum Perc Perf 26:671–692
52. Fink PW, Kelso JAS, Jirsa VK (in press) Perturbation-induced false starts as a test of the Jirsa–Kelso Excitator Model. J Motor Behav (in press)
53. Fontaine RB, Lee TD, Swinnen SP (1997) Learning a new bimanual coordination pattern: Reciprocal influences of intrinsic and to-be-learned patterns. Can J Exp Psychol 51(1):1–9
54. Foo P, Kelso JAS, deGuzman GC (2000) Functional stabilization of unstable fixed points: Human pole balancing using time-to-balance information. J Exp Psychol Hum Perc Perf 26:1281–1297
55. Fox Keller E (2007) A clash of two cultures. Nature 445:603
56. Frank TD, Daffertshofer A, Peper CE, Beek PJ, Haken H (2000) Towards a comprehensive theory of brain activity: Coupled oscillator systems under external forces. Physica D 144:62–86
57. Freeman WJ, Holmes MD (2005) Metastability, instability, and state transition in neocortex. Neural Netw 18:497–504
58. Friston KJ (1997) Transients, metastability, and neuronal dynamics. Neuroimage 5:164–171
59. Fuchs A, Jirsa VK (eds) (2008) Coordination: Neural, behavioral and social dynamics. Springer, Heidelberg

60. Fuchs A, Kelso JAS (1994) A theoretical note on models of interlimb coordination. J Exp Psychol Hum Perc Perf 20:1088–1097
61. Fuchs A, Kelso JAS, Haken H (1992) Phase transitions in the human brain: Spatial mode dynamics. Int J Bifurc Chaos 2:917–939
62. Fuchs A, Deecke L, Kelso JAS (2000) Phase transitions in human brain revealed by large SQuID arrays: Response to Daffertshofer, Peper and Beek. Phys Lett A 266:303–308
63. Fuchs A, Jirsa VK, Kelso JAS (2000) Theory of the relation between human brain activity (MEG) and hand movements. NeuroImage 11:359–369
64. Gleick J (2003) Isaac Newton. Pantheon, New York
65. Goldfield EC, Schmidt RC, Fitzpatrick P (1999) Coordination dynamics of abdomen and chest during infant breathing: A comparison of full-term and preterm infants at 38 weeks postconceptional age. Ecol Psychol 11:209–233
66. Golubitsky M, Stewart I, Buono P-L, Collins JJ (1999) Symmetry in locomotor central pattern generators and animal gaits. Nature 401:693–695
67. Gray CM, König P, Engel AK, Singer W (1989) Oscillatory responses in cat visual cortex exhibit inter-columnar synchroniation which reflects global stimulus properties. Nature 338:334–337
68. Graybiel AM, Saka E (2003) The basal ganglia and the control of action. In: Gazzaniga MS (ed) The New Cognitive Neurosciences, 3rd edn. MIT Press, Cambridge, pp 495–510
69. Greene HS (2000) Information Theory and Quantum Physics. Springer, Berlin
70. Grillner S (1985) Neurobiological bases of rhythmic motor acts in vertebrates. Science 228:143–149
71. Grillner S (2003) The motor infrastructure: from ion channels to neuronal networks. Nat Rev Neurosci 4:573–586
72. Grillner S (2008) Personal communication
73. Haken H (1977/1983) Synergetics: Nonequilibrium phase transitions and self-organization in physics, chemistry and biology. Springer, Berlin
74. Haken H (1988) Information and self-organization. Springer, Berlin
75. Haken H (1996) Principles of Brain Functioning. Springer, Berlin
76. Haken H, Kelso JAS, Bunz H (1985) A theoretical model of phase transitions in human hand movements. Biol Cybern 51:347–356
77. Haken H, Kelso JAS, Fuchs A, Pandya A (1990) Dynamic pattern recognition of coordinated biological motion. Neural Netw 3:395–401
78. Hock HS, Kelso JAS, Schöner G (1993) Bistability, hysteresis, and phase transitions in the perceptual organization of apparent motion. J Exp Psychol Hum Perc Perf 19:63–80
79. Huys R, Daffertshofer A, Beek PJ (2004) Multiple time scales and subsystem embedding in the learning of juggling. Hum Mov Sci 23:315–336
80. Huys R, Studenka BE, Rheaume NL, Zelaznik HN, Jirsa VK (in press) Distinct timing mechanisms produce discrete and continuous movements. Public Library of Science (in press)
81. Jantzen KJ, Kelso JAS (2007) Neural coordination dynamics of human sensorimotor behavior: A Review. In: Jirsa VK, McIntosh R (eds) Handbook of Brain Connectivity. Springer, Heidelberg, pp 421–461
82. Jensen O, Colgin LL (2007) Cross-frequency coupling between neuronal oscillations. Trends Cogn Sci 11:267–269
83. Jeka JJ, Schoner G, Dijkstra TMH, Ribeiro P, Lackner JR (1997) Coupling of fingertip somatosensory information to head and body sway. Exp Brain Res 113:475–483
84. Jirsa VK, Haken H (1997) A derivation of a macroscopic field theory of the brain from the quasi-microscopic neural dynamics. Physica D 99:503–526
85. Jirsa VK, Kelso JAS (2000) Spatiotemporal pattern formation in neural systems with heterogeneous connection topologies. Phys Rev E 62:8462–8465
86. Jirsa VK, Kelso JAS (2005) The excitator as a minimal model for the coordination dynamics of discrete and rhythmic movements. J Motor Behav 37:35–51
87. Jirsa VK, Fuchs A, Kelso JAS (1998) Connecting cortical and behavioral dynamics: bimanual coordination. Neural Comput 10:2019–2045
88. Jirsa VK, Friedrich R, Haken H, Kelso JAS (1994) A theoretical model of phase transitions in the human brain. Biol Cybern 71:27–35
89. Jirsa VK, Fink PW, Foo P, Kelso JAS (2000) Parametric stabilization of biological coordination: A theoretical model. J Biol Phys 26:85–112
90. Jirsa VK, McIntosch AR (eds) (2007) Handbook of brain connectivity. Springer, Heidelberg
91. Kay BA, Warren WH Jr (2001) Coupling of posture and gait: mode locking and parametric excitation. Biol Cybern 85:89–106
92. Kay BA, Kelso JAS, Saltzman EL, Schöner G (1987) The space-time behavior of single and bimanual rhythmical movements: Data and a limit cycle model. J Exp Psychol Hum Perc Perf 13:178–192
93. Kay BA, Saltzman EL, Kelso JAS (1991) Steady-state and perturbed rhythmical movements: Dynamical modeling using a variety of analytic tools. J Exp Psychol Hum Perc Perf 17:183–197
94. Kelso JAS (1981) On the oscillatory basis of movement. Bull Psychon Soc 18:63
95. Kelso JAS (1984) Phase transitions and critical behavior in human bimanual coordination. Am J Physiol Regul Integr Comp 15:R1000-R1004
96. Kelso JAS (1990) Phase transitions: Foundations of behavior. In: Haken H, Stadler M (eds) Synergetics of cognition. Springer, Berlin, pp 249–268
97. Kelso JAS (1991) Behavioral and neural pattern generation: The concept of Neurobehavioral Dynamical System (NBDS). In: Koepchen HP, Huopaniemi T (eds) Cardiorespiratory and Motor Coordination. Springer, Berlin
98. Kelso JAS (1994) Elementary coordination dynamics. In: Swinnen S, Heuer H, Massion J, Casaer P (eds) Interlimb Coordination: Neural Dynamical and Cognitive Constraints, pp 301–318. Academic Press, San Diego
99. Kelso JAS (1994) The informational character of self-organized coordination dynamics. Hum Mov Sci 13:393–413
100. Kelso JAS (1995) Dynamic Patterns: The Self-organization of Brain and Behavior. MIT Press, Cambridge. [Paperback edition 1997:4th Printing]
101. Kelso JAS (2000) Fluctuations in the coordination dynamics of brain and behavior. In: Arhem P, Blomberg C, Liljenstrom H (eds) Disorder versus order in brain function: Essays in Theoretical Biology. World Scientific, Singapore

102. Kelso JAS (2000) Principles of dynamic pattern formation and change for a science of human behavior. In: Bergman LR, Cairns RB, Nilsson L-G, Nystedt L (eds) Developmental science and the holistic approach. Erlbaum, Mahwah, pp 63–83
103. Kelso JAS (2000) The self-organized dynamics of human skill learning. Dynamical Neuroscience VIII, New Orleans (available from NIMH, Bethesda, Maryland)
104. Kelso JAS (2002) The complementary nature of coordination dynamics: Self-organization and the origins of agency. J Nonlinear Phenom Complex Syst 5:364–371
105. Kelso JAS (2007) Synergies. Scholarpedia (Computational Neuroscience/Dynamical Systems)
106. Kelso JAS (2007) The Haken–Kelso–Bunz Model. Scholarpedia (Computational Neuroscience/Dynamical Systems)
107. Kelso JAS (2008) Synergies: Atoms of brain and behavior. In: Sternad D (ed) A multidisciplinary approach to motor control. Springer, Heidelberg
108. Kelso JAS, Fuchs A (1995) Self-organizing dynamics of the human brain: Critical instabilities and Sil'nikov chaos. Chaos 5(1):64–69
109. Kelso JAS, Haken H (1995) New laws to be expected in the organism: Synergetics of brain and behavior. In: Murphy M, O'Neill L (eds) What is Life? The Next 50 Years. Cambridge University Press, Cambridge
110. Kelso JAS, Engstrøm DA (2006) The Complementary Nature. MIT Press, Cambridge
111. Kelso JAS, Tognoli E (2007) Toward a complementary neuroscience: Metastable coordination dynamics of the brain. In: Kozma R, Perlovsky L (eds) Neurodynamics of Higher-level Cognition and Consciousness. Springer, Heidelberg, pp 39–60
112. Kelso JAS, Southard D, Goodman D (1979) On the nature of human interlimb coordination. Science 203:1029–1031
113. Kelso JAS, Scholz JP, Schöner G (1986) Nonequilibrium phase transitions in coordinated biological motion: Critical fluctuations. Phys Lett A 118:279–284
114. Kelso JAS, Scholz JP, Schöner G (1988) Dynamics governs switching among patterns of coordination in biological movement. Phys Lett A 134:8–12
115. Kelso JAS, DelColle JD, Schöner G (1990) Action-Perception as a pattern formation process. In: Jeannerod M (ed) Attention and Performance XIII. Erlbaum, Hillsdale, pp 139–169
116. Kelso JAS, Buchanan JJ, Wallace SA (1991) Order parameters for the neural organization of single, multijoint limb movement patterns. Exp Brain Res 85:432–444
117. Kelso JAS, Fuchs A, Jirsa VK (1999) Traversing scales of brain and behavioral organization. I-III. In: Uhl C (ed) Analysis of neurophysiological brain functioning. Springer, Berlin, pp 73–125
118. Kelso JAS, Buchanan JJ, DeGuzman GC, Ding M (1993) Spontaneous recruitment and annihilation of degrees of freedom in biological coordination. Phys Lett A 179:364–368
119. Kelso JAS, Holt KG, Rubin P, Kugler PN (1981) Patterns of human interlimb coordination emerge from the properties of non-linear oscillatory processes: Theory and data. J Motor Behav 13:226–261
120. Kelso JAS, Tuller B, Bateson EV, Fowler CA (1984) Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. J Exp Psychol Hum Perc Perf 10:812–832
121. Kelso JAS, Schöner G, Scholz JP, Haken H (1987) Phase-locked modes, phase transitions and component oscillators in coordinated biological motion. Phys Scr 35:79–87
122. Kelso JAS, Fink P, DeLaplain CR, Carson RG (2001) Haptic information stabilizes and destabilizes coordination dynamics. Proc R Soc B 268:1207–1213
123. Kelso JAS, Bressler SL, Buchanan S, DeGuzman GC, Ding M, Fuchs A, Holroyd T (1991) Cooperative and critical phenomena in the human brain revealed by multiple SQUIDS. In: Duke D, Pritchard W (eds) Measuring Chaos in the Human Brain. World Scientific, New Jersey, pp 97–112
124. Kelso JAS, Bressler SL, Buchanan S, DeGuzman GC, Ding M, Fuchs A, Holroyd T (1992) A phase transition in human brain and behavior. Phys Lett A 169:134–144
125. Kelso JAS, Fuchs A, Holroyd T, Lancaster R, Cheyne D, Weinberg H (1998) Dynamic cortical activity in the human brain reveals motor equivalence. Nature 392:814–818
126. Knight CA (2004) Neuromotor issues in the learning and control of golf skill. Res Q Exerc Sport 75:9–15
127. Koch C (2005) The Quest for Consciousness. Roberts and Co, Englewood
128. Kostrubiec V, Zanone PG (2002) Memory dynamics: Distance between the new task and existing behavioral patterns affects learning and interference in bimanual coordination. Neurosci Lett 331:193–197
129. Kostrubiec V, Tallet J, Zanone P-G (2006) How a new behavioral pattern is stabilized with learning determines its persistence and flexibility in memory. Exp Brain Res 170:238–244
130. Kryukov VI (1991) An attention model based on principle of dominanta. In: Holden AV, Kryukov VI (eds) Neurocomputers and Attention. 1. Neurobiology, synchronization and chaos. Manchester University Press, Manchester, pp 319–352
131. Kuhn TS (1962) The Structure of Scientific Revolutions. University of Chicago Press, Chicago
132. Kuramoto Y (1984) Chemical oscillations, waves, and turbulences. Springer, Berlin
133. Lagarde J, Kelso JAS (2006) Binding of movement, sound and touch: Multimodal coordination dynamics. Exp Brain Res 173:673–688
134. Lagarde J, Peham C, Licka T, Kelso JAS (2005) Coordination dynamics of the horse-rider system. J Motor Behav 37:418–424
135. Lagarde J, Deguzman GC, Oullier O, Kelso JAS (2006) Interpersonal interactions during boxing: Data and model. J Sport Exerc Psychol 28:S108
136. Latash ML, Anson JG (2006) Synergies in health and disease: Relations to adaptive changes in motor coordination. Phys Ther 86:1151–1160
137. Lee TD, Blandin Y, Proteau L (1996) Effects of task instructions and oscillation frequency on bimanual coordination. Psychol Res 59:100–106
138. Lewis MD (2000) The promise of dynamic systems approaches for an integrated account of human development. Child Dev 71:36–43
139. Liese T, Cohen A (2007) Nonlinear oscillators at our fingertips. Am Math Mon 114:14–28
140. Magne C, Kelso JAS (2008) A dynamical framework for human skill learning. In: Guadagnoli M, Etnyre B (eds) Brain, Behavior and Movement. Elsevier, North Holland
141. Mainzer K (1994) Thinking in Complexity. Springer, Berlin
142. Marin L, Bardy BG, Bootsma RJ (1999) Level of gymnastic skill

as an intrinsic constraint on postural coordination. J Sports Sci 17:615–626
143. Mayville JM, Bressler SL, Fuchs A, Kelso JAS (1999) Spatiotemporal reorganization of electrical activity in the human brain associated with a timing transition in rhythmic auditory-motor coordination. Exp Brain Res 127:371–381
144. Mayville JM, Fuchs A, Ding M, Cheyne D, Deecke L, Kelso JAS (2001) Event-related changes in neuromagnetic activity associated with syncopation and synchronization timing tasks. Hum Brain Mapp 14:65–80
145. Mechsner F, Kerzel D, Knoblich G, Prinz W (2001) Perceptual basis of bimanual coordination. Nature 414:69–73
146. Meyer-Lindenberg A, Ziemann U, Hajak G, Cohen L, Berman KF (2002) Transitions between dynamical states of differing stability in the human brain. Proc Nat Acad Sci USA 99:10948–10953
147. Monno A, Chardenon A, Temprado JJ, Zanone PG, Laurent M (2000) Effects of attention on phase transitions between bimanual coordination patterns: A behavioral and cost analysis in humans. Neurosci Lett 283:93–96
148. Néda Z, Ravasz E, Vicsek T, Brechet Y, Barabasi AL (2000) Physics of the rhythmic applause. Phys Rev E 61:6987–6992
149. Néda Z, Ravasz E, Vicsek T, Brechet Y, Barabasi AL (2000) Self-organization in the concert hall: The dynamics of rhythmic applause. Nature 403
150. Newell KM, Molenaar PCM (1998) Applications of nonlinear dynamics to developmental process modelling. Erlbaum, Mahwah
151. Newell KM, Liu Y-T, Mayer-Kress G (2008) Landscapes beyond the HKB model. In: Fuchs A, Jirsa VK (eds) (2008) Coordination: Neural, behavioral and social dynamics. Springer, Heidelberg
152. Nicolis G, Prigogine I (1977) Self-organization in nonequilibrium systems. Wiley, New York
153. Nicolis G, Prigogine I (1993) Exploring complexity. WH Freeman, San Francisco
154. Nourrit D, Delignieres D, Caillou N, Deschamps T, Lauriot B (2003) On discontinuities in motor learning: a longitudinal study of complex skill acquisition on a ski-simulator. J Motor Behav 35:151–170
155. Nowak A, Vallacher RR (1998) Dynamical social psychology. Guilford Publications, New York
156. Oullier O, Jantzen KJ (2008) Neural indices of behavioral instability in coordination dynamics. In: Fuchs A, Jirsa VK (eds) Coordination: Neural, Behavioral and Social Dynamics. Springer, Heidelberg, pp 205–227
157. Palut Y, Zanone P-G (2005) A dynamical analysis of tennis: concepts and data. J Sports Sci 23:1021–1032
158. Palva S, Palva JM (2007) New vistas for α-frequency band oscillations. Trends Neurosci 30:150–158
159. Pellecchia GL, Shockley K, Turvey MT (2005) Concurrent cognitive task modulates coordination dynamics. Cogn Sci 29:531–557
160. Perlovsky L, Kozma R (eds) (2007) Neurodynamics of Higher-level Cognition and Consciousness. Springer, Heidelberg
161. Pikovsky A, Rosenblum M, Kurths J (2001) Synchronization: A universal concept in nonlinear science. Cambridge University Press, Cambridge
162. Port RF, van Gelder T (eds) (1995) Mind as motion: Explorations in the dynamics of cognition. MIT Press, Cambridge
163. Prigogine I, Stengers I (1984) Order out of chaos: man's new dialogue with nature. Bantam Books, London
164. Rosenbaum DA (2005) The Cinderella of psychology: The neglect of motor control in the science of mental life and behavior. Am Psychol 60:308–317
165. Saltzman EL, Kelso JAS (1987) Skilled actions: A task dynamic approach. Psychol Rev 94:84–106
166. Schaal S, Sternad D, Osu H, Kawato M (2004) Rhythmic arm movement is not discrete. Nat Neurosci 7:1136–1143
167. Schalow G (2002) Recovery from spinal cord injury achieved by 3 months of coordination dynamic therapy. Electromyogr Clin Neurophysiol 42:367–376
168. Schalow G (2005) Phase and frequency coordination between neuron firing as an integrative mechanism of human CNS self-organization. Electromyogr Clin Neurophysiol 45:369–83
169. Schalow G, Jaiqma P (2005) Cerebral palsy improvement achieved by coordination dynamics therapy. Electromyogr Clin Neurophysiol 45:433–445
170. Schmidt RC, Carello C, Turvey MT (1990) Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. J Exp Psychol Hum Perc Perf 16:227–247
171. Scholz JP, Kelso JAS (1990) Intentional switching between patterns of bimanual coordination is dependent on the intrinsic dynamics of the patterns. J Motor Behav 22:98–124
172. Scholz JP, Kelso JAS, Schöner G (1987) Nonequilibrium phase transitions in coordinated biological motion: Critical slowing down and switching time. Phys Lett A 8:390–394
173. Schöner G, Kelso JAS (1988) A dynamic pattern theory of behavioral change. J Theor Biol 135:501–524
174. Schöner G, Kelso JAS (1988) Dynamic pattern generation in behavioral and neural systems. Science 239:1513–1520. Reprinted in: Kelner KL, Koshland DE Jr (eds) Molecules to Models: Advances in Neuroscience, pp 311–325
175. Schöner G, Haken H, Kelso JAS (1986) A stochastic theory of phase transitions in human hand movement. Biol Cybern 53:247–257
176. Schöner G, Jiang W-Y, Kelso JAS (1990) A synergetic theory of quadrupedal gaits and gait transitions. J Theor Biol 142:359–391
177. Schöner G, Zanone PG, Kelso JAS (1992) Learning as change of coordination dynamics: Theory and experiment. J Motor Behav 24:29–48
178. Schrödinger E (1944) What is Life? The physical aspect of the living cell. Camridge University Press, Cambridge
179. Seifert L, Chollet D, Bardy BG (2004) Effect of swimming velocity on arm coordination in the front crawl: a dynamic analysis. J Sports Sci 22:651–660
180. Shannon CE, Weaver W (1949) The mathematical theory of communication. University of Illinois Press, Chicago
181. Sheets-Johnstone M (1999) The primacy of movement. John Benjamin, Amsterdam
182. Sheets-Johnstone M (2004) Preserving integrity against colonization. Phenomenol Cogn Sci 3:249–261
183. Shockley K, Turvey MT (2006) Dual-task influences on retrieval from semantic memory and coordination dynamics. Psychon Bull Rev 13:985–990
184. Singer W (1999) Neural synchrony: a versatile code for the definition of relations. Neuron 24:49–65
185. Spencer RM, Zelaznik H, Diedrichson J, Ivry RB (2003) Disrupted timing of discontinous but not continuous movements by cerebellar lesions. Science 300:1437–1439

186. Spivey MJ (2007) The continuity of mind. Oxford University Press, New York
187. Sporns O (2004) Complex neural dynamics. In: Jirsa VK, Kelso JAS (eds) Coordination Dynamics: Issues and trends. Springer, Berlin, pp 197–215
188. Sternad D, Dean WJ (2005) Rhythmic and discrete elements in multijoint coordination. Brain Res 989:151–172
189. Sternad D, de Rugy A, Pataky T, Dean WJ (2002) Interaction of discrete and rhythmic movements over a wide range of periods. Exp Brain Res 147:162–174
190. Summers JJ, Maeder S, Hiraga CY, Alexander JRM (2008) Coordination dynamics and attentional costs of continuous and discontinuous bimanual circle drawing movements. Hum Mov Sci (in press)
191. Swinnen SP (2002) Intermanual coordination: From behavioural principles to neural-network interactions. Nat Rev Neurosci 3:350–361
192. Swinnen S, Heuer H, Massion J, Casaer P (eds) (1994) Interlimb Coordination: Neural, Dynamical and Cognitive Constraints. Academic Press, San Diego
193. Temprado JJ, Zanone PG, Monno A, Laurent M (1999) Attentional load associated with performing and stabilizing preferred bimanual patterns. J Exp Psychol Hum Perc Perf 25:1579–1594
194. Temprado JJ, Monno A, Zanone PG, Kelso JAS (2002) Attentional demands reflect learning-induced alterations of bimanual coordination dynamics. Eur J Neurosci 16:1–6
195. Thelen E, Smith LB (eds) (1994) Dynamic Approach to Development. MIT Press, Cambridge
196. Thelen E, Kelso JAS, Fogel A (1987) Self-organizing systems and infant motor development. Dev Rev 7:39–65
197. Thelen E, Skala KD, Kelso JAS (1987) The dynamic nature of early coordination: Evidence from bilateral leg movements in young infants. Dev Psychol 23:179–186
198. Thompson E (2007) Mind in life. Harvard, Cambridge
199. Tognoli E, Kelso JAS (2008) Brain coordination dynamics: true and false faces of phase synchrony and metastability. Ms under review and available from the authors
200. Treffner PJ, Kelso JAS (1999) Dynamic encounters: Long memory during functional stabilization. Ecol Psychol 11:103–137
201. Treffner PJ, Turvey MT (1996) Symmetry, broken symmetry and handedness in bimanual coordination dynamics. Exp Brain Res 107:463–478
202. Tuller B, Case P, Ding M, Kelso JAS (1994) The nonlinear dynamics of speech categorization. J Exp Psychol Hum Perc Perf 20:1–16
203. Turvey MT (2004) Impredicativity, dynamics and the perception-action divide. In: Jirsa VK, Kelso JAS (eds) (2004) Coordination Dynamics: Issues and Trends. Springer, Berlin
204. Turvey MT (2007) Action and perception at the level of synergies. Hum Mov Sci 26:657–697
205. Uhl C, Friedrich R, Haken H (1995) Analysis of spatiotemporal signals of complex systems. Phys Rev E 51:3890–3900
206. Vallacher RR, Nowak A (1997) The emergence of dynamical social psychology. Psychol Inq 8:73–99
207. van Den Berg C, Beek PJ, Wagenaar RC, Van Wieringen PC (2000) Coordination disorders in patients with Parkinson's disease. Exp Brain Res 134:174–186
208. van Geert P (1994) Dynamic systems of development: Change between complexity and chaos. Prentice Hall, New York
209. van Gelder TJ (1998) The dynamical hypothesis in cognitive science. Behav Brain Sci 21:1–14
210. van Mourik AM (2006) Structure from randomness. A window into coordination. Ph D Thesis, Free University of Amsterdam
211. van Mourik AM, Dafferthofer A, Beek PJ (2006) Determinsitic and stochastic features of rhythmic human movement. Biol Cybern 94:233–244
212. Varela FJ, Lachaux J-P, Rodriguez E, Martinerie J (2001) The brainweb: Phase synchronization and large-scale integration. Nat Rev Neurosci 2:229–239
213. Volman MJM (1997) Rhythmic coordination dynamics in children with and without a developmental coordination disorder. Ph D Thesis, University of Groningen
214. von der Malsburg C (1981) The correlation theory of brain function. Internal Report, pp 81–82. MPI Biophysical Chemistry, Goettingen
215. von Holst E (1938/73) The behavioral physiology of man and animals. In: Martin R (ed) The collected papers of Erich von Holst. University of Miami Press, Coral Gables
216. Wallenstein GV, Kelso JAS, Bressler SL (1995) Phase transitions in spatiotemporal patterns of brain activity and behavior. Physica D 84:626–634
217. Warren WH (2006) The dynamics of perception and action. Psychol Rev 113:358–389
218. Wimmers RH, Beek PJ, van Wieringen PCW (1992) Phase transitions in rhythmic tracking movements: a case of unilateral coupling. Hum Mov Sci 11:217–226
219. Winfree AT (2002) On emerging coherence. Science 298:2336–2337
220. Yuste R, MacLean JN, Smith J, Lansner A (2005) The cortex as a central pattern generator. Nat Rev Neurosci 6:477–483
221. Zaal FT, Bingham GP, Schmidt RC (2000) Visual perception of mean relative phase and phase variability. J Exp Psychol Hum Perc Perf 26:1209–1220
222. Zanone PG, Kelso JAS (1997) The coordination dynamics of learning and transfer: Collective and component levels. J Exp Psychol Hum Perc Perf 23:1454–1480
223. Zanone PG, Kelso JAS (1992) The evolution of behavioral attractors with learning: Nonequilibrium phase transitions. J Exp Psychol Hum Perc Perf 18/2:403–421

Books and Reviews

Jirsa VK, Kelso JAS (eds) (2004) Coordination Dynamics: Issues and Trends. Springer, Berlin

Kugler PN, Kelso JAS, Turvey MT (1980) Coordinative structures as dissipative structures I Theoretical lines of convergence. In: Stelmach GE, Requin J (eds) Tutorials in motor behavior. North Holland, Amsterdam

Murphy M, O'Neill L (eds) (1995) What is Life? The Next 50 Years. Cambridge University Press, Cambridge

Port RF, van Gelder T (eds) (1995) Mind as Motion: Explorations in the Dynamics of Cognition. MIT Press, Cambridge

Strogatz SH (1994) Nonlinear dynamics and chaos. Addison-Wesley, Reading

Tschacher W, Dauwalder JP (eds) (2003) The Dynamical Systems Approach to Cognition: Concepts and Empirical Paradigms Based on Self-organization, Embodiment and Coordination Dynamics. World Scientific, Singapore

Turvey MT (1990) Coordination. Am Psychol 45:938–953

Corporate and Municipal Bond Market Microstructure in the U.S.

MICHAEL S. PIWOWAR
Securities Litigation and Consulting Group, Inc., Fairfax, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Early Corporate and Municipal Bond Market Microstructure Research
Fixed Income Pricing Service (FIPS) Research
Municipal Bond Market Research
Transaction Reporting and Compliance Engine (TRACE) Research
The Links Between Bond Market Microstructure Research and Other Finance and Economics Research
Bibliography

Glossary

ABS Automated Bond System. The original automated limit-order market for bonds operated by the NYSE that executed orders according to strict price/time priority. ABS was replaced by the NYSE Bonds Platform in 2007.

Agency trade A bond transaction executed by a broker-dealer on behalf of another party. A broker-dealers is compensated by a commission on an agency trade.

Broker A firm that acts as an intermediary by executing agency trades.

Broker-dealer A firm that engages in both agency trades and principal trades.

Broker's broker A broker-dealer that exclusively executes agency trades of municipal bonds with other broker-dealers. Broker's brokers do not execute principal trades and they do not trade directly with public investors.

Commission A form of compensation that a customer pays a broker-dealer for executing an agency trade. Broker-dealers must explicitly disclose the commission to the customer as a separate item on the customer's trade confirmation.

Dealer A firm that engages in principal trades for its own account.

FINRA Financial Industry Regulatory Authority. The self-regulatory organization (SR0) created in July 2007 from the consolidation of NASD and the member regulation, enforcement and arbitration functions of the NYSE. FINRA rules are approved by the SEC and enforced by themselves.

FIPS Fixed Income Pricing Service. The electronic system operated by the National Association of Securities Dealers (NASD) from 1994 through 2002 to collect and disseminate real-time quotations and hourly trade reports for a subset of high-yield corporate bonds. FIPS was retired in July 2002 with the implementation of TRACE.

Market maker A specific designation made by a regulatory authority for a broker-dealer that holds itself out to trade securities by publishing regular or continuous quotations to buy (bid) or sell (offer). Currently, there are no broker-dealers regulated as market makers in the US corporate or municipal bond markets.

Mark-up and mark-down A form of compensation that a customer pays a broker-dealer for executing a principal trade. Customers pay a mark-up when they buy a bond from a broker-dealer; they pay a mark-down when they sell a bond to a broker-dealer. Unlike commissions, mark-ups and mark-downs do not need to be disclosed on customer trade confirmations.

MSRB Municipal Securities Rulemaking Board. The self-regulatory organization (SRO) charged with primary rulemaking authority over broker-dealers in connection with their municipal bond transactions. MSRB rules are approved by the SEC and enforced by FINRA (formerly NASD).

NASD Formerly known as the National Association of Securities Dealers. The self-regulatory organization (SRO) charged with, among other things, primary rulemaking authority over broker-dealers in connection with their corporate bond transactions. In July 2007, NASD and the member regulation, enforcement and arbitration functions of the NYSE consolidated to form FINRA.

NYSE New York Stock Exchange. Operates the NYSE Bonds Platform (formerly ABS) trading system for exchange-listed corporate bonds.

OTC securities Over the-counter securities. Securities that are not traded on an organized exchange.

Principal trade A bond transaction executed by a broker-dealer for its proprietary account. The broker-dealer is compensated by a mark-up or mark-down on a principal trade.

Riskless principal trade A principal trade in which a broker-dealer purchases a bond to satisfy a previously received order to buy, or a broker-dealer sells a bond to satisfy a previously received order to sell. The trans-

action is riskless to the broker-dealer because the firm does not bear any inventory (price) risk.

RTTRS (or TRS) (Real-Time) Transaction Reporting System. MSRB's municipal bond transaction reporting and dissemination system.

Serial offering A bond issuance in which several different bonds are offered with different, often consecutive, maturities. Municipal bonds are typically issued in serial offerings.

SRO Self-regulatory Organization. A non-governmental industry association that has statutory authority to regulate members through the promulgation and enforcement of rules and regulations governing business practices. The SEC oversees SRO activities and approves SRO rules.

SEC US Securities and Exchange Commission. The primary governmental overseer and regulator of US securities markets, including the corporate and municipal bond markets. Broker-dealers and SROs are overseen by the SEC's Division of Trading and Markets (formerly Division of Market Regulation).

TRACE (formerly NASD TRACE) Transaction Reporting and Compliance Engine. FINRA's corporate bond transaction reporting and dissemination system.

Definition of the Subject

The subject of this article is the microstructure of the US corporate and municipal bond markets. ► Treasury Market, Microstructure of the U.S. provide a complementary discussion of the microstructure of the US Treasury bond market.

Market microstructure is broadly defined as the study of the economics of markets and trading. Market microstructure research covers a wide range of interrelated topics including market structure and design issues (e. g., trading systems and rules); price formation and price discovery; strategic trading behavior; market quality, liquidity, and trading costs (explicit and implicit); information, disclosure, and transparency; and consequences of regulatory policy (intended and unintended).

While much has been written on the microstructure of equity markets since the mid-1980s, the bond markets have only recently started receiving attention from academic researchers. The development of research in both markets can largely be attributed to the availability of quality intraday trade, quote, and/or order data ("tick" data) to empirical researchers.

The seminal theoretical work in market microstructure was conducted contemporaneously with the early equity market microstructure research, and much of the underlying economics is general enough to be appropriate for the bond markets. As a result, the significant contributions of bond market research so far have been almost exclusively empirical in nature. The last study featured in this article by Green, Hollifield, and Schurhoff [23] is a notable exception.

Conversely, the empirical methods developed specifically for the structure and design of equity markets are not well-suited for the bond markets. Accordingly, many of the important contributions of bond market microstructure research stem from not only the results and conclusions, but also from the development of new empirical methods. This article will provide details on some of these methods as well as discuss the important results, conclusions, and policy implications.

But, before moving on to a detailed discussion of bond market microstructure research, an important question needs to be answered. Why should we care about the bond markets? We should care because the bond markets provide an important source of capital for issuers and an important source of securities for investors. In other words, the bond markets are large. How large are they? The answer to this question depends on one's definition of size.

Figure 1 shows that an astonishingly large number, approximately 1.5 million, corporate and municipal bonds are outstanding. The vast majority of these are municipal bonds, which are typically issued in serial offerings consisting of a set of up to 20 (or more) bonds issued at the same time with different maturities. Thus, the number of bonds dwarfs the number of equities.

In terms of total dollar amounts outstanding, Fig. 1 shows that US corporate and municipal bond markets combined are roughly half the size of the US equity markets. The average daily trading volume in these bond markets is about $36 billion, which is about 1/3 of the average daily trading volume of $115 billion in the equity markets. While the discussion of the microstructure of the Treasury bond markets is left to ► Treasury Market, Microstructure of the U.S., it is worth noting that total US bond market trading volume (corporate, municipal, and Treasury) exceeds US equity market trading volume. Thus, no matter what measure is used, it is apparent that the bond markets offer important sources of capital for issuers and securities for investors.

The remainder of this article proceeds as follows. Section "Introduction" provides a historical overview of the US corporate and municipal bond markets. Section "Early Corporate and Municipal Bond Market Microstructure Research" through "The Links Between Bond Market Microstructure Research and Other Finance and Economics Research" review the significant contributions to the bond

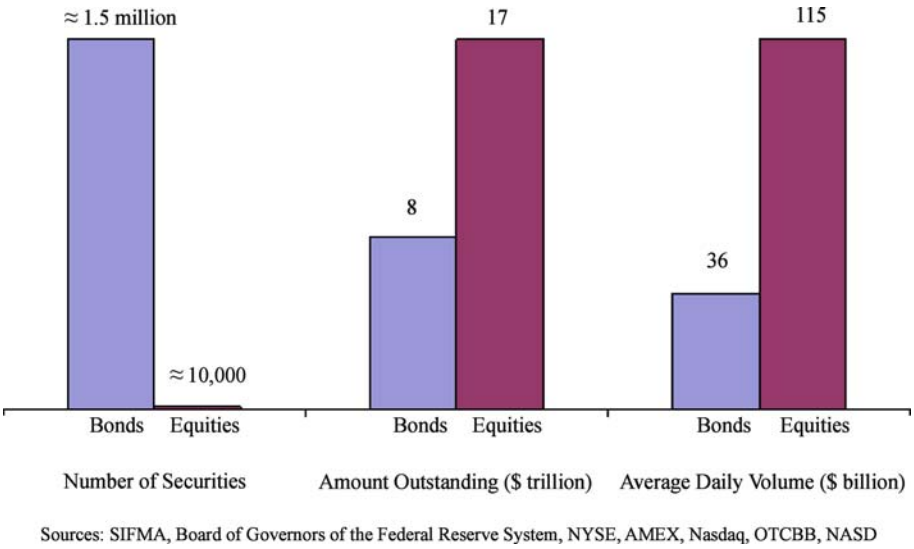

Corporate and Municipal Bond Market Microstructure in the U.S., Figure 1
Comparison US municipal corporate and bond markets with US equity markets

market microstructure literature. Section "Early Corporate and Municipal Bond Market Microstructure Research" reviews the early corporate and municipal bond market microstructure research. Section "Fixed Income Pricing Service (FIPS) Research" reviews the research enabled by the National Association of Securities Dealer's (NASD's) Fixed Income Pricing Service that began in May 1994. Section "Municipal Bond Market Research" reviews the municipal bond market research enabled by the Municipal Securities Rulemaking Board's (MSRB's) transaction data. Section "Transaction Reporting and Compliance Engine (TRACE) Research" reviews the research enabled by the NASD's Transaction Reporting and Compliance Engine (TRACE) system that began in July 2002. Section "The Links Between Bond Market Microstructure Research and Other Finance and Economics Research" provides examples of how bond market microstructure research is linked to other areas of finance and economics research.

Introduction

Today, virtually all US corporate and municipal bond trading occurs in over the-counter (OTC) dealer markets with transparent prices. But, that was not always the case. In the early 20th century there were active and transparent markets for both corporate bonds and municipal bonds on the New York Stock Exchange (NYSE). Then, bond trading began migrating to opaque OTC dealer markets. In the late 20th century, post-trade transparency was added to the both the corporate and municipal bond OTC markets.

What factors are responsible for the evolution of the bond markets over the past century? What caused the migration of trading in the early 20th century? How (and why) was post-trade transparency added to the bond markets in the late 20th century? The brief history of US corporate and municipal bond markets below provides answers to these questions.

The Early 20th Century

Biais and Green [9] provide a fascinating historical overview of the US corporate and municipal bond markets. Early 20th century NYSE bond trading took place among the "bond crowd". Bond trading originally took place in the same trading room as stock trading, with the bond crowd organizing around three trading booths in the "bond corner" of the Exchange. In 1928, the NYSE opened a separate trading room, the "bond room", in response to increases in trading volumes. Trading in the bond room was separated into four different crowds. US corporate and municipal bonds were traded in either the "active" crowd or the "inactive" crowd. The inactive crowd was also known as the "cabinet" crowd because bond orders were written on slips of paper and filed in the bond cabinets. Foreign bonds and Government securities each had their own bond crowds. A small number of active bonds were traded on the floor in an open outcry market.

NYSE bond trading was "order-driven". The exchange collected, posted, and matched public customer orders.

Public investors paid commissions to brokers to facilitate their NYSE bond trades. All NYSE bond brokers could observe the book of available orders and the recent trades, and inform their customers about them. Thus, NYSE bond trading enjoyed a very high level of "pre-trade transparency" and "post-trade transparency". Pre-trade transparency refers to the dissemination of information about trading interests. Pre-trade information can include price (bid and ask) and depth quotations, as well as limit order prices and sizes. Post-trade transparency refers to the dissemination of information about past trades. While post-trade information includes not only prices, such as trade execution times and volumes, post-trade transparency in the bond markets is sometimes referred to as simply "price transparency". Madhavan [38] and Harris [24] provide excellent discussions all the different dimensions of transparency as well as the related market microstructure literature.

In the late 1920s, municipal bond trading migrated to the over the-counter (OTC) market. Corporate bond trading migrated to the OTC market in the 1940s. Biais and Green [9] examine a number of potential explanations for the decline in municipal and corporate bond trading on the NYSE. They find that the decline of exchange trading in bonds was not due to a decline in the supply of bonds outstanding or a decline in listings in response to costly rules and regulations promulgated by the newly created SEC.

Biais and Green [9] find that the migration of bond trading from the NYSE to the OTC markets coincided with changes in the investor base. In the late 1920s, retail investor interest in municipal bonds waned, as they became more attracted to the higher returns on equities. As retail interest in municipal bonds waned, institutions became the dominant investor in the market. During the 1940s, a similar shift in the relative importance of retail investors and institutional investors occurred in the corporate bond market. Biais and Green [9] conclude that the migration of bond trading from the NYSE to the OTC markets was an evolution in response to the changing investor base.

Biais and Green [9] provide evidence that institutions fared better in OTC bond markets and argue that the dealers were happy to accommodate this new class of dominant investors. Because liquidity was no longer concentrated on a centralized transparent exchange, retail investors were effectively forced into trading with dealers in these decentralized opaque OTC markets. Not surprisingly, retail investors faredjt worse in these markets. Both municipal and corporate bond transaction costs increased significantly for retail investors.

The Late 20th Century and Early 21st Century

While the most significant change in the bond markets in the early 20th century was a migration of trading from the exchange to OTC, the most significant change in the late 20th century was the introduction of price transparency. Unlike trading migration, bond market transparency was not caused by market forces. Rather, transparency was added to the bond markets by direct regulatory intervention.

The Municipal Securities Rulemaking Board (MSRB) introduced price transparency to the municipal bond market. The MSRB was created by Congress in 1975 as the self-regulatory organization (SRO) charged with primary rulemaking authority over broker-dealers in connection with their municipal bond transactions.

The MSRB began publicly disseminating municipal bond price information in January 1995. "Interdealer Daily Reports" provided statistics on total interdealer market activity reported for the previous day, as well as information about price and volume for each security that was "frequently traded" on that day. The MSRB defined frequently traded securities to be securities with four or more interdealer transactions on a particular day. The Interdealer Daily Report included the total par value traded, the daily high and low price, and the average price of trades having a par value between $100,000 and $1 million for each frequently traded issue. Transaction price information on securities with three or fewer interdealer transactions on a particular day ("infrequently traded" securities) was not disseminated.

In August 1998, the MSRB began producing "Combined Daily Reports". The Combined Daily Reports merged information from customer and interdealer transactions and provided daily high, low, and average prices for frequently traded securities of municipal securities on a one-day delayed basis. The frequently traded threshold was four transactions per day, taking into account both customer and interdealer transactions.

In January 2000, the MSRB began publicly disseminating transaction details on individual trades in frequently traded securities through "Daily Transaction Reports". Trade information on infrequently traded securities was still not disseminated until October 2000, when the MSRB began producing "Monthly Comprehensive Reports". These reports provided information on a one-month delayed basis for all transactions from the previous month, including infrequently traded issues.

By June 2003, the MSRB was publicly disseminating transaction details for all trades in all securities (frequently traded and infrequently traded) on a one-day lag basis

through "T+1 Daily Reports". In January 2005, the MSRB began disseminating prices on a real-time basis through its Real-Time Transaction Reporting System (RTTRS or TRS).

The National Association of Securities Dealers (NASD) introduced price transparency to the corporate bond market. NASD (now FINRA) is the self-regulatory organization (SRO) charged with primary rulemaking authority over broker-dealers in connection with their corporate bond transactions. Regulatory concern for price transparency spiked in the late 1980s and early 1990s.

At that time, the high-yield corporate bond market faced unprecedented instability, highlighted by insider trading scandals and the ultimate collapse of the dominant dealer and underwriter Drexel Burnham Lambert. Concern over future market instability, along with the recognition of a need for better monitoring, led to regulatory intervention that provided a small degree of transparency in this market segment. The Fixed Income Pricing Service (FIPS) began in 1994. FIPS was the result of the SEC encouraging NASD to develop an electronic reporting and dissemination facility for non-convertible high-yield corporate bonds.

But, FIPS only provided partial transparency for this particular segment of the corporate bond market. While every trade in FIPS-eligible bonds was reported to FIPS, only summary information on a small subset (50 bonds) of the most active bonds was disseminated to the public. Alexander, Edwards, and Ferri [2] point out that some members of the SEC staff at that time feared that adding price transparency to less active bonds could possibly harm the market. FIPS added both pre-trade transparency and post-trade transparency to the market by disseminating quotations and hourly trade summaries, respectively. The hourly trade summaries contained high and low prices as well as total trading volume.

Many bond market participants and some SEC staff felt that FIPS added a sufficient amount of transparency to the corporate bond market. SEC Chairman Arthur Levitt disagreed. In 1998, he gave a speech entitled *The Importance of Transparency in America's Debt Market* in which he famously quipped "The sad truth is that investors in the corporate bond market do not enjoy the same access to information as a car buyer or a homebuyer or, dare I say, a fruit buyer". To address the lack of price transparency in the corporate bond market, he called on NASD to take several related actions. He called on NASD to adopt rules requiring dealers to report all corporate bond transactions; to develop a system to receive all corporate bond transaction information; to create a database of the transactions, and in conjunction, create a surveillance program to better detect fraud in corporate bonds; and, to disseminate the bond transaction prices to the public in order to help them make better investment decisions.

NASD responded by developing the Transaction Reporting and Compliance Engine (TRACE) system, which began operation in July 2002. Corporate bond dealers were required to report all transaction in TRACE-eligible securities. TRACE-eligible securities included investment grade and high-yield debt, convertible and non-convertible debt, and publicly issued debt and privately issued (Rule 144A) debt. While all TRACE-eligible transactions were reported to TRACE from the beginning of its operation, the dissemination of the trade information was phased-in over time. The phase-in approach was adopted by NASD, and approved by the SEC, because of industry concerns that adding transparency to the bond market would somehow harm liquidity.

The TRACE phase-in approach began with the dissemination of trade information on the largest, highest-rated bonds first. Price transparency was introduced to smaller and lower-rated bonds over time. By February 2005, prices were transparent on effectively 99% of trades, and by the end of that year, pricing information on all TRACE-eligible trades was being disseminated on a real-time basis to the public.

By the beginning of the 21st century, investors (and market microstructure researchers) were able to access an unprecedented amount of information about the OTC municipal and corporate bond markets from the post-trade transparency brought by TRS and TRACE, respectively. It is worth noting that bond trading never completely migrated to the OTC markets. The NYSE continues to list and trade some bonds. The NYSE developed the Automated Bond System (ABS), a computerized limit-order market for bonds, in an effort to encourage the migration of trading back to the exchange. The displayed public limit orders on ABS provided pre-trade transparency for some bonds.

However, the vast majority of bonds are not listed on the NYSE ABS or any other exchange, so all of the trading in these bonds occurs in the OTC markets. Moreover, for many of the bonds that are listed on the NYSE ABS, a majority of their trades still occur over the-counter. Therefore, the early 21st century bond markets can be characterized as dealer markets with a high degree of post-trade transparency, but with virtually no pre-trade transparency. It remains to be seen whether market forces, regulatory initiatives, or some combination of the two will eventually lead to the introduction of some form of pre-trade transparency, the emergence of bond market-makers, and/or

a migration of trading back to an order-driven market in the US corporate and municipal bond markets.

Early Corporate and Municipal Bond Market Microstructure Research

Early bond market microstructure researchers were forced to rely on data sources that covered only certain segments of the market because comprehensive bond market data simply did not exist. Bond dealers kept their own trading records because there was no central reporting facility. Bond dealers were not required to publicly disseminate their trades. Researchers, as well as investors and regulators, were able to see snapshots of parts of the bond markets, but no one was able to see the complete picture.

But, even with the data limitations, early bond market researchers found creative ways to tease out useful information. Their initial findings shed the first light on the opaque bond markets. For example, Schultz [42] provides indirect evidence that higher bond market trading costs may be attributable to the lack of transparency. Variants of their original empirical methods continue to be used by more recent researchers.

Any encyclopedic article on corporate bond market microstructure research would be incomplete if it did not mention the efforts of the Fixed Income Research Program (FIRP), and more importantly its founder Professor Arthur Warga, in promoting early bond market research. Art Warga's influence on the development of the bond market microstructure literature extends beyond his own research. He collected, consolidated, cleaned, and organized various fragmented sources of bond market data to create the Fixed Income Securities Database (FISD), which he made accessible to academic and regulatory researchers. Many within the market microstructure research community informally refer to the FISD as simply the "Warga database".

Warga (1991)

Warga [44] uses an econometric model to investigate bond pricing discrepancies that arise when researchers (and commercial bond pricing services) use data from the two different sources that were generally available in 1990. The one source was exchange data in the form of actual transaction prices from the NYSE Automated Bond System (ABS). The other source was OTC dealer data in the form trader-quoted prices.

Warga [44] denotes the unobserved, true value of bond i as P_i^* and the unobserved, true bid-ask spread as BA_i. He assumes that P_i^* is the midpoint of BA_i. He also assumes that the prices/quotes observed in both markets are unbiased in the sense that they deviate from the true unobserved prices/quotes by a random error term. Then, for month-end NYSE transaction prices (P_{NY}):

$$P_{NY_i} = P_i^* + u_i \,,$$

and, for month-end Lehman Brothers bid quotes (P_B):

$$P_{B_i} = P_i^* - \frac{1}{2} BA_i + \zeta_i \,.$$

Combining these two equations and letting $\varepsilon_\iota = \zeta_\iota + \mu_\iota$ yields:

$$P_{B_i} - P_{NY_i} = -\frac{1}{2} BA_i + \varepsilon_i \,.$$

Squaring both sides results in:

$$\left(P_{B_i} - P_{NY_i}\right)^2 = \frac{1}{4}\left(BA_i\right)^2 - \left(BA_i\right)\varepsilon_i + \varepsilon_i^2 \,.$$

Assuming the random error terms are orthogonal to prices/quotes, the expected squared price discrepancies is:

$$E\left[\left(P_{B_i} - P_{NY_i}\right)^2\right] = \frac{1}{4}\left(BA_i\right)^2 + \sigma_{\varepsilon i}^2 \,,$$

where σ_ε^2 equals the variance of the discrepancy.

Warga [44] regresses the squared price discrepancies on six observable liquidity-related variables – bond rating, duration, NYSE dollar trading volume, bond age, issue amount outstanding, and the time of trade of the last trade price on the NYSE – with the following equation:

$$\left(P_{B_i} - P_{NY_i}\right)^2 = \alpha_0 + \alpha_1 MOODYS + \alpha_2 DURTN + \alpha_3 OUTSTD + \alpha_4 DVOL + \alpha_5 AGE + \alpha_6 TIME + \omega \,.$$

Warga [44] finds that squared discrepancies are larger for bonds with lower credit ratings, higher duration, smaller issue sizes, lower trading volume, and trade prices that occurred earlier in the day. While he finds that these variables are capable of explaining some of the observed variation in price discrepancies, he also concludes that commingling exchange and dealer bond pricing data does not induce any biases.

Hong and Warga (2000)

Hong and Warga [29] use a benchmark method to estimate average daily effective spreads with newly available trade data. They obtain exchange market data from the NYSE ABS and OTC dealer market data from the Capital Access International (CAI) database. CAI obtains trading

data on insurance companies, mutual funds, and pension funds from various regulatory filings.

They calculate the effective spread for a given bond on a given day as the dollar-volume-weighted average price transacted at the ask minus the dollar-volume-weighted average price transacted at the bid:

$$\sum_{i=1}^{N} P_i^{\mathrm{A}} W_i^{\mathrm{A}} - \sum_{j=1}^{M} P_j^{\mathrm{B}} W_j^{\mathrm{B}} ,$$

where P_i^{A} is the price of transaction i occurring at the ask, W_i^{A} is the dollar-value weight of transaction i, and N is the number of transactions occurring at the ask for a given bond on a given day. Similarly, P_j^{B} is the price of transaction j occurring at the bid, W_j^{B} is the dollar-value weight of transaction j, and M is the number of transactions occurring at the bid for a given bond on a given day.

Hong and Warga [29] find that the average daily effective spreads for corporate bond transactions occurring on the NYSE ABS that involve at least 10 bonds is about \$0.21 for investment grade bonds and about \$0.19 for high-yield bonds. For corporate bond trades occurring in the OTC dealer market, they find that the average daily effective spreads is about \$0.13 for investment grade bonds and about \$0.19 for high-yield bonds. They note that these spread estimates are smaller than previous estimates based on data from an earlier period, which is consistent with evidence that corporate bond spreads may have declined over time.

Hong and Warga [29] also find that OTC dealer market spreads exhibit much larger dispersion than NYSE ABS spreads. The standard deviations of daily effective spreads for the dealer market are two to three times larger than those for the exchange market. This result suggests that investors, particularly uninformed retail investors, could benefit from more transparency in the OTC markets.

Schultz (2001)

Schultz [42] also uses CAI institutional trade data. He develops an econometric model similar to Warga [44] to estimate average round trip corporate bond trading costs from institutional trade data and estimated contemporaneous bid quotes.

Schultz [42] estimates daily corporate bond bid quotes from the month-end bid quotes available from the Warga database. He develops a threestep estimation procedure that uses daily Treasury bond bid quotes, based on the observation that most of the day-to-day changes in investment grade corporate bond prices are explained by changes in the level of risk-free interest rates.

The first step is to calculate a predicted month-end quote for each corporate bond by taking its previous month-end quote and multiplying it by the change in the price of Treasury bonds of similar duration. The second step is to subtract the predicted month-end quote from the actual month-end quote. This calculation yields the monthly pricing error from predicting that the change in the corporate bond prices is exactly the same as the change in Treasury bond prices. The monthly pricing error is converted to an average daily pricing error by dividing it by the number of trading days in the month. The third step is to estimate the bid quote for a particular within-month trade date by starting with the previous end-of-month quote and adding on the average daily pricing error times the number trading days since the previous month end.

Schultz [42] finds that his bid quote estimates are accurate for investment grade bonds, but not for high-yield bonds. This is not surprising, since changes in high-yield prices are more often due to changes in firm-specific factor than changes in Treasury bond prices. Therefore, he does not attempt to estimate trading costs for high-yield bonds with this methodology.

For investment grade bonds, Schultz [42] estimates round-trip transactions costs by regressing the difference between the CAI trade prices and his estimate of the contemporaneous bid quote on a dummy variable that takes the value of 1 for buys and 0 for sells:

$$\Delta_i = \alpha_0 + \alpha_1 D_i^{\mathrm{Buy}} + \varepsilon_i ,$$

where Δ_i is the price of trade i minus estimated bid price and D_i^{Buy} equals one if trade i is a buy and zero otherwise. The coefficient α_i is an estimate of the average round-trip transaction costs. His estimate of the average round-trip transaction costs across all trades is about \$0.27 per \$100 of par value.

Schultz [42] also examines the determinants of corporate bond trading costs with the following regression:

$$\begin{aligned}\Delta_i = {} & \alpha_0 + \alpha_1 D_i^{\mathrm{Buy}} + \alpha_2 S_i + \alpha_3 D_i^{\mathrm{Inst}} + \alpha_4 D_i^{\mathrm{Deal}} \\ & + \alpha_5 D_i^{\mathrm{Inst}} D_i^{\mathrm{Deal}} + \alpha_6 D_i^{\mathrm{Inst}} S_i + \alpha_7 D_i^{\mathrm{Deal}} S_i + \varepsilon_i ,\end{aligned}$$

where S_i is the signed (positive for buys, negative for sells) natural logarithm of the dollar trade size, D_i^{Inst} is a dummy variable that takes a value of one for buys and negative one for sells by one of the 20 most active institutions and a value of zero otherwise, and D_i^{Deal} is a dummy variable that takes a value of one for buys and negative one for sells if the trade involves one of the 12 active dealers and

a value of zero otherwise. For the interactive term for active institutions and dealers, $D_i^{\text{Inst}} D_i^{\text{Deal}}$, the product of the dummies is positive for buys and negative for sells when the trade involves both an active institution and an active dealer and a value of zero otherwise.

Schultz [42] finds that institutional corporate bond trading costs decline with trade size. He does not find any evidence that trading costs are related to credit rating. But, this is not surprising given the fact that his analysis is limited to the four investment grade rating classes (Aaa, Aa, A, Baa).

Schultz [42] finds that trading costs are lower when a large bond dealer is used. In other words, small bond dealers charge more than large ones. He also finds that inactive institutions pay more than twice as much as active institutions to trade the same bonds. Schultz [42] attributes this result to the lack of transparency in the corporate bond market during his sample period. In an opaque market, obtaining price information is costly, so only active institutions will find it worthwhile to bear them.

Chakravarty and Sarkar (2003)

Chakravarty and Sarkar [12] use CAI data and benchmark methods to calculate "traded bid ask-spreads" over one-day, two-day, and five-day windows in the corporate, municipal, and Treasury bond markets. Similar to Hong and Warga [29], they define the traded bid-ask spread per day as the difference between its mean daily selling price and its mean daily buying price.

To check the sensitivity of their estimates to the requirement of one buy trade and one sell trade for each bond day, Chakravarty and Sarkar [12] calculate spreads over non-overlapping two-day and five-day windows. Their two-day traded bid-ask spread is calculated as the difference between the two-day means of the selling prices and the buying prices.

Chakravarty and Sarkar [12] find that the mean traded bid-ask spread per day per \$100 par value is \$0.21 for corporate bonds, \$0.23 for municipal bonds, and \$0.08 for Treasury bonds. In all three markets, they find that spreads increase with longer time-to-maturity, lower credit ratings, and lower trading volume. These results suggest that spreads are positively related to interest rate risk and credit risk, and negatively related to trading activity. For corporate bonds, Chakravarty and Sarkar [12] find that spreads increase with age.

Chakravarty and Sarkar [12] pool observations across all three bond markets for cross-market comparisons. After controlling for credit risk, Chakravarty and Sarkar [12] find no significant difference in the spreads of corporate bonds and Treasury bonds, but they find that municipal bonds have higher spreads.

Fixed Income Pricing Service (FIPS) Research

FIPS provided new price and volume data that allowed market microstructure researchers to conduct studies that were not previously possible. Alexander, Edwards, and Ferri [1] use FIPS volume data to test various hypotheses about bond liquidity. Alexander, Edwards, and Ferri [2] use FIPS returns and equity returns to tease out new evidence on agency conflicts between stockholders and bondholders. Hotchkiss and Ronen [31] use FIPS returns and equity returns to examine the relative informational efficiency of the corporate bond market. Somewhat surprisingly, they find that that informational efficiency of the corporate bond market is similar to the stock market.

Alexander, Edwards, and Ferri (2000a)

Alexander, Edwards, and Ferri [1] examine the determinants of trading volume of FIPS high-yield bonds using a pooled time-series cross-sectional approach. They use the following linear specification:

$$\begin{aligned} \textit{Trading Volume}_{it} &= \beta_0 + \beta_1 \text{Ln}\,(\textit{Size}_{it}) + \beta_2 \textit{Age}_{it} \\ &+ \beta_3 \textit{Private Equity}_{it} + \beta_4 \textit{Credit Rating}_{it} \\ &+ \beta_5 \textit{Duration}_{it} + \beta_6 \textit{Price Variability} + \varepsilon_{it}\,, \end{aligned}$$

where the dependent variable, *Trading Volume*, is measured in three different ways for each bond i in each month t. The three trading volume measures are the natural log of the average daily number of trades (Ln*(Trades)*), the natural log of the average daily number of bonds traded (Ln*(Bonds)*), and average daily turnover (*Turnover*). Ln*(Size)* is the natural logarithm of the issue's par value outstanding, *Age* is a dummy variable equal to one if the issue has been outstanding for less than two years, *Private Equity* is a dummy variable equal to one if the issue has no public equity outstanding in any part of the month, *Credit Rating* is a dummy variable equal to one if the issue is rated below B- by Standard & Poor's at any point during the month, *Duration* is the bond's modified duration, and *Price Variability* is the monthly average absolute value of the daily percentage change in volume-weighted price.

They find consistent results for all three measures of trading volume. Larger issues and younger issues are more heavily traded. They point out that the age result extends earlier empirical results that found that the liquidity of Treasury securities drops off a few months after issuance.

Alexander, Edwards, and Ferri [1] also find that bonds of firms without public equity trade more frequently than bonds of firms with public equity. This last finding is inconsistent with a disclosure hypothesis that predicts that more relaxed disclosure rules for firms without public equity will lead lower liquidity, as measured by trading volume. However, it is consistent with the competing substitution hypothesis that predicts that high-yield bonds of private firms will attract trading volume that otherwise would have occurred in the equity.

Alexander, Edwards, and Ferri (2000b)

Alexander, Edwards, and Ferri [2] use equity data and FIPS bond data to investigate the relationship between a firm's stock return and its bond return. They examine the long-term co-movement between a firm's bond returns and its stock returns. They also examine stock and bond returns around events typically associated with agency conflicts to see whether their co-movements provide evidence of agency conflicts.

To examine the long-term co-movement between a firm's bond returns and its stock returns, Alexander, Edwards, and Ferri [2] use three regression approaches. The first approach is a time-series regression model:

$$RB_{it} = \beta_0 + \beta_1 XRS_{it} + \beta_2 XRS_{it-1} + \beta_3 RBIND_{it} + \beta_4 RBIND_{it-1} + \varepsilon_{it} ,$$

where RB_{it} is the bond return for firm i on day t, XRS is the current (t) and lagged ($t-1$) excess stock return, and $RBIND$ is the current (t) and lagged ($t-1$) high-yield bond index return, and ε_{it} is the residual bond return for firm i on day t. The second approach is a pooled time-series cross-sectional model that uses the regression equation above and follows the pooling technique of Greene (1993). The third approach is a cross-sectional regression model that follows the approach of Fama and MacBeth. For each sample day, the excess bond returns are regressed on the current and lagged excess stock returns:

$$XRB_{it} = \beta_0 + \beta_1 XRS_{it} + \beta_2 XRS_{it-1} + \varepsilon_{it} .$$

The estimates of β_1 in each of the three regressions show whether the stock and bond returns tend to co-move together (positive), in the opposite direction (negative), or not at all (insignificant). Alexander, Edwards, and Ferri [2] find that all three regressions produce similar results. The β_1 estimates are positive and statistically significant, indicating that excess bond returns are positively correlated with excess stock returns. But, Alexander, Edwards, and Ferri [2] point out the that the magnitudes of the coefficients suggest that the correlation is economically small.

To examine the behavior of stock and bond returns around events typically associated with agency conflicts, Alexander, Edwards, and Ferri [2] look at cumulative excess stock and bond returns around announcements of corporate events that are typically associated with wealth transfers from bondholders to stockholders, or vice versa. Events include debt issuances and redemptions, stock issuances and repurchases, dividend changes, new credit agreements, and others. They use Wilcoxon rank-sum tests to determine whether the means of the cumulative excess bond returns around potentially wealth-transferring events are significantly different from the returns otherwise. They find that the means are significantly different and that the mean cumulative excess bond returns around the wealth-transferring events is negative, while the returns are at other times are positive.

Thus, Alexander, Edwards, and Ferri [2] show that wealth-transferring corporate events (from bondholders to stockholders, or vice versa) can cause a firm's bond returns to diverge from its typical positive (weak) co-movement with its stock returns. In addition, they point out that this result is a likely factor in the weak long-term time-series correlations observed between stock and bond returns.

Hotchkiss and Ronen (2002)

Hotchkiss and Ronen [31] use FIPS data for 55 high-yield bonds to examine the informational efficiency of the corporate bond market relative to the market for the underlying stock. They find that stocks do not lead bonds in reflecting firm-specific information. They also find that pricing errors for bonds are no worse than for the underlying stocks, even on an intraday level.

Hotchkiss and Ronen [31] use a vector autoregression (VAR) approach applied to daily and hourly returns with a return-generating process that includes an interest rate risk factor and an equity market (systematic) risk factor:

$$RB_t = \alpha_t + \sum_{i=1}^{nb} \beta_i^{\mathrm{B}} RB_{t-i} + \sum_{i=0}^{ni} \beta_i^{\mathrm{L}} RL_{t-i} + \sum_{i=0}^{ns} \beta_i^{\mathrm{M}} RM_{t-i} + \varepsilon_t ,$$

where RB_t is the FIPS bond portfolio return, RL_t is the Lehman Intermediate Government Bond Index return, and RM_t is the S&P 500 Index return. The number of lags for the bond, interest rate, and stock returns are $nb = 3$, $ni = 0$, and $ns = 4$, respectively. Hotchkiss and Ronen [31]

include lagged bond returns (RB_{t-i}) to consider autocorrelation-adjusted bond returns. They also consider a specification that replaces the Lehman index return with the default risk-free return, RD_t, as interest rate factor:

$$RB_t = \alpha_t + \sum_{i=1}^{nb} \beta_i^{\mathrm{B}} RB_{t-i} + \sum_{i=0}^{ni} \beta_i^{\mathrm{D}} RD_{t-i} + \sum_{i=0}^{ns} \beta_i^{\mathrm{M}} RM_{t-i} + \varepsilon_t \ .$$

Finally, they add the underlying stock (firm-specific) return, RS_t :

$$RB_t = \alpha_t + \sum_{i=1}^{nb} \beta_i^{\mathrm{B}} RB_{t-i} + \sum_{i=0}^{ni} \beta_i^{\mathrm{D}} RD_{t-i} + \sum_{i=0}^{ns} \beta_i^{\mathrm{M}} RM_{t-i} + \sum_{i=0}^{ns} \beta_i^{\mathrm{S}} RS_{t-i} + \varepsilon_t \ .$$

With these three regressions, Hotchkiss and Ronen [31] find that high-yield bond returns exhibit a very strong interest rate risk component and that this component is significantly greater for higher-rated bonds. They also find that high-yield bond returns exhibit a very strong systematic risk component and that this component is slightly weaker for higher-rated bonds.

To test whether stock returns lead bond returns, Hotchkiss and Ronen [31] conduct Granger causality tests at the daily and hourly levels. They estimate the VAR for the variable set $z_t = [RB_t, RS_t]'$ using the specification:

$$z_t = B_1 z_{t-j} + B_2 z_{t-j} + \mu_t \ ,$$

where RB_t is the bond return and RS_t is the stock return, for day (hour) t, B_i are conformable matrices, and μ_t is a disturbance vector. To test whether stock returns Granger cause bond returns they estimate the following bivariate VAR model using ordinary least squares (OLS):

$$RB_t = c_1 + \sum_{i=1}^{j} a_i RB_{t-i} + \sum_{i=1}^{j} b_i RS_{t-i} + \nu_{1,t} \ ,$$

where c is a constant, as and bs are coefficients, ν_t is the disturbance vector, and j is the lag length. The null hypothesis is that stock returns do not Granger cause bond returns, or that $H_0 = \left[b_i\right] = 0$, for all i. Tests of whether bond returns Granger cause stock returns are conducted in a similar way. F-tests indicate that lagged stock returns are not significant in explaining bond returns. Thus, stocks do not lead bonds in reflecting firm-specific information. The Granger causality test results also indicate that lagged bond returns are not significant in explaining stock returns.

Hotchkiss and Ronen's [31] interpretation of the Granger causality test results is that the contemporaneous correlations between stock returns and bond returns are best described as a joint reactions to common factors. This motivates an additional investigation of the comparative reaction of stocks and bonds to firm-specific information. To conduct this investigation, they examine how quickly firm-specific information contained in earnings announcements are incorporated into bond prices relative to stock prices. First, they compare reported earnings to the median of analysts' earnings forecasts and calculate the log forecast error:

$$FE_i = \ln\left(A_i / F_i\right) \ ,$$

where FE_i is the log forecast error for firm i, A_i is the announced earnings per share, and F_i is the forecast earnings per share. Next, they run the following regressions to examine whether earnings information is reflected in bond returns or stock returns:

$$RB_{[-1,t]} = \alpha_0 + \alpha_1 * FE + \alpha_2 * RM_{[-1,t]} + \varepsilon$$
$$RS_{[-1,t]} = \alpha_0 + \alpha_1 * FE + \alpha_2 * RM_{[-1,t]} + \varepsilon \ ,$$

where RB and RS are the bond and stock returns, respectively, for the period starting at day (hour) -1 prior to the announcement and ending at day $+7$ (hour $+14$) after the announcement, and RM is the market (S&P 500 Index) return. Both the daily and hourly regression results indicate that all information is quickly impounded into both bond prices and stock prices.

Finally, Hotchkiss and Ronen [31] compare the market quality for the high-yield FIPS bonds to the underlying stocks by examining whether price errors of different magnitudes are associated with the different markets. The estimate the following market quality measure:

$$MQ_i = 1 - 2 * \left(\sigma_{si}^2 / \sigma_{Ri}^2\right) \ ,$$

where σ_{si}^2 is the variance of the pricing error described in Hasbrouck [27] and σ_{Ri}^2 is the variance of the return. The intuitive interpretation of this measure is the proportion of the total return variance that is due to fundamental variance. In general, they find that the market quality measure for bonds is no worse than for the underlying stocks.

Municipal Bond Market Research

With the MSRB's introduction of central reporting and price transparency to the municipal bond market, mi-

crostructure researchers were able to make use of a comprehensive source quality transaction-level municipal bond data for the first time. This new data provided researchers the opportunity to develop new methods, examine existing microstructure issues in greater detail, and identify new avenues of research.

Two prominent municipal bond studies, Harris and Piwowar [26] and Green, Hollifield, and Schurhoff [22], develop and use very different methods to examine various economic aspects of trading in the municipal bond market. These two studies provide independent sets of similar and robust results that support two important conclusions related to retail investors. The first is that municipal bonds are expensive for retail investors to trade. Harris and Piwowar [26] and Green, Hollifield, and Schurhoff [22] both find that, unlike in equity markets, municipal bond trading costs decrease with trade size.

The second is that "complexity" is costly for retail investors. Harris and Piwowar [26] find that "instrument complexity" makes municipal bonds more expensive to trade. Instrument complexity is measured in terms of attached features, such as calls, puts, sinking funds, credit enhancement, nonstandard interest payment frequencies, and nonstandard interest accrual methods. Green, Hollifield, and Schurhoff [22] find that "deal complexity" also increases trading costs. Bond dealers charge higher markups on more difficult trades.

Harris and Piwowar (2006)

Harris and Piwowar [26] estimate municipal bond trading costs using an econometric model. They denote the unobserved "true value" of the bond at the time t as V_t and assume that the price of a trade, P_t, is equal to V_t plus or minus a price concession that depends on whether the trade is buyer-initiated or seller-initiated. The absolute customer transaction cost, $c(S_t)$, is estimated as the effective half-spread, measured as a percentage of the price.

I_t^{D} is an indicator variable that takes a value of 1 if the trade was an interdealer trade or 0 if the trade was a customer trade. Q_t is an indicator variable that takes a value of 1 if the customer was a buyer, -1 if the customer was a seller, or 0 if it was an interdealer trade. This results in:

$$\begin{aligned} P_t &= V_t + Q_t P_t c(S_t) + I_t^{\mathrm{D}} P_t \delta_t \\ &= V_t \left(1 + \frac{Q_t P_t c(S_t) + I_t^{\mathrm{D}} P_t \delta_t}{V_t}\right) . \end{aligned}$$

The continuously compounded bond price and "true value" returns between trades t and s, r_{ts}^{P} and r_{ts}^{V} respectively, are found by taking logs of both sides, making two small approximations, and subtracting the same expression for trade s:

$$r_{ts}^{\mathrm{P}} = r_{ts}^{\mathrm{V}} + Q_t c(S_t) - Q_s c(S_s) + I_t^{\mathrm{D}} \delta_t - I_s^{\mathrm{D}} \delta_s .$$

The "true value" return r_{ts}^{V} is represented with a factor model by decomposing it into the linear sum of a time drift, a short-term municipal bond factor return, a long-term municipal bond factor return, and a bond-specific valuation factor, ε_{ts}:

$$\begin{aligned} r_{ts}^{\mathrm{V}} = Days_{ts} \left(5\% - CouponRate\right) \\ + \beta_{\mathrm{Avg}} SLAvg_{ts}^{+} \beta_{\mathrm{Dif}} SLDif_{ts}^{+} \varepsilon_{ts} , \end{aligned}$$

where $Days_{ts}$ counts the number of calendar days between trades t and s, $CouponRate$ is the bond coupon rate. $SLAvg_{ts}$ and $SLDif_{ts}$ are the average and difference, respectively, of continuously compounded short- and long-duration factor returns between trades t and s. The first term models the continuously compounded bond price return that traders expect when interest rates are constant and the bond's coupon interest rate differs from a notional five percent bond, the median coupon rate in their sample. The two index returns model municipal bond value changes due to changes in interest rates and tax-exempt yield spreads. Harris and Piwowar [26] use repeat sales methods to estimate these indices. They assume that the bond-specific valuation factor ε_{ts} has mean zero and variance given by

$$\sigma_{\varepsilon_{ts}}^2 = N_{ts}^{\mathrm{Sessions}} \sigma_{\mathrm{Sessions}}^2$$

where $N_{ts}^{\mathrm{Sessions}}$ is the total number of full and partial trading sessions between trades t and s.

To model customer transaction costs, Harris and Piwowar [26] consider several alternative functional forms that are flexible enough to model very high average trading costs for small trade sizes and very low average trading costs for large trade sizes. Harris and Piwowar [26] choose following parsimonious expression:

$$c(S_t) = c_0 + c_1 \frac{1}{S_t} + c_2 \log S_t + \kappa_t ,$$

where the first three terms specify the cost function that represents average trade costs and κ_t represents the unexplained variation in the observed customer trading costs. The constant term allows total transaction costs to grow in proportion to size. The second term captures fixed costs per trade and the third term allows the costs per bond to vary by size.

The Harris and Piwowar [26] time-series estimation model is obtained by combining the last four equations:

$$r_{ts}^{\mathrm{P}} - Days_{ts}\left(5\% - CouponRate\right) = c_0\left(Q_t - Q_s\right) + c_1\left(Q_t\frac{1}{S_t} - Q_s\frac{1}{S_s}\right) + c_2\left(Q_t\log S_t - Q_s\log S_s\right) + \beta_{\mathrm{SLAvg}}SLAvg_{ts}^{+}\beta_{\mathrm{SLDif}}SLDif_{ts}^{+}\eta_{ts}\ ,$$

where the expression for the regression term, η_{ts}, is given by:

$$\eta_{ts} = \varepsilon_{ts} + Q_t\kappa_t - Q_s\kappa_s + I_t^{\mathrm{D}}\delta_t - I_s^{\mathrm{D}}\delta_s\ .$$

The mean of the error term is zero and its variance is:

$$\sigma_{ts}^2 = N_{ts}^{\mathrm{Sessions}}\sigma_{\mathrm{Sessions}}^2 + D_{ts}\sigma_\delta^2 + \left(2 - D_{ts}\right)\sigma_\kappa^2\ ,$$

where D_{ts} represents the number (0, 1, or 2) of interdealer trades involved in trades t and s. For each bond, Harris and Piwowar [26] separately estimate their time-series transaction cost estimation model using an iterated least squares method, with the weight given by the inverse of the estimates of σ_{ts}^2. For a wide range of trade sizes, they calculate weighted cross-sectional mean cost estimates across all municipal bonds. Each bond's weight is given by the inverse of its estimation error variance at that trade size.

Harris and Piwowar [26] find that retail-size municipal bond trades are substantially more expensive than similar-sized equity trades. Average effective spreads in municipal bonds are almost 2% for representative retail-size trades ($20,000). They point out that this is the equivalent of almost 4 months of total annual return for a bond with a 6% yield-to-maturity.

Harris and Piwowar [26] also find that retail-size municipal bond trades are more expensive than institutional-size trades. Unlike in equities, municipal bond transaction costs decrease with trade size. Harris and Piwowar [26] also find that, unlike in equities, municipal bond transaction costs do not depend on trade frequency. They attribute these results to the lack of price transparency in the municipal bond market during their sample period.

To investigate how estimated transaction costs vary across municipal bonds, Harris and Piwowar [26] conduct cross-sectional weighted least squares regressions for various trade sizes. The dependent variable is the estimated average transaction costs in a given municipal bond at a given trade size. The weight for each bond observation is given by the inverse of the estimation error variance of its cost estimate. The independent variables include measures of credit quality, age, and instrument complexity.

Harris and Piwowar [26] show that bond trading costs increase with credit risk, time to maturity, and time since issuance. They also find that trading costs increase with instrument complexity, and that retail investors are more adversely affected by instrument complexity than institutional investors. They conjecture that investors and issuers might benefit if simpler bonds were issued.

Green, Hollifield, and Schurhoff (2007a)

Green, Hollifield, and Schurhoff [22] focus on trades that can reasonably be assumed to represent two sides of a single intermediated transaction, and employ a structural model to decompose the cost faced by a customer into a portion that represents the cost the dealer incurs and a portion attributable to the dealer's market power. They formulate and estimate a simple structural bargaining model that allows them to estimate measures of dealer bargaining power and relate it to characteristics of the trades.

Green, Hollifield, and Schurhoff [22] use a theoretical model to seek evidence that the high costs of trading are due to dealer market power and to find out how the exercise of market power depends on the characteristics of the trade. They develop a simple theoretical model of the interaction between dealers and their customers in which the expected profits to the dealer reflect both the dealer's costs and his bargaining power relative to the customer. Both of these, in turn, can be parametrized as functions of observable variables, and estimated as a Stochastic Frontier Model. The dealer's cost is the stochastic frontier, which represents the expected mark-up the customer would obtain if dealers were always driven by their reservation values, as they would be if the provision of dealer services were perfectly competitive. The observed markup, expressed in excess returns over a municipal bond index, can be written as:

$$\frac{p_i - p_i^*}{p_i^*} - R_{\mathrm{index},i} = \left[\frac{c\left(X_i,\theta\right)}{p_i^*} - E\left(R_{\mathrm{index},i}\middle|X_i\right)\right] + \varepsilon_i + \xi_i\ ,$$

where p_i is the dealer's selling price, p_i^* is the dealer's purchase price, and $R_{\mathrm{index},i}$ is the municipal bond market index return.

The first term on the right-hand side of the equation represents the dealer's costs in excess of the expected municipal bond index return, where X_i is a set of conditioning variables observable to the buyer and seller and θ is a set of parameters to be estimated. They refer to this term as the cost of intermediation.

The second and third terms capture how the observed markup can differ from the dealer's cost of intermediation.

The second term, ε_i, is a symmetric, normally-distributed error term:

$$\varepsilon_i \equiv \frac{e_i}{p_i^*} - \eta_i \ ,$$

reflecting a zero-mean forecast error:

$$\eta_i = R_{\text{index},i} - E\left(R_{\text{index},i} \middle| X_i\right) \ .$$

The third term, ξ_i, is a one-sided, exponentially-distributed error term:

$$\xi \equiv \frac{\rho_i \left[E\left(p_i \middle| X_i\right) - c\left(X_i, \theta\right) - v_i\right]}{p_i^*} \ ,$$

reflecting the distribution of sellers' reservation values (v_i) and dealer bargaining power.

Green, Hollifield, and Schurhoff [22] estimate restricted and unrestricted versions of the following regression model via maximum likelihood:

$$\frac{p_i - p_i^*}{p_i^*} - R_{\text{index},i} = \theta_0 + \sum_{l=1}^{L} \theta_l X_{il} + \varepsilon_i + \xi_i \ ,$$

with $l = 1, \ldots, L$ conditioning variables. The residual ε_i is normally distributed with standard deviation $b_0 \Pi_{k=1}^{K} e^{b_{ik} Z_k}$, with Z_{ik} for $k = 1, \ldots, K$ conditioning variables. The residual ξ_i is exponentially distributed with mean and standard deviation $a_0 \Pi_{k=1}^{K} e^{a_{ik} Z_k}$. In the "market power" version of their model, all of the parameters are unrestricted. In the restricted ("no market power") model, all of the parameters on the one-sided error are constrained to zero: $a_0 = a_1 = \ldots = a_k = 0$.

The data used by Green, Hollifield, and Schurhoff [22] includes both customer trades and interdealer trades. But, because their data does not identify the specific broker-dealer associated with a given trade, they must infer their trades and profits indirectly by studying pairs of trades that appear to be direct exchanges of bonds through a single dealer. They assume that a customer buy transaction of a given size of a given bond that occurs within a very short time of customer sell transaction of the same size in the same bond are most likely related. The reasonableness of this assumption is confirmed by Harris and Piwowar [26], whose data contains dealer identities.

Green, Hollifield, and Schurhoff [22] find that municipal bond dealers earn lower average markups on larger trades, even though larger trades lead the dealers to bear more risk of losses. Their results suggest that municipal bond dealers exercise substantial market power, particularly in retail-sized transactions. Their measures of market power decrease in trade size and increase in variables that indicate the complexity of the trade for the dealer.

Transaction Reporting and Compliance Engine (TRACE) Research

TRACE not only brought unprecedented transparency to corporate bond market investors, it also provided an unprecedented opportunity for market microstructure researchers to examine new issues. Chief among them was the "natural experiment" of adding price transparency to an opaque market. Three prominent studies (collectively, "the TRACE studies") that examined the introduction of price transparency to the corporate bond market were Edwards et al. [17], Bessembinder, Maxwell, and Venkataraman [7], and Goldstein, Hotchkiss, and Sirri [20].

These TRACE studies were very complementary in terms of their contributions to the market microstructure literature. To understand the full impact of this collective research, it is important to remember that they were written at a time when many market participants and some regulators were concerned that public dissemination of bond pricing data might have an adverse impact on liquidity. Using different experimental designs and empirical methods, the TRACE studies produced similar results, conclusions, and implications for regulatory policymakers. Overall, the results in all three TRACE studies show that public investors benefit significantly from the introduction of price transparency.

Edwards et al. [17] estimate transaction costs for all corporate bonds that trade at least nine times between January 2003 and January 2005. Their TRACE data set includes all reported OTC trades in corporate bonds, whether transparent or not. Consistent with the results of Harris and Piwowar [26] for the municipal bond market, Edwards et al. [17] find that corporate bonds are expensive for retail investors to trade and that corporate bond transaction costs decrease significantly with trade size. They find that effective spreads in corporate bonds average 1.24% of the price of representative retail-sized trades ($20,000). They point out that this is the equivalent of over 2 months of the total annual return for a bond with a 6% yield to maturity, or 52 cents per share for a $40 stock. In cross-sectional tests, Edwards et al. [17] find that transaction costs are lower for highly rated bonds, recently issued bonds, and bonds close to maturity.

Edwards et al. [17] find that costs are lower for bonds with transparent trade prices, and they drop when the TRACE system starts to publicly disseminate their prices. Their results suggest that introduction of price transparency results in a drop in customer trading costs of at least 5 basis points (bps). In 2003, public investors traded approximately $2 trillion in bonds for which prices were not disseminated. If the prices for these bonds had been

TRACE-transparent, a quick back-of the-envelope calculation shows investors could have saved a minimum of $1 billion that year. Edwards et al. [17] point out that the $1 billion figure represents a lower bound for two reasons. First, because many unsophisticated investors were unaware that prices became available, and because learning how to use the price data takes, time, the long-run benefits are undoubtedly much greater. Second, they do not capture the initial reduction in trading costs at the initiation of TRACE. Bessembinder, Maxwell, and Venkataraman [7] find that sophisticated institutional investors benefited from an immediate reduction in trading costs of about $1 billion.

Bessembinder, Maxwell, and Venkataraman [7] estimate their trade execution costs for a sample of institutional (insurance company) trades in corporate bonds before and after the initiation of public transaction reporting for some bonds through the TRACE system in July 2002. They find that the average reduction in one-way trading costs or bonds eligible for TRACE transaction reporting is about 5 to 8 bps. This translates to a reduction in trade execution costs of about 50%. Moreover, they find a 20% reduction for bonds not eligible for TRACE reporting. Bessembinder, Maxwell, and Venkataraman [7] interpret their results as suggesting that better pricing information regarding some bonds also improves valuation and execution cost monitoring for related bonds. They find no evidence that market quality deteriorated in other dimensions.

Bessembinder, Maxwell, and Venkataraman [7] also find that larger trading cost reductions for less liquid and lower-rated bonds, and for larger trades. They estimate that their results equate to annual trading cost reductions of roughly $1 billion per year for the entire corporate bond market, reinforcing that market design can have first-order effects, even for relatively sophisticated institutional customers.

Goldstein, Hotchkiss, and Sirri [20] design and construct a controlled experiment to examine the impact of introducing price transparency on liquidity for BBB-rated corporate bonds. They selected the 120 BBB-rated bonds for which the NASD began disseminating trade data on April 14, 2003. They simultaneously selected a control sample of non-disseminated bonds.

Goldstein, Hotchkiss, and Sirri [20] find that DRT spreads decrease for most BBB-rated corporate bonds whose prices become transparent, and that this effect is strongest for intermediate trade sizes. The only caveat to this result is that they do not find any significant transparency effect for the most thinly-traded bonds. Overall, Goldstein, Hotchkiss, and Sirri [20] conclude that their finds indicate that the introduction of post-trade price transparency has a neutral or positive effect on market liquidity.

The similar results and conclusions in the three complementary TRACE studies collectively generate important policy implications. Foremost, policymakers should take comfort in the fact that there are few, if any, instances in the combined results that show any harm to investors from introducing price transparency to securities markets. To the contrary, the results show that both retail and institutional investors benefit from price transparency. The empirical results from the TRACE studies support the well-founded economic theoretical arguments that transparency should lower transaction costs, especially for smaller trades.

Speeches and testimony by US bond market regulators, such as those listed in the bibliography, show that these studies critically informed the debate over adding price transparency to the US bond markets. Moreover, they continue to provide important lessons for policy makers in bond markets outside of the United States. The bibliography also contains a partial listing of international reports, discussion papers, and conference proceedings that prominently cite the TRACE studies.

Edwards, Harris, and Piwowar (2007)

Edwards et al. [17] apply the Harris and Piwowar [26] econometric approach to corporate bonds. They also extend the approach by allowing liquidity to be time varying. This extension allows them to examine how the introduction of price transparency affects corporate bond transaction costs.

They model the unobserved value return r_{ts}^{V} by decomposing it into the linear sum of a time drift, an average bond index return, differences between index returns for long and short term bonds and for high and low quality bonds, and a bond-specific valuation factor, ε_{ts}.

$$r_{ts}^{V} = Days_{ts}\left(DriftRate\right) + \beta_1 AveIndexRet_{ts} + \beta_2 DurationDif_{ts} + \beta_3 CreditDif_{ts} + \varepsilon_{ts}\,,$$

where $Days_{ts}$ counts the number of calendar days between trades t and s, $DriftRate$ is the bond coupon rate subtracted from five percent, $AveIndexRet_{ts}$ is the index return for the average bond between trades t and s and $DurationDif_{ts}$ and $CreditDif_{ts}$ are the corresponding differences between index returns for long and short term bonds and high and low credit risk bonds. The first term accounts for the continuously compounded bond price return that traders expect when interest rates are constant and the bond's

coupon interest rate differs from five percent. The three factor returns account for bond value changes due to shifts in interest rates and credit spreads. Edwards et al. [17] estimate the bond indices using repeat sale regression methods with terms that account for bond transaction costs. Finally, the bond-specific valuation factor ε_{ts} has mean zero and variance given by

$$\sigma^2_{\varepsilon_{ts}} = N_{ts}^{\text{Sessions}} \sigma^2_{\text{Sessions}} ,$$

where N_{ts}^{Sessions} is the number of trading sessions and fractions of trading sessions between trades t and s.

Edwards et al. [17] model customer transaction costs using the following additive expression:

$$c(S_t) = c_0 + c_1 \frac{1}{S_t} + c_2 \log S_t + c_3 S_t + c_4 S_t^2 + \kappa_t ,$$

where κ_t represents variation in the actual customer transaction cost that is unexplained by the average transaction cost function. This variation may be random or due to an inability of the average transaction cost function to represent average trade costs for all trade sizes. They assume κ_t has zero mean and variance given by σ^2_κ.

The first three terms of the cost function are the same as in Harris and Piwowar [26], where the constant term allows total transaction costs to grow in proportion to size, the second term characterizes any fixed costs per trade, and the third term allows for costs per bond to vary by trade size. The two additional terms allow more flexibility in the costs to vary by size. Because corporate bonds trade more frequently than municipal bonds, Edwards et al. [17] did not need to be as concerned about degrees of freedom as Harris and Piwowar [26].

Combining the last three equations produces the Edwards et al. [17] version of the Harris and Piwowar [26] transaction cost estimation model:

$$\begin{aligned} r^{\text{P}}_{ts} - Days_{ts}\left(DriftRate\right) &= c_0 \left(Q_t - Q_s\right) \\ &+ c_1 \left(Q_t \frac{1}{S_t} - Q_s \frac{1}{S_s}\right) + c_2 \left(Q_t \log S_t - Q_s \log S_s\right) \\ &+ c_3 \left(Q_t S_t - Q_s S_s\right) + c_4 \left(Q_t S_t^2 - Q_s S_s^2\right) \\ &+ \beta_1 AveIndexRet_{ts} + \beta_2 DurationDif_{ts} + \beta_3 CreditDif_{ts} \\ &+ \eta_{ts} , \end{aligned}$$

where the left hand side is simply the continuously compounded bond return expressed as the equivalent rate on a notional five percent coupon bond. Edwards et al. [17] estimate their time-series model in the same way as Harris and Piwowar [26].

Edwards et al. [17] extend Harris and Piwowar [26] by introducing a pooled time-series regression model that they use to estimate average transaction costs for each day for a class of bonds. With this model, they are able to estimate the daily average transaction costs for bonds that became transparent in 2003, and compare these estimates to those for comparable bonds that were either TRACE-transparent throughout 2003 or never TRACE-transparent in 2003.

The pooled time-series regression model that Edwards et al. [17] use to estimate daily transaction costs differs in two respects from the time-series regression model that they use to estimate average transaction costs for a given bond. First, they specify separate average transaction cost functions, $c_T\left(S_t\right)$, for each day T in the sample. Second, to minimize the total number of parameters to be estimated, they use the three-parameter average cost function:

$$c_T\left(S_t\right) = c_{0T} + c_{1T} \frac{1}{S_t} + c_{2T} \log S_t + \kappa_t .$$

For a given bond, the change in value between bond trades is modeled as:

$$\log V_t - \log V_s = f_s r_S + \sum_{J=S+1}^{T-1} r_J + f_t r_T + e_{st} ,$$

where S is the day on which trade s took place and T is the day on which a subsequent trade t took place, r_J is the common index return (to be estimated) for day J and f_s and f_t, respectively, are the fractions of the S and T trading days overlapped by the period spanned by transactions s and t. This portion of the specification is the same as appears in many paired trade regression index estimation procedures. With these changes, the regression model is

$$\begin{aligned} r^{\text{P}}_{ts} - Days_{ts}\left(5\% - CouponRate\right) &= c_{0T} Q_t - c_{0S} Q_s \\ &+ c_{1T} Q_t \frac{1}{S_t} - c_{1S} Q_s \frac{1}{S_s} + c_{2T} Q_t \log S_t - c_{2S} Q_s \log S_s \\ &+ f_s r_S + \sum_{J=S+1}^{T-1} r_J + f_t r_T + \eta_{ts} . \end{aligned}$$

They use iterated weighted least squares methods, where the weights are equal to the predicted values of the regression of the squared residuals on the independent variables appearing in the residual variation expression. Edwards et al. [17] estimate the model using a three-month wide sliding window that moves forward one month at a time. The time series of coefficient estimates are assembled from the center months of each of the sliding regressions. They compute transaction costs for various transaction sizes by evaluating the estimated transaction cost functions at the given transaction sizes. Using

the estimated variance-covariance matrix of the estimators, they also compute daily standard errors of the various daily transaction cost estimates.

Bessembinder, Maxwell, and Venkataraman (2006)

Bessembinder, Maxwell, and Venkataraman [7] develop and estimate an indicator variable regression model:

$$\Delta P = a + wX_t + \gamma SQ_t^* + \alpha S\Delta Q + \omega_t \,,$$

where ΔP is the change in the price of the bond from time $t-1$ to time t, a is the regression intercept, w is the fraction of public information that is observable in the data with realizations X_t, γS is the informational component of the spread, αS is the non-informational component of the spread (where $\alpha = (1-\gamma)$), Q_t^* is the market maker's estimate of bond value due to surprises in order flow, ΔQ is the change in indicator variable Q (which takes a value of 1 if the trade is a customer buy and -1 if it is a customer sell) from time $t-1$ to time t, and ω_t is the regression error term.

Bessembinder, Maxwell, and Venkataraman [7] develop this regression model in the following way. $E_t(V)$ is the market-maker's estimate of the bond's unobserved true value (V) at time t conditional on whether the observed trade is a customer buy or a customer sell. Transaction prices are given by:

$$P_t = E_t(V) + \alpha SQ_t \,.$$

The market maker's estimate of bond value at time t, $E_t(V)$, is her estimate from the prior period, $E_{t-1}(V)$, updated to reflect surprises in order flow, $Q_t - E_{t-1}(Q_t)$, and public information revealed since the prior period, η_τ Substitution yields:

$$E_t(V) = E_{t-1}(V) + \gamma SQ_t^* + \eta_t \,,$$

where $Q_t^* = Q_t - E_{t-1}(Q_t)$. To allow for the possibility that bond market order flow is positively autocorrelated, Bessembinder, Maxwell, and Venkataraman [7] assume that it follows a simple AR1 process, so that $E_{t-1}(Q_t) = \rho(Q_{t-1})$. The change in the price of the bond from time $t-1$ to time t is:

$$P_t - P_{t-1} = \gamma SQ_t^* + \alpha SQ_t - \alpha SQ_{t-1} + \eta_t \,.$$

Substituting ΔP for $P_t - P_{t-1}$ and ΔQ for $Q_t - Q_{t-1}$, this expression can be rewritten as:

$$\Delta P = \gamma SQ_t^* + \alpha S\Delta Q + \eta_t \,.$$

To incorporate observable public information that affects bond value, they assume that a fraction w of public information becomes observable in the data with realizations X_t, while the remaining portion $(1-w)$ is due to unobservable innovations U_t that represent statistical noise. Substitution yields their regression model:

$$\Delta P = wX_t + \gamma SQ_t^* + \alpha S\Delta Q + \omega_t \,,$$

where $\omega = (1-w)\,U_t$. Bessembinder, Maxwell, and Venkataraman [7] show that their model is equivalent to the Madhavan et al. [39] model. Moreover, in the special case of no autocorrelation in order flow ($\rho = 0$), their model is equivalent to Huang and Stoll [33], Schultz [42] and Warga [44].

Goldstein, Hotchkiss, and Sirri (2007)

Goldstein, Hotchkiss, and Sirri [20] use two different methods to estimate transaction costs for a sample of BBB-rated bonds. Their first method involves identifying "dealer round-trip" (DRT) transaction chains. These transaction chains involve a dealer purchasing a bond from a customer and then selling that same bond to another customer within a specified period of time. DRT spreads are calculated as the difference between the customer buy price at the end of the transaction chain and the customer sell price at the beginning of the chain. Their DRT method is similar to the methods used in the municipal bond studies of Green et al. [22] and Biais and Green [9], except their data contains individual dealer identifiers. Goldstein, Hotchkiss, and Sirri [20] estimate DRT spreads for transaction chains that occur with one-day, five-day, and unlimited time intervals. They estimate DRT spreads for various trade size groups.

Their second method is a regression method similar to Warga [44] and Schultz [42]. For each trade size group, Goldstein, Hotchkiss, and Sirri [20] estimate spreads by regressing the difference between the transaction price for a customer (T) and an estimated bid price (B) on a dummy variable that equals one for customer buys and zero for customer sells:

$$[T-B]_i = \alpha_0 + \alpha_1 D_i^{\text{Buy}} + \varepsilon_i \,,$$

where estimated bid prices are obtained from Reuters dealer bid price estimates from the end of the day prior to the transaction. Reuters estimates are based on daily quotes obtained directly from individual dealers.

Goldstein, Hotchkiss, and Sirri [20] estimate a second regression to consider the effect of dissemination while controlling for other bond characteristics impacting the

spread:

$$[T-B]_i = \alpha_0 + \alpha_1 D_i^{\text{Buy}} + \alpha_2 D_i^{\text{DisseminatedBond}} + \alpha_3 D_i^{\text{Post-disseminationPeriod}} + \alpha_4 D_i^{\text{DisseminationBond}^* \text{Post-disseminationPeriod}} + \alpha_5 X_5 + \cdots + \alpha_{10} X_{10} + \varepsilon_i \,,$$

where additional dummies are included for disseminated bonds, transactions that occur in the post-dissemination period, and the interaction of these two dummies for transaction in disseminated bonds that occur in the post-dissemination period. As in Schultz [42] the additional dummies are expressed a $+1$ for buys and -1 for sells. Variables $X_5, \ldots, X_{10}$ are six bond characteristics related to spreads: trade size, time-to-maturity, age, issue amount, average daily volume over the prior 30 days, and days since last trade.

The Links Between Bond Market Microstructure Research and Other Finance and Economics Research

The discussion of bond market research has thus far been presented solely within the framework of the market microstructure literature. However, some of the most interesting bond market research is connected to other areas of finance and economics. The instrument complexity results of Harris and Piwowar [26], for example, provide evidence to support Carlin's [11] formal model of strategic price complexity in security design for retail markets. Additionally, Chen, Lesmond, and Wei [13] provide an example of how bond market microstructure research is linked to asset pricing models in finance. Green, Hollifield, and Schurhoff [23] develop a theoretical model that is analogous to the costly consumer search models in the broader economics literature.

Chen, Lesmond, and Wei (2007)

Beginning with Amihud and Mendelson [3], market microstructure research has consistently shown that a liquidity premium exists in equity markets. Recently, bond market microstructure researchers have begun investigating whether a liquidity premium also exists in bond markets. One such paper is Chen, Lesmond, and Wei [13]. Their investigation of the link between bond liquidity and corporate yield spreads provides important implications for the default risk literature.

Chen, Lesmond, and Wei [13] investigate whether liquidity is priced in corporate yield spreads. They use several approaches, including a regression approach that is an extension to the Lesmond, Ogden, and Trzcinka [36] (LOT) approach developed for equities. The LOT approach assumes that a zero return day (or a non-trading day) is observed when the true price changes by less than the transaction costs. Because marginal informed investors will only trade on information if the trade yields expected profits net of transaction costs, an individual bond's trading costs represent a threshold that must be exceeded before its return will reflect new information. The premise of this approach is that if the value of the information is insufficient to exceed the costs of trading, then the marginal investor will either reduce trading or not trade, causing a zero return.

Chen, Lesmond, and Wei [13] extend the LOT approach to corporate bonds by applying a two-factor return-generating model to estimate bond trading costs:

$$R_{j,t}^* = \beta_{j1} Duration_{j,t}^* \Delta R_{ft} + \beta_{j2} Duration_{j,t}^* \Delta S\&P\ Index_t + \varepsilon_{j,t} \,,$$

where $R_{j,t}^*$ is the unobserved "true" return for bond j on day t that investors would bid given zero trading costs. The daily change in the 10-year risk-free interest rate, ΔR_{ft}, is the factor that is more important for investment grade bonds, while the second factor, $\Delta S\&P\ Index_t$, the daily return on the Standard & Poor's 500 equity index, is more important for high-yield bonds. Both factors are scaled by $Duration_{j,t}$, the bond's duration.

Chen, Lesmond, and Wei [13] then apply the Amihud and Mendelson [3] liquidity premium to bonds. In Amihud and Mendelson [3], observed asset prices differ from true values because of a liquidity premium that compensates investors for liquidity costs. Chen, Lesmond, and Wei [13] state the liquidity effects on bond returns as:

$$R_{j,t} = R_{j,t}^* - \alpha_{i,j} \,,$$

where $R_{j,t}$ is the measured return, $\alpha_{2,j}$ is the effective buy-side cost, and $\alpha_{1,j}$ is the effective sell-side cost for bond j. The resulting liquidity constraint is:

$$\begin{aligned} R_{j,t} &= R_{j,t}^* - \alpha_{1j} && \text{if } R_{j,t}^* < \alpha_{1j} && \text{and } \alpha_{1j} < 0 \\ R_{j,t} &= 0 && \text{if } \alpha_{1j} \le R_{j,t}^* \le \alpha_{2j} \\ R_{j,t} &= R_{j,t}^* - \alpha_{2j} && \text{if } R_{j,t}^* > \alpha_{2j} && \text{and } \alpha_{2j} > 0 \,. \end{aligned}$$

Combining the liquidity constraint with the return generating model, Chen, Lesmond, and Wei [13] us a maximum likelihood method outlined in LOT to estimate transaction costs. They specify the log-likelihood function

as:

$$\begin{aligned}\mathrm{Ln}L = &\sum_1 \mathrm{Ln}\frac{1}{\left(2\pi\sigma_j^2\right)^{1/2}} \\ &- \sum_1 \frac{1}{2\sigma_j^2}\left(R_j + \alpha_{1,j} - \beta_{j1}Duration_{j,t} * \Delta R_{ft}\right. \\ &\qquad \left. -\beta_{j2}Duration_{j,t} * \Delta S\&P\ Index\right)^2 \\ &+ \sum_2 \mathrm{Ln}\frac{1}{\left(2\pi\sigma_j^2\right)^{1/2}} \\ &+ \sum_2 \frac{1}{2\sigma_j^2}\left(R_j + \alpha_{2,j} - \beta_{j1}Duration_{j,t} * \Delta R_{ft}\right. \\ &\qquad \left. -\beta_{j2}Duration_{j,t} * \Delta S\&P\ Index\right)^2 \\ &+ \sum_0 \mathrm{Ln}\left(\Phi_{2,j} - \Phi_{1,j}\right)\ ,\end{aligned}$$

where $\Phi_{i,j}$ represents the cumulative distribution function for each bond-year evaluated at:

$$\frac{\left(\alpha_{1,j} - \beta_{j1}Duration_{j,t} * \Delta R_{ft}\right)}{\sigma_j} - \frac{\left(\beta_{j2}Duration_{j,t} * \Delta S\&P\ Index_t\right)}{\sigma_j}\ .$$

Σ_1 (region 1) represents the negative nonzero measured returns, Σ_2 (region 2) represents the positive nonzero measured returns, and Σ_0 (region 0) represents the zero measured returns. The difference in the buy-side and sell-side cost estimates, $\alpha_{2,j} - \alpha_{1,j}$, represents round-trip trading costs.

The model's implicit assumption that information motivates bond trades and that information is efficiently impounded into bond prices is supported by the results of Hotchkiss and Ronen [31]. The error term captures noise trading and trades due to unanticipated public information.

In addition to LOT estimates, Chen, Lesmond, and Wei [13] use bid-ask spreads and zero-returns as liquidity cost measures. The bid-ask spreads the use are bond-year proportional bid-ask spreads, calculated as the average of quarterly proportional spreads. Quarterly proportional spreads are calculated from quarterly bid-ask spreads obtained from Bloomberg consensus quotes among market participants, divided by the average bid and ask price. Zero-returns are simply the percentage of days with returns equal to zero.

They find that liquidity costs are related to credit rating. Liquidity costs are much higher for high-yield bonds than for investment grade bonds. They also find that liquidity costs are related to maturity. Liquidity costs for long-maturity bonds are higher than for short-maturity bonds.

They also find that yield spreads generally increase with maturity for investment grade bonds. But, they find that yield spreads generally decrease with maturity for high-yield bonds. They point out the endogeneity issue stemming from the Helwege and Turner [28] finding that relatively safer firms within the same high-yield credit rating category tend to issue longer-term bonds. This endogeneity issue causes the average yield spread to decline with maturity for high-yield bonds.

To investigate whether liquidity is priced in corporate yield spreads, Chen, Lesmond, and Wei [13] first run the following regression specification for investment grade bonds and high-yield bonds separately:

$$\begin{aligned}Yield\ Spread_{it} &= \eta_0 + \eta_1 Liquidity_{it} + \eta_2 Maturity_{it} \\ &+ \eta_3 Amount\ Outstanding_{it} + \eta_4 Coupon_{it} \\ &+ \eta_5 Treasury\ Rate_{it} + \eta_6 10Yr - 2Yr\ Treasury\ Rate_{it} \\ &+ \eta_7 EuroDollar_{it} + \eta_8 Volatility_{it} + \eta_9 Bond\ Rating_{it} \\ &+ \eta_{10} PreTax\ Coverage\ Dummy_{it} \\ &+ \eta_{11} Operating\ Income/Sales_{it} + \eta_{12} Debt/Assets_{it} \\ &\qquad + \eta_{13} Debt/Capitalization_{it} + \varepsilon_{it}\ ,\end{aligned}$$

where the subscript *it* refers to bond *i* in year *t*. *Liquidity* refers to the three liquidity cost measures – bid-ask spread, zero-returns, or the LOT estimate. Additional variables control for bond-specific, firm-specific, and macroeconomic factors. *Maturity* is the number of years until the bond matures relative to the year being analyzed, and *Amount Outstanding* is natural logarithm of the dollar amount outstanding, *Coupon* is the bond coupon rate. *Treasury Rate* is the 1-year Treasury Note rate, *10Yr-2Yr Treasury Rate* is the difference between the 10-year and 2-year Treasury rates, and *Eurodollar* is the 30-day Eurodollar rate minus the 3-month T-Bill rate. *Volatility* is the equity volatility for each issuer and *Bond Rating* is a credit rating scale that ranges from 1 (AAA rating) to 10 (BBB- rating) for investment grade bonds and from 1(BB+ rating) to 12 (D rating) for high-yield bonds. *Pre-Tax Coverage Dummy* represents four dummy variables corresponding to groupings of pre-tax income, *Operating Income/Sales*, *Debt/Assets*, and *Debt/Capitalization* are each firm's respective accounting ratios.

They find that all three liquidity measures are positively related to the yield spread in both the investment grade and high-yield samples. The liquidity coefficients are statistically significant at the 1% level in every scenario. This provides strong evidence that liquidity is priced in corporate yield spreads. This finding is robust to control-

ling for issuer influences with issuer fixed-effects regressions. The only caveat is that they achieve slightly weaker results for the zero-return liquidity cost measure than for bid-ask spreads and LOT estimates. This finding is also robust to controlling for potential endogeneity problems arising from the contemporaneous measurement of the yield spread, liquidity costs, and credit rating. They perform this robustness check by employing a simultaneous equations model using three equations that correspond to each of the potentially endogenous variables:

$$\begin{aligned}
&Yield\ Spread_{it} = \eta_0 + \eta_1 Liquidity_{it} + \eta_2 Maturity_{it} \\
&+ \eta_3 Coupon_{it} + \eta_4 Treasury\ Rate_{it} \\
&+ \eta_5 10Yr - 2Yr\ Treasury\ Rate_{it} + \eta_6 EuroDollar_{it} \\
&+ \eta_7 Volatility_{it} + \eta_8 Bond\ Rating_{it} \\
&+ \eta_9 PreTax\ Coverage\ Dummy_{it} \\
&+ \eta_{10} Operating\ Income/Sales_{it} + \eta_{11} Debt/Assets_{it} \\
&\qquad + \eta_{12} Debt/Capitalization_{it} + \varepsilon_{it}\,,
\end{aligned}$$

$$\begin{aligned}
&Liquidity_{it} = \eta_0 + \eta_1 Maturity_{it} + \eta_2 Age_{it} \\
&+ \eta_3 Amount\ Outstanding_{it} + \eta_4 Bond\ Rating_{it} \\
&\quad + \eta_5 Bond\ Volatility_{it} + \eta_6 Yield\ Spread_{it} + \varepsilon_{it}\,,
\end{aligned}$$

$$\begin{aligned}
&Credit\ Rating_{it} = \eta_0 + \eta_1 Treasury\ Rate_{it} \\
&+ \eta_2 10Yr - 2Yr\ Treasury\ Rate_{it} \\
&+ \eta_3 PreTax\ Coverage\ Dummy_{it} \\
&+ \eta_4 Operating\ Income/Sales_{it} \\
&+ \eta_5 Debt/Assets_{it} + \eta_6 Debt/Capitalization_{it} \\
&\qquad + \eta_7 Yield\ Sprad_{it} + \varepsilon\,.
\end{aligned}$$

The model is estimated using twostage least squares. The simultaneous equation model estimation results show that the potential endogeneity does not affect the relation between liquidity and yield spreads for either the investment grade or the high-yield bonds.

Thus, Chen, Lesmond, and Wei [13] find extremely consistent and robust evidence that liquidity is a key determinant in corporate yield spreads. This finding provides at least a partial explanation for the findings of Collin-Dufresne, Goldstein, and Martin [15] and others who show that default risk does not completely explain corporate yield spreads.

Green, Hollifield, and Schurhoff (2007b)

Green, Hollifield, and Schurhoff [23] examine secondary market trading in newly issued municipal bonds for the first 60 trading days of their lives. They begin by descriptively documenting the price behavior of newly issued municipal bonds. They show that municipal bonds are underpriced when issued. But, unlike equities, the average price rises slowly over a period of several days. Green, Hollifield, and Schurhoff [23] also find that the observed price patterns are complex. High levels of price dispersion are observed for small trade sizes in the aftermarket for new municipal bond issues. While some small traders purchase bonds on attractive terms, others do not. In contrast, there is very little price dispersion for large trade sizes. Virtually all of the large traders purchase bonds on attractive terms.

They argue that the price level and dispersion patterns are the result of bond dealers discriminating between informed and uninformed customers. Accordingly, Green, Hollifield, and Schurhoff [23] develop and estimate a mixed distribution model for the markups that uninformed and informed investors pay when they purchase newly issued bonds. Their model incorporates investor search costs, i. e., the costs in terms of time and effort needed for investors to become informed about new bond issues.

The mixed distribution model of Green, Hollifield, and Schurhoff [23] is analogous to economic models of costly consumer search, such as the gametheoretic model of Shilony [43] that focuses on advertising and price competition among retail stores in homogeneous product markets. Shilony [43] assumes that all stores are required to advertise, but the advertising is segmented (e. g., signs are posted on store windows). Consumers have a preference for the particular store that that offers them free access to the advertising (e. g., the store right outside their house or the one that they regularly visit) and they will pay more for a product at this store even if it does not offer the lowest price.

The institutional mechanisms of the primary market and the structure of the secondary market for municipal bonds fits particularly well with the informational interpretation of Shilony's [43] model. Every investor has free information about the price that will be charged by his broker. Also, because all firms must disseminate their last sale information on a real-time basis, the investor can choose to look on www.investinginbonds.com or some other free website to find the range of prices charged by all brokers. But, this last sale information does not identify which broker charged the lowest price. To find this out, the investor must incur some cost.

Green, Hollifield, and Schurhoff [23] begin with the assumption that there are both observable and unobservable sources of heterogeneity in the costs investors face in gathering and using information about prices of new munici-

pal issues. They assume that for investor i, the difference between the benefit and the cost of learning about a new issue is z_i^* with:

$$z_i^* = w_i\delta + \mu_i \ ,$$

where w_i is a vector of conditioning variables, δ is a parameter vector, and μ_ι is an error term. The error term is observed by the investor, but not by the econometrician. Investor i becomes informed about the price of a new issue if and only if $z_i^* \geq 0$. They do not observe z_i^*, but they do observe w_i and the price the investor pays for the bond.

An investor who is uninformed about the reoffering price for a new bond is willing to pay the percentage markup y_U of:

$$y_{Ui} = x_i\beta + \varepsilon_{Ui} \ ,$$

where x_i is a vector of conditioning variables, β is a parameter vector, and ε_{Ui} is an error term. Similarly, an investor who is informed about the underwriter's pricing of a new bond is willing to pay the percentage markup y_I of:

$$y_{Ii} = x_i\gamma + \varepsilon_{Ii} \ ,$$

where x_i is a vector of conditioning variables, γ is a parameter vector, and ε_{Ii} is an error term. The uncertainty about the percentage markup is expected to be lower when the investor is informed than when the investor is uninformed:

$$\sigma_I < \sigma_U \ .$$

They use this condition to empirically identify the informed versus uninformed distributions from which the observed markups, y_i, are drawn:

$$y_i = \begin{cases} y_{Ui} & \text{if} \quad z_i^* < 0 \ , \\ y_{Ii} & \text{if} \quad z_i^* \geq 0 \ . \end{cases}$$

Green, Hollifield, and Schurhoff [23] use iterated expectations to show that investors take the markup into account when deciding whether to become informed about an upcoming bond issue or not:

$$\begin{aligned} E\left(y_i \,|w_i, x_i\right) &= E\left(y_i \,|\text{Informed}_i, w_i, x_i\right) \\ &\qquad \Pr\left(\text{Informed}_i \,|w_i\right) \\ &+ E\left(y_i \,|\text{Uninformed}_i, w_i, x_i\right) \Pr\left(\text{Uninformed}_i \,|w_i\right) . \end{aligned}$$

They estimate their model under the assumption that the error terms are drawn independently and identically from a multivariate normal distribution:

$$\begin{pmatrix} u_i \\ \varepsilon_{Ui} \\ \varepsilon_{Ii} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho_U\sigma_U & \rho_I\sigma_I \\ \rho_U\sigma_U & \sigma_U^2 & 0 \\ \rho_I\sigma_I & 0 & \sigma_I^2 \end{bmatrix}\right),$$

where ρ_U is the correlation between u_i and ε_{Ui} and ρ_I is the correlation between u_i and ε_{Ii}. Denoting the cumulative standard normal distribution as Φ and the standard normal density as φ, Green, Hollifield, and Schurhoff [23] show that the condition that investor i becomes informed if and only if $z_i^* \geq 0$ implies that:

$$\begin{aligned} \Pr\left(\text{Informed}_i \,|w_i\right) &= \Pr\left(z_i^* \geq 0 \,|w_i\right) \\ &= \Pr\left(u_i \geq -w_i\delta \,|w_i\right) \\ &= \Phi\left(w_i\delta\right) \ , \end{aligned}$$

and

$$\Pr\left(\text{Uninformed}_i \,|w_i\right) = 1 - \Phi\left(w_i\delta\right) \ .$$

By combining equations and using the distributional assumptions of the error terms, Green, Hollifield, and Schurhoff [23] show that

$$E\left(y_i \,|\text{Informed}_i, w_i, x_i\right) = x_i\gamma + \rho_I\sigma_I \frac{\phi\left(w_i\delta\right)}{\Phi\left(w_i\delta\right)} \ ,$$

and

$$E\left(y_i \,|\text{Uninformed}_i, w_i, x_i\right) = x_i\beta + \rho_U\sigma_U \frac{-\phi\left(w_i\delta\right)}{1 - \Phi\left(w_i\delta\right)} .$$

Therefore, the expected markup is:

$$\begin{aligned} E\left(y_i \,|w_i, x_i\right) &= \left(x_i\gamma + \rho_I\sigma_I \frac{\phi\left(w_i\delta\right)}{\Phi\left(w_i\delta\right)}\right) \Phi\left(w_i\delta\right) \\ &+ \left(x_i\beta + \rho_U\sigma_U \frac{-\phi\left(w_i\delta\right)}{1 - \Phi\left(w_i\delta\right)}\right) 1 - \Phi\left(w_i\delta\right) \ . \end{aligned}$$

Green, Hollifield, and Schurhoff [23] estimate this equation as a switching regression. Their results are consistent with two pricing regimes for newly issued municipal bonds. Uninformed investors pay higher average prices than informed investors and there is very little variation in prices paid by informed investors. They also find that the upward trend in prices after issuance is related to the change in the mix if informed and uninformed investors. Informed investors buy soon after issuance, while uninformed investors buy later on.

With respect to the decision about whether to become informed about a new municipal bond issue, they find that large buyers are more likely to become informed than small buyers. To examine how much money is left on the table by uninformed investors who pay large markups to dealers, Green, Hollifield, and Schurhoff [23] classify each transaction into either the Informed or Uninformed regime. The classification is based on whether the expected

benefit from becoming informed about a new bond issue is greater than the cost of doing so:

$$\text{Informed}_i = 1 \Leftrightarrow E\left(z_i^* \,|y_i, w_i, x_i\right) \geq 0,$$
$$\text{Uninformed}_i = 1 \Leftrightarrow E\left(z_i^* \,|y_i, w_i, x_i\right) < 0\,.$$

The difference in the expected markup between an informed investor and an uninformed investor is:

$$E\left(y_{Ui} y_{Ii} \,|\text{Uninformed}_i, w_i, x_i\right) = x_i\left(\beta - \gamma\right) + \left(\rho_U \sigma_U - \rho_I \sigma_I\right) \frac{-\phi\left(w_i \delta\right)}{1 - \Phi\left(w_i \delta\right)}\,.$$

They define the money left on the table in each transaction with an uninformed investor as

$$\Delta_i = \begin{cases} \max\left\{E\left(y_{Ui} y_{Ii} \,|\text{Uninformed}_i, w_i, x_i\right), 0\right\}, & \text{if Uninformed}_i = 1, \\ 0, & \text{else}\,. \end{cases}$$

They denote the estimates of Δ_i as $\widehat{\Delta}_i$ and obtain a cumulative measure across all sales transactions i in a given bond issue j, and then across all issues in a bond deal:

$$\widehat{\text{Money Left on the Table}} = \sum_{\text{Issues}\, j} \sum_{i \in j} \widehat{\Delta}_i.$$

Green, Hollifield, and Schurhoff [23] find that money left on the table by uninformed investors represents a significant fraction of the overall expected profits to the underwriters and dealers.

Bibliography

Primary Literature

1. Alexander G, Edwards A, Ferri M (2000a) The determinants of trading volume of high-yield corporate bonds. J Financ Mark 3:177–204
2. Alexander G, Edwards A, Ferri M (2000b) What does Nasdaq's high-yield bond market reveal about bondholderstockholder conflicts? Financ Manag 29:23–39
3. Amihud Y, Mendelson H (1986) Asset pricing and the bid-ask spread. J Financ Econ 17:223–249
4. Amihud Y, Mendelson H (1991) Liquidity, maturity, and yields on US Treasury securities. J Financ 46:1411–1425
5. Barclay M, Hendershott T, Kotz K (2006) Automation versus intermediation: Evidence from Treasuries going off the run. J Financ 61:2395–2414
6. Bernhardt D, Dvoracek V, Hughson E, Werner I (2005) Why do larger orders receive discounts on the London Stock Exchange? Rev Financ Stud 18:1343–1368
7. Bessembinder H, Maxwell W, Venkataraman K (2006) Optimal market transparency: Evidence from the initiation of trade reporting in corporate bonds. J Financ Econ 82:251–288
8. Biais B, Glosten L, Spatt C (2005) Market microstructure: A survey of microfoundations, empirical results, and policy implications. J Financ Mark 8:217–264
9. Biais B, Green R (2005) The microstructure of the bond market in the 20th century. Working paper, Carnegie Mellon University
10. Blume M, Keim D, Patel S (1991) Returns and volatility of low-grade bonds. J Financ 46:49–74
11. Carlin B (2007) Strategic price complexity in retail financial markets. Working paper, UCLA
12. Chakravarty S, Sarkar A (2003) Trading costs in three US bond markets. J Fixed Income 13:39–48
13. Chen L, Lesmond D, Wei J (2007) Corporate yield spreads and bond liquidity. J Financ 52:119–149
14. Cornell B, Green K (1991) The investment performance of low-grade bond funds. J Financ 46:29–48
15. Collin-Dufresne P, Goldstein R, Martin S (2001) The determinants of credit spread changes. J Financ 41:2177–2207
16. Downing C, Zhang F (2004) Trading activity and price volatility in the municipal bond market. J Financ 59:899–931
17. Edwards A, Harris L, Piwowar M (2007) Corporate Bond Market Transaction Costs and Transparency. J Financ 62:1421–1451
18. Elton E, Green C (1998) Tax and liquidity effects in pricing government bonds. J Financ 53:1533–1562
19. Fenn G (2000) Speed of issuance and the adequacy of disclosure in the 144a high-yield debt market. J Financ Econ 56:383–405
20. Goldstein M, Hotchkiss E, Sirri E (2007) Transparency and liquidity: A controlled experiment on corporate bonds. Rev Financ Stud 20:235–273
21. Green R (2007) Issuers, underwriter syndicates, and aftermarket transparency. Presidential Address to the American Finance Association
22. Green R, Hollifield B, Schurhoff N (2007a) Financial intermediation and the costs of trading in an opaque market. Rev Financ Stud 20:275–314
23. Green R, Hollifield B, Schurhoff N (2007b) Dealer intermediation and price behavior in the aftermarket for new bond issues. J Financ Econ 86:643–682
24. Harris L (2003) Trading and exchanges: market microstructure for practitioners. Oxford University Press, New York
25. Harris L, Piwowar M (2004) Municipal bond liquidity. Working paper, Available at SSRN: http://ssrn.com/abstract=503062
26. Harris L, Piwowar M (2006) Secondary trading costs in the municipal bond market. J Financ 61:1361–1397
27. Hasbrouck J (1993) Assessing the quality of a security market: A new approach to transaction-cost measurement. Rev Financ Stud 6:191–212
28. Helwege J, Turner C (1999) The slope of the credit yield curve for speculative-grade issuers. J Financ 54:1869–1884
29. Hong G, Warga A (2000) An empirical study of bond market transactions. Financ Anal J 56:32–46
30. Hong G, Warga A (2004) Municipal marketability. J Fixed Income 14:86–95
31. Hotchkiss E, Ronen T (2002) The informational efficiency of the corporate bond market: An intraday analysis. Rev Financ Stud 15:1325–1354
32. Hotchkiss E, Warga A, Jostova G (2002) Determinants of corporate bond trading: A comprehensive analysis. Working paper, Boston College

33. Huang R, Stoll H (1997) The components of the bid-ask spread: A general approach. R Financ Stud 10:995–1034
34. Kalimipalli M, Warga A (2002) Bid/ask spread, volatility and volume in the corporate bond market. J Fixed Income 12:31–42
35. Karolyi GA (2004) The world of cross-listings and cross-listings of the world: Challenging conventional wisdom. Working paper, Ohio State University
36. Lesmond D, Ogden J, Trzcinka C (1999) A new estimate of transaction costs. Rev Financ Stud 12:1113–1141
37. Levitt A (1998) The importance of transparency in America's debt market. Remarks at the Media Studies Center, New York, September 9, 1998
38. Madhavan A (2000) Market microstructure: A survey. J Financ Mark 3:205–208
39. Madhavan A, Richardson M, Roomans M (1997) Why do security prices change? A transaction-level analysis of NYSE stocks. R Financ Stud 10:1035–1064
40. Reiss P, Werner I (1996) Transaction costs in dealer markets: Evidence from the London Stock Exchange. In: Lo A (ed) The industrial organization and regulation of the securities industry. University of Chicago Press, Chicago
41. Sarig O, Warga A (1989) Bond price data and bond market liquidity. J Financ Quant Anal 24:367–378
42. Schultz P (2001) Corporate bond trading costs: A peek behind the curtain. J Financ 56:677–698
43. Shilony Y (1977) Mixed pricing in oligopoly. J Econ Theory 14:373–88
44. Warga A (1991) Corporate bond price discrepancies in the dealer and exchange markets. J Fixed Income 1:7–16
45. Warga A (1992) Bond returns, liquidity and missing data. J Financ Quant Anal 27:605–616

Books and Reviews

Greene W (1993) Econometric Analysis, 2nd edn. Macmillan, New York

O'Hara M (1997) Market microstructure theory. Basil Blackwell, Cambridge

Shultz B (1946) The securities market and how it works. Harper, New York

US Securities and Exchange Commission (2004) Report on transactions in municipal securities. http://www.sec.gov/news/studies/munireport2004.pdf

Speeches, Public Statements, and Testimony by US Bond Market Regulators and Participants

September 9, 1998, The Importance of Transparency in America's Debt Market, Remarks of SEC Chairman Arthur Levitt at the Media Studies Center, New York, NY, http://www.sec.gov/news/speech/speecharchive/1998/spch218.htm

April 23, 1999, Remarks of SEC Chairman Arthur Levitt Before the Bond Market Association, San Francisco, CA, http://www.sec.gov/news/speech/speecharchive/1999/spch268.htm

October 28, 1999, Electronic Trading Technology's Impact on the Fixed-Income Markets, Remarks of SEC Commissioner Laura S. Unger at The Bond Market Association, Fifth Annual Legal and Compliance Seminar, New York, NY, http://www.sec.gov/news/speech/speecharchive/1999/spch313.htm

January 8, 2002, Remarks Before the Bond Market Association Legal and Compliance Conference, by Annette L. Nazareth, SEC Director, Division of Market Regulation, http://www.sec.gov/news/speech/spch532.htm

April 25, 2002, Remarks Before the Annual Meeting of the Bond Market Association, by SEC Chairman Harvey L. Pitt, New York, NY, http://www.sec.gov/news/speech/spch553.htm

February 3, 2004, Legal & Compliance Conference – The Bond Market Association, by SEC Commissioner Cynthia A. Glassman, New York, NY, http://www.sec.gov/news/speech/spch020304cag.htm

June 17, 2004, US Senate Committee on Banking, Housing, and Urban Affairs, Hearing on "An Overview of the Regulation of the Bond Markets", http://banking.senate.gov/public

September 9, 2004, Remarks before the 2004 Bond Attorney's Workshop of the National Association of Bond Lawyers, by Martha Mahan Haines, SEC Assistant Director, Office of Municipal Securities, Division of Market Regulation, Chicago, IL, http://www.sec.gov/news/speech/spch090904mmh.htm

October 1, 2004, Keynote Address before the NASD Conference on Fixed Income Markets, by Annette L. Nazareth, SEC Director, Division of Market Regulation, New York City, NY, http://www.sec.gov/news/speech/spch100104aln.htm

December 16, 2004, Second MTS Conference on Financial Markets: "The Organization and Performance of Fixed-Income Markets", by Chester Spatt, SEC Chief Economist and SEC Director, Office of Economic Analysis, Vienna, Austria, http://www.sec.gov/news/speech/spch121604cs.htm

February 1, 2005, Remarks before the Bond Market Association 2005 Legal & Compliance Conference, by SEC Commissioner Paul S. Atkins, New York, NY, http://www.sec.gov/news/speech/spch020105psa.htm

February 2, 2005, Keynote Address at the Bond Market Association 2005 Legal & Compliance Conference, by NASD Vice Chairman Doug Shulman, New York, NY, http://www.finra.org/PressRoom/SpeechesTestimony/DouglasShulman/P013225

April 20, 2005, Remarks before the Bond Market Association, by SEC Chairman William H. Donaldson, New York, NY, http://www.sec.gov/news/speech/spch042005whd.htm

April 20, 2005, Navigating the Changing Climate in Fixed-Income Products, Remarks of NASD President Doug Shulman before the Bond Market Association, New York, NY, http://www.finra.org/PressRoom/SpeechesTestimony/DouglasShulman/P013842

May 6, 2005, Broad Themes in Market Microstructure, by Chester Spatt, SEC Chief Economist and SEC Director, Office of Economic Analysis, Cambridge, MA, http://www.sec.gov/news/speech/spch050605css.htm

June 21, 2005, Developing Bond Markets in APEC: Key Lessons from the US Experience, Remarks of SEC Commissioner Roel C. Campos before the ABAC/ADBI/PECC Conference, Tokyo, Japan, http://www.sec.gov/news/speech/spch062105rcc-2.htm

September 13, 2005, Remarks for Promethee: Transatlantic Dialogue and Regulatory Convergence: Panel on Financial Markets, by Ethiopis Tafara, SEC Director, Office of International Affairs, Paris, France, http://www.sec.gov/news/speech/spch091305et.htm

September 15, 2005, Keynote Address before the 30th Bond Attorney's Workshop of the National Association of Bond Lawyers, by SEC Commissioner Roel C. Campos, Chicago, IL, http://www.sec.gov/news/speech/spch091505rcc.htm

November 17, 2005, Address by NASD President Doug Shulman to the NASD Fall Securities Conference, San Francisco, CA, http://www.finra.org/PressRoom/SpeechesTestimony/DouglasShulman/P015554

January 6, 2006, Discussion: An Overview of Bond Market Transparency, by Chester Spatt, SEC Chief Economist and SEC Director, Office of Economic Analysis, Boston, MA, http://www.sec.gov/news/speech/spch010606css.htm

February 7, 2006, Remarks before the TBMA Legal and Compliance Conference, by SEC Commissioner Annette L. Nazareth, New York, NY, http://www.sec.gov/news/speech/spch020706aln.htm

May 19, 2006, The Continuing Evolution of the Bond Market and Other Observations: Remarks Before the Bond Market Association's 30th Annual Conference, by SEC Commissioner Cynthia A. Glassman, New York, NY, http://www.sec.gov/news/speech/2006/spch051906cag.htm

May 19, 2006, Remarks Before the Bond Market Association's 30th Annual Conference, by NASD President Doug Shulman, New York, NY, http://www.finra.org/PressRoom/SpeechesTestimony/DouglasShulman/P016651

Miscellaneous Sources of Information on International Corporate Bond Markets

April, 2004, Markets in Financial Instruments Directive (MiFID), Directive 2004/39/EC of the European Parliament and of the Council, Article 65, http://europa.eu.int/eur-lex/pri/en/oj/dat/2004/l_145/l_14520040430en00010044.pdf

May 2004, International Organization of Securities Commissions (IOSCO), "Transparency of Corporate Bond Markets", Report of the Technical Committee of IOSCO, http://www.iosco.org/library/pubdocs/pdf/IOSCOPD168.pdf

September 2005, UK Financial Services Authority (FSA), "Trading Transparency in the UK Secondary Bond Markets", FSA Discussion Paper 05/5, http://www.fsa.gov.uk/pubs/discussion/dp05_05.pdf

November, 2005, "Developing Bond Markets in Asia Conference", jointly by the Asian Office of the Bank for International Settlements (BIS) and the People's Bank of China (PBC), Kunming, China, http://www.bis.org/publ/bppdf/bispap26.htm

May 2006, Centre for Economic Policy Research (CEPR), "European corporate bond markets: Transparency, liquidity, efficiency", http://www.cepr.org/PRESS/TT_CorporateFULL.pdf

Correlated Equilibria and Communication in Games

Françoise Forges
Ceremade, Université Paris-Dauphine, Paris, France

Article Outline

Glossary
Definition of the Subject
Introduction
Correlated Equilibrium: Definition and Basic Properties
Correlated Equilibrium and Communication
Correlated Equilibrium in Bayesian Games
Related Topics and Future Directions
Acknowledgment
Bibliography

Glossary

Bayesian game An interactive decision problem consisting of a set of n players, a set of types for every player, a probability distribution which accounts for the players' beliefs over each others' types, a set of actions for every player and a *von Neumann–Morgenstern utility function* defined over n-tuples of types and actions for every player.

Nash equilibrium In an n-person strategic form game, a strategy n-tuple from which unilateral deviations are not profitable.

von Neumann–Morgenstern utility function A *utility function* which reflects the individual's preferences over lotteries. Such a utility function is defined over outcomes and can be extended to any lottery λ by taking expectation with respect to λ.

Pure strategy (or simply strategy) A mapping which, in an interactive decision problem, associates an action with the information of a player whenever this player can make a choice.

Sequential equilibrium A refinement of the *Nash equilibrium* for n-person multistage interactive decision problems, which can be loosely defined as a strategy n-tuple together with beliefs over past information for every player, such that every player maximizes his expected utility given his beliefs and the others' strategies, with the additional condition that the beliefs satisfy (possibly sophisticated) Bayes updating given the strategies.

Strategic (or normal) form game An interactive decision problem consisting of a set of n players, a set of strategies for every player and a (typically, *von Neumann–Morgenstern) utility function* defined over n-tuples of strategies for every player.

Utility function A real valued mapping over a set of outcomes which reflects the preferences of an individual by associating a utility level (a "payoff") with every outcome.

Definition of the Subject

The correlated equilibrium is a game theoretic solution concept. It was proposed by Aumann [1,2] in order to cap-

ture the strategic correlation opportunities that the players face when they take into account the extraneous environment in which they interact. The notion is illustrated in Sect. "Introduction". A formal definition is given in Sect. "Correlated Equilibrium: Definition and Basic Properties". The correlated equilibrium also appears as the appropriate solution concept if preplay communication is allowed between the players. As shown in Sect. "Correlated Equilibrium and Communication", this property can be given several precise statements according to the constraints imposed on the players' communication, which can go from plain conversation to exchange of messages through noisy channels. Originally designed for static games with complete information, the correlated equilibrium applies to any strategic form game. It is geometrically and computationally more tractable than the better known Nash equilibrium. The solution concept has been extended to dynamic games, possibly with incomplete information. As an illustration, we define in details the communication equilibrium for Bayesian games in Sect. "Correlated Equilibrium in Bayesian Games".

Introduction

Example

Consider the two-person game known as "chicken", in which each player i can take a "pacific" action (denoted as p^i) or an "aggressive" action (denoted as a^i):

	p^2	a^2
p^1	$(8, 8)$	$(3, 10)$
a^1	$(10, 3)$	$(0, 0)$

The interpretation is that player 1 and player 2 simultaneously choose an action and then get a payoff, which is determined by the pair of chosen actions according to the previous matrix. If both players are pacific, they both get 8. If both are aggressive, they both get 0. If one player is aggressive and the other is pacific, the aggressive player gets 10 and the pacific one gets 3. This game has two pure Nash equilibria (p^1, a^2), (a^1, p^2) and one mixed Nash equilibrium in which both players choose the pacific action with probability 3/5, resulting in the expected payoff 6 for both players. A possible justification for the latter solution is that the players make their choices as a function of independent extraneous random signals. The assumption of independence is strong. Indeed, there may be no way to prevent the players' signals from being correlated.

Consider a random signal which has no effect on the players' payoffs and takes three possible values: low, medium or high, occurring each with probability 1/3. Assume that, before the beginning of the game, player 1 distinguishes whether the signal is high or not, while player 2 distinguishes whether the signal is low or not. The relevant interactive decision problem is then the extended game in which the players can base their action on the private information they get on the random signal, while the payoffs only depend on the players' actions. In this game, suppose that player 1 chooses the aggressive action when the signal is high and the pacific action otherwise. Similarly, suppose that player 2 chooses the aggressive action when the signal is low and the pacific action otherwise. We show that these strategies form an equilibrium in the extended game. Given player 2's strategy, assume that player 1 observes a high signal. Player 1 deduces that the signal cannot be low so that player 2 chooses the pacific action; hence player 1's best response is to play aggressively. Assume now that player 1 is informed that the signal is not high; he deduces that with probability 1/2, the signal is medium (i. e., not low) so that player 2 plays pacific and with probability 1/2, the signal is low so that player 2 plays aggressive. The expected payoff of player 1 is 5.5 if he plays pacific and 5 if he plays aggressive; hence, the pacific action is a best response. The equilibrium conditions for player 2 are symmetric. To sum up, the strategies based on the players' private information form a Nash equilibrium in the extended game in which an extraneous signal is first selected. We shall say that these strategies form a "correlated equilibrium". The corresponding probability distribution over the players' actions is

$$\begin{array}{ccc} & p^2 & a^2 \\ p^1 & \frac{1}{3} & \frac{1}{3} \\ a^1 & \frac{1}{3} & 0 \end{array} \tag{1}$$

and the expected payoff of every player is 7. This probability distribution can be used directly to make private recommendations to the players before the beginning of the game (see the canonical representation below).

Correlated Equilibrium: Definition and Basic Properties

Definition

A game in strategic form $G = (N, (\Sigma^i)_{i\in N}, (u^i)_{i\in N})$ consists of a set of players N together with, for every player $i \in N$, a set of strategies (for instance, a set of actions) Σ^i and a (von Neumann–Morgenstern) utility function $u^i\colon \Sigma \to \mathbb{R}$, where $\Sigma = \prod_{j\in N} \Sigma^j$ is the set of all strategy profiles. We assume that the sets N and Σ^i, $i \in N$, are finite.

A correlation device $d = (\Omega, \boldsymbol{q}, (\mathcal{P}^i)_{i\in N})$ is described by a finite set of signals Ω, a probability distribution $\boldsymbol{q}$ over

Ω and a partition $\mathcal{P}^i$ of Ω for every player $i \in N$. Since Ω is finite, the probability distribution $\boldsymbol{q}$ is just a real vector $\boldsymbol{q} = (\boldsymbol{q}(\omega))_{\omega\in\Omega}$ such that $\boldsymbol{q}(\omega) \geq 0$ and $\sum_{\omega\in\Omega} \boldsymbol{q}(\omega) = 1$.

From G and d, we define the extended game G_d as follows:

- ω is chosen in Ω according to $\boldsymbol{q}$
- every player i is informed of the element $P^i(\omega)$ of $\mathcal{P}^i$ which contains ω
- G is played: every player i chooses a strategy σ^i in Σ^i and gets the utility $u^i(\sigma)$, $\sigma = (\sigma^j)_{j\in N}$.

A (pure) strategy for player i in G_d is a mapping $\alpha^i\colon \Omega \to \Sigma^i$ which is $\mathcal{P}^i$-measurable, namely, such that $\alpha^i(\omega') = \alpha^i(\omega)$ if $\omega' \in P^i(\omega)$. The interpretation is that, in G_d, every player i chooses his strategy σ^i as a function of his private information on the random signal ω which is selected before the beginning of G.

According to Aumann [1], a *correlated equilibrium* of G is a pair (d, α), which consists of a correlation device $d = (\Omega, \boldsymbol{q}, (\mathcal{P}^i)_{i\in N})$ and a Nash equilibrium $\alpha = (\alpha^i)_{i\in N}$ of G_d. The equilibrium conditions of every player i, conditionally on his private information, can be written as:

$$\begin{aligned}&\sum_{\omega'\in P^i(\omega)} q(\omega'|P^i(\omega))u^i(\alpha(\omega'))\\ &\geq \sum_{\omega'\in P^i(\omega)} q(\omega'|P^i(\omega))u^i(\tau^i, \alpha^{-i}(\omega'))\,,\\ &\qquad \forall i \in N,\ \forall \tau^i \in \Sigma^i,\ \forall \omega \in \Omega\colon q(\omega) > 0\,, \qquad (2)\end{aligned}$$

where $\alpha^{-i} = (\alpha^j)_{j\neq i}$.

A mixed Nash equilibrium $\rho = (\rho^i)_{i\in N}$ of G can be viewed as a correlated equilibrium of G. By definition, every ρ^i is a probability distribution over Σ^i, the finite set of pure strategies of player i. Let us consider the correlation device $d = (\Omega, \boldsymbol{q}, (\mathcal{P}^i)_{i\in N})$ in which $\Omega = \Sigma = \prod_{j\in N} \Sigma^j$, $\boldsymbol{q}$ is the product probability distribution induced by the mixed strategies (i. e., $\boldsymbol{q}((\sigma^j)_{j\in N}) = \prod_{j\in N} \rho^j(\sigma^j)$) and for each i, $\mathcal{P}^i$ is the partition of Ω generated by Σ^i (i. e., for $\omega, \nu \in \Omega$, $\nu \in P^i(\omega) \Leftrightarrow \nu^i = \omega^i$). Let $\alpha^i\colon \Sigma \to \Sigma^i$ be the projection over Σ^i (i. e., $\alpha^i(\sigma) = \sigma^i$). The correlation device d and the strategies α^i defined in this way form a correlated equilibrium. As we shall see below, this correlated equilibrium is "canonical" .

Canonical Representation

A *canonical* correlated equilibrium of G is a correlated equilibrium in which $\Omega = \Sigma = \prod_{j\in N} \Sigma^j$ while for every player i, the partition $\mathcal{P}^i$ of Σ is generated by Σ^i and $\alpha^i\colon \Sigma \to \Sigma^i$ is the projection over Σ^i. A canonical correlated equilibrium is thus fully specified by a probability distribution $\boldsymbol{q}$ over Σ. A natural interpretation is that a mediator selects $\sigma = (\sigma^j)_{j\in N}$ according to $\boldsymbol{q}$ and privately recommends σ^i to player i, for every $i \in N$. The players are not forced to obey the mediator, but σ is selected in such a way that player i cannot benefit from deviating unilaterally from the recommendation σ^i, i. e., $\tau^i = \sigma^i$ maximizes the conditional expectation of player i's payoff $u^i(\tau^i, \sigma^{-i})$ given the recommendation σ^i. A probability distribution $\boldsymbol{q}$ over Σ thus defines a canonical correlated equilibrium if and only if it satisfies the following linear inequalities:

$$\begin{aligned}&\sum_{\sigma^{-i}\in\Sigma^{-i}} \boldsymbol{q}(\sigma^{-i}|\sigma^i)u^i(\sigma^i, \sigma^{-i})\\ &\geq \sum_{\sigma^{-i}\in\Sigma^{-i}} \boldsymbol{q}(\sigma^{-i}|\sigma^i)u^i(\tau^i, \sigma^{-i})\,,\\ &\qquad \forall i \in N,\ \forall \sigma^i \in \Sigma^i\colon \boldsymbol{q}(\sigma^i) > 0,\ \forall \tau^i \in \Sigma^i\end{aligned}$$

or, equivalently,

$$\begin{aligned}&\sum_{\sigma^{-i}\in\Sigma^{-i}} \boldsymbol{q}(\sigma^i, \sigma^{-i})u^i(\sigma^i, \sigma^{-i})\\ &\geq \sum_{\sigma^{-i}\in\Sigma^{-i}} \boldsymbol{q}(\sigma^i, \sigma^{-i})u^i(\tau^i, \sigma^{-i})\,,\\ &\qquad \forall i \in N,\ \forall \sigma^i,\ \tau^i \in \Sigma^i \qquad (3)\end{aligned}$$

The equilibrium conditions can also be formulated *ex ante*:

$$\begin{aligned}&\sum_{\sigma\in\Sigma} \boldsymbol{q}(\sigma)u^i(\sigma) \geq \sum_{\sigma\in\Sigma} \boldsymbol{q}(\sigma)u^i(\alpha^i(\sigma^i), \sigma^{-i})\,,\\ &\qquad \forall i \in N,\ \forall \alpha^i\colon \Sigma^i \to \Sigma^i\end{aligned}$$

The following result is an analog of the "revelation principle" in mechanism design (see, e. g., Myerson [41]): *let (α, d) be a correlated equilibrium associated with an arbitrary correlation device $d = (\Omega, \boldsymbol{q}, (\mathcal{P}^i)_{i\in N})$. The corresponding "correlated equilibrium distribution", namely, the probability distribution induced over Σ by $\boldsymbol{q}$ and α, defines a canonical correlated equilibrium.* For instance, in the introduction, (1) describes a canonical correlated equilibrium.

Duality and Existence

From the linearity of (3), duality theory can be used to study the properties of correlated equilibria, in particular to prove their existence without relying on Nash's [45] theorem and its fixed point argument (recall that every

mixed Nash equilibrium is a correlated equilibrium). Hart and Schmeidler [30] establish the existence of a correlated equilibrium by constructing an auxiliary two person zero-sum game and applying the minimax theorem. Nau and McCardle [47] derive another elementary proof of existence from an extension of the "no arbitrage opportunities" axiom that underlies subjective probability theory. They introduce jointly coherent strategy profiles, which do not expose the players as a group to arbitrage from an outside observer. They show that a strategy profile is jointly coherent if and only if it occurs with positive probability in some correlated equilibrium. From a technical point of view, both proofs turn out to be similar. Myerson [44] makes further use of the linear structure of correlated equilibria by introducing dual reduction, a technique to replace a finite game with a game with fewer strategies, in such a way that any correlated equilibrium of the reduced game induces a correlated equilibrium of the original game.

Geometric Properties

As (3) is a system of linear inequalities, the set of all correlated equilibrium distributions is a convex polytope. Nau et al. [49] show that if it has "full" dimension (namely, dimension $|\Sigma| - 1$), then all Nash equilibria lie on its relative boundary. Viossat [60] characterizes in addition the class of games whose correlated equilibrium polytope contains a Nash equilibrium in its relative interior. Interestingly, this class of games includes two person zero-sum games but is not defined by "strict competition" properties. In two person games, all extreme Nash equilibria are also extreme correlated equilibria [13,25]; this result does not hold with more than two players. Finally, Viossat [59] proves that having a unique correlated equilibrium is a robust property, in the sense that the set of n person games with a unique correlated equilibrium is open. The same is not true for Nash equilibrium (unless $n = 2$).

Complexity

From (3), correlated equilibria can be computed by linear programming methods. Gilboa and Zemel [24] show more precisely that the complexity of standard computational problems is "NP-hard" for the Nash equilibrium and *polynomial* for the correlated equilibrium. Examples of such problems are: "Does the game G have a Nash (resp., correlated) equilibrium which yields a payoff greater than r to every player (for some given number r)?" and "Does the game G have a unique Nash (resp., correlated) equilibrium?". Papadimitriou [50] develops a polynomial-time algorithm for finding correlated equilibria, which is based on a variant of the existence proof of Hart and Schmeidler [30].

Foundations

By re-interpreting the previous canonical representation, Aumann [2] proposes a decision theoretic foundation for the correlated equilibrium in games with complete information, in which Σ^i for $i \in N$, stands merely for a set of *actions* of player i. Let Ω be the space of all states of the world; an element ω of Ω thus specifies all the parameters which may be relevant to the players' choices. In particular, the action profile in the underlying game G is part of the state of the world. A partition $\mathcal{P}^i$ describes player i's information on Ω. In addition, every player i has a prior belief, i. e., a probability distribution, $\boldsymbol{q}^i$ over Ω. Formally, the framework is similar as above except that the players possibly hold different beliefs over Ω. Let $\alpha^i(\omega)$ denote player i's action at ω; a natural assumption is that player i knows the action he chooses, namely that α^i is $\mathcal{P}^i$ -measurable. According to Aumann [2], player i is *Bayes-rational* at ω if his action $\alpha^i(\omega)$ maximizes his expected payoff (with respect to $\boldsymbol{q}^i$) given his information $P^i(\omega)$. Note that this is a separate rationality condition for every player, not an equilibrium condition. Aumann [2] proves the following result: *Under the common prior assumption (namely, $\boldsymbol{q}^i = \boldsymbol{q}$, $i \in N$), if every player is Bayes-rational at every state of the world, the distribution of the corresponding action profile α is a correlated equilibrium distribution.* The key to this decision theoretic foundation of the correlated equilibrium is that, under the common prior assumption, Bayesian rationality amounts to (2).

If the common prior assumption is relaxed, the previous result still holds, with *subjective* prior probability distributions, for the *subjective* correlated equilibrium which was also introduced by Aumann [1]. The latter solution concept is defined in the same way as above, by considering a device $(\Omega, (\boldsymbol{q}^i)_{i\in N}, (\mathcal{P}^i)_{i\in N})$, with a probability distribution $\boldsymbol{q}^i$ for every player i, and by writing (2) in terms of $\boldsymbol{q}^i$ instead of $\boldsymbol{q}$. Brandenburger and Dekel [10] show that (a refinement of) the subjective correlated equilibrium is equivalent to (correlated) rationalizability, another well-established solution concept which captures players' minimal rationality. Rationalizable strategies reflect that the players commonly know that each of them makes an optimal choice given some belief. Nau and McCardle [48] reconcile objective and subjective correlated equilibrium by proposing the no arbitrage principle as a unified approach to individual and interactive decision problems. They argue that the objective correlated equilibrium concept applies to a game that is revealed by the players' choices,

while the subjective correlated equilibrium concept applies to the "true game"; both lead to the same set of jointly coherent outcomes.

Correlated Equilibrium and Communication

As seen in the previous section, correlated equilibria can be achieved in practice with the help of a mediator and emerge in a Bayesian framework embedding the game in a full description of the world. Both approaches require to extend the game by taking into account information which is not generated by the players themselves. Can the players reach a correlated equilibrium without relying on any extraneous correlation device, by just communicating with each other before the beginning of the game?

Consider the game of "chicken" presented in the introduction. The probability distribution

$$\begin{array}{ccc} & p^2 & a^2 \\ p^1 & 0 & \frac{1}{2} \\ a^1 & \frac{1}{2} & 0 \end{array} \tag{4}$$

describes a correlated equilibrium, which amounts to choosing one of the two pure Nash equilibria, with equal probability. Both players get an expected payoff of 6.5. Can they safely achieve this probability distribution if no mediator tosses a fair coin for them? The answer is positive, as shown by Aumann et al. [3]. Assume that before playing "chicken", the players independently toss a coin and simultaneously reveal to each other whether heads or tails obtains. Player 1 tells player 2 "h^1" or "t^1" and, at the same time, player 2 tells player 1 "h^2" or "t^2". If both players use a fair coin, reveal correctly the result of the toss and play (p^1, a^2) if both coins fell on the same side (i. e., if (h^1, h^2) or (t^1, t^2) is announced) and (a^1, p^2) otherwise (i. e., if (h^1, t^2) or (t^1, h^2) is announced), they get the same effect as a mediator using (4). Furthermore, none of them can gain by unilaterally deviating from the described strategies, even at the randomizing stage: the two relevant outcomes, $[(h^1, h^2)$ or $(t^1, t^2)]$ and $[(h^1, t^2)$ or $(t^1, h^2)]$, happen with probability 1/2 provided that one of the players reveals the toss of a fair coin. This procedure is known as a "jointly controlled lottery" . An important feature of the previous example is that, in the correlated equilibrium described by (4), the players know each other's recommendation. Hence, they can easily reproduce (4) by exchanging messages that they have selected independently. In the correlated equilibrium described by the probability distribution (1), the private character of recommendations is crucial to guarantee that (p^1, p^2) be played with positive probability. Hence one cannot hope that a simple procedure of direct preplay communication be sufficient to generate (1). However, the fact that direct communication is necessarily public is typical of two-person games.

Given the game $G = (N, (\Sigma^i)_{i\in N}, (u^i)_{i\in N})$, let us define a (bounded) "cheap talk" extension ext(G) of G as a game in which T stages of costless, unmediated preplay communication are allowed before G is played. More precisely, let M_t^i be a finite set of messages for player i, $i \in N$, at stage t, $t = 1, 2, \ldots T$; at every stage t of ext(G), every player i selects a message in M_t^i; these choices are made simultaneously before being revealed to a subset of players at the end of stage t. The rules of $ext(G)$ thus determine a set of "senders" for every stage t (those players i for whom M_t^i contains more than one message) and a set of "receivers" for every stage t. The players perfectly recall their past messages. After the communication phase, they choose their strategies (e. g., their actions) as in G; they are also rewarded as in G, independently of the preplay phase, which is thus "cheap" . Communication has an indirect effect on the final outcome in G, since the players make their decisions as a function of the messages that they have exchanged. Specific additional assumptions are often made on ext(G), as we will see below.

Let us fix a cheap talk extension ext(G) of G and a Nash equilibrium of $ext(G)$. As a consequence of the previous definitions, the distribution induced by this Nash equilibrium over Σ defines a correlated equilibrium of G (this can be proved in the same way as the canonical representation of correlated equilibria stated in Sect. "Correlated Equilibrium: Definition and Basic Properties"). The question raised in this section is whether the reverse holds.

If the number of players is two, the Nash equilibrium distributions of cheap talk extensions of G form a subset of the correlated equilibrium distributions: the convex hull of Nash equilibrium distributions. Indeed, the players have both the same information after any direct exchange of messages. Conversely, by performing repeated jointly controlled lotteries like in the example above, the players can achieve any convex combination (with rational weights) of Nash equilibria of G as a Nash equilibrium of a cheap talk extension of G. The restriction on probability distributions whose components are rational numbers is only needed as far as we focus on *bounded* cheap talk extensions.

Bárány [4] establishes that, if the number of players of G is at least four, every (rational) correlated equilibrium distribution of G can be realized as a Nash equilibrium of a cheap talk extension ext(G), provided that ext(G) allows the players to publicly check the record of communication under some circumstances. The equilibria of ext(G) constructed by Bárány involve that a receiver gets the same

message from two different senders; the message is nevertheless not public thanks to the assumption on the number of players. At every stage of ext(G), every player can ask for the revelation of all past messages, which are assumed to be recorded. Typically, a receiver can claim that the two senders' messages differ. In this case, the record of communication surely reveals that either one of the senders or the receiver himself has cheated; the deviator can be punished (at his minmax level in G) by the other players.

The punishments in Bárány's [4] Nash equilibria of ext(G) need not be credible threats. Instead of using double senders in the communication protocols, Ben-Porath [5,6] proposes a procedure of random monitoring, which prescribes a given behavior to every player in such a way that unilateral deviations can be detected with probability arbitrarily close to 1. This procedure applies if there are at least three players, which yields an analog of Bárány's result already in this case. If the number of players is exactly three, Ben-Porath [6] needs to assumes, as Bárány [4], that public verification of the record of communication is possible in ext(G) (see Ben-Porath [7]). However, Ben-Porath concentrates on (rational) correlated equilibrium distributions which allow for strict punishment on a Nash equilibrium of G; he constructs *sequential* equilibria which generate these distributions in ext(G), thus dispensing with incredible threats. At the price of raising the number of players to five or more, Gerardi [22] proves that every (rational) correlated equilibrium distribution of G can be realized as a sequential equilibrium of a cheap talk extension of G which does not require any message recording. For this, he builds protocols of communication in which the players base their decisions on majority rule, so that no punishment is necessary.

We have concentrated on two extreme forms of communication: mediated communication, in which a mediator performs lotteries and sends private messages to the players and cheap talk, in which the players just exchange messages. Many intermediate schemes of communication are obviously conceivable. For instance, Lehrer [36] introduces (possibly multistage) "mediated talk": the players send private messages to a mediator, but the latter can only make deterministic public announcements. Mediated talk captures real-life communication procedures, like elections, especially if it lasts only for a few stages. Lehrer and Sorin [37] establish that whatever the number of players of G, every (rational) correlated equilibrium distribution of G can be realized as a Nash equilibrium of a single stage mediated talk extension of G. Ben-Porath [5] proposes a variant of cheap talk in which the players do not only exchange verbal messages but also "hard" devices such as urns containing balls. This extension is particularly useful in two-person games to circumvent the equivalence between the equilibria achieved by cheap talk and the convex hull of Nash equilibria. More precisely, the result of Ben-Porath [5] stated above holds for two-person games if the players first check together the content of different urns, and then each player draws a ball from an urn that was chosen by the other player, so as to guarantee that one player only knows the outcome of a lottery while the other one only knows the probabilities of this lottery.

The various extensions of the basic game G considered up to now, with or without a mediator, implicitly assume that the players are fully rational. In particular, they have unlimited computational abilities. By relaxing that assumption, Urbano and Vila [55] and Dodis et al. [12] build on earlier results from cryptography so as to implement any (rational) correlated equilibrium distribution through unmediated communication, including in two-person games.

As the previous paragraphs illustrate, the players can modify their intitial distribution of information by means of many different communication protocols. Gossner [26] proposes a general criterion to classify them: a protocol is "secure" if under all circumstances, the players cannot mislead each other nor spy on each other. For instance, given a cheap talk extension ext(G), a protocol P describes, for every player, a strategy in ext(G) and a way to interpret his information after the communication phase of ext(G). P induces a correlation device $d(P)$ (in the sense of Sect. "Correlated Equilibrium: Definition and Basic Properties"). P is secure if, *for every game* G and every Nash equilibrium α of $G_{d(P)}$, the following procedure is a Nash equilibrium of ext(G): communicate according to the strategies described by P in order to generate $d(P)$ and make the final choice, in G, according to α. Gossner [26] gives a tractable characterization of secure protocols.

Correlated Equilibrium in Bayesian Games

A Bayesian game $\Gamma = (N, (T^i)_{i\in N}, p, (A^i)_{i\in N}, (v^i)_{i\in N})$ consists of: a set of players N; for every player $i \in N$, a set of types T^i, a probability distribution p^i over $T = \prod_{j\in N} T^j$, a set of actions A^i and a (von Neumann–Morgenstern) utility function $v^i : T \times A \to \mathbb{R}$, where $A = \prod_{j\in N} A^j$. For simplicity, we make the common prior assumption: $p^i = p$ for every $i \in N$. All sets are assumed finite. The interpretation is that a virtual move of nature chooses $t = (t^j)_{j\in N}$ according to p; player i is only informed of his own type t^i; the players then choose simultaneously an action. We will focus on two possible extensions of Aumann's [1] solution concept to Bayesian

games: the strategic form correlated equilibrium and the communication equilibrium. Without loss of generality, the definitions below are given in "canonical form" (see Sect. "Correlated Equilibrium: Definition and Basic Properties").

Strategic Form Correlated Equilibrium

A (pure) strategy of player i in Γ is a mapping $\sigma^i : T^i \to A^i$, $i \in N$. The strategic form of Γ is a game $G(\Gamma)$, like the game G considered in Sect. "Correlated Equilibrium: Definition and Basic Properties", with sets of pure strategies $\Sigma_i = A_i^{T_i}$ and utility functions u^i over $\Sigma = \prod_{j\in N} \Sigma^j$ computed as expectations with respect to p: $u^i(\sigma) = E[v^i(t, \sigma(t))]$, with $\sigma(t) = (\sigma^i(t^i))_{i\in N}$. A strategic form correlated equilibrium, or simply, a correlated equilibrium, of a Bayesian game Γ is a correlated equilibrium, in the sense of Sect. "Correlated Equilibrium: Definition and Basic Properties", of $G(\Gamma)$. A canonical correlated equilibrium of Γ is thus described by a probability distribution Q over Σ, which selects an N-tuple of pure strategies $(\sigma^i)_{i\in N}$. This lottery can be thought of as being performed by a mediator who privately recommends σ^i to player i, $i \in N$, before the beginning of Γ, i. e., before (or in any case, independently of) the chance move choosing the N-tuple of types. The equilibrium conditions express that, once he knows his type t^i, player i cannot gain in unilaterally deviating from $\sigma^i(t^i)$.

Communication Equilibrium

Myerson [41] transforms the Bayesian game Γ into a mechanism design problem by allowing the mediator to collect information from the players before making them recommendations. Following Forges [15] and Myerson [42], a canonical communication device for Γ consists of a system $\boldsymbol{q}$ of probability distributions $\boldsymbol{q} = (\boldsymbol{q}(.|t))_{t\in T}$ over A. The interpretation is that a mediator invites every player i, $i \in N$, to report his type t^i, then selects an N-tuple of actions a according to $\boldsymbol{q}(.|t)$ and privately recommends a^i to player i. The system $\boldsymbol{q}$ defines a communication equilibrium if none of the players can gain by unilaterally lying on his type or by deviating from the recommended action, namely if

$$\sum_{t^{-i}\in T^{-i}} p(t^{-i}|t^i) \sum_{a\in A} q(a|t) v^i(t, a)$$
$$\geq \sum_{t^{-i}\in T^{-i}} p(t^{-i}|t^i) \sum_{a\in A} q(a|s^i, t^{-i}) v^i(t, \alpha^i(a^i), a^{-i}),$$
$$\forall i \in N,\ \forall t^i,\ s^i \in T^i,\ \forall \alpha^i : A^i \to A^i$$

Correlated Equilibrium, Communication Equilibrium and Cheap Talk

Every correlated equilibrium of the Bayesian game Γ induces a communication equilibrium of Γ, but the converse is not true, as the following example shows.

Consider the two-person Bayesian game in which $T^1 = \{s^1, t^1\}$, $T^2 = \{t^2\}$, $A^1 = \{a^1, b^1\}$, $A^2 = \{a^2, b^2\}$, $p(s^1) = p(t^1) = \frac{1}{2}$ and payoffs are described by

		a^2	b^2
s^1	a^1	$(1, 1)$	$(-1, -1)$
	b^1	$(0, 0)$	$(0, 0)$

		a^2	b^2
t^1	a^1	$(0, 0)$	$(0, 0)$
	b^1	$(-1, -1)$	$(1, 1)$

In this game, the communication equilibrium $\boldsymbol{q}(a^1, a^2|s^1) = \boldsymbol{q}(b^1, b^2|t^1) = 1$ yields the expected payoff of 1 to both players. However the maximal expected payoff of every player in a correlated equilibrium is 1/2. In order to see this, one can derive the strategic form of the game (in which player 1 has four strategies and player 2 has two strategies). Let us turn to the game in which player 1 can cheaply talk to player 2 just after having learned his type. In this new game, the following strategies form a Nash equilibrium: player 1 truthfully reveals his type to player 2 and plays a^1 if s^1, b^1 if t^1; player 2 chooses a^2 if s^1, b^2 if t^1. These strategies achieve the same expected payoffs as the communication equilibrium.

As in Sect. "Correlated Equilibrium and Communication", one can define cheap talk extensions ext(Γ) of Γ. A wide definition of ext(Γ) involves an *ex ante* preplay phase, before the players learn their types, and an *interim* preplay phase, after the players learn their types but before they choose their actions. Every Nash equilibrium of ext(Γ) induces a communication equilibrium of Γ. In order to investigate the converse, namely whether cheap talk can simulate mediated communication in a Bayesian game, two approaches have been developed. The first one (Forges [17], Gerardi [21,22], Vida [58]) proceeds in two steps, by reducing communication equilibria to correlated equilibria before applying the results obtained for strategic form games (see Sect. "Correlated Equilibrium and Communication"). The second approach (Ben-Porath [6], Krishna [33]) directly addresses the question in a Bayesian game.

By developing a construction introduced for particular two person games (Forges [14]), Forges [17] shows that every communication equilibrium outcome of a Bayesian game Γ with at least four players can be achieved as a correlated equilibrium outcome of a two stage *interim* cheap

talk extension $\mathrm{ext}_{\mathrm{int}}(\Gamma)$ of Γ. No punishment is necessary in $\mathrm{ext}_{\mathrm{int}}(\Gamma)$: at the second stage, every player gets a message from three senders and uses majority rule if the messages are not identical. Thanks to the underlying correlation device, each receiver is able to privately decode his message. Vida [58] extends Forges [17] to Bayesian games with three or even two players. In the proof, he constructs a correlated equilibrium of a long, but almost surely finite, *interim* cheap talk extension of Γ, whose length depends both on the signals selected by the correlation device and the messages exchanged by the players. No recording of messages is necessary to detect and punish a cheating player. If there are at least four players in Γ, once a communication equilibrium of Γ has been converted into a correlated equilibrium of $\mathrm{ext}_{\mathrm{int}}(\Gamma)$, one can apply Bárány's [4] result to $\mathrm{ext}_{\mathrm{int}}(\Gamma)$ in order to transform the correlated equilibrium into a Nash equilibrium of a further, *ex ante*, cheap talk preplay extension of Γ. Gerardi [21] modifies this ex ante preplay phase so as to postpone it at the *interim* stage. This result is especially useful if the initial move of nature in Γ is just a modelling convenience. Gerardi [22] also extends his result for at least five person games with complete information (see Sect. "Correlated Equilibrium and Communication") to any Bayesian game with full support (i. e., in which all type profiles have positive probability: $p(t) > 0$ for every $t \in T$) by proving that every (rational) communication equilibrium of Γ can be achieved as a *sequential* equilibrium of a cheap talk extension of Γ.

Ben-Porath [6] establishes that if Γ is a three (or more) person game with full support, every (rational) communication equilibrium of Γ which strictly dominates a Nash equilibrium of Γ for every type t^i of every player i, $i \in N$, can be implemented as a Nash equilibrium of an *interim* cheap talk extension of Γ in which public verification of past record is possible (see also Ben-Porath [7]). Krishna [33] extends Ben-Porath's [5] result on two person games (see Sect. "Correlated Equilibrium and Communication") to the incomplete information framework. The other results mentioned at the end of Sect. "Correlated Equilibrium and Communication" have also been generalized to Bayesian games (see [26,37,56]).

Related Topics and Future Directions

In this brief article, we concentrated on two solution concepts: the strategic form correlated equilibrium, which is applicable to any game, and the communication equilibrium, which we defined for Bayesian games. Other extensions of Aumann's [1]solution concept have been proposed for Bayesian games, as the agent normal form correlated equilibrium and the (possibly belief invariant) Bayesian solution (see Forges [18,19] for definitions and references). The Bayesian solution is intended to capture the players' rationality in games with incomplete information in the spirit of Aumann [2] (see Nau [46] and Forges [18]). Lehrer et al. [38] open a new perspective in the understanding of the Bayesian solution and other equilibrium concepts for Bayesian games by characterizing the classes of equivalent information structures with respect to each of them. Comparison of information structures, which goes back to Blackwell [8,9] for individual decision problems, was introduced by Gossner [27] in the context of games, both with complete and incomplete information. In the latter model, information structures basically describe how extraneous signals are selected as a function of the players' types; two information structures are equivalent with respect to an equilibrium concept if, in every game, they generate the same equilibrium distributions over outcomes.

Correlated equilibria, communication equilibria and related solution concepts have been studied in many other classes of games, like multistage games (see, e. g., [15,42]), repeated games with incomplete information (see, e. g., [14,16]) and stochastic games (see, e. g., [53,54]). The study of correlated equilibrium in repeated games with imperfect monitoring, initiated by Lehrer [34,35], proved to be particularly useful and is still undergoing. Lehrer [34] showed that if players are either fully informed of past actions or get no information (" standard-trivial" information structure), correlated equilibria are equivalent to Nash equilibria. In other words, all correlations can be generated internally, namely by the past histories, on which players have differential information. The schemes of internal correlation introduced to establish this result are widely applicable and inspired those of Lehrer [36] (see Sect. "Correlated Equilibrium and Communication"). In general repeated games with imperfect monitoring, Renault and Tomala [52] characterize communication equilibria but the amount of correlation that the players can achieve in a Nash equilibrium is still an open problem (see, e. g., [28,57] for recent advances).

Throughout this article, we defined a correlated equilibrium as a *Nash* equilibrium of an extension of the game under consideration. The solution concept can be strengthened by imposing some refinement, i. e., further rationality conditions, to the Nash equilibrium in this definition (see, e. g., [11,43]). Refinements of communication equilibria have also been proposed (see, e. g., [22,23,42]). Some authors (see, e. g., [39,40,51]) have also developed notions of *coalition proof* correlated equilibria, which resist not only to unilateral deviations, as in this article, but

even to multilateral ones. A recurrent difficulty is that, for many of these stronger solution concepts, a useful canonical representation (as derived in Sect. "Correlated Equilibrium: Definition and Basic Properties") is not available.

Except for two or three references, we deliberately concentrated on the results published in the game theory and mathematical economics literature, while substantial achievements in computer science would fit in this survey. Both streams of research pursue similar goals but rely on different formalisms and techniques. For instance, computer scientists often make use of cryptographic tools which are not familiar in game theory. Halpern [29] gives an idea of recent developments at the interface of computer science and game theory (see in particular the section "implementing mediators") and contains a number of references.

Finally, the assumption of full rationality of the players can also be relaxed. Evolutionary game theory has developed models of learning in order to study the long term behavior of players with bounded rationality. Many possible dynamics are conceivable to represent more or less myopic attitudes with respect to optimization. Under appropriate learning procedures, which express for instance that agents want to minimize the regret of their strategic choices, the empirical distribution of actions converge to correlated equilibrium distributions (see, e. g., [20,31,32] for a survey). However, standard procedures, as the "replicator dynamics", may even eliminate all the strategies which have positive probability in a correlated equilibrium (see [61]).

Acknowledgment

The author wishes to thank Elchanan Ben-Porath, Frédéric Koessler, R. Vijay Krishna, Ehud Lehrer, Bob Nau, Indra Ray, Jérôme Renault, Eilon Solan, Sylvain Sorin, Bernhard von Stengel, Tristan Tomala, Amparo Urbano, Yannick Viossat and, especially, Olivier Gossner and Péter Vida, for useful comments and suggestions.

Bibliography

Primary Literature

1. Aumann RJ (1974) Subjectivity and correlation in randomized strategies. J Math Econ 1:67–96
2. Aumann RJ (1987) Correlated equilibrium as an expression of Bayesian rationality. Econometrica 55:1–18
3. Aumann RJ, Maschler M, Stearns R (1968) Repeated games with incomplete information: an approach to the nonzero sum case. Reports to the US Arms Control and Disarmament Agency, ST-143, Chapter IV, 117–216 (reprinted In: Aumann RJ, Maschler M (1995) Repeated Games of Incomplete Information. M.I.T. Press, Cambridge)
4. Bárány I (1992) Fair distribution protocols or how players replace fortune. Math Oper Res 17:327–340
5. Ben-Porath E (1998) Correlation without mediation: expanding the set of equilibrium outcomes by cheap pre-play procedures. J Econ Theor 80:108–122
6. Ben-Porath E (2003) Cheap talk in games with incomplete information. J Econ Theor 108:45–71
7. Ben-Porath E (2006) A correction to "Cheap talk in games with incomplete information". Mimeo, Hebrew University of Jerusalem, Jerusalem
8. Blackwell D (1951) Comparison of experiments. Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley and Los Angeles, pp 93–102
9. Blackwell D (1953) Equivalent comparison of experiments. Ann Math Stat 24:265–272
10. Brandenburger A, Dekel E (1987) Rationalizability and correlated equilibria. Econometrica 55:1391–1402
11. Dhillon A, Mertens JF (1996) Perfect correlated equilibria. J Econ Theor 68:279–302
12. Dodis Y, Halevi S, Rabin T (2000) A cryptographic solution to a game theoretic problem. CRYPTO 2000: 20th International Cryptology Conference. Springer, Berlin, pp 112–130
13. Evangelista F, Raghavan TES (1996) A note on correlated equilibrium. Int J Game Theory 25:35–41
14. Forges F (1985) Correlated equilibria in a class of repeated games with incomplete information. Int J Game Theory 14:129–150
15. Forges F (1986) An approach to communication equilibrium. Econometrica 54:1375–1385
16. Forges F (1988) Communication equilibria in repeated games with incomplete information. Math Oper Res 13:191–231
17. Forges F (1990) Universal mechanisms. Econometrica 58:1341–1364
18. Forges F (1993) Five legitimate definitions of correlated equilibrium in games with incomplete information. Theor Decis 35:277–310
19. Forges F (2006) Correlated equilibrium in games with incomplete information revisited. Theor Decis 61:329–344
20. Foster D, Vohra R (1997) Calibrated learning and correlated equilibrium. Game Econ Behav 21:40–55
21. Gerardi D (2000) Interim pre-play communication. Mimeo, Yale University, New Haven
22. Gerardi D (2004) Unmediated communication in games with complete and incomplete information. J Econ Theory 114:104–131
23. Gerardi D, Myerson R (2007) Sequential equilibria in Bayesian games with communication. Game Econ Behav 60:104–134
24. Gilboa I, Zemel E (1989) Nash and correlated equilibria: some complexity considerations. Game Econ Behav 1:80–93
25. Gomez-Canovas S, Hansen P, Jaumard B (1999) Nash Equilibria from the correlated equilibria viewpoint. Int Game Theor Rev 1:33–44
26. Gossner O (1998) Secure protocols or how communication generates correlation. J Econ Theory 83:69–89
27. Gossner O (2000) Comparison of information structures. Game Econ Behav 30:44–63
28. Gossner O, Tomala T (2007) Secret correlation in repeated games with signals. Math Oper Res 32:413–424
29. Halpern JY (2007) Computer science and game theory. In: Durlauf SN, Blume LE (eds) The New Palgrave dictio-

nary of economics, 2nd edn. Palgrave Macmillan. The New Palgrave dictionary of economics online. http://www.dictionaryofeconomics.com/article?id=pde2008_C000566. Accessed 24 May 2008
30. Hart S, Schmeidler D (1989) Existence of correlated equilibria. Math Oper Res 14:18–25
31. Hart S, Mas-Colell A (2000) A simple adaptive procedure leading to correlated equilibrium. Econometrica 68:1127–1150
32. Hart S (2005) Adaptative heuristics. Econometrica 73:1401–1430
33. Krishna RV (2007) Communication in games of incomplete information: two players. J Econ Theory 132:584–592
34. Lehrer E (1991) Internal correlation in repeated games. Int J Game Theory 19:431–456
35. Lehrer E (1992) Correlated equilibria in two-player repeated games with non-observable actions. Math Oper Res 17:175–199
36. Lehrer E (1996) Mediated talk. Int J Game Theory 25:177–188
37. Lehrer E, Sorin S (1997) One-shot public mediated talk. Game Econ Behav 20:131–148
38. Lehrer E, Rosenberg D, Shmaya E (2006) Signaling and mediation in Bayesian games. Mimeo, Tel Aviv University, Tel Aviv
39. Milgrom P, Roberts J (1996) Coalition-proofness and correlation with arbitrary communication possibilities. Game Econ Behav 17:113–128
40. Moreno D, Wooders J (1996) Coalition-proof equilibrium. Games Econ Behav 17:80–112
41. Myerson R (1982) Optimal coordination mechanisms in generalized principal-agent problems. J Math Econ 10:67–81
42. Myerson R (1986a) Multistage games with communication. Econometrica 54:323–358
43. Myerson R (1986b) Acceptable and predominant correlated equilibria. Int J Game Theory 15:133–154
44. Myerson R (1997) Dual reduction and elementary games. Game Econ Behav 21:183–202
45. Nash J (1951) Non-cooperative games. Ann Math 54:286–295
46. Nau RF (1992) Joint coherence in games with incomplete information. Manage Sci 38:374–387
47. Nau RF, McCardle KF (1990) Coherent behavior in noncooperative games. J Econ Theory 50(2):424–444
48. Nau RF, McCardle KF (1991) Arbitrage, rationality and equilibrium. Theor Decis 31:199–240
49. Nau RF, Gomez-Canovas S, Hansen P (2004) On the geometry of Nash equilibria and correlated equilibria. Int J Game Theory 32:443–453
50. Papadimitriou CH (2005) Computing correlated equilibria in multiplayer games. Proceedings of the 37th ACM Symposium on Theory of Computing. STOC, Baltimore, pp 49–56
51. Ray I (1996) Coalition-proof correlated equilibrium: a definition. Game Econ Behav 17:56–79
52. Renault J, Tomala T (2004) Communication equilibrium payoffs in repeated games with imperfect monitoring. Game Econ Behav 49:313–344
53. Solan E (2001) Characterization of correlated equilibrium in stochastic games. Int J Game Theory 30:259–277
54. Solan E, Vieille N (2002) Correlated equilibrium in stochastic games. Game Econ Behav 38:362–399
55. Urbano A, Vila J (2002) Computational complexity and communication: coordination in two-player games. Econometrica 70:1893–1927
56. Urbano A, Vila J (2004a) Computationally restricted unmediated talk under incomplete information. J Econ Theory 23:283–320
57. Urbano A, Vila J (2004b) Unmediated communication in repeated games with imperfect monitoring. Game Econ Behav 46:143–173
58. Vida P (2007) From communication equilibria to correlated equilibria. Mimeo, University of Vienna, Vienna
59. Viossat Y (2005) Is having a unique equilibrium robust? to appear in J Math Econ
60. Viossat Y (2006) The geometry of Nash equilibria and correlated equilibria and a generalization of zero-sum games. Mimeo, S-WoPEc working paper 641, Stockholm School of Economics, Stockholm
61. Viossat Y (2007) The replicator dynamics does not lead to correlated equilibria. Game Econ Behav 59:397–407

Books and Reviews

Forges F (1994) Non-zero sum repeated games and information transmission. In: Megiddo N (ed) Essays in Game Theory in Honor of Michael Maschler. Springer, Berlin, pp 65–95

Mertens JF (1994) Correlated- and communication equilibria. In: Mertens JF, Sorin S (eds) Game Theoretic Methods in General Equilibrium Analysis. Kluwer, Dordrecht, pp 243–248

Myerson R (1985) Bayesian equilibrium and incentive compatibility. In: Hurwicz L, Schmeidler D, Sonnenschein H (eds) Social Goals and Social Organization. Cambridge University Press, Cambridge, pp 229–259

Myerson R (1994) Communication, correlated equilibria and incentive compatibility. In: Aumann R, Hart S (eds) Handbook of Game Theory, vol 2. Elsevier, Amsterdam, pp 827–847

Sorin S (1997) Communication, correlation and cooperation. In: Mas Colell A, Hart S (eds) Cooperation: Game Theoretic Approaches. Springer, Berlin, pp 198–218

Correlated Percolation

Antonio Coniglio[1,2,3] , Annalisa Fierro[2]

[1] Dipartimento di Scienze Fisiche, Università di Napoli "Federico II", Complesso Universitario di Monte Sant'Angelo, Naples, Italy

[2] INFM-CNR Coherentia, Università di Napoli "Federico II", Naples, Italy

[3] INFN Udr di Napoli, Università di Napoli "Federico II", Naples, Italy

Article Outline

Definition of the Subject
Introduction
Random Percolation
Percolation in the Ising Model
Hill's Clusters
Clusters in Weak and Strong Gels
Scaling Behavior of the Viscosity

Future Directions
Appendix: Random Cluster Model and Ising Droplets
Bibliography

Definition of the Subject

Cluster concepts have been extremely useful in elucidating many problems in physics. Percolation theory provides a generic framework to study the behavior of the cluster distribution. In most cases the theory predicts a geometrical transition at the percolation threshold, characterized in the percolative phase by the presence of a spanning cluster, which becomes infinite in the thermodynamic limit. Standard percolation usually deals with the problem when the constitutive elements of the clusters are randomly distributed. However correlations cannot always be neglected. In this case correlated percolation is the appropriate theory to study such systems. The origin of correlated percolation could be dated back to 1937 when Mayer [76] proposed a theory to describe the condensation from a gas to a liquid in terms of mathematical clusters (for a review of cluster theory in simple fluids see [88]). The location for the divergence of the size of these clusters was interpreted as the condensation transition from a gas to a liquid. One of the major drawbacks of the theory was that the cluster number for some values of thermodynamical parameters could become negative. As a consequence the clusters did not have any physical interpretation [50]. This theory was followed by Frenkel's phenomenological model [54], in which the fluid was considered as made of non interacting physical clusters with a given free energy. This model was later improved by Fisher [50], who proposed a different free energy for the clusters, now called droplets, and consequently a different scaling form for the droplet size distribution. This distribution, which depends on two geometrical parameters, σ and τ, has the nice feature that the mean droplet size exhibits a divergence at the liquid-gas critical point. Interestingly the critical exponents of the liquid gas critical point can be expressed in terms of the two parameters, σ and τ, and are found to satisfy the standard scaling relations proposed at that time in the theory of critical phenomena.

Introduction

Fisher's droplet model was very successful, to describe the behavior of a fluid or of a ferromagnet near the critical point, in terms of geometrical clusters. However the microscopic definition of such a cluster in a fluid or ferromagnet was still a challenge. While the exact definition in a continuum fluid model is still an open problem, a proper definition in the Ising model or lattice gas model has been provided. A first attempt to define a cluster in the Ising model which had the same properties of Fisher's droplet model was to consider a cluster as set of parallel spins. In two dimensions, in fact, these clusters seemed to have the properties of Fisher's droplets, i. e. the mean cluster size of these clusters were found to diverge at the Ising critical point on the basis of numerical analysis [8]. This result was later proved rigorously [36,37]. However the critical exponents for the mean cluster size in $2d$ was found to be larger than the corresponding critical exponent of the susceptibility [100], contrary to the requirement of Fisher's droplet model. Moreover numerical simulations in $3d$ and analytical results on the Bethe lattice showed that the critical point and the percolation point of such clusters were different. It was clear then that the clusters made of nearest neighbors parallel spins were too big to describe correlated regions. A different definition of clusters was then proposed [30] obtained by breaking the clusters of parallel spins by introducing fictitious bonds with a probability p_b between parallel spins. The new clusters are defined as a maximal set of parallel spins connected by bonds. For a particular choice of $p_b \equiv p = 1 - e^{-2J/k_B T}$ it was shown that these clusters (Coniglio–Klein droplets) have the same properties of Fisher's droplets, namely their size diverges at the Ising critical point with Ising exponents. Note that the bonds are only fictitious and do not change the energy of the spins. They only have the role of breaking the clusters made of parallel spins. Some years earlier Kasteleyn and Fortuin defined a random cluster model, obtained starting from an Ising model and by changing the spin interaction J in $J = \infty$, with probability p, and J into $J = 0$, with probability $1 - p$. They showed that the partition function of this modified model, called the random cluster model, coincides with the partition function of the original Ising model. In the random cluster model the clusters are defined as a maximal set of spins connected by infinite interactions. Although these clusters have the same properties of the droplet model, they were defined in the random cluster model, and for this reason these clusters were not associated with the droplets of the Ising model. It was only after Swendsen and Wang [99] introduced a cluster dynamics based on the Kasteleyn and Fortuin formalism, that it was formally shown [40] that the distribution of the Coniglio–Klein (CK) droplets are the same as the distribution of Kasteleyn–Fortuin (KF) clusters in the random cluster model. For this reason often the CK droplets and the KF clusters are identified, however the different meanings should be kept in mind.

A further development was obtained when the fractal structure of the droplets was studied not only for the

Ising model but for the full hierarchy of the q-state Potts model, which in the limit $q = 1$ gives the random percolation problem. It was shown that the critical droplets of the Potts model have the same structure, made of links and blobs, as found for the clusters in the random percolation problem. One of the consequences of this study was a better understanding of scaling and universality in terms of geometrical cluster and fractal dimension [74].

The cluster approach to the phase transitions leads also to a deeper understanding of why critical exponents do not depend on dimensionality above the upper critical dimension, and coincide with mean field exponents. It was in fact suggested [26] that, at least for random percolation, the mean field behavior is due to the presence of an infinite multiplicity of critical clusters at the percolation point. This suggests that similar results may be also extended to thermal problems.

Although the original interest in the field of correlated percolation was the study of critical phenomena in terms of geometrical concepts, later it was suggested that correlated percolation could be applied to the sol-gel transition, in particular when correlation was too large to be neglected. In many cases in fact the sol-gel transition, which is based on long range connectivity and percolation transition, interferes with large density fluctuation or critical point. The interplay between percolation points and critical points gives rise to interesting phenomena which are well understood within the concepts of correlated percolation. Correlated percolation has been studied also in systems with different types of long range correlation [108], and has been applied to many other fields such as nuclear physics [69], Gauge Theory [52] and O(n) models [7], fragmentation [15], urban growth [70], random resistor network [6], interacting colloids [12], biological models [1].

In Sect. "Random Percolation" we introduce random percolation concepts. In Sect. "Percolation in the Ising Model" in the context of the Ising model it is shown how clusters have to be defined in order to describe correlated regions corresponding to spin fluctuations. In Subsects. "Ising Clusters" and "Ising Droplets" the Ising clusters and droplets are respectively introduced, and in Subsect. "Ising Droplets Above $d = 4$" it is shown how the mapping between thermal properties and connectivity breaks down below T_c above $d = 4$. In Subsect. "Generalization to the q-State Potts Model" the results found for the Ising model are extended to the q-state Potts model, and in Subsect. "Fractal Structure in the Potts Model: Links and Blobs" the fractal structure is studied in terms of links and blobs. In Subsect. "Fortuin–Kasteleyn–Random Cluster Model" Fortuin–Kasteleyn–Random Cluster Model is presented, and the connection with the Coniglio–Klein droplets is further developed in Appendix A. In Sect. "Hill's Clusters" the possibility to extend the definition of droplets to simple fluids is discussed. In Sect. "Clusters in Weak and Strong Gels" the mechanism, leading to the formation of bound states in gelling systems, is considered, and in Sect. "Scaling Behavior of the Viscosity" the effect that finite bond lifetime has on the behavior of viscosity in weak or colloidal gels. Finally future directions and open problems are discussed in Sect. "Future Directions".

Random Percolation

In this section we define some connectivity quantities and present some results in the context of random percolation, which we will use in the following sections, where the correlated percolation will be presented.

Consider a d-dimensional hypercubic lattice of linear dimension L. Suppose that each edge has a probability p of being occupied by a bond. For small values of p, small clusters made of sites connected by nearest-neighbor bonds are formed. Each cluster is characterized by its size or mass s, the number of sites in the cluster. For large values of p in addition to small clusters we expect a macroscopic cluster that connects the opposite boundaries. This spanning cluster becomes infinite as the system size becomes infinite. For an infinite system there exists a percolation threshold p_c below which only finite clusters are present.

In order to describe the percolation transition [13,58,95], one defines: an order parameter, $P_\infty(p)$, as the density of sites in the infinite cluster, the mean cluster size, $S(p)$, of the finite clusters, and the average number of clusters, $K(p)$.

These quantities can be related to the average number of clusters of s sites per site, $n(s, p)$, and near the percolation threshold the critical behavior is characterized by critical exponents:

$$K(p)|_{\text{sing}} = \sum n(s,p)|_{\text{sing}} \sim |p - p_c|^{2-\alpha_p} , \tag{1}$$

$$P_\infty(p) = 1 - \sum s n(s,p) \sim \begin{cases} 0 & \text{if } p < p_c \\ (p - p_c)^{\beta_p} & \text{if } p > p_c , \end{cases} \tag{2}$$

$$S(p) = \sum s^2 n(s,p) \sim |p - p_c|^{-\gamma_p} , \tag{3}$$

where the sum is over all finite clusters, and in Eq. (1) only the singular part has been considered. Finally one can define the pair connectedness function p^f_{ij} as the probability

that i and j are in the same finite cluster through

$$\xi^2(p) = \frac{\sum r_{ij}^2 p_{ij}^f}{\sum p_{ij}^f} . \tag{4}$$

The connectedness length, $\xi(p)$, which is the critical radius of the finite clusters, diverges as

$$\xi \sim |p - p_c|^{-\nu_p} . \tag{5}$$

The critical exponents defined in Eqs. (1)–(4) are not all independent. Scaling relations can be derived among them as for ordinary second phase transitions. These scaling laws are intimately related to the property of the incipient infinite cluster of being a self similar fractal [74] to all length scales. The mass, s^*, of a typical cluster of linear dimension, ξ, scales as $s^* \sim \xi^{D_p}$, where D_p is the fractal dimension of the cluster.

Scaling and Hyperscaling

To obtain scaling laws, following Kadanoff's original idea, we perform [26] the following three steps: (i) divide the system into cells of linear dimension b, (ii) coarse grain by some suitable rule, (iii) rescale the lengths by a factor b. The result is a renormalized system where the size of the large clusters s has been reduced by factor b^{D_p} and all lengths by a factor b:

$$L' = \frac{L}{b} , \quad \xi = \frac{\xi'}{b} , \quad s' = \frac{s}{b^{D_p}} . \tag{6}$$

Assuming that the large clusters do not interpenetrate, the sum over the large clusters in an interval between $(s, s + \Delta s)$ must be the same before and after rescaling, i. e.

$$N(s, \xi)\Delta s \sim N(s', \xi')\Delta s' \tag{7}$$

where $N(s, \xi)/L^d = \overline{n}(s, \xi)$ is the number of clusters of s sites per unit volume. Dividing by the volume L^d, from (6) we obtain

$$\overline{n}(s, \xi) = b^{-d-D_p}\overline{n}(sb^{-D_p} , \ \xi b^{-1}) . \tag{8}$$

Choosing $b = s^{1/D_p}$ from (8) we obtain

$$n(s, p) = s^{-\tau_p} f((p - p_c)s^{\sigma_p}) \tag{9}$$

where $n(s, p) = \overline{n}(s, \xi)$ and

$$\tau_p = \frac{\mathrm{d}}{D_p} + 1, \quad \sigma_p = \frac{1}{\nu_p D_p} . \tag{10}$$

Equation (9) exhibits the scaling form postulated by Stauffer [13,58,95]. From (1), (2) and (10) we have:

$$2 - \alpha_p = \frac{\tau_p - 1}{\sigma_p} , \quad \beta_p = \frac{\tau_p - 2}{\sigma_p} , \quad -\gamma_p = \frac{\tau_p - 3}{\sigma_p} \tag{11}$$

and

$$\tau_p = 2 + \frac{\beta_p}{\beta_p + \gamma_p} , \quad \sigma_p = \frac{1}{\beta_p + \gamma_p} , \tag{12}$$

from which the following scaling relations are obtained:

$$\alpha_p + 2\beta_p + \gamma_p = 2 , \tag{13}$$

$$\frac{1}{\nu_p}(\beta_p + \gamma_p) = D_p . \tag{14}$$

From (10), (11) one can also find relations which contain the Euclidean dimensionality d called hyperscaling relation:

$$2 - \alpha_p = \nu_p d , \tag{15}$$

$$d - \frac{\beta_p}{\nu_p} = D_p . \tag{16}$$

Equation (16) was originally suggested in [68]. In $2d$ exact results give $\tau_p = 187/91$ and $\sigma_p = 36/91$, and in $3d$ the best estimates $\tau_p \simeq 2.18$, $\sigma_p \simeq 0.45$. From mean field theory [57] we know that for any d above the upper critical dimension $d_c = 6$, the critical exponents coincide with the mean field ones, namely $-\alpha_p = \beta_p = \gamma_p = 1$, $\nu_p = \sigma_p = 1/2$ and $\tau_p = 5/2$. These exponents satisfy the scaling relation (13), but fail to satisfy the hyperscaling relation (15) except for $d = 6$.

Moreover, while Eq. (14), for any $d > 16$, shows that the fractal dimension is stacked at the value $D_p = 4$, the hyperscaling relation (16) breaks down for $d > 6$.

Breakdown of Hyperscaling

By following a less conventional scaling approach, here we want to propose a geometrical interpretation of hyperscaling, why it breaks down above d_c, and why the hyperscaling breakdown occurs when the mean field becomes valid [26].

Let us assume that the singular behavior comes only from the critical clusters. Say N_ξ is the number of such clusters in a volume of the order ξ^d. The singular part of the cluster number is given by

$$\frac{N_\xi}{\xi^d} \sim \xi^{\frac{\alpha_p - 2}{\nu_p}} . \tag{17}$$

At the same time, the density of sites in the infinite cluster $P_\infty \sim |p - p_c|^{\beta_p}$ scales as the total mass of the spanning clusters $N_\xi s^*$ in a volume of linear dimension ξ divided by the volume ξ^d, namely

$$\frac{N_\xi \xi^{D_p}}{\xi^d} \sim \xi^{\frac{-\beta_p}{\nu_p}} , \tag{18}$$

where we have used $s^* \sim \xi^{D_p}$. Similarly the mean cluster size:

$$\frac{N_\xi \xi^{2D_p}}{\xi^d} \sim \xi^{\frac{\gamma_p}{\nu_p}} . \tag{19}$$

These equations lead to the scaling relations, Eqs. (13) and (14). Now if N_ξ is of the order of unity, we recover the hyperscaling relations [48], Eqs. (15) and (16), while if N_ξ diverges hyperscaling breaks down. We know that for dimension d above $d_c = 6$, $n(s, p) \sim s^{-5/2} e^{-(p-p_c)^2 s}$ for large s. Therefore $N_\xi = \xi^d \sum n(s, p) \sim \xi^{d-6}$, where $\xi \sim |p - p_c|^{-1/2}$. This calculation shows that, above d_c, N_ξ diverges and hyperscaling breaks down, and from Eqs. (17) and (18) the hyperscaling relations are replaced by $2 - \alpha_p = 6\nu_p$ and $D_p = 6 - \beta_p/\nu_p$, which in fact are satisfied for mean field exponents.

The more standard scaling approach of the previous section must be modified taking into account that for $d > 6$ the large number of clusters will be reduced by a factor b^{6-d}, then Eq. (7) will be modified as $N(s, \xi)\Delta s = b^{6-d} N(s', \xi')\Delta s'$ which still leads to all the Eqs. (7)–(14), except that d is replaced everywhere by 6. In particular, both Eqs. (14) and (16) give a fractal dimension $D_p = 4$.

The multiplicity of infinite clusters above d_c was numerically shown in [42,53]. The average (finite) number N_ξ of distinct clusters below d_c have been estimated theoretically and calculated numerically [3,63,94].

Consider now a critical cluster for $d > d_c$ just below p_c and its center of mass, 0. Say ξ_1 is the distance from 0, below which the cluster has not been penetrated by the other critical clusters. This length can be obtained by equating the mass density inside the region of radius ξ_1 to the mass density inside the region of radius ξ, $N_\xi \xi^{4-d} = \xi_1^{4-d}$, which gives $\xi_1 \sim \xi^{2/(d-4)}$.

If $\rho(r)$ is the density profile defined as the mass density of all the critical clusters at a distance r from 0, we expect that the density profile behaves as a power law r^{d-4} for $r < \xi_1$, as it should be for an object with fractal dimension $D_p = 4$ and as a constant for $r > \xi_1$ due to the penetration of the other critical clusters. Consequently we can make the following scaling Ansatz [2]:

$$\rho(r) = \frac{1}{r^{d-4}} f\left(\frac{r}{\xi_1}\right) , \tag{20}$$

where $f(x) \sim$ const for $x < 1$ and $\sim x^{d-4}$ for $x > 1$.

In conclusion, while for $d < 6$ the density of the order parameter fluctuates over a distance of the order ξ, for $d > 6$, where the mean field holds, the fluctuations are damped by the presence of infinitely many interpenetrating clusters, and the density of the order parameter crosses over from a power law (fractal) regime to a homogeneous regime at a distance $\xi_1 \ll \xi$.

The mean field solution is therefore a consequence of the presence of infinitely many interpenetrating clusters which suppress the spatial fluctuation of the order parameter. The condition for the validity of mean field theory is then given by $N_\xi \gg 1$. Using Eqs. (18) and (14) this condition implies

$$N_\xi^{-1} \sim \frac{\xi^{\frac{\gamma_p}{\nu_p}}}{\xi^d \xi^{\frac{-\beta_p}{\nu_p}}} \sim \frac{\langle \Delta M^2 \rangle}{\xi^d M^2} \ll 1 , \tag{21}$$

where M and $\langle \Delta M^2 \rangle$ are the order parameter and the fluctuations of the order parameter (here we used that the mean cluster size $S(p)$ has the same critical behavior as the fluctuations of the order parameter [18,34]). Interestingly enough Eq. (21) coincides with Ginzburg criterion for the validity of mean field theory.

Cluster Structure

Nodes and Links In the previous sections we have shown that the Incipient Infinite Cluster (IIC) is a fractal. Here we want to show in more details the internal structure of the IIC. A very useful nodes and links picture for the infinite cluster just above p_c was introduced by Skal and Shklowskii [89] and de Gennes [44]. In this picture the infinite cluster consists of a superlattice made of nodes, separated by a distance of the order of ξ, connected by macrobonds. Just below p_c the structure of the very large cluster, the IIC, was expected to have the same structure as the macrobonds.

Later on, in 1977, Stanley [90] made the important observation that in general for each configuration of bonds at p_c the IIC can be partitioned into three categories. By associating an electric unit resistance to each bond, and applying a voltage between the ends of the cluster, one distinguishes the dangling bonds which do not carry current (yellow bonds). The remaining bonds are the backbone bonds. The backbone can be partitioned in singly connected bonds (red bonds) and all the others, the multiply connected bonds, which lump together in "blobs" (blue bonds). The red bonds, which carry the whole current, have also the property that if one is cut the cluster breaks

into two parts. This partition in three types of bonds is very general and can be done for any cluster or aggregate.

The next major problem was to determine whether the blobs are or not relevant. In the nodes and links picture the assumption is that the blobs are irrelevant and only links are relevant. A further elaboration [90] assumed that the backbone close to p_c would reduce to a self-avoiding walk chain, which implies that the blobs are not relevant. This self-avoiding walk Ansatz received a large amount of attention, since it predicted a value for the crossover exponent of the dilute Heisenberg ferromagnetic model near the percolation threshold in $2d$, in good agreement with the experimental data, although the prediction for the dilute Ising crossover exponent did not agree as well with the data [9].

Syerpinsky Gasket: A Model Without Links In 1981 a completely alternative model was proposed by Gefen et al. [55]. Based on the observation that in a computer simulation the red bonds were hardly seen, they proposed an alternative model, the Syerpinsky gasket, that represents the opposite extreme of the nodes and links picture. It has a self-similar structure but only multiply connected bonds are present. A great advantage of this model is that it can be solved exactly. It also gives good prediction for the fractal dimension of the backbone, but it fails to predict the correct value for the dilute Ising crossover exponent [9].

Nodes, Links and Blobs Motivated by all these conflicting models, some rigorous results were presented which led unambiguously to the nodes, links and blobs picture of the infinite cluster [23,24], in which both links and blobs are relevant below $d = 6$, while only links are relevant above $d = 6$ or in the mean field. In particular the following relation was proven for any p and for any lattice in any dimension:

$$p\frac{\mathrm{d}p_{ij}}{\mathrm{d}p} = \lambda_{ij} \tag{22}$$

where p_{ij} is the probability that i and j are connected, λ_{ij} is the average number of red bonds between i and j, such that if one is cut, i and j would have been disconnected. From Eq. (22) it is possible to calculate the average number L_{ij} of red bonds between i and j under the condition that i and j are in the same cluster:

$$L_{ij} = \frac{\lambda_{ij}}{p_{ij}} \,. \tag{23}$$

From the scaling form of $p_{ij} = r_{ij}^{-d+2-\eta_p} f(r_{ij}/\xi)$ it follows

$$L_{ij} = r_{ij}^{\frac{1}{\nu_p}} f_1\left(\frac{r_{ij}}{\xi}\right) \,, \tag{24}$$

where $f_1(x)$ is related via Eqs. (22) and (23) to $f(x)$ and goes to a constant for $x \ll 1$. In particular, by putting $r_{ij} = \xi$ in Eq. (24) we obtain

$$L_{\mathrm{R}} \sim \xi^{\frac{1}{\nu_p}} \,, \tag{25}$$

where $L_{\mathrm{R}} \equiv L(r_{ij} = \xi)$ is the average number of red bonds between two points separated by a distance of the order of ξ. From Eq. (25) it follows that the fractal dimension of the red bonds is $D_{\mathrm{R}} = 1/\nu_p$. An immediate consequence is that not only the red bonds are relevant but also the number of bonds L_{B} in the blobs diverge. For more details, see [23,24]. Later Eq. (25) was confirmed numerically by Pike and Stanley [83] in $d = 2$. Although the links are much less in number than the backbone bonds, they can be detected experimentally, in fact it can be shown [23,25] that only the links determine the critical behavior of the dilute Ising model at p_c leading to a crossover exponent 1 in any d. While for a dilute Heisenberg system the crossover exponent is related to the resistivity exponent, in agreement with the experimental data of [9].

In conclusion we can write the following relations

$$y_{\mathrm{H}} = D_p, \quad y_{\mathrm{T}} = D_{\mathrm{R}} \tag{26}$$

where $y_{\mathrm{H}} = d - \beta_p/\nu_p$ is the so-called magnetic field scaling exponent and $y_{\mathrm{T}} = 1/\nu_p$ is the thermal scaling exponent in the renormalization group language. This result is quite interesting as it shows that the scaling exponents can be expressed in terms of geometric quantities: The fractal dimension D_p of the entire incipient infinite cluster, and the fractal dimension D_{R} of the subset made of red bonds.

Surfaces and Interfaces

The study of the structure of the surfaces and interfaces of the large clusters below p_c has not received as much attention as the study of the internal structure of the IIC. This problem is relevant to the study of the dielectric constant of random composite materials, the viscosity of a gel, the conductivity of a random superconducting network, and the relative termite diffusion model.

For simplicity, let us consider a random superconducting network in which superconducting bonds are present with probability p and normal bonds carrying a unit resistance with probability $1 - p$. For small values of p we have finite superconducting clusters in a background of normal resistor. As $p \to p_c$, the superconductivity Σ diverges. For a finite cell of linear dimension L just below p_c, the typical configurations are characterized by two very large clusters almost touching, each one attached to one of two opposite faces. Inside these clusters there are islands of normal

resistors. If a unit voltage is applied between the opposite face of the hypercube, there is no current flowing through the bonds in the island. We call these "dead" bonds, analogous to the dead ends of the percolating cluster. The remaining normal bonds connect one superconducting cluster to the other. These bonds are made of "bridges", also called "antired" bonds, which have the property that if one is replaced by a superconducting bond, a percolating superconducting cluster is formed, and the remaining multiple "connecting" bonds. Similar to the red bonds, it can be proved [26] that the fractal dimensionality of the antired bonds is $1/\nu_p$. The proof is based on the following relation which can be proved for any lattice in any dimension

$$(1-p)\frac{\mathrm{d}p_{ij}}{\mathrm{d}p} = \mu_{ij}\,, \tag{27}$$

where p_{ij} is the pair connectedness function (the probability that sites i and j belong to the same cluster) and μ_{ij} is the average number of antired bonds between i and j. These are defined as non-active bonds, such that if one is made active, i and j become connected.

The above considerations suggest that just below p_c the system can be imagined as a superlattice made of large critical clusters whose centers are separated by a distance of the order ξ. The surfaces of these clusters almost touch, and are connected by bridges made of single bonds and other paths made of more than one bond [33].

Finally we mention the following result which relates the size of the critical cluster s^* and the size of the entire perimeter t^* [13,58,95]

$$t^* = \frac{1-p}{p}s^* - As^{*\sigma_p}\,, \tag{28}$$

where $\sigma_p = 1/(\nu_p D_p)$ is the critical exponent which appears in the cluster number Eq. (9) and A is a constant. The last term $s^{*\sigma_p}$ which appears also in Fisher's droplet model [50] is usually interpreted as the surface of the droplet. However, if it was a surface, σ_p should satisfy the following boundary $(d-1)/d \le \sigma_p \le 1$. The upper bound corresponds to the fully rarefied droplets and the lower bound to compact droplets. Surprisingly enough for the percolation problem σ_p is strictly smaller than $(d-1)/d$. This paradox can be solved by using a result [25], which shows that $As^{*\sigma_p}$ is equal to a number of antired bonds between critical clusters separated by a distance of order ξ. Since the subset of antired bonds is only a subset of the entire perimeter, it explains why $\sigma_p < (d-1)/d$. This result gives the best geometrical interpretation of the thermal scaling exponent y_T. It in fact shows that $y_T = D_{AR}$, where D_{AR} is the fractal dimension of the antired bonds namely that part of the surface which contributes to the surface tension.

Percolation in the Ising Model

In this section we want to extend the percolation problem to the case in which the particles are correlated. The simplest model to consider is the lattice gas or Ising model. In the following we will use the Ising terminology. We know that the Ising model exhibits a thermodynamic transition for zero external field, $H = 0$, at a critical temperature T_c. The question that we ask is how the percolation properties are modified due to the presence of correlation. We first consider the case when the clusters are made of nearest-neighbor down spins (Subsect. "Scaling and Hyperscaling"). Later in Subsect."Breakdown of Hyperscaling" we will modify the cluster definition in such a way that these new clusters describe the thermal fluctuations namely we require that the clusters satisfy the same properties as the droplets in Fisher's droplet model [50]. Namely, i) the size of the clusters must diverge at the Ising critical points, ii) the linear dimension of the clusters must diverge with the same exponent as the correlation length, and iii) the mean cluster size must diverge with the same exponent as the susceptibility.

These conditions are satisfied if the cluster size distribution for zero external field has the following form

$$n(s,T) = s^{-\tau} f((T-T_c)s^{\sigma})\,. \tag{29}$$

The parameters σ and τ are related to critical exponents α, β and γ through Eqs. (10) and (11), where now α, β and γ are the Ising critical exponents. In particular for $d = 2$, $\sigma = 8/15 \simeq 0.53$ and $\tau = 31/15 \simeq 2.07$, and for $d = 3$, $\sigma \simeq 0.64$ and $\tau \simeq 2.21$.

Ising Clusters

The Hamiltonian of the Ising model is given by:

$$\mathcal{H} = -J\sum_{\langle ij\rangle} S_i S_j - H\sum_i S_i \tag{30}$$

where $S_i = \pm 1$ are the spin variables, J is the interaction between two nearest-neighbor (nn) spins and H is the magnetic field.

From the thermodynamic point of view the only quantities of interest are those which can be obtained from the free energy and those were the only quantities that Onsager was concerned with in his famous solution of the $2d$ Ising model. However one can look at the Ising model from a different perspective by studying the connectivity properties using concepts such as clusters which have been

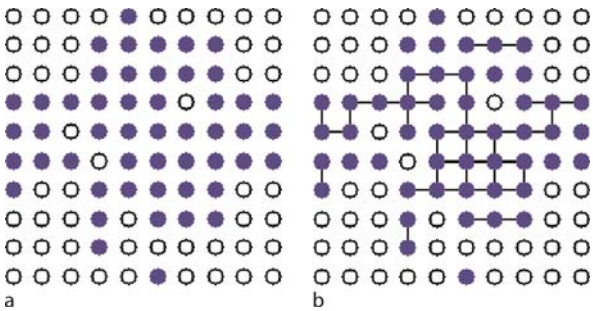

Correlated Percolation, Figure 1
a Ising configuration at T_c: "down" spins are represented by *filled circles*. b Correct clusters are obtained from the configuration given in (a) by putting bonds between occupied sites with probability $p = 1 - e^{-2\beta J}$

systematically elaborated in percolation theory [92]. There are two reasons for approaching the problem also from the connectivity point of view. One reason is that it gives a better understanding of the mechanism of the phase transition [50]. In fact, concepts like universality and scaling have been better understood in terms of geometrical clusters and fractal dimensions [27]. A second reason is that there are physical quantities amenable to experimental observations, which are associated to the connectivity properties and cannot be obtained from the free energy. It is very important to note however that the definition of connectivity, and therefore the definition of the cluster, is not always the same, but may depend on the particular observable associated with it.

In the Ising model, for a given configuration of spins it is rather natural to define a cluster as a maximal set of *nn* down parallel spins[1] (Fig. 1). For some time these clusters were believed to be responsible for the correlations present in the Ising model. This idea was also based on numerical results which showed evidence that in two dimensions the mean cluster size diverges at the thermal critical point [8]. However the idea that the clusters could describe thermal correlations was definitively abandoned when it was shown, by numerical simulations in the three-dimensional Ising model [78] and by exact solution on the Bethe lattice [22], that the percolation point appeared in the low density phase of down spins on the coexistence curve at a temperature T_p before the critical point T_c is reached ($T_p < T_c$).

Later it was suggested by topological arguments [21] that only in two dimensions does the critical point coincide with the percolation point, but not necessarily in higher dimensions. The arguments followed two steps: in the first step it was argued that an infinite cluster of up spins is a necessary condition for having a spontaneous magnetization. This implies a percolation transition of down spins on the coexistence curve $T_p \leq T_c$, in the second step it was argued that due to topological reasons in two dimensions it is not possible to have an infinite cluster of up spins coexisting with an infinite cluster of down spins, which implies $T_p \geq T_c$. Combining with the previous inequalities one obtains in two dimensions $T_p = T_c$.

Later these results were proven rigorously [36,37] along with many other results relating connectivity and thermodynamic quantities. For more details we refer to the original papers.

It is clear that the Ising clusters, defined as a group of *nn* parallel spins, do not have the property of describing correlated regions corresponding to spin fluctuations as originally expected. In fact even in two dimensions, where the thermal critical point coincides with the percolation point, the Ising clusters were not suitable for such description. Series expansion showed that the mean cluster size diverges with an exponent, $\gamma^* = 1.91 \pm 0.001$, rather different from the susceptibility exponent, $\gamma = 1.75$ [100]. Later it has been shown exactly that $\gamma^* = 91/48$ [97].

Ising Droplets

From the properties mentioned in Subsect. "Ising Clusters", it appears that the Ising clusters are too big to describe the proper droplets. The reason is that there are two contributions to the Ising clusters. One is due to correlations but there is another contribution purely geometrical due to the fact that two *nn* spins even in the absence of correlation have a finite probability of being parallel. The last contribution becomes evident in the limit of infinite temperature and zero external field. In this case, in fact, although there is no correlation and the susceptibility is zero, the cluster size is different from zero. In fact in $3d$ at infinite temperature there is even an infinite cluster of "up" and "down" spins.

Binder [8] proposed to cut the infinite cluster in order to have $T_p = T_c$ in $d = 3$, but he did not give the microscopic prescription to do it. Later Coniglio and Klein [30] proposed to reduce the cluster size by introducing fictitious bonds between *nn* parallel spins with probability p_b (Fig. 1). These new clusters are made of *nn* parallel spins connected by bonds. The original Ising cluster will either reduce its size or will break into smaller clusters. If $p_b = 1$, we obtain the Ising clusters again. This case is known as the site correlated percolation problem because one looks

[1] The Ising Hamiltonian, Eq. (30), is equivalent to the lattice gas Hamiltonian $\mathcal{H}_{LG} = -J' \sum_{\langle ij \rangle} n_i n_j - \mu \sum_i n_i$, with $n_i = (1 - S_i)/2$, $J' = 4J$ and $\mu = 2H - 4J$. In the lattice gas terminology an Ising cluster is a maximal set of *nn* occupied sites.

at the properties of the Ising clusters just as in the random percolation problem. The main difference is that in random percolation the occupied sites are randomly distributed, while in this case the down (or up) spins are correlated according to the Ising Hamiltonian. In the infinite temperature limit one recovers random percolation. The case $p_b \neq 1$ is called site-bond correlated percolation [39].

A Hamiltonian formalism was proposed to study site correlated percolation [79]. This formalism was generalized in [30] to study site-bond correlated percolation. In this case for the zero external field the Hamiltonian is given by the following dilute Ising s-state Potts Model (DIPM)[2]:

$$-\mathcal{H}_{\mathrm{DP}} = J_b \sum_{\langle ij \rangle} (\delta_{\sigma_i \sigma_j} - 1)(S_i S_j + 1) + J \sum_{\langle ij \rangle} S_i S_j \,, \quad (31)$$

where $\sigma_i = 1, \ldots, s$ are Potts variables and the sum is over all nearest neighbor sites. In the same way as the s-state Potts model in the limit $s = 1$ [110] describes the random bond percolation model, the DIPM describes percolation in the Ising model where the clusters are made of parallel spins connected by bonds with probability, $p_b = 1 - e^{-2\beta J_b}$.

In particular the average number of clusters G, that plays the role of the free energy in the percolation problem is given by $G = \mathrm{d}F/\mathrm{d}s|_{s=1}$, where

$$-\beta F = \lim_{N \to \infty} \frac{1}{N} \ln \left(\sum_{\{\sigma_i S_i\}} e^{-\beta \mathcal{H}_{\mathrm{DP}}} \right) . \quad (32)$$

At that time the DIPM was investigated in a different context by Berker et al. [81]. The model exhibits the interesting properties that by choosing $J_b = J$ it coincides with a pure $s + 1$-state Potts model. Therefore in the limit $s = 1$ the DIPM coincides with the $s = 2$ Potts model namely with the Ising model. Consequently F becomes the Ising model free energy and G has a singularity at the Ising critical point. This argument immediately suggested that the site-bond correlated percolation for $J_b = J$ namely with the bond probability given by

$$p_b \equiv p = 1 - e^{-2\beta J} \,, \quad (33)$$

should reproduce the same critical behavior of the Ising model. Namely the percolation quantities become critical at the Ising critical point in the same way as the corresponding thermal quantities.

[2]Originally in [30] the Hamiltonian of the DIPM, $\mathcal{H}_{\mathrm{DP}}$, was expressed in terms of the lattice gas variables n_i, and the Ising droplets were defined as *nn* occupied sites connected by bonds, corresponding to *nn* down spins.

In fact using real space renormalization group arguments, it was possible to show that the size of the clusters of parallel spins connected by bonds with probability, p_b, given by Eq. (33), diverges at the Ising critical point with Ising exponents, exhibiting thus the same properties as the droplets in Fisher's model. These clusters were called droplets to distinguish them from the Ising clusters.

Droplets in Two and Three Dimensions

This site-bond correlated percolation problem has been studied by real space renormalization groups in two dimensions [30,35], by ϵ expansion, near six dimensions [31] and by Monte Carlo in two and three dimensions [64,82,85,92].

The renormalization group analysis shows that in $2d$ the Ising critical point is a percolation point for down or up spins connected by bonds for all values of bond probability such that $1 \leq p_b < 1 - e^{-2\beta J}$. The fractal dimension $D^* = (\gamma^*/\nu + 2)/2 = 187/96$ [97] being higher than the fractal dimension $D = (\gamma/\nu + 2)/2 = 15/8$ for the value of $p_b \equiv p = 1 - e^{-2\beta J}$.

In the renormalization group language this means that there are two fixed points, one corresponding to the universality class of the Ising cluster, the other one corresponding to the droplets. In the first one, the variable J_b is irrelevant, namely the scaling exponent associated with it, $y_b < 0$. In the second fixed point associated with the droplets instead $y_b > 0$. The result is that the Ising critical point is a percolation point for a range of values of p_b, at the first sight seems counter-intuitive. In fact if the Ising critical point corresponds to the onset of percolation for Ising clusters ($p_b = 1$), one would expect that for $p_b < 1$ the clusters would not percolate anymore. The puzzle can be clarified by studying the fractal structure of the Ising clusters and the droplets at T_c [27]. In fact it can be shown that y_b is the scaling exponent of the red bonds, namely $L_R \sim l^{y_b}$ where L_R is the number of red bonds between two connected sites separated by a distance of the order l, consequently the droplets, characterized by $y_b > 0$, are made of links and blobs, like in random percolation. Due to the presence of links the cluster breaks apart and does not percolate anymore as the bond probability decreases. On the contrary the Ising clusters ($p_b = 1$), characterized by $y_b < 0$, are made only of blobs and no links, therefore by decreasing the bond probability the infinite cluster does not break and still percolates, until $p_b = p$.

In $3d$ at the Ising critical point, T_c, there is an analogous line of anomalous percolation points for clusters of down spins connected by bonds, for all values of bond probability such that $1 \leq p_b < 1 - e^{-2\beta J}$, although the

probability P_∞ for a down spin to be in the infinite cluster is different from zero. More precisely the quantity $p_{ij} - P_\infty^2$ decays as a power law, where p_{ij} is the probability that i and j are connected. For more details see [29]. As p_b decreases towards $p = 1 - e^{-2\beta J}$ there is a crossover towards a different power law characterized by the Ising exponent, while P_∞ goes to 0.

Droplets in an External Field

By keeping the same definition of droplets given above, in the case of an Ising model in an external field $H > 0$ one finds a phase diagram in the H, T plane or in the M, T plane, with a percolation line of "down" spins ending at the Ising critical point (see Fig. 2). Along the percolation line one finds critical exponents in the universality class of random percolation with a cross-over to Ising critical exponents as the Ising critical point is approached. The Ising critical point being a higher order critical point for the percolation transition. This percolation line, also known as the Kertesz line, has received some attention [66,93,95,105] (see for more details the review by Sator [88]). Although the Ising free energy has no singularity along this line some physical interpretation is given to the Kertesz line [17].

On the other hand this line disappears if the droplet definition is modified in the presence of an external field [40,104], according to Kasteleyn and Fortuin formalism [65] and the Swendsen and Wang approach [99]. In this approach the field is treated as a new interaction between each spin and a ghost site. Consequently for positive H (negative H) an "up" ("down") spin can be connected to the ghost spin with a probability $p_H = 1 - e^{-2\beta|H|}$. Droplets now are defined as a maximal set of spins connected by bonds where as before the bonds between nearest-neighbor parallel spins have probability p_b given by Eq. (33) and p_H between spins and the ghost spin. Note that two far away spins can be easily connected through the ghost spin. In this way the presence of a positive (negative) magnetic field implies always the presence of an infinite cluster of "up" ("down") spins.

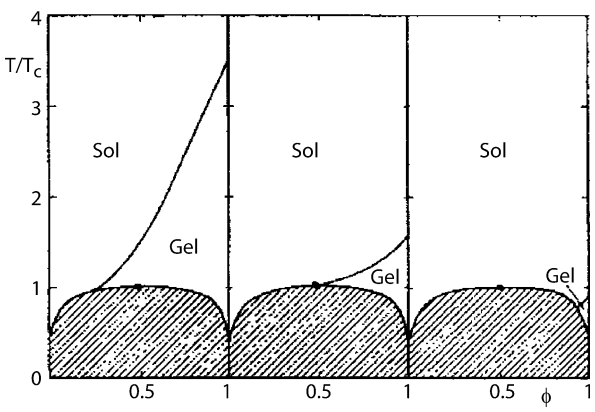

Correlated Percolation, Figure 2
Monte Carlo simulations of the 3*d* lattice gas model for three values of the bond probability $p_b = 1 - e^{-2c\beta J}$ with the constant $c = 2.25, 1, 0.564$ from *left* to *right*. Φ is the density of down spins. The Gel and the Sol indicates the percolation and non percolation phase. From [59]

Exact Relations Between Connectivity and Thermal Properties

Interestingly it was also shown [40] that the droplets so defined have the same statistics as the clusters in the random cluster model introduced by Kasteleyn and Fortuin (KF) [65] (see Subsect. "Fortuin–Kasteleyn–Random Cluster Model"), although the CK droplets and the KF clusters have a different meaning. Using the relations between the connectivity properties of the random cluster model and the thermal properties of the Ising model, it was finally possible to prove that in any dimension and for any temperature T and external field $H > 0$, provided that the extension of the droplet definition in the external field is considered, the following relations between connectivity and thermal properties hold [40]:

$$\begin{cases} \rho_\infty = m \\ p_{ij} = g_{ij} \end{cases} \tag{34}$$

where ρ_∞ is the density of up spins in the percolating droplet, m is the magnetization per site, p_{ij} is the probability that i and j are connected (through both finite or infinite droplet) and $g_{ij} = \langle S_i S_j \rangle$.

In particular, for $T > T_c$ and zero external field $H \to 0$, we have that the magnetization $m = 0$ and g_{ij} coincides with the spin-spin pair correlation function. Consequently $\rho_\infty = 0$, namely the probability for a spin to be in an infinite droplet is zero, and therefore p_{ij} coincides with the probability that two spins i and j are in the same finite droplet. For $T < T_c$ instead we have $\rho_\infty = m > 0$, and $p_{ij} = p_{ij}^f + p_{ij}^\infty$, where p_{ij}^f (p_{ij}^∞) is the probability that spins in i and j are in a finite (infinite) droplet. From Eq. (34) it follows for $T < T_c$:

$$p_{ij}^f + p_{ij}^\infty - \rho_\infty^2 = \langle S_i S_j \rangle - m^2 \,. \tag{35}$$

By summing over i and j we have

$$S + (\Delta\rho_\infty)^2 = \chi \,, \tag{36}$$

where S is the mean cluster size of the finite clusters, $(\Delta\rho_\infty)^2$ is the fluctuation of the density of the infinite cluster and χ is the susceptibility. These exact results show that above T_c mean cluster size and susceptibility coincide, while below T_c there are two contributions to the susceptibility, one due to the mean cluster size and the second related to the fluctuation of the density of the infinity cluster. Monte Carlo calculations [85] show that both terms have the same critical behavior as also occurs in random percolation [18,34], so the mean cluster size S diverges like the susceptibility. We expect that this is the case for dimensions up to $d = 4$, the upper critical dimensionality of the Ising model. In the mean field, as we will see, the mean cluster size below T_c diverges with an exponent different from the susceptibility.

One very interesting application based on the KF approach was produced by Swendsen and Wang [99,106], who elaborated a cluster dynamics which drastically reduced the slowing down near the critical point of the Ising and Potts model (see also [109] for further developments).

The droplet definition can be extended to the *nn* antiferromagnetic Ising model [4] and to the Ising model with any ferromagnetic interaction J_{ij} between sites i and j [64]. In this case the CK clusters are defined as set of parallel spins connected by bonds present between i and j with probability $p_{ij} = 1 - \exp[-2\beta J_{ij}]$. It can be shown that also in this case the relations (34) between connectivity and thermal quantities hold.

Ising Droplets Above $d = 4$

In Subsect. "Ising Droplets" we have reported the relations Eqs. (34) and (35), which are exact and are valid in any dimension including the mean field. As a matter of fact in the mean field the percolation order parameter and the magnetization are identical and go to zero with the exponent $\beta = 1/2$, while the mean cluster size above T_c coincides with the susceptibility and diverges with the exponent $\gamma = 1$. The same is true for the connectedness length above T_c, which coincides with the correlation length, and diverges with an exponent $\nu = 1/2$. However below T_c the mean cluster size diverges with an exponent $\gamma' = 1/2$ and the correlation length with an exponent $\nu' = 1/4$. The result is a consequence that the two terms in Eq. (35), the probability that two sites are in the same finite droplet, p^f_{ij}, and the correlation of the infinite droplet density at site i and j, $p^\infty_{ij} - \rho^2_\infty$, do not scale in the same way, giving rise to two lengths, diverging respectively with exponents ν' and ν.

These somehow anomalous results are probably a consequence that the Ising model has an upper critical dimension $d_c = 4$ while the DIPM which describes the droplet problem has an upper critical dimension $d_c = 6$ [31]. In fact there are arguments that for $4 \le d \le 6$ below T_c the critical exponents are $\nu' = 1/(d-2)$, $\gamma' = 2/(d-2)$, $\beta = 1/2$, $\eta = 0$ and fractal dimension $D_p = 1/2(d+2)$, with an upper critical dimension $d_c = 6$. Of course for $T > T_c$ the exponents are $\gamma = 1$, $\nu = 1/2$ and $\eta = 0$.

Due to the breakdown of the mapping between thermal fluctuations and mean cluster size below T_c above $d = 4$, it is not possible to extend easily the geometrical picture, employed in random percolation, to explain the breakdown of hyperscaling in the Ising model. For a study of droplets inside the metastable region see [5].

Generalization to the q-State Potts Model

All the results found for the Ising case have been extended [32] to the q-state Potts model. This model is defined by the following Hamiltonian:

$$-\mathcal{H}_q = qJ \sum_{\langle ij\rangle} \delta_{\sigma_i\sigma_j} , \tag{37}$$

where the spin variables σ_i can assume q values, $\sigma_i = 1, \ldots, q$. This model coincides with the Ising model for $q = 2$, reproduces the random percolation problem in the limit $q = 1$ and the tree percolation model in the limit $q = 0$ [110]. The geometrical approach developed in the previous sections for the Ising model, can be extended to the q-state Potts model. In particular one can define the site-bond Potts correlated percolation, where clusters are made of *nn* spins in the same state, connected by bonds with bond probability p_b. By choosing $p_b = p = 1 - e^{-q\beta J}$, it is possible to show that these clusters percolate at the Potts critical temperature $T_c(q)$, with percolation exponents identical to the thermal exponents and therefore behave as the critical droplets.

The formalism is based on the following diluted Potts model [32,102]:

$$-\mathcal{H}^q_{DP} = J_b \sum_{\langle ij\rangle} (\delta_{\tau_i\tau_j} - 1)\delta_{\sigma_i\sigma_j} + qJ \sum_{\langle ij\rangle} \delta_{\sigma_i\sigma_j} , \tag{38}$$

where the second term, which controls the distribution of spin variables, is the q-state Potts Hamiltonian, whereas the first term contains auxiliary Potts variables $\tau_i = 1, 2, \ldots, s$ and controls the bonds distribution.

As in the Ising case, Hamiltonian (37) in the limit $s \to 1$ describes the site-bond Potts correlated percolation problem with p_b given by $p_b = 1 - e^{-q\beta J_b}$. The droplets are obtained in the particular case $J_b = qJ$. For this value in fact Hamiltonian (37) for $s \to 1$ coincides with the q-state Potts model.

Correlated Percolation, Table 1
Fractal dimensions, for $d = 2$, of the whole cluster (D), of the Hull (D_H), and of the red bonds (D_R) for the Potts droplets. It is also reported the thermal power exponent y_T

q	D	y_T	D_H	D_R
0	2	0	2	5/4
1	91/48	3/4	7/4	3/4
2	15/8	1	5/3	13/24
3	28/15	6/5	8/5	7/20
4	15/8	3/2	3/2	0

Once the Ising and Potts model has been mapped onto a percolation problem, we can extend some of the results of random percolation to thermal problems.

Fractal Structure in the Potts Model: Links and Blobs

Like in random percolation, also in the q-state Potts model it can be shown that at $T_c(q)$ the critical droplets have a fractal structure made of links and blobs, with a fractal dimension $D(q) = d - \beta(q)/\nu(q)$, where $\beta(q)$ and $\nu(q)$ are respectively the order parameter and correlation length exponent. Therefore $D(q)$ coincides with the magnetic scaling exponent $y_H(q)$. However the fractal dimension of the red bonds $D_R(q)$ does not coincide with the thermal scaling exponent $y_T(q)$, associated with the thermal variable J, like in random percolation. Instead $D_R(q)$ is found to coincide with the bond probability scaling exponent y_b associated with the variable J_b in Hamiltonian (37) [27].

Like for random percolation the fractal dimension of the red bonds coincides with the fractal dimension of the antired bonds. Using the mapping from the Potts model to the Coulomb gas [87], it is possible to obtain the exact value of the fractal dimension of the red bonds and of the external perimeter or hull [27]. For further exact results see also [10,97].

From Table 1 it appears that the exact value of $D(q)$ does not vary substantially with q, for $d = 2$. This observation can be understood by noting that, using this geometrical approach, the driving mechanism of the critical behavior can be viewed as coalescence of clusters just like in random percolation. Then one would expect for any q that the fractal dimension should be close to the fractal dimension of the critical clusters in the percolation problem. This also explains the observation of Suzuky [98], known as strong universality, that for a large class of models the ratio γ/ν or β/ν do not vary appreciably. Since these ratios of critical exponents for fixed d depend only on the magnetic scaling exponent, which is identical to the fractal dimension, the strong universality is a consequence of the quasi-universal feature of the fractal dimension as discussed above. Unlikely the fractal dimension of the whole cluster, $D_R(q)$ and $D_H(q)$ do change substantially and characterize the different models as function of q. Particularly sensitive to q is the fractal dimension of the red bonds, which has its largest value at $q = 0$ (tree percolation), where the backbone is made only of links. As q approaches q_c the cluster becomes less ramified until the red bonds vanish ($D_R(4) = 0$). This results in a drastic structural change from a links and blobs picture to a blobs picture only, anticipating a first order transition. Interestingly, the fractal dimension of the red bonds for $q = 0$, $D_R = 5/4$, has been related to the abelian sandpile model [47]. The reason why $D_R(q)$ is so model-dependent is due to the fact that the fractal set of the red bonds is only a small subset of the entire droplet, and therefore this "detail" is strongly model dependent. Also the thermal exponent $y_T(q)$ is strongly model-dependent, however so far the geometrical characterization in terms of a fractal dimension for such exponent has not been found except $q = 1$ (random percolation).

Fortuin–Kasteleyn–Random Cluster Model

We will present here the random cluster model introduced by Kasteleyn and Fortuin. Let us consider the q-state Potts model on a d-dimensional hypercubic lattice. By freezing and deleting each interaction of the Hamiltonian (see "Appendix: Random Cluster Model and Ising Droplets"), they managed to write the partition function of the Potts model, $Z = \sum_{\{\sigma_i\}} e^{-\beta \mathcal{H}_q}$, in the following way

$$Z = \sum_C p^{|C|}(1-p)^{|A|} q^{N_C} , \tag{39}$$

where C is a configuration of bonds defined in the same hypercubic lattice, just like a bond configuration in the standard percolation model, $|C|$ and $|A|$ are respectively the number of bonds present and absent in the configuration C, and N_C is the number of clusters in the configuration C.

In conclusion, in the KF formalism the partition function of the Potts model is identical to the partition function (39) of a correlated bond percolation model [62,65] where the weight of each bond configuration C is given by

$$W(C) = p^{|C|}(1-p)^{|A|} q^{N_C} \tag{40}$$

which coincides with the weight of the random percolation except for the extra factor q^{N_C}. They called this particular correlated bond percolation model the random cluster model. Clearly for $q = 1$ the cluster model coincides with the random percolation model.

Kasteleyn and Fortuin have related the percolation quantities associated with the random cluster model to the corresponding thermal quantities in the q-state Potts model [65]. In particular for the Ising case, $q = 2$,

$$|\langle S_i \rangle| = \langle \gamma_i^{\infty} \rangle_W \tag{41}$$

and

$$\langle S_i S_j \rangle = \langle \gamma_{ij} \rangle_W \,, \tag{42}$$

where $\langle \ldots \rangle$ is the Boltzmann average and $\langle \ldots \rangle_W$ is the average over bond configurations in the bond correlated percolation with weights given by (40). Here $\gamma_i^{\infty}(C)$ is equal to 1 if the spin at i belongs to the infinite cluster, 0 otherwise; $\gamma_{ij}(C)$ is equal to 1 if the spins at sites i and j belong to the same cluster, 0 otherwise.

Interestingly the connectivity properties in the KF random cluster model can be related to the CK droplets:

$$\rho_{\infty} = \langle \gamma_i^{\infty} \rangle_W \,, \tag{43}$$

$$p_{ij} = \langle \gamma_{ij} \rangle_W \,, \tag{44}$$

where ρ_{∞} and p_{ij} are defined in Subsect. "Ising Droplets Above $d = 4$". From Eqs. (41)–(44) it follows Eqs. (34).

Hill's Clusters

In this section we discuss the possibility to extend the definition of droplets to simple fluids. In 1955 Hill [60] introduced the concept of physical clusters in a fluid in an attempt to explain the phenomenon of condensation from a gas to a liquid. In a fluid made of particles interacting via a pair potential $u(r)$ physical clusters are defined as a group of particles pairwise bounded. A pair of particles is bounded if in the reference frame of their center of mass their total energy is less than zero. Namely their relative kinetic energy plus the potential energy is less than zero. The probability that two particles at distance r are bounded can be calculated [60] and is given by

$$p_{\mathrm{H}}(r) = \frac{4}{\pi} \int_0^{\sqrt{-\beta u(r)}} x^2 \mathrm{e}^{-x^2} \mathrm{d}x \,. \tag{45}$$

More recently it was noted [16] that the bond probability Eq. (45) calculated for the interaction of the three-dimensional *nn* lattice gas model is almost coincident with the bond probability p of Eq. (33). This implies that Hill's physical clusters for the $3d$ lattice gas almost coincide with the droplets defined by Coniglio and Klein, and in fact Hill's clusters percolate along a line almost indistinguishable from the droplets percolation line (see Fig. 2).

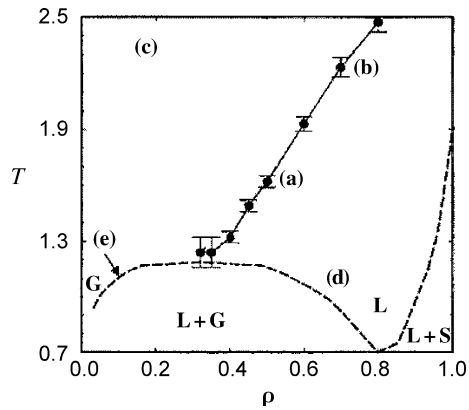

Correlated Percolation, Figure 3
Phase diagram of the Lennard–Jones fluid using molecular dynamics. The full line corresponds to percolation of cluster following Hill's definition. From [17]

In order to calculate percolation quantities in a fluid, in [38] the authors developed a theory based on Mayer's expansion. In particular, using this theory they calculated analytically for a potential made of a hard core plus an attractive interaction, the percolation line of Hill's physical clusters in a crude mean field approximation and compared with the liquid gas coexistence curve. They found that the percolation line ended just below the critical point in the low density phase but not exactly at the critical point. For further developments of the theory see [56].

Very recently, Campi et al. [17], using molecular dynamics have calculated the percolation line of Hill's physical clusters for a Lennard–Jones potential. The results showed a percolation line ending close or at the critical point (Fig. 3) suggesting that Hill's clusters are good candidates to describe the density fluctuations like the droplets in the lattice gas model, although there is no proof of relations analogous to those valid for the droplets in the lattice gas such as Eq. (34), which would prove that their size would diverge exactly at the critical point with thermal exponents.

Although Hill's clusters may represent the critical fluctuation near the critical point, we may wonder whether they have a physical meaning away from the critical point. In particular we may wonder whether we can detect experimentally the percolation line in the phase diagram. In a Lennard–Jones fluid, molecular dynamics shows that quantities such as viscosity or the diffusion coefficient do not seem to exhibit any anomalous behavior through the percolation line [17]. In some colloids instead the percolation line is detected through a steep increase of the viscosity. What would be the difference in the two cases? The difference may rely on their lifetime. The possibility to de-

tect the percolation line of these clusters is expected to depend on the lifetime of the clusters which in turn depends on the bond lifetime. The larger the cluster lifetime, the larger the increase of the viscosity, and the better the percolation line can be detected. In Sect. "Scaling Behavior of the Viscosity" we will discuss the behavior of the viscosity as function of the lifetime of the clusters.

Clusters in Weak and Strong Gels

In the previous section we have shown the case in which the probability of having a bond between two particles coincides with the probability that the two particles form a bound state defined according to Hill's criterion. Now we want to show another mechanism leading to the formation of bound states, which is more appropriate to gels. The importance of connectivity in gels was first emphasized by Flory [51]. The application of percolation theory to gels was later suggested by de Gennes [43] and Stauffer [91,96]. Here we consider a system made of monomers in a solvent. Following [39] we shall assume that the monomers can interact with each other in two ways. One is the usual van der Waals interaction, and the other is a directional interaction that leads to a chemical bond. A simple model for such a system is a lattice gas model where an occupied site represents a monomer and an empty site a solvent. For simplicity we can put the monomer-solvent interaction and the solvent-solvent interaction equal to zero, and include such interactions in an effective monomer-monomer interaction. The monomer-monomer interaction ε_{ij} can reasonably be approximated by a nearest neighbor interaction

$$\varepsilon_{ij} = \begin{cases} -W \\ -E \end{cases} \tag{46}$$

where $-W$ is the van der Waals type of attraction and $-E$ is the bonding energy. Of course, this second interaction, which is the chemical interaction, occurs only when the monomers are in particular configurations. For simplicity we can suppose that there is one configuration which corresponds to the interaction of strength E, and Ω configurations which correspond to the interaction of strength W. We expect $E \gg W$ and $\Omega \gg 1$. It can be easily calculated [39] that such a system is equivalent to a lattice gas model with an effective *nn* interaction $-\varepsilon$ given by

$$e^{\beta\varepsilon} = e^{\beta E} + \Omega e^{\beta W} . \tag{47}$$

Therefore from the static point of view the system exhibits a coexistence curve and a critical temperature which characterizes the thermodynamics of the system. However the system microscopically behaves rather different from a standard lattice gas. In fact in a configuration in which two monomers are *nn*, in a standard lattice gas they feel one interaction, while in the system considered here with some probability p_b they feel a strong chemical interaction $-E$ and with probability $1 - p_b$ they feel a much smaller interaction $-W$. The probability p_b can be easily calculated and is given by

$$p_b = \frac{e^{\beta E}}{e^{\beta E} + \Omega e^{\beta W}} . \tag{48}$$

In conclusion, the system from the static point of view is equivalent to a lattice gas with interaction ε given by (47). However we can also study the percolation line of the clusters made by monomers connected by chemical bonds. This can be done by introducing bonds between *nn* particles in the lattice gas with *nn* interactions, the bonds being present with probability p_b given by Eq. (48). By changing the solvent the effective interaction W changes and one can realize three cases topologically similar to those of Fig. 2, where the percolation line ends at the critical point or below the critical point in the low density or high density phase (for more details see [39]).

The lifetimes of the bonds are of the order of $e^{\beta E}$. Since E is very large the lifetime could be very large. For an infinite bond lifetime the bonded clusters are permanent and the viscosity diverges due to the divergence of the mean cluster size (see for example [96]), and the percolation line can be easily detected. We consider three particular physical systems which could be rather emblematic of a general situation where the percolation line has been detected:

a) Microemulsions of water in oil [20]
b) Triblock copolymers in unicellular systems [71,72]
c) Gelatin water methanol systems [101].

In Figs. 4, 5 and 6 we show respectively the phase diagram of the systems a), b), c), where the coexistence curve in the temperature-concentration diagram is shown together with "percolation lines".

In particular, in a) the system consists of three components AOT/water/decane. For the temperature and the concentration of interest, the system can be considered as made of small droplets of oil surrounded by water in a solvent. The droplets interact via a hard core potential plus short range attractive interaction. Because of the entropic nature of the attractive interaction, the coexistence curve is "upside-down" with the critical point being the minimum instead of the maximum (Fig. 4). The broken line is characterized by a steep increase of conductivity.

In b) the system is made of triblock copolymers unicellular in water solution, c is the volume fraction of the uni-

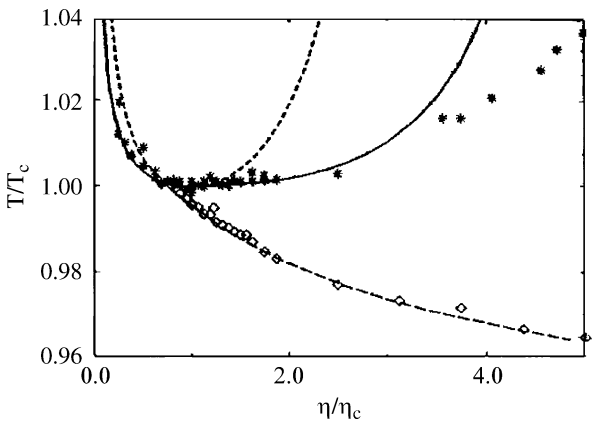

Correlated Percolation, Figure 4
Experimental points in AOT/water/decane from [20] together with the *coexistence curve* and *spinodal curve* based on the Baxter's model. The *percolation line* where the conductivity exhibits a steep increase has been fitted with the Baxter's model

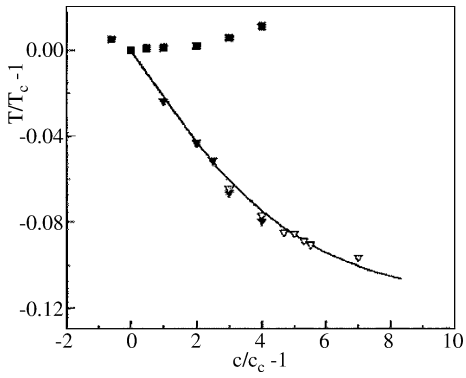

Correlated Percolation, Figure 5
***L*/64 water system. Experimental points of the *coexistence curve* and *percolation line*, where the viscosity exhibits a steep increase. From [72]**

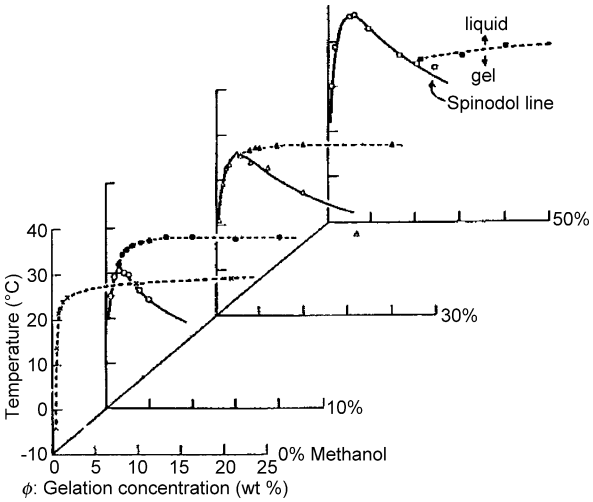

Correlated Percolation, Figure 6
Sol-Gel transition temperature (*solid symbols*) and the spinodal temperature (*open symbols*) of gelatin-water-methanol mixtures as function of gelatin concentration. At the sol-gel transition the viscosity diverges. From [101]

cells (Fig. 5). The line is characterized by a steep increase in the viscosity.

In c) the system is made of gelatin dissolved in water + methanol; ϕ is the gelatin concentration. The broken lines are characterized by the divergence of the viscosity and correspond to the sol-gel transition. Each line represents a different value of the methanol concentration, which has been chosen in such a way that the line ends at the consolute point or, below it, in the low or high density phase (Fig. 6).

In all these experiments the consolute point is characterized by a thermodynamical singularity, where the correlation length and compressibility diverge. The other lines are usually ascribed to a "percolation" transition. However it is important to precisely identify which are the relevant clusters in the three different systems. Also we would like to understand why, in system c), the viscosity diverges at the percolation transition, while in b) it reaches a plateau, and why in a) and b) the "percolation" lines end on the coexistence curve close to the critical point in the low density region. It is also important to realize that for each phenomenon is very important to define the proper cluster, which is responsible for the physical phenomenon. In the conductivity experiments in microemulsion the proper clusters are made of "touching" spheres similar to nearest neighbor particles in a lattice gas model. The viscoelastic properties of microemulsions may be more suitably described by clusters made of spheres pairwise bonded.

From the cluster properties of the lattice gas model we expect the infinite cluster is a necessary condition for a critical point therefore the percolation line ends just below the critical point in the low density region, as observed in the experiments described above and more recently in numerical simulations of models of interacting colloids [84].

In weak reversible gelatin the clusters are made of monomers (or polymers) bonded by a strong interaction which leads to chemical bond. In this case the bond probability can be changed by changing the solvent and therefore the percolation line, by properly choosing the solvent, can end on the coexistence curve at or below the critical point.

The reason why the viscosity in the gel experiments diverges at the percolation point, while it reaches a plateau in

colloids, is due to the lifetime of the bonds which is much longer in the first system than in the second [46,73,86]. In low density colloids the proper cluster to describe colloidal gelation also appear to be related to strong bonds with large bond lifetime [14,41,86]. When the relaxation time is much smaller than the bond lifetime the dynamics is dominated by the clusters, otherwise a crossover is expected towards a regime due to the crowding of the particles [41]. Percolation line of clusters pairwise bonded can also be defined in fluids, but due to the negligible lifetime cannot be detected.

Scaling Behavior of the Viscosity

If the lifetime of the chemical bonds is infinite, the viscosity exhibits a divergence at the percolation threshold as recently shown in different models [11,45,103]

$$\eta \sim \xi^{\tilde{k}} , \tag{49}$$

where ξ is the linear dimension of the critical cluster which diverges at the percolation threshold with the exponent ν.

The relation between the diffusion coefficient $D(R)$ of a cluster of radius R and the viscosity η would be given by the Stokes–Einstein relation for a cluster radius much larger than ξ

$$D(R) \sim \frac{1}{R\eta} . \tag{50}$$

For cluster radius R smaller than ξ it has been proposed [75] that the viscosity will depend also on R in such a way as to satisfy a generalized Stokes–Einstein relation Eq. (50) with $\eta = \eta(R)$. When $R = \xi$ the viscosity $\eta(\xi) = \eta$, and from Eq. (50) one obtains the following scaling behavior for R:

$$D(R) \sim R^{-(1+\tilde{k})} \tag{51}$$

therefore the relaxation time $\tau(R)$ for a cluster of radius R is

$$\tau(R) \sim R^{1+\tilde{k}} . \tag{52}$$

If τ is the lifetime of a typical cluster, then a cluster of radius R will contribute to the viscosity if $\tau(R) < \tau$, and therefore:

$$\eta \sim \xi^{\tilde{k}} f\left(\frac{\tau}{\xi^{1+\tilde{k}}}\right) \sim \begin{cases} \xi^{\tilde{k}} & \tau > \xi^{1+\tilde{k}} \\ \tau^{\frac{\tilde{k}}{1+\tilde{k}}} & \tau < \xi^{1+\tilde{k}} \end{cases} \tag{53}$$

which implies that the viscosity will exhibit a steep increase followed by a plateau. The higher τ is, the higher the plateau.

The viscosity data on microemulsion (Fig. 5) shows in fact such a plateau, suggesting that the mechanism for the appearance of the plateau is linked to the bond lifetime which in turn is related to the cluster relaxation time.

Future Directions

In conclusion, we have discussed the interplay between the percolation line and critical point in systems where thermal correlations play an important role. The problem of defining the droplets in spin models is satisfactorily solved. However there are still some open problems. Above $d = 4$ in the Ising model the definition of droplets presents some difficulties, probably related to the upper critical dimension for the percolation problem. That critical dimension is six. This type of difficulty does not allow for a trivial extension of the arguments used in the random percolation problem, to explain the hyperscaling breakdown. Another open problem is the characterization of the thermal scaling exponent $1/\nu$, in terms of the fractal dimension of some subset of the critical droplet, as occurs in the random percolation problem.

In the last decade the KF, CK approach has been extended to frustrated systems. Interestingly this approach has led to a new frustrated percolation model, with unusual properties relevant to spin glasses and other glassy systems [77,80]. However the precise definition of clusters, which are able to characterize the critical droplets for spin glasses, is still missing.

Although some advances have been obtained towards a droplet definition in Lennard Jones systems [17], a general definition for continuum models of fluids still needs to be formulated.

Appendix: Random Cluster Model and Ising Droplets

In 1969 Kasteleyn and Fortuin (KF) [65] introduced a correlated bond percolation model, called the random cluster model, and showed that the partition function of this percolation model was identical to the partition function of the q-state Potts model. They also showed that the thermal quantities in the Potts model could be expressed in terms of connectivity properties of the random cluster model. Much later in 1980 Coniglio and Klein [30] independently have used a different approach with the aim to define the proper droplets in the Ising model. It was only later that it was realized that the two approaches were related, although the meaning of the clusters in the two approaches is different. We will discuss these two approaches here, and show that their statistical properties are the same.

Random Cluster Model

Let us consider an Ising system of spins $S_i = \pm 1$ on a lattice with nearest-neighbor interactions and, when needed, let us assume periodic boundary conditions in both directions. All interactions have strength J and the Hamiltonian is

$$\mathcal{H}(\{S_i\}) = -\sum_{\langle i,j\rangle} J(S_i S_j - 1)\,, \tag{54}$$

where $\{S_i\}$ represents a spin configuration and the sum is over *nn* spins. The main point in the KF approach is to replace the original Ising Hamiltonian with an annealed diluted Hamiltonian

$$\mathcal{H}'(\{S_i\}) = -\sum_{\langle i,j\rangle} J'_{ij}(S_i S_j - 1)\,, \tag{55}$$

where

$$J'_{ij} = \begin{cases} J' & \text{with probability } p \\ 0 & \text{with probability } (1-p)\,. \end{cases} \tag{56}$$

The parameter p is chosen such that the Boltzmann factor associated with an Ising configuration of the original model coincides with the weight associated with a spin configuration of the diluted Ising model

$$\begin{aligned} \mathrm{e}^{-\beta\mathcal{H}(\{S_i\})} &\equiv \prod_{\langle i,j\rangle} \mathrm{e}^{\beta J(S_i S_j - 1)} \\ &= \prod_{\langle i,j\rangle} \left(p\mathrm{e}^{\beta J'(S_i S_j - 1)} + (1-p)\right)\,, \end{aligned} \tag{57}$$

where $\beta = 1/k_\mathrm{B}T$, k_B is the Boltzmann constant and T is the temperature. In order to satisfy (57) we must have

$$\mathrm{e}^{\beta J(S_i S_j - 1)} = p\mathrm{e}^{\beta J'(S_i S_j - 1)} + (1-p)\,. \tag{58}$$

We take now the limit $J' \mapsto \infty$. In such a case $\mathrm{e}^{\beta J'(S_i S_j - 1)}$ equals the Kronecker delta $\delta_{S_i S_j}$ and from (58) p is given by

$$p = 1 - \mathrm{e}^{-2\beta J}\,. \tag{59}$$

From (57), by performing the products we can write

$$\mathrm{e}^{-\beta\mathcal{H}(\{S_i\})} = \sum_C W_\mathrm{KF}(\{S_i\}, C)\,, \tag{60}$$

where

$$W_\mathrm{KF}(\{S_i\}, C) = p^{|C|}(1-p)^{|A|} \prod_{\langle i,j\rangle \in C} \delta_{S_i S_j}\,. \tag{61}$$

Here C is a configuration of interactions where $|C|$ is the number of interactions of strength $J' = \infty$ and $|A|$ the number of interactions of strength 0. $|C| + |A| = |E|$, where $|E|$ is the total number of edges in the lattice.

$W_\mathrm{KF}(\{S_i\}, C)$ is the statistical weight associated a) with a spin configuration $\{S_i\}$ and b) with a set of interactions in the diluted model where $|C|$ edges have ∞ strength interactions, while all the other edges have 0 strength interactions. The Kronecker delta indicates that two spins connected by an ∞ strength interaction must be in the same state. Therefore the configuration C can be decomposed in clusters of parallel spins connected by infinite strength interactions.

Finally the partition function of the Ising model Z is obtained by summing the Boltzmann factor (60) over all the spin configurations. Since each cluster in the configuration C gives a contribution of 2, we obtain:

$$Z = \sum_C p^{|C|}(1-p)^{|A|} 2^{N_C}\,, \tag{62}$$

where N_C is the number of clusters in the configuration C.

In conclusion, in the KF formalism the partition function (62) is equivalent to the partition function of a correlated bond percolation model [62,65] where the weight of each bond configuration C is given by

$$W(C) = \sum_{\{S_i\}} W_\mathrm{KF}(\{S_i\}, C) = p^{|C|}(1-p)^{|A|} 2^{N_C} \tag{63}$$

which coincides with the weight of the random percolation except for the extra factor 2^{N_C}. Clearly all percolation quantities in this correlated bond model are weighted according to Eq. (63) coincide with the corresponding percolation quantities of the KF clusters made of parallel spins connected by an ∞ strength interaction, whose statistical weight is given by (61). Moreover using (61) and (60) Kasteleyn and Fortuin have proved that [65]

$$|\langle S_i\rangle| = \langle \gamma_i^\infty\rangle_W \tag{64}$$

and

$$\langle S_i S_j\rangle = \langle \gamma_{ij}\rangle_W\,, \tag{65}$$

where $\langle\ldots\rangle$ is the Boltzmann average and $\langle\ldots\rangle_W$ is the average over bond configurations in the bond correlated percolation with weights given by (63). Here $\gamma_i^\infty(C)$ is equal to 1 if the spin at site i belongs to the spanning cluster, 0 otherwise; $\gamma_{ij}(C)$ is equal to 1 if the spins at sites i and j belong to the same cluster, 0 otherwise.

Connection Between the Ising Droplets and the Random Cluster Model

In the approach followed by Coniglio and Klein [30], given a configuration of spins, one introduces at random connecting bonds between *nn* parallel spins with probability p_b, antiparallel spins are not connected with probability 1. Clusters are defined as maximal sets of parallel spins connected by bonds. The bonds here are fictitious, they are introduced only to define the clusters and do not modify the interaction energy as in the FK approach. For a given realization of bonds we distinguish the subsets C and B of *nn* parallel spins respectively connected and not connected by bonds and the subset D of *nn* antiparallel spins. The union of C, B and D coincides with the total set of *nn* pair of spins E. The statistical weight of a configuration of spins and bonds is [28,40]

$$W_{\mathrm{CK}}(\{S_i\}, C) = p_b^{|C|}(1-p_b)^{|B|}\mathrm{e}^{-\beta\mathcal{H}(\{S_i\})} , \tag{66}$$

where $|C|$ and $|B|$ are the number of *nn* pairs of parallel spins respectively in the subset C and B not connected by bonds.

For a given spin configuration, using the Newton binomial rule, we have the following sum rule

$$\sum_C p_b^{|C|}(1-p_b)^{|B|} = 1 . \tag{67}$$

From Eq. (67) follows that the Ising partition function, Z, may be obtained by summing (66) over all bond configurations and then over all spin configurations.

$$Z = \sum_{\{S_i\}}\sum_C W_{\mathrm{CK}}(\{S_i\}, C) = \sum_{\{S_i\}} \mathrm{e}^{-\beta\mathcal{H}(\{S_i\})} . \tag{68}$$

The partition function of course does not depend on the value of p_b which controls the bond density. By tuning p_b instead it is possible to tune the size of the clusters. For example by taking $p_b = 1$ the clusters would coincide with nearest-neighbor parallel spins, while for $p_b = 0$ the clusters are reduced to single spins. By choosing the droplet bond probability $p_b = 1 - \mathrm{e}^{-2\beta J} \equiv p$ and observing that $\mathrm{e}^{-\beta\mathcal{H}(\{S_i\})} = \mathrm{e}^{-2\beta J|D|}$, where $|D|$ is the number of antiparallel pairs of spins, the weight (66) simplifies and becomes:

$$W_{\mathrm{CK}}(\{S_i\}, C) = p^{|C|}(1-p)^{|A|} , \tag{69}$$

where $|A| = |B| + |D| = |E| - |C|$.

From (69) we can calculate the weight $W(C)$ that a given configuration of connecting bonds C between *nn* parallel spins occurs. This configuration C can occur in many spin configurations. So we have to sum over all spin configurations compatible with the bond configuration C, namely

$$W(C) = \sum_{\{S_i\}} W_{\mathrm{CK}}(\{S_i\}, C) \prod_{\langle i,j\rangle\in C} \delta_{S_iS_j} , \tag{70}$$

where, due to the product of the Kronecker delta, the sum is over all spin configurations compatible with the bond configuration C. From (59) and (70) we have

$$\begin{aligned} W(C) &= \sum_{\{S_i\}} p^{|C|}(1-p)^{|A|} \prod_{\langle i,j\rangle\in C} \delta_{S_iS_j} \\ &= p^{|C|}(1-p)^{|A|}2^{N_C} . \end{aligned} \tag{71}$$

Consequently in (68) by taking first the sum over all spins compatible with the configuration C, the partition function Z can be written as in the KF formalism (62).

$$Z = \sum_C p^{|C|}(1-p)^{|A|}2^{N_C} . \tag{72}$$

In spite of the strong analogies the CK clusters and the KF clusters have a different meaning. In the CK formalism the clusters are defined directly in a given configuration of the Ising model as parallel spin connected by fictitious bonds, while in the KF formalism clusters are defined in the equivalent random cluster model. However, due to the equality of the weights (69) and (61) the statistical properties of both clusters are identical [40] and due to the relations between (61) and (63) both coincide with those of the correlated bond percolation whose weight is given by (63). More precisely, any percolation quantity $g(C)$ which depends only on the bond configuration has the same average

$$\langle g(C)\rangle_{\mathrm{KF}} = \langle g(C)\rangle_{\mathrm{CK}} = \langle g(C)\rangle_W , \tag{73}$$

where $\langle\ldots\rangle_{\mathrm{KF}}$, $\langle\ldots\rangle_{\mathrm{CK}}$ are the average over spin and bond configurations with weights given by (61) and (69) respectively and $\langle\ldots\rangle_W$ is the average over bond configurations in the bond correlated percolation with weights given by (63). In view of (73) it follows [40]

$$|\langle S_i\rangle| = \langle\gamma_i^\infty\rangle_{\mathrm{CK}} \tag{74}$$

and

$$\langle S_iS_j\rangle = \langle\gamma_{ij}\rangle_{\mathrm{CK}} . \tag{75}$$

We end this section noting that in order to generate an equilibrium CK droplet configuration in a computer simulation, it is enough to equilibrate a spin configuration of the Ising model and then introduce at random fictitious bonds between parallel spins with a probability given by (59).

Bibliography

Primary Literature

1. Abete T, de Candia A, Lairez D, Coniglio A (2004) Phys Rev Lett 93:228301
2. Aharony A, Gefen Y, Kapitulnik A (1984) J Phys A 17:l197; Alexander S, Grest GS, Makanishi H, Witten TA (1984) J Phys A 17:L185
3. Aizenman M (1997) Nucl Phys B 485:551
4. Amitrano C, di Liberto F, Figari R, Peruggi F (1983) J Phys A Math Gen 16:3925
5. Anghel M, Tobochnik J, Klein W, Gould H, Alexander FJ, Johnson G (2000) Phys Rev Lett 85:1270; Padoa SC, Sciortino F, Tartaglia P (1998) Phys Rev E 57:3797; Neerman DW, Coniglio A, Klein W, Stauffer D (1984) J Stat Phys 36:477
6. Bastiaansen PJM, Knops HJF (1997) J Phys A Math Gen 30:1791
7. Bialas P, Blanchard P, Fortunato S, Gandolfo D, Satz H (2000) Nucl Phys B 583:368; Blanchard P, Digal S, Fortunato S, Gandolfo D, Mendes T, Satz H (2000) J Phys A Math Gen 33:8603
8. Binder K (1976) Ann Phys NY 98:390
9. Birgeneau RJ, Cowley RA, Shirane G, Guggenheim HJ (1976) Phys Rev Lett 37:940; Birgeneau RJ, Cowley RA, Shirane G, Guggenheim HJ (1980) Phys Rev B 21:317
10. Blote HWJ, Knops YMM, Nienhuis B (1992) Phys Rev Lett 68:3440
11. Broderix K, Löwe H, Müller P, Zippelius A (2001) Phys Rev E 63:011510
12. Bug ALR, Safran SA, Grest GS, Webman I (1985) Phys Rev Lett 55:1896; Safran SA, Webman I, Grest GS (1985) Phys Rev A 32:506
13. Bunde A, Havlin S (1991) Percolation I. In: Bunde A, Havlin S (eds) Fractals and disordered systems. Springer, New York, pp 51–95
14. Campbell AI, Anderson VJ, van Duijneveldt JS, Bartlett P (2005) Phys Rev Lett 94:208301
15. Campi X, Krivine H (2005) Phys Rev C 72:057602; Krivine H, Campi X, Sator N (2003) Phys Rev C 67:044610; Mader CM, Chappars A, Elliott JB, Moretto LG,, Phair L, Wozniak GJ (2003) Phys Rev C 68:064601
16. Campi X, Krivine H, Puente A (1999) Physica A 262:328
17. Campi X, Krivine H, Sator N (2001) Physica A 296:24
18. Chayes JT, Chayes L, Grimmet GR, Kesten H, Schonmann R (1989) Ann Probab 17:1277
19. Chayes L, Coniglio A, Machta J, Shtengel K (1999) J Stat Phys 94:53
20. Chen SH, Rouch J, Sciortino F, Tartaglia P (1994) J Phys Cond Matter 6:10855
21. Coniglio A (1975) J Phys A 8:1773
22. Coniglio A (1976) Phys Rev B 13:2194
23. Coniglio A (1981) Phys Rev Lett 46:250
24. Coniglio A (1982) J Phys A 15:3829
25. Coniglio A (1983) In: Proceedings of Erice school on ferromagnetic transitions. Springer, New York
26. Coniglio A (1985) Finely divided matter. In: Boccara N, Daoud M (eds) Proc. les Houches Winter conference. Springer, New York
27. Coniglio A (1989) Phys Rev Lett 62:3054
28. Coniglio A (1990) In: Stanley HE, Ostrowsky W (eds) Correlation and connectivity – Geometric aspects of physics, chemistry and biology, vol 188. NATO ASI series. Kluwer, Dordrecht
29. Coniglio A, Figari R (1983) J Phys A Math Gen 16:L535
30. Coniglio A, Klein W (1980) J Phys A 13:2775
31. Coniglio A, Lubensky T (1980) J Phys A 13:1783
32. Coniglio A, Peruggi F (1982) J Phys A 15:1873
33. Coniglio A, Stanley HE (1984) Phys Rev Lett 52:1068
34. Coniglio A, Stauffer D (1980) Lett Nuovo Cimento 28:33
35. Coniglio A, Zia RVP (1982) J Phys A Math Gen 15:L399
36. Coniglio A, Nappi C, Russo L, Peruggi F (1976) Comm Math Phys 51:315
37. Coniglio A, Nappi C, Russo L, Peruggi F (1977) J Phys A 10:205
38. Coniglio A, De Angelis U, Forlani A, Lauro G (1977) J Phys A Math Gen 10:219; Coniglio A, De Angelis U, Forlani A (1977) J Phys A Math Gen 10:1123
39. Coniglio A, Stanley HE, Klein W (1979) Phys Rev Lett 42:518; Coniglio A, Stanley HE, Klein W (1982) Phys Rev B 25:6805
40. Coniglio A, di Liberto F, Monroy G, Peruggi F (1989) J Phys A 22:L837
41. Coniglio A, Abete T, de Candia A, del Gado E, Fierro A (2007) J Phys Condens Matter 19:205103; de Candia A, del Gado E, Fierro A, Sator N, Coniglio A (2005) Phys A 358:239; Coniglio A, de Arcangelis L, del Gado E, Fierro A, Sator N (2004) J Phys Condens Matter 16:S4831; Gimel JC, Nicolai T, Durand D (2001) Eur Phys J E 5:415
42. de Arcangelis L (1987) J Phys A 20:3057
43. de Gennes PG (1975) J Phys Paris 36:1049; de Gennes PG (1979) Scaling concepts in polymer physics. Cornell University Press, Ithaca
44. de Gennes PG (1976) La Recherche 7:919
45. del Gado E, de Arcangelis L, Coniglio A (2000) Eur Phys J E 2:359
46. del Gado E, Fierro A, de Arcangelis L, Coniglio A (2004) Phys Rev E 69:051103
47. Dhar D (1999) Physica A 263:4
48. Dunn AG, Essam JW, Ritchie DS (1975) J Phys C 8:4219; Cox MAA, Essam JW (1976) J Phys C 9:3985
49. Essam JW (1980) Rep Prog Phys 43:833
50. Fisher ME (1967) Physics NY 3:225; Fisher ME (1967) J Appl Phys 38:981; Fisher ME, Widom B (1969) J Chem Phys 50:3756; Fisher ME (1971) In: Green MS (ed) Critical Phenomena. Proc. of the international school of physics "Enrico Fermi" course LI, Varenna on lake Como (Italy). Academic, New York, p 1
51. Flory PJ (1941) J Am Chem Soc 63:3083; Flory PJ (1979) Principles of polymer chemistry. Cornell University Press. Ithaca
52. Fortunato S, Satz H (2000) Nucl Phys B Proc Suppl 83:452
53. Fortunato S, Aharony A, Coniglio C, Stauffer D (2004) Phys Rev E 70:056116
54. Frenkel J (1939) J Chem Phys 7:200; Frenkel J (1939) J Chem Phys 7:538
55. Gefen Y, Aharony A, Mandelbrot BB, Kirkpatrick S (1981) Phys Rev Lett 47:1771
56. Given JA, Stell G (1991) J Phys A Math Gen 24:3369
57. Harris AB, Lubensky TC, Holcomb W, Dasgupta C (1975) Phys Rev Lett 35:327
58. Havlin S, Bunde A (1991) Percolation II. In: Bunde A, Havlin S (eds) Fractals and disordered systems. Springer, New York, pp 97–149
59. Heermann DW, Stauffer D (1981) Z Phys B 44:339
60. Hill TL (1955) J Chem Phys 23: 617
61. Hong DC, Stanley HE, Coniglio A, Bunde A (1986) Phys Rev B 33:4564

62. Hu CK (1984) Phys Rev B 29:5103; Hu CK (1992) Phys Rev Lett 69:2739; Hu CK, Mak KS (1989) Phys Rev B 40:5007
63. Hu CK, Lin CY (1996) Phys Rev Lett 77:8
64. Jan N, Coniglio A, Stauffer D (1982) J Phys A 15:L699
65. Kasteleyn PW, Fortuin CM (1969) J Phys Soc Japan Suppl 26:11; Fortuin CM, Kasteleyn PW (1972) Phys Utrecht 57:536
66. Kertesz J (1989) Physica A 161:58
67. Kertesz J, Coniglio A, Stauffer D (1983) Clusters for random and interacting percolation. In: Deutscher G, Zallen R, Adler J (eds) Percolation structures and processes, Annals of the Israel Physical Society 5. Adam Hilger, Bristol, pp 121–147. The Israel Physical Society, Jerusalem
68. Kirkpatrick S (1978) AIP Conference Proc. 40:99
69. Ma YG (1999) Phys Rev Lett 83:3617; Ma YG, Han DD, Shen WQ, Cai XZ, Chen JG, He ZJ Long JL, Ma GL, Wang K, Wei YB, Yu LP, Zhang HY, Zhong C, Zhou XF, Zhu ZY (2004) J Phys G Nucl Part Phys 30:13
70. Makse HA, Havlin S, Stanley HE (1995) Nature 377:608; Makse HA, Andrade JS Jr, Batty M, Havlin S, Stanley HE (1998) Phys Rev E 58:7054
71. Mallamace F, Chen SH, Liu Y, Lobry L, Micali N (1999) Physica A 266:123
72. Mallamace F, Gambadauro P, Micali N, Tartaglia P, Liao C, Chen SH (2000) Phys Rev Lett 57:5431
73. Mallamace F, Chen SH, Coniglio A, de Arcangelis L, del Gado E, Fierro A (2006) Phys Rev E 73:020402
74. Mandelbrot BB (1982) The fractal geometry of nature. Freeman, San Francisco
75. Martin JE, Douglas A, Wilcoxon JP (1988) Phys Rev Lett 61:262
76. Mayer JE (1937) J Chem Phys 5:67; Mayer JE, Ackermann PG (1937) J Chem Phys 5:74; Mayer JE, Harrison SF (1938) J Chem Phys 6:87; Mayer JE, Mayer MG (1940) Statistical mechanics. Wiley, New York
77. Monroy G, Coniglio A, di Liberto F, Peruggi F (1991) Phys Rev B 44:12605; Coniglio A (2000) Physica A 281:129
78. Muller-Krhumbaar H (1974) Phys Lett A 50:27
79. Murata KK (1979) J Phys A 12: 81
80. Newman CM, Machta J, Stein DL (2007) J Stat Phys 130:113
81. Nienhuis B, Berker AN, Riedel EK, Shick M (1979) Phys Rev Lett 43:737
82. Odagaki T, Ogita N, Matsuda H (1975) J Phys Soc Japan 39:618
83. Pike R, Stanley HE (1981) J Phys A 14:L169
84. Romano F, Tartaglia P, Sciortino F (2007) J Phys Condens Matter 19:322101; Zaccarelli E (2007) J Phys Condens Matter 19:323101
85. Roussenq J, Coniglio A, Stauffer D (1982) J Phys Paris 43:L703
86. Saika-Voivod I, Zaccarelli E, Sciortino F, Buldyrev SV, Tartaglia P (2004) Phys Rev E 70:041401
87. Saleur H, Duplantier B (1987) Phys Rev Lett 58:2325; Duplantier B, Saleur H (1989) Phys Rev Lett 63:2536
88. Sator N (2003) Physics Reports 376:1
89. Skal AS, Shklovskii BI (1975) Sov Phys Semicond 8:1029
90. Stanley HE (1977) J Phys A 10:1211
91. Stauffer D (1976) J Chem Soc Faraday Trans 72:1354
92. Stauffer D (1981) J Phys Lett 42:L49
93. Stauffer D (1990) Physica A 168:614
94. Stauffer D (1997) Physica A 242:1. for a minireview on the multiplicity of the infinite clusters
95. Stauffer D, Aharony A (1994) Introduction to percolation theory. Taylor and Francis, London
96. Stauffer D, Coniglio A, Adam M (1982) Adv Pol Sci 44:103. For a review on percolation and gelation (special volume Polymer Networks, ed. Dusek K)
97. Stella AL, Vanderzande C (1989) Phys Rev Lett 62:1067
98. Suzuki M (1974) Progr Theor Phys Kyoto 51:1992
99. Swendsen RH, Wang JS (1987) Phys Rev Lett 58:86
100. Sykes MF, Gaunt DS (1976) J Phys A 9:2131
101. Tanaka T, Swislow G, Ohmine I (1979) Phys Rev Lett 42:1557
102. Temesvary T (1984) J Phys A Math Gen 17:1703; Janke W, Schakel AMJ (2004) Nucl Phys B 700:385; Qian X, Deng Y, Blote HWJ (2005) Phys Rev E 72:056132; Deng Y, Guo W, Blote HWJ (2005) Phys Rev E 72:016101; Balog I, Uzelac K (2007) Phys Rev E 76:011103
103. Vernon DC, Plischke M, Joos B (2001) Phys Rev E 64:031505
104. Wang JS (1981) Physica A 161:249
105. Wang JS (1989) Physica A 161:249
106. Wang JS, Swendsen R (1990) Physica A 167:565
107. Webman I, Safran S, Bug ALR (1986) Phys Rev A 33:2842
108. Weinrib A (1984) Phys Rev B 29:387; Weinrib A, Halperin BI (1983) Phys Rev B 27:413; Sahimi M, Knackstedt MA, Sheppard AP (2000) Phys Rev E 61:4920; Sahimi M, Mukhopadhyay (1996) Phys Rev E 54:3870; Makse HA, Havlin S, Schwartz M, Stanley HE (1996) Phys Rev E 53:5445
109. Wolff U (1988) Phys Rev Lett 60:1461; Wolff U (1989) Phys Lett B 228:379; Wolff U (1989) Phys Rev Lett 62:361
110. Wu F (1982) Rev Mod Phys 54:235

Books and Reviews

Grimmett G (1989) Percolation. Springer, Berlin
Sahimi M (1994) Application of percolation theory. Taylor and Francis, London

Correlations in Complex Systems

Renat M. Yulmetyev[1,2], Peter Hänggi[3]
[1] Department of Physics, Kazan State University, Kazan, Russia
[2] Tatar State University of Pedagogical and Humanities Sciences, Kazan, Russia
[3] University of Augsburg, Augsburg, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Correlation and Memory in Discrete Non-Markov Stochastic Processes
Correlation and Memory in Discrete Non-Markov Stochastic Processes Generated by Random Events
Information Measures of Memory in Complex Systems
Manifestation of Strong Memory in Complex Systems
Some Perspectives on the Studies of Memory in Complex Systems
Bibliography

Glossary

Correlation A correlation describes the degree of relationship between two or more variables. The correlations are viewed due to the impact of random factors and can be characterized by the methods of probability theory.

Correlation function The correlation function (abbreviated, as CF) represents the quantitative measure for the compact description of the wide classes of correlation in the complex systems (CS). The correlation function of two variables in statistical mechanics provides a measure of the mutual order existing between them. It quantifies the way random variables at different positions are correlated. For example in a spin system, it is the thermal average of the scalar product of the spins at two lattice points over all possible orderings.

Memory effects in stochastic processes through correlations Memory effects (abbreviated, as ME) appear at a more detailed level of statistical description of correlation in the hierarchical manner. ME reflect the complicated or hidden character of creation, the propagation and the decay of correlation. ME are produced by inherent interactions and statistical after-effects in CS. For the statistical systems ME are induced by contracted description of the evolution of the dynamic variables of a CS.

Memory functions Memory functions describe mutual interrelations between the rates of change of random variables on different levels of the statistical description. The role of memory has its roots in the natural sciences since 1906 when the famous Russian mathematician Markov wrote his first paper in the theory of Markov Random Processes. The theory is based on the notion of the instant loss of memory from the prehistory (memoryless property) of random processes.

Information measures of statistical memory in complex systems From the physical point of view time scales of correlation and memory cannot be treated as arbitrary. Therefore, one can introduce some statistical quantifiers for the quantitative comparison of these time scales. They are dimensionless and possess the statistical spectra on the different levels of the statistical description.

Definition of the Subject

As commonly used in probability theory and statistics, a correlation (also so called correlation coefficient), measures the strength and the direction of a linear relationship between two random variables. In a more general sense, a correlation or co-relation reflects the deviation of two (or more) variables from mutual independence, although correlation does not imply causation. In this broad sense there are some quantifiers which measures the degree of correlation, suited to the nature of data. Increasing attention has been paid recently to the study of statistical memory effects in random processes that originate from nature by means of non-equilibrium statistical physics. The role of memory has its roots in natural sciences since 1906 when the famous Russian mathematician Markov wrote his first paper on the theory of Markov Random Processes (MRP) [1]. His theory is based on the notion of an instant loss of memory from the prehistory (memoryless property) of random processes. In contrast, there are an abundance of physical phenomena and processes which can be characterized by statistical memory effects: kinetic and relaxation processes in gases [2] and plasma [3], condensed matter physics (liquids [4], solids [5], and superconductivity [6]) astrophysics [7], nuclear physics [8], quantum [9] and classical [9] physics, to name only a few. At present, we have a whole toolbox available of statistical methods which can be efficiently used for the analysis of the memory effects occurring in diverse physical systems. Typical such schemes are Zwanzig–Mori's kinetic equations [10,11], generalized master equations and corresponding statistical quantifiers [12,13,14,15,16,17,18], Lee's recurrence relation method [19,20,21,22,23], the generalized Langevin equation (GLE) [24,25,26,27,28,29], etc.

Here we shall demonstrate that the presence of statistical memory effects is of salient importance for the functioning of the diverse natural complex systems. Particularly, it can imply that the presence of large memory times scales in the stochastic dynamics of discrete time series can characterize catastrophical (or pathological for live systems) violation of salutary dynamic states of CS. As an example, we will demonstrate here that the emergence of strong memory time scales in the chaotic behavior of complex systems (CS) is accompanied by the likely initiation and the existence of catastrophes and crises (Earthquakes, financial crises, cardiac and brain attack, etc.) in many CS and especially by the existence of pathological states (diseases and illness) in living systems.

Introduction

A common definition [30] of a correlation measure $\rho(X, Y)$ between two random variables X and Y with the mean values $E(X)$ and $E(Y)$, and fluctuations $\delta X = X - E(X)$ and $\delta Y = Y - E(Y)$, dispersions $\sigma_X^2 = E(\delta X^2) = E(X^2) - E(X)^2$ and $\sigma_Y^2 = E(\delta Y^2) = E(Y^2) - E(Y)^2$ is

defined by:

$$\rho(X,Y) = \frac{E(\delta X\,\delta Y)}{\sigma_X\,\sigma_Y}\,,$$

where E is the expected value of the variable. Therefore we can write

$$\rho(X,Y) = \frac{[E(XY) - E(X)\,E(Y)]}{(E(X^2) - E(X)^2)^{1/2}\,(E(Y^2) - E(Y)^2)^{1/2}}\,.$$

Here, a correlation can be defined only if both of the dispersions are finite and both of them are nonzero. Due to the Cauchy–Schwarz inequality, a correlation cannot exceed 1 in absolute value. Consequently, a correlation assumes it maximum at 1 in the case of an increasing linear relationship, or −1 in the case of a decreasing linear relationship, and some value in between in all other cases, indicating the degree of linear dependence between the variables. The closer the coefficient is either to −1 or 1, the stronger is the correlation between the variables. If the variables are independent then the correlation equals 0, but the converse is not true because the correlation coefficient detects only linear dependencies between two variables.

Since the absolute value of the sample correlation must be less than or equal to 1 the simple formula conveniently suggests a single-pass algorithm for calculating sample correlations. The square of the sample correlation coefficient, which is also known as the coefficient of determination, is the fraction of the variance in σ_x that is accounted for by a linear fit of x_i to σ_y. This is written

$$R^2_{xy} = 1 - \frac{\sigma^2_{y|x}}{\sigma^2_y}\,,$$

where $\sigma^2_{y|x}$ denotes the square of the error of a linear regression of x_i on y_i in the equation $y = a + bx$,

$$\sigma^2_{y|x} = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2$$

and σ^2_y denotes just the dispersion of y.

Note that since the sample correlation coefficient is symmetric in x_i and y_i, we will obtain the same value for a fit to y_i:

$$R^2_{xy} = 1 - \frac{\sigma^2_{x|y}}{\sigma^2_x}\,.$$

This equation also gives an intuitive idea of the correlation coefficient for random (vector) variables of higher dimension. Just as the above described sample correlation coefficient is the fraction of variance accounted for by the fit of a 1-dimensional linear submanifold to a set of 2-dimensional vectors (x_i, y_i), so we can define a correlation coefficient for a fit of an m-dimensional linear submanifold to a set of n-dimensional vectors. For example, if we fit a plane $z = a + bx + cy$ to a set of data (x_i, y_i, z_i) then the correlation coefficient of z to x and y is

$$R^2 = 1 - \frac{\sigma^2_{z|xy}}{\sigma^2_z}\,.$$

Correlation and Memory in Discrete Non-Markov Stochastic Processes

Here we present a non-Markov approach [31,32] for the study of long-time correlations in chaotic long-time dynamics of CS. For example, let the variable x_i be defined as the R-R interval or the time distance between nearest, so called R peaks occurring in a human electrocardiogram (ECG). The generalization will consist in taking into account non-stationarity of stochastic processes and its further applications to the analysis of the heart-rate-variability.

We should bear in mind, that one of the key moments of the spectral approach in the analysis of stochastic processes consists in the use of normalized time correlation function (TCF)

$$a_0(t) = \frac{\langle\langle \mathbf{A}(T)\,\mathbf{A}(T+t)\rangle\rangle}{\langle \mathbf{A}(T)^2\rangle}\,. \tag{1}$$

Here the time T indicates the beginning of a time serial, $\mathbf{A}(t)$ is a state vector of a complex system as defined below in Eq. (5) at t, $|\mathbf{A}(t)|$ is the length of vector $\mathbf{A}(t)$, the double angular brackets indicate a scalar product of vectors and an ensemble averaging. The ensemble averaging is, of course needed in Eq. (1) when correlation and other characteristic functions are constructed. The average and scalar product becomes equivalent when a vector is composed of elements from a discrete-time sampling, as done later. Here a continuous formalism is discussed for convenience. However further, since Sect. "Correlation and Memory in Discrete Non-Markov Stochastic Processes" we shall consider only a case of discrete processes.

The above-stated designation is true only for stationary systems. In a non-stationary case Eq. (1) is not true and should be changed. The concept of TCF can be generalized in case of discrete non-stationary sequence of signals. For this purpose the standard definition of the correlation coefficient in probability theory for the two random signals X and Y must be taken into account

$$\rho = \frac{\langle\langle \mathbf{XY}\rangle\rangle}{\sigma_X\,\sigma_Y}\,,\quad \sigma_X = \langle|\mathbf{X}|\rangle\,,\quad \sigma_Y = \langle|\mathbf{Y}|\rangle\,. \tag{2}$$

In Eq. (2) the multi-component vectors **X**, **Y** are determined by fluctuations of signals x and y accordingly, σ_X^2, σ_Y^2 represent the dispersions of signals **x** and **y**, and values $|\mathbf{X}|$, $|\mathbf{Y}|$ represent the lengths of vectors **X**, **Y**, correspondingly. Therefore, the function

$$a(T,t) = \frac{\langle\langle \mathbf{A}(T)\,\mathbf{A}(T+t)\rangle\rangle}{\langle|\mathbf{A}(T)|\rangle\,\langle|\mathbf{A}(T+t)|\rangle} \tag{3}$$

can serve as the generalization of the concept of TCF (1) for non-stationary processes $\mathbf{A}(T+t)$. The non-stationary TCF (3) obeys the conditions of the normalization and attenuation of correlation

$$a(T,0) = 1\,, \quad \lim_{t\to\infty} a(T,t) = 0\,.$$

Let us note, that in a real CS the second limit, typically, is not carried out due possible occurrence nonergodocity (meaning that a time average does not equal its ensemble average). According to the Eqs. (1) and (3) for the quantitative description of non-stationarity it is convenient to introduce a function of non-stationarity

$$\gamma(T,t) = \frac{\langle|\mathbf{A}(T+t)|\rangle}{\langle|\mathbf{A}(T)|\rangle} = \left\{\frac{\sigma^2(T+t)}{\sigma^2(T)}\right\}^{1/2}\,. \tag{4}$$

One can see that this function equals the ratio of the lengths of vectors of final and initial states. In case of stationary process the dispersion does not vary with the time (or its variation is very weak). Therefore the following relations

$$\sigma(T+t) = \sigma(T)\,, \quad \gamma(T,t) = 1 \tag{5}$$

hold true for the stationary process.

Due to the condition (5) the following function

$$\Gamma(T,t) = 1 - \gamma(T,t) \tag{6}$$

is suitable in providing a dynamic parameter of non-stationarity. This dynamic parameter can serve as a quantitative measure of non-stationarity of the process under investigation. According to Eqs. (4)–(6) it is reasonable to suggest the existence of three different elementary classes of non-stationarity

$$|\Gamma(T,t)| = |1-\gamma(T,t)| = \left\{\begin{array}{ll} \ll 1, & \text{weak non-stationarity} \\ \sim 1, & \text{intermediate non-stationarity} \\ \gg 1, & \text{strong non-stationarity} \end{array}\right\}\,. \tag{7}$$

The existence of dynamic parameter of non-stationarity makes it possible to determine, on-principle, the type of non-stationarity of the underlying process and to find its spectral characteristics from the experimental data base. We intend to use Eqs. (4), (6), (7) for the quantitative description of effects of non-stationarity in the investigated temporary series of R-R intervals of human ECG's for healthy people and patients after myocardial infarction (MI).

Statistical Theory of Non-Stationary Discrete Non-Markov Processes in Complex Systems

Here we shall extend the original results of the statistical theory of discrete non-Markov processes in complex systems, developed recently in [31], for the case of non-stationary processes. The theory [31] is developed on the basis of first principles and represents a discrete finite-difference analogy for complex systems of well known Zwanzig–Mori's kinetic equations [10,11,12,13,14,15,16,17,18] used in the statistical physics of condensed matter.

We examine a discrete stochastic process $X(T+t)$, where $t = m\tau$

$$X = \{x(T), x(T+\tau), x(T+2\tau), \ldots, x(T+k\tau), \ldots, x(T+(N-1)\tau)\}\,, \tag{8}$$

where T is the beginning of the time and τ is a discretization time. The normalized time correlation function (TCF)

$$a(t) = \frac{1}{(N-m)\,\sigma^2}\sum_{j=0}^{N-1-m} \delta x(T+j\tau)\,\delta x(T+(j+m)\,\tau) \tag{9}$$

yields a convenient measure to analyze the dynamic properties of complex systems. Herein, we used the variance σ^2, the fluctuation $\delta x(T+j\tau)$, which in terms of the the mean value $\langle x\rangle$ reads:

$$\delta x_j = \delta x(T+j\tau) = x(T+j\tau) - \langle x\rangle\,,$$
$$\sigma^2 = \frac{1}{(N-m)}\sum_{j=0}^{N-1-m}\{\delta x(T+j\tau)\}^2\,, \tag{10}$$

$$\langle x\rangle = \frac{1}{(N-m)}\sum_{j=0}^{N-1-m} x(T+j\tau)\,. \tag{11}$$

The discrete time t is given as $t = m\tau$.

In general, the mean value, the variance and TCF in (9), (10) and (11) is dependent on the numbers m and N. All indicated values cease to depend on numbers m and N for stationary processes when $m \ll N$. The definition of TCF in Eq. (9) is true only for stationary processes.

Next, we shall try to take into account this important dependence. With this purpose we shall form two k-dimensional vectors of state by the process (8):

$$\begin{aligned} \mathbf{A}_k^0 &= (\delta x_0, \delta x_1, \delta x_2, \ldots, \delta x_{k-1}) \,, \\ \mathbf{A}_{m+k}^m &= (\delta x_m, \delta x_{m+1}, \delta x_{m+2}, \ldots, \delta x_{m+k-1}) \,. \end{aligned} \tag{12}$$

When a vector of a state is composed of elements from a discrete-time sampling, the average and scalar product in Eq. (1) become equivalent. In an Euclidean space of vectors of state (12) TCF $a(t)$

$$a(t) = \frac{\langle \mathbf{A}_{N-1-m}^0 \, \mathbf{A}_{N-1}^m \rangle}{(N-m)\{\sigma(N-m)\}^2} = \frac{\langle \mathbf{A}_{N-1-m}^0 \, \mathbf{A}_{N-1}^m \rangle}{|\mathbf{A}_{N-1-m}^0|^2} \tag{13}$$

describes the correlation of two different states of the system ($t = m\tau$). Here the brackets $\langle \ldots \rangle$ indicate the scalar product of the two vectors. The dimension dependence of the corresponding vectors is also taken into account in the variance $\sigma = \sigma(N-m)$. As a matter of fact TCF $a(t) = \cos\vartheta$, where ϑ is the angle between the two vectors from Eq. (12). Let's introduce a unit vector of dimension $(N-m)$ in the following way:

$$\mathbf{n} = \frac{\mathbf{A}_{N-1-m}^0}{\sqrt{(N-m)\sigma^2}} \,. \tag{14}$$

Then, the TCF $a(t)$ (9) is given by

$$a(t) = \langle \mathbf{n}(0) \, \mathbf{n}(t) \rangle \,. \tag{15}$$

From the above discussion it is evident that Eqs. (13)–(15) are true for the stationary processes only. In case of non-stationary processes it is necessary to redefine TCF, taking into account the non-stationarity in the variance σ^2 in a line with Eqs.(2)–(7). For this purpose we shall redefine a unit vector of the final state as following

$$\mathbf{n}(t) = \frac{\mathbf{A}_{N-1}^m(t)}{|\mathbf{A}_{N-1}^m(t)|} \,. \tag{16}$$

For non-stationary processes it is convenient to write the TCF as the scalar product of the two unit vectors of the initial and final states

$$a(t) = \langle \mathbf{n}(0) \, \mathbf{n}(t) \rangle = \frac{\langle \mathbf{A}_{N-1-m}^0(0) \, \mathbf{A}_{N-1}^m(t) \rangle}{|\mathbf{A}_{N-1-m}^0(0)| \, |\mathbf{A}_{N-1}^m(t)|} \,. \tag{17}$$

Now we shall turn to the the dynamics of a non-stationary stochastic process. The equation of motion of a the random process x_j can be written in a finite-difference form for $0 \le j \le N-1$ [31,32] in the following way

$$\frac{\mathrm{d}x_j}{\mathrm{d}t} \Rightarrow \frac{\Delta \delta x_j}{\Delta t} = \frac{\delta x_j(t+\tau) - \delta x_j(t)}{\tau} \,. \tag{18}$$

Then it is convenient to define the discrete evolution single step operator $\hat{U}$ as following:

$$x(T+(j+1)\,\tau) = \hat{U}(T+(j+1)\,\tau, T+j\tau)\, x(T+j\tau). \tag{19}$$

In the case of stationary process we can rewrite the equation of motion (18) in a more simple form

$$\frac{\Delta \delta x_j}{\Delta t} = \tau^{-1}\{\hat{U}(\tau) - 1\}\, \delta x_j \,. \tag{20}$$

The invariance of the mean value $\langle x \rangle$ is taken into account in an Eq. (20)

$$\langle x \rangle = \hat{U}(\tau)\langle x \rangle \,, \quad \{\hat{U}(\tau) - 1\}\langle x \rangle = 0 \,. \tag{21}$$

In case of a non-stationary process it is necessary to turn to the equation of motion for vector of the final state $\mathbf{A}_{m+k}^m(t)$ $(k = N-1-m)$

$$\frac{\Delta \mathbf{A}_{m+k}^m(t)}{\Delta t} = i\hat{L}(t,\tau)\, \mathbf{A}_{m+k}^m(t) \,, \tag{22}$$

where Liouville's quasioperator is

$$\hat{L}(t,\tau) = (i\tau)^{-1}\{\hat{U}(t+\tau, t) - 1\} \,. \tag{23}$$

It is well known that, in general, a stochastic trajectory does not obey a linear equation, so the general evolution operator and Liouville's quasioperator should probably be non-linear. Furthermore, in statistical physics the Liouville's operator acts upon the probability densities of dynamical variables, as well upon the variables itself like in Mori's paper [12]. The evolution of the density would be indeed linear. But Mori used the Liouville operator in the quantum equation of motion in [12]. In line with Mori [12] Eqs. (20), (22) can be considered as formal and exact equations of the motion of a complex system.

Thus, due to the Eqs. (17), (22) and (23) we may take into account the non-stationarity of the stochastic process. Towards this goal let's introduce the linear projection operator in Euclidean space of the state vectors

$$\Pi\, \mathbf{A}(t) = \frac{\mathbf{A}(0)\rangle\langle \mathbf{A}(0)\, \mathbf{A}(t)\rangle}{|\mathbf{A}(0)|^2} \,, \quad \Pi = \frac{\mathbf{A}(0)\rangle\langle \mathbf{A}(0)}{\langle \mathbf{A}(0)\, \mathbf{A}(0)\rangle} \,, \tag{24}$$

where angular brackets in numerator present the boundaries of action for the scalar product.

For the analysis of the dynamics of the stochastic process $\mathbf{A}(t)$ the vector $\mathbf{A}_k^0(0)$ from (12) can be considered as a vector of the initial state $\mathbf{A}(0)$, and vector $\mathbf{A}_{m+k}^m(t)$ from (12) at value $m+k = N-1$ can be considered as the vector of the final state $\mathbf{A}(t)$.

It is necessary to note that the projection operator (24) has the required property of idem-potency $\Pi^2 = \Pi$. The

presence of operator Π allows one to introduce the mutually supplementary projection operator P:

$$P = 1 - \Pi\,, \quad P^2 = P\,, \quad \Pi P = P\Pi = 0\,. \tag{25}$$

It is necessary to remark, that both projectors Π and P are linear and can be recorded for the fulfillment of operations in the particular Euclidean space. Due to the property (17) and Eq. (4) it is easy to obtain the required TCF:

$$\begin{aligned} \Pi\,\mathbf{A}(t) &= \Pi\,\mathbf{A}^m_{m+k}(t) \\ &= \mathbf{A}^0_k(0)\langle \mathbf{n}^0_k(0)\,\mathbf{n}^m_{k+m}(t)\rangle \gamma_1(t) \\ &= \mathbf{A}^0_k(0)\,a(t)\,\gamma_1(t)\,, \\ \gamma_1(t) &= \frac{|\mathbf{A}^m_{m+k}(t)|}{|\mathbf{A}^0_m(0)|}\,. \end{aligned} \tag{26}$$

Therefore the projector Π generates a unit vector along the vector of the final state $\mathbf{A}(t)$ and makes its projection onto the initial state vector $\mathbf{A}(0)$.

The existence of a pair of two mutually supplementary projection operators Π and P allows one to carry out the splitting of Euclidean space of vectors $A(\mathbf{A}(0), \mathbf{A}(t) \in A)$ into a straight sum of two mutually supplementary subspaces in the following way

$$A = A' \dot{+} A''\,, \quad A' = \Pi A\,, \quad A'' = PA\,. \tag{27}$$

Substituting Eq. (27) in Eq. (23) we find Liouville's quasioperator $\hat{L}$ in a matrix form

$$\hat{L} = \hat{L}_{11} + \hat{L}_{12} + \hat{L}_{21} + \hat{L}_{22}\,, \tag{28}$$

where the matrix elements are introduced

$$\begin{aligned} &\hat{L}_{11} = \Pi\hat{L}\Pi\,, \quad \hat{L}_{12} = \Pi\hat{L}P\,, \\ &\hat{L}_{21} = P\hat{L}\Pi\,, \quad \hat{L}_{22} = P\hat{L}P\,. \end{aligned} \tag{29}$$

The Euclidean space of values of Liouville's quasioperator $W = \hat{L}A$ will be generated by the vectors $\mathbf{W}$ of dimension $k-1$

$$\begin{aligned} &(\mathbf{W}(0) \in W\,,\ \mathbf{W}(t) \in W) \\ &W = W' \dot{+} W''\,, \quad W' = \Pi W\,, \quad W'' = PW\,. \end{aligned} \tag{30}$$

Matrix elements $\hat{L}_{ij}$ of the contracted description

$$\hat{L} = \begin{pmatrix} \hat{L}_{11} & \hat{L}_{12} \\ \hat{L}_{21} & \hat{L}_{22} \end{pmatrix} \tag{31}$$

are acting in the following way:

$\hat{L}_{11}$– from a subspace A' to subspace W',

$\hat{L}_{12}$– from A'' to W',

$\hat{L}_{21}$– from W' to W'' and

$\hat{L}_{22}$– from A'' to W''.

The projection operators Π and P provide the contracted description of the stochastic process. Splitting the dynamic Eq. (22) into two equations in the two mutually supplementary Euclidean subspaces (see, for example [11]), we find

$$\frac{\Delta\mathbf{A}'(t)}{\Delta t} = i\hat{L}_{11}\,\mathbf{A}'(t) + i\hat{L}_{12}\,\mathbf{A}''(t)\,, \tag{32}$$

$$\frac{\Delta\mathbf{A}''(t)}{\Delta t} = i\hat{L}_{21}\,\mathbf{A}'(t) + i\hat{L}_{22}\,\mathbf{A}''(t)\,. \tag{33}$$

Following [31,32] it is necessary to eliminate first the irrelevant part $\mathbf{A}''(t)$ in order to simplify Liouville's Eq. (22) and then to write a closed equation for relevant part $\mathbf{A}'(t)$. According to [32] that can be achieved by a series of successive steps (for example, see Eqs. (32)–(36) in [32]). First a solution to Eq. (33) for the first step can be obtained in a form

$$\begin{aligned} \frac{\Delta\mathbf{A}''(t)}{\Delta t} &= \frac{\mathbf{A}''(t+\tau) - \mathbf{A}''(t)}{\tau} \\ &= i\hat{L}_{21}\,\mathbf{A}'(t) + i\hat{L}_{22}\,\mathbf{A}''(t)\,, \\ \mathbf{A}''(t+\tau) &= \mathbf{A}''(t) + i\tau\,\hat{L}_{21}\,\mathbf{A}'(t) + i\tau\,\hat{L}_{22}\,\mathbf{A}''(t) \\ &= \{1 + i\tau\,\hat{L}_{22}\}\,\mathbf{A}''(t) + i\tau\,\hat{L}_{21}\,\mathbf{A}'(t) \\ &= U_{22}(t+\tau, t)\,\mathbf{A}''(t) + i\tau\,\hat{L}_{21}(t+\tau, t)\,\mathbf{A}'(t)\,. \end{aligned} \tag{34}$$

We next can derive a finite-difference kinetic equation of a non-Markov type for TCF $a(t = m\tau)$

$$\frac{\Delta a(t)}{\Delta t} = \lambda_1 a(t) - \tau\Lambda_1 \sum_{j=0}^{m-1} M_1(t - j\tau)\,a(j\tau)\,. \tag{35}$$

Here, λ_1 is a eigenvalue, Λ_1 is a relaxation parameter of Liouville's quasioperator $\hat{L}$

$$\begin{aligned} \lambda_1 &= i\,\frac{\langle \mathbf{A}^0_k(0)\,\hat{L}\,\mathbf{A}^0_k(0)\rangle}{|\mathbf{A}^0_k(0)|^2}\,, \\ \Lambda_1 &= \frac{\langle \mathbf{A}^0_k(0)\,\hat{L}_{12}\,\hat{L}_{21}\,\mathbf{A}^0_k(0)\rangle}{|\mathbf{A}^0_k(0)|^2} = \frac{\langle \mathbf{A}^0_k(0)\,\hat{L}^2\,\mathbf{A}^0_k(0)\rangle}{|\mathbf{A}^0_k(0)|^2}\,, \end{aligned} \tag{36}$$

The angular brackets indicate here a scalar product of new state vectors. Function $M_1(t - j\tau)$ on the right side of Eq. (35) represents a modified memory function (MF) of the first order

$$M_1(t - j\tau) = \frac{\gamma_1(t - j\tau)}{\gamma_1(t)}\,m_1(t - j\tau)\,. \tag{37}$$

For stationary processes the function $\gamma_1(t)$ approaches unity. Then the memory functions $M_1(t)$ and $m_1(t)$ co-

incide with each other. The latter equation is the first kinetic finite-difference equation for TCF. It is remarkable, that the non-Markovity, discretization and non-stationarity of stochastic process can be considered explicitly. Due to the presence of non-stationarity both in TCF and in the first memory function this equation generalizes our results recently obtained in [31].

Following the projection technique described above, we arrive at a chain of connected kinetic finite-difference equations of a non-Markov type for the normalized short memory functions $m_n(t)$ in Euclidean space of state vectors of dimension $(k-n)$ $(t = m\tau, n \geq 1)$

$$\begin{aligned}\frac{\Delta m_n(t)}{\Delta t} &= \lambda_{n+1}\, m_n(t) - \tau \Lambda_{n+1} \\ &\quad \times \sum_{j=0}^{m-1} m_{n+1}(j\tau)\, m_n(t-j\tau) \\ &\quad \times \left\{ \frac{\gamma_{n+1}(j\tau)\gamma_{n+1}(t-j\tau)}{\gamma_n(t)} \right\} ,\end{aligned} \tag{38}$$

$$\begin{aligned} m_{n+1}(t) &= \frac{\langle \mathbf{W}_{n+1}(0)\, \mathbf{W}_{n+1}(t)\rangle}{|\mathbf{W}_{n+1}(0)||\mathbf{W}_{n+1}(t)|} , \\ \gamma_n(j\tau) &= \left\{ \frac{|\mathbf{W}_n(j\tau)|}{|\mathbf{W}_n(0)|} \right\} . \end{aligned} \tag{39}$$

Here, $\gamma_n(j\tau)$ is the nth order of the non-stationarity function.

The set of all memory functions $m_1(t), m_2(t), m_3(t), \ldots$ allows one to describe non-Markov processes and statistical memory effects in the considered non-stationary system. For the particular case we obtain a more simple form for the set of equations for the first three short memory functions, namely $(t = m\tau)$:

$$\begin{aligned}\frac{\Delta a(t)}{\Delta t} &= -\tau \Lambda_1 \sum_{j=0}^{m-1} m_1(j\tau) \left\{ \frac{\gamma_1(j\tau)\gamma_1(t-j\tau)}{\gamma_1(t)} \right\} \\ &\quad \times a(t-j\tau) + \lambda_1 a(t) , \\ \frac{\Delta m_1(t)}{\Delta t} &= -\tau \Lambda_2 \sum_{j=0}^{m-1} m_2(j\tau) \left\{ \frac{\gamma_2(j\tau)\gamma_2(t-j\tau)}{\gamma_2(t)} \right\} \\ &\quad \times m_1(t-j\tau) + \lambda_2 m_1(t) , \\ \frac{\Delta m_2(t)}{\Delta t} &= -\tau \Lambda_3 \sum_{j=0}^{m-1} m_3(j\tau) \left\{ \frac{\gamma_3(j\tau)\gamma_3(t-j\tau)}{\gamma_3(t)} \right\} \\ &\quad \times m_2(t-j\tau) + \lambda_3 m_2(t) .\end{aligned} \tag{40}$$

Here the relaxation parameters Λ_1, Λ_2 and Λ_3 have already been determined and the non-stationarity functions $\gamma_n(t)$ have been introduced earlier. By analogy with Eq. (6) we can introduce a set of dynamic parameters of non-stationarity (PNS) for the arbitrary nth relaxation level

$$\Gamma_n(T,t) = 1 - \gamma_n(t) = 1 - \gamma_n(T,t) . \tag{41}$$

The whole set of values of dynamic PNS $\gamma_n(t)$ determines the broad spectrum of non-stationarity effects of the considered process.

The obtained equations are similar to the well known Zwanzig–Mori's kinetic equations [10,11,12,13,14,15,16,17,18] used in non-equilibrium statistical physics of condensed matters. Let us point out three essential distinctions of our Eqs. (40) from the results in [10,11,12]. In Zwanzig–Mori's theory the key moment in the analysis of considered physical systems is the presence of a Hamiltonian and an operation of a statistical averaging carried out with the help of quantum density operator or classic Gibbs distribution function [33]. In our examined case, both the Hamiltonian and the distribution function are absent. There are exact classic or quantum equations of motion in physics; so Liouville's equation and Liouville's operator are useful in many applications. The motion of individual particles and whole statistic system is described by variables varying in continuous time. Therefore, for physical systems it is possible to use effectively the methods of integro-differential calculus, based on the mathematically accustomed (but from the physical point of view difficult for understanding) representation of infinitesimal variations of values of coordinates and time. By nature, the monitored time evolution of most complex systems is discrete. As well known, discretization is inherent in a wide variety both of classical and quantum complex systems. This forces us to abandon the concept of an infinite small values and continuity and instead turn to discrete-difference schemes. And, at last, the third feature is connected with incorporating the issue of non-stationary processes into our theory. The Zwanzig–Mori theory is typically applied only for stationary processes. Due to the introduction of normalized vectors of states and the use of the appropriate projection technique [13] our theory allows to take into account non-stationary processes as well. The latter ones can be described by the non-Markov kinetic equations together with the introduction of the set of non-stationarity functions.

The non-stationary theory [32] put forward here differs from the stationary case [31]. The external structure of the kinetic equations remains invariant; they represent the kinetic equations with memory. However, the functions and the parameters, which are included in these equations, appreciably differ from each other. As we already remarked above, non-stationarity effects enter both, in the functions $\gamma_n(t)$ and in spectral and kinetic parameters.

Correlation and Memory in Discrete Non-Markov Stochastic Processes Generated by Random Events

Here we shall find a chain of the kinetic interconnected finite-difference equations for a discrete correlation function $a(n)$ and memory functions $M_s(n)$ in the linear scale of events $E = \{\xi_1, \xi_2, \xi_3, \ldots, \xi_N\}$.

The Basic Assumptions and Concepts of the Theory of Discrete Non-Markov Stochastic Processes of the Events Correlations

As an example we shall consider the time variations of the total X-ray flux of an astrophysical object at a succession of events:

$$E = \{\xi_1, \xi_2, \xi_3, \ldots, \xi_k, \ldots, \xi_N\} , \tag{42}$$

where ξ_i is an event, which occurs at time instant t_i, where $i = 1, \ldots, N$ counts the event number.

The average value $\langle E \rangle$, fluctuations $\delta\xi$ and dispersion σ^2 for the set of N events are obtained as:

$$\begin{aligned} \langle E \rangle &= \frac{1}{N} \sum_{i=1}^{N} \xi_i , \delta\xi_i = \xi_i - \langle E \rangle , \\ \sigma^2 &= \frac{1}{N} \sum_{i=1}^{N} \delta\xi_i^2 = \frac{1}{N} \sum_{i=1}^{N} \{\xi_i - \langle E \rangle\}^2 . \end{aligned} \tag{43}$$

According to [35,36,37,38], for the description of the dynamical properties of the studied system we introduce the correlation dependence of the discrete set of events (see Eq. (42)) using the CF:

$$a(n) = \frac{1}{(N-m)\,\sigma^2} \sum_{i=1}^{N-m} \delta\xi_i \, \delta\xi_{i+m} . \tag{44}$$

Here $n = m\Delta n$, $\Delta n = 1$ is the discretization step. The function $a(n)$, which emerges in this way, is the "event" correlation function (ECF). The normalized ECF must obey the conditions of normalization and of the attenuation of correlation, i. e.: $\lim_{n\to 1} a(n) = 1$, $\lim_{n\to\infty} a(n) = 0$. We remark, however, that the second condition for the case the physical complex systems is typically not observed (at $N \gg 0$). It is necessary to note that in [18] the correlation function for the aftershock events has been introduced:

$$C(n + n_W, n_W) = \frac{[\langle t_{n+n_W} \, t_{n_W} \rangle - \langle t_{n+n_W} \rangle \langle t_{n_W} \rangle]}{\left(\sigma^2_{n+n_W} \sigma^2_{n_W}\right)^{1/2}} ,$$

where the averages and the variance are given by

$$\begin{aligned} \langle t_m \rangle &= \frac{1}{N} \sum_{k=o}^{N-1} t_{m+k} , \\ \langle t_m \, t'_m \rangle &= \frac{1}{N} \sum_{k=o}^{N-1} t_{m+k} \, t'_{m+k} , \quad \text{and} \\ \sigma^2_m &= \langle t^2_m \rangle - \langle t_m \rangle^2 , \end{aligned}$$

respectively.

By the direct analogy of [31,32,35] we use the finite-difference Liouville's equation of motion in the event scale for describing the evolution of discrete set of events Eq. (11), (13):

$$\frac{\Delta\xi_i(n)}{\Delta n} = i\,\widehat{L}(n,1)\,\xi_i(n) . \tag{45}$$

Here $\xi_i(n+1) = U(n+1,n)\,\xi_i(n)$, $U(n+1,n)$ is the "event" evolution operator. It determines the shift in linear event scale to one step Δn. The evolution operator $U(n+1,n)$ and Liouville's quasioperator $\widehat{L}(n,1)$ can be made explicit by writing: $\widehat{L}(n,1) = (i\Delta n)^{-1}\,(U(n+1,n) - 1)$.

Let's represent the set of values of the dynamical variable $\delta\xi_j = \delta\xi(j\Delta n)$, $j = 1, \ldots, N$ as the k-component vector of system state in linear Euclidean space:

a) the vector of initial state of studied complex system:

$$\mathbf{A}^1_k = \{\delta\xi_1, \delta\xi_2, \delta\xi_3, \ldots, \delta\xi_k\} , \tag{46}$$

b) the vector of final system's state, which is shifted on the m events along the event scale:

$$\mathbf{A}^m_{m+k} = \{\delta\xi_{m+1}, \delta\xi_{m+2}, \delta\xi_{m+3}, \ldots, \delta\xi_{m+k}\} , \tag{47}$$

where $1 \le k \le N$. The vectors of initial and final states, which are submitted in a similar way, are very convenient for analyzing the dynamics of the observed discrete stochastic processes with the help of discrete non-Markov processes.

To represent the ECF in a more compact form, we use the expression for the scalar product of vectors $\langle \mathbf{A}^1_k \cdot \mathbf{A}^m_{m+k} \rangle = \sum_{j=1}^{k} A^1_j \, A^m_{m+j}$, and the Eqs. (44), (46) and (47):

$$a(n) = \frac{\langle \mathbf{A}^1_k(1)\, \mathbf{A}^m_{k+m}(n) \rangle}{\langle |\mathbf{A}^1_k(1)|^2 \rangle} . \tag{48}$$

Construction of Chain of Finite-Difference Non-Markov Kinetic Equations for the Events Correlation

Let us consider the finite-difference Liouville's equation (Eq. (44)) for the vector of final system states:

$$\frac{\Delta \mathbf{A}^m_{m+k}(n)}{\Delta n} = i\,\widehat{L}(n,1)\,\mathbf{A}^m_{m+k}(n)\,. \tag{49}$$

We introduce the projection operator Π, which projects the final vector $\mathbf{A}^m_{m+k}(n)$ on the direction of initial vector, and also the orthogonal operator P. The operators Π and P possess the following properties: $\Pi = |\mathbf{A}^1_k(1)\rangle\langle\mathbf{A}^1_k(1)|/\langle|\mathbf{A}^1_k(1)|^2\rangle$, $\Pi^2 = \Pi$, $\mathrm{P} = 1 - \Pi$, $\mathrm{P}^2 = \mathrm{P}$, $\Pi\mathrm{P} = \mathrm{P}\Pi = 0$. They are idempotent and mutually complementary.

The initial ECF $a(n)$ (Eq. (48)) can be derived by means of projecting the vector of final states $\mathbf{A}^m_{m+k}(n)$ on the vector of initial state $\mathbf{A}^1_k(1)$:

$$\Pi\,\mathbf{A}^m_{m+k}(n) = \frac{\mathbf{A}^1_k(1)\langle\mathbf{A}^1_k(1)\,\mathbf{A}^m_{m+k}(n)\rangle}{\langle|\mathbf{A}^0_k|^2\rangle} = \mathbf{A}^1_k(1)\,a(n)\,. \tag{50}$$

The operators Π and P split Euclidean vector space $A(k)$ into two mutually orthogonal subspaces:

$$\begin{aligned} &A(k) = A'(k) + A''(k)\,, \quad A'(k) = \Pi A(k)\,,\\ &A''(k) = \mathrm{P}A(k)\,, \quad \mathbf{A}^m_{m+k} \in A(k)\,. \end{aligned} \tag{51}$$

As a result the finite-difference Liouville's Eq. (47) can be represented as a system of 2 equations into mutually orthogonal linear subspaces:

$$\frac{\Delta \mathbf{A}'(n)}{\Delta n} = i\,\widehat{L}_{11}\,\mathbf{A}'(n) + i\,\widehat{L}_{12}\,\mathbf{A}''(n)\,, \tag{52}$$

$$\frac{\Delta \mathbf{A}''(n)}{\Delta n} = i\,\widehat{L}_{21}\,\mathbf{A}'(n) + i\,\widehat{L}_{22}\,\mathbf{A}''(n)\,. \tag{53}$$

Here $\widehat{L}_{ij} = \Pi_i\,\widehat{L}\,\Pi_j$ are the matrix elements of Liouville's quasioperator:

$$\begin{aligned} &\widehat{L} = \widehat{L}_{11} + \widehat{L}_{12} + \widehat{L}_{21} + \widehat{L}_{22}\,,\\ &\widehat{L}_{11} = \Pi\widehat{L}\Pi\,, \quad \widehat{L}_{12} = \Pi\widehat{L}\mathrm{P}\,,\\ &\widehat{L}_{21} = \mathrm{P}\widehat{L}\Pi\,, \quad \widehat{L}_{22} = \mathrm{P}\widehat{L}\mathrm{P}\,. \end{aligned} \tag{54}$$

To solve the system of Eqs. (52), (53) we eliminate the non-reducible part, which contains $\mathbf{A}''(n)$ and derive the self-contained equation for the reducible part $\mathbf{A}'(n)$. In doing so we solve the Eq. (52) step-by-step and shall substitute the obtained solution into the Eq. (53). As a result we arrive at the closed kinetic equation:

$$\begin{aligned} \frac{\Delta \mathbf{A}'(n+m\Delta n)}{\Delta n} &= i\,\widehat{L}_{11}\,\mathbf{A}'(n+m\Delta n)\\ &\quad + i\,\widehat{L}_{12}\{1 + i\Delta n\,\widehat{L}_{22}\}^m\,\mathbf{A}''(n)\\ &\quad - \widehat{L}_{12}\sum_{j=1}^{m}\{1 + i\Delta n\,\widehat{L}_{22}\}^j\,\Delta n\\ &\quad \times \widehat{L}_{21}\,\mathbf{A}'(n + [m-j]\Delta n)\,. \end{aligned} \tag{55}$$

By use of projection operators Π and P we found the closed finite-difference kinetic equation of non-Markov type for the initial ECF:

$$\frac{\Delta a(n)}{\Delta n} = i\lambda_1\,a(n) - \Delta n\,\Lambda_1\sum_{j=1}^{m} M_1(j\Delta n)\,a(n - j\Delta n)\,. \tag{56}$$

As $\Delta n = 1$, solution of the last equation must be following:

$$a(n+1) = \{i\lambda_1 + 1\}\,a(n) - \Lambda_1\sum_{j=1}^{m} M_1(j)\,a(n-j)\,. \tag{57}$$

Here λ_1 is the proper value of Liouville's quasioperator $\widehat{L}$, Λ_1 is the relaxation parameter, which dimension is square of frequency, $M_1(j\Delta n)$ is the normalized memory function of the first order:

$$\lambda_1 = \frac{\langle A^1_k(1)\,\widehat{L}\,A^1_k(1)\rangle}{\langle|A^1_k(1)|^2\rangle}\,,$$

$$\Lambda_1 = \frac{\langle A^1_k\,\widehat{L}_{12}\,\widehat{L}_{21}\,A^1_k(1)\rangle}{|A^1_k(1)|^2\rangle}\,,$$

$$M_1(j\Delta n) = \frac{\langle A^1_k(1)\,\widehat{L}_{12}(1 + i\Delta n\,\widehat{L}_{22})^j\,\widehat{L}_{21}\,A^1_k(1)\rangle}{\langle A^1_k(1)\,\widehat{L}_{12}\,\widehat{L}_{21}\,A^1_k(1)\rangle}\,.$$

To obtain the finite-difference kinetic equation for the normalized event memory function of first order and, further, for the higher $(s-1)$th orders as well, we have to repeat the foregoing procedure step-by-step. However, we shall make use of the Gram–Schmidt orthogonalization procedure [16]:

$$\langle\mathbf{W}_s\mathbf{W}_p\rangle = \delta_{sp}\,\langle|\mathbf{W}_s|^2\rangle\,. \tag{58}$$

Where δ_{sp} is a Kronecker's symbol. Now we shall derive the recurrence formula $\mathbf{W}_s = \mathbf{W}_s(n)$ for defining the set of the orthogonal dynamic variables:

$$\begin{aligned} &\mathbf{W}_0 = \mathbf{A}^1_k\,,\\ &\mathbf{W}_1 = \{i\widehat{L} - \lambda_1\}\mathbf{W}_0\,,\\ &\mathbf{W}_2 = \{i\widehat{L} - \lambda_2\}\mathbf{W}_1 - \Lambda_1\mathbf{W}_0, \ldots \end{aligned} \tag{59}$$

According to the foregoing formulas we can introduce the succession of projection operators $\Pi_s = \Pi_1^{(s)}$ and the set of mutually complementary projectors $\mathrm{P}_s = 1 - \Pi_s$, which possess the following properties:

$$\Pi_s = \frac{|\mathbf{W}_s\rangle\langle\mathbf{W}_s|}{\langle|\mathbf{W}_s|^2\rangle}\,, \qquad \Pi_s^2 = \Pi_s\,,$$
$$\mathrm{P}_s^2 = \mathrm{P}_s\,, \qquad \Pi_s\,\mathrm{P}_s = \mathrm{P}_s\,\Pi_s = 0\,,$$
$$\Pi_s\,\Pi_p = \delta_{sp}\,\Pi_s\,, \qquad \mathrm{P}_s\,\mathrm{P}_p = \delta_{sp}\,\mathrm{P}_s\,.$$

Each of these operators pairs Π_s, P_s splits the corresponding Euclidean vector space $\mathbf{W}_s$ into the two mutual complementary subspaces: $W_s = W_s' + W_s'$, $W_s' = \Pi_s W_s$, $W_s'' = P_s W_s$. Using the projection operator technique for the next orthogonal variables $\mathbf{W}_s$, we shall obtain the chain of interconnected kinetic finite-difference equations of the non-Markov type for the normalized correlation functions of the $(s-1)$th order:

$$\frac{\Delta M_1(n)}{\Delta n} = i\,\lambda_2\,M_1(n) - \Lambda_2\sum_{j=1}^{m} M_2(j)\,M_1(n-j)\,,$$
$$\ldots\,,$$
$$\frac{\Delta M_{s-1}(n)}{\Delta n} = i\,\lambda_s\,M_{s-1}(n) - \Lambda_s\sum_{j=1}^{m} M_{s-1}(j)\,M_s(n-j)\,. \tag{60}$$

In these equations the normalized events memory function of the first order: $M_1(n) = \langle\mathbf{W}_1(1 + i\Delta n\widehat{L})^m\mathbf{W}_1\rangle/\langle|\mathbf{W}_1|^2\rangle$, memory function of the $(s-1)$th order: $M_{s-1}(n) = \langle\mathbf{W}_{s-1}(1 + i\Delta n\widehat{L})^m\mathbf{W}_{s-1}\rangle/\langle|\mathbf{W}_{s-1}|^2\rangle$, the proper value of the Liouville's quasioperator $\widehat{L}$: $\lambda_s = \langle\mathbf{W}_s\widehat{L}\mathbf{W}_s\rangle/\langle|\mathbf{W}_s|^2\rangle$ and the relaxation parameter $\Lambda_s = \langle|\mathbf{W}_s|^2\rangle/\langle|\mathbf{W}_{s-1}|^2\rangle$ are introduced.

The foregoing finite-difference kinetic Eqs. (60) present the generalization of the statistical theory [31,32,35] for the case of event correlations in discrete stochastic evolution of non-Hamilton complex systems.

Information Measures of Memory in Complex Systems

As an information measures of memory it is useful to apply different dimensionless quantifiers. As a first measure we use the frequency dependence of non-Markovity parameter. This measure was introduced in [31] and it is defined as:

$$\varepsilon_i(\nu) = \left\{\frac{\mu_{i-1}(\nu)}{\mu_i(\nu)}\right\}^{1/2}\,. \tag{61}$$

Here, $\mu_i(\nu)$ denotes the frequency power spectrum of memory function of the ist order $M_i(n)$: $\mu_i(\nu) = |\Delta n\sum_{n=1}^{N} M_i(n)\cos(2\pi n\nu)|^2$. The non-Markovity parameter $\varepsilon_i(\nu)$ along with the memory functions enables us to characterize quantitatively the statistical memory effects in discrete complex systems of various nature. Because the functions $\mu_i(\nu)$ exist for each of the ith levels of relaxation, we obtain the statistical spectrum of parameters: $\varepsilon_i(\nu)$, $i = 1, 2, 3, \ldots$.

Alternatively, a study of 'memory' in physiological time series for electroencephalographic (EEG) and magnetoencephalographic (MEG) signals, both of healthy subjects and patients (including epilepsy patients) has been based on the detrended-fluctuation analysis (DFA) [39,40].

The characterization of memory *per se* is based on a set of dimensionless statistical quantifiers which are capable for measuring the memory strength which is inherent to the complex dynamics.

According to [41] a second set an information memory measure can be constructed as follows:

$$\delta_i(\nu) = \left|\frac{\tilde{M}_i'(\nu)}{\tilde{M}_{i+1}'(\nu)}\right|\,.$$

Here, $\mu_i(\nu) = |\tilde{M}_i(\nu)|^2$ denotes the power spectrum of the corresponding memory function $M_i(t)$, $\tilde{M}_i'(\nu) = \mathrm{d}\tilde{M}_i(\nu)/\mathrm{d}\nu$ and $\tilde{M}_i(\nu)$ is the Fourier transform of the memory function $M_i(t)$. The measures $\varepsilon_i(\nu)$ are suitable for the quantification of the memory effects on a relative scale whereas the second set $\delta_i(\nu)$ proves to be useful for quantifying the amplification of relative memory effects occurring on different complexity levels. Both measures provide statistical criteria for comparison between the relaxation time scales and memory time scales of the process under consideration. For values obeying $\{\varepsilon, \delta\} \gg 1$ one can observe a complex dynamics characterized by the short-ranged temporal memory scales. In the memoryless limit these processes assume a δ-like memory with parameters $\varepsilon, \delta \to \infty$. When $\{\varepsilon, \delta\} > 1$ one deals with a situation with moderate memory strength, and the case where both $\varepsilon, \delta \sim 1$ typically constitutes a more regular and robust random process exhibiting strong memory features.

Manifestation of Strong Memory in Complex Systems

A fundamental role of the strong and weak memory in the functioning of the human organism and seismic phenomena can be illustrated by the example of some situations examined next. We will consider some examples of

the time series for both living and for seismic systems. It is necessary to note that a comprehensive analysis of the experimental data includes the calculation and the presentation of corresponding phase portraits in some planes of the dynamic orthogonal variables, the autocorrelation time functions, the memory time functions and their frequency power spectra, etc. However, we start out by calculating two statistical quantifiers, characterizing two informational measures of memory: the parameters $\epsilon_1(\omega)$ and $\delta_1(\omega)$.

Figures 1 and 3 present the results of experimental data of pathological states of human cardiovascular systems (CVS). Figure 2 depicts the analysis for the seismic observation. Figures 4 and 5 indicate the memory effects for the patients with Parkinson disease (PD), and the last two Figs. 6, 7 demonstrate the key role of the strength of memory in the case of time series of patients suffering from photosensitive epilepsy which are contrasted with signals taken from healthy subjects. All these cases convincingly display the crucial role of the statistical memory in the functioning of complex (living and seismic) systems.

A characteristic role of the statistical memory can be detected from Fig. 1 for the typical representatives taken from patients from four different CVS-groups: (a) for healthy subject, (b) for a patient with rhythm driver migration, (c) for a patient after myocardial infarction (MI), (d) for a patient after MI with subsequent sudden cardiac death (SSCD). All these data were obtained from the short time series of the dynamics of RR-intervals from the electric signals of the human ECG's. It can be seen here that significant memory effects typically lead to the long-time correlations in the complex systems. For healthy we observe weak memory effects while and large values of the measure memory $\epsilon_1(\omega = 0) \approx 25$. The strong memory and the long memory time (approximately, 10 times more) are being observed with the help of 3 patient groups: with RDM (rhythm driver migration) (b), after MI (c) and after MI with SSCD (d).

Figure 2 depicts the strong memory effects presented in seismic phenomena. By a transition from the steady state of Earth ((a), (b) and (c)) to the state of strong earthquake (EQ) ((d), (e), and (f)) a remarkable amplification

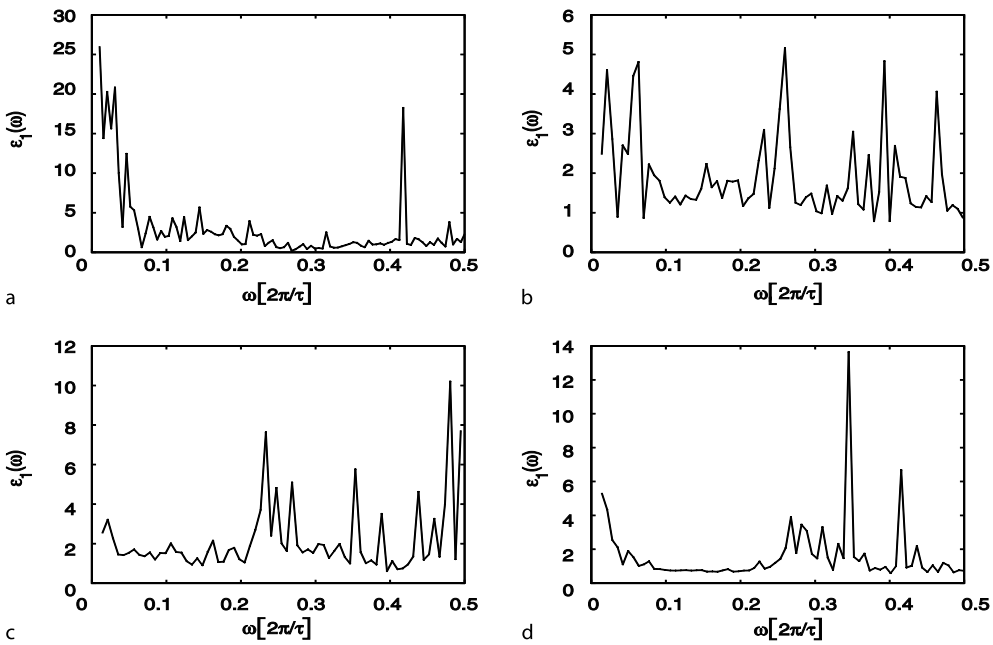

Correlations in Complex Systems, Figure 1

Frequency spectrum of the first information measure of memory (first point in the statistical spectrum on non-Markovity parameter) $\varepsilon_1(\omega)$ for the fourth cardiac patient groups from the short time series of RR-intervals: healthy subject (a), patient with rhythm driver migration (RDM) (b), patient after myocardial infarction (MI) (c), and patient after MI with subsequent sudden cardiac death (SCD) (d). The frequency is marked in terms of units of τ^{-1}. All spectra reveal the miscellaneous faces of statistical memory's strength. For the healthy subject one can see Markov effects and weak memory. For other three cases of cardiac diseases we note the diverse displays of strong memory. The strong memory has been accompanied by the spikes of the weak memory: for RDM on the all frequency regions, for patient with MI for the middle and high frequencies and for patient after MI with SSCD only for high frequencies. From Fig. 7 in [104]

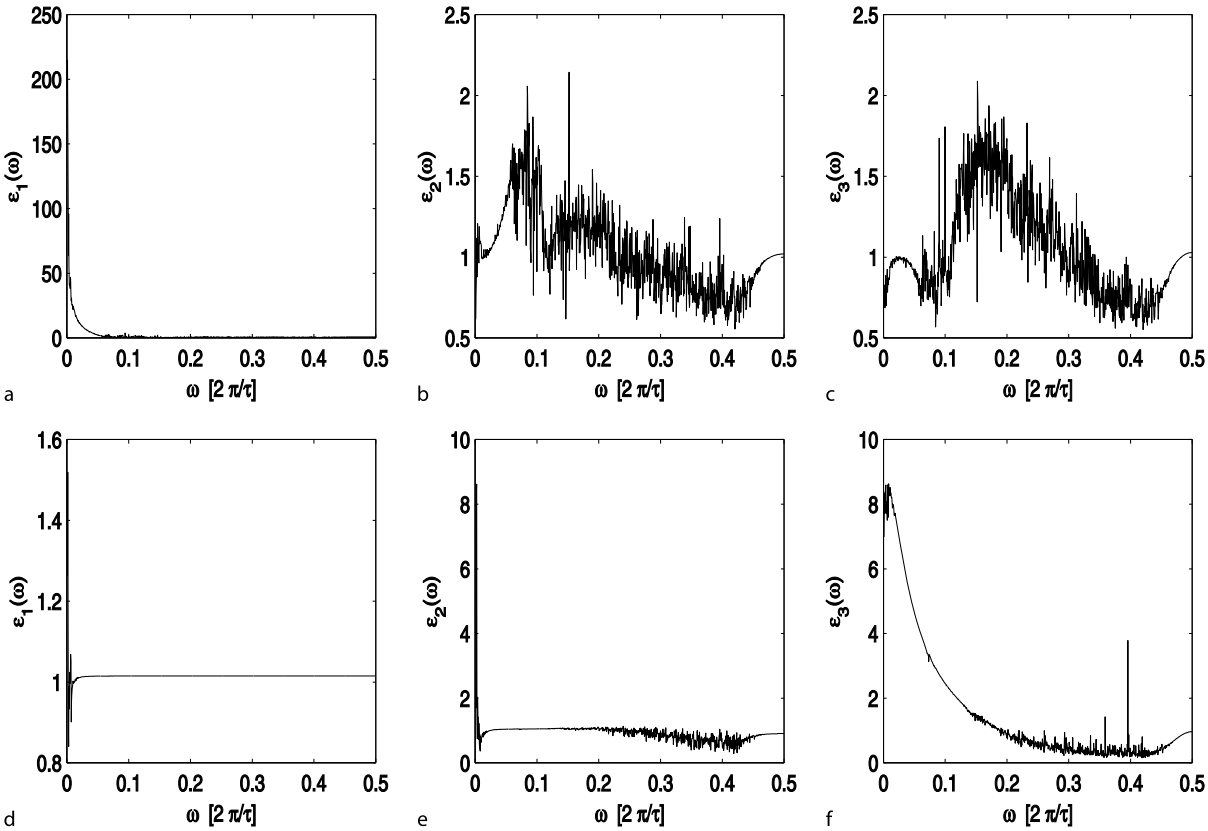

Correlations in Complex Systems, Figure 2

Frequency spectra of the first three points of the first measure of memory (non-Markovity parameters) $\varepsilon_1(\omega)$, $\varepsilon_2(\omega)$, and $\varepsilon_3(\omega)$ for the seismic phenomena: a, b, c long before the strong Earthquake (EQ) for the steady state of Earth and d, e, f during the strong EQ. Markov and quasi-Markov behavior of seismic signals with manifestation of the weak memory is observed only for ε_1 in state before the strong EQ. All remaining cases b, c, d and e relate to non-Markov processes. Strong non-Markovity and strong memory is typical for case d (state during the strong EQ). In behavior of $\varepsilon_2(\omega)$ and $\varepsilon_3(\omega)$ one can see a transition from quasi-Markovity (at low frequencies) to strong non-Markovity (at high frequencies). From Fig. 6 in [105]

of memory effects is highly visible. The term amplification refers to the appearance of strong memory and the prolongation of the memory correlation time in the seismic system. Recent study show that discrete non-Markov stochastic processes and long-range memory effects play a crucial role in the behavior of seismic systems. An approach, permitting us to obtain an algorithm of strong EQ forecasting and to differentiate technogenic explosions from weak EQs, can be developed thereupon.

Figure 3 demonstrates an intensification of memory effects of one order at the transition from healthy people ((a), (b) and (c)) to patient suffering from myocardial infarction. The figures were calculated from the long time series of the RR-intervals dynamics from the human ECG's. The zero frequency values $\epsilon_1(\omega = 0)$ at $\omega = 0$ sharply reduced, approximately of the size of one order for patient as compared to healthy subjects.

Figures 4 and 5 illustrate the behavior for patients with Parkinson's disease. Figure 4 shows time recording of the pathological tremor velocity in the left index finger of a patient with Parkinson's disease (PD) for eight diverse pathological cases (with or without medication, with or without deep brain stimulation (DBS), for various DBS, medication and time conditions). Figure 5, arranged in accordance with these conditions, displays a wide variety of the memory effects in the treatment of PD's patients. Due to the large impact of memory effects this observation permits us to develop an algorithm of exact diagnosis of Parkinson's disease and a calculation of the quantitative parameter of the quality of treatment. A physical role of the strong and long memory correlation time enables us to extract a vital information about the states of various patient on basis of notions of correlation and memory times.

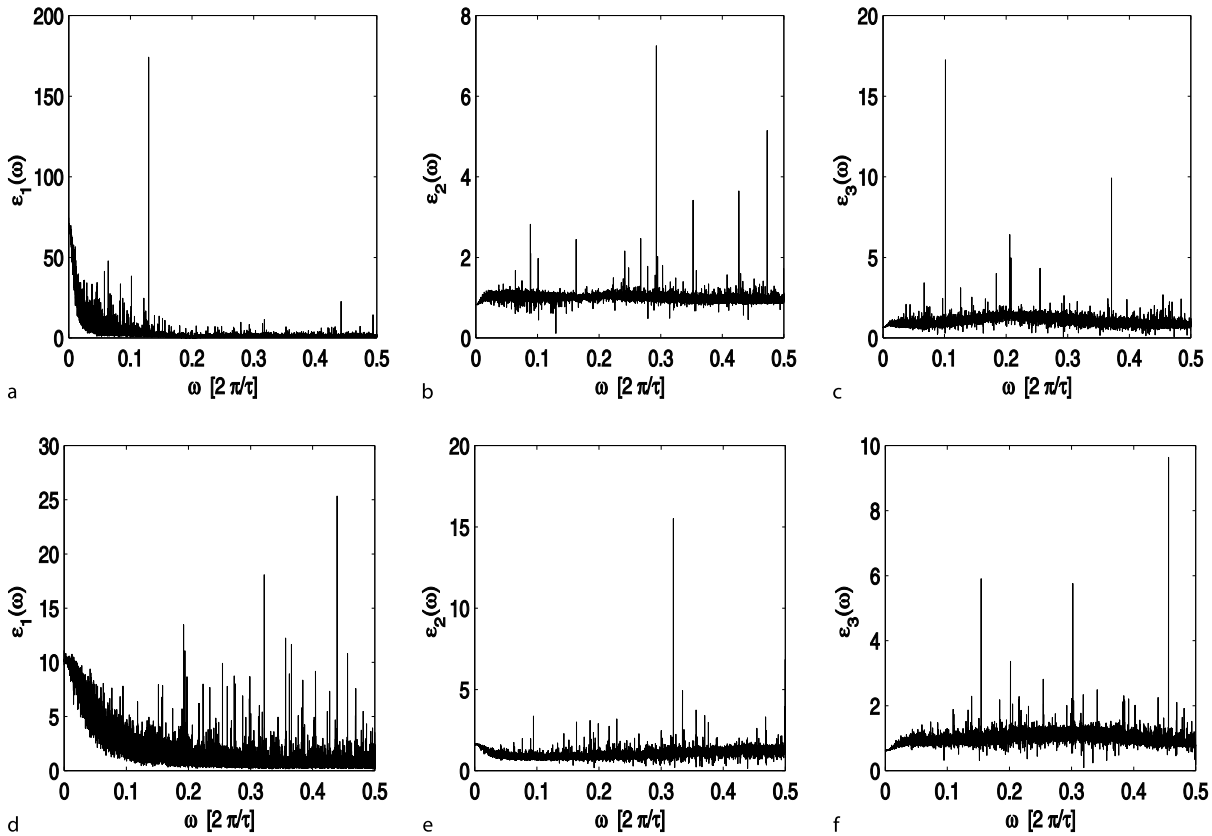

Correlations in Complex Systems, Figure 3

The frequency dependence of the first three points of non-Markovity parameter (NMP) for the healthy person (a), (b), (c) and patient after myocardial infarction (MI) (d), (e), (f) from the time dynamics of RR-intervals of human ECG's for the case of the long time series. In the spectrum of the first point of NMP $\varepsilon_1(\omega)$ there is an appreciable low-frequency (long time) component, which concerns the quasi-Markov processes. Spectra NMP $\varepsilon_2(\omega)$ and NMP $\varepsilon_3(\omega)$ fully comply with non-Markov processes within the whole range of frequencies. From Fig. 6 in [106]

According to Figs. 6 and 7 specific information about the physiological mechanism of photosensitive epilepsy (PSE) was obtained from the analysis of the strong memory effects via the registration the neuromagnetic responses in recording of magnetoencephalogram (MEG) of the human brain core. Figure 6 presents the topographic dependence of the first level of the second memory measure $\delta_1(\omega = 0; n)$ for the healthy subjects in the whole group (upper line) vs. patients (lower line) for red/blue combination of the light stimulus. This topographic dependence of $\varepsilon_1(\omega = 0; n)$ depicted in Fig. 6 clearly demonstrates the existence of long-range time correlation. It is accompanied by a sharp increase of the role of the statistical memory effects in the all MEG's sensors with sensor numbers $n = 1, 2, \ldots, 61$ of the patient with PSE in comparison with healthy peoples. A sizable difference between the healthy subject and a subject with PSE occurs.

To emphasize the role of strong memory one can continue studying the topographic dependence in terms of the novel informational measure, the index of memory, defined as:

$$\nu(n) = \frac{\delta_1^{\text{healthy}}(0; n)}{\delta_1^{\text{patient}}(0; n)}\,, \tag{62}$$

see in Fig. 7.

This measure quantifies the detailed memory effects in the individual MEG sensors of the patient with PSE versus the healthy group. A sharp increase of the role of the memory effects in the stochastic behavior of the magnetic signals is clearly detected in sensor numbers $n = 10, 46, 51, 53$ and 59. The observed points of MEG sensors locate the regions of a protective mechanism against PSE in a human organism: frontal (sensor 10),

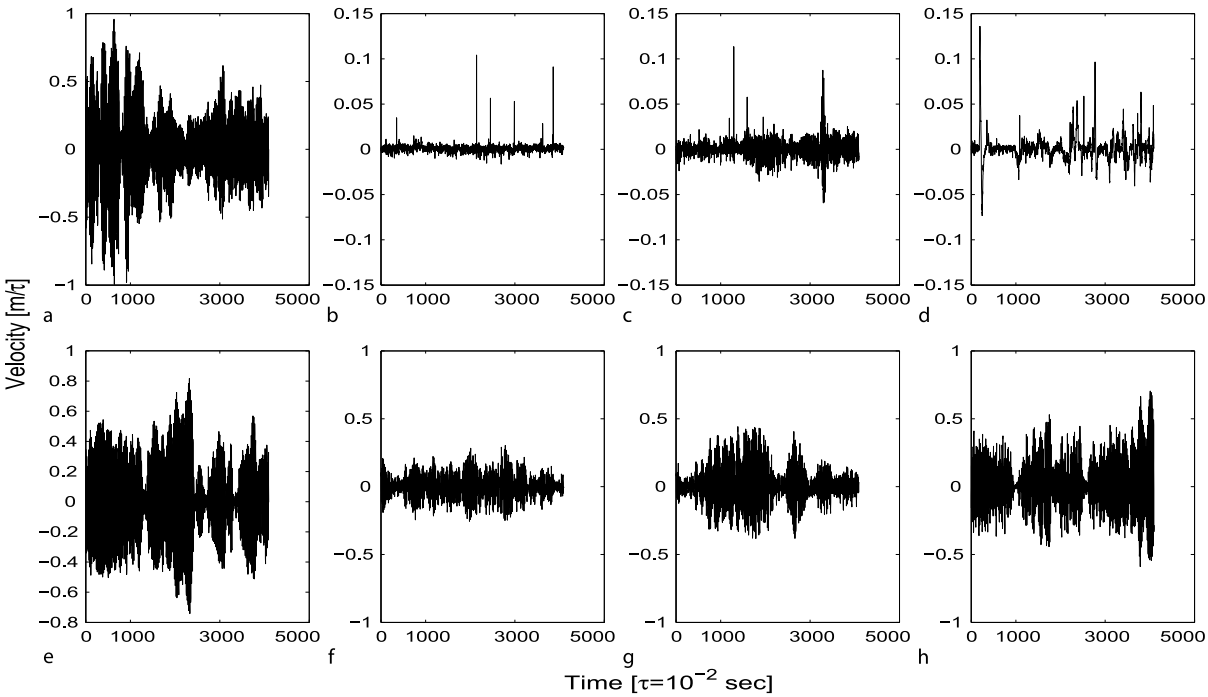

Correlations in Complex Systems, Figure 4

Pathological tremor velocity in the left index finger of the sixth patient with Parkinson's disease (PD). The registration of Parkinsonian tremor velocity is carried out for the following conditions: a "OFF-OFF" condition (no any treatment), b "ON-ON" condition (using deep brain stimulation (DBS) by electromagnetic stimulator and medicaments), c "ON-OFF" condition (DBS only), d "OFF-ON" condition (medicaments (L-Dopa) only), e–h the "15 OFF", "30 OFF", "45 OFF", "60 OFF" conditions – the patient's states 15 (30, 45, 60) minutes after the DBS is switched off, no treatment. Let's note the scale of the pathological tremor amplitude (see the vertical scale). Such representation of the time series allows us to note the increase or the decrease of pathological tremor. From Fig. 1 in [107]

occipital (sensors 46, 51 and 53) and right parietal (sensor 59) regions. The early activity in these sensors may reflect a protective mechanism suppressing the cortical hyperactivity due to the chromatic flickering.

We remark that some early steps towards understanding the normal and various catastrophical states of complex systems have already been taken in many fields of science such as cardiology, physiology, medicine, neurology, clinical neurophysiology, neuroscience, seismology and so forth. With the underlying systems showing fractal and complicated spatial structures numerous studies applying the linear and nonlinear time series analysis to various complex systems have been discussed by many authors. Specifically the results obtained shows evidence of the significant nonlinear structure evident in the registered signals in the control subjects, whereas nonlinearity for the patients and catastrophical states were not detected. Moreover the couplings between distant parts and regions were found to be stronger for the control subjects. These prior findings are leading to the hypothesis that the real normal complex systems are mostly equipped with significantly nonlinear subsystems reflecting an inherent mechanism which stems against a synchronous excitation vs. outside impact or inside disturbances. Such nonlinear mechanisms are likely absent in the occurrence of catastrophical or pathological states of the complex systems.

From the physical point of view our results can be used as a toolbox for testing and identifying the presence or absence of various memory effects as they occur in complex systems. The set of our memory quantifiers is uniquely associated with the appearance of memory features in the chaotic behavior of the observed signals. The registration of the behavior belonging to these indicators, as elucidated here, is of beneficial use for detecting the catastrophical or pathological states in the complex systems. There exist alternative quantifiers of different nature as well, such as the Lyapunov's exponent, Kolmogorov–Sinai entropy, correlation dimension, etc., which are widely used in nonlinear dynamics and relevant applications. In the present context, we have found out that the employed memory measures are not only convenient for the analysis but are

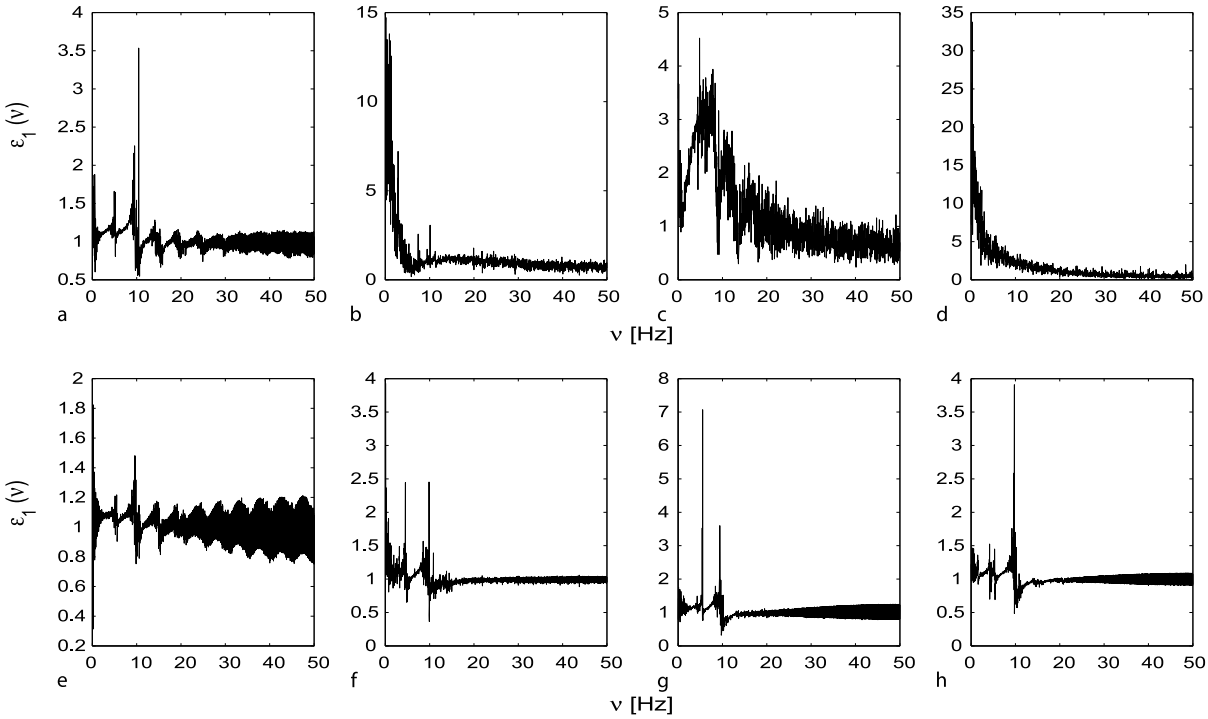

Correlations in Complex Systems, Figure 5

The frequency dependence of the first point of the non-Markovity parameter $\varepsilon_1(\nu)$ for pathological tremor velocity in the patient. As an example, the sixth patient with Parkinson's disease is chosen. The figures are submitted according to the arrangement of the initial time series. The characteristic low-frequency oscillations are observed in frequency dependence (a, e–h), which get suppressed under medical influence (b–d). The non-Markovity parameter reflects the Markov and non-Markov components of the initial time signal. The value of the parameter on zero frequency $\varepsilon_1(0)$ reflects the total dynamics of the initial time signal. The maximal values of parameter $\varepsilon_1(0)$ correspond to small amplitudes of pathological tremor velocity. The minimal values of this parameter are characteristic of significant pathological tremor velocities. The comparative analysis of frequency dependence $\varepsilon_1(\nu)$ allows us to estimate the efficiency of each method of treatment. From Fig. 5 in [107]

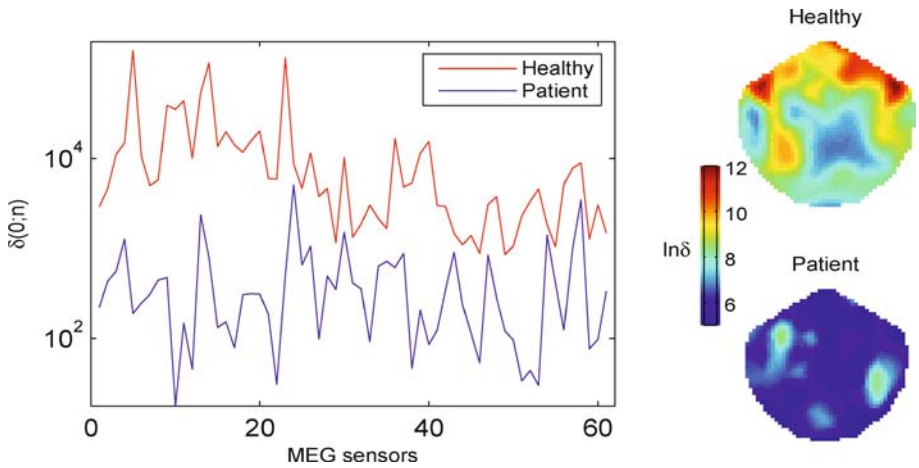

Correlations in Complex Systems, Figure 6

The topographic dependence of the first point of the second measure of memory $\delta_1(\omega = 0; n)$ for the healthy on average in the whole group (*upper line*) vs. patient (*lower line*) for R/B combination of the light stimulus. One can note the singular weak memory effects for the healthy on average in sensors with No. 5, 23, 14, 11 and 9

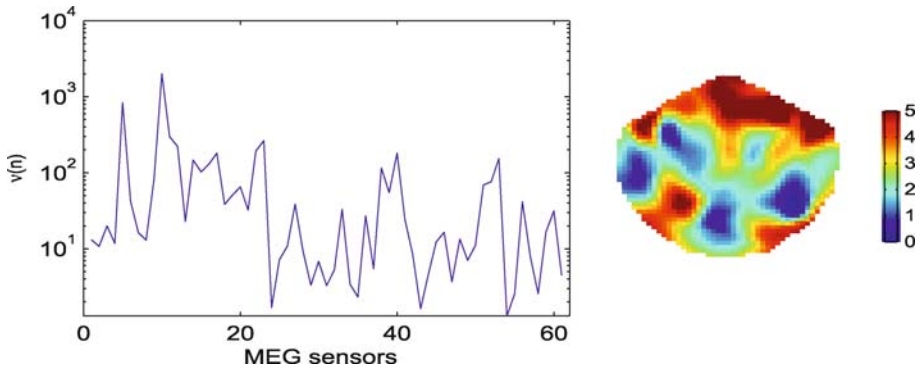

Correlations in Complex Systems, Figure 7
The topographic dependence of the memory index $\nu(n) = \nu_1(n;0)$ for the the whole group of healthy on average vs. patient for an R/B combination of the light stimulus. Strong memory in patient vs. healthy appears clearly in sensors with No. 10, 5, 23, 40 and 53

also ideally suitable for the identification of anomalous behavior occurring in complex systems. The search for other quantifiers, and foremost, the ways of optimization of such measures when applied to the complex discrete time dynamics presents a real challenge. Especially this objective is met when attempts are made towards the identification and quantification of functioning in complex systems. This work presents initial steps towards the understanding of basic foundation of anomalous processes in complex systems on the basis of a study of the underlying memory effects and connected with this, the occurrence of long lasting correlations.

Some Perspectives on the Studies of Memory in Complex Systems

Here we present a few outlooks on the fundamental role of statistical memory in complex systems. This involves the issue of studying cross-correlations. The statistical theory of stochastic dynamics of cross-correlation can be created on the basis of the mentioned formalism of projection operators technique in the linear space of random variables. As a result we obtain the cross-correlation memory functions (MF's) revealing the statistical memory effects in complex systems. Some memory quantifiers will appear simultaneously which will reflect cross-correlation between different parts of CS. Cross-correlation MF's can be very useful for the analysis of the weak and strong interactions, signifying interrelations between the different groups of random variables in CS. Besides that the cross-correlation can be important for the problem of phase synchronization, which can find a unique way of studying of synchronization phenomena in CS that has a special importance when studying aspects of brain and living systems dynamics.

Some additional information about the strong and weak memory effects can be extracted from the observation of correlation in CS in the random event's scales. Similar effects are playing a crucial role in the differentiation between stochastic phenomena within astrophysical systems, for example, in galaxies, pulsars, quasars, microquasars, lacertides, black holes, etc. One of the most important area of application of developed approach is a bispectral and polyspectral analysis for the diverse CS. From the mathematical point of view a correct definition of the spectral properties in the functional space of random functions is quite important. A variety of MF's arises in the quantitative analysis of the fine details of memory effects in a nonlinear manner. The quantitative control of the treatment quality in the diverse areas of medicine and physiology may be one of the important biomedical application of the manifestation of the strong memory effects.

These and other features of memory effects in CS call for an advanced development of brain studies on the basis of EEG's and MEG's data, cardiovascular, locomotor and respiratory human systems, in the development of the control system of information flows in living systems. An example is the prediction of strong EQ's and the clear differentiation between the occurrence of weak EQ's and the technogenic explosions, etc.

In conclusion, we hope that the interested reader becomes invigorated by this presentation of correlation and memory analysis of the inherent nonlinear system dynamics of varying complexity. He can find further details how significant memory effects typically cause long time correlations in complex systems by inspecting more closely some of the published items in [42–103].

There are the relationships between standard fractional and polyfractal processes and long-time correlation

in complex systems, which were explained in [39,40,44,45,46,49,53,54,60,62,64,76,79,83,84,94] in detail.

Example of using the Hurst exponent over time for testing the assertion that emerging markets are becoming more efficient can be found in [51].

While over 30 measures of complexity have been proposed in the research literature one can distinguish [42,55,66,81,89,99] with the specific designation of long-time correlation and memory effects.

The papers [48,57] are focused on long range correlation processes that are nonlocal in time and whence show memory effects.

The statistical characterization of the nonstationarities in real-world time series is an important topic in many fields of research and some numerous methods of characterizing nonstationary time series were offered in [59,65,84].

Long-range correlated time series have been widely used in [52,61,63,68,74] for the theoretical description of diverse phenomena.

Example of the study an anatomy of extreme events in a complex adaptive system can be found in [67].

Approaches for modeling long-time and long-range correlation in complex systems from time series are investigated and applied to different examples in [50,56,69,70,73,75,80,82,86,100,101,102].

Detecting scale invariance and its fundamental relationships with statistical structures is one of the most relevant problems among those addressed correlation analysis [47,71,72,91].

Specific long-range correlation in complex systems are the object of active research due to its implications in the technology of materials and in several fields of scientific knowledge with the use of quantified histograms [78], decrease of chaos in heart failure [85], scaling properties of ECG's signals fluctuations [87], transport properties in correlated systems [88] etc.

It is demonstrated in [43,92,93] how ubiquity of the long-range correlations is apparent in typical and exotic complex statistical systems with application to biology, medicine, economics and to time clustering properties [95,98].

The scale-dependent wavelet and spectral measures for assessing cardiac dysfunction have been used in [97].

In recent years the study of an increasing number of natural phenomena that appear to deviate from standard statistical distributions has kindled interest in alternative formulations of statistical mechanics [58,101].

At last, papers [77,90] present the samples of the deep and multiple interplay between discrete and continuous long-time correlation and memory in complex systems and the corresponding modeling the discrete time series on the basis of physical Zwanzig–Mori's kinetic equation for the Hamilton statistical systems.

Bibliography

Primary Literature

1. Markov AA (1906) Two-dimensional Brownian motion and harmonic functions. Proc Phys Math Soc Kazan Imp Univ 15(4):135–178; in Russian
2. Chapman S, Couling TG (1958) The mathematical theory of nonuniform gases. Cambridge University Press, Cambridge
3. Albeverio S, Blanchard P, Steil L (1990) Stochastic processes and their applications in mathematics and physics. Kluwer, Dordrecht
4. Rice SA, Gray P (1965) The statistical mechanics of simple liquids. Interscience, New York
5. Kubo R, Toda M, Hashitsume N, Saito N (2003) Statistical physics II: Nonequilibrium statistical mechanics. In: Fulde P (ed) Springer Series in Solid-State Sciences, vol 31. Springer, Berlin, p 279
6. Ginzburg VL, Andryushin E (2004) Superconductivity. World Scientific, Singapore
7. Sachs I, Sen S, Sexton J (2006) Elements of statistical mechanics. Cambridge University Press, Cambridge
8. Fetter AL, Walecka JD (1971) Quantum theory of many-particle physics. Mc Graw-Hill, New York
9. Chandler D (1987) Introduction to modern statistical mechanics. Oxford University Press, Oxford
10. Zwanzig R (2001) Nonequilibrium statistical mechanics. Cambridge University Press, Cambridge
11. Zwanzig R (1961) Memory effects in irreversible thermodynamics. Phys Rev 124:983–992
12. Mori H (1965) Transport, collective motion and Brownian motion. Prog Theor Phys 33:423–455; Mori H (1965) A continued fraction representation of the time correlation functions. Prog Theor Phys 34:399–416
13. Grabert H, Hänggi P, Talkner P (1980) Microdynamics and nonlinear stochastic processes of gross variables. J Stat Phys 22:537–552
14. Grabert H, Talkner P, Hänggi P (1977) Microdynamics and time-evolution of macroscopic non-Markovian systems. Z Physik B 26:389–395
15. Grabert H, Talkner P, Hänggi P, Thomas H (1978) Microdynamics and time-evolution of macroscopic non-Markovian systems II. Z Physik B 29:273–280
16. Hänggi P, Thomas H (1977) Time evolution, correlations and linear response of non-Markov processes. Z Physik B 26:85–92
17. Hänggi P, Talkner P (1983) Memory index of first-passage time: A simple measure of non-Markovian character. Phys Rev Lett 51:2242–2245
18. Hänggi P, Thomas H (1982) Stochastic processes: Time-evolution, symmetries and linear response. Phys Rep 88:207–319
19. Lee MH (1982) Orthogonalization process by recurrence relations. Phys Rev Lett 49:1072–1072; Lee MH (1983) Can the velocity autocorrelation function decay exponentially? Phys Rev Lett 51:1227–1230
20. Balucani U, Lee MH, Tognetti V (2003) Dynamic correlations. Phys Rep 373:409–492

21. Hong J, Lee MH (1985) Exact dynamically convergent calculations of the frequency-dependent density response function. Phys Rev Lett 55:2375–2378
22. Lee MH (2000) Heisenberg, Langevin, and current equations via the recurrence relations approach. Phys Rev E 61:3571–3578; Lee MH (2000) Generalized Langevin equation and recurrence relations. Phys Rev E 62:1769–1772
23. Lee MH (2001) Ergodic theory, infinite products, and long time behavior in Hermitian models. Phys Rev Lett 87(1–4):250601
24. Kubo R (1966) Fluctuation-dissipation theorem. Rep Progr Phys 29:255–284
25. Kawasaki K (1970) Kinetic equations and time correlation functions of critical fluctuations. Ann Phys 61:1–56
26. Michaels IA, Oppenheim I (1975) Long-time tails and Brownian motion. Physica A 81:221–240
27. Frank TD, Daffertshofer A, Peper CE, Beek PJ, Haken H (2001) H-theorem for a mean field model describing coupled oscillator systems under external forces. Physica D 150:219–236
28. Vogt M, Hernandez R (2005) An idealized model for nonequilibrium dynamics in molecular systems. J Chem Phys 123(1–8):144109
29. Sen S (2006) Solving the Liouville equation for conservative systems: Continued fraction formalism and a simple application. Physica A 360:304–324
30. Prokhorov YV (1999) Probability and mathematical statistics (encyclopedia). Scien Publ Bolshaya Rossiyskaya Encyclopedia, Moscow
31. Yulmetyev R et al (2000) Stochastic dynamics of time correlation in complex systems with discrete time. Phys Rev E 62:6178–6194
32. Yulmetyev R et al (2002) Quantification of heart rate variability by discrete nonstationarity non-Markov stochastic processes. Phys Rev E 65(1–15):046107
33. Reed M, Samon B (1972) Methods of mathematical physics. Academic, New York
34. Graber H (1982) Projection operator technique in nonequilibrium statistical mechanics. In: Höhler G (ed) Springer tracts in modern physics, vol 95. Springer, Berlin
35. Yulmetyev RM (2001) Possibility between earthquake and explosion seismogram differentiation by discrete stochastic non-Markov processes and local Hurst exponent analysis. Phys Rev E 64(1–14):066132
36. Abe S, Suzuki N (2004) Aging and scaling of earthquake aftershocks. Physica A 332:533–538
37. Tirnakli U, Abe S (2004) Aging in coherent noise models and natural time. Phys Rev E 70(1–4):056120
38. Abe S, Sarlis NV, Skordas ES, Tanaka HK, Varotsos PA (2005) Origin of the usefulness of the natural-time representation of complex time series. Phys Rev Lett 94(1–4):170601
39. Stanley HE, Meakin P (1988) Multifractal phenomena in physics and chemistry. Nature 335:405–409
40. Ivanov P Ch, Amaral LAN, Goldberger AL, Havlin S, Rosenblum MG, Struzik Z, Stanley HE (1999) Multifractality in human heartbeat dynamics. Nature 399:461–465
41. Mokshin AV, Yulmetyev R, Hänggi P (2005) Simple measure of memory for dynamical processes described by a generalized Langevin equation. Phys Rev Lett 95(1–4):200601
42. Allegrini P et al (2003) Compression and diffusion: A joint approach to detect complexity. Chaos Soliton Fractal 15: 517–535
43. Amaral LAN et al (2001) Application of statistical physics methods and concepts to the study of science and technology systems. Scientometrics 51:9–36
44. Arneodo A et al (1996) Wavelet based fractal analysis of DNA sequences. Physica D 96:291–320
45. Ashkenazy Y et al (2003) Magnitude and sign scaling in power-law correlated time series. Physica A Stat Mech Appl 323:19–41
46. Ashkenazy Y et al (2003) Nonlinearity and multifractality of climate change in the past 420,000 years. Geophys Res Lett 30:2146
47. Azbel MY (1995) Universality in a DNA statistical structure. Phys Rev Lett 75:168–171
48. Baldassarri A et al (2006) Brownian forces in sheared granular matter. Phys Rev Lett 96:118002
49. Baleanu D et al (2006) Fractional Hamiltonian analysis of higher order derivatives systems. J Math Phys 47:103503
50. Blesic S et al (2003) Detecting long-range correlations in time series of neuronal discharges. Physica A 330:391–399
51. Cajueiro DO, Tabak BM (2004) The Hurst exponent over time: Testing the assertion that emerging markets are becoming more efficient. Physica A 336:521–537
52. Brecht M et al (1998) Correlation analysis of corticotectal interactions in the cat visual system. J Neurophysiol 79: 2394–2407
53. Brouersa F, Sotolongo-Costab O (2006) Generalized fractal kinetics in complex systems (application to biophysics and biotechnology). Physica A 368(1):165–175
54. Coleman P, Pietronero L (1992) The fractal structure of the universe. Phys Rep 213:311–389
55. Goldberger AL et al (2002) What is physiologic complexity and how does it change with aging and disease? Neurobiol Aging 23:23–26
56. Grau-Carles P (2000) Empirical evidence of long-range correlations in stock returns. Physica A 287:396–404
57. Grigolini P et al (2001) Asymmetric anomalous diffusion: An efficient way to detect memory in time series. Fractal-Complex Geom Pattern Scaling Nat Soc 9:439–449
58. Ebeling W, Frommel C (1998) Entropy and predictability of information carriers. Biosystems 46:47–55
59. Fukuda K et al (2004) Heuristic segmentation of a nonstationary time series. Phys Rev E 69:021108
60. Hausdorff JM, Peng CK (1996) Multiscaled randomness: A possible source of 1/f noise in biology. Phys Rev E 54: 2154–2157
61. Herzel H et al (1998) Interpreting correlations in biosequences. Physica A 249:449–459
62. Hoop B, Peng CK (2000) Fluctuations and fractal noise in biological membranes. J Membrane Biol 177:177–185
63. Hoop B et al (1998) Temporal correlation in phrenic neural activity. In: Hughson RL, Cunningham DA, Duffin J (eds) Advances in modelling and control of ventilation. Plenum Press, New York, pp 111–118
64. Ivanova K, Ausloos M (1999) Application of the detrended fluctuation analysis (DFA) method for describing cloud breaking. Physica A 274:349–354
65. Ignaccolo M et al (2004) Scaling in non-stationary time series. Physica A 336:595–637
66. Imponente G (2004) Complex dynamics of the biological rhythms: Gallbladder and heart cases. Physica A 338:277–281

67. Jefferiesa P et al (2003) Anatomy of extreme events in a complex adaptive system. Physica A 318:592–600
68. Karasik R et al (2002) Correlation differences in heartbeat fluctuations during rest and exercise. Phys Rev E 66:062902
69. Kulessa B et al (2003) Long-time autocorrelation function of ECG signal for healthy versus diseased human heart. Acta Phys Pol B 34:3–15
70. Kutner R, Switala F (2003) Possible origin of the non-linear long-term autocorrelations within the Gaussian regime. Physica A 330:177–188
71. Koscielny-Bunde E et al (1998) Indication of a universal persistence law governing atmospheric variability. Phys Rev Lett 81:729–732
72. Labini F (1998) Scale invariance of galaxy clustering. Phys Rep 293:61–226
73. Linkenkaer-Hansen K et al (2001) Long-range temporal correlations and scaling behavior in human brain oscillations. J Neurosci 21:1370–1377
74. Mercik S et al (2000) What can be learnt from the analysis of short time series of ion channel recordings. Physica A 276:376–390
75. Montanari A et al (1999) Estimating long-range dependence in the presence of periodicity: An empirical study. Math Comp Model 29:217–228
76. Mark N (2004) Time fractional Schrodinger equation. J Math Phys 45:3339–3352
77. Niemann M et al (2008) Usage of the Mori–Zwanzig method in time series analysis. Phys Rev E 77:011117
78. Nigmatullin RR (2002) The quantified histograms: Detection of the hidden unsteadiness. Physica A 309:214–230
79. Nigmatullin RR (2006) Fractional kinetic equations and universal decoupling of a memory function in mesoscale region. Physica A 363:282–298
80. Ogurtsov MG (2004) New evidence for long-term persistence in the sun's activity. Solar Phys 220:93–105
81. Pavlov AN, Dumsky DV (2003) Return times dynamics: Role of the Poincare section in numerical analysis. Chaos Soliton Fractal 18:795–801
82. Paulus MP (1997) Long-range interactions in sequences of human behavior. Phys Rev E 55:3249–3256
83. Peng C-K et al (1994) Mosaic organization of DNA nucleotides. Phys Rev E 49:1685–1689
84. Peng C-K et al (1995) Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos 5:82–87
85. Poon CS, Merrill CK (1997) Decrease of cardiac chaos in congestive heart failure. Nature 389:492–495
86. Rangarajan G, Ding MZ (2000) Integrated approach to the assessment of long range correlation in time series data. Phys Rev E 61:4991–5001
87. Robinson PA (2003) Interpretation of scaling properties of electroencephalographic fluctuations via spectral analysis and underlying physiology. Phys Rev E 67:032902
88. Rizzo F et al (2005) Transport properties in correlated systems: An analytical model. Phys Rev B 72:155113
89. Shen Y et al (2003) Dimensional complexity and spectral properties of the human sleep EEG. Clinic Neurophysiol 114:199–209
90. Schmitt D et al (2006) Analyzing memory effects of complex systems from time series. Phys Rev E 73:056204
91. Soen Y, Braun F (2000) Scale-invariant fluctuations at different levels of organization in developing heart cell networks. Phys Rev E 61:R2216–R2219
92. Stanley HE et al (1994) Statistical-mechanics in biology – how ubiquitous are long-range correlations. Physica A 205: 214–253
93. Stanley HE (2000) Exotic statistical physics: Applications to biology, medicine, and economics. Physica A 285:1–17
94. Tarasov VE (2006) Fractional variations for dynamical systems: Hamilton and Lagrange approaches. J Phys A Math Gen 39:8409–8425
95. Telesca L et al (2003) Investigating the time-clustering properties in seismicity of Umbria-Marche region (central Italy). Chaos Soliton Fractal 18:203–217
96. Turcott RG, Teich MC (1996) Fractal character of the electrocardiogram: Distinguishing heart-failure and normal patients. Ann Biomed Engin 24:269–293
97. Thurner S et al (1998) Receiver-operating-characteristic analysis reveals superiority of scale-dependent wavelet and spectral measures for assessing cardiac dysfunction. Phys Rev Lett 81:5688–5691
98. Vandewalle N et al (1999) The moving averages demystified. Physica A 269:170–176
99. Varela M et al (2003) Complexity analysis of the temperature curve: New information from body temperature. Eur J Appl Physiol 89:230–237
100. Varotsos PA et al (2002) Long-range correlations in the electric signals that precede rupture. Phys Rev E 66:011902
101. Watters PA (2000) Time-invariant long-range correlations in electroencephalogram dynamics. Int J Syst Sci 31:819–825
102. Wilson PS et al (2003) Long-memory analysis of time series with missing values. Phys Rev E 68:017103
103. Yulmetyev RM et al (2004) Dynamical Shannon entropy and information Tsallis entropy in complex systems. Physica A 341:649–676
104. Yulmetyev R, Hänggi P, Gafarov F (2000) Stochastic dynamics of time correlation in complex systems with discrete time. Phys Rev E 62:6178
105. Yulmetyev R, Gafarov F, Hänggi P, Nigmatullin R, Kayumov S (2001) Possibility between earthquake and explosion seismogram processes and local Hurst exponent analysis. Phys Rev E 64:066132
106. Yulmetyev R, Hänggi P, Gafarov F (2002) Quantification of heart rate variability by discrete nonstationary non-Markov stochastic processes. Phys Rev E 65:046107
107. Yulmetyev R, Demin SA, Panischev OY, Hänggi P, Timashev SF, Vstovsky GV (2006) Regular and stochastic behavior of Parkinsonian pathological tremor signals. Physica A 369:655

Books and Reviews

Badii R, Politi A (1999) Complexity: Hierarchical structures and scaling in physics. Oxford University Press, New York

Elze H-T (ed) (2004) Decoherence and entropy in complex systems. In: Selected lectures from DICE 2002 series: Lecture notes in physics, vol 633. Springer, Heidelberg

Kantz H, Schreiber T (2004) Nonlinear time series analysis. Cambridge University Press, Cambridge

Mallamace F, Stanley HE (2004) The physics of complex systems (new advances and perspectives). IOS Press, Amsterdam

Parisi G, Pietronero L, Virasoro M (1992) Physics of complex systems: Fractals, spin glasses and neural networks. Physica A 185(1–4):1–482
Sprott JC (2003) Chaos and time-series analysis. Oxford University Press, New York
Zwanzig R (2001) Nonequilibrium statistical physics. Oxford University Press, New York

Cosmic Gravitational Background, Stochastic

Carlo Ungarelli
CNR-Institute of Geosciences and Earth Resources, Pisa, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Production Mechanisms of Relic GWs
Future Directions
Bibliography

Glossary

Cosmic microwave background Quasi-isotropic thermal, background of electromagnetic radiation of cosmological (i. e. not astrophysical origin). The observed spectrum of this radiation is consistent with a black body with temperature $T = 2.725\,\mathrm{K}$ and represents the best experimental confirmation of the Planck spectrum. According to the standard cosmological model, the dynamics of the primordial universe is determined, via the Einstein gravitational equations, by a hot relativistic plasma of electrons, photons and baryons. The photons interact with the plasma through Thomson scattering. As the universe expanded, the plasma cools until it becomes favorable for electrons to combine with protons and form hydrogen atoms. This occurs at a temperature of about 3×10^3 K (corresponding to an average energy of about 0.3 eV). At this point, the photons scatter off the now neutral atoms and start to travel freely through space. This process is called recombination or decoupling (referring to electrons combining with nuclei and to the decoupling of matter and radiation, respectively). Since the universe is expanding the photon temperature keeps decreasing reaching the current value of $T = 2.725\,\mathrm{K}$. Accordingly, the photons observed today (i. e. the CMB photons) come from a spherical surface, called the **surface of last scattering**, corresponding to the time of decoupling.

Friedmann–Robertson–Walker model Model based upon an isotropic and homogeneous metric describing an expanding universe corresponding to a solution of the Einstein gravitational equation with a perfect fluid source (described in terms of its energy density ρ and its pressure p). The line element depends on the three-dimensional constant curvature k and the scale factor $a(t)$ describing the expansion: $\mathrm{d}s^2 = \mathrm{d}t^2 - a^2(t)(\mathrm{d}r^2/(1 - kr^2) + r^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2))$. The expansion rate is given by $H = \dot{a}/a$. Given the equation of state $p = p(\rho)$, the solution for the scale factor is obtained from the **Friedmann equation** $H^2 = 8\pi G\rho/3$.

Inflation Primordial epoch in the evolution of the Universe characterized by an accelerating expansion ($\ddot{a} > 0$). In particular, to achieve acceleration, a non-standard source is needed (i. e. a perfect fluid with an effective equation of state such that $\rho + 3p < 0$). The existence of such a phase is postulated in order to explain the observed quasi-isotropy and homogeneity of the Universe on large scales, as well as to provide a mechanism to generate primordial seeds for structure formation.

Definition of the Subject

As first predicted by Albert Einstein [1], propagation effects at finite speed in the dynamical gravitational equations yield to the existence of wave-like solutions of the gravitational field linearized around the flat-vacuum configuration. Such solutions correspond to the so-called gravitational waves (GWs). At the theoretical level, the work of Bondi [2] (who considered self-gravitating systems like binaries of compact objects) showed that GWs carry energy, while observationally the discovery of the 1916 + 13 binary pulsar system by Hulse and Taylor [3] (in particular, the speed-up of the binary pulsar period), provided an indirect evidence for the existence of GWs.

On the experimental side, after the pioneering work of Weber in the 1960s and the subsequent development of meter-scale, resonant bar detectors (ALLEGRO, AURIGA EXPLORER, NAUTILUS and NIOBE), in more recent years a number of new detectors have been designed and built to search for GWs. Kilometer-scale interferometers (GEO [4], LIGO [5], TAMA [6] and VIRGO [7]), sensitive to GWs in the frequency range between 10 Hz and 1 kHz, are already producing scientific runs at the design sensitivity (or close to it, as for VIRGO). The Laser Interferometric Space Antenna (LISA) [8], a space-based, three-arms

interferometer will be aimed at detecting GWs in the frequency range between 0.1 mHz and 10 mHz, while a next generation of space-detectors especially designed to detect primordial GWs has already being planned [9,10].

Gravitational waves can be produced by violent dynamical phenomena involving compact astrophysical objects (like neutron stars and/or black holes), as well as by various mechanisms related to the dynamics of the primordial universe. Those primordial GWs produce a background of gravitational radiation that can be considered as a sort of "cosmological noise", i. e. the gravitational counterpart of the cosmic microwave background radiation (CMB). The detection of such primordial gravitational radiation can be exploited to investigate physical processes occurring during the early stages of the evolution of the universe, when the typical energy scales are not accessible experimentally in any other way. The reason why primordial GWs are characterized by this unique feature is the well-known fact that in an expanding universe particles decoupled from the primordial plasma at a temperature T_D carry information about the state of the Universe at T_D. In particular, for gravitons it is possible to estimate the decoupling temperature assuming that the gravitons are kept in thermal equilibrium through two-body scattering processes (see e. g. [11]). In this case, the interaction rate is[1]. $\Gamma_G \sim T^5/M_{pl}^4$, where M_{pl} is the Planck mass ($\sim 10^{19}$ GeV). Since the energy density for relativistic particles scales with the fourth power of the temperature, assuming an adiabatic expansion the expansion rate H scales as $H \sim \dot{T}/T$. Hence, thermal equilibrium is maintained provided that the typical time-scale for graviton interaction is smaller than the expansion time-scale, i. e. $\Gamma_G > H$. Using the Friedmann equation, $H \propto T^2/M_{pl}$ this yields

$$\left(\frac{\Gamma_G}{H}\right) \sim \left(\frac{T}{M_{pl}}\right)^3 . \tag{1}$$

Gravitons are therefore decoupled just below the Planck scale. This implies that any relic GWs retain in their spectrum (i. e., amplitude and frequency) unique information about the state of the primordial universe and the corresponding physical process occurring at extremely high energy scales.

The first papers addressing possible mechanisms for the production of a cosmological background of gravitational radiation date back to the 1970s. Grishchuck [12] and Starobinsky [13] showed that metric tensor perturbations in an expanding, homogeneous and isotropic universe can be parametrically amplified provided there is a stage in the evolution of the universe when the characteristic time scale during which the metric changes is much smaller than the intrinsic wave period of the perturbations. This phenomenon, known as parametric amplification of vacuum fluctuations, is characteristic of inflationary models and its phenomenological consequences have been widely investigated (see e. g. [14]). Furthermore, as Witten [15] and Hogan [16] first pointed out, cosmological first-order phase transitions (e. g. expected to occur at the QCD scale ($T \sim 100$ MeV) and/or at the electro-weak scale ($T \sim 100$ GeV)) may be associated to the production of gravitational radiation. In particular, Hogan and Witten showed that in a cosmological, first-order phase transition, both the collision of detonation waves triggered by the nucleation of the low-temperature phase and the generation of acoustic noise in the relativistic plasma could produce a relic background of gravitational waves. Since those pioneering works, detailed computations of various types of gravitational wave signals have been performed, and data analysis techniques for extracting this kind of signal have been developed.

Introduction

Without loss of generality, a cosmological gravitational background is expected to be Gaussian, isotropic, stationary and unpolarized (for a detailed discussion see e. g. [17]). All the physical properties are then encoded into the so-called *spectrum*, given by the following, dimensionless quantity

$$\Omega_{gw}(f) = \frac{1}{\rho_c} f \, \tilde{\rho}_{gw}(f) , \tag{2}$$

where f is the frequency, $\tilde{\rho}_{gw}(f)$ is the gravitational wave energy density per unit frequency, i. e.

$$\rho_{gw} = \int_0^\infty \mathrm{d}f \, \tilde{\rho}_{gw}(f) , \tag{3}$$

and $\rho_c = 3H_0^2/(8\pi G)$ is the critical energy density ($H_0 = h_0 \times 100$ km/(s Mpc) being the current value of the Hubble constant, with h_0 parameterizing the current experimental uncertainty). Hence, to characterize any stochastic background of gravitational radiation, the combination $h_0^2 \, \Omega_{gw}$ is adopted.

As a signal to be detected, a stochastic background of gravitational waves is described in terms of a random process that cannot be disentangled from the detector noise. Most of the theoretical predictions yield values for the GWs spectrum much lower than the noise level

[1] In the following, we will adopt units for which temperatures, masses and energies are all measured in electron-volts (eV) or multiples of it and distances are expressed as inverse of energies

of any current or planned earth-based detector; on the other hand, the instrumental noise features of GW detectors are not sufficiently well understood to search for excess noise. For these reasons, an optimal detection strategy has been developed [18] based upon the cross-correlation between two or more widely separated detectors (so as to minimize the common noise sources). In particular, the cross-correlation improves the sensitivity (in terms of the spectrum) by a factor $(\Delta f\, T_{\rm obs})^{1/2}$, where Δf is the effective detector bandwidth and $T_{\rm obs}$ is the total observation time. Using the design LIGO sensitivity, assuming four months of observation time and a constant spectrum a stochastic background of GWs could be detected at 90% confidence level provided that $h_0^2\,\Omega_{\rm gw} > 5\times 10^{-6}$ [19] (the recent analysis of LIGO data [20] sets a 95% confidence level upper limit $h_0^2\,\Omega_{\rm gw} < 6.5\times 10^{-5}$), while the advanced-LIGO design sensitivity sets a lower detection limit $\Omega_{\rm gw} \sim 8\times 10^{-9}$ [19]. As for the planned, three-arm space-based interferometer LISA, a possible strategy for detecting a stochastic background of gravitational radiation exploits some combinations of the signals from the three spacecraft that are insensitive to any GW signal (the so-called Sagnac combination) [21]. Such combinations therefore provide a calibration channel to measure the instrumental noise power spectrum. However, beside primordial GWs, in the frequency range 0.1–10 mHz an astrophysical background is also expected due to the incoherent superposition of GWs emitted by large populations (both galactic and extra-galactic) of binary systems of compact objects. Such an astrophysical background provides an additional noise contribution for the detection of any primordial gravitational radiation in the LISA frequency band.

In general the intensity and the spectral content of a generic cosmological background of gravitational radiation depends upon the producing mechanism and the underlying theoretical model. However, it is possible to characterize primordial gravitational radiation in terms of both some observational constraints and some model-independent features.

Observational Bounds

There are three main observational bounds that a cosmological background of gravitational radiation must satisfy: the nucleosynthesis bound, the COBE/CMB bound and the pulsar-timing bound.

The *nucleosynthesis bound* [22] is related to the highly-accurate predictions of the standard cosmological model pertaining to the primordial abundances of light elements. Such abundances depend upon the ratio between the number density of neutrons $n_{\rm n}$ and the number density of protons $n_{\rm p}$. In conditions of thermal equilibrium $n_{\rm n}/n_{\rm p} \sim {\rm e}^{-\Delta m/T}$, where $\Delta m = m_{\rm n} - m_{\rm p} \sim 1.3\,{\rm MeV}$. In particular, the process responsible for maintaining thermal equilibrium is $p + e^- \rightleftarrows n + \nu_e$, as long as the corresponding rate Γ is greater than the expansion rate H. When this condition is no longer fulfilled, the ratio $n_{\rm n}/n_{\rm p}$ freezes out at a value ${\rm e}^{-\Delta m/T_{\rm f}}$, and this determines the number of neutrons yielding the formation of light elements (mainly He^4). Since the Hubble constant is given by the Friedmann equation $H^2 = 8\pi\, G\,\rho/3$, where the energy density ρ incorporates all the possible contributions at the nucleosynthesis epoch, its value at the freeze-out depends also upon a possible contribution due to a gravitational radiation component of cosmological origin. Hence, a constraint on the value of the freeze-out temperature yields an upper-limit for the energy density associated to any extra-contribution other than standard model particles. Assuming that the only extra contribution is due to cosmological gravitons, the corresponding spectrum must satisfy the following bound [23]

$$\int_0^\infty {\rm d}f\, h_0^2 \Omega_{\rm gw}(f) \lesssim 5\times 10^{-6} \tag{4}$$

COBE/CMB bound. On large scales, the spectrum amplitude of a stochastic background of cosmological origin must be compatible with the measurements of the CMB temperature fluctuations. Indeed, at large wavelengths, a stochastic background of gravitational radiation produces a dynamical red-shift on the frequencies of the CMB photons, thus generating fluctuations in their temperature (according to the so-called Sachs–Wolfe effect [24]). Those fluctuations have been measured firstly by the COBE experiment [25] and more recently by WMAP [26]. Since on large scales also density perturbations (inhomogeneities in the matter distribution) generate fluctuations in the CMB temperature, any gravitational wave spectrum must satisfy the following bound [27]

$$h_0^2\,\Omega_{\rm gw}(f) \lesssim \left(\frac{H_0}{f}\right)^2 \left\langle \frac{\delta T}{T}\right\rangle^2 , \tag{5}$$

where $\langle \delta T/T\rangle \sim 5\times 10^{-6}$ is the CMB temperature rms fluctuation (taking only the quadrupole contribution). The frequency range over which the constraint (5) holds corresponds to modes that are currently inside the causal horizon today ($f > 3\times 10^{-8}\,{\rm Hz}$) and were associated to super-horizon scales at the surface of last-scattering ($f < 10^{-16}\,{\rm Hz}$). At the maximum frequency ($10^{-16}\,{\rm Hz}$) the bound therefore yields an upperlimit for any primordial GW spectrum $h_0^2\,\Omega_{\rm gw} \lesssim 10^{-14}$. This upperlimit is

particularly relevant for gravitational backgrounds produced by the amplification of vacuum fluctuations occurring during a primordial, inflationary era: the corresponding spectra are indeed characterized by a wide frequency window ranging from $f \sim H_0 \sim 3 \times 10^{-18}$ Hz up to GHz values.

Recently, it has been pointed out [28] that since any cosmological background of gravitational radiation behaves as a gas of free-streaming particles, in addition to its effect on the expansion rate at the nucleosynthesis epoch, also the growth of density perturbations is affected. In particular, if the primordial metric perturbations yielding to a background of gravitational radiation are not adiabatic (as occurs for the parametrically amplified perturbations during inflation), then the effects on the CMB photons are different from those associated to adiabatic massless neutrinos. Using the recent WMAP data, this feature allows us to put a constraint on the spectrum of primordial gravitational radiation down to frequencies corresponding to the comoving horizon at decoupling ($f \sim 10^{-15}$ Hz). At those frequencies, the analysis carried out in [28] establishes a current 95% confidence-level upper limit $h_0^2\, \Omega_{\rm gw} \lesssim 8.4 \times 10^{-6}$, while future CMB experiments (like Planck) could reach sensitivities such that $h_0^2\, \Omega_{\rm gw} \lesssim 1.4 \times 10^{-6}$ [28].

Pulsar timing bound. A consistent fraction of known pulsars with a period in the range [1.5–30] ms (the so-called millisecond pulsars) are characterized by an extremely stable integrated pulse profile and therefore may be considered as natural clocks. Those astrophysical systems are natural detectors of gravitational radiation: a GW traveling between a detector and a pulsar produces a fluctuation in the arrival time of the pulse proportional to the wave amplitude. For a total observation time $T_{\rm obs}$ and an uncertainty $\Delta\tau_{\rm arr}$ in the arrival time such detectors would be sensitive to GWs whose amplitude is $h \sim \Delta\tau_{\rm arr}/T_{\rm obs}$ at frequencies $f \sim T_{\rm obs}^{-1}$. The highest sensitivity is reached for continuous signals (like a stochastic background) with observation times of one or two years (corresponding to frequencies $10^{-9} - 10^{-8}$ Hz). The most stringent constraint on a stochastic background of gravitational radiation comes from the analysis of the data concerning PSR B1855+09 [29] which gives the following 95% confidence-level bound at $\bar{f} = 4.4 \times 10^{-9}$ Hz

$$h_0^2\, \Omega_{\rm gw}(\bar{f}) < 1.0 \times 10^{-8}\,. \tag{6}$$

Characteristic Frequency and Amplitude

Beside the observational constraints discussed above, predictions concerning relic GWs are poorly constrained both because of our lack of knowledge about physics beyond the standard model and because of large uncertainties in the understanding of the various production mechanisms concerning primordial GWs. Nevertheless, it is possible to outline some general features describing the typical frequency and amplitude spectrum of cosmological gravitational radiation (see e. g. [11]).

The typical frequency at which a gravitational signal could be detected can be estimated [11] disentangling the dynamical effects associated with the production mechanism from the kinematical contribution depending on the red-shift associated to the cosmic time at which the signal is produced. Assuming that a graviton is produced at a certain temperature $T_{\rm P}$ with a frequency $f_{\rm P}$ during the epoch where the Universe was either radiation or matter dominated, the frequency at which it could be currently detected is $f_0 = f_{\rm P}/(1 + z_{\rm P})$. To compute the red-shift $z_{\rm P}$ one exploits the adiabatic expansion of the universe, i. e. the fact that the entropy per comoving volume $S \propto g_{\rm S}(T)\, a^3(t)\, T^3$ is conserved (here $a(t)$ is the time-dependent scale factor of the Universe and $g_{\rm S}(T)$ accounts for the effective number of degrees of freedom at temperature T). The result is ([11,30])

$$f_0 \sim 8 \times 10^{-14}\, f_{\rm P} \left(\frac{100}{g_{\rm S}(T_{\rm P})}\right)^{1/3} \left(\frac{1\,{\rm GeV}}{T_{\rm P}}\right) \tag{7}$$

The dynamics of the production mechanism enters into the estimate of the frequency $f_{\rm P}$. It is natural to express the wavelength $\lambda_{\rm P}$ corresponding to such frequency in terms of the horizon scale $H_{\rm P}^{-1}$ at the time of production, i. e. $\lambda_{\rm P} = \xi H_{\rm P}^{-1}$, with $\xi \leqslant 1$. Assuming that the gravitational radiation is produced during the era where the energy density is dominated by relativistic radiation, $H_{\rm P}^2 \propto g(T_{\rm P})\, T_{\rm P}^4/M_{\rm pl}^2$, $g(T_{\rm P}) \sim g_{\rm S}(T_{\rm P})$ (here $g(T_{\rm P})$ is the number of degrees of freedom at temperature $T_{\rm P}$). This yields [11]

$$f_0 \sim 1.6 \times 10^{-7}\, \frac{1}{\xi} \left(\frac{100}{g(T_{\rm P})}\right)^{-1/6} \left(\frac{T_{\rm P}}{1{\rm GeV}}\right) {\rm Hz}\,. \tag{8}$$

Hence, the physics at TeV scale could be probed by space-borne interferometers like LISA, while earth-based interferometers could investigate phenomena whose typical energy scale is in the range 10^8–10^{10} GeV.

As for the characteristic intensity of primordial gravitational radiation, despite a strong dependence upon the production mechanism, it is still possible to draw some general considerations [11]. When primordial gravitational waves are produced at the Planck scale by collisions and decay together with other relativistic particles that are

currently observed (i. e. CMB photons), it is reasonable to assume $\rho_{gw}(t_P) \sim \rho_\gamma(t_P)$. In this case, taking into account the different red-shift characterizing photons and gravitons (the latter decoupling just below the Planck scale), the present value for the GW spectrum is estimated to be [30]

$$h_0^2\,\Omega_{gw} \sim 1.7 \times 10^{-5} \left(\frac{100}{g(T_P)}\right)^{1/3} \Omega_{gw}(t_P)\,, \tag{9}$$

where $\Omega_{gw}(t_P)$ is the fraction of energy density produced in GWs. Another possible scenario incorporates all possible mechanisms of primordial production of GWs different from the ones giving rise to CMB photons. Also in this case it is possible to relate the GW spectrum to the CMB spectrum. For instance, one may consider the spectrum of gravitational radiation produced by the amplification of vacuum fluctuations during an inflationary epoch; in particular, it is possible to estimate the spectrum amplitude for a frequency corresponding to the highest curvature scale H_{max} reached during the inflationary phase (see e. g. [11]). The result is

$$h_0^2\,\Omega_{gw} \sim \left(\frac{H_{max}}{M_{pl}}\right)^2 h_0^2\,\Omega_\gamma\,, \tag{10}$$

where $h_0^2\,\Omega_\gamma \sim 10^{-5}$ is the current fraction (in terms of the critical density) of radiation energy density. Slow-roll inflationary models are characterized by a typical Hubble scale during the acceleration era that cannot exceed a value of about $8 \times 10^{-6}\,M_{pl}$ (to be consistent with the CMB anisotropy measurements) [31], yielding to an almost constant spectrum in the frequency window $[10^{-16}, 10^9]$ Hz $h_0^2\,\Omega_{gw} \sim 10^{-15}$ [32]. While this is too low to be detected both by earth-based interferometers (even in their advanced configurations) and by LISA (because of the presence of a strong astrophysical background) [33], a space detector sensitive in the frequency band $[0.1, 1]$ Hz (which is almost free from astrophysical backgrounds) could reach the desired sensitivity [34]. Pre-big-bang models [35], string-theory-based cosmological models, are characterized by an inflationary phase where the curvature grows reaching a maximum threshold value $H_{max} \sim 10^{-1}\,M_{pl}$. Such a feature in particular produces a stochastic background characterized by a steep-blue spectrum [36] ($\Omega_{gw} \sim f^3$); the COBE bound is easily satisfied, and, for a certain range of the free parameters of the model, the spectrum can reach a maximum value $h_0^2\,\Omega_{gw} \sim 10^{-7}$ [37] (compatible with the nucleosynthesis bound) in the frequency range $[10^{-4}, 10^3]$ Hz. This signal could be in principle detected by Advanced LIGO.

Production Mechanisms of Relic GWs

The main mechanisms that can produce a primordial stochastic background of gravitational radiation can be divided into two broad classes:

- Parametric amplification of metric tensor perturbations occurring during an early epoch of inflationary expansion;
- Causal process (mainly phase transitions) occurring during the primordial stages of the evolution of the Universe.

Parametric Amplification of Vacuum Fluctuations

In a cosmological model characterized by different dynamical epochs, in each of them any quantum field has different normal modes corresponding to different sets of creation and annihilation operators. Those sets are connected by the so-called Bogoliubov transformations [38] that mix positive frequency modes (i. e. annihilation operators) with negative frequency modes (i. e. creation operators). As a consequence, in an expanding universe, an initial vacuum state (with respect to a certain set of annihilation/creation operators), due to the dynamical evolution of the metric, in a different epoch corresponds to a multi-particle state. The physical idea underlying quantum production of particles in an expanding universe can be qualitatively explained as follows [11]. For the sake of simplicity, let's consider a cosmological model with two different phases (A and B), where the transition between them occurs over a time scale $\Delta\tau \sim H^{-1}(\tau)$. For a mode whose physical frequency at the time of transition is f_τ there exist two alternative regimes: abrupt/sudden transition ($2\pi f_\tau \Delta\tau \ll 1$) or smooth, adiabatic transition ($2\pi f_\tau \Delta\tau \gg 1$). For modes characterized by a sudden transition, while the physical state associated to the field in the phase A has not significantly changed, the corresponding particle number must be evaluated with respect to the set of annihilation and creation operators relative to the phase B. Furthermore, due to the mixing between positive frequency and negative frequency modes, the vacuum state relative to the phase A corresponds to a multi-particle state relative to the phase B. On the other hand, for modes for which the transition is adiabatic, no particle production takes place. Applying this analysis to a cosmological scenario with a primordial phase of accelerated expansion driven by a scalar field, vacuum metric fluctuations are amplified thus producing, in the final decelerating radiation/matter dominated phase a final, multi-graviton phase that gives rise to a stochastic background of gravitational radiation. The corresponding spectrum

is given by [11]

$$\Omega_{\rm gw} = \frac{16\pi^4}{\rho_{\rm c}}\, n_{\rm f}\, f^4 \,, \tag{11}$$

where $n_{\rm f}$ is the number of gravitons per cell of phase space. As mentioned before the frequency spectrum of such backgrounds extends over a huge range of frequencies, from $f_{\rm min} \sim 3 \times 10^{-18}$ Hz (corresponding to a wavelength compatible with the size of our observable universe) up to a maximum frequency in the GHz range (this ultraviolet cut-off is fixed by the maximum curvature attained during the inflationary phase). As for the strength of the spectrum, the predictions are sensitive to the mechanism that produces the inflationary stage (see previous section).

Causal Mechanisms: Phase Transitions

During the cosmological evolution of the Universe, phase transitions corresponding to symmetry breaking of particle physics fundamental interactions are expected to take place. As for the generation of relic GWs, the presence and type of topological defects and the order of the transition are of particular relevance.

Cosmic Strings Cosmic strings are topological defects forming as a consequence of the spontaneous breaking of a global $U(1)$ gauge symmetry [39,40]. They are characterized by a large mass-per-unit-length μ (related to the scale at which the symmetry is broken), a closed shape and a typical length-scale that can be larger or smaller than the Hubble radius $r_{\rm H} = 1/H$. Since their tension is equal to their mass per unit length, they have relativistic oscillations, therefore shrinking in size and emitting gravitational radiation. Assuming that a multi-scale network of such strings has been formed, loops whose radius is smaller than $r_{\rm H}$ oscillate, radiate away GWs, shrink and eventually disappear, replaced by loops broken off from loops whose radius is larger than $r_{\rm H}$. The typical scale for the GW spectrum is determined by the loop radius: since in the network loops with different sizes are present, the corresponding (almost Gaussian) spectrum is nearly flat over a frequency range $[10^{-8}, 10^{10}]$ Hz. A simple dimensional analysis shows that the spectrum can be estimated as $\Omega_{\rm gw} \sim \gamma\mu/M_{\rm pl}^2$, where γ is a dimensionless parameter encoding the radiation efficiency. Numerical simulations of cosmic strings predict $h_0^2\,\Omega_{\rm gw} \sim 10^{-9} \div 10^{-8}$ for $\mu/M_{\rm pl}^2 < 10^{-6}$ [41]. Cosmic strings may also produce intense bursts of gravitational radiation [42]: in particular, the strongest bursts are produced at sections of strings characterized by large Lorentz boosts. Though the analysis of [42] suggests that such bursts could be detected even with the current LIGO design sensitivity, a more recent analysis [43] points out that large-scale earth-based interferometers at current sensitivity (such as LIGO) are unlikely to detect such GW radiation when the string network is characterized by a classical distribution of loop sizes. On the other hand, if the size of the string loops is much smaller than its classical value ($l_{\rm classical} \sim \gamma\mu t/M_{\rm pl}^2$) and it is determined by gravitational back-reaction, the burst amplitude is enhanced and those signals could be detectable with the Advanced LIGO design sensitivity [43].

Cosmic strings can also be produced in some classes of inflationary models based upon string theory [44]. Those strings, dubbed cosmic superstrings, have two key features distinguishing them from the usual field-theoretic strings: (a) Due to their probabilistic interaction and to the higher dimensionality of the underlying theory, their reconnection probability p is less than 1 ($10^{-3} < p < 1$ [45], while for field theoretic strings $p = 1$); (b) different kinds of strings can be formed. As well as strong bursts of gravitational radiation, cosmic superstrings can also generate a stochastic background of GWs, as a result of the incoherent superposition of cusp bursts from a strings network. Taking into account various effects, Siemens and collaborators [46] have investigated the detectability of such a background considering a wide range of experiments. In particular, they have found that interferometric GW detectors are sensitive to portions of the parameter space $[p, \mu/M_{\rm pl}]$, complementary to ones constrained by the various experimental bounds.

First-Order Phase Transitions In a first-order phase transition, the universe is locally trapped into a meta-stable, high-temperature unbroken symmetry phase. Such a meta-stable configuration is separated from the stable vacuum phase by a barrier in the (temperature-dependent) potential describing the order parameter of the transition. At the quantum level, a transition from the meta-stable to the stable vacuum state takes place through tunneling. As a consequence, random nucleation of bubbles occurs, where the configuration inside a generic bubble corresponds to the true vacuum, while outside it is represented by the meta-stable phase. When the temperature drops below a critical threshold, bubbles have enough volume to expand (the latent heat released during the transition is converted into kinetic energy for the bubbles); they approach a velocity close to the speed of light, collide and eventually yield to a universe in the unbroken-symmetry phase. The occurrence of such collisions breaks the bubbles spherical symmetry and GWs (as well as other kinds of radiation) are produced ([15,16,47]). The corresponding GW spectrum is described by two parameters, the bubble

nucleation rate per unit volume β and the ratio α between the false vacuum energy density and the radiation energy density released at the critical, transition temperature T_{tr}. Since the gravitational radiation is produced by collision of bubbles, the spectrum is strongly peaked at the frequency characteristic of the collision rate ($2\pi f_P \sim \beta$). A detailed calculation yields [47]

$$f_P \sim 5 \times 10^{-8} \left(\frac{\beta}{H_{tr}}\right) \left(\frac{T_{tr}}{1\,\mathrm{GeV}}\right) \left(\frac{g_{tr}}{100}\right)^{1/6} \mathrm{Hz} \quad (12)$$

where H_{tr} and g_{tr} are the Hubble parameter and the number of degrees of freedom at the transition temperature T_{tr}. For the electroweak phase transition (EWPT) $\beta/H_{tr} \sim 10^2 \div 10^3$, $T_{tr} \sim 100\,\mathrm{GeV}$ and the peak frequency is therefore in the range $10^{-4} \div 5 \times 10^{-3}$ Hz. A calculation of the peak of the spectrum gives [47]

$$h_0^2\, \Omega_{gw}(f_P) \sim 10^{-6} \xi^2 \frac{\alpha^2}{1+\alpha^2} \frac{v_b^3}{0.24 + v_b^3} \left(\frac{H_{tr}}{\beta}\right)^2 \left(\frac{100}{g_{tr}}\right)^{1/3} \quad (13)$$

where ξ is the fraction of latent heat converted into bubble-wall kinetic energy and v_b is the Hubble expansion velocity. If the Higgs mass is larger than the W-masses (as required by current phenomenological constraints), non-perturbative analysis based upon lattice simulations establishes that within the Standard Model the EWPT cannot be first-order. In the class of the minimal supersymmetric extensions of the standard model, provided that the Higgs mass lies in the range 110–115 GeV and the right-handed stop (i. e. the supersymmetric partner of the top quark) has a mass between 105 and 165 GeV, a first-order EWPT can take place. In this case the GWs spectrum is too weak. Within the larger class of non-minimal supersymmetric extensions of the standard model the strength of the transition can be enhanced: in particular, considering a class of models where an additional gauge singlet in the Higgs sector (thus also obtaining a model explaining the observed baryon asymmetry), for the GW spectrum peak values of $h_0^2\, \Omega_{gw}(f_p) \sim 10^{-15} \div 10^{-10}$ can be reached [48], corresponding to peak frequencies $f_p \sim 10\,\mathrm{mHz}$.

During a first-order phase transition, the onset of turbulent anisotropic eddies induced in the background fluid by the rapid expansion and collision of bubbles may also produce a GW background ([47,48,49,50]). Again within the non-minimal supersymmetric extensions of the standard model there are regions of the parameter space [48] where the GW spectrum can reach peak values $h_0^2\, \Omega_{gw} \sim 10^{-10}$ for frequencies $f_P \sim 10^{-3}$ Hz. Cosmic turbulence with additional magnetic fields (affecting the turbulent energy spectrum) can also generate a GW background at the end of a first-order phase transition. In [51] a scenario is analyzed where a first-order phase transition occurs before neutrino decoupling (i. e. much later than EWPT); the corresponding GW spectrum presents a peak in the frequency window relevant for the LISA experiment.

Finally, phenomenologically interesting GW spectra can be produced in non-minimal inflationary models (the so-called extended inflation models) where the accelerating phase ends with bubble collision [52]: the peak frequency of the spectrum lies in the frequency range of ground-based detectors. A detailed analysis [53] has shown that in two-field models where one field performs the first-order phase transition and the other drives the accelerating phase by slowly rolling towards the minimum of the potential, if the phase transition takes place well before the end of inflation, the spectrum has a peak in the frequency range $[10, 10^3]$ Hz and can be detectable by large-scale, earth-based interferometers.

Future Directions

The search for primordial gravitational waves is rather challenging both experimentally and theoretically. While so little is known about the earliest evolution of the Universe (thus rendering the predictions for a cosmological background of gravitational waves as mere indications), the experimental effort to detect such a signal is rather challenging. Nevertheless, the development of space-based detectors, combined with the possibility that future CMB experiments could detect the tensor contributions of primordial metric perturbations hopefully will provide some concrete chances to detect primordial gravitational waves.

Bibliography

Primary Literature

1. Einstein A (1916) Sitzungsber Preuss Akad Wiss Berlin (Math Phys) 1916:688–696
2. Bondi H (1957) Nature 179:1072; (1960) ibid 186:535
3. Hulse RA, Taylor JH (1975) Astrophys J 95:L51–L53
4. http://www.geo600.uni-hannover.de
5. http://www.ligo.org
6. http://www.tama.mtk.nao.ac.jp
7. http://www.virgo.pi.infn.it
8. http://lisa.jpl.nasa.gov
9. http://science.hq.nasa.gov/universe/science/bang.html
10. Seto N, Kawamura S, Nakamura T (2001) Phys Rev Lett 87:221103
11. Maggiore M (2000) Phys Rep 331:283
12. Grishchuk LP (1975) Sov Phys JEPT 40:409
13. Starobinski AA (1979) JEPT Lett 30:682

14. Rubakov VA, Sazhin M, Veryaskin A (1982) Phys Lett B 115:189; Fabbri R, Pollock M (1983) Phys Lett B 125:445; Abbott LF, Harari DD (1986) Nucl Phys B 264:487; Allen B (1988) Phys Rev D 37:2078; Sahni V (1990) Phys Rev D 42:453
15. Witten E (1984) Phys Rev D 30:272
16. Hogan CJ (1986) Mon Not R Astr Soc 218:629
17. Allen B (1995) Les Houches 1995, Relativistic gravitation and gravitational radiation. In: Proceedings of Astrophysical Sources of Gravitational Radiation, 26 Sep–6 Oct, Les Houches, France, e-Print: gr-qc/9604033, p. 373–417
18. Christensen N (1992) Phys Rev D 46:5250; Flanagan EE (1993) Phys Rev D 48:2389; Allen B, Romano JD (1999) Phys Rev D 59:102001
19. Mandic V, Buonanno A (2006) Phys Rev D 73:063008
20. Abbott B et al (2007) Astrophys J 659:918
21. Armstrong JW, Estabrook FB, Tinto M (2001) Class Quantum Grav 18:4059; Tinto M, Armstrong JW, Estabrook FB (2001) Phys Rev D 63(R):021101
22. Schwartzmann VF (1969) JETP Lett 9:184
23. Copi CJ, Schramm DN, Turner MS (1997) Phys Rev D 55:3389
24. Sachs R, Wolfe A (1967) Astrophys J 147:73
25. Bennett C et al (1996) Astrophys J 464:L1
26. Hinshaw G et al (2007) Astrophys J Suppl 170:288
27. Allen B, Koranda S (1994) Phys Rev D 50:3713
28. Smith T, Pierpaoli E, Kamionkowski M (2006) Phys Rev Lett 97:021301
29. Thorsett S, Dewey R (1996) Phys Rev D 53:3468
30. Kamionkowski M, Kosowsky A, Turner M (1994) Phys Rev D 49:2837
31. Kinney WH, Kolb EW, Melchiorri A, Riotto A (2006) Phys Rev D 74:023502
32. Friedman BC, Cooray A, Melchiorri A (2006) Phys Rev D 74:123509
33. Ungarelli C, Vecchio A (2001) Phys Rev D 63:064030
34. Ungarelli C, Corasaniti PS, Mercer RA, Vecchio A (2005) Class Quantum Grav 22:S955
35. Gasperini M, Veneziano G (1993) Astropart Phys 1:317; (1993) Mod Phys Lett A 8:3701; (1994) Phys Rev D 50:2519; (2003) Phys Rep 373:1
36. Brustein R, Gasperini M, Giovannini M, Veneziano G (1995) Phys Lett B 361:45
37. Buonanno A, Maggiore M, Ungarelli C (1997) Phys Rev D 55:3330
38. Birrell ND (1982) Davis PCW Quantum fields in curved space. Cambridge University Press, Cambridge
39. Vilenkin A (1981) Phys Lett B 107; (1985) Phys Rep 121:263
40. Vilenkin A, Shellard EPS (1994) Cosmic strings and other Topological Defects. Cambridge University Press, Cambridge
41. Caldwell RR, Allen B (1992) Phys Rev D 45:3447; Caldwell RR, Battye RA, Shellard EPS (1996) Phys Rev D 54:7146
42. Damour T, Vilenkin A (2000) Phys Rev Lett 85:3761; Phys Rev D 64:064008 (2001); Phys Rev D 71:063510 (2005)
43. Siemens X et al (2006) Phys Rev D 73:105001
44. Jones N et al (2002) High J Energy Phys 07:051; Sarangi S, Henry SH Tye (2002) Phys Lett B 536:185; Dvali G, Vilenkin A (2004) Cosmol J Astropart Phys 03:010; Jones N et al (2003) Phys Lett B 563:6; Copeland EJ et al (2004) J High Energy Phys 06:013
45. Jackson MG et al (2005) J High Energy Phys 10:013
46. Siemens X, Mandic V, Creighton J (2007) Phys Rev Lett 98:111101
47. Kosowsky A, Turner MS, Watkins R (1992) Phys Rev D 45:4514; Phys Rev Lett 69:2026 (1992); Kosowsky A, Turner MS (1993) Phys Rev D 47:4372; Kamionkowski M, Kosowsky A, Turner MS (1994) Phys Rev D 49:2837
48. Apreda R, Maggiore M, Nicolis A, Riotto A (2001) Class Quant Grav 18:L155; Nucl Phys B 631:342 (2002)
49. Kosowsky A, Mack A, Kahniashvili T (2002) Phys Rev D 66:024030
50. Gogoberidze G, Kahniashvili T, Kosowsky A (2007) Phys Rev D 76:083002
51. Dolgov AD, Grasso D (2002) Phys Rev Lett 88:011301; Dolgov AD, Grasso D, Nicolis A (2002) Phys Rev D 66:103505
52. Turner MS, Wilczek F (1990) Phys Rev Lett 65:3080
53. Baccigalupi C, Amendola L, Fortini P, Occhionero F (1997) Phys Rev D 56:4610

Books and Reviews

Buonanno A (2006) Gravitational Waves. In: Proceedings of Les Houches Summer School-Session 86: Particle Physics and Cosmology: The Fabric of Spacetime. Les Houches, France (to appear). e-Print: arXiv:0709.4682 [gr-qc]

Gasperini M (2007) Elements of String Cosmology. Cambridge University Press, Cambridge

Maggiore M (2007) Gravitational Waves. Vol 1: Theory and Experiments. Oxford University Press, Oxford

Mukhanov V (2005) Physical foundations of cosmology. Cambridge University Press, Cambridge

Schutz BF (1985) A First Course in General Relativity. Cambridge University Press, Cambridge

Cosmic Strings

Ana Achúcarro[1,2], Carlos J. A. P. Martins[3,4,5]

[1] Institute Lorentz of Theoretical Physics, University of Leiden, Leiden, The Netherlands

[2] Department of Theoretical Physics, University of the Basque Country UPV-EHU, Leioa, Spain

[3] Centro de Astrofísica, Universidade do Porto, Porto, Portugal

[4] Centro de Física do Porto, Porto, Portugal

[5] DAMTP, CMS, University of Cambridge, Cambridge, UK

Article Outline

Glossary
Definition of the Subject
Introduction
String Formation
String Evolution
Astrophysical and Cosmological Consequences
Field Theory Strings with More Degrees of Freedom
Cosmic Superstrings
Future Directions
Bibliography

Glossary

Branes, D-branes Brane – after membrane – is the name given generically to the "solitons" (the non-perturbative states) of Superstring theory. The most important class is that of Dirichlet- or D-branes, which are dynamical hypersurfaces in which the fundamental strings can end. The number of spatial dimensions of the brane is sometimes indicated explicitly (e. g. D3-branes, D7-branes). A D1-brane is also known as a D-string.

Cosmic strings Cosmic strings are non-dissipative linear concentrations of energy that form at phase transitions in the early universe at which axial symmetries are broken. They are extremely long and thin, their length could be many times the size of the observable universe while their thickness is usually subatomic.

Cusps Closed loops of string undergo periodic oscillations, with a period related to the size of the loop. The dynamical equations predict that during each oscillation there may be a few points at which the string instantaneously doubles back on itself, dubbed cusps. In the neighborhood of the cusp, the string velocity approaches the speed of light. Such an event generates an intense pulse of gravitational and other forms of radiation, strongly beamed in the direction of motion of the cusp (like the sound produced by the cracking of a whip).

Deficit angle The spacetime metric around a straight static cosmic string looks like Minkowski space in cylindrical coordinates, except that the azimuthal coordinate has a range smaller than the usual 2π. This means that the spacetime is actually conical with a global deficit angle $\alpha = 8\pi G\mu/c^2$ if μ is the mass per unit length of the string. Pictorially, an angular wedge of width α is removed from the plane orthogonal to the string and the remaining edges identified.

Horizon, cosmological Roughly speaking, the size of the observable universe at any given time. It is the maximum distance light could have traveled since the Big Bang, and therefore the maximum distance over which physical phenomena can be causally correlated, since information cannot travel faster than light.

Goto–Nambu action An Action principle – in the sense of Lagrange – that describes the relativistic motion of an idealized, infinitely thin (Goto–Nambu) cosmic string with no internal structure and whose energy per unit length equals its tension. It is a generalization of the Action for a relativistic point particle, being proportional to the invariant area of the surface swept out in spacetime by the string motion. It provides a very good approximation to the dynamics of a relativistic magnetic flux line whenever the effects of the string core and of massive excitations transverse to the string can be ignored.

Inflation, cosmological A near-exponential expansion of the (extremely smooth) early universe in which quantum fluctuations would have been frozen and blown up to cosmological size, providing the initial inhomogeneities that would later grow, under the effect of gravity, into large scale structures such as galaxies and clusters visible in the universe today.

Kibble mechanism The inevitable formation of cosmic strings, if they are stable, in rapid phase transitions. After a symmetry-breaking phase transition the choice of a minimum depends on random fluctuations, and will be different in different regions. Neighboring regions in different ground states will have defects separating them. Below a critical temperature the field fluctuations are not enough to change between ground states and the domains effectively freeze. In a cosmological context, the typical scale of defects at formation must be smaller than the causal horizon.

Scaling The notion that a single lengthscale characterizes the evolution of a string network, so that it always looks the same (in a statistical sense) when relevant quantities are re-scaled by that lengthscale. A particular case relevant for cosmology is linear scaling, where this scale is the horizon size. A structureless (Goto–Nambu) cosmic string network usually evolves in such a scale-invariant manner, with about 40 strings crossing any horizon volume throughout the history of the universe. The long string network is diluted by the creation of small closed loops which oscillate and decay into a stochastic background of radiation, cosmic rays and gravitational waves.

Superstrings, cosmic Superstring theory is a candidate model for a quantum theory of gravity unified with all other interactions. It is based on the idea that the fundamental constituents of nature are one-dimensional "strings" whose vibrational modes produce all known particles and interactions (the prefix "super" refers to a symmetry between bosonic and fermionic excitations, known as supersymmetry). These fundamental superstrings are not to be confused with cosmic strings, but superstrings can also be cosmic in some particular models, in which case they are known as cosmic superstrings.

VOS model Shorthand for Velocity-dependent One-Scale model, a quantitative and physically simple analytic model for the evolution of Goto–Nambu string networks. Comparison with numerical simulations

shows that the model provides an accurate description of the large-scale features of the network through two averaged (macroscopic) quantities, a characteristic length scale (or correlation length) and the root-mean square string velocity.

Worldsheet A point particle's position as a function of time traces a line in spacetime known as a worldline. A one-dimensional object such as a string traces a two-dimensional worldsheet.

Definition of the Subject

Cosmic strings are linear concentrations of energy that form whenever phase transitions in the early universe break axial symmetries, as originally shown by Kibble [31]. They are the result of frustrated order in the quantum fields responsible for elementary particles and their interactions. For about two decades, motivation for their study was provided by the possibility that they could be behind the density inhomogeneities that led to the observed large-scale structures in the universe. Precision observations, particularly of the cosmic microwave background radiation, have limited strings to a sub-dominant role in structure formation. Instead, the inhomogeneities appear to be consistent with a period of cosmological inflation, but it turns out that particle-physics models of the early universe that predict a period of inflation very often also predict the generation of cosmic strings at the end of it [27,28].

More recently, interest has been revived with the realization that there may be strong links between field theory cosmic strings and fundamental strings. The latter are the supposed ultimate building blocks of matter, and in their original context of superstring theory were thought to be microscopic. However, in its modern version – sometimes referred to as M-theory – it is possible and perhaps even mandatory to have macroscopic (cosmological-sized) fundamental strings [37,53]. Their behavior is expected to be quite similar to that of field theory cosmic strings, although there are some important differences so they may in principle be observationally distinguishable. Being relics of the phase transitions that produced them, cosmic strings provide us with a unique window into the early universe. If they are stable and survive for a significant amount of time (possibly even up to the present day), they may leave an imprint in many astrophysical and cosmological observables, and provide us with information on fundamental physics and the very early universe that would otherwise be inaccessible to us. On the other hand, gaining a quantitative understanding of their properties, interactions, evolution and consequences represents a significant challenge because of their intrinsic complexity. Their non-linearity is particularly noteworthy, with highly non-trivial feedback mechanisms between large (cosmological) and small (microscopic) scales affecting the network dynamics. Considerable reliance, therefore, must be placed on numerical simulations, which are technically difficult and computationally costly. A complementary approach is the use of analytic or semi-analytic models, usually to describe the large-scale features of the networks.

The basic picture of the cosmological evolution of string networks that has emerged for the simplest, Goto–Nambu, networks is of a scaling solution with about 40 long strings always stretching across each horizon volume plus a population of loops (other string types can lead to a different behavior). It is then possible to estimate their cosmological implications quantitatively. For example, these strings continuously source gravitational perturbations on sub-horizon scales. The one parameter in these models is the energy scale of the phase transition at which the strings are created. The astrophysical consequences of strings stem from the non-trivial gravitational field around a string [58]. Particles in the vicinity of a static straight string feel no gravitational acceleration, because in general relativity tension is a negative source of gravity and, since tension equals energy per unit length, their effects cancel. The space-time around the string is locally, but not globally, flat. In fact the space is conical, with a deficit angle

$$\alpha = 8\pi \frac{G\mu}{c^2} , \tag{1}$$

where μ is the mass per unit length; the simple way to picture this is to imagine a plane in which an angular wedge α has been removed and the edges glued together. For cosmologically interesting strings the deficit angle ranges from a few seconds of arc to a few millionths of a second of arc.

Introduction

To understand the cosmological evolution and effects of cosmic strings we start in this section with a quick summary of basic cosmology concepts that will be needed later and a description of the simplest type of cosmic strings in which the main features are already apparent. But, first, a warning about units: from now on we will set the speed of light to unity, $c = 1$, so we measure distances in light-travel time, and masses in units of Energy (and vice-versa, using $E = mc^2$). Boltzmann's constant is set to unity, so temperature is measured in units of mass/energy (using $E = K_B T$). Finally, we set Planck's constant to unity, $\hbar = 1$, and measure all lengths and masses in units of Planck's length and mass/energy: $l_P = 1.62 \times 10^{-35}$ m,

$M_P = 2.18 \times 10^{-8}\,\mathrm{kg} = 1.22 \times 10^{19}\ \mathrm{GeV/c^2}$. In these units, Newton's constant is given by $G = M_P^{-2}$.

The early universe is very smooth. To a very good approximation it is a homogeneous and isotropic spacetime described by a single variable: the rate of expansion of its three-dimensional spatial sections. In Einstein's general relativity this spacetime is described by the flat Friedmann–Robertson–Walker (FRW) metric

$$\mathrm{d}s^2 = \mathrm{d}t^2 - a^2(t)[\mathrm{d}\vec{x} \cdot \mathrm{d}\vec{x}]\,, \tag{2}$$

where $\vec{x}$ are fixed (comoving) spatial coordinates and $a(t)$ is the scale factor that determines the fractional or *Hubble expansion rate*

$$H(t) = \frac{1}{a(t)}\frac{\mathrm{d}a}{\mathrm{d}t}\,. \tag{3}$$

The time coordinate t is known as cosmological time; to analyze cosmic string evolution we will also need a different time parametrization known as conformal time, τ. They are related by $\mathrm{d}\tau = \mathrm{d}t/a(t)$, leading to the metric

$$\mathrm{d}s^2 = a^2(\tau)[\mathrm{d}\tau^2 - \mathrm{d}\vec{x} \cdot \mathrm{d}\vec{x}]\,. \tag{4}$$

The age of the universe is currently estimated to be about 13.7 billion years [55]. The universe starts very hot and dense and is cooled by the expansion, with the temperature decreasing as $T(t) \sim a(t)^{-1}$. The Hubble expansion rate is determined by the energy contents of the universe. In a universe dominated by radiation or very relativistic matter (the hottest, earliest stages), the scale factor evolves as $a(t) \sim t^{1/2}$ and the energy density in radiation as $\rho_{\mathrm{radiation}} \sim a(t)^{-4} \sim t^{-2}$. The energy density of non-relativistic matter is inversely proportional to volume $\rho_{\mathrm{matter}} \sim a(t)^{-3}$ and eventually takes over (after about 4000 years), leading to a period of matter domination, during which $a(t) = t^{2/3}$ and therefore $\rho_{\mathrm{matter}} \sim t^{-2}$. More recently – about five billion years ago – we have entered an epoch of accelerated expansion due possibly to a cosmological constant or some unknown form of *dark energy* whose energy density is constant in time $\rho_{\mathrm{dark\ energy}} \sim$ const. Dark energy should not be confused with dark matter, an unknown form of matter whose presence we can detect through its gravitational effects but that does not interact with electromagnetic fields and so in particular does not emit light – hence the adjective "dark"–. In the currently accepted cosmological model the energy density in the universe today would be dominated by dark energy (about 74%), followed by about 22% dark matter and only about 4% of regular (baryonic) matter [55]. Dark matter is widely believed to be a particle still to be discovered.

The universe today is far from smooth, but the structure we observe on the scale of clusters of galaxies is consistent with the gravitational collapse of tiny primordial density inhomogeneities $\delta\rho/\rho \sim 10^{-5}$ at the time the cosmic microwave background (CMB) radiation was emitted. The CMB is the oldest radiation we observe, dating back to the time when the universe was only about 380 000 years old. At this epoch the primordial plasma cooled enough to allow the formation of the first atoms (a process known as recombination), and it became transparent to photons (which is referred to as decoupling). Before that moment, the photons behave like a fluid that is strongly coupled to the protons and electrons. An overdense region in the baryon fluid would like to contract but the photon pressure pushes it back, causing both fluids to oscillate. These oscillations are imprinted in the cosmic microwave background and can be detected today in the form of Doppler peaks in its power spectrum.

The spectrum of density inhomogeneities has been accurately measured in the CMB and found to be near scale-invariant and of the right magnitude to produce the structure we observe. The perturbations to the FRW metric can be classified as *scalar* (overall changes to the Newtonian gravitational potential), *vector* (associated with velocity and/or rotational effects) and *tensor* (transverse traceless perturbations to the spatial metric, such as gravitational waves). Each of these affects the CMB in different ways, so their relative contributions can in principle be observationally distinguished. Finally, Thomson scattering of the anisotropic distribution of the CMB photons is particularly important during decoupling and recombination, and induces a partial linear polarization of the scattered radiation, at a level that is around ten percent of the anisotropy. Detection of this polarization signal is at the borderline of the sensitivity of ongoing experiments at the time of writing, but is expected to become standard with forthcoming experiments.

The energy density of a network of cosmic strings in the linear scaling regime is $\rho_{\mathrm{strings}} \sim t^{-2}$ and therefore it remains a constant fraction of the dominant form of energy during matter or radiation domination. Numerical estimates for the simplest, Goto–Nambu, networks suggest the fraction is around $100G\mu$. Provided the string mass is not close to the Planck scale, this is small enough not to disturb the cosmological evolution; at the same time, for a broad range of values of $G\mu$ this is large enough to be detectable in precision experiments today. Other string types (see Sect. "Field Theory Strings with More Degrees of Freedom") may have larger or smaller fractions or qualitatively different signatures. In particular, networks that do not reach linear scaling may come to dominate

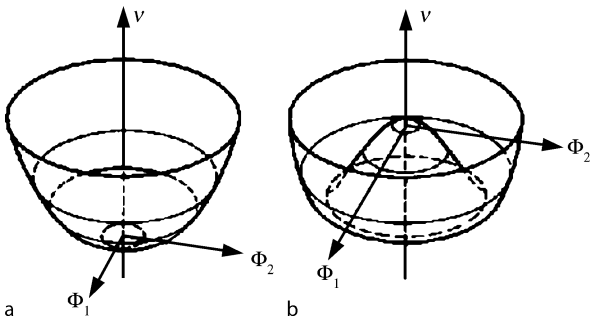

Cosmic Strings, Figure 1
The effective potential energy *V* for a simple string-forming field theory model. The a and b plots correspond to the high and low temperature configurations, respectively. For simplicity the complex field $\mathbf{\Phi}$ has been split into two real scalar fields, $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$

the energy density (which rules them out) or to disappear completely.

The simplest field theory model that produces cosmic strings has a single complex scalar field $\mathbf{\Phi}$ (this is shorthand for a function $\mathbf{\Phi}(t,\vec{x})$ with complex values that do not change under coordinate transformations). Let us assume that the Hamiltonian determining the field dynamics is invariant under an axial symmetry such as a phase rotation, $\mathbf{\Phi} \to \mathbf{\Phi}e^{i\theta}$. For example, take the potential energy

$$\int d^3x\, V = \int d^3x \frac{\lambda}{2}\left(|\mathbf{\Phi}|^2 - \eta^2\right)^2 \tag{5}$$

where λ is a dimensionless coupling constant and η is an energy scale related to the temperature of the symmetry-breaking transition. This has a set of degenerate ground states: the minimum of the potential in field space is the circle $|\mathbf{\Phi}| = \eta$, known as the *vacuum manifold*. Any configuration $\mathbf{\Phi}(t,\vec{x}) = \text{const.} = \eta e^{i\chi}$ with χ real and constant is a possible ground state or *vacuum*, irrespective of the value of the phase χ.

Figure 1 illustrates what happens. At high temperature the field fluctuations are large enough to make the central peak around $|\mathbf{\Phi}| = 0$ irrelevant, and the effective potential is symmetric and has a minimum there. As the temperature falls the energy will eventually be too low to permit fluctuations over the peak, at which point the field will tend to settle towards one of the ground states. The random choice of minimum in this condensation process then breaks the original axial symmetry. This is the case, for instance, in superfluid ^{4}He.

When a large system goes through a phase transition like this, each part of it has to make this random choice, which need not be the same everywhere. The minimization of gradient terms in the energy of the system tends to make it evolve towards increasingly more uniform configurations, but causality (the principle that no information can travel faster than light) imposes that this evolution can only happen at a limited rate. As a result one expects many domains, each with an uncorrelated choice of ground state. Where these domains meet there is some probability of forming linear defects – cosmic strings – around which the phase angle varies by 2π (or possibly multiples thereof). This is the Kibble mechanism. Notice that the field vanishes at the string's core, so there is trapped potential energy (as well as gradient energy). These strings are known as *global* strings because the axial symmetry that is broken below the phase transition is "global", that is, the transformation $\mathbf{\Phi} \to \mathbf{\Phi}e^{i\theta}$ is independent of position.

The next step is to consider charged scalar fields interacting with an electromagnetic field. The best known example of a symmetry–breaking transition of this kind is the condensation of Cooper pairs in a superconductor, that has the effect of making photons massive below the critical temperature (in this case the axial symmetry is of the "local" of "gauge" type). The cosmic strings that result are magnetic flux tubes that do not dissipate because the magnetic field is massive outside the string core.

This type of vortex was first discussed by Abrikosov [1] in the context of type II superconductors. Nielsen and Olesen [45] generalized these ideas to the relativistic quantum field theory models used in particle physics, in particular the Abelian Higgs model which is a relativistic version of the Landau–Ginzburg model of superconductivity, governed by the action

$$S = \int d^4x \left[|\partial_\mu \mathbf{\Phi} - iqA_\mu \mathbf{\Phi}|^2 - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} - \frac{\lambda}{2}(|\mathbf{\Phi}|^2 - \eta^2)^2 \right]. \tag{6}$$

A_μ is the gauge field and $\mathbf{\Phi}$ is a complex scalar of charge q ($q = 2e$ in superconductors, where $\mathbf{\Phi}$ is the Cooper pair wavefunction). The second term is the usual Maxwell action for the electromagnetic field, $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. The energy per unit length of a straight, static string lying on the z-axis is

$$E = \int d^2x \left[|\partial_x \mathbf{\Phi} - iqA_x \mathbf{\Phi}|^2 + |\partial_y \mathbf{\Phi} - iqA_y \mathbf{\Phi}|^2 + \frac{1}{2}B^2 + \frac{\lambda}{2}(|\mathbf{\Phi}|^2 - \eta^2)^2 \right] \tag{7}$$

where $B = \partial_x A_y - \partial_y A_x$ is the z-component of the magnetic field. Finite energy configurations must have $|\mathbf{\Phi}| = \eta$

(the *vacuum manifold* is still a circle) but the phase of $\mathbf{\Phi}$ is undetermined provided the gradient terms and the magnetic field go to zero fast enough. This condition allows for finite energy solutions $A_t = A_r = A_z = 0$, $\mathbf{\Phi}(r,\theta) \sim \eta e^{in\theta}$, $A_\theta(r,\theta) \sim n/(qr)$, as $r \to \infty$, in which the total magnetic flux in the plane perpendicular to the string is quantized,

$$\int d^2x B = \oint \vec{A} \cdot \vec{d}l = \frac{2\pi n}{q}$$

and n is the winding number of the string. If the constants λ and q are such that fluctuations in the scalar field $\mathbf{\Phi}$ and the gauge field A_μ have equal masses, it is possible to show that the string saturates an inequality of the form

$$\text{Energy per unit length} \geq \text{constant x} \mid \text{magnetic flux} \mid$$

known as the Bogomolnyi bound [7]. In this case, parallel strings at close range exert no force on each other and there are static multivortex solutions [56]. If the mass of the scalar excitations is lower (higher) than that of the gauge excitations, parallel strings will attract (repel).

More complicated particle physics models – in particular those describing the early universe – involve gauge symmetries that generalize the electromagnetic interaction, mediated by photons, to more complicated interactions such as the electroweak or Grand Unified (GUT) interactions. The messenger fields that play the role of the photons may be massless in the early universe and become massive following a symmetry-breaking transition, and cosmic strings carry the magnetic flux of these other massive gauge fields (not the electromagnetic field).

From a cosmological point of view, the gauge field has the important effect of making the gradient terms decay exponentially fast away from the string so the energy per unit length of these strings is finite. Abrikosov–Nielsen–Olesen strings have no long-range interactions, so their evolution is dominated by their tension and is well described in the thin string or Goto–Nambu approximation.

Field continuity implies that a string of this kind cannot simply come to an end: it must form a closed loop or extend to infinity, and it cannot break into segments. For this reason, strings, once formed, are hard to eliminate. In the absence of energy loss mechanisms, the strings would eventually dominate the energy density of the universe. On the other hand, the strings can decay into radiation, they may cross and exchange partners, and they may also cross themselves, forming a closed loop which may shrink and eventually disappear. The outcome of these competing mechanisms is that the network is expected to reach a scale-invariant (or *scaling*) regime, where the network's characteristic length scale is proportional to the size of the horizon. We will discuss string evolution in more detail in Sect. "String Evolution". If a random tangle of strings was formed in the early universe, there would always be some strings longer than the horizon, so a few would remain even today. Because cosmological phase transitions typically happen in the very early universe, cosmic strings contain a lot of trapped energy, and can therefore significantly perturb the matter distribution. To first order there is a single parameter quantifying the effects of strings, its energy per unit length. In the simple relativistic strings, the mass per unit length and the string tension are equal, because of Lorentz invariance under boosts along the direction of the string (but this need not be true for more elaborate models, see Sect. "Field Theory Strings with More Degrees of Freedom"). Cosmic strings are exceedingly thin, but very massive. Typically, for strings produced around the epoch of grand unification, the mass per unit length would be of order $\mu \sim 10^{21}$ kg m^{-1} and their thickness 10^{-24} m. The gravitational effects of strings are effectively governed by the dimensionless parameter $G\mu$, where G is Newton's constant. For GUT-scale strings, this is 10^{-6}, while for electroweak-scale strings it is 10^{-34}.

String Formation

Spontaneous symmetry breaking is a ubiquitous feature of our theories of fundamental particle interactions. Cosmic strings are formed in many symmetry-breaking phase transitions. If the symmetry is broken from a group G down to a subgroup H, the set of degenerate vacuum or ground states is the manifold $M = G/H$, and the topology of this manifold determines the types of defect that can form. In our previous examples M was a circle; in general, strings can form if M contains closed curves that cannot be contracted within M (the technical term is that M is not simply connected, or that its first homotopy group is nontrivial) [31].

The Kibble mechanism described in Sect. "Introduction" relates the initial density of strings to the size ξ of the domains over which the field is correlated,

$$\rho_{\text{string}} \sim C \frac{1}{\xi^2}$$

with C a constant of order one reflecting the probability that a string forms when three or more domains meet. The correlation length cannot grow faster than the speed of light so in the early universe an obvious upper bound

on ξ is the size of the horizon at the time of the phase transition. If the dynamics of the phase transition is known, ξ can be estimated more accurately. In a first order phase transition ξ is given by the typical distance between bubble nucleation sites, which depends on the nucleation rate. In second order phase transitions ξ depends on the critical exponents and the rate of cooling through the critical temperature, T_c, as shown by Zurek [65].

Vortex lines or topological strings can therefore appear in a wide range of physical contexts, from cosmic strings in the early universe through disclinations in room-temperature nematic liquid crystals, to magnetic flux tubes in some superconductors and vortex lines in low-temperature superfluid helium. These systems provide us with a range of opportunities to test aspects of the cosmic string formation and evolution scenario experimentally.

The Kibble mechanism in first order transitions was confirmed in experiments on nematic liquid crystals [15,16]. The Kibble–Zurek scenario for second order transitions has been experimentally verified in Superfluid ^{3}He [11,50] and in Josephson Tunneling Junction arrays [41].

The ^{3}He experiments in a rotating cryostat in Helsinki also confirmed the scale invariance of the initial distribution of loops, $n(R) \sim R^{-4}$, where $n(R)\mathrm{d}R$ is the number density of loops with radii between R and $R + \mathrm{d}R$, as predicted by Vachaspati and Vilenkin [61].

More recently, the formation of a defect network following the annihilation of ^{3}He–A / ^{3}He–B boundary layers has been observed [14]. The precise type of defects is still under investigation but this system constitutes an interesting analogue to the formation of strings from the annihilation of branes in brane inflation scenarios.

There are also a few systems where the string density disagrees with the Kibble–Zurek predictions. In ^{4}He, [21] the reasons are understood: the strings are fuzzy and the network does not survive long enough to be detected [36]. In the case of superconducting films the results are somewhat inconclusive [40] and also it is not completely clear what the expected density of flux quanta should be after a temperature quench; an alternative formation mechanism with different vortex clustering properties has been proposed in [24]. In fact the formation of defects in systems with gauge fields is clearly very relevant to cosmology but is still not completely understood (see Kibble [33] for a recent discussion).

String Evolution

The motion of a cosmic string with worldsheet coordinates σ^a and background space-time coordinates x^μ in a metric $g_{\mu\nu}$ is obtainable from a variational principle applied to the Goto–Nambu action [22,43] ($a = 0, 1; \mu = 0, 1, 2, 3$)

$$S = \mu \times \text{Area} = \mu \int \mathrm{d}\tau\, \mathrm{d}\sigma |\det g_{ab}|^2$$
$$= \mu \int \mathrm{d}\tau\, \mathrm{d}\sigma \left| \det \begin{pmatrix} \dot{x}^\rho \dot{x}^\nu g_{\rho\nu} & \dot{x}^\rho x^{\nu\prime} g_{\rho\nu} \\ x^{\rho\prime} \dot{x}^\nu g_{\rho\nu} & x^{\rho\prime} x^{\nu\prime} g_{\rho\nu} \end{pmatrix} \right|^{\frac{1}{2}} \quad (8)$$

where μ is again the string mass per unit length, and with dots and primes respectively denoting derivatives with respect to the time-like (σ^0) and space-like ($\sigma^1 = \sigma$) coordinates on the worldsheet. g_{ab} is called the induced metric. We are interested in strings in a FRW background space-time (see Eq. (4)) and can choose worldsheet coordinates that make the induced metric diagonal

$$\sigma^0 = \tau\,, \quad \dot{\mathbf{x}} \cdot \mathbf{x}' = 0\,, \quad (9)$$

The choice of conformal time coordinate simplifies the microscopic evolution equations, although as we shall see later on physical time is a more natural choice for the macroscopic evolution (see Sect. "Introduction" for definitions of the two time choices). It is also useful to define the coordinate energy per unit σ,

$$\epsilon^2 = \frac{\mathbf{x}'^2}{1 - \dot{\mathbf{x}}^2}\,. \quad (10)$$

Then the usual variational techniques can be used to show that the microscopic string equations of motion are

$$\ddot{\mathbf{x}} + 2\frac{\dot{a}}{a}\dot{\mathbf{x}}(1 - \dot{\mathbf{x}}^2) = \frac{1}{\epsilon}\left(\frac{\mathbf{x}'}{\epsilon}\right)' \quad (11)$$

and

$$\dot{\epsilon} + 2\epsilon\frac{\dot{a}}{a}\dot{\mathbf{x}}^2 = 0\,. \quad (12)$$

For simplicity we are neglecting effects such as cusps and a frictional force due to particle scattering (which for heavy strings is only relevant during a transient period very early in the network's evolution). The first one is just a wave equation with a particular damping term (provided by the expansion of the universe). The damping also has the effect of reducing the coordinate energy per unit σ.

As was mentioned earlier the expansion of the universe stretches the strings, so in the absence of energy loss mechanisms their energy would grow with the scale factor and the string network would eventually become the dominant component of the universe's energy density – which would be in conflict with observational results. Such decay mechanisms do exist (at least for the simplest models), being ultimately due to radiation losses and to the fact that whenever strings interact they will reconnect [42,51]. In particular closed loops may be formed, and these subsequently

oscillate and eventually decay. This decay is thought to be mainly into gravitational radiation, but other forms of radiation are also produced very efficiently.

Provided the decay rate is high enough, the network will not have pathological consequences, but will instead reach a linear scaling solution, where the string density is a constant fraction of the background density and on large scales the network looks the same (in a statistical sense) at all times. Scaling is in fact an attractor solution, as has been shown both using numerical simulations and analytic models. Physically, the reason for this is that if one has a high density of strings then the number of string interactions increases and therefore loop production becomes more efficient and the decay rate increases. Conversely if the density is too low then there are few interactions and the decay rate is correspondingly lower. Numerical simulations confirm this broad picture, but also reveal that string evolution is a complex non-linear process, involving non-trivial interactions between various different scales.

There have been thus far two generations of numerical simulations of Goto–Nambu cosmic string networks in expanding universes. The first (Albrecht and Turok [2], Bennett and Bouchet [9], Allen and Shellard [3]) dates from around 1990, at the peak of the interest in cosmic strings as possible seeds for the large-scale structures we observe today. In the last few years, the renewed interest in strings in the context of models with extra dimensions led to a second generation of simulations (Martins and Shellard [39], Ringeval et al. [49], Olum and Vanchurin [46]), which build upon previous knowledge and exploit the dramatic improvements in hardware and software in the intervening decade and a half to achieve a much higher resolution.

A different approach is provided by full field theory simulations [62]. These are closer to the microphysics of the defects and provide important information on the interactions of the defects and their energy loss mechanisms, but their shorter dynamic range means that they are not optimal for understanding the non-linear feedback mechanisms between widely different scales which affect the dynamics of the network. From this point of view they play a very important role as calibrators, both for Goto–Nambu simulations and for analytic models. One can also carry out Minkowski space simulations (either of Goto–Nambu or field theory type). Neglecting the expansion of the universe is numerically desirable, since such simulations are much easier to implement and evolve much faster. However, the expansion plays a non-trivial role in the network dynamics, so these results should not be naively extrapolated to realistic cosmological scenarios.

Initial conditions for the numerical simulations are usually set up using the Vachaspati–Vilenkin algorithm [61]. One often adds to it random initial velocities, since these tend to enhance the rate of relaxation. All simulations agree on the broad, large-scale features of string networks, and in particular on the fact that the linear scaling solution is an attractor for the evolution. In Goto–Nambu simulations, the initial fraction of the total energy in the form of closed loops is around 20%, but in the linear scaling regime this fraction is around 50% or even slightly more. On the other hand, in field theory simulations this fraction tends to be somewhat smaller.

The first-generation simulations suggested a dynamical picture where the long-string network lost energy to large, long-lived loops, with sizes of order the correlation length. Refinements had each loop self-intersecting into around 10 daughter loops, but loop production from the long strings was essentially monochromatic. The second-generation simulations, however, reveal a quite different picture. Large loops do self-intersect (and indeed the number of daughter loops produced by each one seems to be around 20), but there is also a direct production of large quantities of small loops from slowly moving long-strings with fractal-like substructure. In other words, the loop production is in fact bi-modal. All three second-generation simulations agree on this broad picture, though not on which of the two loop production scales is dominant.

The second-generation simulations present some tentative evidence for the scaling of small-scale features of the network. An open question is whether or not this is expected to happen, given that gravitational backreaction (which would provide a characteristic scale) is not included in any of the network simulations carried out to date. One possible explanation stems from the fact that large loops are not scaled up versions of small loops. Indeed, small loops tend to be nearly circular, whereas large loops are not only far from circular but even far from planar. In other words, the self-intersection probability for a given loop depends on its size, and this may be sufficient to dynamically select a preferred scale. Incidentally, the loop fragmentation processes in these networks highlight the fact that there is a steady flow of energy from large to small scales which is entirely analogous to a Richardson cascade in turbulence. (In this case energy enters via long strings falling inside the horizon, and leaves via radiative decays.)

Figure 2 shows some relevant quantities characterizing Goto–Nambu string networks in the linear scaling regime. These are always plotted as a function of relative scale, to the network's correlation length. The top left panel shows the correlation function for the velocity vectors. A first striking feature is that in the expanding universe cosmic string velocities are anti-correlated on scales around the

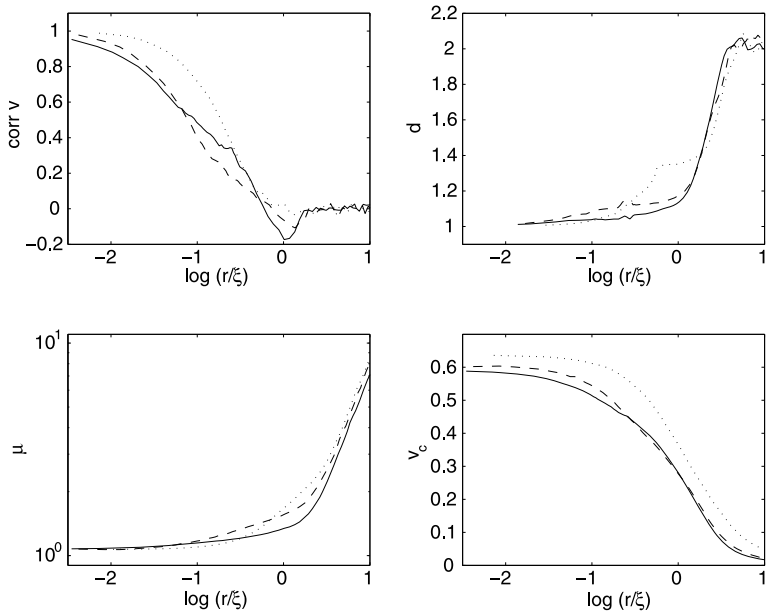

Cosmic Strings, Figure 2

Characteristic small-scale properties of cosmic string networks in the linear scaling regime, for matter (*solid lines*), radiation (*dashed*) and Minkowski spacetime (*dotted*) runs. In all plots the horizontal axis represents the logarithm of the physical lengthscale relative to the correlation length of the network. The simulations leading to these results are described in [39]. Top panels show the correlation function for the velocity vectors and the fractal dimension, *bottom panels* show coarse-grained mass per unit length and coherent velocity. The plotted quantities are described in the main text

correlation length (which is smaller than but comparable to the causal horizon), but such a feature is not present in Minkowski space. This anti-correlation is the result of a 'memory' of the network for recent reconnection events, and it is ultimately due to the damping effect of expansion. The top right panel depicts the fractal dimension of the network: this interpolates between $d = 1$ (straight segments) on small scales and $d = 2$ (Brownian network) on large scales, but it does so in a non-trivial way (which is again different depending on whether or not there is expansion) and over a wide range of scales. The fractal dimension evolves with time, decaying on any given physical scale: the strings continually become smoother on any scale, so as to minimize energy. Finally, the bottom panels show the renormalized (or 'coarse-grained') string mass per unit length on the left, and the corresponding coherent velocity on the right panel – notice that the effect of expansion is to reduce the velocities on any given scale.

The technical difficulty and computational cost of numerical simulation provide strong motivation for alternative analytic approaches, which essentially abandon the detailed statistical physics of the string network to concentrate on its thermodynamics. The best example is the velocity-dependent one-scale (VOS) model [38], which builds on previous work by Kibble and Bennett [8,32] and has demonstrated quantitative success when compared with both field theory and Goto–Nambu numerical simulations. The 'one-scale' assumption is that the network has a single characteristic lengthscale, which coincides with the string correlation length and the string curvature radius. This is an approximation which can be tested numerically.

The first assumption in this analysis is to localize the string so that we can treat it as a one-dimensional line-like object. This is clearly a good assumption for gauged strings, such as magnetic flux lines, but may seem more questionable for strings possessing long-range interactions, such as global strings or superfluid vortex lines. However, good agreement between the VOS model and simulations has been found in both cases. The second step is to average the microscopic string equations of motion to derive the key evolution equations for suitable macroscopic quantities, specifically its energy E and RMS velocity v defined by

$$E = \mu a(\tau) \int \epsilon \, \mathrm{d}\sigma \ , \quad v^2 = \langle \dot{\mathbf{x}}^2 \rangle = \frac{\int \dot{\mathbf{x}}^2 \epsilon \, \mathrm{d}\sigma}{\int \epsilon \, \mathrm{d}\sigma} \ . \qquad (13)$$

Notice that the energy is an 'extensive' quantity but the RMS velocity is an averaged quantity (and the averaging is weighted by the coordinate energy ϵ). In keeping with the

above coordinate choices, the microscopic quantities (on the right-hand side of both equations) are defined in terms of conformal time, but it turns out that the macroscopic evolution that we are now considering is best described in terms of physical time – please refer to the cosmology review for the explicit relation between the two.

Any string network divides fairly neatly into two distinct populations, *viz.* long (or 'infinite') strings and small closed loops. In the following we will focus on the long strings. The long string network is a Brownian random walk on large scales and can be characterized by a correlation length L, which can be used to replace the energy $E = \rho V$ in long strings in our averaged description, that is,

$$\rho \equiv \frac{\mu}{L^2} \,. \tag{14}$$

A phenomenological term must then be included to account for the loss of energy from long strings by the production of loops, which are much smaller than L. A *loop chopping efficiency* parameter $\tilde{c}$ is introduced to characterize this loop production as

$$\left(\frac{\mathrm{d}\rho}{\mathrm{d}t}\right)_{\text{to loops}} = \tilde{c} v \frac{\rho}{L} \,. \tag{15}$$

In this approximation, we would expect the loop parameter $\tilde{c}$ to be a constant; comparison with numerical simulations suggests $\tilde{c} \sim 0.23$.

From the microscopic string equations of motion, one can then average to derive the evolution equation for the correlation length L,

$$2\frac{\mathrm{d}L}{\mathrm{d}t} = 2HL(1+v^2) + \tilde{c}v \,, \tag{16}$$

where H is the Hubble expansion rate defined in Eq. (3). The first term in (16) is due to the stretching of the network by the Hubble expansion which is modulated by the redshifting of the string velocity, while the second is the loop production term. One can also derive an evolution equation for the long string velocity with only a little more than Newton's second law

$$\frac{\mathrm{d}v}{\mathrm{d}t} = \left(1-v^2\right)\left(\frac{k(v)}{L} - 2Hv\right) \,. \tag{17}$$

The first term is the acceleration due to the curvature of the strings and the second is the damping term from the Hubble expansion. Note that strictly speaking it is the curvature radius R which should appear in the denominator of the first term. In the present context we are identifying $R = L$. The function $k(v)$ is the *momentum parameter*, defined by

$$k(v) \equiv \frac{\langle (1-\dot{\mathbf{x}}^2)(\dot{\mathbf{x}}\cdot\mathbf{u})\rangle}{v(1-v^2)} \,, \tag{18}$$

with $\dot{\mathbf{x}}$ the microscopic string velocity and $\mathbf{u}$ a unit vector parallel to the curvature radius vector. For most relativistic regimes relevant to cosmic strings it is sufficient to define it as follows:

$$k_{\mathrm{r}}(v) = \frac{2\sqrt{2}}{\pi}\frac{1-8v^6}{1+8v^6} \,, \tag{19}$$

while in the opposite case ($v \to 0$), we have the non-relativistic limit $k_0 = 2\sqrt{2}/\pi$.

Scale-invariant attractor solutions of the form $L \propto t$ (or $L \propto H^{-1}$) together with $v = \text{const.}$, only appear to exist when the scale factor is a power law of the form

$$a(t) \propto t^{\beta} \,, \quad 0 < \beta = \text{const.} < 1 \,. \tag{20}$$

This condition implies that

$$L \propto t \propto H^{-1} \,, \tag{21}$$

with the proportionality factors dependent on the expansion rate β. It is useful to introduce the following parameters to describe the relative correlation length and density, defining them respectively as

$$L = \gamma t \,, \quad \zeta \equiv \gamma^{-2} = \rho t^2/\mu \,. \tag{22}$$

By looking for stable fixed points in the VOS equations, we can express the actual scaling solutions in the following implicit form:

$$\gamma^2 = \frac{k(k+\tilde{c})}{4\beta(1-\beta)} \,, \quad v^2 = \frac{k(1-\beta)}{\beta(k+\tilde{c})} \,, \tag{23}$$

where k is the constant value of $k(v)$ given by solving the second (implicit) equation for the velocity. It is easy to verify numerically that this solution is well-behaved and stable for all realistic parameter values.

If the scale factor is not a power law, then simple scale-invariant solutions like (23) do not exist. Physically this happens because the network dynamics are unable to adapt rapidly enough to the changes in the background cosmology. An example of this is the transition between the radiation and matter-dominated eras. Indeed, since this relaxation to a changing expansion rate is rather slow, realistic cosmic string networks are strictly speaking *never* in scaling during the matter-dominated era. Another example is the onset of dark energy domination around the

present day. In this case, the network is gradually slowed down by the accelerated expansion, and asymptotically it becomes frozen in comoving coordinates. The corresponding scaling laws for the correlation length and velocity are $L \propto a$ and $v \propto a^{-1}$.

Despite its success in describing the large-scale features of string networks, the VOS model has the shortcoming of not being able to account for the small-scale features developing on the strings as the network evolves, as clearly shown by numerical simulations. This small-scale structure is in the form of wiggles and kinks, and can be phenomenologically characterized by its fractal properties, as we have sketched above. As a first analytic simplification, the string wiggles can be characterized through a renormalized string mass per unit length that is larger than the bare (Goto–Nambu) mass. This effectively corresponds to considering a model with a non-trivial equation of state (the relation between the string tension and the mass per unit length), which turns out to be one among a larger class of models known as elastic string models. This kind of description has interesting parallels with the coarse-graining approaches that are typical of condensed matter.

A more radical approach is to explicitly abandon the one-scale assumption. This is done in the three-scale model [6], which distinguishes between the characteristic lengthscale (which is simply a measure of the total string energy in a given volume) and the persistence length (which is defined in terms of the invariant length along the string and corresponds to the correlation length or inter-string distance). Additionally there is a third lengthscale which approximately describes a typical scale of the small-scale wiggles. This kind of description is in principle highly flexible, though this can be considered a blessing and a curse. The downside is that one is forced to introduce a large number of (almost free) phenomenological parameters over which one has limited control even when comparing the model with simulations.

Having said that, the three-scale model does confirm, at least qualitatively, the expectations for the behavior of string networks. Scaling of the large scales (in this case the characteristic and persistence lengths) is found to be an attractor, just as in the VOS model. Depending on the behavior of small-scale structures, the two large length scales may reach scaling simultaneously or the former may do so before the latter – a behavior that has been seen in numerical simulations. As for the behavior of the small-scale structures, their evolution timescale is typically slower, and generically they only reach scaling due to the effects of gravitational backreaction (not included in numerical simulations). In the absence of gravitational backreaction, scaling of the small-scale characteristic length is contingent on the removal of a sufficiently large amount of small-scale structure from the long strings by radiation and loop production, which in the model is controlled by a parameter whose detailed behavior is not known.

Finally, an interesting and rather different approach starts out with the assumption that there is a range of scales where stretching due to the expansion is the dominant dynamical effect, even on scales well below the cosmological horizon. A sufficient condition for this is that one is assuming that the rate of string intercommutations is fixed in horizon units. This turns out to be sufficient to allow the construction of a statistical-type description based on two-point correlation functions [48]. Their results are to a first approximation dependent on a critical exponent which physically is related to the coherent string velocity on a given scale. Comparison with numerical simulations shows, as expected, that the best agreement is found around and just below the horizon scale.

A second assumption is that loop production at those scales is sufficiently localized to be describable as a perturbation. When loop production is thus folded into the analysis, the picture that ultimately emerges is of a complicated fragmentation cascade. In particular, this model provides supporting evidence for the two-population loop distribution picture outlined above and clearly seen in high-resolution simulations. There is a population of correlation-length sized loops, produced by direct long-string intercommutation, and a second population with sizes a few orders of magnitude below (quite possibly near the gravitational backreaction scale) and due to loop fragmentation. Whether or not the smoothing provided by gravitational radiation is necessary to yield scaling of the loop sizes is again not entirely clear at the moment, but it is in principle a question for which this formalism could provide an answer.

Astrophysical and Cosmological Consequences

As was mentioned in Sect. "Definition of the Subject", the spacetime around a straight cosmic string is flat. A string lying along the z-direction has an equation of state $p_z = -\rho$, $p_x = p_y = 0$ and therefore there is no source term in the relativistic version of the Poisson equation for the Newtonian gravitational potential

$$\nabla^2\phi = 4\pi G(\rho + p_x + p_y + p_z) = 0\,. \tag{24}$$

A straight string exhibits no analogue of the Newtonian pull of gravity on any surrounding matter. However, this does not mean the string has no gravitational impact at all. On the contrary, we will see that a moving string has dramatic effects on nearby matter or propagating microwave

background photons. It is not difficult to derive the spacetime metric about such a straight static string [58]. It has the simple form

$$ds^2 = dt^2 - dz^2 - dr^2 - r^2\, d\theta^2 \,, \tag{25}$$

which looks like Minkowski space in cylindrical coordinates, except for the fact that the azimuthal coordinate θ has a restricted range $0 \le \theta \le 2\pi(1 - 4G\mu)$. That is, the spacetime is actually conical with a global deficit angle

$$\alpha = 8\pi G\mu \,, \tag{26}$$

where an angular wedge of width α is removed and the remaining edges identified.

This deficit angle implies that the string acts as a cylindrical gravitational lens, creating double images of sources behind the string (such as distant galaxies), with a typical angular separation $\delta\theta$ of order α and no distortion [60]. This is illustrated in Fig. 3. A long string would yield a distinctive lensing pattern. We should expect to see an approximately linear array of lensed pairs, each separated in the transverse direction. In each lensing event the two images would be identical and have essentially the same magnitude, i. e. brightness. (Except if we happen to see only part of one of the images.) This is a very unusual signature, because most ordinary gravitational lenses produce an odd number of images of substantially different magnitudes. A number of string lensing event candidates have been discussed in the past, but no confirmed one is currently known.

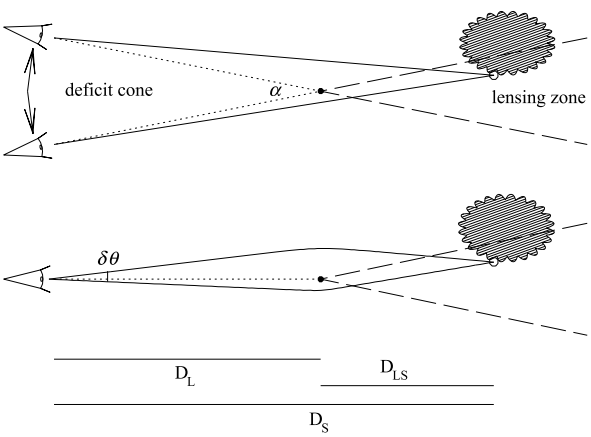

Cosmic Strings, Figure 3
An illustration of the mechanism behind lensing by cosmic strings. The *thick black dot* represents a cosmic string perpendicular to the page. The spacetime metric around the string can be obtained by removing the angular wedge of width α and identifying the edges. An observer can thus see double images of objects located on a certain zone behind the string. This zone is enclosed by the *dashed lines*, while the *solid lines* depict light rays and the angular separation of the two images, $\delta\theta$, will depend on the distances of the source and the observer to the string as well as on the deficit angle. Reprinted, with permission, from [35]

However, the above simple picture is complicated in practice by the fact that cosmic strings are not generally straight or static. Whenever strings exchange partners kinks are created that straighten out only very slowly, so we expect a lot of small-scale structure on the strings. Viewed from a large scale, the effective tension and energy per unit length will no longer be equal. Since the total length of a wiggly string between two points is greater, it will have a larger effective energy per unit length, U, while the effective tension T, the average longitudinal component of the tension force, is reduced, so $T < \mu < U$. This means that there is a non-zero gravitational acceleration towards the string, proportional to $U - T$. Moreover, the strings acquire large velocities, generally a significant fraction of the speed of light, which introduces further corrections to the deficit angle.

Another effect is the formation of over-dense wakes behind a moving cosmic string [52]. When a string passes between two objects, these are accelerated towards each other to a velocity

$$u_\perp = 4\pi G\mu v \,, \tag{27}$$

where v is the string velocity. Matter therefore collides in a sheet-like structure, leaving a wake behind the moving string. This was the basic mechanism underlying the formation of large-scale structures in cosmic string models. This model has significant attractions, such as the early formation of nonlinear structures, and one can get a good match to the observed galaxy power spectrum in models with a large cosmological constant. However, as we shall discuss, it fails to reproduce the power spectrum of CMB anisotropies observed by COBE, WMAP and other experiments; cosmic strings, therefore, can only play a subdominant role in structure formation (albeit still significant, at the ten to twenty percent level). Cosmic strings create line-like discontinuities in the cosmic microwave background signal [23,34]. For the same reason that wakes form behind a cosmic string, the CMB source on the surface of last scattering is boosted towards the observer, so there is a relative CMB temperature shift across a moving string (a red-shift of the radiation ahead of it, and a blue-shift of that behind), given by

$$\frac{\delta T}{T} \sim 8\pi G\mu v_\perp \,. \tag{28}$$

where $v_\perp$ is the component of the string velocity normal to the plane containing the string and the line of

sight. This is known as the Kaiser–Stebbins effect. This simple picture is again complicated in an expanding universe with a wiggly string network and relativistic matter and radiation components. The energy-momentum tensor of the string acts as a source for the metric fluctuations, which in turn create the temperature anisotropies. The problem can be recast using Green's (or transfer) functions which project forward the contributions of strings at early times to today. The actual quantitative solution of this problem entails a sophisticated formalism to solve the Boltzmann equation and then to follow photon propagation along the observer's line of sight. At the time of writing, the most recent comparisons [12] between full-sky maps of cosmic string-induced anisotropies and WMAP data yield a cosmological constraint on the models with

$$G\mu < \text{few} \times 10^{-7} \,, \tag{29}$$

with only a weak dependence on the background cosmology – in particular, on the magnitude of the cosmological constant.

Apart from their scale-invariance (which follows from the network's attractor scaling solution discussed in the previous section), cosmic defect-induced fluctuations appear to be the antithesis of the standard inflation paradigm, because they are causal or active (they are generated inside the horizon and over an extended period), there are also large vector and tensor contributions, and they are distinctly non-Gaussian. All these characteristics leave clear signatures in the cosmic microwave background angular power spectrum, chief of which is a much broader primary Doppler peak and little evidence of secondary oscillations. Unlike inflation, active defect sources act incoherently with extra large-scale power from the vectors and tensors. Moreover, their isocurvature nature provides a partial explanation for why the broad primary peak ends up at larger multipoles (typically $\ell \sim 300$ as opposed to $\ell \sim 200$ in a flat cosmology). Isocurvature perturbations correspond to equal and opposite perturbations in the radiation and matter densities – as opposed to equal fractional perturbations in the number densities of the two components for adiabatic perturbations. On the basis of knowledge from present simulations, therefore, cosmic defects alone are extremely unlikely to have been the seeds for large-scale structure formation. However, they cannot be ruled out entirely. For example, admixtures of inflationary power spectra with significant cosmic defect contributions (at a level around 20%, see Fig. 4) do provide a satisfactory fit to present data. This is interesting among other reasons because it is the sort of level at which the non-Gaussian signatures of cosmic strings should still be discernible, although their distinct line-like discontinuities are only clearly identifiable on small angular scales around a few arc minutes.

Accelerated cosmic strings are sources of gravitational radiation [59]. Consequently, a network of long strings and closed loops produces a stochastic gravitational wave background [25] over a wide range of frequencies and with a spectrum which (at least to a first approximation) has equal power on all logarithmic frequency bins. Another distinctive signal would come from the cusps, the points at which the string instantaneously doubles back on itself, approaching the speed of light. Such an event generates an intense pulse of gravitational and other types of radiation, strongly beamed in the direction of motion of the cusp [18]. If massive cosmic strings do indeed exist, both these pulses and the stochastic background are likely to be among the most prominent signals seen by the gravitational-wave detectors now in operation or planned, in particular LIGO and LISA.

A stringent, though indirect, limit on the string energy per unit length comes from observations of the timing of millisecond pulsars. Gravitational waves between us and a pulsar would distort the intervening space-time, and so cause random fluctuations in the pulsar timing. The fact that pulsar timing is extremely regular places an upper limit on the energy density in gravitational waves, and hence on the string scale. The upper limit [29] is of order $G\mu < 10^{-7}$, though there is still considerable uncertainty because this depends on assumptions about the evolution of small-scale structure.

Although gravitational waves are thought to be the main decay byproduct of the evolution of the simplest cosmic string networks, direct decay into particle radiation is extremely efficient and there are claims that it could be the dominant energy-loss mechanism responsible for scaling [62]. In more complicated models, there are certainly other decay channels. If the strings are global (rather than local), then they will preferentially produce Goldstone bosons instead. In axion models, these Goldstone bosons acquire a small mass and become the axions (a prime dark matter candidate). One can estimate the number density of axions using analytic models of cosmic string evolution such as the VOS model. In another class of string models known as superconducting (since they have additional current-like degrees of freedom), then the decay products can include electromagnetic radiation.

Finally, we should mention the claims that cosmic strings could be responsible for a number of high energy astrophysical and cosmological enigmas, including ultra-high energy cosmic rays, gamma ray bursts, and baryo-

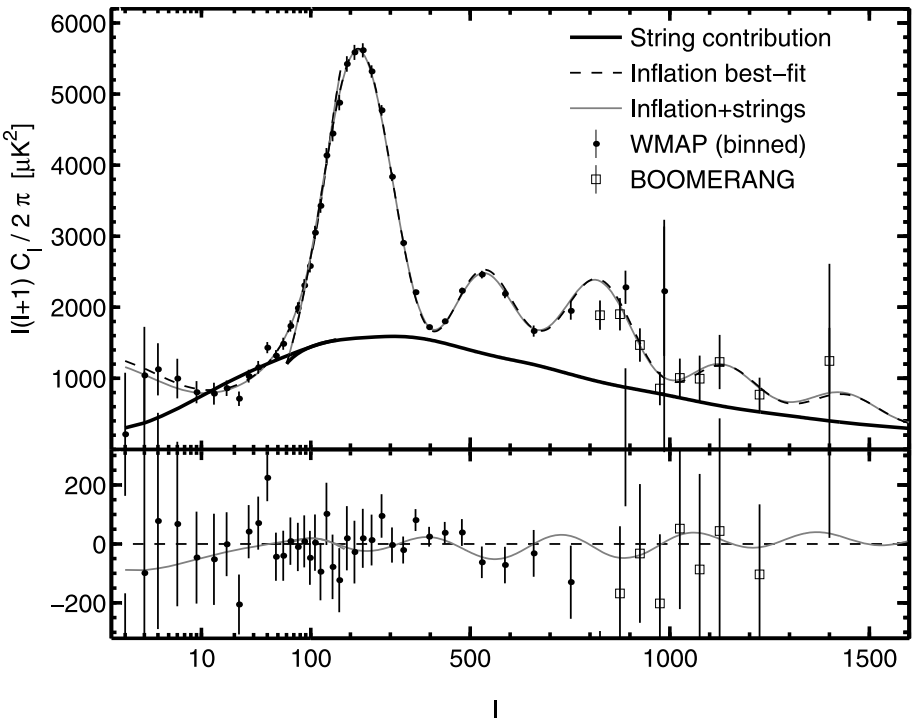

Cosmic Strings, Figure 4

The CMB temperature power spectrum contribution from cosmic strings, normalized to match the WMAP data at $\ell = 10$, as well as the best-fit cases from inflation only (model PL) and inflation plus strings (PL+S). These are compared to the WMAP and BOOMERANG data. The *lower plot* is a repeat but with the best-fit inflation case subtracted, highlighting the deviations between the predictions and the data. Reprinted, with permission, from [13]

genesis (the creation of the matter-antimatter asymmetry of the universe). Since cosmic defects can produce high energy particles, they could contribute to the observed cosmic ray spectrum, notably at ultra-high energies $E \geq 10^{11}$ GeV where the usual acceleration mechanisms seem inadequate [26]. Many ideas have been explored, such as particle emission from cosmic string cusps, but most have been found to produce a particle flux well below current observational limits. However, among the interesting scenarios deserving further study are those with hybrid defects (such as monopoles connected by strings) or vortons.

Field Theory Strings with More Degrees of Freedom

Kibble's original idea was to consider strings in grand unification scenarios, in which the strong and electroweak forces become unified at an energy scale of around 10^{15-16} GeV. Other studies have shown that practically any viable Supersymmetric Grand Unified Theory has a pattern of symmetry breaking transitions that leads to the possibility of cosmic string formation at some point in its history [28]. In particular, in these models the inflationary phase generically ends with a phase transition at which strings are produced. Other studies suggest that this is also a feature of brane inflation models.

So far we have discussed the field theory realization of the simplest model of cosmic string, the Abrikosov–Nielsen–Olesen string, in which the mass per unit length equals the tension and there is no internal structure apart from the magnetic field. In fact, the situation can be much more complex in the early Universe, and realistic particle physics models lead to networks with a much richer phenomenology. The added complexity makes these networks much harder to study, whether by analytic or numerical methods, and consequently they are not as well understood as the simplest case.

We can only give here a brief description of the possible complications. The list below is not complete and furthermore there are strings that fit more than one category. For a more detailed discussion we refer the reader to the reviews by Vilenkin and Shellard, Hindmarsh and Kibble, Carter, and Achúcarro and Vachaspati, where references to the original literature can also be found.

Wiggly Strings; Varying Tension String Networks; Cycloops

The name wiggly strings is sometimes used to refer to any type of string whose mass per unit length is different from its tension. We already mentioned in Sect. "Astrophysical and Cosmological Consequences" that small structure (wiggles) on the string produces a renormalized effective mass per unit length $U > \mu$ and an effective tension $T < \mu$. There are other effects that can affect the mass and tension, for instance the presence of currents along the strings.

A particular kind of small structure is found in extra dimensional models. If spacetime has more than the three spatial dimensions we observe, strings may be able to wrap around the extra dimensions in different ways leading to a renormalized four-dimensional tension and mass per unit length. The effective tension can of course vary along the strings. In extreme cases, the extra dimensional wrapping effects concentrate around certain points along the string which behave like 'beads' (see hybrid networks below) and are called cycloops.

As discussed towards the end of Sect. "String Evolution", the additional degrees of freedom (which can be thought of as a mass current) make the evolution of the networks highly non-trivial. The one-scale assumption is no longer justified: the correlation length, inter-string distance and string curvature radius become distinct lengthscales. Depending on the exact interplay between the bare strings and the mass current wiggles, these lengthscales can evolve differently, and some of them might be scaling while the others are not. The presence of extra dimensions provides a further energy flux mechanism (as energy may be lost into or gained from the extra dimensions) which will affect the string dynamics, but at the time of writing its exact effects have not been studied in detail.

Non-topological/Embedded/Electroweak/Semilocal Strings

In the Abrikosov–Nielsen–Olesen case, the scalar field is zero at the core of the string, and the symmetry is unbroken there. The zero field is protected by the topological properties of the vacuum manifold (the non-contractible circle) and the string is called topological. In more realistic models, the criterion for topological string production is a non-simply connected vacuum manifold, however complicated. These strings are unbreakable and stable.

On the other hand there are examples in which there is no topological protection but the strings are nevertheless stable. The scalar field configuration at the core can be deformed continuously into a ground state, so these non-topological strings can break, their magnetic flux can spread out, or be converted to a different type of flux. But whether this happens is a dynamical question that depends on the detailed masses and couplings of the particles present, on the temperature, etc.

The best studied examples of non-topological strings look like Abrikosov–Nielsen–Olesen strings 'embedded' in a larger model such as the Glashow–Salam–Weinberg model of electroweak interactions [44,57]. These *electroweak* strings carry magnetic flux of the Z boson, but the strings would only be stable for unphysical values of the Z boson mass. They are closely related to *semilocal* strings, another example of embedded strings where the symmetry breaking involves both local and global symmetries intertwined in a particular way. For low scalar mass these can be remarkably stable.

In general, non-topological strings are not resilient enough for the networks to survive cosmological evolution. If the strings are unstable to spreading their magnetic flux, the network will not form. If the strings are breakable the network may form initially but it will quickly disappear (see hybrid networks below for a concrete example). A remarkable exception to this rule are semilocal strings with very low scalar mass: the network forms as a collection of segments which then grow and reconnect to form longer strings or loops. These evolve like a network of Abrikosov–Nielsen–Olesen strings plus a small population of segments and there is some evidence of scaling [5].

Dressed/Superconducting Strings/Vortons

In realistic particle physics models, a stable string will trap in its core any particles or excitations whose mass is lower inside due to the interactions with the scalar field. These *dressed* strings have a more complicated core structure. In extreme cases, the mass of these trapped particles is zero in the core and they lead to persistent currents along the strings, which are then known as *superconducting* [63]. In some cases, the decay of a loop of superconducting string can be stopped by these currents, leading to long-lived remnants called *vortons* [20] that destroy scaling; typical vortons will either dominate the energy density of the universe (contrary to observations) or contribute to the dark matter if they are sufficiently light.

Hybrid Networks

Hybrid networks contain more than one type of defect, such as for instance strings of different kinds or composite defects combining strings, monopoles and/or domain walls.

Composite Defects The production of strings may be accompanied by the production of other defects such as monopoles or domain walls, before or afterwards, that change the behavior of the network as a whole. These networks can have radically different scaling properties – in particular, linear scaling may not exist at all. Consider for instance a sequence of breakings of the form $G \to H \to K$ in which a symmetry group G first breaks down to a subgroup H which subsequently breaks to an even smaller subgroup K at a lower temperature. Two cases are particularly relevant for strings:

- The first breaking produces stable magnetic monopoles, the second confines – totally or partly – the magnetic field to flux tubes (strings) leading to a network of monopoles connected by strings. This can happen either as string segments, with monopoles at the ends, which eventually contract and disappear or as a network of strings carrying heavy "beads" (the monopoles) which can lead to a scaling solution.
- The first breaking produces stable strings, the second makes domain walls attached to the strings (e. g. in *axion* models). The network is made of pancake-like structures that contract under the wall tension and eventually disappear, although in some cases there may be long-lived remnants.

Non-Abelian/(p, q) Strings Another type of hybrid network contains different types of strings whose intercommutation leads to three-point junctions and bridges. These networks are also very different from the simplest ones but the current consensus is that they also seem to reach a scaling solution during cosmological evolution.

In the Non-abelian case, the magnetic flux carried by the string is not just a number but can have different internal "orientations". These become relevant when the strings cross, limiting the ways in which they can reconnect.

Hybrid networks containing several interacting string types are also found in superstring models (see next section). The most interesting type, usually referred to as (p, q) strings, contains two types of string each carrying different type of flux that is separately conserved: fundamental and solitonic or D-strings, roughly corresponding to electric and magnetic flux tubes. The numbers p and q refer to the units of each kind of flux carried by the strings. Since the mass per unit length depends on these fluxes, (p, q) networks are expected to have a hierarchy of different tensions, as well as junctions and bridges. In fact, junctions and bridges will also form in any model in which parallel strings have an attractive interaction, such as Abrikosov–Nielsen–Olesen strings with extremely low scalar to vector mass ratios.

The existence of string junctions and the hierarchy of string tensions make the evolution of these networks considerably more complicated than that of the simple Goto–Nambu strings. Relatively simple analyses suggest that the heavier strings will gradually decay into the lighter ones, and scaling is eventually reached for the strings at the low end of the spectrum (the heavier ones eventually disappear), although this is still under discussion. Naive expectations that the network might be slowed down to non-relativistic speeds and eventually freeze have so far not been supported by the (admittedly simplistic) numerical simulations performed so far. Further work is needed to understand the general conditions under which scaling is (or is not) an attractor.

Cosmic Superstrings

Superstring theory is to date the only candidate model for a consistent quantum theory of gravity that includes all other known interactions. In string theory, the fundamental constituents of nature are not point-like particles but one-dimensional "strings" whose vibrational modes produce all elementary particles and their interactions. Two important features of the theory are supersymmetry (a symmetry between bosonic and fermionic excitations that keeps quantum effects under control) and the presence of extra dimensions above the four spacetime dimensions that we observe.

It is not yet known how to formulate the theory in its full generality but some weak-coupling regimes are well understood. In these, the fundamental strings live in a 10-dimensional spacetime, of which 6 dimensions are "compactified", resulting in an effective 4-dimensional spacetime we live in. There is another regime, M-theory, in which the fundamental objects are two-dimensional "membranes" and the background spacetime is 11-dimensional. These regimes are related to one another by duality transformations that interchange the role of fluctuation quanta and non-perturbative, soliton-like states (branes), so the expectation is that all regimes are different limits of a unique, underlying theory usually referred to as superstring/M-theory, or just M-theory for short.

Before the discovery of D-branes, the "solitons" of superstring theory, the question of whether fundamental superstrings could ever reach cosmological sizes was analyzed and the possibility discarded [64]. There were two main problems. First, the natural mass per unit length of fundamental strings is close to the Planck scale and would correspond to deficit angles of order 2π, which would have been observed. Second, the strings were inherently unstable to either breaking or – depending on the

type of string – becoming the boundary of domain walls that would quickly contract and disappear. The discovery of branes and their role in more exotic compactifications where the six compact dimensions have strong gravitational potentials (and redshifts) has changed this picture. It is now believed that networks of cosmic superstrings could be a natural outcome of brane-antibrane annihilation, especially if the branes are responsible for a period of cosmic inflation [17,30,53].

An important difference with previous scenarios is that these strings are located in regions of the compactified dimensions with very strong gravitational redshift effects ("warping") that reduce the effective mass per unit length of the strings to a level with deficit angles in the region of 10^{-12} to 10^{-7}, compatible with current observations. Another important difference is a much lower probability that the strings intercommute when they cross, estimated to be 10^{-3} to 10^{-1}, depending on the type of strings. The lower intercommutation rates lead to much denser networks. Estimates of the corresponding enhancement in the emission of gravitational radiation by cusps puts these strings in a potentially observable window by future gravitational wave detectors [19,54].

The networks are hybrid, consisting of fundamental strings and D-strings, the latter being either one-dimensional D-branes or perhaps the result of a higher dimensional D-brane where all but one dimension are wrapped around some "holes" (cycles) in the compactified space. There may also be cycloops.

As in the case of hybrid field theory strings, whether or not superstring networks eventually reach a scaling regime is an open question. Analytic studies and numerical simulations of simplified cases suggest that scaling is certainly possible, though contingent on model parameters that at the time of writing are not well understood. In this case, in addition to the presence of junctions and a non-trivial spectrum of string tensions, a third factor can affect the evolution of these networks. If the strings are actually higher-dimensional branes partially wrapped around some extra dimensions, then energy and momentum can in principle leak into or out of these extra dimensions [4]. Since the effective damping force affecting the ordinary and extra dimensions is different, one might generically expect that this will be the case. Depending on its sign and magnitude, such an energy flow can in principle prevent scaling, either by freezing the network (if too much energy leaks out) or by making the strings dominate the universe's energy density (if too much energy leaks in, though this is less likely than the opposite case). In this sense, a somewhat delicate balance may be needed to ensure scaling. At a phenomenological level, further work will be required in order to understand the precise conditions under which each of these scenarios occurs. At a more fundamental level, it is quite likely that which of the scenarios is realized will depend on the underlying compactifications and/or brane inflation models, and that may eventually be used as a discriminating test between string theory realizations.

Future Directions

One of the most exciting prospects is the discovery of magnetic-type CMB polarization (usually referred to as B-modes) as this would reveal the presence of vector and/or tensor modes. Cosmic string models may be further constrained in the near future because B-modes are predicted to have amplitudes comparable to the electric-type E-modes (at large angular scales). At high resolution, one could also hope to observe defects directly through the B-mode signal, against a relatively unperturbed background. Conversely, the detection of vector modes would provide strong evidence against inflation without cosmic defects. Polarization data will also strongly constrain a significant isocurvature contribution to the mainly adiabatic density fluctuations. Isocurvature perturbations can be a signature of more complicated physics during inflation, such as the effects of two or more scalar fields, or the formation of defects at the end of inflation.

Ongoing and future CMB experiments, especially at high resolution, will be probing the degree of Gaussianity of the primordial fluctuations. The detection of significant and unambiguous non-Gaussianity in the primary CMB signal would be inconsistent with simple (so called single field slow-roll) inflation. More general inflationary models can accommodate certain types of non-Gaussianity, and one can also envisage non-Gaussianity from excited initial states for inflation. It is interesting to note that given the existing bounds on $G\mu$, current CMB experiments do not have the sensitivity or resolution to detect cosmic string signatures directly, in particular the Kaiser–Stebbins effect in CMB maps. However, with high-resolution sensitivities becoming available in the near future, direct constraints (or detections) will be possible. This is just one example of the interesting new science that future high-resolution CMB experiments might uncover in the years ahead. In particular, ESA'a Planck Surveyor [47], scheduled for launch in late 2009, may be able to provide significant breakthroughs.

A deeper understanding of the evolution and consequences of string networks will require progress on both numerical simulations and analytic modelings. At the time of writing there is still no numerical code that includes all the relevant physics, even for the simplest (Goto–Nambu)

strings. Inclusion of gravitational backreaction is particularly subtle, and may require completely new approaches. The expected improvements in the available hardware and software will allow for simulations with much longer evolution timespan and spatial resolution, which are needed in order to understand the non-linear interactions between large and small scales all the way down to the level of the constituent quantum fields. This in turn will be a valuable input for more detailed analytic modeling, that must accurately describe the non-trivial small-scale properties of the string networks as well as the detailed features of the loop populations. Better modeling is also needed to describe more general networks – three crucial mechanisms for which at present there is only a fairly simplistic description are the presence of junctions, a non-trivial spectrum of string tensions, and the flow of energy-momentum into extra dimensions.

At a more fundamental level, a better understanding of the energy loss mechanisms and their roles in the evolution of the networks is still missing [10] and it will require new developments in the theory of quantum fields out of equilibrium. Such theoretical developments are also needed to understand defect formation in systems with gauge fields, and could be tested experimentally in superconductors.

The early universe is a unique laboratory, where the fundamental building blocks of nature can be probed under the most extreme conditions, that would otherwise be beyond the reach of any human-made laboratory. Cosmic strings are particularly interesting for this endeavor: they are effectively living fossils of earlier cosmological phases, where physical conditions may have been completely different. The serendipitous discovery of cosmic defects or other exotic phenomena in forthcoming cosmological surveys would have profound implications for our understanding of cosmological evolution and of the physical processes that drove it. The search continues while, in the meantime, the absence of cosmic string signatures will remain a powerful theoretical tool to discriminate between fundamental theories. The possibility that something as fundamental as superstring theory may one day be validated in the sky, using tools as mundane as spectroscopy or photometry, is an opportunity than neither astrophysicists nor particle physicists can afford to miss.

Bibliography

Primary Literature

1. Abrikosov AA (1957) On the magnetic properties of superconductors of the second group. J E T P 32:1442–1452
2. Albrecht A, Turok N (1989) Evolution of cosmic string networks. Phys Rev D40:973
3. Allen B, Shellard EPS (1990) Cosmic string evolution: a numerical simulation. Phys Rev Lett 64:119–122
4. Avgoustidis A, Shellard EPS (2005) Cosmic string evolution in higher dimensions. Phys Rev D71:123513
5. Achucarro A, Salmi P, Urrestilla J (2007) Semilocal cosmic string networks. Phys Rev D75:121703
6. Austin D, Copeland EJ, Kibble TWB (1993) Evolution of cosmic string configurations. Phys Rev D48:5594–5627
7. Bogomol'nyi EB (1976) The stability of classical solutions. Yad Fiz 24:861–870
8. Bennett DP (1986) The Evolution Of Cosmic Strings, Phys Rev D33:873 (Erratum-ibid.D34:1235, 1986, Erratum-ibid.D34:3932, 1986)
9. Bennett DP, Bouchet FR (1990) High resolution simulations of cosmic string evolution. Phys Rev D41:2408
10. Borsanyi S, Hindmarsh M (2007) Semiclassical decay of topological defects. e-Print arXiv:0712.0300
11. Bauerle C, Bunkov YM, Fisher SN, Godfrin H, Pickett GR (1996) Laboratory simulation of cosmic string formation in the early universe using superfluid He-3. Nature 382:332–334
12. Bevis N, Hindmarsh M, Kunz M, Urrestilla J (2007) CMB power spectrum contribution from cosmic strings using field-evolution simulations of the Abelian Higgs model. Phys Rev D75:065015
13. Bevis N, Hindmarsh M, Kunz M, Urrestilla J (2008) Fitting CMB data with cosmic strings and inflation. Phys Rev Lett 100:021301
14. Bradley DI, Fisher SN, Guénault AM, Haley RP, Kopu J, Martin H, Pickett GR, Roberts JE, Tsepelin V (2008) Relic topological defects from brane annihilation simulated in superfluid ^{3}He. Nat Phys 4:46
15. Bowick MJ, Chandar L, Schiff EA, Srivastava AM (1994) The cosmological Kibble mechanism in the laboratory – string formation in liquid crystals. Science 263:943–945
16. Chuang I, Durrer R, Turok N, Yurke B (1991) Cosmology in the laboratory – defect dynamics in liquid crystals. Science 251:1336–1367
17. Copeland EJ, Myers RC, Polchinski J (2004) Cosmic F- and D-strings. JHEP 0406:013
18. Damour T, Vilenkin A (2000) Gravitational wave bursts from cosmic strings. Phys Rev Lett 85:3761–3764
19. Damour T, Vilenkin A (2005) Gravitational radiation from cosmic (super)strings: bursts, stochastic background, and observational windows. Phys Rev D71:063510
20. Davis RL, Shellard EPS (1989) Cosmic vortons. Nucl Phys B323:209
21. Dodd ME, Hendry PC, Lawson NS, McClintock PVE, Williams CDH (2004) Non-appearance of vortices in fast mechanical expansions of liquid He-4 through the lambda transition. Phys Rev Lett 81:3703–3706
22. Goto T (1971) Relativistic quantum mechanics of one-dimensional mechanical continuum and subsidiary condition of dual resonance model. Prog Theor Phys 46:1560–1569
23. Gott JR (1985) Gravitational lensing effects of vacuum string: exact results. Ap J 288:422
24. Hindmarsh M, Rajantie A (2000) Defect formation and local gauge invariance. Phys Rev Lett 85:4660
25. Hogan CJ, Rees M (1984) Gravitational interactions of cosmic strings. Nature 311:109

26. Hill CT, Schramm DN, Walker TP (1987) Ultrahigh-energy cosmic rays from superconducting strings. Phys Rev D36:1007
27. Jeannerot R (1997) Inflation in supersymmetric unified theories. Phys Rev D56:6205–6216
28. Jeannerot R, Rocher J, Sakellariadou M (2003) How generic is cosmic string formation in SUSY GUTs. Phys Rev D68: 103514
29. Jenet FA, Hobbs GB, van Straten W, Manchester RN, Bailes m, Verbiest JPW, Edwards RT, Hotan AW, Sarkissian JM, Ord SM (2006) Upper bounds on the low-frequency stochastic gravitational wave background from pulsar timing observations: current limits and future prospects. Astrophys J 653:1571–1576
30. Jones NT, Stoica H, Tye S-H (2003) The production, spectrum and evolution of cosmic strings in brane inflation. Phys Lett B563:6–14
31. Kibble TWB (1976) Topology of cosmic domains and strings. J Phys A9:1387
32. Kibble TWB (1985) Evolution Of A System Of Cosmic Strings, Nucl Phys B 252:227 (Erratum-ibid. B 261:750,1985)
33. Kibble TWB (2007) Phase-transition dynamics in the lab and the universe. Phys Today 60:57
34. Kaiser N, Stebbins A (1984) Microwave anisotropy due to cosmic strings. Nature 310:391
35. Kuijken K, Siemens X, Vachaspati T (2008) Microlensing by cosmic strings. arXiv:0707.2971. preprint, MNRAS 384:161
36. Karra G, Rivers RJ (1998) A reexamination of quenches in ^{4}He. Phys Rev Lett 98:3707
37. Majumdar M, Davis A-C (2002) Cosmological creation of D-branes and anti-D-branes. JHEP 0203:056
38. Martins CJAP, Shellard EPS (2002) Extending the velocity-dependent one-scale string evolution model. Phys Rev D65:043514
39. Martins CJAP, Shellard EPS (2006) Fractal properties and small-scale structure of cosmic string networks. Phys Rev D73:043515
40. Maniv A, Polturak E, Koren G (2003) Phys Rev Lett 91:197001
41. Monaco R, Aaroe M, Mygind J, Rivers RJ, Koshelets VP (2006) Spontaneous fluxon production in annular josephson tunnel junctions in the presence of a magnetic field. Phys Rev B74:144513
42. Moriarty K, Myers E, Rebbi C (1988) Dynamical interactions of cosmic strings and flux vortices in superconductors. Phys Lett B207:411
43. Nambu Y (1970) Lectures at the Copenhagen Summer School. (in press)
44. Nambu Y (1977) String-like configurations in the Weinberg–Salam theory. Nucl Phys B130:505
45. Nielsen HB, Olesen P (1973) Vortex line models for dual strings. Nucl Phys B61:45–61
46. Olum K, Vanchurin V (2007) Cosmic string loops in the expanding universe. Phys Rev D75:063521
47. ESA Planck. http://www.rssd.esa.int/Planck/
48. Polchinski J, Rocha JV (2006) Analytic study of small-scale structure on cosmic strings. Phys Rev D74:083504
49. Ringeval C, Sakellariadou M, Bouchet F (2007) Cosmological evolution of cosmic string loops. JCAP 0702:023
50. Ruutu VMH, Eltsov VB, Gill AJ, Kibble TWB, Krusius M, Makhlin YG, Placais B, Volovik GE, Xu W (1996) Vortex formation in neutron irradiated superfluid He-3 as an analogue of cosmological defect formation. Nature 382:334–336
51. Shellard EPS (1987) Cosmic string interactions. Nucl Phys B283:624
52. Silk J, Vilenkin A (1984) Cosmic strings and galaxy formation. Phys Rev Lett 53:1700
53. Sarangi S, Tye S-H (2002) Cosmic string production towards the end of brane inflation. Phys Lett B536:185–192
54. Siemens X, Mandic V, Creighton J (2007) Gravitational wave stochastic background from cosmic (super)strings. Phys Rev Lett 98:111101
55. Spergel DN, Bean R, Dore O, Nolta MR, Bennet CL, Dunkley J, Hinshaw G, Jarosik N, Komatsu E, Page L, Peiris HV, Verde L, Halpern M, Hill RS, Kogut A, Limon M, Meyer SS, Odegard N, Tucker GS, Wieland JL, Wollack E, Wright EL (2007) Wilkinson microwave anisotropy probe (WMAP) three year results: implications for cosmology. Astrophys J Suppl 170:377
56. Taubes CH (1981) The existence of multi-monopole solutions to the non-abelian, Yang–Mills–Higgs equations for arbitrary simple gauge groups. Comm Math Phys 80:343
57. Vachaspati T (1992) Vortex solutions in the Weinberg–Salam model. Phys Rev Lett 68:1977–1980 (Erratum-ibid. B 26:750,1985)
58. Vilenkin A (1981) Gravitational field of vacuum domain walls and strings. Phys Rev D23:852
59. Vilenkin A (1981) Gravitational radiation from cosmic strings. Phys Lett B107:47
60. Vilenkin A (1984) Cosmic strings as gravitational lenses. Ap J L51:282
61. Vachaspati T, Vilenkin A (1984) Formation and evolution of cosmic strings. Phys Rev D30:2036
62. Vincent G, Antunes ND, Hindmarsh M (1998) Numerical simulations of string networks in the Abelian–Higgs model. Phys Rev Lett 80:2277
63. Witten E (1985) Superconducting strings. Nucl Phys B249:557–592
64. Witten E (1985) Cosmic superstrings. Phys Lett B153:243–246
65. Zurek WH (1985) Cosmological experiments in superfluid helium. Nature 317:505

Books and Reviews

Achúcarro A, Vachaspati T (2000) Semilocal and electroweak strings. Phys Rept 327:347–436

Carter B (1997) Brane dynamics for treatment of cosmic strings and vortons. arXiv:hep-th/9705172

Davis A-C, Kibble TWB (2005) Fundamental cosmic strings. Contemp Phys 46:313–322

Durrer R, Kunz M, Melchiorri A (2002) Cosmic structure formation with topological defects. Phys Rept 364:1–81

Hindmarsh M, Kibble TWB (1995) Cosmic strings. Rept Prog Phys 58:477–562

Kibble TWB (1980) Some implications of a cosmological phase transition. Phys Rept 67:183

Polchinski J (2004) Introduction to cosmic F- and D-strings. In: Lectures given at NATO advanced study institute and EC summer school on string theory: from gauge interactions to cosmology. Cargese, 7–19 Jun 2004. Cargese, pp 229–253

Vilenkin A, Shellard EPS (2000) Cosmic strings and other topological defects. Cambridge University Press, Cambridge

Zurek WH (1996) Cosmological experiments in condensed matter systems. Phys Rep 276:177–221

Cost Sharing

MAURICE KOSTER
University of Amsterdam,
Amsterdam, Netherlands

Article Outline

Glossary
Definition of the Subject
Introduction
Cooperative Cost Games
Non-cooperative Cost Games
Continuous Cost Sharing Models
Future Directions
Bibliography

Glossary

Core The *core* of a cooperative cost game $\langle N, c\rangle$ is the set of all coalitionally stable vectors of cost shares.

Cost function A *cost function* relates to each level of output of a given production technology the minimal necessary units of input to generate it. It is non-decreasing function $c\colon X \to \mathbb{R}_+$, where X is the (ordered) space of outputs.

Cost sharing problem A cost sharing problem is an ordered pair (q, c), where $q \in \mathbb{R}^N_+$ is a profile of individual demands of a fixed and finite group of agents $N = \{1, 2, \ldots, n\}$ and c is a cost function.

Game theory The branch of applied mathematics and economics that studies situations where players make decisions in an attempt to maximize their returns. The essential feature is that it provides a formal modeling approach to social situations in which decision makers interact.

Cost sharing rule A *cost sharing rule* is a mapping that assigns to each cost sharing problem under consideration a vector of non-negative cost shares.

Demand game Strategic game where agents place demands for output strategically.

Demand revelation game Strategic game where agents announce their maximal contribution strategically.

Strategic game An ordered triple $G = \langle N, (A_i)_{i\in N}, (\precsim_i)_{i\in N}\rangle$, where

- $N = \{1, 2, \ldots, n\}$ is the set of *players*,
- A_i is the set of available *actions* for player i,
- $\precsim_i$ is a *preference relation* over the set of possible consequences C of action.

Definition of the Subject

Throughout we will use a fixed set of agents $N = \{1, 2, \ldots, n\}$, where n is a given natural number. For subsets S, T of N, we write $S \subset T$ if each element of S is contained in T; $T\backslash S$ denotes the set of agents in T except those in S. The *power set* of N is the set of all subsets of N; each coalition $S \subset N$ will be identified with the element $\mathbf{1}_S \in \{0, 1\}^N$, the vector with ith coordinate equal to 1 precisely when $i \in S$. Fix a vector $x \in \mathbb{R}^N$ and $S \subset N$. The projection of x on $\mathbb{R}^S$ is denoted by x_S, and $x_{N\backslash S}$ is sometimes more conveniently denoted by x_{-S}. For any $y \in \mathbb{R}^S$, (x_{-S}, y) stands for the vector $z \in \mathbb{R}^N$ such that $z_i = x_i$ if $i \in N\backslash S$ and $z_i = y_i$ if $i \in S$. We denote $x(S) = \sum_{i\in S} x_i$. The vector in $\mathbb{R}^S$ with all coordinates equal to zero is denoted by $\mathbf{0}_S$. Other notation will be introduced when necessary.

This article focuses on different approaches in the literature through a discussion of a couple of basic and illustrative models, each involving a single facility for the production of a finite set M of outputs, commonly shared by a fixed set $N := \{1, 2, \ldots, n\}$ of agents. The feasible set of outputs for the technology is identified with a set $X \subset \mathbb{R}^M_+$. It is assumed that the users of the technology may freely dispose over any desired quantity or level of the outputs; each agent i has some demand $x_i \in X$ for output. Each profile of demands $x \in X^N$ is associated to its *cost* $c(x)$, i. e. the minimal amount of the idiosyncratic input commodity needed to fulfill the individual demands. This defines the *cost function* $c\colon X^N \to \mathbb{R}_+$ for the technology, comprising all the production externalities. A *cost sharing problem* is an ordered pair (x, c) of a demand profile x and a cost function c. The interpretation is that x is produced and the resulting cost $c(x)$ has to be shared by the collective N. Numerous practical applications fit this general description of a cost sharing problem.

In mathematical terms a cost sharing problem is equivalent to a *production sharing* problem where output is shared based on the profile of inputs. However, although many concepts are just as meaningful as they are in the cost sharing context, results are not at all easily established using this mathematical duality. In this sense consider [68] as a warning to the reader, showing that the strategic analysis of cost sharing solutions is quite different from surplus sharing solutions. This monograph will center on cost sharing problems. For further reference on production sharing see [51,67,68,91].

Introduction

In many practical situations managers or policy-makers deal with private or public enterprises with multiple users.

A production technology facilitates its users, causing externalities that have to be shared. Applications are numerous, ranging from environmental issues like pollution, fishing grounds, to sharing multipurpose reservoirs, road systems, communication networks, and the Internet. The essence in all these examples is that a manager cannot directly influence the behavior of the users, but only indirectly by addressing the externalities through some decentralization device. By choosing the right instrument the manager may help to shape and control the nature of the resulting individual and aggregate behavior. This is what is usually understood as the mechanism design or implementation paradigm. The state-of the-art literature shows for a couple of simple but illustrative cost sharing models that one cannot push these principles too far, as there is often a trade-off between the degree of distributive justice and economic efficiency. Then this is what makes choosing 'the' right solution an ambiguous task, certainly without a profound understanding of the basic allocation principles. Now first some examples will be discussed.

Example 1 The water-resource management problem of the Tennessee Valley Authority (TVA) in the 1930s is a classic in the cost-sharing literature. It concerns the construction of a dam in a river to create a reservoir, which can be used for different purposes like flood control, hydro-electric power, irrigation, and municipal supply. Each combination of purposes requires a certain dam height and accompanying construction costs have to be shared by the purposes. Typical for the type of problem is that up to a certain critical height there are economies of scale as marginal costs of extra height are decreasing. Afterwards, marginal costs increase due to technological constraints. The problem here is to allocate the construction costs of a specific dam among the relevant purposes.

Example 2 Another illustrative cost sharing problem dating back from the early days in the cost sharing literature [69,70] deals with landing fee schedules at airports, so-called *airport problems*. These were often established to cover the costs of building and maintaining the runways. The cost of a runway essentially depends on the size of the largest type of airplane that has to be accommodated – a long runway can be used by smaller types as well. Suppose there are m types of airplanes and that c_i is the cost of constructing a landing strip suitable for type i. Moreover, index the types from small to large so that $0 = c_0 < c_1 < c_2 < \cdots < c_m$. In the above terminology the technology can be described by $X = \{0, 1, 2, \ldots, m\}$, and the cost function $c\colon X^N \to \mathbb{R}_+$ is defined by $c(x) = c_k$ where $k = \max\{x_i \mid i \in N\}$ is the maximal service level required in x. Suppose that in a given year, N_k is the set landings of type k airplanes, then the set of users of the runway is $N = \cup_k N_k$. The problem is now to apportion the full cost $c(x)$ of the runway to the users in N, where x is the demand vector given by $x_i = \ell$ if $i \in N_\ell$.

Airport problems describe a wide range of cost sharing problems, ranging from sharing the maintenance cost of a ditch system for irrigation projects [1], to sharing the dredging costs in harbors [14].

Example 3 A joint project involves a number of activities for which the estimated durations and precedence relations are known. Delay in each of these components affects the period in which the project can be realized. Then a cost sharing problem arises when the joint costs due to the accumulated delay are shared among the individuals causing the delays. See [21].

Example 4 In many applications the production technology is given by a network $G = (V, E)$ with nodes V and set of costly edges $E \subset V \times V$, and cost function $c\colon E \to \mathbb{R}_+$. The demands of the agents are now parts of the infrastructure, i. e. subsets of E. Examples include the sharing the cost of infrastructure for supply of energy and water, or transport systems.

For example, the above airport problem can be modeled as such with

$$V = \{1, 2, \ldots, m\} \cup \{\odot\} \ ,$$
$$E = \{(\odot, 1), (1, 2), \ldots, (m-1, m)\} \ .$$

Graphically, the situation is depicted by the line graph in Fig. 1.

Imagine that the runway starts at the special node $\odot$, and that the edges depict the different pieces of runway served to the players. An airplane of type k is situated at node k, and needs all edges towards $\odot$. The edge left to the kth node is called $e_k = (k-1, k)$, and the corresponding cost is $c(e_k) = c_k - c_{k-1}$. The demand of an airplane at node k is now described by the edges on the path from node k to $\odot$.

Example 5 In more general network design problems, a link facilitates a *flow*; for instance, in telecommunication it is data flowing through the network, in road systems it is traffic. [49] discusses a model where a network planner allocates the fixed cost of a network based on the individual

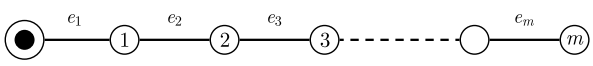

Cost Sharing, Figure 1
Graphical representation of an airport problem

demands being flows. [77,124] discuss *congested* telecommunication networks, where the cost of a link depends on the size of the corresponding flow. Then these positive network externalities lead to a concentration of flow, and thus to *hub-like* networks. Economies of scale require cooperation of the users, and the problem now is to share the cost of these so-called *hub-like* networks.

Example 6 As an insurance against the uncertainty of the future net worths of its constituents, firms are often regulated to hold an amount of riskless investments, i. e. its *risk capital*. Given that returns of normal investments are higher, the difference with the riskless investments is considered as a cost. The sum of the risk capitals of each constituent is usually larger than the risk capital of the firm as a whole, and the allocation problem is to apportion this diversification effect observed in risk measurements of financial portfolios. See Denault [28].

Solving Cost Sharing Problems: Cost Sharing Rules

A *vector of cost shares* for the cost sharing problem (x, c) is an element $y \in \mathbb{R}^N$ with the property that $\sum_{i \in N} y_i = c(x)$. This equality is also called *budget-balancing* condition. Central issue addressed in the cost sharing literature is how to determine the appropriate y. The vast majority of the cost sharing literature is devoted to a mechanistic way of sharing joint costs; given a class of cost sharing problems $\mathcal{P}$, a (simple) formula computes the vector of cost shares for each of its elements. This yields a *cost sharing rule* $\mu : \mathcal{P} \to \mathbb{R}^N$ where $\mu(P)$ is the vector of cost shares for each $P \in \mathcal{P}$.

At this point it should be clear to the reader that many formula's will see to a split of joint cost and heading for *the* solution to cost sharing problems is therefore an ambiguous task. The least we want from a solution is that it is consistent with some basic principles of fairness or justice and, moreover, that it creates the right incentives. Clearly, the desirability of solution varies with the context in which it is used, and so will the sense of appropriateness. Moreover, the different parties involved in the decision making process will typically hold different opinions; accountants, economists, production managers, regulators and others all are looking at the same institutional entity from different perspectives. The existing cost sharing literature is about exploring boundaries of what can be thought of desirable features of cost sharing rules. More important than the rules themselves, are the properties that each of them is consistent with. Instead of building a theory on single instances of cost sharing problems, the cost sharing literature discusses structural invariance properties over *classes* of problems. Here the main distinction is made on the basis of the topological properties of the technology, whether the cost sharing problem allows for a *discrete* or *continuous* formulation. For each type of models, divisible or indivisible goods, the state-of-the-art cost sharing literature has developed into two main directions, based on the way individual preferences over combinations of cost shares and (levels of) service are treated. On the one hand, there is a stream of research in which individual preferences are not explicitly modeled and demands are considered *inelastic*. Roughly, it accommodates the large and vast growing *axiomatic* literature (see e. g. [88,129]) and the theory on cooperative cost games [105,130,136,145]. Secondly, there is the literature on cost sharing models where individual preferences are explicitly modeled and demands are elastic. The focus is on non-cooperative *demand games* in which the agents are assumed to choose their demands strategically, see e. g. [56,91,141].

As an interested reader will soon find out, in the literature there is no shortage of plausible cost sharing techniques. Instead of presenting a kind of summary, this article focuses on the most basic and most interesting ones, and in particular their properties with respect to strategic interplay of the agents.

Outline

The article is organized as follows. Section "Cooperative Cost Games" discusses cost sharing problems from the perspective of cooperative game theory. Basic concepts like core, Shapley value, nucleolus and egalitarian solution are treated. Section "Non-cooperative Cost Games" introduces the basic concepts of non-cooperative game theory including dominance relations, preferences, and Nash-equilibrium. Demand games and demand revelation games are introduced for discrete technologies with concave cost function. This part is concluded with two theorems, the strategic characterization of the Shapley value and constrained egalitarian solution as cost sharing solution, respectively. Section "Continuous Cost Sharing Models" introduces the continuous production model and it consists of two parts. First the simple case of a production technology with homogeneous and perfectly divisible private goods is treated. Prevailing cost sharing rules like proportional, serial, and Shapley–Shubik are shortly introduced. We then give a well-known characterization of additive cost sharing rules in terms of corresponding rationing methods, discuss the related cooperative and strategic games. The second part is devoted to the heterogeneous output model and famous solutions like Aumann–Shapley, Shapley–Shubik, and serial rules. We fi-

nalize with Sect. "Future Directions" where some future directions of research are spelled out.

Cooperative Cost Games

A discussion of cost sharing solutions and incentives needs a proper framework wherein the incentives are formalized. In the seminal work of von Neumann and Morgenstern [140] the notion of a *cooperative game* was introduced as to model the interaction between actors/players who coordinate their strategies in order to maximize joint profits. Shubik [122] was one of the first to apply this theory in the cost sharing context.

Cooperative Cost Game

A cooperative cost game among players in N is a function $c\colon 2^N \to \mathbb{R}$ with the property that $c(\varnothing) = 0$; for non-empty sets $S \subset N$ the value $c(S)$ is interpreted as the cost that would arise should the individuals in S work together and serve only their own purposes. The class of all cooperative cost games for N will be denoted by CG.

Any general class $\mathcal{P}$ of cost sharing problems can be embedded in CG as follows. For the cost sharing problem $(x, c) \in \mathcal{P}$ among agents in N define the *stand-alone* cost game $c_x \in CG$ by

$$c_x(S) := \begin{cases} c(x_S, 0_{N\setminus S}) & \text{if } S \subset N, S \neq \varnothing \\ 0 & \text{if } S = \varnothing\,. \end{cases} \tag{1}$$

So $c_x(S)$ can be interpreted as the cost of serving only the agents in S.

Example 7 The following numerical example will be frequently referred to. An airport is visited by three airplanes in the set $N = \{1, 2, 3\}$, which can be accommodated at cost $c_1 = 12$, $c_2 = 20$, and $c_3 = 33$, respectively. The situation is depicted in Fig. 2.

The corresponding cost game c is determined by associating each coalition S of airplanes to the minimum cost of the runway needed to accommodate each of its members. Then the corresponding cost game c is given by the table below. Slightly abusing notation we denote $c(i)$ to indicate $c(\{i\})$, $c(ij)$ for $c(\{i, j\})$ and so forth.

S	$\varnothing$	1	2	3	12	13	23	123
$c(S)$	0	12	20	33	20	33	33	33

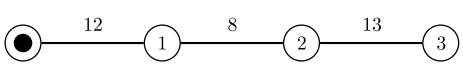

Cost Sharing, Figure 2
Airport game

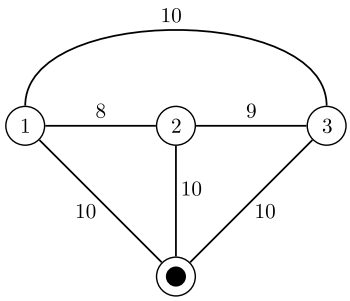

Cost Sharing, Figure 3
Minimum cost spanning tree problem

Note that, since we identified coalitions of players in N with elements in 2^N, we may write c to denote the cooperative cost game. By the binary nature of the demands the cost function for the technology formally *is* a cooperative cost game. For $x = (1, 0, 1)$ the corresponding cost game c_x is specified by

S	$\varnothing$	1	2	3	12	13	23	123
$c_x(S)$	0	12	0	33	12	33	33	33

Player 2 is a *dummy player* in this game, for all $S \subset N\setminus\{2\}$ it holds $c_x(S) = c_x(S \cup \{2\})$.

Example 8 Consider the situation as depicted in the Fig. 3 below, where three players, each situated at a different node, want to be connected to the special node $\odot$ using the indicated costly links. In order to connect themselves to $\odot$ a coalition S may use only links with $\odot$ and the direct links between its members, and then only if the costs are paid for. For instance, the minimum cost of connecting player 1 in the left node to $\odot$ is 10, and the cost of connecting players 1 and 2 to $\odot$ are 18 – the cost of the direct link from 2 and the indirect link between 1 and 2. Then the associated cost game is given by

S	$\varnothing$	1	2	3	12	13	23	123
$c(S)$	0	10	10	10	18	20	19	27

Notice that in this case the network technology exhibits positive externalities. The more players want to be connected, the lower the per capita cost.

For those applications where the cost $c(S)$ can be determined irrespective of the actions taken by its complement $N\setminus S$ the interpretation of c implies *sub-additivity*, i. e. the property that for all $S, T \subset N$ with $S \cap T = \varnothing$ implies $c(S \cup T) \leq c(S) + c(T)$. This is for instance an essential feature of the technology underlying natural monopolies (see, e. g., [13,120]). Note that the cost games in Example 7 and 8 are sub-additive. This is a general property for airport games as well as minimum cost spanning tree games.

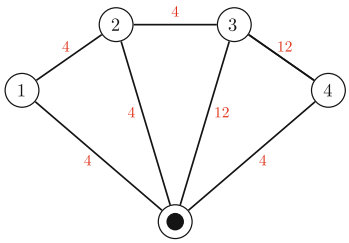

Cost Sharing, Figure 4
Non-concave MCST game

Sometimes the benefits of cooperation are even stronger. A game is called *concave* (or *sub-modular*) if for all $S, T \subset N$ we have

$$(S \cup T) + c(S \cap T) \leq c(S) + c(T) \,. \tag{2}$$

At first this seems a very abstract property, but one may show that it is equivalent with the following that

$$c(S \cup \{i\}) - c(S) \geq c(T \cup \{i\}) - c(T) \tag{3}$$

for all coalitions $S \subset T \subset N \setminus \{i\}$. This means that the marginal cost of a player i with respect to larger coalitions is non-increasing, i. e. the technology exhibits positive externalities. Concave games are also frequently found in the network literature, see [63,75,93,121].

Example 9 Although sub-additive, minimum cost spanning tree games are not always concave. Consider the following example due to [17]. The numbers next to the edges indicate the corresponding cost. We assume a complete graph and that the invisible edges cost 4. Note that in this game every three-player coalition is connected at cost 12, whereas $c(34) = 16$. Then $c(1234) - c(234) = 16 - 12 = 4$ whereas $c(134) - c(34) = -4$. So the marginal cost of player 1 is not decreasing with respect to larger coalitions.

Incentives in Cooperative Cost Games

The objective in cooperative games is to share the profits or costs savings of cooperation. Similar to the general framework, a *vector of cost shares* for a cost game $c \in CG$ is a vector $x \in \mathbb{R}^N$ such that $x(N) = c(N)$. The question is what cost share vectors make sense if (coalitions of) players have the possibility to opt out thereby destroying cooperation on a larger scale.

In order to ensure that individual players join, a proposed allocation x should at least be *individual rational* so that $x_i \leq c(i)$ for all $i \in N$. In that case no player has a justified claim to reject x as proposal, since going alone yields a higher cost. The set of all such elements is called the *imputation set*. If, in a similar fashion, $x(S) \leq c(S)$ for all $S \subset N$ then x is called *stable*; under proposal x no coalition S has a strong incentive to go alone, as it is not possible to redistribute the cost shares afterwards and make every defector better of. The *core* of a cost game c, notation core(c), consists of all stable vectors of cost shares for c. If cooperation on a voluntary basis by the grand coalition N is conceived as a desirable feature then the core and certainly the imputation set impose reasonable conditions for reaching it. Nevertheless, the core of a game can be empty.

Call a collection $\mathcal{B}$ of coalitions *balanced*, if there is a vector of positive weights $(\lambda_S)_{S \in \mathcal{B}}$ such that for all $i \in N$

$$\sum_{S \in \mathcal{B}, S \ni i} \lambda_S = 1 \,.$$

A cost game c is *balanced* if it holds for each balanced collection $\mathcal{B}$ of coalition that

$$\sum_{S \in \mathcal{B}} \lambda_S c(S) \geq c(N) \,.$$

It is the celebrated theorem below which characterizes all games with non-empty cores.

Theorem 1 (Bondareva–Shapley [20,117]) *The cost game c is balanced if and only if the core of c is non-empty.*

Concave cost games are balanced, see [119]. Concavity is not a necessary condition for non-emptiness of the core, since minimum cost spanning tree games are balanced as well.

Example 10 Consider the two-player game c defined by $c(12) = 10, c(1) = 3, c(2) = 8$. Then core($c$) $= \{(x, 10 - x) | 2 \leq x \leq 3\}$. Note that, opposed to the general case, for two-player games sub-additivity is equivalent with non-emptiness of the core.

Cooperative Solutions

A solution on a subclass A of CG is a mapping $\mu : A \to \mathbb{R}^N$ that assigns to each $c \in A$ a vector of cost shares $\mu(c)$; $\mu_i(c)$ stands for the charge to player i.

The Separable Cost Remaining Benefit Solution Common practice among civil engineers to allocate costs of multipurpose reservoirs is the following solution. The *separable cost* for each player (read purpose) $i \in N$ is given by $s_i = c(N) - c(N \setminus \{i\})$, and the *remaining benefit* by $r_i = c(i) - s_i$. The separable cost remaining benefit solution charges each player i for the separable cost s_i and the non-separable costs $c(N) - \sum_{j \in N} s_j$ are then allocated in

proportion to the remaining benefits r_i, leading to the formula

$$SCRB_i(c) = s_i + \frac{r_i}{\sum_{j\in N} r_j}\left[c(N) - \sum_{j\in N} s_j\right] . \quad (4)$$

In this formula it is assumed that c is at least sub-additive to ensure that the r_i's are all positive. For the two-player game c in Example 10 the solution is given by $SCRB(c) = (2 + \frac{1}{2}(10-9), 7 + \frac{1}{2}(10-9)) = (2\frac{1}{2}, 7\frac{1}{2})$. In earlier days the solution was as well known as 'the alternate cost avoided method' or 'alternative justifiable expenditure method'. For references see [144].

Shapley Value One of the most popular and oldest solution concepts in the literature on cooperative games is due to Shapley [116], and named *Shapley-value*. Roughly it measures the average marginal impact of players. Consider an ordering of the players $\sigma : N \to N$ so that $\sigma(i)$ indicates the ith player in the order. Let $\sigma^*(i)$ be the set of the first i players according to σ; so $\sigma^*(1) = \{\sigma(1)\}, \sigma^*(2) = \{\sigma(1), \sigma(2)\}$, etc. The *marginal cost share vector* $m^\sigma(c) \in \mathbb{R}^N$ is defined by $m^\sigma_{\sigma(1)}(1) = c(\sigma(1))$ and for $i = 2, 3, \ldots, n$

$$m^\sigma_{\sigma(i)}(c) = c(\sigma^*(i)) - c(\sigma^*(i-1)) . \quad (5)$$

So according to m^σ each player is charged with the increase in costs when joining the coalition of players before her. Then the *Shapley-value* for c is defined as the average of all $n!$ marginal vectors, i. e.

$$\Phi(c) = \frac{1}{n!}\sum_\sigma m^\sigma(c) . \quad (6)$$

Example 11 Consider the airport game in Example 7. Then the marginal vectors are given by

σ	(123)	(132)	(213)	(231)	(312)	(321)
$\mu^\sigma(c)$	(12,8,13)	(12,0,21)	(0,20,13)	(0,20,13)	(0,0,33)	(0,0,33)

Hence the Shapley value of the corresponding game is $\Phi(c) = (4, 8, 21)$. Following [69,107], for airport games this allocation is easily interpreted as the allocation according to which each player pays an equal share of the cost of only those parts of the runway she uses. Then $c(e_1)$ is shared by all three players, $c(e_2)$ only by players 2 and 3, and, finally, $c(e_3)$ is paid in full by player 3. This interpretation extends to the class of standard fixed tree games, where instead of the *lattice* structure of the runway, there is a cost of a tree network to be shared, see [63].

If cost game is concave then the Shapley-value is in the core. Since then each marginal vector specifies a core-element, and in particular the Shapley-value as a convex combination of these. Reconsider the minimum cost spanning tree game c in Example 9, a non-concave game with non-empty core and $\Phi(c) = (2\frac{2}{3}, 2\frac{2}{3}, 6\frac{2}{3}, 4)$. Note that this is not a stable cost allocation since the coalition $\{2, 3\}$ would profit by defecting, $c(23) = 8 < 9\frac{1}{3} = \Phi_2(c) + \Phi_3(c)$. [50] show that in general games $\Phi(c) \in \text{core}(c)$ precisely when c is *average concave*. Although not credited as a core-selector, the classic way to defend the Shapley-value is by the following properties.

Symmetry Two players i, j are called *symmetric* in the cost game c if for all coalitions S not containing i, j it holds $c(S \cup \{i\}) = c(S \cup \{j\})$. A solution μ is *symmetric* if symmetric players in a cost game c get the same cost shares. If the cost game does not provide any evidence to distinguish between two players, symmetry is the property endorsing equal cost shares.

Dummy A player i in a cost game c is *dummy* if $c(S \cup \{i\}) = c(S)$ for all coalitions S. A solution μ satisfies *dummy* if $\mu_i(c) = 0$ for all dummy players i in c. So when a player has no impact on costs whatsoever, she can not be held responsible.

Additivity A solution is *additive* if for all cost games c_1, c_2 it holds that

$$\mu(c_1) + \mu(c_2) = \mu(c_1 + c_2) . \quad (7)$$

For accounting reasons, in multipurpose projects it is a common procedure to subdivide the costs related to the different activities (players) into cost categories, like salaries, maintenance costs, marketing, et cetera. Each category ℓ is associated with a cost game c_ℓ where $c_\ell(S)$ is the total of category ℓ cost made for the different activities in S; then $c(S) = \sum_\ell c_\ell(S)$ is the joint cost for S. Suppose a solution is applied to each of the cost categories separately, then under an additive solution the aggregate cost share of an activity is independent from the particular cross-section in categories.

Theorem 2 (Shapley [116]) *Φ is the unique solution on CG which satisfies all three properties dummy, symmetry, additivity.*

Note that *SCRB* satisfies dummy and symmetry, but that it does not satisfy additivity. The Shapley-value is credited with other virtues, like the following due to [144].

Consider the practical situation that several division managers simultaneously take steps to increase efficiency by decreasing joint costs, but one division manager establishes a greater relative improvement in the sense that its

marginal contribution to the cost associated with all possible coalitions increases. Then it is more than reasonable that this division should not be penalized. In a broader context this envisions the idea that each player in the cost game should be credited with the merits of 'uniform' technological advances.

Strong Monotonicity Solution μ is *strongly monotonic* if for any two cost games $c, \overline{c}$ it holds for all $i \in N$ that $c(S \cup \{i\}) - c(S) \geq \overline{c}(S \cup \{i\}) - \overline{c}(S)$ for all $S \subseteq N \setminus \{i\}$ implies $\mu_i(c) \geq \mu_i(\overline{c})$.

Anonymity is the classic property for solutions declaring independence of solution with respect to the name of the actors in the cost sharing problem. See e. g., [3,91,106]. Formally, the definition is as follows. For a given permutation $\pi : N \to N$ and $c \in CG$ define $\pi c \in CG$ by $\pi c(S) = c(\pi(S))$ for all $S \subseteq N$.

Anonymity Solution μ is *anonymous* if for all permutations π of N, and all $i \in N$, $\mu_{\pi(i)}(\pi c) = \mu_i(c)$ for all cost games c.

Theorem 3 (Young [144]) *The Shapley-value is the unique anonymous and strongly monotonic solution.*

[99] introduced the *balanced contributions axiom* for the model of *non-transferable utility games*, or games without side-payments, see [118]. Within the present context of CG, a solution μ satisfies the *balanced contributions axiom* if for any cost game c and for any non-empty subset $S \subseteq N$, $\{i, j\} \subseteq S \in N$ it holds that

$$\mu_i(S, c) - \mu_i(S \setminus \{j\}, c) = \mu_j(S, c) - \mu_j(S \setminus \{i\}, c) . \quad (8)$$

The underlying idea is the following. Suppose that players agree on using solution μ and that coalition S forms. Then $\mu_i(S, c) - \mu_i(S \setminus \{j\}, c)$ is the amount player i gains or loses when S is already formed and player j resigns. The balanced contributions axiom states that the gains and/or losses by other player's withdrawal from the coalition should be the same.

Theorem 4 (Myerson [99]) *There is a unique solution on CG that satisfies the balanced contributions axiom, and that is Φ.*

The balanced contribution property can be interpreted in a bargaining context as well. In the game c and with solution μ a player i can object against player j to the solution $\mu(c)$ when the cost share for j increases when i steps out of the cooperation, i. e. $\mu_j(N, c) \geq \mu_j(N \setminus \{i\})$. In turn, a *counter objection* by player j to this objection is an assertion that player i would suffer more when j ends cooperation, i. e. $\mu_j(N, c) - \mu_j(N \setminus \{i\}) \leq \mu_i(N, c) - \mu_i(N \setminus \{j\})$. The balanced contribution property is equivalent to the requirement that each objection is balanced by a counter objection. For an excellent overview of ideas developed in this spirit, see [74].

Another *marginalistic* approach is by [44]. Denote for $c \in CG$ the game restricted to the players in $S \subseteq N$ by (S, c). Given a function $P : CG \to \mathbb{R}$ which associates a real number $P(N, c)$ to each cost game c with player set N, the *marginal cost* of a player i is defined to be $D^i P(c) = P(N, c) - P(N \setminus \{i\}, c)$. Such a function P with $P(\emptyset, c) = 0$ is called *potential* if $\sum_{i \in N} D^i P(N, c) = c(N)$.

Theorem 5 (Hart & Mas-Colell [44]) *There exists a unique potential function P, and for every $c \in CG$ the resulting payoff vector $DP(N, c)$ coincides with $\Phi(c)$.*

Egalitarian Solution The Shapley-value is one of the first solution concepts proposed within the framework of cooperative cost games, but not the most trivial. This would be to neglect all asymmetries between the players and split total costs equally between them. But as one can expect egalitarianism in this pure form will not lead to a stable allocation. Just consider the two-player game in Example 10 where pure egalitarianism would dictate the allocation $(5, 5)$, which violates individual rationality for player 1.

In order to avoid these problems of course we can propose to look for the most egalitarian allocation *within* the core (see [7,31]). Then in this line of thinking what is needed in Example 10 is a minimal transfer of cost 2 to player 2, leading to the final allocation $(3, 7)$ – the *constrained egalitarian solution*. Although in the former example is was clear what allocation to chose, in general we need a tool to evaluate allocations for the degree of egalitarianism. The earlier mentioned papers all suggest the use of Lorenz-order (see, e. g. [8]). More precisely, consider two vectors of cost shares x and x' such that $x(N) = x'(N)$. Assume that these vectors are ordered in decreasing order so that $x_1 \geq x_2 \geq \cdots \geq x_n$ and $x'_1 \geq x'_2 \geq \cdots \geq x'_n$. Then x *Lorenz-dominates* x' – read x is more egalitarian than x' – if for all $k = 1, \ldots, n-1$ it holds that

$$\sum_{i=1}^{k} x_i \leq \sum_{i=1}^{k} x'_i , \quad (9)$$

with at least one strict inequality. That is, x is better for those paying the most.

Example 12 Consider the three allocations of cost 15 among three players $x = (6, 5, 4), x' = (6, 6, 3)$, and $x'' = (7, 4, 4)$. Firstly, x Lorenz-dominates x'' since $x_1 =$

$6 < 7 = x_1''$, and $x_1 + x_2 = x_1' + x_2'$. Secondly, x Lorenz-dominates x' since $x_1 = x_1', x_1 + x_2 < x_1' + x_2'$. Notice, however, that on the basis of only Eq. (9) we can not make any judgment what is the more egalitarian of the allocations x' and x''. Since $x_1' = 6 < 7 = x_1''$ but $x_1' + x_2' = 6 + 6 > 7 + 4 = x_1'' + x_2''$. The Lorenz-order is only a partial order.

The constrained egalitarian solution is the set of Lorenz-undominated allocations in the core of a game. Due to the partial nature of the Lorenz-order there may be more than one Lorenz-undominated elements in the core. And what if the core is empty? The constrained egalitarian solution is obviously not a straight-forward solution. The original idea of constrained egalitarianism as in [30], focuses on the Lorenz-core instead of the core. It is shown that there is at most one such allocation, that may exist even when the core of the underlying game is empty.

For concave cost games c, the allocation is well defined and denoted by $\mu^{\mathrm{E}}(c)$. In particular this holds for airport games. Intriguingly, empirical studies [1,2] show there is a tradition in using the solution for this type of problems.

For concave cost games c, there exists an algorithm to compute $\mu^{\mathrm{E}}(c)$. This method, due to [30], performs the following consecutive steps. First determine the maximal set S_1 of players minimizing the per capita cost $c(S)/|S|$, where $|S|$ is the size of the coalition S. Then each of these players in S_1 pays $c(S_1)/|S_1|$. In the next step determine the maximal set S_2 of players in $N \backslash S_1$ minimizing $c_2(S)/|S|$, where c_2 is the cost game defined by $c_2(S) = c(S_1 \cup S) - c(S_1)$. The players in S_2 pay $c_2(S_2)/|S_2|$ each. Continue in this way just as long as not everybody is allocated a cost share. Then in at most n steps this procedure results in an allocation of total cost, the constrained egalitarian solution. In short the algorithm is as follows

- **Stage 0:** Initialization, put $S_0^* = \varnothing, x^* = 0_N$, go to stage $t = 1$.
- **Stage *t***: Determine

$$S_t \in \arg\max_{S \neq \varnothing} \frac{c(S \cup S_{t-1}^*) - c(S_{t-1}^*)}{|S|} .$$

Put $S_t^* = S_{t-1}^* \cup S_t$ and for $i \in S_t$,

$$x_i^* := \frac{c(S_t^*) - c(S_{t-1}^*)}{|S_t|} .$$

If $S_t^* = N$, we are finished, put $\mu^{\mathrm{E}}(c) = x^*$. Else repeat the stage with $t := t + 1$.

For example, this algorithm can be used to calculate the constrained egalitarian solution for the airport game in Example 7. In the first step we determine $S_1 = \{1, 2\}$, together with cost shares 10 for the players 1 and 2. Player 3 is allocated the remaining cost in the next step; hence the corresponding final allocation is $\mu^{\mathrm{E}}(c) = (10, 10, 13)$.

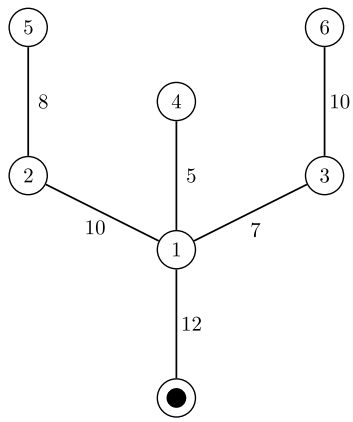

Cost Sharing, Figure 5
Standard fixed tree

Example 13 Consider the case where six players share the cost of the following *tree* network that connects them to $\odot$. The *standard fixed tree game c* for this network associates to each coalition of players the minimum cost of connecting cost of connecting each member to $\odot$, where it may use all the links of the tree. This type of games is known to be concave we can use the above algorithm to calculate $\mu^{\mathrm{E}}(c)$. In the first step we determine $S_1 = \{1, 3, 4\}$ and each herein pays 8. Then in the second step the game remains where the edges e_1, e_3, e_4 connecting S_1 have been paid for. Then it easily follows that $S_2 = \{2, 5\}$, so that players 2 and 5 pay 9 each, leaving 10 as cost share for player 6. Thus, we find $\mu^{\mathrm{E}}(c) = (8, 9, 8, 8, 9, 10)$.

Nucleolus Given a cost game $c \in CG$ the *excess* of a coalition $S \subset N$ with respect to a vector $x \in \mathbb{R}^N$ is defined as $e(S, x) = x(S) - c(S)$; it measures dissatisfaction of S under proposal x. Arrange the excesses of all coalitions $S \neq N, \varnothing$ in decreasing order and call the resulting vector $\vartheta(x) \in \mathbb{R}^{2^n-2}$. A vector of cost shares x will be preferred to a vector $\mathbf{y}$, notation $x \succ y$, whenever $\vartheta(x)$ is smaller than $\vartheta(y)$ in the lexicographic order, i. e. there exists i^* such that for $i \leq i^* - 1$ it holds $\vartheta_i(x) = \vartheta_i(y)$ and $\vartheta_{i^*}(x) < \vartheta_{i^*}(y)$. Schmeidler [115] showed that in the set of individual rational cost sharing vectors there is a unique element that is maximal with respect to $\succ$, which is called the *nucleolus*. This allocation, denoted by $\nu(c)$, is based on the idea of egalitarianism that the largest complaints of coalitions should consistently be minimized. The concept gained much popularity as a core-selector, i. e. it is

a one-point solution contained in the core when it is non-empty. This contrasts with the constrained egalitarian solution which might not be well defined, and the Shapley-value which may lay outside the core.

Example 14 Consider in Example 7 the excesses of the different coalitions with respect to the constrained egalitarian solution $\mu^{\mathrm{E}}(c) = (10, 10, 13)$ and the nucleolus $\nu(c) = (6, 7, 20)$:

S	1	2	3	12	13	23
$e(S, \mu^{\mathrm{E}}(c))$	-2	10	-20	0	-10	-10
$e(S, \nu(c))$	-6	-13	-13	-7	-7	-6

Then the ordered excess vectors are

$$\vartheta(10, 10, 13) = (10, 0, -2, -10, -10, -20)\,,$$
$$\vartheta(6, 7, 20) = (-6, -6, -7, -7, -13, -13)\,.$$

Note that indeed $\vartheta(\nu(c)) \succ \vartheta(\mu^{\mathrm{E}}(c))$ since

$$\vartheta_1(6, 7, 20) = -6 < 10 = \vartheta_1(10, 10, 13)\,.$$

The nucleolus of standard fixed tree games may be calculated as a particular *home-down* allocation, as was pointed out by Maschler et al. [75].

For standard fixed tree games and minimum cost spanning tree games the special structure of the technology makes it possible to calculate the nucleolus in polynomial time, i. e. with a number of calculations bounded by a multiple of n^2 (see [39]). Sometimes one may even express the nucleolus through a nice formula; Legros [66] showed a class of cost sharing problems for which the nucleolus equals the SCRB solution. But in general calculations are hard and involve solving a linear program with a number of inequalities which is exponential in n. [124] suggests to use the nucleolus on the cost game corresponding to hub-games.

Instead of the direct comparison of excesses like above, the literature also discusses weighted excesses as to model the asymmetries of justifiable complaints within coalitions. For instance the *per capita nucleolus* minimizes maximal excesses which are divided by the number of players in the coalition (see [105]).

Cost Sharing Rules Induced by Solutions

Most of the above numerical examples deal with cost sharing problems which have a natural and intuitive representation as a cost game. Then basically on such domains of cost sharing problems there is no difference between cost sharing rules and solutions. It may seem that the cooperative solutions are restricted to this kind of situations. But

$$\begin{tikzcd} (x, c) \arrow[r, "f"] \arrow[dr, dashed, "\overline{\mu}"'] & c_x \arrow[d, "\mu"] \\ & \mu(c_x) \end{tikzcd}$$

Cost Sharing, Figure 6
Induced cost sharing rules

recall that each cost sharing problem (x, c) is associated its stand-alone cost game $c_x \in CG$, as in Eq. (1). Now let μ be a solution on a subclass of $A \subset CG$ and B a class of cost sharing problems (x, c) for which $c_x \in A$. Then a cost sharing rule $\overline{\mu}$ is defined on B through

$$\overline{\mu}(x, c) = \mu(c_x)\,. \tag{10}$$

The general idea is illustrated in the diagram on the left. For example, since the Shapley value is defined on the class of all cost games, it defines a cost sharing rule $\overline{\Phi}$ on the class of all cost sharing problems. The cost sharing rule $\overline{\mu}^{\mathrm{E}}$ is defined on the general class of cost sharing problems with corresponding concave cost game. Cost sharing rules derived in this way, *game-theoretical rules* according to [130], will be most useful below.

Non-cooperative Cost Games

Formulating the cost sharing problem through a cooperative cost game assumes inelastic demands of the players. It might well be that for some player the private merits of service do not outweigh the cost share that is calculated by the planner. She will try to block the payment when no service at no cost is a preferred outcome. Another aspect is that the technology may operate at a sub-optimal level if benefits of delivered services are not taken into account. Below the focus is on a broader framework with elastic demands, which incorporates preferences of a player are defined over *combinations* of service levels and cost shares. The theory of non-cooperative games will provide a proper framework in which we can discuss individual aspirations and efficiency of outcomes on a larger scale.

Strategic Demand Games

At the heart of this non-cooperative theory is the notion of a *strategic game*, which models an interactive decision-making process among a group of *players* whose decisions may impact the consequences for others. Simultaneously, each player i independently chooses some available action

a_i and the so realized *action profile* $a = (a_1, a_2, \ldots, a_n)$ is associated with some consequence $f(a)$. Below we will have in mind demands or offered contributions as actions, and consequences are combinations of service levels with cost shares.

Preferences Over Consequences Denote by A the set of possible action profiles and C as the set of all consequences of action. Throughout we will assume that players have preferences over the different consequences of action. Moreover, such preference relation can be expressed by a *utility function* $u_i : C \to \mathbb{R}$ such that for $z, z' \in C$ it holds $u_i(z) \leq u_i(z')$ if agent i weakly prefers z' to z. Below the set of consequences for agent $i \in N$ will consist of pairs (x, y) where x is the level of service and y a cost share, so that a utilities are specified through multi-variable functions, $(x, y) \mapsto u_i(x, y)$.

Preferences Over Action Profiles In turn define for each agent i and all $a \in A$, $U_i(a) = u_i(f(a))$; then U_i assigns to each action profile the utility of its consequence. We will say that the action profile a' is weakly preferred to a by agent i if $U_i(a) \leq U_i(a')$; U_i is called agent i's utility function *over action profiles*.

Strategic Game and Nash Equilibrium A strategic game is an ordered triple $G = \langle N, (A_i)_{i \in N}, (U_i)_{i \in N} \rangle$ where

- $N = \{1, 2, \ldots, n\}$ is the set of *players*,
- A_i is the set of available *actions* for player i,
- U_i is player i's utility function over action profiles.

Rational players in a game will choose optimal actions in order to maximize utility. The most commonly used concept in game theory is that of *Nash-equilibrium*, a profile of strategies from where unilateral deviation by a single player does not pay. It can be seen as a *steady state* of action in which players hold correct beliefs about the actions taken by others and act rationally. An important assumption here is the level at which the players understand the game; usually it is taken as a starting point that players *know* the complete description of the game, including the action spaces and preferences of others.

Nash Equilibrium (Nash 1950) An action profile a^* in a strategic game $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$ is a Nash-equilibrium if, for every player i it holds $u_i(a^*) \geq u_i(a_i, a^*_{-i})$ for every $a_i \in A_i$.

The literature discusses several refinements of this equilibrium concept. One that will play a role in the games below is that of *strong Nash equilibrium* due to Aumann [9]; it is a Nash equilibrium a^* in a strategic game G such that for all $S \subset N$ and action profile a_S there exists a player $i \in S$ such that $u_i(a_S, a^*_{N \setminus S}) \leq u_i(a^*)$. This means that a strong Nash-equilibrium guarantees stability against coordinated deviations, since within the deviating coalition there is at least one agent who does not strictly improve.

Example 15 Consider the following two-player strategic game with $N = \{1, 2\}$, $A_1 = \{T, B\}$ and $A_2 = \{L, M, R\}$. Let the utilities be as in the table below

	L	***M***	***R***
T	5,4	2,1	3,2
B	4,3	5,2	2,5

Here player 1 chooses a row, and player 2 a column. The numbers in the cells summarize the individual utilities corresponding to the action profiles; the first number is the utility of player 1, the second that of player 2. In this game there is a unique Nash-equilibrium, which is the action profile (T, L).

Dominance in Strategic Games In the game $G = \langle N, (A_i)_{i \in N}, (U_i)_{i \in N} \rangle$, the action $a_i \in A_i$ is *weakly dominated* by $a'_i \in A_i$ if $U_i(a_i, a_{-i}) \leq U_i(a'_i, a_{-i})$ for all $a_{-i} \in A_{-i}$, with strict inequality for some profile of actions a_{-i}. If strict inequality holds for all a_{-i} then a_i is strictly dominated by $\tilde{a}_i$. Rational players will not use strictly dominated strategies, and, as far as prediction of play is concerned, these may be eliminated from the set of possible actions. If we do this elimination step for each player, then we may reconsider whether some actions are dominated within the reduced set of action profiles. This step-by-step reduction of action sets is called the procedure of *successive elimination of (strictly) dominated strategies*. The set of all action profiles surviving this procedure is denoted by D^∞.

Example 16 In Example 15 action M of player 2 is strictly dominated by L and R. Player 1 has no dominated actions. Now eliminate M from the actions for player 2. Then the reduced game is

	L	***R***
T	5,4	3,2
B	4,3	2,5

Notice that action L for player 1 was not dominated in the original game, for the reason that B was the better of the two actions against M. But if M is never played, T is strictly better than B. Now eliminate B, yielding the reduced game

	L	***R***
T	5,4	3,2

In this game, L dominates R; hence the only action profile surviving the successive elimination of strictly dominated strategies is (T, L).

A stronger notion than dominance is the following. Call an action $a_i \in A_i$ *overwhelmed* by $a'_i \in A_i$ if

$$\max\{U_i(a_i, a_{-i}) | a_{-i} \in A_{-i}\} < \min\left\{U_i(a'_i, a_{-i}) | a_{-i} \in A_{-i}\right\} .$$

Then $\mathcal{O}^\infty$ is the set of all action profiles surviving the successive elimination of overwhelmed actions. This notion is due to [34,37]. In Example 15 the action M is overwhelmed by L, not by R. Moreover, the remaining actions in $\mathcal{O}^\infty$ are B, T, L, and R.

Demand Games

Strategic games in cost sharing problems arise when we assume that the users of the production technology choose their demands strategically and a cost sharing rule sees to an allocation of the corresponding service costs. The action profiles A_i are simply specified by the demand spaces of the agents, and utilities are specified over combinations of (level of) received service and accompanying cost shares. Hence utilities are defined over consequences of action, $u_i(q_i, x_i)$ denotes i's utility at receiving service level q_i and cost share x_i; u_i is increasing in the level of received service x_i, and decreasing in the allocated cost y_i. Now assume a cost function c and a cost sharing rule μ. Then given a demand profile $a = (a_1, a_2, \ldots, a_n)$ the cost sharing rule determines a vector of cost shares $\mu(a, c)$, and in return also the corresponding utilities over demands $U_i(a) = u_i(a_i, \mu_i(a, c))$. Observe that agents influence each others utility via the cost component. The *demand game* for this situation is then the strategic game

$$G(\mu, c) = \langle N, (A_i)_{i\in N}, (U_i)_{i\in N}\rangle \ . \tag{11}$$

Example 17 Consider the airport problem in Example 7. Each player now may request service (1) or not (0). Then the cost function is fully described by the demand of the largest player. That is, $c(x) = 33$ if 3 requires service, $c(x) = 20$ for all profiles with $x_2 = 1, x_3 = 0$ and $c(x) = 12$ if $x = (1, 0, 0)$, $c(0, 0, 0) = 0$. Define the cost sharing rule $\overline{\Phi}(x, c) = \Phi(c_x)$, that is, $\overline{\Phi}$ calculates the Shapley-value for the underlying cost game c_x as in Eq. (1). Assume that the players' preferences over ordered pairs of service level and cost shares are fully described by

$$\begin{aligned} u_1(q_1, x_1) &= 8q_1 - x_1 \ , \\ u_2(q_2, x_2) &= 6q_2 - x_2 \ , \\ u_3(q_3, x_3) &= 30q_3 - x_3 \ . \end{aligned}$$

Here q_i takes values 0 (no service) or 1 (service) and x_i stands for the allocated cost. So player 1 prefers to be served at unit cost instead of not being served at zero cost, $u_1(0, 0) = 0 < 7 = u_1(1, 1)$. The infrastructure is seen as an excludable public good, so those with demand 0 do not get access to the technology. Each player now actively chooses to be served or not, so her action set is specified by $A_i = \{0, 1\}$. Recall the definition of c_x as in Eq. (1). Then given a profile of such actions $a = (a_1, a_2, a_3)$ and cost shares $\overline{\Phi}(a, c)$, utilities of the players in terms of action profiles become $U_i(a) = u_i(a_i, \Phi_i(a, c))$, so that

$$\begin{aligned} U_1(a) &= 8a_1 - \overline{\Phi}_1(a, c) \ , \\ U_2(a) &= 6a_1 - \overline{\Phi}_2(a, c) \ , \\ U_3(a) &= 30a_3 - \overline{\Phi}_3(a, c) \ . \end{aligned}$$

Now that we provided all details of the demand game $G(\overline{\Phi}, c)$, let us look for (strong) Nash-equilibria. Suppose that the action profile $a^* = (1, 0, 1)$ is played in the game. Then in turn the complete infrastructure is realized just for players 1 and 3 and the cost allocation is given by $(6, 0, 27)$. Then the vectors of individual utilities is given by $(2, 0, 3)$. Now if we consider unilateral deviations from a^*, what happens to the individual utilities?

$$\begin{aligned} U_1(0, 0, 1) &= 0 < 2 = U_1(1, 0, 1) \ , \\ U_2(1, 1, 1) &= 6 - \overline{\Phi}_2((1, 1, 1), c) \\ &= 6 - \Phi(c) = 6 - 8 \\ &= -2 < 0 = U_2(1, 0, 1) \ , \\ U_3(1, 0, 0) &= 0 < 3 = U_3(1, 0, 1) \ . \end{aligned}$$

This means that for each player unilateral deviation does not pay, a^* is a Nash-equilibrium. The first inequality shows as well why the action profile $(0, 0, 1)$ is not. It is easy to see that the other Nash equilibrium of this game is the action profile $(0, 0, 0)$, no player can afford the completion of the infrastructure just for herself. Notice however that this zero profile is not a strong Nash equilibrium as players 1 and 3 may well do better by choosing for service at the same time, ending up in $(1, 0, 1)$. The latter profile is the unique strong Nash-equilibrium of the game.

Similar considerations in the demand game $G(\overline{\mu}^{\mathrm{E}}, c)$ induced by the constrained egalitarian solution lead to the unique strong Nash-equilibrium $(0, 0, 0)$, nobody wants service.

With cost sharing rules as decentralization tools, the literature postulates Nash-equilibria of the related demand game as the resulting behavioral mode. This is a delicate step because – as the example above shows – it is easy

to find games with many equilibria, which causes a selection problem. And what can we do if there are no equilibria? This will not be the topic of this text and the interested reader is referred to any standard textbook on game theory, for instance see [103,104,109]. If there is a unique equilibrium then it is taken as the prediction of actual play.

Demand Revelation Games

For a social planner one way to retrieve the level at which to operate the production facility is via a pre-specified demand game. Another way is to ask each of the agents for the maximal amount that she is willing to contribute in order to get service, and then, contingent on the reported amounts, install an adequate level of service together with a suitable allocation of costs. Opposed to demand games ensuring the level of service, in a demand revelation game each player is able to ensure a maximum charge for service.

The approach will be discussed under the assumption of a discrete production technology with binary demands, so that the cost function c for the technology is basically the characteristic function of a cooperative game. Moreover assume that the utilities of the agents in N are quasi-linear and given by

$$u_i(q_i, x_i) = \alpha_i q_i - x_i \tag{12}$$

where $q_i \in \{0, 1\}$ denotes the service level, x_i stands for the cost share, and α_i is a non-negative real number. [93] discusses this framework and assume that c is concave, and [64,148] moreover take c as the joint cost function for the realization of several discrete public goods.

Demand Revelation Mechanisms Formally, a *revelation mechanism* M assigns to each profile η of reported maximal contributions a set $S(\eta)$ of agents receiving service and $x(\eta)$ a vector of monetary compensations. Here we will require that these monetary compensations are cost shares; given some cost sharing rule μ the vector $x(\eta)$ is given by $\mu(\mathbf{1}_{S(\eta)}, c)$ where c is the relevant cost function.

Moreover, note that by restricting ourselves to cost share vectors we implicitly assume non-positive monetary transfers. The *budget balance* condition is crucial here, otherwise mechanisms of a different nature must be considered as well, see [23,40,41]. There are other ways that a planner may use to determine a suitable service level is by demanding pre-payment from the players, and determine a suitable service level on the basis of these labeled contributions, see [64,148].

Many mechanisms come to mind, but in order to avoid too much arbitrariness from the planner's side, the more sensible ones will grant the players some control over the outcomes. We postulate the following properties:

- **Voluntary Participation (VP)** Each agent i can guarantee herself the welfare level $u_i(0, 0)$ (no service, no payment) by reporting truthfully the maximal willingness to pay, which is α_i under Eq. (12).
- **Consumer Sovereignty (CS)** For each agent i a report y_i exists so that she receives service, irrespective of the reports by others.

Now suppose that the planner receives the message $\eta = \alpha$, known to her as the profile of true player characteristics. Then for economic reasons she could choose to serve a coalition S of players that maximizes the *net benefit* at α, $\pi(S, \alpha) = \alpha(S) - c(S)$. However, problems will arise when some player i is supposed to pay more than α_i, so the planner should be more careful than that. She may want to choose a coalition S with maximal $\pi(S, \alpha)$ such that $\mu(\mathbf{1}_S, c) \leq \alpha$ holds; such set S is called *efficient*. But in general the planner cannot tell whether the players reported truthfully or not, then what should she do then? One option is that she applies the above procedure thereby naively holding each reported profile η for the *true* player characteristics. In other words, she will pick a coalition that solves the following optimization problem

$$\begin{aligned} \max_{S \subset N} \quad & \pi(S, \eta) = \eta(S) - c(S) \\ \text{s.t.} \quad & \mu(\mathbf{1}_S, c) \leq \eta \,. \end{aligned} \tag{13}$$

Denote such a set by $S(\mu, \eta)$. If this set is unique then the demand revelation mechanism $M(\mu)$ selects $S(\mu, \eta)$ to be served at cost shares determined by $x(\eta) = \mu(\mathbf{1}_S, c)$. This procedure will be explained through some numerical examples.

Example 18 Consider the airport problem and utilities of players over service levels and cost shares as in Example 7. Moreover assume the planner uses the Shapley cost sharing rule $\overline{\Phi}$ as in Example 17 and that she receives the true profile of preferences from the players, $\alpha = (8, 6, 30)$. Calculate for each coalition S the net benefits at α:

S	∅	1	2	3	12	13	23	N
$\pi(S, \alpha)$	0	-4	-14	-3	-6	8	3	11

Not surprisingly, the net benefits are the highest for the grand coalition. But if N were selected by the mechanism the corresponding cost shares are given by $\overline{\Phi}(\mathbf{1}_N, c) = (4, 8, 21)$, and player 2 is supposed to contribute more than she is willing to. Then the second highest net benefits are generated by serving $S = \{1, 3\}$, with cost shares $\overline{\Phi}(\mathbf{1}_S, c) = (6, 0, 27)$. Then $\{1, 3\}$ is the solution to Eq. (13).

What happens if some of the players misrepresent their preferences, for instance like in $\eta = (13, 6, 20)$? The planner determines the *conceived* net benefits

S	$\varnothing$	1	2	3	12	13	23	N
$\pi(S, \eta)$	0	1	-14	-13	-1	0	-7	6

Again, if the planner served the coalition with the highest net benefit, N, then player 2 would refuse to pay. Second highest net benefit corresponds to the singleton $S = \{1\}$, and this player will get service under $M(\overline{\Phi})$ since $\eta_3 = 13 > 12 = c(1, 0, 0)$.

Example 19 Consider the same situation but now with $\overline{\mu}^{\mathrm{E}}$ instead of $\overline{\Phi}$ as cost sharing rule. Now consider the following table, restricted to coalitions with non-negative net benefits (all other will not be selected):

S	$\varnothing$	13	23	N
$\pi(S, \alpha)$	0	8	3	11
$\overline{\mu}^{\mathrm{E}}(\mathbf{1}_S, c)$	(0,0,0)	(12,0,21)	$(0, \frac{33}{2}, \frac{33}{2})$	(10, 10, 13)

Here only the empty coalition $S = \varnothing$ satisfies the requirement $\overline{\mu}^{\mathrm{E}}(\mathbf{1}_S, c) = (0, 0, 0) \leq (8, 6, 30)$; hence according to the mechanism $M(\overline{\mu}^{\mathrm{E}})$ nobody will get service.

In general, the optimization problem Eq. (13) does not give unique solutions in case of which a planner should still further specify what she does in those cases. For concave cost functions, consider the following sequence of coalitions

$$S_1 = N, \;\; S_t = \{\, i \in S_{t-1} \mid \eta_i \geq \mu_i(\mathbf{1}_{S_t}, c)\} \;.$$

So, starting with the grand coalition N, at each consecutive step those players are removed whose maximal contributions are not consistent with the proposed cost share – until the process settles down. The set of remaining players defines a solution to Eq. (13) and taken to define $S(\mu, \eta)$.

Strategyproofness The essence of a demand revelation mechanism M is that its rules are set up in such a way that it provides enough incentives for the players not to lie about their true preferences. We will now discuss the most common non-manipulability properties of revelation mechanisms in the literature.

Fix two profiles $\alpha', \alpha \in \mathbb{R}^N_+$, where α corresponds to the true maximal willingness' to pay. Let (q', x') and (q, x) be the allocations implemented by the mechanism M on receiving the messages α' and α, respectively. The mechanism M is called *strategy-proof* if it holds for all $i \in N$ that $\alpha'_{N\setminus\{i\}} = \alpha_{N\setminus\{i\}}$ implies $u_i(q'_i, x'_i) \leq u_i(q_i, x_i)$.

So, given the situation that the other agents report truthfully, unilateral deviation by agent i from the true preference never produces better outcomes for her. Similarly, M is *group strategy-proof* if deviations of groups of agents does not pay for all deviators, i. e., for all $T \subset N$ the fact that $\alpha'_{N\setminus T} = \alpha_{N\setminus T}$ implies $u_i(q'_i, x'_i) \leq u_i(q_i, x_i)$ for all $i \in T$. So, under a (group) strategy-proof mechanism, there is no incentive to act untruthfully by misrepresenting the true preferences and this gives a benevolent planner control over the outcome.

Cross-Monotonicity Cost sharing rule μ is called *cross-monotonic* if the cost share of an agent i is not increasing if other agents demand more service, in case of a concave cost function c. Formally, if $\overline{x} \geq x$ and $\overline{x}_i = x_i$ then $\mu_i(\overline{x}, c) \leq \mu_i(x, c)$; each agent is granted a (fair) share of the positive externality due to an increase in demand by others.

Proposition 1 (Moulin & Shenker [93]) *The only group strategy-proof mechanisms $M(\mu)$ satisfying VP and CS are those related to cross-monotonic cost sharing rules μ.*

There are many different cross-monotonic cost sharing rules, and thus just as many mechanisms that are group strategy-proof. Examples include the mechanisms $M(\overline{\Phi})$ and $M(\overline{\mu}^{\mathrm{E}})$, because $\overline{\Phi}$ and $\overline{\mu}^{\mathrm{E}}$ are cross-monotonic. However, the nucleolus is not cross-monotonic and does therefore not induce a strategy-proof mechanism.

Above we discussed two instruments a social planner may invoke to implement a desirable outcome without knowing the true preferences of the agents. Basically, the demand revelation games define a so-called *direct mechanism*. The announcement of the maximal price for service pins down the complete preferences of an agent. So in fact the planner decides upon the service level based on a complete profile of preferences. In case of a cross-monotonic cost sharing rule, under the induced mechanism truth-telling is a weakly dominant strategy; announcing the true maximal willingness to pay is optimal for the agent regardless of the actions of others. This means good news for the social planner as the mechanism is self-organizing: the agents need not form any conjecture about the behavior of others in order to know what to do. In the literature [24] such a mechanism is called *straightforward*. The demand games define an *indirect mechanism*, as by reporting a demand the agents do no more than signaling their preferences to the planner.

Although in general there is a clear distinction between direct and indirect mechanisms, in the model presented in this section these are nevertheless strongly connected. Focus on a demand game $G(\mu, c)$; recall that in this game the agents simultaneously and independently decide upon requesting service or not and costs are shared using the rule μ amongst those agents receiving service. Sup-

pose that for each profile of utility functions as in Eq. (12) the resulting game $G(\mu, c)$ has a unique (strong) Nash-equilibrium. Then this equilibrium can be taken to define a mechanism. That is, the mechanism elicits u and chooses the unique equilibrium *outcome* of the reported demand game. Then this mechanism is equivalent with the demand revelation mechanism. Observe that indeed the strong equilibrium $(1, 0, 1)$ in the game $G(\overline{\Phi}, c)$ in Example 17 corresponds to players chosen by $M(\overline{\Phi})$ under truthful reporting. And where no player is served in the strong equilibrium of $G(\overline{\mu}^{\mathrm{E}}, c)$, none of the players is selected by $M(\overline{\mu}^{\mathrm{E}})$. It is a general result in implementation theory due to [24] that a direct mechanism constructed in this way is (group) strategy-proof *provided* the underlying space of preferences is *rich*. It is easily seen that the above sets of preferences meet the requirements. To stress importance of such a structural property as richness, it is instructive to point at what is yet to come in Sect. "Uniqueness of Nash-Equilibria in $\mathcal{P}^1$-Demand Games". Here, the strategic analysis of the demand game induced by the proportional rule shows uniqueness of Nash equilibrium on the domain of preferences $\mathcal{L}^*$ if only costs are convex. However, this domain is *not* rich, and the direct mechanism defined in the same fashion as above by the Nash equilibrium selection is *not* strategyproof.

Efficiency and Strategy-Proof Cost Sharing Mechanisms

Suppose *cardinal utility* for each agent, so that inter-comparison of utility is allowed. Proceeding on the net benefit of a coalition, we may define its *value* at α by

$$v(N, \alpha) = \max_{S \subset N} \pi(S, \alpha) \,, \tag{14}$$

where $\pi(S, \alpha)$ is the net benefit of S at α. A coalition S such that $v(N, \alpha) = \pi(S, \alpha)$ is called *efficient*. It will be clear that a mechanism $M(\mu)$ that is defined through the optimization problem Eq. (13) will not implement an efficient coalition of served players, due to the extra constraint on the cost shares.

For instance, in Example 7 the value of the grand coalition at $\alpha = (8, 6, 30)$ is given by $v(N, \alpha) = \alpha(N) - c(N) = 44 - 33 = 11$. At the same profile the implemented outcome by mechanism $M(\overline{\Phi})$ gives rise to a total surplus of $38 - 30 = 8$ for the grand coalition – which is not optimal. The mechanism $M(\overline{\mu}^{\mathrm{E}})$ performs even worse as it leads to the stand alone surplus 0, none is served.

This observation holds for far more general settings, and, moreover, it is a well known result from implementation theory that – under non-constant marginal cost – any strategy-proof mechanism based on full coverage of total costs will not always implement efficient outcomes. For the constant marginal cost case see [67,71]. Then, if there is an unavoidable loss in using demand revelation mechanisms, can we still tell which mechanisms are more efficient? Is it a coincidence that in the above examples the Shapley value performs better than the egalitarian solution?

The *welfare loss* due to $M(\mu)$ at a profile of true preferences α is given by

$$L(\mu, \alpha) = v(N, \alpha) - \{\alpha(S(\mu, \alpha)) - c(S)\} \,. \tag{15}$$

For instance, with $\alpha = (8, 6, 30)$ in the above examples we calculate $L(\overline{\Phi}, \alpha) = 11 - 8 = 3$ and $L(\overline{\mu}^{\mathrm{E}}, \alpha) = 11 - 0 = 11$. An overall measure of quality of a cost sharing rule μ in terms of efficiency loss is defined by

$$\gamma(\mu) = \sup_{\alpha} L(\mu, \alpha) \,. \tag{16}$$

Theorem 6 (Moulin & Shenker [93]) *Among all mechanisms $M(\mu)$ derived from cross-monotonic cost sharing rules μ, the Shapley rule $\overline{\Phi}$ has the unique smallest maximal efficiency loss, or $\gamma(\overline{\Phi}) < \gamma(\mu)$ if $\mu \neq \overline{\Phi}$.*

Notice that this makes a strong case for the Shapley-value against the egalitarian solution. The story does, however, not end here.

[98] considers a model where the valuations of the agents for the good are independent random variables, drawn from a distribution function F satisfying the *monotone hazard condition*. This means that the function defined by $h(x) = f(x)/(1 - F(x))$ is non-decreasing, where F has f as density function. It is shown that the constrained egalitarian solution maximizes the probability that all members of any given coalition accept the cost shares imputed to them. Moreover, [98] characterized the solution in terms of efficiency. Suppose for the moment that the planner calculated cost share vector $\mathbf{x}$ for the coalition S, and that its members are served conditional on acceptance of the proposed cost shares. The probability that all members of S accept the shares is given by $P(x) = \prod_{i \in S}(1 - F(x_i))$, and if we assume that the support of F is $(0, m)$ then the expected surplus from such an offer can be calculated as follows

$$W(x) = P(x) \cdot \left[\sum_{i \in S} \int_{x_i}^{m} \frac{u_i}{1 - F(x_i)} \mathrm{d}F(u_i) - c(S) \right] . \tag{17}$$

The finding of [98] is that for *log-concave* f, i. e. $x \mapsto \ln(f(x))$ is concave [5], the mechanism based on the constrained egalitarian solution not only maximizes the probability that a coalition accepts the proposal, but it maximizes its expected surplus as well. Formally, the result is the following.

Theorem 7 (Mutuswami (2004)) *If the profile of valuations $(u_i)_{i\in N}$ are independently drawn form a common distribution function F with log-concave and differentiable density function f, then $W(\overline{\mu}^{\mathrm{E}}(\mathbf{1}_S, c)) \geq W(\mu(\mathbf{1}_S, c))$ for all cross monotonic solutions μ and all $S \subset N$.*

Extension of the Model: Discrete Goods

Suppose the agents consume idiosyncratic goods produced in indivisible units. Then given a profile of demands the cost associated with the joint production must be shared by the users. Then this model generalizes the binary good model discusses so far and it is a launch-pad to the continuous framework in the next section. In this discrete good setting [86] characterizes the cost sharing rules which induce strategyproof social choice functions defined by the equilibria of the corresponding demand game. As it turns out, these rules are *basically* the *sequential stand alone rules*, according to which costs are shared in an incremental fashion with respect to a fixed ordering of the agents. This means that such a rule charges the first agent for her stand alone costs, the second for the stand alone cost for the first two users minus the stand alone cost of the first, et cetera. Here the word 'basically' refers to all discrete cost sharing problems other than those with binary goods. Then here the sufficient condition for strategyproofness is that the underlying cost sharing rule be cross monotonic, which admits other rules than the sequential ones – like $\overline{\mu}^{\mathrm{E}}$ and $\overline{\Phi}$.

Continuous Cost Sharing Models

Continuous Homogeneous Output Model, $\mathcal{P}^1$

This model deals with production technologies for one single perfectly divisible output commodity. Moreover, we will restrict ourselves to private goods. Many ideas below have been studied for *public goods* as well, for further references see, e. g., [33,72,82].

The demand space of an individual is given by $X = \mathbb{R}_+$. The technology is described by a non-decreasing cost function $c\colon \mathbb{R}_+ \to \mathbb{R}$ such that $c(0) = 0$, i. e. there are no fixed costs. Given a profile of demands $x \in \mathbb{R}^N_+$ costs $c(x(N))$ have to be shared. Moreover, the space of cost functions will be restricted to those c being *absolutely continuous*. Examples include the differentiable and Lipschitz-continuous functions. Absolute continuity implies that aggregate costs for production can be calculated by the total of marginal costs

$$c(y) = \int_0^y c'(t)\,\mathrm{d}t\,.$$

Denote the set of all such cost functions by C^1 and the related cost sharing problems by $\mathcal{P}^1$. Several cost sharing rules on $\mathcal{P}^1$ have been proposed in the literature.

Average Cost Sharing Rule This is the most popular and oldest concept in the literature and advocates Aristotle's principle of proportionality.

$$\mu^{\mathrm{AV}}(x, c) = \begin{cases} \frac{x}{x(N)} \cdot c(x(N)) & \text{if } x \neq 0_N \\ 0 & \text{if } x = 0_N \end{cases} \tag{18}$$

Shapley–Shubik Rule Each cost sharing problem $(x, c) \in \mathcal{P}^1$ is related to the stand-alone cost game c_x such that $c_x(S) = c(x(S))$. Then the Shapley–Shubik rule is determined by application of the Shapley-value to this game:

$$\mu^{\mathrm{SS}}(x, c) = \Phi(c_x)\,.$$

Serial Rule This rule, due to Moulin and Shenker [91], determines cost shares by considering particular intermediate cost levels. More precisely, given $(x, c) \in \mathcal{P}^1$ it first relabels the agents by increasing demands such that $x_1 \leq x_2 \leq \cdots \leq x_n$. The intermediate production levels are

$$\begin{aligned} y^1 &= nx_1,\ y^2 = x_1 + (n-1)x_2,\ \ldots\,, \\ y^k &= \sum_{j=1}^{k-1} x_j + (n-k+1)x_k,\ \ldots\,, \\ y^n &= x(N)\,. \end{aligned}$$

These levels are chosen such that at each new level one agent more is fully served his demand: at y^1 each agent is handed out x_1, at y^2 agent 1 is given x_1 and the rest x_2, etc. The serial cost shares are now given by

$$\mu_i^{\mathrm{SR}}(x, c) = \sum_{\ell=1}^{k} \frac{c(y^\ell) - c(y^{\ell-1})}{n - \ell + 1}\,.$$

So according to μ^{SR} each agent pays a fair share of the incremental costs in each stage that she gets new units.

Example 20 Consider the cost sharing problem (x, c) with $x = (10, 20, 30)$ and $c(y) = \frac{1}{2}y^2$. Then first calculate the intermediate production levels $y^0 = 0$, $y^1 = 30$, $y^2 = 50$, and $y^3 = 60$. Then the cost shares are calculated as follows

$$\mu_1^{\mathrm{SR}}(x, c) = \frac{c(y^1) - c(y^0)}{3} = 150\,,$$

$$\mu_2^{\mathrm{SR}}(x,c) = \mu_1^{\mathrm{SR}}(x,c) + \frac{c(y^2)-c(y^1)}{2}$$
$$= 150 + \frac{1250-450}{2} = 550\,,$$
$$\mu_3^{\mathrm{SR}}(x,c) = \mu_2^{\mathrm{SR}}(x,c) + c(y^3) - c(y^2) = 550 + 550$$
$$= 1100\,.$$

The serial rule has attracted much attention lately in the network literature, and found its way in fair queuing packet scheduling algorithms in routers [27].

Decreasing Serial Rule De Frutos [26] proposes serial cost shares where demands of agents are put in decreasing order. Resulting is the *decreasing serial rule*. Consider a demand vector $x \in \mathbb{R}_+^N$ such that $x_1 \le x_2 \le \cdots \le x_n$. Define recursively the numbers y^ℓ for $\ell = 1, 2, \ldots, n$ by $y^\ell = \ell x_\ell + x_{\ell+1} + \cdots + x_n$, and put $y^{n+1} = 0$. Then the decreasing serial rule is defined by

$$\mu_i^{\mathrm{DSR}}(x,c) = \sum_{\ell=i}^{n} \frac{c(y^\ell) - c(y^{\ell+1})}{\ell}\,. \tag{19}$$

Example 21 For the cost sharing problem in Example 20 calculate $y^1 = 90$, $y^2 = 70$, $y^1 = 60$, then

$$\mu_3^{\mathrm{DSR}}(x,c) = \frac{c(y^3)-c(y^4)}{3} = \frac{4050-0}{3} = 1350\,,$$
$$\mu_2^{\mathrm{DSR}}(x,c) = \mu_3^{\mathrm{DSR}}(x,c) + \frac{c(y^2)-c(y^3)}{2}$$
$$= 1350 + \frac{2450-4050}{2} = 550\,,$$
$$\mu_1^{\mathrm{DSR}}(x,c) = \mu_2^{\mathrm{DSR}}(x,c) + (c(y^1)-c(y^2))$$
$$= 550 + (1800-2450) = -100\,.$$

Notice that here the cost share of agent 1 is negative, due to the convexity of c! This may be considered as an undesirable feature of the cost sharing rule. Not only are costs increasing in the level of demand, in case of a convex cost function each agent contributes to the negative externality. It seems fairly reasonable to demand a non-negative contribution in those cases, so that none profits for just being there. The mainstream cost sharing literature includes positivity of cost shares into the very definition of a cost sharing rule. Here we will add it as a specific property:
Positivity μ is *positive* if $\mu(x,c) \ge 0_N$ for all (x,c) in its domain. All earlier discussed cost sharing rules have this property, except for the decreasing serial rule.

The decreasing serial rule is far more intuitive in case of economies of scale, in presence of a concave cost function. The larger agents now are credited with a lower price per unit of the output good. [47,57] propose variations on the serial rule that coincide with the increasing (decreasing) serial rule in case of a convex (concave) cost function, meeting the positivity requirement.

Marginal Pricing Rule A popular way of pricing an output of a production facility is *marginal cost pricing*. The price of the output good is set to cover the cost producing one extra unit. It is frequently used in the domain of public services and utilities. However, a problem is that for concave cost functions the method leads to budget deficits. An adapted form of marginal cost pricing splits these deficits equally over the agents. The *marginal pricing rule* is defined by

$$\mu_i^{\mathrm{MP}}(x,c) = x_i c'(x(N)) + \tfrac{1}{n}\left[c(x(N)) - x(N)c'(x(N))\right]. \tag{20}$$

Note that in case of convex cost functions agents can receive negative cost shares, just like it is the case with decreasing serial cost sharing.

Additive Cost Sharing and Rationing

The above cost sharing rules for homogeneous production models share the following properties:

Additivity $\mu(x, c_1 + c_2) = \mu(x, c_1) + \mu(x, c_2)$ for all relevant cost sharing problems. This property carries the same flavor as the homonymous property for cost games.

Constant Returns $\mu(x,c) = \vartheta x$ for linear cost functions c such that $c(y) = \vartheta y$ for all y. So if the agents do not cause any externality, the fixed marginal cost is taken as a price for the good.

It turns out that the class of all positive cost sharing rules with these properties can be characterized by solutions to *rationing problems* – which are the most basic of all models of distributive justice. A rationing problem amongst the agents in N consists of a pair $(x,t) \in \mathbb{R}_+^N \times \mathbb{R}_+$ such that $x(N) \ge t$; t is the available amount of some (in)divisible good and x is the set of demands. The inequality sees to the interpretation of rationing as not every agent may get all she wants.

A *rationing method* r is a solution to rationing problems, such that each problem (x,t) is assigned a vector of shares $r(x,t) \in \mathbb{R}_+^N$ such that $\mathbf{0}_N \le r(x,t) \le x$. The latter restriction is a weak solidarity statement assuring that everybody's demand be rationed in case of shortages.

For $t \in \mathbb{R}_+$ define the special cost function Γ_t by $\Gamma_t(y) = \min\{y, t\}$. The cone generated by these base

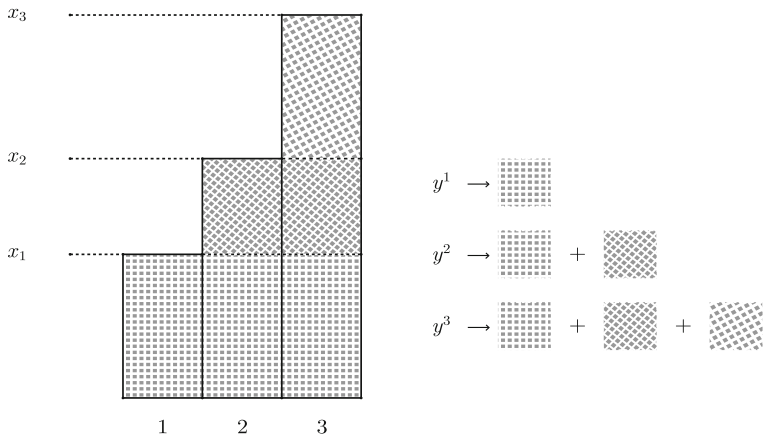

Cost Sharing, Figure 7
Intermediate production levels

functions lays dense in the space of all *absolutely continuous* cost functions c; if we know what the values $\mu(x, \Gamma_t)$ are, then basically we know $\mu(x, c)$. Denote by $\mathcal{M}$ the class of all cost sharing rules with the properties positivity, additivity, and constant returns.

Theorem 8 (Moulin & Shenker [88,92]) *Consider the following mappings associating rationing methods with cost sharing rules and vice versa,*

$$r \mapsto \mu\colon \mu(x, c) = \int_0^{x(N)} c'(t) dr(x, t) ,$$
$$\mu \mapsto r\colon r(x, t) = \mu(x, \Gamma_t) .$$

These define an isomorphism between $\mathcal{M}$ and the space of all monotonic rationing methods.

So each monotonic rationing method relates to a cost sharing rule and vice versa. In this way μ^{P} is directly linked with the proportional rationing method, μ^{SR} to the *uniform gains* method, and μ^{SS} to the *random priority* method. Properties of rationing methods lead to properties of cost sharing rules and vice versa [61].

Incentives in Cooperative Production

Stable Allocations, Stand-Alone Core Suppose again, like in the framework of cooperative cost games, that (coalitions of) agents can decide to leave the cost sharing and organize their own production facility. Under the ability to replicate the technology the question arises whether cost sharing rules induce stable cost share vectors.

Theorem 9 (Moulin [85]) *For concave cost functions c, if μ is an anonymous and cross-monotonic cost sharing rule then $\mu(x, c) \in \mathrm{core}(c_x)$.*

Under increasing returns to scale, this implies that $\mu^{\mathrm{P}}, \mu^{\mathrm{SR}}$ are core-selectors but μ^{MC} is not. [137] associates to each cost sharing problem (x, c) a *pessimistic* one, (x, c^*); here $c^*(y)$ reflects the maximum of marginal cost on $[0, x(N)]$ to produce y units,

$$c^*(y) = \begin{cases} \sup\left\{\int_T c'(t)\mathrm{d}t \mid T \subset [0, x(N)], \lambda(T) = y\right\} & \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{if } y \le x(N) , \\ c(y) \qquad\qquad\qquad\qquad\qquad\qquad \text{else} . \end{cases} \tag{21}$$

Here λ denotes the Lebesgue measure.

Theorem 10 (Koster [60]) *For any cost sharing problem (x, c), it holds $\mathrm{core}(c_x^*) = \{\mu(x, c) \mid \mu \in \mathcal{M}\}$.*

In particular this means that for $\mu \in \mathcal{M}$ it holds that $\mu(x, c) \in \mathrm{core}(c_x)$ whenever c is concave, since this implies $c^* = c$. This result appeared earlier as a corollary to Theorem 8, see [88]. [47,57,61] show *non-linear* cost sharing rules yielding core elements for concave cost functions as well, so additivity is only a sufficient condition in the above statement. For average cost sharing, one can show more, $\mu^{\mathrm{P}}(x, c) \in \mathrm{core}(c_x)$ for all x precisely when the average cost $c(y)/y$ is decreasing in y.

Strategic Manipulation Through Reallocation of Demands In the cooperative production model, there are other ways that agents may use to manipulate the final allocation. In particular, note that the serial procedure gives the larger demanders an advantage in case of positive externalities; as marginal costs decrease, the price paid by the larger agents per unit of output is lower than that of the smaller agents. In the other direction larger demanders are

punished if costs are convex. Then as the examples below show, this is just why the serial ideas are vulnerable to misrepresentation of demands, since combining demands and redistribute the output afterwards can be beneficial.

Example 22 Consider the cost function c given by $c(y) = \min\{5y, 60 + 2y\}$. Such cost functions are part of daily life, whenever one has to decide upon telephone or energy supply contracts: usually customers get to choose between a contract with high fixed cost and a low variable cost, and another with low or no fixed cost and a high variable price. Now consider the two cost sharing problems (x, c) and (x', c) where $x = (10, 20, 30), x' = (0, 30, 30)$. The cost sharing problem (x', c) arises from (x, c) if agent 2 places a demand on behalf of agent 1 – without letting agent 3 know. The corresponding average and serial cost shares are given by

$$\mu^{\mathrm{P}}(x, c) = (30, 60, 90) \quad \mu^{\mathrm{P}}(x', c) = (0, 90, 90)$$
$$\mu^{\mathrm{SR}}(x, c) = (40, 60, 80) \quad \mu^{\mathrm{SR}}(x', c) = (0, 90, 90)\,.$$

Notice that the total of average cost shares for 1 and 2 is the same in both cost sharing problems. But if the serial rule were used, these agents can profit by *merging* their demands; if agent 2 returns agent 1's demand and requires a payment from agent 1 between €30 and €40, then both agents will have profited by such *merging* of demand.

Example 23 Consider the five-agent cost sharing problems (x, c) and $(\overline{x}, c)$ with $x = (1, 2, 3, 0, 0), \overline{x} = (1, 2, 1, 1, 1)$ and convex cost function $c(y) = \frac{1}{2}y^2$. $(\overline{x}, c)$ arises out of (x, c) if agent 3 splits her demand over agents 4 and 5 as well. Then

$$\mu^{\mathrm{P}}(x, c) = (6, 12, 18, 0, 0) \quad \mu^{\mathrm{P}}(\overline{x}, c) = (6, 12, 6, 6, 6)\,,$$
$$\mu^{\mathrm{SR}}(x, c) = (3, 11, 22, 0, 0) \quad \mu^{\mathrm{SR}}(\overline{x}, c) = (5, 16, 5, 5, 5)\,.$$

The aggregate of average cost shares for agents 3, 4, and 5 does not change. But notice that according to the serial cost shares, there is a clear advantage for the agents. Instead of paying 22 in the original case, now the total of payments equals 15. Agent 3 may consider a transfer between 0 and 7 to 3 and 4 for their collaboration and still be better of. In general, in case of a convex cost function the serial rule is vulnerable with respect to manipulation of demands through splitting.

Note that in the above cases the proportional cost sharing rule does prescribe the same cost shares. It is a non-manipulable rule: reshuffling of demands will not lead to different aggregate cost shares. The rule does not discriminate between units, when a unit is produced is irrelevant. It is actually a very special feature of the cost sharing rule that is basically not satisfied by any other cost sharing rule.

Theorem 11 *Assume that N contains at least three agents. The proportional cost sharing rule is the unique rule that charges nothing for a null demand and meets any one of the following properties:*

- *Independence of merging and splitting,*
- *No advantageous reallocation,*
- *Irrelevance of reallocation.*

The second property shows even a stronger property than merging and splitting: agents may redistribute the demands in any preferred way without changing the aggregate cost shares of the agents involved. The third property states that in such cases the cost shares of the other agents do not change. Then this makes proportional cost sharing compelling in situations where one is not capable of detecting the true demand characteristics of individuals.

Demand Games for $\mathcal{P}^1$

Consider demand games $G(\mu, c)$ as in Eq. (11), Sect. "Demand Games", where now μ is a cost sharing rule on $\mathcal{P}^1$. These games with uncountable strategy spaces are more complex than the demand games that we studied before.

The set of consequences for players is now given by $C = \mathbb{R}^2_+$, combinations of levels of production and costs (see Sect. "Strategic Demand Games"). Then an individual i's preference relation is *convex* if for the corresponding utility function u_i and all pairs $z, z' \in C$ it holds

$$u_i(z) = u_i(z') \implies u(tz + (1-t)z) \geq u_i(z) \quad \text{for all } t \in [0, 1]\,. \tag{22}$$

This means that a weighted average of the consequences is weakly preferred to both consequences, if these are equivalent. Such utility functions u_i are called *quasi-concave*. An example of convex preferences are those related to *linear utility* functions of type $u_i(x, y) = \alpha x - y$. Moreover, *strictly convex* preferences are those with strict inequality in Eq. (22) for $0 < t < 1$; the corresponding utility functions are *strictly quasi-concave*. Special classes of preferences are the following.

- $\mathcal{L}$: the class of all convex and continuous preferences utility functions that are non-decreasing in the service component x, non-increasing in the cost component y, non-locally satiated and decreasing on $(x, c(x))$ for x large enough. The latter restriction is no more than assuring that agents will not place requests for unlimited amounts of the good.

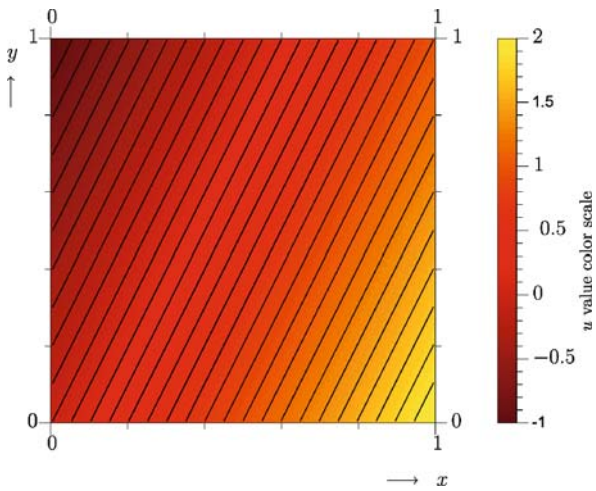

Cost Sharing, Figure 8
Linear, convex preferences, $u(x, y) = 2x - y$. The contours indicate indifference curves, i. e. sets of type $\{(x, y) | u(x, y) = k\}$, the k-level curve of u

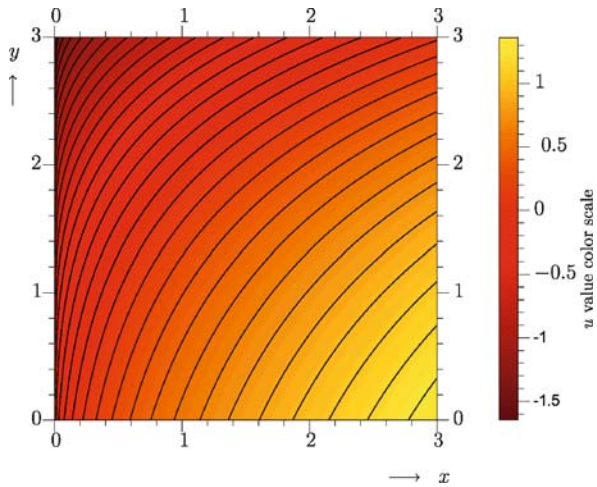

Cost Sharing, Figure 9
Strictly convex preferences, $u(x, y) = \sqrt{x} - e^{0.5y}$. The straight line connecting any two points on the same contour lays in the lighter area – with higher utility. Fix the y-value, then an increase of x yields higher utility, whereas for fixed x an increase of y causes the utility to decrease

- $\mathcal{L}^*$: the class of *bi-normal* preferences in $\mathcal{L}$. Basically, if such a preference is represented by a differentiable utility function u then the slope dy/dx of the indifference contours is non-increasing in x, non-decreasing in y. For a concise definition see [141]. Examples include Cobb–Douglas utility functions and also those of type $u_i(x, y) = \alpha(x) - \beta(y)$ where α and β are concave and convex functions, respectively. A typical plot of level curves of such utility functions is in Fig. 9. Note that the approach differs from the standard literature where agents have preferences over endowments. Here costs are 'negative' endowments. In the latter interpretation, the condition can be read as that the *marginal rate of substitution* is non-positive. At equal utility, an increase of the level of output has to be compensated by a decrease in the level of input good.

Nash-Equilibria of Demand Games in a Simple Case

Consider a production facility shared by three agents $N = \{1, 2, 3\}$ with cost function $c(y) = \frac{1}{2}y^2$. Assume that the agents have quasi-linear utilities in $\mathcal{L}$, i. e. $u_i(x_i, y_i) = \alpha_i x_i - y_i$ for all pairs $(x_i, y_i) \in \mathbb{R}^2_+$. Below the Nash-equilibrium in the serial and proportional demand game is calculated in two special cases. This numerical example is based on [91].

Proportional Demand Game Consider the corresponding proportional demand game, $G(\mu^{\mathrm{P}}, c)$, with utility over actions given by

$$U_i^{\mathrm{P}}(x) = \alpha_i x_i - \mu^{\mathrm{P}}(x, c) = \alpha_i x_i - \tfrac{1}{2} x_i x(N) \,. \tag{23}$$

In a Nash-equilibrium x^* of $G(\mu^{\mathrm{P}}, c)$ each player i gives a best response on x^*_{-i}, the action profile of the other agents. That is, player i chooses $x_i^* \in \arg\max_t U_i^{\mathrm{P}}(t, x^*_{-i})$. Then first order conditions implies for an interior solution

$$\alpha_i - \tfrac{1}{2} x^*(N) - \tfrac{1}{2} x_i^* = 0 \tag{24}$$

for all $i \in N$. Then $x^*(N) = \frac{1}{2}(\alpha_1 + \alpha_2 + \alpha_3)$ and $x_i^* = 2\alpha_i - \frac{1}{2}(\alpha_1 + \alpha_2 + \alpha_3)$.

Serial Demand Game Consider the same production facility and the demand game $G(\mu^{\mathrm{SR}}, c)$, corresponding to the serial rule. Then the utilities over actions are given by

$$U_i^{\mathrm{SR}}(x) = \alpha_i x_i - \mu^{\mathrm{SR}}(x, c) \,. \tag{25}$$

Now suppose $\overline{x}$ is a Nash equilibrium of this game, and assume without loss of generality that $\overline{x}_1 \le \overline{x}_2 \le \overline{x}_3$. Then player 1 with the smallest equilibrium demand maximizes the expression

$$U_1^{\mathrm{SR}}((t, \overline{x}_2, \overline{x}_3) = \alpha_1 t - c(3t)/3 = \alpha_1 t - \tfrac{3}{2} t^2$$

at $\overline{x}_1$, from which we may conclude that $\alpha_1 = 3\overline{x}_1$. In addition, in equilibrium, player 2, maximizes

$$U_2^{\mathrm{SR}}(\overline{x}_1, t, \overline{x}_3) = \alpha_2 t - (\tfrac{1}{3} c(3\overline{x}_1) + \tfrac{1}{2}(c(\overline{x}_1 + 2t) - c(3\overline{x}_1)),$$

for $t \ge \overline{x}_1$, yielding $\alpha_2 = \overline{x}_1 + 2\overline{x}_2$. Finally, the equilibrium condition for player 3 implies $\alpha_3 = \overline{x}(N)$. Then it is not hard to see that actually this constitutes *the* serial equilibrium.

Comparison of Proportional and Serial Equilibria (I)

Now let's compare the serial and the proportional equilibrium in the following two cases:

(i) $\alpha_1 = \alpha_2 = \alpha_3 = \alpha$
(ii) $\alpha_1 = \alpha_2 = 2, \alpha_3 = 4$.

Case (i): Then we get $x^* = (\frac{1}{2}\alpha, \frac{1}{2}\alpha, \frac{1}{2}\alpha)$ and $\overline{x}_i = (\frac{1}{3}\alpha, \frac{1}{3}\alpha, \frac{1}{3}\alpha)$ for all i. The resulting equilibrium payoff vectors are given by

$$U^{\mathrm{P}}(x^*) = (\tfrac{1}{8}\alpha^2, \tfrac{1}{8}\alpha^2, \tfrac{1}{8}\alpha^2) \text{ and}$$
$$U^{\mathrm{SR}}(\overline{x}) = (\tfrac{1}{6}\alpha^2, \tfrac{1}{6}\alpha^2, \tfrac{1}{6}\alpha^2) .$$

Not only the average outcome is less efficient than its serial counterpart, it is also Pareto-inferior to the latter.

Case (ii): The proportional equilibrium is a boundary solution, $x^* = (0, 0, 2)$ with utility profile $U^{\mathrm{P}}(x^*) = (0, 0, 8)$. The serial equilibrium strategies and utilities are given by

$$\overline{x} = (\tfrac{2}{3}, \tfrac{2}{3}, \tfrac{8}{3}), U^{\mathrm{SR}}(\overline{x}) = (\tfrac{2}{3}, \tfrac{2}{3}, 4) .$$

Notice that the serial equilibrium is now less efficient, but is not Pareto-dominated by the proportional utility distribution.

Uniqueness of Nash-Equilibria in $\mathcal{P}^1$-Demand Games

In the above demand games there is a unique Nash-equilibrium which serves as a prediction of actual play. This need not hold for any game. In the literature strategic characterizations of cost sharing rules are discussed in terms of uniqueness of equilibrium in the induced cost games, relative to specific domains of preferences and cost functions. Below we will discuss the major findings of [141]. These results concern a broader cost sharing model with the notion of a cost function as a differentiable function $\mathbb{R}_+ \to \mathbb{R}_+$. So in this paragraph such cost function can decrease, fixed cost need not be 0. This change in setup is not crucial to the overall exposition since the characterizations below are easily interpreted within the context of $\mathcal{P}^1$.

Demand Monotonicity The mapping $t \mapsto \mu_i((t, x_{-i}), c)$ is non-decreasing; μ is strictly demand monotonic if this mapping is increasing whenever c is increasing.

Smoothness The mapping $x \mapsto \mu(x, c)$ is continuously differentiable for all continuously differentiable $c \in C^1$.

Recall that $\mathcal{L}^*$ is the domain of all bi-normal preferences.

Theorem 12 (Watts [141]) *Fix a differentiable cost function c and a demand monotonic and smooth cost sharing rule μ. A cost sharing game $G(\mu, c)$ has a unique equilibrium whenever agents' preferences belong to $\mathcal{L}^*$, only if, for all $x = (x_1, \ldots, x_n)$*

- *Every principal minor of the matrix W with rows w^i is non-negative for all*

$$w^i \in \left\{ \left(\frac{\partial \mu_i}{\partial x_1}, \ldots, \frac{\partial \mu_i}{\partial x_n} \right), \left(\frac{\partial^2 \mu_i}{\partial x_1 \partial x_1}, \ldots, \frac{\partial^2 \mu_i}{\partial x_i \partial x_n} \right) \right\} . \tag{26}$$

- *The determinant of the Hessian matrix corresponding to the mapping $x \mapsto \mu(x, c)$ is strictly positive.*

A sufficient condition to have uniqueness of equilibrium is that the principle minor of the matrix W is strictly positive.

The impact of this theorem is that one can characterize the class of cost functions yielding unique equilibria if the domain of preferences is $\mathcal{L}^*$.

- $G(\mu^{\mathrm{SR}}, c), G(\mu^{\mathrm{DSR}}, c)$: Necessary condition for uniqueness of equilibrium is that c is strictly convex, i. e. $c'' > 0$. A sufficient condition is that c is increasing and strictly convex. Actually, [141] also shows that the conclusions for the serial rule do not change when $\mathcal{L}$ is used instead of $\mathcal{L}^*$. As will get more clear below, the serial games have unique strategic properties.
- $G(\mu^{\mathrm{P}}, c)$: The necessary and sufficient conditions are those for the serial demand game, including $c'(y) > c(y)/y$ for all $y \neq 0$. Notice, that the latter property does not pose additional restrictions on cost functions within the framework of $\mathcal{P}^1$.
- $G(\mu^{\mathrm{SS}}, c)$: Necessary condition is $c'' > 0$. In general it is hard to establish uniqueness if more than 2 players are involved.
- $G(\mu^{\mathrm{MP}}, c)$: even in 2 player games uniqueness is not guaranteed. For instance, uniqueness is guaranteed for cost functions $c(y) = y^\alpha$ only if $1 < \alpha \leq 3$. For $c(y) = y^4$ there are preference profiles in $\mathcal{L}^*$ such that multiple equilibria reside.

Decreasing Returns to Scale The above theorem basically shows that uniqueness of equilibrium in demand games related to $\mathcal{P}^1$ can be achieved for preferences in $\mathcal{L}^*$ if only costs are convex, i. e. the technology exhibits decreasing returns to scale. Starting point in the literature to characterize cost sharing rules in terms of their strategic properties is the seminal paper by Moulin and Shenker [91]. Their finding is that on $\mathcal{L}$ basically μ^{SR} is the only cost sharing rule of which the corresponding demand game passes the unique equilibrium test like in Theorem 12.

Call a smooth and strictly demand monotonic cost sharing rule μ *regular* if it is *anonymous*, so that the name of an agent does not have an impact on her cost share.

Theorem 13 (Moulin & Shenker [91]) *Let c be a strictly convex continuously differentiable cost function, and let μ be a regular cost sharing rule. The following statements are equivalent:*

- $\mu = \mu^{\mathrm{SR}}$,
- *for all profiles $(u_1, u_2, \ldots, u_n)$ of utilities in $\mathcal{L}$, $G(\mu, c)$ has at most one Nash-equilibrium,*
- *for all profiles $(u_1, u_2, \ldots, u_n)$ of utilities in $\mathcal{L}$ every Nash-equilibrium of $G(\mu, c)$ is also a strong equilibrium, i. e., no coalition can coordinate in order to improve the payoff for all its members.*

This theorem makes a strong case for the serial cost sharing rule, especially when one realizes that the serial equilibrium is the unique element surviving successive elimination of strictly dominated strategies. Then this equilibrium may naturally arise through evolutive or eductive behavior, it is a robust prediction of non-cooperative behavior. Recent experimental studies are in line with this theoretical support, see [22,108]. Proposition 1 in [141] shows how easy it is to construct preferences in $\mathcal{L}$ such that regular cost sharing rules other than μ^{SR} give rise to multiple equilibria in the corresponding demand game, even in two-agent cost sharing games.

Besides other fairness concepts in the distributive literature the most compelling is *envy-freeness*. An allocation passes the no envy test if no player prefers her own allocation less than that of other players. Formally, the definition is as follows.

No Envy Test Let x be a demand profile and y a vector of cost shares. Then the allocation $(x_i, y_i)_{i \in N}$ is *envy-free* if for all $i, j \in N$ it holds $u_i(x_i, y_i) \geq u_i(x_j, y_j)$.

It is easily seen that the allocations associated with the serial equilibria are all envy-free.

Increasing Returns to Scale As Theorem 12 already shows, uniqueness of equilibrium in demand games for all utility profiles in $\mathcal{L}^*$ is in general inconsistent with concave cost functions.

Theorem 14 (de Frutos [26]) *Let c be a strictly concave continuously differentiable cost function, and let μ be a regular cost sharing rule. The following statements are equivalent:*

- $\mu = \mu^{\mathrm{DSR}}$ *or* $\mu = \mu^{\mathrm{SR}}$.
- *For all utility profiles $u = (u_i)_{i \in N}$ in $\mathcal{L}$ the induced demand game $G(\mu, c)$ has at most one Nash equilibrium, or*

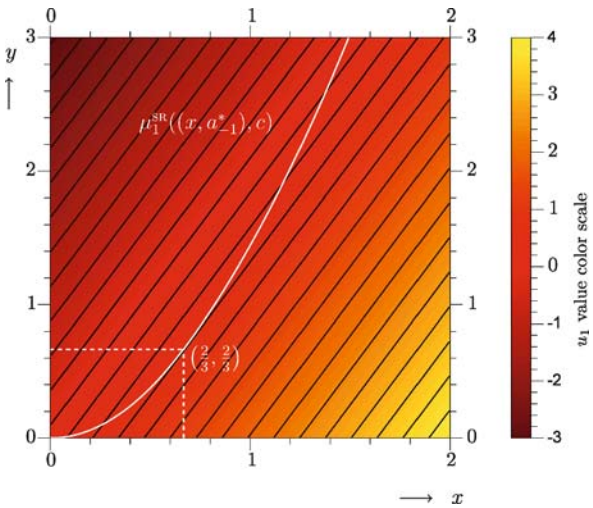

Cost Sharing, Figure 10
Scenario (ii). The indifference curves of agent 1 together with the curve $\kappa : t \mapsto \mu_1^{\mathrm{SR}}((t, \bar{x}_{-1}), c)$. Best response of player 1 against $\bar{x}_{-1} = (\frac{2}{3}, \frac{8}{3})$ is the value of x where the graph of κ is tangent to an indifference curve of u_1

- *For all utility profiles $u = (u_i)_{i \in N}$ in $\mathcal{L}$, every Nash equilibrium of the game $G(\mu, c)$ is a strong Nash equilibrium as well.*

Moreover, if the curvature of the indifference curves is bigger than that of the curve generated by the cost sharing rule as in Fig. 10 then the second and third statement are equivalent with $\mu = \mu^{\mathrm{DSR}}$.

Theorem 15 (Moulin [85]) *Assume agents have preferences in $\mathcal{L}^*$. The serial cost sharing rule is the unique continuous, cross-monotonic and anonymous cost sharing rule for which the Nash-equilibria of the corresponding demand games all pass the no-envy test.*

Comparison of Serial and Proportional Equilibria (II)

Just as in the earlier analysis in Sect. "Efficiency and Strategy-Proof Cost Sharing Mechanisms", performance of cost sharing rules can be measured by the related surpluses in the Nash-equilibria of the corresponding demand games. Assume in this section that the preferences of the agents are quasi-linear in cost shares, and represented by functions $U_i(x_i, y_i) = u_i(x_i) - y_i$. Moreover, assume that u_i is non-decreasing and concave, $u_i(0) = 0$. Then the surplus at the demand profile x and utility profile is the number $\sum_{i \in N} u_i(x_i) - c(x(N))$. Define the *efficient surplus* or *value* of N relative to c and U by

$$v(c, U) = \sup_{x \in \mathbb{R}_+^N} \sum_{i \in N} u_i(x_i) - c(x(N)) \,. \tag{27}$$

Denote the set of Nash-equilibria in the demand game $G(\mu, c)$ with profile of preferences U by $\text{NE}(\mu, c, U)$. Given c, μ the *guaranteed (relative) surplus* of the cost sharing rule μ for N is defined by

$$\gamma(c, \mu) = \inf_{U, x \in \text{NE}(\mu, c, U)} \frac{\sum_{i \in N} u_i(x_i) - c(x(N))}{v(c, U)} . \quad (28)$$

Here the infimum is taken over all utility profiles discussed above. This measure is also called the *price of anarchy* of the game, see [65]. Let C^* be the set of all convex increasing cost functions with $\lim_{y \to \infty} c(y)/y = \infty$. Then [89] shows that for the serial and the proportional rule the guaranteed surplus is at least $1/n$. But sometimes the distinction is eminent. Define the number

$$\delta(y) = \frac{yc''(y)}{c'(y) - c'(0)} ,$$

which is a kind of elasticity. The below theorem shows that on certain domains of cost functions with bounded δ the serial rule prevails over the proportional rule. For large n the guaranteed surplus at μ^{SR} is of order $1/\ln(n)$, that of μ^{AV} of order $1/n$. More precise, write $K_n = 1 + \frac{1}{3} + \cdots + \frac{1}{2n-1} \approx 1 + \frac{\ln n}{2}$, then:

Theorem 16 (Moulin [89]) *For any convex increasing cost function c with $\lim_{y \to \infty} c(y)/y = \infty$ it holds that*

- *If c' is concave and* $\inf \{\delta(y) | y \geq 0\} = p > 0$, *then*

$$\gamma(c, \mu^{\text{SR}}) \geq \frac{p}{K_n} \qquad \gamma(c, \mu^{\text{AV}}) \leq \frac{4}{n+3}$$

- *If c' is convex and* $\sup \{\delta(y) | y \geq 0\} = p < \infty$ *then*

$$\gamma(c, \mu^{\text{SR}}) \geq \frac{1}{2p-1} \frac{1}{K_n}$$
$$\frac{4}{n+3} \leq \gamma(c, \mu^{\text{AV}}) \leq \frac{4(2p-1)}{n}$$

A Word on Strategy-Proofness in $\mathcal{P}^1$

Recall the discussion on strategyproofness in Sect. "Strategyproofness". The serial demand game has a unique strong Nash equilibrium in case costs are convex and preferences are drawn from $\mathcal{L}$. Suppose the social planner aims at designing a mechanism to implement the outcomes associated with these equilibria. [91] show an efficient way to implement this *serial choice function* by an *indirect* mechanism. It is defined through a multistage game which mimics the way the serial Nash equilibria are calculated. It is easily seen that here each agent has a unique dominant strategy, in which demands result from optimization of the true preferences. Then this gives rise to a strategyproof mechanism.

Note that the same approach can not be used for the proportional rule. The strategic properties of the proportional demand game are weaker than that of the serial demand game in several aspects. First of all, it is not hard to find preference profiles in $\mathcal{L}$ leading to multiple equilibria. Whereas uniqueness of equilibrium can be repaired by restricting $\mathcal{L}$ to $\mathcal{L}^*$, the resulting equilibria are, in general, not strong (like the serial counterparts). In the proportional equilibria there is overproduction, see e. g. the example in Sect. "Proportional Demand Game" where a small uniform reduction of demands yields higher utility for all the players. Besides, a single-valued Nash equilibrium selection corresponds to a strategyproof mechanism *provided* the underlying domain of preferences is rich, and $\mathcal{L}^*$ is not. Though richness is not a necessary condition, the proportional rule is not consistent with a strategyproof demand game.

Bayesian $\mathcal{P}^1$-Demand Games

Recall that at the basis of a strategic game there is the assumption that each player *knows* all the ingredients of the game. However, as [56] argues, production cost and output quality may vary unpredictably as a consequence of the technology and worker quality. Besides that, changes in the available resources and demands will have not foreseen influences on individual preferences. On top of that, the players may have asymmetrical information regarding the nature of uncertainty. [56] study the continuous homogeneous cost sharing problem within the context of a *Bayesian demand game* [43], where these uncertainties are taking into account. The qualifications of the serial rule in the stochastic model are roughly the same as in the deterministic framework.

Continuous Heterogeneous Output Model, $\mathcal{P}^n$

The analysis of continuous cost sharing problems for multi-service facilities are far more complex than the single-output model. The literature discusses two different models, one where each agent i demands a different good, and one where agents may require mixed bundles of goods. As the reader will notice, the modeling and analysis of solutions differs in abstraction and complexity. In order to concentrate on the main ideas, here we will stick to the first model, where goods are identified with agents. This means that a demand profile is a vector $x \in \mathbb{R}^N_+$, where x_i denotes the demand of agent i for good i. From now we deal with technologies described by continuously differ-

entiable cost functions $c\colon \mathbb{R}^N_+ \to \mathbb{R}_+$, non-decreasing and $c(0_N) = 0$; the class of all such functions is denoted by C^n.

Extensions of Cost Sharing Rules The single-output model is connected to the multi-output model via the homogeneous cost sharing problems. Suppose that for $c \in C^n$ there is a function $c_0 \in C$ such that $c(x) = c_0(x(N))$ for all x. For instance, such functions are found if we distinguish between production of blue and red cars; the color of the car does not affect the total production costs. Essentially, a homogeneous cost sharing problem (x, c) may be solved as if it were in $\mathcal{P}^1$. If μ is the compelling solution on $\mathcal{P}^1$ then any cost sharing rule on $\mathcal{P}^n$ should determine the same solution on the class of *homogeneous* problems therein. Formally the cost sharing rule on $\mathcal{P}^n$ extends $\overline{\mu}$ on $\mathcal{P}^1$ if for all homogeneous cost sharing problems (x, c) it holds $\overline{\mu}(x, c) = \mu(x, c_0)$. In general a cost sharing rule μ on $\mathcal{P}^1$ allows for a whole class of extensions. Below we will focus on extensions of $\mu^{\mathrm{SR}}, \mu^{\mathrm{P}}, \mu^{\mathrm{SS}}$.

Measurement of Scale Around the world quantities of goods are measured by several standards. Length is expressed in inches or centimeters, volume in gallons or liters, weight in ounces to kilos. Here measurement conversion involves no more than multiplication with a fixed scalar. When such linear scale conversions do not have any effect on final cost shares, a cost sharing rule is called *scale invariant*. It is an *ordinal* rule if this invariance extends to essentially all transformations of scale. Scale invariance captures the important idea that the relative cost shares should not change, whether we are dividing 1 Euro or 1,000 Euros. Ordinality may be desirable, but for many purposes too strong as a basic requirement. Formally, a *transformation of scale* is a mapping $f\colon \mathbb{R}^N_+ \to \mathbb{R}^N_+$ such that $f(x) = (f_1(x_1), f_2(x_2), \ldots, f_n(x_n))$ for all x and each of the coordinate mappings f_j is differentiable and strictly increasing.

Ordinality A cost sharing rule μ on $\mathcal{P}^n$ is *ordinal* if for all transformations of scale f and all cost sharing problems $(x, c) \in \mathcal{P}^n$, it holds that

$$\mu(x, c) = \mu(f(x), c \circ f^{-1}) \,. \tag{29}$$

Scale Invariance A cost sharing rule μ on $\mathcal{P}^n$ is scale invariant if Eq. (29) holds for all linear transforms f. Under a scale invariant cost sharing rule final cost shares do not change by changing the units in which the goods are measured.

Path-Generated Cost Sharing Rules Many cost sharing rules on $\mathcal{P}^n$ calculate the cost shares for $(x, c) \in \mathcal{P}^n$ by the total of marginal costs along some *production path* from 0 toward x. Here a path for x is a non-decreasing mapping $\gamma^x\colon \mathbb{R}_+ \to \mathbb{R}^N_+$ such that $\gamma(0) = 0_N$ and there is a $T \in \mathbb{R}_+$ with $\gamma^x(T) = x$. The cost sharing rule *generated by the path collection* $\gamma = \{\gamma^x | x \in \mathbb{R}^N_+\}$ is defined by

$$\mu^\gamma_i(x, c) = \int_0^\infty \partial_i c(\gamma^x(t))(\gamma^x_i)'(t)\mathrm{d}t \,. \tag{30}$$

Special path-generated cost sharing rules are the *fixed-path cost sharing rules*; a single path $\gamma^*\colon \mathbb{R}_+ \to \mathbb{R}^N_+$ with the property that $\lim_{t\to\infty} \gamma^*_i(t) = \infty$ defines the whole underlying family of paths. More precisely, the *fixed path cost sharing rule* μ generated by γ^* is the path-generated rule for the family of paths $\{\gamma^x | x \in \mathbb{R}^N_+\}$ defined by $\gamma^x(t) = \gamma^*(t) \wedge x$, the vector with coordinates $\min\{\gamma^*_i(t), x_i\}$. So the paths are no more than the projections of $\gamma^*(t)$ on the cube $[0, x]$. Below we will see many examples of (combinations of) such fixed-path methods.

Aumann–Shapley Rule The early characterizations by [15,79] on this rule set off a vast growing literature on cost sharing models with variable demands. [16] suggested to use the Aumann–Shapley rule to determine telephone billing rates in the context of sharing the cost of a telephone system. This extension of proportional cost sharing calculates marginal costs along the path $\gamma^{\mathrm{AS}}(t) = tx$ for $t \in [0, 1]$. Then

$$\mu^{\mathrm{AS}}_i(x, c) = x_i \int_0^1 \partial_i c(tx)\mathrm{d}t \,. \tag{31}$$

The Aumann–Shapley rule can be interpreted as the Shapley-value of the *non-atomic* game where each unit of the good is a player, see [11]. It is the uniform average over marginal costs along all increasing paths from 0_N to x. The following is a classic result in the cost sharing literature:

Theorem 17 (Mirman & Tauman [79], Billera & Heath [15]) *There is only one additive, positive, and scale invariant cost sharing rule on $\mathcal{P}^n$ that extends the proportional rule, and this is μ^{AS}.*

Example 24 If c is positively homogeneous, i. e. $c(\alpha y) = \alpha c(y)$ for $\alpha \geq 0$ and all $y \in \mathbb{R}^N_+$, then

$$\mu^{\mathrm{AS}}_i(x, c) = \frac{\partial_i c}{\partial x_i}(x)$$

i. e. μ^{AS} calculates the marginal costs of the ith good at the final production level x. The risk measures (cost functions) as in [28] are of this kind.

Friedman–Moulin Rule This serial extension [36] calculates marginal costs along the diagonal path, i. e. $\gamma^{\mathrm{FM}}(t) = t\mathbf{1}_N \wedge x$

$$\mu_i^{\mathrm{FM}}(x, c) = \int_0^{x_i} \partial_i c(\gamma^x(t))\mathrm{d}t \,. \tag{32}$$

This fixed path cost sharing rule is demand monotonic. As far as invariance with the choice of unit is concerned, the performance is bad as it is not a scale invariant cost sharing rule.

Moulin–Shenker Rule This fixed-path cost sharing rule is proposed as an ordinal serial extension by [125]. Suppose that the partial derivatives of $c \in C^n$ are bounded away from 0, i. e. there is a such that $\partial_i c(x) > a$ for all $x \in \mathbb{R}_+^N$. The Moulin–Shenker rule μ^{MS} is generated by the path γ^{MS} as solution to the system of ordinary differential equations

$$\gamma_i'(t) = \begin{cases} \frac{\sum_{j\in N} \partial_j c(\gamma(t))}{\partial_i c(\gamma(t))} & \text{if } \gamma_i(t) < x_i \,, \\ 0 & \text{else} \,. \end{cases} \tag{33}$$

The interpretation of this path is that at each moment the total expenditure for production of extra units for the different agents is equal; if good 2 is twice as expensive as good 1, then the production device γ^{MS} will produce twice as much of the good 1. The serial rule embraces the same idea – as long as an agent desires extra production, the corresponding incremental costs are equally split. This makes μ^{MS} a natural extension of the serial rule. Call t_i^* the completion time of production for agent i, i. e. $\gamma_i^{\mathrm{MS}}(t) < x_i$ if $t < t_i^*$ and $\gamma_i^{\mathrm{MS}}(t_i^*) = x_i$. Assume without loss of generality that these completion times are ordered such that $0 = t_0^* \le t_1^* \le t_2^* \le \cdots \le t_n^*$, then the Moulin–Shenker rule is given by

$$\mu_i^{\mathrm{MS}}(x, c) = \sum_{\ell=1}^{i} \frac{c(\gamma^{\mathrm{MS}}(t_\ell^*)) - c(\gamma^{\mathrm{MS}}(t_{\ell-1}^*))}{n - \ell + 1} \,. \tag{34}$$

Note that the path varies with the cost function, and that this is the reason why μ^{MS} is a non-additive solution. Such solutions – though intuitive – are in general notoriously hard to analyze. There are two axiomatic characterizations of the Moulin–Shenker rule. The first is by [125], in terms of the *serial principle* and the technical condition that a cost sharing rule be a partial differentiable functions of the demands. The other characterization by [60] is more in line with the ordinal character of this *serial extension*.

Continuity A cost sharing rule μ on $\mathcal{P}^n$ is *continuous* if $q \mapsto \mu(q, c)$ is continuous on $\mathbb{R}_+^N$ for all c. Continuity is weaker than partial differentiability, as it requires stability of the solution with respect to small changes in the demands.

Upperbound A cost sharing rule μ satisfies *upperbound* if for all $(q, c) \in \mathcal{P}^n$, $i \in N$

$$\mu_i(q, c) \le \max_{y\in[\mathbf{0},q]} \partial_i c(y) \,.$$

An upperbound provides each agent with a conservative and rather pessimistic estimate of her cost share, based on the maximal value of the corresponding marginal cost toward the aggregate demand.

Suppose that d is a demand profile smaller than q. A reduced cost sharing problem is defined by $(q - d, c^d)$ where c^d is defined by $c^d(y) = c(y + d) - c(d)$. So c^d measures the incremental cost of production beyond the level d.

Self–consistency A cost sharing rule μ is *self-consistent* if for all cost sharing problems $(q, c) \in \mathcal{P}^n$ with $q_{N\setminus S} = \mathbf{0}_{N\setminus S}$ for some $S \subseteq N$, and $d \le q$ such that $\mu_i(d, c) = \mu_j(d, c)$ for all $\{i, j\} \subset S$, then $\mu(q, c)_S = \mu(d, c)_S + \mu(q - d, c^d)_S$.

So, self-consistency is expresses the idea that if cost shares of agents with non-zero demand differ, then this is not due to the part of the problem that they are equally charged for, but due to the asymmetries in the related reduced problem. The property is reminiscent of the *step-by-step negotiation property* in the bargaining literature, see [54].

Theorem 18 (Koster [60]) *There is only one continuous, self-consistent and scale invariant cost sharing rule satisfying upper bounds, which is the Moulin–Shenker rule.*

Shapley–Shubik Rule For each demand profile x the stand-alone cost game c_x is defined as before. Then the Shapley–Shubik rule is no more than the Shapley value of this game, i. e. $\mu^{\mathrm{SS}}(x, c) = \Phi(c_x)$. The Shapley–Shubik rule is ordinal.

A Numerical Example Consider the cost sharing problem (x, c) with $N = \{1, 2\}$, $x = (5, 10)$, and $c \in C^2$ is given by $c(t_1, t_2) = \mathrm{e}^{2t_1+t_2} - 1$ on $[0, 10] \times [0, 10]$. We calculate the partial derivatives $\partial_1 c(t_1, t_2) = 2\mathrm{e}^{2t_1+t_2} = 2\partial_2 c(t_1, t_2)$ for all $(t_1, t_2) \in \mathbb{R}_+^2$. The Aumann–Shapley path is given by $\gamma(t) = (5t, 10t)$ for $t \in [0, 1]$ and

$$\mu_1^{\mathrm{AS}}(x, c) = \int_0^1 \partial_1 c(5t, 10t) \cdot 5\mathrm{d}t$$

$$= \int_0^1 2e^{20t} \cdot 5dt = \tfrac{1}{2}\left(e^{20} - 1\right)$$

$$\mu_2^{AS}(x, c) = \int_0^1 \partial_2 c(5t, 10t) \cdot 10dt$$

$$= \int_0^1 e^{20t} \cdot 10dt = \tfrac{1}{2}\left(e^{20} - 1\right) .$$

The Friedman–Moulin rule uses the path

$$\gamma^{FM}(t) = t\mathbf{1}_N \wedge q = \begin{cases} (t, t) & \text{if } 0 \le t < 5 , \\ (5, 5 + t) & \text{if } 5 \le t , \end{cases}$$

and the corresponding cost shares are calculated as follows:

$$\mu_1^{FM}(x, c) = \int_0^5 \partial_1 c(t, t)dt$$

$$= \int_0^5 2e^{3t}dt = \tfrac{2}{3}\left(e^{15} - 1\right) ,$$

$$\mu_2^{FM}(x, c) = \int_0^5 \partial_2 c(t, t)dt + \int_5^{10} \partial_2 c(5, 5 + t)dt$$

$$= \tfrac{1}{3}\left(e^{15} - 1\right) + e^{20} - e^{15} .$$

Note that both discussed cost sharing rules use one and the same path for all cost sharing problems with demand profile x. This is characteristic for *additive* cost sharing rules (see e. g. [35,42]).

Now turn to the Moulin–Shenker rule. Since $\partial_1 c = 2\partial_2 c$ everywhere on $[0, 10] \times [0, 10]$, according the solution γ^{MS} of Eq. (33), until one of the demands is reached, for each produced unit of good 1 two units of good 2 are produced. In particular there is a parametrization $\overline{\gamma}$ of γ^{MS} such that $\overline{\gamma}(t) = (t, 2t)$ for $0 \le t \le 5$. The corresponding cost shares are equal since $\overline{\gamma}$ reaches both coordinates of x at the same time, so $\mu_1^{MS}(x, c) = \mu_2^{MS}(x, c) = \frac{1}{2}c(\overline{\gamma}(5)) = \frac{1}{2}c(5, 10) = \frac{1}{2}(e^{20} - 1)$. Now suppose that the demands are summarized by $x^* = (10, 10)$. In order to calculate $\mu^{MS}(x^*, c)$, notice that there is a parametrization of γ^* of the corresponding path γ^{MS} such that

$$\gamma^*(t) = \begin{cases} (t, 2t) & \text{if } t \le 5 , \\ (t, 10) & \text{for } 5 < t \le 10 , \end{cases}$$

Notice that this path extends $\overline{\gamma}$ just to complete service for agent 1, so that – like before – agent 2 only contributes while $t < 5$. Then the cost shares are given by

$$\mu_2^{MS}(x^*, c) = \tfrac{1}{2}c(\gamma^*(5)) = \tfrac{1}{2}c(5, 10) = \tfrac{1}{2}(e^{20} - 1) ,$$

$$\mu_1^{MS}(x^*, c) = \mu_2^{MS}(x^*, c) + c(\gamma^*(10)) - c(\gamma^*(5))$$

$$= e^{30} - \tfrac{1}{2}(e^{20} + 1) .$$

For x^* the cost sharing rules μ^{AS} and μ^{FM} use essentially the same symmetric path $\gamma(t) = (t, t)$, so that it is easily calculated that $\mu^{AS}(x^*, c) = \mu^{FM}(x^*, c) = (\frac{2}{3}\left(e^{30} - 1\right), \frac{1}{3}\left(e^{30} - 1\right)$.

Axiomatic Characterization of Fixed-Path Rules Recall demand monotonicity as a weak incentive constraint for cost sharing rules. Despite the attention that the Aumann–Shapley rule received, it fails to meet this standard. To see this consider the following

$$c(y) = \frac{y_1 y_2}{y_1 + y_2} , \quad \text{and} \quad \mu_1^{AS}(x, c) = \frac{x_1 x_2^2}{(x_1 + x_2)^2} .$$

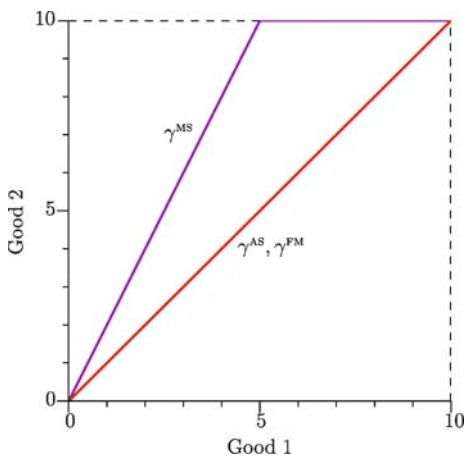

Cost Sharing, Figure 11
Paths for $\mu^{MS}, \mu^{AS}, \mu^{FM}$

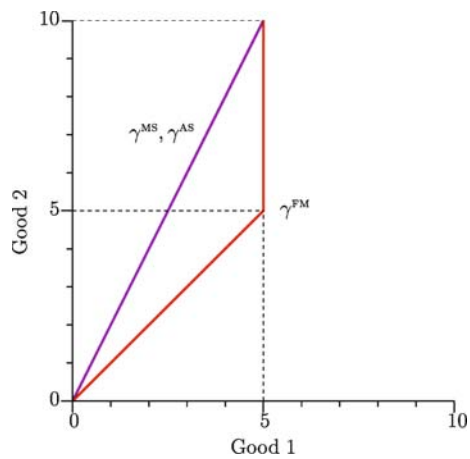

Cost Sharing, Figure 12
Paths for $\mu^{MS}, \mu^{AS}, \mu^{FM}$

Then the latter expression is not monotonic in x_1. One may show that the combination of properties in Theorem 17 are incompatible with demand monotonicity. Now what kind of rules are demand monotonic? The classification of all such rules is too complex. We will restrict our attention to the additive rules with the *dummy* property, which comprises the idea that a player pays nothing if her good is for free:

Dummy If $\partial_i c(y) = 0$ for all y then $\mu_i(x, c) = 0$ for all cost sharing problems $(x, c) \in \mathcal{P}^n$.

Theorem 19 (Friedman [35])

- *A cost sharing rule μ satisfies dummy, additivity and demand monotonicity if and only if it is an (infinite) convex combination of rules generated by fixed paths which do not depend on the cost structure.*
- *A cost sharing rule μ satisfies dummy, additivity and scale invariance if and only if it is an (infinite) convex combination of rules generated by scale invariant fixed paths which do not depend on the cost structure.*

This theorem has some important implications. The Friedman–Moulin rule is the unique serial extension with the properties additivity, dummy and demand monotonicity. As we mentioned before, μ^{FM} is not scale invariant. The only cost sharing rules satisfying all four of the above properties are *random order values*, i. e. a convex combination of marginal vectors of the stand-alone cost game [142]. μ^{SS} is the special element in this class of rules, giving equal weight to each marginal vector. Consider the following weak fairness property:

Equal Treatment Consider $(x, c) \in \mathcal{P}^n$. If $c_x(S \cup \{i\}) = c_x(S \cup \{j\})$ for all i, j and $S \subset N \setminus \{i, j\}$ then $\mu_i(x, c) = \mu_j(x, c)$.

Within the class of random order values, μ^{SS} is the unique cost sharing rule satisfying equal treatment.

Strategic Properties of Fixed-Path Rules [34] shows that the fixed-path cost sharing rules essentially have the same strategic properties as the serial rule. The crucial result in this respect is the following.

Theorem 20 (Friedman [34]) *Consider the demand game $G(\mu, c)$ where μ is a fixed path cost sharing rule and $c \in C^n$ has strictly increasing partial derivatives. Then the corresponding set $\mathcal{O}^\infty$ of action profiles surviving the successive elimination of overwhelmed actions consists of a unique element.*

As a corollary one may prove that the action profile in $\mathcal{O}^\infty$ is actually the unique Nash-equilibrium of the game $G(\mu, c)$, and that it is strong as well. Moreover, [34] shows that this Nash-equilibrium can be reached through some learning dynamics. Then this means that the demand games induced by μ^{FM} and μ^{MS} have strong strategic properties. Notice that the above theorem is only one-way. There are other cost sharing rules, like the *axial serial rule*, having the same strategic characteristics.

Future Directions

So far, a couple of standard stylized models have been launched providing a theoretical basis for defining and studying cost sharing principles at a basic level. The list of references below indicates that this field of research is full in swing, both in theoretical and applied directions. Although it is hard to make a guess where developments lead to, a couple of future directions will be highlighted.

Informational Issues

So far most of the literature is devoted to deterministic cost sharing problems. The cost sharing problems we face in practice are shaped by unsure events. Despite its relevance, this stochastic modeling in the literature is quite exceptional, see [56,138].

The presented models assume the information of costs for every contingent demand profile. Certainly within the continuous framework this seems too much to ask for. Retrieving the necessary information is hindered not only by technical constraints, but leads to new costs as well. [48] discusses data envelopment in cost sharing problems. A stochastic framework will be useful to study such *estimated* cost sharing problems. Other work focusing on informational coherence in cost sharing problems is [126]. Related work is [4], discussing mixtures of discrete and continuous cost sharing problems.

Budget Balance

In this overview, the proposed mechanisms are based on cost *sharing* rules. Another stream in implementation theory – at the other extreme of the spectrum – deals with cost *allocation* rules with no restrictions on the budget. [89] compares the size of budget deficits relative to the overall efficiency of a mechanism.

Performance

Recall the performance indices measuring the welfare impact of different cost sharing rules. [90] focuses on the continuous homogeneous production situations, with cost functions of specific types. There is still a need for a more

general theory. In particular this could prove to be indispensable for analyzing the quality of cost sharing rules in a broader set-up, the heterogeneous and Bayesian cost sharing problems.

Non-linear Cost Sharing Rules

Most of the axiomatic literature is devoted to the analysis of cost sharing rules as linear operators. The additivity property is usually motivated as an accounting convention, but it serves merely as a tool by which some mathematical representation theorems apply. Besides the practical motivation, it is void of any ethical content. As [88] underlines, there are hardly results on non-additive cost sharing rules – one of the reasons is that the mathematical analysis becomes notoriously hard. But – as a growing number of authors acknowledges – the usefulness of these mathematical techniques alone cannot justify the contribution of the property.

I thank Hervé Moulin as a referee of this article, and his useful suggestions.

Bibliography

1. Aadland D, Kolpin V (1998) Shared irrigation cost: an empirical and axiomatical analysis. Math Soc Sci 35:203–218
2. Aadland D, Kolpin V (2004) Environmental determinants of cost sharing. J Econ Behav Organ 53:495–511
3. Albizuri MJ, Zarzuelo JM (2007) The dual serial cost-sharing rule. Math Soc Sci 53:150–163
4. Albizuri MJ, Santos JC, Zarzuelo JM (2003) On the serial cost sharing rule. Int J Game Theory 31:437–446
5. An M (1998) Logconcavity versus logconvexity, a complete characterization. J Econ Theory 80:350–369
6. Archer A, Feigenbaum J, Krishnamurthy A, Sami R, Shenker S (2004) Approximation and collusion in multicast costsharing. Games Econ Behav 47:36–71
7. Arin J, Iñarra E (2001) Egalitarian solutions in the core. Int J Game Theory 30:187–193
8. Atkinson AB (1970) On the measurement of inequality. J Econ Theory 2:244–263
9. Aumann RJ (1959) Acceptable points in general cooperative *n*-person games. In: Kuhn HW, Tucker AW (eds) Contributions to the Theory of Games, vol IV. Princeton University Press, Princeton
10. Aumann RJ, Maschler M (1985) Game theoretic analysis of a bankruptcy problem from the Talmud. J Econ Theory 36:195–213
11. Aumann RJ, Shapley LS (1974) Values of Non-atomic Games. Princeton University Press, Princeton
12. Baumol W, Bradford D (1970) Optimal departure from marginal cost pricing. Am Econ Rev 60:265–283
13. Baumol W, Panzar J, Willig R (1988) Contestable Markets & the Theory of Industry Structure. 2nd edn Hartcourt Brace Jovanovich. San Diego
14. Bergantino A, Coppejans L (1997) A game theoretic approach to the allocation of joint costs in a maritime environment: A case study. Occasional Papers, vol 44. Department of Maritime Studies and International Transport, University of Wales, Cardi
15. Billera LJ, Heath DC (1982) Allocation of shared costs: A set of axioms yielding a unique procedure. Math Oper Res 7:32–39
16. Billera LJ, Heath DC, Raanan J (1978) Internal telephone billing rates: A novel application of non-atomic game theory. Oper Res 26:956–965
17. Bird CG (1976) On cost allocation for a spanning tree: A game theoretic approach. Networks 6:335–350
18. Binmore K (2007) Playing for real: A text on game theory. Oxford University Press, Oxford
19. Bjorndal E, Hamers H, Koster M (2004) Cost allocation in a bank ATM network. Math Methods Oper Res 59:405–418
20. Bondareva ON (1963) Some applications of linear programming to the theory of cooperative games. Problemy Kybernetiki 10:119–139 (in Russian)
21. Brânzei R, Ferrari G, Fragnelli V, Tijs S (2002) Two approaches to the problem of sharing delay costs in joint projects. Ann Oper Res 109:359–374
22. Chen Y (2003) An experimental study of serial and average cost pricing mechanisms. J Publ Econ 87:2305–2335
23. Clarke EH (1971) Multipart pricing of public goods. Publ Choice 11:17–33
24. Dasgupta PS, Hammond PJ, Maskin ES (1979) The implementation of social choice rules: Some general results on incentive compatibility. Rev Econ Stud 46:185–216
25. Davis M, Maschler M (1965) The kernel of a cooperative game. Nav Res Logist Q 12:223–259
26. de Frutos MA (1998) Decreasing serial cost sharing under economies of scale. J Econ Theory 79:245–275
27. Demers A, Keshav S, Shenker S (1990) Analysis and simulation of a fair queueing algorithm. J Internetworking 1:3–26
28. Denault M (2001) Coherent allocation of risk capital. J Risk 4_7–21
29. Dewan S, Mendelson H (1990) User delay costs and internal pricing for a service facility. Manag Sci 36:1502–1517
30. Dutta B, Ray D (1989) A concept of egalitarianism under participation constraints. Econometrica 57:615–635
31. Dutta B, Ray D (1991) Constrained egalitarian allocations. Games Econ Behav 3:403–422
32. Flam SD, Jourani A (2003) Strategic behavior and partial cost sharing. Games Econ Behav 43:44–56
33. Fleurbaey M, Sprumont Y (2006) Sharing the cost of a public good without subsidies. Université de Montréal, Cahier
34. Friedman E (2002) Strategic properties of heterogeneous serial cost sharing. Math Soc Sci 44:145–154
35. Friedman E (2004) Paths and consistency in additive cost sharing. Int J Game Theory 32:501–518
36. Friedman E, Moulin H (1999) Three methods to share joint costs or surplus. J Econ Theory 87:275–312
37. Friedman E, Shenker S (1998) Learning and implementation on the Internet. Working paper 199821, Rutgers University, Piscataway
38. González-Rodríguez P, Herrero C (2004) Optimal sharing of surgical costs in the presence of queues. Math Methods Oper Res 59:435–446
39. Granot D, Huberman G (1984) On the core and nucleolus of minimum cost spanning tree games. Math Program 29:323–347

40. Green J, Laffont JJ (1977) Characterization of satisfactory mechanisms for the revelation of preferences for public goods. Econometrica 45:427–438
41. Groves T (1973) Incentives in teams. Econometrica 41:617–663
42. Haimanko O (2000) Partially symmetric values. Math Oper Res 25:573–590
43. Harsanyi J (1967) Games with incomplete information played by Bayesian players. Manag Sci 14:159–182
44. Hart S, Mas-Colell A (1989) Potential, value, and consistency. Econometrica 57:589–614
45. Haviv M (2001) The Aumann–Shapley price mechanism for allocating congestion costs. Oper Res Lett 29:211–215
46. Hougaard JL, Thorlund-Petersen L (2000) The stand-alone test and decreasing serial cost sharing. Econ Theory 16:355–362
47. Hougaard JL, Thorlund-Petersen L (2001) Mixed serial cost sharing. Math Soc Sci 41:51–68
48. Hougaard JL, Tind J (2007) Cost allocation and convex data envelopment. mimeo University of Copenhagen, Copenhagen
49. Henriet D, Moulin H (1996) Traffic-based cost allocation in a network. J RAND Econ 27:332–345
50. Iñarra E, Isategui JM (1993) The Shapley value and average convex games. Int J Game Theory 22:13–29
51. Israelsen D (1980) Collectives, communes, and incentives. J Comp Econ 4:99–124
52. Jackson MO (2001) A crash course in implementation theory. Soc Choice Welf 18:655–708
53. Joskow PL (1976) Contributions of the theory of marginal cost pricing. Bell J Econ 7:197–206
54. Kalai E (1977) Proportional solutions to bargaining situations: Interpersonal utility comparisons. Econometrica 45:1623–1630
55. Kaminski M (2000) 'Hydrolic' rationing. Math Soc Sci 40:131–155
56. Kolpin V, Wilbur D (2005) Bayesian serial cost sharing. Math Soc Sci 49:201–220
57. Koster M (2002) Concave and convex serial cost sharing. In: Borm P, Peters H (eds) Chapters in game theory. Kluwer, Dordrecht
58. Koster M (2005) Sharing variable returns of cooperation. CeNDEF Working Paper 05–06, University of Amsterdam, Amsterdam
59. Koster M (2006) Heterogeneous cost sharing, the directional serial rule. Math Methods Oper Res 64:429–444
60. Koster M (2007) The Moulin–Shenker rule. Soc Choice Welf 29:271–293
61. Koster M (2007) Consistent cost sharing. mimeo, University of Amsterdam
62. Koster M, Tijs S, Borm P (1998) Serial cost sharing methods for multi-commodity situations. Math Soc Sci 36:229–242
63. Koster M, Molina E, Sprumont Y, Tijs ST (2002) Sharing the cost of a network: core and core allocations. Int J Game Theory 30:567–599
64. Koster M, Reijnierse H, Voorneveld M (2003) Voluntary contributions to multiple public projects. J Publ Econ Theory 5:25–50
65. Koutsoupias E, Papadimitriou C (1999) Worst-case equilibria. In: 16th Annual Symposium on Theoretical Aspects of Computer Science, Trier. Springer, Berlin, pp 404–413
66. Legros P (1986) Allocating joint costs by means of the Nucleolus. Int J Game Theory 15:109–119
67. Leroux J (2004) Strategy-proofness and efficiency are incompatible in production economies. Econ Lett 85:335–340
68. Leroux J (2006) Profit sharing in unique Nash equilibrium: characterization in the two-agent case, Games and Economic Behavior 62:558–572
69. Littlechild SC, Owen G (1973) A simple expression for the Shapley value in a simple case. Manag Sci 20:370–372
70. Littlechild SC, Thompson GF (1977) Aircraft landing fees: A game theory approach. Bell J Econ 8:186–204
71. Maniquet F, Sprumont Y (1999) Efficient strategy-proof allocation functions in linear production economies. Econ Theory 14:583–595
72. Maniquet F, Sprumont Y (2004) Fair production and allocation of an excludable nonrival good. Econometrica 72:627–640
73. Maschler M (1990) Consistency. In: Ichiishi T, Neyman A, Tauman Y (eds) Game theory and applications. Academic Press, New York, pp 183–186
74. Maschler M (1992) The bargaining set, kernel and nucleolus. In: Aumann RJ, Hart S (eds) Handbook of Game Theory with Economic Applications, vol I. North–Holland, Amsterdam
75. Maschler M, Reijnierse H, Potters J (1996) Monotonicity properties of the nucleolus of standard tree games. Report 9556, Department of Mathematics, University of Nijmegen, Nijmegen
76. Maskin E, Sjöström T (2002) Implementation theory. In: Arrow KJ, Sen AK, Suzumura K (eds) Handbook of Social Choice and Welfare, vol I. North–Holland, Amsterdam
77. Matsubayashi N, Umezawa M, Masuda Y, Nishino H (2005) Cost allocation problem arising in hub-spoke network systems. Europ J Oper Res 160:821–838
78. McLean RP, Pazgal A, Sharkey WW (2004) Potential, consistency, and cost allocation prices. Math Oper Res 29:602–623
79. Mirman L, Tauman Y (1982) Demand compatible equitable cost sharing prices. Math Oper Res 7:40–56
80. Monderer D, Shapley LS (1996) Potential games. Games Econ Behav 14:124–143
81. Moulin H (1987) Equal or proportional division of a surplus, and other methods. Int J Game Theory 16:161–186
82. Moulin H (1994) Serial cost-sharing of an excludable public good. Rev Econ Stud 61:305–325
83. Moulin H (1995) Cooperative microeconomics: A game–theoretic introduction. Prentice Hall, London
84. Moulin H (1995) On additive methods to share joint costs. Japan Econ Rev 46:303–332
85. Moulin H (1996) Cost sharing under increasing returns: A comparison of simple mechanisms. Games Econ Behav 13:225–251
86. Moulin H (1999) Incremental cost sharing: Characterization by coalition strategy-proofness. Soc Choice Welf 16:279–320
87. Moulin H (2000) Priority rules and other asymmetric rationing methods. Econometrica 68:643–684
88. Moulin H (2002) Axiomatic cost and surplus-sharing. In: Arrow KJ, Sen AK, Suzumura K (eds) Handbook of Social Choice and Welfare. Handbooks in Economics, vol 19. North–Holland, Amsterdam, pp 289–357
89. Moulin H (2006) Efficient cost sharing with cheap residual claimant. mimeo Rice University, Housten

90. Moulin H (2007) The price of anarchy of serial, average and incremental cost sharing. mimeo Rice University, Housten
91. Moulin H, Shenker S (1992) Serial cost sharing. Econometrica 60:1009–1037
92. Moulin H, Shenker S (1994) Average cost pricing versus serial cost sharing; an axiomatic comparison. J Econ Theory 64:178–201
93. Moulin H, Shenker S (2001) Strategy-proof sharing of submodular cost: budget balance versus efficiency. Econ Theory 18:511–533
94. Moulin H, Sprumont Y (2005) On demand responsiveness in additive cost sharing. J Econ Theory 125:1–35
95. Moulin H, Sprumont Y (2006) Responsibility and cross-subsidization in cost sharing. Games Econ Behav 55:152–188
96. Moulin H, Vohra R (2003) Characterization of additive cost sharing methods. Econ Lett 80:399–407
97. Moulin H, Watts A (1997) Two versions of the tragedy of the commons. Econ Des 2:399–421
98. Mutuswami S (2004) Strategy proof cost sharing of a binary good and the egalitarian solution. Math Soc Sci 48:271–280
99. Myerson RR (1980) Conference structures and fair allocation rules. Int J Game Theory 9:169–182
100. Myerson RR (1991) Game theory: Analysis of conflict. Harvard University Press, Cambridge
101. Nash JF (1950) Equilibrium points in *n*-person games. Proc Natl Acad Sci 36:48–49
102. O'Neill B (1982) A problem of rights arbitration from the Talmud. Math Soc Sci 2:345–371
103. Osborne MJ (2004) An introduction to game theory. Oxford University Press, New York
104. Osborne MJ, Rubinstein A (1994) A Course in Game Theory. MIT Press, Cambridge
105. Peleg B, Sudhölter P (2004) Introduction to the theory of cooperative games, Series C: Theory and Decision Library Series. Kluwer, Amsterdam
106. Pérez-Castrillo D, Wettstein D (2006) An ordinal Shapley value for economic environments. J Econ Theory 127:296–308
107. Potters J, Sudhölter P (1999) Airport problems and consistent allocation rules. Math Soc Sci 38:83–102
108. Razzolini L, Reksulak M, Dorsey R (2004) An experimental evaluation of the serial cost sharing rule. Working paper 0402, VCU School of Business, Dept. of Economics, Richmond
109. Ritzberger K (2002) Foundations of non-cooperative game theory. Oxford University Press, Oxford
110. Rosenthal RW (1973) A class of games possessing pure-strategy Nash equilibria. J Econ Theory 2:65–67
111. Roth AE (ed) (1988) The Shapley value, essays in honor of Lloyd S Shapley. Cambridge University Press, Cambridge, pp 307–319
112. Samet D, Tauman Y, Zang I (1984) An application of the Aumann–Shapley prices for cost allocation in transportation problems. Math Oper Res 9:25–42
113. Sánchez SF (1997) Balanced contributions axiom in the solution of cooperative games. Games Econ Behav 20:161–168
114. Sandsmark M (1999) Production games under uncertainty. Comput Econ 14:237–253
115. Schmeidler D (1969) The Nucleolus of a characteristic function game. SIAM J Appl Math 17:1163–1170
116. Shapley LS (1953) A value for n-person games. Ann Math Study, vol 28. Princeton University Press, Princeton, pp 307–317
117. Shapley LS (1967) On balanced sets and cores. Nav Res Logist Q 14:453–460
118. Shapley LS (1969) Utility comparison and the theory of games. In: Guilbaud GT (ed) La decision: Aggregation et dynamique des ordres de preference. Editions du Centre National de la Recherche Scientifique. Paris pp 251–263
119. Shapley LS (1971) Cores of convex games. Int J Game Theory 1:1–26
120. Sharkey W (1982) Suggestions for a game–theoretic approach to public utility pricing and cost allocation. Bell J Econ 13:57–68
121. Sharkey W (1995) Network models in economics. In: Ball MO et al (eds) Network routing. Handbook in Operations Research and Management Science, vol 8. North–Holland, Amsterdam
122. Shubik M (1962) Incentives, decentralized control, the assignment of joint cost, and internal pricing. Manag Sci 8:325–343
123. Skorin-Kapov D (2001) On cost allocation in hub-like networks. Ann Oper Res 106:63–78
124. Skorin-Kapov D, Skorin–Kapov J (2005) Threshold based discounting network: The cost allocation provided by the nucleolus. Eur J Oper Res 166:154–159
125. Sprumont Y (1998) Ordinal cost sharing. J Econ Theory 81:126–162
126. Sprumont Y (2000) Coherent cost sharing. Games Econ Behav 33:126–144
127. Sprumont Y (2005) On the discrete version of the Aumann–Shapley cost-sharing method. Econometrica 73:1693–1712
128. Sprumont Y, Ambec S (2002) Sharing a river. J Econ Theory 107:453–462
129. Sprumont Y, Moulin H (2005) Fair allocation of production externalities: Recent results. CIREQ working paper 28-2005, Montreal
130. Sudhölter P (1998) Axiomatizations of game theoretical solutions for one-output cost sharing problems. Games Econ Behav 24:42–71
131. Suijs J, Borm P, Hamers H, Koster M, Quant M (2005) Communication and cooperation in public network situations. Ann Oper Res 137:117–140
132. Tauman Y (1988) The Aumann–Shapley prices: A survey. In: Roth A (ed) The shapley value. Cambridge University Press, Cambridge, pp 279–304
133. Thomas LC (1992) Dividing credit-card costs fairly. IMA J Math Appl Bus & Ind 4:19–33
134. Thomson W (1996) Consistent allocation rules. Mimeo, Economics Department, University of Rochester, Rochester
135. Thomson W (2001) On the axiomatic method and its recent applications to game theory and resource allocation. Soc Choice Welf 18:327–386
136. Tijs SH, Driessen TSH (1986) Game theory and cost allocation problems. Manag Sci 32:1015–1028
137. Tijs SH, Koster M (1998) General aggregation of demand and cost sharing methods. Ann Oper Res 84:137–164
138. Timmer J, Borm P, Tijs S (2003) On three Shapley-like solutions for cooperative games with random payoffs. Int J Game Theory 32:595–613
139. van de Nouweland A, Tijs SH (1995) Cores and related solution concepts for multi-choice games. Math Methods Oper Res 41:289–311

140. von Neumann J, Morgenstern O (1944) Theory of games and economic behavior. Princeton University Press, Princeton
141. Watts A (2002) Uniqueness of equilibrium in cost sharing games. J Math Econ 37:47–70
142. Weber RJ (1988) Probabilistic values for games. In: Roth AE (ed) The Shapley value. Cambridge University Press, Cambridge
143. Young HP (1985) Producer incentives in cost allocation. Econometrica 53:757–765
144. Young HP (1985) Monotonic solutions of cooperative games. Int J Game Theory 14:65–72
145. Young HP (1985) Cost allocation: Methods, principles, applications. North-Holland, Amsterdam
146. Young HP (1988) Distributive justice in taxation. J Econ Theory 44:321–335
147. Young HP (1994) Cost allocation. In: Aumann RJ, Hart S (eds) Handbook of Game Theory, vol II. Elsevier, Amsterdam, pp 1193–1235
148. Young HP (1998) Cost allocation, demand revelation, and core implementation. Math Soc Sci 36:213–229

Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of

Rowena Lohman
Cornell University, Ithaca, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Highlights of Earthquake Geodesy
Modeling of Geodetic Observations
Future Directions
Bibliography

Glossary

Aseismic Occurring without detectable radiated seismic energy.

Cascadia The region of the Pacific Northwest dominated by the Cascade Range and affected by subduction of the Juan de Fuca plate beneath North America.

Coseismic Occurring during an earthquake.

Elastic A form of behavior of a solid when subjected to stress. Elastic solids deform in response to stress by an amount proportional to a constant known as the "rigidity". When any applied stress is removed, an elastic solid recovers its original shape.

Geodesy The study of the shape and area of the Earth, including large-scale variations that affect the rotation dynamics of the planet of the whole, down to smaller length scales of earthquakes, landslides, etc.

Forward model A description of what a model of some process would predict about behavior of the system, e. g., how a given distribution of subsurface slip on a fault during an earthquake should affect observations of ground deformation at the surface.

GPS Global Positioning System. A network of satellites that transmit a signal that can be used by receivers (small transportable and/or permanent affixed to the ground) to infer three-dimensional positions.

InSAR Interferometric Synthetic Aperture Radar. The combination of Synthetic Aperture Radar imagery (generally acquired from airborne or satellite-based platforms) to infer changes in ground deformation, digital elevation models, variations in atmospheric water vapor, etc.

Interseismic deformation Occurs in the time period between earthquakes, usually associated with gradual increase in elastic stress to be released in future earthquakes.

Inverse theory The approach to determining the values for parameters of a given physical model that best describe observations of the system of interest.

Leveling The field of geodesy involved in the determination of variations in angle from horizontal between nearby fixed points on the Earth's surface, usually converted to changes in elevations.

Locked zone The portion of the fault zone that does not slip during the interseismic period, therefore accumulating stress and eventually rupturing coseismically.

Paleoseismology The study of individual earthquakes that occurred in the past, usually before the advent of instrumental recordings of seismic events.

Plate tectonics The theory governing how discrete plates on the Earth's surface move relative to each other over geologic time.

Postseismic deformation Deformation occurring in the hours to years following an earthquake.

Seismic cycle The combination of strain build-up and release that occurs on plate margins and along faults within plates, accommodated by processes within the coseismic, postseismic and interseismic time scales.

Seismogenic The region of a fault zone that is capable of producing earthquakes. Also refers to effects caused by an earthquake.

Subduction The process by which one tectonic plate descends beneath another, usually accompanied by volcanism and seismicity.

Triangulation The field of geodesy related to measuring horizontal angles and changes in angles between networks of fixed points.

Trilateration The field of geodesy related to measuring distance and changes in distance between networks of fixed points.

Viscoelastic A material behavior that is a combination of viscous and elastic behavior, resulting in some permanent deformation when the material is subjected to changes in stress.

Viscosity A material property describing its ability to flow in response to an applied stress. A measure of the response of a material to a stress, resulting in permanent deformation. The deformation rate of a viscous material depends on both the viscosity and applied stress.

Definition of the Subject

The seismic cycle consists of the processes associated with the accumulation and release of stress on seismogenic faults. The cycle is commonly divided into 3 periods: the coseismic interval for events occurring during an earthquake, the postseismic interval after an earthquake, and the interseismic period in between earthquakes (Fig. 1). Some of this deformation during these different periods is related directly to the motion on the fault during the earthquake – the ground is translated in one direction or another, there is crushing of rock, rotation or heaving of blocks of earth, and landslides triggered by the shaking. There is also ground deformation that results because of secondary effects – movement of ground water within the crushed and strained rock, cascades of earthquakes triggered by stress changes during the first event, continued slip on the fault interface, as well as flow of deeper, more ductile layers of the crust and mantle in response to the changes in stress. By examining these different behaviors and studying crustal deformation we can learn more about the underlying cause.

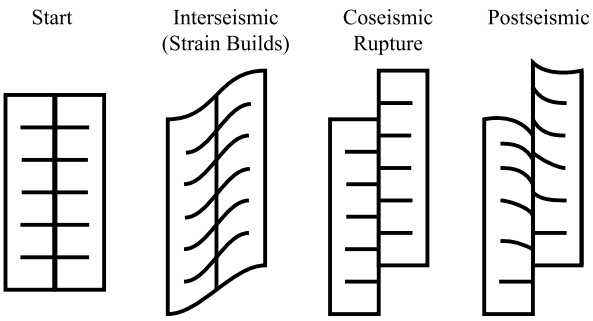

Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 1
Cartoon illustrating map view (view from overhead) of ground deformation during the seismic cycle for a strike-slip fault (after [73]). Similar models exist for other fault types

Earthquakes and other motion along fault zones are some of the ways that the Earth's crust accommodates far-field forcings due to plate tectonic motions. One of the prime features of interest as we study the seismic cycle is the magnitude of motion along these fault zones at different temporal and length scales. However, the only place where we can directly observe this motion is within the shallowest (< several meters) parts of the fault zone, through maps of features that are offset across the fault. The field of paleoseismology involves the search for information about previous earthquakes on a fault zone, often through trenches dug across the fault and the use of radiogenic dating to determine how frequently earthquakes have occurred in the past.

In order to infer what is occurring throughout the whole seismogenic zone (often the upper ~15 km within the continental crust), we rely on an arsenal of tools that include seismology (the study of how seismic waves travel through the earth) and geodesy (the study of changes to the shape of the earth's surface). In this chapter, we will explore how we can draw conclusions about fault zone slip at depths far greater than are directly accessible to us, based on how the earth's surface deforms during, before and after earthquakes.

Introduction

Ground shaking due to earthquakes can be felt over most of the Earths' surface, both on land and underwater. Earthquakes are also associated with volcanic activity, landslides and tsunami – they are often concentrated near discrete tectonic plate boundaries but also occupy diffuse zones where the Earth's crust is deforming [12,57]. The seismic energy that is released during an earthquake passes through deep portions of the Earth and helps us to learn about material properties we could not illuminate any other way, including details of the structure of the Earth's crust, core and everything in between. However, the destructive cost of large earthquake requires that the primary goal of earthquake research is that of determining when and where damaging earthquakes will occur [99]. To achieve this goal, we first have to understand what happens during each earthquake, i. e., where it was located, how big it was, how it's location relates to the distribution of previous earthquakes, etc.

One of the most common methods that we use to study earthquakes relies on the feature most apparent to humans – the rapid and often destructive movements of

the ground that occur during the earthquake. Seismology studies how the ground shakes during earthquakes and how we can use that information to better understand seismogenesis. However, this ground shaking is also always accompanied by some amount of permanent deformation of the ground. After all, the primary driving force behind most earthquakes is the slow motion of tectonic plates relative to each other. In this chapter we will not cover seismology, although combinations of seismic data and observations of geodetic displacements are often very powerful tools. Instead, we cover the geodetic observations of ground deformation during earthquakes and some of the methods that we use to interpret this deformation in terms of what actually happened on and around the fault zone.

Highlights of Earthquake Geodesy

Geodesy is defined as the branch of mathematics concerned with the shape and area of the Earth. In this chapter we will examine how the use of geodesy to quantify changes in the shape of the Earth can help us learn about earthquakes. We begin with a brief history of how the use of geodesy to study earthquakes has evolved, along with a description of some of the more interesting individual earthquakes and other seismic cycle behaviors that have been observed geodetically. We then explain the mechanics of how we use these measurements of surface deformation to understand processes deep beneath the Earth's surface.

Observations of Coseismic Displacements

The history of earthquake geodesy begins even before we really understood what earthquakes were. The key observations began when scientists began to associate earthquakes with observable deformation of the terrain. In Charles Lyell's Principles of Geology [46], he noted that earthquakes often accompany abrupt changes in the ground surface. The 1819 Ran of Cutch, India [64] and 1855 Wairarapa, New Zealand earthquakes were some of the first events where the accompanying ground deformation was observed. In 1835, during Charles Darwin's voyage on the *H.M.S. Beagle*, he experienced a large earthquake near Concepcion, Chile. During his reconnaissance of the area, he noted that the coastline had risen several meters in areas, exposing barnacles that had previously been underwater. He also found fossils hundreds of meters above sea level, indicating that numerous earthquakes had raised the cliffs over many millennia.

Several large earthquakes in the second half of the 19th century helped to advance the theory that earthquakes were caused by motion on faults. The 1872 Owens Valley, California, earthquake [19], the 1888 Amuri/Marlborough, New Zealand, earthquake [50], the 1891 Nobi, Japan, earthquake [36] and the 1893 Baluchistan earthquake [21] were each accompanied by visible deformation of the ground surface. These observations, and the fact that much of the deformation was consistent with regional, long-term topographic relief, helped counter arguments that earthquakes were primarily caused by volcanic activity [48].

Around the same time interval, the practice of surveying was coming into more widespread use, notably in India where Britain was involved in mapping the areas under its control. Surveyors installed permanent "monuments" or benchmarks, which they could then return to and re-survey at a later date. The May 17, 1892 Tapanuli, Sumatra, earthquake occurred during a triangulation survey by JJA Muller [58]. They noted changes in their surveyed angles between benchmarks that were consistent with two meters of deformation on a branch of the great Sumatran fault. In the foothills of the Himalaya, triangulation/leveling surveys measured deformation due to the 1897 Assam earthquake [64] and the 1905 Kangra earthquake [53]. In Italy, the 1915 Avezzano earthquake [94] was also spanned by early geodetic surveys.

In North America, the first great earthquake to be surveyed was the 1906 earthquake that destroyed most of the city of San Francisco, CA [40]. HF Reid [72] used three sets of triangulation surveys across San Andreas Fault to show that there had been approximately 3.2 meters of slip across the fault. During the 1920s and 1930s, there was a great deal of leveling work done in Japan which captured deformation associated with the 1927 Tango earthquake, and the 1944 and 1946 earthquakes associated with oceanic plate subduction in southwestern Japan [63,93]. Tide gauges along the coasts have also proved useful in earthquake studies, especially near subduction zones.

There was a boom in the use of the three main ground-based geodetic techniques (triangulation, trilateration and leveling) in the middle of the twentieth century, with observations of the 1940 Imperial Valley and 1962 Tehachapi earthquakes in California and the first observation of shallow interseismic creep on faults in the San Francisco Bay Area and in the Salton Trough/Imperial Valley region in Southern California. Next, the development of space-based geodesy began to allow for more precise and spatially extensive surveys of areas before and after earthquakes. Very Long Baseline Interferometry (VLBI) stations scattered around the world placed strong constraints on the relative motion of individual tectonic plates, as did the widespread use of Global Positioning System (GPS) [22,83].

GPS can be used either in continuous mode at permanent stations, which allows for high precision observations that can catch temporal changes in deformation, or in survey/campaign mode, where researchers take GPS receivers out to fixed benchmark stations in the field (sometimes previously studied using earlier surveying methods). Some of the first large earthquakes to be studied using GPS were the Ms 6.6 Superstition Hills, CA, earthquake in 1987 [39] and the 1989 Mw 7.1 Loma Prieta, CA earthquake [3]. The precise locations and descriptions of ground deformation allowed researchers to begin to examine how individual earthquakes fit in with the long-term topography, including the estimation of possible recurrence intervals [35,41].

InSAR The advent of Interferometric Synthetic Aperture Radar (InSAR), and it's application to the 1992 Landers, California, earthquake [17,49], ushered in a new era of geodesy with its unparalleled spatial density of deformation observations [23,76]. Synthetic Aperture Radar (SAR) images are acquired from a variety of platforms that send out radar signals, including satellites and airplanes (Fig. 2), and contain information about the phase and amplitude of the reflected return from the earth's surface.

A SAR interferogram is the difference in phase between two SAR images, and reflects a variety of factors that change the path length between the satellite and the ground. After correcting for some of these factors, such as topography (Fig. 3), the resulting interferometric phase is a function of any ground deformation that occurred between the two SAR acquisitions, in addition to variations in atmospheric water vapor, the ionosphere, etc. [11,20,100]. The high spatial resolution of InSAR (pixels commonly 5 m × 20 m or smaller) combined with its large areal coverage (∼100 km) produces images with so many pixels that they can become unwieldy to deal with computationally, especially when multiple interferograms are studied at once. Most researchers use various downsampling methods to reduce the number of data points considered without significantly reducing the amount of information retained [43].

The interferometric phase is only sensitive to ground deformation towards or away from the satellite, in a direction known as the satellite line of sight (LOS). Also, since the interferometric phase can only be measured as a fraction of the radar wavelength, an interferogram does not immediately give us an absolute measurement of the magnitude of ground deformation (Fig. 4). We need to add up the interferometric "fringes" to solve for the total deformation field. This "unwrapped" deformation field, along with information about the line-of-sight direction, can be combined in attempts to model the earthquake source parameters, or for models of deformation throughout other parts of the seismic cycle.

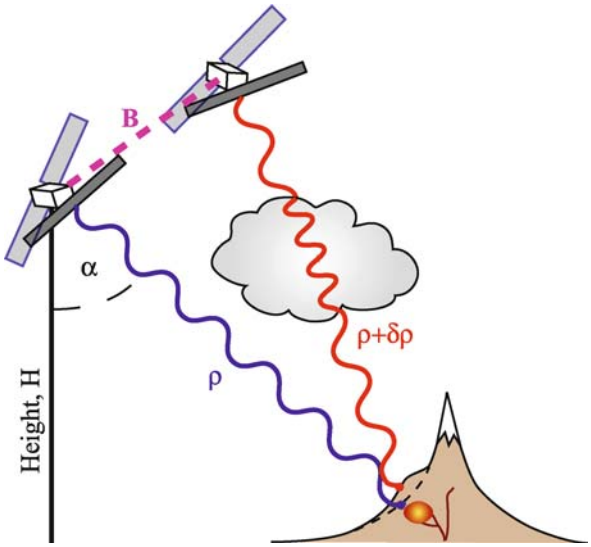

Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 2
Formation of a Synthetic Aperture Radar (SAR) interferogram: Two satellite radar images of the ground surface, separated by a spatial baseline, *B*, are combined to solve for the $\delta\rho$, the difference in line length (ρ) between the satellite and the ground along the satellite viewing direction (α). Knowledge of the satellite orbital path and the relative distance between the two image acquisitions (*B*) is necessary to convert $\delta\rho$ to elevation. If the two images are obtained simultaneously, or over a small time interval, the interferogram will only reflect topographic relief throughout the imaged area. However, if changes in ground surface elevation (e. g., the volcanic inflation shown in this figure between the *blue* and *red*), or changes in the atmosphere/ionosphere (e. g., water vapor content) occur between the two image acquisitions, this will also be reflected in the final interferogram

InSAR observations have improved the spatial complexity of these observations, allowing over a dozen large earthquakes to be studied in detail by a wide range of observers. Some noteworthy events that had various portions of their coseismic-postseismic activity covered by combinations of InSAR and GPS include the 1992 Landers, CA, earthquake [13,17,49], the 1995 Antofagasta, Chile, earthquake [70,77], the 1997 Hector Mine, CA earthquake [85], the 1999 Chi-Chi, Taiwan earthquake [8,98], the 2003 Bam, Iran earthquake [18], and the 2005 Nias–Simeulue earthquake [28]. In other cases, InSAR can help to determine the location of small- to moderate-sized earthquakes in areas with little other geophysical or field-based information [44].

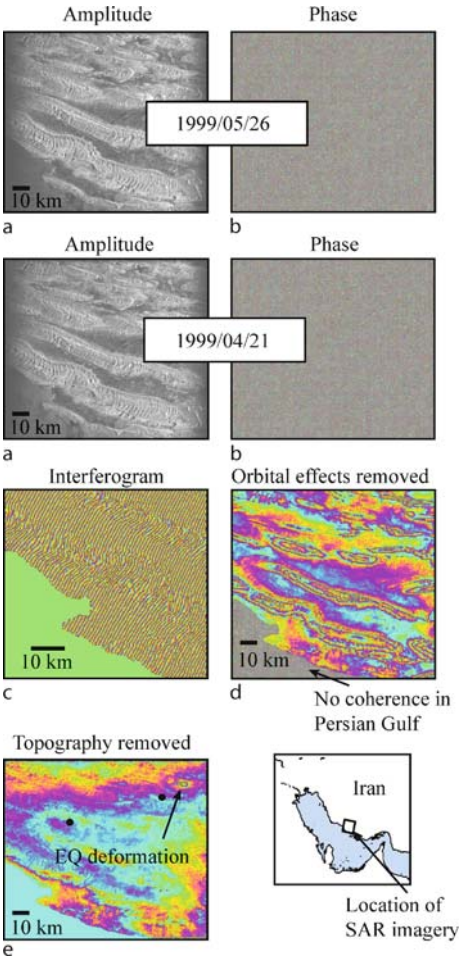

Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 3

Overview of steps in forming and interpreting an interferogram. An interferogram requires the combination of two sets of amplitude a and phase b observations from two separate image acquisitions (here, on two different dates separated by a month over an area along the Persian Gulf, Iran). While phase in the individual SAR images appear to be random noise, when the two images are combined, the phase changes vary coherently to form the interferogram in c. If we remove the effects of the satellite orbital geometry (the "curved earth effect"), we are left with d, which reflects both topography, ground surface deformation and any atmospheric changes between the two image acquisitions. The area in the *lower left* still looks like white noise due to the fact that water (here, the Persian Gulf) changes its reflectivity over short time scales and the phase does not remain coherent. If we remove the effects of topography e we are left with a map that clearly shows the effects of a small Mw 5 earthquake that occurred during this time interval (*black dots* show the seismically-determined locations for this event). Other features are due primarily to changes in atmospheric water vapor

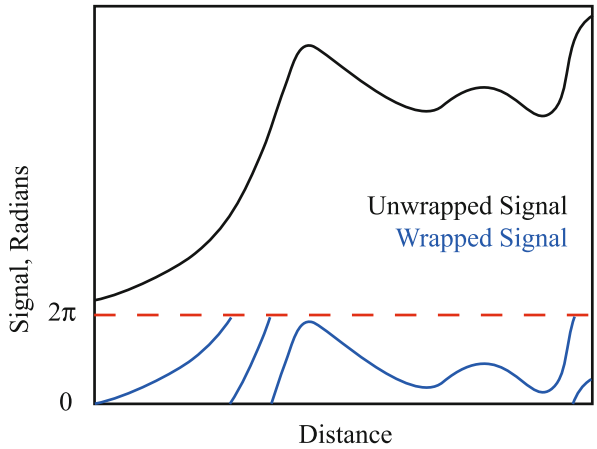

Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 4

Illustration of how the inherent 2π ambiguity affects the observed deformation. The "true" ground deformation vs. distance profile (*black* signal) would appear as the segmented *blue curve* ("*wrapped*" signal) when viewed using InSAR, since interferometry only retains information about the relative phase within a 2π cycle, not about the absolute phase, or number of cycles between the satellite and the ground. Reconstructing the original deformation field requires a process known as phase unwrapping, and there will also be an ambiguity as to the absolute value of the deformation field as a whole

Observations Throughout the Seismic Cycle

Early Models As the body of information about earthquakes grew, scientists were able to begin hypothesizing about the processes controlling their occurrence. Reid's observations of the 1906 San Francisco earthquake led him to propose his elastic rebound theory [73], where he hypothesized that the crust behaves like an elastic solid driven from the far field at a constant rate, which ruptures in earthquakes at periodic intervals to allow the two blocks to slide past each other (Fig. 1). In the interseismic period, the ground deforms smoothly in a manner that depends on the relative plate velocities, the thickness of the elastic plate, the elastic plate rigidity, etc. The size of the largest potential earthquake on a fault would depend on the length and depth of the elastic zone, and the timing until the next earthquake would depend on how much strain had built up since the last one.

This simple model, where each earthquake releases all of the built-up stress along the fault zone at a regular interval, does not seem to hold in the real world. On most faults, the magnitude of fault slip and rupture area appears to vary significantly between earthquakes (Fig. 5). Additionally, large amounts of accelerated deformation and other types of postseismic behavior are observed to

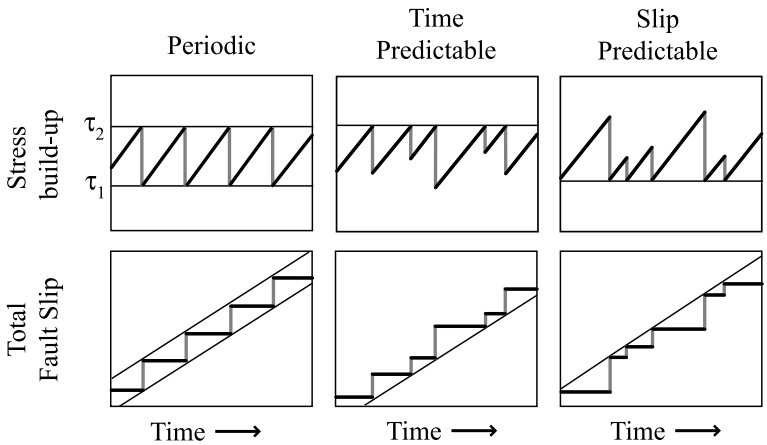

Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 5
Cartoon illustrating theories governing the temporal distribution of large earthquakes [88], each assuming a constant increase in stress vs. time in the interseismic interval (*black lines*), and coseismic behavior (*gray*) that depends on the accumulated stress in various ways. In the periodic model [73], earthquakes occur at constant intervals and always have the same magnitude, releasing the same amount of built-up stress. In the time-predictable model, earthquakes always occur when a maximum stress, τ_2, is reached, but the slip in each event varies. The time of the next earthquake depends on how much stress was released in the previous event. In the slip-predictable model, earthquakes occur at varying times, but always release the amount of built-up stress down to the minimum level, τ_1. The amount of slip, therefore, depends on the time since the previous earthquake

occur immediately after earthquakes, indicating that stress release is not accommodated in a simple manner. Observations such as these can help us improve and expand on the models that we utilize to explain earthquake occurrence.

Postseismic Behavior Earthquakes may primarily release built-up strain due to plate motions, but they also produce stress increases in both the near-field within the crust and within the underlying mantle. Observations of postseismic deformation, which is driven by the preceding coseismic stress changes [25,79,92], span a wide range of behaviors that may be explained by equally large range of constitutive properties. For some shallow strike-slip earthquakes, the observed postseismic deformation is as large as the fault slip during the earthquake [38,86].

Observed postseismic behaviors include poroelastic deformation [13,33], where fluid flow following gradients in coseismic stress changes results in ground deformation, frictional afterslip [5,28,55,81,82] and viscoelastic relaxation of the lower crust [25,26,68,74,79]. Afterslip is sometimes triggered not only on the fault that caused the earthquake, but also on surrounding faults within the stress field [1,12]. Another very noticeable consequence of an earthquake is the series of aftershocks following it, which may be triggered by combinations of stress changes as the lower crust relaxes after an earthquake as well as those due to motion during the earthquake itself [90].

In cases where the coseismic slip distribution (and it's associated stress change) is well-constrained by combinations of geodetic and seismic data, we can explore models of crustal and mantle properties, or place bounds on laboratory-derived rock mechanics laws [7], in order to fit the postseismic response to the coseismic stress change. For instance, the distribution of afterslip places constraints on the frictional behavior of the fault zone [32,55]. However, any such study requires good understand of the processes occurring during the earthquake itself.

Interseismic Behavior Although it is usually less dramatic than deformation occurring during and immediately after earthquakes, interseismic deformation can also tell us a great deal about the fault zone. The depth of the "locked" or "coupled" zone that will eventually rupture seismically [24,29], the rate at which stress is accumulating along the fault zone [59,92], and even variations in crustal elastic properties when rocks of different types are brought into contact across the fault zone [14], can all be addressed by examination of interseismic deformation across a fault. In addition, data types that span a finite amount of time, such as InSAR or campaign GPS observations, will always contain some amount of interseismic strain that may need to be removed before studying measurements spanning an earthquake. The steadiest interseismic deformation is mainly due to flow of the lower crust and mantle beneath the elastic upper crust, but there are also observations of

steady creep in the shallowest portions of several strike-slip fault zones [31,37,47,73].

Transient Behavior One of the most intriguing fault behaviors observed recently are isolated deformation events that are often not directly associated with an earthquake at all. These "slow" or "silent" earthquakes have been observed around the world in subduction zones, especially within the Cascadia subduction zone ([51,54], Fig. 6), Japan [27,56,65], Mexico [45], and Chile [64]. Along-strike variations in the amount of plate convergence that is taken up by these aseismic slip processes vs. large coseismic events may persist over very long timescales, as is indicated by the correlation of potentially aseismic "seismic gaps" in subduction zones and gravity anomalies [89,95]. Although most transient fault zone activity has been noticed along subduction zones, there are also observations of transient aseismic creep along the San Andreas Fault [60] and within the Imperial Valley, CA [42]. Further observations will help us understand how great of a contribution combinations of steady interseismic and transient aseismic behaviors make to the release of plate motion across many of the major tectonic boundaries around the world.

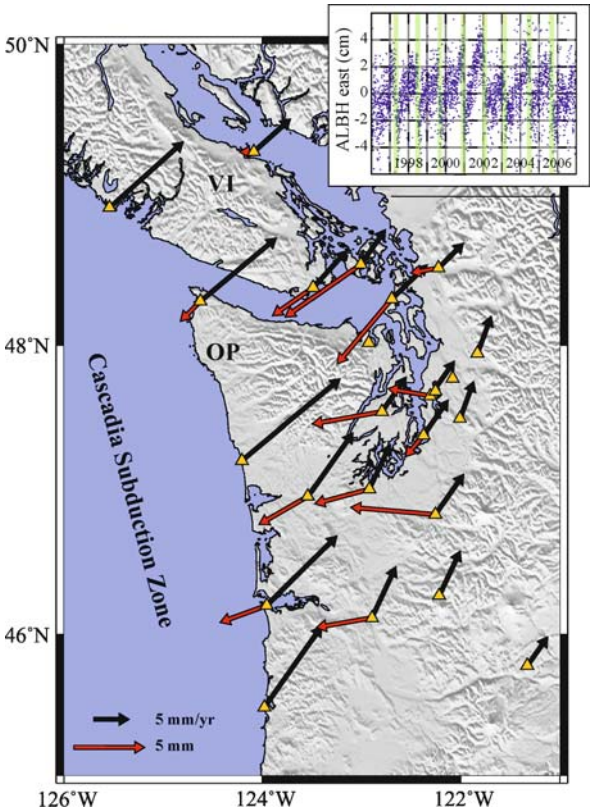

Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 6

Episodic slow slip events in Cascadia. OP=Olympic Peninsula, VI=Vancouver Island. *Black arrows* indicate interseismic convergence, *red arrows* indicate motion during one of the slow slip events recorded by continuous GPS (*yellow triangles*). *Inset* shows detrended component of deformation recorded by stations ALBH on Vancouver Island, with more than a decade of regular slow slip events (*vertical green bars*). Figure after [54]

Modeling of Geodetic Observations

Overview of Inverse Theory

All of the observations discussed above involve measurements of surface deformation. We can learn nothing about processes occurring below the surface, within fault zones, without the ability to infer how those observations were generated. Inverse theory spans the range of problems relating to identifying which model of real-world behavior is consistent with observed data. Some of the simplest inverse problems include that of fitting a line to a series of data points, or even calculating the mean value of a set of observations. Good treatments of the field of inverse theory and it's applications to geophysical problems in general can be found in books by Parker [66] and Menke [52], as well as in key papers by Backus and Gilbert [4] and Tarantola and Valette [91].

Before a geophysical inversion can be performed (i. e., determining the size of an earthquake based on ground deformation) the "forward model" must be defined and understood. A forward model is a description of how we believe the system of interest, here, the Earth's crust, will behave in response to the process we are studying (i. e., fault slip). For inversions for fault slip, the forward model is our best guess at how the deformation field produced by motion on the fault surface will propagate through the solid earth to the ground surface where we can observe it. Laboratory, seismological and field observations support the idea that the crust behaves as an elastic solid on the short timescales associated with earthquakes and their immediate aftermath, although parts of the crust deform viscously at longer time scales under prolonged stress. The response of the elastic crust surrounding a plane that undergoes fault slip can be described mathematically [62]. Figure 7 shows the predicted ground deformation for strike-slip motion on fault "patches" at varying depths. The ground deformation response is linear, i. e., the deformation from multiple sources is just the sum of deformation from the individual fault patches. Therefore, we can predict the expected ground deformation from arbitrarily complicated fault zone geometries and slip distributions.

In an inverse problem, we consider the behavior of a forward model or family of forward models from a num-

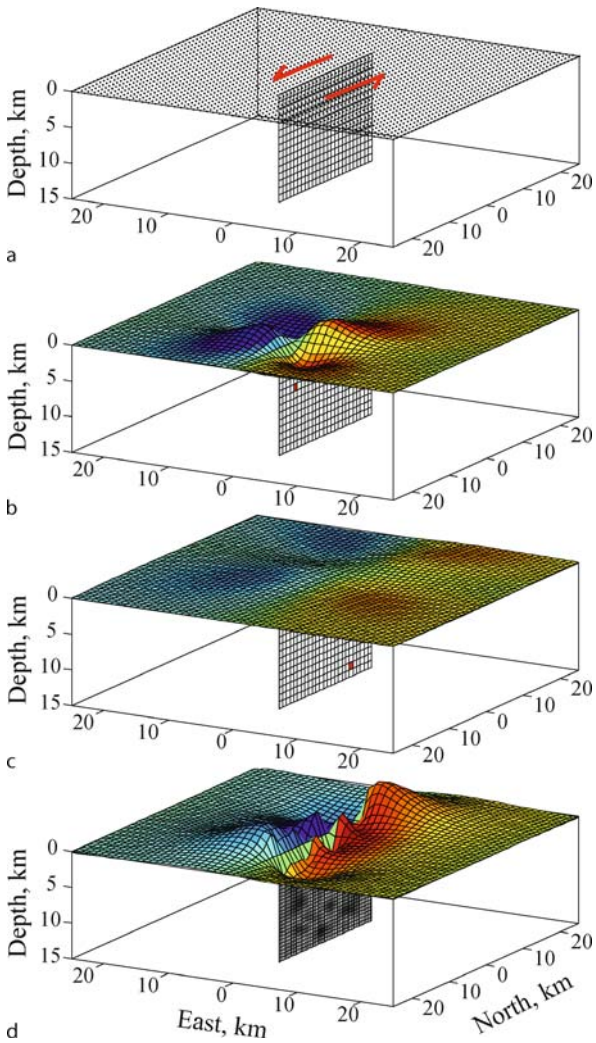

Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 7

Predicted ground surface deformation for a vertical, left-lateral strike-slip fault a, using Okada [62]. Fault slip on a shallow fault patch (*red*, panel b) produces a larger amount of deformation over a smaller area than does fault slip on a deeper region of the fault plane c. Here, *deformation of the grid* indicates vertical deformation and *color* indicates the amount of northward (*red*) or southward (*blue*) deformation. The amount of total deformation from a complicated fault slip distribution is achieved by summing the components from all parts of the fault d

ber of possible sources that could potentially explain our data, and we find the combination these sources that best fits the data. In this section, the family of forward models is represented by slip on a fault plane that has been divided up into a number of fault patches, and the best-fit model would be the distribution of fault slip magnitude along this fault plane that best matches the observed ground deformation.

Linear vs. Nonlinear Inverse problems can be divided up into two major families: Linear and Nonlinear. In a linear problem, we solve for the sum of forward models that best fits the data, without any size constraints on the individual contributions from each forward model (fault patch). One example of a linear problem is where we find the best-fit fault slip distribution on a fixed fault plane, without any constraints such as requiring that the fault slip has a positive or negative sign (i. e., right-lateral vs. left lateral). Linear problems are essentially extensions of the problem of fitting a line to a set of data, and are quite easy to solve [52].

Nonlinear problems exist when the functional shape and form of the predicted data varies with the parameter choice. An inversion for the location of the fault plane itself, where quantities such as the x, y, z, location or fault plane strike and dip are varied, is very nonlinear. The process of solving a nonlinear problem involves finding a model that best fits the data in some pre-defined way, usually by finding the model that produces the lowest sum of squared residuals with the data (least-squares fit). Nonlinear problems may have multi-dimensional misfit functions with poorly defined or multiple minima, and the only way to find the absolute or "global" minimum is by exhaustively searching the parameter space.

Methods for searching the parameter space in a nonlinear problem fall into two main camps. Gradient-based methods determine the slope and gradient of the misfit surface (often in many dimensions) and follow it downhill in an iterative way. These methods work well if the misfit surface has a single, well defined minimum. If the search begins near one of the local minima, it is possible that the true, global minimum will never be found. Global methods involve variations that range from simply sampling the parameter space very densely, to iterative methods that track many initial samples or families of samples in gradient-based searches. One method that performs very well in the ~9-dimensional parameter space of finding the best-fit fault plane and slip that fits data for a small earthquake is the Neighborhood Algorithm [44,78]. In the Neighborhood Algorithm (NA), the misfit values at many initial points are used to iteratively focus in on regions of the parameter space with lower misfit values. The NA method has the strength of being able to track and define multiple minima instead of just choosing one. A survey of several other nonlinear optimization methods applied to earthquake location problems can be found in Cervelli [6].

Smoothing The specific problem of inverting for fault slip on a particular fault plane has an interesting twist to it that is shared by many geophysical inverse problems. Just

as an inversion for the best-fit line to data usually requires at least two data points, other inversions also effectively require the existence of more data than unknown quantities. For fault-slip problems the question does not boil down to simple numbers of data points vs. the number of fault patches – parts of the fault plane are better-resolved than others, and some data points contribute more information than is provided by their neighbors. For instance, if we had a thousand GPS receivers within 100 feet of the fault zone, but extended our target fault plane down to 100 km, the GPS data would be able to tell us little about what was going on at depth, even if we only divided the fault plane up into a few fault patches. Conversely, data points very close together and far from the fault contribute essentially the same information about what occurred – combining them can help to reduce noise, but it does not help us to isolate variations in fault slip along the fault.

Because of these inherent variations in resolution, both in model and data, an inversion can run into problems if we simply solve for the best-fit fault slip distribution on a finely-divided fault plane. Since the forward models for two fault patches at great depth are very similar on the surface, the inversion cannot determine the relative strength of fault slip that should be assigned to each. The difference between [0 1] and [-1000 1001] is very small when you propagate it up to the surface, often smaller than the noise in the data (or even machine precision!). Therefore, the "best-fit" model would have unrealistically large variations that have nothing to do with what really occurred during the earthquake (Fig. 8). This effect is often known as "checkerboarding". While some of the features in the inferred slip distribution are related to the earthquake, much of the complexity is due to the attempt of the inversion to fit noise within the data.

There are several methods for dealing with this effect, all of which place some bounds on how large or spatially rough the variations in fault slip can be. The simplest method (although one of the hardest to optimize) is to parametrize the fault plane in a way that the fault patches are never so small that there are large tradeoffs in their predicted surface deformation [70]. The extreme of this would be to only use one fault patch. In most cases where we have a lot of data, this solution would not fit the data well nor tell us much about the earthquake.

Another method is to place a penalty during the least squares inversion on fault slip solutions that are very large (usually involving large variations) or that are spatially rough. These "regularized" inversions result in spatially smooth slip distributions that usually fit the data almost as well as the rough, unregularized inversions. Regularization always involves some choice of how much weight needs to be placed on the roughness penalty vs. the fit to the data, which can be a difficult procedure. Too large of a penalty weight and the slip distribution will be too smooth and will not fit the data (the logical extension of this is the single fault patch model discussed above). Too small of a weighting and the slip distribution will be arbitrarily rough, with unrealistic changes in sign throughout the fault zone.

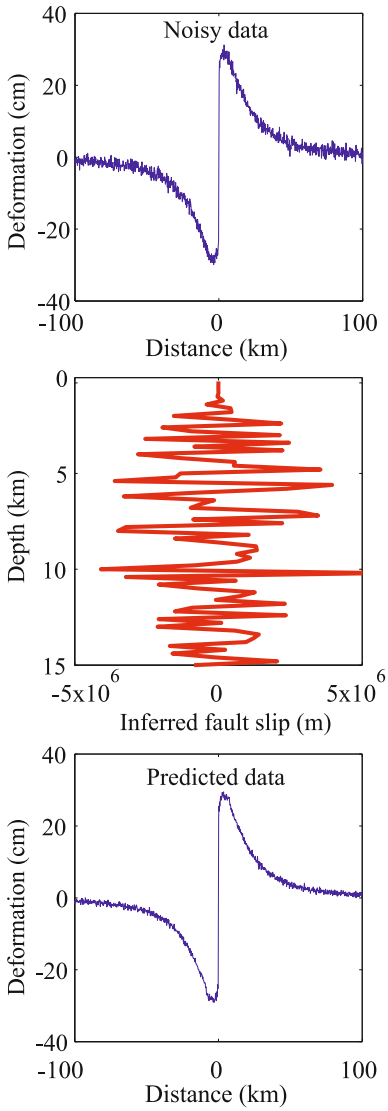

Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 8

Example of how noisy data can affect fault slip inversion. *Top:* Deformation from a vertical strike slip fault, with added noise. *Middle:* Best fit slip distribution inferred from noisy data. The large fluctuations mostly cancel each other out at the surface, producing (*bottom*) predicted data that matches both the underlying deformation signal and the noise

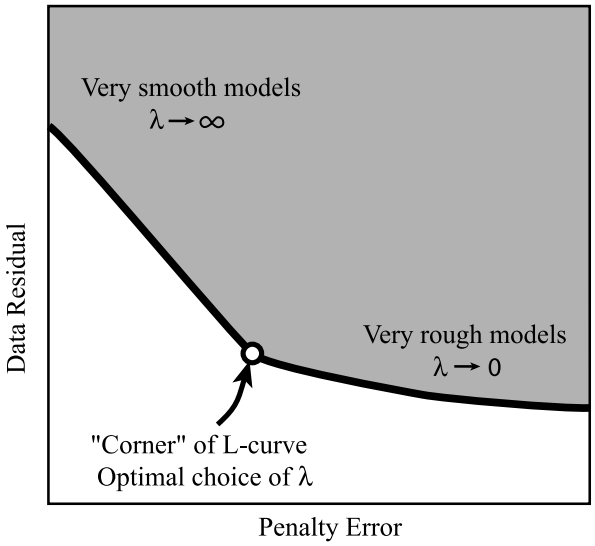

Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 9
Typical L-curve. Each *point* on the curve corresponds to a slip distribution inferred using a different value of λ. In most cases, the "penalty error" will be a measure of model roughness

There are two main families of methods for choosing the appropriate smoothing weight (apart from random guessing, which often performs surprisingly well!). In general, as the amount of regularization is increased, the fault slip distribution becomes smoother and the fit to the data worsens. Ideally, we would choose the value where the proportion of the slip distribution that is just fitting data noise is as small as possible, without smoothing so much that we fit neither the noise nor the underlying signal due to the earthquake. The first method requires choosing the smoothing weight from a plot of data fit vs. model roughness, often referred to as an "L-curve" because of the characteristic shape of such curves (Fig. 9). The smoothing value at the corner of the "L" generally fits the data well without being "too rough". This type of parameter choice is slightly arbitrary and not easy to automate, but it does allow the researcher to include some intuition about characteristics of the earthquake.

The other family relies on the concept that a good choice of smoothing would be able to reproduce another, independent set of data spanning the earthquake fairly well. Too much smoothing and the slip model would fit neither the original nor the additional data sets well. If the choice of smoothing were too small, the resulting complex slip distribution would mainly be fitting noise in the original data set and would not, therefore, fit the independent noise in the second data set.

Of course, we rarely have the luxury of multiple data sets – if they do exist, we should use them for the main inversion! Data resampling procedures, known variously as the bootstrap, jacknife or cross-validation [10] are used to simulate the existence of multiple, independent data sets in cases where only one exists. Du et al. [9] compared cross-validation and other techniques for choosing smoothing parameters for geodetic data spanning the 1983 Borah Peak earthquake. Another powerful parameter choice method is the Akaike Bayesian Information Criterion (ABIC) that can be used to choose smoothing weights or other inversion characteristics [1,30,97].

Noise Although finding the best-fit model to our data is certainly important, knowing how confident we are in that estimate is just as crucial. Studies of postseismic behavior, for instance, rely on good estimates of the coseismic slip distribution [25], and are improved further when we know what constraints we can place on that slip distribution. A variety of techniques exist for estimating these confidence limits – for linear problems, and when we know the character of the noise (magnitude, spatial correlation), it is quite simple to propagate data errors through to error bounds on the inversion results [52]. However, for nonlinear problems, and for cases where we are not quite sure how much of our signal is noise (usually the case with InSAR data), we need to rely on other methods for estimating the noise.

The same data resampling procedures described above can be used to generate multiple sets of inversion results that should reflect the noise structure of the data, even when it is not understood ahead of time. For nonlinear problems, such as the 3-D location of the best fit fault plane and earthquake mechanism for a particular earthquake, knowledge about the data noise can be used to construct multiple synthetic data sets that can then each be inverted using the nonlinear method of choice. Any conclusions about the problem (i. e., error bounds on fault slip) should be drawn from the ensemble as a whole [44].

Case Examples

Here we examine case examples for two earthquakes. In each case, we review the data and discuss characteristics of the inferred slip distributions.

Mw 7.6 Manyi, Tibet, Earthquake The 1997 strike-slip earthquake that occurred Tibet produced a very long rupture (170 km) with up to 7 meters of offset. It was such a long rupture that it takes three overlapping SAR tracks (Fig. 10) to cover its full length! Here, we show the effects

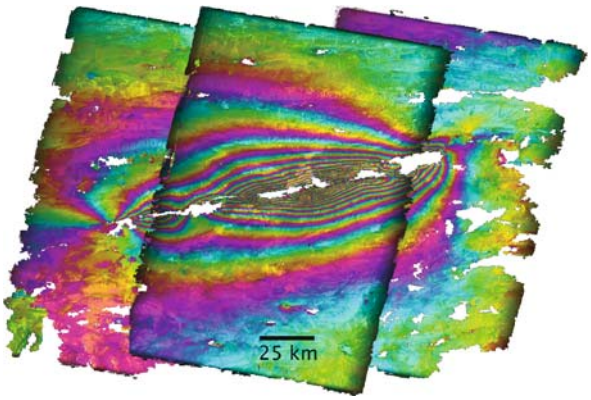

Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 10
Three overlapping interferograms spanning the 1997 Manyi earthquake

of variations in the spatial smoothing weights (λ) on an inversion of the deformation field for fault slip. Figure 11 illustrates the differences in inferred slip distributions for various magnitudes of λ, resulting in a spectrum between very smooth slip distributions that do not fit the data very well, to impossibly rough slip distributions that only fit the data marginally better than the "optimal" model (center panel). Our optimal model predicts up to 10 meters of fault slip with most slip occurring in the upper 10 km. Here, we chose the optimal value of λ using cross-validation, but the other methods discussed above can also be applied to this problem.

Asymmetries in the profiles of deformation across the fault during the Manyi earthquake have led some researchers to think that the elastic behavior of the crust may be different on one side of the fault vs. the other [67]. The magnitude of horizontal ground motion for a straight, vertical, strike-slip fault should be symmetric across the fault, if there are no other complications. However, variations in fault plane dip can affect the resulting ground deformation as much as variations in crustal properties, requiring that such studies carefully consider the fault geometry used in their inversions.

Mw 7.1 Hector Mine, CA, Earthquake The 1999 Hector Mine earthquake [16,85] occurred in the Mojave Desert of Southern California, and was widely studied due to it's proximity to large population centers and the San Andreas fault, it's very complete data coverage (GPS, InSAR, seismic data), and the fact that it occurred only a few years after another very large earthquake on a neighboring fault (1992 Landers earthquake). The combination of the two earthquakes within such a short time interval suggests that series of earthquakes may occur separated by long intervals of quiescence instead of events always occurring at semi-regular time intervals of strain buildup and release [61,69].

In Fig. 12, we show a subset of the available InSAR data spanning the Hector Mine earthquake. Both continuous and campaign GPS data were also recorded on either side of the Y-shaped rupture. In Fig. 13, we show the re-

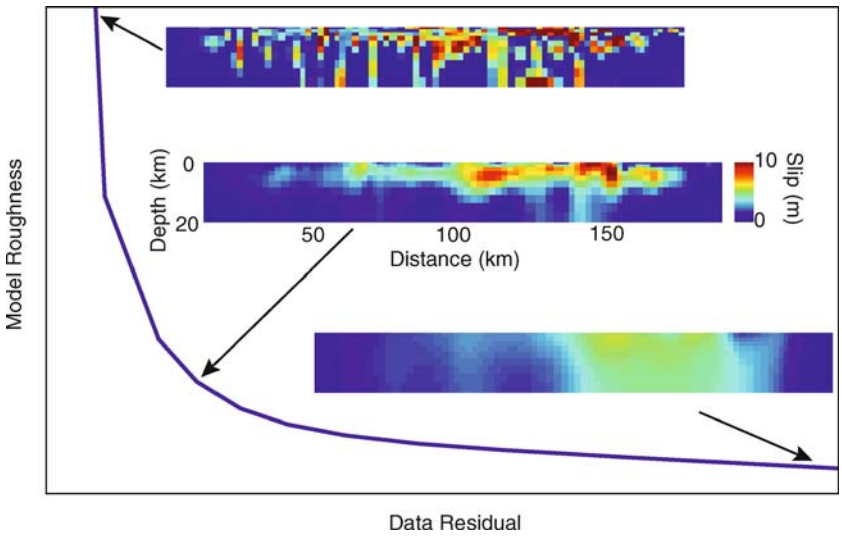

Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 11
L-curve for the Manyi EQ, showing slip distributions inferred using 3 different values of smoothing, each with the same color scale. Small λ (*top panel*) produces an unrealistically rough slip distribution, whereas very large λ (*bottom panel*) produces a smooth model that doesn't fit the data

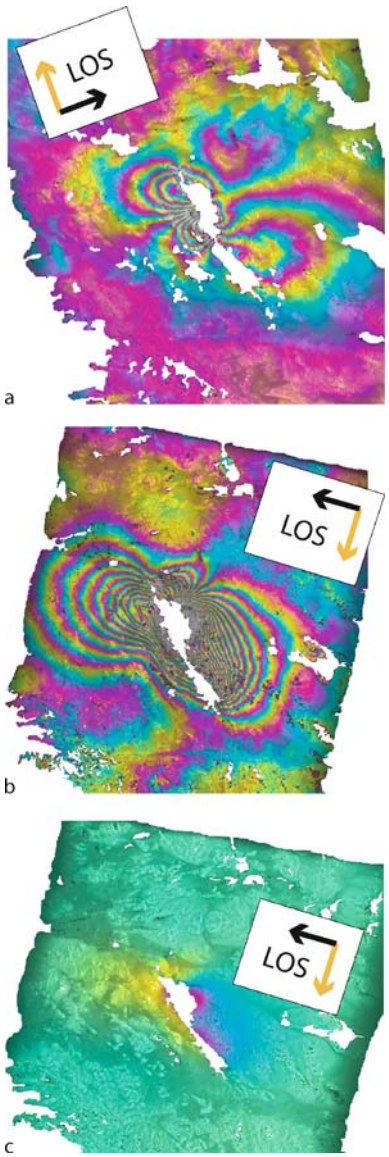

Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 12
Three representations of the Hector Mine EQ. a "Ascending" and b "descending" wrapped interferograms, where the satellite is moving in the direction indicated by the *yellow arrow* and viewing in the direction of the *black arrow*. Note how the same deformation field looks very different depending on the viewing direction. Color cycles correspond to ∼3 cm of deformation. c Unwrapped version of interferogram in b. Now the color scale corresponds to 3 meters and spans the entire dynamic range of the data

sults of inverting the GPS and all the InSAR data for the best-fit fault slip distribution, with up to 6 meters of fault slip. Note how the maximum fault slip is not at the surface, but peaks at a few km depth. If we plot the average slip vs.

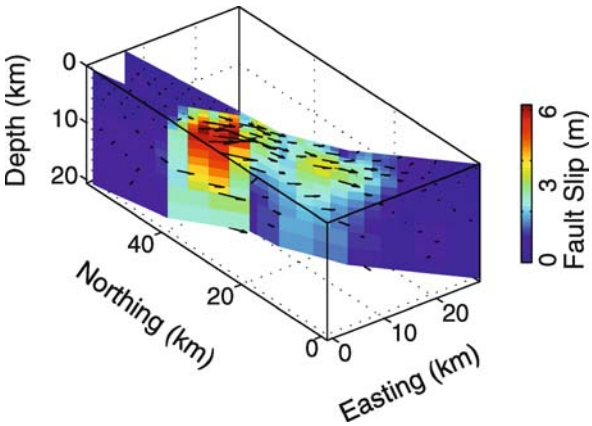

Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 13
3D representation of inferred fault slip that occurred during the Hector Mine earthquake

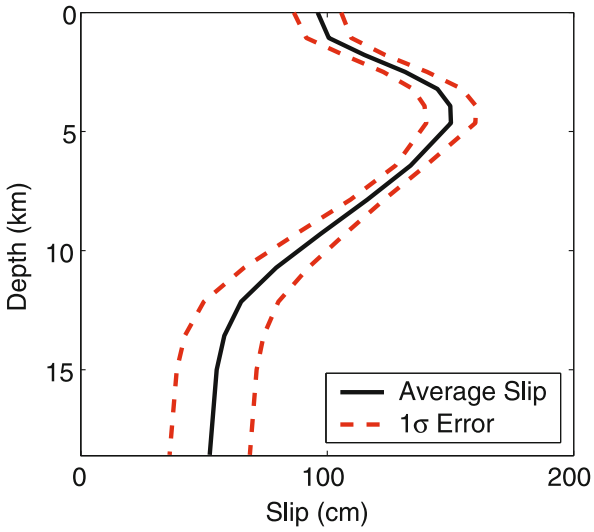

Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of, Figure 14
Fault slip during the Hector Mine earthquake, averaged along strike and plotted vs. depth. *Red dashed lines* indicate the expected error bounds introduced by atmospheric noise

depth (Fig. 14), the profile clearly shows how the shallow slip deficit is robust even given the amount of noise in the data.

Future Directions

The spatial and temporal coverage of geodetic data sets such as InSAR and GPS is increasing to the point where we can observe new types of fault zone behavior and solve for characteristics on a scale and precision that could hardly

have been imagined a few decades ago. As we increase our ability to understand and describe the kinematics of crustal deformation, we can begin to explore the dynamic processes driving seismic and volcanic deformation in tectonically active regions around the world. Three of the main fields that show the most promise in the near future are discussed below:

Transient Deformation and Slow, Silent Earthquakes

The episodic slow slip events that show up with amazing regularity (∼14 months) in Cascadia ([51,54], Fig. 6), and that are also observed in Japan [27,56,65] and Mexico [45], are some of the most exciting new types of fault zone behavior that scientists have observed in recent years. The idea that earthquakes contained energy that was released by slower ruptures was noticed using seismic data far earlier than it was ever observed geodetically (1960 Chilean earthquake [34]). The slow slip events are often associated with heightened levels of spatially-correlated, low-level seismicity, or seismic tremor [87]. Currently, the process that allows these slow earthquakes to initiate and propagate at the observed speeds is still unknown, although they may be occurring at depths where dehydration of subducted sediments becomes important.

Apart from their intrinsic interest (since we always seek to explain processes that we do not understand), these slow/silent events release a not insignificant amount of the accumulated strain across the fault boundaries. In some ways this is a boon to people living nearby – any plate tectonic motion that is accommodated aseismically means that there is that much less accumulated strain that could be released in a destructive earthquake. However, there are indications that aseismic slip events are often closely followed by earthquakes [42,75], indicating that the changes in stress during the slow event can push other, nearby, regions past the brink until they rupture coseismically. This, also, can be seen in a positive light – perhaps close monitoring of aseismic deformation along active plate boundaries can serve as an early warning system, or at least a signal to raise the forecasted earthquake hazard whenever heightened activity is observed.

Reduction or Modeling of Atmospheric Noise

One of the largest hurdles for researchers approaching InSAR data is the atmospheric noise present in all interferograms. GPS data is affected by this problem as well, but the atmospheric signal can be solved for or averaged out due to the long observation times allowed by GPS. An area of active research is the modeling and removing of the atmospheric noise signal from InSAR data, using a variety of tools that range from other satellite-based observations of atmospheric water vapor content to simplistic models of correlations between elevation and interferometric phase [96].

The first-order layering of the atmosphere results in profiles of water vapor that tend to decrease vs. elevation in a manner that often appears essentially linear when two SAR images are combined in an interferogram. However, lateral variations with distance from water bodies and gradients from one side of a mountain belt to the other, result in the fact that the appropriate elevation vs. signal can vary quite a bit across a typical SAR image. For some target problems, such as the location of a small earthquake or the fault slip distribution for an earthquake in a relatively flat area, these elevation-dependent signals will likely affect the inversion to only a limited extent. However, in studies where the signal of interest is correlated with elevation, such as fold growth or subsidence caused by lake loading, great care should be taken when removing overly simple models of atmospheric water vapor from the data.

A more promising, albeit still problematic, approach is the modeling of water vapor content based on observations from satellite-based platforms. Water vapor measurements are made by a variety of satellites, but two in particular have the spatial scale and temporal resolution that make them potentially very useful in InSAR applications. The Moderate Resolution Imaging Spectroradiometer (MODIS) instruments on both the Terra and Aqua satellites acquire almost daily observations over most of the Earth's surface at fairly high resolution (1 km spacing). They are not acquired at exactly the same time as the SAR images, but they can still be used to track seasonal trends or to seek individual features that may persist between the SAR and MODIS image acquisition times. While these cannot be used to quantitatively remove the atmospheric signal, they can be very useful from a qualitative standpoint in determining whether a particular interferometric signal may or may not be tectonic in origin.

The Medium Resolution Imaging Spectrometer (MERIS) is physically located on the same satellite (ENVISAT) that acquires much of the SAR imagery used today. Since MERIS observations are made at essentially the same time as the SAR imagery, they see the same atmosphere and can not only be used in the same qualitative way as MODIS, but also show promise of allowing the actual removal of atmospheric contributions to the interferometric phase. Some difficulties lie in the fact that MERIS and MODIS measurements of atmospheric water vapor can only be made in cloud-free images, and that "double-bounce" effects of the signal bouncing off the base of both visible and invisible clouds can bias the ob-

servations and introduce spurious water vapor features. Still, these satellite-based methods show great promise. The most robust method for reducing the contribution from atmospheric noise to our inversions remains the use of multiple independent data sets, whenever available.

Data Assimilation

As mentioned in Sect. "Highlights of Earthquake Geodesy", the vast quantity of data now available to us can prove troublesome – if it takes hours to perform one forward model of a possible earthquake scenario with the available data, then the inverse problem quickly becomes unmanageable (or at least can begin to extend past the length of a normal graduate thesis). Parallel computing methods and the rapid decrease in cost for computing resources now makes the operation of large, multi-processor machines feasible within individual research departments. These large machines allow us to approach inverse problems with the "big hammer" of stochastic, or Monte Carlo, methods, which rely on the use of many randomly generated simulations of a system [6]. However, the expected continuing increase of both spatial and temporal coverage of deformation observations requires that we need to continue developing new tools that allow us to capitalize on data time series as well as individual snapshots of ground deformation.

Kalman filters and related methods are one very powerful tool now in widespread use, especially within the GPS community [51,84]. The incorporation of InSAR data into such methods is slightly more difficult, in part because of the large number of data points, but also because of the varying character of noise between interferograms and difficulties in dissociating the description of noise vs. modeling of the geophysical signal of interest. Data assimilation methods developed in the atmospheric sciences, which also deal with data sets of varying spatial and temporal scales and resolutions, are a potentially rich source of tools that the geodetic community can explore in the near future.

Bibliography

Primary Literature

1. Akaike H (1980) Bayesian statistics. In: Bernardo JM, DeGroot MH, Lindley DV, Smith AFM (eds) Likelihood and the Bayes procedure. University Press, Valencia, pp 143–166
2. Allen CR et al (1972) Displacements on the Imperial, Superstition Hills, and San Andreas faults triggered by the Borrego Mountain Earthquake. US Geol Surv Prof Pap 787:87–104
3. Arnadottir T, Segall P (1994) The 1989 Loma Prieta earthquake imaged from inversion of geodetic data. J Geophys Res 99:21835–21855
4. Backus G, Gilbert F (1970) Uniqueness in the inversion of inaccurate gross earth data. Phil Trans R Soc Lond 266:123–192
5. Burgmann R et al (2002) Time-dependent afterslip on and deep below the Izmit earthquake rupture. Bull Seism Soc Amer 92:126–137
6. Cervelli P et al (2001) Estimating source parameters from deformation data, with an application to the March 1997 earthquake swarm off the Izu Peninsula, Japan. J Geophys Res 106:11217–11237
7. Dieterich JH (1992) Earthquake nucleation on faults with rate- and state- dependent strength. Tectonophysics 211:115–134
8. Dominguez S, Avouac JP, Michel R (2003) Horizontal coseismic deformation of the 1999 Chi-Chi earthquake measured from SPOT satellite images: Implications for the seismic cycle along the western foothills of central Taiwan. J Geophys Res 108. doi: 10.1029/2001JB000951
9. Du Y, Aydin A, Segall P (1992) Comparison of various inversion techniques as applied to the determination of a geophysical deformation model for the 1983 Borah Peak earthquake. Bull Seism Soc Amer 82:1840–1866
10. Efron B, Tibshirani R (1993) An introduction to the bootstrap. In: Monographs on statistics and applied probability, vol 83. Chapman and Hall, London
11. Emardson TR, Simons M, Webb FH (2003) Neutral atmospheric delay in interferometric synthetic aperture radar applications: Statistical description and mitigation. J Geophys Res 108. doi: 10.1029/2002JB001781
12. England P, Jackson J (1989) Active deformation of the continents. Annu Rev Earth Planet Sci 17:197–226
13. Fialko Y (2004) Evidence of fluid-filled upper crust from observations of post-seismic deformation due to the 1992 Mw 7.3 Landers earthquake. J Geophys Res 109. doi: 10.1029/2004JB002985
14. Fialko Y (2006) Interseismic strain accumulation and the earthquake potential on the southern San Andreas fault system. Nature 441:968–971
15. Fialko Y et al (2002) Deformation on nearby faults induced by the 1999 Hector Mine Earthquake. Science 297:1858–1862
16. Fialko Y, Simons M, Agnew D (2001) The complete (3-D) surface displacement field in the epicentral area of the 1999 Mw 7.1 Hector Mine earthquake, California, from space geodetic observations. Geophys Res Lett 28:3063–3066
17. Freymueller J, King NE, Segall P (1994) The Co-seismic slip distribution of the Landers earthquake. Bull Seism Soc Amer 84:646–659
18. Funning GJ et al (2005) Surface displacements and source parameters of the 2003 Bam (Iran) earthquake from Envisat advanced synthetic aperture radar imagery. J Geophys Res 110. doi: 10.1029/2004JB003338
19. Gilbert GK (1890) Lake Bonneville. In: US Geol Surv Monograph, vol 1. Washington
20. Goldstein R (1995) Atmospheric limitations to repeat-track radar interferometry. Geophys Res Lett 22:2517–2520
21. Griesbach CL (1893) Notes on the earthquake in Baluchistan on the 20th December 1892. Geol Survey India Rec 26
22. Hager BH, King RW, Murray MH (1991) Measurement of crustal deformation using the global positioning system. Annu Rev Earth Planet Sci 19:351–382
23. Hanssen RA (2001) Radar interferometry: Data interpretation and error analysis. Kluwer, Dordrecht

24. Harris RA, Segall P (1987) Detection of a locked zone at depth on the Parkfield, California segment of the San Andreas Fault. J Geophys Res 92:7945–7962
25. Hearn EH (2002) Dynamics of Izmit earthquake postseismic deformation and loading of the Duzce earthquake hypocenter. Bull Seism Soc Amer 92:172–193
26. Hetland EA, Hager BH (2003) Postseismic relaxation across the central Nevada seismic belt. J Geophys Res 108. doi: 10.1029/2002JB002257
27. Hirose H et al (1999) A slow thrust slip event following the two 1996 Hyuganada earthquakes beneath the Bungo Channel, southwest Japan. Geophys Res Lett 26:3237–3240
28. Hsu YJ et al (2006) Frictonal afterslip following the Mw 8.7, 2005 Nias-Simeulue earthquake, Indonesia. Science 312. doi: 10.1126/science.1126960
29. Ito T, Hashimoto M (2004) Spatiotemporal distribution of interplate coupling in southwest Japan from inversion of geodetic data. J Geophys Res 109. doi: 10.1029/2002JB002358
30. Jackson DD, Matsuura M (1985) A Bayesian approach to nonlinear inversion. J Geophys Res 90:581–591
31. Johanson IA, Burgmann R (2005) Creep and quakes on the northern transition zone of the San Andreas Fault from GPS and InSAR data. Geophys Res Lett 32. doi: 10.1029/2005GL023150
32. Johnson KM, Burgmann R, Larson K (2006) Frictional properties on the San Andreas fault near Parkfield, California, inferred from models of afterslip following the 2004 earthquake. Bull Seism Soc Amer 96:S321–S338
33. Jonsson S et al (2003) Post-earthquake ground movements correlated to pore-pressure transients. Nature 424:179–183
34. Kanamori H, Stewart GS (1972) A slow earthquake. Phys Earth Planet Int 18:167–175
35. King GCP, Stein RS, Rundle JB (1988) The growth of geological structures by repeated earthquakes 1: Conceptual framework. J Geophys Res 93:13307–13318
36. Koto B (1983) On the cause of the great earthquake in central Japan, 1891. J Coll Sci Imp Univ Japan 5:296–353
37. Langbein J, Gwyther RL, Hart RHG, Gladwin MT (1999) Slip-rate increase at Parkfield in 1993 detected by high-precision EDM and borehole tensor strainmeters Source. Geophys Res Lett 26(16):2529–2532
38. Langbein J, Murray JR, Snyder HA (2006) Coseismic and initial postseismic deformation from the 2004 Parkfield, California, Earthquake, observed by Global Positioning System, Electronic Distance Meter, Creep Meters, and Borehole Strainmeters. Bull Seism Soc Amer 96:304–320
39. Larsen S et al (1992) Global Positioning System measurements of deformations associated with the 1987 Superstition Hills earthquake – evidence for conjugate faulting. J Geophys Res 97:4885–4902
40. Lawson AC et al (1908) The California earthquake of April 18, 1906 – report of the state earthquake investigation committee. Carnegie Insitute, Washinton
41. Lin J, Stein RS (1989) Coseismic folding, earthquake recurrence and the 1987 source mechanism at Whittier Narrows, Los Angeles Basin, California. J Geophys Res 94:9614–9632
42. Lohman RB, McGuire JJ (2007) Earthquake swarms driven by aseismic creep in the Salton Trough, California. J Geophys Res 112. doi: 10.1029/2006JB004596
43. Lohman RB, Simons M (2005) Some thoughts on the use of InSAR data to constrain models of surface deformation: Noise structure and data downsampling. Geochem Geophys Geosyst 6. doi: 10.1029/2004GC000841
44. Lohman RB, Simons M (2005) Locations of selected small earthquakes in the Zagros mountains. Geochem Geophys Geosyst 6. doi: 10.1029/2004GC000849
45. Lowry AR et al (2001) Transient fault slip in Guerrero, southern Mexico. Geophys Res Lett 28:3753–3756
46. Lyell C (1837) Principles of Geology, 5th edn. Murray, London
47. Lyons S, Sandwell D (2003) Fault creep along the southern San Andreas from interferometric synthetic aperture radar, permanent scatterers and stacking. J Geophys Res 108. doi: 10.1029/2002JB001831
48. Mallet R (1862) The first principles of observational seismology. Chapman and Hall, London
49. Massonnet D, Rossi M, Carmona C, Adragna F, Pelzer G, Feigl K, Rabaute T (1993) The displacement field of the Landers earthquake mapped by radar interferometry. Nature 364: 138–142
50. McKay A (1890) On the earthquake of September 1888, in the Amuri and Marlborough districts of the South Island. NZ Geol Surv Rep Geol Explor 1885–1889 20:78–1007
51. McGuire J, Segall P (2003) Imaging of aseismic fault slip transients recorded by dense geodetic networks. Geophys J Int 155:778–788
52. Menke W (1989) Geophysical data analysis: Discrete inverse theory. Academic Press, London
53. Middlemiss CS (1910) The Kangra earthquake of 4th April, 1905. Geol Surv India Mem 37
54. Miller MM et al (2002) Periodic slow earthquakes from the Cascadia subduction zone. Science 295:2423
55. Miyazaki S et al (2004) Space time distribution of afterslip following the 2003 Tokachi-oki earthquake: Implications for variations in fault zone frictional properties. Geophys Res Lett 31. doi: 10.1029/2003GL019410
56. Miyazaki S, McGuire JJ, Segall P (2003) A transient subduction zone slip episode in southwest Japan observed by the nationwide GPS array. J Geophys Res 108. doi: 10.1029/2001JB000456
57. Molnar P, Tapponnier P (1975) Cenozoic tectonics of Asia: Effects of a continental collision. Science 189:419–425
58. Muller JJA (1895) De verplaasting van eenige traiangulatie pilaren in de residenti Tapanuli (Sumatra) tengevolge de aardbeving van 17 Mei 1892. Natuurwet Tijdscht Ned Indie 54: 299–307
59. Murray J, Segall P (2002) Testing time-predictable earthquake recurrence by direct measurement of strain accumulation and release. Nature 419:298–291
60. Nadeau RM, McEvilly TV (2004) Periodic pulsing of characteristic microearthquakes on the San Andreas Fault. Science 303:202–222
61. Nur A, Hagai R, Beroza G (1993) The nature of the Landers-Mojave earthquake line. Science 261:201–203
62. Okada Y (1985) Surface deformation due to shear and tensile faults in a half space. Bull Seism Soc Amer 75:1135–1154
63. Okudo T (1950) On the mode off the vertical land-deformation accompanying the great Nankaido earthquakes. Bull Geogr Surv Inst 2:37–59
64. Oldham RD (1928) The Cutch (Kacch) earthquake of 16th June 1819, with a revision of the great earthquake of 12th June 1897. Geol Surv India Mem 46:71–147

65. Ozawa S et al (2002) Detection and monitoring of ongoing aseismic slip in the Tokai region, central Japan. Science 298:1009–1012
66. Parker RL (1977) Understanding inverse theory. Annu Rev Earth Planet Sci 5:35–64
67. Peltzer G, Crampe F, King G (1999) Evidence of nonlinear elasticity of the crust from the Mw 7.6 Manyi (Tibet) earthquake. Science 286:272–276
68. Pollitz FF (2003) Transient rheology of the uppermost mantle beneath the Mojave desert, California. Earth Plan Sci Lett 215:89–104
69. Pollitz FF, Sacks IS (2002) Stress triggering of the 1999 Hector Mine earthquake by transient deformation following the 1992 Landers Earthquake. Bull Seism Soc Amer 92:1487–1496
70. Pritchard ME et al (2002) Co-seismic slip from the 1995 July 30 Mw=8.1 Antofagasta, Chile, earthquake as constrained by InSAR and GPS observations. Geophys J Int 150:362–376
71. Pritchard ME, Simons M (2006) An aseismic slip pulse in northern Chile and along-strike variations in seismogenic behavior. J Geophys Res 111. doi: 10.1029/2006JB004258
72. Reid HF (1910) The mechanics of the earthquake. In: Lawson AC (ed) The California earthquake of April 18, 1906. Carnegie Institute, Washington
73. Reid HG (1911) The elastic-rebound theory of earthquakes. Univ Calif Pub Bull 6:413–444
74. Reilinger R (1986) Evidence for postseismic viscoelastic relaxation following the 1959 M=7.5 Hebgen Lake, Montana, earthquake. J Geophys Res 91:9488–9494
75. Roeloffs EA (2006) Evidence for aseismic deformation rate changes prior to earthquakes. Annu Rev Earth Planet Sci 34:591–627
76. Rosen PA et al (2000) Synthetic aperture radar interferometry. Proc IEEE 88:333–382
77. Ruegg JC et al (1996) The Mw=8.1 Antofagasta (North Chile) earthquake July 30, 1995: First results from teleseismic and geodetic data. Geophys Res Lett 23:917–920
78. Sambridge M (1998) Geophysical inversion with a neighborhood algorithm – I: Searching a parameter space. Geophys J Int 138:479–494
79. Savage JC, Lisowski M, Svarc JL (1994) Postseismic deformation following the 1989 (M=7.1) Loma Prieta, California, earthquake. J Geophys Res 99:13757–13765
80. Schmidt DA et al (2005) Distribution of aseismic slip rate on the Hayward fault inferred from seismic and geodetic data. J Geophys Res 110. doi: 10.1029/2004JB003397
81. Scholz CH (1998) Earthquakes and friction laws. Nature 391: 37–42
82. Segall P, Burgmann R, Matthews M (2000) Time-dependent triggered afterslip following the 1989 Loma Prieta earthquake. J Geophys Res 105:S615–S634
83. Segall P, Davis JL (1997) GPS Applications for Geodynamics and Earthquake Studies. Annu Rev Earth Planet Sci 25:301–336
84. Segall P, Matthews M (1997) Time dependent inversion of geodetic data. J Geophys Res 102:22391–22409
85. Simons M, Fialko Y, Rivera L (2002) Coseismic deformation from the 1999 Mw 7.1 Hector Mine, California, earthquake as inferred from InSAR and GPS observations. Bull Seism Soc Amer 92:1390–1402
86. Sharp RV et al (1982) Surface faulting in the Imperial Valley. In: Sharp RV, Lienkaemper JJ, Bonilla MG, Burke DB, Cox BF, Herd DG, Miller DM, Morton DM, Ponti DJ, Rymer MJ, Tinsley JC, Yount JC, Kahle JE, Hart EW, Sieh K (eds) The Imperial Valley, California, earthquake of October 15, 1979. US Geol Surv Prop Pap 1254:119–144
87. Shelley DR et al (2006) Low frequency earthquakes in Shikoku, Japan, and their relationship to episodic tremor and slip. Nature 442:188–191
88. Shimazaki K, Nakata T (1980) Time-Predictable Recurrence Model for Large Earthquakes. Geophys Res Lett 7(4):279–282
89. Song AT, Simons M (2003) Large trench-parallel gravity variations predict seismogenic behavior in subduction zones. Science 301:630–633
90. Stein R (1999) The role of stress transfer in earthquake occurrence. Nature 402:605–609
91. Tarantola A, Valette B (1982) Inverse problems = quest for information. J Geophys 50:159–170
92. Thatcher W (1984) The earthquake deformation cycle at the Nankai trough, southwest Japan. J Geophys Res 89:3087–3101
93. Tsuboi C (1932) Investigation on the deformation of the earth's crust in the Tango district connected with the Tango earthquake of 1927 (part 4). Bull Earthq Res Inst Tokyo Univ 10: 411–434
94. Ward S, Valensise GR (1989) Fault parameters and slip distribution of the 1915 Avezzano, Italy, earthquake derived from geodetic observations. Bull Seism Soc Amer 79:690–710
95. Wells RE et al (2003) Basin-centered asperities in great subduction zone earthquakes: A link between slip, subsidence and subduction erosion. J Geophys Res 108. doi: 10.1029/2002JB002072
96. Williams S, Bock Y, Fang P (1998) Integrated satellite interferometry: Tropospheric noise, GPS estimates and implications for interferometric synthetic aperture radar products. J Geophys Res 103:27051–27067
97. Yabuki T, Matsuura M (1992) Geodetic data inversion using a Bayesian information criterion for spatial distribution of fault slip. Geophys J Int 109:363–375
98. Yang M et al (2000) Geodetically observed surface displacements of the 1999 Chi-Chi, Taiwan earthquake. Earth Planet Space 52:403–413
99. Yeats RS, Sieh K, Allen CF (1997) The geology of earthquakes. Oxford University Press, New York
100. Zebker HA, Rosen PA, Hensley S (1997) Atmospheric effects in interferometric synthetic aperture radar surface deformation and topographic maps. J Geophys Res 102:7547–7563

Books and Reviews

Tse ST, Rice JR (1986) Crustal earthquake instability in relation to the depth variation of frictional slip properties. J Geophys Res 91:9452–9472

Cryosphere Models

Roger G. Barry
NSIDC, CIRES, University of Colorado, Boulder, USA

Article Outline

Glossary
Definition of the Subject

Introduction
Snow Cover
Floating Ice
Glaciers
Ice Sheets
Frozen Ground and Permafrost
Future Directions
Bibliography

Glossary

Cryosphere All forms of terrestrial snow and ice.

Newtonian viscous body A body whose stress at each point is linearly proportional to its strain rate at that point.

Definition of the Subject

The cryosphere comprises all terrestrial forms of snow and ice – snow cover, floating ice, glaciers, ice sheets, frozen ground and permafrost. It is a critical element of the climate system because of its high reflectivity, its insulating effects on the land and ocean, and its storage of water on short and long time scales. Numerical models of components of the cryosphere have been developed over the last 30 years or so, and some elements of these are now incorporated in coupled climate models and earth system models.

Introduction

Currently there are no comprehensive models of the entire cryosphere. Rather there is a wide range of models of components of the cryosphere – snow cover, floating ice, glaciers, ice sheets, frozen ground and permafrost – and various components are treated with varying degrees of detail in coupled atmosphere-ocean-land models. Cryospheric processes are generally parametrized in such earth system models.

Models developed for each of the main cryospheric components are discussed in turn.

Snow Cover

Snow cover is observed in situ at hydromeoteorlogical stations, from daily depth measurements, (monthly) snow courses and in special automated networks such as the western United States Snow Telemetry (SNOTEL) network of snow pressure pillows. Its extent is also observed and mapped daily (since June 1999) over the Northern Hemisphere from operational satellites of the National Oceanic and Atmospheric Administration (NOAA) in the USA. Snow covers about 47 million km^2 at maximum in January and there is only a small area in South America in July. Hemispheric snow water equivalent estimates are routinely made from passive microwave data (1979-present) with a 25-km resolution (Armstrong et al. [4]).

There are numerous models of the formation and disappearance of snow cover. Many have a hydrological focus aimed at estimating seasonal runoff. Some use simple temperature degree-day formulations while others incorporate a full energy balance calculation. Dozier and Painter [14] examine the use of multispectral and hyperspectral remote sensing to estimate the snow's spectral albedo, along with other properties such as grain size, contaminants, temperature, liquid water content, and depth or water equivalent.

The U.S. Army Cold Regions Research and Engineering Laboratory Model SNTHERM (SNow THERmal Model) is a 1-D energy balance model for snow and soil that is forced by meteorologically determined surface fluxes [41]. It simulates in-snow properties and processes, such as heat conduction, water flow, melt, vapor flow, compaction, grain growth, and solar absorption (see Fig. 1). The output provides snow depth, profiles of snow temperature, water content, density, grain size, and surface fluxes of sensible heat and evaporation. Surface boundary conditions require: incoming solar and longwave radiation; wind speed, air temperature and humidity at some reference height; and precipitation. The model will estimate solar and longwave radiation from cloud cover, if data on these variables are not available. Lower boundary conditions include soil textural properties (currently clay or sand used as defaults), wetness and temperature profile.

A comparative study of three snow models with different complexities was carried out by Jin et al. [39] to assess how a physically detailed snow model can improve snow modeling within general circulation models. The three models were (a) SNTHERM; (b) a simplified three-layer model, Snow–Atmosphere–Soil Transfer (SAST), which includes only the ice and liquid-water phases; and (c) the snow submodel of the Biosphere –Atmosphere Transfer Scheme (BATS), which calculates snowmelt from the energy budget and snow temperature by the force–restore method. SNTHERM gave the best match to observations with the SAST simulation being close. BATS captured the major processes in the upper layers of a snow pack where solar radiation is the main energy source and gave satisfactory seasonal results.

CROCUS is a model of the Centre d'Etudes de la Neige, Grenoble [10]. It is a 1-D physical model that determines mass and energy balance for a snow cover and is used for operational avalanche forecasting. The snow cover is represented as a pile of layers parallel to the

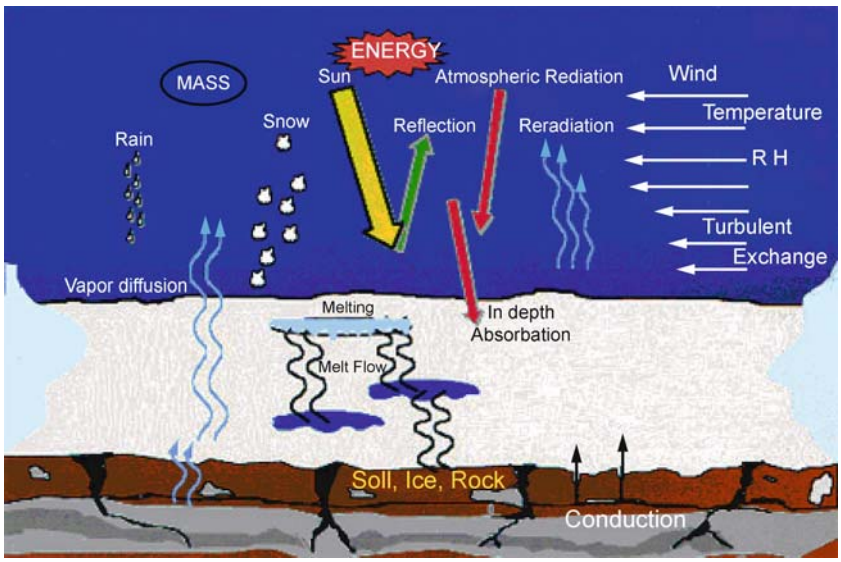

Cryosphere Models, Figure 1
The snowpack energy balance as characterized by SNTHERM (US Army Corps of Engineers)

ground. Energy exchanges are projected orthogonally to the slope. The model describes the evolution of the internal state of the snow cover as a function of meteorological conditions. The variables describing the snow cover are temperature, density, liquid water content, and snow type of each layer. To match the natural layers, the thickness and number of layers are adjusted by the model. The model simulates the heat conduction, melting/refreezing of snow layers, settlement, metamorphism, and percolation. It simulates dry and wet snow metamorphism with experimental laws derived from laboratory data. Snow grains are characterized by their size and type. This allows an accurate albedo of the snow cover to be calculated.

Bartelt and Lehning [7] also present a 1-D physical model of the snow pack (SNOWPACK) with equations for heat transfer, water transport, vapor diffusion and mechanical deformation. New snow, snow drift and ablation are treated. The snow layers are treated in terms of height, density and microstructure (grain size, shale and bonding). The model is used for avalanche warnings in Switzerland.

Interception Models

A physically-based snowfall interception model that scales snowfall interception processes from branch to canopy is now available [25]. It takes account of the persistent presence and subsequent unloading of intercepted snow in cold climates. To investigate how snow is intercepted at the forest stand scale, measurements of wind speed, air temperature, above- and below-canopy snowfall, accumulation of snow on the ground and the load of snow intercepted by a suspended, weighed, full-size conifer were collected from spruce and pine stands in the southern boreal forest. Interception efficiency is found to be particularly sensitive to snowfall amount, canopy density and time since snowfall. Further work resulted in process-based algorithms describing the accumulation, unloading and sublimation of intercepted snow in forest canopies (Pomeroy et al. [70]). These algorithms are unique in that they scale up the physics of interception and sublimation from small scales, where they are well understood, to forest stand-scale calculations of intercepted snow sublimation.

Blowing Snow Models

Physically-based treatments of blowing snow and wind are used to develop a distributed model of blowing snow transport and sublimation over complex terrain for an Arctic tundra basin by Essery et al. [18]. A reasonable agreement with results from snow surveys is obtained when sublimation processes are included. Sublimation typically removes 15–45% of the seasonal snow cover. The model is able to reproduce the distributions of snow mass, classified by vegetation type and landform, which can be approximated by lognormal distributions. The representation used for the downwind development of blowing snow with changes in wind speed and surface characteristics is shown to have a moderating influence on snow redistribution.

Spatial fields of snow depth have power spectra in one and two dimensions that occur in two frequency intervals separated by a scale break between 7 and 45 m [84]. The break in scaling is controlled by the spatial distribution of vegetation height when wind redistribution is minimal and by the interaction of the wind with surface concavities and vegetation when wind redistribution is dominant.

In mountainous regions, wind plays a prominent role in determining snow accumulation patterns and turbulent heat exchanges, strongly affecting the timing and magnitude of snowmelt runoff. Winstral and Marks [90] use digital terrain analysis to quantify aspects of the upwind topography related to wind shelter and exposure. They develop a distributed time-series of snow accumulation rates and wind speeds used to force a distributed snow model. Terrain parameters were used to distribute rates of snow accumulation and wind speeds at an hourly time step for input to ISNOBAL, an energy and mass balance snow model. ISNOBAL forced with accumulation rates and wind fields generated from the terrain parametrizations accurately models the observed snow distribution (including the formation of drifts and scoured wind-exposed ridges) and snowmelt runoff. By contrast, ISNOBAL forced with spatially constant accumulation rates and wind speeds taken from the sheltered meteorological site at Reynolds Mountain in southwest Idaho, a typical snow-monitoring site, overestimated peak snowmelt inputs and tended to underestimate snowmelt inputs prior to the runoff peak.

Avalanche Models

Avalanches range in size from sluffs with a volume of $< 10\,m^3$ to extreme releases of 10^7–$10^8\,m^3$; corresponding impact pressures range from $< 10^3$–10^6 Pa. There are two main types – loose snow avalanches and slab avalanches. Commonly, they begin with the failure of snow layers with densities less than $300\,kg\,m^{-3}$. An avalanche path comprises a starting zone, the track, and a runout–deposition zone. Loose snow avalanches are initiated when the angle of repose is exceeded – about 45°. The angle increases as temperatures rise due to increased cohesion. Slush avalanches can occur on slopes <10°. Downslope propagation continues to the kinetic angle of repose at about 17° [52,69]. Slab avalanches occur when a cohesive slab is released over an extensive plane of weakness on slopes of 35–40°. Slab thicknesses are 0.1–4 m and have a mean density of $\sim 200\,kg\,m^{-3}$. Bed surface temperatures are near 0°C.

The variables of interest for forecasting are velocity, run-out distance and impact pressure. The acceleration of an avalanche is resisted by surface friction, air drag at the front and upper boundary, and ploughing at the advancing front and underneath surface. According to Perla [69], maximum velocities range from 20–$30\,m\,s^{-1}$ for path lengths up to 500 m and slope angles of 25–35°. The mean run-out length on 67 Colorado avalanche paths was 380 m (Bovis and Mears [9]). On occasion, the run-out may cross a valley floor and continue up the facing slope. Impact pressures are a maximum at 1–2 m above the surface and range in value from about 1–10×10^5 Pa (Perla [69]).

Land Surface Schemes in GCMs

Snow cover is treated in Land Surface Models (LSMs), but snow and ice albedo parametrizations differ widely in their complexity [5]. A Snow Model Intercomparison was conducted using 24 snow cover models developed in ten different countries [17]. The models differ as to being multilayer or not, the inclusion of a soil model, variable heat conductivity, variable snow density, and the treatment of liquid storage. Only four of the models met all five criteria.

27 atmospheric general circulation models (GCMs) were run under the auspices of the Atmospheric Model Intercomparison Project (AMIP). The AMIP models reproduce a seasonal cycle of snow extent similar to the observed cycle. However, GCMs tend to underestimate autumn and winter snow extent (especially over North America) and overestimate spring snow extent (especially over Eurasia). The majority of models displays less than half of the observed interannual variability. No temporal correlation is found between simulated and observed snow extent, even when only months with extremely high or low values are considered [20]. The second generation AMIP-II simulations gave better results [21].

Floating Ice

Lake ice formation is dependent on the density characteristics of fresh water, which reaches a maximum density at 4°C. As a water body cools in the autumn it becomes isothermal at 4°C. Further cooling of the surface allows a less dense layer to form and eventually frazil ice or sheet ice forms depending on the wind conditions. Snow accumulation on lake ice depresses the ice surface below the water level, causing the snow to become saturated and leading to the formation of white snow-ice (in contrast with the black lake water ice). In rivers the flow motion leads to frazil ice, which builds up into pancakes. A 1-D energy balance model of lake ice growth is described by Liston and Hall [47] that treats lake-ice freeze-up, break-up, total ice thickness and ice type (Fig. 2). The model is

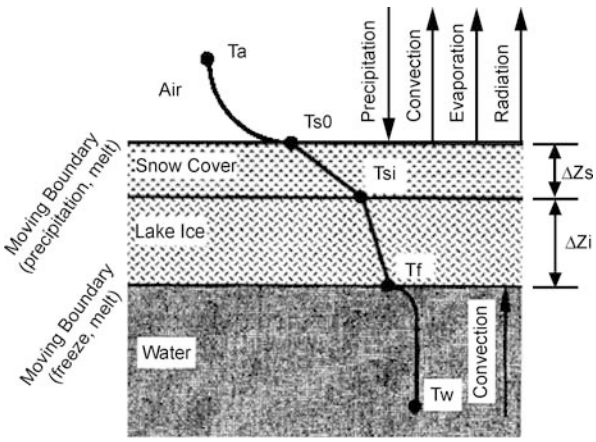

Cryosphere Models, Figure 2
The elements of a model of lake ice [47]

forced by daily atmospheric data on precipitation, wind speed and air temperature.

Sea ice grows thermodynamically by freezing of sea water at near −1.8°C due to the salinity, by the accumulation of snow cover on its surface, and by dynamic processes such as ridging and rafting. It decays thermodynamically, by wave action and by export to areas of warmer ocean. Ice types include new ice, young ice, first year and multi-year ice (World Meteorological Organization [91]). Typical first year ice thickness in the spring in the Arctic is about 1.5–2 m; multi year unridged ice may be 3–4 m thick. Ridging produces keels that may extend to 20–30 m depth. Ice draft (below the sea surface) is measured by upward looking sonars that are moored or deployed on submarines in the Arctic. Ice extent and concentration – the fractional coverage of ice – are determined by aerial reconnaissance, and primarily by satellite remote sensing (optical, passive microwave, synthetic aperture radar, scatterometry and laser altimetry). Weekly or 10-day charts of ice conditions are produced by national operational ice services (see [94]).

Modeling sea ice in either a stand-alone model or a GCM involves the solution of the following equations [19]:

i. for momentum, to obtain the ice velocity fields;
ii. for thermodynamic processes to obtain net ice growth/melt; and
iii. conservation equations including deformation ands transport of ice, plus the thermodynamic sources and sinks.

Ice dynamics is based on five stresses: wind stress, water stress, internal ice stress, Coriolis force, and the stress from the tilt of the sea surface. The Coriolis force and the tilt term are an order of magnitude less than the other three terms. The air and water stresses assume a constant turning angle of 25° in the Arctic and −25° in the Antarctic [28]. Internal ice stress is highly variable depending on ice conditions. It can be negligible when the ice cover is not compact and there are "free-drift" conditions, but it can be the largest force when there is thick, compact ice cover. The force due to ice resistance to deformation involves the relationship between stress and strain rate, which is termed the rheology (Flato [19]). Early work assumed that stress is linearly dependent on strain rate as in a linear viscous fluid [12]. Pritchard et al. [71] used an elastic-plastic rheology where the stress is linearly dependent on strain up to a yield strength where failure occurs. Hibler [27] developed a viscous-plastic model with an elliptical yield curve; the pre-yield stress states are linearly related to the strain rate. Advances by Hunke and Dukowicz [35] address the response of the ice on the timescales associated with wind forcing through an elastic viscous-plastic (EVP) rheology. The model was modified so that it reduces to the viscous–plastic model at these timescales, whereas at shorter timescales the adjustment process takes place by a mathematically efficient elastic wave mechanism. Recently, Lagrangian [33,45] sea ice models using a granular rheology have been developed (Tremblay and Mysak [83]; Overland et al. [65]). They have advantages in that they model individual sea ice "floes", but are also computationally intensive and are still in their infancy.

Ice dynamics have been extensively treated by Hibler [27]. He couples the dynamics to the ice thickness characteristics by allowing the ice interaction to become stronger as the ice becomes thicker and/or contains a lower area percentage of thin ice. The dynamics in turn causes high/low oceanic heat losses in regions of ice divergence/convergence. The ice is considered to interact in a plastic manner with the plastic strength depending on the ice thickness and concentration. These in turn evolve according to continuity equations that include changes in ice mass and percent of open water due to advection, ice deformation and thermodynamic effects. Anisotropic dynamic behavior of sea ice has also been investigated [13,30], though such approaches are computationally intensive and currently are not commonly used in models. The standard model treats sea ice as a visco–plastic material that flows plastically under typical stress conditions but behaves as a linear viscous fluid where strain rates are small and the ice becomes nearly rigid. The standard viscous–plastic model has poor dynamic response to forcing on a daily timescale. Models do not generally account for high-frequency (sub-daily) inertial

and tidal effects on dynamics, though research has shown that such effects can be important in the evolution of the ice cover [26,43]. The thermodynamics and dynamics are coupled through the ice thickness distribution. Essentially, deformation leads to pressure ridging and the formation of open water areas while thermodynamic processes act to ablate ridges and remove open water by ice formation in winter and create thinner ice/open water in summer. Thus, deformation acts to spread out the thickness distribution by promoting thick and thin ice categories while thermodynamic processes work towards a central ice thickness value [28].

Sea ice models typically feature processes of ice thermodynamics and dynamics although the earliest studies essentially used only thermodynamic ice growth and decay. The steady state Stefan relationship is written after Hibler and Flato [29]:

$$\rho_{\mathrm{I}} L \frac{\mathrm{d}H}{\mathrm{d}t} \approx \frac{k_{\mathrm{i}}}{H}(T_{\mathrm{m}} - T_{\mathrm{B}}) ,$$

where L = latent heat of fusion, T_{m} = melting point of the ice, T_{B} = upper boundary temperature of the ice, H = ice thickness, k_{i} = ice conductivity and ρ_{I} = ice density. Ice growth/melt at the underside is a result of the difference between the upward ocean heat flux and the heat conducted away from the ocean/ice interface into the ice. The first 1-D model of sea ice thermodynamics was developed by Maykut and Untersteiner (1971) [51]. A fuller treatment was made by Parkinson and Washington [66]. The model had four layers – ice, snow, ocean, and atmosphere – and 200 km horizontal resolution.

The incorporation of detailed thermodynamic processes includes the presence of snow on the sea ice, leads and polynyas, melt ponds, the effect of internal brine-pocket melting on surface ablation, the storage of sensible and latent heat inside the snow-ice system, and the transformation of snow into slush ice when the snow-ice interface sinks below the waterline due to the weight of snow. Models with enthalpy conservation improve the thermodynamic component of sea ice models [8]. These are starting to be included in larger-scale climate models.

An intermediate one-dimensional thermodynamic sea ice model developed by Ebert and Curry [15] includes leads and a surface albedo parametrization that interacts strongly with the state of the surface, and explicitly includes meltwater ponds (see Fig. 3). Four important positive feedback loops were identified: (1) the surface albedo feedback, (2) the conduction feedback, (3) the lead-solar flux feedback, and (4) the lead fraction feedback. The destabilizing effects of these positive feedbacks were mitigated by two strong negative feedbacks: (1) the outgoing longwave flux feedback, and (2) the turbulent flux feedback. A review of thermodynamic models is given by Steele and Flato [80].

Conservation equations are needed for ice area (concentration) and ice volume (thickness).

$$\partial h/\partial t = -\nabla \bullet (\boldsymbol{u} h) + S_h$$
$$\partial A/\partial t = -\nabla \bullet (\boldsymbol{u} A) + S_A ,$$

where $\boldsymbol{u}$ is the ice velocity vector and S_h and S_A are source terms for mean ice thickness and concentration, respectively. The second equation must also have the constraint that the area $A \leq 1$. The ice strength is parametrized only in terms of ice thickness h and A [27].

An atmospheric GCM was coupled to a global 1-degree, 20-level ocean GCM with dynamic and thermodynamic sea ice by Washington and Meehl [86] and run with increasing atmospheric CO_2. The Coupled Model Intercomparison Project (CMIP) allows a comparison of predicted Arctic sea ice [53]. Of the 12 models, only seven include sea ice motion and only four of these have a prognostic solution to the momentum equation. Apart from errors and approximations in the sea ice representation, the models also suffer from errors in the atmospheric and oceanic forcing fields. While the northern hemisphere ice extent in winter is well simulated overall, the ice thickness does not capture the proper spatial distribution with thicker ice toward North America and Greenland and thinner ice in the Eurasian basin. The simulations for the southern hemisphere show a wider range of extents and thickness. Flato [19] examines the sea ice extent simulated by two GCMs for AD 1900–2100 with the 'business as usual' scenario of greenhouse gases and aerosol concentrations. Both show a progressive decrease in ice in both hemispheres although the two models differ significantly in the initial southern hemisphere ice extent.

Martin and Gerdes [50] make a comparison of sea ice drift results from different Arctic Ocean Model Intercomparison Project (AOMIP) sea ice-ocean coupled models and observations for 1979–2001. The models are capable of reproducing realistic drift pattern variability. However, one class of models has a realistic mode at drift speeds around 3 cm s^{-1} and a short tail toward higher speeds. Another class shows unrealistically a more even frequency distribution with large probability of drift speeds of 10 to 20 cm s^{-1}. Reasons for these differences lie in discrepancies of wind stress forcing as well as sea ice model characteristics and sea ice-ocean coupling. Hunke and Holland [36] underscore the sensitivity of Arctic sea ice and ocean to small changes in forcing parameters. A comparison of three sets of forcing data, all variants of

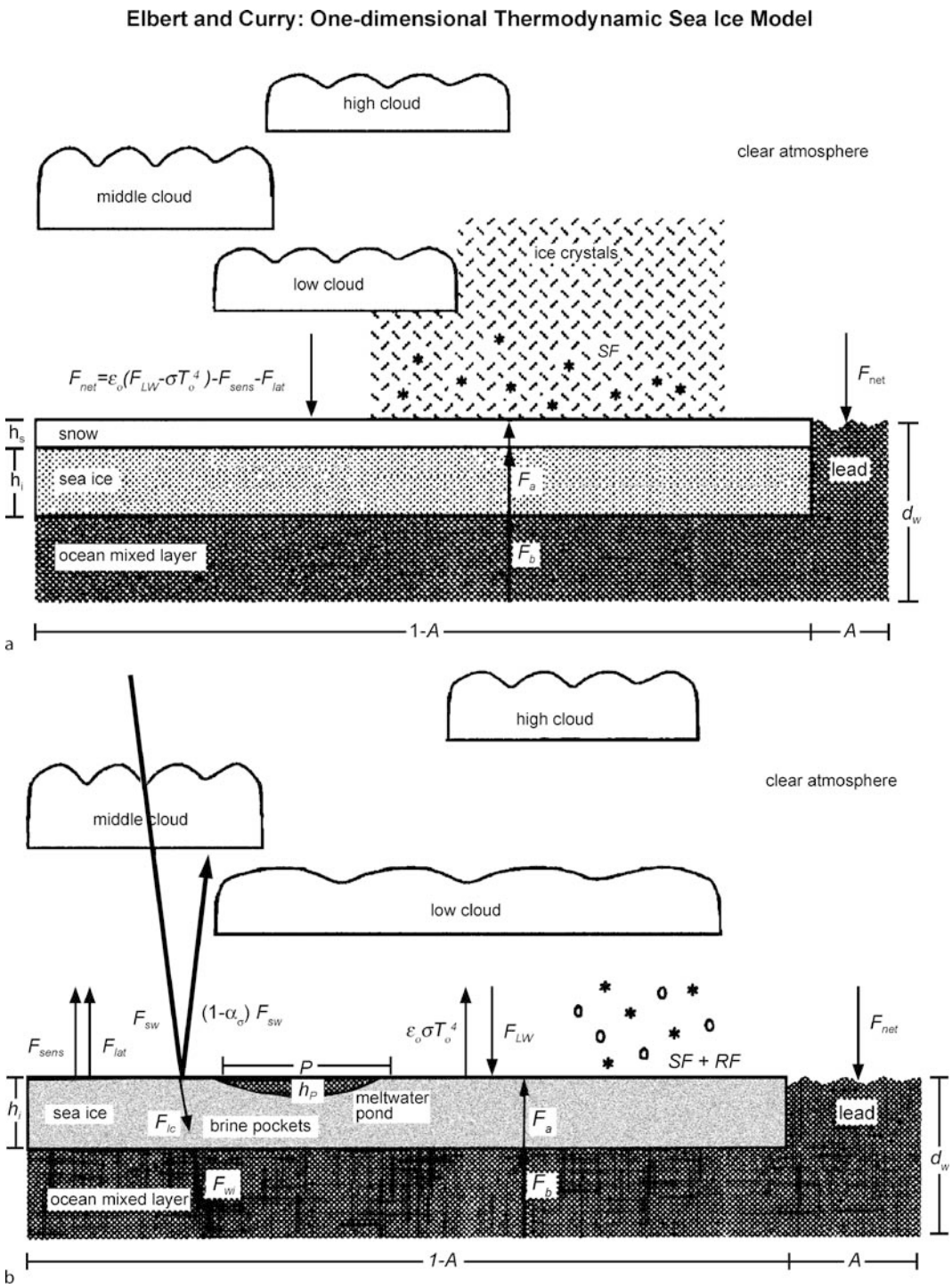

Cryosphere Models, Figure 3
The configuration of a one-dimensional thermodynamic model of sea ice a winter; b summer [15]

National Centers for Environmental Prediction (NCEP) forcing, give significant differences in ice thickness and ocean circulation using a global, coupled, sea ice-ocean model.

A study of GCMs used for the IPCC Fourth Assessment Report shows that while they produce reasonably similar ice extents in the Arctic, their equilibrium ice thickness values have a wide range due to differences in downwelling infrared radiation [16]. Holland et al. [32] found that in some scenarios of future CO_2 concentrations the sea ice cover can respond non-linearly with large decreases in extent within only 5–10 years, indicating that the current fitted to observations linear trends may not hold in the future. Stroeve et al. [81] showed that the IPPC models substantially underestimate the observed decline in Arctic sea ice extent compared to observations over the past 50 years. Hence their application in future scenarios is questionable.

Johnson et al. [40] examine the simulated sea ice concentration from nine ice-ocean numerical models in the AOMIP. The models have similar characteristics in winter (100% cover is produced), and most models reproduce an observed minimum in sea ice concentration for September 1990.

An assessment of coupled climate models with respect to the development of Arctic sea ice thickness during the 20th century is examined by Gerdes and Koeberle [22]. Model behavior is compared with results from an ocean–sea ice model using the AOMIP atmospheric forcing for the period 1948–2000. The hindcast exhibits virtually no trend in Arctic ice volume over its integration period 1948–2000. Most of the coupled climate models show a negative trend over the 20th century that accelerates towards the end of that century.

Glaciers

Glaciers are built up from snow that persists over many years. Initial densification leads to firn (densities of 400–830 kg m^{-3}) and at some depth, where the air passages between grains are sealed off (∼15–70 m according to wetness), to glacier ice with a density of 830–917 kg m^{-3} [67]. The glacier has upper accumulation and lower ablation areas, that are annually varying, and the ice slowly flows downhill towards the glacier terminus. Some glaciers occasionally display surges when the ice advances rapidly for a year or two and then stabilizes or retreats.

Glacier models consider either the mass balance and the rate of change of total mass, or the glacier dynamics and interactions between the ice and the bed. The flow velocity is modeled along the centerline of the glacier.

Glacier flow is determined from a relation between the shear strain rate (ε_{xy}) and shear stress (τ_{xy}) known as Glen's flow law [23]:

$$\dot{(\varepsilon_{xy})} = A\tau_{xy}^{n} ,$$

where $n \sim 3$. A depends on ice temperature, impurities and crystal orientation.

Recommended values of A decrease from 6.8×10^{15} s^{-1} kPa^{-3} at 0°C to 3.6×10^{-18} at –50°C (see Table 5.2 in [67]). Stress causes ice to deform by extension/compression, and shear leading to rotation.

Ice flows only by internal deformation when the bed is frozen, but where temperate conditions exist at the base, sliding becomes important. The sliding law relates basal velocity, shear stress, water pressure and the glacier bed characteristics. Weertman's [88] theory of sliding involved regelation and plastic deformation. Regelation operates over small bumps in the bed (<1 m dimension). All the ice is at pressure melting point. There is excess pressure on the upstream side of the bump so that the ice there is colder than on the downstream side. This causes heat to flow towards the upstream side through the bump and surrounding ice. The heat transferred melts ice on the upstream side and melt water flows around the bump, refreezing on the downstream side because it is colder than the ice there. Ice also deforms plastically. Near a bump, the longitudinal stress in the ice and, therefore, the strain are above average. The greater the distance over which the stress is enhanced, the greater is the ice velocity. This mechanism works best over larger bumps. Both processes are equally effective at the "controlling obstacle" size, about 1–10 cm.

Paterson [67] shows that the sliding velocity:

$$u = \text{constant}\,(\tau^{0.5}/R)^4 ,$$

where R = roughness and τ = basal shear stress.

Water from surface ablation penetrates to the glacier bed and has been shown to lift the ice by as much as 40 cm on the Unteraargletscher, Switzerland (Iken et al. [38]). During rapid uplift events the glacier velocity increased 3–6 times. When the water pressure at the bed exceeds a certain value (the separation pressure) that depends on the bed roughness, cavities form in the lee of bumps. When the water pressure exceeds a second critical value, sliding becomes unstable.

A numerical ice flow model has recently been used to study the advance of tidewater glaciers into a deep fiord [56]. The results suggest that irrespective of the calving criterion and the accumulation rate in the catchment, the glacier cannot advance into deep water (>300 m) unless sedimentation at the glacier front is included.

Using a first-order theory of glacier dynamics, Oerlemans [61] related changes in glacier length to changes in air temperature. He constructed a temperature history for different parts of the world from 169 records of glacier length. The reconstructed warming in the first half of the 20th century is 0.5°C. The warming signals from glaciers at low and high elevations appear to be very similar.

Ice Sheets

There are currently two major ice sheets in Greenland and Antarctica, but during the Last Glacial Maximum about 20,000 years ago there were massive ice sheets over northern North America and Fennoscandia. The Antarctic ice sheet covers 12.4 million km^2 and reaches thicknesses in excess of 4500 m; it represents a sea level equivalent of 64.3 m. The Greenland ice sheet has an area of 1.8 million km^2 and a maximum thickness of 3200 m; it has a sea level equivalent of 7.5 m. Both ice sheets have a parabolic profile and gentle slopes (1–2°) away from the margins. Greenland has a suite of snow facies; from the interior outward these are: the dry-snow zone, percolation zone, wet-snow zone, superimposed ice zone, and the ablation area. The equilibrium line is between the ablation area and the higher zones. At its base, Greenland is close to sea level except for fringing coastal mountains, through which the ice reaches the sea in some 20 major outlet glaciers. Antarctica is mostly dry snow and is little affected by melting except at the margins and in the Antarctic Peninsula. In the Antarctic there are major ice shelves (e. g. the Ross Ice Shelf and the Filchner–Ronne Ice Shelf) that buttress large sections of the ice sheet (see Fig. 4) and extend around 44% of the coastline of the continent.

The major problems in ice sheet dynamics, following Paterson (see p. 238 in [66]), are (i) to calculate the distribution of ice thickness and velocity that will maintain a steady state, given the accumulation and ablation rates and surface temperature. The flow parameters, ice thickness at the ice divide, and geothermal heat flux must be specified. Flow lines, the time that ice takes to travel along them, and the age of the ice at different depths, can be calculated from the velocity distribution. (ii) to determine how the system will react to changes in accumulation, ablation or surface temperature.

The earliest ice sheet models assumed that ice deformed as a Newtonian viscous body. Orowan [64] and Nye [59] assumed that ice behaved as a perfectly plastic material, but Glen [23] established the relationship between strain rate and stress in ice as non-linear (see above).

The surface profile of an ice sheet on a horizontal bed of half width L, thickness h and thickness at the centre H is:

$$h^2 = \frac{2\tau_0}{\rho g}(L - x) ,$$

where ρ = density, g = acceleration due to gravity, τ_0 = the basal shear stress (∼0–100 kPa), and $(L - x)$ is the distance from the edge measured along a flow line. The equation describes a parabola. The ice thickness at the centre is $H = (2\tau_0 L/\rho g)^{0.5}$.

The mass balance (B) can be expressed in a mass conservation equation as:

$$B = \frac{\partial q}{\partial x} : + \frac{\partial h}{\partial t} ,$$

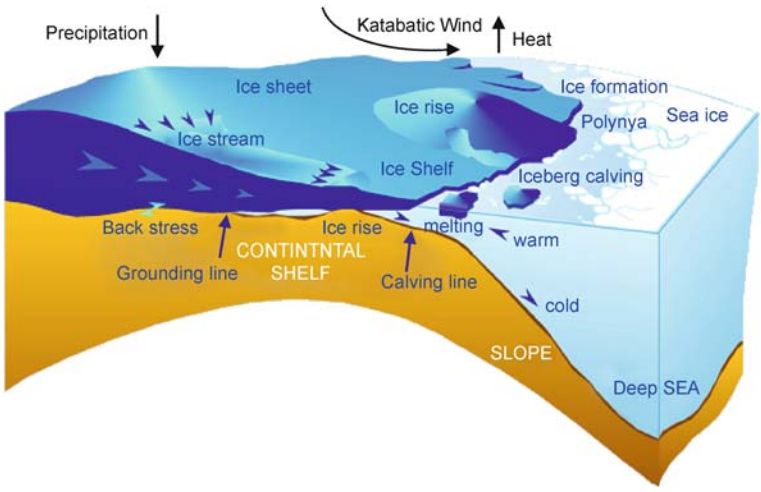

Cryosphere Models, Figure 4
Schematic of an ice sheet and shelf showing the processes at work (after Wikipedia)

where h = ice thickness, t = time, and the flux $q = h\,\boldsymbol{u}$, where $\boldsymbol{u}$ is the velocity averaged over the ice thickness (see p. 246 in Paterson [67]).

If the flow term $\partial q/\partial x$ is small, then the surface elevation will vary in response to the local accumulation/ablation, which will determine the profile. 'Balance velocities' are steady state velocities that are calculated from accumulation rate and ice thickness. Paterson [67] shows that for a 1000 km radius circular ice sheet, of perfectly plastic ice with a yield stress of 199 kPa, an accumulation rate of 150 mm of ice/year, and ablation by iceberg calving at the margin, the ice in the centre would be 4700 m thick. Balance velocities would increase from 1,5 m a^{-1} at 100 km from the centre to 45 m a^{-1} at 900 km and the travel time for ice to move from the center to the edge would be 150,000 years.

In Greenland and Antarctica much of the ice transport is accomplished by fast flowing ice streams – regions where the ice flow is much faster than on either side. Most occupy deep channels with beds below sea level and terminate either as a floating glacier tongue (an outlet glacier) or become part of an ice shelf. Morgan et al. [54] indicate that while ice streams and outlet glaciers account for only 13% of the coastline of Antarctica, they drain about 90% of the accumulation of the interior.

A hierarchy of land ice models is presented by van der Veen and Payne [85]. The simple lamellar flow model, involves a balance between driving stress and basal drag. The surface and bed topography must be nearly level for lamellar flow, which is a good approximation to conditions in the interior of an ice sheet. In cases where an ice stream is bounded by a rock wall or stagnant ice on one or both sides, lateral drag needs to be incorporated. The proportion of driving stress that is supported by drag at the bed is termed the shape factor [60]; it is less than one for narrow ice streams. An important issue in Antarctica is the interaction between the ice sheet and ice shelves. The peripheral ice shelves are thought to exert a back stress that stabilizes the inland ice sheet where it is grounded below sea level, as in most of West Antarctica [82]. Where the bedrock slopes down towards the ice sheet interior, the grounding line is unstable; If the grounding line initially retreats, the ice at the grounding line becomes thicker due to the bedrock slope, and the creep thinning (thinning associated with along-flow gradients in the ice velocity) increases causing the grounding line to retreat further – a positive feedback. The dynamics of ice sheet grounding lines is examined by Schoof [76]. A boundary layer theory for ice flux through the transition zone shows that the flux increases sharply with ice thickness at the grounding line. He finds that marine ice sheets have well-defimed, discrete equilibrium profiles, and steady grounding lines cannot be stable on reverse bed slopes. Also, marine ice sheets with overdeepened beds may undergo hysteresis with variations in sea level, accumulation rate, bed slipperiness and ice viscosity.

Two general types of an ice sheet model have been developed. One is prognostic, based on the original work by Budd et al. [11]; the other category is diagnostic, addressing specific aspects of ice sheet processes. Prognostic models involve four sets of equations (van der Veen and Payne [85]). These are: (i) diagnostic equations for the horizontal velocity components as functions of local ice geometry and ice rheology (Glen's law); (ii) prognostic equations for the evolution of internal ice temperature, given appropriate boundary conditions at the upper and lower ice surfaces; (iii) a diagnostic equation for ice vertical velocity via the divergence of the horizontal velocity; and (iv) a prognostic equation for ice thickness based on the snow accumulation, snow/ice melt and the divergence/convergence of horizontal ice flow. The effects of bedrock depression under the changing weight of the ice load must also be taken into account. Such models have been used to reconstruct ice sheet history over glacial cycles, as well as to assess the responses to future climate change.

Diagnostic models do not address time evolution of the ice sheet and treat the internal stress regime in much greater detail, particularly the contributions of longitudinal and lateral stresses. Recently, models have been developed that do not assume negligible vertical shear. Huybrechts and de Wolde [38] have combined prognostic model elements with a detailed diagnostic model to study the multi-century behaviour of the Antarctic and Greenland ice sheets. A fully dynamic 3-D thermo-mechanical ice sheet model was coupled to a two-dimensional climate model.

A model validation exercise was undertaken by the European Ice Sheet Modeling Initiative (EISMINT). Payne et al. [68] examined the effects of thermo-mechanical coupling while MacAyeal et al. [49] test ice shelf models for the Ross Ice Shelf. Overall, the models agreed in the main features that were simulated.

Frozen Ground and Permafrost

The surface layers of soil and rock may be seasonally or perennially frozen. Perennially frozen ground or permafrost is frozen for at least two successive summers. The ground need not contain ice but may be rock below 0°C. Ground ice may be segregated, in veins, or massive in occurrence. The spatial extent of permafrost ranges from

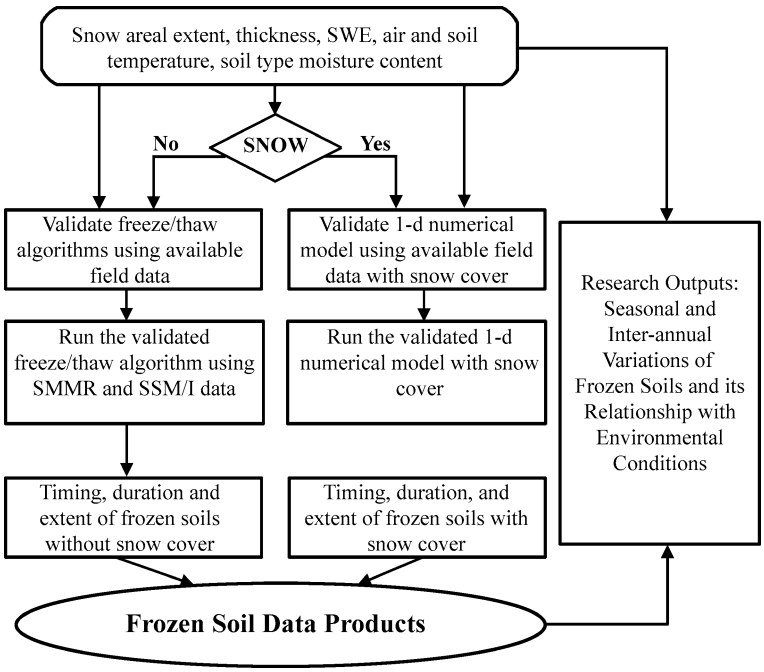

Cryosphere Models, Figure 5
The flowchart of a model processing the effects of snow cover on frozen soil, used to study the timing, duration, number of days, and areal extent of near-surface soil freeze/thaw status [92]

continuous (>90% of the surface is underlain), to discontinuous (50–90% of the surface), sporadic (10–50%) and isolated (<10%). Continuous permafrost is associated with mean annual temperatures below about −7°C. Its thickness ranges from a few meters up to 1500 m in Yakutia. Subsea permafrost also occurs offshore in the Eurasian shelf seas and in the Beaufort Sea; it is a relic of previous glacial intervals when the sea bed was exposed by sea level lowering of up to 135 m under low temperature conditions.

Ground temperatures are largely determined by heat conduction, although in areas of seasonal freezing and discontinuous permafrost localized circulation of groundwater may need to be considered. The thermal properties of the ground vary with the mineral composition, organic content, moisture content (as vapor, water and ice), and temperature, as well as the overlying vegetation and snow cover. A frozen soil algorithm has been developed [92] to detect the near-surface soil freeze/thaw cycle over snow-free and snow-covered land in the United States (see Fig. 5).

The conductive heat transfer is given [89] as:

$$G = -K(\mathrm{d}T/\mathrm{d}z)$$

where K = the thermal conductivity ($W\ m^{-1}\ K^{-1}$).

For steady state conditions, the temperature at depth T_z is written:

$$T_z = T_s + (G/K)z$$

where T_s = surface temperature.

The heat conduction equation is:

$$\partial T/\partial t = \kappa \partial^2 T/\partial z^2$$

where κ = thermal diffusivity ($m^2\ s^{-1}$), the coefficient of heat diffusion. To simulate soil freezing and thawing processes, soil water phase change has to be considered (for details, see Lunardini [48]).

Permafrost models can be broadly classified as either equilibrium or process-based transient models according to their underlying methodology. Equilibrium models are based on empirical and semi-empirical relationships between permafrost occurrence and topoclimatic factors (altitude, slope, aspect) and mean air temperature, freezing and thawing indices, and snow cover. They are often used to predict the lateral "boundaries" of permafrost distribution, or to estimate "average" geocryological parameters. thawing indices, snow cover and solar radiation, for example [77]. A 'frost index' model was developed by Nelson and Outcalt [55] for the Arctic and has been applied to

mountain areas [31]. A wide spectrum of explicit equilibrium permafrost models is available to estimate the thickness of the active layer. The simplest approaches are based on several variations of the analytical Stefan solution to the heat conduction problem with phase change. These methods have been used to estimate regional-scale active layer thickness [56,79]. Kudryavtsev et al. [42] developed a more comprehensive equilibrium permafrost model with analytical solutions that has been adapted and used with Geographical Information System (GIS) technology to estimate active layer thickness at regional [74,78] and circum-arctic [2] scales. The advantages of equilibrium models are their relative simplicity and low data requirements. The major drawback of such models is their inability to resolve seasonal and inter-annual variability [77], which is frequently required for ecological and hydrological studies in the Arctic. These limitations have led to recent spatial adaptations of process-based transient numerical models.

Process-oriented transient models detail one-dimensional heat transfer in soils with phase change driven by either surface temperature [24] or the energy balance components [46]. They account for the major physical processes governing development of the ground thermal regime, simulate soil freezing/thawing processes, and provide insight into the response of soil thermal regime to changes in environmental conditions. These models can provide good results for simulating active layer thickness and permafrost temperatures when driven with known boundary conditions and forcing parameters measured at site-specific locations [73,93]. Recently, such one-dimensional heat transfer models with phase change have been used to simulate regional-scale soil thermal regime [62,63]. However, their adaptation from point-specific to regional-scale is not a straightforward process. It requires simplification, careful selection of climate forcing data, and treatment of surface and subsurface parameters with largely unknown distributions over the modeled domain.

Most GCMs do not treat permafrost dynamics. However, Nikolsky et al. [58] show that in the Community Land Model (CLM3) GCM improvements can be made to the representation of permafrost dynamics and their climate feedbacks. They do this by increasing the total soil depth by adding new layers, incorporating a surface organic soil layer, and modifying the model's numerical scheme to include unfrozen water dynamics and more realistic treatment of the model phase changes between ice and water.

The Community Climate System Model (CCSM) has a 5-layer snow model over a 10 layer 3.4 m deep soil model that treats thermal and hydrologic frozen soil processes. A projection made for the 21st century shows severe degradation of the permafrost in the Northern Hemisphere.

Some models address only the active layer that is the top layer of soil that thaws during the summer and freezes again during the autumn. The thaw depth can be analyzed by the Stefan solution for heat transfer in a medium with phase change (Anisimov et al. [3]):

$$z = [(2n\lambda tT)/(\rho wL)]^{0.5}$$

where z = active layer thickness (m), n = the ratio of seasonal ground surface and air temperature degree-day sums, λ = thermal conductivity of thawed soil (W m^{-1} K^{-1}), t = warm season duration > 0°C (s), T = mean warm season temperature (°C), ρ = soil density (kg m^{-3}), w = relative water content (decimal proportion) and L = latent heat of fusion (J kg^{-1}).

A dynamic 3-D terrain model is currently being developed and tested in Svalbard (Humlum [34]). The model takes topographic data, terrain surface characteristics (geomorphology and vegetation) and meteorological variables (air temperature, wind speed and direction, and cloud cover) as input and provides output on phenomena such as terrain surface net radiation balance, snow cover thickness and duration, glacier mass, active layer thickness, stable permafrost thickness and the amount of summer melt water discharge.

A stochastic model was developed by Anisimov et al. [3] and used to calculate the probability density function of active-layer thickness (ALT). Equations for the mean, variance, and higher moments of ALT were derived by applying stochastic averaging to a semi-empirical model of seasonal thawing. The stochastic model was applied in a case study in the Kuparuk River basin, north-central Alaska.

Shiklomanov et al. [77] compare three models of active layer thickness (ALT) for northern Alaska. One model (NSIDC) is very accurate in the topographically homogeneous Coastal Plain but overestimates (ALT) in the Brooks Range Foothills. The UAF-GIPL 2.0 model reproduced site-specific active layer values well but overestimated ALT on the Coastal Plain. Large differences in ALT fields mainly result from differences in model approaches for characterizing largely unknown spatial distribution of surface (vegetation, snow) and subsurface (soil properties and moisture) conditions.

Data set limitations are a major problem (Anisimov et al. [3]). A permafrost model, forced with available climate data sets, was used to calculate the large-scale char-

acteristics of permafrost in northern Eurasia. Zonal-mean air and ground temperatures, depth of seasonal thawing, and area occupied by near-surface permafrost in Eurasia north of 45° N were analyzed. The 0.5–1.0 °C difference in zonal-mean air temperature between the data sets translates into a 10–20% uncertainty in estimates of near-surface permafrost area, which is comparable to the extent of changes projected for the following several decades.

GCMs have been used to simulate changes in permafrost conditions with global warming. Anisimov and Nelson [1] were the first to study this. Most recently, Saito et al. [74] use a coupled global climate model at high horizontal resolution (0.5° land mesh) with a five-layer, 4.0 m deep soil to evaluate changes in the distribution of frozen ground and subsurface hydrothermal regimes under global warming. Two types of frozen ground were classified according to monthly soil temperatures: "permafrost" for regions with a maximum active layer thickness less than 4 m and "seasonally frozen ground." Approximately 60% of present-day permafrost would degrade into seasonally frozen ground by 2100 in the circum-Arctic basins.

Future Directions

In the next 5–10 years we can expect to see more comprehensive treatment of cryospheric processes in climate models. Already, steps are being taken to incorporate ice sheet processes and to enhance the treatment of frozen ground and permafrost. Increased model resolution will place new demands for cryospheric data sets for boundary conditions and as validation data. Mountain glaciers, as well as lake ice, will need to be incorporated especially in regional climate models.

Bibliography

Primary Literature

1. Anisimov OA, Nelson FE (1996) Permafrost distribution in the Northern Hemisphere under scenarios of climate change. Glob Planet Chang 14:59–72
2. Anisimov OA, Shiklomanov NI, Nelson FE (1997) Global warming and active-layer thickness: results from transient general circulation models. Glob Planet Chang 15:61–77
3. Anisimov OA, Shiklomanov NI, Nelson FE (2002) Variability of seasonal thaw depth in permafrost regions: a stochastic modeling approach. Ecol Model 153:217–227
4. Armstrong RL, Brodzik MJ, Knowles K, Savoie M (2005) Global monthly EASE-Grid snow water equivalent climatology. Digital media. National Snow and Ice Data Center, Boulder
5. Barry RG (1996) The parameterization of surface albedo for sea ice and its snow cover. Progr Phys Geog 20:61–77
6. Barry RG (2002) The role of snow and ice in the global climate system: A review. Polar Geog 24:235–246
7. Bartelt P, Lehning M (2002) A physical SNOWPACK model for the Swiss Avalanche Warning Services. Part I: Numerical model. Cold Reg Sci Technol 35(3):123–145
8. Bitz CM, Lipscomb WH (1999) An energy-conserving thermodynamic model of sea ice. J Geophys Res 105:15669–15677
9. Bovis MJ, Mears AI (1976) Statistical prediction of snow avalanche runout from terrain variables in Colorado. Arct Alp Res 8:115–120
10. Brun E, David P, Sudul M, Brunot G (1992) A numerical model to simulate snow cover stratigraphy for operational avalanche forecasting. J Glaciol 38:13–22
11. Budd WF, Jenssen D, Radok U (1971) Derived physical charcteristics of the Antarctic ice sheet. ANARE Interim Report Series A (IV) Glaciology Pobl
12. Campbell WJ (1965) The wind-driven circulation of ice and water in a polar ocean. J Geophys Res 70:3279–3301
13. Coon MD, Knoke GS, Echert DS, Pritchard RS(1998) The architecture of anisotropic elastic-plastic sea ice mechanics constitutive law. J Geophys Res 103(C10):21915–21925
14. Dozier J, Painter TH (2004) Multispectral and hyperspectral remote sensing of alpine snow properties. Annu Rev Earth Planet Sci 32:465–494
15. Ebert EE, Curry JA (1993) An intermediate one-dimensional thermodynamic sea ice model for investigating ice-atmosphere interactions. J Geophys Res 98(C6):10085–10110
16. Eisenman I, Untersteiner N, Wettlaufer JS (2007) On the reliability of simulated Arctic sea ice in global climate models. Geophys Res Lett 34:L10501, doi:10.1029/2007GL029914
17. Essery R, Yang Z-L (2001) An overview of models participating in the snow model intercomparison project (SnowMIP). In: 8th Scientific Assembly of IAMAS, Innsbruck. http://www.cnrm.meteo.fr/snowmip/. Accessed 22 Aug 2008
18. Essery R, Long L, Pomeroy JW (1999) A distributed model of blowing snow over complex terrain. Hydrol Process 13:2423–2438
19. Flato GM (2004) Sea-ice modelling. In: Bamber JL, Payne AJ (eds) Mass balance of the cryosphere: Observations and modelling of contemporary and future change. Cambridge University Press, Cambridge, pp 367–390
20. Frei A, Robinson DA (1995) Evaluation of snow extent and its variability in the Atmospheric Model Intercomparison Project. J Geophys Res 103(D8):8859–8871
21. Frei A, Miller JA, Robinson DA (2003) Improved simulations of snow extent in the second phase of the Atmospheric Model Intercomparison Project (AMIP-2). J Geophys Res 108(D12):4369, doi:10.1029/2002JD003030
22. Gerdes R, Koeberle C (2007) Comparison of Arctic sea ice thickness variability in IPCC Climate of the 20th Century experiments and in ocean–sea ice hindcasts. J Geophys Res 112(C4)C04S13
23. Glen J (1955) The creep of polycrystalline ice. Proc Roy Soc Lond A228:519–538
24. Goodrich LE (1982) The influence of snow cover on the ground thermal regime. Can Geotech J 19:421–432
25. Hedstrom N, Pomeroy JW (1998) Measurements and modelling of snow interception in the boreal forest Hydrol. Processes 12:1611–1525
26. Heil P, Hibler WD III (2002) Modeling the high-frequency component of Arctic sea ice drift and deformation. J Phys Oceanogr 32:3039–3057

27. Hibler WD III (1979) A dynamic-thermodynamic sea ice model. J Phys Oceanogr 9:815–846
28. Hibler WD III (2004) Modelling the dynamic response of sea ice. In: Bamber JL, Payne AJ (eds) Mass balance of the cryosphere: Observations and modelling of contemporary and future change. Cambridge University Press, Cambridge, pp 227–334
29. Hibler WD III, Flato GM (1992): Sea ice models. In: Trenberth K (ed) Climate System Modeling. Cambridge University Press, New York, pp 413–436
30. Hibler WD III, Schulson EM (2000) On modeling the anisotropic failure and flow of flawed sea ice. J Geophys Res 105(C7):17105–17120
31. Hoelzle M, Mittaz C, Etzelmueller B, Haeberli W (2001) Surface energy fluxes and distribution models of permafrost in European mountain areas: An overview of current developments. Permafr Periglac Process 12:53–68
32. Holland MM, Bitz CM, Tremblay H (2006) Future abrupt reductions in the summer Arctic sea ice. Geophys Res Lett 33:L23503. doi:10.1029/2006GL028024
33. Hopkins MA (1996) On the mesoscale interaction of lead ice and floes. J Geophys Res 101:18315–18326
34. Humlum O (2007) Modeling energy balance, surface temperatures, active layer depth and permafrost thickness around Longyeardalen, Svalbard. http://www.unis.no/research/geology/Geo_research/Ole/Modelling.htm. Accessed 22 Aug 2008
35. Hunke EC, Dukowicz JK (1997) An elastic–viscous–plastic model for sea ice dynamics. J Phys Oceanogr 27:1849–1867
36. Hunke EC, Holland MM (2007) Global atmospheric forcing data for Arctic ice-ocean modeling. J Geophys Res 112:C04S14
37. Huybrechts P, de Wolde J (1999) The dynamic response of the Greenland and Antarctic ice sheets to multiple-century climatic warming. J Climate 12:2169–2188
38. Iken A, Roethlisberger H, Flotron A, Haeberli W (1983) The uplift of the Unteraargletscher at the beginning ot the melt season – a consequence of water storage at the bed. J Glaciol 30:15–25
39. Jin J, Gao X, Yang Z-L, Bales RC, Sorooshian S, Dickinson RE, Sun SF, Wu GX (1999) Comparative analyses of physically based snowmelt models for climate simulations. J Climate 12:2643–2657
40. Johnson M, Gaffigan S, Hunke E, Gerdes R (2007) A comparison of Arctic Ocean sea ice concentration among the coordinated AOMIP model experiments. J Geophys Res 112:C04S11
41. Jordan R (1991) A one-dimensional temperature model for a snow cover. Technical documentation for SNTHERM Special Technical Report 91-16. US Army Cold Regions Research and Engineering Laboratory, Hanover
42. Kudryavtsev VA et al (1974) Fundamentals of frost forecasting in geological engineering investigations. Nauka, Moscow (in Russian). English translation US Armt Cold Regions Res Engr Lan, Hannover, Draft translation 1977
43. Kwok R, Cunningham GF, Hibler III WD (2003) Sub-daily sea ice motion and deformation from RADARSAT observations. Geophys Res Lett 30(23):2218 doi:10.1029/2003GL018723
44. Lawrence DM, Slater AG (2007) A projection of severe near-surface permafrost degradation during the 21st century. Geophys Res Lett 32:L24401
45. Lindsay RW, Stern HL (2005) A new Lagrangian model of Arctic sea ice. J Phys Oceanogr 34:272–283
46. Ling F, Zhang T-J (2004) A numerical model for surface energy balance and thermal regime of the active layer and permafrost containing unfrozen water. Cold Regions Sci Technol 38:1–15
47. Liston GE, Hall DK (1995) An energy-balance model of lake-ice evolution. J Glaciol 41(138):373–382
48. Lunardini V (1988) Freezing of soil with an unfrozen water content and variable thermal properties. US Army Cold Regions Res Engineering Lab, Hanover, p 31
49. MacAyeal DR et al (1996) An ice-shelf model test based on the Ross ice shelf. Antarct Ann Glaciol 23:46–51
50. Martin Y, Gerdes R (2007) Sea ice drift variability in Arctic Ocean Model Intercomparison Project models and observations. J Geophys Res 112(C4):C04S10
51. Maykut G, Untersteiner N (1971) Some results from a time-dependent thermodynamic mode; of sea ice. J Geophys Res 76:1550–75
52. McClung D, Schaerer P(2006) The Avalanche Handbook. The Mountaineers, Seattle
53. Meehl GA, Boer GA, Covet C, Latif M, Stouffer RJ (1997) Intercomparison makes for a better climate model. EOS 78:445–446
54. Morgan VI, Jacka TH, Akermasn GJ, Clarke AL (1982) Outlet glacier and mass budget studies in Enderby, Kemp and MacRobertson Lands, Antarctica. Ann Glaciol 3L:204–210
55. Nelson FE, Outcalt DSI (1987) A computational method for prediction and regionalization of permafrrost. Arct Alp Res 19:279–88
56. Nelson FE et al (1997) Estimating Active-Layer Thickness over a Large Region: Kuparuk River Basin, Alaska, USA. Arct Alp Res 29:367–378
57. Nick EM, van derr Veen CJ, Oerlemans J (2007) Controls on advance of tidewater glaciers: Results from numerical modeling applied to Columbia Glavier. J Geophys Res 112:G03S24
58. Nicolsky DJ, Romanovsky VE, Alexeev VA, Lawrence DM (2007) Improved modeling of permafrost dynamics in a GCM Land Surface Scheme. Geophys Res Lett 34(8):L08591
59. Nye J (1951) The flow of glaciers and ice sheets as a problem in plasticity. Proc Roy Soc Lond A 207:554–572
60. Nye J (1965) The flow of a glacier in a channel of rectangular, elliptic or parabolic cross-section. J Glaciol 5:661–690
61. Oerlemans J (2005) Extracting a climate signal from 169 glacier records. Science 308:675–677
62. Oelke C et al (2003) Regional-scale modeling of soil freeze/thaw over the Arctic drainage basin. J Geophys Res 108(D10):4314
63. Oelke C, Zhang T-J (2004) A model study of circum-Arctic soil temperatures. Permafr Periglac Process 15:103–121
64. Orowan E (1949) Remarks at the joint meeting of the British Glaciological Society, the British Rheologists Club and the Institute of Metals. J Glaciol 1:231–236
65. Overland JE, McNutt SL, Salo S, Groves J, Li S (1998) Arctic sea ice as a granular plastic. J Geophys Res 104(C10):21845–21867
66. Parkinson CL, Washington WM (1979) A large-scale numerical model of sea ice. J Geophys Res 84:311–337
67. Paterson WSB (1994) The physics of glaciers. Pergamon, Elsevier Science, New York, p 480
68. Payne AJ et al (2000) Results from the EISMINT Phase 2 simplofoed geometry experiments: the effevts of thermomechanical coupling. J Glaciol 46(153):227–238
69. Perla RI (1980) Avalanche release, motion, and impact. In: Colbeck SC (ed) Dynamics of snow and ice masses. Academic Press, New York, pp 397–462

70. Pomeroy JW, Parviainen J, Hedstrom N, Gray DM (1998) Coupled modelling of forest snow interception and sublimation. Hydrol Process 12:2317–2337
71. Pritchard RS, Coon M, McPhee MG, Leavitt E (1977) Winter ice dynamics in the nearshore Beaufort Sea. AIDJEX Bull.37, Applied Physics Lab, University of Washington, Seattle, pp 37–93
72. Raymond CF (1980) Temperate valley glaciers. In: Colbeck SC (ed) Dynamics of snow and ice masses. New York. Academic Press, pp 79–139
73. Romanovsky VE, Osterkamp TE, Duzbury NS (1997) An evaluation of three numerical models used in simulations of the active layer and permafrost temperature regimes. Cold Regions Sci Technol 26:195–201
74. Saito K, Kimoto M, Zhang T, Takata K, Emori S (2007) Evaluating a high-resolution climate model: Simulated hydrothermal regimes in frozen ground regions and their change under the global warming scenario. J Geophys Res 112:F02S11
75. Sazonava TS, Romanovsky V (2003) A model for regional-scale estimation of temporal and spatial variability of active layer thickness and mean annual ground temperatures. Permafr Periglac Proc 14:125–139
76. Schoof C (2007) Ice sheet grounding line dynamics: Steady states, stability, and hysteresis. J Geophys Res 112:F03S28
77. Shiklomanov NI et al (2007) Comparison of model-produced active layer fields: Results for northern Alaska. J Geophys Res 112(F2):F02S10
78. Shiklomanov NI, Nelson FE (1999) Analytic representation of the active layer thickness field, Kuparuk River Basin, Alaska. Eccol Model 123:105–125
79. Shiklomanov NI, Nelson FE (2002) Active-layer mapping at regional scales: a 13-year spatial time series for the Kuparuk region, north-central Alaska. Permafrost Periglac Proc 13:219–230
80. Steele M, Flato GM (2000) Sea ice growth and modeling: A survey. In: Lewis EL et al (eds) The freshwater budget of the Arctic. Kluwer, Dordrecht, pp 549–587
81. Stroeve J et al (2007) Arctic sea ice decline: Faster than forecast. Geophys Res Lett 34:L09501, doi:10.1029/2007GL029703
82. Thomas RH (1979) The dynamics of marine ice sheets. J Glaciol 24:167–177
83. Tremblay L-B, Mysak LA (1997) Modeling sea ice as a granular material, including the dilatancy effect. J Phys Oceanogr 27:2342–2360
84. Trujillo E, Ramirez JA, Elder KJ (2007) Topographic, meteorologic and canopy controls on the scaling characteristics if the spatial distribution of snow depth fields. Water Resour Res 43:W07409
85. van der Veen CJ, Payne AJ (2004) Modelling land-ice dynamics. In: Bamber JL, Payne AJ (eds) Mass balance of the cryosphere: Observations and modelling of contemporary and future change. Cambridge University Press, Cambridge, pp 169–225
86. Washington WM, Meehl GA (1996) High-latitude climate change in a global coupled ocean-atmosphere-sea ice model with increased atmospheric CO_2. J Geophys Res 101(D8):12795–12802
87. Washington WM, Semtner AJ, Parkinson C, Morrison L (1976) On the development of a seasonal change sea-ice model. J Oceanogr 6:679–685
88. Weertman J (1957) On the sliding of glaciers. J Glaciol 5:287–303
89. Williams PJ, Smith MW (1989} The frozen earth. Cambridge University Press, Cambridge, p 306
90. Winstral A, Marks D (2002) Simulating wind fields and snow redistribution using terrain-based parameters to model snow accumulation and melt over a semi-arid mountain catchment. Hydrol Process 16:3585–3603
91. World Meteorological Organization (2007) WMO sea ice nomenclature, no 269. WMO, Geneva
92. Zhang T-J, Armstrong RL, Smith J (2003) Investigation of the near-surface soil freeze-thaw cycle in the contiguous United States: Algorithm development and validation. J Geophys Res 108(D22):8860, GCP 21-1 – 21-14
93. Zhang T-J et al (2005) Spatial and temporal variability in active layer thickness over the Russian Arctic drainage basin. J Geophys Res 110:D16101
94. Joint Commission on Oceanography and Marine Meteorology (2007) http://www.ipy-ice-portal.org/. Accessed 22 Aug 2008

Books and Reviews

Bamber JL, Payne AJ (eds) (2004) Mass balance of the cryosphere: Observations and modelling of contemporary and future change. Cambridge University Press, Cambridge, p 644

Curvelets and Ridgelets

Jalal Fadili[1], Jean-Luc Starck[2]
[1] GREYC CNRS UMR 6072, ENSICAEN, Caen Cedex, France
[2] Laboratoire Astrophysique des Interactions Multi-échelles UMR 7158, CEA/DSM-CNRS-Universite Paris Diderot, SEDI-SAP, Gif-sur-Yvette Cedex, France

Article Outline

Glossary
Definition of the Subject
Introduction
Ridgelets
Curvelets
Stylized Applications
Future Directions
Bibliography

Glossary

WT1D The one-dimensional Wavelet Transform as defined in [53]. See also ► Numerical Issues When Using Wavelets.

WT2D The two-dimensional Wavelet Transform.

Discrete ridgelet trasnform (DRT) The discrete implementation of the continuous Ridgelet transform.

Fast slant stack (FSS) An algebraically exact Radon transform of data on a Cartesian grid.

First generation discrete curvelet transform (DCTG1) The discrete curvelet transform constructed based on the discrete ridgelet transform.

Second generationdiscrete curvelet transformx (DCTG2) The discrete curvelet transform constructed based on appropriate bandpass filtering in the Fourier domain.

Anisotropic elements By anistropic, we mean basis elements with elongated effective support; i. e. length > width.

Parabolic scaling law A basis element obeys the parabolic scaling law if its effective support is such that width $\approx$ length^2.

Definition of the Subject

Despite the fact that wavelets have had a wide impact in image processing, they fail to efficiently represent objects with highly anisotropic elements such as lines or curvilinear structures (e. g. edges). The reason is that wavelets are non-geometrical and do not exploit the regularity of the edge curve.

The Ridgelet and the Curvelet [16,17] transforms were developed as an answer to the weakness of the separable wavelet transform in sparsely representing what appears to be simple building atoms in an image, that is lines, curves and edges. Curvelets and ridgelets take the form of basis elements which exhibit high directional sensitivity and are highly anisotropic [9,18,32,68]. These very recent geometric image representations are built upon ideas of multiscale analysis and geometry. They have had an important success in a wide range of image processing applications including denoising [42,64,68], deconvolution [38,74], contrast enhancement [73], texture analysis [2], detection [44], watermarking [78], component separation [70,71], inpainting [37,39] or blind source separation [6,7]. Curvelets have also proven useful in diverse fields beyond the traditional image processing application. Let's cite for example seismic imaging [34,42,43], astronomical imaging [48,66,69], scientific computing and analysis of partial differential equations [13,14]. Another reason for the success of ridgelets and curvelets is the availability of fast transform algorithms which are available in non-commercial software packages following the philosophy of reproducible research, see [4,75].

Introduction

Sparse Geometrical Image Representation

Multiscale methods have become very popular, especially with the development of the wavelets in the last decade. Background texts on the wavelet transform include [23,53,72]. An overview of implementation and practical issues of the wavelet transform can also be found in ► Numerical Issues When Using Wavelets.

Despite the success of the classical wavelet viewpoint, it was argued that the traditional wavelets present some strong limitations that question their effectiveness in higher-dimension than 1 [16,17]. Wavelets rely on a dictionary of roughly isotropic elements occurring at all scales and locations, do not describe well highly anisotropic elements, and contain only a fixed number of directional elements, independent of scale. Following this reasoning, new constructions have been proposed such as the ridgelets [9,16] and the curvelets [17,18,32,68]. Ridgelets and curvelets are special members of the family of multiscale orientation-selective transforms, which has recently led to a flurry of research activity in the field of computational and applied harmonic analysis. Many other constructions belonging to this family have been investigated in the literature, and go by the name contourlets [27], directionlets [76], bandlets [49,62], grouplets [54], shearlets [47], dual-tree wavelets and wavelet packets [40,46], etc.

Throughout this paper, the term 'sparsity' is used and intended in a weak sense. We are aware that practical images and signals may not be supported in a transform domain on a set of relatively small size (sparse set). Instead, they may only be compressible (nearly sparse) in some transform domain. Hence, with a slight abuse of terminology, we will say that a representation is sparse for an image within a certain class, if it provides a compact description of such an image.

Notations

We work throughout in two dimensions with spatial variable $\mathbf{x} \in \mathbb{R}^2$ and $\boldsymbol{\nu}$ a continuous frequency-domain variable. Parentheses $(.,.)$ are used for continuous-domain function evaluations, and brackets $[.,.]$ for discrete-domain array indices. The hat $\hat{}$ notation will be used for the Fourier transform.

Ridgelets

The Continuous Ridgelet Transform

The two-dimensional continuous ridgelet transform in $\mathbb{R}^2$ can be defined as follows [10]. We pick a smooth univariate function $\psi : \mathbb{R} \to \mathbb{R}$ with sufficient decay and satisfying the admissibility condition

$$\int |\hat{\psi}(\boldsymbol{\nu})|^2/|\boldsymbol{\nu}|^2 \, \mathrm{d}\boldsymbol{\nu} < \infty \,, \tag{1}$$

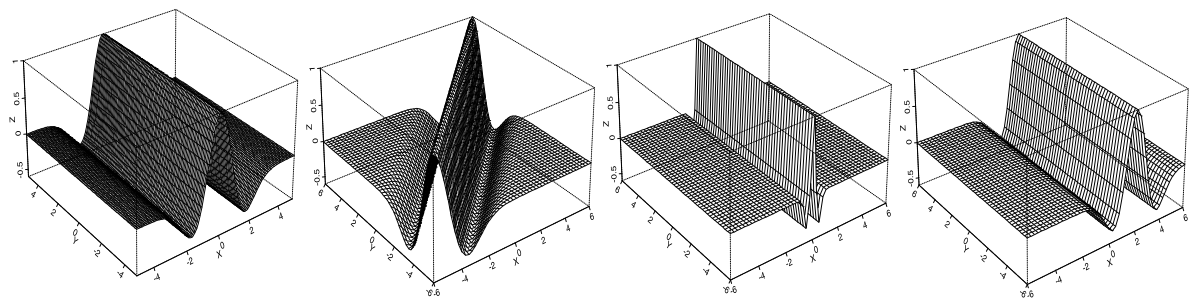

Curvelets and Ridgelets, Figure 1
Few Ridgelets examples – The second to fourth graphs are obtained after simple geometric manipulations of the first ridgelet, namely rotation, rescaling, and shifting

which holds if, say, ψ has a vanishing mean $\int \psi(t)\mathrm{d}t = 0$. We will suppose a special normalization about ψ so that $\int_0^\infty |\hat{\psi}(\boldsymbol{\nu})|^2 \boldsymbol{\nu}^{-2}\mathrm{d}\boldsymbol{\nu} = 1$.

For each scale $a > 0$, each position $b \in \mathbb{R}$ and each orientation $\theta \in [0, 2\pi)$, we define the bivariate *ridgelet* $\psi_{a,b,\theta} : \mathbb{R}^2 \to \mathbb{R}$ by

$$\begin{aligned}\psi_{a,b,\theta}(\mathbf{x}) &= \psi_{a,b,\theta}(x_1, x_2)\\ &= a^{-1/2} \cdot \psi((x_1 \cos\theta + x_2 \sin\theta - b)/a)\,;\end{aligned} \tag{2}$$

A ridgelet is constant along lines $x_1 \cos\theta + x_2 \sin\theta = \text{const}$. Transverse to these ridges it is a wavelet. Figure 1 depicts few examples of ridgelets. The second to fourth panels are obtained after simple geometric manipulations of the ridgelet (left panel), namely rotation, rescaling, and shifting.

Given an integrable bivariate function $f(\mathbf{x})$, we define its ridgelet coefficients by

$$\mathcal{R}_f(a, b, \theta) := \langle f, \psi_{a,b,\theta}\rangle = \int_{\mathbb{R}^2} f(\mathbf{x})\overline{\psi}_{a,b,\theta}(\mathbf{x})\,\mathrm{d}\mathbf{x}\,.$$

We have the exact reconstruction formula

$$f(\mathbf{x}) = \int_0^{2\pi}\int_{-\infty}^{\infty}\int_0^{\infty} \mathcal{R}_f(a, b, \theta)\psi_{a,b,\theta}(\mathbf{x})\frac{\mathrm{d}a}{a^3}\mathrm{d}b\frac{\mathrm{d}\theta}{4\pi} \tag{3}$$

valid almost everywhere for functions which are both integrable and square integrable. This formula is stable and one can prove a Parseval relation [16].

Ridgelet analysis may be constructed as wavelet analysis in the Radon domain. The rationale behind this is that the Radon transform translates singularities along lines into point singularities, for which the wavelet transform is known to provide a sparse representation. Recall that the Radon transform of an object f is the collection of line integrals indexed by $(\theta, t) \in [0, 2\pi) \times \mathbb{R}$ given by

$$Rf(\theta, t) = \int_{\mathbb{R}^2} f(x_1, x_2)\delta(x_1 \cos\theta + x_2 \sin\theta - t)\,\mathrm{d}x_1\mathrm{d}x_2\,, \tag{4}$$

where δ is the Dirac distribution. Then the ridgelet transform is precisely the application of a 1-dimensional wavelet transform to the slices of the Radon transform where the angular variable θ is constant and t is varying. Thus, the basic strategy for calculating the continuous ridgelet transform is first to compute the Radon transform $Rf(t, \theta)$ and second, to apply a one-dimensional wavelet transform to the slices $Rf(\cdot, \theta)$. Several digital ridgelet transforms (DRTs) have been proposed, and we will describe three of them in this section, based on different implementations of the Radon transform.

The RectoPolar Ridgelet Transform A fast implementation of the Radon transform can be proposed in the Fourier domain, based on the projection-slice-theorem. First the 2D-FFT of the given image is computed. Then the resulting function in the frequency domain is to be used to evaluate the frequency values in a polar grid of rays passing through the origin and spread uniformly in angle. This conversion from Cartesian to Polar grid could be obtained by interpolation, and this process is well known by the name gridding in tomography. Given the polar grid samples, the number of rays corresponds to the number of projections, and the number of samples on each ray corresponds to the number of shifts per such angle. Applying one dimensional inverse Fourier transform for each ray, the Radon projections are obtained.

The above described process is known to be inaccurate due to the sensitivity to the interpolation involved. This implies that for a better accuracy, the first 2D-FFT employed should be done with high-redundancy.

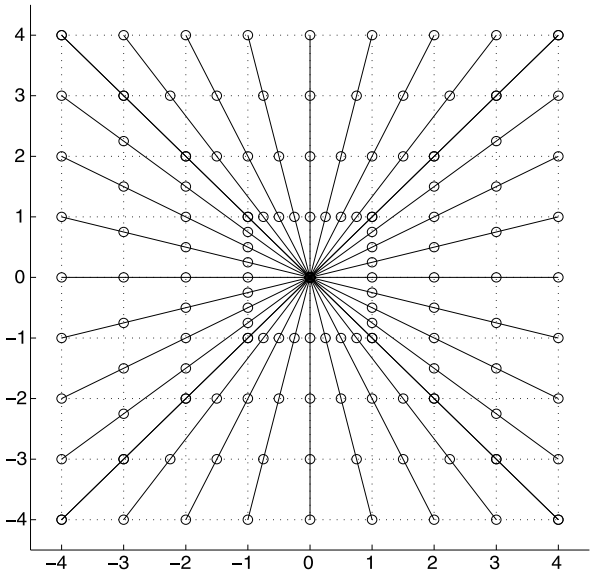

Curvelets and Ridgelets, Figure 2
Illustration of the pseudo-polar grid in the frequency domain for an n by n image ($n = 8$)

An alternative solution for the Fourier-based Radon transform exists, where the polar grid is replaced with a pseudo-polar one. The geometry of this new grid is illustrated in Fig. 2. Concentric circles of linearly growing radius in the polar grid are replaced by concentric squares of linearly growing sides. The rays are spread uniformly not in angle but in slope. These two changes give a grid vaguely resembling the polar one, but for this grid a direct FFT can be implemented with no interpolation. When applying now 1D-FFT for the rays, we get a variant of the Radon transform, where the projection angles are not spaced uniformly.

For the pseudo-polar FFT to be stable, it was shown that it should contain at least twice as many samples, compared to the original image we started with. A by-product of this construction is the fact that the transform is organized as a 2D array with rows containing the projections as a function of the angle. Thus, processing the Radon transform in one axis is easily implemented. More details can be found in [68].

One-Dimensional Wavelet Transform To complete the ridgelet transform, we must take a one-dimensional wavelet transform (WT1D) along the radial variable in Radon space. We now discuss the choice of the digital WT1D.

Experience has shown that compactly-supported wavelets can lead to many visual artifacts when used in conjunction with nonlinear processing, such as hard-thresholding of individual wavelet coefficients, particularly for decimated wavelet schemes used at critical sampling. Also, because of the lack of localization of such compactly-supported wavelets in the frequency domain, fluctuations in coarse-scale wavelet coefficients can introduce fine-scale fluctuations. A frequency-domain approach must be taken, where the discrete Fourier transform is reconstructed from the inverse Radon transform. These considerations lead to use band-limited wavelet, whose support is compact in the Fourier domain rather than the time-domain [28,29,68]. In [68], a specific overcomplete wavelet transform [67,72] has been used. The wavelet transform algorithm is based on a scaling function ϕ such that $\hat{\phi}$ vanishes outside of the interval $[-\nu_c, \nu_c]$. We define the Fourier transform of the scaling function as a re-normalized B_3-spline

$$\hat{\phi}(\nu) = \frac{3}{2} B_3(4\nu),$$

and $\hat{\psi}$ as the difference between two consecutive resolutions

$$\hat{\psi}(2\nu) = \hat{\phi}(\nu) - \hat{\phi}(2\nu).$$

Because $\hat{\psi}$ is compactly supported, the sampling theorem shows than one can easily build a pyramid of $n + n/2 + \cdots + 1 = 2n$ elements, see [72] for details.

This WT1D transform enjoys the following useful properties:

- The wavelet coefficients are directly calculated in the Fourier space. In the context of the ridgelet transform, this allows avoiding the computation of the one-dimensional inverse Fourier transform along each radial line.
- Each sub-band is sampled above the Nyquist rate, hence, avoiding aliasing –a phenomenon typically encountered by critically sampled orthogonal wavelet transforms [65].
- The reconstruction is trivial. The wavelet coefficients simply need to be co-added to reconstruct the input signal at any given point. In our application, this implies that the ridgelet coefficients simply need to be co-added to reconstruct Fourier coefficients.

This wavelet transform introduces an extra redundancy factor. However, we note that the goal in this implementation is not data compression or efficient coding. Rather, this implementation would be useful to the practitioner whose focuses on data analysis, for which it

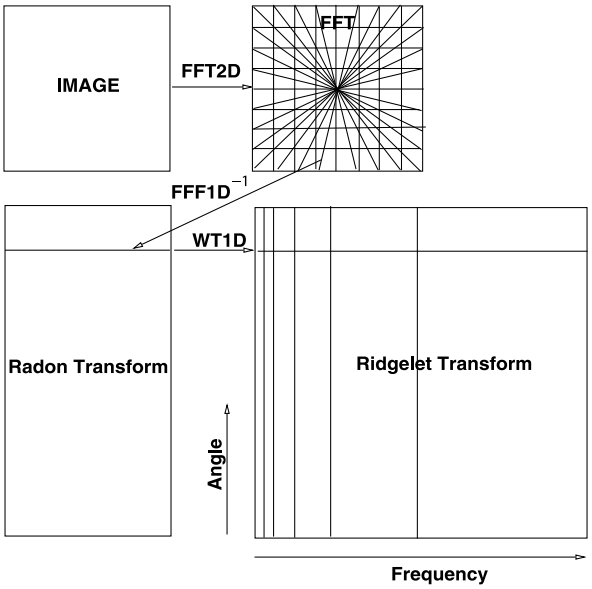

Curvelets and Ridgelets, Figure 3
Discrete ridgelet transform flowchart. Each of the $2n$ radial lines in the Fourier domain is processed separately. The 1-D inverse FFT is calculated along each radial line followed by a 1-D nonorthogonal wavelet transform. In practice, the one-dimensional wavelet coefficients are directly calculated in the Fourier space

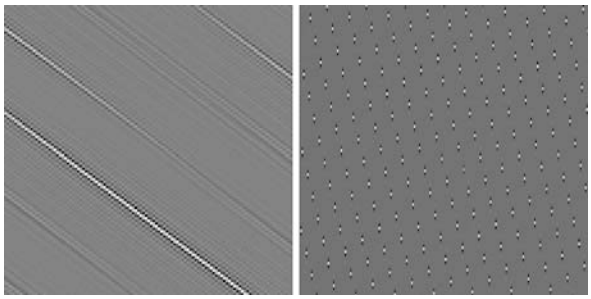

Curvelets and Ridgelets, Figure 4
The backprojection of a ridgelet coefficient by the FFT-based ridgelet transform (*left*), and by the OFRT (*right*)

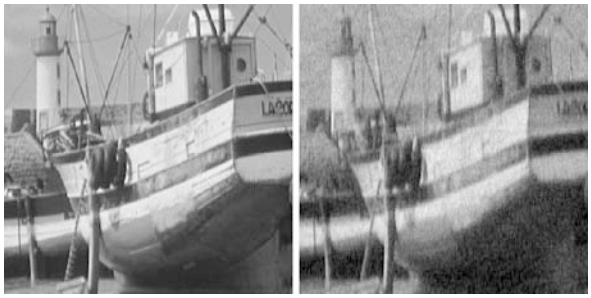

Curvelets and Ridgelets, Figure 5
Part of original noise-free `Boat` image (*left*), and reconstruction after hard thresholding its OFRT coefficients (*right*)

is well-known that over-completeness through (almost) translation-invariance can provide substantial advantages.

Assembling all above ingredients together gives the flowchart of the discrete ridgelet transform (DRT) depicted in Fig. 3. The DRT of an image of size $n \times n$ is an image of size $2n \times 2n$, introducing a redundancy factor equal to 4.

We note that, because this transform is made of a chain of steps, each one of which is invertible, the whole transform is invertible, and so has the exact reconstruction property. For the same reason, the reconstruction is stable under perturbations of the coefficients.

Last but not least, this discrete transform is computationally attractive. Indeed, the algorithm we presented here has low complexity since it runs in $O(n^2 \log n)$ flops for an $n \times n$ image.

The Orthonormal Finite Ridgelet Transform

The orthonormal finite ridgelet transform (OFRT) has been proposed [26] for image compression and filtering. This transform is based on the finite Radon transform [55] and a 1D orthogonal wavelet transform. It is not redundant and reversible. It would have been a great alternative to the previously described ridgelet transform if the OFRT were not based on an awkward definition of a line. In fact, a line in the OFRT is defined algebraically rather that geometrically, and so the points on a 'line' can be arbitrarily and randomly spread out in the spatial domain. Figure 4 shows the back-projection of a ridgelet coefficient by the FFT-based ridgelet transform (left) and by the OFRT (right). It is clear that the backprojection of the OFRT is nothing like a ridge function.

Because of this specific definition of a line, the thresholding of the OFRT coefficients produces strong artifacts. Figure 5 shows a part of the original image `Boat`, and its reconstruction after hard thresholding the OFRT of the noise-free `Boat`. The resulting image is not smoothed as one would expect, but rather a noise has been added to the noise-free image as part of the filtering!

Finally, the OFRT presents another limitation: the image size must be a prime number. This last point is however not too restrictive, because we generally use a spatial partitioning when denoising the data, and a prime number block size can be used. The OFRT is interesting from the conceptual point of view, but still requires work before it can be used for real applications such as denoising.

The Fast Slant Stack Ridgelet Transform

The Fast Slant Stack (FSS) [3] is a Radon transform of data on a Cartesian grid, which is algebraically exact and geometrically more accurate and faithful than the previously described methods. The back-projection of a point in Radon space is exactly a ridge function in the spatial domain (see Fig. 6). The transformation of an $n \times n$ image is a $2n \times 2n$ image. n line integrals with angle between $\left[-\frac{\pi}{4}, \frac{\pi}{4}\right]$ are calculated from the zero padded image on the y-axis, and n line integrals with angle between $\left[\frac{\pi}{4}, \frac{3\pi}{4}\right]$ are computed by zero padding the image on the x-axis. For a given angle inside $\left[-\frac{\pi}{4}, \frac{\pi}{4}\right]$, $2n$ line integrals are calculated by first shearing the zero-padded image, and then integrating the pixel values along all horizontal lines (resp. vertical lines for angles in $\left[\frac{\pi}{4}, \frac{3\pi}{4}\right]$). The shearing is performed one column at a time (resp. one line at a time) by using the 1D FFT. Figure 7 shows an example of the image shearing step with two different angles ($5\frac{\pi}{4}$ and $-\frac{\pi}{4}$). A DRT based on the FSS transform has been proposed in [33]. The connection between the FSS and the Linogram has been investigated in [3]. A FSS algorithm is also proposed in [3], based on the 2D Fast Pseudo-polar Fourier transform which evaluates the 2-D Fourier transform on a non-Cartesian (pseudo-polar) grid, operating in $O(n^2 \log n)$ flops.

Figure 8 left exemplifies a ridgelet in the spatial domain obtained from the DRT based on FSS implementation. Its

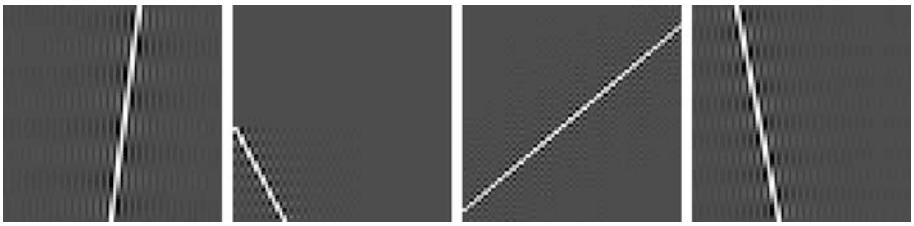

Curvelets and Ridgelets, Figure 6
Backprojection of a point at four different locations in the Radon space using the FSS algorithm

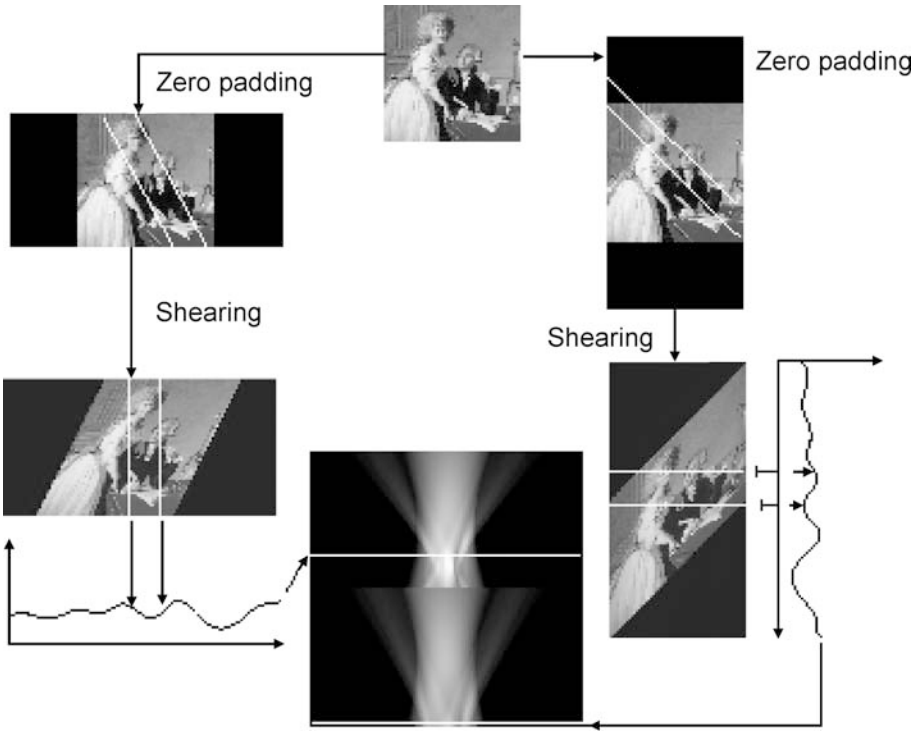

Curvelets and Ridgelets, Figure 7
Slant Stack Transform of an image

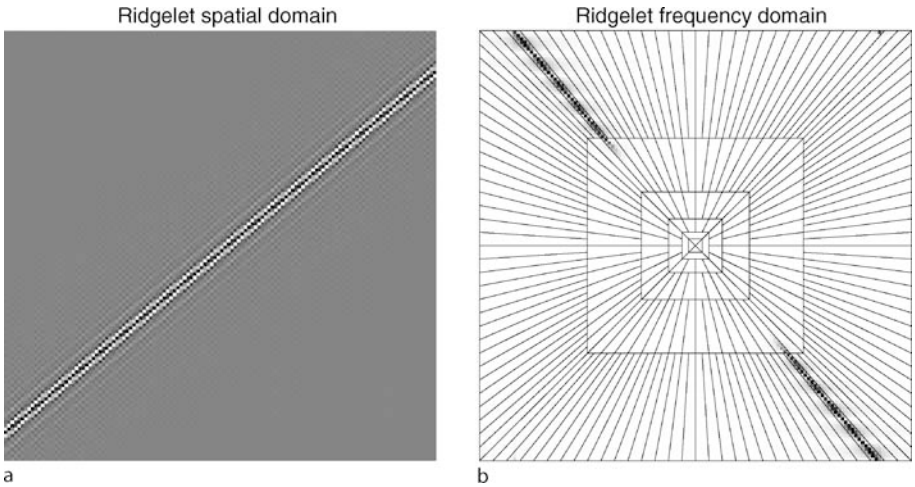

Curvelets and Ridgelets, Figure 8
a Example of a ridgelet obtained by the Fast Slant Stack implementation. b Its FFT superimposed on the DRT frequency tiling

Fourier transform is shown on Fig. 8 right superimposed on the DRT frequency tiling [33]. The Fourier transform of the discrete ridgelet lives in an angular wedge. More precisely, the Fourier transform of a discrete ridgelet at scale j lives within a dyadic square of size $\sim 2^j$.

Local Ridgelet Transforms

The ridgelet transform is optimal for finding global lines of the size of the image. To detect line segments, a partitioning must be introduced [9]. The image can be decomposed into overlapping blocks of side-length b pixels in such a way that the overlap between two vertically adjacent blocks is a rectangular array of size b by $b/2$; we use overlap to avoid blocking artifacts. For an n by n image, we count $2n/b$ such blocks in each direction, and thus the redundancy factor grows by a factor of 4.

The partitioning introduces redundancy, as a pixel belongs to 4 neighboring blocks. We present two competing strategies to perform the analysis and synthesis:

1. The block values are weighted by a spatial window w (analysis) in such a way that the co-addition of all blocks reproduce exactly the original pixel value (synthesis).
2. The block values are those of the image pixel values (analysis) but are weighted when the image is reconstructed (synthesis).

Experiments have shown that the second approach leads to better results especially for restoration problems, see [68] for details. We calculate a pixel value, $f[i_1, i_2]$ from its four corresponding block values of half-size $m = b/2$, namely, $B_1[k_1, l_1]$, $B_2[k_2, l_1]$, $B_3[k_1, l_2]$ and $B_4[k_2, l_2]$ with $k_1, l_1 > b/2$ and $k_2 = k_1 - m$, $l_2 = l_1 - m$, in the following way:

$$\begin{aligned} f_1 &= w(k_2/m)B_1[k_1, l_1] + w(1 - k_2/m)B_2[k_2, l_1] \\ f_2 &= w(k_2/m]B_3[k_1, l_2] + w(1 - k_2/m)B_4[k_2, l_2] \\ f[i_1, i_2] &= w(l_2/m)f_1 + w(1 - l_2/m)f_2. \end{aligned} \tag{5}$$

where $w(x) = \cos^2(\pi x/2)$ is the window. Of course, one might select any other smooth, non-increasing function satisfying, $w(0) = 1$, $w(1) = 0$, $w'(0) = 0$ and obeying the symmetry property $w(x) + w(1 - x) = 1$.

Sparse Representation by Ridgelets

The continuous ridgelet transform provides sparse representation of both smooth functions (in the Sobolev space W_2^2) and of perfectly straight lines [11,31]. We have just seen that there are also various DRTs, i. e. expansions with countable discrete collection of generating elements, which correspond either to frames or orthobases. It has been shown for these schemes that the DRT achieves near-optimal M-term approximation – that is the non-linear approximation of f using the M highest ridgelet coefficients in magnitude - to smooth images with discontinuities along straight lines [16,31]. In summary, ridgelets provide sparse presentation for piecewise smooth images away from global straight edges.

Curvelets

The First Generation Curvelet Transform

In image processing, edges are curved rather than straight lines and ridgelets are not able to efficiently represent such

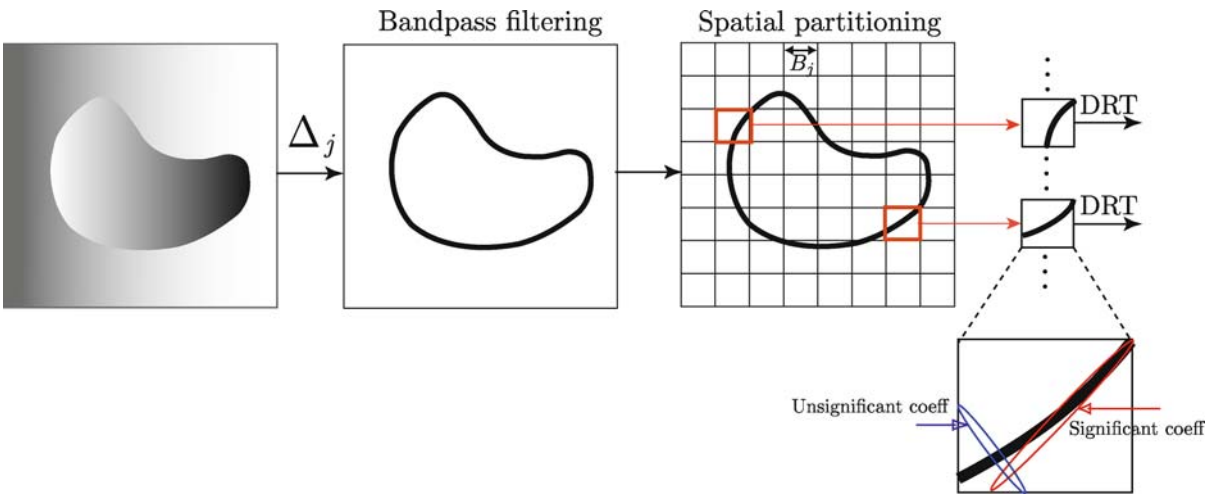

Curvelets and Ridgelets, Figure 9
Local ridgelet transform on bandpass filtered image. At fine scales, curved edges are almost straight lines

images. However, one can still deploy the ridgelet machinery in a localized way, at fine scales, where curved edges are almost straight lines (see Fig. 9). This is the idea underlying the first generation curvelets (termed here CurveletG1) [18].

First Generation Curvelets Construction The CurveletG1 transform [18,32,68] opens the possibility to analyze an image with different block sizes, but with a single transform. The idea is to first decompose the image into a set of wavelet bands, and to analyze each band by a local ridgelet transform as illustrated on Fig. 9. The block size can be changed at each scale level. Roughly speaking, different levels of the multiscale ridgelet pyramid are used to represent different sub-bands of a filter bank output. At the same time, this sub-band decomposition imposes a relationship between the width and length of the *important* frame elements so that they are anisotropic and obey approximately the parabolic scaling law width $\approx$ length^2.

The First Generation Discrete Curvelet Transform (DCTG1) of a continuum function $f(\mathbf{x})$ makes use of a dyadic sequence of scales, and a bank of filters with the property that the bandpass filter Δ_j is concentrated near the frequencies $[2^{2j}, 2^{2j+2}]$, i. e.

$$\Delta_j(f) = \Psi_{2j} * f, \quad \widehat{\Psi}_{2j}(\boldsymbol{\nu}) = \widehat{\Psi}(2^{-2j}\boldsymbol{\nu}) .$$

In wavelet theory, one uses a decomposition into dyadic sub-bands $[2^j, 2^{j+1}]$. In contrast, the sub-bands used in the discrete curvelet transform of continuum functions has the nonstandard form $[2^{2j}, 2^{2j+2}]$. This is nonstandard feature of the DCTG1 well worth remembering (this is where the approximate parabolic scaling law comes into play).

The DCTG1 decomposition is the sequence of the following steps:

- *Sub-band Decomposition.* The object f is decomposed into sub-bands.
- *Smooth Partitioning.* Each sub-band is smoothly windowed into "squares" of an appropriate scale (of side-length $\sim 2^{-j}$).
- *Ridgelet Analysis.* Each square is analyzed via the DRT.

In this definition, the two dyadic sub-bands $[2^{2j}, 2^{2j+1}]$ and $[2^{2j+1}, 2^{2j+2}]$ are merged before applying the ridgelet transform.

Digital Implementation It seems that the isotropic "à trous" wavelet transform (► Numerical Issues When Using Wavelets), [72] is especially well-adapted to the needs of the digital curvelet transform. The algorithm decomposes an n by n image $f[i_1, i_2]$ as a superposition of the form

$$f[i_1, i_2] = c_J[i_1, i_2] + \sum_{j=1}^{J} w_j[i_1, i_2],$$

where c_J is a coarse or smooth version of the original image f and w_j represents 'the details of f' at scale 2^{-j}. Thus, the algorithm outputs $J + 1$ sub-band arrays of size $n \times n$.
A sketch of the DCTG1 implementation is given in Algorithm 1.
The side-length of the localizing windows is doubled *at every other* dyadic sub-band, hence maintaining the fundamental property of the curvelet transform which says

Require: Input $n \times n$ image $f[i_1, i_2]$, type of DRT (see above).
Apply the à trous isotropic WT2D with J scales,
Set $B_1 = B_{\min}$,
for $j = 1, \ldots, J$ **do**
  Partition the sub-band w_j with a block size B_j and apply the DRT to each block,
  if j modulo $2 = 1$ **then**
    $B_{j+1} = 2B_j$,
  else
    $B_{j+1} = B_j$.
  end if
end for

Curvelets and Ridgelets, Algorithm 1
DCTG1

that elements of length about $2^{-j/2}$ serve for the analysis and synthesis of the jth sub-band $[2^j, 2^{j+1}]$. Note also that the coarse description of the image c_J is left intact. In the results shown in this paper, we used the default value $B_{\min} = 16$ pixels in our implementation. Figure 10 gives an overview of the organization of the DCTG1 algorithm.

This implementation of the DCTG1 is also redundant. The redundancy factor is equal to $16J + 1$ whenever J scales are employed. The DCTG1 algorithm enjoys exact reconstruction and stability, as each step of the analysis (decomposition) algorithm is itself invertible. One can show that the computational complexity of the DCTG1 algorithm we described here based on the DRT of Fig. 3 is $O(n^2(\log n)^2)$ for an $n \times n$ image.
Figure 11 shows a few curvelets at different scales, orientations and locations.

Sparse Representation by First Generation Curvelets
The CurveletG1 elements can form either a frame or a tight frame for $L_2(\mathbb{R}^2)$ [17], depending on the WT2D used and the DRT implementation (rectopolar or FSS Radon transform). The frame elements are anisotropic by construction and become successively more anisotropic at progressively higher scales. These curvelets also exhibit directional sensitivity and display oscillatory components across the 'ridge'. A central motivation leading to the curvelet construction was the problem of non-adaptively representing piecewise smooth (e. g. C^2) images f which have discontinuity along a C^2 curve. Such a model is the so-called cartoon model of (non-textured) images. With the CurveletG1 tight frame construction, it was shown in [17] that for such f, the M-term non-linear approximations f_M of f obey,

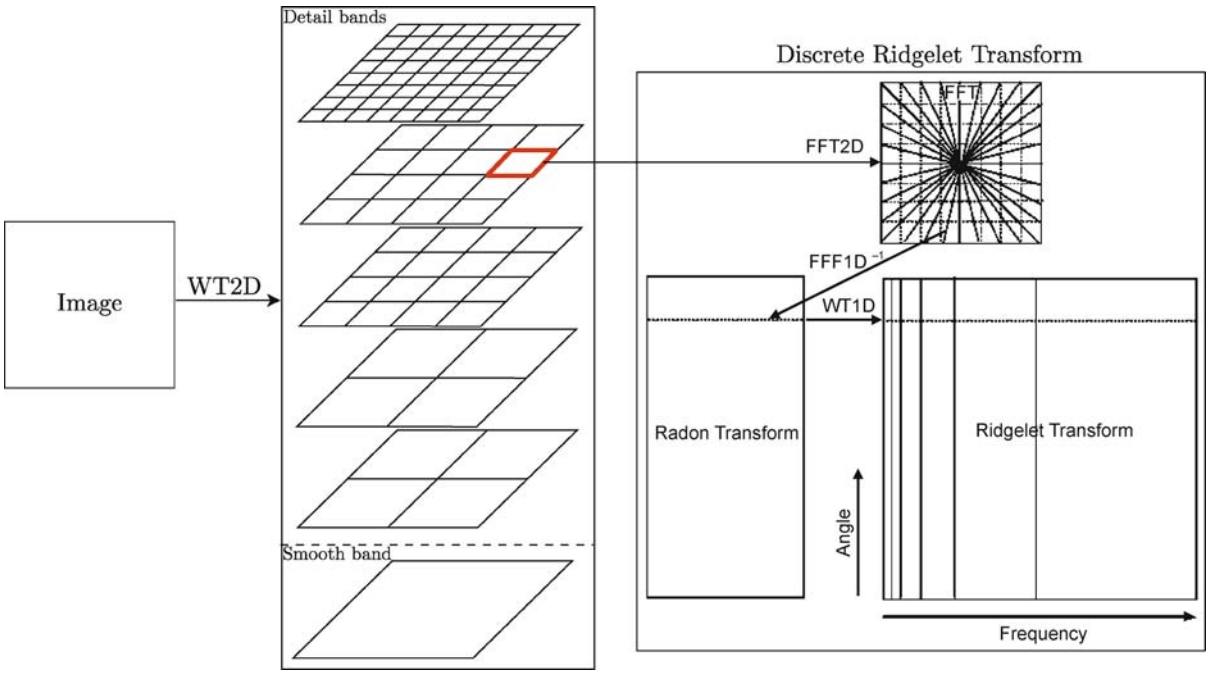

Curvelets and Ridgelets, Figure 10
First Generation Discrete Curvelet Transform (DCTG1) flowchart. The figure illustrates the decomposition of the original image into sub-bands followed by the spatial partitioning of each sub-band. The ridgelet transform is then applied to each block

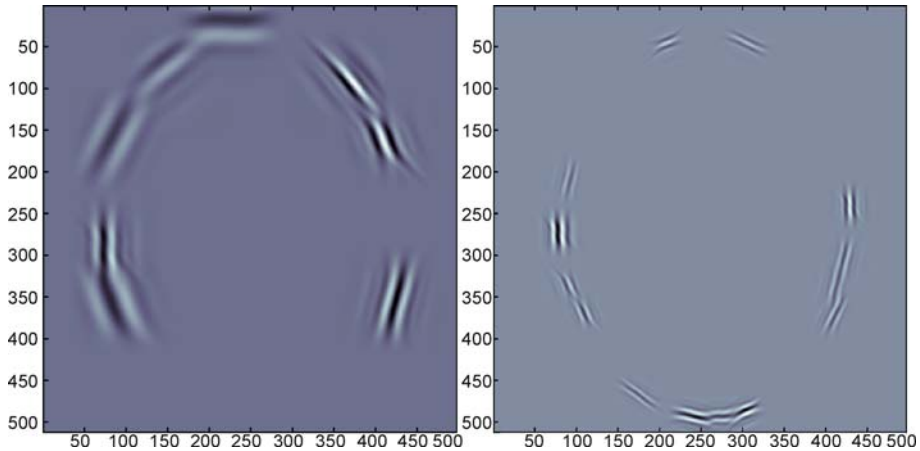

Curvelets and Ridgelets, Figure 11
A few first generation curvelets

for each $\kappa > 0$,

$$\|f - f_M\|_2 \le C_\kappa M^{-2+\kappa}\,, \quad M \to +\infty\,.$$

The M-term approximations in the CurveletG1 are almost rate optimal, much better than M-term Fourier or wavelet approximations for such images, see [53].

The Second Generation Curvelet Transform

Despite these interesting properties, the CurveletG1 construction presents some drawbacks. First, the construction involves a complicated seven-index structure among which we have parameters for scale, location and orientation. In addition, the parabolic scaling ratio width $\approx$ length^2 is not completely true (see Subsect. "First Generation Curvelets Construction"). In fact, CurveletG1 assumes a wide range of aspect ratios. These facts make mathematical and quantitative analysis especially delicate. Second, the spatial partitioning of the CurveletG1 transform uses overlapping windows to avoid blocking effects. This leads to an increase of the redundancy of the DCTG1. The computational cost of the DCTG1 algorithm may also be a limitation for large-scale data, especially if the FSS-based DRT implementation is used.

In contrast, the second generation curvelets (CurveletG2) [15,20] exhibit a much simpler and natural indexing structure with three parameters: scale, orientation (angle) and location, hence simplifying mathematical analysis. The CurveletG2 transform also implements a tight frame expansion [20] and has a much lower redundancy. Unlike the DCTG1, the discrete CurveletG2 implementation will not use ridgelets yielding a faster algorithm [15,20].

Second Generation Curvelets Construction

Continuous Coronization The second generation curvelets are defined at scale 2^{-j}, orientation θ_l and position $\mathbf{x}_{\mathbf{k}}^{j,l} = R_{\theta_l}^{-1}(2^{-j}k_1, 2^{-j/2}k_2)$ by translation and rotation of a mother curvelet φ_j as

$$\varphi_{j,l,\mathbf{k}}(\mathbf{x}) = \varphi_j(R_{\theta_l}(\mathbf{x} - \mathbf{x}_{\mathbf{k}}^{j,l}))\,, \tag{6}$$

where R_{θ_l} is the rotation by θ_l radians. θ_l is the equispaced sequence of rotation angles $\theta_l = 2\pi 2^{-\lfloor j/2 \rfloor} l$, with integer l such that $0 \le \theta_l \le 2\pi$ (note that the number of orientations varies as $1/\sqrt{\text{scale}}$). $\mathbf{k} = (k_1, k_2) \in \mathbb{Z}^2$ is the sequence of translation parameters. The waveform φ_j is defined by means of its Fourier transform $\hat{\varphi}_j(\boldsymbol{\nu})$, written in polar coordinates in the Fourier domain

$$\hat{\varphi}_j(r,\theta) = 2^{-3j/4}\hat{w}(2^{-j}r)\hat{v}\left(\frac{2^{\lfloor j/2 \rfloor}\theta}{2\pi}\right). \tag{7}$$

The support of $\hat{\varphi}_j$ is a polar *parabolic wedge* defined by the support of $\hat{w}$ and $\hat{v}$, the radial and angular windows (both smooth, nonnegative and real-valued), applied with scale-dependent window widths in each direction. $\hat{w}$ and $\hat{v}$ must also satisfy the partition of unity property [15]. See the frequency tiling in Fig. 12a.

In continuous frequency $\boldsymbol{\nu}$, the CurveletG2 coefficients of data $f(\mathbf{x})$ are defined as the inner product

$$c_{j,l,\mathbf{k}} := \langle f\varphi_{j,l,\mathbf{k}}\rangle = \int_{\mathbb{R}^2} \hat{f}(\boldsymbol{\nu})\hat{\varphi}_j(R_{\theta_l}\boldsymbol{\nu})e^{i\mathbf{x}_{\mathbf{k}}^{j,l}\cdot\boldsymbol{\nu}}\,\mathrm{d}\boldsymbol{\nu}\,. \tag{8}$$

This construction implies a few properties: (i) the CurveletG2 defines a tight frame of $L_2(\mathbb{R}^2)$, (ii) the effective length and width of these curvelets obey the parabolic

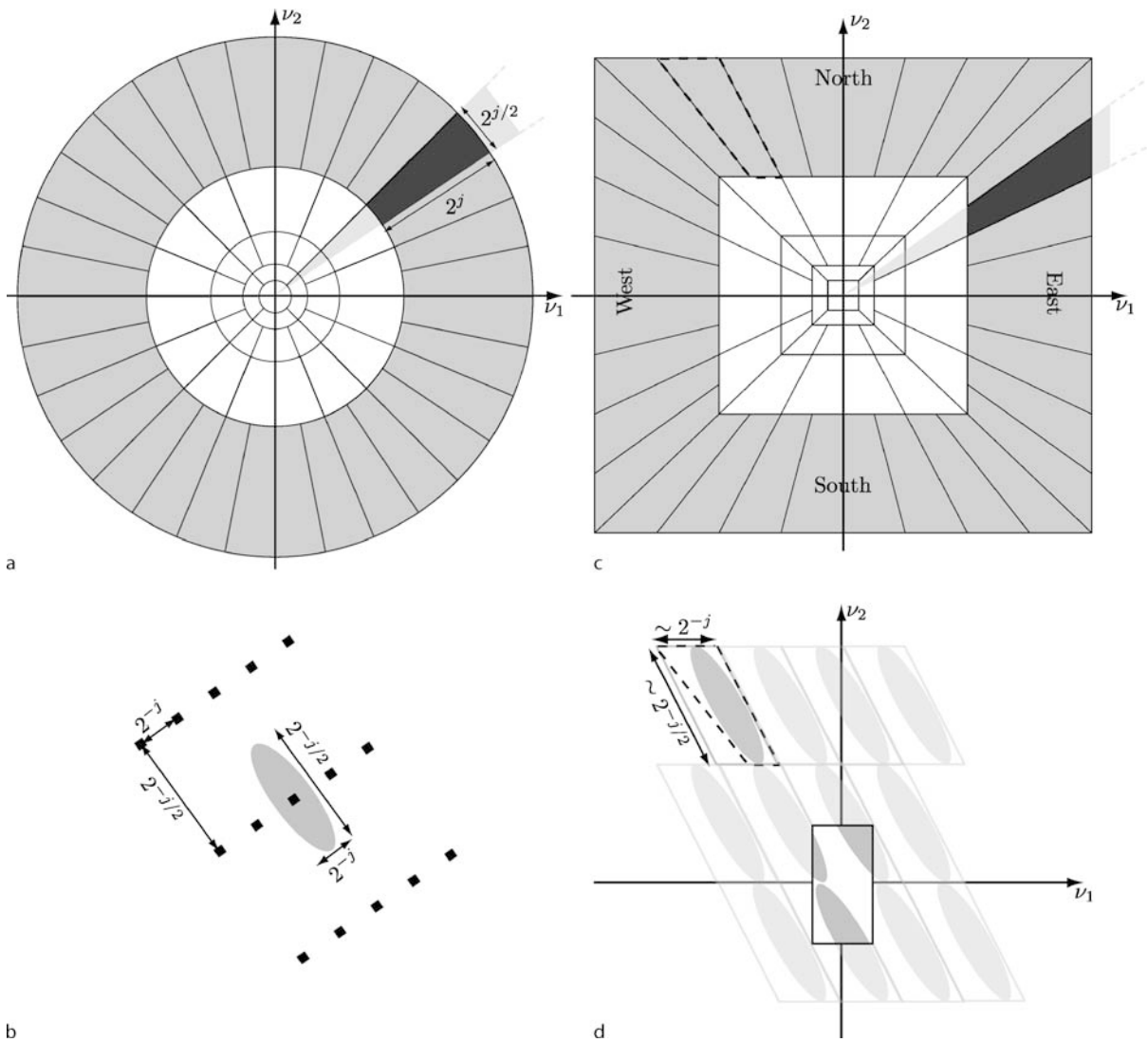

Curvelets and Ridgelets, Figure 12

a Continuous curvelet frequency tiling. The *dark gray* area represents a wedge obtained as the product of the radial window (annulus shown in *gray*) and the angular window (*light gray*). b The Cartesian grid in space associated to the construction in a whose spacing also obeys the parabolic scaling by duality. c Discrete curvelet frequency tiling. The window $\hat{u}_{j,l}$ isolates the frequency near trapezoidal wedge such as the one shown in dark gray. d The wrapping transformation. The *dashed line* shows the same trapezoidal wedge as in c. The parallelogram contains this wedge and hence the support of the curvelet. After periodization, the wrapped Fourier samples can be collected in the rectangle centered at the origin

scaling relation $(2^{-j} = \text{width}) = (\text{length} = 2^{-j/2})^2$, (iii) the curvelets exhibit an oscillating behavior in the direction perpendicular to their orientation. Curvelets as just constructed are complex-valued. It is easy to obtain real-valued curvelets by working on the symmetrized version $\hat{\varphi}_j(r,\theta) + \hat{\varphi}_j(r,\theta+\pi)$.

Discrete Coronization The discrete transform takes as input data defined on a Cartesian grid and outputs a collection of coefficients. The continuous-space definition of the CurveletG2 uses coronae and rotations that are not especially adapted to Cartesian arrays. It is then convenient to replace these concepts by their Cartesian counterparts. That is concentric squares (instead of concentric circles) and shears (instead of rotations), see Fig. 12c.

The Cartesian equivalent to the radial window $\hat{w}_j(\boldsymbol{\nu}) = \hat{w}(2^{-j}\boldsymbol{\nu})$ would be a bandpass frequency-localized window which can be derived from the differ-

ence of separable low-pass windows $H_j(\boldsymbol{\nu}) = \hat{h}(2^{-j}\nu_1)\hat{h}(2^{-j}\nu_2)$ (h is a 1D low-pass filter):

$$\hat{w}_j(\boldsymbol{\nu}) = \sqrt{H_{j+1}^2(\boldsymbol{\nu}) - H_j^2(\boldsymbol{\nu})}, \forall j \geq 0\,, \quad \text{and} \quad \hat{w}_0(\boldsymbol{\nu}) = \hat{h}(\nu_1)\hat{h}(\nu_2)\,.$$

Another possible choice is to select these windows inspired by the construction of Meyer wavelets [20,57]. See [15] for more details about the construction of the Cartesian $\hat{w}_j$'s.

Let's now examine the angular localization. Each Cartesian coronae has four quadrants: East, North, West and South. Each quadrant is separated into $2^{\lfloor j/2 \rfloor}$ orientations (wedges) with the same areas. Take for example the East quadrant ($-\pi/4 \leq \theta_l < \pi/4$). For the West quadrant, we would proceed by symmetry around the origin, and for the North and South quadrant by exchanging the roles of ν_1 and ν_2. Define the angular window for the lth direction as

$$\hat{v}_{j,l}(\boldsymbol{\nu}) = \hat{v}\left(2^{\lfloor j/2 \rfloor}\frac{\nu_2 - \nu_1 \tan\theta_l}{\nu_1}\right), \tag{9}$$

with the sequence of equi-spaced slopes (and not angles) $\tan\theta_l = 2^{-\lfloor j/2 \rfloor} l$, for $l = -2^{\lfloor j/2 \rfloor}, \ldots, 2^{\lfloor j/2 \rfloor} - 1$. We can now define the window which is the Cartesian analog of $\hat{\varphi}_j$ above,

$$\hat{u}_{j,l}(\boldsymbol{\nu}) = \hat{w}_j(\boldsymbol{\nu})\hat{v}_{j,l}(\boldsymbol{\nu}) = \hat{w}_j(\boldsymbol{\nu})\hat{v}_{j,0}(S_{\theta_l}\boldsymbol{\nu}), \tag{10}$$

where S_{θ_l} is the shear matrix. From this definition, it can be seen that $\hat{u}_{j,l}$ is supported near the trapezoidal wedge $\{\boldsymbol{\nu} = (\nu_1, \nu_2) | 2^j \leq \nu_1 \leq 2^{j+1}, -2^{-j/2} \leq \nu_2/\nu_1 - \tan\theta_l \leq 2^{-j/2}\}$. The collection of $\hat{u}_{j,l}(\boldsymbol{\nu})$ gives rise to the frequency tiling shown in Fig. 12c. From $\hat{u}_{j,l}(\boldsymbol{\nu})$, the digital CurveletG2 construction suggests Cartesian curvelets that are translated and sheared versions of a mother Cartesian curvelet $\hat{\varphi}_j^D(\boldsymbol{\nu}) = \hat{u}_{j,0}(\boldsymbol{\nu}), \varphi_{j,l,\mathbf{k}}^D(\mathbf{x}) = 2^{3j/4}\varphi_j^D\left(S_{\theta_l}^T\mathbf{x} - \mathbf{m}\right)$ where $\mathbf{m} = (k_1 2^{-j}, k_2 2^{-j/2})$.

Digital Implementation The goal here is to find a digital implementation of the Second Generation Discrete Curvelet Transform (DCTG2), whose coefficients are now given by

$$c_{j,l,\mathbf{k}} := \left\langle f\varphi_{j,l,\mathbf{k}}^D \right\rangle = \int_{\mathbb{R}^2} \hat{f}(\boldsymbol{\nu})\hat{\varphi}_j^D(S_{\theta_l}^{-1}\boldsymbol{\nu})e^{iS_{\theta_l}^{-T}\mathbf{m}\cdot\boldsymbol{\nu}}\mathrm{d}\boldsymbol{\nu}. \tag{11}$$

To evaluate this formula with discrete data, one may think of (i) using the 2D FFT to get $\hat{f}$, (ii) form the windowed frequency data $\hat{f}\hat{u}_{j,l}$ and (iii) apply the the inverse Fourier transform. But this necessitates to evaluate the FFT at the sheared grid $S_{\theta_l}^{-T}\mathbf{m}$, for which the classical FFT algorithm is not valid. Two implementations were then proposed [15], essentially differing in their way of handling the grid:

- A tilted grid mostly aligned with the axes of $\hat{u}_{j,l}(\boldsymbol{\nu})$ which leads to the Unequi-Spaced FFT (USFFT)-based DCTG2. This implementation uses a nonstandard interpolation. Furthermore, the inverse transform uses conjugate gradient iteration to invert the interpolation step. This will have the drawback of a higher computational burden compared to the wrapping-based implementation that we will discuss hereafter. We will not elaborate more about the USFFT implementation as we never use it in practice. The interested reader may refer to [15] for further details and analysis.
- A grid aligned with the input Cartesian grid which leads to the wrapping-based DCTG2.

The wrapping-based DCTG2 makes a simpler choice of the spatial grid to translate the curvelets. The curvelet coefficients are essentially the same as in (11), except that $S_{\theta_l}^{-T}\mathbf{m}$ is replaced by $\mathbf{m}$ with values on a rectangular grid. But again, a difficulty rises because the window $\hat{u}_{j,l}$ does not fit in a rectangle of size $2^j \times 2^{j/2}$ to which an inverse FFT could be applied. The *wrapping* trick consists in periodizing the windowed frequency data $\hat{f}\hat{u}_{j,l}$, and reindexing the samples array by wrapping around a $\sim 2^j \times 2^{j/2}$ rectangle centered at the origin, see Fig. 12d to get a gist of the wrapping idea.

The wrapping-based DCTG2 algorithm can be summarized as in Algorithm 2.

The DCTG2 implementation can assign either wavelets or curvelets at the finest scale. In the CurveLab toolbox [75], the default choice is set to wavelets at the finest scale, but this can be easily modified directly in the code.

Require: Input $n \times n$ image $f[i_1, i_2]$, coarsest decomposition scale, curvelets or wavelets at the finest scale.
1: Apply the 2D FFT and obtain Fourier samples $\hat{f}[i_1, i_2]$.
2: **for** each scale j and angle l **do**
3: Form the product $\hat{f}[i_1, i_2]\hat{u}_{j,l}[i_1, i_2]$.
4: Wrap this product around the origin.
5: Apply the inverse 2D FFT to the wrapped data to get discrete DCTG2 coefficients.
6: **end for**

Curvelets and Ridgelets, Algorithm 2
DCTG2 via wrapping

We would like to apologize to the expert reader as many technical details are (deliberately) missing here on the CurveletG2 construction. For instance, low-pass coarse component, window overlapping, windows over junctions between quadrants. This paper is intended to give an overview of these recent multi-scale transforms, and the genuinely interested reader may refer to the original papers of Candès, Donoho, Starck and co-workers for further details (see bibliography).

The computational complexity of the wrapping-based DCTG2 analysis and reconstruction algorithms is that of the FFT $O(n^2 \log n)$, and in practice, the computation time is that of 6 to 10 2D FFTs [15]. This is a faster algorithm compared to the DCTG1. The DCTG2 fast algorithm has participated to make the use of the curvelet transform more attractive in many applicative fields (see Sect. "Stylized Applications" for some of them). The DCTG2, as it is implemented in the CurveLab toolbox [75], has reasonable redundancy, at most ~ 7.8 (much higher in 3D) if curvelets are used at the finest scale. This redundancy can even be reduced down to 4 (and 8 in 3D) if we replace in this implementation the Meyer wavelet construction, which introduces a redundancy factor of 4, by another wavelet pyramidal construction, similar to the one presented in Sect. "One–dimensional Wavelet Transform" which has a redundancy less than 2 in any dimension. Our experiments have shown that this modification does not modify the results in denoising experiments. DCTG2 redundancy is anyway much smaller than the DCTG1 one which is $16J + 1$. As stated earlier, the DCTG2 coefficients are complex-valued, but a real-valued DCTG2 with the same redundancy factor can be easily obtained by properly combining coefficients at orientations θ_l and $\theta_l + \pi$.

The DCTG2 can be extended to higher dimensions [21]. In the same vein as wavelets on the interval [53], the DCGT2 has been recently adapted to handle image boundaries by mirror extension instead of periodization [25]. The latter modification can have immediate implications in image processing applications where the contrast difference at opposite image boundaries may be an issue (see e. g. the denoising experiment discussion reported in Sect. "Stylized Applications").

We would like to make a connection with other multiscale directional transforms directly linked to curvelets. The contourlets tight frame of Do and Vetterli [27] implements the CurveletG2 idea directly on a discrete grid using a perfect reconstruction filter bank procedure. In [51], the authors proposed a modification of the contourlets with a directional filter bank that provides a frequency partitioning which is close to the curvelets but with no redundancy. Durand in [35] recently introduced families of non-adaptive directional wavelets with various frequency tilings, including that of curvelets. Such families are non-redundant and form orthonormal bases for $L_2(\mathbb{R}^2)$, and have an implementation derived from a single nonseparable filter bank structure with nonuniform sampling.

Sparse Representation by Second Generation Curvelets It has been shown by Candès and Donoho [20] that with the CurveletG2 tight frame construction, the M-term nonlinear approximation error of C^2 images except at discontinuities along C^2 curves obey

$$\|f - f_M\|_2 \leq CM^{-2}(\log M)^3 .$$

This is an asymptotically optimal convergence rate (up to the $(\log M)^3$ factor), and holds uniformly over the $C^2 - C^2$ class of functions. This is a remarkable result since the CurveletG2 representation is non-adaptative. However, the simplicity due to the non-adaptivity of curvelets has a cost: curvelet approximations loose their near optimal properties when the image is composed of edges which are not exactly C^2. Additionally, if the edges are C^α-regular with $\alpha > 2$, then the curvelets convergence rate exponent remain 2. Other adaptive geometric representations such as bandlets are specifically designed to reach the optimal decay rate $O(M^{-\alpha})$ [49,62].

Stylized Applications

Denoising

Elongated Feature Recovery The ability of ridgelets to sparsely represent piecewise smooth images away from discontinuities along lines has an immediate implication on statistical estimation. Consider a piecewise smooth image f away from line singularities embedded in an additive white noise of standard deviation σ. The ridgelet-based thresholding estimator is nearly optimal for recovering such functions, with a mean-square error (MSE) decay rate almost as good as the minimax rate [12].

To illustrate these theoretical facts, we simulate a vertical band embedded in white Gaussian noise with large σ. Figure 14 (top left) represents such a noisy image. The parameters are as follows: the pixel width of the band is 20 and the signal-to-noise ratio (SNR) is set to 0.1. Note that it is not possible to distinguish the band by eye. The wavelet transform (undecimated wavelet transform) is also incapable of detecting the presence of this object; roughly speaking, wavelet coefficients correspond to weighted averages over approximately isotropic neighborhoods (at different scales) and those wavelets clearly do not correlate very well with the very elongated structure (pattern) of the object to be detected.

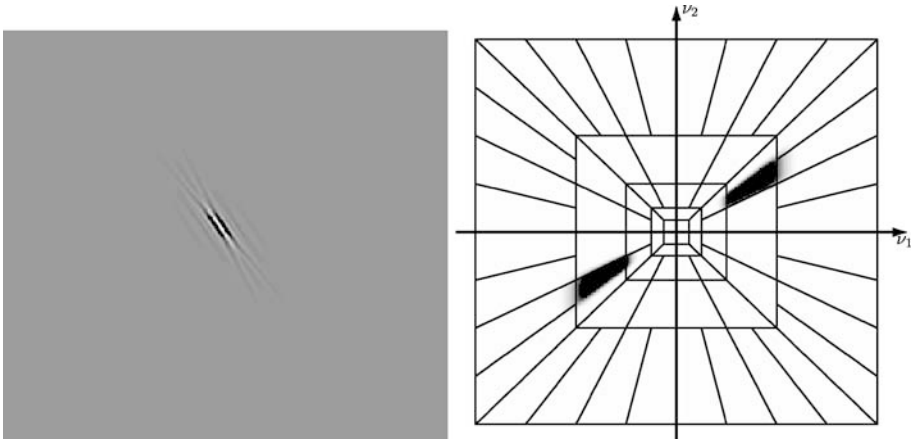

Curvelets and Ridgelets, Figure 13
An example of second generation real curvelet. *Left*: curvelet in spatial domain. *Right*: its Fourier transform

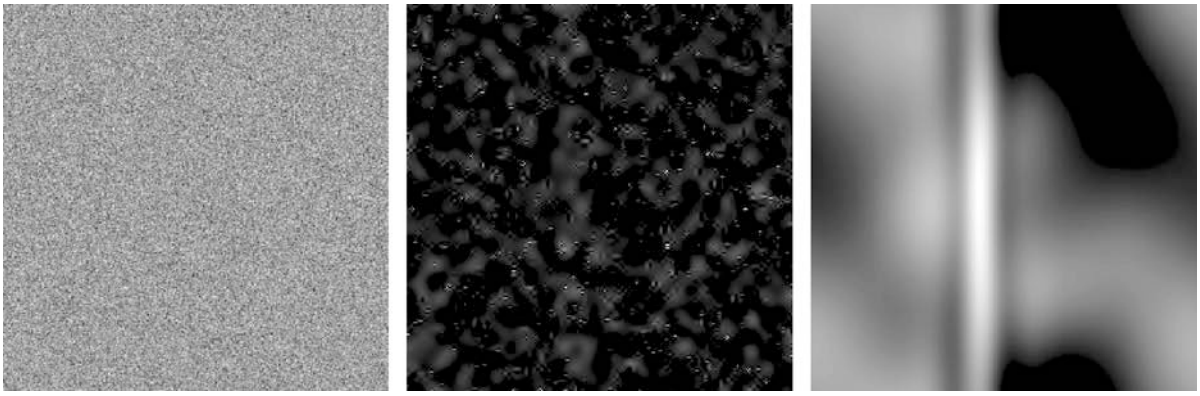

Curvelets and Ridgelets, Figure 14
Original image containing a vertical band embedded in white noise with relatively large amplitude (*left*). Denoised image using the undecimated wavelet transform (*middle*). Denoised image using the DRT based on the rectopolar Radon transform (*right*)

Curve Recovery Consider now the problem of recovering a piecewise C^2 function f apart from a discontinuity along a C^2 edge. Again, a simple strategy based on thresholding curvelet tight frame coefficients yields an estimator that achieves a MSE almost of the order $O(\sigma^{4/3})$ uniformly over the $C^2 - C^2$ class of functions [19]. This is the optimal rate of convergence as the minimax rate for that class scales as $\sigma^{4/3}$ [19]. Comparatively, wavelet thresholding methods only achieves a MSE of order $O(\sigma)$ and no better. We also note that the statistical optimality of the curvelet thresholding extends to a large class of ill-posed linear inverse problems [19].

In the experiment of Fig. 15, we have added a white Gaussian noise to "War and Peace", a drawing from Picasso which contains many curved features. Figure 15 bottom left and right shows respectively the restored images by the undecimated wavelet transform and the DCTG1. Curves are more sharply recovered with the DCTG1.

In a second experiment, we compared the denoising performance of several digital implementations of the curvelet transform; namely the DCTG1 with the rectopolar DRT, the DCTG1 with the FSS-based DRT and the wrapping-based DCTG2. The results are shown in Fig. 16, where the original 512×512 `Peppers` image was corrupted by a Gaussian white noise $\sigma = 20$ (PSNR = 22dB).

Although the FSS-based DRT is more accurate than the rectopolar DRT, the denoising improvement of the former (PSNR=31.31dB) is only 0.18 dB better than the latter (PSNR=31.13dB) on `Peppers`. The difference is almost undistinguishable by eye, but the computation time is 20 higher for the DCTG1 with the FSS DRT. Consequently, it appears that there is a little benefit of using the

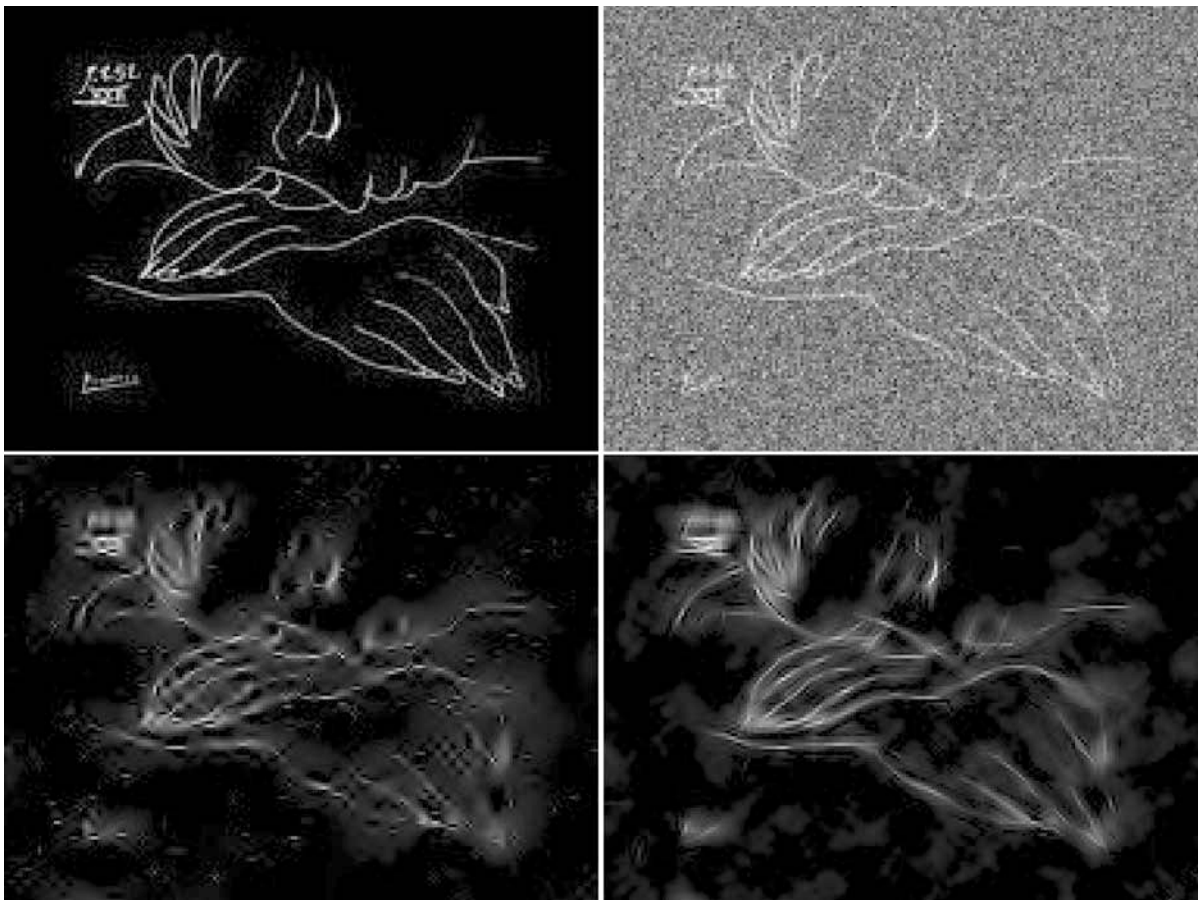

Curvelets and Ridgelets, Figure 15
The Picasso picture `War and Peace` (*top left*), the same image contaminated with a Gaussian white noise (*top right*). The restored images using the undecimated wavelet transform (*bottom left*) and the DCTG1 (*bottom right*)

FSS DRT in the DCTG1 for restoration applications. Denoising using the DCTG2 with the wrapping implementation gives a PSNR=30.62 dB which is ~ 0.7 dB less than the DCTG1. But this is the price to pay for a lower redundancy and a much faster transform algorithm. Moreover, the DCTG2 exhibits some artifacts which look like 'curvelet ghosts'. This is a consequence of the fact that the DCTG2 makes a central use of the FFT which has the side effect of treating the image boundaries by periodization.

Linear Inverse Problems

Many problems in image processing can be cast as inverting the linear degradation equation $y = Hf + \varepsilon$, where f is the image to recover, y the observed image and ε is a white noise of variance $\sigma^2 < +\infty$. The linear mapping H is generally ill-behaved which entails ill-posedness of the inverse problem. Typical examples of linear inverse problems include image deconvolution where H is the convolution operator, or image inpainting (recovery of missing data) where H is a binary mask.

In the last few years, some authors have attacked the problem of solving linear inverse problems under the umbrella of sparse representations and variational formulations, e. g. for deconvolution [19,24,38,41,74] and inpainting [37,39]. Typically, in this setting, the recovery of f is stated as an optimization problem with a sparsity-promoting regularization on the representation coefficients of f, e. g. its wavelet or curvelet coefficients. See [37,38,39,74] for more details.

In Fig. 17 first row, we depict an example of deconvolution on `Barbara` using the algorithm described in [38] with the DCTG2 curvelet transform. The original, degraded (blurred with an exponential kernel and noisy) and restored images are respectively shown on the left, middle

Curvelets and Ridgelets, Figure 16
Comparison of denoising performance of several digital implementations of the curvelet transform. *Top left*: original image. *Top right*: noisy image $\sigma = 20$. *Bottom left*: denoised with DCTG1 using the rectopolar DRT. *Bottom middle*: denoised with DCTG1 using the FSS DRT. Bottom right: denoised with DCTG2 using the wrapping implementation

and right. The second row gives an example of inpainting on `Claudia` image using the DCTG2 with 50% missing pixels.

Contrast Enhancement

The curvelet transform has been successfully applied to image contrast enhancement by Starck et al. [73]. As the curvelet transform capture efficiently edges in an image, it is a good candidate for multiscale edge enhancement. The idea is to modify the curvelet coefficients of the input image in order to enhance its edges. The curvelet coefficients are typically modified according to the function displayed in the left plot of Fig. 18. Basically, this plot says that the input coefficients are kept intact (or even shrunk) if they have either low (e. g. below the noise level) or high (strongest edges) values. Intermediate curvelet coefficient values which correspond to the faintest edges are amplified. An example of curvelet-based image enhancement on `Saturn` image is given in Fig. 18.

Morphological Component Separation

The idea to morphologically decompose a signal/image into its building blocks is an important problem in signal and image processing. Successful separation of a signal content has a key role in the ability to effectively analyze it, enhance it, compress it, synthesize it, and more. Various approaches have been proposed to tackle this problem.

The Morphological Component Analysis method (MCA) [70,71] is a method which allows us to decompose a single signal into two or more layers, each layer containing only one kind of feature in the input signal or image. The separation can be achieved when each kind of feature is sparsely represented by a given transformation in the dictionary of transforms. Furthermore, when a transform sparsely represents a part in the signal/image, it yields nonsparse representation on the other content type. For instance, lines and Gaussians in a image can be separated using the ridgelet transform and the wavelet transform. Locally oscillating textures can be separated from the piecewise smooth content using the local discrete cosine trans-

Curvelets and Ridgelets, Figure 17
Illustration of the use of curvelets (DCTG2 transform) when solving two typical linear inverse problems: deconvolution (*first row*), and inpainting (*second row*). First row: deconvolution of `Barbara` image, original (*left*), blurred and noisy (*middle*), restored (*right*). Second row: inpainting of `Claudia` image, original (*left*), masked image (*middle*), inpainted (*right*)

Curvelets and Ridgelets, Figure 18
Curvelet contrast enhancement. Left: enhanced vs original curvelet coefficient. Middle: original `Saturn` image. Right: result of curvelet-based contrast enhancement

form and the curvelet transform [70]. A full description of MCA is given in [70].

The first row of Fig. 19 illustrates a separation result when the input image contains only lines and isotropic Gaussians. Two transforms were amalgamated in the dictionary; namely the à trous WT2D and the DRT. The left, middle and right images in the first row of Fig. 19 represent respectively, the original image, the reconstructed

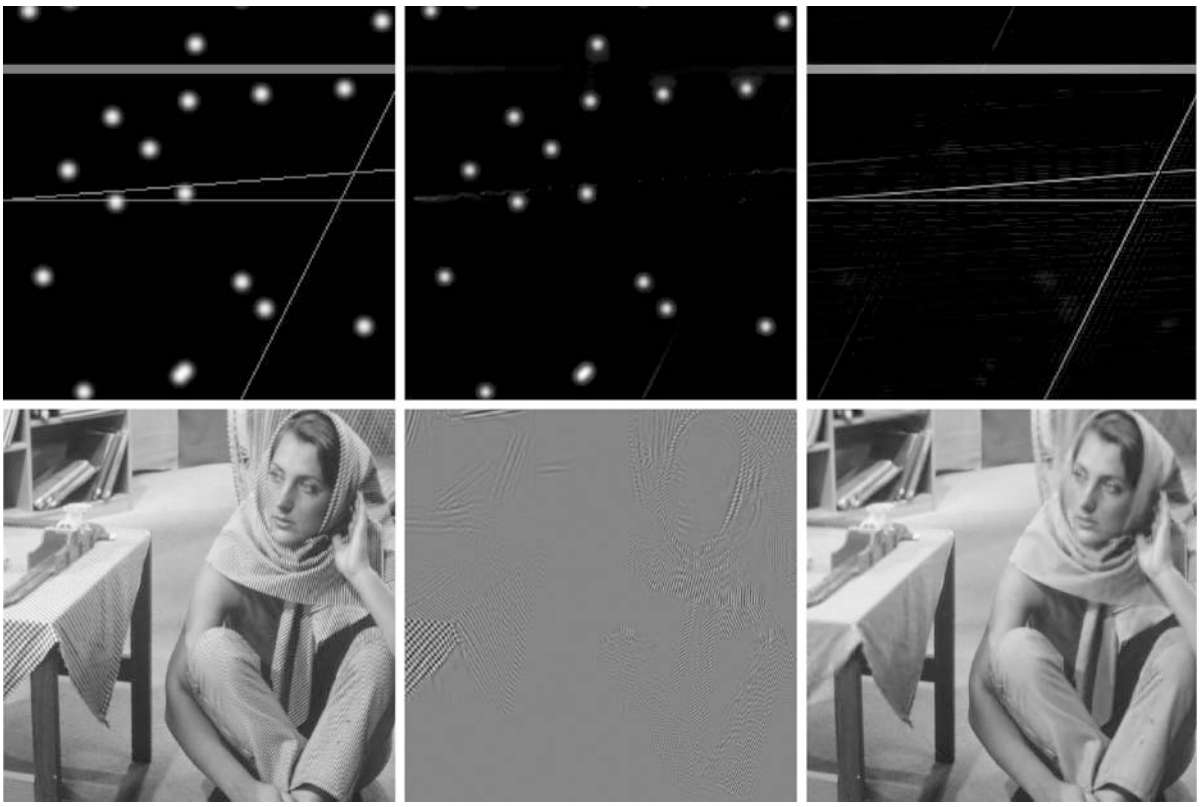

Curvelets and Ridgelets, Figure 19
***First row*, from *left* to *right*: original image containing lines and Gaussians, separated Gaussian component (*wavelets*), separated line component (*ridgelets*). *Second row*, from *left* to *right*: original `Barbara` image, reconstructed local discrete cosine transform part (*texture*), and piecewise smooth part (*curvelets*)**

component from the à trous wavelet coefficients, and the reconstructed layer from the ridgelet coefficients. The second row of Fig. 19 shows respectively the `Barbara` image, the reconstructed local cosine (textured) component and the reconstructed curvelet component. In the `Barbara` example, the dictionary contained the local discrete cosine and the DCTG2 transforms.

Future Directions

In this paper, we gave an overview of two important geometrical multiscale transforms; namely ridgelets and curvelets. We illustrate their potential applicability on a wide range of image processing problems. Although these transforms are not adaptive, they are strikingly effective both theoretically and practically on piecewise images away from smooth contours.

However, in image processing, the geometry of the image and its regularity is generally not known in advance. Therefore, to reach higher sparsity levels, it is necessary to find representations that can adapt themselves to the geometrical content of the image. For instance, geometric transforms such as wedgelets [30] or bandlets [49,50,62] allow to define an adapted multiscale geometry. These transforms perform a non-linear search for an optimal representation. They offer geometrical adaptivity together with stable algorithms. Recently, Mallat [54] proposed a more biologically inspired procedure named the grouplet transform, which defines a multiscale association field by grouping together pairs of wavelet coefficients.

In imaging science and technology, there is a remarkable proliferation of new data types. Beside the traditional data arrays defined on uniformly sampled cartesian grids with scalar-valued samples, many novel imaging modalities involve data arrays that are either (or both):

- acquired on specific "exotic" grids such as in astronomy, medicine and physics. Examples include data defined on spherical manifolds such as in astronomical

imaging, catadioptric optical imaging where a sensor overlooks a paraboloidal mirror, etc.
- or with samples taking values in a manifold. Examples include vector fields such as those of polarization data that may rise in astronomy, rigid motions (a special Euclidean group), definite-positive matrices that are encountered in earth science or medical imaging, etc.

The challenge faced with this data is to find multiscale representations which are sufficiently flexible to apply to many data types and yet defined on the proper grid and respect the manifold structure. Extension of wavelets, curvelets and ridgelets for scalar-valued data on the sphere has been proposed recently by [66]. Construction of wavelets for scalar-valued data defined on graphs and some manifolds was proposed by [22]. The authors in [63][see references therein] describe multiscale representations for data observed on equispaced grids and taking values in manifolds such as: the sphere, the special orthogonal group, the positive definite matrices, and the Grassmannian manifolds. Nonetheless many challenging questions are still open in this field: extend the idea of multiscale geometrical representations such as curvelets or ridgelets to manifold-valued data, find multiscale geometrical representations which are sufficiently general for a wide class of grids, etc. We believe that these directions are one of the hottest topics in this field.

Most of the transforms discussed in this paper can handle efficiently smooth or piecewise smooth functions. But sparsely representing textures remains an important open question, mainly because there is no consensus on how to define a texture. Although Julesz [45] stated simple axioms about the probabilistic characterization of textures. It has been known for some time now that some transforms can sometimes enjoy reasonably sparse expansions of certain textures; e. g. oscillatory textures in bases such as local discrete cosines [70], brushlets [56], Gabor [53], complex wavelets [46]. Gabor and wavelets are widely used in the image processing community for texture analysis. But little is known on the decay of Gabor and wavelet coefficients of "texture". If one is interested in synthesis as well as analysis, the Gabor representation may be useless (at least in its strict definition). Restricting themselves to locally oscillating patters, Demanet and Ying have recently proposed a wavelet-packet construction named WaveAtoms [77]. They showed that WaveAtoms provide optimally sparse representation of warped oscillatory textures.

Another line of active research in sparse multiscale transforms was initiated by the seminal work of Olshausen and Field [59]. Following their footprints, one can push one step forward the idea of adaptive sparse representation and requires that the dictionary is not fixed but rather optimized to sparsify a set of exemplar signals/images, i. e. patches. Such a learning problem corresponds to finding a sparse matrix factorization and several algorithms have been proposed for this task in the literature; see [1] for a good overview. Explicit structural constraints such as translation invariance can also be enforced on the learned dictionary [5,58]. These learning-based sparse representations have shown a great improvement over fixed (and even adapted) transforms for a variety of image processing tasks such as denoising and compression [8,36,52], linear inverse problems (image decomposition and inpainting) [61], texture synthesis [60].

Bibliography

1. Aharon M, Elad M, Bruckstein AM (2006) The K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Process 54(11):4311–4322
2. Arivazhagan S, Ganesan L, Kumar TS (2006) Texture classification using curvelet statistical and co-occurrence feature. In: Proceedings of the 18th International Conference on Pattern Recognition, Hong Kong (ICPR 2006), vol 2. pp 938–941
3. Averbuch A, Coifman RR, Donoho DL, Israeli M, Waldén J (2001) Fast Slant Stack: A notion of Radon transform for data in a cartesian grid which is rapidly computible, algebraically exact, geometrically faithful and invertible. SIAM J Sci Comput
4. BeamLab 200 (2003) http://www-stat.stanford.edu/beamlab/
5. Blumensath T, Davies M (2006) Sparse and shift-invariant representations of music. IEEE Trans Speech Audio Process 14(1):50–57
6. Bobin J, Moudden Y, Starck JL, Elad M (2006) Morphological diversity and source separation. IEEE Trans Signal Process 13(7):409–412
7. Bobin J, Starck JL, Fadili J, Moudden Y (2007) Sparsity, morphological diversity and blind source separation. IEEE Transon Image Process 16(11):2662–2674
8. Bryt O, Elad M (2008) Compression of facial images using the k-svd algorithm. J Vis Commun Image Represent 19(4):270–283
9. Candès EJ (1998) Ridgelets: theory and applications. Ph D thesis, Stanford University
10. Candès EJ (1999) Harmonic analysis of neural networks. Appl Comput Harmon Anal 6:197–218
11. Candès EJ (1999) Ridgelets and the representation of mutilated sobolev functions. SIAM J Math Anal 33:197–218
12. Candès EJ (1999) Ridgelets: Estimating with ridge functions. Ann Statist 31:1561–1599
13. Candès EJ, Demanet L (2002) Curvelets and Fourier integral operators. C R Acad Sci, Paris, Serie I(336):395–398
14. Candès EJ, Demanet L (2005) The curvelet representation of wave propagators is optimally sparse. Com Pure Appl Math 58(11):1472–1528
15. Candès EJ, Demanet L, Donoho DL, Ying L (2006) Fast discrete curvelet transforms. SIAM Multiscale Model Simul 5(3):861–899

16. Candès EJ, Donoho DL (1999) Ridgelets : the key to high dimensional intermittency? Philos Trans Royal Soc of Lond A 357:2495–2509
17. Candès EJ, Donoho DL (1999) Curvelets – a surprisingly effective nonadaptive representation for objects with edges. In: Cohen A, Rabut C, Schumaker LL (eds) Curve and Surface Fitting: Saint-Malo. Vanderbilt University Press, Nashville
18. Candès EJ, Donoho DL (2000) Curvelets and curvilinear integrals. Approx J Theory 113:59–90
19. Candès EJ, Donoho DL (2000) Recovering edges in ill-posed inverse problems: Optimality of curvelet frames. Ann Stat 30:784–842
20. Candès EJ, Donoho DL (2002) New tight frames of curvelets and optimal representations of objects with piecewise-C2 singularities. Comm Pure Appl Math 57:219–266
21. Candès E,Ying L, Demanet L (2005) 3d discrete curvelet transform. In: Wavelets XI conf., San Diego. Proc. SPIE, vol 5914, 591413; doi:10.1117/12.616205
22. Coifman RR, Maggioni M (2006) Diffusion wavelets. Appl Comput Harmon Anal 21:53–94
23. Daubechies I (1992) Ten Lectures on Wavelets. Society for Industrial and Applied Mathematics, Philadelphia
24. Daubechies I, Defrise M, De Mol C (2004) An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Com Pure Appl Math 57:1413–1541
25. Demanet L, Ying L (2007) Curvelets and wave atoms for mirror-extended images. In: Wavelets XII conf., San Diego. Proc. SPIE, vol 6701, 67010J; doi:10.1117/12.733257
26. Do MN, Vetterli M (2003) The finite ridgelet transform for image representation. IEEE Trans Image Process 12(1):16–28
27. Do MN, Vetterli M (2003) Contourlets. In: Stoeckler J, Welland GV (eds) Beyond Wavelets. Academic Press, San Diego
28. Donoho DL (1997) Fast ridgelet transforms in dimension 2. Stanford University, Stanford
29. Donoho DL (1998) Digital ridgelet transform via rectopolar coordinate transform. Technical Report, Stanford University
30. Donoho DL (1999) Wedgelets: nearly-minimax estimation of edges. Ann Stat. 27:859–897
31. Donoho DL (2000) Orthonormal ridgelets and linear singularities. SIAM J Math Anal 31(5):1062–1099
32. Donoho DL, Duncan MR (2000) Digital curvelet transform: strategy, implementation and experiments. In: Szu HH, Vetterli M, Campbell W, Buss JR (eds) Proc. Aerosense, Wavelet Applications VII, vol 4056. SPIE, pp 12–29
33. Donoho DL, Flesia AG (2002) Digital ridgelet transform based on true ridge functions. In: Schmeidler J, Welland GV (eds) Beyond Wavelets. Academic Press, San Diego
34. Douma H, de Hoop MV (2007) Leading-order seismic imaging using curvelets. Geophys 72(6)
35. Durand S (2007) M-band filtering and nonredundant directional wavelets. Appl Comput Harmon Anal 22(1):124–139
36. Elad M, Aharon M (2006) Image denoising via sparse and redundant representations over learned dictionaries. IEEE Trans Image Process 15(12):3736–3745
37. Elad M, Starck JL, Donoho DL, Querre P (2006) Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA). J Appl Comput Harmon Anal 19:340–358
38. Fadili MJ, Starck JL (2006) Sparse representation-based image deconvolution by iterative thresholding. In: Murtagh F, Starck JL (eds) Astronomical Data Analysis IV. Marseille
39. Fadili MJ, Starck JL, Murtagh F (2006) Inpainting and zooming using sparse representations. Comput J
40. Fernandes F, Wakin M, Baraniuk R (2004) Non-redundant, linear-phase, semi-orthogonal, directional complex wavelets. In: IEEE Conf. on Acoustics, Speech and Signal Processing, Montreal, Canada, vol 2, pp 953 956
41. Figueiredo M, Nowak R (2003) An EM algorithm for wavelet-based image restoration. IEEE Trans Image Process 12(8):906–916
42. Hennenfent G, Herrmann FJ (2006) Seismic denoising with nonuniformly sampled curvelets. IEEE Comput Sci Eng 8(3):16–25
43. Herrmann FJ, Moghaddam PP, Stolk CC (2008) Sparsity- and continuity-promoting seismic image recovery with curvelet frames. Appl Comput Harmon Anal 24(2):150–173
44. Jin J, Starck JL, Donoho DL, Aghanim N, Forni O (2005) Cosmological non-gaussian signatures detection: Comparison of statistical tests. Eurasip J 15:2470–2485
45. Julesz B (1962) Visual pattern discrimination. Trans RE Inform Theory 8(2):84–92
46. Kingsbury NG (1998) The dual-tree complex wavelet transform: a new efficient tool for image restoration and enhancement. In: European Signal Processing Conference, Rhodes, Greece, pp 319–322
47. Labate D, Lim W-Q, Kutyniok G, Weiss G (2005) Sparse multidimensional representation using shearlets. In: Wavelets XI, vol 5914. SPIE, San Diego, pp 254–262
48. Lambert P, Pires S, Ballot J, García RA, Starck JL, Turck-Chièze S (2006) Curvelet analysis of asteroseismic data. i. method description and application to simulated sun-like stars. Astron Astrophys 454:1021–1027
49. Le Pennec E, Mallat S (2000) Image processing with geometrical wavelets. In: International Conference on Image Processing, Thessaloniki, Greece
50. Le Pennec E, Mallat S (2005) Bandelet Image Approximation and Compression. Multiscale SIAM Mod Simul 4(3):992–1039
51. Lu Y, Do MN (2003) CRIPS-contourlets: A critical sampled directional multiresolution image representation. In: Wavelet X, San Diego. Proc. SPIE, vol 5207. pp 655–665
52. Mairal J, Elad M, Sapiro G (2008) Sparse representation for color image restoration. IEEE Trans Image Process 17(1):53–69
53. Mallat S (1998) A Wavelet Tour of Signal Processing. Academic Press, London
54. Mallat S (2008) Geometrical grouplets. Appl Comput Harmon Anal
55. Matus F, Flusser J (1993) Image representations via a finite Radon transform. IEEE Trans Pattern Anal Mach Intell 15(10):996–1006
56. Meyer FG, Coifman RR (1997) Brushlets: a tool for directional image analysis and image compression. Appl Comput Harmon Anal 4:147–187
57. Meyer Y (1993) Wavelets: Algorithms and Applications. SIAM, Philadelphia
58. Olshausen BA (2000) Sparse coding of time-varying natural images. In: Int. Conf. Independent Component Analysis and Blind Source Separation (ICA), Barcelona, pp 603–608
59. Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive-field properties by learning a sparse code for natural images. Nature 381(6583):607–609

60. Peyré G (2007) Non-negative sparse modeling of textures. In: SSVM. Lecture Notes in Computer Science. Springer, pp 628–639
61. Peyré G, Fadili MJ, Starck JL (2007) Learning adapted dictionaries for geometry and texture separation. In: Wavelet XII, San Diego. Proc. SPIE, vol 6701, 67011T; doi:10.1117/12.731244
62. Peyré G, Mallat S (2007) A review of bandlet methods for geometrical image representation. Num Algorithms 44(3):205–234
63. Rahman I, Drori I, Stodden VC, Donoho DL, Schröder P (2005) Multiscale representations fpr manifold-valued data. Multiscale Mod Simul 4(4):1201–1232
64. Saevarsson B, Sveinsson J, Benediktsson J (2003) Speckle reduction of SAR images using adaptive curvelet domain. In: Proceedings of the IEEE International Conference on Geoscience and Remote Sensing Symposium, IGARSS '03, vol 6., pp 4083–4085
65. Simoncelli EP, Freeman WT, Adelson EH, Heeger DJ (1992) Shiftable multi-scale transforms [or what's wrong with orthonormal wavelets]. IEEE Trans Inf Theory 38(2):587–607
66. Starck JL, Abrial P, Moudden Y, Nguyen M (2006) Wavelets, ridgelets and curvelets on the sphere. Astron Astrophys 446:1191–1204
67. Starck JL, Bijaoui A, Lopez B, Perrier C (1994) Image reconstruction by the wavelet transform applied to aperture synthesis. Astron Astrophys 283:349–360
68. Starck JL, Candès EJ, Donoho DL (2002) The curvelet transform for image denoising. IEEE Trans Image Process 11(6):131–141
69. Starck JL, Candès EJ, Donoho DL (2003) Astronomical image representation by the curvelet tansform. Astron Astrophys 398:785–800
70. Starck JL, Elad M, Donoho DL (2004) Redundant multiscale transforms and their application for morphological component analysis. In: Hawkes P (ed) Advances in Imaging and Electron Physics, vol 132. Academic Press, San Diego, London, pp 288–348
71. Starck JL, Elad M, Donoho DL (2005) Image decomposition via the combination of sparse representation and a variational approach. IEEE Trans Image Process 14(10):1570–1582
72. Starck JL, Murtagh F, Bijaoui A (1998) Image Processing and Data Analysis: The Multiscale Approach. Cambridge University Press, Cambridge
73. Starck JL, Murtagh F, Candès E, Donoho DL (2003) Gray and color image contrast enhancement by the curvelet transform. IEEE Trans Image Process 12(6):706–717
74. Starck JL, Nguyen MK, Murtagh F (2003) Wavelets and curvelets for image deconvolution: a combined approach. Signal Process 83(10):2279–2283
75. The Curvelab Toolbox (2005) http://www.curvelet.org
76. Velisavljevic V, Beferull-Lozano B, Vetterli M, Dragotti PL (2006) Directionlets: Anisotropic multi-directional representation with separable filtering. IEEE Trans Image Process 15(7):1916–1933
77. Ying L, Demanet L (2007) Wave atoms and sparsity of oscillatory patterns. Appl Comput Harmon Anal 23(3):368–387
78. Zhang Z, Huang W, Zhang J, Yu H, Lu Y (2006) Digital image watermark algorithm in the curvelet domain. In: Proceedings of the International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP'06), Pasadena. pp 105–108

Cytoskeleton and Cell Motility

Thomas Risler[1,2]
[1] Institut Curie, Section de Recherche, Laboratoire Physicochimie Curie (CNRS-UMR 168), Paris, France
[2] Université Pierre et Marie Curie Paris VI, Paris, France

Article Outline

Glossary
Definition of the Subject
Introduction
The Diversity of Cell Motility
The Cell Cytoskeleton
Filament-Driven Motility
Motor-Driven Motility
Putting It Together: Active Polymer Solutions
Extensions and Future Directions
Acknowledgments
Bibliography

Glossary

Cell Structural and functional elementary unit of all life forms. The cell is the smallest unit that can be characterized as living.

Eukaryotic cell Cell that possesses a nucleus, a small membrane-bounded compartment that contains the genetic material of the cell. Cells that lack a nucleus are called *prokaryotic cells* or *prokaryotes*.

Domains of life *archaea*, *bacteria* and *eukarya* – or in English *eukaryotes*, and made of eukaryotic cells – which constitute the three fundamental branches in which all life forms are classified. Archaea and bacteria are prokaryotes. All multicellular organisms are eukaryotes, but eukaryotes can also be single-cell organisms. Eukaryotes are usually classified into four kingdoms: animals, plants, fungi and protists.

Motility Spontaneous, self-generated movement of a biological system.

Cytoskeleton System of protein filaments crisscrossing the inner part of the cell and which, with the help of the many proteins that interact with it, enables the cell to insure its structural integrity and morphology, exert forces and produce motion.

Amoeboid motility Crawling locomotion of a eukaryotic cell by means of protrusion of its leading edge.

Molecular motor Motor of molecular size. In this context, protein or macromolecular complex that converts a specific source of energy into mechanical work.

Filament Here, extended unidimensional structure made of an assembly of repeated protein units that hold together via physical interactions (without covalent bonds). A filament will be either a single *polymer* (or here *biopolymer*), a linear assembly of such polymers, or a linear assembly of molecular motors.

Active gel Cross-linked network of linear or branched polymers interacting by physical means, and that is dynamically driven out of equilibrium by a source of energy.

Definition of the Subject

We, as human beings, are made of a collection of cells, which are most commonly considered as the elementary building blocks of all living forms on earth [5]. Whether they belong to each of the three domains of life (*archaea*, *bacteria* or *eukarya*), cells are small membrane-bounded compartments that are capable of homeostasis, metabolism, response to their environment, growth, reproduction, adaptation through evolution and, at the cellular as well as multicellular level, organization. In addition, spontaneous, self-generated movement – also known as *motility* – is one of the properties that we most closely associate with all life forms. Even in the case of apparently inanimate living forms on macroscopic scales, like most plants and fungi, constitutive cells are constantly remodeling their internal structure for the entire organism to perform its metabolism, growth and reproduction [29]. In animals like human beings, cell motility is at the basis of most – if not all – essential processes participating in their lifetime, from their development, maintenance, to eventual death. It is indeed crucially involved for example in embryonic development (where individual as well as collective motions of cells underly morphogenesis), wound healing and recovery from injuries (where cellular migration is essential for tissue repair and regeneration), as well as immune response and most of disease progressions. There, on a biomedical point of view, cellular motility is involved in processes as diverse as neutrophils (white blood cells) and macrophages (cells that ingest bacteria) progressions, axonal regrowth after injuries, multiple sclerosis and cancer metastases. In addition, motility defects of the animal cells themselves can lead to a variety of inherited health problems, including male infertility, deafness and chronic inflammatory diseases.

Introduction

Cell movement was observed and reported for the first time as early as 1674, when Anthony van Leeuwenhoek brought a glass bead that served him as a primitive microscope close to a drop of water taken from a pool. His astonishment was immediate, as he later reported: "… the motion of these animalcules in the water was so swift and various, upwards, downwards and round about, that it was wonderful to see …" [29]. The organisms he saw were probably ciliated *protozoa* – unicellular non-photosynthetic eukaryotic organisms – a fraction of a millimeter in length, swimming by the agitated but coordinated motion of sometimes thousands of hairlike cilia on their surface (see Fig. 1). Despite this very early observation, only relatively recent advances of the past few decades in microscopy, molecular biology and biochemistry have enabled the discovery of the basic underlying molecular mechanisms by which cells are able to feel their environment, exert forces and move in a directed way in search for nutrients or any other task they need to perform. The cytoskeleton, defined as the system of protein filaments that enable the cell to insure its structural integrity and morphology, exert forces and produce motion, was first observed by H. E. Huxley and J. Hanson in

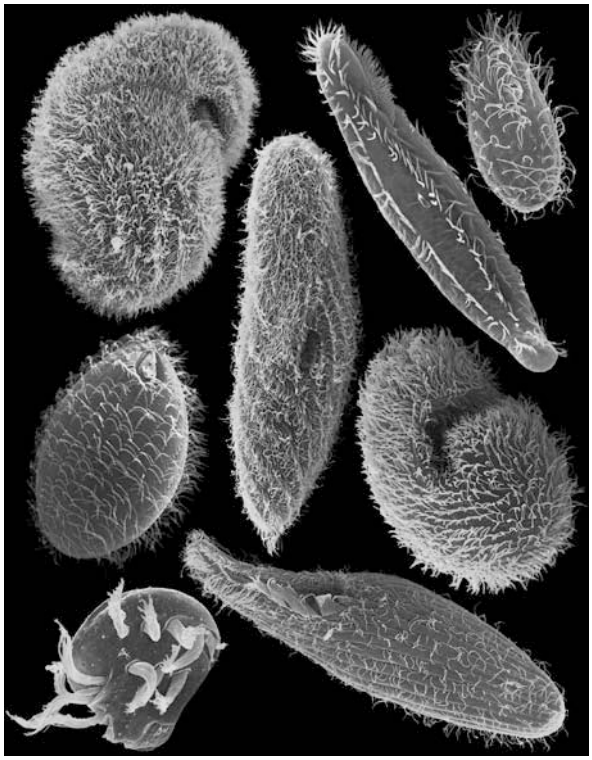

Cytoskeleton and Cell Motility, Figure 1
Electron micrographs of different species of ciliated *protozoa*. Almost all members of the protozoan group are non-pathogenic free-living organisms. Source: Foissner, W. and Zankl, A. (unpublished)

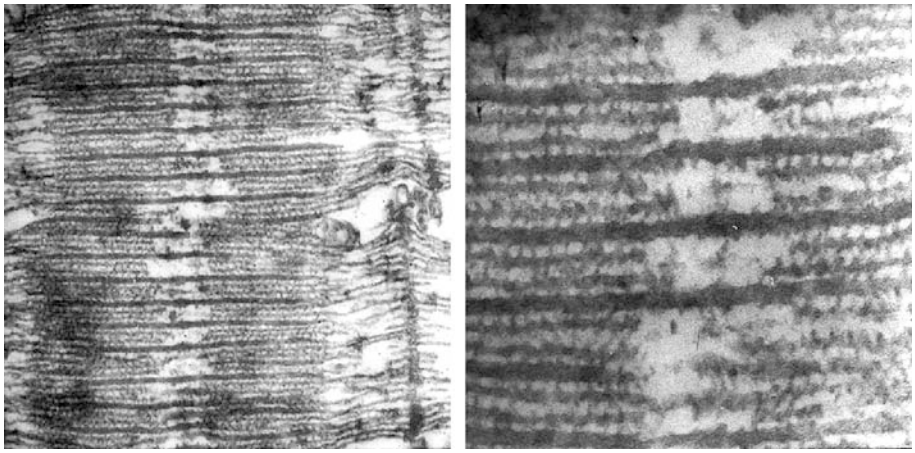

Cytoskeleton and Cell Motility, Figure 2

Skeletal-muscle thick and thin filaments as seen by H. E. Huxley in 1957 [112] (Reproductions from the original 1953 paper [93] were poor). *Left panel*: Thin longitudinal sections of rabbit psoas muscle fibers, showing a single layer of a filament lattice, with individual thick and thin filaments as well as crossbridges between them. *Right panel*: Higher-magnification view of a thin longitudinal section. The relative dimensions were distorted due to axial compression during sectioning: crossbridges' axial spacing is ≃40 nm and thick filaments' diameter is ≃12 nm. Source: reprinted from [114] with permission from Blackwell Publishing Ltd. (based on an original micrograph of 1957, see also [112])

1953, when they discovered the double array of filaments in cross-striated muscles using electron-microscopy techniques [93,110,111]. In parallel with A. F. Huxley and R. Niedergerke, but independently, they published the next year the "sliding-filament model", which explains muscle contraction via the relative sliding of two different types of filaments, originally called "thick" and "thin" filaments [109,115] (see Fig. 2). This, with the help of further genetic, biochemical and crystallographic studies, dated the beginning of a scientific understanding of the subcellular mechanisms that underly cell motility.

In addition to the characterization of the biochemical composition and organization of these subcellular structures, tremendous advances in the past two decades on both physical micro-manipulation and fluorescence-microscopy techniques have enabled the characterization of the processes involved with minute details. On the one hand, thanks to the help of micro-pipettes, atomic-force microscopes and optical tweezers, one can characterize the forces that cells are able to exert as well as their responses to applied stimuli. On the other hand, fluorescence-microscopy techniques provide information on the microscopic dynamics of single molecules in vivo. Finally, combined with biochemistry and gene-expression control, as well as the micro-fabrication of bio-mimetic artificial systems in in vitro assays, these techniques have enabled the study of simplified systems, where some specific aspects of the processes involved can be characterized separately.

In addition to these biochemical and behavioral characterizations, understanding the generic principles that underly cell motility needed an integrated approach to explain how this complex molecular machinery can self-organize and lead to a coherent, purposeful movement at the cellular level. Nothing better than a eukaryotic cell can indeed be categorized as a complex system, in that its behavior integrates the coordinated interplay of more than ten thousand different protein types, numbering together millions and representing 60% of its dry mass [5]. The cytoskeletal machinery is made of hundreds of different molecular players. For example, in 2003, about 160 proteins were known to bind to actin, one of the major biopolymers participating in cell structure and dynamical behavior [53]. Knowledge about this biomolecular machinery is constantly evolving, and its undergoing complexity can be appreciated by consulting up-to-date information on available databases[1]. Therefore, understanding this complexity and describing how it is integrated at the cellular level was made by biophysical studies, both on experimental and theoretical grounds, which helped to identify the generic principles behind cell motility. At the molecular level first, the conversion of chemical energy stored in covalent bounds into mechanical work relies on out-of-equilibrium thermodynamic principles and asymmetrical properties – or polarity – of the structures involved, and happens in a highly-fluctuating environment of brownian particles [13,233]. On larger length scales, the appearance of coordinated motion in large collections of proteins relies on collective phenomena, self-organi-

[1] See, for example, http://www.cytoskeletons.com/database.php.

zation and dynamical symmetry breakings [129]. On yet larger length scales, swimming of microorganisms has attracted the attention of physicists for years [228,263], and morphogenesis and pattern formations in cellular tissues rely on self-organization phenomena, as was envisioned for the first time by Turing in 1952 [141,269]. Therefore, in addition to biophysical experimental techniques, variety of theoretical physics' disciplines spanning the theory of stochastic processes, statistical physics, out-of-equilibrium thermodynamics, hydrodynamics, nonlinear dynamics and pattern formation have contributed and still contribute to our understanding of cell motility.

The present article will mainly focus on the eukaryotic cytoskeleton and cell-motility mechanisms. Bacterial motility as well as the composition of the prokaryotic cytoskeleton will be only briefly mentioned. The article is organized as follows. In Sect. "The Diversity of Cell Motility", we will first present an overview of the diversity of cellular motility mechanisms, which might at first glance be categorized into two different types of behaviors, namely "swimming" and "crawling". Intracellular transport, mitosis – or cell division – as well as other extensions of cell motility that rely on the same essential machinery will be briefly sketched. In Sect. "The Cell Cytoskeleton", we will introduce the molecular machinery that underlies cell motility – the cytoskeleton – as well as its interactions with the external environment of the cell and its main regulatory pathways. Sections "Other Cytoskeleton-Associated Proteins" to "The Prokaryotic Cytoskeleton" are more detailed in their biochemical presentations; readers primarily interested in the theoretical modeling of cell motility might want to skip these sections in a first reading. We will then describe the motility mechanisms that rely essentially on polymerization-depolymerization dynamics of cytoskeleton filaments in Sect. "Filament-Driven Motility", and the ones that rely essentially on the activity of motor proteins in Sect. "Motor-Driven Motility". Finally, Sect. "Putting It Together: Active Polymer Solutions" will be devoted to the description of the integrated approaches that have been developed recently to try to understand the cooperative phenomena that underly self-organization of the cell cytoskeleton as a whole.

The Diversity of Cell Motility

Swimming

At the cellular level, viscous hydrodynamic forces are several orders of magnitude higher than inertial forces. Therefore, simple reciprocal motions cannot produce forward motion, and cellular swimming patterns need to be asymmetric in space and time for the cell to advance. This hydrodynamic problem faced by cells attempting to swim have been eloquently summarized by Purcell as "life at low Reynolds number" [228]. To solve this problem, bacteria use the rotation of a short helical or corckscrew-shaped *flagellum*, which is a relatively rigid structure made of a collection of hundreds of identical protein subunits called *flagellins* [20]. The swimming of a single bacterium can be impressively rapid, as bacteria such as *Escherichia coli* - the common intestinal bacterium – swim at speeds of 20 to 30 micrometers per second, for the cell itself is only about two-micrometer long and half a micrometer in diameter. The bacterium possesses multiple flagella that gather together during swimming, and can fly apart as the bacterium switches direction. Other bacteria such as *Vibrio cholerae* - the causative agent of cholera – use a single flagellum located at one of their pole, but the propulsion mechanism relies always on the presence of a rotary molecular motor located in the cell membrane, and which is sensitive to modifications of the chemical environment of the cell. Under normal conditions, the bacterium changes direction in an intermittent chaotic way by reversing the rotational direction of its motors, a phenomenon known as *tumbling*. When placed in a concentration gradient of nutrients however, the cell can adapt its tumbling frequency to swim towards nutrient-rich regions, a phenomenon known as *chemotaxis* [181].

Even though it shares the same name, the eukaryotic flagellum shares little structures and propulsion mechanisms with its bacterial counterpart. It is indeed at least ten times larger than a bacterial flagellum in both diameter and length, and instead of being a rigid passive structure animated by a remote motor, it bears its motor activity along its length. Propulsion occurs by the propagation of a bending wave along the flagellum as a result of the relative sliding of a group of about 10 long parallel filaments, which are engulfed in the cell's plasma membrane and are animated by hundreds of motor proteins in a coordinated manner. Eukaryotic cells also use another type of protrusions to swim, the *cilia*, which are much like flagella in their internal structure, but which are shorter and work usually in numbers, covering sometimes the whole cell surface as in the case of paramecia or other ciliated protozoa (see Fig. 1). Their beating pattern is then coordinated at the cellular level, most often in a wave-type of manner known as the *metachronal wave*. Beating cilia are also found in animals, as for example in humans where ciliated cells play major roles in several organs like the brain, the retina, the respiratory tract, the Fallopian tube and the kidney [117].

Other strategies of swimming include the elegant movement used by *Eutreptiella* - called *metaboly* – which

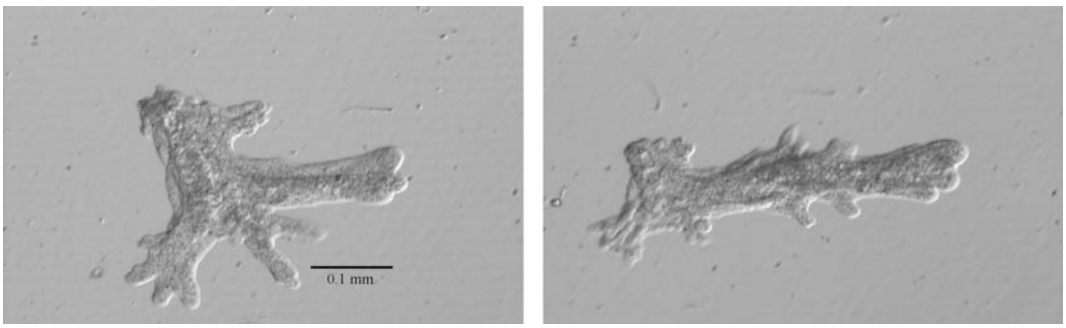

Cytoskeleton and Cell Motility, Figure 3
Two pictures of *amoeba proteus* displaying different shapes of its pseudopodia. Note the dramatic change in cell shape during locomotion. Source: courtesy of Sutherland Maciver

consists in gradually changing the contour of the cell surface to locally increase the drag exerted by the viscous fluid around and move the cell forward [67]. Other organisms like most motile species of *Chrysophytes* – a group of marine photosynthetic protozoa – possess a flagellum attached at their front instead of their back. The flagellum is covered with stiff hairs projecting from its side that allow the cell to move forward as a planar wave propagates from the base to the tip of the flagellum [29]. Finally, one should mention yet another type of motility – namely *walking* – in which cells use also cilia and flagella animated in a coordinated manner to enable the cell to literally "walk" over surfaces. Walking motility relies on the same essential biochemical structures as the ones employed in swimming with collections of cilia.

Crawling

Cell crawling is the common mechanism employed by most eukaryotic animal cells as they move through animal tissues, constituted of other cells or filaments of the extracellular matrix [5,29]. In contrast to swimming cells, crawling cells in general do not employ conspicuous motile organelles that are external to the cell, and which can be studied in isolation. In general however, they either move by means of wormlike cycles of extensions and contractions of their cell body or of some specific protrusions, or slide without visible means of protrusion, a process also referred to as *gliding*. Most of crawling mechanisms rely on the protrusion of specific dynamic extensions at the leading edge of the cell, but gliding seems to rely sometimes on different mechanisms [67,96]. Even though gliding mechanisms are widespread in bacteria, algae and parasitic protozoa, we still do not know for sure the molecular machinery as well as the essential mechanisms that underly these different phenomena [29,96].

The best characterized crawling mechanism is the so-called *amoeboid motility*, referring to the locomotion of all eukaryotic cells that move by means of protrusion of their leading edge. Originally, the term was referring uniquely to the crawling mechanism of *Amoeba proteus*, a particular species of *amoebae*, whose protrusions are stubby three-dimensional projections called *pseudopodia*[2] (see Fig. 3)[3]. But other types of protrusions exist, that are classified with respect to their shape and dimensional organization. Two-dimensional protrusions are the flat veil-shaped projections called *lamellipodia*, as they occur in fibroblasts' or fish epidermal keratocytes' motility for wound healing[4]. One-dimensional projections are the long thin projections called either *filopodia* or *microspikes*, and which occur for example in neuronal-growth-cone progressions[5]. Filopodia usually protrude as small extensions of a lamellipodium, and are used by the cell to extend its lamellipodium in a given direction [136,257] (see Fig. 4). To

[2]Note that some zoologists also use the term "pseudopodia" or "pseudopods" rather generally to refer to a variety of cell-surface protrusions. These include the different types of protrusions described here as playing a role in amoeboid motility, but also the long extended processes that some cell types use only as feeding apparatus, like *axopodia*.

[3]A short video of a locomoting *Amoeba proteus* can be seen on the following website: http://www.bms.ed.ac.uk/research/others/smaciver/A.prot.Loc.mov.

[4]Fibroblasts are the cells that synthesize and maintain the extracellular matrix in most animal connective tissues. They provide a structural framework (stroma) for many tissues, and play a crucial role in wound healing. Keratocytes are epithelial cells that have been characterized in the epidermis of fish and frogs, and that have been named so because of their abundant keratin filaments. They are specialized in wound healing, and are one of the most spectacular example of fast and persistent locomotion in cells, with velocities up to 30 μm/min [5,158].

[5]Growth cones are structures that are found at the tip of axons and dendrites, by means of which neuron cells extend.

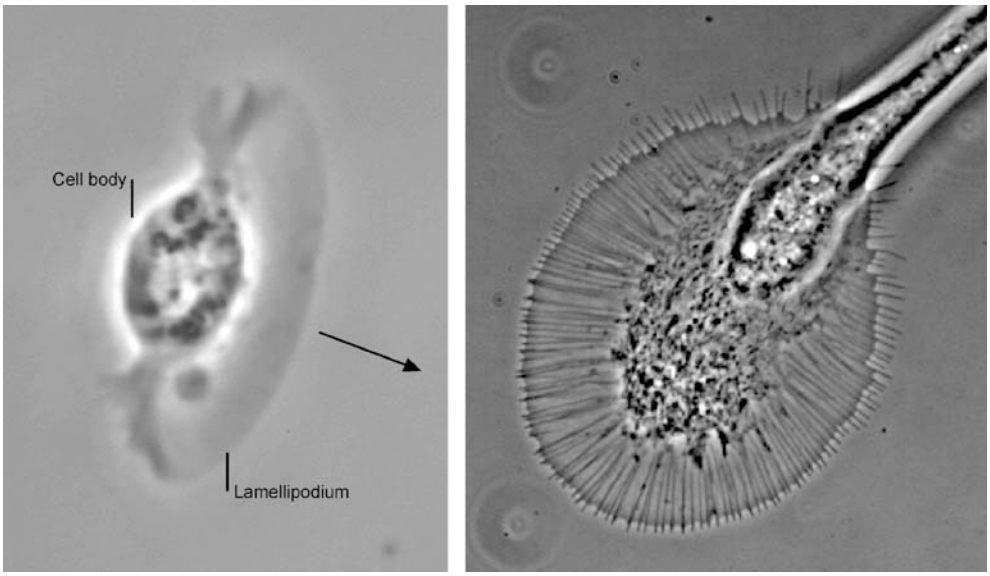

Cytoskeleton and Cell Motility, Figure 4
Two examples of lamellipodia. *Left panel*: A living fish keratocyte extends its leading lamellipodium during crawling. This is a phase-contrast micrograph, a single frame from a video sequence (the whole movie can be seen on the following webpage: http://cmgm.stanford.edu/theriot/movies.htm.). The lamellipodium and the cell body are labeled. The *large arrow* indicates the direction of motion. Source: reprinted from [36] with permission from Nature Publishing Group. Right panel: Snail neuronal growth cone by means of which the nerve fiber elongates at its tip. Clearly visible are the radially-aligned bundle structures that project into filopodia at the leading edge of the lamellipodium. Source: courtesy of Feng-quan Zhou; reprinted from [291] with permission from Rockefeller University Press

these must be added the spherical membrane protrusions called *blebs*, which occur as a result of cortical contractility, and which have been proposed recently to participate in the initiation of lamellipodium formation and the elaboration of cell polarity for directed motion [211], as well as in the amoeboid motility itself for example in *Dictyostelium*, a model species of *amoebae* [289]. Finally, one should mention a motility mechanism that can be classified as *rolling*, in which some organisms such as *helizoa* use coordinated shortening and lengthening of long radiating needlike extensions called *axopodia* to roll over surfaces. Axopodia happen also to be sticky extensions that are most often used for catching preys in numerous protozoa.

Depending on authors, the process of amoeboid motility can be decomposed into three to five steps that occur simultaneously. First the cell makes a protrusion, where the membrane is pushed forward by means of the polymerization of cytoskeletal filaments. Then the protrusion adheres to the substrate via the formation of anchoring points, and subsequent contraction of the cell cytoskeleton drags the cell body forward. Finally at the rear end, the cell de-adheres and retracts [156,192]. Of these five steps, the two last ones – namely de-adhesion and retraction – involve similar structures and mechanisms as the formation of the anchoring points and cell body drag, which led originally Abercrombie to describe his observation as a three-step cycle [1]. The speed of amoeboid motility can range from less than a micrometer per hour to more than one micrometer per second, depending on cell type and stimulation[6].

Extensions of Cell Motility

In addition to moving the whole cell body, the machinery that is responsible for cellular movement can be employed for quite different tasks, which are as essential to the cell survival and reproduction as its motility per se. As we have earlier stated, even in the case of macroscopically non-aminated live forms, the constitutive cells need constantly to displace their internal organelles for their metabolism to be maintained [241]. When looked under the light microscope, mitochondria, vesicles, lysosomes and ingested particles display a rapid and sporadic movement that is interspaced with relatively long periods of quiescence. Velocities are typically of the order of micrometers per second, as the fastest known organelle transport is performed in the

[6]Cell-motility videos can be seen at http://cellix.imba.oeaw.ac.at.

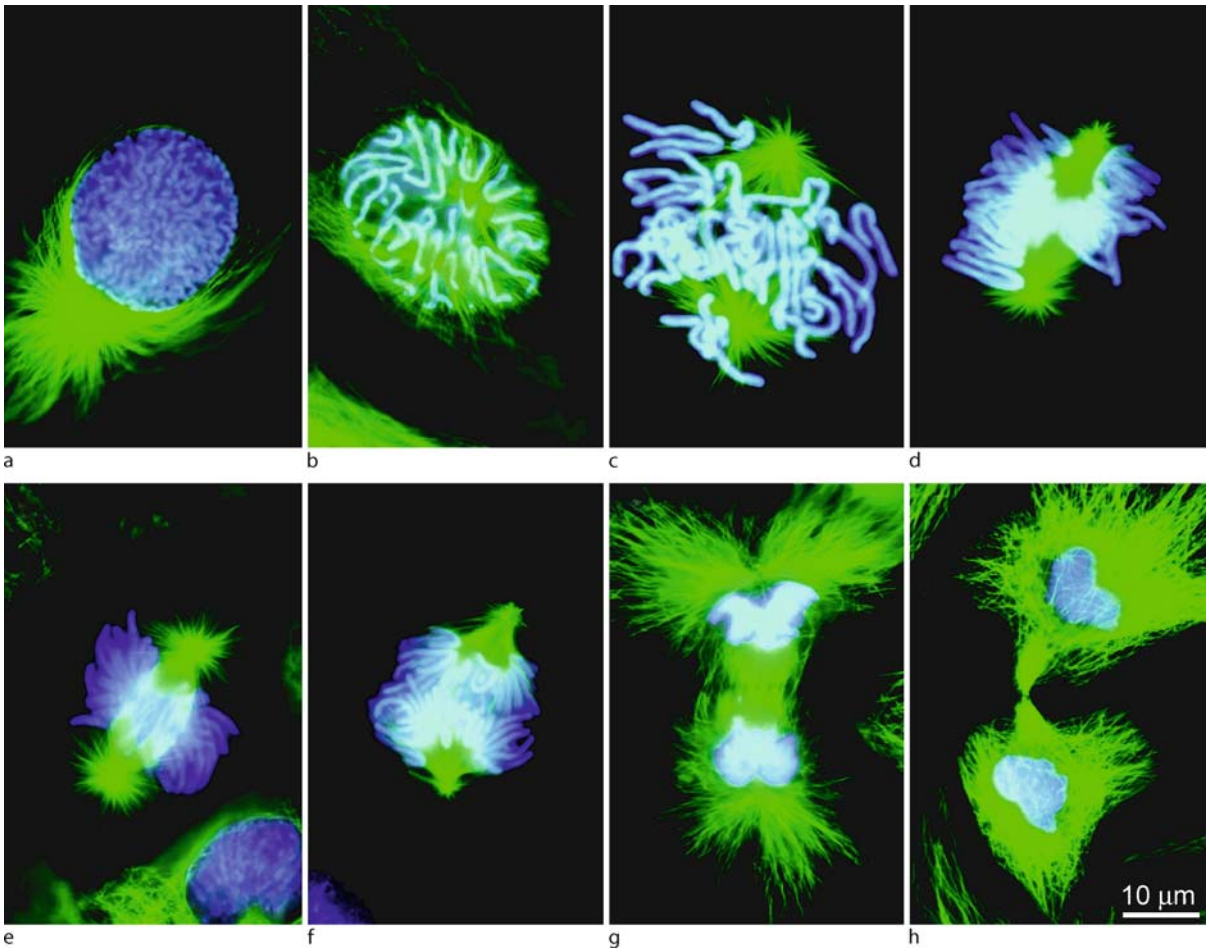

Cytoskeleton and Cell Motility, Figure 5

a to h Fluorescence micrographs of mitosis in fixed newt lung cells stained with antibodies to reveal the microtubules (MT, *green*) see Sect. "Biopolymers", and with a dye (Hoechst 33342) to reveal the chromosomes (*blue*). The spindle forms as the separating astral MT arrays, associated with each centrosome (a to c), interact with the chromosomes. Once the chromosomes are segregated into daughter nuclei (f and g), new MT-based structures known as stem-bodies form between the new nuclei (g). These play a role in cytokinesis (h), the actual cleavage of the two daughter cells. Source: reprinted from [237] with permission from The American Association for the Advancement of Science

green algae *Chara*, whose chloroplasts are transported at velocities that can achieve 60 micrometers per second [39]. Of all cell types, the need for organelle transport is best illustrated by the mammalian motor neurons whose longest extensions – the axons – even though typically only a few micrometers in diameter, can reach lengths up to one meter. Characteristic times that would be required for a mitochondrion to naturally diffuse that distance in such a geometry range from 10 to 100 years. Instead, membrane vesicles and organelles are actively transported in both directions at speeds of about one to five micrometers per second, which allows the whole journey to be made in just a few days [29]. Finally, the probably most-spectacular event of intracellular transport occurs during the essential process of eukaryotic mitosis, by which duplicated chromosomes are segregated from the mother cell and delivered to each of the nascent daughter cells. For this process to occur, major structural reorganizations of the whole-cell cytoskeleton are needed, during which a large and complex cellular structure – the *mitotic spindle* - assembles and drives the chromosomes apart in a coordinated manner [193] (see Fig. 5).

Cell motility can also occur by means of molecular machineries located outside the cell that needs to move. This is the case in particular for mammalian pathogene bacteria such as *Listeria monocytogenes* and *Shigella* and *Rick-*

ettsia species, but also for some viruses like *vaccinia virus*. These organisms propel themselves within and across the cells they invade by utilizing the cytoskeleton of their hosts [41,83,87,266]. Among these organisms, *Listeria* in particular has become a model organism for studying actin-based motility, a simplified version of the whole amoeboid motility, and which can be seen as a representation of just the first step of this complex process in the original Abercrombie description [36,194,222]. Other particular systems use different specific structures from purely cell-cytoskeleton-based motility. Among these, vertebrate skeletal muscles contrast with standard cellular motility, in that they are structured in enormous multinucleated cells that evolved specifically to generate extremely rapid, repetitive and forceful movements. The cytoplasm of these giant cells is crammed full of a highly-organized, almost-crytsalline array of cytoskeletal filaments, whose only function is to produce contractile forces [29,113]. Another essential event of cell division, namely *cytokinesis* – the actual cleavage of the two daughter cells – also involves the contraction of relatively-sliding cytokeletal filaments, this time under the form of a dividing ring [82].

Finally, other types of cells use mechanisms that do not rely on their cytoskeleton for their motility. Of the most spectaculars is the motility based on the stored mechanochemical energy in some supra-molecular springs, which can then contract at velocities as high as eight centimeters per second [179]. Yet another mechanism relies on some stored purely-elastic energy that allows some insect-eating plants to catch their preys. This is the case for example of the Venus flytrap *Dionaea muscipula*, whose leafs can close in about 100 ms, one of the fastest movements in all plant kingdom. To achieve such a performance, the plant relies on a snap-buckling instability, whose onset is actively controlled by the plant after the arrival of a fly has triggered some biochemical response via the disturbance of mechano-sensitive hairs located inside the trap [69].

The Cell Cytoskeleton

The eukaryotic *cytoskeleton* is defined as the system of protein filaments that enable the cell to insure its structural integrity and rigidity, regulate its shape and morphology, exert forces and produce motion. As a framework that insures structural integrity, the cytoskeleton is mainly constituted of a cohesive meshwork of protein filaments that extend throughout the cytoplasm of the cell. But being the essential structure that produces movement at the cellular level, and thereby needing to be highly adaptable to extracellular stimuli or rapid environmental changes, the cytoskeleton has evolved into a highly-dynamic structure. In fact, cytoskeletal filaments constantly grow and shrink, associate and dissociate via multiple linkages, organize on large scales into a dynamic network, and serve as an intricated set of tracks to motor proteins that transport cargos from one part of the cell to the other, or slide filaments with respect to one another to produce contractile forces. This section is devoted to the biochemical description of this very-complex structure. In addition, its interaction with the cell's external world and its regulatory pathways will be briefly presented, as well the prokaryotic cytoskeleton which, even though biochemically different, appears more and more to resemble its eukaryotic counterpart on a functional point of view.

Biopolymers

How can a eukaryotic cell, with a diameter of 10 microns or more, be spatially organized by cytoskeletal proteins that are typically 2000 times smaller in linear dimensions? The answer lies in *polymerization*, this ability of the elementary protein subunits (called *monomers*) to assemble via physical interactions into extended linear structures that are made of a large number of them, typically thousands (called thereby *polymers*, or here more precisely *biopolymers*). There are three types of biopolymers in a given eukaryotic cell, namely *actin filaments*, *microtubules* and *intermediate filaments* (see Fig. 6). Although they are classified according to their respective thickness, more interesting for cellular structures and functions are their rigidity, which at thermodynamic equilibrium is characterized by their persistence length L_p [7] [81].

Actin filaments – or F-actin – have a persistence length that is usually accepted to be of the order of 15 to 17 μm [81,210], even though it has been reported that actin rigidity should depend on the way it is decorated, ranging from 9 ± 0.5 μm for bare F-actin to 20 ± 1 μm for tropomyosin-bound actin filaments in skeletal-muscle structures [119]. Actin filaments are two-stranded helical polymers, 5 to 9 nm in diameter, and are built from dimer pairs of globular-actin monomers – or G-actin – that are polar in nature [5,29]. The two halves of an actin monomer are separated by a cleft that can bind adenosine triphos-

[7] The persistence length L_p is defined as follows: consider a thin flexible rod of fixed length L, submitted to thermal forces. Its shape is completely specified by the tangent angle $\theta(s)$ in three dimensions along the arc length of the rod [154]. The persistence length L_p is defined as the characteristic arc length above which thermal fluctuations of the angle $\theta(s)$ become uncorrelated. Specifically, $\langle \cos[\Delta\theta(s)] \rangle = \exp(-s/L_p)$, where $\Delta\theta(s)$ is the three-dimensional angle change over the arc length s. L_p is related to the rod's material Young modulus E and its geometrical moment of inertia I by $L_p = EI/k_B T$, where $k_B T$ represents thermal energy [81].

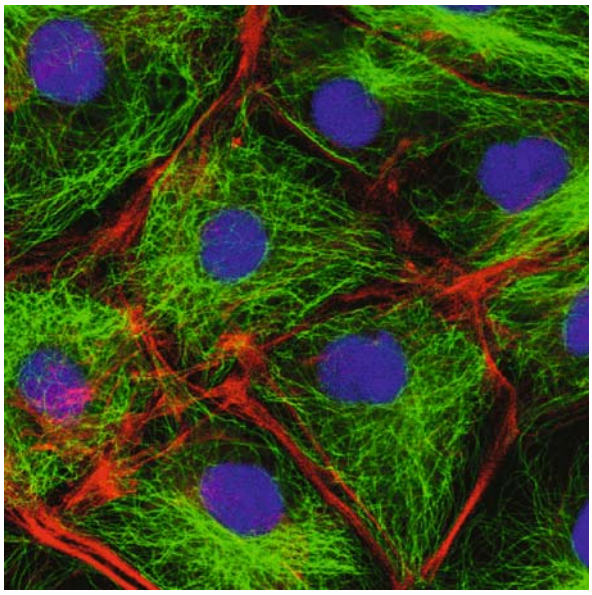

Cytoskeleton and Cell Motility, Figure 6
Animal cells as seen in the fluorescence microscope after fixation and labeling with specific probes. Actin filaments are stained in *red*, microtubules in *green*, and the nuclei in *blue*. Source: courtesy of Mark Shipman, James Blyth and Louise Cramer, MRC-Laboratory Molecular Cell Biology and Cell Biology Unit, UCL, London UK (unpublished)

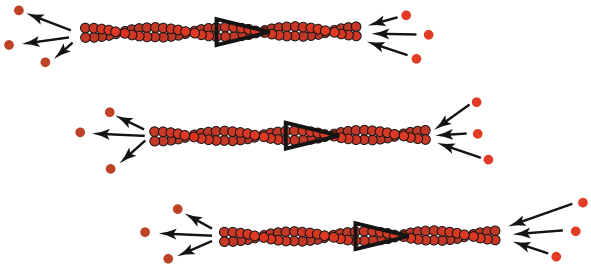

Cytoskeleton and Cell Motility, Figure 7
Schematic representation of a treadmilling actin filament. The arrows indicate the polarity of the filaments. Monomers are added to the plus end and removed from the minus end at the same rate, such that while the filament's length remains constant, its center of mass is advancing. Top to bottom shows three subsequent times. Source: courtesy of Karsten Kruse

phate (ATP) or its hydrolyzed form adenosine diphosphate (ADP)[8]. This is responsible for the existence of two distinct ends to the whole filament, namely a fast growing end – called "plus end" or "barbed end" – where mostly ATP-bound monomers are located, and a slow growing end – known as "minus end" or "pointed end" – that is rich in ADP-bound monomers. The minus end has a critical actin-monomer concentration that is roughly six times as high as that of the plus end. At steady state, and with the help of monomeric diffusion, this drives the phenomenon of *treadmilling*, a dynamic evolution of the actin filament where actin monomers are added to the plus end, and removed from the minus end at the same rate. During this process, the total length of the treadmilling filament is kept constant, while its center of mass is displaced at a constant velocity, even though each individual polymerized monomer do not move on average[9] (see Fig. 7). For pure actin at physiological concentrations, this process is rather slow and occurs at velocities of the order of a few micrometers per hour. But as we shall see in the following, specialized actin-binding proteins allow the cell to increase this speed substantially, which makes actin forces exerted via polymerization-depolymerization mechanisms one of the key players in cellular motility.

Actin is the most abundant protein in a eukaryotic cell (several grams per liter), and has been highly conserved throughout evolution. It organizes into a variety of structures, namely linear bundles, two-dimensional networks or three-dimensional gels, and is mainly concentrated in a layer located just beneath the plasma membrane and called the *actin cortex*. Of primary importance for cell motility are the two-dimensional highly cross-linked networks that actin forms in lamellipodia, and the linear bundles that are found in filopodia and which protrude from the lamellipodia in a directed way. There, as we shall see in Sect. "Filament-Driven Motility", actin polymerization plays a crucial role in driving cell motility. Finally, one should mention the cortical rings that contract during the process of cytokinesis to cleave the two daughter cells, as well as the formation of *stress fibers*, which are force-producing structures that are attached to anchoring points, and which enable the cell to exert traction forces on the substrate on which the cell is crawling[10] (see Sect. "Cell Anchoring and Regulatory Pathways").

Microtubules are the stiffest of all polymers, with persistence lengths ranging from 100 µm up to 6 mm [213]. They are hollow cylinders with an outer diameter of $\simeq$25 nm and are made of tubulin subunits arranged in 13 adjacent protofilaments. Tubulin is a heterodimer formed of α- and β-subunits, which can bind either

[8]The hydrolysis reaction of ATP (ATP $\rightleftharpoons$ ADP + P_i, where P_i designates inorganic phosphate) breaks a high-energy chemical bond – here a phosphoanhydride bond – to drive many chemical reactions in the cell.

[9]Animated movies of this process can be seen at http://www.uni-leipzig.de/~pwm/kas/actin/actin.html or http://cellix.imba.oeaw.ac.at/actin-polymerisation-drives-protrusion.

[10]Illustrations of these structures can be found at http://cellix.imba.oeaw.ac.at. See also [136,257].

guanosine triphosphate (GTP) or guanosine diphosphate (GDP)[11]. Microtubules share some important properties with actin filaments, in that they are polar, treadmill, and can exert forces [51]. They typically organize radially from a single microtubule-organizing center called the *centrosome*, and connect to the actin cortex with their plus ends towards the cell edge. In addition to giving the cell its structural rigidity and shape, they actively participate in regulating the actin cortex dynamics, focal-adhesion assembly and disassembly, and in some cell types participate in determining the cell polarity and its subsequent migrating direction (see Sect. "Cell Anchoring and Regulatory Pathways").

Intermediate filaments are the most flexible polymers of the cell cytoskeleton, with persistence lengths of the order of 0.3 to 1.0 μm. They range in diameter from 7 to 12 nm, in-between that of actin and microtubules. There are different classes of intermediate filaments such as *vimentin*, *desmin*, *keratin*, *lamin* and *neurofilaments*, and they constitute together a large and heterogeneous family, of which different cells possess different members. Unlike actin filaments and microtubules, they are not polar, do not treadmill, and are therefore thought to contribute essentially to the structural and elastic properties of the cell, but little to its dynamics and motility. One particular example of intermediate-filament structure is the *nuclear lamina*, located just beneath the inner nuclear membrane, and that is responsible for its structural integrity.

Molecular Motors

Molecular motors constitute the subset of proteins and macromolecular complexes that convert a given source of energy into mechanical work [5,103]. The energy they need is generally stored into either of two forms by the cell: high-energy chemical bonds, such as the phosphoanhydride bonds found in ATP and GTP, and asymmetric ion gradients across membranes. Known molecular motors can be classified into roughly five categories, namely (1) rotary motors, (2) linear-stepper motors, (3) assembly-disassembly motors, (4) extrusion nozzles, and (5) prestressed springs. A nice table of the major different cell-movements' categories with the different cellular structures and molecular motors they rely on, can be found in [67].

All known biological rotary motors use ion-gradient-based sources of energy, and most of them use electrochemical forces based on hydrogen-ion (or proton) gradients, also known as *proton-motive forces*. This is the case for example for the propulsion motor of bacteria that is responsible for their flagella to rotate [20,181], as well as for the surprising rotary motor F0F1-ATPase that is responsible for ATP synthase in mitochondria and bacteria [290]. This rotary machine usually converts the electrochemical energy stored in proton-concentration gradients, first into mechanical motion, and then back into chemical energy under the form of ATP. But the motor is also reversible, in that it can harness the chemical energy of ATP to produce or maintain the transmembrane electrochemical gradient of proton concentration. This reversibility is best seen in bacteria, when they switch from aerobic to anaerobic conditions [5].

Most of the motors used in amoeboid motility are linear-stepper motors [103,246,258]. We shall therefore focus on this class of motor proteins in the remaining of the present article. They walk on the linear tracks formed by the polymerized cytoskeletal filaments, and can be classified into two different categories, namely *processive* and *non-processive* molecular motors, sometimes designated as "porters" and "rowers" [159]. The processivity is linked to the *duty ratio*, the proportion of time that the molecule spends attached to the filament as compared to the whole motor cycle, namely one ATP-hydrolysis cycle [102] (see Fig. 8). Typically, porters are individual walkers that carry cargos across the cell, and therefore most often participate in intra-cellular traffic. Rowers however work in numbers, and are usually involved in generating contractile forces, like it is the case in skeletal-muscle fibers, stress fibers or contractile rings that form during cytokinesis [82,109,113,115]. Structurally, all these motor proteins can be divided into a motor domain, called the *head*, and a *tail* or *base*. The head is the site of conformational change of the protein during ATP-hydrolysis, and with which the motor attaches to the filament. The tail connects the motor to its cargo or to other motors. Processive motors are (homo-)dimers, such that as one head is attached to the filament, the other can move to a new binding site. In that case, the two tails of the associated monomers wind up together to hold to each other. Non-processive motors can also be found in dimeric forms, one of the two heads being then just unused.

Motor Families

Eukaryotic cytoskeletal motor proteins are divided into three superfamilies, namely *myosins*, *kinesins* and *dyneins*. The motor proteins known longest belong to the myosin superfamily [21], because of their high concentration in skeletal muscles. All myosin motors walk on actin filaments through a general four-step process: binding,

[11] Similarly to ATP, GTP is a stored source of energy for the cell that is consumed via a hydrolysis reaction, here $GTP \rightleftharpoons GDP + P_i$.

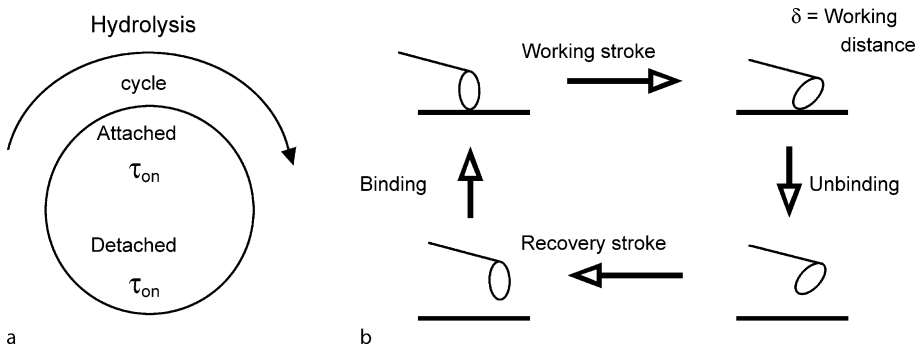

Cytoskeleton and Cell Motility, Figure 8

a ATP-hydrolysis cycle, with the respective durations τ_{on} and τ_{off} of the attached and detached states of the motor. These durations define the duty ratio r as $r = \tau_{on}/(\tau_{on} + \tau_{off})$. b During the attached phase, the head of the motor makes a working stroke of working distance δ. The motor then unbinds from the filament, and makes a recovery stroke during the detached phase. By recovering its initial conformation while detached, the motor avoids stepping backwards and so progresses by a distance equal to the working stroke during each cycle. Source: reprinted from [102] with permission from Nature Publishing Group

power-stroke, unbinding, and recovery-stroke[12] [102] (see Fig. 8). Today, they are classified into 18 different classes, with possibly dozens of different members in each class, even in a single organism. The skeletal-muscle myosins belong to the Myosin II family; they have long tails that form dimeric α-helices and associate into the so-called "thick filaments" originally observed by H. E. Huxley and J. Hanson, while the "thin filaments" are F-actin polymers [93,110,111] (see Fig. 2 and 9). Most myosin molecules are plus-ended directed (to the exception of Myosin VI), and non-processive (to the exception of Myosin V, which is involve in vesicular transport). Their very diverse mechanical features, in terms of step sizes, duty ratios and stepping speeds, are very fine-tuned to their functions[13] [102].

Kinesin proteins share very similar structural features with myosins in their head domain and are therefore thought to have branched from a common ancestor with myosins, but diverge in their tail structures [273]. They walk on microtubules instead of actin filaments, are processive, and are involved mainly in intracellular transport like the transport of organelles along nerve axons[14]. The kinesin superfamily has been divided into 14 families, and a number of "orphans" that are so far ungrouped [189]. Most kinesin motors are plus-ended directed, like the conventional kinesin I that founded the family [274]. Members of the Kinesin-13 family are unconventional, in that they can processively induce microtubule depolymerization, a process that is essential to chromosome segregation during mitosis[15] [107] (see next Section).

Dynein proteins are less well-characterized. It is also unknown whether they share a common ancestor with myosins and kinesins, or whether they are the result of convergent evolution. Two major groups of dyneins exist: axonemal dyneins, which drive the bending of eukaryotic cilia and flagella by inducing the relative sliding of microtubules [226], and cytoplasmic dyneins, which are involved in organelle and vesicular transport, as well as cell division [133]. Most dyneins are minus-ended directed, and interestingly, some dyneins can be non-processive at high, but processive at low ATP concentrations.

Other Cytoskeleton-Associated Proteins

The coordination of the numerous different processes that happen during amoeboid motility rely on a tight regulation of the activity of the cell cytoskeleton, as well as its anchoring to the substrate. In particular, as we shall see below, the protrusion of the leading edge of the cell – the first step of amoeboid motility – relies on the formation of a highly-cross-linked and dynamic network of actin filaments. Its formation and dynamical regulation are carried out with the help of numerous accessory proteins [224,257]. Following Pollard's presentation [223,225] (see Fig. 10), we can focus on the main proteins that are involved in the formation, structure and dynamics of

[12] Animated movies of myosin skeletal fibers' detailed motion can be seen at http://www.scripps.edu/cb/milligan/research/movies/myosin.mov, or http://valelab.ucsf.edu.

[13] Up-to-date information about myosin motors can be found at http://www.proweb.org.

[14] Animated movie of kinesin's detailed motion can be seen at http://www.scripps.edu/cb/milligan/research/movies/kinesin.mov or http://valelab.ucsf.edu.

[15] Up-to-date information about kinesin motors can be found at http://www.proweb.org.

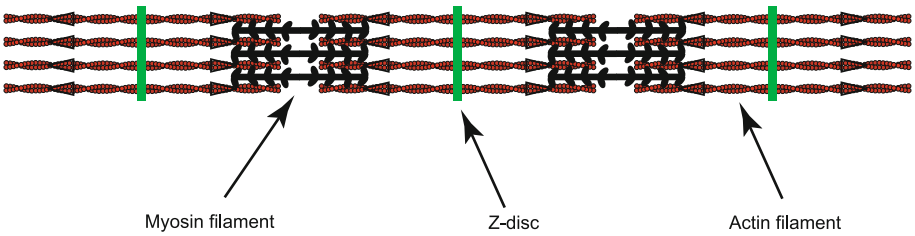

Cytoskeleton and Cell Motility, Figure 9
Schematic representation of muscle myofibrils, the basic contractile fibers of skeletal muscles. Actin and myosin filaments are periodically arranged in a polarity-alternated fashion. Between two "Z discs" is found the elementary structure that is periodically repeated, the *sarcomere*, and where relative sliding of actin and myosin filaments leads to contraction. Source: courtesy of Karsten Kruse

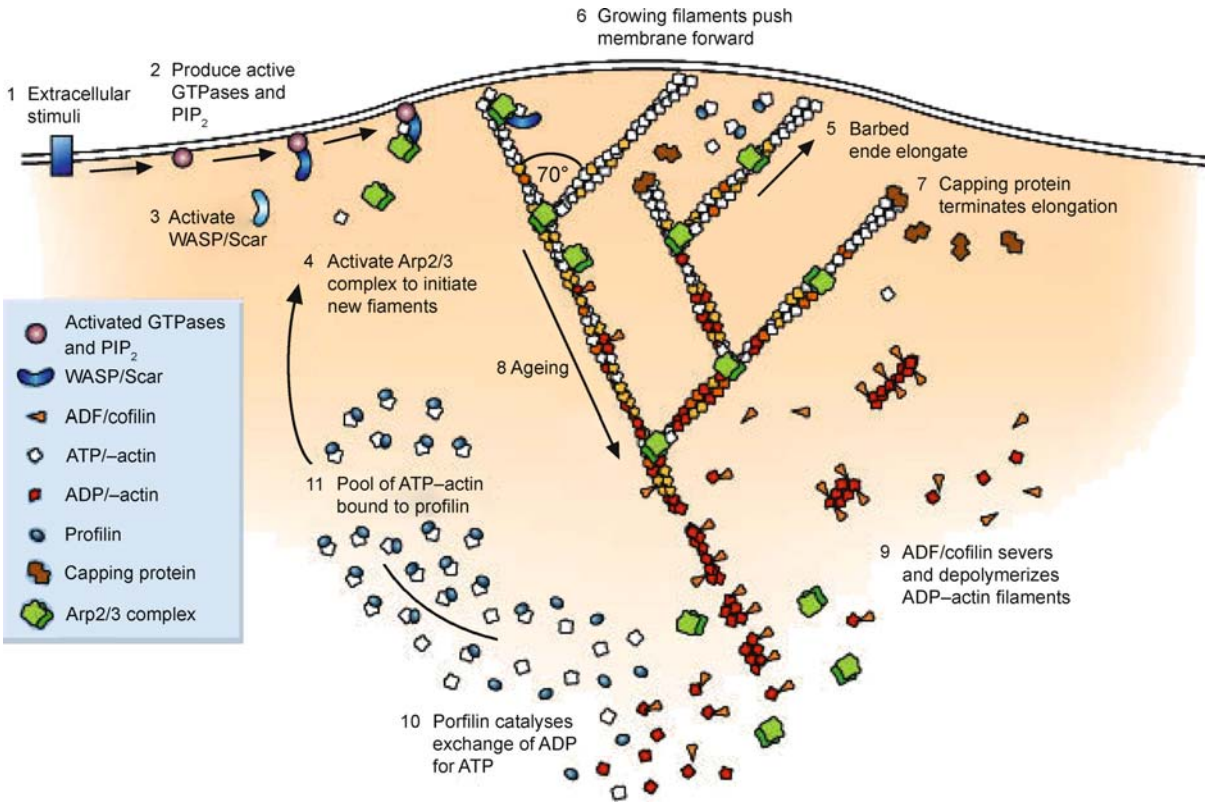

Cytoskeleton and Cell Motility, Figure 10
Dynamical organization of the actin network at the leading edge of a protruding lamellipodium. (1) External cues activate signalling pathways that lead to GTPases and PIP_2 activation (2). These then activate proteins of the WASP family (3), which in turn activate Arp2/3 complexes that initiate new filaments as branching from existing ones (4). Each new filament grows rapidly (5), fed by a high concentration of profilin-bound actin stored in the cytoplasm, and this pushes the plasma membrane forward (6). Capping proteins bind to the growing ends, terminating elongation (7). Actin-depolymerizing factor (ADF)/cofilin sever and depolymerize the ADP filaments, mainly in the "older regions" of the filaments (8, 9). Profilin re-enters the cycle at this point, promoting dissociation of ADP and binding of ATP to dissociated subunits (10). ATPactin binds to profilin, refilling the pool of subunits available for assembly (11). Source: reprinted from [223] with permission from Nature Publishing Group (image based on an original figure of [224])

the actin network. Nucleation of the network starts after biochemical signals have been integrated via G-protein-linked membrane receptors, namely small GTPases and PIP_2 pathways [174] (see Sect. "Cell Anchoring and Regulatory Pathways"). Then members of the Wiscott Aldrich syndrome protein (WASP) family that are anchored to the cell's plasma membrane (like Scar [175]), activate Arp2/3 (for actin-related proteins 2 and 3) complexes that are re-

sponsible for the nucleation and maintenance of branching points in the network[16]. Then, in order to promote growth of the actin gel, recycling of G-actin monomers as well as the creation of new F-actin plus ends are stimulated by mainly two types of proteins: (1) Actin-binding proteins – such as profilin – that bind to actin monomers, catalyze the exchange of ADP for ATP, and inhibits ATP hydrolysis, a process that is antagonized by monomer-sequestering proteins – like thymosineβ-4 – that stabilize ADP-bound G-actin. (2) Actin-depolymerizing factors (ADF) – such as cofilin (or ADF/cofilin) – that sever and depolymerize ADP-actin filaments, thereby increasing the pool of available G-actin monomers. The structure of the network is further controlled by capping proteins that can bind to F-actin plus ends to terminate their growth, and thus limit the increase of free-growing plus ends. Finally, cross-linked structures are formed with the help of actin cross-linkers like filamin, and actin-bundling proteins like fascin, fimbrin and α-actinin. α-actinin and filamin are most present in lamellipodium structures, as fimbrin and fascin and most observed in filopodia [257]. Finally, the same as well as other actin-binding proteins (like espin, fascin, fimbrin and villin) exist in other structures where actin-bundles are formed, like bristles, microvilli and stereocilia [234]. In skeletal muscles, tropomyosin strengthens the actin filaments and prevents myosin motors from binding to actin when muscles need to be at rest[17]. For a relatively recent review on the actin-binding proteins, see [53].

Microtubule-associated proteins (MAPs) have been classified into two types, and participate to microtubules' stability and organization. MAPs of Type I are large filamentous proteins that comprise a microtubule-binding domain and a projection domain, thereby controlling the spacing of microtubules. MAPs of Type II have similar structures and cross-link microtubules to membranes, intermediate filaments or other microtubules. In addition, both types of MAPs promote microtubule assembly and stability, and compete with motor proteins for binding sites, such that they participate in microtubule-transport regulation. Other MAPs that do not belong to these classes are denoted XMAPs, as they have been originally identified in the Xenopus-frog eggs. Among these are the plus-end-binding proteins (or +TIPs) that bind to the microtubule growing ends and participate in their stability, and the highly-conserved stathmin or oncoprotein 18 which, instead, destabilizes microtubules [10,249]. The best understood microtubule end-binding proteins are the MCAKs (for mitotic centromere-associated kinesins), also known as Kin I kinesins, which are unusual kinesins in that, instead of moving along the surface of microtubules like other kinesin proteins do, they bind to microtubules' ends and trigger depolymerization in a processive way [47]. In particular, they depolymerize microtubules during mitosis to drive chromosome segregation [180]. For a review, see [104].

Cell Anchoring and Regulatory Pathways

The two first steps of cell crawling in the Abercrombie classification consist in the protrusion of the leading edge and its adhesion to the substrate [1,156]. Although they were thought to be largely independent processes, evidences are accumulating that adhesion and protrusion are highly interrelated [9,24,90,236]. Protrusion results primarily from actin polymerization at the leading edge of the migrating cell (see Sect. "Filament-Driven Motility"), and is regulated by the small GTPases Rho, Rac and Cdc42 [25,235]. As Rho is known to activate actomyosin contractility, Rac and Cdc42 induce actin polymerization and the formation of actin-filled protrusions such as lamellipodia and filopodia [91]. Through these pathways, the cell can respond to the chemical composition of its environment, an example of chemotaxis: as a function of the gradients of chemoattractants or chemorepellants in its environment, the cell regulates its sites of fastest actin polymerization in order to move towards or away from the source[18] [215]. Of primary importance for the accurate spatial regulation of these processes is the role of microtubules, which in some cell types play a crucial role in determining cell polarity and directional migration [136,204,236,255,256]. Microtubules have been proposed to activate Rac and Rho, the latest via the release of the GDP-GTP exchange factor GEF-H1 during microtubule depolymerization [143,283].

Adhesion occurs via the formation of adhesion sites, which rely primarily on molecules such as *integrins*, also involved in regulating the cell behavior via different signal-transduction pathways [24,78,190]. An important aspect of that process is that it allows the cell to "feel" the mechanical properties of its environment. This has been shown to be important for the migration of fibroblasts, in that they seem to migrate preferentially towards regions of

[16] In addition to Arp2 and Arp3, which are members of the Actin related proteins (Arp) family in that they have sequences and structures that are similar to actin, the Arp2/3 complex contains five other smaller proteins.

[17] Up-to-date informations can be found at http://www.bms.ed.ac.uk/research/others/smaciver/Cyto-Topics/actinpage.htm.

[18] See the movies associated with ref [215], as well as the one of a neutrophil cell that chases a bacterium at http://www.biochemweb.org/fenteany/research/cell_migration/movement_movies.html.

stiff substrates, a process referred to as *durotaxis* [169]. In addition, the mechanical properties of the cell's environment have been proposed to be relevant for tissue-growth directionality, as well as cell differentiation [56,243]. Adhesion sites can be roughly divided into two broad categories, namely *focal complexes*, which locate beneath microspikes or filopodia, and *focal adhesions*, which locate at the termini of stress fibers and serve in long-term anchorage [136] (See Fig. 11). Interestingly, these adhesion sites are also regulated by small GTPases of the Rho, Rac and Cdc42 families: focal complexes are signaled via Rac1 and Cdc42, and can either turnover on a minute time-scale or differentiate into long-lived focal adhesions via the intervention of RhoA [136]. A schematic representation of the integrated role of the small GTPases in regulating cell migration can be seen in Fig. 12. Such regulations are crucial

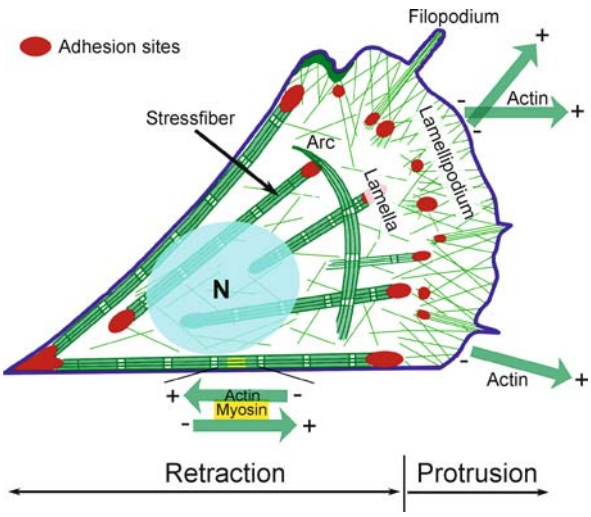

Cytoskeleton and Cell Motility, Figure 11
Schematic representation of the actin cytoskeleton in a polarised fibroblast. The different organisational forms of actin filaments and their relations to adhesion sites to the substrate are depicted: diagonal actin filament meshwork in the lamellipodium, with associated radial bundles that sometimes protrude into filopodia; contractile bundles of actin (stress fibers) in the cell body and at the cell edge; and a loose actin network throughout the cell. Arc-shaped bundles are sometimes observed that move inwards under the dorsal cell surface (arc). The diagram shows an idealized cell: in reality, actin arrays are interconnected in various combinations and geometries. Adhesion sites are indicated in red. The flat region behind the lamellipodium and in front of the nucleus (N) is termed the lamella. At the cell front, in lamellipodia and filopodia, actin filaments are all polarized in one direction, with their fast-growing ends directed forward for producing pushing forces and inducing protrusion; in the cell body, actin filaments form bipolar assemblies with myosin proteins (stress fibers) for retraction. Source: courtesy of Vic Small; modified from [136] with permission from Elsevier Limited

for cell migration to occur optimally. Indeed, whereas adhesion sites are necessary at the leading edge of the cell to provide anchoring points on which the cell can exert traction forces, these need to be released at the rear for the cell to move forward. This results in a biphasic response of the cell-migration speed as a function of adhesive-ligand concentrations, in that too-low or too-high ligand concentrations prevent either the traction forces to be exerted, or the rear to be released [131,156]. How these regulatory pathways lead to a spatio-temporal feedback mechanism between actomyosin regulation and the focal-adhesion system is still under investigation [24,90].

The Prokaryotic Cytoskeleton

This section is completely independent of the rest of the article. Readers not interested in the biochemical composition of the prokaryotic cytoskeleton might want to skip this section.

As cytoskeletal proteins' structures are highly conserved throughout the three domains of life (archaea, bacteria and eukarya), prokaryotic cytoskeletal proteins differ strongly in their sequences from their eukaryotic counterparts. For this reason, and the fact that prokaryotes have a relatively simple organization as compared to eukaryotic cells, it was long thought that they were lacking a cytoskeleton. It is only in the 1990s that prokaryotic homologs of tubulin, actin and intermediate filaments started to be discovered. The first bacterial cytoskeletal proteins to be brought to knowledge was the protein FtsZ, whose relation to tubulin was discovered independently by three groups in 1992 [43,200,232]. Later, it was found that FtsZ could assemble into protofilaments that can be either straight or curved as a function on the state of the nucleotides, similarly to microtubules [58,104], and that its structure at the level of protein folding was nearly identical to that of tubulin [171,206]. The second prokaryotic cytoskeletal proteins to be discovered were MreB and ParM also in 1992, and were shown to be distant relative of the actin superfamily by sophisticated sequence-alignment techniques [26]. Later, it is only in 2001 that MreB was proven to be capable of self-assembly into cytoskeletal filaments that resemble much closely F-actin structures [126,276]. Finally, an homolog of intermediate filaments has been found recently in the bacterium *Caulobacter crescentus*, but only in this particular species so far [16]. Since it is responsible for giving the bacterium its crescent shape, it was given the name of *crescentin*.

Despite their sequencial differences with their eukaryotic counterparts, prokaryotic cytoskeletal proteins share with them strong homologies in their structural as well as

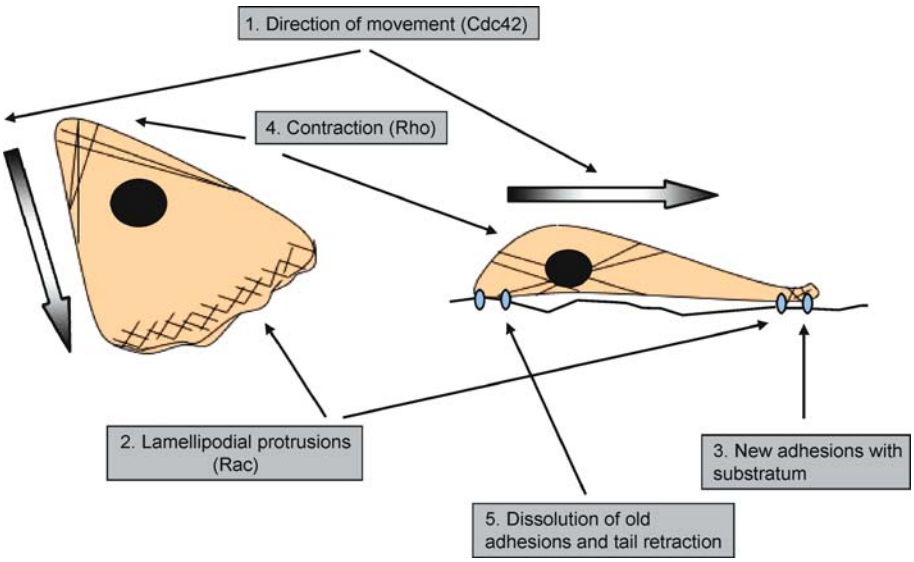

Cytoskeleton and Cell Motility, Figure 12
Schematic representation of the integrated roles of Rho, Rac and Cdc42 proteins in regulating cell migration. By inducing actin-filament assembly, filopodia and focal-complexes formations, Cdc42 regulates the direction of migration (1). Rac induces actin polymerization at the cell periphery (2) and promotes lamellipodia protrusion. It also induces the formation of focal complexes at the leading edge (3). Rho plays a role in regulating longer-lived structures, namely activating actomyosin contraction in the stress fibers located in the cell body and at the rear (4), as well as promoting the assembly of focal-adhesion complexes. Source: courtesy of Alan Hall; reprinted from [229] with permission from Elsevier Limited

functional properties [8,188]. They are classified into four groups [253]: (i) Actin homologs are constituted by MreB and MreB homologs, ParM, and MamK. MreBs play an important role in a number of cellular functions, such as regulation of cell shape, chromosome segregation, establishment of cell polarity and organization of membranous organelles. ParM proteins are involved essentially in plasmid partitioning, and MamK is involved in the subcellular organization of membrane-bounded organelles. Similarly to actin, MreB and ParM protein families present polymerization-depolymerization dynamics that are driven by ATP hydrolysis. Less is known about MamK. (ii) Tubulin homologs contain FtsZ and the BtubA/B proteins, which constitute two other families of GTPases as compared to tubulin. As FtsZ is crucially involved in cytokinesis via its ability to form contractile rings and spiral structures, the role of BtubA/B proteins, which are much less widespread in the bacterial kingdom, has less been characterized so far. (iii) The intermediate filaments' homolog crescentin has only been found in *Caulobacter crescentus* and, as for its eukaryotic counterparts, is mainly involved in cell shape and structural integrity. (iv) Finally, the large MinD/ParA superfamily is made of prokaryotic cytoskeletal proteins that have no counterparts in eukaryotes. They however have the ability to organize into polymeric filaments, and present ATPase activity. Proteins of the MinD group are involved in placement of the bacterial and plasmid division sites, whereas proteins of the ParA subgroup are primarily involved in DNA partitioning.

Interestingly, cytoskeletal proteins seem to have been strongly conserved throughout evolution in each of the three separate domains of life, but differ quite substantially across domains [57]. Bacterial FtsZs proteins are 40–50% identical in sequence across species, and share even the same amount of similarities with their archaeal counterparts. Bacterial MreBs are generally 40% conserved. Among eukaryotes, the conservation is even stronger: it reaches 75–85% for tubulin and 88% for actin, one of the most conserved protein in the eukaryotic domain. In the case of archaea, MreB and actin homologs have not yet been identified for sure [57].

Filament-Driven Motility

Many kind of movements in eukaryotic cells are driven by polymerization-depolymerization mechanisms of cytoskeleton filaments, for which motor proteins per se are not required. Instead, the chemical energy stored in high-energy hydrogen bounds (under the form of ATP or GTP) is converted into movement via treadmilling mechanisms [264]. Two types of filaments have this ability, namely microtubules and actin filaments.

Microtubule Growth and Catastrophes

Microtubules are the stiffest of all cytoskeletal filaments, which confers them the ability to organize and stabilize both the cell structure and its transport network for internal communication and distribution. Depending on the cell need, they constantly reorganize or exert forces on the cell membrane or other organelles they transport. This is the case for example during the organization of the mitotic spindle, the structure formed prior to chromosomal segregation during mitosis [248]. There, chromosomes gather in a plane halfway from two microtubule-organizing centers – the centrosomes – that are located at each pole of the future dividing cell, and from which the microtubules tear the chromosome pairs apart [6,118] (see Fig. 6, especially panels C to F). For this mechanism to happen, microtubules constantly exert pulling and pushing forces both on the chromosomes and the cell membrane, a mechanism that allows for correct positioning of the site of cell-division [288]. Coupled with kinesin-motor activity, correct positioning of the centrosomes relies crucially on the ability of microtubules to grow and shrink spontaneously, a dynamics that provides feedback to centrosome positioning and leads to oscillations orthogonal to the cell spindle axis [85,86]. Such a mechanism is also responsible for the correct positioning of the nucleus in the fission yeast *Schizosaccharomyces pompe* [268].

Understanding microtubules' polymerization-depolymerization dynamics started with the first observation that they display phases of relatively slow growth, alternated with phases of rapid shrinkage [191]. Changes from one type of behavior to the other are referred to as *catastrophes* for the conversion from growing to shrinking, and *rescues* for the opposite transition. Observations of this behavior were further made in culture cells [244] and cellular extracts [19], which confirmed the existence of such dynamic instabilities in vivo. During mitotic-spindle formation, it has later been shown that the specialized structures that connect the microtubules to the chromosomes, the *kinetochores*, can "capture" and stabilize growing microtubules, preventing them from undergoing catastrophes [95]. For a review, see [134].

Further characterization of microtubules' biomechanical properties came from experimental studies of the forces produced by their polymerization-based growth. Analyzing force-induced microtubule buckling [50], microtubule forces were characterized as being potentially as high as those produced by motor proteins – typically a few pico-Newtons [123] – and to be able to deform membranes [74] or center asters in mirofabricated chambers [62,101], a mechanism that imitates nucleus positioning in fission yeast. For reviews, see [51,104].

Actin Gels

As earlier stated, the first step of amoeboid motility in the original Abercrombie classification occurs via protrusion of the leading edge of the cell. This mechanism relies mainly on the polymerization dynamics of actin filaments [194,214,225]. Actin polymerization is known to play a primary role at the plasma membrane, where it is nucleated by proteins of the WASP family via Arp2/3 complexes (see Sect. "Other Cytoskeleton-Associated Proteins"). It has also been proposed to be responsible for driving endocytosis and the movement of endosomes, both in cultured cells and yeast [132,186].

Our understanding of eukaryotic actin-based motility has grandly benefitted from the motility mechanism of the bacterium *Listeria monocytogenes*. This pathogene moves at velocities of the order of several micro-meters per minute by nucleating the formation of an actin "comet-tail" that, while polymerizing thanks the host's cytoskeletal machinery, pushes the pathogene forward [265] (see Fig. 13). This particular motility mechanism, studied in in vitro assays, has allowed for the identification of the minimal set of proteins needed to actin-based motility, as well as the role of several of the main actin-related proteins [170,284]. It has also been used as a probe for the cell cytoskeleton network structures and visco-elastic properties in a position-dependent manner [153], and has shed light into the basic elementary principles of actin-based motility [71,87,194,214].

Except for very recent reports [68], nearly no force measurement has yet been done on single actin filaments. Due to their smaller bending rigidity, the corresponding stall force is expected to be orders of magnitude smaller than that of microtubules because of buckling phenomena. Instead, large forces can only be obtained when highly cross-linked actin filaments work as a whole and form a relatively rigid network, as it is the case in filopodia protrusion. Forces generated during actin-based propulsion have been measured on polymerizing actin gels, in particular using in vitro assays based on artificial biomimetic systems. Forces in the range of a few nano-Newtons have been found for gel comets originating from 2-μm-size polystyrene beads [182].

Other bio-mechanical characterizations of actin network's properties concern the study of its gel-like viscoelastic properties. In particular, transitions between a solid-like elastic material and a solution-like viscous material have been observed [73]. These could rely in part on

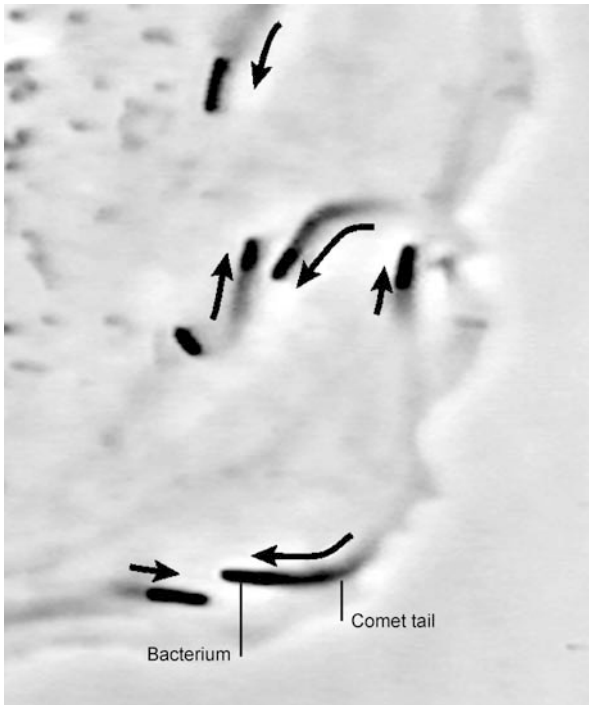

Cytoskeleton and Cell Motility, Figure 13
Listeria **bacteria moving in a cell via actin-based motility. This is a phase-contrast micrograph, a single frame from a video sequence (the whole movie can be seen on the following webpage: http://cmgm.stanford.edu/theriot/movies.htm). The kidney epithelial cell was infected about five hours before the acquisition of this video sequence. All of the bacteria in this cell are clonal descendants of a single individual. A bacterium and its associated comet tail are labeled. Bacteria are moving in the direction of the arrows. Source: reprinted from [36] with permission from Nature Publishing Group**

the biochemical-dependent mechanical properties of the actin filaments themselves [119], on the generic properties of such semiflexible-filament networks [75], or on the activity of motor proteins that help disentangling the network and thereby lead to its fluidization [106]. It has also been observed that cross-linked actin networks display an increase of their elastic modulus as a function of the stress applied, a nonlinear behavior known as *stress-stiffening* [63,259]. This might explain partly the interestingly rich properties of cellular rheology [99,135], which have been partly reproduced in in vitro measurements [77]. Among these, dynamical scaling of the stress stiffening (see e. g. [76]) has been proposed to be the signature of underlying self-similar mechanical properties of the cell cytoskeleton [18,61]. Finally, the intermediate filaments as well as the biochemical environment or preparation of the actin network have been proposed to play an important role in modulating its rheological properties [75,122]. This might contribute to the observed local changes in the elasticity of the cell as it moves, a crucial aspect for driving its motility [9,73].

Modeling Polymerization Forces

With general thermodynamic considerations, growth velocities of polymerizing filaments can be understood as follows: if k_{on} and k_{off} are the association and dissociation constants for monomers at the polymer tip, and δ is the distance a filament grows under addition of a single monomer, a typical growth velocity of the polymerizing filament is given by: $v = \delta[k_{\text{on}} - k_{\text{off}}]$. When experiencing a force f opposing polymerization, like the cellular membrane resistance at the leading edge of the advancing cell, filament-growth velocity becomes:

$$v(f) = \delta\left[k_{\text{on}} \exp\left(-q\frac{f a_1}{k_{\text{B}} T}\right) - k_{\text{off}} \exp\left((1-q)\frac{f a_1}{k_{\text{B}} T}\right)\right]. \quad (1)$$

In this expression, $k_{\text{B}}T$ represents thermal energy, $f a_1$ is the most probable work needed to add a monomer in the presence of the force f, and q is a parameter describing how much the force f influences the on-rate as compared to the off-rate. Under these assumptions, the maximal force a given filament can produce via polymerization, or *stall force*, is expressed as $f_s = (k_{\text{B}}T/a_1) \cdot \ln(k_{\text{on}}/k_{\text{off}})$. Even though good overall agreement with experimental data was obtained for individual microtubules while choosing $13a_1 = a$ and $q = 1$ (with a being the size of a tubulin monomer) [50], the so-derived stall force was too large as compared with experimental measurements. This led to revising the dynamics of the microtubule-polymerizing end, proposing $a_1 \simeq a$ and $q \simeq 0.22$ as better fitting parameters, pointing to a rich dynamics of microtubule polymerization [140].

To understand the origin of polymerization forces, the standard microscopic model relies on the ratchet mechanism, a rectified Brownian motion originally introduced in this context by Peskin et al. [219] to explain filopodia protrusion, *Listeria* propulsion as well as protein translocation. Filopodia protrusion in particular is thought to rely essentially on actin-polymerization forces: when reaching the cell membrane, growing F-actin filaments feel a force opposing their growth, and therefore exert a force on the membrane. Because of thermal fluctuations and membrane's as well as actin-filaments' finite bending rigidities, some space is constantly opened between the growing filament and the membrane. From time to time, an additional

monomer can thereby be added to the growing filament, which pushes the membrane forward[19]. In the simplest case of a single stiff protofilament, the distribution P of distances x between the filament and the membrane is given by the following Fokker–Planck equation:

$$\begin{aligned}\partial_t P(x) &= D\partial_x^2 P(x) + f\frac{D}{k_B T}\partial_x P(x) \\ &\quad + k_{on}P(x+\delta) - k_{off}P(x) \quad \text{for } x < \delta \\ \partial_t P(x) &= D\partial_x^2 P(x) + f\frac{D}{k_B T}\partial_x P(x) \\ &\quad + k_{on}[P(x+\delta) - P(x)] \\ &\quad - k_{off}[P(x) - P(x-\delta)] \quad \text{for } x > \delta\,,\end{aligned} \tag{2}$$

where notations are similar to the ones used in Eq. (1), and to which must be added the effective diffusion coefficient D for the distances x between the filament and the membrane. The time-dependence of P is implicit. Using vanishing-current conditions at the leading edge $x = 0$, the stall force can be obtained and is given by an expression analog to that of microtubules models, with $a_1 = \delta$ being the size of a G-actin monomer. Including bending fluctuations of the growing filament, this led to the Elastic Brownian Ratchet Model [195], a generalization of which is the Tethered Elastic Brownian Ratchet Model [197] that considers that some filaments are attached to the membrane via protein complexes (as it has been observed in the *Listeria*-propulsion mechanism for example) and therefore do not exert polymerization forces. When typical parameter values are plugged into these models, single actin-filament force generation is estimated to be of the order of 5–7 pN [198]. Taken into account that at the leading edge several hundreds of actin filaments per micron work together to drive the cell forward, the resulting force is of the order of nanonewtons per micron [197], a force large enough to tackle the membrane load and resistance. However, it has since then been claimed that motor proteins, called *end-tracking motors*, should be required to explain observed forces in the case of *Listeria* propulsion for example [49]. This work has been reviewed in [194].

Lateral interactions between filaments in an actin network have been investigated via models that take into account the branching structure of the network [37,38]. In particular, Autocatalytic Models assume that new branches are generated from existing ones, which leads to a growth velocity that is independent of the load [38]. To investigate the consequences of these models, two approaches have been followed, namely stochastic simulations of the growing actin network, tracking each filament position and orientation [37], and deterministic rate equations that include growth, capping and branching rates, and which led to a comparison between ratchet and autocalytic models [38]. Experimental tests of the two models have been performed in in-vitro systems, using *Listeria* propulsion as well biomimetic systems [271] (see next paragraph). While some *Listeria* studies favored the Tethered Elastic Brownian Ratchet Model [184], some studies using biomimetic systems favored the Autocatalytic Model [285], and several others neither of them. A possible explanation for these apparently contradictory results may be that different experimental studies led to analyzing different regimes of the force-velocity curve.

A Model System for Studying Actin-Based Motility: The Bacterium *Listeria monocytogenes*

As earlier stated, our understanding of eukaryotic actin-based motility has grandly benefited from the motility mechanism of the bacterium *Listeria monocytogenes* (see Sect. "Actin Gels"). While velocities of *Listeria* bacteria in a homogeneous environment are typically constant, some mutants progress in a saltatory manner [155]. This observation has been later reproduced in in vitro motility assays using latex beads coated with the bacterium transmembrane protein ActA (that further recruits Arp2/3) [35], or directly with VCA proteins, a sub-domain of WASP that is responsible for actin-branching and polymerization nucleation [22]. Such biomimetic systems have allowed for the direct measurement of the characteristic polymerization force that is produced by an actin gel [182], and for the overall study of actin-based motility mechanisms in simple and well-controlled conditions [194,222,271].

Theoretical understanding of such actin-based propulsion mechanisms has come from two different angles, namely molecular and mesoscopic, continuum models. We have already reviewed the molecular models that rely on brownian-ratchet mechanisms. They, in particular, have led to force-velocity curves that are consistent with some observations of *Listeria* motion [184,197]. Continuum models describe the actin network as a compressible elastic gel with an elastic modulus of about 5000 Pa [79,182,207] (see also Sect. "Macroscopic Phenomenological Approaches: The Active Gels"). When growing over a curved surface like the bacterium *Listeria* or a coated bead, the gel deforms as it grows by monomer additions on the particle surface, which in turn generates a stress that pushes the particle forward [79,207]. Monomer transport to the inner surface of the growing gel is purely diffusive, with a diffusion constant that has been estimated to be of the order of 2 $\mu m^2/s$ for actin monomers

[19] For an animated illustration, see http://www.jhu.edu/cmml/movies/anim/eBRatchet2.swf.

in an ActA-produced gel [221]. When originally initiated on a spherical object like a rigid bead, the growth of the gel layer starts isotropic, but ruptures into a comet-type growth because of mechanical instability. The instability relies on a positive feedback that involves creation of a tensile stress as the gel grows because of geometrical effects, and enhancement of the depolymerization rate or rupture of the gel in regions of enhanced tensile stress [252]. The instability occurs less rapidly with increasing bead size, which explains why movement is more often observed with small beads. This mechanism of tensile-stress accumulation and rupture can also explain the saltatory motion observed with *Listeria* mutants and coated beads in some conditions [22]: rapid phases of motion are due to the sudden rupture of the gel that pushes the particle forward, as slow phases of motion correspond to progressive build-up of lateral tensile stress. Depending on the size of the bead as well as the concentration of proteins at its surface, this dynamic instability can be present or not, which explains the observation of both continuous and saltatory regimes with coated beads as well as *Listeria* bacteria [22,23].

To further explore the properties of the actin gel and the *Listeria*-propulsion mechanism, experiments with soft objects like liposomes [80,270], endosomes [262] and oil droplets [28] have been performed. They show that the actin gel squeezes the object, compressing its sides and pulling its rear, an effect that gives it a pear-like shape (see Fig. 14). Analysis of the contour of the deformed objects provides informations on the distribution of the normal stress on the surface of the object. This could in principle allow for the derivation of the total force exerted on the load in the case for example of oil droplets, where interfacial tension is measurable and normal stress can be deduced from Laplace's law [28]. But in fact, assuming a constant surface tension, the integration of the normal stress over the surface of the droplet gives a zero net value of the force that is independent of the droplet shape, the latest being regulated by the variation of the polymerization velocity with normal stress. Instead, the distribution of actin-polymerization promoters on the surface of the droplet follows the gel elastic deformations, which in turn creates pressure variations inside the droplet, and thereby surface-tension gradients along its surface (see Fig. 14); and these are at the origin of the final non-zero net pushing force [28]. Finally, direct observation of the actin comet during its growth on coated beads has shown that the actin gel constantly undergoes deformations that depend on the protein composition of the motility medium they are placed in [212]. As a function of bead size and the concentration of cross-linkers or regulatory proteins, the bead velocity can be limited either by diffusion of the monomers to the coated surface, by polymerization velocity at the surface of the bead, or by the elastic stress built up in the gel. These findings, supported by experimental results, buttress the idea that actin-based movement is governed by the mechanical properties of the actin network, themselves tightly regulated by the proteins that are involved in actin dynamics and assembly (see Sect. "Other Cytoskeleton-Associated Proteins").

Another Example of Filament-Driven Amoeboid Motility: The Nematode Sperm Cell

Even though, as described above, protrusion of the leading edge in amoeboid motility is most commonly actin-driven, other cells, the nematode sperm cells, use another cytoskeletal protein to drive their motility: the Major Sperm Protein (MSP) [240]. Nematodes constitute one of the most common phyla of all animal kingdom, with over 80.000 different described species, and their sperm is

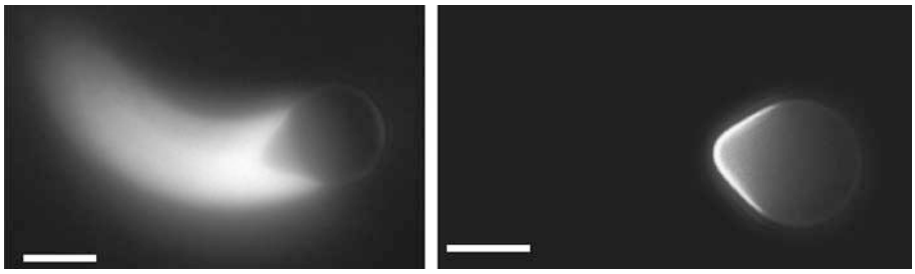

Cytoskeleton and Cell Motility, Figure 14

Actin-based propulsion with liquid drops. Oil drops are covered with VCA, placed in cell extracts that are supplemented with actin, and observed by fluorescence microscopy. *Left panel*: Note the bright actin comet and the pear-like shape of the droplet due to squeezing forces exerted by the actin gel. Scale bar is 4 μm. *Right panel*: VCA is labeled with fluorescin isothiocyanate (FITC). Note the inhomogeneous distribution of the actin-polymerization promoter on the surface of the droplet. Scale bar is 3 μm. Source: reprinted from [28] with permission from The American Physical Society

thought to be the only eukaryote cell type that do not possess the globular protein G-actin. These cells offer an ideal to study cell crawling since, dislike actin, MSP is a simpler, more specialized protein that do not possess as many regulatory or associated proteins, and in particular is not known to bind any molecular motor (at the exclusion of end-tracking proteins). MSPs being also apolar [31], the nematode motility constitutes one of the simplest of all cytoskeleton-driven motility mechanisms known to operate in vivo [240,286].

Similarly to what has been done with biomimetic systems to study actin-based propulsion, motility assays using vesicles derived from the leading-edge of nematode sperm cells of *Ascaris* species, have shed light into the mechanisms at play [120]. In the presence of ATP, growth of MSP fibers are capable of pushing the vesicle forward, their polymerization being driven by specialized proteins located within the vesicle membrane, a mechanism that ressembles very much the *Listeria*-propulsion mechanism. But, contrary to ATP-driven actin treadmilling, MSPs assemble into apolar filaments and lack a nucleotide binding site for ATP hydrolysis. To power membrane protrusion, it has recently been proposed that motor end-tracking proteins processively polymerize MSP filaments, while keeping the elongating filaments' ends in contact with membrane-associated proteins [48]. For cell progression however, a second force is required, namely a traction force that pulls the cell body forward once the advancing lamellipodium has been anchored to the substrate on which the cell is crawling. In actin-based amoeboid motility, this process is motor-driven, but nematode sperm cells use instead the sensitivity of their MSP to pH, whose decrease provokes reorganization, depolymerization and in fine contraction of the network [121,187,286]. Polarity in the cell is maintained by an influx of protons close to the cell body, which creates a pH gradient in the lamellipodium and powers this process [137].

To quantitatively understand the mechanism underlying this motility, both microscopic and phenomenological models have been proposed. In the proposed microscopic models, mechanisms underlying the traction-force generation by solely cytoskeletal disassembly can be qualitatively understood as follows [27,286]: because of pH gradient, MSP filaments tend to bundle at the front, and split apart and disassemble at the rear [187]. A bundle with N filaments being much stiffer than N isolated individuals (with an effective persistence length of $N^2 L_p$ as each individual has a persistence length L_p), it pushes the cell membrane at the leading edge where filaments are bundled, while splitting filaments exert contractile forces at the rear. Indeed, because of entropic effects, filaments tend to retract once split apart. Finally, even further decrease in pH creates weakening of the attachments and dissociation of the filaments for monomeric MSPs to be recycled at the front [286] (see also [196]). In the proposed phenomenological approach however [124], the sensitivity to pH is described as influencing the equilibrium swelling properties of the gel only. As the gel treadmills towards the rear end where acidic conditions are found, it tends to contract by an isotropic multiplicative factor Λ that is position-dependent. General elasticity theory of continuous media allows to express the strain tensor as:

$$u_{\alpha\beta} = \tfrac{1}{2}(1-\Lambda^2)\delta_{\alpha\beta} + \tfrac{1}{2}(\partial_\alpha u_\beta + \partial_\beta u_\alpha)\,, \qquad (3)$$

where u_α are the components of the displacement vector (with $\alpha = x, z$). Assuming linear elasticity theory[20], the stress tensor is then obtained as $\sigma_{\alpha\beta} = \lambda u_{\gamma\gamma}\delta_{\alpha\beta} + 2\mu u_{\alpha\beta}$, where λ and μ are the Lamé coefficients, which further leads to a position-dependent tensile stress as was introduced phenomenologically in [27]. While traveling through the lamellipodium, tangential stress builds up, which leads to rupture of the adhesion points once a critical force has been passed, and eventually drags the cell body forward. Therefore, within this framework, only one parameter is directly controlled by the pH – namely Λ – and the pH in particular does not need to influence directly adhesion strength.

Motor-Driven Motility

Generic Considerations

Despite the major role played by polymerization forces in cellular motility, and in particular as we have previously seen in amoeboid motility, a vast amount of diverse motile processes in eukaryotic cells is driven by motor proteins (see Sect. "Molecular Motors"). Theoretical studies of molecular motors started with the cross-bridge model published independently by A. F. Huxley and H. E. Huxley to explain the relative sliding of myosin filaments with respect to actin filaments in cross-striated muscle fibers [108,112]. This approach was later formalized by Hill [97], who introduced the notion of different "states" of a motor protein, each of these corresponding to a thermodynamic-equilibrium state. Interpretation of these different states was given in terms of different conformations of the motor protein and its interaction with the filament, or in terms of the state of the hydrolysis reaction of ATP, or both [54,159]. Justification for considering different thermodynamic-equilibrium states relied on

[20] For an introduction to the elasticity of continuous media, see, e. g., [154].

the observation that for the transient response of muscles, the fastest response was known to be in the range of milliseconds, as thermal equilibrium on molecular characteristic length scales of 10 nm occurs after at most a few hundreds of nanoseconds. In this class of models, progression of the motor along the filament relies on asymmetric transition rates of the particle between the different states, for which asymmetry of the filament and energy consumption by the motor is required. Typically, after one cycle of conformational states, the motor protein has progressed by one or several allowed binding sites on the periodic lattice represented by the cytoskeletal filament. In between, up to five or six different states could be involved [97,172]. Experimental confirmation came later with the direct observation of walking steps displayed by advancing molecular motors [258]. Such observations were first obtained studying kinesin motors in in-vitro motility assays, and later with myosin motors displacing a filament that was attached to two glass beads placed in laser-trap potentials [32]. For a review, see [246].

Another class of models relies on the generalization of Feynman's famous "thermal ratchet", in which the presence of different heat baths (namely thermal baths at different temperatures) can rectify the brownian motion of a given particle and lead to its directed motion [64]. For motor proteins, as we have already discussed, temperature inhomogoneities in the system cannot hold long enough to ground the mechanism. Instead, various different isothermal rectifying models have been discussed to describe the underlying mechanisms of different biophysical processes [13,177]. Among these, one can mention the translocation of proteins and force-generation by linear molecular motors (which includes cytoskeletal motors, but also motors acting on DNA or RNA, like DNA-polymerases, RNA-polymerases and helicases), the ion transport in ion pumps, and the rotary-motor processes such as the one found in the F0F1-ATPase or the bacterial flagellar motor. Such isothermal rectifying processes and their underlying physical principles have been extensively reviewed in [129,233]. They all rely on a Langevin type of description of an overdamped particle of position x, moving in a spatially-periodic potential $W(x)$ that reflects the motor-filament interaction, and subjected to a viscous friction with coefficient ξ and a fluctuating force $f(t)$ that reflects the stochasticity of thermal fluctuations:

$$\xi \frac{\mathrm{d}x}{\mathrm{d}t} = -\partial_x W(x) + f(t) \,. \tag{4}$$

To rectify brownian motion, three different approaches have been mainly followed, namely (i) random forces $f(t)$ whose fluctuations do not satisfy the fluctuation-dissipation (FD) relation, (ii) fluctuating potentials $W(x, t)$ that are time-dependent, and (iii) particle fluctuating between states, where different states indexed by $i = 1, \ldots, N$ reflects different conformations of the protein and interactions with the filament.

In the following, no attempt will be made to extensively present the literature on molecular motors. We shall instead only briefly sketch the generic considerations of the main proposed models, and focus more closely on a particular example of them, the two-state model, which has allowed for an understanding of the appearance of spontaneous oscillations in systems of coupled motors. This generic mechanism has been proposed to underly axonemal beating, the generic mechanism that powers eukaryotic flagellar and ciliary-based motilities.

Phenomenological Description Close to Thermodynamic Equilibrium

Sufficiently close to thermal equilibrium, out-of-equilibrium perturbations can be described using a generic linear-response theory that introduces generalized forces which drive generalized currents [45]. In the context of molecular motors, the generalized forces that drive the system out of equilibrium are the mechanical force f_{ext} acting on the motor (including drag), and the chemical-potential difference $\Delta\mu$ of the chemical reaction ATP $\rightleftharpoons$ ADP + P_i that drives motor power [129,217]. Linear-response theory then gives:

$$v = \lambda_{11} f_{\mathrm{ext}} + \lambda_{12} \Delta\mu \tag{5}$$

$$r = \lambda_{21} f_{\mathrm{ext}} + \lambda_{22} \Delta\mu \,, \tag{6}$$

where the coefficients λ_{ij} are phenomenological response coefficients. Here λ_{11} and λ_{22} can be viewed respectively as a standard and generalized mobilities, and λ_{12} and λ_{21} as mechano-chemical couplings. Onsager relations impose that $\lambda_{12} = \lambda_{21}$, and the Second Law of Thermodynamics insures that the dissipation rate is positive: $T\dot{S} = f_{\mathrm{ext}} v + r\Delta\mu \geq 0$. Whenever both of the two terms that appear in this inequality are positive, the system is passive, but it works as a motor when $f_{\mathrm{ext}} v < 0$, and as generator of chemical energy when $r\Delta\mu < 0$. The latter function is not known for linear motors, but is the common mode of operation of F0F1-ATPase, the protein complex that synthetizes ATP from electro-chemical energy that is stored in proton gradients [5,290] (see Sect. "Molecular Motors"). The reversibility of this rotary engine can

be related to the predicted reversibility that comes out of linear-response theory: in the absence of an external force, reversing the chemical potential difference $\Delta\mu$ should reverse the sign of the velocity v without a need for a change in the mechanism.

Hopping and Transport Models

Within the first class of models that we have mentioned earlier, namely hopping models between different discrete equilibrium states of the motor-filament system, generic transition rates and periodicity in theses transition rates, related to periodicity of the filament, are generally assumed. Within this framework, one can calculate the mean velocity v and the diffusion coefficient D of the molecular motor from analyzing the generic associated Master Equation [46]. For non-zero mean velocity to occur, at least one of the transitions between states must break detailed balance, a feature that can be associated with chemical energy consumption. In the simplest case of only two possible states of the motor protein, one can derive simple compact expressions for v and D [66]. Their dependence on the external force further leads to the derivation of the force-velocity curve, as well as a simple expression for the stall force, namely the force at which the motor protein ceases to progress on average.

To describe protein trafficking on a filament where many motors are simultaneously engaged, like it is commonly the case for example in organelle transport by kinesin proteins along microtubules, one can reduce the number of states that a motor can occupy to one per filament binding site. Motors are then represented by particles that move on a one-dimensional lattice with homogeneous transition rates, to which attachment and detachment rates from and toward the bulk can be added. This description belongs to a class of driven lattice-gas models that are used to study various transport phenomena, like ionic transport in solids or traffic flow with bulk on-off ramps [247]. In the simplest case of the absence of particle attachment and detachment, the model reduces to the Asymmetric Simple Exclusion Process (ASEP) [250], originally introduced to describe the translation of messenger RNA by ribosomes [173]. Including attachment-detachment rates, the next simplest case describes the space surrounding the filament as a reservoir of uniformly-distributed particles [147,216]. A second possibility is to include the dynamics of unbound particles explicitly, for example on a cubic lattice [161]. Boundary terms can also play an important role, and different possible choices have been considered depending on the biological situation [138].

Consequences of these models are illustrated by various important phenomena. Among these, one can find the followings: anomalous transport due to repeated attachments and detachments [4,161,205], domain walls that separate regions of high and low motor densities in the filament [161,216], phase separation in systems with two motor species [60], and phase transitions when cooperative binding-unbinding is introduced [139]. For a recent review on these collective traffic phenomena, see [40] and references therein.

The Two-State Model

One model that proved to be particularly useful for describing the rectification of brownian motion via coupling to chemical hydrolysis reactions, is the so-called "two-state model". In this description, the molecular motor switches stochastically between two different interaction states with the filament, that are described by two different asymmetric and l-periodic potentials W_1 and W_2 representing polarity and periodicity of the filament [14,15,128,129,178,217,218,227] (see Fig. 15a). The dynamics of this system can be conveniently represented in terms of two coupled Fokker–Planck equations that describe the evolution of the probability density $P_i(x,t)$ of the motor to be in state $i = 1, 2$ at position x at time t. Explicitly, we have:

$$\begin{aligned} \partial_t P_1 + \partial_x J_1 &= -\omega_1 P_1 + \omega_2 P_2 \\ \partial_t P_2 + \partial_x J_2 &= \omega_1 P_1 - \omega_2 P_2 \,. \end{aligned} \tag{7}$$

The currents $J_i (i = 1, 2)$ result from diffusion, interaction with the potentials W_i, and the external force f_{ext}:

$$J_i = \mu_i [-k_\text{B} T \partial_x P_i - P_i \partial_x W_i + P_i f_{\text{ext}}] \,. \tag{8}$$

The transition rates $\omega_i(x) (i = 1, 2)$ between the two states are driven out of equilibrium by ATP consumption, whose strength can be represented by a single parameter Ω using the following form:

$$\omega_1(x) = \omega_2(x) \exp\left[\frac{W_1(x) - W_2(x)}{k_\text{B} T}\right] + \Omega \Theta(x) \,, \tag{9}$$

where $\Theta(x)$ is a l-periodic function of integral one over one period. For $\Omega = 0$, detailed balance is satisfied. Within this formalism, it has been shown that both spatial symmetry and detailed balance need to be broken for directed motion to occur, which has been quantified in terms of an effective potential W_{eff} [227].

Coupled Motors and Spontaneous Oscillations

Directly interesting for eukaryotic cellular motility driven by cilia and flagella, is the motion of motors with respect

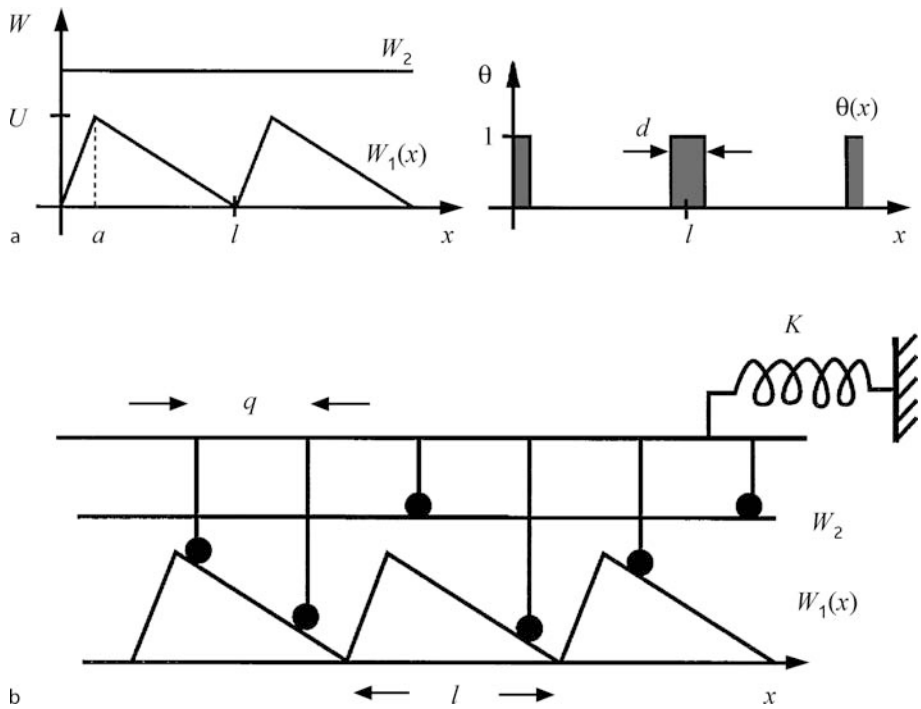

Cytoskeleton and Cell Motility, Figure 15

Schematic representation of the two-state model as used in [128] for the calculation of the collective behavior of rigidly-coupled non-processive motors. a, *left panel*: W_1 represents the attached state of the motor to the filament and is therefore asymmetric and *l*-periodic; W_2 represents the detached state where no interaction with the filament occurs. a, *right panel*: Illustration of the *l*-periodic function $\Theta(x)$ as it appears in Eq. (9) and which represents ATP consumption in the system, here at preferred locations of the motor on the filament, or at preferred configurations of the motor protein for ATP hydrolysis. b Schematic representation of a collection of rigidly-coupled motors with periodicity *q* interacting with the filament, and coupled to its framework by a spring of stiffness *K*. Under some conditions, this model leads to spontaneous oscillations. Both structures and behaviors are reminiscent of skeletal-muscle myofibrils' oscillations or axonemal beating (see Sect. "Axonemal Beating"). Source: reprinted from [128] with permission from The American Physical Society

to their associated cytoskeletal filament when a collection of them is rigidly coupled. Such structures are typical of skeletal-muscle structures (where Myosin II motors associate into the so-called "thick filaments"), or of the axonemal structure that drives oscillatory motions in cilia and flagella. Such systems have been studied using the two-state thermal-ratchet model [127,128], and a crossbridge model [30,97,108,112,279,280]. Here we shall discuss essentially the case of an ensemble of motors that are rigidly linked to each other and walk collectively on a cytoskeletal filament whose interaction with the motors is described by the two-state model [129]. In the case of randomly distributed motors, or motors distributed periodically with a period q that is incommensurate with the filament period l, the probability density $P(\xi, t)$ of finding a particle at position $\xi = x \bmod l$ in either state $i = 1$ or 2, approaches the value $1/l$ in the case of an infinitely-large number of motors. In a mean-field approximation, equations of motion for the probability densities read:

$$\begin{aligned} \partial_t P_1 + v\partial_\xi P_1 &= -\omega_1 P_1 + \omega_2 P_2 \\ \partial_t P_2 + v\partial_\xi P_2 &= \omega_1 P_1 - \omega_2 P_2 \,. \end{aligned} \tag{10}$$

The force-velocity curve can then be computed using the fact that $f_{\text{ext}} = \eta v - f$, where f_{ext} is the external force applied, η is the friction coefficient per motor protein, and f is the force per motor protein exerted by the potentials:

$$f = -\int_0^a \mathrm{d}\xi (P_1 \partial_\xi W_1 + P_2 \partial_\xi W_2) \,. \tag{11}$$

Expressing P_2 as $P_2 = 1/l - P_1$, and P_1 as a series expansion in powers of the velocity v, one finds a generic series expansion for the force-velocity curve f_{ext} as a function of v in the steady state. As a function of the distance to thermal equilibrium Ω, controlled by ATP consumption by the motors, and which appears as a control

parameter for the dynamics with a critical value Ω_c, the curve $f_{ext}(v)$ can be strictly monotonic for $\Omega \leq \Omega_c$, or present some multi-valuated regions for $\Omega > \Omega_c$, where two stable velocity regimes exist for a given external force. For symmetric potentials, the system is quiescent with $v = 0$ at zero force for $\Omega \leq \Omega_c$, but present two possible opposite spontaneous velocities for $\Omega > \Omega_c$, a spontaneous symmetry breaking that is characteristic of second-order phase transitions with characteristic mean-field exponents. Such a reversible spontaneous movement has been observed in a motility assay with NK11 proteins, a mutant of the kinesin protein Ncd that has lost its directionality [55]. In addition, when the external force is varied, a hysteresis is found for $\Omega > \Omega_c$, an experimental observation of which has been reported for a myosin II motility assay under near-stalling conditions induced by electric fields [239]. Numerical simulations of both situations have been performed using the two-state model with a finite number of motors and in the presence thermal noise [17].

Axonemal Beating

The previous and related models have been used to describe the spontaneous oscillations that have been observed in skeletal-muscle myofibrils' oscillations or axonemal beating [72,209,287]. In these cases, it has been proposed that the coupling of the motor backbone to a spring prevents spontaneous steady-state velocities to occur, but instead leads to spontaneous oscillations [128] (see Fig. 15b). In the case of axonemal beating, of most relevance for eukaryotic cell swimming, the elastic force results from bending of the microtubules and leads to self-organization of the dynein motors. This collective behavior has been proposed to explain the bending waves of cilia and flagella [30,176] and analyzed in the framework of the two-state model [33,34]. Close to the oscillatory instability, wave-patterns can be computed, whose frequencies and shapes depend on the filament length and boundary conditions, and which are in good agreement with observed flagellar beating patterns [278].

In the case of cilia however, beating patterns are typically assymetric, like it is at best exemplified in the case of the two cilia of the green alga *Chlamydomonas* [29], an observation that cannot be accounted for by investigating beating patterns at the oscillatory instability only. Using the same underlying model, it has been proposed that this assymetry originates from the presence of transverse external flow that occurs as the organism is swimming [89]. In that case, the cilium tends to beat faster and quite straight in the direction of the flow, whereas it comes back slower and more curved against it, a beating pattern that evokes power and recovery strokes. Hydrodynamics has also been proposed to be responsible for dynamic coupling of adjacent cilia, which results in both spontaneous symmetry breaking and synchronization of their beating pattern [89]. This effect could be at the basis of the observed beating waves that propagate for example on the surface of *paramecia*[21] as they swim, which originate from a constant phase difference in the beating of the adjacent cilia, and which have been called *metachronal waves* [29,88,89]. This could also underly the breaking of symmetry that occurs in mammalian development during gastrulation, and which is responsible for left-right asymmetry. In that case, it has been shown that beating of cilia located in a transiently-formed epithelial chamber known as the *node*, create a directional flow which transports signaling molecules preferentially to one side [29,117,185]. There, beating patterns are unusual in that cilia swirl in vortical fashion rather than beat [208], and hydrodynamic-driven synchronization of these three-dimensional beating patterns has also been studied [281].

Putting It Together: Active Polymer Solutions

The last part of this review is devoted to the presentation of some generic descriptions of the cell cytoskeleton, when considered as a network of long protein filaments that are cross-linked by a variety of smaller proteins. As already discussed, filamentous proteins that are involved in the cell-cytoskeleton dynamics are mostly F-actin and microtubules (made of G-actin and tubulin monomers), with which interact cross-linkers that can be either passive and stationary (such as α-actinin), or active and mobile, consisting then of clusters of molecular motors (mostly myosin and kinesin motors) (see Sect. "The Cell Cytoskeleton"). To model these systems, different complementary approaches have been developed, namely computer simulations [202,203,260], and analytical descriptions that can be roughly divided into three categories, namely microscopic, mesoscopic, and macroscopic or phenomenological hydrodynamic descriptions. The first analytical approaches that have been developed correspond to mesoscopic descriptions. There, starting from a microscopic description of the filaments, the effect of active cross-linkers is described via motor-induced relative velocities of paired filaments, where the form of such velocities is inferred from general symmetry considerations [144,145,148,201,251]. Microscopic approaches start from what is known about the properties of the different molecular players involved and their interactions, and aim

[21] An illustration of a paramecium can be seen in Fig. 1, center.

to build large-scaled coarse-grained theories from statistical physics' principles [2,3,11,12,146,165,166,168]. Finally, macroscopic hydrodynamic approaches have adopted a more phenomenological point of view: they harness the generic symmetry and dynamical properties of the players involved, to derive directly effective continuous theories in terms of a few coarse-grained fields [94,125,130,150,151,152,157,231,245,254,267,282,295]. Recently, attempts have been made to bridge microscopic to macroscopic models, and compare what results in being similar and different in the two types of approaches [3,164,166].

Interests for describing the cytoskeleton as an ensemble of filamentous polymers actively connected by cross-linkers have come to the scene since self-organizations of motor-filament mixtures were observed experimentally [203,260,272]. Among these, complex patterns that include asters, vortices, spirals and connected poles or networks have been observed in confined quasi-two-dimensional systems in in vitro experiments [203,260] (see Fig. 16). Patterns where shown to be selected in a way that is dependent on motor and ATP concentrations, and numerical simulations based on microscopic models of rigid rods connected by active elements have shown to be capable of reproducing the experimental results [260]. Further experiments were performed on systems that resemble more closely a living cell, and which, while being simplified versions of it, still exhibit some of its behaviors. Along these lines, formations of bipolar spindles that do not contain any microtubule-organizing center were observed using cell extracts [116], and cell fragments that contain only the actin cortex where found to self-propagate on a substrate, with coexistence of locomoting and stationary states [59,277].

Mesoscopic Approaches

Theoretical modeling of the cell cytoskeleton have benefitted from the knowledge accumulated in equilibrium statistical physics of polymer solutions and liquid crystals [44]. However, the cell cytoskeleton is an active medium for which new analysis techniques needed to be developed in order to describe, for example, its ability to actively self-organize, exert forces and create motion. First, theoretical models have aimed to describe pattern formations in sys-

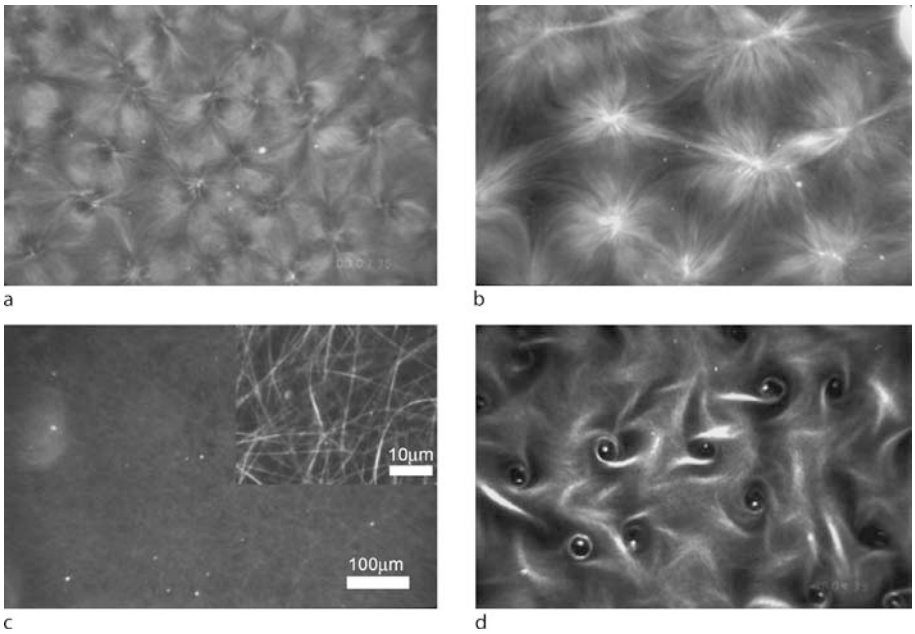

Cytoskeleton and Cell Motility, Figure 16

Different large-scale patterns formed through self-organization of microtubules and kinesin motors as reported in [203]. Initially uniform mixtures of proteins heated to 37°C displayed different patterns after 7 min of self-organization. Patterns are shown at equal magnification; the samples differ in kinesin concentration. a A lattice of asters and vortices obtained at 25 g ml^{-1} kinesin concentration. b An irregular lattice of asters obtained at 37.5 g ml^{-1} kinesin concentration. c Microtubules form bundles at 50 g ml^{-1} kinesin concentration (scale bar, 100 µm); insert: at higher magnification (scale bar, 10 µm). d A lattice of vortices obtained at a kinesin concentration smaller than 15 g ml^{-1}. Source: courtesy of François Nédélec; reprinted from [203] with permission from Nature Publishing Group

tems of actively-driven rigid filaments in one-dimensional geometries [144]. Such configurations are represented in vivo for example by stress fibers that are important for cell-force generation, contractile rings that form during cytokinesis, or the formation of filopodia for forward protrusion during amoeboid motility. There, dynamical equations that govern individual filaments were introduced using "mesoscopic" mean-field models, where the relative sliding of paired polar filaments is described by an effective relative velocity that is induced by many individual events of motor activity. General constraints on these relative velocity fields are imposed by symmetry considerations that rely on the orientational polarity of the filaments (see, e. g., [144]). In such systems, polarity sorting [201], contraction [144,251], as well a propagating waves [148] emerge from the models. Interestingly, it has been shown that relative velocity of filaments of the same orientation is important for contraction to occur [144], a phenomenon that has been suggested to rely on motor-density inhomogeneities along the filament that create inhomogeneous filament interactions along their lengths [147]. More generally, the whole bifurcation diagram of this generic one-dimensional model has been established, and motor-distribution dynamics has been introduced that lead to contractile states with the generation of contractile forces, of most relevance for stress fibers' as well as contractile rings' dynamics [145]. Interestingly, the simplest version of these models, when only one possible polymer orientation is considered, has been mapped to hopping models that describe driven-diffusive systems [199]. In the absence of active cross-linkers, the model reduces to a class of hopping models known as the Zero Range Process (ZRP), for which exact analytical solutions of the steady state as well as one-dimensional phase transitions have been described [92]. In the generic case however, the dynamics defines a new class of driven-diffusive systems, which can still be mapped in some cases to the ZRP analytic solution, even though with a different criterium for condensation to occur [199].

Microscopic Approaches

Microscopic approaches to describing the cell cytoskeleton dynamical behavior model explicitly all the considered different processes and interactions that occur between the different molecular players, and aim to derive effective dynamical equations for the different density fields that enter the description, by coarse-graining the microscopic dynamics. Most studies that have done so model the motor-filament system as an ensemble of rigid rods of fixed lengths, which interact via point-like cross-linkers that can induce relative sliding as well as rotational motions [2,3,146,165,166]. Exceptions to this rule are theoretical descriptions of the mechanical response of active-filament solutions to high frequency stimuli [162,168]. There, anomalous fluctuations occur that are dominated by the bending modes of the filaments in combination with the activity of the cross-linkers. Inspired by polymer physics at thermodynamic equilibrium, excluded-volume interactions as well as entanglements are taken into account in the description. In particular, the system exhibits accelerated relaxation at long times due to directed reptation that relies on active phenomena.

Attempts at deriving the motor-mediated interaction between filaments from microscopic descriptions have been performed in [2,3,11,12,165]. See also the review [163]. In [11,12], a generalization of the Maxwell model of binary collisions in a gas is used to describe the dynamics of polar rods whose inelastic and anisotropic interactions reflect the presence of active crosslinkers. Orientational instabilities lead to bundling as well as the formation of asters and vortices patterns. In [2,3,165], filaments are described as rigid rods of fixed length, and hydrodynamics is obtained by coarse-graining the Smoluchowsky equation for rods in solution, coupled via excluded-volume and motor-mediated interactions. There are two main motor-mediated mechanisms for force exchange among filaments. First, active crosslinkers induce bundling of filaments, building up density inhomogeneities. Second, they induce filament sorting as a function of their polarization state. As a result, phase diagrams are derived that show instabilities of the homogeneous states at high filaments' and crosslinkers' densities. In particular, all homogeneous states are rendered unstable by the same mechanism of filament bundling, a fact reminiscent of the effect described in [144] where the interaction between filaments of the same orientation has been shown to be important for contraction to occur. Interestingly, the broken directional symmetry of the polarized state yields an effective drift velocity that describes filament advection. This convective-type term describes a genuine out-of-equilibrium contribution that is structurally not present in phenomenological descriptions based on systematic linear expansions close to thermodynamic equilibrium (see below). Such a term is reminiscent of the one introduced in earlier studies of self-propelled nematic particles [230,254,267], as well as of the explicit flow of the solvent taken into account in [150,151,282]. Other effects of higher-order nonlinear terms have also been discussed in [292,293], where pattern selection between stripe patterns and periodic asters occurs via nonlinear interactions.

Macroscopic Phenomenological Approaches: The Active Gels

The third category of approaches that have been developed to try to understand the dynamical behavior of the cell cytoskeleton as a whole, are effective phenomenological theories which rely on the hypothesis that large length- and time-scales behaviors of the cytoskeleton are largely independent of the microscopic details that underly its dynamics, but depend instead only on a few macroscopic fields that capture the relevant behavior. Sufficiently close to thermodynamic equilibrium, these relevant fields describe the *hydrodynamic modes* (or *slow modes*) of the dynamics, namely the modes whose relaxation rates go to zero at long wavelengths. As for equilibrium systems close to a critical point, such hydrodynamic modes correspond to the *conserved densities* on the one hand, and the *order parameters* that break continuous symmetries on the other hand [45,100]. To write generic theories for the dynamics of these hydrodynamic modes, standard approaches consist in writing systematic expansions in the different couplings that are allowed by the symmetry properties of the system. In the vicinity of a critical point, which occurs generally when a continuous symmetry is spontaneously broken after a second-order phase transition has been traversed, the concept of renormalization group has given a theoretical framework to identify universality classes: starting from the full nonlinear expansions of the underlying stochastic dynamics, only a few relevant parameters matter for the asymptotic scaling laws that occur at the transition [65,100,294]. Even though originally developed to study equilibrium critical points, the renormalization-group concept has allowed for the characterization of some out-of-equilibrium universality classes (see e. g. [70,238,261] and the review [98]).

Away from such remarkable points however, any term allowed by symmetry in a systematic expansion is a priori relevant. The standard approach for systems close to thermodynamic equilibrium consists in writing generalized thermodynamic forces and fluxes that are related to each other by linear-response theory. Constraints on the generic coupling constants to linear order emerge from the spatio-temporal symmetries of the system, and correspond to the Onsager relations and the Curie principle [45]. Inspired by the dynamics of liquid crystals [44,183], a hydrodynamic theory has been developed that describes the cytoskeleton as a visoelastic polar gel, driven out of equilibrium by a source of chemical energy [150,151,152,282,295] (this work has been reviewed in [130]). Among other applications, such or similar approaches have been applied to the description of pattern formation in motor-microtubule mixtures [157,245], as well as the collective dynamics of self-propelled particles [52,94,254,267].

Originally, this hydrodynamic theory has been presented as a generic theory for active viscoelastic materials made of polar filaments, referred to as *active polar gels* [151]. Within the framework of the previously-described general formalism, here applied to the cytoskeleton, conserved quantities are the different number densities that enter the dynamics, namely the number densities of subunits in the gel, of free monomers, and of respectively bound and unbound motors to the filaments. To these must be added the solvent density and the total mechanical momentum. Source terms in the conservation equations correspond to polymerization and depolymerization of cytoskeleton filaments, attachment and detachment of motor proteins to the filaments, and the potential presence of an external force. Order parameters correspond to orientational order parameters that originate from the polarity of the filaments. Namely, they correspond to momenta of the local polarization vector of individual filaments $\boldsymbol{u}$, and most often only the first momentum $\boldsymbol{p} = \langle \boldsymbol{u} \rangle$ is considered that represents the locally-averaged polarity in the gel[22]. To these must be added a crucial parameter that drives the system out of equilibrium, and which originates from the actively-maintained source of chemical energy in the cell, corresponding to out-of-equilibrium concentrations of ATP versus ADP and P_i. This parameter $\Delta\mu$ is expressed as the difference in chemical potentials of ATP versus ADP plus P_i: $\Delta\mu = \mu_{\mathrm{ATP}} - (\mu_{\mathrm{ADP}} + \mu_{\mathrm{P}_i})$. After identification of the different conjugated generalized fluxes and forces, that are split into dissipative and reactive parts as a function of their properties under time-reversal symmetry, the constitutive equations that specify the dynamics are written in terms of a generalized Maxwell model, which describes the viscoelastic dynamical properties of the gel. Under its simplest form and for nonpolar viscoelastic gels, the Maxwell model writes

$$\left(1 + \tau \frac{D}{Dt}\right) \sigma'_{\alpha\beta} = 2\eta \left(\nu_{\alpha\beta} - \frac{1}{d}\delta_{\alpha\beta}\nu_{\gamma\gamma}\right) + \bar{\eta}\delta_{\alpha\beta}\nu_{\gamma\gamma}\,, \tag{12}$$

in d dimensions. Here $\nu_{\alpha\beta}$ and $\sigma'_{\alpha\beta}$ are the symmetric parts respectively of the velocity-gradient tensor $\partial_\alpha \nu_\beta$ and the viscous stress tensor, η and $\bar{\eta}$ are respectively the shear and bulk viscosities, and $\tau = E/\eta$ is the viscoelastic relaxation time that is related to the Young elastic modulus E

[22]The next momentum $q_{\alpha\beta} = \langle u_\alpha u_\beta - d^{-1}\, \boldsymbol{p}^2\, \delta_{\alpha\beta} \rangle$, where d is the dimension of space, is a symmetric traceless tensor of order two that corresponds to nematic order.

and to the shear viscosity η. This relaxation time describes the crossover between an elastic behavior at short times that resembles that of a solid gel and a viscous behavior at long times that resembles that of a fluid[23]. Finally, D/Dt represents a convective corotational derivative that takes into account invariance with respect to translations and rotations in the system. In the general framework of active polar gels close to thermodynamic equilibrium, generic linear couplings are added to this model following the general procedure described above. Extensive presentations of the complete formalism can be found in refs. [130,151], which also include discussions about its limitations and some of its possible extensions. In particular, in [130], extensions that aim to include the contributions of rotational viscoelasticity, of some nonlinear couplings and of passive as well as active sources of noise are briefly discussed. The main developed extension so far concerns the generalizations of this formalism to multi-component active gels that are now being developed [125,167]. These allow in particular to take into account the possible permeation of the cytosol through the cytoskeletal gel, which affects force balance in the system, and which might be of importance for cell motility.

Despite its recent development, this generic theory of active polar gels has been applied to the description of some systems that are of particular relevance for cell motility or experimental situations observed in vitro. Its first application was the study of topological defects in the polarity field of the gel that lead to the formation of patterns such as asters, vortices and spirals [150]. As a function of two dimensionless parameters representing the relative strength of the coupling to the chemical potential $\Delta\mu$ and to the bend and splay moduli of the polar gel, a phase diagram was derived where vortices and asters give rise to rotating spirals via dynamic instabilities. These relate to the spatial patterns that have been observed in vitro [203,260], as well as to the creation of spontaneous motion, of most relevance for cell motility. In a different geometry, namely a cylinder of finite diameter and length, the formalism has been applied to establish a phase diagram of ring formation that contains phases of one or multiple rings, and which can be quiescent or oscillating [295]. This is relevant for understanding the formation and localization of cortical rings that form prior to cytokinesis and for which double-ring formation has been observed with certain plant cells [84]. To understand the generation of active flows that might be of relevance for cell crawling, a generic phase diagram has been derived for a two-dimensional active polar film that is compressible [282]. Compressibility here might refer to different thicknesses in a three-dimensional incompressible gel that is described in two dimensions after integration of the density fields over its thickness. Within this framework, density fluctuations couple generically to polarity splay, and different topological phases of the gel-polarity organization are found that could correspond to some of the previously-observed patterns in the experimental literature. Finally, the description of spontaneous movements of thin layers of active gels has been applied to the study of cell locomotion on a solid substrate that occurs via the protrusion of the actin-filled lamellipodium at the leading edge of the cell [152]. Reducing the lamellipodium description to a two-dimensional gel protruding in one dimension, and with a spatially-dependent thickness, the steady-state thickness profile as well as the flow and force fields have been computed. One particularly striking aspect of cell crawling that is described by this formalism is the presence of a retrograde flow of the gel as the cell is crawling. This aspect has been quantified in earlier experiments performed on fish epidermal keratocytes [275]. It has been shown that while the cell is crawling, treadmilling of actin filaments happens faster than global motion of the cell, such that the actin cortex is moving rearward with respect to the substrate, in a direction opposite to the movement of the cell [42,131,160]. Similar questions have been addressed using different theoretical frameworks in [7,142,242].

Comparisons of the Different Approaches to Describing Active Polymer Solutions

With these different ways of approaching the description of the dynamics of the cell cytoskeleton as a whole, a natural question is to ask to what extent these different approaches are similar and different, and which aspects of the cytoskeleton or cell behavior can be or not described by each of the theories. For answering these questions, connections between the different approaches have been made, first between mesoscopic and hydrodynamic descriptions [149,164]. In [164], a generalization of the mesoscopic model introduced in [144,148] is developed to obtain a set of continuum equations in unconfined geometries. A phase-diagram is derived that results from the stability analysis of the homogeneous state of actively crosslinked polymers, taking into account excluded-volume interactions and estimates of entanglement in two and three dimensions. It is found that an instability occurs as the bundling rate between filaments of the same orientation is

[23] Note that only one relaxation time is assumed to characterize the system, as some experiments suggest that a power-law distribution of relaxation times is better suited to describe cytoskeleton dynamics, potentially because of some scale-invariant dynamical properties in the system [18,61].

increased, which at low filament density happens first via a density-fluctuation instability, and at high filament density via an orientational-fluctuation instability. In the presence of a finite sorting rate between filaments of different orientations, propagating modes appear that reflect oscillatory behavior. In [149], the continuum theory is related to nonlocal descriptions of filament-motor systems, since filaments can transmit stresses over finite distances. The effective parameters of the continuum theory are recovered from the previously-published mesoscopic description [144], even though with missing coefficients that are thought to correspond to microsocpic multi-particle interactions, not described in [144]. Effects of polymerization-depolymarization dynamics via effective source and sink terms in the local filament densities are also discussed (see also [125,167]) – like it is the case in the effective macroscopic theories – as well as the role of polarity. In particular, it is found that nonpolar arrangements of filaments do not exhibit oscillatory instabilities and propagating modes, which might be of relevance for muscle sarcomeric structures. As seen previously, in these systems, spontaneous oscillations that have been observed correspond more to oscillatory instabilities of rigidly-coupled collective motors than to solitary-wave solutions, as they are found in systems of active polar filaments.

Microscopic theories present the advantage of being able in principle to give rise to a full description of a given system with arbitrary precision and specificity, and to take into account the nonlinear effects that are of direct relevance for the system's behavior. However, they rely on the microscopic knowledge that one has on the system under consideration, and are therefore limited by the available information on the different agents. In addition, they end up with effective descriptions that are model-dependent, in that the different parameters of the so-obtained theory, which describe its physical behavior, depend on the interactions that are taken into account at the microscopic level. Also, an important aspect of active cytoskeleton dynamics that is usually not described in such microscopic approaches is the very important phenomenon of treadmilling that relies on polymerization-depolymerization dynamics of cytoskeletal polymers, and which we have seen to be of crucial importance for some mechanisms of cellular motility such as *Listeria* propulsion or nematode-sperm-cell locomotion (see Sect. "Filament-Driven Motility"). However, despite the absence of these effects, which are taken into account effectively in macroscopic hydrodynamic descriptions, such microscopic approaches allow for the derivation of the forces exchanged between the motors and the filaments from microscopic knowledge, while they appear as effective parameters of unknown explicit origin in effective macroscopic descriptions. Thereby, questions can be addressed that concern the role played by the specific physical properties of motor-filament interactions at the microscopic level in controlling the system behavior on large scales. Indeed, the richness of the observed self-organized structures raises the question of how much is generic, and how much is specific in cytoskeleton behavior. For example, experiments have shown that very different self-organizing structures occur with processive as opposed to non-processive motor proteins: at high motor concentrations, microtubule-kinesin mixtures self-organize in a variety of spatial patterns [203,260], as homogeneous states are more robust with acto-myosin systems [106], an effect that can be thought of as the influence of motor processivity on the dynamical large-scale parameters [164].

Extensions and Future Directions

Cell motility is a complex and integrated process that relies on self-organization of the cytoskeleton, carefully and precisely orchestrated by the cell with the help of numerous different types of molecular players. If one includes the subcellular movements that are responsible for intracellular traffic and material exchange between the inner parts and external parts of the cell, cell-motility mechanisms are found to ground the activity of all life forms on earth. When looked under the microscope, motility mechanisms and structural changes of the diverse cell types appear so vast and various that a comprehensive understanding of their underlying mechanisms seems to be an overwhelming challenge. However, as we have seen from the literature covered in the present article, our understanding of cell motility has tremendously progressed over the past two decades. On the one hand, complexity has even further emerged, since the biochemical characterization of the molecular players involved has revealed that at least hundreds of different protein types participate in the structural and dynamical organization of the cell cytoskeleton. On the other hand, despite the existence of such very complex regulation processes that rely on the integrated interplay of the whole set of different molecular players, the characterization of the cell cytoskeleton has revealed that its main structures and functions are due to just a few types of key proteins, namely three types of biopolymers and three superfamilies of molecular motors. Even more striking is the evolutionary conservation of the main molecular players involved in building the cytoskeleton dynamical filaments, both within the eukaryotic domain of life on the direct sequence point of view, and even across the three domains of life when structural and func-

tional properties are considered. These striking observations indicate that the different underlying mechanisms of cell motility all rely on generic principles that can be understood on a biophysical point of view. In addition, further help from micro-manipulation and fluorescence-microscopy techniques, as well as the development of simplified systems based on gene-expression control and biomimetic artificial systems, has enabled the experimental biophysical investigation of the different specific aspects of the processes at play.

The theoretical analyses reviewed in this article have shown that central concepts that underly the cytoskeleton dynamics are self-organization and dynamic instabilities, here grounded on out-of-equilibrium nonlinear dynamics', thermodynamics' and statistical physics' principles. Such concepts are at the basis of all the theoretical approaches that have been developed to understand the mechanisms of diverse phenomena such as polymerization-depolymerization force and movement generation, molecular motors' individual behaviors and collective phenomena, as well as the generic behaviors of active-polymer solutions which lead to a description of the cytoskeleton dynamics as a whole. On all of these topics, microscopic as well coarse-grained effective macroscopic approaches have been developed. As already discussed, they both have their advantages, powers and limitations, and represent important complementary steps in the ultimate goal of an integrated description of the universal principles that underly cell motility.

As we have seen in this article, our understanding of cell motility and cell cytoskeleton dynamics has grandly benefitted from the interplay between experiments and modeling, each for its own reasons guiding the other in its directions of investigation. To further understand the integrated processes at play in cell motility, such fruitful interactions will certainly be further required and developed. On the theoretical point of view, bridges between understanding simplified systems or some particular aspects of cell motility and the phenomenon of cell motility as a whole at the global cellular level, have already started being investigated, but further developments of these two different ways of approaching the cytoskeleton dynamics as well as understanding the links that ultimately relate them are required. Another important aspect whose understanding represents a challenge is the potentially crucial role of noise that has been so far most of the time absent from the macroscopic effective theoretical approaches. Indeed, noise in nonlinear dynamical systems is known to potentially have important constructive effects, whose main representatives are stochastic resonance, coherence resonance and noise-induced transitions, as well as the extensive gallery of different spatially-extended phenomena such as array-enhanced stochastic and coherence resonance, or noise-enhanced synchronization of nonlinear oscillators (see e. g. [220]). Such phenomena have already been recognized to play an important role in some biological cytoskeleton-based pattern formations (see e. g. [105]), and could play a crucial role in driving other cytoskeletal self-organization phenomena, especially close to dynamical instabilities, where the effect of noise is highest. Finally, having at hand the underlying biochemical and biophysical mechanisms of cell forces and motility, a great challenge is to understand self-organization at yet larger scales, namely in animal tissues, where collections of cells present integrated coherent behaviors that drive diverse key processes such as morphogenesis, wound healing, immune response, tumor development and metastases formations. There, the same scheme involving "microscopic" as well as effective "macroscopic" approaches can certainly play an equally major role, "microscopic" approaches then potentially integrating the whole knowledge acquired at the level of a single cell, and partially reviewed in this article.

Acknowledgments

To write this article, I grandly benefited from Karsten Kruse's habilitation thesis, which constituted a very good starting point as an extensive review of the existing biophysical literature on the cytoskeleton. I acknowledge Cécile Sykes and Jean-François Joanny for discussions, careful reading of the manuscript and constructive criticisms and suggestions. I also acknowledge Andrew Callan-Jones for pointing to me important references.

Bibliography

Primary Literature

1. Abercrombie M (1980) The crawling movement of metazoan cells. Proc R Soc Lond B Biol Sci 108:387–393
2. Ahmadi A, Liverpool TB, Marchetti MC (2005) Nematic and polar order in active filament solutions. Phys Rev E Stat Nonlin Soft Matter Phys 72(6 Pt 1):060901
3. Ahmadi A, Marchetti MC, Liverpool TB (2006) Hydrodynamics of isotropic and liquid crystalline active polymer solutions. Phys Rev E Stat Nonlin Soft Matter Phys 74(6 Pt 1):061913
4. Ajdari A (1995) Transport by active filaments. Europhys Lett 31(2):69–74
5. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002) Molecular biology of the cell, 4th edn. Garland, New York
6. Albertson DG (1984) Formation of the first cleavage spindle in nematode embryos. Dev Biol 101(1):61–72
7. Alt W,Dembo M (1999) Cytoplasm dynamics and cell motion: two-phase flow models. Math Biosci 156(1–2):207–228

8. Amos LA, van den Ent F, Lowe J (2004) Structural/functional homology between the bacterial and eukaryotic cytoskeletons. Curr Opin Cell Biol 16(1):24–31
9. Ananthakrishnan R, Ehrlicher A (2007) The forces behind cell movement. Int J Biol Sci 3(5):303–317
10. Andersen SS (2000) Spindle assembly and the art of regulating microtubule dynamics by MAPs and Stathmin/OP18. Trends Cell Biol 10(7):261–267
11. Aranson IS, Tsimring LS (2005) Pattern formation of microtubules and motors: Inelastic interaction of polar rods. Phys Rev E Stat Nonlin Soft Matter Phys 71(5 Pt 1):050901
12. Aranson IS, Tsimring LS (2006) Theory of self-assembly of microtubules and motors. Phys Rev E Stat Nonlin Soft Matter Phys 74(3 Pt 1):031915
13. Astumian RD (1997) Thermodynamics and kinetics of a brownian motor. Science 276(5314):917–922
14. Astumian RD, Bier M (1994) Fluctuation driven ratchets: Molecular motors. Phys Rev Lett 72(11):1766–1769
15. Astumian RD, Bier M (1996) Mechanochemical coupling of the motion of molecular motors to ATP hydrolysis. Biophys J 70(2):637–653
16. Ausmees N, Kuhn JR, Jacobs-Wagner C (2003) The bacterial cytoskeleton: an intermediate filament-like function in cell shape. Cell 115(6):705–713
17. Badoual M, Jülicher F, Prost J (2002) Bidirectional cooperative motion of molecular motors. Proc Natl Acad Sci USA 99(10):6696–6701
18. Balland M, Desprat N, Icard D, Fereol S, Asnacios A, Browaeys J, Henon S, Gallet F (2006) Power laws in microrheology experiments on living cells: Comparative analysis and modeling. Phys Rev E Stat Nonlin Soft Matter Phys 74(2 Pt 1):021911
19. Belmont LD, Hyman AA, Sawin KE, Mitchison TJ (1990) Real-time visualization of cell cycle-dependent changes in microtubule dynamics in cytoplasmic extracts. Cell 62(3):579–589
20. Berg HC (2003) The rotary motor of bacterial flagella. Annu Rev Biochem 72:19–54
21. Berg JS, Powell BC, Cheney RE (2001) A millennial myosin census. Mol Biol Cell 12(4):780–794
22. Bernheim-Groswasser A, Wiesner S, Golsteyn RM, Carlier M-F, Sykes C (2002) The dynamics of actin-based motility depend on surface parameters. Nature 417(6886):308–311
23. Bernheim-Groswasser A, Prost J, Sykes C (2005) Mechanism of actin-based motility: a dynamic state diagram. Biophys J 89(2):1411–1419
24. Bershadsky A, Kozlov M, Geiger B (2006) Adhesion-mediated mechanosensitivity: a time to experiment, and a time to theorize. Curr Opin Cell Biol 18(5):472–481
25. Bishop AL, Hall A (2000) Rho GTPases and their effector proteins. Biochem J 348(Pt 2):241–255
26. Bork P, Sander C, Valencia A (1992) An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins. Proc Natl Acad Sci USA 89(16):7290–7294
27. Bottino D, Mogilner A, Roberts T, Stewart M, Oster G (2002) How nematode sperm crawl. J Cell Sci 115(Pt 2):367–384
28. Boukellal H, Campas O, Joanny J-F, Prost J, Sykes C (2004) Soft listeria: Actin-based propulsion of liquid drops. Phys Rev E Stat Nonlin Soft Matter Phys 69(6 Pt 1):061906
29. Bray D (2001) Cell movements, 2nd edn. Garland, New York
30. Brokaw CJ (1975) Molecular mechanism for oscillation in flagella and muscle. Proc Natl Acad Sci USA 72(8):3102–3106
31. Bullock TL, Roberts TM, Stewart M (1996) 2.5 A resolution crystal structure of the motile major sperm protein (MSP) of Ascaris suum. J Mol Biol 263(2):284–296
32. Bustamante C, Macosko JC, Wuite GJ (2000) Grabbing the cat by the tail: manipulating molecules one by one. Nat Rev Mol Cell Biol 1(2):130–136
33. Camalet S, Jülicher F (2000) Generic aspects of axonemal beating. New J Phys 2:1–23
34. Camalet S, Jülicher F, Prost J (1999) Self-organized beating and swimming of internally driven filaments. Phys Rev Lett 82(7):1590–1593
35. Cameron LA, Footer MJ, van Oudenaarden A, Theriot JA (1999) Motility of ActA protein-coated microspheres driven by actin polymerization. Proc Natl Acad Sci USA 96(9): 4908–4913
36. Cameron LA, Giardini PA, Soo FS, Theriot JA (2000) Secrets of actin-based motility revealed by a bacterial pathogen. Nat Rev Mol Cell Biol 1(2):110–119
37. Carlsson AE (2001) Growth of branched actin networks against obstacles. Biophys J 81(4):1907–1923
38. Carlsson AE (2003) Growth velocities of branched actin networks. Biophys J 84(5):2907–2918
39. Chaen S, Inoue J, Sugi H (1995) The force-velocity relationship of the ATP-dependent actin-myosin sliding causing cytoplasmic streaming in algal cells, studied using a centrifuge microscope. J Exp Biol 198(Pt 4):1021–1027
40. Chowdhury D, Schadschneider A, Nishinari K (2005) Physics of transport and traffic phenomena in biology from molecular motors and cells to organisms. Phys Life Rev 2(4):318–352
41. Cossart P, Bierne H (2001) The use of host cell machinery in the pathogenesis of Listeria monocytogenes. Curr Opin Immunol 13(1):96–103
42. Coussen F, Choquet D, Sheetz MP, Erickson HP (2002) Trimers of the fibronectin cell adhesion domain localize to actin filament bundles and undergo rearward translocation. J Cell Sci 115(Pt 12):2581–2590
43. de Boer P, Crossley R, Rothfield L (1992) The essential bacterial cell-division protein FtsZ is a GTPase. Nature 359(6392): 254–256
44. de Gennes P-G, Prost J (1993) The physics of liquid crystals, 2nd edn. Clarendon Press, Oxford
45. de Groot SR, Mazur P (1984) Non-equilibrium thermodynamics. Dover Publications Inc, New York
46. Derrida B (1983) Velocity and diffusion constant of a periodic one-dimensional hopping model. J Stat Phys 31(3):433
47. Desai A, Verma S, Mitchison TJ, Walczak CE (1999) Kin I kinesins are microtubule-destabilizing enzymes. Cell 96(1): 69–78
48. Dickinson RB, Purich DL (2007) Nematode sperm motility: nonpolar filament polymerization mediated by end-tracking motors. Biophys J 92(2):622–631
49. Dickinson RB, Caro L, Purich DL (2004) Force generation by cytoskeletal filament end-tracking proteins. Biophys J 87(4):2838–2854
50. Dogterom M, Yurke B (1997) Measurement of the force-velocity relation for growing microtubules. Science 278(5339): 856–860
51. Dogterom M, Kerssemakers JWJ, Romet-Lemonne G, Janson ME (2005) Force generation by dynamic microtubules. Curr Opin Cell Biol 17(1):67–74
52. Dombrowski C, Cisneros L, Chatkaew S, Goldstein RE, Kessler

JO (2004) Self-concentration and large-scale coherence in bacterial dynamics. Phys Rev Lett 93(9):098103
53. dos Remedios CG, Chhabra D, Kekic M, Dedova IV, Tsubakihara M, Berry DA, Nosworthy NJ (2003) Actin binding proteins: Regulation of cytoskeletal microfilaments. Physiol Rev 83(2):433–473
54. Duke T, Leibler S (1996) Motor protein mechanics: a stochastic model with minimal mechanochemical coupling. Biophys J 71(3):1235–1247
55. Endow SA Higuchi H (2000) A mutant of the motor protein kinesin that moves in both directions on microtubules. Nature 406(6798):913–916
56. Engler AJ, Sen S, Sweeney HL, Discher DE (2006) Matrix elasticity directs stem cell lineage specification. Cell 126(4): 677–689
57. Erickson HP (2007) Evolution of the cytoskeleton. Bioessays 29(7):668–677
58. Erickson HP, Taylor DW, Taylor KA, Bramhill D (1996) Bacterial cell division protein FtsZ assembles into protofilament sheets and minirings, structural homologs of tubulin polymers. Proc Natl Acad Sci USA 93(1):519–523
59. Euteneuer U, Schliwa M (1984) Persistent, directional motility of cells and cytoplasmic fragments in the absence of microtubules. Nature 310(5972):58–61
60. Evans MR, Foster DP, Godreche C, Mukamel D (1995) Spontaneous symmetry breaking in a one dimensional driven diffusive system. Phys Rev Lett 74(2):208–211
61. Fabry B, Maksym GN, Butler JP, Glogauer M, Navajas D, Fredberg JJ (2001) Scaling the microrheology of living cells. Phys Rev Lett 87(14):148102
62. Faivre-Moskalenko C, Dogterom M (2002) Dynamics of microtubule asters in microfabricated chambers: The role of catastrophes. Proc Natl Acad Sci USA 99(26):16788–16793
63. Fernandez P, Pullarkat PA, Ott A (2006) A master relation defines the nonlinear viscoelasticity of single fibroblasts. Biophys J 90(10):3796–3805
64. Feynman RP, Leighton RB, Sands M (1963) The Feynman lectures on physics, vol 1. Addison-Wesley, Reading
65. Fisher ME (1998) Renormalization group theory: Its basis and formulation in statistical physics. Rev Mod Phys 70(2): 653–681
66. Fisher ME, Kolomeisky AB (1999) The force exerted by a molecular motor. Proc Natl Acad Sci USA 96(12):6597–6602
67. Fletcher DA, Theriot JA (2004) An introduction to cell motility for the physical scientist. Phys Biol 1(1–2):T1–10
68. Footer MJ, Kerssemakers JWJ, Theriot JA, Dogterom M (2007) Direct measurement of force generation by actin filament polymerization using an optical trap. PNAS 104(7):2181–2186
69. Forterre Y, Skotheim JM, Dumais J, Mahadevan L (2005) How the venus flytrap snaps. Nature 433(7024):421–425
70. Frey E, Täuber UC, Cauber U (1994) 2-loop renormalization-group analysis of the Burgers-Kardar-Parisi-Zhang equation. Phys Rev E 50(2):1024–1045
71. Frischknecht F, Way M (2001) Surfing pathogens and the lessons learned for actin polymerization. Trends Cell Biol 11(1):30–38
72. Fujita H, Ishiwata S (1998) Spontaneous oscillatory contraction without regulatory proteins in actin filament-reconstituted fibers. Biophys J 75(3):1439–1445
73. Fukui Y (2002) Mechanistics of amoeboid locomotion: signal to forces. Cell Biol Int 26(11):933–944
74. Fygenson DK, Marko JF, Libchaber A (1997) Mechanics of microtubule-based membrane extension. Phys Rev Lett 79(22):4497–4500
75. Gardel ML, Shin JH, MacKintosh FC, Mahadevan L, Matsudaira PA, Weitz DA (2004) Elastic behavior of cross-linked and bundled actin networks. Science 304(5675):1301–1305
76. Gardel ML, Shin JH, MacKintosh FC, Mahadevan L, Matsudaira PA, Weitz DA (2004) Scaling of F-actin network rheology to probe single filament elasticity and dynamics. Phys Rev Lett 93(18):188102
77. Gardel ML, Nakamura F, Hartwig JH, Crocker JC, Stossel TP, Weitz DA (2006) Prestressed F-actin networks cross-linked by hinged filamins replicate mechanical properties of cells. Proc Natl Acad Sci USA 103(6):1762–1767
78. Geiger B, Bershadsky A, Pankov R, Yamada KM (2001) Transmembrane crosstalk between the extracellular matrix–cytoskeleton crosstalk. Nat Rev Mol Cell Biol 2(11):793–805
79. Gerbal F, Chaikin P, Rabin Y, Prost J (2000) An elastic analysis of Listeria monocytogenes propulsion. Biophys J 79(5): 2259–2275
80. Giardini PA, Fletcher DA, Theriot JA (2003) Compression forces generated by actin comet tails on lipid vesicles. Proc Natl Acad Sci USA 100(11):6493–6498
81. Gittes F, Mickey B, Nettleton J, Howard J (1993) Flexural rigidity of microtubules and actin filaments measured from thermal fluctuations in shape. J Cell Biol 120(4):923–934
82. Glotzer M (2001) Animal cell cytokinesis. Annu Rev Cell Dev Biol 17:351–386
83. Goldberg MB (2001) Actin-based motility of intracellular microbial pathogens. Microbiol Mol Biol Rev 65(4):595–626
84. Granger C, Cyr R (2001) Use of abnormal preprophase bands to decipher division plane determination. J Cell Sci 114(Pt 3): 599–607
85. Grill SW, Gonczy P, Stelzer EH, Hyman AA (2001) Polarity controls forces governing asymmetric spindle positioning in the caenorhabditis elegans embryo. Nature 409(6820):630–633
86. Grill SW, Kruse K, Jülicher F (2005) Theory of mitotic spindle oscillations. Phys Rev Lett 94(10):108104
87. Gruenheid S, Finlay BB (2003) Microbial pathogenesis and cytoskeletal function. Nature 422(6933):775–781
88. Gueron S, Levit-Gurevich K, Liron N, Blum JJ (1997)Cilia internal mechanism and metachronal coordination as the result of hydrodynamical coupling. Proc Natl Acad Sci USA 94(12):6001–6006
89. Guirao B, Joanny J-F (2007) Spontaneous creation of macroscopic flow and metachronal waves in an array of cilia. Biophys J 92(6):1900–1917
90. Gupton SL, Waterman-Storer CM (2006) Spatiotemporal feedback between actomyosin and focal-adhesion systems optimizes rapid cell migration. Cell 125(7):1361–1374
91. Hall A (2005) Rho GTPases and the control of cell behaviour. Biochem Soc Trans 33(Pt 5):891–895
92. Hanney T, Evans MR (2004) Condensation transitions in a two-species zero-range process. Phys Rev E Stat Nonlin Soft Matter Phys 69(1 Pt 2):016107
93. Hanson J, Huxley HE (1953) Structural basis of the cross-striations in muscle. Nature 172(4377):530–532
94. Hatwalne Y, Ramaswamy S, Rao M, Simha RA (2004) Rheology of active-particle suspensions. Phys Rev Lett 92(11):118101
95. Hayden JH, Bowser SS, Rieder CL (1990) Kinetochores capture astral microtubules during chromosome attachment to

the mitotic spindle: Direct visualization in live newt lung cells. J Cell Biol 111(3):1039–1045
96. Heintzelman MB (2003) Gliding motility: The molecules behind the motion. Curr Biol 13(2):R57–9
97. Hill TL (1974) Theoretical formalism for the sliding filament model of contraction of striated muscle, part I. Prog Biophys Mol Biol 28:267–340
98. Hinrichsen H (2006) Non-equilibrium phase transitions. Physica A 369(1):1–28
99. Hoffman BD, Massiera G, Van Citters KM, Crocker JC (2006) The consensus mechanics of cultured mammalian cells. Proc Natl Acad Sci USA 103(27):10259–10264
100. Hohenberg PC, Halperin BI (1977) Theory of dynamic critical phenomena. Rev Mod Phys 49(3):435–479
101. Holy TE, Dogterom M, Yurke B, Leibler S (1997) Assembly and positioning of microtubule asters in microfabricated chambers. Proc Natl Acad Sci USA 94(12):6228–6231
102. Howard J (1997) Molecular motors: Structural adaptations to cellular functions. Nature 389(6651):561–567
103. Howard J (2001) Mechanics of Motor Proteins and the Cytoskeleton. Sinauer Associates Inc, Sunderland
104. Howard J, Hyman AA (2003) Dynamics and mechanics of the microtubule plus end. Nature 422(6933):753–758
105. Howard M, Rutenberg AD (2003) Pattern formation inside bacteria: Fluctuations due to the low copy number of proteins. Phys Rev Lett 90(12):128102
106. Humphrey D, Duggan C, Saha D, Smith D, Käs J (2002) Active fluidization of polymer networks through molecular motors. Nature 416(6879):413–416
107. Hunter AW, Caplow M, Coy DL, Hancock WO, Diez S, Wordeman L, Howard J (2003) The kinesin-related protein MCAK is a microtubule depolymerase that forms an ATP-hydrolyzing complex at microtubule ends. Mol Cell 11(2):445–457
108. Huxley AF (1957) Muscle structure and theories of contraction. Prog Biophys Biophys Chem 7:255–318
109. Huxley AF, Niedergerke R (1954) Structural changes in muscle during contraction; interference microscopy of living muscle fibres. Nature 173(4412):971–973
110. Huxley HE (1953) Electron microscope studies of the organisation of the filaments in striated muscle. Biochim Biophys Acta 12(3):387–394
111. Huxley HE (1953) X-ray analysis and the problem of muscle. Proc R Soc Lond B Biol Sci 141(902):59–62
112. Huxley HE (1957) The double array of filaments in cross-striated muscle. J Biophys Biochem Cytol 3(5):631–648
113. Huxley HE (1996) A personal view of muscle and motility mechanisms. Annu Rev Physiol 58:1–19
114. Huxley HE (2004) Fifty years of muscle and the sliding filament hypothesis. Eur J Biochem 271(8):1403–1415
115. Huxley HE, Hanson J (1954) Changes in the cross-striations of muscle during contraction and stretch and their structural interpretation. Nature 173(4412):973–976
116. Hyman AA, Karsenti E (1996) Morphogenetic properties of microtubules and mitotic spindle assembly. Cell 84(3):401–410
117. Ibanez-Tallon I, Heintz N, Omran H (2003) To beat or not to beat: roles of cilia in development and disease. Hum Mol Genet 12(Spec no 1):R27–35
118. Inoue S, Salmon ED (1995) Force generation by microtubule assembly/disassembly in mitosis and related movements. Mol Biol Cell 6(12):1619–1640
119. Isambert H, Venier P, Maggs AC, Fattoum A, Kassab R, Pantaloni D, Carlier MF (1995) Flexibility of actin filaments derived from thermal fluctuations. Effect of bound nucleotide, phalloidin, and muscle regulatory proteins. J Biol Chem 270(19):11437–11444 (1995)
120. Italiano JE Jr, Roberts TM, Stewart M, Fontana CA (1996) Reconstitution in vitro of the motile apparatus from the amoeboid sperm of ascaris shows that filament assembly and bundling move membranes. Cell 84(1):105–114
121. Italiano JE Jr, Stewart M, Roberts TM (1999) Localized depolymerization of the major sperm protein cytoskeleton correlates with the forward movement of the cell body in the amoeboid movement of nematode sperm. J Cell Biol 146(5):1087–1096
122. Janmey PA, Euteneuer U, Traub P, Schliwa M (1991) Viscoelastic properties of vimentin compared with other filamentous biopolymer networks. J Cell Biol 113(1):155–160
123. Janson ME, Dogterom M (2004) Scaling of microtubule force-velocity curves obtained at different tubulin concentrations. Phys Rev Lett 92(24):248101
124. Joanny J-F, Jülicher F, Prost J (2003) Motion of an adhesive gel in a swelling gradient: A mechanism for cell locomotion. Phys Rev Lett 90(16):168102
125. Joanny J-F, Jülicher F, Kruse K, Prost J (2007) Hydrodynamic theory for multi-component active polar gels. New J Phys 9:422
126. Jones LJ, Carballido-Lopez R, Errington J (2001) Control of cell shape in bacteria: Helical, actin-like filaments in bacillus subtilis. Cell 104(6):913–922
127. Jülicher F, Prost J (1995) Cooperative molecular motors. Phys Rev Lett 75(13):2618–2621
128. Jülicher F, Prost J (1997) Spontaneous oscillations of collective molecular motors. Phys Rev Lett 78(23):4510–4513
129. Jülicher F, Ajdari A, Prost J (1997) Modeling molecular motors. Rev Mod Phys 69(4):1269–1281
130. Jülicher F, Kruse K, Prost J, Joanny JF (2007) Active behavior of the cytoskeleton. Phys Rep 449(1–3):3–28
131. Jurado C, Haserick JR, Lee J (2005) Slipping or gripping? fluorescent speckle microscopy in fish keratocytes reveals two different mechanisms for generating a retrograde flow of actin. Mol Biol Cell 16(2):507–518
132. Kaksonen M, Sun Y, Drubin DG (2003) A pathway for association of receptors, adaptors, and actin during endocytic internalization. Cell 115(4):475–487
133. Karki S, Holzbaur EL (1999) Cytoplasmic dynein and dynactin in cell division and intracellular transport. Curr Opin Cell Biol 11(1):45–53
134. Karsenti E, Vernos I (2001) The mitotic spindle: A self-made machine. Science 294(5542):543–547
135. Kasza KE, Rowat AC, Liu J, Angelini TE, Brangwynne CP, Koenderink GH, Weitz DA (2007) The cell as a material. Curr Opin Cell Biol 19(1):101–107
136. Kaverina I, Krylyshkina O, Small JV (2002) Regulation of substrate adhesion dynamics during cell motility. Int J Biochem Cell Biol 34(7):746–761
137. King KL, Essig J, Roberts TM, Moerland TS (1994) Regulation of the ascaris major sperm protein (MSP) cytoskeleton by intracellular pH. Cell Motil Cytoskeleton 27(3):193–205
138. Klumpp S, Lipowsky R (2004) Asymmetric simple exclusion processes with diffusive bottlenecks. Phys Rev E Stat Nonlin Soft Matter Phys 70(6 Pt 2):066104

139. Klumpp S, Lipowsky R (2004) Phase transitions in systems with two species of molecular motors. Europhys Lett 66(1):90–96
140. Kolomeisky AB, Fisher ME (2001) Force-velocity relation for growing microtubules. Biophys J 80(1):149–154
141. Kondo S (2002) The reaction-diffusion system: A mechanism for autonomous pattern formation in the animal skin. Genes Cells 7(6):535–541
142. Kozlov MM, Mogilner A (2007) Model of polarization and bistability of cell fragments. Biophys J 93(11):3811–3819
143. Krendel M, Zenke FT, Bokoch GM (2002) Nucleotide exchange factor GEF-H1 mediates cross-talk between microtubules and the actin cytoskeleton. Nat Cell Biol 4(4):294–301
144. Kruse K, Jülicher F (2000) Actively contracting bundles of polar filaments. Phys Rev Lett 85(8):1778–1781
145. Kruse K, Jülicher F (2003) Self-organization and mechanical properties of active filament bundles. Phys Rev E Stat Nonlin Soft Matter Phys 67(5 Pt 1):051913
146. Kruse K, Jülicher F (2006) Dynamics and mechanics of motor-filament systems. Eur Phys J E Soft Matter 20(4):459–465
147. Kruse K, Sekimoto K (2002) Growth of fingerlike protrusions driven by molecular motors. Phys Rev E Stat Nonlin Soft Matter Phys 66(3 Pt 1):031904
148. Kruse K, Camalet S, Jülicher F (2001) Self-propagating patterns in active filament bundles. Phys Rev Lett 87(13):138101
149. Kruse K, Zumdieck A, Jülicher F (2003) Continuum theory of contractile fibres. Europhys Lett 64(5):716–722
150. Kruse K, Joanny JF, Jülicher F, Prost J, Sekimoto K (2004) Asters, vortices, and rotating spirals in active gels of polar filaments. Phys Rev Lett 92(7):078101
151. Kruse K, Joanny JF, Jülicher F, Prost J, Sekimoto K (2005) Generic theory of active polar gels: A paradigm for cytoskeletal dynamics. Eur Phys J E Soft Matter 16(1):5–16
152. Kruse K, Joanny JF, Jülicher F, Prost J (2006) Contractility and retrograde flow in lamellipodium motion. Phys Biol 3(2): 130–137
153. Lacayo CI, Theriot JA (2004) Listeria monocytogenes actin-based motility varies depending on subcellular location: A kinematic probe for cytoarchitecture. Mol Biol Cell 15(5):2164–2175
154. Landau LD, Lifschitz EM (1995) The theory of elasticity. Butterworth-Heinemann, Boston
155. Lasa I, Gouin E, Goethals M, Vancompernolle K, David V, Vandekerckhove J, Cossart P (1997) Identification of two regions in the N-terminal domain of ActA involved in the actin comet tail formation by Listeria monocytogenes. EMBOJ 16(7): 1531–1540
156. Lauffenburger DA, Horwitz AF (1996) Cell migration: A physically integrated molecular process. Cell 84(3):359–369
157. Lee HY, Kardar M (2001) Macroscopic equations for pattern formation in mixtures of microtubules and molecular motors. Phys Rev E Stat Nonlin Soft Matter Phys 64(5 Pt 2):056113
158. Lee J, Ishihara A, Jacobson K (1993) The fish epidermal keratocyte as a model system for the study of cell locomotion. Symp Soc Exp Biol 47:73–89
159. Leibler S, Huse DA (1993) Porters versus rowers: A unified stochastic model of motor proteins. J Cell Biol 121(6): 1357–1368
160. Lin CH, Espreafico EM, Mooseker MS, Forscher P (1996) Myosin drives retrograde F-actin flow in neuronal growth cones. Neuron 16(4):769–782
161. Lipowsky R, Klumpp S, Nieuwenhuizen TM (2001) Random walks of cytoskeletal motors in open and closed compartments. Phys Rev Lett 87(10):108101
162. Liverpool TB (2003) Anomalous fluctuations of active polar filaments. Phys Rev E Stat Nonlin Soft Matter Phys 67(3 Pt 1):031909
163. Liverpool TB (2006) Active gels: Where polymer physics meets cytoskeletal dynamics. Philos Trans A Math Phys Eng Sci 364(1849):3335–3355
164. Liverpool TB, Marchetti MC (2003) Instabilities of isotropic solutions of active polar filaments. Phys Rev Lett 90(13):138102
165. Liverpool TB, Marchetti MC (2005) Bridging the microscopic and the hydrodynamic in active filament solutions. Europhys Lett 69(5):846–852
166. Liverpool TB, Marchetti MC (2006) Rheology of active filament solutions. Phys Rev Lett 97(26):268101
167. Liverpool TB, Marchetti MC (2008) Hydrodynamics and rheology of active polar filaments. In: Lenz P (ed) Cell motility. Springer, New York
168. Liverpool TB, Maggs AC, Ajdari A (2001) Viscoelasticity of solutions of motile polymers. Phys Rev Lett 86(18):4171–4174
169. Lo CM, Wang HB, Dembo M, Wang YL (2000) Cell movement is guided by the rigidity of the substrate. Biophys J 79(1): 144–152
170. Loisel TP, Boujemaa R, Pantaloni D, Carlier MF (1999) Reconstitution of actin-based motility of Listeria and Shigella using pure proteins. Nature 401(6753):613–616
171. Lowe J, Amos LA (1998) Crystal structure of the bacterial cell-division protein FtsZ. Nature 391(6663):203–206
172. Lymn RW, Taylor EW (1971) Mechanism of adenosine triphosphate hydrolysis by actomyosin. Biochemistry 10(25): 4617–4624
173. MacDonald CT, Gibbs JH, Pipkin AC (1968) Kinetics of biopolymerization on nucleic acid templates. Biopolymers 6(1):1–5
174. Machesky LM, Insall RH (1999) Signaling to actin dynamics. J Cell Biol 146(2):267–272
175. Machesky LM, Mullins RD, Higgs HN, Kaiser DA, Blanchoin L, May RC, Hall ME, Pollard TD (1999) Scar, a WASp-related protein, activates nucleation of actin filaments by the Arp2/3 complex. Proc Natl Acad Sci USA 96(7):3739–3744
176. Machin KE (1958) Wave propagation along flagella. J Exp Biol 35:796
177. Magnasco MO (1993) Forced thermal ratchets. Phys Rev Lett 71(10):1477–1481
178. Magnasco MO (1994) Molecular combustion motors. Phys Rev Lett 72(16):2656–2659
179. Mahadevan L, Matsudaira P (2000) Motility powered by supramolecular springs and ratchets. Science 288(5463): 95–100
180. Maney T, Wagenbach M, Wordeman L (2001) Molecular dissection of the microtubule depolymerizing activity of mitotic centromere-associated kinesin. J Biol Chem 276(37): 34753–34758
181. Manson MD, Armitage JP, Hoch JA, Macnab RM (1998) Bacterial locomotion and signal transduction. J Bacteriol 180(5):1009–1022
182. Marcy Y, Prost J, Carlier M-F, Sykes C (2004) Forces generated during actin-based propulsion: A direct measurement by micromanipulation. Proc Natl Acad Sci USA 101(16):5992–5997

183. Martin PC, Parodi O, Pershan P (1972) Unified hydrodynamic theory for crystals, liquid crystals and normal fluids. Phys Rev A 6:2401
184. McGrath JL, Eungdamrong NJ, Fisher CI, Peng F, Mahadevan L, Mitchison TJ, Kuo SC (2003) The force-velocity relationship for the actin-based motility of Listeria monocytogenes. Curr Biol 13(4):329–332
185. McGrath JL, Somlo S, Makova S, Tian X, Brueckner M (2003) Two populations of node monocilia initiate left-right asymmetry in the mouse. Cell 114(1):61–73
186. Merrifield CJ, Moss SE, Ballestrem C, Imhof BA, Giese G, Wunderlich I, Almers W (1999) Endocytic vesicles move at the tips of actin tails in cultured mast cells. Nat Cell Biol 1(1):72–74
187. Miao L, Vanderlinde O, Stewart M, Roberts TM (2003) Retraction in amoeboid cell motility powered by cytoskeletal dynamics. Science 302(5649):1405–1407
188. Michie KA, Lowe J (2006) Dynamic filaments of the bacterial cytoskeleton. Annu Rev Biochem 75:467–492
189. Miki H, Setou M, Kaneshiro K, Hirokawa N (2001) All kinesin superfamily protein, KIF, genes in mouse and human. Proc Natl Acad Sci USA 98(13):7004–7011
190. Miranti CK, Brugge JS (2002) Sensing the environment: A historical perspective on integrin signal transduction. Nat Cell Biol 4(4):E83–90
191. Mitchison TJ, Kirschner M (1984) Dynamic instability of microtubule growth. Nature 312(5991):237–242
192. Mitchison TJ, Cramer LP (1996) Actin-based cell motility and cell locomotion. Cell 84(3):371–379
193. Mitchison TJ, Salmon ED (2001) Mitosis: A history of division. Nat Cell Biol 3(1):E17–21
194. Mogilner A (2006) On the edge: modeling protrusion. Curr Opin Cell Biol 18(1):32–39
195. Mogilner A, Oster G (1996) Cell motility driven by actin polymerization. Biophys J 71(6):3030–3045
196. Mogilner A, Oster G (2003) Cell biology. Shrinking gels pull cells. Science 302(5649):1340–1341
197. Mogilner A, Oster G (2003) Force generation by actin polymerization II: The elastic ratchet and tethered filaments. Biophys J 84(3):1591–1605
198. Mogilner A, Oster G (2003) Polymer motors: Pushing out the front and pulling up the back. Curr Biol 13(18):R721–33
199. Mohanty PK, Kruse K (2007) Driven diffusive systems of active filament bundles. J Stat Phys 128(1–2):95–110
200. Mukherjee A, Dai K, Lutkenhaus J (1993) Escherichia coli cell division protein FtsZ is a guanine nucleotide binding protein. Proc Natl Acad Sci USA 90(3):1053–1057
201. Nakazawa H, Sekimoto K (1996) Polarity sorting in a bundle of actin filaments by two-headed myosins. J Phys Soc Jpn 65(8):2404–2407
202. Nedelec F, Surrey T, Maggs AC (2001) Dynamic concentration of motors in microtubule arrays. Phys Rev Lett 86(14): 3192–3195
203. Nedelec FJ, Surrey T, Maggs AC, Leibler S (1997) Self-organization of microtubules and motors. Nature 389(6648):305–308
204. Nelson WJ (2003) Adaptation of core mechanisms to generate cell polarity. Nature 422(6933):766–774
205. Nieuwenhuizen TM, Klumpp S, Lipowsky R (2004) Random walks of molecular motors arising from diffusional encounters with immobilized filaments. Phys Rev E 69(6):061911
206. Nogales E, Wolf SG, Downing KH (1998) Structure of the alpha beta tubulin dimer by electron crystallography. Nature 391(6663):199–203
207. Noireaux V, Golsteyn RM, Friederich E, Prost J, Antony C, Louvard D, Sykes C (2000) Growing an actin gel on spherical surfaces. Biophys J 78(3):1643–1654
208. Okada Y, Takeda S, Tanaka Y, Izpisua Belmonte J-C, Hirokawa N (2005) Mechanism of nodal flow: A conserved symmetry breaking event in left-right axis determination. Cell 121(4):633–644
209. Okamura N, Ishiwata S (1988) Spontaneous oscillatory contraction of sarcomeres in skeletal myofibrils. J Muscle Res Cell Motil 9(2):111–119
210. Ott A, Magnasco M, Simon A, Libchaber A (1993) Measurement of the persistence length of polymerized actin using fluorescence microscopy. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 48(3):R1642–R1645
211. Paluch E, Sykes C, Prost J, Bornens M (2006) Dynamic modes of the cortical actomyosin gel during cell locomotion and division. Trends Cell Biol 16(1):5–10
212. Paluch E, van der Gucht J, Joanny J-F, Sykes C (2006) Deformations in actin comets from rocketing beads. Biophys J 91(8):3113–3122
213. Pampaloni F, Lattanzi G, Jonas A, Surrey T, Frey E, Florin E-L (2006) Thermal fluctuations of grafted microtubules provide evidence of a length-dependent persistence length. Proc Natl Acad Sci USA 103(27):10248–10253
214. Pantaloni D, Le Clainche C, Carlier MF (2001) Mechanism of actin-based motility. Science 292(5521):1502–1506
215. Parent CA, Devreotes PN (1999) A cell's sense of direction. Science 284(5415):765–770
216. Parmeggiani A, Franosch T, Frey E (2003) Phase coexistence in driven one-dimensional transport. Phys Rev Lett 90(8):086601
217. Parmeggiani A, Jülicher F, Ajdari A, Prost J (1999) Energy transduction of isothermal ratchets: Generic aspects and specific examples close to and far from equilibrium. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 60(2 Pt B):2127–2140
218. Peskin CS, Ermentrout GB, Oster GF (1994) Cell mechanics and cellular engineering. Springer, New York
219. Peskin CS, Odell GM, Oster GF (1993) Cellular motions and thermal fluctuations: the brownian ratchet. Biophys J 65(1):316–324
220. Pikovsky A, Rosenblum M, Kurths J (2001) Synchronization – A unified approach to nonlinear science. Cambridge University Press, Cambridge
221. Plastino J, Lelidis I, Prost J, Sykes C (2004) The effect of diffusion, depolymerization and nucleation promoting factors on actin gel growth. Eur Biophys J 33(4):310–320
222. Plastino J, Sykes C (2005) The actin slingshot. Curr Opin Cell Biol 17(1):62–66
223. Pollard TD (2003) The cytoskeleton, cellular motility and the reductionist agenda. Nature 422(6933):741–5
224. Pollard TD, Blanchoin L, Mullins RD (2000) Molecular mechanisms controlling actin filament dynamics in nonmuscle cells. Annu Rev Biophys Biomol Struct 29:545–576
225. Pollard TD, Borisy GG (2003) Cellular motility driven by assembly and disassembly of actin filaments. Cell 112(4):453–465
226. Porter ME (1996) Axonemal dyneins: Assembly, organization, and regulation. Curr Opin Cell Biol 8(1):10–17

227. Prost J, Chauwin JF, Peliti L, Ajdari A (1994) Asymmetric pumping of particles. Phys Rev Lett 72(16):2652–2655
228. Purcell EM (1977) Life at low reynolds number. Am J Phys 45:3
229. Raftopoulou M, Hall A (2004) Cell migration: Rho GTPases lead the way. Dev Biol 265(1):23–32
230. Ramaswamy S, Simha RA, Toner J (2003) Active nematics on a substrate: Giant number fluctuations and long-time tails. Europhys Lett 62(2):196–202 Apr
231. Ramaswamy S, Toner J, Prost J (2000) Nonequilibrium fluctuations, traveling waves, and instabilities in active membranes. Phys Rev Lett 84(15):3494–3497
232. Raychaudhuri D, Park JT (1992) Escherichia coli cell-division gene FtsZ encodes a novel GTP-binding protein. Nature 359(6392):251–254
233. Reimann P (2002) Brownian motors: Noisy transport far from equilibrium. Phys Rep 361:57–265
234. Revenu C, Athman R, Robine S, Louvard D (2004) The coworkers of actin filaments: From cell structures to signals. Nat Rev Mol Cell Biol 5(8):635–646
235. Ridley AJ (2001) Rho family proteins: Coordinating cell responses. Trends Cell Biol 11(12):471–477
236. Ridley AJ, Schwartz MA, Burridge K, Firtel RA, Ginsberg MH, Borisy G, Parsons JT, Horwitz AR (2003) Cell migration: Integrating signals from front to back. Science 302(5651): 1704–1709
237. Rieder CL, Khodjakov A (2003) Mitosis through the microscope: advances in seeing inside live dividing cells. Science 300(5616):91–96
238. Risler T, Prost J, Jülicher F (2004) Universal critical behavior of noisy coupled oscillators. Phys Rev Lett 93(17):175702
239. Riveline D, Ott A, Jülicher F, Winkelmann DA, Cardoso O, Lacapere JJ, Magnusdottir S, Viovy JL, Gorre-Talini L, Prost J (1998) Acting on actin: The electric motility assay. Eur Biophys J 27(4):403–408
240. Roberts TM, Stewart M (2000) Acting like actin. The dynamics of the nematode major sperm protein (MSP) cytoskeleton indicate a push-pull mechanism for amoeboid cell motility. J Cell Biol 149(1):7–12
241. Rogers SL, Gelfand VI (2000) Membrane trafficking, organelle transport, and the cytoskeleton. Curr Opin Cell Biol 12(1): 57–62
242. Rubinstein B, Jacobson K, Mogilner A (2005) Multiscale two-dimensional modeling of a motile simple-shaped cell. Multiscale Model Simul 3(2):413–439
243. Saez A, Ghibaudo M, Buguin A, Silberzan P, Ladoux B (2007) Rigidity-driven growth and migration of epithelial cells on microstructured anisotropic substrates. Proc Natl Acad Sci USA 104(20):8281–8286
244. Sammak PJ, Borisy GG (1988) Direct observation of microtubule dynamics in living cells. Nature 332(6166):724–726
245. Sankararaman S, Menon GI, Sunil Kumar PB (2004) Self-organized pattern formation in motor-microtubule mixtures. Phys Rev E Stat Nonlin Soft Matter Phys 70(3 Pt 1):031905
246. Schliwa M, Woehlke G (2003) Molecular motors. Nature 422(6933):759–765
247. Schmittmann B, Zia RKP (1995) Statistical mechanics of driven diffusive systems. In: Domb C, Lebowitz JL (eds) Phase transitions and critical phenomena, vol 17. Academic Press, London
248. Scholey JM, Brust-Mascher I, Mogilner A (2003) Cell division. Nature 422(6933):746–752
249. Schuyler SC, Pellman D (2001) Microtubule "plus-end-tracking proteins": The end is just the beginning. Cell 105(4): 421–424
250. Schütz GM (2001) Exactly solvable models for many-body systems. In: Domb C, Lebowitz JL (eds) Phase transitions and critical phenomena, vol 19. Academic Press, London
251. Sekimoto K, Nakazawa H (1998) Contraction of a bundle of actin filaments: 50 years after Szent-Gyorgyi, vol 1. World Scientific, Singapore, p 394
252. Sekimoto K, Prost J, Jülicher F, Boukellal H, Bernheim-Grosswasser A (2004) Role of tensile stress in actin gels and a symmetry-breaking instability. Eur Phys J E Soft Matter 13(3): 247–259
253. Shih Y-L, Rothfield L (2006) The bacterial cytoskeleton. Microbiol Mol Biol Rev 70(3):729–754
254. Simha RA, Ramaswamy S (2002) Hydrodynamic fluctuations and instabilities in ordered suspensions of self-propelled particles. Phys Rev Lett 89(5):058101
255. Small JV, Geiger B, Kaverina I, Bershadsky A (2002) How do microtubules guide migrating cells? Nat Rev Mol Cell Biol 3(12):957–964
256. Small JV, Kaverina I (2003) Microtubules meet substrate adhesions to arrange cell polarity. Curr Opin Cell Biol 15(1): 40–47
257. Small JV, Stradal T, Vignal E, Rottner K (2002) The lamellipodium: Where motility begins. Trends Cell Biol 12(3): 112–120
258. Spudich JA (1994) How molecular motors work. Nature 372(6506):515–518
259. Storm C, Pastore JJ, MacKintosh FC, Lubensky TC, Janmey PA (2005) Nonlinear elasticity in biological gels. Nature 435(7039):191–194
260. Surrey T, Nedelec F, Leibler S, Karsenti E (2001) Physical properties determining self-organization of motors and microtubules. Science 292(5519):1167–1171
261. Täuber UC, Howard M, Vollmayr-Lee BP (2005) Applications of field-theoretic renormalization group methods to reaction-diffusion problems. J Phys A 38(17):R79–R131
262. Taunton J, Rowning BA, Coughlin ML, Wu M, Moon RT, Mitchison TJ, Larabell CA (2000) Actin-dependent propulsion of endosomes and lysosomes by recruitment of N-WASp. J Cell Biol 148(3):519–530
263. Taylor GI (1951) Analysis of the swimming of microscopic organisms. Proc R Soc A 209:447
264. Theriot JA (2000) The polymerization motor. Traffic 1(1): 19–28
265. Theriot JA, Mitchison TJ, Tilney LG, Portnoy DA (1992) The rate of actin-based motility of intracellular Listeria monocytogenes equals the rate of actin polymerization. Nature 357(6375):257–260
266. Tilney LG, Portnoy DA (1989) Actin filaments and the growth, movement, and spread of the intracellular bacterial parasite, Listeria monocytogenes. J Cell Biol 109(4 Pt 1):1597–1608
267. Toner J, Tu YH (1998) Flocks, herds, and schools: A quantitative theory of flocking. Phys Rev E 58(4):4828–4858
268. Tran PT, Marsh L, Doye V, Inoue S, Chang F (2001) A mechanism for nuclear positioning in fission yeast based on microtubule pushing. J Cell Biol 153(2):397–411
269. Turing AM (1952) The chemical basis of morphogenesis. Phil Trans Roy Soc (Lond) 237:37

270. Upadhyaya A, Chabot JR, Andreeva A, Samadani A, van Oudenaarden A (2003) Probing polymerization forces by using actin-propelled lipid vesicles. Proc Natl Acad Sci USA 100(8):4521–4526
271. Upadhyaya A, van Oudenaarden A (2003) Biomimetic systems for studying actin-based motility. Curr Biol 13(18): R734–744
272. Urrutia R, McNiven MA, Albanesi JP, Murphy DB, Kachar B (1991) Purified kinesin promotes vesicle motility and induces active sliding between microtubules in vitro. Proc Natl Acad Sci USA 88(15):6701–6705
273. Vale RD, Milligan RA (2000) The way things move: Looking under the hood of molecular motor proteins. Science 288(5463):88–95
274. Vale RD, Reese TS, Sheetz MP (1985) Identification of a novel force-generating protein, kinesin, involved in microtubule-based motility. Cell 42(1):39–50
275. Vallotton P, Danuser G, Bohnet S, Meister J-J, Verkhovsky AB (2005) Tracking retrograde flow in keratocytes: News from the front. Mol Biol Cell 16(3):1223–1231
276. Van den Ent F, Amos LA, Lowe J (2001) Prokaryotic origin of the actin cytoskeleton. Nature 413(6851):39–44
277. Verkhovsky AB, Svitkina TM, Borisy GG (1999) Self-polarization and directional motility of cytoplasm. Curr Biol 9(1): 11–20
278. Vernon GG, Woolley DM (2004) Basal sliding and the mechanics of oscillation in a mammalian sperm flagellum. Biophys J 87(6):3934–3944
279. Vilfan A, Frey E, Schwabl F (1998) Elastically coupled molecular motors. Eur Phys J B 3(4):535–546
280. Vilfan A, Frey E, Schwabl F (1999) Force-velocity relations of a two-state crossbridge model for molecular motors. Europhys Lett 45(3):283–289
281. Vilfan A, Jülicher F (2006) Hydrodynamic flow patterns and synchronization of beating cilia. Phys Rev Lett 96(5):058102
282. Voituriez R, Joanny JF, Prost J (2006) Generic phase diagram of active polar films. Phys Rev Lett 96(2):028102
283. Waterman-Storer CM, Worthylake RA, Liu BP, Burridge K, Salmon ED (1999) Microtubule growth activates Rac1 to promote lamellipodial protrusion in fibroblasts. Nat Cell Biol 1(1):45–50
284. Welch MD, Rosenblatt J, Skoble J, Portnoy DA, Mitchison TJ (1998) Interaction of human Arp2/3 complex and the Listeria monocytogenes ActA protein in actin filament nucleation. Science 281(5373):105–108
285. Wiesner S, Helfer E, Didry D, Ducouret G, Lafuma F, Carlier M-F, Pantaloni D (2003) A biomimetic motility assay provides insight into the mechanism of actin-based motility. J Cell Biol 160(3):387–398
286. Wolgemuth CW, Miao L, Vanderlinde O, Roberts T, Oster G (2005) MSP dynamics drives nematode sperm locomotion. Biophys J 88(4):2462–2471
287. Yasuda K, Shindo Y, Ishiwata S (1996) Synchronous behavior of spontaneous oscillations of sarcomeres in skeletal myofibrils under isotonic conditions. Biophys J 70(4):1823–1829
288. Yeh E, Yang C, Chin E, Maddox P, Salmon ED, Lew DJ, Bloom K (2000) Dynamic positioning of mitotic spindles in yeast: Role of microtubule motors and cortical determinants. Mol Biol Cell 11(11):3949–3961
289. Yoshida K, Soldati T (2006) Dissection of amoeboid movement into two mechanically distinct modes. J Cell Sci 119(Pt 18):3833–3844
290. Yoshida M, Muneyuki E, Hisabori T (2001) ATP synthase – A marvellous rotary engine of the cell. Nat Rev Mol Cell Biol 2(9):669–677
291. Zhou FQ, Cohan CS (2001) Growth cone collapse through coincident loss of actin bundles and leading edge actin without actin depolymerization. J Cell Biol 153(5):1071–1084
292. Ziebert F, Zimmermann W (2004) Pattern formation driven by nematic ordering of assembling biopolymers. Phys Rev E 70(2):022902
293. Ziebert F, Zimmermann W (2005) Nonlinear competition between asters and stripes in filament-motor systems. Eur Phys JE 18(1):41–54
294. Zinn-Justin J (2002) Quantum field theory and critical phenomena, 4th edn. Oxford University Press, Oxford
295. Zumdieck A, Lagomarsino MC, Tanase C, Kruse K, Mulder B, Dogterom M, Jülicher F (2005) Continuum description of the cytoskeleton: ring formation in the cell cortex. Phys Rev Lett 95(25):258103

Books and Reviews

Alberts B et al (2002) The cytoskeleton. The mechanics of cell division. In: Gibbs S (ed) Molecular biology of the cell. Garland, New York

Bray D (2000) In: Day M (ed) Cell movements. Garland, New York

Howard J (2001) Mechanics of motor proteins and the cytoskeleton. Sinauer Associates Inc, Sunderland

Lenz P (ed) (2008) Cell motility. Biological and medical physics, biomedical engineering. Springer, New York

Reviews of special interest and that cover the subjects treated in this article can be found in the following references: [9,57,67,114,129,130,136,163,222,223,233,236,253,257].

D

Data and Dimensionality Reduction in Data Analysis and System Modeling

Witold Pedrycz[1,2]
[1] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada
[2] Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Article Outline

Glossary
Definition of the Subject
Introduction
Information Granularity and Granular Computing
Data Reduction
Dimensionality Reduction
Co-joint Data and Dimensionality Reduction
Conclusions
Future Directions
Acknowledgments
Appendix: Particle Swarm Optimization (PSO)
Bibliography

Glossary

Reduction process A suite of activities leading to the reduction of available data and/or reduction of features.

Feature selection An algorithmic process in which a large set of features (attributes) is reduced by choosing a certain relatively small subset of them. The reduction is a combinatorial optimization task which is NP-complete. Given this, quite often it is realized in a suboptimal way.

Feature transformation A process of transforming a highly dimensional feature space into a low-dimensional counterpart. These transformations are linear or nonlinear and can be guided by some optimization criterion. The commonly encountered methods utilize Principal Component Analysis (PCA) which is an example of a linear feature transformation.

Dimensionality reduction A way of converting a large data set into a representative subset. Typically, the data are grouped into clusters whose prototypes are representatives of the overall data set.

Curse of dimensionality A phenomenon of a rapid exponential increase of computing related with the dimensionality of the problem (say, the number of data or the number of features) which prevents us from achieving an optimal solution. The curse of dimensionality leads to the construction of sub-optimal solutions.

Data mining A host of activities aimed at discovery of easily interpretable and experimentally sound findings in huge data sets.

Biologically-inspired optimization An array of optimization techniques realizing searches in highly-dimensional spaces where the search itself is guided by a collection of mechanisms (operators) inspired by biological search processes. Genetic algorithms, evolutionary methods, particle swarm optimization, and ant colonies are examples of biologically-inspired search techniques.

Definition of the Subject

Data and dimensionality reduction are fundamental pursuits of data analysis and system modeling. With the rapid growth of sizes of data sets and the diversity of data themselves, the use of some reduction mechanisms becomes a necessity. Data reduction is concerned with a reduction of sizes of data sets in terms of the number of data points. This helps reveal an underlying structure in data by presenting a collection of groups present in data. Given a number of groups which is very limited, the clustering mechanisms become effective in terms of data reduction. Dimensionality reduction is aimed at the reduction of the number of attributes (features) of the data which leads to a typically small subset of features or brings the data from a highly dimensional feature space to a new one of a far

lower dimensionality. A joint reduction process involves data and feature reduction.

Introduction

In the information age, we are continuously flooded by enormous amounts of data that need to be stored, transmitted, processed and understood. We expect to make sense of data and find general relationships within them. On the basis of data, we anticipate to construct meaningful models (classifiers or predictors). Being faced with this ongoing quest, there is an intensive research along this line with a clear cut objective to design effective and computationally feasible algorithms to combat a curse of dimensionality which comes associated with the data. Data mining has become one of the dominant developments in data analysis. The term intelligent data analysis is another notion which provides us with a general framework supporting thorough and user-oriented processes of data analysis in which reduction processes play a pivotal role.

Interestingly enough, the problem of dimensionality reduction and complexity management is by no means a new endeavor. It has been around for a number of decades almost from the very inception of computer science. One may allude here to pattern recognition, data visualization as the two areas in which we were faced with inherent data dimensionality. This has led to a number of techniques which as of now are regarded classic and are used quite intensively. There have been a number of approaches deeply rooted in classic statistical analysis. The ideas of principal component analysis, Fisher analysis and alike are the techniques of paramount relevance. What has changed quite profoundly over the decades is the magnitude of the problem itself which has forced us to the exploration of new ideas and optimization techniques involving advanced techniques of global search including tabu search and biologically-inspired optimization mechanisms.

In a nutshell, we can distinguish between reduction processes involving (a) data and (b) features (attributes). Data reduction is concerned with grouping data and revealing their structure in the form of clusters (groups). Clustering is regarded as one of the fundamental techniques within the domain of data reduction. Typically, we start with thousands of data points and arrive at 10–15 clusters. Feature or attribute reduction deals with a (a) transformation of the feature space into another feature space of a far lower dimensionality or (b) selection of a subset of features that are regarded to be the most essential with respect to a certain predefined objective function. Considering the underlying techniques of feature transformation, we encounter a number of classic linear statistical techniques such as e. g., principal component analysis or more advanced nonlinear mapping mechanisms realized by e. g., neural networks.

The criteria used to assess the quality of the resulted (reduced) feature space give rise to the two general categories, namely *filters* and *wrappers*. Using filters we consider some criterion that pertains to the statistical characteristics of the selected attributes and evaluate them with this respect. In contrast, when dealing with wrappers, we are concerned with the effectiveness of the features as a vehicle to carry out classification so in essence there is a mechanism (e. g., a certain classifier) which effectively evaluates the performance of the selected features with respect to their discriminating capabilities.

In addition to feature and data reduction being regarded as two separate processes, we may consider their combinations. A general roadmap of the dimensionality reduction is outlined in Fig. 1. While the most essential components have been already described, we note that all reduction processes are guided by various criteria. The reduction activities are established in some formal frameworks of information granules.

In the study, we will start with an idea of information granularity and information granules, Sect. "Introduction", in which we demonstrate their fundamental role of the concept that leads to the sound foundations and helps establish a machinery of data and feature reduction. We start with data reduction (Sect. "Information Granularity and Granular Computing") where we highlight the role of clustering and fuzzy clustering as a processing vehicle leading to the discovery of structural relationships within the data. We also cover several key measures used to the evaluation of reduction process offering some thoughts on fuzzy quantization which brings a quantitative characterization of the reconstruction process. In this setting, we show that a tandem of encoding and decoding is guided by well-grounded optimization criteria. Dimension reduction is covered in Sect. "Data Reduction". Here we start with linear and nonlinear transformations of the feature space including standard methods such as a well-known Principal Component Analysis (PCA). In the sequel, we stress a way in which biologically inspired optimization leads to the formation of the optimal subset of features. Some mechanisms of co-joint data and dimensionality reduction are outlined in Sect. "Dimensionality Reduction" in which we discuss a role of biclustering in these reduction problems.

In this paper, we adhere to the standard notation. Individual data are treated as n-dimensional vectors of real numbers, say $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ etc. We consider a collection of

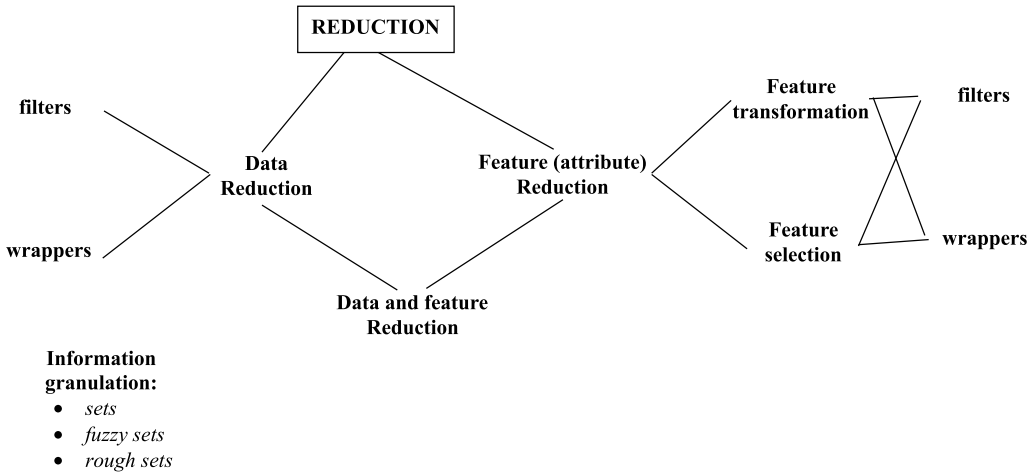

Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 1
A general roadmap of reduction processes; note various ways of dealing with data and dimensionality along with a way of providing some evaluation mechanisms (performance indexes)

N data points which could be arranged in a matrix form where data occupy consecutive rows of the matrix. Furthermore we will be using the terms attribute and feature interchangeably.

Information Granularity and Granular Computing

Information granules permeate numerous human endeavors [1,19,20,22,23,24,25,26]. No matter what problem is taken into consideration, we usually express it in a certain conceptual framework of basic entities, which we regard to be of relevance to the problem formulation and problem solving. This becomes a framework in which we formulate generic concepts adhering to some level of abstraction, carry out processing, and communicate the results to the external environment. Consider, for instance, image processing. In spite of the continuous progress in the area, a human being assumes a dominant and very much uncontested position when it comes to understanding and interpreting images. Surely, we do not focus our attention on individual pixels and process them as such but group them together into semantically meaningful constructs – familiar objects we deal with in everyday life. Such objects involve regions that consist of pixels or categories of pixels drawn together because of their proximity in the image, similar texture, color, etc. This remarkable and unchallenged ability of humans dwells on our effortless ability to construct information granules, manipulate them and arrive at sound conclusions. As another example, consider a collection of time series. From our perspective we can describe them in a semi-qualitative manner by pointing at specific regions of such signals. Specialists can effortlessly interpret ECG signals. They distinguish some segments of such signals and interpret their combinations. Experts can interpret temporal readings of sensors and assess the status of the monitored system. Again, in all these situations, the individual samples of the signals are not the focal point of the analysis and the ensuing signal interpretation. We always granulate all phenomena (no matter if they are originally discrete or analog in their nature). Time is another important variable that is subjected to granulation. We use seconds, minutes, days, months, and years. Depending which specific problem we have in mind and who the user is, the size of information granules (time intervals) could vary quite dramatically. To the high level management time intervals of quarters of year or a few years could be meaningful temporal information granules on basis of which one develops any predictive model. For those in charge of everyday operation of a dispatching plant, minutes and hours could form a viable scale of time granulation. For the designer of high-speed integrated circuits and digital systems, the temporal information granules concern nanoseconds, microseconds, and perhaps microseconds. Even such commonly encountered and simple examples are convincing enough to lead us to ascertain that (a) information granules are the key components of knowledge representation and processing, (b) the level of granularity of information granules (their size, to be more descriptive) becomes crucial to the problem description and an overall strategy of problem solving, (c) there is no universal level of granularity of information; the size of granules is problem-oriented and user dependent.

What has been said so far touched a qualitative aspect of the problem. The challenge is to develop a computing framework within which all these representation and processing endeavors could be formally realized. The common platform emerging within this context comes under the name of Granular Computing. In essence, it is an emerging paradigm of information processing. While we have already noticed a number of important conceptual and computational constructs built in the domain of system modeling, machine learning, image processing, pattern recognition, and data compression in which various abstractions (and ensuing information granules) came into existence, Granular Computing becomes innovative and intellectually proactive in several fundamental ways

- It identifies the essential commonalities between the surprisingly diversified problems and technologies used there which could be cast into a unified framework we usually refer to as a granular world. This is a fully operational processing entity that interacts with the external world (that could be another granular or numeric world) by collecting necessary granular information and returning the outcomes of the granular computing.
- With the emergence of the unified framework of granular processing, we get a better grasp as to the role of interaction between various formalisms and visualize a way in which they communicate.
- It brings together the existing formalisms of set theory (interval analysis) [8,10,15,21], fuzzy sets [6,26], rough sets [16,17,18], etc. under the same roof by clearly visualizing that in spite of their visibly distinct underpinnings (and ensuing processing), they exhibit some fundamental commonalities. In this sense, Granular Computing establishes a stimulating environment of synergy between the individual approaches.
- By building upon the commonalities of the existing formal approaches, Granular Computing helps build heterogeneous and multifaceted models of processing of information granules by clearly recognizing the orthogonal nature of some of the existing and well established frameworks (say, probability theory coming with its probability density functions and fuzzy sets with their membership functions).
- Granular Computing fully acknowledges a notion of variable granularity whose range could cover detailed numeric entities and very abstract and general information granules. It looks at the aspects of compatibility of such information granules and ensuing communication mechanisms of the granular worlds.
- Interestingly, the inception of information granules is highly motivated. We do not form information granules without reason. Information granules arise as an evident realization of the fundamental paradigm of abstraction.

Granular Computing forms a unified conceptual and computing platform. Yet, it directly benefits from the already existing and well-established concepts of information granules formed in the setting of set theory, fuzzy sets, rough sets and others. A selection of a certain formalism from this list depends upon the problem at hand. While in dimensionality reduction set theory becomes more visible, in data reduction we might encounter both set theory and fuzzy sets.

Data Reduction

In data reduction, we transform large collections of data into a limited, quite small family of their representatives which capture the underlying structure (topology). Clustering has become one of the fundamental tools being used in this setting. Depending upon the distribution of data, we may anticipate here the use of set theory or fuzzy sets. As an illustration, consider a two-dimensional data portrayed in Fig. 2.

The data in Fig. 2a exhibit a clearly delineated structure with three well separated clusters. Each cluster can be formalized as a set of elements. A situation illustrated in Fig. 2b is very different: while there are some clusters visible, they overlap to some extent and a number of points are "shared" by several clusters. In other words, some data may belong to more than a single cluster and hence the emergence of fuzzy sets as a formal setting in which we can formalize the resulting information granules.

In what follows, we briefly review the essence of fuzzy clustering and then move on with the ideas of the evaluation of quality of the results.

Fuzzy C-means as an Algorithmic Vehicle of Data Reduction Through Fuzzy Clusters

Fuzzy sets can be formed on a basis of numeric data through their clustering (groupings). The groups of data give rise to membership functions that convey a global more abstract view at the available data. With this regard Fuzzy C-Means (FCM) is one of the commonly used mechanisms of fuzzy clustering [2,3,20].

Let us review its formulation, develop the algorithm and highlight the main properties of the fuzzy clusters. Given a collection of n-dimensional data set $\{\boldsymbol{x}_k\}, k =$

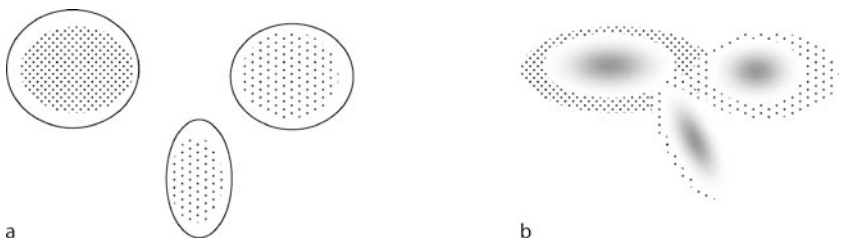

Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 2
Data clustering and underlying formalism of set theory suitable for handling well-delineated structures of data (a) and the use of fuzzy sets in capturing the essence of data with significant overlap (b)

$1, 2, \ldots, N$, the task of determining its structure – a collection of "c" clusters, is expressed as a minimization of the following objective function (performance index) Q being regarded as a sum of the squared distances between data and their representatives (prototypes)

$$Q = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^{m} ||\boldsymbol{x}_k - \boldsymbol{v}_i||^2 . \tag{1}$$

Here $\boldsymbol{v}_i$ s are n-dimensional prototypes of the clusters, $i = 1, 2, \ldots, c$ and $U = [u_{ik}]$ stands for a partition matrix expressing a way of allocation of the data to the corresponding clusters; u_{ik} is the membership degree of data $\boldsymbol{x}_k$ in the ith cluster. The distance between the data $\boldsymbol{z}_k$ and prototype $\boldsymbol{v}_i$ is denoted by $||.||$. The fuzzification coefficient m (> 1.0) expresses the impact of the membership grades on the individual clusters.

A partition matrix satisfies two important and intuitively appealing properties

$$\begin{aligned} &\text{(a)} \quad 0 < \sum_{k=1}^{N} u_{ik} < N, \quad i = 1, 2, \ldots, c \\ &\text{(b)} \quad \sum_{i=1}^{c} u_{ik} = 1, \quad k = 1, 2, \ldots, N . \end{aligned} \tag{2}$$

Let us denote by $\boldsymbol{U}$ a family of matrices satisfying (a)–(b). The first requirement states that each cluster has to be nonempty and different from the entire set. The second requirement states that the sum of the membership grades should be confined to 1.

The minimization of Q completed with respect to $U \in \boldsymbol{U}$ and the prototypes $\boldsymbol{v}_i$ of $V = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_c\}$ of the clusters. More explicitly, we write it down as follows

$$\min Q \quad \text{with respect to} \quad U \in \mathbf{U},\ \boldsymbol{v}_1,\ \boldsymbol{v}_2, \ldots, \boldsymbol{v}_c \in \mathbf{R}^n . \tag{3}$$

From the optimization standpoint, there are two individual optimization tasks to be carried out separately for the partition matrix and the prototypes. The first one concerns the minimization with respect to the constraints given the requirement of the form (2b) which holds for each data point $\boldsymbol{x}_k$. The use of Lagrange multipliers converts the problem into its constraint-free version. The augmented objective function formulated for each data point, $k = 1, 2, \ldots, N$, reads as

$$V = \sum_{i=1}^{c} u_{ik}^{m} d_{ik}^2 + \lambda \left(\sum_{i=1}^{c} u_{ik} - 1 \right) \tag{4}$$

where $d_{ik}^2 = ||\boldsymbol{x}_k - \boldsymbol{v}_i||^2$.

It is now instructive to go over the optimization process. Proceeding with the necessary conditions for the minimum of V for $k = 1, 2, \ldots, N$, one has

$$\frac{\partial V}{\partial u_{st}} = 0 \qquad \frac{\partial V}{\partial \lambda} = 0 \tag{5}$$

$s = 1, 2, \ldots, c$, $t = 1, 2, \ldots, N$. Now we calculate the derivative of V with respect to the elements of the partition matrix in the following way

$$\frac{\partial V}{\partial u_{st}} = m u_{st}^{m-1} d_{st}^2 + \lambda . \tag{6}$$

Given this relationship, using (5) we calculate u_{st} to be equal to

$$u_{st} = -\left(\frac{\lambda}{m}\right)^{\frac{1}{m-1}} d_{st}^{\frac{2}{m-1}} . \tag{7}$$

Given the normalization condition $\sum_{j=1}^{c} u_{jt} = 1$ and plugging it into (2) one has

$$-\left(\frac{\lambda}{m}\right)^{\frac{1}{m-1}} \sum_{j=1}^{c} d_{jt}^{\frac{2}{m-1}} = 1 \tag{8}$$

we compute

$$-\left(\frac{\lambda}{m}\right)^{\frac{1}{m-1}} = \frac{1}{\sum\limits_{j=1}^{c} d_{jt}^{\frac{2}{m-1}}} . \tag{9}$$

Inserting this expression into (7), we obtain the successive entries of the partition matrix

$$u_{st} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{st}^2}{d_{jt}^2} \right)^{\frac{1}{m-1}}} . \tag{10}$$

The optimization of the prototypes $\boldsymbol{v}_i$ is carried out assuming the Euclidean distance between the data and the prototypes that is $||\boldsymbol{x}_k - \boldsymbol{v}_i||^2 = \sum_{j=1}^{n} (x_{kj} - v_{ij})^2$. The objective function reads now as follows $Q = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^m \sum_{j=1}^{n} (x_{kj} - v_{ij})^2$ and its gradient with respect to $\boldsymbol{v}_i$, $\nabla_{\boldsymbol{v}_i} Q$ made equal to zero yields the system of linear equations

$$\sum_{k=1}^{N} u_{ik}^m (x_{kt} - v_{st}) = 0 \tag{11}$$

$s = 1, 2, \ldots, c;\ t = 1, 2, \ldots, n.$

Thus

$$v_{it} = \frac{\sum_{k=1}^{N} u_{ik}^m x_{kt}}{\sum_{k=1}^{N} u_{ik}^m} . \tag{12}$$

Overall, the FCM clustering is completed through a sequence of iterations where we start from some random allocation of data (a certain randomly initialized partition matrix) and carry out the following updates by adjusting the values of the partition matrix and the prototypes. The iterative process is continued until a certain termination criterion has been satisfied. Typically, the termination condition is quantified by looking at the changes in the membership values of the successive partition matrices. Denote by $U(t)$ and $U(t+1)$ the two partition matrices produced in two consecutive iterations of the algorithm. If the distance $||U(t+1) - U(t)||$ is less than a small predefined threshold ε (say, $\varepsilon = 10^{-5}$ or 10^{-6}), then we terminate the algorithm. Typically, one considers the Tchebyschev distance between the partition matrices meaning that the termination criterion reads as follows

$$\max_{i,k} |u_{ik}(t+1) - u_{ik}(t)| \leq \varepsilon . \tag{13}$$

The key components of the FCM and a quantification of their impact on the form of the produced results are summarized in Table 1.

The fuzzification coefficient exhibits a direct impact on the geometry of fuzzy sets generated by the algorithm. Typically, the value of "m" is assumed to be equal to 2.0. Lower values of m (that are closer to 1) yield membership functions that start resembling characteristic functions of sets; most of the membership values become localized around 1 or 0. The increase of the fuzzification coefficient ($m = 3, 4$, etc.) produces "spiky" membership functions with the membership grades equal to 1 at the prototypes and a fast decline of the values when moving away from the prototypes. Several illustrative examples of the membership functions are included in Fig. 3. In addition to the varying shape of the membership functions, observe that the requirement put on the sum of membership grades imposed on the fuzzy sets yields some rippling effect: the membership functions are not unimodal but may exhibit some ripples whose intensity depends upon the distribution of the prototypes and the values of the fuzzification coefficient.

The membership functions offer an interesting feature of evaluating an extent to which a certain data point is shared between different clusters and in this sense become

Data and Dimensionality Reduction in Data Analysis and System Modeling, Table 1
The main features of the Fuzzy C-Means (FCM) clustering algorithm

Feature of the FCM algorithm	Representation and optimization aspects
Number of clusters (c)	Structure in the data set and the number of fuzzy sets estimated by the method; the increase in the number of clusters produces lower values of the objective function however given the semantics of fuzzy sets one should maintain this number quite low (5–9 information granules)
Objective function Q	Develops the structure aimed at the minimization of Q; iterative process supports the determination of the local minimum of Q
Distance function $\|\|.\|\|$	Reflects (or imposes) a geometry of the clusters one is looking for; essential design parameter affecting the shape of the membership functions
Fuzzification coefficient (m)	Implies a certain shape of membership functions present in the partition matrix; essential design parameter. Low values of "m" (being close to 1.0) induce characteristic function. The values higher than 2.0 yield spiky membership functions
Termination criterion	Distance between partition matrices in two successive iterations; the algorithm terminated once the distance below some assumed positive threshold (ε) that is $\|\|U(\text{iter}+1) - U(\text{iter})\|\| < \varepsilon$

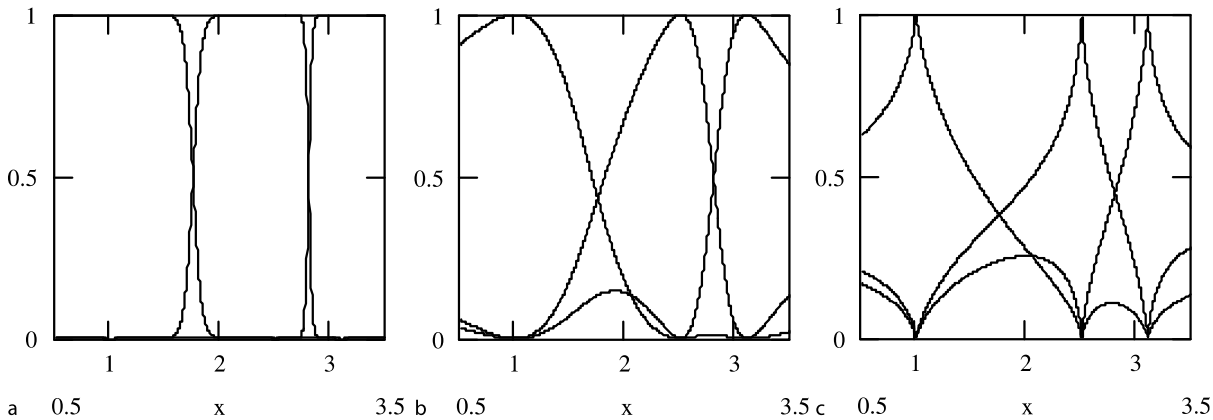

Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 3
Examples of membership functions of fuzzy sets; the prototypes are equal to 1, 3.5 and 5 while the fuzzification coefficient assumes values of 1.2 (a), 2.0 (b) and 3.5 (c). The intensity of the rippling effect is affected by the values of "*m*" and increases with the higher values of "*m*"

difficult to allocate to a single cluster (fuzzy set). Let us introduce the following index which serves as a certain separation measure between the clusters

$$\varphi(u_1, u_2, \ldots, u_c) = 1 - c^c \prod_{i=1}^{c} u_i \tag{14}$$

where $u_1, u_2, \ldots, u_c$ are the membership degrees for some data point. If only one of membership degrees, say $u_i = 1$, and the remaining are equal to zero, then the separation index attains its maximum equal to 1. On the other extreme, when the data point is shared by all clusters to the same degree equal to 1/c, then the value of the index drops down to zero. This means that there is no separation between the clusters as reported for this specific point.

While the number of clusters is typically limited to a few information granules, we can easily proceed with successive refinements of fuzzy sets. This can be done by splitting fuzzy clusters of the highest heterogeneity. Let us assume that we have already constructed "*c*" fuzzy clusters. Each of them can be characterized by the performance index

$$V_i = \sum_{k=1}^{N} u_{ik}^m ||\boldsymbol{x}_k - \boldsymbol{v}_i||^2 \tag{15}$$

$i = 1, 2, \ldots, c$. The higher the value of V_i, the more heterogeneous the ith cluster. The one with the highest value of V_i, that is the one for which we have $i_0 = \text{argmax}_i V_i$ is refined by being split into two clusters. Denote the set of data associated with the i_0th cluster by $\boldsymbol{X}(i_0)$,

$$\boldsymbol{X}(i_0) = \{x_k \in \boldsymbol{X} | u_{i_0 k} = \max_i u_{ik}\} . \tag{16}$$

We cluster the elements in $\boldsymbol{X}(i_0)$ by forming two clusters which leads to two more specific (detailed) fuzzy sets. This gives rise to a hierarchical structure of the family of fuzzy sets as illustrated in Fig. 4. The relevance of this construct in the setting of fuzzy sets is that it emphasizes the essence of forming a hierarchy of fuzzy sets rather than working with a single level structure of a large number of components whose semantics could not be retained.

The process of further refinements is realized in the same by picking up the cluster of the highest heterogeneity and its split into two consecutive clusters.

Evaluation of Performance of Data Reduction

There are two fundamental ways of evaluating the quality of data reduction achieved through clustering or fuzzy clustering. In the first category of methods, we assess the

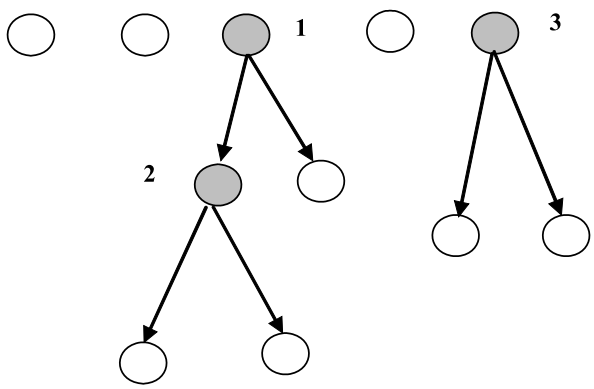

Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 4
Successive refinements of fuzzy sets through fuzzy clustering applied to the clusters of the highest heterogeneity. The numbers indicate the order of the splits

quality of the obtained clusters in terms of their compactness, low fuzziness (entropy measure of fuzziness, in particular), structural separability, and alike. We also look for a "plausible" number of clusters which can be descriptive of the structure of the overall data. The second category of methods stresses the aspects of data representation by its reduced counterpart. In this case, a reconstruction criterion plays a pivotal role. We elaborate on this method by looking at it as a useful vehicle of fuzzy quantization. Furthermore we contrast fuzzy quantization with its set-based counterpart.

Exploring Fuzzy Quantization: Moving Beyond a Winner-Takes-All Scheme

The crux of the considerations here dwells upon the following observation. While the set-based codebook developed through clustering [5,7] leads to a decoding scheme which decodes the result using a single element of the codebook (which in essence becomes a manifestation of the well-known concept of the *winner-takes-all* strategy), here we are interested in the exploitation of the nonzero degrees of membership of several elements (fuzzy sets) of the codebook while representing the input datum. In other words, rather than using a single prototype as a sole representative of a collection of neighboring data, our conjecture is that involving several prototypes at different degrees of activation (weights) could be beneficial to the ensuing decoding. Along the line of abandoning the winner-takes-all principle, let us start with a collection of weights $u_i(\boldsymbol{x})$, $i = 1, 2, \ldots, c$. Adhering to the vector notation, we have $\boldsymbol{u}(\boldsymbol{x}) = [u_1(\boldsymbol{x})\ u_2(\boldsymbol{x}), \ldots, u_c(\boldsymbol{x})]^{\mathrm{T}}$. The weights (membership degrees) express an extent to which the corresponding datum $\boldsymbol{x}$ is encoded in the language of the given prototypes (elements of the codebook) should be involved in the decoding (decompression) scheme. We require that these membership degrees are in-between 0 and 1 and sum up to 1. Thus at the encoding end of the overall scheme, we represent each vector $\boldsymbol{x}$ by $c-1$ values of $u_i(\boldsymbol{x})$. The decoding is then based upon a suitable aggregation of the degrees of membership and the prototypes. Denote this operation by $\hat{\boldsymbol{x}} = D(\boldsymbol{u}(\boldsymbol{x}), \boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_c)$ where $\hat{\boldsymbol{x}}$ denotes the result of the decoding. On the other hand, the formation of the membership degrees (encoding) can be succinctly described in the form of the mapping $E(\boldsymbol{x}, \boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_c)$.

The overall development process is split into two fundamental phases, namely

(a) local optimization activities confined to the individual datum $\boldsymbol{x}_k$ that involve (i) encoding each $\boldsymbol{x}_k$ leading to the vector of the membership degrees (u_i), (ii) decoding being realized in terms of the membership degrees, and
(b) global optimization activities concerning the formation of the codebook in which case we take all data into consideration thus bringing the design to the global level.

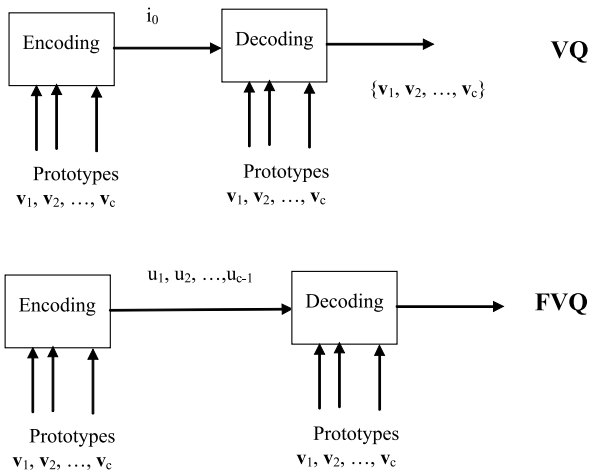

Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 5
VQ and FVQ – a view contrasting the essence of the process and showing the key elements of the ensuing encoding and decoding

The overall process described above will be referred to as fuzzy vector quantization (FVQ) or a fuzzy granulation-degranulation. The latter terminology emphasizes a position of fuzzy sets played in this construct. In contrast, the winner-takes-all strategy is a cornerstone of the vector quantization (VQ). To contrast the underlying computing in the VQ and FVQ schemes, we portray the essential computational facets in Fig. 5.

We focus on the detailed formulas expressing the coding and decoding schemes. Let us emphasize that the coding and decoding emerge as solutions to the well-defined optimization problems. At this point let us assume that the codebook – denoted here as $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_c\}$ has been already formed and become available.

Encoding Mechanism A way of encoding (representing) the original datum $\boldsymbol{x}$ is done through the collection of the degrees of activation of the elements of the codebook. We require that the membership degrees are confined to the unit interval and sum up to 1. We determine their individual values by minimizing the following performance index

$$Q_1(\boldsymbol{x}) = \sum_{i=1}^{c} u_i^m ||\boldsymbol{x} - \boldsymbol{v}_i||^2 \tag{17}$$

subject to the following constraints already stated above, that is

$$u_i(\boldsymbol{x}) \in [0,1], \quad \sum_{i=1}^{c} u_i(x) = 1 \,. \tag{18}$$

The distance function is denoted by $||.||^2$. The fuzzification coefficient (m, $m > 1$), standing in the above expression is used to adjust the level of contribution of the impact of the prototypes on the result of the encoding. The collection of "c" weights $\{u_i(\boldsymbol{x})\}$ is then used to encode the input datum $\boldsymbol{x}$. These membership degrees along with the corresponding prototypes are afterwards used in the decoding scheme.

The minimization of (17) is straightforward and follows a standard way of transforming the problem to unconstrained optimization using Lagrange multipliers. Once solved, the resulting weights (membership degrees) read as

$$u_i(\boldsymbol{x}) = \frac{1}{\sum \left(\frac{||\boldsymbol{x}-\boldsymbol{v}_i||}{||\boldsymbol{x}-\boldsymbol{v}_j||}\right)^{\frac{2}{m-1}}} \,. \tag{19}$$

The fuzzification coefficient used in the formulation of the encoding problem plays an important role. To demonstrate this, let us show how the values of the coefficient impact the distribution of the generated membership degrees. The plots of these degrees in the two-dimensional case ($n = 2$) and $c = 3$ elements of the codebook are included in Fig. 6. The prototypes have been set up to $\boldsymbol{v}_1 = [3.0\ 1.5]^{\mathrm{T}}, \boldsymbol{v}_2 = [2.5\ 0.8]^{\mathrm{T}}, \boldsymbol{v}_3 = [1.0\ 3.7]^{\mathrm{T}}$. The values of "$m$" close to 1 result in a visible set-like character of the distribution of the membership degrees.

The Decoding Mechanism The decoding is concerned with the "reconstruction" of $\boldsymbol{x}$, denoted here by $\hat{\boldsymbol{x}}$. It is based on some aggregation of the elements of the codebook and the associated membership grades $u(\boldsymbol{x})$. The proposed way of forming $\hat{\boldsymbol{x}}$ is accomplished through the minimization of the following expression

$$Q_2(\hat{\boldsymbol{x}}) = \sum_{i=1}^{c} u_i^m ||\hat{\boldsymbol{x}} - \boldsymbol{v}_i||^2 \,. \tag{20}$$

Given the Euclidean distance, the problem of unconstrained optimization leads to a straightforward solution expressed as a combination of the prototypes weighted by the membership degrees, that is

$$\hat{\boldsymbol{x}} = \frac{\sum_{i=1}^{c} u_i^m \boldsymbol{v}_i}{\sum_{i=1}^{c} u_i^m} \,. \tag{21}$$

Interestingly, all prototypes contribute to the decoding process and this way of computing stands in a sharp contrast with the winner-takes-all decoding scheme encountered in the VQ where $\hat{\boldsymbol{x}} = \boldsymbol{v}_l$ where l denotes the index of the winning prototype that was identified during the decoding phase. Some other VQ decoding schemes are available in which several dominant elements of the codebook are involved. Nevertheless because of their binary involvement in the decoding phase, the possible ways of aggregation to be realized by the decoder are quite limited.

The quality of the reconstruction depends on a number of essential parameters including the size of the codebook as well as the value of the fuzzification coefficient. The impact of these particular parameters is illustrated in Fig. 7 in which we quantify the distribution of the decoding error by showing values of the Hamming distance between $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$. Evidently, what could have been anticipated, all $\boldsymbol{x}$'s positioned in a close vicinity of the prototypes exhibit low values of the decoding error. Low values of "m" lead to significant jumps of the error which result from the abrupt switching in-between the prototypes used in the decoding process. For the higher value of the fuzzification coefficient, say $m = 2.0$, the jumps are significantly

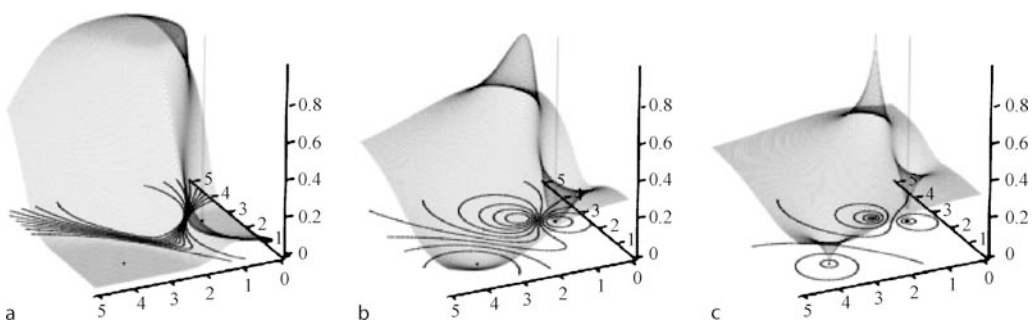

Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 6
Plots of $u_1(x)$ (3D and contour plots) for selected values of the fuzzification coefficient: a $m = 1.2$, b $m = 2.0$, c $m = 3.5$

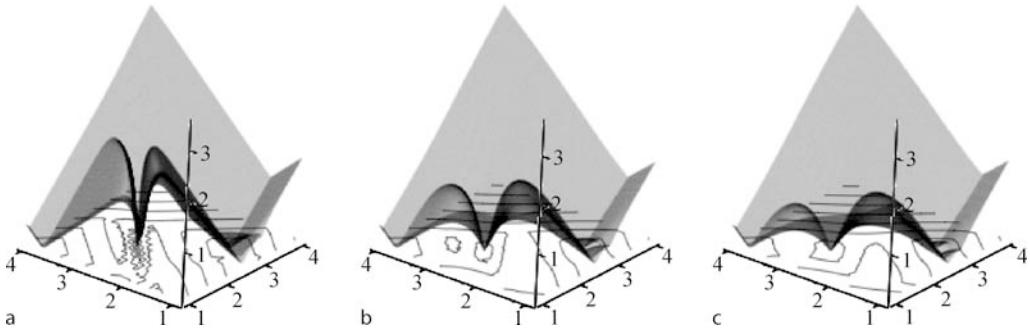

Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 7
Plots of the decoding error expressed in the form of the Hamming distance between x and $\hat{x}$ (3D and contour plots) for selected values of the fuzzification coefficient: a $m = 1.2$, b $m = 2.0$, c $m = 3.5$

reduced. Further reduction is achieved with the larger values of "m", see Fig. 7c.

Obviously in the regions which are quite remote from the location of the prototypes, the error increases quite quickly. For instance, this effect is quite visible for the values of $\boldsymbol{x}$ close to $[4.0\ 4.0]^{\mathrm{T}}$.

Dimensionality Reduction

As indicated earlier, dimensionality reduction is concerned with a reduction of the set of features. One of the two fundamental directions that are taken here involves transformations of the original, highly dimensional feature vectors. The other group of methods focuses on a selection of an "optimal" subset of original attributes.

Linear and Nonlinear Transformations of the Feature Space

Making use of such transformations, we position the original vectors in some other spaces of far lower dimensionality than the original one. There are a number of statistical methods devoted to the formation of optimal linear transformations. For instance, the commonly used Principal Component Analysis (PCA) [11] deals with a linear mapping of original data in such a way that the reduced data vectors $\boldsymbol{y}$ are expressed as $\boldsymbol{y} = \mathrm{T}\boldsymbol{x}$ where T denotes a transformation matrix and $\boldsymbol{y}$ is positioned in the reduced space of dimensionality "m" where $m \ll n$. The transformation matrix is constructed using eigenvectors associated with the largest eigenvalues of the covariance matrix of the data, that is $\Xi = \sum_{k=1}^{N} (\boldsymbol{x}_k - \boldsymbol{m})(\boldsymbol{x}_k - \boldsymbol{m})^{\mathrm{T}}$ where $\boldsymbol{m}$ is a mean vector of all data. The PCA method does not consider the labels of data in case they become available. To accommodate class information, one may consider Fisher discriminant analysis [3].

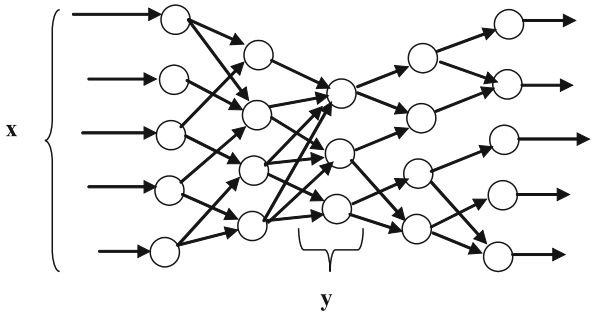

Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 8
Neural network in nonlinear dimensionality reduction; the inner shortest layer produces a reduced version of the original data (y)

Neural networks can serve as nonlinear transformations of the original feature space. More specifically, the dimensionality reduction is accomplished through a sand glass topology of the network as illustrated in Fig. 8. Notice that the shortest hidden layer of the network produces the reduced version of data. The optimization criterion deals with the minimal reconstruction error. For each input $\boldsymbol{x}$, we strive for the output of the network (NN) to be made as close as possible to the input, that is $\mathrm{NN}(\boldsymbol{x}) \approx \boldsymbol{x}$. The learning (adjustments) of the weights helps minimize the error (objective function) of the form $\sum_{k=1}^{N} ||\mathrm{NN}(\boldsymbol{x}_k) - \boldsymbol{x}_k||^2$.

Given the evident nonlinear nature of neurocomputing, we often allude to these transformations as a nonlinear component analysis (NLCA).

Neural networks are also used as a visualization vehicle in which we display original data in some low-dimensional space. The typical architecture comes in the form of Kohonen self-organizing maps [13]. Data that are distant in an original space remain distant in the low-dimensional

space. Likewise, if the two data points are originally close to each other, this closeness is also present in the transformed feature space. While Kohonen maps realize a certain form of clustering, their functioning is quite distinct from the FCM algorithm described before. In essence, self-organizing maps bring a high involvement of humans in the final interpretation of results. For instance, the number of clusters is not predefined in advance, as encountered in the FCM, but, its choice is left to the end user to determine on a basis of the structure displayed on the map.

The Development of Optimal Subsets of Features

An objective is to select a subset of attributes so that a certain predefined performance index (selection criterion) becomes optimized (either minimized or maximized). The choice of a subset of attributes is a combinatorial problem. As a matter of fact, it is NP complete. The only optimal solution comes through exhaustive search. Evidently, such an enumeration of possibilities is feasible for problems of a very low dimensionality. In essence, the practicality of such enumeration is out of question. This has led to numerous techniques which rely on some sort of heuristics which help reduce the size of the search space. While the heuristics could be effective, its usage could easily build some bias into the search process. Subsequently, the results produced by such methods are suboptimal. In addition to a number of heuristics (which contribute to the effectiveness of the search process), there are more advanced optimization techniques such as those available in the realm of Evolutionary Computing [14] such as genetic algorithms, evolutionary strategies, particle swarm optimization are examples of biologically inspired techniques that are aimed at structural optimization using which we construct an optimal (or sub-optimal) collection of attributes. Any potential solution is represented in a form of a chromosome that identifies attributes contributing to the formation of the optimal space. We may encounter either binary or real-number chromosomes whose size is equal to the number of features. In case of the binary chromosome, we choose attributes which are identified as 1, see Fig. 9a. For the required "m" features, we choose the first ones encountered in the chromosome. For the real-number coding of the chromosome, we typically have far more alternatives to take advantage of. For instance, we could consider the first "m" largest entries of the chromosome. Then the coordinates of these entries form a reduced subset of the features; refer to Fig. 9b.

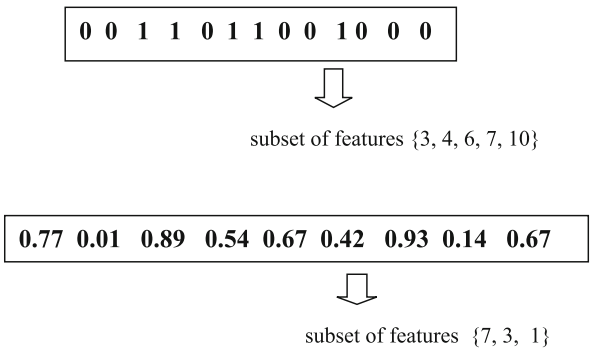

Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 9
Examples of forming a subset of features for a given chromosome: a binary coding, and b real-number coding (here we are concerned with a 3 dimensional feature space by choosing three dominant features)

Optimization Environment in Dimensionality Reduction: Wrappers and Filters

Attribute selection is guided by some criterion. There are numerous ways of expressing a way in which a collection of features is deemed "optimal". There is a generally accepted taxonomy of the methods in which we distinguish between so-called wrappers and filters. Filters embrace criteria that directly express the quality of the subset of features in terms of their homogeneity, compactness and alike. Wrappers, as the name implies, evaluate the quality of the subset of features after being used by some mechanism of classification or prediction. In other words, the assessment of the attributes is done once they have been "wrapped" in some construct. This means that the quality of the feature space is expressed indirectly. The essence of filters and wrappers in illustrated schematically in Fig. 10.

Co-joint Data and Dimensionality Reduction

So far our investigations were concerned with data or dimensionality reduction carried out individually. An interesting question arises as to the possibility of carrying out concurrent reduction in both ways. One among possible alternatives can be envisioned in the form of so-called bi-clustering (or co-clustering). The crux of this process is to carry our simultaneous clustering for data and features. The result is a collection of groups of data and groups of features.

We start with a basic notation. Let us denote by D_1, $D_2, \ldots, D_c$ subsets of data. Similarly, by $F_1, F_2, \ldots, F_c$ we denote subsets of features. We may regard D_i and F_i as two subsets of indexes of data and features $\{1, 2, \ldots, N\}$ and $\{1, 2, \ldots, n\}$. $\{D_i\}$ and $\{F_i\}$ form partitions so that the fundamental requirements of disjointness and total space

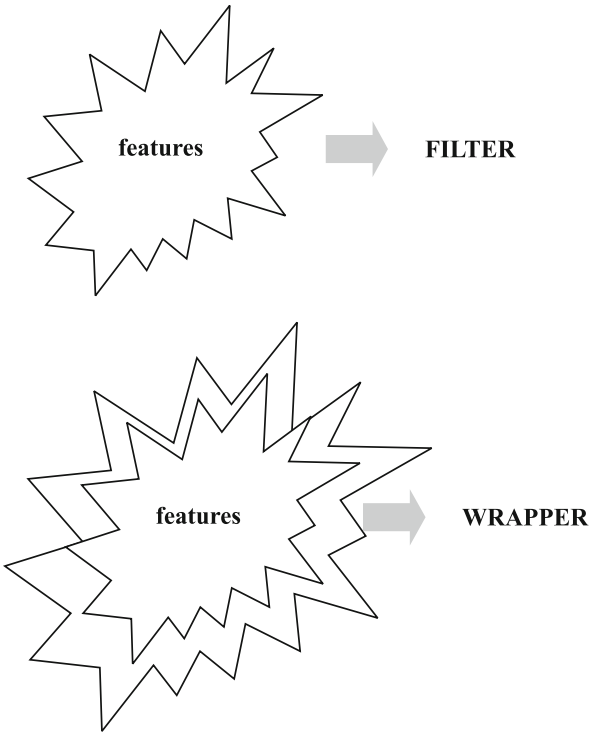

Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 10
Filters and wrappers in the assessment of the process of dimensionality reduction

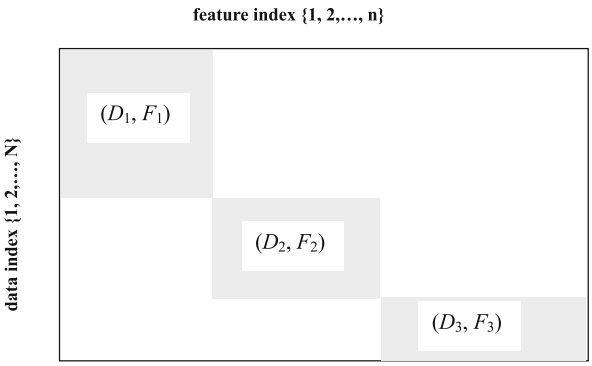

Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 11
An example heatmap of biclusters ($c = 3$)

coverage are satisfied. By biclusters of data set we mean a collection of pairs $(D_i,\ F_i)\ i = 1, 2, \ldots, c$.

Graphically, biclusters are visualized in the form of so-called heatmaps where each bicluster is displayed as a certain region in the data and feature space, Fig. 11. Here the data are represented as a two-dimensional array X.

There have been a number of biclustering algorithms. The method proposed by Hartigan [9] and Cheng and Church [4] are just two examples of biclustering algorithms. The objective function introduced by Hartigan is concerned with the overall variance of the "c" biclusters

$$Q = \sum_{k=1}^{c} \sum_{i \in D_k} \sum_{j \in F_k} (x_{ij} - m_k)^2 \tag{22}$$

where m_k is the mean of the kth bicluster

$$m_k = \frac{1}{|D_k||F_k|} \sum_{i \in D_k} \sum_{j \in F_k} x_{ij} \tag{23}$$

where $|.|$ denotes a cardinality of clusters in data space and feature space. The minimization of Q leads to the development of the biclusters.

The idea of Cheng and Church is to form bi-clusters in such a way so that we minimize the mean squared residue of each bi-cluster. More specifically, denote an average taken over data in D_i and features in F_i as follows

$$\begin{aligned} \mu_{ij} &= \frac{1}{|D_i|} \sum_{k \in D_i} x_{kj} \\ \nu_{ik} &= \frac{1}{|F_i|} \sum_{j \in F_i} x_{kj}\,. \end{aligned} \tag{24}$$

The residue of x_{ij} taken with respect to the kth bicluster is defined as

$$r_{ij} = x_{ij} - \mu_{ij} - \nu_{ij} + m_k$$

and then

$$V_k = \sum_{i \in D_k} \sum_{j \in F_k} r_{ij}\,. \tag{25}$$

The overall minimization of $V_1 + V_2 + \ldots + V_c$ gives rise to the formation of biclusters.

Biologically-inspired optimization can be used here to construct biclusters. Again, as before, a formation of a suitable structure (content) of a chromosome becomes crucial to the effectiveness of the obtained structure. One possible encoding could be envisioned as follows, refer to Fig. 12. Note that a chromosome consists of two parts where the first portion encodes the data while the other one deals with the features.

Assuming that a chromosome has entries in-between 0 and 1 and "c" biclusters are to be developed, we convert entries in the interval $[0, 1/c)$ to identify elements belonging to the first biclusters. Those entries that assume values in $[1/c, 2/c)$ indicate elements (either data or features) belonging to the second bicluster, etc.

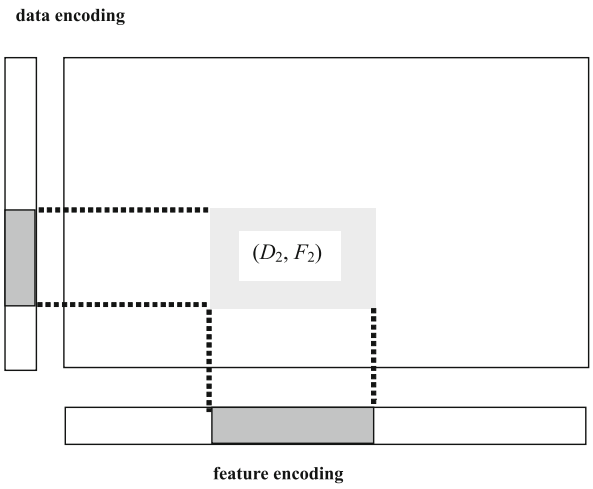

Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 12
Encoding of biclusters in terms of the chromosome with entries confined to the unit interval

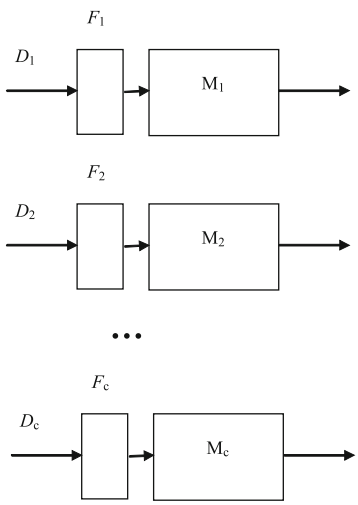

Data and Dimensionality Reduction in Data Analysis and System Modeling, Figure 13
System modeling realized through a collection of local models developed for individual biclusters

Biclustering brings an interesting issue of simultaneous reduction of data and dimensionality. Each cluster of data is expressed in terms of its own (reduced) feature space. This interesting phenomenon gives rise to a collection of models which are locally formed by using subset of data and a subset of features. Each model is developed on a basis of the given bicluster, see Fig. 13.

In this way, the complexity of modeling has been substantially reduced as instead of global models (which are more difficult to construct), we develop a family of compact models involving far fewer attributes and dealing with less data.

Conclusions

Reduction problems in data are essential to a number of fundamental developments of data analysis and system modeling. These activities are central to data understanding and constitute a core of generic activities of data mining and human-centric data interpretation. They are also indispensable in system modeling which otherwise being carried out for high-dimensional data make the development of the models highly inefficient and lead to the models that are of lower quality in particular when it comes to their generalization capabilities. The two facets of the reduction process involve data and features (attributes) and as such they tend to be highly complementary. When discussing individual techniques aimed at either data or dimensionality reduction, we indicated that there are interesting alternatives of dealing with these two aspects making use of biclustering.

Future Directions

There are a number of directions in which the area can expand either in terms of the reduction concept itself or the techniques being sought. Several of them might offer some tangible benefits to the future pursuits:

- The use of stratified approaches to data and dimensionality reduction in which the reduction processes are carried out on a hierarchical basis meaning that further reductions are contemplated bases on results developed at the lower level of generality.
- More active involvement and integration of techniques of visualization of results. While those are in place now, we may envision their more vital role in the future.
- Distributed mode of reduction activities where we simultaneously encounter several sources of data and intend to process them at the same time.
- Exploitation of joint data and dimensionality reduction approaches in which one tackles both the reduction of the number of data and the number of attributes. These two are quite intertwined so their mutual handling could be advantageous. In particular, one could envision their role in the formation of a suite of simpler models.
- Given the NP complete nature of many reduction problems, the use of some heuristics is inevitable. In this way, the role of techniques of biologically inspired optimization will be anticipated and it is very likely that their role will become more visible.

Acknowledgments

Support from the Natural Sciences and Engineering Research Council of Canada (NSERC) and Canada Research Chair (CRC) is gratefully acknowledged.

Appendix: Particle Swarm Optimization (PSO)

In the studies on dimensionality and data reduction, we consider the use of Particle Swarm Optimization (PSO). PSO is an example of a biologically-inspired and population-driven optimization. Originally, the algorithm was introduced by Kennedy and Eberhart [12] where the authors strongly emphasized an inspiration coming from the swarming behavior of animals as well as some inspiration coming from the as well as human social behavior, cf. also [19]. In essence, a particle swarm is a population of particles – possible solutions in the multidimensional search space. Each particle explores the search space and during this search adheres to some quite intuitively appealing guidelines navigating the search process: (a) it tries to follow its previous direction, and (b) it looks back at the best performance both at the level of the individual particle and the entire population. In this sense there is some collective search of the problem space along with some component of memory incorporated as an integral part of the search mechanism.

The performance of each particle during its movement is assessed by means of some performance index. A position of a swarm in the search space, is described by some vector $\boldsymbol{z}(t)$ where "t" denotes consecutive discrete time moments. The next position of the particle is governed by the following update expressions concerning the particle, $\boldsymbol{z}(t+1)$ and its speed, $\boldsymbol{v}(t+1)$

$$\boldsymbol{z}(t+1) = \boldsymbol{z}(t) + \boldsymbol{v}(t+1)$$

// update of position of the particle

$$\boldsymbol{v}(t+1) = \xi \boldsymbol{v}(t) + \phi_1(\boldsymbol{p} - \boldsymbol{x}(t)) + \phi_2(\boldsymbol{p}_{\text{total}} - \boldsymbol{x}(t))$$

// update of speed of the particle

where $\boldsymbol{p}$ denotes the best position (the lowest performance index) reported so far for this particle, $\boldsymbol{p}_{\text{total}}$ is the best position overall developed so far across the whole population. ϕ_1 and ϕ_2 are random number drawn from the uniform distribution $U[0,2]$ that help build a proper mix of the components of the speed; different random numbers affect the individual coordinates of the speed. The second expression governing the change in the velocity of the particle is particularly interesting as it nicely captures the relationships between the particle and its history as well as the history of the overall population in terms of their performance reported so far.

There are three components determining the updated speed of the particle. First, the current speed $\boldsymbol{v}(t)$ is scaled by the inertial weight (ξ) smaller than 1 whose role is to articulate some tendency to a drastic change of the current speed. Second, we relate to the memory of this particle by recalling the best position of the particle achieved so far. Thirdly, there is some reliance on the best performance reported across the whole population (which is captured by the last component of the expression governing the speed adjustment).

Bibliography

Primary Literature

1. Bargiela A, Pedrycz W (2003) Granular Computing: An Introduction. Kluwer, Dordrecht
2. Bezdek JC (1992) On the relationship between neural networks, pattern recognition and intelligence. Int J Approx Reason 6(2):85–107
3. Duda RO, Hart PE, Stork DF (2001) Pattern Classification, 2nd edn. Wiley, Ney York
4. Cheng Y, Church GM (2000) Biclustering of expression data. Proc 8th Int Conf on Intelligent Systems for Molecular Biology, pp 93–103
5. Gersho A, Gray RM (1992) Vector Quantization and Signal Compression. Kluwer, Boston
6. Gottwald S (2005) Mathematical fuzzy logic as a tool for the treatment of vague information. Inf Sci 172(1–2):41–71
7. Gray RM (1984) Vector quantization. IEEE Acoust Speech Signal Process 1:4–29
8. Hansen E (1975) A generalized interval arithmetic. Lecture Notes in Computer Science, vol 29. Springer, Berlin, pp 7–18
9. Hartigan JA (1972) Direct clustering of a data matrix. J Am Stat Assoc 67:123–129
10. Jaulin L, Kieffer M, Didrit O, Walter E (2001) Applied Interval Analysis. Springer, London
11. Jolliffe IT (1986) Principal Component Analysis. Springer, New York
12. Kennedy J, Eberhart RC (1995) Particle swarm optimization, vol 4. Proc IEEE Int Conf on Neural Networks. IEEE Press, Piscataway, pp 1942–1948
13. Kohonen T (1989) Self Organization and Associative Memory, 3rd edn. Springer, Berlin
14. Mitchell M (1996) An Introduction to Genetic Algorithms. MIT Press, Cambridge
15. Moore R (1966) Interval Analysis. Prentice Hall, Englewood Cliffs
16. Pawlak Z (1982) Rough sets. Int J Comput Inform Sci 11:341–356
17. Pawlak Z (1991) Rough Sets. Theoretical Aspects of Reasoning About Data. Kluwer, Dordrecht
18. Pawlak Z, Skowron A (2007) Rough sets: some extensions. Inf Sci 177(1):28–40
19. Pedrycz W (ed) (2001) Granular Computing: An Emerging Paradigm. Physica, Heidelberg
20. Pedrycz W (2005) Knowledge-based Clustering. Wiley, Hoboken
21. Warmus M (1956) Calculus of approximations. Bull Acad Pol Sci 4(5):253–259

22. Zadeh LA (1979) Fuzzy sets and information granularity. In: Gupta MM, Ragade RK, Yager RR (eds) Advances in Fuzzy Set Theory and Applications. North Holland, Amsterdam, pp 3–18
23. Zadeh LA (1996) Fuzzy logic = Computing with words. IEEE Trans Fuzzy Syst 4:103–111
24. Zadeh LA (1997) Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets Syst 90:111–117
25. Zadeh LA (1999) From computing with numbers to computing with words-from manipulation of measurements to manipulation of perceptions. IEEE Trans Circ Syst 45:105–119
26. Zadeh LA (2005) Toward a generalized theory of uncertainty (GTU) – an outline. Inf Sci 172:1–40
27. Zimmermann HJ (1996) Fuzzy Set Theory and Its Applications, 3rd edn. Kluwer, Norwell

Books and Reviews

Baldi P, Hornik K (1989) Neural networks and principal component analysis: learning from examples without local minima. Neural Netw 2:53–58
Bortolan G, Pedrycz W (2002) Fuzzy descriptive models: an interactive framework of information granulation. IEEE Trans Fuzzy Syst 10(6):743–755
Busygin S, Prokopyev O, Pardalos PM (2008) Biclustering in data mining. Comput Oper Res 35(9):2964–2987
Fukunaga K (1990) Introduction to Statistical Pattern Recognition. Academic Press, Boston
Gifi A (1990) Nonlinear Multivariate Analysis. Wiley, Chichester
Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323
Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97(1–2):273–324
Lauro C, Palumbo F (2000) Principal component analysis of interval data: a symbolic data analysis approach. Comput Stat 15:73–87
Manly BF, Bryan FJ (1986) Multivariate Statistical Methods: A Primer. Chapman and Hall, London
Mitra P, Murthy CA, Pal SK (2002) Unsupervised feature selection using feature similarity. IEEE Trans Pattern Anal Mach Intell 24(3):301–312
Monahan AH (2000) Nonlinear principal component analysis by neural networks: Theory and applications to the Lorenz system. J Clim 13:821–835
Muni DP, Das Pal NR (2006) Genetic programming for simultaneous feature selection and classifier design. IEEE Trans Syst Man Cybern Part B 36(1):106–117
Pawlak Z (1991) Rough Sets – Theoretical Aspects of Reasoning about Data. Kluwer, Dordrecht
Pedrycz W, Vukovich G (2002) Feature analysis through information granulation and fuzzy sets. Pattern Recognit 35:825–834
Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290(1):2323–2326
Setiono R, Liu H (1977) Neural-network feature selector, IEEE Trans. Neural Netw 8(3):654–662
Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290(1):2319–2323
Watada J, Yabuuchi Y (1997) Fuzzy principal component analysis and its application. Biomed Fuzzy Human Sci 3:83–92

Data-Mining and Knowledge Discovery: Case-Based Reasoning, Nearest Neighbor and Rough Sets

Lech Polkowski[1,2]
[1] Polish–Japanese Institute of Information Technology, Warsaw, Poland
[2] Department of Mathematics and Computer Science, University of Warmia and Mazury, Olsztyn, Poland

Article Outline

Glossary
Definition of the Subject
Introduction
Rough Set Theory: Extensions
Rough Set Theory: Applications
Nearest Neighbor Method
Case-Based Reasoning
Complexity Issues
Future Directions
Bibliography

Glossary

Knowledge This is a many-faceted and difficult notion to define and it is used very frequently without any attempt at definition as a notion that explains per se. One can follow J. M. Bocheński in claiming that the world is a system of states of things, related to themselves by means of the network of relations; things, their features, relations among them and states are reflected in knowledge: things in objects or notions, features and relations in notions (or, concepts), states of things in sentences. Sentences constitute knowledge. Knowledge allows its possessor to classify new objects, model processes, make predictions etc.

Reasoning Processes of reasoning include an effort by means of which sentences are created; various forms of reasoning depend on the chosen system of notions, symbolic representation of notions, forms of manipulating symbols etc.

Knowledge representation This is a chosen symbolic system (language) by means of which notions are encoded and reasoning is formalized.

Boolean functions An n-ary Boolean function is a mapping $f\colon \{0,1\}^n \to \{0,1\}$ from the space of binary sequences of length n into the doubleton $\{0,1\}$. An equivalent representation of the function f is as a formula ϕ_f of propositional calculus; a Boolean function

can be thus represented either in DNF form or in CNF form. The former representation: $\vee_{i\in I} \wedge_{j\in J_i} l^i_j$, where l^i_j is a literal, i. e., either a propositional variable or its negation, is instrumental in Boolean Reasoning: when the problem is formulated as a Boolean function, its solutions are searched for as prime implicants $\bigwedge_{j\in J_i} l^i_j$. Applications are to be found in various reduct induction algorithms.

Information systems One of the languages for knowledge representation is the attribute-value language in which notions representing things are described by means of attributes (features) and their values; information systems are pairs of the form (U, A) where U is a set of objects – representing things – and A is a set of attributes; each attribute a is modeled as a mapping $a\colon U \to V_a$ from the set of objects into the value set V_a. For an attribute a and its value v, the descriptor $(a = v)$ is a formula interpreted in the set of objects U as $[(a = v)] = \{u \in U\colon a(u) = v\}$. Descriptor formulas are the smallest set containing all descriptors and closed under sentential connectives $\vee, \wedge, \neg, \Rightarrow$. Meanings of complex formulas are defined recursively: $[\alpha \vee \beta] = [\alpha] \cup [\beta]$, $[\alpha \wedge \beta] = [\alpha] \cap [\beta]$, $[\neg\alpha] = U \setminus [\alpha]$, $[\alpha \Rightarrow \beta] = [\neg\alpha \vee \beta]$. In descriptor language each object $u \in U$ can be encoded over a set B of attributes as its information vector $\mathrm{Inf}_B(u) = \{(a = a(u))\colon a \in B\}$.

Indiscernibility The Leibnizian Principle of Identity of Indiscernibles affirms that two things are identical in case they are indiscernible, i. e., no available operator acting on both of them yields distinct values; in the context of information systems, indiscernibility relations are induced from sets of attributes: given a set $B \subseteq A$, the indiscernibility relation relative to B is defined as $\mathrm{Ind}(B) = \{(u, u')\colon a(u) = a(u')$ for each $a \in B\}$. Objects u, u' in relation Ind(B) are said to be B-indiscernible and are regarded as identical with respect to knowledge represented by the information system (U, B). The class $[u]_B = \{u'\colon (u, u') \in \mathrm{Ind}(B)\}$ collects all objects identical to u with respect to B.

Exact, inexact notion An exact notion is a set of objects in the considered universe which can be represented as the union of a collection of indiscernibility classes; otherwise, the set is inexact. In this case, there exist a boundary about the notion consisting of objects which can be with certainty classified neither into the notion nor into its complement (Pawlak, Frege).

Decision systems A particular form of an information system, this is a triple (U, A, d) in which d is the decision, the attribute not in A, that does express the evaluation of objects by an external oracle, an expert. Attributes in A are called conditional in order to discern them from the decision d.

Classification task The problem of assigning to each element in a set of objects (test sample) of a class (of a decision) to which the given element should belong; it is effected on the basis of knowledge induced from the given collection of examples (the training sample). To perform this task, objects are mapped usually onto vectors in a multi-dimensional real vector space (feature space).

Decision rule A formula in descriptor language that does express a particular relation among conditional attributes in the attribute set A and the decision d, of the form: $\bigwedge_{a\in A}(a = v_a) \Rightarrow (d = v)$ with the semantics defined in (Glossary: "Indiscernibility"). The formula is true in case $[\bigwedge_{a\in A}(a = v_a)] = \bigcap_{a\in A}[(a = v_a)] \subseteq [(d = v)]$. Otherwise, the formula is partially true. An object o which matches the rule, i. e., $a(o) = v_a$ for $a \in A$ can be classified to the class $[(d = v)]$; often a partial match based on a chosen distance measure has to be performed.

Distance functions (metrics) A metric on a set X is a non-negative valued function $\rho\colon X \times X \to R$ where R is the set of reals, which satisfies conditions: 1. $\rho(x, y) = 0$ if and only if $x = y$. 2. $\rho(x, y) = \rho(y, x)$. 3. $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$ for each z in X (the triangle inequality); when in 3. $\rho(x, y)$ is bound by $\max\{\rho(x, z), \rho(z, y)\}$ instead of by the sum of the two, one says of non-archimedean metric.

Object closest to a set For a metric ρ on a set X, and a subset Y of X, the distance from an object x in X and the set Y is defined as $\mathrm{dist}(x, Y) = \inf\{\rho(x, y)\colon y \in Y\}$; when Y is a finite set, then infimum inf is replaced with minimum min.

A nearest neighbor For an object $x_0 \in X$, this is an object $n(x_0)$ such that $\rho(x_0, n(x_0)) = dist(x_0, X \setminus \{x_0\})$; $n(x_0)$ may not be unique. In plain words, $n(x_0)$ is the object closest to x_0 and distinct from it.

K-nearest neighbors For an object $x_0 \in X$, and a natural number $K \geq 1$, this is a set $n(x_0, K) \subseteq X \setminus \{x_0\}$ of cardinality K such that for each object $y \in X \setminus [n(x_0, K) \cup \{x_0\}]$ one has $\rho(y, x_0) \geq \rho(z, x_0)$ for each $z \in n(x_0, K)$. In plain words, objects in $n(x_0, K)$ for a set of K objects that are closest to x_0 among objects distinct from it.

Rough inclusion A ternary relation μ on a set $U \times U \times [0, 1]$ which satisfies conditions: 1. $\mu(x, x, 1)$. 2. $\mu(x, y, 1)$ is a binary partial order relation on the set X. 3. $\mu(x, y, 1)$ implies that for each object z in U: if $\mu(z, x, r)$ then $\mu(z, y, r)$. 4. $\mu(x, y, r)$ and $s < r$ imply that $\mu(x, y, s)$. The formula $\mu(x, y, r)$ is read as "the object x is a part in object y to a degree at least r".

The partial containment idea does encompass the idea of an exact part, i. e., mereological theory of concepts (Leśniewski, Leonard–Goodman).

Similarity An extension and relaxing of an equivalence relation of indiscernibility. Among similarity relations, we single out the class of tolerance relations τ (Poincaré, Zeeman) which are reflexive, i. e., $\tau(x, x)$, and symmetric, i. e., $\tau(x, y)$ implies necessarily $\tau(y, x)$. The basic distinction between equivalences and tolerances is that the former induce partitions of their universes into disjoint classes whereas the latter induce covering of their universes by their classes; for this reason they are more difficult in analysis. The symmetry condition may be over-restrictive, for instance, rough inclusions are usually not symmetric.

Case A case is informally a part of knowledge that describes a certain state of the world (the context), along with a query (problem) and its solution, and a description of the outcome, i. e., the state of the world after the solution is applied.

Retrieve A process by means of which a case similar in a sense to the currently considered is recovered from the case base.

Reuse A process by means of which the retrieved case's solution is re-used in the current new problem.

Revise A process by means of which the retrieved solution is adapted in order to satisfactorily solve the current problem.

Retain A process by means of which the adapted solution is stored in the case base as the solution to the current problem.

Definition of the Subject

Basic ideas of Rough Set Theory were proposed by Zdzisław Pawlak [60,62], as a formal set of notions aimed at carrying out tasks of reasoning, in particular about classification of objects, in conditions of uncertainty. Conditions of uncertainty are imposed by incompleteness, imprecision and ambiguity of knowledge. Originally, the basic notion proposed was that of a knowledge base, understood as a collection $\mathcal{R}$ of equivalence relations on a universe of objects; each relation $r \in \mathcal{R}$ induces on the set U a partition ind_r into equivalence classes. Knowledge so encoded is meant to represent the classification ability. As objects for analysis and classification come most often in the form of data, a useful notion of an information system is commonly used in knowledge representation; knowledge base in that case is defined as the collection of indiscernibility relations. Exact concepts are defined as unions of indiscernibility classes whereas inexact concepts are only approximated from below (lower approximations) and from above (upper approximations) by exact ones. Each inexact concept is thus perceived as a pair of exact concepts between which it is sandwiched. In terms of indiscernibility it is possible to define dependencies among attributes: the best case, of functional dependence, comprises instances in which a set say B of attributes depends on a set C of attributes, meaning that values of attributes from B are function-dependent on values of attributes in C; a weaker form of dependence is partial dependence in which there is a non-functional relation between values of attributes in B and those of C. The most important case of dependency is the one between conditional attributes in a decision system and the decision; these dependencies are expressed by means of decision rules which altogether constitute a classifier. From the application point of view, one of most important tasks is to induce from data a satisfactory classifier. A plethora of ideas were put forth in this area. First, the idea of knowledge reduction appeared; here, the notion of a reduct in an information system comes foremost: a reduct $B \subseteq A$ is a minimal with respect to inclusion subset of A such that indiscernibility induced by B is the indiscernibility induced by A. Finding a reduct means that one can dispense with attributes in $A \setminus B$ and induce decision rules from conditional attributes restricted to B. Further developments in reduction area encompass notions of a relative reduct, a local reduct, a dynamic reduct, an entropy reduct and so on. Subsequent developments include discretization algorithms, similarity measures, granulation of knowledge.

Research on classifiers has led to the emergence of a number of systems for rule induction that have been applied in many tasks in areas of, e. g., data classification, medical diagnosis, economic analysis, signal classification; in some cases rough set classifiers have been coupled with fuzzy, wavelet or neural reasoning systems to produce hybrid systems.

Matching rules to new objects is done on the basis of closeness of the object and the rule measured often in terms of common descriptor fraction in the object description and in the rule and in many classifiers it is the closest rule that sets the value of decision class to the object.

The idea of value assignment to a new object by the closest already classified object is at the heart of the Nearest Neighbor Method. The idea goes back to Fix and Hodges [21,22], Skellam [88] and Clark and Evans [12]. Nearest neighbor technique in its simplest form consists of imposing a metric ρ on the object set in question and assigning the class value $c(x_0)$ to a currently examined test object x_0 by selecting the closest to x_0 with re-

spect to the metric ρ training object $n(x_0)$ and letting $c(x_0) = c(n(x_0))$, the class already assigned to $n(x_0)$. This simplest variant is also named 1-nearest neighbor method. A generalization consists of a choice of a natural number k and selecting from among training objects the set of k closest to x_0 objects, $n(x_0) = \{x_1, .., x_k\}$, eventually breaking ties randomly. Then the class value $c(x_0)$ is defined on the basis of class values $c(x_1), \ldots, c(x_k)$ by majority voting; eventual ties are broken randomly. Nearest neighbor methods can be perceived as heuristics implementing the idea of the Parzen window: a region about x_0 (the "window"), e. g., rectangular, in which the count of objects leads to an estimate of density of probability distribution of objects in space. The problem of choosing the "right" window is in case of nearest neighbor techniques delegated to the training set itself and the class value is assigned without any probability assessment. Nearest neighbor idea is a basis as well for prototype methods, in particular k-means and clustering methods and is also applied in other approaches, e.g, rough set classifiers. Applications of this technique can be found in many areas like satellite images analysis, plant ecology, forestry, cluster identification in medical diagnosis etc.

Difficulties in extracting from data an explicit model on basis of which one could make inferences, already observed with the nearest neighbor method, led to emergence of the ideology of Case-Based Reasoning, traced back to Schank and Abelson's [84] discussion of methods for describing the knowledge about situations. Closely related to reasoning by analogy, experiential learning, and philosophical and psychological theories of concepts claiming that real-world concepts are best classified by sets of representative cases (Wittgenstein [117]), Case-Based Reasoning emerged as a methodology for problem solving based on cases, i. e., capsules of knowledge containing the description of the current state of the world at the moment the problem is posed (the situation), the solution to the problem, and the state of the world after the solution is implemented. Solved cases stored in case base form a repository to which new cases are addressed. A new case is matched against the case base and cases in the case base most similar to the given case are retrieved; retrieval is usually based on a form of analogy reasoning, nearest neighbor selection, inductive procedures, or template retrieval. Once the similar case is retrieved, its solution is reused, or adapted, to provide a solution for the new case; various strategies are applied in the reuse process, from the simplest null strategy, i. e., using directly the retrieved solution in the new problem to derivational strategies that take the procedure that generated the retrieved solution and modify it to yield a new solution to the current problem. Finally, the new completed case is retained (or, stored) in the case base. Case-based reasoning has been applied in many systems in the area of expert systems, to mention CYRUS (Kolodner [37]), MEDIATOR (Simpson [87]), PERSUADER (Sycara [106]) and many others (Watson and Marir [114]). CBR systems have found usage in many areas of expertise like complex diagnosis, legal procedures, civil engineering, manufacture planning.

Introduction

Three paradigms covered in this article have been arranged in the previous section in the order corresponding to the specificity of their scope: rough sets are a paradigm intended to reason under uncertainty from data and they exploit the idea of the nearest neighbor in common with many other paradigms; rough set techniques make use of metrics and similarity relations in their methods and consequently we discuss nearest neighbor method as based on metrics and finally case-based reasoning as based on similarity ideas.

Introduced in Pawlak [60] rough set theory is based on ideas that go back to Gottlob Frege, Gottfried Wilhelm Leibniz, Jan Łukasiewicz, Stanislaw Leśniewski, to mention a few names of importance. Its characteristics are that they divide notions (concepts) into two classes: exact as well as inexact. Because of the Fregean idea (Frege [23]), an inexact concept should possess a boundary into which objects that can be classified with certainty neither to the concept nor to its complement fall. This boundary to a concept is constructed from indiscernibility relations induced by attributes (features) of objects (see Glossary: "Indiscernibility"). Given a set of objects U and attributes in a set A, each set B of attributes does induce the B-indiscernibility relation $ind(B)$ (see "Glossary: Indiscernibility"), according to the Leibnizian Principle of Identity of Indiscernibles, see [98]. Each class $[u]_B = \{v \in U : (u, v) \in \mathrm{ind}(B)\}$ of $\mathrm{ind}(B)$ is B-definable, i. e., the decision problem whether $v \in [u]_B$ is decidable: $[u]_B$ is exact (see "Glossary"). Unions of classes $[u]_B$ for some u, B are exact as well. Concepts $\subseteq U$ that are not of this form are inexact, they possess a non-empty boundary. To express the B-boundary of a concept X induced by the set B of attributes, approximations over B are introduced, i. e.,

$$\underline{B}X = \bigcup\{[u]_B : [u]_B \subseteq X\} \quad \text{(the } B\text{-lower approximation)}$$

$$\overline{B}X = \bigcup\{[u]_B : [u]_B \cap X \neq \emptyset\} \quad \text{(the } B\text{-upper approximation)}\,.$$

The difference $Bd_B X = \overline{B}X \setminus \underline{B}X$ is the B-boundary of X; when non-empty it does witness that X is inexact over B.

The information system (U, A) (see Glossary: "Information systems") does represent knowledge about a certain aspect of the world. This knowledge can be reduced: a reduct B of the set A of attributes is a minimal subset of A with the property that $\mathrm{ind}(B) = \mathrm{ind}(A)$. An algorithm for finding reducts based on Boolean reasoning was proposed in Skowron and Rauszer [90]; given input (U, A) with $U = \{u_1, \ldots, u_n\}$ it starts with the discernibility matrix,

$$M_{U,A} = [c_{i,j} = \{a \in A\colon a(u_i) \neq a(u_j)\}]_{1 \geq i,j \leq n}$$

and builds the Boolean function,

$$f_{U,A} = \bigwedge_{c_{i,j} \neq \emptyset, i<j} \bigvee_{a \in c_{i,j}} \overline{a}\,,$$

where $\overline{a}$ is the Boolean variable assigned to the attribute $a \in A$.

The function $f_{U,A}$ is converted to its DNF form:

$$f^*_{U,A}\colon \bigvee_{j \in J} \bigwedge_{k \in K_j} \overline{a_{j,k}}\,.$$

Then: sets of the form $R_j = \{a_{j,k}\colon k \in K_j\}$ for $j \in J$, corresponding to prime implicants (see Glossary: "Boolean functions"), are all reducts of A. Choosing a reduct R, and forming the reduced information system (U, R) one is assured that no information encoded in (U, A) has been lost.

Moreover, as $\mathrm{ind}(R) \subseteq \mathrm{ind}(B)$ for any subset B of the attribute set A, one can establish a functional dependence of B on R: as for each object $u \in U$, $[u]_R \subseteq [u]_B$, the assignment $f_{R,B}\colon \mathrm{Inf}_R(u) \to \mathrm{Inf}_B(u)$ (see Glossary: "Information systems") is functional, i. e., values of attributes in R determine functionally values of attributes on U, object-wise. Thus, any reduct determines functionally the whole data system.

A decision system (U, A, d) (see Glossary: "Decision systems") encodes information about the external classification d (by an oracle, expert etc.). Methods based on rough sets aim as well at finding a description of the concept d in terms of conditional attributes in A in the language of descriptors (see Glossary: "Information systems"). The simpler case is when the decision system is deterministic, i. e., $\mathrm{ind}(A) \subseteq \mathrm{ind}(d)$. In this case the relation between A and d is functional, given by the unique assignment $f_{A,d}$ or in the decision rule form (see Glossary: "Decision rule"), as the set of rules: $\bigwedge_{a \in A}(a = a(u)) \Rightarrow (d = d(u))$. In place of A any reduct R of A can be substituted leading to shorter rules. In the contrary case, some classes $[u]_A$ are split into more than one decision class $[v]_d$ leading to ambiguity in classification. In that case, decision rules are divided into certain (or, exact) and possible; to induce the certain rule set, the notion of a δ-reduct was proposed in Skowron and Rauszer [90]; it is called a relative reduct in Bazan et al. [7]. To define δ-reducts, first the generalized decision δ_B is defined: for $u \in U$, $\delta_B(u) = \{v \in V_d\colon d(u') = v \wedge (u, u') \in \mathrm{ind}(B)$ for some $u' \in U\}$. A subset B of A is a δ-reduct to d when it is a minimal subset of A with respect to the property that $\delta_B = \delta_A$.

δ-reducts can be obtained from the modified Skowron and Rauszer algorithm [90]: it suffices to modify the entries $c_{i,j}$ to the discernibility matrix, by letting $c^d_{i,j} = \{a \in A \cup \{d\}\colon a(u_i) \neq a(u_j)\}$ and then setting $c'_{i,j} = c^d_{i,j} \setminus \{d\}$ in case $d(u_i) \neq d(u_j)$ and $c'_{i,j} = \emptyset$ in case $d(u_i) = d(u_j)$. The algorithm described above input with entries $c'_{i,j}$ forming the matrix $M^\delta_{U,A}$ outputs all δ-reducts to d encoded as prime implicants of the associated Boolean function $f^\delta_{U,A}$.

An Example of Reduct Finding and Decision Rule Induction

We conclude the first step into rough sets with a simple example of a decision system, its reducts and decision rules.

Table 1 shows a simple decision system.

Reducts of the information system $(U, A = \{a_1, a_2, a_3, a_4\})$ can be found from the discernibility matrix $M_{U,A}$ in Table 2; by symmetry, cells $c_{i,j} = c_{j,i}$ with $i > j$ are not filled. Each attribute a_i is encoded by the Boolean variable i.

After reduction by means of absorption rules of sentential calculus: $(p \vee q) \wedge p \Leftrightarrow p$, $(p \wedge q) \vee p \Leftrightarrow p$, the DNF form $f^*_{U,A}$ is $1 \wedge 2 \wedge 3 \vee 1 \wedge 2 \wedge 4 \vee 1 \wedge 3 \wedge 4$. Reducts of A in the information system (U, A) are: $\{a_1, a_2, a_3\}$, $\{a_1, a_2, a_4\}$, $\{a_1, a_3, a_4\}$.

δ-reducts of the decision d in the decision system Simple, can be found from the modified discernibility matrix $M^\delta_{U,A}$ in Table 3.

Data-Mining and Knowledge Discovery: Case-Based Reasoning, Nearest Neighbor and Rough Sets, Table 1
Decision system Simple

Obj.	a_1	a_2	a_3	a_4	d
u_1	1	0	0	1	0
u_2	0	1	0	0	1
u_3	1	1	0	0	1
u_4	1	0	0	1	1
u_5	0	0	0	1	1
u_6	1	1	1	1	0

Data-Mining and Knowledge Discovery: Case-Based Reasoning, Nearest Neighbor and Rough Sets, Table 2
Discernibility matrix $M_{U,A}$ for reducts in (U,A)

Obj.	u_1	u_2	u_3	u_4	u_5	u_6
u_1	$\emptyset$	$\{1,2,4\}$	$\{2,4\}$	$\emptyset$	$\{1\}$	$\{2,3\}$
u_2	–	$\emptyset$	$\{1\}$	$\{1,2,3\}$	$\{2,4\}$	$\{1,3,4\}$
u_3	–	–	$\emptyset$	$\{2,4\}$	$\{2,4\}$	$\{3,4\}$
u_4	–	–	–	$\emptyset$	$\{1\}$	$\{2,3\}$
u_5	–	–	–	–	$\emptyset$	$\{1,2,3\}$
u_6	–	–	–	–	–	$\emptyset$

Data-Mining and Knowledge Discovery: Case-Based Reasoning, Nearest Neighbor and Rough Sets, Table 3
Discernibility matrix $M^{\delta}_{U,A}$ for δ-reducts in (U, A, d)

Obj.	u_1	u_2	u_3	u_4	u_5	u_6
u_1	$\emptyset$	$\{1,2,4\}$	$\{2,4\}$	$\emptyset$	$\{1\}$	$\emptyset$
u_2	–	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\{1,3,4\}$
u_3	–	–	$\emptyset$	$\emptyset$	$\emptyset$	$\{3,4\}$
u_4	–	–	–	$\emptyset$	$\emptyset$	$\{2,3\}$
u_5	–	–	–	–	$\emptyset$	$\{1,2,3\}$
u_6	–	–	–	–	–	$\emptyset$

From the Boolean function $f^{\delta}_{U,A}$ we read off δ-reducts $R_1 = \{a_1, a_2, a_3\}$, $R_2 = \{a_1, a_2, a_4\}$, $R_3 = \{a_1, a_3, a_4\}$.

Taking R_1 as the reduct for inducing decision rules, we read the following certain rules:

$$r_1 : (a_1 = 0) \wedge (a_2 = 1) \wedge (a_3 = 0) \Rightarrow (d = 1)\,;$$
$$r_2 : (a_1 = 1) \wedge (a_2 = 1) \wedge (a_3 = 0) \Rightarrow (d = 1)\,;$$
$$r_3 : (a_1 = 0) \wedge (a_2 = 0) \wedge (a_3 = 0) \Rightarrow (d = 1)\,;$$
$$r_4 : (a_1 = 1) \wedge (a_2 = 1) \wedge (a_3 = 1) \Rightarrow (d = 0)\,;$$

and two possible rules

$$r_5 : (a_1 = 1) \wedge (a_2 = 0) \wedge (a_3 = 0) \Rightarrow (d = 0)\,;$$
$$r_6 : (a_1 = 1) \wedge (a_2 = 0) \wedge (a_3 = 0) \Rightarrow (d = 1)\,,$$

each with certainty factor $= .5$ as there are two objects with $d = 0$.

Consider a new object v: $\mathrm{Inf}_A(v) = \{(a_1 = 0), (a_2 = 1), (a_3 = 1), (a_4 = 0)\}$.

The Nearest Neighbor Approach

We define the metric ρ on objects encoded by means of their information vectors (see Glossary: "Information systems") as the reduced Hamming metrics on information vectors over the reduct R_1; for an object x we let $\mathrm{Desc}(x)$ to denote the set of descriptors that occur in the information vector $\mathrm{Inf}_{R_1}(x)$ of x. The metric is then $\rho_{R_1}(x, y) = 1 - (|\,\mathrm{Desc}(x) \cap \mathrm{Desc}(y)|)/3$; the nearest neighbors of v with respect to ρ_{R_1} are u_2 and u_6 described by, respectively, rules r_1, r_4; however, these rules point to distinct decision classes: 1, 0. The majority voting must be implemented by random choice between class values 0, 1 with probability of error of .5. We can still resort to the idea of the most similar case.

The Case-Based Reasoning Approach

This ambiguity may be resolved if we treat objects as cases defined over the set A with solutions as the values of the decision d; thus formally a case is of the form $(x, d(x))$. Similarity among cases will be defined by means of the Hamming distance reduced over the whole set A; thus, $\rho_A(x, y) = 1 - (|\,\mathrm{Desc}(x) \cap \mathrm{Desc}(y)|)/4$. The most similar case to v is u_6 ($\rho_A(v, u_6) = 1/4$, $\rho(v, u_2) = 1/2$) hence we adapt the solution $d = 0$ to the case u_6 as the solution to v and we classify v as $d = 0$; the new case in our case base is then $(v, 0)$.

Rough Set Theory: Extensions

Basic principles of the rough set approach: knowledge reduction, indiscernibility, functional and partial dependence, decision rules are exposed in Sect. "Introduction". In this section, we augment our discussion with more detailed and specific aspects, dividing the exposition into parts concerned with particular aspects.

Reducts

The notion of a reduct as a minimal with respect to set inclusion subset of the set of attributes A preserving a certain property of the whole attribute set, has undergone an evolution from the basic notions of a reduct and δ-reduct to more specialized variants. The problem of finding a reduct of minimal length is NP-hard [90], therefore one may foresee that no polynomial algorithm is available for computing reducts. Thus, the algorithm based on the discernibility matrix has been proposed with stop rules that permits one to stop the algorithm and obtain a partial set of reducts [7]. In order to find stable reducts, robust to perturbations in data, the idea of a dynamic reduct was proposed in Bazan et al. [7]; for a decision system $D = (U, A, d)$, a family F of subsystems of the form (U_1, A, d) with $U_1 \subseteq U$ is derived (e. g., by random choice) and given $\varepsilon \in (0, 1)$, a generalized dynamic ε-reduct B of D is a set B which is also found as a reduct in at least $(1 - \varepsilon)$-fraction of subsystems in F.

In order to precisely discriminate between certain and possible rules, the notion of a positive region along with

the notion of a relative reduct was proposed and studied in Skowron and Rauszer [90].

Positive region $\mathrm{pos}_B(d)$ is the set $\{u \in U\colon [u]_B \subseteq [u]_d\} = \bigcup_{v\in V_d} \underline{B}[(d = v)]$; $\mathrm{pos}_B(d)$ is the greatest subset of X of U such that (X, B, d) is deterministic; it generates certain rules. Objects in $U \setminus \mathrm{pos}_B(d)$ are subjected to ambiguity: given such u, and the collection $v_1, .., v_k$ of decision d values on the class $[u]_B$, the decision rule describing u can be formulated as, $\bigwedge_{a\in B}(a = a(u)) \Rightarrow \bigvee_{i=1,\ldots,k}(d = v_i)$; each of the rules $\bigwedge_{a\in B}(a = a(u)) \Rightarrow (d = v_i)$ is possible but not certain as only for a fraction of objects in the class $[u]_B$ the decision take the value v_i on.

Relative reducts are minimal sets B of attributes with the property that $\mathrm{pos}_B(d) = \mathrm{pos}_A(d)$; they can also be found by means of discernibility matrix $M^*_{U,A}$ [90]: $c^*_{i,j} = c^d_{i,j} \setminus \{d\}$ in case either $d(u_i) \neq d(u_j)$ and $u_i, u_j \in \mathrm{pos}_A(d)$ or $\mathrm{pos}(u_i) \neq \mathrm{pos}(u_j)$ where pos is the characteristic function of $\mathrm{pos}_A(d)$; otherwise, $c^*_{i,j} = \emptyset$.

For a relative reduct B, certain rules are induced from the deterministic system $(\mathrm{pos}_B(d), A, d)$, possible rules are induced from the non-deterministic system $(U \setminus \mathrm{pos}_B(d), A, d)$. In the last case, one can find δ-reducts to d in this system and turn the system into a deterministic one $(U \setminus \mathrm{pos}_B(d), A, \delta)$ inducing certain rules of the form $\bigwedge_{a\in B}(a = a(u)) \Rightarrow \bigvee_{v\in\delta(u)}(d = v)$.

Localization of reducts was taken care of by means of the notion of a local reduct in [7]: given an object u_i, a local reduct to decision d is a minimal set $B(u_i)$ such that for each object u_j, if $d(u_i) \neq d(u_j)$ then $a(u_i) \neq a(u_j)$ for some $a \in B(u_i)$. An algorithm for finding a covering of the set of objects by local reducts for each of the objects is given in [7].

A more general notion of a reduct as a minimal set of attributes preserving a certain characteristic of the distribution of objects into decision classes, e. g., frequencies of objects, entropy of classification distribution, etc., is discussed in [94].

Decision Rules

Already defined as descriptor formulas of the form $\bigwedge_{a\in B}(a = v_a) \Rightarrow (d = v)$ in descriptor language of the decision system (U, A, d), decision rules provide a description of the decision d in terms of conditional attributes in the set A. Forming a decision rule is searching in the pool of available semantically non-vacuous descriptors for their combination that describes as well as possible a chosen decision class. The very basic idea of inducing rules consists of considering a set B of attributes: the lower approximation $\mathrm{pos}_B(d)$ allows for rules which are certain, the upper approximation $\bigcup_{v\in V_d} \overline{B}[(d = v)]$ adds rules which are possible.

We write down a decision rule in the form $\phi/B, u \Rightarrow (d = v)$ where ϕ/B is a descriptor formula $\bigwedge_{a\in B}(a = a(u))$ over B. A method for inducing decision rules in the systematic way of Pawlak and Skowron [63] and Skowron [89] consists of finding the set of all δ-reducts $\mathbf{R} = \{R_1, \ldots, R_m\}$, and defining for each reduct R_j and each object $u \in U$, the rule $\phi/R_j, u \Rightarrow (d = d(u))$. Rules obtained by this method are not minimal usually in the sense of the number of descriptors in the premise ϕ.

A method for obtaining decision rules with a minimal number of descriptors [63,89], consists of reducing a given rule $r\colon \phi/B, u \Rightarrow (d = v)$ by finding a set $R_r \subseteq B$ consisting of irreducible attributes in B only, in the sense that removing any $a \in R_r$ causes the inequality $[\phi/R_r, u \Rightarrow (d = v)] \neq [\phi/R_r \setminus \{a\}, u \Rightarrow (d = v)]$ to hold. In case $B = A$, reduced rules $\phi/R_r, u \Rightarrow (d = v)$ are called optimal basic rules (with a minimal number of descriptors). The method for finding all irreducible subsets of the set A (Skowron [89]) consists of considering another modification of the discernibility matrix: for each object $u_k \in U$, the entry $c'_{i,j}$ into the matrix $M^\delta_{U,A}$ for δ-reducts is modified into $c^k_{i,j} = c'_{i,j}$ in case $d(u_i) \neq d(u_j)$ and $i = k \vee j = k$, otherwise $c^k_{i,j} = \emptyset$. Matrices $M^k_{U,A}$ and associated Boolean functions $f^k_{U,A}$ for all $u_k \in U$ allow for finding all irreducible subsets of the set A and in consequence all basic optimal rules (with a minimal number of descriptors).

Decision rules are judged by their quality on the basis of the training set and by quality in classifying new as yet unseen objects, i. e., by their performance on the test set. Quality evaluation is done on the basis of some measures: for a rule $r\colon \phi \Rightarrow (d = v)$, and an object $u \in U$, one says that u matches r in case $u \in [\phi]$. $\mathrm{match}(r)$ is the number of objects matching r. Support $\mathrm{supp}(r)$ of r is the number of objects in $[\phi] \cap [(d = v)]$; the fraction $\mathrm{cons}(r) = (\mathrm{supp}(r))/(\mathrm{match}(r))$ is the consistency degree of r $\mathrm{cons}(r) = 1$ and means that the rule is certain.

Strength, $\mathrm{strength}(r)$, of the rule r is defined (Michalski et al. [49], Bazan [6], Grzymala-Busse and Ming Hu [29]), as the number of objects correctly classified by the rule in the training phase; relative strength is defined as the fraction $\mathrm{rel-strength}(r) = (\mathrm{supp}(r))/(|[(d = v)]|)$. Specificity of the rule r, $\mathrm{spec}(r)$, is the number of descriptors in the premise ϕ of the rule r [29].

In the testing phase, rules vie among themselves for object classification when they point to distinct decision classes; in such a case, negotiations among rules or their sets are necessary. In these negotiations rules with better characteristics are privileged.

For a given decision class $c\colon d = v$, and an object u in the test set, the set $\mathrm{Rule}(c, u)$ of all rules matched by u and pointing to the decision v, is characterized globally by $\mathrm{Support}(\mathrm{Rule}(c, u)) = \sum_{r \in \mathrm{Rule}(c,u)} \mathrm{strength}(r) \cdot \mathrm{spec}(r)$. The class c for which $\mathrm{Support}(\mathrm{Rule}(c, u))$ is the largest wins the competition and the object u is classified into the class $c\colon d = v$ [29].

It may happen that no rule in the available set of rules is matched by the test object u and partial matching is necessary, i. e., for a rule r, the matching factor $\mathrm{match-fact}(r, u)$ is defined as the fraction of descriptors in the premise ϕ of r matched by u to the number $\mathrm{spec}(r)$ of descriptors in ϕ. The rule for which the partial support $\mathrm{Part-Support}(\mathrm{Rule}(c, u)) = \sum_{r \in \mathrm{Rule}(c,u)} \mathrm{match-fact}(r, u) \cdot \mathrm{strength}(r) \cdot \mathrm{spec}(r)$ is the largest wins the competition and it does assign the value of decision to u [29].

In a similar way, notions based on relative strength can be defined for sets of rules and applied in negotiations among them [7].

Dependency

Decision rules are particular cases of dependencies among attributes or their sets; certain rules of the form $\phi/B \Rightarrow (d = v)$ establish functional dependency of decision d on the set B of conditional attributes. Functional dependence of the set B of attributes on the set C, $C \mapsto B$, in an information system (U, A) means that $\mathrm{Ind}(C) \subseteq \mathrm{Ind}(B)$. Minimal sets $D \subseteq C$ of attributes such that $D \mapsto B$ can be found from a modified discernibility matrix $M_{U,A}$: letting $\langle B \rangle$ to denote the global attribute representing B: $\langle B \rangle(u) = \langle b_1(u), \ldots, b_m(u) \rangle$ where $B = \{b_1, \ldots, b_m\}$, for objects u_i, u_j, one sets $c_{i,j} = \{a \in C \cup \{\langle B \rangle\}\colon a(u_i) \neq a(u_j)\}$ and then $c^B_{i,j} = c_{i,j} \setminus \{\langle B \rangle\}$ in case $\langle B \rangle$ is in $c_{i,j}$; otherwise $c^B_{i,j}$ is empty. The associated Boolean function $f^B_{U,A}$ gives all minimal subsets of C on which B depends functionally; in particular, when $B = \{b\}$, one obtains in this way all subsets of the attribute set A on which b depends functionally (Skowron and Rauszer [90]). A number of contributions by Pawlak, Novotny and Pawlak and Novotny are devoted to this topic in an abstract setting of semi-lattices: Novotny and Pawlak [56], Pawlak [61].

Partial dependence of the set B on the set C of attributes takes place where there is no functional dependence $C \mapsto B$; in that case, some measures of a degree to which B depends on C were proposed in Novotny and Pawlak [55]: the degree can be defined, e. g., as the fraction $\gamma_{B,C} = (|\mathrm{pos}_C B|)/|U|$ where the C-positive region of B is defined in analogy to already discussed positive region for decision, i. e., $\mathrm{pos}_C(B) = \{u \in U\colon [u]_C \subseteq [u]_B\}$; then, B depends on C partially to the degree $\gamma_{B,C}\colon C \mapsto_{\gamma_{B,C}} B$. The relation $C \mapsto_r B$ of partial dependency is transitive in the sense: if $C \mapsto_r B$ and $D \mapsto_s C$, then $D \mapsto_{\max\{0, r+s-1\}} B$, where $t(r, s) = \max\{0, r + s - 1\}$ is the Łukasiewicz t-norm (or, tensor product) [55].

A logical characterization of functional dependence between attribute sets was given as a fragment of intuitionistic logic in Rauszer [80].

Similarity

Analysis based on indiscernibility has allowed for extracting the most important notions of rough set theory; further progress has been obtained inter alia by departing from indiscernibility to more general similarity relations (see Glossary: "Similarity"). There have been various methods for introducing similarity relations.

An attempt at introducing some degrees of comparison among objects with respect to particular concepts consisted of defining rough membership functions in Pawlak and Skowron [64]: for an object u, an attribute set B and a concept $X \subseteq U$, the value $\mu_{B,X}(u)$ of the rough membership function $\mu_{B,X}$ was defined as $\mu_{B,X}(u) = (|[u]_B \cap X|)/(|[u]_B|)$. Informally, one can say that objects u, v are ε, B-similar with respect to the concept X in case $|\mu_{B,X}(u) - \mu_{B,X}(v)| < \varepsilon$. This relation is reflexive and symmetric, i. e., it is a tolerance relation, see Poincaré [66] and Zeeman [125], for each ε, B, X.

Tolerance relations were introduced into rough sets in Polkowski, Skowron, and Żytkow [77] and also studied in Krawiec et al. [43], Skowron and Stepaniuk [91]. It is possible to build a parallel theory of rough sets based on tolerance or similarity relations in analogy to indiscernibility relations. $\mu_{B,X}$ does characterize partial containment of objects in U into concepts X; a further step consists of considering general relations of partial containment in the form of predicates "to be a part to a degree".

A general form of partial containment was proposed as an extension of mereological theory of concepts by Leśniewski [44]; mereology takes the predicate "to be a part of" as its primitive notion, requiring of it to be irreflexive and transitive on the universe of objects U. The primitive notion of a (proper) part is relaxed to the notion of an ingredient (an improper part) $\mathrm{ing} = \mathrm{part} \cup =$.

The extension consists of considering the predicate "to be a part to a degree of", formally introduced as the generic notion of a rough inclusion μ in Polkowski and Skowron [74,75], a ternary predicate with the semantic domain of $U \times U \times [0, 1]$ and satisfying the requirements that (1) $\mu(u, v, 1)$ is a binary relation of ingredient on objects in U, i. e., $\mu(u, v, 1)$ sets an exact scheme of decom-

position of objects in U into (possibly improper) parts. (2) $\mu(u, v, 1)$ implies informally that the object u is "inside" object v hence for any other object w: $\mu(w, u, r)$ implies $\mu(w, v, r)$. (3) $\mu(u, v, r)$ implies $\mu(u, v, s)$ for any $s < r$.

Similarity of objects with respect to μ can be judged by the degree $\mathrm{sim}(u, v) = \max\{arg\max \mu(u, v, r), arg\max \mu(v, u, s)\}$; $\mathrm{sim}(u, v) = 1$ means that u, v are identical with respect to the parts; the smaller $\mathrm{sim}(u, v)$ the less similar u, v are.

Rough inclusions in an information system (U, A) can be induced in some distinct ways as in Polkowski [67,68,69,71]; Archimedean t-norms, i. e., t-norms $t(x, y)$ that are continuous and have no idempotents, i. e., values x with $t(x, x) = x$ except $0, 1$ offer one way; it is well-known from Mostert and Shields [50] and Faucett [18] that up to isomorphism, there are two Archimedean t-norms: the Łukasiewicz t-norm $L(x, y) = \max\{0, x + y - 1\}$ and the product (Menger) t-norm $P(x, y) = x \cdot y$. Archimedean t-norms admit a functional characterization shown in Ling [48]: $t(x, y) = g(f(x) + f(y))$, where the function $f\colon [0, 1] \to R$ is continuous decreasing with $f(1) = 0$, and $g\colon R \to [0, 1]$ is the pseudo-inverse to f, i. e., $f \circ g = id$. The t-induced rough inclusion μ_t is defined [68] as $\mu_t(u, v, r) \Leftrightarrow g((|\mathrm{DIS}(u, v)|)/(|A|)) \geq r$ where $\mathrm{DIS}(u, v) = \{a \in A\colon a(u) \neq a(v)\}$. With the Łukasiewicz t-norm, $f(x) = 1 - x = g(x)$ and $\mathrm{IND}(u, v) = U \times U \setminus \mathrm{DIS}(u, v)$, the formula becomes: $\mu_L(u, v, r) \Leftrightarrow (|\mathrm{IND}(u, v)|)/(|A|) \geq r$; thus in the case of Łukasiewicz logic, μ_L becomes the similarity measure based on the Hamming distance between information vectors of objects reduced modulo $|A|$; from a probabilistic point of view it is based on the relative frequency of descriptors in information sets of u, v. This formula permeates data mining algorithms and methods, see Klösgen and Żytkow [36].

Another method exploits residua of t-norms [69]. For a continuous t-norm $t(x, y)$, the *residuum*, $x \Rightarrow_t y$ is defined as $\max\{z\colon t(x, z) \leq y\}$. Rough inclusions are defined in this case with respect to objects selected as standards, e. g., objects u representing desired (in the given domain) combinations of attribute values. Given a standard s, the rough inclusion $\nu_t^{\mathrm{IND},s}$ is defined:

$$\nu_t^{\mathrm{IND},s}(u, v, r) \text{ iff } \frac{|\mathrm{IND}(u, s)|}{|A|} \Rightarrow \frac{|\mathrm{IND}(v, s)|}{|A|} \geq r\,.$$

In case of the t-norm L, $\nu_L^{\mathrm{IND},s}(x, y, r)$ iff $|\mathrm{IND}(y, s)| - |\mathrm{IND}(x, s)| \geq (1 - r)|A|$.

In case of the t-norm P, $\nu_P^{\mathrm{IND},s}(x, y, 1)$ iff $|\mathrm{IND}(x, s)| \leq |\mathrm{IND}(y, s)|$, and $\nu_P^{\mathrm{IND},s}(x, y, r)$ iff $|\mathrm{IND}(x, s)| \geq |\mathrm{IND}(y, s)| \geq r \cdot |A|$.

Finally, in the case of $t = \min$, we have $\nu_{\min}^{\mathrm{IND},s}(x, y, r)$ iff $|\mathrm{IND}(y, s)|/|\mathrm{IND}(x, s)| \geq r$.

All these rough inclusions satisfy the transitivity property [68,71]: From $\mu(u, v, r)$ and $\mu(v, w, s)$ it follows that $\mu(u, w, t(r, s))$, where μ is induced by the t-norm t.

Each metric d on the set of objects U [see Glossary: "Distance functions (metrics)"] induces the rough inclusion: $\mu_d(u, v, r)$ if and only if $d(u, v) \leq 1 - r$; conversely, any symmetric rough inclusion μ does induce the metric: $d_\mu(u, v) \leq r \Leftrightarrow \mu(u, v, 1 - r)$. For instance in the case of the t-norm L, the metric $d_L(u, v) = (|\mathrm{DIS}(u, v)|)/(|A|)$, i. e., it is the Hamming distance on information vectors of objects reduced modulo $|A|$.

Let $\phi\colon [0, 1] \to [0, 1]$ be an increasing mapping with $\phi(1) = 1$; for a rough inclusion μ, we define a new predicate μ_ϕ by letting, $\mu_\phi(u, v, r) \Leftrightarrow \mu(u, v, \phi(r))$. μ_ϕ is a rough inclusion. It is transitive with respect to the t-norm $T_\phi(r, s) = \phi^{-1}[t(\phi(r), \phi(s))]$ whenever μ is transitive with respect to the t-norm t. A particular case takes place when one replaces $(|\mathrm{DIS}|)/(|A|)$ with a sum $\sum_{a \in \mathrm{DIS}} w_a$ where $\{w_a\colon a \in A\}$ is a convex family of positive weights. Further, one can replace weights w_a with values of rough inclusions μ_a defined on sets of values V_a. The corresponding formulas will be

$$\mu_L(u, v, r) \Leftrightarrow \sum_{a \in \mathrm{DIS}(u,v)} \mu_a^*(a(u), a(v)) \geq r,$$

$$\mu_P(u, v, r) \Leftrightarrow \exp\left[-\sum_{a \in \mathrm{DIS}(u,v)} \mu_a^*(a(u), a(v))\right] \geq r,$$

where

$$\mu_a^*(u, v) = \frac{\mathrm{argmax}\, \mu_a(a(u), a(v), r)}{\sum_{b \in \mathrm{DIS}} \mathrm{argmax}\, \mu_b(b(u), b(v), r)}\,.$$

This discussion shows that rough inclusions encompass basically all forms of distance functions applied in Data Mining.

Other methods of introducing similarity relations into the realm of information systems, besides rough inclusions, include methods based on templates and on quasi-distance functions in Nguyen S. H. [53]; a template is any conjunct of the form $T\colon \bigwedge_{a \in B}(a \in W_a)$ where B is a set of attributes, and $W_a \subseteq V_a$ is a subset of the value set V_a of the attribute a. Semantics of templates is defined as with descriptor formulas (see Glossary: "Information systems"). Templates are judged by some parameters: length, i. e., the number of generalized descriptors $(a \in W_a)$; support, i.e, number of matching objects; approximated length, applength, i. e., $\sum_{a \in B}(1)/(|W_a \cap a(U)|)$. Quality of the template is given by a combination of some of the parameters, e. g., quality(T)=support(T)+applength(T), etc. Templates are used

in classification problems in an analogous way to decision rules.

A quasi-metric (a similarity measure) [53] is a family $\Delta\colon \{\Delta_a\colon a \in A\}$ of functions where $\Delta_a(u,u) = 0$ and $\Delta_a(u,v) = \Delta_a(v,u)$ for $u,v \in U$. By means of these functions tolerance relations are built with the help of standard metric-forming operators like max, $\sum$: $\tau_1(u,v) \Leftrightarrow \max_a\{\Delta_a(u,v) \le \varepsilon\}$, $\tau_2(u,v) \Leftrightarrow \sum_a \Delta_a(u,v) \le \varepsilon\}$ for a given threshold ε are examples of such similarity relations. These similarity relations are applied in finding classifiers in [53].

Discretization

The important problem of treating continuous values of attributes has been resolved in rough sets with the help of the discretization of attributes technique, common to many paradigms like decision trees etc.; for a decision system (U,A,d), a cut is a pair (a,c), where $a \in A, c$ in reals. The cut (a,c) induces the binary attribute $b_{a,c}(u) = 1$ if $a(u) \ge c$ and it is 0, otherwise. Given a finite sequence $p_a = c_0^a < c_1^a < \cdots < c_m^a$ of reals, the set V_a of values of a is split into disjoint intervals: $(\leftarrow, c_0^a), [c_0^a, c_1^a), \ldots, [c_m^a, \rightarrow)$; the new attribute $D_a(u) = i$ when $b_{c_{i+1}^a} = 0, b_{c_i^a} = 1$, is a discrete counterpart to the continuous attribute a. Given a collection $P = \{p_a\colon a \in A\}$ (a cut system), the set $D = \{D_a\colon a \in A\}$ of attributes transforms the system (U,A,d) into the discrete system (U, D_P, d) called the P-segmentation of the original system. The set P is consistent in case the generalized decision in both systems is identical, i. e., $\delta_A = \delta_{D_P}$; a consistent P is irreducible if P' is not consistent for any proper subset $P' \subset P$; P is optimal if its cardinality is minimal among all consistent cut systems, see Nguyen Hung Son [51], Nguyen Hung Son and Skowron [52].

MD-heuristics based on maximal discernibility of objects and decision trees for optimal sets of cuts were proposed in [51,52]; see also [7].

Granulation of Knowledge

The issue of granulation of knowledge as a problem on its own, has been posed by L.A. Zadeh [124]. Granulation can be regarded as a form of clustering, i. e., grouping objects into aggregates characterized by closeness of certain parameter values among objects in the aggregate and greater differences in those values from aggregate to aggregate. The issue of granulation has been a subject of intensive studies within the rough set community in, e. g., T.Y. Lin [47], Polkowski and Skowron [75,76], Skowron and Stepaniuk [92], Y. Y. Yao [123].

The rough set context offers a natural venue for granulation, and indiscernibility classes were recognized as *elementary granules* whereas their unions serve as *granules of knowledge*; these granules and their direct generalizations to various similarity classes induced by binary relations were subject to research by T.Y. Lin [46], Y.Y. Yao [121]. Granulation of knowledge by means of rough inclusions was studied as well in Polkowski and Skowron [74,75], Skowron and Stepaniuk [92], Polkowski [67,68,69,71].

For an information system (U,A), and a rough inclusion μ on U, granulation with respect to similarity induced by μ is formally performed (Polkowski, op.cit.; op.cit.) by exploiting the class operator *Cls* of mereology [44]. The class operator is applied to any non-vacuous property F of objects (i. e. a distributive entity) in the universe U and produces the object *ClsF* (i. e., the collective entity) representing wholeness of F. The formal definition of *Cls* is: assuming a part relation in U and the associated ingredient relation in *ClsF* does satisfy conditions (1) if $u \in F$ then u is an ingredient of *ClsF*; (2) if v is an ingredient of *ClsF* then some ingredient w of v is an ingredient as well of a T that is in F; in plain words, each ingredient of *ClsF* has an ingredient in common with an object in F. An example of part relation is the proper subset $\subset$ relation on a family of sets; then the subset relation $\subseteq$ is the ingredient relation, and the class of a family F of sets is its union $\bigcup F$. The merit of class operator is in the fact that it always projects hierarchies onto the collective entity plane containing objects.

For an object u and a real number $r \in [0,1]$, we define the granule $g_\mu(u,r)$ about u of the radius r, relative to μ, as the class $ClsF(u,r)$, where the property $F(u,r)$ is satisfied with an object v if and only if $\mu(v,u,r)$ holds.

It was shown [68] that in the case of a transitive μ, v is an ingredient of the granule $g_\mu(u,r)$ if and only if $\mu(v,u,r)$. This fact allows for writing down the granule $g_\mu(u,r)$ as a distributive entity (a set, a list) of objects v satisfying $\mu(v,u,r)$.

Granules of the form $g_\mu(u,r)$ have regular properties of a neighborhood system [69]. Granules generated from a rough inclusion μ can be used in defining a compressed form of the decision system: a granular decision system [69]; for a granulation radius r, and a rough inclusion μ, we form the collection $U_{r,\mu}^G = \{g_\mu(u,r)\}$. We apply a strategy $\mathcal{G}$ to choose a covering $\mathrm{Cov}_{r,\mu}^G$ of the universe U by granules from $U_{r,\mu}^G$. We apply a strategy S in order to assign the value $a^*(g)$ of each attribute $a \in A$ to each granule $g \in \mathrm{Cov}_{r,\mu}^G\colon a^*(g) = S(\{a(u)\colon u \in g\})$. The granular counterpart to the decision system (U,A,d) is a tuple $(U_{r,\mu}^G, \mathcal{G}, S, \{a*\colon a \in A\}, d^*)$. The heuristic principle that H: *objects, similar with respect to conditional at-*

tributes in the set A, should also reveal similar (i. e., close) decision values, and therefore, granular counterparts to decision systems should lead to classifiers satisfactorily close in quality to those induced from original decision systems that is at the heart of all classification paradigms, can be also formulated in this context [69]. Experimental results bear out the hypothesis [73].

Rough Set Theory: Applications

Applications of rough set theory are mainly in classification patterns recognition under uncertainty; classification is performed on test sets by means of judiciously chosen sets of decision rules or generalized decision rules (see Subsects. "Decision Rules", "Similarity", and "Granulation of Knowledge" of Sect. "Rough Set Theory. Extensions") induced from training samples.

Classification

Classification methods can be divided according to the adopted methodology, into classifiers based on reducts and decision rules, classifiers based on templates and similarity, classifiers based on descriptor search, classifiers based on granular descriptors, hybrid classifiers.

For a decision system (U, A, d), classifiers are sets of decision rules. Induction of rules was a subject of research in rough set theory since its beginning. In most general terms, building a classifier consists of searching in the pool of descriptors for their conjuncts that describe sufficiently well decision classes. As distinguished in Stefanowski [99], there are three main kinds of classifiers searched for: *minimal*, i. e., consisting of the minimum possible number of rules describing decision classes in the universe, *exhaustive*, i. e., consisting of all possible rules, *satisfactory*, i. e., containing rules tailored to a specific use. Classifiers are evaluated globally with respect to their ability to properly classify objects, usually by *error* which is the ratio of the number of correctly classified objects to the number of test objects, *total accuracy* being the ratio of the number of correctly classified cases to the number of recognized cases, and *total coverage*, i.e, the ratio of the number of recognized test cases to the number of test cases.

Minimum size algorithms include the LEM2 algorithm due to Grzymala-Busse [28] and covering algorithm in the RSES package [82]; exhaustive algorithms include, e. g., the LERS system due to Grzymala-Busse [27], systems based on discernibility matrices and Boolean reasoning by Skowron [89], Bazan [6], Bazan et al. [7], implemented in the RSES package [82].

Minimal consistent sets of rules were introduced in Skowron and Rauszer [90]; they were shown to coincide with rules induced on the basis of local reducts in Wróblewski [120]. Further developments include dynamic rules, approximate rules, and relevant rules as described in [6,7] as well as local rules [6,7], effective in implementations of algorithms based on minimal consistent sets of rules. Rough set-based classification algorithms, especially those implemented in the RSES system [82], were discussed extensively in [7]; [101] contains a discussion of rough set classifiers along with some attempt at analysis of granulation in the process of knowledge discovery.

In [6], a number of techniques were verified in experiments with real data, based on various strategies:

- **`discretization`** of attributes (codes: N-no discretization, S-standard discretization, D-cut selection by dynamic reducts, G-cut selection by generalized dynamic reducts);
- **`dynamic selection of attributes`** (codes: N-no selection, D-selection by dynamic reducts, G-selection based on generalized dynamic reducts);
- **`decision rule choice`** (codes: A-optimal decision rules, G-decision rules on the basis of approximate reducts computed by Johnson's algorithm, simulated annealing and Boltzmann machines etc., N-without computing of decision rules);
- **`approximation of decision rules`** (codes: N-consistent decision rules, P-approximate rules obtained by descriptor dropping);
- **`negotiations among rules`** (codes: S-based on strength, M-based on maximal strength, R-based on global strength, D-based on stability).

Any choice of a strategy in particular areas yields a compound strategy denoted with the alias being concatenation of symbols of strategies chosen in consecutive areas, e. g., NNAND etc.

We record here in Table 4 an excerpt from the comparison (Tables 8, 9, 10 in [6]) of the best of these strategies with results based on other paradigms in classification for two sets of data: Diabetes and Australian credit from UCI Repository [109].

An adaptive method of classifier construction was proposed in [121]; reducts are determined by means of a genetic algorithm [7] and in turn reducts induce subtables of data regarded as classifying agents; the choice of optimal ensembles of agents is done by a genetic algorithm.

Classifiers constructed by means of similarity relations are based on templates matching a given object or closest to it with respect to a certain distance function, or on coverings of the universe of objects by tolerance classes and assigning the decision value on basis of some of them [53];

Data-Mining and Knowledge Discovery: Case-Based Reasoning, Nearest Neighbor and Rough Sets, Table 4
A comparison of errors in classification by rough set and other paradigms

Paradigm	System/method	Diabetes	Austr. credit
Stat. Methods	Logdisc	0.223	0.141
Stat. Methods	SMART	0.232	0.158
Neural Nets	Backpropagation2	0.248	0.154
Neural Networks	RBF	0.243	0.145
Decision Trees	CART	0.255	0.145
Decision Trees	C4.5	0.270	0.155
Decision Trees	ITrule	0.245	0.137
Decision Rules	CN2	0.289	0.204
Rough Sets	NNANR	0.335	0.140
Rough Sets	DNANR	0.280	0.165
Rough Sets	Best result	0.255 (DNAPM)	0.130 (SNAPM)

Data-Mining and Knowledge Discovery: Case-Based Reasoning, Nearest Neighbor and Rough Sets, Table 5
Accuracy of classification by template and similarity methods

Paradigm	System/method	Diabetes	Austr. credit
Rough Sets	Simple.templ./Hamming	0.6156	0.8217
Rough Sets	Gen.templ./Hamming	0.742	0.855
Rough Sets	Simple.templ./Euclidean	0.6312	0.8753
Rough Sets	Gen.templ./Euclidean	0.7006	0.8753
Rough Sets	Match.tolerance	0.757	0.8747
Rough Sets	Clos.tolerance	0.743	0.8246

we include in Table 5 excerpts from classification results in [53].

A combination of rough set methods with the k-nearest neighbor idea is a further refinement of the classification based on similarity or analogy in Wojna [119]. In this approach, training set objects are endowed with a metric, and the test objects are classified by voting by k nearest training objects for some k that is subject to optimization.

A far-reaching step in classification based on similarity or analogy is granular classification in Polkowski and Artiemjew [73]; in this approach, a granular counterpart (see Subsect. "Granulation of Knowledge" of Sect. "Rough Set Theory. Extensions") $(U^{G}_{r,\mu}, \mathcal{G}, S, \{a*: a \in A\}, d^*)$ to a decision system (U, A, d) is constructed; strategy S can be majority voting, and G either a random choice of a covering or a sequential process of selection. To the granular system standard classifiers like LEM2 or RSES classifiers have been applied. This approach does involve a compression of both universes of objects and rule sets. Table 6 shows a comparison of results with other rough set methods on Australian credit data.

Descriptor search is applied in the LEM2 algorithm [28] aimed at issuing a set of rules with a minimal number of descriptors. This system is applied in medical diagnosis, particularly in the problem of missing values, i.e, data in which some attribute values are lacking.

A number of software systems for inducing classifiers were proposed based on rough set methodology, among them LERS by Grzymala-Busse [27] ; TRANCE due to Kowalczyk [42]; RoughFamily by Słowiński and Stefanowski [95,96]; TAS by Suraj [102]; PRIMEROSE due to Tsumoto [108]; KDD-R by Ziarko [126]; RSES by Skowron et al. [82]; ROSETTA due to Komorowski, Skowron et al. [39]; RSDM by Fernandez-Baizan et al. [19]; GROBIAN due to Duentsch and Gediga [17]; RoughFuzzyLab by Swiniarski [104] and MODLEM by Stefanowski [100].

Rough set techniques were applied in many areas of data exploration, among them in exemplary works:

Processing of audio signals: Czyżewski et al. [14], Kostek [40].
Handwritten digit recognition: Nguyen Tuan Trung [54].
Pattern recognition: Skowron and Swiniarski [93].
Signal classification: Wojdyłło [118].
Image processing: Swiniarski and Skowron [105].
Synthesis and analysis of concurrent processes; Petri nets: Suraj [103].
Conflict analysis, game theory: Deja [15], Polkowski and Araszkiewicz [72].
Rough neural computation modeling: Polkowski [70].
Self organizing maps: Pal, Dasgupta and Mitra [57].
Software engineering: Peters and Ramanna [65].
Multicriteria decision theory; operational research Greco et al. [26].
Programming in logic: Vitoria [112].
Missing value problem: Grzymala-Busse [28].
Learning cognitive concepts: M. Semeniuk-Polkowska [85].

Nearest Neighbor Method

The nearest neighbor method goes back to Fix and Hodges [21,22], Skellam [88] and Clark and Evans [12], who used the idea in statistical tests of significance in nearest neighbor statistics in order to discern between random patterns and clusters. The idea is to assign a class value to a tested object on the basis of the class value of the near-

Data-Mining and Knowledge Discovery: Case-Based Reasoning, Nearest Neighbor and Rough Sets, Table 6
Best results for Australian credit by some rough set-based algorithms; in case ∗, reduction in object size is 49.9 percent, reduction in rule number is 54.6 percent; in case ∗∗, resp., 19.7, 18.2; in case ∗ ∗ ∗, resp., 3.6, 1.9

Source	Method	Accuracy	Coverage
(Bazan,1998)	SNAPM(0.9)	error=0.130	–
(Nguyen S.H.,2000)	simple.templates	0.929	0.623
(Nguyen S.H., 2000)	general.templates	0.886	0.905
(Nguyen S.H.,2000)	closest.simple.templates	0.821	1.0
(Nguyen S.H., 2000)	closest.gen.templates	0.855	1.0
(Nguyen S.H., 2000)	tolerance.simple.templ.	0.842	1.0
(Nguyen S.H., 2000)	tolerance.gen.templ.	0.875	1.0
(Wroblewski, 2004)	adaptive.classifier	0.863	–
(Polkowski Artiemjew, 2007)	granular*.r=0.642	0.8990	1.0
(Polkowski Artiemjew, 2007)	granular**.r=0.714	0.964	1.0
(Polkowski Artiemjew, 2007)	granular***.concept.r=0.785	0.9970	0.9995

est with respect to a chosen metric already classified object. The method was generalized to k-nearest neighbors in Patrick and Fischer [59].

The idea of a closest classified object permeates as already mentioned, many rough set algorithms and heuristics; it is also employed by many other paradigms as a natural choice of strategy. The underlying assumption is that data are organized according to some form of continuity and finding a proper metric or similarity measure secures the correctness of the principle that close in that metric or satisfactorily similar objects should have close class values.

To reach for a formal justification of these heuristics, we recall the rudiments of statistical decision theory in its core fragment, the Bayesian decision theory, see, e. g., Duda, Hart and Stork [16]. Assuming that objects in question are assigned to k decision classes $c_1, \ldots, c_k$, with probabilities $p(c_1), \ldots, p(c_k)$ (priors) and the probability that an object with feature vector x will be observed in the class c_i is $p(x|c_i)$ (the likelihood), the probability of the class c_i when feature vector x has been observed (posterior) is given by the Bayes formula of Bayes [5]: $p(c_i|x) = (p(x|c_i) \cdot p(c_i))/p(x)$ where the evidence $p(x) = \sum_{j=1}^{k} p(x|c_j) \cdot p(c_j)$. Bayesian decision theory assigns to an object with feature vector x the class c_i with $p(c_i|x) > p(c_j|x)$ for all $j \neq i$. The Bayesian error in this case is $p^B = 1 - p(c_i|x)$. This decision rule may be conveniently expressed by means of discriminants assigned to classes: to the class c_i the discriminant $g_i(x)$ is assigned which by the Bayes formula can be expressed as an increasing function of the product $p(x|c_i) \cdot p(c_i)$, e. g., $g_i(x) = \log p(x|c_i) + \log p(c_i)$; an object with features x is classified to c_i for which $g_i(x) = \max\{g_j(x) \colon j = 1, 2, \ldots, k\}$. The most important case happens when conditional densities $p(x|c_i)$ are distributed normally under $N(\mu_i, \sum_i)$; under simplifying assumptions that prior probabilities are equal and features are independent, the Bayes decision rule becomes the following recipe: for an object with feature vector x, the class c_i is assigned for which the distance $||x - \mu_i||$ from x to the mean μ_i is the smallest among all distances $||x - \mu_j||$ [16]. In this way the nearest neighbor method can be formally introduced and justified.

In many real cases, estimation of densities $p(x|c_i)$ is for obvious reasons difficult; nonparametric techniques were proposed to bypass this difficulty. The idea of a Parzen window in Parzen [58], see [16], consists of considering a sample S of objects in a d-space potentially increasing to infinity their number, and a sequence $R_1, R_2, \ldots, R_n, \ldots$ of d-hypercubes of edge lengths h_n hence of volumes $V_n = h_n^d$, R_n containing k_n objects from the sample; the estimate $p_n(x)$ for density induced from R_n is clearly equal to $p_n(x) = (k_n \cdot |S|^{-1})/(V_n)$. Under additional conditions that $\lim_{n\to\infty} V_n = 0$ and $\lim_{n\to\infty} n \cdot V_n = \infty$, the sequence $p_n(x)$ does converge to density $p(x)$ in case the latter is continuous [16].

The difficulty with this approach lies in the question of how to choose regions R_i and their parameters and the idea of nearest neighbors returns: the modification rests on the idea that the sampling window should be dependent on the training sample itself and adapt to its structure; the way of implementing this idea can be as follows: one could center the region R about the object with features x and resize it until it absorbs k training objects (k nearest neighbors); if k_i of them falls into the class c_i then the estimate for probability density is $p(c_i|x) = (k_i)/(k)$ [16]. Letting k variable with $lim_{n\to\infty} \frac{k}{n} = 0$ and $lim_{n\to\infty} k = \infty$ secures that the estimate sequence $p_n(c|x)$ would converge in probability to $p(c_i|x)$ at continuity points of the lat-

ter [16]. The k-nearest neighbor method finds its justification through considerations like the one recalled above.

The 1-nearest neighbor method is the simplest variant of k-nearest neighbors; in the sample space, with the help of a selected metric, it does build neighborhoods of training objects, splitting the space into cells composing together the Voronoi tesselation (Voronoi diagram) of the space, see [79]. The basic theorem by Cover and Hart [13] asserts that the error P in classification by the 1-nearest neighbor method is related to the error P^B in classification by the Bayesian decision rule by the inequality: $P^B \leq P \leq P^B \cdot (2 - (k-1)/(k) \cdot P^B)$, where k is the number of decision classes. Although the result is theoretically valid under the assumption of infinite sample, yet it can be regarded as an estimate of limits of error by the 1-nearest neighbor method as at most twice the error by Bayes classifier also in case of finite reasonably large samples.

A discussion of analogous results for the k-nearest neighbor method can be found in Ripley [81].

Metrics [see Glossary: "Distance functions (metrics)"] used in nearest neighbor-based techniques can be of varied form; the basic distance function is the Euclidean metric in a d-space: $\rho_E(x, y) = [\sum_{i=1}^{d}(x_i - y_i)^2]^{1/2}$ and its generalization to the class of Minkowski metrics $L_p(x, y) = [\sum_{i=1}^{d} |x_i - y_i|^p]^{1/p}$ for $p \geq 1$ with limiting cases of $L_1 = \sum_{i=1}^{d} |x_i - y_i|$ (the Manhattan metric) and $L_\infty(x, y) = \max\{|x_i - y_i| : i = 1, 2, \ldots, d\}$. These metrics can be modified by scaling factors (weights) applied to coordinates, e. g., $L_1^w(x, y) = \sum_{i=1}^{d} w_i \cdot |x_i - y_i|$ is the Manhattan metric modified by the non-negative weight vector w [119] and subject to adaptive training.

Metrics like the above can be detrimental to the nearest neighbor method in the sense that the nearest neighbors are not invariant with respect to transformations like translations, shifts, rotations. A remedy for this difficulty was proposed as the notion of the tangent distance by Simard, Le Cun and Denker [86]. The idea consists of replacing each training as well as each test object, represented as a vector x in the feature space R^k, with its invariance manifold, see Hastie, Tibshirani and Friedman [32], consisting of x along with all its images by allowable transformations: translation, scaling of axes, rotation, shear, line thickening; instead of measuring distances among object representing vectors x, y, one can find shortest distances among invariance manifolds induced by x and y; this task can be further simplified by finding for each x the tangent hyperplane at x to its invariance manifold and measuring the shortest distance between these tangents. For a vector x and the matrix T of basic tangent vectors at x to the invariance manifold, the equation of the tangent hyperplane is $H(x)\colon y = x + Ta$, where a in R^k. The simpler version of the tangent metric method, so-called "one-sided", assumes that tangent hyperplanes are produced for training objects x whereas for test objects x' no invariance manifold is defined but the nearest to x' tangent hyperplane is found and x' is classified as x that defined the nearest tangent hyperplane. In this case the distance to the tangent hyperplane is given by $\arg\min_a \rho_E(x', x + Ta)$ in case the Euclidean metric is chosen as the underlying distance function. In case of a "two-sided" variant, the tangent hyperplane at x' is found as well and x' is classified as the training object x for which the distance between tangent hyperplanes $H(x), H(x')$ is the smallest among distances from $H(x')$ to tangent hyperplanes at training objects.

A simplified variant of the tangent distance method was proposed by Abu-Mostafa, see [32], consisting of producing all images of training and test objects representations in the feature space and regarding them as respectively training and test objects and using standard nearest neighbor techniques.

Among problems related to metrics is also the dimensionality problem; in high-dimensional feature spaces, nearest neighbors can be located at large distances from the test point in question, thus violating the basic principle justifying the method as window choice; as pointed out in [32], the median of the radius R of the sphere about the origin containing the nearest neighbor of n objects uniformly distributed in the cube $[-1/2, 1/2]^p$ is equal to $(1)/(V_p^p) \cdot (1 - 1/2^{1/N})^{1/p}$ where $V_p \cdot r^p$ is the volume of the sphere of radius r in p-space, and it approaches the value of 0.5 with an increase of p for each n, i.e, a nearest neighbor is asymptotically located on the surface of the cube. A related result in Aggarwal et al. [3] asserts that the expected value of the quotient $k^{-1/2} \cdot (\max\{|x|_p, |y|_p\} - \min\{|x|_p, |y|_p\})/(\min\{|x|_p, |y|_p\})$ for vectors x, y uniformly distributed in the cube $(0, 1)^k$ is asymptotically ($k \to \infty$) equal to $C \cdot (2 \cdot p + 1)^{-1/2}$ (this phenomena are collectively known as the dimensionality curse).

To cope with this problem, a method called discriminant adaptive nearest neighbor was proposed by Hastie and Tibshirani [31]; the method consists of adaptive modification of the metric in neighborhoods of test vectors in order to assure that probabilities of decision classes do not vary much; the direction in which the neighborhood is stretched is the direction of the least change in class probabilities.

This idea that the metric used in nearest neighbor finding should depend locally on the training set was also used in constructions of several metrics in the realm of deci-

sion systems, i. e., objects described in the attribute-value language; for nominal values, the metric *VDM* (Value Difference Metric) in Stanfill and Waltz [97] takes into account conditional probabilities $P(d = v|a_i = v_i)$ of decision value given the attribute value, estimated over the training set *Trn*, and on this basis constructs in the value set V_i of the attribute a_i a metric $\rho_i(v_i, v_i') = \sum_{v \in V_d} |P(d = v|a_i = v_i) - P(d = v|a_i = v_i')|$. The global metric is obtained by combining metrics ρ_i for all attributes $a_i \in A$ according to one of many-dimensional metrics, e. g., Minkowski metrics.

This idea was also applied to numerical attributes in Wilson and Martinez [116] in metrics *IVDM* (Interpolated VDM) and *WVDM* (Windowed VDM), see also [119]. A modification of the *WVDM* metric based again on the idea of using probability densities in determining the window size was proposed as the *DBVDM* metric [119].

Implementation of the k-nearest neighbor method requires satisfactorily fast methods for finding k nearest neighbors; the need for accelerating search in training space led to some indexing methods. The method due to Fukunaga and Narendra [24] splits the training sample into a hierarchy of clusters with the idea that in a search for the nearest neighbor the clusters are examined in a top-down manner and clusters evaluated as not fit in the search are pruned from the hierarchy. The alternative is the bottom-up scheme of clustering due to Ward [113]. Distances between test objects and clusters are evaluated as distances between test objects and clusters centers that depend on the chosen strategy.

For features that allow one to map objects onto vectors in a vector space, and clusters in the form of hypercubes a number of tree-based indexing methods exist: k-d trees due to Bentley [8] and quad trees in Finkel and Bentley [20] among them; for high-dimensional feature spaces where the dimensionality curse plays a role, more specialized tree-based indexing methods were proposed: X-trees by Berchtold, Keim and Kriegel [9], SR-trees in Katayama and Satoh [34], TV-trees by Lin, Jagadish and Faloustos [45]. For general, also non-numerical features, more general cluster regions are necessary and a number of indexing methods like BST due to Kalantari and McDonald [33], GHT by Uhlmann [110], GNAT in Brin [10], SS-tree due to White and Jain [115] and M-trees in Ciaccia, Patella and Zezula [11] were proposed. A selection of cluster centers for numerical features may be performed as mean values of feature values of vectors in the cluster; in a general case it may be based on minimization of variance of distance within the cluster [119]. Search for k nearest neighbors of a query object is done in the tree in depth-first order from the root down.

Applications of the k nearest neighbor method include satellite images recognition, protein sequence matching, spatial databases, information retrieval, stock market and weather forecasts, see Aha [4] and Veloso [111].

Case-Based Reasoning

Nearest neighbor techniques can be seen abstractly as tasks of classification of an entity *e* by means of recovering from the base of already classified entities the nearest one *E* to *e* and adopting the class of *E* as the class of *e*. Regarded in this light, classification and more generally, problem solving, tasks may be carried out as tasks of recovering from the base of cases already solved the case most similar to the case being currently solved and adopting its solution, possibly with some modifications, as the solution of the current problem. The distinction between this general idea and nearest neighbor techniques is that in the former, new cases are expected to be used in classification of the next cases.

This is the underlying idea of Case-Based Reasoning. The Case-Based Reasoning methodology emerged in response to some difficulties that were encountered by Knowledge-Based Reasoning systems that dominated in the first period of growth of Artificial Intelligence like DENDRAL, MYCIN, or PROSPECTOR. These systems relied on knowledge of the subject in the form of decision rules or models of entities; the problems were of varied nature: complexity of knowledge extraction; difficulty of managing large amounts of information, difficulties with maintenance and refreshing knowledge, see [114]. In spite of search for improvements and a plethora of ideas and concepts aimed at solving the mentioned difficulties, new ideas based on analogy reasoning came forth. Resorting to analogous cases eliminates the need for modeling knowledge as knowledge is cases along with their solutions and methods; implementing consists in identifying features determining cases; large volumes of information can be managed; learning is acquiring new cases [114].

Case-Based Reasoning (CBR) can be traced back to few sources of origin; on a philosophical level, L. Wittgenstein [117] is quoted in Aamodt and Plaza [2] due to his idea of natural concepts and the claim that they be represented as collections of cases rather than features. An interest in cognitive aspects of learning due partially to the contemporary growth of interest in theories of language inspired in large part by work of N. Chomsky and in psychological aspects of learning and exploiting the concept of situation, was reflected in the machine-learning area by works of Schank and Abelson and Schank [84], [83], regarded as pioneering in the CBR area in [114]. Schank and

Abelson's idea was to use scripts as the tool to describe memory of situation patterns. The role of memory in reasoning by cases was analyzed by Schank [83] leading to memory organization packets (MOP) and by Porter as the case memory model. These studies led to the first CBR-based systems: CYRUS by Kolodner [37], MEDIATOR by Simpson [87], PROTOS by Porter and Bareiss [78], among many others; see [114] for a more complete listing.

The methodology of CBR systems was described in [2] as consisting of four distinct processes repeated in cycles: (1) RETRIEVE: the process consisting of matching the currently being solved case against the case base and fetching cases most similar according to adopted similarity measure to the current case; (2) REUSE: the process of making solutions of most similar cases fit to solve the current case; (3) REVISE: the process of revision of fetched solutions to adapt it to the current case taking place after the REUSED solution was proposed, evaluated and turned to be REVISED; (4) RETAIN: the process of RETAINING the REVISED solution in the case base along with the case as the new solved case that in turn can be used in solving future cases.

Whereas the process of reasoning by cases can be adapted to model reasoning by nearest neighbors yet the difference between the two is clear from Aamodt and Plaza's description of the CBR process: in CBR, case base is incrementally enlarged and retained cases become its valid members used in the process of solving new cases; in nearest neighbors the sharp distinction between training and test objects is kept throughout the classification process.

From an implementation point of view, the problem of representing cases is the first to comment upon. According to Kolodner [38], case representation should secure functionality and ease of information acquisition from the case. The structure of a case is a triple (problem, solution, outcome): the problem is a query along with the state of the world (situation) at the moment the query is posed; the solution is the proposed change in the state of the world and the outcome is the new state after the query is answered. The solution may come along with the method in which it was obtained which is important when revision is necessary. In representing cases many formalisms were used like frames/objects, rules, semantic nets, predicate calculi/first order logics etc.

Retrieval of cases from the case base requires a form of case indexing; Kolodner [38] recommended hand-chosen indices as more human-oriented than automated methods but many automated strategies for indexing were introduced and verified, among them (listed in [114]):

Checklist-based indexing in Kolodner [38] that indexes cases by features and dimensions like MEDIATOR indexing which takes into account types and functions of objects being disputed and relationships among disputants;

Difference-based indexing that selects features discerning best among cases like in CYRUS by Kolodner [37];

Methods based on inductive learning and rule induction which extract features used in indexing;

Similarity methods which produce abstractions of cases sharing common sets of features and use remaining features to discern among cases;

Explanation-based methods which select features by inspecting each case and deciding on the set of relevant features.

The retrieval mechanism relies on indexing and memory organization to retrieve cases; the process of selection of the most similar case uses some measures of similarity and a strategy to find next matching cases in the case base. Among these strategies the nearest neighbor strategy can be found, based on chosen similarity measures for particular features; an exemplary similarity measure is the one adopted in the system ReMind in Kolodner [38]:

$$\mathrm{sim}(I, R) = \frac{\sum_{\text{features f}} w_f \cdot \mathrm{sim}_f(f(I), f(R))}{\sum_{\text{features f}} w_f}$$

where w_f is a weight assigned to the feature f, sim_f is the chosen similarity measure on values of feature f, I is the new case and R is the retrieved case. Other methods are based on inductive learning, template retrieval etc., see [114].

Memory organization should provide a trade-off between wealth of information stored in the case base and efficiency of retrieval; two basic case memory models emerged: the dynamic memory model by Schank [83], Kolodner [37] and the category-exemplar model due to Porter and Bareiss [78]. On the basis of Schank's MOP (Memory Organization Packets), Episodic Memory Organization Packets in Kolodner [37] and Generalized Episodes (GE) in Koton [41] were proposed; a generalized episode consists of cases, norms and indices where norms are features shared by all cases in the episode and indices are features discerning cases in the episode. The memory is organized as a network of generalized episodes, cases, index names and index values. New cases are dynamically incorporated into new episodes. The category-exemplar model divides cases (exemplars) into categories, according to case features, and indexes cases by case links from

categories to cases in them; other indices are feature links from features to categories or cases and difference links from categories to similar cases that differ to a small degree in features. Categories are related within a semantic network.

Reusing and adaptation follows two main lines: structural adaptation, see [38]; in this methodology, the solution of the retrieved case is directly adapted by the adaptation rules. Derivational adaptation in Simpson [87] in which the methods that produced the retrieved solution undergo adaptation in order to yield the solution for the new case is the other. Various techniques were proposed and applied in adaptation process, see, e. g., [114].

Applications based on CBR up to the early 1990s are listed in [38]; to mention a few: JUDGE (Bain): the system for sentencing in murder, assault and man-slaughter; KICS (Yang and Robertson): the system for deciding about building regulations; CASEY (Koton): the system for heart failure diagnosis; CADET (Sycara K): the system for assisting in mechanical design; TOTLEC (Costas and Kashyap): the system for planning in the process of manufacturing design; PLEXUS (Alterman): the system for plan adaptation; CLAVIER (Hennessy and Hinkle): the system implemented at Lockheed to control and modify the autoclave processes in part manufacturing; CaseLine (Magaldi): the system implemented at British Airways for maintenance and repair of the Boeing fleet.

Complexity Issues

Complexity of Computations in Information Systems

The basic computational problems: (DM) Computing the discernibility matrix $M_{\mathrm{U,A}}$ from an information system (U, A); (MLA) The membership in the lower approximation; (RD) Rough definability of sets are of polynomial complexity: (DM) and (RD) in time $\mathrm{O}(n^2)$; (MLA) in time $\mathrm{O}(n)$ [90].

The core, $CORE(U, A)$, of an information system (U, A), is the set of all indispensable attributes, i. e., $CORE(U, A) = \{a \in A\colon \mathrm{ind}(A) \neq \mathrm{ind}(A \setminus \{a\})\}$. As proved in [90], $CORE(U, A) = \{a \in A\colon c_{i,j} = \{a\}$ for some entry $c_{i,j}$ into the discernibility matrix. Thus, finding the core requires $\mathrm{O}(n^2)$ time [90].

The reduct membership Problem, i. e., checking whether a given set B of attributes is a reduct, requires $\mathrm{O}(n^2)$ time [90].

The reduct set Problem, i. e., finding the set of reducts, is polynomially equivalent to the problem of converting a conjunctive form of a monotone Boolean function into the reduced disjunctive form [90]. The number of reducts in an information system of n attributes can be exponential with the upper bound of $\binom{n}{\lfloor n/2 \rfloor}$ reducts [89].

The problem of finding a reduct of minimal cardinality is NP-hard [90]; thus, heuristics for reduct finding are in the general case necessary; they are based on the Johnson algorithm, simulated annealing etc. Genetic algorithms-based hybrid algorithms give short reducts in relatively short time [120].

Complexity of Template-Based Computations

Templates (see Subsect. "Similarity" of Sect. "Rough Set Theory. Extensions") were used in rule induction based on similarity relations [53]. Decision problems related to templates are [53]:

```
TEMPLATE SUPPORT (TS)
Instance: information system(U, A);
          natural numbers s,l
Q.: does there exist a template of
    length l and support s?

OPTIMAL TEMPLATE SUPPORT (OTS)
Input: information system(U, A); natural
       number l
Output: a template T of length l and maximal
        support.
```

The problem TS is NP-complete and the problem OTS is NP-hard [53].

```
TEMPLATE QUALITY PROBLEM (TQP)
Instance: an information system(U, A);
          a natural number k
Q.: does there exist a template of
    quality greater than k?
```

In the case $\mathrm{quality}(T) = \mathrm{support}(T) + \mathrm{length}(T)$, the problem TQP can be solved in polynomial time; in the case $\mathrm{quality}(T) = \mathrm{support}(T) \cdot \mathrm{length}(T)$, the problem TQP is conjectured to be NP-hard.

In [53] some methods for template extraction of satisfactory quality are discussed.

Complexity of Discretization

Discretization (see Subsect. "Discretization" of Sect. "Rough Set Theory. Extensions") offers a number of decision problems. In the case of numeric values of attributes, the decision problem is to check for an irreducible set of cuts.

```
IRREDUCIBLE SET OF CUTS (ISC)
Instance: a decision system(U, A,d)
          with |A| ≥ 2; a natural number k
Q.: does there exist an irreducible set
    of cuts of cardinality less than k?
```

ISC is NP-complete and its optimization version, i. e., if there exists an optimal set of cuts, is NP-hard [53]; in the case $|A| = 1$, ISC is of complexity $O(|U| \cdot log|U|)$ [53].

For non-numerical values of attributes, the counterpart of a cut system is a partition of the value set; for a partition P_a of an attribute value set V_a into k sets, the rank $r(P_a)$ is k; for a family P: $\{P_a : a \in A\}$ of partitions for all value sets of attributes in a decision system (U, A, d), the rank of P is $r(P) = \sum_{a \in A} r(P_a)$; the notion of a consistent partition P mimics that of a consistent cut system. The corresponding problem is

```
Symbolic Value Partition Problem (SVPP)
Input: a decision system (U, A,d)
       a set B of attributes; a natural
       number k
Output: a B-consistent partition of rank
        less or equal k.
```

The problem SVPP is NP-complete [53].

Problems of generating satisfactory cut systems require some heuristics; Maximal Discernibility Heuristics (MD) [51] is also discussed in [7,53].

Complexity of Problems Related to K-Nearest Neighbors

In the case of n training objects in d-space, the search for the nearest neighbor ($k = 1$) does require $O(dn^2)$ time and $O(n)$ space. A parallel implementation [16] is $O(1)$ in time and $O(n)$ in space. This implementation checks for each of Voronoi regions induced from the training sample whether the test object falls inside it and thus would receive its class label by checking for each of the d "faces" of the region whether the object is on the region side of the face.

The large complexity associated with storage of n training objects, called for eliminating some "redundant" training objects; a technique of editing in Hart [30] consists of editing away, i. e., eliminating from the training set of objects that are surrounded in the Voronoi diagram by objects in the same decision class; the complexity of the editing algorithm is $O(d^3 \cdot n^{\lfloor d/2 \rfloor} \cdot \log n)$.

The complexity of problems related to Voronoi diagrams has been researched in many aspects; the Voronoi diagram for n points in 2-space can be constructed in $O(n \cdot \log n)$ [79]; in d-space the complexity is $\Theta(n^{\lfloor d/2 \rfloor})$ (Klee [35]). Such also is the complexity of finding the nearest neighbor in the corresponding space. There were proposed other editing techniques based on graphs: Gabriel graphs in Gabriel and Sokal [25] and relative neighborhood graphs in Toussaint [107].

Future Directions

It seems reasonable to include the following among problems that could occupy the attention of researchers notwithstanding the paradigm applied:

Compression of knowledge encoded in the training set in order to treat large volumes of data.

Granulation of knowledge as a compression means. This should call for efficient formal granulation models based on similarity relations rather than on metrics that would preserve large parts of information encoded in original data.

Search for the most critical decision/classification rules among the bulk induced from data. This search should lead to a deeper insight into relations in data.

Missing values problem, important especially for medical and biological data; its solution should also lead to new deeper relations and dependencies among attributes.

Working out the methods for analyzing complex data like molecular, genetic and medical data that could include signals, images.

Bibliography

Primary Literature

1. Aamodt A (1991) A knowledge intensive approach to problem solving and sustained learning. Dissertation, University Trondheim, Norway. University Microfilms PUB 92–08460
2. Aamodt A, Plaza E (1994) Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI Communications 7:39–59
3. Aggarwal CC, Hinneburg A, Keim DA (2001) On the surprising behavior of distance metrics in high dimensional space. In: Proceedings of the eighth international conference on database theory, London, pp 420–434
4. Aha DW (1998) The omnipresence of case-based reasoning in science and applications. Knowl-Based Syst 11:261–273
5. Bayes T (1763) An essay towards solving a problem in the doctrine of chances. Philos Trans R Soc (London) 53:370–418

6. Bazan JG (1998) A comparison of dynamic and non-dynamic rough set methods for extracting laws from decision tables. In: Polkowski L, Skowron A (eds) Rough sets in knowledge discovery, vol 1. Physica, Heidelberg, pp 321–365
7. Bazan JG et al (2000) Rough set algorithms in classification problems. In: Polkowski L, Tsumoto S, Lin TY (eds) Rough set methods and applications. New developments in knowledge discovery in information systems. Physica, Heidelberg, pp 49–88
8. Bentley JL (1975) Multidimensional binary search trees used for associative searching. Commun ACM 18:509–517
9. Berchtold S, Keim D, Kriegel HP (1996) The X-tree: an index structure for high dimensional data. In: Proceedings of the 22nd International Conference on Very Large Databases VLDB'96 1996 Mumbai, Morgan Kaufmann, San Francisco, pp 29–36
10. Brin S (1995) Near neighbor search in large metric spaces. In: Proceedings of the 21st International Conference on Very Large Databases VLDB'95 Zurich, Morgan Kaufmann, San Francisco, pp 574–584
11. Ciaccia P, Patella M, Zezula P (1997) M-tree: an efficient access method for similarity search in metric spaces. In: Proceedings of the 23rd International Conference on Very Large Databases VLDB'97, Athens, Morgan Kaufmann, San Francisco, pp 426–435
12. Clark P, Evans F (1954) Distance to nearest neighbor as a measure of spatial relationships in populations. Ecology 35:445–453
13. Cover TM, Hart PE (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory IT-13(1):21–27
14. Czyżewski A et al (2004) Musical phrase representation and recognition by means of neural networks and rough sets. In: Transactions on rough sets, vol 1. Lecture Notes in Computer Science, vol 3100. Springer, Berlin, pp 254–278
15. Deja R (2000) Conflict analysis. In: Polkowski L, Tsumoto S, Lin TY (eds) Rough set methods and applications. New developments in knowledge discovery in information systems. Physica, Heidelberg, pp 491–520
16. Duda RO, Hart PE, Stork DG (2001) Pattern classification. Wiley, New York
17. Düntsch I, Gediga G (1998) GROBIAN. In: Polkowski L, Skowron A (eds) Rough sets in knowledge discovery 2. Physica, Heidelberg, pp 555–557
18. Faucett WM (1955) Compact semigroups irreducibly connected between two idempotents. Proc Am Math Soc 6:741–747
19. Fernandez-Baizan MC et al (1998) RSDM: Rough sets data miner. A system to add data mining capabilities to RDBMS. In: Polkowski L, Skowron A (eds) Rough sets in knowledge discovery 2. Physica, Heidelberg, pp 558–561
20. Finkel R, Bentley J (1974) Quad trees: a data structure for retrieval and composite keys. Acta Inf 4:1–9
21. Fix E, Hodges JL Jr (1951) Discriminatory analysis: Nonparametric discrimination: Consistency properties. USAF Sch Aviat Med 4:261–279
22. Fix E, Hodges JL Jr (1952) Discriminatory analysis: Nonparametric discrimination: Small sample performance. USAF Sch Aviat Med 11:280–322
23. Frege G (1903) Grundlagen der Arithmetik II. Jena, Hermann Pohle
24. Fukunaga K, Narendra PM (1975) A branch and bound algorithm for computing k-nearest neighbors. IEEE Trans Comput 24:750–753
25. Gabriel KR, Sokal RR (1969) A new statistical approach to geographic variation analysis. Syst Zool 18:259–278
26. Greco S, Matarazzo B, Słowiński R (1999) On joint use of indiscernibility, similarity and dominance in rough approximation of decision classes. In: Proceedings of the 5th international conference of the decision sciences institute, Athens, Greece, pp 1380–1382
27. Grzymala-Busse JW (1992) LERS – a system for learning from examples based on rough sets. In: Słowiński R (ed) Intelligent decision support. Handbook of Advances and Applications of the Rough Sets Theory. Kluwer, Dordrecht, pp 3–18
28. Grzymala-Busse JW (2004) Data with missing attribute values: Generalization of indiscernibility relation and rule induction. In: Transactions on rough sets, vol 1. Lecture Notes in Computer Science, vol 3100. Springer, Berlin, pp 78–95
29. Grzymala-Busse JW, Ming H (2000) A comparison of several approaches to missing attribute values in data mining. In: Lecture notes in AI, vol 2005. Springer, Berlin, pp 378–385
30. Hart PE (1968) The condensed nearest neighbor rule. IEEE Trans Inf Theory IT-14(3):515–516
31. Hastie T, Tibshirani R (1996) Discriminant adaptive nearest-neighbor classification. IEEE Pattern Recognit Mach Intell 18:607–616
32. Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Springer, New York
33. Kalantari I, McDonald G (1983) A data structure and an algorithm for the nearest point problem. IEEE Trans Softw Eng 9:631–634
34. Katayama N, Satoh S (1997) The SR-tree: an index structure for high dimensional nearest neighbor queries. In: Proceedings of the 1997 ACM SIGMOD international conference on management of data, Tucson, AZ, pp 369–380
35. Klee V (1980) On the complexity of d-dimensional Voronoi diagrams. Arch Math 34:75–80
36. Klösgen W, Żytkow J (eds) (2002) Handbook of data mining and knowledge discovery. Oxford University Press, Oxford
37. Kolodner JL (1983) Maintaining organization in a dynamic long-term memory. Cogn Sci 7:243–80
38. Kolodner JL (1993) Case-based reasoning. Morgan Kaufmann, San Mateo
39. Komorowski J, Skowron A et al (1998) The ROSETTA software system. In: Polkowski L, Skowron A (eds) Rough sets in knowledge discovery 2. Physica, Heidelberg, pp 572–575
40. Kostek B (2007) The domain of acoustics seen from the rough set perspective. In: Transactions on rough sets, vol VI. Lecture Notes in Computer Science, vol 4374. Springer, Berlin, pp 133–151
41. Koton P (1989) Using experience in learning and problem solving. Ph D Dissertation MIT/LCS/TR-441, MIT, Laboratory of Computer Science, Cambridge
42. Kowalczyk W (1998) TRANCE: A tool for rough data analysis, classification and clustering. In: Polkowski L, Skowron A (eds) Rough sets in knowledge discovery 2. Physica, Heidelberg, pp 566–568
43. Krawiec K et al (1998) Learning decision rules from similarity based rough approximations. In: Polkowski L, Skowron A (eds) Rough sets in knowledge discovery, vol 2. Physica, Heidelberg, pp 37–54
44. Leśniewski S (1916) Podstawy Ogólnej Teoryi Mnogosci (On

the Foundations of Set Theory), in Polish. The Polish Scientific Circle, Moscow; see also a later digest: (1982) Topoi 2:7–52
45. Lin KI, Jagadish HV, Faloustos C (1994) The TV-tree: an index structure for high dimensional data. VLDB J 3:517–542
46. Lin TY (1997) From rough sets and neighborhood systems to information granulation and computing with words. In: 5th European Congress on Intelligent Techniques and Soft Computing, 1997 Aachen, Verlagshaus Mainz, Aachen, pp 1602–1606
47. Lin TY (2005) Granular computing: Examples, intuitions, and modeling. In: Proceedings of IEEE 2005 conference on granular computing GrC05, Beijing, China. IEEE Press, pp 40–44, IEEE Press, New York
48. Ling C-H (1965) Representation of associative functions. Publ Math Debrecen 12:189–212
49. Michalski RS et al (1986) The multi-purpose incremental learning system AQ15 and its testing to three medical domains. In: Proceedings of AAAI-86. Morgan Kaufmann, San Mateo, pp 1041–1045
50. Mostert PS, Shields AL (1957) On the structure of semigroups on a compact manifold with a boundary. Ann Math 65:117–143
51. Nguyen SH (1997) Discretization of real valued attributes: Boolean reasoning approach. Ph D Dissertation, Warsaw University, Department of Mathematics, Computer Science and Mechanics
52. Nguyen SH, Skowron A (1995) Quantization of real valued attributes: Rough set and Boolean reasoning approach. In: Proceedings 2nd annual joint conference on information sciences, Wrightsville Beach, NC, pp 34–37
53. Nguyen SH (2000) Regularity analysis and its applications in Data Mining. In: Polkowski L, Tsumoto S, Lin TY (eds) Rough set methods and applications. New developments in knowledge discovery in information systems. Physica, Heidelberg, pp 289–378
54. Nguyen TT (2004) Handwritten digit recognition using adaptive classifier construction techniques. In: Pal SK, Polkowski L, Skowron A (eds) Rough – neural computing. Techniques for computing with words. Springer, Berlin, pp 573–586
55. Novotny M, Pawlak Z (1988) Partial dependency of attributes. Bull Pol Acad Ser Sci Math 36:453–458
56. Novotny M, Pawlak Z (1992) On a problem concerning dependence spaces. Fundam Inform 16:275–287
57. Pal SK, Dasgupta B, Mitra P (2004) Rough-SOM with fuzzy discretization. In: Pal SK, Polkowski L, Skowron A (eds) Rough – neural computing. Techniques for computing with words. Springer, Berlin, pp 351–372
58. Parzen E (1962) On estimation of a probability density function and mode. Ann Math Stat 33(3):128–152
59. Patrick EA, Fisher FP (1970) A generalized k-nearest neighbor rule. Inf Control 16(2):128–152
60. Pawlak Z (1982) Rough sets. Int J Comput Inf Sci 11:341–356
61. Pawlak Z (1985) On rough dependency of attributes in information systems. Bull Pol Acad Ser Sci Tech 33:551–559
62. Pawlak Z (1991) Rough sets: theoretical aspects of reasoning about data. Kluwer, Dordrecht
63. Pawlak Z, Skowron, A (1993) A rough set approach for decision rules generation. In: Proceedings of IJCAI'93 workshop W12. The management of uncertainty in AI. also: ICS Research Report 23/93 Warsaw University of Technology
64. Pawlak Z, Skowron A (1994) Rough membership functions. In: Yaeger RR, Fedrizzi M, Kasprzyk J (eds) Advances in the Dempster–Schafer theory of evidence. Wiley, New York, pp 251–271
65. Peters J, Ramanna S (2004) Approximation space for software models. In: Transactions on rough sets, vol I. Lecture Notes in Computer Science, vol 3100. Springer, Berlin, pp 338–355
66. Poincaré H (1902) Science et hypothese and l'Hypothese.Flammarion, Paris
67. Polkowski L (2003) A rough set paradigm for unifying rough set theory and fuzzy set theory. In: Proceedings RSFDGrC03, Chongqing, China, 2003. Lecture Notes in AI, vol 2639. Springer, Berlin, pp 70–78; also: Fundam Inf 54:67–88
68. Polkowski L (2004) Toward rough set foundations. Mereological approach. In: Proceedings RSCTC04, Uppsala, Sweden. Lecture Notes in AI, vol 3066. Springer, Berlin, pp 8–25
69. Polkowski L (2005) Formal granular calculi based on rough inclusions. In: Proceedings of IEEE 2005 conference on granular computing GrC05, Beijing, China. IEEE Press, New York, pp 57–62
70. Polkowski L (2005) Rough-fuzzy-neurocomputing based on rough mereological calculus of granules. Int J Hybrid Intell Syst 2:91–108
71. Polkowski L (2006) A model of granular computing with applications. In: Proceedings of IEEE 2006 conference on granular computing GrC06, Atlanta, USA May 10-12. IEEE Press, New York, pp 9–16
72. Polkowski L, Araszkiewicz B (2002) A rough set approach to estimating the game value and the Shapley value from data. Fundam Inf 53(3/4):335–343
73. Polkowski L, Artiemjew P (2007) On granular rough computing: Factoring classifiers through granular structures. In: Proceedings RSEISP'07, Warsaw. Lecture Notes in AI, vol 4585, pp 280–290
74. Polkowski L, Skowron A (1994) Rough mereology. In: Proceedings of ISMIS'94. Lecture notes in AI, vol 869. Springer, Berlin, pp 85–94
75. Polkowski L, Skowron A (1997) Rough mereology: a new paradigm for approximate reasoning. Int J Approx Reason 15(4):333–365
76. Polkowski L, Skowron A (1999) Towards an adaptive calculus of granules. In: Zadeh LA, Kacprzyk J (eds) Computing with words in information/intelligent systems, vol 1. Physica, Heidelberg, pp 201–228
77. Polkowski L, Skowron A, Żytkow J (1994) Tolerance based rough sets. In: Lin TY, Wildberger M (eds) Soft Computing: Rough sets, fuzzy logic, neural networks, uncertainty management, knowledge discovery. Simulation Councils Inc., San Diego, pp 55–58
78. Porter BW, Bareiss ER (1986) PROTOS: An experiment in knowledge acquisition for heuristic classification tasks. In: Proceedings of the first international meeting on advances in learning (IMAL), Les Arcs, France, pp 159–174
79. Preparata F, Shamos MI (1985) Computational geometry: an introduction. Springer, New York
80. Rauszer C (1985) An equivalence between indiscernibility relations in information systems and a fragment of intuitionistic logic. Bull Pol Acad Ser Sci Math 33:571–579
81. Ripley BD (1997) Pattern recognition and neural networks. Cambridge University Press, Cambridge
82. Skowron A et al (1994) A system for data analysis. http://logic.mimuw.edu.pl/~rses/
83. Schank RC (1982) Dynamic memory: A theory of reminding

and learning in computers and people. Cambridge University Press, Cambridge
84. Schank RC, Abelson RP (1977) Scripts, plans, goals and understanding. Lawrence Erlbaum, Hillsdale
85. Semeniuk-Polkowska M (2007) On conjugate information systems: A proposition on how to learn concepts in humane sciences by means of rough set theory. In: Transactions on rough sets, vol VI. Lecture Notes in Computer Science, vol 4374. Springer, Berlin, pp 298–307
86. Simard P, Le Cun Y, Denker J (1993) Efficient pattern recognition using a new transformation distance. In: Hanson SJ, Cowan JD, Giles CL (eds) Advances in neural information processing systems, vol 5. Morgan Kaufmann, San Mateo, pp 50–58
87. Simpson RL (1985) A A computer model of case-based reasoning in problem solving: An investigation in the domain of dispute mediation. Georgia Institute of Technology, Atlanta
88. Skellam JG (1952) Studies in statistical ecology, I, Spatial pattern. Biometrica 39:346–362
89. Skowron A (1993) Boolean reasoning for decision rules generation. In: Komorowski J, Ras Z (eds) Proceedings of ISMIS'93. Lecture Notes in AI, vol 689. Springer, Berlin, pp 295–305
90. Skowron A, Rauszer C (1992) The discernibility matrices and functions in decision systems. In: Słowiński R (ed) Intelligent decision support. Handbook of applications and advances of the rough sets theory. Kluwer, Dordrecht, pp 311–362
91. Skowron A, Stepaniuk J (1996) Tolerance approximation spaces. Fundam Inf 27:245–253
92. Skowron A, Stepaniuk J (2001) Information granules: towards foundations of granular computing. Int J Intell Syst 16:57–85
93. Skowron A, Swiniarski RW (2004) Information granulation and pattern recognition. In: Pal SK, Polkowski L, Skowron A (eds), Rough – Neural Computing. Techniques for computing with words. Springer, Berlin, pp 599–636
94. Slezak D (2000) Various approaches to reasoning with frequency based decision reducts: a survey. In: Polkowski L, Tsumoto S, Lin TY (eds) Rough set methods and applications. New developments in knowledge discovery in information systems. Physica, Heidelberg, pp 235–288
95. Słowiński R, Stefanowski J (1992) "RoughDAS" and "RoughClass" software implementations of the rough set approach. In: Słowiński R (ed) Intelligent decision support: Handbook of advances and applications of the rough sets theory. Kluwer, Dordrecht, pp 445–456
96. Słowiński R, Stefanowski J (1998) Rough family – software implementation of the rough set theory. In: Polkowski L, Skowron A (eds) Rough sets in knowledge discovery 2. Physica, Heidelberg, pp 580–586
97. Stanfill C, Waltz D (1986) Toward memory-based reasoning. Commun ACM 29:1213–1228
98. Mackie M (2006) Stanford encyclopedia of philosophy: Transworld identity http://plato.stanford.edu/entries/identity-transworld Accessed 6 Sept 2008
99. Stefanowski J (1998) On rough set based approaches to induction of decision rules. In: Polkowski L, Skowron A (eds) Rough sets in knowledge discovery 1. Physica, Heidelberg, pp 500–529
100. Stefanowski J (2007) On combined classifiers, rule induction and rough sets. In: Transactions on rough sets, vol VI. Lecture Notes in Computer Science, vol 4374. Springer, Berlin, pp 329–350
101. Stepaniuk J (2000) Knowledge discovery by application of rough set models. In: Polkowski L, Tsumoto S, Lin TY (eds) Rough set methods and applications. New developments in knowledge discovery in information systems. Physica, Heidelberg, pp 138–233
102. Suraj Z (1998) TAS: Tools for analysis and synthesis of concurrent processes using rough set methods. In: Polkowski L, Skowron A (eds) Rough sets in knowledge discovery, vol 2. Physica, Heidelberg, pp 587–590
103. Suraj Z (2000) Rough set methods for the synthesis and analysis of concurrent processes. In: Polkowski L, Tsumoto S, Lin TY (eds) Rough set methods and applications. New developments in knowledge discovery in information systems. Physica, Heidelberg, pp 379–490
104. Swiniarski RW (1998) RoughFuzzyLab: A system for data mining and rough and fuzzy sets based classification. In: Polkowski L, Skowron A (eds) Rough sets in knowledge discovery 2. Physica, Heidelberg, pp 591–593
105. Swiniarski RW, Skowron A (2004) Independent component analysis, principal component analysis and rough sets in face recognition. In: Transactions on rough sets, vol I. Lecture Notes in Computer Science, vol 3100. Springer, Berlin, pp 392–404
106. Sycara EP (1987) Resolving adversial conflicts: An approach to integrating case-based and analytic methods. Georgia Institute of Technology, Atlanta
107. Toussaint GT (1980) The relative neighborhood graph of a finite planar set. Pattern Recognit 12(4):261–268
108. Tsumoto S (1998) PRIMEROSE. In: Polkowski L, Skowron A (eds) Rough sets in knowledge discovery 2. Physica, Heidelberg, pp 594–597
109. UCI Repository http://www.ics.uci.edu./mlearn/databases/ University of California, Irvine, Accessed 6 Sept 2008
110. Uhlmann J (1991) Satisfying general proximity/similarity queries with metric trees. Inf Process Lett 40:175–179
111. Veloso M (1994) Planning and learning by analogical reasoning. Springer, Berlin
112. Vitoria A (2005) A framework for reasoning with rough sets. In: Transactions on rough sets, vol IV. Lecture Notes in Computer Science, vol 3100. Springer, Berlin, pp 178–276
113. Ward J (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244
114. Watson I, Marir F (1994) Case-based reasoning: A review http://www.ai-cbr.org/classroom/cbr-review.html Accessed 6 Sept 2008; see also: Watson I (1994). Knowl Eng Rev 9(4):327–354
115. White DA, Jain R (1996) Similarity indexing with the SS-tree. In: Proceedings of the twelve international conference on data engineering, New Orleans LA, pp 516–523
116. Wilson DR, Martinez TR (1997) Improved heterogeneous distance functions. J Artif Intell Res 6:1–34
117. Wittgenstein L (1953) Philosophical investigations. Blackwell, London
118. Wojdyłło P (2004) WaRS: A method for signal classification. In: Pal SK, Polkowski L, Skowron A (eds) Rough – neural computing. Techniques for computing with words. Springer, Berlin, pp 649–688
119. Wojna A (2005) Analogy-based reasoning in classifier construction. In: Transactions on rough sets, vol IV. Lecture Notes in Computer Science, vol 3700. Springer, Berlin, pp 277–374
120. Wróblewski J (1998) Covering with reducts – a fast algorithm

for rule generation. In: Lecture notes in artificial intelligence, vol 1424. Springer, Berlin, pp 402–407
121. Wróblewski J (2004) Adaptive aspects of combining approximation spaces. In: Pal SK, Polkowski L, Skowron A (eds) Rough – neural computing. Techniques for computing with words. Springer, Berlin, pp 139–156
122. Yao YY (2000) Granular computing: Basic issues and possible solutions. In: Proceedings of the 5th Joint Conference on Information Sciences I. Assoc Intell Machinery, Atlantic NJ, pp 186–189
123. Yao YY (2005) Perspectives of granular computing. In: Proceedings of IEEE 2005 Conference on Granular Computing GrC05, Beijing, China. IEEE Press, New York, pp 85–90
124. Zadeh LA (1979) Fuzzy sets and information granularity. In: Gupta M, Ragade R, Yaeger RR (eds) Advances in fuzzy set theory and applications. North-Holland, Amsterdam, pp 3–18
125. Zeeman EC (1965) The topology of the brain and the visual perception. In: Fort MK (ed) Topology of 3-manifolds and selected topics. Prentice Hall, Englewood Cliffs, pp 240–256
126. Ziarko W (1998) KDD-R: Rough set-based data mining system. In: Polkowski L, Skowron A (eds) Rough sets in knowledge discovery 2. Physica, Heidelberg, pp 598–601

Books and Reviews

Avis D, Bhattacharya BK (1983) Algorithms for computing d-dimensional Voronoi diagrams and their duals. In: Preparata FP (ed) Advances in computing research: Computational geometry. JAI Press, Greenwich, pp 159–180
Bocheński JM (1954) Die Zeitgenössischen Denkmethoden. A. Francke, Bern
Dasarathy BV (ed) (1991) Nearest neighbor (NN) norms: NN Pattern classification techniques. IEEE Computer Society, Washington
Friedman J (1994) Flexible metric nearest-neighbor classification. Technical Report, Stanford University
Polkowski L (2002) Rough sets. Mathematical foundations. Physica, Heidelberg
Russell SJ, Norvig P (2003) Artificial intelligence. A modern approach, 2nd edn. Prentice Hall Pearson Education, Upper Saddle River
Toussaint GT, Bhattacharya BV, Poulsen RS (1984) Application of voronoi diagrams to nonparametric decision rules. In: Proceedings of Computer Science and Statistics: The Sixteenth Symposium on the Interface. North Holland, Amsterdam, pp 97–108
Watson I (1997) Applying case-based reasoning. Techniques for enterprise systems. Morgan Kaufmann, Elsevier, Amsterdam

Data-Mining and Knowledge Discovery, Introduction to

Peter Kokol
Department of Computer Science, University of Maribor, Maribor, Slovenia

Data mining and knowledge discovery is the principle of analyzing large amounts of data and picking out relevant information leading to the knowledge discovery process for extracting meaningful patterns, rules and models from raw data making discovered patterns understandable. Applications include medicine, politics, games, business, marketing, bioinformatics and many other areas of science and engineering. It is an area of research activity that stands at the intellectual intersection of statistics, computer science, machine learning and database management. It deals with very large datasets, tries to make fewer theoretical assumptions than has traditionally been done in statistics, and typically focuses on problems of classification, prediction, description and profiling, clustering, and regression. In such domains, data mining often uses decision trees or neural networks as models and frequently fits them using some combination of techniques such as bagging, boosting/arcing, and racing. Data mining techniques include data visualization, neural network analysis, support vector machines, genetic and evolutionary algorithms, case based reasoning, etc. Other activities in data mining focus on issues such as causation in large-scale systems, and this effort often involves elaborate statistical models and, quite frequently, Bayesian methodology and related computational techniques. Also data mining covers evaluation of the top down approach of model building, starting with an assumed mathematical model solving dynamical equations, with the bottom up approach of time series analysis, which takes measured data as input and provides as output a mathematical model of the system. Their difference is termed measurement error. Data mining allows characterization of chaotic dynamics, involves Lyapunov exponents, fractal dimension and Kolmogorov–Sinai entropy. Data mining comes in two main directions: directed and undirected. Directed data mining tries to categorize or explain some particular target field, while undirected data mining attempts to find patterns or similarities among groups of records without the specific goal.

Pedrycz (see ► Data and Dimensionality Reduction in Data Analysis and System Modeling) points out that data and dimensionality reduction are fundamental pursuits of data analysis and system modeling. With the rapid growth of the size of data sets and the diversity of data themselves, the use of some reduction mechanisms becomes a necessity. Data reduction is concerned with a reduction of sizes of data sets in terms of the number of data points. This helps reveal an underlying structure in data by presenting a collection of groups present in data. Given the number of groups which are very limited, the clustering mechanisms become effective in terms of data reduction. Dimensionality reduction is aimed at the reduction of the number of attributes (features) of the data which leads to a typically

small subset of features or brings the data from a highly dimensional feature space to a new one of a far lower dimensionality. A joint reduction process involves data and feature reduction.

Brameier (see ► Data-Mining and Knowledge Discovery, Neural Networks in) describes neural networks or, more precisely, artificial neural networks as mathematical and computational models that are inspired by the way biological nervous systems process information. A neural network model consists of a larger number of highly interconnected, simple processing nodes or units which operate in parallel and perform functions collectively, roughly similar as in biological neural networks. Artificial neural networks like their biological counterpart, are adaptive systems which learn by example. Learning works by adapting free model parameters, i. e., the signal strength of the connections and the signal flow, to external information that is presented to the network. In other terms, the information is stored in the weighted connections. If not stated explicitly, let the term "neural network" mean "artificial neural network" in the following. In more technical terms, neural networks are non-linear statistical data modeling tools. Neural networks are generally well suited for solving problem tasks that involve classification of (necessarily numeric) data vectors, pattern recognition and decision-making. The power and usefulness of neural networks have been demonstrated in numerous application areas, like image processing, signal processing, biometric identification – including handwritten character, fingerprint, face, and speech recognition – robotic control, industrial engineering, and biomedicine. In many of these tasks neural networks outperform more traditional statistical or artificial intelligence techniques or may even achieve human-like performance. The most valuable characteristics of neural networks are adaptability and tolerance to noisy or incomplete data. Another important advantage is in solving problems that do not have an algorithmic solution or for which an algorithmic solution is too complex or time-consuming to be found. Brameier describes his chapter as providing a concise introduction to the two most popular neural network types used in application, back propagation neural networks and self-organizing maps The former learn high-dimensional non-linear functions from given input-output associations for solving classification and approximation (regression) problems. The latter are primarily used for data clustering and visualization and for revealing relationships between clusters. These models are discussed in context with alternative learning algorithms and neural network architectures.

Polkowski (see ► Data-Mining and Knowledge Discovery: Case-Based Reasoning, Nearest Neighbor and Rough Sets) describes rough set theory as a formal set of notions aimed at carrying out tasks of reasoning, in particular about classification of objects, in conditions of uncertainty. Conditions of uncertainty are imposed by incompleteness, imprecision and ambiguity of knowledge. Applications of this technique can be found in many areas like satellite images analysis, plant ecology, forestry, conflict analysis, game theory, and cluster identification in medical diagnosis. Originally, the basic notion proposed was that of a knowledge base, understood as a collection of equivalence relations on a universe of objects; each relation induces on the set a partition into equivalence classes. Knowledge so encoded is meant to represent the classification ability. As objects for analysis and classification come most often in the form of data, a useful notion of an information system is commonly used in knowledge representation; knowledge base in that case is defined as the collection of indiscernibility relations. Exact concepts are defined as unions of indiscernibility classes whereas inexact concepts are only approximated from below (lower approximations) and from above (upper approximations) by exact ones. Each inexact concept is, thus, perceived as a pair of exact concepts between which it is sandwiched.

Povalej, Verlic and Stiglig (see ► Discovery Systems) point out that results from simple query for "discovery system" on the World Wide Web returns different types of discovery systems: from knowledge discovery systems in databases, internet-based knowledge discovery, service discovery systems and resource discovery systems to more specific, like for example drug discovery systems, gene discovery systems, discovery system for personality profiling, and developmental discovery systems among others. As illustrated variety of discovery systems can be found in many different research areas, but they focus on knowledge discovery and knowledge discovery systems from the computer science perspective.

A decision tree in data mining or machine learning is a predictive model or a mapping process from observations about an item to conclusions about its target value. More descriptive names for such tree models are classification tree (discrete outcome) or regression tree (continuous outcome). In these tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. Podgorelec and Zorman (see ► Decision Trees) point out that the term "decision trees" has been used for two different purposes: in decision analysis as a decision support tool for modeling decisions and their possible consequences to select the best course of action in situations where one faces uncertainty, and in machine learning or data mining as a predictive model;

that is, a mapping from observations about an item to conclusions about its target value. Their article concentrates on the machine learning view.

Orlov, Sipper and Hauptman (see ► Genetic and Evolutionary Algorithms and Programming: General Introduction and Application to Game Playing) cover genetic and evolutionary algorithms, which are a family of search algorithms inspired by the process of (Darwinian) evolution in Nature. Common to all the different family members is the notion of solving problems by evolving an initially random population of candidate solutions, through the application of operators like crossover and mutation inspired by natural genetics and natural selection, such that in time "fitter" (i. e., better) solutions emerge. The field, whose origins can be traced back to the 1950s and 1960s, has come into its own over the past two decades, proving successful in solving multitudinous problems from highly diverse domains including (to mention but a few): optimization, automatic programming, electronic-circuit design, telecommunications, networks, finance, economics, image analysis, signal processing, music, and art.

Berkhin and Dhillon (see ► Knowledge Discovery: Clustering) discuss the condition where data found in scientific and business applications usually do not fit a particular parametrized probability distribution. In other words, the data are complex. Knowledge discovery starts with exploration of this complexity in order to find inconsistencies, artifacts, errors, etc. in the data. After data are cleaned, it is usually still extremely complex. Descriptive data mining deals with comprehending and reducing this complexity. Clustering is a premier methodology in descriptive unsupervised data mining. A cluster could represent an important subset of the data such as a galaxy in astronomical data or a segment of customers in marketing applications. Clustering is important as a fundamental technology to reduce data complexity and to find data patterns in an unsupervised fashion. It is universally used as a first technology of choice in data exploration.

Džeroski, Panov and Zenko (see ► Machine Learning, Ensemble Methods in) cover ensemble methods, which are machine learning methods that construct a set of predictive models and combine their outputs into a single prediction. The purpose of combining several models together is to achieve better predictive performance, and it has been shown in a number of cases that ensembles can be more accurate than single models. While some work on ensemble methods has already been done in the 1970s, it was not until the 1990s, and the introduction of methods such as bagging and boosting that ensemble methods started to be more widely used. Today, they represent a standard machine learning method which has to be considered whenever good predictive accuracy is demanded.

Liu and Zhao (see ► Manipulating Data and Dimension Reduction Methods: Feature Selection) cover feature selection, which is the study of algorithms for reducing dimensionality of data for various purposes. One of the most common purposes is to improve machine learning performance. The other purposes include simplifying data description, streamlining data collection, improving comprehensibility of the learned models, and helping gain insight through learning. The objective of feature selection is to remove irrelevant and/or redundant features and retain only relevant features. Irrelevant features can be removed without affecting learning performance. Redundant features are a type of irrelevant features. The distinction is that a redundant feature implies the co-presence of another feature; individually, each feature is relevant, but the removal of either one will not affect learning performance. As a plethora of data are generated in every possible means with the exponential decreasing costs of data storage and computer processing power, data dimensionality increases on a scale beyond imagination in cases ranging from transactional data to high-throughput data. In many fields such as medicine, health care, Web search, and bioinformatics, it is imperative to reduce high dimensionality such that efficient data processing and meaningful data analysis can be conducted in order to mine nuggets from high-dimensional, massive data.

Data-Mining and Knowledge Discovery, Neural Networks in

Markus Brameier
Bioinformatics Research Center, University of Aarhus, Århus, Denmark

Article Outline

Glossary
Definition of the Subject
Introduction
Neural Network Learning
Feedforward Neural Networks
Backpropagation
Other Learning Rules
Other Neural Network Architectures
Self-organizing Maps
Future Directions
Bibliography

Glossary

Artificial neural network An artificial neural network is a system composed of many simple, but highly interconnected processing nodes or neurons which operate in parallel and collectively. It resembles biological nervous systems in two basic functions: (1) Experiential knowledge is acquired through a learning process and can be retrieved again later. (2) The knowledge is stored in the strength (weights) of the connections between the neurons.

Artificial neuron An artificial neuron receives a number of inputs, which may be either external inputs to the neural network or outputs of other neurons. Each input connection is assigned a weight, similar to the synaptic efficacy of a biological neuron. The weighted sum of inputs is compared against an activation level (threshold) to determine the activation value of the neuron.

Activation function The activation or transfer function transforms the weighted inputs of a neuron into an output signal. Activation functions often have a "squashing" effect. Common activation functions used in neural networks are: threshold, linear, sigmoid, hyperbolic, and Gaussian.

Learning rule The learning rule describes the way a neural network is trained, i. e., how its free parameters undergo changes to fit the network to the training data.

Feedforward network Feedforward neural networks are organized in one or more layers of processing units (neurons). In a feedforward neural network the signal is allowed to flow one-way only, i. e., from inputs to outputs. There are no feedback loops, i. e., the outputs of a layer do not affect its inputs.

Feedback networks In feedback or recurrent networks signals may flow in both directions. Feedback networks are dynamic such that they have a state that is changing continuously until it reaches an equilibrium point.

Definition of the Subject

Neural networks (NNs) or, more precisely, *artificial neural networks* (ANNs) are mathematical and computational models that are inspired by the way biological nervous systems process information. A neural network model consists of a larger number of highly interconnected, simple processing nodes or units which operate in parallel and perform functions collectively, roughly similar to biological neural networks. ANNs, like their biological counterpart, are adaptive systems that learn by example. Learning works by adapting free model parameters, i. e., the signal strength of the connections and the signal flow, to external information that is presented to the network. In other terms, the information is stored in the weighted connections. If not stated explicitly, let the term "neural network" mean "artificial neural network" in the following.

In more technical terms, neural networks are non-linear statistical data modeling tools. Neural networks are generally well suited for solving problem tasks that involve classification, pattern recognition and decision making. The power, and usefulness of neural networks have been demonstrated in numerous application areas, like image processing, signal processing, biometric identification – including handwritten character, fingerprint, face, and speech recognition – robotic control, industrial engineering, and biomedicine. In many of these tasks neural networks outperform more traditional statistical or artificial intelligence techniques or may even achieve human-like performance. The most valuable characteristics of neural networks are adaptability and tolerance to noisy or incomplete data. Another important advantage is in solving problems that do not have an algorithmic solution or for which an algorithmic solution is too complex or time-consuming to be found.

The first neural network model was developed by McCulloch and Pitts in the 1940s [32]. In 1958 Rosenblatt [41] described the first learning algorithm for a single neuron, the perceptron model. After Rumelhart et al. [46] invented the popular backpropagation learning algorithm for multi-layer networks in 1986, the field of neural networks gained incredible popularity in the 1990s.

Neural networks is regarded as a method of *machine learning*, the largest subfield of *artificial intelligence* (AI). Conventional AI mainly focuses on the development of expert systems and the design of intelligent agents. Today neural networks also belong to the more recent field of *computational intelligence* (CI), which also includes *evolutionary algorithms* (EAs) and *fuzzy logic*.

Introduction

Herein I provide a concise introduction to the two most popular neural network types used in application, backpropagation neural networks (BPNNs) and self-organizing maps (SOMs). The former learn high-dimensional non-linear functions from given input–output associations for solving classification and approximation (regression) problems. The latter are primarily used for data clustering and visualization and for revealing relationships between clusters. These models are discussed in context with alternative learning algorithms and neural network architectures.

Biological Motivation

Artificial neural networks are inspired by biological nervous systems, which are highly distributed and interconnected networks. The human brain is principally composed of a very large number of relatively simple neurons (approx. 100 billion), each of which is connected to several thousand other *neurons*, on average. A neuron is a specialized cell that consists of the cell body (the *soma*), multiple spine-like extensions (the *dendrites*) and a single nerve fiber (the *axon*). The axon connects to the dendrites of another neuron via a synapse. When a neuron is activated, it transmits an electrical impulse (activation potential) along its axon. At the synapse the electric signal is transformed into a chemical signal such that a certain number of neurotransmitters cross the synaptic gap to the post synaptic neuron, where the chemical signal is converted back to an electrical signal to be transported along the dendrites. The dendrites receive signals from the axons of other neurons. One very important feature of neurons is that they react delayed. A neuron combines the strengths (energies) of all received input signals and sends out its own signal ("fires") only if the total signal strength exceeds a certain critical activation level. A synapse can either be excitatory or inhibitory. Input signals from an excitatory synapse increase the activation level of the neuron while inputs from an inhibitory synapse reduce it. The strength of the input signals critically depends on modulations at the synapses. The brain learns basically by adjusting number and strength of the synaptic connections.

Artificial neural networks copy only a small amount of the biological complexity by using a much smaller number of simpler neurons and connections. Nevertheless, artificial neural networks can perform remarkably complex tasks by applying a similar principle, i. e., the combination of simple and local processing units, each calculating a weighted sum of its inputs and sending out a signal if the sum exceeds a certain threshold.

History and Overview

The history of artificial neural networks begins with a discrete mathematical model of a biological neural network developed by pioneers McCulloch and Pitts in 1943 [32]. This model describes neurons as threshold logic units (TLUs) or binary decision units (BDNs) with multiple binary inputs and a single binary output. A neuron outputs 1 (is activated) if the sum of its unweighted inputs exceeds a certain specified threshold, otherwise it outputs 0. Each neuron can only represent simple logic functions like OR or AND, but any boolean function can be realized by combinations of such neurons.

In 1958 Rosenblatt [41] extended the McCulloch–Pitts model to the perceptron model. This network was based on a unit called the perceptron, which produces an output depending on the weighted linear combination of its inputs. The weights are adapted by the perceptron learning rule. Another single-layer neural network that is based on the McCulloch–Pitts neuron is the ADALINE (ADAptive Linear Element) which was invented in 1960 by Widrow and Hoff [55] and employs a Least-Mean-Squares (LMS) learning rule.

In 1969 Minsky and Papert [33] provided mathematical proofs that single-layer neural networks like the perceptron are incapable of representing functions which are linearly inseparable, including in particular the exclusive-or (XOR) function. This fundamental limitation led the research on neural networks to stagnate for many years, until it was found that a perceptron with more than one layer has far greater processing power.

The backpropagation learning method was first described by Werbos in 1974 [53,54], and further developed for multi-layer neural networks by Rumelhart et al. in 1986 [46]. Backpropagation networks are by far the most well known and most commonly used neural networks today.

Recurrent auto-associative networks were first described independently by Anderson [2] and Kohonen [21] in 1977. Invented in 1982, the Hopfield network [17] is a recurrent neural network in which all connections are symmetric. All neurons are both input and output neurons and update their activation values asynchronously and independently from each other. For each new input the network converges dynamically to a new stable state. A Hopfield network may serve as an associative, i. e., content-addressable, memory.

The Boltzmann machine by Ackley et al. [1] can be seen as an extension of the Hopfield network. It uses a stochastic instead of a deterministic update rule that simulates the physical principle of annealing. The Boltzmann machine is one of the first neural networks to demonstrate learning of an internal representation (hidden units).

It was also in 1982 when Kohonen first published his self-organizing maps [22], a neural networks model based on unsupervised learning and competitive learning. SOMs produce a low-dimensional representation (feature map) of high-dimensional input data while preserving their most important topological features.

Significant progress was made in the 1990s in the field of neural networks which attracted a great deal of attention both in research and in many application domains (see Sect. "Definition of the Subject"). Faster computers allowed a more efficient solving of more complex problems.

Hardware implementations of larger neural networks were realized on parallel computers or in neural network chips with multiple units working simultaneously.

Characteristics of Neural Networks

Interesting general properties of neural networks are that they (1) mimic the way the brain works, (2) are able to learn by experience, (3) make predictions without having to know the precise underlying model, (4) have a high fault tolerance, i. e., can still give the correct output to missing, noisy or partially correct inputs, and (5) can work with data they have never seen before, provided that the underlying distribution fits the training data.

Computation in neural networks is local and highly distributed throughout the network such that each node operates by itself, but tries to minimize the overall network (output) error in cooperation with other nodes. Working like a cluster of interconnected processing nodes, neural networks automatically distribute both the problem and the workload among the nodes in order to find and converge to a common solution. Actually, parallel processing was one of the original motivations behind the development of artificial neural networks.

Neural networks have the topology of a directed graph. There are only one-way connections between nodes, just like in biological nervous systems. A two-way relationship requires two one-way connections. Many different network architectures are used, often with hundreds or thousands of adjustable parameters.

Neural networks are typically organized in layers which each consist of a number of interconnected nodes (see Fig. 2 below). Data patterns are presented to the system via the *input layer*. This non-computing layer is connected to one or more *hidden layers* where the actual processing is done. The hidden layers then link to an *output layer* which combines the results of multiple processing units to produce the final response of the network.

The network acts as a high-dimensional vector function, taking one vector as input and returning another vector as output. The modeled functions are general and complex enough to solve a large class of non-linear classification and estimation problems. No matter which problem domain a neural network is operating in, the input data always have to be encoded into numbers which may be continuous or discrete.

The high number of free parameters in a neural network and the high degree of collinearity between the neuron outputs let individual parameter settings (weight coefficients) become meaningless and make the network for the most part uninterpretable. High-order interactions between neurons also do not permit simpler substructures in the model to be identified. Therefore, an extraction of the acquired knowledge from such *black box* predictors to understand the underlying model is almost impossible.

Neural Network Learning

The human brain learns by practice and experience. The learned knowledge can change if more information is received. Another important element of learning is the ability to infer knowledge, i. e., to make assumptions based on what we know and to apply what we have learned in the past to similar problems and situations. One theory of the physiology of human learning by repetition is that repeated sequences of impulses strengthen connections between neurons and form memory paths. To retrieve the learned information, nerve impulses follow these paths to the correct information. If we get out of practice these paths may diminish over time and we forget what we have learned.

The information stored in a neural network is contained in its free parameters. In general, the network architecture and connections are held constant and only the connection weights are variable during training. Once the numbers of (hidden) layers and units have been selected, the free parameters (weights) are set to fit the model or function represented by the network to the training data, following a certain *training algorithm* or *learning rule*. A neural network learns, i. e., acquires knowledge, by adjusting the weights of connections between its neurons. This is also referred to as *connectionist learning*. Training occurs iteratively in multiple cycles during which the training examples are repeatedly presented to the network. During one *epoch* all data patterns pass through the network once.

The three major *learning paradigms* include supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning, in general, means learning by an external teacher using global information. The problem the network is supposed to solve is defined through a set of training examples given. The learning algorithm searches the solution space F, the class of possible functions, for a function $f^* \in F$ that matches this set of input–output associations $(\vec{x}, \vec{y})$ best. In other words, the mapping implied by the sample data has to be inferred.

Training a neural network means to determine a set of weights which minimizes its prediction error on the training set. The *cost* or *error function* $E\colon F \to \mathbb{R}$ measures the error between the desired output values $\vec{y}$ and the predicted network outputs $f(\vec{x})$ over all input vec-

tors $\vec{x}$. That means, it calculates how far away the current state f of the network is from the optimal solution f^* with $E(f^*) \leq E(f)\, \forall f \in F$.

A neural network cannot perfectly learn a mapping if the input data does not contain enough information to derive the desired outputs. It may also not converge if there is not enough data available (see also below).

Unsupervised learning uses no external teacher and only local information. It is distinguished from supervised learning by the fact that there is no a priori output. In unsupervised learning we are given some input data $\vec{x}$, and the cost function to be minimized can be any function of $\vec{x}$ and the network output $f(\vec{x})$. Unsupervised learning incorporates self-organization, i. e., organizes the input data by using only their inherent properties to reveal their emergent collective properties.

A neural network learns *offline* if learning phase and operation (application) phase are separated. A neural network learns *online* if both happens at the same time. Usually, supervised learning is performed offline, whereas unsupervised learning is performed online.

In *reinforcement learning*, neither inputs $\vec{x}$ nor outputs $\vec{y}$ are given explicitly, but are generated by the interactions of an agent within an environment. The agent performs an action $\vec{y}$ with costs c according to an observation $\vec{x}$ made in the environment. The aim is to discover a policy or plan for selecting actions that minimizes some measure of the expected total costs.

Overtraining and Generalization

The overall motivation and most desirable property of neural networks is their ability to generalize to new unknown data, i. e., to classify patterns correctly on which they have not been trained. Minimizing the network error on the training examples only, does not automatically minimize the real error of the unknown underlying function. This important problem is called *overfitting* or *overtraining*.

A regular distribution of training examples over the input data space is important. Generalization is reasonable only as long as the data inputs remain inside the range for which the network was trained. If the training set only included vectors from a certain part of the data space, predictions on other parts are random and likely wrong.

Overtraining may occur also when the iterative training algorithm is run for too long and if the network is too complex for the problem to solve or the available quantity of data. A larger neural network with more weights models a more complex function and invariably achieves a lower error, but is prone to overfitting. A network with less weights, on the other hand, may not be sufficiently powerful to model the underlying function.

A simple heuristic, called *early stopping*, helps to ensure that the network will generalize well to examples not in the training set. One solution is to check progress during training against an independent data set, the *validation set*. As training progresses, the training error naturally decreases monotonically and, providing training is minimizing the true error function, also the validation error decreases. However, if the validation error stops dropping or even starts to increase again, this is an indication that the network is starting to overfit the data. Then the optimization process has become stuck in a local minima and training should be stopped. The weights that produced the minimum validation error are then used for the final model.

In this case of overtraining, the size of the network, i. e., the number of hidden units and/or hidden layers, may be decreased. Neural networks typically involve experimenting with a large number of different configurations, training each one a number of times while observing the validation error. A problem with repeated experimentation is that the validation set is actually part of the training process. One may just find a network by chance that happens to perform well on the validation set. It is therefore normal practice to reserve a third set of examples for testing the final model on this *test set*.

In many cases a sufficient amount of data is not available, however. Then we have to get around this problem by resampling techniques, like *cross validation*. In principle, multiple experiments are conducted, each using a different division of the available data into training and validation set. This should remove any sampling bias. For small data sets, where splitting the data would leave too few observations for training, *leave-one-out validation* may be used to determine when to stop training or the optimal network size.

Machine-learning techniques, like neural networks, require both positive and negative training examples for solving classification problems. Because they minimize an overall error, the proportion of positive and negative examples in the training set is critical. Ideally, the relation should be close to the (usually unknown) real distribution in the data space. Otherwise, it may bias the network's decision to be more often wrong on unknown data.

Feedforward Neural Networks

In *feedforward neural networks* the information is passed in only one direction (forward) from the inputs, through the hidden nodes (if any) to the output nodes. There are

no connections backwards to neurons of upper layers i. e., there are no feedback loops or cycles in the network.

In feedforward networks with a single-layer of weights, the inputs are directly connected to the output units (*single-layer neural networks*). *Multi-layer feedforward networks* use additional intermediate layers of hidden units. Neural networks with two or more processing layers may have far greater processing power than networks with only one layer. Single-layer neural networks are only capable of learning linearly separable patterns and functions.

The Perceptron

The most simple kind of feedforward neural network is a *perceptron*, which consists of a single pseudo input layer and one or more processing nodes in the output layer. All inputs are weighted and fed directly to the output neuron(s) (see Fig. 1). Each node calculates the sum of the products of weights and inputs. If this value is above some threshold (typically 0) the neuron takes the activated value 1, otherwise it outputs 0. Neurons with this kind of activation function are called *threshold units*. In the literature the term *perceptron* often refers to networks consisting of only one (output) neuron.

More formally, the perceptron is a linear binary classifier that maps a binary n-dimensional input vector $\vec{x} \in \{0,1\}^n$ to a binary output value $f(\vec{w} \cdot \vec{x}) \in \{0,1\}$ calculated as

$$f(s) = \begin{cases} 1 & \text{if } s > T \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $s = \vec{w} \cdot \vec{x} = \sum_{i=1}^{n} w_i x_i$ is the input sum and $\vec{w}$ is a vector of real-valued weights. The constant threshold T does not depend on any input value.

In addition to the network topology, the learning rule is an important component of neural networks. Perceptrons can be trained by a simple learning algorithm, called the *delta rule* or *perceptron learning rule*. This realizes a simple stochastic *gradient descent* where the weights of the network are adjusted depending on the error between the predicted outputs of the network and the example outputs. The delta rule changes the weight vector such that the output error is minimized. McClelland and Rumelhart [30] proved that a neural network using the delta rule can learn associations whenever the inputs are linearly independent.

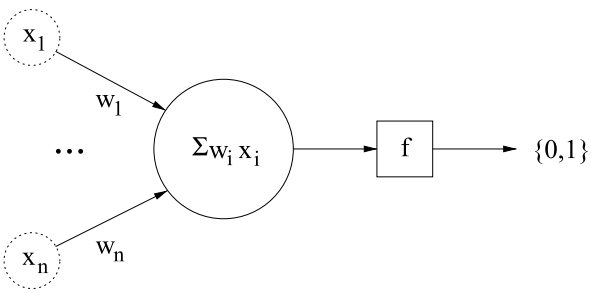

Data-Mining and Knowledge Discovery, Neural Networks in, Figure 1
Principle structure of single-unit perceptron network

All neurons of a perceptron share the same structure and learning algorithm. Each weight w_{ij}, representing the influence of input x_i on neuron j, is updated at time t according to the rule:

$$w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij} \tag{2}$$

$$\Delta w_{ij} = \alpha(o_i - y_i)x_{ij} \,. \tag{3}$$

The network learns by updating the weight vector after each iteration (training example) by an amount proportional to the difference between given output o_i and calculated output $y_i = f(s_i)$. The *learning rate* α is a constant with $0 < \alpha < 1$ and regulates the learning speed.

The training data set is *linearly separable* in n-dimensional data space if its two-classes of vectors $\vec{x}$ can be separated by an $(n-1)$-dimensional hyperplane. If the training examples are not linearly separable, the perceptron learning algorithm is not guaranteed to converge. Linear classifiers, like single-unit perceptrons, are only able to learn, i. e., perfectly classify, linearly separable patterns because they can only implement a simple decision surface (single hyperplane) [33]. The same is true for single-layer neural networks with more than one output unit. This makes these *linear neural networks* unable to learn, for example, the XOR function [33]. Nevertheless a problem that is thought to be highly complex may still be solved as well by a linear network as by a more powerful (non-linear) neural network.

Single-Layer Neural Networks

In general, the state of a neuron is represented by its activation value. An *activation* or *transfer function* f calculates the activation value of a unit from the weighted sum s of its inputs. In case of the perceptron f is called *step* or *threshold function*, with the activation value being 1 if the network sum is greater than a constant T, and 0 otherwise (see Eq. (1)). Another common form of non-linear activation function is the *logistic* or *sigmoid function*:

$$f(s) = \frac{1}{1 + \mathrm{e}^{-s}} \,. \tag{4}$$

This enables a neural network to compute a continuous output between 0 and 1 instead of a step function. With

this choice, a single-layer network is identical to a logistic regression model. If the activation functions is linear, i. e., the *identity*, then this is just a multiple linear regression and the output is proportional to the total weighted sum s.

Multi-Layer Neural Networks

The limitation that non-linearly separable functions cannot be represented by a single-layer network with fixed weights can be overcome by adding more layers. A multi-layer network is a feedforward network with two or more layers of computational units, interconnected such that the neurons' outputs of one layer serve as inputs only to neurons of the directly subsequent layer (see Fig. 2). The input layer is not considered a real layer with processing neurons.

The number of units in the input layer is determined by the problem, i. e., the dimension of the input data space. The number of output units also depends on the output encoding (see Subsect.“ Application Issues”).

By using hidden layers, the partitioning of the data space can be more effective. In principle, each hidden unit adds one hyperplane to divide the space and discriminate the solution. Only if the outputs of at least two neurons are combined in a third neuron, the XOR problem is solvable. Important issues in multi-layer NN design are, thus, the specification of the number of hidden layers and the number of units in these layers (see also Subsect. “Application Issues”). Both numbers determine the complexity of functions that can be modeled. There is no theoretical limitation on the number of hidden layers, but usually one or two are used.

The *universal approximation theorem* for neural networks states that any continuous function that maps intervals of real numbers to an output interval of real numbers can be approximated arbitrarily closely by a multi-layer neural network with only one hidden layer and certain types of non-linear activation functions. This gives, however, no indication about how fast or likely a solution is found. Networks with two hidden layers may work better for some problems. However, more than two hidden layers usually provide only marginal benefit compared to the significant increase in training time.

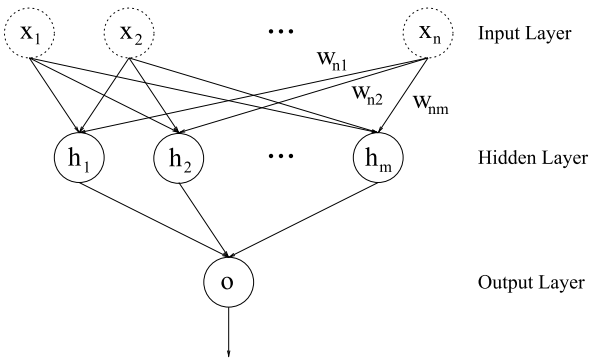

Data-Mining and Knowledge Discovery, Neural Networks in, Figure 2
Principle structure of multi-layer feedforward neural network

Any multi-layer network with fixed weights and linear activation function is equivalent to a single-layer (linear) network: In the case of a two-layer linear system, for instance, let all input vectors to the first layer form matrix X and W_1 and W_2 be the weight matrices of the two processing layers. Then the output $Y_1 = W_1 \cdot X$ of the first layer is input to the second layer, which produces output $Y_2 = W_2 \cdot (W_1 \cdot X) = (W_2 \cdot W_1) \cdot X$. This is equivalent to a single-layer network with weight matrix $W = W_2 \cdot W_1$.

Only a multi-layer network that is non-linear can provide more computational power. In many applications these networks use a sigmoid function as non-linear activation function. This is the case at least for the hidden units. For the output layer, the sigmoid activation function is usually applied with classification problems, while a linear transfer function is applied with regression problems.

Error Function and Error Surface

The *error function* derives the overall network error from the difference of the network's output y_{ij} and target output o_{ij} over all examples i and output units j. The most-common error function is the *sum squared error*:

$$E = \frac{1}{2} \sum_i \sum_j (o_{ij} - y_{ij})^2 \,. \tag{5}$$

Neural network training performs a search within the space of solution, i. e., all possible network configurations, towards a global minimum of the error surface. The *global minimum* is the best overall solution with the lowest possible error.

A helpful concept for understanding NN training is the *error surface*: The n weight parameters of the network model form the n dimensions of the search space. For any possible state of the network or configuration of weights the error is plotted in the $(n+1)$th dimension. The objective of training is to find the lowest point on this n-dimensional surface. The error surface is seldom smooth. Indeed, for most problems, the surface is quite rugged with numerous hills and valleys which may cause the network search to run into a *local minimum*, i. e., a Suboptimum solution.

The speed of learning is the rate of convergence between the current solution and the global minimum.

In a linear network with a sum-squared error function, the error surface is an multi-dimensional parabola, i. e., has only one minimum. In general, it is not possible to analytically determine where the global minimum of the error surface is. Training is essentially an exploration of the error surface. Because of the probabilistic and often highly non-linear modeling by neural networks, we cannot be sure that the error could not be lower still, i. e., that the minimum we found is the absolute one. Since the shape of the error space cannot be known a priori, neural network analysis requires a number of independent runs to determine the best solution. When different initial values for the weights are selected, different network models will be derived.

From an initially random configuration of the network, i. e., a random point on the error surface, the training algorithm starts seeking for the global minimum. Small random values are typically used to initialize the network weights. Although neural networks resulting from different initial weights may have very different parameter settings, their prediction errors usually do not vary dramatically. Training is stopped when a maximum number of epochs has expired or when the network error does not improve any further.

Backpropagation

The best known and most popular training algorithm for multi-layer networks is *backpropagation*, short for backwards error propagation and also referred to as the *generalized delta rule* [46]. The algorithm involves two phases:

Forward pass. During the first phase, the free parameters (weights) of the network are fixed. An example pattern is presented to the network and the input signals are propagated through the network layers to calculate the network output at the output unit(s).

Backward pass. During the second phase, the model parameters are adjusted. The error signals at the output units, i. e., the differences between calculated and expected outputs, are propagated back through the network layers. In doing so, the error at each processing unit is calculated and used to make adjustments to its connecting weights such that the overall error of the network is reduced by some small amount.

After iteratively repeating both phases for a sufficiently large number of training cycles (epochs) the network will converge to a state where its output error is small enough. The backpropagation rule involves the repeated use of the *chain rule*, saying that the output error of a neuron can be ascribed partly to errors in the weights of its direct inputs and partly to errors in the outputs of higher-level (hidden) nodes [46].

Moreover, backpropagation learning may happen in two different modes. In *sequential mode* or *online mode* weight adjustments are made example by example, i. e., each time an example pattern has been presented to the network. The *batch mode* or *offline mode* adjustments are made epoch by epoch, i. e., only after all example patterns have been presented. Theoretically, the backpropagation algorithm performs gradient descent on the total error only if the weights are updated epoch-wise. There are empirical indications, however, that a pattern-wise update results in faster convergence. The training examples should be presented in random order. Then the precision of predictions will be more similar over all inputs.

Backpropagation learning requires a differentiable activation function. Besides adding non-linearity to multi-layer networks, the sigmoid activation function (see Eq. (4)) is often used in backpropagation networks because it has a continuous derivative that can be calculated easily:

$$f'(s) = f(s)(1 - f(s)) \,. \tag{6}$$

We further assume that there is only one hidden layer in order to keep notations and equations clear. A generalization to networks with more than one hidden layer is straightforward.

The backpropagation rule is a generalization of the delta learning rule (see Eq. (3)) to multi-layer networks with non-linear activation function. For an input vector $\vec{x}$ the output $y = f(s)$ is calculated at each output neuron of the network and compared with the desired target output o, resulting in an error δ. Each weight is adjusted proportionally to its effect on the error. The weight of a connection between a unit i and a unit j is updated depending on the output of i (as input to j) and the error signal at j:

$$\Delta w_{ij} = \alpha \delta_j y_i \,. \tag{7}$$

For an output node j the error signal (error surface gradient) is given by:

$$\delta_j = (o_j - y_j) f'(s_j) = (o_j - y_j) y_j (1 - y_j) \,. \tag{8}$$

If the error is zero, no changes are made to the connection weight. The larger the absolute error, the more the responsible weight is changed, while the sign of the error determines the direction of change.

For a hidden neuron j the error signal is calculated recursively using the signals of all directly connected output neurons k.

$$\delta_j = f'(s_j) \sum_k \delta_k w_{jk} = y_j (1 - y_j) \sum_k \delta_k w_{jk} \,. \tag{9}$$

The partial derivative of the error function with respect to the network weights can be calculated purely locally, such that each neuron needs information only from neurons directly connected to it. A theoretical foundation of the backpropagation algorithm can be found in [31].

The backpropagation algorithm performs a *gradient descent* by calculating the gradient vector of the error surface at the current search point. This vector points into the direction of the steepest descent. Moving in this direction will decrease the error and will eventually find a new (local) minimum, provided that the *step size* is adapted appropriately. Small steps slow down learning speed, i. e., require a larger number of iterations. Large steps may converge faster, but may also overstep the solution or make the algorithm oscillate around a minimum without convergence of the weights. Therefore, the step size is made proportional to the slope δ, i. e., is reduced when the search point approaches a minimum, and to the learning rate α. The constant α allows one to control the size of the gradient descent step and is usually set to be between 0.1 and 0.5. For practical purposes, it is recommended to choose the learning rate as large as possible without leading to oscillation.

Momentum

One possibility to avoid oscillation and to achieve faster convergence is in the addition of a *momentum term* that is proportional to the previous weight change:

$$\Delta w_{ij}(t+1) = \alpha \delta_j y_i + \beta \Delta w_{ij}(t) \,. \tag{10}$$

The algorithm increases learning speed step size if it has taken several steps in the same direction. This gives it the ability to overcome obstacles in the error surface, e. g., to avoid and escape from local minima, and to move faster over larger plateaus.

Finding the optimum learning rate α and momentum scale parameter β, i. e., the best trade-off between longer training time and instability, can be difficult and might require many experiments. Global or local adaptation techniques use, for instance, the partial derivative to automatically adapt the learning rate. Examples here are the Delta-Bar-Delta rule [18] and the SuperSAB algorithm [51].

Other Learning Rules

The backpropagation learning algorithm is computationally efficient in that its time complexity is linear in the number of weight parameters. Its learning speed is comparatively low, however, on the basis of epochs. This may result in long training times, especially for difficult and complex problems requiring larger networks or larger amounts of training data.

Another major limitation is that backpropagation does not always converge. Still, it is a widely used algorithm and has its advantages: It is relatively easy to apply and to configure and provides a quick, though not absolutely perfect solution. Its usually pattern-wise error adjustment is hardly affected by data that contains a larger number of redundant examples. Standard backpropagation also generalizes equally well on small data sets as more advanced algorithms, e. g., if there is insufficient information available to find a more precise solution.

There are many variations of the backpropagation algorithm, like resilient propagation (Rprop) [42], quick propagation (Quickprop) [13], conjugate gradient descent [6], Levenberg–Marquardt [16], Delta-Bar-Delta [18], to mention the most popular. All these second-order algorithms are designed to deal with some of the limitations on the standard approach. Some work substantially faster in many problem domains, but require more control parameters than backpropagation, which makes them more difficult to use.

Resilient Propagation

Resilient propagation (Rprop) as proposed in [42,43] is a variant of standard backpropagation with very robust control parameters that are easy to adjust. The algorithm converges faster than the standard algorithm without being less accurate.

The size of the weight step Δw_{ij} taken by standard backpropagation not only depends on the learning rate α, but also on the size of the partial derivative (see Eq. (7)). This may have an unpredictable influence during training that is difficult to control. Therefore, Rprop uses only the sign of derivative to adjust the weights. It necessarily requires learning by epoch, i. e., all adjustments take place after each epoch only.

One iteration of the Rprop algorithm involves two steps, the adjustment of the step size and the update of the weights. The amount of weight change is found by the following update rule:

$$\Delta_{ij}(t) = \begin{cases} \eta^+ \Delta_{ij}(t-1) & \text{if } d_{ij}(t-1) \cdot d_{ij}(t) > 0 \\ \eta^- \Delta_{ij}(t-1) & \text{if } d_{ij}(t-1) \cdot d_{ij}(t) < 0 \\ \Delta_{ij}(t-1) & \text{otherwise} \end{cases} \tag{11}$$

with $0 < \eta^- < 1 < \eta^+$ and derivative $d_{ij} = -\delta_j y_i$. Every time t the derivative term changes its sign, indicating that the last update (at time $t-1$) was too large and the algo-

rithm has jumped over a local minimum, the update value (step size) $\Delta_{ij}(t-1)$ is decreased by a constant factor η^-.

The rule for updating the weights is straightforward:

$$w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij}(t) \tag{12}$$

$$\Delta w_{ij}(t) = \begin{cases} -\Delta_{ij}(t) & \text{if } d_{ij}(t) > 0 \\ +\Delta_{ij}(t) & \text{if } d_{ij}(t) < 0 \\ 0 & \text{otherwise} \end{cases} . \tag{13}$$

One advantage of the Rprop algorithm, compared to, for example, Quickprop [13], is its small set of parameters that hardly requires adaptation. Standard values for decrease factor η^- and increase factor η^+ are 0.5 and 1.2, respectively. To avoid too large or too small weight values, this is bounded above by $\Delta_{\max}$ and bounded below by $\Delta_{\min}$, set by default to 50 and 10^{-6}. The same initial value $\Delta_0 = 0.1$ is recommended for all Δ_{ij}. While the choice of the parameter settings is not critical, for most problems no other choice is needed to obtain the optimal or at least a nearly optimal solution.

Application Issues

The architecture of a neural network, i.e, the number of (hidden) neurons and layers, is an important decision. If a neural network is highly redundant and overparameterized, it might adapt too much to the data. Thus, there is a trade-off between reducing bias (fitting the training data) and reducing variance (fitting unknown data). The most-common procedure is to select a network structure that has more than enough parameters and neurons and then to avoid overfitting only over the training algorithm (see Subsect. "Overtraining and Generalization").

There is no general best network structure for a particular type of application. There are only general rules for selecting the network architecture: (1) The more complex the relationships between input and output data are, the higher the number of hidden units should be selected. (2) If the modeled process is separable into multiple stages, more than one hidden layer may be beneficial. (3) An upper bound for the total number of hidden units may be set by the number of data examples divided by the number of input and output units and multiplied by a scaling factor. A simpler rule is to start with one hidden layer and half as many hidden units as there are input and output units.

One would expect that for a given data set there would be an optimal network size, lying between a minimum of one hidden neuron (high bias, low variance) and a very large number of neurons (low bias, high variance). While this is true for some data sets, in many cases increasing the number of hidden nodes continues to improve prediction accuracy, as long as cross validation is used to stop training in time.

For classification problems, the neural network assigns to each input case a class label or, more generally, estimates the probability of the case to fall into each class. The various output classes of a problem are normally represented in neural networks using one of two techniques, including *binary encoding* and *one-out-of-n encoding*.

A binary encoding is only possible for two-class problems. A single unit calculates class 1 if its output is above the acceptance threshold. If the output is below the rejection threshold, class 0 is predicted. Otherwise, the output class is undecided. Should the network output be always defined, both threshold values must be equal (e. g. 0.5).

In one-out-of-n encoding one unit is allocated for each class. A class is selected if the corresponding output is above the acceptance threshold and all the other outputs are below the rejection threshold. If this condition is not met, the class is undecided. Alternatively, instead of using a threshold, a *winner-takes-all* decision may be made such that the unit with the highest output gives the class.

For regression problems, the objective is to estimate the value of a continuous output variable, given the input variables. Particularly important issues in regression are output scaling and interpolation. The most-common NN architectures produce outputs in a limited range. Scaling algorithms may be applied to the training data to ensure that the target outputs are in the same range. Constraining the network's outputs limits its generalization performance. To overcome this, a linear activation function may be used for the output units. Then there is often no need for output scaling at all, since the units can in principle calculate any value.

Other Neural Network Architectures

This section summarizes some alternative NN architectures that are variants or extensions of multi-layer feed-forward networks.

Cascade Correlation Networks

Cascade-correlation is a neural network architecture with variable size and topology [14]. The initial network has no hidden layer and grows during training by adding new hidden units one at a time. In doing so, a near minimal network topology is built. In the cascade architecture the outputs from all existing hidden neurons in the network are fed into a new neuron. In addition, all neurons – including the output neurons – receive all input values.

For each new hidden unit, the learning algorithm tries to maximize the correlation between this unit's output and

the overall network error using an ordinary learning algorithm, like, e. g., backpropagation. After that the input-side weighs of the new neuron are frozen. Thus, it does not change anymore and becomes a permanent feature detector.

Cascade correlation networks have several advantages over multi-layer perceptrons: (1) Training time is much shorter already because the network size is relatively small. (2) They require only little or no adjustment of parameters, especially not in terms of the number of hidden neurons to use. (3) They are more robust and training is less likely to become stuck in local minima.

Recurrent Networks

A network architecture with cycles is adopted by *recurrent* or *feedback neural networks* such that outputs of some neurons are fed back as extra inputs. Because past outputs are used to calculate future outputs, the network is said to "remember" its previous state. Recurrent networks are designed to process sequential information, like time series data. Processing depends on the state of the network at the last time step. Consequently, the response to the current input depends on previous inputs.

Two similar types of recurrent network are extensions of the multi-layer perceptron: *Jordan networks* [19] feed back all network outputs into the input layer; *Elman networks* [12] feed back from the hidden units. State or context units are added to the input layer for the feedback connections which all have constant weight one. At each time step t, an input vector is propagated in a standard feedforward fashion, and then a learning rule (usually backpropagation) is applied. The extra units always maintain a copy of the previous outputs at time step $t-1$.

Radial Basis Function Networks

Radial basic function (RBF) networks [8,34,37,38] are another popular variant of two-layer feedforward neural networks which uses radial basis functions as activation functions. The idea behind radial basis functions is to approximate the unknown function $f(\vec{x})$ by a weighted sum of non-linear basis functions ϕ, which are often Gaussian functions with a certain standard deviation σ.

$$f(\vec{x}) = \sum_i w_i \phi(||\vec{x} - \vec{c}_i||) \tag{14}$$

The basis functions operate on the Euclidean distance between n-dimensional input vector $\vec{x}$ and center vector $\vec{c}_i$. Once the center vectors $\vec{c}_i$ are fixed, the weight coefficients w_i are found by simple linear regression.

The architecture of RBF networks is fixed to two processing layers. Each unit in the hidden layer represents a center vector and a basis function which realizes a non-linear transformation of the inputs. Each output unit calculates a weighted sum (linear combination) of the non-linear outputs from the hidden layer. Only the connections between hidden layer and output layer are weighted.

The use of a linear output layer in RBF networks is motivated by *Cover's theorem* on the separability of patterns. The theorem states that if the transformation from the data (input) space to the feature (hidden) space is non-linear and the dimensionality of the feature space is relatively high compared to that of the data space, then there is a high likelihood that a non-separable pattern classification task in the input space is transformed into a linearly separable one in the feature space.

The center vectors are selected from the training data, either randomly or uniformly distributed in the input space. In principle, as many centers (and hidden units) may be used as there are data examples. Another method is to group the data in space using, for example, k-means clustering, and select the center vectors close to the cluster centers.

RBF learning is considered a curve-fitting problem in high-dimensional space, i. e., approximates a surface with the basis functions that fits and interpolates the training data points best. The basis functions are well-suited to online learning applications, like adaptive process control. Adapting the network to new data and changing data statistics only requires a retraining by linear regression which is fast. RBF networks are more local approximators than multi-layer perceptrons. New training data from one region of the input space have less effect on the learned model and its predictions in other regions.

Self-organizing Maps

A *self-organizing map* (SOM) or *Kohonen map* [22,23] applies an *unsupervised* and *competitive* learning scheme. That means that the class labels of the data vectors are unknown or not used for training and that each neuron improves through competition with other neurons. It is a non-deterministic machine-learning approach to data clustering that implements a mapping of the high-dimensional input data into a low-dimensional feature space. In doing so, SOMs filter and compress information while preserving the most relevant features of a data set. Complex, non-linear relationships and dependencies are revealed between data vectors and between clusters and are transformed into simple geometric distances. Such an ab-

straction facilitates both visualization and interpretation of the clustering result.

Typically, the units of a SOM network are arranged in a two-dimensional regular grid, the *topological feature map*, which defines a two-dimensional Euclidean distance between units. Each unit is assigned a center vector from the n-dimensional data space and represents a certain cluster.

Algorithm 1 describes the basic principle behind SOM training. Starting with an initially random set of center vectors, the algorithm iteratively adjusts them to reflect the clustering of the training data. In doing so, the two-dimensional order of the SOM units is imposed on the input vectors such that more similar clusters (center vectors) in the input space are closer to each other on the two-dimensional grid structure than more different clusters. One can think of the topological map to be folded and distorted into the n-dimensional input space, so as to preserve as much as possible the original structure of the data.

Algorithm 1 (Self-organizing Map)

1. Initialize the n-dimensional center vector $\vec{c}_i \in \mathbb{R}^n$ of each cluster randomly.
2. For each data point $\vec{p} \in \mathbb{R}^n$ find the nearest center vector $\vec{c}_w$ (called the "winner") in n-dimensional space according to a distance metric d.
3. Move $\vec{c}_w$ and all centers $\vec{c}_i$ within its local neighborhood closer to $\vec{p}$:

$$\vec{c}_i(t+1) = \vec{c}_i(t) + \alpha(t) \cdot h_{r(t)} \cdot (\vec{p} - \vec{c}_i(t))$$

   with learning rate $0 < \alpha < 1$ and neighborhood function h depending on a neighborhood radius r.
4. $$\alpha(t+1) = \alpha(t) - \Delta\alpha \quad \text{where} \quad \Delta\alpha = \alpha_0/t_{\max}$$
   $$r(t+1) = r(t) - \Delta r \quad \text{where} \quad \Delta r = r_0/t_{\max}$$
5. Repeat steps 2.–4. for each epoch $t = 1, .., t_{\max}$.
6. Assign each data point to the cluster with the nearest center vector.

Each iteration involves randomly selecting a data point $\vec{p}$ and moving the closest center vector a bit in the direction of $\vec{p}$. Only distance metric d (usually Euclidean) defined on the data space influences the selection of the closest cluster. The adjustment of centers is applied not just to the winning neuron, but to all the neurons of its neighborhood. The *neighborhood function* is often Gaussian. A simple definition calculates $h = 1$ if the Euclidean distance between the grid coordinates of the winning cluster w and a cluster i is below a radius r, i. e., $||(x_w, y_w) - (x_i, y_i)|| < r$, and $h = 0$ otherwise.

Both learning rate α and *neighborhood radius* r monotonically decrease over time. Initially quite large areas of the network are affected by the neighborhood update, leading to a rather rough topological order. As epochs pass, less neurons are altered with lower intensity and finer distinctions are drawn within areas of the map.

Unlike hierarchical clustering [52] and k-means clustering [29], which are both deterministic – apart from the randomized initialization in k-means clustering – and operate only locally, SOMs are less likely to become stuck in local minima and have a higher robustness and accuracy.

Once the network has been trained to recognize structures in the data, it can be used as a visualization tool and for exploratory data analysis. The example map in Fig. 3 shows a clustering of time series of gene expression values [48]. Clearly higher similarities between neighboring clusters are revealed when comparing the mean vectors.

If neurons in the feature map can be labeled, i. e., if common meaning can be derived from the vectors in a clusters, the network becomes capable of classification. If the winning neuron of an unknown input case has not been assigned a class label, labels of clusters in close or direct neighborhood may be considered. Ideally, higher similarities between neighboring data clusters are reflected in similar class labels. Alternatively, the network output is undefined in this case. SOM classifiers also make use of the distance of the winning neuron from the input case. If this distance exceeds a certain maximum threshold, the SOM is regarded as undecided. In this way, a SOM can be used for detecting novel data classes.

Future Directions

To date, neural networks are widely accepted as an alternative to classical statistical methods and are frequently used in medicine [4,5,11,25,27,44,45] with many applications related to cancer research [10,26,35,36,49]. In the first place, these comprise diagnostics and prognosis (i. e. classification) tasks, but also image analysis and drug design. Cancer prediction is often based on clustering of gene expression data [15,50] or microRNA expression profiles [28] which may involve both self-organizing maps and multi-layer feedforward neural networks.

Another broad application field of neural networks today is bioinformatics and, in particular, the analysis and classification of gene and protein sequences [3,20,56]. A well-known successful example is protein (secondary) structure prediction from sequence [39,40].

Even though the NN technology is clearly established today, the current period is rather characterized by stagnation. This is partly because of a redirection of research

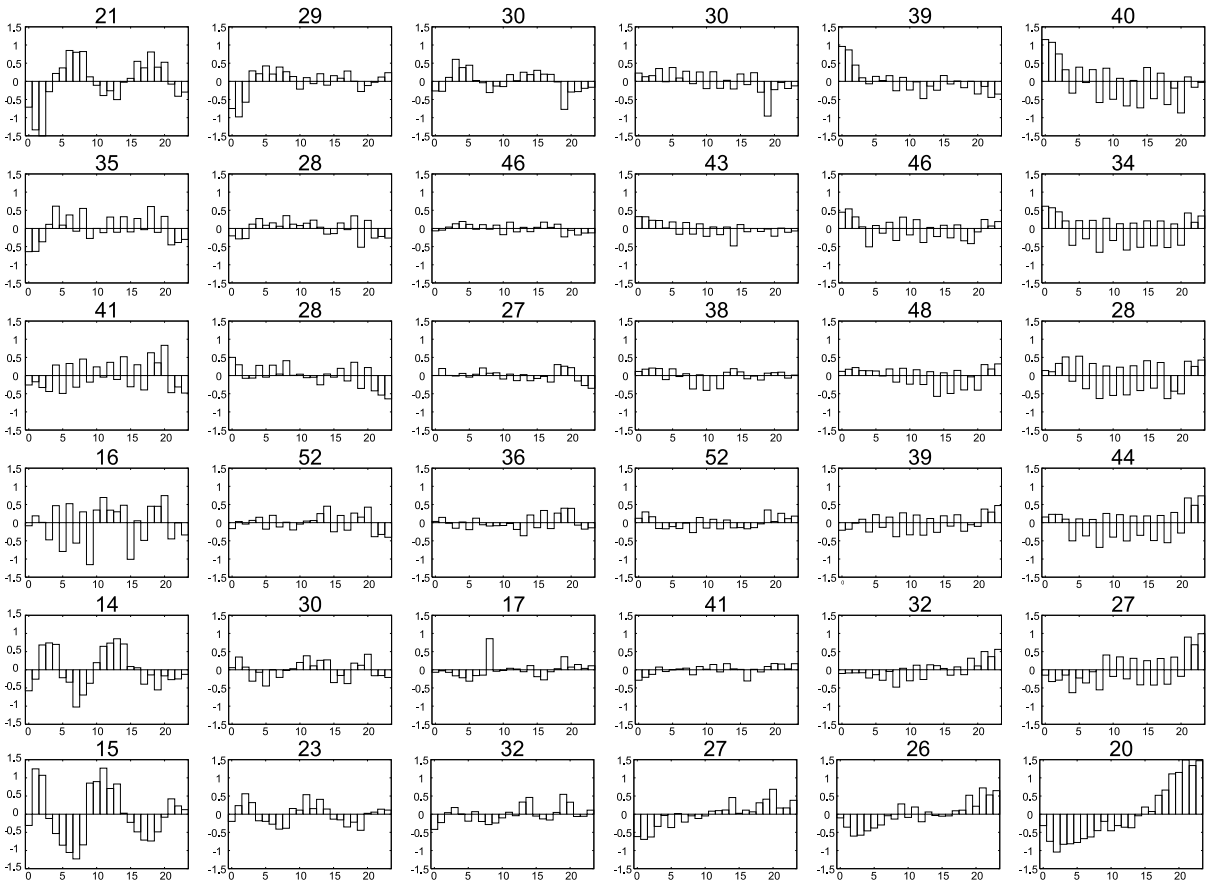

Data-Mining and Knowledge Discovery, Neural Networks in, Figure 3
6 x 6 SOM example clustering of gene expression data (time series over 24 time points) [48]. Mean expression vector plotted for each cluster. Cluster sizes indicate the number of vectors (genes) in each cluster

to newer and often – but not generally – more powerful paradigms, like the popular *support vector machines* (SVMs) [9,47], or to more open and flexible methods, like *genetic programming* (GP) [7,24]. In many applications, for example, in bioinformatics, SVMs have already replaced conventional neural networks as the state-of-the-art black-box classifier.

Bibliography

Primary Literature

1. Ackley DH, Hinton GF, Sejnowski TJ (1985) A learning algorithm for Boltzman machines. Cogn Sci 9:147–169
2. Anderson JA et al (1977) Distinctive features, categorical perception, and probability learning: Some applications of a neural model. Psychol Rev 84:413–451
3. Baldi P, Brunak S (2001) Bioinformatics: The machine learning approach. MIT Press, Cambridge
4. Baxt WG (1995) Applications of artificial neural networks to clinical medicine. Lancet 346:1135–1138
5. Begg R, Kamruzzaman J, Sarkar R (2006) Neural networks in healthcare: Potential and challenges. Idea Group Publishing, Hershey
6. Bishop CM (1995) Neural networks for pattern recognition. Oxford University Press, London
7. Brameier M, Banzhaf W (2007) Linear genetic programming. Springer, New York
8. Broomhead DS, Lowe D (1988) Multivariable functional interpolation and adaptive networks. Complex Syst 2:321–355
9. Cristianini N (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, London
10. Dybowski R (2000) Neural computation in medicine: Perspectives and prospects. In: Malmgren H, Borga M, Niklasson L (eds) Proceedings of the Conference on Artificial Neural Networks in Medicine and Biology (ANNIMAB). Springer, Berlin, pp 26–36
11. Dybowski R, Gant V (2001) Clinical Applications of Artificial Neural Networks. Cambridge University Press, London
12. Elman JL (1990) Finding structure in time. Cogn Sci 14:179–211
13. Fahlman SE (1989) Faster learning variations on backpropagation: An empirical study. In: Touretzky DS, Hinton GE, Sejnowski TJ (eds) Proceedings of the (1988) Connectionist Models Summer School. Morgan Kaufmann, San Mateo, pp 38–51

14. Fahlman SE, Lebiere C (1990) The cascade-correlation learning architecture. In: Touretzky DS (ed) Advances in Neural Information Processing Systems 2. Morgan Kaufmann, Los Altos
15. Golub TR et al (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537
16. Hagan MT, Menhaj M (1994) Training feedforward networks with the Marquardt algorithm. IEEE Trans Neural Netw 5(6):989–993
17. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci 79(8):2554–2558
18. Jacobs RA (1988) Increased rates of convergence through learning rate adaptation. Neural Netw 1:295–307
19. Jordan MI (1986) Attractor dynamics and parallelism in a connectionist sequential machine. In: Proceedings of the Eighth Annual Conf of the Cogn Sci Society. Lawrence Erlbaum, Hillsdale, pp 531–546
20. Keedwell E, Narayanan A (2005) Intelligent bioinformatics: The application of artificial intelligence techniques to bioinformatics problems. Wiley, New York
21. Kohonen T (1977) Associative Memory: A System-Theoretical Approach. Springer, Berlin
22. Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biol Cybern 43:59–69
23. Kohonen T (1995) Self-organizing maps. Springer, Berlin
24. Koza JR (1992) Genetic programming: On the programming of computer programs by natural selection. MIT Press, Cambridge
25. Lisboa PJG (2002) A review of evidence of health benefit from artificial neural networks in medical intervention. Neural Netw 15:11–39
26. Lisboa PJG, Taktak AFG (2006) The use of artificial neural networks in decision support in cancer: a systematic review. Neural Netw 19(4):408–415
27. Lisboa PJG, Ifeachor EC, Szczepaniak PS (2001) Artificial neural networks in biomedicine. Springer, Berlin
28. Lu J et al (2005) MicroRNA expression profiles classify human cancers. Nature 435:834–838
29. MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, vol 1. University of California Press, Berkeley, pp 281–297
30. McClelland JL, Rumelhart DE (1986) Parallel distributed processing: Explorations in the microstructure of cognition. MIT Press, Cambridge
31. McClelland J, Rumelhart D (1988) Explorations in parallel distributed processing. MIT Press, Cambridge
32. McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 5:115–133
33. Minsky ML, Papert SA (1969/1988) Perceptrons. MIT Press, Cambridge
34. Moody J, Darken CJ (1989) Fast learning in networks of locally-tuned processing units. Neural Comput 1:281–294
35. Naguib RN, Sherbet GV (1997) Artificial neural networks in cancer research. Pathobiology 65(3):129–139
36. Naguib RNG, Sherbet GV (2001) Artificial neural networks in cancer diagnosis, prognosis, and patient management. CRC Press, Boca Raton
37. Poggio T, Girosi F (1990) Regularization algorithms for learning that are equivalent to multi-layer networks. Science 247:978–982
38. Poggio T, Girosi F (1990) Networks for approximation and learning. Proc IEEE 78:1481–1497
39. Quian N, Sejnowski TJ (1988) Predicting the secondary structure of globular proteins using neural network models. J Mol Biol 202:865–884
40. Rost B (2001) Review: Protein secondary structure prediction continues to rise. J Struct Biol 134:204–218
41. Rosenblatt F (1958) The perceptron: A probabilistic model for information storage and organization in the brain. Psychol Rev 65(6):386–408
42. Riedmiller M, Braun H (1992) Rprop – a fast adaptive learning algorithm. In: Proceedings of the International Symposium on Computer and Information Science VII
43. Riedmiller M, Braun H (1993) A direct adaptive method for faster backpropagation learning: The Rprop algorithm. In: Proceedings of the IEEE International Conference on Neural Networks. IEEE Press, Piscataway, pp 586–591
44. Ripley BD, Ripley RM (2001) Neural Networks as statistical methods in survival analysis. In: Dybowski R, Gant V (eds) Clinical Applications of Artificial Neural Networks. Cambridge Univ Press, London
45. Robert C et al (2004) Bibliometric overview of the utilization of artificial neural networks in medicine and biology. Scientometrics 59:117–130
46. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by backpropagating errors. Nature 323:533–536
47. Schlökopf B, Smola AJ (2001) Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT Press, Cambridge
48. Spellman PT et al (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 9(12):3273–97
49. Taktak AFG, Fisher AC (2007) Outcome prediction in cancer. Elsevier Science, London
50. Tamayo P et al (1999) Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. PNAS 96(6):2907–2912
51. Tollenaere T (1990) SuperSAB: Fast adaptive backpropagation with good scaling properties. Neural Netw 3:561–573
52. Ward JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58(301):236–244
53. Werbos PJ (1974) Beyond regression: New tools for prediction and analysis in the behavioral science. Ph D Thesis, Harvard University
54. Werbos PJ (1994) The roots of backpropagation. Wiley, New York
55. Widrow B, Hoff ME (1960) Adaptive switching circuits. In: IRE WESCON Convention Record, Institute of Radio Engineers (now IEEE), vol 4. pp 96–104
56. Wu CH, McLarty JW (2000) Neural networks and genome informatics. Elsevier Science, Amsterdam

Books and Reviews

Abdi H (1994) A neural network primer. J Biol Syst 2:247–281
Bishop CM (2008) Pattern recognition and machine learning. Springer, Berlin
Fausett L (1994) Fundamentals of neural networks: Architectures, algorithms, and applications. Prentice Hall, New York

Freeman JA, Skapura DM (1991) Neural networks: Algorithms, applications, and programming techniques. Addison, Reading
Gurney K (1997) An Introduction to neural networks. Routledge, London
Hastie T, Tibshirani R, Friedman JH (2003) The elements of statistical learning. Springer, Berlin
Haykin S (1999) Neural networks: A comprehensive foundation. Prentice Hall, New York
Hertz J, Krogh A, Palmer R (1991) Introduction to the theory of neural computation. Addison, Redwood City
Kröse B, van der Smagt P (1996) An introduction to neural networks. University of Amsterdam, Amsterdam
Masters T (1993) Practical neural network recipes in C++. Academic Press, San Diego
Masters T (1995) Advanced algorithms for neural networks: A C++ sourcebook. Wiley, New York
Parks R, Levine D, Long D (1998) Fundamentals of neural network modeling. MIT Press, Cambridge
Patterson D (1996) Artif neural networks. Prentice Hall, New York
Peretto P (1992) An introduction to the modeling of neural networks. Cambridge University Press, London
Ripley BD (1996/2007) Pattern recognition and neural networks. Cambridge University Press, London
Smith M (1993) Neural networks for statistical modeling. Van Nostrand Reinhold, New York
Wasserman PD (1989) Neural Computing: Theory and Practice. Van Nostrand Reinhold, New York
Wasserman PD (1993) Advanced methods in neural computing. Van Nostrand Reinhold, New York
De Veaux RD, Ungar LH (1997) A brief introduction to neural networks. Technical Report, Williams College, University of Pennsylvania
Hinton GE (1992) How neural networks learn from experience. Sci Am 267:144–151
Lippman RP (1987) An introduction to computing neural networks. IEEE ASSP Mag 4(2):4–22
Reilly D, Cooper L (1990) An overview of neural networks: Early models to real world systems. Neural Electron Netw 2:229–250

Decision Trees

Vili Podgorelec, Milan Zorman
University of Maribor, Maribor, Slovenia

Article Outline

Glossary
Definition of the Subject
Introduction
The Basics of Decision Trees
Induction of Decision Trees
Evaluation of Quality
Applications and Available Software
Future Directions
Bibliography

Glossary

Accuracy The most important quality measure of an induced decision tree classifier. The most general is the overall accuracy, defined as a percentage of correctly classified instances from all instances (correctly classified and not correctly classified). The accuracy is usually measured both for the training set and the testing set.

Attribute A feature that describes an aspect of an object (both training and testing) used for a decision tree. An object is typically represented as a vector of attribute values. There are two types of attributes: continuous attributes whose domain is numerical, and discrete attributes whose domain is a set of predetermined values. There is one distinguished attribute called decision class (a dependent attribute). The remaining attributes (the independent attributes) are used to determine the value of the decision class.

Attribute node Also called a test node. It is an internal node in the decision tree model that is used to determine a branch from this node based on the value of the corresponding attribute of an object being classified.

Classification A process of mapping instances (i. e. training or testing objects) represented by attribute-value vectors to decision classes. If the predicted decision class of an object is equal to the actual decision class of the object, then the classification of the object is accurate. The aim of classification methods is to classify objects with the highest possible accuracy.

Classifier A model built upon the training set used for classification. The input to a classifier is an object (a vector of known values of the attributes) and the output of the classifier is the predicted decision class for this object.

Decision node A leaf in a decision tree model (also called a decision) containing one of the possible decision classes. It is used to determine the predicted decision class of an object being classified that arrives to the leaf on its path through the decision tree model.

Instance Also called an object (training and testing), represented by attribute-value vectors. Instances are used to describe the domain data.

Induction Inductive inference is the process of moving from concrete examples to general models, where the goal is to learn how to classify objects by analyzing a set of instances (already solved cases) whose classes are known. Instances are typically represented as attribute-value vectors. Learning input consists of a set of such vectors, each belonging to a known class, and the output consists of a mapping from attribute val-

ues to classes. This mapping should accurately classify both the given instances (a training set) and other unseen instances (a testing set).

Split selection A method used in the process of decision tree induction for selecting the most appropriate attribute and its splits in each attribute (test) node of the tree. The split selection is usually based on some impurity measures and is considered the most important aspect of decision tree learning.

Training object An object that is used for the induction of a decision tree. In a training object both the values of the attributes and the decision class are known. All the training objects together constitute a training set, which is a source of the "domain knowledge" that the decision tree will try to represent.

Testing object An object that is used for the evaluation of a decision tree. In a testing object the values of the attributes are known and the decision class is unknown for the decision tree. All the testing objects together constitute a testing set, which is used to test an induced decision tree – to evaluate its quality (regarding the classification accuracy).

Training set A prepared set of training objects.

Testing set A prepared set of testing objects.

Definition of the Subject

The term decision trees (abbreviated, DT) has been used for two different purposes: in decision analysis as a decision support tool for modeling decisions and their possible consequences to select the best course of action in situations where one faces uncertainty, and in machine learning or data mining as a predictive model; that is, a mapping from observations about an item to conclusions about its target value. This article concentrates on the machine learning view of DT.

More descriptive names for DT models are classification tree (discrete outcome) or regression tree (continuous outcome). In DT structures, leafs represent classifications and branches represent conjunctions of features that lead to those classifications. The machine learning technique for inducing a DT classifier from data (training objects) is called decision tree learning, or decision trees.

The main goal of classification (and regression) is to build a model that can be used for prediction [16]. In a classification problem, we are given a data set of training objects (a training set), each object having several attributes. There is one distinguished attribute called a decision class; it is a dependent attribute, whose value should be determined using the induced decision tree. The remaining attributes (the independent attributes, we will denote them just as attributes in further text) are used to determine the value of the decision class. Classification is thus a process of mapping instances (i. e. training or testing objects) represented by attribute-value vectors to decision classes.

The aim of DT learning is to induce such a DT model that is able to accurately predict the decision class of an object based on the values of its attributes. The classification of an object is accurate if the predicted decision class of the object is equal to the actual decision class of the object. The DT is induced using a training set (a set of objects where both the values of the attributes and decision class are known) and the resulting DT is used to determine decision classes for unseen objects (where the values of the attributes are known but the decision class is unknown). A good DT should accurately classify both the given instances (a training set) and other unseen instances (a testing set).

DT have a wide range of applications, but they are especially attractive in data mining [16]. Because of their intuitive representation, the resulting classification model is easy to understand, interpret and criticize [5]. DT can be constructed relatively fast compared to other methods [28]. And lastly, the accuracy of classification trees is comparable to other classification models [19,28]. Every major data mining tool includes some form of classification tree model construction component [17].

Introduction

In the early 1960s Feigenbaum and Simon presented EPAM (the Elementary Perceiver And Memorizer) [12] – a psychological theory of learning and memory implemented as a computer program. EPAM was the first system to use decision trees (then called discrimination nets). It memorized pairs of nonsense syllables by using discrimination nets in which it could find images of syllables it had seen. It stored as its cue only enough information to disambiguate the syllable from others seen at the time the association was formed. Thus, old cues might be inadequate to retrieve data later, so the system could "forget". Originally designed to simulate phenomena in verbal learning, it has been later adapted to account for data on the psychology of expertise and concept formation.

Later in the 1960s Hunt et al. built further on this concept and introduced CLS (Concept Learning System) [23] that used heuristic lookahead to construct trees. CLS was a learning algorithm that learned concepts and used them to classify new cases. CLS was the precursor to decision trees; it lead to Quinlan's ID3 system. ID3 [38] added the idea of using information content to choose the attribute

to split; it also initially chooses a window (a subset) of the training examples, and tries to learn a concept that will correctly classify all based on that window; if not, it increases window size. Quinlan later constructed C4.5 [41], an industrial version of ID3. Utgoff's ID5 [53] was an extension of ID3 that allowed many-valued classifications as well as incremental learning. From the early 1990s both the number of researchers and applications of DT have grown tremendously.

Objective and Scope of the Article

In this article an overview of DT is presented with the emphasis on a variety of different induction methods available today. Induction algorithms ranging from the traditional heuristic-based techniques to the most recent hybrids, such as evolutionary and neural network-based approaches, are described. Basic features, advantages and drawbacks of each method are presented. For the readers not very familiar with the field of DT this article should be a good introduction into this topic, whereas for more experienced readers it should broaden their perspective and deepen their knowledge.

The Basics of Decision Trees

Problem Definition

DT are a typical representative of a symbolic machine learning approach used for the classification of objects into decision classes, whereas an object is represented in a form of an attribute-value vector *(attribute*$_1$, *attribute*$_2$, … *attribute*$_N$, *decision class)*. The attribute values describe the features of an object. The attributes are usually identified and selected by the creators of the dataset. The decision class is one special attribute whose value is known for the objects in the learning set and which will be predicted based on the induced DT for all further unseen objects. Normally the decision class is a feature that could not be measured (for example some prediction for the future) or a feature whose measuring is unacceptably expensive, complex, or not known at all.

Examples of attributes-decision class objects are: patient's examination results and the diagnosis, raster of pixels and the recognized pattern, stock market data and business decision, past and present weather conditions and the weather forecast. The decision class is always a discrete-valued attribute and is represented by a set of possible values (in the case of regression trees a decision class is a continuous-valued attribute). If the decision should be made for a continuous-valued attribute, the values should be discretized first (by appropriately transforming continuous intervals into corresponding discrete values).

Formal Definition Let $A_1, \ldots, A_n$, C be random variables where A_i has domain $\mathrm{dom}(A_i)$ and C has domain $\mathrm{dom}(C)$; we assume without loss of generality that $\mathrm{dom}(C) = \{c_1, c_2, \ldots, c_j\}$. A DT classifier is a function

$$\mathrm{dt} : \mathrm{dom}(A_1) \times \cdots \times \mathrm{dom}(A_n) \mapsto \mathrm{dom}(C)$$

Let $P(A', C')$ be a probability distribution on $\mathrm{dom}(A_1) \times \cdots \times \mathrm{dom}(A_n) \times \mathrm{dom}(C)$ and let $t = \langle t.A_1, \ldots, t.A_n, t.C\rangle$ be a record randomly drawn from P; i. e., t has probability $P(A', C')$ that $\langle t.A_1, \ldots, t.A_n\rangle \in A'$ and $t.C \in C'$. We define the misclassification rate R_{dt} of classifier dt to be $P(\mathrm{dt}\,\langle t.A_1, \ldots, t.A_n\rangle \neq t.C)$. In terms of the informal introduction to this article, the training database D is a random sample from P, the A_i correspond to the attributes, and C is the decision class.

A DT is a directed, acyclic graph T in the form of a tree. Each node in a tree has either zero or more outgoing edges. If a node has no outgoing edges, then it is called a decision node (a leaf node); otherwise a node is called a test node (or an attribute node).

Each decision node N is labeled with one of the possible decision classes $c \in \{c_1, \ldots, c_j\}$. Each test node is labeled with one attribute $A_i \in \{A_1, \ldots, A_n\}$, called the splitting attribute. Each splitting attribute A_i has a splitting function f_i associated with it. The splitting function f_i determines the outgoing edge from the test node, based on the attribute value A_i of an object O in question. It is in a form of $A_i \in Y_i$ where $Y_i \subset \mathrm{dom}(A_i)$; if the value of the attribute A_i of the object O is within Y_i, then the corresponding outgoing edge from the test node is chosen.

The problem of DT construction is the following. Given a data set $D = \{t_1, \ldots, t_d\}$ where the t_i are independent random samples from an unknown probability distribution P, find a decision tree classifier T such that the misclassification rate $R_T(P)$ is minimal.

Inducing the Decision Trees

Inductive inference is the process of moving from concrete examples to general models, where the goal is to learn how to classify objects by analyzing a set of instances (already solved cases) whose decision classes are known. Instances are typically represented as attribute-value vectors. Learning input consists of a set of such vectors, each belonging to a known decision class, and the output consists of a mapping from attribute values to decision classes. This mapping should accurately classify both the given instances and other unseen instances.

A decision tree is a formalism for expressing such mappings [41] and consists of test nodes linked to two or more sub-trees and leafs or decision nodes labeled with a decision class. A test node computes some outcome based on the attribute values of an instance, where each possible outcome is associated with one of the sub-trees. An instance is classified by starting at the root node of the tree. If this node is a test, the outcome for the instance is determined and the process continues using the appropriate sub-tree. When a leaf is eventually encountered, its label gives the predicted decision class of the instance.

The finding of a solution with the help of DT starts by preparing a set of solved cases. The whole set is then divided into (1) a training set, which is used for the induction of a DT classifier, and (2) a testing set, which is used to check the accuracy of an obtained solution. First, all attributes defining each case are described (input data) and among them one attribute is selected that represents a decision class for the given problem (output data). For all input attributes specific value classes are defined. If an attribute can take only one of a few discrete values then each value takes its own class; if an attribute can take various numeric values then some characteristic intervals must be defined, which represent different classes. Each attribute can represent one internal node in a generated DT, also called a test node or an attribute node (Fig. 1). Such a test node has exactly as many outgoing edges as its number of different value classes. The leafs of a DT are decisions and represent the value classes of the decision attribute – decision classes (Fig. 1). When a decision has to be made for an unsolved case, we start with the root node of the DT classifier and moving along attribute nodes select outgoing edges where values of the appropriate attributes in the unsolved case matches the attribute values in the DT classifier until the leaf node is reached representing the decision class. An example training database is shown in Table 1 and a sample DT is shown in Fig. 1.

Decision Trees, Table 1
An example training set for object classification

#	Attributes			Decision class
	Color	Edge	Dot	Shape
1	green	dotted	no	triangle
2	green	dotted	yes	triangle
3	yellow	dotted	no	square
4	red	dotted	no	square
5	red	solid	no	square
6	green	solid	yes	triangle
7	green	solid	yes	square
8	yellow	dotted	no	triangle
9	yellow	solid	no	square
10	red	solid	no	square
11	green	solid	yes	square
12	yellow	dotted	yes	square
13	yellow	solid	no	square
14	red	dotted	yes	triangle

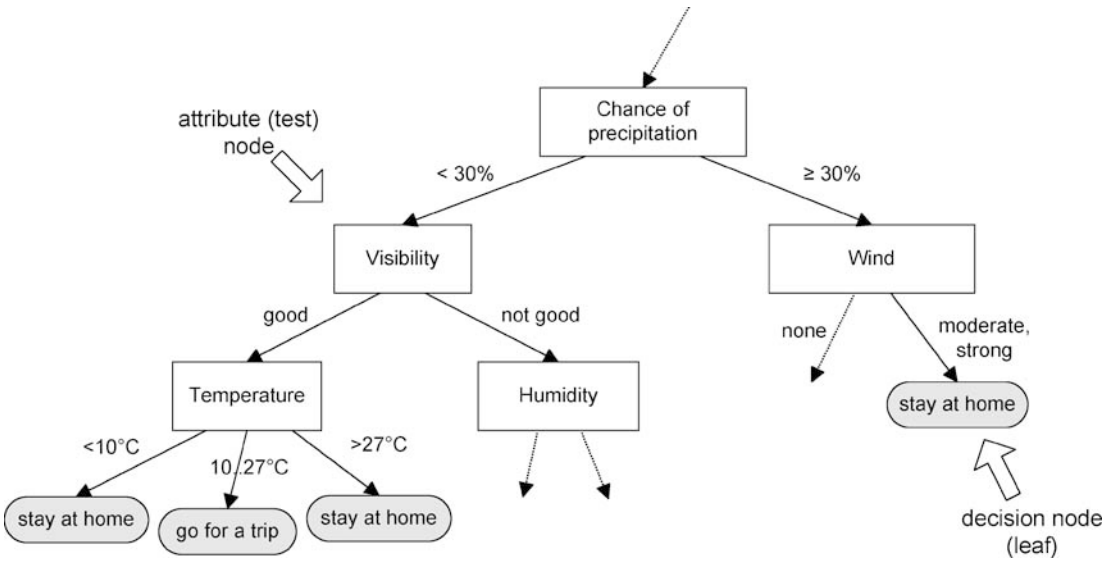

Decision Trees, Figure 1
An example of a (part of a) decision tree

The DT classifier is very easy to interpret. From the tree shown in Fig. 1 we can deduce for example the following rules:

- If the chance of precipitation is less than 30% and the visibility is good and the temperature is in the range of 10–27°C, then we should go for a trip,
- If the chance of precipitation is less than 30% and the visibility is good and the temperature is less than 10°C or more than 27°C, then we should stay at home,
- If the chance of precipitation is 30% or more and the wind is moderate or strong, then we should stay at home.

A DT can be built from a set of training objects with the "divide and conquer" principle. When all objects are of the same decision class (the value of the output attribute is the same) then a tree consists of a single node – a leaf with the appropriate decision. Otherwise an attribute is selected and a set of objects is divided according to the splitting function of the selected attribute. The selected attribute builds an attribute (test) node in a growing DT classifier, for each outgoing edge from that node the inducing procedure is repeated upon the remaining objects regarding the division until a leaf (a decision class) is encountered.

From a geometric point of view a set of n attributes defines a n-dimensional space, where each data object represents a point. A division of data objects regarding the attribute's class suits the definition of decision planes in the same space. Those planes are hyper-planes which are orthogonal to the selected attribute – DT divides a search space into hyper-rectangles, each of them represents one of the possible decision classes; of course more rectangles can also represent the same decision class.

Induction of Decision Trees

In 1986 Quinlan introduced an algorithm for inducing decision trees called ID3 [39,40]. In 1993 ID3 was upgraded with an improved algorithm C4.5 [41] that is still regarded as the reference model to build a DT based on the traditional statistical approach. Both algorithms ID3 and C4.5 use the statistical calculation of information gain from a single attribute to build a DT. In this manner an attribute that adds the most information about the decision upon a training set is selected first, followed by the next one that is the most informative from the remaining attributes, etc.

The method for constructing a DT as paraphrased from Quinlan [41], pp. 17–18, is as follows:

If there are j classes denoted $\{c_1, c_2, \ldots, c_j\}$, and a training set D, then

- *If D contains one or more objects which all belong to a single class c_i, then the decision tree is a leaf identifying class c_i*
- *If D contains no objects, the decision tree is a leaf determined from information other than D*
- *If D contains objects that belong to a mixture of classes, then a test is chosen, based on a single attribute, that has one or more mutually exclusive outcomes $\{o_1, o_2, \ldots, o_n\}$. D is partitioned into subsets $D_1, D_2, \ldots, D_n$, where D_i contains all the objects in D that have outcome o_i of the chosen test. The same method is applied recursively to each subset of training objects.*

Split Selection

The most important aspect of a traditional DT induction strategy is the way in which a set is split, i. e. how to select an attribute test that determines the distribution of training objects into sub-sets upon which sub-trees are built consequently. This process is called split selection. Its aim is to find an attribute and its associated splitting function for each test node in a DT. In the following text some of the most widely used split selection approaches will be presented.

Let D be the learning set of objects described with attributes $A_1, \ldots, A_n$ and a decision class C. Let n denote the number of training objects in D, n_i the number of training objects of decision class c_i, n_j the number of training objects with the jth value of splitting attribute, and n_{ij} the number of training objects of decision class c_i and jth value of the splitting attribute. The relative frequencies (probabilities) of training objects in D are as follows:

$p_{ij} = \frac{n_{ij}}{n}$ is the relative frequency of training objects of decision class c_i and jth value of splitting attribute within the training set D, $p_i = \frac{n_i}{n}$ is the relative frequency of training objects of decision class c_i within the training set D, $p_j = \frac{n_j}{n}$ is the relative frequency of training objects with the jth value of splitting attribute within the training set D, and $p_{i|j} = \frac{n_{ij}}{n_j}$ is the relative frequency of training objects of decision class c_i and jth value of splitting attribute regarding all the training objects with jth value of splitting attribute within the training set D.

Entropy, Information Gain, Information Gain Ratio
The majority of classical DT induction algorithms (such as ID3, C4.5, …) are based on calculating entropy from the information theory [45] to evaluate splits. In information theory entropy is used to measure the unreliability of a message as the source of information; the more infor-

mation a message contains, the lower is the entropy. Two splitting criteria are implemented:

- Gain criterion, and
- Gain ratio criterion.

The gain criterion [41] is developed in the following way:

For any training set D, n_i is the number of training objects of decision class c_i within the training set D. Then consider the "message" that a randomly selected object belongs to decision class c_i. The "message" has probability $p_i = \frac{n_i}{n}$, where n is the total number of training objects in D. The information conveyed by the message (in bits) is given by

$$I = -\log_2 p_i = -\log_2 \frac{n_i}{n} .$$

Entropy E of an attribute A of a training object w with the possible attribute values $a_1, \ldots, a_m$ and the probability distribution $p(A(w) = a_i)$ is thus defined as

$$E_A = -\sum_j p_j \cdot \log_2 p_j .$$

Let E_C be entropy of decision class distribution, E_A be entropy of values of an attribute A, and E_{CA} be entropy of combined decision class distribution and attribute values as follows:

$$E_C = -\sum_i p_i \cdot \log_2 p_i$$

$$E_A = -\sum_j p_j \cdot \log_2 p_j$$

$$E_{CA} = -\sum_i \sum_j p_{ij} \cdot \log_2 p_{ij} .$$

The expected entropy of decision class distribution regarding the attribute A is thus defined as

$$E_{C|A} = E_{CA} - E_A .$$

This expected entropy $E_{C|A}$ measures the reduction of entropy that is to be expected in the case when A is selected as the splitting attribute. The information gain is thus defined as

$$I_{\text{gain}}(A) = E_C - E_{C|A} .$$

In each test node an attribute A is selected with the highest value $I_{\text{gain}}(A)$. The gain criterion [41] selects a test to maximize this information gain.

The gain criterion has one significant disadvantage in that it is biased towards tests with many outcomes. The gain ratio criterion [41] was developed to avoid this bias; it is defined as

$$I_{\text{gain ratio}}(A) = \frac{I_{\text{gain}}(A)}{E_A} .$$

If the split is near trivial, split information will be small and this ratio will be unstable. Hence, the gain ratio criterion selects a test to maximize the gain ratio subject to the constraint that the information gain is large.

Gini Index The gain ratio criterion compares with CART's impurity function approach [5], where impurity is a measure of the class mix of a subset and splits are chosen so that the decrease in impurity is maximized. This approach led to the development of the Gini index [5].

The impurity function approach considers the probability of misclassifying a new sample from the overall population, given that the sample was not part of the training sample, T. This probability is called the misclassification rate and is estimated using either the resubstitution estimate (or training set accuracy), or the test sample estimate (test set accuracy). The node assignment rule selects i to minimize this misclassification rate. In addition, the Gini index promotes splits that minimize the overall size of the tree.

$$\text{gini}(A) = -\sum_j p_j \sum_i p_{i|j}^2 - \sum_i p_i^2 .$$

Chi-Square Various types of chi-square tests are often used in statistics for testing the significance. The Chi-square (chi-square goodness-of-fit) test compares the expected frequency e_{ij} with the actual frequency n_{ij} of training objects, which belong to decision class c_i and having jth value of the given attribute [55]. It is defined as

$$\chi^2(A) = \sum_i \sum_j \frac{(e_{ij} - n_{ij})^2}{e_{ij}}$$

where

$$e_{ij} = \frac{n_j n_i}{n} .$$

The higher value of χ^2 means a clearer split. Consequentially, the attribute with the highest χ^2 value is selected.

J-Measure The J-measure heuristics has been introduced by Smyth and Goodman [48] as an information-theoretical model of measuring the information content of a rule.

The crossover entropy J_j is appropriate for selecting a single value of a given attribute A for constructing a rule; it is defined as

$$J_j(A) = p_j \sum_i p_{i|j} \log \frac{p_{i|j}}{p_i} .$$

The generalization over all possible values of a given attribute A measures the purity of the attribute

$$J(A) = \sum_j J_j(A) .$$

The J-measure is also used to reduce over-fitting in the process of pre-pruning (see the following sections).

DT Induction Example: Classifying Geometrical Objects

Let us have a number of geometrical objects: squares and triangles. Each geometrical object is described with three features: color describes the color of an object (it can be either green, yellow or red), edge describes the line type of an object's edge (it can be either solid or dotted), and dot describes whether there is a dot in an object. See Table 1 for details.

We want to induce a DT that will predict the shape of an unseen object based on the three features. The decision class will thus be **shape**, having two possible values: *square* or *triangle*. The three features represent the tree attributes: **color** (with possible values *green*, *yellow* and *red*), **edge** (with possible values *solid* and *dotted*), and **dot** (with possible values *yes* and *no*). Table 1 resembles the training set with 14 training objects; in Fig. 2 the training objects are also visually represented.

In the training set (Table 1, Fig. 2) there are 14 training objects: five triangles and nine squares. We can calculate the class distributions:

$$p(\text{square}) = \frac{9}{14} \text{ and } p(\text{triangle}) = \frac{5}{14} .$$

The entropy of decision class (shape) distribution is thus

$$E_{\text{shape}} = -\sum_i p_i \cdot \log_2 p_i$$
$$= -\frac{9}{14} \cdot \log_2 \frac{9}{14} - \frac{5}{14} \cdot \log_2 \frac{5}{14} = 0.940 .$$

The entropy can be reduced by splitting the whole training set. For this purpose either of the attributes (color, edge, or dot) can be used as a splitting attribute. Regarding the information gain split selection approach, the entropy of each attribute is calculated first, then the information gain of each attribute is calculated based on these entropies, and finally the attribute is selected that has the highest information gain.

To demonstrate it, let us calculate the entropy for the attribute color. Using different possible values of color (green, yellow, and red) the whole training set is split into three subsets (Fig. 3). The entropies are as follows:

$$E_{\text{color=green}} = -\frac{2}{5} \cdot \log_2 \frac{2}{5} - \frac{3}{5} \cdot \log_2 \frac{3}{5} = 0.971$$

$$E_{\text{color=yellow}} = -\frac{4}{4} \cdot \log_2 \frac{4}{4} - \frac{0}{4} \cdot \log_2 \frac{0}{4} = 0.0$$

$$E_{\text{color=red}} = -\frac{3}{5} \cdot \log_2 \frac{3}{5} - \frac{2}{5} \cdot \log_2 \frac{2}{5} = 0.971 .$$

Consequentially, the entropy of the attribute color based on the whole training set is

$$E_{\text{color}} = \sum_i p_i E_i = \frac{5}{14} \cdot 0.971 + \frac{4}{14} \cdot 0.0 + \frac{5}{14} \cdot 0.971$$
$$= 0.694 .$$

The information gain of the attribute color is thus

$$I_{\text{gain}}(\text{color}) = E_{\text{shape}} - E_{\text{shape|color}}$$
$$= 0.940 - 0.694 = 0.246 .$$

Similarly, the information gain can be calculated also for the remaining two attributes, edge and dot. The information gain for all three attributes are thus

$$I_{\text{gain}}(\text{color}) = 0.246$$

$$I_{\text{gain}}(\text{edge}) = 0.151$$

$$I_{\text{gain}}(\text{dot}) = 0.048 .$$

As we can see the attribute color has the highest information gain and is therefore chosen as the splitting attribute

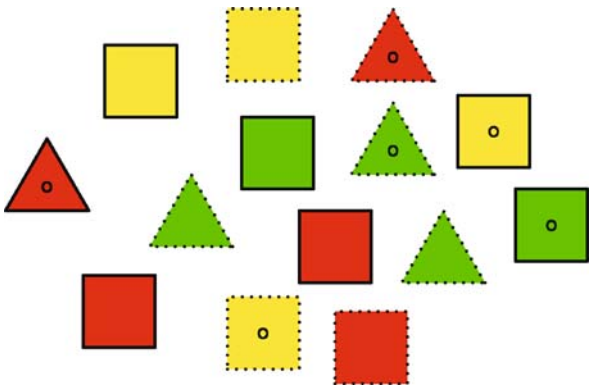

Decision Trees, Figure 2
Visual representation of training objects from Table 1

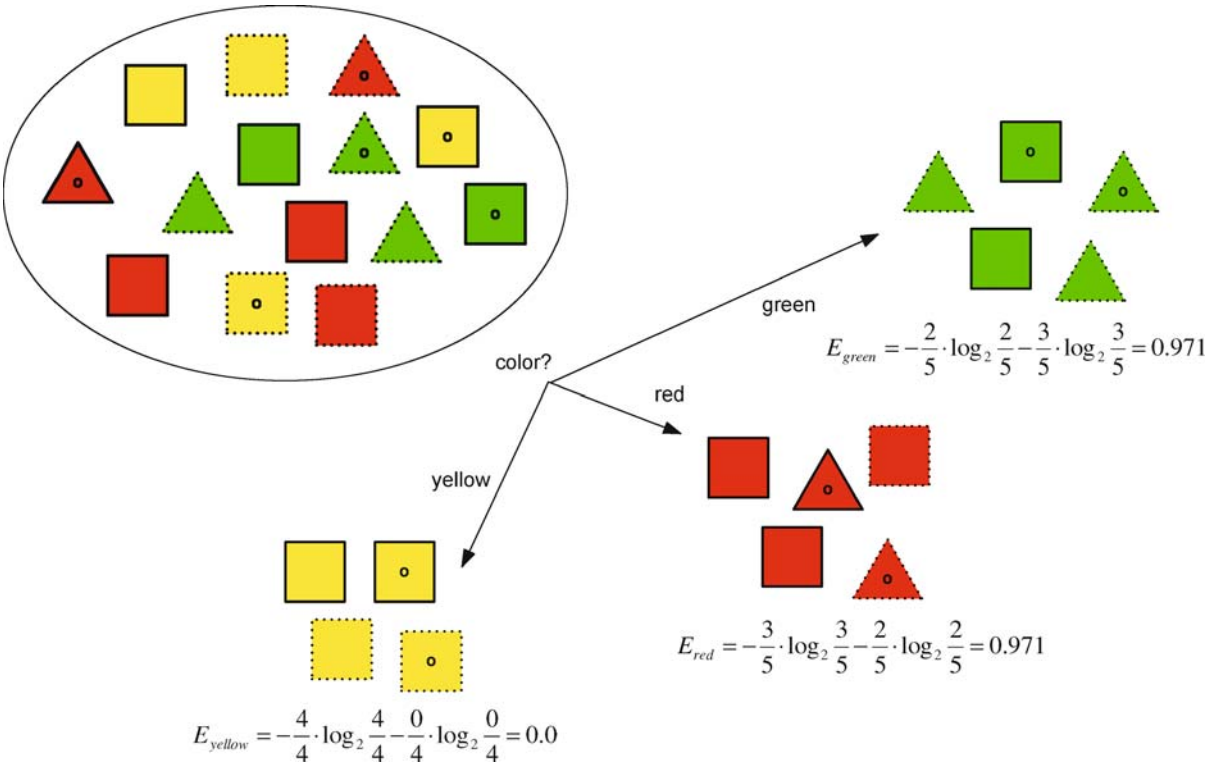

Decision Trees, Figure 3
Splitting the data set based on the attribute color

at the root node of the DT classifier. The whole process of splitting is recursively repeated on all subsets. The resulting DT classifier is shown in Fig. 4.

Tests on Continuous Attributes

In the above example all attributes were discrete-valued. For the continuous attributes a split threshold needs to be determined whenever the attribute is to be used as a splitting attribute. The algorithm for finding appropriate thresholds for continuous attributes [5,33,41] is as follows:

The training objects are sorted on the values of the attribute. Denote them in order as $\{w_1, w_2, \ldots, w_k\}$. *Any threshold value lying between* w_i *and* w_{i+1} *will have the same effect, so there are only k-1 possible splits, all of which are examined.*

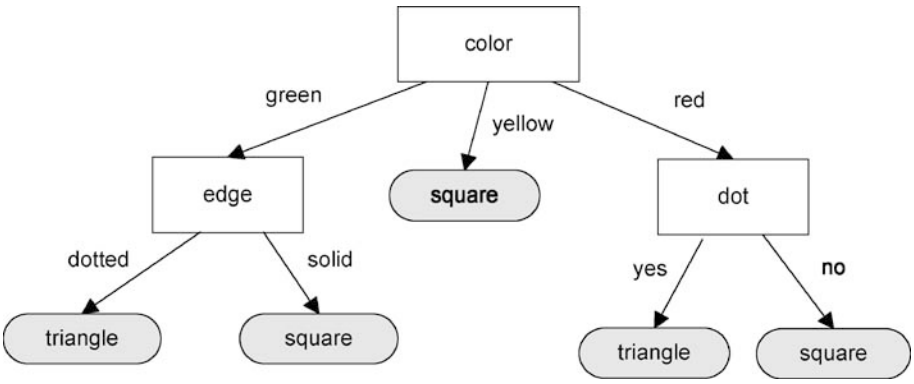

Decision Trees, Figure 4
The resulting DT classifier for the classification of geometrical objects

Some approaches for building DT use discretization of continuous attributes. Two ways are used to select discretized classes:

- equidistant intervals; The Number Of Classes Is Selected First And Then Successive Equidistant Intervals Are Determined Between Absolute Lower And Upper Bounds, And
- percentiles; Again The Number Of Classes Is Selected First And Then Successive Intervals Are Determined Based On The Values Of The Appropriate Attribute In The Training Set So That All Intervals Contain The Same Number Of Training Objects.

Dynamic Discretization of Attributes In the MtDeciT 2.0 tool authors implemented an algorithm for finding subintervals [58], where the distribution of training objects is considered and there are more than two subintervals possible. The approach is called dynamic discretization of continuous attributes, since the subintervals are determined dynamically during the process of building a DT. This technique first splits the interval into many subintervals, so that every training object's value has its own subinterval. In the second step it merges together smaller subintervals that are labeled with the same outcome into larger subintervals. In each of the following steps three subintervals are merged together: two "stronger" subintervals with one "weak" interval, where the "weak" interval lies between those two "strong" subintervals. Here strong and weak applies to the number of training objects in the subinterval tree (Fig. 5). In comparison to the previous two approaches the dynamic discretization returns more natural subintervals, which result in better and smaller DT classifiers.

In general we differentiate between two types of dynamic discretization:

- General dynamic discretization, and
- Nodal dynamic discretization.

General dynamic discretization uses all available training objects for the definition of subintervals. That is why the general dynamic discretization is performed before the start of building a DT. All the subintervals of all attributes are memorized in order to be used later in the process of building the DT.

Nodal dynamic discretization performs the definition of subintervals for all continuous attributes which are available in the current node of a DT. Only those training objects that came in the current node are used for setting the subintervals of the continuous attributes.

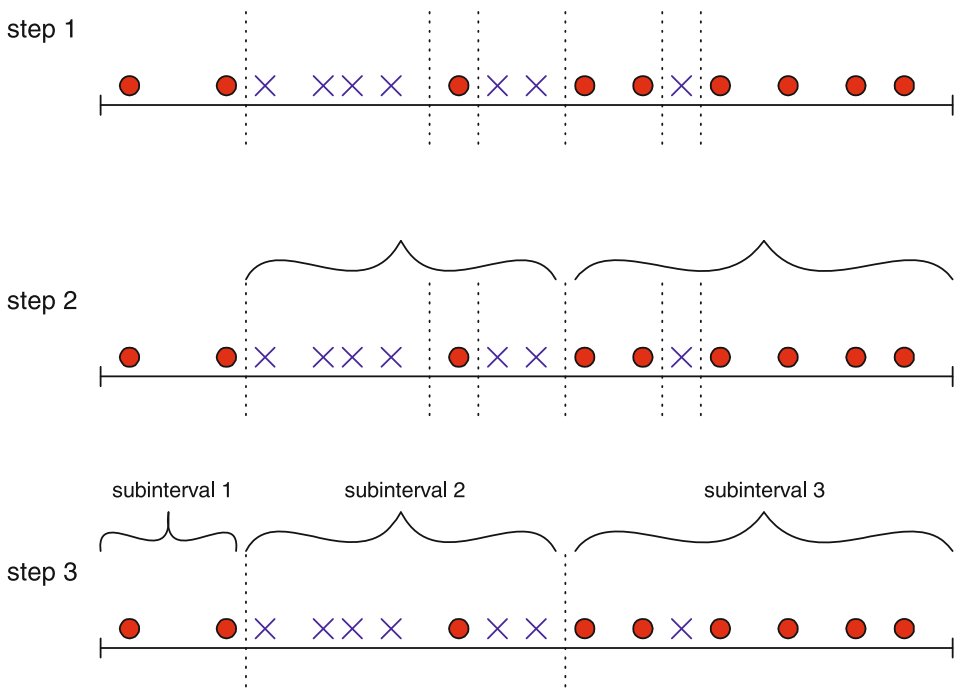

Decision Trees, Figure 5
Dynamic discretization of continuous attribute, which has values between 60 and 100. Shapes represent different attribute's values

In a series of tests the authors showed that nodal dynamic discretization produces smaller DT with higher accuracy than DT built with general dynamic discretization [56,58]. Nodal dynamic discretization also outperforms classical discretization techniques in the majority of cases.

Oblique Partitioning of Search Space

The so-far presented algorithms are using univariate partitioning methods, which are attractive because they are straightforward to implement (since only one feature is analyzed at a time) and the derived DT is relatively easy to understand. Beside univariate partitioning methods there are also some successful partitioning methods that do not partition the search space axis-parallel based on only one attribute at a time but are forming oblique partition boundaries based on a combination of attributes (Fig. 6).

Oblique partitioning provides a viable alternative to univariate methods. Unlike their univariate counterparts, oblique partitions are formed by combinations of attributes. The general form of an oblique partition is given by

$$\sum_{i=1}^{d} \beta_i x_i \leq C$$

where β_i represents the coefficient of the ith attribute. Because of their multivariate nature, oblique methods offer far more flexibility in partitioning the search space; this flexibility comes at a price of higher complexity, however. Consider that given a data set containing n objects described with d attributes, there can be $2 \cdot \sum_{i=0}^{d} \binom{n-1}{i}$ oblique splits if $n > d$ [50]; each split is a hyper-plane that divides the search space into two non-overlapping halves. For univariate splits, the number of potential partitions is much lower, but still significant, $n \cdot d$ [29]. In short, finding the right oblique partition is a difficult task.

Given the size of the search space, choosing the right search method is of critical importance in finding good partitions. Perhaps the most comprehensive reference on this subject is [5] on classification and regression trees (CART). Globally, CART uses the same basic algorithm as Quinlan in C4.5. At the decision node level, however, the algorithm becomes extremely complex. CART starts out with the best univariate split. It then iteratively searches for perturbations in attribute values (one attribute at a time) which maximize some goodness metric. At the end of the procedure, the best oblique and axis-parallel splits found are compared and the better of these is selected.

Although CART provides a powerful and efficient solution to a very difficult problem, it is not without its disadvantages. Because the algorithm is fully deterministic, it has no inherent mechanism for escaping from local optima. As a result, CART has a tendency to terminate its partition search at a given node too early. The most fundamental disadvantage of CART (and of the traditional approach of inducing DT in general) is that the DT induction process can cause the metrics to produce misleading results. Because traditional DT induction algorithms choose what is locally optimal for each decision node, they inevitably ignore splits that score poorly alone, but yield better solution when used in combination. This problem is illustrated by Fig. 7. The solid lines indicate the splits

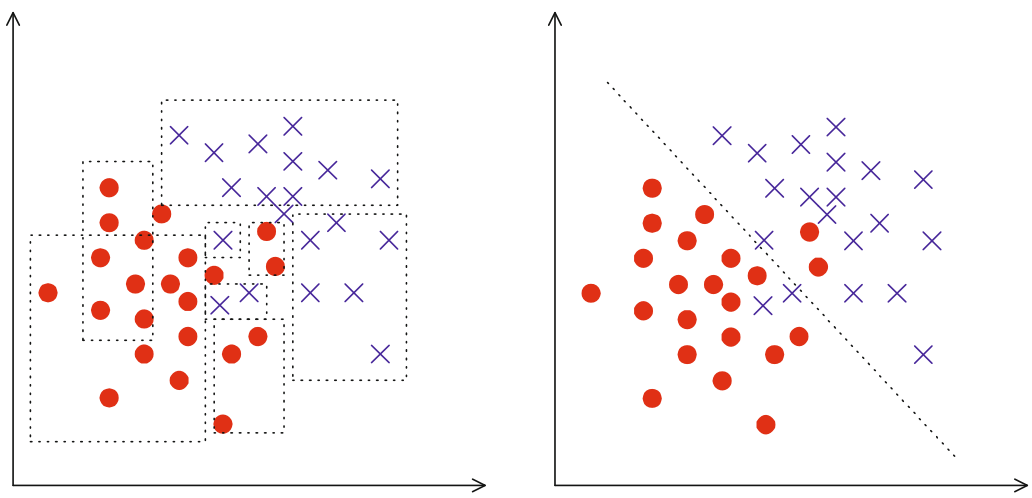

Decision Trees, Figure 6
Given axes that show the attribute values and shape corresponding to class labels: i axis-parallel and ii oblique decision boundaries

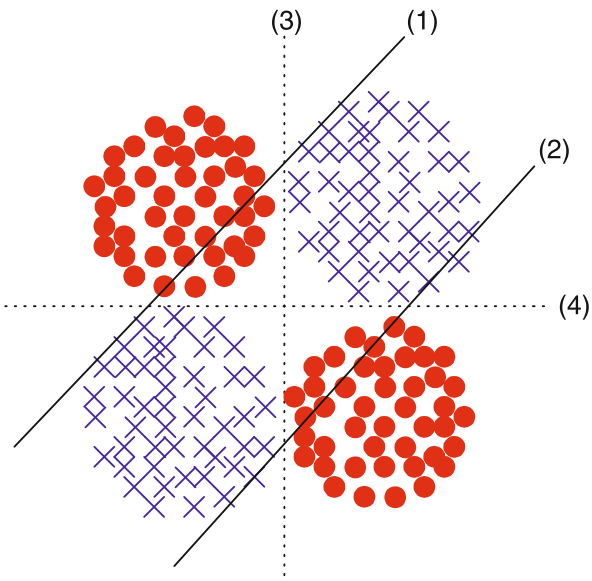

Decision Trees, Figure 7
CART generated splits (*solid lines* – 1,2) minimize impurity at each decision node in a way that is not necessarily optimal regarding the natural structure of data (denoted by *dotted lines* – 3,4)

found by CART. Although each split optimizes the impurity metric, the end product clearly does not reflect the best possible partitions (indicated by the dotted lines). However, when evaluated as individuals, the dotted lines register high impurities and are therefore not chosen. Given this, it is apparent that the sequential nature of DT can prevent the induction of trees that reflect the natural structure of the data.

Pruning Decision Trees

Most of the real-world data contains at least some amount of noise and outliers. Different machine-learning approaches elaborate different levels of sensitivity to noise in data. DT fall into the category of methods, sensitive to noise and outliers. In order to cure that, some additional methods must be applied to DT in order to reduce the complexity of the DT and increase its accuracy. In DTs we call this approach pruning.

Though pruning can be applied in different stages, it follows the basic idea, also called the Ockham's razor. William of Ockham (1285–1349), one of the most influential medieval philosophers has been given credit for the rule, which said that "one should not draw more conclusions about a certain matter than minimum necessary and that the redundant conclusions should be removed or shaved off". That rule was a basis for latter interpretation in the form of the following statement: "If two different solutions solve the same problem with the same accuracy, then the better of the two is the shorter solution."

Mapping this rule to DT pruning allows us to prune or shave away those branches in the DT, which do not decrease the classification accuracy. By doing so, we end up with a DT, which is not only smaller, but also has higher (or in the worse case scenario at least the same) accuracy as the original DT.

DT classifiers aim to refine the training sample T into subsets which have only a single class. However, training samples may not be representative of the population they are intended to represent. In most cases, fitting a DT until all leafs contain data for a single class causes over-fitting. That is, the DT is designed to classify the training sample rather than the overall population and accuracy on the overall population will be much lower than the accuracy on the training sample.

For this purpose most of the DT induction algorithms (C4.5, CART, OC1) use pruning. They all grow trees to maximum size, where each leaf contains single-class data or no test offers any improvement on the mix of classes at that leaf, and then prune the tree to avoid over-fitting. Pruning occurs within C4.5 when the predicted error rate is reduced by replacing a branch with a leaf. CART and OC1 use a proportion of the training sample to prune the tree. The tree is trained on the remainder of the training sample and then pruned until the accuracy on the pruning sample can not be further improved.

In general we differentiate between two pruning approaches: prepruning, which takes place during DT induction and postpruning, applied to already-induced DT in order to reduce complexity.

We will describe three representatives of pruning – one prepruning and two postpruning examples.

Prepruning Prepruning (called also stopping criteria) is a very simple procedure, performed during the first phase of induction. Early stopping of DT construction is based on criteria, which measures the percentage of the dominant class in the training subset in the node. If that percentage is higher than the preset threshold, then the internal node is transformed to a leaf and marked with the dominant class label. Such early stopping of DT construction reduces the size and complexity of DT, reduces the possibility of over-fitting, making it more general. However, there is also danger of oversimplification when the preset threshold is too low. This problem is especially present in training sets where frequencies of class labels are not balanced. An improper threshold can cause some special training objects, which are very important for solv-

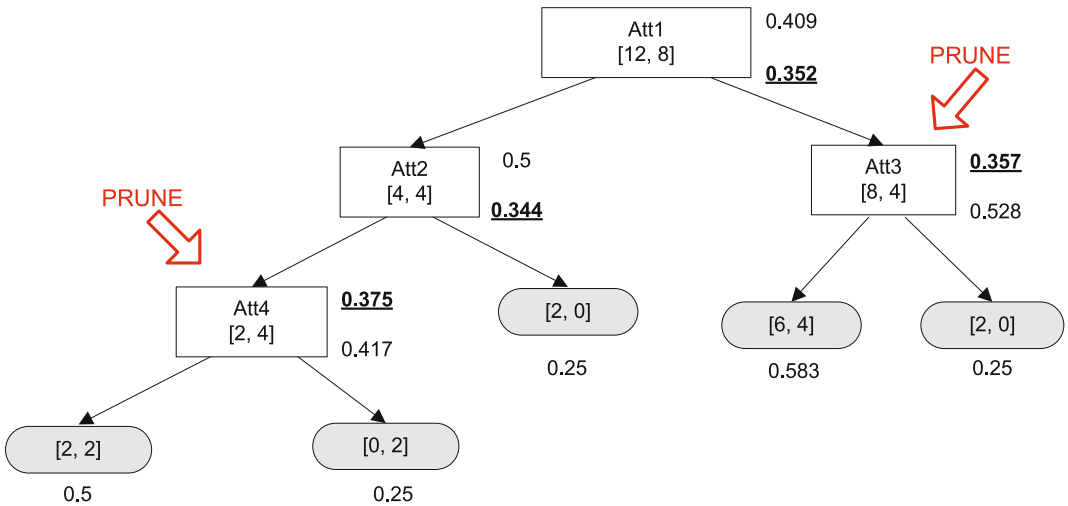

Decision Trees, Figure 8
Unpruned decision tree with static error estimates

ing the classification problem, to be discarded during the prepruning process.

Postpruning The opposite to prepruning, postpruning operates on DT that is already constructed. Using simple frequencies or error estimation, postpruning approaches calculate whether or not substituting a subtree with a leaf (pruning or shaving-off according to Ockham's razor) would increase the DT's accuracy or at least reduce its size without negatively effecting classification accuracy.

Both the following approaches are practically the same up to a point, where we estimate error in a node or subtree.

Static Error Estimate Pruning Static error estimate pruning uses frequencies of training objects' class labels in nodes to estimate error, caused by replacing an internal node with a leaf. If the error estimate in the node is lower or equal to the error computed in the subtree, then we can prune the subtree in question.

Let C be a set of training objects in node V, and m the number of all possible class labels. N is the number of training objects in C and O is the most frequent class label in C. Let n be the number of training objects in C that belong to class O. The static error estimate (also called Laplace error estimate) in node V is $E(V)$:

$$E(V) = \frac{N - n + m - 1}{N + m} .$$

Let us compute the subtree error of V using the next expression

$$\text{SubtreeError}(V) = \sum_{i=1}^{\text{NoOfChildNodes}} P_i * \text{Err}(V_i) .$$

Since we do not have information about class distribution, we use relative frequencies in the place of probabilities P_i. $\text{Err}(V_i)$ is the error estimate of child node V_i.

Now we can compute the new error estimate $\text{Err}(V)$ in node V by using the next expression:

$$\text{Err}(V) = \min\left(E(V), \text{SubtreeError}(V)\right) .$$

If $E(V)$ is less than or equal to $\text{SubtreeError}(V)$, then we can prune the subtree of V, replace it by a leaf, label it class O and assign the value of $E(V)$ to $\text{Err}(V)$. In the opposite case we only assign the value of $\text{SubtreeError}(V)$ to $\text{Err}(V)$.

Static error estimate pruning is a recursive bottom-up approach, where we start to calculate error estimates in leafs and perform error propagation and eventual pruning actions toward parent nodes until the root node is reached.

In Fig. 8 we have an unpruned DT with two possible decision classes, O_1 and O_2. Inside each node there are frequencies of training objects according to their class label, written in brackets (first for O_1 and second for O_2). Beside each internal node (squares) there are two numbers. The top one is static error estimate in the node and the bottom one is error estimate in the subtree. The number below each leaf (ellipses) is its error estimate.

We can see that static error estimates are lower than subtree errors for two nodes: Att3 and Att4. Therefore, we prune those two subtrees and we get a smaller but more accurate DT (Fig. 9).

Reduced Error Pruning Reduced error pruning implements a very similar principle as static error estimate pruning. The difference is in error estimation. Reduced error

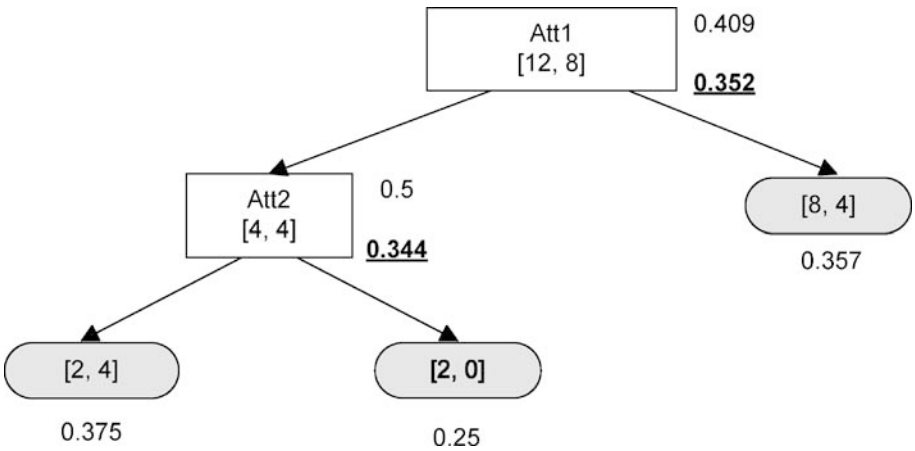

Decision Trees, Figure 9
Pruned decision tree with static error estimates

pruning uses the pruning set to estimate errors in nodes and leafs. The pruning set is very similar to training set, since it contains objects, described with attributes and class label. In the first stage we try to classify each pruning object. When a leaf is reached, we compare DT's classification with actual class label and mark correct/incorrect decision in the leaf. At the end there is a number of errors and number of correct classifications recorded in each leaf.

In a recursive bottom-up approach, we start to calculate error estimates in leafs, perform error propagation and eventual pruning actions toward parent nodes until the root node is reached.

Let SubtreeError(V) be a sum of errors from all subtrees of node V.

$$\text{SubtreeError}(V) = \begin{cases} V \in \text{Leafs} \Rightarrow \text{NoOfErrorsInLeaf} \\ V \notin \text{Leafs} \Rightarrow \sum_{i=1}^{\text{NoOfChildNodes}} \text{Err}(V_i) \,. \end{cases}$$

Let C be a set of training objects in node V and O the most frequent class label in C. Let N be the number of all pruning objects that reached the node V and n be the number of pruning objects in V that belong to class O. The error estimate in node V is defined by $E(V)$:

$$E(V) = N - n \,.$$

Now we can compare the error estimate in the node with the error estimate in the subtree and compute a new error estimate Err(V) in node V:

$$\text{Err}(V) = \min\big(E(V),\ \text{SubtreeError}(V)\big) \,.$$

If $E(V)$ is less than or equal to SubtreeError(V), then we can prune the subtree of V, replace it by a leaf labeled with class O and assign the value of $E(V)$ to Err(V). In the opposite case we only assign the value of SubtreeError(V) to Err(V).

The drawback of this pruning approach is the fact that it is very hard to construct a pruning set that is capable of "visiting" each node in a DT at least once. If we cannot assure that, the pruning procedure will not be as efficient as other similar approaches.

Deficiencies of Classical Induction

The DT has been shown to be a powerful tool for decision support in different areas. The effectiveness and accuracy of classification of DTs have been a surprise for many experts and their greatest advantage is in simultaneous suggestion of a decision and the straightforward and intuitive explanation of how the decision was made. Nevertheless, the classical induction approach also contains several deficiencies.

One of the most obvious drawbacks of classical DT induction algorithms is poor processing of incomplete, noisy data. If some attribute value is missing, classical algorithms do not perform well on processing of such an object. For example, in Quinlan's algorithms before C4.5 such data objects with missing data were left out of the training set – this fact of course resulted in decreased quality of obtained solutions (in this way the training set size and the information about the problem's domain were reduced). Algorithm C4.5 introduced a technique to overcome this problem, but this is still not very effective. In real world problems, especially in medicine, missing data is very common – therefore, effective processing of such data is of vital importance.

The next important drawback of classical induction methods is the fact that they can produce only one DT classifier for a problem (when the same training set is used). In many real-world situations it would be of great

benefit if more DT classifiers would be available and a user could choose the most appropriate one for a single case. As it is possible for training objects to miss some attribute value, the same goes also for a testing object – there can be a new case where some data is missing and it is not possible to obtain it (unavailability of some medical equipment, for example, or invasiveness of a specific test for a patient). In this way another DT classifier could be chosen that does not include a specific attribute test to make a decision.

Let us mention only one more disadvantage of classical induction methods, namely the importance of different errors. Between all decisions possible there are usually some that are more important than the others. Therefore, a goal is to build a DT in such a way that the accuracy of classification for those most important decisions is maximized. Once again, this problem is not solved adequately in classical DT induction methods.

Variant Methods

Although algorithms such as ID3, C4.5, and CART make up the foundation of traditional DT induction practice, there is always room for improvement of the accuracy, size, and generalization ability of the generated trees. As would be expected, many researchers have tried to build on the success of these techniques by developing better variations of them.

Alternatively, a lot of different methods for the induction of DT have been introduced by various researchers, which try to overcome the deficiencies noted above. Most of them are based on so-called soft methods, like evolutionary techniques or neural networks, sometimes several methods are combined in a hybrid algorithm.

A vast number of techniques have also been developed that help to improve only a part of the DT induction process. Such techniques include evolutionary algorithms for optimizing split functions in attribute nodes, dynamic discretization of attributes, dynamic subset selection of training objects, etc.

Split Selection Using Random Search Since random search techniques have proven extremely useful in finding solutions to non-deterministic polynomial complete (NP-complete) problems [30], naturally they have been applied to DT induction. Heath [20,21] developed a DT induction algorithm called SADT that uses a simulated annealing process to find oblique splits at each decision node. Simulated annealing is a variation of hill climbing which, at the beginning of the process, allows some random downhill moves to be made [42]. As a computational process, simulated annealing is patterned after the physical process of annealing [25], in which metals are melted (at high temperatures) and then gradually cool until some solid state is reached.

Starting with an initial hyper-plane, SADT randomly perturbs the current solution and determines the goodness of the split by measuring the change in impurity (ΔE). If ΔE is negative (i. e. impurity decreases), the new hyper-plane becomes the current solution; otherwise, the new hyper-plane becomes the current split with probability $e^{-(\Delta E/T)}$ where T is the temperature of the system. Because simulated annealing mimics the cooling of metal, its initially high temperature falls with each perturbation. At the start of the process, the probability of replacing the current hyper-plane is nearly 1. As the temperature cools, it becomes increasingly unlikely that worse solutions are accepted. When processing a given data set, Heath typically grows hundreds of trees, performing several thousands perturbations per decision node. Thus, while SADT has been shown to find smaller trees than CART, it is very expensive from a computational standpoint [21].

A more elegant variation on Heath's approach is the OC1 system [29]. Like SADT, OC1 uses random search to find the best split at each decision node. The key difference is that OC1 rejects the brute force approach of SADT, using random search only to improve on an existing solution. In particular, it first finds a good split using a CART-like deterministic search routine. OC1 then randomly perturbs this hyper-plane in order to decrease its impurity. This step is a way of escaping the local optima in which deterministic search techniques can be trapped. If the perturbation results in a better split, OC1 resumes the deterministic search on the new hyper-plane; if not, it re-perturbs the partition a user-selectable number of times. When the current solution can be improved no further, it is stored for later reference. This procedure is repeated a fixed number of times (using a different initial hyper-plane in each trial). When all trials have been completed, the best split found is incorporated into the decision node.

Incremental Decision Tree Induction The DT induction algorithms discussed so far grow trees from a complete training set. For serial learning tasks, however, training instances may arrive in a stream over a given time period. In these situations, it may be necessary to continually update the tree in response to the newly acquired data. Rather than building a new DT classifier from scratch, the incremental DT induction approach revises the existing tree to be consistent with each new training instance. Utgoff implemented an incremental version of ID3, called

ID5R [53]. ID5R uses an *E-Score* criteria to estimate the amount of ambiguity in classifying instances that would result from placing a given attribute as a test in a decision node. Whenever the addition of new training instances does not fit the existing tree, the tree is recursively restructured such that attributes with the lowest E-Scores are moved higher in the tree hierarchy. In general, Utgoff's algorithm yields smaller trees compared to methods like ID3, which batch process all training data. Techniques similar to ID5R include an incremental version of CART [8]. Incremental DT induction techniques result in frequent tree restructuring when the amount of training data is small, with the tree structure maturing as the data pool becomes larger.

Decision Forests Regardless of the DT induction method utilized, subtle differences in the composition of the training set can produce significant variances in classification accuracy. This problem is especially acute when cross-validating small data sets with high dimensionality [11]. Researchers have reduced these high levels of variance by using decision forests, composed of multiple trees (rather than just one). Each tree in a forest is unique because it is grown from a different subset of the same data set. For example, Quinlan's windowing technique [41] induces multiple trees, each from a randomly selected subset of the training data (i. e., a window). Another approach was devised by Ho [22], who based each tree on a unique feature subset. Once a forest exists, the results from each tree must be combined to classify a given data instance. Such committee-type schemes for accomplishing this range from using majority rules voting [21,37] to different statistical methods [46].

A Combination of Decision Trees and Neural Networks When DT and neural networks are compared, one can see that their advantages and drawbacks are almost complementary. For instance knowledge representation of DT is easily understood by humans, which is not the case for neural networks; DT have trouble dealing with noise in training data, which is again not the case for neural networks; DT learn very fast and neural networks learn relatively slow, etc. Therefore, the idea is to combine DT and neural networks in order to combine their advantages. In this manner, different hybrid approaches from both fields have been introduced [7,54,56].

Zorman in his MtDecit 2.0 approach first builds a DT that is then used to initialize a neural network [56,57]. Such a network is then trained using the same training objects as the DT. After that the neural network is again converted to a DT that is better than the original DT [2]. The source DT classifier is converted to a disjunctive normal form – a set of normalized rules. Then the disjunctive normal form serves as the source for determining the neural network's topology and weights. The neural network has two hidden layers, the number of neutrons on each hidden layer depends on rules in the disjunctive normal form. The number of neurons in the output layer depends on how many outcomes are possible in the training set. After the transformation, the neural networks is trained using back-propagation. The mean square error of such a network converges toward 0 much faster than it would in the case of randomly set weights in the network. Finally, the trained neural network has to be converted into a final DT. The neural network is examined in order to determine the most important attributes that influence the outcomes of the neural network. A list containing the most important attributes is then used to build the final DT that in the majority of cases performs better than the source DT. The last conversion usually causes a loss of some knowledge contained in the neural network, but even so most knowledge is transformed into the final DT. If the approach is successful, then the final DT has better classification capabilities than the source DT.

Using Evolutionary Algorithms to Build Decision Trees Evolutionary algorithms are generally used for very complex optimization tasks [18], for which no efficient heuristic method exist. Construction of DT is a complex task, but heuristic methods exist that usually work efficiently and reliably. Nevertheless, there are some reasons justifying an evolutionary approach. Because of the robustness of evolutionary techniques they can be successfully used also on incomplete, noisy data (which often happens in real life data because of measurement errors, unavailability of proper instruments, risk to patients, etc.). Because of evolutionary principles used to evolve solutions, solutions can be found which can be easily overlooked otherwise. Also the possibility of optimizing the DT classifier's topology and the adaptation of split thresholds is an advantage.

There have been several attempts to build a DT with the use of evolutionary techniques [6,31,35], one the most recent and also very successful in various applications is genTrees, developed by Podgorelec and Kokol [35,36,37]. First an initial population of (about one hundred) semi-random DT is seeded. A random DT classifier is built by randomly choosing attributes and defining a split function. When a pre-selected number of test nodes is placed into a growing tree then a tree is finalized with decision nodes, which are defined by choosing the most fit decision class in each node regarding the training set. After the initialization the population of decision trees evolves through

many generations of selection, crossover and mutation in order to optimize the fitness function that determines the quality of the evolved DT.

According to the fitness function the best trees (the most fit ones) have the lowest function values – the aim of the evolutionary process is to minimize the value of local fitness function (LFF) for the best tree. With the combination of selection that prioritizes better solutions, crossover that works as a constructive operator towards local optimums, and mutation that works as a destructive operator in order to keep the needed genetic diversity, the searching for the solution tends to be directed toward the global optimal solution. The global optimal solution is the most appropriate DT regarding the specific needs (expressed in the form of fitness function). As the evolution repeats, more qualitative solutions are obtained regarding the chosen fitness function.

One step further from Podgorelec's approach was introduced by Sprogar with his evolutionary vector decision trees – VEDEC [49]. In his approach the evolution of DT is similar to the one by Podgorelec, but the functionality of DT is enhanced in the way that not only one possible decision class is predicted in the leaf, but several possible questions are answered with a vector of decisions. In this manner, for example a possible treatment for a patient is suggested together with the diagnosis, or several diagnoses are suggested at the same time, which is not possible in ordinary DT.

One of the most recent approaches to the evolutionary induction of DT-like methods is the AREX algorithm developed by Podgorelec and Kokol [34]. In this approach DT are extended by so-called decision programs that are evolved with the help of automatic programming. In this way a classical attribute test can be replaced by a simple computer program, which greatly improves the flexibility, at the cost of computational resources, however. The approach introduces a multi-level classification model, based on the difference between objects. In this manner "simple" objects are classified on the first level with simple rules, more "complex" objects are classified on the second level with more complex rules, etc.

Evolutionary Supported Decision Trees Even though DT induction isn't as complicated from the parameter settings' point of view as some other machine-learning approaches, getting the best out of the approach still requires some time and skill. Manual setting of parameters takes more time than induction of the tree itself and even so, we have no guarantee that parameter settings are optimal. With MtDeciT3.1Gen Zorman presented [59] an implementation of evolutionary supported DT. This approach did not change the induction process itself, since the basic principle remained the same. The change was in evolutionary support, which is in charge of searching the space of induction parameters for the best possible combination. By using genetic operators like selection, crossover and mutation evolutionary approaches are often capable of finding optimal solutions even in the most complex of search spaces or at least they offer significant benefits over other search and optimization techniques.

Enhanced with the evolutionary option, DT don't offer us more generalizing power, but we gain on better coverage of method parameter's search space and significant decrease of time spent compared to manual parameter setting. Better exploitation of a method usually manifests in better results, and the goal is reached.

The rule of the evolutionary algorithm is to find a combination of parameters settings, which would enable induction of DT with high overall accuracy, high class accuracies, use as less attributes as possible, and be as small as possible.

Ensemble Methods An extension of individual classifier models are ensemble methods that use a set of induced classifiers and combine their outputs into a single classification. The purpose of combining several individual classifiers together is to achieve better classification results.

All kinds of individual classifiers can be used to construct an ensemble, however DT are used often as one of the most appropriate and effective models. DT are known to be very sensitive to changes in a learning set. Consequentially, even small changes in a learning set can result in a very different DT classifier which is a basis for the construction of efficient ensembles. Two well-known representatives of ensemble methods are Random forests and Rotation forests, which use DT classifiers, although other classifiers (especially in the case of Rotation forests) could be used as well.

Dealing with Missing Values

Like most of the machine-learning approaches, the DT approach also assumes that all data is available at the time of induction and usage of DT. Real data bases rarely fit this assumption, so dealing with incomplete data presents an important task. In this section some aspects of dealing with missing values for description attributes will be given.

Reasons for incomplete data can be different: from cases where missing values could not be measured, human errors when writing measurement results, to usage of data bases, which were not gathered for the purpose of machine learning, etc.

Missing values can cause problems in three different stages of induction and usage of DT:

- Split selection in the phase of DT induction,
- Partition of training set in a test node during the phase of DT induction, and
- Selecting which edge to follow from a test node when classifying unseen cases during the phase of DT usage.

Though some similarities can be found between approaches for dealing with missing data for listed cases, some specific solutions exist, which are applicable only to one of the possible situations.

Let us first list possible approaches for the case of dealing with missing values during split selection in the phase of DT induction:

- Ignore – the training object is ignored during evaluation of the attribute with the missing value,
- Reduce – the evaluation (for instance entropy, gain or gain ratio) of the attribute with missing values is changed according to the percentage of training objects with missing values,
- Substitute with DT – a new DT with a reduced number of attributes is induced in order to predict the missing value for the training object in question; this approach is best suited for discrete attributes,
- Replace – the missing value is replaced with mean (continuous attributes) or most common (discrete attributes) value, and
- New value – missing values for discrete attributes are substituted with new value ("unknown") which is then treated the same way as other possible values of the attribute.

The next seven approaches are suited for partitioning of the training set in a test node during the phase of DT induction:

- Ignore – the training object is ignored during partition of the test set,
- Substitute with DT – a new DT with a reduced number of attributes is induced in order to predict the missing value for the training object in question – according to the predicted value we can propagate the training object to one of the successor nodes; this approach is best suited for discrete attributes,
- Replace – the missing value is replaced with the mean (continuous attributes) or most common (discrete attributes) value,
- Probability – the training object with the missing value is propagated to one of the edges according to probability, proportional to strength of subsets of training objects,
- Fraction – a fraction of the training object with a missing value is propagated to each of the subsets; the size of the fraction is proportional to the strength of each subset of training objects,
- All – the training object with the missing value is propagated to all subsets, coming from the current node, and
- New value – training objects with new value ("unknown") are assigned to a new subset and propagated further, following a new edge, marked with the previously mentioned new value.

The last case for dealing with missing attribute values is during the decision-making phase, when we are selecting which edge to follow from a test node when classifying unseen cases:

- Substitute with DT – a new DT with a reduced number of attributes is induced in order to predict the missing value for the object in question – according to predicted value we can propagate the unseen object to one of the successor nodes; this approach is best suited for discrete attributes,
- Replace – the missing value is replaced with the mean (continuous attributes) or most common (discrete attributes) value,
- All – the unseen object with a missing value is propagated to all subsets, coming from the current node; the final decision is subject to voting by all leaf nodes we ended in,
- Stop – when facing a missing value, the decision-making process is stopped and the decision with the highest probability according to the training subset in the current node is returned, and
- New value – if there exists a special edge marked with value "unknown", we follow that edge to the next node.

Evaluation of Quality

Evaluation of an induced DT is essential for establishing the quality of the learned model. For the evaluation it is essential to use a set of unseen instances – a testing set. A testing set is a set of testing objects which have not been used in the process of inducing the DT and can therefore be used to objectively evaluate its quality.

A quality is a complex term that can include several measures of efficiency. For the evaluation of the DT's quality the most general measure is the overall accuracy, defined as a percentage of correctly classified objects from all objects (correctly classified and not correctly classified).

Accuracy can be thus calculated as:

$$ACC = \frac{T}{T + F}$$

where T stands for "true" cases (i. e. correctly classified objects) and F stands for "false" cases (i. e. not correctly classified objects).

The above measure is used to determine the overall classification accuracy. In many cases, the accuracy of each specific decision class is even more important than the overall accuracy. The separate class accuracy of ith single decision class is calculated as:

$$ACC_{k,i} = \frac{T_i}{T_i + F_i}$$

and the average accuracy over all decision classes is calculated as

$$ACC_k = \frac{1}{M} \cdot \sum_{i=1}^{M} \frac{T_i}{T_i + F_i}$$

where M represents the number of decision classes.

There are many situations in real-world data where there are exactly two decision classes possible (i. e. positive and negative cases). In such a case the common measures are sensitivity and specificity:

$$\text{Sens} = \frac{TP}{TP + FN}, \quad \text{Spec} = \frac{TN}{TN + FP}$$

where TP stands for "true positives" (i. e. correctly classified positive cases), TN stands for "true negatives" (i. e. correctly classified negative cases), FP stands for "false positives" (i. e. not correctly classified positive cases), and FN stands for "false negatives" (i. e. not correctly classified negative cases).

Besides accuracy, various other measures of efficiency can be used to determine the quality of a DT. Which measures are used depends greatly on the used dataset and the domain of a problem. They include complexity (i. e. the size and structure of an induced DT), generality (differences in accuracy between training objects and testing objects), a number of used attributes, etc.

Applications and Available Software

DT are one of the most popular data-mining models because they are easy to induce, understand, and interpret. High popularity means also many available commercial and free applications. Let us mention a few.

We have to start with Quinlan's C4.5 and C5.0/See5 (MS Windows/XP/Vista version of C5.0), the classic version of DT tool. It is commercial and available at http://www.rulequest.com/see5-info.html.

Another popular commercial tool is CART 5, based on original regression trees algorithm. It is implemented in an intuitive Windows-based environment. It is commercially available at http://www.salford-systems.com/1112.php.

Weka (Waikato Environment for Knowledge Analysis) environment contains a collection of visualization tools and algorithms for data analysis and predictive modeling. Among them there are also some variants of DTs. Beside its graphical user interfaces for easy access, one of its great advantages is the possibility of expanding the tool with own implementations of data analysis algorithms in Java. A free version is available at http://www.cs.waikato.ac.nz/ml/weka/.

OC1 (Oblique Classifier 1) is a decision tree induction system written in C and designed for applications where the instances have numeric attribute values. OC1 builds DT that contain linear combinations of one or more attributes at each internal node; these trees then partition the space of examples with both oblique and axis-parallel hyperplanes. It is available free at http://www.cs.jhu.edu/~salzberg/announce-oc1.html.

There exist many other implementations, some are available as a part of larger commercial packages like SPSS, Matlab, etc.

Future Directions

DT has reached the stage of being one of the fundamental classification tools used on a daily basis, both in academia and in industry. As the complexity of stored data is growing and information systems become more and more sophisticated, the necessity for an efficient, reliable and interpretable intelligent method is growing rapidly. DT has been around for a long time and has matured and proved its quality. Although the present state of research regarding DT gives reasons to think there is nothing left to explore, we can expect major developments in some trends of DT research that are still open. It is our belief that DT will have a great impact on the development of hybrid intelligent systems in the near future.

Bibliography

Primary Literature

1. Babic SH, Kokol P, Stiglic MM (2000) Fuzzy decision trees in the support of breastfeeding. In: Proceedings of the 13th IEEE Symposium on Computer-Based Medical Systems CBMS'2000, Houston, pp 7–11
2. Banerjee A (1994) Initializing neural networks using decision trees In: Proceedings of the International Workshop on Computational Learning Theory and Natural learning Systems, Cambridge, pp 3–15
3. Bonner G (2001) Decision making for health care professionals: use of decision trees within the community mental health setting. J Adv Nursing 35:349–356

4. Breiman L (1996) Bagging predictors. Mach Learn 24:123–140
5. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, Belmont
6. Cantu-Paz E, Kamath C (2000) Using evolutionary algorithms to induce oblique decision trees. In: Proceedings of the Genetic and Evolutionary Computation Conference GECCO-2000, Las Vegas, pp 1053–1060
7. Craven MW, Shavlik JW (1996) Extracting tree-structured representations of trained networks. In: Advances in Neural Information Processing Systems, vol 8. MIT Press, Cambridge
8. Crawford S (1989) Extensions to the CART algorithm. Int J Man-Mach Stud 31(2):197–217
9. Cremilleux B, Robert C (1997) A theoretical framework for decision trees in uncertain domains: Application to medical data sets. In: Lecture Notes in Artificial Intelligence, vol 1211. Springer, London, pp 145–156
10. Dantchev N (1996) Therapeutic decision frees in psychiatry. Encephale-Revue Psychiatr Clinique Biol Therap 22(3):205–214
11. Dietterich TG, Kong EB (1995) Machine learning bias, statistical bias and statistical variance of decision tree algorithms. Mach Learn, Corvallis
12. Feigenbaum EA, Simon HA (1962) A theory of the serial position effect. Br J Psychol 53:307–320
13. Freund Y (1995) Boosting a weak learning algorithm by majority. Inf Comput 121:256–285
14. Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: Machine Learning: Proc. Thirteenth International Conference. Morgan Kauffman, San Francisco, pp 148–156
15. Gambhir SS (1999) Decision analysis in nuclear medicine. J Nucl Med 40(9):1570–1581
16. Gehrke J (2003) Decision Tress. In: Nong Y (ed) The Handbook of Data Mining. Lawrence Erlbaum, Mahwah
17. Goebel M, Gruenwald L (1999) A survey of data mining software tools. SIGKDD Explor 1(1):20–33
18. Goldberg DE (1989) Genetic Algorithms in Search, Optimization, and Machine Learning. Addison Wesley, Reading
19. Hand D (1997) Construction and assessment of classification rules. Wiley, Chichester
20. Heath D et al (1993) k-DT: A multi-tree learning method. In: Proceedings of the Second International Workshop on Multistrategy Learning, Harpers Fery, pp 138–149
21. Heath D et al (1993) Learning Oblique Decision Trees. In: Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence IJCAI-93, pp 1002–1007
22. Ho TK (1998) The Random Subspace Method for Constructing Decision Forests. IEEE Trans Pattern Anal Mach Intell 20(8):832–844
23. Hunt EB, Marin J, Stone PT (1966) Experiments in Induction. Academic Press, New York, pp 45–69
24. Jones JK (2001) The role of data mining technology in the identification of signals of possible adverse drug reactions: Value and limitations. Curr Ther Res-Clin Exp 62(9):664–672
25. Kilpatrick S et al (1983) Optimization by Simulated Annealing. Science 220(4598):671–680
26. Kokol P, Zorman M, Stiglic MM, Malcic I (1998) The limitations of decision trees and automatic learning in real world medical decision making. In: Proceedings of the 9th World Congress on Medical Informatics MEDINFO'98, 52, pp 529–533
27. Letourneau S, Jensen L (1998) Impact of a decision tree on chronic wound care. J Wound Ostomy Conti Nurs 25:240–247
28. Lim T-S, Loh W-Y, Shih Y-S (2000) A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Mach Learn 48:203–228
29. Murthy KVS (1997) On Growing Better Decision Trees from Data, Ph D dissertation. Johns Hopkins University, Baltimore
30. Neapolitan R, Naimipour K (1996) Foundations of Algorithms. DC Heath, Lexington
31. Nikolaev N, Slavov V (1998) Inductive genetic programming with decision trees. Intell Data Anal Int J 2(1):31–44
32. Ohno-Machado L, Lacson R, Massad E (2000) Decision trees and fuzzy logic: A comparison of models for the selection of measles vaccination strategies in Brazil. Proceedings of AMIA Symposium 2000, Los Angeles, CA, US, pp 625–629
33. Paterson A, Niblett TB (1982) ACLS Manual. Intelligent Terminals, Edinburgh
34. Podgorelec V (2001) Intelligent systems design and knowledge discovery with automatic programming. Ph D thesis, University of Maribor
35. Podgorelec V, Kokol P (1999) Induction f medical decision trees with genetic algorithms. In: Proceedings of the International ICSC Congress on Computational Intelligence Methods and Applications CIMA. Academic Press, Rochester
36. Podgorelec V, Kokol P (2001) Towards more optimal medical diagnosing with evolutionary algorithms. J Med Syst 25(3): 195–219
37. Podgorelec V, Kokol P (2001) Evolutionary decision forests – decision making with multiple evolutionary constructed decision trees. In: Problems in Applied Mathematics and Computational Intelligence. WSES Press, pp 97–103
38. Quinlan JR (1979) Discovering rules by induction from large collections of examples. In: Michie D (ed) Expert Systems in the Micro Electronic Age, University Press, Edingburgh, pp 168–201
39. Quinlan JR (1986) Induction of decision trees. Mach Learn 1:81–106
40. Quinlan JR (1987) Simplifying decision trees. Int J Man-Mach Stud 27:221–234
41. Quinlan JR (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco
42. Rich E, Knight K (1991) Artificial Intelligence, 2nd edn. McGraw Hill, New York
43. Sanders GD, Hagerty CG, Sonnenberg FA, Hlatky MA, Owens DK (2000) Distributed decision support using a web-based interface: prevention of sudden cardiac death. Med Decis Making 19(2):157–166
44. Schapire RE (1990) The strength of weak learnability. Mach Learn 5:197–227
45. Shannon C, Weaver W (1949) The mathematical theory of communication. University of Illinois Press, Champagn
46. Shlien S (1992) Multiple binary decision tree classifiers. Pattern Recognit Lett 23(7):757–763
47. Sims CJ, Meyn L, Caruana R, Rao RB, Mitchell T, Krohn M (2000) Predicting cesarean delivery with decision tree models. Am J Obstet Gynecol 183:1198–1206
48. Smyth P, Goodman RM (1991) Rule induction using information theory, In: Piatsky-Scharpiro G, Frawley WJ (eds) Knowledge Discovery in Databases, AAAI Press, Cambridge, pp 159–176
49. Sprogar M, Kokol P, Hleb S, Podgorelec V, Zorman M (2000) Vector decision trees. Intell Data Anal 4(3–4):305–321
50. Tou JT, Gonzalez RC (1974) Pattern Recognition Principles. Addison-Wesley, Reading

51. Tsien CL, Fraser HSF, Long WJ, Kennedy RL (1998) Using classification tree and logistic regression methods to diagnose myocardial infarction. In: Proceedings of the 9th World Congress on Medical Informatics MEDINFO'98, 52, pp 493–497
52. Tsien CL, Kohane IS, McIntosh N (2000) Multiple signal integration by decision tree induction to detect artifacts in the neonatal intensive care unit. Artif Intell Med 19(3):189–202
53. Utgoff PE (1989) Incremental induction of decision trees. Mach Learn 4(2):161–186
54. Utgoff PE (1989) Perceptron trees: a case study in hybrid concept representations. Connect Sci 1:377–391
55. White AP, Liu WZ (1994) Bias in information-based measures in decisions tree induction. Mach Learn 15:321–329
56. Zorman M, Hleb S, Sprogar M (1999) Advanced tool for building decision trees MtDecit 2.0. In: Arabnia HR (ed) Proceedings of the International Conference on Artificial Intelligence ICAI-99. Las Vegas
57. Zorman M, Kokol P, Podgorelec V (2000) Medical decision making supported by hybrid decision trees. In: Proceedings of the ICSC Symposia on Intelligent Systems & Applications ISA'2000, ICSC Academic Press, Wollongong
58. Zorman M, Podgorelec V, Kokol P, Peterson M, Lane J (2000) Decision tree's induction strategies evaluated on a hard real world problem. In: Proceedings of the 13th IEEE Symposium on Computer-Based Medical Systems CBMS'2000, Houston, pp 19–24
59. Zorman M, Sigut JF, de la Rosa SJL, Alayón S, Kokol P, Verliè M (2006) Evolutionary built decision trees for supervised segmentation of follicular lymphoma images. In: Proceedings of the 9th IASTED International conference on Intelligent systems and control, Honolulu, pp 182–187

Books and Reviews

Breiman L, Friedman JH, Olsen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, Belmont
Han J, Kamber M (2006) Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco
Hand D, Manilla H, Smyth P (2001) Principles of Data Mining. MIT Press, Cambridge
Kantardzic M (2003) Data Mining: Concepts, Models, Methods, and Algorithms. Wiley, San Francisco
Mitchell TM (1997) Machine Learning. McGraw-Hill, New York
Quinlan JR (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco
Ye N (ed) (2003) The Handbook of Data Mining. Lawrence Erlbaum, Mahwah

Delay and Disruption in Complex Projects

Susan Howick[1], Fran Ackermann[1], Colin Eden[1], Terry Williams[2]
[1] Strathclyde Business School, University of Strathclyde, Glasgow, UK
[2] School of Management, Southampton University, Southampton, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Disruption and Delay
Analyzing D&D and Project Behavior
Cascade Model Building Process
Implications for Development
Future Directions
Bibliography

Glossary

Cause map A cause map is similar to a cognitive map however it is not composed of an individuals perception but rather the views/statements from a number of participants. It follows the same formalisms as cognitive mapping but does not reflect cognition as it is composite.

Cognitive map A cognitive map is a representation of an individuals perception (cognition) of an issue. It is graphically depicted illustrating concepts/statements connected together with arrows representing causality. They are created using a set of established formalisms.

Complex project A complex project is a project in which the project behaviors and outcomes are difficult to predict and difficult to explain post-hoc.

Disruption and delay Disruption and delay (D&D) is primarily the consequence of interactions which feed on themselves as a result of an initial disruption or delay or portfolio of disruptions and delays.

Project A project is a temporary endeavor undertaken to create a unique product or service [1].

Definition of the Subject

There are many examples of complex projects suffering massive time and cost overruns. If a project has suffered such an overrun there may be a need to understand why it behaved the way it did. Two main reasons for this is (i) to gain learning for future projects or (ii) because one party of the project wishes to claim compensation from another party and thus is trying to explain what occurred during the project. In the latter case, system dynamics has been used for the last 30 years to help to understand why projects behave the way they do. Its success in this arena stems from its ability to model and unravel complex dynamic behavior that can result in project overruns. Starting from the first use of system dynamics in a claim situation in the late 1970's [2], it has directly influenced claim results worth millions of dollars. However, the number

of claims which system dynamics has been involved in is still small as it is not perceived by project management practitioners as a standard tool for analyzing projects. System dynamics has a lot to offer in understanding complex projects, not only in a post-mortem situation, but it could also add value in the pre-project analysis stage and during the operational stage of a project.

Introduction

In this chapter we discuss the role of system dynamics (SD) modeling in understanding, and planning, a complex project. In particular we are interested in understanding how and why projects can go awry in a manner that seems surprising and often very difficult to unravel.

When we refer to projects we mean "a temporary endeavor undertaken to create a unique product or service" [1]. Projects are a specific undertaking, which implies that they are "one-shot", non-repetitive, time-limited, and, when complex, frequently bring about revolutionary (rather than evolutionary) improvements, start (to some extent) without precedent, and are risky with respect to customer, product, and project. If physical products are being created in a project, then the product is in some way significantly different to previous occasions of manufacturing (for example, in its engineering principles, or the expected operating conditions of the product, etc.), and it is this feature that means there is a need to take a project orientation.

Complex projects often suffer massive cost overruns. In recent decades those that have been publicized relate to large public construction projects, for example airports, bridges, and public buildings. Some examples include Denver's US$5 billion airport that was 200% overspent [3], the 800 million Danish Kroner Oresund bridge that was 68% overspent [4], and the UK's Scottish Parliament, which was 10 times the first budget [5]. The Major Projects Association [6] talks of a calamitous history of cost overruns of very large projects in the public sector. Flyvberg et al., [7] describe 258 major transportation infrastructure projects showing 90% of projects overspent. Morris and Hough [8] conclude that "the track record of projects is fundamentally poor, particularly for the larger and more difficult ones. … Projects are often completed late or over budget, do not perform in the way expected, involve severe strain on participating institutions or are canceled prior to their completion after the expenditure of considerable sums of money." (p.7).

"Complex" projects are ones in which the project behaviors and outcomes are difficult to predict and difficult to explain post-hoc. Complex projects, by their nature, comprise multiple interdependencies, and involve nonlinear relationships (which are themselves dynamic). For example, choices to accelerate might involve the use of additional overtime which can affect both learning curves and productivity as a result of fatigue – each of which are non-linear relationships. In addition many of the important features of complex projects are manifested through 'soft' relationships – for example managers will recognize deteriorating morale as projects become messy and look a failure, but assessing the impact of morale on levels of mistakes and rate of working has to be a matter of qualitative judgment. These characteristics are amenable particularly to SD modeling which specializes in working with qualitative relationships that are non-linear [9,10,11].

It is therefore surprising that simulation modeling has not been used more extensively to construct post-mortem analyzes of failed projects, and even more surprising because of SD's aptitude for dealing with feedback. Nevertheless the authors have been involved in the analysis of 10 projects that have incurred time and cost overruns and PA Consulting Group have claimed to have used SD to explain time and cost overruns for over 30 litigation cases [12]. Although in the mid-1990's, attempts to integrate SD modeling with more typical approaches to project management were emerging, their use has never become established within the project management literature or practice [13,14,15]. In addition, recognition that the trend towards tighter project delivery and accelerated development times meant that parallelism in project tasks was becoming endemic, and the impact of increasing parallelism could result in complex feedback dynamics where vicious cycles exist [16]. These vicious cycles are often the consequence of actions taken to enforce feedback control designed to bring a project back on track.

As project managers describe their experiences of projects going wrong they will often talk of these "vicious cycles" occurring, particularly with respect to the way in which customer changes seem to generate much more rework than might be expected, and that the rework itself then generates even more rework. Consider a small part of a manager's description of what he sees going on around him:

> "For some time now we've been short of some really important information the customer was supposed to provide us. As a consequence we've been forced to progress the contract by making engineering assumptions, which, I fear, have led to more mistakes being made than usual. This started giving us more rework than we'd planned for. But, of course, rework on some parts of the project has meant re-

opening work that we thought we'd completed, and that, in turn has reopened even more past work. Engineering rework has led to the need for production work-arounds and so our labour in both engineering and production have been suffering stop/starts and interruptions – and each time this happens they take time to get back up to speed again. This has led to productivity dropping because of unnecessary tasks, let alone productivity losses from the workforce getting fed-up with redoing things over and over again and so just becoming demoralized and so working slower. Inevitably all the rework and consequential productivity losses have put pressure on us to accelerate the project forcing us to have to make more engineering assumptions and do work-arounds."

Figure 1 shows a 'cause map' of the arguments presented by this project manager – the words used in the map are those used by the project manager and the arrows represent the causality described by the manager. This description is full of vicious cycles (indeed there are 35 vicious cycles discussed – see Fig. 1) all triggered by a shortage of customer furnished information and resulting in the rework cycle [17,18,19,20,21] and the need to accelerate in order to keep the project on schedule. Using traditional project management models such as Critical Path Method/Network Analysis cannot capture any of the dynamics depicted in Fig. 1, but SD simulation modeling is absolutely appropriate [22].

So, why has SD modeling been so little used? Partly it is because in taking apart a failed project the purpose is usually associated with a contractor wishing to make a claim for cost-overruns. In these circumstances the traditions of successful claims and typical attitudes of courts tend to determine the approach used. A 'measured-mile' approach is common, where numerical simplicity replaces the need for a proper understanding [23].

It was not until the early 1980's that the use of simulation modeling became apparent from publications in the public-domain. The settlement of a shipbuilding claim [2] prompted interest in SD modeling and [24], in the same year, reported on the use of management science modeling for the same purpose. It was not surprising that this modeling for litigation generated interest in modeling where the purpose was oriented to learning about failed projects (indeed the learning can follow from litigation modeling [25], although it rarely does).

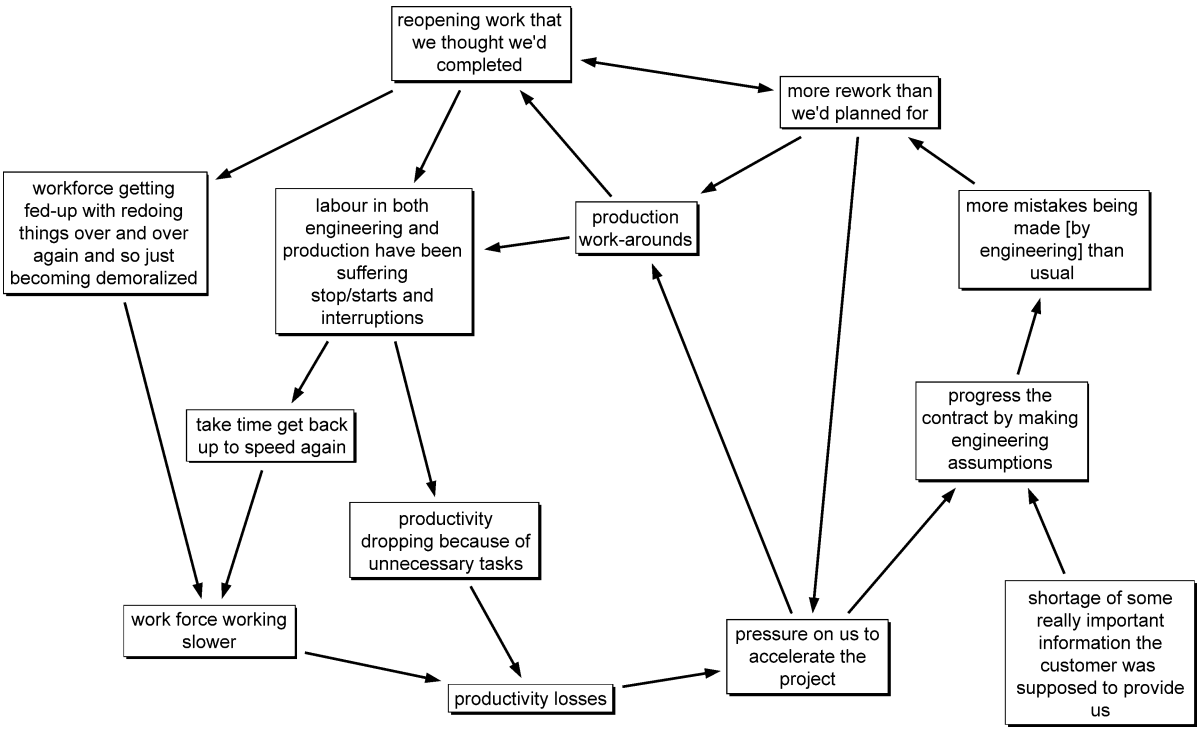

Delay and Disruption in Complex Projects, Figure 1
Cause map showing the interactions described by a project manager and illustrating the feedback loops resulting from the complex dynamics behavior of a project under duress. The *arrows* represent causality

As Fig. 1 demonstrates, it is not easy to understand fully the complex dynamic behavior of a project under duress. Few would realize that 35 feedback loops are encompassed in the description that led to Fig. 1. Indeed one of the significant features of complex projects is the likelihood of underestimating the complexity due to the dynamics generated by disruptions. [6] has reported on the more specific difficulty of understanding feedback behavior and research in the field of managerial judgment reinforces the difficulties of biases unduly influencing judgment [27].

In the work presented here we presume that there is a customer and a contractor, and there is a bidding process usually involving considerations of liquidated damages for delays and possibly strategic reputational consequences for late delivery. Thus, we expect the project to have a clear beginning and an end when the customer (internal or external) signs off a contract. Finally, we do not explore the whole project business life cycle, but that part where major cost overruns occur: thus, we start our consideration when a bid is to be prepared, consider development and manufacturing or construction, but stop when the product of the project is handed over to the customer.

Thus, in this chapter we shall be concerned specifically with the use of SD to model the consequences of disruptions and delays. Often these disruptions are small changes to the project, for example design changes [28]. The work discussed here is the consequence of 12 years of constructing detailed SD simulation models of failed complex projects. The first significant case was reported in Ackermann et al. [10] and Bennett et al. [29]. In each case the prompt for the work was the reasonable prospect of the contractor making a successful claim for damages. In all the cases the claim was settled out of court and the simulation model played a key role in settling the dispute.

The chapter will firstly consider why modeling disruption and delay (D&D) is so difficult. It will discuss what is meant by the term D&D and the typical consequences of D&D. This will be examined using examples from real projects that have suffered D&D. The contribution of SD modeling to the analysis of D&D and thus to the explanation of project behavior will then be discussed. A process of modeling which has been developed over the last 12 years and one that provides a means of modeling and explaining project behavior will be introduced. This process involves constructing both qualitative cause maps and quantitative system dynamics models. The chapter will conclude by considering potential future developments for the use of SD in modeling complex projects.

Disruption and Delay

(The following contains excerpts from Eden at al. [22] which provides a full discussion on the nature of D&D).

The idea that small disruptions can cause serious consequences to the life of a major project, resulting in massive time and cost overruns, is well established. The terms 'disruption and delay' or 'delay and disruption' are also often used to describe what has happened on such projects. However, although justifying the direct impact of disruptions and delays is relatively easy, there has been considerable difficulty in justifying and quantifying the claim for the indirect consequences. Our experience from working on a series of such claims is that some of the difficulty derives from ambiguity about the nature of disruption and delay (D&D). We now consider what we mean by D&D before moving onto considering the types of consequences that can result from the impact of D&D.

What is a Disruption?

Disruptions are events that prevent the contractor completing the work as planned. Many disruptions to complex projects are planned for at the bid stage because they may be expected to unfold during the project. For example, some level of rework is usually expected, even when everything goes well, because there will always be 'normal' errors and mistakes made by both the contractor and client. The disruption and delay that follows would typically be taken to be a part of a risk factor encompassed in the base estimate, although this can be significantly underestimated [30]. However, our experience suggests that *there are other types of disruptions that can be significant in their impact and are rarely thought about during original estimating*. When these types of disruptions do occur, their consequences can be underestimated as they are often seen by the contractor as aberrations with an expectation that their consequences can be controlled and managed. The linkage between risk assessment and the risks as potential triggers of D&D is often missed [31]. Interferences with the flow of work in the project is a common disruption. For example, when a larger number of design comments than expected are made by the client an increased number of drawings need rework. However it also needs to be recognized that these comments could have been made by the contractors own methods engineering staff. In either case, the additional work needed to respond to these comments, increases the contractor's workload and thus requires management to take mitigating actions if they still want to deliver on time. These mitigating actions are usually regarded as routine and capable of easily bringing the contract back to plan,

even though they can have complex feedback ramifications.

Probably one of the most common disruptions to a project comes when a customer or contractor causes changes to the product (a Variation or Change Order). For example, the contractor may wish to alter the product after engineering work has commenced and so request a direct change. However, sometimes changes may be made unwittingly. For example, a significant part of cost overruns may arise where there have been what might be called 'give-aways'. These may occur because the contractor's engineers get excited about a unique and creative solution and rather than sticking to the original design, produce something better but with additional costs. Alternatively, when the contractor and customer have different interpretations of the contract requirements unanticipated changes can occur. For example, suppose the contract asks for a door to open and let out 50 passengers in 2 minutes, but the customer insists on this being assessed with respect to the unlikely event of dominantly large, slow passengers rather than the contractor's design assumptions of an average person. This is often known as 'preferential engineering'. In both instances there are contractor and/or customer requested changes that result in the final product being more extensive than originally intended.

The following example, taken from a real project and originally cited in Eden et al. [30], illustrates the impact of a client induced change to the product:

> **Project 1:** The contract for a 'state of the art' train had just been awarded. Using well-established design principles – adopted from similar train systems – the contractor believed that the project was on track. However within a few months problems were beginning to emerge. The client team was behaving very differently from previous experience and using the contract specification to demand performance levels beyond that envisioned by the estimating team. One example of these performance levels emerged during initial testing, 6 months into the contract, and related to water tightness. It was discovered that the passenger doors were not sufficiently watertight. Under extreme test conditions a small (tiny puddle) amount of water appeared. The customer demanded that there must be no ingress of water, despite acknowledging that passengers experiencing such weather would bring in more water on themselves than the leakage.
>
> The contractor argued that no train had ever met these demands, citing that most manufacturers and operators recognized that a small amount of water would always ingress, and that all operators accepted this. Nevertheless the customer interpreted the contract such that new methods and materials had to be considered for sealing the openings. The dialog became extremely combative and the contractor was forced to redesign. An option was presented to the customer for their approval, one that would have ramifications for the production process. The customer, after many tests and after the verdict of many external experts in the field, agreed to the solution after several weeks. Not only were many designs revisited and changed, with an impact on other designs, but also the delays in resolution impacted the schedule well beyond any direct consequences that could be tracked by the schedule system (www.Primavera.com) or costs forecasting system.

What is a Delay?

Delays are any events that will have an impact on the final date for completion of the project. Delays in projects come from a variety of sources. One common source is that of the client-induced delay. Where there are contractual obligations to comment upon documents, make approvals, supply information or supply equipment, and the client is late in these contractually-defined duties, then there may be a client-induced delay to the expected delivery date (although in many instances the delay is presumed to be absorbed by slack). But also a delay could be self-inflicted: if the sub-assembly designed and built did not work, a delay might be expected.

The different types of client-induced delays (approvals, information, etc.) have different effects and implications. Delays in client approval, in particular, are often ambiguous contractually. A time to respond to approvals may not have been properly set, or the expectations of what was required within a set time may be ambiguous (for example, in one project analyzed by the authors the clients had to respond within n weeks – but this simply meant that they sent back a drawing after n weeks with comments, then after the drawing was modified, they sent back the same drawing after a further m weeks with more comments). Furthermore, excessive comments, or delays in comments can cause chains of problems, impacting, for example, on the document approval process with sub-contractors, or causing over-load to the client's document approval process.

If a delay occurs in a project, it is generally considered relatively straightforward to cost. However, ramifications resulting from delays are often not trivial either to under-

stand or to evaluate. Let us consider a delay only in terms of the CPM (Critical Path Method), the standard approach for considering the effects of delays on a project [32]. The consequences of the delay depend on whether the activities delayed are on the Critical Path. If they *are* on the Critical Path, or the delays are sufficient to cause the activities to become on the critical path, it is conceptually easy to compute the effect as an Extension Of Time (EOT) [33]. However, even in this case there are complicating issues. For example; what is the effect on other projects being undertaken by the contractor? When this is not the first delay, then to which schedule does the term "critical path" refer? To the original, planned programme, which has already been changed or disrupted, or to the "as built", actual schedule? Opinions differ here. It is interesting to note that, "the established procedure in the USA [of using as-built CPM schedules for claims] is almost unheard of in the UK" [33].

If the delay is *not* on the Critical Path then, still thinking in CPM terms, there are only indirect costs. For example, the activities on the Critical Path are likely to be resource dependent, and it is rarely easy to hire and fire at will – so if non-critical activities are delayed, the project may need to work on tasks in a non-optimal sequence to keep the workforce occupied; this will usually imply making guesses in engineering or production, requiring later re-work, less productive work, stop/starts, workforce overcrowding, and so on.

The following example, taken from a real project, illustrates the impact of a delay in client furnished information to the project:

> **Project 2:** A state of the art vessels project had been commissioned which demanded not only the contractor meeting a challenging design but additionally incorporating new sophisticated equipment. This equipment was being developed in another country by a third party. The client had originally guaranteed that the information on the equipment would be provided within the first few months of the contract – time enough for the information to be integrated within the entire design. However time passed and no detailed specifications were provided by the third party – despite continual requests from the contractor to the client.
>
> As the project had an aggressive time penalty the contractor was forced to make a number of assumptions in order to keep the design process going. Further difficulties emerged as information from the third party trickled in demanding changes from the emerging design. Finally manufacturing which had been geared up according to the schedule were forced to use whatever designs they could access in order to start building the vessel.

Portfolio Effect of many Disruptions

It is not just the extent of the disruption or delay but the number of them which may be of relevance. This is particularly the case when a large number of the disruptions and/or delays impact immediately upon one another thus causing a portfolio of changes. These portfolios of D&D impacts result in effects that would probably not occur if only one or two impacts had occurred. For example, the combination of a large number of impacts might result in overcrowding or having to work in poor weather conditions (see example below). In these instances it is possible to identify each individual item as a contributory cause of extra work and delay but not easy to identify the combined effect.

The following example, taken from another real project, illustrates the impact of a series of disruptions to the project:

> **Project 3:** A large paper mill was to be extended and modernized. The extension was given extra urgency by new anti-pollution laws imposing a limit on emissions being enacted with a strict deadline.
>
> Although the project had started well, costs seemed to be growing beyond anything that made sense given the apparent minor nature of the disruptions. Documents issued to the customer for 'information only' were changed late in the process. The customer insisted on benchmarking proven systems, involving visits to sites working with experimental installations or installations operating under different conditions in various different countries. In addition there were many changes of mind about where equipment should be positioned and how certain systems should work. Exacerbating these events was the circumstance of both the customer's and contractor's engineers being co-located, leading to 'endless' discussions and meetings slowing the rate of both design and (later) commissioning.
>
> Relations with the customer, who was seen by the contractor to be continually interfering with progress of the project, were steadily deteriorating. In addition, and in order to keep the construction work going, drawings were released to the construction team before being fully agreed. This meant that construction was done in a piecemeal fashion, often inefficiently (for example, scaffolding would be

put up for a job, then taken down so other work could proceed, then put up in the same place to do another task for which drawings subsequently had been produced). As the construction timescale got tighter and tighter, many more men were put on the site than was efficient (considerable overcrowding ensued) and so each task took longer than estimated.

As a result the project was behind schedule, and, as it involved a considerable amount of external construction work, was vulnerable to being affected by the weather. In the original project plan (as used for the estimate) the outer shell (walls and roof) was due to be completed by mid-Autumn. However, the project manager now found himself undertaking the initial construction of the walls and roofing in the middle of winter! As chance would have it, the coldest winter for decades, which resulted in many days being lost while it was too cold to work. The combination of the particularly vicious winter and many interferences resulted in an unexpectedly huge increase in both labour hours and overall delay. Overtime payments (for design and construction workers) escalated. The final overspend was over 40% more than the original budget.

Consequences of Disruptions and Delays

Disruption and delay (D&D) is primarily the consequence of interactions which feed on themselves as a result of an initial disruption or delay or portfolio of disruptions and delays. If an unexpected variation (or disruption) occurs in a project then, if no intervention was to take place, a delivery delay would occur. In an attempt to avoid this situation, management may choose to take actions to prevent the delay (and possible penalties). In implementing these actions, side-effects can occur which cause further disruptions. These disruptions then cause further delays to the project. In order to avoid this situation, additional managerial action is required. Thus, an initial disruption has led to a delay, which has led to a disruption, which has led to a further delay. A positive feedback loop has been formed, where both disruption and delay feed back on themselves causing further disruptions and delays. Due to the nature of feedback loops, a powerful vicious cycle has been created which, if there is no alternative intervention, can escalate with the potential of getting 'out of control'. It is the dynamic behavior caused by these vicious cycles which can cause severe disruption and consequential delay in a project.

The dynamic behavior of the vicious cycles which are responsible for much of the D&D in a project make the costing of D&D very difficult. It is extremely difficult to separate each of the vicious cycles and evaluate their individual cost. Due to the dynamic behavior of the interactions between vicious cycles, the cost of two individual cycles will escalate when they interact with one another, thus disruptions have to be costed as part of a portfolio of disruptions.

Returning to Project 2, the vessel case, as can be seen in Fig. 2, the client caused both disruptions (continuous changes of mind) and delays (late permission to use a particular product). Both of these caused the contractor to undertake rework, and struggle with achieving a frozen (fixed) design. These consequences in turn impacted upon staff morale and also developed as noted above dynamic behavior – where rework resulted in more submissions of designs, which led to further comments, some of which were inconsistent and therefore led to further rework. As mentioned in the introduction, the rework cycle [17,18,19,20,21] can be a major driver of escalating feedback within a complex project.

Managerial Actions and the Consequences of D&D

The acceleration of disrupted projects to avoid overall project delays is common practice by managers who are under pressure from the client and/or their own senior management to deliver on time. However, the belief that this action will always help avoid delays is naive as it does not take into account an appreciation of the future consequences that can be faced. For example, one typical action designed to accelerate a project is to hire new staff. In doing so, some of the difficulties which may follow are:

- New staff take time to become acquainted with both the project and thus their productivity is lower than that of an existing skilled worker.
- New staff require training on the project and this will have an impact on the productivity of existing staff.
- Rather than hiring new staff to the organization, staff may be moved from other parts of the organization. This action results in costs to other projects as the other project is short of staff and so may have to hire workers from elsewhere, thereby suffering many of the problems discussed above.

Many of the outcomes of this action and other similar actions can lead to a reduction in expected productivity levels. Low productivity is a further disruption to the project through a lack of expected progress. If management identifies this lack of progress, then further managerial actions

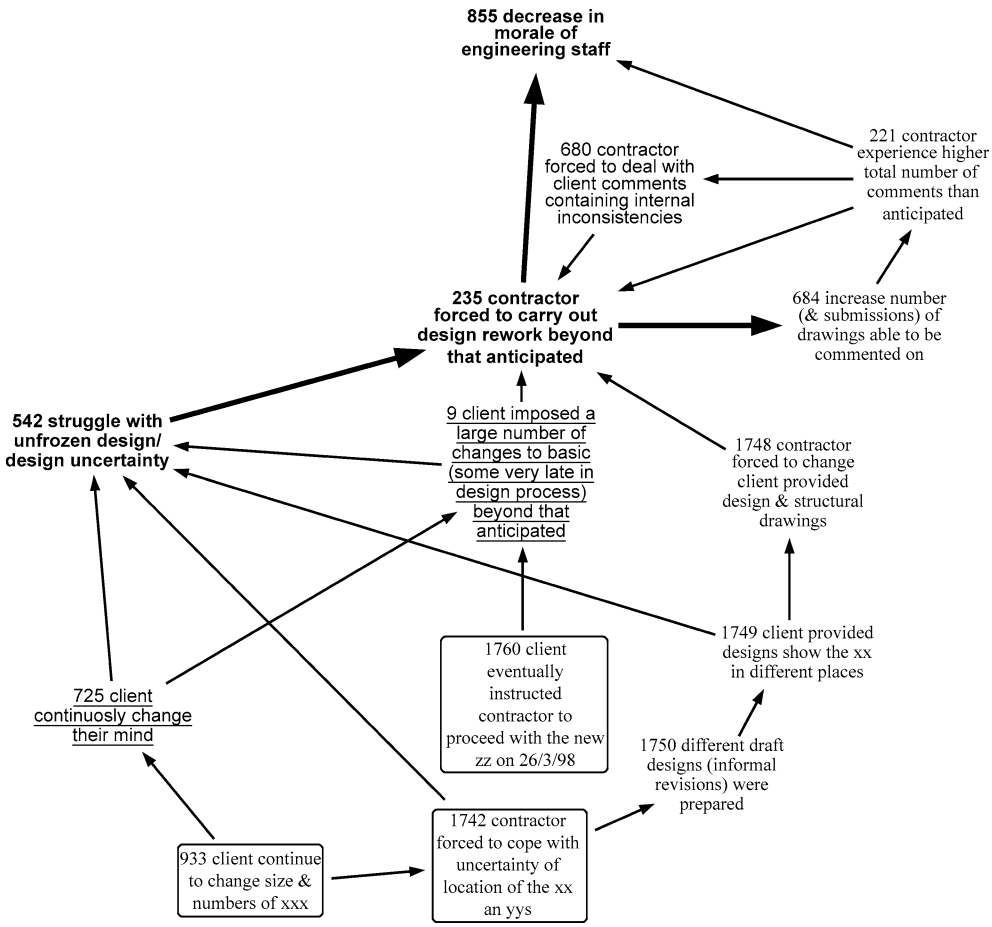

Delay and Disruption in Complex Projects, Figure 2

Excerpt from a cause map showing some of the consequences of disruption and delay in Project 2. *Boxed* statements are specific illustrations with statements *underlined* representing generic categories (e. g. changes of mind). Statements in *bold text* represent the SD variables with the remainder providing additional context. *All links* are causal however those in *bold illustrate sections* of a feedback loop. *The numbers* at the beginning of concept are used as reference numbers in the model

may be taken in an attempt to avoid a further delay in delivery. These actions often lead to more disruptions, reinforcing the feedback loop that had been set up by the first actions.

Two other common managerial actions taken to avoid the impact of a disruption on delivery are (i) the use of overtime and (ii) placing pressure on staff in an attempt to increase work rate. Both of these actions can also have detrimental effects on staff productivity once they have reached particular levels. Although these actions are used to increase productivity levels, effects on fatigue and morale can actually lead to a lowering of productivity via a slower rate of work and/or additional work to be completed due to increased levels of rework [21,34]. This lowering of productivity causes a delay through lack of expected progress on the project, causing a further delay to delivery. Management may then attempt to avoid this by taking other actions which in turn cause a disruption which again reinforces the feedback loop that has been set up.

Analyzing D&D and Project Behavior

The above discussion has shown that whilst D&D is a serious aspect of project management, it is a complicated phenomenon to understand. A single or a series of disruptions or delays can lead to significant impacts on a project which cannot be easily thought through due to human difficulties in identifying and thinking through feedback loops [26,35]. This makes the analysis of D&D and the resulting project behavior particularly difficult to explain.

SD modeling has made a significant contribution to increasing our understanding of why projects behave in the way they do and in quantifying effects. There are two situations in which this is valuable: the claim situation, where one side of the party is trying to explain the project's behavior to the other (and, usually, why the actions of the other party has caused the project to behave in the way it has) and the post-project situation, where an organization is trying to learn lessons from the experience of a project. In the case of a claim situation, although it has been shown that SD modeling can meet criteria for admissibility to court [36], there are a number of objectives which SD, or any modeling method, would need to address [37]. These include the following:

1. Prove causality – show what events triggered the D&D and how the triggers of D&D caused time and cost overruns on the project.
2. Prove the 'quantum' – show that the events that caused D&D created a specific time and cost over-run in the project. Therefore, there is a need to replicate over time the hours of work due to D&D that were over-and-above those that were contracted, but were required to carry out the project.
3. Prove responsibility – show that the defendant was responsible for the outcomes of the project. Also to demonstrate the extent to which plaintiff's management of the project was reasonable and the extent that overruns could not have been reasonably avoided.
4. All of the above have to be proved in a way which will be convincing to the several stakeholders in a litigation audience.

Over the last 12 years the authors have developed a model building process that aims to meet each of these purposes. This process involves constructing qualitative models to aid the process of building the 'case' and thus help to prove causality and responsibility (purposes 1 and 3). In addition, quantitative system dynamics models are involved in order to help to prove the quantum (purpose 2). However, most importantly, the process provides a structured, transparent, formalized process from "real world" interviews to resulting output which enables multiple audiences, including multiple non-experts as well as scientific/expert audiences to appreciate the validity of the models and thus gain confidence in these models and the consulting process in which they are embedded (purpose 4). The process is called the 'Cascade Model Building Process'. The next section describes the different stages of the model building process and some of the advantages of using the process.

Cascade Model Building Process

(The following contains excerpts from Howick et al. [38], which contains a full description of the Cascade Model Building process).

The 'Cascade Model Building Process' involves four stages (see Fig. 3) each of which are described below.

Stage 1: Qualitative Cognitive and Cause Map

The qualitative cognitive maps and /or project cause map aim to capture the key events that occurred on the project, for example a delay as noted above in the vessel example in Project 2. The process of initial elicitation of these events can be achieved in two ways. One option is to interview, and construct cognitive maps [39,40,41] for each participant's views. Here the aim is to gain a deep and rich understanding that taps the wealth of knowledge of each individual. These maps act as a preface to getting the group together to review and assess the total content represented as a merged cause map [42] in a workshop setting. The second option is to undertake group workshops where participants can contribute directly, anonymously and simultaneously, to the construction of a cause map. The participants are able to 'piggy back' off one another, triggering new memories, challenging views and developing together a comprehensive overview [43]. As contributions from one participant are captured and structured to form a causal chain, this process triggers thoughts from others and as a result a comprehensive view begins to unfold. In Project 1, this allowed the relevant design engineers (not just those whose responsibility was the water tight doors, but also those affected who were dealing with car-body structure, ventilation, etc.), methods personnel and construction managers to surface a comprehensive view of the different events and consequences that emerged.

The continual development of the qualitative model, sometimes over a number of group workshops, engenders clarity of thought predominantly through its adherence to the coding formalisms used for cause mapping [44]. Members of the group are able to debate and consider the impact of contributions on one another. Through bringing the different views together it is also possible to check for coherency – do all the views fit together or are there inconsistencies? This is not uncommon as different parts of the organizations (including different discipline groups within a division e. g. engineering) encounter particular effects. For example, during an engineering project, manufacturing can often find themselves bewildered by engineering processes – why are designs so late. However, the first stage of the cascade process enables the views from

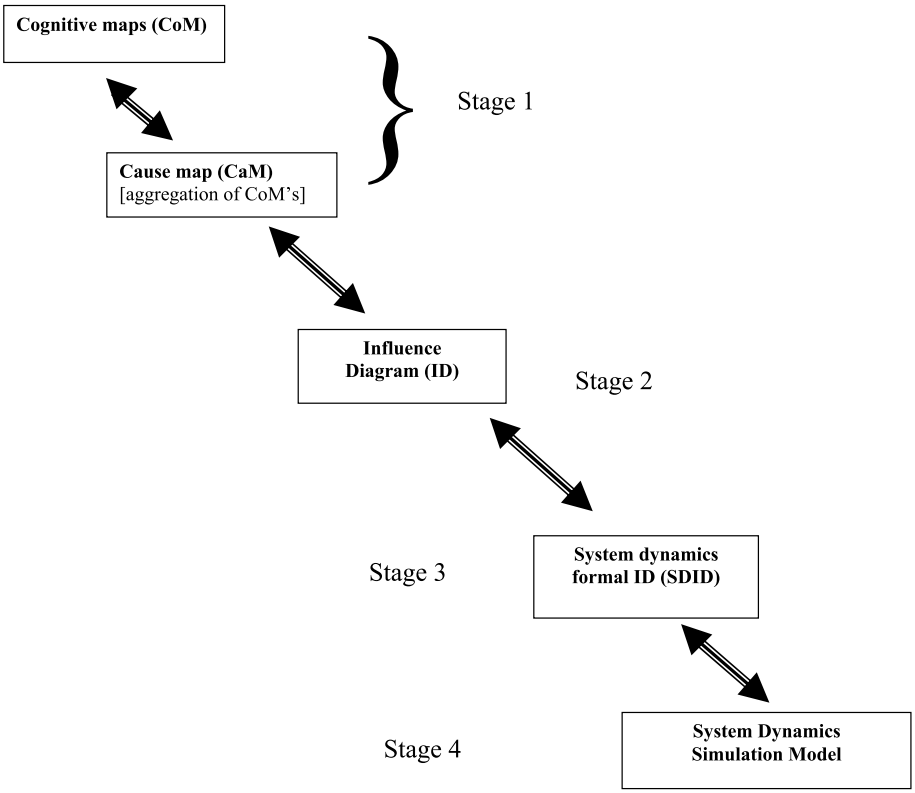

Delay and Disruption in Complex Projects, Figure 3
The Cascade Model Building Process

engineering, methods, manufacturing, commissioning etc. to be integrated. Arguments are tightened as a result, inconsistencies identified and resolved and detailed audits (through analysis and features in the modeling software) undertaken to ensure consistency between both modeling team and model audience. In some instances the documents generated through reports about the organizational situation can be coded into a cause map and merged into the interview and workshop material [45].

The cause map developed at this stage is usually large – containing up to 1000 nodes. Computer supported analysis of the causal map can inform further discussion. For example, it can reveal those aspects of causality that are central to understanding what happened. Events that have multiple consequences for important outcomes can be detected. Feedback loops can be identified and examined. The use of software facilitates the identification of sometimes complex but important feedback loops that follow from the holistic view that arises from the merging of expertise and experience across many disciplines within the organization.

The resulting cause map from stage 1 can be of particular use in proving causality. For example, Fig. 4 represents some of the conversations made regarding the water ingress situation described in the above case. In this figure, consequences such as additional engineering effort and engineering delays can be traced back to events such as client found water seeping out of door.

Stage 2: Cause Map to Influence Diagram

The causal model produced from stage 1 is typically very extensive. This extensiveness requires that a process of 'filtering' or 'reducing' the content be undertaken – leading to the development of an Influence Diagram (ID) (the second step of the cascade process). Partly this is due to the fact that many of the statements captured whilst enabling a detailed and thorough understanding of the project, are not relevant when building the SD model in stage 4 (as a result of the statements being of a commentary like nature rather than a discrete variable). Another reason is that for the most part SD models comprise fewer variables/auxiliaries to help manage the complexity (necessary for good modeling as well as comprehension).

The steps involved in moving from a cause map to an ID are as follows:

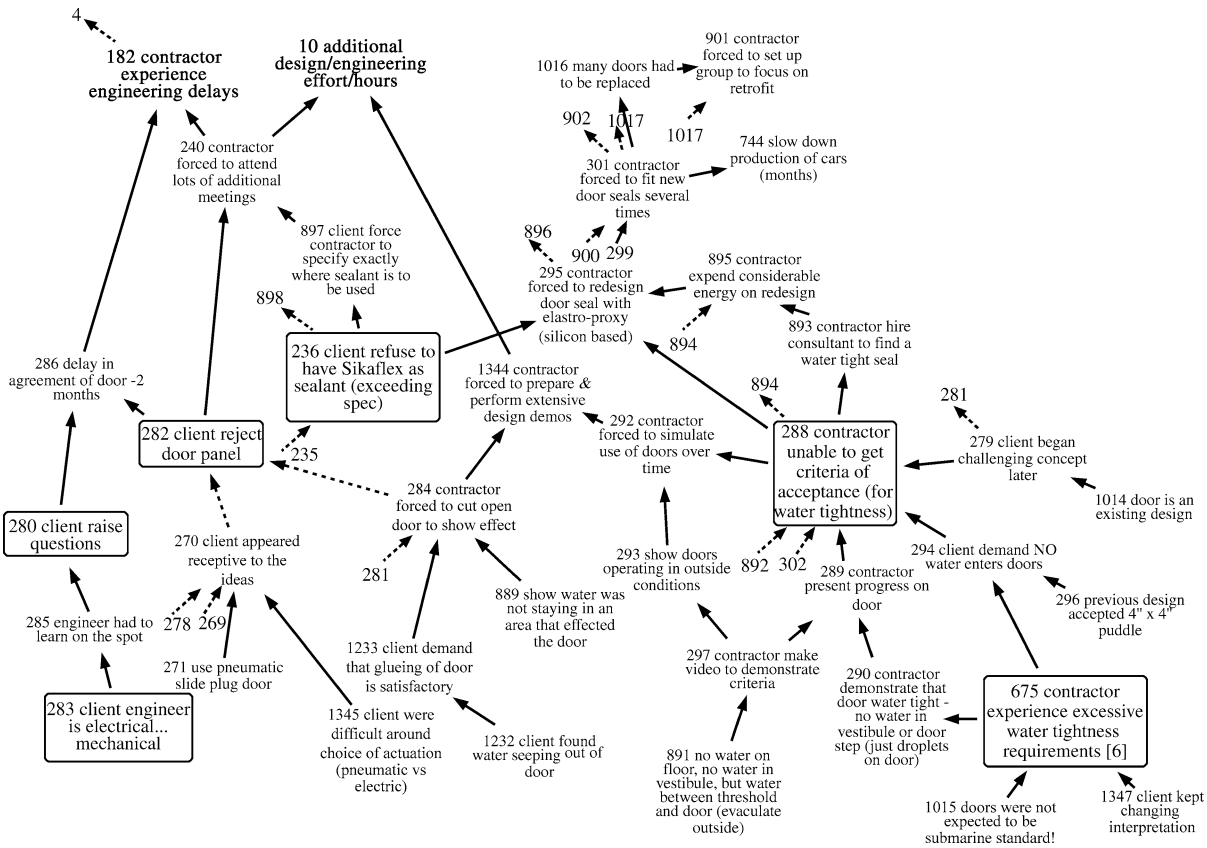

Delay and Disruption in Complex Projects, Figure 4

Excerpt from a cause map showing some of the conversations regarding the water ingress situation in Project 1. As with Fig. 2, statements that have borders are the illustrations, those with *bold font* represent variables with the remainder detailing context. *Dotted arrows* denote the existence of further material which can be revealed at anytime

Step 1: Determining the core/endogenous variables of the ID

(i) Identification of feedback loops: It is not uncommon to find over 100 of these (many of these may contain a large percentage of common variables) when working on large projects with contributions from all phases of the project.

(ii) Analysis of feedback loops: Once the feedback loops have been detected they are scrutinized to determine a) whether there are nested feedback 'bundles' and b) whether they traverse more than one stage of the project. Nested feedback loops comprise a number of feedback loops around a particular topic where a large number of the variables/statements are common but with variations in the formulation of the feedback loop. Once detected, those statements that appear in the most number of the nested feedback loops are identified as they provide core variables in the ID model.

Where feedback loops straddle different stages of the process for example from engineering to manufacturing note is taken. Particularly interesting is where a feedback loop appears in one of the later stages of the project e. g. commissioning which links back to engineering. Here care must be taken to avoid chronological inconsistencies – it is easy to link extra engineering hours into the existing engineering variable however by the time commissioning discover problems in engineering, the majority if not all engineering effort has been completed.

Step 2: Identifying the triggers/exogenous variables for the ID The next stage of the analysis is to look for triggers – those statements that form the exogenous variables in the ID. Two forms of analysis provide clues which can subsequently be confirmed by the group:

(i) The first analysis focuses on starting at the end of the chains of argument (the tails) and laddering up

(following the chain of argument) until a branch point appears (two or more consequences). Often statements at the bottom of a chain of argument are examples which when explored further lead to a particular behavior e. g. delay in information, which provides insights into the triggers.

(ii) The initial set of triggers created by (i) can be confirmed through a second type of analysis – one which takes two different means of examining the model structure for those statements that are central or busy. Once these are identified they can be examined in more detail through creating hierarchical sets based upon them and thus "tear drops" of their content. Each of these teardrops is examined as possible triggers.

Step 3: Checking the ID Once the triggers and the feedback loops are identified care is taken to avoid double counting – where one trigger has multiple consequences some care must be exercised in case the multiple consequences are simple replications of one another.

The resulting ID is comparable to a 'causal loop diagram' [46] which is often used as a pre-cursor to a SD model. From the ID structure it is possible to create "stories" where a particular example triggers an endogenous variable which illustrates the dynamic behavior experienced.

Stage 3: Influence Diagram to System Dynamics Influence Diagram (SDID)

When a SD model is typically constructed after producing a qualitative model such as an ID (or causal loop diagram), the modeler determines which of the variables in the ID should form the stocks and flows in the SD model, then uses the rest of the ID to determine the main relationships that should be included in the SD model. However when building the SD model there will be additional variables/constants that will need to be included in order to make it 'work' that were not required when capturing the main dynamic relationships in the ID. The SDID is an influence diagram that includes all stocks, flows and variables that will appear in the SD model and is, therefore a qualitative version of the SD model. It provides a clear link between the ID and the SD model.

The SDID is therefore far more detailed than the ID and other qualitative models normally used as a pre-cursor to a SD model.

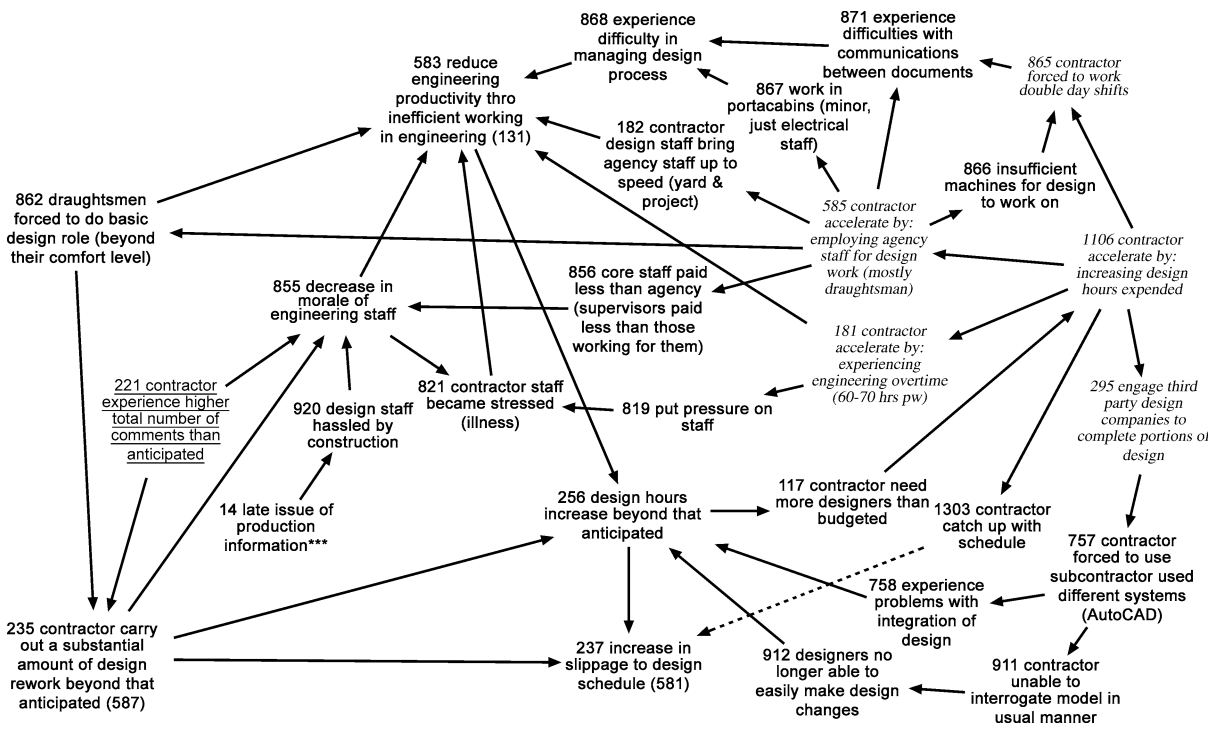

Delay and Disruption in Complex Projects, Figure 5

A small section of an ID from Project 2 showing mitigating actions (*italics*), triggers (*underline*) and some of the feedback cycles

Methods have been proposed to automate the formulation of a SD model from a qualitative model such as a causal loop diagram [47,48,49] and for understanding the underlying structure of a SD model [50]. However, these methods do not allow for the degree of transparency required to enable the range of audiences involved in a claim situation, or indeed as part of an organizational learning experience, to follow the transition from one model to the next. The SDID provides an intermediary step between an ID and a SD model to enhance the transparency of the transition from one model to another for the audiences. This supports an auditable trail from one model to the next.

The approach used to construct the SDID is as follows: The SDID is initially created in parallel with the SD model. As a modeler considers how to translate an ID into a SD model, the SDID provides an intermediary step. For each variable in the ID, the modeler can do either of the following:

(i) Create one variable in the SD & SDID: If the modeler wishes to include the variable as one variable in the SD model, then the variable is simply recorded in both the SDID and the SD model as it appears in the ID.
(ii) Create multiple variables in the SD & SDID: To enable proper quantification of the variable, additional variables need to be created in the SD model. These variables are then recorded in both the SD model and SDID with appropriate links in the SDID which reflect the structure created in the SD model.

The SDID model forces all qualitative ideas to be placed in a format ready for quantification. However, if the ideas are not amenable to quantification or contradict one another, then this step is not possible. As a result of this process, a number of issues typically emerge including the need to add links and statements and the ability to assess the overall profile of the model though examining the impact of particular categories on the overall model structure. This process can also translate back into the causal model or ID model to reflect the increased understanding.

Stage 4: The System Dynamics Simulation Model

The process of quantifying SD model variables can be a challenge, particularly as it is difficult to justify subjective estimates of higher-level concepts such as "productivity" [51]. However, moving up the cascade reveals the causal structure behind such concepts and allows quantification at a level that is appropriate to the data-collection opportunities available. Figure 6, taken from the ID for Project 1, provides an example. The quantitative model

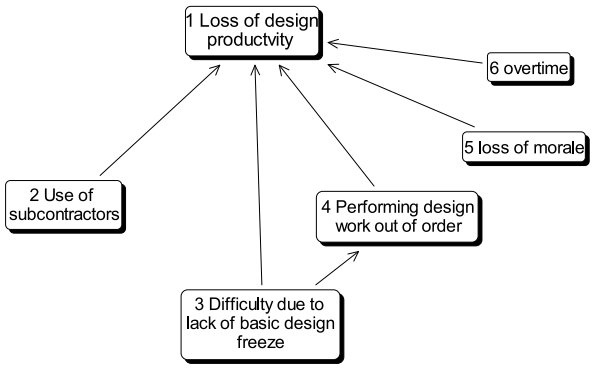

Delay and Disruption in Complex Projects, Figure 6
Section of an ID from Project 1 showing the factors affecting productivity

will require a variable "productivity" or "morale", and the analyst will require estimation of the relationship between it and its exogenous and (particularly) endogenous causal factors. But while the higher-level concept is essential to the quantitative model, simply presenting it to the project team for estimation would not facilitate justifiable estimates of these relationships.

Reversing the Cascade

The approach of moving from stage 1 through to stage 4 can increase understanding and stimulate learning for all parties. However, the process of moving back up the cascade can also facilitate understanding between the parties. For example, in Fig. 7 the idea that a company was forced to use subcontractors and thus lost productivity might be a key part of a case for lawyers. The lawyers and the project team might have come at Fig. 7 as part of their construction of the case. Moving back up from the ID to the Cause Map (i. e. Fig. 7 to Fig. 8) as part of a facilitated discussion not only helps the parties to come to an agreed definition of the (often quite ill-defined) terms involved, it also helps the lawyers understand how the project team arrived at the estimate of the degree of the relationship. Having established the relationship, moving through the SDID (ensuring well-defined variables etc.) to the SD model enables the analysts to test the relationships to see whether

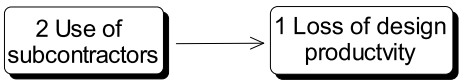

Delay and Disruption in Complex Projects, Figure 7
Section of an ID from Project 1 indicating the influence of the use of subcontractors on productivity

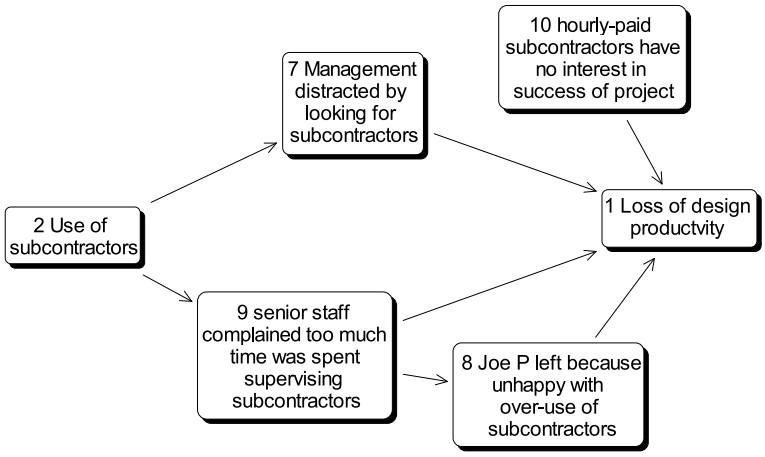

Delay and Disruption in Complex Projects, Figure 8
Section of a Cause Map from Project 1 explaining the relationship between the use of subcontractors and productivity

any contradictions arise, or if model behaviors are significantly different from actuality, and it enables comparison of the variables with data that might be collected by (say) cost accountants. Where there are differences or contradictions, the ID can be re-inspected and if necessary the team presented with the effect of the relationship within the SD model explained using the ID, so that the ID and the supporting cause maps can be re-examined to identify the flaws or gaps in the reasoning. Thus, in this example, as simulation modelers, cost accountants, lawyers and engineers approach the different levels of abstraction, the cascade process provides a unifying structure within which they can communicate, understand each other, and equate terms in each others discourse.

Advantages of the Cascade

The Cascade integrates a well-established method, cause mapping, with SD. This integration results in a number of important advantages for modeling to explain project behavior:

Achieving Comprehensiveness Our experience suggests that one of the principal benefits of using the cascade process derives from the added value gained through developing a rich and elaborated qualitative model that provides the structure (in a formalized manner) for the quantitative modeling. The cascade process immerses users in the richness and subtlety that surrounds their view of the projects and ensures involvement and ownership of all of the qualitative and quantitative models. The comprehensiveness leads to a better understanding of what occurred, which is important due to the complex nature of D&D, and enables effective conversations to take place across different organizational disciplines.

The process triggers new contributions as memories are stimulated and both new material and new connections are revealed. The resultant models thus act as organizational memories providing useful insights into future project management (both in relation to bids and implementation). These models provide more richness and therefore an increased organizational memory when compared to the traditional methods used in group model building for system dynamics models (for example [52]). However this outcome is not untypical of other problem structuring methods [53].

Testing the Veracity of Multiple Perspectives The cascade's bi-directionality enabled the project team's understandings to be tested both numerically and from the perspective of the coherency of the systemic portrayal of logic. By populating the initial quantitative model with data [10] rigorous checks of the validity of assertions were possible.

In a claim situation, blame can be the fear of those participating in accounting for history and often restricts contributions [44]. When initiating the cascade process, the use of either interviews or group workshops increases the probability that the modeling team will uncover the rich story rather than partial explanations or as is often the case with highly politicized situations, 'sanitized' explanations. By starting with 'concrete' events that can be verified, and exploring their multiple consequences, the resultant model provides the means to reveal and explore the different experiences of various stakeholders in the project.

Modeling Transparency By concentrating the qualitative modeling efforts on the capture and structuring of multiple experiences and viewpoints the cascade process initially uses natural language and rich description as the medium which facilitates generation of views and enables a more transparent record to be attained.

There are often insightful moments as participants viewing the *whole* picture realize that the project is more complex than they thought. This realization results in two advantages. The first is a sense of relief that they did not act incompetently given the circumstances i. e. the consequences of D&D took over – which in turn instills an atmosphere more conducive to openness and comprehensiveness (see [44]). The second is learning – understanding the whole, the myriad and interacting consequences and in particular the dynamic effects that occurred on the project (that often acts in a counter-intuitive manner) provides lessons for future projects.

Common Understanding Across many Audiences Claim situations involve numerous stakeholders, with varying backgrounds. The cascade process promotes ownership of the models from this mixed audience. For example, lawyers are more convinced by the detailed qualitative argument presented in the cause map (stage 1) and find this part of greatest utility and hence engage with this element of the cascade. However, engineers get more involved in the construction of the quantitative model and evaluating the data encompassed within it.

A large, detailed system dynamics model can be extremely difficult to understand for many of the stakeholders in a claim process [54]. However, the rich qualitative maps developed as part of the cascade method are presented in terms which are easier for people with no modeling experience to understand. In addition, by moving back up the cascade, the dynamic results that are output by the simulation model are given a grounding in the key events of the project, enabling the audience to be given fuller explanations and reasons for the D&D that occurred on the project.

Using the cascade method, any structure or parameters that are contained in the simulation model can be easily, and quickly, traced back to information gathered as a part of creating the cognitive maps or cause maps. Each contribution in these maps can then normally be traced to an individual witness who could defend that detail in the model. This auditable trail can aid the process of explaining the model and refuting any attacks made on the model.

Clarity The step-by-step process forces the modeler to be clear in what statements mean. Any illogical or inconsistent statements highlighted, require the previous stage to be revisited and meanings clarified, or inconsistencies cleared up. This results in clear, logical models.

Confidence Building As a part of gaining overall confidence in a model, any audience for the model will wish to have confidence in the structure of the model (for example [55,56,57,58]). When assessing confidence levels in a part of the structure of a SD model, the cascade process enables any member of the 'client' audience to clearly trace the structure of the SD model directly to the initial natural language views and beliefs provided from individual interviews or group sessions.

Scenarios are also an important test in which the confidence of the project team in the model can be considerably strengthened. Simulation is subject to the demands to reproduce scenarios that are recognizable to the managers capturing a portfolio of meaningful circumstances that occur at the same time, including many qualitative aspects such as morale levels. For example, if a particular time-point during the quantitative simulation is selected, clearly the simulated values of all the variables, and in particular the relative contributions of factors in each relationship, can be output from the model. If we consider Fig. 6, the simulation might show that at a particular point in a project, loss of productivity is 26% and that the loss due to:

"Use of subcontractors" is 5%.

"Difficulty due to lack of basic design freeze" is 9%.

"Performing design work out of order" is 3%.

"loss of morale" is 5%.

"overtime" is 4%.

Asking the project team their estimates of loss of productivity at this point in time, and their estimation of the relative contribution of these five factors, will help to validate the model. In most cases this loss level is best captured by plotting the relative levels of productivity against the time of critical incidents during the life the project. Discussion around this estimation might reveal unease with the simple model described in Fig. 6, which will enable discussion around the ID and the underlying cause map, either to validate the agreed model, or possibly to modify it and return up the cascade to further refine the model. In this scenario, validation of the cascade process provides a unifying structure within which the various audiences can communicate and understand each other.

The Cascade Model Building Process provides a rigorous approach to explaining why a project has behaved in

a certain way. The cascade uses rich, qualitative stories to give a grounding in the key events that drive the behavior of the project. In addition, it provides a quantifiable structure that allows the over time dynamics of the project to be described. The Cascade has therefore contributed significantly in understanding why projects behave in the way they do.

This chapter has focused on the role of SD modeling in explaining the behavior of complex projects. The final two sections will consider the implications of this work and will explore potential future directions for the use of SD modeling of projects.

Implications for Development

So what is the current status of SD modeling of projects? What is the research agenda for studying projects using SD? Below we consider each aspect of the project life-cycle in turn, to suggest areas where SD modeling may be applied, and to consider where further work is needed.

The first area is pre-project risk analysis. Risk analysis traditionally looks at risks individually, but looking at the systemicity in risks has clear advantages [59]. Firstly, the use of cause mapping techniques by an experienced facilitator, aided by software tools, is a powerful means of drawing out knowledge of project risk from an individual manager (or group of managers), enhancing clarity of thought, allowing investigation of the interactions between risks, and enhancing creativity. It is particularly valuable when used with groups, bringing out interactions between the managers and helping to surface cultural differences. And it clearly enables analysis of the systemicity, in particular identification of feedback dynamics, which can help explicate project dynamics in the ways discussed above. The influence of such work has led to the ideas of cause maps, influence diagrams and SD to be included into risk practice standard advice (the UK "PRAM" Guide, edition 2 [60] – absent from Edition 1). In one key example [31], the work described above enabled the team to develop a 'Risk Filter' in a large multinational project-based organization, for identifying areas of risk exposure on future projects and creating a framework for their investigation. The team reviewed the system after a few years; it had been used by 9 divisions, on over 60 major projects, and completed by 450 respondents; and it was used at several stages during the life of a project to aid in the risk assessment and contribute to a project database. The system allowed investigation of the interactions between risks, and so encouraged the management of the causality of relationships between risks, rather than just risks, thus focusing attention on those risks and causality that create the most frightening ramifications on clusters of risks, as a system, rather than single items. This also encouraged conversations about risk mitigation across disciplines within the organization. Clearly cause mapping is useful in risk analysis, but there are a number of research questions that follow, for example:

- In looking at possible risk scenarios, what are appropriate methodologies to organize and facilitate heterogeneous groups of managers? And how technically can knowledge of systemicity and scenarios be gathered into one integrated SD model and enhance understanding? [61]
- How can SD models of possible scenarios be populated to identify key risks? How does the modeling cascade help in forward-looking analysis?
- There are many attempts to use Monte-Carlo simulation to model projects, without taking the systemic issues into account – leading to models which can be seriously misleading [62]. SD models can give a much more realistic account of the effect of risks – but how can essentially deterministic SD models as described above be integrated into a stochastic framework to undertake probabilistic risk analyzes of projects which acknowledges the systemicity between the risks and the systemic effects of each risk?
- The use of SD is able to identify structures which give projects a propensity for the catastrophic systemic effects discussed in the Introduction. In particular, the three dimensions of structural complexity, uncertainty, and severe time-limitation in projects can combine together to cause significant positive feedback. However, defining metrics for such dimensions still remains an important open question. While a little work has been undertaken to give operational measures to the first of these (for example [63,64]), and de Meyer et al. [65] and Shenhar and Dvir [66] suggest selecting the management strategy based on such parameters, there has been little success so far in quantifying these attributes. The use of the SD models discussed above needs to be developed to a point where a project can be parametrized to give quantitatively its propensity for positive feedback.
- Finally, SD modeling shows that the effects of individual risks can be considerably greater than intuition would indicate, and the effects of clusters of risks particularly so. How can this be quantified so that risks or groups of risks can be ranked in importance to provide prioritization to managers? Again, Howick et al. [61] gives some initial indications here, but more work is needed.

The use of SD in operational control of projects has been less prevalent (Lyneis et al., [12] refers to and discusses examples of where it has been used). For a variety of reasons, SD and the traditional project management approach do not match well together. Traditional project-management tools look at the project in its decomposed pieces in a structured way (networks, work breakdown structures, etc.); they look at operational management problems at a detailed level; SD models aggregate into a higher strategic level and look at the underlying structure and feedback. Rodrigues and Williams [67] describe one attempt at an integrated methodology, but there is scope for research into how work with the SD paradigm can contribute to operational management of projects, and Williams [68] provides some suggestions for hybrid methods.

There is also a more fundamental reason why SD models do not fit in easily into conventional project management. Current project management practice and discourse is dominated by the "Bodies of Knowledge" or BoKs [69], which professional project management bodies consider to be the core knowledge of managing projects [1,70], presenting sets of normative procedures which appear to be self-evidently correct. However, there are three underlying assumptions to this discourse [71].

- Project Management is self-evidently correct: it is rationalist [72] and normative [73].
- The ontological stance is effectively positivist [74].
- Project management is particularly concerned with managing scope in individual parts [75].

These three assumptions lead to three particular *emphases* in current project management discourse and thus in the BoKs [71]:

- A heavy emphasis on planning [73,76].
- An implication of a very conventional control model [77].
- Project management is generally decoupled from the environment [78].

The SD modeling work provided explanations for why some projects severely over-run, which clash with these assumptions of the current dominant project management discourse.

- Unlike the third assumption, the SD models show behavior arising from the complex interactions of the various parts of the project, which would *not* be predicted from an analysis of the individual parts of the project [79].
- Against the first assumption, the SD models show project behavior which is complex and non-intuitive, with feedback exacerbated through management response to project perturbations, conventional methods provide unhelpful or even disbeneficial advice and are not necessarily self-evidently correct.
- The second assumption is also challenged. Firstly, the models differ from the BoKs in their emphasis on, or inclusion of, "soft" factors, often important links in the chains of causality. Secondly, they show that the models need to incorporate not only "real" data but management perceptions of data and to capture the socially constructed nature of "reality" in a project.

The SD models tell us why failures occur in projects which exhibit complexity [63] – that is, when they combine *structural complexity* [80] – many parts in complex combinations – and *uncertainty*, in project goals and in the means to achieve those goals [81]. Goal uncertainty in particular is lacking in the conventional project management discourse [74,82], and it is when uncertainty affects a structurally complex traditionally-managed project that the systemic effects discussed above start to occur. But there is a third factor identified in the SD modeling. Frequently, events arise that compromise the plan at a faster rate than that at which it is practical to re-plan. When the project is heavily *time-constrained*, the project manager feels forced to take acceleration actions. A structurally complex project when perturbed by external uncertainties can become unstable and difficult to manage, and under time-constraints dictating acceleration actions when management has to make very fast and sometimes very many decisions, the catastrophic over-runs described above can occur. Work from different direction seeking to establish characteristics that cause complexity projects come up with similar characteristics (for example [66]). But the SD modeling explains *how* the tightness of the time-constraints strengthen the power of the feedback loops which means that small problems or uncertainties can cause unexpectedly large effects; it also shows how the type of under-specification identified by Flyvberg et al. [4] brings what is sometimes called "double jeopardy" – under-estimation (when the estimate is elevated to the status of a project control-budget) which leads to acceleration actions that then cause feedback which causes much greater over-spend than the degree of under-estimation.

Because of this, the greatest contribution that SD has made – and perhaps can make – is to increase our understanding of why projects behave in the way they do. There are two situations in which this is valuable: the claim situation, where one side of the party is trying to explain the project's behavior to the others (and, usually, why the actions of the other party has caused the project to behave

in the way it has) and the post-project situation, where an organization is trying to learn lessons from the experience of a project.

The bulk of the work referred to in this chapter comes in the first of these, the claim situation. However, while these have proved popular amongst SD modelers, they have not necessarily found universal acceptance amongst the practicing project-management community. Work is needed therefore in a number of directions. These will be discussed in the next section.

Future Directions

We have already discussed the difficulty that various audiences can have in comprehending a large, detailed system dynamics model [54], and that gradual explanations that can be given by working down (and back up) the cascade to bring understanding to a heterogeneous group (which might include jurors, lawyers, engineers and so on) and so link the SD model to key events in the project. While this is clearly effective, more work is needed to investigate the use of the cascade. In particular, ways in which the cascade can be most effective in promoting understanding, in formalizing the methodology and the various techniques mentioned above to make it replicable, as well as how best to use SD here (Howick [54] outlines nine particular challenges the SD modeler faces in such situations). Having said this, it is still the case that many forums in which claims are made are very set in conventional project-management thinking, and we need to investigate more how the SD methods can be combined with more traditional methods synergistically, so that each supports the other (see for example [83]).

Significant unrealized potential of these methodologies are to be found in the post-project "lessons learned" situation. Research has shown many problems in learning generic lessons that can be extrapolated to other projects, such as getting to the root causes of problems in projects, seeing the underlying systemicity, and understanding the narratives around project events (Williams [84], which gives an extensive bibliography in the area). Clearly, the modeling cascade, working from the messiness of individual perceptions of the situation to an SD model, can help in these areas. The first part of the process (Fig. 3), working through to the cause map, has been shown to enhance understanding in many cases; for example, Robertson and Williams [85] describe a case in an insurance firm, and Williams [62] gives an example of a project in an electronics firm, where the methodology was used very "quick and dirty" but still gave increased understanding of why a (in that case successful) project turned out as it did, with some pointers to lessons learned about the process. However, as well as formalization of this part of the methodology and research into the most effective ways of bringing groups together to form cause maps, more clarity is required as to how far down the cascade to go and the additional benefits that the SD modeling brings. "Stage 4" describes the need to look at quantification at a level that is appropriate to the data-collection opportunities available, and there might perhaps be scope for SD models of parts of the process explaining particular aspects of the outcomes. Attempts to describe the behavior of the whole project at a detailed level may only be suitable for the claims situation; there needs to be research into what is needed in terms of Stages 3 and 4 for gaining lessons from projects (or, if these Stages are not carried out, how the benefits such as enhanced clarity and validity using the cause maps, can be gained).

One idea for learning lessons from projects used by the authors, following the idea of simulation "learning labs", was to incorporate learning from a number of projects undertaken by one particular large manufacturer into a simulation learning "game" [25]. Over a period of 7 years, several hundred Presidents, Vice-Presidents, Directors and Project Managers from around the company used the simulation tool as a part of a series of senior management seminars, where it promoted discussion around the experience and the effects encountered, and encouraged consideration of potential long-term consequences of decisions, enabling cause and effect relationships and feedback loops to be formed from participants' experiences. More research is required here as to how such learning can be made most effective.

SD modeling has brought a new view to project management, enabling understanding of the behavior of complex projects that was not accessible with other methods. The chapter has described methodology for where SD has been used in this domain. This last part of the chapter has looked forward to a research agenda into how the SD work needs to be developed to bring greater benefits within the project-management community.

Bibliography

1. Project Management Institute (2000) A guide to the Project Management Body of Knowledge (PMBOK). Project Management Institute, Newtown Square
2. Cooper KG (1980) Naval ship production: a claim settled and a framework built. Interfaces 10:20–36
3. Szyliowicz JS, Goetz AR (1995) Getting realistic about megaproject planning: the case of the new Denver International Airport. Policy Sci 28:347–367
4. Flyvberg B, Bruzelius N, Rothengatter W (2003) Megaprojects

and risk: an anatomy of ambition. Cambridge University Press, Cambridge
5. Scottish Parliament (2003) Corporate body issues August update on Holyrood. Parliamentary News Release 049/2003
6. Major Projects Association (1994) Beyond 2000: A source book for major projects. Major Projects Association, Oxford
7. Flyvberg B, Holm MK, Buhl SL (2002) Understanding costs in public works projects: error or lie? J Am Plan Assoc 68:279–295
8. Morris PWG, Hough GH (1987) The anatomy of major projects. A study of the reality of project management. Wiley, Chichester
9. Forrester J (1961) Industrial dynamics. Productivity Press, Portland
10. Ackermann F, Eden C, Williams T (1997) Modeling for litigation: mixing qualitative and quantitative approaches. Interfaces 27:48–65
11. Lyneis JM, Ford DN (2007) System dynamics applied to project management: a survey, assessment, and directions for future research. Syst Dyn Rev 23:157–189
12. Lyneis JM, Cooper KG, Els SA (2001) Strategic management of complex projects: a case study using system dynamics. Syst Dyn Rev 17:237–260
13. Ford DN (1995) The dynamics of project management: an investigation of the impacts of project process and coordination on performance. Massachusetts Institute of Technology, Boston
14. Rodrigues A, Bowers J (1996) The role of system dynamics in project management. Int J Proj Manag 14:213–220
15. Rodrigues A, Bowers J (1996) System dynamics in project management: a comparative analysis with traditional methods. Syst Dyn Rev 12:121–139
16. Williams TM, Eden C, Ackermann F (1995) The vicious circles of parallelism. Int J Proj Manag 13:151–155
17. Cooper KG (1993) The rework cycle: benchmarks for the project manager. Proj Manag J 24:17–21
18. Cooper KG (1993) The rework cycle: How it really works.. and reworks… PMNETwork VII:25–28
19. Cooper KG (1993) The rework cycle: why projects are mismanaged. PMNETwork VII:5–7
20. Cooper KG (1993) The rework cycle: benchmarks for the project manager. Proj Manag J 24:17–21
21. Cooper KG (1994) The $2,000 hour: how managers influence project performance through the rework cycle. Proj Manag J 25:11–24
22. Eden C, Williams TM, Ackermann F, Howick S (2000) On the nature of disruption and delay. J Oper Res Soc 51:291–300
23. Eden C, Ackermann F, Williams T (2004) Analysing project cost overruns: comparing the measured mile analysis and system dynamics modelling. Int J Proj Manag 23:135–139
24. Nahmias S (1980) The use of management science to support a multimillion dollar precedent-setting government contact litigation. Interfaces 10:1–11
25. Williams TM, Ackermann F, Eden C, Howick S (2005) Learning from project failure. In: Love P, Irani Z, Fong P (eds) Knowledge management in project environments. Elsevier, Oxford
26. Sterman JD (1989) Modelling of managerial behavior: misperceptions of feedback in a dynamic decision making experiment. Manag Sci 35:321–339
27. Kahneman D, Slovic P, Tversky A (1982) Judgment under uncertainty: heuristics and biases. Cambridge University Press, Cambridge
28. Williams TM, Eden C, Ackermann F, Tait A (1995) The effects of design changes and delays on project costs. J Oper Research Society 46:809–818
29. Bennett PG, Ackermann F, Eden C, Williams TM (1997) Analysing litigation and negotiation: using a combined methodology. In: Mingers J, Gill A (eds) Multimethodology: the theory and practice of combining management science methodologies. Wiley, Chichester, pp 59–88
30. Eden C, Ackermann F, Williams T (2005) The amoebic growth of project costs. Proj Manag J 36(2):15–27
31. Ackermann F, Eden C, Williams T, Howick S (2007) Systemic risk assessment: a case study. J Oper Res Soc 58(1):39–51
32. Wickwire JM, Smith RF (1974) The use of critical path method techniques in contract claims. Public Contract Law J 7(1):1–45
33. Scott S (1993) Dealing with delay claims: a survey. Int J Proj Manag 11(3):143–153
34. Howick S, Eden C (2001) The impact of disruption and delay when compressing large projects: going for incentives? J Oper Res Soc 52:26–34
35. Diehl E, Sterman JD (1995) Effects of feedback complexity on dynamic decision making. Organ Behav Hum Decis Process 62(2):198–215
36. Stephens CA, Graham AK, Lyneis JM (2005) System dynamics modelling in the legal arena: meeting the challenges of expert witness admissibility. Syst Dyn Rev 21:95–122.35
37. Howick S (2003) Using system dynamics to analyse disruption and delay in complex projects for litigation: Can the modelling purposes be met? J Oper Res Soc 54(3):222–229
38. Howick S, Eden C, Ackermann F, Williams T (2007) Building confidence in models for multiple audiences: the modelling cascade. Eur J Oper Res 186:1068–1083
39. Eden C (1988) Cognitive mapping: a review. Eur J Oper Res 36:1–13
40. Ackermann F, Eden C (2004) Using causal mapping: individual and group: traditional and new. In: Pidd M (ed) Systems modelling: theory and practice. Wiley, Chichester, pp 127–145
41. Bryson JM, Ackermann F, Eden C, Finn C (2004) Visible thinking: unlocking causal mapping for practical business results. Wiley, Chichester
42. Shaw D, Ackermann F, Eden C (2003) Approaches to sharing knowledge in group problem structuring. J Oper Res Soc 54:936–948
43. Ackermann F, Eden C (2001) Contrasting single user and networked group decision support systems. Group Decis Negot 10(1):47–66
44. Ackermann F, Eden C, Brown I (2005) Using causal mapping with group support systems to elicit an understanding of failure in complex projects: some implications for organizational research. Group Decis Negot 14(5):355–376
45. Eden C, Ackermann F (2004) Cognitive mapping expert views for policy analysis in the public sector. Eur J Oper Res 152:615–630
46. Lane (2000) Diagramming conventions in system dynamics. J Oper Res Soc 51(2):241–245
47. Burns JR (1977) Converting signed digraphs to Forrester schematics and converting Forrester schematics to differential equations. IEEE Trans Syst Man Cybern SMC 7(10):695–707
48. Burns JR, Ulgen OM (1978) A sector approach to the formulation of system dynamics models. Int J Syst Sci 9(6):649–680
49. Burns JR, Ulgen OM, Beights HW (1979) An algorithm for con-

verting signed digraphs to Forrester's schematics. IEEE Trans Syst Man Cybern SMC 9(3):115–124
50. Oliva R (2004) Model structure analysis through graph theory: partition heuristics and feedback structure decomposition. Syst Dyn Rev 20(4):313–336
51. Ford D, Sterman J (1998) Expert knowledge elicitation to improve formal and mental models. Syst Dyn Rev 14(4):309–340
52. Vennix J (1996) Group model building: facilitating team learning using system dynamics. Wiley, Chichester
53. Rosenhead J, Mingers J (2001) Rational analysis for a problematic world revisited. Wiley, Chichester
54. Howick S (2005) Using system dynamics models with litigation audiences. Eur J Oper Res 162(1):239–250
55. Ackoff RL, Sasieni MW (1968) Fundamentals of operations research. Wiley, New York
56. Rivett P (1972) Principles of model building. Wiley, London
57. Mitchell G (1993) The practice of operational research. Wiley, Chichester
58. Pidd M (2003) Tools for thinking: modelling in management science. Wiley, Chichester
59. Williams TM, Ackermann F, Eden C (1997) Project risk: systemicity, cause mapping and a scenario approach. In: Kahkonen K, Artto KA (eds) Managing risks in projects. E & FN Spon, London, pp 343–352
60. APM Publishing Ltd (2004) Project risk analysis and management guide. APM Publishing Ltd, High Wycombe, Bucks
61. Howick S, Ackermann F, Andersen D (2006) Linking event thinking with structural thinking: methods to improve client value in projects. Syst Dyn Rev 22(2):113–140
62. Williams TM (2004) Learning the hard lessons from projects – easily. Int J Proj Manag 22(4):273–279
63. Williams TM (1999) The need for new paradigms for complex projects. Int J Proj Manag 17:269–273
64. Shenhar AJ (2001) One size does not fit all projects: exploring classical contingency domains. Manag Sci 47:394–414
65. De Meyer A, Loch CH, Rich MT (2002) Managing project uncertainty: from variation to chaos. MIT Sloan Mgmt Rev 43(2):60–67
66. Shenhar AJ, Dvir D (2004) How project differ and what to do about it. In: Pinto J, Morris P (eds) Handbook of managing projects. Wiley, New York, pp 1265–1286
67. Rodrigues A, Williams TM (1997) Systems dynamics in software project management: towards the development of a formal integrated framework. Eur J Inf Syst 6:51–66
68. Williams TM (2002) Modelling complex projects. Wiley, Chichester
69. Dixon M (ed) (2000) The Association for Project Management (APM) Body of Knowledge (BoK), 4th edn. Association for Project Management, High Wycombe
70. Stevens M (2002) Project management pathways. Association for Project Management, High Wycombe
71. Williams TM (2005) Assessing and building on project management theory in the light of badly over-run projects. IEEE Trans Eng Manag 52(4):497–508
72. Lundin RA (1995) Editorial: temporary organizations and project management. Scand J Mgmt 11:315–317
73. Packendorff J (1995) Inquiring into the temporary organization: new directions for project management research. Scand J Mgmt 11:319–333
74. Linehan C, Kavanagh D (2004) From project ontologies to communities of virtue. Paper presented at the 2nd International Workshop, Making projects critical, University of Western England, 13–14th December 2004
75. Koskela L, Howell G (2002) The theory of project management: explanation to novel methods. In: Proceedings 10th Annual Conference on Lean Construction, IGLC-10, August 2002, Gramado, Brazil
76. Koskela L, Howell G (2002) The underlying theory of project management is obsolete. In: Proc. PMI (Project Management Institute) Research Conference, Seattle 2002, pp 293–301
77. Hodgson DE (2004) Project work: the legacy of bureaucratic control in the post-bureaucratic organization. Organization 11:81–100
78. Malgrati A, Damiani M (2002) Rethinking the new project management framework: new epistemology, new insights. In: Proc. PMI (Project Management Institute) Research Conference, Seattle 2002, pp 371–380
79. Lindkvist L, Soderlund J, Tell F (1998) Managing product development projects: on the significance of fountains and deadlines. Org Stud 19:931–951
80. Baccarini D (1996) The concept of project complexity – a review. Int J Proj Manag 14:201–204
81. Turner JR, Cochrane RA (1993) Goals-and-methods matrix: coping with projects with ill defined goals and/or methods of achieving them. Int J Proj Manag 11:93–102
82. Engwall M (2002) The futile dream of the perfect goal. In: Sahil-Andersson K, Soderholm A (eds) Beyond project management: new perspectives on the temporary-permanent dilemma. Libe Ekonomi, Copenhagen Business School Press, Malmo, pp 261–277
83. Williams TM (2003) Assessing extension of time delays on major projects. Int J Proj Manag 21(1):19–26
84. Williams TM (2007) Post-project reviews to gain effective lessons learned. Project Management Institute, Newtown Square
85. Robertson S, Williams T (2006) Understanding project failure: using cognitive mapping in an insurance project. Proj Manag J 37(4):55–71

Dependency and Granularity in Data-Mining

Shusaku Tsumoto, Shoji Hirano
Department of Medical Informatics, Shimane University, School of Medicine, Enya-cho Izumo City, Shimane, Japan

Article Outline

Definition of the Subject
Introduction
Contingency Table from Rough Sets
Rank of Contingency Table (2×2)
Rank of Contingency Table ($m \times n$)
Rank and Degree of Dependence
Degree of Granularity and Dependence
Conclusion

Acknowledgment
Bibliography

Definition of the Subject

The degree of granularity of a contingency table is closely related with that of dependence of contingency tables. We investigate these relations from the viewpoints of determinantal devisors and determinants. From the results of determinantal divisors, it seems that the devisors provide information on the degree of dependencies between the matrix of all elements and its submatrices and that an increase in degree of granularity may lead to an increase in dependency. However, another approach shows that a constraint on the sample size of a contingency table is very strong, which leads to an evaluation formula in which an increase of degree of granularity gives a decrease of dependency.

Introduction

Independence (dependence) is a very important concept in data mining, especially for feature selection. In rough sets [2], if two attribute-value pairs, say $[c = 0]$ and $[d = 0]$ are independent, their supporting sets, denoted by C and D do not have an overlapping region ($C \cap D = \phi$), which means that an attribute independent to a given target concept may not appear in the classification rule for the concept.

This idea is also frequently used in other rule discovery methods: let us consider deterministic rules, described as *if-then* rules, which can be viewed as classic propositions ($C \rightarrow D$). From the set-theoretical point of view, a set of examples supporting the conditional part of a deterministic rule, denoted by C, is a subset of a set whose examples belong to the consequence part, denoted by D. That is, the relation $C \subseteq D$ holds and deterministic rules are supported only by positive examples in a dataset [4].

When such a subset relation is not satisfied, indeterministic rules can be defined as if-then rules with probabilistic information [6]. From the set-theoretical point of view, C is not a subset, but closely overlapped with D. That is, the relations $C \cap D \neq \phi$ and $|C \cap D|/|C| \geq \delta$ will hold in this case[1]. Thus, probabilistic rules are supported by a large number of positive examples and a small number of negative examples.

On the other hand, in a probabilistic context, independence of two attributes means that one attribute (a_1) will not influence the occurrence of the other attribute (a_2), which is formulated as $p(a_2|a_1) = p(a_2)$.

Although independence is a very important concept, it has not been fully and formally investigated as a relation between two attributes. Tsumoto introduces linear algebra into formal analysis of a contingency table [5]. The results give the following interesting results: First, a contingency table can be viewed as comparison between two attributes with respect to information granularity. Second, algebra is a key point of analysis of this table. A contingency table can be viewed as a matrix and several operations and ideas of matrix theory are introduced into the analysis of the contingency table. Especially for the degree of independence, rank plays a very important role in extracting a probabilistic model from a given contingency table.

This paper presents a further investigation into the degree of independence of a contingency matrix.

Intuitively and empirically, when two attributes have many values, the dependence between these two attributes becomes low. However, from the results of determinantal divisors, it seems that the devisors provide information on the degree of dependencies between the matrix of all elements and its submatrices and that an increase in degree of granularity may lead to an increase in dependency. The key of the resolution of these conflicts is to consider constraints on the sample size.

In this paper we show that a constraint on the sample size of a contingency table is very strong, which leads to the evaluation formula in which the increase of degree of granularity gives a decrease of dependency. The paper is organized as follows: Sect. "Contingency Table from Rough Sets" shows preliminaries. Sect. "Rank of Contingency Table (2×2)" discusses the former results. Sect. "Rank of Contingency Table ($m \times n$)" gives the relations between rank and submatrices of a matrix. Finally, Sect. "Degree of Granularity and Dependence" concludes this paper.

Contingency Table from Rough Sets

Notations

In the subsequent sections, the following notation, introduced in [3], is adopted: Let U denote a nonempty, finite set called the universe and A denote a nonempty, finite set of attributes, that is, $a: U \rightarrow V_a$ for $a \in A$, where V_a is called the domain of a. Then, a decision table is defined as an information system, $A = (U, A \cup \{\mathcal{D}\})$, where $\{\mathcal{D}\}$ is a set of given decision attributes. The atomic formulas over $B \subseteq A \cup \{\mathcal{D}\}$ and V are expressions of the form $[a = v]$, called descriptors over B, where $a \in B$ and $v \in V_a$. The set $F(B, V)$ of formulas over B is the least set containing all atomic formulas over B and closed with respect to disjunc-

[1] The threshold δ is the degree of the closeness of overlapping sets, which will be given by domain experts. For more information, please refer to Sect. "Rank of Contingency Table (2×2)".

Dependency and Granularity in Data-Mining, Table 1
Contingency table ($m \times n$)

	A_1	A_2	$\cdots$	A_n	Sum
B_1	x_{11}	x_{12}	$\cdots$	x_{1n}	$x_{1\cdot}$
B_2	x_{21}	x_{22}	$\cdots$	x_{2n}	$x_{2\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
B_m	x_{m1}	x_{m2}	$\cdots$	x_{mn}	$x_{m\cdot}$
Sum	$x_{\cdot 1}$	$x_{\cdot 2}$	$\cdots$	$x_{\cdot n}$	$x_{\cdot\cdot} = \lvert U\rvert = N$

tion, conjunction and negation. For each $f \in F(B, V)$, f_A denotes the meaning of f in A, that is, the set of all objects in U with property f, defined inductively as follows:

1. If f is of the form $[a = v]$ then, $f_A = \{s \in U | a(s) = v\}$.
2. $(f \wedge g)_A = f_A \cap g_A; (f \vee g)_A = f_A \vee g_A; (\neg f)_A = U - f_a$.

Contingency Table ($m \times n$)

Definition 1 Let R_1 and R_2 denote multinomial attributes in an attribute space A which have m and n values.

A contingency table is a table of a set described by the following formulas: $|[R_1 = A_j]_A|$, $|[R_2 = B_i]_A|$, $|[R_1 = A_j \wedge R_2 = B_i]_A|$, $|[R_1 = A_1 \wedge R_1 = A_2 \wedge \ldots \wedge R_1 = A_m]_A|$, $|[R_2 = B_1 \wedge R_2 = A_2 \wedge \ldots \wedge R_2 = A_n]_A|$ and $|U|$ ($i = 1, 2, 3, \ldots, n$ and $j = 1, 2, 3, \ldots, m$).

This table is arranged into the form shown in Table 1, where: $|[R_1 = A_j]_A| = \sum_{i=1}^{m} x_{1i} = x_{\cdot j}$, $|[R_2 = B_i]_A| = \sum_{j=1}^{n} x_{ji} = x_{i\cdot}$, $|[R_1 = A_j \wedge R_2 = B_i]_A| = x_{ij}$, $|U| = N = x_{\cdot\cdot}$ ($i = 1, 2, 3, \ldots, n$ and $j = 1, 2, 3, \ldots, m$).

Example 1 Let us consider an information table shown in Table 2.

The relationship between b and e can be examined by using the corresponding contingency table. First, the frequencies of four elementary relations, called *marginal distributions*, are counted: $[b = 0]$, $[b = 1]$, $[e = 0]$, and $[e = 1]$. Then, the frequencies of four kinds of conjunction are counted: $[b = 0] \wedge [e = 0]$, $[b = 0] \wedge [e = 1]$, $[b = 1] \wedge [e = 0]$, and $[b = 1] \wedge [e = 1]$.

Dependency and Granularity in Data-Mining, Table 2
A small dataset

a	b	c	d	e
1	0	0	0	1
0	0	1	1	1
0	1	2	2	0
1	1	1	2	1
0	0	2	1	0

Dependency and Granularity in Data-Mining, Table 3
Corresponding contingency table

	$b = 0$	$b = 1$	
$e = 0$	1	1	2
$e = 1$	2	1	3
	3	2	5

Then, the following contingency table is obtained (Table 3).

From this table, accuracy and coverage for $[b = 0] \rightarrow [e = 0]$ are obtained as $1/(1 + 2) = 1/3$ and $1/(1 + 1) = 1/2$.

One of the important observations from granular computing is that a contingency table shows the relations between two attributes with respect to intersection of their supporting sets. For example, in Table 3, both b and e have two different partitions of the universe and the table gives the relation between b and e with respect to the intersection of supporting sets. It is easy to see that this idea can be extended into an $n \times n$ contingency table, which can be viewed as an $n \times n$-matrix. When two attributes have a different number of equivalence classes, the situation may be a little complicated. But, in this case, due to knowledge about linear algebra, we have only to consider the attribute which has a smaller number of equivalence classes and the surplus number of equivalence classes of the attributes with larger number of equivalence classes can be projected into other partitions. In other words, an $m \times n$ matrix or contingency table includes a projection from one attribute to the other one.

Rank of Contingency Table (2×2)

Preliminaries

Definition 2 A corresponding matrix $C_{T_{a,b}}$ is defined as a matrix the elements of which are equal to the value of the corresponding contingency table $T_{a,b}$ of two attributes a and b, except for marginal values.

Definition 3 The rank of a table is defined as the rank of its corresponding matrix. The maximum value of the rank is equal to the size of (square) matrix, denoted by r.

Example 2 Let the table given in Table 3 be defined as $T_{b,e}$. Then, $C_{T_{b,e}}$ is:

$$\begin{pmatrix} 1 & 1 \\ 2 & 1 \end{pmatrix}.$$

Since the determinant of $C_{T_{b,e}}$ $\det(C_{T_{b,e}})$ is not equal to 0, the rank of $C_{T_{b,e}}$ is equal to 2. It is the maximum value ($r = 2$), so b and e are statistically dependent.

Independence when the Table is 2 × 2

From the application of linear algebra, several results are obtained. (The proof is omitted). First, it is assumed that a contingency table is given as $m = 2$, $n = 2$ in Table 1. Then the corresponding matrix ($C_{T_{R_1,R_2}}$) is given as:

$$\begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} .$$

Proposition 1 *The determinant of* $\det(C_{T_{R_1,R_2}})$ *is equal to* $|x_{11}x_{22} - x_{12}x_{21}|$.

Proposition 2 *The rank will be:*

$$\text{rank} = \begin{cases} 2, & \text{if} \det(C_{T_{R_1,R_2}}) \neq 0 \\ 1, & \text{if } \det(C_{T_{R_1,R_2}}) = 0 . \end{cases}$$

If the rank of $\det(C_{T_{b,e}})$ is equal to 1, according to the theorems of linear algebra, it is obtained that one row or column will be represented by the other column. That is,

Proposition 3 *Let r_1 and r_2 denote the rows of the corresponding matrix of a given 2×2 table, $C_{T_{b,e}}$. That is,*

$$r_1 = (x_{11}, x_{12}) , \quad r_2 = (x_{21}, x_{22}) .$$

Then, r_1 can be represented by r_2: $r_1 = kr_2$, where k is given as:

$$k = \frac{x_{11}}{x_{21}} = \frac{x_{12}}{x_{22}} = \frac{x_{1\cdot}}{x_{2\cdot}} .$$

From this proposition, the following theorem is obtained.

Theorem 1 *If the rank of the corresponding matrix is 1, then two attributes in a given contingency table are statistically independent.*

Thus,

$$\text{rank} = \begin{cases} 2, & \textit{dependent} \\ 1, & \textit{statistical independent} . \end{cases}$$

Rank of Contingency Table ($m \times n$)

In the case of a general square matrix, the results in the 2×2 contingency table can be extended. It is especially important to observe that conventional statistical independence is supported only when the rank of the corresponding matrix is equal to 1. Let us consider the contingency table of c and a in Table 2, which is obtained as follows. Thus, the corresponding matrix of this table is:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} ,$$

whose determinant is equal to 0. It is clear that its rank is 2.

Dependency and Granularity in Data-Mining, Table 4
Contingency table for *a* and *c*

	$a = 0$	$a = 1$	
$c = 0$	0	1	1
$c = 1$	1	1	2
$c = 2$	2	0	2
	3	2	5

It is interesting to see that if the case of $[d = 0]$ is removed, then the rank of the corresponding matrix is equal to 1 and two rows are equal. Thus, if the value space of d into $\{1, 2\}$ is restricted, then c and d are statistically independent. This relation is called *contextual independence* [1], which is related to conditional independence.

However, another type of weak independence is observed: let us consider the contingency table of a and c. The table is obtained as Table 4.

Its corresponding matrix is:

$$C_{T_{a,c}} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 0 \end{pmatrix} ,$$

Since the corresponding matrix is not square, the determinant is not defined. But it is easy to see that the rank of this matrix is two. In this case, even removing any attribute-value pair from the table will not generate statistical independence. Finally, the relation between rank and independence in a $n \times n$ contingency table is obtained.

Theorem 2 *Let the corresponding matrix of a given contingency table be a square $n \times n$ matrix. If the rank of the corresponding matrix is 1, then two attributes in a given contingency table are statistically independent. If the rank of the corresponding matrix is n, then two attributes in a given contingency table are dependent. Otherwise, two attributes are contextually dependent, which means that several conditional probabilities can be represented by a linear combination of conditional probabilities. Thus,*

$$\text{rank} = \begin{cases} n & \textit{dependent} \\ 2, \ldots, n-1 & \textit{contextual independent} \\ 1 & \textit{statistical independent} . \end{cases}$$

This theorem can be generalized into $m \times n$ matrix. If the corresponding matrix of a given contingency table is not square and of the form $m \times n$, then its rank is at most $\min(m, n)$.

For example, since $C_{T_{a,c}}$ is 3×2, the rank is at most 2. Actually, from the calculation of subdeterminants shown in the next section, this matrix has a rank of 2.

Theorem 3 *Let the corresponding matrix of a given contingency table be an $m \times n$ matrix. The rank of this matrix*

is less than $\min(m, n)$. *If the rank of the corresponding matrix is 1, then two attributes in a given contingency table are statistically independent. If the rank of the corresponding matrix is n, then two attributes in a given contingency table are dependent. Otherwise, two attributes are contextually dependent, which means that several conditional probabilities can be represented by a linear combination of conditional probabilities. Thus,*

$$\text{rank} = \begin{cases} \min(m, n) & \textit{dependent} \\ 2, \ldots, \min(m, n) - 1 & \textit{contextual independent} \\ 1 & \textit{statistical independent}\,. \end{cases}$$

In the cases of $m \neq n$, we need a discussion on submatrix and subdeterminant in the next section.

Rank and Degree of Dependence

Submatrix and Subdeterminant

The next interest is the structure of a corresponding matrix with $1 \leq \text{rank} \leq n - 1$. First, let us define a submatrix (a subtable) and subdeterminant.

Definition 4 Let A denote a corresponding matrix of a given contingency table $(m \times n)$. A corresponding submatrix $A^{i_1 i_2 \ldots i_r}_{j_1 j_2 \ldots j_s}$ is defined as a matrix which is given by an intersection of r rows and s columns of A $(i_1 < i_2 < \ldots < i_r, j_1 < j_2 < \ldots < j_r)$.

Definition 5 A subdeterminant of A is defined as a determinant of a submatrix $A^{i_1 i_2 \ldots i_r}_{j_1 j_2 \ldots j_s}$, which is denoted by $\det(A^{i_1 i_2 \ldots i_r}_{j_1 j_2 \ldots j_s})$.

Let us consider the contingency table given as Table 1. Then, a subtable for $A^{i_1 i_2 \ldots i_r}_{j_1 j_2 \ldots j_s}$ is given as Table 5.

Rank and Subdeterminant

Let δ_{ij} denote a co-factor of a_{ij} in a square corresponding matrix of A. Then,

$$\Delta_{ij} = (-1)^{i+j} \det\left(A^{1,2,\ldots,i-1,i+1,\ldots,n}_{1,2,\ldots,j-1,j+1,\ldots,n}\right)\,.$$

Dependency and Granularity in Data-Mining, Table 5
A subtable ($r \times s$)

	A_{j_1}	A_{j_2}	$\cdots$	A_{j_r}	**Sum**
B_{i_1}	$x_{i_1 j_1}$	$x_{i_1 j_2}$	$\cdots$	$x_{i_1 j_r}$	$x_{i_1 \cdot}$
B_{i_2}	$x_{i_2 j_1}$	$x_{i_2 j_2}$	$\cdots$	$x_{i_2 j_r}$	$x_{i_2 \cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
B_{i_r}	$x_{i_r j_1}$	$x_{i_r j_2}$	$\cdots$	$x_{i_r j_n}$	$x_{i_r \cdot}$
Sum	$x_{\cdot 1}$	$x_{\cdot 2}$	$\cdots$	$x_{\cdot n}$	$x_{\cdot\cdot} = \lvert U \rvert = N$

It is notable that a co-factor is a special type of submatrix, where only the ith-row and j-column are removed from a original matrix.

By the use of co-factors, the determinant of A is defined as:

$$\det(A) = \sum_{j=1}^{n} a_{ij} \Delta_{ij}\,,$$

which is called *Laplace expansion*.

From this representation, if $\det(A)$ is not equal to 0, then $\Delta_{ij} \neq 0$ for $\{a_{i1}, a_{i2}, \ldots, a_{in}\}$ which are not equal to 0. Thus, the following proposition is obtained.

Proposition 4 $\det(A)$ *is not equal to 0 if at least one co-factor of* $a_{ij} (\neq 0)$, Δ_{ij} *is not equal to 0.*

It is notable that the above definition of a determinant gives the relation between an original matrix A and its submatrices (co-factors). Since cofactors give a square matrix of size $n - 1$, the above proposition gives the relation between a matrix of size n and submatrices of size $n - 1$. In the same way, we can discuss the relation between a corresponding matrix of size n and submatrices of size $r (1 \leq r < n - 1)$.

Rank and Submatrix

Let us assume that corresponding matrix and submatrix are square ($n \times n$ and $r \times r$, respectively).

Theorem 4 *If the rank of a corresponding matrix of size* $n \times n$ *is equal to r, the determinant of at least one submatrix of size* $r \times r$ *is not equal to 0. That is, there exists a submatrix* $A^{i_1 i_2 \ldots i_r}_{j_1 j_2 \ldots j_r}$, *which satisfies* $\det(A^{i_1 i_2 \ldots i_r}_{j_1 j_2 \ldots j_r}) \neq 0$.

Corollary 1 *If the rank of a corresponding matrix of size* $n \times n$ *is equal to r, all the determinants of the submatrices whose number of columns and rows are larger than* $r + 1 (\leq n)$ *are equal to 0.*

Example 3 Let us consider the corresponding matrix mentioned in the above section, $C_{T_{a,c}}$.

The submatrices of this matrix are:

$$\begin{pmatrix} 1 & 1 \\ 2 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

Since all the subderminants are not equal to 0, the rank of this corresponding matrix is equal to 2.

Example 4 Let us consider the following corresponding matrix:

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}.$$

The determinant of A is:

$$\begin{aligned}\det(A) &= 1 \times (-1)^{1+1} \det \begin{pmatrix} 5 & 6 \\ 8 & 9 \end{pmatrix} \\ &\quad + 2 \times (-1)^{1+2} \det \begin{pmatrix} 4 & 6 \\ 7 & 9 \end{pmatrix} \\ &\quad + 3 \times (-1)^{1+3} \det \begin{pmatrix} 4 & 5 \\ 7 & 8 \end{pmatrix} \\ &= 1 \times (-3) + 2 \times 6 + 3 \times (-3) = 0 \,.\end{aligned}$$

Thus, the rank of A is smaller than 2. All the subdeterminants of A are:

$$\det \begin{pmatrix} 5 & 6 \\ 8 & 9 \end{pmatrix} = -3 \,, \qquad \det \begin{pmatrix} 4 & 6 \\ 7 & 9 \end{pmatrix} = -6 \,,$$
$$\det \begin{pmatrix} 4 & 5 \\ 7 & 8 \end{pmatrix} = -3 \,, \qquad \det \begin{pmatrix} 1 & 2 \\ 7 & 8 \end{pmatrix} = -6 \,,$$
$$\det \begin{pmatrix} 1 & 3 \\ 7 & 9 \end{pmatrix} = -12 \,, \qquad \det \begin{pmatrix} 2 & 3 \\ 8 & 9 \end{pmatrix} = -6 \,,$$
$$\det \begin{pmatrix} 1 & 2 \\ 4 & 5 \end{pmatrix} = -3 \,, \qquad \det \begin{pmatrix} 1 & 3 \\ 4 & 6 \end{pmatrix} = -6 \,,$$
$$\det \begin{pmatrix} 2 & 3 \\ 5 & 6 \end{pmatrix} = -3 \,.$$

Since all the subdeterminants of A are not equal to 0, the rank of A is equal to 2.

Actually, since

$$\begin{pmatrix} 4 & 5 & 6 \end{pmatrix} = \frac{1}{2} \left\{ \begin{pmatrix} 1 & 2 & 3 \end{pmatrix} + \begin{pmatrix} 7 & 8 & 9 \end{pmatrix} \right\} \,,$$

and $(7\ 8\ 9)$ cannot be represented by $k(1\ 2\ 3)$ (k: integer), the rank of this matrix is equal to 2.

Thus, one attribute-value pair is statistically dependent on other two pairs, statistically independent of the other attribute. In other words, if two pairs are fixed, the one remaining attribute-value pair will be statistically independently determined.

Determinantal Divisors

From the subdeterminants of all submatrices of size 2, all the subdeterminants of a corresponding matrix has the greatest common divisor, equal to 3.

From the recursive definition of the determinants, it is show that the subdeterminants of size $r+1$ will have the greatest common divisor of the subdeterminants of size r as a divisor. Thus,

Theorem 5 *Let $d_k(A)$ denote the greatest common divisor of all the subdeterminants of size k, $\det(A^{i_1 i_2 \ldots i_k}_{j_1 j_2 \ldots j_r})$. $d_1(A), d_2(A), \ldots, d_n(A)$ are called determinantal divisors. From the definition of Laplace expansion,*

$$d_k(A) | d_{k+1}(A) \,.$$

In the example of the above subsection, $d_1(A) = 1$, $d_2(A) = 3$ and $d_3(A) = 0$.

Example 5 Let us consider $C_{T_{a,c}}$ as an example. $d_1(C_{T_{a,c}}) = 1$ and $d_2(C_{T_{a,c}}) = 1$.

Example 6 Let us consider the following corresponding matrix:

$$B = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & \mathbf{11} & 9 \end{pmatrix} \,.$$

Calculation gives: $d_1(B) = 1$, $d_2(B) = 3$ and $d_3(B) = 18$.

It is notable that a simple change of a corresponding matrix gives a significant change to the determinant, which suggests a change of structure with regard to dependence/independence.

The relation between $d_k(A)$ gives an interesting constraint.

Proposition 5 *Since $d_k(A) | d_{k+1}(A)$, the sequence of the divisors is a monotonically increasing one:*

$$d_1(A) \le d_2(A) \le \ldots \le d_r(A) \,,$$

where r denotes the rank of A.

The sequence of B illustrates this: $1 < 3 < 18$.

Let us define a ratio of $d_k(A)$ to $d_{k-1}(A)$, called *elementary divisors*, where C denotes a corresponding matrix and $k \le \operatorname{rank} A$:

$$e_k(C) = \frac{d_k(C)}{d_{k-1}(C)} (d_0(C) = 0) \,.$$

Elementary divisors may give the increase of dependency between two attributes. For example, $e_1(B) = 1$, $e_2(B) = 3$, and $e_3(B) = 6$. Thus, a transition from 2×2 to 3×3 have a higher impact on the dependency of two attributes.

It is trivial to see that $\det(B) = e_1 e_2 e_3$, which can be viewed as a decomposition of the determinant of a corresponding matrix.

Divisors and Degree of Dependence

Since the determinant can be viewed as the degree of dependence, this result is very important. If values of all the

subdeterminants (size r) are very small (nearly equal to 0) and $d_r(A) \simeq 1$, then the values of the subdeterminants (size $r+1$) are very small. This property may hold until the r reaches the rank of the corresponding matrix. Thus, the sequence of the divisors of a corresponding matrix gives a hidden structure of a contingency table.

Also, these results show that $d_1(A)$ and $d_2(A)$ are very important to estimate the rank of a corresponding matrix. Since $d_1(A)$ is given only by the greatest common divisor of all the elements of A, $d_2(A)$ are much more important components. This also intuitively suggests that the subdeterminants of A with size 2 are principal components of a corresponding matrix from the viewpoint of statistical dependence.

Recall that statistical independence of two attributes is equivalent to a corresponding matrix with a rank of 1. A matrix of rank 2 gives a context-dependent independence, which means three values of two attributes are independent, but two values of two attributes are dependent.

Subdeterminants and Degree of Dependence

Since the determinants give the degree of dependence, the degree of dependence can be evaluated by the values of subdeterminants.

For the above examples (A), since

$$\det\begin{pmatrix} 1 & 3 \\ 7 & 9 \end{pmatrix} = -12$$

gives the maximum value, the first and the third attribute-value pairs for two attributes are dependent on each other.

On the other hand, concerning B, since

$$\det\begin{pmatrix} 2 & 3 \\ 11 & 9 \end{pmatrix} = -15$$

gives the maximum value, the second and the third attribute-value pairs for two attributes are dependent on each other.

This discussion can be extended into the dependency between attribute-value pairs and a corresponding attribute. Let us consider 3×2 submatrices of A, which removes one column of the matrix

$$A_1 = \begin{pmatrix} 2 & 3 \\ 5 & 6 \\ 8 & 9 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1 & 3 \\ 4 & 6 \\ 7 & 9 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 1 & 2 \\ 4 & 5 \\ 7 & 8 \end{pmatrix}.$$

From the discussions in the above subsections, a set of the subderminants of 2×2 submatrices of A_j, denoted by D_{A_j} are obtained as:

$$\begin{aligned} D_{A_1} &= \{-3, -6, -3\} \\ D_{A_2} &= \{-12, -6, -6\} \\ D_{A_3} &= \{-3, -6, -3\}\,. \end{aligned}$$

Thus, the first and third attribute value pairs are more dependent than the second value pairs, concerning the classification of attributes for the rows.

Elementary Divisors and Elementary Transformation

Let us define the following three elementary (row/column) transformations of a corresponding matrix:

1. Exchange two rows (columns), i_0 and j_0 ($P(i_0, j_0)$).
2. Multiply -1 to a row (column) i_0 ($T(i_0; -1)$).
3. Multiply t to a row (column) j_0 (i_0) and add it to a row i_0 (j_0). ($W(i_0, j_0, t)$).

Then, three transformations have several interesting characteristics.

Proposition 6 *Matrices corresponding to three elementary transformations are regular.*

Proposition 7 *Three elementary transformations do not change the rank of a corresponding matrix.*

Proposition 8 *Let $\tilde{A}$ denote a matrix transformed by finite steps of three operations. Then,*

$$\operatorname{rank} \tilde{A} = \operatorname{rank} A\,, \quad d_r(\tilde{A}) = d_r(A)\,,$$

where r denotes the rank of matrix A.

Then, from the application of linear algebra, the following interesting result is obtained.

Theorem 6 *With the finite steps of elementary transformations, a given corresponding matrix is transformed into*

$$\tilde{A} = \left(\begin{array}{cccc|c} e_1 & & & & \\ & e_2 & & & \\ & & \ddots & & O \\ & & & e_r & \\ & & O & & O \end{array}\right)$$

where $e_j = \frac{d_j(A)}{d_{j-1}(A)}$ ($d_0(A) = 1$) and r denotes the rank of a corresponding matrix. Then, the determinant is decomposed into the product of e_j.

$$d_r(\tilde{A}) = d_r(A) = e_1 e_2 \ldots e_r\,.$$

Degree of Granularity and Dependence

From Theorem 6, it seems that the increase of the degree of granularity gives that of the dependence between two attributes.

However, our empirical observations are different from the above intuitive analysis. Thus, there should be a strong constraint which suppresses the above effects on the degree of granularity.

Let us assume that the determinant of a given contingency matrix gives the degree of the dependence of the matrix. Then, from the application of linear algebra, we obtain the following theorem.

Theorem 7 *Let A denote an $n \times n$ contingency matrix, which includes N samples. If the rank of A is equal to n, then there exists a matrix B $(n \times n)$ which satisfies*

$$BA = \begin{pmatrix} \rho_1 & & & O \\ & \rho_2 & & \\ & & \ddots & \\ O & & & \rho_n \end{pmatrix} = P\,,$$

where $\rho_1 + \rho_2 + \ldots + \rho_n = N$.

It is notable that the value of determinants of P is larger than A:

$$\det A \leq \det P\,.$$

Example 7 Let us consider B as an example (Example 6). Let C denote the orthogonal matrix for transformation of B. Since the cardinality of B is equal to 48, the diagonal matrix which gives the maximum determinant is equal to:

$$\begin{pmatrix} 16 & 0 & 0 \\ 0 & 16 & 0 \\ 0 & 0 & 16 \end{pmatrix}.$$

On the other hand, the determinant of B is equal to 18. Thus, $\det B = 18 < 16^3 = 4096$. Then, C is obtained from the following equation:

$$C \times \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 11 & 9 \end{pmatrix} = \begin{pmatrix} 16 & 0 & 0 \\ 0 & 16 & 0 \\ 0 & 0 & 16 \end{pmatrix}.$$

Thus,

$$C = \begin{pmatrix} -56/3 & 40/3 & -8/3 \\ 16/3 & -32/3 & 16/3 \\ 8 & 8/3 & -8/3 \end{pmatrix}.$$

It is notable that the determinant of C is equal to 2048/9. Also, since $\det B = 18$, we do not have any diagonal matrix whose determinant is equal to 18 and whose sum of all the elements is equal to 48.

It is easy to see that the transformed matrix P has a very nice property for calculating the determinant.

Proposition 9 *The determinant of the transformed matrix P is equal to the multiplication of ρ_1 to ρ_n. That is,*

$$\det P = \rho_1 \rho_2 \ldots \rho_n\,.$$

Then, the following constraint will be have the special meaning:

$$\rho_1 + \rho_2 + \ldots + \rho_n = N\,, \tag{1}$$

because the following inequality holds in general:

$$\frac{\rho_1 + \rho_2 + \ldots + \rho_n}{n} \geq \sqrt[n]{\rho_1 \rho_2 \ldots \rho_n} \tag{2}$$

where the equality holds when $\rho_1 = \rho_2 = \ldots = \rho_n$.

Since the above inequality can be transformed into:

$$\rho_1 \rho_2 \ldots \rho_n \leq \left(\frac{\rho_1 + \rho_2 + \ldots + \rho_n}{n} \right)^n,$$

the following inequality is obtained:

$$\det P = \rho_1 \rho_2 \ldots \rho_n \leq \left(\frac{\rho_1 + \rho_2 + \ldots + \rho_n}{n} \right)^n, \tag{3}$$

where the equality holds when $\rho_1 = \rho_2 = \ldots = \rho_n$.

From the Theorem 7 and Eq. (1), the following theorem is obtained.

Theorem 8 *When a contingency matrix A holds $AB = P$, where P is a diagonal matrix, the following inequality holds:*

$$\det A \leq \left(\frac{N}{n} \right)^n.$$

Proof

$$\begin{aligned} \det A &= \det(PB^{-1}) \\ &\leq \det P \\ &= \rho_1 \rho_2 \ldots \rho_n \\ &\leq \left(\frac{\rho_1 + \rho_2 + \ldots + \rho_n}{n} \right)^n = \left(\frac{N}{n} \right)^n, \end{aligned} \tag{4}$$

where the former equality holds when $\det B^{-1} = \det B = 1$ and the latter equality holds when $\rho_1 = \rho_2 = \ldots = \rho_n = \frac{N}{n}$. □

Example 8 Let us consider the following contingency matrices D and E:

$$D = \begin{pmatrix} 1 & 2 & 3 & 0 \\ 4 & 5 & 6 & 0 \\ 7 & \mathbf{11} & 9 & 0 \\ 0 & 0 & 0 & \mathbf{1} \end{pmatrix}$$

$$E = \begin{pmatrix} 1 & 2 & 3 & 0 \\ 4 & 5 & 6 & 0 \\ 7 & \mathbf{10} & 9 & 0 \\ 0 & 0 & 0 & \mathbf{1} \end{pmatrix} .$$

The numbers of examples of D and E are 49 and 48, respectively, which can be comparable to that of B. Then, from Theorem 8,

$$\det D = 18 < (49/4)^4 = \frac{5764801}{256} \sim 22\,518$$
$$\det E = 12 < (48/4)^4 = 20\,736 .$$

Thus, the maximum value of the determinant of A is at most $\left(\frac{N}{n}\right)^n$. Since N is constant for the given matrix A, the degree of dependence will decrease very rapidly when n becomes very large. That is,

$$\det A \sim n^{-n} .$$

Thus,

Corollary 2 *The determinant of A will converge to 0 when n increases to infinity*

$$\lim_{n\to\infty} \det A = 0 .$$

This result suggests that when the degree of granularity becomes higher, the degree of dependence will become lower due to constraints on the sample size.

However, it is notable that N/n is very important. If N is very large, a rapid decrease will be observed when N is close to n. Even when N is 48 as shown in Example 8, $n = 3, 4$ may give a strong dependency between two attributes. For the behavior of $(N/n)^n$, we can apply the technique of real analysis, which will be our future work.

Conclusion

In this paper, a contingency table is interpreted from the viewpoint of granular computing and statistical independence. Matrix algebra is a key point of the analysis of a contingency table and the degree of independence, and rank plays a very important role in extracting a probabilistic model. From the correspondence between contingency table and matrix, the following results are obtained: First, the value of determinants gives the degree of of dependency between attribute-value pairs for a set of submatrices with the same size. Second, from the characteristics of the determinants, the larger the rank a corresponding matrix has, the more the two attributes are dependent. This result is shown by a monotonicity of a sequence of determinantal divisors. Third, elementary divisors give a decomposition of the determinant of a corresponding matrix. Finally, the constraint on the sample size of a contingency table is very strong, which leads to an evaluation formula in which an increase of degree of granularity gives the decrease of dependency.

Acknowledgment

This work was supported by the Grant-in-Aid for Scientific Research (13131208) on Priority Areas (No.759) "Implementation of Active Mining in the Era of Information Flood" by the Ministry of Education, Science, Culture, Sports, Science and Technology of Japan.

Bibliography

1. Butz C (2002) Exploiting contextual independencies in web search and user profiling. In: Proceedings of World Congress on Computational Intelligence (WCCI'2002) (CD-ROM)
2. Pawlak Z (1991) Rough sets. Kluwer, Dordrecht
3. Skowron A, Grzymala-Busse J (1994) From rough set theory to evidence theory. In: Yager R, Fedrizzi M, Kacprzyk J (eds) Advances in the Dempster–Shafer Theory of Evidence. Wiley, New York, pp 193–236
4. Tsumoto S (2000) Knowledge discovery in clinical databases and evaluation of discovered knowledge in outpatient clinic. Inf Sci 124:125–137
5. Tsumoto S (2003) Statistical independence as linear independence. In: Skowron A, Szczuka M (eds) Electronic Notes in Theoretical Computer Science, vol 82. Elsevier
6. Tsumoto S, Tanaka H (1996) Automated discovery of medical expert system rules from clinical databases based on rough sets. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining 96. AAAI Press, Palo Alto, pp 63–69

Development, Complex Dynamic Systems of

Paul van Geert
The Heymans Institute, Groningen, The Netherlands

Article Outline

Glossary
Definition of the Subject
Introduction

Adaptation and Adaptive Agent Systems
Main Theories in the Field and Short Historical Background
The Study of Individual Development in Light of Complexity Theory
An Overview of the Human Life Span in Light of the Theory of Complex Dynamic Systems
Future Directions
Bibliography

Glossary

Development In the context of developmental psychology, development implies the process of increasing knowledge, skill, capacity and so forth across the life span, in an ordered and directional process, leading to a stable state of maturity. Development implies an increase in complexity of the developing person or system.

Education A co-adaptive process involving asymmetrical relationships between educators (parents, teachers, ...) and young persons (children, pupils, ...); the process of the upbringing of children, by means of teaching, providing resources, models, teaching, guidance and so forth

Learning and teaching The process of gaining knowledge or skills, often in the context of help by a more competent person, who enables the learning through teaching, consisting of guidance, transmission of knowledge, structuring, promoting and confining the learner's zones of action, often in the context of explicit learning and teaching goals (e. g. teaching and learning how to write, teaching and learning how to weld iron on a construction site, etc.)

Complex developmental system A developmental system consisting of components such as the persons involved, material and cultural artifacts, properties attributable to individuals, that shows development as defined earlier, through the interactions between and interdependence of the components

Dynamic systems (developmental) A way of describing how one state of a developmental system changes into another state, as defined by a developmental state space (the whole of possible states distinguishable in the system, described by the set of dimensions or variables needed to specify the system as being developmental)

Developmental states, stages and ages A developmental state is any possible state in the developmental state space, which is defined by the dimensions used to describe development; with continuous dimensions, the number of possible states is continuous and infinite; stages are states characterized by a stability that lasts over a sufficiently long time span (a few years) and by a pattern of mutually dependent properties, i. e. values of the developmental dimensions, stages are characteristic of the founding theories, such as Piaget's theory; ages are periods in the lifespan characterized by sufficient stability of the properties to qualify as properties characteristic of that period, ages often coincide with stages as distinguished in classical theories but often comprise additional properties required for coherent description.

Developmental domains Physical, neurological and sensorimotor development; cognitive development; language development; emotional development, social and personality development; self-, gender- and identity development, moral development; domains follow characteristic paths of development, can be distinguished from one another on the basis of their components and laws of change, but closely interact with one another and form interdependencies on the level of action and developmental time scales.

Developmental time scales Development takes place across various time scales, characterized by their characteristic event duration, the laws or principles that govern change on that particular time scale and interdependencies with other time scales; in order of descending duration the time scales relevant to human development are the scale of biological evolution; the time scale of socio-cultural historical development; the time scale of development across the human life span; the time scale of action and real-time experience and the time scale of underlying neurobiological processes.

Definition of the Subject

Developmental psychology concerns the study of developmental changes in human beings across the life span. Developmental changes are broadly defined as changes in the organism that are mostly progressive – in terms of increasing complexity, adaptation to the environment, efficiency of actions and operations and so forth. The subject of developmental psychology is the individual person, embedded in a particular social, cultural and material context. Although it is now widely accepted that development is a life-long process, the main developmental events take place during the first part of the life span, which co-occur with physical growth and development and coincide roughly with the age from birth to adulthood. Developmental changes through adulthood and old-age are often referred to as processes of aging and imply processes of

loss, for instancing due to aging of the nervous system, and processes compensating for such losses.

Developmental psychology is also concerned with problematic development, for instance in the form of developmental psychopathology, studying the life-span development of conditions such as autism, attention deficit disorders, oppositional behavior and so forth.

The aim of studying development from the perspective of complex dynamic systems is to apply the concepts of complexity and dynamic systems to the phenomena, theories and explanations currently found in developmental psychology, including the educational sciences, in order to arrive at a comprehensive theoretical approach on the subject that focuses on the mechanisms and forms of change.

Introduction

Terms of Change

Developmental psychology deals with various terms of change, some of which have already been defined under the Glossary terms: Development; education; learning. Other terms referring to developmental change that are worth considering are the life span (the period between conception and death in a single individual); socio-cultural evolution (the historical process of changes in human cultures and societies as they pertain to the life span and development of individuals); and variability and fluctuation. The latter refer to non-permanent changes in a developmentally relevant property or variable that are in principle occurring over the short-term (as compared to the long-term of the preceding forms of change), variability and fluctuation typically occur around one (or several) central tendencies or pivotal points.

Etymologically, development means unwrapping or unfolding, as in the unwrapping or unfolding of a book roll, or the unwrapping or unfolding of a flower bud [317,330,341,345]. Development thus carries an implicit notion of an inner logic in the sequence of the unfolding, a notion of potentiality (what is in there must come out) and a notion of finality (the unfolding comes to an end when the folded object is spread out). Although the historical meaning of development can of course not determine how we see or define development in scientific discourse, the deliberate application of this term in a particular context – instead of words like maturation, learning etc. – implies that we wish to refer to a phenomenon that is characterized to a more than a trivial extent, by these notions of inner logic, potentiality and finality. Development implies a directed process of change towards or unfolding of a mature state. It is a directed process, from an immature to a mature state, implying increasing complexity in terms of a system that differentiates (incorporates more and more elements, features, knowledge ...) and at the same time integrates (constructs connections between the components).

Readers familiar with dynamic systems will immediately recognize these notions as metaphorical representations of self-organizing dynamics. The inner logic corresponds with the evolution term or the change function that governs the dynamics, and the potentiality and finality refer to self-organization or the systems tendency to move towards a particular attractor state. The notion of increasing developmental complexity refers to theories of complexity and emergence [60,146,147,362]. In short, given its core assumptions, developmental psychology is a natural domain of application for the approach of nonlinear, complex dynamic systems. Development, moreover, concerns a complex dynamic system characterized by adaptation in various senses, namely adaptation of the developing individual to the environment, adaptation of the environment to the developing individual (in the sense of education, but also in the sense of long-lasting historical and cultural adaptations as a result of intergenerational effects on development). In short, in order to describe human development, we need three grounding notions: complexity, dynamic systems and adaptation.

Complexity

A Working Definition of Complexity In the context of our discussion of the complex dynamics of developmental processes, we shall use the following working definition of complexity.

- A complex system consists of many components or elements; the magnitude of "many" typically depends on the nature of the system
- The components are interacting; the interactions occur on the basis of a few, simple interaction principles, with a system-characteristic degree of connectedness among the components
- The components change because of their interactions with other components
- And are thus interdependent
- The complex system shows characteristic higher-order properties (exceeding the properties or behaviors of the single components, implying characteristic patterns of related behaviors among many components)
  - Examples are sub-systems, trajectories of long-term change or development, and events at various time scales (see further)
- That are emergent on the interactions, i. e. they occur through self-organization

- Complex systems naturally divide and organize into sub-systems
  - Sub-systems are also complex in the sense of the current definition
  - Sub-systems are defined by the strength of the connections between the components of the sub-system, and these strengths may vary over a broad range, thus allowing the possibility of hard- as well as soft-assembled subsystems
- Complex systems are characterized by patterns and mechanisms of change that occur on various interdependent time scales
  - A characteristic distinction is that between short-term versus long-term processes, that are interdependent, with long-term processes determining the constraints and parameters of the short-term processes, and short-term processes determining the constraints and parameters of the long-term change
  - Patterns of change are characterized by non-linear phenomena such as the emergence of attractors, phase transitions, "tipping point" or "domino" effects, slow and gradual change, surge or peak phenomena, bimodal states and fluctuations, and so forth
- Complex social or human systems are characterized by the embeddedness of the observer
  - The observer is a member of the system he or she studies, and thus makes self-referring statements when explaining and describing the complex system
  - To the embedded observer, complexity often corresponds with various cognitive states, that relate to properties of the complex system, such as ambiguity, fuzziness, contradictions, superposition and entanglement, surprise and so forth, that are not necessarily in principle reducible to unambiguous, crisp, non-contradictory and independent statements or beliefs.

An Application of Complexity to Development

Networks of Interacting and Interdependent Components and Subsystems They depend on the level of organization (and corresponding time scale) or system component chosen. For instance, for a student of language development, the major component chosen is that of language. To understand its development, one must reckon with the fact that language must be subdivided into various subsystems, e. g. the lexicon, syntax, meaning and so forth. Each subsystem, for instance the lexicon, can be further subdivided into components, such as the lemma's (words) in the lexicon. The subsystems interact, but they also interact with non-linguistic systems, such as the members of the community of language users (e. g. the parents, siblings), the perceptual-motor system of the child itself, and so forth. A theory of the development of such collection of elements or components will need to specify the dynamic relationships between them. For instance, across development, the development of language is dynamically related to the development of social understanding (e. g. theory-of-mind) and vice versa. Hence, the developing language of a child is a network of interactions among components. Interdependency means that connected components cannot be treated as independent variables, or independent components. For instance, a child's current linguistic skill is dependent on its effective environment, in terms of learning opportunities, and its the effective environment is interdependent on the child's mind (in terms of the language addressed to the child).

In addition to the notion of subsystems, we can also invoke the term "levels of organization", which refers to levels of particular types of patterns or structures that are stable at their characteristic time scale. An example is the life-span history of a person with a number of characteristic properties, including those of the person's characteristic life spaces; another example at a shorter time scale is the example of a counting strategy in which a child uses his fingers to count and make simple artihmetic calculations.

In a complex dynamic system like development, all phenomena are interconnected. A major theoretical insight from dynamic systems is that the patterns of action, thinking, or development in the long term, result not from any single factor (plus some additional "noise"), but from the local and temporal confluence of many factors, operating on many time scales [19,271]. In order to study a phenomenon, for instance the development of abstract thinking, or of language, that phenomenon must be isolated for study. However, although it is possible to take a phenomenon out of a complex system, it is not possible to take the complex system out of the phenomenon. That is to say, while isolating a phenomenon for study, the model must explicitly account for the complex system properties from which the phenomenon is taken.

Time Scales Time scales are characteristic durations of processes or phenomena, corresponding with characteristic levels of organization. An example most directly pertaining to development is the time scale of developmental phenomena, spanning developmental events taking place over years or decades and limited by the duration of the human life span

These time scales are

- Scale of biological evolution: practically speaking, this time scale accounts for static constraints, i. e. constraints that do not change across the number of generations that developmental theories customarily address; exceptionally, rapid evolutionary changes can occur based on intergenerational links between developmental processes, such as certain food tolerances under high survival pressure (famine)
- Scale of socio-cultural historical development
  - The constraints and affordances at this time scale change relatively rapidly, it is virtually impossible to formulate an a-historical developmental psychology, i. e. a developmental psychology that takes the historical conditions as a constant; Baltes, Reese and Lippsit [13] made a distinction between normative-historical and normative age-related influences, for instance, the influence of war (normative historical) on a generation of adolescents (normative age-related); other examples are the influence of computers and the internet on young children growing up with them, in contrast with parents who did not use computers at that age.
  - Socio-cultural development can be conceived of as as a Complex Dynamic System in the following way. Its components are agents (many, of different ages, forming intergenerational networks), cultural tools, social organizations and channels of communication and interaction and so forth. The emergent phenomena emerging from the interactions between the agents are: the historical production of the life space, historical production of tools, products for human action, continuous innovation and elaboration of the human life space of human environment or Umwelt, which are emergent phenomena of mass interaction in the social networks of human culture and human society.
  - Scale of development across the human life span: this time scale encompasses changes describable only at the level of the human life span and involves processes such as development in the sense of increasing complexity, skill, and so forth, and processes of aging which refer to losses and negative changes, e. g. decrease of information-processing speed as a consequence of aging. These changes form the topic of the current article.
  - Scale of action and real-time experience: this time scale encompasses actions as goal-driven or intentional behaviors and require the duration characteristic of actions, lasting from seconds to minutes to hours; the dynamics are explained by means of theories of adaptive agents (see further).
  - Scale of underlying fast neurobiological processes: the time scale of rapid processes in the brain, nervous, motor and visceral system.

Theories of human development differ with respect to the time scale(s) they wish to emphasize. For instance, evolutionary developmental theory explains development from the perspective of innate properties that served the fitness of humans in evolutionary times; socio-cultural and Vygotskyan theory of development emphasized the cultural tools available to individuals and that depend on the historical evolution of their society and the actual socio-economic position they occupy, enabling them to use or not use those tools to various extents; Piaget's theory which emphasizes development as a partial life-span trajectory (from birth to about twelve years of age), dependent on processes taking place at the time scale of action.

Self-Organization, Emergence Self-organization is the spontaneous increase in order, complexity or structure, i. e. structure increases not because it is imposed or transferred by an external source, as in transmission (e. g. transmission of certain knowledge items from one person to another through communication). An example is increasing differentiation and integration in cognitive skills and performance across development that occurs spontaneously out of the interactions among all the components involved. It is not imposed on the human life span by the unfolding of a genetic blueprint or by transmission of knowledge and skill from a teacher to an apprentice. Genes, environment and the person's actions are all interdependent components, the interactions of which lead to a self-organization in terms of developmental levels [314,340].

A common theme among proponents of (complex) dynamic systems theory in development is that they view the developmental system as a self-organizing system, showing attractor states, non-linearity in its behavior, emergence and so forth [185,186,188,190,287,314,337,345,348].

Emergence is the spontaneous appearance or evolving of a new property of a system, in the form of a coherence or pattern on a global or macro level of organization. Emergence is related to the viewpoint of the observer (e. g. [76]) and implies a certain degree of surprise from the side of the observer of the system [121,122]. In the field of developmental psychology, emergence relates to the fundamental issue of whether the mature state of human skills, knowledge and so forth is an emergent phenomenon or not. Developmental viewpoints have tra-

ditionally emphasized the emergent nature and have focused on the appearance of (subjective) novelty in a system. A prime example of an emergentist view on development is Piaget's theory. Vygotsky's theory emphasized the emergence of novel forms or innovations on the time scale of cultural-historical evolution through the contributions of individuals and collectives of people. Vygotsky primarily defined ontogenesis as an appropriation by the individual of tools and symbols generated through cultural-historical innovations. The notion of a zone of proximal development may imply emergence, if viewed as an instance of a proper dynamic system [341,352]. Modern theories, such as nativism, assign emergence to the level of phylogeny, i. e. the evolution of the human species leading to heritable biological preconditions for development (as in the concept of core knowledge). Traditional theories of learning equally denied the primacy of emergence for development by emphasizing the role of transmission and appropriation, more particularly transmission of knowledge, skills etc. by instructors [348]. As such, the question of emergence is at the heart of developmental psychology. Notwithstanding the centrality of this question, development originally (see etymology of the term) implied the unfolding of intrinsic properties, i. e. the unfolding or uncovering of what is already there [330,345] whereas emergence implies the coming about of something truly new.

Development as an Increase in Complexity of the Developing System A developing system, for instance a child in a particular familial, cultural and historical context, is characterized by an increase in the system's complexity, often with an asymptotic level of complexity implied as the "final state" of the developmental trajectory. The description of development thus requires a descriptive framework or manifold specifying the space of complexity. Imagine such a space as a multi-dimensional space consisting of all the descriptive dimensions or features needed to specify a distinction between any possible developmental states or levels [330,331]. The simplest possible description entails a single developmental scale or "ruler" specifying the relevant developmental order. For instance, the complexity or structure of a child's thinking and problem solving is given by its position on an ordered scale of cognitive accomplishments, which are often inferred from a relevant content theory. An example is Fischer's theory of iterative embeddings of components, such as single representations of objects or properties and the relations between them, describing an orderly structure of increasingly complex levels [102,103]. A child's developmental level is assessed by letting the child perform a number of actions that map on the developmental scale at issue, which is mostly done in the form of a standardized task setting, i. e. a test.

A typical and enduring problem of developmental psychology concerns the relationship between the structure of the developmental scale with the structure of the child's "mind". The identification of one with the other – the structure of the test and the associations between test results on the one hand with the structure of the mind on the other hand – is a tacit but common stance for many (developmental) psychologists and is an example of an essentialist and primarily Aristotelian view on the nature of the human mind. It does not reckon with the fact that the relationships involved – those between child, observer, theory and observation instrument – are in themselves an example of a complex adaptive system, and not an instance of a straightforward measurement problem, where a property of an object (a child) is measured by a measurement operation that has no effect on the property being measured.

The expression of knowledge in an activity (nonverbal, verbal, symbolic) is a matter of particular stabilities and patternings of the actions (or expressions). For instance the infant's searching for a hidden object in the correct hiding place in the case of "object permanence", or the older child's verbal communications about perceived events in the case of causal understanding or Theory of Mind, reveal the knowledge in the form of certain stabilities of the pattern of reaching (it is not perturbed by a replacement of the hiding object for instance) or in the pattern of certain verbal explanations, which can take various concrete forms. Because the pattern is a temporal, self-sustaining pattern over a state space of possibilities (e. g. reaching and grasping possibilities; the space of words and grammatical relationships), the knowledge is said to be soft-assembled, i. e. existing for the moment and context in which it is actually expressed [314]. This model is very different from a standard mentalistic one, in which the mind entertains a symbolic or conceptual entity – such as the object concept or a Theory-of-Mind – which is then expressed in the form of or linked to overt activities, which borrow their meaning from the mental contents to which they are attached. In this sense, knowledge has no existence outside its expression in the form of real-time action. The stable patterns that express a particular form of knowledge (e. g. the object concept) emerge through coordinations of many components, part of which are "internal" or linked to the individual, part of which are linked to the world in which the action at issue takes place and are given through perception ([271]). Many of the internal components are non-cognitive in the classical sense, and include goals, concerns and emotions [19,303]. Through

development, the rules and components of the coordinations change, giving rise to actions, reasoning, and so forth that can be described in terms of formal structures that increase in complexity. Such formal descriptions (e. g. that children's thinking involves systems of relationships between components), however, do not refer to the underlying causes of the developing knowledge expressions and skills, but only to their abstract form. They are ways or frameworks for comparing knowledge and skills at various levels of development.

Development as Increasing Complexity Applied to Language Theory and Theory of Language Development As explained above, a theory of a domain of development – for instance language, cognition, … – serves as a model, eventually as an implicit model, of the state space in which the developmental trajectory is situated. Thus, in order to define what counts as development, the researcher must take a definition or description of the developmental domain and specify changes along developmental time as changes in the quality or quantity of the features that figure in the definition. However, a theory of, for instance, language, need not in itself contain the elements for a complete developmental state space description, i. e. a description of all the possible developmental states relevant to language development. Simply stated, a theory of language can contain a specific description of a state of development, i. e. a mature or ideal state, but be entirely underdetermined in terms of the possible paths that lead to this final state. For instance, assume that using passive constructions is a feature of mature language. Given that passive constructions form part of the description of mature language or language per se, the developmental route towards passive construction use is logically confined to a two-step process, namely no-passive-construction followed by passive-construction. This simple from-nothing-to-all switching process vastly underestimates the variety of observable developmental steps leading to passive constructions. That is, it falls short in specifying the potential developmental state space.

Transformational generative grammar, originating in Chomsky's work, made a major contribution to solving this problem by advancing a theory of language that logically entailed a theory of the possible developmental steps towards mature language [6]. By so doing, transformational generative grammar included a description of the developmental state space. Such a description should not be identified with a description of the developmental process per se, since that process can be any of the possible trajectories through the state space. However, if all states of the state space are conditionally ordered, such that for any state there is only one possible state it can emerge from, the state space description trivially becomes a description of the developmental trajectory, since it is the only one possible. A comparable situation occurs if we take Piaget's theory of cognition, which logically entailed a description of the developmental state space. The possible states are conditionally ordered, and only one trajectory is logically possible (see [330,332,333,334,335,336] also for discussion of developmental models based on Galperin's and Erikson's work). However, if the formal theory of a domain, e. g. linguistic theory, defines – or even simply constrains – the state space for developmental processes, the empirical and theoretical validity of developmental findings becomes conditional on the validity of the formal theory at issue. In short, the developmental findings answer the question of how a developing system reaches a particular developmental outcome, as defined by the domain-theory at issue. However, by studying development as such, it is possible to directly contribute to the formal domain theory itself. For instance, according to genetic epistemology, by studying cognitive development one obtains a better understanding of the nature of human knowledge in general, i. e. of what uniquely defines human knowledge [241].

In a similar vein, by addressing the logical problem of learnability of language as a human competence, Chomsky contributed to the formal definition and specification of that competence [62,63]. The earlier notion of transformational generative grammar has now been replaced by a linguistic theory entailed in the so-called minimalist program, which "… proposes that the computational system central to human language is a perfect solution to the task of relating sound and meaning" [64,181]. A major question is of course what linguistic theory has to say about meaning, how it is structured and what defines its developmental state space. An interesting development in this regard is the work relating catastrophe theory to semantics, thus defining semantics in dynamic systems terms [275,316,369,370]. Meanings are represented in terms of attractor states in morpho-dynamic fields, and such type of representation might be linked to dynamic field theories that were developed in the context of developmental research by Schöner and others [273,273,274,296]. If meaning can indeed be described as a morpho-dynamic field, language development amounts, at least in part, to the dynamic unfolding of this field. Formal theories of how such unfolding can take place can thus form a theoretical basis for a complex dynamic systems theory of language development. Relating meaning and sound, the solution of which is formally described by the minimalist program, leaves the question of why or for what purpose a person would want to relate meaning

and sound in a concrete situation. Transformational generative linguistics, which has dominated the field of linguistics in the second part of the 20th century, sees this question as related to the actual use of language, the "performance", that is not part of fundamental linguistic theory. This stance relates to the essentially anti-functionalist view of transformational generative linguistics. However, from an evolutionary and developmental point of view, language is the outcome of complex dynamic and adaptive processes, and it is hard to avoid the conclusion that this adaptive evolution has not fundamentally shaped language down to its deepest layers.

Simulation studies and mathematical models of iterative processes in language evolution and acquisition provide evidence that the formal structure of a language is shaped by the dynamics of language transmission and appropriation by individuals, and more particularly, by its use in social interaction [48,65,66,168,169,229,231,290].

An important feature of these models is the very close dynamic interaction they propose between learning, culture and biological evolution. Over the course of language evolution, these three components transform each other in a process generally known as co-evolution. In that sense, human biology is deeply transformed by human culture and vice versa. For instance, the biological pre-adaptation for language acquisition – in whatever form one wishes to specify it – is the result of a dynamic systems process occurring over the intergenerational (i. e. evolutionary and historical) and intra-generational (i. e. ontogenetic) timescales. In that sense, the dynamic systems approach can help explain – in principle – how language structure emerged through self-organization over the course of generations [165,228]. The evolutionary processes that have shaped language were modified by the fact that they had to pass through the highly specific constraints and opportunities of transmission and appropriation actions in individual agents. They have resulted in language being an essential aspect of the psychological life space of individuals and being appropriated in an extremely rapid and robust way, given the complexities of the task of language development. The question is how language functions in the psychological life space of individuals, and thus, under which constraints and opportunities language develops.

A classical theory in that regard is Vygotsky's, whose work is still of theoretical importance to the field of developmental psychology. Vygotsky saw language as a complex set of evolutionarily and historically developed tools, that individuals use to act with and solve problems [328,361]. Hence, language development can at least partly be understood in accordance with the dynamics of tool use and its development [200]. This view leads to the idea that language is a cognitive niche or a material scaffolding structure that the child and its environment construct during the developmental process, as an additional dimension and partition of the psychological life space (e. g. the name of an object as a feature of the object, relating the object through the linguistic relationships of the word; linguistic forms as objects of action in themselves, related in complex ways to other features of the complex world, an issue which relates to the so-called grounding problem [24,68,69,70,71,75,357,358,375]).

Development and the Embeddedness of the Observer As explained earlier (in the section on developmental level, order and structure), the assessment of a child's developmental level is not a simple measurement issue, of a determinate property (the developmental level) being tapped by an otherwise neutral measuring instrument (a test or observation). For instance, Elbers [92] showed that children bring specific expectations of answers to particular questioning situations and use the (non-)reactions of the test administrators as information, thus turning the alleged test or measurement situation into a social dynamics. For a young child, the test is an educational situation, and the child will react to the situation on the basis of the expected consensual frames [107]. From a measurement-theoretical point of view, this is not a trivial issue. The adult's intended act of measurement co-determines the measured content in a a direct and objective sense, and there is no measurement outside the context in which the adult's probing changes the probe. Exceptions are observations in natural, free contexts, e. g. observations of spontaneous behavior as in language developmental studies, which are observations of relatively unconstrained social interactions. However, the result of these observations should also not be seen as measurements of the "true" level of some developmental property, but as observations of the dynamics of social interactions, involving the dynamics of what children already know and are able to do in an environment that is adapted and reacts to the child's possibilities and, equally important, anticipates on the child's growing potentials.

Another fundamental issue, which is typical of complex human systems, is that in the field of human development, ontological and epistemological issues become entangled. The researcher's view on the cognitive development of a child to an adult, for instance, requires an ontology (a theory of being, in this case of the mental contents, knowledge etc. of the developing persons under study) of a process that ends with an epistemology, i. e. a theory of how humans know, including the researcher. The researcher's level of understanding of the cognitive develop-

ment of children towards adults is in itself the endpoint (roughly speaking) of his own, personal developmental trajectory.

For most developmental scholars, there is no problem with this entanglement, if any such entanglement is observed at all. For instance, a cognitive developmental theorist, working in the spirit of Piaget's stage theory and using his own mature, formal operational thinking as a tool for understanding the world, can fruitfully and without any internal inconsistency, study the emergence of formal-operational thought starting from its roots in the baby's non-operational, actional and sensorimotor way of understanding the world. However, the understanding of the developmental process will thus be determined by its endpoint, also because that is the tool with which this understanding is accomplished [330]. If we assume that development continues, also because scholars invest in making historical change processes come true and contribute to transforming systems of understanding, the representation of the process of development by developmental scholars will shift, as their own developmental levels (forms of scientific understanding) shift over the course of their lifetime, or over the course of historical time.

Dynamic Systems

A Definition of "Dynamic System" Dynamic systems theory is an approach to the description and explanation of *change*. A simple definition is "a means of describing how one state develops into another state over the course of time" [365], which can be expressed mathematically as

$$y_{t+1} = f(y_t) \tag{1}$$

expressing that the next state (at time $t+1$) is a function f of the preceding state, at time t. In a slightly different notation

$$\Delta y/\Delta t = f(y)\,. \tag{2}$$

The equation states that the change of a system, denoted by y, over some amount of time, denoted by Δt, is a function f of the state of y. The function f is also referred to as the evolution term or evolution "law". That is, it is important that f specifies some causal principle of change.

An important property of the current equation is that it represents recursive relationships. Thus, y_t leads to y_{t+1}, and according to the same principle, y_{t+1} generates y_{t+2} and so on.

A system can be described as a set of entities that are related to one another and that influence one another, and a state of the system is the set of properties of its components at any particular moment in time. The properties of the system are expressed in terms of dimensions or variables, for instance the variable y from the preceding equations. Dynamic systems can consist of any number of such variables. For instance, if y represents a child's current lexical knowledge, and z represents the child's knowledge of syntactic rules, the dynamic system consists of two dimensions, and the child's current developmental state is a point in this two-dimensional space. The space of developmental dimensions is the developmental state space, and development can then be defined as a trajectory across the developmental state space.

Properties of Dynamic Systems Applied to Development

Iterativeness The iterative or recursive nature of a dynamic system refers to the map or flow that the system instantiates, and which, in qualitative and metaphorical terms can be rephrased as "explaining after by before" [352]. The application to development is – apparently – trivial, in the sense that a developmental process is a process that transforms a current state to arrive at a new state, the "state" being any point or region in the developmental state or phase space, as defined earlier (or alternatively, the symbol string corresponding with a set of properties describing the current properties of a developing person in a co-developing environment). Examples in classical developmental theories are Piaget's notion of assimilation, implying that the representation of given information is a function of the current forms of understanding and representing, or Vygotsky's notion of the zone of proximal development, in which the next attainable level of development is a function of the level already consolidated.

Dynamic Rule/Principle/Function The dynamic rule or function describes the way a current state of the system is mapped onto, or transformed into, another state at some later time. In principle, this dynamic rule corresponds with the basic mechanism of development that a developmental theory entails. In principle, this basic mechanism is any causal mechanism that operates on the current developmental state and that brings about a particular change of that state, including a 0-change, which occurs if the developmental state has reached stability.

Classical developmental theories were usually explicit about the mechanisms operating on developmental states. For instance, for Piaget, the mechanism is one of adaptation, with constituent mechanisms of assimilation and accommodation. The working of these mechanisms is ex-

plicitly determined by the properties of the current developmental state. For instance, the child assimilates the information it obtains from its actions to the cognitive structures it currently maintains. The same holds true for accommodation, which is the driving force behind developmental change. A similar logic applies to the fundamental Vygotskyan notion of the zone of proximal development, which implies that a more competent person adapts, in terms of help given, to the current level of development of an apprentice, and by giving this level-adapted help, stimulates a process of interiorization in the apprentice that leads to a new and higher developmental level. Other examples are Werner's notions of differentiation and integration, which are mechanisms explicitly operating on the cognitive, behavioral and emotional structures that are present in a child (or in a child-environment system).

Providing such models of mechanisms boils down to specifying an implicit function for development, i. e. given a state so-and-so, application of the mechanisms or rules will result in a different state, and through iteration or recurrence, to a developmental process. In principle, the mechanism or developmental function itself implies not only the present developmental state of the system, but also any influences – coupled to or indepenent of the state – that are incorporated by the mechanism and that moderate development. For instance, a mechanism explaining the growth of a child's lexicon not only operates on the current state of the lexicon, but also implies external influences, such as the environment's lexicon, teaching activities and so forth.

The majority of modern developmental studies have replaced these implicit developmental functions by explicit ones, which are based on samples of independent subjects. For instance, the time dimension, which is a fundamental kinematical variable for specifying change, can be used in an explicit developmental function, assigning a developmental level (of whatever kind is required) to a particular value of developmental time (most particularly age, but also duration of experience, for instance). Although these explicit models claim to achieve generality (or as it is often called, generalization), they achieve this result at a devastating cost. The cost is that they bypass the actual process, and are in fact completely ignorant as to which causal mechanisms explain actual developmental processes, which are processes applying to concrete, individual systems. For instance, lexical growth is a process occurring with a particular child in a particular environment, which is to a considerable extent dependent on the child's actions. For instance, the linguistic environment tends to adapt itself to the language use and understanding of the child. An explicit model provides a model of the lexicon as a function of a given time (age), other variables such as maternal language proficiency and an additional stochastic influence. The model is achieved by averaging over many individuals, and thus risks to lose all information about actual processes (unless the processes are virtually uniform over all individuals, which is rarely, if ever, the case). This problem, which refers to the fact that such explicit models of associations between independent and dependent variables are not in any way logically related to implicit models describing mechanisms operating on developmental states, has recently come under the attention of statisticians and methodologists working on individual developmental trajectories [136,216] (e. g. Molenaar, Hamaker, ...). The reason why such implicit functions are central to developmental science is that they attempt to specify the causal processes that operate in real time, and thus provide models or prescriptions of actions by practitioners.

Attractor An attractor is a set of points in the phase or state space towards which a system will evolve, given its dynamic function, if it is in the basin of attraction.

The application of the notion of attractor to a developmental system implies that the system will asymptotically evolve towards a particular state if it is under a particular set of conditions (the basin of attraction). More precisely, if the system is in an attractor state and gets perturbed, it will spontaneously return to the attractor state, unless effectively counteracted by some external force. For instance, a particular child, in a particular development, will tend to show certain stabilities or stable patterns in its behavior. The criterion of return after perturbation, under certain limits concerning the strengths of the perturbation, is an important criterion for distinguishing attractor states from accidental states for which the system has no particular "preference" (if any such states exist). Many developmentally relevant attractor states are likely to be rather idiosyncratic, i. e. dependent on individual and local circumstances. Other states may be relatively general and predictable on the basis of broad criteria such as age. An example of the latter type of state are the classical developmental states or stages, for instance the states distinguished by Piaget, by neo-Piagetian scholars, stages distinguished in the progression towards mature language and so forth. An example of a clinical application of the concept of attractor is the notion of resilience, which refers to a child's capacity to spontaneously move back to a healthy psychological state after having experienced highly stressful or adverse experiences. Although resilience can easily be defined as a personality property of the child or as an individual capacity, it is probably more realistic to view

it as a an attractor state of the developmental system in which the child is embedded [209]. A comparable example is the emergence and self-sustainment of highly problematic teaching-learning patterns in children with developmental disorders such as ADHD [353].

Developmental State Space and State Space Grids The use of dynamic systems theory to development applies a fundamentally geometric way of thinking to the study, description and explanation of development. Developmentalists are used to thinking in terms of variables that they identify with psychologically real properties of the mind, which implies that their frame of reference is a model of the mind or the brain, i. e. a model with a topology that is similar to the topology of the mind or the brain (since the topology of the mind is difficult to imagine, and the topology of the brain is at least seemingly given by modern brain imaging studies, developmental psychologists are increasingly turning towards the brain as a model; however, see [50] for a critical discussion). From a dynamic systems point of view, however, a developing system is a geometric manifold or space, consisting of all the dimensions used to describe the system (and this number can eventually be very small), including the evolution laws or rules that specify the change of positions (developmental states) in this space. There is no implicit reference to topological similarity with the mind or the brain. The use of such geometric, state space descriptions can free the researcher from "unsolicited" ontological claims, i. e. implicit claims about the nature and composition of the human mind that are relatively standard in mainstream developmental psychological investigations [350].

The notion of state spaces, in particular categorical state spaces, has been promoted by several researchers, mainly working in the field of social interaction and social development [127,128,129,148,149,191,192].

Static Versus Dynamic Models in Development Developmental psychology, has almost exclusively focused on *static* models and has implicitly assumed that change, for instance developmental change in an individual, could be approximated by stretching static relationships over the time axis [352]. A characteristic expression of a static relationship takes the form

$$y_i = f(x_i) \tag{3}$$

with y a dependent variable and x an independent variable, which, for any possible value x_i generates a corresponding value for the dependent variable y. The variable x can also be time, but the use of time as such does not turn the model into a dynamic model.

A difficulty arising with this definition of a static model is that any dynamic model that is expressed as a function of time

$$y_i = f(t_i) \tag{4}$$

must strictly speaking be qualified as a static model [300]. Hence, we should confine static models to those where the x-variable is not time. However, in Sect. "A Definition of 'Dynamic System'" it was claimed that the f in the dynamic equation must specify some causal principle of change, with the implicit assumption that this causal principle, however general, is theoretically justified and based on what we know about how things change. Hence, a model of the form of Eq. (4), $y_i = f(t_i)$, that applies to an empirical sample (e. g. a regression model of time applying to a cross-sectional sample, or a sample with two or a few consecutive measurements), can easily be transformed into a model of change by taking the first derivative of the model. By doing so, however, one does not automatically arrive at a meaningful dynamic model, since the function term f featuring in a descriptive time-serial model is not necessarily *descriptively adequate*. A dynamic model can be characterized as developmentally descriptively adequate if the mechanism implied in the dynamic model (1) corresponds with a developmentally plausible mechanism, (2) in principle applies to the whole developmental time scale of the developmental phenomenon in question [193].

A static system describes a particular value of the variable of interest as a function of the value of another variable (or set of such variables). For instance, for any possible age, or for any level of the mother's lexical knowledge, or for a combination of age and maternal lexicon, the static system or model will generate a predicted or expected size of the lexicon, without any reference to recursiveness.

This distinction between static and dynamic type models has considerable consequences [152,300]. Whereas a dynamic model recursively generates a time series (a state and the next state and the next ...), a static model generates a sample or population of individuals that are in principle independent of one another (an individual with age i and lexicon i, an individual with age j and lexicon j, and so forth). Statements about associations between variables across populations do not necessarily apply to the mechanisms that apply to change in the individuals in the population. However, the behavioral sciences, including developmental psychology, often implicitly take a relationship between variables that holds across a sample as a representation of some dynamic rule or principle (also known as the homology or ergodicity error) [136,216,224]. For example, a study showed that early math skills in 5

to 6 year olds have the greatest predictive power for later school achievement, whereas socio-emotional behaviors, on the other hand, had little or no predictive power, irrespective of gender and socioeconomic background [90]. From such findings, it is easy to infer that increasing early math achievement, e. g. through preschool teaching programs, will thus lead to better school achievement at a later age, implying also that attempts to increase socio-emotional skills should be reduced since they do not relate to academic achievement. However, there exists no logical or direct relationship between the static relationship (how is it associated across a population) and the dynamic relationship (how can something be increased or decreased in individuals).

Adaptation and Adaptive Agent Systems

Adaptation is the process of adapting something to something else, usually in the context of an organism adapting to its environment. An important question concerns the relationship between adaptation and development. There is no doubt that adaptation and complex adaptive systems play a major role in the process of development (for further discussion, see Sects. "Theory Of Complex Adaptive Systems (CAS): Developmental Agent Models" and "Theory Of Complex Adaptive Systems (CAS): Epigenetic Robotics"). A classical theory such as Piaget's conceived of development as a process of adaptation, which shows a clear pattern of increasing complexity. However, adaptation, especially in the sense of organism-environment adaptation, need not be a process of increasing complexity, adaptation can also mean loss of specialization, decreasing complexity etc., if the latter is better adapted to the organism's current environment. In developmental processes, there are also processes of loss of knowledge and of complexity, depending on changes in the environmental circumstances (e. g. language loss [176,234]). An encompassing theory might claim that an adaptive process as it applies to a growing organism (literally as in embryogenesis, metaphorically as in cognitive or language development), must by necessity show an increase in complexity (and size, etc.), given the constraints on reproduction (reproductive activities produce an offspring that is less complex (of smaller size etc.) than the progenitor).

The theory of complex systems refers to complex adaptive systems, which are collections of interacting components (agents) that adapt to each other. A developmental system, e. g. a child in his or her environment, can be conceived of as a complex adaptive system, with the agents or components adapting to each other. Examples of mutual adaptations are given in dynamic reinterpretations of Vygotsky's theory [341,352] or in the theory of transactional development [263,265].

An adaptive system is not necessarily a goal-driven or teleological system. Human beings and organisms in general, however, are complex adaptive systems that are goal-driven or teleological, and in addition to adapting to their environment, they also wish to control their environment [77,167]. Developmental theory that explains the mutuality between the long-term time scale of development and the short-term time scale of action, thus needs a theory of adaptive, goal-driven or concern-driven agents in order to explain the level of action, and the developmental level of changes in goals and concerns [302,303] (for further explanation see Sects "Theory Of Complex Adaptive Systems (CAS): Developmental Agent Models" and "Theory Of Complex Adaptive Systems (CAS): Epigenetic Robotics").

Main Theories in the Field and Short Historical Background

Founding Historical Theories

A brief look into any arbitrary collection of handbooks on developmental psychology illustrates the field's historical concern with the question of whether development implies the unfolding of what is already given at birth (which refers to the original meaning of development as unwrapping) or whether development implies a start from zero and a process of appropriating of whatever is necessary to become a mature person. The main figures embodying these standpoints are John Locke (1632–1704) and Jean-Jacques Rousseau (1712–1778) respectively. Although it is no longer stated in this naïve form, theorists and researchers struggle with this issue even at the present day (see for instance the discussion on gene-environment relationships; [261]). Biologically-inspired theories – seeing development as the unfolding of a biologically given program – are associated with the work of Charles Darwin (1809–1882) and G. Stanley Hall (1844–1824). Historically, however, developmental psychology is based on the confluence of many theoretical strands.

Main Theoretical Viewpoints

Biologically Inspired Explanations The notion that the important components or aspects of the human condition in its mature form are in fact innate and not appropriated thanks to external influences, or constructed on the basis of one's own action, has received a major impetus by the work of Noam Chomsky in generative linguistics. Chomsky argued that language – which is obvi-

ously learned from the input received from the environment, in that no child will for instance learn French if confronted with a Dutch language environment and input – cannot be learned or acquired without an innate language acquisition device, which defines the major properties or degrees of freedom, of human language (see the section on Language theory and theory of language development nativism and modern nativism for further discussion [62,63]). The major argument was that language is in fact underdetermined by the input if the learning device has no preset clues about what the input means or implies in terms of structural relationships among identifiable components. A similar line of thought is followed by proponents of *core knowledge theory*, such as Elizabeth Spelke, who claim that the major components of human knowledge about the world – such as the notion of space, number, causality and so forth – must be innately given (see [295] for an overview).

Evolutionary developmental psychology [32,235] is a theory that applies Darwinian principles of evolutionary adaptation to explain the evolutionary emergence of epigenetic programs that evolved under specific selection pressures [33,96,115,194,195]. Examples include early fantasy play, parental investment and cognitive development. Since the epigenetic programs evolved under historical environmental conditions that are no longer present in contemporary environments, mismatches between such programs and the requirements of contemporary life may lead to perturbed development. Evolutionary developmental psychology can be seen as an offspring of ethological theory, a biological theory that tries to understand the adaptive functions of behavior of an organism. Ethological theory, primarily through the work of Konrad Lorenz, has given rise to theories about critical periods, i. e. specific ages at which the development of a particular category of action or skill, such as language, is particularly stimulated and beyond which that development cannot take place. Modern theories of development tend to speak about sensitive periods or windows of opportunity, and eschew the notion of critical periods and the impossibility of development if the period, for some reason or another, is missed [8,49,159,319].

Developmental behavioral genetics attempts to explain development and particular developmental trajectories on the basis of the person's specific genetic endowment [244,245]. Recently, major progress has been made in the study of the effects of "generalist genes" on development [246,275,276]. In the context of learning disabilities, Plomin and Kovas describe "generalist genes" as genes that affect not only the disability but also the normal variation in the behavior at issue (e. g. reading), that affect all aspects of the disability and related disabilities (or normal behavior) and not just a particular aspect.

The influence of genes is not unidirectional: genes and environments are to a considerable extent linked with transactional relationships. Specific genes act so as to make the person more sensitive to or selective to particular environments, whereas environments have an influence on the activation of particular genes or moderate the effect of genetic influence [5,261].

Psycho-dynamically Inspired Theories Psychodynamical theories originated in the work of Sigmund Freud (1856–1939). According to Freud, human behavior and action are determined by energetic forces resulting from basic human drives, which, for Freud, were primarily sexual in origin. The actions required for fulfilling these drives conflict with the exigencies of reality, and actions serve to resolve this perennial conflict, leading to particular psychological structures (such as the unconscious, the Id or superego, etc.). The principle of drive-determined conflicting actions operates from birth on, and leads to a series of psycho-social stages (the oral, anal and genital phases, from birth to about 7 years). Although Freud can be criticized for his narrow view on the importance of sexual drives for the explanation of human behavior and development, his theory is one of the few that actually tries to understand human action and development from the goals and concerns that human beings try to accomplish, and from the problems they encounter in doing so.

In modern developmental psychology, psychodynamic theories mainly survive through the work of Erik Erikson (1902–1994) who emphasized the social, cultural and environmental aspects of the conflicts between the person's drives and the challenges of the environment. Erikson saw the human life cycle as a sequence of basic conflicts, arising out of the demands that culture and society make on the growing individual and that depend on that individual's psychological and biological maturation. These conflicts, for instance the conflict between identity versus confusion which is typical of adolescence form a conditional sequence, in that the solution of an earlier conflict is a condition for the way in which later conflicts will be manifested and solved [115].

Socio-culturally Inspired Theories These theories find their origin in the work of Lev Semenovich Vygotsky (1896–1934) who developed his theories during the early stages of the development of the Soviet state, which may help account for the particular properties of this theory [251,328,359,360,366]. For Vygotsky, development is a process that builds upon the biological predispositions

given by biological evolution and that consists of an appropriation of cultural tools for action, including mental action. This appropriation of cultural tools – skills relating to tools in the literal sense as well as symbolic tools such as language and historically developed concepts – leads to the involvement of the person in his culture and is a requirement for his ability to contribute to the further development of his culture and society. Vygotsky emphasized the intertwining of processes on distinct time-scales, such as the historical time scale of societal evolution, the developmental time scale of the human life span and the short-term time scale of human action and tool use. Vygotsky's primary developmental mechanism consists of a combination of principles. One is that of interiorization, which states that children can interiorize or appropriate the actions and skills they first perform with the help of more competent others, such as parents, teachers or more competent peers. The second is that of the zone of proximal development, which refers to the distance between what a child can accomplish on his own and what it can accomplish with the help of others, such that interiorization is likely to take place and thus, development is further advanced. Development of an individual child takes place in direct and intensive interactions and transactions with other people, and is a deeply socially embedded and social process.

To Vygotsky, development of higher levels of thinking, including abstract thinking, is made possible by the appropriation of language-grounded concepts, such as time or causality, or names for numerical digits, and by the availability and use of physical symbolic systems such as writing.

The principle of social embeddedness in the culture has been used by various authors who have extended Vygotsky's work (examples are [255,324,367]. The theory is now more widely known as the socio-cultural theory of development, thus emphasizing the two main forces acting on and shaping human development.

Because of the emphasis on the mechanism of development and the direct causes of change, Vygotsky's major developmental principles lend themselves relatively easily to assimilation in a dynamic systems or agent-based framework [300,340,343,344].

Piaget and Cognitive Approaches to Development For Jean Piaget (1896–1980), cognition is a particular biological adaptation that allows the human organism to control and predict the environment through understanding and mental models. However, being a biological adaptation, it is not innately given but must develop, and it will do so on the basis of primarily biological mechanisms that operate on the child's understanding of the world. These mechanisms are summarized by the term adaptation and involve processes of assimilation and processes of accommodation. It is through these processes that operate under the form of human activities that the child actually constructs his understanding of the world, in a series of stage-like steps, leading to a level of understanding that is self-sustaining under its interaction with the environment. Assimilation involves the transformation of information to the form of the cognitive structures already present, whereas accommodation refers to changes in the cognitive structures to accommodate the information given. Assimilation and accommodation form a dialectical pair of forces, allowing the child to move beyond the direct sensory and motor givenness of experiences, and to come to abstraction. Cognitive structures are interrelated structures of schemes, concepts, skills and so forth, which, through their interrelationships are self-sustaining, at least for a given period of time, after which the working of the adaptive mechanism transforms them in new structures. These sequential structures correspond with Piaget's main stages, the sensorimotor, pre-operational, concrete-operational and formal operational stage. Cognitive structures are constructed, and are thus characterized by structural properties that are not or cannot be inductively inferred from the information given. A core example of such a structural property is the property of reversibility, which develops around the age of about six years and which marks the transition from pre-operational to operational structures. Reversibility is the property that assigns a reversible operation to any operation the cognitive system may have or develop, and it does so by implication. With reversibility, cognitive systems obtain properties of mathematical groups of operations, which greatly increases their power, enabling them, among others, to take their own operations as topic of reflection.

As with Vygotsky, Piaget was mainly interested in the nature of the developmental mechanism and considerably less so in the actual trajectories of development (although the textbook-representation of Piaget focuses mainly on his theory of stages and only superficially on the developmental mechanisms; the reason for this skewed focus is that modern developmental psychology is more a collection of facts about age differences in phenomena and associations of variables within samples than a science of development that primarily addresses the developmental mechanisms).

Piaget's theory was not so much inspired by the wish to learn about the developmental trajectories followed by children than by his interest in the nature of human

thought and understanding, which he tried to capture by studying their development. Piaget was first of all a genetic epistemologist, interested in the nature of human knowledge. For Piaget, developmental science is an intellectually reflective science, i. e. a reflection on the nature of the thinking that makes that reflection possible. This interest puts Piaget in the ranks of complexity theorists who view complex systems as systems in which the observer and knower of the system himself is embedded [304].

Because of his focus on the mechanism of development, that is on the process of change itself, Piaget's basic theoretical notions lend themselves naturally to dynamic systems modeling [343,344]. Piaget's notion of development as construction of cognitive systems is closely related to dynamic systems notions of self-organization. Given the constraints and degrees of freedom present in the organism and in the environment, cognitive systems corresponding to the major developmental stages are self-organizing structures that develop towards equilibrium, i. e. they are self-sustaining, until, through the accumulation of experiences and most notably cognitive conflicts, a critical point is reached at which the structure tends to change rapidly towards another equilibrium, through a cascading process of inter-related events of change.

Neo-Piagetian theorists started from the major assumptions of Piaget regarding structural relationships in cognitive systems and the constructive nature of development, and added principles from modern cognitive science, such as biological maturation and brain development, limited working memory and other constraints on information processing [58,102]. Neo-Piagetan theorists have also employed dynamic systems modeling to explain processes of cognitive development, including stepwise growth and temporary regressions preceding developmental accelerations [103].

In addition to Piaget's theory, the theory of information processing has made an important contribution to developmental science. Information processing theory and its recent offspring in cognitive science and neurocognitive science starts with a basic model of human information processing, containing components such as input and effector components, short-and long-term memories and so forth, and studies development as the changes in those components. Changes concern quantitative changes in the operation of the components, for instance processing speed or size of working memory, but also qualitative, i. e. content-related changes, which amount to changes in the information-processing rules [171,210,279]. Information processing has been dynamically modeled by simulation architectures such as ACT-R (adaptive control of thought-rational [3,4,161]).

Learning-Theoretically Inspired Approaches Learning theory is concerned with how experiences shape future behavior and focuses on the addition of new knowledge and skills to what is already present. Classical learning theory remains close to the observable properties of behavior, e. g. under which environmental conditions particular behaviors occur. Learning and behavior are seen as being under the control of the environment. For the explanation of development, two main principles of learning are important. The first is the principle of contiguity and stems from the field of *respondent (or classical) conditioning*, which is mainly associated with the pioneering works of I.P. Pavlov (1849–1936). A stimulus that is contiguous (immediately precedes and overlaps) with a stimulus that evolves a particular response, will automatically obtain the response-retrieving properties of the latter. The second is the principle of functional or operant learning and is mainly associated with the work of B.F. Skinner (1904–1990). Behaviors have different consequences, and if the consequence of a behavior is positively evaluated by the organism, its future frequency will increase. Under the influence of the particular context, every behavior shows a spontaneous variation, and this variation allows for selective consequences, which may alter the frequency of the behavioral variant at issue, which on its turn will also vary spontaneously and be selectively reinforced. Through this principle of *operant conditioning*, behaviors can be shaped towards entirely new forms by applying reinforcements to each variant that comes closer to the goal behavior than other variants [283,285]. The principle is very similar to "survival-of-the-fittest" principles acting in evolutionary biology. If developmental processes must be seen as a result of many such learning processes over the long-term, it is likely that such processes will interfere, for instance in terms of contexts and reinforcements, and thus that complexity principles, involving networks of many interacting components, will apply. The question is to what extent non-linearities, attractor states and phase shifts may arise under these conditions, although principles of operant learning can be very easily transformed in a dynamic systems format [338]. The interaction of many levels of reinforcement is addressed in learning theory with relation to the matching law [142,143]. The *Matching Law* holds that the proportional distribution of behaviors (e. g. how much on-task versus off-task behavior in a child during a math lesson) will evolve towards a value that reflects the reinforcements provided by these behaviors (e. g. how pleasurable or goal-effective, on average, is on-task versus off-task behavior). The Matching Law explains the emergence of what in dynamic systems terms is called a specific attractor state and can be used to explain developmental

changes (see [85,301] for an application in the field of social interaction; [30,278,286] for an application in teaching-learning processes and [38,223] for applications with developmental disorders).

Finally *social learning theory*, which originated with the work of Bandura, shows that children learn by imitating other people. Imitation depends for instance on the effect that the imitated behavior has for other people, on the social power and status of the person imitated and on beliefs on whether the imitating person will be able to accomplish the imitation or not (self-efficacy) [1,14,16,52,124,131,132,260]. Imitation tends to be a process that is far more complex than mere mimicry, and even in very young children extends to relatively abstract properties of the imitated behaviors, such as the intentions of the person who is imitated [53,57,320]. Thus, even with a seemingly simple act as imitation, one needs to conceive of the imitating person as a complex system, operating on many levels at the same time. The principle of modeling and imitation, which is maybe better referred to as behavioral contagion, can be fruitfully used in dynamic systems models of social behavior in children, in a developmental context [302,303].

Systems and Dynamically Inspired Theories Developmental systems theory, as it is applied to developmental psychology, explains development as a transactional process, implying transactions (that is, mutual transformations) of individuals and their contexts or life spaces, or of biological and environmental influences [108,184,264]. The theory has its roots in developmental systems theory as a specific approach to evolutionary biology [125,128,194,195], and the theory of nonlinear, complex dynamic systems, which is the focus of the current chapter.

Dynamic systems theory of development is a general approach, relying on the basic definition of a dynamic system as a system explaining the change of a variable, or why a next state follows from a preceding state, for "state" defined as any value on one or many variables describing the developmental state space (see [265,348,352,365] for general introductions in the field of development). In developmental psychology, the term "dynamic systems theory" has a somewhat confusing meaning. It refers to a theory of development based on processes of physically and socially embedded and embodied action, which originates from the work of Thelen and Smith [314]. Another approach is inspired by ecological models from biology and applies principles of resource-dependent change and mutual relationships of competition and support between components of the developing system, and is based on the work of van Geert [337,339]. Finally, dynamic systems theory refers to a variety of applications of general concepts from dynamic systems theory, such as self-organization, attractors, state space, etc. to developmental phenomena, for instance in the work of Lewis, Granic, Hollenstein, Fogel, Dishion and others (these theories will be discussed later).

Content-Driven Approaches A quick glance through the scholarly journals in developmental psychology shows that the field as a whole is theory-poor, in the sense that well-established, non-trivial and fundamental theories of developmental change that lay the groundwork for our understanding of developmental processes are virtually absent from most of the empirical work. Developmental psychology is primarily a science of relatively isolated fields or themes of development. Examples are attachment, theory-of-mind, aggression and bullying, and so forth. The standard research approach is to measure a collection of variables over samples of children of various ages, describe the change of the dependent variable based on averages for the distinct age groups and show how the dependent variable is associated with a host of independent variables (for instance, changes in average levels of theory-of-mind scores or reactions in experimental situations, which are associated with measures on language, intelligence, socio-economic status and so forth.

Conclusion

In its present state the field of developmental psychology is a scattered collection of approaches and fields that do not unite into a common framework or emerge from a common framework explaining the fundamental mechanisms of developmental change. The field lacks a generally accepted notion and theory of development in general, except for relatively trivial collections of principles taken from various (historical) theoretical approaches, such as Piaget's theory, learning theory, knowledge about biological underpinnings of development and so forth, that form the inevitable introductory part of the field's basic textbooks. Theoretical unification may take place by liberally employing principles from dynamic systems theory, theory of complex systems and of complex adaptive systems.

The Study of Individual Development in Light of Complexity Theory

Theory of Development and Complex Dynamic Systems Models

Development and the Theory of Embodied Action According to Thelen and Smith [314], current psycholog-

ical theories tend to invoke "ghostly" things to explain behavior, namely internal representations and concepts. The representationalist stance, also known as the computational-representational understanding of mind, or information-processing theory, states that thinking can best be understood in terms of representational structures in the mind and of computational procedures that operate on these structures [311]. That is, in order to explain thinking you need internal entities. For instance, if a baby watches an object being hidden by an adult, and then watches how the object is moved to another hiding place, the baby will still look in the first hiding place in an attempt to retrieve the hidden object. This so-called A-not-B error is traditionally explained by the absence of a fully developed object concept, i. e. an internal representation of the object concept [287,292]. According to Thelen and Smith, invoking the notion of "concept" to explain a particular behavior or action related to that concept, is a categorical error, as if one explains the color of the red traffic light by the workings of its inherent redness. It is this categorical mistake that Thelen and others seek to repair by explaining phenomena such as the A-not-B-error by a theory of situated action, in which an embodied subject (not an epistemic subject, such as in Piaget's case) acts with the help of and under the constraints of a physical world that includes the external environment and the physical properties of the body [292].

In essence, cognition, thinking and action are explained as dynamic patterns unfolding from the continuous, "here-and-now" interaction between the person and the immediate environment. A particularly clear description, in the context of cognition and intelligence comes from Linda Smith [287]:

> The embodiment hypothesis is the idea that intelligence emerges in the interaction of an organism with an environment and as a result of sensory-motor activity. The continual coupling of cognition to the world through the body both adapts cognition to the idiosyncrasies of the here and now, makes it relevant, and provides the mechanism for developmental change. (p. 205)

The dynamic system at issue is the continuous coupling between the organism and its environment. This system shows a short-term time-evolution that takes the form of intelligent action, which changes the body and the brain through processes of learning and adaptation, thus giving rise to a long-term evolution we call "development".

In order to explain the short-term dynamics of action, thinking and knowledge in concrete contexts and the long-term dynamics of development, including the interdependencies between these two time scales, Thelen and Smith invoke the notion of dynamic fields, which will be explained in a later section.

Development and Resource-Dependent Competition-Support Systems As for any complex system, a developmental system can be viewed as a collection of elements or components. These components are related through functional relationships, implying that one component can change another, and vice versa. The system is embedded in an environment with which it is also functionally related (it can affect the environment and can be affected by it). The notions of "system" or "environment" refer to distinctions arbitrarily made by the describer of the system. For instance, a developmental system can be defined as a single variable or component, for instance a particular child's social cognition, or a child's lexical knowledge. This single-variable system then defines an environment consisting of any other component that functionally affects the component at issue. For instance, for the social-cognition system of a child, the child's emotional repertoire, intelligence, linguistic knowledge all form part of that system's environment, in addition to components that are not internal to the child, e. g. The child's peers, family, culture etc. In this sense, the notion of environment does not refer to "environment" in the usual sense, namely the child's current life space, family environment etc. (although such components do belong to the system-dynamic definition of environment). A system consisting of two variables, for instance a child's lexical knowledge as one component and the language addressed to the child by the mother as the second component, is embedded in an environment that consists of all sorts of internal and external phenomena that are related to the two variables at issue (such as the cognitive systems of the child and of the mother, their emotions, but also the material and cultural artifacts of their homes, other family members, etc.). Such developmental systems can be easily extended towards any number of explicitly defined and related components, which then define a complementary system set, which is the environment of the system in the abstract sense just introduced.

If the components through which a developmental system is defined can be described as (roughly) numerical variables, the system can be treated as and modeled as a dynamic ecological system. That is, if the system's components can be described by means of variables that specify the level of the component, the system dynamics amount to relationships affecting those levels. For instance, a child's lexical knowledge can be described as a particular number of words actually known or under-

stood by the child. The child's social cognition can be described by means of a variable that distinguishes possible levels of development of social cognition (see Fischer's notion of developmental rulers, capturing the same idea [105,257,350]). This representation by means of numerical variables forms part of the abstract description of the system, and does not necessarily need to correspond with a homologous empirical measure of the variable in question. It suffices that the abstract numerical dimension and the empirical measure of the variable at issue are sufficiently analogous to warrant an eventual empirical test of the resulting dynamic model describing the system. The descriptive use of abstract numerical variables to describe components of the system also makes no implicit ontological claims with regard to the nature of those variables. For instance, describing a child's social cognitive knowledge as an underlying numerical variable does not entail any claim about that knowledge being localizable (e. g. in the child's brain), as internal or symbolic contents. The variable in question can equally well refer to a distributed, soft-assembled property of a child's actions that depends on causal loops between the child and its environment (see [350] for discussion).

The description of the developmental system as a dynamic ecological system makes use of the following general assumptions (for thorough discussions of these principles [334,337,345]). First, development is defined as the growth or increase in level of more developmentally advanced or complex variables and the decline or decrease of less developmentally advanced variables. The growth of a variable (e. g. a child's lexicon, a child's level of social cognitive understanding, and so on) is an auto-catalytic process, in that it depends on the level already attained. Thus, if l is the current level of some developmental variable (e. g. a child's level of lexical knowledge, of social cognitive skill, etc.), the growth or change of l over some time duration is expressed as

$$\Delta l = rl$$

for r any rate or change parameter. In an ecological system, of which developmental systems are examples, growth or change depends on the availability of resources, which are limited. For instance, a resource factor for lexical growth is the language spoken in the environment, but also the child's auditory system that helps it pick up acoustic signals that form the physical basis of spoken and heard language. Given that resources are limited, the growth parameter r approaches a zero limit as l approaches the level that is sustained by the available resources (a simple example is the resource "lexicon of the environment" which limits the number of lexical items that can be learned by a child living in that environment). The effect of the limited resources is expressed mathematically as follows

$$\Delta l = r(1 - l/K)l$$

for K the limit level of l under the given resource conditions. This equation thus corresponds to the well-known and basic logistic or Verhulst equation (see [337] for further explanation). If applied to knowledge-related variables, it basically states that the growth of knowledge depends on what one already knows, and on what one does not know yet, given what there is to know in the current context of the particular person in a particular environment. In the simplest possible case of a single-dimensional developmental system, the resource component corresponds with the system's environment.

For a system description to be really explanatory interesting, it should contain various coupled components, corresponding with the major dimensions of a particular developmental system. For instance, a model of the child's entire cognitive system should consist of variables referring to the major components of the cognitive system, e. g. language, conceptual knowledge, emotions and appraisal components and so forth. Another example concerns a model of an educational system, which should at least consist of a variable that changes as a result of environmental stimulation and interaction on the one hand, and a variable that describes the educational influence on the other hand. These two variables should be coupled, since it is relatively obvious that, as the child's growth (at least partially) depends on educational influences, the educational influences themselves will depend on the developmental or growth level that the child has already attained (one is not going to teach higher algebra to a child who has not even mastered elementary arithmetics, for instance). Coupled variables, each of them dependent on background resource factors, form the heart of a dynamic ecological system. To explain the principle of coupled variables, let us take an example from the field of social and personality development during adolescence (the example is taken from [193]).

The adolescent's main conflicts with the parents are related to the adolescent's wish for growing autonomy and the parents' reluctance to grant this autonomy too easily or too rapidly. The tendency to increase autonomy depends on resources such as physical maturation and cognitive and social skills, which are clearly increasing during the stage of adolescence. However, the tendency to increase autonomy is coupled to a complementary variable, which is the adolescent's connectedness with the parents. We shall represent autonomy and connectedness as two coupled numerical variables, which can thus increase or

decrease in level. In psychologically healthy families, the level (including the quality) of the connectedness between parents and adolescent can have a positive effect on the adolescent's wish for autonomy, i. e. increase the growth of autonomy. The other way round, however, autonomy of the adolescent – and later adult – will have a positive effect on the quality of the connectedness and thus – given the principle that differences in quality are implicitly quantified – on the level of the connectedness (it is easy to see that if parents cannot accept the increasing autonomy of their children growing to adulthood, the level of connectedness that exists between parents and children is in fact going back). However, the actual striving towards autonomy (including participation in risk behaviors, staying out late at night, visiting places the parents do not like etc.) causes conflicts that tend to temporally reduce the good relationships between parents and children, i. e. temporally reduce the levels of connectedness they feel towards each other. Hence, the growth in autonomy (through conflicts) is related to the decrease in connectedness, whereas the level of connectedness (in a healthy family situation) is positively related to the growth of autonomy. With both autonomy and connectedness related to their own resource factors, K_A and K_C, the dynamic relationships between Autonomy A and Connectedness C are described as follows:

$$\Delta C = r_C(1 - C/K_C) - a\Delta AC + bAC$$
$$\Delta A = r_A(1 - A/K_A) + cCA\,.$$

The first part of the equation describing change in connectedness (ΔC) relates to its resource-dependent growth, the second part relates to its being negatively affected by the change or increase in autonomy, which takes place mainly through conflicts; the third part relates to the positive effect of autonomy on the level of connectedness. The first part in the equation describing the growth of Autonomy (ΔA) describes its resource-dependent growth, the second part describes the growth due to support from Connectedness.

A stochastic version of this model produces developmental trajectories as described in the literature on adolescent development. As autonomy grows, connectedness shows a temporary decline, from which it restores and then shows a gradual increase, more or less parallel with the gradual increase in connectedness. The system then stabilizes around an attractor level, with stochastic fluctuations (see Fig. 1). The local regression and restoration is characteristic of U-shaped growth, which is a typical developmental phenomenon [59,117,226,305,337].

The model is an example of the ecological relationships that can hold for any couple of "growers", i. e. relatively autonomous components of a developmental system. These relationships can be symmetrical supportive, symmetrical competitive, asymmetrical competitive and supportive (as in a predator-prey relationship in a biological model) and, finally, conditional (if a particular, minimal level of a component is a necessary precondition for another component to start growing). The relationships are represented in Fig. 2.

A developmental system is characterized by relationships between any of its components and forms a web of relationships, formally similar to the foodwebs described in biological models [252]. Figure 3 represents an imaginary web of relationships between components in a developmental system.

A major difference between ecological web models used in biology and those used in developmental psychology is that the latter are considerably less supported by em-

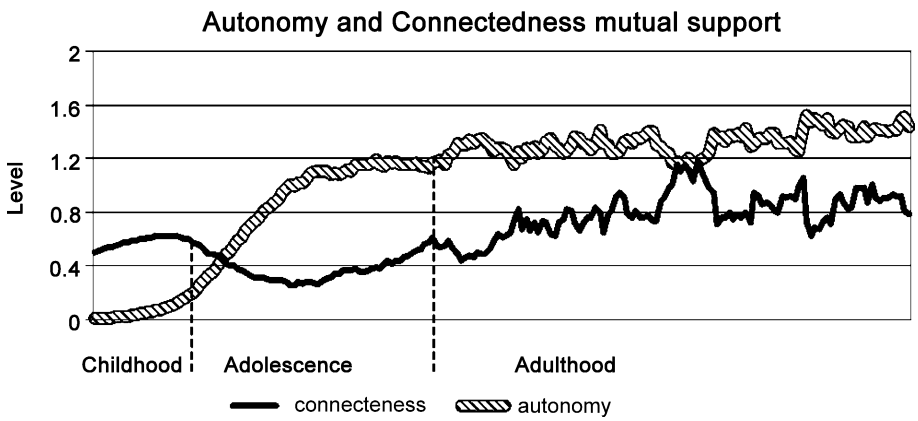

Development, Complex Dynamic Systems of, Figure 1
Developmental trajectories generated by a stochastic version of the autonomy-connectedness dynamics

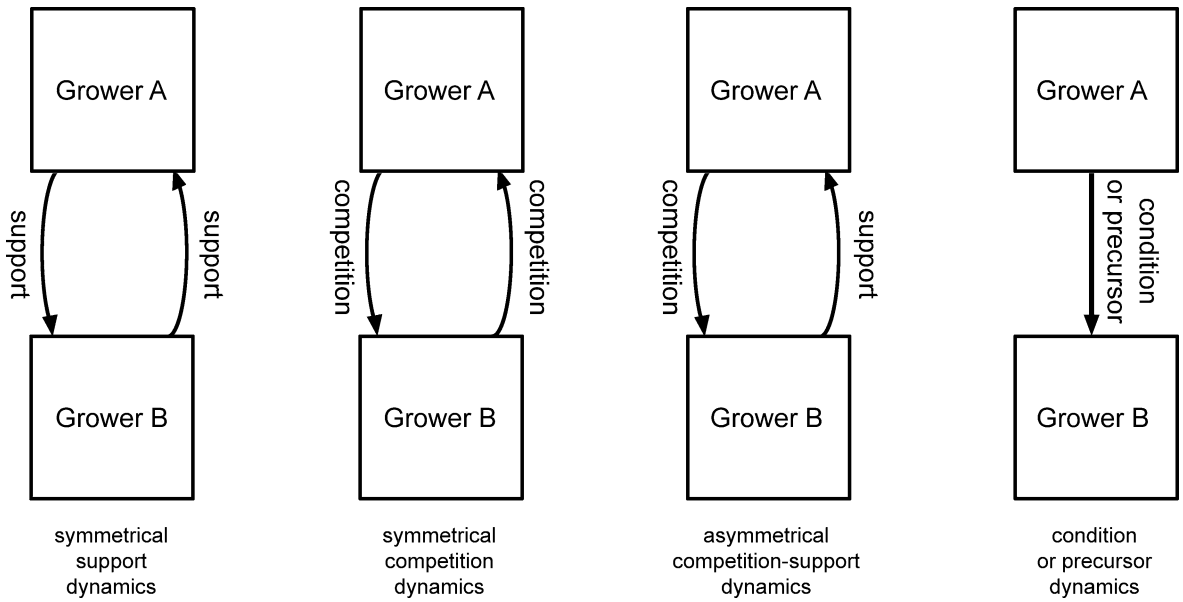

Development, Complex Dynamic Systems of, Figure 2
Four types of dynamic relationships between "growers"

pirical data than the first. The lack of empirical support is due to two factors mainly: the first is that the components distinguished in developmental models are considerably more fuzzy and less tangible than those used in biological models, with the associated difficulty of measurement precision. The second reason is that developmental psychologists are considerably less accustomed than biologists to study real developmental systems over sufficient time spans. Developmentalists mainly focus on statistical relationships between distributions of variables in samples, which tell very little if anything about the dynamics of the developmental system. Exceptions to this rule are the studies carried out in the field of language development, where single-case studies are the rule rather than the exception. By conceiving of language as a developmental system, consisting of components such as semantics, syntax, phonetics etc., or components on a lower level of organization, such as prepositions, adjectives, verbs, nouns etc., ecological network models can be specified simulating the growth of linguistic variables in a single child (see for instance [254,337,340] on lexical development in relation to the growth of plurals; [259] on the growth of closed-class words; and [20] on the pattern of growth and decline of sentences of various sentence length).

A study by Bassano and van Geert [259] illustrates the process of the emergence of three, developmentally successive syntactic generators, the holophrastic, combinatorial and syntactic generators. The holophrastic generator is basically a "one-word grammar", i. e. the set of early grammatical principles that generate utterances with a characteristic word length of one. The combinatorial generator is the developmentally more advanced set of principles that generate combinations of words, typically two per utterance. The syntactic generator is the set of principles that use the syntactic rules of sentence formation typical of mature language. Bassano and van Geert assume a series of asymmetric relationships between a less and a more advanced developmental structure, for instance, the holophrastic and the combinatorial generator. The less advanced structure has a conditional and supportive relationship with the more advanced structure. For instance, a minimum level of one-word productions is needed for the combinatorial generator to emerge, and, in addition, the level of one-word production supports the growth of the combinatorial generator, i. e. the production of two-to-three-word sentences based on a simple combinatorial principle. The more advanced structure on the other hand has a competitive relationship with the less advanced structure. For instance, the use of two-to-three-word sentences negatively affects the use of one-word sentences (see Fig. 4)

A mathematical model of these relationships with three connected growers (the holophrastic, combinatorial and syntactic grower) provides a good fit of the empirical data (model is shown in comparison with the smoothed data; see Fig. 5)

Another example of a model based on growth relationships between components of the developing system

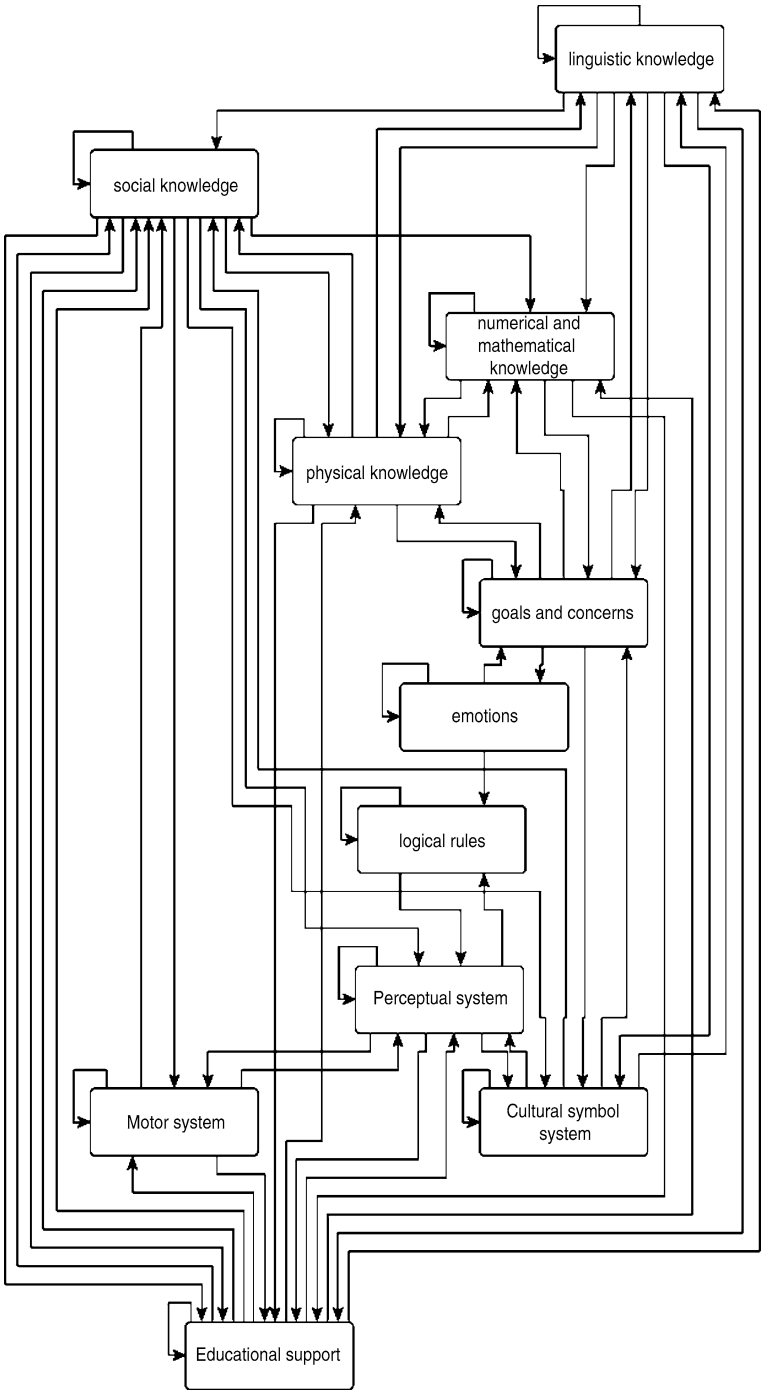

Development, Complex Dynamic Systems of, Figure 3
An imaginary developmental growth web or network. *Arrows* represent either positive (supportive) or negative (competitive) relationships, in addition to conditional relationships. The *reflexive arrows* refer to the nodes' autocatalytic and resource-dependent growth

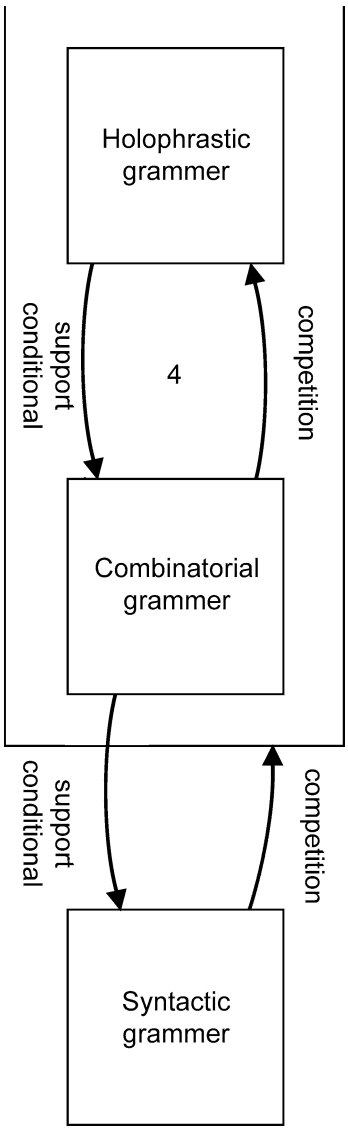

Development, Complex Dynamic Systems of, Figure 4
Asymmetric growth relationships between three developmental levels of grammar (after [259])

is Fischer's model of development through tiers and levels [103]. The model describes development as the emergence of skills, which are context- and content-specific. Skills are general formats of behavioral control and perception, and they are characterized by general structural properties that develop over the life span through a series of more or less discontinuous changes. Discontinuity is primarily observable in what Fischer calls optimal performance, which is a subject's skill level under optimal conditions, including support from other persons (for instance, more competent persons, as in a context of teaching and learning). Fischer distinguishes three major structural types of skills, which he calls tiers: the tier of action (roughly from birth to two years), the tier of representations (from two to about 12 years), and the tier of abstractions (from 12 years up). Tiers are further subdivided into levels, for instance the tier of abstractions goes through a sequence of single abstractions, then mappings, then systems, and finally, principles. Fischer's model has an additive structure, in that earlier levels do not disappear – as in Bassano and van Geert's model of language development – but are conserved and in fact transformed through the emergence of a more advanced level. The postulated structure of relationships between a preceding and succeeding level of development is as follows (see Fig. 6). First, the preceding level acts as a precursor or condition for the emergence of the succeeding level, e. g. single abstractions are necessary precursors of mappings, because a mapping is a relationship between single abstractions. Second, the succeeding level competes with the preceding level, in that the change (growth) in the succeeding level has a negative influence on the preceding level. Third, the succeeding level supports the preceding level: its positive effect on the preceding level is proportional to its magnitude of occurrence ("level" in the quantitative sense).

Dynamic models constructed according to these growth relationships predict typical developmental patterns, for instance in reflective judgment [103,170] (see Fig. 7). In addition, such growth models have predicted developmentally discontinuous trajectories at high growth rates and smooth trajectories at low growth rates, consistent with the data [104].

Development and Dynamic Field Theory Thelen, Smith and co-workers explain the short-term dynamics of action, thinking and knowledge and the long-term dynamics of development, by means of dynamic fields. A dynamic field is defined by an abstract metric dimension (or a space consisting of various such dimensions) that describes the main variable (or variables) of an action or thinking process. For instance, a major variable in the aforementioned object search task in which babies make the A-not-B error is the spatial position of the hiding objects and hiding places, which also defines the major variable of the child's action, which is the place toward which the infant will reach in order to retrieve the hidden object. For each point of this metric variable, there exists a particular activation value, which, in the case of the current example, would mean a likelihood that the child reaches at a particular location. The whole of activation values forms an activation field, the form of which changes on the basis of "inputs" from various sources. In the present example,

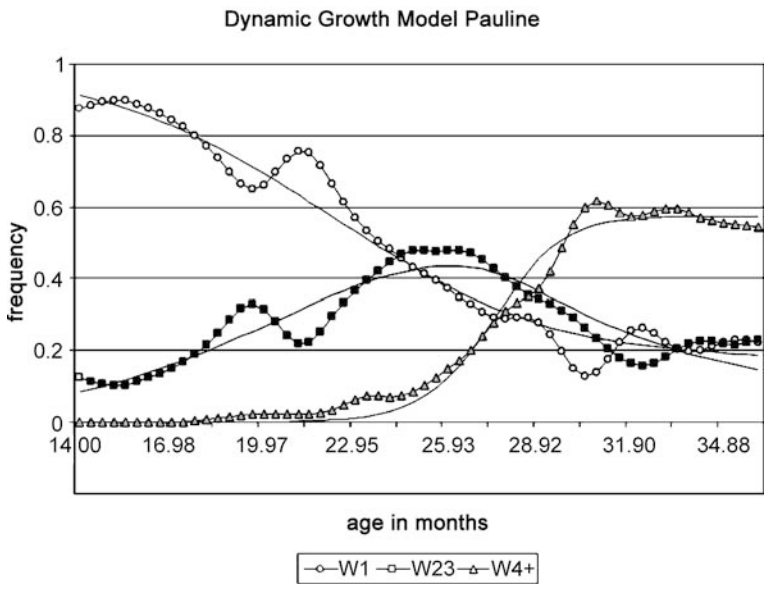

Development, Complex Dynamic Systems of, Figure 5
Smoothed data of the development of one-word, two-to-three-word and four-plus-word sentences in a French girl, Pauline, with fitted growth model based on relationships of competition and support (after [259])

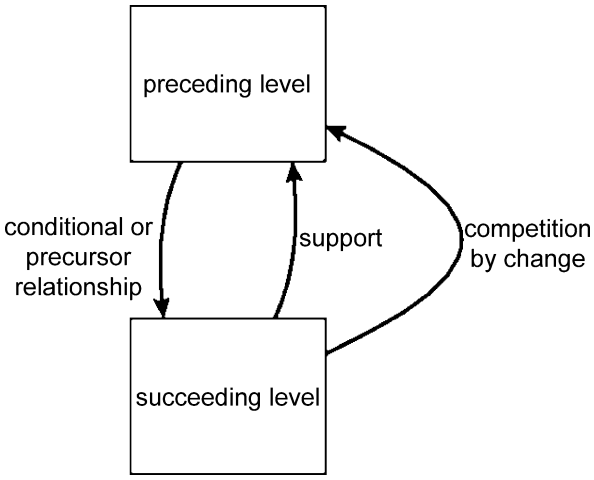

Development, Complex Dynamic Systems of, Figure 6
Growth relationships in Fischer's model of tiers and levels of development

inputs come from the child's perception of the environment, for instance the position of the hiding places, from events taking place in the environment, for instance the adult moving a hidden object to another place; and finally, from the child's memory [273,274,313], which is carried by the neural network that constitutes the brain and that changes as a result of experiences, maturation and self-organizing processes [188,189].

The inputs to a dynamic field are not just linearly superposed: they show cooperative and competitive interactions, leading to self-stabilization of activation patterns [99]. The mathematical properties of that field can be defined rigorously and allow for a dynamic systems model that is no longer just metaphorical. The dynamic field theory that describes the dynamics of this field thus bridges the "representational gap" that exists in current dynamic systems models [296]. This "representational gap" refers to the fact that developmental dynamic systems models of embodied and embedded action have no use for concepts and representations as mental entities that act as mental causes of behavior [74].

In addition to the development of the object concept, dynamic field theory has been applied to development of habituation [273] and development of working memory [274].

Dynamic fields can also be specified for abstract properties of the developmental state space in order to model long-term changes and mechanisms of development [343]. The starting point is the geometric notion of development as specified earlier, i. e. the developing system defined as a manifold of dimensions or variables, describing all of its relevant developmental properties. Since all those dimensions can be ordered along a scale of developmental progress (a developmental "ruler"), the developmental state space is thus characterized by a principal component (in the statistical sense) that can be used to specify any kind

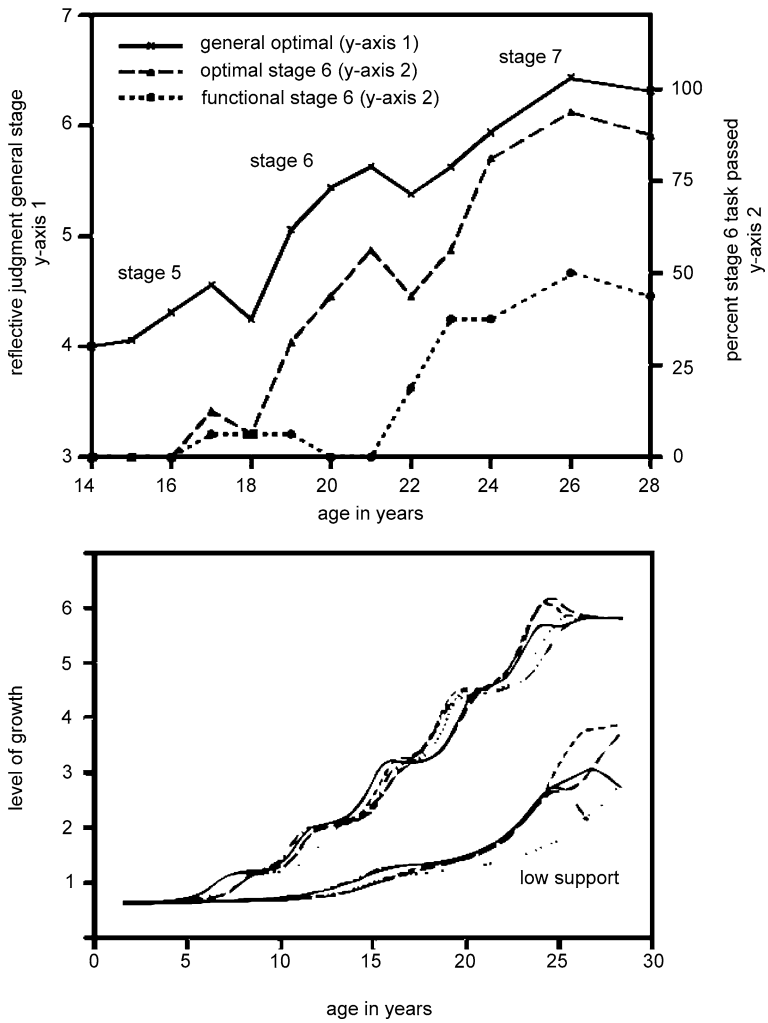

Development, Complex Dynamic Systems of, Figure 7

Fischer's model of structural stages in cognitive development, with empirical data from reflective judgment. The dynamic model based on growth relationships generates the pattern of stepwise change with intermittent regressions (after [103])

of developmental progress or succession. Any point or region in the developmental state space can be mapped onto the principal component of the space, i. e. the general developmental distance introduced above. Any point on this metric distance dimension has a certain likelihood of being "visited" by the developing system. The actual likelihood, i. e. the actual activation field, is determined by inputs from the child's momentary experience of the context, retrieved memories, actions from other persons, and so forth. These likelihoods can be represented as a vector field, with an activation vector for each point in the developmental principal component or distance dimension. The vector field can specify a single peak, in which case the developmental state of the individual is crisp and uni-modal (the classical ideal), or by a landscape of peaks, in which case the developmental state of the individual is multi-modal, fluctuating and fuzzy (which is more like reality; see Fig. 8). For instance, if a child alternates between solving a problem either in a less or in a more developmentally advanced way, it occupies two regions in the developmental space between which it shifts randomly.

Development can then be represented as the change of the vector field over time, beginning with a dominant mode in the lower and ending with a dominant mode in the upper regions. The short-term dynamics of development consist of the individual's actions, experiences and interactions in real time. These real-time actions have a lasting effect on the structure of the vector field. The effect is moderated through two mechanisms that have already been described by the founding fathers of develop-

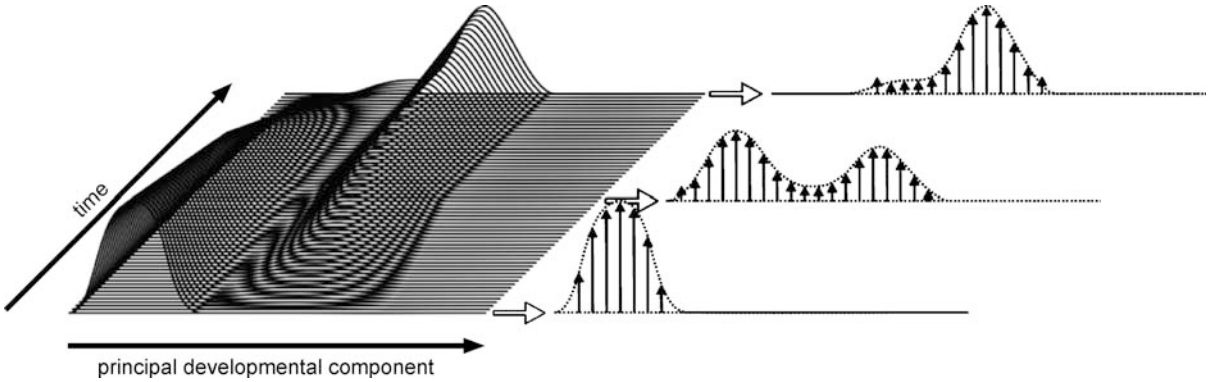

Development, Complex Dynamic Systems of, Figure 8
Probability functions of developmental levels assigned to potential actions or experiences of a child across time. The probability wave moves from a dominant mode on the left to a bimodal mode in the middle to a dominant mode on the right (see the three probability functions with vectors at the *right*)

mental psychology, Piaget and Vygotsky. They see development as the result of what I have freely termed *conservative* and *progressive* forces. These forces operate as follows (see [343,344] for an explanation of the model). It is assumed that any activation of components of the developing system in the form of a particular action, experience or event, have a consolidating effect on those components, and hence on the developmental level(s) that they represent. The consolidation depends on the functional success of the action or experience in question, i. e. on its short-term dynamics. The consolidation takes place in the form of increasing the vector values at the levels corresponding with the action or experience in question. The consolidation function spreads out to nearby regions and becomes negative (reducing vector values) for regions farther away on the developmental distance dimension. The consolidation function amounts to some sort of familiarity effect, which decreases with increasing distance from the actual, i. e. familiar level.

The developing system is also driven by a second force, namely novelty, which is a general term for novelty (new things) per se, curiosity, interest, goal-related activity and so forth. Novelty is a function that increases with increasing distance from the familiar. Assuming that familiarity and novelty are governed by their own characteristic parameters, there is a point on the developmental distance dimension where the combination of both has a maximal value or optimum. The vector values corresponding with this point are also upgraded, with an upgrade function the form of which is in principle similar to the conservative upgrade function. The updated vector field will generate new short-term actions and experiences, which will cause the vector field to update again, and so forth.

Simulations based on this model of development show that, depending on the values of the main parameters (familiarity and novelty parameters, rate of vector field upgrading, nature of information activating vector loadings and so forth), a rich landscape of developmental phenomena can be achieved, ranging from stepwise growth as described in the Piagetian and neo-Piagetian theories, to models of overlapping waves of strategies (Siegler), microgenetic fluctuations in performance, and so forth. An interesting effect of these dynamic field models is that they are able to explain discontinuous changes in development.

Theory of Complex Adaptive Systems (CAS): Developmental Agent Models Although development is a prime example of a long-term adaptive process – and according to Piaget, for instance, adaptation is the central developmental mechanism – there is very little work on complex adaptive systems in the field of development. An adaptive system can be defined as a collection of agents and artifacts that interact with one another and through that interaction are aiming at reaching their goals, concerns or interests [7,17,94,288]. Agents pursue their goals by means of their action repertoires, their knowledge of the world and the information they obtain through acting. In a complex adaptive system, agents are interdependent, yet autonomous in achieving their goals. From a developmental viewpoint, agents learn from their experiences and show long-term change in their goals, knowledge, action repertoire and skills, a long-term process we call development [269,303,352]. In a situated and embodied agent, development not only concerns the change in the person, but also changes in the person's niches or preferred environments [68,69].

A characteristic property of agent-based models is that they conceive of agents as being equipped with relatively simple rules, instead of complex internal representational worlds, carrying out extensive computations before act-

ing [288]. Developmental psychology, on the other hand, emphasizes the complexity and the richness of a child's emerging knowledge and skills. However, the two approaches need not be in conflict with one another, in that very simple rules of agent interaction can in fact emerge on the basis of complex, self-organizing processes requiring the interplay of the environment, experience, knowledge and so forth. An example of how agent-based modeling and development come together is Steenbeek and van Geert's work on the development of social status, social power and social skills in young children [301,302,303]. The model describes the dyadic interaction of children by conceiving of children as agents with particular interests or goals with regard to playing alone or playing with other children. The agents in the model have a representation of the social value (popularity) of the other agents in their group, and try to optimize their positive appraisals of the interaction situation by drawing actions from a repertoire of either solitary action or actions aimed at other children. The principles governing the actions – for instance the choice among actions – are very simple at the level of the agent-based model, but such simple principles are in fact emerging out of a complex multitude of components and influences. An example of such emergent simplicity in social interaction concerns the emergence of goals and interests of the agent through the interactions themselves (see for instance [66]).

The agent-model described by Steenbeek and van Geert [299,301,302] simulates patterns of interactions that are empirically validated by data on play interactions between children from various sociometric statuses. The model can also be linked with a model of long-term changes in the peer selection preferences of children and the emergence of social statuses such as popularity or rejectedness. A similar model is used to explain the spreading of risk behaviors among groups of adolescents and the formation of friendship groups or cliques [12].

One possible drawback of agent models for development is that the nonlinear and complex patterns that typically result from interactions between agents do so if the number of interacting agents is high (see [11] for many examples). The typical number of agents in developmental models is small, e. g. two in dyadic interaction, or ten to twenty in small group interactions (e. g. a group of friends, a peer group). In spite of this limitation, agent models are typically suited for simulating short-term temporal patterns of interaction among persons of various developmental levels, and it is from these patterns that the interacting persons learn, adapt and develop.

Agent models have been used to study and simulate language development in connection with language evolution, by modeling generations of language learning, teaching and communicating agents. Such models explain, among others, why language tends to evolve towards a capacity that is optimally adapted to biological and social learning principles, and thus tends to become an increasingly "innate" type of capacity (examples of such studies are [66,165,168,169,290].

Theory of Complex Adaptive Systems (CAS): Epigenetic Robotics A different type of agent models in developmental studies concerns the epigenetic robotics models. The embodied and embedded aspects of developing agents are literally implemented in the form of robotic models, i. e. artificial autonomous agents that wander about, act in real environments and interact with other agents, including humans, and must learn from their experiences. Metta and Berthouze [211] define epigenetic robotics as the study of how a realistically embedded and embodied robot-model of a person, including a brain, sensory and effector organs, changes and develops in interaction with a real world (for introductions and reviews, see [26,27,28,182,202,211,267]. In fact, "… Beyond a certain level, it becomes extremely difficult to study realistic interactions between the agent and the environment without including a real body and real people in it" (p. 130 in [267]). Epigenetic robotics serves two purposes. The first is of a technical nature, and is aimed at designing self-organizing and learning robots that serve practical goals, in cases where direct programming of the robot is too complicated (e. g. [376]). The second is of more concern to the present article, which is to understand developmental processes in humans by studying simulated but embodied (robotic) agents in real environments. Epigenetic robotics approaches have so far dealt with the following aspects of development.

The first refers to the observation that human learning and development very strongly depends on social scaffolding and socially situated processes of cognition and perception. An important issue is joint attention, which occurs at a very early age (approximately around nine months) and involves the infants capacity to infer an object or event of attention to an adult by using the adult's gaze, pointing, etc., and to share that topic of attention with the adult. Joint attention is a typical human capacity, and greatly facilitates the process of cognitive and social learning. A host of robotics studies have been carried out, showing that processes of joint attention or closely related to it can be implemented in a robotic system and greatly enhance the processes of cognitive, linguistic and social learning [163,225,268,306,307]. Joint attention is closely related to empathy, i. e. the ability to in-

tuitively understand the minds of others by picking up their intentions, an ability which is related to specific neurophysiological structures, the so-called mirror-neurons, in human and primate brains (for studies using principles of epigenetic robotics, see [39,212]; and for applications to impaired attentional processes, see [239]). Socially-situated learning rests heavily on imitation learning, or learning through emulation, which means transforming the perceived actions into one's own action repertoires, aiming at the inferred goal or intention of the perceived action (for studies on imitation learning in epigenetic robots, see [47,83,116]. An important aspect of perceptual development concerns the infant's very early capacity to integrate information from the various senses and perceive the world in a multi-modal way. Multi-modal perception can also be accomplished by epigenetic robots through associations of information from various sensory organs [106,248]. Epigenetic robot studies have tested embodied and socially-situated processes of cognitive and language development [86,87,197,200,377] and emotional development and communication [44,45,46,56].

The Forms of Development: Fluctuations, Variability, Continuity, Discontinuity and Critical States

Development and the Notion of Stages Classical developmental theories (e. g. Piaget, Erikson) typically view development as occurring in stages, that is, the course from the initial developmental state to some sort of end state is seen as a stepwise path, or a path moving across various qualitatively distinct states. Piaget's theory, for instance, describes a first stage as a sensorimotor level of thought, after which children proceed to a second level called pre-operational, then to concrete-operational thinking. Development finally stabilizes at a formal operational level, which is characteristic of adult thinking.

Recent stage-oriented theorists, in particular the neo-Piagetians, occupy a considerably more sophisticated standpoint [103]. They see "stages" as in fact qualitatively different forms of thought, or skill in general, that are developmentally ordered but are also context- and domain-specific [58,82,103]. A child may function on stage (or level) 1 in domain A (e. g. simple mathematical operations) and on level 2 in domain B (e. g. social relationships). Within a domain, such stages – or one should say levels – can fluctuate with varying context, because context is a part of a person's skill (e. g. a child who faces a particular problem context may function on level 2 with help and on level 1 without help). The levels or stages may fluctuate strongly over the short-term time scale of a problem-solving event, in a process that Fischer has called "scalloping" [103,130]. Overall, however, there is also a fuzzy but nevertheless convincing ordering in the level or stages. Two-year olds, for instance will show a very different mixture and frequency of context- and domain-specific levels than adults, and are thus characterized by a different major-stage category than adults are (see [80] for an example). The notion of stages advocated by neo-Piagetian theorists reflects the complexity of the developmental system by viewing stages all the way down, in a complex, hierarchical and dynamic organization.

The notion of stage (level, phase,. . .) entails an idea of internal coherence, a relatively stable structure of interdependent elements such as skills, habits, processes and so forth. The notion of stage is thus highly reminiscent of a basic notion from dynamic systems, namely the notion of attractor. Starting from the theory of complex systems, we can follow the assumption that such systems tend to self-organize into islands of relative stability rather than remain unconnected collections of features where any combination of such features is as likely and (in)stable as any other. From this, we can reach the conclusion that stages, defined in the dynamic and complex way explained above, should be the default option for a system as complex as human development.

Developmental Stages and the Theory of Bifurcations Developmental attractor states do not need to be overall states in the sense of stages, but can amount to any stable pattern of coordinated knowledge or skills at any level of aggregation. The transition from one such state to another represents a discontinuity, since none of the intermediary states, if there are any, is stable. Developmental researchers have used the framework of catastrophe theory to answer their questions about developmental discontinuity (for an overview see [327,351]). By testing for empirical indicators of the so-called catastrophe flags (structural properties of discontinuities in general), they have tried to show that developmental transitions are instances of the so-called cusp catastrophe and thus entail a clear form of discontinuity. Examples of phenomena studied are the transition between non-conservation and conservation understanding in young children [139,150,151,157,158,326], reaching and grasping in infants [373,374] and syntactic development [259,329]. The results show that rapid, jump-wise development takes place in a variety of domains. However it remains unclear whether these changes are real discontinuities in the bifurcation sense. In addition, they seem to occur in some children but not in all. A problem with discontinuities is that the empirical detection depends on the definition given by the researcher (see [329] for discussion). Finally, different attractor patterns can occur si-

multaneously and in that sense show a form of superposition [343,352]. Children can act according to a less advanced skill pattern and only little later act in accordance with a more developmentally advanced pattern. If such patterns co-occur, they will lead to increased intra-individual variability in performance, which has been shown in the above-mentioned catastrophe-theoretical studies and in studies of language development [20]. As development proceeds, the less advanced pattern will disappear in some cases, but remain unchanged in others. In language development, for instance, less advanced grammatical principles disappear, but on the level of cognition as a whole, earlier patterns – one should not particularly call them less advanced – can survive and be used in contexts where the more advanced patterns are not directly applicable (see the scalloping principle mentioned earlier). Finally, developmentally divergent levels of skill or knowledge can even occur in the same action pattern, for instance if the child's verbal explanation refers to a less advanced level and his non-verbal gestures to a more advanced level of understanding (or vice versa, see [119,156].

Because behavior and action are in themselves highly variable phenomena and because this variability is intrinsic to behavior and not a matter of added measurement error or noise, the continuity issue amounts to the question whether bands of fluctuation or variability in behavior develop continuously or discontinuously [329]. The growth models described above represent development or growth in a variable as a single point on a dimension, but this point should be seen as an estimation of a central point of what in reality amounts to a bandwidth of fluctuation.

Are Developmental Stage Transitions Like Physical Phase Transitions? One possibility is that what has traditionally been called "stages" are comparable, in that sense, to the phases of physical matter (gaseous, liquid, solid) and depend, in essence, on a single parameter or a confluence of parameters. Developmental stages form attractor states because they rely on network-like structures, i. e. on structures of relationships between the components of the system. A developmental attractor state is represented by habitual, coherent patterns of performance, skill or action that self-organize spontaneously in the person's habitual contexts, niches or living spaces. These patterns consist of mutually-supportive and sustaining features. To give a simple example, Piaget's sensorimotor stage defines thought in the form of external action on objects. For instance, reaching to and grasping an object requires the coordination in real-time of a myriad of components or aspects, including the coordination of the muscles in the arm and hand, the coordination of vision and movement, the coordination of vision of the object and vision of the own arm and hand, and so forth. These patterns are not innately given, but self-organize through processes that eventually amount to discontinuous changes (e. g. Wimmer's studies of early prehension development, Wimmers, Beek, and Savelsbergh, 1998a; Wimmers et al., 1998b; [372]). The characteristic feature of these sensorimotor patterns is that their contextual self-organization (e. g. in the form of reaching to and grasping a particular object) emerges on the basis of dominant driving forces or control parameters that are of a sensory and motor nature (see for instance dynamic field theory described earlier). In addition, the sensory and motor control parameters of infant action are likely to be biologically preadapted to important features of the environment, such as object-person distinctions, numerosity, etc. (see for instance [295,372,373,374]).

Are Developmental Transitions Caused by Self-organized Criticality? Irrespective of the stage theory under consideration, the durations of sequential stages tend to increase in a logarithmic manner [340]. One might ask if the distribution of stage durations relates to the power law distribution characteristic of self-organizing phenomena [242,354], and more particularly, to self-organized criticality [10]. The phenomenon of self-organized criticality emerges in complex systems, consisting of many components that entertain local relationships. We have already seen that the embodied-embedded brain is such a system, consisting of many components (perceptions, thoughts, actions, memories, tools, environments) that are temporally and functionally connected. This complex system is under a certain external "tension": there are problems to solve, goals to achieve. The person does so by means of the complex system of skills, knowledge, sensory and motor systems. As not every action is successful, the person will adapt and learn from his experiences and from being taught by other people. This complex, interconnected system exchanging information with the world is a good example of a system that shows self-organized criticality. Its attractor states are critical states, i. e. states for which any external influence can cause patterns of change with a wide variety of magnitude and duration, dissipating the stress that has been build up in the system. Note the major difference with a phase transition model: in a phase transition model the attractor states are the phases, whereas in a critical transition model the attractor states are those where a transition might occur.

The magnitudes and durations of changes are statistically distributed according to a power law distribution, with very few large-scale changes and increasing numbers

of smaller scale changes. It is tempting to see development as an example of such a self-organized criticality: a succession of meta-stable states punctuated by changes of various magnitude (e. g. a relatively small change in a relatively context-specific problem solving strategy, versus an avalanche of changes in many aspects and domains of cognitive performance, the latter characteristic of what would count as a stage transition).

If for some reason something changes in one skill (or knowledge, ability, action pattern, habit) it is likely to affect other skills (habits, etc.) to the extent that these two developmental components are interrelated. However, the second component, affected by the first, can eventually affect a third one to which it is connected, and so forth. In principle, such changes can remain quite limited, but they can also grow into an avalanche of changes that affects the whole developmental system. If we assume that in a developing system the "weakest", i. e. the least adapted or effective skills (habits, knowledge) are eliminated (or altered) more easily than better adapted or more effective skills, we wind up with a system that closely resembles the Bak-Sneppen model of biological evolution through punctuated equilibria [9,35]. This model of evolution changes through many events of extinction and speciation, interspersed by periods of stasis.

The pattern of evolution with many small and only a few major extinction-speciation events is clearly reminiscent of the course of human development, with many small and a few major changes. The principle of eliminating or altering the weakest component is also applied in a routine for solving hard optimization problems, called extremal optimization [36]. The solution patterns are characterized by shifts following the power law distribution. In a certain sense, (cognitive) development is like solving a hard optimization problem, an adaptation of knowledge and skills to the complexities of reality. It would thus not be surprising that the general dynamic structure of cognitive development follows a pattern very close to that of the extremal optimization process, including the power law distribution of the changes.

An Overview of the Human Life Span in Light of the Theory of Complex Dynamic Systems

Preliminary Remark

A discussion of the issues of dynamics, self-organization, complexity and so forth in the context of human development requires that the reader has some basic knowledge of how current developmental psychology describes the human life span from the viewpoint of developmental processes. In order to provide such knowledge, I will present an overview of selected themes and topics discussed in mainstream handbooks on developmental psychology [25,51,73,162,227,276,280,356]. The themes are chosen for several reasons, mostly because they are considered of great importance to development, but also because they provide interesting possibilities for applying a complex dynamic systems framework on issues that are usually not seen in this light. The overview presented is therefore, by necessity, only fragmentary and exemplary. Interested readers are referred to a host of introductory handbooks (see above for some suggestions). Because of their importance for starting and guiding development, I shall concentrate more on early processes of development than on the later ones, and confine myself to the period between birth and adolescence.

Most handbooks give a review in terms of stages or phases in the life span, with subdivisions based on major topics or fields of development such as physical and motor development, social and personality development, cognitive and language development as the main sections.

Handbooks often start with a theoretical introduction, discussing major theories such as Piaget, Vygotsky, Freud and so on. These theories mainly or exclusively refer to the work of historical figures who laid the foundations of the field. In addition to such theoretical perspectives, handbooks also discuss basic questions such as the nature-nurture relationship.

Prenatal Development

Prenatal development covers the period from conception to birth. Normal fetal development occurs through three stages. The germinal stage covering the first two weeks after conception is a period of cell division and implantation in the uterus. The second or embryonic period is a period of emergence of essential organ systems, such as the central nervous system, the heart etc. and lasts from week 3 to week 8. The third or fetal stage lasts from week 9 to week 38 in case of full term birth. During the embryonic and fetal stages, the embryo or fetus is sensitive to teratogenic influences, i. e. influences of substances in the mother's body that negatively affect the growth of specific organic systems, due to disease or intake of substances such as alcohol. The influence of teratogenic substances depends on the level of development of the embryo or fetus and is thus related to periods of greater or lesser sensitivity to such influences. During the fetal stage, the child becomes increasingly sensitive to sensory stimulation, which is limited and modified by the child's position inside the body in the amniotic fluid. Of particular importance is the communication with the mother's men-

tal and physical state through the exchange of hormonal and chemical substances that reflect the mother's current psychological state. Such influences can affect the child's later neurophysiological reaction to stress, for instance. Already before birth, the unborn child and the pregnant mother entertain a transactional relationship. The child is affected by the mother through the sensory and neurochemical links described above, whereas the unborn child affects the mother's behavior, moods and evaluations. The effect of the child on the mother ranges from direct effects, for instance the physical stress of pregnancy, to indirect effects via individual, family and cultural evaluations of the mother's current pregnancy. Of particular importance for later development are eventual problems during the birth process, such as hypoxia of the fetus during birth, or premature birth. Most of such relatively minor birth problems are related to diffuse and (very) mild effects on performance, for instance cognitive and academic performance at school age.

Infancy and the Preschool Years (Birth to 4 Years)

General Introduction Infancy is the period that lasts from birth to 24 months. It is the age of sensorimotor functioning, more precisely action that is mainly limited to a direct coupling between sensory and motor systems in an action context. Action itself emerges out of the newborn's reflexes and reflexive actions. The basic sensory capacities are relatively well developed, in that they suffice to allow the infant to make sense of the environment and perceive the environment "as it is". That is, the infant is capable of perceiving the environment as a structure of functional affordances, e. g. as a three-dimensional space with identifiable objects and events that the infant can relate to in its actions. For example, very soon after birth, the infant is able to locate objects in space, for instance by following moving targets with the eyes and turning to objects identified by their sound. Given the early and seemingly automatic adaptation of action of the infant to the core proporties of the physical and social world, various developmentalists have endowed the infant with innate core knowledge of the main features of the world. These core knowledge systems comprise objects, actions, number, space and social partners [295]. The notion of identifiable core knowledge systems has been criticised from a dynamic systems point of view as not presenting a model of the real time mechanisms creating the developmental expression of such knowledge in real, physical situations, and which, in the words of Smith [291] are "general, probabilistic, emergent and distributed across several levels of analyzes".

Cognition and Intelligence Infant *cognition and intelligence* is deeply sensorimotor, that is, intelligent action takes place in the form of real physical action. Infant action, involving looking, grasping, repetitive actions, quitting the action and so forth, must be understood as dynamic processes, the patterning of which unfolds in real time, thanks to the continuous feedback loops between the perceiving and acting infant and the affordances of the infant's proximal environment. By "patterning" one can understand correlational regularities in the action sequences, for instance regularities such as following a moving object with the head and eyes, where the object gets temporally occluded by other objects, for instance. The patterning is a relatively high level patterning, however, in the sense that the regularities are interpretable as expressions of or as being consistent with real-world properties such as the permanence of objects or the existence of causal relations among events. In the mentalistic framework, interpretive patterns such as object permanence or causality were viewed as internally represented structures, generating or producing the observable actions. According to dynamic systems theorists, these patterns are not the cause of, but emerge through the real-time interaction between the infant and the proximal environment (see for instance the studies on the infant's understanding of object-permanence in the A-not-B-error experiment [292,296]. As development proceeds, the dynamics of this real-time unfolding of understanding and acting on the world will become increasingly complex, e. g. by incorporating verbal actions, by enriching memory and by enriching and changing the perceptual organization of the environment.

The infancy-to-toddler phase is characterized by a succession of two major modes of thought. The first is the sensorimotor form, in which thought is entirely expressed through sensory and motor action as described above. In the second phase, thought incorporates forms of representation and symbolization, thanks to the blending of thought and sensorimotor action. In addition to using language for and in thought, for instance in the form of private speech [361], symbolization also typically entails symbolic or make-believe play, which is related to skills needed for thinking about other people's minds (so-called Theory-of-Mind [112]; ToM is further discussed in the section on childhood).

The distinction between perception and symbolization should not be taken too strictly. According to the tradition of Ecological psychology [118], perception means picking up invariants and properties of the environment that "resonate" with the functional abilities of the perceiving organism (this view is also closely related to Thelen and Smith's dynamic systems theory of development [314]).

A typical example of the abstractness of direct perception in infants, in addition to the examples already given on early core knowledge of space, time, causality and so forth, is the perception of the goal-directedness of actions of other persons by infants younger than 1 year of age [31].

Motor Development In the *motor domain*, the infant develops capacities such as sitting alone, crawling, standing and walking. Although such milestones are usually associated with age averages, such as walking around the age of 12 months, individual differences in onset of such motor skills are considerable, for instance between 9 and 18 months for walking. In the motor domain, the disappearance of the early stepping reflex and its replacement by real stepping at a later age provides an illustration of dynamic systems thinking. Thelen and Fisher [314] argued that the disappearance is not due to central, cortical processes, and that it is in fact not really disappearing. Due to more "peripheral" biodynamic processes, namely differential growth in body mass and muscle strength, the reflex is inhibited biomechanically, and "reappears" as soon as the baby's legs are held under water for instance.

Language Development *Language* begins as vocalization, which around the age of six months changes into babbling, i. e. combining vocalizations into larger clusters. Around the 12th month, the first recognizable words begin to appear, notably words referring to caretakers, the meaning of which is probably assigned by the adults instead of actually intended by the children, who form such sound patterns (mama, babab etc.) more or less automatically. Meaning assignment forms an interesting illustration of the transactional nature of early developmental processes, with the adult automatically assigning a meaning to sound patterns uttered automatically by infants, which then sets out an iterative pattern of meaning assignments and language production, resulting in the child's understanding of referential meaning and linguistic significance (around 18 months). Language production in infancy typically consists of one-word sentences expressing various semantic features and nuances (holophrastic speech). Word combinations, mostly in the form of two-word sentences, become abundant around the age of 24 months. Meanwhile, the child's language production and comprehension gradually assimilates the syntactic features of the child's ambient language. The process is typically dynamical, in that it consists of an iterative trajectory of assimilations of syntactic features given the child's current state of language production. Various quantitative features of language development can be described by means of growth processes of linguistic and non-linguistic components connected into a web of mutually supportive or competitive relationships, under the constraints of limited and specific resources [337,340,346]. In addition to continuous growth processes, language development is also characterized by discontinuities, corresponding to the emergence of new strategies of language production [21] or new linguistic forms or categories [259,329].

Our understanding of the dynamics of language development is crucially dependent on how we understand language per se, although – in the spirit of genetic epistemology – our understanding of the nature of language can be greatly enhanced by the study of language development. For a further discussion of this relationship, which is a typical issue of developmental psychological theory formation, see the Sect. "Development as Increasing Complexity Applied to Language Theory and Theory of Language Development."

Brain Development As regards *brain development*, after the prenatal period, productions of new neurons virtually stops or is greatly reduced. Synaptic connections between neurons are abundant in the beginning, and will be selectively lost in a process called synaptic pruning, which relies on experience and practice. During infancy, the two hemispheres become increasingly specialized (lateralization process). Meanwhile, the brain, or more precisely the cortex, remains highly adaptive, a property known as brain plasticity . Brain specialization occurs through complex, non-linear dynamic processes involving interactions inside the brain and interactions between brain and body and body and environment [187,188,189]. Brain plasticity decreases as the person grows older, but does not disappear [175].

Brain plasticity however, is a typical dynamic property which is nonlinearly dependent on the brain's developmental history. The change in plasticity is not linear or curvilinear, as the notion of a gradual decline in plasticity suggests. Rather, there are nonlinear peaks of plasticity, known as critical periods or sensitive periods. These critical or sensitive periods, in which the brain is particularly sensitive to particular experiences, are in themselves also self-organizing and dynamic phenomena [49]. They emerge epigenetically from the brain's development and are thus co-dependent on biological brain growth and the unfolding of experiences, including teaching and learning over developmental time [173,319]. According to a now obsolete view on sensitive periods – critical periods – the sensitive period is like a time window of opportunity that, if missed, will never come back and will leave the person with an irreparable gap in development. This view relates sensitive periods to relatively isolated processes of growth

in the brain that unilaterally govern the developmental process. It is incompatible with the view, for which there is now abundant evidence, that sensitive periods are self-organizational states integrating interdependent phenomena of brain growth, experience and environment [159].

A dramatic illustration of how brain plasticity – and development as a whole, for that matter – always passes through the short-term dynamics of action, is the development of children after hemispherectomy. Hemispherectomy is the surgical removal of a brain hemisphere, mostly as a last possibility for curing major and highly frequent epileptic insults that cannot be treated pharmacologically [22,154,355].

Social and Emotional Development Basic *emotions* such as anger, sadness and happiness, are already present during early infancy, which suggests that basic emotions are innate patterns. However, dynamic systems theorists have challenged this interpretation and view basic emotions as the earliest patterns of emotional expression that arise through self-organization of components from various sources (motor, contextual, ...) and become stable attractor patterns [54,55]. Infants are able to interpret emotions of other persons early in infancy. The ability to understand the relation between an other person's emotional expressions and that person's goals and actions is already developed at the end of the first year [218,240].

Temperament is a person's habitual pattern of emotional reaction, activity level, attention and self-regulation. Temperament in the sense of such stable characteristics occurs from early infancy on, with about 2/3rds of the infants falling in the categories "easy child", "difficult child", "slow starter". Temperament, as an invariant property overarching contextual variability in reaction patterns, is a typical short-term form of stability. Temperament changes over the long-term process of development, but it does so to varying degrees, depending on the person (e. g. extreme patterns are less likely to change) and age (e. g. early temperament tends to change more easily than temperament at a later age). In that sense, temperament is a higher-order short-term attractor pattern of behavior and emotion, which over the long-term tends to shift across the temperamental state space with context-, person- and age-specific velocities. The classical study of Thomas and Chess on temperament development [318] provides an example of a dynamic person-context transaction, known as the goodness-of-fit hypothesis.

A third theme that is of particular importance for early social and emotional development is the development of *attachment*. Around 6 to 8 months, infants begin to develop a strong affectional bond with familiar people, the mother or primary caretaker in the first place. This strong tie serves as a source of emotions, such as comfort and joy while the object of attachment is present and sadness or discomfort when he or she is absent. Attachment thus serves as a basic model for the emotionally close relationships that will develop and last throughout life. As with temperament, attachment shows a certain stability across short-term fluctuations in contexts, and a partitioning into characteristic patterns. About 2/3ds of the infants show a pattern of secure attachment; other infants show avoidant, resistant or disorganized patterns of attachment. In a dynamic systems framework, these patterns should be interpreted as short-term attractor states, showing a gradual long-term development, with considerable individual differences in the amount of displacement over the developmental state space [342].

Childhood (5 to 12 Years)

Cognitive Development A characteristic feature of cognitive development during this phase is that it makes a transition to a new mode of thought, concrete operational thinking, a term stemming from the work of Piaget (as do most of the basic terms used for stages or phases as represented in the major handbooks). A major feature of this type of cognition is that it shows reversibility, a notion that is similar to the notion of inverse operations in mathematical groups. Hence, for every action the child can think of, it automatically knows there is an imaginable operation that undoes the effect of the first. The emergence of a property such as reversibility amounts to an increase in the complexity of the cognitive system.

From a dynamic point of view, such formal properties must be given a concrete temporal meaning, for instance in the form of a pattern of reasoning about a possible inverse operation that is not explicitly given in perception.

Neo-Piagetian theorists such as Fischer and Case have postulated comparable qualitative extensions of the child's cognitive system. In Fischer's theory, for instance, the cognitive system which is able to represent relations between elements (e. g. a relationship is-a-brother between me and my brother) is transformed into one which can represent relationships between relationships; see for instance [103].

As regards the theoretical interpretation of what such increases in or extensions of complexity of the cognitive system actually mean, various considerations should be taken into account. The first is that properties such as reversibility and comparable formal properties of the cognitive system refer to action potentialities of a system that is based on or consists of an embodied neural network in a concrete, spatiotemporal world. Hence, the gener-

ative cause of child's actions and reasoning is the system consisting of a concrete context or niche on the one hand and the child – as embodied and organic neural-network on the other hand [350]. This generative cause generates a stream of action to which certain formal properties can be ascribed, such as reversibility (and comparable properties). Second, a major feature of dynamic theories is that they view this generative cause as a dynamic interdependence of many factors (memory, perception, action, changes brought about by action in the context, recent experiences, long-term effects of experiences and so forth) on many levels (from the micro-level of neuronal activity via the meso-level of bodily activity to the macro-level of physical-cultural environments). There is no single factor that can be held responsible for an emergent phenomenon such as reversibility of cognitive representations [19,103,271,293,368].

Examples of developmental acquisitions typical of this phase are conservation, classification, and seriation. Conservation, for instance, is the child's ability to understand that physical properties, such as the amount of liquid in a glass, are conserved under certain operations, such as pouring the liquid into another glass. The emergence of conservation understanding is a typical example of a discontinuous development, i. e. it tends to occur in an all-or-none fashion and can be described as a bifurcation or cusp catastrophe [325,327]. However, the developmental emergence of a phenomenon such as conservation is characterized by a considerable intra- and inter-individual variability. That is, children tend to dramatically fluctuate in their type of conservation answer (yes or no) over test occasions during the period of transition. Moreover, children also tend to differ considerably from each other in terms of the path they take towards conservation understanding. The emergence of conservation forms an interesting example of a phenomenon of development that resembles the phenomenon of phase transition in physics.

During childhood, an important aspect of the development of information processing concerns the development of executive functions, functions that critically depend on the frontal lobe. Important skills related to executive functions are inhibitory control (resisting habits, temptations, or distractions), working memory (mentally holding and using information), and cognitive flexibility (adjusting to change). These skills are of crucial importance for and to a considerable extent dependent on scholarly learning during childhood [79,84]. They are closely related to issues of time management (when to do what, how long to retain information), adaptability and goal-and-desire structures (how to focus on which goals and when). The dynamic manipulation of goals and tools (physical and symbolic) and the patterning that emerges as a consequence of this process is a characteristic feature of dynamic accounts of cognition and cognitive development [19,322]. By "patterning" I understand the unique structure of components and relationships in a child's ability repertoire as witnessed through context-specific and context-supported actions. This ability repertoire greatly extends during childhood, as the child goes to school and acquires skills in major domains such as math, reading and writing, conceptual knowledge and so forth. These new domains of thought and symbolic action were prepared and prefigured during the preceding developmental stage (see for instance the discussion on core knowledge), but show an explosive development in (schoolgoing) children during this developmental stage. The acquisition of these fundamental domains of skill and knowledge requires massive recruitments of brain regions [298].

Social Cognition In this section I shall concentrate on an interesting aspect of social cognition which nicely illustrates the complexities of development and which is also increasingly accepted as being of applied and clinical importance, namely Theory-of-Mind. Theory-of-Mind (ToM) is the ability to attribute mental states to oneself and others and to use these attributions in understanding, predicting and explaining behavior of oneself and others [19,214] or the book 'Autism: mind and brain' [111]. A typical aspect of ToM is the ability to entertain first-order beliefs, i. e. beliefs about the beliefs (thoughts, knowledge, ...) of other persons, which typically develops around the age of 4.

There is supportive but indirect evidence of two 'approaches' to ToM: an intuitive (or automatic) and a reflective (or controlled) route [196]. Indirect evidence for an intuitive, neurophysiologically-based understanding of ToM-related properties of other persons comes from the rapidly growing literature on the neuronal systems that underlie the spontaneous understanding of human actions and psychological states of others. An example of such systems is the mirror neuron system (for the relationship between the mirror neuron system and autism, see for instance [113,135,153,183,232,371] but see [81] for critical remarks). There is neuropsychological evidence that specific parts of the brain, such as the medial prefrontal cortex and the temporal-parietal junction are involved in the processing of ToM-related information [110,174,196,266].

In a dynamic systems framework, the ability as described under the definition of ToM above does not emerge from an internal symbolic structure, generating prescriptions for acting under certain conditions (e. g. a false belief experiment). ToM, as it is actually expressed,

in real time, emerges out of a coordination of many components, including perceived similarities in bodily appearance and action between the child and other persons, goals and emotions, inhibitions of associations (e. g. between what I know and what another person eventually knows), automatic simulations of actions and emotions perceived in others, language and linguistic terms for expressions of states of mind and so forth. This ability grows throughout childhood, until it entails abilities such as the ability to think about how other people think you think (second-order beliefs).

Other aspects of social and socio-cognitive development that are worth mentioning in this regard are the views children develop on their own person in relation to others. These views concern the child's self-concept, the sense of competence and self efficacy, gender- and gender-role concepts and so forth. From a dynamic standpoint, these concepts are relatively stable and adaptive patterns of action and emotional evaluation featuring in situations or events that elicit self-referential aspects are in general self-sustaining (e. g. when confronted with a math assignment, a child must ask himself "will I be able to solve that assignment", "will it require little or great effort", etc.). In a social interaction the child may react with intense negative emotion when confronted with statements about his capacities, lack of fit with the gender stereotype and so forth. For examples of the interdependence between short- and long-term time scales in this particular domain, see [301,303].

Self-referential emotions and concerns play an important role in the dynamics of group relationships among peers. Children in this particular age range tend to form patterns of social interaction, creating particular interaction positions such as popular versus rejected children. These so-called sociometric structures (because they formed the subject of an approach to children's social interactions known as sociometry) tend to self-organize as a result of interactions driven by concerns, emotions and perceptions of actions of other children in the group (for an explicit dynamic model of this process, see [208,300,302,303]).

Adolescence

Biological Maturation and the Problem of Timing Puberty and adolescence are ages of accelerated change and development. A major aspect concerns sexual maturity during puberty, which involves major changes in primary sexual development (the person becomes biologically capable of reproduction) and secondary sexual development (the person develops bodily features characteristic of one's sex). These biological changes are related to a spurt in bodily growth, the functioning of the endocrine system and changes in brain development [29,78]. Biological changes are directly related to social changes in terms of interaction, interaction forms and preferences, and changes in cultural expectations.

The structure of interactions between biological, social and cultural changes provides an interesting example of the importance of timing (the temporal ordering of related events). Sexual maturation is an extremely important event. From the point of view of the evolutionary time scale, it must occur at the "right" time, which means that it must run parallel with the timing, i. e. the temporal sequencing, of additional events, required for successful reproduction. The timing is a result of events that preceded it, and on the other hand timing provides a condition for later events. For instance, pubertal timing is responsive to ecological conditions earlier in life, which might be beneficiary as well as adverse. Such conditions may lead to either inhibiting or accelerating onset of puberty. The effects of earlier or later puberty, on the other hand, are non-linear, being most notable in the extremes (both positive and negative [43,95,97]. Differences in the timing of biological maturation are gender specific and occur mostly with girls [207]. The problematic or beneficiary effects of early or late timing of onset or puberty are not a matter of a linear ordering of events. For instance, it is not the eventually early menarche itself that leads girls to experience adverse effects on other variables, such as psychosocial adaptation or birth weight of the first offspring. What matters is that such timing issues cause a problematic (or beneficiary) coordination, that is, mutual ordering, of different time lines or event sequences. Examples of such event sequences are timing of romantic dating and first sexual intercourse, the nature of the partners and potential providers who may or may not be able to support the offspring, and so forth. From the point of view of the participants, i. e. the people who are actually involved in these issues, timing relates to the subjective order of meaningful life-events in their own life and in the life of other persons with whom they are intimately connected (educators, parents, peers, romantic friends,...; see for instance [206]). In short, biological maturation during puberty provides an excellent example of timing problems in a complex dynamic system.

Autonomy, Identity and Connectedness The processes of biological and social development described in the preceding section are related to a renegotiation of the person's place in the familial, social and peer network. According to the historically important and psycho-analytically inspired theory of Erik Erikson (1902–1994), the main chal-

lenge during adolescence is to develop a personal identity (see [41,42] for an overview). For Erikson,"At times, identity refers to a structure or a configuration, at other points it refers to a process. Still on other occasions identity is viewed as both a conscious subjective experience as well as an unconscious entity" [179]. Erikson himself saw the conscious feeling of having a personal identity as "… based on two simultaneous observations: the perception of the selfsameness and continuity of one's one existence in time and space and the perception of the fact that others recognize one's sameness and continuity" [100]. Self-sameness means the existence of invariant properties over change that is both developmental (long-term) and actional (short-term). Erikson's view of identity is an example of an approach that respects the complex nature of the phenomenon under study and by doing so accepts the superposition of its features (it is both a process and an entity, it is an invariant defined by change, it is of the person but only insofar as it is also of other persons, and so on; see Sect. "A Working Definition of Complexity"). Although the concept of identity is an essential feature of the study of adolescent development, its complexity and in particular the acceptance of its complexity is a constant source of unease among researchers. Lichtwarck-Aschoff et al. [193] suggested a description – and the associated empirical study – of identity as a concept distributed across a two dimensional, categorical state space. One dimension concerns the distinction between dynamic and static approaches to phenomena, the other concerns the distinction between short- and long-term processes.

Identity development is closely related to the tension between autonomy and connectedness, both in terms of the short-term dynamics of action and social interaction and the long-term dynamics of development. As children become sexually mature they renegotiate their relationship with their parents by claiming a greater level of personal autonomy and self-determination. Meanwhile, if the relationship with the parents is positive, they do not want to loose the connectedness they feel with their parents, and these opposite tendencies create an interesting developmental dynamics. Lichtwarck-Aschoff et al. [193] have used this tension as the basic component of a long-term dynamic growth model of autonomy and connectedness, in order to account for the classical phenomenon of parent-child conflict during adolescence and for the considerable inter-individual variation in the form and outcome of this process (see also Sect. Development and Resource-Dependent Competition-Support Systems" for an explanation of the model; for comparable models see [180]).

Future Directions

Although human development is a prime example of a complex dynamic system, the theory of complex dynamic systems is not the mainstream approach to studying development. Human development is an example of one such system in which the observer – the researcher, the parent, the educator, … – are part and parcel of the complex system itself, thus leading to the epistemological and conceptual complexities and entanglements that are characteristic of systems in which the observer is an essential agent, in one way or another. Maybe it is because of this entanglement and epistemological complexity that psychology, and the study of development for that matter, still adheres to an approach of linearizing processes and disentangling causal contributions from identifiable factors. Changes are observable, for instance in the form of theories and approaches that focus on developmental systems, holistic interactions, dynamic systems, and so forth. Unfortunately, these approaches are still mainly theoretical: their contribution to changing the habits of empirical research and theory formation are at best very modest. The statistical sophistication that is required from most manuscripts that are submitted for publication in developmental journals stands in a glaring contrast to the almost complete lack of requirements regarding theoretical depth and reflection. In a book review, the philosopher and psychologist Harré referred to the situation as follows:

> It is a remarkable feature of mainstream academic psychology that, alone among the sciences, it should be almost wholly immune to critical appraisal as an enterprise. Methods that have long been shown to be ineffective or worse are still used on a routine basis by hundreds, perhaps thousands of people. Conceptual muddles long exposed to view are evident in almost every issue of standard psychology journals. ([138], p. 1303)

The blessing of our being so closely involved in the processes of development and education (all persons living have gone through a process of bio-psycho-social development themselves and cannot be but daily witnesses of ongoing developmental processes in others) is maybe also its curse. In order to study the process scientifically we probably feel obliged to discard all the intuitive knowledge and models we use in daily life and by doing so surrender to approaches that are critically incompatible with the complexity and dynamics of the phenomenon that we wish to understand. This problem is not easy to solve and will continue to be a very hard problem for the coming decades.

Meanwhile, it is the firm belief of the current author that progress can be made by reformulating the questions and methods on the basis of an approach of complex adaptive systems. The first steps that scholars in this field will have to take will most likely appear utterly naïve and limited to scholars studying other fields where conceptual and mathematical theory building has gone hand in hand with the development of rigorous empirical methods. Students of development will have to invest much more time and intellectual energy in the descriptive and exploratory study of individual developmental trajectories, studied with an intensity that is sufficient to capture the nature of the underlying process and with a thorough understanding of the fact that their focusing on the individual alone is in itself an important factor in the development of that individual. In addition, researchers should not be afraid – and in fact should promote the use – of so-called toy models that are highly simplified representations of the assumed, basic dynamics of some sort of phenomenon. It should not be expected that these toy models will provide us with stunning empirical fits with great data sets (but model fitting in itself is not the primary goal of scientific research, see [126]). However, if we ever wish to understand the intricacies of even relatively simple dynamic processes, we will have to study them in ways that can be made conceptually transparent, and this means, among others, taking toy models serious, as long as they are based on good theoretical considerations. An important consideration for a model is that it should be descriptively adequate, that is, that it captures the "essential" features of the phenomenon it addresses (the term descriptive adequacy stems from Chomskyan linguistics [61]). Since it is difficult to agree on what the essential features of some sort of phenomenon are, it is easier to define a model as not descriptively adequate if it leaves out or is incompatible with at least one feature of the phenomenon that is generally accepted as "essential". A descriptively adequate approach to human development and education should be one that is deeply compatible with the major features of development, namely its complexity and its dynamic nature.

Bibliography

Primary Literature

1. Abramovitch R, Grusec JE (1978) Peer imitation in a natural setting. Child Dev 49:60–65
2. Allen P, Strathern M (2003) Evolution, emergence, and learning in complex systems. Emerg 5(4):8–33
3. Anderson JR (1996) ACT: A simple theory of complex cognition. Am Psychol 51:355–365
4. Anderson JR, Bothell D, Byrne MD, Douglass S, Lebiere C, Qin Y (2004) An integrated theory of the mind. Psychol Rev 111(4):1036–1060
5. Asbury K, Wachs T, Plomin R (2005) Environmental moderators of genetic influence on verbal and nonverbal abilities in early childhood. Intell 33(6):643–661
6. Atkinson M (1986) Learning and models of development. In: van Geert P (ed) Theory building in developmental psychology. North Holland, Amsterdam, pp 105–139)
7. Axelrod R (1997) The complexity of cooperation: agent-based models of competition and collaboration. Princeton University Press, Princeton
8. Bailey DB, Bruer JT, Symons FJ, Lichtman JW (2001) Critical thinking about critical periods. Brookes Publishing, Baltimore
9. Bak P, Sneppen K (1993) Punctuated equilibrium and criticality in a simple model of evolution. Phys Rev Lett 71(24):4083–4086
10. Bak P (1996) How Nature works: the science of self-organized criticality. Springer, New York
11. Ball P (2005) Critical mass: how one thing leads to another. Arrow Books/Random House, London
12. Ballato L, van Geert P, Schumacher N (2007) A dynamic systems model of friendship interaction and risk behavior in adolescence: short-and-long term aspects of influence and selection mechanisms. Internal paper. Groningen: Department of Developmental Psychology
13. Baltes PB, Reese H, Lipsitt L (1980) Lifespan developmental psychology. Ann Rev Pyschol 31:65–110
14. Bandura A, Bussey K (2004) On broadening the cognitive, motivational, and sociostructural scope of theorizing about gender development and functioning: comment on Martin, Ruble, and Szkrybalo (2002). Psycholog Bull 130:691–701
15. Bandura A, Walters R (1977) Social learning and personality development. Holt, Rinehart and Winston, London
16. Bandura A, Ross D, Ross SA (1963) A comparative test of the status envy, social power, and secondary reinforcement theories of identificatory learning. J Abnorm Soc Psychol 67:527–534
17. Bankes S, Lempert R (2004) Robust reasoning with agent-based modeling. Nonlin Dyn Psychol Life Sci 8(2):259–277
18. Bankes S, Lempert R, Popper S (2002) Making computational social science effective: epistemology, methodology, and technology. Soc Sci Comput Rev 20(4):377–388
19. Barsalou L, Breazeal C, Smith L (2007) Cognition as coordinated non-cognition. Cogn Process 8(2):79–91
20. Bassano D, van Geert P (2007) Modeling continuity and discontinuity in utterance length: a quantitative approach to changes, transitions and intra-individual variability in early grammatical development. Dev Sci 10(5):588–612
21. Baron-Cohen S, Tager-Flusberg H, Cohen DJ (1993) The impairment of ToMM: some issues. In: Baron-Cohen S, Tager-Flusberg H, Cohen DJ (eds) Understanding other minds. Perspectives form autism. Oxford University Press, pp 102–105
22. Battro AM (2001) Half a brain is enough: the story of Nico. Cambridge University Press, New York
23. Baum WM, Davison M (2004) Choice in a variable environment: visit patterns in the dynamics of choice. J Exp Anal Behav 81:85–127
24. Belpaeme T, Cowley S (2007) Foreword: extending symbol grounding. Interact Stud 8(1):1–6

25. Berger KS (2003) The developing person through childhood and adolescence, 6th edn. Worth Publishers, New York
26. Berthouze L, Metta G (2005) Epigenetic robotics: modelling cognitive development in robotic systems, Cogn Syst Res 6(3):189–192
27. Berthouze L, Ziemke T (2003) Epigenetic robotics-modelling cognitive development in robotic systems. Connect Sci 15(4)147:150
28. Berthouze L, Metta G, Sun R (2005) Editorial: Epigenetic robotics: Modelling cognitive development in robotic systems. Cogn Syst Res 6(3):189–192
29. Bianchi-Berthouze N, Kleinsmith A (2003) A categorical approach to affective gesture recognition. Connect Sci 15(4):259–269
30. Billington E, DiTommaso N (2003) Demonstrations and applications of the matching law in education. J Behav Educ 12(2):91–104
31. Biro S, Leslie A (2007) Infants' perception of goal-directed actions: development through cue-based bootstrapping. Dev Sci 10(3):379–398
32. Bjorklund D, Pellegrini A (2000) Child development and evolutionary psychology. Child Dev 71(6):1687–1708
33. Bjorklund D, Smith P (2003) Evolutionary developmental psychology: Introduction to the special issue. J Exp Child Psychol 85(3):195–198
34. Boccia M, Pedersen C (2001) Animal models of critical and sensitive periods in social and emotional development. In: Critical thinking about critical periods. Brookes Publishing, Baltimore, pp 107–127
35. Boettcher S, Paczuski M (1996) Exact Results for Spatiotemporal Correlations in a Self-Organized Critical Model of Punctuated Equilibrium. Phys Rev Lett 76(3):348–351
36. Boettcher S, Percus A (2000) Nature's way of optimizing. Artif Intell 119:275–286
37. Bongers R, Smitsman A, Michaels C (2003) Geometrics and dynamics of a rod determine how it is used for reaching. J Motor Behav 35(1):4–22
38. Borrero J, Vollmer T (2002) An application of the matching law to severe problem behavior. J Appl Behav Anal 35(1):13–27
39. Borenstein E, Ruppin E (2005) The evolution of imitation and mirror neurons in adaptive agents, Cogn Syst Res 6(3):229–242
40. Bortfeld H, Whitehurst G (2001) Sensitive periods in first language acquisition. In: Critical thinking about critical periods. Brookes Publishing, Baltimore, pp 173–192
41. Bosma H, Kunnen E (2001) Determinants and mechanisms in ego identity development: a review and synthesis. Dev Rev 21(1):39–66
42. Bosma H, Kunnen E (2001) Identity and emotion: Development through self-organization. Cambridge University Press, New York
43. Boyce WT, Ellis BJ (2005) Biological sensitivity to context: I. An evolutionary-developmental theory of the origins and functions of stress reactivity. Dev Psychopathol 17:271–301
44. Breazeal C (2003) Emotion and sociable humanoid robots. Int J Human–Comput Stud 59(1):119–155
45. Breazeal C, Brooks R (2005) Robot emotion: a functional perspective. In: Who needs emotions: the brain meets the robot. Oxford University Press, New York, pp 271–310
46. Breazeal C, Scassellati B (2000) Infant-like social interactions between a robot and a human caregiver. Adapt Behav 8(1):49–74
47. Breazeal C, Scassellati B (2002) Robots that imitate humans. Trends Cogn Sci 6(11):481–487
48. Brighton H (2002) Compositional syntax from cultural transmission. Artif Life 8:25–54
49. Bruer J (2001) A critical and sensitive period primer. In: Critical thinking about critical periods. Brookes Publishing, Baltimore, pp 3–26
50. Bruer J (2002) Avoiding the pediatrician's error: how neuroscientists can help educators (and themselves). Nature Neurosci 5:1031–1033
51. Bukatko D, Daehler MW (2001) Child development: a thematic approach, 4th edn. Houghton, Mifflin and Company, Boston
52. Bussey K, Bandura A (1984) Influence of gender constancy and social power on sex-linked modeling. J Personal Soc Psychol 47:1292–1302
53. Call J, Carpenter M (2002) Three sources of information in social learning. In: Dautenhahn K, Nehaniv CL (eds) Imitation in animals and artifacts. MIT Press, Cambridge, pp 211–2228
54. Camras L, Witherington D (2005) Dynamical systems approaches to emotional development. Dev Rev 25(3):328–350
55. Camras L, Oster H, Bakeman R, Meng Z, Ujiie T, Campos J (2007) Do infants show distinct negative facial expressions for fear and anger? Emotional expression in 11-month-old European American, Chinese, and Japanese infants. Infancy 11(2):131–155
56. Cañamero L, Gaussier P (2005) Emotion understanding: robots as tools and models. In: Emotional development: recent research advances. Oxford University Press, New York, pp 235–258
57. Carpenter M, Call J, Tomasello M (2005) Twelve- and 18-month-olds copy actions in terms of goals. Dev Sci 8:13–20
58. Case R (1992) The mind's staircase: Exploring the conceptual underpinnings of children's thought and knowledge. Lawrence Earlbaum, Hillsdale
59. Cashon CH, Cohen LB (2004) Beyond U-shaped development in infants' processing of faces: an information-processing account. J Cogn Dev 5:59–80
60. Casti JL (1994) Complexification: explaining a paradoxical world through the science of surprise. Harper Collins, New York
61. Chomsky N (1957) Syntactic structures. The Hague, Mouton
62. Chomsky N (1959) A review of Bf Skinner's verbal behavior. Lang 35(1):26–58
63. Chomsky N (1965) Aspects of the theory of syntax. MIT Press, Oxford
64. Chomsky N (1995) The minimalist program. MIT Press, Cambridge MA
65. Christiansen MH, Dale R (2003) Language evolution and change. In: Arbib MA (ed) Handbook of brain theory and neural networks, 2nd edn. MIT Press, Cambridge, pp 604–606
66. Christiansen MH, Kirby S (2003) Language evolution: consensus and controversies. Trends Cogn Sci 7(7):300–307
67. Cilliers P (2002) Why we cannot know complex things completely. Emerg 4(1/2):77–84
68. Clark A (1997) Being there: putting brain, body and world together again. MIT Press, Cambridge
69. Clark A (2006) Language, embodiment, and the cognitive niche. Trends Cogn Sci 10:370–374

70. Clark A (2006) Material symbols. Philos Psychol 19(3):291–307
71. Clark A, Chalmers D (1998) The extended mind. Anal 58:7–19
72. Clowes R (2007) Semiotic symbols and the missing theory of thinking. Interact Stud 8(1):105–124
73. Cole M Cole SR (1993) The development of children, 2nd edn. Scientific American Books, New York
74. Colunga E, Smith L (2005, April) From the Lexicon to Expectations About Kinds: A Role for Associative Learning. Psychol Rev 112(2):347–382
75. Cowley S (2007) How human infants deal with symbol grounding. Interact Stud 8(1):83–104
76. Cruchtfield JP (1994) The calculi of emergence: computation, dynamics and induction. Physica D 75:11–54
77. Cziko G (2000) The things we do using the lessons of Bernard and Darwin to understand the what, how, and why of our behavior. MIT Press, Cambridge
78. Dahl R, Spear L (2004) Adolescent brain development: Vulnerabilities and opportunities. New York Academy of Sciences, New York
79. Davidson M, Amso D, Anderson L, Diamond A (2006) Development of cognitive control and executive functions from 4 to 13 years: evidence from manipulations of memory, inhibition, and task switching. Neuropsychol 44(11):2037–2078
80. Dawson-Tunik T, Commons M, Wilson M, Fischer K (2005, June) The shape of development. Eur J Dev Psychol 2(2):163–195
81. de Hamilton CA, Brindley R, Frith U (2007) Imitation and action understanding in autistic spectrum disorders: How valid is the hypothesis of a deficit in the mirror neuron system? Neuropsychologia 45(8):1859–1868
82. Demetriou A, Kyriakides L (2006, June) The functional and developmental organization of cognitive developmental sequences. Br J Educational Psychol 76(2):209–242
83. Demiris Y, Johnson M (2003) Distributed, predictive perception of actions: a biologically inspired robotics architecture for imitation and learning. Connect Sci 15(4):231–243
84. Diamond A, Barnett WS, Thomas J, Munro S (2007) The early years: preschool program improves cognitive control, Sci 318:1387–1388
85. Dishion T, Bullock B, Granic I (2002, September) Pragmatism in modeling peer influence: Dynamics, outcomes and change processes. Dev Psychopathol 14(4):969–981
86. Dominey PF (2007) Towards a construction-based framework for development of language, event perception and social cognition: insights from grounded robotics and simulation. Neurocomput 70(13):2288–2302
87. Dominey PF, Boucher J-D (2005) Developmental stages of perception and language acquisition in a perceptually grounded robot, Cogn Syst Res 6(3):243–259
88. Dooley K (2002) Organizational complexity In: Warner M (ed) International encyclopedia of business and management. Thompson Learning, London, pp 5013–5022
89. Dooley K, Corman S (2002) Agent-based, genetic, and emergent computational models of complex systems. In: Kiel LD (ed) Encyclopedia of life support systems (EOLSS). UNESCO/EOLSS, Oxford UK, http://www.eolss.net
90. Duncan G, Dowsett C, Claessens A, Magnuson K, Huston A, Klebanov P et al (2007) School readiness and later achievement. Dev Psychol 43(6):1428–1446
91. Eggleston R, McCreight K, Young M (2005) Distributed cognition and situated behavior. Modeling human behavior with integrated cognitive architectures: comparison, evaluation, and validation. Lawrence Erlbaum, Mahwah, pp 177–235
92. Elbers E (2004) Conversational asymmetry and the child's perspective in developmental and educational research. Int J Disabil Dev Educ 51(2):201–215
93. Eley T, Sugden K, Corsico A, Gregory A, Sham P, McGuffin P et al (2004) Gene-environment interaction analysis of serotonin system markers with adolescent depression. Mol Psychiatry 9(10):908–915
94. Elliott E, Kiel L (2004) Agent-based modeling in the social and behavioral sciences. Nonlin Dyn Psychol Life Sci 8(2):121–130
95. Ellis BJ (2004) Timing of pubertal maturation in girls: An integrated life history approach. Psychol Bull 130:920–958
96. Ellis B, Bjorklund D (2005) Origins of the social mind: evolutionary psychology and child development. Guilford Press, New York
97. Ellis BJ, Essex MJ, Boyce WT (2005) Biological sensitivity to context: II. Empirical explorations of an evolutionary-developmental theory. Dev Psychopathol 17:303–328
98. Eoyang G (2004) The practitioner's landscape. Emerg: Complex Organ 6(1/2):55–60
99. Erlhagen W, Schöner G (2002, July) Dynamic field theory of movement preparation. Psychol Rev 109(3):545–572
100. Erikson EH (1968) Identity: youth and crisis. Norton, New York
101. Fioretti G, Visser B (2004) A cognitive interpretation of organizational complexity. Emerg: Complex Organ 6(1/2):11–23
102. Fischer KW (1980) A theory of cognitive development: the control and construction of hierarchies of skills. Psychol Rev 87:477–531
103. Fischer KW, Bidell TR (2006) Dynamic development of action, thought, and emotion. In: Damon W, Lerner RM (eds) Theoretical models of human development. Handbook of child psychology, vol 1, 6th edn. Wiley, New York, pp 313–399
104. Fischer KW, Kennedy B (1997) Tools for analyzing the many shapes of development: the case of self-in-relationships in Korea. In: Change and development: issues of theory, method, and application. Lawrence Erlbaum, Mahwah, pp 117–152
105. Fischer KW, Rose SP (1999) Rulers, models, and nonlinear dynamics: measurement and method in developmental research. In: Savelsbergh G, van der Maas H, van Geert P (eds) Nonlinear developmental processes, Royal Netherlands Academy of Arts and Sciences, Amsterdam, pp 197–212
106. Fitzpatrick P, Arsenio A, Torres-Jara E (2006) Reinforcing robot perception of multi-modal events through repetition and redundancy and repetition and redundancy. Interact Stud 7(2):171–196
107. Fogel A (1993) Developing through relationships: Origins of communication, self, and culture. University of Chicago Press, Chicago
108. Ford DH Lerner RM (1992) Developmental systems theory: an integrative approach. Sage Publications, Thousand Oaks
109. Förster J, Liberman N, Friedman R (2007) Seven principles of goal activation: a systematic approach to distinguishing goal priming from priming of non-goal constructs. Personal Soc Psychol Rev 11(3):211–233
110. Frith U, Frith CD (2003) Development and neurophysiology of mentalizing. Philos Trans Royal Soc Lond Biol Sci 358:459–473
111. Frith U, Hill E (2003) Autism: Mind and brain. Oxford University Press, New York

112. Frith U, Morton J, Leslie A (1991) The cognitive basis of a biological disorder: autism. Trends Neurosci 14(10):433–438
113. Gallese V (2006) Intentional attunement: A neurophysiological perspective on social cognition and its disruption in autism. Brain Res 1079:15–24
114. Gallistel CR, Gibbon J (2000) Time, rate, and conditioning. Psychol Rev 107:289–344
115. Geary D (2006) Evolutionary developmental psychology: current status and future directions. Dev Rev 26(2):113–119
116. Gergely G (2003) What should a robot learn from an infant? Mechanisms of action interpretation and observational learning in infancy. Connect Sci 15(4):191–209
117. Gershkoff-Stowe L, Thelen E (2004) U-shaped changes in behavior: a dynamic systems perspective. J Cogn Dev 5:11–36
118. Gibson J (1966) The senses considered as perceptual systems. Houghton Mifflin, Oxford
119. Goldin-Meadow S, Wagner S (2005) How our hands help us learn. Trends Cogn Sci 9(5):234–241
120. Goldspink C (2007) Transforming education: evidential support for a complex systems approach. Emerg: Complex Organ 9(1/2):77–92
121. Goldstein J (1999) Emergence as a construct: History and issues. Emergence 1:49–72
122. Goldstein J (2002, October) The Singular Nature of Emergent Levels: Suggestions for a Theory of Emergence. Nonlinear Dyn Psychol Life Sci 6(4):293–309
123. Goldstone R, Janssen M (2005) Computational models of collective behavior. Trends Cogn Sci 9(9):424–430
124. Gottfried AW, Gottfried AE (1974) Influence of social power vs status envy modeled behaviors on children's preferences for models. Psycholog Rep 34:1147–1150
125. Gottlieb G (2001) A developmental psychobiological systems view: early formulation and current status. MIT Press, Cambridge
126. Gottman JM, Murray JD, Swanson CC, Tyson R, Swanson KR (2002) The mathematics of marriage: dynamic nonlinear models. MIT Press, Cambridge
127. Granic I, Hollenstein T (2003) Dynamic systems methods for models of developmental psychopathology. Dev Psychopathol 15(3):641–669
128. Granic I, Lamey A (2002) Combining dynamic systems and multivariate analyses to compare the mother-child interactions of externalizing subtypes. J Abnorm Child Psychol 30(3):265–283
129. Granic I, Hollenstein T, Dishion T, Patterson G (2003) Longitudinal analysis of flexibility and reorganization in early adolescence: a dynamic systems study of family interactions. Dev Psychol 39(3):606–617
130. Granott N (2002) How microdevelopment creates macrodevelopment: Reiterated sequences, backward transitions, and the Zone of Current Development. Microdevelopment: Transition processes in development and learning. Cambridge University Press, New York, pp 213–242
131. Grusec JE (1992) Social learning theory and developmental psychology: the legacies of Robert Sears and Albert Bandura. Dev Psychol 28:776–786
132. Grusec JE, Abramovitch R (1982) Imitation of peers and adults in a natural setting: a functional analysis. Child Dev 53:636–642
133. Guerin S, Kunkle D (2004) Emergence of constraint in self-organizing systems. Nonlin Dyn Psychol Life Sci 8(2):131–146
134. Gulyás L (2002) On the transition to agent-based modeling: Implementation strategies from variables to agents. Soc Sci Comput Rev 20(4):389–399
135. Hadjikhani N, Joseph RM, Snyder J, Tager-Flusberg H (2006) Anatomical differences in the mirror neuron system and social cognition network in autism. Cereb Cortex 16:1276–1282
136. Hamaker EL, Dolan CV, Molenaar PCM (2005) Statistical modeling of the individual: rationale and application of multivariate stationary time series analysis. Multivar Behav Res 40:207–233
137. Hammond S, Sanders M (2002) Dialogue as social self-organization: an introduction. Emerg 4(4):7–24
138. Harré R (2000) Acts of living. Sci 289(25):1303–1304
139. Hartelman PA, van der Maas HLJ, Molenaar PCM (1998) Detecting and modelling developmental transitions. Br J Dev Psychol 16:97–122
140. Henrickson L (2002) Old wine in a new wineskin: college choice, college access using agent-based modeling. Soc Sci Comput Rev 20(4):400–419
141. Hernández Blasi C, Bjorklund D (2003) Evolutionary developmental psychology: a new tool for better understanding human ontogeny. Human Dev 46(5):259–281
142. Herrnstein RJ (1961) Relative and absolute strength of response as a function of frequency of reinforcement. J Exp Anal Behav 4:267–272
143. Herrnstein RJ (1970) On the law of effect. J Exp Anal Behav 13:243–266
144. Herrnstein RJ, Prelec D (1991) Melioration: A theory of distributed choice. J Econom Perspect 5:137–156
145. Hodgson G (2000) The concept of emergence in social sciences: its history and importance. Emerg 2(4):65–77
146. Holland JH (1995) Hidden order: How adaptation builds complexity. Addison-Wesley, Reading
147. Holland JH (1998) Emergence. From chaos to order. Addison-Wesley, Redwood City
148. Hollenstein T (2007) State space grids: analyzing dynamics across development. Int J Beh Dev 31(4):384–396
149. Hollenstein T, Lewis M (2006) A state space analysis of emotion and flexibility in parent-child interactions. Emot 6(4):656–662
150. Hosenfeld B, van der Maas HLJ, van den Boom DC (1997a) Detecting bimodality in the analogical reasoning performance of elementary schoolchildren. Int J Behav Dev 20:529–547
151. Hosenfeld B, van der Maas HLJ, van den Boom DC (1997b) Indicators of discontinuous change in the development of analogical reasoning. J Exp Child Psychol 64:367–395
152. Howe ML, Lewis MD (2005) The importance of dynamic systems approaches for understanding development. Dev Rev 25:247–251
153. Iacoboni M, Dapretto M (2006) The mirror neuron system and the consequences of its dysfunction. Nat Rev Neurosci 7:942–951
154. Immordino-Yang MH (2005) A tale of two cases: emotion and affective prosody after hemispherectomy. ProQuest Information Learning
155. Immordino-Yang MH (2007) A tale of two cases. Lessons for education from the study of two boys living with half their brains. Mind Brain Educ 1(2):66–83
156. Iverson J, Goldin-Meadow S (2005) Gesture paves the way for language development. Psychol Sci 16(5):367–371
157. Jansen BRJ, van der Maas HLJ (2001b) Evidence for the Phase

transition from rule I to rule II on the balance scale task. Developmental Review, 21, 450–494
158. Jansen BRJ, van der Maas HLJ (2002b) The development of children's rule use on the balance scale task. J Exp Child Psychol 81:383–416
159. Johnson M (2005) Sensitive periods in functional brain development: problems and prospects. Dev Psychobiol 46(3):287–292
160. Jones G (2007) Agent-based modeling: Use with necessary caution. Am J Public Health 97(5):780–781
161. Jones G, Ritter FE, Wood DJ (2000) Using a cognitive architecture to examine what develops. Psychol Sci 11(2):1–8
162. Kail RV (2001) Children and their development, 2nd edn. Prentice-Hall, Englewood Cliffs
163. Kaplan F, Hafner V (2006) The challenges of joint attention. Interact Stud 7(2):135–169
164. Ke J, Holland J (2006) Language origin from an emergentist perspective. Appl Linguist 27(4):691–716
165. Kello CT (2004) Characterizing the evolutionary dynamics of language. Trends Cogn Sci 8:392–394
166. Kendal J (2006) Review of: Galef BG Jr, Heyes CM (2004) Social learning and imitation. Learn Behav 32(1):1–140. Interact Stud 7(2):273–288
167. Kerpelman J, Pittman J, Lamke L (1997) Toward a microprocess perspective on adolescent identity development: an identity control theory approach. J Adolesc Res 12(3):325–346
168. Kirby S (2002) Natural language from artificial life. Artif Life 8:185–215
169. Kirby S, Smith K, Brighton H (2004) From UG to universals: linguistic adaptation through iterated learning. Stud Lang 28:587–607
170. Kitchener K, Lynch C, Fischer K, Wood P (1993) Developmental range of reflective judgment: the effect of contextual support and practice on developmental stage. Dev Psychol 29(5):893–906
171. Klahr D, Wallace JG (1976) Cognitive development: an information processing approach. Lawrence Erlbaum, Hillsdale
172. Knafo A, Plomin R (2006) Prosocial behavior from early to middle childhood: genetic and environmental influences on stability and change. Dev Psychol 42(5):771–786
173. Knudsen E (2004) Sensitive periods in the development of the brain and behavior. J Cogn Neurosci 16(8):1412–1425
174. Kobayashi C, Glover GH, Temple E (2007) Children's and adults' neural bases of verbal and nonverbal 'theory of mind'. Neuropsychologia, 45, 1522–1532
175. Kolb B, Whishaw IQ (1998) Brain plasticity and behavior. Ann Rev Psychol 49:43–64
176. Kouritzin S (1999) Face[t]s of first language loss. Lawrence Erlbaum, Mahwah
177. Kovas Y, Plomin R (2007) Learning abilities and disabilities: generalist genes, specialist environments. Current Dir Psychol Sci 16(5):284–288
178. Kovas Y, Petrill S, Plomin R (2007) The origins of diverse domains of mathematics: generalist genes but specialist environments. J Educ Psychol 99(1):128–139
179. Kroger J (2004) Identity in adolescence. The balance between self and other, 3rd edn. Routledge, East Sussex
180. Kunnen E, Bosma H (2000) Development of meaning making: a dynamic systems approach. New Ideas Psychol 18(1):57–82
181. Lasnik H (2002) The minimalist program in syntax. Trends Cogn Sci 6(10):432–437
182. Lee K, Park N, Song H (2005) Can a robot be perceived as a developing creature? Effects of a robot's long-term cognitive developments on its social presence and people's social responses toward it. Human Commun Res 31(4):538–563
183. Lepage JF, Théoret H (2006) EEG evidence for the presence of an action observation-execution matching system in children. Eur J Neurosci 23:2505–2510
184. Lerner RM (2006) Developmental science, developmental systems, and contemporary theories of human development. Wiley, Hoboken
185. Lewis M (1995, March) Cognition-emotion feedback and the self-organization of developmental paths. Hum Dev 38(2):71–102
186. Lewis M (2000, January) The promise of dynamic systems approaches for an integrated account of human development. Child Dev 71(1):36–43
187. Lewis MD (2005) Bridging emotion theory and neurobiology through dynamic systems modeling. Behav Brain Sci 28:169–245
188. Lewis MD (2005) Self-organizing individual differences in brain development. Dev Rev 25(3):252–277
189. Lewis MD, Todd R (2007) The self-regulating brain: cortical-subcortical feedback and the development of intelligent action. Cogn Dev 22(4):406–430
190. Lewis MD, Lamey A, Douglas L (1999) A new dynamic systems method for the analysis of early socioemotional development. Dev Sci 2(4):457–475
191. Lewis MD, Zimmerman S, Hollenstein T, Lamey A (2004) Reorganization in coping behavior at 1½ years: Dynamic systems and normative change. Dev Sci 7(1):56–73
192. Lewis T, Maurer D (2005) Multiple sensitive periods in human visual development: evidence from visually deprived children. Dev Psychobiol 46(3):163–183
193. Lichtwarck-Aschoff A, van Geert P, Bosma H, Kunnen S (2008) Time and identity: a framework for research and theory formation. Dev Rev 28(3):370–400
194. Lickliter R, Honeycutt H (2003) Developmental dynamics: toward a biologically plausible evolutionary psychology. Psycholog Bull 129:819–835
195. Lickliter R, Honeycutt H (2003) Evolutionary approaches to cognitive development: status and strategy. J Cogn Dev 4(4):459–473
196. Lieberman MD (2007) Social cognitive neuroscience: A review of core processes. Ann Rev Psychol 58:259–289
197. Lindblom J, Ziemke T (2006) The social body in motion: cognitive development in infants and androids. Connect Sci 18(4):333–346
198. Locke J (2007) Bimodal signaling in infancy: Motor behavior, reference, and the evolution of spoken language. Interact Stud 8(1):159–175
199. Lockman JJ (2000) A perception–action perspective on tool use development. Child Dev 71(1):137–144
200. Lopes L, Chauhan A (2007) How many words can my robot learn? An approach and experiments with one-class learning. Interact Stud 8(1):53–81
201. Louro M, Pieters R, Zeelenberg M (2007) Dynamics of multiple-goal pursuit. J Personal Soc Psychol 93(2):174–193
202. Lungarella M, Metta G, Pfeifer R, Sandini G (2003) Developmental robotics: a survey. Connect Sci 15(4):151–190

203. MacDorman K (2007) Afterword: life after the symbol system metaphor. Interact Stud 8(1):1–999
204. MacDorman K, Ishiguro H (2006) Toward social mechanisms of android science: a CogSci Workshop, 25 and 26 July 2005, Stresa, Italy. Interact Stud 7(2):289–296
205. Maestripieri D, Roney J (2006,). Evolutionary developmental psychology: contributions from comparative research with nonhuman primates. Dev Rev 26(2):120–137
206. Magnusson D (1999) On the individual: a person-oriented approach to developmental research. Eur Psychol 4(4):205–218
207. Magnusson D (2001) The holistic-interactionistic paradigm: some directions for empirical developmental research. Eur Psychol 6(3):153–162
208. Martin LC, Fabes RA, Hanish LD, Hollenstein T (2005) Social dynamics in the preschool. Dev Rev 25:299–327
209. Masten A (2007) Resilience in developing systems: progress and promise as the fourth wave rises. Dev Psychopath 19(3):921–930
210. McShane J (1991) Cognitive development: an information processing approach. Blackwell, Oxford
211. Metta G, Berthouze L (2006) Epigenetic robotics: modelling cognitive development in robotic systems. Interact Stud 7(2):129–134
212. Metta G, Sandini G, Natale L, Craighero L, Fadiga L (2006) Understanding mirror neurons: a bio-robotic approach. Interact Stud 7(2):197–232
213. Mezic I, Gruenewald P, Gorman D, Mezic J (2007) Agent-based modeling: use with necessary caution: reply. Am J Public Health 97(5):781–782
214. Mitchell P (1997) Introduction to theory of mind. Children, autism and apes. Arnold, Londen
215. Moldoveanu M (2004) An intersubjective measure of organizational complexity: a new approach to the study of complexity in organizations. Emerg: Complex Organ 6(3):9–26
216. Molenaar P (2004) A manifesto on psychology as idiographic science: bringing the person back into scientific psychology, this time forever. Meas: Interdiscipl Res Perspect 2(4):201–218
217. Molenaar P, Huizenga H, Nesselroade J (2003) The relationship between the structure of interindividual and intraindividual variability: a theoretical and empirical vindication of developmental systems theory. In: Understanding human development: dialogues with lifespan psychology. Kluwer, Dordrecht
218. Montague DPF, Walker-Andrews AS (2001) Peekaboo: a new look at infants' perception of emotion expressions. Dev Psychol 37:826–838
219. Moore D (2003) Trying to fix the development in evolutionary developmental psychology. Am J Psychol 116(2):299–307
220. Mosekilde E (1998) Topics in nonlinear dynamics: applications to physics biology and economic systems. World Scientific, Singapore
221. Moss S, Edmonds B (2005) Sociology and simulation: statistical and qualitative cross-validation. Am J Sociol 110(4):1095–1131
222. Munakata Y, Casey BJ, Diamond A (2004) Developmental cognitive neuroscience: progress and potential. Trends Cogn Sci 8(3): 122–128
223. Murray L, Kollins S (2000) Effects of methylphenidate on sensitivity to reinforcement in children diagnosed with attention deficit hyperactivity disorder: an application of the matching law. J Appl Behav Anal 33(4):573–591
224. Musher-Eizenman DR, Nesselroade JR, Schmitz B (2002) Perceived control and academic performance: a comparison of high- and low-performing children on within-person change patterns. Int J Behav Dev 26:540–547
225. Nagai Y, Hosoda K, Morita A, Asada M (2003) A constructive model for the development of joint attention. Connect Sci 15(4):211–229
226. Namy LL, Campbell AL, Tomasello M (2004) The changing role of iconicity in non-verbal symbol learning: a U-shaped trajectory in the acquisition of arbitrary gestures. J Cogn Dev 5:37–57
227. Newman BM Newman PR (2006) Development through life: a psychosocial approach. Thomson Wadsworth, Belmont
228. Niyogi P, Berwick RC (2001) A dynamical systems model for language change. Complex Syst 11(3):161–204
229. Nowak MA, Komarova NL (2001) Towards an evolutionary theory of language. Trends Cogn Sci 5(7):288–295
230. Nowak MA, Komarova NL, Niyogi P (2001) Evolution of universal grammar. Sci 291:114–118
231. Nowak MA, Komarova NL, Niyogi P (2002) Computational and evolutionary aspects of language. Nature 417:611—617
232. Oberman LM, Ramachandran VS (2007) The simulating social mind: The role of the mirror neuron system and simulation in the social and communicative deficits of autism spectrum disorders. Psychol Bull 133:310–327
233. Oyama S, Griffiths PE, Gray RD (2001) Cycles of contingency: developmental systems and evolution. MIT Press, Cambridge
234. Pan B, Gleason J (1986) The study of language loss: models and hypotheses for an emerging discipline. Appl Psycholinguist 7(3):193–206
235. Pellegrini A, Bjorklund D (2004) The ontogeny and phylogeny of children's object and fantasy play. Human Nature 15(1):23–43
236. Pepper SC (1926) Emergence, The J Philos 23(9):241–245
237. Pepper SC (2004) Emergence. Emerg: Complex Organ 6(4):66–245
238. Pessa E (2004) Quantum connectionism and the emergence of cognition. In: Globus GG, Pribram KH, Vitiello G (eds) Brain and being: at the boundary between science, philosophy, language and arts. Benjamins, Amsterdam, pp 129–148
239. Bjorne P, Balkenius C (2005) A model of attentional impairments in autism: first steps toward a computational theory. Cogn Syst Res 6(3):193–204
240. Phillips AT, Wellman HM, Spelke ES (2002) Infants' ability to connect gaze and emotional expression to intentional action. Cogn 85:53–78
241. Piaget J (1950) Introduction à l'épistémologie génétique; I La pensée mathématique. – II La pensée physique. – III La pensée biologique, la pensée psychologique et la pensée sociologique. Presses Universitaires de France, Paris
242. Pincus D, Guastello SJ (2005) Nonlinear Dynamics and Interpersonal Correlates of Verbal Turn-Taking Patterns in a Group Therapy Session. Small Group Res 36:635–677
243. Plomin R (2004) Genetics and developmental psychology. Merrill-Palmer Quarterly 50(3):341–352
244. Plomin R (2007) Genetics and developmental psychology. Wayne State University Press, Detroit
245. Plomin R, Asbury K (2005) Nature and nurture: genetic and environmental influences on behavior. Ann Am Acad Political Soc Sci 600:86–98

246. Plomin R, Kovas Y (2005) Generalist genes and learning disabilities. Psycholog Bull 131(4):592–617
247. Popper KR (1976) Unended quest: an intellectual autobiography. Fontana, London
248. Prince CG, Hollich GJ (2005) Synching models with infants: a perceptual-level model of infant audio-visual synchrony detection. Cogn Syst Res 6(3):205–228
249. Raczynski S (2006) A self-destruction game. Nonlin Dyn Psychol Life Sci 10(4):471–483
250. Richardson K (2004) Systems theory and complexity: Part 1. Emerg: Complex Organ 6(3):75–79
251. Rieber RW, Carton AS (eds) (1987) The collected works of LS Vygotsky: vol I Problems of general psychology. Plenum, New York, pp 37–285
252. Ripa J, Ives AR (2003) Food web dynamics in correlated and autocorrelated environments, Theor Popul Biol 64(3):369–384
253. Robertson D (2005) Review of agent-based modeling toolkits. Acad Manag Learn Educ 4(4):525–527
254. Robinson BF, Mervis CB (1998) Disentangling early language development: modeling lexical and grammatical acquisition using and extension of case-study methodology. Dev Psychol 34:363–375
255. Rogoff B (1990) Apprenticeship in thinking. Oxford University Press, New York
256. Ronald A, Happé F, Plomin R (2006) Genetic research into autism. Sci 311(5763):952–952
257. Rose SP, Fischer KW (1998) Models and rulers in dynamical development. Brit J Dev Psychol 16:123–131
258. Rubino C (2002) The consolations of uncertainty: time, change, and complexity. Emerg 4(1/2):200–206
259. Ruhland R, van Geert P (1998) Jumping into syntax. Brit J Dev Psychol 16:65–95
260. Rushton JP (1975) Generosity in children: immediate and long-term effects of modeling, preaching, and moral judgment. J Personal Soc Psychol 31:459–466
261. Rutter M (2007) Gene-environment interdependence. Dev Sci 10:1–12
262. Sallach D (2003) Social theory and agent architectures: prospective issues in rapid-discovery social science. Soc Sci Comput Rev 21:179–195
263. Sameroff AJ (1975) Transactional models in early social relations. Human Dev 18(1):65–79
264. Sameroff AJ (2000) Developmental systems and psychopathology. Dev Psychopathol 12:297–312
265. Sameroff AJ, MacKenzie M (2003) Research strategies for capturing transactional models of development: the limits of the possible. Dev Psychopathol 15(3):613–640
266. Saxe R, Wexler A (2005) Making sense of another mind: The role of the right temporo-parietal junction. Neuropsychologia 43:1391–1399
267. Schlesinger M (2003) A lesson from robotics: modeling infants as autonomous agents. Adapt Behav 11(2):97–107
268. Schlesinger M, Casey P (2003) Where infants look when impossible things happen: simulating and testing a gaze-direction model. Conn Sci 15(4):271–280
269. Schlesinger M, Parisi D (2001) The agent-based approach: a new direction for computational models of development. Dev Rev 21(1):121–146
270. Schlindwein S, Ison R (2004) Human knowing and perceived complexity: implications for systems practice. Emerg: Complex Organ 6(3):27–32
271. Schöner G, Dineva E (2007) Dynamic instabilities as mechanisms for emergence. Dev Sci 10(1):69–74
272. Schöner G, Thelen E (2006) Using dynamic field theory to rethink infant habituation. Psychol Rev 113:273–299
273. Schutte AR, Spencer JP (2002) Generalizing the dynamic field theory of the A-not-B error beyond infancy: three-year-olds' delay- and experience-dependent location memory biases. Child Dev 73:377–404
274. Schutte AR, Spencer JP, Schöner G (2003) Testing the dynamic field theory: working memory for locations becomes more spatially precise over development. Child Dev 74(5):1393–1417
275. Sebeok TA (2001) Biosemiotics: its roots, proliferation, and prospects. Semiot 134:61–78
276. Seifert KL, Hoffnung RJ (1991) Child and adolescent development, 2nd edn. Houghton, Mifflin and Company, Boston
277. Shahan T, Podlesnik C (2007) Divided attention and the matching law: sample duration affects sensitivity to reinforcement allocation. Learn Behav 35(3):141–148
278. Shriver M, Kramer J (1997) Application of the generalized matching law for description of student behavior in the classroom. J Behav Educ 7(2):131–149
279. Siegler RS (1983) Information processing approaches to development. In: Mussen PH (ed) Recent advances in cognitive developmental theory. Wiley, New York
280. Sigelman CK, Rider EA (2006) Life-span human development, 5th edN. Thomson Wadsworth, Belmont
281. Simon H (1962) The architecture of complexity. Proc Am Philos Soc ISSN 0003–049X 106(6):467–482
282. Simon H, Cilliers P (2005) The architecture of complexity. Emerg: Complex Organ, 7(3/4), 138–154
283. Skinner BF (1938) The behavior of organisms: an experimental analysis. Appleton, New York
284. Skinner BF (1945) The operational analysis of psychological terms. Psychol Rev :52:270–77
285. Skinner BF (1950) Are theories of learning necessary? Psychol Rev 57:193–216
286. Skinner C, Robinson S, Johns G, Logan P, Belfiore P (1996) Applying Herrnstein's matching law to influence students' choice to complete difficult academic tasks. J Exp Educ 65(1):5–17
287. Smith L (2005, September) Cognition as a dynamic system: Principles from embodiment. Dev Rev 25(3):278–298
288. Smith E, Conrey F (2007) Agent-based modeling: a new approach for theory building in social psychology. Personal Soc Psychol Rev 11(1):1–18
289. Smith K (2002) Natural selection and cultural selection in the evolution of communication. Adapt Behav 10(1):25–44
290. Smith K, Kirby S, Brighton H (2003) Iterated learning: a framework for the emergence of language. Artif Life 9:371—386
291. Smith LB (1999) Not 'either', not 'or', not 'both'. Dev Sci 2(2):162–163
292. Smith LB, Thelen E, Titzer R, McLin D (1999) Knowing in the context of acting: the task dynamics of the A-not-B error. Psychol Rev 106:235–260
293. Smith LB, Breazeal C (2007) The dynamic lift of developmental process. Dev Sci 10(1):61–68

294. Spear L (2000, August) Neurobehavioral changes in adolescence. Curr Dir Psychol Sci 9(4):111–114
295. Spelke E, Kinzler K (2007) Core knowledge. Dev Sci 10(1):89–96
296. Spencer JP, Schöner G (2003) Bridging the representational gap in the dynamic systems approach to development. Dev Sci 4(6):392–412
297. Spencer JP, Smith LB, Thelen E (2001) Tests of a dynamic systems account of the A-not-B error: The influence of prior experience on the spatial memory abilities of two-year-olds. Child Dev 72(5):1327–1346
298. Stanislas D (2007) A few steps toward a science of mental life. Mind Brain Educ 1(1):28–47
299. Steenbeek H (2006) Modeling dyadic child-peer interactions: sociometric status, emotional expressions and instrumental actions during play. Doctoral Dissertation, University of Groningen
300. Steenbeek H, van Geert P (2005) A dynamic systems model of dyadic interaction during play of two children. Eur J Dev Psychol 2(2):105–145
301. Steenbeek H, van Geert P (2007) Do you still like to play with him? Variability and the dynamic nature of children's sociometric ratings. Neth J Psychol 63(3):86–101
302. Steenbeek H, van Geert P (2007) The empirical validation of a dynamic systems model of interaction; do children of different sociometric statuses differ in their dyadic play interactions? Dev Sci 11(2):253–281
303. Steenbeek H, van Geert P (2008) A dynamic systems approach to dyadic interaction in children: emotional expression, action, dyadic play, and sociometric status. Dev Rev 27(1):1–40
304. Sterman J (1994) Learning in and about complex systems. Syst Dyn Rev 10(2):291–330
305. Strauss S, Stavy R (1982) U-shaped behavioral growth. Academic, New York
306. Striano T, Henning A, Stahl D (2006) Sensitivity to interpersonal timing at 3 and 6 months of age. Interact Stud 7(2):251–271
307. Striano T, Henning A, Vaish A (2006) Selective looking by 12-month-olds to a temporally contingent partner. Interact Stud 7(2):233–250
308. Sulis W (2004) Archetypal dynamical systems and semantic frames in vertical and horizontal emergence. Emerg: Complex Organ 6(3):52–64
309. Sun R (2001) Cognitive science meets multi-agent systems: a prolegomenon. Philos Psychol 14:5–28
310. Taptiklis T (2005) After managerialism. Emerg: Complex Organ 7(3/4):2–14
311. Thagard P (1996) Mind: introduction to cognitive science. MIT Press, Cambridge
312. Thelen E, Fisher D (1982) Newborn stepping: an explanation for a 'disappearing' reflex. Dev Psychol 18(5):760–775
313. Thelen E, Schöner G, Scheier C, Smith L (2001) The dynamics of embodiment: A field theory of infant perseverative reaching. Behav Brain Sci 24(1):1–86
314. Thelen E, Smith L (1994) A dynamic systems approach to the development of cognition and action. MIT Press, Cambridge
315. Thelen E, Whitmyer V (2005) Using dynamic field theory to conceptualize the interface of perception, cognition, and action. In: Action as an organizer of learning and development, vol 33 in the Minnesota Symposia on Child Psychology. Lawrence Erlbaum, Mahwah, pp 243–280
316. Thom R (1990) Semiophysics: a sketch, Addison-Wesley, Redwood City
317. Thomae H (1959) Entwicklungsbegriff und Entwicklungstheorie. In: Bergius R, Thomae H (eds) Handbuch der Psychologie, Band 3: Entwicklungspsychologie. Verlag für Psychologie, Göttingen
318. Thomas A, Chess S (1977) Temperament and development. Brunner/Mazel, Oxford
319. Thomas M, Johnson M (2006) The computational modeling of sensitive periods. Dev Psychobiol 48(4):337–344
320. Thompson DE, Russell J (2004) The ghost condition: imitation versus emulation in young children's observational learning. Dev Psychol 40:882–889
321. Thompson R (2001) Sensitive periods in attachment? In: Critical thinking about critical periods. Brookes Publishing, Baltimore, pp 83–106
322. Tschacher W, Haken H (2007) Intentionality in non-equilibrium systems? The functional aspects of self-organized pattern formation. New Ideas Psychol 25:1–15
323. Uylings H (2006) Development of the human cortex and the concept of 'critical' or 'sensitive' periods. Lang Learn 56(1):59–90
324. Valsiner J (1987) Culture and the development of children's action: A cultural-historical theory of developmental psychology. Wiley, Oxford
325. van der Maas HL (1993) Catastrophe analysis of stagewise cognitive development. Faculty of Psychology, University of Amsterdam, Amsterdam (Academic dissertation)
326. van der Maas H, Jansen B, Raijmakers M (2004) Developmental patterns in proportional reasoning. Cambridge University Press, New York
327. van der Maas HL, Molenaar PC (1992) Stagewise cognitive development: An application of catastrophe theory. Psychol Rev 99:395–417
328. van der Veer R, Valsiner J (1991) Understanding Vygotsky: a quest for synthesis. Blackwell, Oxford
329. van Dijk M, van Geert P (2007) Wobbles, humps and sudden jumps: a case study of continuity, discontinuity and variability in early language development. Infant Child Dev 16(1):7–33
330. van Geert P (1986) The concept of development. In: van Geert P (ed) Theory building in developmental psychology. North Holland, Amsterdam, pp 3–50
331. van Geert P (1987) Developmental theories as cognitive systems. Cogn Syst 2:5–39
332. van Geert P (1987) The concept of development and the structure of developmental theories. In: Baker WJ, Hyland ME (eds) Current issues in theoretical psychology: selected/edited proceedings of the founding conference of the International Society for Theoretical Psychology held in Plymouth UK, 30 August–2 September 1985. North-Holland, Oxford, pp 379–392
333. van Geert P (1987) The structure of developmental theories. A generative approach. Hum Dev 30(3):160–177
334. van Geert P (1987) The structure of Erikson's model of the Eight Ages of Man. A generative approach. Hum Dev 30(5):236–254
335. van Geert P (1987) The structure of Galperin's theory of the formation of mental acts: a generative approach. Human Dev 30(6):355–381
336. van Geert P (1988) A graph theoretical approach to the structure of developmental models. Human Dev 31(2):107 135

337. van Geert P (1991) A dynamic systems model of cognitive and language growth. Psychol Rev 98:3–53
338. van Geert P (1993) A dynamic systems model of cognitive growth: competition and support under limited resource conditions. In: Smith LB, Thelen E (eds) A dynamic systems approach to development: applications. MIT Press, Cambridge, pp 265–331
339. van Geert P (1994) Dynamic systems of development. Change between complexity and chaos. Harvester, New York
340. van Geert P (1994) Vygotskian dynamics of development. Human Dev 37:346–365
341. van Geert P (1995) Dimensions of change: a semantic and mathematical analysis of learning and development. Human Dev 38:322–331
342. van Geert PLC (1996) The development of attachment and attachment-related competences. A dynamic model. In: Koops W, Hoelsma J, van den Boom J (eds) New developments in attachment research. North Holland, pp 181–199
343. van Geert P (1998) A dynamic systems model of basic developmental mechanisms: Piaget, Vygotsky and beyond. Psychol Rev 105(4):634–677
344. van Geert P (2000) The dynamics of general developmental mechanisms: from Piaget and Vygotsky to dynamic systems models. Current Direct Psychol Sci 9(2):64–68
345. van Geert P (2003) Dynamic systems approaches and modeling of developmental processes. In: Valsiner J, Conolly KJ (eds) Handbook of developmental Psychology. Sage, London, pp 640–672
346. van Geert P (2007) The Dynamic Systems approach in the study of L1 and L2 acquisition: an introduction. Mod Lang J 91:179–199
347. van Geert P (2007) Transmission, self-organization and the emergence of language: a dynamic systems point of view. In: Schönpflug U (ed) Cultural transmission: developmental, psychological, social, and methodological perspectives. Cambridge University Press, Cambridge UK (in press)
348. van Geert P (2008) Nonlinear-Complex-Dynamic-Systems in developmental psychology. In: Guastello S, Koopmans M, Pincus D (eds) Chaos and Complexity Now and in the Future: Recent Advances in the Theory of Nonlinear Dynamical Systems Psychology. Cambridge University Press, Cambridge
349. van Geert P (2008) The Dynamic Systems approach in the study of L1 and L2 acquisition: an introduction. Mod Lang J 92(2):179–199
350. van Geert P, Fischer KW (2008). Dynamic systems and the quest for individual-based models of change and development. In: Spencer JP, Thomas MSC, McClelland J (eds) Toward a new grand theory of development? Connectionism and dynamic systems theory re-considered. Oxford University Press, Oxford (in press)
351. van Geert PLC, Savelsbergh G, van der Maas H (1999) Transitions and non-linear dynamics in developmental psychology. In Savelsbergh G, Maas H, van Geert PLC (eds) Non-linear developmental processes. Elsevier Science Publishers, New York
352. van Geert P, Steenbeek H (2005) The dynamics of scaffolding. New Ideas Psychol 23(3):115–128
353. van Geert P, Steenbeek H (2008) Brains and the dynamics of 'wants' and 'cans': a commentary on Immordino-Yang's "A tale of two cases". Mind Brain Educ 2(2):62–66
354. van Orden GC, Holden JG, Turvey MT (2005) Human Cognition and 1/f Scaling. J Exp Psychol Gen 134:117–123
355. Vargha-Khadem F, Carr LJ, Isaacs E, Brett E (1997) Onset of speech after left hemispherectomy in a nine-year-old boy. Brain: J Neurol 120:159–182
356. Vasta R, Haith MM, Miller SA (1995) Child psychology: The modern science, 2nd edn. Wiley, Oxford
357. Viger C (2007) The acquired language of thought hypothesis: a theory of symbol grounding. Interact Stud 8(1):125–142
358. Vogt P, Divina F (2007) Social symbol grounding and language evolution. Interact Stud 8(1):31–52
359. Vygotsky LS (1978) Cole M, John V-Steiner, Scribner S, Souberman E (eds) Mind in society: the development of higher psychological processes. Harvard University Press, Cambridge
360. Vygotsky LS (1986) Kozulin A (ed) Thought and language. MIT Press, Cambridge
361. Vygotsky LS (1987) Thinking and speech. In: Rieber RW, Carton AS (eds) The collected works of LS Vygotsky: vol I Problems of general psychology. Plenum, New York, pp 37–285
362. Waldrop MM (1992) Complexity: The emerging science at the edge of order and chaos. Penguin, London
363. Weaver W (1948) Science and complexity, Am Sci 36:536–544
364. Weber-Fox C, Neville H (2001) Sensitive periods differentiate processing of open- and closed-class words: an ERP study of bilinguals. J Speech Lang Hear Res 44(6):1338–1353
365. Weisstein EW (1999) CRC concise encyclopedia of mathematics. Chapman Hall/CRC, Boca Raton
366. Wertsch JV (1985) Vygotsky and the social formation of mind. Harvard University Press, Cambridge
367. Wertsch JV (1991) Voices of the mind: a sociocultural approach to mediated action. Harvard University Press, Cambridge
368. Westermann G, Mareschal D, Johnson M, Sirois S, Spratling M, Thomas M (2007) Neuroconstructivism. Dev Sci 10(1):75–83
369. Wildgen W (1986) Processual semantics of the verb. J Semant 5(4):321–344
370. Wildgen W (1982) Catastrophe theoretic semantics-: an elaboration and application of René Thom's theory. Benjamins, Amsterdam
371. Williams JHG, Waiter GD, Gilchrist A, Perrett DI, Murray AD, Whiten A (2006) Neural mechanisms of imitation and 'mirror neuron' functioning in autistic spectrum disorder. Neuropsychologia 44:610–621
372. Wimmers RH, Beek PJ, Van Wieringen PC (1992) Phase transitions in rhythmic tracking movements: A case of unilateral coupling. Hum Mov Sci 11:217–226
373. Wimmers RH, Savelsbergh GJP, Beek PJ, Hopkins B (1998) Evidence for a phase transition in the early development of prehension. Dev Psychobiol 32:235–248
374. Wimmers RH, Savelsbergh GJP, van der Kamp J, Hartelman P (1998b) A developmental transition in prehension modeled as a cusp catastrophe. Dev Psychobiol 32:23–35
375. Worgan S, Damper R (2007) Grounding symbols in the physics of speech communication. Interact Stud 8(1):7–30
376. Yavuz H (2007) An integrated approach to the conceptual design and development of an intelligent autonomous mobile robot. Robot Autonom Syst 55(6):498–512
377. Yoshikawa Y, Asada M, Hosoda K, Koga J (2003) A constructivist approach to infants' vowel acquisition through mother-infant interaction. Connect Sci 15(4):245–258

Books and Reviews

Jansen BRJ, van der Maas HLJ (2001a) Evidence for the Phase transition from rule I to rule II on the balance scale task. Dev Rev 21:450–494

Jansen BRJ, van der Maas HLJ (2002a) The development of children's rule use on the balance scale task. J Exp Child Psychol 81:383–416

van der Maas HLJ, Jansen BRJ (2003) What response times tell of children's behavior on the balance scale task. J Exp Child Psychol 85:141–177

Development, Evolution, and the Emergence of Novel Behavior

Amy K. Gardiner, David F. Bjorklund
Florida Atlantic University, Boca Raton, USA

Article Outline

Glossary
Definition of the Subject
Introduction
The Influence of Developmental Timing on Evolution: Heterochrony
Epigenesis, Developmental Systems, and Plasticity
Gene × Environment × Development Interactions
Developmental Systems and Evolutionary Change
Epigenetic Theories of Evolution
Conclusion
Bibliography

Glossary

Heterochrony an evolutionary change in the timing of individual developmental events.

Plasticity the ability of individuals to respond flexibly in biologically or behaviorally adaptive ways to changes in the enviroment.

Developmental Systems Theory the idea that development unfolds via the bidirectional interaction of genes and enviroment at all levels of the developmental system, including genetic, cellular, structural, behavioral, and cultural.

Epigenesis an emergent process by which an organism's structure and function chage from relatively undifferentiated states to increasingly specialized, differentiated forms throughout ontogeny.

Epigenetic inheritance the non-genetic transfer of information from one generation to another.

Definition of the Subject

Ontogeny, or individual development, results from the bidirectional interactions of genes and environment. It is this interaction that allows inherited traits to become expressed in the phenotypes of adult organisms. While each individual will develop along its own unique trajectory, most members of a species are very much the same because they all inherit a species-typical genotype and a species-typical environment. When this enviroment changes, individuals must adopt or they will fail to survive. Individuals with enough plasticity to respond to new environments by developing novel phenotypes will be more likely to survive than those without such resilience. In this way, developmental change can have substantial impact on evolution by providing the grist upon which natural selection acts. Successful developmental systems will be selected and inherited, and evolution may thus be seen as a series of ontogenies.

Introduction

There has been a resurgence of interest in evolution among psychologists and other students of behavior. Darwin's great insight that not only morphology but also behavior and "mind" have evolutionary histories has captured the attention of a myriad of scholars in disciplines from anthropology through zoology. An important issue in evolutionary psychology concerns how evolved, inherited characteristics become expressed in the phenotypes of adults. It seems obvious to some that such characteristics do not appear fully formed in the adult but emerge during *ontogeny* (the development of the individual), requiring a developmental analysis (e. g., [11,40,99]). On the flip side, developmental analyzes can be used to provide insights into the processes of evolution. *Phylogeny* (the development of the species) can be viewed as a series of ontogenies: The many ancestors of extant creatures each themselves developed, and it was changes during the course of these ontogenies that produced evolutionary change. In the words of West-Eberhard (p. 89 in [124]), "The evolution of the phenotype is synonymous with the evolution of development". Natural selection has had as much or more of an impact on the early stages of development as it has had on the adult, and as a result modifications during the fetal, infant, or juvenile periods establish new contexts for further selection and thus the evolution of the species. From this perspective, the influences of both genetic and environmental mechanisms are expressed through the process of development, yielding phenotypic variation that selection might then act upon.

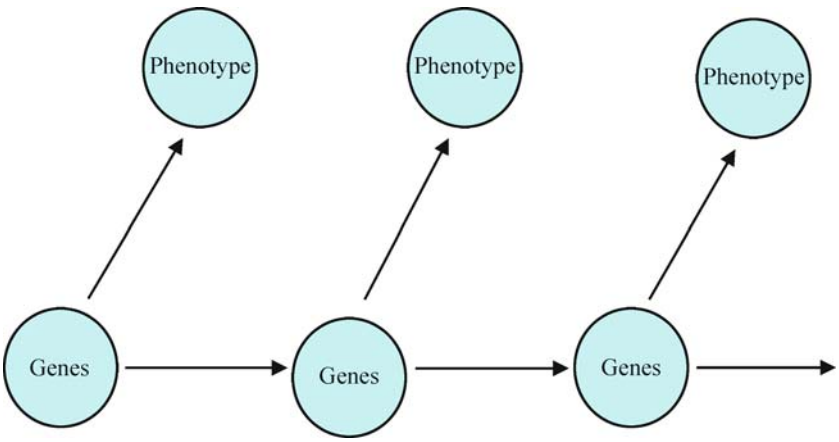

Development, Evolution, and the Emergence of Novel Behavior, Figure 1
Genocentric view of inheritance (see [123] and canonical view of Modern Synthesis)

Arguing that an animal's ancestors developed and that a species' ontogeny had an influence on its phylogeny may seem self-evident; however, a closer examination reveals that this perspective was contrary to conventional wisdom in mainstream biology throughout most of the 20th century, and continues to be controversial for some today. Even before the rediscovery of Mendal's work, biologists realized that organisms could only inherit parental characteristics through the germ line (see Fig. 1), making what happened during the course of development irrelevant for the evolution of the species [123]. This viewpoint became dogma with the establishment of the Modern Synthesis [32,76,105], with any claims about the potential role of ontogeny on phylogeny being described as "Lamarckian", a label anathema to any biologist. There were a few theorists over the century who attempted to integrate development with evolution (e. g., [32,39,45,51,88]), but for the most part, development was seen by evolutionary biologists as an epiphenomenon – of critical importance to the individual but irrelevant for evolution.

In the waning years of the 20th century, with the increasing awareness that genes are always expressed in a developmental context, some biologists began to propose a significant role of ontogeny on phylogeny, albeit one quite different from that proposed by Lamarck. The new field of evolutionary developmental biology, or *Evo-Devo*, examines the role of ontogeny (especially developmental genetics) on phylogeny across a wide range of species (e. g., [24,102,124]). Evolutionary-oriented psychologists also took note, proposing that changes in behavior over the course of development could contribute to evolutionary change (e. g., [12,44,46,48,55,72,95,97,99]). These theorists recognized that the connection between genes and behavior (or genes and morphology, for that matter) was not a direct one. Rather, new forms and function emerge via a dynamic interaction between genes and environments, with alterrations in environment, especially during early development, having potentially profound effects on the adult phenotype. With respect to behavior, this perspective implies that there is a great deal of *plasticity*, defined as the ability to modify behavior as a result of environmental input. This plasticity permits an organism to change its behavior to adapt to novel environments, or perhaps to enter new environments, where they will experience new selection pressures. Despite this plasticity, most members of a species end up resembling one another to a significant degree, an outcome that is due to the fact that individuals inherit not only a species-typical genome but also a species-typical environment.

In this chapter, we examine the influence of development on evolution. We first briefly examine the possible role of developmental timing (heterochrony) on evolution. The bulk of our chapter deals with how an epigenetic account of gene-environment interactions describes the course of development and evolution. We specifically explore the developmental systems perspective of ontogeny (e. g., [46,49,50,95] and theories of *epigenetic inheritance*, or *epigenetic theories of evolution* (e. g., [57,61,62,78]). We focus toward the end of our chapter on the possible role that developmental plasticity may have had on the evolution of human social cognition, particularly through maternal effects (e. g., [7,10]).

The Influence of Developmental Timing on Evolution: Heterochrony

Prior to the Modern Synthesis, development was seen as playing a critical role in evolution. This can be seen in

the popularity of Ernst Haeckel's *biogenetic law*, which is best captured by the phrase *ontogeny recapitulates phylogeny* (see [45,51,80,106] for historical reviews). Briefly, Haeckel proposed that the development of the individual goes through, or repeats, the same stages as the evolution of the species, which is most clearly seen during embryonic development. Haeckel proposed that new stages are added to the adult stage of ancestral organisms, making evolution a progressive phenomenon. Findings in the early 20th century seriously questioned Haeckel's assumptions, however. For example, although teeth are seen in evolution prior to tongues, they develop in the reverse order in modern-day mammals [32]. Biologists soon came to recognize that the timing of some aspects of individual development could vary in a number of ways from ancestral development, a phenomenon termed *heterochrony* [32,51,82,83,107,108]. In addition to accelerations of developmental timing relative to an ancestor (implicit in Haeckel's recapitulation theory), development could also be slowed, or retarded (termed *paedomorphosis*). Modifications of developmental timing result in different phenotypes, which, if they prove adaptive, eventually result in evolutionary change [32]. "Evolution by heterochrony" implies that factors that influence the onset or offset of a developmental process can have cascading effects on an organism, resulting in substantial changes in form or function without requiring substantial changes in the genome.

One particular type of developmental retardation that caught the attention of many evolutionary biologists is termed *neoteny* (literally, "holding youth") and refers to a reduced rate of development and the retention of juvenile characteristics into adulthood [51,56,89,106]. For instance, Gould ([51], p. 375) stated "the early stages of ontogeny are a storehouse of potential adaptations, for they contain countless shapes and structures that are lost through later allometries. When development is retarded, a mechanism is provided (via retention of fetal growth rates and proportions) for bringing these features forward to later ontogenetic stages."

One oft-cited example of neoteny is the salamander species axolotl (see [51]). Axolotls begin life in water as tadpoles, their larval state. They metamorphose into adults as land-dwelling, air-breathing newts. However, when conditions in the water are favorable, the tadpoles will become sexually mature. The development of their reproductive system follows the species-typical pattern, while the development of their "gills-to-lungs" transformation is delayed, resulting in a sexually mature larval organism. Some of the offspring of these larval parents may then progress through the typical developmental sequence, becoming air-breathing newts, while their parents remain tadpoles (albeit sexually mature ones).

Changing the timing of certain developmental events can result in modifications in the juvenile and adult organism, some of which may be adaptive and lead, eventually, to evolutionary changes. However, changes in developmental rates *permit* evolutionary innovations rather than cause them. The origins of those changes must be found in genetics (through mutations, for example) and/or pressures in the environment, with heterochrony being a response to these pressures. But development should not be viewed as a simple unfolding of maturationally paced events that once set in motion (perhaps earlier or later than in an ancestor) have an inevitable outcome. The key to understanding evolutionary change, we argue, is to understand developmental change, and it is to this topic we now turn.

Epigenesis, Developmental Systems, and Plasticity

The phenotypic expression of genetically inherited traits involves many components, including DNA, RNA, proteins, and multiple levels of environmental surroundings. The traditional biological view of development described the genotype as a blueprint for the phenotypes of organisms. This "central dogma" asserted that genetic information flowed in one direction only, from DNA to RNA to proteins. Subsequent evidence revealed that the process of development is extensively complex and that environmental influences on gene expression play a critical role in how inherited traits emerge within organisms (see [48]). This is expressed in *developmental systems theory*, which posits that development progresses via a system of interacting levels, from the genetic through the cultural.

Epigenesis and Development

Central to developmental systems theory is the concept of *epigenesis*, which can be defined as "an emergent process by which an organism's structure and function change from relatively undifferentiated states to increasingly specialized, differentiated forms throughout ontogeny" ([86], p. 105). An epigenetic view of development describes ontogeny as a process of continuous, bidirectional interaction between components at all levels of the developmental system, including the genetic, cellular, phenotypic, behavioral, and cultural. Components of the system include genetic activity, structural maturation, activity emanating from structure, known as function, and the larger environment (see e. g., [48,49,95,124]). Genes are not given a privileged role in development, but are considered as one integral part of the developmental system that require in-

put from and interaction with other components of the system to function appropriately. From this perspective, no dichotomy is drawn between genetic and environmental influences on particular traits. Rather than distinguishing specific, individual contributions of genes and environment, or nature and nurture, the epigenetic perspective emphasizes the interaction between genes and environment as they jointly contribute to the emergence of phenotypic characteristics. Development is not construed as simply a combination of nature and nurture, but as the emergent process of nature *via* nurture.

Gottlieb [46,47] differentiated two types of epigenesis: predetermined and probabilistic. *Predetermined epigenesis* is reminiscent of the central dogma of biology, describing a unidirectional relationship between structure and function: genes → structure → function. In contrast, *probabilistic epigenesis* describes a bidirectional relationship between structure and function, with interaction among adjacent levels of the system: genes ↔ structural maturation ↔ function. This produces increasingly complex organization over the course of development as new structures and functions emerge from interactions between genes and environment.

These interactions occur throughout development, beginning at conception, when the first somatic cells of an organism emerge. Each somatic cell contains the entire inherited genotype. However, different parts of the genotype are expressed in different parts of the organism's body, and only a handful of genes are expressed at any single point in development. Genetic activity requires environmental interaction and is therefore determined by the environment in which genes exist. Gene × Environment interactions begin as the genotype of a newly conceived organism interacts with the cytoplasmic environment of the mother's ovum. These interactions continue throughout development as genes find themselves in increasingly complex biological and social environments. This bidirectional interaction between different levels of organization that occurs continuously over time, hypothesized by developmental systems theory, is illustrated in Fig. 2.

Central to the concept of probabilistic epigenesis is *developmental plasticity*, the flexibility that organisms show in response to their environments during development. While a genotype may be inherited as a replication of parental genes, individuals are unlikely to experience the precise environment of previous generations or even of members within their own generation. Therefore, the developmental of each individual follows a slightly different course than that of its parents and other conspecifics. The environment may also change within a single generation, requiring organisms to adapt or fail to survive. Demands of a shifting environment can be met by a genotype that has the potential to respond to environmental novelties by interacting with the environment in ways that facilitate successful development. Plasticity thus ensures individual survival by allowing organisms to adapt to environmental changes encountered during ontogeny and ensures species survival by allowing individuals to develop successfully in

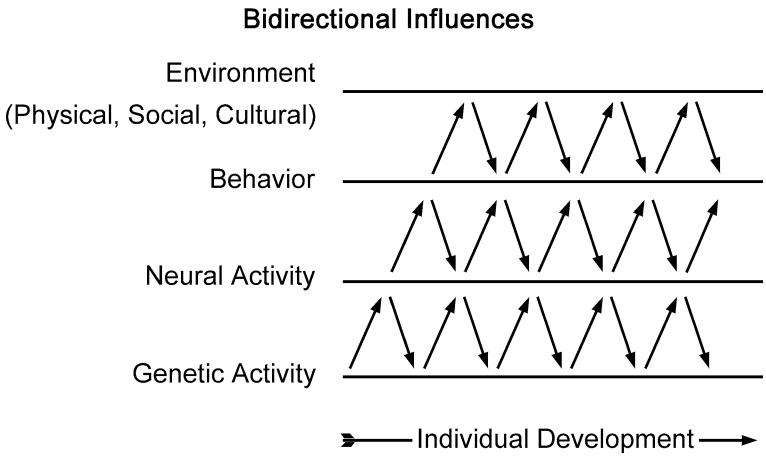

Development, Evolution, and the Emergence of Novel Behavior, Figure 2
A simplified schematic of the developmental systems approach, showing a hierarchy of four mutually interacting components (from [46]). Activity at one level influences activity at adjacent levels. For instance, genetic activity may affect the generation or activity of neurons, which in turns influences behavior. That behavior then has some outcome on the environment. The response of environment, in turn, influences behavior, which affects brain activity, which itself alters genetic activity (turns on or off a particular gene, for example)

environments that may be different from those of their ancestors.

Not all aspects of the environment are equally likely to change. Certain environmental features, such as gravity and patterned light, will remain reliably constant across generations, whereas other environmental characteristics, such as climate, are more apt to vary. Likewise, developmental plasticity does not exist to the same degree across all traits, as some traits show higher levels of plasticity than others. Traits that evolved within reliable environments show low levels of plasticity and vary little across generations. They are considered highly *canalized* and require extremely atypical environments to develop abnormally (e. g., for humans, language). In contrast, traits that evolved within variable environments show high levels of plasticity and may vary considerably across populations or generations (e. g., literacy).

Because of the dynamic nature of the developmental system, the specific ontogeny of any individual is impossible to predict with complete accuracy. Development can be predicted to a certain degree, however, as implied by the concept of probabilistic epigenesis. While organisms may encounter atypical environments and experience environmental change that lead to atypical development, most individuals inherit a species-typical environment in addition to a species-typical genome. Organisms can expect to develop in a species-typical way so long as environmental conditions do not depart radically from those experienced by their ancestors. When an organism is faced with drastic changes, however, atypical development is expected. The fact that many organisms exhibit developmental plasticity that allows them to survive when confronted with species-atypical environments provides evidence that environmental change was a recurrent problem faced by many species during their evolution, a problem that plasticity evolved to solve.

Developmental Systems

The entire system of interacting gene and environment components constitutes a developmental system. The ontogeny of intersensory integration in bobwhite quails provides an excellent illustration of a developmental system in action. The auditory system of bobwhite quail chicks that are hatched and reared in a species-typical environment begins functioning before hatching, while the visual system begins functioning after hatching. Normally hatched chicks show an auditory preference for the maternal call of their species at both 24 and 48 hours after birth, and at 72 hours require simultaneous auditory and visual information (seeing the mother while hearing her call) to show a preference. This typical pattern of development can be altered when the initial functioning of these sensory systems is manipulated.

Lickliter [69] removed a portion of the eggshell of bobwhite quail chicks and exposed the birds to patterned light 24–36 hours prior to hatching, thus shifting the initial stimulation and functioning of the visual system to an earlier point in development, corresponding to the time when the auditory system normally begins to function. These chicks showed no preference for the maternal call of their own species over a chicken call and no preference for a stuffed bobwhite hen over a stuffed quail hen at either 24 or 48 hours after hatching. This suggests interference of normal auditory perception because of the early exposure to visual information. However, when tested with both auditory and visual information in the form of a stuffed adult bobwhite hen with the bobwhite maternal call versus a stuffed quail hen with the chicken call, chicks showed a preference for the bobwhite hen at both 24 and 48 hours after hatching. As normally hatched chicks do not require both sources of information until 72 hours of age to show preference, this early activation of the visual system effectively accelerated the chicks' integration of the auditory and visual sensory systems. These findings indicate that the sequence of functional onset of these systems is an important factor in the early perceptual organization of this species. Genes were surely involved in this "imprinting" behavior, but the young bird's tendency to approach the maternal call was mediated by the timing of perceptual experience and not the simple outcome of a genetically programmed "instinct".

Further work with bobwhite quail hatchlings provided more details about their developmental system in regard to prenatal auditory learning. Lickliter and Hellewell [70] found that bobwhite quail chicks that were repeatedly exposed to a particular maternal call during the 24 hours before hatching showed a preference for this call over an unfamiliar variant until at least 24 hours after hatching, while unexposed chicks showed no preference. The addition of visual information appeared to interfere with this auditory learning when the maternal call was presented concurrently with patterned light. Chicks exposed to light at the same time that they heard the maternal call showed no preference for this call when tested after hatching. However, when the visual information was presented non-concurrently, chicks were able to learn the maternal call and showed a preference after hatching, suggesting that the timing of the visual information might be important.

Another factor to consider is the social environment in which the chicks were incubated. All chicks in the Lickliter and Hellewell [70] study were incubated communally, sur-

rounded by their broodmates during the pre-hatching period, and thus received sensory stimulation in the form of vocalizations uttered by their broodmates. To investigate if the social environment had an effect on bobwhite prenatal auditory learning, Lickliter and Lewkowicz [71] incubated some chicks in isolation and others communally. All chicks received auditory information of the particular maternal call. None of the chicks incubated communally received visual information, but some of the chicks incubated in isolation were exposed to light at the same time that they heard the call. Chicks incubated communally and chicks that were isolated but exposed to visual information preferred this call to an unfamiliar variant. Chicks incubated in isolation without visual information did not show a preference. Based on findings from these studies, Lickliter and Lewkowicz [71] concluded that bobwhite quail chicks need an optimal level of background stimulation to learn auditory information prenatally, but too much (broodmates, light, and sound) or too little (just sound) can prevent learning. The source of the sensory stimulation, broodmates or light, is less important than the stimulation itself. These studies provide a clear example of a developmental system and demonstrate how small changes to the system can produce noticeable perturbations in development.

Plasticity

Plasticity is essential for development. Especially for traits that develop in unpredictable environments, a degree of plasticity is necessary for the interactions between genes and environments to produce viable phenotypes.

Plasticity and Morphological Change Many animals display truly remarkable degrees of plasticity, at least in comparison to mammals. For instance, beginning in high-school biology class, we are told that sex is determined genetically, associated with genes on the X and Y chromosomes. This is true for all mammals and birds, and sex is similarly determined for most other animal species. However, whether one is male or female for some lizards, turtles, and crocodilians is not determined by specific genes but by the temperature at which eggs are incubated. In most turtles in which sex is determine by temperature, lower temperatures produce males and higher temperatures females, with this pattern being reversed for crocodilians [17,30]. Temperature also determines the camouflaging pattern on the wings of some butterflies. For instance, in the warm, wet season, the wings of the butterfly *Bicyclus anynana* develop many eyespots, which helps it blend into the folliage of the summer environment. In cooler, dry weather, *Bicyclus anynana* develop fewer or no spots, which helps them blend into the mostly brown leaf litter [16]. Temperatures seem to affect a gene (Distal-less) associated with the production proteins, which in turn affects the generation of wing spots (described in [23]).

Early diet can also play a pivotal role in the development of some organisms. The moth *Nemoria arizonaria* lays its eggs on oak trees during different times of the year, when floral environments and available diets are different. Based on the diet ingested during the first few days after hatching, the caterpillars that emerge will develop one of two morphologies. The spring brood feeds on oat catkins and comes to resemble the catkins, while the summer brood feeds on oak leaves and resembles oak twigs. These morphologies allow the caterpillars to blend into the surrounding environment, protecting them from predators [52].

In each of these examples, important aspects of an animal's morphology are determined by environmental conditions, not by commonly-shared genes. A male and female alligator may be genetic twins, something that is impossible in mammals. Genes evolved to be sensitive to differences in some environmental conditions during particular times in development, which resulted in adaptive outcomes during the species' natural selective history, what Boyce and Ellis [14] refer to as *conditional adaptations*. Although the action of genes still determines an animal's morphology, variation in the environment, not in the genome, is responsible for important aspects (e. g., sex) of an animal's physique.

Plasticity and Behavioral Development Behavioral plasticity early in development has important consequences for animals. As we saw in the examples of bobwhite quails (e. g., [69]), important aspects of an animal's behavior (and thus presumably their brains) can be altered as a result of species-atypical perceptual experiences. The plasticity of brain and behavior can also serve a protective function, in that the effects of deleterious early environments can be reversed with a radical change of circumstances. For example, six months of social isolation in rhesus monkeys, which typically results in permanent social and sexual dysfunction, can be reversed if the animals are provided with appropriate "therapy" (in this case, six months of daily interaction beginning at 6-months of age with a younger conspecific [115]). Similarly, the placement of children into foster or adoptive homes who had spent their early years in stultifying institutions, results in significant improvements of intellectual and social functioning (e. g., [63,91,92,110]). The relatively slow development of the human brain over the preschool years provides a de-

gree of plasticity that permits a reversal of the effects of an early harmful environment, with the degree to which such effects can be reversed declining with age (e. g., [91,92]).

Plasticity also appears to play an important role in the ontogeny of children's social competencies. Huether [60] proposed a framework for understanding the interaction between early environment and neural plasticity and how it affects social development. According to this framework, individuals can experience at least two types of responses to environmental stress: controllable and uncontrollable. The type of response experienced depends on the degree of control an individual subjectively perceives he or she can assert on a situation through his or her own actions. A controllable stress response is in reaction to a situation one feels he or she can control, whereas an uncontrollable stress response is in reaction to a situation one feels he or she cannot control. Both responses begin as nonspecific patterns of arousal of cortical and limbic structures in the brain. Controllable responses direct arousal to specific neuronal pathways involved in behavioral response to the given stressor and increase noradrenergic output to stabilize these pathways. In contrast, uncontrollable responses escalate arousal and activate the hypothalamic-pituitary-adrenal (HPA) system and adrenal glucocorticoid secretion, which can destabilize previously established pathways and provide opportunity for neural reorganization through the acquisition of new coping strategies. Behavioral strategies that succeed in making a stressor subjectively controllable are reinforced and facilitated, whereas strategies that repeatedly fail and make a stressor subjectively uncontrollable will be eliminated or become a source of constant dysregulation. Both controllable and uncontrollable stress experiences are needed to establish individual patterns of coping that represent the sum of previous experience and to contribute to normal brain development, but it is repeated experience of controllable stress that allows children to develop appropriate, successful social strategies.

Flinn [37] notes that stress hormones such as glucocorticoids divert bodily resources from important health functions such as immune response and that chronic stress, especially during childhood, can have negative long-term effects on health. Therefore, uncontrollable stress responses, which activate glucocorticoid secretion, appear to come at a cost. Evolutionarily, if these responses were adaptive, it is expected that they would provide benefits that outweigh such costs. Flinn argues that the benefits of psychosocial stress during childhood exceed the costs because navigating relationships is of paramount importance in the socially dynamic human environment. From this perspective, elevated levels of stress hormones in response to social situations can be advantageous if they facilitate neural reorganization that allows individuals to cope with a variable, unpredictable environment.

The relationships between social experiences, stress responses, and neural organization are extensively complex. In humans especially, the effects of early stress experiences on neural organization and social ontogeny are difficult to study because of methodological considerations. To understand the relationship between social experiences and stress response, Flinn measures the stress hormone cortisol, a glucocorticoid, through collection of children's saliva samples. In a long-term study, Flinn [37] collected saliva from the children of a small rural community on the Caribbean island of Dominica at least twice a day over a period of more than 10 years. He found that cortisol levels were generally highest in response to social challenges, particularly those involving conflict or change within children's families. Children living in families with high levels of marital conflict were more likely to show abnormal, elevated levels of cortisol than children living in more harmonious family environments. Longitudinal results support the conclusion that interactions within the family are a critical source of psychosocial stress in childhood. Children who experienced early family trauma in the form of parental divorce, death, or abuse have higher cortisol levels at age 10 and higher morbidity than children who do not experience early family trauma.

Flinn suggests that early experiences of family trauma result in permanent dysregulation of the HPA system, which leads to impaired regulation of cortisol levels in response to social challenges, as well as deleterious effects on brain regions involved in important neural, metabolic, and immune system health. However, temporary cortisol response to psychosocial challenges is important because it facilitates neural reorganization in response to the constantly shifting social landscapes of human communities. Because human relationships are of such great importance, the benefits of the stress response system outweigh the costs.

Both Huether [60] and Flinn [37] note that not all children will develop identically given the same social circumstances. Genetically-based individual differences between children in how they respond to social challenges, such as temperament, will have a significant effect on how they interact with their environments, as well as the development of their social competencies. These interactions between genes and environment are seen in studies of rhesus monkeys. For instance, in one study rhesus monkeys were classified as being either highly reactive to novel social experiences and displaying elevated cortisol levels or classified as less reactive [15]. When living in a spacious,

low-stress environment, low-reactive animals had slightly fewer injuries over the course of the year than high-reactive individuals. When the monkeys were moved to a confined, high-stress environment, the relation between stress reactivity and injuries reversed. Now, the high-reactive monkeys showed substantial increases in injuries, whereas there was no change in the rate of injuries for the low-reactive monkeys.

Other research with monkeys (e. g., see [112,114]) and children (e. g., see [5,42,111]) has shown that individual differences in temperament or reaction to stress interacts with the quality of the rearing environment to produce differential reaction to parental influence or to specific psychosocial outcomes (e. g., depression, antisocial behavior). In the past decade, researchers have identified specific genes associated with certain behavioral outcomes under some environmental contexts, providing greater insight into the nature of gene × environment × development interactions, and we examine briefly several examples of this research in the next section.

Gene × Environment × Development Interactions

Gene × environment × development interactions are a prominent feature of the developmental systems approach. Throughout development, an individual's inherited genotype interacts with the environment in the epigenetic process that leads to the emergence of phenotypic characteristics (see [9]). Several studies of the interaction between genes and environment in relation to human and non-human primate parenting behavior illustrate this effect.

Caspi et al. [25] examined the relationship between maltreatment during childhood and possession of a particular allele related to the production of Monoamine oxidase A (MAOA), a compound that metabolizes several neurotransmitters. Levels of MAOA are governed by different alleles of a gene located on the X chromosome and have been shown to be associated with aggression. Specifically, low levels of MAOA are associated with high levels of aggression. Caspi et al. [25] found that boys with low MAOA levels showed elevated amounts of antisocial behavior, but only if they had experienced maltreatment during childhood. In the group that had not experienced maltreatment, boys with low MAOA levels actually showed lower amounts of antisocial behavior than boys with high MAOA levels. Another study by Caspi and colleagues [26] found that different alleles of the 5-HTT serotonin transporter gene and childhood maltreatment were related to depression in young adulthood. Individuals with the LL (long-long) variant of the gene showed no relationship between maltreatment and depression, but individuals with the LS (long-short) or SS (short-short) variants who had experienced maltreatment were more likely to report depressive episodes during young adulthood than those who had not experienced maltreatment. In other research, adolescents who experienced high degrees of maternal rejection were more apt to be classified as clinically depressed than adolescents whose mothers did not reject them, but only if they had one combination of a particular gene that influences dopamine transfer [53].

These and other studies (e. g., [27,87]) suggest that early rearing environments may moderate genetic effects on behavior through interaction with particular gene variants. This conclusion is tentatively drawn, however, because these findings are correlational and because parenting behavior (and consequently early rearing environment) may naturally vary with genotype, thereby making true gene-environment interactions difficult to discern.

These difficulties can be overcome in animal studies. Suomi [114] investigated the relationship between the LL and LS alleles of the 5-HTT gene and parenting behavior in rhesus monkeys by examining the differences in aggression between peer-raised and mother-raised individuals. Peer-raised monkeys with the LS variant showed high levels of aggression, while those with the LL variant did not. There was no relation between allelic variation and aggression in mother-raised monkeys, suggesting a maternal buffering effect. Similar maternal-buffering effects for individuals with the LS allele were found for hypothalamic-pituitary-adrenal responses to social separation [3] and for neonatal neurobehavioral development [29].

Suomi [113] discussed additional findings from cross-fostering studies that investigated gene × environment × development interactions in regard to the temperament of rhesus monkey infants and the type of maternal parenting style received during the first months of life. Monkeys were selectively bred to fall within the normal range of temperamental reactivity or within the high reactive range. High-reactive infants tend to have fearful and anxious reactions to stressful situations, begin to use their mothers as a secure base from which to explore relatively late in development, and are initially shy in peer relationships. The monkeys were cross-fostered with mothers who were either in the normal range of maternal style or showed an unusually nurturant style. After six months with their mothers, the infants were placed in peer groups. High-reactive infants reared by normal nurturant mothers displayed deficits in exploration and fell to the bottom of peer hierarchies. In contrast, high-reactive infants reared by high-nurturant mothers appeared behaviorally precocious, leaving their mothers to explore earlier and explor-

ing more than all other groups, and rising to high-ranking positions within peer hierarchies. When females in this study became mothers themselves, they displayed a maternal style consistent with that of their foster mothers, regardless of the maternal style of their biological mothers. These findings suggest that the interactions between the infants' genetically inherited temperament and their early experiences of maternal care had long-term effects on the infants' developmental trajectories, even persisting into the subsequent generation.

Developmental Systems and Evolutionary Change

Evolution is the change in organisms (often defined in terms of changes in gene frequencies) and the differentiation of species that occur over geologic time. This process is borne out through the differential survival of individual organisms, known as natural selection. Evolution by natural selection is a relatively simple process that depends on the following four factors: 1) *superfecundity* – within a species, there are more members in each generation than can actually survive; 2) *variation* – there are significant differences in physical and behavioral traits between individuals within a species; 3) *inheritability* – traits are passed on from parents to offspring, making variation heritable; and 4) *selection* – the most adaptive traits, those that best promote survival and reproduction in the local ecology, increase in frequency within the population of a species, while maladaptive traits, which fail to support survival and reproduction, decrease. A good measure of the adaptability of a trait is the degree to which it promotes an organism's *inclusive fitness*, which is the reproductive success of an individual including its own offspring and the offspring of others with whom it shares genes [54]. Among all of the variation within a generation, those individuals that have traits that maximize their inclusive fitness will spread the most copies of their genes to future generations, thus increasing the frequencies of these traits within the population of a species.

Natural selection does not generate variation; it merely acts on the variation that exists between the phenotypes of each generation. There are several mechanisms through which variation may be created. Traditionally, the source of phenotypic variety has been identified as random mutation at the genotypic level. In this description of evolutionary change, a mutation alters the genetic makeup of an individual and this creates a novel phenotype within the population. In contrast, epigenetic theories of evolution recognize the prominent role of gene × environment × development interactions in phenotypic creation, and invoke phenotypic plasticity as the primary generator of novel variation. That is, the creation of novel phenotypes, whether produced by genetic variation or induced via environmental events, provides the material upon which natural selection acts. Along these lines, West-Eberhard [124] proposed that "adaptive evolution – phenotypic improvement due to selection – . . . is a two step process: first the generation of variation by development, then the screening of that variation by selection" (p. 139). West-Eberhard proposed that "new phenotypic subunits begin and evolve as products of developmental plasticity. They originate when an environmental or genetic perturbation causes a shift in genes expression, and they are consolidated under selection for improved regulation and form" (p. 129).

From this perspective, evolution should not be viewed simply as changes in gene frequencies over time, as is the canonical way, but rather as changes in developmental systems. According to Oyama ([96], p. 29), "What is transmitted between generations is not traits, or blueprints, or symbolic representations of traits, but developmental *means* (or *resources*, or *interactants*). These means include genes, the cellular machinery necessary for their functioning, and the larger developmental context, which may include a maternal reproductive system, parental care, or other interaction with conspecifics, as well as relations with other aspects of the animate and inanimate worlds. This context, which is actually a system of partially nested contexts, changes with time, partly as a result of the developmental processes themselves."

Individuals respond to environmental change by altering their gene expression in ways that diverge from those of the parent generation. Those changes in the developmental system that facilitated by plasticity. If an organism encounters a species-atypical environment and possesses appropriate levels of plasticity to adapt to this environment, its development will diverge in a species-atypical manner, creating a novel phenotype. If this novel phenotype provides the organism with enhanced survival and reproductive fitness relative to a typical phenotype or the phenotype of a less-plastic individual, the species-atypical developmental system will spread within the population. With each environmental change, this process repeats. Thus, developmental plasticity in response to environmental variation facilitates evolutionary change, often preceding genetic mutation. Such plastic responses lead to phenotypic variation within a population and provide the grist upon which natural selection acts [44,45]. Although the underlying genetic contribution of species-atypical behaviors and traits is still unknown, novel experiences may activate previously dormant genes, deactivate genes, alter the onset, offset, or amount of activity of genes, or result in adaptive outcomes for some but not other alleles

of a gene, all actions consistent with the Modern Synthesis.

Traits evolve in response to selection pressures within the environment. Each adaptive trait evolved within a specific *environment of evolutionary adaptedness*, which is the collection of environmental factors that provided pressure for the selection of a particular variation of a trait. Those individuals that have or are able to develop the trait under pressure are more likely to survive than those that do not or cannot.

Selection pressures may exist as passive components of the environment, such as light or gravity, which cannot be changed due to the actions of organisms. Or, selection pressures may arise from the modification of the environment by organisms through a process known as *niche construction* [68,94], defined as "the activities, choices, and metabolic processes of organisms, through which they define, choose, modify, and partly create their own niches" ([65], pp. 132–133). Environmental modification can result from basic life processes of organisms such as consumption and excretion, or through the direct construction of habitats. These habitats may represent an "extended phenotype" of the creatures that build them [31], such as the web of the spider or the beaver's dam. Through niche construction, organisms are able to create or modify selection pressures within their environments. The niche construction efforts of organisms may constitute the *ecological inheritance* [93] of subsequent generations that live in environments at least partially created by the actions of previous generations. If environmental modification occurs in consistent patterns across generations, niche construction can have significant effects on evolutionary outcomes by creating pressure for genetic change [64,65].

Epigenetic Theories of Evolution

Epigenetic inheritance is the non-genetic transfer of information across generations. Epigenetic theories of evolution have been discussed since the late 1800s, the oldest account named for James Mark Baldwin but proposed contemporaneously by Conway Lloyd Morgan and H. F. Osborn. The *Baldwin effect* occurs when acquired behaviors are transmitted non-genetically via social learning. Baldwin [2] suggested that organisms that survive environmental stress are able to cope because they have sufficient developmental plasticity to allow behavioral adjustment in response to stressors. These individuals reproduce and their offspring learn from them how to survive in the given environment. Eventually the once-acquired behavior comes to be expressed in all members of the population, a process Baldwin termed *organic selection*.

The selective advantage of a plastic developmental system was demonstrated by Waddington in several classic experiments (described in his 1975 collection of essays). For example, in one set of experiments Waddington exposed pupal fruit flies (*Drosophila melanogaster*) to heat shock. Although many individuals did not survive this exposure, some that did survive responded by developing wings with few or no cross veins. Waddington selectively bred those flies without cross veins and exposed their offspring to heat shock. This second generation also responded by developing wings with few or no cross veins. After 14 generations of selecting for this phenotype, some of the flies developed the selected phenotype without being exposed to heat shock. Waddington termed this phenomenon *genetic assimilation*. He was also able to demonstrate genetic assimilation for an adaptive trait by feeding fruit flies a diet high in salt. The surviving flies developed larger anal papillae, which facilitate the excretion of salt, than flies fed a normal diet. Waddington bred the survivors, and after 21 generations produced flies that developed large anal papillae even when grown on a low-salt medium. In both of these cases, a response to an environmental change was facilitated by the organisms' developmental plasticity and allowed the organisms to adapt and survive in their new environments.

Since Waddington's experiments, epigenetic inheritance has been demonstrated in a variety of species (see [62,124] for reviews). For example, the asexually reproducing water flea *Daphnia cucullata* will grow a protective helmet if exposed to the larva of predators, and offspring and grandoffspring of helmeted members of this species will also develop the protective trait, even without exposure to a predatory environment [1]. Ho, Tucker, Keeley, and Saunder [58] exposed fruit flies to ether, causing the flies to produce two sets of wings. This novel phenotype appeared in offspring of affected females, but not of affected males, suggesting transgenerational epigenetic inheritance of changes in the chemical machinery of the female gametes. This does not imply a change to the nuclear DNA of any female, and it should be stressed that none of the epigenetic theories presented relies on Lamarckian ideas to explain evolutionary change.

There is also solid evidence of epigenetic inheritance in mammals, dating back to the 1960s [33,103]. For example, in one early study mice from either the C57BL or BALB strains that had been raised by foster parents from the BALB strain performed better on four operant conditioning performance measures than C57BL foster-raised mice, and this effect was maintained into the subsequent generation [103]. These results suggested to Ressler [103] that "a nongenetic system of inheritance based upon trans-

mission of parental influences is potentially available to all mammals" (p. 267).

More recently, a research program by Michael Meany and his colleagues has provided impressive evidence of epigenetic inheritance in rats, with these effects being mediated by maternal behavior (see [28,84]). In rats, licking/grooming (LG) is highly associated with arched-back nursing (ABN), in which a mother arches her back and splays her legs outward during nursing. Offspring of mothers that engage in high levels of LG-ABN experience less stress than offspring of mothers that engage in low levels of LG-ABN. This effect is mediated by gene expression (see [84]), but is probably not due to genetic inheritance, as demonstrated in a cross-fostering study by Francis and colleagues [38], in which offspring of low LG-ABN mothers were raised by high LG-ABN mothers and vice versa. The stress reactions of the offspring were similar to the behavior of their foster mothers, not their biological mothers. The females of this second generation then served as foster mothers and displayed LG-ABN behavior characteristic of their foster mothers. The same effect was seen in this third generation. The experiment demonstrates that changes in gene expression can be moderated by maternal behavior and that these changes can persist across generations.

Meany (Fig. 1, p. 1179 in [84]) presented a schema illustrating how stress in an environment could affect maternal anxiety and subsequent maternal care, which in turn influences neural development of the offspring. Neural development affects expression of genes related to chemicals involved in stress responses (e. g., corticotropin-releasing factor) and production of neurotransmitters and hormones such as dopamine and oxytocin, which are involved in emotional expression and regulation. This pattern would affect the behavior of the female animal when she becomes a mother, influencing the neural development and behavior of her offspring. This led Meaney [84] to suggest that "Individual differences in behavioral and neuroendocrine responses to stress in rats are, in part, derived from naturally occurring variations in maternal care. Such effects might serve as a possible mechanism by which selected traits are transmitted from one generation to another" (pp. 1170–1171).

Other research with rhesus macaques has investigated the possibility that abuse of offspring is transmitted from mothers to daughters. In a cross-fostering study, Maestripieri [75] placed daughters of abusive mothers with non-abusive foster mothers and daughters of non-abusive mothers with abusive foster mothers during the first two days after birth. These groups were compared to macaques raised by abusive and non-abusive biological mothers. Both biological and foster daughters raised by abusive mothers were abused, while daughters of non-abusive mothers were not abused. Subsequent parenting styles of these macaques showed that 9 of the 16 females raised by abusive mothers became abusive toward their own offspring, whereas none of the females raised by non-abusive mothers became abusive. This suggests that transfer of abusive parenting from mother to daughter is mediated by early experience rather than genetic inheritance, but that individual differences in risk or protective factors can also impact whether a female will eventually display an abusive parenting style. Other studies have shown that abusive mothers invest more time in their infants than non-abusive mothers and are consistent in the level of abuse shown toward individual offspring, suggesting that this behavior represents inappropriate means of controlling infants and is pathological [74,76,77]. Early experience of abuse may lead to such pathological behavior because of long-term effects of early abusive experiences on stress responses to provoking situations (see discussion in [76]).

These studies by Maestripieri and colleagues, as well as other studies discussed above, illustrate the significant impact that mothers have on the development of their offspring. In mammals, maternal effects on the phenotypes of offspring begin prenatally and continue throughout the juvenile period, portions of an individual's life in which the developmental system is most flexible. A mother's contribution to the development of her offspring can significantly affect her ontogeny and thereby affect the phenotypes that serve as the raw material to be filtered by natural selection. Thus, maternal behavior shapes not only ontogeny, but phylogeny as well [7].

Epigenetic Inheritance, the Enculturation Hypothesis, and the Evolution of Human Social Cognition

Given the emphasis placed on the role of sociality as the impetus for the evolution of human intelligence, some theorists have suggested that one avenue for such changes may have been via epigenetic inheritance, specially via maternal effects [7,10]. As work with rats (e. g., Francis et al., [38]) and monkeys (e. g., [114]) has shown that socioemotional responsivity can be transmitted across generations via maternal effects, might social cognitive abilities similarly be transmitted? Although experimental research of the nature done by Meany, Suomi, and their colleagues with rats and monkeys cannot be performed with humans to address this question, some experimentation can be done with humans' closest genetic relatives, chimpanzees (*Pan troglodytes*). According to the *enculturation hypothesis* [22], when chimpanzees are reared much as hu-

man children, specially spoken to and treated as intentional agents – as individuals who are motivated by their own wishes and desires and who understand that the behavior of others is similarly motivated – they develop some social-cognitive abilities in a direction characteristic of human cognitive abilities, particularly the ability to understand others as intentional agents and to imitate others.

Imitation is an important human ability that facilitates cultural transmission by allowing individuals to understand and assume the goals of others and thereby faithfully replicate the behavior of previous generations [118]. Through imitation, children are able to inherit accumulated cultural knowledge, and a ratchet effect ensues as each generation contributes its own innovations to the culture and passes them on [117]. Much inherited cultural knowledge has significant survival value, such as how to use objects as tools, and shared cultural norms and values create important group bonds. Therefore, intentional understanding and imitation may be considered cognitive adaptations that facilitate the larger human adaptation of culture.

Enculturation effects are best seen in social learning, particularly imitation. Although chimpanzees and the other great apes display impressive social-learning abilities both in the wild (e. g., [125,126]) and in captivity (e. g., [125]), most do not seem to involve *true imitation*, which requires observers to understand the intentions of others, to adopt the another's goals as their own, and to reproduce important aspects of the observed behavior when attempting to achieve the goals [118,120]. Rather, most sophisticated chimpanzee social learning seems to be achieved via *emulation*, where observers attend to the effects on the environment produced by another's actions and reproduce those effects through different means, without understanding the effects as a goal of the observed individual [116,119]. For instance, Horner and Whiten ([59], see also [20,90]) had a human adult demonstrate how to retrieve a treat from within a puzzle box, first performing an irrelevant action that did not contribute to retrieval and then performing a relevant action that facilitated retrieval. Chimpanzees performed the most efficient solution by ignoring the irrelevant action and only replicating the relevant action, while 3- and 4-year-old children performed both the irrelevant and relevant actions. It is unclear why children chose a non-optimal solution, but it could be that a tendency to copy actions overrides emulative capabilities [73,127], perhaps because of the paramount importance of cultural learning.

In some ways, enculturated chimpanzees are more similar to human preschool children than to their nonenculturated conspecifics. For example, in investigating the imitation of novel behaviors by infants, Gergely, Bekkering, and Király [41] questioned whether one could differentiate true imitation from emulation. In their terms, were infants imitating rationally? To assess this, a model demonstrated a behavior to 14-month-old infants – using her head to press a button to turn on a light. This is a novel action (one would usually use one's hand) and, if babies copied it, would indicate that they were reproducing the exact behaviors the model had witnessed to achieve a goal. Gergely and his colleagues hypothesized that infants may have believed that the model did not use her hands when she could have because using her head may have provided some advantage in turning on the light. In their study, infants watched as a model turned on the light with her head either (a) with her hands free, or (b) with her hands wrapped in a blanket, that is, her hands were occupied and thus were not available to turn on the light. Infants in the "hands free" condition used their head to turn on the light. But the pattern was reversed in the "hands occupied" condition. Now most of the infants (79%) used their hands to turn on the light. That is, when there was a reason why the model did not use her hands (they were wrapped in a blanket), the babies focused on the goal (turn on the light), not the means (use your head), and used their hands as the most efficient way to turn on the light. When no such reason was available, they copied the model's behavior exactly, reflective of imitative learning. Nearly identical results were found in a similarly designed study with enculturated chimpanzees, suggesting that the imitation these animals display is indeed "rational", reflecting an understanding of the intentions and goals of the model [18].

Other studies have demonstrated that enculturated apes are more apt to display *deferred imitation* (imitating a model after a significant delay) than nonenculturated apes, often comparable to that of human preschool children. Tomasello, Savage-Rumbaugh, and Kruger [121] demonstrated a variety of actions on objects, such as prying open a paint can, and tested the immediate and deferred imitation abilities of both enculturated and mother-reared chimpanzees. The enculturated group replicated actions at significantly higher rates than the mother-reared group and even outperformed 1.5- and 2.5-year-old children on the deferred trials. Subsequent work by Bjorklund, Bering, and their colleagues also found that enculturated chimpanzees displayed deferred imitation ([6,8,13], see [12] for review). In an initial study, apes were allowed to interact with the target objects for a period of several minutes and then observed a human perform actions on the objects, such as placing a plastic nail in a form board and striking it with a plastic hammer. After a delay of 10 minutes, the chimpanzees and orangutans were given the

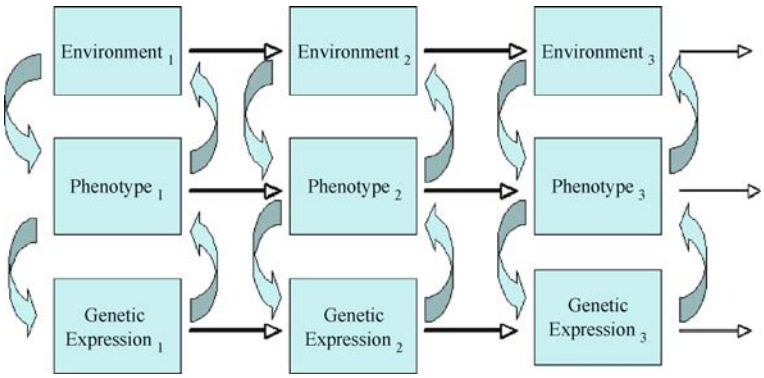

Development, Evolution, and the Emergence of Novel Behavior, Figure 3
An epigenetic view of inheritance

objects again [6]. These enculturated apes showed greater-than-chance levels of deferred imitation.

The emergence of such human-like behavior in enculturated non-human primates in the domains of social learning indicates that these apes have enough developmental plasticity to respond to human "mothering" in a manner approximating some aspects of human social cognition. This also suggests that our common ancestor had a similar level of plasticity that allowed for phenotypes to develop in response to environments increasing in social complexity and to lay the foundation for the modern human mind. These abilities would have emerged through of an epigenetic process in which the inherited cognitive and behavioral systems of the young animal are sensitive to the ecology in which it is raised [7,12,67]. One important component of that ecology is the infant's mother, who herself would need the social-cognitive plasticity to modify her behavior in response to environmental pressures. In fact, should a group of mothers experience the same novel environment that induced greater attention to their infants' intentions, their offspring should be similarly affected, resulting in a cohort of individuals with superior social-learning abilities, that in turn would treat their own offspring as intentional beings, promoting these skills in the population and, as a result, providing new selection pressures.

Conclusion

Although evolution and development both examine change over time, throughout most of the 20th century what happened during the lifetime on an individual was believed to have no consequence on evolution (other than reproducing the next generation). The new field of Evo-Devo and the recognition that genes are always expressed in a developmental context has changed this perspective. Development matters to evolution, just as an animal's phylogenetic history matters to its ontogeny. Epigenetic theories of development and evolution realize that activities of the organism and events in the wider environment feedback upon the genome, so that both development and evolution are best conceived as dynamical systems with bidirectional relations between different levels of organization within and outside of the organism. This suggests that the genocentric view of evolution presented in Fig. 1 requires modification to something more resembling the situation depicted in Fig. 3: Activity of the genome, the form and function of the individual (the phenotype in Fig. 3), and the environment all interact, with the nature of the interaction changing over time, resulting in gene × environment × development effects that determine both the course on an individual life and of a species.

Bibliography

1. Agrawal AA, Laforsch C, Tollrian R (1999) Transgenerational induction of defences in animals and plants. Nature 401:60–63
2. Baldwin JM (1902) Development and evolution. McMillan, New York
3. Barr CS, Newman TK, Shannon C, Parker C, Dvoskin RL, Becker ML et al (2004) Rearing condition and rh5-HTTLPR interact to influence limbic-hypothalmic-pituitary-adrenal axis response to stress in infant macaques. Biol Psychiatry 55:733–738
4. Belsky J, Steinberg L, Draper P (1991) Childhood experience, interpersonal development, and reproductive strategy: An evolutionary theory of socialization. Child Dev 62:647–670
5. Belsky J, Bakermans-Kranenburg MJ, van IJzendoorn MH (2007) For better and worse: Differential susceptibility to environmental influences. Curr Dir Psychol Sci 16:300–304
6. Bering JM, Bjorklund DF, Ragan P (2000) Deferred imitation of object-related actions in human-reared juvenile chimpanzees and orangutans. Dev Psychobiol 36:218–232
7. Bjorklund DF (2006) Mother knows best: Epigenetic inheritance, maternal effects, and the evolution of human intelligence. Dev Rev 26:213–242

8. Bjorklund DF, Bering JM, Ragan P (2000) A two-year longitudinal study of deferred imitation of object manipulation in an enculturated juvenile chimpanzee (Pan troglodytes) and orangutan (Pongo pygmaeus). Dev Psychobiol 37:229–237
9. Bjorklund DF, Ellis BJ, Rosenberg JS (2007) Evolved probabilistic cognitive mechanisms: An evolutionary approach to gene X environment X development interachtions. Adv Child Dev Beh 35:1–36
10. Bjorklund DF, Grotuss J, Csinady A (2008) Maternal effects, social cognitive development, and the evolution of human intelligence. In: Maestripieri D, Mateo J (eds) Maternal effects in mammals. Chicago University Press, Chicago
11. Bjorklund DF, Pellegrini AD (2002) The origins of human nature: Evolutionary developmental psychology. American Psychological Association, Washington
12. Bjorklund DF, Rosenberg JS (2005) The role of developmental plasticity in the evolution of human cognition. In: Ellis BJ, Bjorklund DF (eds) Origins of the social mind: Evolutionary psychology and child development (pp 45–75). Guilford, New York
13. Bjorklund DF, Yunger JL, Bering JM, Ragan P (2002) The generalization of deferred imitation in enculturated chimpanzees (Pan troglodytes). Anim Cogn 5:49–58
14. Boyce WT, Ellis BJ (2005) Biological sensitivity to context: I An evolutionary-developmental theory of the origins and functions of stress reactivity. Dev Psychopathol 17:271–301
15. Boyce WT, O'Neill-Wagner P, Price CS, Haines M, Suomi SJ (1998) Crowding stress and violent injuries among behaviorally inhibited rhesus macaques. Health Psychol 17:285–289
16. Brakefield PM, Gates J, Keys D, Kesbeke F, Wijngaarden PJ, Montelro A, French V, Carroll SB (1996) Development, plasticity and evolution of butterfly eyespot patterns. Nature 384:236–242
17. Bull JJ (1980) Sex determination in reptiles. Q Rev Biol 55:3–21
18. Buttelman D, Carpenter M, Call J, Tomasello M (2007) Enculturated chimpanzees imitate rationally. Dev Sci 10:F31-F38
19. Call J, Carpenter M (2002) Three sources of information in social learning. In: Dautenham K, Nehaniv C (eds) Imitation in Animals and Artifacts. MIT Press, Cambridge, pp 211–228
20. Call J, Carpenter M, Tomasello M (2005) Copying results and copying actions in the process of social learning: chimpanzees (Pan troglodytes) and human children (Homo sapiens). Anim Cogn 8:151–163
21. Call J, Tomasello M (1994) The production and comprehension of referential pointing by orangutans. J Comp Psychol 108:301–317
22. Call J, Tomasello M (1996) The effects of humans on the cognitive development of apes. In: Russon AE, Bard KA, Parker ST (eds) Reaching into thought: The minds of the great apes. Cambridge University Press, New York, pp 371–403
23. Carroll SB (2005) Endless forms most beautiful: The new science of Evo Devo. Norton, New York
24. Carroll SB, Grenier J, Weatherbee S (2005) From DNA to diversity: Molecular genetics and the evolution of animal design, 2nd edn. Blackwell Science, Medford
25. Caspi A, McClay J, Moffitt TW, Mill J, Martin J, Craig IW et al (2002) Role of genotype in the cycle of violence in maltreated children. Science 297:851–854
26. Caspi A, Sudgen K, Moffitt TE, Taylor A, Craig IW, Harrington H et al (2003) Influence of life stress on depression: moderation of a polymorphism in the 5-HTT gene. Science 301:386–389
27. Caspi A, Williams B, Kim-Cohen J, Craig IW, Milne BJ, Poulton R, Schalkwyk LC, Taylor A, Helen Werts H, Terrie E Moffitt TE (2007) Moderation of breastfeeding effects on the IQ by genetic variation in fatty acid metabolism. Proc Natl Acad Sci 47:18860–18865
28. Champagne FA, Curley JP (2008) The trans-generational influence of maternal care on offspring gene expression and behavior in rodents. In: Maestripieri D, Mateo J (eds) Maternal effects in mammals. Chicago University Press, Chicago
29. Champoux M, Bennett A, Shannon C, Higley JD, Lesch KP, Suomi SJ (2002) Serotonin transporter gene polymorphism, differential early rearing, and behavior in rhesus monkey neonates. Mol Psychiatr 7:1058–1063
30. Crews D (2003) Sex determination: where environment and genetics meet. Evol Dev 5:50–55
31. Dawkins R (1982) The extended phenotype: The long reach of the gene. Oxford University Press, Oxford
32. de Beer G (1958) Embryos and ancestors (3rd Ed). Clarendon Press, Oxford
33. Denenberg VH, Rosenberg KM (1967) Non-genetic transmission of information. Nature 216:549–550
34. Dobzhansky T (1937) Genetics and the origins of species. Columbia University Press, New York
35. Ellis BJ (2004) Timing of pubertal maturation in girls: An integrated life history approach. Psychol Bull 130:920–958
36. Ellis BJ (2005) Determinants of pubertal timing: An evolutionary developmental approach. In: Ellis BJ, Bjorklund DF (eds) (2005) Origins of the social mind: Evolutionary psychology and child development. Guilford Press, New York, pp 164–188
37. Flinn MV (2006) Evolution and ontogeny of stress response to social challenges in the human child. Dev Rev 26:138–174
38. Francis DD, Diorio J, Liu D, Meaney MJ (1999) Nongenomic transmission across generations in maternal behavior and stress response in the rat. Science 286:1155–1158
39. Garstang W (1922) The theory of recapitulation: A critical re-statement of the biogenetic law. J Linn Soc Lond Zool 35:81–101
40. Geary DC, Bjorklund DF (2000) Evolutionary developmental psychology. Child Dev 71:57–65
41. Gergely G, Bekkering H, Kiraly I (2002) Rational imitation in preverbal infants. Nature 415:755
42. Gilissen R, Bakermans-Kranenburg MJ, van IJzendoorn MH, van der Veer R (2008) Parent-child relationship, temperament, and physiological reactions to fear-inducing film clips: Further evidence for differential susceptibility. J Exp Child Psychol 99:182195
43. Goodall J (1986) The chimpanzees of Gombe. Belknap Press of Harvard University, Cambridge
44. Gottlieb G (1987) The developmental basis of evolutionary change. J Comp Psychol 101:262–271
45. Gottlieb G (1992) Individual development and evolution: The genesis of novel behavior. Oxford University Press, New York
46. Gottlieb G (1998) Normally occurring environmental and behavioral influences on gene activity: From central dogma to probabilistic epigenesis. Psychol Rev 105:792–802
47. Gottlieb G (2000) Environmental and behavioral influences on gene activity. Curr Dir Psychol Sci 9:93–97
48. Gottlieb G (2002) Developmental-behavioral initiation of evolutionary change. Psychol Rev 109:211–218

49. Gottlieb G (2007) Probabilistic epigenesis. Dev Sci 10:1–11
50. Gottlieb G, Wahlsten D, Lickliter R (2006) The significance of biology for human development: A developmental psychobiological systems view. In: Damon W, Lerner RM (Gen Eds), Handbook of Child Psychology (6th edition), Lerner RM (Vol Ed), vol 1, Theoretical models of human development. Wiley, New York, pp 210–257
51. Gould SJ (1977) Ontogeny and phylogeny. Harvard University Press, Cambridge
52. Greene E (1996) Effect of light quality and larval diet on morph induction in the polymorphic caterpillar Nemoria arizonaria (Lepidoptera: Geometridae). Biol J Linn Soc 58:277–285
53. Haeffel GJ, Getchell M,. Koposov R a., Yrigollen CY, DeYoung CG, Klinteberg B, Oreland L, Ruchkin VV, Grigorenko EL (2008) Association between polymorphisms in the dopamine transporter gene and depression: evidence for a gene-environment interaction in a sample of juvenile detainees. Psychol Sci 19:62–69
54. Hamilton WD (1964) The genetical theory of social behavior. J Theor Biol 7:1–52
55. Harper L (2005) Epigenetic inheritance and the intergenerational transfer of experience. Psychol Bull 131:340–360
56. Hattori K (1998) Drivers of intelligence evolution in Homo: Sexual behavior, food acquisition and infant neoteny. Mank Q 39:127–146
57. Ho M-W (1998) Evolution. In: Greenberg G, Haraway MM (eds) Comparative psychology: A handbook. Garland, New York, pp 107–119
58. Ho M-W, Tucker C, Keeley D, Saunder PT (1983) Effects of successive generations of ether treatment on penetrance and expression of the bithorax phenocopy in Drosophila melanogaster. J Exp Zool 225:357–368
59. Horner V, Whiten A (2005) Causal knowledge and imitation/emulation switching in chimpanzees (Pan troglodytes) and children (Homo sapiens). Anim Cogn 8:164–181
60. Huether G (1998) Stress and the adaptive self-organization of neural connectivity during early childhood. Int J Dev Neurosci 16:297–306
61. Jablonka E (2001) The systems of inheritance. In: Oyama S, Griffiths PE, Gray RD (eds) Cycles of contingency: Developmental systems and evolution (pp 99–116). MIT Press, Cambridge
62. Jablonka E, Lamb M (1995) Epigenetic inheritance and evolution: The Lamarckian dimension. Oxford University Press, New York
63. Juffer F, IJzendoorn M (2005) Behavior problems and mental health referrals of international adoptees: A meta-analysis. J Am Med Assoc, 293:2501–2515
64. Laland KN, Odling-Smee FJ, Feldman MW (1999) Evolutionary consequences of niche construction and their implications for ecology. Proc Natl Acad Sci 96:10242–10247
65. Laland KN, Odling-Smee FJ, Feldman MW (2000) Niche construction, biological evolution, and cultural change. Behav Brain Sci 23:131–175
66. Laland KN, Odling-Smee FJ, Feldman MW (2001) Cultural niche construction and human evolution. J Evol Biol 14:22–33
67. Leavens DA, Russell JL, Hopkins WD (2005) Intentionality as measured in the persistence and elaboration of communication by chimpanzees (Pan troglodytes). Child Dev 76:291–306
68. Lewontin RC (1983) Gene, organism and environment. In: Bendall DS (ed) Evolution from molecules to men. Cambridge University Press, Cambridge, pp 273–286
69. Lickliter R (1990) Premature visual stimulation accelerates intersensory functioning in bobwhite quail neonates. Dev Psychobiol 23:15–27
70. Lickliter R, Hellewell TB (1992) Contextual determinates of auditory learning in bobwhite quail embryos and hatchlings. Dev Psychobiol 25:17–31
71. Lickliter R, Lewkowicz DJ (1995) Intersensory experience and early perceptual development: Attenuated prenatal sensory stimulation affects postnatal auditory and visual responsiveness in bobwhite quail chicks (Colinus virginianus). Dev Psychol 31:609–618
72. Lickliter R, Schneider SM The role of development in evolutionary change: A view from comparative psychology. Int J Comp Psychol, in press
73. Lyons De, Young AG, Keil FC (2007) The hidden structure of overlimitation. Proc Natl Acad Sci 104:19751–19756
74. Maestripieri D (1998) Parenting styles of abusive mothers in group-living rhesus macaque monkeys. Anim Behav 55:1–11
75. Maestripieri D (2005) Early experience affects the intergenerational transmission of infant abuse in rhesus monkeys. Proc Natl Acad Sci 102:9726–9729
76. Maestripieri D, Carroll KA (1998) Child abuse and neglect: usefulness of the animal data. Psychol Bull 3:211–223
77. Maestripieri D, Tomaszycki M, Carroll KA (1999) Consistency and change in behavior of rhesus macaque abusive mothers with successive infants. Dev Psychobiol 34:29–35
78. Mameli M (2004) Nongenetic selection and nongenetic inheritance. Br J Philos Sci 55:35–71
79. Mayr E (1942) Systematics and the origins of species from the viewpoint of a zoologist. Columbia University Press, New York
80. Mayr E (1982) The growth of biological thought: Diversity, evolution, and inheritance. Belknap Press, Cambridge
81. McGrew WC, Tutin CEG (1978) Evidence for a social custom in wild chimpanzees? Man 13:243–251
82. McKinney ML (1998) Cognitive evolution by extending brain development: On recapitulation, progress, and other heresies. In: Langer J, Killen M (eds) Piaget, evolution, and development. Erlbaum, Mahwah, pp 9–31
83. McKinney ML (2000) Evolving behavioral complexity by extending development. In: Parker ST, Langer J, McKinney ML (eds) Biology, brains, and behavior: The evolution of human development. School of American Research Press, Santa Fe, pp 25–40
84. Meany MJ (2001) Maternal care, gene expression, and the transmission of individual differences in stress reactivity across generations. Ann Rev Neurosci 24:1161–1192
85. Miles L (1990) The cognitive foundations for reference in signing orangutan. In: Parker ST, Gibson KR (eds). "Language" and intelligence in monkeys and apes. Cambridge University Press, Cambridge, pp 511–539
86. Miller DB (1998) Epigenesis. In: Breenberg G, Haraway MM (eds) Comparative psychology: A handbook. Garland, New York, pp 105–106
87. Moffitt TW, Caspi A, Rutter M (2006) Measured gene-environment interactions in Psychology: Concepts, research strategies, and implications for research, intervention, and public understanding of genetics. Perspect Psychol Sci 1:5–27

88. Montagu MFA (1962) Time, morphology, and neoteny in the evolution of man. In: Montagu MFA (ed) Culture and the evolution of man. Oxford University Press, New York, pp 324–342
89. Montagu A (1989) Growing young, 2nd edn. Bergin and Garvey, Grandy
90. Nagell K, Olguin RS, Tomasello M (1993) Processes of social learning in the tool use of chimpanzees (Pan troglodytes) and human children (Homo sapiens). J Comp Psychol 107:174–186
91. Nelson CA, Zeanah CH, Fox NA, Marshall PJ, Smyke AT, Guthrie D (2007) Cognitive recovery in socially deprived young children: The Bucharest Early Intervention Project. Science 319:1937–1940
92. O'Connor TG, Rutter M, Beckett C, Keaveny L, Kreppner JM, and the English and Romanian Adoptees Study Team (1999) The effects of global severe privation on cognitive competence: Extension and longitudinal follow-up. Child Dev 71:376–390
93. Odling-Smee FJ (1988) Niche constructing phenotypes. In: Plotkin HC (ed) The role of Behaviour in Evolution. MIT Press, Cambridge, pp 73–132
94. Odling-Smee FJ, Laland KN, Feldman MW (2003) Niche construction: The neglected process in evolution. Princeton University Press, Princeton
95. Oyama S (2000a) The ontogeny of information: Developmental systems and evolution, 2nd edn. Duke University Press, Durham
96. Oyama S (2000b) Evolution's eye: A systems view of biology-culture divide. Duke University Press, Durham
97. Oyama S, Griffiths PE, Gray RD (eds) (2001) Cycles of contingency: Developmental systems and evolution. MIT Press, Cambridge
98. Patterson F (1978) Linguistic capabilities of a lowland gorilla. In: Peng F (ed) Sign language and language acquisition in man and ape. Westview Press, Boulder, pp 161–201
99. Ploeger A, van der Maas HLJ, Rajimakers MEJ (2008) Is evolutionary psychology a metatheory for psychology? A discussion of four major issues in psychology from an evolutionary developmental perspective. Psychol Inq 19:1–18
100. Povinelli DJ, Nelson K, Boysen S (1992) Comprehension of role reversal in chimpanzees: Evidence of empathy? Anim Behav 43:633–640
101. Povinelli DJ, Reaux DJ, Bierschwale DT, Allain AD, Simon BB (1997) Exploitation of pointing as a referential gesture in young children, but not adolescent chimpanzees. Cogn Dev 12:423–461
102. Raff RA (1996) The shape of life: Genes, development, and the evolution of animal form. Chicago University press, Chicago
103. Ressler RH (1966) Inherited enviromental influences on the operant behaviour of mice. J Comp Phys Psychol 61:264–267
104. Savage-Rumbaugh ES (1986) Ape language: From conditioned response to symbol. Columbia University Press, New York
105. Savage-Rumbaugh ES, McDonald K, Sevcik RA, Hopkins WD, Rubert E (1986) Spontaneous symbol acquisition and communicative use by pygmy chimpanzees (Pan paniscus). J Exp Psychol Gen 115:211–235
106. Schwartz JH (1999) Sudden origins: Fossils, genes, and the emergence of species. Wiley, New York
107. Shea BT (1989) Heterochrony in human evolution: The case for neoteny revisit(ed). Yearb Phys Anthr 32:69–101
108. Shea BT (2000) Current issues in the investigation of evolution by heterochrony, with emphasis on the debate over human neoteny. In: Parker ST, Langer J, McKinney ML (eds) Biology, brains, and behavior: The evolution of human development. School of American Research Press, Santa Fe, pp 181–213
109. Simpson GG (1944) Tempo and mode in evolution. Columbia University Press, New York
110. Skeels HM (1966) Adult status of children with contrasting early life experiences. Monogr Soc Res Child Dev 31(3, Serial No. 105)
111. Stright AD, Gallagher KC, Kelly K (2008) Infant temperament moderates relations between maternal parenting in early childhood and children's adjustment in first grade. Child Dev 79:186–200
112. Suomi S (1995) Influence of attachment theory on ethological studies of biobehavioral development in nonhuman primates. In: Goldberg S, Muir R, Kerr J (eds) Attachment theory: Social, developmental and clinical perspectives. Analytic Press, Hillsdale, pp 185–202
113. Suomi SJ (1999) Attachment in rhesus monkeys. In: Cassidy J, Shaver PR (eds) Handbook of attachment: Theory, research, and clinical applications. Guilford Press, New York
114. Suomi SJ (2004) How gene-environment interactions shape biobehavioral development: Lessons from studies with rhesus monkeys. Res Human Dev 1:205–222
115. Suomi SJ, Harlow H (1972) Social rehabilitation of isolate-reared monkeys. Dev Psychol 6:487–496
116. Tomasello M (1998) Emulation learning and cultural learning. Behav Brain Sci 21:703–704
117. Tomasello M (2000) Culture and cognitive development. Curr Dir Psychol Sci 9:37–40
118. Tomasello M, Carpenter M, Call J, Behne T, Moll H (2005) Understanding and sharing intentions: the origins of cultural cognition. Behav Brain Sci 28:675–735
119. Tomasello M, Davis-Dasilva M, Camak L, Bard K (1987) Observational learning of tool-use by young chimpanzees. Hum Evol 2:175–183
120. Tomasello M, Kruger AC, Ratner HH (1993) Cultural learning. Behav Brain Sci 16:495–552
121. Tomasello M, Savage-Rumbaugh S, Kruger AC (1993) Imitative learning of actions on objects by children, chimpanzees, and enculturated chimpanzees. Child Dev 64:1688–1705
122. Waddington CH (1975) The evolution of an evolutionist. Cornell University Press, Ithaca
123. Weismann A (1892) Das Keimplasma: Eine Theorie der Vererbung. Gustav Fischer, Jena
124. West-Eberhard MJ (2003) Developmental plasticity and evolution. Oxford University Press, New York
125. Whiten A (2007) Pan African culture: Memes and genes in wild chimpanzees. Proc Natl Acad Sci 104:17559–17560
126. Whiten A, Goodall J, McGrew WC, Nishida T, Reynolds V, Sugiyama Y, Tutin CEG, Wrangham RW, Boesch C (1999) Cultures in chimpanzees. Nature 399:682–685
127. Whiten A, Custance DM, Gomez JC, Teixidor P, Bard KA (1996) Imitative learning of artificial fruit processing in children (Homo sapiens) and chimpanzees (Pan troglodytes). J Comp Psychol 110:3–14
128. Yunger JL, Bjorklund DF (2004) An assessment of generalization of imitation in two enculturated orangutans (Pongo pygmaeus). J Comp Psychol 118:242–246

Diagrammatic Methods in Classical Perturbation Theory

GUIDO GENTILE
Dipartimento di Matematica, Università di Roma Tre, Roma, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Examples
Trees and Graphical Representation
Small Divisors
Multiscale Analysis
Resummation
Generalizations
Conclusions and Future Directions
Bibliography

Glossary

Dynamical system Let $W \subseteq \mathbb{R}^N$ be an open set and $f\colon W \times \mathbb{R} \to \mathbb{R}^N$ be a smooth function. The ordinary differential equation $\dot{x} = f(x,t)$ on W defines a continuous dynamical system. A discrete dynamical system on W is defined by a map $x \to x\prime = F(x)$, with F depending smoothly on x.

Hamiltonian system Let $\mathcal{A} \subseteq \mathbb{R}^d$ be an open set and $\mathcal{H}\colon \mathcal{A} \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ be a smooth function ($\mathcal{A} \times \mathbb{R}^d$ is called the phase space). Consider the system of ordinary differential equations $\dot{q}_k = \partial\mathcal{H}(q,p,t)/\partial p_k$, $\dot{p}_k = -\partial\mathcal{H}(q,p,t)/\partial q_k$, for $k = 1,\ldots,d$. The equations are called Hamilton equations, and $\mathcal{H}$ is called a Hamiltonian function. A dynamical system described by Hamilton equations is called a Hamiltonian system.

Integrable system A Hamiltonian system is called integrable if there exists a system of coordinates $(\alpha, A) \in \mathbb{T}^d \times \mathbb{R}^d$, called angle-action variables, such that in these coordinates the motion is $(\alpha, A) \to (\alpha + \omega(A)t, A)$, for some smooth function $\omega(A)$. Hence in these coordinates the Hamiltonian function $\mathcal{H}$ depends only on the action variables, $\mathcal{H} = \mathcal{H}_0(A)$.

Invariant torus Given a continuous dynamical system we say that the motion occurs on an invariant d-torus if it takes place on a d-dimensional manifold and its position on the manifold is identified through a coordinate in $\mathbb{T}^d$. In an integrable Hamiltonian system all phase space is filled by invariant tori. In a quasi-integrable system the KAM theorem states that most of the invariant tori persist under perturbation, in the sense that the relative Lebesgue measure of the fraction of phase space filled by invariant tori tends to 1 as the perturbation tends to disappear. The persisting invariant tori are slight deformations of the unperturbed invariant tori.

Quasi-integrable system A quasi-integrable system is a Hamiltonian system described by a Hamiltonian function of the form $\mathcal{H} = \mathcal{H}_0(A) + \varepsilon f(\alpha, A)$, with (α, A) angle-action variables, ε a small real parameter and f periodic in its arguments α.

Quasi-periodic motion Consider the motion $\alpha \to \alpha + \omega t$ on $\mathbb{T}^2$, with $\omega = (\omega_1, \omega_2)$. If ω_1/ω_2 is rational, the motion is periodic, that is there exists $T > 0$ such that $\omega_1 T = \omega_2 T = 0 \bmod 2\pi$. If ω_1/ω_2 is irrational, the motion never returns to its initial value. On the other hand it densely fills $\mathbb{T}^2$, in the sense that it comes arbitrarily close to any point of $\mathbb{T}^2$. We say in that case that the motion is quasi-periodic. The definition extends to $\mathbb{T}^d$, $d > 2$: a linear motion $\alpha \to \alpha + \omega t$ on $\mathbb{T}^d$ is quasi-periodic if the components of ω are rationally independent, that is if $\omega \cdot \nu = \omega_1\nu_1 + \cdots + \omega_d\nu_d = 0$ for $\nu \in \mathbb{Z}^d$ if and only if $\nu = 0$ ($a \cdot b$ is the standard scalar product between the two vectors a, b). More generally we say that a motion on a manifold is quasi-periodic if, in suitable coordinates, it can be described as a linear quasi-periodic motion. The vector ω is usually called the frequency or rotation vector.

Renormalization group By renormalization group one denotes the set of techniques and concepts used to study problems where there are some scale invariance properties. The basic mechanism consists in considering equations depending on some parameters and defining some transformations on the equations, including a suitable rescaling, such that after the transformation the equations can be expressed, up to irrelevant corrections, in the same form as before but with new values for the parameters.

Torus The 1-torus $\mathbb{T}$ is defined as $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$, that is the set of real numbers defined modulo 2π (this means that x is identified with y if $x - y$ is a multiple of 2π). So it is the natural domain of an angle. One defines the d-torus $\mathbb{T}^d$ as a product of d 1-tori, that is $\mathbb{T}^d = \mathbb{T} \times \cdots \times \mathbb{T}$. For instance one can imagine $\mathbb{T}^2$ as a square with the opposite sides glued together.

Tree A graph is a collection of points, called nodes, and of lines which connect the nodes. A walk on the graph is a sequence of lines such that any two successive lines in the sequence share a node; a walk is nontrivial if it contains at least one line. A tree is a planar graph with

no closed loops, that is, such that there is no nontrivial walk connecting any node to itself. An oriented tree is a tree with a special node such that all lines of the tree are oriented toward that node. If we add a further oriented line connecting the special node to another point, called the root, we obtain a rooted tree (see Fig. 1 in Sect. "Trees and Graphical Representation").

Definition of the Subject

Recursive equations naturally arise whenever a dynamical system is considered in the regime of perturbation theory; for an introductory article on perturbation theory see ► Perturbation Theory. A classical example is provided by Celestial Mechanics, where perturbation series, known as Lindstedt series, are widely used; see Gallavotti [21] and ► Perturbation Theory in Celestial Mechanics.

A typical problem in Celestial Mechanics is to study formal solutions of given ordinary differential equations in the form of expansions in a suitable small parameter, the perturbation parameter. In the case of quasi-periodic solutions, the study of the series, in particular of its convergence, is made difficult by the presence of the small divisors – which will be defined later on. Under some non-resonance condition on the frequency vector, one can show that the series are well-defined to any order. The first proof of such a property was given by Poincaré [53], even if the convergence of the series remained an open problem up to the advent of KAM theory – an account can be found in Gallavotti [17] and in Arnold et al. [2]; see also ► Kolmogorov–Arnold–Moser (KAM) Theory. KAM is an acronym standing for Kolmogorov [47], Arnold [1] and Moser [51], who proved in the middle of last century the persistence of most of invariant tori for quasi-integrable systems.

Kolmogorov and Arnold proofs apply to analytic Hamiltonian systems, while Moser's approach deals also with the differentiable case; the smoothness condition on the Hamiltonian function was thereafter improved by Pöschel [54]. In the analytic case, the persisting tori turn out to be analytic in the perturbation parameter, as explicitly showed by Moser [52]. In particular, this means that the perturbation series not only are well-defined, but also converge. However, a systematic analysis with diagrammatic techniques started only recently after the pioneering, fundamental works by Eliasson [16] and Gallavotti [18], and were subsequently extended to many other problems with small divisors, including dynamical systems with infinitely many degrees of freedom, such as nonlinear partial differential equations, and non-Hamiltonian systems. Some of these extensions will be discussed in Sect. "Generalizations".

From a technical point of view, the diagrammatic techniques used in classical perturbation theory are strongly reminiscent of the Feynman diagrams used in quantum field theory: this was first pointed out by Gallavotti [18]. Also the multiscale analysis used to control the small divisors is typical of renormalization group techniques, which have been successfully used in problems of quantum field theory, statistical mechanics and classical mechanics; see Gallavotti [20] and Gentile & Mastropietro [38] for some reviews.

Note that there exist other renormalization group approaches to the study of dynamical systems, and of KAM-like problems in particular, different from that outlined in this article. By confining ourselves to the framework of problems of KAM-type, we can mention the paper by Bricmont et al. [11], which also stressed the similarity of the technique with quantum field theory, and the so called dynamical renormalization group method – see MacKay [50] – which recently produced rigorous proofs of persistence of quasi-periodic solutions; see for instance Koch [46] and Khanin et al. [45].

Introduction

Consider the ordinary differential equation on $\mathbb{R}^d$

$$Du = G(u) + \varepsilon F(u) \,, \tag{1}$$

where D is a pseudo-differential operator and G, F are real analytic functions. Assume that (1) admits a solution $u^{(0)}(t)$ for $\varepsilon = 0$, that is $Du^{(0)} = G(u^{(0)})$. The problem we are interested in is to investigate whether there exists a solution of (1) which reduces to $u^{(0)}$ as $\varepsilon \to 0$. For simplicity assume $G = 0$ in the following.

The first attempt one can try is to look for solutions in the form of power series in ε,

$$u(t) = \sum_{k=0}^{\infty} \varepsilon^k u^{(k)}(t) \,, \tag{2}$$

which, inserted into (1), when equating the left and right hand sides order by order, gives the list of recursive equations $Du^{(0)} = 0$, $Du^{(1)} = F(u^{(0)})$, $Du^{(2)} = \partial_u F(u^{(0)}) u^{(1)}$, and so on. In general to order $k \geq 1$ one has

$$Du^{(k)} = \sum_{s=0}^{k-1} \frac{1}{s!} \partial_u^s F(u^{(0)}) \sum_{\substack{k_1+\cdots+k_s=k-1 \\ k_i \geq 1}} u^{(k_1)} \ldots u^{(k_s)} \,, \tag{3}$$

where $\partial_u^s F$, the sth derivative of F, is a tensor with $s+1$ indices (s must be contracted with the vectors

$u^{(k_1)}, \ldots, u^{(k_s)}$), and the term with $s = 0$ in the sum has to be interpreted as $F(u^{(0)})$ and appears only for $k = 1$.

For instance for $F(u) = u^3$ the first orders give

$$\begin{aligned} Du^{(1)} &= u^{(0)3} , \\ Du^{(2)} &= 3u^{(0)2}u^{(1)} , \\ Du^{(3)} &= 3u^{(0)2}u^{(2)} + 3u^{(0)}u^{(1)2} , \\ Du^{(4)} &= 3u^{(0)2}u^{(3)} + 6u^{(0)}u^{(1)}u^{(2)} + u^{(1)3} , \end{aligned} \tag{4}$$

as is easy to check.

If the operator D can be inverted then the recursions (3) provide an algorithm to compute the functions $u^{(k)}(t)$. In that case we say that (2) defines a *formal power series*: by this we mean that the functions $u^{(k)}(t)$ are well-defined for all $k \geq 0$. Of course, even if this can be obtained, there is still the issue of the convergence of the series that must be dealt with.

Examples

In this section we consider a few paradigmatic examples of dynamical systems which can be described by equations of the form (1).

A Class of Quasi-integrable Hamiltonian Systems

Consider the Hamiltonian system described by the Hamiltonian function

$$\mathcal{H}(\alpha, A) = \tfrac{1}{2}A^2 + \varepsilon f(\alpha) , \tag{5}$$

where $(\alpha, A) \in \mathbb{T}^d \times \mathbb{R}^d$ are angle-action variables, with $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$, f is a real analytic function, 2π-periodic in each of its arguments, and $A^2 = A \cdot A$, if (here and henceforth) $\cdot$ denotes the standard scalar product in $\mathbb{R}^d$, that is $a \cdot b = a_1 b_1 + \cdots + a_d b_d$. Assume also for simplicity that f is a trigonometric polynomial of degree N.

The corresponding Hamilton equations are (we shorten $\partial_x = \partial/\partial_x$)

$$\begin{aligned} \dot{\alpha} &= \partial_A \mathcal{H}(\alpha, A) = A , \\ \dot{A} &= -\partial_\alpha \mathcal{H}(\alpha, A) = -\varepsilon \partial_\alpha f(\alpha) , \end{aligned}$$

which can be written as an equation involving only the angle variables:

$$\ddot{\alpha} = -\varepsilon \partial_\alpha f(\alpha) , \tag{6}$$

which is of the form (1) with $u = \alpha$, $G = 0$, $F = -\partial_\alpha f$, and $D = \mathrm{d}^2/\mathrm{d}t^2$.

For $\varepsilon = 0$, $\alpha^{(0)}(t) = \alpha_0 + \omega t$ is a solution of (6) for any choice of $\alpha_0 \in \mathbb{T}^d$ and $\omega \in \mathbb{R}^d$. Take for simplicity $\alpha_0 = 0$: we shall see that this choice makes sense. We say that for $\varepsilon = 0$ the Hamiltonian function (5) describes a system of d *rotators*.

We call ω the *frequency vector*, and we say that ω is irrational if its components are rationally independent, that is if $\omega \cdot \nu = 0$ for $\nu \in \mathbb{Z}^d$ if and only if $\nu = 0$. For irrational ω the solution $\alpha^{(0)}(t)$ describes a *quasi-periodic motion* with frequency vector ω, and it densely fills $\mathbb{T}^d$.

Then (3) becomes

$$\begin{aligned} \ddot{\alpha}^{(k)} &= -\left[\varepsilon f(\alpha)\right]^{(k)} \\ &:= -\sum_{s=0}^{k-1} \frac{1}{s!} \partial_\alpha^{s+1} f(\omega t) \sum_{\substack{k_1+\cdots+k_s=k-1 \\ k_i \geq 1}} \alpha^{(k_1)} \ldots \alpha^{(k_s)} . \end{aligned} \tag{7}$$

We look for a quasi-periodic solution of (6), that is a solution of the form $\alpha(t) = \omega t + h(\omega t)$, with $h(\omega t) = O(\varepsilon)$. We call h the *conjugation function*, as it "conjugates" (that is, maps) the perturbed solution $\alpha(t)$ to the unperturbed solution ωt. In terms of the function h (6) becomes

$$\ddot{h} = -\varepsilon \partial_\alpha f(\omega t + h) , \tag{8}$$

where ∂_α denotes derivative with respect to the argument. Then (8) can be more conveniently written in Fourier space, where the operator D acts as a multiplication operator.

If we write

$$h(\omega t) = \sum_{\nu \in \mathbb{Z}^d} \mathrm{e}^{i\omega\cdot\nu t} h_\nu , \qquad h_\nu = \sum_{k=1}^{\infty} \varepsilon^k h_\nu^{(k)} , \tag{9}$$

and insert (9) into (8) we obtain

$$\begin{aligned} (\omega \cdot \nu)^2 h_\nu^{(k)} &= \left[\varepsilon \partial_\alpha f(\alpha)\right]_\nu^{(k)} \\ &:= \sum_{s=0}^{k-1} \sum_{\substack{k_1+\cdots+k_s=k-1 \\ k_i \geq 1}} \sum_{\substack{\nu_0+\nu_1+\cdots+\nu_s=\nu \\ \nu_i \in \mathbb{Z}^n}} \frac{1}{s!} \\ &\quad \cdot (i\nu_0)^{s+1} f_{\nu_0} h_{\nu_1}^{(k_1)} \ldots h_{\nu_s}^{(k_s)} . \end{aligned} \tag{10}$$

These equations are well-defined to all orders provided $[\varepsilon \partial_\alpha f(\alpha)]_\nu^{(k)} = 0$ for all ν such that $\omega \cdot \nu = 0$. If ω is an irrational vector we need $[\varepsilon \partial_\alpha f(\alpha)]_0^{(k)} = 0$ for the equations to be well-defined. In that case the coefficients $h_0^{(k)}$ are left undetermined, and we can fix them arbitrarily to vanish (which is a convenient choice).

We shall see that under some condition on ω a quasi-periodic solution $\alpha(t)$ exists, and densely fills a d-dimensional manifold. The analysis carried out above for $\alpha_0 = 0$ can be repeated unchanged for all values of $\alpha_0 \in \mathbb{T}^d$: α_0 represents the *initial phase* of the solution, and by varying α_0 we cover all the manifold. Such a manifold can be parametrized in terms of α_0, so it represents an invariant torus for the perturbed system.

A Simplified Model with No Small Divisors

Consider the same equation as (8) with $D = \mathrm{d}^2/\mathrm{d}t^2$ replaced by -1, that is

$$h = \varepsilon \partial_\alpha f(\omega t + h) \,. \tag{11}$$

Of course in this case we no longer have a differential equation; still, we can look again for quasi-periodic solutions $h(\omega t) = O(\varepsilon)$ with frequency vector ω. In such a case in Fourier space we have

$$\begin{aligned} h_\nu^{(k)} &= \left[\varepsilon \partial_\alpha f(\alpha)\right]_\nu^{(k)} \\ &:= \sum_{s=0}^{k-1} \sum_{\substack{k_1+\cdots+k_s=k-1 \\ k_i \geq 1}} \sum_{\substack{\nu_0+\nu_1+\cdots+\nu_s=\nu \\ \nu_i \in \mathbb{Z}^n}} \frac{1}{s!} (i\nu_0)^{s+1} \\ &\quad \cdot f_{\nu_0} h_{\nu_1}^{(k_1)} \ldots h_{\nu_s}^{(k_s)} \,. \end{aligned} \tag{12}$$

For instance if $d = 1$ and $f(\alpha) = \cos\alpha$ the equation, which is known as the *Kepler equation*, can be explicitly solved by the Lagrange inversion theorem [55], and gives

$$h_\nu^{(k)} = \begin{cases} \dfrac{-i(-1)^{k+(\nu-k)/2}}{2^k((k-\nu)/2)!((k+\nu)/2)!} \,, & |\nu| \leq k, \\ & \nu + k \text{ even} \,, \\ 0, & \text{otherwise} \,. \end{cases} \tag{13}$$

We shall show in Sect. "Small Divisors" that a different derivation can be provided by using the forthcoming diagrammatic techniques.

The Standard Map

Consider the finite difference equation

$$D\alpha = -\varepsilon \sin\alpha \,, \tag{14}$$

on $\mathbb{T}$, where now D is defined by

$$D\alpha(\psi) := 2\alpha(\psi) - \alpha(\psi + \omega) - \alpha(\psi - \omega) \,. \tag{15}$$

By writing $\alpha = \psi + h(\psi)$, (14) becomes

$$Dh = -\varepsilon \sin(\psi + h) \,, \tag{16}$$

which is the functional equation that must be solved by the *conjugation function* of the *standard map*

$$\begin{cases} x' = x + y + \varepsilon \sin x \,, \\ y' = y + \varepsilon \sin x \,. \end{cases} \tag{17}$$

In other words, by writing $x = \psi + h(\psi)$ and $y = \omega + h(\psi) - h(\psi - \omega)$, with (ψ, ω) solving (17) for $\varepsilon = 0$, that is $(\psi', \omega') = (\psi + \omega, \omega)$, we obtain a closed-form equation for h, which is exactly (16).

In Fourier space the operator D acts as $D\colon \mathrm{e}^{i\nu\psi} \to 4\sin^2(\omega\nu/2)\mathrm{e}^{i\nu\psi}$, so that, by expanding h according to (9), we can write (16) as

$$\begin{aligned} h_\nu^{(k)} &= \frac{1}{4\sin^2(\omega\nu/2)} \sum_{s=0}^{k-1} \sum_{\substack{k_1+\cdots+k_s=k-1 \\ k_i \geq 1}} \sum_{\substack{\nu_0+\nu_1+\cdots+\nu_s=\nu \\ \nu_i \in \mathbb{Z}}} \frac{1}{s!} \\ &\quad \cdot (i\nu_0)^{s+1} f_{\nu_0} h_{\nu_1}^{(k_1)} \ldots h_{\nu_s}^{(k_s)} \,, \end{aligned} \tag{18}$$

where $\nu_0 = \pm 1$ and $f_{\pm 1} = 1/2$.

Note that (17) is a discrete dynamical system. However, when passing to Fourier space, (18) acquires the same form as for the continuous dynamical systems previously considered, simply with a different kernel for D. In particular if we replace D with 1 we recover the Kepler equation.

The number ω is called the *rotation number*. We say that ω is irrational if the vector $(2\pi, \omega)$ is irrational according to the previous definition.

Trees and Graphical Representation

Take ω to be irrational. We study the recursive equations

$$\begin{cases} h_\nu^{(k)} = g(\omega \cdot \nu) \left[\varepsilon \partial_\alpha f(\alpha)\right]_\nu^{(k)} \,, & \nu \neq 0 \,, \\ \left[\varepsilon \partial_\alpha f(\alpha)\right]_0^{(k)} = 0 \,, & \nu = 0 \,, \end{cases} \tag{19}$$

where the form of g depends on the particular model we are investigating. Hence one has either $g(\omega \cdot \nu) = (\omega \cdot \nu)^{-2}$ or $g(\omega \cdot \nu) = 1$ or $g(\omega \cdot \nu) = (2\sin(\omega\nu/2))^{-2}$ according to models described in Sect. "Examples".

For $\nu \neq 0$ we have equations which express the coefficients $h_\nu^{(k)}$, $\nu \in \mathbb{Z}^d$, in terms of the coefficients $h_\nu^{(k')}$, $\nu \in \mathbb{Z}^d$, with $k' < k$, provided the equations for $\nu = 0$ are satisfied for all $k \geq 1$. Recursive equations, such as (19), naturally lead to a graphical representation in terms of trees.

Trees

A connected *graph* $\mathcal{G}$ is a collection of points (nodes) and lines connecting all of them. Denote with $N(\mathcal{G})$ and $L(\mathcal{G})$ the set of nodes and the set of lines, respectively. A path between two nodes is the minimal subset of $L(\mathcal{G})$ connecting the two nodes. A graph is planar if it can be drawn in a plane without graph lines crossing.

A *tree* is a planar graph $\mathcal{G}$ containing no closed loops. Consider a tree $\mathcal{G}$ with a single special node v_0: this introduces a natural partial ordering on the set of lines and

nodes, and one can imagine that each line carries an arrow pointing toward the node v_0. We add an extra oriented line ℓ_0 exiting the special node v_0; the added line will be called the *root line* and the point it enters (which is not a node) will be called the *root* of the tree. In this way we obtain a *tree* θ defined by $N(\theta) = N(\mathcal{G})$ and $L(\theta) = L(\mathcal{G}) \cup \ell_0$. A *labeled tree* is a rooted tree θ together with a label function defined on the sets $L(\theta)$ and $N(\theta)$.

We call *equivalent* two rooted trees which can be transformed into each other by continuously deforming the lines in the plane in such a way that the lines do not cross each other. We can extend the notion of equivalence also to labeled trees, by considering equivalent two labeled trees if they can be transformed into each other in such a way that the labels also match. In the following we shall deal mostly with nonequivalent labeled trees: for simplicity, where no confusion can arise, we call them just trees.

Given two nodes $v, w \in N(\theta)$, we say that $w \prec v$ if v is on the path connecting w to the root line. We can identify a line ℓ through the node v it exits by writing $\ell = \ell_v$.

We call *internal nodes* the nodes such that there is at least one line entering them, and *end-points* the nodes which have no entering line. We denote with $L(\theta)$, $V(\theta)$ and $E(\theta)$ the set of lines, internal nodes and end-points, respectively. Of course $N(\theta) = V(\theta) \cup E(\theta)$.

The number of unlabeled trees with k nodes (and hence with k lines) is bounded by 2^{2k}, which is a bound on the number of random walks with $2k$ steps [38].

For each node v denote by $S(v)$ the set of the lines entering v and set $s_v = |S(v)|$. Hence $s_v = 0$ if v is an end-node, and $s_v \geq 1$ if v is an internal node. One has

$$\sum_{v \in N(\theta)} s_v = \sum_{v \in V(\theta)} s_v = k - 1\,; \tag{20}$$

this can be easily checked by induction on the order of the tree. An example of unlabeled tree is represented in Fig. 1.

For further details on graphs and trees we refer to the literature; cf. for instance Harary [43].

Labels and Diagrammatic Rules

We associate with each node $v \in N(\theta)$ a *mode* label $\nu_v \in \mathbb{Z}^d$, and with each line $\ell \in L(\theta)$ a *momentum* label $\nu_\ell \in \mathbb{Z}^d$, with the constraint

$$\nu_{\ell_v} = \sum_{\substack{w \in N(\theta) \\ w \preceq v}} \nu_w = \nu_v + \sum_{\ell \in S(v)} \nu_\ell\,, \tag{21}$$

which represents a *conservation rule* for each node.

Call $\mathcal{T}_{k,\nu}$ the set of all trees θ with k nodes and momentum ν associated with the root line. We call k and ν the order and the momentum of θ, respectively.

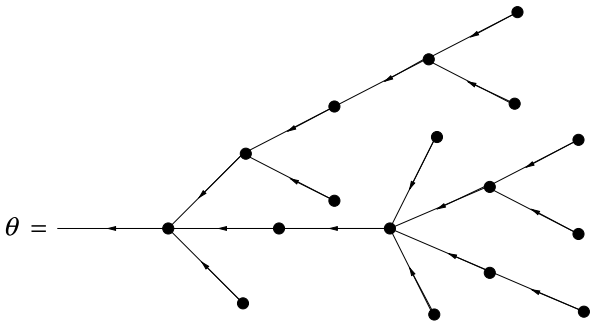

Diagrammatic Methods in Classical Perturbation Theory, Figure 1
An unlabeled tree with 17 nodes

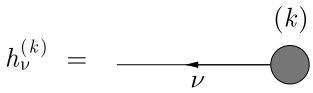

Diagrammatic Methods in Classical Perturbation Theory, Figure 2
Graph element

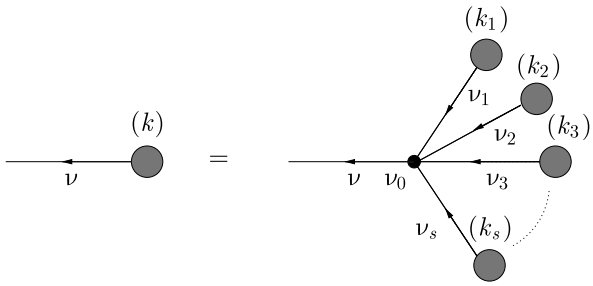

Diagrammatic Methods in Classical Perturbation Theory, Figure 3
Graphical representation of the recursive equations

We want to show that trees naturally arise when studying (19). Let $h_\nu^{(k)}$ be represented with the graph element in Fig. 2 as a line with label ν exiting from a ball with label (k).

Then we can represent (19) graphically as depicted in Fig. 3. Simply represent each factor $h_{\nu_i}^{(k_i)}$ on the right hand side as a graph element according to Fig. 2. The lines of all such graph elements enter the same node v_0. This is a graphical expedient to recall the conservation rule: the momentum ν of the root line is the sum of the mode label ν_0 of the node v_0 plus the sum of the momenta of the lines entering v_0.

The first few orders $k \leq 4$ are as depicted in Fig. 4. For each node the conservation rule (21) holds: for instance for $k = 2$ one has $\nu = \nu_1 + \nu_2$, for $k = 3$ one has $\nu = \nu_1 + \nu_{\ell_1}$ and $\nu_{\ell_1} = \nu_2 + \nu_3$ in the first tree and $\nu = \nu_1 + \nu_2 + \nu_3$ in

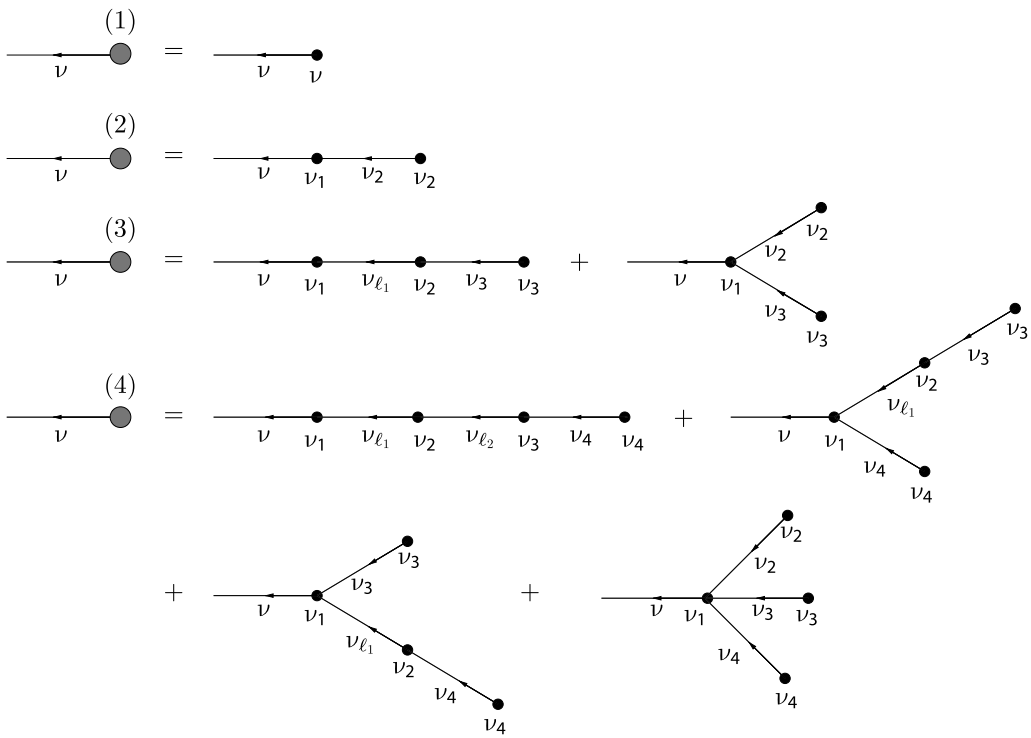

Diagrammatic Methods in Classical Perturbation Theory, Figure 4
Trees of lower orders

the second tree, and so on. Moreover one has to sum over all possible choices of the labels ν_v, $v \in N(\theta)$, which sum up to ν.

Given any tree $\theta \in \mathcal{T}_{k,\nu}$ we associate with each node $v \in N(\theta)$ a *node factor* F_v and with each line $\ell \in L(\theta)$ a *propagator* g_ℓ, by setting

$$F_v := \frac{1}{s_v!}(i\nu_v)^{s_v+1} f_{\nu_v}\,, \qquad g_\ell := g(\omega\cdot\nu_\ell)\,, \tag{22}$$

and define the *value* of the tree θ as

$$\mathrm{Val}(\theta) := \Big(\prod_{v\in N(\theta)} F_v\Big)\Big(\prod_{\ell\in L(\theta)} g_\ell\Big)\,. \tag{23}$$

The propagators g_ℓ are scalars, whereas each F_v is a tensor with $s_v + 1$ indices, which can be associated with the $s_v + 1$ lines entering or exiting v. In (23) the indices of the tensors F_v must be contracted: this means that if a node v is connected to a node v' by a line ℓ then the indices of F_v and $F_{v'}$ associated with ℓ are equal to each other, and eventually one has to sum over all the indices except that associated with the root line.

For instance the value of the tree in Fig. 4 contributing to $h_\nu^{(2)}$ is given by

$$\mathrm{Val}(\theta) = (i\nu_1)^2 f_{\nu_1}(i\nu_2)\, f_{\nu_2} g(\omega\cdot\nu) g(\omega\cdot\nu_2)\,,$$

with $\nu_1 + \nu_2 = \nu$, while the value of the last tree in Fig. 4 contributing to $h_\nu^{(4)}$ is given by

$$\mathrm{Val}(\theta) = \frac{(i\nu_1)^4}{3!} f_{\nu_1}(i\nu_2)\, f_{\nu_2}(i\nu_3)\, f_{\nu_3}(i\nu_4)\, f_{\nu_4} \cdot g(\omega\cdot\nu) g(\omega\cdot\nu_2) g(\omega\cdot\nu_3) g(\omega\cdot\nu_4)\,,$$

with $\nu_1 + \nu_2 + \nu_3 + \nu_4 = \nu$.

It is straightforward to prove that one can write

$$h_\nu^{(k)} = \sum_{\theta\in\mathcal{T}_{k,\nu}} \mathrm{Val}(\theta)\,, \qquad \nu \neq 0\,, \quad k \geq 1\,. \tag{24}$$

This follows from the fact that the recursive equations (19) can be graphically represented through Fig. 3: one iterates the graphical representation of Fig. 3 until only graph elements of order $k = 1$ appear, and if θ is of order 1 (cf. Fig. 4) then $\mathrm{Val}(\theta) = (i\nu) f_\nu g(\omega\cdot\nu)$.

Each line $\ell \in L(\theta)$ can be seen as the root line of the tree consisting of all nodes and lines preceding ℓ. The choice $h_0^{(k)} = 0$ for all $k \geq 1$ implies that no line can have zero momentum: in other words we have $\nu_\ell \neq 0$ for all $\ell \in L(\theta)$.

Therefore in order to prove that (9) with $h_\nu^{(k)}$ given by (24) solves formally, that is order by order, the Eqs. (19),

we have only to check that $[\varepsilon\partial_\alpha f(\omega t + h(\omega t))]_0^{(k)} = 0$ for all $k \geq 1$.

If we define $g_\ell = 1$ for $\nu_\ell = 0$, then also the second relation in (19) can be graphically represented as in Fig. 3 by setting $\nu = 0$ and requiring $h_0^{(k)} = 0$, which yields that the sum of the values of all trees on the right hand side must vanish. Note that this is not an equation to solve, but just an identity that has to be checked to hold at all orders.

For instance for $k = 2$ (the case $k = 1$ is trivial) the identity $[\varepsilon\partial_\alpha f(\omega t + h(\omega t))]_0^{(2)} = 0$ reads (cf. the second line in Fig. 4)

$$\sum_{\nu_1+\nu_2=0} (i\nu_1)^2 f_{\nu_1}(i\nu_2) f_{\nu_2} g(\omega\cdot\nu_2) = 0\,,$$

which is found to be satisfied because the propagators are even in their arguments.

Such a cancellation can be graphically interpreted as follows. Consider the tree with mode labels ν_1 and ν_2, with $\nu_1 + \nu_2 = 0$: its value is $(i\nu_1)^2 f_{\nu_1}(i\nu_2) f_{\nu_2} g(\omega\cdot\nu_2)$. One can detach the root line from the node with mode label ν_1 and attach it to the node with mode label ν_2, and reverse the arrow of the other line so that it points toward the new root line. In this way we obtain a new tree (cf. Fig. 5): the value of the new tree is $(i\nu_1) f_{\nu_1}(i\nu_2)^2 f_{\nu_2} g(\omega\cdot\nu_1)$, where $g(\omega\cdot\nu_1) = g(-\omega\cdot\nu_2) = g(\omega\cdot\nu_2)$, so that the values of the two trees contain a common factor $(i\nu_1) f_{\nu_1}(i\nu_2) f_{\nu_2} g(\omega\cdot\nu_2)$ times an extra factor which is $(i\nu_1)$ for the first tree and $(i\nu_2)$ for the second tree. Hence the sum of the two values gives zero.

The cancellation mechanism described above can be generalized to all orders. Given a tree θ one considers all trees which can be obtained by detaching the root line and attaching to the other nodes of the tree, and by reversing the arrows of the lines (when needed) to make them point toward the root line. Then one sums together the values of all the trees so obtained: such values contain a common factor times a factor $i\nu_\nu$, if ν is the node which the root line exits (the only nontrivial part of the proof is to check that the combinatorial factors match each other: we refer to Gentile & Mastropietro [37] for details). Hence the sum gives zero, as the sum of all the mode labels vanishes.

For instance for $k = 3$ the cancellation operates by considering the three trees in Fig. 5: such trees can be considered to be obtained from each other by shifting the root line and consistently reversing the arrows of the lines.

Diagrammatic Methods in Classical Perturbation Theory, Figure 5
Trees to be considered together to prove that $[\varepsilon\partial f(\alpha)]_0^2 = 0$

In such a case the combinatorial factors of the node factors are different, because in the second tree the node factor associated with the node with mode label ν_2 contains a factor 1/2: on the other hand if $\nu_1 \neq \nu_3$ there are two nonequivalent trees with that shape (with the labels ν_1 and ν_3 exchanged between themselves), whereas if $\nu_1 = \nu_3$ there is only one such tree, but then the first and third trees are equivalent, so that only one of them must be counted. Thus, by using that $\nu_1 + \nu_2 + \nu_3 = 0$ – which implies $g(\omega\cdot(\nu_2+\nu_3)) = g(-\omega\cdot\nu_1)$ and $g(\omega\cdot(\nu_1+\nu_2)) = g(-\omega\cdot\nu_3)$ – in all cases we find that the sum of the values of the trees gives a common factor $(i\nu_1) f_{\nu_1}(i\nu_2)^2 f_{\nu_2}(i\nu_3) f_{\nu_3} g(\omega\cdot\nu_3) g(\omega\cdot\nu_1)$ times a factor 1 or 1/2 times $i(\nu_1 + \nu_2 + \nu_3)$, and hence vanishes: once more the property that g is even is crucial.

Small Divisors

We want to study the convergence properties of the series

$$h(\omega t) = \sum_{\nu\in\mathbb{Z}^d} e^{i\omega\cdot\nu t} h_\nu\,, \quad h_\nu = \sum_{k=1}^{\infty} \varepsilon^k h_\nu^{(k)}\,, \tag{25}$$

which has been shown to be well–defined as a formal power series for the models considered in Sect. "Examples".

Recall that the number of unlabeled trees of order k is bounded by 2^{2k}. To sum over the labels we can confine ourselves to the mode labels, as the momenta are uniquely determined by the mode labels. If f is a trigonometric polynomial of degree N, that is $f_\nu = 0$ for all ν such that $|\nu| := |\nu_1| + \cdots + |\nu_d| > N$, we have that $h_\nu^{(k)} = 0$ for all $|\nu| > kN$ (which can be easily proved by induction), and moreover we can bound the sum over the mode labels of any tree of order k by $(2N+1)^{dk}$. Finally we can bound

$$\prod_{\nu\in N(\theta)} |\nu_\nu|^{s_\nu+1} \leq \prod_{\nu\in N(\theta)} N^{s_\nu+1} \leq N^{2k}\,, \tag{26}$$

because of (20).

For the model (11), where $g_\ell = 1$ in (22), we can bound

$$\left|h_\nu^{(k)}\right| \leq \sum_{\nu\in\mathbb{Z}^d} \left|h_\nu^{(k)}\right| \leq 2^{2k}(2N+1)^{dk} N^{2k} \Phi^k\,, \qquad \Phi = \max_{|\nu|\leq N} |f_\nu|\,, \tag{27}$$

which shows that the series (25) converges for ε small enough, more precisely for $|\varepsilon| < \varepsilon_0$, with

$$\varepsilon_0 := C_0(4N^2\Phi(2N+1)^d)^{-1}\,, \tag{28}$$

where $C_0 = 1$. Hence the function $h(\omega t)$ in that case is analytic in ε. For $d = 1$ and $f(\alpha) = \cos\alpha$, we can easily

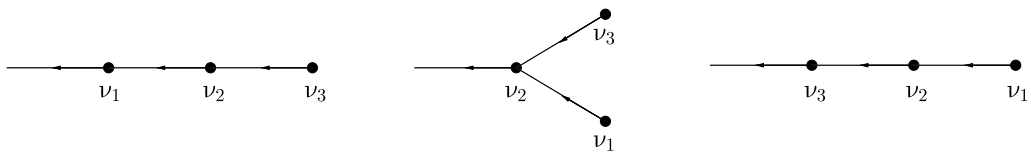

Diagrammatic Methods in Classical Perturbation Theory, Figure 6
Trees to be considered together to prove that $[\varepsilon\partial f(\alpha)]_0^3 = 0$

provide an exact expression for the coefficients $h_\nu^{(k)}$: all the computational difficulties reduce to a combinatorial check, which can be found in Gentile & van Erp [42], and the formula (13) is recovered.

However for the models where $g_\ell \neq 1$, the situation is much more involved: the propagators can be arbitrarily close to zero for ν large enough. This is the so-called *small divisor problem*. The series (25) is formally well–defined, assuming only an irrationality condition on ω. But to prove the convergence of the series, we need a stronger condition. For instance one can require the *standard Diophantine condition*

$$|\omega \cdot \nu| > \frac{\gamma}{|\nu|^\tau} \qquad \forall \nu \neq 0\,, \tag{29}$$

for suitable positive constants γ and τ. For fixed $\tau > d-1$, the sets of vectors which satisfy (29) for some constant $\gamma > 0$ has full Lebesgue measure in $\mathbb{R}^d$ [17]. We can also impose a weaker condition, known as the *Bryuno condition*, which can be expressed by requiring

$$\mathcal{B}(\omega) := \sum_{k=0}^{\infty} \frac{1}{2^k} \log \frac{1}{\min_{0<|\nu|\le 2^k} |\omega \cdot \nu|} < \infty\,. \tag{30}$$

We call *Diophantine vectors* and *Bryuno vectors* the vectors satisfying (29) and (30), respectively. Note that any Diophantine vector satisfies (30).

If we assume only analyticity on f – that is, we remove the assumption that f be a trigonometric polynomial, – then we need a Diophantine condition, such as (29) or (30), also to show that the series (25) are well-defined as formal power series: the condition is needed, as one can easily check, in order to sum over the mode labels.

In the case of the standard map $|\omega \cdot \nu|$ must be replaced with $\min_{p\in\mathbb{Z}} |\omega\nu - p|$, and the function $\mathcal{B}(\omega)$ can be expressed in terms of the *best approximants* of the number ω [15].

For the models with small divisors considered in Sect. "Examples" convergence of the series (25) can be proved for ω satisfying (29) or (30). But this requires more work, and, in particular, a detailed discussion of the product of propagators in (24). In the following we shall consider the case of Diophantine vectors, by referring to the bibliography for the Bryuno vectors (cf. Sect. "Generalizations").

Multiscale Analysis

Consider explicitly the case $g(\omega \cdot \nu) = (\omega \cdot \nu)^{-2}$ in (19) and ω satisfying (29). We can introduce a label characterizing the size of the propagator: we say that $\nu \in \mathbb{Z}^d$ is on *scale*

$$\begin{cases} n \ge 1\,, & \text{if } 2^{-n}\gamma \le |\omega \cdot \nu| < 2^{-(n-1)}\gamma\,, \\ n = 0\,, & \text{if } \gamma \le |\omega \cdot \nu|\,, \end{cases} \tag{31}$$

where γ is the constant appearing in (29), and we say that a line ℓ has a scale label $n_\ell = n$ if ν_ℓ is on scale n. If $\mathfrak{N}_n(\theta)$ denotes the number of lines $\ell \in L(\theta)$ with scale $n_\ell = n$, then we can bound in (23)

$$\Big| \prod_{\ell\in L(\theta)} g_\ell \Big| \le \gamma^{-2k} \prod_{n=0}^{\infty} 2^{2n\mathfrak{N}_n(\theta)}\,, \tag{32}$$

so that the problem is reduced to bounding $\mathfrak{N}_n(\theta)$.

The product of propagators gives problems when the small divisors "accumulate". To make more precise the idea of accumulation we introduce the notion of cluster. Once all lines of a tree θ have been given their scale labels, for any $n \ge 0$ we can identify the maximal connected sets of lines with scale not larger than n. If at least one among such lines has scale equal to n we say that the set is a *cluster* on scale n. For instance consider the tree in Fig. 1, and assign the mode labels to the nodes: this uniquely fixes the momenta, and hence the scale labels, of the lines. Suppose that we have found the scales as in Fig. 7a. Then a cluster decomposition as in Fig. 7b follows. Given a cluster T call $L(T)$ the set of lines of θ contained in T, and denote by $N(T)$ the set of nodes connected by such lines. We define $k_T = |N(T)|$ the order of the cluster T.

Any cluster has either one or no exiting line, and can have an arbitrary number of entering lines. We call *self-energy clusters* the clusters which have one exiting line and only one entering line and are such that both lines have the same momentum. This means that if T is a self-energy

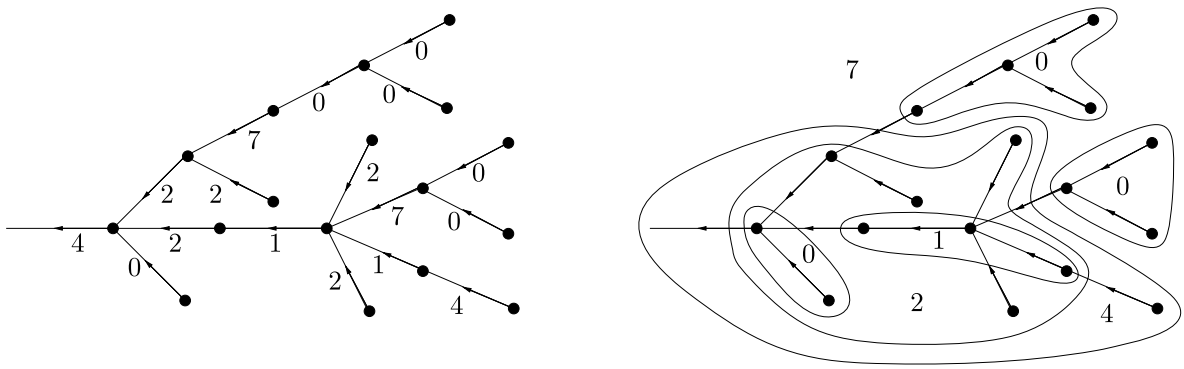

a b

Diagrammatic Methods in Classical Perturbation Theory, Figure 7
Example of clusters with their scales

cluster and ℓ_1 and ℓ_2 are the lines entering and exiting T, respectively, then $\nu_{\ell_1} = \nu_{\ell_2}$, so that

$$\sum_{v \in N(T)} \nu_v = 0 \,. \tag{33}$$

By definition of scale, both lines ℓ_1 and ℓ_2 have the same scale, say n, and that, by definition of cluster, one has $n_\ell < n$ for all $\ell \in L(T)$.

We define the *value* of the self-energy cluster T whose entering line has momentum ν as the matrix

$$\mathcal{V}_T(\omega \cdot \nu) := \Big(\prod_{v \in N(T)} F_v \Big) \Big(\prod_{\ell \in L(T)} g_\ell \Big) \,, \tag{34}$$

where all the indices of the node factors must be contracted except those associated with the line ℓ_1 entering T and with the line ℓ_2 exiting T.

We can extend the notion of self-energy clusters also to a single node, by saying that v is a self-energy cluster if $s_v = 1$ and the line entering v has the same momentum as the exiting line. In that case (34) has to be interpreted as $\mathcal{V}_T(\omega \cdot \nu) = F_v$: in particular it is independent of $\omega \cdot \nu$.

The simplest self-energy cluster one can think of consists of only one node v, but then (33) implies $\nu_v = 0$, so that $F_v = 0$, see (22), hence the corresponding value is zero. Thus the simplest non-trivial self-energy clusters contain at least two nodes, and are represented by the clusters T_1, T_2 and T_3 of Fig. 8. In all cases one has $\nu_1 + \nu_2 = 0$. By using the definition (34) one has $\mathcal{V}_{T_1}(x) = (i\nu_1)^2 f_{\nu_1}(i\nu_2)^2 f_{\nu_2} g(\omega \cdot \nu_2 + x)$ and $\mathcal{V}_{T_2}(x) = \mathcal{V}_{T_3}(x) = (i\nu_1)^3 f_{\nu_1}(i\nu_2) f_{\nu_2} g(\omega \cdot \nu_2)/2$, with $x = \omega \cdot \nu$. Hence for $x = 0$ the sum of the three values gives zero. It is not difficult to see that also the first derivatives of the values of all self-energy clusters of order 2 sum up to zero, that is the sum of the values of all possible self-energy clusters of order $k = 2$ gives zero up to order x^2. (The only *caveat* is that, if we want to derive $\mathcal{V}_T(x)$, the sharp multiscale decomposition in (31) can be a little annoying, hence it can be more convenient to replace it with a smooth decomposition through C^∞ compact support functions; we refer to the literature for details). This property generalizes to all orders k, and the underlying cancellation mechanism is essentially the same that ensures the validity of the second relation in (19).

The reason why it is important to introduce the self-energy clusters is that if we could neglect them then the product of small divisors would be controlled. Indeed, let us denote by $\mathfrak{R}_n(\theta)$ the number of lines on scale n which do exit a self-energy cluster, and set $\mathfrak{N}_n^*(\theta) = \mathfrak{N}_n(\theta) -$

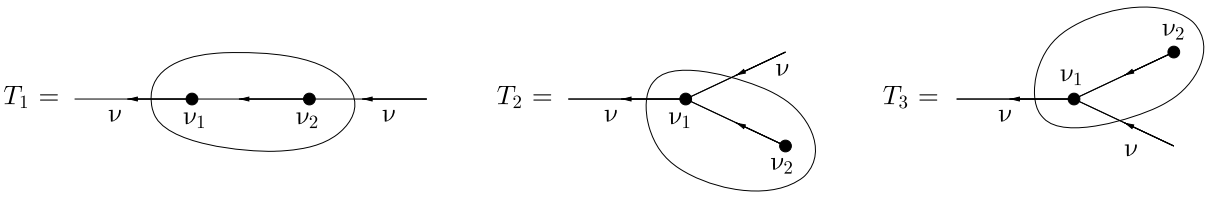

Diagrammatic Methods in Classical Perturbation Theory, Figure 8
Self-energy clusters of order $k = 2$

$\mathfrak{R}_n(\theta)$. Then an important result, known as the *Siegel–Bryuno lemma*, is that

$$\mathfrak{N}_n^*(\theta) \le c2^{-n/\tau}k\,, \tag{35}$$

for some constant c, where k is the order of θ and τ is the Diophantine exponent in (29).

The bound (35) follows from the fact that if $\mathfrak{N}_n^*(\theta) \neq 0$ then $\mathfrak{N}_n^*(\theta) \le E(n,k) := 2Nk2^{-(n-2)/\tau} - 1$, which can be proved by induction on k as follows. Given a tree θ let ℓ_0 be its root line, let $\ell_1, \ldots, \ell_m$, $m \ge 0$, be the lines on scales $\ge n$ which are the closest to ℓ_0, and let $\theta_1, \ldots, \theta_m$ the trees with root lines $\ell_1, \ldots, \ell_m$, respectively (cf. Fig. 9 – note that by construction all lines ℓ in the subgraph T have scales $n_\ell < n$, so that if $n_{\ell_0} \ge n$ then T is necessarily a cluster). If either ℓ_0 is not on scale n or it is on scale n but exits a self-energy cluster then $\mathfrak{N}_n^*(\theta) = \mathfrak{N}_n^*(\theta_1) + \cdots + \mathfrak{N}_n^*(\theta_m)$ and the bound $\mathfrak{N}_n^*(\theta) \le E(n,k)$ follows by the inductive hypothesis. If ℓ_0 does not exit a self-energy cluster and $n_{\ell_0} = n$ then $\mathfrak{N}_n^*(\theta) = 1 + \mathfrak{N}_n^*(\theta_1) + \cdots + \mathfrak{N}_n^*(\theta_m)$, and the lines $\ell_1, \ldots, \ell_m$ enter a cluster T with $k_T = k - (k_1 + \cdots + k_m)$, where $k_1, \ldots, k_m$ are the orders of $\theta_1, \ldots, \theta_m$, respectively. If $m \ge 2$ the bound $\mathfrak{N}_n^*(\theta) \le E(n,k)$ follows once more by the inductive hypothesis. If $m = 0$ then $\mathfrak{N}_n^*(\theta) = 1$; on the other hand for ℓ_0 to be on scale $n_{\ell_0} = n$ one must have $|\omega \cdot \nu_{\ell_0}| < 2^{-n+1}\gamma$ (see (31)), which, by the Diophantine condition (29), implies $Nk \ge |\nu_{\ell_0}| > 2^{(n-1)/\tau}$, hence $E(n,k) > 1$. If $m = 1$ call ν_1 and ν_2 the momenta of the lines ℓ_0 and ℓ_1, respectively. By construction T cannot be a self-energy cluster, hence $\nu_1 \neq \nu_2$, so that, by the Diophantine condition (29),

$$2^{-n+2}\gamma \ge |\omega \cdot \nu_1| + |\omega \cdot \nu_2| \ge |\omega \cdot (\nu_1 - \nu_2)| > \frac{\gamma}{|\nu_1 - \nu_2|^\tau}\,, \tag{36}$$

because $n_{\ell_0} = n$ and $n_{\ell_1} \ge n$. Thus, one has

$$Nk_T \ge \sum_{\nu \in N(T)} |\nu_\nu| \ge |\nu_1 - \nu_2| > 2^{(n-2)/\tau}\,, \tag{37}$$

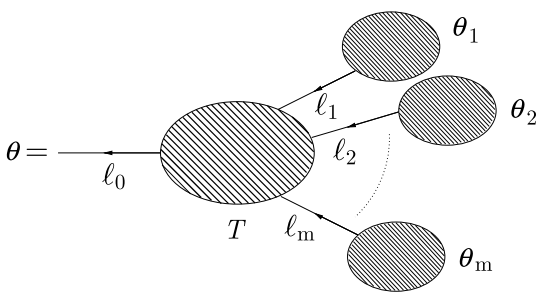

Diagrammatic Methods in Classical Perturbation Theory, Figure 9
Construction for the proof of the Siegel–Bryuno lemma

hence T must contain "many nodes". In particular, one finds also in this case $\mathfrak{N}_n^*(\theta) = 1 + \mathfrak{N}_n^*(\theta_1) \le 1 + E(n,k_1) \le 1 + E(n,k) - E(n,k_T) \le E(n,k)$, where we have used that $E(n,k_T) \ge 1$ by (37).

The argument above shows that small divisors can accumulate only by allowing self-energy clusters. That accumulation really occurs is shown by the example in Fig. 10, where a tree θ of order k containing a chain of p self-energy clusters is depicted. Assume for simplicity that $k/3$ is an integer: then if $p = k/3$ the subtree θ_1 with root line ℓ is of order $k/3$. If the line ℓ entering the rightmost self-energy cluster T_p has momentum ν, also the lines exiting the p self-energy clusters have the same momentum ν. Suppose that $|\nu| \approx Nk/3$ and $|\omega \cdot \nu| \approx \gamma/|\nu|^\tau$ (this is certainly possible for some ν). Then the value of the tree θ grows like $a_1^k(k!)^{a_2}$, for some constants a_1 and a_2: a bound of this kind prevents the convergence of the perturbation series (25).

If no self-energy clusters could occur (so that $\mathfrak{R}_n(\theta) = 0$) the Siegel–Bryuno lemma would allow us to bound in (32)

$$\prod_{n=0}^{\infty} 2^{2n\mathfrak{N}_n(\theta)} = \prod_{n=0}^{\infty} 2^{2n\mathfrak{N}_n^*(\theta)} \le \exp\left(C_1 k \sum_{n=0}^{\infty} n2^{-n/\tau}\right) \le C_2^k\,, \tag{38}$$

for suitable constants C_1 and C_2. In that case convergence of the series for $|\varepsilon| < \varepsilon_0$ would follow, with ε_0 defined as in (26) with $C_0 = \gamma^2/C_2$. However, there are self-energy clusters and they produce factorials, as the example in Fig. 10 shows, so that we have to deal with them.

Resummation

Let us come back to the Eq. (10). If we expand $g(\omega \cdot \nu)[\varepsilon\partial_\alpha f(\alpha)]_\nu^{(k)}$ in trees according to the diagrammatic rules described in Sect. "Trees and Graphical Representation", we can distinguish between contributions in which the root line exits a self-energy cluster T, that we can write as

$$g(\omega \cdot \nu) \sum_{T:\, k_T < k} \mathcal{V}_T(\omega \cdot \nu) h_\nu^{(k-k_T)}\,, \tag{39}$$

and all the other contributions, that we denote by $[\varepsilon\partial_\alpha f(\alpha)]_\nu^{(k)*}$. We can shift the contributions (39) to the left hand side of (10) and divide by $g(\omega \cdot \nu)$, so to obtain

$$D(\omega \cdot \nu) h_\nu^{(k)} - \sum_{T:\, k_T < k} \mathcal{V}_T(\omega \cdot \nu) h_\nu^{(k-k_T)} = [\varepsilon\partial_\alpha f(\omega t + h)]_\nu^{(k)*}. \tag{40}$$

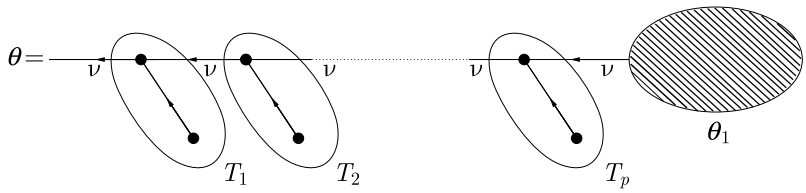

Diagrammatic Methods in Classical Perturbation Theory, Figure 10
Example of accumulation of small divisors because of the self-energy clusters

where $D(\omega\cdot\nu) = 1/g(\omega\cdot\nu) = (\omega\cdot\nu)^2$. By summing over k and setting

$$M(\omega\cdot\nu;\varepsilon) = \sum_{k=1}^{\infty}\varepsilon^k \sum_{T:\,k_T=k} \mathcal{V}_T(\omega\cdot\nu)\,, \tag{41}$$

then (40) gives

$$\begin{aligned} &\mathcal{D}(\omega\cdot\nu)h_\nu = [\varepsilon\partial_\alpha f(\omega t+h)]^*_\nu,\\ &\mathcal{D}(\omega\cdot\nu) := D(\omega\cdot\nu) - M(\omega\cdot\nu;\varepsilon)\,. \end{aligned} \tag{42}$$

The motivation for proceeding in this way is that, at the price of changing the operator D into $\mathcal{D}$, hence of changing the propagators, lines exiting self-energy clusters no longer appear. Therefore, in the tree expansion of the right hand side of the equation, we have eliminated the self-energy clusters, that is the source of the problem of accumulation of small divisors.

Unfortunately the procedure described above has a problem: $M(\omega\cdot\nu;\varepsilon)$ itself is a sum of self-energy clusters, which can still contain some other self-energy clusters on lower scales. So finding a good bound for $M(\omega\cdot\nu;\varepsilon)$ could have the same problems as for the values of the trees.

To deal with such a difficulty we modify the prescription by proceeding recursively, in the following sense. Let us start from the momenta ν which are on scale $n=0$. Since there are no self-energy clusters with exiting line on scale $n=0$, for such ν one has $M(\omega\cdot\nu;\varepsilon)=0$. Next, we consider the momenta ν which are on scale $n=1$, and we can write (42), where now all self-energy clusters T whose values contribute to $M(\omega\cdot\nu;\varepsilon)$ cannot contain any self-energy clusters, because the lines $\ell\in T$ are on scale $n_\ell=0$. Then, we consider the momenta ν which are on scale $n=2$: again all the self-energy clusters contributing to $M(\omega\cdot\nu;\varepsilon)$ do not contain any self-energy clusters, because the lines on scale $n=0,1$ cannot exit self-energy clusters by the construction of the previous step, and so on. The conclusion is that we have obtained a different expansion for $h(\omega t)$, that we call a *resummed series*,

$$h(\omega t) = \sum_{\nu\in\mathbb{Z}^d} \mathrm{e}^{i\omega\cdot\nu t}h_\nu\,, \qquad h_\nu = \sum_{k=1}^{\infty}\varepsilon^k h_\nu^{[k]}(\varepsilon)\,, \tag{43}$$

where the self-energy clusters do not appear any more in the tree expansion and the propagators must be defined recursively: the propagator g_ℓ of a line ℓ on scale $n_\ell=n$ and momentum $\nu_\ell=\nu$ is the matrix

$$g_\ell := g^{[n]}(\omega\cdot\nu;\varepsilon) = \big(D(\omega\cdot\nu) - \mathcal{M}^{[n]}(\omega\cdot\nu;\varepsilon)\big)^{-1}\,, \tag{44}$$

with

$$\mathcal{M}^{[n]}(\omega\cdot\nu;\varepsilon) := \sum_{T:\,n_T<n} \varepsilon^{k_T}\mathcal{V}_T(\omega\cdot\nu)\,, \tag{45}$$

where the value $\mathcal{V}_T(\omega\cdot\nu)$ is written in accord with (34), with all the lines $\ell'\in L(T)$ on scales $n_{\ell'}<n$ and the corresponding propagators $g_{\ell'}$ expressed in terms of matrices $\mathcal{M}^{[n_{\ell'}]}(\omega\cdot\nu_{\ell'};\varepsilon)$ as in (44). By construction, the new propagators depend on ε, so that the coefficients $h_\nu^{[k]}(\varepsilon)$ depend explicitly on ε: hence (43) is not a power series expansion.

The coefficients $h_\nu^{[k]}(\varepsilon)$ still admit a tree expansion

$$\begin{aligned} &h_\nu^{[k]}(\varepsilon) = \sum_{\theta\in\mathcal{T}^{\mathcal{R}}_{k,\nu}} \mathrm{Val}(\theta), \quad \nu\neq 0, k\geq 1\,,\\ &\mathrm{Val}(\theta) := \Big(\prod_{v\in N(\theta)} F_v\Big)\Big(\prod_{\ell\in L(\theta)} g^{[n_\ell]}(\omega\cdot\nu_\ell;\varepsilon)\Big)\,, \end{aligned} \tag{46}$$

which replaces (24). In particular $\mathcal{T}^{\mathcal{R}}_{k,\nu}$ is defined as the set of *renormalized trees* of order k and momentum ν, where "renormalized" means that the trees do not contain any self-energy clusters. Now the propagators are matrices: the contractions of the indices in (46) yields that if ℓ connects v to v' the indices of F_v and $F_{v'}$ associated with ℓ must be equal to the column and row indices of g_ℓ, respectively.

Since for any tree $\theta\in\mathcal{T}^{\mathcal{R}}_{k,\nu}$ one has $\mathfrak{N}_n(\theta)=\mathfrak{N}^*_n(\theta)$, we can bound the product of propagators according to (36) *provided the propagators on scale n can still be bounded proportionally to* 2^{2n}. This is certainly not obvious, because of the extra term $\mathcal{M}^{[n]}(\omega\cdot\nu;\varepsilon)$ appearing in (44).

It is a remarkable cancellation that $\mathcal{M}^{[n]}(x;\varepsilon)$ vanishes up to second order, that is $\mathcal{M}^{[n]}(x;\varepsilon)=O(x^2)$, so that, by

taking into account also that $\mathcal{V}_T(x) \neq 0$ requires $k_T \geq 2$, we can write, for some constant C,

$$\mathcal{M}^{[n]}(x;\varepsilon) = \varepsilon^2 x^2 \overline{\mathcal{M}}^{[n]}(x;\varepsilon)\,, \qquad \left\| \overline{\mathcal{M}}^{[} n](x;\varepsilon) \right\| \leq C\,, \tag{47}$$

where $\|\cdot\|$ denotes – say – the uniform norm. The cancellation leading to (47) can be proved as discussed in Sect. "Multiscale Analysis" – where it has been explicitly proved for $k = 2$ in the absence of resummation. Now the propagators are matrices, but one can prove (by induction on n) that $\mathcal{M}^{[n]}(x;\varepsilon) = (\mathcal{M}^{[n]}(-x;\varepsilon))^T = (\mathcal{M}^{[n]}(x;\varepsilon))^\dagger$, with T and $\dagger$ denoting transposition and adjointness, respectively, and this is enough to see that the same cancellation mechanism applies [23].

Thus (47) implies that

$$\left\| g^{[n]}(x;\varepsilon) \right\| = \left\| \left(x^2 - \varepsilon^2 x^2 \overline{\mathcal{M}}^{[n]}(x;\varepsilon) \right)^{-1} \right\| \leq \frac{2}{x^2}\,, \tag{48}$$

and the same argument as used in Sect. "Small Divisors" implies that the series in (43) for h_ν converges for $|\varepsilon| < \varepsilon_0$, where ε_0 is defined as in (26) with $C_0 = 2\gamma^2/C_2$. Moreover (47) also implies that $h(\omega t)$ is analytic in ε (notwithstanding that the expansion is not a power expansion), so that we can say a posteriori that the original power series (25) also converges.

Generalizations

The case where the function $G(u)$ in (1) does not vanish is notationally more involved, and it is explicitly discussed in Gentile [28].

Extensions of the results described above to Hamiltonian functions more general than (5) require only some notational complications, and can be found in Gentile & Mastropietro [37] for anisochronous systems and in Bartuccelli & Gentile [3] for isochronous systems. The case (5) with f a real analytic function was first discussed by using trees in Chierchia & Falcolini [12].

Lower-Dimensional Tori

A tree formalism for hyperbolic lower-dimensional tori was introduced and studied in Gallavotti [19], Gentile [26] and Gallavotti, Gentile & Mastropietro [25], for systems consisting of a set of rotators interacting with a pendulum. In that case also the stable and unstable manifolds (*whiskers*) of the tori were studied with diagrammatic techniques.

For Hamiltonian functions of the form (5) solutions of the form (43) describe quasi-periodic motions with frequency vector ω on a d-dimensional manifold. If we fix t, say $t = 0$, and keep α_0 as a parameter, we obtain a parametrization of the manifold in terms of $\alpha_0 \in \mathbb{T}^d$, hence the manifold is an invariant torus. We say that the torus is a *maximal torus*.

We can study the problem of persistence of *lower-dimensional tori* by considering unperturbed solutions $\alpha(t) = \alpha_0 + \omega t$, where the components of ω are rationally dependent. For instance we can imagine that there exist s linearly independent vectors $\widehat{\nu}_1, \ldots, \widehat{\nu}_s \in \mathbb{Z}^d$ such that $\omega \cdot \widehat{\nu}_k = 0$ for $k = 1, \ldots, s$. In that case, we say that the unperturbed torus is a *resonant* torus of order s. We can imagine performing a linear change of variables which transforms the frequency vector into a new vector, that we still denote with ω, such that $\omega = (\bar{\omega}, 0)$, where $\bar{\omega} \in \mathbb{R}^r$ and $0 \in \mathbb{R}^s$, with $r + s = d$, and $\bar{\omega}$ is irrational. This naturally suggests that we write $\alpha = (\bar{\alpha}, \beta)$, with $\bar{\alpha} \in \mathbb{T}^r$ and $\beta \in \mathbb{T}^s$. More generally, here and henceforth in this subsection for any vector $\nu \in \mathbb{R}^d$ we denote by $\bar{\nu}$ the vector in $\mathbb{R}^r$ whose components are the first r components of ν. For instance for the initial phase α_0 we write $\alpha_0 = (\bar{\alpha}_0, \beta_0)$.

In general the resonant torus is destroyed by the perturbation, and only some lower-dimensional tori persist under perturbation. We assume on $\bar{\omega}$ one of the Diophantine conditions of Sect. "Small Divisors" in $\mathbb{R}^r$, for instance $|\bar{\omega} \cdot \bar{\nu}| > \gamma/|\bar{\nu}|^\tau$ for all $\bar{\nu} \in \mathbb{Z}^r$, $\bar{\nu} \neq 0$.

To prove the existence of a maximal torus for the system with Hamiltonian function (5) we needed no condition on the perturbation f. On the contrary to prove existence of lower-dimensional tori, we need some *non-degeneracy condition*: by defining

$$f_0(\beta) = \int_{\mathbb{T}^r} \frac{\mathrm{d}\bar{\alpha}}{(2\pi)^r} f(\bar{\alpha}, \beta)\,, \tag{49}$$

if $\partial_\beta f_0(\beta_*) = 0$ for some β_* then we assume that the matrix $\partial^2_\beta f_0(\beta_*)$ is positive definite (more generally one could assume it to be non-singular, that is $\det \partial^2_\beta f_0(\beta_*) \neq 0$).

The formal analysis can be carried out as in Sect. "Trees and Graphical Representation", with the only difference being that now the compatibility conditions $[\varepsilon \partial_\alpha f(\alpha)]^{(k)}_\nu = 0$ have to be imposed for all ν such that $\bar{\nu} = 0$, because $\omega \cdot \nu = \bar{\omega} \cdot \bar{\nu}$. It turns out to be an identity only for the first r components (this is trivial for $k = 1$, whereas it requires some work for $k > 1$). For the last s components, for $k = 1$ it reads $\partial_\beta f_0(\beta_0) = 0$, hence it fixes β_0 to be a stationary point for $f_0(\beta)$, while for higher values of k it fixes the corrections of higher order of these values (to do this we need the non-degeneracy condition). Thus, we are free to choose only $\bar{\alpha}_0$ as a free parameter, since the last s components of α_0 have to be fixed.

Clusters and self-energy clusters are defined as in Sect. "Multiscale Analysis". Note that only the first r components $\bar{\nu}$ of the momenta ν intervene in the definition of the scales – again because $\omega \cdot \nu = \bar{\omega} \cdot \bar{\nu}$. In particular, in the definition of self-energy clusters, in (33) we must replace ν_v with $\bar{\nu}_v$. Thus, already to first order the value of a self-energy cluster can be non-zero: for $k_T = 1$, that is for T containing only a node v with mode label $(\bar{\nu}_v, \tilde{\nu}_v) = (0, \tilde{\nu}_v)$, the matrix $\mathcal{V}_T(x;\varepsilon)$ is of the form

$$\begin{aligned} \mathcal{V}_T(x) &= \begin{pmatrix} 0 & 0 \\ 0 & b_{\tilde{\nu}_v} \end{pmatrix}, \\ (b_{\tilde{\nu}_v})_{i,j} &= \mathrm{e}^{i\tilde{\nu}_v \cdot \beta_0} (i\tilde{\nu}_{v,i})(i\tilde{\nu}_{v,j}) f_{(0,\tilde{\nu}_v)} , \end{aligned} \tag{50}$$

with $i, j = r+1, \ldots, d$. If we sum over $\tilde{\nu}_v \in \mathbb{Z}^s$ and multiply times ε, we obtain

$$\begin{aligned} M_0 &:= \begin{pmatrix} 0 & 0 \\ 0 & \varepsilon B \end{pmatrix}, \\ B_{i,j} &= \sum_{\tilde{\nu} \in \mathbb{Z}^s} \mathrm{e}^{i\tilde{\nu} \cdot \beta_0} (i\tilde{\nu}_i)(i\tilde{\nu}_j) f_{(0,\tilde{\nu})} = \partial_{\beta_i} \partial_{\beta_j} f_0(\beta_0) . \end{aligned} \tag{51}$$

The $s \times s$ block B is non-zero: in fact, the non-degeneracy condition yields that it is invertible.

To higher orders one finds that the matrix $\mathcal{M}^{[n]}(x;\varepsilon)$, with $x = \omega \cdot \nu = \bar{\omega} \cdot \bar{\nu}$, is a self-adjoint matrix and $\mathcal{M}^{[n]}(x;\varepsilon) = (\mathcal{M}^{[n]}(-x;\varepsilon))^T$, as in the case of maximal tori. Moreover the corresponding eigenvalues $\lambda_i^{[n]}(x;\varepsilon)$ satisfy $\lambda_i^{[n]}(x;\varepsilon) = O(\varepsilon^2 x^2)$ for $i = 1, \ldots, r$ and $\lambda_i^{[n]}(x;\varepsilon) = O(\varepsilon)$ for $i = r+1, \ldots, d$; this property is not trivial because of the off-diagonal blocks (which in general do not vanish at orders $k \geq 2$), and to prove it one has to use the self-adjointness of the matrix $\mathcal{M}^{[n]}(x;\varepsilon)$. More precisely one has $\lambda_i^{[n]}(x;\varepsilon) = \varepsilon a_i + O(\varepsilon^2)$ for $i > r$, where $a_{r+1}, \ldots, a_d$ are the s eigenvalues of the matrix B in (51). From this point on the discussion proceeds in a very different way according to the sign of ε (recall that we are assuming that $a_i > 0$ for all $i > r$).

For $\varepsilon < 0$ one has $\lambda_i^{[n]}(x;\varepsilon) = a_i \varepsilon + O(\varepsilon^2) < 0$ for $i > r$, so that we can bound the last s eigenvalues of $x^2 - \mathcal{M}^{[n]}(x;\varepsilon)$ with x^2, and the first r with $x^2/2$ by the same argument as in Sect. "Resummation". Hence we obtain easily the convergence of the series (43); of course, analyticity at the origin is prevented because of the condition $\varepsilon < 0$. We say in that case that the lower-dimensional tori are *hyperbolic*. We refer to Gallavotti & Gentile [22] and Gallavotti et al. [23] for details.

The case of *elliptic* lower-dimensional tori – that is $\varepsilon > 0$ when $a_i > 0$ for all $i > r$ – is more difficult. Essentially the idea is as follows (we only sketch the strategy: the details can be found in Gentile & Gallavotti [36]). One has to define the scales recursively, by using a variant, first introduced in Gentile [27], of the resummation technique described in Sect. "Resummation". We say that ν is on scale 0 if $|\bar{\omega} \cdot \bar{\nu}| \geq \gamma$ and on scale $[\geq 1]$ otherwise: for ν on scale 0 we write (42) with $M = M_0$, as given in (51). This defines the propagators of the lines ℓ on scale $n = 0$ as

$$g_\ell = g^{[0]}(\omega \cdot \nu_\ell) = \left((\bar{\omega} \cdot \bar{\nu}_\ell)^2 - M_0\right)^{-1} . \tag{52}$$

Denote by λ_i the eigenvalues of M_0: given ν on scale $[\geq 1]$ we say that ν is on scale 1 if $2^{-1}\gamma \leq \min_{i=1,\ldots,d} \sqrt{|(\bar{\omega} \cdot \bar{\nu})^2 - \lambda_i|}$, and on scale $[\geq 2]$ if $\min_{i=1,\ldots,d} \sqrt{|(\bar{\omega} \cdot \bar{\nu})^2 - \lambda_i|} < 2^{-1}\gamma$. For ν on scale 1 we write (42) with M replaced by $\mathcal{M}^{[0]}(\omega \cdot \bar{\nu};\varepsilon)$, which is given by M_0 plus the sum of the values of all self-energy clusters T on scale $n_T = 0$. Then the propagators of the lines ℓ on scale $n_\ell = 1$ is defined as

$$g_\ell = g^{[1]}(\omega \cdot \nu_\ell) = \left((\bar{\omega} \cdot \bar{\nu}_\ell)^2 - \mathcal{M}^{[0]}(\bar{\omega} \cdot \bar{\nu}_\ell;\varepsilon)\right)^{-1} . \tag{53}$$

Call $\lambda_i^{[n]}(x;\varepsilon)$ the eigenvalues of $\mathcal{M}^{[n]}(x;\varepsilon)$: given ν on scale $[\geq 2]$ we say that ν is on scale 2 if $2^{-2}\gamma \leq \min_{i=1,\ldots,d} \sqrt{|(\bar{\omega} \cdot \bar{\nu})^2 - \lambda_i^{[0]}(\bar{\omega} \cdot \bar{\nu};\varepsilon)|}$, and on scale $[\geq 3]$ if $\min_{i=1,\ldots,d} \sqrt{|(\bar{\omega} \cdot \bar{\nu})^2 - \lambda_i^{[0]}(\bar{\omega} \cdot \bar{\nu};\varepsilon)|} < 2^{-2}\gamma$. For ν on scale 2 we write (42) with M replaced by $\mathcal{M}^{[1]}(\bar{\omega} \cdot \bar{\nu};\varepsilon)$, which is given by $\mathcal{M}^{[0]}(\bar{\omega} \cdot \bar{\nu};\varepsilon)$ plus the sum of the values of all self-energy clusters T on scale $n_T = 1$. Thus, the propagators of the lines ℓ on scale $n_\ell = 2$ will be defined as

$$g_\ell = g^{[2]}(\omega \cdot \nu_\ell) = \left((\bar{\omega} \cdot \bar{\nu}_\ell)^2 - \mathcal{M}^{[1]}(\bar{\omega} \cdot \bar{\nu}_\ell;\varepsilon)\right)^{-1} , \tag{54}$$

and so on. The propagators are self-adjoint matrices, hence their norms can be bounded through the corresponding eigenvalues. In order to proceed as in Sects. "Multiscale Analysis" and "Resummation" we need some Diophantine conditions on these eigenvalues. We can assume for some $\tau' > \tau$

$$\left| |\bar{\omega} \cdot \bar{\nu}| - \sqrt{|\lambda_i^{[n]}(\bar{\omega} \cdot \bar{\nu};\varepsilon)|} \right| > \frac{\gamma}{|\bar{\nu}|^{\tau'}} \qquad \forall \bar{\nu} \neq 0 , \tag{55}$$

for all $i = 1, \ldots, d$ and $n \geq 0$. These are known as the *first Melnikov conditions*.

Unfortunately, things do not proceed so plainly. In order to prove a bound like (35), possibly with a different τ' replacing τ, we need to compare the propagators of the

lines entering and exiting clusters T which are not self-energy clusters. This requires replacing (36) with

$$2^{-n+2}\gamma \geq \left|\bar\omega\cdot(\bar\nu_1-\bar\nu_2) \pm \sqrt{|\lambda_i^{[n]}(\bar\omega\cdot\bar\nu_1;\varepsilon)|} \pm \sqrt{|\lambda_j^{[n]}(\bar\omega\cdot\bar\nu_2;\varepsilon)|}\right| > \frac{\gamma}{|\bar\nu_1-\bar\nu_2|^{\tau'}}\,, \tag{56}$$

for all $i, j = 1, \ldots, d$ and choices of the signs $\pm$, and hence introduces further Diophantine conditions, known as the *second Melnikov conditions*.

The conditions in (56) turn out to be too many, because for all $n \geq 0$ and all $\bar\nu \in \mathbb{Z}^r$ such that $\bar\nu = \bar\nu_1 - \bar\nu_2$ there are infinitely many conditions to be considered, one per pair $(\bar\nu_1, \bar\nu_2)$. However we can impose both the conditions (55) and (56) not for the eigenvalues $\lambda_i^{[n]}(\bar\omega\cdot\bar\nu;\varepsilon)$, but for some quantities $\underline{\lambda}_i^{[n]}(\varepsilon)$ independent of $\bar\nu$ and then use the smoothness of the eigenvalues in x to control $(\bar\omega\cdot\bar\nu)^2 - \lambda_i^{[n]}(\bar\omega\cdot\bar\nu;\varepsilon)$ in terms of $(\bar\omega\cdot\bar\nu)^2 - \underline{\lambda}_i^{[n]}(\varepsilon)$. Eventually, beside the Diophantine condition on $\bar\omega$, we have to impose the Melnikov conditions

$$\begin{aligned} &\left||\bar\omega\cdot\bar\nu| - \sqrt{|\underline{\lambda}_i^{[n]}(\varepsilon)|}\right| > \frac{\gamma}{|\bar\nu|^{\tau'}}\,,\\ &\left||\bar\omega\cdot\bar\nu| \pm \sqrt{|\underline{\lambda}_i^{[n]}(\varepsilon)|} \pm \sqrt{|\underline{\lambda}_j^{[n]}(\varepsilon)|}\right| > \frac{\gamma}{|\bar\nu|^{\tau'}}\,, \end{aligned} \tag{57}$$

for all $\bar\nu \neq 0$ and all $n \geq 0$. Each condition in (57) leads us to eliminate a small interval of values of ε. For the values of ε which are left define $h(\omega t)$ according to (43) and (46), with the new definition of the propagators. If ε is small enough, say $|\varepsilon| < \varepsilon_0$, then the series (43) converges. Denote by $\mathfrak{E} \subset [0, \varepsilon_0]$ the set of values of ε for which the conditions (57) are satisfied. One can prove that $\mathfrak{E}$ is a *Cantor set*, that is a perfect, nowhere dense set. Moreover $\mathfrak{E}$ has large relative Lebesgue measure, in the sense that

$$\lim_{\varepsilon\to 0} \frac{\mathrm{meas}(\mathfrak{E}\cap[0,\varepsilon])}{\varepsilon} = 1\,, \tag{58}$$

provided τ' in (57) is large enough with respect to τ. The property (58) yields that, notwithstanding that we are eliminating infinitely many intervals, the measure of the union of all these intervals is small.

If $a_i < 0$ for all $i > r$ we reason in the same way, simply exchanging the role of positive and negative ε. On the contrary if $a_i = 0$ for some $i > r$, the problem becomes much more difficult. For instance if $s = 1$ and $a_{r+1} = 0$, then in general perturbation theory in ε is not possible, not even at a formal level. However, under some conditions, one can still construct fractional series in ε, and prove that the series can be resummed [24].

Other Ordinary Differential Equations

The formalism described above extends to other models, such as skew-product systems [29] and systems with strong damping in the presence of a quasi-periodic forcing term [32].

As an example of skew-product system one can consider the linear differential equation $\dot x = \big(\lambda A + \varepsilon f(\omega t)\big)\, x$ on $\mathrm{SL}(2,\mathbb{R})$, where $\lambda \in \mathbb{R}$, ε is a small real parameter, $\omega \in \mathbb{R}^n$ is an irrational vector, and $A, f \in \mathfrak{sl}(2,\mathbb{R})$, with A is a constant matrix and f an analytic function periodic in its arguments. Trees for skew-products were considered by Iserles and Nørsett [44], but they used expansions in time, hence not suited for the study of global properties, such as quasi-periodicity.

Quasi-periodically forced one-dimensional systems with strong damping are described by the ordinary differential equations $\ddot x + \gamma\dot x + g(x) = f(\omega t)$, where $x \in \mathbb{R}$, $\varepsilon = 1/\gamma$ is a small real parameter, $\omega \in \mathbb{R}^n$ is irrational, and f,g are analytic functions (g is the "force"), with f periodic in its arguments.

We refer to the bibliography for details and results on the existence of quasi-periodic solutions.

Bryuno Vectors

The diagrammatic methods can be used to prove the any unperturbed maximal torus with frequency vector which is a Bryuno vector persists under perturbation for ε small enough [31]. One could speculate whether the Bryuno condition (30) is optimal. In general the problem is open. However, in the case of the standard map – see Sect. "The Standard Map", – one can prove [6,15] that, by considering the radius of convergence ε_0 of the perturbation series as a function of ω, say $\varepsilon_0 = \rho_0(\omega)$, then there exists a universal constant C such that

$$|\log\rho_0(\omega) + 2\mathcal{B}(\omega)| \leq C\,. \tag{59}$$

In particular this yields that the invariant curve with rotation number ω persists under perturbation if and only if ω is a Bryuno number. The proof of (59) requires the study of a more refined cancellation than that discussed in Sect. "Resummation". We refer to Berretti & Gentile [6] and Gentile [28] for details.

Extensions to Bryuno vectors for lower-dimensional tori can also be found in Gentile [31]. For the models considered in Sect. "Other Ordinary Differential Equations" we refer to Gentile [29] and Gentile et al. [33].

Partial Differential Equations

Existence of quasi-periodic solutions in systems described by one-dimensional nonlinear partial differential equations (finite-dimensional tori in infinite-dimensional systems) was first studied by Kuksin [48], Craig and Wayne [14] and Bourgain [9]. In these systems, even the case of periodic solutions yields small divisors, and hence requires a multiscale analysis. The study of persistence of periodic solutions for nonlinear Schrödinger equations and nonlinear wave equations, with the techniques discussed here, can be found in Gentile & Mastropietro [39], Gentile et al. [40], and Gentile & Procesi [41].

The models are still described by (1), with $G(u) = 0$, but now D is given by $D = \partial_t^2 - \Delta + \mu$ in the case of the wave equation and by $D = i\partial_t - \Delta + \mu$ in the case of the Schrödinger equation, where Δ is the Laplacian and $\mu \in \mathbb{R}$. In dimension 1, one has $\Delta = \partial_x^2$. If we look for periodic solutions with frequency ω it can be convenient to pass to Fourier space, where the operator D acts as

$$D\colon \mathrm{e}^{i\omega nt + imx} \to \left(-\omega^2 n^2 + m^2 + \mu\right) \mathrm{e}^{i\omega nt + imx}\,, \quad (60)$$

for the wave equation; a similar expression holds for the Schrödinger equation. Therefore the kernel of D can be arbitrarily close to zero for n and m large enough.

Then one can consider, say, (1) for $x \in [0, \pi]$ and $F(u) = u^3$, with Dirichlet boundary conditions $u(0) = u(\pi) = 0$, and study the existence of periodic solutions with frequency ω close to some of the unperturbed frequencies. We refer to the cited bibliography for results and proofs.

Conclusions and Future Directions

The diagrammatic techniques described above have been applied also in cases where no small divisors appear; cf. Berretti & Gentile [4] and Gentile et al. [34]. Of course, such problems are much easier from a technical point of view, and can be considered as propaedeutic examples to become familiar with the tree formalism. Also the study of lower-dimensional tori becomes easy for $r = 1$ (periodic solutions): in that case one has $|\bar{\omega} \cdot \bar{\nu}| \geq |\bar{\omega}|$ for all $\bar{\nu} \neq 0$, so that the product of the propagators is bounded by $|\bar{\omega}|^{-2k}$, and one can proceed as in Sect. "Small Divisors" to obtain analyticity of the solutions.

In the case of hyperbolic lower-dimensional tori, if ω is a two-dimensional Diophantine vector of constant type (that is, with $\tau = 1$) the conjugation function h can be proved to be Borel summable [13]. Analogous considerations hold for the one-dimensional systems in the presence of friction and of a quasiperiodic forcing term described in Sect. "Other Ordinary Differential Equations"; in that case one has Borel summability also for one-dimensional ω, that is for periodic forcing [33]. It would be interesting to investigate whether Borel summability could be obtained for higher values of τ.

Recently existence of finite-dimensional tori in the nonlinear Schrödinger equation in higher dimensions was proved by Bourgain [10]. It would be interesting to investigate how far the diagrammatic techniques extend to deal with such higher dimensional generalizations. The main problem is that (the analogues of) the second Melnikov conditions in (57) cannot be imposed.

In certain cases the tree formalism was extended to non-analytic systems, such as some quasi-integrable systems of the form (5) with f in a class of C^p functions for some finite p [7,8]. However, up to exceptional cases, the method described here seems to be intrinsically suited in cases in which the vector fields are analytic. The reason is that in order to exploit the expansion (3), we need that F be infinitely many times differentiable and we need a bound on the derivatives. It is a remarkable property that the perturbation series can be given a meaning also in cases where the solutions are not analytic in ε.

An advantage of the diagrammatic method is that it allows rather detailed information about the solutions, hence it could be more convenient than other techniques to study problems where the underlying structure is not known or too poor to exploit general abstract arguments.

Another advantage is the following. If one is interested not only in proving the existence of the solutions, but also in explicitly constructing them with any prefixed precision, this requires performing analytical or numerical computations with arbitrarily high accuracy. Then high perturbation orders have to be reached, and the easiest and most direct way to proceed is just through perturbation theory: so the approach illustrated here allows a unified treatment for both theoretical investigations and computational ones.

The resummation technique described in Sect. "Resummation" can also be used for computational purposes. With respect to the naive power series expansion it can reduce the computation time required to approximate the solution within a prefixed precision. It can also provide accurate information on the analyticity properties of the solution. For instance, for the Kepler equation, Levi-Civita at the beginning of the last century described a resummation rule (see [49]), which gives immediately the radius of convergence of the perturbation series. Of course, in the case of small divisor problems, everything becomes much more complicated.

Bibliography

Primary Literature

1. Arnold VI (1963) Proof of a theorem of A. N. Kolmogorov on the preservation of conditionally periodic motions under a small perturbation of the Hamiltonian (Russian). Uspehi Mat Nauk 18(5):13–40
2. Arnold VI, Kozlov VV, Neĭshtadt AI (1988) Dynamical systems III, Encyclopaedia of Mathematical Sciences, vol 3. Springer, Berlin
3. Bartuccelli MV, Gentile G (2002) Lindstedt series for perturbations of isochronous systems: a review of the general theory. Rev Math Phys 14(2):121–171
4. Berretti A, Gentile G (1999) Scaling properties for the radius of convergence of a Lindstedt series: the standard map. J Math Pures Appl 78(2):159–176
5. Berretti A, Gentile G (2000) Scaling properties for the radius of convergence of a Lindstedt series: generalized standard maps. J Math Pures Appl 79(7):691–713
6. Berretti A, Gentile G (2001) Bryuno function and the standard map. Comm Math Phys 220(3):623–656
7. Bonetto F, Gallavotti G, Gentile G, Mastropietro V (1998) Quasi linear flows on tori: regularity of their linearization. Comm Math Phys 192(3):707–736
8. Bonetto F, Gallavotti G, Gentile G, Mastropietro V (1998) Lindstedt series, ultraviolet divergences and Moser's theorem. Ann Scuola Norm Sup Pisa Cl Sci 26(3):545–593
9. Bourgain J (1998) Quasi-periodic solutions of Hamiltonian perturbations of 2D linear Schrödinger equations. Ann of Math 148(2):363–439
10. Bourgain J (2005) Green's function estimates for lattice Schrödinger operators and applications. Annals of Mathematics Studies 158. Princeton University Press, Princeton
11. Bricmont J, Gawędzki K, Kupiainen A (1999) KAM theorem and quantum field theory. Comm Math Phys 201(3):699–727
12. Chierchia L, Falcolini C (1994) A direct proof of a theorem by Kolmogorov in Hamiltonian systems. Ann Scuola Norm Sup Pisa Cl Sci 21(4):541–593
13. Costin O, Gallavotti G, Gentile G, Giuliani A (2007) Borel summability and Lindstedt series. Comm Math Phys 269(1):175–193
14. Craig W, Wayne CE (1993) Newton's method and periodic solutions of nonlinear wave equations. Comm Pure Appl Math 46(11):1409–1498
15. Davie AM (1994) The critical function for the semistandard map. Nonlinearity 7(1):219–229
16. Eliasson LH (1996) Absolutely convergent series expansions for quasi periodic motions. Math Phys Electron J 2(4):1–33
17. Gallavotti G (1983) The elements of mechanics. Texts and Monographs in Physics. Springer, New York
18. Gallavotti G (1994) Twistless KAM tori. Comm Math Phys 164(1):145–156
19. Gallavotti G (1994) Twistless KAM tori, quasi flat homoclinic intersections, and other cancellations in the perturbation series of certain completely integrable Hamiltonian systems. A review. Rev Math Phys 6(3):343–411
20. Gallavotti G (2001) Renormalization group in statistical mechanics and mechanics: gauge symmetries and vanishing beta functions, Renormalization group theory in the new millennium III. Phys Rep 352(4–6):251–272
21. Gallavotti G (2006) Classical mechanics. In: Françoise JP, Naber GL, Tsun TS (eds) Encyclopedia of Mathematical Physics, vol 1. Elsevier, Oxford
22. Gallavotti G, Gentile G (2002) Hyperbolic low-dimensional invariant tori and summations of divergent series. Comm Math Phys 227(3):421–460
23. Gallavotti G, Bonetto F, Gentile G (2004) Aspects of ergodic, qualitative and statistical theory of motion. Texts and Monographs in Physics. Springer, Berlin
24. Gallavotti G, Gentile G, Giuliani A (2006) Fractional Lindstedt series. J Math Phys 47(1):1–33
25. Gallavotti G, Gentile G, Mastropietro V (1999) A field theory approach to Lindstedt series for hyperbolic tori in three time scales problems. J Math Phys 40(12):6430–6472
26. Gentile G (1995) Whiskered tori with prefixed frequencies and Lyapunov spectrum. Dyn Stab Syst 10(3):269–308
27. Gentile G (2003) Quasi-periodic solutions for two-level systems. Comm Math Phys 242(1-2):221–250
28. Gentile G (2006) Brjuno numbers and dynamical systems. Frontiers in number theory, physics, and geometry. Springer, Berlin
29. Gentile G (2006) Diagrammatic techniques in perturbation theory. In: Françoise JP, Naber GL, Tsun TS (eds) Encyclopedia of Mathematical Physics, vol 2. Elsevier, Oxford
30. Gentile G (2006) Resummation of perturbation series and reducibility for Bryuno skew-product flows. J Stat Phys 125(2):321–361
31. Gentile G (2007) Degenerate lower-dimensional tori under the Bryuno condition. Ergod Theory Dyn Syst 27(2):427–457
32. Gentile G, Bartuccelli MV, Deane JHB (2005) Summation of divergent series and Borel summability for strongly dissipative differential equations with periodic or quasiperiodic forcing terms. J Math Phys 46(6):1–21
33. Gentile G, Bartuccelli MV, Deane JHB (2006) Quasiperiodic attractors, Borel summability and the Bryuno condition for strongly dissipative systems. J Math Phys 47(7):1–10
34. Gentile G, Bartuccelli MV, Deane JHB (2007) Bifurcation curves of subharmonic solutions and Melnikov theory under degeneracies. Rev Math Phys 19(3):307–348
35. Gentile G, Cortez DA, Barata JCA (2005) Stability for quasi-periodically perturbed Hill's equations. Comm Math Phys 260(2):403–443
36. Gentile G, Gallavotti G (2005) Degenerate elliptic tori. Comm Math Phys 257(2):319–362
37. Gentile G, Mastropietro V (1996) Methods for the analysis of the Lindstedt series for KAM tori and renormalizability in classical mechanics. A review with some applications. Rev Math Phys 8(3):393–444
38. Gentile G, Mastropietro V (2001) Renormalization group for one-dimensional fermions. A review on mathematical results. Renormalization group theory in the new millennium III. Phys Rep 352(4–6):273–437
39. Gentile G, Mastropietro V (2004) Construction of periodic solutions of nonlinear wave equations with Dirichlet boundary conditions by the Lindstedt series method. J Math Pures Appl 83(8):1019–1065
40. Gentile G, Mastropietro V, Procesi M (2005) Periodic solutions for completely resonant nonlinear wave equations with Dirichlet boundary conditions. Comm Math Phys 256(2):437–490
41. Gentile G, Procesi M (2006) Conservation of resonant periodic solutions for the one-dimensional nonlinear Schrödinger equation. Comm Math Phys 262(3):533–553

42. Gentile G, van Erp TS (2005) Breakdown of Lindstedt expansion for chaotic maps. J Math Phys 46(10):1–20
43. Harary F (1969) Graph theory. Addison-Wesley, Reading
44. Iserles A, Nørsett SP (1999) On the solution of linear differential equations in Lie groups. Royal Soc Lond Philos Trans Ser A Math Phys Eng Sci 357(1754):983–1019
45. Khanin K, Lopes Dias J, Marklof J (2007) Multidimensional continued fractions, dynamical renormalization and KAM theory. Comm Math Phys 270(1):197–231
46. Koch H (1999) A renormalization group for Hamiltonians, with applications to KAM tori. Ergod Theory Dyn Syst 19(2):475–521
47. Kolmogorov AN (1954) On conservation of conditionally periodic motions for a small change in Hamilton's function (Russian). Dokl Akad Nauk SSSR 98:527–530
48. Kuksin SB (1993) Nearly integrable infinite-dimensional Hamiltonian systems. Lecture Notes in Mathematics, vol 1556. Springer, Berlin
49. Levi-Civita T (1954) Opere matematiche. Memorie e note. Zanichelli, Bologna
50. MacKay RS (1993) Renormalisation in area-preserving maps. Advanced Series in Nonlinear Dynamics, 6. World Scientific Publishing, River Edge
51. Moser J (1962) On invariant curves of area-preserving mappings of an annulus. Nachr Akad Wiss Göttingen Math-Phys Kl. II 1962:1–20
52. Moser J (1967) Convergent series expansions for quasi-periodic motions. Math Ann 169:136–176
53. Poincaré H (1892–1899) Les méthodes nouvelles de la mécanique céleste. Gauthier-Villars, Paris
54. Pöschel J (1982) Integrability of Hamiltonian systems on Cantor sets. Comm Pure Appl Math 35(5):653–696
55. Wintner A (1941) The analytic foundations of celestial mechanics. Princeton University Press, Princeton

Books and Reviews

Berretti A, Gentile G (2001) Renormalization group and field theoretic techniques for the analysis of the Lindstedt series. Regul Chaotic Dyn 6(4):389–420

Gentile G (1999) Diagrammatic techniques in perturbations theory, and applications. Symmetry and perturbation theory. World Science, River Edge

Differential Games

Marc Quincampoix
Université de Bretagne Occidentale, Brest, France

Article Outline

Glossary
Definition of the Subject
Introduction
Qualitative and Quantitative Differential Games
Existence of a Value for Zero Sum Differential Games
Nonantagonist Differential Games
Miscellaneous
Future Directions
Bibliography

Glossary

Dynamics This is the law which governs the evolution of the system: for differential games it is a differential equation.

Strategies This is the way a player chooses his control as a function of the state of the system and of the action of his opponents.

Information This is the set of parameters known by the player in order to build his strategy.

Definition of the Subject

Differential games is a mathematical theory which is concerned with problems of conflicts modelized as game problems in which the state of the players depends on time in a continuous way. The positions of the players are the solution to differential equations. Differential games can be described from two different points of view, depending mainly on the field of application. Firstly, they can be considered as games where time is continuous. This aspect is often considered for applications in economics or management sciences. Secondly, they also can be viewed as control problems with several controllers having different objectives. In this way, differential games are a part of control theory with conflicts between the players. The second aspect often concerns classical applications of control theories: engineering sciences.

The importance of the subject was emphasized by J. von Neuman in 1946 in his pioneer book "Theory of Games and Economic Behaviour" [35]: *We repeat most emphatically that our theory is thoroughly static. A dynamic theory would unquestionably be more complete and therefore preferable*. But at that time, static aspects of game theory were mostly treated. The true birth of the domain was pursuit differential games (motivated by military applications during the "Cold War") developed in the 1950s concurrently by R. Isaacs in the USA and L. Pontryagin in the Soviet Union. Now differential games have a wide range of applications from economics to engineering sciences but also more recently in biology and behavioral ecology. The present article is focused on two-player zero sum and antagonist differential games.

Introduction

The best-known example of a differential game is the pursuit game "Lion and Man" where in an infinite plane a Lion wants to catch a Man. This example is elementary enough

for describing clearly the problematic and we will use this example as an illustration throughout the article. The position of the Man at time t is $y(t)$ while the Lion's position is $z(t)$. At any time the Man can choose his velocity $y'(t)$ in any direction but with a maximum intensity, while the Lion can choose his velocity $z'(t)$ with an intensity less than or equal to L (here $L > M$). The objective of the Lion is to catch the Man as soon as possible, the aim of the Man is to escape the Lion as long as possible. In this very intuitive game, we wish to introduce the important elements needed for defining a differential game: *the dynamics, the actions of the players, the objectives, the rules of the game (the strategies)*. The *dynamics* is then

$$y'(t) = u(t)\,, \quad u(t) \in U\,, \quad z'(t) = v(t) \in V\,; \quad t \geq 0 \tag{1}$$

where $u(t)$ is the *action* chosen by the first player (the Man), $u(t)$ belongs to a set $U = \{u \mid \|u\| \leq M\}$, similarly $v(t)$ belongs to the set $V = \{v \mid \|v\| \leq L\}$. The *objectives* of the two players are of antagonist nature: The Lion wants to minimize the time before capture while the Man wants to maximize it. The rule of the game is the way the two players choose their actions: here we can assume that they choose their instantaneous velocity as a function of the current positions

$$u(t) = \bar{u}(x(t), y(t))\,, \quad v(t) = \bar{v}(x(t), y(t))\,. \tag{2}$$

Clearly the time of capture is T in a function of the strategies $\bar{u}$ and $\bar{v}$; the Lion wants to minimize this function, the Man wants to maximize it. Solving the game means finding the *optimal value* namely the time of capture which is minimum over all strategies $\bar{v}$ and maximum over all strategies $\bar{u}$. In the elementary example of "Lion and Man," one can easily check that

$$\bar{v}(y, z) = L\frac{y - z}{\|y - z\|}\,,$$

$$\bar{u}(y, z) = -M\frac{y - z}{\|y - z\|}\,,$$

$$T = \frac{\|y(0) - z(0)\|}{L - M}\,,$$

which is a rather intuitive solution: the Lion runs in a direction towards the Man at his maximal possible speed L, while the Man runs in the opposite direction with his maximal speed M. We leave this illustrative example which has allowed us to underline the concepts of differential games.

The present article is restricted to the two player case. A state variable x is governed by a differential equation

$$x'(t) = f(x(t), u(t), v(t))\,. \tag{3}$$

The players act on the state variable x by choosing two controls u and v that are functions of the time variable t taking their values in spaces U and V. (Note that for the Lion and the Man $x = (y, z)$ and each player controls a part of the state: the game is said to be *separated*). For making the notations clearer, the first player which chooses u is called Ursula and the second player which chooses v is called Victor.

A classification of differential games according to the level of conflict between the players is very relevant. The case where the two players have opposite objectives concerns *antagonist games* or *zero-sum games*, and in the case of a nonopposite objective these are *cooperative or nonantagonist* differential games.

The objectives of the game can be quantitative – players want to minimize or maximize a *payoff* (a function depending on the initial state and their action, for instance the time of capture in the Lion and Man case) – or qualitative for instance one player wants the state to stay in a subset of the state space while the other player wants the state to reach a given target. In Isaacs' terminology, these games are *games of degree* or *games of kind*. The definition of the rules of the games (or strategies) is one of the most difficult topics in differential games. This problem does not exist neither static games nor for elementary pursuit games like the Lion and Man game. To illustrate this fact, suppose that we have to solve the game with a strategy of *feedback form*, i. e. when controls depend on the current state (as in Eq. (2), then we have to solve the differential equation

$$x'(t) = f(x(t), \bar{u}(x(t)), \bar{v}(x(t))$$

which has neither uniqueness nor existence of solutions, even if Eq. (3) enjoys good properties of existence and uniqueness of solutions. This point is crucial and is of pure mathematical nature. A way to overcome this difficulty could be to impose regularity properties on the feedback strategies. In the – young – history of differential games, this was the "time of paradoxes" in the 1960s and 1970s, where specific regularity of feedback was adequate for a class of problem but led to "paradoxes" (for quantitative games nonexistence of the value, for qualitative games existence of an initial position starting from which players are not able either to win or to lose) in slightly modified problems. This was solved in the 1970s by enlarging the class of feedback strategies allowing the choice of the control to depend not only on the current state space but also on passed values of the control of the opponents (this is the so-called nonanticipative strategies introduced by Varaiya, Roxin, Elliot, and Kalton [16,31,33,34], another class of strategies was introduced by Krasovski [21]).

For differential games, another important feature is to express the fact that the two players act simultaneously on the system. One way to translate this fact in a mathematical way is to prove that the order of the actions of the player do not modify the final result. Section "Qualitative and Quantitative Differential Games" is concerned with this question. For instance, in the case of quantitative games like the Lion and Man, by interchanging the order of operations "minimum" and "maximum" we must obtain the same result. This is the problem of existence of the value which is of first importance for differential games. We discuss this question in Sect. "Existence of a Value for Zero Sum Differential Games". The fifth section is devoted to cooperative quantitative games. Some other aspects or applications of differential games are evoked in Sect. "Miscellaneous". In the last section, we give short descriptions of present domains of research such as problems of information or impulsive differential games.

Qualitative and Quantitative Differential Games

Throughout this article, we will make suppositions such that as soon as the initial position is fixed and the controls u and v are given, we have the existence of a unique associated trajectory defined for every time (this could be easily obtained with assumptions on the function f). We also denote X the state space to which the state variable x belongs.

A nonanticipative strategy for the first player associates to any action of the second player (to any control v) a control u of the first player such that *at any time* v depends only on past values of v. It could be shown that every regular feedback strategy is such a nonanticipative strategy. So a nonanticipative strategy of the first player is a map α from $\mathcal{V}$ (the set of measurable control $v[0,+\infty] \mapsto V$) to $\mathcal{U}$ (the set of measurable controls u), which has furthermore the nonanticipative property. Similarly, one can define a strategy of the second player as a nonanticipative function β from $\mathcal{U}$ to $\mathcal{V}$.

We will now present a rather general description of a target game viewed as a game of kind (qualitative game) or as a game of degree. For this, we consider a given set in the state space – called the target – that Victor wants the state of the system to reach while Ursula wants to avoid the target. For the Lion and Man pursuit game, the target can be considered as the set of (y, z) such that $y = z$.

Qualitative Target Games

The problems consist of finding the victory domains, i. e. the set of initial positions such that one player can win. This leads to the precise following definition.

Victor Victory domain W_{V} is the set of initial position x_0 (not in C) such that he can find a strategy β such that there exists a time T such that whatever is the control u chosen by Ursula, the target is reached before the time T (cf. [1]).

In fact, to make this definition mathematically correct we have to add a small number ε and to say that the trajectory reaches a ε neighborhood of C:

$$W_{\mathrm{V}} := \{x_0 \notin C,\ \exists\beta, \exists T, \varepsilon > 0,\ \text{such that}$$
$$\forall u \in \mathcal{U}, \exists\tau > 0,\ \mathrm{dist}(x[x_0, u, \beta(u)](\tau), C) \le \varepsilon\}\,.$$

In the above formula dist means the distance and $t \mapsto x[x_0, u, \beta(u)](t)$ denotes the trajectory associated with the control u and the strategy β.

In a parallel way, the victory domain W_{U} of Ursula is the set of initial positions x_0 for which Ursula can find a strategy α such that whatever is the control played by Victor, the associate trajectory never reaches the target C.

$$W_{\mathrm{U}} := \{x_0 \notin C,\ \exists\alpha,\ \text{such that}$$
$$\forall v \in \mathcal{V},\ x[x_0, \alpha(v), v](t) \notin C,\ \forall t \ge 0\}\,.$$

In Fig. 1, the reader can see a figure corresponding to a specific differential game.

Before going further, it could be surprising to the reader that the game is not "symmetric" due to the presence of $\varepsilon > 0$. This is a mathematical problem that exceeds the scope of the present paper, for a better understanding the reader can imagine that $\varepsilon = 0$ and see [10] for deeper analysis.

The main problem of such a qualitative game is the following alternative problem. Roughly speaking, if one player does not win, the other player must win. Of course this appears to be the minimal requirement so that the modelization of the problem is correctly formulated. Nevertheless, the alternative is not an easy problem; for instance, it is not obvious that the intuitive notion of feedback strategies is not suitable for giving a positive answer to this question. The alternative can be expressed by the fact that the sets CW_{U} and W_{V} form a partition of the whole space X:

$$X = C \cup W_{\mathrm{U}} \cap W_{\mathrm{V}}, \quad \text{and} \quad W_{\mathrm{U}} \cap W_{\mathrm{V}} = \emptyset\,.$$

The above alternative is valid under several technical conditions (for instance C is an open set) we will not describe and under the following crucial condition – called *Isaacs' condition*

$$\min_{v\in V}\max_{u\in U}\langle f(x,u,v), p\rangle = \max_{u\in U}\min_{v\in V}\langle f(x,u,v), p\rangle \quad (4)$$

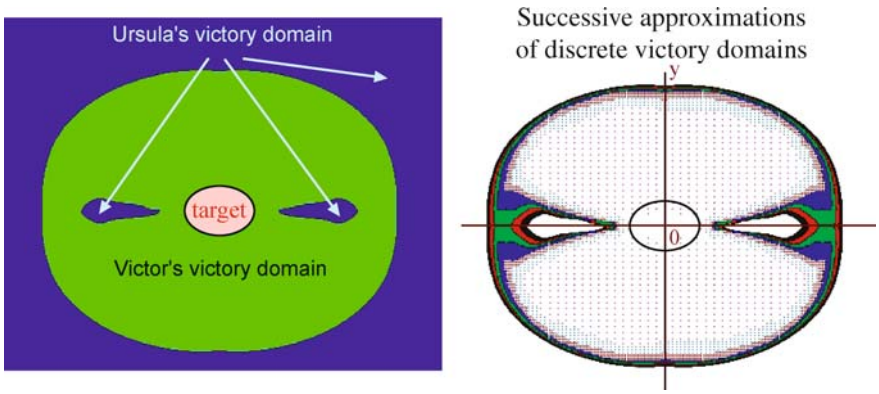

Differential Games, Figure 1
On the left: Ursula and Victor's victory domains. On the right, successive approximations of Victor's victory domain. $\binom{x'(t)}{y'(t)} = w_p\binom{-1}{0} + \frac{w_p}{0.8}\binom{-y(t)}{x(t)}v(t) + w_e \min\left(2\sqrt{x^2+y^2}, 1\right)\binom{\cos(u(t))}{\sin(u(t))}$ **where** $u \in [-\pi, \pi]$ **and** $v \in [-1, 1]$**. Numerical values for the pursuer and evader's velocities are** $w_p = 1$ **and** $w_e = 1.1$

for any direction p. This could be understood as the existence of a saddle point for the *static game* with payoff $\langle f(x, u, v), p\rangle$ ($\langle\cdot,\cdot\rangle$ is the scalar product). The expression of Eq. (4) above is called the *Hamiltonian* of the game; it is a function $(x, p) \mapsto H(x, p)$.

The alternative theorem was first obtained by Krassovskii and Subbotin [21] for a slightly different notion of strategy, and by Cardaliaguet [10] for a nonanticipative strategy. It is worth pointing out that even for the elementary Lion and Man game played in a circular "arena" (instead of the whole plane) the problem is not obvious [19]. For more complex forms of "arenas" and for general differential games with restricted space domain the alternative problem was only solved in 2001 [14].

Once the alternative property is known, the second interesting step is to describe the victory domains. In several examples, Isaacs discovered that the boundaries of the domains satisfy a geometric equation called *the Isaacs equations*

$$H(x, \nu_x) = 0 \quad \text{for any } x \text{ on the boundary of the victory domain} \tag{5}$$

where ν_x denotes the normal of the victory domain. Hypersurfaces satisfying such an equation are *semipermeable barriers*: one player can prevent the other from crossing the barrier. From this property it is possible to obtain much information on the winning strategies. But unfortunately the victory domains are seldom regular enough to have a normal. Very often they have "corners" even for very elementary games. So there are two different ways to treat this question. A first approach initiated by Isaacs himself, and developed by Breakwell, Bernhard, and Melykian consists of a precise and fine geometrical analysis of semipermeable barriers [6,8,24]. When this approach is possible, this gives very precise information on the behavior of the players. Unfortunately, this is hardly possible in high-dimension games or/and when the victory domain is not smooth enough. A second approach consists of proving that the boundary of the victory domain satisfies (5) in a suitable generalized sense [11,28] and to use this information to approach numerically the victory domain (cf. [13] for a detailed exposition of this method).

Quantitative Target Games

Here the goal of the player is of quantitative nature: Victor wants to minimize the time to reach the target C while Ursula wants to maximize it. For describing the game, we associate to any trajectory of the dynamics $t \in [0, +\infty) \mapsto x(t)$ the first time $x(\cdot)$ reaches C:

$$\vartheta_C(x(\cdot)) = \inf\{t \geq 0 \mid x(t) \in C\},$$

with, by convention, $\vartheta_C(x(\cdot)) = +\infty$ if $x(\cdot)$ does not reach C.

The modelization problem of such a quantitative game is to express the fact that both players act *simultaneously* on the system. So roughly speaking, we must check that if Ursula chooses her strategy before the second player, the result is the same as that in the case where Victor chooses his strategy first. This means that the following *value functions* of the game do coincide.

$$\vartheta^\flat(x_0) = \inf_\beta \sup_{u\in\mathcal{U}} \vartheta_C(x[x_0, u, \beta(u)]) \quad \text{(lower value)},$$

$$\vartheta^\sharp(x_0) = (\sup_\alpha \inf_{v\in\mathcal{V}} \vartheta_C(x[x_0, \alpha(v), v])) \quad \text{(upper value)}.$$

When $\vartheta^\flat = \vartheta^\sharp$, one says that the game has a value. Because the question of existence of value in a differential game is an essential question in game theory, Sect. "Existence of a Value for Zero Sum Differential Games" is devoted to the exposition of this feature.

It is worth pointing out, that in an opposite sense to many static games, this quantitative differential game is not in a *normal form*, namely a player does not play a strategy against a strategy of his opponents. This is another main difference with static games. In fact, it is sometimes possible to write the game in normal form allowing the nonanticipative strategies to have a small delay [15]. The normal form will be used in Sect. "Nonantagonist Differential Games" for nonzero sum games.

The *Dynamic programming principle* says, roughly speaking, that the game maintains the same structure if the time changes. Indeed, if starting from time 0 the players play the games until time t, then at time t both players are facing a differential game of the same nature as the initial game. This can be expressed by the dynamic programming equation

$$\begin{cases} \vartheta^\flat(x_0) = \inf_\beta \sup_{u\in\mathcal{U}}\{t + \vartheta^\flat(x[x_0, u, \beta(u)](t))\} \\ \vartheta^\sharp(x_0) = \sup_\alpha \inf_{v\in\mathcal{V}}\{t + \vartheta^\sharp(x[x_0, \alpha(v), v](t))\} \end{cases} \tag{6}$$

which is available as soon as the trajectories do not reach the target.

Existence of a Value for Zero Sum Differential Games

Using the dynamic programming principle, it is possible to deduce an infinitesimal characterization of the value functions (formally we subtract the two sides of one line of (6), divide by t and we let t tend to 0^+). Under the Isaacs condition (4), when the value functions are differentiable, it is not difficult to show that the values $\vartheta^\sharp$ and $\vartheta^\flat$ are solutions to a partial differential equation: the following is the *Hamilton–Jacobi–Isaacs equation*

$$H(x, D\vartheta(x)) = 1 \text{ in } \mathbb{R}^N \backslash C\,. \tag{7}$$

This fact was noticed at the very beginning of the history of differential games. Furthermore, if the Hamilton–Jacobi equation has a continuously differentiable solution, then this solution is the value function and the game can be solved using feedback strategies (for instance the Lion and Man game). This is the famous *Isaacs verification theorem*. This is a completely satisfactory solvation of the game when the value is smooth. Unfortunately, early on it was noticed that the values are not differentiable and so the previous approach is not possible.

Before going further, it is worth pointing out that this is not only a mathematical question of regularity of the value functions. Indeed the smoothness of the value function is closely related to the number of optimal strategies in the game. Most differential games, even very elementary ones have not smooth values. The reader can convince himself by considering the Lion and Man game in a circular arena with furthermore a round pillar in the center of the arena (which of course the players cannot cross).

One of the more substantial advancements in the theory of the differential game in the 1980s was due to the use of viscosity solution theory [3]. In fact, even when the values $\vartheta^\sharp$ and $\vartheta^\flat$ are not smooth, they are both solutions in a *viscosity sense* to the Hamilton–Jacobi–Isaacs equation (7). Moreover, with very weak assumptions (Lipschitzian continuity of the data) the partial differential Eq. (7) has a unique solution. Consequently the values coincide. This existence of the value result was due to Evans and Souganidis [17] and can be viewed as a generalization of Isaacs verification theorem. We refer the reader to [3] for viscosity solutions. In several cases, for instance in the state constraint games case, the value is neither smooth nor Lipschitz (even not continuous), so it is necessary to extend the Evans and Souganidis scheme. The main idea is to reduce the quantitative game to a qualitative game in a higher dimensional space [7,14]; this is the object of the *Viability theory* approach to differential games. We refer the reader to [23], pp. 3–37, for a survey of this method. Another advantage of this *reduction to qualitative* method lies in the fact that numerical analysis tools of the qualitative approach [13] can be used to build algorithms computing the value.

At this point, it is important to note that the scheme for obtaining the existence of the value is valid for a wide class of payoff. This fact can be explained by considering a game on a finite time interval $[0, T]$ such that at any initial condition $x(t_0) = x_0$ and any pair of controls u and v is associated the cost

$$C(t_0, x_0, u, v) := \int_{t_0}^{T} L(x(s), u(s), v(s))ds + g(x(T)) \tag{8}$$

where L and g are given and $x(\cdot)$ is the solution to (3) on $[t_0, T]$ with initial condition $x(t_0) = x_0$. The payoff (the cost) is then the sum of an integral cost and of a terminal cost. Victor wants to minimize this cost while Ursula wants to maximize it. This leads to the following values

$$V^\flat(t_0, x_0) := \inf_\beta \sup_{u\in\mathcal{U}} C(t_0, x_0, u, \beta(u))\,,$$

$$V^\sharp(t_0, x_0) := \sup_\alpha \inf_{v\in\mathcal{V}} C(t_0, x_0, \alpha(v), v)$$

which are both the solution to the following Hamilton–Jacobi–Isaacs equation

$$\begin{cases} \frac{\partial V}{\partial t}(t,x) + H(t,x,\frac{\partial V}{\partial x}(t,x)) = 0 & \\ & \text{for all} \quad (t,x) \in [0,T[\times\mathbb{R}^N \\ V(T,x) = g(x) & \text{for all } x \in \mathbb{R}^N . \end{cases} \tag{9}$$

Here the Isaacs condition takes the following form

$$\min_{v\in V}\max_{u\in U}\{L(x,u,v) + \langle f(x,u,v),p\rangle\} = \max_{u\in U}\min_{v\in V}\{L(x,u,v) + \langle f(x,u,v),p\rangle\}$$

and the Hamiltonian $H(x,p)$ is equal to the above expression (we refer to [7,27] in the discontinuous or constrained case).

Nonantagonist Differential Games

This section is devoted to a two player game on $[0,T]$ with different objectives: Ursula wants to maximize a cost

$$C_1(t_0,x_0,u,v) := \int_{t_0}^{T} L_1(x(s),u(s),v(s))\mathrm{d}s + g_1(x(T)),$$

while Victor wants to maximize a payoff

$$C_2(t_0,x_0,u,v) := \int_{t_0}^{T} L_2(x(s),u(s),v(s))\mathrm{d}s + g_2(x(T)).$$

Observe that if $C_2 + C_1 = 0$ (i. e. $L_1 + L_2 = 0$ and $g_1 + g_2 = 0$), the game reduces to the game studied in Sect. "Existence of a Value for Zero Sum Differential Games". Here we consider the general case where $C_1 + C_2$ is nonnecessarily equal to 0. In a nonantagonist game, it is important to play a strategy against a strategy, so we introduced the concept of nonanticipative strategies with delay on the time interval $[0,T]$. A nonanticipative strategy with delay for the first player associates to any action of the second player (to any control v) a control u of the first player such that there exists a delay $r > 0$ such that at any time t if two controls v_1 and v_2 coincide on $[0,t]$ then the associated controls $u_1 = \alpha(v_1)$ and $u_2 = \alpha(v_2)$ do coincide on $[0,t+r]$. The nonanticipative strategy with delay β for the second player is defined in a similar way. Clearly any nonanticipative strategy with delay is nonanticipative. Moreover, it is possible to prove that for an initial condition (t_0,x_0) for any pair (α,β) of nonanticipative strategies with delays there exists a *unique pair* of control (u,v) satisfying

$$\alpha(v) = u\,, \quad \beta(u) = v\,.$$

Hence it is possible to associate a trajectory to (α,β) which is the trajectory associated to (u,v) and we denote $C_1(t_0,x_0,\alpha,\beta)$ and $C_1(t_0,x_0,\alpha,\beta)$ the associated costs. This enables us to write the game in a normal form.

The antagonist differential game problem consists of finding *Nash Equilibria* defined as follows. Fix (t_0,x_0) an initial condition, a pair of real numbers (e_1,e_2) is a Nash equilibria payoff if and only if there exists a pair of nonanticipative strategies with delays $(\bar\alpha,\bar\beta)$ such that

$$e_1 = C_1(t_0,x_0,\bar\alpha,\bar\beta)\,, \quad e_2 = C_2(t_0,x_0,\bar\alpha,\bar\beta)$$

and such that for any other pair of nonanticipative strategies with delay (α,β) the following inequalities hold true

$$C_1(t_0,x_0,\bar\alpha,\bar\beta) \geq C_1(t_0,x_0,\alpha,\bar\beta)\,,$$
$$C_2(t_0,x_0,\bar\alpha,\bar\beta) \geq C_2(t_0,x_0,\bar\alpha,\beta)\,.$$

In a completely rigorous mathematical sense the above definition must be understood with a small $\varepsilon > 0$ error (the correct statement is: For all $\epsilon > 0$, there exists $(\bar\alpha,\bar\beta)$ such that

$$|e_i - C_i(t_0,x_0,\bar\alpha,\bar\beta)| \leq \epsilon\,, \quad i = 1,2$$

and for any other pair of strategies (α,β) we have

$$C_1(t_0,x_0,\bar\alpha,\bar\beta) \geq C_1(t_0,x_0,\alpha,\bar\beta) - \epsilon,$$
$$C_2(t_0,x_0,\bar\alpha,\bar\beta) \geq C_2(t_0,x_0,\bar\alpha,\beta) - \epsilon$$

cf. [12]). Nash equilibria were studied in detail in [20] for another concept of strategy. Now the remaining part of this section concerns first the characterization of the Nash equilibria through zero-sum games and second the existence of Nash equilibria. For doing this, we introduce the following value function associated with two auxiliary zero-sum games where one player is the minimizer while its opponent is the maximizer.

$$V_1(t_0,x_0) = \inf_{\beta}\sup_{u\in\mathcal{U}(t_0)} C_1(t_0,x_0,u,\beta(u))\,,$$
$$V_2(t_0,x_0) = \inf_{\alpha}\sup_{v\in\mathcal{V}(t_0)} C_2(t_0,x_0,\alpha(v),v)\,.$$

Recall that under the Isaacs condition, the value of the above auxiliary games does exist. Then the characterization of Nash payoff can be expressed: the pair (e_1,e_2) is a Nash equilibrium payoff if and only if there exist two controls (u,v) such that

$$e_1 = C(t_0,x_0,u,v), \quad e_2 = C(t_0,x_0,u,v)$$

and for all time $t \in [t_0,T]$

$$e_1 \geq V_1(t,x[t_0,x_0,u,v](t))\,,$$
$$e_2 \geq V_2(t,x[t_0,x_0,u,v](t))\,.$$

Observe that this is not an intuitive result (once again a correct statement is, a $\epsilon > 0$ is needed to have a rigorous

statement: For any ϵ there exists a pair (u, v) satisfying the above inequality up to an error ϵ). ¿From this result, it is possible – but not easy – to obtain the existence of a Nash equilibrium pair under Isaacs condition.

Observe that at a first glance, it is very surprising to obtain the existence of the Nash equilibria assuming Isaacs condition which means the existence of a saddle point in static games. One could assume instead the existence of Nash equilibria of static games. Unfortunately, except for very specific games like some linear quadratic differential games, this method fails. Another way to understand this lies in the fact that Nash equilibria are very unstable (cf. pp. 57–67 in [23]). We refer the reader to the bibliography for noncooperative games in specific cases and applications.

Miscellaneous

This part is devoted to the description of Stochastic differential games (games with randomness in the dynamics) and Worst Case Design where games are used as a tool to prevent the worst case caused by nature or uncertainty.

Stochastic Differential Games

The dynamics is given by a stochastic differential equation

$$\mathrm{d}X_t = f(X_t, u_t, v_t)\mathrm{d}t + \sigma(X_t, u_t, v_t)\mathrm{d}W_t\,, \tag{10}$$

where W is a Brownian motion on a given probability space. Victor and Ursula choose the controls u and v which are adapted processes (with respect to the filtration generated by the Brownian motion). The notion of nonanticipative strategies is similar to the determinist case but the strategies are nonanticipative in time *almost surely*. The rigorous description of the game is rather technical due to the need of stochastic analysis. The reader can refer to the nice article [30] for a precise formulation. We want just to stress some important aspects of stochastic zero-sum games. Of course the notions of qualitative and quantitative games are relevant. The qualitative aspect is not yet well-studied. Oppositely the quantitative aspect is now well studied. Consider for instance a payoff of the form

$$C(t_0, x_0, u, v) := E\left[\int_{t_0}^{T} L(x(s), u(s), v(s))\mathrm{d}s + g(x(T))\right]\,,$$

where E is the expectation. The existence of the value was obtained in [18], by adopting the same scheme as described in Subsect. "Quantitative Target Games": The values are the unique viscosity solution of a – second order – Hamilton–Jacobi–Isaacs equation which has a unique solution. The nonantagonist case could be solved similarly to the zero-sum case up to hard technicalities due to stochastic analysis [9].

Worst Case Design

This concerns a zero-sum quantitative game where a controller wants to drive a system against another action on the system (uncertainty, nature) considered as a second player. The controller wants to find the most *robust* strategy which prevents the worst case of nature. Clearly the question of existence of a value is not relevant in this case. Only one value is interesting in this case.

The linear quadratic case was extensively studied in the book [4]. In the fully nonlinear case, the Hamilton–Jacobi–Isaacs equation becomes infinite dimensional. An alternative approach can be made by setting Hamilton–Jacobi equations on the space of closed sets [29].

Future Directions

Among several future directions of investigations, let us mention first the case of impulsive differential games where the player can at any time either follow a continuous dynamic of the form (3) or make a jump. These games combine the effects and difficulties of static and differential games. Only very preliminary results are available (pp. 223–249 in [23]) for pursuit games when both players cannot "jump" simultaneously. Another interesting class of pursuit impulsive game, is the case where each player can choose at every time point between two dynamics. Once again, this research domain is still wide open.

Another important and very active field of research is devoted to information phenomena. This is well understood in the case of (discrete) repeated games (cf. [2]) with the help of mixed strategies but still open for differential games. In more concrete terms: Consider the two-player quantitative game of Sect. "Qualitative and Quantitative Differential Games", with dynamics (3) and cost (8). We consider that the initial position belongs to a set of $I \times J$ possible initial positions

$$x_0^{i,j}, \quad i = 1\,, \ldots I, \quad j = 1 \ldots J\,.$$

At the initial time the game is played in two steps, first the integer i is chosen according to a probability p (belonging to the set $\Delta(I)$ of probabilities on $\{1, 2 \ldots I\}$) while j is chosen according to a probability $q \in \Delta(J)$; second the index i is communicated to Ursula only and the index j is communicated to the player j. As previously stated Victor wants to minimize the cost C and Ursula want to maximize it. Each player has a *private information* (the knowledge

of i respectively j) which cannot allow him/her to compute the current position of the game. Each player can only observe his player behavior (choice of controls) trying to deduce from this observation his missing information. The problem consists of checking the existence of a value for suitably defined mixed strategies.

Very recently, this specific game was solved, but the more general question of taking into account the information phenomena in differential games has been little studied and a wide field of research is just starting to rise in this direction.

Bibliography

Primary Literature

1. Aubin JP (1991) Viability Theory. Birkhäuser, Basel
2. Aumann RJ, Maschler M (1995) Repeated games with incomplete information. MIT Press, Cambridge
3. Bardi M, Capuzzo Dolcetta I (1997) Optimal control and viscosity solutions of Hamilton–Jacobi–Bellman equations. Systems and Control: Foundations and Applications, vol xvii. Birkhäuser, Boston, pp 570
4. Basar T, Bernhard P (1995) H^{∞}-optimal control and related minimax design problems. A dynamic game approach, 2nd edn. Systems and Control: Foundations and Applications. Birkhäuser, Basel
5. Berkovitz LD (1994) A theory of differential games. In: Basar T et al (eds) Advances in dynamic games and applications, Birkhäuser, Basel; Ann Int Soc Dyn Games 1:3–22
6. Bernhard P (1988) Differential games in Systems and control Encyclopedia. In: Singh MG (ed) Theory Technology Application. Pergamon Press, Oxford
7. Bettiol P, Cardaliaguet P, Quincampoix M (2006) Zero-sum state constrained differential games: Existence of value for Bolza problem. Int J Game Theor 34(4):495–527
8. Breakwell JV (1977) Zero-sum differential games with terminal payoff. In: Hagedorn P, Knobloch HW, Olsder GH (eds) Differential Game and Applications. Lecture Notes in Control and Information Sciences, vol 3. Springer, Berlin
9. Buckdahn R, Cardaliaguet P, Rainer C (2004) Nash equilibrium payoffs for nonzero-sum stochastic differential games. SIAM J Control Optim 43(2):624–642
10. Cardaliaguet P (1996) A differential game with two players and one target. SIAM J Control Optim 34(4):1441–1460
11. Cardaliaguet P (1997) Nonsmooth semi-permeable barriers, Isaacs equation, and application to a differential game with one target and two players. Appl Math Optim 36:125–146
12. Cardaliaguet P, Plasckacz S (2003) Existence and uniqueness of a Nash equilibrium feedback for a simple non zero-sum differential game. Int J Game Theor 32(4):561–593
13. Cardaliaguet P, Quincampoix M, Saint-Pierre P (1999) Numerical methods for optimal control and numerical games. In: Bardi M, Parthasarathy T, Raghavan TES (eds) Annals of International Society of Dynamical Games. Birkhäuser, Basel, pp 177–249
14. Cardaliaguet P, Quincampoix M, Saint-Pierre P (2001) Pursuit differential games with state constraints. SIAM J Control Optim 39(5):1615–1632
15. Cardaliaguet P, Quincampoix M (2008) Determinist Differential games under probability knowledge of initial condition. Int Game Theor Rev 6(1):1–16
16. Elliot N, Kalton N (1972) The existence of value in differential games. Mem Am Math Soc 126:67
17. Evans LC, Souganidis PE (1984) Differential games and representation formulas for solutions of Hamilton–Jacobi equations. Indiana Univ Math J 282:487–502
18. Fleming W, Souganidis P (1989) On the existence of value functions of two-player, zero-sum stochastic differential games. Indiana Univ Math J 38(2):293–314
19. Flynn J (1973) Lion and Man: The Boundary Constraints. SIAM J Control 11:397
20. Klejmenov AF (1993) Nonantagonistic positional differential games. Nauka, Ekaterinburg (in Russian)
21. Krassovski NN, Subbotin AI (1988) Game-Theorical Control Problems. Springer, New York
22. Isaacs R (1965) Differential Games. Wiley, New York
23. Jorgensen S, Quincampoix M, Vincent T (2007) Advances in Dynamic Games Theory. Annals of International Society of Dynamic Games. Birkhäuser, Basel
24. Melikyan AA (1998) Generalized characteristics of first order PDEs. Applications in optimal control and differential games. Birkhäuser, Boston
25. Petrosjan LA (2004) Cooperation in games with incomplete information. Nonlinear analysis and convex analysis. Yokohama Publ, pp 469–479
26. Pontryagin N (1968) Linear Differential Games, I, II. Sov Math Doklady 8(3–4):769–771; 910–913
27. Plaskacz S, Quincampoix M (2000) Discontinuous Mayer control problem under state-constraints. Topol Method Nonlinear Analysis 15:91–100
28. Quincampoix M (1992) Differential inclusions and target problems. SIAM J Control Optim 30(2):324–335
29. Quincampoix M, Veliov V (2005) Optimal control of uncertain systems with incomplete information for the disturbance. Siam J Control Optim 43(4):1373–1399
30. Rainer C (2007) On two different approaches to nonzero-sum stochastic differential games. Appl Math Optim 56:131–144
31. Roxin E (1969) The axiomatic approach in differential games. J Optim Theor Applt 3:153–163
32. Subbotin AI (1995) Generalized solutions of first-order PDEs, The dynamical optimization perspective. Translated from the Russian. Systems and Control: Foundations and Applications. Birkhäuser, Boston
33. Varaiya P (1967) The existence of solution to a differential game. SIAM J Control Optim 5:153–162
34. Varaiya P, Lin J (1967) Existence of saddle points in differential game. SIAM J Control Optim 7(1):141–157
35. Von Neumann J, Morgenstern O (1946) Theory of Games and Economic Behaviour. Princeton University Press, Princeton

Books and Reviews

Bardi M, Raghavan TES, Parthasarath T (1999) Stochastic and differential games. Theory and numerical methods, Dedicated to Prof. Subbotin AI. Annals of the International Society of Dynamic Games, vol 4. Birkhäuser, Boston

Basar T, Olsder GJ (1999) Dynamic noncooperative game theory, 2nd edn. Classics in Applied Mathematics, vol 23. SIAM, Society for Industrial and Applied Mathematics, Philadelphia

Blaquiére A, Gérard F, Leitman G (1969) Quantitative and Qualitative Games. Academic Press, New York
Dockner E, Jrgensen S, Van Long N, Sorger G (2000) Differential games in economics and management science. Cambridge University Press, New York
Hajek O (1975) Pursuit games. Academic Press, New York
Patsko VS, Turova VL (2000) Numerical study of differential games with the homicidal chauffeur dynamics. Russian Academy of Sciences, Institute of Mathematics and Mechanics, Ekaterinburg
Petrosyan LA (1993) Differential games of pursuit. Series on Optimization, vol 2. World Scientific, Singapore

Diffusion of Innovations, System Dynamics Analysis of the

Peter M. Milling[1], Frank H. Maier[2]
[1] Industrieseminar der Universität Mannheim, Mannheim University, Mannheim, Germany
[2] International University in Germany, Bruchsal, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Principle Structures to Model the Diffusion of Innovations
Representing Managerial Decision Making in Innovation Diffusion Models
Managerial Implications
Future Directions
Bibliography

Glossary

Adopters The cumulated number of persons who have bought a product over time.

Diffusion The spread of a new product, process or concept in the market. The process of bringing innovation into wide use.

Invention The process of bringing new technology into being.

Innovator A customer with general interest in innovations making his buying decision independent of others.

Innovation The process of bringing new technology into use.

Installed base Installed base is defined as the amount of users in a network system.

Imitator An imitator buy a new product because he observed or communicated with customers who have already bought the product. The buying decision of imitators is influenced by the adoption of other customers.

Network effects A product is characterized by a network effect, if the utility of that product is a function of the installed base. The utility increases with the installed base.

Definition of the Subject

The article describes how system dynamics-based models can contribute to the understanding and improved management of the diffusion of innovations. It emphasizes the importance of an integrated feedback-oriented view of the different stages of innovation processes. The aim is to generate insight in the complexity and the dynamics of innovation processes. Based on the classical Bass model of innovation diffusion, the system dynamics perspective is introduced. In a systematic approach several structures to model the complexity and dynamics of managerial decision-making in the context of the diffusion of innovation are described and analyzed. Aspects covered consider market structure, network externalities, dynamic pricing, manufacturing related decisions and the link between research and development and the diffusion of a new product in the market place. The article concludes with managerial implications.

Introduction

Continuous activities to renew a company's range of products are crucial for the survival in a competitive environment. However, to improve the competitive position or the competitive advantage, ongoing innovation activity through the development, test, and introduction of new products is necessary. At least since the 1970s, it could be observed that new and technically more complex and sophisticated products have to be developed in a shorter time span. Resources have to be allocated to research and development (R&D) projects that are expected to be economically successful. New products have to be introduced to global markets with severe competition. Decisions about the adequate time to market and appropriate pricing, advertising, and quality strategies have to be made.

The complexity and difficulties to manage innovation activities partly derive from the comprehensiveness of the innovation processes. To be competitive, companies have to be successful in all stages of the innovation process, i. e., the process of invention, innovation, and diffusion. This becomes obvious when new product failure rates and innovation costs are analyzed. Figure 1 illustrates the cas-

cading process of innovation activity and the related innovation costs.

For one successful new product in the market place, 64 promising ideas must be channeled through the process of invention and innovation. The cost at each stage of the invention and innovation process increases from a $1000 to $5 million per attempt. Not only is failure more expensive in later stages – which requires an effective management to reduce the failure rates – successful new products have to earn all necessary resources for the whole process. This requires the following: (1) to manage R&D projects and processes effectively and efficiently – including thorough and educated assessment of the economic potential of a new product – to reduce failure rates in later stages and (2) to increase management attention in the final stages since failures in late stages of the process are much more expensive.

Models of innovation diffusion can support the complex and highly dynamic tasks. The article will briefly examine how system dynamics-based analysis of innovation diffusion can contribute to the understanding of the structures and forces driving the processes of innovation and diffusion. It will show how system dynamics models can support the decision-making and how they can help to reduce failures in the later stages of innovation activities.

Principle Structures to Model the Diffusion of Innovations

Traditional Innovation Diffusion Models from a System Dynamics Perspective

In literature discusses plenty of models about the diffusion of innovations. Many models are based on Frank M. Bass' model of innovation diffusion. In this model, product purchases result from two distinct forms of buying behavior, i. e., innovative purchases and imitative purchases. According to the original Bass model, innovative purchases of a period can be calculated as a fraction of the remaining market potential $(N - X_{t-1})$ with N being the market potential and $X_{t-1} = \sum_{\tau=0}^{t-1} S_\tau$ representing the accumulation of all past purchases of the product S_τ until period $t - 1$.

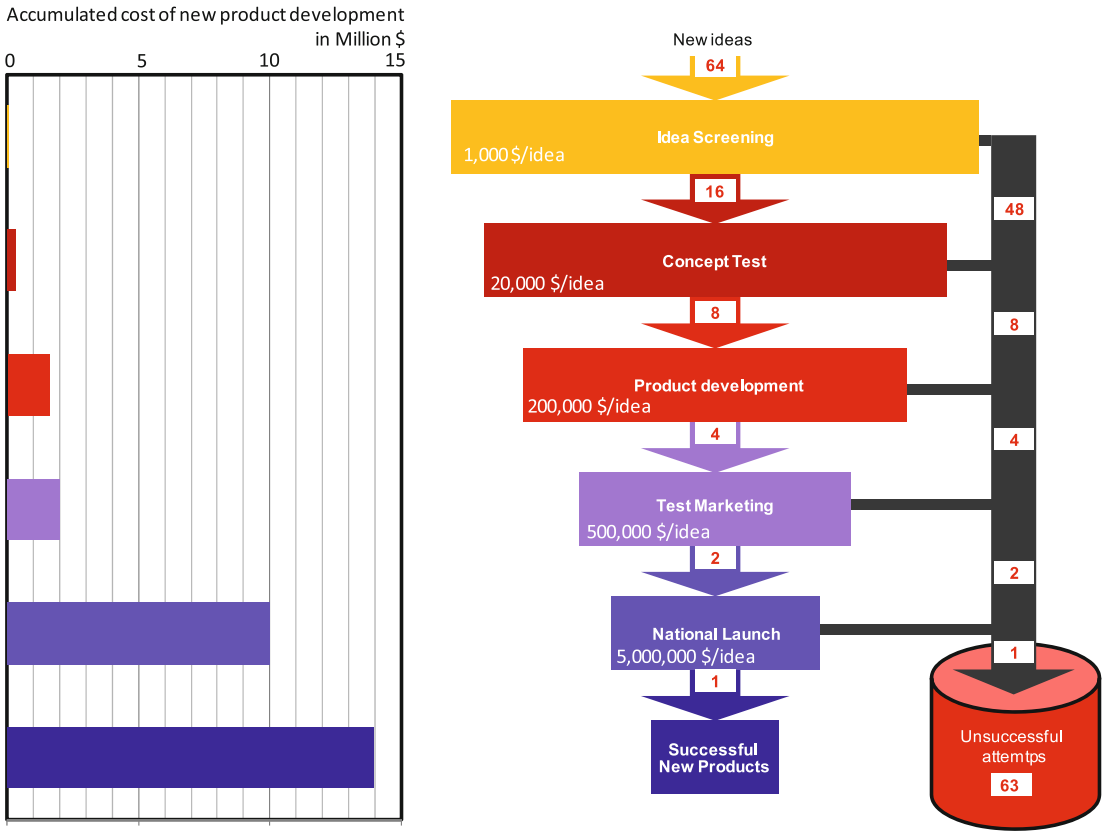

Diffusion of Innovations, System Dynamics Analysis of the, Figure 1
Outcome of activities along the process of invention and innovation

According to this, innovative purchases S_t^{inno} can be calculated as

$$S_t^{\text{inno}} = \alpha \cdot \left(N - \sum_{\tau=0}^{t-1} S_\tau \right) \tag{1}$$

where α represents the coefficient of innovation. In the original model, this coefficient is a constant essentially representing the fraction of innovators of the remaining market potential at any point of time. Imitative purchases, however, are influenced by the number of purchases in the past. Potential adopters of an innovation make their purchasing decision depending on the spread of the product in the market place. The more customers have adopted the product in the past, the higher is the social pressure to purchase the product as well. Imitative demand of a period S_t^{imit} hence can be calculated as

$$S_t^{\text{imit}} = \beta \cdot \frac{\sum_{\tau=0}^{t-1} S_\tau}{N} \cdot \left(N - \sum_{\tau=0}^{t-1} S_\tau \right) \tag{2}$$

with β representing the coefficient of imitation – a probability that a purchase takes place by someone who observed the use of a product. Together, the total purchases in a period S_t^{total} equal $S_t^{\text{inno}} + S_t^{\text{imit}}$ and hence are calculated as

$$S_t^{\text{total}} = S_t^{\text{inno}} + S_t^{\text{imit}} = \alpha \cdot \left(N - \sum_{\tau=0}^{t-1} S_\tau \right) + \beta \cdot \frac{\sum_{\tau=0}^{t-1} S_\tau}{N} \cdot \left(N - \sum_{\tau=0}^{t-1} S_\tau \right) . \tag{3}$$

Innovative and imitative purchases together create the typical product life cycle behavior of the diffusion of an innovation in the market place as shown in Fig. 2.

The model above is a simple mathematical representation of the product life cycle concept, a key framework in business management. It describes the time pattern a product follows through subsequent stages of introduction, growth, maturity, and decline. Because of its mathematical simplicity and its ability to represent the diffusion of an innovation, the Bass model has been used for parameter estimation and therefore serves as a base for projections of future sales. Although the concept is a powerful heuristic, many models generating this typical behavior do not consider e. g., actual economic environment, competition, capital investment, cost and price effects. Innovation diffusion models, which do not comprise the relevant decision variables, exhibit a significant lack of policy content. They do not explain how structure conditions behavior. They cannot indicate how actions of a firm can promote but also impede innovation diffusion. For an improved understanding of innovation dynamics generated by feedback structures that include managerial decision variables or economic conditions, the system dynamics approach is highly suitable.

Equations (1) to (3) can easily be transformed into the system dynamics terminology. $(N - X_{t-1})$ represents the stock of the remaining market potential at any point in time and X_{t-1} represents the accumulation of all product purchases over time. The sales of a period S_t^{total} are the flows connecting these two stocks as shown in the Fig. 3.

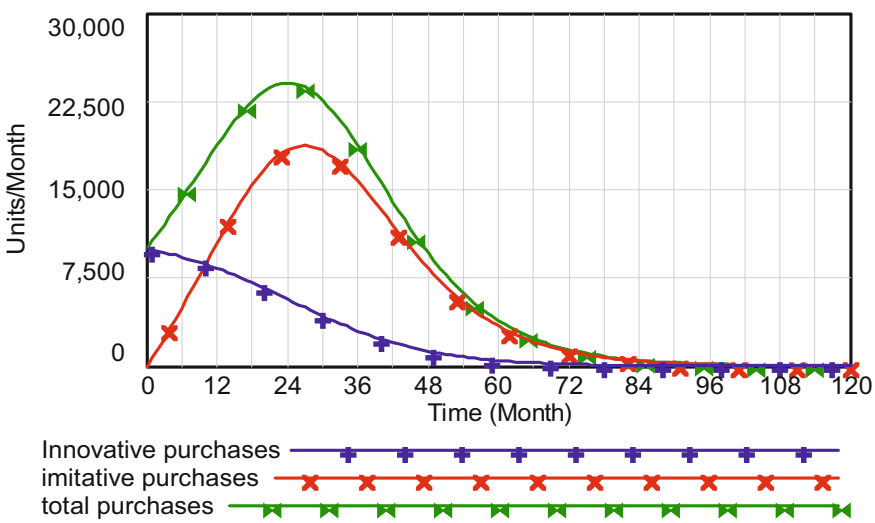

Diffusion of Innovations, System Dynamics Analysis of the, Figure 2
Product life cycle behavior generated by the Bass model

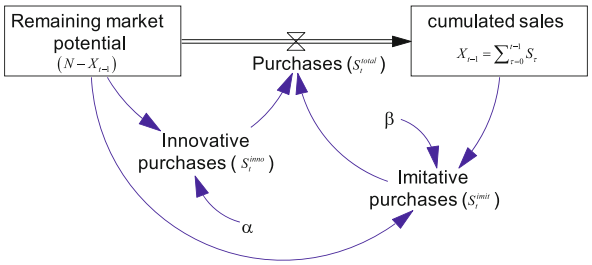

Diffusion of Innovations, System Dynamics Analysis of the, Figure 3
Stock-flow view of the Bass model

The coefficients α and β represent the probability of a purchase taking place; they are constants in the original Bass model and independent of any decisions or changes over time. For this reason, the model has been criticized and subsequent models have been developed that make the coefficients depending on variables like price or advertising budget. Most of the extensions, however, include no feedback between the diffusion process and these decision variables. This is a severe shortcoming since in the market place, diffusion processes are strongly influenced by feedback. What in classical innovation diffusions models typically is referred to as word-of-mouth processes is nothing else than a reinforcing feedback process. Adopters of an innovation – represented by the cumulated sales X_{t-1} – communicate with potential customers $(N - X_{t-1})$ and – by providing information about the product – influence their behavior. However, feedback in innovation diffusion goes beyond the pure word-of-mouth processes. It also involves the decision processes of a company and the outcome generated by the purchasing decision of the customers like the sales volume generated.

Figure 4 describes as a causal loop diagram the diversity of potential influences of corporate decision variables (marked with hexagons) on demand of the products by making the probability of a purchase – the coefficients α and β – depending on decision variables. It also shows how corporate decisions are interconnected through several feedback structures and influence the diffusion of a new product in the market place. Although be-

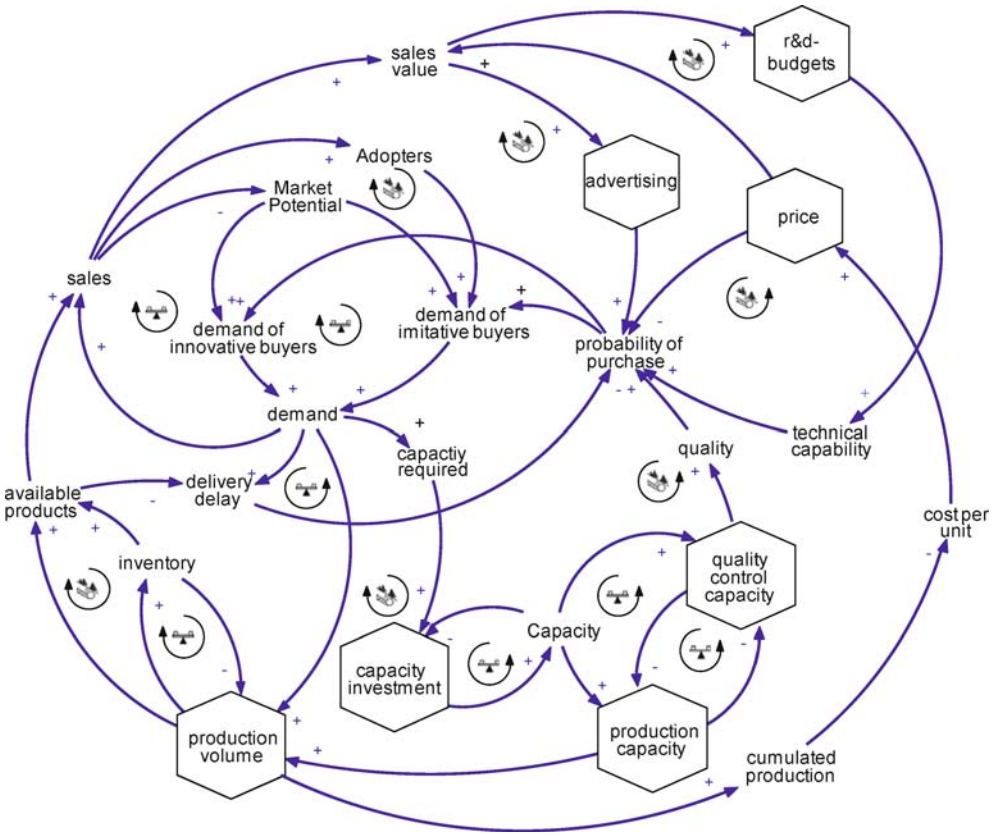

Diffusion of Innovations, System Dynamics Analysis of the, Figure 4
Feedback structures driving innovation processes

ing far from a comprehensive structure of potential feedback, the figure gives an impression of the complex dynamic nature of innovation diffusion processes.

Decision variables like pricing or advertising directly influence the purchase probability of a potential customer. The higher the advertising budgets and the lower the price, the higher will be demand for the products of a company. Furthermore, there are indirect and/or delayed effects on the speed of the spread of a new product in the market. Actual sales of a product may be limited by insufficient production and inventory levels which increases delivery delays (perceived or actual) and therefore reduce demand. Growing demand, however, motivates the company to expand its capacity and to increase the volume of production. This leads to higher cumulated production and through experience curve effects to decreasing costs per unit, lower prices, and further increased demand. Other influences might reflect that a certain percentage of total available production capacity has to be allocated to ensure the quality of the output – either by final inspection or during the production process. Quality control then will improve product quality, which directly affects demand.

Models developed in this manner can serve as simulators to analyze the consequences of strategies and to improve understanding. They can show e. g., how pricing and investment strategies depend on each other and quantify the impact of intensified quality control on production and sales. They are suitable tools to investigate the effects resulting from the impact of a particular management problem on the dynamic complexity of innovation diffusion. Creating an understanding of the processes and interactions is the main purpose of system dynamics-based innovation diffusion models. Subsequently, a base structure of a system dynamics-based model will be described.

Base Structure of a System Dynamics-Based Model of Innovation Diffusion

First, a model will be discussed that maps the diffusion of an innovation in a monopolistic situation or can serve as an industry level model. Secondly, competition between potential and existing companies is introduced. Thirdly, substitution between successive product generations is considered. Each step adds complexity to the model. This approach allows for a better understanding of the forces driving the spread of a new product in the market.

In the following, the coarse structure of a model generating the life cycle in the market of a new product is presented and analyzed in its dynamic implications in Sect. "Representing Managerial Decision Making in Innovation Diffusion Models". Figure 5 gives an aggregated view the main model structure. It also introduces – in contrast to the mathematical terms known from the Bass model, variable names, which are informative and consistent with the use in system dynamics models.

The diffusion of a new product is generated by the behavior of the before mentioned two different types of buyers: innovators and imitators. If the potential customers (PC) – i. e., the remaining market potential of a product – decide to purchase, either as innovators or as imitators, they become adopters *(ADOP)*. The variables *PC* and *ADOP* and their associated transfer rates are the basic variables of the core diffusion process. The untapped market *(UM)* covers latent demand that can be activated by appropriate actions and leads to an increase in the number of potential customers and therefore increases the remaining market potential. Besides the growth resulting from the influx from the untapped market, a decline in market volume can be caused by the loss of potential customers to competitors. This lost demand *(LD)* turned to competing products that are more attractive, e. g., products of a higher level of technological sophistication, quality or lower price.

The differentiation into the two buying categories "innovators" and "imitators" refers to the Bass model of innovation diffusion as described in Subsect. "Traditional Innovation Diffusion Models from a system dynamics Perspective". The distinction is made because these two types of buyers react differently to prices charged, product quality offered, advertisements or the market penetration already achieved. The term "innovator" refers to customers who make their purchasing decision without being influenced by buyers who already purchased the product, the adopters. In the beginning of an innovation diffusion process, innovators take up the new product because they are interested in innovations. The number of innovators is a function of the potential customers. Mathematically, the purchasing decision of innovators D^{Inno} is defined by a coefficient of innovation α times the number of potential customers *PC*.

$$D^{\text{inno}}_{(t)} = \alpha_{(t)} \cdot PC_{(t)} \tag{4}$$

with:

$D^{\text{inno}}_{(t)}$	Demand from innovators
$\alpha_{(t)}$	Coefficient of innovation
$PC_{(t)}$	Potential customers .

The purchasing decision of "imitators" is calculated differently. Imitators buy a new product because they observe or communicate with customers who have already adopted the product. They imitate the observed buying

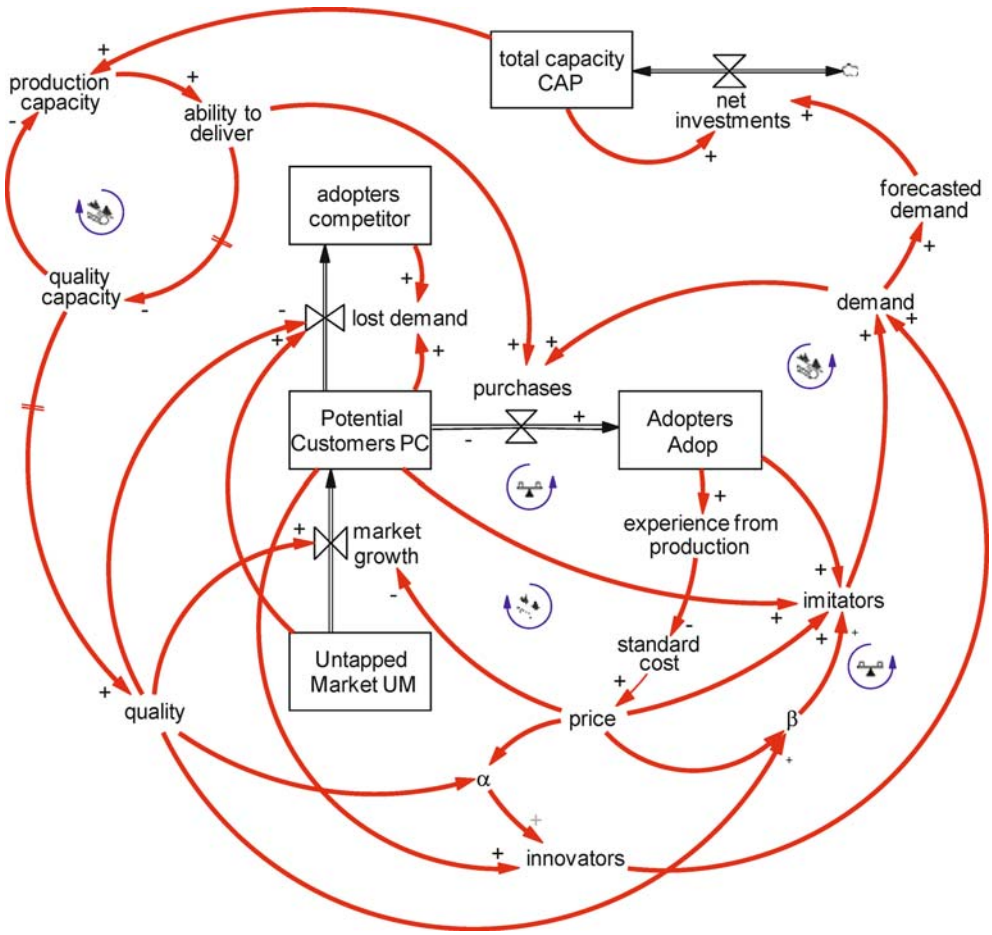

Diffusion of Innovations, System Dynamics Analysis of the, Figure 5
Coarse structure of the innovation diffusion model

behavior. Innovators initiate new product growth, but the diffusion gains momentum from the word-of-mouth process between potential customers and the increasing level of adopters. The driving force behind the imitation process is communication – either personal communication between an adopter and someone who still does not own the product or just observation of someone already owning and using the product. Although, the Bass model describes how the imitators' purchases can be calculated – as shown in Eq. (2) – the equation can also be derived from a combinatorial analysis of the number of possible contacts between the adopters and the potential customers. If N is the total number of people in a population consisting of potential customers PC and adopters $ADOP$, the amount of possible combinations C_N^k is

$$C_N^k = \binom{N}{k} = \frac{N!}{k!(N-k)!} \,. \tag{5}$$

Here we are only interested in paired combinations ($k = 2$) between the elements in N

$$\begin{aligned} C_N^2 = \binom{N}{2} &= \frac{N!}{2!(N-2)!} \\ &= \frac{N(N-1)}{2!} = \frac{1}{2}(N^2 - N) \,. \end{aligned} \tag{6}$$

Since N represents the sum of elements in PC and in $ADOP$, ($N = PC + ADOP$), the number of combinations between potential customers and adopters is

$$\begin{aligned} &= \frac{1}{2}\left[(PC + ADOP)^2 - (PC + ADOP)\right] \\ &= \frac{1}{2}\left[PC^2 + 2 \cdot PC \cdot ADOP + ADOP^2 - PC - ADOP\right] \end{aligned} \tag{7}$$

and after regrouping and collecting terms, we get

$$= \frac{1}{2}\Big(\underbrace{2 \cdot PC \cdot ADOP}_{\substack{\text{Communication between} \\ PC \text{ and } ADOP}} + \underbrace{PC^2 - PC}_{\substack{\text{Communication} \\ \text{within } PC}} + \underbrace{ADOP^2 - ADOP}_{\substack{\text{Communication} \\ \text{within } ADOP}} \Big) . \tag{8}$$

Internal communications, both within *PC* and *ADOP*, generate no incentive to purchase the new product and are neglected; the process of creating imitative buying decisions in Eq. (9) is, therefore, reduced to the first term in Eq. (8), the information exchange between potential customers and adopters.

$$D^{\text{imit}}_{(t)} = \beta^* \cdot PC_{(t)} \cdot ADOP_{(t)} \tag{9}$$

with:

$D^{\text{imit}}_{(t)}$	Demand from imitators
$\beta^*_{(t)}$	Coefficient of imitation $= \dfrac{\beta_{(t)}}{N}$
$ADOP_{(t)}$	Adopters
N	Initial market potential.

The coefficient of imitation $\beta^*_{(t)}$ represents the original coefficient of innovation β from the Bass model divided by the initial market potential N. β can be interpreted as the probability that the possible contacts between members in *PC* and *ADOP* have been established, relevant information has been exchanged, and a purchasing decision is made.

The sum of the demand of innovators and imitators in each period, $D_{(t)}$, establishes the basic equation for the spread of a new product in the market. Together with the state variables of potential customers and adopters the flows of buyers (innovators and imitators) constitute the core model of innovation diffusion, which generates the typical s-shaped pattern of an adoption process over time.

$$\begin{aligned} D_{(t)} &= D^{\text{inno}}_{(t)} + D^{\text{imit}}_{(t)} \\ &= \alpha_{(t)} \cdot PC_{(t)} + \beta^*_{(t)} \cdot PC_{(t)} \cdot ADOP_{(t)} . \end{aligned} \tag{10}$$

Although Eqs. (3) and (10) are based on different interpretations and explanations, they are structurally identical since *PC* equals $(N - X_{t-1})$ and *ADOP* equals $X_{t-1} = \sum_{\tau=0}^{t-1} S_\tau$. The only difference is that the coefficients of innovation and imitation, in the context of the model based on (10) are now a variable – rather than a constant – depending on corporate decision variables like price or quality. Furthermore, corporate decisions are not just set as predefined time paths; they are endogenously calculated and depend on the outcome of the diffusions process itself. Model simulations of this extended innovation diffusion model will be discussed in Sect. "Representing Managerial Decision Making in Innovation Diffusion Models".

Extending the Base Structure to Include Competition

In the model described above, competition is not modeled explicitly. The model only assumes a potential loss in demand, if price, quality or ability to deliver are not within the customers' expectations. The internal corporate structures of competition are not explicitly represented. To generate diffusion patterns that are influenced by corporate decisions and the resulting dynamic interactions of the different competitors in a market, a more sophisticated way to incorporate competition is needed. Therefore, a subscript i $(i = 1, 2, \ldots, k)$ representing a particular company is introduced as a convenient and efficient way to model the different competitors. In a competitive innovation diffusion model the calculation of innovative and imitative demand of a company has to be modified. Equation (4) that determines the innovative demand in a monopolistic situation becomes Eq. (11) – in the following discussion, the time subscript *(t)* is omitted for simplicity. The coefficient of innovation α has to be divided by the number of competitors N to ensure that each company will have the same share of innovative demand as long as there is no differentiation among the competitors' products through, e. g., through pricing or advertising. The subscript i in the coefficient of innovation is necessary because it considers that the decisions of an individual company regarding product differentiation influences its proportion of innovative buyers. A third modification is necessary, because the number of competitors may vary over time. Therefore, the term φ_i represents a factor to model different dates of market entry. It takes the value 1 if a company i is present at the market, otherwise it is 0. Hence, the demand of company i is 0, as long as it is not present at the market and $\sum_{i=1}^{k} \varphi_i$ represents the actual number of competitors. The variable potential customers *PC* has no subscript because all companies in the market compete for a common group of potential customers, whereas innovative demand has to be calculated for each company.

$$D^{\text{inno}}_i = \frac{\alpha_i}{NC} \cdot PC \cdot \varphi_i \tag{11}$$

with:

α_i	coefficient of innovation for company i
NC	number of active competitors $= \sum_{i=1}^{k} \varphi_i$
φ_i	factor of market presence company i

i subscript representing the companies $i = (1, 2, \ldots, k)$.

The buying decisions of imitators are influenced by observation of, or communication with the adopters *(ADOP)*. In a competitive environment two alternative approaches can be used to calculate imitative demand. These different approaches are a result of different interpretations of the object of the communication processes. In the first interpretation, the 'product related communication', the adopters of a particular company's product communicate information about the product they have purchased e. g., an electronic device like a MP3 player of a particular company. In this case, the calculation of imitative demand has to consider the number of potential contacts between the potential customers *PC* and the adopters of the products of company i ($ADOP_i$) as shown in Eq. (12).

$$D_i^{\text{imit}} = \frac{\beta_i}{N} \cdot ADOP_i \cdot PC \cdot \varphi_i \tag{12}$$

with:

β_i coefficient of imitation for company i.

The second interpretation about the object of communication is the 'product form-related communication'. Here, the adopters communicate information about a product form, for example, DVD players in general and not about an MP3 player of a particular company. The equation to calculate imitative demand for the model of product form related communication is shown in Eq. (13). The sum of adopters for each company i $\left(\sum_{i=1}^{k} ADOP_i\right)$ represents the total number of adopters in the market. The product of the total adopters and the potential customers then represents the total number of potential contacts in the market. Imitative demand of a company i depends on the share of total adopters $\frac{ADOP_i}{\sum_{i=1}^{k} ADOP_i}$ this company holds.

$$D_i^{\text{imit}} = \frac{\beta_i}{N} \cdot \frac{ADOP_i}{\sum_{i=1}^{k} ADOP} \cdot PC \cdot \sum_{i=1}^{k} ADOP_i \cdot \phi_i \,. \tag{13}$$

If the term that represents a company's share of the total adopters of a market $\frac{ADOP_i}{\sum_{i=1}^{k} ADOP_i}$ is raised to the power of γ as in Eq. (14), weaker ($0 < \gamma < 1$) or stronger ($\gamma > 1$) influences of a company's share of total adopters on demand can be represented explicitly. For $\gamma = 1$, Eq. (14) is identical to Eq. (13).

$$D_i^{\text{imit}} = \frac{\beta_i}{N} \cdot \left(\frac{ADOP_i}{\sum_{i=1}^{k} ADOP}\right)^{\gamma} \cdot PC \cdot \sum_{i=1}^{k} ADOP_i \cdot \phi_i \tag{14}$$

with:

γ factor representing customers' resistance to "Me-too"-pressure.

Figure 6 shows the effects of a company's share of the total adopters for different γ. For a given share of total adopters

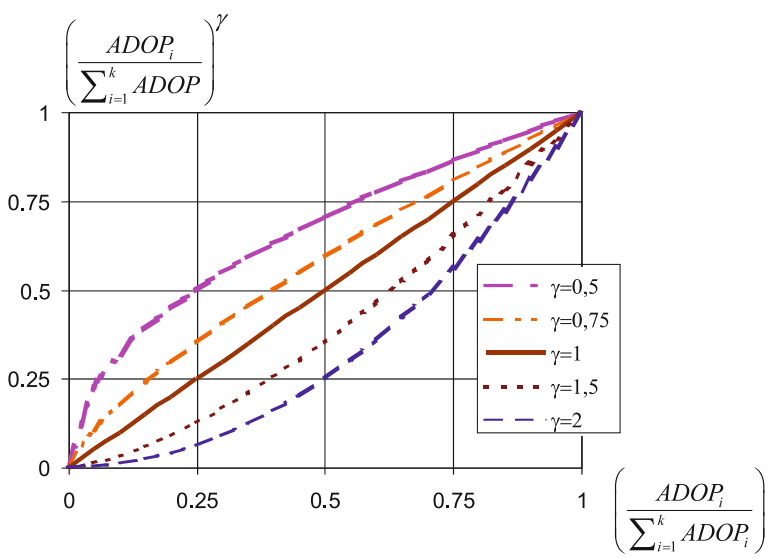

Diffusion of Innovations, System Dynamics Analysis of the, Figure 6
Effects of a company's share of adopters for different γ

this means: the higher γ, the lower is the value of the term

$$\left(\frac{ADOP_i}{\sum_{i=1}^{k} ADOP_i}\right)^{\gamma}$$

and the stronger is the importance of a high share of total adopters. The parameter γ can be interpreted as a measure of the importance of customer loyalty or as resistance to "me-too" pressure.

Figure 7 illustrates the coarse structure of an oligopolistic innovation diffusion model as described by Eqs. (11) and (14). The hexahedron at the top represents the stock of potential customers *PC* for the whole market. The blocks with the different shading represent for each company *i* the level of adopters, i. e., the cumulated sales of the company. The total number of adopters of the product form corresponds to the addition of these blocks.

Since the sales are calculated separately for each company *i* there are n outflows from the stock of potential customers to the adopters. Again, sales comprise innovative and imitative demand, which are influenced by the coefficient of innovation α_i and imitation β_i. Both coefficients are influenced by managerial decisions of each company *i* like pricing, advertising, quality, market entry timing, etc. and measure the relative influence of the decisions compared to the competitor's decisions. Therefore, the values

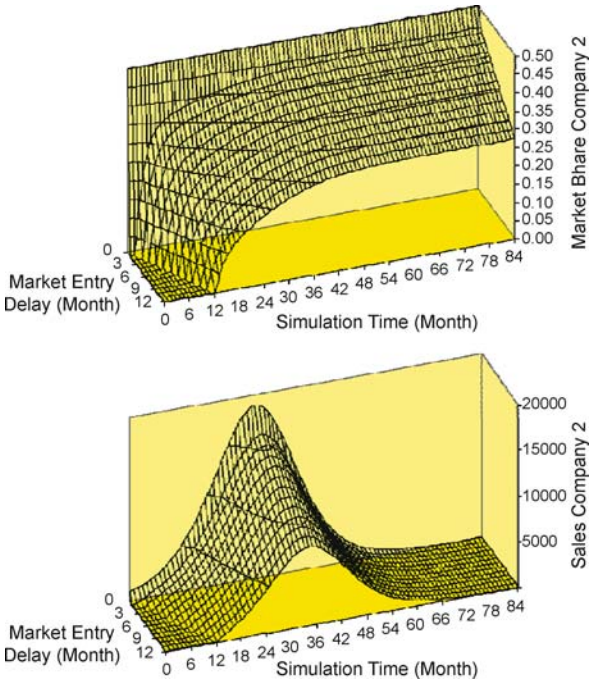

Diffusion of Innovations, System Dynamics Analysis of the, Figure 8
Follower's market share and sales for different market entry times

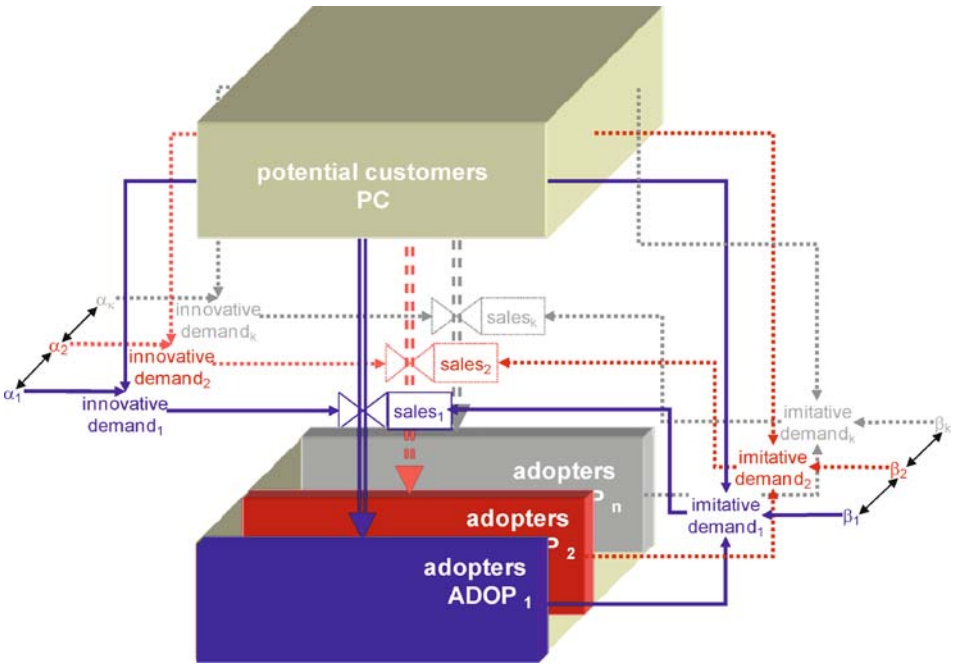

Diffusion of Innovations, System Dynamics Analysis of the, Figure 7
Coarse structure of an oligopolistic innovation diffusion model

α_i and β_i not only depend on the decisions of company i, they also depend on the competitor's decisions. Both variables are crucial for the speed and the maximum volume of demand for the products of a company i.

Figure 8 shows the results of simulations based on Eq. (11) for innovative demand and Eq. (14) for imitative demand with the effects of a market entry delay of the second company – the influences of other decision variables are switched off. Several model simulations have been made assuming a market entry delay of company 2 between 0 and 12 months.

The plots in Fig. 8 show the development of market share and sales of the second company over time. Since there is no further product differentiation, both competitors have the same market share when they enter the market at the same time. With each month delay of the second company the market share that can be achieved at the end of the simulation decreases. A three months delay reduces the finally achieved market share to 40%; a 12-month delay even causes a decrease in market share down to approximately 25%. Accordingly, the maximum sales volume decreases significantly with each month delay in market entry time.

Representing Network Externalities

In the following, we will investigate the diffusion of a specific type of goods in order to show the importance of understanding the diffusion of goods with network effects (based on [22]). The trend towards an information society has stressed the relevance of goods satisfying information and communication needs. Many products of this market segment such as electronic mail contain attributes that necessitate a specific examination, since the diffusion of goods showing network effects differs from that of conventional products. The main difference between conventional products and products with network effects is that the utility of the latter cannot be regarded as a constant value. With regard to these products, utility is an endogenous variable which results in a specific diffusion behavior. Two effects are responsible for this particular behavior: the bandwagon effect and the penguin effect. A refined system dynamics model supports a better understanding of this special diffusion process.

The fact that the utility is not constant can be reasoned by a concept commonly referred to as "network effect". A product is characterized by a network effect, if the utility of that product is a function of the installed base, which is defined as the amount of users in a network system. The utility increases with the installed base. This leads to a virtual interdependency of the users inside the network. The interdependency is based on the bandwagon effect and the penguin effect. Starting from the fundamental diffusion model of Bass the characteristics of network effects are integrated into the model in order to simulate the diffusion behavior.

Many definitions for network effects (sometimes also called "positive demand externalities") can be found in the literature. Basically, it can be stated that network externalities exist if the utility of a product for a customer depends on the number of other customers who have also bought and use this product. These network externalities can be indirect or direct, whereby we concentrate on the latter. A typical example for a product with direct network effects is the telephone. The utility that a telephone system can generate increases with the amount of feasible communication connections. Other examples are e-mail, fax, instant messaging, etc. each of which satisfies communication needs. But none of these products can generate any utility for a user on its own. In order to create utility the existence of other users adopting this product is required. Accordingly, the product's utility changes dynamically with the number of users, i. e., the installed base.

The installed base B_t determines the utility of products influenced by network externalities. In terms of direct network externalities the utility is based on B_t exclusively since utility can only be generated by interconnections within the underlying network solely. Accordingly, the utility of a product with direct network externalities is a function of the number of feasible connections I_t. The number of connections is determined by the number of users on the one hand and the technological restriction of the network on the other hand, whereby the latter one represents the number of users being able to communicate via the network simultaneously (for instance, classical telephone system $r = 2$, telephone conferencing $r \leq 2$). Thus, U_t can be calculated by the formula:

$$U_t = U_t\left(I_t\right) = \sum_{k=2}^{n} \binom{B_t}{r} = \sum_{k=2}^{n} \frac{B_t!}{r!\left(B_t - r\right)!}\,, \quad \text{whereby } r,\ B_t > 0\,. \tag{15}$$

Since the achievable utility of a product with direct network externalities depends exclusively on the network size the adoption process depends on the decision of potential users influencing the diffusion process significantly. This leads to two different effects: Firstly, the utility for each user grows exponentially with an increasing amount of actual users according to the formula. This implies that the more people are attracted the more are part of the network leading to an exponential growth of the diffusion process which is referred to as the "bandwagon effect": The higher

the number of people the more are decoyed as well resulting in a reinforcing process. Although this effect occurs with conventional products as well, in case of products with direct network externalities, it is much stronger since the exponentially growing utility has a greater impact on the diffusion process.

Secondly, utility must be created by the utilization of the users first in order to establish a new communication network which determines the diffusion process significantly since products influenced by direct network externalities cannot generate an original utility by itself. Accordingly, early adopters of a network are confronted with the risk that they cannot derive sufficient benefit from the network so that they must rely on potential users to follow entering the network in the future. Therefore, the adoption decision of a potential user depends on the future decision of other potential users. All in all, this leads to a hesitating behavior of all potential adopters resulting in an imaginary network barrier, which is based on the risk of backing the wrong horse, which is also known as the "penguin effect".

Finally, another important aspect must be considered when analyzing the diffusion process of products with network externalities. In terms of conventional products the decision to buy a product is the final element of the decision process. Contrary to that, concerning products with network externalities the adoption process is not finished with the decision to enter a network since the subsequent utilization of the product is important for the diffusion process. If the expected utility of the communication product cannot be achieved users may stop using this product leading to a smaller installed base and a lower network utility which is important for other users that may stop their utilization as well and for the adoption process of potential users.

In the following, the basic structure of the underlying model will be described. Analogously to the model presented in the preceding paragraphs there exists a group of potential users (in this model, we only focus on the core diffusion process without considering competitors or latent demand in order to keep the complexity of the model low.) If these potential users decide to adopt the communication product, they become part of the installed base *B*. The adoption process is illustrated by the adoption rate *AR*, which is primarily influenced by the variable *word of mouth*. In order to consider the average utility per user – as it is necessary for analyzing products with network externalities – the imitation coefficient β has been endogenized contrary to the classical Bass model. Therefore, the variable β is influenced by the "word-of-mouth" effect which depends on the average utility per user. If actual utility is bigger than the desired utility all individuals in contact with users adopt and buy the product. If it is smaller, however, only a fraction adopts. The size of this fraction depends on the distance between actual and desired utility.

Figure 9 depicts two simulation runs showing the system behavior of the diffusion process of conventional

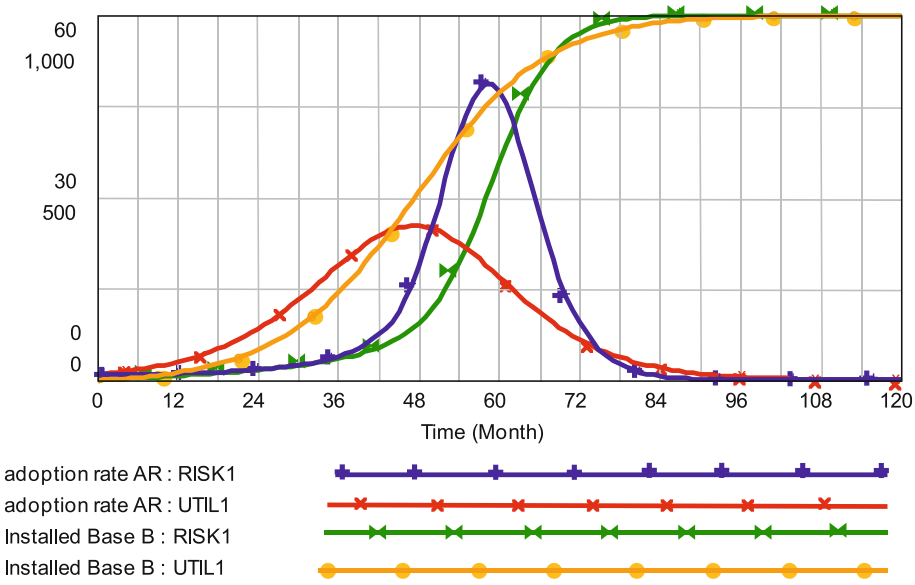

Diffusion of Innovations, System Dynamics Analysis of the, Figure 9
Comparison of diffusion behavior

products and products influenced by direct network externalities. Graph UTIL1 represents the *adoption rate*, i. e., the amount of buyers of a conventional product per period. The graph UTIL1 shows the behavior of the variable *installed base B* which is the accumulation of *adoption rate* (note that the graphs have a different scale). The graphs RISK1 show the system behavior for products influenced by direct network externalities, i. e., the *adoption rate AR* and the corresponding *installed base B*.

A comparison of both simulation runs shows that diffusion needs longer to take off in terms of products influenced by direct network externalities, but showing a steeper proceeding of *Installed Base B* in later periods. This behavior can be verified comparing the adoption rates of the two runs: although adoption starts later with an endogenously generated adoption fraction, it nevertheless has a higher amplitude. This behavior can be interpreted as the penguin effect and the bandwagon effect.

Finally, it has to be taken into account that some users of the installed base might quit to use the product since they are disappointed from its utility. Accordingly, it is an important issue to find ways in order to raise the patience of users to stay within the network. That gives potential users the chance to follow into the network which will increase the utility for the user as well.

From the simulation analysis the following conclusions can be drawn. The importance of the installed base for a success diffusion process is shown. Without a sufficient amount of users it is not possible to generate a utility on a satisfying level which prevents potential users to enter the network or even making users leave the network. Accordingly, ways must be found to increase the utility that a network creates for a user in order to reach the critical mass. This can be done in several ways of which some will be discussed briefly. One possible way is to increase the installed base by compatibility to other networks. Furthermore, the risk to back the wrong horse can be mitigated by product pre-announcements in order to lower the imaginary network barrier by making potential users familiar with the product. Another possibility is to increase the group of relevant users, i. e., to enlarge the average group size within the network, since not all users are equally important for a potential user. Furthermore, the technological potential can be improved by introducing multilateral interconnections between the members of a network.

Representing Managerial Decision Making in Innovation Diffusion Models

Subsequently the basic structures of innovation diffusion processes described above will be extended and simulated to demonstrate the impact of managerial decision-making on the diffusion of innovations. The model used for the simulations serves as a simulator to determine how individual strategies can accelerate or hamper market penetration and profit performance. The models are not designed to predict the basic market success or failure of innovations. Although, they are rather comprehensive, several assumptions apply here as for all models. E.g., in all model runs, the basic market acceptance of the innovation is assumed. Furthermore, the simulations assume for the moment that no competition exists.

Dynamic Pricing Without Direct Competition

In a first step the basic model from Subsect. "Base Structure of a System Dynamics-Based Model of Innovation Diffusion" is extended to generate dynamic cost behavior as suggested in Fig. 5. Standard costs are the basis for the calculation of prices – an important decision variable. Experience curve effects are modeled based on cumulated production in order to map the long-term behavior of standard cost. The actual costs of a product in a certain period are derived from the standard cost modified for variations resulting from capacity utilization.

The concept of experience curve effects suggests a direct relationship between cumulated production $X_{(t)}$ and average standard cost per unit $c^s_{(t)}$, adjusted for inflation; where c^s defines standard unit cost at the planned level of production. Every doubling of $X_{(t)}$ is associated with a cost reduction in real terms by a constant percentage according to:

$$c^s_{(t)} = c^n \left(\frac{X_{(t)}}{n} \right)^{-\delta} \tag{16}$$

where c^n stands for the cost of unit n ($n \subseteq X$) and δ represents a constant depending on the experience rate. For many businesses experience rates of 10% to 20% have been observed and ample empirical evidence for this relationship is available.

The costs of a product in each period of time $C_{(t)}$ are a function of cumulated production $X_{(t)}$ and capacity utilization determined by the production volume of a period $x_{(t)}$ as defined in Eq. (17). Figure 10 shows the behavior of the dynamic cost function

$$C_{(t)} = \Phi \left(X_{(t)}, x_{(t)} \right) \ . \tag{17}$$

Furthermore, the model comprises elements of (i) market development, (ii) product pricing and its impact on the profits from producing and selling the products, i. e., the operating results, and (iii) resource allocation, e. g., capital investment, production volume, and quality control.

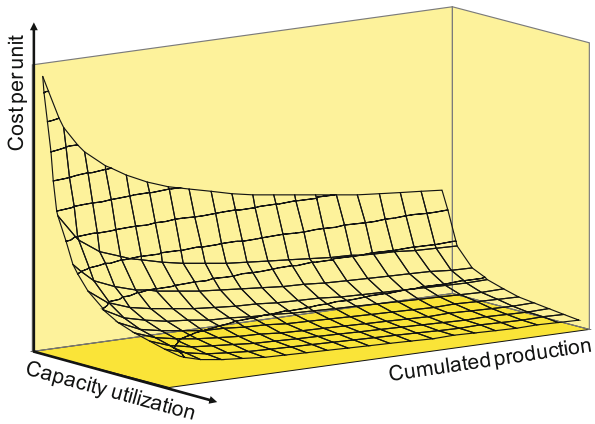

Diffusion of Innovations, System Dynamics Analysis of the, Figure 10
Dynamic cost function

Pricing and quality affects the coefficients of innovation α and imitation β from Eq. (10). Figure 11 shows the run of a model version including market development.

The time behavior of production, demand, and operating results duplicate usual characteristics of the life cycle of a successful innovation. After the product launch, additional customers can be gained from an untapped market as diffusion and thereby product awareness proceeds and prices decline. The maximum of demand from imitators – the quantitatively most important fraction of demand – is reached when the amount of possible communications between potential customers and adopters reaches its maximum. The decreasing level of potential customers and the depletion of the untapped market cause the decline towards the end of the simulation. The behavior also shows that demand rises much faster than the company can increase its production capacity. The behavior of Fig. 11 will serve as reference mode for further analysis.

Pricing strategies and decisions are additional important elements, which require an extension of the model. The problem of the "right price" for a new product is essential but still unsolved in the area of innovation management. Difficulties to recommend the optimal pricing policy derive in particular from the dynamics in demand interrelations, cost development, potential competition, and the risk of substitution through more advanced products. Regardless of this complex framework, several attempts in management science try to derive and to apply optimal pricing policies. However, they are faced with difficulties, both mathematical and practical. Their results are too complicated to support actual pricing decisions. Therefore simulation studies found more frequently their way into management science.

The extended model includes four predefined pricing policies to investigate their impact on market development on operating results:

Myopic profit maximization assuming perfect information about cost and demand. The optimal price p^{opt} is de-

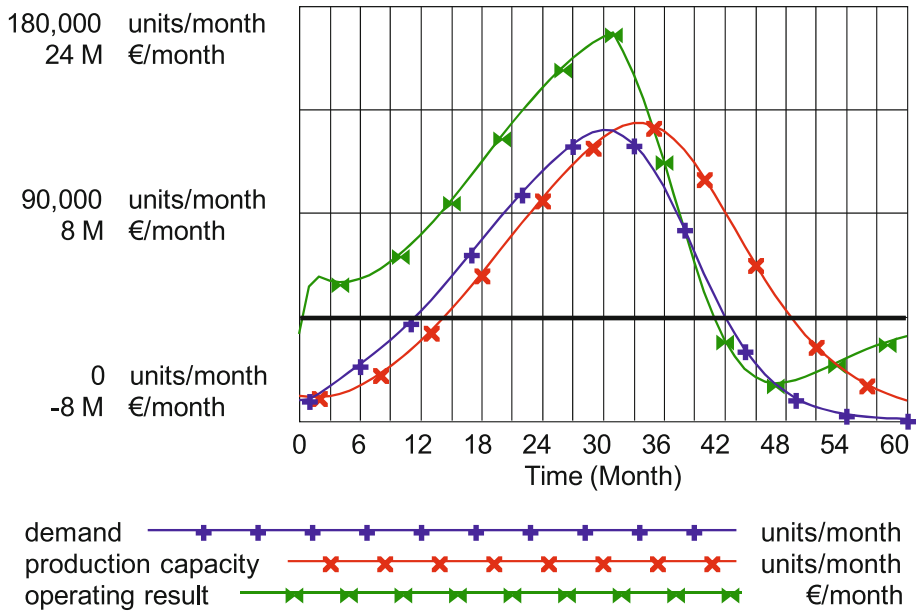

Diffusion of Innovations, System Dynamics Analysis of the, Figure 11
Reference mode of the basic innovation diffusion model

rived from elasticity of demand $\varepsilon_{(t)}$ and per unit standard cost $c^{s}_{(t)}$ considering the impact of short term capacity utilization:

$$p^{\text{opt}}_{(t)} = c^{s}_{(t)} \cdot \frac{\varepsilon_t}{\varepsilon_t - 1} \,. \tag{18}$$

Skimming price strategy aims at serving innovative customers with high reservation prices and then subsequently reduces prices. The model applies a simple decision rule modifying $p^{\text{opt}}_{(t)}$ through an exponential function that raises the price during the first periods after market introduction:

$$p^{\text{skim}}_{(t)} = p^{\text{opt}}_{(t)} \cdot \left(1 + a \cdot \mathrm{e}^{\frac{-t}{T}}\right) \,. \tag{19}$$

Full cost coverage, i. e., standard cost per unit plus a profit margin π to assure prices above cost level even during the early stages of the life cycle:

$$p^{\text{fcc}}_{(t)} = c^{s}_{(t)} \cdot \pi \,. \tag{20}$$

Penetration pricing aims at rapidly reaching high production volumes to benefit from the experience curve and to increase the number of adopters. It uses a similar policy as for the skimming price, but instead of a surcharge it decreases prices early after market introduction:

$$p^{\text{pen}}_{(t)} = c^{s}_{(t)} \cdot \pi \cdot \left(1 - a \cdot \mathrm{e}^{\frac{-t}{T}}\right) \,. \tag{21}$$

The simulation runs shown in Fig. 12 give an overview of the development of profits, cumulated profits, and sales for the four pricing strategies discussed above. The model assumes the following: (1) there is an inflow from the un-

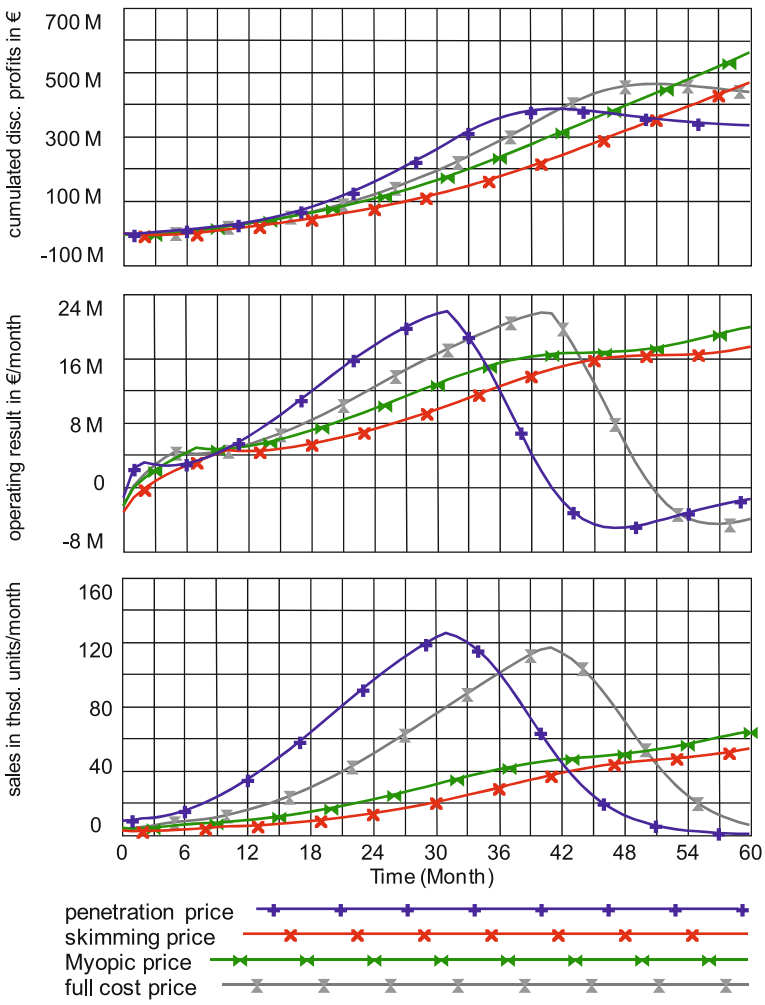

Diffusion of Innovations, System Dynamics Analysis of the, Figure 12
Comparison of the outcome of pricing strategies

tapped market, which depends on the dynamic development of prices; (2) there is no risk of competition; (3) repeat purchases do not occur. Taking profits into account, Fig. 12 indicates that – over the time horizon observed –, the classic pricing rule of profit optimization leads to superior results from a financial point of view. However, if judged by the market development, the strategy of penetration prices is the appropriate strategy. This strategy allows rapid penetration of the market by setting relatively low prices, especially in the early stages of the life cycle. The combined price and diffusion effects stimulate demand and reduce the risk of losing potential customers to upcoming substitution products.

Figure 12 also indicates a disadvantage of the penetration strategy. Since the market is already completely satisfied after period 54, there is only little time to develop and introduce a new product in the market successfully. The slower market growth of the skimming and optimum price strategy leaves more time for the development of a new product, but the attractive profit situation and the slow development also increase the risk that competitors might enter the market. In a dynamic demand situation where prices influence market growth, where substitution or competition can occur, and where delivery delays eventually accelerate the decision of potential buyers to turn to other products, a strategy of rapid market penetration seems to be the most promising one. It will, therefore, be the basis for the following simulation runs investigating manufacturing's role in innovation management.

Linking Manufacturing-Related Decision Variables

The role of manufacturing is important for the successful management of innovations. Manufacturing has to provide sufficient capacity to produce the goods sold. The investments to adjust capacity influence a company's ability to meet demand and deliver on time. It is assumed that the necessary financial resources for the investments are available. The aggregated capacity provided by the company includes both, machinery equipment and production personnel. Since the manufacturing function also has to ensure the quality of the output through dedicating a portion of its total available capacity to quality control, the capacity resources can be used to either manufacture the products or to assure the desired level of quality. Capacity allocation to improve quality takes away capacity for production. This additional feedback structure – as indicated in Fig. 5 – maps the allocation of resources for quality control to the achieved ability to meet product demand. If manufacturing capacity does not meet demand, a temporary reduction of capacity for quality assurance seems a plausible strategy. Quality control resources than are allocated to manufacturing rather than testing whether quality standards are met. In this scenario, it would be expected that total cost remain unchanged and the additional manufacturing capacity gained through the reallocation can be used to provide more products to the customers, increase sales, and improve the overall results.

Figure 13 shows the simulation assuming the same scenario as in the base mode together with penetration prices and reduced quality resources if demand exceeds production capacity. It also shows a quality index plotted as an additional variable. Quality is defined to be 1, if the actual quality capacity equals a standard value of quality resources necessary. It is assumed that 10% of total production capacity is necessary to assure 100% quality. For values above the 10%-level, quality is better; for values be-

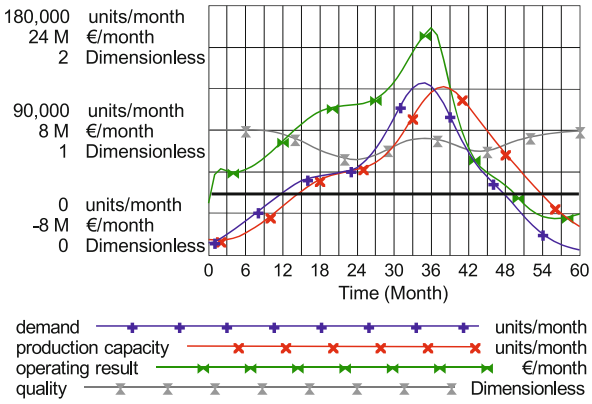

Diffusion of Innovations, System Dynamics Analysis of the, Figure 13
Reduced quality control

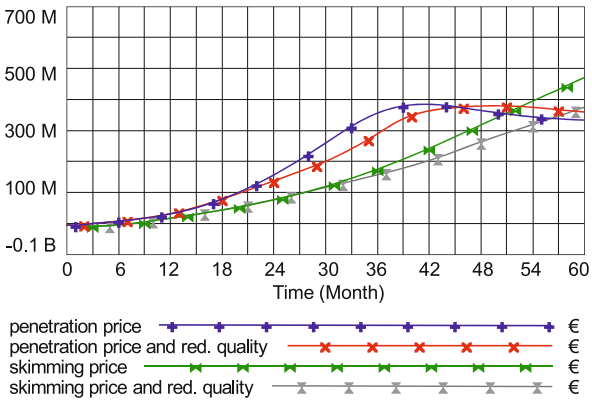

Diffusion of Innovations, System Dynamics Analysis of the, Figure 14
Cumulated discounted profits – penetration vs. skimming pricing in combination with quality control strategies

low, it is poorer. The simulation indicates that the policy of reduced quality resources successfully decreases the discrepancy between demand and production as seen in the reference mode of Fig. 11. This results from the increased proportion of capacity used for production and an additional effect caused by lower product quality, which then decreases demand. Although the maximum sales are nearly the same in the simulation of reduced quality control strategy, the peek demand occurs around 5 months later. Instead of gaining higher sales only the shape of the life cycle changed. However, operating results had improved, in particular the sharp decline of profits in the base mode of the simulation could be slowed down and losses could be avoided. The reduced quality control strategy caused a slower capacity build-up and therefore, when product sales declined capacity adjustment was easier to achieve. From the financial point of view the strategies of penetration prices and reduced quality control fit quiet well.

The results are different if a strategy of quality reduction is used in combination with a strategy of skimming pricing. Figure 14 compares the outcome of cumulated discounted profits for the strategy of reduced quality and penetration prices or skimming prices with the development of the reference mode – the simulations without quality adjustment. The behavior indicates that in the case

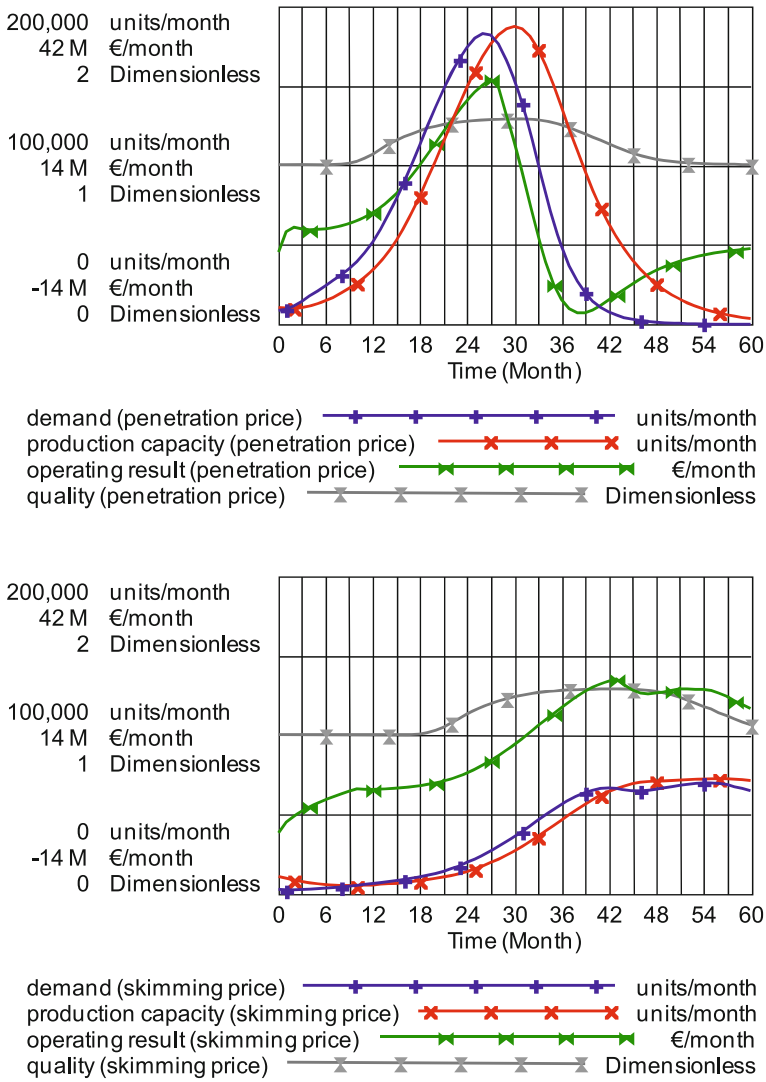

Diffusion of Innovations, System Dynamics Analysis of the, Figure 15
Emphasized quality in all innovation stages

of skimming prices, quality reductions slow down the development of the market and cumulated profits significantly.

The simulation results raise the question whether emphasizing quality when demand is higher than capacity would be a more appropriate way to react. As the upper part of Fig. 15 points out, the strategy of emphasized quality leads to an accelerated product life cycle in the case of the penetration pricing strategy. Tremendous capacity build-up is necessary after the introduction of the new product. As demand declines, a plenty of capacity is idle, causing significant losses during the downswing of the product life cycle.

Emphasizing quality turns out to be more effective in the case of skimming prices. The additional demand gained from quality improvements also accelerates the product life cycle, but at a much slower rate and leads to improved cumulated profits. Emphasizing quality in combination with skimming or optimum prices leads to improved cumulated profits, compared to both, the simulation without quality reaction and the quality reduction run.

The simulations show the importance of a detailed judgment of strategic choices. Strategies must be consistent with each other and with the real world structures mapped by the model. The simulations above assume a sit-

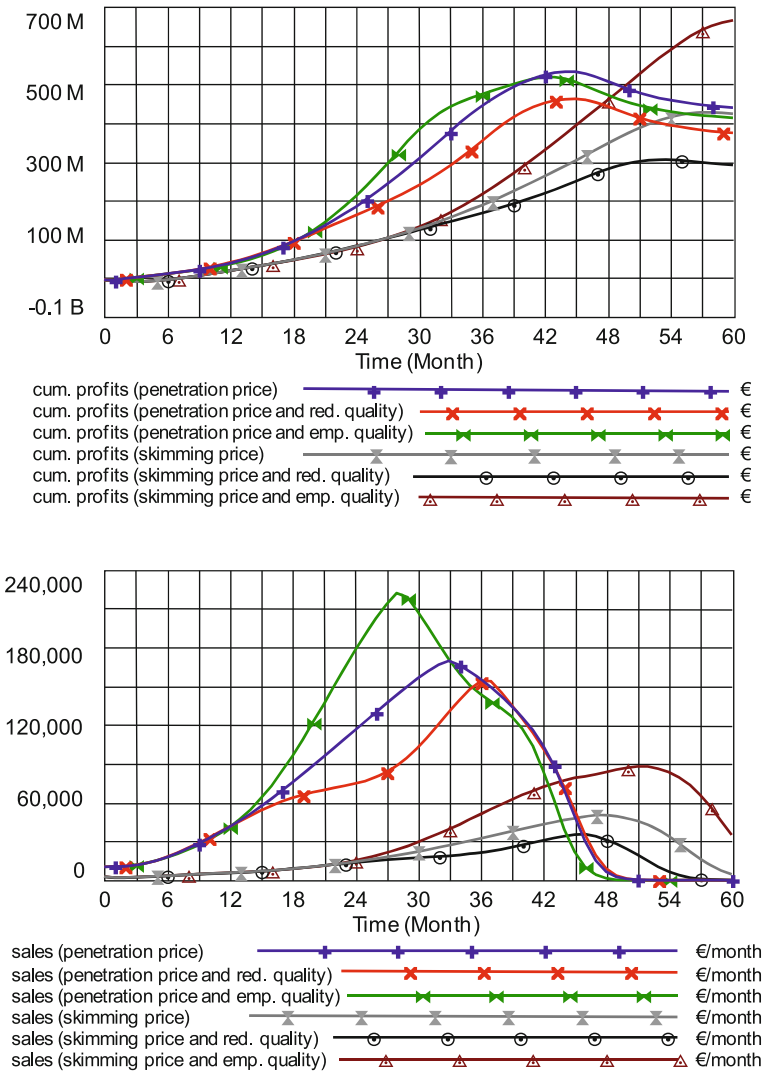

Diffusion of Innovations, System Dynamics Analysis of the, Figure 16
Behavior of the base model including simple competitive structures

uation without existing or potential competition. In such an environment there is no interest paid for fast market penetration. Hence, a penetration pricing strategy is the most unfavorable alternative. However, this changes if structural elements are added to the model that incorporate competition – even as in the simple structure from Fig. 5, which considers the loss of demand to a competitor. Lost demand therefore is represented as a process equivalent to the imitative demand from Eq. (9). The calculation of lost demand starts in period 15 through an initial switching of a potential customer to the competitor. This switch starts a process that drives demand for the competitors' products and is influenced through the quality the company offers. If the company provides poor quality, more potential customers and market potential from the untapped market will directly move to the competitor. The accumulation of lost demand corresponds to the number of adopters the competitors gained over time. Simulations with these additional structures give some additional insights (Fig. 16).

Penetration pricing leads again to the fastest market development. In the competitive surrounding, however, emphasizing quality accelerates the market development and leads to better performance than quality reductions. This is in contrast to the simulations without competition shown in Fig. 13 to Fig. 15. Skimming prices in combination with reduced quality control shows the poorest financial and market performance. A strategy of reduced quality control causes in the competitive environment the demand to increase at a slower rate than in the base run, where no quality adjustments were made when demand exceeded capacity. In both cases, the skimming and the penetration price scenario, quality reductions lead to the poorest performance.

Linking R&D and New Product Development

The models discussed above are able to generate under different conditions the typical diffusion patterns of new products in the market place. However, these models do not consider the stage of new product development. New products have to be developed before they can be introduced into the market. A costly, lengthy, and risky period of R&D has to be passed successfully. The diverging trends of shortening product life cycles and increasing R&D costs show the importance of an integrated view of all innovation stages. In the remainder, a comprehensive model comprising both, the process of R&D and an oligopolistic innovation diffusion with subsequent product generations is used to investigate the interrelations between the stages of innovation processes. The integration of both modules is shown in Fig. 17.

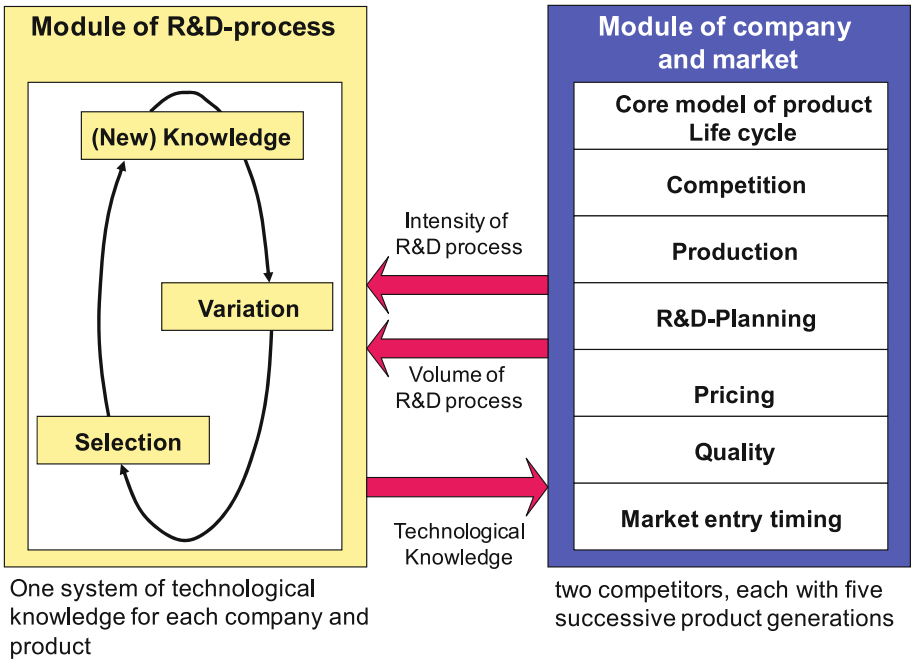

Diffusion of Innovations, System Dynamics Analysis of the, Figure 17
Linking R&D-processes with corporate and market structures

The volume and the intensity of the research and development activities feed the R&D-process. The number of research personnel determines the volume. Since R&D personnel requires resources like laboratory equipment, material for experiments etc., the intensity of R&D depends on the budget available for each person working in the R&D sector. This information is calculated in a more comprehensive model in the sector of R&D planning, which also includes policies about resource allocation within the research and development stages, i. e., mainly the question of how much to spend on which new product development project.

Depending on the volume and the intensity of R&D, the technological knowledge of each product generation for each company evolves over time. The module of the R&D-process feeds back the current state of the technological knowledge for each company and product generation.

The basic assumptions of the model are as follows. The model maps the structures of two competitors. Both competitors can introduce up to five successive product generations. The initial values of the model ensure that all competitors start from the same point. All firms have already introduced the first product generation and share the market equally. The resources generated by the first product are used to develop subsequent product generations. In the base run each company follows the same set of strategies. Therefore, except for minor differences resulting from the

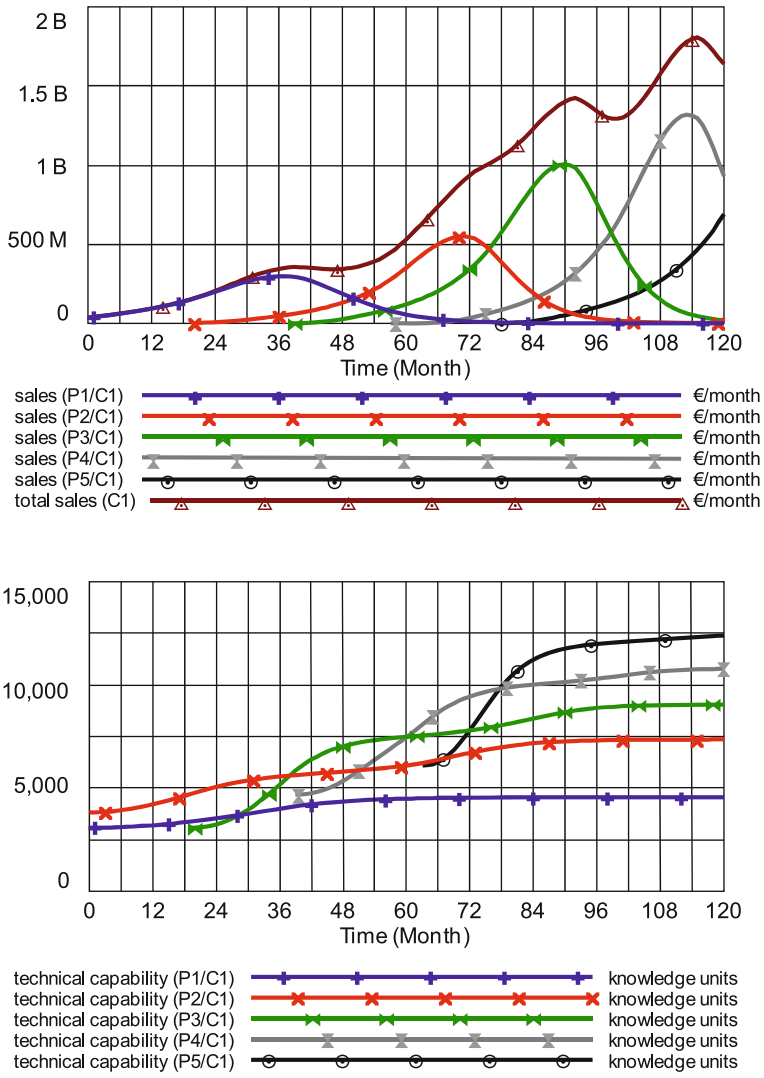

Diffusion of Innovations, System Dynamics Analysis of the, Figure 18
Exemplary behavior of the integrated innovation model

stochastic nature of the R&D-process, they show the same behavior over time. Figure 18 provides a simulation run of the model with all modules and sectors coupled.

The curves show for a single company the development of the sales of the products and the total sales. They emphasize the importance of a steady flow of new and improved products. Without on-time replacement of older products, the total sales of the products will flatten or deteriorate like in the simulation around periods 44, 92, and 116. The model also generates the typical s-shaped curves of technological development (lower part of Fig. 18). Each product generation has a higher technological potential and the knowledge developed for the preceding product generations partly can be used by the successive product generations. For this reason the subsequent product generations start at a level different from zero.

In a dynamic environment such as the computer industry, where investments in R&D and manufacturing equipment are high, the product life cycles are short, and time-to-market as well as time-to-volume are essential variables, it is important to understand the dynamic consequences of decisions and strategies in the different areas. Figure 19 describes some of the important feedback loops linking the process of invention to the processes of innovation and diffusion.

Central element in the figure is the calculation of the sales of a company according to Eqs. (11) and (14). The coefficients of innovation and imitation are influenced by the multiplier of relative competitive advantage, which depends on the relative technical capability and the price advantage of a company. The technical capability of the products is influenced by the strength of its R&D-processes and the total amount of R&D expenditures. Empirical studies in Germany have shown that measures like sales volume, profits or R&D budgets of earlier periods are quite common as a basis for R&D budgeting. However, using historic sales volume as a basis to determine R&D budgets invokes the positive feedback loop "competing by technical capability". With an increasing number of products sold and growing value of sales the budget and the number of personnel for R&D grow. This leads to an improved competitive position, if the technical capabilities of a product increases. The higher the sales volume, the better is the resulting competitive position. This produces increasing coefficients of innovation and imitation and leads to higher sales. This budgeting strategy is implemented in the model for the next simulation runs.

The second loop "price competition" links pricing strategies to sales volume. The actual price of a product is influenced by three factors. The first factor, standard

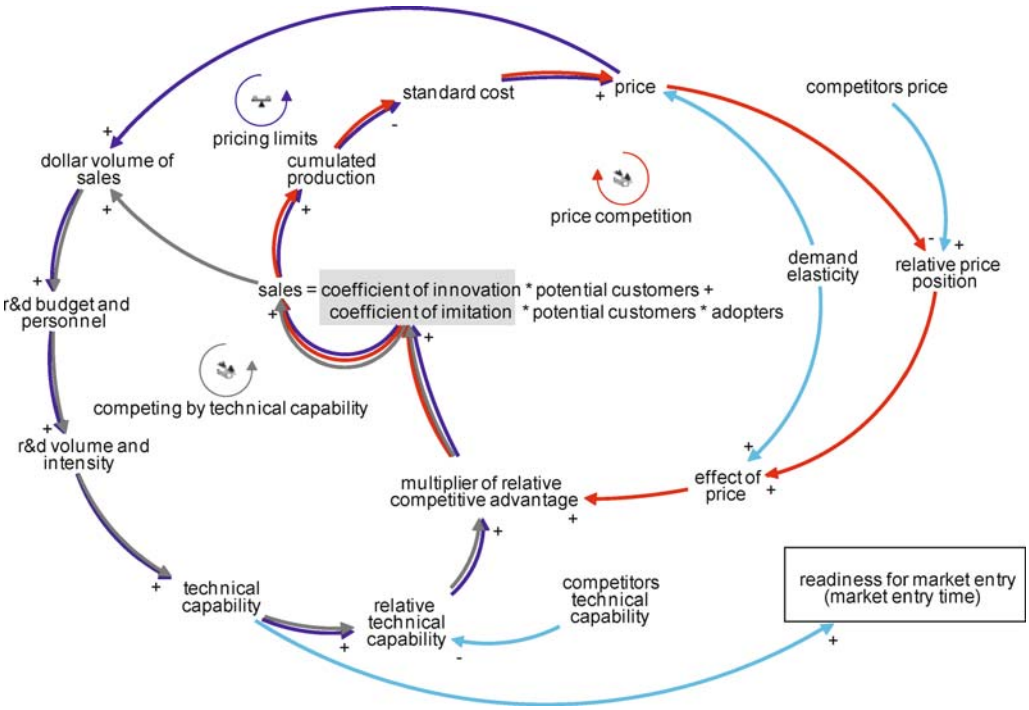

Diffusion of Innovations, System Dynamics Analysis of the, Figure 19
Feedback structure influencing the diffusion process

costs, is endogenous. As cumulated production increases, the experience gained from manufacturing causes declining standard costs. The second and third elements influencing the calculation of prices are exogenous: parameters which define the pricing strategy and demand elasticity. Caused by increasing cumulated production, standard costs fall over the life cycle and prices are also declining. Lower prices affect the relative price and improve the effect of price on the coefficients of innovation and imitation, which leads to increased sales and higher cumulated production.

The loop "pricing limits" reduces the effects of the reinforcing loops described above to some extent. The standard cost and price reductions induce – ceteris paribus – a decrease in the sales volume and set off all the consequences on the R&D-process, the technical know-how, the market entry time and sales shown in the first feedback loop – but in the opposite direction. Additionally, since standard cost cannot be reduced endlessly this feedback loop will show a goal seeking behavior.

With equivalent initial situations and the same set of strategies, both companies behave in an identical way for all product generations. If one company has a competitive advantage, the reinforcing feedback loops suggest that this company will achieve a dominating position. In the simulation shown in Fig. 20, both competitors have the same competitive position for the first product generation. But the first company will be able to enter the market 2 months

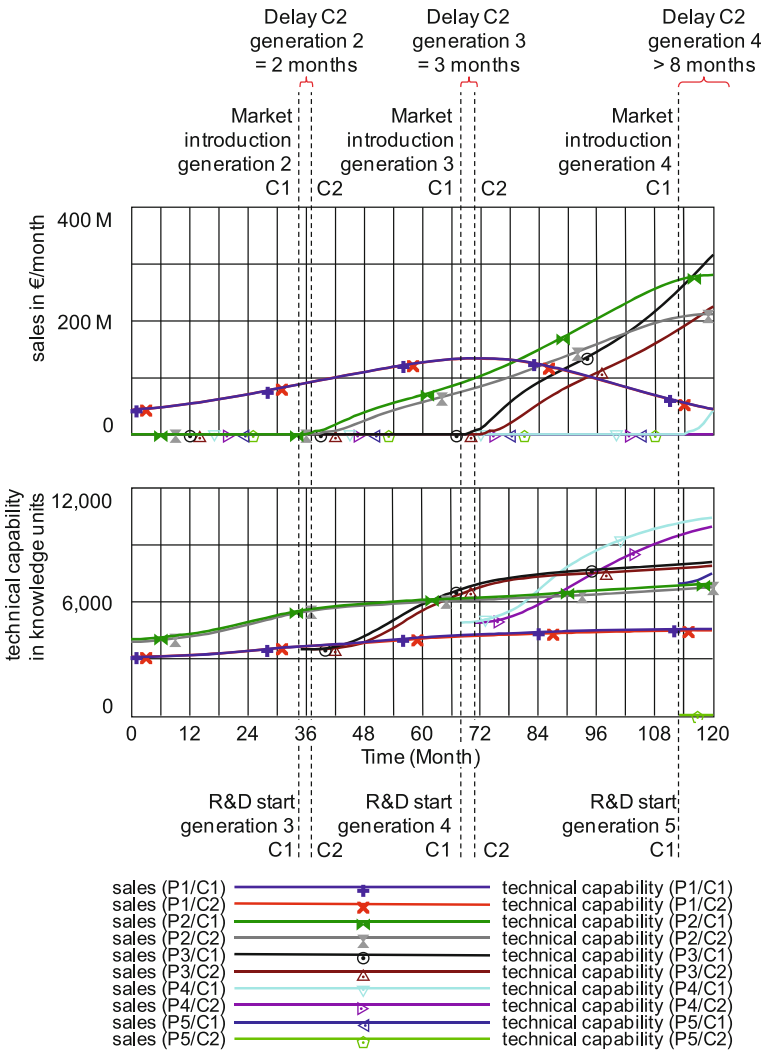

Diffusion of Innovations, System Dynamics Analysis of the, Figure 20
Reinforcing effects of initial competitive advantage

earlier than the competitor, because the initial outcome of the R&D process is slightly better than the second company's second product generation. Both competitors follow a strategy of skimming prices and demand elasticity has the value −2.

The initial gain in the outcome of the R&D-process initiates a process of sustained and continuing competitive advantage for the first company. It will improve continuously, since the positive feedback loop "competing by technical capability" dominates. The first company's advantage in the market introduction leads to an increasing readiness for market entry. It is able to launch the third product generation 3 months earlier than the follower and will introduce the fourth product generation in period 112. The follower is not able to introduce its fourth generation during the time horizon of the simulation, i. e., the pioneers advantage has extended to more than 8 months. The first company's competitive advantage is a result of the slightly higher initialization of the knowledge system and the dominance of the positive feedback loops, which causes shortened time-to-market and higher sales volume over all successive product life cycles. Additionally, the technical capabilities of both competitors' product generations show the same reinforcing effect. The difference

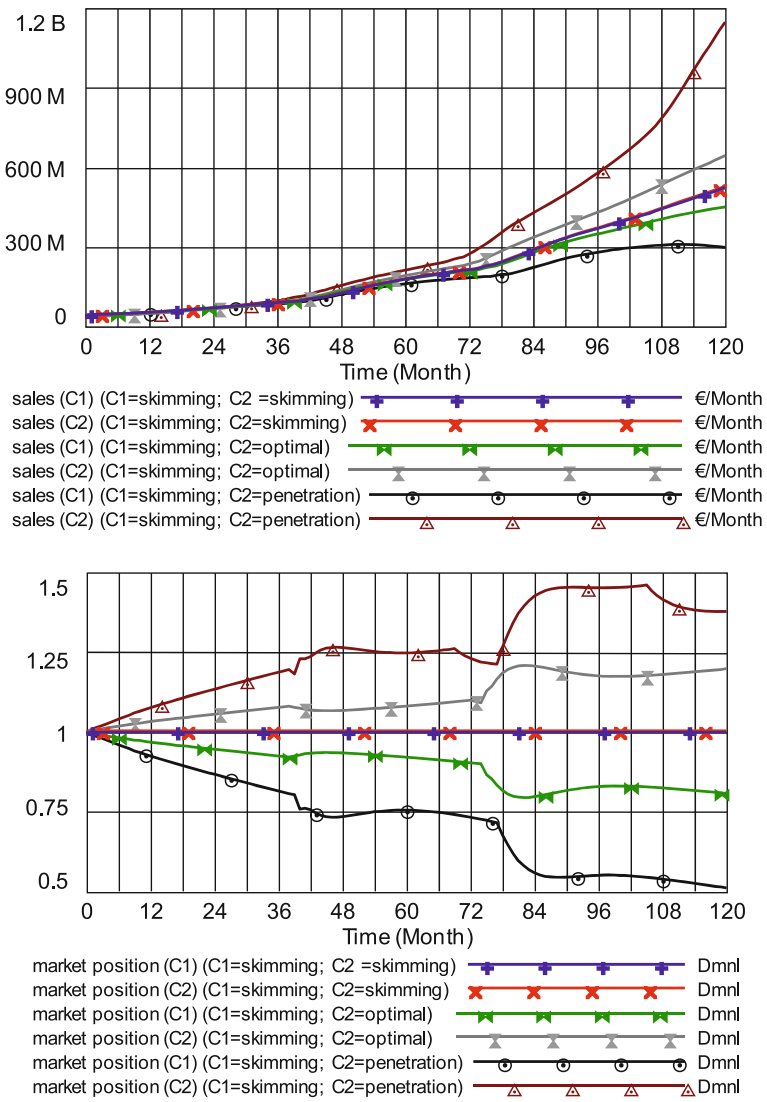

Diffusion of Innovations, System Dynamics Analysis of the, Figure 21
Sales volume and market position for different pricing strategies

between the technical capability of both competitors increases in favor of company 1 until they approach the boundaries of the technology.

Although literature discusses a variety of models to find optimal pricing strategies, these models usually only consider the market stage of a new product and neglect the interactions with the development stage of a new product. Pricing decisions not only drive the diffusion of an innovation, but they also have a strong impact on the resources available for research and development. Since the comprehensive innovation model links the stages of developing and introducing a new product, the following simulations will show the impact of pricing strategies on performance in a competitive environment. In the analysis shown the first company uses the strategy of skimming price for all product generations. The second company alternatively uses a skimming price strategy in the first model run, myopic profit maximization strategy in the second run, and the strategy of penetration prices in the third run. The initial conditions are identical, except the price strategy settings. Sales volume, market position, and cumulated discounted profits are used to judge the advantages of the alternative pricing strategies. Market entry time is used as a measure of time-to-market.

The logic behind the skimming price strategy is to sell new products with high profit margins in the beginning of a life cycle to receive high returns on investment, achieve short pay off periods, and high resources for the R&D-process. However, in a dynamic competitive setting the strategy of myopic profit maximization and penetration prices achieve better results (Fig. 21). Company 1 which uses a skimming price strategy achieves the lowest sales volume. Myopic profit maximization prices and penetration prices of the second competitor causes the sales to increase stronger through the combined price and diffusion effect.

The results are confirmed if the variable market position – an aggregate of the market share a company has for its different products – is taken into account. For values greater than 1 the market position is better than the one of the competitor. Using the penetration strategy, company 2 can improve its market share, achieve higher sales vol-

Diffusion of Innovations, System Dynamics Analysis of the, Table 1
Consequences of pricing strategies on market entry time

	Product generation 2			Product generation 3			Product generation 4		
Pricing strategy C2*	C1	C2	Delay C1 to C2	C1	C2	Delay C1 to C2	C1	C2	Delay C1 to C2
Skimming prices	38	38	0	71	71	0	n.i.	n.i.	–
Profit maximization	36	35	1	71	69	2	n.i.	118	> 2
Penetration prices	37	35	2	74	66	8	n.i.	102	> 8

*C1 uses skimming prices in all simulations; **n.i. = product was not introduced

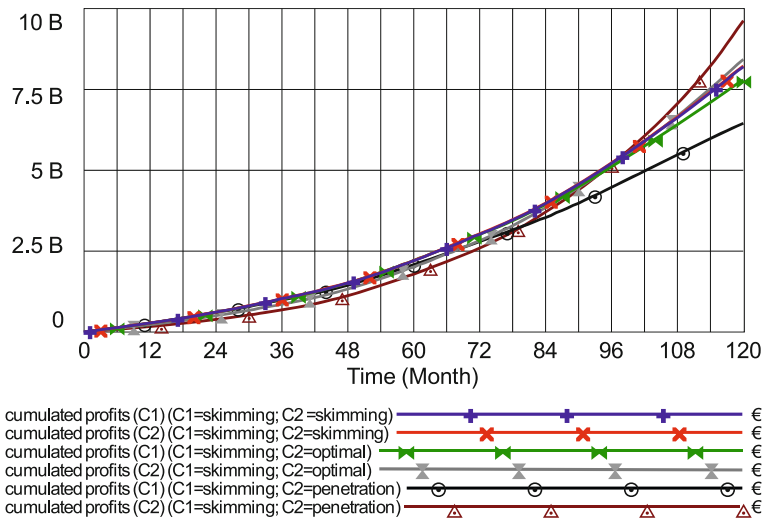

Diffusion of Innovations, System Dynamics Analysis of the, Figure 22
Time path of cumulated profits

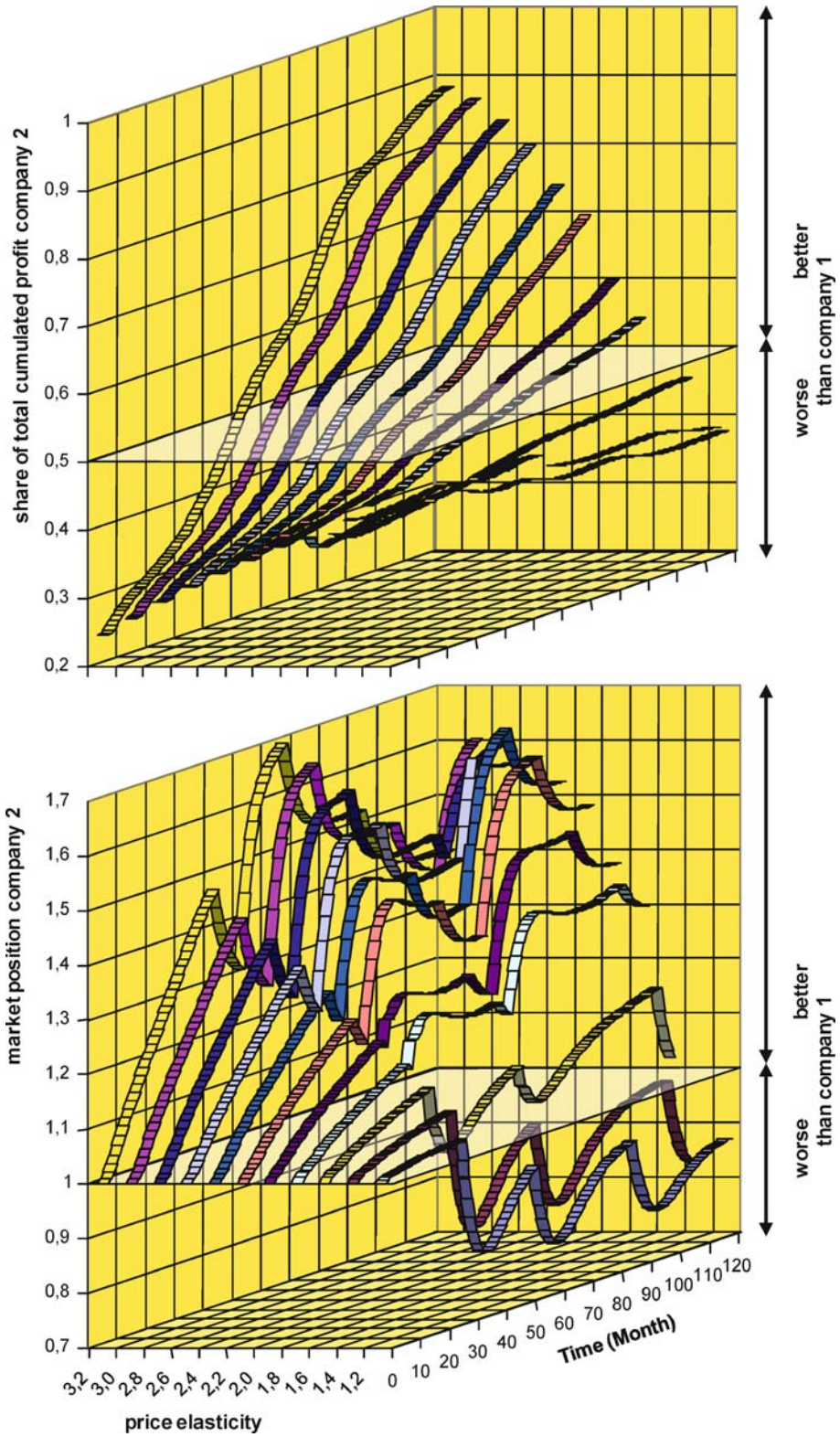

Diffusion of Innovations, System Dynamics Analysis of the, Figure 23
Impact of demand elasticity on performance measures

ume and therefore has more resources available for R&D. This enables it to launch new products earlier than company 1. As shown in Table 1, the advantage of time-to-market increases from product generation to product generation.

The improvement in time-to-market for the first company's second product generation results from the slightly higher sales volume compared to the use of skimming pricing strategies for both competitors. The second company achieves the strongest improvements in time-to-market if it uses a penetration pricing strategy.

In terms of cumulative profits (Fig. 22) one would expect that skimming prices should generate the highest cumulated profits, however, this is not true. Penetration prices generate the highest results followed by skimming prices. The strategy of myopic profit maximization shows the least favorable outcome.

The simulations so far assumed a price response function with a constant price elasticity ε of -2. Since price elasticity influences both, the demand for a product as well as the price level (cf. Fig. 19), the influence of price elasticities have to be investigated before recommendations can be made. Assuming that company 1 uses a strategy of skimming prices and the second competitor follows a strategy of penetration pricing, Fig. 23 shows the time path of cumulated discounted profits and market position for ε between -3.2 and -1.2.

Due to the different profit margins – resulting from myopic profit maximization being the basis for price calculation – the use of the absolute value of the cumulated profits is not appropriate. Therefore, the second company's share of the total cumulated profits is used for evaluation purposes. The measure is calculated as

$$\left(\frac{\text{cum.profits}_2}{\sum_{i=1}^{2}\text{cum.profits}_i}\right).$$

The first graph in Fig. 23 shows that the initial disadvantage of the second company rises with increasing demand elasticity. However, its chance of gaining an advantage increases as well. In the case of lower demand elasticities ($\varepsilon > -1.7$) firm 2 cannot make up the initial disadvantage during the whole simulation. For demand elasticities ($\varepsilon > -1.4$) the cumulated profits ratio even deteriorates. Considering the market position the picture is similar. For demand elasticities $\varepsilon > -1.6$ the penetrations strategy leads to a loss in the market position in the long run. The improvements resulting from the introduction of the successive product generations are only temporary.

Managerial Implications

The simulations above lead to the insight that general recommendations for strategies are not feasible in such complex and dynamic environments. The specific structures like competitive situation, demand elasticity, or strategies followed by the competitors have to be taken into account. Recommendations only can be given in the context of the specific situation. Furthermore, the evaluation of strategies depends on the objectives of a company. If a firm wants to enhance its sales volume or the market share, the strategy of penetration pricing is the superior one. Viewing cumulative profits and the readiness for market entry as prime objectives, the strategy of skimming prices is the best. However, these recommendations hold only for high demand elasticities. Furthermore, the model does not consider price reactions of competitors. The evaluation of improved strategic behavior would become even more difficult. The outcome and the choice of a particular strategy depend on many factors that influence the diffusion process. The dynamics and the complexity of the structures make it almost unfeasible to find optimal solutions. Improvements of the system behavior gained through a better understanding, even if they are incremental, are steps into the right direction.

Future Directions

The series of models presented here are designed in a modular fashion. They offer the flexibility to be adapted to different types of innovations, to different structures, initial conditions and situations. The models provide the opportunity to investigate courses of action in the setting of a management laboratory. They allow one to investigate different strategies and to learn in a virtual reality. They emphasize the process of learning in developing a strategy rather than the final result. To support learning processes, the models could be combined with an easy-to-use interface and serve as a management flight simulator which allows one to gain experience and understanding from playing.

Although the models cover a variety of different aspects in the management of innovations, they still can be extended. Besides more detailed mapping of corporate structures behind managerial decision processes the structures representing the diffusion process can be extended in various ways. Although some research already discusses the problems of mapping the substitution among successive product generations, this area deserves further attention. In particular in high-tech industries with short product life cycles the interrelations between successive product generations strongly influence the overall success of

a company. Furthermore, the diffusion structures could be extended to include cross-buying and up-buying behavior of customers and by that link models of innovation diffusion to the field of customer equity marketing.

Bibliography

Primary Literature

1. Bass FM (1969) A New Product Growth Model for Consumer Durables. Manag Sci 15:215–227
2. Bental B, Spiegel M (1995) Network Competition, Product Quality, and Market Coverage in the presence of network externalities. J Ind Econ 43(2):197–208
3. Boston Consulting Group (1972) Perspectives on Experience. Boston Consulting Group Inc., Boston
4. Brockhoff K (1987) Budgetierungsstrategien füur Forschung und Entwicklung. Z Betriebswirtschaft 75:846–869
5. Brynjolfsson E, Kemerer CF (1996) Network Externalities in Microcomputer Software: An Econometric Analysis of the Spreadsheet Market. Manag Sci 42(12):1627–1647
6. Church J, Gandal N (1993) Complementary network externalities and technological adoption. Int J Ind Organ 11:239–260
7. Farrell J, Saloner G (1986) Installed Base and Compatibility: Innovation, Product Preannouncements, and Predation. Am Econ Review 76(5):940–955
8. Forrester JW (1961) Industrial Dynamics. MIT Press, Cambridge
9. Jeuland AP, Dolan RJ (1982) An Aspect of New Product Planning: Dynamic Pricing. In: Zoltners AA (ed) TIMS Studies in the Management Sciences 18. North Holland, Amsterdam, pp 1–21
10. Katz ML, Shapiro C (1985) Network Externalities, Competition, and Compatibility. Am Econ Review 75(3):424–440
11. Kotler P (1994) Marketing Management. Prentice-Hall, Englewood Cliffs
12. Leibenstein H (1950) Bandwagon, Snob, and Veblen Effects in the Theory of Consumers' Demand. Q J Econ 64(2):183–207
13. Mahajan V, Muller E (1979) Innovation Diffusion and New Product Growth Models in Marketing. J Mark 43:55–68
14. Maier FH (1998) New Product Diffusion Models in Innovation Management – A System Dynamics Perspective. Syst Dyn Review 14:285–308
15. Milling P (1996) Modeling Innovation Processes for Decision Support and Management Simulation. Syst Dyn Review 12(3):221–234
16. Milling P, Maier F (1996) Invention, Innovation und Diffusion. Duncker & Humboldt, Berlin
17. Milling P, Maier F (2004) R&D, Technological Innovations and Diffusion. In: Barlas Y (ed) System Dynamics: Systemic Feedback Modeling for Policy Analysis. In: Encyclopedia of Life Support Systems (EOLSS), Developed under the Auspices of the UNESCO. EOLSS-Publishers, Oxford, p 39; http://www.eolss.net
18. Schmalen H (1989) Das Bass-Modell zur Diffusionsforschung – Darstellung, Kritik und Modifikation. Schmalenbachs Z betriebswirtschaftliche Forsch (zfbf) 41:210–225
19. Senge PM (1994) Microworlds and Learning Laboratories. In: Senge PM et al (eds) The Fifth Discipline Fieldbook. Doubleday, New York, pp 529–531
20. Sterman JD (1992) Teaching Takes Off – Flight Simulators for Management Education. In: OR/MS Today, October 1992. Lionheart, Marietta, pp 40–44
21. Sterman JD (2000) Business Dynamics. Irwin McGraw-Hill, Boston
22. Thun JH, Größler A, Milling P (2000) The Diffusion of Goods Considering Network Externalities – A System Dynamics-Based Approach. In: Davidsen P, Ford DN, Mashayekhi AN (eds) Sustainability in the Third Millennium. Systems Dynamics Society, Albany, pp 204.1–204.14
23. Xie J, Sirbu M (1995) Price Competition and Compatibility in the Presence of Positive Demand Externalities. Manag Sci 41(5):909–926

Books and Reviews

Abernathy JW, Utterback JM (1988) Patterns of Industrial Innovation. In: Burgelman RA, Maidique MA (eds) Strategic Management of Technology and Innovation. Homewood, Irwin, pp 141–148

Bailey NTJ (1957) The Mathematical Theory of Epidemics. Griffin, London

Bass FM (1980) The Relationship between Diffusion Rates, Experience Curves, and Demand Elasticities for Consumer Durable Technological Innovations. J Bus 53:50–67

Bower JL, Hout TM (1988) Fast-Cycle Capability for Competitive Power. Harvard Bus Review 66(6):110–118

Bye P, Chanaron J (1995) Technology Trajectories and Strategies. Int J Technol Manag 10(1):45–66

Clarke DG, Dolan RJ (1984) A Simulation Analysis of Alternative Pricing Strategies for Dynamic Environments. J Bus 57:179–200

Dumaine B (1989) How Managers can Succeed Through Speed. Fortune 4 February 13:30–35

Easingwood C, Mahajan V, Muller E (1983) A Non-Uniform Influence Innovation Diffusion Model of New Product Acceptance. Mark Sci 2:273–295

Fisher JC, Pry RH (1971) A Simple Substitution Model of Technological Change. Technol Forecast Soc Chang 3:75–88

Ford DN, Sterman JD (1998) Dynamic Modeling of Product Development Processes. Syst Dyn Review 14:31–68

Forrester JW (1968) Industrial Dynamics – After the First Decade. Manag Sci 14:389–415

Forrester JW (1981) Innovation and Economic Change. Futures 13(4):323–331

Georgescu-Roegen N (1971) The Entropy Law and the Economic Process. Harvard University Press, Cambridge

Graham AK, Senge PM (1980) A Long Wave Hypothesis of Innovation. Technol Forecast Soc Chang 17:283–311

Homer JB (1983) A Dynamic Model for Analyzing the Emergence of New Medical Technologies. Ph.D. Thesis. MIT Sloan School of Management

Homer JB (1987) A Diffusion Model with Application to Evolving Medical Technologies. Technol Forecast Soc Chang 31(3):197–218

Kern W, Schröder HH (1977) Forschung und Entwicklung in der Unternehmung. Rowohlt Taschenbuch Verlag, Reinbek

Linstone HA, Sahal D (eds) (1976) Technological Substitution. Forecast Techniques Appl. Elsevier, New York

Maier FH (1992) R&D Strategies and the Diffusion of Innovations. In: Vennix JAM (ed) Proceedings of the 1992 International System Dynamics Conference. System Dynamics Society, Utrecht, pp 395–404

Maier FH (1995) Die Integration wissens- und modellbasierter Konzepte zur Entscheidungsunterstützung im Innovationsmanagement. Duncker & Humboldt, Berlin
Maier FH (1995) Innovation Diffusion Models for Decision Support in Strategic Management. In: Shimada T, Saeed K (eds) System Dynamics '95. vol II. System Dynamics Society, Tokyo, pp 656–665
Maier FH (1996) Substitution among Successive Product Generations – An Almost Neglected Problem in Innovation Diffusion Models. In: Richardson GP, Sterman JD (eds) System Dynamics '96. System Dynamics Society, Boston, pp 345–348
Mansfield EJ, Rapoport J, Schnee S, Wagner S, Hamburger M (1981) Research and Innovation in the Modern Corporation: Conclusions. In: Rothberg RR (ed) Corporate Strategy and Product Innovation. Norton, New York, pp 416–427
Meieran ES (1996) Kostensenkung in der Chip-Fertigung. Siemens Z Special FuE Frühjahr:6–10
Milling P (1986) Diffusionstheorie und Innovationsmanagement. In: Zahn E (ed) Technologie- und Innovationsmanagement. Duncker & Humboldt, Berlin, pp 49–70
Milling PM (1986) Decision Support for Marketing New Products. In: Aracil J, Machuca JAD, Karsky M (eds) System Dynamics: On the Move. The System Dynamics Society, Sevilla, Spain, pp 787–793
Milling PM (1987) Manufacturing's Role in Innovation Diffusion and Technological Innovation. In: Proceedings of the 1987 International Conference of the System Dynamics Society. The System Dynamics Society, Shanghai, China, pp 372–382
Milling PM (1989) Production Policies for High Technology Firms. In: Murray-Smith D, Stephenson J, Zobel RN (eds) Proceedings of the 3rd European Simulation Congress. Society for Computer Simulation International, Edinburgh, pp 233–238
Milling PM (1991) An Integrative View of R&D and Innovation Processes. In: Mosekilde E (ed) Modelling and Simulation. Simulation Councils, San Diego, pp 509–514
Milling PM (1991) Quality Management in a Dynamic Environment. In: Geyer F (ed) The Cybernetics of Complex Systems – Self-organization, Evolution, and Social Change. InterSystems Publications, Salinas, pp 125–136
Milling PM, Maier FH (1993) Dynamic Consequences of Pricing Strategies for Research and Development and the Diffusion of Innovations. In: Zepeda E, Machuca JAD (eds) The Role of Strategic Modeling in International Competitiveness – System Dynamics '93. The System Dynamics Society, Cancun, Mexico, pp 358–367
Milling PM, Maier FH (1993) The Impact of Pricing Strategies on Innovation Diffusion and R&D Performance. Syst Dyn Int J Policy Model 6:27–35
Norton JA, Bass FM (1987) A Diffusion Theory Model of Adoption and Substitution for Successive Generations of High-Technology Products. Manag Sci 33:1069–1086
Norton JA, Bass FM (1992) Evolution of Technological Generations: The Law of Capture. Sloan Manag Review 33:66–77
Paich M, Sterman JD (1993) Boom, Bust, and Failure to Learn in Experimental Markets. Manag Sci 39:1439–1458
Pearl R (1924) Studies in Human Biology. Williams and Wilkins, Baltimore
Robinson B, Lakhani C (1975) Dynamic Price Models for New Product Planning. Manag Sci 21:1113–1122
Roberts EB (1964) The Dynamics of Research and Development. Harper and Row Publishers, New York Evanston London
Roberts EB (1978) Research and Development Policy Making, In: Roberts EB (ed) Managerial Applications of System Dynamics. Productivity Press, Cambridge, Massachusetts, pp 283–292
Roberts EB (1978) A Simple Model of R&D Project Dynamics. In: Roberts EB (ed) Managerial Applications of System Dynamics. Productivity Press, Cambridge, Massachusetts, pp 293–314
Rogers EM (1983) Diffusion of Innovation, 3rd edn. The Free Press, New York
Schumpeter JA (1961) Konjunkturzyklen – Eine theoretisch, historische und statistische Analyse des kapitalistischen Prozesses, Erster Band. Vandenhoeck & Ruprecht, Göttingen
Steele LW (1989) Managing Technology. McGraw-Hill, New York
Sterman JD (1989) Modeling Managerial Behavior: Misperceptions of Feedback in a Dynamic Decision Making Environment. Manag Sci 35:321–339
Sterman JD (1994) Learning in and about Complex Systems. Syst Dyn Review 10:291–330
Weil HB, Bergan TB, Roberts EB (1978) The Dynamics of R&D Strategy. In: Roberts EB (ed) Managerial Applications of System Dynamics. Productivity Press, Cambridge, Massachusetts, pp 325–340

Discovery Systems

Petra Povalej, Mateja Verlic, Gregor Stiglic
Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia

Article Outline

Glossary
Definition of the Subject
Introduction
Knowledge Discovery and Data Mining Process
Application Domain Understanding
Data Understanding
Data Preparation and Identification of DM Technology
Applying Data Mining
Interpretation and Evaluation of Results
Utilization of Results
Knowledge Discovery Frameworks and Tools
Conclusions and Future Directions
Bibliography

Glossary

Accuracy (rate) Used for evaluating quality of induced model.

Average class accuracy One of the simplest metrics for estimating the quality of a model. Classification accuracy is calculated for each class of the target variable and then the average of all accuracies per class is calculated.

Aggregation Process of combining two or more objects into single one. Typical statistical aggregation functions for quantitative attributes are sum and average.

Attribute A property or characteristic of data object, which may vary in time and also from object to object. Attributes have usually assigned values or symbols for the purpose of analysis. Other frequently used names for an attribute are variable and feature.

Binarization Transformation of continuous or discrete attributes into binary attributes. Binary attributes have only two possible values.

Classification Classification of data objects is a process of assigning classes or class labels to data objects. It is a type of predictive modeling and it is used for predicting discrete target variable.

Classification accuracy See description under *Accuracy (rate)*.

Classifier A model based on data used for classification.

Confusion matrix A matrix of results from testing model versus predicted class values. It is very useful visual tool for understanding results of testing a classification model.

Data cleaning Step of KDDM usually involving detection and correction of data quality problems, removal of noise, defining outliers, and dealing with missing values.

Data mining A technology combining traditional data analysis methods and sophisticated algorithms for automatically processing large volumes of data and finding and extracting novel, useful and usually hidden patterns.

Data object A record of attribute values about an object or person. Other common names for data object are data record, case, point, sample, observation or entity.

Data preprocessing/preparation A phase of KDDM related to the preparation and transformation of data for data mining. It comprises of several techniques for selecting relevant data and attributes and creating or changing the attributes.

Data set A collection of similar or related data objects. Data objects are usually collected for a particular study.

Dimensionality reduction Reduction in the number of attributes. It is used for eliminating irrelevant features and noise by creating new attributes as a combination of the old attributes. Feature subset selection or feature selection is other type of dimensionality reduction, where dimensionality is reduced by selecting and using only a subset of old attributes.

Discretization A transformation of continuous numerical attributes into categorical or discrete attributes.

Ensemble Also known as committee or multiple classifier system is a group of classifiers. Ensemble approaches exploit the classification abilities of multiple classifiers. The integration of classifiers usually enhances the performance of final classification.

Feature A feature (variable) is a synonym for attribute. It is frequently used in data-mining domain. See *Attribute*.

Feature extraction A process of creating new features from the original raw data. It is highly domain-specific.

Knowledge discovery and data mining (KDDM) The "umbrella" term for the overall process of knowledge discovery.

Knowledge discovery (KD) Nontrivial process of mapping low-level data into other more meaningful forms that are easier to understand, like patterns, rules, summaries or even graphs.

Noise Result of erroneous measurements. It can involve distortion of values or addition of unauthentic data objects. Unlike outlier, noise is not legitimate data.

Outlier An anomalous object or atypical value of an attribute. Outliers can be legitimate data objects or values. Detecting outliers is especially important in fraud detection or network intrusion detection.

Pattern In KDDM defined as a high-level description of a subset of data and can be in many forms, e. g. statistical or predictive models of data, relationships among parts of data sets, association rules, clusters, graphs, summaries, or classification rules, tree structures, linear equations etc.

Precision Fraction of positive samples correctly classified as positive among all samples classified as positive.

Recall See *Sensitivity*.

Regression Regression is a type of predictive modeling used for predicting continuous target variable.

Sampling A process of selecting a subset of data or sample, for the data analysis. Basic sampling techniques are simple random sampling with or without replacement, stratified sampling and adaptive or progressive sampling.

Sensitivity (recall) Proportion of samples correctly classified as positive (true positives) of all positive samples tested. If sensitivity is 1, all positive samples have been identified as positive.

Specificity Proportion of samples classified as negative of all negative samples tested.

Definition of the Subject

By definition, to *discover* is to see, get knowledge of, learn of, find or find out; gain sight or knowledge of something

previously unseen or unknown [18], therefore a discovery system can be defined as a system that supports the process of finding new knowledge. Results of a simple query for *discovery system* on the World Wide Web returns different types of discovery systems: from knowledge discovery systems in databases, internet-based knowledge discovery, service discovery systems and resource discovery systems to more specific, like for example drug discovery systems [10], gene discovery systems [43], discovery system for personality profiling [48], and developmental discovery systems [17] among others. As illustrated variety of discovery systems can be found in many different research areas, but we will focus on knowledge discovery and knowledge discovery systems from the computer science perspective. Inconsistent definitions of terms knowledge discovery (abbreviated, KD), knowledge discovery in databases (abbreviated, KDD) and data mining (abbreviated, DM) frequently confuse potential a novice in the field of knowledge discovery.

DM is a technology combining traditional data analysis methods and sophisticated algorithms for automatically processing large volumes of data and finding and extracting novel, useful and usually hidden patterns. DM is also associated with other names, such as knowledge extraction, information harvesting, data archaeology, data pattern processing, and knowledge discovery in databases. DM and KDD are often used as synonyms, although DM is an integral part of KDD.

KD is a process of mapping low-level data into other more meaningful forms that are easier to understand [21] and knowledge discovery systems are systems that implement knowledge discovery process. Also, KD can be defined as a process that seeks new knowledge about specific problem or application domain.

KDD is a KD process applied to databases [30]. In the most popular definition by Fayyad et al. [21] KDD is defined as a non-trivial process of identifying valid, novel, potentially useful, and understandable patterns in data. Note that although databases are considered as a primary source of data in KDD, the definition itself is not limited to one type of data sources only.

The umbrella terms *knowledge discovery* and *data mining* (abbreviated, KDDM) was proposed as the most appropriate name for the overall process of KD [30] instead of *knowledge discovery process* (abbreviated, KDP) and it will be used from now on to describe the overall process of knowledge discovery from different sources of data including methods for storing and accessing data, scalability of algorithms for large data sets, interpretation and visualization of results, and modeling of human-machine interaction [21].

The importance of KD systems lies in discovering the knowledge that may otherwise remain hidden and to integrate extracted knowledge into decision support systems used by scientist and businesses. The need for automated or at least semi-automated data analysis arose when computers revolutionized our way of living and making business. The digital information era brought advancements and benefits but also a very serious pitfall - data overload or data flood. Waste amounts of data, currently measured in terabytes, are collected daily, much more than human analysts can examine in foreseeable time. Still, the importance of human involvement in KD is not to be neglected. KD can only assist people with finding hidden patterns and relationships in data; it cannot tell whether discovered patterns are valuable for the organization.

Introduction

Knowledge discovery is not a new approach to analyzing data. People soon discovered that data need to be analyzed to make sense. The term *knowledge discovery in databases* was introduced in 1989, when first workshop on KDD was organized [35]. Soon after that, in 1991, Frawley [23] published a definition of KDD, which was later revised by Fayyad et al. [21]. The revised definition of KDD became very popular in KD community.

Classical approach to data analysis involves one or more analysts who need to be closely acquainted with the data. The analysts act as an interface between the data and the end-users of gained knowledge. Although this approach is perhaps appropriate for small amounts of data or with data with small number of variables or attributes, it has many drawbacks if larger amount of data or large number of variables are involved. Unfortunately, these drawbacks cannot be afforded in the modern high-paced business world. First, analyzing large amounts of data manually is too slow to effectively apply the results of the analysis. Second, this approach is too expensive because experts need to be very familiar with the data (their expertise is valuable) and third, the level of subjectivity of results is too high. Hence, the urge for at least semi-automated data analysis became even stronger.

Several computational approaches to data analysis have been proposed as an alternative to the classical approach. They are based on the idea that while on one hand computers contribute to data overload, they can on the other hand assist humans in extracting useful knowledge by finding meaningful patterns from growing volumes of data. For example, statistical, artificial intelligence and machine learning and even cognitive approaches can all deal in the same problem domain but from different perspec-

tives and in different ways using different models, methods, and techniques.

In contrast to single-disciplinary approaches, knowledge discovery in databases is an interdisciplinary approach. It evolved from several research fields and it incorporates many techniques and findings from statistics, artificial intelligence, machine learning, pattern recognition, databases, and data visualization among others. All those disciplines are combined in KDD with one common goal: extracting high-level knowledge from low-level data.

Before the beginning of the 21st century Piatetsky-Shapiro [36] had already identified three generations in KD systems. First-generation KD systems (early-1990s) were intended for expert users and provided only single data mining technique for data analysis per system. They had very weak support for the overall KDDM process and had only limited commercial success. The main research was focused on the development of new data mining algorithms. Second-generation KD systems (mid-1990s), called suites, were collections of multiple data analysis methods with the support for data cleaning, preprocessing and visualization. They were certainly a step forward, but the idea of KDDM process model was still not implemented. Third-generation KD systems (late-1990s) introduced different approach by addressing specific business problems like fraud detection. They provided interface to hide complexity of underlying data mining techniques and introduced first KDDM process models. Now, when we are approaching the second decade of the 21st century, it is possible to identify new generation of KD systems. This new generation of KD systems aims at integration and interoperability of modern KDDM models through use of popular industrial standards like eXtensible Markup Language (abbreviated, XML) and Predictive Model Markup Language (abbreviated, PMML) or some other approaches [30].

After the first workshop on KDD the idea of KDDM model evolved simultaneously with the development of KD systems. In 1996 Fayyad et al. [22] laid the foundation for KDDM process model by proposing the basic structure of the model. After the publication of their book the development of KDDM models followed two different streams: data-centric and human-centric. Data-centric (or data-driven) models focused on the iterative and interactive nature of the data analysis task, while the human-centric models focused on a series of knowledge-intensive tasks with complex interactions between human and the database [30]. Furthermore, the models can be broadly divided into academic, usually not considering industrial aspects of KDDM projects, and industrial, which address specific industrial issues. All process models consist of multiple sequential steps with loops and interactions; they differ in the number and the scope of proposed specific steps. Also, all models emphasize the iterative nature of the model.

Although many models have been developed, only five of them left significant impact and were applied in at least several real KDDM projects [30]. The nine step academic data-centric KDDM model by Fayyad et al. [22] in 1996 was the first reported model. In the same year the CRISP-DM (Cross-Industry Standard Process for DM) model was proposed by a consortium of four companies: SPSS, NCR, Daimler Chrysler and OHRA, but its mature version was released in 2000. The second model, this time from industrial area, was suggested in 1998 by Cabena et al. [9] and it consists of five steps. Third model was an eight-step academic model by Anand and Buchner [1], developed at about the same time as the second model. In 2000 the fourth model, already mentioned CRISP-DM model consisting of six steps, was officially released [44]. This model is still strongly supported by the industry, CRISP-DM Special Interest Group and the European Commission. The last, fifth model consisting of six steps, was proposed by Cios et al. [11]. This model adopted CRISP-DM model to the needs of academic research community [30]. The study by Kurgan and Musilek [30] offers detailed description and an exhaustive comparison of the major five existing KDDM models. They also introduced a generic model, which tries to capture main steps and tasks of the mentioned KDDM models. In following sections and subsections this generic model was used in order to avoid favoring any particular KDDM model. Using KDDM models for knowledge discovery is essential to ensure that useful knowledge is discovered; blind application of data mining methods, notoriously known as data dredging or data fishing can easily lead to the discovery of potentially interesting but meaningless patterns.

Other pioneers, who were not explicitly mentioned above but also contributed significantly to the development of contemporary KDDM, are P. Smyth [21,26] and H. Manilla [26,32].

In the following sections knowledge discovery process will be described in more details. Each step of the process will be briefly explained to show how a particular step contributes to the results of the entire KDDM process. Detailed description of the KDDM process will be followed by a section on KDDM frameworks, which will include a review of contemporary frameworks. In next section several applications of KDDM will be mentioned so that we can recognize the real value of applying KD to the real world problems. In the last section some guidelines for the future development of KDDM will be suggested.

Knowledge Discovery and Data Mining Process

The simplest definition of KDDM is an overall process of transforming raw data into knowledge. As already mentioned in the introductory section, KDDM process includes, among others, methods for this transformation, which is the central part of KDDM process. Fayyad et al. [21] defined this transformation as KDD and described it as 'the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data'. Before we describe KDDM in more details, it is important to understand the meaning of basic terms in this definition. **Data** is a set of facts and a **pattern** is an expression (high-level description) that describes a subset of the data or a model applicable to this subset. Patterns, as mentioned in definition, can take many forms; for example, they can be statistical or predictive models of data, relationships among parts of data sets, classification rules, association rules, summaries, linear equations, clusters, graphs, tree structures, or recurrent patterns in time series. In this context a pattern is considered as **knowledge** if it exceeds a certain level of **interestingness**, usually set by user or even data itself. The notion of interestingness was introduced by Fayyad et al. [21] and it represents an overall measure of pattern's value, novelty, usefulness, and understandability. Because functions and thresholds for measuring interestingness are chosen by user, this definition of knowledge is purely user oriented and domain specific. As we may intuitively know, knowledge is much more than just patterns or relationships in data. For instance, researchers from the field of knowledge management define knowledge as a mix of experience, values and

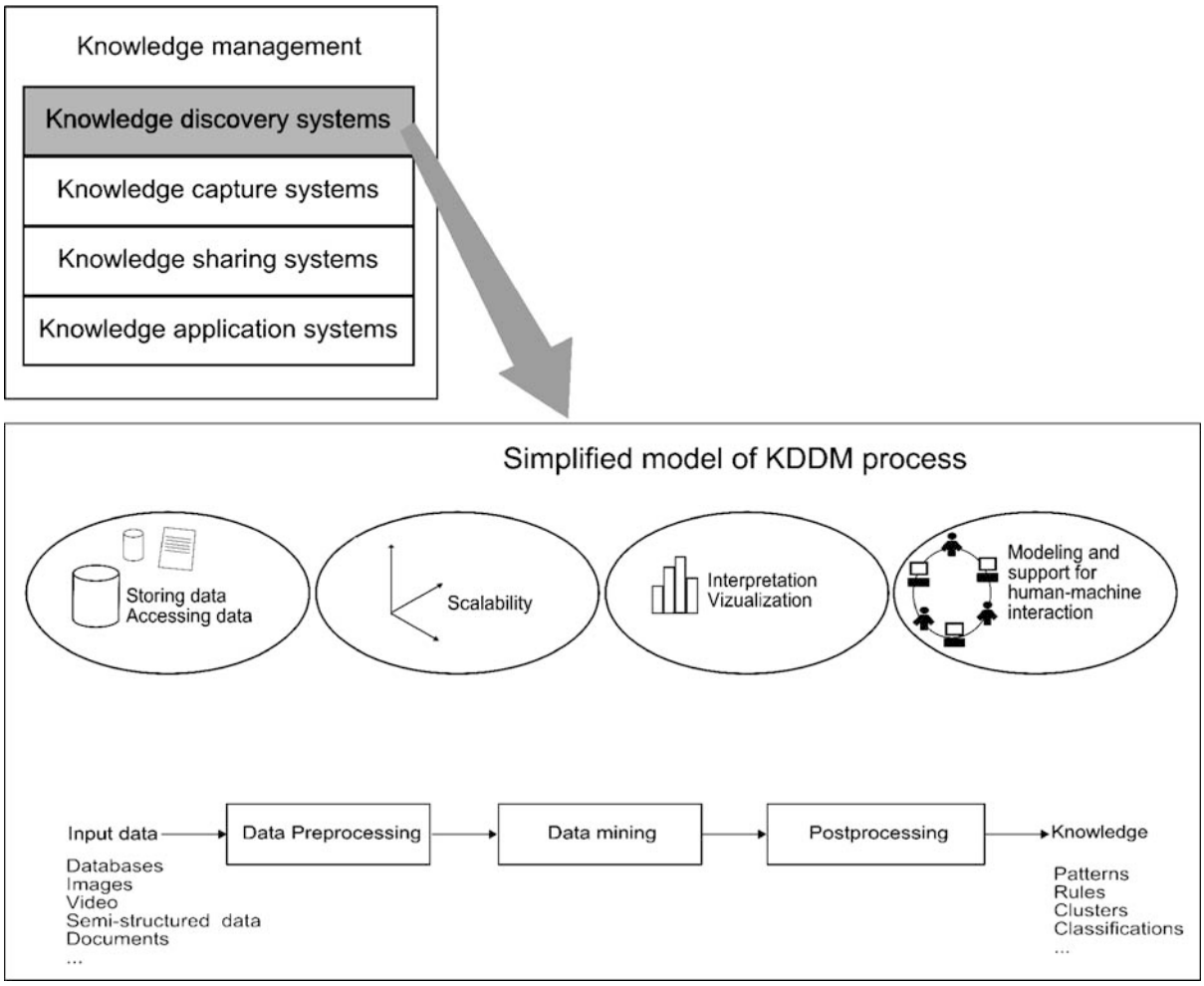

Discovery Systems, Figure 1
Relation between knowledge management, knowledge discovery systems, and data mining

insights that provides some sort of framework for evaluating new experiences and information. It is embedded not only in documents but also in organizational routines, processes, practices and norms [14]. This broader definition of knowledge is certainly better than the definition of knowledge as specific patterns, but unfortunately it is not very useful in the context of KDDM.

Figure 1 shows the relationship between knowledge management, knowledge discovery systems, which implement KDDM process, and data mining. KDDM process is essential part of an even broader process named knowledge management. While knowledge management deals with aspects of gathering, organizing, sharing and analyzing intellectual capital – knowledge [4], KDDM focuses on the overall process of knowledge discovery from data, including methods for storing and accessing data, scalability of algorithms for large data sets, interpretation and visualization of results, and modeling of human-machine interaction. Simplified model of KDDM process clearly illustrates the central role of DM in KDDM process.

KDDM is a process of several steps which transforms inputs in form of raw data into outputs – applicable knowledge. Different KDDM models define steps involved in this process, but the number and the scope of particular step vary from model to model. We decided to describe the process of KDDM using the generalized model of KDDM process proposed by Kurgan and Musilek [30]. Figure 2 shows the steps of generalized model around the schematic diagram of general sequential structure of KDDM process [12] in the center. The sequential structure includes many feedback loops between steps, because the process of revision can show that some of the steps need to be repeated.

The generalized KDDM model has six steps:

(1) Application domain understanding,
(2) Data understanding,
(3) Data preparation and identification of DM technology,
(4) Data mining,
(5) Evaluation and
(6) Knowledge consolidation and deployment.

Each step will be described in more details in the following sections.

Application Domain Understanding

Before analyzing data it is recommended to define the goals or objectives of the analysis. First, we have to understand the domain objectives that are important for the end-user, for example a customer. We have to carefully consider constraints, assumptions and any other factors that may affect our search for new knowledge. Second, some project management efforts are needed to identify human and technical resources, decompose KDDM project into tasks, estimate the duration of each task, define inputs, outputs, and dependencies to successfully apply data mining techniques. Each task should also be associated with particular KD goal. Third, at this point business goals or KD goals need to be translated into more spe-

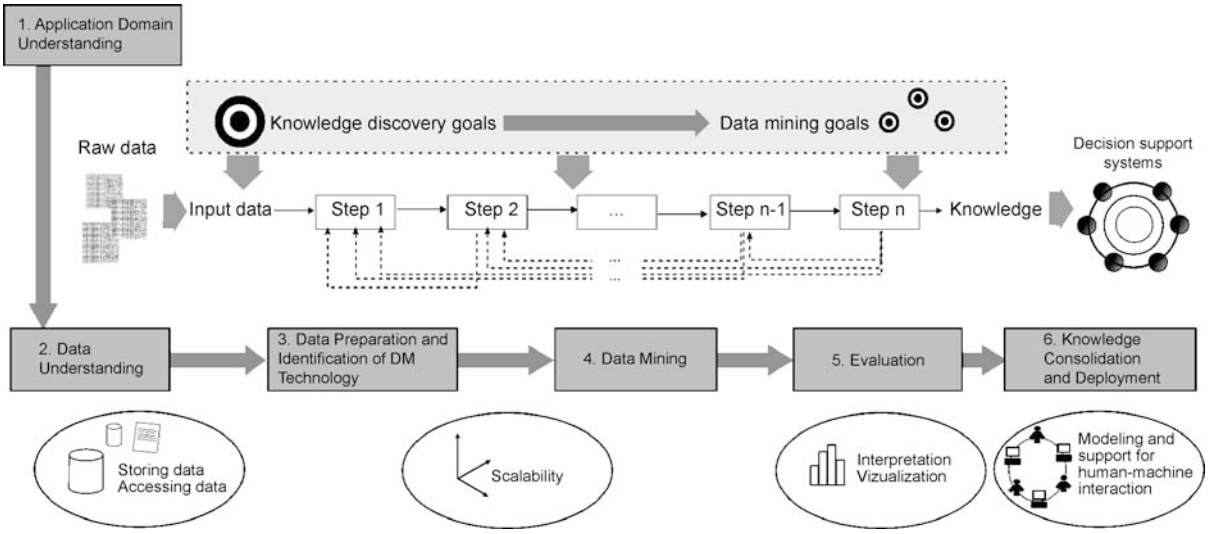

Discovery Systems, Figure 2
Generalized model of KDDM process

cific DM goals, so that later one the most appropriate DM techniques can be chosen and applied.

Furthermore, understanding relevant prior knowledge can also help to effectively execute further steps of the process.

Data Understanding

When knowledge discovery goals and more specific data mining goals are known, we can start to collect data included in the resources we included in the plan. Resources may include transactions data, customer histories, demographic information, protein sequences, microarray expression data or weather data. In this phase fusion of data from multiple sources has to be performed and we have to familiarize with data, so we can more easily identify potential data issues. Diversity of resources contributes to diversity of formats in which data can be stored, for example we can find data in flat files, relational tables, spreadsheets or even semi-structured documents at either centralized or distributed locations.

A **data set** can be considered as a set of **data objects** (also known as records, cases, points, samples, observations, entities) and data objects are described by **attributes** (also known as variables, characteristics, features, dimensions). For example, in files with a record data set data objects are written as records in rows and each field or column corresponds to an attribute (Table 1).

Attribute is a property or characteristic of a data object that may vary from object to object and even from one time to another. We assign numbers or symbols to attributes so we can analyze characteristics of object. As it can be seen from the example, different types of values can be assigned to attributes. Which type of value, either symbolic or numerical, will be associated with attribute is defined by a rule or function called a **measurement scale**.

Note that properties of an attribute and the properties of the values for measuring it do not have to be the same. For example, patient's identification number (ID) and age can be both represented as numeric values, usually integers. While it is quite reasonable to calculate average age of the patients, it is senseless to calculate average ID, because the purpose of using IDs is to distinguish data objects. Also, it can happen that an attribute is measured in such a way that measurements do not capture all the properties of the attribute. Thus, knowing the type of an attribute is important. It tells us which properties of the measured values are consistent with the properties of the attribute, so we can avoid senseless actions.

According to the properties of numbers that correspond to the properties of attribute we can define four types of attributes: nominal, ordinal, interval, and ratio. **Nominal** attributes are described with symbols or numbers which distinguish one object from another. Nominal attributes from the example in Table 1 are identification number, gender and name. **Ordinal** attributes provide information to order the objects. Examples of ordinal attributes are level of tiredness and blood sugar. For **interval** attributes, the difference between values are important and unit of measurement exists. An example of interval attribute is body temperature in the Celsius scale. For **ratio** attributes, both differences and rations are important. Examples of ratio attributes are age, height and weight.

Nominal and ordinal attributes are **categorical** or **qualitative** attributes, while interval and ratio attributes are **quantitative** or **numeric** attributes. Qualitative attributes take on values that are names or labels and do not have properties of numbers. Even if their values are numeric, they should be treated more like symbols. Quantitative attributes are represented by numbers and have most of the properties of numbers.

If an attribute can have any value between two specified values, then it is **continuous**; but if it can take only a limited or finite number of values it is **discrete**. Example of continuous attributes is weight, while gender is an example of discrete attribute.

Discrete attributes are often represented using integer values. **Binary attributes** are special case of discrete attributes. They can have only two values like yes/no, true/false, male/female or 1/0 and are usually represented with Boolean variables or as integer variables. Gender, level of tiredness and blood sugar from Table 1 are ex-

Discovery Systems, Table 1

Sample data set containing information of patients. Each row in a table is a record and attributes are person identification number, name, age, gender, height, weight, level of tiredness, blood sugar, and body temperature

Person ID	Name	Age	Gender	Height (cm)	Weight (kg)	Level of tiredness	Blood sugar	Body temp. (°C)
…	…	…	…	…	…	…	…	…
13678	John Doe	37	M	183.5	84.5	Low	Normal	35.7
13679	Jane Smith	28	F	158.2	75.3	Medium	High	37.1
…	…	…	…	…	…	…	…	…

amples of discrete attributes. Continuous attributes have values which are real numbers and are typically represented as floating point variables. Height, weight and body temperature from Table 1 are examples of continuous attributes.

Different types of data sets exist. Most common types are data sets with record data (e. g. transaction data, data matrix, and sparse data matrix), graph-based data (e. g. data with relationships, data as graphs) and ordered data (e. g. sequential data, time series data, and spatial data). Data sets may even display some special characteristics. For example, when we are working with time series data, we should consider temporal autocorrelation; or when we are dealing with spatial data, we should not ignore spatial autocorrelation. Furthermore, some data sets might even include explicit relationships within data (e. g. linked documents in the WWW). Many data sets share three general characteristics that significantly affect data mining techniques: dimensionality, sparsity, and resolution.

The **dimensionality** of a data set is the number of attributes that are used to describe data. Sometimes it is possible to come across the term **curse of dimensionality**. This term is used for difficulties that usually occur during analysis of high-dimensional data and therefore dimensionality reduction is an important part of data pre-processing phase.

In some data sets only few attributes of an object have non-zero values. In this case we can talk about **sparsity**. Attributes, for which only non-zero values are important, are called **asymmetric attributes**. Sparsity in data set can be considered as an advantage because we don't have to store, manipulate or analyze non-important (zero) values. Some data mining algorithms work well only for sparse data sets.

Data is often obtained at different levels of resolution. The properties of data at different resolution levels are different. For example, we measure weight of babies in grams, while the weight of adults is measured in kilograms. If the resolution is too fine, a pattern might not be visible or too much noise is present; on the other hand, if resolution is too coarse, the pattern may disappear.

In this phase we also have to deal with several data-related issues: the types of data in our selected data sets, the quality of data, how to improve data quality or how to modify data for the needs of data mining methods. Data is never perfect: values may be missing, duplicate objects may exist, inconsistencies may occur due to human error, hardware and software limitations, or flaws in the data collection process. Detection and correction of data quality problems is often called **data cleaning**. When dealing with data quality we have to consider measurement and data collection issues and also issues related to applications. We will briefly describe some of most frequently occurring data errors: noise, outliers and missing values.

We can encounter a variety of problems related to **measurement error**, when recorded value differs from 'true' value to a certain extent, and **data collection error**, which refers to omitting data objects or attribute values or including data object in inappropriate way. Both errors can be systematic or random. Good news is that for some types of data errors well-developed techniques for detecting and correcting these errors exist.

Noise is an example of measurement error. Often it occurs randomly and it may involve distortion of values or addition of unauthentic objects. Usually, the term noise is used in connection with temporal or spatial data, but is not uncommon in other types of data as well. The elimination of noise is difficult, therefore **robust** (insensitivity for noise) data mining algorithms are desired. Precision, bias and accuracy are also important factors for measuring the quality of data.

Outliers are data objects that are different from most of the other data objects or values of an attribute that are atypical for this attribute. Outliers are considered anomalous objects or values. It is important to distinguish between noise and outliers. Unlike noise, outliers can be legitimate data objects or values and may sometimes be interesting, especially in fraud detection or network intrusion detection.

Missing values are not uncommon in data sets. A data object can have one or more missing attribute values; sometimes the values were not collected or attributes are not applicable to all objects or attributes are optional. Usually, for the sake of simplicity, all attributes (fields) are stored, but during the data analysis all missing values should be included.

The problem of **duplicate data** should also be taken into an account in data cleaning. Data duplication can occur due to inconsistent values and these values must be resolved. However, duplication can also occur due to accidentally combining data objects, that are similar but not the same. For example, two or even more persons can have identical names.

Data Preparation and Identification of DM Technology

Data pre-processing phase is related to preparation and transformation of data into more suitable form for data mining. This particular step is by far the most time-consuming part of the KDDM process [12].

Data pre-processing strategies and techniques can be roughly divided into two groups – strategies and techniques for:

(1) Selecting relevant data object and attributes, and
(2) Creating or changing the attributes.

We will familiarize with terms aggregation, sampling, dimensionality reduction, feature subset selection, feature creation, discretization, binarization, and variable transformation. In data pre-processing the term feature or variable is frequently used as a synonym for attribute.

The idea behind **aggregation** is to combine two or more objects into a single object. Typically, quantitative attributes are aggregated by simple statistic functions like a sum or an average. Qualitative attributes are more difficult to deal with. They can be either omitted or summarized in some way. If aggregation is used, then resulting smaller data sets require less computer resources (memory and processing time) and thus more expensive data mining techniques can be used. Aggregation can also serve as a high-level view of the data. However, aggregation can also lead to potential loss of interesting details.

Sampling is an approach for selecting a subset of data objects for the data analysis. It has been long used in statistics and it can be also very useful for data mining. In statistics sampling is used because obtaining the entire set of data of interest is too expensive or even impossible; in data mining sampling is used because processing of all the data would be too expensive or time consuming. Sampling can, similar to aggregation, reduce the size of data set and thus more expensive data mining algorithms can be used. Sampling is effective only if the sample is **representative**; in other words, a sample should have very similar or even identical characteristics as the original (entire) set of data or **population**. Different sampling approaches have been developed. One of the basic sampling techniques is **simple random sampling**, which has two variations:

(1) Sampling without replacement and
(2) Sampling with replacement.

In random sampling all items in the population have equal opportunity to be selected. In sampling without replacement an item that was selected is removed from the population. In sampling with replacement selected item is not removed from the population, so the same item can be selected more than once. Because simple random sampling is not suitable for less frequent types of objects (rare cases), different sampling method is needed for such data sets. **Stratified sampling** takes into account differing frequencies of objects types. It starts with predefined groups of objects. In one version equal number of objects is drawn from each group, regardless of potentially different sizes of groups. In another version, the number of objects drawn from each group is proportional to the size of corresponding group. After selecting sampling technique, we have to decide on the **sample size**. Sample size is very important, because if we choose large sample sizes the probability of having a representative sample is higher, but we lose advantage of sampling. On the other hand, if we use a smaller sample size we may miss some patterns or even detect erroneous patterns. Because defining proper sample size can be difficult, we can sometimes use **adaptive** or **progressive sampling**. In this kind of approach, sample size is progressively increased until sufficient size is obtained, but we need a way to determine if the sample size is large enough.

Data sets can have a large number of attributes or features, even tens of thousands attributes. Because many data mining algorithms are more suitable for data sets with smaller number of attributes, reduction in the number of attributes or **dimensionality reduction** is desired. Dimensionality reduction has many benefits, a key one is that data mining algorithms work better with lower dimensionality of data. This is partly due to elimination of irrelevant features and noise reduction. Dimensionality reduction can also lead to more understandable model with fewer attributes, data can be more easily visualized and data mining algorithms require less computer resources. Dimensionality reduction techniques usually reduce the dimensionality by creating new attributes as a combination of the old attributes. Dimensionality of data sets can be also reduced by selecting and using only a subset of old attributes. This reduction is known as **feature subset selection** or **feature selection**. Redundant features duplicate much or all information contained in one or more features while **irrelevant features** contain little or no information that could be used in data mining. Redundant and irrelevant features can even reduce classification accuracy and the quality of clustering and eliminating them can only be beneficial. Some of features can be eliminated immediately, but otherwise selecting the best subset of features requires a systematic approach. Three standard approaches to feature selection exist: embedded (feature selection occurs naturally as a part of data mining), filter (features are selected before data mining), and wrapper (learning algorithm itself is used as part of the evaluation function for selecting features). Sometimes we want to create new set of attributes from original ones, especially if new features capture important information more effectively. Creating new features from the original raw data is known as **feature extraction**, but unfortunately this approach is highly domain-specific. For some cases data need to be mapped to a new space to reveal interesting and important features,

for example, using a Fourier transformation in time series analysis. In situation, when the features in the original data set have the necessary information but they are not in the form suitable for data mining, constructing one or more features from the original ones can be more useful. The most common approach to **feature construction** is using domain expertise.

Several data mining algorithms, especially in classification or association analysis, require data in the form of categorical attributes of binary attributes. Normally attributes in data sets are not only categorical or binary and transformation of attributes is needed. Transforming continuous attributes into categorical attributes is **discretization** and transformation of continuous or discrete attributes into binary attributes is **binarization**. There are many different approaches to discretization and binarization, but they all try to produce the best result for data mining algorithm we intend to use for the data analysis. Another type of transformation is **variable transformation** that is applied to all values of an attribute or variable. Two important types of variable transformations are simple functional transformations and normalization. For **simple functional transformations** a simple mathematical function is applied individually to each value. For example, if x is a variable such simple functions are x^k, $\log x$, e^x, $\sqrt{x}$, $1/x$, $\sin x$, or $|x|$. Note that variable transformations should be applied with caution, because they change the nature of data (e. g. changing magnitudes, order, or negative values). Standardization or normalization of a variable is concerned with transforming entire set of values in such a way, that they share particular property. This transformation is often necessary if are combining different attributes (variables) in some way and we want to omit the dominance of large values.

Some researchers recommend choosing DM technology before data preparation, others say that methods should be selected we have prepared, but in practice these two things are intermingled. Data need to be prepared at least in such a way that we can decide which DM methodology is most appropriate for selected DM task. Among core DM tasks are predictive modeling, anomaly detection, association analysis and cluster analysis.

With **predictive modeling** we build a model for the target, or dependent variable as a function of other independent variables. Well-known types of predictive modeling are **classification**, which is used for predicting discrete target variable, and **regression**, which is used for predicting continuous target variables. The goal of predictive modeling is to minimize the error between predicted and true values of the target variable [47]. **Anomaly detection** is identification of data samples, known as **anomalies** or **outliers** that are significantly different from the rest of data. The main goal of this task is to identify the real outliers not the noise. **Association analysis** is used to discover strongly associated features and patterns that occur in data because of these associations. Discovered patterns are usually in the form of implication rules or feature subsets. The main goal of this analysis is efficient extraction of the most interesting patterns, for example, which items are frequently bought together by customers. **Cluster analysis** is a task of finding groups or **clusters** of data samples. Data samples in a cluster are more similar to each other than samples of other clusters.

After identifying DM tasks, we can choose one or more of appropriate algorithms for selected DM task.

Applying Data Mining

When applying DM process we can generally pursue two different types of objectives: either to verify some user predefined hypotheses (**verification**) or find new knowledge upon which a user can base decisions using (**knowledge discovery**). Both verification and knowledge discovery are done by building a model of a real world based on collected data in a specific application domain. The result of model building is a description of patterns and relationships in data, which can be used for verification of predefined knowledge (hypothesis), discovery of new knowledge for either **prediction**, where the system autonomously finds patterns for predicting future behavior of some entities (class attributes), or **description**, where the system finds patterns for a presentation in a human-understandable form. The boundary between prediction and description is blurred – some prediction models can also be understandable to some degree.

After selecting specific goals of DM process and the type of prediction, which is the most important for mining the data from the data set, a DM model type can be chosen. Several traditional statistical models exist, for example discriminant analysis, general linear models and logistic regression, but there are also some models from the fields of machine learning and pattern recognition, such as neural net to perform regression, a decision tree or a decision rules for the classification.

Furthermore, several algorithms are available for building selected model. For example, we may build a neural net using backpropagation or radial basis functions; for inducing a decision tree we can choose among C4.5 [39], CART (classification and regression trees) [7], CHAID (CHi-squared automatic interaction detector) [29], QUEST (quick, unbiased, efficient, statistical tree) [31] and many other algorithms. Algorithms tend

to differ primarily in the goodness-of-fit criterion used to evaluate model fit or to find the good fit in a process of searching.

In predictive models, the values (classes) we are predicting with a model are called **response/dependent/target variables** and the values used as an input to build a model are called **prediction/independent variables**. The process of building predictive models requires a well-defined training and validation protocol in order to ensure the most accurate and robust predictions. This kind of protocol is called **supervised learning**. The idea of supervised learning is to *train* a model on a portion of data from the data set (**a training set**), then test and validate it on remainder of the data (**test set**). For each observation in a data set values predicted by the model are compared with actual (known) values of the response variable. For contrast, some of the descriptive techniques like clustering can build a model without known values of the response variable. These types of techniques are sometimes referred to as **unsupervised learning**.

It is important to remember that no model or algorithm should be used exclusively. Model building is an iterative process. The nature of the problem and the data itself affect the choice of models and algorithms. There is no best model or algorithm in general. Selection of the most appropriate model type and algorithm can be crucial for the success of DM process. Therefore for a given problem different models should be explored to find the one producing the most useful solution.

In the following sections, some of the types of models and algorithms used to mine data will be briefly examined. In general, most of the model types can be viewed as an extension or a hybrid of some primary types of models. Advanced descriptions of models can be found in other sections of this chapter.

K-Nearest Neighbor (k-NN)

K-nearest neighbor (k-NN) is a classification technique that defines how to classify an observed item on the basis of the majority class of k items in the neighborhood (Fig. 3).

The idea of k-NN is based on measuring a distance between attributes in the database. If the attributes are numerical the calculation of distances is easy, but categorical attributes require special handling (e. g. what is the distance between blue and gray?). The next task is a selection of the neighborhood of pre-classified cases to be used for classification of a new case and possibly the influence of neighbors considering the distance from the new case. For example, nearest neighbors can have a higher influence (weight) on classification of a new case than the neighbors that are father away.

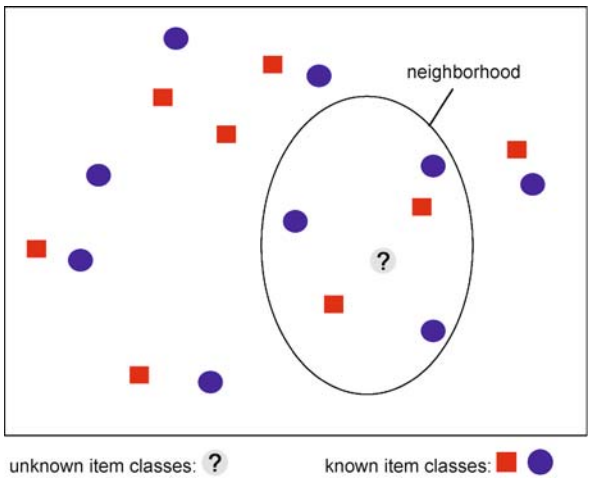

Discovery Systems, Figure 3
k-Nearest neighbor. ? has unknown class; ● and ■ are known classes of the neighbors. In most simple algorithm the class of unknown item ? would be ● since 3 items in the neighborhood have class ● and 2 items have class ■

Since k-NN models are computationally demanding they are mostly useful when only a few prediction variables are included. Different data types of variables can be included into the model under only one basic requirement: the metric for measuring distance has to be defined. The advantage of k-NN models is their understandability.

Memory–Based Reasoning (MBR)

MBR is one of the memory based technologies that belongs to the family of nearest-neighbor-like models. It is used on databases with large number of cases where the data is kept in a system's memory in order to speed up computations. As in k-NN model, MBR algorithms search for the most similar problem from the database of pre-classified cases and apply its class to a new case. There is no learning phase and no adaptive or learning parameters to manipulate.

A basic advantage of MBR algorithms is their effectiveness (on a proper database) that is comparable (and sometimes even better) to neural networks. The disadvantages of MBR algorithms are: a lack of generalization (a large number of good examples are needed), no data compression, large computational demands and consequently large memory requirements that all make them unsuitable for on-line learning.

It has been often used for object recognition and classification of free text samples taken from very large databases.

Rule Induction

Rule induction algorithms derive a set of independent rules from the database of pre-defined cases to classify a new case. Decision rules can be in a simple form, like:

$$\textbf{IF} \quad condition_j \quad \textbf{THEN} \quad \text{CLASS} = class_i \ .$$

Example of a simple rule is: *IF body temperature is higher than 38 degrees THEN patient is ill*. More complex rules can include conjunction (logical AND) and/or disjunction (logical OR), like:

$$\textbf{IF} \quad condition_a \quad \textbf{AND} \quad condition_b \quad \textbf{OR} \quad condition_C \quad \textbf{THEN} \quad \text{CLASS} = class_i \ .$$

An example of a more complex rule is: *IF body temperature is higher than 37.5 degrees AND diarrhea is true THEN patient is ill*. More complex dependent rules can form a decision tree.

Rules can also include ELSE statement, like:

$$\textbf{IF} \quad condition_j \quad \textbf{THEN} \quad \text{CLASS} = class_i \quad \textbf{ELSE} \quad \text{CLASS} = class_k .$$

An example of IF-THAN-ELSE decision rule is: *IF body temperature is higher than 37.5 degrees AND diarrhea is true THEN patient is ill* ELSE *patient is healthy*.

Disadvantages of rule induction that have to be considered are:

(1) A set of induced rules does not necessarily cover all possible situations (it is possible that the system will not be able to classify a new case),
(2) Two or more induced rules my conflict in their predictions. A usual solution to the second disadvantage is to assign a confidence to each rule and for classification of a new case use the most confident one or alternatively to use weighted voting of rules according to their confidence.

An important advantage of decision rule induction is certainly their simplicity and understandability.

Decision Trees

Decision trees are a way of representing a series of rules that lead to classification of known classes. For example, we might wish to diagnose young children with asthma on the basis of gathered data about family history, blood examinations and other indicators of asthma in adults. Figure 4 shows a simple decision tree that solves this problem while illustrating basic components of a decision tree: internal nodes, branches and external nodes (also called leaves).

Decision trees are induced on the basis of iterative splitting of data from the training set into discrete groups, where the goal is to minimize the 'distance' among groups at each split. Therefore one of the distinctions among decision tree algorithms is the splitting criteria. The most commonly used splitting criteria are: information gain and information gain ratio used by Quinlan [38] in ID3, C4.5 and C5.0 algorithms; chi square goodness of fit [46] used in CHAID (SPSS answer tree) [29]; gini index used by Breiman in CART algorithm [7]; j-measure or cross-entropy introduced by Smyth and Goodman [45], etc.

Decision trees models are extensively used in data mining for prediction purposes such as decision support and knowledge extraction in various fields. Decision trees provide very important advantage compared to many of other data mining models - the possibility of explaining the decisions in a way understandable by humans. A main disadvantage of classic decision tree induction is sensitivity to noise (missing or corrupted data).

Decision trees are able to process both numerical and categorical variables. However, on the basis of the type of prediction variable we distinguish between **classification trees** used to predict categorical variables, and **regression trees** for continuous prediction variable.

Neural Networks

A neural network is a powerful model modeled after the human brain. As the human brain consist of millions of neurons interconnected by synapses, neural networks are formed from large number of artificial neurons (nodes) connected to each other to form a network. It can be used to capture and represent complex relationships between prediction variables (input) and target variables (output) or to find patterns in data. Neural networks may be applied to classification problems (where the output is categorical variable) as well to regression (where the output variable is continuous).

A neural network (Fig. 5) starts with an **input layer**, where each node (neuron) corresponds to a predictor variable. Each input node is connected to every node in a **hidden layer**. The nodes in a hidden layer can be connected to nodes in another hidden layer or to an **output layer**. The output layer consists of one or more response variables. Each connection has a **weight** assigned which reflects strength of a connection between the connected neurons. Like in human brain the strength of a connection can

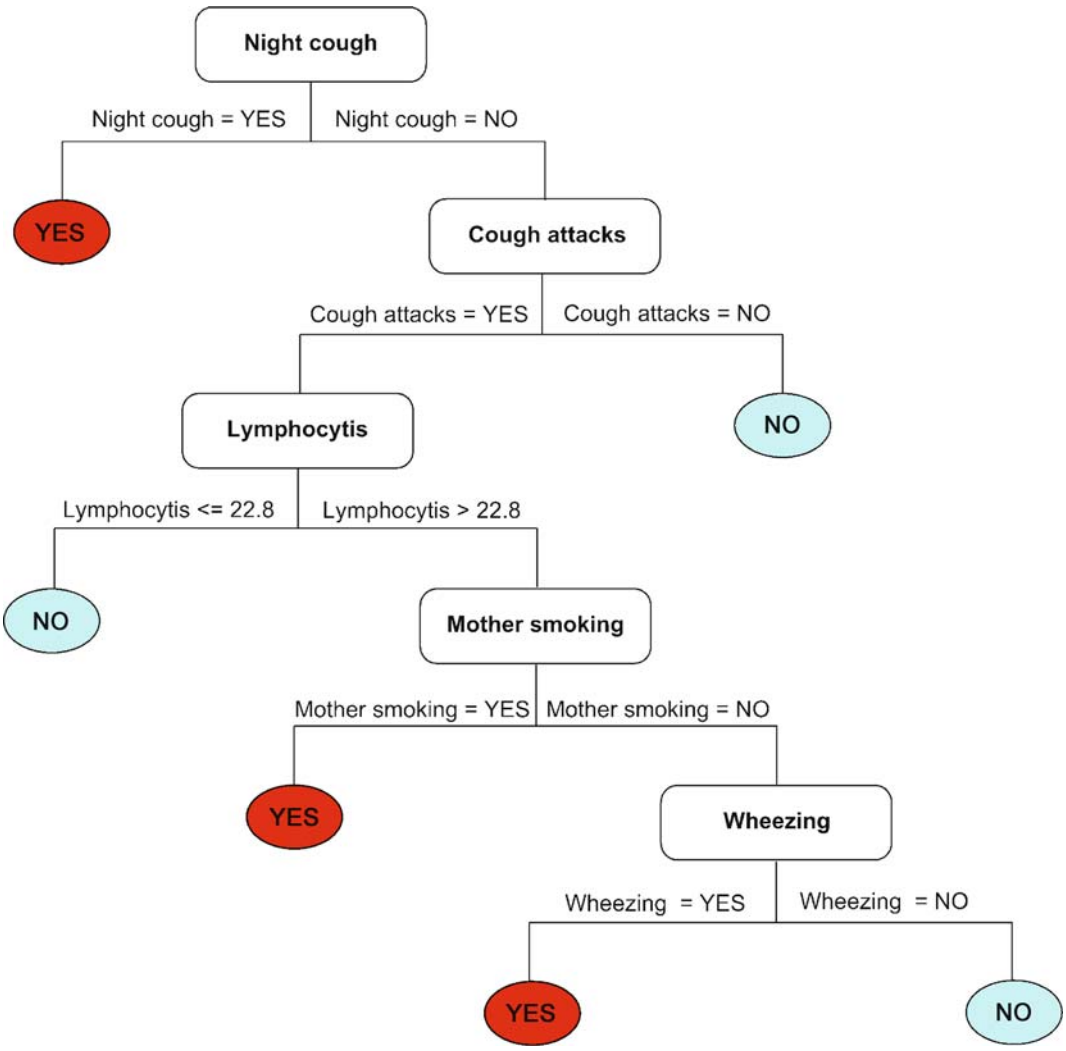

Discovery Systems, Figure 4
A simple decision tree

change in response to a presented stimulation or an obtained output, which enables the network to learn. After the input layer, each node (in the first hidden layer) gets a set of inputs which are then multiplied with connection weights and summed together to form a linear combination of values of nodes from the previous layer. An activation function is applied to the linear combination of values and the output is passed to the node(s) in the next layer.

In the beginning of model building (learning a network) the connection weights are unknown parameters which have to be learned on the basis of training data using one of the training algorithms. One of the most known training algorithms is the backpropagation algorithm [42,51], however newer methods include genetic algorithms [41], quasi-Newton [15], conjugate gradient [49], and many others.

The mayor advantages of neural networks are: good results on large databases, low sensitivity to noise and data distribution (on large databases or when proper mechanisms for preventing overfitting are used), applicability to multivariate non-linear problems, easy to implement to run on parallel computers, applicability to several problem domains.

Neural networks also have some disadvantages that have to be mentioned. The complexity of neural network model makes it hard to interpret the results. They belong to a group of so called black box models which means that the relationships among prediction and target vari-

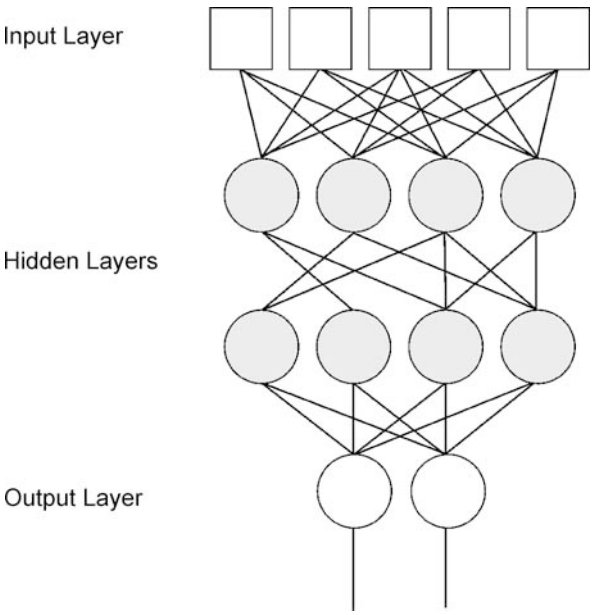

Discovery Systems, Figure 5
A neural network with two hidden layers

ables cannot be represented in humanly understandable (symbolic) form. Therefore they are often not applicable to problems that demand an explanation of induced models.

Second disadvantage of neural networks is their proneness to overfitting due to large number of adjustable parameters which can adopt to every data if the size of a neural network is large enough. In order to reduce overfitting some mechanisms such as weight decay and/or cross validation have to be used.

The next disadvantage is an extensive amount of computational time needed for training a network (especially on large databases). However, as stated before they can be run on massively parallel computers with each node simultaneously doing its own calculations.

Neural networks have been successfully applied to a broad spectrum of data mining applications, such as: character recognition for optical character recognition (OCR) applications, target recognition used in military applications, machine diagnosis, medical diagnosis for image data (MRIs, X-rays), voice recognition, intelligent searching for internet search engines, fraud detection, quality control, etc.

Ensemble Approaches

In recent years there has been a growing interest in the area of combining classifiers into ensembles also known as **committees** or **multiple classifier systems** (Fig. 6). The intuitive concept of ensemble approaches is that no single classifier can claim to be uniformly superior to any other, and that the integration of several single approaches will enhance the performance of final classification [53]. Hence, using the classification capabilities of multiple classifiers, where each classifier may make different and perhaps complementary errors, tend to yield an improved performance over single classifiers. Some ensemble approaches are actively trying to perturb some aspects of the training set, such as training samples, attributes or classes, in order to ensure classifier diversity. One of the most

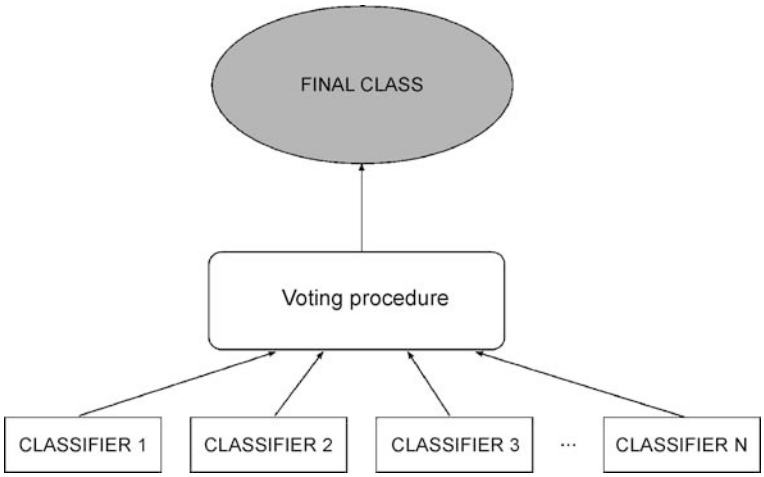

Discovery Systems, Figure 6
Combining classifiers into an ensemble

popular perturbation approaches are bootstrap aggregation (bagging) and boosting.

Bagging was first introduced by Breimen [6] in 1996 manipulates the training samples and forms replicate training sets. The final classification is based on a majority vote. Boosting, introduced by Freund and Schapire in 1996 [24] combines classifiers with weighted voting and is more complex since the distribution of training samples in training set is adaptively changed according to the performance of sequentially constructed classifiers.

In general ensemble approaches can be divided into three groups:

(1) Ensemble approaches that combine different independent classifiers (such as: Bayesian voting, majority voting, etc.),
(2) Ensemble learning approaches which construct a set of classifiers on the basis of one base classifier with perturbation of a training set (such as: bagging, boosting, windowing, etc.) and
(3) A combination of (1) and (2).

A detailed empirical study is presented in [19].

Hybrid Approaches

The hybrid approaches rest on the assumption that only in the synergetic combination of single models can unleash their full power [28]. Each of the single model types has its advantages, but also inherent limitations and disadvantages, which must be taken into account when using the particular model type. Therefore the logical step is to combine different models (classic approaches) to overcome the disadvantages and limitations of a model.

According to the way of combining different models into a hybrid, four basic types of hybrids can be differentiated:

Sequential hybrid the inputs in the hybrid are fed as inputs in model 1. The outputs from model 1 represent inputs for model 2 and the outputs from model 2 are also outputs of the sequential hybrid (Fig. 7a). Both models in the hybrid are autonomous and can be also used independently.

External hybrid model 1 represents the main method which forwards the data to model 2. Model 2 processes the data and sends the results back to model 1 (Fig. 7b). Model 2 totally depends on model 1 and its outputs cannot be used as hybrid outputs.

Embedded hybrid has the most powerful connection between both models which are so strong interlaced, that none of them can function independently (Fig. 7c).

Parallel hybrid models in the parallel hybrid are totally independent. They both operate on the same inputs, but produce separate outputs. The outputs of the hybrid are determined within the arbitrary mechanism, embedded in the hybrid (Fig. 7d).

These four types of hybrids differ in type and strength of connection between the embedded methods. Such hybrids can also be building blocks for making more complex hybrids.

For example, in 1999 Boz [8] presented a method for converting a trained backpropagation neural network into a decision tree. In this approach the neural network was used to determine the list of attributes that influence the outcomes of the network the most. That list of attributes

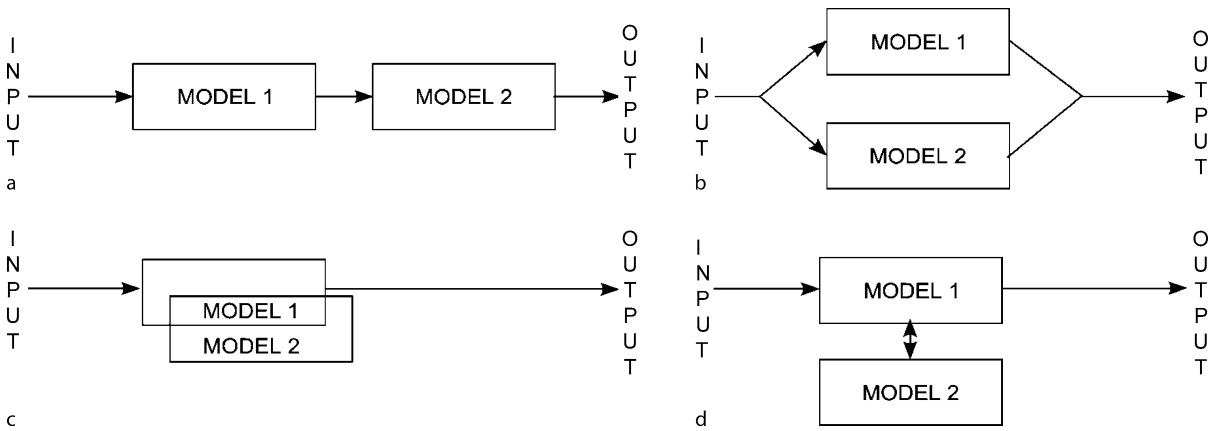

Discovery Systems, Figure 7
Basic types of hybrids: a sequential hybrid, b external hybrid, c embedded hybrid and d parallel hybrid

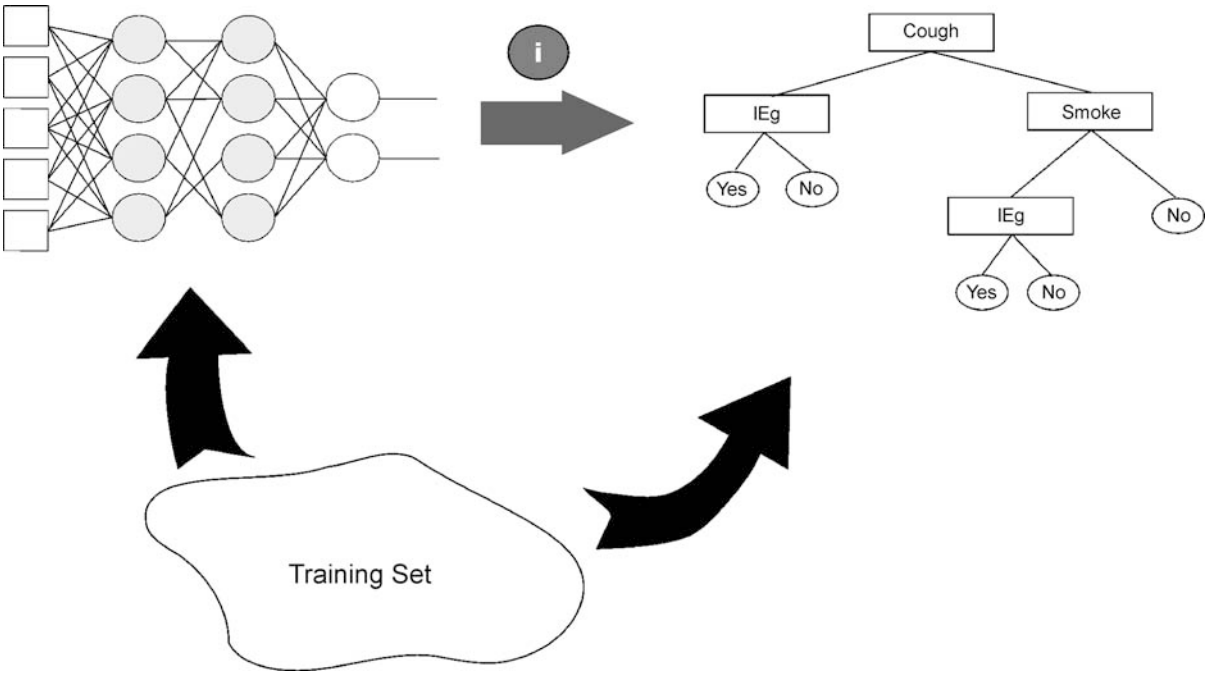

Discovery Systems, Figure 8
External hybrid of a decision tree and neural network

was then used in the process of the decision tree induction instead of the splitting criteria [Fig. 8].

Interpretation and Evaluation of Results

Evaluation of an induced model is a very important step of KDDM process. Testing of an induced model is crucial for evaluation. For that purpose, an initial data set is usually randomly split into two data sets: a training set (70% of randomly chosen samples) and a test set (30% of samples). When a training set is large it can be divided into two sets: a training set and a validation set. A validation set is used to assess how well the model fits the data, to adjust the model, or to select the best model among those that have been validated.

Beside random selection of training and testing samples from the database, the distribution of classes in training set is also very important for effective learning. In most cases it is best to preserve the natural distribution of classes (as it appears in initial data set) in the training set. However, in real world problems (especially from the medical field) we often come across the data sets with very unequal distribution of classes (for example: less than 10% of samples belong to a specific class). In cases like that the use of natural distribution is not appropriate simply because there are not enough training samples to effectively learn a model of basic concepts. Equalization of class distribution is therefore a better solution [50].

The quality of the induced model is most commonly evaluated with the **accuracy rate**. The accuracy rate is calculated separately on the training set and the test set. Training accuracy indicates the extent of concepts learned, whereas the accuracy of correctly classified samples from test set provides us more information about the predictive abilities of a model and its applicability to practice. Accuracy is calculated as follows:

$$\text{accuracy} = \frac{\text{num of correctly classified samples}}{\text{num of all samples}} .$$

Unfortunately, the accuracy by itself is not necessarily the right metric for selecting the best model. As mentioned before, the distribution of target variable classes in real world data sets is often very biased. In those cases we can obtain a model with high accuracy; however the accuracy of classifying the samples of non-dominant class can be catastrophically low. Therefore use of other metrics that expose more information of the types of errors and the costs associated with them are a necessity.

Average class accuracy is one of the simple metrics that gives more information about model's quality. The classification accuracy is calculated separately for each class of a target variable and the average across all classes

Discovery Systems, Table 2
A sample confusion matrix for three classes

Predicted	*Actual*		
	Class 1	Class 2	Class 3
Class 1	30	2	0
Class 2	10	25	3
Class 3	5	7	43

is calculated:

$$\text{average class accuracy} = \frac{\sum_c \text{accuracy}_c}{\text{num of classes}},$$

where c is one of the possible classes and accuracy_c is:

$$\text{accuracy}_c = \frac{\text{num of correctly classified objects in class } c}{\text{num of all objects in class } c}.$$

When accuracy is equal or similar to average class accuracy, one can speculate that the model does not have problems with classification of samples belonging to a specific class. Knowing the accuracies for each class of target variable gives us additional information about problems in the database or model.

For classification problems a **confusion matrix** [37] is a very useful tool for understanding the results of testing a model. A confusion matrix shows the results of actual versus predicted class values. Table 2 presents a sample of a confusion matrix where the columns show actual classes and the rows show predicted classes. Therefore the diagonal shows all correctly classified samples (98 samples out of 125). Class accuracies can also be calculated. For example, 30 samples out of 45 belonging to Class 1 were correctly classified (class accuracy for Class 1 is 67%); 10 samples were misclassified as Class 2 and 5 samples were misclassified as Class 3).

The confusion matrix for binary classification can be interpreted with two commonly used matrices: **specificity** and **sensitivity**. They are most often used to evaluate a performance of medical test. Sensitivity is the proportion of samples that were classified as positive of all the positive samples tested. In the case of classifying people with illness it can be seen as *the probability that the test is positive given that the patient is sick*. Consequently, the higher the sensitivity, the fewer sick people are undetected.

Specificity is the proportion of samples that were classified as negative of all the negative samples tested. In the case of classifying people with illness it can be seen as *the probability that the test is negative given that the patient is not sick*. Consequently, the higher the specificity, the fewer healthy people are classified as sick.

Discovery Systems, Table 3
A binary confusion matrix

Predicted	*Actual*	
	Positive	Negative
Positive	**TP** (*True Positive*)	**FP** (*False Positive*)
Negative	**FN** (*False Negative*)	**TN** (*True Negative*)

From the definitions above we can conclude the following: *Sensitivity and specificity are inversely related (higher sensitivity leads to lower specificity).*

Sensitivity and specificity are calculated as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}$$

where the notations are explained in Table 3.

The relationship between specificity and sensitivity can be visualized with the **ROC graph** [5].

ROC (receiver operating characteristics) graph [5] is a two dimensional plot with the false positive rate (1-specificity) on the x-axis and the true positive rate (sensitivity) on y-axis (Fig. 9). The best possible prediction model would yield a point (0,1) in the upper left corner of ROC graph, which represents 100% sensitivity (all true positive samples correctly classified) and 100% specificity (no false positive samples classified). On the contrary, the point (1,0) in the lower right corner of ROC graph represents the worst possible prediction model with 0% sensitivity and

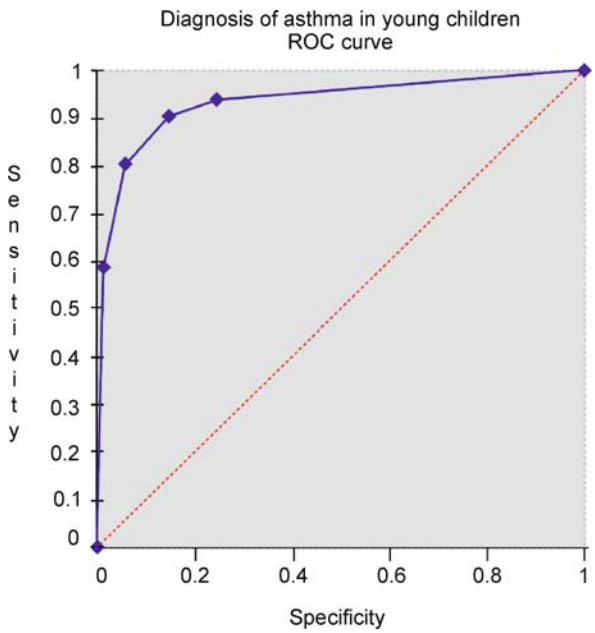

Discovery Systems, Figure 9
ROC graph

0% specificity. A model with completely random classification would give a point along diagonal line from the left bottom to the top right corner ($y = x$). The diagonal line divides the ROC space into two areas: points above diagonal line represent good models and points below diagonal line represent bad models. A more precise way of characterizing this is simply to calculate the **area under the ROC curve** [3]. The closer the area is to 0.5, the worse is the model, and the closer it is to 1.0, the better is the model.

All presented evaluation techniques express the quality of induced model in dependence of data gathered in the data set. However, no matter how well the model performs on data sets used for testing, there is no 100% guarantee that the model will perform well in the real world. One of the reasons can be found in the quality of data used to build the model (a lot of noise, missing data, implicit assumptions about choosing prediction variables, etc.). When the training data fails to match the real world, the model can learn incorrect concepts and will therefore perform very poor in the practice. For that purpose it is very important to test the model in the real world. For example: if the model is used for diagnosing asthma in young children, it should be tested on a small set of symptomatic children.

An evaluation of a model by a **domain expert** (usually also an end-user) is also very important before presenting the model to the real world. In that case symbolic models (such as: decision rules, decision trees, etc.) have essential advantage over the so called black-box models (such as: neural networks, support vector machines, etc.) – the possibility to interpret the relationships among prediction and target variables.

Data mining extracts already known and expected information from the database, but on the other hand, useful relationships between variables that are non-intuitive are the jewels that data mining hopes to find. In order to enable a domain expert to evaluate the model and possibly to extract some new knowledge, it is very important to choose a proper visualization of a model. This requires understanding end-user's needs. Visualizing a model allows a user to discuss and explain the logic behind the model with colleagues, customers, and other users. It also builds the user's trust in the results. Symbolic models are relatively easy to visualize, however black-box models can pose more problems but novel solutions are starting to appear.

Utilization of Results

Once a data mining model is built and evaluated it can be used in one of the following ways: (1) analyzing a model and extracting knowledge or (2) applying a model in practice for a decision support.

In (1) the analyst (usually a domain expert) studies the model and the results in order to extract some new knowledge about the relationships among variables (for example: comparison of different clusters, interpretation of induced rules, etc).

In (2) the model is applied into practice and used on new data in order to perform some classification or regression tasks. Prediction models can be used for decision support independently. For instance, a model for diagnosing asthma in young children can support the physician in making a diagnosis. However, usually models are used in some business process and therefore they are most commonly incorporated into business applications. For instance, a model for predicting a risky loan applicant can be integrated into the loan application.

Some prediction models turn out to be very time and recourse consuming in practical use (for example: monitoring credit card transactions, fraud detections, etc). In those cases parallel systems can provide an effective solution.

Since all practical knowledge and experiences of domain experts are often not captured in prediction models built in data mining process, they may be used in combination with a model and as such applied into practice. Usually a combination of expert knowledge and knowledge discovered in data mining process can be very effective.

Since we are all aware that over time all systems evolve (all things are changing – nothing is static) it is very important to continuously monitor the performance of a model used in practice. When a significant reduction in performance of a model is detected, the model has to be retrained or sometimes even completely rebuilt.

Knowledge Discovery Frameworks and Tools

As already stated, knowledge discovery is a process of searching through large amounts of data and picking out relevant information. It is very often used in science but has become a very important process in the business and other areas of everyday life, too. Due to a huge demand for knowledge discovery and data mining tools, there are a lot of different tools and frameworks that are becoming available over the Internet. Most of them are coming from academic circles and can be obtained on open source or general public license basis. In this chapter we mostly focus on knowledge discovery and data mining frameworks and tools that are freely available and are used by a wide range of users from different fields.

The following tools that cover the biggest share of KD tools users will be presented:

- Weka/Pentaho
- YALE/Rapid-i
- Orange
- Tanagra.

Weka/Pentaho

Originally created in 1993, the Weka project [52] was established by the University of Waikato as a platform for research and testing of advanced machine learning techniques. Since then, Weka has developed a large and loyal following in both academic and industry circles. It is so popular because of easy tailoring and extending of powerful analytical techniques with advanced visualization and industry-specific algorithms. In 2006 it was acquired by Pentaho [20], which is an even larger suite of tools that provides a full spectrum of open source business intelligence (BI) capabilities including reporting, analysis, dashboards, data mining, data integration, and a BI platform that have made it the world's most popular open source BI suite. Pentaho is also the primary sponsor and owner of popular open source projects including Mondrian, JFreeReport, Kettle, and Weka. By this acquisition, Weka got even wider range of users and established itself in the BI knowledge discovery circles.

YALE/RapidMiner

Similar to Weka, YALE [33] was also transformed into a more commercial, but still freely available set of tools for machine learning based tasks that can be used in knowledge discovery process. Both have their roots in academic circles and are based on most recent research papers. YALE was renamed into RapidMiner after the acquisition of YALE by Rapid-I in May 2007.

YALE/RapidMiner is a rapid prototyping system for knowledge discovery and data mining. With more than 400 data mining operators it is one of the most comprehensive open-source data mining tools. It is widely used by a large number of organizations covering a wide range of different branches.

As the main focus of the YALE authors has been flexibility, it is reasonable to expect most advantage for YALE in this area. But flexibility in creating experiments, reusing previous work and preprocessing data to work on, were main advantages only in the early versions of YALE compared to Weka. However, visualization of results is still better supported in YALE.

Orange

Orange [16] is component-based data mining software. It includes a range of preprocessing, modeling and data exploration techniques. It is based on C++ components, that are accessed either directly, through Python scripts, or through GUI objects called Orange Widgets.

Orange core objects and Python modules support various data mining tasks that span from data preprocessing to modeling and evaluation. Among other it supports techniques for:

- Data input, providing the support for various popular data formats,
- Data manipulation and preprocessing, like sampling, filtering, scaling, discretization, construction of new attributes, and alike,
- Methods for development of classification models, including classification trees, naïve Bayesian classifier, instance-based approaches, logistic regression and support vector machines,
- Regression methods, including linear regression, regression trees, and instance-based approaches,
- Various wrappers, like those for calibration of probability predictions of classification models,
- Ensemble approaches, like boosting and bagging,
- Various state-of-the-art constructive induction methods, including function decomposition,
- Association rules and data clustering methods,
- Evaluation methods, different hold-out schemes and range of scoring methods for prediction models including classification accuracy, AUC, Brier score, and alike. Various hypothesis testing approaches are also supported,
- Methods to export predictive models to PMML.

Orange is based on C++, which could be the main advantage when fast knowledge discovery tools are needed in comparison to before mentioned Java based frameworks.

Tanagra

Tanagra [40] is a free data mining and knowledge discovery software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. In comparison to other presented tools, Tanagra contains more statistical functions than other frameworks, but is aimed mainly to academic circles. The main purpose of Tanagra project is to give researchers and students an easy-to-use data mining software, conforming to the present standards of the software development in this domain (especially in the design of its graphical user

interface and the way in which it is used), and allowing one to analyze either real or synthetic data.

Another approach to data analysis is R [13], which is heavily used for statistical analysis of data and also features a number of machine learning algorithms. Being very efficient with handling data, it lacks a graphical user interface for building projects and uses a rather nonintuitive syntax one has to get used to. Using the extensions that are constantly becoming available one can expect that R will feature most of the knowledge discovery functionality in the future.

Conclusions and Future Directions

The aggressive rate of data that is stored on modern computer systems has far outpaced our ability to process and analyze this data. This way data is often deposited to merely rest in peace, never to be accessed again. Novel applications of knowledge discovery systems in supporting scientific discoveries, business exploitation or complex decision making, have awaken the growing commercial interest in knowledge discovery and data mining techniques. That has stimulated new interest in automatic knowledge extraction from examples stored in large databases.

Knowledge discovery can be broadly defined as automated discovery of novel and useful information from databases. Unfortunately there is no method or formal model that would be able to define the optimal knowledge discovery solution that should be applied for a certain problem. This problem will become even more important in the future as a growing amount of collected data will not allow time costs that are caused by empirical evaluation of all possible knowledge discovery techniques to find the most appropriate one. The future directions and possible solutions for this problem were defined in [34] where different meta-learning techniques along with a knowledge-driven framework for a KD system that contains a limited number of data mining techniques are presented.

Another interesting direction for future discoveries is application of KD techniques in the field of structured knowledge or textual documents that can be transformed into formally structured form. This field of research is heading toward advanced techniques from recent text mining research like sentiment analysis and tries to combine them with classical data mining techniques. This will only be possible when we will be able to transform unstructured text into machine readable form that would formalize parts of text and extract the underlying messages in the text.

Despite the rapid growth, KD systems are still evolving and will represent an important part of the future intelligent systems in various fields. It is obvious that KD systems are also becoming an important part of the large business systems that are fortunately available through different forms of open source initiatives that can help in their sustainable development for the future.

Bibliography

Primary Literature

1. Anand S, Buchner A (1998) Decision support using data mining. Financial Time Management, London
2. Baeck T (1996) Evolutionary algorithms in theory and practice. Oxford University Press, New York
3. Barley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 30(7):1145–1159
4. Becerra-Fernandez I, Gonzalez A, Sabherwal R (2004) Knowledge management: Challenges, solutions, and technologies. Prentice Hall, Upper Saddle River
5. Beck JR, Shultz E (1986) The use of relative operating characteristic (ROC) curves in test performance evaluation. Arch Pathol Lab Med 110:13–20
6. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
7. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth International Group, Belmont
8. Boz O (2000) Converting a trained neural network to a decision tree dectext – decision tree extractor. Ph D thesis, Computer Science and Engineering, Lehigh University. http://citeseer.ist.psu.edu/boz00converting.html. Accessed 12 Nov 2007
9. Cabena P, Hadjinian P, Stadler R, Verhees J, Zanasi A (1998) Discovering data mining: From concepts to implementation. Prentice Hall, Upper Saddle River
10. Caspase Drug Discovery Systems. drug discovery system. http://www.biomol.com/Online_Catalog/Online_Catalog/Products/36/?categoryId=420. Accessed 6 Nov 2007
11. Cios K, Teresinska A, Konieczna S, Potocka J, Sharma S (2000) Diagnosing myocardial perfusion from PECT bull's-eye maps – a knowledge discovery approach. IEEE Eng Med Biol Mag, Special Issue Med Data Mining Knowl Discov 19(4):17–25
12. Cios KJ, Pedrycz W, Swiniarski RW, Kurgan LA (2007) Data mining. A knowledge discovery approach. Springer, New York
13. Dalgaard P (2002) Introductory statistics with R. Springer, New York
14. Davenport TH, Prusak L (1997) Information ecology: Mastering the information and knowledge environment. Oxford University Press, New York
15. Dennis JE Jr, Schnabel RB (1989) A view of unconstrained optimization. In: Nemhauser GL, Runnooy Kan AHG, Todd MJ (eds) Handbook in operations research and management science, vol 1 Optimization. Elsevier, Amsterdam
16. Demsar J, Zupan B (2004) Orange: From experimental machine learning to interactive data mining. White Paper. Faculty of Computer and Information Science, University of Ljubljana. http://www.ailab.si/orange
17. Developmental Discovery System (TM). Developmental discovery system. http://www.gotofocus.com/. Accessed 6 Nov 2007

18. Dictionary.com Unabridged (v 1.1). discover. http://dictionary.reference.com/browse/discover. Accessed 5 Nov 2007
19. Dietterich TG (2000) Ensemble methods in machine learning. In: First International Workshop on Multiple Classifier Systems. Lecture Notes in Computer Science. Springer, New York, pp 1–15
20. Dixon J (2005) Pentaho Open Source Business Intelligence Platform Technical White Paper. Pentaho Corporation, Orlando. http://sourceforge.net/project/showfiles.php?group_id=140317
21. Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases (a survey). AI Mag 17(3):37–54
22. Fayyad U, Piatesky-Shapiro G, Smyth P, Uthurusamy R (eds) (1996) Advances in knowledge discovery and data mining. AAAI Press, Menlo Park
23. Frawley W, Piatesky-Shapiro G, Matheus C (1991) Knowledge discovery in databases: An overview. In: Piatesky-Shapiro G, Frowley W (eds) Knowledge Discovery in Databases. AAAI/MIT Press, pp 1–27, Menlo Park
24. Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: Proceedings Thirteenth International Conference on Machine Learning. Morgan Kaufman, San Francisco, pp 148–156
25. Goldberg DE (1989) Genetic algorithms in search, optimization, and machine learning. Addison, Reading
26. Hand D, Mannila H, Smyth P (eds) (2001) Principles of data mining. MIT Press, Cambridge
27. Holland JH (1975) Adaptation in natural and artificial systems. MIT Press, Cambridge
28. Iglesias CJ (1996) The role of hybrid systems in intelligent data management: The case of fuzzy/neural hybrids. Control Eng Pract 4(6):839–845
29. Kass GV (1980) An exploratory technique for investigating large quantities of categorical data. Appl Stat 29:119–127
30. Kurgan L, Musilek P (2006) A survey of Knowledge Discovery and Data Mining process models. Knowl Eng Rev 21(1):1–24
31. Loh W, Shih Y (1997) Split selection methods for classification trees. Stat Sinica 7:815–840
32. Mannila H (2000) Theoretical frameworks of data mining. SIGKDD Explor 1:30–32
33. Mierswa I, Wurst M, Klinkenberg R, Scholz M, Euler T (2006) YALE: Rapid Prototyping for Complex Data Mining Tasks. In: Proc of the 12th ACMSIGKDD. International Conference on Knowledge Discovery and Data Mining, Philadelphia, pp 1–6
34. Pechenizkiy M, Tsymbal A, Puuronen S (2005) Meta-knowledge management in multistrategy process-oriented knowledge discovery systems. Technical Report, Dublin, Trinity College Dublin, Department of Computer Science, TCD-CS-2005–30, p 12
35. Piatetsky-Shapiro G (1991) Knowledge discovery in real databases: A report on the IJCAI-89 Workshop. AI Mag 11(5): 68–70
36. Piatetsky-Shapiro G (1999) The data mining industry coming to age. IEEE Intel Syst 14(6):32–33
37. Provost F, Fawcett T, Kohavi R (1998) The case against accuracy estimation for comparing classifiers. In: Proceedings of the Fifteenth International Conference on Machine Learning, (ICML-98), San Francisco
38. Quinlan JR (1986) Induction of decision trees. In: Machine Learning, vol 1. Kluwer, Hingham
39. Quinlan R (1993) C4.5: Programs for machine learning. Morgan Kaufmann, San Francisco
40. Rakotomalala R (2005) TANAGRA: Un logiciel gratuit pour l'enseignement et la recherche. In: Proc of the 5th Journees d'Extraction et Gestion des Connaissances 2:697–702
41. Reeves CR (ed) (1993) Modern heuristic techniques for combinatorial problems. Wiley, New York
42. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536
43. Sano M, Katoa Y, Taira K (2005) Functional gene-discovery systems based on libraries of hammerhead and hairpin ribozymes and short hairpin RNAs. Mol Biosyst 1:27–35
44. Shearer C (2000) The CRISP-DM model: the new blueprint for data mining. J Data Wareh I5(4):13–19
45. Smyth P, Goodman RM (1991) Rule induction using information theory. In: Piatetsky-Schapiro G, Frawley WJ (eds) Knowledge Discovery in Databases. AAAI Press, pp 159–176, Menlo Park
46. Snedecor GW, Cochran WG (1989) Statistical methods, 8th edn. Iowa State University Press, Ames
47. Tan P, Steinbach M, Kumar V (2005) Introduction to data mining. Addison, Boston
48. The Discovery System. discovery system for personality profiling. http://www.insights.com/core/English/TheDiscoverySystem/default.shtm. Accessed 6 Nov 2007
49. Towsey M, Alpsan D, Sztriha L (1995) Training a neural network with conjugate gradient methods. IEEE Proc Neural Netw 1:373–378
50. Weiss GM, Provost F (2001) The effect of class distribution on classifier learning. Technical Report ML-TR 43, Department of Computer Science, Rutgers University
51. Werbos PJ (1994) The roots of backpropagation. Wiley, New York
52. Witten IH, Frank E (2005) Data mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco
53. Wolpert D, Macready W (1997) No free lunch theorems for optimization. IEEE Trans Evol Comput 1(1):67–82

Books and Reviews

Berthold M, Hand DJ (2003) Intelligent data analysis: An introduction, 2nd edn. Springer, New York

Lin TY, Ohsuga S, Liau CJ, Hu X, Tsumoto S (eds) (2005) Foundations of data mining and knowledge discovery. Studies in Computational Intelligence, vol 6. Springer, New York

Discrete Control Systems

Taeyoung Lee[1], Melvin Leok[2], Harris McClamroch[1]

[1] Department of Aerospace Engineering, University of Michigan, Ann Arbor, USA

[2] Department of Mathematics, Purdue University, West Lafayette, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Discrete Lagrangian and Hamiltonian Mechanics
Optimal Control of Discrete Lagrangian and Hamiltonian Systems
Controlled Lagrangian Method for Discrete Lagrangian Systems
Future Directions
Acknowledgments
Bibliography

Glossary

Discrete variational mechanics A formulation of mechanics in discrete-time that is based on a discrete analogue of *Hamilton's principle*, which states that the system takes a trajectory for which the action integral is stationary.

Geometric integrator A numerical method for obtaining numerical solutions of differential equations that preserves geometric properties of the continuous flow, such as symplecticity, momentum preservation, and the structure of the configuration space.

Lie group A differentiable manifold with a group structure where the composition is differentiable. The corresponding *Lie algebra* is the tangent space to the Lie group based at the identity element.

Symplectic A map is said to be symplectic if given any initial volume in phase space, the sum of the signed projected volumes onto each position-momentum subspace is invariant under the map. One consequence of symplecticity is that the map is volume-preserving as well.

Definition of the Subject

Discrete control systems, as considered here, refer to the control theory of discrete-time Lagrangian or Hamiltonian systems. These discrete-time models are based on a discrete variational principle, and are part of the broader field of geometric integration. Geometric integrators are numerical integration methods that preserve geometric properties of continuous systems, such as conservation of the symplectic form, momentum, and energy. They also guarantee that the discrete flow remains on the manifold on which the continuous system evolves, an important property in the case of rigid-body dynamics.

In nonlinear control, one typically relies on differential geometric and dynamical systems techniques to prove properties such as stability, controllability, and optimality. More generally, the geometric structure of such systems plays a critical role in the nonlinear analysis of the corresponding control problems. Despite the critical role of geometry and mechanics in the analysis of nonlinear control systems, nonlinear control algorithms have typically been implemented using numerical schemes that ignore the underlying geometry.

The field of discrete control systems aims to address this deficiency by restricting the approximation to the choice of a discrete-time model, and developing an associated control theory that does not introduce any additional approximation. In particular, this involves the construction of a control theory for discrete-time models based on geometric integrators that yields numerical implementations of nonlinear and geometric control algorithms that preserve the crucial underlying geometric structure.

Introduction

The dynamics of Lagrangian and Hamiltonian systems have unique geometric properties; the Hamiltonian flow is symplectic, the total energy is conserved in the absence of non-conservative forces, and the momentum maps associated with symmetries of the system are preserved. Many interesting dynamics evolve on a non-Euclidean space. For example, the configuration space of a spherical pendulum is the two-sphere, and the configuration space of rigid body attitude dynamics has a Lie group structure, namely the special orthogonal group. These geometric features determine the qualitative behavior of the system, and serve as a basis for theoretical study.

Geometric numerical integrators are numerical integration algorithms that preserve structures of the continuous dynamics such as invariants, symplecticity, and the configuration manifold (see [14]). The exact geometric properties of the discrete flow not only generate improved qualitative behavior, but also provide accurate and efficient numerical techniques. In this article, we view a geometric integrator as an intrinsically discrete dynamical system, instead of concentrating on the numerical approximation of a continuous trajectory.

Numerical integration methods that preserve the simplecticity of a Hamiltonian system have been studied (see [28,36]). Coefficients of a Runge–Kutta method are carefully chosen to satisfy a simplecticity criterion and order conditions to obtain a symplectic Runge–Kutta method. However, it can be difficult to construct such integrators, and it is not guaranteed that other invariants of the system, such as a momentum map, are preserved. Alternatively, variational integrators are constructed by dis-

cretizing Hamilton's principle, rather than discretizing the continuous Euler–Lagrange equation (see [31,34]). The resulting integrators have the desirable property that they are symplectic and momentum preserving, and they exhibit good energy behavior for exponentially long times (see [2]). Lie group methods are numerical integrators that preserve the Lie group structure of the configuration space (see [18]). Recently, these two approaches have been unified to obtain Lie group variational integrators that preserve the geometric properties of the dynamics as well as the Lie group structure of the configuration space without the use of local charts, reprojection, or constraints (see [26,29,32]).

Optimal control problems involve finding a control input such that a certain optimality objective is achieved under prescribed constraints. An optimal control problem that minimizes a performance index is described by a set of differential equations, which can be derived using Pontryagin's maximum principle. Discrete optimal control problems involve finding a control input for a discrete dynamic system such that an optimality objective is achieved with prescribed constraints. Optimality conditions are derived from the discrete equations of motion, described by a set of discrete equations. This approach is in contrast to traditional techniques where a discretization appears at the last stage to solve the optimality condition numerically. Discrete mechanics and optimal control approaches determine optimal control inputs and trajectories more accurately with less computational load (see [19]). Combined with an indirect optimization technique, they are substantially more efficient (see [17,23,25]).

The geometric approach to mechanics can provide a theoretical basis for innovative control methodologies in geometric control theory. For example, these techniques allow the attitude of satellites to be controlled using changes in its shape, as opposed to chemical propulsion. While the geometric structure of mechanical systems plays a critical role in the construction of geometric control algorithms, these algorithms have typically been implemented using numerical schemes that ignore the underlying geometry. By applying geometric control algorithms to discrete mechanics that preserve geometric properties, we obtain an exact numerical implementation of the geometric control theory. In particular, the method of controlled Lagrangian systems is based on the idea of adopting a feedback control to realize a modification of either the potential energy or the kinetic energy, referred to as potential shaping or kinetic shaping, respectively. These ideas are applied to construct a real-time digital feedback controller that stabilizes the inverted equilibrium of the cart-pendulum (see [10,11]).

In this article, we will survey discrete Lagrangian and Hamiltonian mechanics, and their applications to optimal control and feedback control theory.

Discrete Lagrangian and Hamiltonian Mechanics

Mechanics studies the dynamics of physical bodies acting under forces and potential fields. In Lagrangian mechanics, the trajectory of the object is derived by finding the path that extremizes the integral of a Lagrangian over time, called the action integral. In many classical problems, the Lagrangian is chosen as the difference between kinetic energy and potential energy. The Legendre transformation provides an alternative description of mechanical systems, referred to as Hamiltonian mechanics.

Discrete Lagrangian and Hamiltonian mechanics has been developed by reformulating the theorems and the procedures of Lagrangian and Hamiltonian mechanics in a discrete time setting (see, for example, [31]). Therefore, discrete mechanics has a parallel structure with the mechanics described in continuous time, as summarized in Fig. 1 for Lagrangian mechanics. In this section, we describe discrete Lagrangian mechanics in more detail, and we derive discrete Euler–Lagrange equations for several mechanical systems.

Consider a mechanical system on a configuration space Q, which is the space of possible positions. The Lagrangian depends on the position and velocity, which are elements of the tangent bundle to Q, denoted by TQ. Let $L\colon TQ \to \mathbb{R}$ be the Lagrangian of the system. The discrete Lagrangian, $L_{\mathrm{d}}\colon Q\times Q \to \mathbb{R}$ is an approximation to the exact discrete Lagrangian,

$$L_{\mathrm{d}}^{\mathrm{exact}}(q_0, q_1) = \int_0^h L(q_{01}(t), \dot{q}_{01}(t))\,\mathrm{d}t\,, \tag{1}$$

where $q_{01}(0) = q_0$, $q_{01}(h) = q_1$, and $q_{01}(t)$ satisfies the Euler–Lagrange equation in the time interval $(0, h)$. A discrete action sum $\mathfrak{G}_{\mathrm{d}}\colon Q^{N+1} \to \mathbb{R}$, analogous to the action integral, is given by

$$\mathfrak{G}_{\mathrm{d}}(q_0, q_1, \ldots, q_N) = \sum_{k=0}^{N-1} L_{\mathrm{d}}(q_k, q_{k+1})\,. \tag{2}$$

The discrete Hamilton's principle states that

$$\delta\mathfrak{G}_{\mathrm{d}} = 0$$

for any δq_k, which yields the *discrete Euler–Lagrange (DEL)* equation,

$$D_2L_{\mathrm{d}}(q_{k-1}, q_k) + D_1L_{\mathrm{d}}(q_k, q_{k+1}) = 0\,. \tag{3}$$

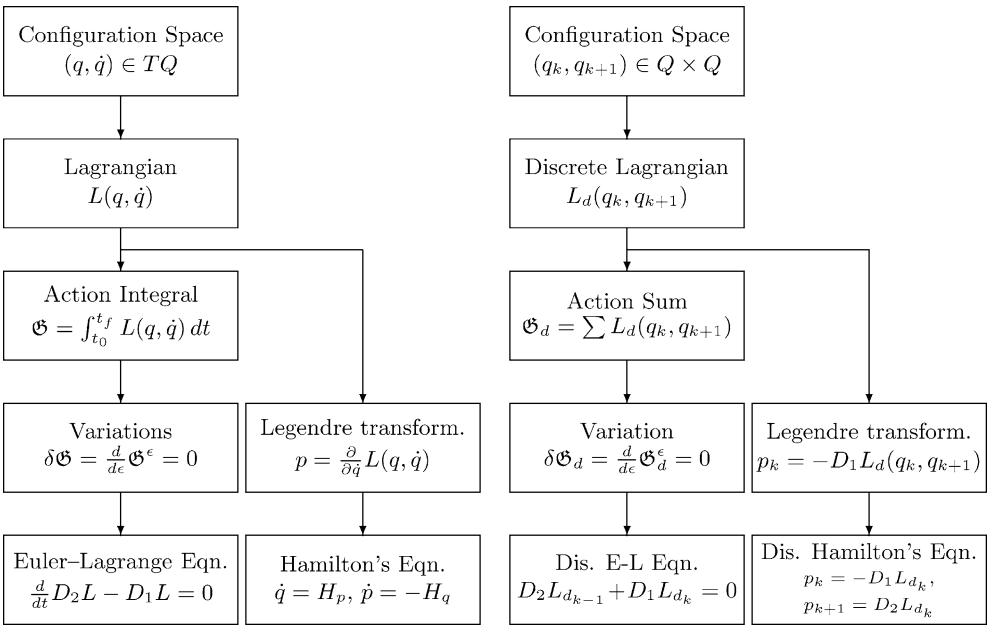

Discrete Control Systems, Figure 1
Procedures to derive continuous and discrete equations of motion

This yields a discrete Lagrangian flow map $(q_{k-1}, q_k) \mapsto (q_k, q_{k+1})$. The discrete Legendre transformation, which from a pair of positions (q_0, q_1) gives a position-momentum pair $(q_0, p_0) = (q_0, -D_1 L_d(q_0, q_1))$ provides a discrete Hamiltonian flow map in terms of momenta.

The discrete equations of motion, referred to as variational integrators, inherit the geometric properties of the continuous system. Many interesting Lagrangian and Hamiltonian systems, such as rigid bodies evolve on a Lie group. Lie group variational integrators preserve the nonlinear structure of the Lie group configurations as well as geometric properties of the continuous dynamics (see [29,32]). The basic idea for all Lie group methods is to express the update map for the group elements in terms of the group operation,

$$g_1 = g_0 f_0 \,, \tag{4}$$

where $g_0, g_1 \in G$ are configuration variables in a Lie group G, and $f_0 \in G$ is the discrete update represented by a right group operation on g_0. Since the group element is updated by a group operation, the group structure is preserved automatically without need of parametrizations, constraints, or re-projection. In the Lie group variational integrator, the expression for the flow map is obtained from the discrete variational principle on a Lie group, the same procedure presented in Fig. 1. But, the infinitesimal variation of a Lie group element must be carefully expressed to respect the structure of the Lie group. For example, it can be expressed in terms of the exponential map as

$$\delta g = \left.\frac{\mathrm{d}}{\mathrm{d}\epsilon}\right|_{\epsilon=0} g \exp \epsilon\eta = g\eta \,,$$

for a Lie algebra element $\eta \in \mathfrak{g}$. This approach has been applied to the rotation group SO(3) and to the special Euclidean group SE(3) for dynamics of rigid bodies (see [24,26,27]). Generalizations to arbitrary Lie groups gives the generalized *discrete Euler–Poincaré (DEP)* equation,

$$\begin{aligned} T_e^* L_{f_0} \cdot D_2 L_\mathrm{d}(g_0, f_0) - \mathrm{Ad}_{f_0}^* \cdot \left(T_e^* L_{f_1} \cdot D_2 L_\mathrm{d}(g_1, f_1)\right) \\ + T_e^* L_{g_1} \cdot D_1 L_\mathrm{d}(g_1, f_1) = 0 \,, \end{aligned} \tag{5}$$

for a discrete Lagrangian on a Lie group, $L_\mathrm{d} \colon G \times G \to \mathbb{R}$. Here $L_f \colon G \to G$ denotes the left translation map given by $L_f g = fg$ for $f, g \in G$, $T_g L_f \colon T_g G \to T_{fg} G$ is the tangential map for the left translation, and $\mathrm{Ad}_g \colon \mathfrak{g} \to \mathfrak{g}$ is the adjoint map. A dual map is denoted by a superscript * (see [30] for detailed definitions).

We illustrate the properties of discrete mechanics using several mechanical systems, namely a mass-spring system, a planar pendulum, a spherical pendulum, and a rigid body.

Example 1 (Mass-spring System) Consider a mass-spring system, defined by a rigid body that moves along a straight frictionless slot, and is attached to a linear spring.

Continuous equation of motion: The configuration space is $Q = \mathbb{R}$, and the Lagrangian $L\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ is given by

$$L(q, \dot{q}) = \frac{1}{2}m\dot{q}^2 - \frac{1}{2}\kappa q^2 , \tag{6}$$

where $q \in \mathbb{R}$ is the displacement of the body measured from the point where the spring exerts no force. The mass of the body and the spring constant are denoted by $m, \kappa \in \mathbb{R}$, respectively. The Euler–Lagrange equation yields the continuous equation of motion.

$$m\ddot{q} + \kappa q = 0 . \tag{7}$$

Discrete equation of motion: Let $h > 0$ be a discrete time step, and a subscript k denotes the kth discrete variable at $t = kh$. The discrete Lagrangian $L_\mathrm{d}\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ is an approximation of the integral of the continuous Lagrangian (6) along the solution of (7) over a time step. Here, we choose the following discrete Lagrangian.

$$\begin{aligned} L_\mathrm{d}(q_k, q_{k+1}) &= hL\left(\frac{q_k + q_{k+1}}{2}, \frac{q_{k+1} - q_k}{h}\right) \\ &= \frac{1}{2h}m(q_{k+1} - q_k)^2 - \frac{h\kappa}{8}(q_k + q_{k+1})^2 . \end{aligned} \tag{8}$$

Direct application of the discrete Euler–Lagrange equation to this discrete Lagrangian yields the discrete equations of motion. We develop the discrete equation of motion using the discrete Hamilton's principle to illustrate the principles more explicitly. Let $\mathfrak{G}_\mathrm{d}\colon \mathbb{R}^{N+1} \to \mathbb{R}$ be the discrete action sum defined as $\mathfrak{G}_\mathrm{d} = \sum_{k=0}^{N-1} L_\mathrm{d}(q_k, q_{k+1})$, which approximates the action integral. The infinitesimal variation of the action sum can be written as

$$\begin{aligned} \delta\mathfrak{G}_\mathrm{d} = \sum_{k=0}^{N-1} &\delta q_{k+1}\left\{\frac{m}{h}(q_{k+1} - q_k) - \frac{h\kappa}{4}(q_k + q_{k+1})\right\} \\ &+ \delta q_k\left\{-\frac{m}{h}(q_{k+1} - q_k) - \frac{h\kappa}{4}(q_k + q_{k+1})\right\} . \end{aligned}$$

Since $\delta q_0 = \delta q_N = 0$, the summation index can be rewritten as

$$\begin{aligned} \delta\mathfrak{G}_\mathrm{d} = \sum_{k=1}^{N-1} \delta q_k\Big\{&-\frac{m}{h}(q_{k+1} - 2q_k + q_{k-1}) \\ &- \frac{h\kappa}{4}(q_{k+1} + 2q_k + q_{k-1})\Big\} . \end{aligned}$$

From discrete Hamilton's principle, $\delta\mathfrak{G}_\mathrm{d} = 0$ for any δq_k. Thus, the discrete equation of motion is given by

$$\begin{aligned} &\frac{m}{h}(q_{k+1} - 2q_k + q_{k-1}) \\ &\qquad + \frac{h\kappa}{4}(q_{k+1} + 2q_k + q_{k-1}) = 0 . \end{aligned} \tag{9}$$

For a given (q_{k-1}, q_k), we solve the above equation to obtain q_{k+1}. This yields a discrete flow map $(q_{k-1}, q_k) \mapsto (q_k, q_{k+1})$, and this process is repeated. The discrete Legendre transformation provides the discrete equation of motion in terms of the velocity as

$$\left(1 + \frac{h^2\kappa}{4m}\right)q_{k+1} = h\dot{q}_k + \left(1 - \frac{h^2\kappa}{4m}\right)q_k , \tag{10}$$

$$\dot{q}_{k+1} = \dot{q}_k - \frac{h\kappa}{2m}q_k - \frac{h\kappa}{2m}q_{k+1} . \tag{11}$$

For a given $(q_k, \dot{q}_k)$, we compute q_{k+1} and $\dot{q}_{k+1}$ by (10) and (11), respectively. This yields a discrete flow map $(q_k, \dot{q}_k) \mapsto (q_{k+1}, \dot{q}_{k+1})$. It can be shown that this variational integrator has second-order accuracy, which follows from the fact that the discrete action sum is a second-order approximation of the action integral.

Numerical example: We compare computational properties of the discrete equations of motion given by (10) and (11) with a 4(5)th order variable step size Runge–Kutta method. We choose $m = 1\,\mathrm{kg}$, $\kappa = 1\,\mathrm{kg/s^2}$ so that the natural frequency is 1 rad/s. The initial conditions are $q_0 = \sqrt{2}\,\mathrm{m}$, $\dot{q}_0 = 0$, and the total energy is $E = 1\,\mathrm{Nm}$. The simulation time is 200π sec, and the stepsize $h = 0.035$ of the discrete equations of motion is chosen such that the CPU times are the same for both methods. Figure 2a shows the computed total energy. The variational integrator preserves the total energy well. The mean variation is 2.7327×10^{-13} Nm. But, there is a notable dissipation of the computed total energy for the Runge–Kutta method

Example 2 (Planar Pendulum) A planar pendulum is a mass particle connected to a frictionless, one degree-of-freedom pivot by a rigid massless link under a uniform gravitational potential. The configuration space is the one-sphere $\mathbb{S}^1 = \{q \in \mathbb{R}^2 \mid \|q\| = 1\}$. While it is common to parametrize the one-sphere by an angle, we develop parameter-free equations of motion in the special orthogonal group SO(2), which is a group of 2×2 orthogonal matrices with determinant 1, i.e. $\mathrm{SO}(2) = \{R \in \mathbb{R}^{2\times2} \mid R^\mathrm{T}R = I_{2\times2}, \det[R] = 1\}$. SO(2) is diffeomorphic to the one-sphere. It is also possible to develop global equations of motion on the one-sphere directly, as shown

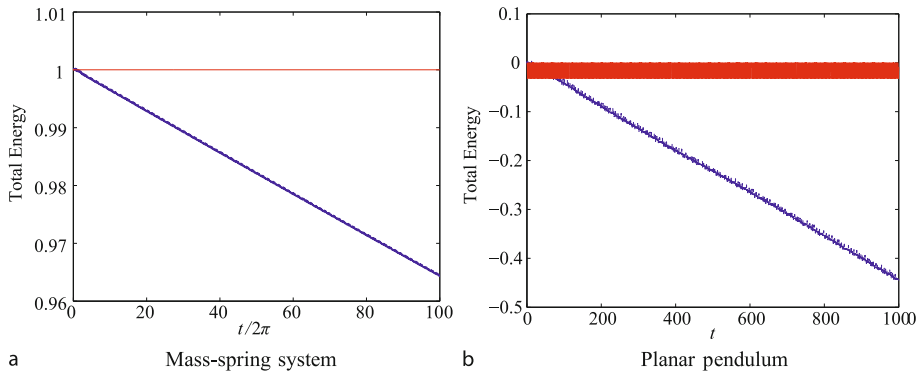

Discrete Control Systems, Figure 2
Computed total energy (RK45: *blue, dotted*, VI: *red, solid*)

in the next example, but here we focus on the special orthogonal group in order to illustrate the key steps to develop a Lie group variational integrator.

We first exploit the basic structures of the Lie group SO(2). Define a hat map $\hat{\cdot}$, which maps a scalar Ω to a 2×2 skew-symmetric matrix $\hat{\Omega}$ as

$$\hat{\Omega} = \begin{bmatrix} 0 & -\Omega \\ \Omega & 0 \end{bmatrix} .$$

The set of 2×2 skew-symmetric matrices forms the Lie algebra $\mathfrak{so}(2)$. Using the hat map, we identify $\mathfrak{so}(2)$ with $\mathbb{R}$. An inner product on $\mathfrak{so}(2)$ can be induced from the inner product on $\mathbb{R}$ as $\langle \hat{\Omega}_1, \hat{\Omega}_2 \rangle = 1/2\mathrm{tr}[\hat{\Omega}_1^{\mathrm{T}} \hat{\Omega}_2] = \Omega_1 \Omega_2$ for any $\Omega_1, \Omega_2 \in \mathbb{R}$. The matrix exponential is a local diffeomorphism from $\mathfrak{so}(2)$ to SO(2) given by

$$\exp \hat{\Omega} = \begin{bmatrix} \cos \Omega & -\sin \Omega \\ \sin \Omega & \cos \Omega \end{bmatrix} .$$

The kinematics equation for $R \in \mathrm{SO}(2)$ can be written in terms of a Lie algebra element as

$$\dot{R} = R\hat{\Omega} . \tag{12}$$

Continuous equations of motion: The Lagrangian for a planar pendulum $L\colon \mathrm{SO}(2)\times\mathfrak{so}(2) \to \mathbb{R}$ can be written as

$$\begin{aligned} L(R, \hat{\Omega}) &= \frac{1}{2} m l^2 \Omega^2 + m g l e_2^{\mathrm{T}} R e_2 \\ &= \frac{1}{2} m l^2 \langle \hat{\Omega}, \hat{\Omega} \rangle + m g l e_2^{\mathrm{T}} R e_2 , \end{aligned} \tag{13}$$

where the constant $g \in \mathbb{R}$ is the gravitational acceleration. The mass and the length of the pendulum are denoted by $m, l \in \mathbb{R}$, respectively. The second expression is used to define a discrete Lagrangian later. We choose the bases of the inertial frame and the body-fixed frame such that the unit vector along the gravity direction in the inertial frame, and the unit vector along the pendulum axis in the body-fixed frame are represented by the same vector $e_2 = [0; 1] \in \mathbb{R}^2$. Thus, for example, the hanging attitude is represented by $R = I_{2\times 2}$. Here, the rotation matrix $R \in \mathrm{SO}(2)$ represents the linear transformation from a representation of a vector in the body-fixed frame to the inertial frame.

Since the special orthogonal group is not a linear vector space, the expression for the variation should be carefully chosen. The infinitesimal variation of a rotation matrix $R \in \mathrm{SO}(2)$ can be written in terms of its Lie algebra element as

$$\delta R = \frac{\mathrm{d}}{\mathrm{d}\epsilon}\bigg|_{\epsilon=0} R \exp \epsilon\hat{\eta} = R\hat{\eta} \exp \epsilon\hat{\eta}\bigg|_{\epsilon=0} = R\hat{\eta} , \tag{14}$$

where $\eta \in \mathbb{R}$ so that $\hat{\eta} \in \mathfrak{so}(2)$. The infinitesimal variation of the angular velocity is induced from this expression and (12) as

$$\begin{aligned} \delta\hat{\Omega} &= \delta R^{\mathrm{T}} \dot{R} + R^{\mathrm{T}} \delta \dot{R} \\ &= (R\hat{\eta})^{\mathrm{T}} \dot{R} + R^{\mathrm{T}} (\dot{R}\hat{\eta} + R\hat{\dot{\eta}}) \\ &= -\hat{\eta}\hat{\Omega} + \hat{\Omega}\hat{\eta} + \hat{\dot{\eta}} = \hat{\dot{\eta}} , \end{aligned} \tag{15}$$

where we used the equality of mixed partials to compute $\delta\dot{R}$ as $\mathrm{d}/\mathrm{d}t(\delta R)$.

Define the action integral to be $\mathfrak{G} = \int_0^T L(R, \hat{\Omega})\mathrm{d}t$. The infinitesimal variation of the action integral is obtained by using (14) and (15). Hamilton's principle yields the following continuous equations of motion.

$$\dot{\Omega} + \frac{g}{l} e_2^{\mathrm{T}} R e_1 = 0 , \tag{16}$$

$$\dot{R} = R\hat{\Omega} . \tag{17}$$

If we parametrize the rotation matrix as

$$R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix},$$

these equations are equivalent to

$$\ddot{\theta} + \frac{g}{l}\sin\theta = 0\,. \tag{18}$$

Discrete equations of motion: We develop a Lie group variational integrator on SO(2). Similar to (4), define $F_k \in \mathrm{SO}(2)$ such that

$$R_{k+1} = R_k F_k\,. \tag{19}$$

Thus, $F_k = R_k^{\mathrm{T}} R_{k+1}$ represents the relative update between two integration steps. If we find $F_k \in \mathrm{SO}(2)$, the orthogonal structure is preserved through (19) since multiplication of orthogonal matrices is also orthogonal. This is a key idea of Lie group variational integrators.

Define the discrete Lagrangian $L_{\mathrm{d}}\colon \mathrm{SO}(2)\times\mathrm{SO}(2) \to \mathbb{R}$ to be

$$\begin{aligned} L_{\mathrm{d}}(R_k, F_k) &= \frac{1}{2h} m l^2 \langle F_k - I_{2\times 2}, F_k - I_{2\times 2}\rangle \\ &\quad + \frac{h}{2} m g l e_2^{\mathrm{T}} R_k e_2 + \frac{h}{2} m g l e_2^{\mathrm{T}} R_{k+1} e_2 \\ &= \frac{1}{2h} m l^2 \mathrm{tr}[I_{2\times 2} - F_k] + \frac{h}{2} m g l e_2^{\mathrm{T}} R_k e_2 \\ &\quad + \frac{h}{2} m g l e_2^{\mathrm{T}} R_{k+1} e_2\,, \end{aligned} \tag{20}$$

which is obtained by an approximation $h\hat{\Omega}_k \simeq R_k^{\mathrm{T}}(R_{k+1} - R_k) = F_k - I_{2\times 2}$, applied to the continuous Lagrangian given by (13).

As for the continuous time case, expressions for the infinitesimal variations should be carefully chosen. The infinitesimal variation of a rotation matrix is the same as (14), namely

$$\delta R_k = R_k \hat{\eta}_k\,, \tag{21}$$

for $\eta_k \in \mathbb{R}$, and the constrained variation of F_k is obtained from (19) as

$$\begin{aligned} \delta F_k &= \delta R_k^{\mathrm{T}} R_{k+1} + R_k^{\mathrm{T}} \delta R_{k+1} \\ &= -\hat{\eta}_k F_k + F_k \hat{\eta}_{k+1} = F_k(\hat{\eta}_{k+1} - F_k^{\mathrm{T}} \hat{\eta}_k F_k) \\ &= F_k(\hat{\eta}_{k+1} - \hat{\eta}_k)\,, \end{aligned} \tag{22}$$

where we use the fact that $F\hat{\eta}F^{\mathrm{T}} = \hat{\eta}$ for any $F \in \mathrm{SO}(2)$ and $\hat{\eta} \in \mathfrak{so}(2)$.

Define an action sum $\mathfrak{G}_{\mathrm{d}}\colon \mathrm{SO}(2)^{N+1} \to \mathbb{R}$ as $\mathfrak{G}_{\mathrm{d}} = \sum_{k=0}^{N-1} L_{\mathrm{d}}(R_k, F_k)$. Using (21) and (22), the variation of the action sum is written as

$$\begin{aligned} \delta\mathfrak{G}_{\mathrm{d}} = \sum_{k=1}^{N-1} \Big\langle & \frac{1}{2h} m l^2 \left(F_{k-1} - F_{k-1}^{\mathrm{T}}\right) \\ & - \frac{1}{2h}\left(F_k - F_k^{\mathrm{T}}\right) - h m g l \widehat{e_2^{\mathrm{T}} R_k e_1},\ \hat{\eta}_k \Big\rangle. \end{aligned}$$

From the discrete Hamilton's principle, $\delta\mathfrak{G}_{\mathrm{d}} = 0$ for any $\hat{\eta}_k$. Thus, we obtain the Lie group variational integrator on SO(2) as

$$\left(F_k - F_k^{\mathrm{T}}\right) - \left(F_{k+1} - F_{k+1}^{\mathrm{T}}\right) - \frac{2h^2 g}{l} \widehat{e_2^{\mathrm{T}} R_{k+1} e_1} = 0\,, \tag{23}$$

$$R_{k+1} = R_k F_k\,. \tag{24}$$

For a given (R_k, F_k) and $R_{k+1} = R_k F_k$, (23) is solved to find F_{k+1}. This yields a discrete map $(R_k, F_k) \mapsto (R_{k+1}, F_{k+1})$. If we parametrize the rotation matrices R and F with θ and $\Delta\theta$ and if we assume that $\Delta\theta \ll 1$, these equations are equivalent to

$$\frac{1}{h}(\theta_{k+1} - 2\theta_k + \theta_k) + \frac{hg}{l}\sin\theta_k = 0\,.$$

The discrete version of the Legendre transformation provides the discrete Hamiltonian map as follows.

$$F_k - F_k^{\mathrm{T}} = 2h\hat{\Omega} - \frac{h^2 g}{l} \widehat{e_2^{\mathrm{T}} R_k e_1}\,, \tag{25}$$

$$R_{k+1} = R_k F_k\,, \tag{26}$$

$$\Omega_{k+1} = \Omega_k - \frac{hg}{2l} e_2^{\mathrm{T}} R_k e_1 - \frac{hg}{2l} e_2^{\mathrm{T}} R_{k+1} e_1\,. \tag{27}$$

For a given (R_k, Ω_k), we solve (25) to obtain F_k. Using this, (R_{k+1}, Ω_{k+1}) is obtained from (26) and (27). This yields a discrete map $(R_k, \Omega_k) \mapsto (R_{k+1}, \Omega_{k+1})$.

Numerical example: We compare the computational properties of the discrete equations of motion given by (25)–(27) with a 4(5)th order variable step size Runge–Kutta method. We choose $m = 1$ kg, $l = 9.81$ m. The initial conditions are $\theta_0 = \pi/2$ rad, $\Omega = 0$, and the total energy is $E = 0$ Nm. The simulation time is 1000 sec, and the step-size $h = 0.03$ of the discrete equations of motion is chosen such that the CPU times are identical. Figure 2b shows the computed total energy for both methods. The variational integrator preserves the total energy well. There is no drift in the computed total energy, and the mean variation is 1.0835×10^{-2} Nm. But, there is a notable dissipation of the computed total energy for the Runge–

Kutta method. Note that the computed total energy would further decrease as the simulation time increases.

Example 3 (Spherical Pendulum) A spherical pendulum is a mass particle connected to a frictionless, two degree-of-freedom pivot by a rigid massless link. The mass particle acts under a uniform gravitational potential. The configuration space is the two-sphere $\mathbb{S}^2 = \{q \in \mathbb{R}^3 \mid \|q\| = 1\}$. It is common to parametrize the two-sphere by two angles, but this description of the spherical pendulum has a singularity. Any trajectory near the singularity encounters numerical ill-conditioning. Furthermore, this leads to complicated expressions involving trigonometric functions.

Here we develop equations of motion for a spherical pendulum using the global structure of the two-sphere without parametrization. In the previous example, we develop equations of motion for a planar pendulum using the fact that the one-sphere $\mathbb{S}^1$ is diffeomorphic to the special orthogonal group SO(2). But, the two-sphere is not diffeomorphic to a Lie group. Instead, there exists a natural Lie group action on the two-sphere. That is the 3-dimensional special orthogonal group SO(3), a group of 3×3 orthogonal matrices with determinant 1, i.e. $\mathrm{SO}(3) = \{R \in \mathbb{R}^{3\times3} \mid R^{\mathrm{T}}R = I_{3\times3},\ \det[R] = 1\}$. The special orthogonal group SO(3) acts on the two-sphere in a transitive way; for any $q_1, q_2 \in \mathbb{S}^2$, there exists a $R \in \mathrm{SO}(3)$ such that $q_2 = Rq_1$. Thus, the discrete update for the two-sphere can be expressed in terms of a rotation matrix as (19). This is a key idea to develop a discrete equations of motion for a spherical pendulum.

Continuous equations of motion: Let $q \in \mathbb{S}^2$ be a unit vector from the pivot point to the point mass. The Lagrangian for a spherical pendulum can be written as

$$L(q, \dot{q}) = \frac{1}{2}ml^2\dot{q}\cdot\dot{q} + mgle_3\cdot q\,, \tag{28}$$

where the gravity direction is assumed to be $e_3 = [0; 0; 1] \in \mathbb{R}^3$. The mass and the length of the pendulum are denoted by $m, l \in \mathbb{R}$, respectively. The infinitesimal variation of the unit vector q can be written in terms of the vector cross product as

$$\delta q = \xi \times q\,, \tag{29}$$

where $\xi \in \mathbb{R}^3$ is constrained to be orthogonal to the unit vector, i.e. $\xi \cdot q = 0$. Using this expression for the infinitesimal variation, Hamilton's principle yields the following continuous equations of motion.

$$\ddot{q} + (\dot{q}\cdot\dot{q})q + \frac{g}{l}(q \times (q \times e_3)) = 0\,. \tag{30}$$

Since $\dot{q} = \omega \times q$ for some angular velocity $\omega \in \mathbb{R}^3$ satisfying $\omega \cdot q = 0$, this can also be equivalently written as

$$\dot{\omega} - \frac{g}{l}q \times e_3 = 0\,, \tag{31}$$

$$\dot{q} = \omega \times q\,. \tag{32}$$

These are global equations of motion for a spherical pendulum; these are much simpler than the equations expressed in term of angles, and they have no singularity.

Discrete equations of motion: We develop a variational integrator for the spherical pendulum defined on $\mathbb{S}^2$. Since the special orthogonal group acts on the two-sphere transitively, we can define the discrete update map for the unit vector as

$$q_{k+1} = F_k q_k \tag{33}$$

for $F_k \in \mathrm{SO}(3)$. The rotation matrix F_k is not uniquely defined by this condition, since $\exp(\lambda\hat{q}_k)$, which corresponds to a rotation about the q_k direction fixes the vector q_k. Consequently, if F_k satisfies (33), then $F_k \exp(\lambda\hat{q}_k)$ does as well. We avoid this ambiguity by requiring that F_k does not rotate about q_k, which can be achieved by letting $F_k = \exp(\hat{f}_k)$, where $f_k \cdot q_k = 0$.

Define a discrete Lagrangian $L_{\mathrm{d}}\colon \mathbb{S}^2 \times \mathbb{S}^2 \to \mathbb{R}$ to be

$$\begin{aligned} L_{\mathrm{d}}(q_k, q_{k+1}) = {} & \frac{1}{2h}ml^2(q_{k+1} - q_k)\cdot(q_{k+1} - q_k) \\ & + \frac{h}{2}mgle_3\cdot q_k + \frac{h}{2}mgle_3\cdot q_{k+1}\,. \end{aligned}$$

The variation of q_k is the same as (29), namely

$$\delta q_k = \xi_k \times q_k \tag{34}$$

for $\xi_k \in \mathbb{R}^3$ with a constraint $\xi_k \cdot q_k = 0$. Using this discrete Lagrangian and the expression for the variation, discrete Hamilton's principle yields the following discrete equations of motion for a spherical pendulum.

$$\begin{aligned} q_{k+1} = {} & \left(h\omega_k + \frac{h^2 g}{2l}q_k \times e_3\right) \times q_k \\ & + \left(1 - \left\|h\omega_k + \frac{h^2 g}{2l}q_k \times e_3\right\|^2\right)^{1/2} q_k\,, \end{aligned} \tag{35}$$

$$\omega_{k+1} = \omega_k + \frac{hg}{2l}q_k \times e_3 + \frac{hg}{2l}q_{k+1} \times e_3\,. \tag{36}$$

Since an explicit solution for $F_k \in \mathrm{SO}(3)$ can be obtained in this case, the rotation matrix F_k does not appear in the equations of motion. This variational integrator on $\mathbb{S}^2$ exactly preserves the unit length of q_k, the constraint

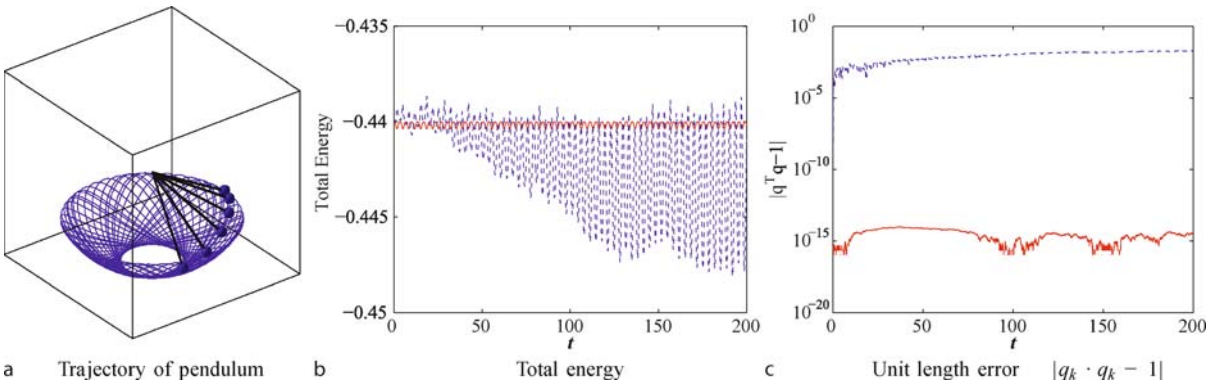

Discrete Control Systems, Figure 3
Numerical simulation of a spherical pendulum (RK45: *blue, dotted*, VI: *red, solid*)

$q_k \cdot \omega_k = 0$, and the third component of the angular velocity $\omega_k \cdot e_3$ which is conserved since gravity exerts no moment along the gravity direction e_3.

Numerical example: We compare the computational properties of the discrete equations of motion given by (35) and (36) with a 4(5)th order variable step size Runge–Kutta method for (31) and (32). We choose $m = 1\,\text{kg}$, $l = 9.81\,\text{m}$. The initial conditions are $q_0 = [\sqrt{3}/2, 0, 1/2]$, $\omega_0 = 0.1[\sqrt{3}, 0, 3]$ rad/sec, and the total energy is $E = -0.44\,\text{Nm}$. The simulation time is 200 sec, and the step-size $h = 0.05$ of the discrete equations of motion is chosen such that the CPU times are identical. Figure 3 shows the computed total energy and the unit length errors. The variational integrator preserves the total energy and the structure of the two-sphere well. The mean total energy deviation is 1.5460×10^{-4} Nm, and the mean unit length error is 3.2476×10^{-15}. But, there is a notable dissipation of the computed total energy for the Runge–Kutta method. The Runge–Kutta method also fails to preserve the structure of the two-sphere. The mean unit length error is 1.0164×10^{-2}.

Example 4 (Rigid Body in a Potential Field) Consider a rigid body under a potential field that is dependent on the position and the attitude of the body. The configuration space is the special Euclidean group, which is a semidirect product of the special orthogonal group and Euclidean space, i. e. $\mathrm{SE}(3) = \mathrm{SO}(3) \circledS \mathbb{R}^3$.

Continuous equations of motion: The equations of motion for a rigid body can be developed either from Hamilton's principle (see [26]) in a similar way as Example 2, or directly from the generalized discrete Euler–Poincaré equation given at (5). Here, we summarize the results. Let $m \in \mathbb{R}$ and $J \in \mathbb{R}^{3\times 3}$ be the mass and the moment of inertia matrix of a rigid body. For $(R, x) \in \mathrm{SE}(3)$, the linear transformation from the body-fixed frame to the inertial frame is denoted by the rotation matrix $R \in \mathrm{SO}(3)$, and the position of the mass center in the inertial frame is denoted by a vector $x \in \mathbb{R}^3$. The vectors $\Omega, v \in \mathbb{R}^3$ are the angular velocity in the body-fixed frame, and the translational velocity in the inertial frame, respectively. Suppose that the rigid body acts under a configuration-dependent potential $U\colon \mathrm{SE}(3) \to \mathbb{R}$. The continuous equations of motion for the rigid body can be written as

$$\dot{R} = R\hat{\Omega}\,, \tag{37}$$

$$\dot{x} = v\,, \tag{38}$$

$$J\dot{\Omega} + \Omega \times J\Omega = M\,, \tag{39}$$

$$m\dot{v} = -\frac{\partial U}{\partial x}\,, \tag{40}$$

where the hat map $\hat{\cdot}$ is an isomorphism from $\mathbb{R}^3$ to 3×3 skew-symmetric matrices $\mathfrak{so}(3)$, defined such that $\hat{x}y = x \times y$ for any $x, y \in \mathbb{R}^3$. The moment due to the potential $M \in \mathbb{R}^3$ is obtained by the following relationship:

$$\hat{M} = \frac{\partial U}{\partial R}^{\mathrm{T}} R - R^{\mathrm{T}} \frac{\partial U}{\partial R}\,. \tag{41}$$

The matrix $\partial U/\partial R \in \mathbb{R}^{3\times 3}$ is defined such that $[\partial U/\partial R]_{ij} = \partial U/\partial [R]_{ij}$ for $i, j \in \{1, 2, 3\}$, where the i, jth element of a matrix is denoted by $[\cdot]_{ij}$.

Discrete equations of motion: The corresponding discrete equations of motion are given by

$$h\widehat{J\Omega_k} + \frac{h^2}{2}\hat{M}_k = F_k J_{\mathrm{d}} - J_{\mathrm{d}} F_k^{\mathrm{T}}\,, \tag{42}$$

$$R_{k+1} = R_k F_k\,, \tag{43}$$

$$x_{k+1} = x_k + hv_k - \frac{h^2}{2m}\frac{\partial U_k}{\partial x_k}\,, \tag{44}$$

$$J\Omega_{k+1} = F_k^{\mathrm{T}} J\Omega_k + \frac{h}{2}F_k^{\mathrm{T}} M_k + \frac{h}{2}M_{k+1}\,, \tag{45}$$

$$mv_{k+1} = mv_k - \frac{h}{2}\frac{U_k}{\partial x_k} - \frac{h}{2}\frac{\partial U_{k+1}}{\partial x_{k+1}}\,, \tag{46}$$

where $J_{\mathrm{d}} \in \mathbb{R}^{3\times 3}$ is a non-standard moment of inertia matrix defined as $J_{\mathrm{d}} = 1/2\mathrm{tr}[J]I_{3\times 3} - J$. For a given $(R_k, x_k, \Omega_k, v_k)$, we solve the implicit Eq. (42) to find $F_k \in \mathrm{SO}(3)$. Then, the configuration at the next step R_{k+1}, x_{k+1} is obtained by (43) and (44), and the moment and force $M_{k+1}, (\partial U_{k+1})/(\partial x_{k+1})$ can be computed. Velocities Ω_{k+1}, v_{k+1} are obtained from (45) and (46). This defines a discrete flow map, $(R_k, x_k, \Omega_k, v_k) \mapsto (R_{k+1}, x_{k+1}, \Omega_{k+1}, v_{k+1})$, and this process can be repeated. This Lie group variational integrator on SE(3) can be generalized to multiple rigid bodies acting under their mutual gravitational potential (see [26]).

Numerical example: We compare the computational properties of the discrete equations of motion given by (42)–(46) with a 4(5)th order variable step size Runge–Kutta method for (37)–(40). In addition, we compute the attitude dynamics using quaternions on the unit three-sphere $\mathbb{S}^3$. The attitude kinematics Eq. (37) is rewritten in terms of quaternions, and the corresponding equations are integrated by the same Runge–Kutta method.

We choose a dumbbell spacecraft, that is two spheres connected by a rigid massless rod, acting under a central gravitational potential. The resulting system is a restricted full two body problem. The dumbbell spacecraft model has an analytic expression for the gravitational potential, resulting in a nontrivial coupling between the attitude dynamics and the orbital dynamics.

As shown in Fig. 4a, the initial conditions are chosen such that the resulting motion is a near-circular orbit combined with a rotational motion. Figure 4b and c show the computed total energy and the orthogonality errors of the rotation matrix. The Lie group variational integrator preserves the total energy and the Lie group structure of SO(3). The mean total energy deviation is 2.5983×10^{-4}, and the mean orthogonality error is 1.8553×10^{-13}. But, there is a notable dissipation of the computed total energy and the orthogonality error for the Runge–Kutta method. The mean orthogonality errors for the Runge–Kutta method are 0.0287 and 0.0753, respectively, using kinematics equation with rotation matrices, and using the kinematics equation with quaternions. Thus, the attitude of the rigid body is not accurately computed for Runge–Kutta methods. It is interesting to see that the Runge–Kutta method with quaternions, which is generally assumed to have better computational properties than the kinematics equation with rotation matrices, has larger total energy error and orthogonality error. Since the unit length of the quaternion vector is not preserved in the numerical computations, orthogonality errors arise when converted to a rotation matrix. This suggests that it is critical to preserve the structure of SO(3) in order to study the global characteristics of the rigid body dynamics.

The importance of simultaneously preserving the symplectic structure and the Lie group structure of the configuration space in rigid body dynamics can be observed numerically. Lie group variational integrators, which preserve both of these properties, are compared to methods that only preserve one, or neither, of these properties (see [27]). It is shown that the Lie group variational integrator exhibits greater numerical accuracy and efficiency. Due to these computational advantages, the Lie group variational integrator has been used to study the dynamics of the binary near-Earth asteroid 66391 (1999 KW_4) in joint work between the University of Michigan and the Jet Propulsion Laboratory, NASA (see [37]).

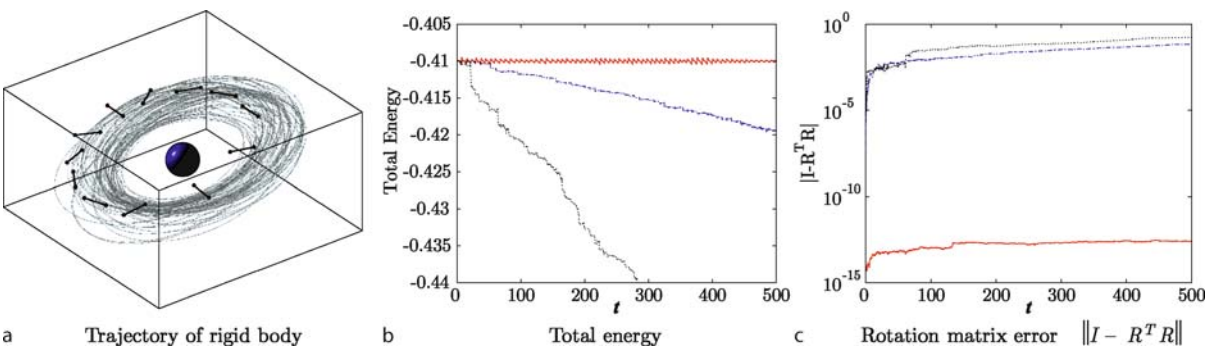

Discrete Control Systems, Figure 4
Numerical simulation of a dumbbell rigid body (LGVI: *red, solid*, RK45 with rotation matrices: *blue, dash-dotted*, RK45 with quaternions: *black, dotted*)

Optimal Control of Discrete Lagrangian and Hamiltonian Systems

Optimal control problems involve finding a control input such that a certain optimality objective is achieved under prescribed constraints. An optimal control problem that minimizes a performance index is described by a set of differential equations, which can be derived using Pontryagin's maximum principle. The equations of motion for a system are constrained by Lagrange multipliers, and necessary conditions for optimality is obtained by the calculus of variations. The solution for the corresponding two point boundary value problem provides the optimal control input. Alternatively, a sub-optimal control law is obtained by approximating the control input history with finite data points.

Discrete optimal control problems involve finding a control input for a given system described by discrete Lagrangian and Hamiltonian mechanics. The control inputs are parametrized by their values at each discrete time step, and the discrete equations of motion are derived from the discrete Lagrange–d'Alembert principle [21],

$$\delta \sum_{k=0}^{N-1} L_{\mathrm{d}}(q_k, q_{k+1}) = \sum_{k=0}^{N-1} [F_{\mathrm{d}}^{-}(q_k, q_{k+1}) \cdot \delta q_k + F_{\mathrm{d}}^{+}(q_k, q_{k+1}) \cdot \delta q_{k+1}] ,$$

which modifies the discrete Hamilton's principle by taking into account the virtual work of the external forces. Discrete optimal control is in contrast to traditional techniques such as collocation, wherein the continuous equations of motion are imposed as constraints at a set of collocation points, since this approach induces constraints on the configuration at each discrete timestep.

Any optimal control algorithm can be applied to the discrete Lagrangian or Hamiltonian system. For an indirect method, our approach to a discrete optimal control problem can be considered as a multiple stage variational problem. The discrete equations of motion are derived by the discrete variational principle. The corresponding variational integrators are imposed as dynamic constraints to be satisfied by using Lagrange multipliers, and necessary conditions for optimality, expressed as discrete equations on multipliers, are obtained from a variational principle. For a direct method, control inputs can be optimized by using parameter optimization tools such as a sequential quadratic programming. The discrete optimal control can be characterized by discretizing the optimal control problem from the problem formulation stage.

This method has substantial computational advantages when used to find an optimal control law. As discussed in the previous section, the discrete dynamics are more faithful to the continuous equations of motion, and consequently more accurate solutions to the optimal control problem are obtained. The external control inputs break the Lagrangian and Hamiltonian system structure. For example, the total energy is not conserved for a controlled mechanical system. But, the computational superiority of the discrete mechanics still holds for controlled systems. It has been shown that the discrete dynamics is more reliable even for controlled system as it computes the energy dissipation rate of controlled systems more accurately (see [31]). For example, this feature is extremely important in computing accurate optimal trajectories for long term spacecraft attitude maneuvers using low energy control inputs.

The discrete dynamics does not only provide an accurate optimal control input, but also enables us to find it efficiently. For the indirect optimal control approach, optimal solutions are usually sensitive to a small variation of multipliers. This causes difficulties, such as numerical ill-conditioning, when solving the necessary conditions for optimality expressed as a two point boundary value problem. Sensitivity derivatives along the discrete necessary conditions do not have numerical dissipation introduced by conventional numerical integration schemes. Thus, they are numerically more robust, and the necessary conditions can be solved computationally efficiently. For the direct optimal control approach, optimal control inputs can be obtained by using a larger discrete step size, which requires less computational load.

We illustrate the basic properties of the discrete optimal control using optimal control problems for the spherical pendulum and the rigid body model presented in the previous section.

Example 5 (Optimal Control of a Spherical Pendulum) We study an optimal control problem for the spherical pendulum described in Example 3. We assume that an external control moment $u \in \mathbb{R}^3$ acts on the pendulum. Control inputs are parametrized by their values at each time step, and the discrete equations of motion are modified to include the effects of the external control inputs by using the discrete Lagrange–d'Alembert principle. Since the component of the control moment that is parallel to the direction along the pendulum has no effect, we parametrize the control input as $u_k = q_k \times w_k$ for $w_k \in \mathbb{R}^3$.

The objective is to transfer the pendulum from a given initial configuration (q_0, ω_0) to a desired configuration (q^d, ω^d) during a fixed maneuver time N, while minimiz-

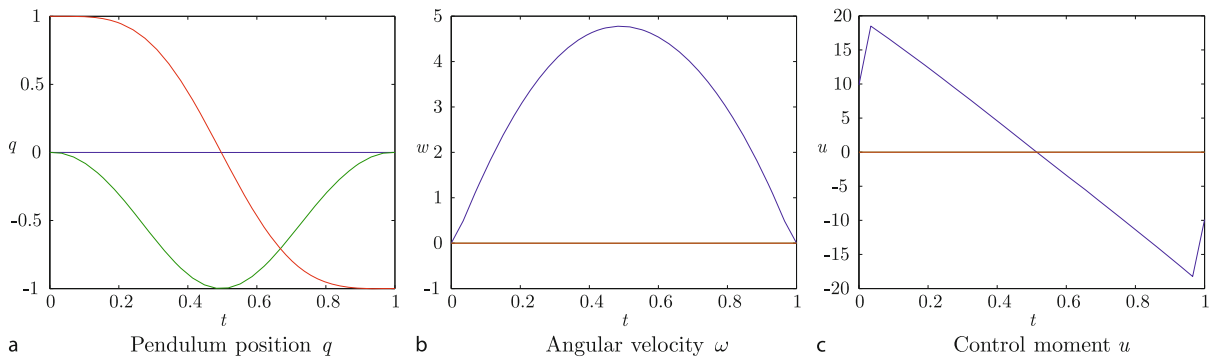

Discrete Control Systems, Figure 5
Optimal control of a spherical pendulum

ing the square of the weighted l_2 norm of the control moments.

$$\min_{w_k} \mathcal{J} = \sum_{k=0}^{N} \frac{h}{2} u_k^{\mathrm{T}} u_k = \sum_{k=0}^{N} \frac{h}{2} (q_k \times w_k)^{\mathrm{T}} (q_k \times w_k) \,.$$

We solve this optimal control problem by using a direct numerical optimization method. The terminal boundary condition is imposed as an equality constraint, and the $3(N+1)$ control input parameters $\{w_k\}_{k=0}^{N}$ are numerically optimized using sequential quadratic programming. This method is referred to as a DMOC (Discrete Mechanics and Optimal Control) approach (see [19]).

Figure 5 shows a optimal solution transferring the spherical pendulum from a hanging configuration given by $(q_0, \omega_0) = (e_3, 0_{3\times 1})$ to an inverted configuration $(q^d, \omega^d) = (-e_3, 0_{3\times 1})$ during 1 second. The time step size is $h = 0.033$. Experiment have shown that the DMOC approach can compute optimal solutions using larger step sizes than typical Runge–Kutta methods, and consequently, it requires less computational load. In this case, using a second-order accurate Runge–Kutta method, the same optimization code fails while giving error messages of inaccurate and singular gradient computations. It is presumed that the unit length errors of the Runge–Kutta method, shown in Example 3, cause numerical instabilities for the finite-difference gradient computations required for the sequential quadratic programming algorithm.

Example 6 (Optimal Control of a Rigid Body in a Potential Field) We study an optimal control problem of a rigid body using a dumbbell spacecraft model described in Example 4 (see [25] for detail). We assume that external control forces $u^f \in \mathbb{R}^3$, and control moment $u^m \in \mathbb{R}^3$ act on the dumbbell spacecraft. Control inputs are parametrized by their values at each time step, and the Lie group variational integrators are modified to include the effects of the external control inputs by using discrete Lagrange–d'Alembert principle.

The objective is to transfer the dumbbell from a given initial configuration $(R_0, x_0, \Omega_0, v_0)$ to a desired configuration $(R^d, x^{\mathrm{d}}, \Omega^{\mathrm{d}}, v^{\mathrm{d}})$ during a fixed maneuver time N, while minimizing the square of the l_2 norm of the control inputs.

$$\min_{u_{k+1}} \mathcal{J} = \sum_{k=0}^{N-1} \frac{h}{2} \left(u_{k+1}^f\right)^{\mathrm{T}} W_f u_{k+1}^f + \frac{h}{2} \left(u_{k+1}^m\right)^{\mathrm{T}} W_m u_{k+1}^m \,,$$

where $W_f, W_m \in \mathbb{R}^{3\times 3}$ are symmetric positive definite matrices. Here we use a modified version of the discrete equations of motion with first order accuracy, as it yields a compact form for the necessary conditions.

Necessary conditions for optimality: We solve this optimal control problem by using an indirect optimization method, where necessary conditions for optimality are derived using variational arguments, and a solution of the corresponding two-point boundary value problem provides the optimal control. This approach is common in the optimal control literature; here the optimal control problem is discretized at the problem formulation level using the Lie group variational integrator presented in Sect. "Discrete Lagrangian and Hamiltonian Mechanics".

$$\begin{aligned}
\mathcal{J}_a = & \sum_{k=0}^{N-1} \frac{h}{2} \left(u_{k+1}^f\right)^{\mathrm{T}} W^f u_{k+1}^f + \frac{h}{2} \left(u_{k+1}^m\right)^{\mathrm{T}} W^m u_{k+1}^m \\
& + \lambda_k^{1,T} \{-x_{k+1} + x_k + h v_k\} \\
& + \lambda_k^{2,T} \left\{-m v_{k+1} + m v_k - h \frac{\partial U_{k+1}}{\partial x_{k+1}} + h u_{k+1}^f\right\} \\
& + \lambda_k^{3,T} \left(\log\mathrm{m}(F_k - R_k^{\mathrm{T}} R_{k+1})\right)^{\vee} \\
& + \lambda_k^{4,T} \left\{-J\Omega_{k+1} + F_k^{\mathrm{T}} J \Omega_k + h\left(M_{k+1} + u_{k+1}^m\right)\right\} ,
\end{aligned}$$

where $\lambda_k^1, \lambda_k^2, \lambda_k^3, \lambda_k^4 \in \mathbb{R}^3$ are Lagrange multipliers. The matrix logarithm is denoted by $\mathrm{logm}\colon \mathrm{SO}(3) \to \mathfrak{so}(3)$ and the vee map $\vee\colon \mathfrak{so}(3) \to \mathbb{R}^3$ is the inverse of the hat map introduced in Example 4. The logarithm form of (43) is used, and the constraint (42) is considered implicitly using constrained variations. Using similar expressions for the variation of the rotation matrix and the angular velocity given in (14) and (15), the infinitesimal variation can be written as

$$\begin{aligned}\delta \mathcal{J}_a = \sum_{k=1}^{N-1} & h\delta u_k^{f,\mathrm{T}} \left\{ W_f u_k^f + \lambda_{k-1}^2 \right\} \\ & + h\delta u_k^{m,\mathrm{T}} \left\{ W_m u_k^m + \lambda_{k-1}^4 \right\} \\ & + z_k^{\mathrm{T}} \left\{ -\lambda_{k-1} + A_k^{\mathrm{T}} \lambda_k \right\} ,\end{aligned}$$

where $\lambda_k = [\lambda_k^1; \lambda_k^2; \lambda_k^3; \lambda_k^4] \in \mathbb{R}^{12}$, and $z_k \in \mathbb{R}^{12}$ represents the infinitesimal variation of $(R_k, x_k, \Omega_k, v_k)$, given by $z_k = [\mathrm{logm}(R_k^{\mathrm{T}} \delta R_k)^\vee; \delta x_k, \delta \Omega_k, \delta v_k]$. The matrix $A_k \in \mathbb{R}^{12\times 12}$ is defined in terms of $(R_k, x_k, \Omega_k, v_k)$. Thus, necessary conditions for optimality are given by

$$u_{k+1}^f = -W_f^{-1} \lambda_k^2 , \tag{47}$$

$$u_{k+1}^m = -W_m^{-1} \lambda_k^4 , \tag{48}$$

$$\lambda_k = A_{k+1}^{\mathrm{T}} \lambda_{k+1} \tag{49}$$

together with the discrete equations of motion and the boundary conditions.

Computational approach: Necessary conditions for optimality are expressed in terms of a two point boundary problem. The problem is to find the optimal discrete flow, multipliers, and control inputs to satisfy the equations of motion, optimality conditions, multiplier equations, and boundary conditions simultaneously. We use a neighboring extremal method (see [12]). A nominal solution satisfying all of the necessary conditions except the boundary conditions is chosen. The unspecified initial multiplier is updated by successive linearization so as to satisfy the specified terminal boundary conditions in the limit. This is also referred to as the shooting method. The main advantage of the neighboring extremal method is that the number of iteration variables is small.

The difficulty is that the extremal solutions are sensitive to small changes in the unspecified initial multiplier values. The nonlinearities also make it hard to construct an accurate estimate of sensitivity, thereby resulting in numerical ill-conditioning. Therefore, it is important to compute the sensitivities accurately to apply the neighboring extremal method. Here the optimality conditions (47) and (48) are substituted into the equations of motion and the multiplier equations, which are linearized to obtain sensitivity derivatives of an optimal solution with respect to boundary conditions. Using this sensitivity, an initial guess of the unspecified initial conditions is iterated to satisfy the specified terminal conditions in the limit. Any type of Newton iteration can be applied. We use a line search with backtracking algorithm, referred to as Newton–Armijo iteration (see [22]).

Figure 6 shows optimized maneuvers, where a dumbbell spacecraft on a reference circular orbit is transferred to another circular orbit with a different orbital radius and inclination angle. Figure 6a shows the violation of the terminal boundary condition according to the number of iterations on a logarithmic scale. Red circles denote outer iterations in the Newton–Armijo iteration to compute the sensitivity derivatives. The error in satisfaction of the terminal boundary condition converges quickly to machine precision after the solution is close to the local minimum at around the 20th iteration. These convergence results are consistent with the quadratic convergence rates expected of Newton methods with accurately computed gradients.

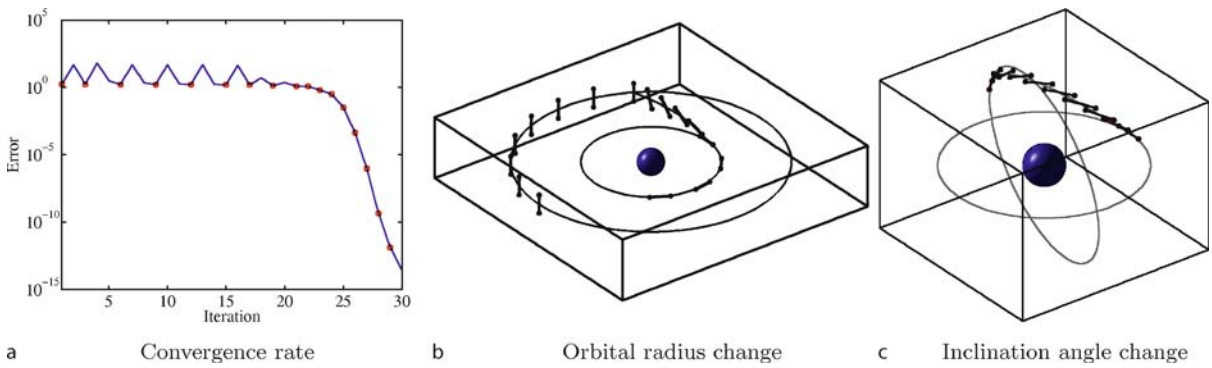

Discrete Control Systems, Figure 6
Optimal control of a rigid body

The neighboring extremal method, also referred to as the shooting method, is numerically efficient in the sense that the number of optimization parameters is minimized. But, in general, this approach may be prone to numerical ill-conditioning (see [3]). A small change in the initial multiplier can cause highly nonlinear behavior of the terminal attitude and angular momentum. It is difficult to compute the gradient for Newton iterations accurately, and the numerical error may not converge. However, the numerical examples presented in this article show excellent numerical convergence properties. The dynamics of a rigid body arises from Hamiltonian mechanics, which have neutral stability, and its adjoint system is also neutrally stable. The proposed Lie group variational integrator and the discrete multiplier equations, obtained from variations expressed in the Lie algebra, preserve the neutral stability property numerically. Therefore the sensitivity derivatives are computed accurately.

Controlled Lagrangian Method for Discrete Lagrangian Systems

The method of controlled Lagrangians is a procedure for constructing feedback controllers for the stabilization of relative equilibria. It relies on choosing a parametrized family of controlled Lagrangians whose corresponding Euler–Lagrange flows are equivalent to the closed loop behavior of a Lagrangian system with external control forces. The condition that these two systems are equivalent results in matching conditions. Since the controlled system can now be viewed as a Lagrangian system with a modified Lagrangian, the global stability of the controlled system can be determined directly using Lyapunov stability analysis.

This approach originated in Bloch et al. [5] and was then developed in Auckly et al. [1]; Bloch et al. [6,7,8,9]; Hamberg [15,16]. A similar approach for Hamiltonian controlled systems was introduced and further studied in the work of Blankenstein, Ortega, van der Schaft, Maschke, Spong, and their collaborators (see, for example, [33,35] and related references). The two methods were shown to be equivalent in Chang et al. [13] and a nonholonomic version was developed in Zenkov et al. [40,41], and Bloch [4].

Since the method of controlled Lagrangians relies on relating the closed-loop dynamics of a controlled system with the Euler–Lagrange flow associated with a modified Lagrangian, it is natural to discretize this approach through the use of variational integrators. In Bloch et al. [10,11], a discrete theory of controlled Lagrangians was developed for variational integrators, and applied to the feedback stabilization of the unstable inverted equilibrium of the pendulum on a cart.

The pendulum on a cart is an example of an underactuated control problem, which has two degrees of freedom, given by the pendulum angle and the cart position. Only the cart position has control forces acting on it, and the stabilization of the pendulum has to be achieved indirectly through the coupling between the pendulum and the cart. The controlled Lagrangian is obtained by modifying the kinetic energy term, a process referred to as kinetic shaping. Similarly, it is possible to modify the potential energy term using potential shaping.

Since the pendulum on a cart model involves both a planar pendulum, and a cart that translates in one-dimension, the configuration space is a cylinder, $\mathbb{S}^1 \times \mathbb{R}$.

Continuous kinetic shaping: The Lagrangian has the form kinetic minus potential energy

$$L(q,\dot{q}) = \frac{1}{2}\left[\alpha\dot{\theta}^2 + 2\beta(\theta)\dot{\theta}\dot{s} + \gamma\dot{s}^2\right] - V(q), \tag{50}$$

and the corresponding controlled Euler–Lagrange dynamics is

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial\dot{\theta}} - \frac{\partial L}{\partial\theta} = 0\,, \tag{51}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial\dot{s}} = u\,, \tag{52}$$

where u is the control input.

Since the potential energy is translation invariant, i. e., $V(q) = V(\theta)$, the *relative equilibria* $\theta = \theta_e$, $\dot{s} = \text{const}$ are unstable and given by non-degenerate critical points of $V(\theta)$. To stabilize the relative equilibria $\theta = \theta_e$, $\dot{s} = \text{const}$ with respect to θ, kinetic shaping is used. The controlled Lagrangian in this case is defined by

$$L^{\tau,\sigma}(q,\dot{q}) = L(\theta,\dot{\theta},\dot{s}+\tau(\theta)\dot{\theta}) + \frac{1}{2}\sigma\gamma(\tau(\theta)\dot{\theta})^2\,, \tag{53}$$

where $\tau(\theta) = \kappa\beta(\theta)$. This velocity shift corresponds to a new choice of the horizontal space (see [8] for details). The dynamics is just the Euler–Lagrange dynamics for controlled Lagrangian (53),

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L^{\tau,\sigma}}{\partial\dot{\theta}} - \frac{\partial L^{\tau,\sigma}}{\partial\theta} = 0\,, \tag{54}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L^{\tau,\sigma}}{\partial\dot{s}} = 0\,. \tag{55}$$

The Lagrangian (53) satisfies the simplified matching conditions of Bloch et al. [9] when the kinetic energy metric coefficient γ in (50) is constant.

Setting $u = -\mathrm{d}\big(\gamma\tau(\theta)\dot{\theta}\big)/\mathrm{d}t$ defines the control input, makes Eqs. (52) and (55) identical, and results in controlled momentum conservation by dynamics (51)

and (52). Setting $\sigma = -1/\gamma\kappa$ makes Eqs. (51) and (54) identical when restricted to a level set of the controlled momentum.

Discrete kinetic shaping: Here, we adopt the following notation:

$$q_{k+1/2} = \frac{q_k + q_{k+1}}{2}\,, \qquad \Delta q_k = q_{k+1} - q_k\,, \quad q_k = (\theta_k, s_k)\,.$$

Then, a second-order accurate discrete Lagrangian is given by,

$$L_\mathrm{d}(q_k, q_{k+1}) = hL(q_{k+1/2}, \Delta q_k/h)\,.$$

The discrete dynamics is governed by the equations

$$\frac{\partial L_\mathrm{d}(q_k, q_{k+1})}{\partial \theta_k} + \frac{\partial L_\mathrm{d}(q_{k-1}, q_k)}{\partial \theta_k} = 0\,, \tag{56}$$

$$\frac{\partial L_\mathrm{d}(q_k, q_{k+1})}{\partial s_k} + \frac{\partial L_\mathrm{d}(q_{k-1}, q_k)}{\partial s_k} = -u_k\,, \tag{57}$$

where u_k is the control input. Similarly, the discrete controlled Lagrangian is,

$$L_\mathrm{d}^{\tau,\sigma}(q_k, q_{k+1}) = hL^{\tau,\sigma}(q_{k+1/2}, \Delta q_k/h)\,,$$

with discrete dynamics given by,

$$\frac{\partial L_\mathrm{d}^{\tau,\sigma}(q_k, q_{k+1})}{\partial \theta_k} + \frac{\partial L_\mathrm{d}^{\tau,\sigma}(q_{k-1}, q_k)}{\partial \theta_k} = 0\,, \tag{58}$$

$$\frac{\partial L_\mathrm{d}^{\tau,\sigma}(q_k, q_{k+1})}{\partial s_k} + \frac{\partial L_\mathrm{d}^{\tau,\sigma}(q_{k-1}, q_k)}{\partial s_k} = 0\,. \tag{59}$$

Equation (59) is equivalent to the *discrete controlled momentum conservation:*

$$p_k = \mu\,,$$

where

$$p_k = -\frac{\partial L_\mathrm{d}^{\tau,\sigma}(q_k, q_{k+1})}{\partial s_k} = \frac{(1+\gamma\kappa)\beta(\theta_{k+1/2})\Delta\theta_k + \gamma\Delta s_k}{h}\,.$$

Setting

$$u_k = -\frac{\gamma\Delta\theta_k\tau(\theta_{k+1/2}) - \gamma\Delta\theta_{k-1}\tau(\theta_{k-1/2})}{h}$$

makes Eqs. (57) and (59) identical and allows one to represent the discrete momentum equation (57) as the discrete momentum conservation law $p_k = p$.

The condition that (56)–(57) are equivalent to (58)–(59) yield the discrete matching conditions. The dynamics determined by Eqs. (56)–(57) restricted to the momentum level $p_k = p$ is equivalent to the dynamics of Eqs. (58)–(59) restricted to the momentum level $p_k = \mu$ if and only if the matching conditions

$$\sigma = -\frac{1}{\gamma\kappa}, \qquad \mu = \frac{p}{1+\gamma\kappa},$$

hold.

Numerical example: Simulating the behavior of the discrete controlled Lagrangian system involves viewing Eqs. (58)–(59) as an implict update map $(q_{k-2}, q_{k-1}) \mapsto (q_{k-1}, q_k)$. This presupposes that the initial conditions are given in the form (q_0, q_1); however it is generally preferable to specify the initial conditions as $(q_0, \dot{q}_0)$. This is achieved by solving the boundary condition,

$$\frac{\partial L}{\partial \dot{q}}(q_0, \dot{q}_0) + D_1 L_\mathrm{d}(q_0, q_1) + F_\mathrm{d}^-(q_0, q_1) = 0\,,$$

for q_1. Once the initial conditions are expressed in the form (q_0, q_1), the discrete evolution can be obtained using the implicit update map.

We first consider the case of kinetic shaping on a level surface, when κ is twice the critical value, and without dissipation. Here, $h = 0.05$ sec, $m = 0.14$ kg, $M = 0.44$ kg, and $l = 0.215$ m. As shown in Fig. 7, the θ dynamics is stabilized, but since there is no dissipation, the oscillations are sustained. The s dynamics exhibits both a drift and oscillations, as potential shaping is necessary to stabilize the translational dynamics.

Future Directions

Discrete Receding Horizon Optimal Control: The existing work on discrete optimal control has been primarily focused on constructing the optimal trajectory in an open loop sense. In practice, model uncertainty and actuation errors necessitate the use of feedback control, and it would be interesting to extend the existing work on optimal control of discrete systems to the feedback setting by adopting a receding horizon approach.

Discrete State Estimation: In feedback control, one typically assumes complete knowledge regarding the state of the system, an assumption that is often unrealistic in practice. The general problem of state estimation in the context of discrete mechanics would rely on good numerical methods for quantifying the propagation of uncertainty by solving the Liouville equation, which describes the evolution of a phase space distribution function advected by

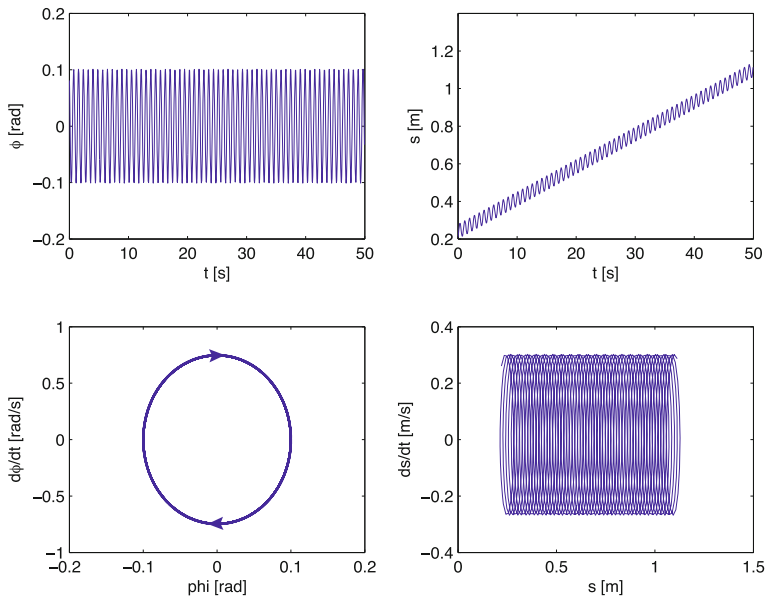

Discrete Control Systems, Figure 7

Discrete controlled dynamics with kinetic shaping and without dissipation. The discrete controlled system stabilizes the θ motion about the equilibrium, but the s dynamics is not stabilized; since there is no dissipation, the oscillations are sustained

a prescribed vector field. In the setting of Hamiltonian systems, the solution of the Liouville equation can be solved by the method of characteristics (Scheeres et al. [38]). This implies that a collocational approach (Xiu [39]) combined with Lie group variational integrators, and interpolation based on noncommutative harmonic analysis on Lie groups could yield an efficient means of propagating uncertainty, and serve as the basis of a discrete state estimation algorithm.

Forced Symplectic–Energy-Momentum Variational Integrators: One of the motivations for studying the control of Lagrangian systems using the method of controlled Lagrangians is that the method provides a natural candidate Lyapunov function to study the global stability properties of the controlled system. In the discrete theory, this approach is complicated by the fact that the energy of a discrete Lagrangian system is not exactly conserved, but rather oscillates in a bounded fashion.

This can be addressed by considering the symplectic-energy-momentum [20] analogue to the discrete Lagrange-d'Alembert principle,

$$\delta \sum_{k=0}^{N_1} L_{\mathrm{d}}(q_k, q_{k+1}, h_k) = \sum_{k=0}^{N-1} [F_{\mathrm{d}}^{-}(q_k, q_{k+1}, h_k) \cdot \delta q_k + F_{\mathrm{d}}^{+}(q_k, q_{k+1}, h_k) \cdot \delta q_{k+1}] \,,$$

where the timestep h_k is allowed to vary, and is chosen to satisfy the variational principle. The variations in h_k yield an Euler–Lagrange equation that reduces to the conservation of discrete energy in the absence of external forces. By developing a theory of controlled Lagrangians around a geometric integrator based on the symplectic-energy-momentum version of the Lagrange–d'Alembert principle, one would potentially be able to use Lyapunov techniques to study the global stability of the resulting numerical control algorithms.

Acknowledgments

TL and ML have been supported in part by NSF Grant DMS-0504747 and DMS-0726263. TL and NHM have been supported in part by NSF Grant ECS-0244977 and CMS-0555797.

Bibliography

Primary Literature

1. Auckly D, Kapitanski L, White W (2000) Control of nonlinear underactuated systems. Commun Pure Appl Math 53:354–369
2. Benettin G, Giorgilli A (1994) On the Hamiltonian interpolation of near to the identity symplectic mappings with application to symplectic integration algorithms. J Stat Phys 74:1117–1143
3. Betts JT (2001) Practical Methods for Optimal Control Using Nonlinear Programming. SIAM, Philadelphia, PA
4. Bloch AM (2003) Nonholonomic Mechanics and Control. In: Interdisciplinary Applied Mathematics, vol 24. Springer, New York

5. Bloch AM, Leonard N, Marsden JE (1997) Matching and stabilization using controlled Lagrangians. In: Proceedings of the IEEE Conference on Decision and Control. Hyatt Regency San Diego, San Diego, CA, 10–12 December 1997, pp 2356–2361
6. Bloch AM, Leonard N, Marsden JE (1998) Matching and stabilization by the method of controlled Lagrangians. In: Proceedings of the IEEE Conference on Decision and Control. Hyatt Regency Westshore, Tampa, FL, 16–18 December 1998, pp 1446–1451
7. Bloch AM, Leonard N, Marsden JE (1999) Potential shaping and the method of controlled Lagrangians. In: Proceedings of the IEEE Conference on Decision and Control. Crowne Plaza Hotel and Resort, Phoenix, AZ, 7–10 December 1999, pp 1652–1657
8. Bloch AM, Leonard NE, Marsden JE (2000) Controlled Lagrangians and the stabilization of mechanical systems I: The first matching theorem. IEEE Trans Syst Control 45:2253–2270
9. Bloch AM, Chang DE, Leonard NE, Marsden JE (2001) Controlled Lagrangians and the stabilization of mechanical systems II: Potential shaping. IEEE Trans Autom Contr 46:1556–1571
10. Bloch AM, Leok M, Marsden JE, Zenkov DV (2005) Controlled Lagrangians and stabilization of the discrete cart-pendulum system. In: Proceedings of the IEEE Conference on Decision and Control. Melia Seville, Seville, Spain, 12–15 December 2005, pp 6579–6584
11. Bloch AM, Leok M, Marsden JE, Zenkov DV (2006) Controlled Lagrangians and potential shaping for stabilization of discrete mechanical systems. In: Proceedings of the IEEE Conference on Decision and Control. Manchester Grand Hyatt, San Diego, CA, 13–15 December 2006, pp 3333–3338
12. Bryson AE, Ho Y (1975) Applied Optimal Control. Hemisphere, Washington, D.C.
13. Chang D-E, Bloch AM, Leonard NE, Marsden JE, Woolsey C (2002) The equivalence of controlled Lagrangian and controlled Hamiltonian systems. Control Calc Var (special issue dedicated to Lions JL) 8:393–422
14. Hairer E, Lubich C, Wanner G (2006) Geometric Numerical Integration, 2nd edn. Springer Series in Computational Mathematics, vol 31. Springer, Berlin
15. Hamberg J (1999) General matching conditions in the theory of controlled Lagrangians. In: Proceedings of the IEEE Conference on Decision and Control. Crowne Plaza Hotel and Resort, Phoenix, AZ, 7–10 December 1999, pp 2519–2523
16. Hamberg J (2000) Controlled Lagrangians, symmetries and conditions for strong matching. In: Lagrangian and Hamiltonian Methods for Nonlinear Control. Elsevier, Oxford
17. Hussein II, Leok M, Sanyal AK, Bloch AM (2006) A discrete variational integrator for optimal control problems in *SO*(3). In: Proceedings of the IEEE Conference on Decision and Control. Manchester Grand Hyatt, San Diego, CA, 13–15 December 2006, pp 6636–6641
18. Iserles A, Munthe-Kaas H, Nørsett SP, Zanna A (2000) Lie-group methods. In: Acta Numerica, vol 9. Cambridge University Press, Cambridge, pp 215–365
19. Junge D, Marsden JE, Ober-Blöbaum S (2005) Discrete mechanics and optimal control. In: IFAC Congress, Praha, Prague, 3–8 July 2005
20. Kane C, Marsden JE, Ortiz M (1999) Symplectic-energy-momentum preserving variational integrators. J Math Phys 40(7):3353–3371
21. Kane C, Marsden JE, Ortiz M, West M (2000) Variational integrators and the Newmark algorithm for conservative and dissipative mechanical systems. Int J Numer Meth Eng 49(10):1295–1325
22. Kelley CT (1995) Iterative Methods for Linear and Nonlinear Equations. SIAM, Philadelphia, PA
23. Lee T, Leok M, McClamroch NH (2005) Attitude maneuvers of a rigid spacecraft in a circular orbit. In: Proceedings of the American Control Conference. Portland Hilton, Portland, OR, 8–10 June 2005, pp 1742–1747
24. Lee T, Leok M, McClamroch NH (2005) A Lie group variational integrator for the attitude dynamics of a rigid body with applications to the 3D pendulum. In: Proceedings of the IEEE Conference on Control Applications. Toronto, Canada, 28–31 August 2005, pp 962–967
25. Lee T, Leok M, McClamroch NH (2006) Optimal control of a rigid body using geometrically exact computations on SE(3). In: Proceedings of the IEEE Conference on Decision and Control. Manchester Grand Hyatt, San Diego, CA, 13–15 December 2006, pp 2710–2715
26. Lee T, Leok M, McClamroch NH (2007) Lie group variational integrators for the full body problem. Comput Method Appl Mech Eng 196:2907–2924
27. Lee T, Leok M, McClamroch NH (2007) Lie group variational integrators for the full body problem in orbital mechanics. Celest Mech Dyn Astron 98(2):121–144
28. Leimkuhler B, Reich S (2004) Simulating Hamiltonian Dynamics. Cambridge Monographs on Applied and Computational Mathematics, vol 14. Cambridge University Press, Cambridge
29. Leok M (2004) Foundations of Computational Geometric Mechanics. Ph D thesis, California Institute of Technology
30. Marsden JE, Ratiu TS (1999) Introduction to Mechanics and Symmetry, 2nd edn. Texts in Applied Mathematics, vol 17. Springer, New York
31. Marsden JE, West M (2001) Discrete mechanics and variational integrators. In: Acta Numerica, vol 10. Cambridge University Press, Cambridge, pp 317–514
32. Marsden JE, Pekarsky S, Shkoller S (1999) Discrete Euler–Poincarée and Lie–Poisson equations. Nonlinearity 12(6):1647–1662
33. Maschke B, Ortega R, van der Schaft A (2001) Energy-based Lyapunov functions for forced Hamiltonian systems with dissipation. IEEE Trans Autom Contr 45:1498–1502
34. Moser J, Veselov AP (1991) Discrete versions of some classical integrable systems and factorization of matrix polynomials. Commun Math Phys 139:217–243
35. Ortega R, Spong MW, Gómez-Estern F, Blankenstein G (2002) Stabilization of a class of underactuated mechanical systems via interconnection and damping assignment. IEEE Trans Autom Contr 47:1218–1233
36. Sanz-Serna JM (1992) Symplectic integrators for Hamiltonian problems: an overview. In: Acta Numerica, vol 1. Cambridge University Press, Cambridge, pp 243–286
37. Scheeres DJ, Fahnestock EG, Ostro SJ, Margot JL, Benner LAM, Broschart SB, Bellerose J, Giorgini JD, Nolan MC, Magri C, Pravec P, Scheirich P, Rose R, Jurgens RF, De Jong EM, Suzuki S (2006) Dynamical configuration of binary near-Earth asteroid (66391) 1999 KW4. Science 314:1280–1283
38. Scheeres DJ, Hsiao F-Y, Park RS, Villac BF, Maruskin JM (2006) Fundamental limits on spacecraft orbit uncertainty and distribution propagation. J Astronaut Sci 54:505–523

39. Xiu D (2007) Efficient collocational approach for parametric uncertainty analysis. Comm Comput Phys 2:293–309
40. Zenkov DV, Bloch AM, Leonard NE, Marsden JE (2000) Matching and stabilization of low-dimensional nonholonomic systems. In: Proceedings of the IEEE Conference on Decision and Control. Sydney Convention and Exhibition Centre, Sydney, NSW Australia; 12–15 December 2000, pp 1289–1295
41. Zenkov DV, Bloch AM, Leonard NE, Marsden JE (2002) Flat nonholonomic matching. In: Proceedings of the American Control Conference. Hilton Anchorage, Anchorage, AK, 8–10 May 2002, pp 2812–2817

Books and Reviews

Bloch AM (2003) Nonholonomic Mechanics and Control. Interdisciplinary Appl Math, vol 24. Springer
Bullo F, Lewis AD (2005) Geometric control of mechanical systems. Texts in Applied Mathematics, vol 49. Springer, New York
Hairer E, Lubich C, Wanner G (2006) Geometric Numerical Integration, 2nd edn. Springer Series in Computational Mathematics, vol 31. Springer, Berlin
Iserles A, Munthe-Kaas H, Nørsett SP, Zanna A (2000) Lie-group methods. In: Acta Numerica, vol 9. Cambridge University Press, Cambridge, pp 215–365
Leimkuhler B, Reich S (2004) Simulating Hamiltonian Dynamics. Cambridge Monographs on Applied and Computational Mathematics, vol 14. Cambridge University Press, Cambridge
Marsden JE, Ratiu TS (1999) Introduction to Mechanics and Symmetry, 2nd edn. Texts in Applied Mathematics, vol 17. Springer
Marsden JE, West M (2001) Discrete mechanics and variational integrators. In Acta Numerica, vol 10. Cambridge University Press, Cambridge, pp 317–514
Sanz-Serna JM (1992) Symplectic integrators for Hamiltonian problems: an overview. In: Acta Numerica, vol 1. Cambridge University Press, Cambridge, pp 243–286

Disordered Elastic Media

Thierry Giamarchi
DPMC-MaNEP, University of Geneva,
Geneva, Switzerland

Article Outline

Glossary
Definition of the Subject
Introduction
Static Properties of Disordered Elastic Media
Pinning and Dynamics
Future Directions
Bibliography

Glossary

Pinning An action exerted by impurities on an object. The object has a preferential position in space, and will move in response to an external force only if this force is large enough.

Scaling The fact that each of two quantities varies in a power-law relationship to the other.

Random manifold A single elastic structure (line, sheet) embedded in a random environment.

Bragg glass A periodic elastic structure embedded in a weakly disordered environment, nearly as ordered as a solid but exhibiting some characteristics normally associated with glasses.

Creep Very slow response at finite temperature of a pinned structure in response to an external force.

Definition of the Subject

Many seemingly different systems, with extremely different microscopic physics, ranging from magnets to superconductors, share the same essential ingredients and can be described under the unifying concept of disordered elastic media. In all these systems, an internal elastic structure, such as an interface between regions of opposite magnetization in magnetic systems, is subject to the effects of disorder existing in the material. A specially interesting feature of all these systems is that these disordered elastic structures can be set in motion by applying an external force on them (e. g. a magnetic field sets in motion a magnetic interface), and that motion will be drastically affected by the presence of the disorder. What properties result from this competition between elasticity and disorder is a complicated problem which constitutes the essence of the physics of disordered elastic media. The resulting physics present characteristics similar to those of glasses. This poses extremely challenging fundamental questions in determining the static and dynamic properties of these systems. Understanding both the static and dynamic properties of these objects is not only an important question from a fundamental point of view, but also has strong practical applications. Indeed, being able to write an interface between two regions of magnetization or polarization, and the speed of writing and stability of such regions, is what conditions, in particular, our ability to store information in such systems, as (for example) recordings on a magnetic hard drive. The physics pertaining to disordered elastic media directly condition how we can use these systems for practical applications.

Introduction

Understanding the statics and dynamics of elastic systems in a random environment is a long-standing problem with important applications for a host of experimental systems. Such problems can be split into two broad categories: (a)

propagating interfaces such as magnetic [1,2,3,4], spintronic [5,6], or ferroelectric [7,8] domain walls, fluid invasion in porous media [9], contact lines in wetting [10], epitaxial growth [11] or crack propagation [12,13]; (b) periodic systems such as vortex lattices in type II superconductors [14,15,16], charge density waves [17,18], magnetic bubbles [19], colloids [20], Wigner crystals of classical particles [21] or of electrons [22,23].

Although all these systems have very different microscopic descriptions, one aspect of their physics is identical at a more macroscopic scale: An object exists that obeys a macroscopic elastic description. For case (a) this is an interface separating two different regions of the system, for example in a magnetic material a domain wall separating regions of opposite magnetization. An example of such an interface is shown in Fig. 1. Since creating an interface costs energy, the interface left to itself would tend to be flat, and there is an elastic cost to its deformations. Since this object lives inside a microscopic crystal with disorder, it is also subjected to potentials that tend to roughen it and pin it in specific regions of space. This interface can be set in motion by applying an external force on it, caused for example by a magnetic field for the magnetic domain wall or an electric field for a ferroelectric. For case (b), that of periodic systems, a similar physics exists. A "crystal" of objects (lines for vortices, points for magnetic bubbles and colloids, or sheets for charge-density waves) exists inside the system. Since these objects repel each other, in the absence of disorder they would tend to form a perfect periodic crystal. In a way similar to the interfaces, such crystals can be set in motion by applying an external force (for example a current, in the case of vortices). The disorder present at the microscopic level tends to pin this crystal. It is important to note that one is dealing here with the physics of a crystal embedded in an external medium containing impurities. The disorder can thus vary at a length scale much smaller than the lattice spacing of the moving crystal, which leads to a physics radically novel compared to that of chemical impurities in a regular solid. Understanding the physics, both static and dynamic, of these objects has thus been a considerable challenge in the last 50 years or so. There are several reasons for pursuing this interest, and for tackling the challenges posed by this field of disordered elastic media.

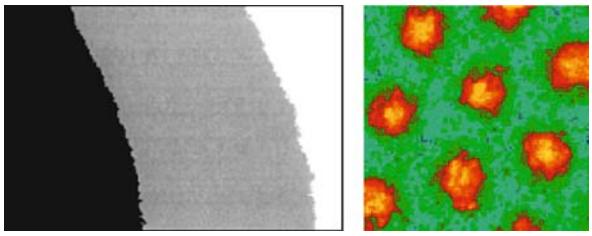

Disordered Elastic Media, Figure 1
Left: **an interface in a magnetic system separating two different magnetic polarizations (*dark and white*). The image is $90 \times 72\,\mu m^2$. The roughness of the domain wall due to the presence of disorder in the system is obvious on the image. Two positions of the interface are shown. *Dark* and *gray* correspond to two consecutive images after the interface has been pulled to the right by applying a magnetic field to the sample favoring the magnetization direction on the left of the interface. Such a magnetic field acts as a force pulling the domain wall. [From [1] (Copyright 1998 by the American Physical Society)]. *Right:* A vortex lattice image, from Scanning Tunneling Microscope, in the superconductor MgB_2. The tips of the vortices on the surface of the sample are shown in the image, and correspond to the *red parts*. The image is about $250\,nm^2$. In a perfectly pure system, the vortex lattice is a periodic arrangement (here in a triangular lattice) of objects of a given size (here the core size of the vortex). Disorder affects this perfectly periodic arrangement over a large distance. [From [24] (Copyright 2002 by the American Physical Society)]**

First, at the fundamental level these systems pose difficult and important questions. It has been known since the 1970s that the presence of disorder is crucial [25] and changes the physics completely. While the resulting models are very difficult to solve, they have contributed to pushing the limits of our understanding of disordered systems, and to developing new techniques of statistical physics to deal with such issues. In particular, it is clear that from the competition between disorder and elasticity emerges a complicated energy landscape with many metastable states. This results in glassy properties [26] such as hysteresis and history dependence of the static configuration. Initially viewed as toy models of glasses, these systems have acquired their own importance and have posed their own challenging questions. Understanding the static properties of such systems has stimulated the development of sophisticated approaches such as replica theory [27], functional renormalization groups [28] and numerical methods. Much progress has recently been accomplished due to both analytical and numerical advances. If the static view allows us to improve our techniques of statistical physics, the dynamic view is even more complicated since most of our theoretical tools fail. These systems thus provide wonderful motivations to develop new techniques to tackle the out-of-equilibrium dynamics of disordered systems and to understand and unify the concepts of out-of-equilibrium physics of glasses.

Second, in addition to this theoretical motivation, the ability to apply these concepts in so many different physical systems is a tremendous motivation and challenge. The various realizations allow us to apply stringent tests to pro-

posed theories and, as we will see, have very often served to kill misguided proposals or to put the theory on the right track in these quite complicated systems. Experiments in these systems can be remarkable by the range they offer. In vortex systems, for example, one can vary the vortex lattice spacing by several orders of magnitude just by changing the magnetic field applied to the sample, something impossible to do on a simple crystal. Similarly, for magnetic domain walls, measurements of the velocity in response to an external force can span about ten orders of magnitude. This interplay and exchange between theory and experiment has fueled the field and contributed greatly to its progress.

Last but not least, the phenomena studied for disordered elastic media have a potential impact for applications. Creating interfaces in magnetic or, more recently, ferroelectric materials is a way to store information (with "0" being one direction of magnetization or polarization and "1" being the reverse). This idea is at the root of information storage in a magnetic hard drive or in ferroelectric or magnetic bubble memory. How well one can store the information is thus directly related to both the static and dynamic properties of such interfaces. In particular, stability of the written information is ensured only if the interface is pinned and will not meander due to outside forces, such as thermal agitation. Similarly, there has been much interest recently on spintronic materials where the magnetic properties can be manipulated by applying electrical currents [5,6]. In the same vein, multiferroic materials [29] allow manipulation of ferroelectric properties by the application of magnetic fields. How the disordered interfaces behave in such materials will certainly determine their possible use for information technology. In a similar way, the vortex lattice in a superconductor is set in motion by the application of a current, while its motion generates a voltage. The dynamic properties of the vortex lattice, and how well it is pinned, thus directly affect the absence of resistance of a superconductor [14,15,16], hence its potential uses. Other examples, such propagation of fractures, clearly show the potential importance of such phenomena for applications.

This chapter presents basic concepts and results in this very active field. Section "Interfaces and Basic Concepts" presents the basic concepts and discusses the static properties of interfaces and domain walls. Section "Periodic Systems and Bragg Glass" deals with the periodic systems and their differences compared to interfaces. Section "Pinning and Dynamics" presents concepts and important questions for the dynamics of disordered elastic media, with focus on depinning in Sect. "Depinning", on large velocity behavior in Sect. "High Velocity Phase" and responses to small external forces in Sect. "Small Applied Force and Creep Motion". Finally, Sect. "Future Directions" discusses future directions of and perspectives on the field.

Static Properties of Disordered Elastic Media

Interfaces and Basic Concepts

This section introduces the basic ingredients of the systems under study and discusses the specific case of interfaces. An interface is a sheet of dimension d living in a space of D dimensions. For realistic interfaces $D = d + 1$, but generalizations are of course possible. If we let r represent the internal coordinate of the interface and z all its transverse directions, the interface position is described by displacement $u(r)$ from a flat configuration. This totally determines the shape of the interface provided that u is univalued, i. e., that there are no overhangs or bubbles. The modelization of a one dimensional interface ($d = 1$) in a two dimensional film is shown in Fig. 2.

Since interface distortions cost elastic energy, a system's zero temperature equilibrium configuration in the absence of disorder is flat. Deviations from this equilibrium position are described by a Hamiltonian $H[u]$ which is a function of the displacement u. For small displace-

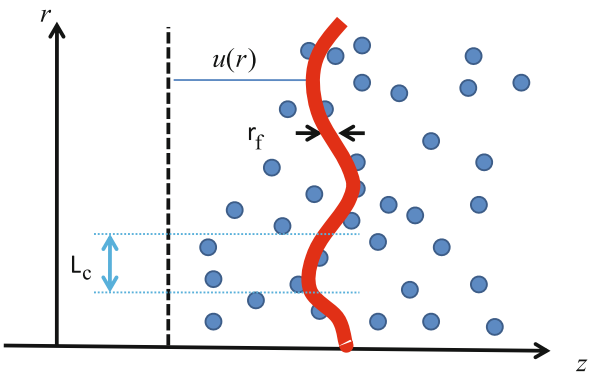

Disordered Elastic Media, Figure 2
A one dimensional interface (such as a magnetic domain wall), shown in *red*, living in a two dimensional space (film). The position of the interface is determined (provided there are no overhangs or bubbles) by the displacement $u(r)$ from a flat configuration, indicated by the *dashed line*. In the absence of disorder, denoted by the *blue dots*, which pin the line in preferred positions in space, the line would be flat. The competition between elasticity and disorder leads to the physics of disordered elastic media and to glassy properties. The thickness of the line, denoted r_f, or the correlation length of the disorder define the Larkin length L_c for which the relative displacements are of the order of r_f, namely $u(L_c) - u(0) \sim r_f$

ments one can make the usual elastic approximation

$$H[u] = \frac{1}{2}\int \frac{\mathrm{d}^d q}{(2\pi)^d} c(q) u_q^* u_q \,, \tag{1}$$

where u_q is the Fourier transform of $u(r)$ and $c(q)$ are the so-called elastic coefficients. If the elastic forces acting on the interface are short ranged then one has $c(q) = cq^2$ which corresponds to

$$H[u] = \frac{c}{2}\int \mathrm{d}^d r (\nabla u(r))^2 \,. \tag{2}$$

For some interfaces where long range interactions play a role, different forms of elasticity are possible. This is particularly the case when dipolar forces [30] are taken into account [8] or for the contact line in wetting [31] and crack propagation [32].

In addition to elastic energy, the interface gains some energy by coupling to the disorder. Two universality classes for the disorder exist (see Fig. 3). The random bond disorder corresponds to impurities that directly attract or repel the interface. In contrast, for random field disorder the pinning energy is affected by all the randomness that the interface has encountered in its previous motion. On a more technical level, random bond disorder couples in a symmetric way to the two order parameters on each side of the domain wall, while random field disorder introduces an asymmetry between these two inequivalent order parameters. If $V(r, z)$ denotes the random potential generated by the impurities the pinning energy:

$$H_{\mathrm{dis}}[u] = \int \mathrm{d}^d r \begin{cases} V(r, u(r)) & \text{random bond} \\ \int_0^{u(r)} \mathrm{d}z V(r, z) & \text{random field.} \end{cases} \tag{3}$$

As is obvious from Fig. 3, even if the microscopic disorder is short-range correlated, as in the case of the random field disorder, the fact that the energy of the system integrates between two positions of the interfaces means that long range correlations exist if one considers the description only in terms of the interface. The full Hamiltonian given by (1) and (3) describes the properties of disordered elastic media, and despite its apparent simplicity hides an extremely rich physics.

The competition between disorder and elasticity manifests itself in several properties of the interface. From an energetic point of view, this competition leads to a complicated energy landscape for the configurations of the system, with many metastable states leading to glassy properties. The competition also manifests itself in the shape of the interface. In particular, it deviates from the flat configuration and becomes rough. From the scaling of the relative displacements correlation function, a roughness exponent ζ can be defined from the correlation function of the displacements

$$B(r) = \overline{\langle [u(r) - u(0)]^2 \rangle} \propto r^{2\zeta} \,, \tag{4}$$

where $\langle\ \rangle$ denotes thermodynamic average and $\overline{\cdots}$ denotes disorder average. There are relations between the shape of the line and the energetic properties. In particular, (4) suggests that displacements would scale with distance as $u(L) \sim L^{\zeta}$. (2) suggests that the energy of a sample of size L fluctuates from sample to sample as

$$\Delta F(L) \sim L^{d-2+2\zeta} \,. \tag{5}$$

Given the complexity of the problem, several approximate methods have been put forward to portray the role

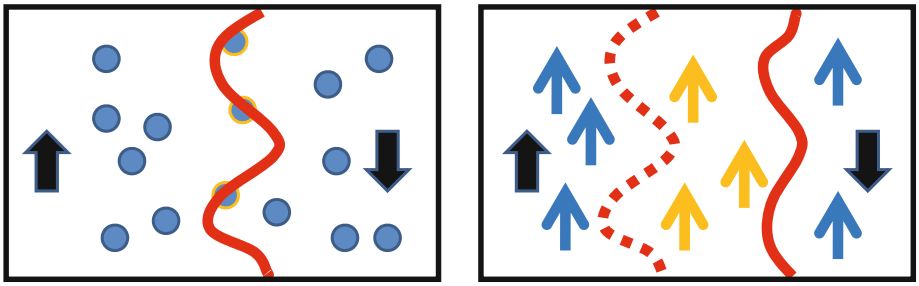

Disordered Elastic Media, Figure 3

The two universality classes of disorder (the names come from the magnetic realization of such systems). In both figures the domain wall in *red* separates two regions with different order parameters, denoted by the two *thick black arrows*. *Left:* random bond disorder. The impurities, denoted in *blue*, couple symmetrically to the two sides of the domain wall, thus only the impurities on the domain wall (denoted with the *orange circle*) contribute to the energy. *Right:* random field disorder. The impurities, denoted by the *blue arrows*, favor one of the two sides. Thus, all the impurities (denotes in *orange*) between two configurations of the domain wall contribute to the energy. This leads to a disorder seen by the domain wall which has long range correlations, even if the microscopic disorder is short-range correlated

of disorder. A remarkable model by which to probe the physics of such systems was introduced by Larkin [25], and goes by the name of the Larkin model. The idea is to focus on short length scale properties. In that case, the displacements are small and one can expand the disorder term in power of the displacements

$$H_{\text{dis}} = \int \mathrm{d}^d r V(r, u(r)) \simeq \int \mathrm{d}^d r [V(r,0) + \nabla_r V(r,0)|_{z=0}\, u(r)] \,. \quad (6)$$

The first term is a trivial constant and the second one indicates that the interface is subject to a random force

$$H_L = \int \mathrm{d}^d r f(r) u(r) \,. \quad (7)$$

Although this model has several pathologies, it has the advantage of being quadratic in the displacement field u and thus of being exactly solvable. It shows that below $d = 4$, disorder plays a major role. The displacements grow as a function of distance and

$$B(r) = r^{4-d} \,. \quad (8)$$

This confirms that there is algebraic roughening of the interface with the displacements growing as a power law of distance. One can define the scaling $u(L) \sim L^{\zeta}$, with the Larkin model giving $\zeta = (4-d)/2$. Below $d = 4$, disorder is relevant and drastically modifies the physical properties of the interface compared to those of the non-disordered one. In addition to the exponent itself, since the displacements grow unboundedly, there exists a length scale, L_c, called the Larkin length, at which the displacements become of the order of the only characteristic scale available, namely either the correlation length of the random potential or the size of the interface r_f, as shown in Fig. 2. Clearly, this is also the length where the applicability of the Larkin model breaks down, since beyond that length the potential $V(r,z)$ is not smooth anymore and thus expansion in powers of u is not justified. Beyond this length, the system will thus truly feel the effects of random potential. One can thus expect metastability, glassy effects and pinning to appear above that length. One has to determine the physics of this regime appearing above the Larkin length, which we will call the random manifold regime. The Larkin length is thus an important length scale for the static properties since it separates two different regimes for the interface. As I will discuss in Sect. "Pinning and Dynamics", the Larkin length also has considerable consequences for the dynamics.

To solve the problem in the random manifold regime is not easy and requires greatly sophisticated techniques of statistical physics. To get a rough idea, one can simply use a scaling argument, known as the Flory argument [14]. At scale L, the elastic energy scales as $cL^{d-2}u(L)^2$. To estimate the disorder term is more complicated but one can assume that if L is large enough, one sums random variables $V(r,z)$. If one considers for example random bond disorder, because the disorder is short-range correlated, one has

$$\overline{V(r_1,u_1)V(r_2,u_2)} = D\delta^m(u_1-u_2)\delta^d(r_1-r_2) \,. \quad (9)$$

Since a $\delta(r)$ function has the dimension of $1/r^d$, this leads to the scaling

$$V(r,u) \sim D^{1/2} u(L)^{-m/2} L^{-d/2} \,, \quad (10)$$

where m is the number of components of u (for an interface $m = 1$). The disorder term thus scales as

$$H_{\text{dis}} = D^{1/2} u(L)^{-m/2} L^{d/2} \,. \quad (11)$$

Balancing the elastic and disorder terms leads to a scaling $u(L) \sim L^{\zeta}$ with $\zeta = \frac{4-d}{4+m}$ for the random bond case and $\zeta = \frac{4-d}{4-m}$ for the random field case. This argument suggests that even in the random manifold, the interface remains rough, with unbounded displacements growing with an *algebraic* roughness. The value of the exponent is characteristic of the universality class of the disorder, and different from the one occurring below $L < L_c$ where the Larkin model applies.

Clearly this simple argument needs to be substantiated by more rigorous calculations. Because the system is subjected to a random potential and the metastability and glassy effects matter, one can use the techniques traditionally used for disordered systems and spin glasses. For a one dimensional interface, this problem can be solved exactly and the roughness exponent $\zeta = 2/3$ obtained [33,34]. Note that this exact value is, of course, slightly different from the mean field estimate. In higher dimensions three main methods have been used to tackle this problem. The first one uses the so-called replica trick [26] to average over the disorder and then a variational approach to solve the corresponding field theory [27]. In this method the initial symmetry between replicas is broken, something familiar in spin-glasses and characteristics of glasses with many metastable minima in the energy. This approximate method gives back the Flory exponent. The second method applies the traditional renormalization technique (RG), so successful for standard critical phenom-

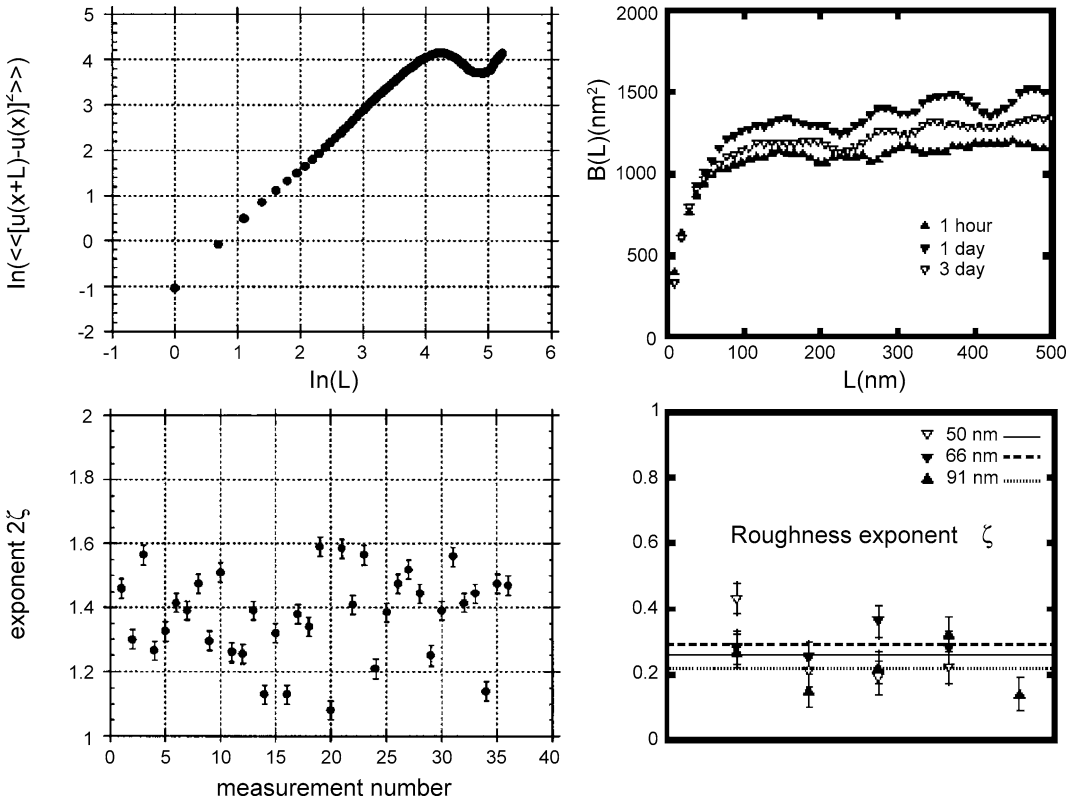

Disordered Elastic Media, Figure 4

Measurements of the roughness exponent ζ in two experimental system. These two examples show the algebraic roughness of the domain walls. The top figures are the correlation function (4) of the displacements $B(r)$, and the bottom ones the measured value of the roughness exponent ζ. *Left:* Magnetic domain walls in thin magnetic films. An exponent of $\zeta \sim 0.6$ is measured compatible with the value $\zeta = 2/3$ expected for a one dimensional wall in a two dimensional space. [From [1] (Copyright 1998 by the American Physical Society)]. *Right:* Ferroelectric domain wall in a ferroelectric film. An exponent of $\zeta = 0.26$ is measured. This value is compatible with the value expected for a two dimensional wall in a three dimensional space in the presence of long range dipolar interactions. [From [8] (Copyright 2005 by the American Physical Society)]

ena. This consists in looking at the problem at larger and larger length scales, eliminating degrees of freedom, while changing the Hamiltonian to ensure that the large length scale physics remain invariant. Usually one can expand the interaction potential, and only a few terms are relevant, which means that the RG consists in the flow of a small number of coupling constants. In the case of disordered elastic systems, the task is considerably more complex since all powers in the expansion of the correlator of the disorder have the same scaling dimension. During the flow the *whole correlator of the disorder* is modified. One must thus follow the renormalization of a whole function, hence the name functional renormalization group (FRG) [28]. This leads to a remarkable property: Beyond a length scale that coincides with the Larkin length, the disorder correlator, initially a smooth analytic function, becomes non-analytic and develops a cusp. The appearance of this non-analyticity is, in this method, the signal of glassy physics. FRG allows us to obtain the roughening exponent in a systematic expansion in $\epsilon = 4 - d$. This has been worked out for the moment up to second order in $\epsilon = 4 - d$, leading to $\zeta = 0.20829804\epsilon + 0.0068582\epsilon^2$ and $\zeta = \epsilon/3$ for the random bond and random field disorder, respectively [35]. Note that for the random field the mean-field (Flory) exponent is exact due to the long range nature of the disorder. In addition to these analytical approaches, a very useful approach is provided by numerical studies of such systems, using either molecular dynamics simulations [36], Monte Carlo techniques [37,38], or specially designed algorithms [39,40]. Numerical approaches are of course quite challenging due to the glassy nature of the system with many metastable minima close to the ground state. However they have proven quite useful in obtaining not only the asymptotic regime but also

the full crossover between the Larkin and random manifold regimes, as well as in incorporating the effects of finite temperature.

These predictions can be verified experimentally. I show in Fig. 4 the roughness exponent as measured in a magnetic and a ferroelectric film. The algebraic growth of the correlation function $B(r)$ is clearly seen. These two experimental situations correspond to two different dimensionalities for the domain walls, due to the different thicknesses of the material and different characteristics of the domain wall.

Now we can say that we have a rather good understanding of the static properties of the interfaces, at least for the simple case of local elasticity shown here. Of course even for the statics this is not the end of the story, since several microscopic systems such as the contact line of a fluid or ferroelectric systems, have long range interactions (dipolar interactions) making even the static properties quite challenging to determine. Other open questions will be discussed in Sect. "Future Directions".

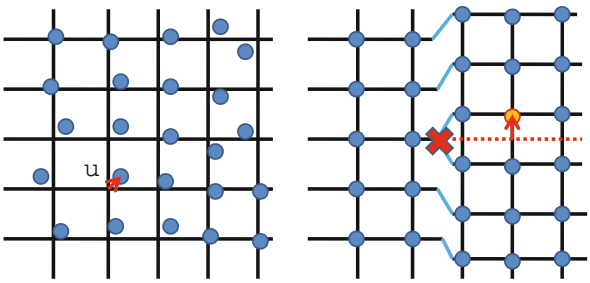

Disordered Elastic Media, Figure 5
For a periodic system the ability to define a displacement $u(r)$ for the objects (*blue dots*) compared to the perfect lattice (here a square lattice corresponding to the intersections of the *black lines*), necessitates the absence of topological defects. *Left:* if there are no topological defects one can associate, for each site R_i^0 of the perfect lattice, a displacement u_i, as indicated by the *red arrow*. *Right:* here there is a topological defect, denoted by the *red cross*, corresponding to the addition of one line of particles. In this case the displacement u_i is not univalued. From the point of view of the particles on the left of the topological defect, the *orange* particle has a displacement u of half a lattice spacing, while one could surmise $u = 0$ looking only at particles on the right of the topological defect

Periodic Systems and Bragg Glass

Similar concepts apply directly to the case of periodic systems. In all these systems the constituent elements (lines for vortices, points for colloids and magnetic bubbles, sheets for phase maxima in the charge density wave systems) form a solid that is embedded into the microscopic system but can have widely different characteristics and, in particular, widely different lattice spacing. For example, in the case of vortices, lattice spacing is controlled by the magnetic field and can easily be varied. An important characteristic is that, in a similar fashion to the interface, this crystal can be embedded in the "external" disorder that corresponds to the imperfections of the real microscopic lattice in which this artificial crystal lives. It is important to note that the variation of the disorder potential can thus occur at length scales much smaller than the lattice spacing.

Each point of the system can be described by an equilibrium position R_i^0 forming a perfect lattice (usually triangular for the vortex lattice), and a displacement u_i relative to this equilibrium position, as shown on Fig. 5. As for the interfaces, the interactions between the objects forming the crystal favor a perfectly ordered crystal. The energy of the system can be expanded for small deviations and lead to a quadratic expansion in u characteristic of an elastic energy. It is important to note that for such an expansion to be valid, it is only necessary for the relative displacements $u_i - u_j$ between two neighbors to be small. The displacements themselves can be arbitrary; for example translating the whole crystal by a uniform displacement does not change the energy.

$$H = \frac{1}{2}\sum_{ij} C_{ij}(u_i - u_j)^2 \,, \tag{12}$$

where the C_{ij} are the elastic coefficients of the system. Since the interactions can be long range, the elastic coefficients are not necessarily limited to nearest neighbor only. Note that for such an expansion to be meaningful, it is necessary for the displacements to be uniquely defined. This assumes that there are no topological defects (dislocations, disclinations) in the crystal. Indeed in the presence of such defects, as shown in Fig. 5 the displacements have two different values when circling around the defect. In order to use the elastic approximation it is thus important to ascertain that topological defects are not generated. I will come back to this crucial point below.

As with the interfaces, the minimum energy configuration is the perfect crystal with all $u_i = 0$. In the absence of disorder this perfect crystal can be affected only by thermal fluctuations. If temperature becomes too large the crystal will melt. A rule of thumb for melting is when the relative displacements between two neighbors become a sizeable fraction of the lattice spacing

$$\langle (u_i - u_{i+1})^2 \rangle = C_L^2 a^2 \,, \tag{13}$$

where $\langle \cdots \rangle$ denotes the thermal average and C_L is a phenomenological constant which turns out to be of the or-

der of $C_L \sim 0.1$ to reproduce reasonable values for the observed melting of solids. This rule of thumb, called the Lindemann criterion for melting, gives, in fact, quite decent results.

In the presence of disorder, one must add to the elastic energy term (12) the energy coming from the random potential created by the disorder. This takes the form

$$H_{\text{dis}} = \int \mathrm{d}^d r V(r)\rho(r) = \int \mathrm{d}^d r V(r) \sum_i \delta(r - R_i^0 - u_i) \tag{14}$$

and this term will clearly tend to disorder the crystal.

The case of a periodic system constitutes a specially important and interesting situation. Indeed, the nature of order and the possible phases are more complex than they are for interfaces. An important question is thus whether these two systems are in the same universality class or not. In a general way, the order in a periodic system is characterized by a positional order, indicating that, knowing the position of a reference particle, one can find a particular particle of the solid at a given position. This positional order can be measured by the structure factor, which is the correlation function of the Fourier transform of the density $S(q) = \langle |\rho_q|^2 \rangle$. In a perfect crystal, the structure factor has *divergent* peaks at the position of the reciprocal vectors K_0 of the perfect lattice. The presence of such divergent peaks indicates a perfect positional order. The fact that one sees peaks also indicates the existence of another type of order, namely the orientational order in a solid. This is illustrated in Fig. 6. The orientational order indicates that if the bonds have a certain orientation in a region of space, this orientation is preserved in the other parts of the solid. Losing the orientational order replaces the peaks in the structure factor by a ring since the orientation of a given peak is not defined any more. In standard solids, both orders are usually lost at the same time and the solid melts to a liquid, usually by a first order phase transition. But we also know that in some cases for pure systems, such as two dimensional solids, the melting may occur as a two step process where the positional order is lost first and only then the orientational order, leading to a so called hexatic phase [41]. A summary of the various cases is shown in Fig. 7. In addition to these standard order parameters, a periodic system is also characterized by a topological order corresponding to the fact that the connectivity of the perfect crystal is preserved by small displacements. Such order is determined by a triangulation of the solid and a determination of the topology. If the topology is identical to that of the perfect lattice, it means that

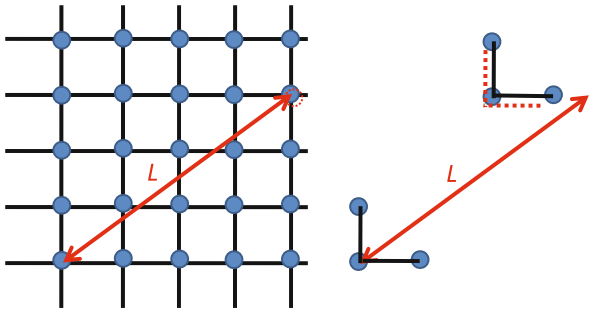

Disordered Elastic Media, Figure 6
A periodic system possesses positional, orientational and topological order. *Left:* positional order. A solid has perfect positional order if, by knowing the position of a reference particle, one can predict the position of a particle at distance *L* as indicated by the *dashed circle* and the *red arrow*. *Right:* orientational order. A solid has orientational order if, by knowing the orientation of the bonds in a region of space, one can predict the orientation at a distance *L* as indicated by the *red dashed lines*. Note that the system need not possess positional order for the orientational order to exist. Topological order: topological order exists if, after a triangulation, the topology (i. e., the number of neighbors) of each point of the solid is fixed and the displacements can be defined in a univalued manner. The picture on the left possesses perfect topological order

the displacements can be defined in a univalued way across the solid. In the liquid, topological defects such as dislocations and disclinations destroy this perfect topological order and the very concept of displacements around an equilibrium position becomes ill defined, since in that case the displacement field is no longer univalued.

As for interfaces, disorder changes the properties of the pure elastic system. In order to take into account the effect of disorder on periodic systems, it is thus important to address two different aspects of the problem: a) the effect of the disorder on an approximation of the real system given by elastic theory; b) whether the disorder is able to generate topological defects, in which case the very idea of an elastic approximation breaks down and another starting point should be found.

As was shown in the groundbreaking paper by Larkin [25], there always exists for $d \leq 4$, a characteristics length scale L_a for which the displacements become of the order of the lattice spacing a of the perfect crystal. The fact that displacements can become as large as the lattice spacing indicates that *perfect* positional order is lost. The question of *how* this destruction takes place and the nature of the resulting phase is a long standing problem. Given the complexity of the question, no solution existed until recently, but it is interesting to see that the community converged, by inference, on closely related models, to a consensus that was accepted for a long time but eventually

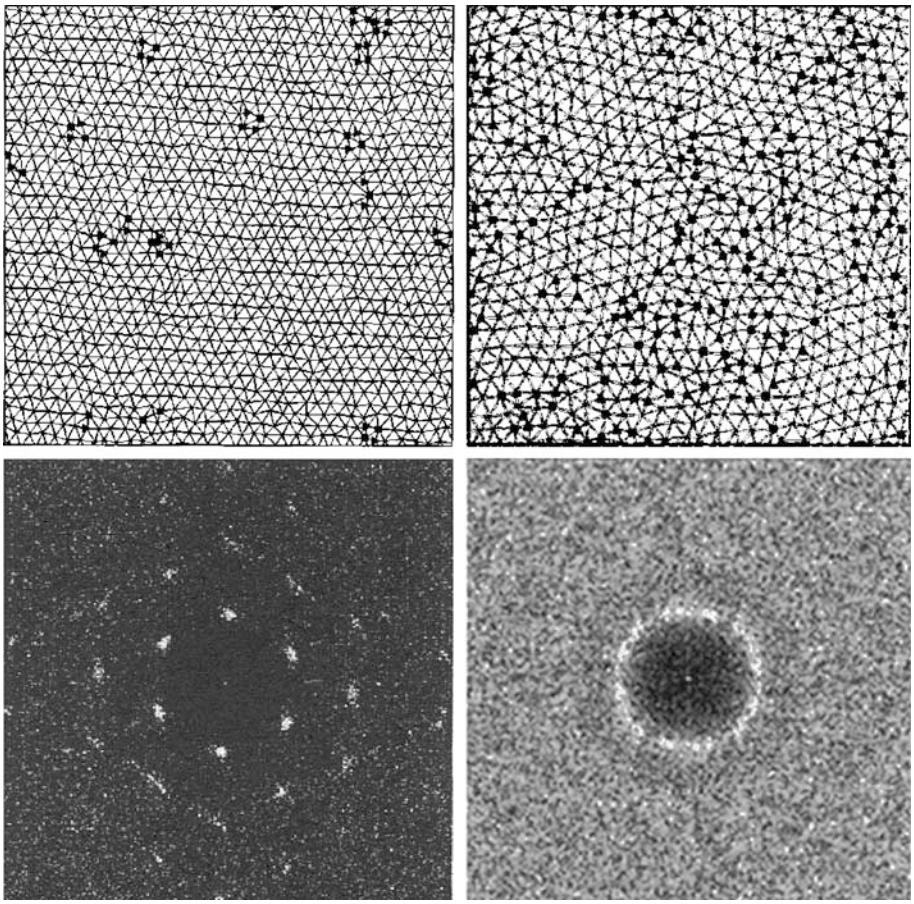

Disordered Elastic Media, Figure 7

Decoration images of vortex lattices, illustrating the difference between solid and liquid phases. The *top figures* are the images in real space, while the *bottom ones* are the structure factor $S(q) = \langle|\rho(q)|\rangle$. *Left:* the system is in a solid-like phase (in fact a Bragg glass phase (see text)). The system possesses good positional order and orientational order. This can be seen both from the pictures in real space and from the structure factor that shows Bragg peaks at the position of the reciprocal vectors K_0 of the perfect underlying lattice. As shown by the *triangulation*, most sites have six neighbors. Topological defects where sites have five or seven neighbors (as indicated by the *triangles* and *square black marks* respectively) do exist, but are paired in 5–7 pairs, making the system free of topological defects at large length scales. *Right:* the system is in a liquid-like phase. Positional order and orientational order are lost. The Bragg peaks are gone and the structure factor has a ring-like structure (indicating the loss of orientational order). Topological defects are proliferating and are unpaired, contributing to the exponential decay of order in the system. [Images from M. Marchevsky, J. Aarts, P.H. Kes (unpublished)]

proved wrong. The route followed was to learn as much as possible from the interfaces. At short distance, one can make an expansion in powers of the displacements, and the system is described by the Larkin model. This ceases to be valid at the Larkin length L_c, for which the displacements become of the order of the size r_f of the particle in the solid. Note that L_c and L_a are in general two different length scales since the size of the particle r_f and the lattice spacing a are usually different. Naturally, one has $L_c < L_a$. For systems such as the vortex lattice or Wigner crystals, the difference can be huge, while for charge-density waves, one expects to have $r_f \sim a$ and thus $L_c \sim L_a$. Below the Larkin length the system is described by the Larkin model and thus by an algebraic growth of the displacements with a roughness exponent of $\zeta = (4 - d)/2$. Above the Larkin length L_c, but for displacements smaller than a (i. e., for lengths smaller than L_a), one can consider that the various objects of the periodic system don't see each other except by their elastic forces. In particular, they do not sample the same random potential given the smallness of the displacements. One thus has a regime very similar to the random manifold regime of the interface. The displacement continues to grow algebraically, $u \sim L^{\zeta}$, albeit with a different exponent. Above L_c, the connection between the growth

of the displacements and the structure factor (the density-density correlations) is non-trivial since the model is non-Gaussian. Indeed the structure factor is given by [42]

$$S(K_0 + q) = \int \mathrm{d}^d q \mathrm{e}^{iqr} C(r) \,, \tag{15}$$

where

$$C(r) = \langle \mathrm{e}^{iK_0 u(r)} \mathrm{e}^{-iK_0 u(0)} \rangle \,. \tag{16}$$

For Gaussian models such as the Larkin model, one had $C(r) = \exp[-K_0^2/2B(r)] \sim \exp[-K_0^2/2r^{2\zeta}]$ indicating a stretched exponential decay of positional order. Such an exponential decay of positional order would lead to non-divergent peaks in the structure factor. The constant finding of algebraic roughening, and the existence of the length scale L_a seemed to suggest that even beyond L_a the roughening is also algebraic, with perhaps another exponent. One would thus naively expect $C(r) \simeq \exp[-K_0^2/2a^2(r/L_a)^{2\zeta}]$ giving Lorentzian like non-divergent Bragg peaks with a width controlled by $1/L_a$. In addition to this exponential loss of positional order occurring even in the elastic theory, one could question the very starting point of the analysis, namely, the elastic limit and the single-valuedness of the displacements. One could indeed expect topological defects (dislocations, disclinations etc.) to be generated at the scale L_a where the displacements are of the order of the lattice spacing a. Indeed there are arguments [43] (incorrect, as we will see) "showing" that *an arbitrarily small disorder* would always generate topological defects at length scale L_a, leading definitely to an exponential loss of positional order beyond this length. All these elements thus seemed to click together to suggest the picture of a crystal broken into little crystallites of size L_a as shown in Fig. 8. A consensus was thus reached in the community that disordered periodic systems would just lose translational order, and some theories for the vortex lattices were built on this incorrect premise. However this picture crumbled on two fronts. On the experimental side, it was in direct contradiction with experiments showing, for example, extremely large regions free of defects [44,45] or a first order melting [46], which was hardly compatible with a very disordered solid in which all positional order would have been lost from the start. On the theory side, our understanding of glassy systems reached a point where better solutions of this problem can be found. The displacements were found to grow in fact only logarithmically [42,47,48,49] with distance $B(r) = A'_d \log(r)$ (or $u(L) \sim \log(L)^{1/2}$). The prefactor A'_d was computed using either a variational approach [42,48,49] or an FRG one [42,49]. Elastic disordered systems have, therefore, a completely different

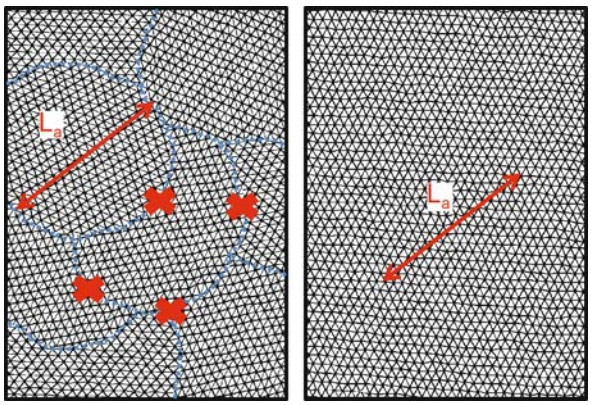

Disordered Elastic Media, Figure 8
Left: **the (incorrect) image of an elastic medium in presence of disorder. The system would be broken into "crystallites" of size L_a characteristic size for which the displacements become of the order of the lattice spacing a ($u(L_a) \sim a$). At the same length scale to release part of the elastic energy due to disorder, the system would prefer to create topological defects (schematically indicated by the *red crosses*). Beyond the size L_a indicated by the *blue dashed line*, positional order would be lost exponentially quickly. *Right:* the Bragg glass, describing the properties of a disordered periodic system in the presence of weak disorder. Although positional order is destroyed at large length scale, and the length scale L_a for which displacements are of order a exists, the system preserves quasi-long range positional order, and perfect topological order. No "crystallite" is thus associated with the length scale L_a, and no topological defects are generated by the disorder**

roughness than interface systems. This a priori surprising behavior can be explained in a qualitative way as shown in Fig. 9. Quite interestingly the structure factor and positional order could still be computed for the full model [42,49], and it was shown that in a quite nontrivial way, the relation $C(r) = \exp[-K_0^2/2B(r)]$ remains essentially applicable, leading to a *power-law* decay of positional order $C(r) \propto (1/r)^\eta$, where the exponent η, is for all practical purposes, a number determined only by the dimension [50]. For example $\eta = 1 - 1.2$ for a three dimensional vortex lattice. The algebraic decay of positional order as well as the value of the exponent indicated that the system still retained *divergent* Bragg peaks in its structure factor and thus, although indeed losing positional order, the loss was very slow and the system was nearly as ordered as a perfect solid. Furthermore, it has been shown [42] that the argument claiming that disorder would always generate dislocations was incorrect and that, on the contrary, due to the slow algebraic decay of the positional order, a three dimensional system is *stable* to the generation of dislocations, at least when the disorder is below a certain threshold. This has led to a physics for a periodic dis-

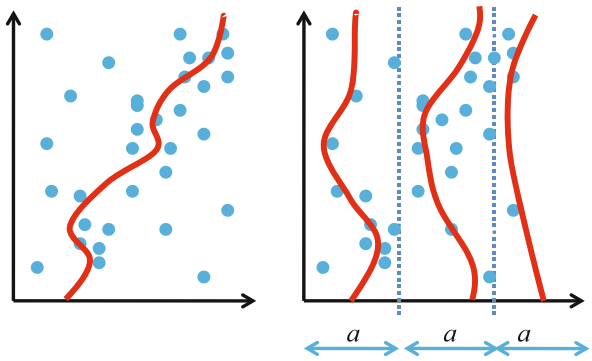

Disordered Elastic Media, Figure 9
Schematic explanation of the difference of roughness between interfaces and periodic systems. *Left:* In an interface, a high degree of roughness is produced by the fact that there are always regions where it is energetically favorable for the line to go, and from there further to another region, endlessly increasing the displacements. *Right:* For a periodic system (here a periodic system of lines of period *a*), since what counts is the *total* energy of the system, there is no interest for one line to make displacements much larger than the interline distance, since it would just steal disorder from the neighbor. Thus, even with the same disorder, displacements would "saturate" (in fact still grow but very slowly with distance) when they reach the interparticle distance

ordered system radically different from the generally accepted consensus. In particular, consider the two following facts: (a) algebraic (quasi-long range) decay of the positional order, and divergent Bragg peaks; (b) absence of topological defect has led to the prediction [42] that the disordered periodic system are in fact in a new state of matter, the Bragg glass. Such a system is a disordered system with glassy properties: an energy landscape with many metastable states and the dynamics of a glass, but which "looks" nearly as ordered as a perfect solid. After the Bragg glass was first predicted, its existence has been supported by further analytical [51,52,53] and numerical [54,55] calculations.

The existence of the Bragg glass phase has important consequences and makes it possible to reconcile several apparently contradictory results on the phase diagram of the vortices. It allows us to explain that very large regions free of dislocations can be observed [44,45] while the system is obviously pinned by the disorder. It also explains [56] the narrow peaks observed in neutron scattering experiments [57,58], with a width given by the experimental resolution, which were indicating an excellent degree of positional order. The power-law nature of the peaks has been directly tested in neutron scattering experiments, proving directly the existence of the Bragg glass phase [59]. In addition to its intrinsic properties, the presence of the Bragg glass phase puts strong constraints on the phase diagram. Indeed, since it is a phase without free topological defects, this phase has to "melt" either when the temperature becomes too high or the disorder too strong, since topological defects have to appear. In vortex systems the latter can be done by changing the magnetic field. The Bragg glass thus provides a very natural explanation [42,60,61] for the existence of a "melting" phase transition as a function of the magnetic field [62,63], such a transition being associated with the destruction of the Bragg glass phase [64,65]. An example is shown in Fig. 10.

This section can cover only a fraction of the physics of periodic systems, and several other questions have been explored. I refer the reader to the above mentioned literature on the subject for more details.

Pinning and Dynamics

Let us now turn to dynamic properties. One of the main interests of such systems is the fact that their dynamics can easily be probed. Indeed most of these systems can be set in motion by an external force acting directly on the interface or on the crystal, and the velocity v versus force F characteristics are directly measurable. As mentioned before, this is of special importance since these characteristics are linked to paramount properties of the systems (voltage-current for vortices, current-voltage for CDW and Wigner crystals, velocity-applied magnetic field for magnetic domain walls). In addition to this practical importance, the dynamics will reflect, even in a more dramatic way than the statics, the competition between disorder and elasticity. In particular, one can expect the dynamics to be dramatically sensitive to the glassy properties and the energy landscape.

The main issues relating to the application of an external force are shown in Fig. 11. In the presence of disorder it is natural to expect that, at zero temperature, the system remains pinned and polarizes only under the action of a small applied force, i. e., it moves until it locks on a local minimum of the tilted energy landscape. At a larger drive, the system follows the force F and acquires a nonzero asymptotic velocity v. So a first set of questions is prompted by the zero temperature properties: what is F_c and how can it be computed? In addition, the $v - F$ curve at $T = 0$ is reminiscent of the curve of an order parameter in a second order phase transition. Here, the system being out of equilibrium, no direct analogy is possible, but this suggests that one could expect $v \sim (F - F_c)^\beta$ with a dynamical critical exponent β. Whether such an analogy to critical phenomena holds and what the physical consequences and calculation of such exponents might be are, of course, important questions.

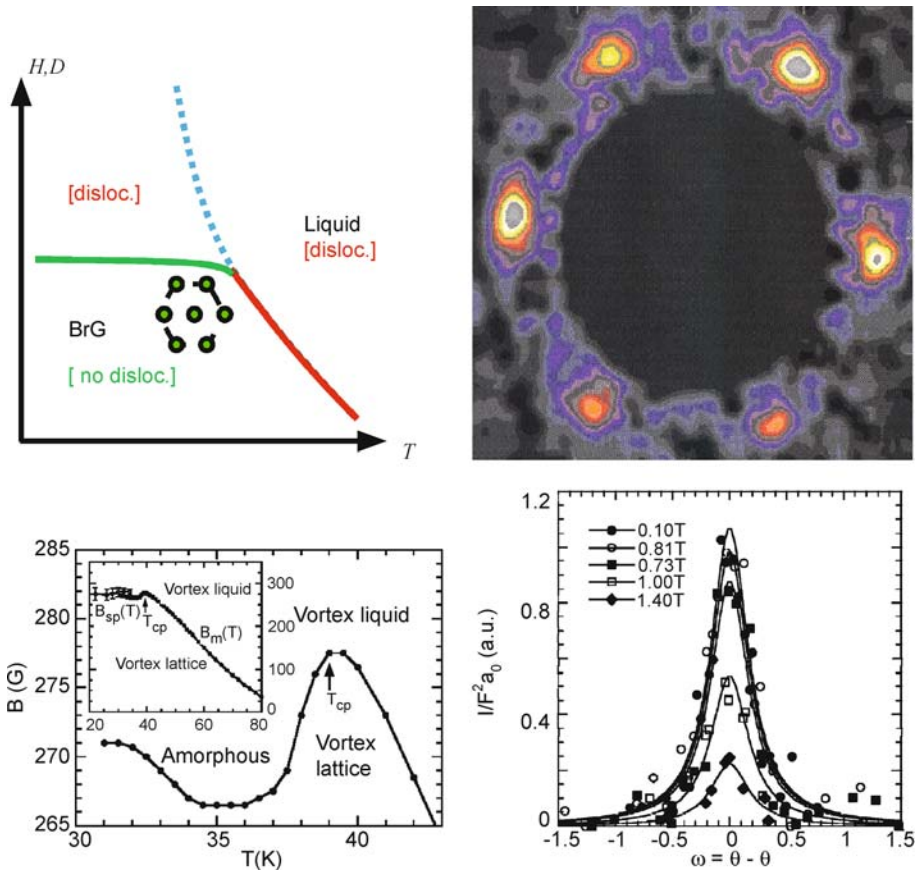

Disordered Elastic Media, Figure 10

Left top: **Schematic theoretical phase diagram for vortices as a function of the temperature *T* and the magnetic field *H*, or the disorder *D*. The Bragg glass (BrG) that has perfect topological order can melt either due to thermal fluctuations (*red line*) or because the disorder becomes too large (*green line*). The existence of the Bragg glass thus implies a single melting curve having a crossover between these two regimes. The melting of the Bragg glass thus explain the existence of a transition as a function of the magnetic field. The *blue dashed line* would be the melting line of the solid in the absence of disorder (After [61]). *Left bottom:* Measured phase diagram for high temperature superconductor BSCCO, showing that both melting transitions with temperature and with magnetic field are indeed the same melting curve. [From [65] (Copyright 2001 by the Nature group)]. *Right:* Neutron diffraction on the superconductor BKBO. The structure factor shows clear Bragg peaks, indicating the good degree of both positional and orientational order despite the disorder present in the sample. As shown in the *bottom diagram* the width of the structure factor does not change when the magnetic field is changed, since it is controlled by the experimental resolution, while the height decreases. This is in agreement with the consequences of a powerlaw divergent Bragg peak, and is thus a test of the existence of the Bragg glass. [From [59] (Copyright 2001 by the Nature group)]**

Another important set of questions pertains to the nature of the moving phase itself, and in particular to behavior at large velocity: to what extent does this moving system resemble the static one? This concerns both the positional order properties and the fluctuations in velocity such as the ones measured in noise experiments. Can we expect novel physics there, or is the system simply "surfing" over the disorder?

Finally, how does the system respond to a very small applied force? We are accustomed to the fact that a normal system when perturbed usually responds linearly to the perturbation. We could thus naively expect $v \propto F$, with a coefficient that would define the "mobility" of the interface. Is this true, or, due to the glassy nature of the system, do we have nonlinear response and more complicated physics?

General Description of the Dynamics

Computing the dynamics is not an easy task. Let me illustrate the method using the case of the interfaces. The displacement field in each point is now a function of the

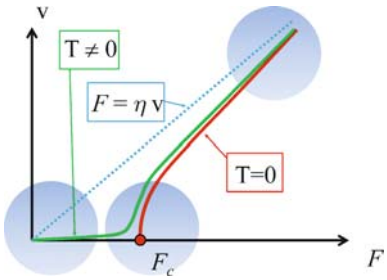

Disordered Elastic Media, Figure 11
The velocity v induced by an external force F of a disordered elastic system. In the absence of pinning and with a damping coefficient η the steady state velocity $v = F/\eta$ is reached. At zero temperature $T = 0$ the system stays pinned until a critical force F_c is reached. At finite temperature a motion can occur even for forces below the threshold $F < F_c$, since the barriers to motion can always be passed by thermal activation. One can distinguish three very different regimes in this curve: large velocity, depinning and small force response (creep)

time $u(r, t)$ and has to obey the equation of motion. The starting point is the equation of motion

$$m\frac{\mathrm{d}^2u(r,t)}{\mathrm{d}t^2} + \eta\frac{\mathrm{d}u(r,t)}{\mathrm{d}t} = \sum F\,, \tag{17}$$

where η is a friction coefficient that phenomenologically describes the dissipation processes that take place inside the object (interface, etc.) when there is motion. Usually one is interested in the steady state motion of the system, in which case in the long time limit, the second order derivative becomes smaller than the first order one and a good approximation is to take $m = 0$ in the above equation.

The forces are of two types. There are the forces deriving from a Hamiltonian

$$F[u(r,t)] = -\frac{\partial H[u]}{\partial u(r,t)}\,. \tag{18}$$

The two main contributions (elastic and disorder) lead in the equation of motion to the elastic forces trying to keep the interface flat, and to the pinning forces. In addition to these forces that would be present in equilibrium, one must add two other forces: The first one is the external force. I consider here only the simple case of a constant external force F. In the presence of this force and in the absence of pinning it is natural to expect the system to reach a steady state velocity $v = F/\eta$. Note that such a state, although time-independent, cannot be described by an equilibrium theory. In particular, the fluctuation dissipation theorem, relating in equilibrium the fluctuations in the absence of a perturbation and the response of the system to an external perturbation, is not in general obeyed any more. The second force is needed if we want to describe the system at a finite temperature. In that case, one must add [66] a Langevin force $\zeta(z, t)$ which is a noise with correlations

$$\langle\zeta(r,t)\zeta(r',t')\rangle = \eta T\delta(r-r')\delta(t-t')\,. \tag{19}$$

The equation of motion thus becomes, in its simplest incarnation,

$$\eta\frac{\mathrm{d}u(r,t)}{\mathrm{d}t} = -\frac{\partial H[u]}{\partial u(r,t)} + F + \zeta(r,t)\,. \tag{20}$$

As is well known, the presence of Langevin noise ensures that, in the absence of an external force F, the time evolution of the system reproduces the thermodynamic ensemble average. In other words, the equal time correlations $\langle u(z,t)u(z',t)\rangle$ obtained by averaging over the thermal noises, are identical to the equilibrium correlation function $\langle u(z)u(z')\rangle_H$ that one would have obtained for a system with the Hamiltonian H at the temperature T.

In the absence of disorder the equation becomes quite simple and is known as the Edwards–Wilkinson equation

$$\eta\frac{\mathrm{d}u(r,t)}{\mathrm{d}t} = -\nabla_r^2 u(r,t) + F + \zeta(r,t)\,. \tag{21}$$

The system thus slides at a constant velocity $v = F/\eta$ and one can see that in the moving frame the interface is at equilibrium, since the change of variable $u(r,t) = \frac{F}{\eta}t + \delta u(r,t)$ gives for the relative displacements $\delta u(r,t)$ exactly the same equation as in equilibrium in the absence of any external force. Even in this simple case, there are several effects of the motion that need to be taken into account, in particular the presence of a cutoff in the system generates terms that would not normally have been incorporated in the original equation of motion and that can modify the behavior of the system. The most well known is the so-called Kardar–Parisi–Zhang (KPZ) term [67].

In the presence of disorder, the equation of motion becomes extremely complicated to solve since the pinning force is a random variable depending on the particular realization of the disorder, and a double averaging must be done, both on the thermal noise and on the disorder. No perfect method exists to treat such an equation. Since we are usually better equipped to deal with integrals than with differential equations, especially with stochastic terms, a convenient rewriting of this equation exists, which formally gives back the equivalent of a path integral and an action. This is the so-called Martin–Siggia–Rose (MSR) formalism [68,69]. I refer the reader to the literature on this relatively specialized method [66]. The advantage is that it allows averaging over the disorder from the start. This in particular paves the way for an FRG treatment of the problem.

Depinning

The first set of questions arises around depinning. Indeed, in the presence of disorder the naive expectation is that the interface is unable to move, at zero temperature, below a certain threshold of force F_c called the pinning force. Computing this pinning force is not easy. In a remarkable feat of physical intuition, Larkin has shown that the pinning force can be directly obtained from the static behavior of the system [70]. Indeed, the idea is that the pinning force is related to the appearance of many metastable states and the presence of random potential. Because it is quadratic in the displacements, the Larkin model does not exhibit a pinning force. The idea would thus be to relate the pinning force to the length scale at which the Larkin model ceases to pertain. As we discussed before, this is the length L_c for which the displacements are of the order of the correlation length of the random potential or the width of the elastic object. At that scale, the elastic plus disorder energy scales as $cL_c^{d-2}r_f^2$ while the additional energy due to the force scales as

$$H_F = \int \mathrm{d}^d r F u(r) \sim F L_c^d r_f \,. \tag{22}$$

Balancing the two terms leads to the famous Larkin collective pinning force

$$F_c = \frac{c r_f}{L_c^2} \,. \tag{23}$$

This is a remarkable relation since it relates a dynamic property to purely static quantities. This intuitive result can be substantiated by considerably more complicated calculations. In the next section we will see another rough estimate based on a large velocity expansion. Finally, starting from the equation of motion (17), it is possible to obtain F_c from an FRG calculation [71,72], confirming, from this microscopic calculation, Larkin's result. Numerical methods have allowed an extremely precise calculation of F_c [39].

Besides the existence of the pinning force itself, the description of depinning is a considerable challenge. A very fruitful line of approach for this problem was suggested by D.S. Fisher [73]. Indeed, looking at the $v - F$ characteristics is strongly reminiscent of the curve of an order parameter as a function of temperature in a second order phase transition (zero for $T > T_c$ and non-zero for $T < T_c$). This strongly suggests using an analogy with a standard critical phenomenon to analyze the depinning. In particular, one can infer from this analogy that a divergent length scale exists at the transition, and that one can define scaling behavior and critical exponents as a function of this length scale. One can define a critical exponent for the velocity $v \sim (F - F_c)^\beta$, for the correlation length $\xi \sim (F - F_c)^{-\nu}$ and a dynamical exponent relating space and time divergences $\tau \sim \xi^z$. These exponents are related by scaling relations, analogous to those of standard critical phenomena and that can be computed by looking at the scaling of the equation of motion. The scaling relations are

$$\nu = \frac{\beta}{2-\zeta} = \frac{1}{z-\zeta} \,. \tag{24}$$

Such scaling behavior is directly confirmed from solutions of the equation of motion, either from FRG or from numerical simulations. The length scale ξ can be identified as the length scale of avalanches. Computing and measuring these exponents is a considerable challenge and sophisticated FRG [35,71,72,74] or numerical [39,75] techniques have been developed for this goal.

In addition to these quantities characterizing the motion of the line, other important physical observables are modified by the application of external force. This is in particular the case of the roughness of the line. Right at depinning $F = F_c$, the line is much more rough than when in equilibrium, since it is on the verge of depinning. There is thus a new roughness exponent ζ_{dep} which can be computed and is $\zeta_{\mathrm{dep}} \sim 1.2$ for a one dimensional interface. This result has two very important consequences. The first one comes from the value of the roughness exponent itself. Since, at least for a line, this value is larger than one, this immediately suggests that close to depinning the elastic model will run into trouble. Indeed when u scales more than linearly with distance, the vary basis of the elastic approximation $\nabla u \ll 1$ is violated at large length scales. The line will thus have to generate defects (overhangs) to heal this fact. The nature of the resulting physics when this is taken into account is a challenging and yet open question. The second observation concerns the steady state aspect of the line. At large length scales, because of finite velocity, the system will average over the disorder. We will come back in more detail to this point in the next section, but for the moment, stick with this simple vision. In this case, beyond the length ξ one can expect the disorder to be irrelevant and thus to recover the pure thermal roughness exponent. The system will thus have the depinning roughness exponent ζ_{dep} for length scales below ξ and the thermal one ζ_{th} for length scales above ξ. This is summarized in Fig. 12. One important question now is what happens at a small but finite temperature. The first effect is, of course, to smooth the $v - F$ characteristics. This leads to the important question of whether one can define a scaling with the temperature of this thermal rounding of the depinning (see e. g., [77] and references therein). Even more interest-

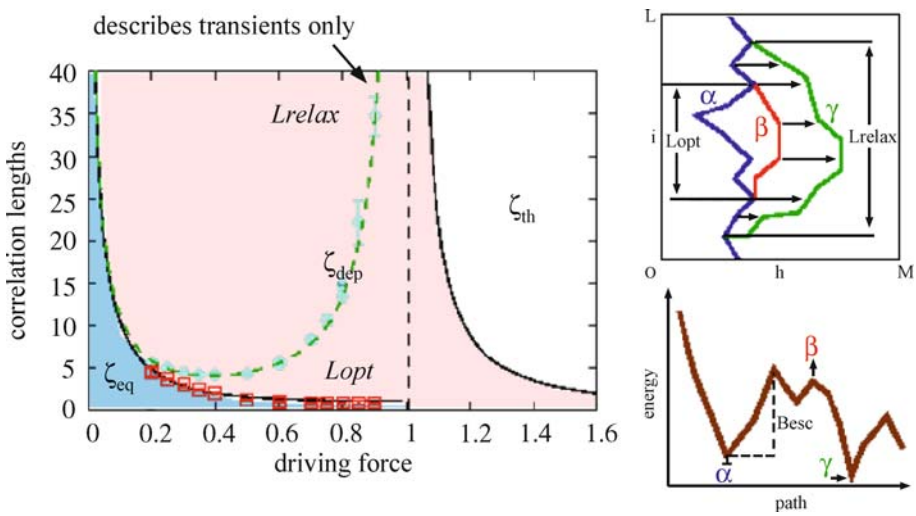

Disordered Elastic Media, Figure 12

Close to depinning the motion proceeds by avalanches between two configurations α and γ. Above F_c, there exists a divergent length scale (L_{opt} on the figure) below which the line is characterized by the roughness exponent ζ_{dep} and above which the line shows the thermal roughness ζ_{th}. A normal critical phenomenon would have had a similar divergent length scale for $F < F_c$. This is not the case for the depinning. A transient divergent length scale L_{relax} does exist, but does not show up in the steady state properties of the line. Contrary to naive expectations from a "standard" critical phenomenon, one observes the equilibrium roughness exponent ζ_{eq} at *short* distances. This shows that the analogy between depinning and "standard" critical phenomena, although very fruitful, must be taken with a grain of salt. On the right, the schematic shape of the line and energy profiles are shown. After [76]

ing is the question of the roughness of the line. The analogy with a critical phenomenon would simply suggest that a similar divergent length scale should exist for $F < F_c$, leading to the standard pattern of "critical regime". However, as indicated in Fig. 12, such a divergent length scale does not exist [76]. This leads to a very puzzling behavior and shows that although the analogy with standard critical phenomena can be extremely fruitful, it must be taken with a grain of salt. The depinning, which is by essence a dynamical transition, possesses its own physics.

High Velocity Phase

Another important set of questions and physics occurs when the interface is moving at large velocity, i. e., for $F \gg F_c$. This is apparently the simplest regime since one could expect that at large velocity one has a control parameter on the theory and that an expansion in $1/v$ is possible. This is indeed the case for the $v - F$ characteristics. A large velocity expansion of the disorder term can be made by going into the moving frame of the elastic media. One can indeed write $u(r, t) = vt + \delta u(r, t)$, where $\delta u(r, t)$ describes the displacements in the moving frame. Because the system surfs over the disorder at very large velocity one can expect the effects of the disorder, and hence $\delta u(r, t)$, to be small. This is confirmed by a well controlled large velocity expansion [78,79]. In particular, the correction due to the disorder to the velocity can be computed and behaves as

$$\frac{F - \eta v}{\eta v} \propto D \left(\frac{1}{\eta v} \right)^{\frac{4-d}{2}} . \tag{25}$$

This shows clearly that the effects of disorder are kept small at large velocity or large force and become increasingly important as the force/velocity gets smaller, in agreement with Fig. 11. The relative correction to the velocity grows in dimensions smaller than $d = 4$, confirming that disorder is relevant below this dimension. Although one cannot extrapolate the perturbative expressions, a crude way to estimate the critical force F_c is when the deviations (25) become of order one. This method gives back the estimate (23) for F_c, obtained from totally different considerations.

The large velocity expansion also allows us to address the physics of the moving phase, namely the shape and properties of the elastic system in the moving frame. A calculation of these effects was performed [80] by computing the displacements δu from the large velocity expansion. This leads to the striking result that at large velocity the effect of disorder disappears and can be absorbed in a simple modification of the temperature of the system, leading to an effective temperature T_{eff}. This has important consequences on the properties of the system in the

moving frame. For an interface, this is consistent with the idea, exposed in the previous section, that at large distance one recovers the thermal roughening exponent. For periodic systems, since there is the possibility of melting, this has the more drastic consequences that driving the system could induce a velocity controlled melting. Indeed, for large velocities, the effective temperature is small, while for smaller velocities it would be large. The system would thus be a disordered system while static, then close to depinning, where the effective temperature would be large, it would be in a melted (liquid) state, and would then recrystallize when moving at larger velocities.

Although the concept of effective temperature is extremely fruitful, in particular for this dynamic recrystallization, the properties of the moving periodic system are richer [81] than those of a simple solid subjected to temperature. Indeed, periodic systems have a structure in the direction *transverse* to the direction of motion. This structure, and the corresponding disorder structure, cannot be averaged by the motion, however large the velocity remains. This leads to the fact that disorder remains even when the system is in motion. In other words, a moving periodic system remains a glass. The way the motion takes place is quite peculiar. The system finds optimal paths [81] which are a compromise between the elastic energy, disorder and the motion. These paths are rough, similar to the way in which a static system in a disordered environment is rough. This is shown in Fig. 13. For periodic systems, pinning effects still manifest themselves in the moving system. The glassy nature and the channel motion has been confirmed both numerically [82,83,84] and experimentally [85]. The channel motion leads to an interesting consequence. Since the effects of disorder are weakening as velocity increases, the channels undergo a transition between a regime for which the particles in different channels are coupled or decoupled [86,87]. In the first case, the system is essentially moving as a "solid" (in fact a moving Bragg glass) since the topological order is perfect even if the system is still distorted by disorder. In the second case, the channels are decoupled, which means that a smectic like structure of the particles inside the channels is expected. These transitions as described in Fig. 13 have also been observed both numerically and experimentally as shown in Fig. 14. An additional consequence of the existence of such channels is the existence of a *transverse* pinning force [81,87]. Indeed, even if the particles themselves are moving along the channels, the channels themselves are pinned if one applies an additional force transverse to the direction of motion. This surprising phenomenon has been numerically confirmed [82,83,84], but observing it in classical periodic systems is still an experimental challenge. Experiments showing the absence of Hall effect in Wigner crystal systems [88] could constitute an experimental proof of such a transverse critical force, but clearly further experimental data would be needed to unambiguously decide on that point.

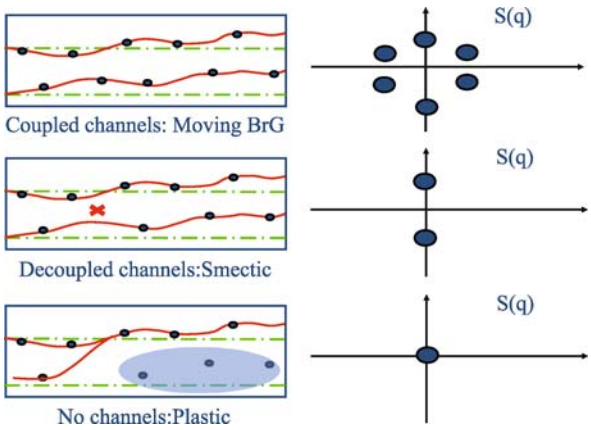

Disordered Elastic Media, Figure 13
Motion of a periodic system. The system develops rough channels which compromise between elastic energy and the transverse component of the disorder that are poorly averaged by the motion. All the particles follow on these channels like cars on highways. Depending on the velocity, the channels are increasingly coupled: *Bottom image:* close to depinning, motion can proceed through a plastic regime where unpinned and pinned regions (denoted by the *blue circle*) coexist; *Middle image:* topological defects can exist between the channels, so although the channels themselves are well formed, the particles in them are essentially decoupled leading to a smectic like behavior; *Top image:* the channels are coupled and the system is a moving Bragg glass with effects of both disorder and elasticity and no topological defects. On the right, the corresponding structure factors are indicated

Small Applied Force and Creep Motion

Finally, let us look at the response of the system to a small external force. At zero temperature, one is below the pinning force, and thus except for a transient motion the system remains pinned. Motion can thus only take place due to thermal activation over the energy barriers. The response to a small external force is thus a method of choice to probe for the nature of the energy landscape of such systems. For usual systems one expects the response to be linear. Indeed, earlier theories of such motion have found a linear response. The idea is to consider that a blob of pinned material has to move in an energy landscape with characteristic barriers Δ as shown in Fig. 15. The external force F tilts the energy landscape, thus making forward motion possible. The barriers are overcome by thermal activation [89] (hence the name: Thermally Assisted

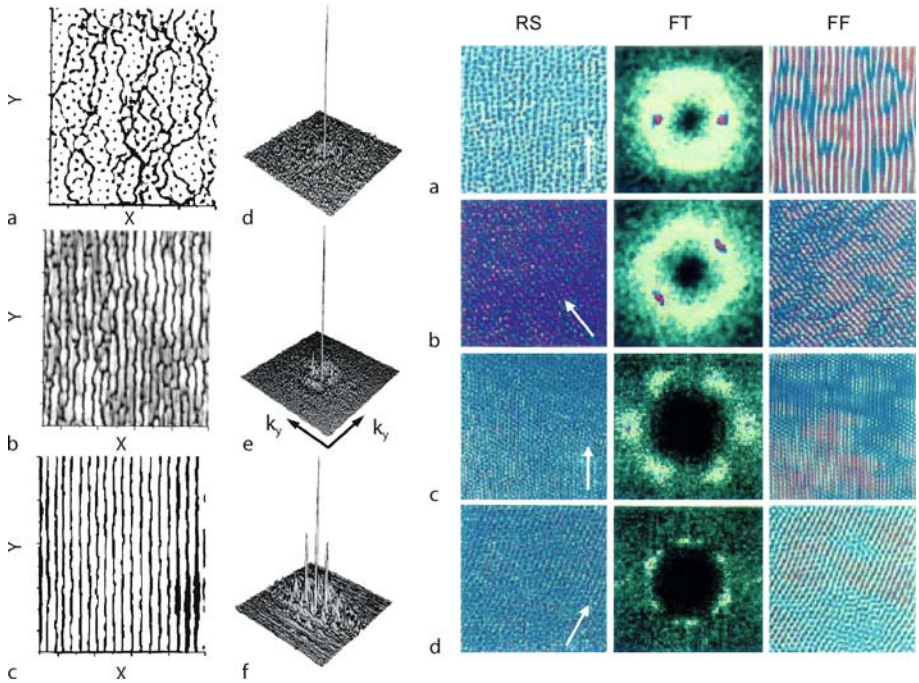

Disordered Elastic Media, Figure 14

Left: **numerical simulations confirming the presence of channels for moving periodic structures and the sequence of transitions depicted in Fig. 13. The *left part* of the image shows the real space trajectories, while the *right part* is the structure factor. The force is increasing from the *top* to the *bottom* of the image. [From [83] (Copyright 1999 by the American Physical Society)]. *Right:* A decoration image of moving vortices also showing the presence of channels. The direction of the applied force is indicated by an *arrow*. The *left column* is the raw decoration image, the *center one* is the Fourier transform giving the structure factor, the *right column* is a filtered version of the image showing the channels more clearly. [From [85](Copyright 1999 by the Nature group)]**

Flux Flow (TAFF)) with an Arrhenius law. If the minima are separated by a distance a the velocity is

$$v \propto e^{-\beta(\Delta - Fa/2)} - e^{-\beta(\Delta + Fa/2)} \simeq e^{-\beta\Delta} F \,. \qquad (26)$$

The response is thus linear, but exponentially small.

However this argument is grossly inadequate for a glassy system. The reason is easy to understand if one remembers that the static system is in a glassy state. In such a state a characteristic barrier Δ does not exist, since barriers are expected to diverge as one gets closer to the ground state of the system. The TAFF formula is thus valid in systems where the glassy aspect is somehow eliminated and the barriers saturate. This could be the case, for example, for a finite size interface. When the glassy nature of the system persists up to arbitrarily large length scales, the theory should be accommodated to take into account divergent barriers. This can be done qualitatively within the framework of the elastic description using scaling arguments [47,90,91,92]. The basic idea rests on two quite strong but reasonable assumptions: (i) the motion is so slow that one can consider, at each stage, that the interface is motionless and use its *static* description; (ii) the scaling for barriers, which is quite difficult to determine, is the same as the scaling of the minimum of energy (metastable states) that can be extracted, again, from the static calculation. If the displacements scale as $u \sim L^{\zeta}$ then the en-

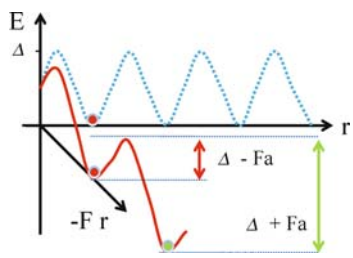

Disordered Elastic Media, Figure 15

In thermally assisted flux flow [89] a region of pinned material is considered as a particle moving in an energy landscape characterized by characteristic barriers Δ, schematized by the *blue dashed periodic potential*, of period a. Applying an external force tilts the energy landscape. The motion over barriers can always proceed by thermal activation. Due to tilt, the barrier to forward motion (*in red*) is smaller than the reverse barrier (*in green*). This results in an exponentially small but linear response when a small external force is applied to the system

ergy of the metastable states (see (2)) scales as given by (5): $E(L) \sim L^{d-2+2\zeta}$. Since the motion is very slow, the effect of the external force is simply to tilt the energy landscape

$$E(L) - F \int \mathrm{d}^d r u(r) \sim L^{d-2+2\zeta} - F L^{d+\zeta} \,. \tag{27}$$

Thus, in order to make the motion to the next metastable state, one needs to move a piece of the pinned system of size

$$L_{\mathrm{opt}} \sim \left(\frac{1}{F}\right)^{\frac{1}{2-\zeta}} \,. \tag{28}$$

The size of the optimal nucleus able to move thus grows as the force decreases. Since the barriers to overcome grow with the size of the object, the minimum barrier to overcome (*assuming* that the scaling of the barriers is *also* given by (5))

$$U_b(F) \sim \left(\frac{1}{F}\right)^{\frac{d-2+2\zeta}{2-\zeta}} \tag{29}$$

leading to the creep formula for the velocity

$$v \propto \exp\left[-\beta U_c \left(\frac{F_c}{F}\right)^{\mu}\right] , \tag{30}$$

where F_c is the depinning force and U_c a characteristic energy scale and the *creep exponent* μ is given by,

$$\mu = \frac{d-2+2\zeta}{2-\zeta} \,. \tag{31}$$

Equations (30) and (31) are quite remarkable. They relate a dynamical property to *static* exponents, and show clearly the glassy nature of the system. The corresponding motion has been called creep since it is a sub-linear response. It is a direct consequence of the divergent barriers in the pinned system.

Of course the derivation above is phenomenological, so it is important to ascertain by more microscopic methods whether the results hold. Although in principle one simply has to solve the equation of motion (20), in practice this is quite complicated. A natural framework for computing perturbation theory in off-equilibrium systems is the MSR formalism. Using this formalism and an FRG analysis, one can confirm the creep formula for velocity [72,93]. Numerical simulations also show the absence of linear response and the existence of a creep response [94]. Creep has also been checked experimentally in various systems. Vortices show a creep behavior with an exponent $\mu = 1/2$ compatible with the existence of the Bragg glass ($d = 3$, $\zeta = 0$) [95]. However, the range of measurable velocities makes it difficult to unambiguously check for this law. One spectacular determination of the creep law was performed in a magnetic film [1]. In such a situation the roughness exponent is known exactly ($\zeta = 2/3$) and thus the value of the creep exponent $\mu = 1/4$ is not an adjustable parameter, making it a much more stringent test. As shown in Fig. 16, the velocity was measured over ten orders of magnitude, a spectacular feat, and a remarkable confirmation of the creep law. Measurements of the creep law have also been performed for domain walls in ferroelectrics where a simultaneous measurement of the creep exponent and of the roughness exponent was performed [7,8]. As shown in Fig. 16 the stretched exponential behavior for the velocity is well verified, and the formula (31) consistent with what is expected for a two dimensional domain wall in the presence of dipolar forces, corresponding to the experimental situation.

The FRG derivation allows us to more deeply probe the physical understanding of the system. In particular it unravels a new phenomenon. Although the velocity itself is dominated by events occurring at the thermal length scale (28), interesting physics takes place beyond this length scale. Indeed, when a portion L_{opt} of the line has been able to move forward by thermal activation over the barriers, it serves as a nucleation center to trigger an avalanche [72] over a much larger length scale L_{av}. The behavior between these two length scales is thus very similar to a depinning phenomenon where temperature plays no role. Although the velocity is dominated by the first process, which is the slow one, the shape of the line reflects this much larger avalanche scale in a way which is compatible with experiments [3]. The creep, being controlled by the time to overcome divergent barriers in the system, has several other consequences, in particular on the issue of the out-of-equilibrium physics of such systems [96] and its connection to the aging of glasses [97].

Future Directions

Both because of experimental drive (no pun intended) but also because of theoretical advances and the development of the proper tools, this field has known several breakthroughs in the last decade or so. There is now a good understanding of static properties for both interfaces and for periodic systems, and most of the misconceptions and folklore have been replaced by solid results. Novel phases have emerged such as the Bragg glass phase. Steady state dynamics has also made significant progress, with the understanding of processes such as creep motion. From the point of view of methods, these systems have made it possible to perfect methods to deal with glassy systems such as replica methods, functional renormalization group as well

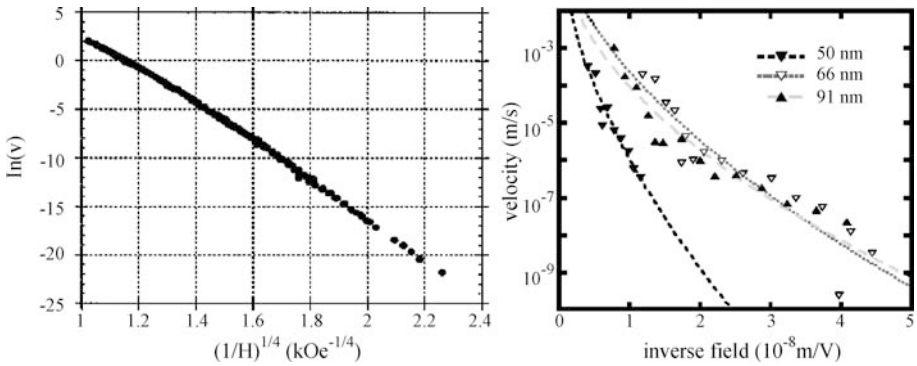

Disordered Elastic Media, Figure 16

Experimental verification of the creep law for magnetic and ferroelectric domain walls. *Left:* Magnetic domain walls. The film is extremely thin, thus the domain is a line in a two dimensional plane, leading to a creep exponent of $\mu = 1/4$. The creep law is observed on about ten orders of magnitude for the velocity. [From [1] (Copyright 1998 by the American Physical Society)]. *Right:* Ferroelectric films. A creep behavior is observed over several orders of magnitude giving a creep exponent of $\mu \sim 0.58$. This value together with the measured roughness exponent $\zeta = 0.26$ leads to an effective dimension of $d = 2.5$, well in agreement with a two dimensional domain wall in presence of dipolar forces. [From [8] (Copyright 2005 by the American Physical Society)]

as special numerical methods. These results have found and continue to find applications in a large variety of experimental domains. Despite these advances, it is clear that many questions remain pending, making it still a very challenging field which is yet in constant progress. Experiments regularly provide new systems and new challenges and stimulate theoretical analysis. Several lines of research are actually open and should carry the bulk of the research in that domain in the future.

From the point of view of the static, although the situation without defects is under control, we know next to nothing when elasticity, disorder and defects are included. For interfaces this means treating the overhangs and bubbles, and for periodic systems all the topological defects such as the dislocations. Although we know now that the situation without defects is robust below a certain threshold of disorder, it is clear that the ability to deal with the strong disorder situation is needed for many experimental systems. This is the case for the high field phase of vortices and strong disorder in the interfaces, both of which are dominated by defects.

In the dynamics, one of the very challenging questions that one has to face is the one of the out-of-equilibrium dynamics, when the system has not yet reached a steady state. A simple example of such a situation would be an interface relaxing slowly from a flat configuration or quenched at low temperatures from a high temperature configuration. Given that these systems are glasses, the time evolution of such cases is highly non-trivial, and should show a generic phenomenon of glasses known as aging. This is directly a situation relevant to many experiments. From the conceptual point of view this is an extremely challenging question since most of the theoretical tools that we have fail to tackle such situations, and thus new tools or new concepts need to be invented.

Last but not least, we have dealt mainly with classical systems here. But disordered elastic systems can also be realized in the quantum worlds as briefly mentioned in the introduction. The question on how to extend the concepts of glasses to quantum systems remains largely open. In particular, one can expect the dynamics to be affected. Indeed, classical systems can pass barriers only by thermal activation, while quantum systems are good at tunneling through barriers. The extension of the above concepts to the quantum world is thus a very challenging direction.

The gold mine of disordered elastic media is thus far from being exhausted. It seems that each nugget we find is only the opening of a new vein with even richer stones. The variety of experimental realization is ever growing, as is the depth of the questions that are now within our grasp.

Bibliography

Primary Literature

1. Lemerle S et al (1998) Phys Rev Lett 80:849
2. Krusin-Elbaum L et al (2001) Nature 410:444
3. Repain V et al (2004) Europhys Lett 68:460
4. Metaxas PJ et al (2007) Phys Rev Lett 99:217208
5. Yamanouchi M et al (2006) Phys Rev Lett 96:096601
6. Yamanouchi M et al (2007) Science 317:1726
7. Tybell T, Paruch P, Giamarchi T, Triscone JM (2002) Phys Rev Lett 89:97601
8. Paruch P, Giamarchi T, Triscone JM (2005) Phys Rev Lett 94:197601
9. Wilkinsion D, Willemsen JF (1983) J Phys A 16:3365
10. Moulinet S, Guthmann C, Rolley E (2002) Eur Phys J E 8:437
11. Barabasi A-L, Stanley HE (1995) In: Fractal Concepts in Surface Growth. Cambridge University Press, Cambridge

12. Bouchaud E et al (2002) J Mech Phys Solids 50:1703
13. Alava M, Nukalaz PKVV, Zapperi S (2006) Adv Phys 55:349
14. Blatter G et al (1994) Rev Mod Phys 66:1125
15. Nattermann T, Scheidl S (2000) Adv Phys 49:607
16. Giamarchi T, Bhattacharya S (2002) In: Berthier C et al (eds) High Magnetic Fields: Applications in Condensed Matter Physics and Spectroscopy. Springer, Berlin, p 314. arXiv:cond-mat/0111052
17. Grüner G (1988) Rev Mod Phys 60:1129
18. Nattermann T, Brazovskii S (2004) Adv Phys 53:177
19. Seshadri R, Westervelt RM (1992) Phys Rev B 46:5150
20. Murray CA, Sprenger WO, Wenk R (1990) Phys Rev B 42:688
21. Coupier G, Guthmann C, Noat Y, Saint Jean M (2005) Phys Rev E 71:046105
22. Andrei EY et al (1988) Phys Rev Lett 60:2765
23. Giamarchi T (2004) In: Quantum phenomena in mesoscopic systems. IOS Press, Bologna. arXiv:cond-mat/0403531
24. Eskildsen MR et al (2002) Phys Rev Lett 89:187003
25. Larkin AI (1970) Sov Phys JETP 31:784
26. Mézard M, Parisi G, Virasoro MA (1987) Spin Glass Theory and beyond. World Scientific, Singapore
27. Mézard M, Parisi G (1991) J de Phys I 4:809
28. Fisher DS (1986) Phys Rev Lett 56:1964
29. Nature group (2007) Nature Material, vol. 6. Focus issue on Multiferroics
30. Nattermann T (1983) J Phys C 16:4125
31. Joanny JF, de Gennes PG (1984) J Chem Phys 81:552
32. Gao H, Rice JR (1989) J Appl Mech 56:828
33. Huse DA, Henley CL (1985) Phys Rev Lett 54:2708
34. Kardar M (1985) Phys Rev Lett 55:2923
35. Le Doussal P, Wiese K, Chauve P (2004) Phys Rev E 69:026112
36. Giamarchi T, Kolton AB, Rosso A (2006) In: Miguel MC, Rubi JM (eds) Jamming, Yielding and Irreversible deformation in condensed matter. Springer, Berlin, p 91. arXiv:cond-mat/0503437
37. Yoshino H (1998) Phys Rev Lett 81:1493
38. Rosso A, Krauth W (2002) Phys Rev B 65:12202
39. Rosso A, Krauth W (2002) Phys Rev E 65:025101R
40. Petaja V et al (2006) Phys Rev E 73:94517
41. Nelson DR (1978) Phys Rev B 18:2318
42. Giamarchi T, Le Doussal P (1995) Phys Rev B 52:1242
43. Fisher DS, Fisher MPA, Huse DA (1990) Phys Rev B 43:130
44. Grier DG et al (1991) Phys Rev Lett 66:2270
45. Kim P, Yao Z, Lieber CM (1996) Phys Rev Lett 77:5118
46. Schilling A, Fisher RA, Crabtree GW (1996) Nature 382:791
47. Nattermann T (1990) Phys Rev Lett 64:2454
48. Korshunov SE (1993) Phys Rev B 48:3969
49. Giamarchi T, Le Doussal P (1994) Phys Rev Lett 72:1530
50. For a scalar displacement or isotropic elasticity the exponent is universal. When the full anistropy of the elastic constants is taken into account in the FRG equations [Bogner S, Emig T, Nattermann T (2001) Phys Rev B 63 174501] the possible variation of the exponent with the magnetic field is still less than a percent.
51. Carpentier D, Le Doussal P, Giamarchi T (1996) Europhys. Lett. 35:379
52. Kierfeld J, Nattermann T, Hwa T (1997) Phys Rev B 55:626
53. Fisher DS (1997) Phys Rev Lett 78:1964
54. Gingras MJP, Huse DA (1996) Phys Rev B 53:15193
55. Otterlo AV, Scalettar R, Zimányi G (1998) Phys Rev Lett 81:1497
56. Giamarchi T, Le Doussal P (1995) Phys Rev Lett 75:3372
57. Yaron U et al (1994) Phys Rev Lett 73:2748
58. Ling XS et al (2001) Phys Rev Lett 86:712
59. Klein T et al (2001) Nature 413:404
60. Ertas D, Nelson DR (1996) Phys C 272:79
61. Giamarchi T, Le Doussal P (1997) Phys Rev B 55:6577
62. Khaykovich B et al (1996) Phys Rev Lett 76:2555
63. Deligiannis K et al (1997) Phys Rev Lett 79:2121
64. Paltiel Y, Zeldov E, Myasoedov Y, Rappaport ML (2000) Phys Rev Lett 85:3712
65. Avraham N et al (2001) Nature 411:451
66. Zinn-Justin J (1989) Quantum field theory and Critical Phenomena. Clarendon Press, Oxford
67. Kardar M, Parisi G, Zhang Y (1986) Phys Rev Lett 56:889
68. Janssen HK (1976) Z Phys B 23:377
69. Martin PC, Siggia ED, Rose HA (1973) Phys Rev A 8:423
70. Larkin AI, Ovchinnikov YN (1979) J Low Temp Phys 34:409
71. Nattermann T, Stepanow S, Tang LH, Leschhorn H (1992) J Phys 2:1483
72. Chauve P, Giamarchi T, Le Doussal P (2000) Phys Rev B 62:6241
73. Fisher DS (1985) Phys Rev B 31:1396
74. Narayan O, Fisher D (1993) Phys Rev B 48:7030
75. Duemmer O, Krauth W (2005) arXiv:cond-mat/0501467
76. Kolton AB, Rosso A, Giamarchi T, Krauth W (2006) Phys Rev Lett 97:057001
77. Bustingorry S, Kolton AB, Giamarchi T (2008) Europhys Lett 81:26005
78. Larkin AI, Ovchinnikov YN (1974) Sov Phys JETP 38:854
79. Schmidt A, Hauger W (1973) J Low Temp Phys 11:667
80. Koshelev AE, Vinokur VM (1994) Phys Rev Lett 73:3580
81. Giamarchi T, Le Doussal P (1996) Phys Rev Lett 76:3408
82. Moon K et al (1997) Phys Rev Lett 77:2378
83. Kolton AB, Grønbech-Jensen DDN (1999) Phys Rev Lett 83:3061
84. Fangohr H, Cox SJ, de Groot PAJ (2001) Phys Rev B 64:64505
85. Pardo F et al (1998) Nature 396:348
86. Balents L, Marchetti C, Radzihovsky L (1998) Phys Rev B 57:7705
87. Le Doussal P, Giamarchi T (1998) Phys Rev B 57:11356
88. Perruchot F et al (2000) Physica B 284:1984
89. Anderson PW, Kim YB (1964) Rev Mod Phys 36:39
90. Ioffe LB, Vinokur VM (1987) J Phys C 20:6149
91. Nattermann T (1987) Europhys Lett 4:1241
92. Feigelman M, Geshkenbein VB, Larkin AI, Vinokur V (1989) Phys Rev Lett 63:2303
93. Chauve P, Giamarchi T, Le Doussal P (1998) Europhys Lett 44:110
94. Kolton AB, Rosso A, Giamarchi T (2005) Phys Rev Lett 94:047002
95. Fuchs DT et al (1998) Phys Rev Lett 80:4971
96. Kolton AB, Rosso A, Giamarchi T (2005) Phys Rev Lett 95:180604
97. Cugliandolo LF, Kurchan J, Bouchaud JP, Mezard M (1998) In: Young AP (ed) Spin Glasses and Random fields. World Scientific, Singapore

Books and Reviews

Barabasi A-L, Stanley HE (1995) Fractal Concepts in Surface Growth. Cambridge University Press, Cambridge

Nelson DR (2002) Defects and Geometry in Condensed Matter Physics. Cambridge University Press, Cambridge

Young AP (ed) (1998) Spin Glasses and Random fields. World Scientific, Singapore

Dispersion Phenomena in Partial Differential Equations

PIERO D'ANCONA
Dipartimento di Matematica, Unversità di Roma "La Sapienza", Roma, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
The Mechanism of Dispersion
Strichartz Estimates
The Nonlinear Wave Equation
The Nonlinear Schrödinger Equation
Future Directions
Bibliography

Glossary

Notations Partial derivatives are written as u_t or $\partial_t u$, $\partial^\alpha = \partial_{x_1}^{\alpha_1} \cdot \partial_{x_n}^{\alpha_n}$, the Fourier transform of a function is defined as

$$\mathcal{F} f(\xi) = \widehat{f}(\xi) = \int \mathrm{e}^{-ix\cdot\xi} f(x)\mathrm{d}x$$

and we frequently use the mute constant notation $A \lesssim B$ to mean $A \leq CB$ for some constant C (but only when the precise dependence of C from the other quantities involved is clear from the context).

Evolution equations Partial differential equations describing physical systems which evolve in time. Thus, the variable representing time is distinguished from the others and is usually denoted by t.

Cauchy problem A system of evolution equations, combined with a set of initial conditions at an initial time $t = t_0$. The problem is *well posed* if a solution exists, is unique and depends continuously on the data in suitable norms adapted to the problem.

Blow up In general, the solutions to a nonlinear evolution equation are not defined for all times but they break down after some time has elapsed; usually the L^∞ norm of the solution or some of its derivatives becomes unbounded. This phenomenon is called *blow up* of the solution.

Sobolev spaces we shall use two instances of Sobolev space: the space $W^{k,1}$ with norm

$$\|u\|_{W^{k,1}} = \sum_{|\alpha|\leq k} \|\partial^\alpha u\|_{L^1}$$

and the space H_q^s with norm

$$\|u\|_{H^s} = \left\|(1-\Delta)^{s/2} u\right\|_{L^q} .$$

Recall that this definition does not reduce to the preceding one when $q = 1$. We shall also use the homogeneous space $\dot{H}_q^s$, with norm

$$\|u\|_{\dot{H}^s} = \left\|(-\Delta)^{s/2} u\right\|_{L^q} .$$

Dispersive estimate a pointwise decay estimate of the form $|u(t,x)| \leq Ct^{-\alpha}$, usually for the solution of a partial differential equation.

Definition of the Subject

In a very broad sense, dispersion can be defined as the spreading of a fixed amount of matter, or energy, over a volume which increases with time. This intuitive picture suggests immediately the most prominent feature of dispersive phenomena: as matter spreads, its size, defined in a suitable sense, decreases at a certain rate. This effect should be contrasted with dissipation, which might be defined as an actual loss of energy, transferred to an external system (heat dissipation being the typical example).

This rough idea has been made very precise during the last 30 years for most evolution equations of mathematical physics. For the classical, linear, constant coefficient equations like the wave, Schrödinger, Klein–Gordon and Dirac equations, the decay of solutions can be measured in the L^p norms, and sharp estimates are available. In addition, detailed information on the profile of the solutions can be obtained, producing an accurate description of the evolution. For nonlinear equations, the theory now is able to explain and quantify several complex phenomena such as the splitting of solutions into the sum of a train of solitons, plus a radiation part which disperses and decays as time increases.

Already in the 1960s it was realized that precise information on the decay of free waves could be a powerful tool to investigate the effects of linear perturbations (describing interactions with electromagnetic fields) and even nonlinear perturbations of the equations. One of the first important applications was the discovery that a nonlinear PDE may admit global small solutions, provided the rate of decay is sufficiently strong compared with the concentration effect produced by the nonlinear terms. This situation is very different from the ODE setting, where dispersion is absent.

Today, dispersive and Strichartz estimates represent the backbone of the theory of nonlinear evolution equations. Their applications include local and global existence

results for nonlinear equations, existence of low regularity solutions, scattering, qualitative study of evolution equations on manifolds, and many others.

Introduction

To a large extent, mathematical physics is the study of qualitative and quantitative properties of the solutions to systems of differential equations. Phenomena which evolve in time are usually described by nonlinear evolution equations, and their natural setting is the Cauchy problem: the state of the system is assigned at an initial time, and the following evolution is, or should be, completely determined by the equations.

The most basic question concerns the local existence and uniqueness of the solution; we should require that, on one hand, the set of initial conditions is not too restrictive, so that a solution exists, and on the other hand it is not too lax, so that a unique solution is specified. For physically meaningful equations it is natural to require in addition that small errors in the initial state propagate slowly in time, so that two solutions corresponding to slightly different initial data remain close, at least for some time (continuous dependence on the data). This set of requirements is called *local well posedness* of the system. *Global* well posedness implies the possibility of extending these properties to arbitrary large times.

However, local well posedness is just the first step in understanding a nonlinear equation; basically, it amounts to a check that the equation and its Cauchy problem are meaningful from the mathematical and physical point of view. But the truly interesting questions concern global existence, asymptotic behavior and more generally the structure and shape of solutions. Indeed, if an equation describes a microscopic system, all quantities measured in an experiment correspond to asymptotic features of the solution, and local properties are basically useless in this context.

The classical tool to prove local well posedness is the energy method; it can be applied to many equations and leads to very efficient proofs. But the energy method is intrinsically unable to give answers concerning the global existence and behavior of solutions.

Let us illustrate this point more precisely. Most nonlinear evolution equations are of the form

$$Lu = N(u)$$

with a linear part L (e. g., the d'Alembert or Schrödinger operator) and a nonlinear term $N(u)$. The two terms are in competition: the linear flow is well behaved, usually with several conserved quantities bounded by norms of the initial data, while the nonlinear term tends to concentrate the peaks of u and make them higher, increasing the singularity of the solution. At the same time, a small function u is made even smaller by a power-like nonlinearity $N(u)$.

The energy method tries to use only the conserved quantities, which remain constant during the evolution of the linear flow. Essentially, this means to regard the equation as an ODE of the form $y'(t) = N(y(t))$. This approach is very useful if the full (nonlinear) equation has a positive conserved quantity. But in general the best one can prove using the energy method is local well posedness, while a blow up in finite time can not be excluded, even if the initial data are assumed to be small, as the trivial example $y' = y^2$ clearly shows.

To improve on this situation, the strength of the linear flow must be used to a full extent. This is where dispersive phenomena enter the picture. A finer study of the operator L shows that there are quantities, different from the standard L^2 type norms used in the energy estimates, which actually *decay* during the evolution. Hence the term L has an additional advantage in the competition with N which can lead to global existence of small solutions. See Sect. "The Nonlinear Wave Equation" for more details.

This circle of ideas was initiated at the beginning of the 1970s and rapidly developed during the 1980s in several papers, (see [26,27,31,32,33,34,38,40,41,42,50] to mention just a few of the fundamental contributions on the subject). Global existence of small solutions was proved for many nonlinear evolution equations, including the nonlinear wave, Klein–Gordon and Schrödinger equations. The proof of the stability of vacuum for the Einstein equations [15,35] can be regarded as an extreme development of this direction of research.

Another early indication of the importance of L^p estimates came from the separate direction of harmonic analysis. In 1970, at a time when general agreement in the field was that the L^2 framework was the only correct one for the study of the wave equation, R. Strichartz ([44,45], see also [10,11,46]) proved the estimate

$$u_{tt} - \Delta u = F\,, \quad u(0,x) = u_t(0,x) = 0 \qquad \Longrightarrow \qquad \|u\|_{L^p(\mathbb{R}^{n+1})} \lesssim \|F\|_{L^{p'}(\mathbb{R}^{n+1})}$$

for $p = 2(n+1)/(n-1)$. This was a seminal result, both for the emphasis on the L^p approach, and for the idea of regarding all variables including time on the same level, an idea which would play a central role in the later developments. At the same time, it hinted at a deep connection

between dispersive properties and the techniques of harmonic analysis.

The new ideas were rapidly incorporated in the mainstream theory of partial differential equations and became an important tool ([22,23,52]); the investigation of *Strichartz estimates* is still in progress. The freedom to work in an L^p setting was revealed as useful in a wide variety of contexts, including linear equations with variable coefficients, scattering, existence of low regularity solutions, and several others.

We might say that this point of view created a new important class of *dispersive equations*, complementing and overlapping with the standard classification into elliptic, hyperbolic, parabolic equations. Actually this class is very wide and contains most evolution equations of mathematical physics: the wave, Schrödinger and Klein–Gordon equations, Dirac and Maxwell systems, their nonlinear counterparts including wave maps, Yang–Mills and Einstein equations, Korteweg de Vries, Benjamin–Ono, and many others. In addition, many basic physical problems are described by systems obtained by coupling these equations; among the most important ones we mention the Dirac–Klein–Gordon, Maxwell–Klein–Gordon, Maxwell–Dirac, Maxwell–Schrödinger and Zakharov systems. Finally, a recent line of research pursues the extension of the dispersive techniques to equations with variable coefficients and equations on manifolds.

The main goal of this article is to give a quick overview of the basic dispersive techniques, and to show some important examples of their applications. It should be kept in mind however that the field is evolving at an impressive speed, mainly due to the contribution of ideas from Fourier and harmonic analysis, and it would be difficult to give a self-contained account of the recent developments, for which we point at the relevant bibliography. In Sect. "The Mechanism of Dispersion" we analyze the basic mechanism of dispersion, with special attention given to two basic examples (Schrödinger and wave equation). Section "Strichartz Estimates" is a coincise exposition of Strichartz estimates, while Sects. "The Nonlinear Wave Equation" and "The Nonlinear Schrödinger Equation" illustrate some typical applications of dispersive techniques to problems of global existence for nonlinear equations. In the last Sect. "Future Directions" we review some of the more promising lines of research in this field.

The Mechanism of Dispersion

The Fourier transform puts in duality oscillations and translations according to the elementary rule

$$\mathcal{F}(f(x+h)) = \mathrm{e}^{-ix\cdot h}\mathcal{F}f\,.$$

The same mechanism lies behind dispersive phenomena, and indeed two dual explanations are possible:

a) in physical space, dispersion is characterized by a finite speed of propagation of the signals; data concentrated on some region tend to spread over a larger volume following the evolution flow. Correspondingly, the L^∞ and other norms of the solution tend to decrease;
b) in Fourier space, dispersion is characterized by oscillations in the Fourier variable ξ, which increase for large $|\xi|$ and large t. This produces cancelation and decay in suitable norms.

We shall explore these dual points of view in two fundamental cases: the Schrödinger and the wave equation. These are the most important and hence most studied dispersive equations, and a fairly complete theory is available for both.

The Schrödinger Equation in Physical Space

The solution $u(t,x)$ of the Schrödinger equation

$$iu_t + \Delta u = 0\,, \quad u(0,x) = f(x)$$

admits several representations which are not completely equivalent. From spectral theory we have an abstract representation as a group of L^2 isometries

$$u(t,x) = \mathrm{e}^{it\Delta} f\,; \tag{1}$$

then we have an explicit representation using the fundamental solution

$$\begin{aligned} u(t,x) &= \frac{1}{(4\pi\, it)^{n/2}}\mathrm{e}^{-i\frac{|x|^2}{4t}} * f \\ &= \frac{1}{(4\pi\, it)^{n/2}}\int_{\mathbb{R}^n} \mathrm{e}^{-i\frac{|x-y|^2}{4t}} f(y)\mathrm{d}y\,; \end{aligned} \tag{2}$$

finally, we can represent the solution via Fourier transform as

$$u(t,x) = (2\pi)^{-n}\int \mathrm{e}^{i(t|\xi|^2+x\cdot\xi)}\widehat{f}(\xi)\mathrm{d}\xi\,. \tag{3}$$

Notice for instance that the spectral and Fourier representations are well defined for $f \in L^2$, while it is not immediately apparent how to give a meaning to (2) if $f \notin L^1$.

From (1) we deduce immediately the *dispersive estimate* for the Schrödinger equation, which has the form

$$|\mathrm{e}^{it\Delta} f| \le (4\pi)^{-n/2}\frac{1}{t^{n/2}}\|f\|_{L^1}\,. \tag{4}$$

Since the L^2 norm of u is constant in time

$$\|\mathrm{e}^{it\Delta} f\|_{L^2} = \|f\|_{L^2}\,, \tag{5}$$

this might suggest that the constant L^2 mass spreads on a region of volume $\sim t^n$ i. e. of diameter t. Apparently, this is in contrast with the notion that the Schrödinger equation has an infinite speed of propagation. Indeed, this is a correct but incomplete statement. To explain this point with an explicit example, consider the evolution of a *wave packet*

$$\pi_{x_0,\xi_0} = \mathrm{e}^{-\frac{1}{2}|x-x_0|^2}\mathrm{e}^{i(x-x_0)\cdot\xi_0} .$$

A wave packet is a mathematical representation of a particle localized near position x_0, with frequency localized near ξ_0; notice indeed that its Fourier transform is

$$\widehat{\pi}_{x_0,\xi_0} = c\mathrm{e}^{-\frac{1}{2}|\xi-\xi_0|^2}\mathrm{e}^{i\xi\cdot x_0}$$

for some constant c. Exact localization both in space and frequency is of course impossible by Heisenberg's principle; however the approximate localization of the wave packet is a very good substitute, since the gaussian tails decay to 0 extremely fast. The evolution of a wave packet is easy to compute explicitly:

$$\mathrm{e}^{it\Delta}\pi_{x_0,\xi_0} = (1+2it)^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\frac{(x-x_0-2t\xi_0)^2}{(1+4t^2)}\right)\mathrm{e}^{i\Gamma}$$

where the real valued function Γ is given by

$$\Gamma = \frac{t(x-x_0)^2 + (x-x_0-t\xi_0)\cdot\xi_0}{(1+4t^2)} .$$

We notice that the particle "moves" with velocity $2\xi_0$ and the height of the gaussian spreads precisely at a rate $t^{n/2}$. The mass is essentially concentrated in a ball of radius t, as suspected. This simple remark is at the heart of one of the most powerful techniques to have appeared in harmonic analysis in recent years, the wave packet decomposition, which has led to breakthroughs in several classical problems (see e. g. [7,8,48,51]).

The Schrödinger Equation in Fourier Space

In view of the ease of proof of (4), it may seem unnecessary to examine the solution in Fourier space, however a quick look at this other representation is instructive and prepares the discussion for the wave equation. If we examine (3), we see that the integral is a typical oscillatory integral which can be estimated using the stationary phase method. The phase here is

$$\phi(t,x;\xi) = t|\xi|^2 + x\cdot\xi \qquad \Longrightarrow \qquad \nabla_\xi\phi = 2t\xi + x$$

so that for each (t,x) we have a unique stationary point $\xi = -x/2t$; the n curvatures at that point are different from zero, and this already suggests a decay of the order $t^{-n/2}$. This argument can be made precise and gives an alternative proof of (4).

We also notice that the Fourier transform in both space and time of $u(t,x) = \mathrm{e}^{it\Delta}f$ is a temperate distribution in $S'(\mathbb{R}^{n+1})$ which can be written, apart from a constant,

$$\tilde{u}(\tau,\xi) = \delta(\tau - |\xi|^2)$$

and is supported on the paraboloid $\tau = |\xi|^2$. This representation of the solution gives additional insight to its behavior in certain frequency regimes, and has proved to be of fundamental importance when studying the existence of low regularity solutions to the nonlinear Schrödinger equation [6,9].

The Wave Equation in Physical Space

For the wave equation, the situation is reversed: the most efficient proof of the dispersive estimate is obtained via the Fourier representation of the solution. This is mainly due to the complexity of the fundamental solution; we recall the relevant formulas briefly. The solution to the homogeneous wave equation

$$w_{tt} - \Delta w = 0 , \quad w(0,x) = f(x) , \quad w_t(0,x) = g(x)$$

can be written as

$$u(t,x) = \cos(t|D|)f + \frac{\sin(t|D|)}{|D|}g$$

where we used the symbolic notations

$$\cos(t|D|)f = \mathcal{F}^{-1}\cos(t|\xi|)\mathcal{F}f ,$$
$$\frac{\sin(t|D|)}{|D|}g = \mathcal{F}^{-1}\frac{\sin(t|\xi|)}{|\xi|}\mathcal{F}g$$

(recall that $\mathcal{F}$ denotes the Fourier transform in space variables). Notice that

$$\cos(t|D|)f = \frac{\partial}{\partial t}\frac{\sin(t|D|)}{|D|}f ,$$

hence the properties of the first operator can be deduced from the second. Moreover, by Duhamel's formula, we can express the solution to the nonhomogeneous wave equation

$$w_{tt} - \Delta w = F(t,x) , \quad w(0,x) = 0 , \quad w_t(0,x) = 0$$

as

$$u(t,x) = \int_0^t \frac{\sin((t-s)|D|)}{|D|}F(u(s,\cdot))\mathrm{d}s .$$

In other words, a study of the operator

$$S(t)g = \frac{\sin(t|D|)}{|D|}g$$

is sufficient to deduce the behavior of the complete solution to the nonhomogeneous wave equation with general data.

The operator $S(t)$ can be expressed explicitly using the fundamental solution of the wave equation. We skip the one-dimensional case since then from the expression of the general solution

$$u(t,x) = \phi(t+x) + \psi(t-x)$$

it is immediately apparent that no dispersion can occur. In odd space dimension $n \geq 3$ we have

$$\frac{\sin(t|D|)}{|D|}g = c_n \left(\frac{1}{t}\partial_t\right)^{\frac{n-3}{2}} \left(\frac{1}{t}\int_{|y|=t} g(x-y)\mathrm{d}\sigma_y\right),$$

and in even space dimension $n \geq 2$

$$\frac{\sin(t|D|)}{|D|}g = c_n \left(\frac{1}{t}\partial_t\right)^{\frac{n-2}{2}} \left(\frac{1}{t}\int_{|y|<t} \frac{g(x-y)}{\sqrt{t^2-|y|^2}}\mathrm{d}y\right)$$

for some constant c_n. These formulas can be expanded and the resulting integrals estimated in order to compute the explicit decay rate of the solution; this approach was followed in [50] (see also [43] for a more concise proof based on the fundamental solution). We shall prove the dispersive estimate using Fourier methods in the next section, but here it is instructive to consider the particularly simple case $n = 3$, when the above representation reduces to *Kirchhoff's formula*

$$\frac{\sin(t|D|)}{|D|}g = \frac{c}{t}\int_{|y|=t} g(x-y)\mathrm{d}\sigma_y\,.$$

Using the divergence theorem it is easy to deduce that

$$\left|\frac{\sin(t|D|)}{|D|}g\right| \lesssim \frac{1}{t}\|\nabla f\|_{L^1}\,.$$

Recall that the rate of decay for the Schrödinger equation in dimension 3 is $\sim t^{-3/2}$; for the wave equation the rate of decay is slower, in addition the estimate shows a loss of one derivative with respect to the initial data. Actually this estimate is optimal, as it is evident from the explicit representation of the solution. In dimension n the correct rate of decay is $t^{-\frac{n-1}{2}}$ and the loss of derivative is of order $(n-1)/2$, as we shall see in the next section.

This result is in apparent contrast with our physical space picture. Indeed the L^2 norm of the first derivatives of the solution remains constant, in view of the energy estimate

$$\|u_t(t,\cdot)\|_{L^2}^2 + \|\nabla u(t,\cdot)\|_{L^2}^2 = \|u_t(0,\cdot)\|_{L^2}^2 + \|\nabla u(0,\cdot)\|_{L^2}^2;$$

on the other hand, the velocity of propagation for the wave equation is exactly 1, so that the constant L^2 energy is spread on a volume which increases at a rate $\sim t^n$; as a consequence, we would expect the same rate of decay $\sim t^{-n/2}$ as for the Schrödinger equation. However, a more detailed study of the shape of the solution reveals that it tends to concentrate in a shell of constant thickness around the light cone. For instance, in odd dimension larger than 3 initial data with support in a ball $B(0,R)$ produce a solution which at time t is supported in the shell $B(0,t+R) \setminus B(0,t-R)$. Also in even dimension the solution tends to concentrate along the light cone $|t| = |x|$, with a faster decreasing tail far from the cone. Thus, the energy spreads over a volume of size t^{n-1}, in accordance with the rate of decay in the dispersive estimate.

The Wave Equation in Fourier Space

Since

$$2i\frac{\sin(t|D|)}{|D|}g = \frac{\mathrm{e}^{it|D|}}{|D|}g - \frac{\mathrm{e}^{-it|D|}}{|D|}g\,,$$
$$2\cos(t|D|)f = \mathrm{e}^{it|D|}f + \mathrm{e}^{-it|D|}f\,,$$

we can unify the study of the various components of the solution by considering the operator

$$\mathrm{e}^{it|D|}f = \mathcal{F}^{-1}\mathrm{e}^{it|\xi|}\mathcal{F}f = (2\pi)^{-n}\int \mathrm{e}^{i(x\cdot\xi+t|\xi|)}\widehat{f}\mathrm{d}\xi\,.$$

The phase of the last integral is

$$\phi(t,x;\xi) = t|\xi| + x\cdot\xi \quad\Longrightarrow\quad \nabla_\xi\phi = t\frac{\xi}{|\xi|} + x$$

and we notice immediately that the structure of stationary points is more complex than for Schrödinger: if $|t| = |x|$, i.e. if the point (t,x) is on the light cone, the phase has a half line of critical points

$$|t| = |x| \quad\Longrightarrow\quad \frac{\xi}{|\xi|} = -\frac{x}{t}$$

which of course are degenerate since only $n-1$ curvatures are different from zero; on the other hand, when (t,x) is not on the light cone, the phase is nondegenerate at all points.

In the nondegenerate case $|t| \neq |x|$, we notice that

$$\nabla_\xi\phi = t\frac{\xi}{|\xi|} + x \quad\Longrightarrow\quad |\nabla_\xi\phi| \geq ||t| - |x||$$

hence by direct integration by parts we obtain

$$\left|\int e^{i(x\cdot\xi+t|\xi|)}\widehat{f}d\xi\right| \le C(f)||t|-|x||^{-(n-1)} . \tag{6}$$

Notice that the singularity $|\xi|$ of the phase makes it impossible to perform the integration by parts more than $n-1$ times, but this is largely sufficient for our purposes here.

In the degenerate case $|t|=|x|$, we can eliminate the degeneracy by passing to polar coordinates:

$$\int e^{i(x\cdot\xi+t|\xi|)}\widehat{f}d\xi = \int_0^\infty e^{it\rho}\rho^{n-1} \times \int_{|\omega|=1} e^{ix\cdot\rho\omega}\widehat{f}(\rho\omega)d\omega d\rho .$$

Now the phase in the inner integral has exactly two isolated critical points for each value of x; by a standard application of the stationary phase method we can obtain an estimate of the form

$$\left|\int e^{i(x\cdot\xi+t|\xi|)}\widehat{f}d\xi\right| \le C(f)|x|^{-\frac{n-1}{2}} . \tag{7}$$

The power $(n-1)/2$ is directly related to the dimension $n-1$ of the unit sphere.

From the above computations it is possible to obtain a very detailed description of the shape of the solution (see e. g. [20]). However, here we are only interested in the rate of decay of the L^∞ norm of the solution. Thus, using estimate (6) when $|x|<|t|/2$ and (7) when $|x|>|t|/2$ we obtain

$$|e^{it|D|}f| \le C(f)|t|^{-\frac{n-1}{2}} .$$

It is also possible to make precise the form of the constant $C(f)$ appearing in the estimate. This can be done in several ways; one of the most efficient exploits the scaling properties of the wave equation (see [10,23]). In order to state the optimal estimate we briefly recall the definition of some function spaces. The basic ingredient is the homogeneous *Paley–Littlewood partition of unity* which has the form

$$\sum_{j\in\mathbb{Z}}\phi_j(\xi)=1 , \quad \phi_j(\xi)=\phi_0(2^{-j}\xi) , \quad \sup\phi_j=\{\xi : 2^{j-1}\le|\xi|\le 2^{j+1}\}$$

for a fixed $\phi_0\in C_0^\infty(\mathbb{R}^n)$ (the following definitions are independent of the precise choice of ϕ_0). To each symbol $\phi_j(\xi)$ we associate by the standard calculus the operator $\phi_j(D)=\mathcal{F}^{-1}\phi_j(\xi)\mathcal{F}$. Now the *homogeneous Besov norm* $\dot{B}^s_{1,1}$ can be defined as follows:

$$\|f\|_{\dot{B}^s_{1,1}} = \sum_{j\in\mathbb{Z}} 2^{js}\|\phi_j(D)f\|_{L^1} \tag{8}$$

and the corresponding Besov space can be obtained by completing C_0^∞ (see [4] for a thorough description of these definitions).

Then the optimal dispersive estimate for the wave equation is

$$|e^{it|D|}f| \le C\|f\|_{\dot{B}^{\frac{n-1}{2}}_{1,1}}|t|^{-\frac{n-1}{2}} . \tag{9}$$

We might add that in odd space dimension a slightly better estimate can be proved involving only the standard Sobolev norm $\dot{W}^{\frac{n-1}{2}}$ instead of the Besov norm.

The singularity in t at the right hand side of (9) can be annoying in applications, thus it would be natural to try to replace $|t|$ by $(1+|t|)$ by some different proof. However, this is clearly impossible since both sides have the same scaling with respect to the transformation $u(t,x)\to u(\lambda t,\lambda x)$ which takes a solution of the homogeneous wave equation into another solution. Moreover, if such an estimate were true, taking the limit as $t\to 0$ we would obtain a false Sobolev embedding. Thus, the only way to get a nonsingular estimate is to replace the Besov norm of the data by some stronger norm.

Indeed, for $|t|<1$ we can use the L^2 conservation which implies, for any $\epsilon>0$,

$$|e^{it|D|}f| \le \|e^{it|D|}f\|_{H^{n/2+\epsilon}} = \|f\|_{H^{n/2+\epsilon}} .$$

Combining the two estimates, we arrive at

$$|e^{it|D|}f| \lesssim \left(\|f\|_{\dot{B}^{\frac{n-1}{2}}_{1,1}} + \|f\|_{H^{n/2+\epsilon}}\right)(1+|t|)^{-\frac{n-1}{2}} . \tag{10}$$

Other Dispersive Equations

Dispersive estimates can be proved for many other equations; see for instance [3] for general results concerning the equations of the form

$$iu_t = P(D)u , \quad P \text{ polynomial}$$

which correspond to integrals of the form

$$\int e^{i(x\cdot xi+P(\xi))}\widehat{f}(\xi)d\xi .$$

Here we would like to add some more details on two equations of particular importance for physics. The *Klein–Gordon* equation

$$u_{tt}-\Delta u+u=0 , \quad u(0,x)=f(x) , \quad u_t(0,x)=g(x)$$

shows the same decay rate as the Schrödinger equation, although for most other properties it is very close to the wave

equation. This phenomenon can be easily explained. In physical space, we notice that the solution has finite speed of propagation less or equal to 1, so that for compactly supported initial data the solution spreads uniformly on a volume of order t^n, and there is no concentration along the light cone as for the wave equation.

On the other hand, the Fourier representation of the solution is

$$u(t,x) = \cos(t\langle D\rangle)f + \frac{\sin(t\langle D\rangle)}{\langle D\rangle}g$$

where we used the symbolic notations

$$\cos(t\langle D\rangle)f = \mathcal{F}^{-1}\cos(t|\langle\xi\rangle|)\mathcal{F}f\,,$$
$$\frac{\sin(t\langle D\rangle)}{\langle D\rangle}g = \mathcal{F}^{-1}\frac{\sin(t|\langle\xi\rangle|)}{|\langle\xi\rangle|}\mathcal{F}g$$

with $\langle\xi\rangle = (1+|\xi|^2)^{1/2}$ i.e. $\langle D\rangle = (1-\Delta)^{1/2}$. Thus, we see that, similarly to the wave equation, the study of the solutions can be reduced to the basic operator

$$\mathrm{e}^{it\langle D\rangle}f = \mathcal{F}^{-1}\mathrm{e}^{it\langle\xi\rangle}\mathcal{F}f = (2\pi)^{-n}\int \mathrm{e}^{i(t\langle\xi\rangle+x\cdot\xi)}\widehat{f}(\xi)\mathrm{d}\xi\,.$$

It is easy to check that the phase has an isolated stationary point if and only if $|t| < |x|$, and no stationary point if $|t| \geq |\xi|$:

$$\phi(t,x;\xi) = t\langle\xi\rangle + x\cdot\xi\,, \quad \nabla_\xi\phi = t\frac{\xi}{\langle\xi\rangle} + x\,,$$

in accord with the claimed dispersion rate of order $t^{-n/2}$. However, the proof is not completely elementary (for a very detailed study, see Section 7.2 of [25]). The optimal estimate is now

$$|\mathrm{e}^{it\langle D\rangle}f| \leq C\|f\|_{B^{\frac{n}{2}}_{1,1}}|t|^{-\frac{n}{2}} \tag{11}$$

(see the Appendix to [18]). Here we used the *nonhomogeneous Besov norm* $B^s_{1,1}$ which is defined exactly as the homogeneous one (8) by using the inhomogeneous version of the Palye–Littlewood partition of unity:

$$\phi_{-1}(\xi) + \sum_{j\in\mathbb{N}}\phi_j(\xi) = 1\,,$$
$$\phi_j(\xi) = \phi_0(2^{-j}\xi) \quad \text{for} \quad j \geq 0,$$
$$\sup\phi_j = \{\xi\colon 2^{j-1} \leq |\xi| \leq 2^{j+1}\}\,,$$
$$\sup\phi_{-1} = \{|x| \leq 1\}$$

for some $\phi_{-1}, \phi_0 \in C_0^\infty(\mathbb{R}^n)$.

A second important example is the *Dirac system*. From a mathematical point of view (and setting the values of the speed of light and Planck's constant equal to 1 for simplicity of notation) this is a 4×4 constant coefficient system of the form

$$iu_t + \mathcal{D}u = 0$$

in the *massless* case, and

$$iu_t + \mathcal{D}u + \beta u = 0$$

in the *massive* case. Here $u\colon \mathbb{R}_t\times\mathbb{R}^3_x \to \mathbb{C}^4$, the operator $\mathcal{D}$ is defined as

$$\mathcal{D} = \frac{1}{i}\sum_{k=1}^{3}\alpha_k\partial_k$$

and the 4×4 *Dirac matrices* can be written

$$\alpha_k = \begin{pmatrix} 0 & \sigma_k \\ \sigma_k & 0\end{pmatrix}, \quad \beta = \begin{pmatrix} I_2 & 0 \\ 0 & -I_2\end{pmatrix}, \quad k = 1,2,3$$

in terms of the *Pauli matrices*

$$I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1\end{pmatrix}, \qquad \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0\end{pmatrix},$$
$$\sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0\end{pmatrix}, \qquad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1\end{pmatrix}.$$

Then the solution $u(t,x) = \mathrm{e}^{it\mathcal{D}}f$ of the massless Dirac system with initial value $u(0,x) = f(x)$ satisfies the dispersive estimate

$$|u(t,x)| \leq \frac{C}{t}\|f\|_{\dot{B}^2_{1,1}}$$

while in the massive case we have

$$|u(t,x)| \leq \frac{C}{t^{\frac{3}{2}}}\|f\|_{B^{\frac{5}{2}}_{1,1}}$$

(see Sect. 7 of [16]).

Strichartz Estimates

Since the Schrödinger flow $\mathrm{e}^{it\Delta}f$ is an L^2 isometry, it may be tempting to regard it as some sort of "rotation" of the Hilbert space L^2. This picture is far from true: there exist small subspaces of L^2 which contain the flow for almost all times. This surprising phenomenon, which holds in greater generality for any unbounded selfadjoint operator on an abstract Hilbert space, is known as *Kato smoothing* and was discovered by T. Kato in 1966 [29], see also [39]. This fact is just a corollary of a quantitative estimate, the *Kato smoothing estimate*, which in the case of the Laplace

operator takes the following form: there exist closed unbounded operators A on L^2 such that

$$\|Ae^{it\Delta}f\|_{L^2L^2} \lesssim \|f\|_{L^2} .$$

As a consequence, the flow belongs for a. e. t to the domain of the operator A, which can be a very small subspace of L^2. Two simple examples of smoothing estimates are the following:

$$V(x) \in L^2 \cap L^{n-\epsilon} \cap L^{n+\epsilon} \quad \Longrightarrow \quad \|V(x)e^{it\Delta}f\|_{L^2L^2} \lesssim \|f\|_{L^2} \tag{12}$$

and

$$\||x|^{-1}e^{it\Delta}f\|_{L^2L^2} \lesssim \|f\|_{L^2} . \tag{13}$$

On the other hand, Strichartz proved in 1977 [46] that

$$p = \frac{2(n+2)}{n} \quad \Longrightarrow \quad \|e^{it\Delta}f\|_{L^p(\mathbb{R}^{n+1})} \lesssim \|f\|_{L^2} \tag{14}$$

thus showing that the smoothing effect could be measured also in the L^p norms. If we try to reformulate (12) in terms of an L^p norm, we see that it would follow from

$$\|e^{it\Delta}f\|_{L^2L^{\frac{2n}{n-2}}} \lesssim \|f\|_{L^2} \tag{15}$$

while (13) would follow from the slightly stronger

$$\|e^{it\Delta}f\|_{L^2L^{\frac{2n}{n-2},2}} \lesssim \|f\|_{L^2} \tag{16}$$

where $L^{p,2}$ denotes a Lorentz space. Both (15) and (16) are true, although their proof is far from elementary. Such estimates are now generally called *Strichartz estimates*. They can be deduced from the dispersive estimates, hence they are in a sense weaker and contain less information. However, for this same reason they can be obtained for a large number of equations, even with variable coefficients, and their flexibility and wide range of applications makes them extremely useful.

The full set of Strichartz estimates, with the exception of the endpoint (see below) was proved in [22], where it was also applied to the problem of global well posedness for the nonlinear Schrödinger equation. The corresponding estimates for the wave equation were obtained in [23], however they turned out to be less useful because of the loss of derivatives. A fundamental progress was made in [30], where not only the most difficult endpoint case was solved, but it also outlined a general procedure to deduce the Strichartz estimates from the dispersive estimates, with applications to any evolution equation. In the following we shall focus on the two main examples, the Schrödinger and the wave equation, and we refer to the original papers for more details and the proofs.

The Schrödinger Equation

The homogeneous Strichartz estimates have the general form

$$\|e^{it\Delta}f\|_{L^pL^q} \lesssim \|f\|_{L^2}$$

for suitable values of the couple (p, q). Notice that we can restrict the L^pL^q norm at the right hand side to an arbitrary time interval I (i. e. to $L^p(I; L^q(\mathbb{R}^n))$). Notice also that by the invariance of the flow under the scaling $u(t.x) \mapsto u(\lambda^2 t, \lambda x)$ the possible couples (p, q) are bound to satisfy

$$\frac{2}{p} + \frac{n}{q} = \frac{n}{2} . \tag{17}$$

The range of possible values of q is expressed by the inequalities

$$2 \le q \le \frac{2n}{n-2} \quad \text{if} \quad n \ge 2 , \quad q \ne \infty , \tag{18}$$

$$2 \le q \le \infty \quad \text{if} \quad n = 1 ; \tag{19}$$

the index p varies accordingly between ∞ and 2 (between 4 and 2 in one space dimension). Such couples (p, q) are called *(Schrödinger) admissible*. It is easy to visualize the admissible couples by plotting the usual diagram in the $(1/p, 1/q)$ space (see Fig 1–3).

Note that the choice $(p, q) = (\infty, 2)$, the point E in the diagrams, corresponds to the conservation of the L^2 norm, while at the other extreme we have for $n \ge 2$ point P corresponding to the couple

$$(p, q) = \left(2, \frac{2n}{n-2}\right) . \tag{20}$$

This is usually called the *endpoint* and is excluded in dimension 2, allowed in dimension ≥ 3. The original Strichartz estimate corresponds to the case $p = q$ and is denoted by S. After the proof in [22] of the general Strichartz estimates, it was not clear if the endpoint case P could be reached or not; the final answer was given in [30] where the endpoint estimates were proved to be true for all dimension larger than three. For the two-dimensional case and additional estimates, see [47]. Notice that all admissible couples can be obtained by interpolation between the endpoint and the L^2 conservation, thus in a sense the endpoint contains the maximum of information concerning the L^p smoothing properties of the flow.

There exist an *inhomogeneous* form of the estimates, which is actually equivalent to the homogeneous one, and can be stated as follows:

$$\left\| \int_0^t e^{i(t-s)\Delta} F(s) ds \right\|_{L^pL^q} \lesssim \|F\|_{L^{\tilde{p}'}L^{\tilde{q}'}} \tag{21}$$

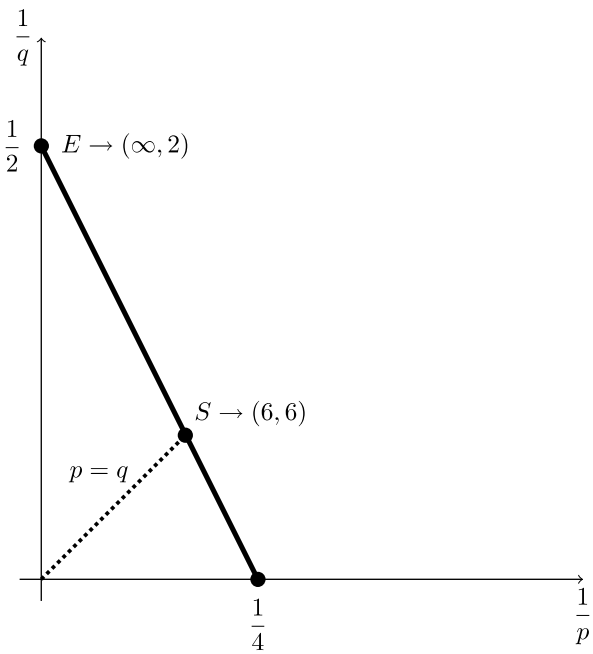

Dispersion Phenomena in Partial Differential Equations, Figure 1
Schrödinger admissible couples, $n = 1$

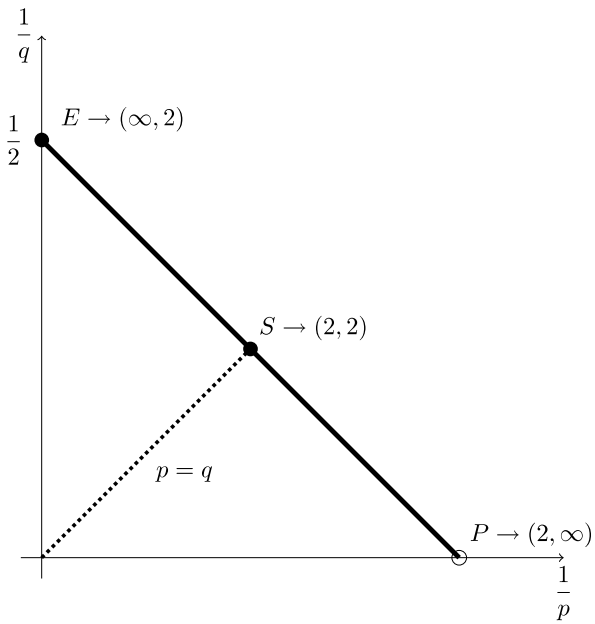

Dispersion Phenomena in Partial Differential Equations, Figure 2
Schrödinger admissible couples, $n = 2$

for all admissible couples (p, q) and $(\tilde{p}, \tilde{q})$, which can be chosen independently. Here p' denotes the conjugate exponent to p. These estimates continue to be true in their *local* version, with the norms at both sides restricted to a fixed time interval $I \subset \mathbb{R}$. Notice that the inhomogeneous global estimates are optimal since other combination of indices are excluded by scale invariance; however, the local Strichartz estimates can be slightly improved allowing for indices outside the admissible region (see [21]).

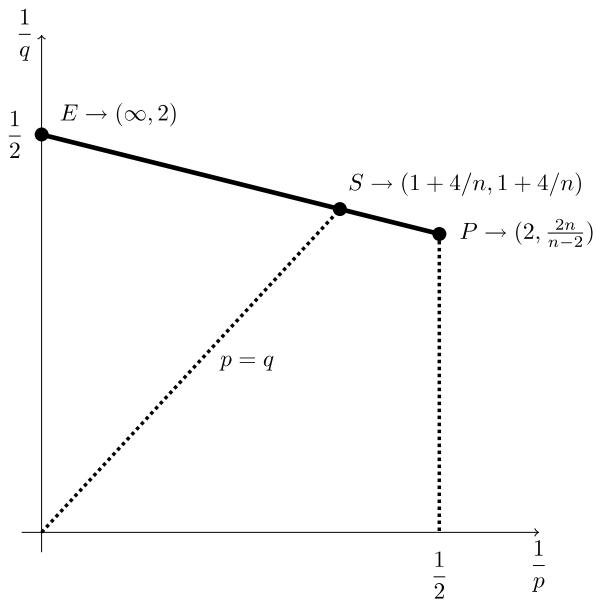

Dispersion Phenomena in Partial Differential Equations, Figure 3
Schrödinger admissible couples, $n \geq 3$

We also mention that a much larger set of estimates can be deduced from the above ones by combining them with Sobolev embeddings; in the above diagrams, all couples in the region bounded by the line of admissibility, the axes and the line $1/p = 2$ can be reached, while the external region is excluded by suitable counterexamples (see [30]).

Strichartz Estimates for the Wave Equation

Most of the preceding section has a precise analogue for the wave equation. Since the decay rate for the dispersive estimates is $t^{-(n-1)/2}$ for the wave equation instead of $t^{-n/2}$ as for the Schrödinger equation, this causes a shift $n \to n-1$ in the numerology of indices. A more serious difference is the loss of derivatives: this was already apparent in the original estimate by Strichartz [44], which can be written

$$\|e^{it|D|} f\|_{L^{\frac{2(n+1)}{n-1}}} \lesssim \|f\|_{\dot{H}^{1/2}}, \quad n \geq 2 .$$

The general homogeneous Strichartz estimates for the wave equation take the form

$$\|e^{it|D|} f\|_{L^p L^q} \lesssim \|f\|_{\dot{H}^{\frac{1}{p}\frac{n+1}{n-1}}} ;$$

now the *(wave) admissible* couples of indices must satisfy the scaling condition

$$\frac{2}{p} + \frac{n-1}{q} = \frac{n-1}{2} \tag{22}$$

and the constraints $2 \le p \le \infty$,

$$2 \le q \le \frac{2(n-1)}{n-3} \quad \text{if} \quad n \ge 3\,, \quad q \ne \infty \tag{23}$$

$$2 \le q \le \infty \quad \text{if} \quad n = 2\,, \tag{24}$$

while p varies accordingly between 2 and ∞ (between 4 and ∞ in dimension 2). Recall that in dimension 1 the solutions of the wave equation do not decay since they can be decomposed into traveling waves. We omit the graphic representation of these sets of conditions since it is completely analogous to the Schrödinger case (apart from the shift $n \to n-1$).

The inhomogeneous estimates may be written

$$\left\| \int_0^t \mathrm{e}^{i(t-s)\Delta} F(s)\mathrm{d}s \right\|_{L^p \dot{H}_q^{-\frac{1}{p}\frac{n+1}{n-1}}} \lesssim \|F\|_{L^{p'} \dot{H}_{q'}^{-\frac{1}{p}\frac{n+1}{n-1}}} \tag{25}$$

in terms of the homogeneous Sobolev norms

$$\|g\|_{\dot{H}_q^s} = \||D|^s g\|_{L^q}\,, \quad |D|^s = (1-\Delta)^{s/2}\,.$$

Other Equations

The general procedure of Keel and Tao can be applied to a large variety of equations. For the convenience of the reader we list here the precise form of the homogeneous Strichartz estimates in some important cases.

For the *Klein–Gordon* equation we have

$$\|\mathrm{e}^{it\langle D\rangle} f\|_{L^p L^q} \lesssim \|f\|_{\dot{H}^{\frac{1}{p}\frac{n+2}{n}}}$$

with (p, q) Schrödinger admissible (recall that the rate of dispersion is $t^{-n/2}$ for Klein–Gordon). For the *massless Dirac* system we have

$$\|\mathrm{e}^{it\mathcal{D}} f\|_{L^p L^q} \lesssim \|f\|_{\dot{H}^{\frac{2}{p}}}\,, \quad n = 3\,,$$

with (p, q) wave admissible in dimension 3, while for the *massive Dirac* system the estimate takes the form

$$\|\mathrm{e}^{it(\mathcal{D}+\beta)} f\|_{L^p L^q} \lesssim \|f\|_{H^{\frac{5}{3p}}}\,, \quad n = 3$$

with (p, q) Schrödinger admissible in dimension 3.

The Nonlinear Wave Equation

In this section we shall illustrate, using the semilinear wave equation as a basic example, what kind of improvements can be obtained over the classical energy method via the additional information given by dispersive techniques.

From the linear theory it is well known that the correct setting of the Cauchy problem for the wave equation

$$u_{tt} - \Delta u = F(t,x)\,, \quad u(t,x)\colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{C}$$

requires two initial data at $t = 0$:

$$u(0,x) = u_0(x)\,, \quad u_t(0,x) = u_1(x)\,.$$

For this equation the energy method is particularly simple. If we multiply the equation by $\overline{u}_t$ we can recast it in the form

$$\partial_t \left(|u_t|^2 + |\nabla u|^2\right) = 2\Re \nabla \cdot (\overline{u}_t \nabla u) + 2\Re(\overline{u}_t F)$$

and integrating this identity on different domains we can obtain several conservation laws. Integrating on a slab $0 \le t \le T$, $x \in \mathbb{R}^n$ we arrive at an estimate of the form

$$\|u_t\|_{L_T^\infty L^2} + \|\nabla u\|_{L_T^\infty L^2} \lesssim \|u_1\|_{L^2} + \|\nabla u_0\|_{L^2} + \|F\|_{L_T^1 L^2} \tag{26}$$

where we are using the shorthand notation

$$L_T^p L^q = L^p(0, T; L^q(\mathbb{R}^n))\,.$$

Estimate (26) is the basic *energy estimate*, which is enough to prove the existence of a global solution $u \in C(\mathbb{R}; H^1(\mathbb{R}^n)) \cap C^1(\mathbb{R}; L^2(\mathbb{R}^n))$ of the linear wave equation for data in $(u_0, u_1) \in H^1 \times L^2$ and a forcing term in $F \in L_T^1 L^2$ for all T. Uniqueness and continuous dependence on the data are simple consequences of the energy estimate. Moreover, if we differentiate the equation k times, the same computations lead to the higher order energy estimates

$$\|u_t\|_{L_T^\infty H^k} + \|u\|_{L_T^\infty H^{k+1}} \lesssim \|u_1\|_{H^k} + \|u_0\|_{H^{k+1}} + \|F\|_{L_T^1 H^k}\,, \quad k = 0, 1, 2, \ldots \tag{27}$$

Consider now the nonlinear wave equation

$$u_{tt} - \Delta u = F(u) \tag{28}$$

for some function $F(s)$ sufficiently smooth. The classical approach to solve it is to look for a fixed point of the mapping $u = \Phi(v)$, where $u(t,x)$ solves the linearized equation

$$u_{tt} - \Delta u = F(v)\,. \tag{29}$$

If we apply the energy estimate to the difference of two solutions $u_1 = \Phi(v_1)$ and $u_2 = \Phi(v_2)$ with the same initial data, we obtain

$$\|\partial_t(u_1 - u_2)\|_{L^\infty_T H^k} + \|u_1 - u_2\|_{L^\infty_T H^{k+1}} \lesssim \|F(v_1) - F(v_2)\|_{L^1_T H^k} .$$

Using a standard Moser-type estimate (see e. g. [25,49])

$$\|f(v_1) - f(v_2)\|_{H^k} \lesssim C(\|v_1\|_{L^\infty} + \|v_2\|_{L^\infty})\|v_1 - v_2\|_{H^k}$$

we arrive at the inequality

$$\|\partial_t(u_1 - u_2)\|_{L^\infty_T H^k} + \|u_1 - u_2\|_{L^\infty_T H^{k+1}} \lesssim C(\|v_1\|_{L^\infty_T L^\infty} + \|v_2\|_{L^\infty_T L^\infty}) \cdot \|v_1 - v_2\|_{L^1_T H^k} .$$

In particular, writing $X_T = L^\infty_T H^k$, and noticing that

$$\|G\|_{L^1_T H^k} \le \|G\|_{X_T} \cdot T ,$$

we obtain

$$\|u_1 - u_2\|_{X_T} \lesssim C(\|v_1\|_{L^\infty_T L^\infty} + \|v_2\|_{L^\infty_T L^\infty}) \cdot \|v_1 - v_2\|_{X_T} \cdot T .$$

We still have to handle the L^∞ norm of v_j, which can be controlled by the H^k norm provided $k > n/2$. In conclusion

$$k > \frac{n}{2} \quad \Longrightarrow \quad \|u_1 - u_2\|_{X_T} \lesssim C(\|v_1\|_{X_T} + \|v_2\|_{X_T}) \cdot \|v_1 - v_2\|_{X_T} \cdot T .$$

At this point the philosophy of the method should be clear: if we agree to differentiate $n/2$ times the equation, we can prove a nonlinear estimate, with a coefficient which can be made small for T small. Thus, for $T \ll 1$ the mapping $v \mapsto u$ is a contraction, and we obtain a local (in time) solution of the nonlinear equation.

The energy method is quite elementary, yet it has the remarkable property that it can be adapted to a huge number of equations: it can be applied to equations with variable coefficients, and even to fully nonlinear equations. An outstanding example of its flexibility is the proof of the local solvability of Einstein's equations ([14]).

The main drawbacks of the energy method are:

1) A high regularity of the data and of the solution is required;
2) The nonlinear nature of the estimate leads to a *local* existence result, leaving completely open the question of global existence.

Actually, the two difficulties are related to each other. Indeed, from the above argument one can check that the lifespan of the local solution is a function of the H^k norm of the initial data. Most equations with a physical origin possess conserved quantities, which are essentially H^k norms of the solution with k low, e. g. $k = 0$ or 1. Assume now that we can prove the local existence for data say in H^1, on some interval $[0, T_0]$ with T_0 depending on the H^1 norm of the data. Since this norm is conserved, we can take T_0 as a new initial time and construct a solution on $[T_0, 2T_0]$, and so on. This is the essence of the *continuation method*. As a consequence, the local existence of low regularity solution with the addition of suitable conservation laws can lead to global existence.

We shall now apply the dispersive information to improve on the preceding result and obtain the global existence of small solution, under suitable assumptions on the nonlinear term. Consider a free wave $w(t, x)$, solution of

$$w_{tt} - \Delta w = 0 , \quad w(0, x) = f(x) , \quad w_t(0, x) = g(x) .$$

Recalling the dispersive estimate (10), here the following simplified version

$$|w(t, x)| \le C(1 + t)^{-\frac{n-1}{2}} \|f\|_{W^{N,1}} , \quad N > n$$

will be sufficient. This suggests modification of the energy norm by adding a term that incorporates the decay information. To this end, we introduce the time-dependent quantity

$$M(t) = \sup_{0 \le \tau \le t} \left\{ \|u(\tau, \cdot)\|_{H^{k+1}} + (1 + \tau)^{\frac{n-1}{2}} \|u(\tau, \cdot)\|_{L^\infty} \right\} ,$$
$$k = \left[\frac{n}{2} \right] .$$

Consider now the nonlinear Eq. (28) with initial data u_0, u_1. We already know that a local solution exists, provided the initial data are smooth enough; we shall now try to extend this solution to a global one. Using Duhamel's principle we rewrite the equation as

$$u(t, x) = w(t, x) + \int_0^t \frac{\sin((t - s)|D|)}{|D|} F(u(s, \cdot))\mathrm{d}s$$

where we are using the symbolic notations of Sect. "Other Dispersive Equations". $\mathcal{F}$ being the space Fourier transform. Note that the above dispersive estimate can be applied in particular to the last term; we obtain

$$\left| \frac{\sin((t - s)|D|)}{|D|} F(s) \right| \le C(1 + t - s)^{-\frac{n-1}{2}} \|F\|_{W^{N,1}} .$$

By this inequality and the energy estimate already proved, in a few steps we arrive at the estimate

$$M(T) \lesssim C_0 + \|F\|_{L^1_T H^k} + (1+T)^{\frac{n-1}{2}} \int_0^T (1+T-s)^{-\frac{n-1}{2}} \|F(u(s))\|_{W^{N,1}} \mathrm{d}s$$

where $C_0 = \|u_0\|_{W^{N,1}} + \|u_1\|_{W^{N,1}}$ and the mute constant is independent of T. By suitable nonlinear estimates of Moser type, we can bound the derivatives of the nonlinear term using only the L^∞ and the H^k norm of u, i. e., the quantity $M(t)$ itself. If we assume that $F(u) \sim |u|^\gamma$ for small u, we obtain

$$M(T) \lesssim C_0 + M(T)^\gamma (1+T)^{\frac{n-1}{2}} \int_0^T (1+T-s)^{-\frac{n-1}{2}} (1+s)^{-\frac{n-1}{2}(\gamma-2)} \mathrm{d}s\,.$$

Now we have

$$\int_0^t (1+t-\tau)^{-\frac{n-1}{2}} \cdot (1+\tau)^{-\frac{n-1}{2}\cdot(\gamma-2)} \mathrm{d}\tau \le C(1+t)^{-\frac{n-1}{2}}$$

provided

$$\gamma > 2 + \frac{2}{n-1}$$

so that, for such values of γ, we obtain

$$M(T) \lesssim C_0 + M(T)^\gamma.$$

For small initial data, which means small values of the constant C_0, (30) implies that $M(T)$ is uniformly bounded as long as the smooth solution exists. By a simple continuation argument, this implies that the solution is in fact global, provided the initial data are small enough.

The Nonlinear Schrödinger Equation

Strichartz estimates for the Schrödinger equation are expressed entirely in terms of L^p norms, without loss of derivatives, in contrast with the case of the wave equations. For this reason they represent a perfect tool to handle semilinear perturbations. Using Strichartz estimates, several important results can be proved with minimal effort. As an example, we chose to illustrate in this Section the critical and subcritical well posedness in L^2 in the case of power nonlinearities (see [13] for a systematic treatment).

Consider the Cauchy problem

$$iu_t - \Delta u = |u|^\gamma\,, \quad u(0,x) = f(x) \tag{31}$$

for some $\gamma \ge 1$. We want to study the local and global solvability in L^2 of this Cauchy problem. Recall that this is a difficult question, still not completely solved; however, at least in the subcritical range, a combination of dispersive techniques and the contraction mapping principle gives the desired result easily.

Instead of Eq. (31), we are actually going to solve the corresponding integral equation

$$u(t,x) = \mathrm{e}^{it\Delta} f + i \int_0^t \mathrm{e}^{i(t-s)\Delta} |u(s)|^\gamma \mathrm{d}s\,.$$

Our standard approach is to look for a fixed point of the mapping $u = \Phi(v)$ defined by

$$\Phi(v) = \mathrm{e}^{it\Delta} f + i \int_0^t \mathrm{e}^{i(t-s)\Delta} |v(s)|^\gamma \mathrm{d}s\,;$$

it is easy to check, using Strichartz estimates, that the mapping Φ is well defined, provided the function v is in the suitable $L^p L^q$. Indeed we can write, for all admissible couples (p,q) and $(\tilde{p},\tilde{q})$

$$\|\Phi(v)\|_{L^p L^q} \lesssim \|f\|_{L^2} + \||\Phi(v)|^\gamma\|_{L^{\tilde{p}'} L^{\tilde{q}'}}$$

and by Hölder's inequality this implies

$$\|\Phi(v)\|_{L^p L^q} \le C_1 \|f\|_{L^2} + C_1 \|v\|^\gamma_{L^{\gamma\tilde{p}'} L^{\gamma\tilde{q}'}}\,. \tag{32}$$

Thus, we see that the mapping Φ operates on the space $L^p L^q$, provided we can satisfy the following set of conditions:

$$\begin{gathered} \tilde{p}'\gamma = p\,, \quad \tilde{q}'\gamma = q, \\ \frac{2}{p} + \frac{n}{q} = \frac{n}{2}, \quad \frac{\tilde{2}}{\tilde{p}} + \frac{n}{\tilde{q}} = \frac{n}{2}, \\ p, \tilde{p} \in [2,\infty]\,, \quad \tilde{q} \in \left[2, \frac{2n}{n-2}\right]\,. \end{gathered} \tag{30}$$

Notice that from the first two identities we have

$$\frac{2\gamma}{p} + \frac{n\gamma}{q} = \alpha - \frac{2}{\tilde{p}} + n - \frac{n}{\tilde{q}}$$

and by the admissibility conditions this implies

$$\frac{n}{2}\gamma = 2 + n - \frac{n}{2}\,,$$

which forces the value of γ to be

$$\gamma = 1 + \frac{4}{n}\,. \tag{33}$$

This value of γ is called L^2 *critical*.

Let us assume for the moment that γ is exactly the critical exponent (33). Then it is easy to check that there exist choices of (p, q) and $(\tilde{p}, \tilde{q})$ which satisfy all the above constraints. To prove that Φ is a contraction on the space

$$X = L^p L^q$$

we take any $v_1, v_2 \in X$ and compute

$$\begin{aligned}\|\Phi(v_1) - \Phi(v_2)\|_X &\lesssim \big\| |v_1|^\gamma - |v_2|^\gamma \big\|_{L^{\tilde{p}'} L^{\tilde{q}'}} \\ &\lesssim \big\| |v_1 - v_2| (|v_1|^{\gamma-1} + |v_2|^{\gamma-1}) \big\|_{L^{\tilde{p}'} L^{\tilde{q}'}} ;\end{aligned}$$

the same computation as above gives

$$\|\Phi(v_1) - \Phi(v_2)\|_X \le C_2 \|v_1 - v_2\|_X \||v_1| + |v_2|\|_X^{\gamma-1} \quad (34)$$

for some constant C_2. Now assume that

$$\|v_i\|_X < \varepsilon , \quad \|f\|_{L^2} < \delta ;$$

using (32) we can write

$$\|\Phi(v)\|_X \le C_1(\delta + \varepsilon^\gamma) < \varepsilon$$

provided ε, δ are so small that

$$C_1 \delta < \frac{\varepsilon}{2} , \quad C_1 \varepsilon^\gamma < \frac{\varepsilon}{2} .$$

On the other hand, by (34) we get

$$\|\Phi(v_1) - \Phi(v_2)\|_X \le C_2 \|v_1 - v_2\|_X (2\varepsilon)^{\gamma-1} < \frac{1}{2} \|v_1 - v_2\|_X$$

provided ε is so small that

$$C_2 (2\varepsilon)^{\gamma-1} < \frac{1}{2} .$$

In conclusion, if the initial data are small enough, the mapping Φ is a contraction on a ball $B(0, \varepsilon) \subset X$ and the unique fixed point is a global small solution of the critical L^2 equation

$$iu_t - \Delta u = |u|^{1+\frac{4}{n}} .$$

Notice that in estimate (32) the first couple can also be taken equal to $(\infty, 2)$, thus the solution we have constructed belongs to $L^\infty L^2$. A limit argument (dominated convergence) in the identity

$$u = \mathrm{e}^{it\Delta} f + i \int_0^t \mathrm{e}^{i(t-s)\Delta} |u(s)|^\gamma \mathrm{d}s ;$$

proves that we have actually $u \in C(\mathbb{R}; L^2(\mathbb{R}^n))$.

We briefly sketch a proof of the local existence in the subcritical case $1 < \gamma < 1 + \frac{4}{n}$. Consider the local space

$$X_T = L_T^p L^q = L^p(0, T; L^q(\mathbb{R}^n)) .$$

If we follow the steps of the preceding proof and choose the same indices, we have now

$$\tilde{p}' \gamma < p , \quad \tilde{q}' \gamma < q .$$

It is easy to check that we can shift the point $(\tilde{p}, \tilde{q})$ so that

$$\tilde{p}' \gamma < p , \quad \tilde{q}' \gamma = q$$

and proceeding as above and using Hölder inequality in time, we obtain the inequality. Indeed, choosing the indices as above and applying Hölder's inequality in time we have

$$\|\Phi(v)\|_{X_T} \le C_1 \|f\|_{L^2} + C_1 T^\lambda \|v\|_{X_T} \quad (35)$$

for some strictly positive $\lambda > 0$. Analogously, estimate (34) will be replaced by

$$\begin{aligned}&\|\Phi(v_1) - \Phi(v_2)\|_{X_T} \\ &\qquad \le C_2 T^\lambda \|v_1 - v_2\|_{X_T} \||v_1| + |v_2|\|_{X_T}^{\gamma-1} . \quad (36)\end{aligned}$$

Now let M be so large and T so small that

$$\|f\|_{L^2} < \frac{M}{2C_1} , \quad C_1 T^\lambda < \frac{M}{2}$$

we obtain from (35) that Φ takes the ball $B(0, M) \subset X_T$ into itself; if in addition we have also, by possibly taking T smaller,

$$C_2 T^\lambda (2M)^{\gamma-1} < \frac{1}{2}$$

we see by (36) that Φ is a contraction, and in conclusion we obtain the existence of a unique local solution to the subcritical equation. By the same argument as above this solution is in $C([0, T]; L^2(\mathbb{R}^n))$.

We notice that in the above proof the quantity M is only determined by the L^2 norm of f, and hence also the lifespan of the local solution is a function of $\|f\|_{L^2}$. If for some reason we know a priori that the L^2 norm of the solution is conserved by the evolution, this remark implies immediately that the solution is in fact global, by an elementary continuation argument. This is the case for instance of the *gauge invariant* equation

$$iu_t - \Delta u = \pm |u|^{\gamma-1} u$$

for which any solution continuous with values in L^2 satisfies $\|u(t)\|_{L^2} = \|u(0)\|_{L^2}$. The above computations apply without modification and we obtain the global well posedness for all subcritical exponents γ.

Future Directions

The dispersive properties of constant coefficient equations are now fairly well understood. In view of the wide range of applications, it would be very useful to extend the techniques to more general classes of equations, in particular for equations with variable coefficients and on manifolds. These problems have been the subject of intense interest in recent years and are actively pursued.

Strichartz estimates have been extended to very general situations, including the fully variable coefficients case, under suitable smoothness, decay and spectral assumptions on the coefficients (see e. g. [36,37] and, for singular coefficients, [18]).

The results concerning *dispersive estimates* are much less complete. They have been proved for perturbations of order zero

$$iu_t - \Delta u + V(x)u = 0\,, \quad u_{tt} - \Delta u + V(x)u = 0$$

with a potential $V(x)$ sufficiently smooth and decaying at infinity (see e. g. [2,19,28,53,54,55]) while for perturbation of order one

$$iu_t - \Delta u + a(x)\cdot\nabla u_{x_j} + V(x)u = 0\,,$$
$$u_{tt} - \Delta u + a(x)\cdot\nabla u_{x_j} + V(x)u = 0$$

very few results are available, (see [16] for the three-dimensional wave and Dirac equations, and [1,17] for the one-dimensional case where quite general results are available).

A closely connected problem is the study of the dispersive properties of equations on manifolds. This part of the theory is advancing rapidly and will probably be a very interesting direction of research in the next years (see e. g. [12] and related papers for the case of compact manifolds, and [5,24] for noncompact manifolds).

Bibliography

1. Artbazar G, Yajima K (2000) The L^p-continuity of wave operators for one dimensional Schrödinger operators. J Math Sci Univ Tokyo 7(2):221–240
2. Beals M (1994) Optimal L^∞ decay for solutions to the wave equation with a potential. Comm Partial Diff Eq 19(7/8):1319–1369
3. Ben-Artzi M, Trèves F (1994) Uniform estimates for a class of evolution equations. J Funct Anal 120(2):264–299
4. Bergh J, Löfström J (1976) Interpolation spaces. An introduction. Grundlehren der mathematischen Wissenschaften, No 223. Springer, Berlin
5. Bouclet J-M, Tzvetkov N (2006) On global strichartz estimates for non trapping metrics.
6. Bourgain J (1993) Fourier transform restriction phenomena for certain lattice subsets and applications to nonlinear evolution equations. I. Schrödinger equations. Geom Funct Anal 3(2):107–156
7. Bourgain J (1995) Some new estimates on oscillatory integrals. In: (1991) Essays on Fourier analysis in honor of Elias M. Stein. Princeton Math Ser vol. 42. Princeton Univ Press, Princeton, pp 83–112
8. Bourgain J (1995) Estimates for cone multipliers. In: Geometric aspects of functional analysis. Oper Theory Adv Appl, vol 77. Birkhäuser, Basel, pp 41–60
9. Bourgain J (1998) Refinements of Strichartz' inequality and applications to 2D-NLS with critical nonlinearity. Int Math Res Not 5(5):253–283
10. Brenner P (1975) On $L_p - L_{p'}$ estimates for the wave-equation. Math Z 145(3):251–254
11. Brenner P (1977) $L_p - L_{p'}$-estimates for Fourier integral operators related to hyperbolic equations. Math Z 152(3):273–286
12. Burq N, Gérard P, Tzvetkov N (2003) The Cauchy problem for the nonlinear Schrödinger equation on a compact manifold. J Nonlinear Math Phys 10(1):12–27
13. Cazenave T (2003) Semilinear Schrödinger equations. Courant Lecture Notes in Mathematics, vol 10. New York University Courant Institute of Mathematical Sciences, New York
14. Choquet-Bruhat Y (1950) Théorème d'existence pour les équations de la gravitation einsteinienne dans le cas non analytique. CR Acad Sci Paris 230:618–620
15. Christodoulou D, Klainerman S (1989) The nonlinear stability of the Minkowski metric in general relativity. In: Bordeaux (1988) Nonlinear hyperbolic problems. Lecture Notes in Math, vol 1402. Springer, Berlin, pp 128–145
16. D'Ancona P, Fanelli L (2006) Decay estimates for the wave and dirac equations with a magnetic potential. Comm Pure Appl Anal 29:309–323
17. D'Ancona P, Fanelli L (2006) L^p – boundedness of the wave operator for the one dimensional Schrödinger operator. Comm Math Phys 268:415–438
18. D'Ancona P, Fanelli L (2008) Strichartz and smoothing estimates for dispersive equations with magnetic potentials. Comm Partial Diff Eq 33(6):1082–1112
19. D'Ancona P, Pierfelice V (2005) On the wave equation with a large rough potential. J Funct Anal 227(1):30–77
20. D'Ancona P, Georgiev V, Kubo H (2001) Weighted decay estimates for the wave equation. J Diff Eq 177(1):146–208
21. Foschi D (2005) Inhomogeneous Strichartz estimates. J Hyperbolic Diff Eq 2(1):1–24
22. Ginibre J, Velo G (1985) The global Cauchy problem for the nonlinear Schrödinger equation revisited. Ann Inst Poincaré H Anal Non Linéaire 2(4):309–327
23. Ginibre J, Velo G (1995) Generalized Strichartz inequalities for the wave equation. J Funct Anal 133(1):50–68
24. Hassell A, Tao T, Wunsch J (2006) Sharp Strichartz estimates on nontrapping asymptotically conic manifolds. Am J Math 128(4):963–1024
25. Hörmander L (1997) Lectures on nonlinear hyperbolic differential equations. Mathématiques & Applications (Mathematics & Applications), vol 26. Springer, Berlin
26. John F (1979) Blow-up of solutions of nonlinear wave equations in three space dimensions. Manuscr Math 28(1–3):235–268
27. John F, Klainerman S (1984) Almost global existence to nonlinear wave equations in three space dimensions. Comm Pure Appl Math 37(4):443–455
28. Journé J-L, Soffer A, Sogge CD (1991) Decay estimates for Schrödinger operators. Comm Pure Appl Math 44(5):573–604

29. Kato T (1965/1966) Wave operators and similarity for some non-selfadjoint operators. Math Ann 162:258–279
30. Keel M, Tao T (1998) Endpoint Strichartz estimates. Am J Math 120(5):955–980
31. Klainerman S (1980) Global existence for nonlinear wave equations. Comm Pure Appl Math 33(1):43–101
32. Klainerman S (1981) Classical solutions to nonlinear wave equations and nonlinear scattering. In: Trends in applications of pure mathematics to mechanics, vol III. Monographs Stud Math, vol 11. Pitman, Boston, pp 155–162
33. Klainerman S (1982) Long-time behavior of solutions to nonlinear evolution equations. Arch Rat Mech Anal 78(1):73–98
34. Klainerman S (1985) Long time behaviour of solutions to nonlinear wave equations. In: Nonlinear variational problems. Res Notes in Math, vol 127. Pitman, Boston, pp 65–72
35. Klainerman S, Nicolò F (1999) On local and global aspects of the Cauchy problem in general relativity. Class Quantum Gravity 16(8):R73–R157
36. Marzuola J, Metcalfe J, Tataru D. Strichartz estimates and local smoothing estimates for asymptotically flat Schrödinger equations. To appear on J Funct Anal
37. Metcalfe J, Tataru D. Global parametrices and dispersive estimates for variable coefficient wave equation. Preprint
38. Pecher H (1974) Die Existenz regulärer Lösungen für Cauchy- und Anfangs-Randwert-probleme nichtlinearer Wellengleichungen. Math Z 140:263–279 (in German)
39. Reed M, Simon B (1978) Methods of modern mathematical physics IV. In: Analysis of operators. Academic Press (Harcourt Brace Jovanovich), New York
40. Segal I (1968) Dispersion for non-linear relativistic equations. II. Ann Sci École Norm Sup 4(1):459–497
41. Shatah J (1982) Global existence of small solutions to nonlinear evolution equations. J Diff Eq 46(3):409–425
42. Shatah J (1985) Normal forms and quadratic nonlinear Klein–Gordon equations. Comm Pure Appl Math 38(5):685–696
43. Shatah J, Struwe M (1998) Geometric wave equations. In: Courant Lecture Notes in Mathematics, vol 2. New York University Courant Institute of Mathematical Sciences, New York
44. Strichartz RS (1970) Convolutions with kernels having singularities on a sphere. Trans Am Math Soc 148:461–471
45. Strichartz RS (1970) A priori estimates for the wave equation and some applications. J Funct Anal 5:218–235
46. Strichartz RS (1977) Restrictions of Fourier transforms to quadratic surfaces and decay of solutions of wave equations. J Duke Math 44(3):705–714
47. Tao T (2000) Spherically averaged endpoint Strichartz estimates for the two-dimensional Schrödinger equation. Comm Part Diff Eq 25(7–8):1471–1485
48. Tao T (2003) Local well-posedness of the Yang–Mills equation in the temporal gauge below the energy norm. J Diff Eq 189(2):366–382
49. Taylor ME (1991) Pseudodifferential operators and nonlinear PDE. Progr Math 100. Birkhäuser, Boston
50. von Wahl W (1970) Über die klassische Lösbarkeit des Cauchy-Problems für nichtlineare Wellengleichungen bei kleinen Anfangswerten und das asymptotische Verhalten der Lösungen. Math Z 114:281–299 (in German)
51. Wolff T (2001) A sharp bilinear cone restriction estimate. Ann Math (2) 153(3):661–698
52. Yajima K (1987) Existence of solutions for Schrödinger evolution equations. Comm Math Phys 110(3):415–426
53. Yajima K (1995) The $W^{k,p}$-continuity of wave operators for Schrödinger operators. J Math Soc Japan 47(3):551–581
54. Yajima K (1995) The $W^{k,p}$-continuity of wave operators for schrödinger operators. III. even-dimensional cases $m \geq 4$. J Math Sci Univ Tokyo 2(2):311–346
55. Yajima K (1999) L^p-boundedness of wave operators for two-dimensional schrödinger operators. Comm Math Phys 208(1):125–152

Distributed Controls of Multiple Robotic Systems, An Optimization Approach

John T. Feddema, David A. Schoenwald, Rush D. Robinett, Raymond H. Byrne
Sandia National Laboratories, Albuquerque, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Cooperative Control
Communication Effects
Conclusions
Future Directions
Acknowledgments
Bibliography

Glossary

Multiple robotic system Multiple robotic systems are a collective of multiple robotic vehicles (e. g. ground-based, aerial, underwater, and/or some combination thereof) wherein each robot has assigned tasks designed to help the entire system achieve its overall goals.

Provably convergent cooperative controls Provably convergent cooperative controls is a distribution of control laws among the agents of a multiple robotic system with the objective that global stability is preserved and a global performance index is optimized. The control laws take into account the physics of the individual robots, the communication protocols and information exchange between the robots, the overall terrain, and the goals/constraints of the entire collective.

Connective stability A system is connectively stable if it is stable in the sense of Lyapunov for all structural perturbations of the interconnection matrix that describes the system (see [63]). These structural perturbations

include attenuation, degradation, and loss of communication between nodes in the network.

Vehicle communication protocols In a multiple robotic system, vehicle communication protocols refer to the type of communication algorithm used by the vehicles to share information amongst each other. These protocols are crucial in determining stability and performance properties.

Time division multiple access (TDMA) TDMA is a channel access method for shared medium (usually radio) networks. It allows several users to share the same frequency channel by dividing the signal into timeslots. The users transmit in rapid succession, one after the other, each using their own timeslot. This allows multiple stations (e. g. robots) to share the same transmission medium (e. g. radio frequency channel) while using only part of its bandwidth.

Carrier sense multiple access (CSMA) CSMA is also a channel access method for shared medium networks where the station first listens to the channel for a predetermined amount of time so as to check for any activity on the channel. If the channel is sensed "idle" then the station is permitted to transmit. If the channel is sensed as "busy" the station defers its transmission. This access method is used in wired TCP/IP networks.

Linear TDMA broadcast The number of time slots in the TDMA network is equal to the total number of nodes in the network.

Polylogarithmic TDMA broadcast The number of time slots in the TDMA network is less than the total number of nodes because of spatial reuse. This method only applies as long as the degree of the network (the number of nodes within communication range) is small (see [66]).

Definition of the Subject

This chapter describes an integrated approach for designing communication, sensing, and control systems for mobile distributed systems. A three-step process is used to create locally optimal distributed controls for multiple robot vehicles. The first step is to define a global performance index whose extremum produces a desired cooperative result. The second step is to partition and eliminate terms in the performance index so that only terms of local neighbors are included. This step minimizes communication amongst robots and improves system robustness. The third step is to define a control law that is the gradient of the partitioned performance index. This control drives the system towards the local extremum of the partitioned performance index. Graph theoretic methods are then used to analyze the input/output reachability and structural controllability and observability of the decentralized system. Connective stability of the resulting controls is evaluated with a vector Lyapunov technique. From this analysis, the trade-off between vehicle responsiveness and the maximum allowable communication delay is theoretically determined. The approach has been tested in simulation and robotics hardware on a number of test problems. Six examples are given. The first problem is the distribution of multiple robots along a line or curve. The second and third problems are the distribution of multiple robots in an unconstrained and constrained two-dimensional plane. The fourth problem is the distribution of multiple robots around an elliptical curve. The fifth problem is for multiple robots to converge on the source of a plume in a plane. The sixth problem is for multiple underwater robots to converge on the source of a plume in a 3D volume. Four different communication protocols are compared for minimum communication delay: a Time Division Multiple Access (TDMA) linear broadcast, a TDMA polylogarithmic broadcast, a TDMA coloring algorithm, and a Collision Sense Multiple Access (CSMA) algorithm.

Introduction

Over the past two decades, considerable research has concentrated on the development of cooperative controls of multiple robotic systems [1]. Much of this work is originally inspired by the observation that social insects such as termites, ants, and bees perform amazing group behaviors when individually executing simple rules [2]. The hypothesis being that it must be possible to create useful "emergent" group behaviors from simple individual behaviors. While this hypothesis has yet to be proven or disproved, many people have demonstrated some interesting group behaviors. For example, researchers have shown that groups of robots can forage for objects [3], maintain formations [4], and push boxes [5].

As the field of cooperative robotics has matured, more complex cooperative behaviors have been developed to clean up hazardous waste [6], explore and map unknown environments [7,8], uniformly cover a specified area with a mobile sensor network [9], track multiple targets [10], and perform cooperative pursuit-evasion maneuvers [11,12,13,14]. Algorithms for resource allocation of heterogeneous agents has been developed [6,15], as well as distributed fault detection methods for alerting operators when individual robots are not moving as expected [16]. Measures of effectiveness for human-robot

teams have been defined [17] to evaluate the mission performance of the cooperative algorithms.

In an effort to create cooperative behaviors that meet a performance specification, researchers have begun to take a system controls perspective and analyze the stability of multiple vehicles when driving in formations. Chen and Luh [18] examined decentralized control laws that drove a set of holonomic mobile robots into a circular formation. A conservative stability requirement for the sample period is given in terms of the damping ratio and the undamped natural frequency of the system. Similarly, Yamaguchi studied line-formations [19] and general formations [20] of nonholonomic vehicles, as did Yoshida et al. [21]. Decentralized control laws using a potential field approach to guide vehicles away from obstacles can be found in [22,23]. In these studies, only continuous time analyses have been performed, assuming that the relative position between vehicles and obstacles can be measured at all time.

In the area of line formations, Stankovac [24] considered the problem of controlling a one dimensional platoon of N identical vehicles moving at a constant velocity and a fixed distance apart. The measured distance and relative velocity of each neighboring vehicle is used to regulate the formation. Seiler [25] analyzed the "string instability" problem where a disturbance in the inter-vehicular spacing is amplified as nearest neighbor feedback propagates through the leader-follower platoon. Barooah [26] showed the system performance can be improved by "mistuning" the individual controller gains.

A variety of control techniques have been used to control general formations. Orgen [27] shows that multiagent coordination is possible if a vector Lyapunov function exists for the formation that is a weighted sum of Lyapunov functions for each individual robot. This same approach is used in this chapter to guarantee convergence and stability. In [28], complex maneuvers are decomposed into a sequence of maneuvers between formation patterns. The formation patterns are arranged in a bidirectional ring topology and their inter-robot distances are controlled with via passivity techniques. In [29], leader-to-formation stability gains control error amplification, relate interconnection topology to stability and performance, and offer safety bounds for different formation topologies. In [30], algebraic conditions are developed that guarantee formation feasibility given individual agent kinematics. Given the kinematics of several agents along with inter-agent constraints, [30] determine whether there exist nontrivial agent trajectories that maintain the constraints. In [31], task functions describe the desired kinematic relationship between vehicle positions and suitable variables that synthetically express some feature of the platoon shape. Multiple task functions are merged using a singularity-robust task-priority inverse kinematics algorithm. The task function approach is similar to the optimization approach followed in this chapter.

Another way of analyzing stability is to investigate the convergence of a distributed algorithm. Beni and Liang [32] prove the convergence of a linear swarm of asynchronous distributed autonomous agents into a synchronously achievable configuration. The linear swarm is modeled as a set of linear equations that are solved iteratively. Their formulation is best applied to resource allocation problems that can be described by linear equations. Liu et al. [33] provide conditions for convergence of an asynchronous swarm in which swarm "cohesiveness" is the stability property under study. Their paper assumes position information is passed between nearest neighbors only and proximity sensors prevent collisions.

Also of importance is the recent research combining graph theory with decentralized controls. Most cooperative mobile robot vehicles have wireless communication, and simulations have shown that a wireless network of mobile robots can be modeled as an undirected graph [34]. These same graphs can be used to control a formation. Desai et al. [35,36] used directed graph theory to control a team of robots navigating terrain with obstacles while maintaining a desired formation and changing formations when needed. When changing formations, the transition matrix between the current adjacency matrix and all possible control graphs are evaluated.

Graph theory can also be used to maintain connectedness during maneuvers. Connectedness is the problem of ensuring that a group of mobile agents stays connected in a communication sense while achieving some performance objective. Ji [37] use a weighted graph Laplacian to generate nonlinear feedback laws that guarantee connectedness during rendezvous and formation-control maneuvers. In [38], the connectivity condition is translated to differentiable constraints on individual agent motion by considering the dynamics of the Laplacian matrix and its spectral properties. Artificial potential fields are used to drive the agents to configurations away from the undesired space of disconnected networks while avoiding collisions with each other.

The ability to scale up to very large sized systems has also been recently investigated. Belta [39] addressed the general problem of controlling a large group of robots required to move as a group. The group is abstracted to a lower dimensional manifold whose dimension is independent of the number of robots. The manifold captures the position and orientation of the ensemble and the shape

spanned by the robots. Decoupled controllers are designed for the group and shape variables, and individual controllers for each robot are designed to keep them within the manifold. In [40], the scalability is improved by representing formation structures by queues, which are smooth line segments consisting of multiple robots. Artificial potential trenches are used to control the formation of the queues. In [41], the complexity of population modeling is avoided by modeling the agent distribution as state probability density functions. The control problem becomes one of maximizing the probability of robotic presence in a given region. In [42], the large dimensional state space is reduced into a small dimensional continuous abstract space which captures essential features of the swarm. High level specifications of the continuous abstraction, given as linear temporal logic formulas over linear predicates, are automatically mapped to provably correct robot control laws.

Lastly, cooperative algorithms for mobile manipulation are also being developed. Tanner et al. [43] presented a motion planning methodology based on nonsmooth Lyapunov functions that account for nonpoint articulated robots. Diffeomorphic transformations are used to map the volume of each articulated robot and obstacle into a point world and a dipolar potential field function is used to ensure asymptotic convergence of the robots. Yamashita [44] developed motion planning algorithms for controlling multiple mobile robots cooperatively transporting a large object in a three dimensional environment. The transportation portion of the problem is solved in a reduced configuration space using a potential field path planner. An A* algorithm is used to avoid local minimum caused by the constraints of object manipulation. Bonaventura [45] modeled the dynamics of complex multiple mobile manipulators constrained by concurrent contacts, including time-varying and nonholonomic constraints.

Over the past several years, the authors have been developing cooperative control systems for military missions such as perimeter surveillance [46], facility reconnaissance [47], and chemical plume tracking [48,49,50]. Multiple mobile robots perform these missions in a coordinated fashion using low bandwidth communication between nodes. We have demonstrated that by using Provably Convergent Cooperative Controls (PC^3) we can guarantee a specified level of performance even under uncertainty. PC^3 involves selecting an overall performance index and using distributed optimization techniques to minimize this performance index. The stability of the control can be proven using a vector Lyapunov function. The scheme is inherently fault tolerant to nodal failure, allowing other nodes to complete a task started by a failed node.

Key to the success of PC^3 is that the communication between nodes is used to synchronize the controls. In most of our work, we have used a TDMA protocol with a small number of nodes (up to 10). To expand this PC^3 capability in larger scaled systems of 100 to 1000 nodes, we must understand the underlying communication layer and ensure that the communication delays are bounded. Therefore, an important aspect of this work is an integrated approach to both communication and controls. Given the advances being made on the bandwidth of networked systems, this is an exciting new area of research, and the interested reader may want to peruse [51,52].

PC^3 uses graph theoretic methods to analyze the input/output reachability and structural controllability and observability of a decentralized system. This analysis could be embedded in each node and be used to automatically reconfigure an ad hoc communication network for the control task at hand. The graph analysis could also be used to create the most efficient communication flow control based upon spatial distribution of the network nodes. Edge coloring algorithms in graph theory tell us that the minimum number of time slots in a planar network is equal to the maximum number of adjacent nodes plus some small number. The more spread out the nodes are, the fewer number of time slots are needed for communication, and the smaller the latency between nodes. In a coupled system, smaller latency results in a more responsive control system. In this chapter, network protocols that propagate this information and distributed algorithms that automatically adjust the number of time slots available for communication were evaluated. For real-time implementation, these protocols and algorithms must be extremely efficient and only updated as network nodes move.

The next section provides a common mathematical framework that can be used to develop cooperative behaviors among mobile robot vehicles. Six examples of the application of this approach are given. The proceeding section describes the effects of communication on the sampling period of a cooperative system. The tradeoff between communication delay, interaction gain, and vehicle responsiveness is evaluated using a vector Lyapunov technique. Four different communication protocols: a Time Division Multiple Access (TDMA) linear broadcast, a TDMA polylogarithmic broadcast, a TDMA coloring algorithm, and a Collision Sense Multiple Access (CSMA) algorithm are compared. The final section summarizes this work and lists areas of suggested future research.

Cooperative Control

In this section, a common mathematical framework is used to describe several cooperative behaviors. This mathematical framework is applied to three generic behaviors: containment, coverage, and converging. The authors believe that this same framework could be applied to the other behaviors, although this has not been proven at the time of this publication.

This mathematic framework is motivated by the fact that most of the laws in physics and mechanics can be derived by finding a stationary point [53] that may be a maximum or minimum of some performance index, in this case, called the Lagrangian integral [54]. In physics, the Lagrangian integral is a function of energy, and the goal is often to find the action that minimizes the total energy or action integral. It should be noted that the Lagrangian of each phenomenon is a function of some gradient term squared. In the analysis that follows, the reader will notice that each behavior's performance index is also a function of some gradient term squared.

Following this same optimization approach, a three-step process for developing cooperative control algorithms has been developed [55]. These three steps are as follows:

1. Define a global performance index as a function of parameters from all entities.
2. Partition and eliminate terms in the performance index so that only terms of local neighbors are included.
3. The local control law is the gradient (or Newton's method) of the partitioned performance index.

The first step requires that one understands the problem well enough that it can be posed as a global optimization problem. This step can be relatively difficult, but as the examples in the remainder of this section will show, with the right simplifying assumptions, rather simple equations can be used to generate complex cooperative behaviors.

The second step, partitioning the performance index, is often used in parallel optimization to reduce the computation time for large-scale problems [56]. In this case, the second step is used to reduce communications between robots and to increase robustness of the distributed system. The control law that would result from step 1 would require that every robot be able to communicate with all the other robots. As the number of robots increase to 100 s and 1000 s, the time delay necessary for communication would make the resulting control infeasible. Instead, partitioning the performance index and eliminating terms that include interaction between distant neighbors results in a control law that only requires communication with nearest neighbors, thus greatly reducing communication delay. Also, using nearest neighbors that change throughout the motion adds an element of robustness. The mathematical formulation of the partition does not specify that robot number 10 must communicate with robot number 6. Instead, the mathematical formulation specifies a group of nearest neighbors that can change based on external forces and environmental conditions. This creates an element of self-organization that allows the system to change and evolve. If a robot fails, a new set of nearest neighbors is formed.

The third step is to solve for the extremum of the partitioned performance index using either a first-order steepest descent or second order method such as the Newton's Method [57].

The remainder of this section shows how these three steps have been used in practice. Six examples are given with details on the problem formulation and the task that was performed.

Example 1: Spreading Apart Along a Line – A Containment Behavior

This first example is a simple one-dimensional problem. The goal is for multiple robots to evenly spread apart along a straight line using only information from the neighboring robots on the right and left. In Fig. 1, the first and last robots are assumed to be stationary while the ones in between are to spread apart a distance d away from each other.

The optimization steps are as follows.

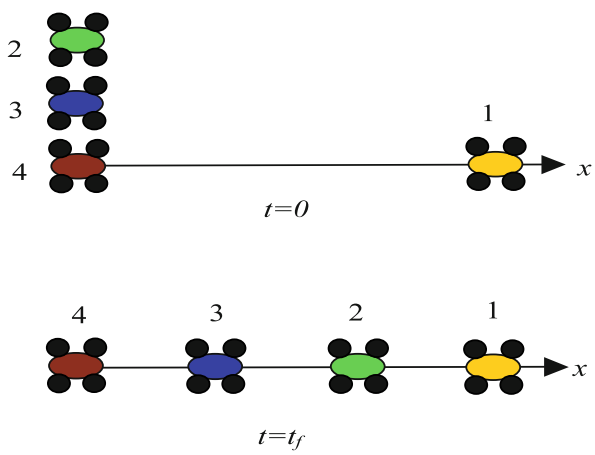

Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 1
One-dimensional control problem. The *top line* is the initial state. The *second line* is the desired final state. Vehicles 1 and 4 are boundary conditions. Vehicles 2 and 3 spread out along the line using only the position of their left and right neighbors

Step 1. Specified as an optimization problem, the objective is to

$$\min_{\bar{x}} v(\bar{x}) \tag{1}$$

where the global performance index is

$$v(\bar{x}) = \frac{1}{2}\sum_{i=1}^{N-1}(d - |x_{i+1} - x_i|)^2 , \tag{2}$$

x_i is the position of robot i, $\bar{x} = [x_1 \ldots x_N]^{\mathrm{T}}$ are the positions of all the robots, and d is the desired distance between each robot. The goal is to minimize the sum of squared errors in distances between every robot.

Step 2. This problem is easily partitioned amongst the interior $N-2$ robots. The distributed objective is to

$$\min_{x_i} v_i(\bar{x}) \quad \forall i = 2, \ldots, N-1 \tag{3}$$

where the partitioned performance index is

$$v_i(\bar{x}) = \frac{1}{2}(d - |x_i - x_{i-1}|)^2 + \frac{1}{2}(d - |x_i - x_{i+1}|)^2 . \tag{4}$$

Because of the additive form of Eq. (2), simultaneously solving Eq. (3) for each robot is the same as minimizing the global performance index in Eq. (2). Therefore, in this case, no terms were eliminated. This is not necessarily true for the other example problems below.

Step 3. A steepest descent control law for the partitioned performance index is given by

$$x_i(k+1) = x_i(k) - \alpha \nabla v_i(\bar{x}(k)) , \quad 0 < \alpha \leq 1 , \tag{5}$$

where

$$\nabla v_i(\bar{x}) = 2x_i - (x_{i+1} + x_{i-1}) \quad \text{if } x_{i-1} < x_i < x_{i+1}. \tag{6}$$

Note that $\nabla v_i(\bar{x}) = 0$ when $x_i = \frac{1}{2}(x_{i+1} + x_{i-1})$. Therefore, the vehicles will disperse along the line until they have reached a position that is exactly in the middle of its nearest neighbors. In [47], it is shown that α is actually more constrained than $0 < \alpha \leq 1$ depending on the responsiveness of the vehicle and the communication sample period.

The control law in Eqs. (5)–(6) have been used to spread robot vehicles apart along a perimeter [46] as shown in Fig. 2, as well as to spread out hopping minefield robots [58] as shown in Fig. 3.

Example 2: Coverage of a Two-Dimensional Space

Next, we consider the example of dispersing robots in a plane in a specified pattern. In Fig. 4, the robots are to move from the configuration on the left to the configuration on the right. The configuration on the right is specified by the distances d_{ij} between robot i and robot j.

Step 1. The objective is to

$$\min_{\bar{x}} v(\bar{x}) \tag{7}$$

where the global performance index is

$$v(\bar{x}) = \frac{1}{2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(d_{ij}^2 - (x_i - x_j)^2 - (y_i - y_j)^2\right)^2 , \tag{8}$$

$$\bar{x}_i = \begin{bmatrix} x_i \\ y_i \end{bmatrix} \in \Re^2$$

Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 2
***Left*: Four robot vehicles are shown guarding a perimeter denoted by the blue line segments. When an intrusion detection sensor denoted by the numbered circles alarms, one robot vehicle attends to the alarm (vehicle near sensor 33) while the others spread apart along the perimeter so that each vehicle is midway between its neighbors. *Right*: Robot vehicles used to perform perimeter surveillance task**

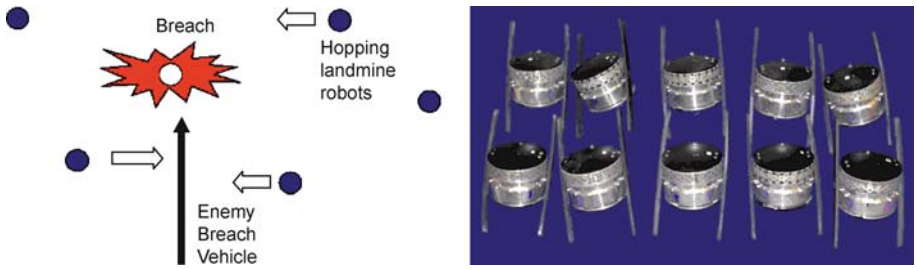

Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 3
***Left*: Hopping landmine robots are filling breach left by enemy vehicle. When a robot is breached, the robots will hop towards the missing robot and settle when each robot is midway between its neighbors. *Right*: Hopping landmine robots used in self-healing minefield tests**

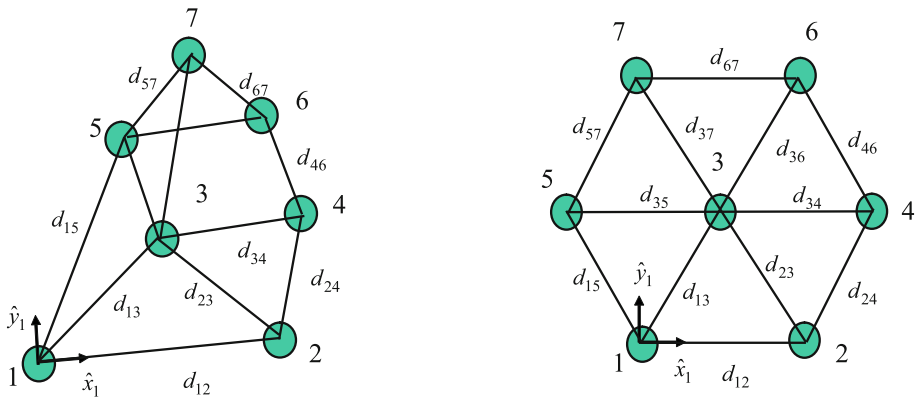

Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 4
***Left*: Initial configuration of robots. *Right*: Desired final configuration**

is the position of robot i in the xy plane, and $\bar{x} = [\bar{x}_1^{\mathrm{T}} \ldots \bar{x}_N^{\mathrm{T}}]^{\mathrm{T}}$ is the position of all the robots in the xy plane. By minimizing the error between the squared desired distance and the squared measured distance between every pair-wise combination of robots, we can drive the robots from an initial pattern to the desired specified pattern. Notice that the global performance index does not specify the orientation or final absolute position of the group of robots.

Step 2. The global performance index is over constrained since it is possible to achieve the same minimum solution without having to minimize the error between every pair-wise combination. The same minimum solution can be achieved by only minimizing the error between neighboring robots. The distributed objective is to

$$\min_{\bar{x}_i} v_i(\bar{x}) \quad \forall i = 1, \ldots, N \tag{9}$$

where the partitioned performance index is

$$v_i(\bar{x}) = \frac{1}{2} \sum_{j \in \mathrm{NN}} \left(d_{ij}^2 - (x_i - x_j)^2 - (y_i - y_j)^2 \right)^2 \tag{10}$$

and NN stands for nearest neighbor.

Step 3. The steepest descent control law for the partitioned performance index is given by

$$\bar{x}_i(k+1) = \bar{x}_i(k) - \alpha \nabla v_i(\bar{x}(k)) \ , \quad 0 < \alpha \leq 1 \tag{11}$$

where

$$\nabla v_i = \begin{bmatrix} \frac{\partial v_i}{\partial x_i} \\ \frac{\partial v_i}{\partial y_i} \end{bmatrix} \in \Re^2 \ , \tag{12}$$

$$\frac{\partial v_i(\bar{x})}{\partial x_i} = -2 \sum_{j \in \mathrm{NN}} \left[d_{ij}^2 - (x_i - x_j)^2 - (y_i - y_j)^2 \right] (x_i - x_j) \ , \tag{13}$$

$$\frac{\partial v_i(\bar{x})}{\partial y_i} = -2 \sum_{j \in \mathrm{NN}} \left[d_{ij}^2 - (x_i - x_j)^2 - (y_i - y_j)^2 \right] (y_i - y_j) \ . \tag{14}$$

Note that $\nabla v_i = 0$ when $d_{ij}^2 = (x_i - x_j)^2 + (y_i - y_j)^2$

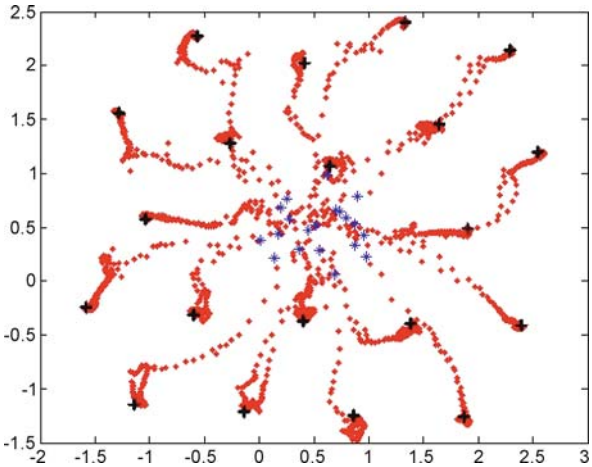

Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 5
Plot of 20 vehicles' trajectories started from a clustered position with the goal of spreading out uniformly through the space (*blue* ✳ indicates initial position, *red marks* indicate trajectory, and *black* + indicate final position)

for $j \in$ NN. In [59], the connective stability of this control law is proven using a vector Lyapunov technique. The control law in Eqs. (11)–(14) has been used to spread apart the hopping minefield robots as shown in Fig. 5. In this case, the specified distances d_{ij} are all equal and the number of nearest neighbors used for control is three.

Example 3: Coverage of a Two-Dimensional Space with Constraints

Next, consider the same problem as in the previous example, except that the robots are constrained to stay within a region that is bounded by line segments as shown in Fig. 6.

Step 1. The objective is to

$$\min_{\bar{x}} v(\bar{x}) \tag{15}$$

where the global performance index is

$$v(\bar{x}) = \frac{1}{2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(d_{ij}^2 - |\bar{x}_i - \bar{x}_j|^2\right)^2 \tag{16}$$

subject to

$$A\bar{x}_i \le b \quad \forall i = 1, \ldots, N \tag{17}$$

where $A \in \Re^{m\times 2}$ and $b \in \Re^m$. Equation (17) specifies the boundary conditions of m straight-line segments.

Step 2. The distributed objective is to

$$\min_{\bar{x}_i} v_i(\bar{x}) \quad \forall i = 1, \ldots, N \tag{18}$$

where the partitioned performance index is

$$v_i(\bar{x}) = \frac{1}{2}\sum_{j\in \mathrm{NN}}\left(d^2 - |\bar{x}_i - \bar{x}_j|^2\right)^2 + \frac{1}{2}\Lambda\sum_{l\in \mathrm{NO}}(A_l\bar{x}_i - b_l)^{-2}. \tag{19}$$

Here, the inequality constraints in Eq. (17) have been added as a weighted penalty function that is the sum of the inverse squared perpendicular distances between robot i and the nearest obstacle (NO) line segments l. The Λ is a scalar used to vary the importance of obstacle avoidance. As before, NN stands for the set of nearest neighbors. Similarly, the set NO is the set of nearest obstacles.

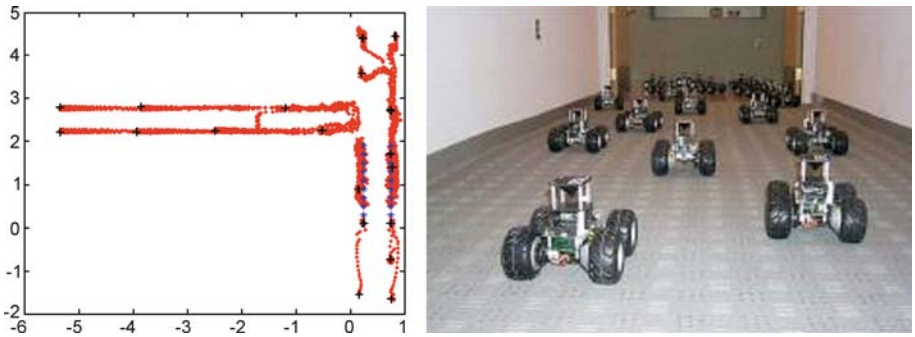

Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 6
***Left*: Plot of 20 vehicles' trajectories started from a clustered position with the goal of spreading apart uniformly through a hallway with a side corridor (*blue* ✳ indicates initial position, *red marks* indicate trajectory, and *black* + indicate final position). *Right*: Robot vehicles used in an indoor communication/navigation network**

Step 3. The steepest descent control law is

$$\bar{x}_i\,(k+1) = \bar{x}_i\,(k) - \alpha\,\nabla v_i\,(\bar{x}(k))\,, \quad 0 < \alpha \le 1 \quad (20)$$

where

$$\nabla v_i(\bar{x}) = -2 \sum_{j\in \mathrm{NN}} (d^2 - |\bar{x}_i - \bar{x}_j|^2)(\bar{x}_i - \bar{x}_j) - \Lambda \sum_{l\in \mathrm{NO}} A_l^{\mathrm{T}}(A_l\bar{x}_i - b_l)^{-3}\,. \quad (21)$$

The control law in Eqs. (20) and (21) has been used to spread out the robot vehicles in a hallway as shown in Fig. 6. The nearest obstacles are determined from IR proximity sensors. The specified distance d between vehicles was chosen to be within the 10 meter acoustic range of the sensors on top of the vehicle (see Fig. 6) [60]. Again, the number of nearest neighbors used for control is three.

Example 4. Forming an Ellipse with Constraints – A Containment Behavior

Next, consider a path following/formation problem where multiple vehicles are to 1) travel towards and spread apart on an ellipse, 2) not drive into each other, and 3) stay away from obstacle line segments. This is shown in Fig. 7.

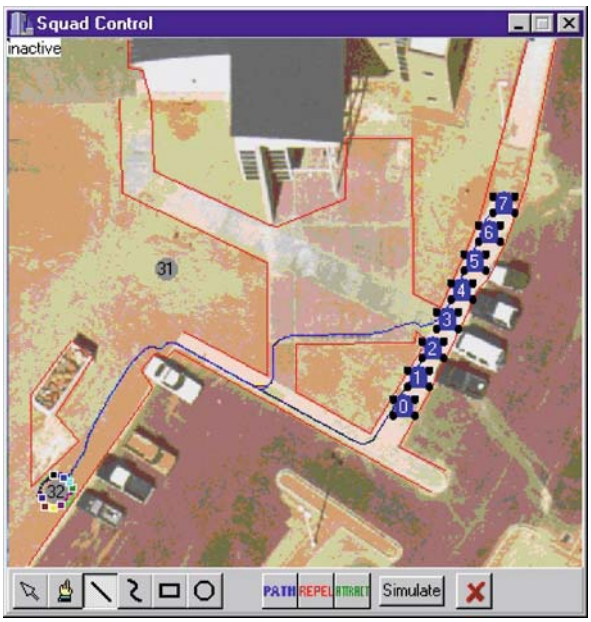

Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 7
Robot path planner drives vehicles towards ellipse while staying away from obstacles, denoted by *red line*, and the other vehicles

Step 1. The objective is to

$$\min_{\bar{x}} v(\bar{x}) \quad (22)$$

where the global performance index is

$$v(\bar{x}) = \frac{1}{2}\sum_{i=1}^{N}\left((\bar{x}_i - \bar{x}_o)^{\mathrm{T}}\begin{bmatrix} 1/\rho^2 & 0 \\ 0 & 1/\sigma^2 \end{bmatrix}(\bar{x}_i - \bar{x}_o) - 1\right)^2 \quad (23)$$

subject to

$$|\bar{x}_{i+1} - \bar{x}_i| > d \quad \forall\, i = 1,\ldots,N-1$$

$$A\bar{x}_i \le b \quad \forall\, i = 1,\ldots,N\,. \quad (24)$$

The global performance index is the squared error of the robot's position from the ellipse. The position of the center of the ellipse is $\bar{x}_o$, and ρ and σ are the elliptical parameters along the x and y-axes. The first constraint ensures that the vehicles stay a distance d apart from each other. The second constraint ensures that the vehicles stay away from the line constraints as in the previous example.

Step 2. The distributed objective is

$$\min_{\bar{x}_i} v_i(\bar{x}) \quad \forall\, i = 1,\ldots,N \quad (25)$$

where the partitioned performance index is

$$v_i(\bar{x}) = \frac{1}{2}\left((\bar{x}_i - \bar{x}_o)^{\mathrm{T}}\begin{bmatrix} 1/\rho^2 & 0 \\ 0 & 1/\sigma^2 \end{bmatrix}(\bar{x}_i - \bar{x}_o) - 1\right)^2 + \frac{1}{2}\sum_{j\in \mathrm{NN}}\left(d^2 - |\bar{x}_i - \bar{x}_j|^2\right)^2 + \frac{1}{2}\Lambda\sum_{l\in \mathrm{NO}}(A_l\bar{x}_i - b_l)^{-2}\,. \quad (26)$$

The two constraints are implemented as penalty functions. The equations are the same as in the previous example.

Step 3. The steepest descent control law is

$$\bar{x}_i\,(k+1) = \bar{x}_i\,(k) - \alpha\,\nabla v_i\,(\bar{x}(k)) \quad (27)$$

where

$$\nabla v_i(\bar{x}) = 2\left((\bar{x}_i - \bar{x}_o)^{\mathrm{T}}\begin{bmatrix} 1/\rho^2 & 0 \\ 0 & 1/\sigma^2 \end{bmatrix}(\bar{x}_i - \bar{x}_o) - 1\right) \cdot \begin{bmatrix} 1/\rho^2 & 0 \\ 0 & 1/\sigma^2 \end{bmatrix}(\bar{x}_i - \bar{x}_o) - 2\sum_{j\in \mathrm{NN}}\left(d^2 - |\bar{x}_i - \bar{x}_j|^2\right)(\bar{x}_i - \bar{x}_j) - \Lambda\sum_{l\in \mathrm{NO}} A_l^{\mathrm{T}}(A_l\bar{x}_i - b_l)^{-3}\,. \quad (28)$$

The control law in Eqs. (27)–(28) has been implemented on a path planner as shown in Fig. 7. The number of nearest neighbors and number of nearest obstacles can be one if the time step is small. The nearest neighbor and obstacle will continually change throughout the motion.

Example 5. Converging on the Source of a Plume – 2D Case

The next example is a plume localization problem. The objective is for multiple vehicles to locate and converge on a source, which could either be acoustic, radio frequency, temperature, or chemical (see Fig. 8). It is assumed that the spatial signature of the source can be approximated by a quadratic surface. The form of this second order equation allows us to easily formulate a convergent control to the extremum of the surface. If the data were fit to a higher order surface with many local extremum, then it would not be possible to guarantee convergence to a single solution.

Step 1. The objective is

$$\max_{\bar{x}} v(\bar{x}) \tag{29}$$

where the global performance index is

$$v(\bar{x}) \cong \sum_{i=1}^{N} a_0 + A_1^{\mathrm{T}}(\bar{x}_i - \bar{x}_0) + \frac{1}{2}(\bar{x}_i - \bar{x}_0)^{\mathrm{T}} A_2 (\bar{x}_i - \bar{x}_0) . \tag{30}$$

The parameters of the quadratic surface are $a_0 \in \Re$, $A_1 \in \Re^2$, and $A_2 \in \Re^{2x2}$. The center of the source is located at $\bar{x}_0 \in \Re^2$.

Step 2. The distributed objective is

$$\max_{\bar{x}_i} v_i(\bar{x}) \quad \forall i = 1, \ldots, N \tag{31}$$

where the partitioned performance index is

$$v_i(\bar{x}) \cong \sum_{j \in \mathrm{NN}} a_{0i} + A_{1i}^{\mathrm{T}}(\bar{x}_j - \bar{x}_i) + \frac{1}{2}(\bar{x}_j - \bar{x}_i)^{\mathrm{T}} A_{2i}(\bar{x}_j - \bar{x}_i) . \tag{32}$$

Each vehicle determines its own estimate of the quadratic surface using information from its nearest neighbors. An alternative approach is to use data from as many neighbors as possible and calculate a least-squares estimate of the quadratic coefficients. References [48,49,50] and [61] describe the least squares fitting algorithm in more detail.

Step 3. The second order Newton's method control law is

$$\bar{x}_i(k+1) = \bar{x}_i(k) - \alpha \; A_{2i}^{-1}\big|_{\bar{x}(k)} A_{1i}\big|_{\bar{x}(k)} \tag{33}$$

where the quadratic coefficients are determined from the solution to the nearest neighbor equations

$$v_i(\bar{x}_j) = a_{0i} + A_{1i}^{\mathrm{T}}(\bar{x}_j - \bar{x}_i) + \frac{1}{2}(\bar{x}_j - \bar{x}_i)^{\mathrm{T}} A_{2i}^{\mathrm{T}}(\bar{x}_j - \bar{x}_i) \quad \forall j \in \mathrm{NN} . \tag{34}$$

The control law in Eqs. (33)–(34) has been implemented on RATLER vehicles (see Fig. 2) that locate an acoustic source and on a set of miniature robotic vehicles (see Fig. 8) that locate a block of dry ice [48,49,50,61]. In both cases, the number of nearest neighbors is six because seven measurements (including itself) are needed to uniquely determine the quadric coefficients A_{1i} and A_{2i}.

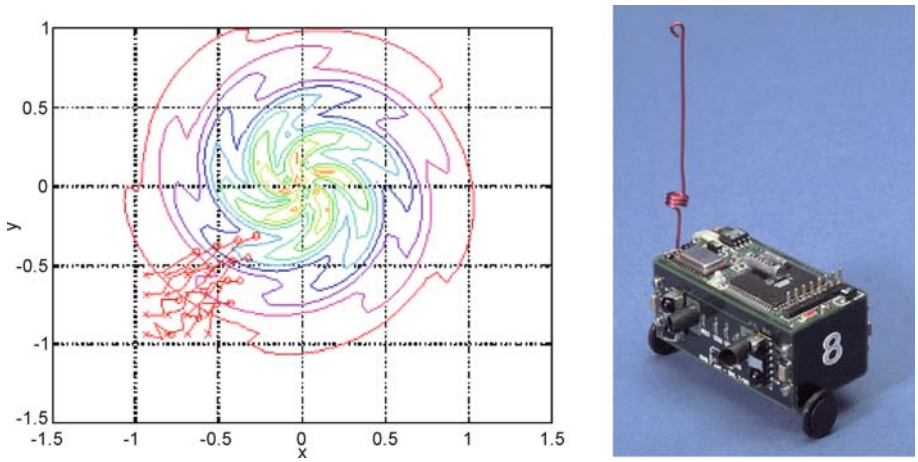

Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 8
***Left*: Multiple vehicles converging on a rotating plume. *Right*: Miniature robot used in the plume localization experiment that located a block of dry ice**

Example 6. Converging on the Source of a Plume – 3D Case

The last example is a three-dimensional plume localization problem. The objective is for multiple vehicles to locate and converge on a source, which could either be acoustic, temperature, or chemical. It is assumed that the spatial signature of the source can be approximated by a quadratic surface.

Step 1. The objective is

$$\max_{\bar{x}} v(\bar{x}) \tag{35}$$

where the global performance index is

$$v(\bar{x}) \cong \sum_{i=1}^{N} a_0 + A_1^{\mathrm{T}}(\bar{x}_i - \bar{x}_0) + \frac{1}{2}(\bar{x}_i - \bar{x}_0)^{\mathrm{T}} A_2 (\bar{x}_i - \bar{x}_0) . \tag{36}$$

The parameters of the quadratic surface are $a_0 \in \Re$, $A_1 \in \Re^3$, and $A_2 \in \Re^{3x3}$. The center of the source is located at $\bar{x}_0 \in \Re^3$.

Step 2. The distributed objective is

$$\max_{\bar{x}_i} v_i(\bar{x}) \quad \forall i = 1, \ldots, N \tag{37}$$

where the partitioned performance index is

$$v_i(\bar{x}) \cong \sum_{j \in \mathrm{NN}} a_{0i} + A_{1i}^{\mathrm{T}}(\bar{x}_j - \bar{x}_i) + \frac{1}{2}(\bar{x}_j - \bar{x}_i)^{\mathrm{T}} A_{2i}(\bar{x}_j - \bar{x}_i). \tag{38}$$

Each vehicle determines its own estimate of the quadratic surface using information from its nearest neighbors.

Step 3. The second order Newton's method control law is

$$\bar{x}_i(k+1) = \bar{x}_i(k) - \alpha \; A_{2i}^{-1}\big|_{\bar{x}(k)} \; A_{1i}\big|_{\bar{x}(k)} \tag{39}$$

where the quadratic coefficients are determined from the solution to the nearest neighbor equations

$$v_i(\bar{x}_j) = a_{0i} + A_{1i}^{\mathrm{T}}(\bar{x}_j - \bar{x}_i) + \frac{1}{2}(\bar{x}_j - \bar{x}_i)^{\mathrm{T}} A_{2i}^{\mathrm{T}}(\bar{x}_j - \bar{x}_i) \quad \forall j \in \mathrm{NN} . \tag{40}$$

For the 3D case, the number of nearest neighbors is nine because ten measurements (including itself) are needed to uniquely determine the quadratic coefficients A_{1i} and A_{2i}. Most recently, this algorithm has been implemented on underwater vehicles that locate and converge in on a 3D plume [62]. Preliminary tests were conducted with a synthetic plume. Synthesized sensor data was calculated as a function of position to debug the algorithm. The underwater robots are shown in Fig. 9, as well as the results of a typical test run.

These six examples demonstrate the utility of this three-step process for creating locally optimal distributed controls for multiple robotic vehicles. The resulting control laws are robust and only require sharing of information between nearest neighbors. The robustness is the result of the self-organizing nature of the control where nearest neighbors are continually changing throughout the motions. If a vehicle is lost or dies, another set of nearest neighbors can be used to complete the task. By using penalty functions to approximate constraints, the control laws are in a form that is identical to the potential field control laws often used for controlling single and multiple robots. The main difference is the switching of potential fields based on the nearest neighbors and the nearest obstacles.

Communication Effects

In this section, previous analysis regarding stable control of multiple vehicles using large-scale decentralized control techniques [47] is extended to include the communications aspects of the problem. A stability analysis shows that the local feedback control gains of the robotic vehicles must be decreased if the communication sample period is increased. Therefore, there is a tight coupling between communications and controls that cannot be ignored. In general, a system will be more responsive and have shorter settling times if the feedback control gains are as large as possible and the communication sample period is as short as possible. This section evaluates the resulting communication sample period of four different communication protocols: a Time Division Multiple Access (TDMA) linear broadcast, a TDMA polylogarithmic broadcast, a TDMA coloring algorithm, and a Collision Sense Multiple Access (CSMA) coloring algorithm. The selection of the best protocol depends on the density of the robot vehicles and the communication radius of each vehicle.

Throughout this section, the one-dimensional dispersion example from the previous section is used to illustrate the design methodology. In [47], it is shown that α in Eq. (5) is actually more constrained than $0 < \alpha \leq 1$ depending on the responsiveness of the vehicle and the communication sample period. Therefore, the next question to ask is that of connective stability. Under what conditions will the overall system be globally asymptotically sta-

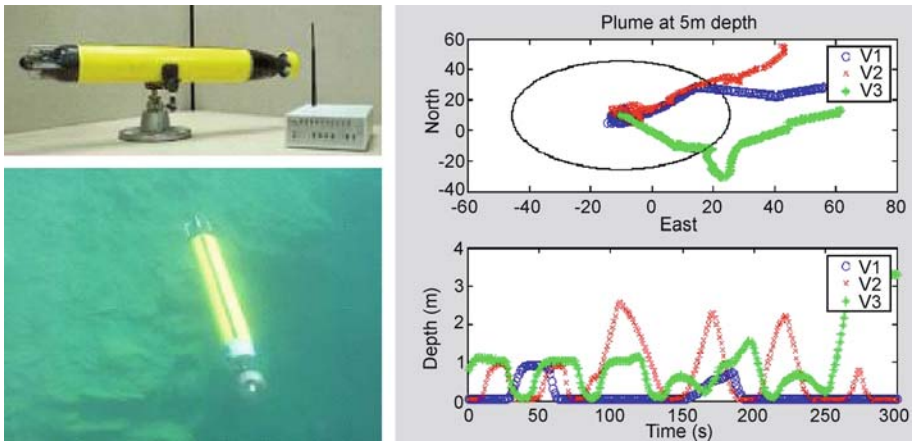

Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 9
***Left*: Nekton Research underwater vehicle used to locate synthetic plume source. *Right*: Underwater synthetic plume test results**

ble even under structural perturbations? Analysis of connective stability is based upon the concept of vector Lyapunov functions, which associates several scalar functions with a dynamic system in such a way that each function guarantees stability in different portions of the state space. The objective is to prove that there exist Lyapunov functions for each of the individual subsystems and then prove that the vector sum of these Lyapunov functions is a Lyapunov function for the entire system.

Stability Analysis

To simplify matters, we will assume that the control function has already been chosen and the closed loop dynamics of the discrete time system can be written as

$$\mathbf{S}: \; x_i(k+1) = g_i(k, x_i) + \tilde{g}_i(k, \bar{x}), \quad i \in \{1, \ldots, N\}, \tag{41}$$

where $\bar{x}(k) \in \Re^n$ is the state of $\mathbf{S}$ (e. g., x, y position, orientation, and linear and angular velocities of all vehicles) at time $k \in T$, $x_i(k) \in \Re^{n_i}$ is the state of the ith subsystem $\mathbf{S}_i$ at time $k \in T$. The function $g_i: T \times \Re^{n_i} \to \Re^{n_i}$ describes the local dynamics of $\mathbf{S}_i$, and the function $\tilde{g}_i: T \times \Re^n \to \Re^{n_i}$ represents the dynamic interaction of $\mathbf{S}_i$ with the rest of the system $\mathbf{S}$. The interconnection function can be written as

$$\tilde{g}_i(k, \bar{x}) = \tilde{g}_i(k, \bar{e}_{i1}x_1, \bar{e}_{i2}x_2, \ldots, \bar{e}_{iN}x_N), \quad i \in \{1, \ldots, N\} \tag{42}$$

where $\bar{e}_{ij} \in B^{n_i \times n_j}$, and the elements of the fundamental interconnection matrix $\bar{E} = (\bar{e}_{ij})$ are

$$(\bar{e}_{ij})_{pq} = \begin{cases} 1, & (x_j)_q \text{ occurs in } (\tilde{g}_i(t, \bar{x}))_p \\ 0, & (x_j)_q \text{ does not occur in } (\tilde{g}_i(t, \bar{x}))_p \end{cases}, \tag{43}$$

where $q \in \{n_j\}$ and $p \in \{n_i\}$.

The structural perturbations of $\mathbf{S}$ are introduced by assuming that the elements of the fundamental interconnection matrix that are one can be replaced by any number between zero and one, i. e.

$$e_{ij} = \begin{cases} [0, 1], & \bar{e}_{ij} = 1 \\ 0, & \bar{e}_{ij} = 0. \end{cases} \tag{44}$$

Therefore, the elements e_{ij} represent the strength of coupling between the individual subsystems. A system is connectively stable if it is stable in the sense of Lyapunov for all possible $E = (e_{ij})$ [63]. In other words, if a system is connectively stable, it is stable even if an interconnection becomes decoupled, i. e. $e_{ij} = 0$, or if interconnection parameters are perturbed, i. e. $0 < e_{ij} < 1$. This is potentially very powerful, as it proves that the system will be stable even if an interconnection is lost through communication failure.

For linear systems, the discrete time dynamics may be written as

$$\mathbf{S}: \; x_i(k+1) = A_{ii}x_i(k) + \sum_{j=1}^{N} e_{ij}A_{ij}x_j(k), \quad i \in \{1, \ldots, N\}, \tag{45}$$

and the Lyapunov function for each individual subsystems is $v_i(x_i) = \left(x_i^{\mathrm{T}} H_i x_i\right)^{1/2}$ where H_i is a positive definite matrix. For the system **S** to be connectively stable, the following test matrix $W = \left(w_{ij}\right)$ must be an M-matrix (i. e., all leading principal minors must be positive) [64]:

$$w_{ij} = \begin{cases} \xi_i, & i = j \\ -e_{ij}\xi_{ij}, & i \neq j \end{cases} \tag{46}$$

where $\xi_i = 1 - \sqrt{1 - \frac{1}{\lambda_M(H_i^*)}}$, $\xi_{ij} = \lambda_M^{1/2}\left(A_{ij}^{\mathrm{T}} A_{ij}\right)$, and

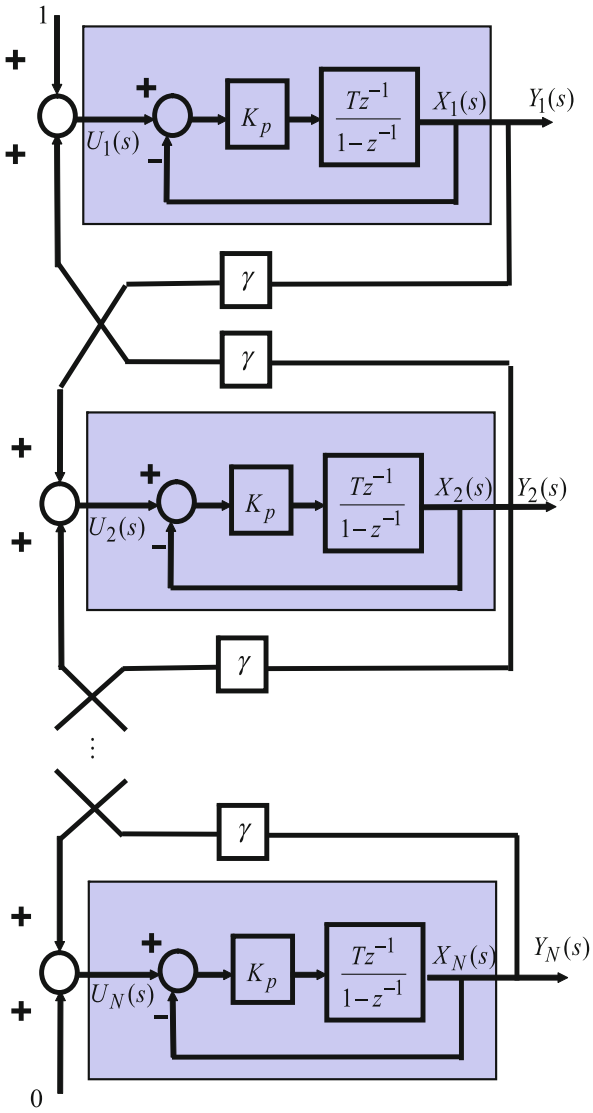

Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 10
Discrete time control block diagram of *N*-vehicle interaction problem

$A_{ii}^{\mathrm{T}} H_i^* A_{ii} - H_i^* = -I$, $\lambda_M(\bullet)$ is the maximum eigenvalue of the corresponding matrices, and the superscript * denotes the Hermitian operator. For the linear dispersion example, we will model the vehicle dynamics as a discrete time integrator with a position feedback loop (see Fig. 10). The proportional control gain is K_{p}, and the sampling period is T.

The sampling period is both the communication and position update sample time. The state equations of the system are

$$\begin{aligned} \mathbf{S}\colon x_1(k+1) &= \left(1 - K_{\mathrm{p}}T\right) x_1(k) + \gamma K_{\mathrm{p}} T x_2(k) \\ x_i(k+1) &= \left(1 - K_{\mathrm{p}}T\right) x_i(k) + \gamma K_{\mathrm{p}} T x_{i-1}(k) \\ &\quad + \gamma K_{\mathrm{p}} T x_{i+1}(k), \quad i \in \{2, \ldots, N-1\} \\ x_N(k+1) &= \left(1 - K_{\mathrm{p}}T\right) x_N(k) + \gamma K_{\mathrm{p}} T x_{N-1}(k)\ . \end{aligned} \tag{47}$$

Note that when comparing Eq. (47) to Eqs. (5) and (6), it is evident that $2\alpha = K_{\mathrm{p}}T$ and $\alpha = \gamma K_{\mathrm{p}} T$. If Eq. (47) is forced to be exactly equivalent to Eqs. (5) and (6), then $\gamma = 1/2$ and $\alpha = K_{\mathrm{p}}T/2$. The following stability test is less restrictive, and the interaction gain γ is less constrained. If $0 < K_{\mathrm{p}}T \leq 1$, the resulting test matrix is

$$W = \begin{bmatrix} K_{\mathrm{p}}T & -K_{\mathrm{p}}T\gamma & 0 & \cdots & 0 \\ -K_{\mathrm{p}}T\gamma & K_{\mathrm{p}}T & -K_{\mathrm{p}}T\gamma & & \vdots \\ 0 & -K_{\mathrm{p}}T\gamma & K_{\mathrm{p}}T & & 0 \\ \vdots & & & \ddots & -K_{\mathrm{p}}T\gamma \\ 0 & \cdots & 0 & -K_{\mathrm{p}}T\gamma & K_{\mathrm{p}}T \end{bmatrix}, \tag{48}$$

and if $1 < K_{\mathrm{p}}T \leq 2$, the test matrix is

$$W = \begin{bmatrix} \left(2 - K_{\mathrm{p}}T\right) & -K_{\mathrm{p}}T\gamma & 0 & \cdots & 0 \\ -K_{\mathrm{p}}T\gamma & \left(2 - K_{\mathrm{p}}T\right) & -K_{\mathrm{p}}T\gamma & & \vdots \\ 0 & -K_{\mathrm{p}}T\gamma & \left(2 - K_{\mathrm{p}}T\right) & & 0 \\ \vdots & & & \ddots & -K_{\mathrm{p}}T\gamma \\ 0 & \cdots & 0 & -K_{\mathrm{p}}T\gamma & \left(2 - K_{\mathrm{p}}T\right) \end{bmatrix}. \tag{49}$$

For $N = 2$, the test matrix is an M-matrix, and the system is connectively stable if

$$|\gamma| < \begin{cases} 1, & 0 < K_{\mathrm{p}}T \leq 1 \\ \frac{2}{K_{\mathrm{p}}T} - 1, & 1 < K_{\mathrm{p}}T \leq 2\ . \end{cases} \tag{50}$$

Figure 11 illustrates the stability region for the case of $N = 2$. The dark region represents stable combinations of the interaction gain γ and $K_p T$ (proportional control gain multiplied by the sampling period). The white region represents unstable combinations of γ and $K_p T$. We refer to the dark region as a stability "house" due to the shape of the stable zone. The size of this stability house varies only with N. As N is increased, the house gets smaller in width but maintains the same height and shape. The size of the stability house is a measure of the robustness of the closed-loop system to parameter variations in interaction gain γ, sampling period T, and proportional control gain K_p. Figure 11 also shows the stability region for $N = 10\,000$.

For this particular example, another way to check the stability of this linear system is to check that the eigenvalues of the system matrix A are within the unit circle. There is a special formula (p. 59 in [65]) for the eigenvalues of A given by

$$\lambda_i(A) = 1 - K_p T + 2K_p T\gamma \cos\left(\frac{i\pi}{N+1}\right), \quad i = 1, \ldots, N. \tag{51}$$

From this formula, we can see that as $N \to \infty$, the cosine term becomes unity. This implies that γ must stay between -0.5 and 0.5 for $K_p T$ less than one in order to maintain stability. For $K_p T$ greater than one, the admissible γ values taper off parabolically (the sloped "roof") until $K_p T = 2$.

Several conclusions can be drawn from this stability analysis. First, asymptotic stability of vehicle positions depends on vehicle responsiveness K_p, communication sampling period T, and vehicle interaction gain γ. If the vehicle is too fast (large K_p) or the sample period is too long (large T) then the vehicles will go unstable. There is a dependence on interaction gain for stability as well. Second, the interaction gains can be used to bunch the vehicles closer together or spread them out. Third, the stability region shrinks as the number of vehicles, N, increases but only to a defined limit.

Communications Sample Period

As noted, the communication sample period greatly affects the stability of the system. As defined in the equations above, this sample period is the time it takes for every node to communicate once. In this section, we will evaluate the communication sample period of four different communication schemes: a Time Division Medium Access (TDMA) linear broadcast, a TDMA polylogarithmic broadcast, a TDMA coloring algorithm, and a Collision Sense Medium Access (CSMA) coloring algorithm. All of these schemes assume that each node has a unique identification number. The TDMA schemes also assume that each node has a synchronized clock that is used to notify each node when it may transmit a message. The CSMA scheme first checks the communication channel for a collision before transmitting a packet.

In order to determine this sample period, one important parameter associated with an ad hoc communication network is the degree of the network. The degree of the network is defined as the maximum number of nodes that any node can communicate with given a limited communication range.

For the network shown in Fig. 12, the degree of the network is

$$\Delta = \max_{i \in \{1,\ldots,N\}} \left\{ \sum_{j=1,\, j \neq i}^{N} \text{range}\left(x_i - x_j, R_c\right) \right\} \tag{52}$$

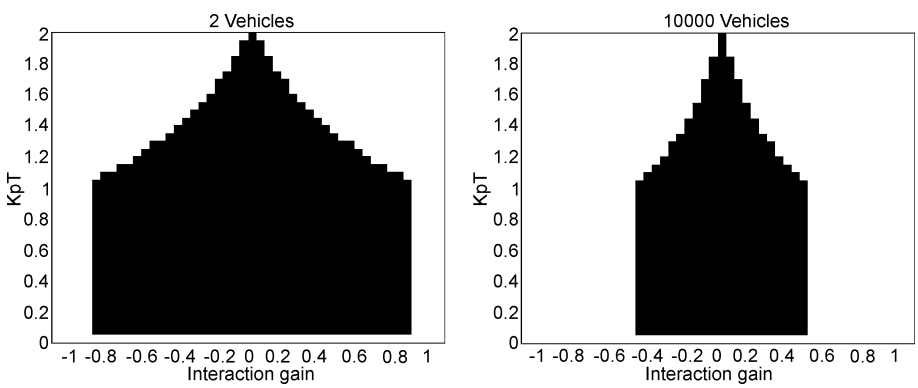

Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 11
***Left*: Stability region for the $N = 2$ vehicle case. *Right*: Stability region for the $N = 10\,000$ vehicle case**

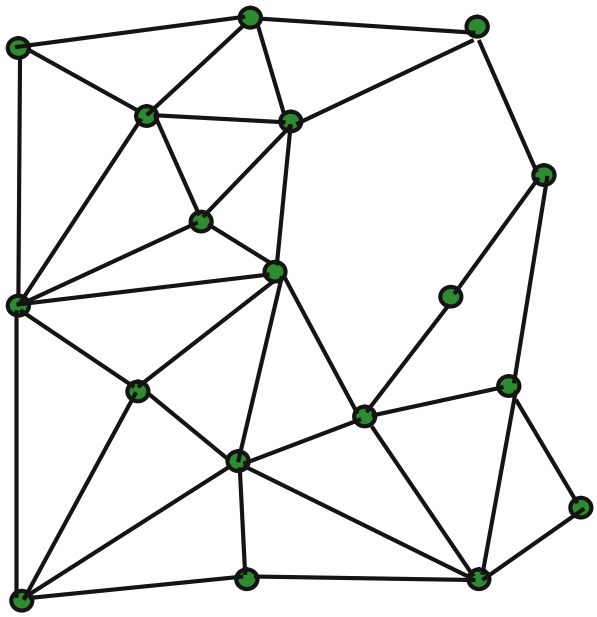

Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 12
Graph of ad hoc communication network. Nodes with connecting lines can communicate with each other

where

$$\text{range}\left(x_i - x_j, R_c\right) = \begin{cases} 1 & \text{if } \left|x_i - x_j\right| < R_c \\ 0 & \text{else} \end{cases} \tag{53}$$

and R_c is the communication radius of each node. Assuming all the robots are evenly spaced along a line of length L and have a density $\delta = \frac{N}{L}$, then the degree of the resulting network is

$$\Delta = 2\left\lfloor \delta R_c \right\rfloor = 2\left\lfloor \frac{N R_c}{L} \right\rfloor . \tag{54}$$

For the TDMA linear broadcast where every node is assigned a unique identification number, the communication sample period time required for every node to send a message is

$$T_{\text{linear}} = \tau N \tag{55}$$

where τ is the time period associated with each communication time slot. Notice that the above expression is proportional to N. This delay time can be shortened by using a polylogarithmic broadcast scheme [66] where each node communicates during multiple time slots. Even though multiple messages are being broadcast at the same time, the message is guaranteed a successful broadcast during

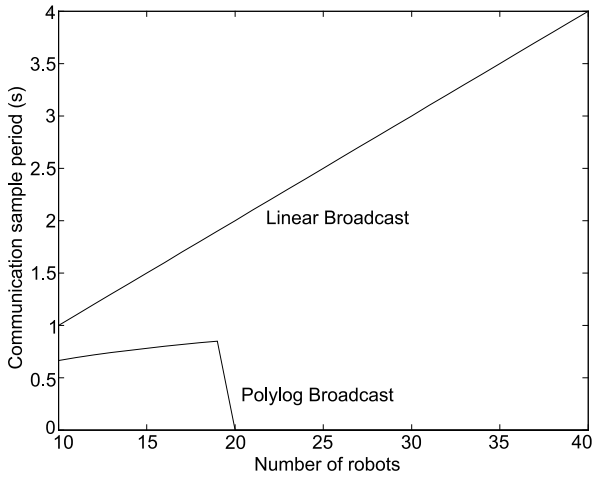

Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 13
Communication sample period for TDMA linear and polylogarithmic broadcasts when $R_c/L = 0.1$ and $\tau = 0.1$ s

one of the time slots as long as the degree of the network is below a certain value. The communication period for a polylogarithmic broadcast is

$$T_{\text{poly log}} = \tau \left(2 \log_2 N\right)^h \tag{56}$$

when

$$\Delta \leq 2^{h+1} - 1 \tag{57}$$

and where

$$h = \left\lfloor \left(\log_2 N\right) / \left(\log_2 \left(\log_2 N\right) + 1\right) \right\rfloor . \tag{58}$$

Notice that this expression is proportional to the $\log_2 N$ instead of N. Figure 13 compares the linear broadcast to the polylogarithmic broadcast for a network spread out over a line and with each node having a communication radius that is one-tenth the length of the line. At 20 robots, the average degree of the network becomes too large and the polylogarithmic broadcast will no longer work.

Even better than the linear broadcast and the polylogarithmic broadcast, a color scheme allows multiple nodes to communicate at the same time by using spatial reuse of time slots. Time slots (called colors) are assigned so that each node has a different color than its first and second nearest neighbors. By using different colors, the hidden node problem, where two nodes speak to an intermediate node at the same time, is eliminated. In graph theory, the minimum number of colors can range from the maximum degree of the network plus one to the square of the

maximum degree of the network plus one.

$$\Delta + 1 \leq k \leq \Delta^2 + 1 \,. \tag{59}$$

However, in a typical planar wireless network the number of colors is typically bounded by

$$k \leq \Delta + \varepsilon \tag{60}$$

where ε is a small number, typically 1 to 5. For a TDMA colored network, the communication sample period is given by

$$T_{\mathrm{TDMA}} = \tau\,(\Delta + \varepsilon) \quad \text{where } \varepsilon \text{ is small} \,. \tag{61}$$

For a CSMA network, the actual communication time is non-deterministic because the packets often collide and a random back-off is used before retransmitting. The average communication time depends on the network utilization, i. e. the percentage of time the network is being used. Modeling the CSMA network as a M/D/1 queue [67], the average communication time is

$$\tau_{\mathrm{CSMA}} = \frac{\rho T_{\mathrm{m}}}{2\,(1-\rho)} + T_{\mathrm{m}} \tag{62}$$

where T_{m} is the time to send a single message (or service time of the queue). For simplicity of comparison with the TDMA network, we will assume that $T_{\mathrm{m}} = \tau$. In reality, the message length T_{m} in a TDMA network must be slightly less than the time slot τ. The utilization factor is

$$\rho = \frac{T_{\mathrm{m}}\,(\Delta + \varepsilon)}{T_{\mathrm{d}}} \tag{63}$$

where T_{d} is the delay time between each new message that a node initiates. For a CSMA network with degree Δ, the communication sample period is approximately given by

$$T_{\mathrm{CSMA}} = \tau_{\mathrm{CSMA}}\,(\Delta + \varepsilon) \,. \tag{64}$$

Figure 14 shows how the network utilization changes as the number of robots increases in our one-dimensional dispersion example. The resulting communication sampling period for both the TDMA and CSMA colored networks are also shown in Fig. 14. Notice that for smaller numbers of vehicles both the colored TDMA and CSMA networks have a shorter communication sample period than both linear and polylogarithmic broadcasts. However, when the utilization of the CSMA network reaches 0.9, the CSMA network starts to substantially degrade in performance. This is caused by flooding the network with messages as the density of the robots increase while the back-off time of communication stays the same. It should be noted that this flooding can be alleviated if the nodes were to adjust how often they send out messages based on the maximum degree of the network. Ideally, the nodes should adjust their communication sample period so that $T_{\mathrm{d}} = T_{\mathrm{CSMA}}$.

These results show that a colored TDMA network appears to perform the best. However, the disadvantage of the colored TDMA network is that there is an initialization time that is required to determine the color of each node whenever the network topology changes. The other algorithms have the advantage that they do not require a network initialization time. Using an algorithm by [68], this initialization time is the time required to broadcast their

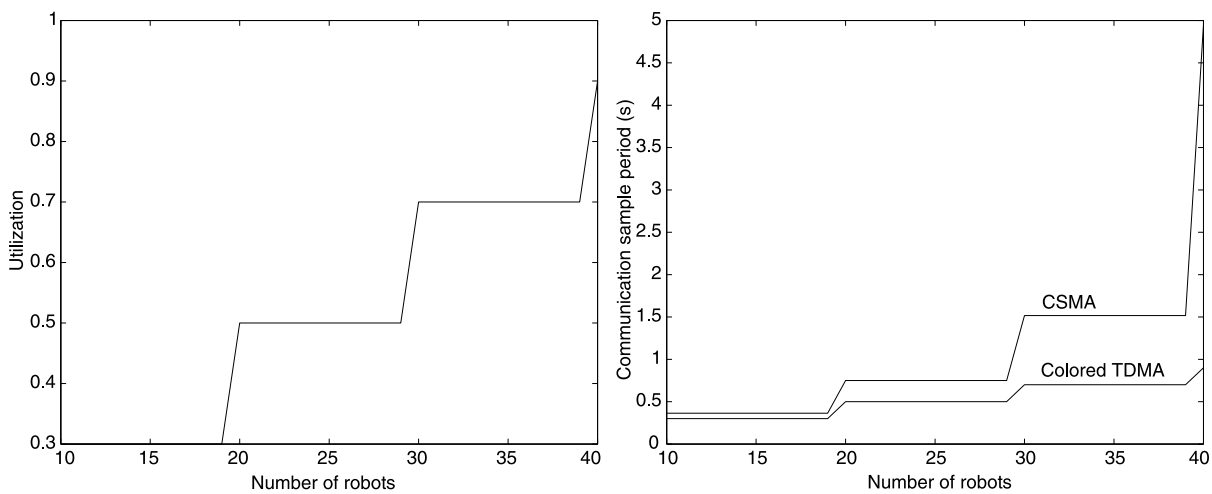

Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 14
***Left*: Utilization for a CSMA network and *right*: communication sample period for TDMA and CSMA reconfigurable coloring when $R_c/L = 0.1$, $\tau = 0.1$ s, $\varepsilon = 1$, and $T_{\mathrm{d}} = 1$ s**

own identification number, their 1st nearest neighbor lists (after which each node can determine 2nd nearest neighbor), and $k+2$ additional messages for coloring. In addition to these messages, each neighboring node must acknowledge the messages containing the neighbor lists and coloring information to ensure that the messages were received. Since time slots are typically not assigned before hand, this initialization process occurs using CSMA protocols, and it should be performed whenever the topology of the network changes, i. e. when robots move. Assuming that all messages are the same length, the resulting initialization time is given by

$$\begin{aligned} t_{\text{init}} &= \tau_{\text{CSMA}}\left[2\left(\Delta+\varepsilon\right)+\left(\Delta+\varepsilon+2\right)\left(\Delta+\varepsilon\right)\right] \\ &\quad + \tau_{\text{CSMA}}\left[\left(\Delta+\varepsilon\right)\left(\Delta+\varepsilon-1\right)\right. \\ &\quad \left. +\left(\Delta+\varepsilon\right)\left(\Delta+\varepsilon+2\right)\left(\Delta+\varepsilon-1\right)\right] \end{aligned} \tag{65}$$

where ε is small. This initialization time is plotted as a function of the number of robots in Fig. 15. This figure shows that this initialization time can be substantial. Adding the initialization time in Fig. 15 to the communication sample period in Fig. 14, we see that the linear, polylogarithmic, and CSMA networks have a shorter sample period than the colored TDMA network. In general, it is best to use the TDMA mode only if the network does not reconfigure; otherwise, it is best to use a CSMA mode of communications.

Utilization Versus Delay in CSMA Networks

If a control system is implemented with a CSMA communications scheme, there is a tradeoff between network utilization and message delay. Ideally, one would like to maximize network utilization while minimizing the delay seen by each message. Modeling the CSMA network as an M/D/1 queue, the normalized delay (where T_{m} is the message length) is given by

$$\frac{\tau_{\text{CSMA}}}{T_{\text{m}}} = \frac{\rho}{2\left(1-\rho\right)} + 1 \tag{66}$$

where τ_{CSMA} is the average communication time. Using the cost function $J\left(\rho\right)$ below, the gains K_1 and K_2 may be used to weight the penalty associated with utilization and delay.

$$J\left(\rho\right) = \frac{K_1}{\rho} + \frac{K_2\rho}{2\left(1-\rho\right)} + K_2\,. \tag{67}$$

The optimal solution is then obtained by minimizing the cost function over $0 \le \rho < 1$

$$J\left(\rho^*\right) = \min_{0\le\rho<1} J\left(\rho\right) = \min_{0\le\rho<1}\left\{\frac{K_1}{\rho} + \frac{K_2\rho}{2\left(1-\rho\right)} + K_2\right\}. \tag{68}$$

Differentiating Eq. (67) with respect to ρ yields

$$\frac{\partial J\left(\rho\right)}{\partial\rho} = -\frac{K_1}{\rho^2} + \frac{2\left(1-\rho\right)K_2 + 2K_2\rho}{4\left(1-\rho\right)^2}\,. \tag{69}$$

Solving Eq. (69) for the minimum gives the following optimal value for network utilization ρ^* based on the weighting values K_1 and K_2.

$$\rho^* = \frac{-2K_1 \pm \sqrt{2K_1K_2}}{\left(K_2-2K_1\right)}, \quad 0 \le \rho^* < 1\,. \tag{70}$$

If utilization and average delay are equally weighted, the optimal value of ρ is $\rho^* = 0.5858$. Optimal utilization values for several values of K_1 and K_2 are summarized in Table 1. These results are intuitive – if average message delay is critical, then network utilization should be kept to a minimum.

This section illustrates the tight coupling that exists between communications and controls when designing large-scale cooperative robotic systems. A connective stability analysis shows that local feedback control gains and

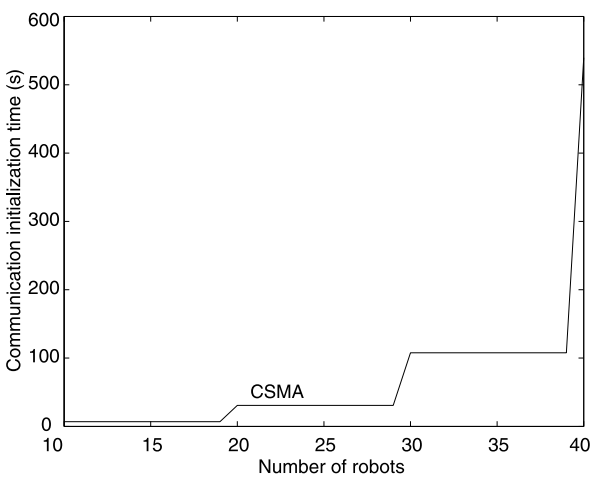

Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Figure 15
Communication overhead for reconfigurable coloring when $R_c/L = 0.1$, $\tau = 0.1$ s, $\varepsilon = 1$, and $T_d = 1$ s

Distributed Controls of Multiple Robotic Systems, An Optimization Approach, Table 1
Optimal utilization for different values of K_1 and K_2

Weighting gains	Optimal utilization, ρ^*
$K_1 = K_2 = 1$	$\rho^* = 0.5858$
$K_1 = 1$, $K_2 = 10$	$\rho^* = 0.309$
$K_1 = 0.1$, $K_2 = 10$	$\rho^* = 0.124$

communication sample periods are inversely related. If the communication sample period increases, then the local feedback control gains must decrease. The communication sample period is a function of the protocol, and the protocol with the shortest communication sample period depends on the density of robots and the communication radius. By assuming worst-case conditions for robot density and communication range, this analysis can be used off-line to determine conservative control gains required for stable control. In the future, it might also be possible to use this analysis on-line to adjust control gains and/or communication range as the robot density changes.

Conclusions

This chapter described an integrated approach to designing communication, sensing, and control systems for fixed and mobile distributed systems. The analysis built upon provably convergent cooperative controls techniques and upon concepts from graph theory as applied to communication networks. A common mathematical framework consisting of three steps was developed for creating decentralized cooperative control laws. The first step is to define a global performance index for the cooperative behavior. The second step is to partition the performance index so that only local interactions are included. The third step is to create a first or second order gradient control law that minimized the partitioned performance index. After these three steps, a vector Lyapunov technique is used to determine the stability constraints on the individual subsystem control gains, interaction control gains, and the communication sampling period.

The communication protocols were evaluated to ensure that they could meet the stability constraint on the communication sampling period. Coloring algorithms from graph theory were used to compare the required performance of TDMA or CSMA communication networks. In general, we found that TDMA networks will outperform CSMA networks if the network is stationary, and there is sufficient time to perform the network coloring algorithms. However, if the network is moving, then a CSMA network will outperform the TDMA network as long as the network utilization stays below 58 percent. Queuing theory was used to determine a closed form solution of the time response of a CSMA network, and Monte Carlo simulations have been used to verify the solution. The vector Lyapunov analysis shows that the collective CSMA networked system will still remain stable as long as the maximum communication time is below the maximum sampling period determined from the vector Lyapunov analysis.

Future Directions

Future work should focus in several directions. First, more complex vehicle dynamics should be considered. This may include the effects of nonlinearities in the vehicle locomotion and/or higher order friction effects. Second, more complicated vehicle formation problems should be considered which may include obstacles and varied terrain features. Third, other collective control applications should be addressed such as distributed electric power grids and other swarm applications. Finally, the communications issue should continue to be studied to look at the effects of sample periods, delays, and interrupted communications on the stability of collectives for other communication protocols. Eventually, real-time algorithms should be developed that evaluate the stability of a networked system, and appropriately adjust the communication protocol and vehicle responsiveness.

Acknowledgments

The authors greatly appreciate the help of Steven Eskridge, John Hurtado, Chris Lewis, John Harrington, and Nekton Research, LLC, in implementing and testing these algorithms on a variety of robot platforms. This work was supported in part by the Sandia National Laboratories. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE–AC04–94AL85000. In addition, this research was partially funded by the Information Processing Technology Office and Microsystems Technology Office of the Defense Advanced Research Projects Agency.

Bibliography

1. Arai T, Pagello E, Parker PE (2002) Guest Editorial: Advances in Multirobot Systems. IEEE Trans Robot Autom 18(5):655–659
2. Bonabeau E, Dorigo M, Theraulaz G (1999) Swarm Intelligence: From Natural to Artificial Systems. Oxford University Press, New York
3. Arkin RC (1992) Cooperation Without Communication: Multiagent Schema-Based Robot Navigation. J Robotic Syst 9(3): 351–364
4. Balch T, Arkin RC (1998) Behavior-Based Formation Control for Multirobot Teams. IEEE Trans Robot Autom 14(6):926–939
5. Kube RC, Zhang H (1993) Collective Robotics: From Social Insects to Robots. Adapt Behav 2(2):189–218
6. Parker LE (1998) ALLIANCE: An Architecture for Fault Tolerant Multirobot Cooperation. IEEE Trans Robot Autom 14(2): 220–240
7. Di Marco M, Carulli A, Giannitrapani A, Vicino A (2003) Simultaneous Localization and Map Building for a Team of Cooperating Robots: A Set Membership Approach. IEEE Trans Robot 19(2):238–249

8. Burgard W, Moors M, Stanchniss C, Schneider FE (2005) Coordinated Multi-Robot Exploration. IEEE Trans Robot 21(3): 376–386
9. Cortes J, Martinez S, Karatas T, Bullo F (2004) Coverage Control for Mobile Sensing Networks. IEEE Trans Robot Autom 20(2):865–875
10. Tang Z, Ozguner U (2005) Motion Planning for Multitarget Surveillance With Mobile Sensor Agents. IEEE Trans Robot 21(5):898–908
11. Vidal R, Shakernia O, Kim HJ, Shim DH, Sastry S (2002) Probabilistic Pursuit-Evasion Games: Theory, Implementation, and Experimental Evaluation. IEEE Trans Robot Autom 18(5): 662–669
12. Isler V, Kannan S, Khanna S (2005) Randomized Pursuit-Evasion in a Polygonal Environment. IEEE Trans Robot 21(5):875–894
13. Cao Z, Tan M, Li L, Gu N, Wang S (2006) Cooperative Hunting by Distributed Mobile Robots Based on Local Interaction. IEEE Trans Robot 22(2):403–407
14. Bopardikar SD, Bullo F, Hespanha JP (2007) Cooperative Pursuit with Sensing Limitations. In: Proceedings of the 2007 American Control Conference. IEEE, Minneapolis
15. Fua CH, Ge SS (2005) COBOS: Cooperative Backoff Adaptive Scheme for Multirobot Task Allocation. IEEE Trans Robot 21(6):1168–1178
16. Daigle MJ, Koutsoukos XD, Biswas G (2007) Distributed Diagnosis in Formations of Mobile Robots. IEEE Trans Robot 23(2):353–369
17. Crandall JW, Cummings ML (2007) Identifying Predictive Metrics for Supervisory Control of Multiple Robots. IEEE Trans Robot 23(5):942–951
18. Chen Q, Luh JYS (1994) Coordination and Control of a Group of Small Mobile Robots. In: Proceedings of the IEEE International Conference on Robotics and Automation, vol 3. pp 2315–2320. IEEE, San Diego
19. Yamaguchi H, Arai T (1994) Distributed and Autonomous Control Method for Generating Shape of Multiple Mobile Robot Group. In: Proceedings of the IEEE International Conference on Intelligent Robots and Systems, vol 2. pp 800–807. IEEE, Munich
20. Yamaguchi H, Burdick JW (1998) Asymptotic Stabilization of Multiple Nonholonomic Mobile Robots Forming Group Formations. In: Proceedings of the 1998 Conference on Robotics & Automation. Leuven, Belgium, pp 3573–3580
21. Yoshida E, Arai T, Ota J, Miki T (1994) Effect of Grouping in Local Communication System of Multiple Mobile Robots. In: Proceedings of the IEEE International Conference on Intelligent Robots and Systems, vol. 2. pp 808–815. IEEE, Munich
22. Molnar P, Starke J (2000) Communication Fault Tolerance in Distributed Robotic Systems. In: Parker LE, Bekey G, Barhen J (eds) Distributed Autonomous Robotic Systems 4. Springer, Berlin, pp 99–108
23. Schneider FE, Wildermuth D, Wolf HL (2000) Motion Coordination in Formations of Multiple Robots Using a Potential Field Approach. In: Parker LE, Bekey G, Barhen J (eds) Distributed Autonomous Robotic Systems 4. Springer, Berlin, pp 305–314
24. Stankovic SS, Stanojevic MJ, Siljak DD (2000) Decentralized Overlapping Control of a Platoon of Vehicles. IEEE Trans Control Syst Technol 8:816–832
25. Seiler P, Pant A, Hedrick K (2004) Disturbance Propagation in Vehicle Strings IEEE Trans Autom Control 29:1835–1841
26. Barooah P, Mehta PG, Hespanha JP (2007) Control of Large Vehicular Platoons: Improving Closed Loop Stability by Mistuning. In: Proceedings of the 2007 American Control Conference. IEEE, Minneapolis
27. Ogren P, Egerstedt M, Hu X (2002) A Control Lyapunov Function Approach to Multiagent Coordination. IEEE Trans Robot Autom 18(5):847–851
28. Lawton JRT, Beard RW (2003) A Decentralized Approach to Formation Maneuvers. IEEE Trans Robot Autom 19(6):933–941
29. Tanner HG, Pappas GJ, Kumar V (2004) Leader-to-Formation Stability. IEEE Trans Robot Autom 20(3):443–455
30. Tabuada P, Pappas GJ, Lima P (2005) Motion Feasibility of Multi-Agent Formations. IEEE Trans Robot 21(3):387–392
31. Antonelli G, Chiaverini S (2006) Kinematic Control of Platoons of Autonomous Vehicles. IEEE Trans Robot 22(6):1285–1292
32. Beni G, Liang P (1996) Pattern Reconfiguration in Swarms – Convergence of a Distributed Asynchronous and Bounded Iterative Algorithm. IEEE Trans Robot Autom 12(3):485–490
33. Liu Y, Passino K, Polycarpou M (2001) Stability Analysis of One-Dimensional Asynchronous Swarms. American Control Conference, Arlington, pp 25–27, 716–721
34. Winfield A (2000) Distributed Sensing and Data Collection via Broken Ad Hoc Wireless Connected Networks of Mobile Robots. In: Parker LE, Bekey G, Barhen J (eds) Distributed Autonomous Robotic Systems 4. Springer, Berlin, pp 273–282
35. Desai JP, Ostrowski J, Kumar V (1998) Controlling Formations of Multiple Mobile Robots. In: Proceedings of the 1998 IEEE Conference on Robotics & Automation. IEEE, Leuven, pp 2864–2869
36. Desai JP, Kumar V, Ostrowski J (2001) Modeling and Control of Formations of Nonholonomic Mobile Robots. IEEE Trans Robot Autom 17(6):905–908
37. Ji M, Egerstedt M (2007) Distributed Coordination Control of Multiagent Systems While Preserving Connectedness. IEEE Trans Robot 23(4):693–703
38. Zavlanos MM, Pappas GJ (2007) Potential Fields for Maintaining Connectivity of Mobile Networks. IEEE Trans Robot 23(4):812–816
39. Belta C, Kumar V (2004) Abstraction and Control of Groups of Robots. IEEE Trans Robot Autom 20(5):865–875
40. Ge SS, Fua CH (2005) Queues and Artificial Potential Trenches for Multrobot Formations. IEEE Trans Robot 21(4):646–656
41. Milutinovic D, Lima P (2006) Modeling and Optimal Centralized Control of a Large-Size Robotic Population. IEEE Trans Robot 22(6):1280–1285
42. Kloetzer M, Belta C (2007) Temporal Logic Planning and Control of Robotic Swarms by Hierarchical Abstractions. IEEE Trans Robot 23(2):320–330
43. Tanner HG, Loizou SG, Kyriakopoulos KJ (2003) Nonholonomic Navigation and Control of Cooperating Mobile Manipulators. IEEE Trans Robot Autom 19(1):53–64
44. Yamashita A, Arai T, Ota J, Asama H (2003) Motion Planning of Multiple Mobile Robots for Cooperative Manipulation and Transportation. IEEE Trans Robot Autom 19(2):223–237
45. Bonaventura CS, Jablokow KW (2005) A Modular Approach to the Dynamics of Complex Multirobot Systems. IEEE Trans Robot 21(1):26–37
46. Lewis C, Feddema JT, Klarer P (1998) Robotic Perimeter Detection System. In: Proceedings of SPIE, vol 3577. Boston, pp 14–21
47. Feddema JT, Lewis C, Schoenwald DA (2002) Decentralized

Control of Cooperative Robotic Vehicles: Theory and Application. IEEE Trans Robot Autom 18(5):852–864
48. Hurtado JE, Robinett III RD (2005) Convergence of Newton's Method via Lyapunov Analysis. AIAA J Guidance Control Dyn 28(2):363–365
49. Hurtado JE, Robinett III RD, Dohrmann CR, Goldsmith SY (2004) Decentralized Control for aSwarm of Vehicles Performing Source Localization. J Intell Robot Syst 41:1–18
50. Robinett RD, Hurtado JE (2004) Stability and Control of Collectives Systems. J Intell Robot Syst 39:43–55
51. Clauset A, Tanner B, Byrne R, Abdallah CT (2007) Controlling Across Complex Networks. Proceedings of the IFAC Time-delay Symposium, Nantes, pp 17–19
52. Hespanha JP, Naghshtabrizi P, Xu Y (2007) A Survey of Recent Results in Networked Control Systems. Proc IEEE Special Issue: Technol Netw Control Syst 95(1):138–162
53. Gelfand IM, Fomin SV (1963) Calculus of Variations. Prentice-Hall, New Jersey
54. Frieden BR (1998) Physics from Fischer Information. Cambridge University Press, Cambridge
55. Feddema JT, Robinett RD, Byrne RH (2003) An Optimization Approach to Distributed Controls of Multiple Robot Vehicles, Workshop on Control and Cooperation of Intelligent Miniature Robots. IEEE/RSJ Int Conf Intell Robot Syst. IEEE, Las Vegas
56. Censor Y, Zenios SA (1997) Parallel Optimization: Theory, Algorithms, and Applications. Oxford University Press, Oxford
57. Luenberger DG (1984) Linear and Nonlinear Programming, 2nd edn. Addison Wesley, Reading
58. Schmitt DJ et al (2003) Intelligent Mobile Land Mine (IMLM) System, Sandia National Laboratories Report. SAND 2003–1186
59. Feddema JT, Schoenwald DA (2002) Stability Analysis of Decentralized Cooperative Controls. In: Shultz AC, Parker LE (eds) Multi-Robot Systems: From Swarms to Intelligent Automata. Kluwer, Boston, pp 133–122
60. Feddema JT, Schoenwald DA (2002) Distributed Communication/Navigation Robot Vehicle Network. In: Proceedings of World Automation Congress, Orlando, pp 9–13. TSI Press, Albuquerque
61. Byrne RH, Adkins DR, Eskridge SE, Harrington JJ, Heller EJ, Hurtado JE (2002) Miniature Mobile Robots for Plume Tracking and Source Localization Research. J Micromechatronics 1(3): 253–261
62. Byrne RH, Eskridge SE, Hurtado JE, Salvage EL (2003) Algorithms and Analysis of Underwater Vehicle Plume Tracing, Sandia National Laboratories Report. SAND pp 2003–2643
63. Siljak DD (1991) Decentralized Control of Complex Systems. Academic Press, San Diego
64. Sezer ME, Siljak DD (1988) Robust Stability of Discrete Systems. Int J Control 48(5):2055–2063
65. Smith GD (1985) Numerical Solution of Partial Differential Equations: Finite Difference Methods, 3rd edn. Oxford University Press, Oxford
66. Basagni S, Bruschi D, Chlamtac I (1999) A Mobility-Transparent Deterministic Broadcast Mechanism for Ad Hoc Networks. IEEE/ACM Trans Netw 7(6):799–807
67. Kleinrock L (1975) Queueing Systems, Volume 1: Theory, 1st edn. Wiley, New York
68. Chlamtac I, Pinter SS (1987) Distributed Nodes Organization Algorithm for Channel Access in a Multihop Dynamic Radio Network. IEEE Trans Comput C 36(6):728–737

Distributed Robotic Teams: A Framework for Simulated and Real-World Modeling

Michael Janssen, Andrew Drenner, Nikolaos Papanikolopoulos
Center for Distributed Robotics,
Department of Computer Science and Engineering,
University of Minnesota,
Minneapolis, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Robotic Team Design
Software Modeling and Simulation
Tasks Well–Suited to Distributed Robotics
Tasks for Optimizing Distributed Robotics
Future Directions
Acknowledgment
Bibliography

Glossary

Marsupial robot A robot that can pick up and store other (usually smaller) robots and then possibly move or recharge those robots.

Marsupial pair One specific robot which can be picked up by another specific robot, as a pair.

Marsupial team Two or more robots which can form one or more marsupial pairs, i. e. one marsupial robot and multiple carrier robots.

Marsupial mechanism A hardware device which enables a robot to carry one or more robots. May be active or passive.

Dead-reckoning A type of robot localization accomplished by measuring the distance of wheel rotation. Prone to drifting error.

Bluetooth Wireless networking with low bandwidth and a short range, typically less than 10 meters. Used in distributed robots because of it's low power requirements.

Homogeneous team A group of robots which are all of the same model, including all of the same sensors. Each robot is replaceable with any other from the team.

Heterogeneous team A group of robots which includes members that have unique sensors. More types of sensors can be included in a heterogeneous team as each robot can have a unique sensor load.

Reverse auction A type of auctioning where the lowest bidder is the winner. Used for task allocation as the lowest cost represents the lowest amount of resources expended for the task.

Definition of the Subject

The field of robotics covers devices that are in use in a wide variety of applications from interplanetary exploration [56] to performing common household tasks (such as vacuuming your floor) [23]. Definitions of what is and is not a robot can vary wildly. Generally speaking, a robot is a device with the ability to sense and interact with its environment. Usually there is some degree of intelligence or autonomy in a robotic system, but this can vary as well from a *tele-robotic* device which is completely controlled by operators from a remote location [58] to a *fully autonomous robot* which can be given a goal and will reach that goal without any human intervention [5]. The phrase *distributed robotics* deals with teams of robots used to accomplish tasks that a single robot either could not achieve or could not achieve within some constraint (cost of the system, time alloted, etc.).

Robotics is an important field because it allows for the collection of information while minimizing potential hazards to humans. Robots are already in use in war zones, searching for survivors in collapsed buildings, and operating in areas which are inhospitable to humans. Distributed robotics are often used to accomplish these tasks faster and at lower cost.

Introduction

As technology pushes toward smaller, faster, and cheaper devices, it is reasonable to expect multiple robots to work together to solve a common goal. The field of distributed robotics aims to solve the problems which appear when robots work together. It work based on robotics, electrical engineering, mechanical engineering, psychology, intelligent agents, and game theory. Certain distributed robotic approaches are inspired by biological systems as well.

Distributed robotics has many advantages over single-robot systems. Having multiple autonomous systems makes it possible to complete complex and large tasks more quickly than a single robot could. Also, building a single robot with the combined sensing of a robotic team may be cost-prohibitive. The many sensors available on the multiple smaller robots present a significantly larger amount of data about the environment than a single robot alone would produce. Having multiple robots is preeminently useful when there is more than one task to complete, as it is possible to split the team to perform both tasks at once, and even reassign team members from one task to another. Ideally, each member of the team can be built at a relatively low cost. This affordability allows for increased redundancy throughout the robotic team.

The advantages brought by a robot team do not come without drawbacks. Having more robots in use multiplies costs, in both time of maintenance, materials cost, and potentially operator-related expenses. Some of these costs can be overcome by building many simple devices rather than a single complex one as well as developing an increased degree of autonomy. The introduction of a "team mechanic" introduces new problems not found in a monolithic robot, specifically in the areas of team coordination, task assignment, and communication. Communication itself may require additional hardware on each device, making the design of such robots more complex, requiring engineering experience in wireless and RF communication. When a team is tele-operated as opposed to autonomous, usually one operator is required for each robot on the team. This makes the cost of operating such a team of robots much higher, especially in situations where those operators could perform other tasks.

In the rest of this chapter, we will present an overview of distributed robotics. In Sect. "Robotic Team Design", design of robotic teams and views on the composition of such teams is considered. Attention is given to the concerns relating to individual members of a team in Sect. "Design Considerations" and the makeup of a full distributed robot system in Sect. "Robotic Team Composition". Section "Software Modeling and Simulation" covers some of the software used currently to simulate distributed teams and presents some of the challenges specific to distributed teams. Particular attention is presented to modeling a marsupial system in Sect. "Marsupial Modeling". We then cover two robotic tasks that are well-suited to robotic teams, along with an approach for each that takes advantage of a team in Sect. "Tasks Well–Suited to Distributed Robotics". Problems which then arise in optimizing distributed robot teams are covered in Sect. "Tasks for Optimizing Distributed Robotics". Finally in Sect. "Future Directions", we consider briefly the future direction of distributed robotics.

Robotic Team Design

Designing a robotic platform requires the balancing of many factors, many of which are at odds. Cost is a major factor in the design, and may be prohibitive to all others. Sensing modalities and on-board processing are also major in design. In addition to these factors, distributed robotics adds the complexity of multiple robots.

Even when not designing the individual robots, some design decisions must be made for each team. The question of homogeneity of the robots is particularly interesting, as choosing to be homogeneous or heterogeneous has differing strengths. Further, marsupial teams should also be considered.

Design Considerations

Each robot in a distributed team would ideally be designed specific to the purpose of the team. In a research setting, this is often not possible due to the team possibly being used in many different situations for many differing goals. There are many advantages which could be exploited, but would cause other areas of the platform to be neglected. Each of these dimensions of robot design is somewhat dependent on the others.

There are a number of important criteria in designing a distributed robotic team. Many of the trade-offs between these criteria are based on mission specific goals and will vary between implementations. Designing a "general purpose" distributed robotic team is a difficult if not impossible task as some of the criteria below will create interdependencies which may not be easily satisfied.

- *Physical Characteristics* – The physical size of the robot, its shape, and its mass will have impact on how the robots can be deployed, traverse terrain, or withstand impacts if it were to fall in its environment. Smaller robots may be able to negotiate cramped spaces allowing exploration of areas that larger robots may be unable to reach. However, smaller robots may also have more difficulty in traversing terrain, less volume for computation, communication, and power storage, and require increased precision in manufacturing.
- *System Longevity* – Running time is crucial to some applications, especially in situations where retrieval of the robots may not be possible and the distributed team must travel to and from the area of activity. The energy density of current battery technology requires that significant percentage of the robot's volume be dedicated to energy storage. This constraint can be hard to satisfy with respect to the robot's physical size. Some robots overcome this limitation by tethering while others are able to self-recharge, either from the environment in the form of solar energy [52] or auto-docking [53], or by being recharged from another robot [42]. The inclusion of these technologies requires additional electronics on the robot.
- *Communication* – Communication is a key component of distributed robotic teams. Wireless networking in the form of the 802.11, 802.15, Bluetooth, and newer emerging protocols such as 802.16 and 802.20 allow for the team to communicate without wires and without rendezvous. Research in ad-hoc networks and mesh networking is very applicable here as many robots in a distributed robot team may not be directly within communication range of each other. When this is the case, there must be a routing protocol in order to get a message to a particular robot. This can be accomplished with simple broadcast of all messages, but it does not scale well to large numbers of robots as it will eventually saturate the available bandwidth. Smarter routing protocols [46] must be used in order to minimize the amount of bandwidth used. Recent networking research in the areas of sensor networks [28] and ad-hoc networking impact this area as well.
  There are also other methods of communication which are in use. Some robotic teams also communicate by visual manipulation through LED lights [22], or can recognize each other through vision [12]. Usually a single network type is used over all robots in the system. Some technologies require additional electronics or are simply too large or expensive to integrate into smaller robots. In addition to the electronic components of the communication, the size, type, directionality, and number of antennas used greatly impacts the range and therefore usefulness of the wireless communication. Proper selection of communication methods can greatly extend the robot's operational lifetime.
  Another aspect of the communication of distributed robot teams is connectivity. It is preferable to keep all robots in the distributed team be able to communicate with all others at all times. Due to the mobile nature of each robot in the team, it may be that a robot's movement needs to be limited in order to maintain the link between the other members of the team. This problem has been addressed in research in a number of ways. One way is to define a set of robots to be the network nodes, and have their only goal to keep the network available. This is undesirable because it sets aside resources which could otherwise be used to perform tasks. Another option is to compute the connectivity for the entire group and detect disconnects, and make movements to reestablish connectivity.
- *Mobility Methods* – The mobility of the robot platform is one dimension that should be considered as well, especially if the robot team is going to be deployed in a real-world environment or a simulated real-world environment [25]. Larger wheel bases and stronger motors are required in these situations to effectively navigate, but are at odds with a small robot design and require more energy for movement. Some type of dead-

reckoning sensors may also be included in the motors allowing for localization. While wheels are not exclusively used in mobile robotics [1,45,47], they are the most prevalent. Stairs and multi-level obstacles are also a mobility consideration, as they are quite common in real-world environments. There are some interesting approaches that have been presented in this area, including wheels that change their size [14].

- *Sensing Capabilities* – Finally, sensing is a key component of any robot's design. Many sensors which are high-fidelity such as laser range-finders, sonar sensors, or camera platforms are also fairly large in size. Some locomotion sensing would be desired for the locomotion, although recent advances in visual odometry [9,34,43] may be used. Sensors also require power, meaning that a large number of sensors may reduce the run time significantly. Sensor expansion is an important design consideration as well. Processing of sensor data may require significant computational power. This computational power is yet another trade-off. If it is onboard, it may require a larger chassis or more energy reserves. If it is done off-board, then there is more reliance on network connectivity and bandwidth.

The Scout 2000, shown in Fig. 1, is a small robotic platform designed for distributed robotics. At 40 mm in diameter and 110 mm in length, the Scout was compact and had little volume for expensive computation or sensing capabilities. The speed of the platform is minimal compared to other larger platforms. However, it also contained an innovative jumping mechanism which allows it to climb stairs [57] which increased both its mobility and cost. The Scout 2000's primary locomotion was accomplished using two wheels, although it could be ballistically deployed from a larger robot, the Ranger. These two robots are part of the Scout team [20]. There were many variations on the Scout platform, designed to give teams of Scouts increased usefulness through innovative sensing and locomotion methods. Some of these methods included the "Actuating Wheel Scout" which was a specially modified Scout that had the ability to increase its wheel diameter from 40 mm to around 120 mm. Other variations included a deployable grappling hook, a miniature pan-tilt camera head, geiger counters, or IR emitters which could illuminate a room for the whole team. The small size of the Scout allowed only minimal onboard processing —instead, relying on a wireless video broadcasts to be processed by a remote monitoring station. Coordination of teams of Scouts was accomplished using a CORBA-based architecture [37] which would dynamically load balance mission objectives across a number of systems.

The Explorer robots, shown in Fig. 2 are a newer evolution of the Scout 2000 robot platform which is designed primarily for educational use and research purposes. They are designed with expandability in mind as well as communication and mobility.

The Explorer includes a larger amount of processing in the form of an ARM processor, and expansions for more sensing connections to the device. It also has an articulated tail in order to change the angle of sensors mounted onboard the robot. The Explorer can broadcast analog video as well as utilize an integrated Bluetooth device, allowing Explorer robots to relay information between robots. While larger than the Scout 2000, the Explorer is still small

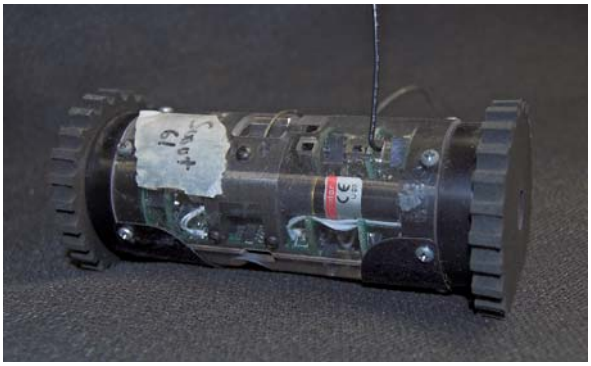

Distributed Robotic Teams: A Framework for Simulated and Real-World Modeling, Figure 1
A Scout 2000 Robot

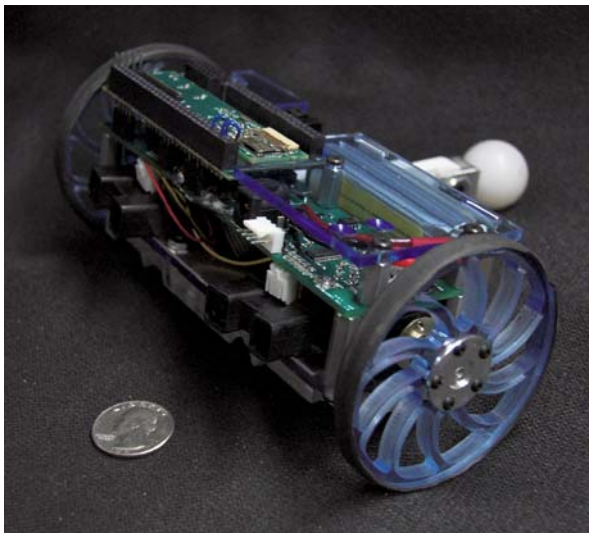

Distributed Robotic Teams: A Framework for Simulated and Real-World Modeling, Figure 2
An Explorer robot

Distributed Robotic Teams: A Framework for Simulated and Real-World Modeling, Figure 3
A TiTAN robot

at 17 cm long and 7 cm in diameter. The Bluetooth compatibility also allows control of the Explorer through standards compliant devices such as the Nintendo Wii-mote.

The TiTAN, pictured in Fig. 3, is a platform developed between the Scout 2000 and the Explorer robots. It is of significantly larger size, sacrificing the small form factor for more features in all of the other dimensions of robot design. TiTAN robots measure 37 cm long and have a 13 cm diameter, 35 times the volume of the original Scout 2000, and 7 times the Explorer robots. In exchange for this increased size, the TiTAN has significantly better sensing, allowing for swappable bays of sensors which include independent tilt control. The increased size allows for payloads on the back and top.

Mobility is also increased, as the much more powerful motors of the TiTAN allow it to travel at almost 6 m/s over flat terrain, and even handle towing a payloads in excess of 100 kg. The TiTAN, like the Scout 2000, also has a set of actuating wheels which increase its wheel base and allow it to traverse over rocks and obstacles which would otherwise stop it due to its low ground clearance. Communication is achieved through 802.11 wifi with a diversity antenna system which provide significantly longer range than bluetooth both indoors and outside. TiTANs also have an optional 900 MHz antenna, allowing them to receive video from the Scout 2000 and other robots on this band. Processing onboard the TiTAN uses a Pentium based PC104 stack. The selection of this stack enables onboard vision processing and a path for future upgrades.

Robotic Team Composition

Whether the team will be homogeneous is often the first question when presented with designing a distributed team. The answer will impact what can actually be done with the team, as many task-allocation methods and distributed robotics methods require the use of a team which is homogeneous. Having a homogeneous team may make tasks more efficient, as any robot's sensing information can be substituted, and information sharing between robots may be more prevalent. Having a homogeneous team also makes it easier to keep a team of robots running – extra robots can be kept to be used while some are being fixed.

Heterogeneous teams, alternately, may be easier and cheaper to create, as existing robots that are used for other purposes can simply be combined into a single robot team. Such teams may also include multiple sensing modalities for a single area, enhancing the ability of the entire team by sharing information between members. Differing sensors on each member of the team may also be an advantage if one sensor is cost-prohibitive but may be only needed in some areas of an environment.

Having a heterogeneous team may also enable marsupial team mechanics, allowing for the benefits of small robotic platforms as well as large robotic platforms. The smaller, carried robots of a marsupial team generally have a smaller footprint and can traverse smaller confined spaces. They are also usually as many or more carried robots as carrier robots. This enables the task at the mission site to be completed in a more timely fashion than would be attainable. The smaller robot teams are slower over long distances as well, which can cause problems for very distant mission areas. Because they are limited in size, they also carry smaller energy stores and may not be able to even complete a round trip to the mission area and back without assistance.

In contrast, the larger robots in a marsupial team can travel very long distances due to their increased wheel base, and may be able to traverse more difficult terrain. Carrying the smaller robots, in total the energy cost of the entire mission may be lowered because of the efficiency of the carrier robot. Additionally, the marsupial device may allow for the recharging of the energy of the smaller carried systems, allowing for a much longer lifetime of the entire system. Having extra space and operating power can also be useful in processing the large amount of data. Powerful computing resources on the larger carrier robot may be able to coordinate communication or store large amounts of sensor data, and also provide a general overview of the scene. More costly and high-fidelity sensors as well as sensors with a higher demand on computing resources may be kept on the carrier robot. Another option is for the larger robot to relay communication back to an operator which would not be in the range of the smaller carried robots' communication systems.

Marsupial teams have been researched by many universities and experiments have been performed. One of the earliest examples of a marsupial team are the team of Sil-

ver Bullet and Bujold [40], which was used to evaluate the usefulness of a marsupial team. In this team, Bujold was tethered to Silver Bullet for all of its power needs, significantly reducing the ability of the marsupial pair. Further research was done with the Yellowjacket team [13], which used a static arm mounted on the marsupial carrier, allowing for the carried robots to dock and undock themselves. It allowed for six carried robots to be docked on the carrier at once.

Marsupial teams need to deal with challenges which are not present in either of the single robot systems alone. The most important of these problems is the storage and deployment process. Some marsupial pairs handle this issue by placing a beacon of some sort on the carrier robot, allowing it to be recognized and docked with. There are many examples of carrier robots which use vision along with markers on the carrier robot in order to either line up or dock [38]. A team of Scout 2000 robots was used in at the University of Minnesota which took advantage of their unique jumping ability to both store and deploy themselves onto a carrier robot [27]. More recent work focuses on using an earlier version of the Explorer robot in order to dock with a prototype robot transportation mechanism referred to as the "Saddlepack". The Explorer-like robots use orange markers which are situated around the location which needs to be approached in order to store the robot in order to position itself in the correct orientation and approach the docking station. Once close enough, dead reckoning can be used to drive onto the forklift and be stored [38].

One example of a complete marsupial system is the Saddlepack system developed by the University of Minnesota. The SaddlePack is a modular system of interconnected robotic platforms. Three basic building blocks make up the SaddlePack system; two relatively large, two-wheeled robotic platforms, the TiTANs, which cooperatively maneuver the system, and a single immobile superstructure that suspends between the TiTANs and serves as communication base, power reserve, and attachment point for modular systems. The TiTANs themselves are actually specially-equipped TiTAN robots. A concept drawing of the SaddlePack can be seen in Fig. 4.

The SaddlePack's ability to support modular functional devices is an expansion of capability. The first of these modular components to be developed is a Modular Robotic Docking Station, MRDS, which is capable of carrying, deploying, retrieving and re-energizing up to six relatively small robots, the Explorers, for expanded search and explore missions at a lesser energy use rate. The SaddlePack can hold up to three of these stations for a total of eighteen extra Explorers for added utility.

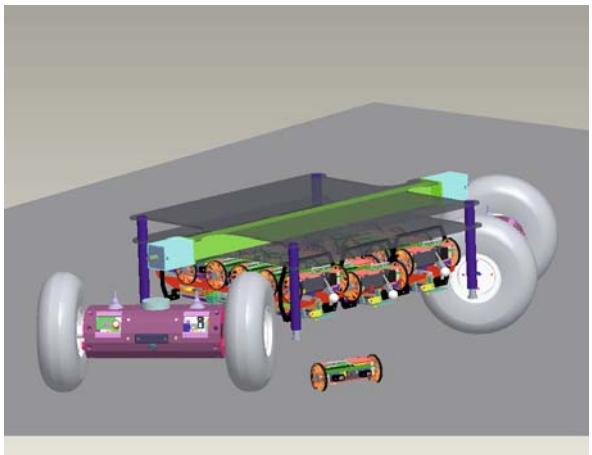

Distributed Robotic Teams: A Framework for Simulated and Real-World Modeling, Figure 4
Concept drawing of the SaddlePack

Ideally, the SaddlePack system will operate autonomously for a minimum of 10 hours due to its large battery store. This means that it must be capable of continuous movement, communication, and computation for that time. The docking station must also have adequate battery power for recharging the deployable systems within the Modular Robotic Docking Station and any future modular functional unit demands.

The docking station itself lacks mobility. The design of this system allows for two pre-existing robots, the TiTANs, to dock with the docking station. This mechanism is general enough that a variety of robots can provide the requisite mobility. The MRDSs are designed to lift robots to and from the ground plane, which is 10 cm below the base of the station to allow for ground clearance, and shuffle the robots within the secure space to replenish their power supply and transport them efficiently between mission areas.

The Saddlepack system is a good example of a marsupial robotic system which includes both the smaller Explorer robots which can be deployed and retrieved by the MRDS. It also includes the possibility of additional mobility of the TiTAN robots which carry the MRDS to the mission area and have a large number of sensors and processing power. This provides in whole a large number of different mobility scenarios. The robots within the MRDS can be carried to the mission zone by the TiTANs, where they can then be deployed to complete tasks which require mobility in cramped spaces. The TiTANs themselves can detach and use their higher processing power and sensing capabilities to further explore the environment. In addi-

tion, the MRDS acts as a relay point. The coordination of all of the separate members of this robotic team is a daunting task.

The Saddlepack concept's core tenet is modularity. The individual MRDS system can be transported by the TiTANs or mounted to a number of other larger commercially available or custom platforms. This enables flexible design and reconfiguration of the team. For example, a modified iRobot Packbot could be used to transport an MRDS system up a flight of stairs, a task that could not be completed with the TiTANs.

Software Modeling and Simulation

Simulating distributed robotics has its own unique set of challenges and advantages. Some of the advantages of simulation are particularly useful in distributed robotics where the complexity of experiments increases. However, simulating such experiments also includes challenges that are not present when simulating a single robot platform in an environment. In addition, simulating marsupial systems presents its own set of unique issues in addition to the issues caused by simulating a multi-robot system.

Simulation is used extensively by robotics researchers for a variety of reasons. First, simulation reduces the cost of algorithmic development for distributed robotic teams. Additionally, simulation removes many of the issues associated with maintaining hardware, noise in sensors, difficult environments, and presents an ideal world that allows for a "theoretically fair" comparison of approaches. It is even more appealing in the field of distributed robotics – the addition of multiple other robots that now must all work simultaneously causes maintenance to take even more time than with single robot platforms. The time for setup of experiments and programming of the robot systems is also multiplied by additional robots. In light of these expanded issues, it is reasonable to be drawn to the use of software modeling tools.

Distributed robotics brings its own set of challenges to modeling. Communication, which is nonexistent with a monolithic robot, is very important when working with multiple coordinated robots. In addition, software modeling of wireless communication is especially tricky due to the physical properties of wireless signals. Many simulation systems do not include a communication component, and many papers may simply assume the existence of a network which can communicate a set of shared data, which can cause distributed behaviors that work in simulation to have unexpected issues in real-world environments. Even in this case, the complexity required to exchange data between multiple running programs may not be reduced in a simulation environment. Along with these modeling issues, simulation software must be made more complex in order to handle multiple robots at once, possibly including concurrent programming when none was used before. As the number of robots used in simulation is raised, the computing power required to simulate the aforementioned robots in an environment increases minimally in a linear fashion. This makes it difficult to simulate large numbers of robots such as those deployed in a large building or over a large area.

There are a number of software packages available for researchers to use. One of the most prevalent in use today is the Player/Stage software [8], which is in active development. Player presents a set of interfaces to the programmer of a robot, which can then be used to control a variety of robots which would support that interface. The `position2d` interface, for example, represents a robot's position on a 2d plane, which can then be manipulated and updated through commands to the interface. The commands are then translated by the Player server software, contacting the driver for the specific hardware the server is running on in order to move the robot. The same program can be used unchanged for any robot which supports the `position2d` interface. In addition to the Player interfaces, a server which represents a simulated set of robots and environment exists in the Stage software. The Stage software simulates 2 1/2 dimensions in relatively low fidelity. This level of fidelity is used specifically in order to lower the amount of processing needed in order to simulate large numbers of robots at once. In this way a robotics researcher can write a Player client program and test it in the Stage simulator before running it on the actual hardware. The Player/Stage software is also easy to extend due to its open source, allowing specific drivers to be written for custom devices.

Player also can communicate with the Gazebo simulator, which provides a full 3-D environment simulation for robots. This provides a higher fidelity than the Stage software, allowing for much more complex interactions. Unfortunately, it also takes much more computing resources than Gazebo in order to perform its physics calculations. It also allows for marsupial actions to occur even without explicit support in the simulator – marsupial carriers actually physically pick up the carried robots and the interactions are modeled through the simulated physics. As the simulator also connects with Stage, the same programs which are programmed for Gazebo or real robots can be used when connecting to Gazebo. Gazebo has a level of complexity which makes it harder to create new simulated devices, but it is still open source and therefore extendable by the programmers.

The Microsoft Robotics Studio (MSRS) was introduced in 2007, and it also contains a simulation system. The simulation used in MSRS is of much higher fidelity than the Stage simulation software, including a full 3-D physics simulation of the environment. MSRS uses a commercial physics simulator which is highly accurate. This increases the accuracy of the situation but also causes a significant increase in computing resources used for the simulation similar to Gazebo. However, specialized hardware is available which the commercial simulator can take advantage of. This may allow for a larger number of objects to be simulated and make the MSRS more suitable to simulating robotic teams. The MSRS also presents a set of interfaces to the programmer through its service-based architecture, but they currently are at a much lower level of abstraction than the Player interfaces. Users can build their own services on top of existing services, as well as program new drivers which provide access through existing services. MSRS communicates through a REST-like HTTP-based protocol.

These are only three of the many simulation packages used for robotics available. They are the most popular, in part because they are both free to use. A number of other simulation systems are used as well, with various advantages and disadvantages [11]. Some researchers also prefer instead to create simple simulations using the MATLAB commercial mathematical programming language.

Marsupial Modeling

Marsupial systems present all of the challenges present in distributed robot simulation, but also have unique challenges which arise when coordinating two robots. Each marsupial system handles the retrieval and deployment of the stored robots in a different way, causing problems when trying to create a generalized solution to a problem in which marsupials are present. It is desirable then to create a high-level abstraction which can be used to deploy and store robots in a marsupial situation. We present one such model in this section.

There are very few robotic simulation systems which include marsupial actions. The easiest example from this class of simulators are full physics simulators such as Gazebo [8] and the one included in the MSRS. However, in current simulators with physics, there are still no high-level marsupial abstractions – the entire sequence of manipulators or motors would have to be activated individually in sequence each time a marsupial action was desired. In addition to these issues involving the modeling of a marsupial mechanism, the implementation of full physics simulation requires much more computing power than the relatively low fidelity environment provided by software such as Stage. This can make it difficult and sometimes impossible to simulate marsupial systems, which rely on the physics of the manipulators in order to store and retrieve robots.

When considering a general model for marsupial mechanisms, very few mechanical considerations need to be made. It should be general enough to possibly describe many different systems while being specific to marsupial actions, as opposed to movement, manipulation, or general sensing actions. While it may be the case that zero or more actuators are used to perform a marsupial action, the desired action is more high-level. Simulating this general device allows us to separate the marsupial actions from actual hardware, removing some of the difficulties in performing marsupial actions. At the same time, it allows for the desirable portability of algorithms by providing an abstraction that may be used by multiple different marsupial hardware systems.

The level of abstraction present here is fairly high, so there are few actions which should be possible for the marsupial abstraction. The first and most important action defines the group of mechanisms, that is *storing* and *deploying*. In our abstract model, we do not care how this is accomplished, only that it can occur. Therefore we consider two states of the model: it is either *open*, ready to store or deploy, or *closed* and it is not ready.

Because we are not interested in the specifics of storage, we focus on what can be stored. All physical systems have a maximum size of object which can be stored due to the physical limitations of the device that we name its *door size*. Objects larger than the door size cannot be stored by the mechanism. Marsupial mechanisms also store a limited amount of objects due to physical constraints on the storage area, which we call the *storage capacity*. Some systems also can sense how much of the storage capacity is currently occupied.

This general model for a marsupial mechanism because it covers a wide range of possible marsupial devices. By limiting the requirements for a marsupial system, we expand the set of devices which can be represented, hoping to impose nothing beyond the definition of a marsupial robotic system. This model was implemented in the Player/Stage software [26] and is currently being used to simulate the prototype SaddlePack marsupial system described in Sect. "Robotic Team Composition".

Tasks Well–Suited to Distributed Robotics

There are a number of tasks which are well-suited to a distributed robot team. Some of these tasks may be possi-

ble to complete with a static deployment of sensors, but they may be suboptimal. Others may be able to be performed in a shorter amount of time with a set of distributed robots, but the addition of more than one robot may require communication throughout the team or distribution of the work load, introducing new issues which must be addressed.

Here we discuss two different tasks which can be accomplished without a distributed robot team. The first is mapping or coverage of an area. These two problems are considered together because they are very closely related. Mapping and coverage are both completable by a single mobile robot, but utilizing a distributed robot team potentially decreases the time to complete both of these tasks. The other task which is considered is optimal placement of sensors with respect to observing activity. This task is possible to complete by placing static sensors, but static sensors cannot respond to dynamic changes in the activity being observed. When considering either of these tasks, we propose improvements and challenges which must be considered specifically for a marsupial team as well.

Mapping/Coverage

Mapping is particularly well suited to distributed robotics, and has been well researched in the field [10,32,48]. Mapping is of particular use in situations where the environment has recently changed and an exhaustive search is desired, for example with in a collapsed building or in an unknown structure. It is generally considered a basic problem in the field of robotics. A distributed robot team can finish the task in a shorter amount of time by either coordination of search areas, or by a distributed behavior which naturally causes the team to explore different areas independently.

Using a distributed team also brings new challenges to the table when mapping. Because multiple robots and therefore multiple sensors are now being used to sense the environment, some type of relationship between those sensors is needed in order to combine the maps into a full map is required. This is generally referred to as the *map merging* problem. Also, exploring an area to be mapped now also contains the other members of the team themselves, which can corrupt the map with spurious obstacles. The problem of conflict management must also be addressed, as robots may need to travel to different areas but may block each other in doing so – consider the case where one robot needs to search a room at the end of a hallway, and another robot is exiting the same hallway. The methods used must employ a protocol to resolve these deadlocks.

In the most basic type of method for distributed mapping, the occupancy grid which is used in mapping with a single robot is used and simply extended to the multi-robot situation. Assuming that global communication is available, the global map with the occupancy grid can be filled in independently by each of the exploring robots and each robot can select the closest longest frontier in order to expand the amount of cells which are explored. This method has problems if the robots start in a similar location and orientation, because the same frontier cells will be chosen by each robot. This may cause the exploration to take the same amount of time as a single robot which is undesirable. Some methods can be used to introduce randomness or choose target frontiers in a different way to reduce this behavior, but it cannot be removed without communication of the target frontiers also. This method does not scale well because it requires communication of the current task at regular intervals which will eventually overload the communications medium if enough robots are added to the team.

Methods which communicate the frontiers which are being traversed have been improved upon from there. Maximizing the information gain of the relative robot's assigned frontiers is one improvement that can be made [6]. The relative robot localization loss in each area is considered [30] in order to facilitate map merging at the end of the method. Limitations can also be added to improve localization, such as incrementally approaching the frontier in order to maintain visual contact [21]. Task assignment methods such as auctioning and market economies can be used as well [60].

Another type of mapping which is considered for single robots is the semantic place map. This type of map usually doesn't take the form of a traditional map, but is more similar to a subway or large train system map in that is conveys the connections between places which can be easily recognized. This type of map is usually built specific to a sensor such as a omnidirectional camera or a high-resolution laser range-finder which has a high density of data. This high volume of data is required in order to sufficiently disambiguate the specific locations in the case of highly similar locations. The locations are connected in a directed graph representing the travel required by a robot in order to travel from one place to another. This method is extended to the distributed robotic team by ensuring that the required sensor for identifying the nodes of the semantic map is present on all of the members of the team. As such, this method is more well-suited to a homogeneous team of robots than a heterogeneous team. The information required in order to recognize each place is then distributed throughout the team in order to

facilitate localization within the semantic map. Some issues may persist if the locations are very similar to each other, but they can be disambiguated at the very worst by their location within the map itself. This may make localization of a new member of the team a longer process.

Recent developments in distributed exploration involve the expansion of random trees of traversable areas and combining and optimizing the trees instead of explicitly mapping the environment. This results in a map which is not only easily usable by the humans observing the exploration, but also by the robots themselves in order to use later for path planning and safe traversal of the mission space. The methods which are in use are based on Rapidly-expanding Random Trees, with new research being done in Sensor-based Random Trees (SRT) [16]. These methods were first developed for single robot exploration, but have been parallelized in order to increase the efficiency of the search by cooperation and modified in order to avoid conflicts between differing members of the robot team. These methods also employ a frontier region similar to the occupancy grid methods, with similar selection problems being handled in much the same way as those methods. The robots build a SRT and then make that information available for others to use while creating their own trees. When the method completes, the trees can be combined into a usable map for other robotic teams which may need to explore the area.

While distributed mapping may construct the map after the exploration is complete, simultaneous localization and mapping (SLAM) methods construct the map while the robot is exploring the area. There are many different approaches to SLAM, and some have been extended to work in multi-robot situation in real time. Some methods are based on the initial work [15] simply combine all landmarks and localization together assuming adequate communication between all robots. This causes a problem as the set of landmarks increases in a large environment, because as some landmarks are removed in order to lower the complexity of the computation required, the landmarks may be sparse and a robot may be out of range. It also does not scale well to a large number of robots. Other methods focus on improving the local SLAM methods by using the other robots in the team as landmarks that are mobile. The simplest way to extend the traditional SLAM method [33] is to only have a single robot move at once, using the stationary robot as a landmark. This again is not reasonable in situations where the robot team is large, and some methods that allow for multiple or all robots in the team to move have been developed [50].

Search is a problem closely related to mapping, but not exactly the same. While mapping focuses on sensing one area of an environment at least once, coverage focuses on sensing over all of the environment. This may be performed in an unknown environment with a single robot, and can be thought of as a search method which will eventually find an object no matter where it is located in the environment. Coverage can be used trivially to construct a map such as the ones discussed with occupancy grids, but a search task may not require the creation of a map, and a complete map may not be sufficient to guarantee that an item searched for may be found. Coverage may also be thought of in terms of painting where a robot is required to paint an entire surface of an object. Optimizing the coverage results in a robot which crosses its path a minimal number of times.

Interestingly, communication is not a required element for many search strategies. Some methods [36] use potential fields along with recognition of other robots in order to search effectively areas that have nor been searched by having other robots and obstacles repel the robot and the frontier of the search attract the robot. However, this method itself acknowledges that the lack of communication degrades the performance of the search activity. Additionally, implicit cooperation methods are used, with some shared state which is discernible through local sensors. Some implicit communication involves using environmental markers such as those used by ants while exploring [31]. Random strategies can also be used within a large team of robots to perform search, and can be shown to be highly scalable without issues to run on a large team of robots [18].

Communication has been shown to increase the performance of search tasks however [51], so utilizing the communication available is desired. Simple methods use a global map or grid which can be communicated to all team members when areas have been searched [59]. The grid size of completed work may be at a much larger granularity as an occupancy grid, and some methods allow for two or more robots to cooperate in coverage of a single cell [24]. Yet another method [55] allows the robot to assign tasks for search only to itself as it discovers new areas, but other robots may be able to remove a task if the area has already been searched. This limited communication allows for robots to make search choices in parallel independently, increasing the scalability of this approach.

These limited communication methods may eliminate some duplicate coverage, but they do not completely eliminate duplicate coverage. More complex methods will share the current task which is being completed as well. This allows for each area of a map to be searched only once.

An early way of doing this was to assign a global coordinator which would collect the possible actions of each robot in the team, and assign each task to each robot in order to maximize global utility [54]. Recent advances however favor a distributed approach, with tasks being generated when new areas are discovered, and can be completed by any robot. This may take the form of an assistance request [17] when the area is particularly large, or auctioning using a method as described in Sect. "Team Coordination/Task Allocation". This task allocation process lowers the amount of communication required in contrast to the coordinator method which requires state and intent information to be communicated continuously. However, many of these methods do not scale well with the size of the robot team because of the large amount of communication needed. Some researchers are addressing this problem by improving on the current strategies by lowering the amount of communication which is required for each of them. One such improvement [49] involves the robots agreeing before the search starts on a strategy for search if communication is lost or is otherwise unavailable due to traffic constraints.

Another possibility which can be considered is static coverage. This is when the entire area is covered by the sensors of at least one robot. It is related in some ways to the art gallery problem [44]. This is a task which is almost impossible to consider for a single robot – it is specifically the domain of distributed robotics. Static coverage is useful in real-world situations for expanding a network of communication to its largest available network range while continuing to provide routing ability even in situations with failing sensors. It is also useful when the environment of the robots is changing rapidly and therefore must be mapped at a frequency which is higher. The placement of robots evenly across an entire environment also allows the completion time of randomly placed tasks to be lower than if the set of robots were initially concentrated on a single point.

Static coverage can also be achieved without communication by utilizing a potential field which repels the robots from each other and from large obstacles such as walls. This results in the robot team being spread throughout the environment. If the environment is a priori known, a coordinator robot may place the location of the robots ahead of time in order to maximize static coverage. The placement methods which have been developed for sensor networks are analogous to the placement needed here.

It is interesting to consider the implications that a marsupial team may have on exploration and search methods. Exploration may be enhanced specifically by a marsupial robotic team. One method would be similar to the centralized assignment method, with the marsupial carrier robot coordinating the actions of the smaller robots in order to efficiently explore the frontiers of the region. This method would effectively let the smaller carried robots explore the area faster than possible before because of the increased mobility of the carrier robots. An extension of this method may shift the control of the method to the carrier robots, having the small robots request a marsupial robot in order to travel distances which would take a long amount of time when using normal locomotion. This would allow any of the distributed exploration or coverage methods to be used without change on the network of carried robots, simply increasing the mobility over long distances.

Another method could be considered which attempts to center the larger marsupial carrier in the space and then a deployment phase happens, and the carried robots carry out their own exploration of the room, returning to the carrier robot when finished. This process would then be repeated in other spaces, or in parallel by multiple carriers with the set of robots they are carrying. In this method, the carrier robot may be used as a landmark for localization, reducing the total error in the map and decreasing the computational requirements for map merging. It is also possible that a higher-level survey of a room may be acquired when the marsupial carrier is entering the room, and subsequently the smaller carried robots would increase the information density in the map acquired by filling in the smaller spaces which can be sensed by a smaller platform.

The marsupial team also brings with it challenges which are not present in the general distributed robot exploration problem. The differing sensors on the carrier and carried robots in a marsupial pair must be considered in an extension of these methods to a marsupial team. The smaller robots also have a level of resolution for mapping and localization which varies by a large amount compared to the larger carrier robots. This may be exploited as in the second method as described above. In addition, the sensor levels of the smaller robots will be at a lower level than the carrier robots. This may be an advantage because two differing heights of maps can be created, but it may also be a detriment – maps which signal the openness of a particular spot for a carried robot may not be accessible for a larger robot. A good example of an obstacle which would pose a risk for the carrier robot but not for the smaller carried robots is a coffee table, which has a wide open space underneath and a solid top. This would not be traversable for the larger carrier robot but would be easy to navigate for the smaller robots.

Coordinating Teams to Optimize Observability of a Situation

Sensor networks are useful for monitoring an area for a variety of purposes. Static camera networks are utilized for monitoring traffic or security applications. Chemical sensors may be deployed throughout a production facility to monitor for chemical leaks. However, the effectiveness of the sensors is a function of how well placed they are. Traditionally, these problems can be characterized as variations on the "Art Gallery Problem" [44] where the task is to determine the number of cameras necessary and where to place them in order to monitor art exhibits in a gallery.

However, in dynamic situations, statically placed networks may miss critical pieces of information. This is especially critical when utilizing directional sensors such as cameras. Camera systems can be utilized for automated facial recognition purposes, human activity monitoring [3], and abandoned object detection [2]. However, if the cameras are not deployed in an effective configuration, the systems will operate poorly, potentially misclassifying activity in a scene or missing an abandoned object.

Distributed robotic teams offer an alternative to solving this type of camera placement problem. Robotic teams offer the ability to relocate the cameras as the situation changes, thus providing a continuously optimizing observation of the situation. The selection of the locations for the cameras first requires monitoring of movement within the scene. When the goal of the observation is to characterize human motion (walking, running, jumping, etc.), the best vantage points for the cameras may be perpendicular to the motion of the individuals in the scene. This can be complicated by the number of individuals moving throughout the scene in a non-uniform fashion. When the goal of the system is facial recognition, the cameras must be relocated to observe the trajectories of individuals moving along the focal axis of the camera.

The framework developed by Bodor et al. [4] offers one method of solving the camera placement problem with a distributed robotic team. In this approach, the goal is to place the robots perpendicular to the path that individuals moving throughout a scene take. However, in observing the path, there are additional constraints, first that the entire path be in the field of view, and second that the camera be as close as possible to the subject. In a perspective projection camera with a fixed focal length, the size of an object is proportional to the distance from the object. By placing the camera as close as possible, the object being observed is larger, and thus more information can be gained. A minimum distance d_0 can be determined based upon the characteristics of the camera and the paths that

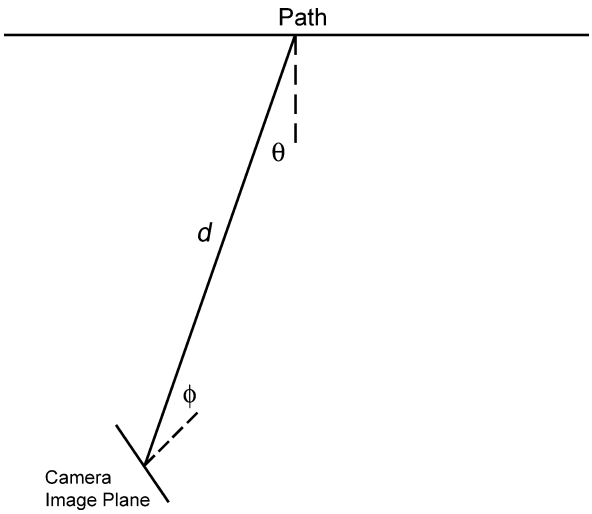

Distributed Robotic Teams: A Framework for Simulated and Real-World Modeling, Figure 5
Angles to minimize for each path

are being monitored. The distance between the ith camera and the jth path can be given by d_{ij} and must be at least d_0 away from any path in the distribution being observed.

In two dimensional space, the angle ϕ can be used to represent the angle between the focal axis of the camera and the line connecting the center of the camera's image plane to the midpoint of the jth path $(X_{\text{midp}_j}, Y_{\text{midp}_j})$ while θ represents the angle between a perpendicular axis run through the midpoint of the path and the image plane of the camera. These angles are depicted in Fig. 5. Considering the entire path distribution and the multi-camera system, this results in the following function which must be optimized with respect to the camera placement and orientation:

$$O_P = \sum_{i}^{\text{cams}} \sum_{j}^{\text{paths}} \frac{d_0^2}{d_{ij}^2} \cos\left(\theta_{ij}\right) \cos\left(\phi_{ij}\right) \ . \tag{1}$$

The individual robots are able to relocate themselves to the positions resulting from the optimization. These positions are derived by changing the variables from (d, θ, ϕ) to absolute positioning and orientation of the camera and path in the real-world $(X_{\text{cam}}, Y_{\text{cam}}, \text{pan}_{\text{cam}})$ using Eqs. (2), (3), and (4) for each camera-path pair. When Eqs. (2), (3), and (4) are substituted into (1), the value of the observability objective function can be determined for a given camera position and pose in world coordinates:

$$d_{ij} = \sqrt{\left(X_{\text{midp}_j} - X_{\text{cam}_i}\right)^2 + \left(Y_{\text{midp}_j} - Y_{\text{cam}_i}\right)^2} \tag{2}$$

$$\theta_{ij} = \cos^{-1}\left(\frac{T_1 + T_2}{T_3 \cdot T_4}\right) \tag{3}$$

where

$$T_1 = \left(Y_{\text{startp}_j} - Y_{\text{midp}_j}\right)\left(X_{\text{midp}_j} - X_{\text{cam}_i}\right)$$
$$T_2 = \left(X_{\text{startp}_j} - X_{\text{midp}_j}\right)\left(Y_{\text{midp}_j} - Y_{\text{cam}_i}\right)$$
$$T_3 = \sqrt{\left(X_{\text{midp}_j} - X_{\text{startp}_j}\right)^2 + \left(Y_{\text{midp}_j} - Y_{\text{startp}_j}\right)^2}$$
$$T_4 = \sqrt{\left(X_{\text{midp}_j} - X_{\text{cam}_i}\right)^2 + \left(Y_{\text{midp}_j} - Y_{\text{cam}_i}\right)^2}$$

$$\phi_{ij} = \cos^{-1}\left(\frac{0.167\,[P_1 - P_2]}{P_3 \cdot P_4}\right) \tag{4}$$

with

$$P_1 = \left(Y_{\text{midp}_j} - Y_{\text{cam}_i}\right)\cos\left(\text{pan}_{\text{cam}_i}\right)$$
$$P_2 = \left(X_{\text{midp}_j} - X_{\text{cam}_i}\right)\sin\left(\text{pan}_{\text{cam}_i}\right)$$
$$P_3 = \sqrt{0.028\left[\cos^2\left(pan_{\text{cam}_i}\right) + \sin^2\left(\text{pan}_{\text{cam}_i}\right)\right]}$$
$$P_4 = \sqrt{\left(X_{\text{midp}_j} - X_{\text{cam}_i}\right)^2 + \left(Y_{\text{midp}_j} - Y_{\text{cam}_i}\right)^2}.$$

Methods of constrained nonlinear optimization allow for the parameters $(X_{\text{cam}}, Y_{\text{cam}}, \text{pan}_{\text{cam}})$ to be derived. This framework can be extended to operation in three dimensions by including two additional angles in the optimization.

Teams of robots can also be used to identify the sources of potential biological, chemical, or radiological contaminants. In [7], a framework is developed where a team of distributed robots can be deployed to explore an environment. As the robots explore the environment, the individual robots take readings of airborne chemical contaminants. These readings are used to estimate the source location and quantity of the chemical. In doing so, these measurements can be used to predict where the plume will travel in order to warn and evacuate those that may be in the path of the plume.

Tasks for Optimizing Distributed Robotics

While many tasks can benefit from a distributed robotic team, the team itself introduces issues and optimizations must be tackled. Many of these tasks are directly related to other fields, such as the issues of communication and routing through a set of mobile communication nodes dispersed throughout an environment. We discuss two tasks which arise from the use of a distributed robotic team. The first is related to the longevity of a robotic team and utilizes the specific properties of a marsupial team in order to achieve the optimization of the team. The other is task allocation, which is related to work in the intelligent agents domain in auctions. The impact of task assignment is considered in the context of marsupial distributed robotic teams as well.

Coordinating Teams to Maximize Longevity

Extended coverage of an operational environment by a distributed robotic team is essential in minimizing risks to humans that have to enter a hazardous area. In order to do this, the robots in the area must have access to sufficient power to remain functional for the mission duration. While this could be as simple as integrating a larger battery system, batteries still offer only a finite power supply and it may be impossible to predict how long a robotic team may be required. There are three options for distributed teams of this nature to remain effective:

1. *Infinite Power* – One way to avoid this difficulty is to utilize tethering to provide "infinite power." This option is attractive for long term deployment, however, tethers can be difficult to manage, require extra coordination/larger motors/power expenditure for movement and relocation, and can snag on objects in the environment trapping the robot in a specific place.
2. *Energy from the Environment* – In extreme conditions, it is necessary for the robot to be designed such that it is able to extract all necessary energy from the environment. This is very common in cases of interplanetary exploration. Here, solar panels are utilized to obtain energy for recharging on board battery systems. This is extremely effective in these scenarios, but may not be effective when responding to hazardous incidents where chemicals in the air, ash from fires, or other debris may coat and block solar panels. Alternatively, robots such as Slugbot [29] are able to extract energy from the environment by identifying sources of sugar which it can use to create a chemical reaction to generate power. This approach while interesting, also has limitations, in that energy is expended foraging for sugar and sugar must be deployed in the environment.
3. *Resupply* – The third approach for extending operational longevity is for the robots to autonomously resupply themselves. This can be done by autonomously docking with fixed charging stations or with mobile docking stations. Fixed docking stations can reduce the complexity of the problem and are found in a number of environments [19]. However, when responding to emergency situations, it is unlikely that there will ex-

ist fixed docking stations, and if they are present, that they will be functional.

Given that the distributed robotic team must be able to operate in arbitrary environments, the problem of longevity can be reduced to the following. Given n distributed robots and d mobile docking stations, how can those d mobile docking stations be distributed in order to maximize the operational longevity of the n robots?

In order to solve this problem, first we define a robot as a system that can be represented as a location (X, Y), orientation θ, velocity V and available energy E. A robot's position at a given time (t) can be represented by the following equations:

$$X(t) = X(t - \delta t) + V\delta t \cos(\theta) \tag{5}$$

$$Y(t) = Y(t - \delta t) + V\delta t \sin(\theta)\,. \tag{6}$$

Thus, the possible locations the robot could be at in a given time can be represented by a circle centered at the robot, with a radius equal to the distance travelled (velocity x time). A mobile docking station utilizes a similar representation to determine where it can be within a given time. The problem then becomes an optimization one. Where can the docking station move so that the most robots in need of power can arrive there at a minimum power expenditure?

Each robot R_i, maintains an estimate for its position $(R_{i_X}(t), R_{i_Y}(t))$, velocity R_{i_V}, and battery life R_{i_E}. The cost to return to a docking station D_i located $(D_{i_X}(t), D_{i_Y}(t))$ at time t can be given by the following equation:

$$C(R_i(t), D(t)) = xe^{x-f} \tag{7}$$

where:

$$x = \frac{\frac{d_i}{R_{i_V}}}{R_{i_E}}$$
$$d_i = dist(R_i(t), D_i(t))$$
$$f = 1\,.$$

Here *dist* is assumed to be the Euclidean distance, however *dist* can be any distance function which in general will be the shortest known traversable path.

Let S_r represent the set of robots that are in need of recharge. There is a number of ways in which this set can be derived. The most straightforward way is for each robot to maintain an estimate of the location of the nearest docking stations that is periodically relayed through a wireless network. The robots can then estimate the amount of energy that is required to return to a docking station. As their onboard reserve is reduced, they can periodically broadcast a message that they are reaching a critical level of power.

This reduces the problem to identifying which movements of the docking station will minimize Eq. (8) at time $t + \delta t$:

$$\sum_{i \in S_r} C(R_i(t), [D_X(t) + D_V \delta t \cos(D_\theta)\,, \quad D_Y(t) + D_V \delta t \sin(D_\theta)])\,. \tag{8}$$

However, simply minimizing this function will result in a global solution that may not be optimal. In order to increase the utility of this function, a clustering method is applied to the members of S_r in order to divide it into reasonably sized spatial groupings. A prioritization is then applied to these clusters and the docking station will then only concern itself with the members of that cluster.

Team Coordination/Task Allocation

Coordination of a team of robots is a key component to creating a usable distributed robot team. Significant research in the field is based on methods developed in the area of intelligent agents, as each robot in a team can be thought of as an independent agent. Task allocation can be done a priori by a single scheduler for all tasks, but this is not desirable as it does not allow for changing conditions such as the loss of a robot or obstacles getting in the way resulting in a task being handled by a robot which is unable to complete the task. Methods have been developed based on auctioning which distribute the tasks fairly to all robots and attempt to approach optimality by minimizing path length, energy cost, or total time of the tasks.

Using auctions to distribute tasks among a group of robots is a common theme in the research in task allocation [39]. In this type of framework, the robots act as agents which are bidding on a task for completion. It is setup as a first-price reverse auction. One task may be auctioned, or multiple tasks may be bid on collectively. This allows for tasks which have good synergy to be completed together, for example two tasks which have completion points very close to each other, to be more optimally assigned. In this auction environment, tasks may be ordered by priority, auctioning off the most important tasks first. These tasks will receive more of the available resources of the entire team of distributed robots because less robots are committed to tasks already. Bids are computed based on the type of optimization preferred: for energy cost, a hilly terrain may require a higher bid because of the difficulty in traversing.

One example of task allocation in a distributed robot team is presented [41]. In this auction-based task alloca-

tion method, one robot is committed to all tasks at the beginning of the auction, and serves as an auctioneer for the tasks. One task at a time is auctioned to the other robots in the team. Tasks are ordered by priority in order to give the first tasks more attention. In this method, after any task is completed, all robots then rebid for tasks. This allows tasks which may be easy to complete after one task is completed. If no bids are received for a task, it may be dropped because no robot can complete the task. This method is particularly useful in dynamic environments, because as tasks are completed, some areas of the environment may be blocked off by other robots or moving obstacles. The rebidding process allows for those tasks to be completed by having them switched to another robot. The re-auctioning rounds include tasks that did not receive bids in the previous rounds, as it may later be possible for these tasks to be completed. In this framework, new tasks can be entered into the system by assigning the task to any of the robots as tasks which had no bid in the previous round, and they will be automatically completed by the robot which is most well suited to complete the task.

It has been shown that even in the simple case of homogeneous tasks and homogeneous robots, task assignment is a NP-hard problem [35]. This makes task allocation especially important when a heterogeneous team of robots is considered. Some members of the team may not be able to complete all of the tasks which need to be completed by the group. More able or powerful members of the group may be able to complete multiple tasks with a lower cost than other members. This makes the problem of assigning tasks much more complex, because the tasks cannot be simply assigned to any robot.

Marsupial robotics adds again this complexity, because tasks can be assigned to heterogeneous robots of the team, and the hybrid and changing nature of the team must be taken into account. One method would be to consider each robot separately and use a previously discussed method, but this is very suboptimal because it does not exploit the advantages which emerge when a marsupial pair is considered. The marsupial pair has enhanced mobility due to its changing size and sensing abilities. The advantages of the marsupial pair should be exploited in order to complete the tasks assigned in a more optimal way than the standard methods. There are two simple extensions to the protocol presented here which may be considered and are presented here.

With communication, each marsupial pair can negotiate as a team in order to place bids for task in the auction systems. This allows for the marsupial team to allocate its resources better. The complicating factor of the consideration that each marsupial carrier as well as each carried robot are potentially part of multiple marsupial pairs, causing there to be *carriers * carried* bidding entities on top of the individual entities for each carrier and carried robot. There is also an issue that must be addressed considering the order of the bidding. If a marsupial pair completes a bid, there is a significant change in the resources available to the entire group. This changes the bids of all of the marsupial pair combinations significantly and many must be recomputed, at least all of the pairs which include the marsupial carrier which has bid and secured the task for completion. This computation may cause the ability of the team to complete the task allocation within a reasonable amount of time to be hampered significantly, which must be considered when proposing such a system.

Another method to be considered is for the bidding carried robots to be able to generate tasks, which would include being deployed at a particular spot. This method would require the carried robots to be able to compute the traversable area for the larger carrier robot. However, this method may be suboptimal as these tasks may be specific to a marsupial pair. It also may be advantageous for a carrier to hand-off the carried robot to another carrier in order to complete the task faster. This makes this method unreasonable because of the large number of possible generated tasks grows without bound, and similarly to the previously discussed method may increase the computational requirements of the bidding process unreasonably.

Describing a task which is to be completed by a group of robots is also an ongoing problem in distributed robotics. In the simplest case, a task which is destined for a team of robots may be decomposable into tasks which are assignable to single robots. However, there may be tasks which are not decomposable, such as two pressure switches being activated at the same time or simultaneously carrying a large object. A task that requires more than one robot to perform needs special consideration. Tasks which may need an unknown number of robots are even more complex. It may be desirable to generate a task which only requires a single robot, but which can be amended to require multiple robots in the future.

Future Directions

Mobile robots are in use today around the world in order to solve problems in areas where it is not safe for humans to enter. They are used to disarm and detonate improvised explosive devices, and in collapsed buildings and natural disasters. Unmanned aerial vehicles are used in the field by soldiers to get a bird's-eye view of the battlefield. The advancement of the field of distributed robotics shows

promise in accomplishing these types of tasks with more efficiency than a single robot alone.

Distributed robotics is a fertile field of research with many open problems still to solve. It also gains from research done in related areas such as networking, intelligent agents, game theory, and control. Robots which are part of a team may require special hardware considerations. The team must be composed and communication between each of the team members must be considered.

Marsupial teams as a subfield of distributed robotics presents its own challenges and rewards. While each marsupial team can take advantage of two separate robot designs in order to maximize the potential of the distributed team, complex hardware and coordination issues must be considered. Marsupials must be considered specially for simulation of the behaviors which they exhibit, and in order to study the effects of differing parameters on the marsupial setup. A marsupial team also greatly increases the challenge of task allocation among the members of a distributed robot team. In exploration and coverage, the marsupial robot can be exploited to utilize the specific features of the multi-resolution sensing. Marsupial teams are a deep subfield of distributed robotics with possibly wide-reaching results yet to be researched.

As hardware technology advances, the improvements will have a large impact on distributed robotics. The trend toward smaller sensors and cheaper designs make it much more feasible to create a large team of robots which must then work together to complete a goal. Advances in batteries which increase energy density and operational lifetime will enhance the functioning lifetime of distributed teams. At the same time, wireless technologies such as WiMax and (other wireless technology) which both increase the bandwidth and range of communications for devices will increase the operational range of robotic teams. These advances will make distributed robotics even more applicable as robots become easier to produce in quantity and can communicate over a longer range. The ubiquity of robotics in the world today may be changed to make distributed robot teams ubiquitous in the near future.

One problem which must be considered is the issue of control. While increasingly tasks which are given to distributed robotic teams are autonomous and teams may be able to generate their own tasks through programming for a specific purpose, the creation and assignment of tasks in a short time frame by a non-expert needs to be addressed. Currently the most common form of control in the field by robot technicians is tele-operation, either in direct view of the robot or aided by the robots sensors or both. This form of control will not scale to distributed robot teams very well. While it is possible for robots being tele-operated to cooperate, it is very unwieldy for the operators and fails at a much higher rate of incidence. In most cases, each single robot must be controlled by a operator which is dedicated to the control of the specific robot in the field. In cases where a large environment must be searched and other tasks must be performed in a short time frame, this is simply not possible. The development of algorithms and interfaces which allow for the control of a large number of robots in a distributed team by a single operator must be developed. The development of these interfaces requires a combination of many of the elements which have been presented here; a level of autonomy of the robots after tasks have been assigned must be expected. The longevity of such a deployed team should not be a burden considered by the operator of a team. The tasks which are assigned by the operator may need to be shifted to a robot which is available and able to perform the task – the interface itself may be task-centric with tasks being defined by the operator and the robot team deciding which robot is best suited to the task. It is certain that this issue with distributed robots' teams will be present in the future, as robots become smaller and are produced in higher quantities.

Acknowledgment

This work has been supported in part by National Science Foundation through grants #IIS-0219863, #CNS-0224363, #CNS-0324864, #CNS-0420836, #IIP-0443945, #IIP-0726109, and #CNS-0708344

Bibliography

Primary Literature

1. Albro JV, Bobrow J (2004) Motion generation for a tumbling robot using a general contact model. In: Proceedings of the 2004 IEEE International Conference on Robotics and Automation, IEEE, New Orleans, pp 3270–3275
2. Bird N, Atev S, Caramelli N, Martin R, Papanikolopoulos N (2006) Real-time, online detection of abandoned objects in public areas. In: Proceedings of the 2006 IEEE International Conference on Robotics and Automation, Orlando, pp 3775–3780
3. Bird N, Masoud O, Papanikolopoulos N, Isaacs A (2005) Detection of loitering individuals in public transportation areas. IEEE Trans Intell Transp Syst 6(2):167–177
4. Bodor R (2005) Multi-camera human activity recognition in unconstrained indoor and outdoor environments. Ph D thesis, University of Minnesota
5. Broggi A, Caraffi C, Fedriga RI, Grisleri P (2005) Obstacle detection with stereo vision for off-road vehicle navigation. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, vol 3, p 65

6. Burgard W, Moors M, Stachniss C, Schneider FE (2005) Coordinated multi-robot exploration. IEEE Trans Robotics 21(3):376–386
7. Christopoulos VN, Roumeliotis S (2005) Adaptive sensing for instantaneous gas release parameter estimation. In: Proceedings of the 2005 IEEE International Conference on Robotics and Automation, Barcelona, Spain, pp 4450–4456
8. Collett TH, MacDonald BA, Gerkey BP (2005) Player 2.0: Toward a practical robot programming framework. In: Proceedings of the Australasian Conference on Robotics and Automation. http://www.araa.asn.au/acra/acra2005/papers/collet.pdf. Accessed 6 June 2008
9. Comport A, Malis E, Rives P (2007) Accurate quadrifocal tracking for robust 3D visual odometry. In: Proceedings of the 2007 IEEE International Conference on Robotics and Automation, IEEE, Roma, Italy, pp 40–45
10. Correll N, Martinoli A (2007) Robust distributed coverage using a swarm of miniature robots. In: Proceedings of the 2007 IEEE International Conference on Robotics and Automation, IEEE, Roma, Italy, pp 379–384
11. Craighead J, Murphy R, Burke J, Goldiez B (2007) A survey of commercial & open source unmanned vehicle simulators. In: Proceedings of the 2007 IEEE International Conference on Robotics and Automation, IEEE, Roma, Italy, pp 852–857
12. Das AK, Fierro R, Kumar V, Ostrowski JP, Spletzer J, Taylor CJ (2002) A vision-based formation control framework. IEEE Trans Robotics Autom 18(5):813–825
13. Dellaert F, Balch T, Kaess M, Ravichandran R, Alegre F, Berhault M, McGuire R, Merrill E, Moshkina L, Walker D (2002) The Georgia Tech yellow jackets: A marsupial team for urban search and rescue. In: AAAI Mobile Robot Competition Workshop, Edmonton, Alberta, pp 44–49
14. Drenner A, Burt I, Kratochvil B, Nelson BJ, Papanikolopoulos N, Yeşsin KB (2002) Communication and mobility enhancements to the scout robot. In: Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems, Lausanne, Switzerland
15. Fenwick JW, Newman PM, Leonard JJ (2002) Cooperative concurrent mapping and localization. In: Proceedings of the 2002 IEEE International Conference on Robotics and Automation, IEEE, Washington, DC, pp 1810–1817
16. Franchi A, Freda L, Oriolo G, Vendittelli M (2007) A randomized strategy for cooperative robot exploration. In: Proceedings of the 2007 IEEE International Conference on Robotics and Automation, IEEE, Roma, Italy, pp 768–774
17. Gage A, Murphy RR (2004) Affective recruitment of distributed heterogenous agents. In: Proceedings of the 19th National Conference on Artificial Intelligence, San Jose, pp 14–19
18. Gage DW (1993) Randomized search strategies with imperfect sensors. In: Proceedings of SPIE Mobile Robots VIII, Boston, vol 2058, pp 270–279
19. Hada Y, Yuta S (2000) A first experiment of long term activity of autonomous mobile robot – result of repetitive base-docking over a week. In: Proceedings of the ISER 2000 7th International Symposium on Experimental Robotics, Waikiki, pp 235–244
20. Hougen DF, Bonney JC, Budenske JR, Dvorak M, Gini M, Krantz DG, Malver F, Nelson B, Papanikolopoulos N, Rybski PE, Stoeter SA, Voyles R, Yesin KB (2000) Reconfigurable robots for distributed robotics. In: Government Microcircuit Applications Conf., Anaheim, CA, pp 72–75
21. Howard A, Matarić M, Sukhatme G (2002) An incremental deployment algorithm for mobile robot teams. In: Proceeedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, EPFL, Switzerland, vol 3, pp 2849–2854
22. Hu H, Kelly I, Keating D, Vinagre D (1998) Coordination of multiple robots via communication. In: Proceedings of SPIE, Boston, pp 94–103
23. iRobot corporation (2008) iRobot Roomba® Vacuuming Robot. http://irobot.com/sp.cfm?pageid=122. Accessed 6 June 2008
24. Iv DL, Srinivasa S, Lee-Shue V (2002) Towards sensor based coverage with robot teams. In: Proceedings of the 2002 IEEE International Conference on Robotics and Automation, Washington, DC, vol 1, pp 961–967
25. Jacoff A, Messina E, Evans J (2000) A standard test course for urban search and rescue robots. In: Proceedings of the 2000 performance metrics for intelligent system workshop, Gaithersburg, August 2000
26. Janssen M, Papanikolopoulos N (2007) Enabling complex behavior by simulating marsupial actions. In: Proceedings of the 15th Mediterranean Conference on Control and Automation, Athens, Greece
27. Kadioglu E, Papanikolopoulos N (2003) A method for transporting a team of miniature robots. In: Proceedings of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems, Las Vegas, NV, vol 3, pp 2297–2302
28. Kamath S, Meisner E, Isler V (2007) Triangulation based multi target tracking with mobile sensor networks. In: Proceedings of the 2007 IEEE International Conference on Robotics and Automation, IEEE, Roma, Italy, pp 3283–3288
29. Kelly I, Holland O, Melhuish C (2000) Slugbot: A robotic predator in the natural world. In: Proceedings of the 5th International Symposium on Artificial Life and Robotics for Human Welfare and Artificial Liferobotics, Oita, Japan, pp 470–475
30. Ko J, Stewart B, Fox D, Konolige K, Limketkai B (2003) A practical, decision-theoretic approach to multi-robot mapping and exploration. In: Proceedings of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, Las Vegas, NV, vol 4, pp 3232–3238
31. Koenig S, Szymanski B, Liu Y (2001) Efficient and inefficient ant coverage methods. Ann Math Artif Intell 31(1):41–76
32. Konolige K, Fox D, Limketkai B, Ko J, Stewart B (2003) Map merging for distributed robot navigation. In: Proceedings of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, Las Vegas, NV, pp 212–217
33. Kurazume R, Hirose S (2000) An experimental study of a cooperative positioning system. Autonomous Robots 8(1):43–52
34. Kwolek B (2007) Visual odometry based on gabor filters and sparse bundle adjustment. In: Proceedings of the 2007 IEEE International Conference on Robotics and Automation, IEEE, Roma, Italy, pp 3573–3578
35. Lagoudakis MG, Markakis E, Kempe D, Keskinocak P, Kleywegt A, Koenig S, Tovey C, Meyerson A, Jain S (2005) Auction-based multi-robot routing. In: Proceedings of Robotics: Science and Systems, Cambridge, USA
36. Lau H (2003) Behavioural approach for multi-robot exploration. In: Proceedings of 2003 Australasian Conference on Robotics and Automation, Brisbane, Australia
37. McMillen C, Stubbs K, Rybski PE, Stoeter SA, Gini M, Papanikolopoulos N (2002) Resource scheduling and load balancing in distributed robotic control systems. In: The 7th international conference on intelligent autonomous systems, Marina del Rey, pp 223–230

38. Min HJ, Drenner A, Papanikolopoulos N (2007) Autonomous docking for an erosi robot based on a vision system with points clustering. In: Proceedings of the 15th Mediterranean Conference on Control and Automation, Athens, Greece
39. Mosteo AR, Montano L (2007) Comparative experiments on optimization criteria and algorithms for auction based multi-robot task allocation. In: Proceedings of the 2007 IEEE International Conference on Robotics and Automation, IEEE, Roma, Italy, pp 3345–3350
40. Murphy RR (2000) Marsupial and shape-shifting robots for urban search and rescue. IEEE Intell Syst 15(2):14–19
41. Nanjanath M, Gini M (2006) Dynamic task allocation for robots via auctions. In: Proceedings of the 2006 IEEE International Conference on Robotics and Automation, Orlando, FL, pp 2781–2786
42. Ngo TD, Raposo H, Schioler H (2007) Being sociable: Multirobots with self-sustained energy. In: Proceedings of the 15th Mediterranean Conference on Control and Automation, Athens, Greece
43. Nistér D, Naroditsky O, Bergen J (2004) Visual odometry. In: Proceedings of the 2004 IEEE International Conference on Computer Vision and Pattern Recognition, IEEE, vol 1, Washington DC, pp 652–659
44. O'Rourke J (1987) Art Gallery Theorems and Algorithms. Oxford University Press, New York
45. Ota Y, Kuga T, Yoneda K (2006) Deformation compensation for continuous force control of a wall climbing quadruped with reduced-dof. In: Proceedings of the 2006 IEEE International Conference on Robotics and Automation, IEEE, Orlando, FL, pp 468–474
46. Perkins CE, Royer EM (1999) Ad-hoc on-demand distance vector routing. In: Proceedings of the 2nd IEEE Workshop on Mobile Computer Systems and Applications, New Orleans, p 90
47. Pongas D, Mistry M, Schaal S (2007) A robust quadruped walking gait for traversing rough terrain. In: Proceedings of the 2007 IEEE International Conference on Robotics and Automation, IEEE, Roma, Italy, pp 1474–1479
48. Pugh J, Martinoli A (2007) The cost of reality: Effects of real-worlde factors on multi-robot search. In: Proceedings of the 2007 IEEE International Conference on Robotics and Automation, IEEE, Roma, Italy, pp 397–404
49. Rekleitis I, Lee-Shue V, New AP, Choset H (2004) Limited communication, multi-robot team based coverage. In: Proceedings of the 2004 IEEE International Conference on Computer Vision and Pattern Recognition, New Orleans, LA, pp 3462–3468
50. Roumeliotis SI, Bekey GA (2002) Distributed multi-robot localization. IEEE Trans Robotics Autom 18(5):781–795
51. Rybski PE, Larson A, Veeraraghavan H, LaPoint M, Gini M (2004) Communication strategies in multi-robot search and retrieval: Experiences with minDART. In: DARS 2004, Toulouse, France, pp 301–310
52. Shillcutt KJ (2000) Solar based navigation for robotic explorers. Ph D thesis, Carnegie Mellon University
53. Silverman MC, Nies D, Jung B, Sukatme GS (2002) Staying alive: A docking station for autonomous robot recharging. In: Proceedings of the 2002 IEEE International Conference on Robotics and Automation, Washington, DC, pp 1050–1055
54. Simmons R, Apfelbaum D, Burgard W, Fox D, Moors M, Thrun S, Younes H (2000) Coordination for multi-robot exploration and mapping. In: Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence, 0-262-51112-6, Austin, pp 852–858
55. Singh K, Fujimura K (1993) Map making by cooperating mobile robots. In: Proceedings of the 1993 IEEE International Conference on Robotics and Automation, IEEE, Atlanta, pp 254–259
56. Southard L, Hoeg TM, Palmer DW, Antol J, Kolacinski RM, Quinn RD (2007) Exploring mars using a group of tumbleweed rovers. In: IEEE international converence on robotics and automation, Roma, Italy, pp 775–780
57. Stoeter SA (2003) Vision-based control of miniature jumping scout robots. Ph D thesis, University of Minnesota
58. Whittaker W, Champeny L (1988) Conception and developmet of two mobile teleoperated systems for TMI-2. In: Proceedings of the international meeting and topicam meeting TMI-2 accident. American Nuclear Society, Washington DC
59. Yamauchi B (1998) Frontier-based exploration using multiple robots. In: Proceedings of the 2nd international converence on autonomous agents, Mineapolis/St. Paul, pp 47–53
60. Zlot R, Stentz A, Dias MB, Thayer S (2002) Multi-robot exploration controlled by a market economy. In: Proceedings of the 2002 IEEE international conference on robotics and automation, IEEE, Washington DC, pp 3016–3023

Books and Reviews

Bräunl T (2006) Embedded robotics: Mobile robot design and applications with embedded systems. Springer, Berlin
Dudek G, Jenkin M (2000) Computational principles of mobile robotics. Cambridge University Press, Cambridge
Schultz AC, Parker LE (eds) (2002) Multi-robot systems: From swarms to intelligent automata. Kluwer Academic, Dordrecht

DNA Computing

Martyn Amos
Manchester Metropolitan University, Manchester, UK

Article Outline

Glossary
Definition of the Subject
Introduction
The DNA Molecule
The First DNA Computation
Models of DNA Computation
Subsequent Work
Assessment
Future Directions
Bibliography

Glossary

DNA Deoxyribonucleic acid. Molecule that encodes the genetic information of cellular organisms.

Enzyme Protein that catalyzes a biochemical reaction.

Nanotechnology Branch of science and engineering dedicated to the construction of artifacts and devices at the nanometer scale.

RNA Ribonucleic acid. Molecule similar to DNA, which helps in the conversion of genetic information to proteins.

Satisfiability (SAT) Problem in complexity theory. An instance of the problem is defined by a Boolean expression with a number of variables, and the problem is to identify a set of variable assignments that makes the whole expression true.

Definition of the Subject

DNA computing (or, more generally, biomolecular computing) is a relatively new field of study that is concerned with the use of biological molecules as fundamental components of computing devices. It draws on concepts and expertise from fields as diverse as chemistry, computer science, molecular biology, physics and mathematics. Although its theoretical history dates back to the late 1950s, the notion of computing with molecules was only physically realized in 1994, when Leonard Adleman demonstrated in the laboratory the solution of a small instance of a well-known problem in combinatorics using standard tools of molecular biology. Since this initial experiment, interest in DNA computing has increased dramatically, and it is now a well-established area of research. As we expand our understanding of how biological and chemical systems process information, opportunities arise for new applications of molecular devices in bioinformatics, nanotechnology, engineering, the life sciences and medicine.

Introduction

In the late 1950s, the physicist Richard Feynman first proposed the idea of using living cells and molecular complexes to construct "sub-microscopic computers." In his famous talk *"There's Plenty of Room at the Bottom"* [18], Feynman discussed the problem of "manipulating and controlling things on a small scale", thus founding the field of nanotechnology. Although he concentrated mainly on information storage and molecular manipulation, Feynman highlighted the potential for biological systems to act as small-scale information processors:

> The biological example of writing information on a small scale has inspired me to think of something that should be possible. Biology is not simply writing information; it is doing something about it. A biological system can be exceedingly small. Many of the cells are very tiny, but they are very active; they manufacture various substances; they walk around; they wiggle; and they do all kinds of marvelous things – all on a very small scale. Also, they store information. Consider the possibility that we too can make a thing very small which does what we want – that we can manufacture an object that maneuvers at that level! [18].

Early Work

Since the presentation of Feynman's vision there there has been an steady growth of interest in performing computations at a molecular level. In 1982, Charles Bennett [8] proposed the concept of a "Brownian computer" based around the principle of reactant molecules touching, reacting, and effecting state transitions due to their random Brownian motion. Bennett developed this idea by suggesting that a Brownian Turing Machine could be built from a macromolecule such as RNA. "Hypothetical enzymes", one for each transition rule, catalyze reactions between the RNA and chemicals in its environment, transforming the RNA into its logical successor.

In the same year, Conrad and Liberman developed this idea further in [15], in which the authors describe parallels between physical and computational processes (for example, biochemical reactions being employed to implement basic switching circuits). They introduce the concept of molecular level "word processing" by describing it in terms of transcription and translation of DNA, RNA processing, and genetic regulation. However, the paper lacks a detailed description of the biological mechanisms highlighted and their relationship with "traditional" computing. As the authors themselves acknowledge, "our aspiration is not to provide definitive answers … but rather to show that a number of seemingly disparate questions must be connected to each other in a fundamental way." [15]

In [14], Conrad expanded on this work, showing how the information processing capabilities of organic molecules may, in theory, be used in place of digital switching components. Particular enzymes may alter the three-dimensional structure (or *conformation*) of other *substrate* molecules. In doing so, the enzyme switches the *state* of the substrate from one to another. The notion of *conformational computing* (q.v.) suggests the possibility of a potentially rich and powerful computational architecture. Following on from the work of Conrad et al., Arkin and Ross show how various logic gates may be constructed using the computational properties of enzymatic reaction mechanisms [5] (see Dennis Bray's article [10] for a review of this work). In [10], Bray also describes work [23,24] showing how chemical "neurons" may be constructed to form the building blocks of logic gates.

Motivation

We have made huge advances in machine miniaturization since the days of room-sized computers, and yet the underlying computational framework (the *von Neumann architecture*) has remained constant. Today's supercomputers still employ the kind of sequential logic used by the mechanical "dinosaurs" of the 1940s [13].

There exist two main barriers to the continued development of "traditional", silicon-based computers using the von Neumann architecture. One is inherent to the machine architecture, and the other is imposed by the nature of the underlying *computational substrate*. A computational substrate may be defined as "*a physical substance acted upon by the implementation of a computational architecture*." Before the invention of silicon integrated circuits, the underlying substrates were bulky and unreliable. Of course, advances in miniaturization have led to incredible increases in processor speed and memory access time. However, there is a limit to how far this miniaturization can go. Eventually "chip" fabrication will hit a wall imposed by the *Heisenberg Uncertainty Principle* (HUP). When chips are so small that they are composed of components a few atoms across, quantum effects cause interference. The HUP states that the act of observing these components affects their behavior. As a consequence, it becomes impossible to know the exact state of a component without fundamentally changing its state.

The second limitation is known as the *von Neumann bottleneck*. This is imposed by the need for the central processing unit (CPU) to transfer instructions and data to and from the main memory. The route between the CPU and memory may be visualized as a two-way road connecting two towns. When the number of cars moving between towns is relatively small, traffic moves quickly. However, when the number of cars grows, the traffic slows down, and may even grind to a complete standstill. If we think of the cars as units of information passing between the CPU and memory, the analogy is complete. Most computation consists of the CPU fetching from memory and then executing one instruction after another (after also fetching any data required). Often, the execution of an instruction requires the storage of a result in memory. Thus, the speed at which data can be transferred between the CPU and memory is a limiting factor on the speed of the whole computer.

Some researchers are now looking beyond these boundaries and are investigating entirely new computational architectures and substrates. These developments include *quantum computing* (q.v.), *optical computing* (q.v.), *nanocomputers* (q.v.) and bio-molecular computers.

In 1994, interest in molecular computing intensified with the first report of a successful non-trivial molecular computation. Leonard Adleman of the University of Southern California effectively founded the field of DNA computing by describing his technique for performing a massively-parallel random search using strands of DNA [1]. In what follows we give an in-depth description of Adleman's seminal experiment, before describing how the field has evolved in the years that followed. First, though, we must examine more closely the structure of the DNA molecule in order to understand its suitability as a computational substrate.

The DNA Molecule

Ever since ancient Greek times, man has suspected that the features of one generation are passed on to the next. It was not until Mendel's work on garden peas was recognized [38] that scientists accepted that both parents contribute material that determines the characteristics of their offspring. In the early 20th century, it was discovered that *chromosomes* make up this material. Chemical analysis of chromosomes revealed that they are composed of both *protein* and *deoxyribonucleic acid*, or *DNA*. The question was, which substance carries the genetic information? For many years, scientists favored protein, because of its greater complexity relative to that of DNA. Nobody believed that a molecule as simple as DNA, composed of only four subunits (compared to 20 for protein), could carry complex genetic information.

It was not until the early 1950s that most biologists accepted the evidence showing that it is in fact DNA that carries the genetic code. However, the physical structure of the molecule and the hereditary mechanism was still far from clear.

In 1951, the biologist James Watson moved to Cambridge to work with a physicist, Francis Crick. Using data collected by Rosalind Franklin and Maurice Wilkins at King's College, London, they began to decipher the structure of DNA. They worked with models made out of wire and sheet metal in an attempt to construct something that fitted the available data. Once satisfied with their double helix model, they published the paper [42] (also see [41]) that would eventually earn them (and Wilkins) the Nobel Prize for Physiology or Medicine in 1962.

DNA Structure

DNA (deoxyribonucleic acid) [43] encodes the genetic information of cellular organisms. It consists of *polymer chains*, commonly referred to as DNA *strands*. Each strand may be viewed as a chain of *nucleotides*, or *bases*, attached

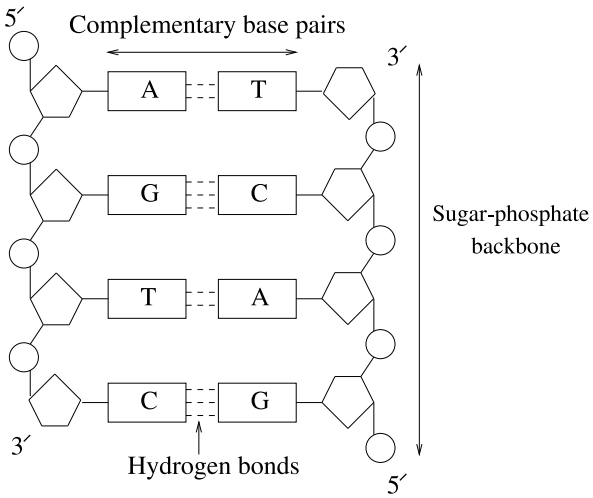

DNA Computing, Figure 1
Structure of double-stranded DNA (from [3])

to a sugar-phosphate "backbone". An n-letter sequence of consecutive bases is known as an n-mer or an *oligonucleotide* (commonly abbreviated to "oligo") of length n. Strand lengths are measured in *base pairs* (b.p.)

The four DNA nucleotides are adenine, guanine, cytosine, and thymine, commonly abbreviated to A, G, C, and T respectively. Each strand, according to chemical convention, has a 5′ and a 3′ end; thus, any single strand has a natural orientation (Fig. 1). The classical double helix of DNA is formed when two separate strands bond. Bonding occurs by the pairwise attraction of bases; A bonds with T and G bonds with C. The pairs (A, T) and (G, C) are therefore known as *complementary* base pairs. The two pairs of bases form *hydrogen bonds* between each other, two bonds between A and T, and three between G and C.

The bonding process, known as *annealing*, is fundamental to our implementation. A strand will only anneal to its complement if they have opposite polarities. Therefore, one strand of the double helix extends from 5′ to 3′, and the other from 3′ to 5′, as depicted in Fig. 1.

Operations on DNA

Some (but not all) DNA-based computations apply a specific sequence of biological operations to a set of strands. These operations are all commonly used by molecular biologists, and we now describe them in more detail.

Synthesis

Oligonucleotides may be synthesized to order by a machine the size of a microwave oven. The synthesizer is supplied with the four nucleotide bases in solution, which are

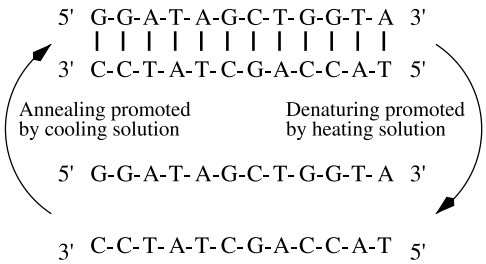

DNA Computing, Figure 2
DNA melting and annealing (from [3])

combined according to a sequence entered by the user. The instrument makes millions of copies of the required oligo and places them in solution in a small vial.

Denaturing, Annealing, and Ligation

Double-stranded DNA may be dissolved into single strands (or *denatured*) by heating the solution to a temperature determined by the composition of the strand [11]. Heating breaks the hydrogen bonds between complementary strands (Fig. 2). *Annealing* is the reverse of melting, whereby a solution of single strands is cooled, allowing complementary strands to bind together (Fig. 2).

In double-stranded DNA, if one of the single strands contains a discontinuity (i. e., one nucleotide is not bonded to its neighbor) then this may be repaired by DNA *ligase* [12]. This particular enzyme is useful for DNA computing, as it allows us to create a unified strand from several strands bound together by their respective complements. Ligase therefore acts as a molecular "cement" or "mortar", binding together several strands into a single strand.

Separation of Strands

Separation is a fundamental operation, and involves the extraction from a test tube of any *single* strands containing a specific short sequence (e. g., extract all strands containing the sequence $GCTA$). For this, we may use a "molecular sieving" process known as *affinity purification*. If we want to extract from a solution single strands containing the sequence x, we may first create many copies of its complement, $\overline{x}$. We attach to these oligos biotin molecules (a process known as "biotinylation") which in turn bind to a fixed matrix. If we pour the contents of the test tube over this matrix, strands containing x will anneal to the anchored complementary strands. Washing the matrix removes all strands that *did not* anneal, leaving only strands containing x. These may then be removed from the matrix.

Gel Electrophoresis

Gel electrophoresis is an important technique for sorting DNA strands by size [12]. Electrophoresis is the movement of charged molecules in an electric field. Since DNA molecules carry a negative charge, when placed in an electric field they tend to migrate toward the positive pole. The rate of migration of a molecule in an *aqueous* solution depends on its shape and electric charge. Since DNA molecules have the same charge per unit length, they all migrate at the same speed in an aqueous solution. However, if electrophoresis is carried out in a *gel* (usually made of agarose, polyacrylamide, or a combination of the two), the migration rate of a molecule is also affected by its *size*. This is due to the fact that the gel is a dense network of pores through which the molecules must travel. Smaller molecules therefore migrate faster through the gel, thus sorting them according to size. In order to sort strands, the DNA is placed in a well cut out of the gel, and a charge applied. After running the gel, the results are visualized by staining the DNA with dye and then viewing the gel under ultraviolet light. Bands of DNA of a specific length may then be cut out of the gel and soaked to free the strands (which may then be used again in subsequent processing steps).

PCR

The DNA *polymerases* perform several functions, including the repair and duplication of DNA. Given a short *primer* (or "tag") oligo, in the presence of nucleotide triphosphates (i. e., "spare" nucleotides), the polymerase extends the primer if and only if the primer is bound to a longer *template* strand.

Primer extension is fundamental to the *Polymerase Chain Reaction*, or PCR [28]. PCR is a process that quickly amplifies the amount of DNA in a given solution. PCR employs polymerase to make copies of a specific region (or *target sequence*) of DNA that lies between two *known* sequences. Note that this target sequence (which may be up to around 3,000 b.p. long) can be unknown ahead of time. In order to amplify template DNA with known regions (perhaps at either end of the strands), we first design forward and backward primers (i. e. primers that go from 5′ to 3′ on each strand. We then add a large excess (relative to the amount of DNA being replicated) of primer to the solution and heat it to denature the double-stranded template. Cooling the solution then allows the primers to anneal to their target sequences. We then add the polymerase, which extends the primers, forming an identical copy of the template DNA.

If we start with a single template, then of course we now have two copies. If we then repeat the cycle of heating, annealing, and polymerizing, it is clear that this approach yields an exponential number of copies of the template (since the number of strands doubles after each cycle). A typical number of cycles would be perhaps 35, yielding (assuming a single template) around 68 billion copies of the *target sequence* (for example, a gene).

The First DNA Computation

We now describe in detail the first ever successful computation performed using strands of DNA. This experiment was carried out in 1994 by Leonard Adleman of the University of Southern California. Adleman was already a highly distinguished computer scientist prior to this work, having been awarded, along with Rivest and Shamir, the 2002 ACM Turing Award for the development of the RSA public-key encryption scheme [32].

Adleman utilized the incredible storage capacity of DNA to implement a brute-force algorithm for the directed Hamiltonian Path Problem (HPP) [1]. The HPP involves finding a path through a graph (or "network") that visits each vertex ("node") exactly once. For an example application, consider a salesperson who wishes to visit several cities connected by rail links. In order to save time and money, she would like to know if there exists an itinerary that visits every city precisely once. We model this situation by constructing a graph where vertices represent cities and edges the rail links. The HPP is a classic problem in the study of *complex networks and graph theory* (q.v.) [20], and belongs to the class of problems referred to as *NP-complete* [19]. To practitioners in the field of *computational complexity* (q.v.), the NP-complete problems are most interesting, since they are amongst the hardest known problems, and include many problems of great theoretical and practical significance, such as network design, scheduling, and data storage.

The instance of the HPP that Adleman solved is depicted in Fig. 3, with the unique Hamiltonian Path (HP) highlighted by a dashed line. Adleman's approach was simple:

1. Generate strands encoding random paths such that the Hamiltonian Path (HP) is represented with high probability. The quantities of DNA used far exceeded those necessary for the small graph under consideration, so it is likely that *many* strands encoding the HP were present.
2. Remove all strands that do not encode the HP.
3. Check that the remaining strands encode a solution to the HPP.

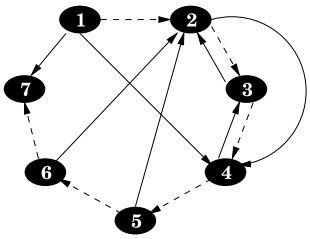

DNA Computing, Figure 3
Instance of the HPP solved by Adleman (from [3])

The individual steps were implemented as follows:
Stage 1: Each vertex and edge was assigned a distinct 20-base sequence of DNA (Fig. 4a). This implies that strands encoding a HP were of length 140 b.p., since there are seven vertices in the problem instance. Sequences representing edges act as 'splints' between strands representing their endpoints (Fig. 4b).

In formal terms, the sequence associated with an edge $i \rightarrow j$ is the $3'$ $10''$-mer of the sequence representing v_i followed by the $5'$ $10''$-mer of the sequence representing v_j. These oligos were then combined to form strands encoding random paths through the graph. An (illegal) example path ($v_1 \rightarrow v_2 \rightarrow v_3 \rightarrow v_4$) is depicted in Fig. 5.

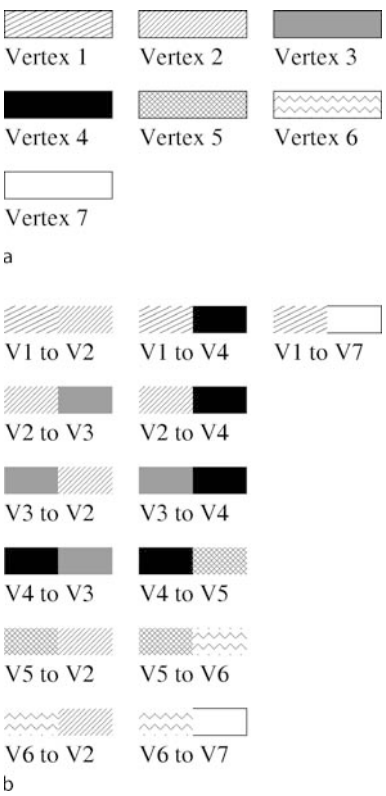

DNA Computing, Figure 4
Adleman's scheme for encoding paths-schematic representation of oligos (from [3])

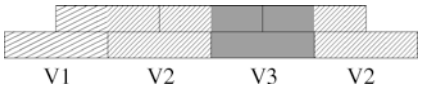

DNA Computing, Figure 5
Example path created in Adleman's scheme (from [3])

Fixed amounts (50 pmol) of each oligo were synthesized, mixed together and then ligase was added to seal any backbone nicks and form completely unified strands from the shorter "building blocks". At the end of this reaction, it is assumed that a complete strand representing the HP is present with high probability. This approach solves the problem of generating an exponential number of different paths using a polynomial number of initial oligos.
Stage 2: PCR was first used to massively amplify the population of strands encoding paths starting at v_1 and ending at v_7. These strands were amplified to such a massive extent that they effectively "overwhelmed" any strands that *did not* start and end at the correct points.

Next, strands that do not encode paths containing exactly n visits were removed. The product of the PCR amplification was run on an agarose gel to isolate strands of length 140 b.p. A series of affinity purification steps was then used to isolate strands encoding paths that visited each vertex exactly once.
Stage 3: PCR was used to identify the strand encoding the unique HP that this problem instance provides. For an n-vertex graph, we run $n - 1$ PCR reactions, with the strand representing v_1 as the left primer and the complement of the strand representing v_i as the right primer in the ith lane. The presence of molecules encoding the unique HP depicted in Fig. 3 should produce bands of length 40, 60, 80, 100, 120, and 140 b.p. in lanes 1 through 6, respectively. This is exactly what Adleman observed.

Models of DNA Computation

Although Adleman provided the impetus for subsequent work on DNA computing, his algorithm was not expressed within a formal model of computation. Richard Lipton realized that Adleman's approach, though seminal, was limited in that it was specific to solving the HPP. Lipton proposed extending Adleman's algorithm 'in a way that allows biological computers to potentially radically change the way we do all computations, not just HPPs' [25].

Satisfiability Model

Lipton's article described a methodology for solving the *satisfiability problem* (SAT) [16] using DNA. In terms of

DNA Computing, Figure 6
Unique Hamiltonian path (from [3])

computational complexity (q.v.), SAT is the "benchmark" NP-complete problem, and may be phrased as follows: given a finite set $V = \{v_1, v_2, \ldots, v_n\}$ of logical variables, we define a *literal* to be a variable, v_i, or its complement, $\overline{v_i}$. If v_i is *true* then $\overline{v_i}$ is *false*, and vice-versa. We define a *clause*, C_j, to be a set of literals $\{v_1^j, v_2^j, \ldots, v_l^j\}$. An instance, I, of SAT consists of a set of clauses. The problem is to assign a Boolean value to each variable in V such that at least one variable in each clause has the value *true*. If this is the case we may say that I has been *satisfied*.

By showing how molecular computers could be applied to this archetypal NP-complete problem, Lipton hoped to show how *any* difficult problem may be solved using this approach. The underlying principle was the same as Adleman's approach: generate all possible solutions to the problem, then gradually *filter out* strands until any remaining strands *must* be a solution to the problem. Lipton proposed solving an instance of SAT by starting with a tube containing strands representing all possible assignments to its variables.

Lipton's main contribution lay in his method for encoding *arbitrary* binary strings as strands of DNA. If we only need a small number different strands as the foundation of our computer, then it would be fairly cheap and quick to specify and order them to be individually synthesized. However, Lipton realized that this approach would quickly become infeasible as the number of variables grew. One of the characteristics of the NP-complete problems is that for even a small increase in the problem size (in this case, the number of variables), the number of possible solutions rises exponentially. For any non-trivial problem, Lipton would require a prohibitively expensive number of 'one-off' strands. For example, if Lipton wanted to solve a problem with ten variables, he would need to order $2^{10} = 1,024$ individual strands.

Lipton needed a way of encoding an *exponential*-sized pool of starting strands, using only a polynomial number of 'building-block' strands. Just as Adleman encoded a large number of possible paths through a graph, using a small number of node and edge strands, Lipton wanted to build a large number of assignment strands, using a small number of 'variable' strands. His key insight was that an arbitrary binary string (in this case encoding the values of a set of variables) could be represented as

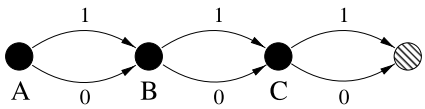

DNA Computing, Figure 7
Lipton's graph representation for a three-bit binary string

a path through a graph, where each node represented one particular bit (or variable).

Figure 7 shows an example of a graph to encode a string of three bits (variables), each represented by a node labeled A, B or C. Each node has two edges emanating from it, labeled either 1 or 0 (the edges leading out of node C lead to a 'dummy' node, which just acts like a railway buffer to end the paths). Any three-bit string can be represented by a path built by taking one of the two possible paths at each node, the value of each bit being defined by the label of the edge that is taken at each step. For example, if we only ever take the 'top' path at each branch, we encode the string 111, and if we only ever take the 'bottom' path, we end up with 000. A path that goes 'top, bottom, top' encodes the string 101, 'top, top, bottom' encodes 110, and so on.

The power of this encoding is that, just like Adleman's approach, we only have to generate one sequence for each node (including the dummy node) and one for each edge. With the correct Watson–Crick encoding, random paths through the graph form spontaneously, just like in Adleman's Hamiltonian Path experiment. Using this approach, Lipton believed he could be confident of starting off with a tube containing all possible binary strings of a given length. Each strand in the tube would encode a possible assignment of values for a set of variables. The next stage was to remove strands encoding those assignments that did *not* result in satisfaction of the formula to be solved.

Because the general form of a SAT formula is a set of clauses combined with the AND operation, it follows that each clause *must* be satisfied: if evaluating only a single clause results in a value of 0, then the *whole* formula evaluates to 0, and the formula is not satisfied. Because the variables in each clause are separated by the OR operation, it only takes a single variable to take the value 1 for the whole clause to be satisfied. Lipton's algorithm was very straightforward: he would proceed one clause at a time, looking at each component of it in turn, and keeping only the strands encoding a sequence of bits that satisfied that

particular clause. The 'winning' strands would then go forward to the *next* clause, where the process would repeat, until there were no more clauses to examine. At the end of this procedure, if Lipton had *any* strands left in his tube, then he would know with certainty that the formula was satisfiable, since only strings satisfying every clause would have survived through the successive filters.

We now show how Lipton's method can be *formally* expressed. In all *filtering* models of DNA computing, a computation consists of a sequence of operations on finite *multi-sets* of strings. Multi-sets are sets that may contain more than one copy of the same element. It is normally the case that a computation begins and terminates with a single multi-set. Within the computation, by applying legal operations of a model, several multi-sets may exist at the same time. We define operations on multi-sets shortly, but first consider the nature of an *initial set*.

An initial multi-set consists of strings which are typically of length $O(n)$ where n is the problem size. As a subset, the initial multi-set should include all possible solutions (each encoded by a string) to the problem to be solved. The point here is that the superset, in any implementation of the model, is supposed to be relatively easy to generate as a starting point for a computation. The computation then proceeds by *filtering out* strings which cannot be a solution.

Within one possible model [2], the following operations are available on sets of strings over some alphabet α:

- *separate*(T, S). Given a set T and a substring S, create two new sets $+(T, S)$ and $-(T, S)$, where $+(T, S)$ is all strings in T containing S, and $-(T, S)$ is all strings in T *not* containing S.
- *merge*$(T_1, T_2, \ldots, T_n)$. Given set $T_1, T_2, \ldots, T_n$, create $\cup(T_1, T_2, \ldots, T_n) = T_1 \cup T_2 \cup \ldots T_n$.
- *detect*(T). Given a set T, return *true* if T is nonempty, otherwise return *false*.

For example, given $\alpha = \{A, B, C\}$, the following algorithm returns *true* only if the initial multi-set contains a string composed entirely of "A"s:

Input(T)
$T \leftarrow -(T, B)$
$T \leftarrow -(T, C)$
Output(*detect*(T))

Although Lipton does not explicitly define his operation set in [25], his solution to SAT may be phrased in terms of the the operations above, described by Adleman in [2]. Lipton employs the *merge*, *separate*, and *detect* operations described above. The initial set T contains many strings, each encoding a single n-bit sequence. All possible n-bit sequences are represented in T. The algorithm proceeds as follows:

(1) Create initial set, T
(2) For each clause do begin
(3) For each literal v_i do begin
(4) if $v_i = x_j$ extract from T strings encoding $v_i = 1$ else extract from T strings encoding $v_i = 0$
(5) End for
(6) Create new set T by merging extracted strings
(7) End for
(8) If T nonempty then I is satisfiable

The pseudo-code algorithm may be expressed more formally thus:

(1) Input(T)
(2) **for** $a = 1$ to $|I|$ **do begin**
(3) **for** $b = 1$ to $|C_a|$ **do begin**
(4) **if** $v_b^a = x_j$ **then** $T_b \leftarrow +(T, v_b^a = 1)$ **else** $T_b \leftarrow +(T, v_b^a = 0)$
(5) **end for**
(6) $T \leftarrow merge(T_1, T_2, \ldots, T_b)$
(7) **end for**
(8) Output(*detect*(T))

Step 1 generates all possible n-bit strings. Then, for each clause $C_a = \{v_1^a, v_2^a, \ldots, v_l^a\}$ (Step 2) we perform the following steps. For each literal v_b^a (Step 3) we operate as follows: if v_b^a computes the positive form then we extract from T all strings encoding 1 at position v_b^a, placing these strings in T_b; if v_b^a computes the negative form we extract from T all strings encoding 0 at position v_b^a, placing these strings in T_b (Step 4); after l iterations, we have satisfied every variable in clause C_a; we then create a new set T from the union of sets $T_1, T_2, \ldots, T_b$ (Step 6) and repeat these steps for clause $C_a + 1$ (Step 7). If any strings remain in T after all clauses have been operated upon, then I is satisfiable (Step 8). Although Lipton did not report an experimental verification of his algorithm, the biological implementation would be straightforward, using affinity purification to implement extraction, pouring of tubes for *merge* and running the solution through a gel to *detect*.

Soon after Lipton's paper, the Proceedings of the Second Annual Workshop on DNA Based Computers contained several papers describing significant developments in molecular computing. These included the parallel filtering model, the sticker model, DNA self-assembly, RNA computing and surface-based computing.

We describe the last four developments (or variants thereof) in the next section. Here, we briefly introduce the parallel filtering model, as it motivates some of the later discussion.

Parallel Filtering Model

The first description of the parallel filtering model appeared in [4], with further analysis in [3]. This model was the first to provide a formal framework for the easy description of DNA algorithms for *any* problem in the complexity class *NP*. The main difference between the parallel filtering model and those previously proposed lies in the *implementation* of the removal of strings. All other models propose *separation* steps, where strings are *conserved*, and may be used later in the computation. Within the parallel filtering model, however, strings that are removed are *discarded*, and play no further part in the computation. This model is the first exemplar of the so-called "mark and destroy" paradigm of molecular computing, and was motivated in part by discussion of the error-prone nature of biotin-based separation techniques. The new operation introduced within this model is:

- *remove*$(U, \{S_i\})$. This operation removes from the set U, in parallel, any string which contains at least one occurrence of any of the substrings S_i.

The proposed implementation of the *remove* operation is as follows: for a given substring S_i, add "tag" strands composed of its Watson–Crick *complement* to the tube U, so that the tags anneal only to strands containing the target subsequence. We then add polymerase to make tagged strands double-stranded from the point of annealing ("mark"). We assume that working strands contain a specific restriction site embedded along them at fixed intervals, so adding the appropriate restriction enzyme removes only marked strands by repeated digestion ("destroy").

Subsequent Work

In this section we describe several other successful laboratory implementations of molecular-based computing, each illustrating a novel approach or new technique. Our objective is not to give an exhaustive description of each experiment, but to give a high-level treatment of the general methodology, so that the reader may approach with confidence the fuller description in the literature.

DNA Addition

One of the first successful experiments reported after Adleman's result was due to Guarnieri et al. in 1996 [21], in which they describe a DNA-based algorithm for binary addition. The method uses single-stranded DNA reactions to add together two nonnegative binary numbers. This application, as the authors note, is very different from previous proposals, which use DNA as the substrate for a massively parallel random search.

Adding binary numbers requires keeping track of the position of each digit and of any "carries" that arise from adding 1 to 1 (remembering that $1 + 1 = 0$ plus carry 1 in binary). The DNA sequences used represent not only binary strings but also allow for carries and the extension of DNA strands to represent answers. Guarnieri et al. use sequences that encode a digit in a given position and its significance, or position from the right. For example, the first digit in the first position is represented by two DNA strands, each consisting of a short sequence representing a "position transfer operator", a short sequence representing the digit's value, and a short sequence representing a "position operator".

DNA representations of all possible two bit binary integers are constructed, which can then be added in pairs. Adding a pair involves adding appropriate complementary strands, which then link up and provide the basis for strand extension to make new, longer strands. This is termed a "horizontal chain reaction", where input sequences serve as templates for constructing an extended result strand. The final strand serves as a record of successive operations, which is then read out to yield the answer digits in the correct order.

The results obtained confirmed the correct addition of $0 + 0$, $0 + 1$, $1 + 0$, and $1 + 1$, each calculation taking between 1 and 2 days of bench work. Although limited in scope, this experiment was (at the time) one of the few experimental implementations to support theoretical results.

Maximal Clique Computation

The problem of finding a Maximal Clique using DNA was addressed by Ouyang et al. in 1997 [30]. A *clique* is a fully connected subgraph of a given graph (Fig. 8). The *maximal clique* problem asks: given a graph, how many vertices are there in the largest clique? Finding the size of the largest clique is an NP-complete problem.

The algorithm proceeds as follows: for a graph G with n vertices, all possible vertex subsets (subgraphs) are rep-

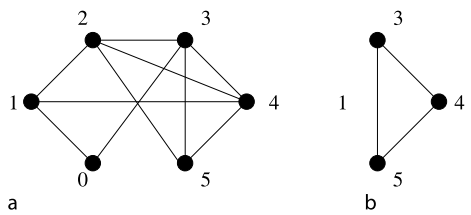

DNA Computing, Figure 8
a Ouyang graph. b Example subgraph (from [3])

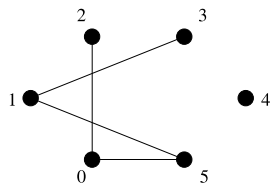

DNA Computing, Figure 9
Complement of Ouyang graph (from [3])

resented by an n-bit binary string $b_{n-1}, b_{n-2}, \ldots, b_0$. For example, given the six-vertex graph used in [30] and depicted in Fig. 8a, the string 111000 corresponds to the subgraph depicted in Fig. 8b, containing v_5, v_4, and v_3.

Clearly, the largest clique in this graph contains v_5, v_4, v_3, and v_2, represented by the string 111100.

The next stage is to find pairs of vertices that are not connected by an edge (and, therefore, by definition, cannot appear together in a clique). We begin by taking the *complement* of G, $\overline{G}$, which contains the same vertex set as G, but which only contains an edge $\{v, w\}$ if $\{v, w\}$ is *not* present in the edge set of G. The complement of the graph depicted in Fig. 8 is shown in Fig. 9.

If two vertices in $\overline{G}$ are connected by an edge then their corresponding bits cannot both be set to 1 in any given string. For the given problem, we must therefore remove strings encoding $***1*1$ (v_2 and v_0), $1****1$ (v_5 and v_0), $1***1*$ (v_5 and v_1) and $**1*1*$ (v_3 and v_1), where * means either 1 or 0. All other strings encode a (not necessarily maximal) clique.

We must then sort the remaining strings to find the largest clique. This is simply a case of finding the string containing the largest number of 1s, as each 1 corresponds to a vertex in the clique. The string containing the largest number of 1s encodes the largest clique in the graph.

The DNA implementation of the algorithm goes as follows. The first task is to construct a set of DNA strands to represent all possible subgraphs. There are two strands per bit, to represent either 1 or 0. Each strand associated with a bit i is represented by two sections, its position P_i and its value V_i. All P_i sections are of length 20 bases. If $V_i = 1$ then the sequence representing V_i is a restriction site unique to that strand. If $V_i = 0$ then the sequence representing V_i is a 10 base "filler" sequence. Therefore, the longest possible sequence is 200 bases, corresponding to the string 000000, and the shortest sequence is 140 bases, corresponding to 111111.

The computation then proceeds by digesting strands in the library, guided by the complementary graph $\overline{G}$. To remove strands encoding a connection i, j in $\overline{G}$, the current tube is divided into two, t_0 and t_1. In t_0 we cut strings encoding $V_i = 1$ by adding the restriction enzyme associated with V_i. In t_1 we cut strings encoding $V_j = 1$ by adding the restriction enzyme associated with V_j. For example, to remove strands encoding the connection between V_0 and V_2, we cut strings containing $V_0 = 1$ in t_0 with the enzyme Afl II, and we cut strings containing $V_2 = 1$ in t_1 with Spe I. The two tubes are then combined into a new working tube, and the next edge in $\overline{G}$ is dealt with.

In order to read the size of the largest clique, the final tube was simply run on a gel. The authors performed this operation, and found the shortest band to be 160 bp, corresponding to a 4-vertex clique. This DNA was then sequenced and found to represent the correct solution, 111100.

Although this is another good illustration of a DNA-based computation, the authors acknowledge the lack of scalability of their approach. One major factor is the requirement that each vertex be associated with an individual restriction enzyme. This, of course, limits the number of vertices that can be handled by the number of restriction enzymes available. However, a more fundamental issue is the exponential growth in the problem size (and thus the initial library), which we shall encounter again.

Chess Games

In [17], Faulhammer et al. describe a solution to a variant of the satisfiability problem that uses RNA rather than DNA as the computational substrate. They consider a variant of SAT, the so-called "Knight problem", which seeks configurations of knights on an $n \times n$ chess board, such that no knight is attacking any other [40].

The authors prefer the "mark and destroy" strategy rather than the repeated use of extraction to remove illegal solutions. However, the use of an RNA library and ribonuclease (RNase) H digestion gives greater flexibility, as one is not constrained by the set of restriction enzymes available. In this way, the RNase H acts as a "universal restriction enzyme", allowing selective marking of virtually any RNA strands for parallel destruction by digestion.

The particular instance solved in [17] used a 3×3 board, with the variables $a - i$ representing the squares. If a variable is set to 1 then a knight is present at that variable's square, and 0 represents the absence of a knight. The 3×3 knight problem may therefore be represented as the following instance of SAT:

$$
\begin{aligned}
&((\neg h \wedge \neg f) \vee \neg a) \wedge ((\neg g \wedge \neg i) \vee \neg b) \\
&\wedge ((\neg d \wedge \neg h) \vee \neg c) \wedge ((\neg c \wedge \neg i) \vee \neg d) \\
&\wedge ((\neg a \wedge \neg g) \vee \neg f) \wedge ((\neg b \vee \neg f) \vee \neg g) \\
&\wedge ((\neg a \wedge \neg c) \vee \neg h) \wedge ((\neg d \wedge \neg b) \vee \neg i)
\end{aligned}
$$

which, in this case, simplifies to

$$((\neg h \wedge \neg f) \wedge \neg a) \wedge ((\neg g \wedge \neg i) \vee \neg b)$$
$$\wedge ((\neg d \wedge \neg h) \vee \neg c) \wedge ((\neg c \wedge \neg i) \vee \neg d)$$
$$\wedge ((\neg a \wedge \neg g) \vee \neg f) .$$

This simplification greatly reduces the number of laboratory steps required. The experiment proceeds by using a series of RNase H digestions of "illegal" board representations, along the lines of the parallel filtering model [4].

Board representations are encoded as follows: the experiment starts with all strings of the form $x_1, \ldots, x_n$, where each variable x_i takes the value 1 or 0; then, the following operations may be performed on the population of strings:

- Cut all strings containing any pattern of specified variables $p_i, \ldots, p_k$
- Separate the "test tube" into several collections of strings (molecules) by length
- Equally divide (i. e., split) the contents of a tube into two tubes
- Pour (mix) two test tubes together
- Sample a random string from the test tube.

The first stage of the algorithm is the construction of the initial library of strands. Each strand sequence follows the template depicted in Fig. 10.

The prefix and suffix regions are included to facilitate PCR. Each variable is represented by one of two unique sequences of length 15 nucleotides, one representing the fact that the variable is set to 1, and the other the fact that it is set to 0. Variable regions are separated by short (5 nucleotide) spacer regions. In order to avoid having to individually generate each individual sequence, a "mix and split" strategy (described in more detail in [17]) is used. The RNA version of the library is then generated by *in vitro* transcription.

The algorithm proceeds as follows:

1. For each square, sequentially, split the RNA library into two tubes, labeled 1 and 2. After digestions have taken place, tube 1 will contain strands that contain a knight at that square, and tube 2 will contain strands that do *not* have knights at that square
2. In tube 1, digest with RNase H strands that have no knight at position **a**, as well as strands that describe a knight at attacking positions **h** and **f**. This implements the logical statement $((\neg h \wedge \neg f) \vee \neg a)$
3. In tube 2, digest strands that have a knight present at position **a**
4. Remove the DNA oligos used to perform the above digestions
5. Go to step 1, repeating with square **b**.

Steps 1 through 4 implement the following: "There may or may not be a knight in square **a**: if there *is*, then it is attacking squares **h** and **f**, so disallow this." The algorithm only needs to be performed for squares **a**, **b**, **c**, **d**, and **f**, as square **e**, by the rules of chess, cannot threaten or be threatened on a board this size, and any illegal interactions that squares **g**, **h**, and **i** may have are with **a**, **b**, **c**, **d**, and **f**, and have already been dealt with. At the conclusion of this stage, any remaining full-length strands are recovered, as they should encode legal boards.

The "mark and destroy" digestion operation is implemented as follows. If we wish to retain (i. e., select) strands encoding variable a to have value 1, DNA oligonucleotides corresponding to the complement of the $a = 0$ sequence are added to the tube, and anneal to all strands encoding $a = 0$. RNase H is then added to the solution. Ribonuclease H (RNase H) is an endoribonuclease which specifically hydrolyzes the phosphodiester bonds of RNA hybridized to DNA. RNase H does not digest single or double-stranded DNA, so his operation therefore leaves intact only those strands encoding $a = 1$, in a fashion similar to the removal operation of the parallel filtering model [4].

The results obtained (described in [17]) were extremely encouraging: out of 43 output strands sampled, only one contained an illegal board. Given that the population sampled encoded 127 knights, this gave an overall knight placement success rate of 97.7%.

Computing on Surfaces

Another experiment that makes use of the "mark and destroy" paradigm is described in [26] (although early work was performed from 1996). The key difference between

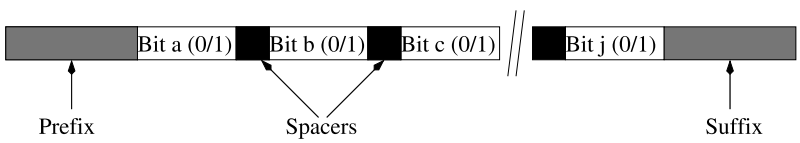

DNA Computing, Figure 10
Template for RNA strands (from [3])

this and previous experiments is that the DNA strands used are tethered to a support rather than being allowed to float freely in solution. The authors argue that this approach greatly simplifies the automation of the (potentially very many) repetitive chemical processes required during the performance of an experiment.

The authors report a DNA-based solution to a small instance of SAT. The specific problem solved is

$$(w \vee x \vee y) \wedge (w \vee \neg y \vee z) \wedge (\neg x \vee y) \wedge (\neg w \vee \neg y) .$$

16 unique DNA strands were synthesized, each one corresponding to one of the $2^4 = 16$ combinations of variable values. The actual encodings are given in Table 1 (taken from [26]).

Each of the 16 sets of strands was then affixed to a specific region of a gold coated surface, so that each solution to the SAT problem was represented as an individual cluster of strands.

The algorithm then proceeds as follows. For each clause of the problem, a cycle of "mark", "destroy", and "unmark" operations is carried out. The goal of each cycle is to destroy the strands that do *not* satisfy the appropriate clause. Thus, in the first cycle, the objective is to destroy strands that do not satisfy the clause $(w \vee x \vee y)$. Destruction is achieved by "protecting", or *marking* strands that *do* satisfy the clause by annealing to them their complementary strands. *E. coli* exonuclease I is then used to digest unmarked strands (i. e., any single-stranded DNA).

By inspection of Table 1, we see that this applies to only two strands, S_0 (0000) and S_1 (0001). Thus, in cycle 1 the complements of the 14 other strands

$$\begin{aligned}(w &= 1(S_8, S_9, S_10, S_11, S_12, S_13, S_14, S_15) ;\\ x &= 1(S_4, S_5, S_6, S_7, S_12, S_13, S_14, S_15) ;\\ y &= 1(S_2, S_3, S_6, S_7, S_10, S_11, S_14, S_15))\end{aligned}$$

were combined and hybridized to the surface before the exonuclease I was added. The surface was then regenerated (the unmark operation) to return the remaining surface-bound oligos ($S_2 - S_15$) to single-stranded form. This process was repeated three more times for the remaining three clauses, leaving a surface containing only strands encoding a legal solution to the SAT problem.

The remaining molecules were amplified using PCR and then hybridized to an addressed array. The results of fluorescence imaging clearly showed four spots of relatively high intensity, corresponding to the four regions occupied by legal solutions to the problem (S_3, S_7, S_8, and S_9).

Although these results are encouraging as a first move toward more error-resistant DNA computing, as the authors themselves acknowledge, there remain serious concerns about the scalability of this approach (1,536 individual oligos would be required for a 36-bit, as opposed to 4-bit, implementation).

DNA Computing, Table 1
Strands used to represent SAT variable values

Strand	Sequence	*wxyz*
S_0	CAACCCAA	0000
S_1	TCTCAGAG	0001
S_2	GAAGGCAT	0010
S_3	AGGAATGC	0011
S_4	ATCGAGCT	0100
S_5	TTGGACCA	0101
S_6	ACCATTGG	0110
S_7	GTTGGGTT	0111
S_8	CCAAGTTG	1000
S_9	CAGTTGAC	1001
S_{10}	TGGTTTGG	1010
S_{11}	GATCCGAT	1011
S_{12}	ATATCGCG	1100
S_{13}	GGTTCAAC	1101
S_{14}	AACCTGGT	1110
S_{15}	ACTGGTCA	1111

Computing with Hairpins

The tendency of DNA molecules to self-anneal was exploited by Sakamoto et al. in [35] for the purposes of solving a small instance of SAT. The authors encode the given formula in "literal strings" which are conjunctions of the literals selected from each SAT clause (one literal per clause). A formula is satisfiable if there exists a literal string that does not contain any variable together with its negation. If each variable is encoded as a DNA subsequence that is the Watson–Crick complement of its negation then any strands containing a variable and its negation self-anneal to form "hairpin" structures. These can be distinguished from non-hairpin structure-forming strands, and removed. The benefit of this approach is that it does not require physical manipulation of the DNA, only temperature cycling. The drawback is that it requires 3^m literal strings for m clauses, thus invoking once again the scalability argument.

Gel-Based Computing

A much larger (20 variable) instance of 3-SAT was successfully solved by Adleman's group in an experiment described in 2002 [9]. This is, to date, the largest known problem instance successfully solved by a DNA-based

computer; indeed, as the authors state, "this computational problem may yet be the largest yet solved by non-electronic means" [9].

The architecture underlying the experiment is related to the Sticker Model described by Roweis et al. [34]. The difference here is that only separation steps are used – the application of stickers is not used. Separations are achieved by using oligo probes immobilized in polyacrylamide gel-filled glass modules, and strands are pulled through them by electrophoresis. Strands are removed (i. e., retained in the module) by virtue of their hybridizing to the immobilized probes, with other strands free to pass through the module and be subject to further processing. Captured strands may be released and transported (again via electrophoresis) to other modules for further processing.

The potential benefits of such an approach are clear; the use of electrophoresis minimizes the number of laboratory operations performed on strands, which, in turn, increases the chance of success of an experiment. Since strands are not deliberately damaged in any way, they, together with the glass modules, are potentially reusable for multiple computations. Finally, the whole process is potentially automatable.

The problem solves was a 20-variable, 24-clause 3-SAT formula Φ, with a unique satisfying truth assignment. These are

$$\begin{aligned}\Phi = {} & (\neg x_{13} \vee x_{16} \vee x_{18}) \wedge (x_5 \vee x_{12} \vee \neg x_9) \\ & \wedge (\neg x_{13} \vee \neg x_2 \vee x_{20}) \wedge (x_{12} \vee x_9 \vee \neg x_5) \\ & \wedge (x_{19} \vee \neg x_4 \vee x_6) \wedge (x_9 \vee x_{12} \vee \neg x_5) \\ & \wedge (\neg x_1 \vee x_4 \vee \neg x_{11}) \wedge (x_{13} \vee \neg x_2 \vee \neg x_{19}) \\ & \wedge (x_5 \vee x_{17} \vee x_9) \wedge (x_{15} \vee x_9 \vee \neg x_{17}) \\ & \wedge (\neg x_5 \vee \neg x_9 \vee \neg x_{12}) \wedge (x_6 \vee x_{11} \vee x_4) \\ & \wedge (\neg x_{15} \vee \neg x_{17} \vee x_7) \wedge (\neg x_6 \vee x_{19} \vee x_{13}) \\ & \wedge (\neg x_{12} \vee \neg x_9 \vee x_5) \wedge (x_{12} \vee x_1 \vee x_{14}) \\ & \wedge (x_{20} \vee x_3 \vee x_2) \wedge (x_{10} \vee \neg x_7 \vee \neg x_8) \\ & \wedge (\neg x_5 \vee x_9 \vee \neg x_{12}) \wedge (x_{18} \vee \neg x_{20} \vee x_3) \\ & \wedge (\neg x_{10} \vee \neg x_{18} \vee \neg x_{16}) \wedge (x_1 \vee \neg x_{11} \vee \neg x_{14}) \\ & \wedge (x_8 \vee \neg x_7 \vee \neg x_{15}) \wedge (\neg x_8 \vee x_{16} \vee \neg x_{10})\end{aligned}$$

with a unique satisfying assignment of

$$\begin{aligned}&x_1 = F, \quad x_2 = T, \quad x_3 = F, \quad x_4 = F, \quad x_5 = F, \\ &x_6 = F, \quad x_7 = T, \quad x_8 = T, \quad x_9 = F, \quad x_{10} = T, \\ &x_{11} = T, \quad x_{12} = T, \quad x_{13} = F, \quad x_{14} = F, \quad x_{15} = T, \\ &x_{16} = T, \quad x_{17} = T, \quad x_{18} = F, \quad x_{19} = F, \quad x_{20} = F\,.\end{aligned}$$

As there are 20 variables, there are $2^{20} = 1,048,576$ possible truth assignments. To represent all possible assignments, two distinct 15 base *value sequences* were assigned to each variable x_k ($k = 1, \ldots, 20$), one representing true (T), X_k^T, and one representing false (F), X_k^F. A mix and split generation technique similar to that of Faulhammer et al. [17] was used to generate a 300-base *library sequence* for each of the unique truth assignments. Each library sequence was made up of 20 value sequences joined together, representing the 20 different variables. These library sequences were then amplified with PCR.

The computation proceeds as follows: for each clause, a glass clause module is constructed which is filled with gel and contains covalently bound probes designed to capture only those library strands that *do satisfy* that clause; strands that do not satisfy the clause are discarded.

In the first clause module ($\neg x_{13} \vee x_{16} \vee x_{18}$) strands encoding X_3^F, X_{16}^F, and X_{18}^T are retained, while strands encoding X_3^T, X_{16}^T, and X_{18}^F are discarded. Retained strands are then used as input to the next clause module, for each of the remaining clauses. The final (24th) clause module should contain only those strands that have been retained in all 24 clause modules and hence encode truth assignments satisfying Φ.

The experimental results confirmed that a unique satisfying truth assignment for Φ was indeed found using this method. Impressive though it is, the authors still regard with scepticism claims made for the potential superiority of DNA-based computers over their traditional silicon counterparts. In a separate interview, Len Adleman stated that "DNA computers are unlikely to become stand-alone competitors for electronic computers. We simply cannot, at this time, control molecules with the deftness that electrical engineers and physicists control electrons" [31]. However, "they [DNA computers] enlighten us about alternatives to electronic computers and studying them may ultimately lead us to the true 'computer of the future'" [9]. In the next Section, we consider how the focus of DNA computing has since shifted away from the solution of "traditional" problems.

Assessment

In February 1995, one of the founding fathers of the theory of *computational complexity* (q.v.) published a short paper on DNA computing. In his article, titled "On the Weight of Computations" [22], Juris Hartmanis sounded a cautionary note amidst the growing interest surrounding DNA computing. By calculating the smallest amount of DNA required to encode *all possible paths* through a graph, he calculated that Adleman's experiment, if scaled up from 7 cities to 200 cities, would require an initial set of DNA strands that would weigh more than the Earth. Hartmanis was quick to point out the value of the initial work –

"Adleman's molecular solution of the Hamiltonian path problem is indeed a magnificent achievement and may inititiate a more intensive exploration of molecular computing and computing in biological systems". However, he also emphasised the long-term futility of hoping that DNA-based computers could ever beat silicon machines in this particular domain. Soon after Adleman's experiment, hopes were expressed that a lot of the promise of molecular computers could be derived from their massive inherent parallelism – each operation is performed at the same time on trillions of DNA strands. However, as Turing showed, computers are not limited to any particular physical construction, and a DNA computer will suffer just as much as its silicon counterpart from the "exponential curse". If we require a molecular algorithm for a difficult problem to run in reasonable *time* then there is a fundamental requirement (unless P is shown to be equivalent to NP) for an exponential amount of *space* (in this case, the amount of DNA required). As Hartmanis observed, "… the exponential function grows too fast and the atoms are a bit too heavy to hope that the molecular computer can break the exponential barrier, this time the weight barrier".

This view would now seem to largely reflect the community consensus. As Ogihara and Ray argued in 2000, "It is foolish to attempt to predict the future of technology, but it may be that the ideal application for DNA computation does not like in computing large NP problems" [29]. In an interview in 2002, nanotechnologist Ned Seeman stated that "I am not a computer scientist, but I suspect it [molecular computing] is not well suited to traditional problems. I certainly feel that it is better suited to algorithmic self-assembly and biologically-oriented applications" [36]. That is not to say that molecular-based computations do not have a future, and in the final Section we briefly discuss some possible future directions that molecular computing may take (and offer pointers to other articles in the current volume).

Future Directions

If molecular computations are to be scalable and/or sustainable then it is clear that they should be performed with a minimum of human intervention. The type of experiment originally performed by Adleman was feasible because it required a relatively small number of biological steps (although it still took a week of bench time to carry out). However, the number of manipulations (and the amount of material) required for a non-trivial computation would quickly render this approach infeasible. Although attempts have been made to automate molecular computations using, for example, microfluidics [39], this does not allow us to avoid the fundamental issue of scalability.

One promising subfield concerns molecular computing using machines (or "automata") constructed from DNA strands and other biomaterials. The construction of *molecular automata* (q.v.) was demonstrated by Benenson et al. in [6]. This experiment builds on the authors' earlier work [7] on the construction of biomolecular machines. In [6], the authors describe the construction of a molecular automaton that uses the process of DNA backbone hydrolysis and strand hybridization, fuelled by the potential free energy stored in the DNA itself.

One aim of researchers in this field is to build automata that may operate within living cells. Such devices may offer possibilities in terms of novel therapeutics, drug synthesis, bio-nanotechnology or *amorphous computing* (q.v.) Theoretical studies in DNA computing based on abstract models of the cell are already well-established (see the article on *membrane computing*), but recent work on *bacterial computing* (q.v.) has illustrated the feasibility of engineering living cells for the purposes of human-defined computation.

Finally, the notion of *biomolecular self-assembly* (q.v.) suggests ways in which the tendency of molecules to form spontaneously ordered structures may be harnessed, either for the purposes of computation or for engineering-based applications (for example, *DNA-templated self-assembly of proteins and nanowires* (q.v.)). *Algorithmic* self-assembly has been demonstrated in the laboratory by Mao et al. [27]. This builds on work done on the self-assembly of periodic two-dimensional arrays (or "sheets") of DNA tiles connected by "sticky" pads [44,45]. The authors of [27] report a one-dimensional algorithmic self-assembly of DNA triple-crossover molecules (tiles) to execute four steps of a logical XOR operation on a string of binary bits.

Triple-crossover molecules contain four strands that self-assemble through Watson–Crick complementarity to produce three double helices in roughly a planar formation. Each double helix is connected to adjacent double helices at points where their strands cross over between them. The ends of the core helix are closed by hairpin loops, but the other helices may end in sticky ends which direct the assembly of the macrostructure. The tiles then self-assemble to perform a computation. The authors of [27] report successful XOR computations on pairs of bits, but note that the scalability of the approach relies on proper hairpin formation in very long single-stranded molecules, which cannot be assumed.

In 2006, Paul Rothemund produced a ground-breaking piece of work, which he called "DNA origami" [33].

His breakthrough was to describe a competely novel method for constructing *arbitrary* two-dimensional DNA-based structures. "When it comes to making shapes out of DNA", explained Lloyd Smith in a commentary article [37], "the material is there, and its properties are understood. What was missing was a convincing, universal design scheme to allow our capabilities to unfold to the full". Rothemund demonstrated the generality of his approach by showing how it could build structures as diverse as a triangle, a five-pointed star, and even a "smiley" face and a map of the Americas. The shapes that he managed to construct were ten times more complex than anything that had been constructed from DNA before. Rothemund's scheme was different to previous tile-based methods in that it used only a *single* long DNA strand at its foundation. By repeatedly folding a long strand of DNA around in a maze-like pattern, a scaffold was formed that traced the outline of the desired object. This structure was then "clipped" into place by short DNA strands, which Rothemund referred to as "staples". He used as the main building block a 7,000 base sequence of DNA taken from the M13 virus. He developed a computer program that would take a human-designed folding path (the "shape"), map it onto the sequence of the DNA strand and then generate the sequences of the staples that would be required to hold it all together. In order to demonstrate the power of his method, he built a DNA map of the Americas, where a single nanometer (one billionth of a meter) represented 200 kilometers in real life (the map was 75 nanometers across). "The results that emerge are stunning", praised Lloyd Smith, after seeing the work. The possible impact of this work may well be huge, with one possible application being a "nanoworkbench". In 2003, John Reif and his team had shown how a simple DNA scaffold could cause proteins to self-assemble into regular, periodic two-dimensional grids [46]. By using his approach to build much more complex holding structures, Rothemund argued that it could now be used to carry out biological experiments at the nanoscale, studying the behavior of complex protein assemblies, such as drugs, in a spatially-controlled environment. As Lloyd Smith observes, "Thus equipped not only with DNA building materials and an understanding of their structural and chemical properties, but also with a versatile general approach to weaving them together, we are arriving at a new frontier in our pursuit of ever-smaller structures. The barrier we have to surmount next is to deploy our knowledge to develop structures and devices that are really useful. Happily, in that endeavor we are now perhaps limited more by our imagination than our ability" [37].

Bibliography

Primary Literature

1. Adleman LM (1994) Molecular computation of solutions to combinatorial problems. Science 266:1021–1024
2. Adleman LM (1995) On constructing a molecular computer. Draft, University of Southern California
3. Amos M (2005) Theoretical and Experimental DNA Computation. Springer, Berlin
4. Amos M, Gibbons A, Hodgson D (1996) Error-resistant implementation of DNA computations. In: Landweber LF, Baum EB (eds) 2nd Annual Workshop on DNA Based Computers. Princeton University, NJ, 10–12 June 1996. American Mathematical Society
5. Arkin A, Ross J (1994) Computational functions in biochemical reaction networks. Biophys J 67:560–578
6. Benenson Y, Adam R, Paz-Livneh T, Shapiro E (2003) DNA molecule provides a computing machine with both data and fuel. Proc Natl Acad Sci 100:2191–2196
7. Benenson Y, Paz-Elizur T, Adar R, Keinan E, Livneh Z, Shapiro E (2001) Programmable and autonomous computing machine made of biomolecules. Nature 414:430–434
8. Bennett CH (1982) The thermodynamics of computation – a review. Int J Theor Phys 21:905–940
9. Braich RS, Chelyapov N, Johnson C, Rothemund PWK, Adleman L (2002) Solution of a 20-variable 3-SAT problem on a DNA computer. Science 296:499–502
10. Bray D (1995) Protein molecules as computational elements in living cells. Nature 376:307–312
11. Breslauer KJ, Frank R, Blocker H, Marky LA (1986) Predicting DNA duplex stability from the base sequence. Proc Natl Acad Sci 83(11):3746–3750
12. Brown TA (1993) Genetics: A Molecular Approach. Chapman and Hall, New York
13. Campbell-Kelly M, Aspray W (2004) Computer: A History of the Information Machine, 2nd edn. Westview Press, Colorado
14. Conrad M (1985) On design principles for a molecular computer. Commun ACM 28:464–480
15. Conrad M, Liberman EA (1982) Molecular computing as a link between biological and physical theory. J Theor Biol 98: 239–252
16. Cook S (1971) The complexity of theorem proving procedures. Proceedings of the 3rd Annual ACM Symposium on Theory of Computing, pp 151–158
17. Faulhammer D, Cukras AR, Lipton RJ, Landweber LF (2000) Molecular computation: RNA solutions to chess problems. Proc Nat Acad Sci 97:1385–1389
18. Feynman RP (1961) There's plenty of room at the bottom. In: Gilbert D (ed) Miniaturization. Reinhold, New York, pp 282–296
19. Garey MR, Johnson DS (1979) Computers and Intractability: A Guide to the Theory of NP-Completeness. WH Freeman and Company, New York
20. Gibbons AM (1985) Algorithmic Graph Theory. Cambridge University Press, Cambridge
21. Guarnieri F, Fliss M, Bancroft C (1996) Making DNA add. Science 273:220–223
22. Hartmanis J (1995) On the weight of computations. Bull Euro Assoc Theor Comput Sci 55:136–138
23. Hjelmfelt A, Schneider FW, Ross J (1993) Pattern recognition in coupled chemical kinetic systems. Science 260:335–337

24. Hjelmfelt A, Weinberger ED, Ross J (1991) Chemical implementation of neural networks and Turing machines. Proc Nat Acad Sci 88:10983–10987
25. Lipton RJ (1995) DNA solution of hard computational problems. Science 268:542–545
26. Liu Q, Wang L, Frutos AG, Condon AE, Corn RM, Smith LM (2000) DNA computing on surfaces. Nature 403:175–179
27. Mao C, LaBean TH, Reif JH, Seeman NC (2000) Logical computation using algorithmic self-assembly of DNA triple-crossover molecules. Nature 407:493–496
28. Mullis KB, Ferré F, Gibbs RA (eds) (1994) The Polymerase Chain Reaction. Birkhauser, Boston
29. Ogihara M, Ray A (2000) DNA computing on a chip. Nature 403:143–144
30. Ouyang Q, Kaplan PD, Liu S, Libchaber A (1997) DNA solution of the maximal clique problem. Science 278:446–449
31. Regalado A (2002) DNA computing. MIT Technology Review. http://www.technologyreview.com/articles/00/05/regalado0500.asp. Accessed 26 May 2008
32. Rivest R, Shamir A, Adleman L (1978) A method for obtaining digital signatures and public key cryptosystems. Comm ACM 21:120–126
33. Rothemund PWK (2006) Folding DNA to create nanoscale patterns. Nature 440:297–302
34. Roweis S, Winfree E, Burgoyne R, Chelyapov NV, Goodman MF, Rothemund PWK, Adleman LM (1996) A sticker based architecture for DNA computation. In: Landweber LF, Baum EB (eds) 2nd Annual Workshop on DNA Based Computers. Princeton University, NJ, 10–12 June 1996. American Mathematical Society
35. Sakamoto K, Gouzu H, Komiya K, Kiga D, Yokoyama S, Yokomori T, Hagiya M (2000) Molecular computation by DNA hairpin formation. Science 288:1223–1226
36. Smalley E (2005) Interview with Ned Seeman. Technology Research News, May 4
37. Smith LM (2006) Nanostructures: The manifold faces of DNA. Nature 440:283–284
38. Stubbe H (1972) History of Genetics – from Prehistoric times to the Rediscovery of Mendel's Laws. MIT Press, Cambridge
39. van Noort D, Gast F-U, McCaskill JS (2002) DNA computing in microreactors. In: Jonoska N, Seeman NC (eds) DNA Computing: 7th International Workshop on DNA-Based Computers. LNCS, vol 2340. Springer, Berlin, pp 33–45
40. Watkins JJ (2004) Across the Board: The Mathematics of Chess Problems. Princeton University Press, Princeton
41. Watson JD, Crick FHC (1953) Genetical implications of the structure of deoxyribose nucleic acid. Nature 171:964
42. Watson JD, Crick FHC (1953) Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature 171: 737–738
43. Watson JD, Hopkins NH, Roberts JW, Steitz JA, Weiner AM (1987) Molecular Biology of the Gene, 4th edn. Benjamin/Cummings, Menlo Park
44. Winfree E, Liu F, Wenzler L, Seeman NC (1998) Design and self-assembly of two-dimensional DNA crystals. Nature 394: 539–544
45. Winfree E (1998) Algorithmic self-assembly of DNA. Ph D thesis, California Institute of Technology
46. Yan H, Park SH, Finkelstein G, Reif JH, LaBean TH (2003) DNA-templated self-assembly of protein arrays and highly conductive nanowires. Science 301:1882–1884

Books and Reviews

Adleman L (1998) Computing with DNA. Sci Am 279:54–61
Amos M (2006) Genesis Machines: The New Science of Biocomputing. Atlantic Books, London
Forbes N (2004) Imitation of Life: How Biology is Inspiring Computing. MIT Press, Cambridge
Gonick L, Wheelis M (1983) The Cartoon Guide to Genetics. Harper Perennial, New York
Jones R (2004) Soft Machines: Nanotechnology and Life. Oxford University Press, Oxford
Păun G, Rozenberg G, Salomaa A (1998) DNA Computing: New Computing Paradigms. Springer, Berlin
Pool R (1995) A boom in plans for DNA computing. Science 268:498–499
Watson J (2004) DNA: The Secret of Life. Arrow Books, London

DNA-Templated Self-Assembly of Protein Arrays and Highly Conductive Nanowires

Baoquan Ding[1], Yan Liu[2], Sherri Rinker[2], Hao Yan[2]
[1] Molecular Foundry, Lawrence Berkeley National Lab, Berkeley, USA
[2] Department of Chemistry and Biochemistry and Biodesign Institute, Arizona State University, Tempe, USA

Article Outline

Glossary
Definition of the Subject
Introduction
DNA Self-Assembly
DNA-Templated Assembly of Protein Arrays
DNA-Templated Highly Conductive Nanowires
Future Directions
Bibliography

Glossary

DNA Deoxyribonucleic acid, is a nucleic acid molecule that contains the genetic instructions used in the development and functioning of all living organisms. Chemically, DNA is a long polymer of simple units called nucleotides, which are held together by a backbone made of alternating sugars and phosphate groups. Attached to each sugar is one of four types of molecules called bases: Adenine (A), Thymine (T), Guanine (G) and Cytosine (C).

Double-crossover (DX) motif The DX motif consists of two DNA double helices linked in two different places.

These molecules are related to intermediates in genetic recombination but, in the molecules used in DNA self-assembly, the linkages are usually between strands of opposite polarity, rather than the same polarity.

Paranemic-crossover DNA A DNA motif that can be formed by reciprocal exchange between strands of the same polarity on two DNA double helices at every possible position.

Definition of the Subject

DNA self-assembly is a very useful and powerful tool for bottom-up nanofabrication. It consists of combining unusual DNA motifs by specific structurally well-defined cohesive interactions (sticky ends) to produce target materials with predictable 3D structure. This method has generated versatile DNA nanostructures including polyhedral catenanes, robust nanomechanical devices, and a variety of periodic and aperiodic arrays in two dimensions. DNA self-assembled structures have been used as the template for different guest functional molecules such as proteins, metallic nano-particles, DNA based nanodevices and highly conductive nanowires. Regular lattices made of DNA could hold copies of large biological molecules in a highly ordered array for x-ray crystallography to determine their structure, an important step in the "rational" design of drugs. Alternatively, the lattices could serve as scaffolding for nanoelectronic components, either as a working device or as a step in the manufacture of a device. Materials could be constructed-either made of DNA or templated by DNA-with precisely designed structures in the nanometer scale. The organizational capabilities of DNA nanotechnology are just beginning to be explored, and the field is expected to be able to organize a variety of species that will lead to exciting and possibly revolutionary materials.

Introduction

DNA is well known as the genetic material for information storage and duplication in biology. But, the applications of DNA are not restricted to biological science. DNA's molecular structure and chemical properties make it centrally important in biology. These properties make it to be highly useful as an engineering material for nanoscale constructions. The diameter of DNA double helix is about 2 nm and its helical repeat is about 3.5 nm. The key feature of DNA that makes it useful as a nanoscale building block is its specificity and programmability in intermolecular interactions. The famous Watson–Crick base pairing consists of Adenine (A) paired specifically with Thymine (T) and Guanine (G) paired specifically with Cytosine (C) through hydrogen bonds within the context of the DNA double helical backbone. It is also possible to get two different double helices to cohere end-to-end by having short overhangs at the ends of the strands; these overhangs are called 'sticky ends'. It has been shown that sticky ends cohere to form the conventional structure of B-DNA. The structure of sticky-ended DNA complexes is well-defined and specifically predictable. Given that the persistence length of DNA is about 50 nm, sticky ends can be used to assemble DNA molecules. DNA can be manipulated using commercially available enzymes for site-selective DNA cleavage, ligation, labeling, transcription, replication, kination, and methylation. DNA nanotechnology is further empowered by well-established methods for purification and structural characterization and by solid-phase synthesis, so that any designed DNA strand can be constructed. The above advantages of DNA make DNA based nanotechnology a rapidly progressing and exciting research field [1,2,3], especially in self-assembled nanostructures [4,5,6], nanorobotics [7,8,9,10,11,12], and nanocomputation [13,14,15].

DNA Self-Assembly

The simple hybridization of single strand DNA to assemble into linear duplex structures is not sufficient for nanoconstruction purposes, because it can not produce precise complex systems in 2D or 3D structures. To do that, it is necessary to work with branched DNA molecules. Branched DNA occurs naturally in living systems. For example, ephemeral branch points are found as Holliday junction intermediates in the process of genetic recombination. Branching can occur when exchange occurs between two DNA strands. A single exchange event can produce a Holliday junction-like molecule. It is a 4-arm branch, which can be stabilized by ensuring that there is no twofold symmetry around its branch point; minimizing sequence symmetry is key to the design of all unusual DNA motifs. A double exchange between DNA double helices produces double crossover (DX) molecules [4], and double exchange between three successive helices produced triple crossover (TX) molecules [5]. Some examples are shown in Fig. 1. Synthetic DNA complexes can be designed to have fixed branch points. Thus, using DNA as the building blocks for complex materials with nanoscale features is simple: Take synthetic branched DNA molecules with programmed sticky ends, and get them to self-assemble into the desired structures.

The first artificial DNA structure is a stick-cube, whose edges are double helices. More complex polyhedra and

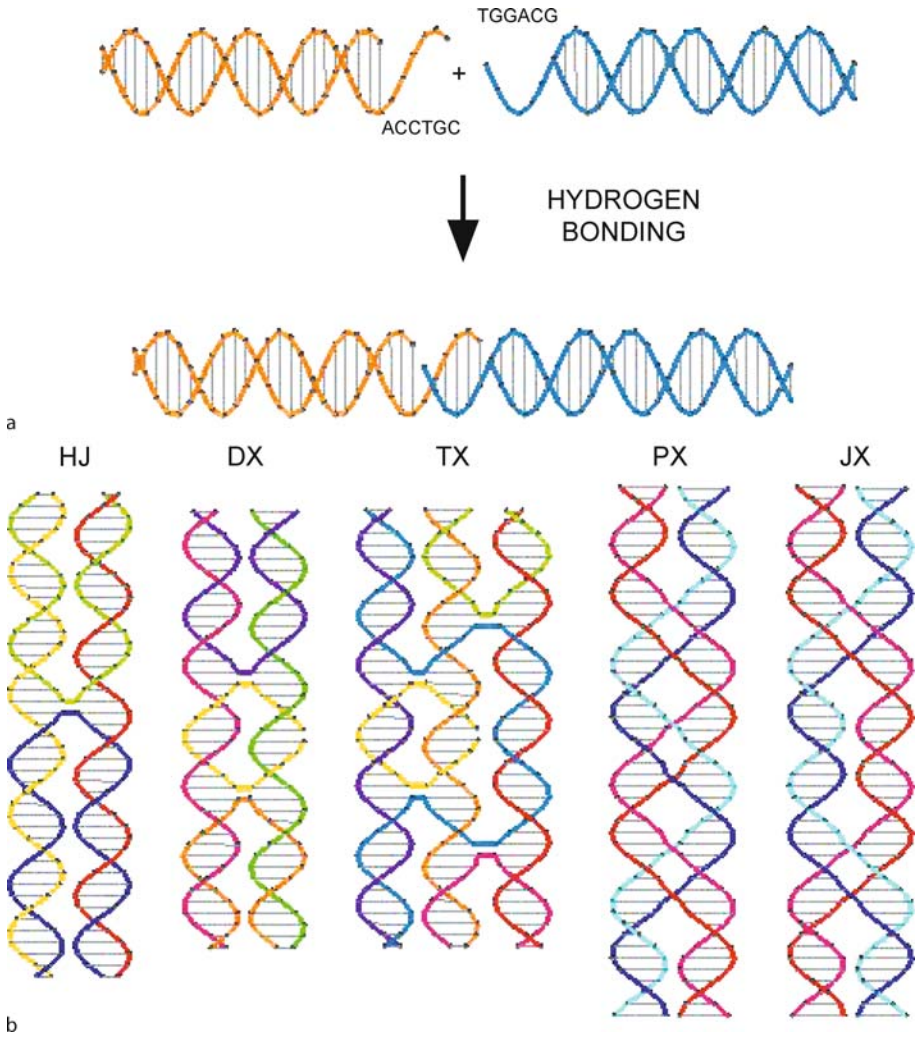

DNA-Templated Self-Assembly of Protein Arrays and Highly Conductive Nanowires, Figure 1
Component of DNA self-assembly. a Sticky Ended Cohesion. Two linear double helical molecules of DNA are shown at the *top of panel* a. The *right end* of the *orange* molecule and *left end* of the *blue* molecule have single-strand extension 'sticky ends' that are totally complementary to each other. The *middle* portion shows that, under the proper conditions, these bind to each other specifically by hydrogen bonding. (5′ and 3′ ends of DNA molecule are not shown). b Key Motifs in DNA self-assembly. On the *left* is a Holliday Junction (HJ), a 4-arm junction that results from a single strand exchange between double helices. To its *right* is a double crossover (DX) molecule, resulting from a double exchange. To the *right* of the DX is a triple crossover (TX) molecule that results from two successive double reciprocal exchanges. The HJ, the DX and the TX molecule all contain exchange between strands of opposite polarity. To the *right* of the TX molecule is a paranemic crossover (PX) molecule, where two double helices exchange strands at every possible points where the helices come into proximity. To the *right* of the PX molecule is a JX_2 molecule that lacks two of the crossover of the PX molecule. The exchange in the PX and JX_2 molecule are between strands of the same polarity

topological structures, followed [2]. DX molecules were the initial stable and rigid building blocks for the construction of periodic assemblies and the formation of the first 2D DNA array. DX, TX and other DNA motifs have been used to produce algorithmic assemblies displaying well designed patterns. These different DNA lattices provide multiple attachment sites for complex programmable structures and have lead to diverse possibilities for templating useful components and interesting chemistry.

Since DX and TX tiles are designed with their helices parallel and coplanar, their 2D lattices grow well in the direction parallel to the helix axis and not as well in the direction perpendicular to it. H. Yan et al. have designed a 4 × 4 cross tile shown in Fig. 2 to get rid of these problem

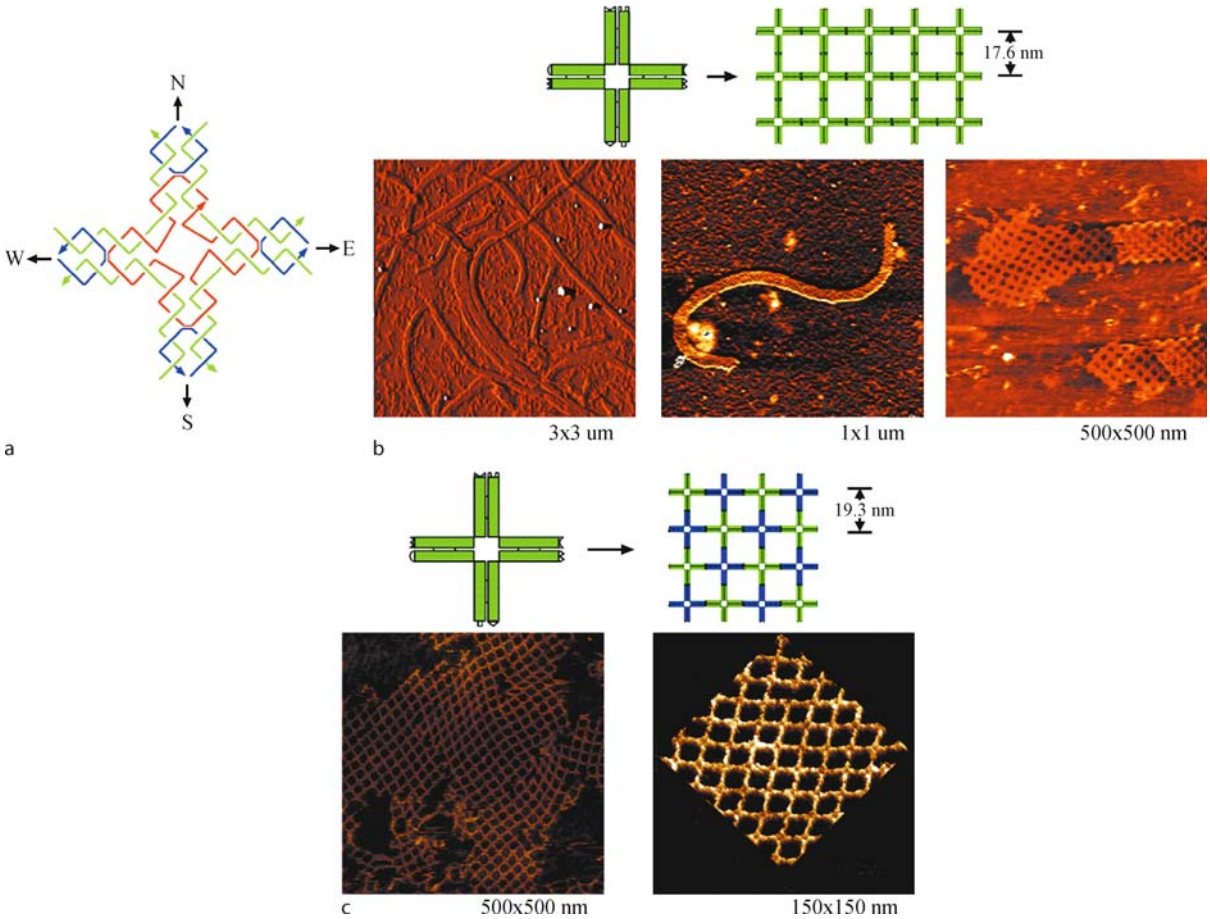

DNA-Templated Self-Assembly of Protein Arrays and Highly Conductive Nanowires, Figure 2

Self-assembly of DNA nanoribbons and nanogrids using 4 × 4 DNA tiles. a The 4 × 4 tile strand structure. The tile contains nine oligonuleotides, shown as simplified backbone traces. One four-arm junction is oriented in each direction (N, S, E, W); the *red strand* participates in all four junctions and contain T_4 loops connecting adjacent junctions. b Self-assembly of nanoribbons with original design. Double-helical domains are illustrated as rectangles, and paired rectangles represent four-arm junctions. Complementary sticky ends are shown as matching geometric shapes. (*Upper right*) Designed structure of self-assembled lattice. (*Bottom*) AFM images of the nanoribbons. The *left panel* shows an amplitude-mode image and the *right* two panels are AFM images in height mode. c Self-assembly of 2D nanogrids with corrugated design. (*Upper left*) The component tile is drawn similar to that in b; positions of sticky ends are changed. The tile has two surfaces; one faces out of the plane (*green*) and the other faces into the plane (*blue*). (*Upper right*) Corrugated self-assembly. (*Bottom*) AFM images of the 2D lattice formed from the corrugated design. The *right panel* is a surface plot of a magnified region from the *left panel*

and grow a lattice with a square aspect ratio [16]. By programming the sticky-end association to vary the assembly strategy, they could control the preferred lattice formation. One strategy produced a high preponderance of uniform-width ribbon structure (Fig. 2b). Their second strategy, referred to as the corrugated design, cause tiles to associate with one another such that the same face of each tile is oriented up and down alternately in neighboring tiles; therefore, the surface curvature inherent in each tile should be canceled out within the assembly (Fig. 2c).

Natural objects have hexagonal 2D packing, but making hexagonal DNA 2D array was long an unfulfilled goal. The simple three-arm junction is floppy and even stiffer motifs were incapable of producing a hexagonal array. In addition, the simple sticky end is susceptible to variations of twist. B. Ding et al. designed a double layer triangle molecule to produce a trigonal-pseudo-hexagonal array [17]. Figure 3 show the structures and AFM image. The honey-comb nature of the arrangement is obvious. A key control experiment was to remove the sticky ends

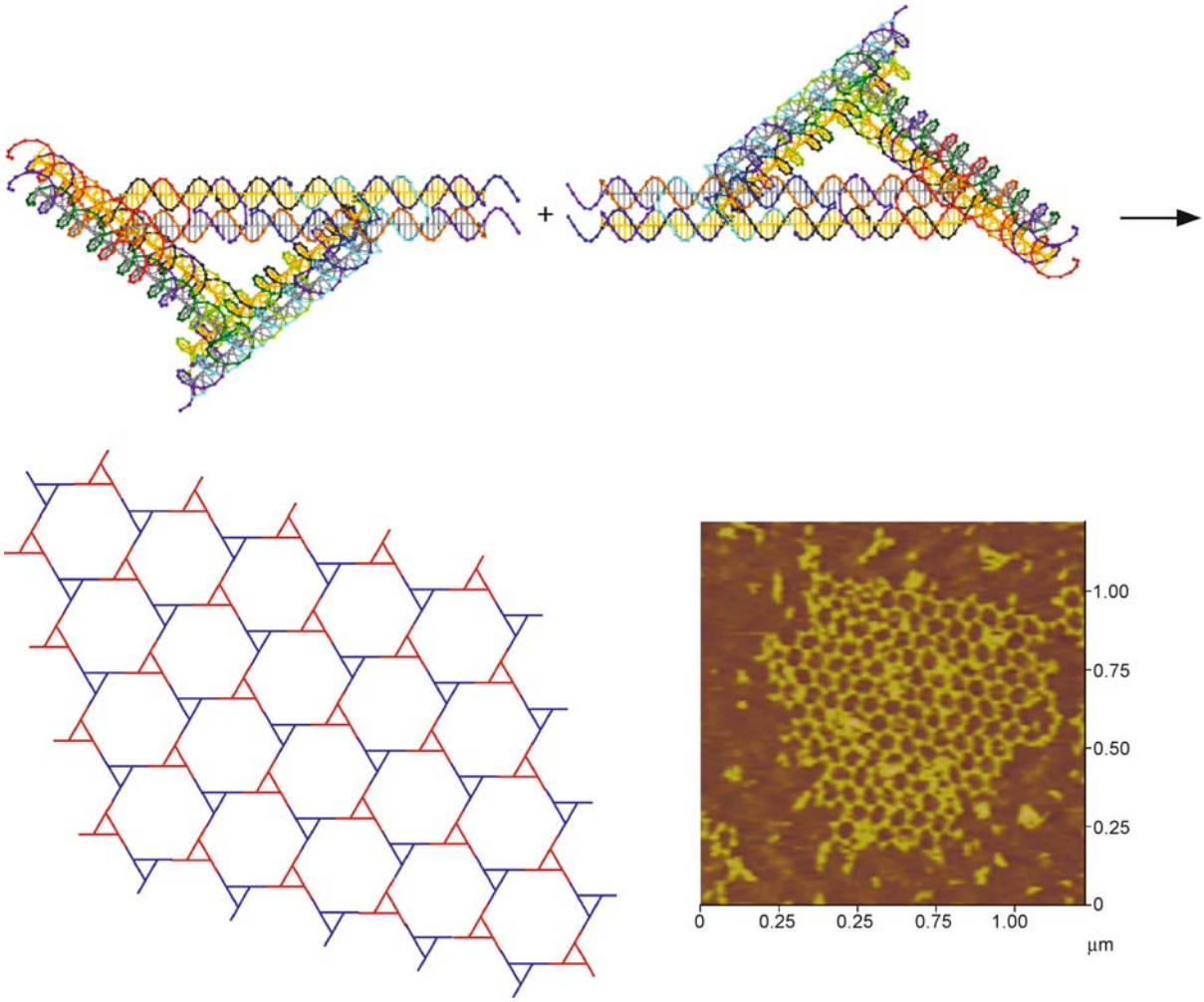

DNA-Templated Self-Assembly of Protein Arrays and Highly Conductive Nanowires, Figure 3
The assembly of a trigonal lattice by DX Cohesion. The *upper part* of the panel shows two triangles with each arm containing Double Crossover structures. They combine in the way shown in the *lower left* to produce a pseudo-hexagonal lattice. The *lower right* shows an AFM image of the array

from one of the two helices in each direction: No array was found, demonstrating that the presence of DX cohesion helps to solve geometrical problems as exist when only a single helix is used.

P. Rothemund illustrated a new method of DNA self-assembly [18], known as DNA origami by folding a long single DNA strand (7249 bases) into well designed shapes such as smiley faces and many other complex 2D patterns. His design is performed in several steps. The first is to build a geometric model of a DNA structure that will approximate the desired shape. The shape is filled by a certain number of parallel helices. Then a single long scaffold strand is folded back and forth so that it comprised one of the two strands in every helix. Short staple strands are then designed, which can hybridize to segments of the scaffold strand and cross over from one row of helix to another like DX molecules. Thus, the staple strands hold the designed pattern together rigidly. Recently, W. Shih extended this ideal and has successfully constructed 3D origami structures [19].

► DNA-Templated Self-Assembly of Protein Arrays and Highly Conductive Nanowires, Figure 4
a Schematic drawing of the TX DNA-templated self-assembly of streptavidin linear array. b AFM images of the DNA-templated protein array. (a) steptavindin alone. (b) bare TX DNA tile arrays. (c) single-layer steptavidin array. (d) double-layer streptavidin array

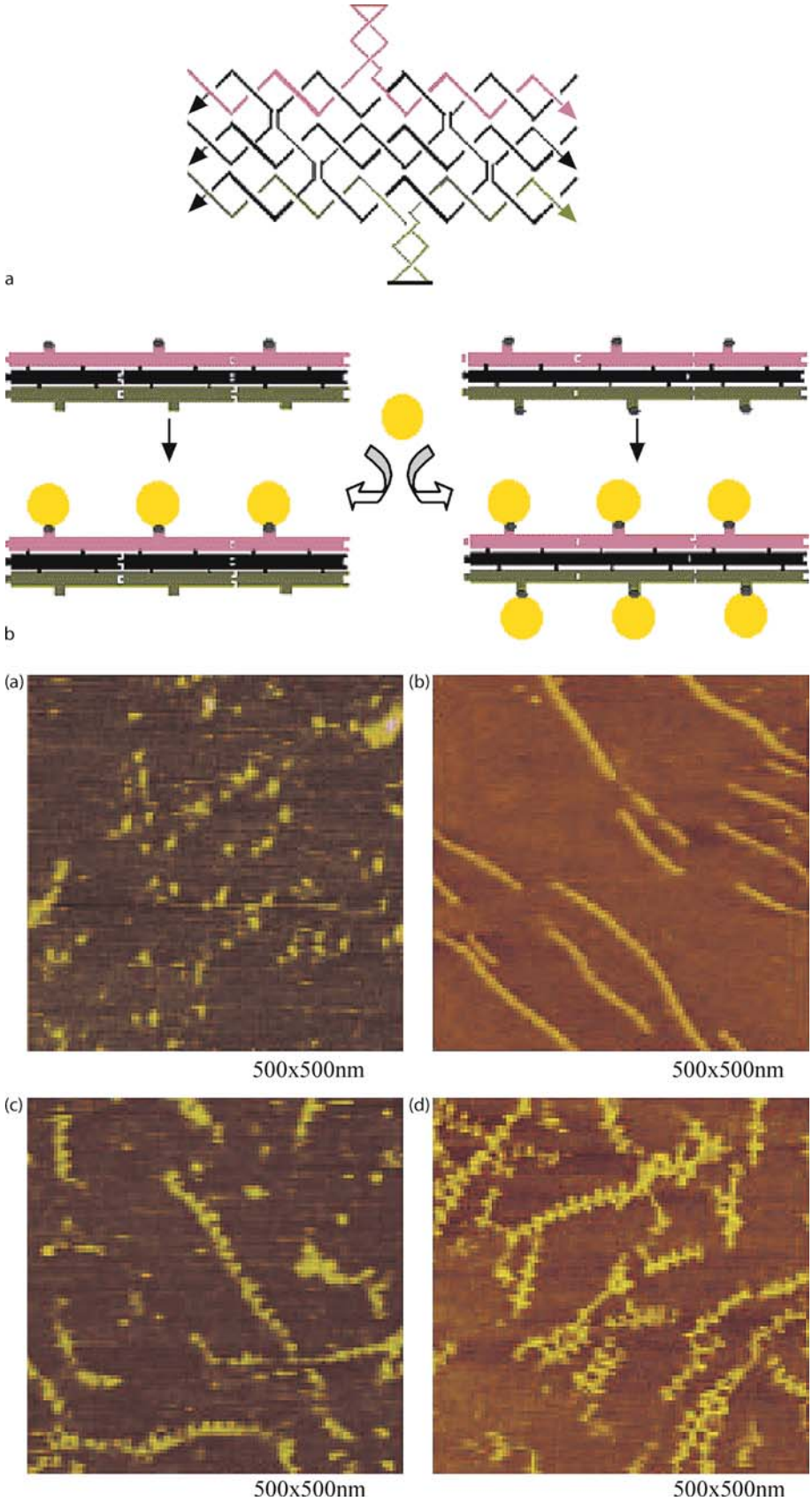
a
b
(a)
(b)
500x500nm
500x500nm
(c)
(d)
500x500nm
500x500nm

DNA-Templated Assembly of Protein Arrays

One of the key goals of DNA self-assembly is to build lattices and other rigid well defined structures from DNA in a controlled way so that DNA can form a designed template on which biological macromolecules can be orderly arrayed. DNA lattice could also bring enzymes and organic molecules into certain proximity, thus catalyze reactions.

H. Li et al. have demonstrated the use of a linear triple crossover (TX) DNA array for the assembly of two forms (single-layer or double-layer) streptavidin arrays through biotin-streptavidin interaction [20]. Figure 4a shows the design. The TX molecule used here consists of seven oligonucleotides hybridized to form three double-stranded helix domains lying in a plane and linked by strand exchange at four immobile crossover points. The TX molecule contains two stem loops protruding on two opposite sides of the TX tile plane. To obtain a single-layer streptavidin array, only the stem loop on one side of each TX tile is modified with biotin; to obtain a double-layer streptavidin array, stem loops on both sides of each TX tiles are modified with biotin. Streptavidin can bind to biotin-modified stem loops specifically and thus self-assemble into single or double-layer arrays respectively. Atomic force microscopy images in Fig. 4b demonstrate the DNA self-assembly TX array templates single-layer and double layer streptavidin particles. The distance between each pair of adjacent particles is about 17 nm, matching the designed parameter.

A 4 × 4 cross tile lattice was used to template the assembly of protein to form 2D array as shown in Fig. 5 [16]. The 4 × 4 DNA tile is modified by incorporating a biotin group into one of the T4 loops at the tile center. When streptavidin was added to the solution of the self-assembled DNA lattice, the interaction of streptavidin with biotin leads to the formation of periodic streptavidin array. Individual protein molecules are visible as separate peaks in the AFM image (Fig. 5). Further design evolution of the 4 × 4 cross tile system to a two tile type (a and b) allowed for more complex structures and patterns [21].

One type of guest molecule attachment has been shown: Biotin-streptavidin interaction is useful to create patterned protein arrays. The design of more diverse surfaces containing different molecular components, each at well-defined and addressable locations remains a challenging problem. As a possible solution, B. Williams et al. have developed a general method to produce high-density peptide arrays that rely on the addressable information encoded in the nucleic acid portion of a DNA-tagged peptide [22]. The idea is to template a capture stand that is protruding out from the DNA self-assembled lattice. Then

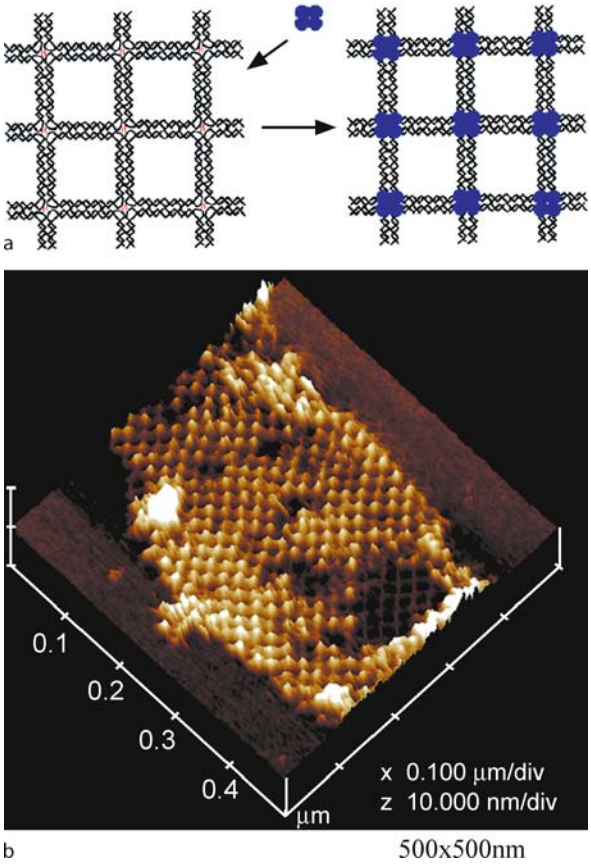

DNA-Templated Self-Assembly of Protein Arrays and Highly Conductive Nanowires, Figure 5

Self-assembly of protein arrays templated by 4 × 4 DNA nanogrids. a Schematic drawing of the DNA nanogrids assembly of streptavidin. (*Left*) The DNA nanogrids, a biotin group labeled as a *red letter* B, are incorporated into one of the loops at the center of each tile. (*Right*) Binding of streptavidin (represented by a *blue tetramer*) to biotin will lead to protein nanoarrays on DNA lattices. b AFM image of the self-assembled protein arrays

DNA-tagged peptide will bind to the pre-designed capture strand through DNA hybridization. The peptide portion will serve as the probe for other protein such as antibody recognition. The DNA template was assembled from ABCD four tile DX motifs. D tile was modified to contain a DNA capturing tag for positioning DNA-peptide fusion at a specific addressable location on the DNA array. The DNA-peptide fusion was constructed by standard peptide coupling chemistry to covalently link the myc-epitope peptide to the 5′-amino-modified DNA strand. The DNA portion of the DNA-tagged peptide is complementary in the sequence to the capture probe on the D tile. Then the anti-myc antibody can be assembled to form 2D structure. The AFM images in Fig. 6 have clearly shown

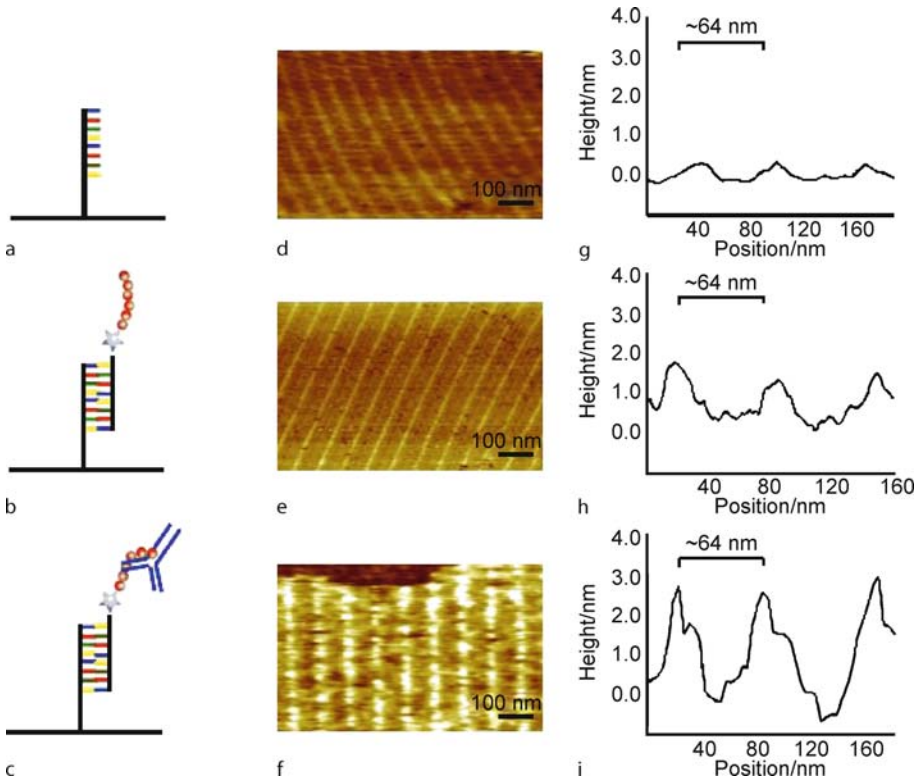

DNA-Templated Self-Assembly of Protein Arrays and Highly Conductive Nanowires, Figure 6
AFM imaging of the peptide nanoarray. a–c Schematic illustration showing the DNA capture probe on the DNA surface, annealed to the myc-peptide fusion, and immunocaptured by the anti-myc antibody, respectively. d–f AFM images were collected for the array before hybridization of the myc-peptide fusion, after hybridization of the myc peptide, and following incubation with the anti-myc antibody, respectively. g–i Corresponding height profiles of the arrays

the binding of DNA-peptide tag and antibody molecules to the DNA template.

DNA-Templated Highly Conductive Nanowires

One potential application of DNA self-assembly is the use of DNA lattices to template assembly of nanoelectronic components, such as metallic nanoparticles [23,24,25,26] and even metallic nanowires. Several groups have demonstrated the successful organization of metallic nanoparticles through DNA template. DNA templated highly conductive nanowired was illustrated by H. Yan et al. [16]. The self-assembled 4×4 nanoribbons serve as the template, then was metalized with silver. The resulting nanowires have been characterized by AFM and SEM (Fig. 7). The metalized nanoriboons have heights of around 35 nm, widths of around 43 nm, and lengths of up to 5 μm. Metal leads were patterned to the wire by electro beam lithography (5 nm Cr followed by 28 nm Au). An SEM image of an actual device is shown in Fig. 7c. A two-terminal current-voltage (I–V) measurement was conducted on this device at room temperature. The I–V curve (Fig. 7c, inset) is linear, demonstrating Ohmic behavior in the range of -0.2 to 0.2 V, and the resistance of this sample is 200 Ω as measured between the two central contacts. This number corresponds to a bulk resistivity of 2.4×10^{-6} Ωm for the silver nanowire. This nanowire is easily reproducible and has markedly higher conductivity than previously reported nanowires [27].

Future Directions

This article has reviewed the current development of structural DNA nanotechnology and DNA templated self-assembly. What is the future direction of this research field? The achievement of several key near-term goals will move DNA nanotechnology to a system with more practical capabilities. First among these goals is the extension of array-making capabilities from 2D to higher order 3D. Also, new DNA templates—to organize guest components in a higher controlled fashion—should be developed for the application of nanoelectronics and nanophotonics. Other

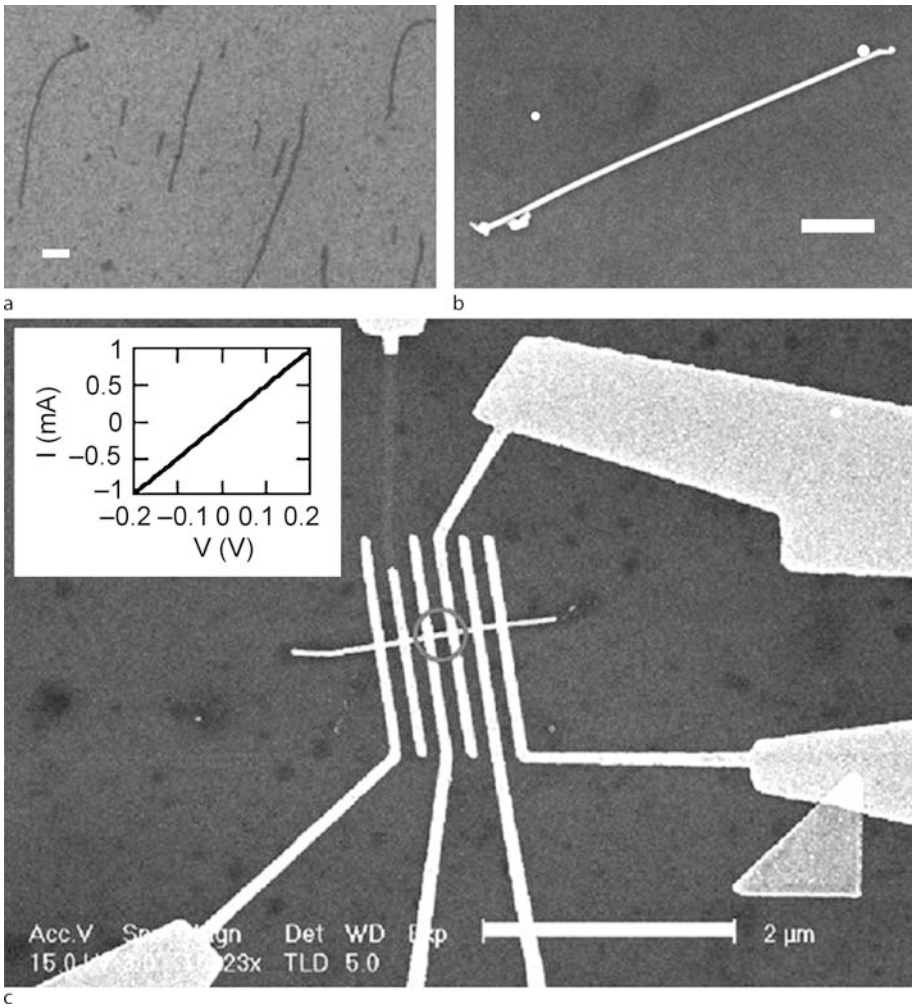

DNA-Templated Self-Assembly of Protein Arrays and Highly Conductive Nanowires, Figure 7
Metalization and conductivity measurements of metalized 4 × 4 ribbons. a SEM images of nonmetalized 4 × 4 DNA nanoribbons (scale bar: 500 nm). b SEM image of silver-seeded silver nanoribbon (scale bar: 500 nm). The change in the signal contrast between a and b is apparent. c SEM image of the actual device (scale bar: 2 μm). (*Inset*) Current-voltage curve of the silver-seeded silver 4 × 4 nanoribbon

goals include further development of binding and release of biological molecules, molecular detection, diagnosis and drug release. It is also a challenge to combine DNA bottom-up assemble with top-down lithography to generate more complex and practical nano-device systems. There are undoubtedly many other directions, and DNA will continue playing its important role in complex nanoscale systems.

Bibliography

1. Seeman NC (2003) DNA in a material world. Nature 421:427–431
2. Seeman NC, Lukeman PS (2005) Nucleic acid nanostructures: Bottom-up control of geometry on the nanoscale. Rep Prog Phys 68:237–270
3. Reif JH (2002) Introduction to self-assembling DNA nanostructures for computation and nanofabrication In: Wang JTL, Wu CH, Wang PP (eds) Computational biology and genome informatics. World Scientific, River Edge
4. Winfree E, Liu F, Wenzler LA, Seeman NC (1998) Design and self-assembly of two-dimensional DNA crystal. Nature 394:539–544
5. LaBean T, Yan H, Kopatsch J, Liu F, Winfree E, Reif JH, Seeman NC (2000) The construction, analysis, ligation and self-assembly of DNA triple crossover complexes. J Am Chem Soc 122:1848–1860
6. Mao C, Sun W, Seeman NC (1999) Two-dimensional DNA hol-

lliday junction arrays visualized by atomic force microscopy. J Am Chem Soc 121:5437–5443
7. Mao C, Sun W, Shen Z, Seeman NC (1999) A DNA nanomechanical device based on the B-Z transition. Nature 397:144–146
8. Yurke B, Turberfield AJ, Mills AP Jr, Simmel FC, Newmann JL (2000) A DNA-fuelled molecular machine made of DNA. Nature 406:605–608
9. Yan H, Zhang X, Shen Z, Seeman NC (2002) A robust DNA mechanical device controlled by hybridization topology. Nature 415:62–65
10. Li JJ, Tan W (2002) A single DNA molecule nanomotor. Nano Lett 2(4):315–318
11. Sherman WB, Seeman NC (2004) A precisely controlled DNA biped walking device. Nano Lett 4:1203–1207
12. Ding B, Seeman NC (2007) Operation of a DNA robot arm inserted into a 2D DNA crystalline substrate. Science 314:1583–1585
13. Adleman LM (1994) Molecular computation of solution to combinatorial problems. Science 266:1021–1024
14. Liu Q, Wang L, Frutos AG, Condon AE, Corn RM, Smith LM (2000) DNA computing on surfaces. Nature 403:175–179
15. Mao C, Labean TH, Reif JH, Seeman NC (2000) Logical computation using algorithmic self-assembly of DNA triple crossover molecules. Nature 407:493–496
16. Yan H, Park SH, Finkelstein G, Reif JH, LaBean TH (2003) DNA-tempated self-assembly of protein arrays and highly conductive nanowires. Science 301:1882–1884
17. Ding B, Sha R, Seeman NC (2004) Pseudohexagonal 2D DNA crystals from double crossover cohesion. J Am Chem Soc 126:10230–10231
18. Rothemund PWK (2006) Folding DNA to create nanoscale shapes and patterns. Nature 440:297–302
19. Douglas SM, Chou JJ, Shih WM (2007) DNA-nanotube-induced alignment of membrane proteins for NMR structure determination. Proc Nat Acad Sci USA 104:6644–6648
20. Li H, Park SH, Rief J, LaBean TH, Yan H (2004) DNA templated self-assembly of protein and nanoparticle linear arrays. J Am Chem Soc 126:418–419
21. Park SH, Yin P, Liu Y, Reif JH, LaBean TH, Yan H (2005) Programmable DNA self-assemblies for nanoscale organization of ligands and proteins. Nano Lett 5(4):729–733
22. William BAR, Lund K, Liu Y, Yan H, Chaput JC (2007) Self-assembled peptide nanoarrays: An approach to studying protein-protein interactions. Angew Chem Int Ed 46:3051–3054
23. Mirkin CA, Letsinger RL, Mucic RC, Storhoff JJ (1996) A DNA-based method for rationally assembling nanoparticles into macroscopic materials. Nature 382:607–609
24. Alivisatos AP, Johnsson KP, Peng X, Wilson TE, Loweth TCJ, Bruchez MP, Schultz PG (1996) Organization of 'nanocrystal molecules' using DNA. Nature 382:609–611
25. Pinto YY, Le JD, Seeman NC, Musier-Forsyth K, Taton TA, Kiehl RA (2005) Sequence-encoded self-assembly of multiple-nanocomponent arrays by 2D DNA scaffolding. Nano Lett 5(12):2399–2402
26. Zheng J, Constantinou PE, Micheel C, Alivisatos AP, Kiehl RA, Seeman NC (2006) Two-dimensional nanoparticle arrays show the organizational power of robust DNA motifs. Nano Lett 6(7):1502–1504
27. Braun E, Eichen Y, Sivan U, Ben-Yoseph G (1998) DNA-templated assembly and electrode attachment of a conducting silver wire. Nature 391:775–778

Drug Design with Artificial Intelligence Methods

Ovidiu Ivanciuc
Department of Biochemistry and Molecular Biology, University of Texas Medical Branch, Galveston, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Genetic Algorithms
Ant Colony Optimization
Particle Swarm Optimization
Artificial Immune Systems
Future Directions
Bibliography

Glossary

Ant colony optimization Ant colony optimization (ACO) is an agent-based algorithm procedure inspired by the function of ant colonies and the ants search for the optimum path to food sources. The virtual agents are called artificial ants or ants, and the optimization problem is represented as a trial and error search for the optimum path on a weighted graph. The pheromone that is deposited by ants on the trail is represented as weights for graph components (vertices or edges). Each ant generates a solution by moving on the graph and by selecting the next step based on the pheromone level. The pheromone level is updated after each cycle (when all ants found a solution) by adding a pheromone quantity proportional to the quality of the solutions to which it belongs.

Antigen An antigen is a molecule (chemical compound, protein or polysaccharide) that induces an immune response. Each pathogen contains specific antigens that are recognized by the immune system. The antigen region that is recognized by the immune system is called an epitope.

Antibody An antibody (or immunoglobulin) is a protein used by the immune system to identify bacteria, viruses and other pathogens or foreign molecules. The antibody region that binds antigens is extremely variable, thus allowing the immune system to recognize a large diversity of pathogens. The ability to recognize antigens is improved through successive cycles of antigen presentation, antibody cloning, and hypermutation of the variable region of the antibody.

Artificial immune systems Artificial immune systems (AIS) represent a class of optimization algorithms inspired by the components and mechanisms of the biological immune system. AIS simulate the learning and memory capabilities of the immune system to develop computational algorithms for pattern recognition, function optimization, classification, process control, and intrusion detection.

Genetic algorithms Genetic algorithms (GA) solve high-dimensional problems through a Darwinian evolution of a population of individuals, in which each individual (chromosome) represents a possible solution. Depending on the type of the optimization problem, chromosomes may represent the solution in a binary, continuous, or hybrid encoding. Each chromosome has a fitness value that measures the quality of the solution. A population of parents evolves to a generation of children by crossover and mutation.

Particle swarm optimization Swarm intelligence (SI) represents a group of distributed intelligence algorithms that solve optimization problems by applying processes inspired by swarming, herding, and flocking of various species. Particle swarm optimization (PSO) simulates the swarming behaviors observed in swarms of bees, flocks of birds, or schools of fish. PSO considers a swarm of particles that start from a random position and have a random velocity. At each step a particle moves to a new position that is determined by its own experience (the best past position) and by the memory of the best particle in the swarm. PSO may be applied to both binary and continuous optimization problems, and its main strength is a fast convergence.

Quantitative structure-activity relationships Quantitative structure-activity relationships (QSAR) represent regression models that define quantitative correlations between the chemical structure of molecules and their physical properties (boiling point, melting point, aqueous solubility), chemical properties and reactivities (chromatographic retention, reaction rate), or biological activities (cell growth inhibition, enzyme inhibition, lethal dose). The fundamental hypotheses of QSAR are that similar chemicals have similar properties, and that small structural changes result in small changes in property values. The general form of a QSAR equation is $P(i) = f(\mathbf{SD}_i)$, where $P(i)$ is a physical, chemical, or biological property of compound i, $\mathbf{SD}_i$ is a vector of structural descriptors of i, and f is a mathematical function such as linear regression, partial least squares, artificial neural networks, or support vector machines. A QSAR model for a property P is based on a dataset of chemical compounds with known values for the property P, and a matrix of structural descriptors computed for all chemicals. The learning (training) of the QSAR model is the process of determining the optimum parameters of the regression function f. After the training phase, a QSAR model may be used to predict the property P for novel compounds that are not present in the learning set of molecules.

Structural descriptor A structural descriptor (SD) is a numerical value computed from the chemical structure of a molecule, which is invariant to the numbering of the atoms in the molecule. Structural descriptors may be classified as constitutional (counts of molecular fragments, such as rings, functional groups, or atom pairs), topological indices (computed from the molecular graph), geometrical (volume, surface, charged-surface), quantum (atomic charges, energies of molecular orbitals), and molecular field (such as those used in CoMFA, CoMSIA, or CoRSA).

Structure-activity relationships Structure-activity relationships (SAR) represent classification models that can discriminate between sets of chemicals that belong to different classes of biological activities, usually active/inactive towards a certain biological receptor. The general form of a SAR equation is $C(i) = f(\mathbf{SD}_i)$, where $C(i)$ is the activity class of compound i (active/inactive, inhibitor/non-inhibitor, ligand/non-ligand), $\mathbf{SD}_i$ is a vector of structural descriptors of i, and f is a classification function such as k-nearest neighbors, linear discriminant analysis, random trees, random forests, Bayesian networks, artificial neural networks, or support vector machines.

Definition of the Subject

Drug design and development represents a complex and expensive process that is based on the creative application of scientific results from various disciplines, including genomics, chemistry, biology, computational chemistry, pharmacology, toxicology, and clinical studies. The average cost of bringing a new drug to market is currently around US$800 million, with a large part of the cost coming from chemical compounds that fail in different stages of development. Computational simulation of biochemical processes may guide the drug discovery process through reliable *in silico* models of biochemical properties (aqueous solubility, octanol-water partition, intestinal absorption, blood-brain barrier transport, excretion), prediction of enzyme-ligand interactions, simulations of cells, tissues and organisms. In this chapter we review the most important applications of artificial intelligence

in structure-activity relationships (SAR) and quantitative structure-activity relationships (QSAR). These techniques are used in different stages of drug design, including large scale screening of chemical libraries, optimization of protein-ligand interactions, modeling the drug transport through membranes, prediction of drug metabolism, mutagenicity, and carcinogenicity. The common goal of artificial intelligence applications in computer-assisted drug design is to identify the best candidates in each step, which may eventually lead to reduced costs for the development of new drugs.

Introduction

Biology is a rich source of inspiration for developing algorithms that solve complex problems by emulating mechanisms and functions of biological systems. Well-known examples of biologically inspired algorithms are artificial neural networks, genetic algorithms, ant colony optimization, DNA computing, particle swarm optimization, and artificial immune systems.

Evolutionary algorithms represent a family of stochastic methods that solve optimization problems by evolving solutions based on Darwinian evolution and concepts of DNA genetics (for details on GA and evolutionary algorithms, see ► Genetic and Evolutionary Algorithms and Programming: General Introduction and Application to Game Playing). The main algorithms from this class are genetic algorithms (GA), genetic programming (GP), and evolutionary programming (EP). The major principles of genetic algorithms were developed by Holland [1], and then further developed by Goldberg [2]. Many applications of chemoinformatics and computational chemistry have a large search space that must be explored to locate the solution. Usually, the brute-force grid search approach cannot be applied but for small systems, and various stochastic methods were developed to find near-optimal solutions. Several examples of high-dimensional problems are the prediction of the biopolymer structure from sequence (peptides, proteins, DNA, RNA), protein-protein docking, protein-ligand docking, conformational search, geometry optimization, design of chemical libraries, and design of chemical compounds with special physico-chemical and biological properties. For other GA applications in chemistry and biology see the reviews by Jones [3], Terfloth [4], and von Homeyer [5]. The most important GA applications in drug development are reviewed in Sect. "Genetic Algorithms".

Dorigo and co-workers developed the ant colony optimization (ACO) algorithm to mimic the foraging behavior of some ant species [6,7,8,9,10]. The main feature modeled in ACO is the ability of an ant population to find the shortest path to a food source using as guide the pheromone trace that is deposited on the path explored by each individual ant. The pheromone accumulates on paths explored more frequently by ants, which indicates that those paths are shorter routes to the food source. ACO has numerous applications, mainly in combinatorial optimization, when their ability to explore large solution spaces is a clear advantage. For theoretical details and applications of agent based simulation, see ► Agent Based Modeling and Simulation, ► Agent Based Modeling, Mathematical Formalism for, and ► Agent Based Modeling and Artificial Life. In Sect. "Ant Colony Optimization" we present an overview of ACO applications in drug design.

The particle swarm optimization (PSO) algorithm proposed by Kennedy and Eberhart is inspired by the social behavior of large groups of individuals, such as bird flocking, fish schooling, and animal herding [11]. Each individual of the group, represented as a particle that moves with a particular velocity through the search space, is a solution for the optimization problem. The movement of each particle is determined by the best position visited by the particle, and by the best position found by the group. The balance between a local and a global search is introduced by weighting the attraction of the best solution of the particle and the best solution of the swarm (for more details on the PSO algorithms, see ► Swarm Intelligence). PSO converges fast and may be used with success to explore high dimensional spaces. The algorithm is simple, with a small number of parameters, and the large number of variants proposed in the literature is a sign of the great interest and vigorous research in this field [12,13]. Swarm intelligence algorithms are used in drug design for diverse applications, including gene expression [14], enzyme-inhibitor docking [15], selection of structural descriptors for QSAR models [16], QSAR with support vector machines optimized with PSO [17], and modeling enzyme inhibitors with artificial neural networks trained with PSO [18]. The most important PSO applications in drug discovery are presented in Sect. "Particle Swarm Optimization".

The immune system protects an organism against infection by identifying and killing pathogens. Recognition cells known as B-cells and T-cells identify the pathogens that enter into the human body. Receptors situated on the surface of the B-cells and T-cells recognize and bind proteins and protein fragments from pathogens, thus forming high affinity antigen-antibody complexes. The learning and memory capabilities of the biological immune system are used in a novel class of machine learning algorithms, the artificial immune systems (AIS) [19,20,21,22,23,24,25,26,27] (for further de-

tails on AIS see ► Immunecomputing. The major AIS algorithms and the most important applications are presented in several books and conference proceedings: *Artificial Immune Systems and Their Applications* edited by Dasgupta [28]; *Artificial Immune Systems: A New Computational Intelligence Approach* by de Castro and Timmis [29]; *Immunocomputing: Principles and Applications*, by Tarakanov, Skormin, and Sokolova [30]; *Immunity-Based Systems* by Ishida [31]; *Artificial Immune Systems: ICARIS 2003* edited by Timmis, Bentley and Hart [32]; *Artificial Immune Systems: ICARIS 2004* edited by Nicosia, Cutello, Bentley, and Timmis [33]; *Artificial Immune Systems: ICARIS 2005* edited by Jacob, Pilat, Bentley, and Timmis [34]; *Artificial Immune Systems: ICARIS 2006* edited by Bersini and Carneiro [35].

AIS models were successfully applied to biological and medical problems, such as classification of gene expression data [36,37,38], identification of breast cancer [39], diagnosis of lung cancer [40], recognition of ECG arrhythmia [41], and interpretation of carotid artery Doppler signals [42]. Protein structure prediction starting from the amino acids sequence is a difficult and computationally intensive task, which was investigated with AIS for models based on Dill's hydrophobic-hydrophilic lattice approach [43] and with three-dimensional models [44]. In Sect. "Artificial Immune Systems" we present a review of the AIS applications in drug design and toxicology.

Genetic Algorithms

Compared with other families of artificial intelligence algorithms, evolutionary algorithms are by far the most popular, with the largest number of publications and with the most diverse applications. GA methods are applied with success to solve diverse drug design problems, such as protein-ligand docking [45], structure-based drug design [46], global optimization of QSAR models based on artificial neural networks [47], computer-aided molecular design [48,49], design of combinatorial libraries [50], and feature selection in QSAR models [51,52]. All these problems are difficult to solve thorough a brute force approach due to the huge search space, but GA are very efficient in finding their global optimum with modest computational resources.

Evolutionary algorithms are used with remarkable results in computer-aided molecular design [48,49] to generate novel molecules with prescribed physical, chemical, or biological properties. In pharmaceutical applications, molecular design is focused on discovering chemical structures that can satisfy all requirements of a successful drug, such as affinity and selectivity for the biological target, and good ADME-Tox (absorption, distribution, metabolism, excretion, and toxicity) properties. The most important part of any molecular design application is a proper encoding of the molecular structure into a chromosome. A straightforward translation of chemicals may be achieved if the molecule can be partitioned into a constant skeleton and a series of substituents, such as the family of chemical compounds **1** (Fig. 1) that has four substituents R_1, R_2, R_3, and R_4. Each molecule from this family may be encoded by a chromosome with four elements ($/R_1/R_2/R_3/R_4/$), with each element recording the index of the respective substituent. Each substitution position has a set of allowed substituents encoded with numbers. For example, compound **2** is represented by the chromosome /3/5/3/1/.

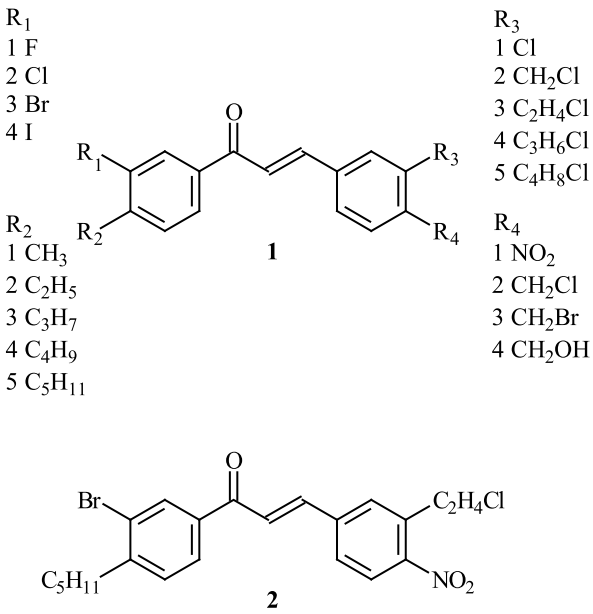

Drug Design with Artificial Intelligence Methods, Figure 1
General formula for a family of chemical compounds (1) that may be encoded with chromosomes having four elements, and an example of molecule (2) from this family

Using this molecular encoding, one can easily define the GA operations of crossover and mutation. The crossover operation involves the exchange of substituents between two parent molecules. For example, parent molecules **3** and **4** generate child molecules **5** and **6** by exchanging substituents R_3 and R_4 (Fig. 2).

The chemical space is also explored with the substituent mutation, as shown in Fig. 3: parent molecule **7** generates child molecule **8** by mutating R_4, and parent molecule **9** generates child molecule **10** by mutating R_1. These examples demonstrate the encoding and evolution of chemical structures in combinatorial libraries of chem-

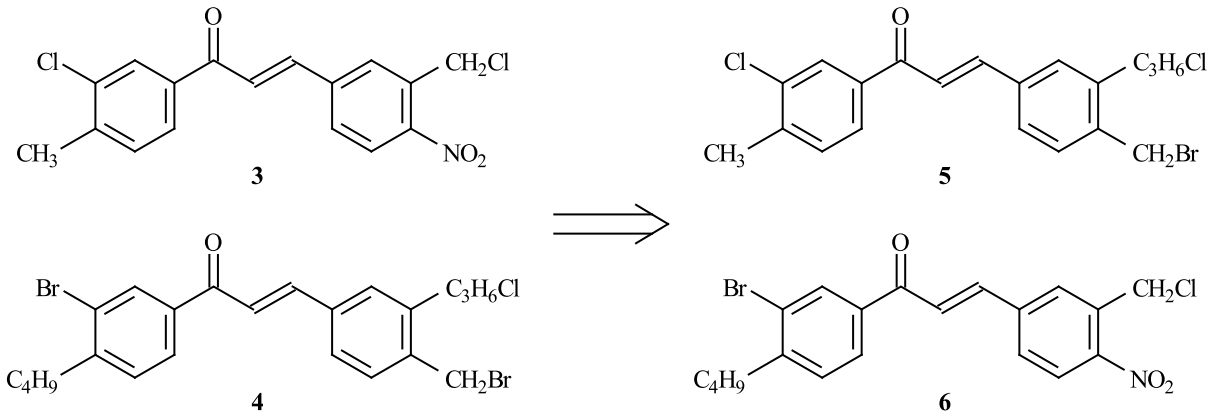

Drug Design with Artificial Intelligence Methods, Figure 2
Example of molecule crossover: parent molecules 3 and 4 generate child molecules 5 and 6 by exchanging substituents R_3 and R_4 (see molecule 1)

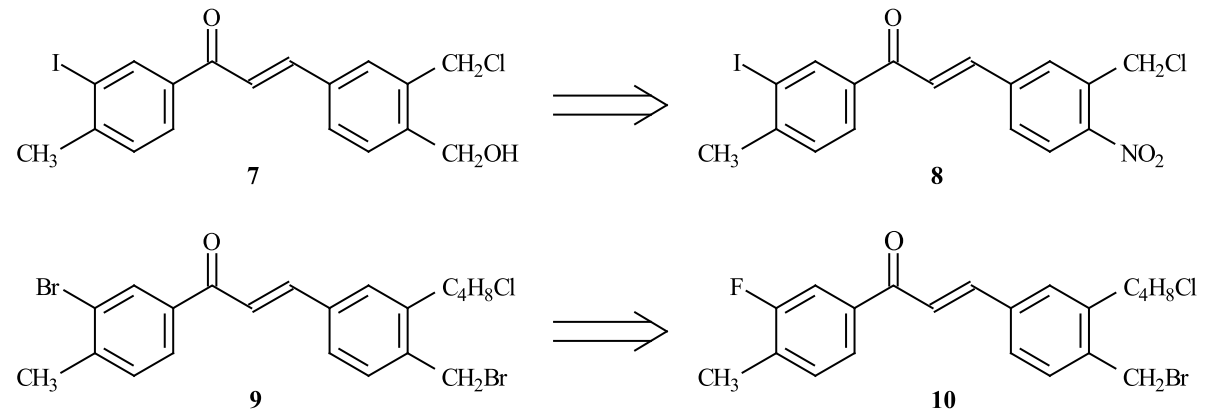

Drug Design with Artificial Intelligence Methods, Figure 3
Examples of molecule mutation: parent molecule 7 generates child molecule 8 by mutating R_4, and parent molecule 9 generates child molecule 10 by mutating R_1

ical compounds [53,54]. The progress in combinatorial chemistry [55,56], virtual screening of chemical libraries, and high throughput techniques dramatically increased the chemical space that can be explored in the quest for molecules with special properties (peptides, nucleic acids, catalysts, pesticides, drugs). Due to the huge chemical space that can be generated through combinatorial chemistry, it is rarely possible to perform an exhaustive synthesis of all possible chemical species. Instead, GA implementations are used to guide the chemical synthesis towards regions containing molecules with target properties [57,58].

An efficient class of reactions that may generate large combinatorial libraries is the Ugi multicomponent reaction (MCR) [59]. Ugi MCRs are one-pot reactions in which three reactants (Fig. 4; U-3CR), four reactants (Fig. 5; U-4CR), or more reactants are converted into the corresponding product without separation and purification of the intermediates. The diversity of chemical structures generated thorough MCR reactions comes from the diversity of the groups R from reactants. Using available chemicals, one can design chemical libraries that are too large to synthesize. Instead, a sample of the combinatorial library is synthesized and evaluated in biological assays, followed by an *in silico* exploration based on GA models [60]. If each reactant type in an U-3CR is a set of 1000 different chemical compounds, then the complete library has 10^9 distinct molecular structures. Similarly, an U-4CR library generated from four sets of 1000 chemicals each consists of 10^{12} distinct compounds. It becomes apparent that the vast chemical space available through combinatorial synthesis is too large even for the *in silico* exploration, which explains why evolutionary algorithms are used to guide the chemical synthesis.

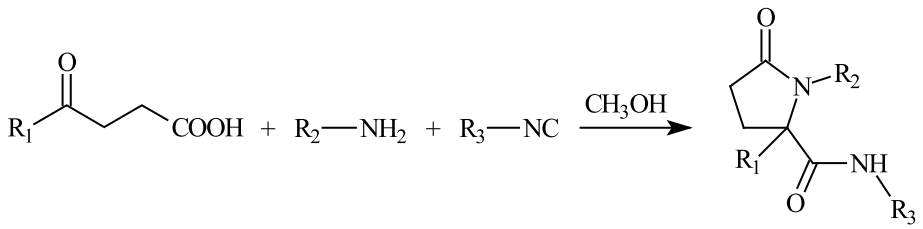

Drug Design with Artificial Intelligence Methods, Figure 4
Example of Ugi 3-component reactions (U-3CR)

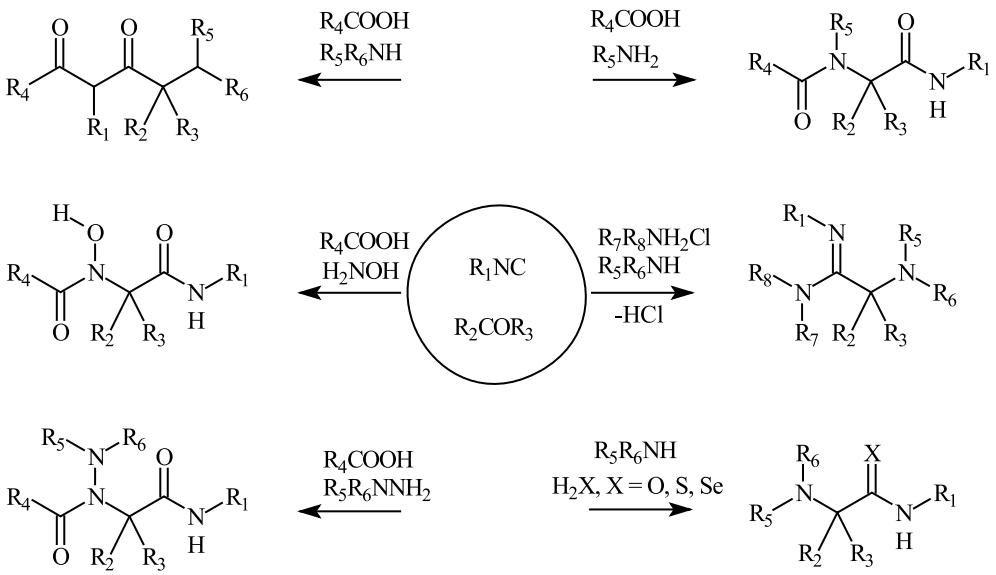

Drug Design with Artificial Intelligence Methods, Figure 5
Examples of Ugi 4-component reactions (U-4CR)

The GA translation of MCR reactions and other combinatorial libraries is straightforward, and in many cases the *in silico* exploration of the chemical space may be performed with standard GA software. This approach has limitations, because the size of the chemical space is fixed by the initial sets of reactants, and the common skeleton remains constant during the simulation. Graph-based GA models solve these limitations by generating molecular structures that are not programmed in the starting building blocks. Such GA systems have crossover and mutation procedures that operate directly on the molecular graph [61], and define a chromosome structure capable to encode a molecular graph [62]. The graph-based GA system proposed by Brown et al. introduces novel crossover and mutation operations for molecular graphs [62] in order to solve the inverse QSAR problem, i. e., to design new chemicals starting from structure-activity models [63]. Four mutations operate on atoms (graph nodes), namely append, prune, insert, and delete (Fig. 6; the site of the transformation is indicated with an arrow). The append mutation adds an atom and its chemical bond to the molecular graph (**11**→**12**). The connecting atom is selected at random from the set of atoms in the molecule that have available valences. The type of the connecting bond is randomly selected from the possible types for the two atoms. The prune mutation removes a terminal atom from the molecular graph (**13**→**14**). The insert mutation selects a bond in the molecular graph, cuts it and inserts a molecular fragment between the two disconnected atoms (**15**→**16**). The molecular fragment is selected from a library, and may consist of a single atom or a more complex subgraph. Additional tests are performed to ensure that the final chromosome (molecular graph) is a valid chemical structure. The delete mutation selects an atom at random, removes it and then reconnects the molecular graph (**17**→**18**). The edge mutations operate on the set of edges in a chromosome (Fig. 7; the site of the transformation is indicated with an arrow). The substi-

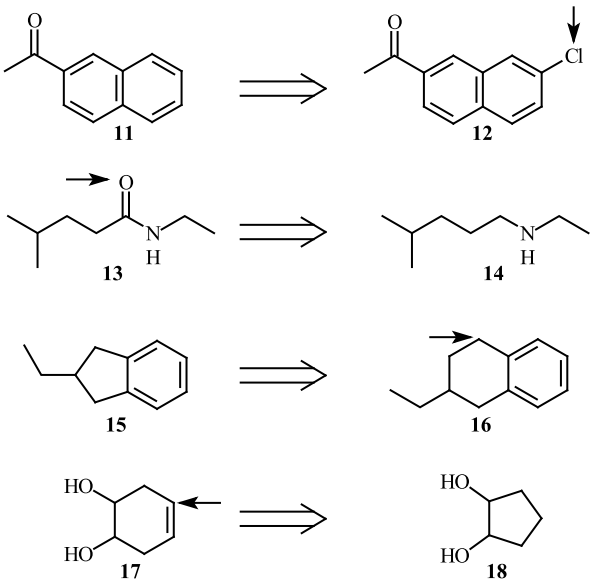

Drug Design with Artificial Intelligence Methods, Figure 6
Examples of the node mutation operators: 11→12, append, 13→14, prune, 15→16, insert, and 17→18, delete

tute mutation selects randomly an edge and then changes its type to another bond type (**19**→**20**). The mutation result must correspond to a correct chemical structure. The add mutation adds a new bond between two atoms (**21**→**22**), thus making possible the generation of cyclic structures. Finally, the delete mutation deletes a bond that was randomly selected (**23**→**24**). The resulting chromosome must represent a connected molecular graph. Two crossover operations are defined for molecular graphs, i. e., multiple crossover and subgraph crossover. The multiple crossover starts from two parent molecules, then each parent molecule is disconnected into two subgraphs, and finally, two child molecules are generated by swapping subgraphs from the parent molecules (Fig. 8). Parent molecule **25** generates subgraphs **27** and **28**, and parent molecule **26** generates subgraphs **29** and **30**. The crossover operation generates child molecule **31** from subgraphs **27** and **30**, and then assembles child molecule **32** from subgraphs **29** and **28**. In the subgraph crossover, a connected subgraph is selected in each parent molecule, and then the subgraphs are combined to obtain the first child molecule (Fig. 9). The combination of the two fragments tries to retain the topology of the two subgraphs. In the second step, a different subgraph is induced in each parent molecule, and the two subgraphs form the second child molecule. The parent molecule **33** generates the induced connected subgraphs **34** and **35**, and the second parent molecule **36** generates the induced connected subgraphs **37** and **38**. The first child molecule **39** is obtained by combining subgraphs **34** and **37**, and the second child molecule **40** is obtained from subgraphs **35** and **38**. The main advantage of the graph-based GA system is its ability to explore chemical structures that are not related to the starting molecules, and to discover novel chemical topologies.

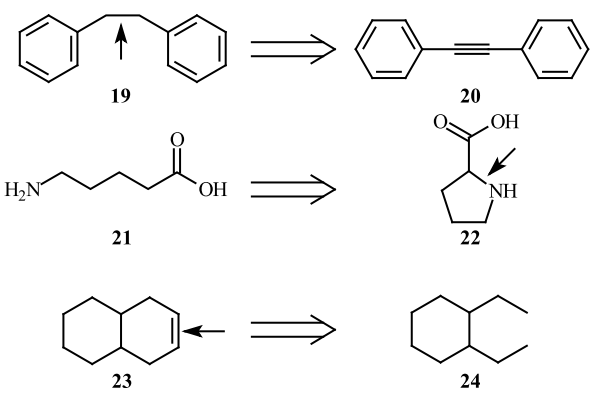

Drug Design with Artificial Intelligence Methods, Figure 7
Examples of the edge mutation operators: 19→20, substitute, 21→22, add, and 23→24, delete

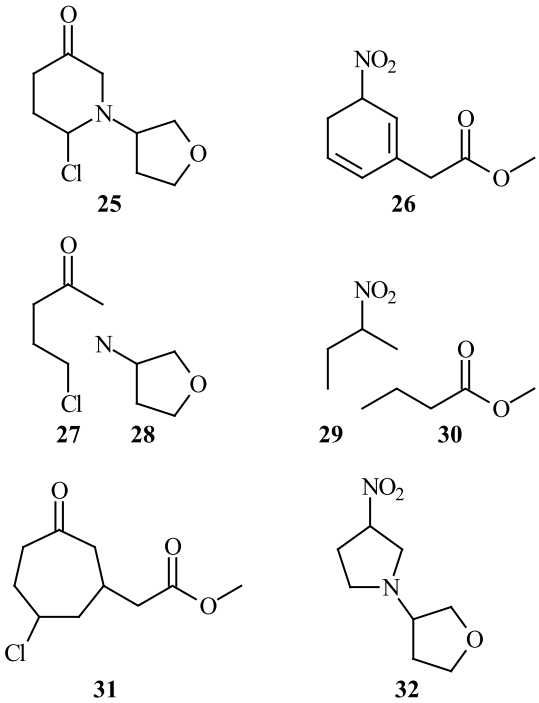

Drug Design with Artificial Intelligence Methods, Figure 8
Example of the multiple crossover: 25 and 26, parent molecules; 27 and 28, disconnected subgraphs of 25; 29 and 30, disconnected subgraphs of 26; 31, child molecule generated from subgraphs 27 and 30; 32, child molecule generated from subgraphs 29 and 28

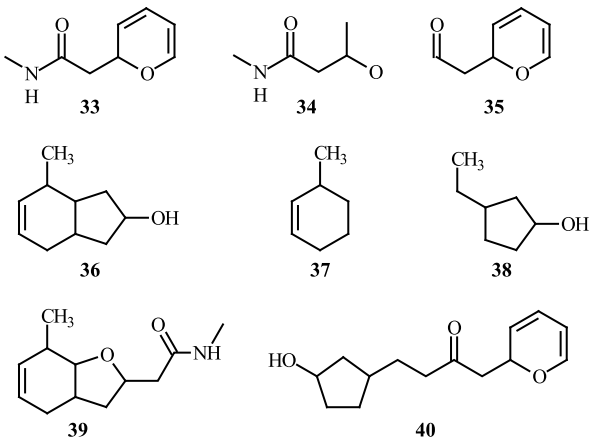

Drug Design with Artificial Intelligence Methods, Figure 9
Example of the subgraph crossover: the first parent molecule 33 and its two induced connected subgraphs 34 and 35; the second parent molecule 36 and its two induced connected subgraphs 37 and 38; the first child molecule 39 generated from subgraphs 34 and 37; the second child molecule 40 generated from subgraphs 35 and 38

Virtual Screening of Chemical Libraries

QSAR models are very useful tools for the identification of structural features that determine various molecular properties, and may even suggest the mechanism of action for biochemical processes. Thus, QSAR models start from the chemical structure and correlate structural descriptors with molecular properties. Once a QSAR model is established, an inverse process becomes possible, namely setting a target value for a molecular property and then finding all possible chemical structures that might exhibit that property value, within a certain range of variation. This process in called inverse QSAR, and it represents an important step in optimizing the drug-like properties of chemical compounds. Lewis proposed an inverse QSAR strategy that may assist medicinal chemists in deciding how to optimize a library of chemical compounds [64]. The starting point is a dataset of chemical compounds with a molecular property, and a corresponding QSAR model. The inverse QSAR strategy involves an iterative application of several steps, namely generation of new structures, structure filtering based on synthetic feasibility or undesired properties, and QSAR filtering. The first step generates a new chemical library by applying simple chemical transformations to the molecules from the initial dataset. Examples of such transformations are modification of the bond order, adding or removing an atom, adding or removing a fragment, or changing C to N or O. The second step filters molecules that have nonspecific reactivity, such as electrophiles, nucleophiles, acylating agents, or redox systems. Synthetic feasibility rules are used to eliminate compounds that are difficult to synthesize or those that are expensive. Finally, QSAR models are used to select candidates for chemical synthesis. The inverse QSAR strategy developed by Lewis was tested for a combinatorial library of 150 inhibitors of human carbonic anhydrase II, that was used to develop a MLR genetic function approximation QSAR, as implemented in Cerius[2]. The best QSAR model is based on five structural descriptors:

$$\mathrm{pIC}_{50} = 7.5 - 0.6\mathrm{PHI} - 5.7\mathrm{Jurs\text{-}RPCG} + 0.2\mathrm{SdsN} + 1.7\mathrm{NaaS} + 0.001\mathrm{Vm}$$
$$n = 150\; r^2 = 0.81\; q^2_{LOO} = 0.80\; F = 127$$

where PHI is the molecular flexibility index, Jurs-RPCG is the charge of the most positive atom divided by the total positive charge, SdsN is the E-state index for sp^2 N, NaaS is the electrotopological count for aromatic S, and Vm is the molecular volume inside the contact surface. This QSAR was used as the starting point for performing automated property optimization.

Ant Colony Optimization

The classical ACO algorithm was successfully modified and adapted in numerous variants to solve specific problems from chemistry and drug design. By far the most important ACO application is represented by the feature selection for QSAR models [65,66]. Several ACO implementations were tested in diverse QSAR models, including multi-linear regression, artificial neural networks, and regression trees. Clustering is routinely used to discover novelty in large chemical datasets, based on structural similarities measured by molecular descriptors. Since similar chemicals usually have similar properties, clustering may suggest groups of molecules that interact with the same biological target. Shelokar et al. proposed a clustering algorithm based on ACO assignment of objects in clusters [67]. Many biochemical problems require optimization of continuous variables, whereas the classical ACO implementation optimizes discrete variables. He et al. demonstrated an ACO extension to continuous variables that may be applied to identify optimum parameters for QSAR models [68]. Korb and co-workers introduced a new protein-ligand docking algorithm, PLANTS (Protein-Ligand ANT System), which uses ACO to find a minimum energy conformation for a protein-ligand complex [69]. Compared with docking algorithms based on GA, PLANTS is faster and finds a larger number of good solutions.

Izrailev and Agrafiotis used an ACO approach to identify the best regression tree models in QSAR [65]. Each

ant represents a regression tree, and the pheromone trail is obtained from a reference tree that represents the topological union of all ant trees simulated. The ACO selection of regression trees was evaluated for three QSAR datasets, namely the antifilarial activity of antimycin analogues, the binding affinities of ligands to benzodiazepine/$GABA_A$ receptors, and the inhibition of dihydrofolate reductase by pyrimidines. Each simulation generated 2000 ant trees and then the tree with the best cross-validation predictions was selected as solution. For all three QSAR datasets the ant tree results were significantly better than those obtained with recursive partitioning and with random trees. Using the same three QSAR datasets, Izrailev and Agrafiotis proposed an ACO procedure (ANTSELECT) for feature selection in artificial neural networks QSAR [66]. A number of 100 independent ANTSELECT simulations were performed for each QSAR dataset, with each simulation containing a population of 2000 ants. Structural descriptors are represented as graph vertices, and an ant generates a path by visiting a number of vertices. All vertices on the path represent the selected structural descriptors that are subsequently used as input to an artificial neural network. Features that give good QSAR models receive a larger quantity of pheromones, thus having greater chances to be selected in subsequent iterations. The QSAR results indicate that the ANTSELECT algorithm provides good solutions if the simulations use a sufficient number of ants to evaluate all features in different combinations. A second requirement is to have a pheromone accumulation that distinguishes between good and bas features. Artificial neural networks are sensitive to the input features, and ANTSELECT provides sets of descriptors that result in models with good predictive power.

Nonsteroidal antiinflammatory drugs (NSAID) treat inflammation and pain by inhibiting both cyclooxygenase-1 and cyclooxygenase-2 (COX2). NSAID have serious side effects, such as gastrointestinal ulceration and bleeding, but the observation that acute and chronic inflammation correlates with higher levels of COX2 prompted several drug design studies to identify selective COX2 inhibitors. Shen et al. proposed a novel ACO procedure for feature selection in a QSAR study of 42 COX2 inhibitors [70]. Starting from 85 structural descriptors, the simulation used 100 ants and 200 iterations to select 3 descriptors for the optimum model. The ACO procedure selected a better set of descriptors, compared with a selection made with an evolutionary algorithm.

The drug binding to human serum albumin (HSA) determines its bioavailability, pharmacokinetics, and therapeutic effect. Many drugs are transported by HSA, but only the free drug has pharmacological effect. Gunturi et al. modeled the HSA binding of 94 diverse drugs starting from a pool of 327 structural descriptors [71]. Since the number of descriptors is too large for a multi-linear regression QSAR, an ACO procedure was implemented to select those features that determine HSA binding. The ACO solutions were cross-validated, and the best QSAR equations with five and six descriptors were selected as final models. The importance of each descriptor was evaluated by the frequency of selection in QSAR models, and it was found that HSA binding depends on hydrophobic interactions, solubility, size, and shape.

Tyrosine kinases are enzymes that transfer a phosphate group from ATP to a tyrosine residue in a protein. These enzymes have important functions in diverse cellular processes, such as metabolism, differentiation, growth, and apoptosis. Shi et al. developed QSAR models for inhibitors of the epidermal growth factor receptor (EGFR), a cell-surface receptor from the tyrosine kinase family [72]. Mutations affecting EGFR expression or activity could result in cancer. The structure of the 61 EGFR inhibitors was encoded with 50 structural descriptors, and ACO was used to select relevant groups of descriptors. The ant population had 100 individuals trained for 200 iterations. The analysis of the descriptors selected with higher frequency by ants reveals the importance of electronic indices, and suggest that electron-donating groups increase the activity of these EGFR inhibitors.

The ability to distinguish between foreign and self proteins is one of the most important characteristics of the immune system. The major histocompatibility complex (MHC) molecules bind short peptides resulting from intracellular processing of foreign and self proteins. The MHC molecule loaded with the peptide migrates to the cell surface where it interacts with T-cell receptors. There are two classes of MHC molecules: (a) MHC class I, which binds peptides derived from endogenously expressed proteins and (b) MHC class II, which binds peptides derived mainly from exogenous or transmembrane proteins. Karpenko et al. devised a novel procedure to predict peptides that bind to MHC II, by using ACO to identify the optimum alignment of a set of variable length peptides [73]. The multiple alignment of all peptides is then utilized to compute a position specific scoring matrix. This matrix assigns different weights to each position and amino acid type, and provides a score for each peptide. Finally, the score is compared with a threshold to determine if the peptide binds or not to MHC II. The predictive power of the scoring matrix was demonstrated on several benchmark datasets, showing that the novel algorithm may be useful to design peptides that bind to MHC II and that may be used in vaccine development.

Major advances in proteomics are a result of significant technological advances in protein purification and mass spectrometry. Another critical component is the automated and reliable protein identification from mass spectrometric data. To improve the protein identification process, Hernandez et al. devised a heuristic algorithm that addresses the difficulties of the current methods, such as poor performance for large databases or for low quality data [74]. The new method based on ACO matches theoretical peptide sequences from a database with a structured representation of the source MS/MS spectrum. Tested with a set of 721 MS/MS spectra, the ACO-based procedure showed success rate of 88.9%, demonstrating that the artificial ants may perform an efficient exploration of the search space.

Particle Swarm Optimization

Particle swarm algorithms are used in diverse biochemistry and drug design applications, to solve problems that require binary or real value optimization. Among the advantages of using PSO in optimization one should count the simple algorithm that translates into small and effective software, fast convergence, small population, and low number of iterations. PSO is applied with success to difficult problems, such as feature selection for gene expression data [14,75], identification of the global minimum geometry of chemical compounds [76], enzyme-inhibitor docking [15], QSAR [16], and protein motif discovery [77].

PSO is an effective replacement of GA for the global optimization of protein-ligand geometry in docking studies. Several PSO modifications of the most popular docking program, AutoDock, were proposed in the literature. The Tribe-PSO algorithm was used in AutoDock to identify the best protein-ligand geometry [78]. In Tribe-PSO the population is divided into several subpopulations or tribes. Each tribe has the same structure and evolution mechanism as the basic PSO model. In the first phase, each tribe evolves independent of the other tribes and converges to an optimum solution. In the second phase, the tribes exchange information regarding the best solution from each tribe, and in the third phase all particles are united into a single population that evolves as a classical PSO model towards the final solution. In a comparative test involving 100 protein-ligand complexes from PDB, over 90% are docked better with Tribe-PSO than with AutoDock. Another PSO modification of AutoDock is SODOCK, which combines the basic PSO model with a local search for the best particle [79]. Compared with four docking methods (GOLD, DOCK, FlexX, and AutoDock) for a set of 37 PDB protein-ligand complexes, SODOCK obtained an average RMSD of 2.29 Å, whereas all other docking programs had an RMSD higher than 3 Å. In a related implementation, PSO@AUTODOCK, AutoDock is combined with a PSO variant that allows larger movements in the search space [15]. Significant improvement is obtained for 12 out of the 37 test complexes, compared with the SODOCK predictions.

Feature selection is an important step in QSAR and in virtual screening of chemical libraries, because almost all QSAR models are sensitive to the presence of irrelevant descriptors. Another benefit of feature selection is the identification of structural descriptors that may explain the mechanism of a particular structure-activity relationship. Agrafiotis and Cedeño used a binary PSO to select descriptors for a QSAR based on multilayer feed-forward artificial neural networks (MLF ANN) [16]. The real value PSO model may also be used for feature selection, as shown for QSAR models based on k-nearest neighbors kernel regression [80]. The target of the PSO model was to find the optimum weight (situated in the range $[0, 1]$) for each structural descriptor. The features with the largest weights were selected in the QSAR model.

In a comparative study for 42 cyclooxygenase inhibitors, Lü et al. found that binary PSO is superior to GA for feature selection in multi-linear regression (MLR) QSAR [81]. Shen et al. showed that the partial least-squares (PLS) QSAR model could be improved by using structural descriptors selected with a binary PSO [82]. Another approach to feature selection is the optimized blockwise variable combination (OBVC) method that performs a descriptor selection guided by PSO followed by PLS modeling of the data [83,84]. Instead of selecting each descriptor independent of the other descriptors, OBVC operates with groups of descriptors. The size and composition of each group of descriptors is optimized with PSO. OBVC was evaluated in QSAR models for the carcinogenic potency of aromatic amines [83] and for inhibitors of lung carcinoma cells [84]. OBVC was also tested for a QSAR dataset consisting of 37 ligands of the $\alpha 6$ benzodiazepine receptor, and more than 70 structural descriptors (topological, geometric, and quantum indices) [85]. Comparative tests show that OBVC exceeds the predictions obtained with MLR, PLS, and hierarchical PLS. OBVC may suggest several combinations of descriptors with comparable prediction statistics, and can assist the discovery of the most important structural descriptors.

PSO is used also to modify and improve QSAR models, such as the piecewise modeling by particle swarm algorithm (PMPSO) which is a QSAR based on piecewise linear models [86]. PMPSO may be useful for datasets with

high structural diversity, when a single linear model for all compounds might not be the best option. A minimum spanning tree model is used to cluster all compounds, and then PSO is applied to divide the tree into predictive piecewise linear models. PMPSO was applied with good results for angiotensin II antagonists. A variant of this QSAR model is the piecewise hypersphere modeling by particle swarm optimization (PHMPSO) which clusters similar compounds in subsets defined as hyperspheres [87]. The position and size of the hyperspheres are optimized with PSO, and then a QSAR model is fitted for the compounds in each hypersphere. PHMPSO was tested with good results for dihydrofolate reductase inhibitors, epidermal growth factor receptor inhibitors, and benzodiazepine receptor ligands [88]. Another PSO-modified algorithm is the optimized sample-weighted PLS (OSWPLS) which uses PSO to weight each object (chemical compound in QSAR) from the training dataset [89]. The weight determines the importance of each object, and the target of the PSO step is to minimize the error of the calibration model.

Training a neural network QSAR consists of (a) finding the best network topology (number of hidden neurons and distribution of the connections between neurons), and (b) optimization of the connection weights. PSO is very efficient in optimizing the ANN weights, as shown in a QSAR study of inhibitors of platelet-derived growth factor receptor phosphorylation [18]. The versatility of the swarm algorithm is practical in a global optimization of ANN QSAR, namely finding the best topology and set of weights. Shen et al. proposed a hybrid use of PSO in training a MLF ANN, namely a binary PSO to determine the optimum network topology and a continuous PSO to find the optimum connection weights [90]. Extensive tests showed that this combination converges quickly and may avoid the overfitting of the learning dataset of chemicals.

The training process of a radial basis function artificial neural network (RBF ANN) consists of selecting the network topology, finding the centers and widths of the RBF neurons, and computing the connection weights between the hidden and output layers. A hybrid particle swarm optimization (HPSO) was used by Zhou et al. to train an RBF network for drug design studies [91]. In the HPSO algorithm, a discrete PSO is used to optimize the network topology, whereas a continuous PSO is used to optimize the network parameters. The new QSAR approach was tested with a dataset of 40 inhibitors of murine P388 leukemia cells and over 70 Cerius^2 descriptors. The HPSO network has the highest predictions: PLS, $r = 0.664$; RBF with parameters optimized with PSO, $r = 0.838$; RBF with parameters optimized with K-means, $r = 0.852$; RBF optimized with HPSO, $r = 0.894$. A similar trend was found for a second QSAR test, performed with 72 cyclooxygenase-2 inhibitors: RBF optimized with PSO, $r = 0.894$; RBF optimized with K-means, $r = 0.903$; RBF optimized with HPSO, $r = 0.921$. The experimental evidence suggests that the hybrid PSO optimization of RBF ANN has a fast convergence to predictive QSAR models.

Zhou et al. proposed a novel version of nonlinear partial least-square method that is based on structural descriptors transformed by an artificial neural network [92]. The structural descriptors represent the ANN input, whereas the output signals from the neurons in the hidden layer represent the nonlinear input for PLS. The ANN weights are trained with PSO. The novel nonlinear QSAR model was tested with good results for two datasets, namely 53 antitumor agents, and 52 benzodiazepine receptor ligands.

As shown in the QSAR studies reviewed here, PSO is an efficient method to optimize linear and nonlinear structure-activity models. A fast convergence to the global minimum depends on the parameters that control the population size, number of iterations, and weights to update the velocity of each particle. Choosing the best parameters that control a PSO model is a meta-optimization problem that was solved by Meissner et al. with the optimized particle swarm optimization (OPSO) model, in which the control parameters are optimized by a meta-swarm [93]. Although OPSO is more complex than a classical PSO because it contains swarms within a swarm, the system converges fast to good QSAR models. OPSO was tested for the prediction of the blood-brain barrier permeation coefficient with a MLF neural network.

Support vector machines (SVM) represent a class of versatile models that can produce nonlinear classification or regression QSAR equations [94]. PSO can be efficiently applied to select the best structural descriptors for SVM models, as demonstrated in a QSAR for P-glycoprotein substrates [17]. The mathematical formalism of SVM was adapted by Lin et al. for the training of MLF ANN [95]. The parameters of the hybrid method SVM-ANN were optimized with PSO, and the new QSAR model was compared with other two algorithms, namely back-propagation ANN (BP-ANN) and ANN optimized with PSO (PSO-ANN). These methods were compared for a dataset of 111 dihydrofolate reductase inhibitors and for another set of 85 cyclooxygenase-2 inhibitors. The results show that SVM-ANN models have better prediction statistics, and that the PSO procedure converges fast to optimum parameters. A similar QSAR model was developed based on radial basis function ANN [96], by defining a nonlinear SVM model (RBF-SVM) representing a kernel transform

based on RBF ANN optimized with PSO. QSAR models obtained for inhibitors of HIV-1 reverse transcriptase demonstrate that RBF-SVM provides better predictions compared to BP-ANN and SVM.

Artificial Immune Systems

The mechanisms and functions of the biological immune system were used as an inspiration for many AIS algorithms, such as the artificial immune network (aiNet) [97,98], the hierarchical artificial immune network (HaiNet) [37], the artificial immune recognition system (AIRS) [99,100,101,102], the clonal selection algorithm (CLONALG) [103,104], the clonal selection classification system (CSCA) [105], IMMUNOS-81 [106], and IMMUNOS-99 [107]. The pattern recognition capabilities of the artificial immune systems may be applied in modeling structure-activity relationships for drug design or for the computational screening of chemical libraries. In the following sections we review several SAR models obtained with AIRS, CLONALG, CSCA, and IMMUNOS. All AIS models were computed with Weka [108].

AIRS – Artificial Immune Recognition System

The AIRS machine learning algorithm developed by Watkins, Timmis, and Boggess is an efficient and popular pattern recognition adaptation of AIS [99,100,101,102]. Brownlee tested AIRS for a wide range of classification problems [109], confirming its utility as a supervised learning classifier. The main characteristics of AIRS are briefly reviewed below.

An antigen is represented as an n-dimensional vector $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$, where each structural descriptor x_i is a real number ($x_i \in R$ for $i = 1, 2, \ldots, n$), and an associated class $y = \{+1, -1\}$. An identical encoding is used for antibodies. An artificial recognition ball (ARB) represents a B-cell, and consists of an antibody, a number of resources, and a stimulation value. The similarity between an ARB and an antigen is measured by the stimulation value. The number of resource from an AIRS model is limited, and ARBs compete for their allocation. Resources are allocated to the most stimulated ARBs by removing them from the least stimulated ARBs, and ARBs without resources are eliminated from the cell population. The ARB population is trained during several cycles of competition for limited resources. In each cycle of ARB training, the best ARB classifiers generate mutated clones that enhance the antigen recognition process, whereas the ARBs with insufficient resources are removed from the population. After training, the top ARB classifiers are selected as memory cells. Finally, the memory cells are used to classify novel antigens (patterns).

The drug design applications reviewed here were obtained with AIRS2, an improved version of AIRS [110]. The AIRS2 algorithm consists of the following steps [109]:

(1) Initialization. In the first phase of the algorithm the system is prepared for the learning process. The training data are normalized between 0 and 1. The Euclidean distance is computed for all pairs of antigens, and then the affinity Af is determined as the ratio between the distance and the maximum distance. The affinity threshold AT is computed as the average affinity for all antigens in the training set. The memory cell pool is populated with randomly selected antigens. At the end of the AIRS algorithm, the memory cell pool represents the recognition ARBs used as classifiers.

(2) Train for all antigens. The AIRS algorithm trains a classifier by passing only once over the entire population of training antigens.

(2.1) Antigen presentation. Each training antigen is presented to the memory cell pool, and each memory cell receives a stimulation value St, $\mathrm{St} = 1 - \mathrm{Af}$. The memory cells with the largest stimulation values are selected, and a number of mutated clones are created and added to the ARB pool. The number of clones NC generated is computed with the formula:

$$\mathrm{NC} = \mathrm{St} \times \mathrm{CR} \times \mathrm{HR}$$

where the clonal rate CR and the hypermutation rate HR are user defined parameters.

(2.2) Competition for limited resources. During this iterative process the algorithm selects those ARBs that have the best recognition capabilities, while optimally allocating the resources to the best ARBs. For each antigen the process trains only those ARBs from the same class with the antigen.

(2.2.1) Perform competition for resources.
The total number of resources is a user defined parameter that limits the number of ARBs.

(2.2.1.1) Stimulation. The selected antigen is presented to all ARBs and the stimulation is computed for each cell in the ARB pool.

(2.2.1.2) Normalization. The ARB stimulation values NSt are normalized.

(2.2.1.3) Allocate limited resources. The amount of resources Rs allocated to each ARB is computed from the normalized stimulation NSt and the clonal rate CR:

$$\mathrm{Rs} = \mathrm{NSt} \times \mathrm{CR}$$

The ARB pool is sorted in the descending order of allocated resources Rs and then resources are removed from the ARB situated at the end of the list until the sum of all allocated resources is lower than the total number of resources.

(2.2.1.4) Remove ARBs with insufficient resources. The ARBs with zero resources are removed from the pool.

(2.2.2) Continue with (2.3) if the stop condition is satisfied. The stop condition for the ARB refinement is met when the average normalized stimulation is higher than a user defined stimulation threshold.

(2.2.3) Generate mutated clones of surviving ARBs. The number of clones generated for each ARB is:

$$NC = St \times CR$$

where St is the stimulation against the antigen, and CR is the clonal rate. The clones undergo a process of hypermutation, during which the elements of the **x** vector are randomly modified to increase the antigen recognition.

(2.2.4) Go to (2.2.1)

(2.3) Memory cell selection. In this step, new ARB classifiers are evaluated for inclusion into the memory cell pool. An ARB is inserted into the memory cell pool if its stimulation value is higher than that of the existing best matching memory cell. The existing best matching memory cell is then removed if the affinity between the candidate ARB and the existing memory cell is less than a cut-off value CutOff computed with the formula:

$$CutOff = AT \times ATS$$

where the affinity threshold AT was computed during the initialization phase, and the affinity threshold scalar ATS is a user defined parameter.

(3) Classification. At the end of the training phase, the memory cell pool represents the AIRS classifier. The classification is performed with a *k*-nearest neighbor method, in which the *k* best matches to a prediction pattern are identified and the predicted class is determined with a majority vote. The parameter *k* is user defined, and may be optimized to maximize the prediction performances.

AIRS was applied with success in several drug design structure-activity relationships that are reviewed here. The classification performance of the AIRS algorithm depends on eight user defined parameters: affinity threshold scalar, clonal rate, hypermutation rate, number of nearest neighbors, initial memory cell pool size, number of instances to compute the affinity threshold, stimulation threshold, and total resources. To illustrate the influence of these parameters, we show the variation of the prediction statistics with the affinity threshold scalar. The statistical indices reported for each AIRS model are: TP_p, true positive in prediction (number of compounds from class +1 classified in class +1); FN_p, false negative in prediction (number of compounds from class +1 classified in class −1); TN_p, true negative in prediction (number of compounds from class −1 classified in class −1); FP_p, false positive in prediction (number of compounds from class −1 classified in class +1); Se_p, prediction selectivity; Sp_p, prediction specificity; Ac_p, prediction accuracy; MCC_p, prediction Matthews correlation coefficient.

Torsade de pointes (TdP) is a polymorphic ventricular arrhythmia that may be caused by drugs that induce the prolongation of the QT interval [111]. QT prolongation and TdP may be caused by a large number of drugs, such as antiarrhythmics, antihistamines, antimicrobials, antidepressants, and antipsychotics. The drug design and development costs may be significantly reduced if, along with other ADME/Tox filters, chemical compounds that have the potential to induce torsade de pointes are eliminated as early as possible. AIRS was applied with success to classify 349 drugs into a subset of 106 drugs that induce torsade de pointes and a subset of 243 drugs that do not induce torsade de pointes [112]. The chemical structure was described with five linear solvation energy relationships (LSER) descriptors, and the prediction of the AIRS models was evaluated with the ten-fold (leave-10%-out) cross-validation. The MCC_p variation with ATS (Table 1) shows that the best predictions are obtained with low values of ATS, in this case ATS = 0.05 (Ac = 0.845, MCC = 0.632). After several steps of optimizations involving the remaining seven parameters, the best AIRS model (Ac = 0.860, MCC = 0.671) has better predictions than 11 other machine learning algorithms.

Drug Design with Artificial Intelligence Methods, Table 1
AIRS prediction statistics for TdP SAR models based on LSER descriptors and computed for various values of the affinity threshold scalar ATS

ATS	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
0.01	76	30	213	30	0.7170	0.8765	0.8281	0.5935
0.05	78	28	217	26	0.7358	0.8930	0.8453	0.6323
0.10	78	28	206	37	0.7358	0.8477	0.8138	0.5710
0.30	71	35	210	33	0.6698	0.8642	0.8052	0.5369
0.50	63	43	203	40	0.5943	0.8354	0.7622	0.4333
0.70	55	51	198	45	0.5189	0.8148	0.7249	0.3394
0.90	49	57	198	45	0.4623	0.8148	0.7077	0.2872

Drug Design with Artificial Intelligence Methods, Table 2
AIRS prediction statistics for TdP SAR models based on 2D/3D descriptors and computed for various values of the affinity threshold scalar ATS

ATS	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
(a) 159 structural descriptors								
0.01	39	46	213	63	0.4588	0.7717	0.6981	0.2173
0.05	42	43	213	63	0.4941	0.7717	0.7064	0.2484
0.09	43	42	218	58	0.5059	0.7899	0.7230	0.2795
0.15	40	45	204	72	0.4706	0.7391	0.6759	0.1924
0.30	40	45	207	69	0.4706	0.7500	0.6842	0.2039
0.50	36	49	203	73	0.4235	0.7355	0.6620	0.1470
0.70	36	49	201	75	0.4235	0.7283	0.6565	0.1396
0.95	38	47	201	75	0.4471	0.7283	0.6620	0.1604
(b) 13 structural descriptors								
0.01	40	45	236	40	0.4706	0.8551	0.7645	0.3327
0.05	42	43	235	41	0.4941	0.8514	0.7673	0.3484
0.09	40	45	227	49	0.4706	0.8225	0.7396	0.2885
0.15	47	38	226	50	0.5529	0.8188	0.7562	0.3558
0.30	30	55	238	38	0.3529	0.8623	0.7424	0.2336
0.50	24	61	243	33	0.2824	0.8804	0.7396	0.1894
0.70	29	56	246	30	0.3412	0.8913	0.7618	0.2668
0.95	30	55	235	41	0.3529	0.8514	0.7341	0.2182

In a related study, AIRS was applied to the classification of 361 drugs (85 induce torsade de pointes, and 276 do not induce torsade de pointes) based on 159 structural indices computed from the molecular structure [113]. The ATS parameter has a significant influence on the AIRS predictions (Table 2a). A series of fivefold (leave-20%-out) cross-validation tests shows that MCC increases from 0.2173 for ATS = 0.01, peaks at 0.2795 for ATS = 0.09, and then decreases to 0.1604 for ATS = 0.95. To investigate the effect of feature selection on the AIRS prediction quality, Weka was used to reduce the number of features to 13 with the combination SubsetEvaluation and BestFirst. Feature selection significantly improves the TdP predictions (Table 2b), with the best predictions obtained for ATS = 0.15 (MCC = 0.356). These results suggest that feature selection should be explored in order to increase the AIRS prediction power.

A good intestinal absorption is a major requirement for oral drugs [114,115], and various computational models were proposed as fast, reliable, and inexpensive *in silico* methods to assess the intestinal permeability of a chemical compound before synthesis [116,117]. The oral absorption of a drug is influence by a large number of variables, such as drug formulation and stability, aqueous solubility, contents of the gastrointestinal tract, residence time in the intestine, intestinal metabolism, rate of passive intestinal permeability, carrier-mediated influx, and active efflux via transporters. The human intestinal absorption (HIA) of 196 drugs (131 drugs that penetrate the human intestine, and 65 drugs that do not penetrate the intestine) was modeled with the AIRS algorithm [118]. The AIRS classifiers were obtained with 159 structural descriptors from five classes, namely constitutional, topological indices, electrotopological state indices, quantum descriptors, and geometrical indices. The influence of the ATS parameter in L20%O cross-validation was investigated for values between 0.01 and 0.95 (Table 3a). As in previous experiments, MCC increases from 0.3174 for ATS = 0.01 to a maximum of 0.3506 for ATS = 0.09, and then decreases to 0.1997 for ATS = 0.95. After optimizing all eight parameters, the best predictions of the AIRS algorithm (Ac = 0.735, MCC = 0.406) are higher than those obtained with seven other machine learning algorithms, namely Bayesian network, naïve Bayes classifier, updateable naïve Bayes classifier, logistic regression, Gaussian radial basis function network, decision tree with naïve Bayes classifiers at the leaves, and random tree. In a feature selection experiment (SubsetEvaluation and BestFirst) the number of structural descriptors was reduced to 21, which improved considerably the AIRS predictions [119]. The results obtained for the ATS parameter (Table 3b) show a significant increase across the entire range of values, with a maximum of 0.53 for ATS = 0.04.

Drug Design with Artificial Intelligence Methods, Table 3
AIRS prediction statistics for HIA SAR models computed for various values of the affinity threshold scalar ATS

ATS	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
(a) 159 structural descriptors								
0.01	105	26	33	32	0.8015	0.5077	0.7041	0.3174
0.04	107	24	33	32	0.8168	0.5077	0.7143	0.3364
0.09	107	24	34	31	0.8168	0.5231	0.7194	0.3506
0.15	107	24	28	37	0.8168	0.4308	0.6888	0.2640
0.30	100	31	30	35	0.7634	0.4615	0.6633	0.2287
0.50	105	26	25	40	0.8015	0.3846	0.6633	0.1997
0.70	105	26	25	40	0.8015	0.3846	0.6633	0.1997
0.95	105	26	25	40	0.8015	0.3846	0.6633	0.1997
(b) 21 structural descriptors								
0.01	113	18	41	24	0.8626	0.6308	0.7857	0.5064
0.04	114	17	42	23	0.8702	0.6462	0.7959	0.5300
0.09	113	18	39	26	0.8626	0.6000	0.7755	0.4796
0.15	111	20	40	25	0.8473	0.6154	0.7704	0.4727
0.30	116	15	30	35	0.8855	0.4615	0.7449	0.3885
0.50	121	10	30	35	0.9237	0.4615	0.7704	0.4500
0.70	120	11	28	37	0.9160	0.4308	0.7551	0.4090
0.95	123	8	28	37	0.9389	0.4308	0.7704	0.4495

P-glycoprotein (Pgp) is responsible for the low cellular accumulation of anticancer drugs, for reduced oral absorption, for low blood-brain barrier penetration, and is involved in hepatic, renal, or intestinal elimination of drugs. Computational methods for the identification of Pgp substrates are useful drug design tools for the early elimination of potential Pgp substrates [120,121]. The immune system classifier AIRS was used to discriminate between 116 Pgp substrates and 85 Pgp nonsubstrates [122]. The SAR models were computed from 159 structural descriptors and the prediction power was estimated with L20%O cross-validation. Low values for the ATS parameter give better predictions, with the highest predictions obtained for ATS = 0.03 (Table 4a). The AIRS model optimized for all eight parameters (Ac = 0.702, MCC = 0.380) is better than five machine learning algorithms (alternating decision tree, Bayesian network, logistic regression with ridge estimator, random tree, and fast decision tree learner), demonstrating that Pgp substrates may be successfully recognized with AIRS. A feature selection step reduces the number of structural descriptors from 159 to 15, and increases the SAR performances over the entire range of ATS values (Table 4b) [119]. The best predictions are obtained with ATS = 0.01 (Ac = 0.736 and MCC = 0.467), but the variation of the prediction statistics is not monotonous with ATS, and no simple rule can be extracted to guide further experiments.

Drug Design with Artificial Intelligence Methods, Table 4
AIRS prediction statistics for Pgp SAR models computed for various values of the affinity threshold scalar ATS

ATS	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
(a) 159 structural descriptors								
0.01	85	31	45	40	0.7328	0.5294	0.6468	0.2671
0.03	85	31	48	37	0.7328	0.5647	0.6617	0.3009
0.07	80	36	46	39	0.6897	0.5412	0.6269	0.2320
0.15	78	38	51	34	0.6724	0.6000	0.6418	0.2709
0.30	80	36	42	43	0.6897	0.4941	0.6070	0.1863
0.50	80	36	44	41	0.6897	0.5176	0.6169	0.2092
0.70	80	36	43	42	0.6897	0.5059	0.6119	0.1978
0.95	80	36	43	42	0.6897	0.5059	0.6119	0.1978
(b) 15 structural descriptors								
0.01	86	30	62	23	0.7414	0.7294	0.7363	0.4668
0.03	85	31	59	26	0.7328	0.6941	0.7164	0.4241
0.07	85	31	54	31	0.7328	0.6353	0.6915	0.3681
0.15	88	28	58	27	0.7586	0.6824	0.7264	0.4403
0.30	87	29	57	28	0.7500	0.6706	0.7164	0.4199
0.50	81	35	56	29	0.6983	0.6588	0.6816	0.3544
0.70	78	38	58	27	0.6724	0.6824	0.6766	0.3509
0.95	80	36	56	29	0.6897	0.6588	0.6766	0.3455

Drug Design with Artificial Intelligence Methods, Table 5
AIRS prediction statistics for BZR SAR models computed for various values of the affinity threshold scalar ATS

ATS	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
(a) 75 structural descriptors								
0.01	63	19	52	29	0.7683	0.6420	0.7055	0.4137
0.05	64	18	53	28	0.7805	0.6543	0.7178	0.4385
0.10	62	20	52	29	0.7561	0.6420	0.6994	0.4008
0.20	60	22	57	24	0.7317	0.7037	0.7178	0.4356
0.25	65	17	56	25	0.7927	0.6914	0.7423	0.4867
0.30	62	20	56	25	0.7561	0.6914	0.7239	0.4485
0.50	63	19	56	25	0.7683	0.6914	0.7301	0.4611
0.95	63	19	56	25	0.7683	0.6914	0.7301	0.4611
(b) 16 structural descriptors								
0.01	64	18	57	24	0.7805	0.7037	0.7423	0.4857
0.05	65	17	57	24	0.7927	0.7037	0.7485	0.4985
0.10	57	25	55	26	0.6951	0.6790	0.6871	0.3742
0.20	62	20	65	16	0.7561	0.8025	0.7791	0.5591
0.25	60	22	62	19	0.7317	0.7654	0.7485	0.4974
0.30	61	21	61	20	0.7439	0.7531	0.7485	0.4970
0.50	59	23	62	19	0.7195	0.7654	0.7423	0.4854
0.95	59	23	61	20	0.7195	0.7531	0.7362	0.4728

Another successful application of AIRS in drug design was reported for the identification of benzodiazepine receptor (BZR) ligands [123]. The structure of the 163 BZR ligands was encoded with 75 structural descriptors, and AIRS classifiers were trained to discriminate between 82 high affinity ligands (class +1, pIC_{50} between 8.92 and 7.80) and 81 low affinity ligands (class −1, pIC_{50} between 7.77 and 5). A scan of the ATS values (Table 5a) shows that the best predictions are obtained for ATS = 0.025 (Ac = 0.7423 and MCC = 0.4867). The feature selection step further reduces the number of structural descriptors to 16 (Table 5b), and results in better predictions (best ATS = 0.20, with Ac = 0.7791 and MCC = 0.5591).

Numerous organic chemicals are environmental pollutants, and a considerable number of studies are dedicated to the computational prediction of their mechanism of aquatic toxicity (MOA). The reliable prediction of MOA has major applications in selecting the appropriate QSAR model, to identify chemicals with similar toxicity mechanism, and in extrapolating toxic effects between different species and exposure regimens [124,125]. The immune system AIRS was applied for the MOA prediction of 187 chemicals (143 non-polar narcotics, and 44 polar narcotics) [126]. The chemical structure was described with five LSER descriptors, and the AIRS predictions were evaluated with the ten-fold cross-validation. The ATS parameter was modified between 0.01 and 0.95 (Table 6), and the

Drug Design with Artificial Intelligence Methods, Table 6
AIRS prediction statistics for MOA SAR models computed for various values of the affinity threshold scalar ATS

ATS	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
0.01	138	5	40	4	0.9650	0.9091	0.9519	0.8674
0.02	138	5	40	4	0.9650	0.9091	0.9519	0.8674
0.05	138	5	40	4	0.9650	0.9091	0.9519	0.8674
0.15	135	8	39	5	0.9441	0.8864	0.9305	0.8120
0.30	139	4	39	5	0.9720	0.8864	0.9519	0.8653
0.50	138	5	39	5	0.9650	0.8864	0.9465	0.8514
0.70	137	6	39	5	0.9580	0.8864	0.9412	0.8379
0.95	137	6	39	5	0.9580	0.8864	0.9412	0.8379

best predictions were obtained for low ATS values (0.01, 0.02, and 0.05), namely Ac = 0.9519 and MCC = 0.8674. Based on the high prediction rates obtained with AIRS, such models may be used to identify the aquatic toxicity mechanism and to select the appropriate computational model for new chemical compounds.

CLONALG – Clonal Selection Algorithm

An AIS algorithm that gives a central role to the clonal selection theory is CLONALG, proposed by de Castro and Von Zuben [103,104]. CLONALG implements several mechanisms of the clonal selection: training of a group of memory cells; identification and cloning of the antibodies with the highest recognition power; death of the antibodies with low recognition power; cloning and hypermutation of the antibodies with high recognition power; evaluation and replacement of the clones; generation and preservation of antibody diversity. The CLONALG algorithm, as implemented by Brownlee, consists of the following steps [105]:

(1) Initialization. The CLONALG algorithm starts by generating a pool of N antibodies, which is subsequently partitioned into the memory antibody pool (MAP) and the remaining antibody pool (RAP). MAP contains m antibodies, and at the end of the training process it will represent the solution of the CLONALG classifier. RAP contains the remaining antibodies, $r = N - m$, and it has the role of adding additional diversity during the learning phase.

(2) Train antibodies. The main part of the CLONALG algorithm is an iterative process of exposing the system to all antigens from the training set for a number of G generations (iterations).

(2.1) Train for each antigen. Repeat steps (2.2)–(2.9) for all antigens in the training set. In each generation, an antigen is selected for training once and only once.

(2.2) Antigen selection. For each generation, an antigen is randomly selected without replacement from the entire pool of antigens.

(2.3) Affinity calculation. The selected antigen interacts with all antibodies, and the affinity is calculated for the interaction between the antigen and every antibody in the system. The affinity measures the similarity between an antigen and an antibody, and is based on the Euclidean distance between the vectors of structural descriptors that characterize the antigen and the antibody.

(2.4) Select antibodies. The antibodies are ranked according to their decreasing affinity towards the antigen, and the top n antibodies are selected for further processing.

(2.5) Clone antibodies. All n antibodies selected in the previous step are cloned proportionally with their affinity. The number of clones computed for an antibody that is ranked ith according to its affinity, with $i \in [1, n]$, is

$$\mathrm{Nc} = \left\lfloor \frac{\mathrm{CF} \times N}{i} + 0.5 \right\rfloor$$

where CF is the clonal factor. The total number of clones generated for the entire system of n antibodies is:

$$\mathrm{NC} = \sum_{i=1}^{n} \mathrm{Nc}\,.$$

(2.6) Affinity maturation. The clones enter the process of affinity maturation, during which random mutations are performed onto each clone in order to increase its affinity towards the antigen. The degree of affinity maturation is inversely proportional to the initial affinity, namely the lower the initial affinity the greater the mutation rate is.

(2.7) Evaluate clones. All clones are exposed to the antigen to compute their affinity.

(2.8) Select candidates. The antibodies with the highest affinity are selected to replace antibodies from MAP that have lower affinities.

(2.9) Replacement. The RAP group of antibodies is ranked according to the decreasing affinity towards the antigen, and the set of s antibodies with the lowest affinity is replaces with random antibodies.

(3) Classification. After training the system for G generations, the MAP group of antigens represents the solution of the CLONALG classifier.

The CLONALG machine learning was tested with success in drug design applications, namely recognition of

glycogen phosphorylase B inhibitors, classification of benzodiazepine receptor ligands, and identification of polar and nonpolar narcotic pollutants. To illustrate the effect of the user defined parameters on the prediction performance of CLONALG, we show the influence of the clonal factor CF on the L20%O cross-validation statistics. The clonal factor is a scaling factor, with values between 0 and 1, that determines the number of clones generated for each selected antibody. Low values for CF result in a local search, whereas for high values the algorithm generates a larger number of clones that may explore a wider region and result in a higher diversity.

CLONALG was applied in drug development for the recognition of glycogen phosphorylase B (GPB) inhibitors, based on a set of 66 compounds and 70 structural descriptors [127]. The subset of active compounds contains 33 chemicals (class +1, pKi between 6.8 and 2.5), whereas the subset of inactive compounds contains the remaining 33 chemicals (class −1, pK_i between 2.4 and 1.3). The prediction performance depends on the number of clones generated, controlled by the values of CF (Table 7a), with best results for CF = 0.50 (Ac = 0.6212 and MCC = 0.2453), whereas low and high values for CF result in lower predictions. A feature selection step drastically reduces the number of structural descriptors from 70 to 2, while the model prediction increases (Ac = 0.6667 and MCC = 0.3339) for several CF values (Table 7b).

Drug Design with Artificial Intelligence Methods, Table 7
CLONALG prediction statistics for GPB SAR models computed for various values of the clonal factor CF

CF	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
(a) 70 structural descriptors								
0.01	20	13	15	18	0.6061	0.4545	0.5303	0.0613
0.05	22	11	17	16	0.6667	0.5152	0.5909	0.1839
0.08	22	11	14	19	0.6667	0.4242	0.5455	0.0937
0.15	20	13	17	16	0.6061	0.5152	0.5606	0.1217
0.25	23	10	15	18	0.6970	0.4545	0.5758	0.1562
0.50	23	10	18	15	0.6970	0.5455	0.6212	0.2453
0.65	24	9	16	17	0.7273	0.4848	0.6061	0.2186
0.95	22	11	14	19	0.6667	0.4242	0.5455	0.0937
(b) 2 structural descriptors								
0.01	21	12	22	11	0.6364	0.6667	0.6515	0.3032
0.05	21	12	23	10	0.6364	0.6970	0.6667	0.3339
0.08	21	12	23	10	0.6364	0.6970	0.6667	0.3339
0.15	21	12	23	10	0.6364	0.6970	0.6667	0.3339
0.25	21	12	23	10	0.6364	0.6970	0.6667	0.3339
0.50	21	12	22	11	0.6364	0.6667	0.6515	0.3032
0.65	21	12	23	10	0.6364	0.6970	0.6667	0.3339
0.95	21	12	22	11	0.6364	0.6667	0.6515	0.3032

Drug Design with Artificial Intelligence Methods, Table 8
CLONALG prediction statistics for BZR SAR models computed for various values of the clonal factor CF

CF	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
(a) 75 structural descriptors								
0.01	56	26	48	33	0.6829	0.5926	0.6380	0.2767
0.05	57	25	50	31	0.6951	0.6173	0.6564	0.3134
0.10	58	24	51	30	0.7073	0.6296	0.6687	0.3380
0.20	56	26	52	29	0.6829	0.6420	0.6626	0.3252
0.45	56	26	53	28	0.6829	0.6543	0.6687	0.3374
0.60	62	20	52	29	0.7561	0.6420	0.6994	0.4008
0.85	57	25	47	34	0.6951	0.5802	0.6380	0.2773
0.95	63	19	47	34	0.7683	0.5802	0.6748	0.3550
(b) 16 structural descriptors								
0.01	42	40	53	28	0.5122	0.6543	0.5828	0.1682
0.05	57	25	53	28	0.6951	0.6543	0.6748	0.3498
0.10	53	29	55	26	0.6463	0.6790	0.6626	0.3255
0.20	57	25	54	27	0.6951	0.6667	0.6810	0.3620
0.45	64	18	52	29	0.7805	0.6420	0.7117	0.4267
0.60	61	21	52	29	0.7439	0.6420	0.6933	0.3880
0.85	51	31	55	26	0.6220	0.6790	0.6503	0.3014
0.95	57	25	55	26	0.6951	0.6790	0.6871	0.3742

The CLONALG immune system was tested for the classification of 163 benzodiazepine receptor (BZR) ligands (82 high affinity ligands and 81 low affinity ligands) which are encoded with 75 structural descriptors [123]. The clonal factor was modified between 0.01 and 0.95 (Table 8a). The prediction MCC increases from 0.2767 for CF = 0.01, peaks at 0.4008 for CF = 0.60, and then decreases to 0.2888 for CF = 0.90. These results indicate that too few or too many clones are detrimental to the antigen recognition. The number of structural descriptors can be significantly reduced to 16 by feature selection (Table 8b), which also results in a slight increase of the prediction quality (MCC = 0.4267 for CF = 0.45). The optimum CF is situated in the middle of the range of CF values, similarly with the results obtained for the identification of GPB inhibitors.

The mechanism of toxic action of polar and nonpolar narcotic pollutants may be efficiently identified with CLONALG classifiers [128]. The dataset consists of 190 compounds (114 nonpolar pollutants, class +1; 76 polar pollutants, class −1), with each chemical characterized by five structural descriptors, namely the octanol-water partition coefficient, the energy of the highest occupied molecular orbital, the energy of the lowest unoccupied molecular orbital, the most negative partial charge on any nonhydrogen atom in the molecule, and the most positive partial charge on a hydrogen atom. The prediction MCC has no clear-cut variation with CF (Table 9), but the optimum

Drug Design with Artificial Intelligence Methods, Table 9
CLONALG prediction statistics for MOA SAR models computed for various values of the clonal factor CF

CF	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
0.01	98	16	73	3	0.8596	0.9605	0.9000	0.8052
0.05	100	14	72	4	0.8772	0.9474	0.9053	0.8116
0.10	103	11	75	1	0.9035	0.9868	0.9368	0.8763
0.15	105	9	68	8	0.9211	0.8947	0.9105	0.8141
0.30	104	10	72	4	0.9123	0.9474	0.9263	0.8503
0.55	110	4	69	7	0.9649	0.9079	0.9421	0.8791
0.70	102	12	73	3	0.8947	0.9605	0.9211	0.8427
0.90	105	9	73	3	0.9211	0.9605	0.9368	0.8720

is still in the middle of the range, as in previous studies, with MCC = 0.8791 for CF = 0.55.

CSCA – Clonal Selection Classification System

The clonal selection classification system, developed by Brownlee, is formulated as a function optimization procedure that maximizes the number of patterns correctly classified and minimizes the number of patterns incorrectly classified [105]. Unlike the AIRS algorithm, in which the system is exposed only once to the set of antigens, CSCA is trained for several generations, and during each generation the entire set of antibodies is exposed to all antigens. The computational steps of the CSCA algorithm are shown in the following procedure:

(1) Initialization. The CSCA algorithm starts by generating a set of N antibodies.
(2) Training. Repeat the training of all antibodies for G generations (iterations).
(2.1) Selection and pruning. The entire group of antibodies is exposed to the antigen set and a fitness score is computed for each antibody. Then all antibodies are selected and the following three evaluation rules are applied to each antibody:
(2.1.1) Remove from the selected set all antibodies with a misclassification score of zero.
(2.1.2) Antibodies that have zero correct classification and misclassification higher than zero are reassigned to the class of the majority. Fitness is recalculated.
(2.1.3) Remove from the selected set and from the base antibody population all antibodies with a fitness scoring lower than a threshold.
(2.2) Cloning and mutation. The selected set of antibodies is cloned and mutated.
(2.3) Insert new antibodies. Insert the clones generated into the main antibody population. A number of n randomly selected antigens from the antigen set are inserted into the main antibody population, where n is the number of antibodies selected in step (2.1).
(3) Final pruning. The antibody population is exposed to the entire antigen population, fitness scores are computed for each antibody, and pruning of antibodies is performed as described in step (2.1.3).
(4) Select classifier. The final antibody population represents the CSCA classifier. To classify a new pattern, the classification antibodies are exposed to the pattern, then the k most similar (highest affinity) antibodies are selected and a majority vote assigns the class of the pattern.

The artificial immune system CSCA was applied in several virtual screening studies, namely identification of estrogen receptor ligands, recognition of dihydrofolate reductase inhibitors, classification of angiotensin converting enzyme inhibitors, detection of benzodiazepine receptor ligands, and SAR for thermolysin inhibitors. To demonstrate the influence of the user defined parameters on the CSCA predictions, we present the influence of the clonal scale factor CSF, tested in L20%O cross-validation. CSF is used to increase or decrease the number of clones generated for each antibody, and has a default value of one. Low values for CSF promote a low diversity of solutions, whereas high CSF values increase the diversity of the recognition cells.

CSCA was applied for the classification of 232 chemical compounds into estrogen receptor (ER) ligands (131 chemicals, class +1) and compounds that do not bind to the estrogen receptor (101 chemicals, class −1) [129]. The chemical structure was represented with 312 topological indices computed with Molconn-Z. The clonal scale factor was modified between 0.1 and 4 (Table 10a), with the best predictions obtained for CSF = 2 (Ac = 0.6207 and MCC = 0.2057), but with no clear trend apparent for the values that give the best predictions. For example, the next best predictions are obtained for CSF = 0.1 (MCC = 0.1935), whereas the lowest predictions are obtained with CSF = 0.7 (MCC = 0.0416). To investigate the influence of feature selection on the classification abilities of CSCA, 29 structural descriptors were selected with SubsetEvaluation and BestFirst, which results in slightly better predictions for a much lower value of CSF (CSF = 0.1, Ac = 0.6336, MCC = 0.2508; Table 10b).

Dihydrofolate reductase (DHFR) inhibitors may be efficiently identified with CSCA, as was demonstrated for a dataset of 397 chemicals (198 compounds in class +1, pIC_{50} between 9.81 and 6.08; 199 compounds in class −1, pIC_{50} between 6.06 and 3.30) [130]. CSCA classifiers computed with 70 structural descriptors are used to eval-

Drug Design with Artificial Intelligence Methods, Table 10
CSCA prediction statistics for ER SAR models computed for various values of the clonal scale factor CSF

CSF	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
(a) 312 structural descriptors								
0.1	98	33	44	57	0.7481	0.4356	0.6121	0.1935
0.3	92	39	35	66	0.7023	0.3465	0.5474	0.0519
0.5	97	34	37	64	0.7405	0.3663	0.5776	0.1149
0.7	92	39	34	67	0.7023	0.3366	0.5431	0.0416
1.0	108	23	29	72	0.8244	0.2871	0.5905	0.1326
2.0	112	19	32	69	0.8550	0.3168	0.6207	0.2057
3.0	99	32	31	70	0.7557	0.3069	0.5603	0.0698
4.0	106	25	30	71	0.8092	0.2970	0.5862	0.1238
(b) 29 structural descriptors								
0.1	91	40	56	45	0.6947	0.5545	0.6336	0.2508
0.3	99	32	45	56	0.7557	0.4455	0.6207	0.2119
0.5	97	34	42	59	0.7405	0.4158	0.5991	0.1651
0.7	94	37	47	54	0.7176	0.4653	0.6078	0.1887
1.0	98	33	47	54	0.7481	0.4653	0.6250	0.2226
2.0	98	33	38	63	0.7481	0.3762	0.5862	0.1338
3.0	96	35	41	60	0.7328	0.4059	0.5905	0.1466
4.0	98	33	42	59	0.7481	0.4158	0.6034	0.1738

Drug Design with Artificial Intelligence Methods, Table 11
CSCA prediction statistics for DHFR SAR models computed for various values of the clonal scale factor CSF

CSF	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
(a) 70 structural descriptors								
0.1	130	68	92	107	0.6566	0.4623	0.5592	0.1212
0.2	138	60	98	101	0.6970	0.4925	0.5945	0.1935
0.5	129	69	92	107	0.6515	0.4623	0.5567	0.1159
0.7	130	68	87	112	0.6566	0.4372	0.5466	0.0961
1.0	119	79	98	101	0.6010	0.4925	0.5466	0.0940
2.0	142	56	87	112	0.7172	0.4372	0.5768	0.1608
3.0	115	83	99	100	0.5808	0.4975	0.5390	0.0786
4.0	132	66	89	110	0.6667	0.4472	0.5567	0.1167
(b) 5 structural descriptors								
0.1	166	32	144	55	0.8384	0.7236	0.7809	0.5656
0.2	155	43	142	57	0.7828	0.7136	0.7481	0.4975
0.5	151	47	150	49	0.7626	0.7538	0.7582	0.5164
0.7	149	49	148	51	0.7525	0.7437	0.7481	0.4963
1.0	152	46	151	48	0.7677	0.7588	0.7632	0.5265
2.0	154	44	150	49	0.7778	0.7538	0.7657	0.5317
3.0	158	40	153	46	0.7980	0.7688	0.7834	0.5670
4.0	148	50	151	48	0.7475	0.7588	0.7531	0.5063

uate the effect of the clonal scale factor on the prediction accuracy. Based on the structure of the CSCA algorithm, it should be expected that higher CSF values are useful in identifying better solutions, because more clones are generated, and the system explores a wider diversity of solutions. However, for dihydrofolate reductase inhibitors, the highest predictions are obtained for CSF = 0.2 (Ac = 0.5945 and MCC = 0.1935; Table 11a). Also, for high CSF values, between 0.7 and 3, MCC decreases markedly. A dramatic increase of the CSCA model quality is obtained with a feature selection that reduces the set of structural descriptors to 5 (Table 11b). The best predictions are obtained for a much higher CSF value, namely CSF = 3, with Ac = 0.7834 and MCC = 0.5670. Further tests should be performed with other SAR datasets in order to find the optimum CSF values for various drug screening experiments.

Another set of experiments with CSCA involved the classification of 114 angiotensin converting enzyme (ACE) inhibitors (57 compounds in class +1, pIC_{50} between 9.94 and 6.41; 57 compounds in class −1, pIC_{50} between 6.37 and 2.14) [131]. The chemical structure was encoded with 56 structural descriptors, and the CSF influence was evaluated for 16 values between 0.1 and 4. For all but one CSF values the CSCA classifiers give the same prediction indices, with Ac = 0.8684 and MCC = 0.7510. The CSCA insensitivity to the CSF variation is unexpected, and more experiments are necessary to fully understand this behavior. A feature selection step decreases the pool of structural descriptors to 12 (Table 12), with a slight decrease in the prediction statistics (CSF = 0.9, Ac = 0.8509, MCC = 0.7027). Usually, feature selection provides a smaller set of structural descriptors that increase the predictions of artificial immune systems. The exception encountered for ACE inhibitors should be further investigated to identify possible explanations and better feature selection procedures.

Drug Design with Artificial Intelligence Methods, Table 12
CSCA prediction statistics for ACE SAR models with 12 structural descriptors computed for various values of the clonal scale factor CSF

CSF	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
0.1	45	12	50	7	0.7895	0.8772	0.8333	0.6692
0.3	43	14	49	8	0.7544	0.8596	0.8070	0.6175
0.5	47	10	48	9	0.8246	0.8421	0.8333	0.6668
0.7	44	13	49	8	0.7719	0.8596	0.8158	0.6340
0.9	47	10	50	7	0.8246	0.8772	0.8509	0.7027
2.0	46	11	43	14	0.8070	0.7544	0.7807	0.5622
3.0	46	11	49	8	0.8070	0.8596	0.8333	0.6676
4.0	46	11	48	9	0.8070	0.8421	0.8246	0.6495

The CSCA immune system was evaluated for the discrimination of 163 benzodiazepine receptor (BZR) ligands (82 high affinity ligands and 81 low affinity ligands) [123]. Starting from a set of 75 structural descriptors, CSF was modified between 0.1 and 4 (Table 13a), with

Drug Design with Artificial Intelligence Methods, Table 13
CSCA prediction statistics for BZR SAR models computed for various values of the clonal scale factor CSF

CSF	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
(a) 75 structural descriptors								
0.1	55	27	57	24	0.6707	0.7037	0.6871	0.3746
0.3	52	30	58	23	0.6341	0.7160	0.6748	0.3513
0.5	52	30	58	23	0.6341	0.7160	0.6748	0.3513
0.7	57	25	57	24	0.6951	0.7037	0.6994	0.3988
1.0	48	34	65	16	0.5854	0.8025	0.6933	0.3971
2.0	48	34	59	22	0.5854	0.7284	0.6564	0.3169
3.0	54	28	54	27	0.6585	0.6667	0.6626	0.3252
4.0	58	24	52	29	0.7073	0.6420	0.6748	0.3501
(b) 16 structural descriptors								
0.1	56	26	55	26	0.6829	0.6790	0.6810	0.3619
0.3	62	20	51	30	0.7561	0.6296	0.6933	0.3890
0.5	53	29	57	24	0.6463	0.7037	0.6748	0.3506
0.7	60	22	54	27	0.7317	0.6667	0.6994	0.3993
1.0	61	21	45	36	0.7439	0.5556	0.6503	0.3050
2.0	65	17	50	31	0.7927	0.6173	0.7055	0.4166
3.0	58	24	55	26	0.7073	0.6790	0.6933	0.3865
4.0	58	24	51	30	0.7073	0.6296	0.6687	0.3380

Drug Design with Artificial Intelligence Methods, Table 14
CSCA prediction statistics for THER SAR models computed for various values of the clonal scale factor CSF

CSF	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
(a) 64 structural descriptors								
0.1	24	14	26	12	0.6316	0.6842	0.6579	0.3162
0.5	24	14	26	12	0.6316	0.6842	0.6579	0.3162
1.0	24	14	26	12	0.6316	0.6842	0.6579	0.3162
2.0	24	14	26	12	0.6316	0.6842	0.6579	0.3162
2.5	19	19	27	11	0.5000	0.7105	0.6053	0.2154
3.0	24	14	26	12	0.6316	0.6842	0.6579	0.3162
3.5	25	13	26	12	0.6579	0.6842	0.6711	0.3422
4.0	24	14	26	12	0.6316	0.6842	0.6579	0.3162
(b) 10 structural descriptors								
0.1	20	18	27	11	0.5263	0.7105	0.6184	0.2410
0.5	19	19	31	7	0.5000	0.8158	0.6579	0.3328
1.0	20	18	24	14	0.5263	0.6316	0.5789	0.1588
2.0	21	17	32	6	0.5526	0.8421	0.6974	0.4124
2.5	19	19	27	11	0.5000	0.7105	0.6053	0.2154
3.0	21	17	31	7	0.5526	0.8158	0.6842	0.3819
3.5	13	25	32	6	0.3421	0.8421	0.5921	0.2127
4.0	16	22	32	6	0.4211	0.8421	0.6316	0.2901

the best results obtained for CSF = 0.7 (Ac = 0.6994 and MCC = 0.3988). A small improvement of the CSCA predictions is obtained by reducing the pool of descriptors to 16 by feature selection (Table 13b). Although the model improvement is not big (Ac = 0.7055 and MCC = 0.4166 for CSF = 2), feature selection is still important because the CSCA model can be computed faster, and the selected descriptors may suggest which molecular features influence the biological activity.

CSCA was also tested for a dataset of 76 thermolysin (THER) inhibitors (38 compounds in class +1, pK_i between 10.17 and 5.55; 38 compounds in class −1, pK_i between 5.16 and 0.52) and 64 structural descriptors [132]. For 14 out of 16 CSF values tested in this experiment, the CSCA classifiers have identical predictions, with Ac = 0.6711 and MCC = 0.3162. The best predictions are obtained for CSF = 3.5, with slightly higher prediction statistics, namely MCC = 0.3422 (Table 14a). Feature selection reduces the number of descriptors to 10, which results in a minor improvement (CSF = 2, Ac = 0.6974, MCC = 0.4124, Table 14b).

IMMUNOS

Carter developed the IMMUNOS-81 artificial immune systems as an instance based classifier with some similarity to *k*-nearest neighbor classifiers [106]. Brownlee extended this algorithm by adding elements from other AIS classifiers, such as cloning and hypermutation, to obtain IMMUNOS-99 [107]. A brief description of the IMMUNOS-99 consists of the following steps:

(1) Initialization. The training group of antigens is divided into groups based on class label.

(2) Train B-cell groups. The final IMMUNOS classifier consists of a B-cell population for each class represented in the training set of antigens. Each B-cell population is generated and trained independent of the other B-cell populations. Steps (2.1) and (2.2) are repeated *C* times, where *C* is the number of antigen classes.

(2.1) Create B-cell population. Generate a B-cell population for the antigen class under training. A fraction of the antigen population from that class is used as seed for the B-cell population.

(2.2) Training. Train the B-cell class for *G* generations (iterations).

(2.2.1) Expose population. The B-cell population is exposed to all antigens from all classes, and an affinity value is computed for each B-cell/antigen comparison. A rank-based scoring is established for each B-cell.

(2.2.2) Compute fitness. A fitness index is computed for each B-cell, based on the rank scores for antigens in the same class and the rank scores for antigens in all other classes. B-cells that recognize better the antigens from the same class have fitness score higher than one,

whereas B-cells that recognize better the antigens from other classes have fitness score lower than one.

(2.2.3) Pruning. A user-defined parameter, between $[0, 1]$, sets the minimum fitness score of a B-cell. All B-cells with fitness scores lower than this threshold are removed from the population.

(2.2.4) Affinity maturation. After pruning, the B-cell population contains only cells that can identify antigens from the same class. To improve the B-cell recognition ability, the system undergoes an affinity maturation process based on cloning and hypermutation.

(2.2.4.1) Order population. The B-cell population is ordered in the descending order of the fitness scores.

(2.2.4.2) Generate clones. Each B-cell is cloned proportional to its fitness rank. The rank ratio for a B-cell is:

$$r_i = \frac{\text{rank}}{S}$$

where r_i is the rank ration of the ith B-cell, rank is the actual index of the B-cell in the ordered sequence, rank $\in [1, S]$, and S is the total number of B-cells in the population (class). The number of clones generated for each B-cell is:

$$NC_i = \left\lfloor \frac{r_i}{\sum_{j=1}^{S} r_j} N + 0.5 \right\rfloor$$

where N is the total number of antigens in the same class.

(2.2.4.3) Mutate clones. The clones are mutated by the inverse of the B-cell rank ratios. As a result of this procedure, clones of B-cells with higher ranks undergo small mutations, whereas clones of B-cells with lower ranks go through large mutations. All clones generated are added to the B-cell population.

(2.2.5) Insert random antigens. In order to increase the diversity of the B-cell population, a random selection of antigens from the same class is added to the B-cell pool. The number of antigens added is equal to the number of B-cells deleted during the pruning process from step (2.2.3). The diversity introduced by the antigen-based B-cells is particularly useful whenever the affinity maturation process converges to a limited number of B-cells.

(3) Final pruning. This step removes B-cells with low fitness after the system finishes the training for each antigen class and for the set number of generations G.

(3.1) Compute fitness. Each B-cell population (class) is exposed to all antigens, one antigen at a time, and only the best matching B-cells receive a score.

(3.2) Pruning. Similarly with the pruning process from step (2.2.3), all B-cells with low fitness scores lower are removed from the population.

(4) Select classifier. The populations of B-cells that survive the final pruning represent the classifier for new, unknown antigens. During the classification process, each B-cell class is exposed to the unknown antigen, and an avidity index is computed. Then the B-cell populations compete for the unknown antigen that takes the class label of the B-cell population with the highest avidity index.

The IMMUNOS-99 system was evaluated in several drug design studies, namely structure-activity relationships for acetylcholinesterase inhibitors, virtual screening of cyclooxygenase-2 inhibitors, recognition of benzodiazepine receptor ligands, and classification of thrombin inhibitors. All examples presented here investigate the influence of the seed population percentage SPP. SPP is a user defined parameter that specifies the percentage of the antigen population from each class that is used as seed for the B-cell population. If SPP = 100% then the initial B-cell population is identical with the antigen population in the same class. The influence of the SPP parameter was investigated in series of L20%O cross-validation experiments. For each drug design dataset, the IMMUNOS-99 classifier was trained for 19 values of the SPP parameter, between 0.05 and 0.95.

IMMUNOS-99 structure-activity relationships were developed for a dataset of 111 acetylcholinesterase (AChE) inhibitors characterized by 63 structural descriptors [133]. The classifiers were trained to discriminate between 55 inhibitors in class +1 (pIC_{50} between 9.52 and 6.87) and 56 inhibitors in class −1 (pIC_{50} between 6.84 and 4.27). The prediction MCC increases from 0.1349 for SPP = 0.05, has a maximum of 0.2847 for SPP = 0.35, and then decreases to 0.2110 for SPP = 0.95 (Table 15a). These results suggest that seeding the B-cell population with less than half of the antigen population improves the prediction statistics. The number of structural descriptors is reduced to 9 by feature selection, which results in a slight decrease in the IMMUNOS-99 predictions (Table 15b).

The virtual screening of cyclooxygenase-2 (COX2) inhibitors may be efficiently done with IMMUNOS-99, as shown for 322 compounds (162 compounds in class +1, pIC_{50} between 9 and 6.60; 160 compounds in class −1, pIC_{50} between 6.59 and 4) [134]. Starting from a set of 74 structural descriptors, several IMMUNOS-99 classifiers were developed to study the influence of the SPP parameter (Table 16a). The results obtained from this series of experiments indicate that the prediction statistics

Drug Design with Artificial Intelligence Methods, Table 15
IMMUNOS-99 prediction statistics for AChE SAR models computed for various values of the seed population percentage SPP

SPP	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
(a) 63 structural descriptors								
0.05	30	25	33	23	0.5455	0.5893	0.5676	0.1349
0.10	45	10	22	34	0.8182	0.3929	0.6036	0.2329
0.15	44	11	18	38	0.8000	0.3214	0.5586	0.1382
0.35	45	10	25	31	0.8182	0.4464	0.6306	0.2847
0.50	45	10	22	34	0.8182	0.3929	0.6036	0.2329
0.70	43	12	23	33	0.7818	0.4107	0.5946	0.2072
0.85	44	11	23	33	0.8000	0.4107	0.6036	0.2286
0.95	44	11	23	33	0.8000	0.4107	0.6036	0.2286
(b) 9 structural descriptors								
0.05	35	20	21	35	0.6364	0.3750	0.5045	0.0118
0.10	41	14	18	38	0.7455	0.3214	0.5315	0.0738
0.15	47	8	15	41	0.8545	0.2679	0.5586	0.1510
0.35	48	7	6	50	0.8727	0.1071	0.4865	-0.0313
0.50	53	2	3	53	0.9636	0.0536	0.5045	0.0415
0.70	55	0	3	53	1.0000	0.0536	0.5225	0.1652
0.85	54	1	3	53	0.9818	0.0536	0.5135	0.0949
0.95	54	1	2	54	0.9818	0.0357	0.5045	0.0541

Drug Design with Artificial Intelligence Methods, Table 16
IMMUNOS-99 prediction statistics for COX2 SAR models computed for various values of the seed population percentage SPP

SPP	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
(a) 74 structural descriptors								
0.05	113	49	72	88	0.6975	0.4500	0.5745	0.1523
0.15	104	58	75	85	0.6420	0.4688	0.5559	0.1124
0.30	93	69	91	69	0.5741	0.5687	0.5714	0.1428
0.45	90	72	96	64	0.5556	0.6000	0.5776	0.1557
0.60	93	69	93	67	0.5741	0.5813	0.5776	0.1553
0.75	90	72	94	66	0.5556	0.5875	0.5714	0.1431
0.85	93	69	89	71	0.5741	0.5563	0.5652	0.1303
0.95	93	69	90	70	0.5741	0.5625	0.5683	0.1366
(b) 12 structural descriptors								
0.05	160	2	25	135	0.9877	0.1562	0.5745	0.2596
0.15	159	3	34	126	0.9815	0.2125	0.5994	0.3041
0.30	158	4	43	117	0.9753	0.2687	0.6242	0.3456
0.45	159	3	41	119	0.9815	0.2562	0.6211	0.3461
0.60	159	3	46	114	0.9815	0.2875	0.6366	0.3744
0.75	159	3	48	112	0.9815	0.3000	0.6429	0.3855
0.85	157	5	50	110	0.9691	0.3125	0.6429	0.3742
0.95	158	4	50	110	0.9753	0.3125	0.6460	0.3852

have similar values for a wide range of the SPP parameter, with a small improvement for SPP = 0.45. The number of structural descriptors was reduced by feature selection to 12 important descriptors, thus improving the predictions of the IMMUNOS-99 classifiers (Table 16b). The best results are obtained for SPP = 0.75 (Ac = 0.6429, MCC = 0.3855), but FP is still too large, i. e., too many inactive compounds are predicted as active.

The IMMUNOS-99 immune system was also tested for the dataset of benzodiazepine receptor (BZR) ligands (82 high affinity ligands and 81 low affinity ligands) [123]. The best predictions for the entire pool of 75 structural descriptors were obtained for SPP = 0.75 (Ac = 0.6564, MCC 0.3506; Table 17a). To evaluate the importance of feature selection, the number of structural descriptors was reduced to 16 and the entire analysis was repeated for the full range of SPP values. Although FN decreases (active compounds predicted inactive), FP increases which results in slightly worse predictions (Table 17b). The best predictions are obtained also for SPP = 0.15 (Ac = 0.6503, MCC = 0.3006), but the results suggest that IMMUNOS-99 predictions do not improve with feature selection.

Drug Design with Artificial Intelligence Methods, Table 17
IMMUNOS-99 prediction statistics for BZR SAR models computed for various values of the seed population percentage SPP

SPP	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
(a) 75 structural descriptors								
0.05	48	34	47	34	0.5854	0.5802	0.5828	0.1656
0.15	35	47	68	13	0.4268	0.8395	0.6319	0.2922
0.25	39	43	67	14	0.4756	0.8272	0.6503	0.3232
0.35	35	47	70	11	0.4268	0.8642	0.6442	0.3233
0.50	36	46	69	12	0.4390	0.8519	0.6442	0.3191
0.65	34	48	69	12	0.4146	0.8519	0.6319	0.2960
0.75	36	46	71	10	0.4390	0.8765	0.6564	0.3506
0.95	36	46	70	11	0.4390	0.8642	0.6503	0.3347
(b) 16 structural descriptors								
0.05	57	25	39	42	0.6951	0.4815	0.5890	0.1808
0.15	55	27	51	30	0.6707	0.6296	0.6503	0.3006
0.25	54	28	51	30	0.6585	0.6296	0.6442	0.2883
0.35	55	27	49	32	0.6707	0.6049	0.6380	0.2763
0.50	53	29	49	32	0.6463	0.6049	0.6258	0.2515
0.65	51	31	51	30	0.6220	0.6296	0.6258	0.2516
0.75	50	32	51	30	0.6098	0.6296	0.6196	0.2394
0.95	50	32	51	30	0.6098	0.6296	0.6196	0.2394

The classification of thrombin (THR) inhibitors with IMMUNOS-99 was investigated for 88 chemicals (44 compounds in class +1, pK_i between 8.48 and 6.70; 44 compounds in class −1, pK_i between 6.68 and 4.36) and 66 structural descriptors [135]. The prediction statistics indicate that the IMMUNOS-99 is not very successful in discriminating thrombin inhibitors from non-inhibitors (Table 18a). In all 19 experiments that explore the influence of the SPP parameter, almost all chemical compounds are

Drug Design with Artificial Intelligence Methods, Table 18
IMMUNOS-99 prediction statistics for THR SAR models computed for various values of the seed population percentage SPP

SPP	TP_p	FN_p	TN_p	FP_p	Se_p	Sp_p	Ac_p	MCC_p
(a) 66 structural descriptors								
0.05	38	6	10	34	0.8636	0.2273	0.5455	0.1179
0.15	43	1	6	38	0.9773	0.1364	0.5568	0.2100
0.25	44	0	2	42	1.0000	0.0455	0.5227	0.1525
0.40	44	0	2	42	1.0000	0.0455	0.5227	0.1525
0.50	43	1	1	43	0.9773	0.0227	0.5000	0.0000
0.65	43	1	2	42	0.9773	0.0455	0.5114	0.0626
0.80	44	0	1	43	1.0000	0.0227	0.5114	0.1072
0.95	43	1	2	42	0.9773	0.0455	0.5114	0.0626
(b) 7 structural descriptors								
0.05	41	3	7	37	0.9318	0.1591	0.5455	0.1432
0.15	44	0	6	38	1.0000	0.1364	0.5682	0.2705
0.25	44	0	3	41	1.0000	0.0682	0.5341	0.1879
0.40	44	0	5	39	1.0000	0.1136	0.5568	0.2454
0.50	44	0	5	39	1.0000	0.1136	0.5568	0.2454
0.65	44	0	5	39	1.0000	0.1136	0.5568	0.2454
0.80	44	0	4	40	1.0000	0.0909	0.5455	0.2182
0.95	44	0	3	41	1.0000	0.0682	0.5341	0.1879

predicted in the class +1 (inhibitors). As a result, FN is small (which is good) but FP is very large (which is bad), and the overall statistics are low. A maximum is identified for SPP = 0.15 (Ac = 0.5568, MCC = 0.2100). Feature selection reduces the pool of descriptors to 7, and results in slightly better models (Table 18b). FP is still too large for the whole range of SPP values, which explains the low values for the statistical indices. Compared with the other three artificial immune systems, IMMUNOS-99 seems to be the most difficult to tune in order to obtain good predictions. Feature selection has no or small effect in improving IMMUNOS-99 models, which suggests that other algorithms should be investigated to reduce the pool of structural descriptors.

Future Directions

Pharmaceutical drug discovery uses computer-assisted molecular design to increase the chances of bringing a drug on the market, and to lower the research and development costs. Computational models are used to simulate the physical, chemical, biological, and toxicological properties of drug candidates, thus replacing expensive and time-consuming large scale experiments. The entire process consists of iterative steps, in which experimental results are used to train computational models, which in turn suggest novel molecules that are synthesized and tested in the laboratory. We reviewed here the most important artificial intelligence algorithms used in drug design, namely genetic algorithms, ant colony optimization, particle swarm optimization, and artificial immune systems. The main advantage of artificial intelligence algorithms is their ability to explore search spaces of high dimensionality, and to identify the global optimum for complex and difficult problems. Genetic algorithms have a long history of applications in QSAR and drug design, and their operation is thoroughly explored. The other artificial intelligence algorithms were adopted only recently, but they already demonstrated strong results that make them competitors for GA. More important, ACO, PSO and AIS bring new simulation capabilities, thus complementing GA. A promising direction of development is a combined use of these artificial intelligence algorithms that could provide better predictions of molecular properties. Another source of improvement might come from the integration of the molecular graph into the artificial intelligence algorithms, which would complement (or even substitute) the use of structural descriptors.

Bibliography

1. Holland J (1975) Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor
2. Goldberg DE (1989) Genetic Algorithms in Search, Optimization & Machine Learning. Addison Wesley, Reading
3. Jones G (1998) Genetic and evolutionary algorithms. In: Schleyer PvR, Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer III HF, Schreiner PR (eds) The Encyclopedia of Computational Chemistry. Wiley, Chichester, pp 1127–1136
4. Terfloth L, Gasteiger J (2001) Neural networks and genetic algorithms in drug design. Drug Discov Today 6:S102–S108
5. von Homeyer A (2003) Evolutionary algorithms and their applications in chemistry. In: Gasteiger J (ed) Handbook of Chemoinformatics, vol 3. Wiley-VCH, Weinheim, pp 1239–1280
6. Dorigo M, Maniezzo V, Colorni A (1996) Ant system: Optimization by a colony of cooperating agents. IEEE Trans Syst Man Cybern Part B Cybern 26:29–41
7. Dorigo M, Gambardella LM (1997) Ant colony system: A cooperative learning approach to the traveling salesman problem. IEEE Trans Evol Comput 1:53–66
8. Dorigo M, Di Caro G, Gambardella LM (1999) Ant algorithms for discrete optimization. Artif Life 5:137–172
9. Dorigo M, Stützle T (2004) Ant Colony Optimization. MIT Press, Cambridge
10. Dorigo M, Blum C (2005) Ant colony optimization theory: A survey. Theor Comput Sci 344:243–278
11. Kennedy J, Eberhart R (1995) Particle swarm optimization. Proceedings of IEEE International Conference on Neural Networks, vol 4. pp 1942–1948
12. Banks A, Vincent J, Anyakoha C (2007) A review of particle swarm optimization Part I: background and development. Nat Comput 6:467–484

13. Banks A, Vincent J, Anyakoha C (2008) A review of particle swarm optimization Part II: hybridisation, combinatorial, multicriteria and constrained optimization, and indicative applications. Nat Comput 7:109–124
14. Chuang LY, Chang HW, Tu CJ, Yang CH (2008) Improved binary PSO for feature selection using gene expression data. Comput Biol Chem 32:29–38
15. Namasivayam V, Günther R (2007) PSO@AUTODOCK: A fast flexible molecular docking program based on swarm intelligence. Chem Biol Drug Des 70:475–484
16. Agrafiotis DK, Cedeño W (2002) Feature selection for structure-activity correlation using binary particle swarms. J Med Chem 45:1098–1107
17. Huang J, Ma G, Muhammad I, Cheng Y (2007) Identifying P-glycoprotein substrates using a support vector machine optimized by a particle swarm. J Chem Inf Model 47:1638–1647
18. Shen Q, Shi WM, Yang XP, Ye BX (2006) Particle swarm algorithm trained neural network for QSAR studies of inhibitors of platelet-derived growth factor receptor phosphorylation. Eur J Pharm Sci 28:369–376
19. Hunt JE, Cooke DE (1996) Learning using an artificial immune system. J Netw Comput Appl 19:189–212
20. de Castro LN, Von Zuben FJ (1999) Artificial immune systems: Part I Basic theory and applications. FEEC/UNICAMP, Brazil
21. de Castro LN, Von Zuben FJ (2000) Artificial immune systems: Part II A survey of applications. FEEC/UNICAMP, Brazil
22. Timmis J, Neal M, Hunt J (2000) An artificial immune system for data analysis. Biosystems 55:143–150
23. Chao DL, Forrest S (2003) Information immune systems. Genet Programm Evolv Mach 4:311–331
24. de Castro LN, Timmis JI (2003) Artificial immune systems as a novel soft computing paradigm. Soft Comput 7:526–544
25. Musilek P, Lau A, Reformat M, Wyard-Scott L (2006) Immune programming. Inf Sci 176:972–1002
26. Timmis J (2007) Artificial immune systems – Today and tomorrow. Nat Comput 6:1–18
27. Forrest S, Beauchemin C (2007) Computer immunology. Immunol Rev 216:176–197
28. Dasgupta D (1999) Artificial Immune Systems and Their Applications. Springer, Berlin
29. de Castro LN, Timmis J (2002) Artificial Immune Systems: A New Computational Intelligence Approach. Springer, Berlin
30. Tarakanov AO, Skormin VA, Sokolova SP (2003) Immunocomputing: Principles and Applications. Springer, Berlin
31. Ishida Y (2004) Immunity-Based Systems. Springer, Berlin
32. Timmis J, Bentley P, Hart E (2003) Artificial Immune Systems: Second International Conference, ICARIS 2003, Edinburgh, September 1–3. Lecture Notes in Computer Science, vol 2787. Springer, Berlin
33. Nicosia G, Cutello V, Bentley PJ, Timmis JI (2004) Artificial Immune Systems: Third International Conference, ICARIS 2004, Catania, September 13–16. Lecture Notes in Computer Science, vol 3239. Springer, Berlin
34. Jacob C, Pilat ML, Bentley PJ, Timmis J (2005) Artificial Immune Systems: 4th International Conference, ICARIS 2005, Banff, August 14–17. Lecture Notes in Computer Science, vol 3627. Springer, Berlin
35. Bersini H, Carneiro J (2006) Artificial Immune Systems: 5th International Conference, ICARIS 2006, Oeiras, September 4–6. Lecture Notes in Computer Science, vol 4163. Springer, Berlin
36. Ando S, Iba H (2004) Classification of gene expression profile using combinatory method of evolutionary computation and machine learning. Genet Programm Evolv Mach 5:145–156
37. Bezerra GB, Cançado GMA, Menossi M, de Castro LN, Von Zuben FJ (2005) Recent advances in gene expression data clustering: A case study with comparative results. Genet Mol Res 4:514–524
38. Tsankova D, Georgieva V, Kasabov N (2005) Artificial immune networks as a paradigm for classification and profiling of gene expression data. J Comput Theor Nanosci 2:543–550
39. Şahan S, Polat K, Kodaz H, Güneş S (2007) A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. Comput Biol Med 37:415–423
40. Polat K, Güneş S (2008) Computer aided medical diagnosis system based on principal component analysis and artificial immune recognition system classifier algorithm. Expert Syst Appl 34:773–779
41. Polat K, Şahan S, Güneş S (2006) A new method to medical diagnosis: Artificial immune recognition system (AIRS) with fuzzy weighted pre-processing and application to ECG arrhythmia. Expert Syst Appl 31:264–269
42. Latifoglu F, Şahan S, Kara S, Güneş S (2007) Diagnosis of atherosclerosis from carotid artery Doppler signals as a real-world medical application of artificial immune systems. Expert Syst Appl 33:786–793
43. Cutello V, Nicosia G, Pavone M, Timmis J (2007) An immune algorithm for protein structure prediction on lattice models. IEEE Trans Evol Comput 11:101–117
44. Anile AM, Cutello V, Narzisi G, Nicosia G, Spinella S (2007) Determination of protein structure and dynamics combining immune algorithms and pattern search methods. Nat Comput 6:55–72
45. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 19:1639–1662
46. Wang R, Gao Y, Lai LH (2000) LigBuilder: A multi-purpose program for structure-based drug design. J Mol Model 6:498–516
47. So SS, Karplus M (1996) Evolutionary optimization in quantitative structure-activity relationship: An application of genetic neural networks. J Med Chem 39:1521–1530
48. Venkatasubramanian V, Chan K, Caruthers JM (1995) Evolutionary design of molecules with desired properties using the genetic algorithm. J Chem Inf Comput Sci 35:188–195
49. Sundaram A, Venkatasubramanian V (1998) Parametric sensitivity and search-space characterization studies of genetic algorithms for computer-aided polymer design. J Chem Inf Comput Sci 38:1177–1191
50. Gillet VJ, Willett P, Bradshaw J, Green DVS (1999) Selecting combinatorial libraries to optimize diversity and physical properties. J Chem Inf Comput Sci 39:169–177
51. Ivanciuc O, Ivanciuc T, Cabrol-Bass D (2002) QSAR for dihydrofolate reductase inhibitors with molecular graph structural descriptors. J Mol Struct (Theochem) 582:39–51
52. Wegner JK, Fröhlich H, Zell A (2004) Feature selection for descriptor based classification models, 2. Human intestinal absorption (HIA). J Chem Inf Comput Sci 44:931–939
53. Weber L (1998) Evolutionary combinatorial chemistry: application of genetic algorithms. Drug Discov Today 3:379–385
54. Weber L (2005) Current status of virtual combinatorial library design. QSAR Comb Sci 24:809–823

55. Gallop MA, Barrett RW, Dower WJ, Fodor SPA, Gordon EM (1994) Applications of combinatorial technologies to drug discovery, 1. Background and peptide combinatorial libraries. J Med Chem 37:1233–1251
56. Gordon EM, Barrett RW, Dower WJ, Fodor SPA, Gallop MA (1994) Applications of combinatorial technologies to drug discovery, 2. Combinatorial organic-synthesis, library screening strategies, and future directions. J Med Chem 37:1385–1401
57. Weber L (1998) Applications of genetic algorithms in molecular diversity. Curr Opin Chem Biol 2:381–385
58. Illgen K, Enderle T, Broger C, Weber L (2000) Simulated molecular evolution in a full combinatorial library. Chem Biol 7:433–441
59. Ugi I, Almstetter M, Bock H, Dömling A, Ebert B, Gruber B, Hanusch-Kompa C, Heck S, Kehagia-Drikos K, Lorenz K, Papathoma S, Raditschnig R, Schmid T, Werner B, von Zychlinski A (1998) MCR XVII. Three types of MCRs and the libraries – Their chemistry of natural events and preparative chemistry. Croat Chem Acta 71:527–547
60. Weber L (2002) Multi-component reactions and evolutionary chemistry. Drug Discov Today 7:143–147
61. Globus A, Lawtonb J, Wipke T (1999) Automatic molecular design using evolutionary techniques. Nanotechnology 10:290–299
62. Brown N, McKay B, Gilardoni F, Gasteiger J (2004) A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. J Chem Inf Comput Sci 44:1079–1087
63. Brown N, McKay B, Gasteiger J (2006) A novel workflow for the inverse QSPR problem using multiobjective optimization. J Comput Aided Mol Des 20:333–341
64. Lewis RA (2005) A general method for exploiting QSAR models in lead optimization. J Med Chem 48:1638–1648
65. Izrailev S, Agrafiotis D (2001) A novel method for building regression tree models for QSAR based on artificial ant colony systems. J Chem Inf Comput Sci 41:176–180
66. Izrailev S, Agrafiotis DK (2002) Variable selection for QSAR by artificial ant colony systems. SAR QSAR Environ Res 13:417–423
67. Shelokar PS, Jayaraman VK, Kulkarni BD (2004) An ant colony approach for clustering. Anal Chim Acta 509:187–195
68. He Y, Chen D, Zhao W (2006) Ensemble classifier system based on ant colony algorithm and its application in chemical pattern classification. Chemom Intell Lab Syst 80:39–49
69. Korb O, Stützle T, Exner TE (2006) PLANTS: Application of ant colony optimization to structure-based drug design. Ant Colony Optimization and Swarm Intelligence. Proceedings. LNCS, vol 4150. Springer, Berlin, pp 247–258
70. Shen Q, Jiang JH, Tao JC, Shen GL, Yu RQ (2005) Modified ant colony optimization algorithm for variable selection in QSAR modeling: QSAR studies of cyclooxygenase inhibitors. J Chem Inf Model 45:1024–1029
71. Gunturi SB, Narayanan R, Khandelwal A (2006) In silico ADME modelling 2: Computational models to predict human serum albumin binding affinity using ant colony systems. Bioorg Med Chem 14:4118–4129
72. Shi WM, Shen Q, Kong W, Ye BX (2007) QSAR analysis of tyrosine kinase inhibitor using modified ant colony optimization and multiple linear regression. Eur J Med Chem 42:81–86
73. Karpenko O, Shi J, Dai Y (2005) Prediction of MHC class II binders using the ant colony search strategy. Artif Intell Med 35:147–156
74. Hernandez P, Gras R, Frey J, Appel RD (2003) Popitam: Towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. Proteomics 3:870–878
75. Shen Q, Shi WM, Kong W, Ye BX (2007) A combination of modified particle swarm optimization algorithm and support vector machine for gene selection and tumor classification. Talanta 71:1679–1683
76. Call ST, Zubarev DY, Boldyrev AI (2007) Global minimum structure searches via particle swarm optimization. J Comput Chem 28:1177–1186
77. Chang BCH, Ratnaweera A, Halgamuge SK, Watson HC (2004) Particle swarm optimisation for protein motif discovery. Genet Programm Evolv Mach 5:203–214
78. Chen K, Li T, Cao T (2006) Tribe-PSO: A novel global optimization algorithm and its application in molecular docking. Chemom Intell Lab Syst 82:248–259
79. Chen HM, Liu BF, Huang HL, Hwang SF, Ho SY (2007) SODOCK: Swarm optimization for highly flexible protein-ligand docking. J Comput Chem 28:612–623
80. Cedeño W, Agrafiotis DK (2003) Using particle swarms for the development of QSAR models based on K-nearest neighbor and kernel regression. J Comput Aided Mol Des 17:255–263
81. Lü JX, Shen Q, Jiang JH, Shen GL, Yu RQ (2004) QSAR analysis of cyclooxygenase inhibitor using particle swarm optimization and multiple linear regression. J Pharm Biomed Anal 35:679–687
82. Shen Q, Jiang JH, Jiao CX, Shen GL, Yu RQ (2004) Modified particle swarm optimization algorithm for variable selection in MLR and PLS modeling: QSAR studies of antagonism of angiotensin II antagonists. Eur J Pharm Sci 22:145–152
83. Lin WQ, Jiang JH, Shen Q, Shen GL, Yu RQ (2005) Optimized block-wise variable combination by particle swarm optimization for partial least squares modeling in quantitative structure-activity relationship studies. J Chem Inf Model 45:486–493
84. Lin L, Lin WQ, Jiang JH, Shen GL, Yu RQ (2005) QSAR analysis of substituted bis[(acridine-4-carboxamide)propyl] methylamines using optimized block-wise variable combination by particle swarm optimization for partial least squares modeling. Eur J Pharm Sci 25:245–254
85. Hu L, Wu H, Lin W, Jiang J, Yu R (2007) Quantitative structure-activity relationship studies for the binding affinities of imidazobenzodiazepines for the $\alpha 6$ benzodiazepine receptor isoform utilizing optimized blockwise variable combination by particle swarm optimization for partial least squares modeling. QSAR Comb Sci 26:92–101
86. Shen Q, Jiang JH, Jiao CX, Huan SY, Shen GL, Yu RQ (2004) Optimized partition of minimum spanning tree for piecewise modeling by particle swarm algorithm. QSAR studies of antagonism of angiotensin II antagonists. J Chem Inf Comput Sci 44:2027–2031
87. Lin WQ, Jiang JH, Shen Q, Wu HL, Shen GL, Yu RQ (2005) Piecewise hypersphere modeling by particle swarm optimization in QSAR studies of bioactivities of chemical compounds. J Chem Inf Model 45:535–541
88. Lin L, Lin WQ, Jiang JH, Zhou YP, Shen GL, Yu RQ (2005) QSAR analysis of a series of 2-aryl(heteroaryl)-2,5-dihydropyrazolo[4,3-c]quinolin-3-(3H)-ones using piecewise

hyper-sphere modeling by particle swarm optimization. Anal Chim Acta 552:42–49
89. Xu L, Jiang JH, Lin WQ, Zhou YP, Wu HL, Shen GL, Yu RQ (2007) Optimized sample-weighted partial least squares. Talanta 71:561–566
90. Shen Q, Jiang JH, Jiao CX, Lin WQ, Shen GL, Yu RQ (2004) Hybridized particle swarm algorithm for adaptive structure training of multilayer feed-forward neural network: QSAR studies of bioactivity of organic compounds. J Comput Chem 25:1726–1735
91. Zhou YP, Jiang JH, Lin WQ, Zou HY, Wu HL, Shen GL, Yu RQ (2006) Adaptive configuring of radial basis function network by hybrid particle swarm algorithm for QSAR studies of organic compounds. J Chem Inf Model 46:2494–2501
92. Zhou YP, Jiang JH, Lin WQ, Xu L, Wu HL, Shen GL, Yu RQ (2007) Artificial neural network-based transformation for nonlinear partial least-square regression with application to QSAR studies. Talanta 71:848–853
93. Meissner M, Schmuker M, Schneider G (2006) Optimized Particle Swarm Optimization (OPSO) and its application to artificial neural network training. BMC Bioinformatics 7:125
94. Ivanciuc O (2007) Applications of support vector machines in chemistry. In: Lipkowitz KB, Cundari TR (eds) Reviews in Computational Chemistry, vol 23. Wiley-VCH, Weinheim, pp 291–400
95. Lin WQ, Jiang JH, Zhou YP, Wu HL, Shen GL, Yu RQ (2007) Support vector machine based training of multilayer feedforward neural networks as optimized by particle swarm algorithm: Application in QSAR studies of bioactivity of organic compounds. J Comput Chem 28:519–527
96. Tang LJ, Zhou YP, Jiang JH, Zou HY, Wu HL, Shen GL, Yu RQ (2007) Radial basis function network-based transform for a nonlinear support vector machine as optimized by a particle swarm optimization algorithm with application to QSAR studies. J Chem Inf Model 47:1438–1445
97. de Castro LN (2004) Dynamics of an artificial immune network. J Exp Theor Artif Intell 16:19–39
98. Bezerra GB, de Castro LN, Von Zuben FJ (2004) A hierarchical immune network applied to gene expression data. In: Nicosia G, Cutello V, Bentley PJ, Timmis JI (eds) Artificial Immune Systems: Third International Conference, ICARIS 2004. Catania, September 13–16. LNCS, vol 3239. Springer, Berlin, pp 14–27
99. Watkins A, Timmis J, Boggess L (2004) Artificial immune recognition system (AIRS): An immune-inspired supervised learning algorithm. Genet Programm Evolv Mach 5:291–317
100. Meng L, van der Putten P, Wang H (2005) A comprehensive benchmark of the artificial immune recognition system (AIRS). Advanced Data Mining and Applications, Proceedings Lecture Notes in Artificial Intelligence, vol 3584. pp 575–582
101. Watkins AB (2001) AIRS: A resource limited artificial immune classifier. Department of Computer Science, vol MS. Mississippi State University, pp 81
102. Watkins AB (2005) Exploiting immunological metaphors in the development of serial, parallel and distributed learning algorithms. Ph D, University of Kent, pp 314
103. de Castro LN, Von Zuben FJ (2000) The clonal selection algorithm with engineering applications. In: Whitley D, Goldberg D, Cantu-Paz E, Spector L, Parmee I, Beyer HG (eds) GECCO-2000: Proceedings of the Genetic and Evolutionary Computation Conference, July 10–12. Las Vegas, Morgan Kaufmann, pp 36–37
104. de Castro LN, Von Zuben FJ (2002) Learning and optimization using the clonal selection principle. IEEE Trans Evol Comput 6:239–251
105. Brownlee J (2005) Clonal selection theory & CLONAG. The clonal selection classification algorithm (CSCA). Centre for Intelligent Systems and Complex Processes (CISCP), Faculty of Information and Communication Technologies (ICT), Swinburne University of Technology (SUT). Victoria
106. Carter JH (2000) The immune system as a model for pattern recognition and classification. J Am Med Inf Assoc 7:28–41
107. Brownlee J (2005) Immunos-81. The misunderstood artificial immune system. Centre for Intelligent Systems and Complex Processes (CISCP), Faculty of Information and Communication Technologies (ICT), Swinburne University of Technology (SUT). Victoria
108. Witten IH, Frank E (2005) Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco, pp 525
109. Brownlee J (2005) Artificial immune recognition system (AIRS). A review and analysis. Centre for Intelligent Systems and Complex Processes (CISCP), Faculty of Information and Communication Technologies (ICT), Swinburne University of Technology (SUT). Victoria
110. Watkins A, Timmis J (2002) Artificial immune recognition system (AIRS): Revisions and refinements. Artificial Immune Systems: First International Conference, ICARIS 2002. University of Kent at Canterbury, pp 173–181
111. Fenichel RR, Malik M, Antzelevitch C, Sanguinetti M, Roden DM, Priori SG, Ruskin JN, Lipicky RJ, Cantilena LR (2004) Drug-induced torsades de pointes and implications for drug development. J Cardiovasc Electrophysiol 15:475–495
112. Ivanciuc O (2006) Artificial immune system classification of drug-induced torsade de pointes with AIRS (artificial immune recognition system). Internet Electron J Mol Des 5:488–502
113. Ivanciuc O (2007) Artificial immune systems in drug design: Structure-activity relationships for torsade de pointes with AIRS (artificial immune recognition system). Internet Electron J Mol Des 6:47–62
114. Stenberg P, Luthman K, Artursson P (2000) Virtual screening of intestinal drug permeability. J Control Release 65:231–243
115. Ponce YM, Pérez MAC, Zaldivar VR, Sanz MB, Mota DS, Torrens F (2005) Prediction of intestinal epithelial transport of drug in (Caco-2) cell culture from molecular structure using in silico approaches during early drug discovery. Internet Electron J Mol Des 4:124–150
116. Linnankoski J, Makela JM, Ranta VP, Urtti A, Yliperttula M (2006) Computational prediction of oral drug absorption based on absorption rate constants in humans. J Med Chem 49:3674–3681
117. Iyer M, Tseng YJ, Senese CL, Liu J, Hopfinger AJ (2007) Prediction and mechanistic interpretation of human oral drug absorption using MI-QSAR analysis. Mol Pharmaceutics 4:218–231
118. Ivanciuc O (2006) Artificial immune system prediction of the human intestinal absorption of drugs with AIRS (artificial immune recognition system). Internet Electron J Mol Des 5:515–529
119. Ivanciuc O (2007) Feature Selection in AIRS (Artificial Immune Recognition System) Structure-Activity Relationships. Internet Electron J Mol Des 6:331–344

120. Crivori P, Reinach B, Pezzetta D, Poggesi I (2006) Computational models for identifying potential P-glycoprotein substrates and inhibitors. Mol Pharmaceutics 3:33–44
121. Kaiser D, Terfloth L, Kopp S, Schulz J, de Laet R, Chiba P, Ecker GF, Gasteiger J (2007) Self-organizing maps for identification of new inhibitors of P-glycoprotein. J Med Chem 50:1698–1702
122. Ivanciuc O (2006) Artificial immune systems in drug design: Recognition of P-glycoprotein substrates with AIRS (artificial immune recognition system). Internet Electron J Mol Des 5:542–554
123. Ivanciuc O (2006) Structure-activity relationships with artificial immune systems: Classification of benzodiazepine receptor ligands with AIRS, CLONALG, CSCA, and IMMUNOS. Internet Electron J Mol Des 5:585–604
124. Verhaar HJM, Solbé J, Speksnijder J, van Leeuwen CJ, Hermens JLM (2000) Classifying environmental pollutants: Part 3. External validation of the classification system. Chemosphere 40:875–883
125. Ivanciuc O (2003) Aquatic toxicity prediction for polar and nonpolar narcotic pollutants with support vector machines. Internet Electron J Mol Des 2:195–208
126. Ivanciuc O (2007) Artificial immune systems in aquatic toxicology: Structure-activity relationships for the mechanism of toxic action with AIRS (artificial immune recognition system). Internet Electron J Mol Des 6:13–28
127. Ivanciuc O (2007) Drug design with artificial immune systems: Structure-activity relationships for glycogen phosphorylase B inhibitors with CLONALG (clonal selection algorithm). Internet Electron J Mol Des 6:311–319
128. Ivanciuc O (2007) Structure-activity relationships in aquatic toxicology with artificial immune systems: Mechanism of toxic action classification of polar and nonpolar narcotic pollutants with CLONALG (clonal selection algorithm). Internet Electron J Mol Des 6:106–114
129. Ivanciuc O (2007) Artificial immune systems structure-activity relationships for estrogen receptor ligands with CSCA (clonal selection classification system). Internet Electron J Mol Des 6:81–89
130. Ivanciuc O (2007) Artificial immune systems in the virtual screening of dihydrofolate reductase inhibitors with CSCA (clonal selection classification system). Internet Electron J Mol Des 6:253–261
131. Ivanciuc O (2007) Drug design with artificial immune systems: Classification of angiotensin converting enzyme inhibitors with CSCA (clonal selection classification system). Internet Electron J Mol Des 6:135–143
132. Ivanciuc O (2007) Artificial immune systems in structure-activity relationships: Classification of thermolysin inhibitors with CSCA (clonal selection classification system). Internet Electron J Mol Des 6:209–217
133. Ivanciuc O (2007) Structure-activity relationships for acetylcholinesterase inhibitors with the IMMUNOS artificial immune system. Internet Electron J Mol Des 6:167–175
134. Ivanciuc O (2007) Virtual screening of cyclooxygenase-2 inhibitors with the IMMUNOS artificial immune system. Internet Electron J Mol Des 6:200–208
135. Ivanciuc O (2007) Structure-activity relationships with the IMMUNOS artificial immune system for thrombin inhibitors. Internet Electron J Mol Des 6:262–270

Drug Design with Artificial Neural Networks

Ovidiu Ivanciuc
Department of Biochemistry and Molecular Biology, University of Texas Medical Branch, Galveston, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Perceptron
Multilayer Feedforward Artificial Neural Network
Radial Basis Function Network
Self-organizing Map
Counterpropagation Neural Network
Graph Machine Neural Networks
Other Neural Networks
Future Directions
Bibliography

Glossary

Artificial neuron An artificial neuron is a mathematical function that simulates in a simplified form the functions of biological neurons. Usually, an artificial neuron has four computational functions, namely receives signals through input connections from other neurons or from the environment, sums the input signals, applies a nonlinear functions (transfer function or activation function) to the sum, and sends the result to other neurons or as output from the neural network.

Counterpropagation neural network The counterpropagation neural network is a hybrid network that consists of a self-organizing map as the hidden layer and an output layer that has as output a computed value for the modeled property. The network implements a supervised learning algorithm that converges to a unique solution.

Multilayer feedforward artificial neural network A multilayer feedforward (MLF) artificial neural network consists of artificial neurons organized in layers. The MLF network has an input layer that receives the structural descriptors for each molecule, an output layer that provides one or more computed properties, and one or more hidden layers situated between the input and the output layers. Each neuron in a hidden layer receives signals from neurons in the preceding layer and sends signals to the neurons in the next layer.

Perceptron A perceptron is a linear classifier that consists of a layer of input neurons and an output neuron. Each connection between an input neuron and the output neuron has a weight. Depending on the sum of the signals received by the output neuron, its output is +1 or −1.

Quantitative structure-activity relationships Quantitative structure-activity relationships (QSAR) represent regression models that define quantitative correlations between the chemical structure of molecules and their physical properties (boiling point, melting point, aqueous solubility), chemical properties and reactivities (chromatographic retention, reaction rate), or biological activities (cell growth inhibition, enzyme inhibition, lethal dose). The fundamental hypotheses of QSAR are that similar chemicals have similar properties, and that small structural changes result in small changes in property values. The general form of a QSAR equation is $P(i) = f(\mathbf{SD}_i)$, where $P(i)$ is a physical, chemical, or biological property of compound i, $\mathbf{SD}_i$ is a vector of structural descriptors of i, and f is a mathematical function such as linear regression, partial least squares, artificial neural networks, or support vector machines. A QSAR model for a property P is based on a dataset of chemical compounds with known values for the property P, and a matrix of structural descriptors computed for all chemicals. The learning (training) of the QSAR model is the process of determining the optimum parameters of the regression function f. After the training phase, a QSAR model may be used to predict the property P for novel compounds that are not present in the learning set of molecules.

Radial basis function network The radial basis function (RBF) neural network has three layers, namely an input layer, a hidden layer with a nonlinear RBF activation function and a linear output layer.

Self-organizing map A self-organizing map (SOM) is an artificial neural network that uses an unsupervised learning algorithm to project a high dimensional input space into a two dimensional space called a map. The topology of the input space is preserved in SOM, and points that are close to each other in the SOM grid correspond to input vectors that are close to each other in the input space. A SOM consists of neurons arranged usually in a rectangular or hexagonal grid. Each neuron has a position on the map and a weight vector of the same dimension as the input vectors.

Structural descriptor A structural descriptor (SD) is a numerical value computed from the chemical structure of a molecule, which is invariant to the numbering of the atoms in the molecule. Structural descriptors may be classified as constitutional (counts of molecular fragments, such as rings, functional groups, or atom pairs), topological indices (computed from the molecular graph), geometrical (volume, surface, charged-surface), quantum (atomic charges, energies of molecular orbitals), and molecular field (such as those used in CoMFA, CoMSIA, or CoRSA).

Structure–activity relationships Structure–activity relationships (SAR) represent classification models that can discriminate between sets of chemicals that belong to different classes of biological activities, usually active/inactive towards a certain biological receptor. The general form of a SAR equation is $C(i) = f(\mathbf{SD}_i)$, where $C(i)$ is the activity class of compound i (active/inactive, inhibitor/non-inhibitor, ligand/non-ligand), $\mathbf{SD}_i$ is a vector of structural descriptors of i, and f is a classification function such as k-nearest neighbors, linear discriminant analysis, random trees, random forests, Bayesian networks, artificial neural networks, or support vector machines.

Definition of the Subject

The fundamental hypothesis of the structure-property models is that the structural features of molecules determine the physical, chemical and biological properties of chemical compounds. The first studies that use structure-activity relationships to explain the biological properties of sets of compounds were published by Kopp [74], Crum-Brown and Frazer [18], Meyer [88], and Overton [97]. Modern structure-activity relationships (SAR) and quantitative structure-activity relationships (QSAR) models are based on the Hansch model that predicts a biological property as a statistical correlation with steric, electronic, and hydrophobic indices [27,35,36,37]. The Hansch model shaped the general scene of structure-activity correlations, and almost all subsequent SAR and QSAR models are variations that extend the Hansch model with novel classes of descriptors or with more powerful statistical models, such as partial least squares (PLS), artificial neural networks (ANN), support vector machines (SVM), or other machine learning algorithm. A structural descriptor is a numerical representation of some important molecular features, such as empirical indices (Hammett and Taft substituent constants), physical properties (octanol-water partition coefficient, dipole moment, aqueous solubility), counts of substructures or substituents, graph descriptors [13,54,125], topological indices [4,12,55], connectivity indices [65,66], electrotopological indices [67], geometrical descriptors (molecular surface and volume),

quantum indices (atomic charges, HOMO and LUMO energies) [60,124], and molecular fields (steric, electrostatic, and hydrophobic). SAR represent classification models that are used when the experimental property is a class label (+1/−1), such as substrate/non-substrate, inhibitor/non-inhibitor, ligand/non-ligand, toxic/non-toxic, or carcinogen/non-carcinogen. Classification models are used to screen chemical libraries and to identify compounds that have a desired biological activity, such as inhibitor for a particular enzyme. QSAR represent regression models that are used for experimental properties with continuous values, such as hydrophobicity, aqueous solubility, membrane penetration, lethal concentration, or inhibition constant for an enzyme.

The drug discovery and development process is lengthy, expensive, and has a high attrition rate. The duration of the drug development phases, namely pre-clinical (1 to 5 years), clinical (5 to 11 years), and approval (0.5 to 2 years) [9], puts at 18 years the upper limit for bringing a drug to market, without considering the duration of the drug discovery phase. For 168 drugs approved between 1992 and 2002, the median clinical trial and approval periods were 5.1 and 1.2 years, respectively, with no tendency of decreasing despite exceptional technological advances [63]. Cost data for 68 randomly selected drugs from 10 pharmaceutical companies show that better discovery and screening programs can reduce the total cost per approved drug by up to $US 242 million [20]. A similar analysis shows that the average out-of-pocket cost per new drug is $US 403 million (2000 dollars), which increases to $US 802 after adding the opportunity cost (the cost of pursuing one choice instead of another) [21]. Depending on the therapy or the developing firm, the cost per new drug varies between $US 500 million to $US 2,000 million [1]. The attrition rate of the chemical compounds is also impressive, considering that the process starts with millions of compounds tested in high throughput screening and eventually ends with one successful drug on the market. From all compounds that enter the clinical trials, 30% fail due to lack of efficacy, 30% fail in toxicological and safety tests, and only 11% finish the trials [73]. A study considering all 548 new chemical entities approved between 1975 and 1999 found that 45 drugs (8.2%) received one or more black box warnings and 16 (2.9%) were withdrawn from the market [76]. Examples of drugs withdrawn from the market due to adverse reactions include troglitazone (Rezulin) in 2000 [25], cerivastatin (Baycol, Lipobay) in 2001 [28], nefazodone (Serzone) in 2003 [17], pemoline (Cylert) in 1999 (Canada) and 2005 (USA) [42], and rofecoxib (Vioxx) in 2004 [22]. For all difficulties, expenses, and failures that are associated with the drug development process, the reward for a successful drug can be substantial. The top 2006 best selling drugs (in $US billions) are lipitor with 14.39 (Pfizer, cholesterol), advair with 6.13 (GlaxoSmithKline, asthma), plavix with 6.06 (Bristol–Myers Squibb, vascular disease), nexium with 5.18 (AstraZeneca, acid reflux), norvasc with 4.87 (Pfizer, hypertension).

Computer-assisted drug design (CADD) uses computational chemistry to increase the chances of finding valuable drug candidates. Among the broad range of computational tools used in CADD, artificial neural networks (ANN) have a special place due to their abilities to model with high accuracies the complex relationships between chemical structures and molecular properties. ANN can be used both for classification (SAR) and regression (QSAR), and they are essential tools in drug discovery cycles to identify active compounds in chemical libraries, and to optimize a wide range of physico-chemical and biological properties, such as enzyme inhibition, target selectivity, or membrane transport. The SAR and QSAR models based on ANN may predict with high accuracy properties for novel chemicals, even before their chemical synthesis. CADD systems based on ANN may increasing the chances of bringing a drug to market, when they are integrated into the chemical, biological, pharmacological and clinical procedures used by the pharmaceutical industry.

Introduction

The simulation of brain mechanisms with artificial neural networks was initiated by McCulloch and Pitts [84,101] and by Rosenblatt [106]. Minsky and Papert investigated the mathematical properties of these simple neural networks in their seminal book *Perceptrons* [92]. Although perceptrons are able to solve some problems, they fail to learn the XOR (exclusive-or) operation. As a result, the research in this field was limited until Hopfield demonstrated that by adding nonlinear functions to each artificial neuron the network could simulate complex behavior [43].

Another milestone in ANN research was the discovery of the backpropagation algorithm for the training of multilayer feedforward artificial neural networks [107,108]. Multilayer feedforward (MLF) networks consist of multiple layers of artificial neurons, all connected in a feedforward way, from the input layer to the output layer. MLF networks have at least one hidden layer of neurons, and the neurons in the hidden and output layers have nonlinear activation functions. The hidden neurons and the nonlinear activation functions are essential in modeling nonlinear relationships between input variables and the output

property. MLF networks represent the most widespread ANN type used to model physico-chemical, biological, and toxicological properties of chemical compounds.

In addition to MLF, other networks used in drug design are radial basis function networks, self-organizing maps, counter-propagation networks, graph machine neural networks, and probabilistic neural networks. Several reviews [49,52,144] and books are recommended for a detailed overview of neural networks and their applications: *Neural Computing* by Wasserman [134], *Neural Networks for Pattern Recognition* by Bishop [10], *Pattern Recognition and Neural Networks* by Ripley [104], *Neural Networks for Chemical Engineers* by Bulsari [15], and *Neural Networks in Chemistry and Drug Design* by Zupan and Gasteiger [145].

The artificial neuron and the perceptron are described in Sect. "Perceptron". MLF networks are used to model physico-chemical properties, such as melting temperature [75], boiling temperature [33], aqueous solubility [46], and *n*-octanol/water partition coefficients [123]. Drug design applications of MLF networks include models for enzyme inhibition [120], human intestinal absorption [136], virtual screening of chemical libraries [102], and genotoxicity [127]. Another important application of MLF networks is in cancer diagnosis from gene expression data [70,119,135]. Section "Multilayer Feedforward Artificial Neural Network" is an overview of the most important MLF network models in structure-property prediction and in drug design.

The radial basis function (RBF) neural network belongs to the class of feedforward artificial neural network, and it has three layers, namely an input layer, a hidden layer with a nonlinear RBF activation function and a linear output layer [93]. RBF networks give predictive models for various properties, such as boiling temperature [81], QSAR for enzyme inhibitors [99,100], ligand-target binding affinity [142], and aquatic toxicity [87]. SAR and QSAR applications of the RBF network are presented in Sect. "Radial Basis Function Network".

The self-organizing map (SOM) was developed by Kohonen as an unsupervised neural network that transforms a high dimensional input space into a two dimensional space [71,72]. SOM maps the objects by trying to preserve the topological properties of the input space. To predict the class of a new molecule, a vector of structural descriptors is sent as input for a trained SOM and projected onto a neuron from the map. The molecule is classified in the same class with the training molecules that are projected onto the same neuron. SOM networks cluster chemical compounds according to their similarity, and thus can be used to identify groups of molecules with similar properties [3]. Applications of SOM are found in virtual screening of chemical libraries [105], in modeling the aquatic toxicity of organic chemicals [98], or in structure-activity studies [11]. These and other SOM applications are examined in Sect. "Self-organizing Map".

The counterpropagation neural network (CPNN) was developed by Hecht-Nielsen by transforming a self-organizing map into a supervised learning neural network [39,40,41]. The input layer of neurons is connected to the hidden layer that is a self-organizing map, and the excited neuron from SOM sends a signal to the output layer. CPNN may be used for classification or for regression. SAR applications of CPNN include the mutagenicity prediction of aromatic amines [59], and the classification of organic compounds based on their mode of toxic action [117]. Applications of the counterpropagation neural network in SAR and QSAR are presented in Sect. "Counterpropagation Neural Network".

Graph machine neural networks learn SAR and QSAR models directly from the molecular graph [32], by translating the chemical structure into the network topology. For each chemical compound presented to the graph machine, the neural network adopts a structure derived from the molecular graph of the chemical. There are several graph machines used with success in SAR and QSAR, such as recursive neural networks [5,89], the Baskin–Palyulin–Zefirov (BPZ) neural device [7], ChemNet [69], and MolNet [47,50]. The BPZ neural device is constructed by analogy with a biological vision system and contains a sensor field, a set of eyes, and a brain. A chemical structure investigated with the BPZ neural device is represented as a molecular matrix that is superimposed over a sensor field that receives the network input. The information received by the sensor field is transmitted to a set of eyes that transforms it into several MLF networks, and sends signals to the next block, the brain. ChemNet encodes a molecular distance matrix between the input and hidden layers, and has as output an atomic property. In MolNet each atom in a molecule has a corresponding input and hidden neuron, and the connections between the input and hidden layers correspond to the weighted interactions between atoms. The output neuron offers the computed value for a molecular property. In Sect. "Graph Machine Neural Networks" we review structure-activity models computed with graph machine neural networks.

Other ANN used in drug design and molecular property modeling are the Bayesian regularized neural network [31,133], the probabilistic neural network [38,77], and the general regression neural network [141]. Their applications in SAR and QSAR are presented in Sect. "Other Neural Networks".

Perceptron

An artificial neuron is a simple model of a biological neuron expressed as a mathematical function. Artificial neurons represent the basic computational units in an artificial neural network. Each artificial neuron has four basic functions (Fig. 1): (a) receives input signals through connections from the neurons in the previous layer or from the environment; (b) makes a summation of the inputs signals; (c) the result of summation is passed to the activation function and transformed with a mathematical function; (d) the result of the activation function is used by a transfer function to produce an output value that is send to the neurons in the next layer of neurons. The artificial neuron depicted in Fig. 1 receives the input signals $x_1, \ldots, x_n$ through connections with the weights $w_1, \ldots, w_n$. The summation of the input signal is computed as the dot product between $\mathbf{w}$ and $\mathbf{x}$ plus a bias parameter b. An activation function Act is applied to the sum Σ to give the output value z. The first artificial neuron was the threshold logic unit proposed by McCulloch and Pitts [84,101]. The activation function was a step function, which has the value zero for a negative argument and one for a positive argument. Other important activation functions are: linear, $\mathrm{Act}(z) = z$; sigmoid, $\mathrm{Act}(z) = 1/(1 + e^{-z})$ [43]; hyperbolic tangent, $\mathrm{Act}(z) = \tanh(z)$; symmetric logarithmoid, $\mathrm{Act}(z) = \mathrm{sign}(z)\ln(1 + |z|)$ [15].

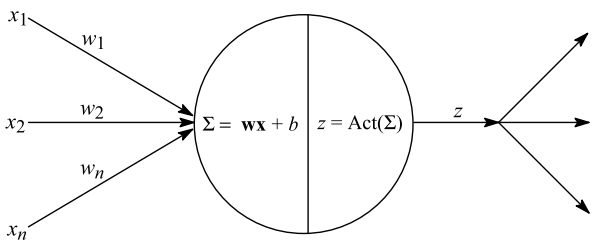

Drug Design with Artificial Neural Networks, Figure 1
Structure and functions of an artificial neuron

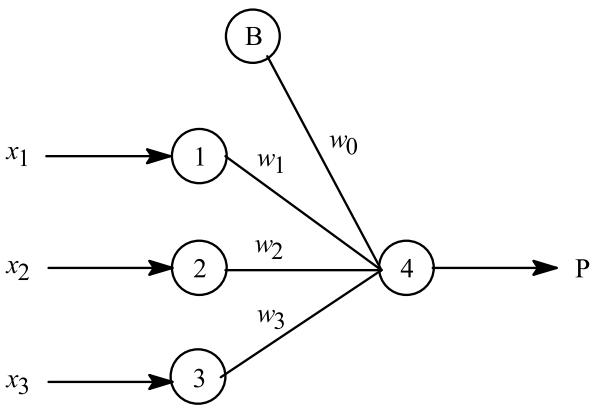

Drug Design with Artificial Neural Networks, Figure 2
Perceptron structure and signal flow

A perceptron is the simplest type of ANN, consisting of a layer of input neurons and an output neuron (Fig. 2) [92,106]. The perceptron shown in Fig. 2 has three input neurons (1, 2, 3), a bias neuron B, and an output neuron 4. An input neuron receives an input signal, scales it and then sends it to the output neuron. A simple perceptron is a linear classifier, which cannot learn simple operations, such as the XOR (exclusive-or) operation. A major development was the introduction of nonlinear activation functions that allows ANN to simulate any nonlinear function [43].

Multilayer Feedforward Artificial Neural Network

A multilayer feedforward artificial neural network, called also a multilayer perceptron (MLP) is a generalization of the simple perceptron obtained by inserting one or more hidden layers of neurons between the input layer and the output layer. Another important characteristic is the use of nonlinear activation functions in the hidden and output layers, although the output neurons can have also a linear activation function. MLF ANN are universal approximators, namely they can approximate any function with arbitrary precision provided that enough neurons are used in the hidden layer.

In SAR and QSAR applications the structural descriptors represent the input data received by the input neurons. As an example, we consider topological descriptors computed from the molecular graph of *t*-butylcyclopropane (Fig. 3a) that are used as input in an MLF ANN with one hidden layer (Fig. 3b). The input layer is formed by neurons 1–4, the hidden layer contains neurons 5–7, and the output layer is represented by neuron 8. The bias neurons B send a signal with the value +1 through weighted connections, and have the role to fine-tune each neuron. The topological descriptors represent the counts of vertex degrees, namely $\mathbf{Deg} = \{3, 2, 1, 1\}$, i. e., there are three atoms with degree 1, two atoms with degree 2, one atom with degree 3, and one atom with degree 4. These numbers are used as input to neurons 1–4, which perform a scaling and then send the values to each neuron from the hidden layer. The output neuron 8 offers the computed value of a property P for *t*-butylcyclopropane.

The MLF ANN application to property modeling consists of two phases, namely training and prediction. During the training (learning) phase the weights of all connections are optimized (adjusted) in order to estimate with high precision the investigated molecular property.

The most popular training method is the backpropagation algorithm [107,108], a variant of the gradient algorithms. More efficient training algorithms are direction-set methods (Powell's method), methods that require the computation of the first derivatives such as conjugate gradient methods (Fletcher–Reeves or Polak–Ribiere), Levenberg–Marquardt, or quasi-Newton (variable metric) methods (Davidon–Fletcher-Powell or Broyden–Fletcher-Goldfarb–Shanno).

MLF ANN can be trained also with artificial intelligence algorithms, such as particle swarm optimization, which was used in a QSAR study of inhibitors of platelet-derived growth factor receptor phosphorylation [113]. Such algorithms can provide a complete optimization of an MLF ANN, namely they can optimize both the network topology and the connection weights. The network topology is defined by the number of hidden layers and by the number of neurons in each hidden layer. The complete optimization of ANN with particle swarm optimization has a fast convergence to the global minimum [85] and avoids the overfitting of the learning dataset of chemicals [112]. The selection of the most important descriptors for the ANN input is an important step in obtaining predictive models. Several artificial intelligence algorithms can be used for an efficient feature selection, such as genetic algorithm [115,116], particle swarm optimization [2], and ant colony optimization [58].

Taskinen and Yliruusi reviewed the MLF ANN applications in predicting several physico-chemical properties of interest in drug development, such as octanol-water partition coefficient, aqueous solubility, boiling temperature, and vapor pressure [121]. The melting temperatures for sulfur-containing organic compounds were predicted from simple molecular graph indices [75]. The melting temperatures of 717 ionic liquids were modeled with MLF ANN, associative neural networks, support vector machines, *k*-nearest neighbors, and multiple linear regression [126]. The structural descriptors evaluated in these models were electrotopological indices [67] and Dragon descriptors [124]. The moderate accuracy of predictions obtained with all algorithms can be explained by the difficulty to encode the solid state characteristics of ionic liquids. Hall and Story used electrotopological indices and an MLF ANN to model the boiling temperatures of 298 compounds and the critical temperatures of 165 compounds [33]. The best results for the boiling temperatures were obtained with a neural network with five hidden neurons with a mean absolute error of 3.93 K, whereas a network with four hidden neurons gave good predictions for the critical temperatures with a mean absolute error of 4.52 K. In a similar study, atom type electrotopological state indices were used to model the boiling temperatures of alkanes, alcohols and chloroalkanes [34].

Huuskonen et al. developed MLF ANN models for the aqueous solubility of a diverse set of 734 organic compounds [46]. Atom type electrotopological state indices were used as input data in a neural network with five hidden neurons. Good results were obtained both in training ($s = 0.52$) and prediction ($s = 0.75$), showing that the aqueous solubility for a large and diverse set of organic compounds can be predicted from topological indices. Tetko et al. developed a neural network model for the octanol-water partition coefficients of 12,908 organic compounds [123]. The model consists of an ensemble of

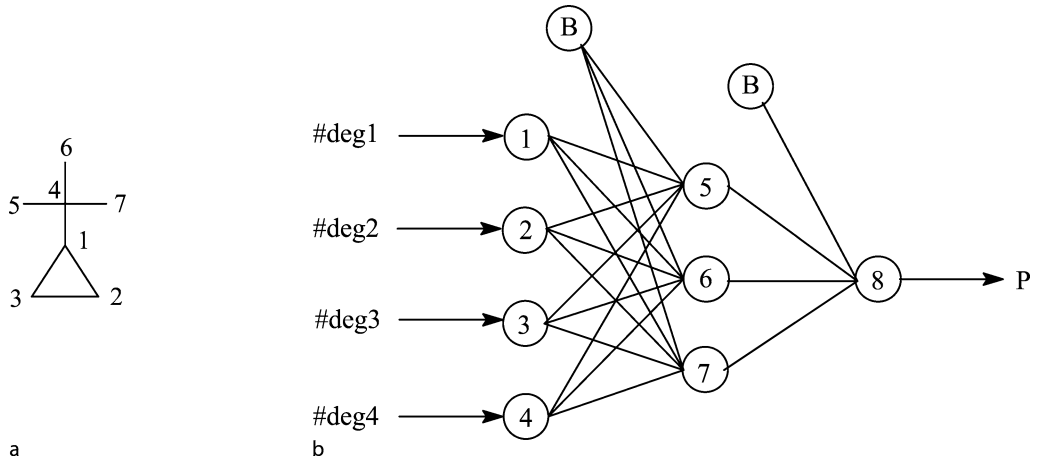

Drug Design with Artificial Neural Networks, Figure 3
ANN prediction of molecular properties from the molecular structure: topological descriptors computed from a the molecular graph of *t*-butylcyclopropane are used as input in b a multilayer feedforward neural network with one hidden layer

50 ANN, each ANN having 75 input neurons and 10 neurons in the hidden layer. The leave-one-out cross-validation indicate that the model gives good predictions, with $r^2 = 0.95$. The ^{13}C NMR chemical shift can be predicted from simple topological parameters, as shown for a dataset of acyclic alkenes [56,57].

To compensate for the lack of toxicological data for aquatic organisms, a number of neural networks QSAR models were proposed for the toxicity of organic chemicals against fathead minnow, rainbow trout, bluegill sunfish, *Daphnia magna*, *Tetrahymena pyriformis*, and *Vibrio fischeri* [61]. These ANN models can be used to predict the environmental impact of novel chemicals, and to prioritize their experimental evaluation in aquatic toxicity assays. Huuskonen tested multiple linear regression and artificial neural networks in QSAR models for toxicity of 140 organic chemicals against fathead minnow [45]. A neural network trained with 14 electrotopological state indices provided a predictive model for the experimental toxicity. Basak et al. modeled the toxic modes of action for 283 chemicals with ANN and topological indices [6]. The toxic compounds were classified as narcotics, electrophiles/proelectrophiles, uncouplers of oxidative phosphorylation, acetylcholinesterase inhibitors, and neurotoxicants. The modes of action SAR models provide reliable predictions, with rates of correct classification between 65% and 95%. Lower predictions are obtained for classes with a low number of compounds.

Dopamine antagonists, serotonin antagonists, and serotonin-dopamine dual antagonists are used as antipsychotics. Based on a training dataset consisting of 1135 dopamine antagonists, 1251 serotonin antagonists, and 386 serotonin-dopamine dual antagonists, Kim et al. tested several machine learning algorithms, including ANN and recursive partitioning [68]. The average classification rate in training was 73.6% and in prediction was 69.8%, indicating that these models can be used in virtual screening to identify new active compounds. Siu and Che used multiple linear regression, partial least squares, and ANN to model the α-amino acids affinities for H^+ and for sodium, copper, and silver cations [114]. The cross-validation tests indicate that the best predictions are obtained with the artificial neural network. These QSAR models are useful in elucidating the binding properties of the α-amino acids.

Votano et al. used ANN, *k*-nearest neighbors, and decision forest to model the mutagenicity of 3363 diverse compounds tested for their Ames genotoxicity [127]. The dataset was split into 2963 training compounds and 400 prediction compounds, and the SAR models were developed with 148 topological indices that included electrotopological state indices and molecular connectivity indices. All three classifiers gave good predictions, with a slight advantage for the neural network, as indicated by the area under the receiver operator characteristic (AUROC) curve, namely AUROC = 0.93 for ANN, AUROC = 0.92 for *k*-nearest neighbors, and AUROC = 0.91 for decision forest.

A good intestinal absorption is a desirable property for drugs, but its experimental determination is costly and time-consuming. Wessel et al. used ANN to estimate the percent human intestinal absorption (% HIA) of 86 drugs [136]. A neural network optimized with a genetic algorithm provided the best results, with an error of 9.4% HIA units for training, 19.7% HIA units the cross-validation, and 16.0% HIA units for an external prediction set.

Sutherland et al. evaluated ANN, PLS, genetic function approximation, and genetic PLS in QSAR models for inhibitors of angiotensin converting enzyme, acetylcholinesterase, benzodiazepine receptor, cyclooxygenase-2, dihydrofolate reductase, glycogen phosphorylase B, thermolysin, and thrombin [120]. ANN gave the best predictions for QSAR models obtained with graph indices and geometrical descriptors. Another large scale evaluation of structure-activity models was made for chemicals that target HIV-reverse transcriptase, COX2, dihydrofolate reductase, estrogen receptor, and thrombin [102]. A comparison of several nonlinear modeling algorithms shows that ANN are successful in identifying active compounds from chemical libraries and can be used in virtual screening. Pharmaceutical companies constantly enrich their collections of chemical compounds by acquisitions that increase the structural diversity of molecules available for high throughput screening. Muresan and Sadowski proposed an ANN system to compute an "in-house likeness" score for compound acquisition [94]. The analysis of several datasets shows that a set of atom-type counts used as input to the ANN models represents an efficient way of identifying structural patterns that are missing from an in-house collection.

Votano et al. compared multiple linear regression, ANN, *k*-nearest neighbors, and support vector machines in QSAR models for the human serum protein binding of 1008 chemicals (808 for training and 200 for prediction) [128]. The best predictions ($r^2 = 0.7$) were obtained with ANN models, indicating that these QSAR can assist the drug design process. Katritzky et al. evaluated ANN and multiple linear regression in QSAR models for 277 inhibitors of glycogen synthase kinase-3 [62]. The descriptors were selected from a pool consisting of subgraph counts, topological indices, geometrical parameters and quantum indices. The QSAR mod-

els highlight the structural factors that influence the inhibition of glycogen synthase kinase-3. Inhibitors of the hERG (human ether-á-go-go-related gene) potassium channel can lead to a prolongation of the QT interval that can trigger torsade de pointes, an atypical ventricular tachycardia. Seierstad and Agrafiotis used ANN to model the nonlinear structure-activity relationship for hERG channel inhibitors [110]. The QSAR models with best predictive power are obtained with a feature selection to prevent over-fitting and by aggregating several ANN into an ensemble model to minimize the instability in predicting novel chemicals. The time- and dose-dependent anti-inflammatory in vivo activity of substituted imidazo[1,2-*a*]pyridines was modeled with artificial neural networks [109]. This study shows that ANN have unique properties to study pharmacodynamic and pharmacokinetic properties of drugs and drug-like compounds. Peptides that have a good binding affinity for major histocompatibility complex (MHC) class I molecules are candidates in the development of new vaccines against cancer and viral infections. Filter et al. modeled the peptide binding to MHC class I allele HLA-A*0201 with artificial neural networks [26]. The ANN system can identify peptides that do not correspond to the usual recognition motifs, as shown for several new melanoma-associated peptides.

Artificial neural networks represent the method of choice in cancer prediction and prognosis [19]. Their use improves the accuracy of predicting cancer susceptibility, recurrence and mortality. Molecular data, such as protein biomarkers and microarrays, together with ANN models can improve the understanding of cancer development and progression. The mechanism of action of drugs against 60 malignant cell lines in the National Cancer Institute drug screening program was investigated with ANN [135]. Out of the 141 drugs, the ANN gives incorrect predictions for only 12 of them, showing that the model can be used to guide the screening program. Khan et al. used ANN and gene expression data to identify the class of small, round blue-cell tumors [64]. Although the classification of these tumors is difficult in clinical practice, the neural networks identified the relevant genes and correctly classified all samples. ANN trained with gene microarray data can identify with success several types of cancers, such as colon cancer [111] and esophageal cancer [138].

Radial Basis Function Network

An RBF neural network is a feedforward network with three layers, in which the neurons in the hidden layer have an RBF activation function [93]. RBF neural networks are universal approximators, can approximate any continuous function with arbitrary precision if enough neurons are allowed in the hidden layer. An early application of the RBF network in structure-property studies was reported by Lohninger for the boiling temperatures of organic compounds [81]. The training process of a radial basis function artificial neural network (RBF ANN) consists of selecting the network topology, finding the centers and widths of the RBF neurons, and computing the connection weights between the hidden and output layers. Zhou et al. showed that the particle swarm optimization is very efficient in training an RBF neural network [143]. The procedure was tested with a dataset of 40 inhibitors of murine P388 leukemia cells and 70 structural descriptors, and the results indicate that the swarm optimization converges fast to a global optimum. Wan and Harrington compared two algorithms for RBF ANN training, namely K-means clustering and linear averaging [131]. The results obtained for two datasets (polychlorobiphenyl mass spectra and Italian olive oil classification) indicate that the linear averaging method gives better predictions.

QSAR models for glycine/NMDA receptor antagonists, which may treat stroke or seizure, were computed for a dataset consisting of 109 compounds [99]. The set of structural descriptors included topological indices, geometric descriptors, quantum indices, and polar surface parameters. The optimum set of descriptors was selected with a genetic algorithm and an RBF network. SAR models for inhibitors of protein tyrosine phosphatase 1B, which are potential agents for the treatment for type 2 diabetes and obesity, were computed for 128 compounds [100]. The classification models were obtained with RBF networks, linear discriminant analysis, and *k*-nearest neighbors. A correct classification rate of 85.7% was obtained for an external prediction set, indicating that the SAR model can be used to find inhibitors in virtual libraries. Neuraminidase is a key target in treating infections with the influenza virus, and many classes of inhibitors are tested for this enzyme. In order to identify the structural features that influence the inhibition of this enzyme, Lü et al. evaluated RBF networks in QSAR models for 46 neuraminidase inhibitors [82]. Five important descriptors were found with a heuristic search among several hundred indices. Prediction tests show that the RBF network results are comparable with those obtained with a multiple linear model, indicating a possible absence of nonlinear structure-activity relationships for these inhibitors.

Zheng et al. compared RBF networks, support vector machines, and partial least square in QSAR models for the Ah receptor binding affinities of polychlorinated, polybrominated, and polychlorinated-brominated

dibenzo-p-dioxins [142]. The best leave-one-out cross-validation predictions are obtained by the RBF network, with $q^2 = 0.88$. Melagraki et al. used RBF networks to model the toxicity of 39 organic compounds against *Vibrio fischeri* [87]. The descriptors used in the simulation include topological indices, electronegativity parameters, and lipophilicity. The good QSAR predictions indicate that the model can be used to evaluate the toxicity of diverse chemicals. QSAR models based on RBF networks were developed for the toxicity of 221 phenols against *Tetrahymena pyriformis* [86]. Prediction tests performed with cross-validation and with an external set show that the RBF network gives more accurate predictions than multiple linear regression.

2	2	2		1	1	1	1
2	2				1	1	1
2	2			1	1	1	1
3	3	3					
3	3		1		1		
3	3	3	1	1	1	1	
			1	1	1		

Drug Design with Artificial Neural Networks, Figure 4
A self-organizing map with three classes

Self-organizing Map

A self-organizing map is an artificial neural network that projects a high-dimensional input space into a low-dimensional (usually two dimensional) output space [71,72]. A SOM is trained with unsupervised learning, meaning that the training molecules do not need a class label or property value. The input vectors are mapped in a two dimensional space by preserving the topological properties of the input space. The output layer in a SOM is a matrix of neurons, with the property that an input vector is projected into a single output neuron. Several similar input vectors may be projected into the same output neuron, or into adjacent output neurons that form clusters of similar objects. This SOM property is used to discover clusters in collections of chemical compounds. The chemical similarity of large chemical libraries is easily evaluated by projecting the molecules onto a SOM that auto-organizes similar compounds by preserving their similarity relationships in the descriptor space. SOM are used to screen libraries to identify clusters of structurally similar compounds that may have similar biological properties.

SOM may be used also for classification by assigning class labels to a trained network (Fig. 4). This example shows an 8×8 SOM with labels attached to the output neurons. The class labels are attached after training and represent the majority class for objects projected onto a particular output neuron. The objects are organized in four clusters, two for class 1, one for class 2, and one for class 3. The SOM predictions are straightforward, namely a new object is projected onto an output neuron and takes the label of the output neuron as its predicted class. Single molecules may be mapped onto a SOM network by projecting the points situated on the molecular surface [3]. The Cartesian coordinates of each surface point are used as input for the SOM network that provides a two-dimensional display of the molecular surface. To add more structural information on the map, output neurons are colored according to the molecular electrostatic potential of the surface points projected onto them.

SOM networks are very efficient in discovering activity clusters in chemical libraries that contain active compounds for several targets. Terfloth and Gasteiger clustered with SOM a number of 299 compounds representing four activity classes, namely 75 5-hydroxytryptamine 5-HT_{1a}-receptor agonists, 75 histamine H_2-receptor antagonists, 74 monoamine oxidase MAO-A inhibitors, and 75 thrombin inhibitors [122]. All classes are generally clustered in distinct regions, but a small number of neurons represent chemicals from different classes. Such degenerate mapping may be solved by increasing the size of the map or by testing other structural descriptors that may discriminate better these activity classes. Drug candidates are routinely screen to determine if they are hERG channel blockers, because such chemicals can cause sudden cardiac death. Roche et al. examined the hERG activity of a chemical library with SOM, PLS, MLF ANN, principal component analysis, and substructure analysis [105]. These classification models offer good predictions for a validation set, with 71% success rate for active compounds and 93% for inactive compounds.

Trypanosoma cruzi is deposited on the skin surface by bugs from the Reduviidae family, and then it infects the human host by penetrating through insect bites, thus causing trypanosomiasis, or Chagas disease. The chronic form of the disease may develop more than 10 years after the infection, and may be fatal. Boiani et al. developed several computational models to evaluate the anti-*Trypanosoma cruzi* activity of *N*-oxide containing heterocycles [11]. A SOM was applied to separate between

activity classes and to select an optimum set of descriptors. Other models tested with success for this dataset are *k*-nearest neighbors and decision tree.

QSAR models have limitations for the domain of chemical structures and biological activities for which they give reliable predictions. For example, interpolation gives better predictions than extrapolation, and extrapolations far away from the training structural space or activity space are not reliable. Also, all compounds used to train a QSAR should have the same mechanism of interaction with the biological target. A SOM can easily identify if a prediction compound is similar or not with the training set of chemicals. If the prediction compound is mapped in the SOM regions occupied by the training compounds, then the QSAR prediction has a high confidence level. SOM is also used to select from a large dataset only those chemicals that act through a common mechanism or form a cluster of compounds with similar structures. Gini et al. applied such an approach to separate molecules into homogeneous groups of chemicals with similar biological and structural properties [30]. These groups were then used to compute local QSAR models for the aquatic toxicity of chemicals against *Pimephales promelas* and *Tetrahymena pyriformis*. A proper validation of a QSAR model gives vital information regarding its prediction error and validity domain. SOM is a good choice to split a dataset into a training set and a validation (prediction) set. Papa et al. applied this method to develop QSAR models for the toxicity of organic chemicals against *Pimephales promelas* [98].

SOM networks are flexible tools to cluster and visualize large chemical libraries. von Korff and Hilpert applied SOM to cluster a library containing 3000 G-protein-coupled receptor (GPCR) ligands for about 130 receptors and 3000 non-GPCR ligands [147]. The chemical structure was described with fingerprints and topological pharmacophores. Cross-validation tests indicate that SOM can separate GPCR ligands from non-GPCR ligands. A possible application of the SAR model is in the design and identification of focused libraries that target a specific GPCR receptor. The supervised self-organizing map (sSOM) is a SOM variant that uses property labels during the training phase. Xiao et al. compared sSOM with *k*-nearest neighbors, PLS and other linear methods in their ability to model a set of 256 inhibitors of dihydrofolate reductase [137]. The predictions tests show that sSOM is insensitive to noise and provides more accurate predictions than linear QSAR methods.

The experimental data obtained in the National Cancer Institute antitumor drug screening program were investigated with SOM to identify clusters of chemicals having the same biological action [103]. SOM clusters highlight four classes of mechanism of action of drugs, namely mitosis, nucleic acid synthesis, membrane transport and integrity, and phosphatase- and kinase-mediated cell cycle regulation. SOM and fuzzy C-means clustering were used by Wang et al. to identify marker genes and to classify tumors based on DNA microarray data [132]. The methods were tested with good results for four datasets, namely leukemia, colon cancer, brain tumors, and National Cancer Institute cancer cell lines. For each dataset the algorithm identified sets of marker genes with predictive power for tumor classification.

Counterpropagation Neural Network

Hecht-Nielsen developed the counter-propagation neural network (CPNN) as a supervised learning network that combines a self-organizing map that sends signals to an output layer of neurons [39,40,41]. CPNN applications in chemistry and in structure-activity are reviewed by Zupan et al. [146] and by Vracko [129]. Recent CPNN studies in molecular modeling and structure-activity relationships are reviewed in this section.

Jezierska et al. used CPNN to select an optimum group of descriptors for mutagenicity QSAR models obtained for 95 aromatic and heteroaromatic amines [59]. Starting from 275 descriptors, CPNN selected six topological indices and the LUMO energy as the best combination of input parameters. The reduced set of descriptors give good predictions, with $R_{cv} = 0.85$ obtained in leave-one-out cross-validation. Numerous organic chemicals are environmental pollutants, and a considerable number of studies are dedicated to the computational prediction of their mechanism of aquatic toxicity (MOA). The reliable prediction of MOA has major applications in selecting the appropriate QSAR model, to identify chemicals with similar toxicity mechanism, and in extrapolating toxic effects between different species and exposure regimens [53]. Spycher et al. developed MOA classifiers for 115 compounds comprising nine MOA classes and 24 descriptors [117]. The four classifiers evaluated, namely CPNN, PLS, logistic regression, and linear discriminant analysis, had an overall correct classification between 52% and 59%, as determined by five-fold cross-validation. Several MOA classes contain a small number of compounds, which might explain the low prediction rate. In a related study, CPNN and logistic regression classifiers were computed for a dataset of 220 phenols with four MOA [118]. The rate of correct predictions in five-fold cross-validation was 92%, much higher than in the previous study that considered nine MOA classes. The higher prediction rate may be a result

of a larger number of compounds in each MOA class and of a more homogeneous set of chemicals.

CPNN and support vector machines (SVM) were compared in a QSAR study for the estrogen receptor binding affinity of 131 compounds [29]. Several methods of feature selection were tested, and the predictions obtained with CPNN and SVM vary significantly with the different sets of selected descriptors. Mazurek et al. investigated CPNN structure-selectivity models for a set of artificial metalloenzymes [83]. These catalysts perform the enantioselective hydrogenation of the acetamidoacrylic acid with a high enantiomeric excess (%ee). The structural descriptors are computed from the geometry of the complexes ligand-metalloenzyme, and the CPNN is trained to predict the %ee. The predictions obtained for an external test set ($R = 0.953$ and RMS $= 16.8$ %ee) provide convincing evidence for the utility of ligand-metalloenzyme descriptors in modeling structure-selectivity relationships. Wagner et al. applied CPNN to structure-activity models for the NF-*K*B inhibition by 103 sesquiterpene lactones [130]. The activity range for NF-*K*B inhibition was partitioned in six activity classes, and the chemical structure was expressed with autocorrelation descriptors based on atomic properties. A ten-fold cross-validation shows that the rate of correct classification is 80.6%, and an external validation with 14 new compounds has 78.6% correct predictions. CPNN and *k*-nearest neighbors classifiers were applied to the classification of plants from the Asteraceae family based on a set of sesquiterpene lactones isolated from individual species [44]. The separation of plants in three tribes and seven subtribes is based on structural descriptors computed from the molecular structure of sesquiterpene lactones. This approach can be applied for the automatic identification of plant species based on chemical analysis and machine learning.

Graph Machine Neural Networks

Usual structure-activity models consider the chemical structure only indirectly, through descriptors computed from various representations of molecules. In these models, the topology of neural networks remains constant for all training and prediction molecules. All training molecules, large or small, are represented by a constant-length vector of descriptors, thus restricting the use of the chemical structure in models. A different approach is taken in artificial neural networks based on graph machines which learn SAR and QSAR models directly from the molecular graph, by translating the chemical structure into the network topology [32]. In this section we review four graph machine neural networks, namely recursive neural networks [5,89,90], the Baskin–Palyulin–Zefirov (BPZ) neural device [7], ChemNet [69], and MolNet [47,50].

Recursive Neural Networks

Recursive neural networks (RecNN) map a directed ordered acyclic graph to a molecular property. All chemical compounds modeled in a RecNN model must be represented as a directed ordered acyclic graph. As an example consider 2-chloro-1-methyl-4-nitrobenzene (Fig. 5a) as a member of series of compounds that have a common skeleton, which in this case is benzene. The molecular structure is transformed into an ordered tree as shown in Fig. 5b. All molecules in the training set are transformed into ordered trees. Using these structured data as input, RecNN reflects the molecular topology as encoded in the ordered tree. The theory and applications of RecNN in SAR and QSAR are presented in several reviews [5,89,90]. RecNN were applied to several structure-property stud-

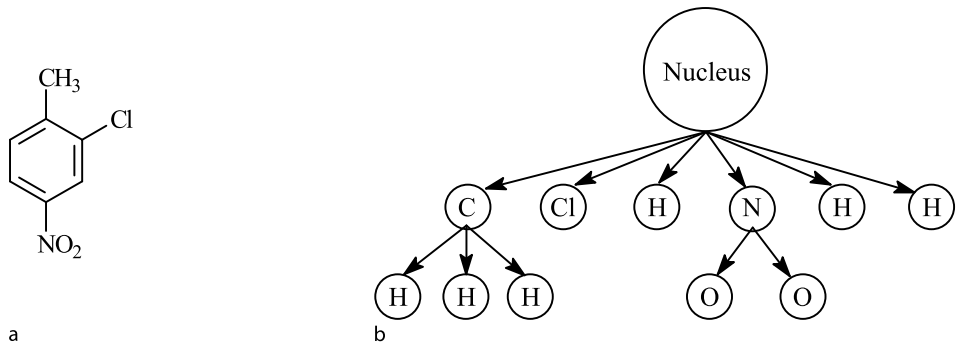

Drug Design with Artificial Neural Networks, Figure 5
Representation of congeneric chemical compounds in recursive neural networks: a 2-chloro-1-methyl-4-nitrobenzene b represented as a tree

ies, including models for the glass transition temperature of polymers [23,24], QSAR for the binding affinities of ligands to the benzodiazepine receptor [91], and for the Gibbs free energy of solvation in water of organic compounds [8].

Baskin–Palyulin–Zefirov Neural Device

The Baskin–Palyulin–Zefirov (BPZ) neural device (Fig. 6) is a computational implementation of a biological vision system and contains a sensor field, a set of eyes, and a brain [7]. A molecule submitted for evaluation to the BPZ neural device is represented as a molecular matrix superimposed over the sensor field. The sensor field receives the network input as a matrix representation of a molecule, and then sends this information to a set of eyes. The structural information is processed in the eyes by several multilayer feedforward neural networks, and the output signal is sent to the brain. The signals received from the eyes are processed in the brain by an MLF ANN and the output represents the computed property for the molecule sent as input to the sensor field. A set of rules is used to translate a molecule into the structure of the sensor field and of the eyes.

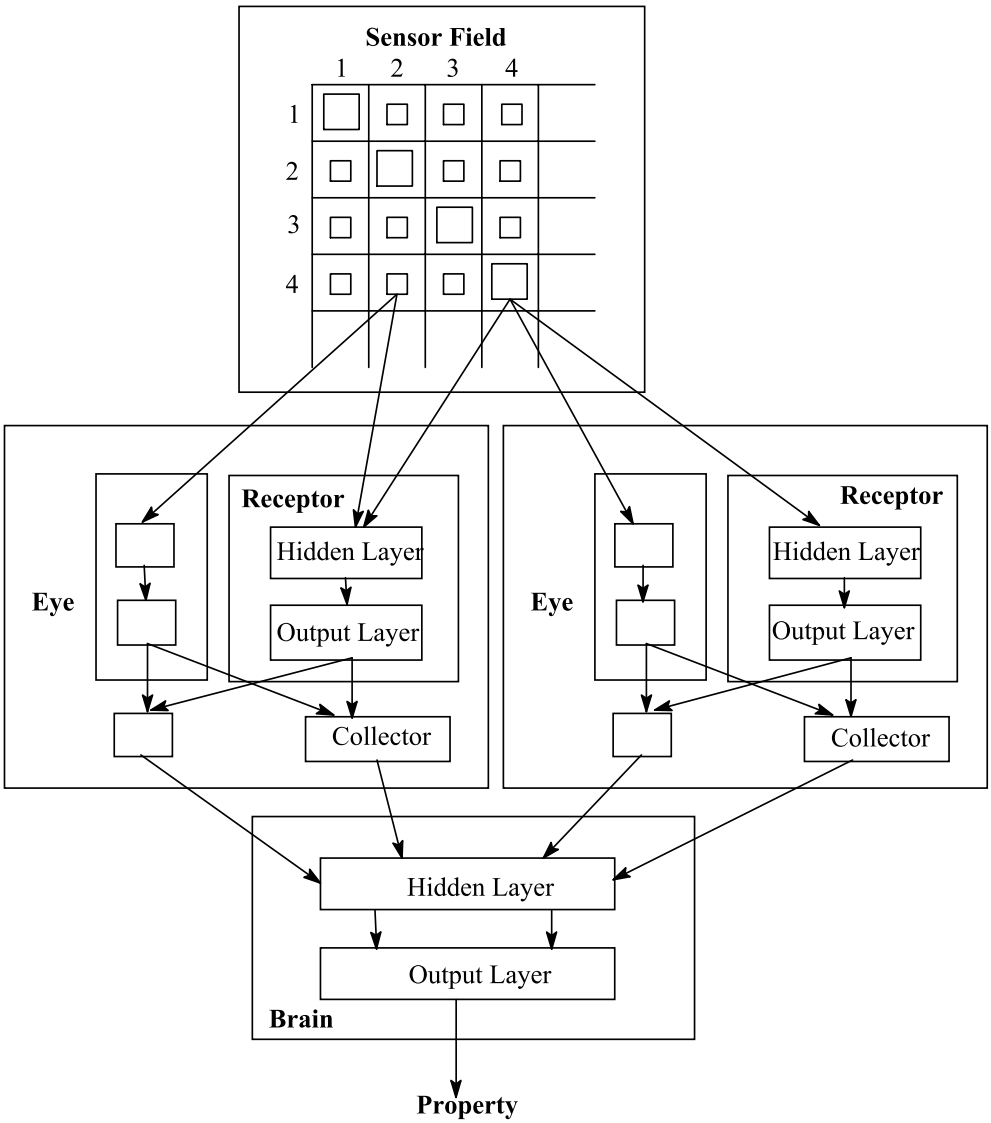

Drug Design with Artificial Neural Networks, Figure 6
The structure of the Baskin–Palyulin–Zefirov neural device

A BPZ neural device with a minimal configuration has one receptor in each eye, and can be used as a template to generate a neural device for any molecule presented to the sensor field (Fig. 7). In the minimal configuration, Eye 1 receives signals from individual atoms, whereas Eye 2 receives from the sensor field signals corresponding to pairs of bonded atoms. The information from the sensor neuron 1 goes to neurons 3 and 4 representing the hidden layer of the receptor R1 from Eye 1. The signal goes then to the output neuron 7 and then to the collector C1 that sends the output to the brain. The information flow is similar in Eye 2, with the only difference that the input signal represents information regarding bonds between atoms. Neurons 11 and 12 form the hidden layer from the brain, whereas neuron 13 is the output neuron that offers the computed molecular property.

ChemNet

ChemNet is an MLF neural network that translates a molecular structure into the network topology to compute an atomic property [69]. Each molecule is represented by its labeled hydrogen-depleted molecular graph. For each atom i from the molecular graph ChemNet has

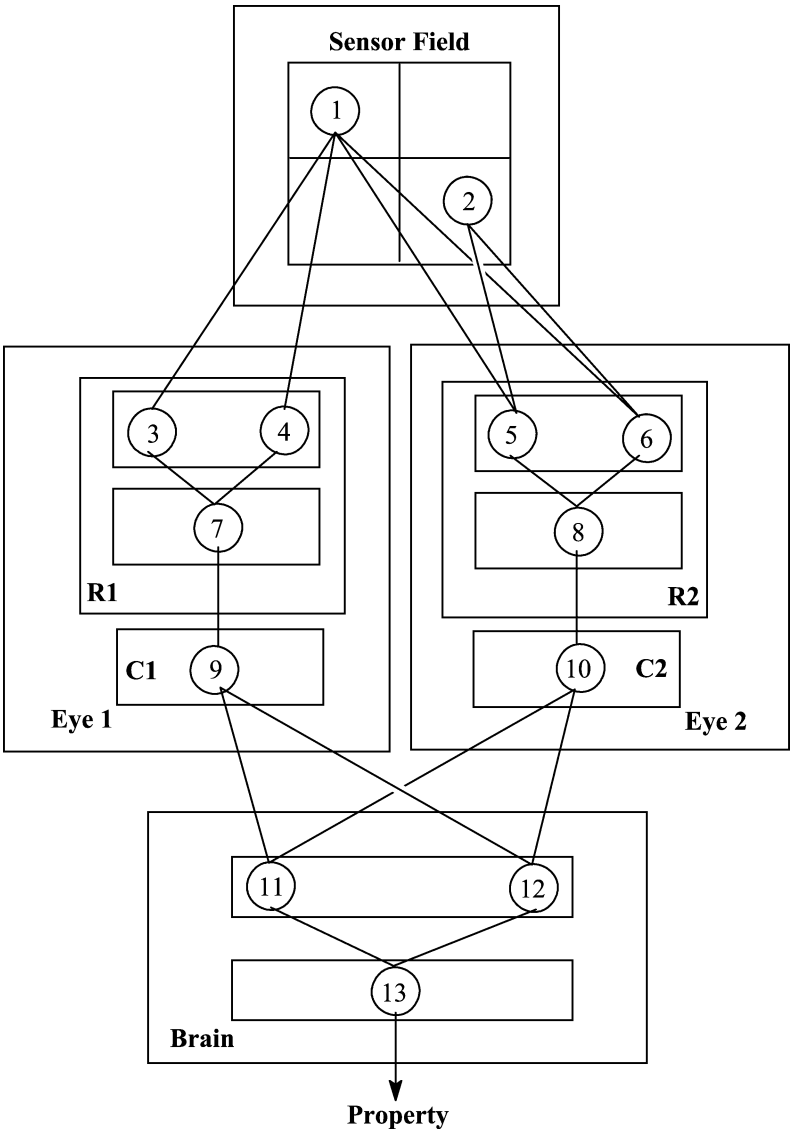

Drug Design with Artificial Neural Networks, Figure 7
A minimal configuration of the Baskin–Palyulin–Zefirov neural device

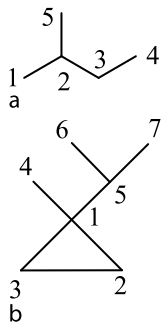

Drug Design with Artificial Neural Networks, Figure 8
Molecular graphs used to generate ChemNet and MolNet neural networks: a 2-methylbutane; b 1-methyl-1-isopropylcyclopropane

two corresponding neurons with the same label *i*, namely one in the input layer and the second one in the hidden layer. The output layer has only one neuron, representing the atom from the molecular graph whose property is computed with ChemNet. The connections between the input and hidden layers correspond to the bonding relationships between pairs of atoms, i. e. two pairs of atoms separated by the same number of bonds have identical connection weights. The ChemNet structure is demonstrated for 2-methylbutane (Fig. 8a). The distance matrix for the molecular graph of 2-methylbutane is:

$$\mathbf{D} = \begin{vmatrix} 0 & 1 & 2 & 3 & 2 \\ 1 & 0 & 1 & 2 & 1 \\ 2 & 1 & 0 & 1 & 2 \\ 3 & 2 & 1 & 0 & 3 \\ 2 & 1 & 2 & 3 & 0 \end{vmatrix}$$

Each neuron from the input and hidden layers of ChemNet corresponds to the atom with the same label from the molecular graph of 2-methylbutane (Fig. 9). Input values represent the number of hydrogen atoms attached to each atom in the molecular graph of 2-methylbutane. All pairs of atoms that are separated by the same number of bonds are characterized by input-hidden connections with identical weights. The distance matrix determines the connections that have the same weight. All ChemNet connections corresponding to pairs of atoms separated by one bond have the same weight (Fig. 9a). The connections between atoms separated by two bonds are shown in Fig. 9b, whereas the connections between atoms separated by three bonds are shown in Fig. 9c. All connections from the bias neuron to the 5 hidden neurons have the same weight (Fig. 9d). The ChemNet structure from Fig. 9 computes a property for atom 1. The connections between the hidden layer and the output neuron correspond to the bonding relationships between atom 1 and all other atoms in the molecule, namely atoms situated at distance 1 (Fig. 9a), distance 2 (Fig. 9b), and distance 3 (Fig. 9c).

MolNet

MolNet is a graph machine that maps the distance matrix in the ANN structure and computes a molecular property using atomic descriptors as input data [47,48,50,51]. The number of neurons in the input and hidden layers is equal to the number of atoms from the molecular graph, and each atom in a molecule has a corresponding neuron in the input and hidden layers. The output layer has only one neuron that provides the computed value for the investigated molecular property. The bonding pattern between two atoms is defined by the ordered sequence of atom types and bond types situated on the shortest path between the two atoms. All identical bonding patterns correspond to network connections with identical weights. An input neuron *i* is connected to the hidden neuron *i* with a connection type that depends on the chemical nature of the corresponding atom *i*. All $i - i$ connections corresponding to the same atomic species have identical weights. The connections from the hidden layer to the output layer (HO connections) are classified according to the type of atoms represented by the hidden neurons. The atoms are partitioned into classes according to their atomic number *Z*, the hybridization state and the degree. All hidden neurons representing atoms of the same type, either in the same molecule or in different molecules, are connected to the output neuron by connections with identical weights. The bias neuron is connected to each hidden neuron with connections (BH connections) partitioned similarly with the connections between the hidden and output layers, i. e. according to the atom types as defined above. The bias neuron is also connected with the output neuron (BO connection). The MolNet structure of 1-methyl-1-isopropylcyclopropane (Fig. 8b) is based the distance matrix computed from the molecular graph:

$$\mathbf{D} = \begin{vmatrix} 0 & 1 & 1 & 1 & 1 & 2 & 2 \\ 1 & 0 & 1 & 2 & 2 & 3 & 3 \\ 1 & 1 & 0 & 2 & 2 & 3 & 3 \\ 1 & 2 & 2 & 0 & 2 & 3 & 3 \\ 1 & 2 & 2 & 2 & 0 & 1 & 1 \\ 2 & 3 & 3 & 3 & 1 & 0 & 2 \\ 2 & 3 & 3 & 3 & 1 & 2 & 0 \end{vmatrix}$$

The MolNet connections between the input and hidden layers corresponding to 1-methyl-1-isopropylcyclopropane are shown in Fig. 10. The neural network is sep-

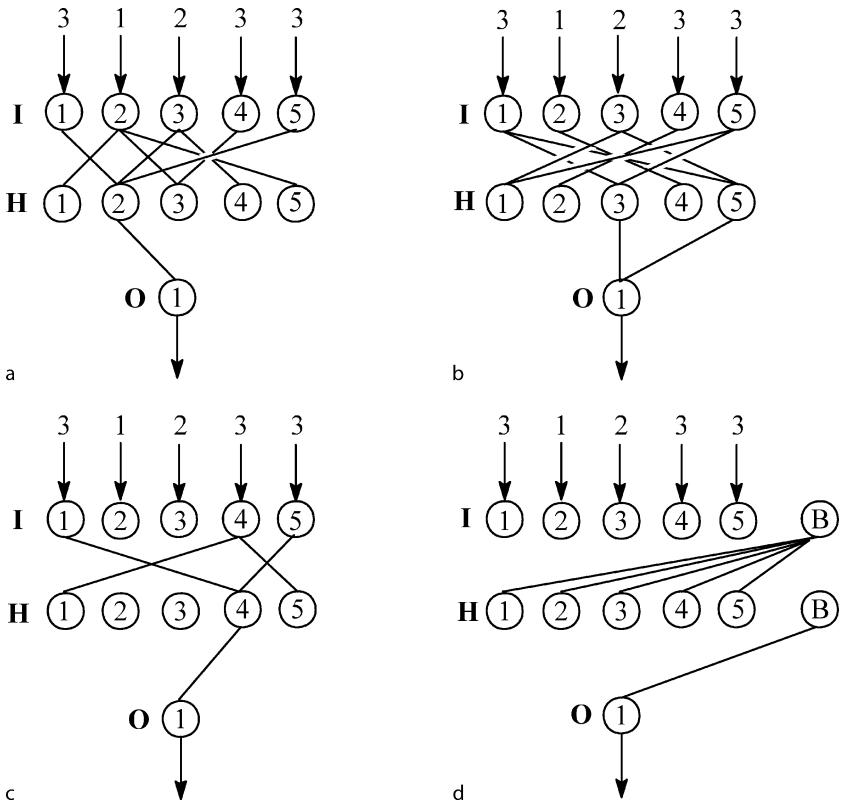

Drug Design with Artificial Neural Networks, Figure 9
ChemNet topology for 2-methylbutane (Fig. 8a) that computes a property of atom 1. Each neuron in the input (I) and hidden (H) layers corresponds to the atom with the same label from the molecular graph of 2-methylbutane. The connections between atoms situated at topological distances 1, 2, and 3 are presented in a, b, and c, respectively, whereas the connections from the bias neuron B are presented in d. Input values represent the number of hydrogen atoms attached to each atom in the molecular graph of 2-methylbutane

arated into four sections, each corresponding to a bonding relationship. Because the molecule considered here has only carbon atoms and single bonds, all connections depicted in a section from Fig. 10 have identical weights. Input values represent vertex degrees, but any atomic index may be used as input to the network. The network consists of connections between atoms with the same label (Fig. 10a), and between atoms situated at topological distances 1 (Fig. 10b), 2 (Fig. 10c), and 3 (Fig. 10d), respectively.

The connections between the hidden and output layers (HO connections) are determined by the structure of the molecule presented to MolNet. HO connections in alkanes are classified according to the degree of the carbon atoms: hidden neurons representing atoms with identical degrees are linked to the output neuron with connections having identical weights. The connections between the bias neuron and the neurons in the hidden layer (BH connections) are classified according to the same rules used for HO connections. MolNet connections between the hidden and output layers for 1-methyl-1-isopropylcyclopropane are presented in Fig. 11.

Other Neural Networks

In this section we present an overview of other artificial neural networks applied in structure-activity models and in drug design. A Bayesian regularized neural network (BRNN) was evaluated in QSAR for K_m values of cytochrome P450 3A4 substrates [133]. Correlations with electrotopological state indices give predictive and robust models. Bruneau and McElroy applied BRNN to models for the distribution coefficient log $D_{7.4}$ computed for a set of 5000 compounds [14]. The neural model obtained is stable and can be applied for compounds in different ionization states. The steady-state volume of distribution V_{ss} for drugs and drug-like chemicals was computed with BRNN for a set of 199 compounds (human V_{ss}) and a sec-

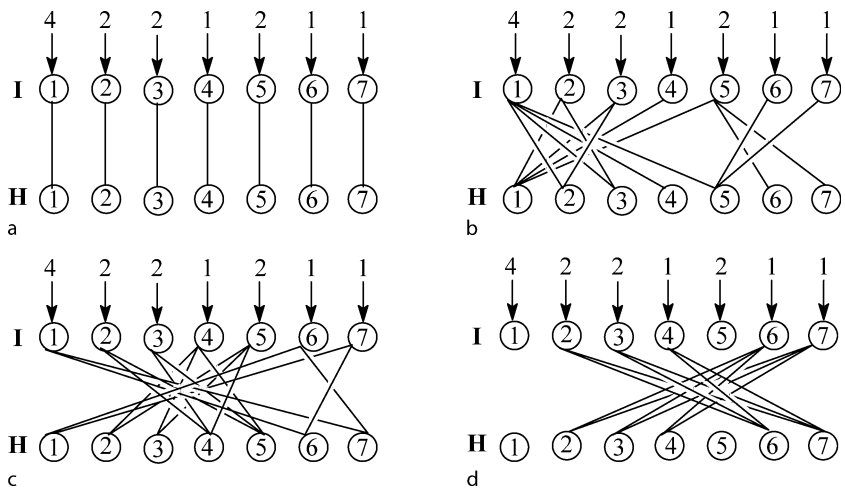

Drug Design with Artificial Neural Networks, Figure 10

MolNet connections between the input (I) and hidden (H) layers for 1-methyl-1-isopropylcyclopropane (Fig. 8b); each neuron corresponds to the carbon atom with the same label from the molecular graph of 1-methyl-1-isopropylcyclopropane. The connections between atoms with the same label are presented in a, whereas the connections between atoms situated at topological distances 1, 2, and 3 are presented in b, c, and d, respectively. Input values represent vertex degrees

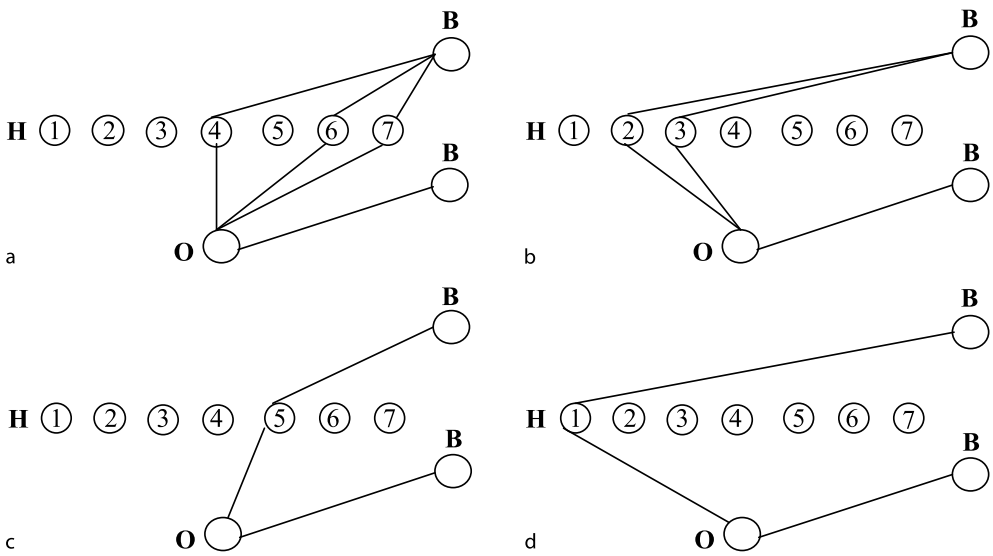

Drug Design with Artificial Neural Networks, Figure 11

MolNet connections between the hidden (H) and output (O) layers for 1-methyl-1-isopropylcyclopropane (Fig. 8b); the bias neuron is labeled with B. The bias-hidden and hidden-output connections to/from hidden neurons representing atoms with the degree 1, 2, 3, and 4 are presented in a, b, c, and d, respectively

ond set of 2086 compounds (rat V_{ss}) [31]. Both models have good predictions statistics, and can be used to filter chemicals in the early stages of the drug discovery process. Caballero et al. computed a BRNN model for the inhibition constant of 46 inhibitors of the cysteine protease cruzain [16]. The neural network was also used to rank the importance of the structural descriptors.

Probabilistic neural networks (PNN) and general regression neural network (GRNN) have similar architectures, with the difference that PNN perform classification whereas GRNN perform regression. Niwa applied PNN to identify the biological target from the chemical structure of their ligands [96]. The neural network was trained with a set of 799 compounds having activities against seven

biological targets. On average, 90% of the compounds were correctly classified. PNN classification models were developed for the genotoxic potential of 85 quinolones and 115 quinolines [38]. The quinolone dataset had 23 genotoxic and 62 nongenotoxic compounds, whereas the quinoline dataset had 44 genotoxic and 71 nongenotoxic chemicals. An ensemble of nine PNN models was developed for each classification model, and the final class attribution (genotoxic/nongenotoxic) was decided by a majority vote of the trained classifiers. Simulated annealing was used to select between three and ten structural descriptors for each PNN classifier. The ensemble PNN model for quinolone derivatives was able to predict correctly 16 of the 23 genotoxic chemicals and 60 of the 62 nongenotoxic compounds, with an overall accuracy of 89.4%, an overall accuracy for genotoxic class of 69.6%, and an overall accuracy for nongenotoxic class of 96.8%. Other structure-activity and drug design applications of PNN include the prediction of estrogen receptor agonists [78], identification of drugs that penetrate the blood-brain barrier [79], estimation of the genotoxicity [77], prediction of P-glycoprotein substrates [140], classification of chemicals that are toxic for *Tetrahymena pyriformis* [139], and identification of factor Xa inhibitors [80]. The general regression neural network was applied for the computation of the total clearance CL_{tot} of 503 compounds [141]. The CL_{tot} of a drug characterizes its bioavailability and elimination, and thus may be used to determine its dose and steady-state concentration. Based on the statistics obtained for a validation dataset of 105 compounds, the best predictions are obtained with the support vector regression followed by GRNN. The percent human intestinal absorption (%HIA) of 86 drug and drug-like chemicals was computed with GRNN and PNN [95]. For an external prediction set, GRNN has a root mean square error of 22.8%HIA and PNN has 80% rate of correct classification. The models for human intestinal absorption are based on topological indices, thus making the GRNN and PNN models strong candidates for screening in large virtual libraries.

Future Directions

Artificial neural networks have a unique appeal in structure-activity predictions because they promise to deliver superior modeling power inspired by the functions and mechanisms of the human brain. Although the initial exaggerated expectations never materialized, and artificial intelligence is still a faraway dream, the drug design applications reviewed here provide compelling evidence for the exceptional modeling abilities of artificial neural networks. Current applications span all major SAR and QSAR properties relevant to drug design and discovery, and new applications are reported in the literature. Future developments should consider an integrated development of optimally predictive models, in which feature selection, network topology optimization, network training, and validation form a single optimization problem. We anticipate significant improvement in the optimization algorithms used in ANN training, inspired by the remarkable results obtained with particle swarm optimization and ant colony optimization. Novel ANN algorithms and procedures should be adopted from the disciplines of computer science and machine learning. It is expected that the continuous increase in computer power will make feasible applications of ANN ensembles for datasets relevant to the pharmaceutical industry. Significant improvement in SAR and QSAR models may come from graph machine applications that incorporate structured data into the ANN topology.

Bibliography

1. Adams CP, Brantner VV (2006) Estimating the cost of new drug development: is it really 802 million dollars? Health Aff 25:420–428
2. Agrafiotis DK, Cedeño W (2002) Feature selection for structure-activity correlation using binary particle swarms. J Med Chem 45:1098–1107
3. Anzali S, Gasteiger J, Holzgrabe U, Polanski J, Sadowski J, Teckentrup A, Wagener M (1998) The use of self-organizing neural networks in drug design. Perspect Drug Discov Design 9–11:273–299
4. Balaban AT, Ivanciuc O (1999) Historical development of topological indices. In: Devillers J, Balaban AT (eds) Topological Indices and Related Descriptors in QSAR and QSPR. Gordon and Breach Science Publishers, Amsterdam, pp 21–57
5. Baldi P, Pollastri G (2004) The principled design of large-scale recursive neural network architectures-DAG-RNNs and the protein structure prediction problem. J Mach Learn Res 4:575–602
6. Basak SC, Grunwald GD, Host GE, Niemi GJ, Bradbury SP (1998) A comparative study of molecular similarity, statistical, and neural methods for predicting toxic modes of action. Environ Toxicol Chem 17:1056–1064
7. Baskin II, Palyulin VA, Zefirov NS (1997) A neural device for searching direct correlations between structures and properties of chemical compounds. J Chem Inf Comput Sci 37:715–721
8. Bernazzani L, Duce C, Micheli A, Mollica V, Sperduti A, Starita A, Tiné MR (2006) Predicting physical-chemical properties of compounds from molecular structures by recursive neural networks. J Chem Inf Model 46:2030–2042
9. Berndt ER, Gottschalk AHB, Strobeck MW (2005) Opportunities for improving the drug development process: Results from a survey of industry and the FDA. National Bureau of Economic Research Workshop on Innovation Policy and the Economy, NBER Working Paper No 11425, Washington, DC

10. Bishop CM (1996) Neural Networks for Pattern Recognition. Oxford University Press, Oxford, 504 pp
11. Boiani M, Cerecetto H, M González, Gasteiger J (2008) Modeling anti-Trypanosoma cruzi activity of *N*-oxide containing heterocycles, J Chem Inf Model 48:213–219
12. Bonchev D (1983) Information Theoretic Indices for Characterization of Chemical Structure. Research Studies Press, Chichester
13. Bonchev D, Rouvray DH (eds) (1991) Chemical Graph Theory. Introduction and Fundamentals. Abacus Press/Gordon & Breach Science Publishers, New York
14. Bruneau P, McElroy NR (2006) $logD_{7.4}$ Modeling using Bayesian regularized neural networks. Assessment and correction of the errors of prediction. J Chem Inf Model 46:1379–1387
15. Bulsari AB (1995) Neural Networks for Chemical Engineers. Elsevier, Amsterdam, 609 pp
16. Caballero J, Tundidor-Camba A, Fernández M (2007) Modeling of the inhibition constant (K_i) of some cruzain ketone-based inhibitors using 2D spatial autocorrelation vectors and data-diverse ensembles of Bayesian-regularized genetic neural networks. QSAR Comb Sci 26:27–40
17. Choi S (2003) Nefazodone (Serzone) withdrawn because of hepatotoxicity. Can Med Assoc J 169:1187–1187
18. Crum-Brown A, Frazer T (1868–1869) On the connection between chemical constitution and physiological action. Part 1. On the physiological action of the ammonium bases, derived from Strychia, Brucia, Thebaia, Codeia, Morphia and Nicotia. Trans Royal Soc Edinburgh 25:257–274
19. Cruz JA, Wishart DS (2006) Applications of machine learning in cancer prediction and prognosis. Cancer Inform 2:59–78
20. DiMasi JA (2002) The value of improving the productivity of the drug development process: faster times and better decisions. Pharmacoeconomics 20(S3):1–10
21. DiMasi JA, Hansen RW, Grabowski HG (2003) The price of innovation: new estimates of drug development costs. J Health Econ 22:151–185
22. Dieppe PA, Ebrahim S, Martin RM, Jüni P (2004) Lessons from the withdrawal of rofecoxib. Br Med J 329:867–868
23. Duce C, Micheli A, Solaro R, Starita A, Tiné MR (2006) Prediction of chemical-physical properties by neural networks for structures. Macromol Symp 234:13–19
24. Duce C, Micheli A, Starita A, Tiné MR, Solaro R (2006) Prediction of polymer properties from their structure by recursive neural networks. Macromol Rapid Commun 27:711–715
25. Faich GA, Moseley RH (2001) Troglitazone (Rezulin) and hepatic injury. Pharmacoepidemiol Drug Saf 10:537–547
26. Filter M, Eichler-Mertens M, Bredenbeck A, Losch FO, Sharav T, Givehchi A, Walden P, Wrede P (2006) A strategy for the identification of canonical and non-canonical MHC I-binding epitopes using an ANN-based epitope prediction algorithm. QSAR Comb Sci 25:350–358
27. Fujita T, Iwasa J, Hansch C (1964) A new substituent constant, π, derived from partition coefficients. J Am Chem Soc 86:5175–5180
28. Furberg CD, Pitt B (2001) Withdrawal of cerivastatin from the world market. Curr Control Trials Cardivasc Med 2:205–207
29. Ghafourian T, Cronin MTD (2006) The effect of variable selection on the non-linear modelling of oestrogen receptor binding. QSAR Comb Sci 25:824–835
30. Gini G, Craciun MV, König C, Benfenati E (2004) Combining unsupervised and supervised artificial neural networks to predict aquatic toxicity. J Chem Inf Comput Sci 44:1897–1902
31. Gleeson MP, Waters NJ, Paine SW, Davis AM (2006) In silico human and rat V_{ss} quantitative structure-activity relationship models. J Med Chem 49:1953–1963
32. Goulon-Sigwalt-Abram A, Duprat A, Dreyfus G (2005) From Hopfield nets to recursive networks to graph machines: Numerical machine learning for structured data. Theor Comput Sci 344:298–334
33. Hall LH, Story CT (1996) Boiling point and critical temperature of a heterogeneous data set: QSAR with atom type electrotopological state indices using artificial neural networks. J Chem Inf Comput Sci 36:1004–1014
34. Hall LH, Story CT (1997) Boiling point of a set of alkanes, alcohols and chloroalkanes: QSAR with atom type electrotopological state indices using artificial neural networks. SAR QSAR Environ Res 6:139–161
35. Hansch C (1969) A quantitative approach to biochemical structure-activity relationships. Acc Chem Res 2:232–239
36. Hansch C, Fujita T (1964) $\rho - \sigma - \pi$ analysis. A method for the correlation of biological activity and chemical structure. J Am Chem Soc 86:1616–1626
37. Hansch C, Maloney PP, Fujita T, Muir RM (1962) Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. Nature 194:178–180
38. He L, Jurs PC, Kreatsoulas C, Custer LL, Durham SK, Pearl GM (2005) Probabilistic neural network multiple classifier system for predicting the genotoxicity of quinolone and quinoline derivatives. Chem Res Toxicol 18:428–440
39. Hecht-Nielsen R (1987) Counterpropagation networks. Appl Optics 26:4979–4984
40. Hecht-Nielsen R (1988) Applications of counterpropagation networks. Neural Netw 1:131–139
41. Hecht-Nielsen R (1990) Neurocomputing, Addison-Wesley, Reading
42. Hogan V (2000) Pemoline (Cylert): market withdrawal. Can Med Assoc J 162:106–106
43. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci USA 79:2554–2558
44. Hristozov D, Da Costa FB, Gasteiger J (2007) Sesquiterpene lactones-based classification of the family Asteraceae using neural networks and *k*-nearest neighbors. J Chem Inf Model 47:9–19
45. Huuskonen J (2003) QSAR modeling with the electrotopological state indices: Predicting the toxicity of organic chemicals. Chemosphere 50:949–953
46. Huuskonen J, Rantanen J, Livingstone D (2000) Prediction of aqueous solubility for a diverse set of organic compounds based on atom-type electrotopological state indices. Eur J Med Chem 35:1081–1088
47. Ivanciuc O (1998) Artificial neural networks applications, Part 9. MolNet prediction of alkane boiling points. Rev Roum Chim 43:885–894
48. Ivanciuc O (1999) Artificial neural networks applications. Part 11. MolNet prediction of alkane densities. Rev Roum Chim 44:619–631
49. Ivanciuc O (1999) Molecular graph descriptors used in neural network models. In: Devillers J, Balaban AT (eds) Topological

Indices and Related Descriptors in QSAR and QSPR. Gordon and Breach Science Publishers, Amsterdam, pp 697–777
50. Ivanciuc O (1999) The neural network MolNet prediction of alkane enthalpies. Anal Chim Acta 384:271–284
51. Ivanciuc O (2000) Artificial neural networks applications. Part 12. The prediction of alkane heat capacities with the MolNet neural network. Rev Roum Chim 45:391–403
52. Ivanciuc O (2001) New neural networks for structure-property models. In: Diudea MV (ed) QSPR/QSAR Studies by Molecular Descriptors. Nova Science Publishers, Huntington, pp 213–231
53. Ivanciuc O (2003) Aquatic toxicity prediction for polar and nonpolar narcotic pollutants with support vector machines. Internet Electron J Mol Des 2:195–208
54. Ivanciuc O (2003) Graph theory in chemistry. In: Gasteiger J (ed) Handbook of Chemoinformatics, vol 1. Wiley-VCH, Weinheim, pp 103–138
55. Ivanciuc O (2003) Topological indices. In: Gasteiger J (ed) Handbook of Chemoinformatics, vol 3. Wiley-VCH, Weinheim, pp 981–1003
56. Ivanciuc O, Rabine J-P, Cabrol DB, Panaye A, Doucet JP (1996) ^{13}C NMR chemical shift prediction of sp^2 carbon atoms in acyclic alkenes using neural networks. J Chem Inf Comput Sci 36:644–653
57. Ivanciuc O, Rabine J-P, Cabrol DB, Panaye A, Doucet JP (1997) ^{13}C NMR chemical shift prediction of the sp^3 carbon atoms in the α position relative to the double bond in acyclic alkenes. J Chem Inf Comput Sci 37:587–598
58. Izrailev S, Agrafiotis DK (2002) Variable selection for QSAR by artificial ant colony systems. SAR QSAR Environ Res 13:417–423
59. Jezierska A, Vračko M, Basak SC (2004) Counter-propagation artificial neural networks as a tool for the independent variable selection: Structure-mutagenicity study on aromatic amines. Mol Divers 8:371–377
60. Jurs P (2003) Quantitative structure-property relationships. In: Gasteiger J (ed) Handbook of Chemoinformatics, vol 3. Wiley-VCH, Weinheim, pp 1314–1335
61. Kaiser KLE (2003) The use of neural networks in QSARs for acute aquatic toxicological endpoints. J Mol Struct (Theochem) 622:85–95
62. Katritzky AR, Pacureanu LM, Dobchev DA, Fara DC, Duchowicz PR, Karelson M (2006) QSAR modeling of the inhibition of glycogen synthase kinase-3. Bioorg Med Chem 14:4987–5002
63. Keyhani S, Diener-West M, Powe N (2006) Are development times for pharmaceuticals increasing or decreasing? Health Aff 25:461–468
64. Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Med 7:673–679
65. Kier LB, Hall LH (1976) Molecular Connectivity in Chemistry and Drug Research. Academic Press, New York
66. Kier LB, Hall LH (1986) Molecular Connectivity in Structure-Activity Analysis. Research Studies Press, Letchworth
67. Kier LB, Hall LH (1999) Molecular Structure Description. The Electrotopological State. Academic Press, San Diego
68. Kim H-J, Choo H, Cho YS, Koh HY, No KT, Pae AN (2006) Classification of dopamine, serotonin, and dual antagonists by decision trees. Bioorg Med Chem 14:2763–2770
69. Kireev DB (1995) ChemNet: A novel neural network based method for graph/property mapping. J Chem Inf Comput Sci 35:175–180
70. Ko D, Xu W, Windle B (2005) Gene function classification using NCI-60 cell line gene expression profiles. Comput Biol Chem 29:412–419
71. Kohonen T (1989) Self-Organization and Associative Memory, 3 edn. Springer, Berlin
72. Kohonen T (1995) Self-Organizing Maps. Springer, Berlin
73. Kola I, Landis J (2004) Can the pharmaceutical industry reduce attrition rates? Nat. Rev. Drug Discov 3:711–715
74. Kopp H (1844) Ueber den Zusammenhang zwischen der chemischen Constitution und einigen physikalischen Eigenschaften bei flüssigen Verbindungen. Ann Chem Pharm 50:71–144
75. Kozioł J (2002) Neural network modeling of melting temperatures for sulfur-containing organic compounds. Internet Electron J Mol Des 1:80–93
76. Lasser KE, Allen PD, Woolhandler SJ, Himmelstein DU, Wolfe SN, Bor DH (2002) Timing of new black box warnings and withdrawals for prescription medications. J Am Med Assoc 287:2215–2220
77. Li H, Ung CY, Yap CW, Xue Y, Li ZR, Cao ZW, Chen YZ (2005) Prediction of genotoxicity of chemical compounds by statistical learning methods. Chem Res Toxicol 18:1071–1080
78. Li H, Ung CY, Yap CW, Xue Y, Li ZR, Chen YZ (2006) Prediction of estrogen receptor agonists and characterization of associated molecular descriptors by statistical learning methods. J Mol Graph Modell 25:313–323
79. Li H, Yap CW, Ung CY, Xue Y, Cao ZW, Chen YZ (2005) Effect of selection of molecular descriptors on the prediction of blood-brain barrier penetrating and nonpenetrating agents by statistical learning methods. J Chem Inf Model 45:1376–1384
80. Lin HH, Han LY, Yap CW, Xue Y, Liu XH, Zhu F, Chen YZ (2007) Prediction of factor Xa inhibitors by machine learning methods. J Mol Graph Modell 26:505–518
81. Lohninger H (1993) Evaluation of neural networks based on radial basis functions and their application to the prediction of boiling points from structural parameters. J Chem Inf Comput Sci 33:736–744
82. Lü WJ, Chen YL, Ma WP, Zhang XY, Luan F, Liu MC, Chen XG, Hu ZD (2008) QSAR study of neuraminidase inhibitors based on heuristic method and radial basis function network. Eur J Med Chem 43:569–576
83. Mazurek S, Ward TR, Novič M (2007) Counter propagation artificial neural networks modeling of an enantioselectivity of artificial metalloenzymes. Mol Divers 11:141–152
84. McCulloch WS, Pitts WH (1943) A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 7:115–133
85. Meissner M, Schmuker M, Schneider G (2006) Optimized Particle Swarm Optimization (OPSO) and its application to artificial neural network training. BMC Bioinformatics 7:125
86. Melagraki G, Afantitis A, Makridima K, Sarimveis H, Igglessi-Markopoulou O (2006) Prediction of toxicity using a novel RBF neural network training methodology. J Mol Model 12:297–305
87. Melagraki G, Afantitis A, Sarimveis H, Igglessi-Markopoulou O, Alexandridis A (2006) A novel RBF neural network training methodology to predict toxicity to Vibrio fischeri. Mol Divers 10:213–221
88. Meyer H (1899) Zur Theorie der Alkoholnarkose, welche

Eigenschaft der Anästhetica bedingt ihre narkotische Wirkung. Arch Exp Pathol Pharmakol 42:109–118
89. Micheli A, Portera F, Sperduti A (2005) A preliminary empirical comparison of recursive neural networks and tree kernel methods on regression tasks for tree structured domains. Neurocomputing 64:73–92
90. Micheli A, Sperduti A, Starita A (2007) An introduction to recursive neural networks and kernel methods for cheminformatics. Curr Pharm Design 13:1469–1495
91. Micheli A, Sperduti A, Starita A, Bianucci AM (2001) Analysis of the internal representations developed by neural networks for structures applied to quantitative structure-activity relationship studies of benzodiazepines. J Chem Inf Comput Sci 41:202–218
92. Minsky ML, Papert SA (1969) Perceptrons. MIT Press, Cambridge
93. Moody J, Darken CJ (1989) Fast learning in networks of locally-tuned processing units. Neural Comput 1:281–294
94. Muresan S, Sadowski J (2005) "In-house likeness": Comparison of large compound collections using artificial neural networks. J Chem Inf Model 45:888–893
95. Niwa T (2003) Using general regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures. J Chem Inf Comput Sci 43:113–119
96. Niwa T (2004) Prediction of biological targets using probabilistic neural networks and atom-type descriptors. J Med Chem 47:2645–2650
97. Overton CE (1901) Studien über die Narkose. Zugleich ein Beitrag zur Allgemeinen Pharmakologie. Gustav Fisher Verlag, Jena
98. Papa E, Villa F, Gramatica P (2005) Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in Pimephales promelas (fathead minnow). J Chem Inf Model 45:1256–1266
99. Patankar SJ, Jurs PC (2002) Prediction of glycine/NMDA receptor antagonist inhibition from molecular structure. J Chem Inf Comput Sci 42:1053–1068
100. Patankar SJ, Jurs PC (2003) Classification of inhibitors of protein tyrosine phosphatase 1B using molecular structure based descriptors. J Chem Inf Comput Sci 43:885–899
101. Pitts WH, McCulloch WS (1947) How we know universals: The perception of auditory and visual forms. Bull Math Biophys 9:127–147
102. Plewczynski D, Spieser SAH, Koch U (2006) Assessing different classification methods for virtual screening. J Chem Inf Model 46:1098–1106
103. Rabow AA, Shoemaker RH, Sausville EA, Covell DG (2002) Mining the National Cancer Institute's tumor-screening database: Identification of compounds with similar cellular activities. J Med Chem 45:818–840
104. Ripley BD (2008) Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge, 416 pp
105. Roche O, Trube G, Zuegge J, Pflimlin P, Alanine A, Schneider G (2002) A virtual screening method for prediction of the hERG potassium channel liability of compound libraries. Chem Bio Chem 3:455–459
106. Rosenblatt F (1962) Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington
107. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536
108. Rumelhart DE, McClelland JL (eds) (1986) Parallel Distributed Processing. MIT Press, Cambridge, 344 pp
109. Saxena AK, Schaper K-J (2006) QSAR analysis of the time- and dose-dependent anti-inflammatory in vivo activity of substituted imidazo[1:2-*a*]pyridines using artificial neural networks. QSAR Comb Sci 25:590–597
110. Seierstad M, Agrafiotis DK (2006) A QSAR model of hERG binding using a large, diverse, and internally consistent training set. Chem Biol Drug Des 67:284–296
111. Selaru FM, Xu Y, Yin J, Zou T, Liu TC, Mori Y, Abraham JM, Sato F, Wang S, Twigg C, Olaru A, Shustova V, Leytin A, Hytiroglou P, Shibata D, Harpaz N, Meltzer SJ (2002) Artificial neural networks distinguish among subtypes of neoplastic colorectal lesions. Gastroenterology 122:606–613
112. Shen Q, Jiang J-H, Jiao C-X, Lin W-Q, Shen G-L, Yu R-Q (2004) Hybridized particle swarm algorithm for adaptive structure training of multilayer feed-forward neural network: QSAR studies of bioactivity of organic compounds. J Comput Chem 25:1726–1735
113. Shen Q, Shi W-M, Yang X-P, Ye B-X (2006) Particle swarm algorithm trained neural network for QSAR studies of inhibitors of platelet-derived growth factor receptor phosphorylation. Eur J Pharm Sci 28:369–376
114. Siu F-M, Che C-M (2006) Quantitative structure-activity (affinity) relationship (QSAR) study on protonation and cationization of α-amino acids. J Phys Chem A 110:12348–12354
115. So S-S, Karplus M (1996) Evolutionary optimization in quantitative structure-activity relationship: An application of genetic neural networks. J Med Chem 39:1521–1530
116. So S-S, Karplus M (1996) Genetic neural networks for quantitative structure-activity relationships: Improvements and application of benzodiazepine affinity for benzodiazepine/$GABA_A$ receptors. J Med Chem 39:5246–5256
117. Spycher S, Nendza M, Gasteiger J (2004) Comparison of different classification methods applied to a mode of toxic action data set. QSAR Comb Sci 23:779–791
118. Spycher S, Pellegrini E, Gasteiger J (2005) Use of structure descriptors to discriminate between modes of toxic action of phenols. J Chem Inf Model 45:200–208
119. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 21:631–643
120. Sutherland JJ, O'Brien LA, Weaver DF (2004) A comparison of methods for modeling quantitative structure-activity relationships. J Med Chem 47:5541–5554
121. Taskinen J, Yliruusi J (2003) Prediction of physicochemical properties based on neural network modelling. Adv Drug Deliv Rev 55:1163–1183
122. Terfloth L, Gasteiger J (2001) Neural networks and genetic algorithms in drug design. Drug Discov Today 6:S102–S108
123. Tetko IV, Tanchuk VY, Villa AEP (2001) Prediction of *n*-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices. J Chem Inf Comput Sci 41:1407–1421
124. Todeschini R, Consonni V (2003) Descriptors from molecular geometry. In: Gasteiger J (ed) Handbook of Chemoinformatics, vol 3. Wiley-VCH, Weinheim, pp 1004–1033

125. Trinajstić N (1992) Chemical Graph Theory. CRC Press, Boca Raton
126. Varnek A, Kireeva N, Tetko IV, Baskin II, Solov'ev VP (2007) Exhaustive QSPR studies of a large diverse set of ionic liquids: How accurately can we predict melting points? J Chem Inf Model 47:1111–1122
127. Votano JR, Parham M, Hall LH, Kier LB, Oloff S, Tropsha A, Xie Q, Tong W (2004) Three new consensus QSAR models for the prediction of Ames genotoxicity. Mutagenesis 19:365–377
128. Votano JR, Parham M, Hall LM, Hall LH, Kier LB, Oloff S, Tropsha A (2006) QSAR modeling of human serum protein binding with several modeling techniques utilizing structure-information representation. J Med Chem 49:7169–7181
129. Vracko M (2005) Kohonen artificial neural network and counter propagation neural network in molecular structure-toxicity studies. Curr Comput-Aided Drug Des 1:73–78
130. Wagner S, Hofmann A, Siedle B, Terfloth L, Merfort I, Gasteiger J (2006) Development of a structural model for NF-*K*B inhibition of sesquiterpene lactones using self-organizing neural networks. J Med Chem 49:2241–2252
131. Wan C, Harrington PB (1999) Self-configuring radial basis function neural networks for chemical pattern recognition. J Chem Inf Comput Sci 39:1049–1056
132. Wang J, Johannsen TH, Myklebost O, Hovig E (2003) Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data. BMC Bioinformatics 4:60
133. Wang Y-H, Li Y, Li Y-H, Yang S-L, Yang L (2005) Modeling K_m values using electrotopological state: Substrates for cytochrome P450 3A4-mediated metabolism. Bioorg Med Chem Lett 15:4076–4084
134. Wasserman PD (1989) Neural Computing. Van Nostrand Reinhold, New York, 230 pp
135. Weinstein JN, Kohn KW, Grever MR, Viswanadhan VN, Rubinstein LV, Monks AP, Scudiero DA, Welch L, Koutsoukos AD, Chiausa AJ, Paull KD (1992) Neural computing in cancer drug development: Predicting mechanism of action. Science 258:447–451
136. Wessel MD, Jurs PC, Tolan JW, Muskal SM (1998) Prediction of human intestinal absorption of drug compounds from molecular structure. J Chem Inf Comput Sci 38:726–735
137. Xiao Y-D, Clauset A, Harris R, Bayram E, Santago P, Schmitt JD (2005) Supervised self-organizing maps in drug discovery, 1. Robust behavior with overdetermined data sets. J Chem Inf Model 45:1749–1758
138. Xu Y, Selaru FM, Yin J, Zou TT, Shustova V, Mori Y, Sato F, Liu TC, Olaru A, Wang S, Kimos MC, Perry K, Desai K, Greenwald BD, Krasna MJ, Shibata D, Abraham JM, Meltzer SJ (2002) Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barrett's esophagus and esophageal cancer. Cancer Res 62:3493–3497
139. Xue Y, Li H, Ung CY, Yap CW, Chen YZ (2006) Classification of a diverse set of Tetrahymena pyriformis toxicity chemical compounds from molecular descriptors by statistical learning methods. Chem Res Toxicol 19:1030–1039
140. Xue Y, Yap CW, Sun LZ, Cao ZW, Wang JF, Chen YZ (2004) Prediction of P-glycoprotein substrates by a support vector machine approach. J Chem Inf Comput Sci 44:1497–1505
141. Yap CW, Li ZR, Chen YZ (2006) Quantitative structure-pharmacokinetic relationships for drug clearance by using statistical learning methods. J Mol Graph Modell 24:383–395
142. Zheng G, Xiao M, Lu XH (2005) QSAR study on the Ah receptor-binding affinities of polyhalogenated dibenzo-*p*-dioxins using net atomic-charge descriptors and a radial basis neural network. Anal Bioanal Chem 383:810–816
143. Zhou Y-P, Jiang J-H, Lin W-Q, Zou H-Y, Wu H-L, Shen G-L, Yu R-Q (2006) Adaptive configuring of radial basis function network by hybrid particle swarm algorithm for QSAR studies of organic compounds. J Chem Inf Model 46:2494–2501
144. Zupan J (2003) Neural networks. In: Gasteiger J (ed) Handbook of Chemoinformatics, vol 3. Wiley-VCH, Weinheim, pp 1167–1215
145. Zupan J, Gasteiger J (1999) Neural Networks in Chemistry and Drug Design. Wiley-VCH, Weinheim
146. Zupan J, Novič M, Gasteiger J (1995) Neural networks with counter-propagation learning-strategy used for modeling. Chemometrics Intell Lab Syst 27:175–187
147. von Korff M, Hilpert K (2006) Assessing the predictive power of unsupervised visualization techniques to improve the identification of GPCR-focused compound libraries. J Chem Inf Model 46:1580–1587

Drug Design with Machine Learning

Ovidiu Ivanciuc
Department of Biochemistry and Molecular Biology, University of Texas, Medical Branch, Galveston, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Decision Trees
Lazy Learning and *k*-Nearest Neighbors
Bayesian Methods
Support Vector Machines
Comparative Studies
Future Directions
Bibliography

Glossary

Bayesian classifier Bayes' theorem of conditional probability is a method of statistical inference that represents the basis of several classification machine learning models used in drug design and chemoinformatics to classify libraries of compounds into active and inactive chemicals. A Bayesian classifier considers each structural feature or descriptor independent of the other descriptors, and the probability that a compound is active is proportional to the ratio of active to inactive compounds that have the same structural feature or have the same value for that descriptor. The final probability that a compound is active is computed as the prod-

uct of all descriptor-based probabilities. Structural descriptors that are real numbers are usually binned prior to their evaluation with a Bayesian classifier.

Decision tree A decision tree is a sequence of rules applied to selected structural descriptors. The training phase comprises the selection of the structural descriptors that are evaluated, the order in which the rules are applied, and the decision taken at each leaf. Usually, each rules evaluates a descriptor ($\geq$ or $<$ than a threshold) and splits the objects into two or more populations. Then each population is selected and the splitting procedure is performed with a new rule, until a stopping condition is met (for example, when all objects in the population belong to the same class). The prediction phase starts from the root node and evaluates each rule on a pathway determined by the outcome (true or false) of the previous rule. When a leaf is reached the algorithm predicts the class of the object (classification trees) or the numerical value of a property (regression trees).

***k*-nearest neighbors** *k*-nearest neighbors (*k*-NN) is a supervised learning algorithm that predicts the property of an object based on a local interpolation model. In classification, the class of a new object is predicted based on the majority vote of its *k* nearest neighbors. In regression, the property value for a new object is predicted as an average value of the property values for its *k* nearest neighbors.

Lazy learning Lazy learning is a memory based local learning that defers the computation until a prediction is requested for an object. The first step is to insert the query object into the space of the training objects, and to identify the training objects located in a set neighborhood. The predicted property of the query object is based on an interpolation of the properties of the objects situated in the neighborhood.

Machine learning Machine learning is an important field of artificial intelligence, and includes a diversity of methods and algorithms that extract rules and functions from large datasets, such as decision trees, lazy learning, *k*-nearest neighbors, Bayesian methods, Gaussian processes, support vector machines, and kernel algorithms. Machine learning algorithms extract information from experimental data by computational and statistical methods and generate a set of rules, functions or procedures that allow them to predict the properties of novel objects that are not included in the learning set.

Quantitative structure-activity relationships Quantitative structure-activity relationships (QSAR) represent regression models that define quantitative correlations between the chemical structure of molecules and their physical properties (boiling point, melting point, aqueous solubility), chemical properties and reactivities (chromatographic retention, reaction rate), or biological activities (cell growth inhibition, enzyme inhibition, lethal dose). The fundamental hypotheses of QSAR are that similar chemicals have similar properties, and that small structural changes result in small changes in property values. The general form of a QSAR equation is $P(i) = f(\mathbf{SD}_i)$, where $P(i)$ is a physical, chemical, or biological property of compound i, $\mathbf{SD}_i$ is a vector of structural descriptors of i, and f is a mathematical function such as linear regression, partial least squares, artificial neural networks, or support vector machines. A QSAR model for a property P is based on a dataset of chemical compounds with known values for the property P, and a matrix of structural descriptors computed for all chemicals. The learning (training) of the QSAR model is the process of determining the optimum parameters of the regression function f. After the training phase, a QSAR model may be used to predict the property P for novel compounds that are not present in the learning set of molecules.

Support vector machines Support vector machines (SVM) are a class of supervised machine learning methods based on the structural risk minimization and the statistical learning theory of Vapnik. SVM may be applied to data classification and regression, using selected objects (support vectors) to generate the SVM model. Nonlinear classification problems are transformed into linear classification problems by using kernel functions that combine the input space into a higher-dimensional feature space in which a hyperplane may discriminate the classes. An SVM classification model computes a maximum margin hyperplane that separates the classes in the feature space. The maximal margin hyperplane maximizes the distance to the hyperplane of the closest patterns from the two classes. An SVM regression model builds a regression tube with the property that all objects inside the tube do not contribute to the overall error of the model. The shape of the regression tube is determined by selected objects (support vectors) situated outside the tube.

Structural descriptor A structural descriptor (SD) is a numerical value computed from the chemical structure of a molecule, which is invariant to the numbering of the atoms in the molecule. Structural descriptors may be classified as constitutional (counts of molecular fragments, such as rings, functional groups, or atom pairs), topological indices (computed from the molec-

ular graph), geometrical (volume, surface, charged-surface), quantum (atomic charges, energies of molecular orbitals), and molecular field (such as those used in CoMFA, CoMSIA, or CoRSA).

Structure-activity relationships Structure-activity relationships (SAR) represent classification models that can discriminate between sets of chemicals that belong to different classes of biological activities, usually active/inactive towards a certain biological receptor. The general form of a SAR equation is $C(i) = f(\mathbf{SD}_i)$, where $C(i)$ is the activity class of compound i (active/inactive, inhibitor/non-inhibitor, ligand/non-ligand), $\mathbf{SD}_i$ is a vector of structural descriptors of i, and f is a classification function such as k-nearest neighbors, linear discriminant analysis, random trees, random forests, Bayesian networks, artificial neural networks, or support vector machines.

Definition of the Subject

The process of drug discovery has the goal to identify lead chemicals that have a significant activity against a selected biological target. A disease state may be the result of changes in the structure and function of cell-signaling receptors, enzymes, hormone receptors, or other functional proteins. The drug target is a protein whose activity is modulated by its interaction with a chemical compound, and thus may control a disease. The lead compounds identified in the drug discovery step are optimized in the drug development phase that results in a small number of chemicals that are evaluated in human clinical trials. The first priority in drug development is to increase the biological activity of a lead compound while preserving its drug-like properties. The lead compound is expanded into a chemical library that conserves the structure responsible for the biological activity (pharmacophore) and adds chemical groups that might improve its activity. Then the chemicals are synthesized and tested in biological assays against the target, which may result in the identification of more active compounds. The cycle "library design – chemical synthesis – biological assay" is repeated several times until a number of good candidates are identified for clinical trials. The design of effective drug candidates is a multi-objective optimization problem, because simultaneous with a good biological activity, the chemicals must pass several other important tests, including pharmacokinetics, pharmacodynamics, toxicity, mutagenicity, metabolism, and excretion.

Computer-assisted drug design (CADD) uses computational chemistry to increase the chances of finding valuable drug candidates. CADD methods include machine learning, structure-activity relationships (SAR), quantitative structure-activity relationships (QSAR), molecular mechanics, quantum mechanics, molecular dynamics, and drug-protein docking. SAR and QSAR are based on the theory that chemical structure determines all physical, chemical and biological properties of a molecule. To obtain structure-activity models, the chemical structure is characterized with structural descriptors and then ML models are used to identify a statistical relationship between the descriptor space and a molecular property. Other considerations for the application of SAR and QSAR in drug design are that similar molecules have similar properties, and that a small modification in a chemical structure result in a small modification in its properties.

Machine learning (ML) procedures are applied in drug discovery in an iterative way, in which experimental activity data are used to train ML models which in turn offer predictions for novel molecules with improved biological properties. Then the molecules designed with ML are tested in biological assays, and the best candidates are selected for further cycles of optimization. ML models may greatly improve the chances of finding lead molecules by screening a larger diversity of chemical topologies. In the classical approach, a number of chemicals in the range 10^3–10^4 is synthesized and assayed to identify lead compounds, whereas the introduction of combinatorial chemistry and high-throughput screening (HTS) increased these numbers to 10^5–10^6. Although the robots offer great speed and reliability in HTS, the chemical compounds still have to be synthesized before screening, thus limiting the diversity of structures evaluated in HTS. This is why ML models are used to enhance the lead identification process with a virtual HTS (vHTS) screening of 10^7–10^8 molecules. Based on experimental results from a HTS campaign, ML models are computed and then used in vHTS experiments. These ML are mainly classification models that predict molecules that have a high probability of interacting with the selected target. vHTS has several obvious advantages compared to HTS, mainly coming from the fact that the chemical compounds are generated only in a computer, which opens the possibility to explore a larger diversity of chemical skeletons and a much higher number of molecules.

Classification (SAR) and regression (QSAR) ML models are applied during the drug development cycles to optimize the biological activity, target selectivity, and other physico-chemical and biological properties of selected chemicals. ML models are essential tools in increasing the chances of bringing a drug to market, but only when integrated with the usual chemical, biological, pharmacological and clinical procedures used by the pharmaceuti-

cal industry. Boid enumerated several successful CADD applications in which computational methods had a decisive contribution to the discovery of a drug, namely norfloxacin (Merck, antibacterial), losartan (Merck, antihypertensive), dorzolamide (Merck, antiglaucoma), ritonavir (Abbott, antiviral), indinavir (Merck, antiviral), donepezil (Esai, anti-Alzheimer's disease), zolmitriptan (AstraZeneca, antimigraine), nelfinavir (Pfizer, antiviral), amprenavir (GlaxoSmithKline, antiviral), zanamivir (GlaxoSmithKline, antiviral), oseltamivir (Roche, antiviral), lopinavir (Abbott, antiviral), imatinib (Novartis, antineoplastic), erlotinib (OSI, antineoplastic) [17]. This is an impressive list that adds convincing evidence to advocate the integration of machine learning and other chemoinformatics methods in drug discovery and development.

Introduction

Machine learning is an important field of artificial intelligence, and includes a diversity of methods and algorithms that extract rules and functions from large datasets, such as linear discriminant analysis (LDA), artificial neural networks (ANN), decision trees, lazy learning, *k*-nearest neighbors, Bayesian methods, Gaussian processes, support vector machines (SVM), and kernel algorithms. Several influential books are recommended for a detailed overview of ML algorithms: *Pattern Recognition and Neural Networks* by Ripley [110], *Machine Learning and Data Mining* by Kononenko and Kukar [83], *Pattern Recognition and Machine Learning* by Bishop [15], *Data Mining: Practical Machine Learning Tools and Techniques* by Witten and Frank [151], *Introduction to Machine Learning* by Alpaydin [4], *The Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman [52], *Pattern Classification* by Duda, Hart, and Stork [35], *Machine Learning* by Mitchell [96], and *Neural Networks for Pattern Recognition* by Bishop [16].

ML algorithms extract information from experimental data by computational and statistical methods and generate a set of rules, functions or procedures that allow them to predict the properties of novel objects that are not included in the learning set. In drug design the task is to learn how chemical structure determines some important drug property, such as physico-chemical properties (aqueous solubility, skin penetration, hydrophobicity, intestinal absorption, blood-brain barrier penetration), biological activity (enzyme inhibition), metabolism, toxicity, mutagenicity, or excretion. Each molecule is associated with a set of structural descriptors and an experimental property. The structural descriptors are used as input for an ML algorithm and the property is the target (output) of the model. A structural descriptor is a structural feature of the chemical structure, and can be a list of substructures or substituents, graph descriptors [19,64,136], topological indices [8,18,65], connectivity indices [74,75] electrotopological indices [76], or indices derived from the molecular geometry and quantum calculations [73,133]. The experimental property that is modeled may be a class label ($+1/-1$), such as soluble/non-soluble, inhibitor/non-inhibitor, ligand/non-ligand, toxic/non-toxic, mutagen/non-mutagen, carcinogen/non-carcinogen. For this type of data one obtains a classification model or SAR. Classification models are mainly used to filter large chemical libraries and reduce them to a small number of chemicals with a high probability of having a desired property (ligand for a biological target, inhibitor for an enzyme, non-toxic, non-mutagen, or non-carcinogen). The experimental property may be also a continuous value, such as inhibition constant to an enzyme, binding constant to a target, hydrophobicity, aqueous solubility, skin penetration, or intestinal absorption, in which case one computes a regression model or QSAR. Regression models, which may be linear or nonlinear, are used to optimize the drug-like properties of a chemical compound. A drug-related property may be considered categorical variable or a continuous variable, depending on the drug development phase in which is applied. For example, in developing CNS (central nervous system) drugs, the blood-brain barrier penetration is used as a categorical variable in the early stages of drug discovery, to identify and eliminate compounds that do not pass the blood-brain barrier. In later stages of drug development this property is used as a continuous variable in QSAR models to optimize the brain concentration of a small number of selected chemicals that show promising pharmacological properties. Both SAR and QSAR models belong to the class of supervised learning algorithms, because the target (output) value is provided together with the input chemical structures and descriptors. Supervised learning algorithms have two distinct phases. The first one is learning or training, in which a ML method is used to learn the relationships between input (structural descriptors) and output (an experimental property provided for each molecule). The result is a model that can be a statistical function or a set of rules. The second phase of a ML algorithm is the prediction, when the trained model is used to predict the property for novel chemicals that were not present in the training set or that are not even synthesized. The predictions obtained are then used as a guide in synthesizing and testing novel chemicals.

In unsupervised learning the dataset contains only chemical structures and structural descriptors, without

output values. The ML objective is to identify how molecules form clusters based on their structural similarity. Obviously, starting from a particular set of structural descriptors, different similarity indices and different clustering algorithms will result in different clusters. Clustering algorithms are applied in drug design to identify similar molecules in chemical libraries. For example, starting from a lead compound one can find similar compounds in catalogs and databases, and thus purchase chemicals that might be useful in the drug discovery process. Another application is to identify which chemical structures are under-represented or missing from a company collection, and to guide compound synthesis and acquisition.

To illustrate two basic approaches in data classification we consider the objects from Fig. 1a. The two classes of objects may represent two populations of molecules, for example enzyme inhibitors form class +1, whereas non-inhibitors form class −1. From the distribution of the two classes one can identify a cluster for class +1 and a distinct cluster for class −1. Such a clear-cut situation is not the norm in drug design application, because classes of molecules may overlap which makes more difficult the property prediction for novel chemicals. The problem considered in Fig. 1a is to use the information provided by the two classes of objects to predict the class of the unknown object marked with "?". The first ML tested is k-nearest neighbors (k-NN), that belongs to the class of lazy learning algorithms, and performs a local approximation of the model that is computed only in the moment of the prediction. An object is classified in the most populated class among its k nearest neighbors, and for two-class problems k must be odd to avoid undecided situations. We test here the simplest variant, 1-nearest neighbor, which considers only the closest neighbor. An inspection of the distances between the prediction object • and all other objects shows that the closest object belongs to class −1, and thus the prediction object is assigned to class −1 (Fig. 1b). The predicted class may change if one considers a larger number of nearest neighbors, and the optimum value for k is usually determined through cross-validation.

An efficient procedure to classify linearly separable classes is represented by a hyperplane H (Fig. 1c) that separates the descriptor space into a region for class +1 and another region for class −1. A new object that is located in the space region that belongs to class +1 will be assigned to class +1 irrespective of its distance to the separating hyperplane. Similarly, if the new object is situated in the space region that belongs to class −1, then the object is predicted in class −1. The unknown object from Fig. 1a is assigned to class +1 by the hyperplane H (Fig. 1c). For the same classification problem we obtained two different predictions for the unknown object, namely class −1 with 1-NN and class +1 with a hyperplane. Such divergent predictions obtained with different ML methods are quite common, and it is not unusual to obtain different predictions for the same object just by changing the ML parameters. These practical aspects of ML should receive proper consideration, and it is a good practice to evaluate a large diversity of ML methods to obtain consensus predictions.

The two ML demonstrated here are very different in their formulation and their properties. k-NN can be used for multi-class discrimination and it is a nonlinear classifier that is useful for classes that cannot be separated with a linear hypersurface. Also, k-NN does not produce a statistical function to replace the data, and the learning phase is missing. The second classifier replaces the data with the equation of a hyperplane (Fig. 1d). This function may be used only for two-class problems in which the objects may be separated by linear hypersurface. A hyperplane H has the formula $\mathbf{w} \cdot \mathbf{x} + b = 0$, where $\mathbf{w}$ is the normal vector to H. A molecule i characterized by a vector of structural descriptors $\mathbf{x}_i$ belongs to class +1 if $\mathbf{w} \cdot \mathbf{x}_i + b > 0$, or to class −1 if $\mathbf{w} \cdot \mathbf{x}_i + b < 0$. The rules for predicting the class for an unknown object k are:

$$\text{class}(k) = \begin{cases} +1 & \text{if } \boldsymbol{w} \cdot \boldsymbol{x}_k + b > 0 \\ -1 & \text{if } \boldsymbol{w} \cdot \boldsymbol{x}_k + b < 0\,. \end{cases} \tag{1}$$

The hyperplane classifier may be extended with kernel functions to nonlinear classification, as it will be shown in Sect. "Support Vector Machines" in which we give an overview of support vector machines and their applications in structure-activity models.

A decision tree performs a mapping of the structural descriptors to a conclusion about a property of the object. The learning process establishes a series of successive rules that are based on the numerical values of a subset of descriptors. Rules are applied in a set sequence and at each branching point a rule is evaluated and its outcome (true or false) determines which branch is selected in the next step. Starting from the root node, the rules are evaluated and depending on the actual numerical values of the descriptors the pathway ends on a leaf that assigns a class label or a value to the investigated property. The class of decision tree algorithms includes the C4.5 decision tree [108], the NBTree decision tree with naïve Bayes classifiers at the leaves [82], the alternating decision tree ADTree [42], random trees and random forests [22]. Decision trees are computationally very efficient, and their predictions are usually better than those obtained with more complicated ML algorithms. Drug design applications of decision trees include diverse

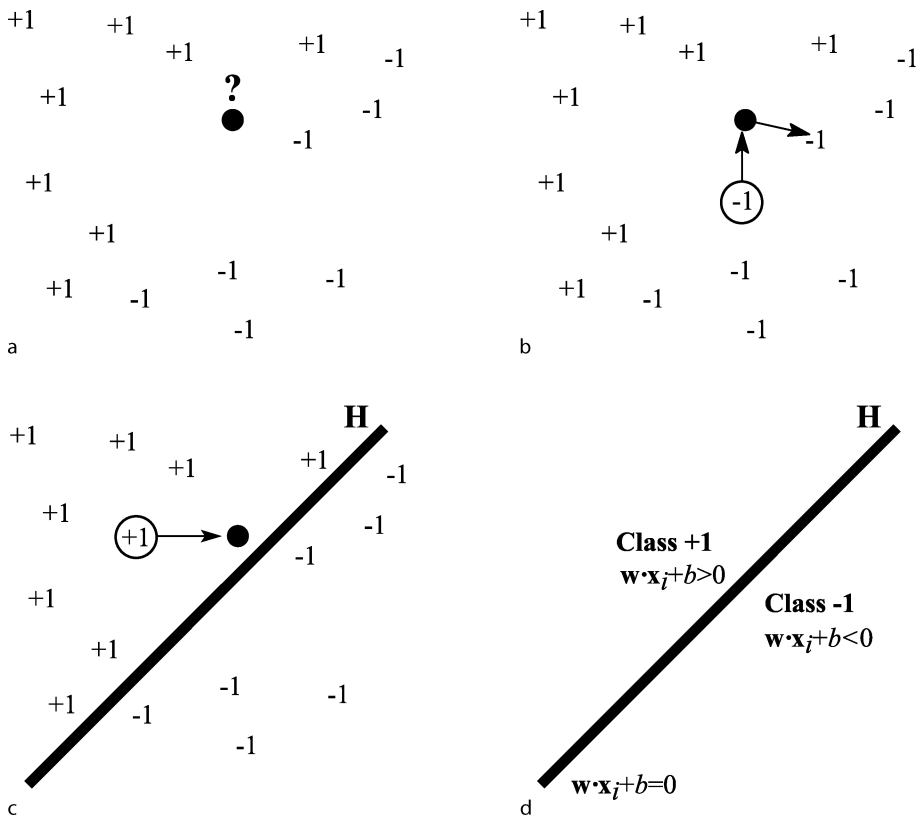

Drug Design with Machine Learning, Figure 1
Classification with machine learning: a a class prediction (+ 1/—1) is sought for the object •; b a *k*-nearest neighbor classifier (*k* = 1) predicts the object • in class —1; c a linear classifier defined by the hyperplane *H* predicts the object • in class + 1; d the classification hyperplane *H* separates the space in region for class + 1 and another region for class —1

applications, such as prediction of drugs that cross the blood-brain barrier [30], structure-activity relationships for estrogen receptor ligands [135], prediction of *P*-glycoprotein substrates [91], evaluation of the cytochrome P450 metabolism [154], identification of drug-like chemicals [114], design of target chemical libraries [31], prediction of protein-protein interaction [98], and modeling the oral absorption of drugs [57]. A review of these studies is presented in Sect. "Decision Trees".

Lazy learning is a memory-based local learning method that stores the training objects and performs all computations only when a prediction request is received by the system [5,6]. The prediction is performed by inserting the unknown object into the prediction space and then identifying all objects situated in its neighborhood. Each prediction is made by computing a local model from the neighboring objects. Lazy learning may be used both for classification and regression, with predictions based on local interpolations of selected objects according to a distance measure. Local learning may perform better than global learning, as shown in a number of SAR and QSAR studies that use various techniques of lazy learning [47,85,159].

The *k*-nearest neighbors (*k*-NN) algorithm is a lazy learning method that can be used for classification and regression [1]. The training set of objects is stored in the memory, and then each prediction is made by computing a local interpolation model. The prediction object is added to the descriptor space in which the training objects are placed and then its *k* nearest neighbors are identified. For classification, the class of the prediction object is assigned to class of the majority of its neighbors. Ties are avoided by selected an odd value for *k*, and the best value for *k* is usually determined by cross-validation. For regression, the predicted property for the query object is an average value of the property values for its *k* nearest neighbors. A distance weighting may be added in such a way to increase the contribution of closer objects and to decrease the influence of more distant objects. Another lazy learning algorithm is K^* which uses an entropy-

based distance function [26]. Basak was an early proponent of the *k*-NN regression for physico-chemical and biological properties [10,11,49]. The method was further refined by Tropsha by optimizing the descriptor selection, *k*, and the distance weighing function [60,160,161]. Other *k*-NN applications were proposed for three-dimensional QSAR [2] and for ligand-based virtual screening of chemical libraries [54]. In Sect. "Lazy Learning and *k*-Nearest Neighbors" we review SAR and QSAR models computed with lazy learning and *k*-nearest neighbors.

Bayes' theorem provides mathematical tools that explain how the probability that a theory is true is affected by a new piece of evidence [12]. Several Bayesian classifiers [71,151] were proposed to estimate the probability that a chemical is active based on a set of descriptors and a chemical library with known class attributes (active/inactive). Bayesian classifiers consider that each descriptor is statistically independent of all other descriptors. Binary descriptors, structural counts, fingerprints and other descriptors that have integer values are directly evaluated with Bayesian classifiers, whereas real number descriptors are discretized and transformed into an array of bins. The probability that a compound is active is proportional to the ratio of active to inactive compounds that have the same structural feature or have the same value for that descriptor. A final probability is computed as the product of all descriptor-based probabilities. Bayesian classifiers have numerous applications in chemoinformatics and structure-activity models, such as for the prediction of multidrug resistance reversal [125], to estimate the phospholipidosis inducing potential [105], to identify chemical compounds similar to natural products [39], to improve high-throughput docking [78,79], and to screen virtual chemical libraries [77]. Drug design applications of Bayesian methods are presented in Sect. "Bayesian Methods".

Support vector machines belong to a heterogeneous group of machine learning methods that use kernels to solve efficiently high-dimensional nonlinear problems. SVM extend the generalized portrait algorithm developed by Vapnik [144] by using elements of statistical learning theory [143] that describe the ML properties that guarantee dependable predictions. Vapnik elaborated further the statistical learning theory in three more recent books, *Estimation of Dependencies Based on Empirical Data* [138], *The Nature of Statistical Learning Theory* [139], and *Statistical Learning Theory* [140]. Vapnik and co-workers developed the current formulation of the SVM algorithm at AT&T Bell Laboratories [20,25,28,33,50,116,141,142].

SVM models have several interesting and appealing properties, namely the maximum margin classification, the kernel transformation of the input space into a feature space where a hyperplane may separate the classes, and a unique solution. The SVM introduction generated an enormous interest, comparable only with that produced by the development of artificial neural networks. Many applications were investigated in a short period of time, simultaneous with developments of novel SVM algorithms, such as least squares SVM [126,127] and other kernel algorithms [21,58,119]. The theory and applications of SVM are presented in a number of books, including *Learning with Kernels* by Schölkopf and Smola [115], *Learning Kernel Classifiers* by Herbrich [53], *An Introduction to Support Vector Machines* by Cristianini and Shawe-Taylor [29], and *Advances in Kernel Methods: Support Vector Learning* by Schölkopf, Burges, and Smola [117]. SVM have numerous applications in drug design, from screening of chemical libraries to QSAR [70], with accurate predictions that frequently surpass those obtained with more established algorithms. In Sect. "Support Vector Machines" we review the most important aspects of SVM models and we present several applications to drug design, SAR and QSAR.

With an ever-increasing number of machine learning algorithms available, it is difficult to select those methods that have better chances of solving a particular SAR or QSAR problem. To address this topic, we review in Sect. "Comparative Studies" several papers that compare diverse ML models for drug design and structure-activity relationships. A comparison of 21 machine learning algorithms is presented for data selected from the National Cancer Institute 60-cell line screening panel (NCI-60) [69]. The structure-anticancer activity models are developed for three cell lines, namely for lung large cell carcinoma NCI-H460, glioma SF-268, and melanoma SK-MEL-5. Finally, we present an assessment of the results obtained in the machine learning competition CoEPrA 2006 (Comparative Evaluation of Prediction Algorithms, http://www.coepra.org/).

Decision Trees

A decision tree represents a series of rules that perform a mapping of the structural descriptors to a prediction for a molecular property. The class of decision tree algorithms includes the C4.5 decision tree [108], the NBTree decision tree with naïve Bayes classifiers at the leaves [82], the alternating decision tree ADTree [42], random trees and random forests [22]. A typical decision tree algorithm, such as C4.5 and related methods, generates a sequence of rules based on splitting the objects into two subgroups based on a selected structural descriptor. A threshold of the descriptor separates the objects into a subgroup of objects

that have a descriptor value lower than the threshold, and another subgroup of objects that have a descriptor value higher than the threshold. The descriptor and the threshold are selected such as to maximize the difference in entropy, which maximizes the separation of objects based on their class, i. e., one subgroup contains mainly object from one class whereas the second subgroup contains mainly objects from the other class. The process is repeated until a proper separation in classes is obtained for all objects.

The structure of a C4.5 decision tree is demonstrated for the classification of narcotic pollutants into polar chemicals (class 1) and nonpolar chemicals (class 2) [66,109,137]. The dataset consists of 190 compounds, from which 76 are polar and 114 are nonpolar. The structural descriptors are the octanol-water partition coefficient $\log K_{ow}$, the energy of the highest occupied molecular orbital E_{HOMO}, the energy of the lowest unoccupied molecular orbital E_{LUMO}, the most negative partial charge on any non-hydrogen atom in the molecule $Q-$, and the most positive partial charge on a hydrogen atom $Q+$. The J48 (C4.5) decision trees computed with Weka [41,151] shows that only two descriptors, namely $Q-$ and E_{LUMO}, can separate the leaning dataset (Fig. 2a). The first rule of this decision tree separates molecules based on their $Q-$ value, namely those with $Q- \leq -0.475$ are sent to the second rule whereas those with $Q- > -0.475$ are classified in class 2. The second rule separates chemicals according to their E_{LUMO} value, namely those with $E_{LUMO} \leq 0.723$ are classified in class 1 whereas those with $E_{LUMO} > 0.723$ are classified in class 2.

The second example considers a J48 (C4.5) decision trees with four classes (Fig. 2b) representing four modes of toxic action of phenols in the *Tetrahymena pyriformis* assay, namely polar narcotics (class 1), oxidative uncouplers (class 2), proelectrophiles (class 3), and soft electrophiles (class 4). The dataset consists of 220 chemicals and seven structural descriptors, namely $\log K_{ow}$, E_{HOMO}, E_{LUMO}, the maximum donor (electrophilic) delocalizability for C atoms D_C-maxE, the average acceptor (nucleophilic) delocalizability for C atoms D_C-avN, the hydrogen bond donor count N_Hdon, and the hydrogen bond acceptor count N_Hacc. Out of the seven descriptors, only $\log K_{ow}$ is not selected to form a splitting rule, whereas E_{HOMO} and N_Hacc are selected in two rules each. Decision trees are efficient ML algorithms that provide accurate and fast predictions, with the added benefit of performing also a selection of the most useful descriptors.

Combinatorial chemistry and high-throughput screening accelerate the drug discovery process by producing vast quantities of biological assay results. The data mining of all these chemical and biological experiments requires faster structure-activity algorithms, as shown by Rusinko et al for a library of monoamine oxidase inhibitors that was evaluated with recursive partitioning [112]. The advantage of RP is that it scales linearly with the number of descriptors, and it can be used for datasets containing more than 10^5 chemicals and 2×10^6 structural descriptors. To generate chemical libraries focused on a specific target, Deng et al developed the structural interaction fingerprints, a class of descriptors that encode ligand-target binding interactions [31]. These fingerprints were used in a decision tree model to identify ligands for MAP kinase p38, showing that the method can be used with success for the structure-based screening of combinatorial chemical libraries. Statistical studies of FDA approved drugs show that there are certain physico-chemical properties that define drug-like structural requirements, independent of the drug targets. Drug-like filters are used to eliminate at early stages those compounds that do not have the general structural features of a drug. Schneider et al applied decision trees to the classification of drug-like compounds based on drug-like indices, hydrophobicity, and molar refractivity [114]. This simple filter correctly identifies 83% of drugs and 39% of the non-drugs. The classification trees suggest several properties that separate drugs from non-drugs, such as a molecular weight higher than 230, a molar refractivity higher than 40, as well as the presence of structural elements such as rings and functional groups. Natural products are frequently screened to identify active compounds for selected targets. A database of 240 Chinese herbs containing 8264 compounds was screened with a random forest algorithm to identify inhibitors for several targets, including cyclooxygenases, lipoxygenases, aldose reductase, and three HIV targets [36]. The screening results show that random forests give dependable predictions even for unbalanced libraries in which inactive compounds are more numerous than active compounds. A literature search confirmed the computational predictions for 83 inhibitors identified in Chinese herbs.

Torsade de pointes (TdP) is a polymorphic ventricular arrhythmia that may be caused by drugs that induce the prolongation of the QT interval by inhibiting the heart potassium channel hERG (human ether-á-go-go). Gepp and Hutter investigated the TdP potential of 339 drugs, and found that a decision tree has a success rate of up to 80% in predicting the correct class [46]. Ekins et al combined a recursive partitioning (RP) tree with Kohonen and Sammon maps in order to identify hERG inhibitors [38]. RP is used as a fast and accurate filter to screen and prioritize drug databases for an in-depth assessment with Ko-

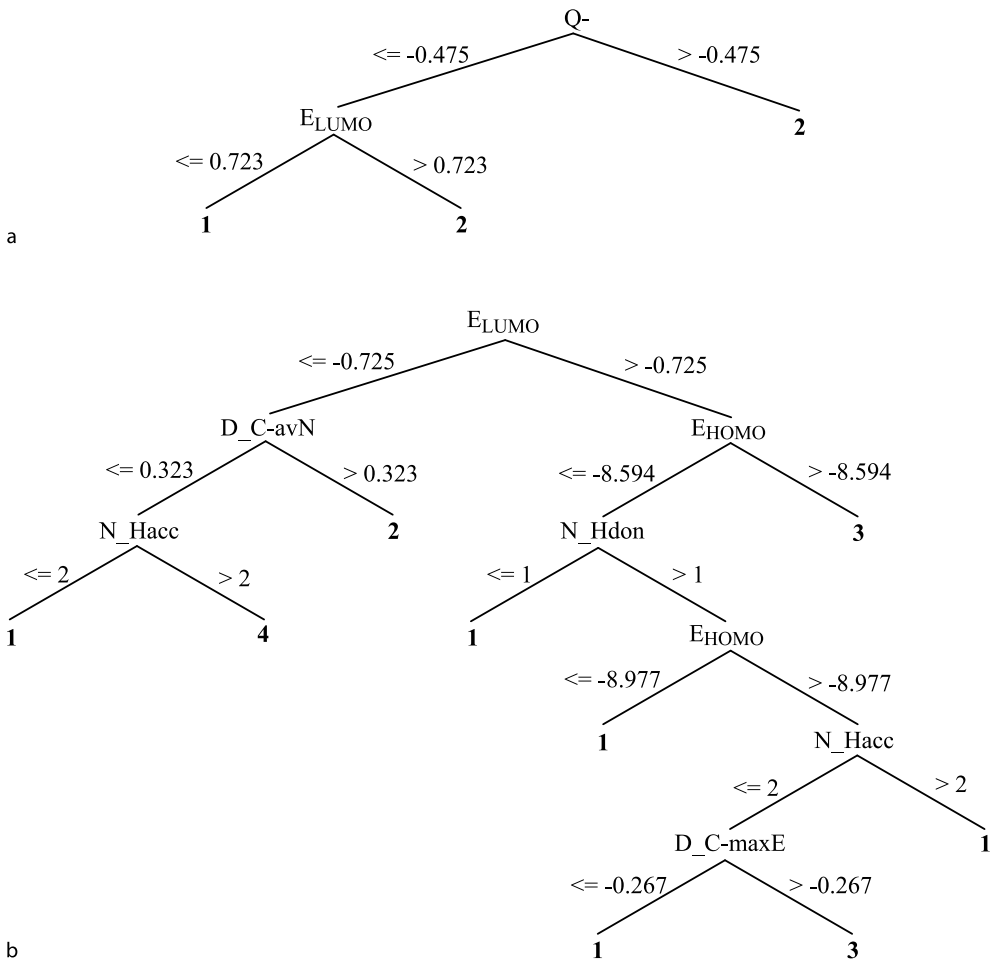

Drug Design with Machine Learning, Figure 2

Classification with J48 (C4.5) decision trees: a polar (class 1) and nonpolar (class 2) narcotic pollutants; b toxicity modes of phenols against the *Tetrahymena pyriformis*, namely polar narcotics (class 1), oxidative uncouplers (class 2), proelectrophiles (class 3), and soft electrophiles (class 4)

honen and Sammon maps. The human intestinal absorption of potential drugs may be investigated with in silico methods that assess the intestinal permeability of a chemical compound before synthesis. Hou et al used recursive partitioning to classify the intestinal absorption (poor or good) of chemical compounds [57]. The SAR model based on decision tree classification has good results for a training set of 481 compounds (95.9% success rate for the poor absorption class and 96.1% for the good absorption class) and a prediction set of 98 chemicals (100% success rate for the poor absorption class and 96.8% for the good absorption class).

All drugs that have a CNS (central nervous system) action must penetrate the blood-brain barrier (BBB), and BBB permeability models are used to filter compounds that cannot pass the barrier. Andres and Hutter found that a decision tree provides fast and accurate results, with an accuracy of 96% for 186 training compounds and 84% for 38 test chemicals [3]. The drugs that pass the BBB are predicted with a higher accuracy (94%) than those that do not pass the BBB (89%). Deconinck et al predicted BBB permeability with classification and regression trees (CART), alone and aggregated in a boosting approach [30]. The training was performed for 147 drugs, and the structural descriptors were computed with Dragon [133]. The percentage of correctly classified drugs in cross-validation is 83.6% for a single tree, and increases to 94% for a boosting model that contains 150 trees. The single tree model may be used for a fast screening of large libraries, but for more reliable predictions a boosting model should be preferred.

The drug affinity for the cytochrome P450 2C9 was predicted with a consensus method based on four ML al-

gorithms [59]. Two of these algorithms are variants of recursive partitioning, the third is a local interpolation method, and the fourth is a subgraph search method. These four models were trained with a set of 276 compounds, and then the SAR models were tested for a prediction set of 50 chemicals. The consensus predictions have an accuracy of 94%, demonstrating that the computational methods may predict which drugs are metabolized by P450 2C9. The metabolic stability of 161 drugs against six isoforms of the cytochrome P450 (1A2, 2C9, 2C19, 2D6, 2E1, and 3A4) was evaluated with a classification tree [154]. The SAR models identified several structural characteristics that determine the P450 specificity, namely 3A4 substrates are larger compounds, 2E1 are smaller molecules, 2C9 substrates are anionic, and 2D6 substrates are cationic.

P-glycoprotein (Pgp) mediated drug efflux is responsible for the low cellular accumulation of anticancer drugs, for reduced oral absorption, and for low blood-brain barrier penetration. Pgp affects also the hepatic, renal, or intestinal elimination of drugs. Li el al. proposed a decision tree model to differentiate Pgp substrates from non-substrates [91]. Four-point 3D pharmacophores obtained from the three-dimensional structure of 163 chemical compounds were subsequently used as descriptors for the classification tree. Nine pharmacophores were selected in the decision tree, giving an accuracy of 87.7% for the training set and 87.6% for the prediction set. These pharmacophores highlight the most important structural features of Pgp substrates, and can be used as a fast filter for chemical libraries. The renal clearance of 130 drugs was investigated with Molconn-Z topological indices [76] and recursive partitioning [32]. RP results for the separation of high-clearance compounds from low-clearance compounds show 88% correct predictions for the training set of 130 compounds and 75% correct predictions for a prediction set of 20 compounds.

The prediction power of classification and regression algorithms can be significantly improved by using an ensemble of models that are combined into a final prediction by an averaging or a voting procedure. Such ML algorithms are called ensemble, committee, or jury methods. Tong et al proposed an ensemble method, decision forest (DF), which is based on classification and regression trees [134]. A DF is obtained in four steps: (a) generate a CART; (b) generate a second CART based only on structural descriptors that were not used in the first CART; (c) repeat the first two steps until no more trees can be generated; (d) predict a property of a chemical compound based on all trees generated. DF was used to predict the estrogen receptor binding activity of organic compounds from Molconn-Z topological indices [135]. After 2000 simulations of a 10-fold cross-validation, the predicted accuracy for a dataset of 232 compounds was 81.9%, and the accuracy for another dataset of 1092 chemicals was 79.7%. The DF models can be used to identify potential endocrine disruptors, which are chemicals that affect the endocrine functions. A similar algorithm is the random forest (RF) [22], which was extensively tested for several structure-activity studies [128]. RF has three properties that are important in SAR, namely the algorithm evaluates the importance of descriptors, identifies similar compounds, and measures the prediction performance. The skin sensitization activity of 131 chemicals was modeled with RF, using Dragon and Molconn-Z indices [90]. Compared with single tree models, RF improve significantly the ability to identify chemical allergens.

Lazy Learning and *k*-Nearest Neighbors

Lazy learning is a local learning method that keeps in the memory the entire set of training objects, and computed an interpolation model based on the neighborhood on the prediction object Atkeson, [56]. The prevalent lazy learning algorithm in SAR and QSAR is the *k*-nearest neighbors [1]. A *k*-NN prediction is based on a descriptor space, which may be the entire set of structural descriptors evaluated in a study or a subset that gives better predictions, a distance function, and a neighborhood k of the prediction object. The *k*-NN classification may be applied to two or more classes (Fig. 3a). The prediction object, shown here as a black circle with an unknown class "?", is inserted into the descriptor space and then its k nearest neighbors are identified. The class of the prediction object is the class of the majority of its neighbors. In the example from Fig. 3a, the predicted class is -1 if $k = 3$, and $+1$ if $k = 5$. The prediction statistics of the *k*-NN classifier depend on the value of k, and its optimum value is usually determined by cross-validation.

A similar procedure is used for *k*-NN regression (Fig. 3b). For regression, each training object has a real number as property value, and the predicted property for the query object is an average value of the property values for its k nearest neighbors. For the situation depicted in Fig. 3b, the predicted value for $k = 3$ is $(6+6+7)/3 = 6.3$, whereas the prediction for $k = 5$ is 5.6. A distance weight is usually added to increase the contribution of closer objects and to decrease the influence of more distant objects. For a distance function d, a possible weighting of the property values is the inverse distance $1/d$.

Lazy learning applications in drug design are still rare, mainly because the community of practitioners only re-

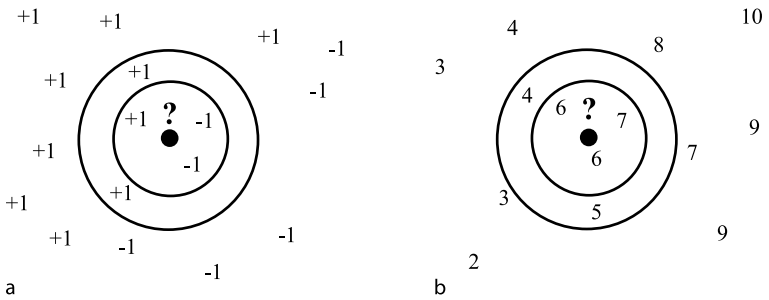

Drug Design with Machine Learning, Figure 3
Applications of the *k*-NN algorithm to a classification and b regression

cently discovered the rich field of machine learning algorithms. Kumar et al used locally linear embedding for nonlinear dimensionality reduction coupled with lazy learning for two QSAR benchmarks, the Selwood dataset and the steroids dataset originally analyzed with CoMFA [85]. Guha et al applied local lazy regression to three QSAR models, namely to 179 artemisinin analogues with antimalarial activity, to 79 platelet-derived growth factor inhibitors, and to 756 inhibitors of dihydrofolate reductase [47]. In all cases, the local models give better predictions compared to global regression models. An automated lazy learning quantitative structure-activity relationship (ALL-QSAR) approach was developed based on locally weighted linear regression models computed from the compounds in training set that are chemically most similar to a test compound [159]. ALL-QSAR was applied with good results for three datasets, namely 48 anticonvulsant compounds, 48 dopamine receptor antagonists, and to model the toxicity of 250 phenols against *Tetrahymena pyriformis*. Sommer and Kramer reported several applications of lazy learning to chemical datasets with a high structural diversity [123]. Collections of noncongeneric compounds are usually difficult to model, but their approach obtained good results for several datasets, namely 739 mutagenic chemicals from the Carcinogenic Potency Database, 342 chemicals tested for biodegradability, and 1481 from the AIDS Antiviral Screen of the NCI Developmental Therapeutics Program.

The fundamental assumptions in *k*-NN applications to SAR and QSAR are that molecules with similar structures are close in the descriptor space, and that similar molecules have similar properties. Basak advocated the use of *k*-NN in QSAR as an alternative to the multiple linear regression [10,11]. In this approach, the chemical compounds are represented with a collection of topological indices, and the Euclidean distance is used to identify the structures most similar with the test molecule. This *k*-NN method was applied with success to model the hydrophobicity of 4067 compounds [10], the mutagenicity of a diverse set of 520 chemicals [11], the boiling temperature of 139 hydrocarbons and the mutagenicity of 95 aromatic and heteroaromatic amines [9]. Good models are obtained with optimized *k* values over a range between 1 and 25, as shown in a QSAR model for vapor pressure [48]. The descriptors selected in the structural space are essential in establishing a suitable similarity measure. General collections of descriptors usually give good predictions, but property tailored structural spaces can increase significantly the predictivity of the model. Tailored structural spaces include only those descriptors that contribute significantly to the model, as demonstrated for hydrophobicity and boiling temperature [49]. A structural space consisting of electrotopological state indices gives reliable predictions for hydrophobicity, Henry's law constant, vapor pressure, and OH-radical bimolecular rate constant [24].

Tropsha applied *k*-NN to numerous QSAR models, demonstrating its value in comparisons with more established algorithms. For a dataset of 29 dopamine antagonists, *k*-NN predictions are much better than those obtained with CoMFA, which represents the standard 3D QSAR model [55]. *k*-NN may be used also as a robust method to select those structural descriptors that are important in modeling a particular biological property, as shown for 58 estrogen receptor ligands [161] and for 48 amino acid derivatives with anticonvulsant activity [120]. In a more elaborate implementation, *k*-NN is used to select descriptors, to find the best value for *k*, and to optimize the function that weights the contribution of its neighbors [60]. The anticancer activity of 157 epipodophyllotoxin derivatives was modeled with molecular connectivity indices and *k*-NN [152]. In a comparison with a CoMFA model obtained for the same dataset it was found that *k*-NN gives better predictions in a cross-validation test. The metabolic stability of 631 drug candidates was evaluated with *k*-NN and a collection of structural descriptors that included molecular connectivity indices and atom

pairs [121]. The prediction in a test set of compounds had accuracy higher than 85%, whereas an external validation set of 107 chemicals was predicted with a 83% success rate.

k-NN is also used as a component in more elaborate drug design algorithms. Ajmani et al combined molecular field analysis and *k*-NN to obtain a robust QSAR model that was tested for three datasets, namely the CoMFA steroids, cyclooxygenase-2 inhibitors, and anticancer compounds [2]. Another QSAR model that incorporates *k*-NN is Complimentary Ligands Based on Receptor Information (CoLiBRI) [103]. CoLiBRI represents both ligands and receptor binding sites in the same space of universal chemical descriptors. In a test consisting of 800 X-ray protein-ligand receptors from PDB, CoLiBRI ranked the correct ligands in the top 1% chemicals selected, and was able to quickly eliminate improbable ligands. The skin permeability coefficients of 110 compounds were predicted with an ensemble consisting of a collection of *k*-NN and ridge regression models [99].

The melting temperature of organic compounds is particularly difficult to model, due mainly to a lack of good descriptors for the interactions in the solid and liquid states. Nigsch et al used several types of molecular descriptors to model with *k*-NN the melting temperature of 4119 diverse organic molecules with [101]. The contribution of selected neighbors was evaluated with four weighting functions (arithmetic and geometric average, inverse distance weighting, and exponential weighting), with better results obtained by using the exponential weighting scheme. The optimized model has an average error of 42.2 °C. In a comparison between global and local (*k*-NN) regression models for the blood-brain distribution of 328 chemicals, it was found that *k*-NN consistently gives better predictions as measured in cross-validation [84].

Bayesian Methods

Although Bayesian classifiers were only recently applied to drug design and structure-activity studies, they have several characteristics that make them particularly useful in screening large chemical libraries. Bayesian classifiers give dependable predictions, can tolerate noise and errors in experimental data, and can be computed fast compared to other ML algorithms, which is particularly important in screening large collections of molecules. Bayes' theorem describes the relationship between the prior and conditional probabilities of two events A and B, when B has a non-zero probability [12]:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} ,$$

where the terms have the following meaning: $P(A)$ is the prior or marginal probability of A, and does not consider any information regarding B; $P(B)$ is the prior (marginal) probability of B; $P(A|B)$ is the conditional (posterior) probability of A given the occurrence of event B; $P(B|A)$ is the conditional (posterior) probability of B given the occurrence of event A. Bayesian classifiers [71,151] evaluate a set of descriptors to predict the activity of a chemical based on the known descriptors and activities of a training set of chemicals. Each descriptor is considered to be statistically independent of all other descriptors. Based on a selected descriptor, the probability that a compound is active is proportional to the ratio of active to inactive compounds that have the same value for that descriptor. The final prediction is obtained by multiplying the descriptor-based probabilities.

A variety of drug discovery and development problems can be solved efficiently with Bayesian classifiers. The virtual screening of large chemical libraries represents a major application of Bayesian classifiers, mainly due to their tolerance to noise in experimental data [146]. Klon and Diller proposed a method to prioritize chemicals for synthesis and screening based on physico-chemical properties [77]. It was found that the Bayesian classifier is superior to the usual pairwise Tanimoto similarity filter. High-throughput docking is an essential tool for structure-based drug design, but the current scoring functions do not provide an accurate ranking of the molecules. Various improvements to the scoring functions were proposed, including Bayesian classifiers [78,79]. Simulations with two biological targets showed that a Bayesian classifier increases the enrichment in active compounds. Finding the potential targets of chemicals is increasingly relevant in drug discovery. Nidhi et al developed a multi-class Bayesian model in which chemicals represented with extended-connectivity fingerprints are evaluated against 964 target classes [100]. The practical utility of the approach is demonstrated by its high success rate, with 77% correct target identification. Adverse drug reactions are investigated during early phases of drug discovery in preclinical safety pharmacology tests for various targets. Multi-class Bayesian classifiers developed for 70 such targets can identify 93% of ligands with a 94% correct classification rate [14]. In a test using chemicals from World Drug Index, the classifier identifies 90% of the adverse drug reactions with a 92% correct classification rate. Adverse reactions of several drugs withdrawn from the market were also predicted with good accuracy.

Watson used 2D pharmacophore feature triplet vectors and Bayesian classifiers to find active compounds in library screening [150]. The good enrichment in active

compounds obtained for several chemogenomic databases demonstrates the practical utility of this approach. Natural products represent an excellent source of starting structures for the design of novel drugs. Ertl trained a Bayesian classifier to provide the natural product-likeness score for a chemical [39]. This score measures the similarity between a molecule and the structural space covered by natural products, and can separate between natural products and synthetic molecules. Usual similarity search in chemical libraries is based on a set of structural descriptors and a similarity (or distance) measure. Bender et al showed that the retrieval rate of active compounds increases considerably by using Bayes affinity fingerprints that describe the ligand bioactivity space [13]. In this space, Bayes scores for ligands from about 1000 activity classes describe the chemical structures. This approach was able to improve with 24% the results from a standard similarity screening.

Bayesian methods are also used in various structure-activity models. Klon et al demonstrated several applications of modified Bayesian classifiers that model continuous numerical data with a Gaussian distribution [80]. Several models developed for absorption, distribution, metabolism and excretion property prediction show that the new classifiers have better performance compared to classical Bayesian classifiers. Multidrug resistance represents the ability of cancer cells to become simultaneously resistant to several drugs. A possible solution to this problem is the use of multidrug resistance reversal (MDDR) agents. Sun found that a Bayesian classifier based on atom types could be used to identify MDDR agents in a set of 424 training compounds [125]. The prediction of MDDR agents in a group of 185 test compounds had a success rate of 82.2%. Phospholipidosis is an adverse effect that is investigated during preclinical studies in animals, and may stop the development of an otherwise promising drug candidate. Pelletier et al proposed to identify phospholipidosis with an in silico Bayesian classifier as a viable alternative to more expensive and time-consuming in vivo tests [105]. Based on two simple descriptors, pK_a and ClogP, the model has a success rate of 83% for a dataset of 201 compounds. UGT is an enzyme family that catalyzes the reaction between glucuronic acid and a chemical that has a nucleophilic group. Drug glucuronidation represents an important mechanism for their elimination, but the selectivity of different UGT isoforms is difficult to predict. The Bayesian classifier suggested by Sorich et al predicts the glucuronidation site for eight UGT isoforms based on the partial charge and Fukui function of the nucleophilic atom and the presence of an aromatic ring attached to the nucleophilic atom [124]. The models have a good predictive ability, with a sensitivity and specificity in the range 75–80%. The screening test to identify bitter compounds presented by Rodgers et al uses selected substructural features identified from 649 bitter chemicals and 13,530 random drug-like compounds [111]. A Bayesian classifier trained with these data can predict correctly 72.1% of the bitter compounds. The bitter taste is determined mainly by substructures representing sugar moieties and highly branched carbon skeletons.

Support Vector Machines

Support vector machines represent a major development for SAR and QSAR models, as suggested by the large number of publications that apply SVM and related kernel methods to drug design. To better understand the mathematical basis of SVM and the parameters that influence their results, we start this section with a brief theoretical presentation of SVM for classification and regression. Several kernel functions are presented together with SVM plots that demonstrate the complex shapes that may be simulated with these functions. The influence of various kernels on QSAR predictions is shown for a dataset of benzodiazepine receptor ligands, and several applications in drug design are reviewed.

Hard Margin Linear SVM Classification

The most simple case of SVM is the classification of two classes of objects that may be separated by a hyperplane (Fig. 4a). It is obvious from this figure that there are many hyperplanes that can separate the two classes of objects, but not all of them give good predictions for novel objects. Intuitively, a hyperplane that maximizes the separation of the two classes, like hyperplane H, is expected to offer the best predictions. The SVM algorithm determines a unique hyperplane H that has the maximum margin, i. e., the maximum distance between hyperplanes H_1 and H_2. Hyperplane H_1, which is determined by three objects from class +1 that are shown inside circles, defines the border with class +1. Hyperplane H_2, which is determined by two objects from class −1 that are shown also inside circles, defines the border with class −1. The objects represented inside circles are called support vectors and they determine the unique solution for the SVM model. By removing all other objects and keeping only the support vectors one obtains the same SVM model. The hyperplane H is defined by $\mathbf{w} \cdot \mathbf{x} + b = 0$, where $\mathbf{w}$ is the normal vector to H, the hyperplane H_1 that borders class +1 is defined by $\mathbf{w} \cdot \mathbf{x} + b = +1$, and the hyperplane H_2 that borders class −1 is defined by $\mathbf{w} \cdot \mathbf{x} + b = -1$. The distance between the origin and the hyperplane H is $|b|/\|\mathbf{w}\|$, and the SVM

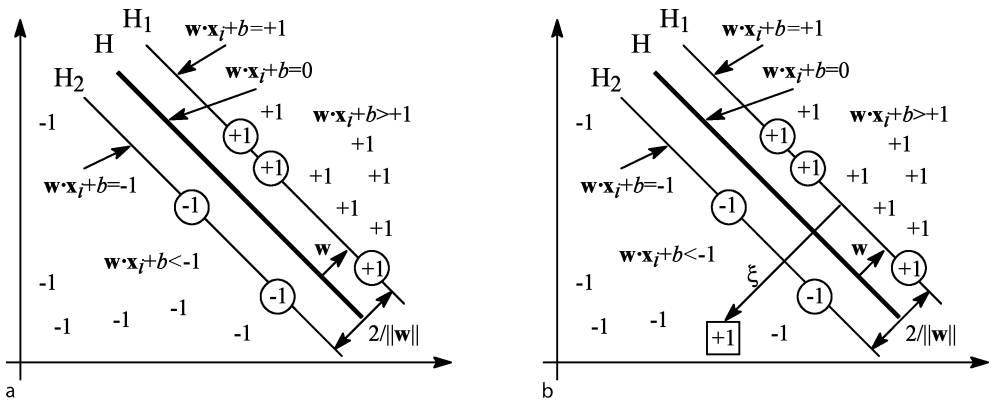

Drug Design with Machine Learning, Figure 4
SVM classification models: a linearly separable data; b linearly non-separable data

margin (the distance between hyperplanes H_1 and H_2) is $2/\|\mathbf{w}\|$. A molecule i characterized by a vector of structural descriptors $\mathbf{x}_i$ belongs to class $+1$ if $\mathbf{w}\cdot\mathbf{x}_i + b \geq +1$, or to class -1 if $\mathbf{w}\cdot\mathbf{x}_i + b \leq -1$. Obviously, support vectors from class $+1$ are determined by $\mathbf{w}\cdot\mathbf{x}_i + b = +1$, whereas support vectors from class -1 are determined by $\mathbf{w}\cdot\mathbf{x}_i + b = -1$.

Based on the formula of the SVM margin, it follows that the maximum separation hyperplane is obtained by maximizing $2/\|\mathbf{w}\|$, which is equivalent to minimizing $\|\mathbf{w}\|^2/2$. The linear SVM model (Fig. 4a) is formulated as:

$$\begin{aligned} &\text{minimize} \quad f(\mathbf{x}) = \frac{\|\mathbf{w}\|^2}{2} \\ &\text{with the constraints} \quad g_i(\mathbf{x}) = y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1 \geq 0\,, \\ &\qquad\qquad\qquad\qquad i = 1,\ldots,m \end{aligned} \tag{2}$$

where y_i is the class ($+1$ or -1) of the object i. This optimization equation is a quadratic programming that is further transformed with a Lagrangian function into its dual formulation. All SVM formulations are solved for the dual formulation, because the solution obtained for this simple case is easily extended to more complex situations. The minimization problem from Eq. (2) is expressed with a Lagrangian function:

$$\begin{aligned} L_P(\mathbf{w}, b, \mathbf{A}) &= f(\mathbf{x}) + \sum_{i=0}^{m} \alpha_i g_i(\mathbf{x}) \\ &= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{m} \alpha_i (y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1) \\ &= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{m} \alpha_i y_i (\mathbf{w}\cdot\mathbf{x}_i + b) + \sum_{i=1}^{m} \alpha_i \\ &= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{m} \alpha_i y_i \mathbf{w}\cdot\mathbf{x}_i \\ &\quad - \sum_{i=1}^{m} \alpha_i y_i b + \sum_{i=1}^{m} \alpha_i\,, \end{aligned} \tag{3}$$

where $\mathbf{A} = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ is the set of Lagrange multipliers for the objects that have $\alpha_i \geq 0$, and P in L_P indicates the primal formulation of the problem. The Lagrangian function L_P must be minimized with respect to $\mathbf{w}$ and b, and maximized with respect to α_i, subject to the constraints $\alpha_i \geq 0$. The dual optimization problem L_D is:

$$\begin{aligned} &\text{maximize} \\ &L_D(\mathbf{w}, b, \mathbf{A}) = \sum_{i=1}^{m} \lambda_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \\ &\text{subject to} \quad \alpha_i \geq 0\,, \quad i = 1,\ldots,m \\ &\text{and} \quad \sum_{i=1}^{m} \alpha_i y_i = 0\,. \end{aligned} \tag{4}$$

The optimization problem is usually solved with the sequential minimal optimization (SMO) proposed by Platt [106]. For a complete formulation of the SVM solution see the books and reviews mentioned in Introduction [70]. The training objects with $\alpha > 0$ constitute the support vectors, whereas the objects with $\alpha = 0$ may be removed from the learning set without affecting the SVM model. The vector $\mathbf{w}$ is obtained as:

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \tag{5}$$

and b is computed as the average value for all support vectors. The SVM model is the optimum separation hyperplane $(\mathbf{w}, b)$ that can now be used to predict the class mem-

bership for new objects. The class of a molecule k with the structural descriptors $\mathbf{x}_k$ is:

$$\text{class}(k) = \begin{cases} +1 & \text{if } \mathbf{w} \cdot \mathbf{x}_k + b > 0 \\ -1 & \text{if } \mathbf{w} \cdot \mathbf{x}_k + b < 0 \,. \end{cases} \tag{6}$$

The classifier is further simplified by substituting $\mathbf{w}$ with its expression (5):

$$\text{class}(k) = \text{sign}\left(\sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x}_k + b\right) , \tag{7}$$

where the summation goes only over the support vectors. The SVM described in the above equations is called a "hard margin" SVM because it does not allow for classification errors.

The class separation with a linear SVM is simple and intuitive. However, SVM are usually used with nonlinear kernels that have a much more complex shape, which might result in too complicated separation surfaces. To demonstrate the shape of the separation surface for linearly separable data we compare results obtained with a linear kernel and several nonlinear kernels that will be introduced later. All calculations were performed with *R* (http://www.r-project.org/) and the kernlab package. In all figures, class + 1 patterns are represented by "+" and class −1 patterns are represented by "−". The SVM hyperplane is depicted with a continuous line, whereas the margins of the SVM hyperplane are shown with dotted lines. Support vectors from class + 1 are represented as "+" inside a circle, and support vectors from the class −1 are depicted as "−" inside a circle.

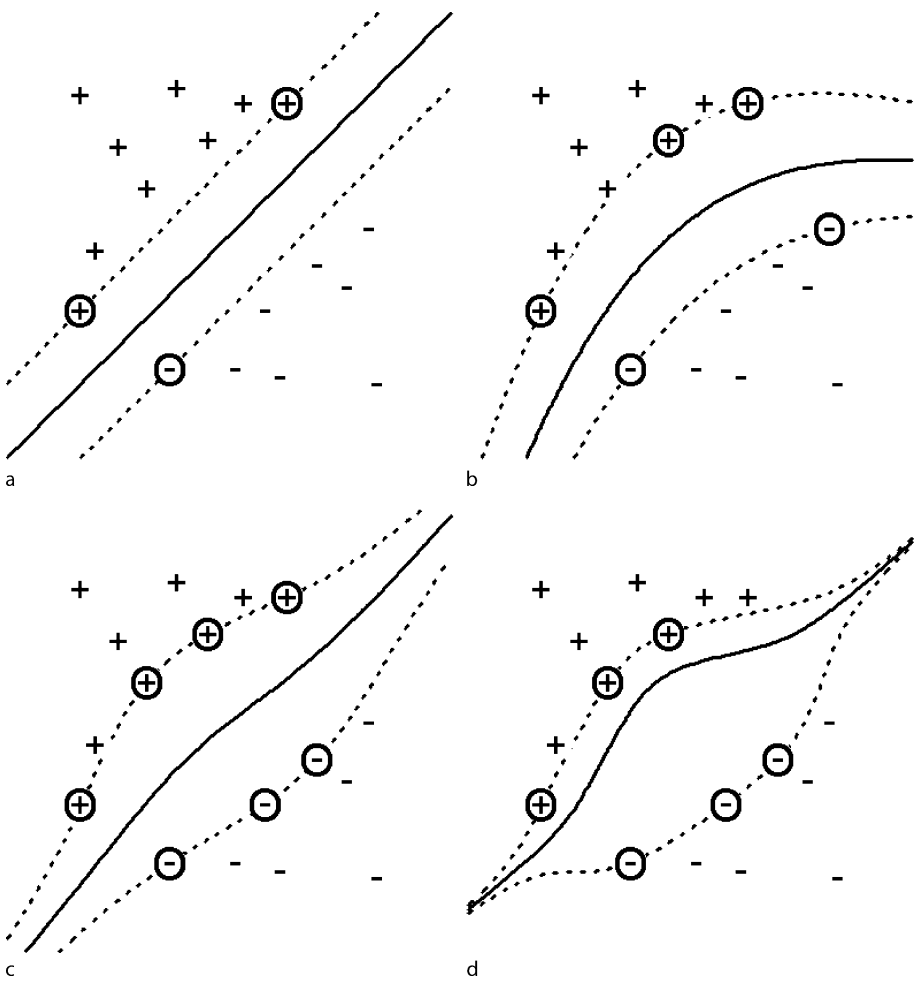

Drug Design with Machine Learning, Figure 5
SVM classification models for linearly separable data: a dot kernel (*linear*); b polynomial kernel, degree 2; c polynomial kernel, degree 3; d polynomial kernel, degree 10

The linear kernel gives an SVM model that has an optimum separation of the two classes (Fig. 5a) with three support vectors, namely two from class + 1 and one from class −1. The hyperplane has the maximum width, and no object is situated inside the margins (represented with dotted lines). The classification of new objects is made by applying Eq. (7) to the three support vectors. The same dataset is modeled with a degree 2 polynomial kernel (Fig. 5b) in which case the SVM model has five support vectors, namely three for class + 1 and two for class −1. The margin width varies, being larger in the middle and smaller towards the extremes. The hyperplane topology is different from that obtained with a linear kernel, and obviously for some objects the two SVM models will give different predictions. The SVM classifier obtained with a degree 3 polynomial kernel (Fig. 5c) has four support vectors from class + 1 and three support vectors from class −1, and the margin becomes smaller towards the ends. Higher order polynomial kernels produce a complex separation hyperplane, as shown also for a degree 10 polynomial kernel (Fig. 5d), in which case the margin almost vanished towards the extremes.

The Gaussian radial basis function (RBF) kernel is the most common option for SVM, and in Fig. 6 we present four SVM models obtained for different σ values. For low σ values the RBF kernel (Fig. 6a, $\sigma = 0.01$) approximates a linear kernel (Fig. 5a) and the two hyperplanes are very similar. As σ increases the nonlinear behavior of the kernel becomes apparent (Fig. 6b, $\sigma = 0.1$), but the SVM model still has a good separation of the data. As σ increases to 1 (Fig. 6c) and then to 5 (Fig. 6d) the nonlinearity of the SVM model becomes more apparent, and the number of support vectors increases. In the last case all data points are support vectors, and the SVM model has a very complex topology. The examples presented here show that the RBF kernel may generate too complex hy-

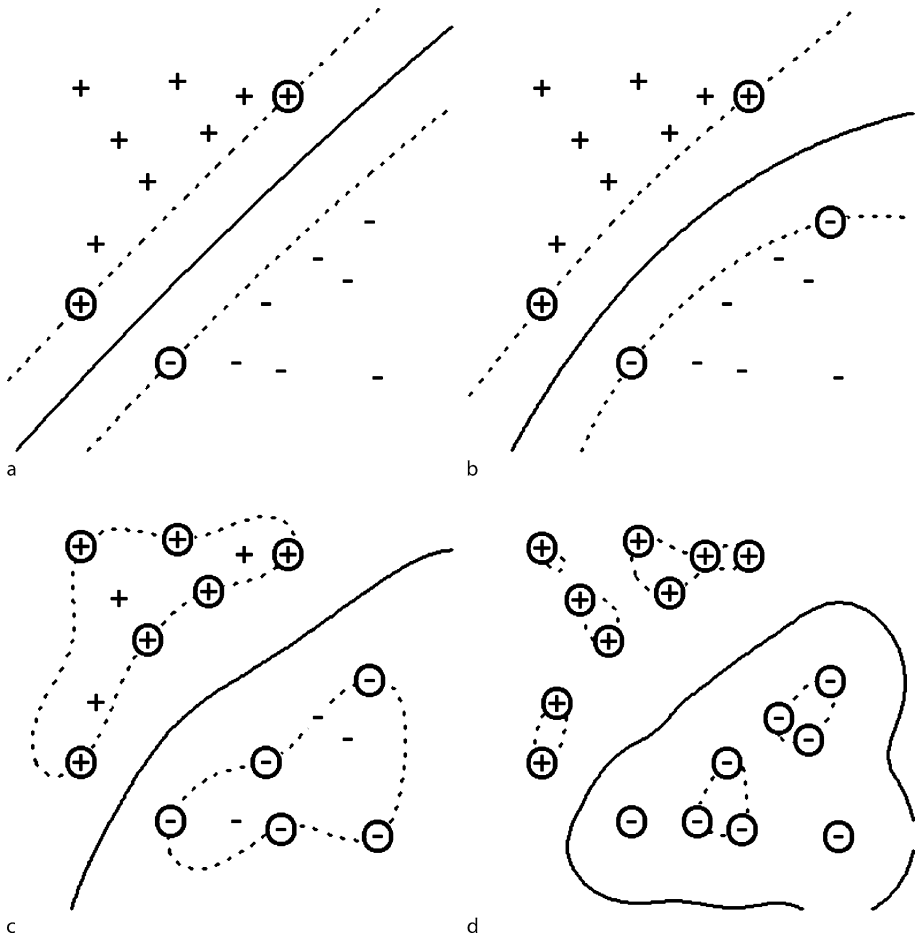

Drug Design with Machine Learning, Figure 6
SVM classification models for linearly separable data with the Gaussian RBF kernel: a $\sigma = 0.01$; b $\sigma = 0.1$; c $\sigma = 1$; d $\sigma = 5$

perplanes that do not properly reflect the structure of the data.

Another important kernel is the hyperbolic tangent (tanh), which is also used in artificial neural networks. Some combinations of parameters result in SVM models almost identical with those obtained with the linear kernel (Fig. 7a) or similar (Fig. 7b). Higher values for the two parameters result in SVM models with smaller margins and nonlinear separation of the two classes (Fig. 7c,d). The SVM models demonstrated here for linearly separable data show that nonlinear kernels, such as RBF and tanh, may give SVM models similar to those obtained with linear kernels, for certain values of their parameters. However, nonlinear kernels may also give SVM models that are too complex and that do not follow the structure of the data. As a general rule, the SVM computed with nonlinear kernels should be compared with a linear kernel SVM.

Soft Margin Linear SVM Classification

The classification problem solved in the previous section considered a two-class situation with classes separated with a linear classifier. However, such cases are rare in drug design applications, and most often the classes of compounds have regions where they superpose. There are several reasons why classes of chemicals cannot be separated with a linear classifier. The most common problem is the identification of the structural descriptors that determine the property. There are thousands of structural descriptors proposed in the literature, but it is difficult to try them all in a SAR or QSAR study. It is always a possibility that the proper type of descriptors is not even discovered. The second problem is to identify the subset of descriptors that give the best predictions, which is a time-consuming task that is usually solved with heuristic al-

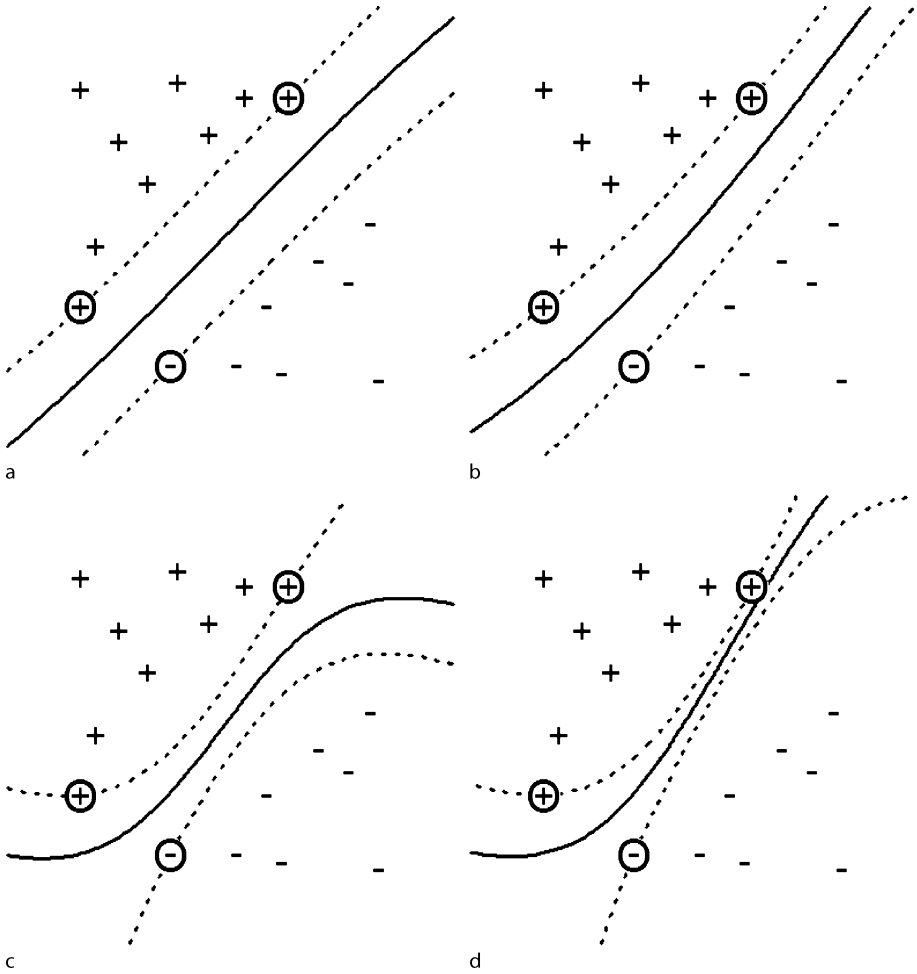

Drug Design with Machine Learning, Figure 7
SVM classification models for linearly separable data with the hyperbolic tangent (tanh) kernel: a $a = 0.1, b = 0$; b $a = 0.1, b = 0.5$; c $a = 0.5, b = 0$; d $a = 0.5, b = 0.5$

gorithms that give a sub-optimal solution. The mapping between the descriptor space and the property might be nonlinear, and obviously a linear classifier fails to model the data. As a final point we have to mention that experimental data may be affected by noise, measurement errors, interference with other biological targets, or unaccounted factors that determine the outcome of the experiments. Although the hard margin linear SVM is limited in its abilities to model real data, the formalism presented in the previous section is the basis for developing more complex SVM models. The linear SVM with soft margin is such an extension that allows classification errors.

The classification problem shown in Fig. 4b is very similar with the linear separable case from Fig. 4a, because all but one object can be separated with a linear hyperplane H. The exception is the object from class +1 that is situated in the −1 space region (shown inside a square). A penalty term is introduced for misclassified data, with the property that it is zero for objects correctly classified, and it has a positive value for objects that are not situated on the correct side of the classifier. The value of the penalty increases when the distance to the corresponding hyperplane increases. The penalty is called a slack variable and it is denoted with ξ. For the situation depicted in Fig. 4b the penalty for the misclassified object is measured starting from hyperplane H_1 because this hyperplane border the +1 region. For +1 molecules situated in the buffer zone between H and H_1, and for −1 molecules situated in the buffer zone between H and H_2, the slack variable takes values between 0 and 1. Such patterns are not considered to be misclassified, but have a penalty added to the objective function. The constraints of the objective function for soft margin linear SVM are:

$$\begin{cases} \mathbf{w}\cdot\mathbf{x}_i + b \geq +1 - \xi_i & \text{if } y_i = +1 \\ \mathbf{w}\cdot\mathbf{x}_i + b \leq -1 + \xi_i & \text{if } y_i = -1 \\ \xi_i > 0\,, \quad \forall i\,. \end{cases} \tag{8}$$

The soft margin SVM should balance two opposite conditions, namely a maximum margin and a minimum of errors that is equivalent with minimizing a sum of slack variables. The optimization problem for a linear SVM with classification errors is;

$$\begin{aligned} &\text{minimize} \quad \frac{\|\mathrm{w}\|^2}{2} + C\sum_{i=1}^{m}\xi_i \\ &\text{with the constraints} \\ &\qquad y_i(\mathrm{w}\cdot\mathrm{x}_i + b) \geq +1 - \xi_i\,, \quad i = 1,\ldots,m \\ &\qquad\qquad \xi_i \geq 0\,, \qquad\qquad i = 1,\ldots,m \end{aligned} \tag{9}$$

where C is called capacity and is a parameter that weights the penalty for classification errors. A large value for the capacity parameter C represents a high penalty for classification errors, which forces the SVM solution to minimize the number of misclassified objects by reducing the margin of the classifier. A small capacity C lowers the value of the penalty term and maximizes the margin, which makes the solution less sensitive to classification errors. The optimization problem is solved with Lagrange multipliers, similar with the procedure outlined for the hard margin linear SVM, when the primal Lagrangian expression is:

$$\begin{aligned} &L_P(\mathbf{w}, b, \mathbf{A}, \mathbf{M}) = \\ &\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i[y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1 + \xi_i] \\ &\qquad - \sum_{i=1}^{m}\mu_i\xi_i \end{aligned} \tag{10}$$

where the Lagrange multipliers $\mathbf{A} = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ are assigned to each constraint $y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq +1 - \xi_i$, and the Lagrange multipliers $\mathbf{M} = (\mu_1, \mu_2, \ldots, \mu_m)$ are assigned to each constraint $\xi_i \geq 0, \forall i = 1, \ldots, m$. The dual optimization problem is:

$$\begin{aligned} &\text{maximize} \\ &L_D(\mathbf{w}, b, \mathbf{A}, \mathbf{M}) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \mathbf{x}_i\cdot\mathbf{x}_j \\ &\text{subject to} \;\; 0 \leq \alpha_i \leq C\,, \quad i = 1,\ldots,m \\ &\text{and} \;\; \sum_{i=1}^{m}\alpha_i y_i = 0\,. \end{aligned} \tag{11}$$

The expression for the vector $\mathbf{w}$ is identical with that obtained for hard margin SVM:

$$\mathbf{w} = \sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i\,. \tag{12}$$

The SVM classifier, which is determined by the pair $(\mathbf{w}, b)$, may be used for the classification of new molecules. The class of a molecule k with the structural descriptors $\mathbf{x}_k$ is:

$$\text{class}(k) = \text{sign}\left(\sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i\cdot\mathbf{x}_k + b\right) \tag{13}$$

where the summation goes only over the support vectors.

Nonlinear Classification with Kernels

Nonlinear relationships between structural descriptors and drug properties may be modeled with SVM equations that replace the linear classifier with nonlinear kernel functions. The SVM extension to nonlinear relationships is

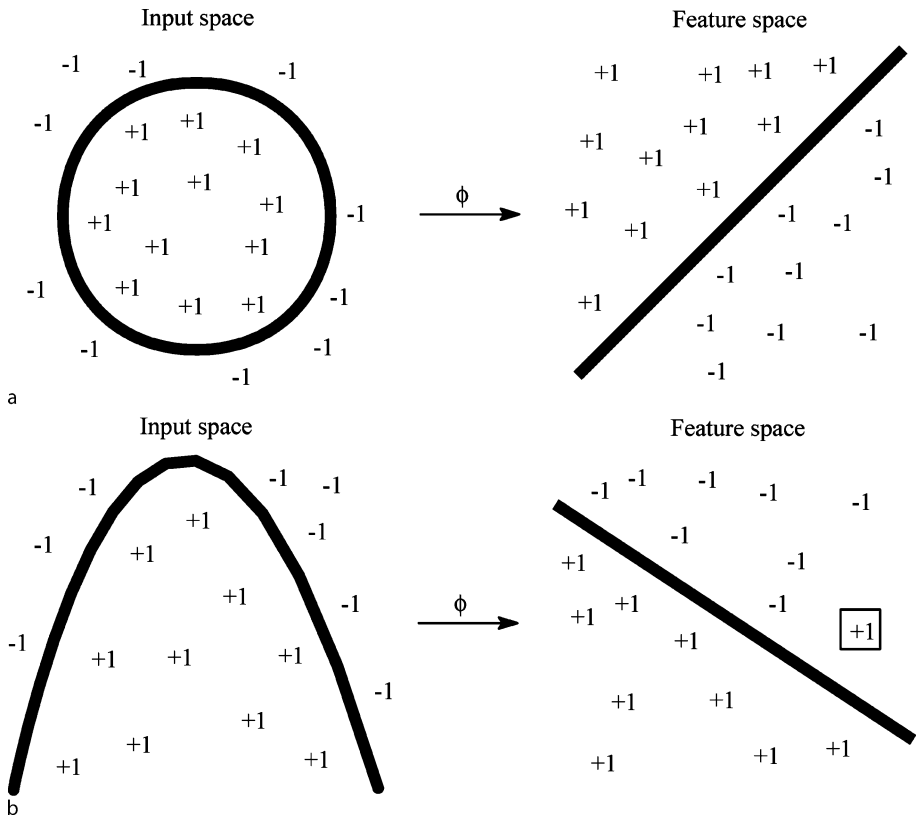

Drug Design with Machine Learning, Figure 8

Patterns that are nonlinear separable in the input space are linear separable in the feature space: a the class separation is complete; b a pattern from class + 1 (depicted in a *square*) is situated in the −1 region of the hyperplane

based on the property of feature functions $\phi(\mathbf{x})$ that transform the input spaces into a feature space of higher dimensionality, in which a linear separation of the classes is possible. This transformation is depicted in Fig. 8a. In the input space, which is represented by the structural descriptors, a circle can discriminate the two classes of objects. Using a proper feature function ϕ the input space is transformed into a feature space in which the two classes can be separated with a linear classifier. Figure. 8b illustrates another situation in which another nonlinear function can discriminate the classes in input space, which is then transformed by the function ϕ into a linear classifier in feature space. The challenge is to find that feature function ϕ that maps the objects into a feature space in which the classes are linearly separable. In real-life applications it is not even desirable to achieve a perfect separation, because the function may map noise or errors in data (see Fig. 8b, where a + 1 object, depicted in a square, is situated in the −1 region of the hyperplane). Due to the fact that in feature space the objects may be separated with a linear classifier, it is possible to use the results for linear SVM presented in the previous sections, which is a fundamental property of SVM.

The mapping from the input space to the feature space may be expressed as:

$$\mathbf{x} = (x_1, x_2, \ldots, x_n) \rightarrow \phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_h(\mathbf{x}))\,. \quad (14)$$

A soft margin SVM in the feature space is obtained by simply substituting the input vectors $\mathbf{x}$ with the feature vectors $\phi(\mathbf{x})$. We have to point here that a good linear classifier in feature space depends on a proper form of the function ϕ. The class of a molecule k with the structural descriptors $\mathbf{x}_k$ is:

$$\begin{aligned}\text{class}(k) &= \text{sign}[\mathbf{w} \cdot \phi(\mathbf{x}_k) + b] \\ &= \text{sign}\left(\sum_{i=1}^{m} \alpha_i y_i \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_k) + b\right)\end{aligned} \quad (15)$$

The prediction for a molecule k requires the computation of the dot product $\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_k)$ for all support vectors $\mathbf{x}_i$.

This is an important observation showing that we do not have to know the actual expression of the feature function ϕ. Furthermore, nonlinear SVM use kernels, which are a special class of functions that compute the dot product $\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_k)$ in the input space:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \,. \tag{16}$$

Several popular kernel functions are presented below. The inner product of two vectors defines the linear (dot) kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j \,. \tag{17}$$

The dot kernel is a linear classifier, and should be used as a reference to demonstrate an eventual improvement of the classification with nonlinear kernels. The polynomial kernel is useful mainly for lower degrees, because for higher degrees it has the tendency to overfit the data:

$$K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i \cdot \mathbf{x}_j)^d \,. \tag{18}$$

The Gaussian radial basis functions (RBF) kernel is very flexible, and depending on the values of the parameter σ it can display a wide range of nonlinearity:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\sigma \|x - y\|^2) \,. \tag{19}$$

The hyperbolic tangent (tanh) function has a sigmoid shape and it is the most used transfer function for artificial neural networks. The corresponding kernel has the formula:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(a\mathbf{x}_i \cdot \mathbf{x}_j + b) \,. \tag{20}$$

The nonlinearity of the anova kernel is controlled by the parameters γ and d:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_i \exp(-\gamma(\mathbf{x}_i - \mathbf{x}_j)) \right)^d \,. \tag{21}$$

Hard Margin Nonlinear SVM Classification

The formulation of the hard margin nonlinear SVM classification is obtained from the linear SVM by substituting input vectors $\mathbf{x}$ with feature functions $\phi(\mathbf{x})$. The dual problem is:

$$\begin{aligned} &\text{maximize} \\ &L_D(\mathbf{w}, b, \mathbf{A}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \\ &\text{subject to } \alpha_i \geq 0 \,, \quad i = 1, \ldots, m \\ &\text{and } \sum_{i=1}^{m} \alpha_i y_i = 0 \,. \end{aligned} \tag{22}$$

The optimum separation hyperplane is determined by the vector $\mathbf{w}$:

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \phi(\mathbf{x}_i) \,. \tag{23}$$

In the final expression of the classifier, the dot product for two feature functions $\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_k)$ is replaced with a kernel function $K(\mathbf{x}_i, \mathbf{x}_k)$:

$$\text{class}(k) = \text{sign} \left(\sum_{i=1}^{m} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}_k) + b \right) \,. \tag{24}$$

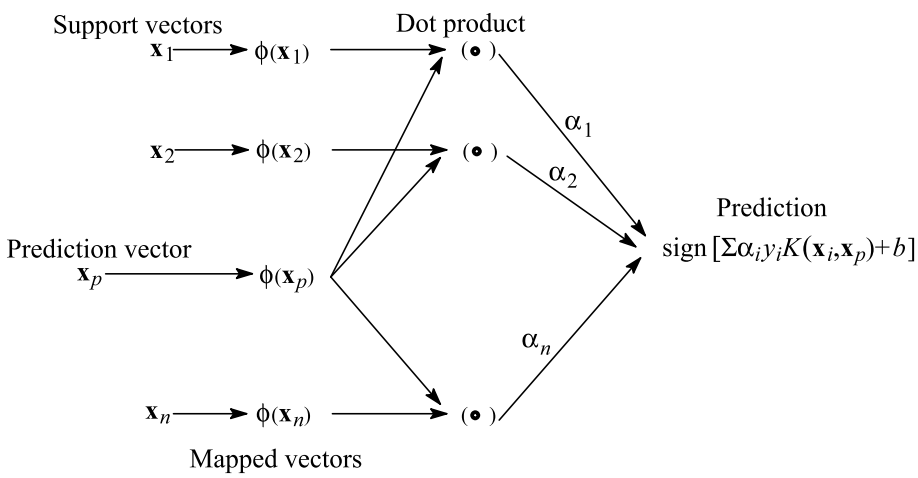

Drug Design with Machine Learning, Figure 9
Network representation of support vector machines

The structure of a nonlinear SVM may be represented in a network format (Fig. 9). The network input is represented by a set of support vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and a prediction object $\mathbf{x}_p$. The input space is transformed with a feature function ϕ into the feature space, and the dot product of feature functions is computed and multiplied with the Lagrangian multipliers α. The output is Eq. (24) in which the dot product of feature functions is substituted with a kernel function K.

Soft Margin Nonlinear SVM Classification

The expression of the soft margin nonlinear SVM classification is obtained from the corresponding linear SVM formula by substituting input vectors $\mathbf{x}$ with feature functions $\phi(\mathbf{x})$. The dual problem is:

$$\begin{aligned} &\text{maximize} \\ &L_D(\mathbf{w}, b, \mathbf{A}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \\ &\text{subject to } \; 0 \le \alpha_i \le C\,, \quad i = 1, \ldots, m \\ &\text{and } \; \sum_{i=1}^{m} \alpha_i y_i = 0 \end{aligned} \tag{25}$$

with a solution identical with that for hard margin nonlinear SVM classification from Eq. (24).

In Fig. 10 we demonstrate the nonlinear SVM classification for the Gaussian RBF kernel. The best separation

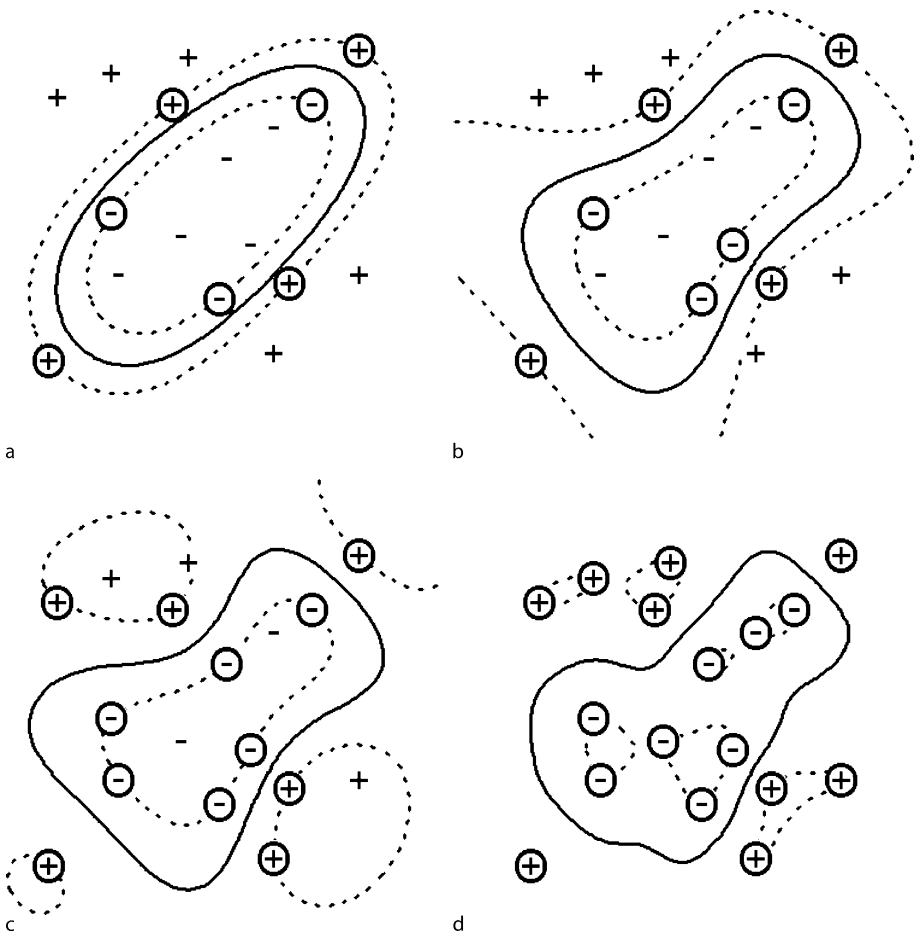

Drug Design with Machine Learning, Figure 10
SVM classification models for linearly non-separable data with the Gaussian RBF kernel: a $\sigma = 0.1$; b $\sigma = 0.5$; c $\sigma = 1$; d $\sigma = 5$

of the two classes is obtained for a low σ value (Fig. 10a, $\sigma = 0.1$). As the value of σ increases the kernel nonlinearity increases and the separation surface becomes more complex. It is clear from this example that the kernel nonlinearity must match the structure of the data. Also, highly nonlinear functions do not translate in better SVM models. In practical applications the parameters that determine the kernel shape must be optimized to provide the best predictions in a cross-validation test.

SVM Regression

Vapnik extended the SVM algorithm for regression [139] by using an ε-insensitive loss function (Fig. 11). The SVM regression (SVMR) determines a function $f(\mathbf{x})$ with the property that for all learning objects $\mathbf{x}$ it has a maximum deviation ε from the target (experimental) values y, and it has a maximum margin. Starting from the training objects, SVMR computes a model representing a tube with radius ε fitted to the data. All object situated inside the regression tube have an error equal to zero. For the hard margin SVMR no object is allowed outside the tube, whereas for the soft margin SVMR uses positive slack variables to penalize the objected situated outside the tube [44,45,93,94,122]. If Fig. 11 the objects situated inside the regression tube are shown as white circles, whereas the objects outside the regression tube are depicted as black circles. The slack variable ξ introduces a penalty proportional with the distance between the object and the margin of the regression tube. Support vectors are selected only from objects situated outside the tube.

To demonstrate the kernel influence on the SVM regression we model the inhibition constant K_i for bovine milk xanthine oxidase [51]. The structural descriptor used is ClogP, the computed hydrophobicity. To model the nonlinear dependence between ClogP and K_i we apply a degree 2 polynomial kernel (Fig. 12) with different values for the ε-insensitive parameter. The molecules situated inside the regression tube are indicated with black triangles, whereas those located outside the regression tube are represented as black dots (support vectors are shown as black dots inside a circle). A low value for ε produces a SVMR model with a larger number of support vectors, but an overall good shape for the relationship between ClogP and K_i (Fig. 12a, $\varepsilon = 0.1$). As ε increases the diameter of the regression tube includes almost all molecules, and thus the model depends on a small number of support vectors because objects situated inside the tube do not influence the regression model.

We explore further two higher order polynomials, namely a degree 5 polynomial kernel (Fig. 13a) and a degree 8 polynomial kernel (Fig. 13b). Both kernels overfit the data, and have shapes that are too complex. The Gaussian RBF kernel with $\sigma = 0.2$ is a good fit for the K_i data (Fig. 13c), but an increase in σ results in a regression that start to fit the errors in the experimental data (Fig. 13d).

The next set of SVM regression QSAR evaluates the tanh kernel (Fig. 14). The first combination of parameters is insensitive to the relationship between ClogP and K_i (Fig. 14a). Although the K_i data are best represented by a downward parabola, the second tanh QSAR has the shape of an upward parabola, and misses the most apparent correlation between ClogP and K_i (Fig. 14b). The third tanh QSAR is overfitted, mainly in its right side (Fig. 14c). The last combination of parameters for the tanh QSAR shows a very good fit of the data and obviously it is a good candidate for the prediction of new molecules (Fig. 14d). Considering the large variation in shape of the tanh QSAR it is apparent that its parameters should be varied over a large range to find the best regression model.

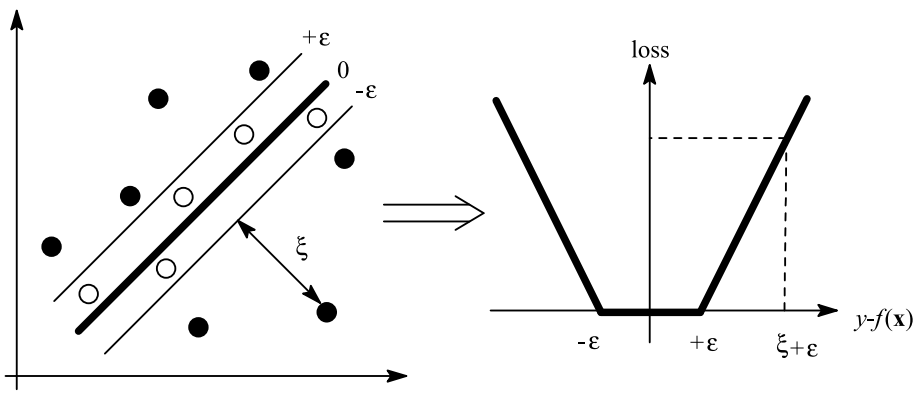

Drug Design with Machine Learning, Figure 11
SVM regression with ε-insensitive loss function

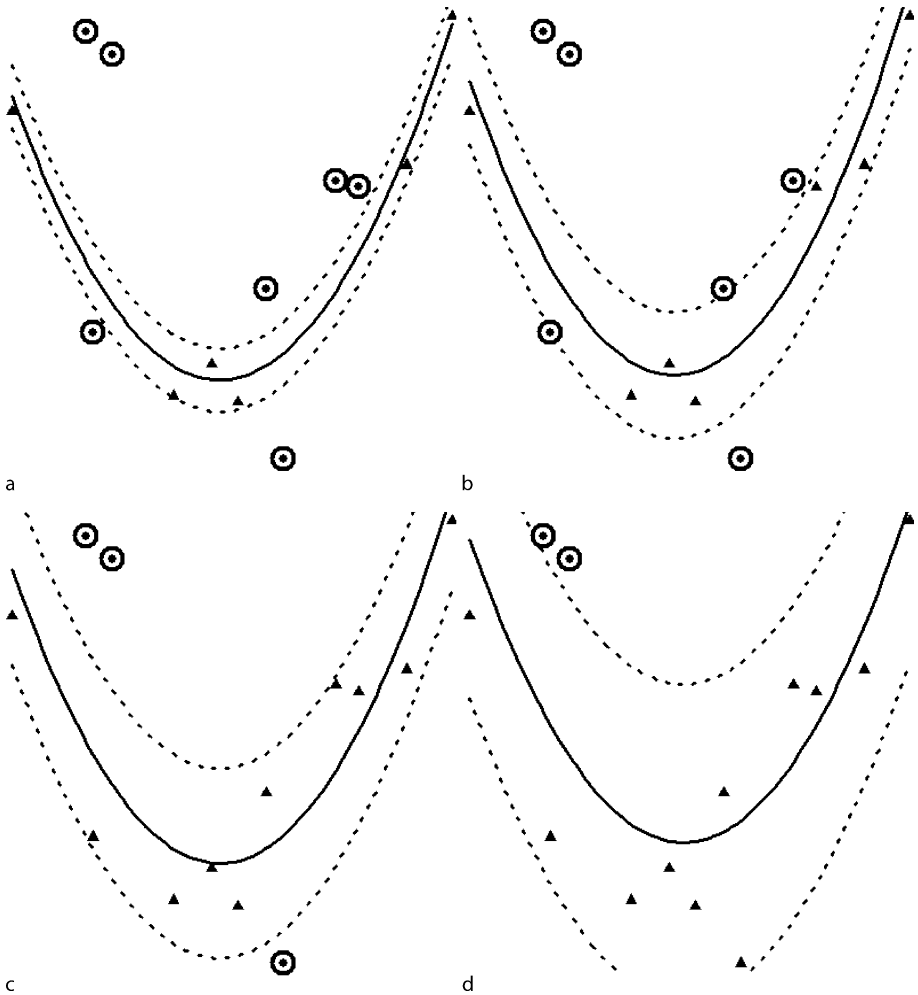

Drug Design with Machine Learning, Figure 12
SVM regression models with a degree 2 polynomial kernel ($C = 1000$): a $\varepsilon = 0.1$; b $\varepsilon = 0.2$; c $\varepsilon = 0.3$; d $\varepsilon = 0.5$

Since the dataset considered is a real QSAR problem, it is interesting to examine the behavior of other, less common, kernels. The anova RBF kernel shows an overfit of the data, which is more prominent in its left side (Fig. 15a). The spline kernel has a too low nonlinearity, and the model has no value in predicting new chemicals (Fig. 15b). The degree 1 Bessel kernel is a very good fit for the correlation between ClogP and K_i (Fig. 15c), whereas a degree 5 Bessel kernel overfits the data in the left side (Fig. 15d). The results presented here are only to illustrate the shape of various kernels, and to demonstrate their ability to model nonlinear relationships. Other combinations of parameters and capacity C may give better or worse correlations, and the most predictive QSAR may be identified only by optimizing the kernel parameters.

Kernel Comparison in QSAR

Several kernels have the property to give SVM models that are universal approximators to arbitrary functions, namely the SVM model can approximate any function to any level of precision. However, the theory cannot guarantee that such a model will be also optimally predictive. The essential requirement for a SAR or QSAR model is its ability to provide reliable predictions for new molecules. The simple data fitting ability with a low error is not enough for an SVM model, because such models are usually overfitted and give bad predictions. The experimental data used in SAR and QSAR are affected by experimental errors and sometimes by the fact that good structural descriptors are missing from the model.

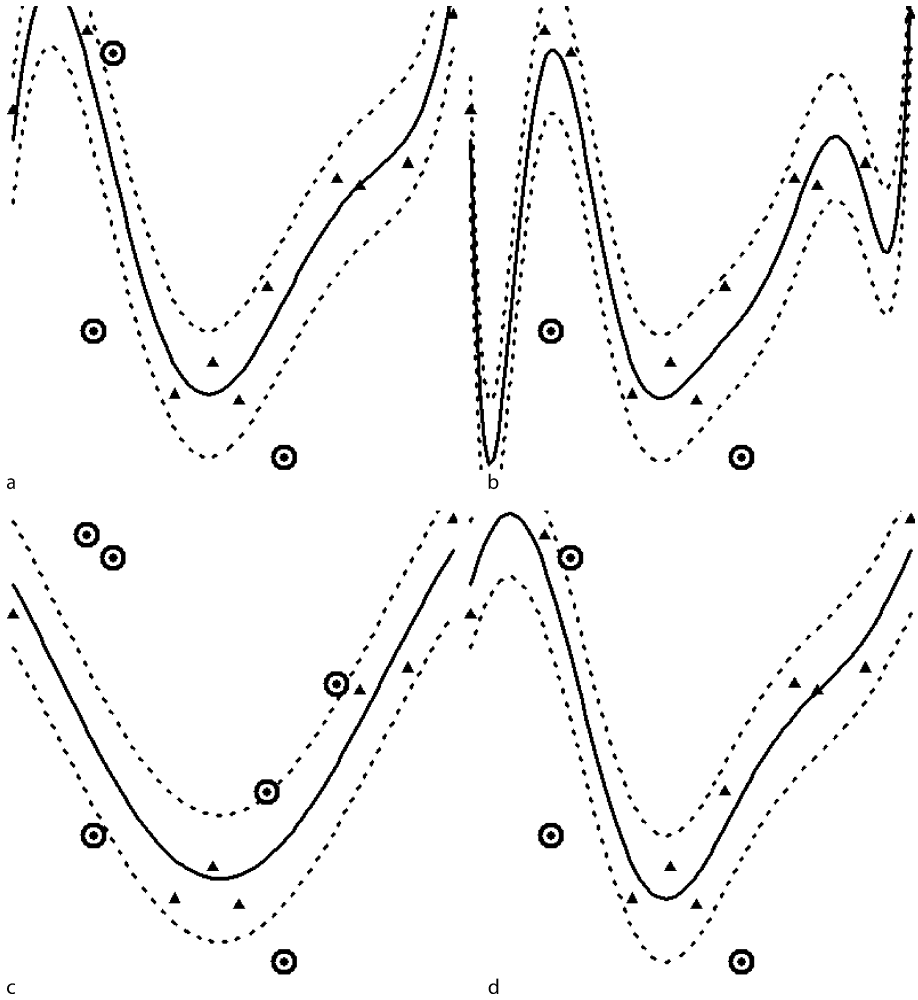

Drug Design with Machine Learning, Figure 13
SVM regression models ($C = 1000$, $\varepsilon = 0.2$): a degree 5 polynomial kernel; b degree 8 polynomial kernel; c Gaussian RBF kernel, $\sigma = 0.1$; d Gaussian RBF kernel, $\sigma = 0.5$

Therefore it is important to evaluate a number of kernels in order to find the most predictive SVM model and to avoid fitting the noise or the errors from the data. To offer a comparative evaluation of several kernels we review here a number of SVM classification and regression models in which five kernels are compared, namely linear, polynomial, Gaussian RBF, tanh, and anova. Each kernel was evaluated for a large range of parameters, and all results are compared for cross-validation. The SVM models were computed with mySVM, an efficient SVM program authored by Rüping (http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/index.html).

The carcinogenic activity of a group of 46 methylated and 32 non-methylated polycyclic aromatic hydrocarbons was modeled with semiempirical quantum indices [61]. From the set of 78 chemicals, 34 are carcinogenic and 44 are non-carcinogenic. The accuracy for leave-10%-out (L10%O) cross-validation tests shows that the best predictions are obtained with the Gaussian RBF kernel: RBF, $\sigma = 0.5$, $AC = 0.86$; anova, $\gamma = 0.5$, $d = 1$, $AC = 0.84$; degree 2 polynomial, $AC = 0.82$; linear, $AC = 0.76$; tanh, $a = 2$, $b = 0$, $AC = 0.66$. All SVM classification models were obtained with $C = 10$. The increase in prediction accuracy for RBF and anova kernel, compared with the linear SVM, indicates that there is a nonlinear relationship between the quantum indices and hydrocarbon carcinogenicity.

A structure-odor classification was performed for 98 tetra-substituted pyrazines representing three odor classes, namely 32 green, 23 nutty, and 43 bell-pepper [62]. L10%O predictions were obtained for three classification tests, in which one odor type is selected as class + 1 and

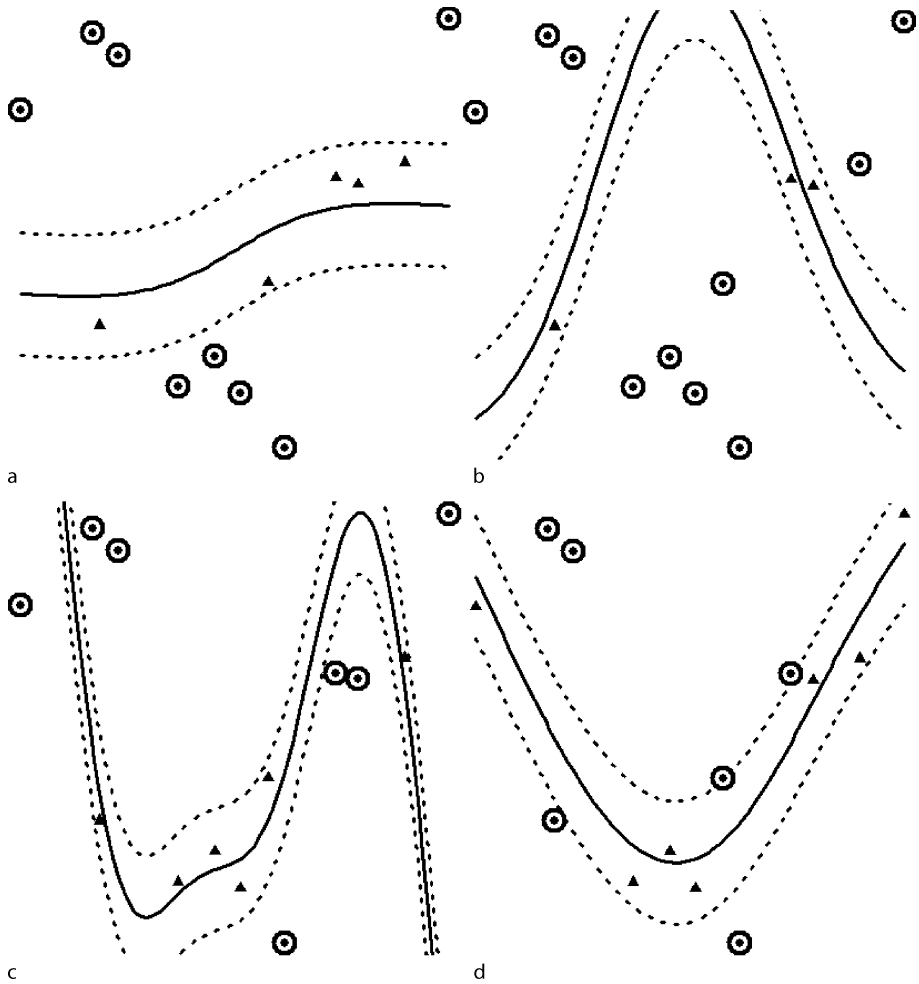

Drug Design with Machine Learning, Figure 14
SVM regression models with a hyperbolic tangent (tanh) kernel ($\varepsilon = 0.2$): a $a = 1, b = 0, C = 1$; b $a = 1, b = 1, C = 1$; c $a = 0.9, b = -1, C = 10$; d $a = 0.5, b = -0.7, C = 10$

the remaining two odor types form class −1. For green aroma compounds the best predictions are obtained with a polynomial kernel: degree 2 polynomial, $C = 1000$, $AC = 0.86$; anova, $C = 100$, $\gamma = 0.5$, $d = 1$, $AC = 0.84$; linear, $C = 10$, $AC = 0.80$; RBF, $C = 10$, $\sigma = 0.5$, $AC = 0.79$; tanh, $C = 10$, $a = 0.5$, $b = 0$, $AC = 0.73$. For nutty aroma compounds the anova kernel has slightly better predictions: anova, $C = 100$, $\gamma = 0.5$, $d = 1$, $AC = 0.92$; linear, $C = 10$, $AC = 0.89$; degree 2 polynomial, $C = 10$, $AC = 0.89$; RBF, $C = 10$, $\sigma = 0.5$, $AC = 0.89$; tanh, $C = 100$, $a = 0.5$, $b = 0$, $AC = 0.79$. The SVM models for bell-pepper aroma compounds show that three kernels (RBF, polynomial, and anova) give good predictions: RBF, $C = 10$, $\sigma = 0.5$, $AC = 0.89$; degree 2 polynomial, $C = 10$, $AC = 0.88$; anova, $C = 10$, $\gamma = 0.5$, $d = 1$, $AC = 0.87$. linear, $C = 10$, $AC = 0.74$; tanh, $C = 100$, $a = 2$, $b = 1$, $AC = 0.68$. In all three SVM experiments the best model outperformed the linear kernel, showing that nonlinear relationships are useful in structure-aroma studies.

The relationships between the structure of chemical compounds and their aqueous toxicity are highly relevant because many chemicals are environmental pollutants. The mechanism of toxic action (MOA) of organic chemicals may be predicted from hydrophobicity and experimental toxicity against *Pimephales promelas* and *Tetrahymena pyriformis* [67]. SVM classification was used to discriminate between 126 nonpolar narcotics and 211 chemicals that have other MOA. The RBF, anova, linear and polynomial kernels have similar statistics: RBF, $\sigma = 1$, $AC = 0.80$; anova, $\gamma = 0.5$, $d = 2$, $AC = 0.80$; linear, $AC = 0.79$; degree 2 polynomial, $AC = 0.79$; tanh,

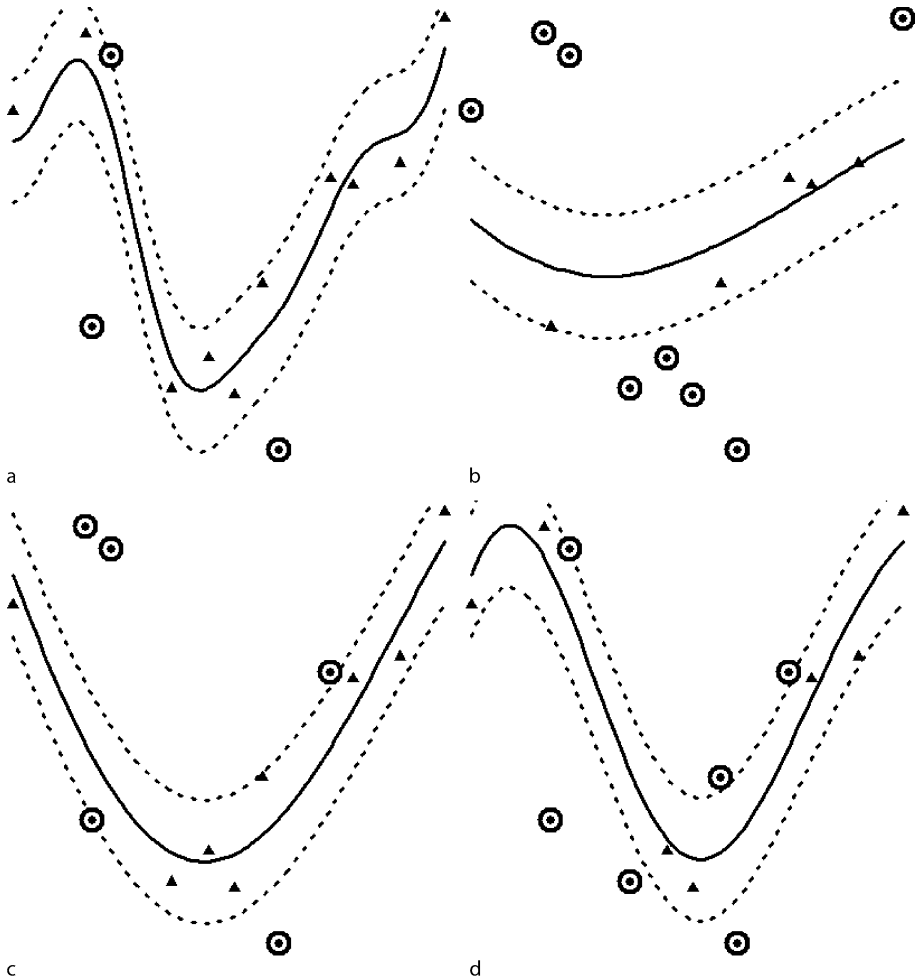

Drug Design with Machine Learning, Figure 15
SVM regression models ($\varepsilon = 0.2$): a anova RBF kernel, $\sigma = 0.2$, degree = 5, $C = 1000$; b spline kernel, $C = 0.05$; c Bessel kernel, $\sigma = 1$, order = 1, degree = 1, $C = 1000$; c Bessel kernel, $\sigma = 0.6$, order = 1, degree = 5, $C = 1000$

$a = 0.5$, $b = 0$, $AC = 0.53$. These results indicate that there is an almost linear relationship between descriptors and MOA. A similar conclusion was obtained for the MOA classification of polar and nonpolar narcotic compounds [66]. In a related study for the classification of MOA for narcotic and reactive compounds [63], the degree 2 polynomial offered the best predictions: degree 2 polynomial, $C = 10$, $AC = 0.92$; anova, $C = 10$, $\gamma = 0.5$, $d = 1$, $AC = 0.87$; linear, $C = 1000$, $AC = 0.86$; RBF, $C = 100$, $\gamma = 0.5$, $AC = 0.83$; tanh, $C = 10$, $a = 0.5$, $b = 0$, $AC = 0.78$.

SVM regression is demonstrated for QSAR models obtained for 52 benzodiazepine receptor ligands [68]. Five structural descriptors were used as SVM input, and the prediction was evaluated with leave-5%-out (L5%O) and L10%O cross-validation. A number of 34 experiments were conducted with five kernels (Tab. 1). The predictions obtained with the linear kernel L are used as baseline to compare the results provided by the other kernels. All four polynomial kernels P give very bad predictions, as indicated by the statistical indices. The RBF kernel R has better predictions than L only for $\sigma = 0.25$ (experiment 6), whereas all the other RBF models are worse than L. Overall, the best results are obtained with the tanh kernel T for $a = 0.5$ and $b = 0$ (experiment 11) which is a clear improvement compared to the linear kernel. Finally, the anova kernel A has uneven predictions. Some are good (experiments 22 and 20) whereas experiments 25–34 have very bad predictions. The SVM regression experiments presented in Tab. 1 show that kernel parameters are of paramount importance in obtaining a predictive QSAR.

Drug Design with Machine Learning, Table 1
Support vector regression statistics for benzodiazepine receptor ligands

Exp	Kernel	p_1	p_2	$q^2_{L5\%O}$	$RMSE_{L5\%O}$	$q^2_{L10\%O}$	$RMSE_{L10\%O}$
1	L			0.261	0.98	0.273	0.97
2	P	2		< -100	> 10	< -100	> 10
3	P	3		< -100	> 10	< -100	> 10
4	P	4		< -100	> 10	< -100	> 10
5	P	5		< -100	> 10	< -100	> 10
6	R	0.25		0.368	0.91	0.370	0.91
7	R	0.5		0.226	1.00	0.324	0.94
8	R	1.0		0.221	1.01	0.310	0.95
9	R	1.5		0.236	1.00	0.258	0.98
10	R	2.0		0.208	1.01	0.205	1.02
11	T	0.5	0.0	0.453	0.84	0.498	0.81
12	T	1.0	0.0	0.416	0.87	0.411	0.87
13	T	2.0	0.0	0.396	0.89	0.394	0.89
14	T	0.5	1.0	0.070	1.10	0.120	1.07
15	T	1.0	1.0	0.389	0.89	< -10	5.36
16	T	2.0	1.0	0.297	0.96	0.348	0.92
17	T	0.5	2.0	−0.365	1.33	−0.405	1.35
18	T	1.0	2.0	0.106	1.08	0.243	0.99
19	T	2.0	2.0	0.345	0.92	0.376	0.90
20	A	0.25	1	0.397	0.89	0.377	0.90
21	A	0.5	1	0.324	0.94	0.331	0.93
22	A	1.0	1	0.398	0.88	0.412	0.87
23	A	1.5	1	0.299	0.95	0.374	0.90
24	A	2.0	1	0.293	0.96	0.339	0.93
25	A	0.25	2	−1.289	1.73	−0.921	1.58
26	A	0.5	2	−1.375	1.76	−0.873	1.56
27	A	1.0	2	−0.606	1.44	−0.465	1.38
28	A	1.5	2	−0.241	1.27	−0.195	1.25
29	A	2.0	2	−0.071	1.18	−0.060	1.17
30	A	0.25	3	−2.998	2.28	−1.934	1.95
31	A	0.5	3	−1.282	1.72	−0.983	1.61
32	A	1.0	3	−0.060	1.17	−0.094	1.19
33	A	1.5	3	0.060	1.11	−0.097	1.19
34	A	2.0	3	−0.062	1.18	−0.073	1.18

In this section we presented comparative SAR and QSAR models obtained with five kernels. In developing SVM models it is important to include the linear kernel as a baseline comparison. If nonlinear kernels do not improve the predictions of the linear kernel, than the structure-activity relationship is linear and a linear SVM should be used. The parameters of nonlinear kernels should be optimized in order to obtain the best prediction. The most predictive SAR and QSAR model may be obtained with anyone of the five kernels investigated, and there is no apparent rule that can predict which kernel is the best. Therefore, it is important to experiment with a large diversity of kernels in order to find a predictive SVM model.

Applications in Drug Design

Torsade de pointes (TdP) is an important adverse drug reaction that is responsible for almost one-third of all drug failures during drug development, and resulted in several drugs being withdrawn from the market. Yap et al compared several ML for their ability to separate TdP compounds from those that do not induce TdP [156]. The prediction accuracy shows that SVM is the most reliable ML: SVM with Gaussian RBF kernel 91.0%, *k*-NN 88.5%, probabilistic neural network 78.2%, and C4.5 decision tree 65.4%. Inhibitors of the hERG potassium channel can lead to a prolongation of the QT interval that can trigger TdP,

an atypical ventricular tachycardia. A dataset of 39 hERG inhibitors (26 for learning and 13 for test) was evaluated by a pharmacophore ensemble coupled with SVM regression [87]. The pharmacophore ensemble encodes the protein conformational flexibility, and the SVM brings the capability of modeling nonlinear relationships. The QSAR model has very good statistics, with $q^2 = 0.89$ and $r^2 = 0.94$ for the test set. A structure-activity SVM study for hERG inhibitors was performed by Tobita et al for 73 drugs [132]. Using a combination of structural descriptors and molecular fragments, the SVM model had a prediction accuracy higher than 90%. Selecting an optimum group of descriptors is an important phase in computing a predictive structure-activity model. The SVM recursive feature elimination algorithm was evaluated by Xue et al for three SAR datasets, namely *P*-glycoprotein substrates, human intestinal absorption, and compounds that cause TdP [153]. The prediction statistics show that a proper descriptor selection may significantly increase the SVM performance.

The drug metabolism by human cytochrome P450 isoforms 3A4, 2D6, and 2C9 was studied by Terfloth et al for 379 drugs and drug analogs [131]. The best SVM model has a cross-validation accuracy of 89% and a prediction accuracy of 83% for 233 validation compounds. Unbalanced data sets, with a majority of inactive compounds and only a few active chemicals, represent the usual situation in library screening. To overcome the ML bias toward the larger class, Eitrich et al used SVM with oversampling to identify cytochrome P450 2D6 inhibitors [37]. An ensemble of SVM models was presented by Yap and Chen as suitable classifiers for inhibitors and substrates of cytochromes P450 3A4, 2D6, and 2C9 [155]. Descriptors computed with Dragon were used to develop predictive models, with MCC (Matthews correlation coefficient) statistics, namely 0.899 for 3A4, 0.884 for 2D6, and 0.872 for 2C9. The classification of cytochrome P450 3A4 inhibitors and non-inhibitors was investigated by Arimoto et al with SVM, *k*-NN, recursive partitioning, logistic regression, and Bayesian classifier [7]. The chemical structure of the 4470 compounds was encoded with fingerprints and MolconnZ topological indices, and the best predictions were obtained with SVM trained with fingerprints. A comparison of single and ensemble classifiers was made by Merkwirth et al for SVM, *k*-NN, and ridge regression [95]. The tests conducted for two libraries of chemicals, one for unspecific protein inhibition and another one for inhibitors of cytochrome P450 3A4, showed that single and ensemble SVM models give similar results, with better predictions than *k*-NN and ridge regression.

Du et al predicted the genotoxicity of 140 thiophene derivatives with SVM and linear discriminant analysis (LDA) [34]. Seven structural descriptors were used as input data for the classification models, and the SVM parameters were identified by grid search. The accuracy of the SVM model (92.9% in training and 92.6% in prediction) is significantly higher than the LDA accuracy (81.4% in training and 85.2% in prediction). Recursive partitioning and SVM were compared in a test for mutagenicity prediction from substructure descriptors [92]. The prediction accuracy for 2199 mutagens is similar for SVM (81.4%) and recursive partitioning (80.2%).

SVM classification was compared with linear discriminant analysis in predicting drug bioavailability [149]. The learning set comprises 167 compounds characterized by five structural descriptors. The SVM parameters were optimized with grid search. The classification results indicate that SVM (accuracy 85.6%) provides better models for bioavailability compared to LDA (accuracy 72.4%). Sakiyama et al studied SAR models for the metabolic stability of drug candidates [113]. The structure-activity models were obtained with random forest, support vector machine, logistic regression, and recursive partitioning. Classification models were obtained for a collection of 1952 compounds characterized by 193 descriptors. All ML models had accuracy > 0.8, with slightly higher results for random forest and SVM.

The identification of drug-like compounds based on atom type counts was investigated with SVM, linear programming machines, linear discriminant analysis, bagged *k*-nearest neighbors, and bagged decision trees C4.5 [97]. The classification dataset consisted of drug-like compounds from World Drug Index and non-drug compounds from Available Chemicals Directory. The smallest errors were obtained with SVM trained with polynomial and Gaussian RBF kernels. A computational filter for chemical compounds with antidepressant activity was developed by Lepp et al based on 21 biological targets related to depression [88]. An SVM model obtained with atom type descriptors gave high enrichment levels, and showed that a compound selected as active interacts on average with 2.3 targets. SVM, artificial neural networks, and multiple linear regression were compared in a model for apoptosis induction by 4-aryl-4-H-chromenes [40]. A leave-one-out cross-validation test showed that SVM gives the best predictions. Briem and Günther used a dataset of 565 kinase inhibitors and 7194 inactive compounds to compare four machine learning algorithms, namely SVM with a Gaussian RBF kernel, artificial neural networks, *k*-nearest neighbors, and recursive partitioning [23]. The best predictions were obtained with SVM, followed closely by

k-NN. The SVM ability to identify active compounds from a chemical library was investigated by Jorissen and Gilson for sets of molecules that interact with the α_{1A} adrenoceptor, and cyclin-dependent kinase, cyclooxygenase-2, factor Xa, and phosphodiesterase-5 [72]. A comparison over ten QSAR datasets was performed to compare the stochastic gradient boosting method (SGB, an ensemble of classification and regression trees) with decision tree, random forest, partial least squares, *k*-NN, Bayes classifier, and SVM [129]. The best results predictions were obtained with random forest, followed by SVM and SGB.

The enzyme-substrate interaction is very specific, which points to the existence of recognition elements located on the enzyme interface. A surface patch ranking (SPR) method was proposed for the identification of those protein residues that determine the substrate specificity [157]. SPR incorporates an SVM module that can highlight the residues important in ligand binding. The method was applied with good results for several homologous enzymes, namely guanylyl/adenylyl cyclases, lactate/malate dehydrogenases, and trypsin/chymotrypsin. Potential applications are residue selection for mutagenesis experiments and functional annotation of proteins. Single base DNA mutations that result in an amino acid substitution are a common cause of monogenic disease. These substitutions reduce the protein stability, as measured by reduction in hydrophobic area, overpacking, backbone strain, and loss of electrostatic interactions [158]. An SVM classifier was trained with a set of mutations causative of disease and another set of non-disease causing mutations. The cross-validation tests show that the SVM identifies 74% of disease mutations, and confirms that loss of protein stability is a determinant factor in monogenic disease.

Comparative Studies

SAR and QSAR models have two major components, namely a set of structural descriptors and a machine learning or ensemble of ML algorithms. The studies reviewed in previous sections show that there is no universal family of descriptors that gives the best predictions for any molecular property, which justifies the continuous development of novel structural descriptors. Similarly, no ML algorithm performs optimally for any property and any set of descriptors. Extensive computational experiments show that for a given property, the best model can be identified only by an empirical comparison of a large number of ML methods. In this section we review several studies that sought to compare the predictive abilities of ML models.

The aqueous solubility of drugs and drug-related compounds was modeled by Schroeter et al with Gaussian process, random forest, SVM, and ridge regression [118]. The domain of applicability for each model was estimate by computing error levels. The aqueous solubility of 988 organic chemicals was modeled with four regression algorithms, namely partial least squares, random forest, SVM, and MLF artificial neural networks [104]. The best predictions were obtained with random forest, which also gave good results for an external validation set of 330 chemicals. The melting temperatures of ionic liquids were modeled with SVM, associative neural networks, *k*-NN, partial least squares, multilayer feedforward (MLF) artificial neural networks, and multiple linear regression [145]. Slightly better results were obtained with SVM, associative neural networks, and MLF artificial neural networks.

The human serum protein binding of 808 compounds was simulated with multiple linear regression, MLF artificial neural networks, *k*-NN, and SVM [148]. An ensemble of MLF artificial neural networks provided the best predictions for an external validation set of 200 compounds. The blood-brain barrier permeability of 415 chemicals was modeled by Li et al with logistic regression, linear discriminate analysis, *k*-NN, C4.5 decision tree, probabilistic neural network, and SVM [89]. It was found that all ML models are improved by using a descriptor set selected with recursive feature elimination. Plewczynski et al compared SAR models for five biological targets obtained with SVM, random forest, artificial neural networks, *k*-NN, trend vectors, Bayesian classifier, and decision tree [107]. There are significant differences in performance depending on the target and objective, i. e., high enrichment or maximum number of actives. The toxicity risk assessment of organic chemicals was evaluated with SVM, *k*-NN, Bayesian classifier, and self-organizing maps (SOM) [147]. Both SOM and SVM can separate toxic from nontoxic compounds based on structural fragments. A new ML algorithm that shows promising results is kScore, which in several tests performed better than SVM, *k*-NN, recursive partitioning, artificial neural networks, Gaussian process, and Bayesian classifier [102].

A comparison of 21 machine learning algorithms is presented for data selected from the National Cancer Institute 60-cell line screening panel (NCI-60) [69]. The SMILES codes of the chemical compounds [130] were used to compute fingerprints with Open Babel (http://openbabel.org/). The descriptor selection (based on CfsSubsetEval and BestFirst) and all machine learning models were computed with Weka [41,151], and all prediction results are for 10-fold (leave-10%-out) cross-validation. The machine learning algorithms used are briefly listed

Drug Design with Machine Learning, Table 2
Machine learning anticancer SAR for lung large cell carcinoma NCI-H460

Exp	Model	TP_p	FN_p	TN_p	FP_p	Ac_p	MCC_p
1	IBk $k = 9W(1-d)$	1386	663	1035	515	0.6727	0.3414
2	IBk $k = 5W(1-d)$	1410	639	1011	539	0.6727	0.3383
3	IBk $k = 9$ NoW	1404	645	1015	535	0.6721	0.3378
4	IBk $k = 7W(1-d)$	1386	663	1027	523	0.6705	0.3364
5	IBk $k = 9W(1/d)$	1399	650	1012	538	0.6699	0.3334
6	IBk $k = 5$ NoW	1433	616	983	567	0.6713	0.3324
7	KStar	1364	685	1034	516	0.6663	0.3299
8	IBk $k = 7$ NoW	1404	645	1001	549	0.6682	0.3290
9	SVM RBF $\sigma = 0.01$	1494	555	926	624	0.6724	0.3286
10	IBk $k = 7W(1/d)$	1396	653	1003	547	0.6666	0.3263
11	IBk $k = 3W(1-d)$	1420	629	982	568	0.6674	0.3252
12	RandomForest $T = 40$	1509	540	902	648	0.6699	0.3217
13	IBk $k = 3$ NoW	1446	603	954	596	0.6669	0.3210
14	IBk $k = 5W(1/d)$	1408	641	985	565	0.6649	0.3210
15	RandomForest $T = 50$	1499	550	903	647	0.6674	0.3171
16	RandomForest $T = 30$	1499	550	902	648	0.6671	0.3164
17	IBk $k = 3W(1/d)$	1428	621	955	595	0.6621	0.3125
18	RandomForest $T = 20$	1487	562	889	661	0.6602	0.3021
19	IBk $k = 1$ NoW	1457	592	911	639	0.6580	0.3000
20	IBk $k = 1W(1/d)$	1457	592	911	639	0.6580	0.3000
21	IBk $k = 1W(1-d)$	1457	592	911	639	0.6580	0.3000
22	RandomForest $T = 10$	1494	555	878	672	0.6591	0.2990
23	JRip	1568	481	810	740	0.6607	0.2972
24	J48	1481	568	884	666	0.6571	0.2959
25	ADTree Ehp	1536	513	836	714	0.6591	0.2956
26	ADTree Ezp	1536	513	836	714	0.6591	0.2956
27	ADTree Erp	1580	469	791	759	0.6588	0.2922
28	NaiveBayes	1130	919	1147	403	0.6327	0.2919
29	NaiveBayesUpdateable	1130	919	1147	403	0.6327	0.2919
30	ADTree Eap	1563	486	806	744	0.6582	0.2919
31	Logistic	1514	535	848	702	0.6563	0.2911
32	BayesNet	1256	793	1054	496	0.6418	0.2903
33	NBTree	1406	643	930	620	0.6491	0.2857
34	RBFNetwork	1578	471	769	781	0.6521	0.2774
35	PART	1404	645	917	633	0.6449	0.2766
36	REPTree	1551	498	785	765	0.6491	0.2723
37	DecisionTable	1549	500	752	798	0.6393	0.2507
38	RandomTree	1384	665	885	665	0.6305	0.2464
39	Ridor	1617	432	642	908	0.6277	0.2201
40	OneR	1745	304	427	1123	0.6035	0.1565
41	ConjunctiveRule	1899	150	163	1387	0.5729	0.0562
42	DecisionStump	2049	0	0	1550	0.5693	0.0000

here, using their notation in Weka: BayesNet, Bayesian network; NaiveBayes, naïve Bayesian classifier [71]; NaiveBayesUpdateable, naïve Bayesian classifier with estimator classes [71]; Logistic, logistic regression with a ridge estimator [86]; RBFNetwork, Gaussian radial basis function network; IBk, k-NN classifier with distance weight (NoW – no distance weight, $W(1/d)$ – weighted with $1/d$, $W(1-d)$ – weighted with $(1-d)$) and k an odd number between 1 and 9 [1]; KStar, K^* lazy learner with entropy-based distance function [26]; ADTree, alternating

Drug Design with Machine Learning, Table 3
Machine learning anticancer SAR for glioma SF-268

Exp	Model	TP_p	FN_p	TN_p	FP_p	Ac_p	MCC_p
1	IBk $k = 9W(1/d)$	1338	682	1177	524	0.6759	0.3530
2	SVM RBF $\sigma = 0.01$	1403	617	1118	583	0.6775	0.3513
3	IBk $k = 7W(1/d)$	1344	676	1165	536	0.6743	0.3490
4	IBk $k = 5W(1/d)$	1351	669	1152	549	0.6727	0.3449
5	IBk $k = 7W(1 - d)$	1325	695	1173	528	0.6713	0.3443
6	IBk $k = 9W(1 - d)$	1301	719	1185	516	0.6681	0.3395
7	IBk $k = 5W(1 - d)$	1330	690	1160	541	0.6692	0.3391
8	IBk $k = 5$ NoW	1372	648	1120	581	0.6697	0.3368
9	IBk $k = 7$ NoW	1353	667	1135	566	0.6686	0.3360
10	KStar	1307	713	1168	533	0.6651	0.3325
11	IBk $k = 3$ NoW	1396	624	1089	612	0.6678	0.3311
12	IBk $k = 3W(1 - d)$	1358	662	1121	580	0.6662	0.3304
13	IBk $k = 9$ NoW	1323	697	1150	551	0.6646	0.3298
14	RandomForest $T = 40$	1408	612	1074	627	0.6670	0.3287
15	RandomForest $T = 50$	1408	612	1072	629	0.6665	0.3275
16	RandomForest $T = 30$	1418	602	1062	639	0.6665	0.3269
17	IBk $k = 3W(1/d)$	1363	657	1108	593	0.6641	0.3254
18	J48	1376	644	1082	619	0.6606	0.3169
19	RandomForest $T = 20$	1403	617	1045	656	0.6579	0.3095
20	Logistic	1404	616	1028	673	0.6536	0.3003
21	NBTree	1240	780	1164	537	0.6461	0.2974
22	RandomForest $T = 10$	1411	609	1015	686	0.6520	0.2965
23	BayesNet	1220	800	1176	525	0.6439	0.2948
24	IBk $k = 1$ NoW	1396	624	1020	681	0.6493	0.2916
25	IBk $k = 1W(1/d)$	1396	624	1020	681	0.6493	0.2916
26	IBk $k = 1W(1 - d)$	1396	624	1020	681	0.6493	0.2916
27	ADTree Erp	1407	613	993	708	0.6450	0.2819
28	RBFNetwork	1519	501	884	817	0.6458	0.2800
29	NaiveBayes	1064	956	1272	429	0.6278	0.2790
30	NaiveBayesUpdateable	1064	956	1272	429	0.6278	0.2790
31	ADTree Ehp	1406	614	986	715	0.6428	0.2774
32	ADTree Ezp	1406	614	986	715	0.6428	0.2774
33	REPTree	1380	640	1009	692	0.6420	0.2771
34	JRip	1418	602	960	741	0.6391	0.2689
35	ADTree Eap	1443	577	936	765	0.6393	0.2684
36	PART	1328	692	1024	677	0.6321	0.2593
37	DecisionTable	1438	582	915	786	0.6324	0.2538
38	RandomTree	1357	663	984	717	0.6291	0.2510
39	Ridor	1705	315	630	1071	0.6275	0.2454
40	DecisionStump	674	1346	1414	287	0.5611	0.1877
41	ConjunctiveRule	665	1355	1419	282	0.5601	0.1869
42	OneR	1257	763	932	769	0.5883	0.1702

decision tree (search type: Eap, expand all paths; Ehp, expand the heaviest path; Ezp, expand the best z-pure path; Erp, expand a random path) [42]; DecisionStump, one-level binary decision tree with categorical or numerical class label; J48, C4.5 decision tree [108]; NBTree, decision tree with naïve Bayes classifiers at the leaves [82]; RandomForest, random forest (the number of random trees is between 10 and 50) [22]; RandomTree, a tree that considers k randomly chosen attributes at each node; REPTree, fast decision tree learner; ConjunctiveRule, conjunc-

Drug Design with Machine Learning, Table 4
Machine learning anticancer SAR for melanoma SK-MEL-5

Exp	Model	TP_p	FN_p	TN_p	FP_p	Ac_p	MCC_p
1	IBk $k = 9W(1-d)$	1379	655	1119	532	0.6779	0.3541
2	IBk $k = 9$ NoW	1401	633	1099	552	0.6784	0.3532
3	IBk $k = 7$ NoW	1422	612	1081	570	0.6792	0.3531
4	IBk $k = 5W(1-d)$	1401	633	1095	556	0.6773	0.3508
5	IBk $k = 7W(1-d)$	1390	644	1104	547	0.6768	0.3506
6	RandomForest $T = 50$	1487	547	1019	632	0.6801	0.3504
7	IBk $k = 7W(1/d)$	1414	620	1079	572	0.6765	0.3479
8	IBk $k = 5W(1/d)$	1417	617	1074	577	0.6760	0.3465
9	IBk $k = 9W(1/d)$	1404	630	1084	567	0.6752	0.3458
10	RandomForest $T = 30$	1478	556	1019	632	0.6776	0.3456
11	IBk $k = 5$ NoW	1427	607	1063	588	0.6757	0.3451
12	RandomForest $T = 20$	1482	552	1013	638	0.6771	0.3443
13	RandomForest $T = 40$	1481	553	1013	638	0.6768	0.3437
14	SVM RBF $\sigma = 0.03$	1414	620	1061	590	0.6716	0.3373
15	KStar	1326	708	1119	532	0.6635	0.3279
16	RandomForest $T = 10$	1481	553	976	675	0.6668	0.3222
17	IBk $k = 3W(1-d)$	1390	644	1053	598	0.6630	0.3204
18	IBk $k = 3$ NoW	1416	618	1025	626	0.6624	0.3171
19	IBk $k = 3W(1/d)$	1393	641	1044	607	0.6613	0.3166
20	IBk $k = 1$ NoW	1427	607	987	664	0.6551	0.3005
21	IBk $k = 1W(1/d)$	1427	607	987	664	0.6551	0.3005
22	IBk $k = 1W(1-d)$	1427	607	987	664	0.6551	0.3005
23	BayesNet	1232	802	1132	519	0.6415	0.2901
24	PART	1348	686	1033	618	0.6461	0.2875
25	RBFNetwork	1541	493	857	794	0.6507	0.2856
26	NBTree	1313	721	1055	596	0.6426	0.2832
27	J48	1378	656	1000	651	0.6453	0.2831
28	Logistic	1426	608	958	693	0.6469	0.2830
29	JRip	1483	551	899	752	0.6464	0.2785
30	RandomTree	1401	633	963	688	0.6415	0.2731
31	NaiveBayes	1068	966	1224	427	0.6220	0.2698
32	NaiveBayesUpdateable	1068	966	1224	427	0.6220	0.2698
33	DecisionTable	1494	540	863	788	0.6396	0.2634
34	REPTree	1431	603	918	733	0.6374	0.2622
35	ADTree Erp	1436	598	889	762	0.6309	0.2478
36	ADTree Ehp	1429	605	892	759	0.6299	0.2459
37	ADTree Ezp	1429	605	892	759	0.6299	0.2459
38	ADTree Eap	1323	711	981	670	0.6252	0.2441
39	Ridor	1469	565	817	834	0.6204	0.2230
40	DecisionStump	1279	755	932	719	0.6000	0.1930
41	OneR	1279	755	932	719	0.6000	0.1930
42	ConjunctiveRule	1404	630	770	881	0.5900	0.1605

tive rule learner; DecisionTable, decision table majority classifier [81]; JRip, a propositional rule learner based on RIPPER [27]; OneR, rule classifier that uses the minimum-error attribute for prediction [56]; PART, a PART decision list that builds a partial C4.5 decision tree in each iteration and transforms the best leaf into a rule ; Ridor, a RIpple–DOwn Rule learner [43]; SVM, LibSVM. More details for each ML may be found in Weka.

The first series of SAR models for anticancer compounds considers the cell line NCI-H460 (lung large cell

Drug Design with Machine Learning, Table 5
Machine learning algorithms with best results in the CoEPrA 2006 (comparative evaluation of prediction algorithms, http://www.coepra.org/) competition

Task	Rank	Machine Learning	Task	Rank	Machine Learning
C1	1	naïve Bayes	R1	1	LS-SVM linear kernel
	2	LS-SVM RBF kernel		2	random forest
	3	CART		3	Gaussian process
	4	C4.5		4	kernel PLS
	5	SVM string kernel		5	SVM RBF kernel
C2	1	SVM linear kernel	R2	1	kernel PLS
	2	kScore		2	kNN
	3	SVM RBF kernel		3	PLS
	4	ClassificationViaRegression		4	SVM string kernel
	5	random forest		5	kScore
C3	1	SVM linear kernel	R3	1	random forest
	2	LS-SVM RBF kernel		2	kernel PLS
	3	SVM		3	SVM string kernel
	4	C4.5		4	PLS
	5	random forest		5	Gaussian process
C4	1	LS-SVM quadratic kernel	R3D2	1	kernel PLS
	2	SVM with AA sequence		2	SVM string kernel
	3	SVM binary encoding of AA		3	random forest
	4	kScore		4	Gaussian process
	5	random forest		5	SVM binary encoding of AA

carcinoma) with 3599 chemicals (2049 in class + 1, 1550 in class −1) and 30 descriptors (Tab. 2). For an easier comparison of the ranking of all ML methods, the results are ordered after the Matthews correlation coefficient MCC. The prediction statistical indices show that *k*-NN, in its several variants evaluated here, constantly gives the best predictions. Three other ML with good predictions are KStar, SVM, and RandomForest.

The second group of structure-activity models is obtained for the cell line SF-268 (glioma) with 3721 chemicals (2020 in class + 1, 1701 in class −1) and 36 descriptors (Tab. 3). A *k*-NN classifier is in the first position, followed by SVM and other *k*-NN variants. In decreasing order of the predictions, the list continues with another lazy learning method, KStar, followed by RandomForest and J48.

The third set of experiments considers the cell line SK-MEL-5 (melanoma) with 3685 chemicals (2034 in class + 1, 1651 in class −1) and 30 descriptors (Tab. 4). All *k*-NN give good predictions, including the top five results. RandomForest ranks better compared with the previous two cell lines, and it is followed by SVM and KStar. As a conclusion of these three sets of experiments, we find that a small group of ML algorithms gives constantly the best predictions. The top results are obtained with *k*-NN, which has good statistics for the whole range of parameters tested here. This is an important result, because *k*-NN is a robust method that is much easier to compute than SVM. Three other ML algorithms have good predictions, namely SVM, KStar, and RandomForest. Although KStar is not used in drug design, the results reported here indicate that this ML algorithm should be added to the toolbox of structure-activity methods. We have to note the absence of the Bayesian classifiers, although they are the preferred method in some library screening experiments.

The studies reviewed in this section show that it is difficult to perform comprehensive comparisons for a large number of machine learning algorithms. To address this issue, we organized a competition for the ML evaluation in blind predictions, CoEPrA 2006 (Comparative Evaluation of Prediction Algorithms, http://www.coepra.org/). For each prediction task the participants received a training dataset (structure, descriptors, and class attribution or activity) and a prediction dataset (structure and descriptors). All predictions received before deadline were compared with the experimental results, and then the ML models were ranked after different statistical indices (Tab. 5). The competition had four classification (C in Tab. 5) tasks and four regression (R in Tab. 5) tasks. All datasets consisted of peptides evaluated for their binding affinity to the major histocompatibility complex (MHC) proteins. To evaluate the general performance of the algorithms tested in competition, for each task we present the ML from the top

five positions. In general, each task has a different winner and order for the top ML. The most frequent ML is SVM, in several variants. The least squares SVM (LS-SVM) is particularly robust, insensitive to noise, and with good performance. Currently, LS-SVM is not used in virtual screening, but the results from the competition show that it should be included in the set of preferred methods. Other predictive ML methods are random forest, Gaussian process, and kScore. The CoEPrA competition offered interesting conclusions with direct implications in drug design, SAR, and QSAR. Similar experiments are necessary to evaluate datasets consisting of drugs and drug-like chemicals.

Future Directions

For a long period of time the machine learning selection available to the computational chemist was restricted to a small number of algorithms. However, recent publications explore a much larger diversity of statistical models, with the intention to help the drug discovery process with reliable predictions. As an example of the current trend, one should consider the fast transition of the support vector machines from small-scale experiments to the preferred method in many industrial applications. Following the SVM success, other kernel methods were adopted in chemoinformatics and drug design. The results reviewed here show that it is not possible to predict which combination of structural descriptors and machine learning will give the most reliable predictions for novel chemicals. Therefore, in order to identify good models, it is always a good idea to compare several types of descriptors and as many ML algorithms as computationally possible. Equally important is a good plan for the experiments, that includes a proper validation and whenever possible tests with external data that were not used in training or cross-validation. Of great importance is an objective and comprehensive evaluation of ML algorithms, which can be performed in settings similar with the CoEPrA competition. Such blind predictions may offer an unbiased ranking of the machine learning algorithms.

Bibliography

1. Aha DW, Kibler D, Albert MK (1991) Instance-based learning algorithms. Mach Learn 6:37–66
2. Ajmani S, Jadhav K, Kulkarni SA (2006) Three-dimensional QSAR using the *k*-nearest neighbor method and its interpretation. J Chem Inf Model 46:24–31
3. Andres C, Hutter MC (2006) CNS permeability of drugs predicted by a decision tree. QSAR Comb Sci 25:305–309
4. Alpaydin E (2004) Introduction to machine learning. MIT Press, Cambridge, p 445
5. Atkeson CG, Moore AW, Schaal S (1997) Locally weighted learning. Artif Intell Rev 11:11–73
6. Atkeson CG, Moore AW, Schaal S (1997) Locally weighted learning for control. Artif Intell Rev 11:75–113
7. Arimoto R, Prasad MA, Gifford EM (2005) Development of CYP3A4 inhibition models: comparisons of machine-learning techniques and molecular descriptors. J Biomol Screen 10:197–205
8. Balaban AT, Ivanciuc O (1999) Historical development of topological indices. In: Devillers J, Balaban AT (eds) Topological indices and related descriptors in QSAR and QSPR. Gordon & Breach Science Publishers, Amsterdam, pp 21–57
9. Basak SC, Grunwald GD (1995) Molecular similarity and estimation of molecular properties. J Chem Inf Comput Sci 35:366–372
10. Basak SC, Bertelsen S, Grunwald GD (1994) Application of graph theoretical parameters in quantifying molecular similarity and structure-activity relationships. J Chem Inf Comput Sci 34:270–276
11. Basak SC, Bertelsen S, Grunwald GD (1995) Use of graph theoretic parameters in risk assessment of chemicals. Toxicol Lett 79:239–250
12. Bayes T (1763) An essay towards solving a problem in the doctrine of chances. Philos Trans Roy Soc London 53:370–418
13. Bender A, Jenkins JL, Glick M, Deng Z, Nettles JH, Davies JW (2006) "Bayes affinity fingerprints" improve retrieval rates in virtual screening and define orthogonal bioactivity space: when are multitarget drugs a feasible concept? J Chem Inf Model 46:2445–2456
14. Bender A, Scheiber J, Glick M, Davies JW, Azzaoui K, Hamon J, Urban L, Whitebread S, Jenkins JL (2007) Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. Chem Med Chem 2:861–873
15. Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin, p 740
16. Bishop CM (1996) Neural networks for pattern recognition. Oxford University Press, Oxford, p 504
17. Boid DB (2007) How computational chemistry became important in the pharmaceutical industry. In: Lipkowitz KB, Cundari TR (eds) Reviews in computational chemistry, vol 23. Wiley, Weinheim, pp 401–451
18. Bonchev D (1983) Information theoretic indices for characterization of chemical structure. Research Studies Press, Chichester
19. Bonchev D, Rouvray DH (eds) (1991) Chemical graph theory. Introduction and fundamentals. Abacus Press/Gordon & Breach Science Publishers, New York
20. Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: Haussler D (ed) Proc of the 5th annual ACM workshop on computational learning theory. ACM Press, Pittsburgh, pp 144–152
21. Bottou L, Chapelle O, DeCoste D, Weston J (2007) Large-scale kernel machines. MIT Press, Cambridge, p 416
22. Breiman L (2001) Random forests. Mach Learn 45:5–32
23. Briem H, Günther J (2005) Classifying "kinase inhibitor-likeness" by using machine-learning methods. Chem Bio Chem 6:558–566
24. Cash GG (1999) Prediction of physicochemical properties from Euclidean distance methods based on electrotopological state indices. Chemosphere 39:2583–2591

25. Chapelle O, Haffner P, Vapnik VN (1999) Support vector machines for histogram-based image classification. IEEE Trans Neural Netw 10:1055–1064
26. Cleary JG, Trigg LE (1995) *K**: an instance-based learner using and entropic distance measure. In: Prieditis A, Russell SJ (eds) Proc of the 12th international conference on machine learning. Morgan Kaufmann, Tahoe City, pp 108–114
27. Cohen WW (1995) Fast effective rule induction. In: Prieditis A, Russell SJ (eds) Proc of the 12th international conference on machine learning. Morgan Kaufmann, Tahoe City, pp 115–123
28. Cortes C, Vapnik V (1995) Support vector networks. Mach Learn 20:273–297
29. Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines. Cambridge University Press, Cambridge
30. Deconinck E, Zhang MH, Coomans D, Vander Heyden Y (2006) Classification tree models for the prediction of blood-brain barrier passage of drugs. J Chem Inf Model 46:1410–1419
31. Deng Z, Chuaqui C, Singh J (2006) Knowledge-based design of target-focused libraries using protein-ligand interaction constraints. J Med Chem 49:490–500
32. Doddareddy MR, Cho YS, Koh HY, Kim DH, Pae AN (2006) In silico renal clearance model using classical Volsurf approach. J Chem Inf Model 46:1312–1320
33. Drucker H, Wu DH, Vapnik VN (1999) Support vector machines for spam categorization. IEEE Trans Neural Netw 10:1048–1054
34. Du H, Wang J, Watzl J, Zhang X, Hu Z (2008) Classification structure-activity relationship (CSAR) studies for prediction of genotoxicity of thiophene derivatives. Toxicol Lett 177: 10–19
35. Duda RO, Hart PE, Stork DG (2000) Pattern classification. 2nd edn. Wiley, New York
36. Ehrman TM, Barlow DJ, Hylands PJ (2007) Virtual screening of chinese herbs with random forest. J Chem Inf Model 47:264–278
37. Eitrich T, Kless A, Druska C, Meyer W, Grotendorst J (2007) Classification of highly unbalanced CYP450 data of drugs using cost sensitive machine learning techniques. J Chem Inf Model 47:92–103
38. Ekins S, Balakin KV, Savchuk N, Ivanenkov Y (2006) Insights for human ether-a-go-go-related gene potassium channel inhibition using recursive partitioning and Kohonen and Sammon mapping techniques. J Med Chem 49:5059–5071
39. Ertl P, Roggo S, Schuffenhauer A (2008) Natural product-likeness score and its application for prioritization of compound libraries. J Chem Inf Model 48:68–74
40. Fatemi MH, Gharaghani S (2007) A novel QSAR model for prediction of apoptosis-inducing activity of 4-aryl-4-H-chromenes based on support vector machine. Bioorg Med Chem 15:7746–7754
41. Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004) Data mining in bioinformatics using Weka. Bioinformatics 20:2479–2481
42. Freund Y, Mason L (1999) The alternating decision tree learning algorithm. In: Bratko I, Dzeroski S (eds) Proc of the 16th international conference on machine learning (ICML (1999)). Morgan Kaufmann, Bled, pp 124–133
43. Gaines BR, Compton P (1995) Induction of ripple-down rules applied to modeling large databases. Intell J Inf Syst 5:211–228
44. Gao JB, Gunn SR, Harris CJ (2003) SVM regression through variational methods and its sequential implementation. Neurocomputing 55:151–167
45. Gao JB, Gunn SR, Harris CJ (2003) Mean field method for the support vector machine regression. Neurocomputing 50:391–405
46. Gepp MM, Hutter MC (2006) Determination of hERG channel blockers using a decision tree. Bioorg Med Chem 14:5325–5332
47. Guha R, Dutta D, Jurs PC, Chen T (2006) Local lazy regression: making use of the neighborhood to improve QSAR predictions. J Chem Inf Model 46:1836–1847
48. Gute BD, Basak SC (2001) Molecular similarity-based estimation of properties: a comparison of three structure spaces. J Mol Graph Modell 20:95–109
49. Gute BD, Basak SC, Mills D, Hawkins DM (2002) Tailored similarity spaces for the prediction of physicochemical properties. Internet Electron J Mol Des 1:374–387
50. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46:389–422
51. Hansch C, Garg R, Kurup A, Mekapati SB (2003) Allosteric interactions and QSAR: on the role of ligand hydrophobicity. Bioorg Med Chem 11:2075–2084
52. Hastie T, Tibshirani R, Friedman JH (2003) The elements of statistical learning. Springer, Berlin, p 552
53. Herbrich R (2002) Learning kernel classifiers. MIT Press, Cambridge
54. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2006) New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. J Chem Inf Model 46:462–470
55. Hoffman B, Cho SJ, Zheng W, Wyrick S, Nichols DE, Mailman RB, Tropsha A (1999) Quantitative structure-activity relationship modeling of dopamine D_1 antagonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and *K*-nearest neighbor methods. J Med Chem 42:3217–3226
56. Holte RC (1993) Very simple classification rules perform well on most commonly used datasets. Mach Learn 11:63–90
57. Hou T, Wang J, Zhang W, Xu X (2007) ADME evaluation in drug discovery. 7. Prediction of oral absorption by correlation and classification. J Chem Inf Model 47:208–218
58. Huang T-M, Kecman V, Kopriva I (2006) Kernel based algorithms for mining huge data sets. Springer, Berlin, p 260
59. Hudelson MG, Ketkar NS, Holder LB, Carlson TJ, Peng C-C, Waldher BJ, Jones JP (2008) High confidence predictions of drug-drug interactions: predicting affinities for cytochrome P450 2C9 with multiple computational methods. J Med Chem 51:648–654
60. Itskowitz P, Tropsha A (2005) *k*-nearest neighbors QSAR modeling as a variational problem: theory and applications. J Chem Inf Model 45:777–785
61. Ivanciuc O (2002) Support vector machine classification of the carcinogenic activity of polycyclic aromatic hydrocarbons. Internet Electron J Mol Des 1:203–218
62. Ivanciuc O (2002) Structure-odor relationships for pyrazines

with support vector machines. Internet Electron J Mol Des 1:269–284
63. Ivanciuc O (2002) Support vector machine identification of the aquatic toxicity mechanism of organic compounds. Internet Electron J Mol Des 1:157–172
64. Ivanciuc O (2003) Graph theory in chemistry. In: Gasteiger J (ed) Handbook of chemoinformatics, vol 1. Wiley, Weinheim, pp 103–138
65. Ivanciuc O (2003) Topological indices. In: Gasteiger J (ed) Handbook of chemoinformatics, vol 3. Wiley, Weinheim, pp 981–1003
66. Ivanciuc O (2003) Aquatic toxicity prediction for polar and nonpolar narcotic pollutants with support vector machines. Internet Electron J Mol Des 2:195–208
67. Ivanciuc O (2004) Support vector machines prediction of the mechanism of toxic action from hydrophobicity and experimental toxicity against pimephales promelas and tetrahymena pyriformis. Internet Electron J Mol Des 3:802–821
68. Ivanciuc O (2005) Support vector regression quantitative structure-activity relationships (QSAR) for benzodiazepine receptor ligands. Internet Electron J Mol Des 4:181–193
69. Ivanciuc O (2005) Machine learning applied to anticancer structure-activity relationships for NCI human tumor cell lines. Internet Electron J Mol Des 4:948–958
70. Ivanciuc O (2007) Applications of support vector machines in chemistry. In: Lipkowitz KB, Cundari TR (eds) Reviews in computational chemistry, vol 23. Wiley, Weinheim, pp 291–400
71. John GH, Langley P (1995) Estimating continuous distributions in Bayesian classifiers. In: Besnard P, Hanks S (eds) UAI '95: Proc of the 11th annual conference on uncertainty in artificial intelligence. Morgan Kaufmann, Montreal, pp 338–345
72. Jorissen RN, Gilson MK (2005) Virtual screening of molecular databases using a support vector machine. J Chem Inf Model 45:549–561
73. Jurs P (2003) Quantitative structure-property relationships. In: Gasteiger J (ed) Handbook of chemoinformatics, vol 3. Wiley, Weinheim, pp 1314–1335
74. Kier LB, Hall LH (1976) Molecular connectivity in chemistry and drug research. Academic Press, New York
75. Kier LB, Hall LH (1986) Molecular connectivity in structure-activity analysis. Research Studies Press, Letchworth
76. Kier LB, Hall LH (1999) Molecular structure description. The electrotopological state. Academic Press, San Diego
77. Klon AE, Diller DJ (2007) Library fingerprints: a novel approach to the screening of virtual libraries. J Chem Inf Model 47:1354–1365
78. Klon AE, Glick M, Davies JW (2004) Combination of a naive Bayes classifier with consensus scoring improves enrichment of high-throughput docking results. J Med Chem 47:4356–4359
79. Klon AE, Glick M, Thoma M, Acklin P, Davies JW (2004) Finding more needles in the haystack: a simple and efficient method for improving high-throughput docking results. J Med Chem 47:2743–2749
80. Klon AE, Lowrie JF, Diller DJ (2006) Improved naïve Bayesian modeling of numerical data for absorption, distribution, metabolism and excretion (ADME) property prediction. J Chem Inf Model 46:1945–1956
81. Kohavi R (1995) The power of decision tables. In: Lavrac N, Wrobel S (eds) ECML-95 8th european conference on machine learning. Lecture Notes in Computer Science, vol 912. Springer, Heraclion, pp 174–189
82. Kohavi R (1996) Scaling up the accuracy of naive-Bayes classifiers: a decision-tree hybrid. In: Simoudis E, Han J, Fayyad UM (eds) Proc of the 2nd international conference on knowledge discovery and data mining (KDD-96). AAAI Press, Menlo Park, pp 202–207
83. Kononenko I, Kukar M (2007) Machine learning and data mining: introduction to principles and algorithms. Horwood, Westergate, p 454
84. Konovalov DA, Coomans D, Deconinck E, Vander Heyden Y (2007) Benchmarking of QSAR models for blood-brain barrier permeation. J Chem Inf Model 47:1648–1656
85. Kumar R, Kulkarni A, Jayaraman VK, Kulkarni BD (2004) Structure-activity relationships using locally linear embedding assisted by support vector and lazy learning regressors. Internet Electron J Mol Des 3:118–133
86. le Cessie S, van Houwelingen JC (1992) Ridge estimators in logistic regression. Appl Statist 41:191–201
87. Leong MK (2007) A novel approach using pharmacophore ensemble/support vector machine (PhE/SVM) for prediction of hERG liability. Chem Res Toxicol 20:217–226
88. Lepp Z, Kinoshita T, Chuman H (2006) Screening for new antidepressant leads of multiple activities by support vector machines. J Chem Inf Model 46:158–167
89. Li H, Yap CW, Ung CY, Xue Y, Cao ZW, Chen YZ (2005) Effect of selection of molecular descriptors on the prediction of blood-brain barrier penetrating and nonpenetrating agents by statistical learning methods. J Chem Inf Model 45:1376–1384
90. Li S, Fedorowicz A, Singh H, Soderholm SC (2005) Application of the random forest method in studies of local lymph node assay based skin sensitization data. J Chem Inf Model 45:952–964
91. Li W-X, Li L, Eksterowicz J, Ling XB, Cardozo M (2007) Significance analysis and multiple pharmacophore models for differentiating *P*-glycoprotein substrates. J Chem Inf Model 47:2429–2438
92. Liao Q, Yao J, Yuan S (2007) Prediction of mutagenic toxicity by combination of recursive partitioning and support vector machines. Mol Divers 11:59–72
93. Mangasarian OL, Musicant DR (2000) Robust linear and support vector regression. IEEE Trans Pattern Anal Mach Intell 22:950–955
94. Mangasarian OL, Musicant DR (2002) Large scale kernel regression via linear programming. Mach Learn 46:255–269
95. Merkwirth C, Mauser HA, Schulz-Gasch T, Roche O, Stahl M, Lengauer T (2004) Ensemble methods for classification in cheminformatics. J Chem Inf Comput Sci 44:1971–1978
96. Mitchell TM (1997) Machine learning. McGraw-Hill, Maidenhead, p 432
97. Müller K-R, Rätsch G, Sonnenburg S, Mika S, Grimm M, Heinrich N (2005) Classifying 'drug-likeness' with kernel-based learning methods. J Chem Inf Model 45:249–253
98. Neugebauer A, Hartmann RW, Klein CD (2007) Prediction of protein-protein interaction inhibitors by chemoinformatics and machine learning methods. J Med Chem 50:4665–4668
99. Neumann D, Kohlbacher O, Merkwirth C, Lengauer T (2006) A fully computational model for predicting percutaneous drug absorption. J Chem Inf Model 46:424–429
100. Nidhi, Glick M, Davies JW, Jenkins JL (2006) Prediction

of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. J Chem Inf Model 46:1124–1133
101. Nigsch F, Bender A, van Buuren B, Tissen J, Nigsch E, Mitchell JBO (2006) Melting point prediction employing *k*-nearest neighbor algorithms and genetic parameter optimization. J Chem Inf Model 46:2412–2422
102. Oloff S, Muegge I (2007) kScore: a novel machine learning approach that is not dependent on the data structure of the training set. J Comput-Aided Mol Des 21:87–95
103. Oloff S, Zhang S, Sukumar N, Breneman C, Tropsha A (2006) Chemometric analysis of ligand receptor complementarity: Identifying complementary ligands based on receptor information (CoLiBRI). J Chem Inf Model 46:844–851
104. Palmer DS, O'Boyle NM, Glen RC, Mitchell JBO (2007) Random forest models to predict aqueous solubility. J Chem Inf Model 47:150–158
105. Pelletier DJ, Gehlhaar D, Tilloy-Ellul A, Johnson TO, Greene N (2007) Evaluation of a published in silico model and construction of a novel Bayesian model for predicting phospholipidosis inducing potential. J Chem Inf Model 47:1196–1205
106. Platt J (1999) Fast training of support vector machines using sequential minimal optimization. In: Schölkopf B, Burges CJC, Smola AJ (eds) Advances in kernel methods – support vector learning. MIT Press, Cambridge, pp 185–208
107. Plewczynski D, Spieser SAH, Koch U (2006) Assessing different classification methods for virtual screening. J Chem Inf Model 46:1098–1106
108. Quinlan R (1993) C4.5: programs for machine learning. Morgan Kaufmann, San Mateo
109. Ren S (2002) Classifying class I and class II compounds by hydrophobicity and hydrogen bonding descriptors. Environ Toxicol 17:415–423
110. Ripley BD (2008) Pattern recognition and neural networks. Cambridge University Press, Cambridge, p 416
111. Rodgers S, Glen RC, Bender A (2006) Characterizing bitterness: identification of key structural features and development of a classification model. J Chem Inf Model 46:569–576
112. Rusinko A, Farmen MW, Lambert CG, Brown PL, Young SS (1999) Analysis of a large structure/biological activity data set using recursive partitioning. J Chem Inf Comput Sci 39:1017–1026
113. Sakiyama Y, Yuki H, Moriya T, Hattori K, Suzuki M, Shimada K, Honma T (2008) Predicting human liver microsomal stability with machine learning techniques. J Mol Graph Modell 26:907–915
114. Schneider N, Jäckels C, Andres C, Hutter MC (2008) Gradual in silico filtering for druglike substances. J Chem Inf Model 48:613–628
115. Schölkopf B, Smola AJ (2002) Learning with kernels. MIT Press, Cambridge
116. Schölkopf B, Sung KK, Burges CJC, Girosi F, Niyogi P, Poggio T, Vapnik V (1997) Comparing support vector machines with gaussian kernels to radial basis function classifiers. IEEE Trans Signal Process 45:2758–2765
117. Schölkopf B, Burges CJC, Smola AJ (1999) Advances in kernel methods: support vector learning. MIT Press, Cambridge
118. Schroeter TS, Schwaighofer A, Mika S, ter Laak A, Suelzle D, Ganzer U, Heinrich N, Müller K-R (2007) Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules. J Comput-Aided Mol Des 21:485–498
119. Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University Press, Cambridge
120. Shen M, LeTiran A, Xiao Y, Golbraikh A, Kohn H, Tropsha A (2002) Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using *k*-nearest neighbor and simulated annealing PLS methods. J Med Chem 45:2811–2823
121. Shen M, Xiao Y, Golbraikh A, Gombar VK, Tropsha A (2003) Development and validation of *k*-nearest-neighbor QSPR models of metabolic stability of drug candidates. J Med Chem 46:3013–3020
122. Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 14:199–222
123. Sommer S, Kramer S (2007) Three data mining techniques to improve lazy structure-activity relationships for noncongeneric compounds. J Chem Inf Model 47:2035–2043
124. Sorich MJ, McKinnon RA, Miners JO, Smith PA (2006) The importance of local chemical structure for chemical metabolism by human uridine 5′-diphosphate-glucuronosyltransferase. J Chem Inf Model 46:2692–2697
125. Sun H (2005) A naive Bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. J Med Chem 48:4031–4039
126. Suykens JAK (2001) Support vector machines: a nonlinear modelling and control perspective. Eur J Control 7:311–327
127. Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J (2002) Least squares support vector machines. World Scientific, Singapore
128. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 43:1947–1958
129. Svetnik V, Wang T, Tong C, A Liaw, Sheridan RP, Song Q (2005) Boosting: an ensemble learning tool for compound classification and QSAR modeling. J Chem Inf Model 45:786–799
130. Swamidass SJ, Chen J, Phung P, Ralaivola L, Baldi P (2005) Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. Bioinformatics 21[S1]:i359–i368
131. Terfloth L, Bienfait B, Gasteiger J (2007) Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates. J Chem Inf Model 47:1688–1701
132. Tobita M, Nishikawa T, Nagashima R (2005) A discriminant model constructed by the support vector machine method for HERG potassium channel inhibitors. Bioorg Med Chem Lett 15:2886–2890
133. Todeschini R, Consonni V (2003) Descriptors from molecular geometry. In: Gasteiger J (ed) Handbook of chemoinformatics, vol 3. Wiley, Weinheim, pp 1004–1033
134. Tong W, Hong H, Fang H, Xie Q, Perkins R (2003) Decision forest: Combining the predictions of multiple independent decision tree models. J Chem Inf Comput Sci 43:525–531
135. Tong W, Xie Q, Hong H, Shi L, Fang H, Perkins R (2004) Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. Env Health Perspect 112:1249–1254
136. Trinajstić N (1992) Chemical graph theory. CRC Press, Boca Raton

137. Urrestarazu Ramos E, Vaes WHJ, Verhaar HJM, Hermens JLM (1998) Quantitative structure-activity relationships for the aquatic toxicity of polar and nonpolar narcotic pollutants. J Chem Inf Comput Sci 38:845–852
138. Vapnik VN (1979) Estimation of dependencies based on empirical data. Nauka, Moscow
139. Vapnik VN (1995) The nature of statistical learning theory. Springer, New York
140. Vapnik VN (1998) Statistical learning theory. Wiley, New York
141. Vapnik VN (1999) An overview of statistical learning theory. IEEE Trans Neural Netw 10:988–999
142. Vapnik V, Chapelle O (2000) Bounds on error expectation for support vector machines. Neural Comput 12:2013–2036
143. Vapnik VN, Chervonenkis AY (1974) Theory of pattern recognition. Nauka, Moscow
144. Vapnik V, Lerner A (1963) Pattern recognition using generalized portrait method. Automat Remote Control 24:774–780
145. Varnek A, Kireeva N, Tetko IV, Baskin II, Solov'ev VP (2007) Exhaustive QSPR studies of a large diverse set of ionic liquids: how accurately can we predict melting points? J Chem Inf Model 47:1111–1122
146. Vogt M, Bajorath J (2008) Bayesian similarity searching in high-dimensional descriptor spaces combined with Kullback–Leibler descriptor divergence analysis. J Chem Inf Model 48:247–255
147. von Korff M, Sander T (2006) Toxicity-indicating structural patterns. J Chem Inf Model 46:536–544
148. Votano JR, Parham M, Hall LM, Hall LH, Kier LB, Oloff S, Tropsha A (2006) QSAR modeling of human serum protein binding with several modeling techniques utilizing structure-information representation. J Med Chem 49:7169–7181
149. Wang J, Du H, Yao X, Hu Z (2007) Using classification structure pharmacokinetic relationship (SCPR) method to predict drug bioavailability based on grid-search support vector machine. Anal Chim Acta 601:156–163
150. Watson P (2008) Naïve Bayes classification using 2D pharmacophore feature triplet vectors. J Chem Inf Model 48:166–178
151. Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco, p 525
152. Xiao Z, Xiao Y-D, Feng J, Golbraikh A, Tropsha A, Lee K-H (2002) Antitumor agents. 213. Modeling of epipodophyllotoxin derivatives using variable selection *k*-nearest neighbor QSAR method. J Med Chem 45:2294–2309
153. Xue Y, Li ZR, Yap CW, Sun LZ, Chen X, Chen YZ (2004) Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. J Chem Inf Comput Sci 44:1630–1638
154. Yamashita F, Hara H, Ito T, Hashida M (2008) Novel hierarchical classification and visualization method for multiobjective optimization of drug properties: application to structure-activity relationship analysis of cytochrome P450 metabolism. J Chem Inf Model 48:364–369
155. Yap CW, Chen YZ (2005) Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. J Chem Inf Model 45:982–992
156. Yap CW, Cai CZ, Xue Y, Chen YZ (2004) Prediction of torsade-causing potential of drugs by support vector machine approach. Toxicol Sci 79:170–177
157. Yu G-X, Park B-H, Chandramohan P, Munavalli R, Geist A, Samatova NF (2005) In silico discovery of enzyme-substrate specificity-determining residue clusters. J Mol Biol 352:1105–1117
158. Yue P, Li Z, Moult J (2005) Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol 353:459–473
159. Zhang S, Golbraikh A, Oloff S, Kohn H, Tropsha A (2006) A novel automated lazy learning QSAR (ALL-QSAR) approach: method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models. J Chem Inf Model 46:1984–1995
160. Zhang S, Golbraikh A, Tropsha A (2006) Development of quantitative structure-binding affinity relationship models based on novel geometrical chemical descriptors of the protein-ligand interfaces. J Med Chem 49:2713–2724
161. Zheng WF, Tropsha A (2000) Novel variable selection quantitative structure-property relationship approach based on the *k*-nearest-neighbor principle. J Chem Inf Comput Sci 40:185–194

Drug Design, Molecular Descriptors in

Alexandru T. Balaban
Texas A&M University, Galveston, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Molecular Properties and Biological Activities
Linear Free Energy Relationships, QSAR and QSPR
Molecular Descriptors
Graph Theoretical Intermezzo
Topological Indices (TIs)
A Real-World Example of Using Molecular Descriptors Including TIs
Three-Dimensional Molecular Descriptors
Molecular Similarity, Dissimilarity, and Clustering of Molecular Descriptors
Hierarchical Ordering of Structures and of Molecular Descriptors
General Problems Connected with 2D-QSAR for Drug Design
Interpretation of Topological Indices
Reverse Engineering of Molecular Descriptors
Future Directions
Bibliography

Glossary

ADMET Absorption, distribution, metabolism, excretion, and toxicity, five factors which determine whether a compound with favorable features resulting from

a computer-assisted drug design will also be a useful medicinal drug.

Local vertex invariant (LOVI) A number associated with a vertex of a molecular graph. It does not depend on the arbitrary numerical labeling of vertices.

Molecular descriptor A mathematical entity (either a number, or a set of numbers characterizing substituents or molecular fragments) associated with a chemical structure, allowing quantitative manipulations of such structures (correlations with properties, clustering, ordering or partial ordering, determination of similarity or dissimilarity). Constitutional isomerism is associated with topological indices or other two-dimensional descriptors, whereas stereoisomerism requires three-dimensional descriptors.

Molecular graph Constitutional chemical formula representing atoms by points (vertices) and covalent bonds by lines (edges). Hydrogen atoms are usually omitted (in hydrogen-depleted graphs) and carbon atoms are not shown explicitly, but other atoms are assigned weights according to the nature of the heteroatom. Multiple bonds are shown explicitly in multigraphs.

Pharmacophore A set of chemical features that determine the biological activity.

Quantitative structure-activity relationship (QSAR) Mathematical correlation (e. g. a mono- or multiparametric equation) between a physical or a chemical property and molecular descriptor(s).

Receptor Physiological macromolecule (usually a protein, but sometimes DNA or RNA), also called target, to which a drug (ligand) binds.

Quantitative structure-property relationship (QSPR) Mathematical correlation (e. g. a mono- or multiparametric equation) between a biological property and molecular descriptor(s).

Quantitative structure-toxicity relationship (QSTR) Mathematical correlation (e. g. a mono- or multiparametric equation) between toxicologic properties and molecular descriptor(s).

Topological index (TI) A number characterizing a molecular constitutional graph, resulting from mathematical operations on the LOVIs, or edges/vertices of the graph.

Definition of the Subject

Empirical observations of association between chemical structures and their physical, chemical, or biological properties have a long history [1], but the real development of mathematical correlations started in the 20th century. For physico-chemical properties, topological indices and their applications were first published by Wiener in 1947 followed soon by Platt and Hosoya, as will be shown later. Quantitative structure-activity relationships (QSARs) have 1962 as official birthdates with the connection between Hammett substituent constants, lipophilicity, and biological activities, as will be discussed below.

The huge development in analytical and synthetic methods, leading to combinatorial chemistry, high-throughput screening, and virtual screening (ligand- or structure-based) on one hand, and on the other hand to the progress in computational hardware and software that allows the processing of millions of virtual structure libraries, have made it possible to harness and combine the power of these workhorses for exploring the infinity of "druglike" structures in order to extract information for drug design. Whereas high-throughput screening detects biologically active compounds from a mixture of existing molecules produced by combinatorial synthesis, virtual screening uses computer programs simulating imagined structures. Virtual screening is of two types: ligand-based when the receptor's data are not known, or structure-based when the target structure is available, especially at high atomic resolution.

Introduction

There are four main reasons why new medicinal drugs have to be continuously explored, tested, and brought to the market: (i) bacterial resistance, exemplified nowadays by the far lower success of older antibiotics as compared to the situation when they were first introduced; (ii) after completion of the Human Genome Project (with the corollaries of genomics, proteomics, and metabolomics) it has become possible to design molecules that interact with specific proteins, alleviating or curing ailments due either to the lack or to the abundance of these proteins; (iii) the increased life span brought to prominence many diseases characteristic of old age, that need new treatments and new types of drugs; and (iv) it may be possible to replace defective proteins by gene therapy, which needs either modified viral vectors or non-viral vectors (cationic lipids) – in both cases by using molecular design [2,3].

The traditional manner in which a new drug was developed consisted in uncovering a lead compound whose molecular structure was then systematically varied by synthetic chemists, followed in each case by sequential testing in vitro (i. e. in cell cultures), in vivo (on lower and then on higher organisms), and finally by clinical testing on humans. An exponential decrease in the number of drug candidates follows each of these steps, but the final cost

still attains staggering proportions – nowadays for each new drug reaching the market, the "big pharma" companies invest hundreds of millions of dollars. It is interesting to note that only half a dozen countries (USA, UK, France, Germany, Switzerland, and Sweden) account for most of the new medicinal drugs that are launched on the market. For lowering the cost and time needed for new drugs, these companies must resort to such strategies as mergers, outsourcing, high-throughput screening of thousands of compounds synthesized in one batch by robots using combinatorial chemistry approaches, and especially by preceding the sequence "chemical laboratory – in vitro – in vivo" with a novel strategy, aptly called "in silico". Taking into account that patents expire after 20 years and that after this interval generic drug manufacturers (located mainly in Asia) are able to sell such drugs at a much lower price, which make them more convenient for insurance companies and for the public at large, it is understandable why new drugs cost so much: the "big pharma" companies, which are practically the only ones investing in drug discovery and testing, must recover their investment in about ten years (this is the lifetime of a new drug from the date of FDA approval till generic drugs appear on the market). A medicinal drug that is withdrawn from the market causes enormous financial losses, also due to court litigations. Two relevant papers are mentioned in this regard [4,5].

Molecular Properties and Biological Activities

A first classification of molecular properties is whether such properties depend on individual molecules in vacuum (e. g. heats of formation, molecular weights, molecular dimensions) or on molecules interacting among themselves in bulk or condensed phases (e. g. normal boiling points, critical pressure and temperature, solubility and partition coefficients) [6]. A second classification is according to the type of properties (physical, chemical, or biological), and of course subdivisions for each of these properties (for instance toxicity, bioactivity in interaction with certain enzymes for certain cell cultures or animals).

All properties are measured and expressed on numerical scales, achieving a metric and ordering of properties. In drug design there is not much incentive in establishing correlations between various properties because one wishes to explore structures that have not yet been prepared. On the other hand, chemical structures are discrete entities and in order to establish a structure-property correlation one has to associate numbers with structural formulas. A simple way to do this is to translate structures into molecular invariants, which can be topological indices [7,8,9,10,11,12,13,14] or assemblies of molecular fragments (substructures). *The present review will concentrate on the former approach.*

All elementary textbooks of Organic Chemistry mention that with increasing *branching*, the normal boiling points (NBPs) of isomeric alkanes decrease whereas the Research Octane Numbers (RONs) increase. The former (NBP) is a molecular physical property, and the latter (RON) is a molecular chemical property. Both are bulk properties that can be measured precisely and accurately (unlike biological activities which are usually measured with poorer reproducibility). On examining the data presented in Table 1 for all isomers of heptane and octane one sees, however, that the intercorrelation factor for RON vs. NBP is $r^2 = 0.6454$ for heptanes and only 0.2068 for octanes; for the combined set it is even lower. On sorting each of these two sets of isomeric alkanes according to decreasing NBP and increasing RON, one obtains different orderings. Therefore "*branching*" means something different for the NBP and for the RON properties, as it was stressed in the first sentence of this paragraph.

On the other hand, a shape parameter such as index J (to be discussed later) presents better correlation factors, as seen in Fig. 1. One cannot deny a certain arbitrariness in the experimentally defined number RON as the amount k (%) of a binary mixture of k parts of 2,2,4-trimethylpentane or isooctane and 100 – k parts of n-heptane that develops the same traction in a standard spark-initiated engine as the gasoline being tested. This number then defines a certain operational type of branching, just as there is another type of branching also defined operationally by dividing into 100 parts (called Kelvin or Celsius degrees) the temperature interval between the phase transitions of water under normal pressure, and comparing the boiling temperatures of various substances at normal pressure on these temperature scales. The conclusion is that for molecular branching one needs many molecular descriptors that can account for the corresponding types of branching defined by operational experiments. And this is also valid for other fuzzy properties such as *cyclicity*, *centricity*, or *complexity*. Yet these notions are deeply ingrained and familiar to all of us, and they serve for allowing us to ascribe to objects a certain metric. If we can also establish a numerical scale for such notions, then we can find quantitative structure-property relationships (QSPR for physical-chemical properties) or quantitative structure-activity relationships (QSAR for biological activities). Of course, in some cases the ordering induced by such a numerical scale is only a partial order.

Drug Design, Molecular Descriptors in, Table 1

Six molecular descriptors and two physico-chemical properties (research octane numbers, and normal boiling points, in °C) of all C_7H_{16} and C_8H_{18} alkanes

Alkanes sorted by *J*	Path code	*n*	*W*	*Z*	*J*	*Q*	$^1\chi$	*P*	RON	NBP
Heptane	6,5,4,3,2,1	7	56	2 1	2.447 5	91	3.4142	7.09205	0	98.4
2-Methylhexane	6,6,4,3,2	7	52	1 8	2.678 3	101	3.2701	6.83472	42.4	90.1
3-Methylhexane	6,6,5,3,1	7	50	1 9	2.831 8	107	3.3081	6.78577	52.0	91.9
2,4-Dimethylpentane	6,7,4,4	7	**48**	1 5	2.953 2	**117**	3.1259	6.48502	83.1	79.2
3-Ethylpentane	6,6,6,3	7	**48**	2 0	2.992 3	**117**	3.3461	6.47501	65.0	89.8
2,3-Dimethylpentane	6,7,6,2	7	**46**	1 7	3.144 2	**125**	3.1807	6.47021	91.1	80.5
2,2-Dimethylpentane	6,8,4,3	7	**46**	1 4	3.154 5	**125**	3.0607	6.46178	92.8	86.0
3,3-Dimethylpentane	6,8,6,1	7	44	1 6	3.360 4	137	3.1213	6.44163	80.8	93.5
2,2,3-Trimethylbutane	6,9,6	7	42	1 3	3.541 2	153	2.9434	5.98502	112.1	80.9
Octane	7,6,5,4,3,2,1	8	84	3 4	2.530 1	140	2.5301	8.3987	−19.0	125.7
2-Methylheptane	7,7,5,4,3,2	8	79	2 9	2.715 8	152	2.7158	8.15952	21.7	117.7
3-Methylheptane	7,7,6,4,3,1	8	76	3 1	2.862 1	160	2.8621	8.11363	26.8	117.0
4-Methylheptane	7,7,6,5,2,1	8	75	3 0	2.919 6	164	2.9196	8.08953	26.7	117.7
2,5-Dimethylhexane	7,8,5,4,4	8	74	2 5	2.927 8	170	2.9278	7.83266	55.5	106.8
3-Ethylhexane	7,7,7,5,2	8	72	3 2	3.074 4	176	3.0744	7.83117	33.5	114.0
2,4-Dimethylhexane	7,8,6,5,2	8	**71**	2 6	3.098 8	178	3.0988	7.81045	65.2	109.0
2,2-Dimethylhexane	7,9,5,4,3	8	**71**	2 3	3.111 8	180	3.1118	7.80573	72.5	109.0
2,3-Dimethylhexane	7,8,7,4,2	8	70	2 7	3.170 8	182	3.1708	7.79459	71.3	112.0
3,4-Dimethylhexane	7,8,8,4,1	8	68	2 9	3.292 5	194	3.2925	7.74181	76.3	117.7
3,3-Dimethylhexane	7,9,7,4,1	8	**67**	2 5	3.354 9	196	3.3734	7.72595	87.3	109.8
3-Ethyl-2-methylpentane	7,8,8,5	8	**67**	2 8	3.373 4	202	3.3549	7.40006	75.5	118.0
2,2,4-Trimethylpentane	7,10,5,6	8	66	1 9	3.388 9	**210**	3.3889	7.39752	100.0	99.2
2,3,4-Trimethylpentane	7,9,8,4	8	65	2 4	3.464 2	**210**	3.4642	7.39677	102.7	114.8
3-Ethyl-3-methylpentane	7,9,9,3	8	64	2 8	3.583 2	220	3.5832	7.38083	80.8	113.5
2,2,3-Trimethylpentane	7,10,8,3	8	63	2 2	3.623 3	222	3.6233	7.36514	109.6	115.6
2,3,3-Trimethylpentane	7,10,9,2	8	62	2 3	3.708 3	234	3.7083	7.32097	106.1	118.3
2,2,3,3-Tetramethylbutane	7,12,9	8	58	1 7	4.020 4	274	4.0240	6.82729	> 100	106.5

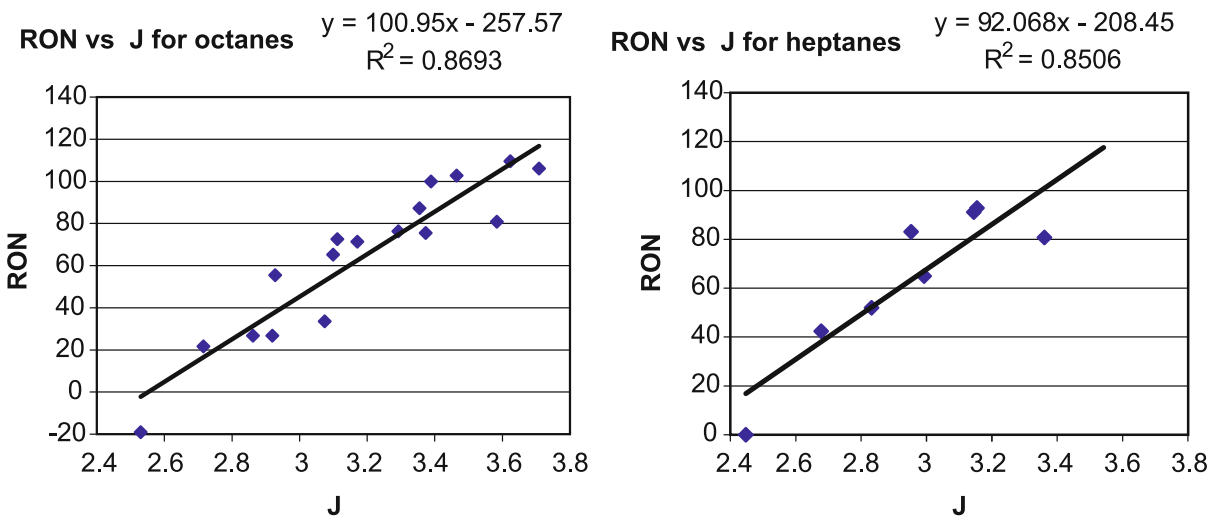

Drug Design, Molecular Descriptors in, Figure 1

Plot of Research Octane Numbers (RON) vs. index *J* for C_7H_{16} and C_8H_{18} alkanes

Linear Free Energy Relationships, QSAR and QSPR

Louis P. Hammett published his influential book *Physical Organic Chemistry* in 1940, but he wrote several other reviews earlier on correlations based on electronic effects of substituents in benzene derivatives, showing how chemical structure determined reaction rates [15,16]. He proved that chemical acid-base equilibria (ionization constants of *meta*- or *para*-substituted benzoic acids) can be used to correlate and predict rates of reactions involving benzylic side-chains influenced by the same substituents in *meta*- or *para*-positions. The Hammett equation (1) links specific rates (k, also called rate constants) with equilibrium constants K in terms of Hammett electronic constants σ and reaction constants ρ for unsubstituted (k_H, K_H) and substituted benzene derivatives (k_X, K_X). The σ_X constants are determined for *meta*- or *para*-substituents from pK_a values of X-substituted benzoic acids; electron-donor substituents have $\sigma < 0$, and electron-acceptor substituents have $\sigma > 0$. Steric effects in *ortho*-positions complicate the situation, but Taft introduced steric parameters that can remedy this problem.

$$\rho \times \sigma = \log(k_X/k_H) \quad \text{with} \quad \sigma = \log(K_X/K_H) \quad \text{in water at } 25\,^\circ\text{C}\,. \tag{1}$$

Not only kinetic data, but also various physical, chemical, and biological activity data obey such linear free energy relationships (LFER). In the early 60s, Corwin Hansch, Albert Leo and T. Fujita proposed a similar LFER Eq. (3) to account for the partition coefficient between water and amphiphilic solvents (hydrophobicity, denoted by π, a structure-dependent physical property) [17,18,19,20].

$$\pi = [C_{org}]/[C_{aq}] \tag{2}$$

where $[C_{org}]$ and $[C_{aq}]$ are molar concentrations of the solute at equilibrium in the organic and aqueous phase, respectively. This was the first approach to modern Quantitative Structure-Property Relationships (QSPR) and it continued with Quantitative Structure-Activity Relationships (QSAR) involving biological activities. By choosing as solvent 1-octanol, the lipophilicity π becomes $P = K_{ow}$ and is determined experimentally as

$$\log \pi = \log P = \log[C_o]/[C_w] = \log[C_{org}] - \log[C_{aq}]\,. \tag{3}$$

Compounds with $\log P > 0$ are lipophilic and those with $\log P < 0$ are lipophobic. Several reviews have been published by Hansch in Chem. Rev. drawing attention to the database of about 10^4 organic reactions with known Hammett ρ data, and a similar number of QSAR data for various biological activities; combining these two treasure troves may offer insight on the mechanism of each biological activity [22,23]. By analogy with the Hammett equation, lipophilicity values can be calculated using a variety of approaches, among which only a few will be mentioned [23,24,25,26].

(i) The Hansch–Fujita hydrophobic substituent constants

$$\pi_X = (1/\rho)\log(P_X/P_H)\,, \tag{4}$$

where P_X and P_H are the partition coefficients for an X-substituted and unsubstituted compound, respectively. For the 1-octanol/water system, ρ is assumed to be 1. By summing up for all i substituents in the molecule, one obtains the lipophilicity of any compound A:

$$\log P_A = \log P_H + \sum_i \pi_{X,i} \tag{5}$$

(ii) The Leo–Hansch hydrophobic fragmental constants (calculated log P, CLOGP) can be obtained by means of a computer program or from the Hansch and Leo's books [24,25].

(iii) The Bodor hydrophobic model (BLOGP) is a nonlinear 18-parameter model based on 10 molecular descriptors such as molecular weight, number of carbon atoms, ovality index based on optimized geometry, charges on oxygen and nitrogen atoms.

(iv) The Gombar hydrophobic model (VLOGP) uses 363 molecular descriptors based on molecular topology (vide infra) such as Kier shape indices and electrotopological indices.

(v) The topological index of hydrophobicity is based on the Randić molecular connectivity (v. i.).

In many cases, the biological activity is a linear or nonlinear function of the lipophilicity, meaning that the determining step is the penetration of the substance through the cell bilayer membrane.

Other molecular descriptors that are being used in drug design involve indicators for: hydrogen bond donors; hydrogen bond acceptors; electrical charges for atoms; conformational flexibility for rotatable chemical bonds; topological surface area; and various quantum-chemical indices, which will be discussed in other chapters of this book.

Molecular Descriptors

One can use two kinds of molecular descriptors, namely computed and experimentally determined ones. Of course, for the latter, one needs to have the substances

whose descriptors one wishes to measure, which makes such descriptors undesirable for drug design, therefore such descriptors will not be discussed here.

The simplest computed molecular descriptors are numerical data on the various types of atoms, chemical bonds, or molecular fragments in the structure. For example, one can enumerate how many carbon, halogen, oxygen, or nitrogen atoms there are in the molecule; how many sp^3- (primary, secondary, tertiary, quaternary), sp^2-, or sp-hybridized atoms of each kind; how many hydrogen-bond donor and hydrogen-bond acceptor groups; how many functional groups such as alcohol, phenol, aldehyde, ketone, ester, amine (primary, secondary, tertiary, quaternary), amide, nitrile, oxime, etc. Using force-field calculations or quantum-chemical programs one can compute and optimize molecular geometries and molecular orbital energies, so that the results of these calculations will provide numerical data that can be used as molecular descriptors. Global dimensions of molecules can be expressed as ovality indices. Molecular weights and calculated $\log P$ values are also important indicators for drug design [27,28,29,30,31].

In the following, we will describe a class of molecular descriptors known as topological indices (TIs) because they are obtained from the constitutional formula of a compound, and therefore they include (in one number) information on the topology of the molecule, i. e. the way atoms are interconnected. Stereochemical information is not provided at this stage, but may be added via special upgrading of the descriptors. Other limitations of TIs are that they are more or less degenerate (i. e. more than one structure corresponds to the same value of the TI), that their interpretation is often unclear, and that with few exceptions one cannot retrieve the molecular structure from the numerical value of the TI; however, the last two drawbacks of TIs will be discussed again later.

Graph Theoretical Intermezzo

A *graph* is a set V of elements (v_i) called vertices and a set E of elements (e_i) called edges, that couple two distinct vertices [32,33,34,35,36]. The number of edges meting at a vertex is called the *degree* of that vertex. The sum of all vertex degrees of a graph is twice the number of edges. A *walk* in a graph is an alternating sequence of vertices and edges, beginning and ending with vertices, in which each edge is incident with the two vertices immediately preceding and following it; an example of a walk is: $v_1e_1v_2e_2v_3e_3v_4e_3v_3$, and it is evident that edges of a walk need not be distinct but can be repeated. A *path* is a walk whose vertices (and thus necessarily all edges) are distinct. In a *connected graph* any two vertices are connected by one or more paths. All graphs discussed in this paper will be connected graphs. A *cycle* (or a closed walk) is a walk with ≥ 3 distinct edges where the initial and final vertices are identical. A *tree* is a connected graph devoid of circuits; if n is its number of vertices (also called the order of the graph) the number of edges is $e = n + 1$. The *cyclomatic number* μ of a graph is the smallest number of edges that have to be deleted in order to convert it into a tree. In a tree every two vertices are joined by a unique path. The numbers of paths in a graph or a tree are limited, but the number of walks is not. *Chemical graphs* have vertices of degree at most 4. In addition, they are graph-theoretically planar, i. e. they can be embedded on a plane without crossing edges. Molecular formulas for covalently-bonded organic compounds (which represent the vast majority of medicinal drugs) can be translated mathematically into molecular graphs, by assimilating atoms with vertices (points) and covalent bonds with edges (lines). Normally, a vertex is not connected to itself, but multiple bonds are a common occurrence in chemical formulas, and the corresponding type of graph is called a multigraph; edges do not have directions, i. e. we will not use directed graphs (digraphs). However, we need to distinguish various kinds of atoms and for this purpose vertex-weighted graphs will be employed. Unless otherwise stated, hydrogen atoms will not be included so that we will have only *hydrogen-depleted-graphs* or *-multigraphs*. Measured physico-chemical or biological properties of molecules are expressed in numerical form with corresponding units, and therefore they have metric properties, i. e. one can arrange these properties in increasing or decreasing order. On the other hand, chemical formulas or the corresponding molecular graphs are not metric, but for QSPR and QSAR studies we need to find ways to assign metric properties to chemical formulas. One of the simplest and computationally cheapest ways to achieve this performance is to associate numerical invariants to graphs and multigraphs. In graph-theoretical parlance, two vertices are called adjacent if they are connected by an edge, and two edges are called adjacent if they share one vertex. A vertex is incident to an edge if it is an endpoint of that edge. The number n of vertices of a graph is called the *order* of the graph. A simple graph has no loops or multiple edges. For the purpose of chemical information, one can associate graphs with various kinds of matrices, resulting in a one-to-one correspondence between vertices and "local vertex invariants" (LOVIs). Two of the most frequently used matrices are the adjacency matrix $\mathbf{A}$, and the distance matrix, $\mathbf{D}$, which are symmetrical matrices with entries a_{ij} and d_{ij}, respectively. For adjacent vertices (i. e. points connected by an edge), $a_{ij} = 1$, and

it is 0 otherwise. It will be noted that **A** is identical to the Hückel matrix for conjugated systems. The row and column sums (written in Fig. 2 at the right of each matrix in ***bold italics***) of **A** are called vertex degrees, v_i, a notion familiar to organic chemists who call vertices of degree 1 "primary carbon atoms", vertices of degree 2 "secondary carbon atoms", etc.

$$v_i = \sum_j a_{ij} \,. \tag{6}$$

Figure 2 presents as examples the **A** and **D** matrices for eight graphs with five carbon atoms: **1**–**3** are the three possible isomers of alkanes C_5H_{12} and **4**–**8** are monocyclic hydrocarbons C_5H_{10}. It is evident from Fig. 2 that vertex degrees as LOVIs are not highly discriminating because the primary atoms of **2** have all the same value ($v_i = 1$) although vertex 4 is not equivalent to vertices 1 and 5; similary, the three secondary vertices of **1** and **4** have all the same value ($v_i = 2$) although vertices 2 and 4 in **1**, or 4 and 5 in **2** are not equivalent to the third vertex. It will be seen from Fig. 2 that both **A** and **D** have the same entries 1, and on their main diagonal the same 0 values; however, all other zero values of **A** are replaced in **D** by integer values higher than 1.

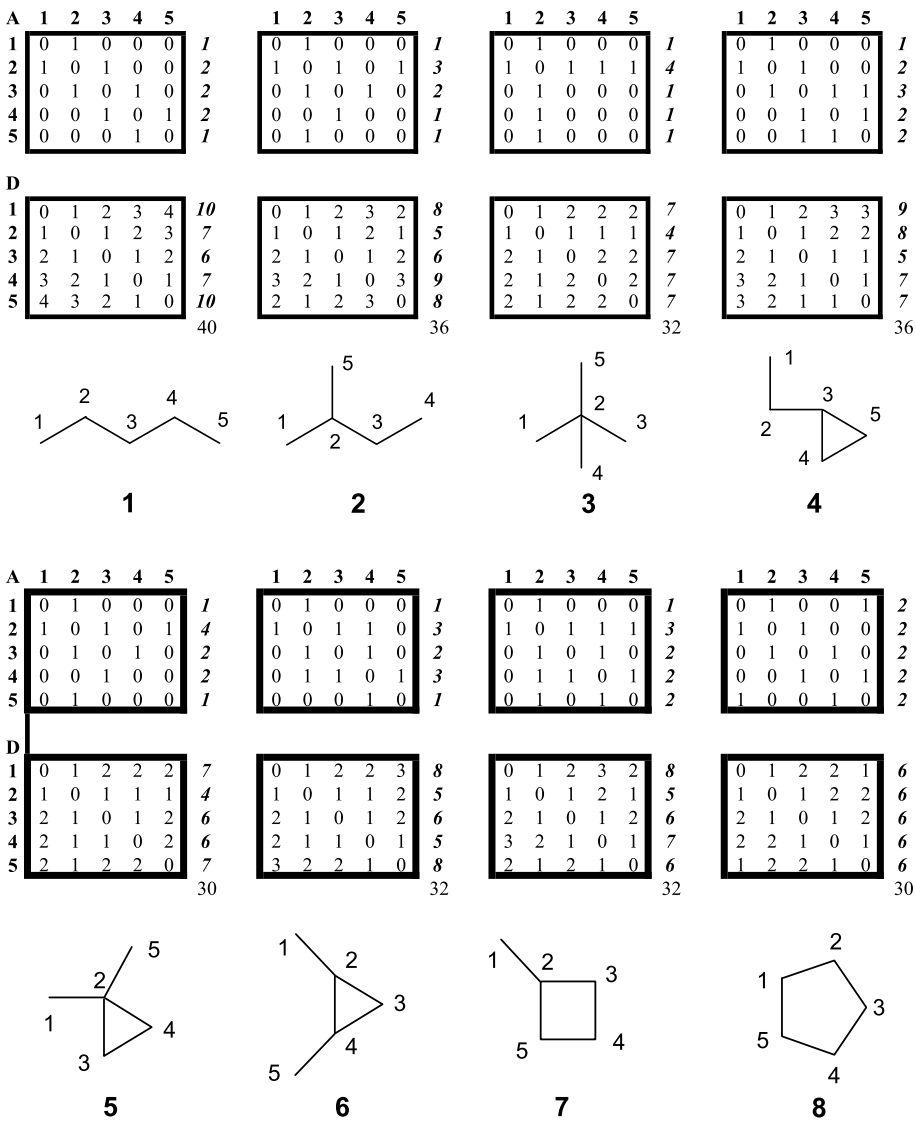

Drug Design, Molecular Descriptors in, Figure 2
Adjacency and distance matrices for eight 5-vertex graphs 1–8

Entries d_{ij} of the distance matrix are the shortest topological distances (i. e. the smallest number of connected edges, or shortest path) between vertices i and j. The row or column sums of **D** are called distance degrees, distance sums, or briefly distasums, d_i:

$$d_i = \sum_j d_{ij} \,. \tag{7}$$

Examination of Fig. 2, where at the right-hand side of each **D** matrix one sees the distasums, shows that now the primary carbon atoms in **2** have different d_i values (8 for vertices 1 and 5, and 9 for vertex 4); also the secondary carbon atoms in **1** (7 for vertices 2 and 4, and 6 for vertex 3) as well as in **4** (7 for vertices 4 and 5, and 8 for vertex 2). The same is true for graphs **5**–**8** where non-equivalent vertices have different d_i values.

Three other symmetrical matrices that were used for devising LOVIs are: (i) the resistance distance matrix whose entries represent the electrical resistance between two vertices if each edge has a resistance of one ohm; for acyclic graphs, there is no difference between the topological distance and the resistance distance; (ii) the reciprocal distance matrix, having as entries reciprocal values of topological distances; Harary indices were devised on the basis of this matrix; (iii) the detour matrix whose entries are the numbers of edges in the longest paths between two vertices. Whereas the previous matrices characterize graphs uniquely, nonisomorphic graphs may have the same detour matrix. For more information of symmetrical and non-symmetrical matrices (distance-valency matrix, Szeged matrix, Cluj matrix, Wiener matrix, hyper-Wiener matrix) the reader should consult three recent reviews [37,38,39].

Topological Indices (TIs)

A topological index (TI) results by applying various mathematical operations to graph constituents, such as combining LOVIs of all graph vertices into a single number. The combination may be a simple operation such as summation, or a more elaborate one. If one would add all vertex degrees of a graph, the result is an integer which has the same value for all constitutional isomers (different substances with the same molecular formula, translated into hydrogen-depleted graphs with the same number n of vertices but with different connectivities):

$$\sum_i v_i = 2n - 2 + 2\mu = 2e \tag{8}$$

where the cyclomatic number μ is obtained from the number n of points and e of edges:

$$\mu = e - n + 1 \,. \tag{9}$$

Thus, all alkanes have $\mu = 0$, and all monocycloalkanes have $\mu = 1$.

When multiple bonds are present and each edge is counted separately, in the corresponding hydrogen-depleted *multigraphs* double bonds are equivalent to rings, and a triple bond counts as two double bonds. For most of the simple organic compounds with the following numbers of C, H, monovalent Hal, and trivalent N atoms, the cyclomatic number can also be found as follows: $\mu = \mathrm{C} + 1 - (\mathrm{H} + \mathrm{Hal} - \mathrm{N})/2$. Any divalent sulfur and oxygen atoms do not affect μ.

Integer TIs based on integer LOVIs are considered to belong to the 1st generation, real-number TIs based on integer LOVIs to the 2nd generation, and real-number TIs based on real-number LOVIs to the 3rd generation. We will proceed in an approximate chronological order according to these generations but this review will not be exhaustive.

The Wiener Index W A young medical student, Harold Wiener, published in 1947 in the Journal of the American Chemical Society two papers on QSPR for normal boiling points and for heats of isomerization or vaporization of alkanes, using a simple molecular descriptor which is now known as the Wiener index (W). Although the initial definition was valid only for alkanes [40], this index was soon afterwards shown by Hosoya [41] to be equal to the half-sum of all entries in the distance matrix **D** of the hydrogen-depleted graph (values of $2W$ can be seen for all constitutional isomers of heptane and octane in Table 1, and for **1**–**8** in Fig. 2):

$$W = \frac{1}{2}\sum_i d_i = \frac{1}{2}\sum_i \sum_j d_{ij} \,. \tag{10}$$

It can be easily deduced that in the alkane series, W increases with the number n of carbon atoms, but one can observe in Fig. 2 that with increasing branching among isomers, W decreases. It will be seen later that this may be a drawback for W because increasing complexity requires both these features (size and branching) to increase.

Also, it can be observed that W does not have a high discriminating ability among isomers (in other words, W has a high degeneracy). Thus, in Table 1 the degeneracy of W is indicated by boldface numbers; in Fig. 2, two pairs of non-isomorphic graphs, namely **5** and **8** with $2W = 30$, as well as **3**, **6** and **7**, with $2W = 32$ share the same value

for W. In conclusion, W is the first TI to have been invented, is easily computed, does not provide for encoding the presence of heteroatoms or multiple bonding, yet it correlates fairly well with many properties. For the initial correlations, Harold Wiener combined index W with a second descriptor, the number of distances of length 3.

The Platt Index and the Gordon–Scantlebury Index Almost simultaneously with Wiener, J.R. Platt published in J. Chem. Phys. (1947) and J. Phys. Chem. (1952) two QSPR studies for additive properties of saturated hydrocarbons (paraffins) using as a molecular descriptor the sum (F) of all edge degrees (number of edges adjacent to each edge) [42,43]. It was shown later by M. Gordon and G.R. Scantlebury [44] that $F/2$ is the number of ways a path of length 2 (i. e. propane subgraph) can be superimposed on the given graph.

The Hosoya Index In 1971 Haruo Hosoya published the first in a series of papers, proposing another molecular descriptor (Z) that he called *topological index* [41]. This name is now used as a class denomination. For a hydrogen-deleted graph G,

$$Z = \sum p(G, k) \tag{11}$$

where $p(G, k)$ is the number of ways in which k edges of the graph may be chosen so that no two of them are adjacent. The summation starts with $p(G, 0) = 1$ and $p(G, 1) = e$ (the number of edges in a simple graph). This index Z shows remarkable correlation ability for many alkane properties.

The Zagreb Indices Looking back at Eq. (8) where summing all vertex degrees affords a trivial result (twice the number of edges), it is easy to understand that there are ways [45] to improve on this by considering the following two new indices, named "Zagreb indices":

$$M_1 = \sum_i v_i^2 \tag{12}$$

$$M_2 = \sum_{(\text{edge } ij)} v_i \times v_j \,. \tag{13}$$

The Schultz Indices were introduced in 1989–1990 [46]. The first of them is the molecular topological index MTI:

$$\text{MTI} = \sum_i E_i \tag{14}$$

where the "intricacy elements" E_i are the elements in the ith row of the $\mathbf{E}$ matrix built from the product of the sum of adjacency and distance matrices with $\boldsymbol{v}$ the $\mathbf{1} \times n$ vertex degree matrix:

$$\mathbf{E} = \boldsymbol{v}(\mathbf{A} + \mathbf{D}) \,. \tag{15}$$

The determinant of the adjacency-plus-distance matrix was proposed as an index called TI for alkanes:

$$\text{TI} = \det |\mathbf{A} + \mathbf{D}| \,. \tag{16}$$

Interestingly, from such indices Schultz went on to define three-dimensional molecular descriptors encoding chirality.

Centric Indices for acyclic graphs were proposed by Balaban in 1979 by summing squares of numbers of vertices of degree 1 pruned stepwise from the periphery towards the graph center, which is either one vertex or a pair of adjacent vertices [47]. Later, generalizations for graph centers of cyclic graphs or of substituents were proposed [48,49,50,51].

Bonchev's Overall Connectivity The sum of all adjacencies (vertex degrees) in a graph is called the total adjacency A:

$$A = \sum_i v_i = \sum_i \sum_j a_{ij} \,. \tag{17}$$

It was shown earlier (Eq. 8) that A is twice the number e of edges. If, according to Danail Bonchev, one sums A values for all connected subgraphs of a given graph (up to a selected threshold for the number of vertices in the subgraph) one obtains the total (or overall) connectivity, or topological complexity, TC [52]. The subgraphs in acyclic graphs (trees) are paths of various lengths or clusters of various sizes. Because in practice the total number of subgraphs increases very rapidly with the graph size, a pragmatic solution is to take into account only subgraphs up to a certain number e of edges as the *partial topological complexity* ${}^eTC$. An advantage of TC and ${}^eTC$ is the fact that such indices increase both with size and with branching, whereas W and the Randić index χ (vide infra) increase with size and decrease with branching.

We start now with 2nd generation TIs.

The Randić Connectivity Index The following desirable qualities of TIs have been enumerated by Randić [53]:

1. Direct structural interpretation
2. Good correlation with at least one property
3. Good discrimination of isomers
4. Locally defined
5. Generalizable to "higher" analogues
6. Linearly independent

7. Simplicity
8. Not based on physico-chemical properties
9. Not trivially related to other TIs
10. Efficiency of construction
11. Based on familiar structural concepts
12. Showing a correct size dependence
13. Gradual change in value with gradual change in structure.

To these qualities, one may add:

14. Low requirement in computer (CPU) time.

The Randić connectivity index χ or ${}^1\chi$ [54] was proposed in 1975 as the sum over all edges $i-j$ of the reciprocal square root of products of vertex degrees:

$$\chi = \sum_{(i-j)} (v_i \times v_j)^{-1/2} . \tag{18}$$

This TI is the most popular and most frequently cited in the literature. It provides good correlations with many physico-chemical properties and biological activities, and fulfills most of the above requirements. It increases with the number of vertices and decreases with increasing branching, like Hosoya's index Z or the NBPs of alkanes.

The Kier–Hall Higher-Order Indices and Related Indices An edge is equivalent to a path of length one and this relation between edges and paths prompted Kier and Hall in 1976 to introduce higher-order indices for various paths of longer lengths (${}^2\chi$, ${}^3\chi$, etc.) or other subgraphs such as clusters with 3 or 4 vertices, or cycles. They then suggested taking into account the presence of heteroatoms in the molecular scaffold proposing "valence connectivity indices" ${}^v\chi$, where weights (defined in accordance to positions in the Periodic System) replace vertex degrees [55]. By their two initial and influential books they "converted" many pharmaceutical chemists to look at TIs as a convenient tool for drug design [11,12].

Kier introduced also a "flexibility index" [56]. By taking into account additional molecular features such as electronegativity, Kier and Hall described "electrotopological state indices" [13], and the corresponding software MOLCONN is distributed by Hall Associates. All these indices were reviewed recently as book chapters [57].

Information Theoretic Indices were first described in 1977 by Bonchev and Trinajstić [58]. On applying Shannon's formula for summands belonging to different equivalence classes before the summation of LOVIs, one can reduce the degeneracy of many TIs. Thus, by sorting topological distances d_{ij} into groups of g_i summands having the same d value and only then applying the global summation, one can convert the 1st-generation Wiener index $W = 1/2 \sum_i i g_i$ into the 2nd-generation corresponding *total* information-based index on the distribution of distances, with much lower degeneracy than W:

$$I_D^W = W \log_2 W - \sum_i i g_i \log_2 i . \tag{19}$$

One can also obtain the *mean* information on the distance degree distribution. A similar approach affords the total and the mean information on other descriptors such as the Hosoya index or the vertex degree distribution [14]. Moreover, the 3D-analog of the Wiener number (WG) introduced by Bonchev and Mekenyan in 1985, and its information-based counterpart, like other information-based TIs, is now part of software packages CODESSA, DRAGON, and OASIS (the last one was developed in Bourgas).

In 1979, Basak and coworkers devised a different type of information-theoretic indices from constitutional formulas that include hydrogen atoms [59,60]. Atoms are partitioned into disjoint subsets, and probabilities p_i are assigned to sets of equivalence classes based on chromatic coloring or to orbits of the automorphism group of the graph. Then, information content (IC), structural information content (SIC) and complementary information content (CIC) are defined as:

$$\mathrm{IC} = -\sum_i p_i \log_2 p_i \tag{20}$$

$$\mathrm{SIC} = \mathrm{IC}/\log_2 n \tag{21}$$

$$\mathrm{CIC} = \log_2 n - \mathrm{IC} . \tag{22}$$

Analogous but different information-theoretic formulas for four indices (U, V, X, Y) were proposed by Balaban and Balaban [61].

Mean Square Distances in Graphs A very simple descriptor is derived from the observation that compact graphs have lower average distances than elongated graphs [62]. By investigating various ways of averaging distances, Balaban proposed the following formula for the mean square distance index $D^{(2)}$:

$$D^{(2)} = \left(\sum_i i^2 g_i \Big/ \sum_i g_i \right)^{1/2} . \tag{23}$$

For acyclic graphs, it was shown that even better correlations with octane numbers or heptanes and octanes could be obtained (instead of summing up all distances) by considering only distances between primary carbon atoms (endpoints).

Average Distance-Based Connectivity *J* (Balaban Index) With the aim of obtaining a TI with low degeneracy that does not increase appreciably with size but depends mainly on branching, Balaban replaced vertex degrees in Eq. (18) by distasums (s_i) and introduced a term that averaged the result with respect to the number of edges and rings, keeping the summation over all edges $i - j$ and the inverse square root operation [63]:

$$J = [e/(\mu + 1)] \sum_{(i-j)} (d_i \times d_j)^{1/2} . \tag{24}$$

Interestingly, it was proved that for an infinite vinylic polymer $H_2C=CH–R$ the J index is a rational multiple of the number $\pi = 3.14159\ldots$, and for polyethylene with R = H, it is exactly π [64]. For accounting for the presence of heteroatoms or multiple bonds, parameters based on electronegativities or covalent radii relatively to carbon seem to work well [65,66]. Unlike previously described topological indices, the J index has a low degeneracy, so that for alkanes, the first isomers with the same J values have 12 carbon atoms [67]. Because of its special nature, index J should not be used in monoparametric correlations for properties that depend on the molecular size but it can be used in such correlations for isomeric sets such as those of Fig. 1.

Eigenvalues of Matrices as TIs It was shown by Lovasz and Pelikan that the largest eigenvalue of the adjacency matrix increases with graph branching [68]. Medeleanu and Balaban investigated several types of eigenvalues for various matrices [69], and similar explorations were reported by Randić and coworkers [70].

The *Harary index* was proposed independently by chemists in Zagreb [71] and Bucharest [72] and is based on the matrix containing reciprocals of topological distances. Such LOVIs are then combined into an index by a Randić-type formula (see operation 4 on the next page). It places larger weights on atom pairs close to each other, unlike TIs based on the distance matrix.

Gutman's Szeged Index Gutman introduced the edge-Szeged matrix where entries ${}^{e}SZ_{ij}$ are products of numbers of vertices on lying closer to i or j after edge ij is removed from the graph (vertices equidistant to i and j are not counted). For trees, the Szeged and Wiener matrices and indices are identical [73].

Diudea Indices Diudea and Balaban devised two different approaches for the same purpose by using "regressive distances" or "regressive vertex degrees" by means of nonsymmetrical matrices whose entries in row i represent sums of topological distances or vertex degrees, respectively, for vertices in shells around vertex i; an exponentially decreasing factor multiplies each successive shell [74,75]. Further work by Diudea produced new matrices (Cluj matrix) and modified Szeged matrices from which other descriptors were obtained [76,77,78].

Walk Counts (Rücker Indices) By raising adjacency matrices to higher powers one can obtain easily total walk counts of graphs. They can be used in correlations with NBP's of alkanes and cycloalkanes [79,80,81].

Indices Based on Path Counts The numbers of p_L paths with various lengths L have been seen to determine Kier–Hall indices ${}^{L}\chi$. On examining NBPs of alkanes, Randić observed that there are regular variations when looking at NBPs of octane isomers arranged according to p_2 and p_3 and this led to the idea of orthogonalized descriptors to be mentioned further below. The enumeration of all paths with lengths from 1 to the maximum value in a graph (called the *path code*) provides distinct sequences for all alkanes up to octane isomers. Among possible indices obtained from path codes, index Q modeled after the Zagreb index as sum of squares of all path numbers (but normalized for the number of rings) has a fairly high degeneracy, but index P appears to provide the best correlations with physical properties [82]. In Table 1 one can see for all constitutional isomers of heptanes and octanes the path code and both indices Q and P along with other four TIs.

$$P = \sum_i \left\{ p_i^{1/2} / \left[i^{1/2} (\mu + 1) \right] \right\} . \tag{25}$$

Triplet LOVIs and Derived Topological Indices Triplet indices were introduced for converting 1st or 2nd generation TIs into 3rd generation TIs, by combining symmetrical or non-symmetrical matrices with two column vectors (one for the main diagonal and the other for the free term) and then solving the resulting system of linear equations [83]. As a result, one can have a wide variety of TIs that can be adapted to the structures under investigation. For instance, the **A** or **D** matrices of neopentane (2,2-dimethylpropane, structure **3** in Fig. 2) may be combined with column vectors [a], [b] that are arbitrary constants, or that convey chemical (e. g. atomic numbers, electronegativities, bond multiplicity, etc.) or topological information (e. g. vertex degrees, distance sums, etc.).

Taking into account that the hydrogen-depleted graph of neopentane has only two kinds of vertices (primary and quaternary carbon atoms, with LOVIs $x_1 = x_3 = x_4 = x_5$, and x_2, respectively), for the simplest case when the two column vectors are constants a and b,

A

	x_1	x_2	x_3	x_4	x_5	
x_1	0	1	0	0	0	1
x_2	1	0	1	1	1	4
x_3	0	1	0	0	0	1
x_4	0	1	0	0	0	1
x_5	0	1	0	0	0	1

D

	x_1	x_2	x_3	x_4	x_5	
x_1	0	1	2	2	2	7
x_2	1	0	1	1	1	4
x_3	2	1	0	2	2	7
x_4	2	1	2	0	2	7
x_5	2	1	2	2	0	7

	x_1	x_2	x_3	x_4	x_5	
x_1	a	1	0	0	0	$=b$
x_2	1	a	1	1	1	$=b$
x_3	0	1	a	0	0	$=b$
x_4	0	1	0	a	0	$=b$
x_5	0	1	0	0	a	$=b$

	x_1	x_2	x_3	x_4	x_5	
x_1	a	1	2	2	2	$=b$
x_2	1	a	1	1	1	$=b$
x_3	2	1	a	2	2	$=b$
x_4	2	1	2	a	2	$=b$
x_5	2	1	2	2	a	$=b$

Drug Design, Molecular Descriptors in, Figure 3

***Upper left and right sides*: adjacency and distance matrix for 2,2-dimethylpropane (structure 3 in Fig. 2) with row sums (vertex degrees and distance sums, respectively). *Lower left and right sides*: addition of column vectors *a* on the main diagonal and *b* as the free term, converts matrices A and D into systems of linear equations whose solutions provide real-number LOVIs, x_1 through x_5, for the five vertices**

one obtains for the triplet system [**A**, a, b] the following LOVIs, as seen from Fig. 3:

$x_1 = x_3 = x_4 = x_5 = b(a-1)/(a^2-4)$ and $x_2 = b(a-4)/(a^2-4)$. Thus, the ratio x_2/x_1 (which for vertex degrees as LOVIs was 4/1) becomes $x_2/x_1 = (a-4)/(a-1)$. For instance, with $a = 5$ and $b = 7$, one obtains $x_1 = 4/3$ and $x_2 = 1/3$ (in this case, a and b are arbitrary constants without any chemical or graph-theoretical significance).

Likewise for the triplet system [**D**, a, b] one obtains the following LOVIs:

$x_1 = x_3 = x_4 = x_5 = b(a-1)/(a^2+6a-4)$ and $x_2 = b(a+2)/(a^2+6a-4)$. Thus, the ratio x_2/x_1 (which for distance sums as LOVIs was 7/4) becomes $x_2/x_1 = (a+2)/(a-1)$. For instance, with $a = 5$ and $b = 7$, one obtains $x_1 = 24/51$ and $x_2 = 49/51$.

After obtaining real-number LOVIs one combines them into TIs by various operations such as:

1. Sum of weights $\sum_i x_i$;
2. Sum of squared weights $\sum_i x_i^2$;
3. Sum of square root of weights $\sum_i x_i^{1/2}$;
4. Sum of cross-products for edge endpoints $\sum_{ij} x_i \times x_j$;

and one denotes the TI thus obtained by the initial of the matrix and of the two vectors followed by the number of the operation, e. g. AZV1 (adjacency matrix A, atomic number Z on the main diagonal, vertex degree V as the free term, and summation of the LOVIs, i. e. operation 1). Interestingly, Basak et al. used the TRIPLET and POLLY programs for a variety of QSAR studies [84].

Estrada's Generalized Topological Index It was found by Ernesto Estrada that many 1st and 2nd-generation TIs (such as Zagreb indices, Randić index, Balaban's J index, Wiener index, Harary index, Schultz MTI index, Gutman index, and hyper-Wiener index) can be derived from a unique generalized formula for generalized topological indices (GTIs) [85,86,87,88,89,90]. Again, a matrix and two vectors are involved, but the result is an all-encompassing powerful tool, because it also optimizes the variable parameters for the database. These GTIs represent points in a 6-or 8-dimensional space of topological parameters, which can be optimized for describing a specific property. From the infinity of possible GTIs defined by the vector-matrix-vector multiplication, by a systematic change of 6 parameters from 0.0 to 0.9 with a step of 0.1, one reduces the set to only 10^6! As an example of the power of this generalization, Estrada and Gutierrez optimized the J index according to 16 octane numbers from the set of isomeric octanes resulting in a correlation factor $R = 0.992$ and standard deviation $s = 3.5$ with a cubic equation [91].

Estrada's Topological–Substructure Molecular Design QSAR is based on the computation of the spectral moments of the bond matrix and is able to identify compounds with a certain biological activity among a database with compounds with many other types of biological activities. Thus, sedatives/hypnotics were found in the Merck database by using a TOSS-MODE discrimination

model [92,93]. By computing fragment contribution models, it was possible to predict that among a virtual library of 327 antioxidants (flavonoid and cinnamic acid derivatives), 70 (from which only about 20 have been described in the literature) should be more active than the most powerful antioxidants in the Brazilian propolis [94].

Orthogonal Descriptors It was mentioned earlier that even when molecular descriptors are intercorrelated, yet they cover different aspects of the multidimensional space determining the variation of properties/activities, a procedure of taking the difference between such descriptors (i. e. a projection of one vector on the other) leads to surprisingly high correlations, as shown by Randić [95,96].

Variable Connectivity Indices are derived from Randić's connectivity indices by optimizing correlations for a given set of data under the assumption that entries in the connectivity matrix may vary. Thus for a compound with a heteroatom, instead of having an entry 1 when adjacencies with this heteroatom are involved, one allows the corresponding entry to be optimized with respect to the final factor of correlation, leading to a weighted entry. Even atoms of the same kind but in different environments (e. g. primary, secondary, tertiary or quaternary carbon atoms) can be allowed to have different weights [97,98]. Pompe observed that in certain cases ("anticonnectivity") the weight can be negative [99]. The field was reviewed recently by Randic and Pompe [100].

Optimal Descriptors Based on SMILES Notation Weininger introduced a simple yet powerful linear notation for encoding chemical structures, called Simplified Molecular Input Line Entry System (SMILES) that can be manipulated by computer programs for interconversion between structural formula and adjacency matrix. Inspired by Randić's optimal (flexible) descriptors for a set of compounds with a certain property/activity expressed numerically, Toropov, Gutman and coworkers assigned optimized numerical values to each individual symbol the SMILES notation for these compounds, including lower and upper case letters, brackets, or numbers, and producing thus satisfactory correlations with the given property/activity [101]. Other studies involved toxicity [102,103] or various biological activities [104].

Gálvez's Charge Indices for Discriminant Analysis for drug design combines the use of molecular connectivity, charge indices, and in-house molecular descriptors accounting for molecular shape [105]. Starting from two data sets, each comprising active and inactive compounds, and using linear discriminant analysis, one can obtain separation between activity classes. The predicted active compounds are tested for the predicted activity or (in case a virtual library was analyzed) synthesized and then tested. It was thus possible to discover novel activities for known compounds and to synthesize new compounds with certain biological activities [106].

A Real-World Example of Using Molecular Descriptors Including TIs

As an example of how these simple molecular descriptors (which are easy and fast to compute for any imaginable structure) can be applied for biomedicinal purposes, we present a study by Lahana and coworkers [107] who described a rational design of immunosuppressive decapeptides using 13 molecular descriptors for each amino acid including four topological and shape indices (Balaban's J and three Kier–Hall indices), along with log P, ellipsoidal and molar volumes, molar refractivity, dipole moment, plus four simple counts of O and N atoms, ethyl and hydroxyl groups. Starting from 24 known bioactivities of active and inactive decapeptides, each descriptor had intervals of favorable and unfavorable ranges of values. The virtual library of decapeptides with structure RXXXRXXXXY (where R = arginine, Y = tyrosine, and X allowing a combinatorial choice involving six natural and one non-natural amino acids), amounted to about a quarter million. This library was reduced 10,000-fold to 26 structures by rapidly computing all their descriptors and selecting those that fitted into the favorable domains. Then in a much slower and elaborate computational effort involving molecular dynamics simulations, a set of five optimal structures was obtained. All five were synthesized and tested for mouse heart allograft model. One decapeptide displayed an immunosuppressive activity approximately 100 times higher than the lead compound [107,108,109].

One should mention that usually the reduction in candidate structures is not as large as in this case. Moreover, often multiparametric correlations involving simple molecular descriptors are followed by more elaborate 3D modeling which is moderately more demanding for CPU time. Only then can one achieve the final stage of molecular docking and rational drug design that is described by other monographs of this Encyclopedia.

Three-Dimensional Molecular Descriptors

The TIs discussed so far provide information on constitution (connectivity), but not on stereochemistry. For many biologically active compounds, this is a serious limitation

because enantiomers with the same constitution usually have different biological activities, as known from the different odors of enantiomers or from the teratogenic consequences of using the racemic mixture of the pharmaceutical Thalidomide, one of the enantiomers of which had tragic results and produced limbless babies. Several ideas have been published on how to proceed from two-dimensional topology to three-dimensional geometry (only reviews are cited) [110,111,112,113]. Unlike the calculations for TIs, such 3D calculations usually require longer computer-time (CPU-intensive).

Cramer defined for his *comparative molecular field analysis* (CoMFA) a 3-D grid so as either to (i) include the most active or the most rigid parmacophore (ligand), or (ii) model the receptor cavity defined by its molecular docking with the potential bonding sites [114]. At each intersection point of this imaginary grid one computes the molecular geometry using force-field (molecular mechanics) methods or Gaussian approximations. Cross-validation is achieved by the leave-one-out procedure from the studied series, and the error is measured by a parameter named PRESS (predicted residual sum squares). Tripos Co. owns the patent and computer programs. The most difficult problems in CoMFA studies are connected with conformational aspects of flexible molecules, and with molecular alignment. Various remedies have been proposed, and the literature grows continuously both for CoMFA and for CoMSIA (*comparative molecular similarity analysis*) [115,116,117,118].

Todeschini proposed the Weighted Holistic Invariant Molecular (WHIM) approach and applied it to molecular surfaces (MS-WHIM) to derive new 3D theoretical descriptors. A 3D QSAR study performed on a series of steroids, comparing the WHIM description to CoMFA fields, showed that MS-WHIM descriptors provide meaningful quantitative structure–activity correlations. The concise number of indices, the ease of their calculation, and their invariance to the coordinate system make MS-WHIM an attractive tool for 3D-QSAR studies [119,120,121,122,123,124,125]. Other descriptors have been developed, which encompass size, shape, symmetry, and atom distribution, such as the EVA descriptor (normal coordinate EigenVAlues) [126] derived from the vibrational frequencies; all of these are independent of structural alignment [127].

Molecular Similarity, Dissimilarity, and Clustering of Molecular Descriptors

On the basis of the postulate that similar structures will have similar biological activities, a proven strategy in finding lead compounds for drug design is to look at structures with descriptors that are close neighbors in large databases [128]. With fuzzy bipolar pharmacophore autocorrelograms, the neighborhood behavior (as distinct from the clustering behavior) provide alternative and complementary criteria for evaluating performances of molecular similarity metrics [129]. The approach initiated by Mezey ("molecular holograms" at various resolution levels) has allowed shape simulations to be made for various biologically important molecules such as amino acids or even macromolecular compounds [130]. The rapidly increasing performance of computers now makes it possible to perform elaborate calculations for quantum-chemical parameters [131,132]. With ab-initio calculations of molecular volumes and areas of projection, it was possible to classify molecular shapes (e. g. spheroids, disk-like, rodlike, starlike, crosslike, egglike) by means of descriptors for ovality, roughness, size-corrected parameters and also (based on higher central momenta) skewness and kurtosis [133]. The functional diversity of compound libraries is discussed in detail by Gorse and Lahana [108].

Molecular similarity is important when examining potential candidates for enzyme inhibitors (a classical example is that of sulfonamides that mimic *para*-aminobenzoic acid and thus act as bactericides). On the other hand, molecular dissimilarity (diversity) is important when looking for lead compounds in large databases such as the Cambridge Structural Database (for X-ray crystallographic data) or the National Cancer Institute Databases. Burden [134] suggested using molecular identification numbers obtained from the two lowest EVAs of the connectivity matrix obtained from the hydrogen-depleted graph. A further step was undertaken by Pearlman who proposed the BCUT (Burden–Chemical Abstracts–University of Texas) approach, which consists of taking the lowest and highest EVAs, but using modified matrices instead of atomic numbers. Relevant data for intermolecular interactions are inserted on the main diagonal, namely atomic charges, polarizabilities, and H-bond donor and acceptor abilities. By means of the program CONCORD for generating 3-D structures from 2-D structures, it is possible to introduce scaled information on the 3-D structure as the off-diagonal entries. With the "auto-choose" algorithm, Pearlman obtained a program for comparing the diversities of large databases at a relatively low CPU cost.

Clustering of TIs has been investigated for large and diverse databases [135,136] and the interesting result was that one can group together most TIs into a few classes, so that one should be able to select molecular descriptors that

correlate with different aspects of the molecular structure. Thus it was found that the largest cluster contains TIs that depend mainly of the size of the graph such as W; a separate cluster contains path connectivities of intermediate order; such indices of higher order appear as a separate cluster; information-based indices are clustered separately; shape descriptors also form a separate cluster; triplet indices of various kinds are distributed among many of these clusters, proving the diversity of such TIs when they are properly selected.

Hierarchical Ordering of Structures and of Molecular Descriptors

On the basis of mathematical arguments, Bertz advocated the idea that line graphs of successive orders provide the most logical ordering of chemical structures, and published a list of ordered alkanes [137]. It was a surprise for him to observe that among all TIs, index J provided the same ordering of isomeric alkanes up to heptanes, and that even for higher alkanes the differences were minimal. In Table 1 heptanes and octane isomers were sorted according to increasing J values [138,139,140]. It is interesting to observe that apart from degeneracy, W values are ordered similarly. Among other criteria for accepting new TIs, a logical ordering of isomers is useful, and the selection of index P (Table 1) took also this criterion into consideration [82].

Basak and coworkers published several papers proving that molecular descriptors obey a hierarchy from the simplest (*topostructural indicators* of the graph constitution) through *topochemical indicators* that add chemical information on heteroatoms and bond multiplicity, then *geometrical indicators*, which contain additional information on the stereochemistry of molecules, to *quantum-chemical descriptors* that are CPU-intensive. In most QSAR studies, it was found that correlation factors increase appreciably from topostructural to topochemical descriptors, but then the improvement was minimal when investing more computational efforts for including geometrical or quantum-chemical information [141]. A thorough review of quantum-chemical descriptors contains details about their applications for QSPR and QSAR [142].

General Problems Connected with 2D-QSAR for Drug Design

With several older and newer journals covering the area of computer-aided drug design (CADD) and with the success of in-silico molecular drug discovery, it is small wonder that all "big pharma" companies sustain groups of chemists, biochemists and computer specialists. A few specific aspects will be discussed briefly here, but the area is too vast for in-depth coverage, and even the bibliography can be only sketchy [143].

Several computer programs are available for QSAR using hundreds of molecular descriptors and selecting a prescribed number of those that lead to the highest correlation factors: Katritzky and Karelson's CODESSA and CODESSAPRO [144], Todeschini's DRAGON, Hall and Kier's MOLCONN, Basak's POLLY and TRIPLET, Bonchev and Mekenyan's OASIS, and Diudea's TOPOCLUJ, among other ones [145]. Some of the newer statistical validation procedures are included such as leave-one-out, or division into training and test sets. For similarity evaluation, one may select k nearest neighbors (KNN), artificial neural networks [146,147], support vector machines [148], genetic (evolutionary) algorithms [149,150] and combinations thereof. Statistical methods include principal component analysis (PCA), partial least squares (PLS), nonlinear iterative partial least-squares (NIPALS), decision trees, etc.

In order for a molecule to be drug-like (i. e. a drug candidate) it must fulfill several criteria, as discussed by Lipinsky's rule-of-five, which states that candidate compounds are likely to have unfavorable absorption, permeation and bioavailability characteristics if they contain more than 5 H-bond donors, more than 10 H-bond acceptors, a log P greater than 5, and/or a molecular mass of more than 500 Da. Only then may a compound comply with ADMET requirements (absorption, distribution, metabolism, excretion, and toxicity) [151]. Modeling structures for predicting ADMET properties may require non-linear techniques such as support vector machines (SVMs) and Bayesian neural networks (BNNs) [152]. Special cases are compounds that need to penetrate the blood-brain-barrier (BBB) [153,154].

Often in developing combinatorial virtual databases one needs to have descriptors for subgraphs (fingerprints, molecular fragments, or simply substituents). Topological descriptors must depend on the position of the "root" vertex and one should also consider the interaction between fragments and the main part of the molecule [155,156].

Interpretation of Topological Indices

In the introductory section about molecular descriptors it was mentioned that TIs do not appear to have a clear physical interpretation. Moreover, Charton argued that TIs cannot express any "fundamental property" and therefore whenever they show good correlation ability with bulk physical properties (e. g. phase change properties such as normal boiling points, melting points, critical tempera-

tures, heats of melting or vaporization, solubilities of noble gases in liquid compounds, or ratios of Van der Waals constants) they must contain "hidden" correlations to the real causes of such phenomena, namely dipole moment, polarizability, hydrogen bonding, and steric effects. The same should be true, according to him, for QSAR of TIs with biological effects [157].

However, the great advantages of TIs are their ease of calculation with low CPU requirements for any chemical constitution, making the search of virtual combinatorial libraries a useful first step in any drug design effort. Subsequent steps for the selected structures may then involve more elaborate computational methods (such as 3D descriptors, CoMFA and related procedures, BCUT, WHIM, etc.), docking into receptor cavities of proteins with known steric features, quantum-chemical descriptors with flexible geometries and variable ranges of acceptable energies [158]. One should note that on going from 1D or 2D topology to 3D geometry [111], several molecular descriptors ("topographic indices" including "chirality topological indices"), modeled after topological indices, have been described. One of the first is the 3D Schultz index [159], followed by Pyka's index [160,161,162]. To follow the example of Lahana et al.'s strategy described above, the major reduction of candidate structures will occur at the first step, so that the CPI-intensive second step becomes feasible. Then the final step will be the chemical synthesis and biological/clinical testing of a restricted number of potential candidates.

Nevertheless, one should mention that there exist several proposals about possible interpretations of TIs. Soon after Wiener published his ideas about what we call now Wiener's TI W [40], Platt provided the hypothesis that $W^{-1/3}$ approximates the mean molecular diameter [43]. Randić et al. [163,164] argued that just as in quantum chemistry the HOMO-LUMO concepts have no interpretation in VB model, or conversely Kekulé valence structures have no place in MO models, TIs need not be interpretable in physical chemistry concepts. Surprisingly, the Pauling bond orders for 40 smaller benzenoids show a better correlation with the connectivity index ${}^1\chi$ than with the Coulson bond orders. This should not mean that ${}^1\chi$ is a kind of bond order. However, recalling that edges in hydrogen-depleted constitutional graphs are paths of length one, this indicates that *paths and walks that lie at the basis of many TIs do define fundamental properties of chemical constitution*. Then row sums of entries in distance matrices (distance sums) must also have a meaning, assigning higher weights to longer paths. Vice-versa, the same sums for reversed distance matrices and reciprocal distance matrices, from which the reversed Wiener index and the Harary index, respectively, are obtained, should also justify the assignment of lower weights to longer paths, where intramolecular interactions, such as inductive electronic effects, are weaker. Weighting edges in molecular graphs shows how chemical bonds have different contributions to the TIs [164]. Additional support for the fact that TIs may have an intrinsic meaning is provided by Estrada's finding of a unique formula unifying all the main TIs [85,86,87,88,89,90]. Other discussions on the interpretation and intercorrelation of TIs may be found in [165,166,167,168,169,170].

Reverse Engineering of Molecular Descriptors

Topological indices have been proved to be useful in drug design for both lead discovery and lead optimization, as attested by numerous published reports. The result of a QSAR study is an equation in terms of several descriptors, obtained with a training and a testing set (the latter can be replaced by the leave-one-out procedure), or with a ridge regression validation. The problem that arises often is then to find related structures that can improve or optimize the result, i. e. to find other structures that have the same set of descriptors. This is equivalent to finding non-isomorphic graphs that have the same TI. For highly or moderately degenerate TIs such as the Wiener, Randić, Hosoya, and centric Balaban indices, this "reverse engineering" problem is mathematically relatively simple as shown by Zefirov and coworkers [171,172,173] or Luque–Ruiz et al. [174], but will result in large numbers of graphs with given TI values or value ranges. For less degenerate TIs such as J, this would be a "harder nut to crack".

In 2005, at an American Chemical Society Symposium on reverse engineering of molecular descriptors, it was discussed whether one may find a way to share information on physico-chemical properties such as $\log P$ or solubility, or even on biological activities or toxicities, without disclosing chemical structures. From the almost 30 million existing substances in the Chemical Abstracts Service database, only few databases with such properties or biological activities are openly available for academic research. Although the "organic chemical universe (under constraints defining chemical stability and synthetic feasibility)" contains about 10^{63} substances, most drug molecules have molecular weights (MW) < 360 Da, and can be obtained from smaller lead compounds with MW ≈ 150 Da by adding substituents. The huge number mentioned above was found assuming that there are about 30 C, N, O, or S atoms linked in a molecule with up to 4

rings and 10 branch points. Reymond and coworkers [175] generated a database of all the 13.9 million virtually possible such organic compounds with MW < 160 Da. Stereoisomers account for 75% of the structures, and heterocycles predominate [175].

On the other hand, the large pharmaceutical companies have proprietary databases including hundreds of thousands of existing compounds. The problem is compounded by the fact that several thousands of large-volume chemical are manufactured around the world and reach the environment, yet the toxicity (including teratogenic, endocrine disruption, fertility and other biological effects) is known only for a small part of these chemicals. Large amounts of plasticizers in polymers, pesticides or fertilizers are probably involved in alarming signals about decreasing spermatogenesis and these signals lead to the widespread negative image of *chemistry* and *synthetic*, as opposed to *organic* or *natural* (as if everything material around us was not chemical). There is therefore an urgent need for computing (at least approximately) the toxicities of these structures by QSTR methods.

Among the views provided at this Symposium, some of them expressed the hope that perhaps chemical identity might be kept apart from physico-chemical or biological properties [176,177,178,179,180], but other authors believed that so far there is no guarantee that one could not retrieve the structure from a collection of properties [181]. This is an open and challenging question.

Future Directions

It was seen in this monograph that a continuous effort existed during the last decades for devising new molecular descriptors and computational techniques for 2D-QSAR to precede the more elaborate 3D-QSAR. The success rate so far is satisfactory, but the field progresses so rapidly that based on data about protein conformations and receptors gathered from X-ray diffraction or from NMR in solution, there is high hope that many more virtual drug-like ligands (enzyme inhibitors or activators) could be designed by computer-assisted methods.

What would greatly facilitate the area of CADD using TIs would be the development of simple descriptors (supplementing the information on covalent bonding) for charges, hydrogen-bond donors and acceptors. A few starting efforts have been published but till now no real success in this area was demonstrated.

The challenge described in the preceding paragraph is only one of the possible applications of better future molecular descriptors.

Bibliography

Primary Literature

1. Devillers J, Balaban AT (eds) (1999) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach, Amsterdam
2. Templeton NS, Lasic DD (eds) (2000) Gene therapy. Therapeutical mechanisms and strategies. Marcel Dekker, New York
3. Ilies MA, Seitz WA, Hohnson BH, Ezell EL, Miller AL, Thompson EB, Balaban AT (2006) J Med Chem 49:3872
4. Kubinyi H (2002) Drug design course (free download from his home page www.kubinyi.de). See also "Why drugs fail" Herman Skolnik Award Lecture, ACS Meeting, San Francisco, September 2006 on that web page
5. Maggiora GM (2006) J Chem Inf Model 46:1535
6. Balaban AT, Pompe M (2007) J Phys Chem 111:2448; erratum, ibid. 4871
7. Karelson M (2000) Molecular descriptors in QSAR/QSPR. Wiley, New York
8. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley, Weinheim
9. Balaban AT (2000) Quantitative structure-activity relationships and computational methods in drug discovery. In: Meyers RA (ed) Encyclopedia of analytical chemistry. Wiley, Chichester, p 7288
10. Randić M (1998) Topological indices. In: Schleyer PVR et al (eds) Encyclopedia of computational chemistry. Wiley, Chichester, p 3018
11. Kier LB, Hall LH (1986) Molecular connectivity in structure-activity analysis. Research Studies Press, Letchworth
12. Kier LB, Hall LH (1976) Molecular connectivity in chemistry and drug research. Academic Press, New York
13. Kier LB, Hall LH (1999) Molecular structure description: the electrotopological state. Academic Press, San Diego
14. Bonchev D (1983) Information theoretic indices for characterization of chemical structure. Wiley, Chichester
15. Hammett LP (1940) Physical organic chemistry, 2nd edn, 1970. McGrawHill, New York
16. Hammett LP (1935) Chem Rev 17:125
17. Hansch C, Maloney PP, Fujita T, Muir RM (1962) Nature 194:258
18. Hansch C, Fujita T (1964) J Am Chem Soc 86:1616
19. Fujita T, Iwasa J, Hansch C (1964) J Am Chem Soc 86:5175
20. Hansch C (1969) Acc Chem Res 2:232
21. Hansch C, Leo A, Hoekman D (1995) Exploring QSAR, hydrophobic, electronic and steric constants. Am. Chem. Soc., Washington
22. Hansch C, Hoekman D, Gao H (1996) Chem Rev 96:1045
23. Hansch C, Gao H, Hoekman D (1998) In: Devillers J (ed) Comparative QSAR. Taylor and Francis, Washington, p 285
24. Hansch C, Leo A (1979) Substituent constants for correlation analysis in chemistry and biology. Wiley/Interscience, New York
25. Hansch C, Leo A (1995) Exploring QSAR: fundamentals and applications in chemistry and biology. Am. Chem. Soc., Washington
26. Kubinyi H (1993) QSAR: Hansch analysis and related approaches. VCH Publishers, Weinhim
27. Verloop A (1976) In: Ariëns EJ (ed) Drug design, vol 3. Academic, New York, p 133

28. Verloop A, Tipker J (1977) In: Buisman JAK (ed) Biological activity and chemical structure. Elsevier, Amsterdam, p 63
29. Verloop A, Tipker J (1977) In: Jerman-Blazić D (ed) QSAR in drug design and toxicology, Hadzi D21–23. Elsevier, Amsterdam, p 97
30. Kubinyi H (1998) Quantitative structure-activity relationships in drug design. In: Schleyer PVR et al (eds) Encyclopedia of computational chemistry. Wiley, Chichester, p 2309
31. Kubinyi H (2006) In: Sener EA, Yalcin I (eds) QSAR and molecular modeling in rational design of bioactive molecules. CADD Society, Ankara, p 30
32. Balaban AT (1976) (ed) Chemical applications of graph theory. Academic Press, London
33. Trinajstić N (1992) Chemical graph theory, 2nd edn. CRC Press, Boca Raton
34. Ivanciuc O, Balaban AT (1998) Graph theory in chemistry. In: Schleyer PVR et al (eds) Encyclopedia of computational chemistry. Wiley, Chichester, p 1169
35. Rouvray DH, Balaban AT (1979) In: Wilson RJ, Beineke LW (eds) Applications of graph theory. Academic Press, London, p 177
36. Balaban AT (1995) J Chem Inf Comput Sci 35:339
37. Ivanciuc O, Ivanciuc T, Diudea MV (1997) SAR QSAR Environ Res 7:63
38. Ivanciuc O, Ivanciuc T (1999) In: Devillers J, Balaban AT (eds) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach, Amsterdam
39. Janežič D, Miličević A, Nikolić S, Trinajstić N (2007) In: Gutman I (ed) Graph theoretical matrices in chemistry. University of Kragujevac (Serbia)
40. Wiener H (1947) J Am Chem Soc 69(17):2636
41. Hosoya H (1971) Bull Chem Soc Jpn 44:2332
42. Platt JR (1947) J Chem Phys 15:419
43. Platt JR (1952) J Phys Chem 56:151
44. Gordon M, Scantlebury GR (1967) J Chem Soc B:1
45. Gutman I, Ruscić B, Trinajstić N, Wilcox CF (1975) J Chem Phys 62:3399
46. Schultz HP, Schultz TP (1998) J Chem Inf Comput Sci 38:853
47. Balaban AT (1979) Theor Chim Acta 5:239
48. Bonchev D, Balaban AT, Mekenyan O (1980) J Chem Inf Comput Sci 20:106
49. Bonchev D, Mekenyan O, Balaban AT (1989) J Chem Inf Comput Sci 29:91
50. Bonchev D (1989) Theochem 185:155
51. Balaban AT, Bertelsen S, Basak SC (1994) MATCH Commun Math Comput Chem 30:55
52. Bonchev D (2001) J Mol Graphics Model 20:65
53. Randić M (1991) J Math Chem 7:155
54. Randić M (1975) J Am Chem Soc 97:6609
55. Kier LB, Hall LH (1976) J Pharm Sci 65:1806
56. Kier LB (1989) Quant Struct-Act Relat 8:221
57. Kier LH, Hall LB (1999) In: Devillers J, Balaban AT (eds) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach, Amsterdam, pp 307, 455, 491
58. Bonchev D, Trinajstić N (1977) J Chem Phys 67:4517
59. Basak SC (1999) In: Devillers J, Balaban AT (eds) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach, Amsterdam, p 563
60. Raychaudhury C, Ray SK, Ghosh JJ, Roy AB, Basak SC (1984) J Comput Chem 5:581
61. Balaban AT, Balaban TS (1991) J Math Chem 8:383
62. Balaban AT (1983) Pure Appl Chem 55:199
63. Balaban AT (1982) Chem Phys Lett 89:399
64. Balaban AT, Ionescu-Pallas N, Balaban TS (1985) MATCH Commun Math Comput Chem 17:121
65. Balaban AT (1986) MATCH Commun Math Comput Chem 21:115
66. Ivanciuc O, Ivanciuc T, Balaban AT (1998) J Chem Inf Comput Sci 38:395
67. Balaban AT, Quintas LV (1983) MATCH Commun Math Comput Chem 14:213
68. Lovasz L, Pelikan J (1973) Periodica Math Hung 3:175
69. Medeleanu M, Balaban AT (1998) J Chem Inf Comput Sci 38:1038
70. Randić M, Vračko M, Novič M (2001) In: Diudea MV (ed) QSPR/QSAR studies by molecular descriptors. Nova Sci. Publ., Huntington, p 147
71. Plavsić D, Nikolić S, Trinajstić N, Mihalić Z (1993) J Math Chem 12:235
72. Ivanciuc O, Balaban TS, Balaban AT (1993) J Math Chem 12:309
73. Khadikar PV, Deshpande NV, Kale PP, Dobrynin AA, Gutman I, Dömötör G (1995) J Chem Inf Comput Sci 35:547
74. Diudea MV, Minailiuc O, Balaban AT (1991) J Comput Chem 12:527
75. Balaban AT, Diudea MV (1993) J Chem Inf Comput Sci 33:421
76. Diudea MV (1997) J Chem Inf Comput Sci 37:292–300
77. Diudea MV (1997) MATCH Commun Math Comput Chem 35:163
78. Diudea MV, Vizitiu AE, Janežič D (2007) J Chem Inf Model 47:864
79. Rücker G, Rücker C (1993) J Chem Inf Comput Sci 33:683
80. Rücker G, Rücker C (1999) J Chem Inf Comput Sci 39:788
81. Rücker G, Rücker C (1994) J Chem Inf Comput Sci 34:534
82. Balaban AT, Beteringhe A, Constantinescu T, Filip PA, Ivanciuc O (2007) J Chem Inf Model 47:716; Vukičević D, Beteringe A, Constantinescu T, Pompe M, Balaban AT (2008) Chem Phys Lett 464:155
83. Filip PA, Balaban TS, Balaban AT (1987) J Math Chem 1:61
84. Basak SC, Gute BD (2001) J Mol Graphics Model 20:95
85. Estrada E (2001) Chem Phys Lett 336:248
86. Estrada E (2003) J Phys Chem A 107:7482
87. Estrada E (2004) J Phys Chem A 108:5468
88. Matamala AR, Estrada E (2005) J Phys Chem A 109:9890
89. Matamala AR, Estrada E (2005) Chem Phys Lett 410:343
90. Estrada E, Matamala AR (2007) J Chem Inf Comput Sci 47:794
91. Estrada E, Gutierrez Y (2001) MATCH Commun Math Comput Chem 44:155
92. Esterada E (2000) SAR QSAR Environ Des 11:55; Estrada E, Gutierrez Y, Gonzales H (2000) J Chem Inf Comput Sci 40:1386
93. Estrada E, Pena A, Garcia-Domenech R (1998) J Comput-Aided Mol Des 12:583
94. Estrada E, Quincoces JA, Patlewicz G (2004) Mol Divers 8:21
95. Randić M (1991) J Chem Inf Comput Sci 31:311
96. Randić M (1991) New J Chem 15:517
97. Randić M (2001) J Mol Graphics Model 20:19
98. Randić M, Pompe M (2001) J Chem Inf Comput Sci 41:573
99. Pompe M, Randić M (2006) J Chem Inf Comput Sci 46:2
100. Randic M, Pompe M (2007) Acta Chim Slov 54:605
101. Toropov AA, Toropova AP, Mukhamedzhnova DV, Gutman I (2005) Indian J Chem 44A:1545
102. Toropov AA, Benfenati E (2007) Comput Biol Chem 31:57
103. Toropov AA, Benfenati E (2007) Eur J Med Chem 42:606

104. Roy K, Toropov AA, Raska I Jr (2007) QSAR & Comp Sci 26:460
105. Gálvez J, Garcia R, Salabert MT, Soler R (1994) J Chem Inf Comput Sci 34:520
106. Gálvez J, de Julian-Ortiz JV, Garcia-Domenech R (2001) J Mol Graph Model 20:84
107. Grassy G, Calas B, Yasri A, Lahana R, Woo J, Iyer S, Kaczorek M, Floc'h R, Buelow R (1998) Nat Biotechnol 16:748
108. Gorse D, Lahana R (2000) Curr Opin Chem Biol 4:287; Gorse A-D (2006) Curr Top Med Chem 6:3; Iyer S, Lahana R, Buelow R (2002) Curr Pharm Des 8:2217; Gorse D, Rees A, Kaczorek M, Lahana R (1999) Drug Disc Today 4:257
109. Grassy G, Kaczorek M, Lahana R, Yasri A (2006) US Patent 7,024,311
110. Kubinyi H, Folkers G, Martin YC (eds) (1998) 3D QSAR in drug design: Ligand–protein interactions and molecular similarity, vols 9–11. Kluwer, Dordrecht
111. Balaban AT (ed) (1997) From chemical topology to three-dimensional geometry. Plenum, New York
112. Balaban AT (1997) J Chem Inf Comput Sci 37:645
113. Kubinyi H, Folkers G, Martin YC (eds) (1998) 3D QSAR in drug design: Recent advances, vols. 12–14. Kluwer, Dordrecht
114. Cramer RD III, Patterson DE, Bunce JD (1988) J Am Chem Soc 110:5959
115. Cramer RD III, Patterson DE, Bunce JD (1989) In: Fauchère JL (ed) Quantitative structure-activity relationships in drug design. Alan R Liss, New York, p 161
116. Cramer RD III, DePriest SA, Patterson DE, Hecht P (1993) In: Kubinyi H (ed) 3D QSAR in Drug Design: Theory, methods and applications. ESCOM, Leiden, p 583
117. Böhm M, Klebe G (2002) J Med Chem 45:1585
118. Martin YC, Kim KH, Lin CT (1996) In: Charton M (ed) Advances in quantitative structure-activity relationships. JAI Press, Greenwich
119. Todeschini R, Lasagni M, Marengo E (1994) J Chemom 8:263
120. Todeschini R, Gramatica P, Marengo E, Provenzani R (1995) Chemom Intell Lab Syst 27:221
121. Todeschini R, Vighi M, Provenzani R, Finizio A, Gramatica P (1996) Chemosphere 32:1527
122. Todeschini R, Gramatica P (1997) Quant Struct-Act Relat 16:113, 120
123. Todeschini R, Gramatica P (1997) SAR QSAR Environ Res 7:89
124. Gramatica P, Consonni V, Todeschini R (1999) Chemosphere 38:1371
125. Gramatica P, Corradi M, Consonni V (2000) Chemosphere 41:763
126. Ferguson AM, Heritage T, Jonathon P, Pack SE, Phillips L, Rogan J, Snaith PJ (1997) J Comp-Aided Mol Des 11:143
127. Fontaine F, Pastor M, Sanz F (2004) J Med Chem 47:2805
128. Johnson M, Maggiora GM (eds) (1990) Concepts and applications of molecular similarity. Wiley, New York; Maggiora GM, Shanmungasundaram V (2004) Methods Mol Biol 275:1
129. Horvath D, Mao B (2003) QSAR Comb Sci 22:498
130. Mezey PG (1993) Shape in chemistry. An introduction to molecular shape and topology. Wiley, New York
131. Carbo-Dorca R, Mezey PG (eds) (1998) Advances in molecular similarity, vol 2. JAI Press, Stamford
132. Karelson M, Lobanov VS, Katritzky AR (1996) Chem Rev 96:1027
133. Zyrianov Y (2005) J Chem Inf Model 45:657
134. Burden FR (1989) J Chem Inf Comput Sci 29:225
135. Basak SC, Balaban AT, Grunwald G, Gute BD (2000) J Chem Inf Comput Sci 40:891; Basak SC, Gute BD, Balaban AT (2004) Croat Chem Acta 77:331
136. Basak SC, Gute BD, Mills D (2006) Arkivoc ix:157
137. Bertz SH (1988) Discret Appl Math 19:65
138. Balaban AT (2002) In: Rouvray DH, King RB (eds) Topology in chemistry: Discrete mathematics of molecules. Horwood Publishing Ltd., Chichester, p 89
139. Ivanciuc O, Ivamciuc T, Cabrol-Bass D, Balaban AT (2000) MATCH Commun Math Comput Chem 42:155
140. Balaban AT (2002) MATCH Commun Math Comput Chem 45:5
141. Basak SC, Mills D, Mumtaz MM, Balasubramanian K (2003) Indian J Chem 42A:1385; Basak SC, Mills D, Balaban AT, Gute BD (2001) J Chem Inf Comput Sci 41:671
142. Katritzky AR, Lobanov VS, Karelson M (1995) Chem Soc Rev 24:279
143. Perun TJ, Propst CL (1989) Computer-aided drug design. Methods and applications. Marcel Dekker, New York
144. Karelson M, Maran U, Wang Y, Katritzky AR (2000) Coll Czech Chem Commun 64:1551
145. Jurs PC (1996) Computer software applications in chemistry, 2nd edn. Wiley, New York
146. Zupan J, Gasteiger J (1999) Neural networks in chemistry and drug design. Wiley, Weinheim
147. Devillers J (1996) Neural networks in QSAR and drug design. Academic Press, London
148. Jorissen RN, Gilson MK (2005) J Chem Inf Model 45:549
149. Devillers J (1996) Genetic algorithms in molecular modeling. Academic Press, London
150. Goldberg DE (1989) Genetic algorithms in search, optimization and machine learning. Addison-Wesley, Reading
151. Liu R, So SS (2001) J Chem Inf Comput Sci 41:1633
152. Gola J, Obrezanova O, Champness E, Segall M (2006) QSAR Comb Sci 25:1172
153. Katritzky AR, Kuanar M, Slavov S, Dobchev DA, Fara DC, Karelson M, Acree WE, Solov'ev VP, Varnek A (2006) Bioorg Med Chem 14:4888
154. Li H, Yap CW, Ung CY, Xue Y, Cao ZW, Chen YZ (2005) J Chem Inf Model 45:1376
155. Balaban AT, Catana C (1994) SAR QSAR Environm Res 2:1
156. Mekenyan O, Bonchev D, Balaban AT (1988) J Math Chem 2:347
157. Charton M (2003) J Comput-Aided Mol Des 17:197; Charton M, Charton B (2003) J Comput-Aided Mol Des 17:211
158. Mekenyan O, Pavlov T, Grancharov V, Todorov M, Schmieder P, Veith G (2005) J Chem Inf Model 45:283
159. Schultz HP, Schultz EB, Schultz TP (1995) J Chem Inf Comput Sci 35:864
160. Pyka A (1993) J Planar Chromatog Mod TLC 6:282
161. Pyka A (1997) J Serb Chem Soc 62:251; Gutman I, Pyka A, ibid. 261
162. Pyka A (1999) J Liq Chromatog Relat Technol 22:41
163. Randić M, Zupan J (2001) J Chem Inf Comput Sci 41:550
164. Randić M, Balaban AT, Basak SC (2001) J Chem Inf Comput Sci 41:593
165. Labanowski JK, Motoc I, Dammkoehler RA (1991) Comput Chem 15:47
166. Bögel H, Dettman J, Randić M (1997) Croat Chem Acta 70:827
167. Stankevich IV, Skvortsova MI, Zefirov NS (1995) Theochem 342:173
168. Hosoya H, Gotoh M, Murajami M, Ikeda S (1999) J Chem Inf Comput Sci 39:192

169. Kier LB, Hall LB (1977) Eur J Med Chem Chim Ther 12:307
170. Estrada E (1999) Chem Phys Lett 312:556
171. Skvortsova MI, Fedyaev KS, Palyulin VA, Zefirov NS (2003) Internet Electron J Mol Des 2:70
172. Skvortseva MI, Baskin II, Slovokhovotova OL, Palyulin VA, Zefirov NS (1993) J Chem Inf Comput Sci 33:630
173. Gordeeva EV, Molchanova MS, Zefirov NS (1990) Tetrahdron Comput Meth 3:389
174. Garcìa GC, Luque-Ruiz I, Gómez MA Doncel AC, Plaza AG (2005) J Chem Inf Model 45:231
175. Fink T, Bruggesser, Reymond J-L (2005) Angew Chem Int Ed Engl 44:1504
176. Bologa C, Allu TK, Olah M, Kappler MA, Oprea TI (2005) J Comput Aided Mol Des 19:625
177. Balaban AT (2005) J Comput Aided Mol Des 19:651
178. Clement OO, Güner OF (2005) J Comput Aided Mol Des 19:731
179. Faulon J-L, Brown WM, Martin S (2005) J Comput Aided Mol Des 19:637
180. Varnek A, Fourches D, Hoonakker F, Solov'ev VP (2005) J Comput Aided Mol Des 19:693
181. Filimonov D, Poroikov V (2005) J Comput Aided Mol Des 19:705

Books and Reviews

Bohacek RS, McMartin C, Guida WC (1996) Med Res Revs 16:3
Bonchev D, Rouvray DH (eds) (2005) Complexity in chemistry, biology, and ecology. Springer, New York
Bonchev D, Buck GA (2007) J Chem Inf Model 47:909
Estrada E, Uriarte E (2001) Curr Med Chem 8:1573
Diudea MV (ed) (2000) QSPR/QSAR studies by molecular descriptors. Nova Science Press, New York
Diudea MV, Florescu MS, Khadikar PV (2006) Molecular topology and its applications. Eficon, Bucharest
Holtje HD, Sippl W (eds) (2001) Rational approach to drug design. Prous Sci, Barcelona
Mannhold R, Kubinyi H, Timmermann H (eds) (1997) Molecular modeling. Methods and principles in medicinal chemistry, vol 5. VCH, Weinheim
Ooms F (2000) Curr Med Chem 7:141
Pogliani L (2000) Chem Rev 100:3827
Randić M (1998) In: Schleyer PvR et al (eds) Encyclopedia of computational chemistry, vol 5. Wiley, Chichester, p 3018
Richon AB, Young SS: An introduction of QSAR methodology. http://www.netsci.org/Science/Compchem/feature19.html
Snyder JP, Snyder FD (2000) In: Gundertofte K, Jørgensen FS (eds) Molecular modeling: prediction of bioactivity. Kluwer, New York
van de Waterbeemd H (1995) Chemometric methods in drug design. VCH, Weinheim

Dynamic Games with an Application to Climate Change Models

Prajit K. Dutta
Department of Economics, Columbia University, New York, USA

Article Outline

Glossary
Definition of the Subject
Introduction
The Dynamic – or Stochastic – Game Model
The Dynamic – or Stochastic – Game: Results
Global Climate Change – Issues, Models
Global Climate Change – Results
Future Directions
Bibliography

Glossary

Players The agents who take actions. These actions can be – depending on application – the choice of capital stock, greenhouse emissions, level of savings, level of Research & Development expenditures, price level, quality and quantity of effort, etc.

Strategies Full contingent plans for the actions that players take. Each strategy incorporates a choice of action not just once but rather a choice of action for every possible decision node for the player concerned.

Payoffs The utility or returns to a player from playing a game. These payoffs typically depend on the strategies chosen – and the consequent actions taken – by the player herself as well as those chosen by the other players in the game.

Game horizon The length of time over which the game is played, i. e., over which the players take actions. The horizon may be finite – if there are only a finite number of opportunities for decision-making – or infinite – when there are an infinite number of decision-making opportunities.

Equilibrium A vector of strategies, one for each player in the game, such that no player can unilaterally improve her payoffs by altering her strategy, if the others' strategies are kept fixed.

Climate change The consequence to the earth's atmosphere of economic activities such as the production and consumption of energy that result in a build-up of greenhouse gases such as carbon dioxide.

Definition of the Subject

The study of dynamic games is an important topic within game theory. Dynamic games involve the study of problems that are a) inherently dynamic in nature (even without a game-theoretic angle) and b) are naturally studied from a strategic perspective. Towards that end the struc-

ture generalizes dynamic programming – which is the most popular model within which inherently dynamic but non-strategic problems are studied. It also generalizes the model of repeated games within which strategic interaction is often studied but which structure cannot handle dynamic problems. A large number of economic problems fit these two requirements.

In this paper we examine the dynamic game model. The structure is discussed in detail as well as its principal results. Then the paper introduces a leading important application, the economics of climate change. It is shown that the problem is best studied as a dynamic commons game. Some recent models and associated results are then discussed.

We begin the analysis with a recall of the familiar model of repeated games (whose main results have been presented elsewhere in this volume). That is followed by the generalization of that framework to the model of dynamic – also known as stochastic or Markovian – games. These games may be thought of as "repeated games with a state variable". The presence of a state variable allows the analysis of situations where there is a fundamental dynamic intrinsic to the problem, a situation – or "state" – that changes over time often on account of the players' past actions. (In contrast to repeated games where an identical stage game is played period after period.) Such a state variable maybe capital stock, level of technology, national or individual wealth or even environmental variables such as the size of natural resources or the stock of greenhouse gases. To provide a concrete illustration of the dynamic game concepts and results, this paper will provide a fairly detailed overview of ongoing research by a number of authors on the very current and important topic of the economics of global climate change.

Section "The Dynamic – or Stochastic – Game Model" recalls the repeated games structure, introduces the subject of dynamic games and presents the dynamic games model. Section "The Dynamic – or Stochastic – Game: Results" presents – mostly with proofs – the main results from the theory of dynamic games. Section "Global Climate Change – Issues, Models" then introduces the problem of climate change, argues why the dynamic game framework is appropriate for studying the problem and presents a family of models that have been recently studied by Dutta and Radner – and in a variant by Dockner, Long and Sorger. Finally, Sect. "Global Climate Change – Results" presents the main results of these analyzes of the climate change problem. Future directions for research are discussed in Sect. "Future Directions" while references are collected in Sect. "Bibliography".

Introduction

In this paper we examine the dynamic game model. The structure is discussed in detail as well as its principal results. Then the paper introduces a leading important application, the economics of climate change. It is shown that the problem is best studied as a dynamic commons game. Some recent models and associated results are then discussed.

The Dynamic – or Stochastic – Game Model

The most familiar model of dynamic interaction is the Repeated Game model (described elsewhere in this volume). In that set-up players interact every period for many periods – finite or infinite in number. At each period they play exactly the same game, i. e., they pick from exactly the same set of actions and the payoff consequence of any given action vector is identical. Put differently, it is as if there is an intrinsic static set-up, a "state" of the system that never changes. The only thing that changes over time is (potentially) every player's response to that fixed state, i. e., players (can) treat the game dynamically if they so wish but there is no inherent non-strategic reason to do so. An impartial "referee" choosing on behalf of the players to achieve some optimization aim indeed would pick the same action every period.

Things change in a set-up where the state can change over time. That is the structure to which we now turn. This set-up was introduced by Shapley [26] under the name of Stochastic Games. It has since also been called Markovian Games – on account of the Markovian structure of the intrinsic problem – or Dynamic Games. We will refer to the set-up (for the most part) as Dynamic Games.

Set-Up

There are I players, and time is discrete. Each period the players interact by picking an action. Their action interaction take place at a given state which state changes as a consequence of the action interaction. There is a payoff that each player receives in each period based on the action vector that was picked and the state.

The basic variables are:

Definition

$\boldsymbol{t}$	Time period $(0, 1, 2, \ldots T)$.
$\boldsymbol{i}$	Players $(1, \ldots, I)$.
$\boldsymbol{s(t)}$	State at the beginning of period t, $s(t) \in S$.
$\boldsymbol{a_i(t)}$	Action taken by player i in period, $a_i(t) \in A_i$
$\boldsymbol{a(t)}$	$(a_1(t), a_2(t), \ldots a_I(t))$ vector of actions taken in period t.
$\boldsymbol{\pi_i(t)}$	$\pi_i(s(t), a(t))$ payoff of player i in period t.

$\boldsymbol{q(t)}$ $q(s(t+1) \mid s(t), a(t))$ conditional distribution of state at the beginning of period $t+1$.

$\boldsymbol{\delta}$ The discount factor, $\delta \in [0, 1)$.

The state variable affects play in two ways as stated above. In any given period, the payoff to a player depends not only on the actions that she and other players take but it also depends on the state in that period. Furthermore, the state casts a shadow on future payoffs in that it evolves in a Markovian fashion with the state in the next period being determined – possibly stochastically – by the state in the current period and the action vector played currently.

The initial value of the state, $s(0)$, is exogenous. So is the discount factor δ and the game horizon, T. Note that the horizon can be finite or infinite. All the rest of the variables are endogenous, with each player controlling its own endogenous variable, the actions. Needless to add, both state as well as action variables can be multi-dimensional and when we turn to the climate change application it will be seen to be multi-dimensional in natural ways.

Example 1 S infinite – The state space can be countably or uncountably infinite. It will be seen that the infinite case, especially the uncountably infinite one, has embedded within it a number of technical complications and – partly as a consequence – much less is known about this case.

S finite – In this case, imagine that we have a repeated game like situation except that there are a finite number of stage games any one of which gets played at a time.

Example 2 When the number of players is one, i. e., $I = 1$, then we have a dynamic programming problem. When the number of states is one, i. e., $\#(S) = 1$, then we have a repeated game problem. (Alternatively, repeated games constitute the special case where the conditional distribution brings a state s always back to itself, regardless of action.) Hence these two very familiar models are embedded within the framework of dynamic games.

Histories and Strategies

Preliminaries – A *history* at time t, $h(t)$, is a list of prior states and action vectors up to time t (but not including $a(t)$)

$$h(t) = s(0)\,,\quad a(0)\,,\quad s(1)\,,\quad a(1)\,,\quad \ldots\,,\quad s(t)\,.$$

Let the set of histories be denoted $H(t)$. A *strategy* for player i at time t, $\boldsymbol{\sigma}_i(t)$, is a complete conditional plan that specifies a choice of action for every history. The choice may be probabilistic, i. e., may be an element of $P(A_i)$, the set of distributions over A_i. So a strategy at time t is

$$\boldsymbol{\sigma}_i(t)\colon H(t) \longrightarrow P(A_i)\,.$$

A strategy for the entire game for player i, $\boldsymbol{\sigma}_i$, is a list of strategies, one for every period: $\boldsymbol{\sigma}_i = \boldsymbol{\sigma}_i(0), \boldsymbol{\sigma}_i(1), \ldots, \boldsymbol{\sigma}_i(t), \ldots$ Let $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2, \ldots \boldsymbol{\sigma}_I)$ denote a vector of strategies, one for each player.

A particular example of a strategy for player i is a pure strategy $\boldsymbol{\sigma}_i$ where $\boldsymbol{\sigma}_i(t)$ is a deterministic choice (from A_i). This choice may, of course, be conditional on history, i. e., may be a map from $H(t)$ to A_i. Another example of a strategy for player i is one where the player's choice $\boldsymbol{\sigma}_i(t)$ may be probabilistic but the conditioning variables are not the entire history but rather only the current state. In other words such a strategy is described by a map from S to $P(A_i)$ – and is called a Markovian strategy. Additionally, when the map is independent of time, the strategy is called a stationary Markovian strategy, i. e., a stationary Markovian strategy for player i is described by a mapping $f_i\colon S \longrightarrow P(A_i)$.

Example 3 Consider, for starters, a pure strategy vector $\boldsymbol{\sigma}$, i. e., a pure strategy choice for every i. Suppose further that q the conditional distribution on states is also deterministic. In that case, there is, in a natural way, a unique history that is generated by $\boldsymbol{\sigma}$:

$$h(t;\boldsymbol{\sigma}) = s(0)\,,\quad a(0;\boldsymbol{\sigma})\,,\quad s(1;\boldsymbol{\sigma})\,,\quad a(1;\boldsymbol{\sigma})\,,\quad \ldots\,,\quad s(t;\boldsymbol{\sigma})$$

where $\boldsymbol{a}(\tau;\boldsymbol{\sigma}) = \boldsymbol{\sigma}(\tau; h(\tau;\boldsymbol{\sigma}))$ and $s(\tau+1;\boldsymbol{\sigma}) = q(s(\tau+1) \mid s(\tau;\boldsymbol{\sigma}), \boldsymbol{a}(\tau;\boldsymbol{\sigma}))$. This unique history associated with the strategy vector $\boldsymbol{\sigma}$ is also called the *outcome path* for that strategy. To every such outcome path there is an associated lifetime payoff

$$R_i(\boldsymbol{\sigma}) = \sum_{t=0}^{T} \delta^t \pi_i(s(\tau;\boldsymbol{\sigma}), \boldsymbol{a}(\tau;\boldsymbol{\sigma}))\,. \tag{1}$$

If $\boldsymbol{\sigma}$ is a mixed strategy, or if the conditional distribution q, is not deterministic, then there will be a joint distribution on the set of histories $H(t)$ generated by the strategy vector $\boldsymbol{\sigma}$ and the conditional distribution q in the obvious way. Moreover, there will be a marginal distribution on the state and action in period t, and under that marginal, an expected payoff $\pi_i(s(\tau;\boldsymbol{\sigma}), a(\tau;\boldsymbol{\sigma}))$. Thereafter lifetime payoffs can be written exactly as in Eq. 1.

Consider the game that remains after every history $h(t)$. This remainder is called a *subgame*. The restriction of the strategy vector $\boldsymbol{\sigma}$ to the subgame that starts after history $h(t)$, is denoted $\boldsymbol{\sigma} \mid h(t)$.

Equilibrium

A strategy vector $\boldsymbol{\sigma}^*$ is said to be a Nash Equilibrium (or NE) of the game if

$$R_i(\boldsymbol{\sigma}^*) \geq R_i(\boldsymbol{\sigma}_i, \boldsymbol{\sigma}^*_{-i})\,, \quad \text{for all } i, \boldsymbol{\sigma}_i\,. \tag{2}$$

A strategy vector $\boldsymbol{\sigma}^*$ is said to be a Subgame Perfect (Nash) Equilibrium of the game – referred to in short as SPE – if not only is Eq. 2 true for $\boldsymbol{\sigma}^*$ but it is true for every restriction of the strategy vector $\boldsymbol{\sigma}^*$ to every subgame $h(t)$, i. e., is true for $\boldsymbol{\sigma}^* \mid h(t)$ as well. In other words, $\boldsymbol{\sigma}^*$ is a SPE if

$$R_i(\boldsymbol{\sigma}^* \mid h(t)) \geq R_i(\boldsymbol{\sigma}_i, \boldsymbol{\sigma}^*_{-i} \mid h(t)), \quad \text{for all } i, \boldsymbol{\sigma}_i, h(t). \tag{3}$$

As is well-known, not all NE satisfy the further requirement of being a SPE. This is because a NE only considers the outcome path associated with that strategy vector $\boldsymbol{\sigma}^*$ – or, when the outcome path is probabilistic, only considers those outcome paths that have a positive probability of occurrence. That follows from the inequality Eq. 2. However, that does not preclude the possibility that players may have no incentive to follow through with $\boldsymbol{\sigma}^*$ if some zero probability history associated with that strategy is reached. (Such a history may be reached either by accident or because of deviation/experimentation by some player.) In turn that may have material relevance because how players behave when such a history is reached will have significance for whether or not a player wishes to deviate against $\boldsymbol{\sigma}^*$. Eq. 3 ensures that – even after a deviation – $\boldsymbol{\sigma}^*$ will get played and that deviations are unprofitable.

Recall the definition of a stationary Markovian strategy (SMS) above. Associated with that class of strategies is the following definition of equilibrium. A stationary Markov strategy vector $\boldsymbol{f}^*$ is a Markov Perfect Equilibrium (MPE) if

$$R_i(\boldsymbol{f}^*) \geq R_i(f_i, f^*_{-i}), \quad \text{for all } i, f_i .$$

Hence, a MPE restricts attention to SMS both on and off the outcome path. Furthermore, it only considers – implicitly – histories that have a positive probability of occurrence under $\boldsymbol{f}^*$. Neither "restriction" is a restriction when T is infinite because when all other players play a SMS player i has a stationary dynamic programming problem to solve in finding his most profitable strategy and – as is well-known – he loses no payoff possibilities in restricting himself to SMS as well. And that best strategy is a best strategy on histories that have zero probabilities of occurrence as well as histories that have a positive probability of occurrence. In particular therefore, when T is infinite, a MPE is also a SPE.

The Dynamic – or Stochastic – Game: Results

The main questions that we will now turn to are:

1. Is there always a SPE in a dynamic – or stochastic – game?
2. Is there a characterization for the set of SPE akin to the Bellman optimality equation of dynamic programming? If yes, what properties can be deduced of the SPE payoff set?
3. Is there a Folk Theorem for dynamic games – akin to that in Repeated Games?
4. What are the properties of SPE outcome paths?

The answers to questions 1–3 are very complete for finite dynamic games, i. e., games where the state space S is finite. The answer is also complete for questions 1 and 2 when S is countably infinite but when the state space is uncountably infinite, the question is substantively technically difficult and there is reason to believe that there may not always be a SPE. The finite game arguments for question 3 is conceptually applicable when S is (countably or uncountably) infinite provided some technical difficulties can be overcome. That and extending the first two answers to uncountably infinite S remain open questions at this point. Not a lot is known about Question 4.

Existence

The first result is due to Parthasarathy [23] and applies to the infinite horizon model, i. e., where $T = \infty$. When T is finite, the result and arguments can be modified in a straightforward way as will be indicated in the remarks following the proof.

Theorem 1 *Suppose that S is countable, A_i are finite sets, and the payoff functions π_i are bounded. Suppose furthermore that T is infinite. Then there is a MPE (and hence a SPE).*

Proof The proof will be presented by way of a fixed point argument. The domain for the fixed point will be the set of stationary Markovian strategies:

$$M_i = \{f_i : S \to P(A_i), \quad \text{s.t. for all } s, \sum_{\boldsymbol{a}_i} f_i(\boldsymbol{a}_i; s) = 1, f_i(\boldsymbol{a}_i; s) \geq 0\} .$$

□

Properties of M_i: In the pointwise convergence topology, M_i is compact. That this is so follows from a standard diagonalization argument by way of which a subsequence can be constructed from any sequence of SMS f_i^n such that the subsequence, call it $f_i^{n'}$ has the property that $f_i^{n'}(s)$ converges to some $f_i^0(s)$ for every s. Clearly, $f_i^0 \in M_i$. The diagonalization argument requires S to be countable and A_i to be finite.

M_i is also clearly convex since its elements are probability distributions on A_i at every state.

The mapping for which we shall seek a fixed point is the best response mapping:

$$B_i(f) = \{g_i \in M_i : R_i(g_i, f_{-i}) \geq R_i(f_i, f_{-i}), \quad \text{for all } f_i.$$

Since the best response problem for player i is a stationary dynamic programming problem, it follows that there is an associated value function for the problem, say v_i, such that it solves the optimality equation of dynamic programming

$$v_i(s) = \max_{\lambda_i} \left\{ \pi_i(s, \lambda_i, f_{-i}(s)) + \delta \sum_{s'} v_i(s') q\left(s' \mid s, \lambda_i, f_{-i}(s)\right) \right\} \quad (4)$$

where

$$\pi_i(s, \lambda_i, f_{-i}(s)) = \sum_{\mathbf{a}_{-i}} \left[\sum_{\mathbf{a}_i} \pi_i(s, a_i, a_{-i}) \lambda_i(a_i) \right] f_i(a_{-i}, s) \quad (5)$$

where $\lambda_i(a_i)$ is the probability of player i picking action a_i whilst $f_i(a_{-i}, s)$ is the product probability of players other than i picking the action vector $\boldsymbol{a}_{-i}$. Similarly.

$$q(s' \mid s, \lambda_i, f_{-i}(s)) = \sum_{\mathbf{a}_{-i}} \left[\sum_{\mathbf{a}_i} q(s' \mid s, a_i, a_{-i}) \lambda_i(a_i) \right] f_i(a_{-i}, s) \,. \quad (6)$$

Additionally, it follows that the best best response, i. e., g_i, solves the optimality equation, i. e.,

$$v_i(s) = \pi_i(s, g_i, f_{-i}(s)) + \delta \sum_{s'} v_i(s') q(s' \mid s, g_i, f_{-i}(s)) \quad (7)$$

where $\pi_i(s, g_i, f_{-i}(s))$ and $q(s' \mid s, g_i, f_{-i}(s))$ have the same interpretations as given by Eqs. 5 and 6.

Properties of B_i: B_i is a correspondence that is convex-valued and upper hemi-continuous. That B_i is convex-valued follows from the fact that we are operating in the set of mixed strategies, that every convex combination of mixed strategies is itself a mixed strategy and that every convex combination of best responses is also a best response.

To show that B_i is upper hemi-continuous, consider a sequence of other players' strategies f_{-i}^n, an associated best response sequence of player i, g_i^n with value function sequence v_i^n. Note that each of these best responses g_i^n satisfies the Eqs. 6 and 7 (for the value function v_i^n). By diagonalization there exist subsequences and subsequential pointwise convergent limits: $f_{-i}^n \longrightarrow f_{-i}^0$, $g_i^n \longrightarrow g_i^0$, and $v_i^n \longrightarrow v_i^0$. It suffices to show that

$$v_i^0(s) = \max_{\lambda_i} \left\{ \pi_i(s, \lambda_i, f_{-i}^0(s)) + \delta \sum_{s'} v_i^0(s') q(s' \mid s, \lambda_i, f_{-i}^0(s)) \right\} \quad (8)$$

and

$$v_i^0(s) = \pi_i(s, g_i^0, f_{-i}^0(s)) + \delta \sum_{s'} v_i^0(s') q(s' \mid s, g_i^0, f_{-i}^0(s))\}. \quad (9)$$

Equation. 9 will be proved by using the analog of Eq. 7, i. e.,

$$v_i^n(s) = \pi_i(s, g_i^n, f_{-i}^n(s)) + \delta \sum_{s'} v_i^n(s') q(s' \mid s, g_i^n, f_{-i}^n(s))\}. \quad (10)$$

Clearly the left-hand side of Eq. 10 converges to the left-hand side of Eq. 9. Lets check the right-hand side of each equation. Evidently

$$\sum_{\mathbf{a}_{-i}} \left[\sum_{\mathbf{a}_i} \pi_i(s, a_i, a_{-i}) g_i^n(a_i) \right] f_i^n(a_{-i}; s) \to \sum_{\mathbf{a}_{-i}} \left[\sum_{\mathbf{a}_i} \pi_i(s, a_i, a_{-i}) g_i^0(a_i) \right] f_i^0(a_{-i}; s)$$

since each component of the sum converges and we have a finite sum. Finally,

$$\begin{aligned} &\left| \sum_{s'} v_i^n(s') q(s' \mid s, g_i^n, f_{-i}^n(s)) - \sum_{s'} v_i^0(s') q(s' \mid s, g_i^0, f_{-i}^0(s)) \right| \\ &\leq \left| \sum_{s'} \left[v_i^n(s') - v_i^0(s') \right] q(s' \mid s, g_i^n, f_{-i}^n(s)) \right| \\ &+ \left| \sum_{s'} v_i^0(s') \left[q(s' \mid s, g_i^n, f_{-i}^n(s)) - q(s' \mid s, g_i^0, f_{-i}^0(s)) \right] \right| . \end{aligned} \quad (11)$$

The first term in the right-hand side of the inequality above goes to zero by the dominated convergence theo-

rem. The second term can be re-written as

$$\sum_{s'} v_i^0(s')q(s' \mid s, a_i, a_{-i}) \cdot \sum_{\mathbf{a}_{-i}} \sum_{\mathbf{a}_i} \left[g_i^n(a_i) f_i^n(a_{-i}; s) - g_i^0(a_i) f_i^0(a_{-i}; s)\right]$$

and goes to zero because each of the finite number of terms in the summation over action probabilities goes to zero. Hence the RHS of Eq. 10 converges to the RHS of Eq. 9 and the proposition is proved.

Remark 1 Note that the finiteness of A_i is crucial. Else, the very last argument would not go through, i. e., knowing that $\left[g_i^n(a_i) f_i^n(a_{-i}; s) - g_i^0(a_i) f_i^0(a_{-i}; s)\right] \to 0$ for every action vector $\boldsymbol{a}$ would not guarantee that the sum would converge to zero as well.

Remark 2 If the horizon were finite one could use the same argument to prove that there exists a Markovian strategy equilibrium, though not a stationary Markovian equilibrium. That proof would combine the arguments above with backward induction. In other words, one would first use the arguments above to show that there is an equilibrium at every state in the last period T. Then the value function so generated, v_i^T, would be used to show that there is an equilibrium in period $T-1$ using the methods above thereby generating the relevant value function for the last two periods, v_i^{T-1}. And so on.

The natural question to ask at this point is whether the restriction of countable finiteness of S can be dropped (and – eventually – the finiteness restriction on A_i). The answer, unfortunately, is not easily. The problems are two-fold:

1. *Sequential Compactness of the Domain Problem* – If S is uncountably infinite, then it is difficult to find a domain M_i that is sequentially compact. In particular, diagonalization arguments do not work to extract candidate strategy and value function limits.
2. *Integration to the Limit Problem* – Note as the other players change their strategies, f_{-i}^n, continuation payoffs to player i change in two ways. They change first because the value function v_i^n changes, i. e., $v_i^n \neq v_i^m$ if $n \neq m$. Second, the expected continuation value changes because the measure over which the value function is being integrated, $q(s' \mid s, \lambda_i, f_{-i}^n(s))$, itself changes, i. e., $q(s' \mid s, \lambda_i, f_{-i}^n(s)) \neq q(s' \mid s, \lambda_i, f_{-i}^m(s))$. This is the well-known – and difficult – integration to the limit problem: simply knowing that v_i^n "converges" to v_i^0 in some sense – such as pointwise – and knowing that the integrating measure q^n "converges" to q^0 in some sense – such as in the weak topology – does not, in general, imply that

$$\int v_i^n \mathrm{d}q^n \to \int v_i^0 \mathrm{d}q^0 \,. \tag{12}$$

(Of course in the previous sentence q^n is a more compact stand-in for $q(s' \mid s, \lambda_i, f_{-i}^n(s))$ and q^0 for $q(s' \mid s, \lambda_i, f_{-i}^0(s))$.) There are a limited number of cases where Eq. 12 is known to be true. These results typically require q^n to converge to q^0 in some strong sense. In the dynamic game context what this means is that very strong convergence restrictions need to be placed on the transition probability q. This is the underlying logic behind results reported in [7,20,22,24].

Such strong convergence properties are typically not satisfied when q is deterministic – which case comprises the bulk of the applications of the theory. Indeed simply imposing continuity when q is deterministic appears not to be enough to generate an existence result. Harris et al. [16] and Dutta and Sundaram [14] contain results that show that there may not be a SPE in finite horizon dynamic games when the transition function q is continuous. Whether other often used properties of q and π_i – such as concavity and monotonicity – can be used to rescue the issue remains an open question.

Characterization

The Bellman optimality equation has become a workhorse for dynamic programming analysis. It is used to derive properties of the value function and the optimal strategies. Moreover it provides an attractive and conceptually simple way to view a multiple horizon problem as a series of one-stage programming problems by exploiting the recursive structure of the optimization set-up. A natural question to ask, since dynamic games are really multi-player versions of dynamic programming, is whether there is an analog of the Bellman equation for these games. Abreu, Pearce and Stachetti – APS (1988), in an important and influential paper, showed that this is indeed the case for repeated games. They defined an operator, hereafter the APS operator, whose largest fixed point is the set of SPE payoffs in a repeated game and whose every fixed point is a subset of the set of SPE payoffs. (Thereby providing a necessary and sufficient condition for SPE equilibrium payoffs in much the same way that the unique fixed point of the Bellman operator constitutes the value function for a dynamic programming problem.) As with the Bellman equation, the key idea is to reduce the multiple horizon problem to a (seemingly) static problem.

In going from repeated to dynamic games there are some technical issues that arise. We turn now to that analysis pointing out along the way where the technical pitfalls are. Again we start with the infinite horizon model, i. e., where $T = \infty$. When T is finite, the result and arguments can be modified in a straightforward way as will be indicated in a remark following the proof.

But first, some definitions. Suppose for now that S is countable.

APS Operator – Consider a compact-valued correspondence, W defined on domain S which takes values that are subsets of $\mathbb{R}^I$. Define the APS operator on W, call it LW, as follows:

$$LW(s) = \left\{ \begin{array}{l} v \in \mathbb{R}^I \colon \exists \widehat{f} \in P(A) \quad \text{and} \\ w \colon S \times A \times S \to W\,, \\ \text{uniformly bounded, s.t.} \;\; v_i = \pi_i(s, \widehat{f}) \\ +\delta \sum_{s'} w_i(s, \widehat{f}, s') q(s' \mid s, \widehat{f}) \\ \geq \pi_i(s, a_i, \widehat{f_{-i}}) + \delta \sum_{s'} w_i(s, a_i, \widehat{f_{-i}}, s') \\ q(s' \mid s, a_i, \widehat{f_{-i}})\,, \quad \text{for all} \;\; a_i, i \end{array} \right\} \tag{13}$$

where, as before,

$$\pi_i(s, \widehat{f}) = \sum_{\mathbf{a}_{-i}} \left[\sum_{\mathbf{a}_i} \pi_i(s, a_i, a_{-i}) \widehat{f}_i(a_i) \right] \widehat{f}_{-i}(a_{-i}; s) \tag{14}$$

and

$$q(s' \mid s, \widehat{f}) = \sum_{\mathbf{a}_{-i}} \left[\sum_{\mathbf{a}_i} q(s' \mid s, a_i, a_{-i}) \widehat{f}_i(a_i) \right] \widehat{f}_{-i}(a_{-i}; s) \tag{15}$$

Finally, let V^* denote the SPE payoffs correspondence, i. e., $V^*(s)$ is the set of SPE payoffs starting from initial state s. Note that the correspondence is non-empty by virtue of Theorem 1.

Theorem 2 *Suppose that S is countable, A_i are finite sets, and the payoff functions π_i are bounded. Suppose furthermore that T is infinite. Then a) V^* is a fixed point of the APS operator, i. e., $LV^* = V^*$. Furthermore, b) consider any other fixed point, i. e., a correspondence $\widetilde{V}$ such that $L\widetilde{V} = \widetilde{V}$. Then it must be the case that $\widetilde{V} \subset V^*$. Finally, c) there is an algorithm that generates the SPE correspondence, V^*.*

Proof of a: Suppose that $v^* \in V^*(s)$, i. e., is a SPE payoff starting from s. Then, by definition, there is a first-period play, f^* and a continuation strategy after every one-period history $h(1)$, $\boldsymbol{\sigma}^*(h(1))$, such that $\{\boldsymbol{f}^*, \boldsymbol{\sigma}^*(h(1))\}$ is a SPE. By definition, then, the payoffs associated with each history-dependent strategy, $\boldsymbol{\sigma}^*(h(1))$, call them $w_i(s, \boldsymbol{f}^*, s')$, satisfy Eq. 13. (Note that π_i bounded implies that the lifetime payoffs $w_i(s, \boldsymbol{f}^*, s')$ are uniformly bounded. In other words, $v^* \in LV^*(s)$. On the other hand, suppose that $v^* \in LV^*(s)$. Then, by definition, there is a first-period play, f^*, and SPE payoffs, $w(s, \boldsymbol{f}^*, s')$, that together satisfy Eq. 13. Let the SPE strategy associated with $w(s, \boldsymbol{f}^*, s')$ be $\boldsymbol{\sigma}^*(h(1))$. It is not difficult to see that the concatenated strategy – $\boldsymbol{f}^*, \boldsymbol{\sigma}^*(h(1))$ – forms a SPE. Since the associated lifetime payoff is v^*, it follows that $v^* \in V^*(s)$. □

Proof of b: Suppose that $L\widetilde{V} = \widetilde{V}$. Let $v \in L\widetilde{V}$. Then, from Eq. 13, there is $\widetilde{f}$, and SPE payoffs, $w(s, \widetilde{f}, s')$ that satisfy the equation. In turn, since $w(s, \widetilde{f}, s') \in \widetilde{V}(s') = L\widetilde{V}(s')$ there is an associated $\overline{f}(s')$ and $w(s', \overline{f}, s'')$ for which Eq. 13 holds. By repeated application of this idea, we can create a sequence of strategies for periods $t = 0, 1, 2, \ldots - \widetilde{f}$, $\overline{f}(s')$, $\overrightarrow{f}(s, s'), \ldots$ such that at each period Eq. 13 holds. Call the strategy so formed, ϕ. This strategy can then not be improved upon by a single-period deviation. A standard argument shows that if a strategy cannot be profitably deviated against in one period then it cannot be profitably deviated against even by deviations in multiple periods. (This idea of "unimprovability" is already present in dynamic programming. Within the context of repeated games, it was articulated by Abreu [1].) □

Proof of c: Note two properties of the APS operator:

Lemma 1 *LW is a compact-valued correspondence (whenever W is compact-valued).*

Proof Consider Eq. 13. Suppose that $v^n \in LW(s)$ for all n, with associated $\widehat{f}^n$ and w^n. By diagonalization, there exists a subsequence s.t. $v^n \to v^0$, $\widehat{f}^n \to \widehat{f}^0$ and $w^n \to w^0$. This arguments uses the countability of S and the finiteness of A_i. From Eq. 14 evidently $\pi_i(s, \widehat{f}^n) \to \pi_i(s, \widehat{f}^0)$ and similarly from Eq. 15 $\sum_{s'} w_i(s, , \widehat{f}^n, s') q(s' \mid s, \widehat{f}^n)$ goes to $\sum_{s'} w_i(s, \widehat{f}^0, s') q(s' \mid s, , \widehat{f}^0)$. Hence the inequality in Eq. 13 is preserved and $v^0 \in LW(s)$. □

It is not difficult to see that – on account of the boundedness of π_i – if W has a uniformly bounded selection, then so does LW. Note that the operator is also monotone in the set-inclusion sense, i. e., if $W'(s) \subset W(s)$ for all s then $LW' \subset LW$.

The APS algorithm finds the set of SPE payoffs by starting from a particular starting point, an initial set $W^0(s)$ that is taken to be the set of all feasible payoffs from initial state s. (And hence the correspondence W^0 is so

defined for every initial state.) Then define, $W^1 = LW^0$. More generally, $W^{n+1} = LW^n$, $n \geq 0$. It follows that $W^1 \subset W^0$. This is because W^1 requires a payoff that is not only feasible but additionally satisfies the incentive inequality of Eq. 13 as well. From the monotone inclusion property above it then follows that, more generally, $W^{n+1} \subset W^n$, $n \geq 0$. Furthermore, $W^n(s)$ is a non-empty, compact set for all n (and s). Hence, $W^\infty(s) = \cap_n W^n(s) = \lim_{n\to\infty} W^n(s)$ is non-empty and compact.

Let us now show that W^∞ is a fixed point of the APS operator, i. e., that $LW^\infty = W^\infty$.

Lemma 2 *$LW^\infty = W^\infty$, or, equivalently, $L(\lim_{n\to\infty} W^n) = \lim_{n\to\infty} LW^n$.*

Proof Clearly, by monotonicity, $L(\lim_{n\to\infty} W^n) \subset \lim_{n\to\infty} LW^n$. So consider a $v \in LW^n(s)$, for all n. By Eq. 13 there is at each n an associated first-period play f^n and a continuation payoff $w^n(s, f^n, s')$ such that the inequality is satisfied and

$$v_i = \pi_i(s, f^n) + \delta \sum_{s'} w_i^n(s, f^n, s') q(s' \mid s, f^n) .$$

By the diagonalization argument, and using the countability of S, we can extract a (subsequential) limit $f^\infty = \lim_{n\to\infty} f^n$ and $w^\infty = \lim_{n\to\infty} w^n$. Clearly, $w^\infty \in W^\infty$. Since equalities and inequalities are maintained in the limit, equally clearly

$$v_i = \pi_i(s, f^\infty) + \delta \sum_{s'} w_i^\infty(s, f^\infty, s') q(s' \mid s, f^\infty)$$

and

$$v_i \geq \pi_i(s, a_i, f_{-i}^\infty) + \delta \sum_{s'} w_i^\infty \cdot (s, a_i, f_{-i}^\infty, s') q(s' \mid s, a_i, f_{-i}^\infty) , \quad \text{for all } a_i , i$$

thereby proving that $v \in L(\lim_{n\to\infty} W^n)(s)$. The lemma is proved. □

Since the set of SPE payoffs, $V^*(s)$, is a subset of $W^0(s)$ – and $LV^*(s) = V^*(s)$ – it further follows $V^*(s) \subset W^\infty(s)$, for all s. From the previous lemma, and part b), it follows that $V^*(s) \supset W^\infty(s)$, for all s. Hence, $V^* = W^\infty$. Theorem 2 is proved. □

A few remarks are in order.

Remark 1 If the game horizon T is finite, there is an immediate modification of the above arguments. In the algorithm above, take W^0 to be the set of SPE payoffs in the one-period game (with payoffs π_i for player i). Use the APS operator thereafter to define $W^{n+1} = LW^n$, $n \geq 0$. It is not too difficult to show that W^n is the set of SPE payoffs for a game that lasts $n + 1$ periods (or has n remaining periods after the first one).

Remark 2 Of course an immediate corollary of the above theorem is that the set of SPE payoffs $V^*(s)$ is a compact set for every initial state s. Indeed one can go further and show that V^* is in fact an upper hemi-continuous correspondence. The arguments are very similar to those used above – plus the Maximum Theorem.

Remark 3 Another way to think of Theorem 2 is that it is also an existence theorem. Under the conditions outlined in the result, the SPE equilibrium set has been shown to be non-empty. Of course this is not a generalization of Theorem 1 since Theorem 2 does not assert the existence of a MPE.

Remark 4 When the state space A_i is infinite or the state space S is uncountably infinite we run into technical difficulties. The complications arise from not being able to take limits. Also, as in the discussion of the Integration to the Limit problem, integrals can fail to be continuous thereby rendering void some of the arguments used above.

Folk Theorem

The folk theorem for Repeated Games – Fudenberg and Maskin [15] following up on earlier contributions – is very well-known and the most cited result of that theory. It proves that the necessary conditions for a payoff to be a SPE payoff – feasibility and individual rationality – are also (almost) sufficient provided the discount factor δ is close enough to 1. This is the result that has become the defining result of Repeated Games. For supporters, the result and its logic of proof are a compelling demonstration of the power of reciprocity, the power of long-term relationships in fostering cooperation through the lurking power of "punishments" when cooperation breaks down. It is considered equally important and significant that such long-term relationships and behaviors are sustained through implicit promises and threats which therefore do not violate any legal prohibitions against explicit contracts that specify such behavior. For detractors, the "anything goes" implication of the Folk Theorem is a clear sign of its weakness – or the weakness of the SPE concept – in that it robs the theory of all predictive content. Moreover there is a criticism, not entirely correct, that the strategies required to sustain certain behaviors are so complex that no player in a "real-world" setting could be expected to implement them.

Be that as it may, the Folk Theorem question in the context of Dynamic Games then is: is it the case that feasibility and individual rationality are also (almost) enough to guarantee that a payoff is a SPE payoff at high enough δ? Two sets of obstacles arise in settling this question. Both emanate from the same source, the fact that the state does not remain fixed in the play of the games, as it does in the case of Repeated Games. First, one has to think long and hard as to how one should define individual rationality. Relatedly, how does one track feasibility? In both cases, the problem is that what payoff is feasible and individually rational depends on the state and hence changes after every history $h(t)$. Moreover, it also changes with the discount factor δ. The second set of problems stems from the fact that a deviation play can unalterably change the future in a dynamic game – unlike a repeated game where the basic game environment is identical every period. Consequently one cannot immediately invoke the logic of repeated game folk theorems which basically work because any deviation has only short-term consequences while the punishment of the deviation is long-term. (And so if players are patient they will not deviate.)

Despite all this, there are some positive results that are around. Of these, the most comprehensive is one due to Dutta [9]. To set the stage for that result, we need a few crucial preliminary results. For this sub-section we will assume that S is finite – in addition to A_i.

Feasible Payoffs Role of Markovian Strategies – Let $F(s,\delta)$ denote the set of "average" feasible payoffs from initial state s and for discount factor δ. By that I mean

$$F(s,\delta) = \{v \in \mathbb{R}^I : \exists \text{ strategy } \boldsymbol{\sigma}$$
$$\text{s.t. } v = (1-\delta)\sum_{t=0}^{T} \delta^t \pi_i(s(t;\boldsymbol{\sigma}), a(t;\boldsymbol{\sigma}))\} .$$

Let $\Phi(s,\delta)$ denote the set of "average" feasible payoffs from initial state s and for discount factor δ that are generated by pure stationary Markovian strategies – PSMS. Recall that a SMS is given by a map f_i from S to the probability distributions over A_i, so that at state $s(t)$ player i chooses the mixed strategy $f_i(s(t))$. A pure SMS is one where the map f_i is from S to A_i. In other words,

$$\Phi(s,\delta) = \{v \in \mathbb{R}^I : \exists \text{ PSMS } f$$
$$\text{s.t. } v = (1-\delta)\sum_{t=0}^{T} \delta^t \pi_i(s(t;f), a(t;f))\} .$$

Lemma 3 *Any feasible payoff in a dynamic game can be generated by averaging over payoffs to stationary Markov strategies, i. e., $F(s,\delta) = co\Phi(s,\delta)$, for all (s,δ).*

Proof Note that $F(s,\delta) = co[\text{extreme points } F(s,\delta)]$. In turn, all extreme points of $F(s,\delta)$ are generated by an optimization problem of the form: $\max_\sigma \sum_{i=1}^{I} \alpha_i v_i(s,\delta)$. That optimization problem is a dynamic programming problem. Standard results in dynamic programming show that the optimum is achieved by some stationary Markovian strategy. □

Let $F(s)$ denote the set of feasible payoffs under the long-run average criterion. The next result will show that this is the set to which discounted average payoffs converge:

Lemma 4 *$F(s,\delta) \to F(s)$, as $\delta \to 1$, for all s.*

Proof Follows from the fact that a) $F(s) = co\Phi(s)$ where $\Phi(s)$ is the set of feasible long-run average payoffs generated by stationary Markovian strategies, and b) $\Phi(s,\delta) \to \Phi(s)$. Part b) exploits the finiteness of S (and A_i). □

The lemmas above simplify the answer to the question: What is a feasible payoff in a dynamic game? Note that they also afford a dimensional reduction in the complexity and number of strategies that one needs to keep track of to answer the question. Whilst there are an uncountably infinite number of strategies – even with finite S and A_i – including the many that condition on histories in arbitrarily complex ways – the lemmas establish that all we need to track are the finite number of PSMS. Furthermore, whilst payoffs do depend on δ, if the discount factor is high enough then the set of feasible payoffs is well-approximated by the set of feasible long-run average payoffs to PSMS.

One further preliminary step is required however. This has to do with the fact that while $v \in F(s,\delta)$ can be exactly reproduced by a period 0 average over PSMS payoffs, after that period continuation payoffs to the various component strategies may generate payoffs that could be arbitrarily distant from v. This, in turn, can be problematical since one would need to check for deviations at every one of these (very different) payoffs. The next lemma addresses this problem by showing that there is an averaging over the component PSMS that is ongoing, i. e., happens periodically and not just at period 0, but which, consequently, generates payoffs that after all histories stays arbitrarily close to v.

For any two PSMS f^1 and f^2 denote a *time-cycle strategy* as follows: for T^1 periods play proceeds along f^1, then it moves for T^2 periods to f^2. After the elapse of the $T^1 + T^2$ periods play comes back to f^1 for T^1 periods and f^2 for T^2 periods. And so on. Define $\lambda^1 = T^1/(T^1 + T^2)$. In the obvious way, denote a general time-cycle strategy to be one that cycles over any finite number of PSMS f^k

where the proportion of time spent at strategy f^k is λ^k and allows the lengths of time to depend on the initial state at the beginning of the cycle.

Lemma 5 *Pick any $v \in \cap_s F(s)$. Then for all $\varepsilon > 0$ there is a time cycle strategy such that its long-run average payoff is within ε of v after all histories.*

Proof Suppose that $v = \sum_k \lambda^k(s) v^k(s)$ where $v^k(s)$ is the long-run average payoff to the kth PSMS when the initial state is s. Ensure that T^k is chosen such that a) the average payoff over those periods under that PSMS – $1/T^k \sum_{t=0}^{T^k-1} \pi_i(s(t; f^k), a(t; f^k))$ – is within ε of $v^k(s)$ for all s. And b) that $T^k(s)/\sum_l T^l(s)$ is arbitrarily close to $\lambda^k(s)$ for all s. □

Since $\Phi(s, \delta) \to \Phi(s)$ it further follows that the above result also holds under discounting:

Lemma 6 *Pick any $v \in \cap_s F(s)$. Then for all $\varepsilon > 0$ there is a time cycle strategy and a discount cut-off $\delta(\varepsilon) < 1$ such that the discounted average payoffs to that strategy are within ε of v for all $\delta > \delta(\varepsilon)$ and after all histories.*

Proof Follows from the fact that $(1-\delta)/(1-\delta^T) \sum_{t=0}^{T^k-1} \delta^t \pi_i(s(t; f^k), a(t; f^k))$ goes to $\frac{1}{T^k} \sum_{t=0}^{T^k-1} \pi_i(s(t; f^k), a(t; f^k))$ as $\delta \to 1$. □

Individually Rational Payoffs Recall that a min-max payoff is a payoff level that a player can guarantee by playing a best response. In a Repeated Game that is defined at the level of the component stage game. Since there is no analog of that in a dynamic game, the min-max needs to be defined over the entire game – and hence is sensitive to initial state and discount factor:

$$m_i(s, \delta) = \min_{\boldsymbol{\sigma}_{-i}} \max_{\boldsymbol{\sigma}_i} R_i(\boldsymbol{\sigma} \mid s, \delta) .$$

Evidently, given (s, δ), in a SPE it cannot be that player i gets a payoff $v_i(s, \delta)$ that is less than $m_i(s, \delta)$. Indeed that inequality must hold at all states for a strategy to be a SPE, i. e., for all $s(t)$ it must be the case that $v_i(s(t), \delta) \geq m_i(s(t), \delta)$. But, whilst necessary, even that might not be a sufficient condition for the strategy to be a SPE. The reason is that if player i can deviate and take the game to, say, s' at $t+1$, rather than $s(t+1)$, he would do so if $v_i(s(t), \delta) < m_i(s', \delta)$ since continuation payoffs from s' have to be at least as large as the latter level and this deviation would be worth essentially that continuation when δ is close to 1. So sufficiency will require a condition such as $v_i(s(t), \delta) > \max_s m_i(s, \delta)$ for all $s(t)$. Call such a strategy *dynamically Individually Rational*. From the previous lemmas, and the fact that $m_i(s, \delta) \to m_i(s)$, as $\delta \to 1$, where $m_i(s)$ is the long-run average min-max level for player i the following result is obvious. The min-max limiting result is due to Mertens and Neyman [21].

Lemma 7 *Pick any $v \in \cap_s F(s)$ such that $v_i > \max_s m_i(s)$ for all s. Then there is a time-cycle strategy which is dynamically Individually Rational for high δ.*

We are now ready to state and prove the main result:

Theorem 3 (Folk Theorem) *Suppose that S and A_i are finite sets. Suppose furthermore that T is infinite and that $\cap_s F(s)$ has dimension I (where I is the number of players). Pick any $v \in \cap_s F(s)$ such that $v_i > \max_s m_i(s)$ for all s. Then, for all $\varepsilon > 0$, there is a discount cut-off $\delta(\varepsilon) < 1$ and a time-cycle strategy that for $\delta > \delta(\varepsilon)$ is a SPE with payoffs that are within ε of v.*

Proof Without loss of generality, let us set $\max_s m_i(s) = 0$ for all i. From the fact that $\cap_s F(s)$ has dimension I it follows that we can find I payoff vectors in that set – v^i, $i = 1, \ldots, I$ – such that for all i a) $v^i \gg 0$, b) $v_i^j > v_i^i$, $j \neq i$, and c) $v_i > v_i^i$. That we can find these vectors such that b) is satisfied follows from the dimensionality of the set. That we can additionally get the vectors to satisfy a) and c) follows from the fact that it is a convex set and hence an appropriate "averaging" with a vector such as $\mathbf{v}$ achieves a) while an "averaging" with $i's$ worst payoff achieves c).□

Now consider the following strategy: Norm – Start with a time-cycle strategy that generates payoffs after all histories that are within ε of v. Choose a high enough δ as required. Continue with that strategy if there are no deviations against it. Punishment – If there is, say if player i deviates, then min-max i for T periods and thereafter proceed to the time-cycle strategy that yields payoffs within ε of v^i after all histories. Re-start the punishment whenever there is a deviation.

Choose T in such a fashion that the payoff to the min-max period plus v_i^i is strictly less than v_i. That ensures there is no incentive to deviate against the norm provided the punishment is carried out. That there is incentive for players $j \neq i$ to punish player i follows from the fact that $v_j^i > v_j^j$ the former payoff being what they get from punishing and the latter from not punishing i. That there is incentive for player i not to deviate against his own punishment follows from the fact that re-starting the punishment only lowers his payoffs. The theorem is proved.

A few remarks are in order.

Remark 1 If the game horizon T is finite, there is likely a Folk Theorem along the lines of the result proved for

Repeated Games by Benoit and Krishna [4]. To the best of my knowledge it remains, however, an open question.

Remark 2 When the state space A_i is infinite or the state space S is uncountably infinite we again run into technical difficulties. There is an analog to Lemmas 3 and 4 in this instance and under appropriate richer assumptions the results can be generalized – see Dutta (1993). Lemmas 5–7 and the Folk Theorem itself does use the finiteness of S to apply uniform bounds to various approximations and those become problematical when the state space is infinite. It is our belief that nevertheless the Folk Theorem can be proved in this setting. It remains, however, to be done.

Dynamics

Recall that the fourth question is: what can be said about the dynamics of SPE outcome paths? The analogy that might be made is to the various convergence theorems – sometimes also called "turnpike theorems" – that are known to be true in single-player dynamic programming models. Now even within those models – as has become clear from the literature of the past twenty years in chaos and cycles theory for example – it is not always the case that there are regularities exhibited by the optimal solutions. Matters are worse in dynamic games.

Even within some special models where the single-player optima are well-behaved, the SPE of the corresponding dynamic game need not be. A classic instance is the neo-classical aggregative growth model. In that model, results going back fifty years show that the optimal solutions converge monotonically to a steady-state, the so-called "golden rule". (For references, see Majumdar, Mitra and Nishimura (2000).) However, examples can be constructed – and may be found in Dutta and Sundaram (1996) and Dockner, Long and Sorger (1998) – where there are SPE in these models that can have arbitrarily complex state dynamics which for some range of discount factor values descend into chaos. And that may happen with Stationary Markov Perfect Equilibrium. (It would be less of a stretch to believe that SPE in general can have complex dynamics. The Folk Theorem already suggests that it might be so.)

There are, however, many questions that remain including the breadth of SPE that have regular dynamics. One may care less for complex dynamic SPE if it can be shows that the "good ones" have regular dynamics. What also remains to be explored is whether adding some noise in the transition equation can remove most complex dynamics SPE.

Global Climate Change – Issues, Models

Issues

The dramatic rise of the world's population in the last three centuries, coupled with an even more dramatic acceleration of economic development in many parts of the world, has led to a transformation of the natural environment by humans that is unprecedented in scale. In particular, on account of the greenhouse effect, *global warming* has emerged as a central problem, unrivaled in its potential for harm to life as we know it on planet Earth. Seemingly the consequences are everywhere: melting and break-up of the world's ice-belts whether it be in the Arctic or the Antarctic; heat-waves that set all-time temperature highs whether it be in Western Europe or sub-Saharan Africa; storms increased in frequency and ferocity whether it be Hurricane Katrina or typhoons in Japan or flooding in Mumbai. In addition to Al Gore's eminently readable book, "An Inconvenient Truth", two authoritative recent treatments are the Stern Review on the Economics of Climate Change, October, 2006 and the IPCC Synthesis Report, November, 2007. Here are three – additional – facts drawn from the IPCC Report:

1. Eleven of the last twelve years (1995–2006) have been amongst the twelve warmest years in the instrumental record of global surface temperatures (since 1850).
2. If we go on with "Business as Usual", by 2100 global sea levels will probably have risen by 9 to 88 cm and average temperatures by between 1.5 and 5.5°C.
   Various factors contribute to global warming, but the major one is an increase in greenhouse gases (GHGs) – primarily, carbon dioxide – so called because they are transparent to incoming shortwave solar radiation but trap outgoing longwave infrared radiation. Increased carbon emissions due to the burning of fossil fuel is commonly cited as the principal immediate cause of global warming. A third relevant fact is:
3. Before the Industrial Revolution, atmospheric CO_2 concentrations were about 270–280 parts per million (ppm). They now stand at almost 380 ppm, and have been rising at about 1.5 ppm annually.

The IPCC Synthesis (2007) says "Warming of the climate system is unequivocal, as is now evident from observations of increases in global average air and ocean temperatures, widespread melting of snow and ice, and rising global average sea level." (IPCC Synthesis Report [17]).

It is clear that addressing the global warming problem will require the coordinated efforts of the world's nations. In the absence of an international government, that coordination will have to be achieved by way of an inter-

national environmental treaty. For a treaty to be implemented, it will have to align the incentives of the signatories by way of rewards for cutting greenhouse emissions and punishments for not doing so. For an adequate analysis of this problem one needs a dynamic and fully strategic approach. A natural methodology for this then is the theory of Subgame Perfect (Nash) equilibria of dynamic games – which we have discussed at some length in the preceding sections.

Although there is considerable uncertainty about the exact costs of global warming, the two principal sources will be a rise in the sea-level and climate changes. The former may wash away low-lying coastal areas such as Bangladesh and the Netherlands. Climate changes are more difficult to predict; tropical countries will become more arid and less productive agriculturally; there will be an increased likelihood of hurricanes, fires and forest loss; and there will be the unpredictable consequences of damage to the natural habitat of many living organisms. On the other hand, emission abatement imposes its own costs. Higher emissions are typically associated with greater GDP and consumer amenities (via increased energy usage). Reducing emissions will require many or all of the following costly activities: cutbacks in energy production, switches to alternative modes of production, investment in more energy-efficient equipment, investment in R&D to generate alternative sources of energy, etc.

The principal features of the global warming problem are:

- *The Global Common* – although the sources of carbon buildup are localized, it is the total stock of GHGs in the global environment that will determine the amount of warming.
- *Near-irreversibility* – since the stock of greenhouse gases depletes slowly, the effect of current emissions can be felt into the distant future.
- *Asymmetry* – some regions will suffer more than others.
- *Nonlinearity* – the costs can be very nonlinear; a rise in one degree may have little effect but a rise in several degrees may be catastrophic.
- *Strategic Setting* – Although the players (countries) are relatively numerous, there are some very large players, and blocks of like-minded countries, like the US, Western Europe, China, and Japan. That warrants a strategic analysis.

The theoretical framework that accommodates all of these features is an *asymmetric dynamic commons* model with the global stock of greenhouse gases as the (common) state variable. The next sub-section will discuss a few models which have most of the above characteristics.

Models

Before presenting specific models, let us briefly relate the climate change problem to the general dynamic game model that we have seen so far, and provide a historical outline of its study. GHGs form – as we saw above – a global common. The study of global commons is embedded in dynamic commons game (DCG). In such a game the state space S is a single-dimensional variable with a "commons" structure meaning that each player is able to change the (common) state. In particular, the transition function is of the form

$$s(t+1) = q\left(s(t) - \sum_{i=1}^{I} a_i(t)\right) .$$

The first analysis of a DCG may be found in [18]. That paper considered the particular functional form is which $q(s(t) - \sum_{i=1}^{I} a_i(t)) = [s(t) - \sum_{i=1}^{I} a_i(t)]^{\alpha}$ for a fixed fraction α. (And, additionally, Levhari and Mirman assumed the payoffs π_i to be logarithmic.) Consequently, the paper was able to derive in closed form a (linear) MPE and was able to analyze its characteristics.

Subsequently several authors – Sundaram [31], Sobel [27], Benhabib and Radner [3], Rustichini [25], Dutta and Sundaram (1992, [14]), Sorger [28] – studied this model in great generality, without making the specific functional form assumption of Levhari and Mirman, and established several interesting qualitative properties relating to existence of equilibria, welfare consequences and dynamic paths.

More recently in a series of papers by Dutta and Radner on the one hand and Dockner and his co-authors on the other, the DCG model has been directly applied to environmental problems including the problem of global warming. We shall describe the Dutta and Radner work in detail and also discuss some of the Dockner, Long and Sorger research. In particular, the transition equation is identical.in the two models (and described below). What is different is the payoff functions.

We turn now to a simplified climate change model to illustrate the basic strategic ideas. The model is drawn from Dutta and Radner [12]. In the basic model there is no population growth and no possibility of changing the emissions producing technologies in each country. (Population growth is studied in Dutta and Radner [11] while certain kinds of technological changes are allowed in Dutta and Radner [10]. These models will be discussed later.) However, the countries may differ in their "sizes", their emissions technologies, and their preferences.

There are I countries. The emission of (a scalar index of) greenhouse gases during period t by country i is de-

noted by $a_i(t)$. [Time is discrete, with $t = 0, 1, 2, \ldots$, ad inf.] Let $A(t)$ denote the global (total) emission during period t;

$$A(t) = \sum_{i=1}^{I} a_i(t) \,. \tag{16}$$

The total (global) stock of greenhouse gases (GHGs) at the beginning of period t is denoted by $g(t)$. (Note, for mnemonic purposes we are denoting the state variable – the amount of "gas" – g.) The law of motion – or transition function q in the notation above – is

$$g(t+1) = A(t) + \boldsymbol{\sigma}\, g(t) \,, \tag{17}$$

where $\boldsymbol{\sigma}$ is a given parameter ($0 < \boldsymbol{\sigma} < 1$). We may interpret $(1 - \boldsymbol{\sigma})$ as the fraction of the beginning-of-period stock of GHG that is dissipated from the atmosphere during the period. The "surviving" stock, $\boldsymbol{\sigma}\, g(t)$, is augmented by the quantity of global emissions, $A(t)$, during the same period.

Suppose that the payoff of country i in period t is

$$\pi_i(t) = h_i[a_i(t)] - c_i g(t) \,. \tag{18}$$

The function h_i represents, for example, what country i's gross national product would be at different levels of its own emissions, holding the global level of GHG constant. This function reflects the costs and benefits of producing and using energy as well as the costs and benefits of other activities that have an impact on the emissions of GHGs, e. g., the extent of forestation. It therefore seems natural to assume that h_i is a strictly concave C^2 function that reaches a maximum and then decreases thereafter.

The parameter $c_i > 0$ represents the marginal cost to the country of increasing the global stock of GHG. Of course, it is not the stock of GHG itself that is costly, but the associated climatic conditions. As discussed below, in a more general model, the cost would be nonlinear.

Histories, strategies – Markovian strategies – and outcomes are defined in exactly the same way as in the general theory above – and will, hence, not be repeated. Thus associated with each strategy vector $\boldsymbol{\sigma}$ is a total discounted payoff for each player

$$v_i(\boldsymbol{\sigma}, g_0) \equiv \sum_{t=0}^{\infty} \delta^t \pi_i(t; \boldsymbol{\sigma}, g_0) \,.$$

Similarly, SPE and MPE can be defined in exactly the same way as in the general theory.

The linearity of the model is undoubtedly restrictive in several ways. It implies that the model is unable to analyze catastrophes or certain kinds of feedback effects running back from climate change to economic costs. It has, however, two advantages: first, its conclusions are simple, can be derived in closed-form and can be numerically calibrated; hence may have a chance of informing policy-makers. Second, there is little consensus on what is the correct form of non-linearity in costs. Partly the problem stems from the fact that some costs are not going to be felt for another fifty to hundred years and forecasting the nature of costs on that horizon length is at best a hazardous exercise. Hence, instead of postulating one of many possible non-linear cost functions, all of which may turn out to be incorrect for the long-run, one can opt instead to work with a cost function which may be thought of as a linear approximation to any number of actual non-linear specifications.

Dockner, Long and Sorger (1998) impose linearity in the emissions payoff function h (whereas in Dutta and Radner it is assumed to be strictly concave) while their cost to g is strictly convex (as opposed to the above specification in which it is linear). The consequent differences in results we will discuss later.

Global Climate Change – Results

In this section we present two sets of results from the Dutta and Radner [12] paper. The first set of results characterize two benchmarks – the global Pareto optima, and a simple MPE, called "Business As Usual" and compares them. The second set of results then characterizes the entire SPE correspondence and – relatedly – the best and worst equilibria. Readers are referred to that paper for further results from this model and for a numerical calibration of the model. Furthermore, for the results that are presented, the proofs are merely sketched.

Global Pareto Optima

Let $x = (x_i)$ be a vector of positive numbers, one for each country. A *Global Pareto Optimum (GPO)* corresponding to x is a profile of strategies that maximizes the weighted sum of country payoffs,

$$v = \sum_i x_i v_i \,, \tag{19}$$

which we shall call *global welfare*. Without loss of generality, we may take the weights, x_i, to sum to I.

Theorem 4 *Let $\hat{V}(g)$ be the maximum attainable global welfare starting with an initial GHG stock equal to g. That*

function is linear in g;

$$\hat{V}(g) = \widehat{u} - wg\,,$$
$$w = \frac{1}{1-\delta\boldsymbol{\sigma}}\sum_i x_i c_i\,, \tag{20}$$
$$\widehat{u} = \frac{\sum_i x_i h_i(\hat{a}_i) - \delta w \hat{A}}{1-\delta}\,.$$

The optimal strategy is to pick a constant action – emission – every period and after all histories, $\hat{a}_i$ where its level is determined by

$$x_i h_i'(\hat{a}_i) = \delta w\,. \tag{21}$$

Proof We shall show by dynamic programming arguments that the Pareto-optimal value function is of the form $\widehat{V} = \sum_{i=1}^I x_i[\widehat{u}_i - w_i g]$. We need to be able to find the constants $\widehat{u}_i$ to satisfy:

$$\sum_{i=1}^I x_i[\widehat{u}_i - w_i g] = \max_{\mathbf{a}_1,\ldots a_I}\sum_{i=1}^I x_i \cdot\left[h_i(a_i) - c_i g + \delta(\widehat{u}_i - w_i(\boldsymbol{\sigma} g + \sum_{j=1}^I a_j))\right]\,. \tag{22}$$

Collecting terms that need maximization we can reduce the equation above to

$$\sum_{i=1}^I x_i\widehat{u}_i = \max_{\mathbf{a}_1,\ldots a_I}\sum_{i=1}^I x_i\left[h_i(a_i) - \delta w_i\sum_{j=1}^I a_j\right] + \delta\sum_{i=1}^I x_i\widehat{u}_i\,. \tag{23}$$

It is clear that the solution to this system is the same for all g; call this (first-best) solution $\widehat{a}_i$. Elementary algebra reveals that

$$\widehat{u}_i = \frac{h_i(\widehat{a}_i) - \delta w_i\sum_{j=1}^I \widehat{a}_j}{1-\delta} \quad\text{and}\quad w_i = \frac{c_i}{1-\delta\boldsymbol{\sigma}}\,.$$

It is also obvious that $x_i h_i'(\hat{a}_i) = \delta w$, where $w = \sum_{i=1}^I x_i w_i$. □

Theorem 4 states that, independently of the level of GHG, g, each country should emit an amount $\hat{a}_i$. The fact that the optimal emission is constant follows from the linearity of the model in g. Notice that on account of the linearity in the gas buildup equation – Eq. 17 – a unit of emission in period t can be analyzed in isolation as a surviving unit of size $\boldsymbol{\sigma}$ in period $t+1$, $\boldsymbol{\sigma}^2$ in period $t+2$, $\boldsymbol{\sigma}^3$ in period $t+2$, and so on. On account of the linearity in cost, these surviving units add $(\sum_i x_i c_i)\times\delta\boldsymbol{\sigma}$ in period $t+1$, $(\sum_i x_i c_i)\times(\delta\boldsymbol{\sigma})^2$ in period $t+2$, and so on, i. e., the marginal lifetime cost is

$$\frac{1}{1-\delta\boldsymbol{\sigma}}\sum_i x_i c_i\,,$$

or w, and that marginal cost is independent of g.

A Markov–Perfect Equilibrium: "Business as Usual"

This MPE shares the feature that the equilibrium emission rate of each country is constant in time, and it is the unique MPE with this property. We shall call it the "Business-as-Usual" equilibrium. Note that in this equilibrium each country takes account of the incremental damage to itself caused by an incremental increase in its emission rate, but does not take account of the damage caused to other countries.

Theorem 5 (Business-as-Usual Equilibrium) *Let g be the initial stock of GHG. For each country i, let a_i^* be determined by*

$$h_i'(a_i^*) = \delta w_i\,, \tag{24}$$
$$w_i = \frac{c_i}{1-\delta\boldsymbol{\sigma}}\,,$$

and let its strategy be to use a constant emission equal to a_i^* in each period; then this strategy profile is a MPE, and country i's corresponding payoff is

$$V_i^*(g) = u_i^* - w_i g\,, \tag{25}$$
$$u_i^* = \frac{h_i(a_i^*) - \delta w_i A^*}{1-\delta}\,.$$

The intuition for the existence of an MPE with constant emissions is similar to the analogous result for the GPO solution. (And indeed for that reason the proof will be omitted.) As long as other countries do not make their emissions contingent on the level of GHGs, country i has a constant marginal lifetime cost to emissions. And that marginal cost is independent of g.

Comparison of the GPO and Business as Usual

The preceding results enable us to compare the emissions in the GPO with those in the Business-as-Usual MPE:

$$\text{GPO}:\ h_i'(\hat{a}_i) = \frac{\delta\sum_j x_j c_j}{x_i(1-\delta\boldsymbol{\sigma})}\,, \tag{26}$$
$$\text{BAU}:\ h_i'(a_i^*) = \frac{\delta c_i}{1-\delta\boldsymbol{\sigma}}\,.$$

Since

$$x_i c_i < \sum_j x_j c_j \,,$$

it follows that

$$\frac{\delta c_i}{1-\delta\boldsymbol{\sigma}} < \frac{\delta \sum_j x_j c_j}{x_i(1-\delta\boldsymbol{\sigma})} \,.$$

Since h_i is concave, it follows that

$$a_i^* > \hat{a}_i \,. \tag{27}$$

Note that this inequality holds except in the trivial case in which all welfare weights are zero (except one). This result is known as the tragedy of the commons – whenever there is some externality to emissions, countries tend to over-emit in equilibrium. In turn, all this follows from the fact that in the BAU equilibrium each country only considers its own marginal cost and ignores the cost imposed on other countries on account of its emissions; in the GPO solution that additional cost is, of course, accounted for. It follows that the GPO is strictly Pareto superior to the MPE for an open set of welfare weights x_i (and leads to a strictly lower steady-state GHG level for all welfare weights).

One can contrast these results with those in Dockner, Long and Sorger (1998) that studies a model in which the benefits are linear in emission – i. e., h_i is linear – but convex in costs $c_i(.)$. The consequence of linearity in the benefit function h is that the GPO and BAU solutions have a "most rapid approach" (MRAP) property – if $(1-\boldsymbol{\sigma})g$, the depreciated stock in the next period, is less than a most preferred g^*, it is optimal to jump the system to g^*. Else it is optimal to wait for depreciation to bring the stock down to g^*. In other words, linearity in benefits implies a "one-shot" move to a desired level of gas g^*, which is thereafter maintained, while linearity in cost (as in the Dutta and Radner model) implies a constant emission rate. What is unclear in the Dockner, Long and Sorger model is why the multiple players would have the same target steady-state g^*. It would appear natural that, with asymmetric payoffs, each player would have a different steady-state. The existence of a MRAP equilibrium would appear problematical consequently. The authors impose a condition that implies that there is not too much asymmetry.

All SPE

We now turn to the second set of results – a full characterization of SPE in Dutta and Radner [12]. We will show that the SPE payoff correspondence has a surprising simplicity; the *set* of equilibrium payoffs at a level g is a simple linear translate of the set of equilibrium payoffs from some benchmark level, say, $g = 0$. Consequently, it will be seen that the *set* of emission levels that can arise in equilibrium from level g is identical to those that can arise from equilibrium play at a GHG level of 0. Note that the fact that the set of equilibrium possibilities is invariant to the level of g is perfectly consistent with the possibility that, in a particular equilibrium, emission levels vary with g. However, the invariance property will make for a particularly simple characterization of the best and worst equilibria.

Let $\Xi(g)$ denote the *set* of equilibrium payoff vectors with initial state g, i. e., each element of $\Xi(g)$ is the payoff to *some* SPE starting from g.

Theorem 6 *The equilibrium payoff correspondence Ξ is linear; there is a compact set $U \subset \Re^I$ such that for every initial state g*

$$\Xi(g) = U - \{w_1 g, w_2 g, \ldots w_I g\}$$

where $w_i = c_i/(1-\boldsymbol{\sigma}\delta)$, $i = 1, \ldots I$. In particular, consider any SPE, any period t and any history of play up until t. Then the payoff vector for the continuation strategies must necessarily be of the form

$$v - (w_1 g_t, w_2 g_t, \ldots w_I g_t) \,.$$

The theorem is proved by way of a bootstrap argument. We presume that a (candidate) payoff set has this invariance and show that the linear structure of the model confirms the conjecture. Consequently, we generate another candidate payoff set – which is also state-invariant. Then we look for a fixed point of that operator. In other words, we employ the APS operator to generate the SPE correspondence. Since that has already been discussed in the previous section, it is skipped here.

We will now use the above result to characterize the best – and the worst – equilibria in the global climate change game. Consider the *second-best problem* (from initial state g and for a given vector of welfare weights $\boldsymbol{x} = (x_i; i = 1, \ldots I)$), i. e., the problem of maximizing a weighted sum of *equilibrium payoffs*:

$$\max \sum_{i=1}^{I} x_i V_i(g) \,, \quad V(g) \in \Xi(g) \,.$$

Note that we consider all possible equilibria, i. e., we consider equilibria that choose to condition on current and past GHG levels as well as equilibria that do not. The result states that the best equilibrium *need not* condition on GHG levels:

Theorem 7 *There exists a constant emission level $\overline{a} \equiv \overline{a}_1, \overline{a}_2, \ldots \overline{a}_I$ – such that no matter what the initial level of GHG, the second-best policy is to emit at the constant rate $\overline{a}$. In the event of a deviation from this constant emissions policy by country i, play proceeds to i's worst equilibrium. Furthermore, the second-best emission rate is always strictly lower than the BAU rate, i. e., $\overline{a} < a^*$. Above a critical discount factor (less than 1), the second-best rate coincides with the GPO emission rate $\widehat{a}$.*

The theorem is attractive for three reasons: first, it says that the best possible equilibrium behavior is no more complicated than BAU behavior; so there is no argument for delaying a treaty (to cut emissions) merely because the status quo is simple. Second, the cut required to implement the second-best policy is an across the board cut – independently of anything else, country i should cut its emissions by the amount $a_i^* - \overline{a}_i$. Third, the second-best is exactly realized at high discount factors, rather than asymptotically approached as the discount factor tends to 1.

Sanctions will be required if countries break with the second-best policy and without loss of generality we can restrict attention to the worst such sanction. We turn now to a characterization of this worst equilibrium (for, say, country i). One definition will be useful for this purpose:

Definition 1 An i-less second-best equilibrium is the solution to a second-best problem in which the welfare weight of i is set equal to zero, i. e., $x_i = 0$.

By the previous theorem, every such problem has a solution in which on the equilibrium path, emissions are a constant. Denote that emission level $a(x_{-i})$:

Theorem 8 *There exists a "high" emission level $\overline{\overline{a}}(i)$ (with $\sum_{j\neq i} \overline{\overline{a}}_j(i) > \sum_{j\neq i} a_j^*$) and an i-less second-best equilibrium $a(x_{-i})$ such that country i's worst equilibrium is:*

1. *Each country emits at rate $\overline{\overline{a}}_j(i)$ for one period (no matter what g is), $j = 1, \ldots I$.*
2. *From the second period onwards, each country emits at the constant rate $a_j(x_{-i})$, $j = 1, \ldots I$.*

And if any country k deviates at either stages 1 or 2, play switches to k's worst equilibrium from the very next period after the deviation.

Put another way, for every country i, a sanction is made up of two emission rates, $\overline{\overline{a}}(i)$ and $a(x_{-i})$. The former imposes immediate costs on country i. The way it does so is by increasing the emission levels of countries $j \neq i$. The effect of this is a temporary increase in incremental GHG but due to the irreversibility of gas accumulation, a permanent increase in country i's costs, enough of an increase to wipe out any immediate gains that the country might have obtained from the deviation. Of course this additional emission also increases country j's costs. For the punishing countries, however, this increase is offset by the subsequent permanent change, the switch to the emission vector $\boldsymbol{a}(x_{-i})$, which permanently increases their quota at the expense of country i's.

Generalizations

The models discussed thus far are base-line models and do not deal with two important issues relating to climate change – technological change and capital accumulation. Technological change is important because that opens access to technologies that do not currently exist, technologies that may have considerably lower "emissions to energy" ratios, i. e., cleaner technologies. Capital accumulation is important because an important question is whether or not curbing GHGs is inimical to growth. The position articulated by both developing countries like Indian and China as well as by developed economies like the United States is that it is: placing curbs on emissions would restrict economic activities and hence restrain the competitiveness of the economy.

In Dutta and Radner [10,13] the following modification was made to the model studied in the previous section. It was presumed that the actual emission level associated with energy usage e_i is $f_i e_i$ where f_i is an index of (un)cleanliness – or *emission factor* – higher values implying larger emissions for the same level of energy usage. It was presumed that the emission factor could be changed at cost but driven no lower than some minimum μ_i. In other words,

$$0 \leq e_i(t) \,, \tag{28}$$

$$\mu_i \leq f_i(t+1) \leq f_i(t) \,. \tag{29}$$

Capital accumulation and population growth is also allowed in the model but taken to be exogenous. The dynamics of those two variables are governed by:

$$g(t) = \boldsymbol{\sigma}\, g(t-1) + \sum_{i=1}^{I} f_i(t) e_i(t) \,, \tag{30}$$

$$K_i(t+1) = H\left[K_i(t)\right] \,, \quad K_i(t) \nearrow \quad \text{and unbounded in } t \,, \tag{31}$$

$$P_i(t+1) = \psi_i P_i(t) + (1-\psi_i)\Psi \,, \quad P_i(t) \leq \Psi \,. \tag{32}$$

The output (gross-domestic product) of country i in period t is

$$h_i[K_i(t), P_i(t), e_i(t)] ,$$

where the function h_i has all of the standard properties mentioned above. The damage due to the stock of GHG, $g(t)$, is assumed to be (in units of GDP):

$$c_i P_i(t) g(t) .$$

The cost of reducing the emission factor from $f_i(t)$ to $f_i(t+1)$ is assumed to be:

$$\varphi_i[f_i(t) - f_i(t+1)] .$$

Immediately it is clear that the state variable now encompasses not just the common stock g but, additionally, the emission factor profile as well as the sizes of population and capital stock. In other words, $s = (g, f, K, P)$. Whilst this significant increase in dimensionality might suggest that it would be difficult to obtain clean characterizations, the papers show that there is some separability. The MPE "Business as Usual" has a separable structure – energy usage $e_i(t)$ and emission factor choice $f_i(t+1)$ – depend solely on country $i's$ capital stock and population alone. It varies by period – unlike in the base-line model discussed above – as the exogenous variables vary. Furthermore, the emission factor $f_i(t+1)$ stays unchanged till the population and capital stock cross a threshold level beyond which the cleanest technology μ_i gets picked. (This bang-bang character follows from the linearity of the model.)

The Global Pareto Optimal solution has similar features – the energy usage in country i is directly driven by the capital stock and population of that country. Furthermore the emission factor choice follows the same bang-bang character as for the MPE. However, there is a tragedy of the common in that in the MPE (versus the Pareto optimum) the energy usage is higher – at every state – and the switch to the cleanest technology happens later.

Future Directions

Within the general theory of dynamic games there are several open questions and possible directions for future research to take. On the existence question, there needs to be a better resolution of the case where the state space S is uncountably infinite. This is not just a technical curiosity. In applications, typically, in order to apply calculus techniques, we take the state variable to be a subset of some real space. The problem is difficult but one hopes that ancillary assumptions – such as concavity and monotonicity – will be helpful. These assumptions come "cheaply" because they are routinely invoked in economic applications.

The characterization result via APS techniques has a similar technical difficulty blocking its path, as the existence question. The folk theorem needs to be generalized as well to the S infinite case. Here it is our belief though that the difficulty is not conceptual but rather one where the appropriate result needs to be systematically worked out. As indicated above, the study of the dynamics of SPE paths is in its infancy and much remains to be done here.

Turning to the global climate change application, this is clearly a question of utmost social importance. The subject here is very much in the public consciousness yet academic study especially within economics is only a few years old. Many questions remain: generalizing the models to account for technological change and endogenous capital accumulation, examination of a carbon tax, of cap and trade systems for emission permits, of an international bank that can selectively foster technological change, … There are – as should be immediately clear – enough interesting important questions to exhaust many dissertations and research projects!

Bibliography

1. Abreu D (1988) On the theory of infinitely repeated games with discounting. Econometrica 56:383–396
2. Abreu D, Pearce D, Stachetti E (1990) Towards a general theory of discounted repeated games with discounting. Econometrica 58:1041–1065
3. Benhabib J, Radner R (1992) The joint exploitation of a productive asset: A game-theoretic approach. Econ Theory 2:155–190
4. Benoit J-P, Krishna V (1987) Finitely repeated games. Econometrica 53:905–922
5. Dockner E, Long N, Sorger G (1996) Analysis of Nash equilibria in a class of capital accumulation games. J Econ Dyn Control 20:1209–1235
6. Dockner E, Nishimura K (1999) Transboundary boundary problems in a dynamic game model. Jpn Econ Rev 50:443–456
7. Duffie D, Geanakoplos J, Mas-Colell A, Mclennan A (1994) Stationary Markov equilibria. Econometrica 62-4:745–781
8. Dutta P (1991) What do discounted optima converge to? A theory of discount rate asymptotics in economic models. J Econ Theory 55:64–94
9. Dutta P (1995) A folk theorem for stochastic games. JET 66:1–32
10. Dutta P, Radner R (2004) Self-enforcing climate change treaties. Proc Nat Acad Sci USA 101-14:5174–5179
11. Dutta P, Radner R (2006) Population growth and technological change in a global warming model. Econ Theory 29:251–270
12. Dutta P, Radner R (2008) A strategic model of global warming model: Theory and some numbers. J Econ Behav Organ (forthcoming)
13. Dutta P, Radner R (2008) Choosing cleaner technologies: Global warming and technological change, (in preparation)

14. Dutta P, Sundaram R (1993) How different can strategic models be? J Econ Theory 60:42–61
15. Fudenberg D, Maskin E (1986) The Folk theorem in repeated games with discounting or incomplete information. Econometrica 54:533–554
16. Harris C, Reny P, Robson A (1995) The existence of subgame perfect equilibrium in continuous games with almost perfect information: A case for extensive-form correlation. Econometrica 63:507–544
17. Inter-Governmental Panel on Climate Change (2007) Climate Change, the Synthesis Report. IPCC, Geneva
18. Levhari D, Mirman L (1980) The great fish war: An example using a dynamic cournot-Nash solution. Bell J Econ 11:322–334
19. Long N, Sorger G (2006) Insecure property rights and growth: The role of appropriation costs, wealth effects and heterogenity. Econ Theory 28:513–529
20. Mertens J-F, Parthasarathy T (1987) Equilibria for Discounted Stochastic Games. Research Paper 8750, CORE. University Catholique de Louvain
21. Mertens, Neyman (1983)
22. Nowak A (1985) Existence of equilibrium stationary strategies in discounted noncooperative stochastic games with uncountable state space. J Optim Theory Appl 45:591–603
23. Parthasarathy T (1973) Discounted, positive and non-cooperative stochastic games. Int J Game Theory 2–1:
24. Rieder U (1979) Equilibrium plans for non-zero sum Markov games. In: Moeschlin O, Pallasche D (ed) Game theory and related topics. North-Holland, Amsterdam
25. Rustichini A (1992) Second-best equilibria for games of joint exploitation of a productive asset. Econ Theory 2:191–196
26. Shapley L (1953) Stochastic Games. In: Proceedings of National Academy of Sciences, Jan 1953
27. Sobel M (1990) Myopic solutions of affine dynamic models. Oper Res 38:847–53
28. Sorger G (1998) Markov-perfect Nash equilibria in a class of resource games. Econ Theory 11:79–100
29. Stern N (2006) Review on the economics of climate change. HM Treasury, London. www.sternreview.org.uk
30. Stern Review on the Economics of Climate Change, October, (2006)
31. Sundaram R (1989) Perfect equilibrium in a class of symmetric dynamic games. J Econ Theory 47:153–177

Dynamics of Cellular Automata in Non-compact Spaces

Enrico Formenti[1], Petr Kůrka[2]
[1] Laboratoire I3S – UNSA/CNRS UMR 6070, Université de Nice Sophia Antipolis, Sophia Antipolis, France
[2] Center for Theoretical Study, Academy of Sciences and Charles University, Prague, Czech Republic

Article Outline

Glossary
Definition of the Subject
Introduction
Dynamical Systems
Cellular Automata
Submeasures
The Cantor Space
The Periodic Space
The Toeplitz Space
The Besicovitch Space
The Generic Space
The Space of Measures
The Weyl Space
Examples
Future Directions
Acknowledgments
Bibliography

Glossary

Almost equicontinuous CA A CA which has at least one equicontinuous configuration.

Attraction basin The set of configurations whose orbit is eventually attracted by an attractor.

Attractor A closed invariant set which attracts all orbits in some of its neighborhood.

Besicovitch pseudometrics A pseudometric that quantifies the upper-density of differences.

Blocking word A word that interrupts the information flow. A configuration containing an infinite number of blocking words both to the right and to the is equicontinuous in the Cantor topology.

Equicontinuous CA A CA in which all configurations are equicontinuous.

Equicontinuous configuration A configuration for which nearby configurations remain close.

Expansive CA Two distinct configurations, no matter how close, eventually separate during the evolution.

Generic space The space of configurations for which upper-density and lower-density coincide.

Sensitive CA In any neighborhood of any configuration there exists a configuration such that the orbits of the two configurations eventually separate.

Spreading set A clopen invariant set propagating both to the left and to the right.

Toeplitz space The space of regular quasi-periodic configurations.

Weyl pseudometrics A pseudometric that quantifies the upper density of differences with respect to all possible cell indices.

Definition of the Subject

In topological dynamics, the assumption of compactness is usually adopted as it has far reaching consequences.

Each compact dynamical system has an almost periodic point, contains a minimal subsystem, and each trajectory has a limit point. Nevertheless, there are important examples of non-compact dynamical systems like linear systems on $\mathbb{R}^n$ and the theory should cover these examples as well. The study of dynamics of cellular automata (CA) in the compact Cantor space of symbolic sequences starts with Hedlund [6] and is by now a firmly established discipline (see e. g., ► Topological Dynamics of Cellular Automata). The study of dynamics of CA in non-compact spaces like Besicovitch or Weyl spaces is more recent and provides an interesting alternative perspective.

The study of dynamics of cellular automata in non-compact spaces has at least two distinct origins. The first concerns the study of dynamical properties on peculiar countable dense sub-spaces of the Cantor space (the space of finite configuration or the space of spatially periodic configurations, for instance). The idea is that on those spaces, some properties are easier to prove than on the full Cantor space. Once a property is proved on such a subspace, one can try to lift it to the original Cantor space by using denseness. Another advantage is that the configurations on these spaces are easily representable on computers. Indeed, computer simulations and practical applications of CA usually take place in these subspaces.

The second origin is connected to the question of suitability of the classical Cantor topology for the study of chaotic behavior of CA and of symbolic systems in general. We briefly recall the motivations. Consider sensitivity to initial conditions for a CA in the Cantor topology. The shift map σ, which is a very simple CA, is sensitive to initial conditions since small perturbations far from the central region are eventually brought to the central part. However, from an algorithmic point of view, the shift map is very simple. We are inclined to regard a system as chaotic if its behavior cannot easily be reconstructed. This is not the case of the shift map whose chaoticity is more an artifact of the Cantor metric, rather than an intrinsic property of the system. Therefore, one may want to define another metric in which sensitive CA not only transport information (like the shift map) but also build/destroy new information at each time step.

This basic requirement stimulated the quest for alternative topologies to the classical Cantor space. This lead first to the Besicovitch topology and then to the Weyl topology in Cattaneo et al. [4] used to investigate almost periodic real functions (see [1,8]). Both these pseudometrics can be defined starting from suitable semi-measures on the set $\mathbb{Z}$ of integers. This way of construction had a *Pandora effect* opening the way to many new interesting topological spaces. Some of them are reported in this paper; others can be found in Cervelle and Formenti (► Algorithmic Complexity and Cellular Automata).

Each topology focuses on some peculiar aspects of the dynamics under study but all of them have a common denominator, namely non-compactness.

Introduction

A given CA over an alphabet A can be regarded as a dynamical system in several topological spaces: Cantor configuration space C_A, the space $\mathcal{M}_A$ of shift-invariant Borel probability measures on $A^{\mathbb{Z}}$, the Weyl space $\mathcal{W}_A$, the Besicovitch space $\mathcal{B}_A$, the generic space $\mathcal{G}_A$, the Toeplitz space $\mathcal{T}_A$ and the periodic space $\mathcal{P}_A$. We refer to various topological properties of these systems by prefixing the name of the space in question. Basic results correlate various dynamical properties of CA in these spaces.

The Cantor topology corresponds to the point of view of an observer who can distinguish only a finite central part of a configuration and sites outside this central part of the configuration are not taken into account. The Besicovitch and Weyl topologies, on the other hand, correspond to a god-like position of someone who sees whole configurations and can distinguish the frequency of differences. In the Besicovitch topology, the centers of configurations still play a distinguished role, as the frequencies of differences are computed from the center. In the Weyl topology, on the other hand, no site has a privileged position. Both Besicovitch and Weyl topologies are defined by pseudometrics. Different configurations can have zero distance and the topological space consists of equivalence classes of configurations which have zero distance.

The generic space $\mathcal{G}_A$ is a subspace of the Besicovich space of those configurations, in which each finite word has a well defined frequency. These frequencies define a Borel probability measure on the Cantor space of configurations, so we have a projection from the generic space $\mathcal{G}_A$ to the space $\mathcal{M}_A$ of Borel probability measures equipped with the weak* topology. This is a natural space for investigating the dynamics of CA on random configurations.

The Toeplitz space $\mathcal{T}_A$ consists of regular quasi-periodic configurations. This means that each pattern repeats periodically but different patterns have different periods. The Besicovitch and Weyl pseudometrics are actually metrics on the Toeplitz space and moreover they coincide on $\mathcal{T}_A$.

Dynamical Systems

A **dynamical system** is a continuous map $F\colon X \to X$ of a nonempty metric space X to itself. The nth it-

eration $F^n: X \to X$ of F is defined by $F^0(x) = x$, $F^{n+1}(x) = F(F^n(x))$. A point $x \in X$ is **fixed**, if $F(x) = x$. It is **periodic**, if $F^n(x) = x$ for some $n > 0$. The least positive n with this property is called the **period** of x. The **orbit** of x is the set $\mathcal{O}(x) = \{F^n(x): n > 0\}$. A set $Y \subseteq X$ is **positively invariant**, if $F(Y) \subseteq Y$ and **strongly invariant** if $F(Y) = Y$. A point $x \in X$ is **equicontinuous** ($x \in \mathcal{E}_F$) if the family of maps F^n is equicontinuous at X, i. e. $x \in \mathcal{E}_F$ iff

$$(\forall \varepsilon > 0)(\exists \delta > 0)(\forall y \in B_\delta(x))(\forall n > 0) \\ (d(F^n(y), F^n(x)) < \varepsilon) .$$

The system (X, F) is **almost equicontinuous** if $\mathcal{E}_F \neq \emptyset$ and **equicontinuous**, if

$$(\forall \varepsilon > 0)(\exists \delta > 0)(\forall x \in X)(\forall y \in B_\delta(x)) \\ (\forall n > 0)(d(F^n(y), F^n(x)) < \varepsilon) .$$

For an equicontinuous system $\mathcal{E}_F = X$. Conversely, if $\mathcal{E}_F = X$ and if X is compact, then F is equicontinuous; this needs not be true in the non-compact case. A system (X, F) is **sensitive** (to initial conditions), if

$$(\exists \varepsilon > 0)(\forall x \in X)(\forall \delta > 0)(\exists y \in B_\delta(x)) \\ (\exists n > 0)(d(f^n(y), f^n(x)) \geq \varepsilon) .$$

A sensitive system has no equicontinuous point. However, there exist systems with no equicontinuity points which are not sensitive. A system (X, F) is **positively expansive**, if

$$(\exists \varepsilon > 0)(\forall x \neq y \in X)(\exists n \geq 0)(d(f^n(x), f^n(y)) \geq \varepsilon) .$$

A positively expansive system on a perfect space is sensitive. A system (X, F) is (topologically) **transitive**, if for any nonempty open sets $U, V \subseteq X$ there exists $n \geq 0$ such that $F^{-n}(U) \cap V \neq \emptyset$. If X is perfect and if the system has a dense orbit, then it is transitive. Conversely, if (X, F) is topologically transitive and if X is compact, then (X, F) has a dense orbit. A system (X, F) is **mixing**, if for any nonempty open sets $U, V \subseteq X$ there exists $k > 0$ such that for every $n \geq k$ we have $F^{-n}(U) \cap V \neq \emptyset$. An ε-chain (from x_0 to x_n) is a sequence of points $x_0, \ldots, x_n \in X$ such that $d(F(x_i), x_{i+1}) < \varepsilon$ for $0 \leq i < n$. A system (X, F) is **chain-transitive**, if for any $\varepsilon > 0$ and any $x, y \in X$ there exists an ε-chain from x to y.

A strongly invariant closed set $Y \subseteq X$ is **stable**, if

$$\forall \varepsilon > 0, \exists \delta > 0, \forall x \in X, (d(x, Y) < \delta \\ \Longrightarrow \forall n > 0, d(F^n(x), Y) < \varepsilon) .$$

A strongly invariant closed stable set $Y \subseteq X$ is an **attractor**, if

$$\exists \delta > 0, \forall x \in X, (d(x, Y) < \delta \Longrightarrow \lim_{n\to\infty} d(F^n(x), Y) = 0.$$

A set $W \subseteq X$ is **inward**, if $F(\overline{W}) \subseteq W^\circ$. In compact spaces, attractors are exactly Ω-limits $\Omega_F(W) = \bigcap_{n>0} F(W)$ of inward sets.

Theorem 1 (Knudsen [10]) *Let (X, F) be a dynamical system and $Y \subseteq X$ a dense, F-invariant subset.*

(1) (X, F) is sensitive iff (Y, F) is sensitive.
(2) (X, F) is transitive iff (Y, F) is transitive.

Recall that a space X is **separable**, if it has a countable dense set.

Theorem 2 (Blanchard, Formenti, and Kůrka [2]) *Let (X, F) be a dynamical system on a non-separable space. If (X, F) is transitive, then it is sensitive.*

Cellular Automata

For a finite alphabet A, denote by $|A|$ the number of its elements, by $A^* := \bigcup_{n\geq 0} A^n$ the set of words over A, and by $A^+ := \bigcup_{n>0} A^n = A^* \setminus \{\lambda\}$ the set of nonempty words. The length of a word $u \in A^n$ is denoted by $|u| := n$. We say that $u \in A^*$ is a subword of $v \in A^*(u \sqsubseteq v)$ if there exists k such that $v_{k+i} = u_i$ for all $i < |u|$. We denote by $u_{[i,j)} = u_i \ldots u_{j-1}$ and $u_{[i,j]} = u_i \ldots u_j$ subwords of u associated to intervals. We denote by $A^{\mathbb{Z}}$ the set of A-***configurations***, or doubly-infinite sequences of letters of A. For any $u \in A^+$ we have a periodic configuration $u^\infty \in A^{\mathbb{Z}}$ defined by $(u^\infty)_{k|u|+i} = u_i$ for $k \in \mathbb{Z}$ and $0 \leq i < |u|$. The **cylinder** of a word $u \in A$ located at $l \in \mathbb{Z}$ is the set $[u]_l = \{x \in A^{\mathbb{Z}}: x_{[l,l+|u|)} = u\}$. The cylinder set of a set of words $U \subseteq A^+$ located at $l \in \mathbb{Z}$ is the set $[U]_l = \bigcup_{u\in U}[u]_l$.

A **subshift** is a nonempty subset $\Sigma \subseteq A^{\mathbb{Z}}$ such that there exists a set $D \subseteq A^+$ of **forbidden words** and $\Sigma = \Sigma_D := \{x \in A^{\mathbb{Z}}: \forall u \sqsubseteq x, u \notin D\}$. A subshift Σ_D is of **finite type** (SFT), if D is finite. A subshift is uniquely determined by its **language**

$$\mathcal{L}(\Sigma) := \bigcup_{n\geq 0} \mathcal{L}^n(\Sigma) , \\ \text{where} \quad \mathcal{L}^n(\Sigma) := \{u \in A^n: \exists x \in \Sigma, u \sqsubseteq x\} .$$

A **cellular automaton** is a map $F: A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ defined by $F(x)_i = f(x_{[i-r,i+r]})$, where $r \geq 0$ is a radius and $f: A^{2r+1} \to A$ is a local rule. In particular the **shift map**

$\sigma: A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ is defined by $\sigma(x)_i := x_{i+1}$. A local rule extends to the map $f: A^* \to A^*$ by $f(u)_i = f(u_{[i,i+2r]})$ so that $|f(u)| = \max\{|u| - 2r, 0\}$.

Definition 3 Let $F: A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ be a CA.

(1) A word $u \in A$ is m-blocking, if $|u| \geq m$ and there exists offset $d \leq |u| - m$ such that $\forall x, y \in [u]_0, \forall n > 0, F^n(x)_{[d,d+m)} = F^n(y)_{[d,d+m)}$.
(2) A set $U \subseteq A^+$ is spreading, if $[U]$ is F-invariant and there exists $n > 0$ such that $F^n([U]) \subseteq \sigma^{-1}([U]) \cap \sigma([U])$.

The following results will be useful in the sequel.

Proposition 4 (Formenti, Kůrka [5]) *Let $F: A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ be a CA and let $U \subseteq A^+$ be an invariant set. Then $\Omega_F([U])$ is a subshift iff U is spreading.*

Theorem 5 (Hedlund [6]) *Let $F: A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ be a CA with local rule $f: A^{2r+1} \to A$. Then F is surjective iff $f: A^* \to A^*$ is surjective iff $|f^{-1}(u)| = |A|^{2r}$ for each $u \in A^+$.*

Submeasures

A pseudometric on a set X is a map $d: X \times X \to [0, \infty)$ which satisfies the following conditions:

1. $d(x, y) = d(y, x)$,
2. $d(x, z) \leq d(x, y) + d(y, z)$.

If moreover $d(x, y) > 0$ for $x \neq y$, then we say that d is a metric. There is a standard method to create pseudometrics from submeasures. A bounded submeasure (with bound $M \in \mathbb{R}^+$) is a map $\varphi: \mathcal{P}(\mathbb{Z}) \to [0, M]$ which satisfies the following conditions:

1. $\varphi(\emptyset) = 0$,
2. $\varphi(U) \leq \varphi(U \cup V) \leq \varphi(U) + \varphi(V)$ for $U, V \subseteq \mathbb{Z}$.

A bounded submeasure φ on $\mathbb{Z}$ defines a pseudometric $d_\varphi: A^{\mathbb{Z}} \times A^{\mathbb{Z}} \to [0, \infty)$ by $d_\varphi(x, y) := \varphi(\{i \in \mathbb{Z}: x_i \neq y_i\})$. The Cantor, Besicovich and Weyl pseudometrics on $A^{\mathbb{Z}}$ are defined by the following submeasures:

$$\varphi_C(U) := 2^{-\min\{|i|:\, i \in U\}},$$

$$\varphi_{\mathcal{B}}(U) := \limsup_{l \to \infty} \frac{|U \cap [-l, l)|}{2l},$$

$$\varphi_{\mathcal{W}}(U) := \limsup_{l \to \infty} \sup_{k \in \mathbb{Z}} \frac{|U \cap [k, k+l)|}{l}.$$

The Cantor Space

The Cantor metric on $A^{\mathbb{Z}}$ is defined by

$$d_C(x, y) = 2^{-k} \quad \text{where} \quad k = \min\{|i|: x_i \neq y_i\},$$

so $d_C(x, y) < 2^{-k}$ iff $x_{[-k,k]} = y_{[-k,k]}$. We denote by $C_A = (A^{\mathbb{Z}}, d_C)$ the metric space of two-sided configurations with metric d_C. The cylinders are clopen sets in C_A. All Cantor spaces (with different alphabets) are homeomorphic. The Cantor space is compact, totally disconnected and perfect, and conversely, every space with these properties is homeomorphic to a Cantor space. Literature about CA dynamics in Cantor spaces is really huge. In this section, we just recall some results and definitions which will be used later.

Theorem 6 (Kůrka [11]) *Let (C_A, F) be a CA with radius r.*

(1) (C_A, F) is almost equicontinuous iff there exists a r-blocking word for F
(2) (C_A, F) is equicontinuous iff all sufficiently long words are r-blocking.

Denote by $\mathcal{E}_F$ the set of equicontinuous points of F. The sets of **equicontinuous directions** and **almost equicontinuous directions** of a CA (C_A, F) (see Sablik [15]) are defined by

$$\mathfrak{E}(F) = \left\{ \frac{p}{q}: p \in \mathbb{Z}, q \in \mathbb{N}^+, \mathcal{E}_{F^q \sigma^p} = A^{\mathbb{Z}} \right\},$$

$$\mathfrak{A}(F) = \left\{ \frac{p}{q}: p \in \mathbb{Z}, q \in \mathbb{N}^+, \mathcal{E}_{F^q \sigma^p} \neq \emptyset \right\}.$$

The Periodic Space

Definition 7 The **periodic space** $\mathcal{P}_A = \{x \in A^{\mathbb{Z}}: \exists n > 0, \sigma^n(x) = x\}$ over an alphabet A consists of shift periodic configurations with Cantor metric d_C.

All periodic spaces (with different alphabets) are homeomorphic. The periodic space is not compact, but it is totally disconnected and perfect. It is dense in C_A. If (C_A, F) is a CA, Then $F(\mathcal{P}_A) \subseteq \mathcal{P}_A$. We denote by $F_{\mathcal{P}}: \mathcal{P}_A \to \mathcal{P}_A$ the restriction of F to $\mathcal{P}_A$, so $(\mathcal{P}_A, F_{\mathcal{P}})$ is a (non-compact) dynamical system. Every $F_{\mathcal{P}}$-orbit is finite, so every point $x \in \mathcal{P}_A$ is $F_{\mathcal{P}}$-eventually periodic.

Theorem 8 *Let F be a CA over alphabet A.*

(1) (C_A, F) is surjective iff $(\mathcal{P}_A, F_{\mathcal{P}})$ is surjective.
(2) (C_A, F) is equicontinuous iff $(\mathcal{P}_A, F_{\mathcal{P}})$ is equicontinuous.

(3) (C_A, F) *is almost equicontinuous iff* $(\mathcal{P}_A, F_\mathcal{P})$ *is almost equicontinuous.*
(4) (C_A, F) *is sensitive iff* $(\mathcal{P}_A, F_\mathcal{P})$ *is sensitive.*
(5) (C_A, F) *is transitive iff* $(\mathcal{P}_A, F_\mathcal{P})$ *is transitive.*

Proof (1a) Let F be surjective, let $y \in \mathcal{P}_A$ and $\sigma^n(y) = y$. There exists $z \in F^{-1}(y)$ and integers $i < j$ such that $z_{[inr, inr+r)} = z_{[jnr, jnr+r)}$. Then $x = (z_{[inr, jnr)})^\infty \in \mathcal{P}_A$ and $F_\mathcal{P}(x) = y$, so $F_\mathcal{P}$ is surjective.
(1b) Let $F_\mathcal{P}$ be surjective, and $u \in A^+$. Then u^∞ has $F_\mathcal{P}$-preimage and therefore u has preimage under the local rule. By Hedlund Theorem, (C_A, F) is surjective.
(2a) Since $\mathcal{P}_A \subset C_A$, the equicontinuity of F implies trivially the equicontinuity of $F_\mathcal{P}$.
(2b) Let $F_\mathcal{P}$ be equicontinuous. There exist $m > r$ such that if $x, y \in \mathcal{P}_A$ and $x_{[-m,m]} = y_{[-m,m]}$, then $F^n(x)_{[-r,r]} = F^n(y)_{[-r,r]}$ for all $n \geq 0$. We claim that all words of length $2m+1$ are $(2r+1)$-blocking with offset $m - r$. If not, then for some $x, y \in A^\mathbb{Z}$ with $x_{[-m,m]} = y_{[-m,m]}$, there exists $n > 0$ such that $F^n(x)_{[-r,r]} \neq F^n(y)_{[-r,r]}$. For periodic configurations $x' = (x_{[-m-nr, m+nr]})^\infty$, $y' = (y_{[-m-nr, m+nr]})^\infty$ we get $F^n(x')_{[-r,r]} \neq F^n(y')_{[-r,r]}$ contradicting the assumption. By Theorem 6, F is C-equicontinuous.
(3a) If (C_A, F) is almost equicontinuous, then there exists a r-blocking word u and $u^\infty \in \mathcal{P}_A$ is an equicontinuous configuration for $(\mathcal{P}_a, F_\mathcal{P})$.
(3b) The proof is analogous as (2b).
(4) and (5) follow from the Theorem 1 of Knudsen. □

The Toeplitz Space

Definition 9 Let A be an alphabet

(1) The **Besicovitch pseudometric** on $A^\mathbb{Z}$ is defined by

$$d_\mathcal{B}(x, y) = \limsup_{l \to \infty} \frac{|\{j \in [-l, l) \colon x_j \neq y_j\}|}{2l} .$$

(2) The **Weyl pseudometric** on $A^\mathbb{Z}$ is defined by

$$d_\mathcal{W}(x, y) = \limsup_{l \to \infty} \max_{k \in \mathbb{Z}} \frac{|\{j \in [k, k+l) \colon x_j \neq y_j\}|}{l} .$$

Clearly $d_\mathcal{B}(x, y) \leq d_\mathcal{W}(x, y)$ and

$$\begin{aligned} d_\mathcal{B}(x, y) < \varepsilon &\iff \exists l_0 \in \mathbb{N}\,, \forall l \geq l_0\,, \\ &\qquad |\{j \in [-l, l] \colon x_j \neq y_j\}| < (2l+1)\varepsilon\,, \\ d_\mathcal{W}(x, y) < \varepsilon &\iff \exists l_0 \in \mathbb{N}\,, \forall l \geq l_0, \forall k \in \mathbb{Z}\,, \\ &\qquad |\{j \in [k, k+l) \colon x_j \neq y_j\}| < l\varepsilon\,. \end{aligned}$$

Both $d_\mathcal{B}$ and $d_\mathcal{W}$ are symmetric and satisfy the triangle inequality, but they are not metrics. Distinct configurations $x, y \in A^\mathbb{Z}$ can have zero distance. We construct a set of **regular quasi-periodic** configurations, on which $d_\mathcal{B}$ and $d_\mathcal{W}$ coincide and are metrics.

Definition 10

(1) The **period** of $k \in \mathbb{Z}$ in $x \in A^\mathbb{Z}$ is $r_k(x) := \inf\{p > 0 \colon \forall n \in \mathbb{Z}, x_{k+np} = x_k\}$. We set $r_k(x) = \infty$ if the defining set is empty.
(2) $x \in A^\mathbb{Z}$ is **quasi-periodic**, if $r_k(x) < \infty$ for all $k \in \mathbb{Z}$.
(3) A **periodic structure** for a quasi-periodic configuration x is a sequence of positive integers $\mathfrak{p} = (p_i)_{i < |\mathfrak{p}| \leq \infty}$, such that $p_i | p_{i+1}$ (p_i divides p_{i+1}), and for every $k \in \mathbb{Z}$, $r_k(x) | p_i$ for some i.
(4) A quasi-periodic configuration $x \in A^\mathbb{Z}$ is **regular**, if for some periodic structure $\mathfrak{p}$ of x we have $\lim_{i \to \infty} q_i(x)/p_i = 0$, where $q_i(x) := |\{k \in [0, p_i) \colon r_k(x) \nmid p_i\}|$ ($r_k(x)$ does not divide p_i).

Clearly every σ-periodic configuration is quasi-periodic and has a finite periodic structure.

Proposition 11

(1) If x, y *are regular quasi-periodic configurations, then* $d_\mathcal{W}(x, y) = d_\mathcal{B}(x, y)$.
(2) If $x \neq y$ *are quasi-periodic configurations, then* $d_\mathcal{W}(x, y) \geq d_\mathcal{B}(x, y) > 0$.

Proof (1) We must show $d_\mathcal{W}(x, y) \leq d_\mathcal{B}(x, y)$. Let $\mathfrak{p}^x, \mathfrak{p}^y$ be the periodic structures for x and y, and let $p_i = k_i^x p_i^x = k_i^y p_i^y$ be the lowest common multiple of p_i^x and p_i^y. Then $\mathfrak{p} = (p_i)_i$ is a periodic structure for both x and y. For each $i > 0$ and for each $k \in \mathbb{Z}$ we have

$$\begin{aligned} &|\{j \in [k - p_i, k + p_i) \colon x_j \neq y_j\}| \\ &\quad \leq 2k_i^x q_i^x + 2k_i^y q_i^y + |\{j \in [-p_i, p_i) \colon x_j \neq y_j\}|\,, \end{aligned}$$

$$\begin{aligned} d_\mathcal{W}(x, y) &\leq \lim_{i \to \infty} \max_{k \in \mathbb{Z}} |\{j \in [k - p_i, k + p_i) \colon x_j \neq y_j\}| \\ &\leq \lim_{i \to \infty} \left(\frac{2k_i^x q_i^x}{2k_i^x p_i^x} + \frac{2k_i^y q_i^y}{2k_i^y p_i^y} \right. \\ &\qquad \left. + \frac{|\{j \in [-p_i, p_i) \colon x_j \neq y_j\}|}{2p_i} \right) \\ &= d_\mathcal{B}(x, y)\,. \end{aligned}$$

(2) Since $x \neq y$, there exists i such that for some $k \in [0, p_i)$ and for all $n \in \mathbb{Z}$ we have $x_{k+np_i} = x_k \neq y_k = y_{k+np_i}$. It follows $d_\mathcal{B}(x, y) \geq 1/p_i$. □

Definition 12 The **Toeplitz space** $\mathcal{T}_A$ over A consists of all regular quasi-periodic configurations with metric $d_{\mathcal{B}} = d_{\mathcal{W}}$.

Toeplitz sequences are constructed by filling in periodic parts successively. For an alphabet A put $\widetilde{A} = A \cup \{*\}$.

Definition 13

(1) The **p**-*skeleton* $S_p(x) \in \widetilde{A}^{\mathbb{Z}}$ of $x \in A^{\mathbb{Z}}$ is defined by

$$S_p(x)_k = \begin{cases} x_k & \text{if} \quad \forall n \in \mathbb{Z}, x_{k+np} = x_k \\ * & \text{otherwise} . \end{cases}$$

(2) The **sequence of gaps** of $x \in \widetilde{A}^{\mathbb{Z}}$ is the unique increasing integer sequence $(t_i)_{a<i<b}$ such that $x_{t_i} = *, x_k \neq *$ for $t_i < k < t_{i+1}$ and $t_{-1} < 0 \leq t_0$.

(3) Let $x, y \in \widetilde{A}^{\mathbb{Z}}$ and let (t_i) be the sequence of gaps of x. The **amalgamation** $T(x, y) \in \widetilde{A}^{\mathbb{Z}}$ of x, y is

$$T(x, y)_i = \begin{cases} x_i & \text{if} \quad x_i \neq * \\ y_j & \text{if} \quad x_i = * \ \& \ i = t_j . \end{cases}$$

The p-skeleton is p-periodic. If p is its smallest period, we say that p is an **essential** period of x. The sequence of gaps may be two-way infinite (then $a = -\infty, b = \infty$), one-way infinite ($a = -\infty, b < \infty$ or $-\infty < a, b < \infty$), finite ($-\infty < a < b < \infty$) or even empty when $x \in A^{\mathbb{Z}}$. If it is nonempty then it must be defined at least on -1 or 0.

Proposition 14 *Let* $\mathbf{2} := \{0, 1\}$ *be the binary alphabet and* $[0, 1]$ *the real unit interval (with standard metric). There exists an isometric embedding* $f\colon [0, 1] \to \mathcal{T}_{\mathbf{2}}$ *such that* $f(0) = 0^{\infty}$ *and* $f(1) = 1^{\infty}$.

Proof Consider a map $h\colon \mathbf{2}^* \to \widetilde{\mathbf{2}}^{\mathbb{Z}}$ defined by $h(\lambda) = *^{\infty}, h(0) = (0*)^{\infty}, h(1) = (*1)^{\infty}, h(x_0 \ldots x_{n-1}x_n) = T(h(x_0 \ldots x_{n-1}), h(x_n))$. Thus

```
   h(0)  =  0*0*0*0*0*0*0*0*0*0*0 ...
   h(1)  =  *1*1*1*1*1*1*1*1*1*1*1 ...
  h(01)  =  0*010*010*010*010*010 ...
  h(10)  =  01*101*101*101*101*101 ...
 h(011)  =  0*0101010*0101010 ...
 h(100)  =  010101*1010101 ...
h(0111)  =  0*010101010101010 ...
h(1000)  =  01010101010101 ...
```

For $x \in \mathbf{2}^{\mathbb{N}}$, the limit (in the Cantor topology) $h(x) = \lim_{n\to\infty} h(x_0, \ldots x_n)$ exists. If no suffix of x is 1^{∞}, then $h(x) \in \mathbf{2}^{\mathbb{Z}}$, otherwise $h(x)$ contains exactly one star and we replace it by 1. Thus for each $x \in \mathbf{2}^{\mathbb{N}}$, $h(x) \in \mathbf{2}^{\mathbb{Z}}$ is a regular quasi-periodic sequence. It can be verified directly that $h(0111\ldots) = (01)^{\infty} = h(1000\ldots)$. Using an easily verifiable formula $T(x, T(y, z)) = T(T(x, y), z)$, we get

$$\begin{aligned} h(x_0 \ldots x_n 01^{\infty}) &= T(h(x_0 \ldots x_n), h(01^{\infty}) \\ &= T(h(x_0 \ldots x_n), h(10^{\infty})) \\ &= h(x_0 \ldots x_n 10^{\infty}) . \end{aligned}$$

For a real number $x \in [0, 1]$ with binary expansion $x = \sum_{i=0}^{\infty} x_i 2^{-i-1}$ put $f(x) = h(x_0 x_1 x_2 x_3 \ldots)$. If $2^n x$ is an integer for some n, then x has two binary expansions, and $f(x)$ is the same for both expansions. If $|x - y| \leq 2^{-m}$, then $x_{[0,m)} = y_{[0,m)}$; therefore $d_{\mathcal{W}}(f(x), f(y)) \leq 2^{-m}$ and $f\colon [0, 1] \to \mathcal{T}_A$ is continuous. For dyadic numbers x, y of the form $k/2^m$ we verify $d_{\mathcal{B}}(f(x), f(y)) = |x - y|$, so f is an isometry. □

Proposition 15 *The Toeplitz space* $\mathcal{T}_A$ *of regular quasi-periodic sequences is pathwise connected and infinite-dimensional.*

Proof Assume that the alphabet A contains letters 0,1 and consider the map $f\colon [0, 1] \to A^{\mathbb{Z}}$ from Proposition 14. For $a \in A$ set $a \cdot 0 = 0$, $a \cdot 1 = a$. Given $u \in \mathbf{2}^{\mathbb{Z}}$ the map $g_u\colon [0, 1] \to \mathbf{2}^{\mathbb{Z}}$ defined by $g_u(x)_i = u_i f(x)_i$ is continuous, $g_u(0) = 0$ and $g_u(1) = u$. Thus $\mathcal{T}_A$ is pathwise connected. To show that $\mathcal{T}_A$ is infinite-dimensional, construct for any n an embedding $f_n\colon [0, 1]^n \to X_{\mathcal{W}}$ of an n-dimensional cube by

$$f_n(x_1, \ldots, x_n) = f(x_1)_0 \ldots f(x_n)_0 f(x_1)_1 \ldots f(x_n)_1 \ldots .$$

Thus $\mathcal{T}_A$ is at least n-dimensional and therefore infinite-dimensional. □

Proposition 16 *Let* $F\colon A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ *be a CA with radius r.*

(1) If $x \in A^{\mathbb{Z}}$ *is a quasi-periodic with periodic structure* $\mathfrak{p}$, *then* $F(x)$ *is quasi-periodic with periodic structure* $\mathfrak{p}$.
(2) If x *is regular quasi-periodic, then* $F(x)$ *is regular.*

Proof (1) For $k \in \mathbb{Z}$ denote by $m := \min\{i\colon \forall j \in [k - r, k + r], r_j(x) | p_i\}$. Then p_m is a period of $F(x)_k$.
(2) We have

$$\begin{aligned} q_i(F(x)) &:= |\{k < p_i\colon r_k(F(x)) \not| \ p_i\}| \\ &\leq (2r + 1) \cdot |\{k < p_i\colon r_k(x) \not| \ p_i\}| \\ &= (2r + 1) \cdot q_i(x) , \end{aligned}$$

so

$$\lim_{i\to\infty} \frac{q_i(F(x))}{p_i} \leq (2r + 1) \cdot \lim_{i\to\infty} \frac{q_i(x)}{p_i} = 0 .$$

□

For a CA F we denote by $F_{\mathcal{T}}$ the restriction of F to $\mathcal{T}_A$.

Theorem 17 *Let F be a CA.*

(1) (C_A, F) *is surjective iff* $(\mathcal{T}_A, F_\mathcal{T})$ *is surjective.*
(2) If $\mathfrak{A}(F) \neq \emptyset$*, then* $(\mathcal{T}_A, F_\mathcal{T})$ *is almost equicontinuous.*
(3) if $\mathfrak{E}(F) \neq \emptyset$*, then* $(\mathcal{T}_A, F_\mathcal{T})$ *is equicontinuous.*
(4) If (C_A, F) *is chain-transitive, then* $(\mathcal{T}_A, F_\mathcal{T})$ *is chain-transitive.*
(5) $(\mathcal{T}_A, F_\mathcal{T})$ *is injective iff it is surjective.*

Proof (1) The proof is the same as in Theorem 8(1).
(2) Assume first that F is almost equicontinuous, so there exists $m > r$ and $u \in A^{2m+1}$ such that for any $x, y \in [u]_{-m}$, $F^n(x)_{[-r,r]} = F^n(y)_{[-r,r]}$ for all $n > 0$. We show that u^∞ is $\mathcal{T}$-equicontinuous. For a given $\varepsilon > 0$ set $\delta = \varepsilon/(4m - 2r + 1)$. If $d_\mathcal{T}(y, x) < \delta$, then there exists l_0 such that for all $l \geq l_0$, $|\{i \in [-l, l]\colon x_i \neq y_i\}| < (2l + 1)\delta$. For $k(2m + 1) \leq j < (k + 1)(2m + 1)$, $F^n(y)_j$ can differ from $F^n(x)_j$ only if y differs from x in some $i \in [k(2m+1)-(m-r), (k+1)m+(m-r))$ Thus a change $x_i \neq y_i$ can cause at most $2m+1+2(m-r) = 4m-2r+1$ changes $F^n(y)_j \neq F^n(x)_j$. We get

$$|\{i \in [-l, l)\colon F^n(x)_i \neq F^n(y)_i\}| \leq 2l\delta(4m - 2r + 1) \leq 2l\varepsilon\,.$$

This shows that $F_\mathcal{T}$ is almost equicontinuous. In the general case that $\mathfrak{A}(F) \neq \emptyset$, we get that $F_\mathcal{T}^q \sigma^p$ is almost equicontinuous for some $p \in \mathbb{Z}$, $q \in \mathbb{N}^+$. Since σ is $\mathcal{T}$-equicontinuous, $F_\mathcal{T}^q$ is almost equicontinuous and therefore $(\mathcal{T}_A, F_\mathcal{T})$ is almost equicontinuous.
(3) The proof is the same as in (2) with the only modification that all $u \in A^m$ are $(2r + 1)$-blocking.
(4) The proof of Proposition 8 from [2] works in this case too.
(5) The proof of Proposition 12 of [3] works in this case also. □

The Besicovitch Space

On $A^\mathbb{Z}$ we have an equivalence $x \approx_\mathcal{B} y$ iff $d_\mathcal{B}(x, y) = 0$. Denote by $\mathcal{B}_A$ the set of equivalence classes of $\approx_\mathcal{B}$ and by $\pi_\mathcal{B} : A^\mathbb{Z} \to \mathcal{B}_A$ the projection. The factor of $d_\mathcal{B}$ is a metric on $\mathcal{B}_A$. This is the **Besicovitch space** on alphabet A. Using prefix codes, it can be shown that every two Besicovitch spaces (with different alphabets) are homeomorphic. By Proposition 11 each equivalence class contains at most one quasi-periodic sequence.

Proposition 18 $\mathcal{T}_A$ *is dense in* $\mathcal{B}_A$.

The proof of Proposition 9 of [3] works also for regular quasi-periodic sequences.

Theorem 19 (Blanchard, Formenti and Kůrka [2]) *The Besicovitch space is pathwise connected, infinite-dimensional, homogenous and complete. It is neither separable nor locally compact.*

The properties of path-connectedness and infinite dimensionality is proved analogously as in Proposition 15. To prove that $\mathcal{B}_A$ is neither separable nor locally compact, Sturmian configurations have been used in [2]. The completeness of $\mathcal{B}_A$ has been proved by Marcinkiewicz [14]. Every cellular automaton $F\colon A^\mathbb{Z} \to A^\mathbb{Z}$ is uniformly continuous with respect to $d_\mathcal{B}$, so it preserves the equivalence $\approx_\mathcal{B}$. If $d_\mathcal{B}(x, y) = 0$, then $d_\mathcal{B}(F(x), F(y)) = 0$. Thus a cellular automaton F defines a uniformly continuous map $F_\mathcal{B} : \mathcal{B}_A \to \mathcal{B}_A$.

Theorem 20 (Blanchard, Formenti and Kůrka [2]) *Let F be a CA on A.*

(1) (C_A, F) *is surjective iff* $(\mathcal{B}_A, F_\mathcal{B})$ *is surjective.*
(2) If $\mathfrak{A}(F) \neq \emptyset$ *then* $(\mathcal{B}_A, F_\mathcal{B})$ *is almost equicontinuous.*
(3) if $\mathfrak{E}(F) \neq \emptyset$*, then* $(\mathcal{B}_A, F_\mathcal{B})$ *is equicontinuous.*
(4) If $(\mathcal{B}_A, F_\mathcal{B})$ *is sensitive, then* (C_A, F) *is sensitive.*
(5) No cellular automaton $(\mathcal{B}_A, F_\mathcal{B})$ *is positively expansive.*
(6) If (C_A, F) *is chain-transitive, then* $(\mathcal{B}_A, F_\mathcal{B})$ *is chain-transitive.*

Theorem 21 (Blanchard, Cervelle and Formenti [3])

(1) No CA $(\mathcal{B}_A, F_\mathcal{B})$ *is transitive.*
(2) A CA $(\mathcal{B}_A, F_\mathcal{B})$ *has either a unique fixed point and no other periodic point, or it has uncountably many periodic points.*
(3) If a surjective CA has a blocking word, then the set of its $F_\mathcal{B}$*-periodic points is dense in* $\mathcal{B}_A$.

The Generic Space

For a configuration $x \in A^\mathbb{Z}$ and word $v \in A^+$ set

$$\underline{\Phi}_v(x) = \liminf_{n\to\infty} |\{i \in [-n, n)\colon x_{[i,i+|v|)} = v\}|/2n\,,$$
$$\overline{\Phi}_v(x) = \limsup_{n\to\infty} |\{i \in [-n, n)\colon x_{[i,i+|v|)} = v\}|/2n\,.$$

For every $v \in A^+$, $\underline{\Phi}_v, \overline{\Phi}_v\colon A^\mathbb{Z} \to [0, 1]$ are continuous in the Besicovitch topology. In fact we have

$$|\overline{\Phi}_v(x) - \overline{\Phi}_v(y)| \leq d_\mathcal{B}(x, y) \cdot |v|\,,$$
$$|\underline{\Phi}_v(x) - \underline{\Phi}_v(y)| \leq d_\mathcal{B}(x, y) \cdot |v|\,.$$

Define the generic space (over the alphabet A) as

$$\mathcal{G}_A = \{x \in A^\mathbb{Z}\colon \forall v \in A^*, \underline{\Phi}_v(x) = \overline{\Phi}_v(x)\}\,.$$

It is a closed subspace of $\mathcal{B}_A$. For $v \in A^*$ denote by $\Phi_v\colon \mathcal{G}_A \to [0, 1]$ the common value of $\underline{\Phi}$ and $\overline{\Phi}$.

Using prefix codes, one can show that all generic spaces (with different alphabets) are homeomorphic. The generic space contains all uniquely ergodic subshifts, in particular all Sturmian sequences and all regular Toeplitz sequences. Thus the proofs in Blanchard Formenti and Kůrka [2] can be applied to the generic space too. In particular the generic space is homogenous. If we regard the alphabet $A = \{0, \ldots, m-1\}$ as the group $\mathbb{Z}_m = \mathbb{Z}/m\mathbb{Z}$, then for every $x \in \mathcal{G}_A$ there is an isometry $H_x : \mathcal{G}_A \to \mathcal{G}_A$ defined by $H_x(y) = x + y$. Moreover, $\mathcal{G}_A$ is pathwise connected, infinite-dimensional and complete (as a closed subspace the full Besicovitch space). It is neither separable nor locally compact. If $F : A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ is a cellular automaton, then $F(\mathcal{G}_A) \subseteq \mathcal{G}_A$. Thus, the restriction of $F_{\mathcal{B}}$ to $\mathcal{G}_A$ defines a dynamical system $(\mathcal{G}_A, F_{\mathcal{G}})$. See also Pivato for a similar approach.

Theorem 22 *Let $F : A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ be a CA.*

(1) (C_A, F) is surjective iff $(\mathcal{G}_A, F_{\mathcal{G}})$ is surjective.
(2) If $\mathfrak{A}(F) \neq \emptyset$, then $(\mathcal{G}_A, F_{\mathcal{G}})$ is almost equicontinuous.
(3) if $\mathcal{E}(F) \neq \emptyset$, then $(\mathcal{G}_A, F_{\mathcal{G}})$ is equicontinuous.
(4) If $(\mathcal{G}_A, F_{\mathcal{G}})$ is sensitive, then (C_A, F) is sensitive.
(5) If F is C-chain transitive, then F is $\mathcal{G}$-chain transitive.

The proofs are the same as the proofs of corresponding properties in [2].

The Space of Measures

By a **measure** we mean a **Borel shift-invariant probability measure** on the Cantor space $A^{\mathbb{Z}}$ (see ▶ Ergodic Theory of Cellular Automata). This is a countably additive function μ on the Borel sets of $A^{\mathbb{Z}}$ which assigns 1 to the full space and satisfies $\mu(U) = \mu(\sigma^{-1}(U))$. A measure on $A^{\mathbb{Z}}$ is determined by its values on cylinders $\mu(u) := \mu([u]_n)$ which does not depend on $n \in \mathbb{Z}$. Thus a measure can be identified with a map $\mu : A^* \to [0, 1]$ subject to **bilateral Kolmogorov compatibility conditions**

$$\sum_{a \in A} \mu(ua) = \sum_{a \in A} \mu(au) = \mu(u) , \quad \mu(\lambda) = 1 .$$

Define the distance of two measures

$$d_{\mathcal{M}}(\mu, \nu) := \sum_{u \in A^+} |\mu(u) - \nu(u)| \cdot |A|^{-2|u|} .$$

This is a metric which yields the topology of weak* convergence on the compact space $\mathcal{M}_A := \mathfrak{M}_\sigma(A^{\mathbb{Z}})$ of shift-invariant Borel probability measures. A CA $F : A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ with local rule f determines a continuous and affine map $F_{\mathcal{M}} : \mathcal{M}_A \to \mathcal{M}_A$ by

$$(F_{\mathcal{M}}(\mu))(u) = \sum_{v \in f^{-1}(u)} \mu(v) .$$

Moreover F and $F\sigma$ determine the same dynamical system on $\mathcal{M}_A$: $F_{\mathcal{M}} = (F\sigma)_{\mathcal{M}}$.

For $x \in \mathcal{G}_A$ denote by $\Phi^x : A^* \to [0, 1]$ the function $\Phi^x(v) = \Phi_v(x)$. For every $x \in \mathcal{G}_A$, Φ^x is a shift-invariant Borel probability measure. The map $\Phi : \mathcal{G}_A \to \mathcal{M}_A$ is continuous with respect to the Besicovich and weak* topologies. In fact we have

$$\begin{aligned} d_{\mathcal{M}}(\Phi^x, \Phi^y) &\leq d_{\mathcal{B}}(x, y) \sum_{u \in A^+} |u| \cdot |A|^{-2|u|} \\ &= d_{\mathcal{B}}(x, y) \sum_{n > 0} n \cdot |A|^{-n} \\ &= d_{\mathcal{B}}(x, y) \cdot |A|/(|A| - 1)^2 . \end{aligned}$$

By a theorem of Kamae [9], Φ is surjective. Every shift-invariant Borel probability measure has a generic point. It follows from the ergodic theorem that if μ is a σ-invariant measure, then $\mu(\mathcal{G}_A) = 1$ and for every $v \in A^*$, the measure of v is the integral of its density Φ_v,

$$\mu(v) = \int \Phi_v(x) \, \mathrm{d}\mu .$$

If F is a CA, we have a commutative diagram $\Phi F_{\mathcal{G}} = F_{\mathcal{M}} \Phi$.

$$\begin{array}{ccc} \mathcal{G}_A & \xrightarrow{F_{\mathcal{G}}} & \mathcal{G}_A \\ \Phi \downarrow & & \downarrow \Phi \\ \mathcal{M}_A & \xrightarrow{F_{\mathcal{M}}} & \mathcal{M}_A \end{array}$$

Theorem 23 *Let F be a CA over A.*

(1) (C_A, F) is surjective iff $(\mathcal{M}_A, F_{\mathcal{M}})$ is surjective.
(2) If $(\mathcal{G}_A, F_{\mathcal{G}})$ has dense set of periodic points, then $(\mathcal{M}_A, F_{\mathcal{M}})$ has dense set of periodic points.
(3) If $\mathcal{A}(F) \neq \emptyset$, then $(\mathcal{M}_A, F_{\mathcal{M}})$ is almost equicontinuous.
(4) If $\mathfrak{E}(F) \neq \emptyset$, then $(\mathcal{M}_A, F_{\mathcal{M}})$ is equicontinuous.

Proof (1) See Kůrka [13] for a proof.
(2) This holds since $(\mathcal{M}_A, F_{\mathcal{M}})$ is a factor of $(\mathcal{G}_A, F_{\mathcal{G}})$.
(3) It suffices to prove the claim for the case that F is almost equicontinuous. In this case there exists a blocking word

$u \in A^+$ and the Dirac measure δ_u defined by

$$\delta_u(v) = \begin{cases} 1/|u| & \text{if} \quad v \sqsubseteq u \\ 0 & \text{if} \quad v \not\sqsubseteq u \end{cases}$$

is equicontinuous for $(\mathcal{M}_A, F_{\mathcal{M}})$.
(4) If (C_A, F) is equicontinuous, then all sufficiently long words are blocking and there exists $d > 0$ such that for all $n > 0$, and for all $x, y \in A^{\mathbb{Z}}$ such that $x_{[-n-d,n+d]} = y_{[-n-d,n+d]}$ we have $F^k(x)_{[-n,n]} = F^k(y)_{[-n,n]}$ for all $k > 0$. Thus there are maps $g_k : A^* \to A^*$ such that $|g_k(u)| = \max\{|u| - 2d, 0\}$ and for every $x \in A^{\mathbb{Z}}$ we have $F^k(x)_{[-n,n]} = F_k(x_{[-n-kd,n+kd]}) = g_k(x_{[-n-d,n+d]})$, where f is the local rule for F. We get

$$\begin{aligned} d_{\mathcal{M}}\big(F^k_{\mathcal{M}}(\mu), F^k_{\mathcal{M}}(\nu)\big) &= \sum_{n=1}^{\infty} \sum_{u \in A^n} \left| \sum_{v \in f^{-k}(u)} (\mu(v) - \nu(v)) \right| \cdot |A|^{-2n} \\ &= \sum_{n=1}^{\infty} \sum_{u \in A^n} \left| \sum_{v \in g_k^{-1}(u)} (\mu(v) - \nu(v)) \right| \cdot |A|^{-2n} \\ &\le \sum_{n=1}^{\infty} \sum_{v \in A^{n+2d}} |\mu(v) - \nu(v)| \cdot |A|^{-2n} \\ &\le |A|^{4d} \cdot d_{\mathcal{M}}(\mu, \nu) \,. \end{aligned}$$

□

The Weyl Space

Define the following equivalence relation on $A^{\mathbb{Z}}$: $x \approx_{\mathcal{W}} y$ iff $d_{\mathcal{W}}(x, y) = 0$. Denote by $\mathcal{W}_A$ the set of equivalence classes of $\approx_{\mathcal{W}}$ and by $\pi_{\mathcal{W}} : A^{\mathbb{Z}} \to \mathcal{W}_A$ the projection. The factor of $d_{\mathcal{W}}$ is a metric on $\mathcal{W}_A$. This is the Weyl space on alphabet A. Using prefix codes, it can be shown that every two Weyl spaces (with different alphabets) are homeomorphic. The Toeplitz space is not dense in the Weyl space (see Blanchard, Cervelle and Formenti [3]).

Theorem 24 (Blanchard, Formenti and Kůrka [2]) *The Weyl space is pathwise connected, infinite-dimensional and homogenous. It is neither separable nor locally compact. It is not complete.*

Every cellular automaton $F : A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ is continuous with respect to $d_{\mathcal{W}}$, so it preserves the equivalence $\approx_{\mathcal{W}}$. If $d_{\mathcal{W}}(x, y) = 0$, then $d_{\mathcal{W}}(F(x), F(y)) = 0$. Thus a cellular automaton F defines a continuous map $F_{\mathcal{W}} : \mathcal{W}_A \to \mathcal{W}_A$. The shift map $\sigma : \mathcal{W}_A \to \mathcal{W}_A$ is again an isometry, so in $\mathcal{W}_A$ many topological properties are preserved if F is composed with a power of the shift. This is true for example for equicontinuity, almost continuity and sensitivity. If $\pi : \mathcal{W}_A \to \mathcal{B}_A$ is the (continuous) projection and F a CA, then the following diagram commutes.

$$\begin{array}{ccc} \mathcal{W}_A & \xrightarrow{F_{\mathcal{W}}} & \mathcal{W}_A \\ \pi \downarrow & & \downarrow \pi \\ \mathcal{B}_A & \xrightarrow{F_{\mathcal{B}}} & \mathcal{B}_A \end{array}$$

Theorem 25 (Blanchard, Formenti and Kůrka [2]) *Let F be a CA on A.*

(1) (C_A, F) is surjective iff $(\mathcal{W}_A, F_{\mathcal{W}})$ is surjective.
(2) If $\mathfrak{A}(F) \neq \emptyset$, then $(\mathcal{W}_A, F_{\mathcal{W}})$ is almost equicontinuous.
(3) if $\mathcal{E}(F) \neq \emptyset$, then $(\mathcal{W}_A, F_{\mathcal{W}})$ is equicontinuous.
(4) If (C_A, F) is chain-transitive, then $(\mathcal{W}_A, F_{\mathcal{W}})$ is chain-transitive.

Theorem 26 (Blanchard, Cervelle and Formenti [3]) *No CA is $(\mathcal{W}_A, F_{\mathcal{W}})$ is transitive.*

Theorem 27 *Let Σ be a subshift attractor of finite type for F (in the Cantor space). Then there exists $\delta > 0$ such that for every $x \in \mathcal{W}_A$ satisfying $d_{\mathcal{W}}(x, \Sigma) < \delta$, $F^n(x) \in \Sigma$ for some $n > 0$.*

Thus a subshift attractor of finite type is a $\mathcal{W}$-attractor. Example 2 shows that it need not be $\mathcal{B}$-attractor. Example 3 shows that the assertion need not hold if Σ is not of finite type.

Proof Let $U \subseteq A^{\mathbb{Z}}$ be a C-clopen set such that $\Sigma = \Omega_F(U)$. Let U be a union of cylinders of words of length q. Set $\widetilde{\Omega}_\sigma(U) = \bigcap_{n \in \mathbb{Z}} \sigma^n(U)$. By a generalization of a theorem of Hurd [7] (see ▶ Topological Dynamics of Cellular Automata), there exists $m > 0$ such that $\Sigma = F^m(\widetilde{\Omega}_\sigma)$. If $d_{\mathcal{W}}(x, \Sigma) < 1/q$ then there exists $l > 0$ such that for every $k \in \mathbb{Z}$ there exists a nonnegative $j < l$ such that $\sigma^{k+j}(x) \in U$. It follows that there exists $n > 0$ such that $F^n(x) \in \widetilde{\Omega}_\sigma(U)$ and therefore $F^{n+m}(x) \in \Sigma$. □

Examples

Example 1 The identity rule $\mathrm{Id}(x) = x$.

$(\mathcal{B}_A, \mathrm{Id}_{\mathcal{B}})$ and $(\mathcal{W}_A, \mathrm{Id}_{\mathcal{W}})$ are chain-transitive (since both $\mathcal{B}_A$ and $\mathcal{W}_A$ are connected). However, (C_A, Id) is not chain-transitive. Thus the converse of Theorem 20(6) and of Theorem 25(4) does not hold.

Example 2 The product rule ECA128 $F(x)_i = x_{i-1} \cdot x_i \cdot x_{i+1}$.

(C_A, F), $(\mathcal{B}_A, F_{\mathcal{B}})$ and $(\mathcal{W}_A, F_{\mathcal{W}})$ are almost equicontinous and the configuration 0^∞ is equicontinuous in all these versions. By Theorem 27, $\{0^\infty\}$ is a $\mathcal{W}$-attractor. However, contrary to a mistaken Proposition 9 in [2], $\{0^\infty\}$ is

Dynamics of Cellular Automata in Non-compact Spaces, Figure 1
The product ECA184

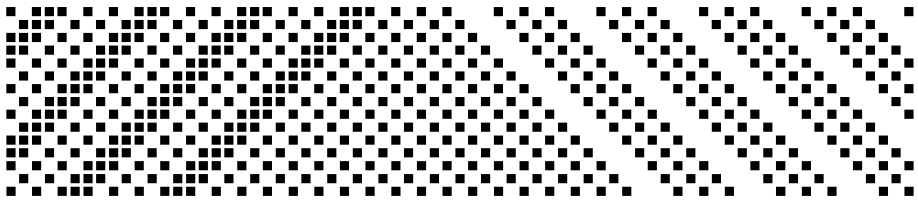

Dynamics of Cellular Automata in Non-compact Spaces, Figure 2
The traffic ECA184

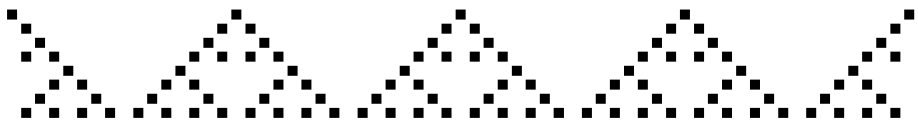

Dynamics of Cellular Automata in Non-compact Spaces, Figure 3
The sum ECA90

not $\mathcal{B}$-attractor. For a given $0 < \varepsilon < 1$ define $x \in A^{\mathbb{Z}}$ by $x_i = 1$ iff $3^n(1-\varepsilon) < |i| \leq 3^n$ for some $n \geq 0$. Then $d_{\mathcal{B}}(x, 0^\infty) = \varepsilon$ but x is a fixed point, since $d_{\mathcal{B}}(F(x), x) = \lim_{n\to\infty} 2n/3^n = 0$ (see Fig. 1).

Example 3 The traffic ECA184 $F(x)_i = 1$ iff $x_{[i-1,i]} = 10$ or $x_{[i,i+1]} = 11$.

No $F^q\sigma^p$ is C-almost equicontinuous, so $\mathfrak{A}(F) = \emptyset$. However, if $d_{\mathcal{W}}(x, 0^\infty) < \delta$, then $d_{\mathcal{W}}(F^n(x), 0^\infty) < \delta$ for every $n > 0$, since F conserves the number of letters 1 in a configuration. Thus 0^∞ is a point of equicontinuity in $(\mathcal{T}_A, F_{\mathcal{T}})$, $(\mathcal{B}_A, F_{\mathcal{B}})$, and $(\mathcal{W}_A, F_{\mathcal{W}})$. This shows that item (2) of Theorems 17, 20 and 25 cannot be converted. The maximal C-attractor $\Omega_F = \{x \in A^{\mathbb{Z}} \colon \forall n > 0, 1(10)^n 0 \not\sqsubseteq x\}$ is not SFT. We show that it does not $\mathcal{W}$-attracts points from any of its neighborhood. For a given even integer $q > 2$ define $x \in A^{\mathbb{Z}}$ by

$$x_i = \begin{cases} 0 & \text{if} \quad \exists n \geq 0\,,\, i = qn+1 \\ 1 & \text{if} \quad \exists n < 0\,,\, i = qn \\ ((01)^\infty)_i & \text{otherwise}\,. \end{cases}$$

Then $d_W(F^k(x), \Omega_F) = 1/q$ for all $k > 0$ (see Fig. 2, where $q = 8$).

Example 4 The sum ECA90 $F(x)_i = (x_{i-1} + x_{i+1}) \bmod 2$.

Both $(\mathcal{B}_A, F_{\mathcal{B}})$ and $(\mathcal{W}_A, F_{\mathcal{W}})$ are sensitive (Cattaneo et al. [4]). For a given $n > 0$ define a configuration z by $z_i = 1$ iff $i = k2^n$ for some $k \in \mathbb{Z}$. Then $F^{2^{n-1}}(z) = (01)^\infty$. For any $x \in A^{\mathbb{Z}}$, we have $d_{\mathcal{W}}(x, x+z) = 2^{-n}$ but $d_{\mathcal{W}}(F^{2^{n-1}}(x), F^{2^{n-1}}(x+z)) = 1/2$. The same argument works for $(\mathcal{B}_A, F_{\mathcal{B}})$.

Example 5 The shift ECA170 $F(x)_i = x_{i+1}$.

Since the system has fixed points 0^∞ and 1^∞, it has uncountable number of periodic points. However, the periodic points are not dense in $\mathcal{B}_A$ ([3]).

Future Directions

One of the promising research directions is the connection between the generic space and the space of Borel probability measures which is based on the factor map Φ. In particular Lyapunov functions based on particle weight functions (see Kůrka [12]) work both for the measure space $\mathcal{M}_A$ and the generic space $\mathcal{G}_A$. The potential of Lyapunov functions for the classification of attractors has not yet been fully explored. This holds also for the connections between attractors in different topologies. While the theory of attractors is well established in compact spaces, in noncompact spaces there are several possible approaches. Finally, the comparison of entropy properties of CA in different topologies may be revealing for classification of CA.

There is even a more general approach to different topologies for CA based on the concept of submeasure on $\mathbb{Z}$. Since each submeasure defines a pseudometric, it would

be interesting to know, whether CA are continuous with respect to any of these pseudometrics, and whether some dynamical properties of CA can be derived from the properties of defining submeasures.

Acknowledgments

We thank Marcus Pivato and Francois Blanchard for careful reading of the paper and many valuable suggestions. The research was partially supported by the Research Program Project "Sycomore" (ANR-05-BLAN-0374).

Bibliography

Primary Literature

1. Besicovitch AS (1954) Almost periodic functions. Dover, New York
2. Blanchard F, Formenti E, Kůrka P (1999) Cellular automata in the Cantor, Besicovitch and Weyl spaces. Complex Syst 11(2):107–123
3. Blanchard F, Cervelle J, Formenti E (2005) Some results about the chaotic behaviour of cellular automata. Theor Comput Sci 349(3):318–336
4. Cattaneo G, Formenti E, Margara L, Mazoyer J (1997) A shift-invariant metric on $S^{\mathbb{Z}}$ inducing a nontrivial topology. Lecture Notes in Computer Science, vol 1295. Springer, Berlin
5. Formenti E, Kůrka P (2007) Subshift attractors of cellular automata. Nonlinearity 20:105–117
6. Hedlund GA (1969) Endomorphisms and automorphisms of the shift dynamical system. Math Syst Theory 3:320–375
7. Hurd LP (1990) Recursive cellular automata invariant sets. Complex Syst 4:119–129
8. Iwanik A (1988) Weyl almost periodic points in topological dynamics. Colloquium Mathematicum 56:107–119
9. Kamae J (1973) Subsequences of normal sequences. Isr J Math 16(2):121–149
10. Knudsen C (1994) Chaos without nonperiodicity. Am Math Mon 101:563–565
11. Kůrka P (1997) Languages, equicontinuity and attractors in cellular automata. Ergod Theory Dyn Syst 17:417–433
12. Kůrka P (2003) Cellular automata with vanishing particles. Fundamenta Informaticae 58:1–19
13. Kůrka P (2005) On the measure attractor of a cellular automaton. Discret Continuous Dyn Syst 2005(suppl):524–535
14. Marcinkiewicz J (1939) Une remarque sur les espaces de a.s. Besicovitch. C R Acad Sc Paris 208:157–159
15. Sablik M (2006) étude de l'action conjointe d'un automate cellulaire et du décalage: une approche topologique et ergodique. Ph D thesis, Université de la Mediterranée

Books and Reviews

Besicovitch AS (1954) Almost periodic functions. Dover, New York
Kitchens BP (1998) Symbolic dynamics. Springer, Berlin
Kůrka P (2003) Topological and symbolic dynamics. Cours spécialisés, vol 11 . Société Mathématique de France, Paris
Lind D, Marcus B (1995) An introduction to symbolic dynamics and coding. Cambridge University Press, Cambridge

Dynamics and Evaluation: The Warm Glow of Processing Fluency

Piotr Winkielman, David E. Huber
Department of Psychology,
University of California, San Diego, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Psychological Evidence for the Role of Fluency in Evaluation
Computational Mechanisms
Modeling Fluency-Affect Interactions: The Influence of Specific Variables
Neural Basis of Fluency–Affect Connection
Future Directions
Bibliography

Glossary

Evaluative responses Reactions reflecting an implicit or explicit assessment of stimulus' goodness or badness. Such reactions can be captured with judgments (liking, preference, choice), behavior (approach, avoidance, desire to continue or terminate), and physiology (peripheral and central). Strong evaluative reactions can develop into full-blown affective states, such as moods and emotion.

Fluency A general term used to describe efficiency of processing on perceptual and conceptual levels. Fluent processing is fast, error-free, and easy – a quality reflected in several dynamical properties (see next). The level of processing fluency can be monitored and influence evaluative as well as cognitive processes.

Processing dynamics Processing dynamic refers to content non-specific parameters characterizing a system's behavior at the level of individual units (e. g., neurons) or networks. Those dynamical parameters include (i) coherence of the signals within the system – the extent to which the signals arriving at a given unit from other units consistently dictate the same state, (ii) settling time – amount of time for the system to achieve a steady state, (iii) volatility – the number of units changing state, (iv) signal-to-noise ratio – overall strength of signals in the network, and (v) or differentiation – ratio of strongly activated to weakly activated units.

Mere-exposure effect Empirical observation that simple repetition (mere-exposure) of an initially neutral stimulus enhances people's liking for it. Because a typical result of simple repetition is enhancement of processing fluency (more efficient re-processing), this classic empirical phenomenon inspired research into evaluative consequences of changes in processing dynamics, and led to investigation of other variables that change processing dynamics, such as priming, duration, clarity, contrast, prototypicality or symmetry.

Definition of the Subject

A major goal of contemporary psychology, cognitive science and neuroscience is to understand the interactions of cognition and emotion – the mutual influences between thinking and feeling [11,41,101]. Sometimes such interactions are quite dramatic – as when emotions rob an individual of sound judgment (e. g., crimes of passions) or when emotions inspire an individual to transgress self-interest (e. g., acts of compassion). But most cognition-emotion interactions are more subtle. For example, seeing a well-balanced piece of art or a symmetric design may invoke a sense of aesthetic pleasure. Simply recognizing a familiar acquaintance on a street may evoke a sense of warm glow. A difficult to hear cell phone conversation may evoke a subtle sense of annoyance. And, in the first few days of being in a foreign country, people's faces just look "weird". This contribution deals with the nature of cognition-emotion interactions in these more subtle *everyday evaluative responses* – basic liking/disliking reactions that can be captured by people's preference judgments, behaviors, and physiological measures. Our special focus is on the role of non-specific, dynamical aspect of processing in the formation of evaluative responses. We show how the mechanisms underlying such responses can be understood using a combination of psychological experimentation and mathematical modeling grounded in physical models.

Introduction

Overview

The structure of our contribution is roughly as follows. First, we distinguish various sources of evaluative responses – non-specific processing dynamics and specific feature-based information. Next, we describe empirical work suggesting that evaluative reactions to processing dynamics can explain several common preference phenomena. Then, we describe some computational models of mechanisms underlying dynamics-affect connection. Finally, we discuss some neural underpinnings of the dynamic-affect connection and close with suggestions for future work.

Dynamical and Featural Information

You walk down a busy street and scan the passing faces. Some you like, some you do not. Why? Psychologists explore this question by focusing on the "how" and "what" of processing. The "how" refers to non-specific dynamical information about the quality of processing, and the "what" refers to specific feature-based information. Let us distinguish these two sources and briefly characterize their relationship.

During any kind of processing, sometimes even before any specific features are extracted from the stimulus, the mental system has access to a nonspecific source of information – the dynamics accompanying the processing of the stimulus. Historically, interest in the non-specific dynamics of processing originated in the field of metacognition [37,47,50]. This work highlighted that people monitor not only the content ("what") of mental representations, but also the "how" of processing, including such non-specific parameters as processing speed, ease, representation strength, volatility, and the degree of match between the incoming information and stored representations. Although there are substantial differences between these various parameters, it is common to refer to these non-specific aspects of processing with the general term of "*fluency*" (for reviews see [34,77]). As we describe shortly, the central idea guiding this review is that high fluency is typically associated with positive evaluations.

Evaluation is also obviously influenced by "what" is extracted – the stimulus' specific *features*. Thus, the overall positivity of a response to a face of a passing stranger will also depend on detection of such features (e. g., a smile or a symmetrical appearance), and on the perceivers idiosyncratic appraisals of these features (e. g., interest in meeting a stranger). There are many available reviews of the experimental and modeling work on featural processing in evaluation and emotion, so we will not repeat it here [1,3,76]. But we want to highlight a couple of things about the relation between dynamical and featural aspects of processing. First, both sources of information are available simultaneously, with each contributing to the net evaluative reaction. For example, positivity from detecting a smile can combine with positivity from high fluency of recognition. Second, the dynamical and featural sources can play off each other. For example, the same feature, such as symmetry, might create a positive reaction because of its implications (e. g., good health), but also make the face easier to recognize. In other words, a feature might not only cre-

ate an evaluative reaction directly, but also indirectly, via its influence on the processing dynamics.

How is it Going? Linking Dynamics and Affect

The idea that dynamical aspects of information processing have affective implications has been discussed in several domains of research. The major proposals focus on the role of dynamics as cue to the quality of the internal state of the system, or as a cue to the quality of an external stimulus.

Feedback About Quality of Internal Processing

At least since Simon [81], psychologists assume that one function of affect to provide information about the internal state of the system. Thus, unless there is an obvious external cause for feeling good or bad, the presence of negative affect conveys that something is internally "wrong", whereas positive affect conveys that things are going "right" (e. g., [8]). More specifically, affect can provide information about the current state of cognitive operations. Thus, high fluency of a perceptual or a conceptual process indicates progress toward, for example, successful recognition of the stimulus or a successful solution of a task. Besides informing the organism that processing is going well, positive affect triggered by high fluency may play a motivational function and reinforce the successful strategy [62,91]. On the other hand, low fluency can be a signal of cognitive error or incompatibility, and play a motivational role in revision of a processing strategy [12,18]. These ideas converge with classic observations that mental states characterized by low coherence, such as cognitive dissonance, are unpleasant, as indicated by self-report as well as physiological measures of affect [24].

Feedback About Quality of External Stimuli

Processing dynamics can also have affective consequences because it informs (probabilistically) whether an external stimulus is good or bad. For example, it's known, at least since Titchener [90], that familiar stimuli elicit a "warm glow." Conversely, illusions of familiarity (oldness) can be produced through unobtrusive inductions of positive affect [20,60]. One reason for this warmth-familiarity link could be biological predispositions for caution in encounters with novel, and thus potentially harmful, stimuli Zajonc [101]. Other accounts suggest that familiarity is just a learned, "fast and frugal" heuristic for easily identifying choices that are in truth objectively better [21]. Similarly, as we discuss next, dynamics could offer a probabilistic cue regarding other valued properties of external stimuli, such as symmetry, prototypicality, etc.

Psychological Evidence for the Role of Fluency in Evaluation

So far, we have focused on theoretical reasons for the dynamics-affect connection. The specific empirical research on the role of dynamical information in affect has centered around five related variables: (i) repetition/mere exposure, (ii) priming, (iii) contrast, clarity, duration, (iv) symmetry and (v) prototypicality. As we show, all these preference phenomena are consistent with the notion that high processing fluency enhances evaluations (for more comprehensive review see [66,99]).

Mere-Exposure/Repetition

The "mere exposure effect" (MEE) is the observation that simple repetition enhances liking for an initially neutral stimulus [101]. Interestingly, all that is required for the MEE is that the stimulus is "merely" shown, however briefly or incidentally, to the individual – no reinforcement is required and the presentation can be even subliminal (for reviews see [5]). The reader has probably experienced this phenomenon many times. Thus, most melodies and paintings "grow on you" with repeated exposure, faces that are simply familiar tend to generate a "warm glow," and advertisers try to increase sales by simply repeating product's name or image. Anecdotes aside, empirical evidence for the mere exposure effects is quite robust. For example in a study by Monahan, Murphy, and Zajonc [51], participants were subliminally exposed to 25 pictures of novel ideographs, and were later asked to report their tonic mood. For some participants, each of the 25 ideographs was different, while for other participants, 5 different ideographs were repeated 5 times each. The results showed that participants who were subliminally exposed to repeated ideographs reported being in a better mood than participants exposed to 25 different ideographs. Additional evidence for the positivity of reactions from the mere exposure effect comes from studies that used facial electromyography (EMG). This technique relies on the observation that positive affective responses manifest themselves in incipient smiles, as reflected by higher activity over the cheek region – zygomaticus major – whereas negative affective responses manifest themselves in incipient frowns, as reflected by higher activity over the brow region – corrugator supercilii [7]. Harmon-Jones and Allen [25] observed that repeatedly presented stimuli elicited stronger EMG activity over the "smiling" region of the participants' face (cheek), indicative of positive affect, without changing the activity over the "frowning" region (brow).

There is now good evidence that the mere-exposure effect reflects changes in processing fluency – the ease of recognition (e. g., [6,34,36,45,79]) . Stimulus repetition speeds up stimulus recognition and enhances judgments of stimulus clarity and presentation duration, which are indicative of processing facilitation (e. g., [22,33]).

Priming

Based on the just mentioned research, we may expect that any variable that facilitates processing should result in increased liking, even under conditions of a single exposure. Several studies confirmed this possibility. In one of these studies (see Study 1 in [65]) participants were exposed to pictures of everyday objects (e. g., a desk, bird, or plane). The processing fluency of these target pictures was facilitated or inhibited by subliminal presentation of visual contours (e. g., [2]). Some target pictures were preceded by matched contours (e. g., contour of a desk followed by a picture of the desk), whereas others were preceded by mismatched contours (e. g., contour of a desk followed by a picture of a bird). Some participants were asked to indicate how much they liked the target pictures; other participants were asked to press a button as soon as they could recognize the object in the picture, thus providing an independent measure of processing ease. The data showed that pictures preceded by matched contours were recognized faster, indicating higher fluency, and were liked more than pictures preceded by mismatched contours.

Importantly, Winkielman and Cacioppo [95] provided evidence for the positivity of reactions caused by priming using the earlier-mentioned technique of facial electromyography (fEMG). High fluency was associated with stronger activity over the zygomaticus region (indicative of positive affect), but was not associated with stronger activity of the corrugator region (indicative of negative affect). This effect occurred in the first 3 seconds after the presentation of the stimulus, which was several seconds before participants made their overt judgments. This suggests a quick link between high fluency and positive affect.

Contrast, Clarity, and Duration

High contrast and clarity have repeatedly been identified as characteristics of aesthetically appealing objects (e. g., [86]). According to our proposal, these properties trigger liking because they facilitates processing. In one study (see Study 2 in [65]) we manipulated fluency through different degrees of figure-ground contrast, taking advantage of the observation that high contrast decreases identification speed [9]. Participants liked the same stimulus more when it was presented with higher contrast, and hence could be processed more fluently. In another study (see Study 3 in [65]) we manipulated fluency through subtle increases in presentation duration, taking advantage of the observation that longer presentation durations facilitate the extraction of information [44]. As expected, participants evaluated the same stimulus more positively when it was presented for a longer duration, even if they were were unaware that duration was manipulated. Winkielman and Cacioppo [95] replicated these results and also found corresponding changes in EMG activity, which suggests that high fluency elicits positive affect on the physiological level.

Symmetry

Humans and non-human animals show a widespread preference for symmetry [67]. This is often attributed to the biological value of symmetry as a signal of mate quality (e. g., [89]). However, we propose that symmetry is appealing at least partly because it facilitates information processing. After all, symmetrical stimuli are structurally simpler, and thus more fluent, than non-symmetrical stimuli. Support for this comes from studies on preference and fluency of abstract shapes [64]. These researchers asked participants to make preference judgments and also same-different judgments for symmetrical and asymmetrical shapes. The results showed that symmetrical shapes are not only more appealing, but also easier to identify than comparable asymmetrical shapes. This finding is compatible with earlier studies by Palmer and his colleagues showing that symmetry is preferred, as long as it facilitates information processing. Specifically, Palmer [58] presented the same symmetrical dot patterns (such that overall amount of information was held constant) in one of three orientations – vertically, diagonally, or horizontally – and asked participants to rate the figural goodness of each of the patterns. He found that dot patterns presented in the vertically symmetrical orientation received the highest figural goodness ratings, followed by those presented in the horizontally symmetrical orientation, with those presented in the diagonally symmetrical orientation receiving the lowest figural goodness ratings. Importantly, the figural goodness ratings paralleled earlier work by Palmer and Hemenway [59] on ease of symmetry detection: symmetry in the dot patterns presented in vertically symmetrical orientations was detected the fastest, followed by the symmetry in the horizontally symmetrical orientations, with the symmetry of the dot patterns presented in diagonally symmetrical orientations being the most difficult to detect. Since each of the patterns in the three orientations con-

tained the same amount of information, this result suggests that symmetry makes any given stimulus more appealing because it facilitates the ability of the perceiver to detect redundant information and, as such, to more easily identify the stimulus.

Prototypicality

Another robust source of preference is prototypicality or "averageness" – in the sense of a stimulus providing the "best representation" of the category, or fitting its central tendency [67]. People show prototypicality preference for living objects, such as faces, fish, dogs and birds, and also for nonliving objects, such as color patches, furniture, wristwatches and automobiles [23,40]. This effect, known since Galton [19], has also been explained as reflecting evolved predisposition to interpret prototypicality as a cue to mate quality [88]. However, there is a more straightforward dynamical explanation. Given that prototypes are the most representative members of their categories, they are also fluent, as reflected in accuracy and speed of classification [66]. This raises the possibility that prototypes are liked *because* they are fluent. Winkielman, Halberstadt, Fazendeiro, and Catty [96] examined this idea in a series of three experiments. Participants first learned a category of random dot patterns (Experiment 1) or of common geometric patterns (Experiment 2) and then were presented with novel patterns varied across different levels of prototypicality. Participants classified these patterns into their respective categories as quickly as possible (measure of fluency), and also rated the attractiveness of each. A close relationship between fluency, attractiveness, and the level of prototypicality was observed. Both fluency and attractiveness increased with prototypicality. Importantly, when fluency was statistically controlled, the relation between prototypicality and attractiveness dropped by half (though it remained significant). This suggests that processing facilitation is important to, but not the sole cause of the "beauty-in-averageness" effect. Finally, Experiment 3 showed that viewing prototypical, rather than non-prototypical patterns elicited significantly greater EMG activity, suggesting that viewing prototypes involves genuine affective reactions.

In combination, the above studies, based on manipulations of repetition, figure-ground contrast, presentation duration, symmetry, and prototypicality consistently show that high perceptual fluency leads to more positive evaluations of the perceived stimuli. However, verbal descriptions of fluency are often vague and so fluency is often difficult to quantify – what exactly does it mean that one stimulus is more fluent than another? The answer to this question has been provided by computational models inspired by physical phenomena.

Computational Mechanisms

There is surprisingly little research on the role of dynamical parameters in cognition and emotion [55,61]. One notable exception is the Neural network approach, or connectionism, in which cognition is viewed in terms of the passage of activation among simple, neuron-like units organized in large, densely interconnected networks [73]. The individual units function as simple processors that can influence each other through connections, which vary in strength and sign (facilitatory or inhibitory). This massively interconnected and parallel architecture gives the neural network approach a certain neurophysiological realism and makes it suitable for a wide variety of applications. For more biological applications, one can conceptualize the network units as actual neurons, whereas for more psychological applications, one can treat the units as blocks of neurons or functional sub-systems [57]. Many different neural network architectures have been proposed that utilize dynamical parameters. Below we primarily focus on a proposal by Lewenstein and Nowak [42], which illustrates the role of dynamical parameters in learning and recognition using a simple attractor neural network [26]. Although this is in some regards an overly simplified model, the conceptual framework of the attractor network has been successfully expanded to more complicated applications, such as the plasticity-stability dilemma [52], and more realistic biological assumptions [54,84]. We address some of these more complex models later.

Fluency in a Hopfield Network

In a typical Hopfield network, representations are encoded as attractors of the network, i. e. states into which the network dynamics converge. The processing of information with the network can be seen as a gradual, evolving process, during which each neuron adjusts to the signal coming from other neurons. Because neurons are reciprocally connected, and because there are a large number of paths connecting one neuron to another, activation can reverberate dynamically through the network over simulated time steps until the network settles on the identified representation. For example, when presented with a to-be-recognized pattern, the network goes through a series of adjustments and after some time approaches a stable state, an attractor, corresponding to the "recognition" of a particular pattern.

Lewenstein and Nowak [42] proposed that a typical Hopfield model can be extended with a simple control

mechanism, which allows the network to monitor the dynamics of its own processing. Such a control mechanism can measure a variety of dynamical parameters, such as settling time, volatility, signal strength, coherence, and so on. These formally related properties can then be used by the network to roughly monitor the quality of its own processing (e. g., is it going well?) as well as estimate the characteristics of the stimuli being processed (e. g., is it familiar).

Studies with this model focused on how monitoring the dynamical properties of cognition can allow the network to estimate proximity to its closest attractor during the recognition process. This, in turn, allows the network to estimate the likelihood that the presented pattern is "known", without requiring full specification for the manner in which the attractor is known. Specifically, two key dynamical properties were identified. The first property is the network's "volatility", or the proportion of neurons changing their state at a given point. When the incoming, "to-be-recognized" pattern matches or closely approximates a known pattern, corresponding to one of the attractors (memories), the network is characterized by a relatively small proportion of neurons changing their state. When the incoming pattern is novel, and thus does not approximate one of the attractors, the network is characterized by a large number of neurons changing their state. The second key dynamical property is the coherence of the signals received by the neurons. In the vicinity of an attractor (old pattern), the signals arriving from other neurons at a given neuron are consistent in that they belong to the same pattern. However, when the network is far from an attractor (new pattern), the signals arriving from other neurons at a given neuron dictate conflicting states and may provide partial matches to a variety of other patterns. A closely related criterion is the signal-to-noise ratio. In the vicinity of the attractor (old pattern), signals from other neurons typically add up, resulting in a relatively large summary signal dictating the state of a given neuron. However, far from an attractor (new pattern), signals from other neurons cancel each other, resulting in a relatively weak summary signal dictating the state of a given neuron. As a consequence, the processing of "old" patterns is characterized by a higher signal-to-noise ratio than the processing of "new" patterns.

Extension to Graded Representations and Incremental Change

Traditional Hopfield networks use simulated neurons that are either "on" or "off", with no graded signal between these states. More realistic simulated neurons use a continuous range of intermediary values, allowing a graded measure for the magnitude and speed of settling into attractor states (e. g., [57]). However, because many applications are focused on learning and representational change, large simulated time steps are used and settling occurs in less than 10 time steps, which makes it difficult to measure relatively subtle differences in settling time. For such applications, fluency is measured rather indirectly as differentiation – the magnitude of the most active units [42]. Providing a more direct measure of fluency based on speed of processing, we have used neural simulations with millisecond time steps, which allows measurement of the time needed to achieve peak activation [28,30]. Not only does this provide a measure of choice preference, but it can be used to indicate reaction times [29]. In these real-time simulations, habituation dynamics are implemented such that activation achieves a peak value, but then falls to a lower value with continued processing. Because well learned representations include stronger connections, and because activation is the driving force behind habituation, familiar representations reach a peak value more quickly as habituation occurs more quickly [31].

Fast Fluency

The modeling work on fluency also shed light on the puzzling phenomenon when the system responds affectively to a pattern before it is fully recognized ("preference without inference", Zajonc [101]). Processing speed, volatility, differentiation, and the onset of habituation are all measurements of fluency that allow the network to estimate whether a pattern is "new" or "old" (i. e., proximity to its closest attractor) prior to explicit identification of the pattern. For instance, it is possible to determine the familiarity of incoming stimuli by monitoring how frequently a mere 10% of the neurons change their state during the very first time step [42]. Similarly, a fast familiarity signal can be based on the early differentiation [53]. It is also worth noting that checking the coherence of incoming signals makes it possible to estimate not only the global novelty of the whole pattern, but also the novelty of fragments in the perceived pattern, such as elements of an object or objects in a scene [103]. As discussed earlier, because familiarity is affectively positive, all these mechanisms explain how one can "like" something before even knowing what it is.

Fluency and Self-Regulation

In addition to quick feedback about the valence of the incoming stimulus, the early pre-recognition of familiarity may be used to control the recognition process, so that known stimuli are processed differently than new ones.

This may be achieved by linking the outcome of pre-recognition based on monitoring the system dynamics to a control parameter (e. g., network's overall noise level) that influences the later stages of the recognition process. A number of specific models that involve a feedback loop between pre-recognition and the noise level have been proposed. For example, in the original model by Lewenstein and Nowak [42], unknown patterns raised the noise level, preventing false "recognition" of unfamiliar patterns – a common problem for neural networks. In another example, by monitoring its own early dynamics a network can switch between recognizing known patterns and learning novel patterns [104]. Yet another implementation of this control mechanism allows a network to recognize the emotional quality of the stimulus in the pre-recognition process and use this emotional pre-recognition to facilitate the recognition of stimuli that are relevant to this emotion [102]. For an extensive model of how such loops are used in self-regulation, see Nowak and Vallacher [55] and also Vallacher and Nowak [91].

Modeling Fluency-Affect Interactions: The Influence of Specific Variables

So far, we have discussed computational models of fluency in terms of more general principles. In this section, we show that such models can be used to precisely specify the processing dynamics that underlie affective responses in several concrete empirical phenomena discussed earlier. To recall, experimental psychological research found that positive affect can be enhanced by repetition, priming, figure-ground contrast, presentation duration, symmetry, and prototypicality. How does this work computationally?

Repetition

Drogosz and Nowak [14] used a dynamic attractor neural network to simulate the effect of repetition on liking and explicit recognition. Specifically, they modeled the results of a study by Seamon, Marsh, and Brody [78] who exposed participants to 50 repetitions of polygons, presented at very brief exposure times ranging from 2 to 48 milliseconds. As in other mere exposure experiments, participants showed an increased preference for repeated polygons, even those presented at 2 and 8 milliseconds. Moreover, their preference increased with increasing exposure times, but reached asymptote at 24 milliseconds. In contrast, explicit recognition was at chance at low durations (2 and 8 milliseconds), and then gradually increased up to 90% recognition at 48 milliseconds. The model by Drogosz and Nowak [14] showed that the relationship between preference and recognition as a function of exposure time can be simulated by assuming that the affective response represents a non-specific signal about the early dynamics of the network, as indexed by the estimated proportion of change in the first time step, whereas the recognition response represents a stabilization of the network on a specific pattern, which takes approximately 6 time steps. A psychological interpretation that can be attached to these simulation data is that at very short presentation durations, the participants only have access to the non-specific fluency signal, which elicits positive affect and influences their preference judgments. With progressively longer presentation duration, the fluency signal (affective response) increases only marginally, whereas the recognition response continues to grow until it reaches nearly perfect performance. The above simulations show that many prior exposures to a pattern establish a relatively strong memory for this pattern, whereas few prior exposures establish a relatively weak memory for the pattern. Test patterns with relatively stronger memories (i. e., stronger attractors) are processed with higher processing fluency (less volatility, more coherent signals) than test patterns with weaker or no memories. These differential fluency signals are picked up early on, as indicated by the simulation, and precede the extraction of stimulus information. Because the fluency signal is hedonically marked, it allows for evaluative responses prior to stimulus recognition, as initially reported by Kunst-Wilson and Zajonc [38].

Computational models of this type can also help us conceptualize the results of studies that used all novel patterns and manipulated the fluency of processing through procedures like figure-ground contrast, presentation duration, symmetry, prototypicality, and priming. To account for these effects, the model requires only minimal modifications. Specifically, the above simulations were carried out in attractor networks composed of neurons with binary states, where a state of the neuron corresponds either to the presence or the absence of the feature preferred by that neuron [26]. However, the same "fluency" criteria (volatility, coherence, differentiation) apply to networks with continuous neurons, where the state of a neuron encodes the degree to which a feature is present or activated [27,57].

Duration, Clarity and Contrast

The influence of these liking-enhancing variables can be conceptualized as reflecting a process in which patterns presented for longer time, greater clarity, and higher contrast are represented by more extreme values of activation. All this leads to stronger signals in the network, more differentiated states of the neurons, and faster settling.

Symmetry

This highly valued feature is easily incorporated because the representation of symmetrical patterns is stronger. This is due to simplicity and redundancy (e. g., in faces the left side of the symmetrical faces is identical to the right) and position-independence in recognition (e. g., symmetrical face looks the same from different angles). In contrast, the representation of asymmetrical features is weaker due to complexity and position-dependence [15,35].

Prototypicality

The effects of prototypicality (responsible for the 'beauty-in-averages' effect) can result from converging exemplars creating a strong attractor for a prototype. As a result, the recognition of a prototype pattern typically involves faster settling time, and less volatility [97]. Recent computational models of fluency using a support vector machine (a nonlinear classifier) have also shown that prototypical faces are located further from a face/non-face classification boundary, which allows for more efficient categorization [72].

Priming

In neural networks, priming corresponds either to the pre-activation of neurons that encode the pattern (activation-based priming) or to temporary changes in weights between the neurons (weight-based priming). The effects of the prime and the actual target sum up in determining the state of neurons. This results in more extreme values of activation (i. e., better differentiation) of the neurons for primed versus non-primed patterns.

As mentioned previously, fluency might be better captured by models that simulate the real-time, millisecond by millisecond processing of information. With such a model, we have explained a variety of empirical priming phenomena by including habituation dynamics (decrease in responding as a result of repetition or strong activation). In this model the presentation of a minimal prime (i. e., a brief duration or a single stimulus) immediately prior to a target induces a positive priming effect through pre-activation that boosts fluency. This is similar to the above mentioned pre-activation model. However, our model also predicted that presentation of an excessive prime (i. e., a long duration or repeated stimulus) immediately prior to a target would eliminate the positive priming effect, or perhaps even induce a negative priming effect. This occurs because habituation to the prime produces a disfluency in processing the target (i. e., the response to the target occurs slowly). This transition from positive to negative priming as a function of prime duration explained a variety of priming phenomena in the domain of word identification [28] and recognition memory [31].

We recently demonstrated in several experiments that this fluency-disfluency dynamic also applies to the domain of evaluation, more specifically, the appearance and disappearance of evaluative priming effects [32]. These experiments explored predictions for situations that should produce or eliminate priming as a function of prime duration, prime-target similarity and target salience. More specifically, when the prime-target similarity is low, such as with extremely valenced prime words and ideograph targets that have no meaning, habituation to the prime does not produce habituation to the target, and so empirically there is no correction effect even with long duration primes. Furthermore, when the target is itself minimal (e. g., a subliminally presented target word), then there is only an assimilative pre-activation effect because, again, habituation to the prime does not cause a change in target processing. In short, the fluency-based priming models can explain not only when the evaluative priming occurs, but also when it disappears.

In sum, the just discussed computational models show that manipulations such as repetition, priming, presentation duration, figure-ground contrast, clarity, prime-target similarity, and prototypicality change fluency in the network dynamics. These changes in fluency can trigger an affective response via the monitoring mechanisms discussed earlier.

Neural Basis of Fluency–Affect Connection

The assumptions guiding the just discussed psychological and computational models are consistent with neuroscience evidence. So far, this work has focused on low-level perceptual effects of stimulus repetition, response of higher-level value-coding system to previously exposed stimuli, and effects of processing coherence and conflict.

Perceptual Response

There is much evidence that novel stimuli elicit a non-specific, undifferentiated activity, which gradually decreases with repetition [82,85]. More specifically, single cell recording and neuroimaging studies suggest that stimulus repetition tends to decrease non-specific activation and leads to more selective firing [13,71]. One interpretation of these data is that stimulus familiarization leads to a gradual differentiation of the neurons that represent the incoming stimulus from neurons that do not represent the stimulus [48,54].

Response of Value-Coding Regions

A specific example of neuroscience research that examined connections between novelty and evaluation comes from Elliot and colleagues. Elliot, Dolan, and Frith [17] reported an fMRI study on the neural substrates of delayed matching to sample, as compared with delayed non-matching to sample. In such a task, participants are initially shown an item and then are subsequently shown a pair of items whereby they have to identify either the item they saw or the one that they did not see. Significantly more activity occurred in the ventromedial OFC in the matching condition (while reprocessing an old item) as compared to the non-matching condition (while processing a novel item). This conclusion is consistent with results of an earlier PET study of the subliminal mere exposure effect by Elliot and Dolan [16]. In their study, when participants made preference judgments, repeated stimuli activated the medial PFC, an area closely connected to the medial OFC and that is also known for its role in reward processing. Notably, these neuroimaging studies showing activation of neural circuits involved in reward to repeated stimuli fit nicely with the earlier reviewed Harmon-Jones and Allen [25] EMG study showing greater zygomaticus activation to merely exposed items. Taken together, they highlight the multilevel hedonic consequences of mere exposure and, more generally, high processing fluency.

Processing Coherence and Conflict

There is also work on the neural basis of mechanisms involved in successful and unsuccessful integration of different cognitive representations [10]. Neuroimaging evidence highlights a particular role of the anterior cingulate cortex (ACC) [18,39]. Though originally thought of as primarily a "cognitive" structure, more recent studies suggest that enhanced ACC due to cognitive conflict is accompanied by negative affect and enhanced arousal [10]. If so, the ACC could provide a neural substrate by which processing coherence on the level of multiple representations translates into negative affect.

Future Directions

Several issues remain essential for further studies. First, we primarily focused on the role of perceptual sources of dynamical information. However, dynamical information is available across the entire processing spectrum, from simple pattern matching to semantic coherence of high-order conceptual content. Though there is some psychological work available on this issue, there is very little computational and neural work that deals with specific mechanisms [97]. Second, emerging evidence in psychology suggests that the impact of fluency is moderated by processing expectations – how much speed, effort, coherence is expected given the stimulus [93]. For example, the simulations by Drogosz and Nowak [14] discussed earlier were conducted using very similar patterns, as is typical in the mere-exposure studies. Accordingly, the absolute processing fluency of a given pattern was a reliable indicator of its "oldness." However, for the fluency signal to be informative in a more realistic situation, in which stimuli differ widely in overall signal strength, the network needs to scale the absolute value of the fluency signal for the particular pattern against the expected value [93] . A comparison between an observed value and an expected value can be derived with a computational likelihood ratio model, which is a class of model that has proven remarkably successful in explaining recognition memory based on familiarity [48,80]. Developing a similar Bayesian approach to fluency would likewise provide a precise account for the role of expectations in the study of affect. Finally, this contribution has emphasized the role of non-specific dynamical information in evaluation and has been silent on the role of specific stimulus features. However, we know very little about the interaction the proposed fluency based "how" and the content based "what" of processing, and exploring this interaction may prove a useful direction for future research.

Bibliography

Primary Literature

1. Anderson NH (1981) Foundations of information integration theory. Academic Press, New York
2. Bar M, Biederman I (1998) Subliminal visual priming. Psychol Sci 9:464–469
3. Beeman M, Ortony A, Monti LA (1995) Emotion-cognition interactions. In: Arbib MA (ed) The handbook of brain theory and neural networks. MIT Press, Cambridge, pp 360–363
4. Berlyne DE (1974) Studies in the new experimental aesthetics: Steps toward an objective psychology of aesthetic appreciation. Hemisphere, Washington
5. Bornstein RF (1989) Exposure and affect: Overview and meta-analysis of research, 1968-1987. Psychol Bull 106:265–289
6. Bornstein RF, D'Agostino PR (1994) The attribution and discounting of perceptual fluency: Preliminary tests of a perceptual fluency/attributional model of the mere exposure effect. Soc Cogn 12:103–128
7. Cacioppo JT, Bush LK, Tassinary LG (1992) Microexpressive facial actions as a function of affective stimuli: Replication and extension. Personality Soc Psychol Bull 18:515–526
8. Carver CS, Scheier MF (1990) Origins and functions of positive and negative affect: A control-process view. Psychol Rev 97:19–35

9. Checkosky SF, Whitlock D (1973) The effects of pattern goodness on recognition time in a memory search task. J Exp Psychol 100:341–348
10. Critchley HD (2005) Neural mechanisms of autonomic, affective, and cognitive integration. J Comp Neurol 493:154–166
11. Damasio AR (1994) Descartes' error: Emotion, reason and the human brain. Grosset/Putnam, New York
12. Derryberry D, Tucker DM (1994) Motivating the focus of attention. In: Niedenthal PM, Kitayama S (eds) The heart's eye. Academic Press, San Diego, pp 167–196
13. Desimone R, Miller EK, Chelazzi L, Lueschow A (1995) Multiple memory systems in the visual cortex. In: Gazzaniga MS (ed) The cognitive neurosciences. MIT Press, Cambridge, pp 475–490
14. Drogosz M, Nowak A (2006) A neural model of mere exposure: The EXAC mechanism. Pol Psychol Bull 37:7–15
15. Enquist M, Arak A (1994) Symmetry, beauty and evolution. Nature 372:169–172
16. Elliott R, Dolan R (1998) Neural response during preference and memory judgments for subliminally presented stimuli: A functional neuroimaging study. J Neurosci 18:4697–4704
17. Elliot R, Dolan RJ, Frith CD (2000) Dissociable functions in the medial and lateral oribitofrontal cortex: Evidence from human neuroimaging studies. Cereb Cortex 10:308–317
18. Fernandez-Duque D, Baird JA, Posner MI (2000) Executive attention and metacognitive regulation. Conscious Cogn 9:288–307
19. Galton F (1878) Composite portraits. J Anthropol Inst G B Irel 8:132–144
20. Garcia-Marques T, Mackie DM (2000) The positive feeling of familiarity: Mood as an information processing regulation mechanism. In: Bless H, Forgas J (eds) The message within: The role of subjective experience in social cognition and behavior. Psychology Press, Philadelphia, pp 240–261
21. Gigerenzer G (2007) Gut feelings: The intelligence of the unconscious. Viking Press, New York
22. Haber RN, Hershenson M (1965) The effects of repeated brief exposures on growth of a percept. J Exp Psychol 69:40–46
23. Halberstadt J, Rhodes G (2000) The attractiveness of nonface averages: Implications for an evolutionary explanation of the attractiveness of average faces. Psychol Sci 4:285–289
24. Harmon-Jones E (2000) A cognitive dissonance theory perspective on the role of emotion in the maintenance and change of beliefs and attitudes. In: Frijda NH, Manstead ARS, Bem S (eds) Emotion and Beliefs. Cambridge University Press, Cambridge, pp 185–211
25. Harmon-Jones E, Allen JB (2001) The role of affect in the mere exposure effect: Evidence from psychophysiological and individual differences approaches. Personality Soc Psychol Bull 27:889–898
26. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci 79:2554–2558
27. Hopfield JJ (1984) Neurons with graded response have collective computational properties like those of two-state neurons. Proc Natl Acad Sci 81:3088–3092
28. Huber DE (2008) Immediate priming and cognitive aftereffects. J Exp Psychol Gen 137:324–347
29. Huber DE, Cousineau D (2004) A race model of perceptual forced choice reaction time. Proceedings of the 25th Annual Conference of the Cognitive Science Society. Erlbaum Associates, Hillsdale, pp 687–692
30. Huber DE, O'Reilly RC (2003) Persistence and accommodation in short-term priming and other perceptual paradigms: Temporal segregation through synaptic depression. Cogn Sci A Multidiscip J 27:403–430
31. Huber DE, Clark TF, Curran T, Winkielman P. Effects of repetition priming on recognition memory: Testing a perceptual fluency-disfluency model. J Exp Psychol Learn Mem Cogn (in press)
32. Huber DE, Winkielman P, Parsa A, Chun WY. Too much of a good thing: Testing a Bayesian model of evaluative priming. Submitted
33. Jacoby LL (1983) Perceptual enhancement: Persistent effects of an experience. J Exp Psychol Learn Mem Cogn 9:21–38
34. Jacoby LL, Kelley CM, Dywan J (1989) Memory attributions. In: Roediger HL, FIM Craik (eds) Varieties of memory and consciousness: Essays in honour of Endel Tulving. Erlbaum, Hillsdale, pp 391–422
35. Johnstone RA (1994) Female preference for symmetrical males as a by-product of selection for mate recognition. Nature 372:172–175
36. Klinger MR, Greenwald AG (1994) Preferences need no inferences?: The cognitive basis of unconscious mere exposure effects. In: Niedenthal PM, Kitayama S (eds) The heart's eye. Academic Press, San Diego, pp 67–85
37. Koriat A (2000) The feeling of knowing: Some metatheoretical implications for consciousness and control. Conscious Cogn 9:149–171
38. Kunst-Wilson WR, Zajonc RB (1980) Affective discrimination of stimuli that cannot be recognized. Science 207:557–558
39. Lane RD, Reiman EM, Axelrod B, Yun L, Holmes A, Schwartz GE (1998) Neural correlates of levels of emotional awareness: Evidence of an interaction between emotion and attention in the anterior cingulate cortex. J Cogn Neurosci 10:525–535
40. Langlois JH, Roggman LA (1990) Attractive faces are only average. Psychol Sci 1:115–121
41. LeDoux JE (1996) The Emotional Brain. Touchstone, New York
42. Lewenstein M, Nowak A (1989) Recognition with self-control in neural networks. Phys Rev 40:4652–4664
43. Losch ME, Cacioppo JT (1990) Cognitive dissonance may enhance sympathetic tonus, but attitudes are changed to reduce negative affect rather than arousal. J Exp Soc Psychol 26: 289–304
44. Mackworth JF (1963) The duration of the visual image. Canadian J Psychol 17:62–81
45. Mandler G, Nakamura Y, Van Zandt BJ (1987) Nonspecific effects of exposure on stimuli that cannot be recognized. J Exp Psychol Learn Mem Cogn 13:646–648
46. Martindale C, Moore K (1988) Priming, prototypicality, and preference. J Exp Psychol Hum Percept Perform 14:661–670
47. Mazzoni G, Nelson TO (1998) Metacognition and cognitive neuropsychology: Monitoring and control processes. Lawrence Erlbaum, Mahwah
48. McClelland JL, Chappell M (1998) Familiarity breeds differentiation: A Bayesian approach to the effects of experience in recognition memory. Psychol Rev 105:724–760
49. Metcalfe J (1993) Novelty monitoring, metacognition, and control in a composite holographic associative recall model: Implications for Korsakoff Amnesia. Psychol Rev 100:3–22

50. Metcalfe J, Shimamura AP (1994) Metacognition: Knowing about knowing. MIT Press, Cambridge
51. Monahan JL, Murphy ST, Zajonc RB (2000) Subliminal mere exposure: Specific, general, and diffuse effects. Psychol Sci 6:462–466
52. Murre JMJ, Phaf RH, Wolters G (1992) CALM: Categorizing and learning module. Neutral Netw 55–82
53. Norman KA, O'Reilly RC (2003) Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. Psychol Rev 110:611–646
54. Norman KA, O'Reilly RC, Huber DE (2000) Modeling hippocampal and neocortical contributions to recognition memory. Poster presented at the Cognitive Neuroscience Society Meeting, San Francisco
55. Nowak A, Vallacher RR (1998) Dynamical social psychology. Guilford Press, New York
56. Oatley K, Johnson-Laird P (1987) Towards a cognitive theory of emotions. Cogn Emot 1:29–50
57. O'Reilly RC, Munakata Y (2000) Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain. MIT Press, Cambridge
58. Palmer SE (1991) Goodness, gestalt, groups, and Garner: Local symmetry subgroups as a theory of figural goodness. In: Pomerantz JR, Lockhead GR (eds) Perception of Structure. APA, Washington
59. Palmer SE, Hemenway K (1978) Orientation and symmetry: Effects of multiple, near, and rotational symmetries. J Exp Psychol Hum Percept Perform 4:691–702
60. Phaf RH, Roteveel M (2005) Affective modulation of recognition bias. Emotion 5(3):309–318
61. Port RT, van Gelder T (1995) Mind as motion: Exploration in the dynamics of cognition. MIT Press, Cambridge
62. Posner MI, Keele SW (1968) On the genesis of abstract ideas. J Exp Psychol 77:353–363
63. Ramachandran VS, Hirstein W (1999) The science of art: A neurological theory of aesthetic experience. J Conscious Stud 6:15–51
64. Reber R, Schwarz N (2006) Perceptual fluency, preference, and evolution. Pol Psychol Bull 37:16–22
65. Reber R, Winkielman P, Schwarz N (1998) Effects of perceptual fluency on affective judgments. Psychol Sci 9:45–48
66. Reber R, Schwarz N, Winkielman P (2004) Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? Personality Soc Psychol Rev 8:364–382
67. Rhodes G (2006) The evolution of facial attractiveness. Annu Rev Psychol 57:199–226
68. Rhodes G, Tremewan T (1996) Averageness, exaggeration, and facial attractiveness. Psychol Sci 7:105–110
69. Rhodes G, Proffitt F, Grady JM, Sumich A (1998) Facial symmetry and the perception of beauty. Psychon Bull Rev 5:659–669
70. Roediger HL (1990) Implicit memory: Retention without remembering. American Psychol 45:1043–1056
71. Rolls ET, Baylis GC, Hasselmo ME, Nalwa V (1989) The effect of learning on the face selective responses of neurons in the cortex in the superior temporal sulcus of the monkey. Exp Brain Res 76:153–164
72. Rosen LH, Bronstad PM, Griffin AM, Hoss RA, Langlois JH. Children, adults, and a computational model identify attractive faces more fluently: Evidence in support of averageness theory of facial attractiveness (under review)
73. Rumelhart DE, McClelland JL (1986) Parallel Distributed Processes: Exploration in Microstructure of Cognition. MIT Press, Cambridge
74. Schachter SE, Singer J (1962) Cognitive, social and physiological determinants of emotional state. Psychol Rev 69:379–399
75. Schacter DL (1992) Understanding implicit memory: A cognitive neuroscience approach. American Psychol 47:559–569
76. Schwarz N (1998) Accessible content and accessibility experiences: The interplay of declarative and experiential information in judgment. Personality Soc Psychol Rev 2:87–99
77. Schwarz N, Clore GL (1996) Feelings and phenomenal experiences. In: Higgins ET, Kruglanski AW (eds) Social Psychology: Handbook of Basic Principles. The Guilford Press, New York
78. Seamon JG, Marsh RL, Brody N (1984) Critical importance of exposure duration for affective discrimination of stimuli that are not recognized. J Exp Psychol Learn Mem Cogn 10:465–469
79. Seamon JG, McKenna PA, Binder N (1998) The mere exposure effect is differentially sensitive to different judgment tasks. Conscious Cogn 7:85–102
80. Shiffrin RM, Steyvers M (1997) A model for recognition memory: REM: Retrieving effectively from memory. Psychon Bull Rev 4(2):145–166
81. Simon HA (1967) Motivational and emotional controls of cognition. Psychol Rev 74:29–39
82. Skarda CA, Freeman WJ (1987) How brains make chaos in order to make sense of the world. Behav Brain Sci 10:161–195
83. Smith ER (1998) Mental representation and memory. In: Gilbert DT, Fiske ST, Lindzey G (eds) The Handbook of Social Psychology pp 269–322; The McGraw-Hill Companies, Boston
84. Smith ER (2000) Subjective experience of familiarity: Functional basis in connectionist memory. In: Bless H, Forgas JP (eds) The message within: The role of subjective experience in social cognition and behavior. Psychology Press, Philadelphia, pp 109–124
85. Sokolov EN (1963) Perception and the orienting reflex. MacMillan, NY
86. Solso RL (1997) Cognition and the visual arts. MIT Press, Cambridge
87. Squire LR (1992) Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. Psychol Rev 99:195–231
88. Symons D (1979) Evolution of human sexuality. Oxford University Press, New York
89. Thornhill R, Gangestad SW (1993) Human facial beauty: Averageness, symmetry, and parasite resistance. Hum Nat 4:237–269
90. Titchener EB (1910) A textbook of psychology. Macmillan, New York
91. Vallacher RR, Nowak A (1999) The dynamics of self-regulation. In: Jr. Wyer RS (ed) Perspectives on behavioral self-regulation. Lawrence Erlbaum Associates, Mahwah, pp 241–259
92. Whittlesea BWA (1993) Illusions of familiarity. J Exp Psychol Learn Mem Cogn 19:1235–1253
93. Whittlesea BWA, Williams LD (2001) The Discrepancy-Attribution Hypothesis: I. The Heuristic Basis of Feelings of Familiarity. J Exp Psychol Learn Mem Cogn 27:3–13
94. Winkielman P, Berntson GG, Cacioppo JT (2001) The psychophysiological perspective on the social mind. In: Tesser A,

Schwarz N (eds) Blackwell Handbook of Social Psychology: Intraindividual Processes. Blackwell, Oxford, pp 89–108
95. Winkielman P, Cacioppo JT (2001) Mind at ease puts a smile on the face: Psychophysiological evidence that processing facilitation leads to positive affect. J Personality Soc Psychol 81:989–1000
96. Winkielman P, Halberstadt J, Fazendeiro T, Catty S (2006) Prototypes are attractive because they are easy on the mind. Psychol Sci 17:799–806
97. Winkielman P, Hooda P, Munakata Y (2004) Neural network model of fluency for average patterns. Unpublished manuscript. University of Denver
98. Winkielman P, Schwarz N (2001) How pleasant was your childhood? Beliefs about memory shape inferences from experienced difficulty of recall. Psychol Sci 2:176–179
99. Winkielman P, Schwarz N, Fazendeiro T, Reber R (2003) The hedonic marking of processing fluency: Implications for evaluative judgment. In: Musch J, Klauer KC (eds) The Psychology of Evaluation: Affective Processes in Cognition and Emotion. Lawrence Erlbaum, Mahwah, pp 189–217
100. Zajonc RB (1968) Attitudinal effects of mere exposure. J Personality Soc Psychol Monogr Suppl 9:1–27
101. Zajonc RB (1998) Emotions. In: Gilbert DT, Fiske ST, Lindzey G (eds) The Handbook of Social Psychology. McGraw-Hill, Boston, pp 591–632
102. Zochowski M, Lewenstein M, Nowak A (1993) A memory which tentatively forgets. J Phys A 26:2099–2112
103. Zochowski M, Lewenstein M, Nowak A (1994) Local noise in neural networks with self-control. International J Neural Syst 5:287–298
104. Zochowski M, Lewenstein M, Nowak A (1995) SMARTNET – A neural net with self-controlled learning. Network 6:93

Books and Reviews

Reber R, Schwarz N, Winkielman P (2004) Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? Personality Soc Psychol Rev 8:364–382
Winkielman P, Schwarz N, Fazendeiro T, Reber R (2003) The hedonic marking of processing fluency: Implications for evaluative judgment. In: Musch J, Klauer KC (eds) The Psychology of Evaluation: Affective Processes in Cognition and Emotion. Lawrence Erlbaum, Mahwah, pp 189–217

Dynamics on Fractals

Raymond L. Orbach
Department of Physics, University of California, Riverside, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Fractal and Spectral Dimensions
Nature of Dynamics on Fractals – Localization
Mapping of Physical Systems onto Fractal Structures
Relaxation Dynamics on Fractal Structures
Transport on Fractal Structures
Future Directions
Bibliography

Glossary

Fractal Fractal structures are of two types: deterministic fractals and random fractals. The former involves repeated applications of replacing a given structural element by the structure itself. The process proceeds indefinitely, leading to *dilation symmetry*: if we magnify part of the structure, the enlarged portion looks the same as the original. Examples are the Mandelbrot–Given fractal and the Sierpinski gasket. A random fractal obeys the same properties (e. g. dilation symmetry), but only in terms of an ensemble average. The stereotypical example is the percolating network. Fractals can be constructed in any dimension, d, with for example, $d = 6$ being the mean-field dimension for percolating networks.

Fractal dimension The fractal dimension represents the "mass" dependence upon length scale (measuring length). It is symbolized by D_f, with the number of sites on a fractal as a function of the measurement length L being proportional to L^{D_f}, in analogy to a homogeneous structure embedded in a dimension d having mass proportional to the volume spanned by a length L proportional to L^d.

Spectral (or fracton) dimension The spectral (or fracton) dimension refers to the dynamical properties of fractal networks. It is symbolized by $\tilde{d}_s$ and can be most easily thought of in terms of the density of states of a dynamical fractal structure (e. g. vibrations of a fractal network). Thus, if the excitation spectrum is measured as a function of energy ω, the density of states for excitations of a fractal network would be proportional to $\omega^{(\tilde{d}_s-1)}$, in analogy to a homogeneous structure embedded in a dimension d having a density of states proportional to $\omega^{(d-1)}$.

Localization exponent Excitations on a fractal network are in general *strongly localized* in the sense that wave functions fall off more rapidly than a simple exponential, $\Psi(r) \sim \exp[-\{[r/\Lambda(\omega)]^{d_\phi}\}]$ where $\Lambda(\omega)$ is an energy dependent localization length, and the exponent d_ϕ is in general greater than unity.

Definition of the Subject

The dynamical properties of fractal networks are very different from homogeneous structures, dependent upon

strongly localized excitations. The thermal properties of fractals depend upon a "spectral dimension" $\tilde{d}_s$ less than the "Euclidean" or embedding dimension d. Fractal randomness introduces statistical properties into relaxation dynamics. Transport takes place via hopping processes, reminiscent of Mott's variable range-rate hopping transport for electronic impurity states in semi-conductors. Fractal dynamics serve as a guide for the behavior of random structures where the short length scale excitations are localized.

Introduction

The study of the dynamics on and of fractal networks [29,30,31] is not an arcane investigation, with little application to real physical systems. Rather, there are many examples of real materials that exhibit fractal dynamics. Examples are the vibrational properties of silica aerogels [21,45], and magnetic excitations in diluted antiferromagnets [18,19,43]. Beyond these explicit physical realizations, one can learn from the very nature of the excitations on fractal structures how localized excitations in homogeneous materials behave. That is, the dynamics of fractal networks serve as a calculable model for transport of localized excitations in random structures that are certainly not mass fractals. Examples are thermal transport in glasses above the so-called "plateau" temperature [2,20,26], and the lifetime of high-energy lattice vibrations in a-Si [32,33,36,37]. As we shall show in subsequent sections, the former is an example of vibrational hopping-type transport (similar in nature to Mott's variable-range hopping [22,23] for localized electronic states), while the latter exhibits counter-intuitive energy dependences for vibrational lifetimes.

The following Sect. "Fractal and Spectral Dimensions" describes in brief terms the concept of the fractal (or mass) dimension, D_f [7,8,9,10,15,40,41,46], and the spectral (or fracton) dimension $\tilde{d}_s$ [5,14,17,34,35]. The latter is related to the anomalous diffusion characteristics of fractals (the so-called "ant in a labyrinth" introduced by de Gennes [8]), and is at the heart of the thermal properties of fractal structures. A conjecture introduced by Alexander and Orbach [5] suggested that the mean-field value (exact at $d = 6$) for $\tilde{d}_s$ for percolating networks, $\tilde{d}_s = 4/3$, might be universal, independent of dimension. It is now generally regarded that this conjecture is only approximate, though a very good approximation for percolating networks in $2 = d < 6$. Were it to be exact, it would mean that the dynamical properties of percolating networks could be expressed in terms of their geometrical properties.

Section "Nature of Dynamics on Fractals–Localization" introduces the nature of the wave function for an excitation embedded in a fractal structure. In general, random self-similar structures generate "super-localized" wave functions, falling off faster than exponential. We will characterize the envelope of the wave function by the functional form [1]

$$\Psi(r) \sim \exp[-\{[r/\Lambda(\omega)]^{d_\phi}\}] ,$$

where $\Lambda(\omega)$ is an energy-dependent localization length, and the exponent d_ϕ is in general greater than unity. For excitations on fractal structures, the localization length $\Lambda(\omega)$ can be calculated analytically, allowing for explicit computation of physical phenomena. This allows analytic expressions for scattering lifetimes and excitation transport.

Mapping of the diffusion problem onto the secular equation for scalar lattice vibrations allows the extraction of the density of vibrational states [5]. It is given by the simple relationship $D(\omega) \sim \omega^{(\tilde{d}_s - 1)}$ displaying the utility of the spectral dimension $\tilde{d}_s$. A similar though slightly more complicated mapping can be applied to diluted Heisenberg antiferromagnets. The silica aerogels [13,16,42,44,45], and $Mn_xZn_{1-x}F_2$ [43] and $RbMn_x Mg_{1-x}F_3$ [18,19] are realizations of such systems, respectively, and have been extensively investigated experimentally.

The random nature of fractal structures introduces statistical complexity into relaxation processes. The direct (single vibration) [38] and Raman (scattering of two vibrations) [24,39] spin-lattice relaxation processes are discussed for random fractal networks in Sect. "Relaxation Dynamics on Fractal Structures". The theoretical results are complex, related to the "Devil's staircase" for Raman relaxation processes. Experiments probing these rather bizarre behaviors would be welcome.

Given the "super-localization" of excitations on fractal networks [1,42], transport properties need to be expressed according to the same concepts as introduced by Mott [22,23] for the insulating phase of doped semiconductors—the so-called "variable range-rate hopping". This is developed in Sect. "Transport on Fractal Structures", and applied to the thermal transport of fractal structures [2,20,26]. The "lessons learned" from these applications suggest an interpretation of the thermal transport of glassy materials above the so-called "plateau" temperature range. This instance of taking the insights from fractal structures, and applying it to materials that are certainly not fractal yet possess localized vibrational states, is an example of the utility of the analysis of fractal dynamics.

Finally, some suggestions, hinted at above, for future experimental investigations are briefly discussed in Sect. "Future Directions".

Fractal and Spectral Dimensions

Fractals can be defined by the nature of the symmetry they exhibit: self-similar geometry or equivalently dilation symmetry [7,8,9,10,15,40,41,46]. This symmetry is most assuredly not translational, leading to a very different set of behaviors for physical realizations that shall be explored in subsequent Sections. Stated most simply, self-similar structures "look the same" on any length scale. Conversely, one cannot extract a length scale from observation of a self-similar structure. There are clearly limits to self-similarity for physical realizations. For example, at the atomic level, dilation symmetry must come to an end. And for finite structures, there will be a largest length scale. There are other length scales that bracket the self-similar regime. Thus, for percolating networks, an example of a random fractal [38,39], the percolation correlation length, $\xi(p)$, where p is the concentration of occupied sites or bonds, defines an upper limit to the regime of fractal geometry. In general, for percolating networks, the range of length scales ℓ over which the structure is fractal is sandwiched between the atomic length a, and the percolation correlation length $\xi(p)$: $a < \ell < \xi(p)$.

Fractal structures can be deterministic or random. The former is a result of applying a structural rule indefinitely, replacing an element of a design by the design itself. Examples are the Mandelbrot–Given fractal and the Sierpinski gasket. The latter arises from a set of probabilistic rules. A simple example is the percolating network where a site or bond (e. g. in $d = 2$, a square grid) is occupied randomly with probability p. The resulting clusters contain a random number of sites or bonds, where the distribution function for the number of finite clusters of size s is denoted by $n_s(p)$. There exists a critical value $p = p_c$ where a connected cluster extends over all space (the "infinite" cluster). For $p < p_c$ only finite clusters exist, with the largest cluster spanned by the length scale $\xi(p)$. $\xi(p)$ diverges at $p = p_c$ as $|p - p_c|^{-\nu}$. For $p > p_c$ finite clusters and the infinite cluster coexist, the number of finite clusters vanishing at p approaches unity. The probability that a site belongs to the infinite cluster varies with occupation probability p as $(p - p_c)^{\beta}$.

The characteristic length $\xi(p)$ for $p > p_c$ is a measure of the largest finite cluster size. Hence, for length scales $\ell > \xi(p)$ the systems looks homogeneous. Thus, the fractal regime is sandwiched between the lattice constant a and the percolation correlation length $\xi(p)$. The remarkable utility of the percolating network becomes clear from these properties: there is a crossover length scale between fractal and homogeneous structural behavior that can be chosen at will by choosing a value for the site or bond occupation probability $p > p_c$. Excitations of structures that map onto percolating networks can change their nature from fractal to continuous as a function of their length scale ℓ, the transition occurring when $\ell_c \sim \xi(p)$. This property makes the percolating network the "fruit fly" of random structures. The flexibility of adjusting the crossover length ℓ_c enables a fit between those properties calculated for a percolating network with the properties of those physical systems that map onto percolating networks.

The fractal dimension, D_f, is defined from the dependence of the "mass" on length scale. For fractals, the number of points or bonds on the infinite cluster within a length scale ℓ (hence, the "mass"), for $\ell \ll \xi(p)$, depends upon ℓ as $M(\ell) \sim \ell^{D_f}$. From above it is straight forward to show $D_f = d - (\beta/\nu)$. For a percolation network [29] in $d = 2$, $D_f = 91/48$. For $d = 3$, $D_f = 2.48 \pm 0.09$. In mean field, $d = 6$, $D_f = 4$.

The spectral dimension, $\tilde{d}_s$, follows from the analysis of diffusion on a fractal network, as first postulated by de Gennes [14]: "An ant parachutes down onto an occupied site of the infinite cluster of a percolating network. At every time unit, the ant makes one attempt to jump to one of its adjacent sites. If that site is occupied, it moves there. If it is empty, the ant stays at its original site. What is the ensemble-averaged square distance that the ant travels in time t?" Gefen et al. [17] found that $\langle r^2(t)\rangle \sim t^{2+\theta}$ where θ depends upon the scaling exponent μ (conductivity exponent), the probability that a site belongs to the infinite cluster scaling exponent β, and the scaling exponent ν for $\xi(p)$: $\theta = (\mu - \beta)/\nu$. Alexander and Orbach [5] defined the spectral (or fracton) dimension as $\tilde{d}_s = 2D_f/(2 + \theta)$. The importance of $\tilde{d}_s$ can be seen from the calculation of the probability of finding a diffusing particle at the starting point at time t, $P_0(t)$, which for compact diffusion is the inverse of the number of visited sites in time t, $V(t) \sim t^{(\tilde{d}_s/2)}$. For scalar elasticity, the vibrational density of states

$$D(\omega) = -(2\omega/\pi)\mathrm{Im}\,\tilde{P}_0(-\omega^2 + i0^+) \sim \omega^{(\tilde{d}_s - 1)}\,,$$

where $\tilde{P}_0(\omega)$ is the Laplace transform of $P_0(t)$. Noting that for homogeneous systems, $D(\omega) \sim \omega^{(d-1)}$, the name spectral dimension for $\tilde{d}_s$ becomes evident. Alexander and Orbach [5] named the vibrational excitations of fractal structures, fractons, in analogy with the vibrational excitations of homogeneous structures, phonons. Hence, they termed $\tilde{d}_s$ the fracton dimension.

Alexander and Orbach noted that for percolation structures in the mean-field limit, $d = 6$, $\tilde{d}_s = 4/3$ precisely. At that time (1982), values for $\tilde{d}_s$ for $d < 6$ appeared close to that value, and they conjectured that $\tilde{d}_s = 4/3$ for $2 \le d \le 6$. We now know that this is only approximate (but remarkably close [24]): $\tilde{d}_s = 1.325 \pm 0.002$ in $d = 2$, and 1.317 ± 0.003 in $d = 3$. Were the conjecture exact, a numerical relationship between structure and transport would be exact [$\theta = (3D_f/2) - 2$], dictated solely by fractal geometry.

For percolating networks, the structure appears fractal for length scales $\ell \le \xi(p)$ and homogeneous for $\ell \ge \xi(p)$. The vibrational excitations in the fractal regime are termed fractons, and in the homogeneous regime phonons. We define the crossover frequency ω_c as that for which the excitation length scale equals $\xi(p)$. Then $\omega_c \sim (p - p_c)^{(\nu D_f/\tilde{d}_s)}$. Continuity leads to a phonon velocity $v(p) \sim (p - p_c)^{\nu}[(D_f/\tilde{d}_s) - 1]$. This then leads to the dispersion relations:

$$[\ell \gg \xi(p), \omega \ll \omega_c\,,\ \text{phonon regime}]\omega \sim v(p)k\,,$$

and

$$[\ell \ll \xi(p), \omega \gg \omega_c\,,\ \text{fracton regime}]\omega \sim k^{(D_f/\tilde{d}_s)}\,.$$

In the latter case, as we shall show in the next Section, vibrational excitations in the fractal regime are localized, so that k should not be thought of as a wave vector but rather the inverse of the localization length $\Lambda(\omega)$.

Nature of Dynamics on Fractals – Localization

The conductance of a d dimensional percolating network of size L, $G(L)$, can be shown to be proportional to L^{β} where the exponent $\beta = (D_f/\tilde{d}_s)(\tilde{d}_s - 2)$. In the Anderson sense [6], localization occurs when $\beta \le 0$ (marginal at $\beta = 0$). The Alexander–Orbach conjecture [5], $\tilde{d}_s \approx 4/3$, leads to β well less than zero for all embedding dimensions d. One can think of this form of localization as geometrical, as opposed to the scattering localization that traditionally is associated with Anderson localization. The localization is strong, the wave function falling off faster than simple exponential.

The wave function has the form [1]: $\langle\Psi(r)\rangle \sim \exp[-\{[r/\Lambda(\omega)]^{d_\phi}\}]$. The exponent d_ϕ is a geometrical exponent describing the fact that an exponential decay along the fractal will be distorted when viewed in real space. In general, $1 \le d_\phi \le d_{\min}$, where $d_{\min}$ is defined by $\ell \sim R^{d_{\min}}$, with ℓ being the shortest path along the network between two points separated by a Pythagorean distance R. Bunde and Roman [11] have found that for percolating networks $d_\phi = 1$. The decay length $\Lambda(\omega)$ is the localization length from Sect. "Introduction", $\Lambda(\omega) \sim \omega^{(-\tilde{d}_s/D_f)}$.

For a random system such as a percolating network, a given realization of $\Psi(r)$ is badly behaved, depending upon the particular choice of origin taken for the position coordinate r. The ensemble average takes the average over all realizations of $\Psi(r)$, and is denoted by $\langle\Psi(r)\rangle$. This has its dangers, as the calculation of physical properties should be performed using a specific realization for $\Psi(r)$, and then ensemble averaged. Calculation of physical properties utilizing an ensemble averaged wave function $\langle\Psi(r)\rangle$ will not yield the same result, and could be very misleading.

Specific realizations for $\Psi(r)$ for two-dimensional bond percolation vibrational networks have been calculated by Nakayama and Yakubo [25,28,47]. The fracton "core" (or largest amplitude) possesses very clear boundaries for the edges of the excitation, with an almost step-like character and a long tail. The tail extends over large distances and the amplitude oscillates in sign. This is required for orthogonality to the uniform translational mode with $\omega = 0$. Clearly these modes do not look like simple exponentials, hence the warning about the use of $\langle\Psi(r)\rangle$ as compared to $\Psi(r)$ for the calculation of matrix elements.

Mapping of Physical Systems onto Fractal Structures

The foregoing description of dynamics of fractal networks would be interesting in and of itself, but its applicability to physical systems makes it of practical importance. Further, the nature of localization, and computation of "hopping-type" transport on fractal structures, has application to materials that are not fractal. Thus, it is worthwhile to consider what systems are fractal, the nature of dynamics on such systems, and the lessons we can learn from them for structures that possess similar characteristics but are not fractal. Specifically, Mott's [22,23] variable range-rate hopping for localized electronic impurity states has direct application to transport of vibrational energy via fracton hopping on fractal networks [2], and we shall argue to thermal transport in glassy materials above the "plateau" temperature [20,26].

The fundamental equations for fractal dynamics spring from the equation for diffusion of a random walker on a network: $\mathrm{d}P_i/\mathrm{d}t = \sum_{j\neq I} w_{ij}(P_j - P_i)$ where P_i is the probability that the ith site is occupied, w_{ij} is the probability per unit time for hopping from site i to site j, and $w_{ij} = w_{ji}$. Introducing new quantities $W_{ij} = w_{ij}$ for $i \neq j$, and $W_{ii} = -\sum_{j\neq I} w_{ij}$, one can rewrite the diffu-

sion equation as $dP_i/dt = \sum_j W_{ij}P_j$ with the defined relation $\sum_j W_{ij} = 0$.

For scalar elasticity, the equation of motion for atomic vibrations is $d^2u_i/dt^2 = \sum_j K_{ij}u_j$, with u_i the atomic displacement at position i, and $K_{ij}(i \neq j)$ is given by $K_{ij} = k_{ij}/m_i$ with k_{ij} the force constant connecting two sites i and j, and m_i the mass of the ith site. The diagonal element W_{ii} is obtained from the balancing of forces at the ith site: $\sum_j K_{ij} = 0$.

The only difference between these two equations is the order of the time derivatives, leading to the equivalence of the eigenvalue problems. We have used this equivalence to extract the density of vibrational states for scalar elasticity, and the definition of the spectral (or fracton) dimension in Sect. "Introduction".

The most direct physical example of a vibrating fractal network is the silica aerogel [13,16,42,44]. Different preparation conditions can change the density of these materials to very small values, indicative of the open porous structure of these materials. Two length scales are present: the fundamental building block of the structure (the silica), and the correlation length of the gel. In between, the clusters possess a fractal structure, while at length scales beyond the correlation length the gel behaves as a homogeneous porous glass. This crossover from fractal to homogeneous is the physical example of our previous discussion for percolating networks when the length scale of excitations passes through the percolation correlation length $\xi(p)$. The thermal, transport, and scattering properties of aerogels have been well studied, and are the classic example of fractal behavior.

There are other examples of structures that map onto the eigenvalue spectrum of the diffusion equation. Ferromagnetic and antiferromagnetic materials can be diluted by non-magnetic impurities. This leads to random site dilution, mapping directly onto the site-diluted percolating network for magnetic interactions. This can be seen from the Heisenberg Hamiltonian for spin systems: $H = 1/2\sum_{i,j} J_{ij}\mathbf{S}_i\mathbf{S}_j$, where $\mathbf{S}_i$ is the vector spin at the ith site, and J_{ij} the isotropic exchange interaction between sites i and j. The equation of motion for the spin operator $S_i^+ = S_i^x + iS_i^x$ becomes for diluted ferromagnets $i\hbar\partial S_i^+/\partial t = \sum_{j\neq i} J_{ij}(S_i^z S_j^+ - S_j^z S_i^+)$. For low levels of spin wave excitations, $S_i^z \approx S_j^z \approx S$, so that one obtains the linear equation of motion $i\hbar\partial S_i^+/\partial t = \sum_{j\neq I} J_{ij}(S_j^+ - S_i^+)$. This is the same form as that for diffusion of a random walker on a network, so that, in analogy with the scalar vibrational network, the eigenvalue spectrum is the same. This mapping allows the spin wave spectrum to exhibit magnons and fracton waves obeying the same properties that phonons and fractons exhibit for scalar vibrations. For antiferromagnets, the change in sign of S_i^z for the two sublattices complicates the linearized equations of motion. They become, $i\hbar\partial S_i^+/\partial t = \sigma_i \sum_{j\neq I} J_{ij}(S_j^+ + S_i^+)$, where $\sigma_i = -1$ for down spins and $\sigma_i = 1$ for up spins. This alters the dynamical equations for diluted antiferromagnets, and requires a separate calculation of the eigenvalue spectrum.

Neutron diffraction experiments exhibit both conventional magnon and fracton excitations. In the case of the site-diluted antiferromagnet $Mn_xZn_{1-x}F_2$ both excitations are observed simultaneously for length scales in the vicinity of $\xi(p)$ [43].

Relaxation Dynamics on Fractal Structures

Electron spin resonance and non-radiative decay experiments can be performed in fractal materials. It is of interest to compare these dynamical properties with those found in translationally ordered structures. At first sight, for vibrational relaxation, it might be thought that the only difference would originate with the differences in the densities of vibrational states $D(\omega)$. However, the localized nature of the vibrational states (fractons) and their random positions in the fractal network, introduces properties that are specific to fractal lattices. There are two cases of interest: one-fracton relaxation, and two-fracton inelastic scattering. The former is referred to as the "direct" relaxation process; the latter the "Raman" relaxation process. In both cases, the main effect of vibrational localization is on the relaxation time profile. Different spatial sites can have very different relaxation rates because of the randomness of their distance from the appropriate fracton sites. The result is a strongly non-exponential time decay.

For one-fracton relaxation [1], a straightforward generalization of the direct relaxation process rate, $W(\omega_0, L)$, for relaxation energy ω_0 arising from interaction with a fracton of the same energy centered a distance L away, is proportional to

$$W(\omega_0, L) \sim \omega_0^{2q-1}\{[\Lambda(\omega_0)]^{(-D_f)}\} \cdot [\coth(\beta\omega_0/2)](1/\delta_L)\exp\{-[L/\Lambda(\omega_0)]^{d_\phi}\}\,.$$

Here, $\beta = 1/k_B T$ and the factor δ_L represents the energy width of the fracton state. In the homogeneous limit, where phonons represent the extended vibrational states, an energy conserving delta function would replace δ_L. There are two limits for δ_L, case (a) when the fracton relaxation rate δ caused by anharmonicity dominates; and case (b) when the electronic relaxation rate caused by the electron-fracton interaction is greater than δ. In the for-

mer case, $\delta_L = \delta$; while in the latter, $\delta_L = W(\omega_0, L)$ itself, leading to a self-consistent determination of the relaxation rate. The latter case requires explicit consideration of the L dependence of δ_L, further complicating the calculation of $W(\omega_0, L)$. The calculation of the time profile of the electronic state population and the average relaxation rate are complex, and the reader is referred to [32] for details.

For case (a), the population of the initial electronic state is found to vary as

$$(\ln t)^{[(D_f - d_\phi)/2d_\phi]} \left[t^{\{-c_1 (\ln t)^{[(D_f/d_\phi)-1]}\}} \right] .$$

Here, c_1 is a constant independent of time, but dependent upon ω_0 and δ. The population of the initial electronic state decays is thus found to be faster than a power law, but slower than exponential or stretched exponential.

For case (b), the population of the initial electronic state is found to vary as

$$(1/t) \left\{ (\ln t)^{[(D_f/d_\phi)-1]} \right\} .$$

The decay is slower than in case (a), and is closer to a power law.

The average relaxation rates for cases (a) and (b) have the same functional form: $\langle W \rangle \sim D(\omega_0)[(\omega_0)^{(2q-1)}] \coth(\beta\omega_0/2)$ where $q = \tilde{d}_s(d_\phi/D_f)$. The temperature dependence is the usual one found for the direct process, but the energy dependence differs, depending on the values of the dimensions and parameters for fractal networks.

Two fracton relaxation is considerably more complex [3,4]. In summary, the time profile of the initial electronic state in the long time regime begins as a stretched exponential, then crosses over into a form similar to case (a) for the one fracton decay process. In the presence of rapid electronic cross relaxation, the time profile is exponential, with a low-temperature relaxation time $\langle 1/T_1 \rangle$ proportional to $T^{\{2\tilde{d}_s[1+2(d_\phi/D_f)]-1\}}$ for Kramers transitions (between half-integer time-reversed spin states), and $T^{\{2\tilde{d}_s[1+2(d_\phi/D_f)]-3\}}$ for non-Kramers transitions (between integer time-reversed spin states).

The complexity of random systems adds an interesting consideration: the nature of the relaxation rate *at a specific site* as compared to the *average* relaxation rate calculated above. The decay profile of single-site spin-lattice relaxation is always exponential. However, the temperature dependence does not simply follow that of the average relaxation rate. Instead, it exhibits irregular statistical fluctuations reflecting the environment of the chosen site. As the temperature is raised, new relaxation channels, involving higher-frequency localized vibrations, will become activated, adding to the relaxation processes responsible for the relaxation at low temperatures. The temperature dependence of the relaxation rate $W(T)$ at different sites will fluctuate with strong correlations in the temperature dependence between them, underlying large accumulative fluctuations.

Using a step function for the vibrational Bose functions, the relaxation rate has a T^2 temperature dependence between randomly spaced steps that occur when new relaxation channels are opened: a "devil's staircase" structure. The full Bose function smoothes out these steps, but the qualitative features remain.

All of these features point to the complexity and richness of electronic relaxation associated with localized vibrational spectra. Their remarkable complexity is worthy of investigation.

Transport on Fractal Structures

The localization of excitations discussed in Sect. "Fractal and Spectral Dimensions" has profound influence on the transport properties of and on fractal structures. For example, in the case of the silica aerogels, the "conventional" form for glasses or amorphous materials for the thermal conductivity $\kappa(T)$ is found [12]: a rapid rise at low temperatures leveling off to a "plateau" as a function of temperature T, and then a rise in $\kappa(T)$ roughly as T^2. At first sight, this is to be "expected" as the aerogels are random, and one could make an analogy with glassy systems. However, there is a fundamental inconsistency in such an analysis. The fracton excitations are localized, and therefore cannot contribute to transport. Thus, one would expect instead a rapid rise in $\kappa(T)$ at low temperatures where the long length scale excitations are phonon-like (and hence delocalized), flattening off to a constant value when all the extended state density of states have reached their duLong–Petit value. The specific heat for these states is then a constant, and $\kappa(T)$ would be a constant independent of temperature T.

So far, this behavior is consistent with experiment. But what happens for temperatures T greater than the plateau temperature region? How can $\kappa(T)$ increase when the only excitations are fractons and therefore localized? The solution to this conundrum lies in the anharmonic forces connecting the silica clusters in the aerogel network. The presence of the anharmonicity allows for a process of fracton hopping [2,20,26], in precisely the same fashion as in Mott's [22,23] variable range-rate hopping for localized electronic impurity states in semiconductors. The fracton can absorb/emit a lower frequency extended state phonon, and *hop* to another fracton location, thereby transmitting excitation energy spatially. This possibility is amusing, for

it means that for fractal networks the anharmonicity *contributes* to thermal transport, whereas in translationally invariant systems the anharmonicity *inhibits* thermal transport. In fact, were there no anharmonicity in a fractal network, the thermal conductivity $\kappa(T)$ would stay at the plateau value for all T greater than the temperature T_p, the onset of the plateau, for all $T > T_p$. In this sense, anharmonicity *stands on its head*. It is essential for transport in the fracton regime, defined by $T \geq T_p \approx \hbar\omega_c/k_B$.

The localized vibrational excitation hopping contribution to the thermal conductivity is $\kappa_{hop}(T) = (k_B/2V)\Sigma_{\lambda'}R^2(\omega_{\lambda'})/\tau_{sl}(\omega_{\lambda'}, T)$. Here sl means strongly localized modes, $R(\omega_{\lambda'})$ is the hopping distance associated with the sl mode, λ', $\tau_{sl}(\omega_{\lambda'}, T)$ the hopping lifetime of the sl mode λ' caused by anharmonic interactions at temperature T, and V the volume of the system. We have already derived the ensemble averaged wave function for fractons, $\langle\Psi(r)\rangle$, in Sect. "Fractal and Spectral Dimensions". We adopt the sl notation [26] because the hopping transport arguments apply to strongly localized states, independent of whether or not they are fractons. The dynamical process for $\kappa_{hot}(T)$ is an sl mode at one site coupling with a low energy extended state to hop to an sl mode at another site, thereby transferring excitation energy spatially.

The evaluation of $\tau_{sl}(\omega_{\lambda'}, T)$ will depend upon the hopping distance for transport of the sl modes, $R(\omega_{\lambda'})$. The argument of Mott [22,23], derived by him for localized electronic states, can be taken over for sl vibrational modes [2]. The volume that contains at least one sl mode is given by $(4\pi/3)D(\omega_{\lambda'})\omega_{sl}[R(\omega_{\lambda'})]^3 = 1$, where $D(\omega_{\lambda'})$ is the density of sl states per unit energy. This condition assures that, for an sl mode at the origin, a second sl mode can be found within a hopping distance $R(\omega_{\lambda'})$. We find $R(\omega_{\lambda'}) \approx (\omega_{\lambda'}/\omega_{sl})\Lambda(\omega_{\lambda'})$. Thus, the most probable hopping distance is *at least* the sl localization length.

The hopping rate $1/\tau_{sl}(\omega_{\lambda'}, T)$ caused by anharmonicity, with anharmonic coupling constant C_{eff} can be found from the Golden Rule, leading to the hopping contribution to the thermal conductivity, with $(\xi_M)^3$ the volume for finding a single sl mode, $\kappa_{hop}(T) = [4^3 3 (C_{eff})^2(k_B)^2 T]/[\pi^3(v_s)^5\rho^3(\xi_M)^3\langle\Lambda(\omega_\lambda)\rangle^2]$. Here $\langle\Lambda(\omega_\lambda)\rangle$ is the average sl length scale, v_s the velocity of sound, and ρ the density. There are very few undetermined parameters in this expression for $\kappa_{hop}(T)$. The anharmonic coupling constant can be found from the shift of the "boson peak" in the Raman spectra with pressure [48,49], typically a factor of ~ 25 times larger than the long length scale third-order elastic coupling constant for extended modes. Making use of experiments on amorphous a-GeS_2 one finds [26,27] $\kappa_{hop}(T) \approx 0.0065T$ (W/mK) using the appropriate values for ρ and v_s, and $\xi_M \approx \langle\Lambda(\omega_\lambda)\rangle = 15$ Å. This contribution is comparable to that observed for other network-forming glasses above the plateau value.

There is a quantum phenomenon associated with sl mode hopping. The hopping vertex is associated with a low energy extended mode coupling with an sl mode to hop to another (spatially separated) sl mode. As the temperature increases, the lifetime of the sl mode will become shorter than $1/\omega$ of the low energy extended mode. This quantum effect leads to a breakdown of the Golden Rule expression, causing a leveling off of the linear-in-T thermal conductivity above the plateau value.

The sl mode hopping contribution to $\kappa(T)$ above the plateau temperature for glasses and amorphous materials appears to provide a quantitative explanation of the observed phenomena. It is a specific example of how dynamics on a fractal network can be used to determine the properties of materials that, while certainly not fractal, have properties that map onto the behavior of fractal structures.

Future Directions

As stated in the Introduction, studying the dynamics on fractal networks is a convenient structure for analyzing the properties of localized states, with all of the frequency and temperature-dependent physical properties determined without arbitrary parameters. For example, the frequency dependence of the localization length scale, $\Lambda(\omega)$, is known precisely on fractal networks, and is used for precise calculation of the hopping transport in Sect. "Relaxation Dynamics on Fractal Structures", *without* adjustable parameters.

If one extrapolates the properties of fractal networks to those random systems that are certainly not fractal, but which exhibit localization, the transport and relaxation properties can be understood. Examples are the thermal transport of glasses and amorphous materials above the plateau temperature, and the lifetime of high energy vibrational states in a-Si. Measurements of these properties may offer opportunities for practical devices in frequency regimes beyond extended state frequencies. For example, the lifetime of high energy vibrational states in a-Si has been shown to *increase* with *increasing vibrational energies* [36,37]. Such behavior is opposite to that expected from extended vibrational states interacting anharmonically. But it is entirely consistent if these states are localized and behave as shown for fractal networks [32,33].

The thesis presented here points to a much broader application of the consequences of fractal geometry than simply of those physical systems which exhibit such structures. The claim is made that using what has been learned from fractal dynamics can point to explication of physical

phenomena in disordered systems that are certainly not fractal, but display properties analogous to fractal structures. In that sense, the future lies in application of the ideas contained in the study of fractal dynamics to that of random systems in general. This very broad class of materials may well have properties of great practical importance, predicted from the dynamics on fractals.

Bibliography

Primary Literature

1. Alexander S, Entin-Wohlman O, Orbach R (1985) Relaxation and nonradiative decay in disordered systems, I. One-fracton emission. Phys Rev B 32:6447–6455
2. Alexander S, Entin-Wohlman O, Orbach R (1986) Phonon-Fracton anharmonic interaction: the thermal conductivity of amorphous materials. Phys Rev B 34:2726–2734
3. Alexander S, Entin-Wohlman O, Orbach R (1986) Relaxation and non-radiative decay in disordered systems, II. Two-fracton inelastic scattering. Phys Rev B 33:3935–3946
4. Alexander S, Entin-Wohlman O, Orbach R (1987) Relaxation and non-radiative decay in disordered systems, III. Statistical character of Raman (two-quanta) spin-lattice relaxation. Phys Rev B 35:1166–1173
5. Alexander S, Orbach R (1982) Density of states on fractals, "fractons". J Phys Lett (Paris) 43:L625–L631
6. Anderson PW (1958) Absence of diffusion in certain random lattices. Phys Rev 109:1492–1505
7. Avnir D (ed) (1989) The Fractal Approach to Heterogeneous Chemistry. Wiley, Chichester
8. Barabaási A-L, Stanley HE (1995) Fractal Concepts in Crystal Growth. Cambridge University Press, Cambridge
9. Bunde A, Havlin S (eds) (1991) Fractals and Disordered Systems. Springer, Berlin
10. Bunde A, Havlin S (eds) (1994) Fractals in Science. Springer-Verlag, Berlin
11. Bunde A, Roman HE (1992) Vibrations and random walks on random fractals: anomalous behavior and multifractality. Philos Magazine B 65:191–211
12. Calemczuk R, de Goer AM, Salce B, Maynard R, Zarembowitch A (1987) Low-temperature properties of silica aerogels. Europhys Lett 3:1205–1211
13. Courtens E, Pelous J, Phalippou J, Vacher R, Woignier T (1987) Brillouin-scattering measurements of phonon-fracton crossover in silica aerogels. Phys Rev Lett 58:128–131
14. de Gennes PG (1976) La percolation: un concept unificateur. Recherche 7:919–927
15. Feder J (1988) Fractals. Plenum Press, New York
16. Freltoft T, Kjems J, Richter D (1987) Density of states in fractal silica smoke-particle aggregates. Phys Rev Lett 59:1212–1215
17. Gefen Y, Aharony A, Alexander S (1983) Anomalous diffusion on percolating clusters. Phys Rev Lett 50:77–80
18. Ikeda H, Fernandez-Baca JA, Nicklow RM, Takahashi M, Iwasa I (1994) Fracton excitations in a diluted Heisenberg antiferromagnet near the percolation threshold: $RbMn_{0.39}Mg_{0.61}F_3$. J Phys: Condens Matter 6:10543–10549
19. Ikeda H, Itoh S, Adams MA, Fernandez-Baca JA (1998) Crossover from Homogeneous to Fractal Excitations in the Near-Percolating Heisenberg Antiferromagnet $RbMn_{0.39}Mg_{0.61}F_3$. J Phys Soc Japan 67:3376–3379
20. Jagannathan A, Orbach R, Entin-Wohlman O (1989) Thermal conductivity of amorphous materials above the plateau. Phys Rev B 39:13465–13477
21. Kistler SS (1932) Coherent Expanded–Aerogels. J Phys Chem 36:52–64
22. Mott NF (1967) Electrons in disordered structures. Adv Phys 16:49–144
23. Mott NF (1969) Conduction in non-crystalline materials III. Localized states in a pseudogap and near extremities of conduction and valence bands. Philos Mag 19:835–852
24. Nakayama T (1992) Dynamics of random fractals: large-scale simulations. Physica A 191:386–393
25. Nakayama T (1995) Elastic vibrations of fractal networks. Japan J Appl Phys 34:2519–2524
26. Nakayama T, Orbach R (1999) Anharmonicity and thermal transport in network glasses. Europhys Lett 47:468–473
27. Nakayama T, Orbach RL (1999) On the increase of thermal conductivity in glasses above the plateau region. Physica B 263–264:261–263
28. Nakayama T, Yakubo K (2001) The forced oscillator method: eigenvalue analysis and computing linear response functions. Phys Rep 349:239–299
29. Nakayama T, Yakubo K (2003) Fractal Concepts in Condensed Matter Physics. Solid-State Sciences. Springer, Berlin
30. Nakayama T, Yakubo K, Orbach R (1994) Dynamical properties of fractal networks: Scaling, numerical simulations, and physical realizations. Rev Mod Phys 66:381–443
31. Orbach R (1986) Dynamics of Fractal Networks. Science 231:814–819
32. Orbach R (1996) Transport and vibrational lifetimes in amorphous structures. Physica B 219–220:231–234
33. Orbach R, Jagannathan A (1994) High energy vibrational lifetimes in a-Si:H. J Phys Chem 98:7411–7413
34. Rammal R (1984) Spectrum of harmonic excitations on fractals. J Phys (Paris) 45:191–206
35. Rammal R, Toulouse G (1983) Random walks on fractal structures and percolation clusters. J Phys Lett (Paris) 44:L13–L22
36. Scholten AJ, Dijkhuis JI (1996) Decay of high-frequency phonons in amorphous silicon. Phys Rev B 53:3837–3840
37. Scholten AJ, Verleg PAWE, Dijkjuis JI, Akimov AV, Meltzer RS, Orbach R (1995) The lifetimes of high-frequency phonons in amorphous silicon: evidence for phonon localization. Solid State Phenom 44–46:289–296
38. Stanley HE (1977) Cluster shapes at the percolation threshold: and effective cluster dimensionality and its connection with critical-point exponents. J Phys A 10:L211–L220
39. Stauffer D, Aharony A (1992) Introduction to Percolation Theory, 2nd edn. Taylor and Francis, London/Philadelphia
40. Stauffer D, Stanley HE (1995) From Newton to Mandelbrot: A primer in Theoretical Physics, 2nd edn. Springer, Berlin
41. Takayasu H (1990) Fractals in the Physical Sciences. Manchester University Press, Manchester
42. Tsujimi Y, Courtens E, Pelous J, Vacher R (1988) Raman-scattering measurements of acoustic superlocalization in silica aerogels. Phys Rev Lett 60:2757–2760
43. Uemura YJ, Birgeneau RJ (1987) Magnons and fractons in the diluted antiferromagnet $Mn_xZn_{1-x}F_2$. Phys Rev B 36:7024–7035
44. Vacher R, Courtens E, Coddens G, Pelous J, Woignier T (1989) Neutron-spectroscopy measurement of a fracton density of states. Phys Rev B 39:7384–7387

45. Vacher R, Woignier T, Pelous J, Courtens E (1988) Structure and self-similarity of silica aerogels. Phys Rev B 37:6500–6503
46. Vicsek T (1993) Fractal Growth Phenomena. World Scientific, Singapore
47. Yakubo K, Nakayama T (1989) Direct observation of localized fractons excited on percolating nets. J Phys Soc Japan 58:1504–1507
48. Yamaguchi M, Nakayama T, Yagi T (1999) Effects of high pressure on the Bose peak in a-GeS_2 studied by light scattering. Physica B 263–264:258–260
49. Yamaguchi M, Yagi T (1999) Anharmonicity of low-frequency vibrations in a-GeS_2 studied by light scattering. Europhys Lett 47:462–467

Recommended Reading

While many of the original articles are somewhat difficult to access, the reader is nevertheless recommended to read them as there are subtle issues that tend to disappear as references are quoted and re-quoted over time. For a very helpful introduction to percolation theory, Ref. 39 is highly recommended. A thorough review of dynamics of/on fractal structures can be found in Ref. 30. Finally, a comprehensive treatment of localization and multifractals, and their relevance to dynamics of/on fractal networks, can be found in Ref. 29.

Dynamics of Hamiltonian Systems

Heinz Hanssmann
Mathematisch Instituut, Universiteit Utrecht,
Utrecht, The Netherlands

Article Outline

Glossary
Definition of the Subject
Introduction
The Phase Space
Systems in One Degree of Freedom
Systems in Two Degrees of Freedom
Symmetry Reduction
Integrable Systems
Perturbation Analysis
Future Directions
Acknowledgments
Bibliography

Glossary

Action angle variables In an *integrable* Hamiltonian system with n degrees of freedom the level sets of regular values of the n integrals are *Lagrangean submanifolds*. In case these level sets are compact they are (unions of) n-tori. In the neighborhood of such an invariant torus one can find a *Darboux chart* with co-ordinates (x, y) such that the integrals (and in particular the Hamiltonian) depend only on the actions y. The values of (x, y) form a product set $\mathbb{T}^n \times \mathbb{Y}$, with $\mathbb{Y} \subseteq \mathbb{R}^n$ open. Every invariant n-torus $y = \text{const}$ in the domain of this chart is parametrized by the angles x.

Canonical 1-form The symplectic form (or canonical 2-form) is closed, satisfying $\mathrm{d}\omega = 0$. If it is furthermore exact, i. e. of the form $\omega = \mathrm{d}\theta$, then $\vartheta = -\theta$ is called a canonical 1-form (unique up to addition of $\mathrm{d}f$, $f \in C^\infty(\mathcal{P})$).

Canonical transformation An invertible transformation that preserves the Poisson bracket, turning canonical co-ordinates into canonical co-ordinates. In the symplectic case also called symplectomorphism.

Cantor set, Cantor dust, Cantor family, Cantor stratification Cantor dust is a separable locally compact space that is perfect, i. e. every point is in the closure of its complement, and totally disconnected. This determines Cantor dust up to homeomorphy. The term Cantor set (originally reserved for the specific form of Cantor dust obtained by repeatedly deleting "the middle third") designates topological spaces that locally have the structure $\mathbb{R}^n \times$ *Cantor dust* for some $n \in \mathbb{N}$. Cantor families are parametrized by such Cantor sets. On the real line $\mathbb{R}$ one can define Cantor dust of positive measure by excluding around each rational number p/q an interval of size $2\gamma/q^\tau$, $\gamma > 0, \tau > 2$. Similar *Diophantine conditions* define Cantor sets in $\mathbb{R}^n$. Since these Cantor sets have positive measure their (Hausdorff)-dimension is n. Where the unperturbed system is stratified according to the co-dimension of occurring (bifurcating) tori, this leads to a Cantor stratification.

Casimir function A function $f\colon \mathcal{P} \longrightarrow \mathbb{R}$ is a Casimir element of the *Poisson algebra* $C^\infty(\mathcal{P})$ if $\{f, g\} = 0$ for all $g \in C^\infty(\mathcal{P})$. This induces a Poisson structure on the level sets $f^{-1}(a)$, $a \in \mathbb{R}$. In case every point has a small neighborhood on which the only Casimir elements are the constant functions, the Poisson structure on $\mathcal{P}$ is non-degenerate, i. e. $\mathcal{P}$ is a *symplectic manifold*.

Center An equilibrium of a vector field is called a center if all eigenvalues of the linearization are nonzero and on the imaginary axis.

Center-saddle bifurcation Under variation of a parameter λ a *center* and a *saddle* meet and vanish.

Chern class Let $\rho\colon \mathcal{R} \longrightarrow C$ be a torus bundle with fiber $\mathbb{T}^n$ and denote by $\mathcal{G}$ the locally constant sheaf of first homotopy groups $\pi_1(\mathbb{T}^n)$ of the fibers. The Chern class of the torus bundle is an element of $H^2(C, \mathcal{G})$ that measures the obstruction to the existence of a global section of ρ – such a section exists if and only if

the Chern class vanishes. An example with a non-vanishing Chern class is the Hopf fibration $S^3 \longrightarrow S^2$.

Conditionally periodic A motion $t \mapsto \alpha(t) \in \mathcal{P}$ is conditionally periodic if there are frequencies $\varpi_1, \ldots, \varpi_k \in \mathbb{R}$ and a smooth embedding $F\colon \mathbb{T}^k \longrightarrow \mathcal{P}$ such that $\alpha(t) = F(\mathrm{e}^{2\pi \mathrm{i} \varpi_1 t}, \ldots, \mathrm{e}^{2\pi \mathrm{i} \varpi_k t})$. One can think of the motion as a superposition of the periodic motions $t \mapsto F(1, \ldots, 1, \mathrm{e}^{2\pi \mathrm{i} \varpi_j t}, 1, \ldots, 1)$. If the frequencies are rationally independent, the motion $t \mapsto \alpha(t) \in \mathcal{P}$ lies dense in im F and this embedded torus is invariant. In case there are *resonances* among the frequencies the motion is restricted to a subtorus. A flow on a torus is parallel or conditionally periodic if there exist co-ordinates in which the vector field becomes constant. In the absence of resonances the flow is called *quasi-periodic*.

Conjugate co-ordinates, canonical co-ordinates Two co-ordinates Q and P of a *Darboux* chart are (canonically) conjugate if $\omega(X_Q, X_P) = \pm 1$, i. e. X_Q and X_P span for every point x in the domain U of the chart a hyperbolic plane in $T_x U$.

Darboux chart, Darboux co-ordinates, Darboux basis In a Darboux basis $\{e_1, \ldots, e_n, f_1, \ldots, f_n\}$ of a symplectic vector space (V, ω) the symplectic product takes the simple form $\omega(e_i, e_j) = 0 = \omega(f_i, f_j)$ and $\omega(e_i, f_j) = \delta_{ij}$. In Darboux co-ordinates $(q_1, \ldots, q_n, p_1, \ldots, p_n)$ of a symplectic manifold $(\mathcal{P}, \omega)$ the symplectic form becomes $\omega = \sum \mathrm{d}q_i \wedge \mathrm{d}p_i$.

Degree of freedom In simple mechanical systems the *phase space* is the cotangent bundle of the configuration space and the dimension of the latter encodes "in how many directions the system can move". For *symplectic manifolds* this notion is immediately generalized to one half of the dimension of the phase space. *Poisson spaces* are foliated by their symplectic leaves and the number of degrees of freedom is defined to be one half of the rank of the Poisson structure.

Diophantine condition, Diophantine frequency vector A frequency vector $\varpi \in \mathbb{R}^n$ is called Diophantine if there are constants $\gamma > 0$ and $\tau > n - 1$ with

$$|\langle k \mid \varpi \rangle| \geq \frac{\gamma}{|k|^\tau} \quad \text{for all} \quad k \in \mathbb{Z}^{n \setminus \{0\}} .$$

The Diophantine frequency vectors satisfying this condition for fixed γ and τ form a Cantor set of half lines. As the "Diophantine parameter" γ tends to zero (while τ remains fixed), these half lines extend to the origin. The complement in any compact set of frequency vectors satisfying a Diophantine condition with fixed τ has a measure of order $O(\gamma)$ as $\gamma \to 0$.

Elliptic A periodic orbit/invariant torus is elliptic if all *Floquet multipliers/exponents* are on the unit circle/imaginary axis. An elliptic equilibrium is called a *center*.

Energy shell The set of points that can be attained given a certain energy. If the phase space is a symplectic manifold, this set is given by the pre-image of that energy value under the Hamiltonian. For more general Poisson spaces this pre-image has to be intersected with a symplectic leaf.

Focus-focus type The (normal) linear behavior generated by a quartet of eigenvalues/Floquet multipliers/exponents.

Frequency-halving bifurcation In the *supercritical* case an invariant torus loses its stability, giving rise to a stable torus with one of its frequencies halved. In the *dual* or *subcritical* case stability is lost through collision with an unstable invariant torus with one of its frequencies halved. The periodic counterpart is called a period-doubling bifurcation.

Generating function A function $S = S(q, y)$ that defines a canonical transformation $(x, y) \mapsto (q, p)$ by means of $x = \partial S/(\partial y)$ and $p = \partial S/(\partial q)$. The generating function always depends on both the old and the new variables, working with dependence on e. g. x instead of y introduces a minus sign.

Group action A mapping $\Gamma\colon G \times \mathcal{P} \longrightarrow \mathcal{P}$ is a group action of the Lie group G on the phase space $\mathcal{P}$ if $\Gamma_e = \mathrm{id}$ and $\Gamma_{g \cdot h} = \Gamma_g \circ \Gamma_h$ for all Lie group elements g and h. For $G = \mathbb{R}$ one can recover the generating vector field by means of $X = \mathrm{d}/(\mathrm{d}t)\Gamma_t$.

Hamiltonian Hopf bifurcation In a two-degree-of-freedom system a center loses its stability: the two pairs of purely imaginary eigenvalues meet in a *Krein collision*. In a three-degree-of-freedom system the same can happen to two pairs $\mathrm{e}^{\pm\lambda}, \mathrm{e}^{\pm\mu}$ of *Floquet multipliers* in a one-parameter family of periodic orbits – parametrized by the energy. For a quasi-periodic Hamiltonian Hopf bifurcation one needs at least four degrees of freedom.

Hamiltonian system Newton's second law states $F = m\ddot{q}$. Suppose that F is a conservative force, with potential V. Write the equations of motion as a system of first order differential equations

$$\dot{q} = \tfrac{1}{m}\, p$$
$$\dot{p} = -\frac{\partial V}{\partial q}$$

that has the total energy $H(q, p) = \langle p \mid p \rangle/(2m) + V(q)$ as a first integral. This can be generalized. Given

a Hamiltonian function $H(q,p)$ one has the Hamiltonian vector field

$$\dot{q} = \frac{\partial H}{\partial p}$$
$$\dot{p} = -\frac{\partial H}{\partial q}$$

with first integral H. Moreover, one can replace $\mathbb{R}^{2n}$ by a *symplectic manifold* or by a *Poisson space*.

Hilbert basis The (smooth) invariants of an action of a compact group are all given as functions of finitely many "basic" invariants.

Hypo-elliptic An equilibrium is called hypo-elliptic if its linearization has both elliptic and hyperbolic eigenvalues, all nonzero. Similar for periodic orbits and invariant tori.

Integrable system A *Hamiltonian system* with n degrees of freedom is (Liouville)-integrable if it has n functionally independent commuting integrals of motion. Locally this implies the existence of a (local) torus action.

Integral (of motion) A conserved quantity.

Iso-energetic Poincaré mapping For a Hamiltonian system a *Poincaré mapping* leaves the *energy shells* invariant. Restricting to the intersection $\Sigma \cap \{H = h\}$ of the *Poincaré section* with an energy shell one obtains an iso-energetic Poincaré mapping.

Isotropic For a subspace $U < V$ of a symplectic vector space (V,ω) are equivalent: (*i*) U is contained in its ω-orthogonal complement, (*ii*) every basis of U can be extended to a *Darboux basis* of V. If U satisfies one and thus both of these conditions it is called an isotropic subspace. A submanifold $\mathcal{B} \subseteq \mathcal{P}$ of a symplectic manifold $(\mathcal{P},\omega)$ is called isotropic if all tangent spaces $T_x\mathcal{B} < T_x\mathcal{P}$ are isotropic subspaces.

Krein collision Two pairs of purely imaginary *Floquet exponents* meet in a double pair on the imaginary axis and split off to form a complex quartet $\pm\Re \pm \mathrm{i}\Im$. This is a consequence of a transversality condition on the linear terms at $1: -1$ *resonance*; an additional non-degeneracy condition on the non-linear part ensures that a ((quasi)-periodic) *Hamiltonian Hopf bifurcation* takes place.

Lagrangean submanifold For a subspace $U < V$ of a symplectic vector space (V,ω) are equivalent: (*i*) U is a maximal isotropic subspace, (*ii*) V has a *Darboux basis* $\{e_1,\ldots,e_n,f_1,\ldots,f_n\}$ with $\mathrm{span}\{e_1,\ldots,e_n\} = U$. If U satisfies one and thus both of these conditions it is called a Lagrangean subspace. A submanifold $\mathcal{B} \subseteq \mathcal{P}$ of a symplectic manifold $(\mathcal{P},\omega)$ is called Lagrangean if all tangent spaces $T_x\mathcal{B} < T_x\mathcal{P}$ are Lagrangean subspaces.

Lie–Poisson structure Let $\mathfrak{g}$ be a Lie algebra with structure constants Γ_{ij}^k. Then $\{\mu_i,\mu_j\} = \pm\sum \Gamma_{ij}^k \mu_k$ defines two *Poisson structures* on the dual space $\mathfrak{g}^*$.

Local bifurcation Bifurcations of equilibria can be studied within a small neighborhood in the product of phase space and parameter space. The same is true for fixed points of a discrete dynamical system, but when suspended to a flow the corresponding bifurcating periodic orbits obtain a *semi-local* character.

Momentum mapping Let $\Gamma: G \times \mathcal{P} \longrightarrow \mathcal{P}$ be a symplectic action of the Lie group G on the *symplectic manifold* $\mathcal{P}$. A mapping $J: \mathcal{P} \longrightarrow \mathfrak{g}^*$ into the dual space of the Lie algebra of G is a momentum mapping for the action if $X_{J^\xi} = \xi_{\mathcal{P}}$ for all $\xi \in \mathfrak{g}$. Here $J^\xi: \mathcal{P} \longrightarrow \mathbb{R}$ is defined by $J^\xi(z) = J(z)\cdot\xi$ and $\xi_{\mathcal{P}}$ is the infinitesimal generator of the action corresponding to ξ. The momentum mapping J is called Ad^*-equivariant provided that $J \circ \Gamma_g = \mathrm{Ad}^*_{g^{-1}} \circ J$.

Monodromy The group homomorphism $\mathcal{M}: \pi_1(C) \longrightarrow SL_n(\mathbb{Z})$ that measures how much the n-torus bundle $\mathcal{R} \longrightarrow C$ deviates from a *principal* torus bundle.

Non-degenerate integrable Hamiltonian system, function In *action angle variables* (x,y) the Hamiltonian H only depends on the action variables y and the equations of motion become $\dot{x} = \varpi(y)$, $\dot{y} = 0$; with frequencies $\varpi(y) = \partial H/\partial y$. The system is non-degenerate at the invariant torus $\{y = y_0\}$ if $D^2H(y_0)$ is invertible. In this case ϖ defines near $\{y = y_0\}$ an isomorphism between the actions and the angular velocities. Other conditions ensuring that most tori have *Diophantine* frequency vectors are iso-energetic non-degeneracy or Rüssmann-like conditions on higher derivatives.

Normal frequency Given an elliptic invariant torus of a Hamiltonian system, one can define the normal linearization on the *symplectic normal* bundle. The eigenvalues of the normal linearization being $\pm\mathrm{i}\Omega_1,\ldots,\pm\mathrm{i}\Omega_m$, one calls the Ω_j the normal frequencies. Under the exponential mapping the eigenvalues of the normal linearization of a periodic orbit are mapped to *Floquet multipliers*.

Normal linearization The linearization within the normal bundle of an invariant submanifold. In the Hamiltonian context these are often *isotropic* and one further restricts to the symplectic normal linear behavior, e. g. to identify *elliptic* and *hyperbolic* invariant tori.

Normal modes A family of periodic orbits parametrized by the energy value h, as $h \to h_0$ shrinking down to an equilibrium that has a pair of eigenvalues $\pm 2\pi\mathrm{i}\lim_{h\to h_0} T_h^{-1}$, where T_h denotes the period(s).

Parabolic An equilibrium of a one-degree-of-freedom system is called parabolic if its linearization is nilpotent but nonzero. An invariant torus is parabolic if its symplectic *normal linearization* has a parabolic equilibrium. In particular the four *Floquet multipliers* of a parabolic periodic orbit in two degrees of freedom are all equal to 1.

Phase space By Newton's second law the equations of motion are second order differential equations. The trajectory is completely determined by the initial positions and the initial velocities, or, equivalently, the initial momenta. The phase space is the set of all possible combinations of initial positions and initial momenta.

Pinched torus The compact *(un)stable manifold* of a saddle in two degrees of freedom with a quartet $\pm\Re \pm \mathrm{i}\Im$ of hyperbolic eigenvalues resembles a torus $\mathbb{T}^2 = S^1 \times S^1$ with one of the fibers $\{x\} \times S^1$ reduced to a point.

Poisson space, Poisson bracket, Poisson structure A Poisson algebra $\mathcal{A}$ is a real Lie algebra that is also a commutative ring with unit. These two structures are related by Leibniz' rule $\{f \cdot g, h\} = f \cdot \{g, h\} + g \cdot \{f, h\}$. A Poisson manifold $\mathcal{P}$ has a Poisson bracket on $C^\infty(\mathcal{P})$ that makes $C^\infty(\mathcal{P})$ a Poisson algebra. If there are locally no *Casimir elements* other than constant functions this leads to a *symplectic* structure on $\mathcal{P}$. Poisson spaces naturally arise in *singular reduction*, this motivates to allow varieties $\mathcal{P}$ where the Poisson bracket is defined on a suitable subalgebra $\mathcal{A}$ of $C(\mathcal{P})$.

Given a Hamiltonian function $H \in \mathcal{A}$ one obtains for $f \in \mathcal{A}$ the equations of motion $(\mathrm{d}/(\mathrm{d}t))f = \{f, H\}$. For canonically *conjugate* co-ordinates (q, p) on $\mathcal{P}$, i. e. with $\{q_i, q_j\} = 0 = \{p_i, p_j\}$ and $\{q_i, p_j\} = \delta_{ij}$, this amounts to

$$\dot{q} = \frac{\partial H}{\partial p}$$
$$\dot{p} = -\frac{\partial H}{\partial q}\,.$$

Poisson symmetry A symmetry that preserves the *Poisson structure*.

Principal fiber bundle A fiber bundle $F\colon \mathcal{P} \longrightarrow C$ admitting a free Lie group action $G \times \mathcal{P} \longrightarrow \mathcal{P}$ that acts transitively on the fibers $F^{-1}(c) \cong G$.

Proper degeneracy For the application of the KAM theorem to a perturbation of an *integrable system* it is necessary that the integrable system is *non-degenerate*, so that the frequencies have maximal rank as function of the actions. If there are global conditions relating the frequencies, so that the *conditionally periodic* motion can be described by a smaller number of these, the system is properly degenerate. *Superintegrable systems* are a particular example.

Quasi-periodic A *conditionally periodic* motion that is not periodic. The closure of the trajectory is an invariant k-torus with $k \geq 2$. A parallel or conditionally periodic flow on a k-torus is called quasi-periodic if the frequencies $\varpi_1, \ldots, \varpi_k$ are rationally independent.

Ramified torus bundle Let a differentiable mapping $f\colon \mathcal{P} \longrightarrow \mathbb{R}^m$ be given. According to Sard's Lemma almost all values $a \in \mathbb{R}^m$ are regular. The connected components of the sets $f^{-1}(a)$, $a \in \operatorname{im} f$ regular, define a foliation of an open subset $\mathcal{R}$ of the $(n+m)$-dimensional manifold $\mathcal{P}$. In the present settings the components of f are the first integrals of an *integrable* Hamiltonian system with compact level sets. Then the regular fibers are n-tori, and their union is a torus bundle. The topology of this bundle depends on the topology of the base space C, the *monodromy* and the *Chern class* of the bundle. In many examples one encounters the connected components of C are contractible, whence monodromy and Chern class are trivial. For the geometry of the bundle one also wants to know how the singular fibers are distributed: where n-tori shrink to normally elliptic $(n-1)$-tori and where they are separated by stable and unstable manifolds of normally hyperbolic $(n-1)$-tori. The singular fibers are not necessarily manifolds, but may be stratified into X_H-invariant strata which are possibly non-compact. One can continue and look for the singularities of these 'regular singular leaves'; the tori of dimension $\leq n-2$ and the normally *parabolic* $(n-1)$-tori in which normally elliptic and normally hyperbolic $(n-1)$-tori meet in a quasi-periodic *center-saddle bifurcation* or a *frequency-halving bifurcation*. The next "layer" is given by quasi-periodic *Hamiltonian Hopf bifurcations* and bifurcations of higher co-dimension.

Reduced phase space Let $\Gamma\colon G \times \mathcal{P} \longrightarrow \mathcal{P}$ be a symplectic action of the Lie group G on the *symplectic manifold* $\mathcal{P}$ with Ad^*-equivariant *momentum mapping* $J\colon \mathcal{P} \longrightarrow \mathfrak{g}^*$. For a regular value $\mu \in \mathfrak{g}^*$ let the action of the isotropy group $G_\mu = \{g \in G \mid \mathrm{Ad}^*_{g^{-1}}(\mu) = \mu\}$ on $J^{-1}(\mu)$ be free and proper. Then the quotient $J^{-1}(\mu)/_{G_\mu}$ is again a symplectic manifold, the reduced phase space. A Γ-invariant Hamiltonian function H on $\mathcal{P}$ leads to a reduced Hamiltonian function H_μ on the reduced phase space.

Relative equilibrium Let H be a Hamiltonian function invariant under the *symplectic* action $G \times \mathcal{P} \longrightarrow \mathcal{P}$ and let $\mu \in \mathfrak{g}^*$ be a regular value of the Ad^*-

equivariant *momentum mapping* $J\colon \mathcal{P} \longrightarrow \mathfrak{g}^*$. Also assume that the isotropy group G_μ under the Ad^* action on $\mathfrak{g}^*$ acts freely and properly on $J^{-1}(\mu)$. Then X_H induces a Hamiltonian flow on the *reduced phase space* $\mathcal{P}_\mu = J^{-1}(\mu)/_{G_\mu}$. The phase curves of the given Hamiltonian system on $\mathcal{P}$ with momentum constant $J = \mu$ that are taken by the projection $J^{-1}(\mu) \longrightarrow \mathcal{P}_\mu$ into equilibrium positions of the reduced Hamiltonian system are called relative equilibria or stationary motions (of the original system).

Remove the degeneracy A perturbation of a *superintegrable system* removes the degeneracy if it is sufficiently mild to define an intermediate system that is still integrable and sufficiently wild to make that intermediate system *non-degenerate*.

Resonance If the frequencies of an invariant torus with *conditionally periodic* flow are rationally dependent this torus divides into invariant subtori. Such resonances $\langle h \mid \varpi \rangle = 0, h \in \mathbb{Z}^k$, define hyperplanes in ϖ-space and, by means of the frequency mapping, also in phase space. The smallest number $|h| = |h_1| + \cdots + |h_k|$ is the order of the resonance. *Diophantine conditions* describe a measure-theoretically large complement of a neighborhood of the (dense!) set of all resonances.

Saddle An equilibrium of a vector field is called a saddle if the linearization has no eigenvalues on the imaginary axis. On a small neighborhood of a saddle the flow is topologically *conjugate* to its linearization.

Semi-local bifurcation Bifurcations of n-tori can be studied in a tubular neighborhood. For $n = 1$ a *Poincaré section* turns the periodic orbit into a fixed point of the *Poincaré mapping* and the bifurcation obtains a *local* character.

Simple mechanical system A quintuple $(T^*M, \langle .. \mid .. \rangle, V, T, \omega)$ consisting of the co-tangent bundle of a Riemannian manifold $(M, \langle .. \mid .. \rangle)$, a potential function $V\colon M \longrightarrow \mathbb{R}$, the kinetic energy $T(\alpha) = \langle \alpha \mid \alpha \rangle$ and the symplectic form $\omega = -\mathrm{d}\vartheta$ derived from the canonical 1-form on the co-tangent bundle T^*M.

Singular reduction If $\Gamma\colon G \times \mathcal{P} \longrightarrow \mathcal{P}$ is a *Poisson symmetry*, then the group action on $\mathcal{P}$ makes $\mathcal{B} = \mathcal{P}/_G$ a *Poisson space* as well. Fixing the values of the resulting *Casimir functions* yields the *reduced phase space*, which turns out to have singular points where the action Γ is not free.

Solenoid Given a sequence $f_j\colon S^1 \longrightarrow S^1$ of coverings $f_j(\zeta) = \zeta^{\alpha_j}$ of the circle S^1 the solenoid $\Sigma_a \subseteq (S^1)^{\mathbb{N}_0}$, $a = (\alpha_j)_{j\in\mathbb{N}_0}$ consists of all $z = (\zeta_j)_{j\in\mathbb{N}_0}$ with $\zeta_j = f_j(\zeta_{j+1})$ for all $j \in \mathbb{N}_0$.

Stratification The decomposition of a topological space into smaller pieces satisfying certain boundary conditions.

Superintegrable system A Hamiltonian system with n degrees of freedom is superintegrable if it has $n+1$ functionally independent integrals of motion such that each of the first $n-1$ of them commutes with all $n+1$. Such a *properly degenerate* system admits generalized action angle co-ordinates $(x_1, \ldots, x_{n-1}, y_1, \ldots, y_{n-1}, q, p)$. In case the non-degeneracy condition $\det D^2H(y) \neq 0$ is satisfied almost everywhere the system is "minimally superintegrable". In the other extreme of a "maximally superintegrable" system all motions are periodic.

Symplectic manifold A $2n$-dimensional manifold $\mathcal{P}$ with a non-degenerate closed two-form ω, i.e. $\mathrm{d}\omega = 0$ and $\omega(u, v) = 0$ for all $v \in T\mathcal{P} \Rightarrow u = 0$. A diffeomorphism ψ of symplectic manifolds that respects the two-form(s) is called a symplectomorphism. Given a Hamiltonian function $H \in C^\infty(\mathcal{P})$ one obtains through $\omega(X_H, ..) = \mathrm{d}H$ the Hamiltonian vector field X_H. For every $x \in \mathcal{P}$ there are co-ordinates (q, p) around x with $\omega = \mathrm{d}q \wedge \mathrm{d}p$. In these *Darboux co-ordinates* X_H reads

$$\dot{q} = \frac{\partial H}{\partial p}$$
$$\dot{p} = -\frac{\partial H}{\partial q}\,.$$

Symplectic form A non-degenerate closed two-form.

Syzygy A constraining equation that is identically fulfilled by the elements of a *Hilbert basis*.

(Un)stable manifold In Hamiltonian systems with one degree of freedom the *stable manifold* and the unstable manifold of an equilibrium often coincide and thus consist of *homoclinic orbits*. In such a case it is called an (un)stable manifold. This carries over to the stable and the unstable manifold of a periodic orbit or an invariant torus in higher degrees of freedom if the system is *integrable*.

Definition of the Subject

Hamiltonian systems are a class of dynamical systems which can be characterized by preservation of a symplectic form. This allows to write down the equations of motion in terms of a single function, the Hamiltonian function. They were conceived in the 19th century to study physical systems varying from optics to frictionless mechanics in a unified way. This description turned out to be particularly efficient for symmetry reduction and perturbation analysis.

Introduction

The best-known Hamiltonian system is the harmonic oscillator. The second order differential equation

$$\ddot{x} + \varpi^2 x = 0$$

models the motion of a point mass attached to a massless spring (Hooke's law) and has the general solution

$$x(t) = x_0 \cos \varpi t + \frac{y_0}{\varpi} \sin \varpi t$$

with initial conditions $(x(0), \dot{x}(0)) = (x_0, y_0)$. Choosing co-ordinates

$$q = \sqrt{\varpi}\, x \quad \text{and} \quad p = \frac{y}{\sqrt{\varpi}} \tag{1}$$

yields the first order system

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} 0 & \varpi \\ -\varpi & 0 \end{pmatrix}\begin{pmatrix} q \\ p \end{pmatrix},$$

the solutions of which move along the circles $p^2 + q^2 =$ const. This last observation lies at the basis of the Hamiltonian formalism, defining

$$H(q, p) = \varpi\, \frac{p^2 + q^2}{2}$$

to be the Hamiltonian function of the system. The Hamiltonian vector field is then defined to be perpendicular to the gradient ∇H, ensuring that H is a conserved quantity. Passing back through (1) yields the energy

$$H(x, y) = \frac{\dot{x}^2}{2} + \varpi^2\, \frac{x^2}{2}$$

of the spring system (the point mass being scaled to $m = 1$).

The conserved quantity H makes two-dimensional systems, said to have one *degree of freedom*, easy to study and many efforts are made to reduce more complex systems to this setting, often using symmetries. In such a case the system is *integrable* and can in principle be explicitly solved. Where the symmetries are only approximately preserving the system, a more geometric understanding allows to analyze how the perturbation from the integrable approximation to the original system alters the dynamics.

The Phase Space

The simplest type of example of a Hamiltonian system is that of a 1-dimensional particle with kinetic energy

$$T = \frac{1}{2}\, mv^2 = \frac{p^2}{2m} \tag{2}$$

and potential energy $V = V(q)$. The canonical equations derived from the Hamiltonian function $H = T + V$ are

$$\dot{q} = \frac{\partial H}{\partial p} = \frac{p}{m}$$
$$\dot{p} = -\frac{\partial H}{\partial q} = -V'(q)$$

and are equivalent with Newton's laws, where $F = -V'$ is the force field with potential V. An immediate consequence is

$$\dot{H} = \frac{\partial H}{\partial q}\dot{q} + \frac{\partial H}{\partial p}\dot{p} = 0\,,$$

the conservation of energy.

The same conclusion holds in $\mathbb{R}^{2n}$, with *canonical co-ordinates* $q_1, \ldots, q_n$, $p_1, \ldots, p_n$, where a Hamiltonian function $H\colon \mathbb{R}^{2n} \longrightarrow \mathbb{R}$ defines the canonical equations

$$\dot{q}_i = \frac{\partial H}{\partial p_i}\,, \quad i = 1, \ldots, n \tag{3a}$$

$$\dot{p}_i = -\frac{\partial H}{\partial q_i}\,, \quad i = 1, \ldots, n\,. \tag{3b}$$

The Hamiltonian vector field $\mathrm{X}H$ defined by these equations satisfies

$$i_{\mathrm{X}H} \lrcorner\, \omega := \omega(\mathrm{X}H, _\,) = \mathrm{d}H \tag{4}$$

where

$$\omega = \mathrm{d}q_1 \wedge \mathrm{d}p_1 + \cdots + \mathrm{d}q_n \wedge \mathrm{d}p_n \tag{5}$$

is the canonical *symplectic form* on $\mathbb{R}^{2n}$.

Hamiltonian mechanics becomes conceptually easier if one abstracts from well-chosen canonical co-ordinates and considers the phase space to be a *symplectic manifold* $(\mathcal{P}, \omega)$. A Hamiltonian function $H\colon \mathcal{P} \longrightarrow \mathbb{R}$ then defines by means of (4) the Hamiltonian vector field $\mathrm{X}H$ on $\mathcal{P}$ in a co-ordinate free way. With Darboux's Theorem [1,2,22,26] one can always return to canonical co-ordinates, locally around any given point of $\mathcal{P}$.

The flow $\varphi\colon \mathbb{R} \times \mathcal{P} \longrightarrow \mathcal{P}$ of a Hamiltonian vector field preserves the symplectic structure, i. e. $\varphi_t^* \omega = \omega$ [2,8,22,26]. An immediate consequence is Liouville's Theorem that the phase space volume $\omega^n = \omega \wedge \cdots \wedge \omega$ is preserved as well. As a corollary one obtains Poincaré's Recurrence Theorem.

Theorem (Poincaré) *Let $\Omega \subseteq \mathcal{P}$ be compact and invariant under φ_t. Then every neighborhood U of every point $a \in \Omega$ has a trajectory that returns to U.*

The proof consists of "a walk in the park" – however small my shoes are, I am bound to step on my own trail if I walk forever in a park of finite size.

Hamiltonian systems have special properties that are untypical for general dissipative systems. An important aspect is that volume preservation excludes the existence of attractors. In dissipative systems the restriction of the flow to an attractor often allows to dramatically lower the dimension of the system; more generally one restricts with the same aim to the non-wandering set. In the present conservative context, if e. g. the energy level sets are compact then the non-wandering set consists of the whole phase space.

One speaks of a *simple mechanical system* [1,2,22] if the phase space $\mathcal{P} = T^*M$ is the cotangent bundle of a Riemannian manifold M, the configuration space, and the Hamiltonian $H = T + V$ is the sum of kinetic and potential energy. Furthermore, the kinetic energy

$$T(\alpha_q) = \langle \alpha_q \mid \alpha_q \rangle_q$$

is given by the Riemannian metric (evaluated at the base point $q \in M$ of $\alpha_q \in T_q^*M$) and the potential energy

$$V(\alpha_q) = V(q)$$

depends only on the configuration $q \in M$. The cotangent bundle has a *canonical* 1-*form* ϑ defined by

$$\vartheta(v_{\alpha_q}) = \alpha_q(T\pi(v_{\alpha_q})) \quad \text{for all} \quad v_{\alpha_q} \in T_{\alpha_q}T^*M$$

where $\pi : T^*M \longrightarrow M$ is the bundle projection to the configuration space and $T\pi : TT^*M \longrightarrow TM$ its derivative, see Fig. 1. From ϑ the symplectic form is obtained as the exterior derivative $\omega = -\mathrm{d}\vartheta$.

Choosing any co-ordinates $q_1, \ldots, q_n$ on the configuration space M and completing them with the conjugate momenta $p_1, \ldots, p_n$ one automatically has a canonical co-ordinate system on the phase space $\mathcal{P} = T^*M$ in which the canonical 1-form reads

$$\vartheta = \sum_{i=1}^{n} p_i \, \mathrm{d}q_i$$

and the symplectic form is given by (5). This freedom of choosing any co-ordinate system is lost on the whole phase space if one insists on the equations of motion to be in canonical form (3).

A significant part of the classical literature [2,12,16,31] is devoted to *generating functions*, a means to generate transformations that turn canonical co-ordinates into canonical co-ordinates, therefore called *canonical transformations*. A contemporary means to obtain canonical transformations is to use the time-1-flow φ_1 of a well-chosen Hamiltonian function as these are better suited for implementation on a computer. Since $\varphi_1^{-1} = \varphi_{-1}$ it is as simple (or complicated) to compute the inverse of such a transformation as it is to compute the transformation itself.

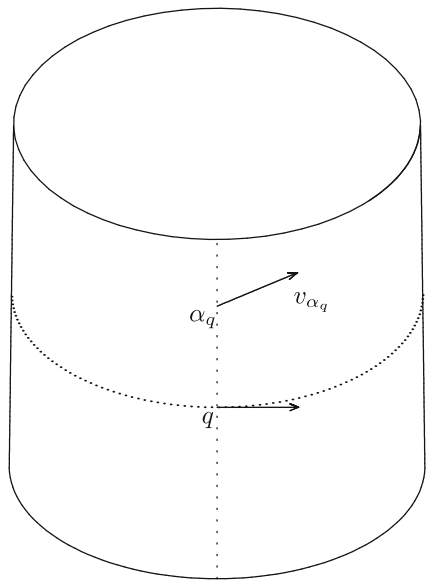

Dynamics of Hamiltonian Systems, Figure 1
Projection of a vector v_{α_q} tangent to T^*M to the tangent space of M at $q = \pi(\alpha_q)$. The resulting tangent vector can be fed into $\alpha_q : T_qM \longrightarrow \mathbb{R}$

In a (not necessarily canonical) co-ordinate system $z_1, \ldots, z_{2n}$ one can use the *Poisson bracket*, defined by

$$\{F, H\} := \omega(\mathrm{X}F, \mathrm{X}H) \,,$$

to write down the equations of motion

$$\dot{z}_j = \{z_j, H\} = \sum_{k=1}^{2n} \frac{\partial H}{\partial z_k} \{z_j, z_k\}$$

which have the canonical form precisely when it is possible to write $(z_1, \ldots, z_n) = (q_1, \ldots, q_n)$ and $(z_{n+1}, \ldots, z_{2n}) = (p_1, \ldots, p_n)$ with Poisson bracket relations

$$\{q_i, q_j\} = 0 \,, \quad \{p_i, p_j\} = 0 \,, \quad \{q_i, p_j\} = \delta_{ij} \,.$$

While a symplectic manifold has necessarily even dimension, one can turn also manifolds of odd dimension into a Poisson space. An important example is the Poisson structure

$$\{F, H\}(z) := \langle \nabla F(z) \times \nabla H(z) \mid z \rangle \tag{6}$$

on $\mathbb{R}^3$ (with its inner product $\langle x \mid y\rangle = x_1y_1 + x_2y_2 + x_3y_3$) for which the equations of motion read

$$\dot{z} = \nabla H(z) \times z\,. \tag{7}$$

The function $R(z) = (1/2)\langle z \mid z\rangle$ is a *Casimir function* of the Poisson structure as it is invariant under every Hamiltonian flow on $\mathbb{R}^3$ since $\{R, H\} = 0$ for all (Hamiltonian) functions $H\colon \mathbb{R}^3 \longrightarrow \mathbb{R}$. This fibrates $\mathbb{R}^3$ into invariant spheres $\{R = (1/2)\rho^2\}$ of radius ρ, with a singular sphere reduced to a point at the origin. The area element σ makes each sphere S^2_ρ a symplectic manifold, and the restriction of the Poisson bracket (6) to S^2_ρ satisfies

$$\{F, H\} = \tfrac{1}{\rho^2}\,\sigma(\mathrm{X}\,F, \mathrm{X}\,H)\,.$$

A similar (though in general slightly more complicated) decomposition into symplectic leaves exists for every Poisson manifold [2,22,27].

A first measure for the complexity of a Hamiltonian system is the number of degrees of freedom. For a simple mechanical system this is the dimension of the configuration space, and accordingly one defines this number as $(1/2)\dim\mathcal{P}$ for a symplectic manifold $\mathcal{P}$. On a Poisson space this number is related to the rank of the Poisson bracket, given by $\mathrm{rank}(\{z_j, z_k\})_{j,k=1,\ldots,m}$ in local coordinates $z_1, \ldots, z_m$. This even number is upper semicontinuous and coincides at each point with the dimension of the symplectic leaf passing through that point. Hence, the number of degrees of freedom is defined as one half of the maximal rank of the Poisson structure.

Systems in One Degree of Freedom

For the simple mechanical systems consisting of a 1-dimensional particle with kinetic energy (2) moving in a potential $V = V(q)$ the trajectories coincide with the level curves $\{H = h\}$ of $H = T + V$. Finding the time parametrizations of the trajectories is thereby reduced to the quadrature

$$\int \frac{\mathrm{d}q}{\sqrt{2m(h - V(q))}}\,,$$

and correspondingly one speaks of an integrable system. Where this time parametrization is not important one can always multiply the Hamiltonian by a strictly positive function, in the present one-degree-of-freedom situation one has then the extra choice [6,13] of performing any coordinate transformation and still writing the equations in canonical form (with respect to the transformed Hamiltonian function).

The phase portraits in a Poisson space $\mathcal{P}$ with one degree of freedom can be obtained by intersecting the level sets of the energy with the symplectic leaves. In particular, a point where the rank of the Poisson structure drops from 2 to 0 is an equilibrium for every Hamiltonian system on $\mathcal{P}$. Equilibria on regular symplectic leaves are called regular equilibria.

There are four types of linearizations of regular equilibria in one degree of freedom. In canonical co-ordinates these are given by quadratic Hamiltonian functions H_2.

Elliptic $H_2(q, p) = (\varpi/2)(p^2 + q^2)$, the nearby motion is periodic with frequency close to ϖ. This is the harmonic oscillator.

Hyperbolic $H_2(q, p) = (a/2)(p^2 - q^2)$, the equilibrium is a *saddle*.

Parabolic $H_2(q, p) = (a/2)p^2$, the higher order terms of the Hamiltonian around the equilibrium determine the character of the flow.

Vanishing $H_2(q, p) \equiv 0$, the linearization contains no information and the flow is given by the higher order terms.

Generic Hamiltonian systems on a symplectic surface have only elliptic and hyperbolic equilibria, see Fig. 2. Where the system depends on external parameters or is defined on a family of symplectic leaves one may also encounter parabolic equilibria. The phenomenon of vanishing linearization is of co-dimension three.

As an example consider the phase space $\mathbb{R}^3$ with Poisson structure (6) and Hamiltonian energy function

$$H(z) = \sum_{i=1}^{3} \frac{a_i}{2} z_i^2 + b_i z_i$$

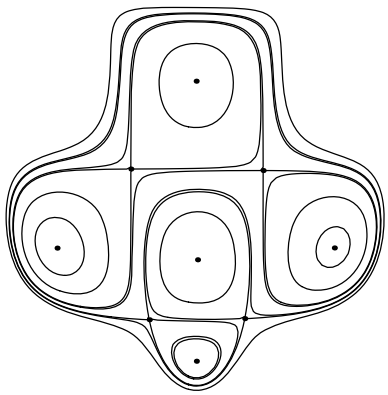

Dynamics of Hamiltonian Systems, Figure 2
Typical recurrent flow in one degree of freedom

depending on the external parameters $(a, b) \in \mathbb{R}^3 \times \mathbb{R}^3$. On each sphere S^2_ρ the points with minimal and maximal energy are elliptic equilibria. All equilibria occur where S^2_ρ touches the quadric $\{H = h\}$ of constant energy. Where this happens with coinciding curvature along a line the equilibrium is parabolic. For $a, b \in \mathbb{R}^3$ in general position a *center-saddle bifurcation* occurs as ρ passes through such a value. If e. g. the three conditions $a_1 = a_2$, $b_1 = b_2 = 0$ hold, then the curvatures at

$$z = \begin{pmatrix} 0 \\ 0 \\ -\frac{b_i}{a_i} \end{pmatrix}$$

coincide along all lines and the linearization at this equilibrium vanishes. The origin $z = 0$ is for all values of $a, b \in \mathbb{R}^3$ a (singular) equilibrium.

Systems in Two Degrees of Freedom

While all recurrent motion is periodic in one degree of freedom, the flow can have a stochastic (or chaotic) character already in two degrees of freedom. The level sets $\{H = h\}$ of the Hamiltonian are now 3-dimensional invariant hypersurfaces, and complicated dynamics is known to be possible from dimension three on. Still, being Hamiltonian imposes certain restrictions.

Leaving aside equilibria with vanishing eigenvalues, there are the following types of linearizations of regular equilibria in two degrees of freedom, with quadratic Hamiltonian H_2 in canonical co-ordinates.

Elliptic The quadratic part

$$H_2(q, p) = \alpha\, \frac{p_1^2 + q_1^2}{2} + \varpi\, \frac{p_2^2 + q_2^2}{2}, \text{ with } |\alpha| \le |\varpi|, \tag{8}$$

is the superposition of two harmonic oscillators. For $\alpha/\varpi \in \mathbb{Q}$ the motion is periodic, but for irrational frequency ratio the trajectories spin densely around invariant tori. In the case $\varpi = -\alpha$ of $1 : -1$ resonance one can add a nilpotent part

$$\frac{p_1^2 + q_1^2 \;+\; p_2^2 + q_2^2}{4} - \frac{p_1 q_2 + p_2 q_1}{2}$$

whence the linear flow becomes unbounded.

Hypo-elliptic The linear vector field with quadratic Hamiltonian

$$H_2(q, p) = a\, \frac{p_1^2 - q_1^2}{2} + \varpi\, \frac{p_2^2 + q_2^2}{2}$$

has one pair of eigenvalues on the real axis and one pair of eigenvalues on the imaginary axis. One also speaks of a saddle-center equilibrium.

Hyperbolic The linearization has no eigenvalues on the imaginary axis. In case the spectrum consists of two pairs of real eigenvalues the standard form of the Hamiltonian is

$$H_2(q, p) = a\, \frac{p_1^2 - q_1^2}{2} + b\, \frac{p_2^2 - q_2^2}{2}$$

and one speaks of a saddle-saddle equilibrium (or real saddle). In the alternative case of a complex quartet $\pm\alpha \pm \mathrm{i}\varpi$ one has

$$H_2(q, p) = \alpha(p_1 q_1 + p_2 q_2) + \varpi(p_1 q_2 - p_2 q_1)$$

and speaks of a *focus-focus* equilibrium (or complex saddle), since the flow on the stable manifold spirals into the equilibrium and the flow on the unstable manifold spirals away from the equilibrium.

The $1 : -1$ resonance is special in that it typically triggers off a *Hamiltonian Hopf bifurcation* during which an elliptic equilibrium becomes hyperbolic, of focus-focus type. Other bifurcations occur where eigenvalues vanish under variation of an external parameter (possibly parametrizing 4-dimensional symplectic leaves in a higher-dimensional Poisson space).

In generic Hamiltonian systems all equilibria are isolated and the periodic orbits form 1-parameter families. In one degree of freedom these families extend between (elliptic) equilibria and/or the *(un)stable manifolds* of (hyperbolic) equilibria, see Fig. 2. In two degrees of freedom the 1-parameter families of periodic orbits are special solutions and one may hope that they similarly organize the dynamics. The normal linear behavior of a periodic orbit with period τ is determined by its Floquet multipliers, the eigenvalues of the linearization $D\varphi_\tau$ of the flow.

The simplest way to find periodic orbits in a Hamiltonian system, e. g. as a starting point for continuation, is to look for the *normal modes* of an equilibrium.

Theorem (Liapunov) *Let $\lambda_\pm = \pm\mathrm{i}\alpha$ be a purely imaginary pair of eigenvalues of the linearization $A = D\,\mathrm{X}\,H(z)$ at an equilibrium $z \in \mathcal{P}$ of a Hamiltonian system $\mathrm{X}\,H$ on a symplectic manifold $\mathcal{P}$ for which no integer multiple $k\lambda_\pm$, $k \in \mathbb{N}$, is an(other) eigenvalue of A. Then $\mathrm{X}\,H$ admits a 1-parameter family $(\gamma_\varepsilon)_{0<\varepsilon\le\varepsilon_0}$ of periodic orbits that approach z as $\varepsilon \to 0$ with periods $T_\varepsilon \xrightarrow{\varepsilon\to 0} 2\pi/\alpha$. The union*

$$\{z\} \cup \bigcup_{0<\varepsilon\le\varepsilon_0} \gamma_\varepsilon$$

forms a smooth 2-dimensional submanifold of $\mathcal{P}$ with boundary γ_{ε_0} that is diffeomorphic to the closed disk in $\mathbb{R}^2$.

For a proof see [1,23,26]; this result immediately generalizes to n degrees of freedom.

If the Hessean $D^2H(z)$ is positive (or negative) definite, the non-resonance condition $\lambda \neq k\lambda_\pm$ for the remaining eigenvalues of A can be dropped, but the resulting families of periodic orbits may no longer form manifolds through z and only form cones with z as vertex instead. In two degrees of freedom the $1:-1$ and $1:-2$ resonances provide examples where the normal mode is lacking, but for $1:-k$ resonant equilibria with $k \geq 3$ it turns out to be generic for the normal modes to exist.

The normal mode of a hypo-elliptic equilibrium in two degrees of freedom is a hyperbolic periodic orbit with a real pair $a, (1/a) \neq \pm 1$ of Floquet multipliers. The periodic orbits born at an elliptic equilibrium also inherit their normal behavior from the "second" pair of eigenvalues and have a pair $e^{\pm i\varpi} \neq \pm 1$ of Floquet multipliers on the unit circle. Coupling two generic systems with one degree of freedom with a sufficiently weak interaction yields periodic orbits that are globally organized in this fashion, each periodic orbit being the superposition of a center or a saddle in one subsystem and a periodic orbit in the other subsystem.

The periodic orbits that are neither elliptic nor hyperbolic are isolated within the 1-parameter families of periodic orbits. The arcs in between consist either completely of elliptic or completely of hyperbolic periodic orbits and may be parametrized by the values of the energy. The following types of bifurcations are triggered off by parabolic periodic orbits, for which the *normal linear* behavior is governed by a Floquet matrix $\left(\begin{smallmatrix}1&1\\0&1\end{smallmatrix}\right)$ or $\left(\begin{smallmatrix}-1&1\\0&-1\end{smallmatrix}\right)$.

The periodic center-saddle bifurcation Under variation of the energy an elliptic and a hyperbolic periodic orbit meet at a parabolic periodic orbit with all Floquet multipliers equal to 1. No periodic orbit remains when further in(or de)creasing the energy. In a suitable projection

$$(H, I)\colon \mathcal{P} \longrightarrow \mathbb{R}^2 \xrightarrow{H} \mathbb{R}$$

the family of periodic orbits forms a fold. See also Fig. 4 below.

The Hamiltonian period-doubling bifurcation Under variation of the energy an elliptic periodic orbit turns hyperbolic (or *vice versa*) when passing through a parabolic periodic orbit with Floquet multipliers -1. Furthermore a family of periodic orbits with twice the period emerges from the parabolic periodic orbit, inheriting the normal linear behavior from the initial periodic orbit. See also Fig. 5 below.

For proofs of this and the following see [20,23].

In generic systems with two degrees of freedom these are the only occurring bifurcations of periodic orbits. In three and more degrees of freedom there is "more space" in the normal direction and also periodic Hamiltonian Hopf bifurcations may occur. A prominent example is the gyroscopic stabilization of the rotating Lagrange top "standing up", cf. [2,8,22].

Along arcs of elliptic periodic orbits the pair $e^{\pm i\varpi}$ of Floquet multipliers passes regularly through roots of unity. Generically this happens on a dense set of parametrizing energy values, but for fixed denominator ℓ in $e^{\pm i\varpi} = e^{\pm 2\pi i k/\ell}$ the corresponding energy values are again isolated. The cases $\ell = 1$ and $\ell = 2$ correspond to the above bifurcations so the "first" case is $\ell = 3$.

Two arcs of hyperbolic periodic orbits emerge at elliptic orbits with Floquet multipliers $e^{\pm 2\pi i k/3}$, both with three times the period. One extends for lower and the other for higher energy values. The periodic orbit with Floquet multipliers $e^{\pm 2\pi i k/3}$ momentarily loses its (normal) stability due to these approaching unstable orbits.

In the case of Floquet multipliers $e^{\pm 2\pi i k/\ell}$ with $\ell \geq 5$ again two arcs of periodic orbits with ℓ times the period emerge, but now one is elliptic and the other hyperbolic, and they both extend to the same side of energy values (either lower or higher). Furthermore there is no momentary loss of (normal) stability as these families detach. For $\ell = 4$ both the $\ell = 3$ and the $\ell \geq 5$ scenarios can occur.

It is at isolated values of the energy that the two arcs of ($\ell \geq 5$)-times periodic orbits (or ($\ell = 4$)-times where appropriate) change from extending along lower energy values to extending along higher energy values. For definiteness let us consider a sub-arc between such values where all these exist for higher energy values. Where this sub-arc contains periodic orbits with Floquet multipliers $e^{\pm 2\pi i k/3}$ (or $e^{\pm 2\pi i k/4}$ where appropriate) the family of 3-times periodic orbits that extends for lower energy values typically vanishes in a nearby periodic center-saddle bifurcation. The family of elliptic periodic orbits born in this bifurcation then extends along higher energy values as well. A possible arc of elliptic periodic orbits is sketched in Fig. 3. For more details on periodic solutions of Hamiltonian systems see the entry ► Periodic Orbits of Hamiltonian Systems in this encyclopedia.

To visualize the flow on the 3-dimensional *energy shells* $\{H = h\}$ one uses iso-energetic Poincaré-sections, i. e. surfaces $\Sigma_h \subset \{H = h\}$ that are everywhere transverse to the vector field X_H. For recurrent points $z \in \Sigma_h$

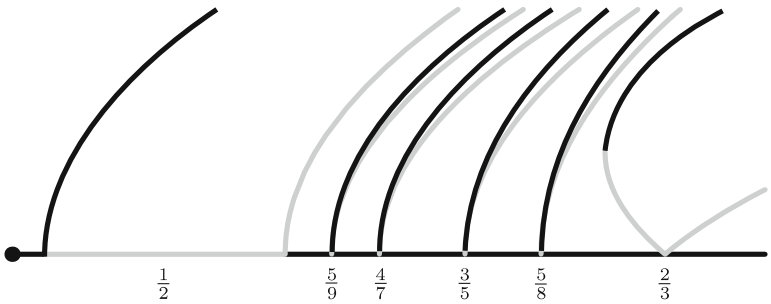

Dynamics of Hamiltonian Systems, Figure 3
Subharmonic branching along a normal mode of an elliptic equilibrium in a response diagram with axes along normal frequency and amplitude. Where normal and internal frequency of the periodic orbit emanating from the equilibrium have a ratio $\varpi = 2\pi(k/\ell)$ with integers $k, \ell \in \mathbb{Z}$ two periodic orbits with ℓ times the period branch off, one elliptic and one hyperbolic (shown in *gray*). For $\ell = 3$ the distance between these widens as the two periodic orbits come into existence in a 3-periodic center-saddle bifurcation "before" the resonance after which the hyperbolic 3-periodic orbit passes through the normal mode at the (2/3)-resonance (shown reflected as the vertical depicts the amplitude). For $\ell = 2$ the resonance leads to a gap within the elliptic normal mode, filled by hyperbolic periodic orbits and with boundaries marked by two Hamiltonian period-doubling bifurcations

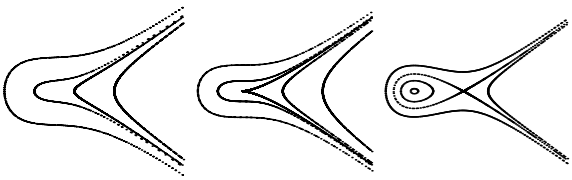

Dynamics of Hamiltonian Systems, Figure 4
Center-saddle bifurcation in an area-preserving mapping

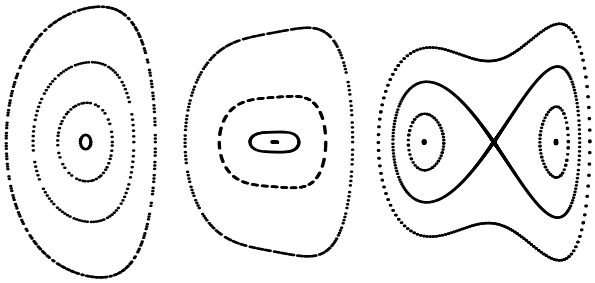

Dynamics of Hamiltonian Systems, Figure 5
Period-doubling bifurcation in an area-preserving mapping

the (first) return time

$$\tau(z) := \min\{\, T > 0 \mid \varphi_T(z) \in \Sigma_h \,\}$$

allows to define the *iso-energetic Poincaré-mapping*

$$\begin{aligned} P_h \colon \Sigma_h &\longrightarrow \Sigma_h \\ z &\mapsto \varphi_{\tau(z)}(z) \end{aligned}.$$

The phase space volume $\omega^2 = \omega \wedge \omega$ coming from the symplectic structure induces an area element σ on Σ_h that is preserved by P_h. On the other hand, every area-preserving mapping can be realized as an iso-energetic Poincaré-mapping of a Hamiltonian system.

A periodic orbit γ of $\mathrm{X}\,H$ with energy h corresponds to a fixed point $x \in \Sigma_h$ (or to a periodic point of period $k \in \mathbb{N}$ if γ has $k-1$ intersections with Σ_h before returning to x). Two of the Floquet multipliers of γ are equal to 1, reflecting that periodic orbits form 1-parameter families in Hamiltonian systems and that moving the initial condition within γ yields that same periodic orbit with a translated time parametrization. The remaining two Floquet multipliers are the eigenvalues of the linearization $DP_h(z)$ of the iso-energetic Poincaré-mapping at a fixed point z.

Because of area-preservation $\det DP_h(z) = 1$, so one eigenvalue of $DP_h(z)$ has to be the inverse of the other eigenvalue. For an elliptic periodic orbit both eigenvalues lie on the unit circle where the inverse equals the complex conjugate. For hyperbolic γ both eigenvalues are real and one sometimes makes the distinction between the direct hyperbolic case of positive eigenvalues and the inverse hyperbolic case of negative eigenvalues. See Fig. 4 for the center-saddle bifurcation triggered off by the double eigenvalue $+1$ and Fig. 5 for the period-doubling bifurcation triggered off by the double eigenvalue -1.

Symmetry Reduction

When a dynamical system admits a symmetry group it is possible to simplify the dynamics. This reduction process is especially rewarding in Hamiltonian systems, where Noether's theorem yields for every continuous symmetry a conserved quantity. One even has the choice between first fixing the values of the conserved quantities and then

reducing what is left of the symmetry, or first reducing the symmetry and then fixing the remaining conserved quantities.

Let $H \in C^\infty(\mathcal{P})$ be a Hamiltonian function on the symplectic manifold $(\mathcal{P}, \omega)$ and G a compact Lie group (i. e. the group operation is smooth) with Lie algebra $\mathfrak{g}$ (the tangent space T_eG at the neutral element $e \in G$, provided with the Lie bracket that measures the non-commutativity of the group operation). Results for more general groups do exist, but some kind of compactness, e. g. that the *group action* be proper, is always needed. Assume that the group action

$$\begin{aligned} G \times \mathcal{P} &\longrightarrow \mathcal{P} \\ (g, z) &\mapsto gz \end{aligned} \tag{9}$$

preserves both the Hamiltonian function H and the symplectic form ω. Then G also preserves the Hamiltonian vector field $\mathrm{X}H$ and the resulting flow commutes with the group action (9), i. e. $\varphi_t \circ g = g \circ \varphi_t$ for all $(t, g) \in \mathbb{R} \times G$. In fact, this allows to combine the flow $\varphi\colon \mathbb{R} \times \mathcal{P} \longrightarrow \mathcal{P}$ of $\mathrm{X}H$ and the action (9) to the action

$$\begin{aligned} (\mathbb{R} \times G) \times \mathcal{P} &\longrightarrow \mathcal{P} \\ ((t, g), z) &\mapsto \varphi_t(gz) \end{aligned} \tag{10}$$

of the Lie group $\mathbb{R} \times G$ on $\mathcal{P}$.

Reduction aims to find a phase space of smaller dimension on which the dynamics can be studied. Identifying points $z \sim gz$ that are transformed into each other by group elements leads to the quotient space $\mathcal{P}/G$ with canonical projection

$$\mathcal{P} \longrightarrow \mathcal{P}/G\,.$$

As $H(z) = H(gz)$ for all $g \in G$ the Hamiltonian induces a function on $\mathcal{P}/G$, again denoted by H. However, the symplectic structure on $\mathcal{P}$ does not induce a symplectic structure on the quotient space. What can be transferred from $\mathcal{P}$ to $\mathcal{P}/G$ is the Poisson structure.

The symplectic form is preserved by G, whence the mappings $z \mapsto gz$ are canonical transformations, satisfying

$$\{F \circ g, H \circ g\} = \{F, H\} \circ g \quad \text{for all } F, H \in C^\infty(\mathcal{P})\,.$$

Thus, the Poisson bracket of G-invariant functions is again G-invariant, defining a Poisson bracket on $\mathcal{P}/G$. The Casimir functions on this Poisson space correspond to those G-invariant functions on $\mathcal{P}$ that are conserved quantities for every G-invariant Hamiltonian function.

As an example consider the free rigid body with a fixed point, subject only to its own inertia. The configuration space is $SO(3)$, all (rigid) rotations about the fixed point, with a Riemannian metric provided by the mass distribution. On the phase space $T^*SO(3)$ the Hamiltonian $H = T$ is given by the resulting kinetic energy.

The mass distribution and hence the Hamiltonian are invariant under rotations, making $SO(3)$ a symmetry group. Using the left trivialization

$$\begin{aligned} \lambda\colon T^*SO(3) &\longrightarrow SO(3) \times T_e^*SO(3) \\ \alpha_g &\mapsto (g, T_e^*L_g(\alpha_g)) \end{aligned}$$

defined by means of the left translation

$$\begin{aligned} L_g\colon SO(3) &\longrightarrow SO(3) \\ h &\mapsto gh \end{aligned}$$

the group action reads

$$\begin{aligned} SO(3) \times (SO(3) \times \mathbb{R}^3) &\longrightarrow SO(3) \times \mathbb{R}^3 \\ (g, (h, \ell)) &\mapsto (gh, \ell) \end{aligned}$$

and reveals $T^*SO(3)/SO(3) \cong \mathbb{R}^3$. Here $\ell \in \mathbb{R}^3 \cong \mathfrak{so}(3)^* = T_e^*SO(3)$ consists of the three components of the angular momentum with respect to a set of axes fixed in the body, whence one calls $\lambda(\alpha_g) = (g, \ell)$ body coordinates. Choosing the body set of axes along the principal axes of the rigid body the Hamiltonian takes the form

$$H(g, \ell) = H(\ell) = \frac{\ell_1^2}{2I_1} + \frac{\ell_2^2}{2I_2} + \frac{\ell_3^2}{2I_3}$$

where I_1, I_2, I_3 are the principal moments of inertia. The Poisson bracket relations inherited from $T^*SO(3)$ are

$$\{\ell_1, \ell_2\} = -\ell_3\,, \quad \{\ell_2, \ell_3\} = -\ell_1\,, \quad \{\ell_3, \ell_1\} = -\ell_2 \tag{11}$$

whence the Poisson structure on $\mathbb{R}^3$ is almost (6) considered in Sect. "The Phase Space", differing only by a minus sign. The Casimir function $R(\ell) = (1/2)(\ell_1^2 + \ell_2^2 + \ell_3^2)$ measures (half of the square of) the length of the angular momentum, fixing this conserved quantity yields the symplectic leaves of the quotient space $\mathbb{R}^3$.

The alternative approach to symmetric Hamiltonian systems first fixes the conserved quantities. For $\xi \in \mathfrak{g}$ the 1-parameter subgroup $\{\exp(s\xi) \mid s \in \mathbb{R}\}$ of G yields a conserved quantity

$$J^\xi\colon \mathcal{P} \longrightarrow \mathbb{R} \tag{12}$$

by Noether's theorem, cf. [2,6,23]. The Hamiltonian vector field X_{J^ξ} has the flow $(s, z) \mapsto \exp(s\xi) \cdot z$ provided by the 1-parameter subgroup.

Here and from now on the assumption is made that the (mild) conditions for Noether's theorem are fulfilled, making (9) a Hamiltonian group action. What has to be avoided is that the flow provided by the 1-parameter subgroup is only locally Hamiltonian, see [1,27,28] for more details.

In the example of the free rigid body a conserved quantity (12) is the component of the angular momentum along an axis fixed in space. Collecting these conserved quantities by means of

$$\begin{aligned} J\colon \mathcal{P} &\longrightarrow \mathfrak{g}^* \\ z \mapsto J(z)\colon \mathfrak{g} &\longrightarrow \mathbb{R} \\ \xi &\mapsto J^\xi(z) \end{aligned} \tag{13}$$

yields the *momentum mapping*. For the rigid body this amounts to fixing a set of axes in space and assigning to $\alpha_g \in T_g^* SO(3)$ the three components $\mu \in \mathbb{R}^3 \cong \mathfrak{so}(3)^*$ of the angular momentum with respect to these axes. Correspondingly, the right trivialization reads

$$\begin{aligned} \varrho\colon T^*SO(3) &\longrightarrow SO(3) \times \mathbb{R}^3 \\ \alpha_g &\mapsto (g, \mu) \end{aligned}$$

where $\mu = T_e^* R_g(\alpha_g)$ is obtained differentiating the right translation $R_g(h) = hg$. In space co-ordinates $\varrho(\alpha_g) = (g, \mu)$ the group action takes the form

$$\begin{aligned} SO(3) \times (SO(3) \times \mathbb{R}^3) &\longrightarrow SO(3) \times \mathbb{R}^3 \\ (g, (h, \mu)) &\mapsto (hg, g(\mu)) \end{aligned}$$

and one sees that the momentum mapping

$$SO(3) \times \mathbb{R}^3 \longrightarrow \mathbb{R}^3 \tag{14}$$

intertwines between the $SO(3)$-actions on the phase space and on $\mathbb{R}^3$.

To formulate the corresponding equivariance property of the momentum mapping in the general case the co-adjoint action

$$\begin{aligned} G \times \mathfrak{g}^* &\longrightarrow \mathfrak{g}^* \\ (g, \mu) &\mapsto \mathrm{Ad}^*_{g^{-1}}(\mu) = \mu \circ \mathrm{Ad}_g^{-1} \end{aligned} \tag{15}$$

is needed, here $\mathrm{Ad}_g\colon \mathfrak{g} \longrightarrow \mathfrak{g}$ is the derivative $T_e(L_g R_{g^{-1}})$ of the inner automorphism $h \mapsto ghg^{-1} = L_g R_{g^{-1}} h$ at the neutral element. The momentum mapping J of the Hamiltonian group action (9) is now called equivariant if it intertwines between (9) and (15), i. e.

$$J(gz) = \mathrm{Ad}^*_{g^{-1}} J(z) \quad \text{for all } (g, z) \in G \times \mathcal{P}\,.$$

In this case the assignment $\xi \mapsto J^\xi \in C^\infty(\mathcal{P})$ of conserved quantities to Lie algebra elements $\xi \in \mathfrak{g}$ turns out to be a Lie algebra homomorphism, cf. [1,22,28]. Providing $\mathfrak{g}^*$ with the *Lie–Poisson bracket* (defined in terms of the commutator on $\mathfrak{g}$) makes (13) a canonical mapping, see [27,28] for more details. The identification $\mathfrak{so}(3)^* \cong \mathbb{R}^3$ turns the Lie–Poisson bracket into (6) and indeed $\{\mu_1, \mu_2\} = \mu_3$ etc. for the spatial representation of the angular momentum. Working with right group actions (instead of acting from the left) would yield a minus sign here and a plus sign in (11).

Let $\mu \in \mathfrak{g}^*$ be a regular value of the equivariant momentum mapping (13). Then $J^{-1}(\mu)$ is a submanifold of $\mathcal{P}$ and the (compact) isotropy subgroup

$$G_\mu = \left\{ g \in G \mid \mathrm{Ad}^*_{g^{-1}}(\mu) = \mu \right\}$$

of the co-adjoint action leaves this manifold invariant as

$$J(gz) = \mathrm{Ad}^*_{g^{-1}} J(z) = \mu \quad \text{for all } (g, z) \in G_\mu \times J^{-1}(\mu)$$

whence the restriction

$$\begin{aligned} G_\mu \times J^{-1}(\mu) &\longrightarrow J^{-1}(\mu) \\ (g, z) &\mapsto gz \end{aligned} \tag{16}$$

defines a Lie group action on the manifold $J^{-1}(\mu)$. Passing to the quotient $J^{-1}(\mu)/G_\mu$ is called symplectic point reduction.

Theorem (Marsden and Weinstein) *Let μ be a regular value of J and assume that the action (16) is free. Then the reduced phase space $\mathcal{P}_\mu = J^{-1}(\mu)/_{G_\mu}$ is a symplectic manifold, with ω_μ uniquely determined by $\omega_{|J^{-1}(\mu)} = \omega_\mu \circ \pi_\mu$ where π_μ is the quotient projection. A G-invariant Hamiltonian function $H \in C^\infty(\mathcal{P})$ induces $H_\mu \in C^\infty(\mathcal{P}_\mu)$ and π_μ intertwines between the flows of X_H and X_{H_μ}.*

For a proof see [1,2,8,28].

In the example of the free rigid body the momentum mapping $SO(3) \times \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ is surjective and all values are regular, with inverse images $SO(3) \times \{\mu\}$. For $\mu \neq 0$ the isotropy subgroup $G_\mu \cong S^1$ consists of all rotations in $\mathbb{R}^3$ about the axis along μ and the reduced phase space

$$SO(3) \times \{\mu\}/G_\mu \cong SO(3)/S^1 \cong S^2$$

can be identified with the sphere of radius $|\mu|$ as

$$\begin{aligned} SO(3) \times \{\mu\} &\longrightarrow S^2_{|\mu|} \\ (g, \mu) &\mapsto g(\mu) \end{aligned}$$

performs the reduction. The symplectic form on $S^2_{|\mu|}$ is a multiple

$$\tfrac{-1}{|\mu|^2} \cdot \sigma$$

of the area element σ (the minus sign is the "same" as in (11)). The isotropy subgroup of $\mu = 0$ is the whole

$SO(3)$ and the quotient space consists of the origin $\ell = g(0)$ in $\mathbb{R}^3$.

As illustrated in this example, symplectic point reduction provides the symplectic leaves of the Poisson space $\mathcal{P}/G$ obtained by directly reducing the action (9) on $\mathcal{P}$. A link between these two procedures of symmetry reduction is symplectic orbit reduction. By equivariance of (13) the inverse image $J^{-1}(G(\mu))$ of the co-adjoint orbit

$$G(\mu) = \left\{ \mathrm{Ad}^*_{g^{-1}}\mu \mid g \in G \right\} \tag{17}$$

is invariant under the whole group G as

$$J(gz) = \mathrm{Ad}^*_{g^{-1}} J(z) \in G(\mu) \quad \text{for all } (g,z) \in G \times J^{-1}(G(\mu)).$$

Hence, the restriction of (9) to $J^{-1}(G(\mu))$ defines a group action and the quotient spaces $J^{-1}(G(\mu))/G$ form a partition of $\mathcal{P}/G$ (since

$$G = \bigcup_{\mu \in \mathfrak{g}^*} G(\mu)$$

and two orbits are either disjoint or equal). On the other hand, the two quotients $J^{-1}(G(\mu))/G$ and $J^{-1}(\mu)/G_\mu$ are symplectomorphic, see [27,28] for more details. In the example of the free rigid body, the inverse image of the co-adjoint orbit $G(\mu)$ is the sphere bundle in $T^*SO(3)$ of radius $|\mu|$ which is given by $SO(3) \times S^2_{|\mu|}$ both in body and space co-ordinates.

The group action (9) is free if every $z \in \mathcal{P}$ is moved away from z by every group element, i. e. all isotropy subgroups

$$G_z = \{\, g \in G \mid gz = z \,\}$$

of (9) (not of the co-adjoint action) are trivial. For free actions the Poisson space $\mathcal{P}/G$ is a Poisson manifold. Non-trivial isotropy groups typically lead to singularities. This can already be seen in the example of S^1 acting on $\mathbb{R}^4 = T^*\mathbb{R}^2 = \mathbb{R}^2 \times \mathbb{R}^2$ by simultaneous rotation in both planes, for the general theory see [8,27] and references therein.

The group $S^1 = SO(2)$ of planar rotations is commutative, whence the isotropy subgroups of the co-adjoint action coincide with the whole group and the generator $l = q_1p_2 - q_2p_1$ of the action

$$\begin{pmatrix} q \\ p \end{pmatrix} \mapsto \begin{pmatrix} \cos\rho & -\sin\rho & 0 & 0 \\ \sin\rho & \cos\rho & 0 & 0 \\ 0 & 0 & \cos\rho & -\sin\rho \\ 0 & 0 & \sin\rho & \cos\rho \end{pmatrix} \begin{pmatrix} q \\ p \end{pmatrix} \tag{18}$$

may be fixed both before and after passing to the quotient. The ring of S^1-invariant functions is generated by the polynomials

$$x = \frac{q_1^2 + q_2^2}{2}, \quad y = \frac{p_1^2 + p_2^2}{2}, \quad z = p_1q_1 + p_2q_2 \quad \text{and} \quad l$$

which are restricted by the relations

$$x \geq 0, \quad y \geq 0, \quad \text{and} \quad R_l(x,y,z) = 0$$

where

$$R_l(x,y,z) = \tfrac{1}{2}z^2 - 2xy + \tfrac{1}{2}l^2$$

defines the *syzygy* between these invariants. In particular, the S^1-invariant Hamiltonian function H may be written as a function $H = H_l(x,y,z)$.

This allows to refrain from local co-ordinates and use (x,y,z) as (global) variables on the reduced phase spaces

$$\mathcal{P}_l = \left\{ (x,y,z) \in \mathbb{R}^3 \mid x \geq 0, y \geq 0, R_l(x,y,z) = 0 \right\} \tag{19}$$

which foliate the Poisson space

$$\mathbb{R}^4/S^1 = \bigcup_{l \in \mathbb{R}} \mathcal{P}_l \times \{l\} \subseteq \mathbb{R}^4$$

with Poisson brackets given in Table 1. Fixing the Casimir function l, the resulting Poisson structure on $\mathbb{R}^3$ can also be written as

$$\{F, H\} = \langle \nabla F \times \nabla H \mid \nabla R_l \rangle \tag{20}$$

(generalizing (6)), revealing R_l to be the second Casimir function.

The fixed point $(q,p) = 0$ of (18) has the whole group S^1 as isotropy group, all other isotropy groups of (18) are trivial. Thus, for $l \neq 0$ the positive sheets $\mathcal{P}_l$ of the two-sheeted hyperboloids $R_l^{-1}(0)$ yield symplectic leaves of the Poisson space $\mathbb{R}^4/S^1$. For $l = 0$ the reduced phase space $\mathcal{P}_0$ is *stratified* into two symplectic strata of dimensions 2 and 0, the positive part $x + y > 0$ of the double cone $R_0^{-1}(0)$ and the vertex of this cone, respectively.

Dynamics of Hamiltonian Systems, Table 1
Poisson structure on $\mathbb{R}^4$

$\{\downarrow, \rightarrow\}$	x	y	z	l
x	0	z	$2x$	0
y	$-z$	0	$-2y$	0
z	$-2x$	$2y$	0	0
l	0	0	0	0

Hence, the vertex is automatically an equilibrium of the reduced Hamiltonian system.

Where an action (9) of a discrete group G preserves Hamiltonian H and symplectic form ω there is no resulting conserved quantity, but one still can pass to the quotient $\mathcal{P}/G$. For instance, on $\mathbb{R}^2$ the Hamiltonians

$$H(q,p) = \frac{p^2}{2} \pm \frac{q^4}{24} + \frac{\lambda q^2}{2}$$

admit the symmetry group $G = \{\pm \mathrm{id}\}$ that also preserves the symplectic form $\mathrm{d}q \wedge \mathrm{d}p$. The reduced Poisson space is the cone (19) with $l = 0$.

A symmetry may preserve only the equations of motion, but neither the Hamiltonian function nor the symplectic form. An example on $\mathbb{R}^2$ is given by

$$H(q,p) = \frac{p^3}{6} + \frac{pq^3}{6} + \lambda p + \mu pq$$

satisfying $H(q,-p) = -H(q,p)$ and $\mathrm{d}q \wedge \mathrm{d}(-p) = -\mathrm{d}q \wedge \mathrm{d}p$. The quotient may be realized as

$$\mathbb{R}^2/\{p \mapsto -p\} = \left\{ (q,p) \in \mathbb{R}^2 \mid p \geq 0 \right\}$$

and inherits outside the boundary the symplectic structure $\mathrm{d}q \wedge \mathrm{d}p$. The q-axis is invariant under the flow, with equation of motion $\dot{q} = \partial H/\partial p$.

Integrable Systems

Let $(\mathcal{P}, \omega)$ be a symplectic manifold of dimension $\dim \mathcal{P} = 2n$ and $H \in C^\infty(\mathcal{P})$ a Hamiltonian function. The Hamiltonian vector field $\mathrm{X}H$ is completely (or Liouville) integrable if there are n functions $F_1, \ldots, F_n \in C^\infty(\mathcal{P})$ with $\{F_i, H\} = 0$, $i = 1, \ldots, n$ and $\{F_i, F_j\} = 0$, $i, j = 1, \ldots, n$ (the F_i are *integrals* in involution) that are functionally independent outside a set of measure zero [1] or on an open and dense set [8], i. e. the closed set

$$\left\{ z \in \mathcal{P} \mid \det\left(X_{F_1}(z), \ldots, X_{F_n}(z)\right) = 0 \right\} \tag{21}$$

is small, e. g. of non-zero co-dimension. Examples abound, among them are all linear systems, all systems with one degree of freedom and uncoupled superpositions of completely integrable systems. Since one may choose $F_1 := H$ a system in two degrees of freedom is integrable as soon as it has a conserved quantity F_2. Where there is an S^1-symmetry, as in the (planar) two body problem and the geodesic flow on a surface of revolution, the momentum mapping $J =: F_2$ provides this conserved quantity.

The simplifying assumptions usually made when modeling e. g. a mechanical system often introduce extra symmetries. Consequently, some of the problems from classical mechanics, like the Lagrange top, turned out to be integrable. The continuous efforts of the 19th century lead to more integrable systems, like the geodesic flow on a triaxial ellipsoid and the Kovalevskaya top. Eventually it became clear that integrable systems are the exception and non-integrable systems are the rule, with as most prominent example the three body problem. However, the discovery of the Toda lattice renewed interest and the list of known integrable systems is still growing.

Theorem (Liouville) *Let the components of $F\colon \mathcal{P} \longrightarrow \mathbb{R}^n$ be n integrals in involution of the Hamiltonian $H = F_1$. For a regular value $c \in \operatorname{im} F \subseteq \mathbb{R}^n$ a compact connected component $\mathcal{R}_c$ of $F^{-1}(c)$ is an $\mathrm{X}H$-invariant manifold that is diffeomorphic to the n-torus $\mathbb{T}^n = \mathbb{R}^n/\mathbb{Z}^n$. The subset $\mathcal{R}_c \subseteq \mathcal{P}$ has an open neighborhood $\mathcal{U}$ on which $\mathrm{X}H$ admits action angle variables $(x, y) \in \mathbb{T}^n \times \mathbb{Y}$, $\mathbb{Y} \subseteq \mathbb{R}^n$ open, i. e. the diffeomorphism*

$$\begin{aligned} \mathcal{U} &\longrightarrow \mathbb{T}^n \times \mathbb{Y} \\ z &\mapsto (x(z), y(z)) \end{aligned}$$

turns the symplectic structure ω into $\sum \mathrm{d}x_i \wedge \mathrm{d}y_i$ and $y\colon \mathcal{U} \longrightarrow \mathbb{Y}$ factors through $F\colon \mathcal{U} \longrightarrow \mathbb{R}^n$.

For a proof see [1,2,8,26].

The last statement makes F independent of the angle variable x. In particular, $H = H(y)$ for the Hamiltonian function (though not $H = H(y_1)$ as in Hamilton–Jacobi theory) and the equations of motion read

$$\begin{aligned} \dot{x} &= \varpi(y) := DH(y) \\ \dot{y} &= 0 \end{aligned}$$

whence the flow $\varphi_t(x, y) = (x + t \cdot \varpi(y), y)$ is easily computed. Thus, constructing action angle variables of an integrable system is equivalent to explicitly solving the equations of motion. The term "completely integrable" indicates that this can be achieved by solving algebraic equations and indefinite integrals.

The invariant torus $\mathcal{R}_c$ is Lagrangean, i. e. an *isotropic* submanifold (the symplectic structure vanishes on $\mathcal{R}_c$) and maximal with that property. The dimension of *Lagrangean submanifolds* is always equal to the number of degrees of freedom. The flow on $\mathcal{R}_c$ is *conditionally periodic*, in the angle variable x it is parallel with frequency vector $\varpi(0)$, assuming that $y = 0 \in \mathbb{Y}$ is the action value of $\mathcal{R}_c$.

Denote by $C \subseteq \operatorname{im} F \subseteq \mathbb{R}^n$ the regular values of F, the complement of the image of (21) under F in $\operatorname{im} F$. Assume from now on that all level sets $F^{-1}(c)$, $c \in C$ are compact, e. g. because $F\colon \mathcal{P} \longrightarrow \mathbb{R}^n$ is proper or all energy level sets are compact. Putting $\mathcal{R} := F^{-1}(C)$ yields an n-torus bundle with fibers $\mathcal{R}_c$, and assuming that all

$F^{-1}(c)$ are connected, the restriction

$$F\colon \mathcal{R} \longrightarrow C \tag{22}$$

can be used as projection mapping of this bundle. In case $\mathcal{P} = T^*M$ is the cotangent bundle of a configuration space M the $\mathbb{T}^n$-bundle is trivial if and only if there are global action angle variables. The first obstruction for a torus bundle to be trivial is *monodromy*. For instance, if the inverse image $S = \mathcal{P}^{\backslash\mathcal{R}}$ of the singular values of F contains an $(n-2)$-parameter family of invariant $(n-2)$-tori with normal linear behavior of focus-focus type, then the bundle (22) has non-trivial monodromy.

Associated to (22) is the homology bundle $H_1(\mathcal{R}/C, \mathbb{Z})$ of $\mathcal{R}$ over C with fiber $H_1(\mathcal{R}_c, \mathbb{Z}) \cong \mathbb{Z}^n$, supplying the lattice that has to be divided out of $H_1(\mathcal{R}_c, \mathbb{R}) \cong \mathbb{R}^n$ to obtain the particular torus $\mathcal{R}_c$ at $c \in C$. For each path $\gamma\colon [0,1] \longrightarrow C$ the lift to $H_1(\mathcal{R}, \mathbb{Z})$ yields a bijective orientation-preserving $\mathbb{Z}$-linear mapping between the lattices at $\gamma(0)$ and $\gamma(1)$. For a closed loop the two lattices coincide and the mapping is represented by a matrix $\mathcal{M}(\gamma) \in SL_n(\mathbb{Z})$ with integer coefficients and determinant 1. This discrete object remains invariant under homotopies, and the resulting homomorphism

$$\mathcal{M}\colon \pi_1(C) \longrightarrow SL_n(\mathbb{Z})$$

from the first homotopy group of C into $SL_n(\mathbb{Z})$ is the monodromy of the n-torus bundle $\mathcal{R}$. In case $\mathcal{M} \equiv \mathrm{id}$ one can uniquely move a chosen basis $Y_1(c), \ldots, Y_n(c)$ of $H_1(\mathcal{R}_c, \mathbb{Z})$ at some chosen point $c \in C$ to all other fibers of the bundle $H_1(\mathcal{R}, \mathbb{Z})$, using paths with $\gamma(0) = c$. This yields Hamiltonian vector fields

$$X_{y_1} = Y_1 \circ F, \ldots, X_{y_n} = Y_n \circ F$$

on $\mathcal{R}$ for which the Hamiltonian functions $y_1, \ldots, y_n \in C^\infty(\mathcal{R})$ are global action variables. The remaining task is to also find global angle variables.

The global actions y_i define Hamiltonian vector fields with periodic flows. This yields a free Hamiltonian group action

$$\mathbb{T}^n \times \mathcal{R} \longrightarrow \mathcal{R}$$

and makes (22) a *principal* torus bundle.

Now principal torus bundles are classified by their *Chern class* in $H^2(C, \mathbb{Z}^n)$, a discrete invariant measuring the obstruction to the existence of a global section

$$\sigma\colon C \longrightarrow \mathcal{R}$$

(i. e. a mapping satisfying $F \circ \sigma = \mathrm{id}$). Such a global section yields in every fiber $\mathcal{R}_c$ the desired "origin" $x = 0$ of the angle variables and then one can let the group $\mathbb{T}^n$ act. For the resulting globally defined x, y to be action angle variables the equation

$$\mathrm{d}x \wedge \mathrm{d}y = \sum_{i=1}^{n} \mathrm{d}x_i \wedge \mathrm{d}y_i = \omega$$

has to be fulfilled, if necessary adapting the section σ accordingly. The cohomology class $[\sigma^*\omega] \in H^2(C, \mathbb{R})$ is a continuous invariant that vanishes if and only if this can be achieved. In the particular case that ω is exact, being defined by means of a canonical 1-form, one has $[\sigma^*\omega] = \sigma^*[\omega] = 0$. See [10] for more details.

A special but important situation occurs if the globally defined Hamiltonian vector fields

$$X_{F_2}, \ldots, X_{F_n} \tag{23}$$

already have periodic flows themselves (and only $F_1 = H$ has to be replaced by a local action y_1 in the construction of action angle variables). This defines an action

$$\mathbb{T}^{n-1} \times \mathcal{P} \longrightarrow \mathcal{P} \tag{24}$$

globally on the phase space (the restriction of which to the regular part $\mathcal{R}$ is free, giving $\mathcal{R}$ the structure of a principal $(n-1)$-torus bundle). Reducing this symmetry yields a one-degree-of-freedom problem on (the symplectic leaves of) the base space in

$$\sigma\colon \mathcal{P} \longrightarrow \mathcal{P}/\mathbb{T}^{n-1}$$

with set Σ of singular values. Constructing action angle variables amounts to finding the time parametrizations of the (relative) periodic trajectories together with the areas encircled by these. Let $\Xi \subseteq \mathcal{P}/\mathbb{T}^{n-1}$ denote the set of (relative) regular equilibria.

The singular part $S = \mathcal{P}^{\backslash\mathcal{R}}$ is the union of the energy level sets containing points of $\sigma^{-1}(\Sigma)$ – here the $n-1$ vector fields (23) are linearly dependent – and those containing points of $\sigma^{-1}(\Xi)$ – where $\mathrm{X}\,H(z)$ is a linear combination of the linear independent vector fields (23). This makes $\mathcal{P}$ a *ramified torus bundle*, with regular fibers in $\mathcal{R}$ forming n-parameter families of Lagrangean n-tori, the distribution of which is determined by the collection S of singular fibers.

In case the action (24) is free, the set Σ is empty and σ makes the whole phase space $\mathcal{P}$ a principal $\mathbb{T}^{n-1}$-bundle. The isotropic $(n-1)$-tori reconstructed from Ξ behave similar to the periodic orbits in a two-degrees-of-freedom system described in Sect. "Systems in Two Degrees of Freedom". Thus, the families of Lagrangean tori shrink down

to elliptic $(n-1)$-tori and are separated by the (un)stable manifolds of ($(n-1)$-parameter families of) hyperbolic $(n-1)$-tori, and the $(n-1)$-tori may undergo bifurcations. However, these bifurcations are more involved than those of periodic orbits for three reasons.

- Normal-internal resonances $\langle k \mid \varpi \rangle = \ell\alpha$ between the internal frequencies $\varpi_1, \ldots, \varpi_{n-1}$ and the *normal frequency* α of elliptic $(n-1)$-tori with $k \in \mathbb{Z}^{n-1}$ and $\ell = 1, 2$ are dense, triggering off quasi-periodic center-saddle bifurcations and *frequency-halving bifurcations*.
- The occurring bifurcations may be degenerate and typically have co-dimensions up to $n-1$. This is a genericity condition on H, within the "universe" of integrable Hamiltonian systems on $\mathcal{P}$.
- It is also generic for heteroclinic bifurcations reconnecting the (un)stable manifolds of $(n-1)$-tori to involve parabolic $(n-1)$-tori.

In case the action (24) has non-trivial isotropy groups, the invariant tori reconstructed from Σ of dimensions $n-2, n-3, \ldots, 2, 1, 0$ (the latter two being periodic orbits and equilibria) and their (un)stable manifolds form $\sigma^{-1}(\Sigma)$.

The description of $\mathcal{P}$ as a ramified n-torus bundle still applies when some (or all) of the vector fields (23) do not have periodic flows and some (or all) of the action variables $y_2, \ldots, y_n$ are only locally defined. The Lagrangean tori in $\mathcal{R}$ form n-parameter families and the singular fibers in S determine how these families fit together. At the $(n-1)$-parameter families of elliptic $(n-1)$-tori the Lagrangean tori shrink down in the same way as periodic orbits shrink down to centers in one degree of freedom. Different families of Lagrangean tori are separated by $(n-1)$-parameter families of hyperbolic $(n-1)$-tori and their (un)stable manifolds.

This picture is repeated in how the $(n-1)$-tori shrink down to $(n-2)$-parameter families of (partially) elliptic $(n-2)$-tori and are separated by $(n-2)$-parameter families of (partially) hyperbolic $(n-2)$-tori and (part of) their (un)stable manifolds. Furthermore there are $(n-2)$-parameter families of hyperbolic $(n-2)$-tori of focus-focus type, together with their (un)stable manifolds these form *pinched n-tori*. These three ways lead to invariant tori of smaller and smaller dimension until ending up with 1-parameter families of periodic orbits and isolated equilibria.

Within the family of all $(n-1)$-tori one encounters quasi-periodic center–saddle and frequency halving bifurcations along $(n-2)$-parameter subfamilies and more generally bifurcations of co-dimension $k \leq n-1$ along $(n-k-1)$-parameter subfamilies. Similarly, invariant $(n-2)$-tori undergoing a quasi-periodic Hamiltonian Hopf bifurcation form $(n-3)$-parameter families and the m-parameter families of invariant m-tori have $(m-k)$-parameter subfamilies where bifurcations of co-dimension $k \leq m$ occur. Such bifurcations are not restricted to those of *semi-local* type, but may also involve coinciding stable and unstable manifolds of different invariant tori. For instance, heteroclinic orbits between hyperbolic $(n-1)$-tori form $(2n-2)$-dimensional submanifolds of the phase space.

Let $F_1, \ldots, F_r \in C^\infty(\mathcal{P})$ be functionally independent integrals in involution. Fixing a point $z_0 \in \mathcal{P}$, the orbit $\mathbb{T}^r(z_0)$ of the $\mathbb{T}^r$-action generated by

$$X_{F_1}, \ldots, X_{F_r} \tag{25}$$

is an r-torus to which the vector fields (25) are tangent. Hence, the symplectic structure ω vanishes on the manifold $\mathbb{T}^r(z_0) \subseteq \mathcal{P}$ whence this torus is isotropic, implying $r \leq n$. Consequently, if a Hamiltonian system $\mathrm{X}H$ on $\mathcal{P}$ is *superintegrable*, having more functionally independent integrals of motion than degrees of freedom, these cannot be all in involution. One therefore also speaks of non-commutative integrability in this context.

The "extra" integral makes superintegrable systems even more exceptional than integrable systems, although it usually can be attributed to a non-commutative symmetry group. An example is the Euler case of a free rigid body for which the four integrals are the energy and the three components of the angular momentum mapping (14) induced by the symmetry group $SO(3)$. For the planar and spatial two body problem (the Kepler system) the symmetry groups $SO(3)$ and $SO(4)$ lead to 3 and 5 independent integrals of motion, respectively. The latter is exceptional *even within the class of superintegrable systems*; replacing the inverse square attraction by another central force (not the harmonic oscillator with its 5 independent integrals of motion due to an $SU(3)$-symmetry) breaks the symmetry down to $SO(3)$ with still 4 independent integrals of motions.

Theorem (Nekhoroshev, Mishchenko and Fomenko) *On the subset $\mathcal{R} \subseteq \mathcal{P}$ let $F\colon \mathcal{R} \longrightarrow \mathbb{R}^{2n-r}$ be a submersion with compact and connected fibers (hence, a fibration). Assume that $\{F_i, F_j\} = P_{ij} \circ F$, $i, j = 1, \ldots, 2n-r$ and that the matrix P with entries $P_{ij}\colon \mathcal{P} \longrightarrow \mathbb{R}$ has rank $2(n-r)$ at all points of $F(\mathcal{P})$. Then every fiber of F is diffeomorphic to $\mathbb{T}^r$ and the fibration F has local trivializations which are symplectic.*

This formulation is taken from [11], where the geometric contents of this theorem is thoroughly described in terms

of this fibration and an associated co-fibration with fibers of dimension $2n - r$.
Thus, every fiber of F has a neighborhood $\mathcal{U}$ with co-ordinates

$$(x, y, q, p) : \mathcal{U} \longrightarrow \mathbb{T}^r \times \mathbb{R}^r \times \mathbb{R}^{n-r} \times \mathbb{R}^{n-r}$$

such that the level sets of F coincide with the level sets of (y, q, p) and

$$\sigma_{|\mathcal{U}} = \sum_{i=1}^{r} \mathrm{d}x_i \wedge \mathrm{d}y_i + \sum_{j=1}^{n-r} \mathrm{d}q_j \wedge \mathrm{d}p_j \,.$$

These co-ordinates are Nekhoroshev's generalized action-angle variables. Where superintegrability is due to a non-commutative symmetry group G, the $2(n - r)$ parameters "live" in the co-adjoint orbits (17).

Perturbation Analysis

An important property of integrable Hamiltonian systems is their behavior under small perturbations. For a satisfactory description of e. g. mechanical systems the simplifying assumptions used to derive the model should not completely change the dynamics, some kind of "robustness" is desirable. An instance where the underlying approximation by a "simpler" system is part of the mathematical treatment is normal form theory.

Using a series expansion of the Hamiltonian function, the aim of normalization is to find co-ordinates in which the terms of the expansion look particularly simple (whence the Hamiltonian vector field takes a particularly simple form as well). This is an algorithmic procedure that inductively pushes a torus symmetry through the series. While the resulting series are typically divergent, a well-chosen truncation yields a normalized approximation for which good estimates are available. Already existing symmetries are preserved and the choices to be made when normalizing a given Hamiltonian are largely dictated ensuring that the combination of acquired and inherited symmetries render the system integrable. This does not always work, the integrability of normal forms is automatic only in two degrees of freedom, cf. [29]. For more details on normal forms in perturbation theory see the entry ► Normal Forms in Perturbation Theory in this encyclopedia.

Normalization often results in a torus action (24) on the phase space $\mathcal{P}$, for which the components $F_2, \ldots, F_n$ of the momentum mapping

$$J : \mathcal{P} \longrightarrow \mathbb{R}^{n-1}$$

define Hamiltonian vector fields (23) with periodic flows. This allows for the detailed description given in the previous section, and the question is what remains of this description when the symmetry (24) is broken.

The most prominent part of the ramified torus bundle defined by an integrable Hamiltonian system are the families of Lagrangean tori. Persistence of invariant tori can only be expected if these have a dynamical meaning. For instance, an invariant 1-torus that consists of a union of equilibria instead of being a periodic orbit is highly unlikely to remain present in a perturbed system. Similarly, an invariant torus with conditionally periodic motion has dynamical meaning if it is the closure of a dense orbit. This excludes resonances

$$\langle k \mid \varpi \rangle = 0 \,, \quad k \in \mathbb{Z}^{m \setminus \{0\}}$$

between the frequencies $\varpi_1, \ldots, \varpi_n$ of the parallel flow on the invariant torus T. The parallel nature of the flow implies that for a non-resonant frequency vector ϖ the time average

$$\int_{-\infty}^{+\infty} f(x(t))\, \mathrm{d}t = \int_{\mathsf{T}} f(x)\, \mathrm{d}x$$

of some function $f : \mathsf{T} \longrightarrow \mathbb{R}$ along the quasi-periodic motion $x(t) = x + t\varpi$ is equal to the space average over the torus.

This space average is approximated "quickly" – taking the time average over finite intervals of time – if ϖ is *Diophantine*, satisfying the strong non-resonance condition

$$|\langle k \mid \varpi \rangle| \geq \frac{\gamma}{|k|^\tau} \quad \text{for all} \quad k \in \mathbb{Z}^{m \setminus \{0\}} \tag{26}$$

with constants $\gamma > 0$ and $\tau > n - 1$. The set $\mathbb{R}^n_{\tau,\gamma}$ of (τ, γ)-Diophantine $\varpi \in \mathbb{R}^n$ has the local structure

$$\mathbb{R} \times \textit{Cantor dust}\,; \tag{27}$$

for $\varpi \in \mathbb{R}^n_{\tau,\gamma}$ also $s\varpi \in \mathbb{R}^n_{\tau,\gamma}$ for all $s \geq 1$, and the intersection $S^{n-1} \cap \mathbb{R}^n_{\tau,\gamma}$ is perfect and totally disconnected. While (Lebesgue)-almost all frequency vectors are non-resonant, the complement of Diophantine frequency vectors is an open and dense set. Still, the relative measure of $\mathbb{R}^n_{\tau,\gamma}$ goes to 1 as $\gamma \to 0$. The celebrated KAM theorem yields persistence of Lagrangean tori with Diophantine frequency vector.

Theorem (Kolmogorov, Arnol'd and Moser) *Let $\mathbb{Y} \subseteq \mathbb{R}^n$ be an open neighborhood of the origin and consider the phase space $\mathcal{P} = \mathbb{T}^n \times \mathbb{Y}$ with symplectic structure $\mathrm{d}x \wedge \mathrm{d}y$. Let the Hamiltonian*

$$H_\varepsilon(x, y) = H_0(y) + \varepsilon H_1(x, y; \varepsilon)$$

be real analytic and non-degenerate, satisfying

$$\det D^2 H_0(y) \neq 0 \quad \textit{for all} \quad y \in \mathbb{Y} \,. \tag{28}$$

Then there exists $\varepsilon_0 > 0$ such that for all $|\varepsilon| < \varepsilon_0$ there is a canonical transformation ϕ_ε near the identity and a measure-theoretically large set $\mathbb{Y}'_\varepsilon \subseteq \mathbb{Y}$ with the property that for $p \in \mathbb{Y}'_\varepsilon$ the transformed Hamiltonian $H_\varepsilon \circ \phi_\varepsilon$ does not depend on $q \in \mathbb{T}^n$.

For a proof see [3,7,25].
The Kolmogorov non-degeneracy condition (28) expresses that the frequency mapping

$$\begin{aligned} \varpi : \mathbb{Y} &\longrightarrow \mathbb{R}^n \\ y &\mapsto DH_0(y) \end{aligned} \tag{29}$$

is locally a diffeomorphism. This allows to pull back the whole geometry defined by (26) into phase space to obtain

$$\mathbb{Y}_\varepsilon = \varpi^{-1}(\mathbb{R}^n_{\tau,\gamma})$$

(and from this the subset $\mathbb{Y}'_\varepsilon \subseteq \mathbb{Y}_\varepsilon$ by omitting points γ-close to the boundary $\partial\mathbb{Y}$). The constant γ is chosen as a function of ε – of order $\mathcal{O}(\sqrt{\varepsilon})$ – to find an optimal balance between the (relative) measure of $\mathbb{Y}'_\varepsilon$ and the deviation of ϕ_ε from the identity.

The subset $\mathbb{T}^n \times \mathbb{Y}'_\varepsilon \subseteq \mathcal{P}$ consists (in the transformed variables) of quasi-periodic tori since $\dot{p}$ vanishes for $p \in \mathbb{Y}'_\varepsilon$. The theorem makes no statement for $p \in \mathbb{Y}^{\setminus \mathbb{Y}'_\varepsilon}$. For $n \geq 3$ this leaves the possibility of Arnol'd diffusion, trajectories that venture off to distant parts of the phase space.

While the KAM theorem concerns the fate of "most" trajectories and for all times, the complementary Nekhoroshev theory concerns all trajectories and states that they stay close to the unperturbed tori for times of the order

$$\exp\left[\left(\frac{\varepsilon_0}{\varepsilon}\right)^{\frac{1}{2n}}\right] .$$

Here analyticity of the Hamiltonian is a necessary ingredient, for finitely differentiable Hamiltonians one only obtains polynomial times. In the above formulation of the KAM theorem the assumption of analyticity of the Hamiltonian can be weakened without essential changes for the result, during the proof one merely has to intersperse an analytic approximation at each iteration step. The diffusion is even superexponentially slow for trajectories starting close to surviving tori, see [30] for more details on this phenomenon of exponential condensation, and see the entry ► Nekhoroshev Theory in this encyclopedia for more details on Nekhoroshev theory.

Where the energy level sets are transversal to the continuous direction in $\mathbb{Y}'_\varepsilon$ one has persistence of most Lagrangean tori on each energy shell, parametrized by *Cantor dust*. The same result is obtained under the condition of iso-energetic non-degeneracy

$$\det \begin{pmatrix} D^2 H_0(y) & \nabla H_0(y) \\ DH_0(y) & 0 \end{pmatrix} \neq 0 \,, \tag{30}$$

which is independent of Kolmogorov's condition. It is generic for an integrable system to satisfy both conditions almost everywhere. However, in applications it is a non-trivial task to actually check this and to determine the hypersurfaces in action space where these determinants vanish.

In two degrees of freedom the energy shells $\{H = h\}$ are 3-dimensional invariant manifolds for regular values h of the Hamiltonian and are separated by each Lagrangean torus. Thus, for initial conditions (q_0, p_0) in the co-ordinates provided by the KAM theorem with $p_0 \notin \mathbb{Y}'_\varepsilon$ the persistent tori parametrized by $\mathbb{Y}'_\varepsilon$ still have dynamical consequences as the trajectory is confined between two such tori, so $|p(t) - p_0|$ admits a bound of order $\mathcal{O}(\sqrt{\varepsilon})$. As a consequence, one obtains (dynamical) stability for all elliptic equilibria in generic Hamiltonian systems with two degrees of freedom.

Indeed, if the Hessian of (8) is (positive or negative) definite, then the Hamiltonian serves as a Liapunov function. In the indefinite case one also includes third and fourth order terms in the analysis and passes to a Birkhoff normal form

$$\alpha I_1 \;+\; \varphi I_2 + \tfrac{\beta}{2} I_1^2 \;+\; \delta I_1 I_2 \;+\; \tfrac{\chi}{2} I_2^2 \tag{31}$$

with $I_1 = (1/2)(p_1^2 + q_1^2)$ and $I_2 = (1/2)(p_2^2 + q_2^2)$. This can be achieved if there are no low order resonances $\varpi = -k\alpha$, $k = 1, 2, 3$ between the two (normal) frequencies of the equilibrium. A second genericity condition on the Hamiltonian is that the linear part of (31) does not divide the quadratic part (in the I_i), ensuring that (30) holds in a whole neighborhood of the elliptic equilibrium.

The Cantor set structure defined by the Diophantine conditions (26) can be used to weaken the necessary non-degeneracy condition. Since the gaps are defined by linear inequalities, the conditions on the first derivatives of the frequency mapping (29) can be replaced by conditions on the curvature or even higher derivatives. Such Rüssmann-like conditions still guarantee that the relative measure of surviving tori tends to 1 as the perturbation strength tends to zero, but at a price. For instance, the highest derivative $L \in \mathbb{N}$ needed in

$$\left\langle \frac{\partial^{|\ell|}\omega}{\partial y} \;\middle|\; |\ell| \leq L \right\rangle = \mathbb{R}^n \tag{32}$$

enters the Diophantine conditions on the frequency vector by means of the inequality $\tau > nL - 1$ on the Diophantine constant τ. For more details see the entry ▶ Kolmogorov–Arnold–Moser (KAM) Theory in this encyclopedia.

The KAM theorem is a semi-local result, valid in the neighborhood of an initial torus that admits action angle variables. A global version is obtained in [5]. The global conjugacy is glued together from convex combinations of local conjugacies using a partition of unity, the key ingredient being a unicity result on the tori obtained in the KAM theorem.

In the integrable approximation the distribution of the n-parameter families of Lagrangean tori is determined by the singular part S of the ramified torus bundle $\mathcal{P}$. Since S consists of families of lower dimensional tori together with their (un)stable manifolds, the persistence of (isotropic) m-tori, $m < n$, becomes important. For $m = 0$ the persistence of equilibria, together with their linear behavior (a superposition of what is possible in one and two degrees of freedom) follows from the implicit mapping theorem as it is generic for the Hamiltonian function that no equilibrium has vanishing (and neither multiple) eigenvalues. In the periodic case $m = 1$ the Diophantine condition (26) ensures $\varpi \neq 0$ and the 1-parameter families of periodic orbits persist as well, together with occurring bifurcations. In 1-parameter families all bifurcations are generically of co-dimension 1 – a genericity condition on the Hamiltonian. Bifurcations of higher co-dimension would not be expected to persist.

For hyperbolic tori the criteria remain valid almost verbatim; the key step is to pass to a center manifold. A technical difficulty is that even for analytic Hamiltonians center manifolds may only be of finite differentiability. While KAM theorems remain true in this context, the analytic context has its advantages – for instance (32) is satisfied for some $L \in \mathbb{N}$ for an analytic frequency mapping ϖ if and only if im ϖ does not lie within a linear hyperplane. An alternative is therefore to prove persistence of hyperbolic tori directly, this also gives a more direct hold on their stable and unstable manifolds.

Elliptic $(n-1)$-tori need one extra parameter to control the normal frequency as well. Similar to the isoenergetic case one can use time re-parametrization and obtain Cantor families of persistent elliptic $(n-1)$-tori parametrized by Cantor dust. Where there are more than one normal frequency to control this can no longer be done in a linear way; a problem solved by Rüssmann-like conditions on the higher derivatives of the frequency vector, see [7] and references therein. In case the mapping of internal frequencies satisfies Kolmogorov's condition, the higher order derivatives are only needed of normal frequencies. Now normal frequencies α_j enter the Diophantine conditions

$$|2\pi\langle k \mid \varpi\rangle \;+\; \langle \ell \mid \alpha\rangle| \geq \frac{\gamma}{|k|^{\tau}} \tag{33}$$

only as combinations $\langle \ell \mid \alpha\rangle$ with $|\ell| \leq 2$. This allows to extend the result to finite-dimensional elliptic tori in infinitely many degrees of freedom, cf. [14,15] and the entry ▶ Perturbation Theory for PDEs in this encyclopedia. For hypo-elliptic tori one may deal with the hyperbolic part by means of a center manifold or use a direct approach. Such m-tori have k additional pairs of purely imaginary Floquet exponents and excitation of normal modes leads for $l = 1, \ldots, k$ to $(m+l)$-parameter families of $(n+l)$-tori inheriting the "remaining" normal linear behavior, see [30] and references therein.

Where (lower-dimensional) m-tori undergo a semi-local bifurcation the m actions y *conjugate* to the toral angles x first of all have to versally unfold the bifurcation scenario. It is generic for the integrable Hamiltonian H that the m-parameter families of m-tori, $1 \leq m \leq n-1$, do not encounter bifurcations of co-dimension higher than m, so this is possible. The curvature of the frequency mapping is then used to ensure Diophanticity of most bifurcating tori, i. e. a Rüssmann-like condition with $L = 2$ is sufficient, cf. [13]. This curvature requirement is not necessary for 2-tori; these may undergo the quasi-periodic analogues of the co-dimension one bifurcations of periodic orbits as co-dimension two bifurcations are isolated within these 2-parameter families and cannot be prevented to disappear in resonance gaps.

Small perturbations of an integrable Hamiltonian thus lead to a Cantorification of the ramified n-torus bundle as sketched in Fig. 6, the stratification of the action space into various subfamilies parametrizing the tori is replaced by a Cantor stratification. Of equal importance are those changes that make sure that the non-integrable perturbed dynamics is indeed qualitatively different from the integrable unperturbed dynamics. While the former persistence results are obtained upon genericity conditions on the unperturbed system, such changes require the perturbation to be generic.

Disintegrating Lagrangean tori lead to invariant m-tori, where $n - m$ is the number of independent resonances $\langle k \mid \varpi\rangle = 0$ of the (internal) frequencies. Most of these lower dimensional tori will be elliptic or hyperbolic. The new hyperbolic tori lie at the basis of a possible scenario for Arnol'd diffusion. One of the effects of a small generic perturbation is that stable and unstable manifolds of hyperbolic periodic orbits no longer coincide, but split and intersect transversely, cf. [1]. This caries over to hy-

Dynamics of Hamiltonian Systems, Figure 6
A typical decomposition of the action space of a nearly integrable system in three degrees of freedom. The 2-dimensional Cantor dust parametrizes elliptic 2-tori that vanish in quasi-periodic center-saddle bifurcations along the fold lines. Between these extend arcs of hyperbolic 2-tori, parametrized by a 2-dimensional set of the form (27). The 1-dimensional Cantor dust along the folds consists of Lebesgue density points of these 2-dimensional parameter sets in the same way that these consist of Lebesgue density points of 3-dimensional sets of the form (27) above and below the surface which parametrize Lagrangean 3-tori

perbolic tori. The splitting of separatrices also leads to transverse intersections of stable and unstable manifolds of neighboring hyperbolic tori in the same energy shell. These hyperbolic tori form a Cantor family, and one of the main problems is to make sure that the transition chain of hyperbolic tori and their heteroclinic connections bridges the occurring gaps, cf. [9]. The dynamics in the gaps of Cantor families of hyperbolic tori can already be studied in the perturbation near resonant singular fibers of the ramified n-torus bundle. On the center manifold these become again (resonant) regular fibers, but the full perturbed motion is superposed by the hyperbolic dynamics in the symplectic normal directions.

The nature of the gaps opened by violations of (33) in families of elliptic tori is twofold. Internal resonances $\langle k \mid \varpi \rangle = 0$ lead to the destruction of the torus. Boundary points of the gaps resulting from normal-internal resonances are related to quasi-periodic bifurcations. The resonance $\alpha = 2\pi \langle k \mid \varpi \rangle$ triggers off a (quasi-periodic) center-saddle bifurcation and resonance gaps

$$|2\pi \langle k \mid \varpi \rangle + 2\alpha| < \frac{\gamma}{|k|^{\tau}}$$

are completely filled by hyperbolic tori that terminate in frequency halving bifurcations.

The maximal tori of superintegrable systems are m-tori in $n > m$ degrees of freedom and their normal behavior vanishes, both linearly and non-linearly. The strategy when studying a perturbation of such a *properly degenerate* system is to find an intermediate system that is also integrable, but non-degenerately so. The perturbation H_1 of a superintegrable Hamiltonian H_0 *removes the degeneracy* if the perturbed Hamiltonian $H_\varepsilon = H_0 + \varepsilon H_1$ can be written in the form

$$H = H_0 + \varepsilon \tilde{H}_1 + \varepsilon^2 H_2$$

where $H_0 + \varepsilon \tilde{H}_1$ is a non-degenerate integrable Hamiltonian. Since $\tilde{H}_1$ is defined in terms of H_1 (e. g. as its average along the unperturbed flow defined by H_0) every genericity condition on the intermediate system puts genericity conditions on the perturbation H_1. This first step of a normal form procedure lies also at the basis of Nekhoroshev theory for superintegrable systems, see [11] for more details.

Future Directions

Many Hamiltonian systems modeling real phenomena have symmetries, and the conditions of the regular reduction theorem of Marsden and Weinstein are often not fulfilled. The regularity assumptions could be successfully removed, cf. [8,24,27], and progress is still made in weakening the compactness conditions. Since the flow $\varphi\colon \mathbb{R} \times \mathcal{P} \longrightarrow \mathcal{P}$ itself is also a group action some condition has to exclude too general situations. It is for this reason that the symmetry (9) is studied in Sect. "Symmetry Reduction" and not the larger symmetry (10).

In an integrable system, action angle variables define a $\mathbb{T}^n$-action in the neighborhood of a given Lagrangean torus. Globally one only has the $\mathbb{R}^n$-action defined by the commuting flows of the integrals $F_1, \ldots, F_n$. The flow of $\mathrm{X}\,H = X_{F_1}$ is the actual object of study, and in general the flows of the vector fields (23) may be as complicated.

On the topological level, the $\mathbb{T}^n$-bundle $\mathcal{R}$ has a monodromy mapping $\mathcal{M}$ and the result of [10] explained in Sect. "Integrable Systems" allows one to uniquely characterize $\mathcal{R}$ by its Chern class only if $\mathcal{M}$ is trivial, such bundles are isomorphic if and only if their Chern classes coincide. An extension of this characterization to the case of non-trivial monodromy has been derived in [18,19]. The topological bundle $\mathcal{R}$ is uniquely characterized by its affine structure and Chern class, and the symplectic bundle $\mathcal{R}$ by its affine structure and Lagrange class.

The global version of the KAM theorem provides a Cantorification of the global bundle $\mathcal{R}$. This makes also non-local properties subject to the perturbation analysis, for instance showing a discrete invariant like monodromy to persist. Globally on the phase space, perturbation not only of the Hamiltonian H but also of the symplectic struc-

ture ω becomes a well-defined problem. For instance, if $\omega = -\mathrm{d}\vartheta$ is exact one may add a small non-exact closed 2-form $\varepsilon\sigma$ whence $\omega + \varepsilon\sigma$ is a non-exact symplectic form on the phase space.

A Lagrangean torus with $n-1$ independent resonances consists of periodic orbits. When the torus breaks up under the perturbation, only finitely many of these are expected to survive. At the same time the trivial normal behavior of these periodic orbits changes, resulting in hyperbolic and elliptic periodic orbits. The latter can serve as starting points for the construction of *solenoids*, cf. [4,21]. This construction should carry over to elliptic tori, where the "encircling" tori emerge from normal-internal resonances and might also result in solenoids that are limits of tori with varying dimension.

The results in [13] address persistence of Diophantine tori involved in a bifurcation and the corresponding gaps trigger off new phenomena, cf. [17]. Internally resonant tori involved in a quasi-periodic bifurcation may result in large dynamical instabilities, especially where multiple parabolic resonances are encountered. The effect is further amplified for tangent (or flat) parabolic resonances, which fail to satisfy the iso-energetic non-degeneracy condition.

The high co-dimensions of bifurcations that may be encountered within families of isotropic tori makes it necessary to study Hamiltonian bifurcations that have been left aside since they generically do not occur for periodic orbits. The coupling of the three types of co-dimension one bifurcations is unavoidable where the resonance gaps defined by (33) with $\ell \neq 0$ intersect. Already for an equilibrium with linearization

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

the nearby dynamics is extremely complicated, with all possible resonances of equilibria in two degrees of freedom occurring in a versal unfolding.

A perturbed superintegrable system can lead to the combination of two bifurcations in both the fast and the slow dynamics. With two different time scales e. g. the dynamics triggered off by two simultaneous violations of (33) appears to be of $(1+1)$-degree-of-freedom rather than having truly 2 degrees of freedom. This might render this problem more accessible.

For more details on Hamiltonian perturbation theory see the entry ► Hamiltonian Perturbation Theory (and Transition to Chaos) in this encyclopedia and references therein. An important subject are variational methods, which can be used to obtain periodic solutions (but also more complicated dynamics); see the entry ► Periodic Orbits of Hamiltonian Systems on this subject and also the entry ► n-Body Problem and Choreographies for a striking application to celestial mechanics.

Acknowledgments

The author thanks Henk Broer, Marius Crainic, Hans Duistermaat, Luuk Hoevenaars, Hessel Posthuma and Ferdinand Verhulst for fruitful discussions and helpful remarks.

Bibliography

Primary Literature

1. Abraham R, Marsden JE (1978) Foundations of Mechanics, 2nd edn. Benjamin
2. Arnol'd VI (1989) Mathematical Methods of Classical Mechanics, 2nd edn. Springer
3. Arnol'd VI, Avez A (1968) Ergodic problems of classical mechanics. Benjamin
4. Birkhoff BD (1935) Nouvelles recherches sur les systèmes dynamiques. Mem Pont Acad Sci Novi Lyncaei 3(1):85–216
5. Broer HW, Cushman R, Fassò F, Takens F (2007) Geometry of KAM tori for nearly integrable Hamiltonian systems. Ergod Theory Dyn Syst 27(3):725–741
6. Broer HW, Dumortier F, van Strien SJ, Takens F (1991) Structures in dynamics. Finite-dimensional deterministic studies. Stud Math Phys 2
7. Broer HW, Huitema GB, Sevryuk MB (1996) Quasi-Periodic Motions in Families of Dynamical Systems: Order amidst Chaos. Lect Notes Math, vol 1645. Springer, Berlin
8. Cushman RH, Bates LM (1997) Global Aspects of Classical Integrable Systems. Birkhäuser
9. Delshams A, de la Llave R, Martínez-Seara T (2006) A Geometric Mechanism for Diffusion in Hamiltonian Systems. Overcoming the Large Gap Problem: Heuristics and Rigorous Verification on a Model. Mem AMS 179(844):1–141
10. Duistermaat JJ (1980) On global action-angle coordinates. Comm Pure Appl Math 33:687–706
11. Fassò F (2005) Superintegrable Hamiltonian systems: Geometry and Perturbation. In: Gaeta G (ed) Symmetry and Perturbation Theory, Cala Gonone 2004. Acta Appl Math 87:93–121
12. Goldstein H (1980) Classical Mechanics. Addison-Wesley
13. Hanßmann H (2007) Local and Semi-Local Bifurcations in Hamiltonian Dynamical Systems – Results and Examples. Lect Notes Math, vol 1893. Springer, Berlin
14. Kappeler T, Pöschel J (2003) KdV & KAM. Erg Math Grenzgeb 3(45)
15. Kuksin SB (1993) Nearly integrable infinite-dimensional Hamiltonian systems. Lect Notes Math, vol 1556. Springer, Berlin
16. Landau LD, Lifshitz EM (1960) Mechanics. In: Course of Theoretical Physics 1. Pergamon
17. Litvak-Hinenzon A, Rom-Kedar V (2002) Parabolic resonances in 3 degree of freedom near-integrable Hamiltonian systems. Phys D 164:213–250
18. Lukina O (to appear) Torus bundles in integrable Hamiltonian systems. Ph D thesis. Rijksuuniversiteit Groningen, Groningen

19. Lukina O, Takens F, Broer HW (2008) Global properties of integrable Hamiltonian systems. Reprint, Rijksuuniversiteit Groningen, Groningen
20. MacKay RS (1993) Renormalisation in Area-preserving Maps. World Scientific
21. Markus L, Meyer KR (1980) Periodic orbits and solenoids in generic Hamiltonian dynamical systems. Am J Math 102(1):25–92
22. Marsden JE, Raţiu TS (1994) Introduction to Mechanics and Symmetry. Springer
23. Meyer KR, Hall GR (1992) Introduction to Hamiltonian Dynamical Systems and the *N*-Body Problem. Appl Math Sci 90
24. Montaldi J, Raţiu TS (2005) Geometric Mechanics and Symmetry: the Peyresq Lectures. Cambridge Univ Press
25. Moser J (1973) Stable and Random Motions in Dynamical Systems: With Special Emphasis on Celestial Mechanics. Princeton Univ Press
26. Moser J, Zehnder EJ (2005) Notes on Dynamical Systems. Courant Lect Notes Math 12, AMS
27. Ortega J-P, Raţiu TS (2004) Momentum Maps and Hamiltonian Reduction. Prog Math 222
28. Raţiu TS, Tudoran R, Sbano L, Sousa Dias E, Terra G (2005) A Crash Course in Geometric Mechanics. In: Montaldi J, Raţiu TS (eds) Geometric Mechanics and Symmetry: the Peyresq Lectures. Cambridge Univ Press, pp 23–156
29. Sanders JA, Verhulst F, Murdock J (2007) Averaging Methods in Nonlinear Dynamical Systems, 2nd edn. Springer
30. Sevryuk MB (2003) The classical KAM theory at the dawn of the twenty-first century. Mosc Math J 3(3):1113–1144
31. Siegel CL, Moser JK (1971) Lectures on Celestial Mechanics. Springer

Books and Reviews

Arnol'd VI, Givental' AB (2001) Symplectic geometry. In: Novikov SP (ed) Dynamical systems IV, Symplectic geometry and its applications. Springer, pp 1–138

Arnol'd VI, Kozlov VV, Neishtadt AI (1988) Mathematical Aspects of Classical and Celestial Mechanics. In: Arnol'd VI (ed) Dynamical Systems III. Springer

Arnol'd VI, Novikov SP (1994) Dynamical systems VII, Integrable Systems. Springer

Bartsch T, Szulkin A (2005) Hamiltonian systems: periodic and homoclinic solutions by variational methods. In: Cañada A, Drábek P, Fonda A (eds) Handbook of differential equations: ordinary differential equations, vol 2. Elsevier, pp 77–146

Bibikov YuN (1979) Local theory of nonlinear analytic ordinary differential equations. Lect Notes Math 702

Birkhoff BD (1966) Dynamical Systems. Am Math Soc (1927); reprint 1966

Bolsinov AV, Fomenko AT (2004) Integrable Hamiltonian Systems – Geometry, Topology, Classification. Chapman & Hall/CRC

Bridges TJ, Furter JE (1993) Singularity Theory and Equivariant Symplectic Maps. Lect Notes Math 1558

Broer HW, Hasselblatt B, Takens F (2007) Handbook of Dynamical Systems, vol 3. North-Holland (to appear)

Broer HW, Hoveijn I, Lunter G, Vegter G (2003) Bifurcations in Hamiltonian systems: Computing Singularities by Gröbner Bases. Lect Notes Math 1806

Broer HW, Sevryuk MB (2007) KAM Theory: quasi-periodicity in dynamical systems. In: Broer HW, Hasselblatt B, Takens F (eds) Handbook of Dynamical Systems, vol 3. North-Holland (to appear)

Buono P-L, Laurent-Polz F, Montaldi J (2005) Symmetric Hamiltonian Bifurcations. In: Montaldi J, Raţiu TS (eds) Geometric mechanics and symmetry. The Peyresq Lecture. Cambridge University Press, Cambridge

Cannas da Silva A, Weinstein A (1999) Geometric models for noncommutative algebras. Berkeley Math Lect Notes 10

Ciocci MC, Litvak-Hinenzon A, Broer HW (2005) Survey on dissipative KAM theory including quasi-periodic bifurcation theory. In: Montaldi J, Raţiu TS (eds) Geometric Mechanics and Symmetry: the Peyresq Lectures. Cambridge Univ. Press, pp 303–355

Cushman R (1992) A Survey of Normalization Techniques Applied to Perturbed Keplerian Systems. Dyn Rep new ser 1:54–112

Dubrovin BA, Krichever IM, Novikov SP (2001) Integrable systems I. In: Novikov SP (ed) Dynamical systems IV, Symplectic geometry and its applications. Springer, pp 177–332

Duistermaat JJ (1984) Bifurcations of periodic solutions near equilibrium points of Hamiltonian systems. In: Salvadori L (ed) Bifurcation Theory and Applications Montecatini 1983. Lect Notes Math 1057:57–105

Efstathiou K (2005) Metamorphoses of Hamiltonian systems with symmetries. Lect Notes Math 1864

Efstathiou K, Sadovskií D (2005) No Polar Coordinates. In: Montaldi J, Raţiu TS (eds) Geometric Mechanics and Symmetry: the Peyresq Lectures. Cambridge Univ. Press, pp 211–301

Eliasson LH, Kuksin SB, Marmi S, Yoccoz J-C (2002) Marmi S, Yoccoz J-C (eds) Dynamical Systems and Small Divisors, Cetraro 1998. Lect Notes Math 1784

Gallavotti G (1983) The elements of mechanics. Springer

Gallavotti G (1994) Twistless KAM tori, quasi flat homoclinic intersections, and other cancellations in the perturbation series of certain completely integrable Hamiltonian systems. A review. Rev Math Phys 6(3):343–411

Guillemin V, Sternberg S (1984) Symplectic techniques in physics. Cambridge Univ. Press

Guillemin V, Sternberg S (1990) Variations on a theme by Kepler. Am Math Soc Coll Pub 42

Haller G (1999) Chaos near Resonance. Appl Math Sci 138

Hitchin NJ, Segal GB, Ward RS (1999) Integrable systems. Twistors, loop groups, and Riemann surfaces, Oxford 1997. Clarendon/Oxford Univ. Press

Hofer H, Zehnder E (1994) Symplectic invariants and Hamiltonian dynamics. Birkhäuser

Holm DD (2005) The Euler-Poincaré variational framework for modeling fluid dynamics. In: Montaldi J, Raţiu TS (eds) Geometric Mechanics and Symmetry: the Peyresq Lectures. Cambridge Univ. Press, pp 157–209

Karasev MV, Maslov VP (1993) Nonlinear Poisson Brackets: Geometry and Quantization. Transl Math Monogr 119

Kuksin SB (2000) Analysis of Hamiltonian PDEs. OLS Math Appl 19

Libermann P, Marle C-M (1987) Symplectic Geometry and Analytical Mechanics. Reidel

Markus L, Meyer KR (1974) Generic Hamiltonian dynamical systems are neither integrable nor ergodic. Mem Am Math Soc 144:1–52

Marsden JE (1992) Lectures on Mechanics. Cambridge Univ. Press

Marsden JE, Misiołek G, Ortega JP, Perlmutter M, Raţiu TS (2007) Hamiltonian reduction by stages. Lect Notes Math 1913

Marsden J, Montgomery R, Raţiu TS (1990) Reduction, symmetry, and phases in mechanics. Mem Am Math Soc 88(436):1–110
Mawhin J, Willem M (1989) Critical Point Theory and Hamiltonian Systems, 2nd edn. Appl Math Sci 74
McDuff D, Salamon D (1995) Introduction to symplectic topology. Clarendon/Oxford Univ. Press
van der Meer JC (1985) The Hamiltonian Hopf bifurcation. Lect Notes Math 1160
Meyer KR (1999) Periodic solutions of the *N*-body problem. Lect Notes Math 1719
Mielke A (1991) Hamiltonian and Lagrangian Flows on Center Manifolds – with Applications to Elliptic Variational Problems. Lect Notes Math 1489
Novikov SP (2001) Dynamical systems IV, Symplectic geometry and its applications. Springer
Olshanetsky MA, Perelomov AM, Reyman AG, Semenov-Tian-Shansky MA (1994) Integrable systems II. In: Arnol'd VI, Novikov SP (eds) Dynamical systems VII, Integrable Systems. Springer, pp 87–259
Oxtoby JC (1971) Measure and category. A survey of the analogies between topological and measure spaces. Grad Texts Math 2
Poincaré H (1957) Les Méthodes Nouvelles de la Mécanique Céleste. Gauthier–Villars (1892,1893,1899); reprint (1957) Dover
Rabinowitz PH (2002) Variational methods for Hamiltonian systems. In: Hasselblatt B, Katok A (eds) Handbook of dynamical systems, vol 1A. North-Holland, pp 1091–1127
Rink B, Tuwankotta T (2005) Stability in Hamiltonian Systems. In: Montaldi J, Raţiu TS (eds) Geometric Mechanics and Symmetry: the Peyresq Lectures. Cambridge Univ. Press, pp 1–22
Souriau J-M (1970) Structure des systèmes dynamiques. Dunod. English edition: Souriau JM (1997) Structure of dynamical systems, a symplectic view of physics. Prog Math 149
Thirring W (1978) A course in mathematical physics I. Classical dynamical systems. Springer
Trofimov VV, Fomenko AT (1994) Geometric and Algebraic Mechanisms of the Integrability of Hamiltonian Systems on Homogeneous Spaces and Lie Algebras. In: Integrable systems II. In: Arnol'd VI, Novikov SP (eds) Dynamical systems VII, Integrable Systems. Springer, pp 261–333

Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation

Khalid Saeed
Worcester Polytechnic Institute, Worcester, USA

Article Outline

Glossary
Definition of the Subject
Introduction
The Alternative Economic Models and Their Limitations
A System Dynamics Model of Resource Allocation, Production and Entitlements
Replicating Income Distribution Patterns Implicit in Models of Alternative Economic Systems
Feedback Loops Underlying Wage and Income Patterns
Possibilities for Poverty Alleviation
Conclusion
Future Directions
Appendix
Bibliography

Glossary

Absentee owners Parties not present on land and capital resources owned by them.

Artisan owners Parties using own labor together with land and capital resources owned by them to produce goods and services.

Behavioral relations Causal factors influencing a decision.

Capital intensive A process or industry that requires large sums of financial resources to produce a particular good.

Capital Machinery equipment, cash and material inputs employed for the production of goods and services.

Capitalist sector A subeconomy in which all resources are privately owned and their allocation to production and renting activities is exclusively carried out through a price system.

Capitalist system An economic system in which all resources are in theory privately owned and their allocation to production and renting activities is exclusively carried out through a price system.

Commercial Pertaining to buying and selling with intent to make profit.

Controlling feedback A circular information path that counters change.

Corporate pertaining to a profit maximizing firm.

Economic dualism Side-by-side existence of multiple subeconomies.

Economic sector A collection of production units with common characteristics.

Entrepreneurship Ability to take risk to start a new business.

Feedback loops Circular information paths created when decisions change information that affects future decisions.

Financial market A mechanism that allows people to easily buy and sell commodities, financial instruments and other fungible items of value at low transaction costs and at prices that reflect efficient markets.

Household income Income accrued to a household from wages, profits and rents received by all its members.

Institutionalist economic models Models attributing performance of economies to institutional relationships and advocating selective government intervention to change the behavior that creates dysfunctions.

Iron law of wages David Ricardo's most well-known argument about wages "naturally" tending towards a minimum level corresponding to the subsistence needs of the workers.

Keynesian A belief that the total spending in the economy is influenced by a host of economic decisions – both public and private.

Labor intensive A process or industry with significant labor costs.

Labor productivity Output per worker or worker-hour.

Labor Economically active persons in an economy.

Marginal factor cost The incremental costs incurred by employing one additional unit of input.

Marginal revenue product The additional income generated by using one more unit of input.

Market economy An economy which relies primarily on interactions between buyers and sellers to allocate resources.

Marxist economic theory A theory highlighting exploitive mechanisms in an economic system and advocating central governance.

Marxist system A centrally run economic system emphasizing in theory the Marxist axiom "from each according to ability to each according to need".

Model An abstract representation of relationships in a real system.

Neoclassical economic theory A theory highlighting constructive market forces in an economic system and advocating consumer sovereignty and a price system as invisible sources of governance.

Non-linear A system whose behavior can't be expressed as a sum of the behaviors of its parts.

Opportunity cost Real value of resources used in the most desirable alternative, or the amount of one commodity foregone when more of another is consumed.

Ordinary differential equation A relation that contains functions of only one independent variable, and one or more of its derivatives with respect to that variable.

Output elasticity Change in output caused by addition of one unit of a production factor.

Perfect market A hypothetical economic system that has a large number of buyers and sellers – all price takers trading a homogeneous product – with complete information on the prices being asked and offered in other parts of the market; and with perfect freedom of entry to and exit from the market.

Political economy Interaction of political and economic institutions, and the political environment.

Production factor A resource input such as land, labor, or capital contributing to production of output.

Productivity The amount of output created (in terms of goods produced or services rendered) per unit input used.

Purchasing power parity The value of a fixed basket of goods and services based on the ratio of a countries' price levels relative to a country of reference.

Revisionist economic models Models recognizing both constructive and exploitive forces and advocating government intervention against exploitation.

Sector A collection of production units with common characteristics.

Self-employment Work for a self-owned production unit without a defined wage.

System dynamics A methodology for studying and managing complex feedback systems, such as one finds in business and other social systems.

Subeconomy A collection of production units and households with common characteristics.

Theories of value How people positively and negatively value things and concepts, the reasons they use in making their evaluations, and the scope of applications of legitimate evaluations across the social world.

Unearned income Income received as rents.

Wage employment Work for a defined wage.

Definition of the Subject

Poverty is perhaps the most widely written about subject in economic development, although there is little agreement over its causes and how to alleviate it. The undisputed facts about poverty are that it is pervasive, growing and that the gap between the rich and the poor is widening.

It is widely believed that the governments – irrespective of their ideological inclinations – have the responsibility to intervene to help the poor. Poverty alleviation is also the key mandate of International Bank for Reconstruction and Development (World Bank) and the many civil society organizations. World Bank places poverty line at purchasing power parity of $1 per day, which has improved a bit in terms of percentage below over the past three decades, except in Africa, but remains large in terms of head count. This threshold is however unrealistic since it targets absolutely basket cases. A poverty line at purchasing power parity of $3 per day, which is close to average purchasing power per capita in the poor countries

shows that both poverty head count and gap between rich and poor have been expanding across board. World Bank web site at http://iresearch.worldbank.org/PovcalNet/jsp/index.jsp allows making such computations for selected countries, regions, years and poverty lines.

Neoclassical economic theory does not explicitly address the process of income distribution among households, although it often views income distribution as shares of profits and wages. In most economic surveys and censuses, however, income distribution is invariably measured in terms of shares of various percentages of the households. The fact that more than 80% of the income is claimed by fewer than 20% of the households who also own most of the capital resources in almost all countries of the world, the theory and the measurement have some common ground. Neoclassical theory has, however, shed little light on the process of concentration of wealth and how can this dysfunction be alleviated.

System dynamics, although rarely used for the design of public policy for addressing poverty, allows us to construct and experiment with models of social systems to understand their internal trends and test policy combinations for changing them. In this paper I have used system dynamics modeling to understand the process of concentration of wealth and re-evaluate the on-going poverty alleviation effort.

The model, which subsumes resource allocation, production and entitlements, explains the many manifestations of income distribution in a market economy. It generates multiple patterns of asset ownership, wage and employment assumed in neo-classical, Marxist and revisionist perspectives on economic growth while it allows ownership to change through the normal course of buying and selling transactions based on rational though, information-bound criteria. Privately owned resources can be employed through hiring wage-labor, rented out or used for self-employment. In addition to the labor market conditions, the wage rate depends also on the opportunity cost of accepting wage employment as workers may be either self-employed or wage-employed. Since this opportunity cost varies with the capital resources owned by the workers, which may support self-employment, the wage rate is strongly affected by the distribution of ownership. Thus, ownership can become concentrated or widely distributed depending on legal and social norms governing transactions in the economy, which the model replicates. Extended experimentation with this model serves as a basis to identify critical policy instruments that make best use of the system potential for resource constrained growth and poverty alleviation through widening participation in the market and improving income distribution.

Introduction

The opening up of the major centrally planned economies of the world has brought to the fore problems concerning the psychological deprivation, inefficiencies of resource allocation and production, and the lack of dynamism experienced in the working of central planning in a socialist system. The accompanying enthusiasm for free market in a capitalist system has, however, hidden many of the dysfunctional aspects of this alternative. It should be recognized that both systems emerged from time-specific and geography-specific empirical evidence. Since their underlying models treat as given specific economic patterns, the institutional roles and the legal norms associated with each system have inherent weaknesses, which create dysfunctions when implemented in different environmental contexts [43,64]. Thus, neither model may furnish an adequate basis for the design of policies for sustainable economic development and poverty alleviation. A search is, therefore, necessary for an organizational framework that might explain the internal trends inherent in each model as special modes of a complex system subsuming the variety of behavioral patterns recognized by specific models before an effective policy for change can be conceived [52].

Using as an experimental apparatus a formal model of the decision structure affecting wage determination, saving and investment behavior, and the disbursement of income, presented earlier in [53], this paper seeks to identify the fundamental economic relations for creating a dynamic and sustainable market system that may also increase opportunities for the poor, whose market entry is often limited by their financial ability and social position [58], to participate in the economy and be entitled to the value it creates. System dynamics provides the technical framework to integrate the various behavioral relations in the system [13,63].

Notwithstanding the many objections to the abstract models of orthodox economics, which are difficult to identify in the real world [28,46], the model of this paper draws on neo-classical economics to construct a basic structure for growth and market clearing. This structure is, however, progressively modified by relaxing its simplifying assumptions about aggregation of sub-economies, wage determination, ownership, income disbursement, saving and investment behavior, financial markets, and technological differentiation between sub-economies to realize the many growth and income distribution patterns addressed in a variety of economic growth models.

The modified model I finally create represents a real world imperfect market in which expectations formed under bounded rational conditions govern the decisions of

the economic actors [59], as recognized in the pioneering works of Kaldor (1969) [24], Kalecki (1965) [25], Wientraub (1956) [66], and Joan Robinson (1978) [45]. The model also subsumes the concept of economic dualism first recognized by Boeke (1947) [7] and developed further by Lewis (1958) [29], Sen (1966) [57], Bardhan (1973) [4] and others to represent the multiple sub-economies that co-exist especially in the developing countries. Such a model is more identifiable with respect to the real world as compared with the time and geography specific concepts propounded by the various, often controversial, theories of economic growth.

Simulation experiments with the model explore entry points into the economic system for creating an egalitarian wage and income distribution pattern through indirect policy instruments. Also explored are the functions of entrepreneurship and innovation and the mechanisms that may increase the energy of those processes toward facilitating economic growth and alleviating poverty.

The Alternative Economic Models and Their Limitations

The economic models used as the bases for designing development policies over the past several decades have ascended largely from time-specific and geography-specific experiences rather than from a careful study of the variety of behavioral patterns occurring over various time periods and across several geographic locations. Among these, the socialist and the capitalist models are most at odds. They differ in their assumptions about ownership and income distribution patterns, the basis for wage determination, the influence of technology on income growth and the functions of entrepreneurship and innovation [21,55].

The neo-classical economic theory, which is the basis for the capitalist model, is silent on the ownership of capital resources, often assuming it in default to be widely distributed [5]. Thus, the labor-wage rate may bear little relationship to the income of households, who are also recipients of profits. It is assumed that private control of productive resources is a means for market entry, which creates unlimited potential for economic growth, although private investment is not often seen to be subject to self-finance due to the assumption that financial markets are perfect. The neo-classical economic theory also postulates that short-run labor-wage rates depend on labor market conditions, while in the long run, they are determined by the marginal revenue product of labor. Neo-classical models of economic growth, however, often make the simplifying assumption that equilibrium continues to prevail in both factor and product markets over the course of growth. Thus, only minor fluctuations may occur in wages, profits and prices in the short run, and these can be ignored.

The belief in the existence of such equilibrium is further strengthened by the Keynesian argument for the ineffectiveness of the market mechanisms due to the dependence of prices on long-term wage contracts and production plans which may not respond easily to short-run changes of the market. As a result of the belief in this theory of wage determination, technological choices that increase labor productivity are expected to have a positive effect on wage rates and household income, because they increase the marginal revenue product of labor. Entrepreneurship is viewed as important for new entry into economic activity, which is open to all, and innovation is supposed to benefit society through increased productivity. With these assumptions, the capitalist system advocates minimal government intervention in the economy. This model is widely presented in the many texts on economic development. Pioneering texts include Hirshleifer (1976) [22] and Kindelberger and Herrick (1977) [27].

Marxist economic theory, which underpins the socialist model, assumes on the other hand that ownership of capital resources is concentrated in a minority excluding the workers and that the majority of households receive no part of the profits. Thus, wage payments have a strong effect on household income. The Marxist theory views private ownership as a source of exploitation and postulates labor-wage rates determined by the consumption necessary for a worker to support production in a grossly labor surplus economy following Ricardo's iron law of wages [32,39]. The labor-wage rate is, thus, based on the real value of the commodities needed for a worker to subsist, which is more or less fixed, irrespective of the contribution of labor to the production process. In such conditions, technological choices that increase labor productivity may indeed only serve to increase the share of the surplus of product per unit of labor appropriated by the capitalists. In this model, entrepreneurship is viewed as an asocial activity and innovation seen to originate from the need to boost falling returns on capital. Attributing the development of these conditions to market failure, the socialist system assigns control of the economy to the government.

There also exist a number of revisionist models of political economy attempting to explain the nature of interdependence of the multiple sub-economies observed to co-exist in many developing countries in violation of the theoretical premises of the neo-classical model according to which all production factors must eventually move to the most efficient sector. These models of-

ten attribute the development of disparities between the various sub-economies to exploitative mechanisms that tend to maintain an upper hand of the stronger influence groups. The revisionist analyses have largely led to making moral appeals for the government policy to target the poor and the disadvantaged in its development effort, which is a stated mandate of the International Bank for Reconstruction and Development (World Bank). Targeting the poor has also been advocated widely by numerous development economists over the past half century. They include such prominent economists as Myrdal (1957) [36], Lipton (1977) [30], Galbraith (1979) [15], and Sen (1999) [58].

Indeed, each economic system can often be endorsed with the help of selected historical evidence, and this has been fully exploited to fuel the traditional debate between the neo-classical and Marxist economic schools. Interesting artifacts of this debate include the normative theories of value suggested by each system to justify the various wage systems, which have little practical significance for development policy [44,62]. This is unfortunate, since contradictions of evidence should clearly indicate the existence of fundamental organizational arrangements in the economic system, which are capable of creating the multiple behavior patterns on which the various economic models are based. Once identified, such arrangements may also serve as entry points for the design of evolutionary changes in an existing pattern. To quote Professor Joan Robinson:

> Each point of view bears the stamp of the period when it was conceived. Marx formed his ideas in the grim poverty of the forties. Marshal saw capitalism blossoming in peace and prosperities in the sixties. Keynes had to find an explanation for the morbid condition of 'poverty in the midst of plenty' in the period between the wars. But each has significance for other times, for in so far as each theory is valid, it throws light upon essential characteristics of the system which have always been present in it and still have to be reckoned with. [43]

Following sections of this paper experiment with a system dynamics model of an economic system, widely found in the developing countries and presented earlier in [53], to understand the variety of economic patterns experienced over time and geography under different legal and social norms. Furthermore, exploratory experimentation with this model helps to outline the basic principles of a market system that can sustain growth, create equitable distribution of benefits and facilitate innovation and productivity improvement, all widely deemed necessary for poverty alleviation.

A System Dynamics Model of Resource Allocation, Production and Entitlements

A system dynamics model subsuming the broad decision rules that underlie resource allocation, production, and income disbursement processes of a developing country economic system was proposed in Saeed (1988) [49] and further experimented with in Saeed (1994) [53]. In this model, capital, labor, and land (which may also be assumed as a proxy for natural resources) are used as production factors. Model structure provides for the functioning of two modes of production, commercial, in which resources are employed on the basis of their profitability and which is managed by the capitalist sector of the economy; and self-employed, in which workers not employed in the commercial mode make a living. These two modes of production have been referred to variously in the literature, for example as oligopolistic and peripheral firms [16], formal and informal sectors [29], and modern and traditional subeconomies [12].

It has been assumed in the model that all workers, whether self-employed using their own or rented capital resources or employed as wage-workers by the capitalist sector, are members of a homogeneous socio-economic group with a common interest, which is to maximize consumption. This group is also the sole supplier of labor in the economy since the small number of working capitalists is ignored. On the other hand, the capitalist sector is assumed to maximize profit while it is also the sole wage-employer in the economy [2,3,57].

It is assumed that private ownership is protected by law, but land and capital assets can be freely bought, sold and rented by their owners. Each buying and selling transaction between the two sectors must be accompanied by a corresponding transfer of the cash value of the assets determined by the going market prices. The model also permits the appearance of technological differences between the capitalist and the self-employed sectors, when more than one technologies embodied in the type of capital used (named traditional and modern in the model) are available and the two sectors cannot employ the preferred technology with equal ease [30,41], or when the self-employed sector is burdened by excess workers not employed by the commercial sector while it lacks the financial capacity to use its preferred technology.

Figure 1 shows how workers and capital might potentially be retained and employed by the two sectors in the model. Rectangles represent stocks, valve symbols

flows and circles intermediate computations following the diagramming convention of system dynamics modeling. The size of each sector is not specified and is determined endogenously by the model, depending on assumptions about the socio-technical environment in which the system functions.

The changes in the quantities of the production factors owned or employed by each sector are governed by the decisions of the producers and the consumers of output and by the suppliers of the production factors acting rationally according to their respective motivations within the bounds of the roles defined for them by the system [59]. The value of production is shared by households on the basis of the quantity of the production factors they contribute and the factor prices they can bargain for [10]. Income share of the workers, less any investment needed to

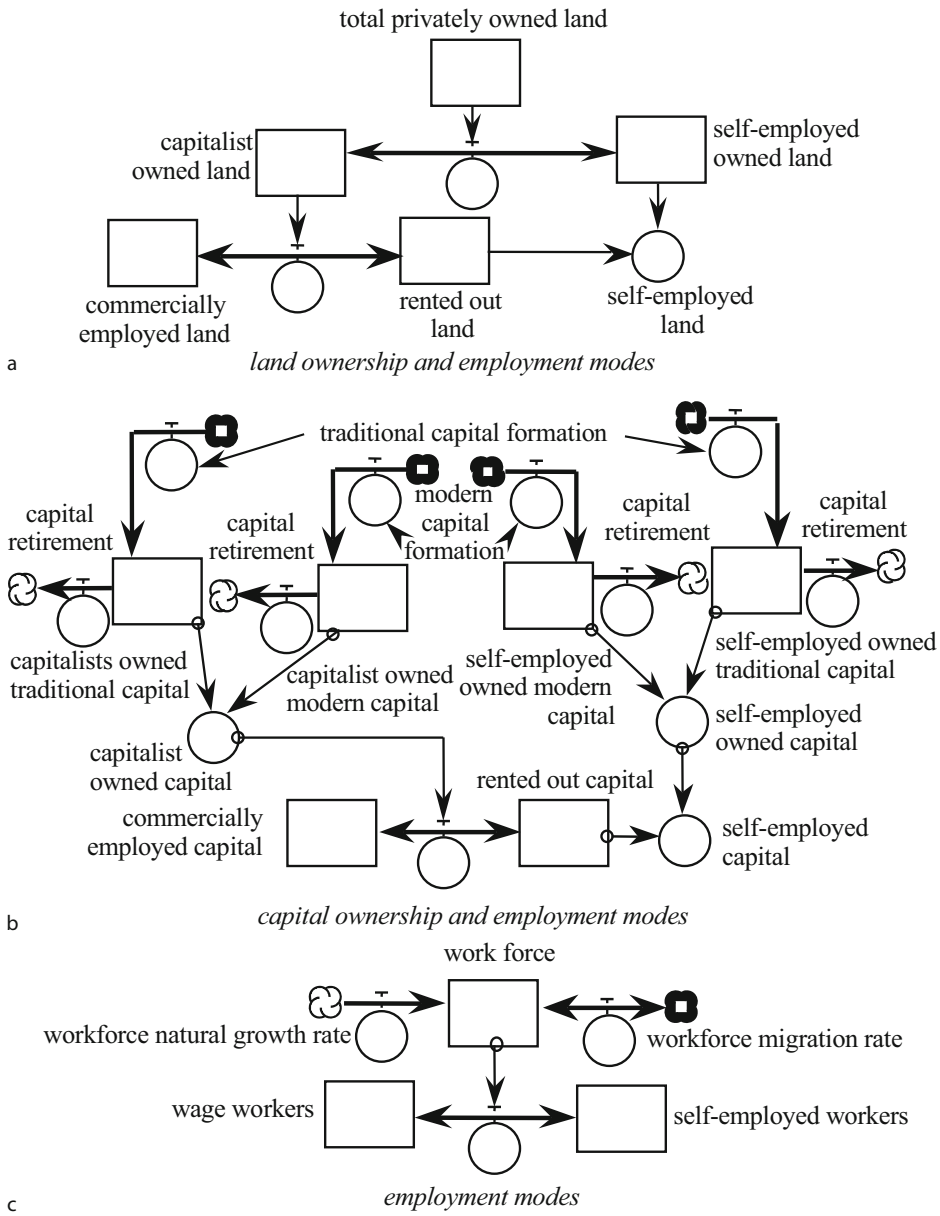

Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 1
Potential worker and capital distribution between capitalist and self employed sectors

maintain self-employment, divided by the total workforce, determines average consumption per worker, which represents the opportunity cost of accepting wage-employment and this is the basis for negotiating a wage [57,62].

Investment and saving rates in the two sectors are decoupled through a balance of internal savings. The financial markets are segmented by sectors and the investment decisions of a sector are not independent of its liquidity position, given by the unspent balance of its savings. Thus, investment decisions depend on profitability criteria, but are constrained by the balance of internal savings of each sector [33,34]. Figure 2 shows the mechanisms of income disbursement, saving and internal finance incorporated into the model.

The saving propensity of all households is assumed not to be uniform. Since capitalist households receive incomes that are much above subsistence, their saving propensity is stable. On the other hand, the saving propensity of the worker households depends on their need to save to support investment for self-employment and on how their absolute level of income compares with their inflexible consumption [23,26,31].

The broad mathematical and behavioral relationships incorporated into the model are given in the Appendix "Model Description". Technical documentation and a machine-readable listing of the model written in DYNAMO code are available from the author on request.

Replicating Income Distribution Patterns Implicit in Models of Alternative Economic Systems

The model is simulated under different assumptions about wages, rents, financial markets and technology and its behavior analyzed in relation to the various theoretical and empirical premises its information relationships represent.

As an arbitrary initial condition, production factors are equally divided between the two sectors and equilibrium in both product and factor markets is assumed to exist under the conditions of a perfect economic system as

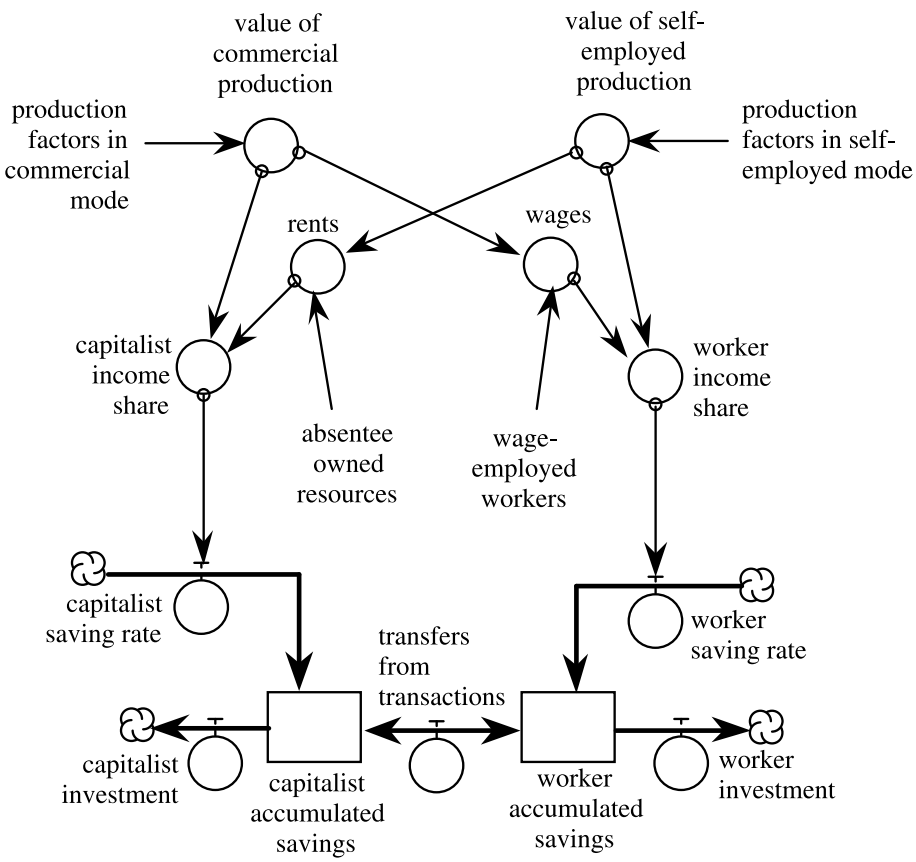

Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 2
Income disbursement process

described in neo-classical economics. Thus, the marginal revenue products of land and capital are initially assumed to be equal to their respective marginal factor costs determined by an exogenously specified interest rate which represents the general pattern of preferences of the community for current as against future consumption [22]. The marginal revenue product of workers is initially set equal to wage rate. The market is initially assumed to be clear and there is no surplus of supply or demand.

Replicating the Theoretical Neo-classical System

This experiment is aimed at understanding internal trends of a system representing the neo-classical economic theory. To transform the model to represent this system, it is assumed that the production factors employed by each sector are owned by it and no renting practice exists [5]. The wage rate is assumed to be determined by the marginal revenue product of workers and the availability of labor instead of the opportunity cost to the workers of supplying wage-labor. Financial markets are assumed to be perfect and investment decisions of the two sectors are uncoupled from their respective liquidity positions. It is also assumed that the technology of production is the same in the two sectors and, in terms of the model, only traditional capital is available to both of them. The only difference between the two sectors is that the capitalist sector can vary all production factors, including labor to come to an efficient mix, while the self-employed sector may absorb all labor not hired by the capitalist sector, while it can freely adjust other production factors to achieve an efficient mix.

The model thus modified stays in equilibrium when simulated as postulated in neo-classical economic theory. When this equilibrium is disturbed arbitrarily by transferring a fraction of the workers from the capitalist to the self-employed sector, the model tends to restore its equilibrium in a manner also similar to that described by the neo-classical economic theory. This is shown in Fig. 3.

The transfer raises the marginal revenue product of workers in the capitalist sector, which immediately proceeds to increase its workforce. The transfer also raises the intensity of workers in the self-employed sector as a result of which the marginal revenue products of land and capital in that sector rise. Hence, it proceeds to acquire more land and capital. These activities continue until the marginal revenue products of the factors and their proportions are the same in the two sectors. Note that while the factor proportions and marginal revenue products of the factors are restored by the model to their original values, the absolute amounts of the various factors are different when new equilibrium is reached. There is, however, no difference in endowments per worker between the capitalist and the self-employed sectors.

Since factor payments are determined purely on the basis of contribution to the production process while the quantities of production factors allocated to each sector depend on economic efficiency, the wages and factor allocations seem to be determined fairly and efficiently, as if by an invisible hand. Ownership in such a situation can either be communal or very widely distributed among households since otherwise the wage bargaining process will not lead to fair wages. Renting of production factors among households is irrelevant since transfer to parties who can efficiently employ them is automatic.

Before anything is said about the empirical validity of the simplifying assumptions made in this model, the historical context of these assumptions must be examined carefully. The simplified model is based on Adam Smith's description of an industrial economy observed at the start of the industrial revolution. This economy was run by artisan-turned capitalists and there were many of these capitalists competing with one another, although, none had the financial muscle to outbid the others except through his/her ability to employ resources efficiently [60].

As far as labor wage rate was concerned, although there were instances of exploitation of workers at a later stage of the industrial revolution, the artisan workers could obtain a wage that was equal to their contribution of labor to the production process, as otherwise they could easily be self-employed since the economy was still quite labor intensive and the tools needed for self-employment may not have cost very much. Also, since ownership of the tools of a trade may have been quite widespread while the contribution of capital resources to the production process was quite small as compared to that of labor, a major part of the income might have accrued to the working households. In such circumstances, the simplifying assumptions of the neo-classical model may appear quite reasonable.

The neo-classical model became irrelevant, however, as the system made progress in the presence of a social organizational framework that legally protected ownership of all types and freely allowed the renting of assets, thus making possible an absentee mode of owning productive resources while technological changes also made the contribution of capital resources to the production process more significant.

Creating Worker Capitalism

It is not only methodologically expedient but also pedagogically interesting to explore what ownership and wage

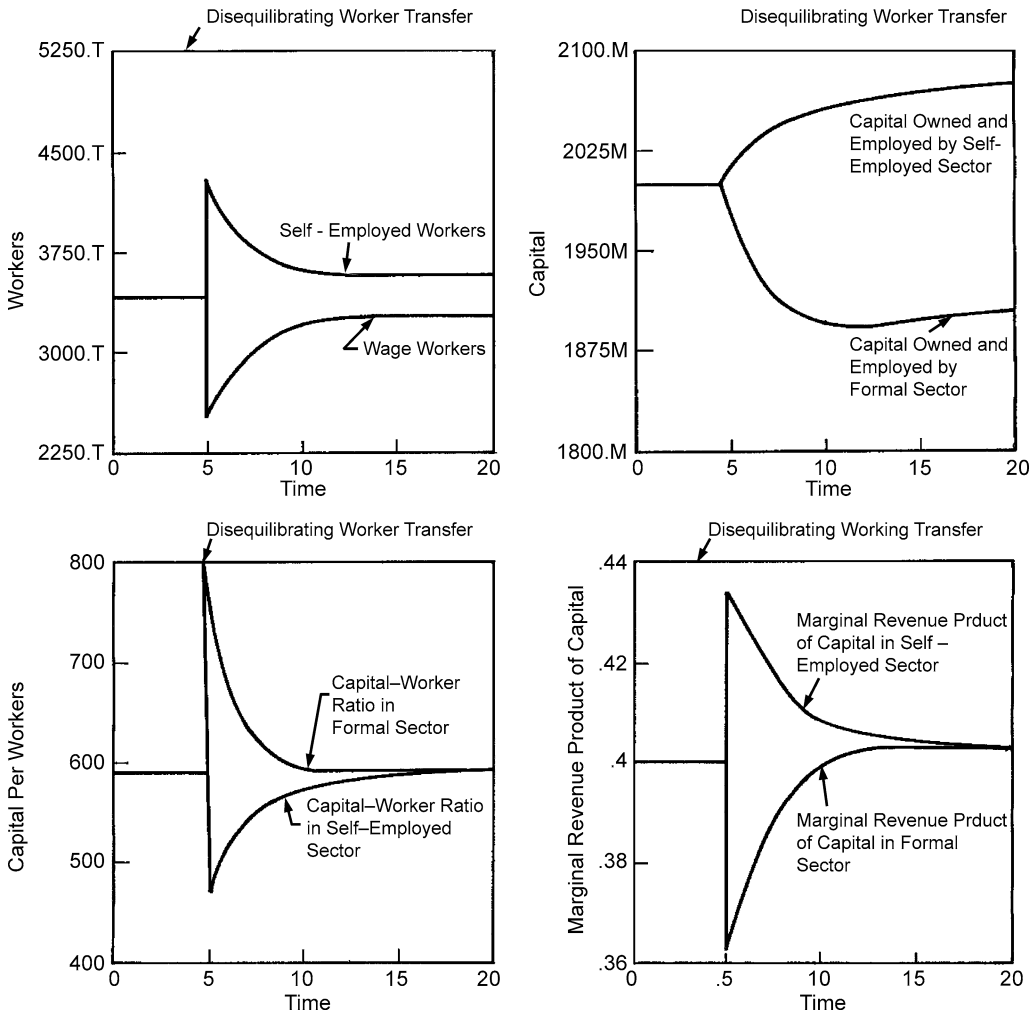

Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 3
Recovery from dis-equilibrium in a neo-classical system

patterns might have emerged if labor-wages were determined through bargaining mechanisms incorporated into the model instead of fair payment equal to the marginal revenue product of workers, while all other assumptions of a perfect market of the experiment of the last section were maintained.

Figure 4 shows a simulation of the model in which wage rate is determined by the average consumption expenditure per worker (as given in Eqs. (1) and (2) of the model described in the Appendix "Model Description") while renting of production factors and financial fragmentation of the households are still not allowed. This change in assumptions disturbs the initial market equilibrium in the model thus activating its internal tendency to seek a new equilibrium. No exogenous disequilibrating changes are needed to generate the dynamic behavior in this simulation and in those discussed hereafter.

As a result of this change, the compensation demanded for working in the capitalist sector becomes much higher than the marginal revenue product of the workers. Thus, wage-workers are laid off and accommodated in the self-employed sector. Consequently, the marginal revenue product of land and capital in the self-employed sector increases and its bids for these resources rise. On the other hand, the decrease in the workforce of the capitalist sector increases its land and capital intensities and hence lowers their marginal revenue products. The falling productivity of these resources increases the opportunity cost of holding them. Since renting is not allowed, the capitalist sector is persuaded to sell the resources to the self-employed who

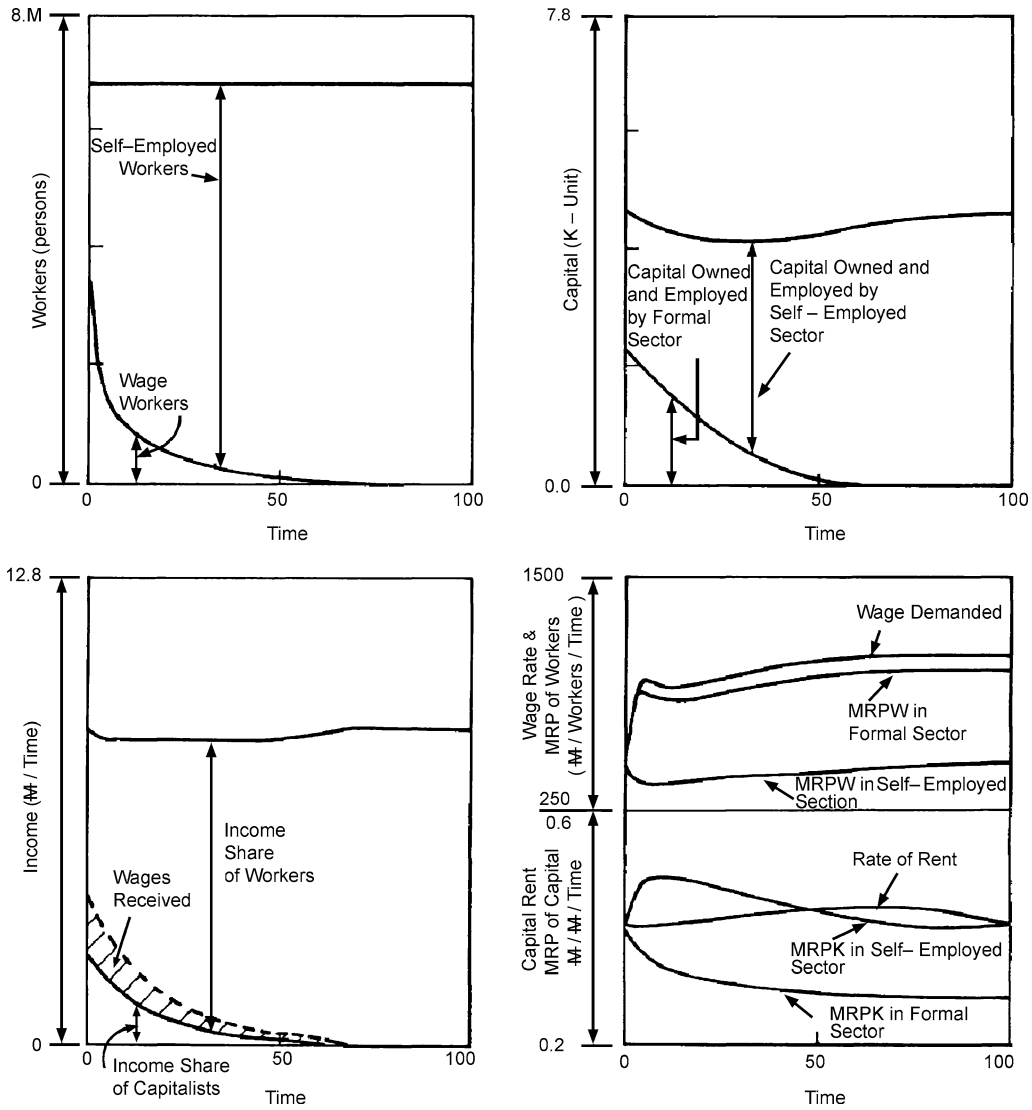

Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 4
The develop of worker capitalism when wages depend on bargaining position of workers

can easily buy them since investment in the model is not subject to internal self-finance.

As the self-employed sector increases its land and capital holdings, its production rises. When increases in the production of this sector exceed the wage income lost due to decreasing wage disbursements from the capitalist sector, the net revenue of the workers, and hence their average consumption, rises. The wage rate is thus pushed up further, which necessitates further reductions in wage-workers. These processes spiral into a gradual transfer of all resources to the self-employed sector.

The marginal revenue products of land and labor in the two sectors tend to equilibrate at different values, but the capitalist sector exists only in theory because towards the end of the simulation almost all the resources are owned and managed by the self-employed. Since no part of the income is obtained by absentee owners, and working households may own and manage resources according to the quantity of labor they can supply, the income distribution may appear to be truly egalitarian.

Even though the above simulation is hypothetical, the wage and income distribution pattern shown by it may be experienced when the separation of resources from the households employing them is socially or legally ruled out or the state allocates capital resources and land ac-

cording to the quantity and quality of labor supplied by a household. Instances of peasant economies having such characteristics have been recorded in history in tribal cultures and, in a somewhat advanced form, in medieval India [35]. Interestingly, such implicit assumptions are also subsumed in the illusive perfect market the neoclassical economic theory is based on.

Appearance of Absentee Ownership

When ownership of resources is legally protected, whether they are productively employed or owned in absentia, many renting and leasing arrangements may appear which may allow a household to own resources without having to employ them for production [47]. This is borne out in the simulation of Fig. 5, in which resources are divided by the capitalist sector between commercial production and renting activities depending on the rates of return in each. Rents depend on long-term averages of the marginal revenue products of the respective factors and on the demand for renting as compared with the supply of rentable assets. In the new equilibrium reached by the model, the commercial mode of production and wage-employment gradually disappear but a substantial part of the resources continues to be owned by the capitalist sector, which rents these out to the self-employed sector.

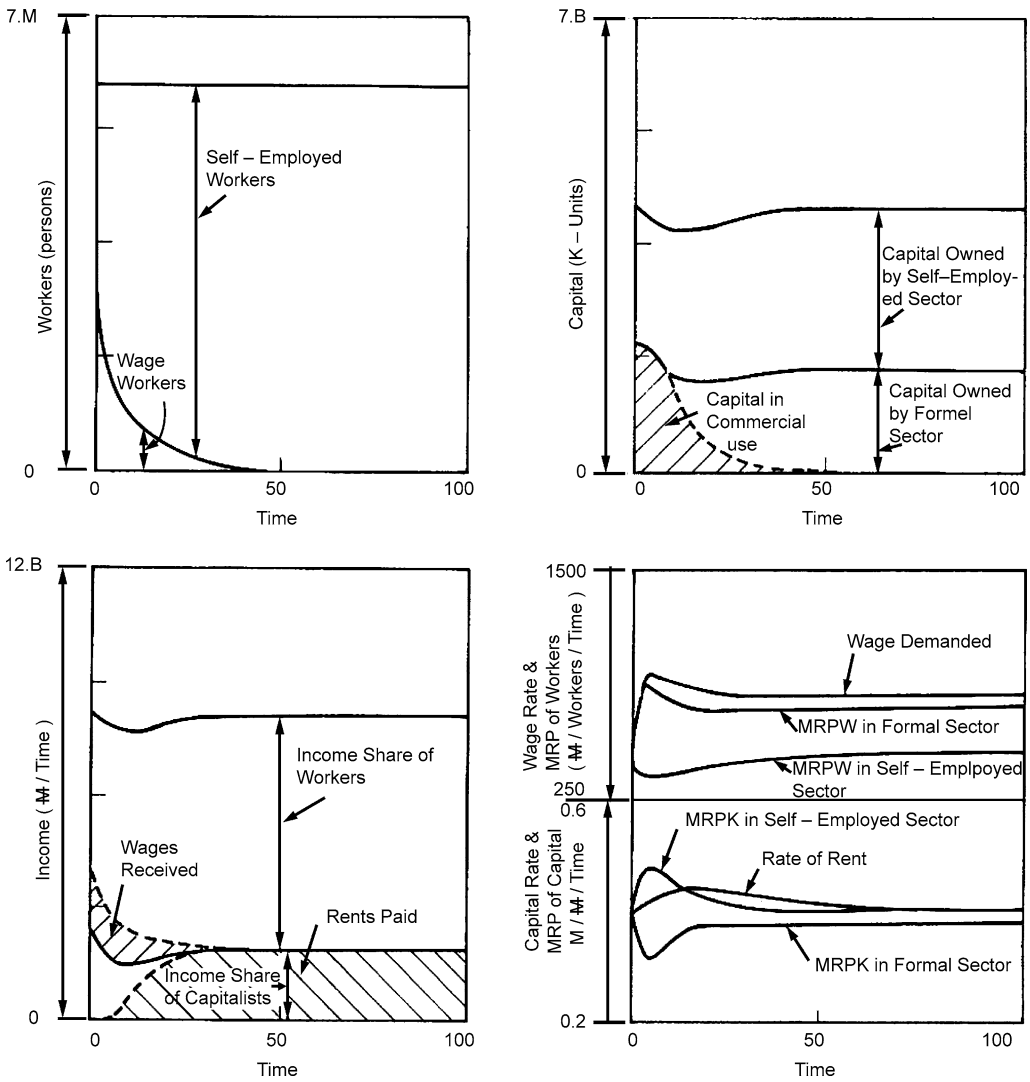

Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 5
The appearance of absentee ownership when renting is also allowed

Such a pattern develops because of the combined effect of wage and tenure assumptions incorporated into the model. When workers are laid off by the capitalist sector in response to a high wage rate, the marginal revenue products of land and capital for commercially employing these resources in this sector fall. However, as the laid-off workers are accommodated in the self-employed sector, the marginal revenue products of land and capital, and hence their demand in this sector, rise. Therefore, rents are bid up and the capitalist sector is able to get enough return from renting land and capital to justify maintaining its investment in these.

Again, the marginal revenue products of the production factors in the commercial mode of production are only hypothetical as that mode is not practiced towards the end of the simulation. The renting mechanism allows the self-employed sector to adjust its factor proportions quickly when it is faced with the accommodation of a large number of workers. When the economy reaches equilibrium, the marginal rates of return of the production factors in the self-employed sector are the same as those at the beginning of the simulation. But, the wage demanded equilibrates at a level lower than that for the exclusively self-employed economy described in the simulation of Figure 4, because a part of the income of the economy is now being obtained by the absentee owners of the capitalist sector in the form of rent.

Note that, although the total income of the economy falls a little during the transition, it rises back to the original level towards the end equilibrium since the technology is uniform, irrespective of the mode of production. Also note that the end equilibrium distribution of income depends on initial distribution of factors when modifying assumptions are introduced, and on the volume of transfers occurring over the course of transition. Thus, an unlimited number of income and ownership distribution patterns would be possible depending on initial conditions and the parameters of the model representing the speeds of adjustment of its variables. The common characteristics of these patterns, however, are the presence of absentee ownership, the absence of a commercial mode of production, and a shadow wage that is less than an exclusively self-employed system.

Separation of Ownership from Workers and the Creation of a Marxist System

The ownership of resources becomes separated from the workers and concentrated in the capitalist sector in the model, irrespective of the initial conditions of resource distribution, when the assumption about the existence of a perfect financial market is also relaxed.

Figure 6 shows the ownership and wage pattern which develops when acquisition of resources by the capitalist and self-employed sectors is made dependent, in addition to their profitability, on the ability to self-finance their purchase. Recall also that the ability to self-finance depends on the unspent balance of savings, and the saving rate of the self-employed sector is sensitive both to the utility of saving in this sector to support investment for self-employment and to the rent burden of this sector compared with the factor contribution to its income from land and capital. The saving rate of the capitalist sector is assumed to be constant.

Such a pattern develops because of an internal goal of the system to employ resources in the most efficient way while the ownership of these resources can only be with the households who have adequate financial ability, which is also not independent of ownership.

Creation of a Dualist System

A dualist system is characterized by the side-by-side existence of both commercial and self-employed modes of production. The former appears to be economically efficient and is often also capital-intensive. The latter is seen to be economically inefficient and is also invariably labor-intensive. The side-by-side existence of these two modes of production in many developing countries has often puzzled observers, since according to the neo-classical economic theory, any inefficient production mode must be displaced by the efficient one.

A stable commercially run capital-intensive production sector existing together with a self-employed labor-intensive sector develops in the model if a technological differentiation is created between the capitalist and self-employed sectors. This is shown in the simulation in Fig. 7, in which an exogenous supply of modern capital is made available after end equilibrium of the simulation in Fig. 6 is reached.

Capital differentiation between the two sectors appears since the scale of the self-employed producers does not allow them to adopt modern technologies requiring indivisible capital inputs. The capitalist sector starts meeting its additional and replacement capital needs by acquiring a mixture of modern and traditional capital while the self-employed sector can use only traditional capital. However, the capital demand of the capitalist sector is met by modern capital as much as the fixed supply permits. The balance of its demand is met by acquiring traditional capital.

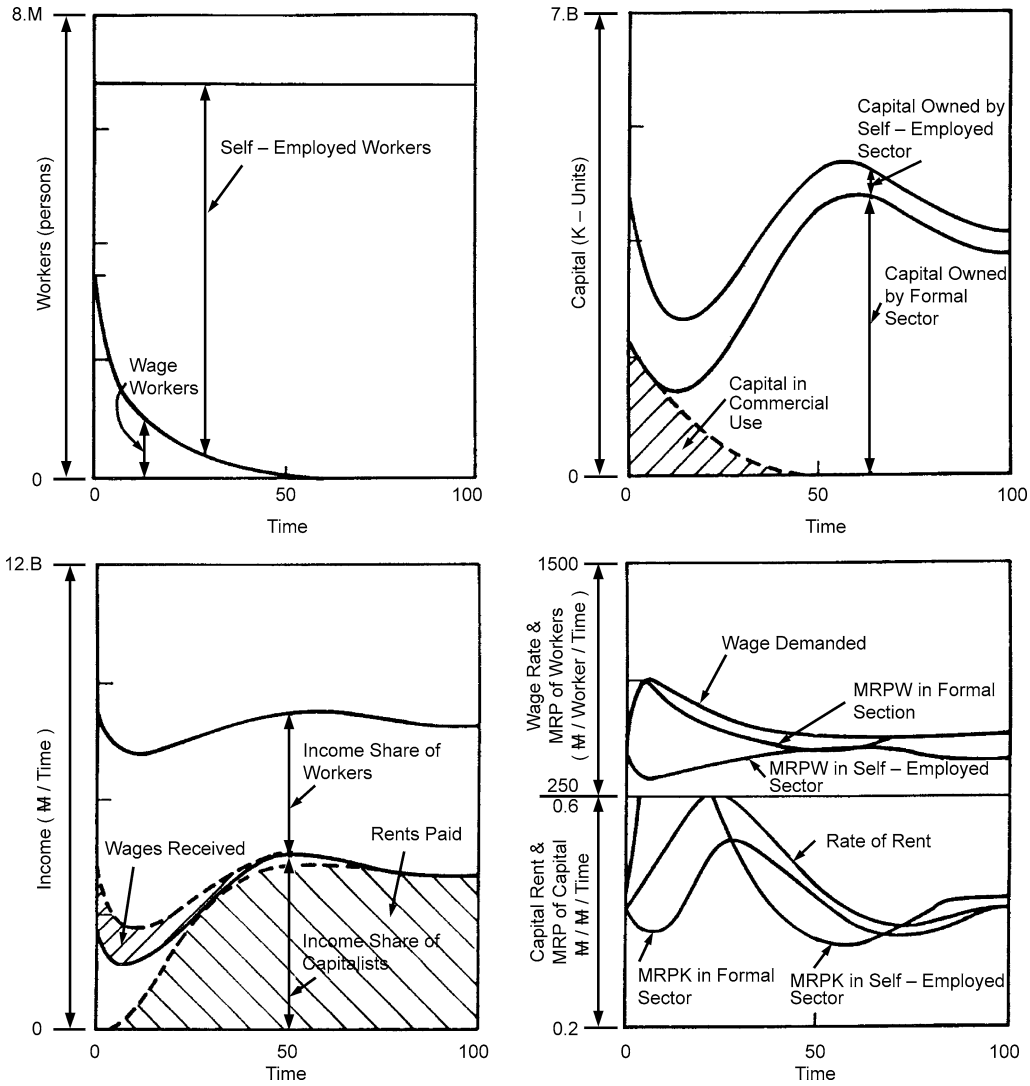

Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 6
Separation of ownership from workers as postulated by Marx system when investment must also be internally financed

The output elasticity of modern capital is assumed to be higher than that of the traditional capital while the use of the former also allows an autonomous increase in output. The output elasticity of land is assumed to remain constant. The assumption of uniform returns to scale is maintained. Thus, the output elasticity of workers decreases when modern capital is introduced. These assumptions serve to represent the high productivity and labor-saving characteristics of the modern capital.

As its capital becomes gradually more modern and potentially more productive, the capitalist sector is able to employ its productive resources with advantage in the commercial mode of production, instead of renting these out, and to employ wage-workers at the going wage rate. The increased productivity and income derived from this make it both economically and financially viable for the capitalist sector to invest more. Thus, its share of resources, when a new equilibrium is reached, is further increased.

Since the output elasticity of workers falls with the increase in the fraction of modern capital, the marginal revenue product of workers in the commercial mode may not rise much with the increase in its output. At the same time, since resources are being transferred away by the capitalist sector from renting to commercial employment, the labor intensity and the demand for renting rises in the self-em-

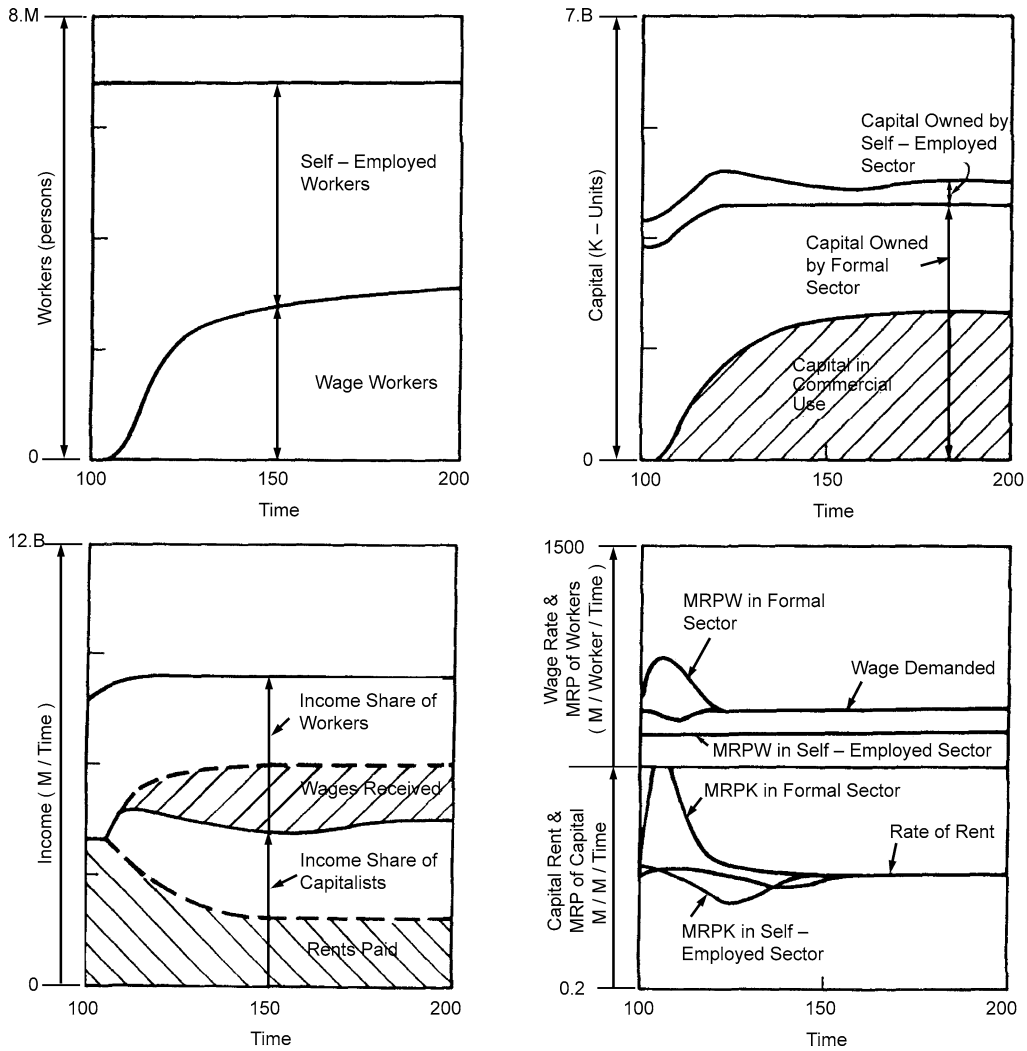

Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 7
Creation of dualist system when technological differentiation develops between the capitalist and self-employed sectors

ployed sector. Hence rents are bid up and it again becomes profitable for the capitalist sector to allocate resources to renting. The amount of resources rented out, however, will depend on the degree of technological differentiation that may be created between the two sectors.

The wage rate reaches equilibrium at a lower level and the rents at higher levels than without technological differentiation. Rents, however, equal marginal revenue products of land and capital, which rise in the capitalist sector because of employing superior technology and in the self-employed sector due to increased labor intensity.

Interestingly, dualist patterns appeared in the developing countries, both in the agricultural and industrial sectors, only after modern capital inputs became available in limited quantities. Labor-intensive peasant agriculture and small-scale industry and services carried out by the self-employed came to exist side-by-side with the commercially run farms and large-scale industry employing wage labor and modern technologies. However, worker income, both in wage-employment and self-employment, remained low [19].

Feedback Loops Underlying Wage and Income Patterns

The internal goal of a dynamic system represented by a set of non-linear ordinary differential equations is created by the circular information paths or feedback loops which are

formed by the causal relations between variables implicit in the model structure. These causal relations exist in the state space independently of time (unless time also represents a state of the system). The existence of such feedback loops is widely recognized in engineering and they are often graphically represented in the so-called block and signal flow diagrams [17,40,65].

While many feedback loops may be implicit in the differential equations describing the structure of a system, only a few of these would actively control the system behavior at any time. The nonlinearities existing in the relationships between the state variables determine which of the feedback loops would actively control the system behavior. A change may occur in the internal goals of a system if its existing controlling feedback loops become inactive while simultaneously other feedback loops present in its structure become active. Such a shift in the controlling feedback loops of a system is sometimes called a structural change in the social sciences and it can result both from the dynamic changes occurring over time in the states of the system and from policy intervention. The realization of a specific wage and income distribution pattern depends not on assumptions about initial conditions but on legal and social norms concerning ownership, renting, financing of investment and the state of technology, determining which feedback loops would be dominant [14,40].

Figure 8 describes the feedback loops, formed by the causal relations implicit in the model structure that appear to polarize income distribution by separating asset ownership from working households and creating a low wage rate, as shown in Fig. 6. An arrow connecting two variables indicates the direction of the causality while a positive or a negative sign shows the slope of the function relating cause to effect. For clarity, only key variables located along each feedback path are shown.

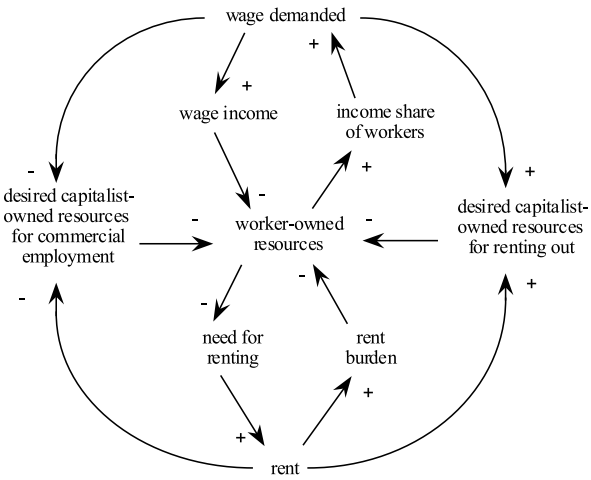

Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 8
Feedback loops creating dysfunctional income distribution trends in the capitalis system

When productive resources can potentially be engaged in commercial or self-employed modes by owners and renters, any autonomous increase in the wage rate would not only decrease the desired capitalist owned resources for commercial employment, it would also concomitantly decrease the utility of investing in resources for self-employment. Thus, while the ownership of resources freed from commercial employment is not transferred to the self-employed sector, the surplus labor released by the commercial sector has to be absorbed in self-employment. As a result, worker income is depressed while the demand for renting rises. Thus, it not only continues to be profitable for the capitalist sector to hold its investments in land and capital, it also gives this sector a financial edge over the self-employed sector, whose savings continue to decline as its rent burden rises. These actions spiral into an expansion of ownership of resources by the capitalist sector even though the commercial mode of production is eliminated due to the high cost of wage labor. This also precipitates a very low wage rate when equilibrium is reached since a low claim to income of the economy creates low opportunity costs for the self-employed workers for accepting wage-employment.

Ironically, the fine distinction between the corporate, artisan and absentee types of ownership is not recognized in the political systems based on the competing neoclassical and Marxist economic paradigms. The former protects all types of ownership; the latter prohibits all. None creates a feasible environment in which a functional form of ownership may help to capture the entrepreneurial energy of the enterprise.

Possibilities for Poverty Alleviation

A functional economic system must incorporate the mechanisms to mobilize the forces of self-interest and entrepreneurship inherent in private ownership of the resources. Yet, it must avoid the conflicts inherent in the inequalities of income and resource ownership that led to the creation of the alternative socialist paradigm, which is devoid of such forces. According to the preceding analysis, the fundamental mechanism which creates the possibility of concentration of resource ownership is the equal protection accorded to the artisan and absentee forms of ownership by the prevailing legal norms. The financial fragmentation of households and the differences in their

saving patterns further facilitate the expansion of absentee ownership. Technological differences between the capitalist and self-employed sectors not only make possible the side-by-side existence of the two modes of production, they also exacerbate the dichotomy between ownership of resources and workership. Apparently, the policy agenda for changing resource ownership and income distribution patterns should strive to limit renting and should additionally prevent the development of financial fragmentation and technological differentiation between the commercial and self-employed production modes if the objective is to minimize the conflicts related to income distribution.

Assisting the Poor

Programs to provide technological, organizational, and financial assistance to the poor have been implemented extensively in the developing countries over the past few decades although they have changed neither income distribution nor wage rate as reflected in many learned writings over these decades as well as the data published by UN and World Bank. This occurred because the increased productivity of the self-employed mode first pushed up wage rate, making renting-out resources more attractive for the capitalist sector than commercial production. However, the consequent decrease in wage payments and increase in rent payments pushed down the income share of the workers, which again suppressed the wage rate. Any efforts to facilitate the small-scale sector to increase its productivity through technological development also failed to affect income distribution since the mechanism of renting allowed the gains of the improved productivity to accrue to the absentee owners of the resources [56]. This experience is verified by the simulation of Fig. 9, which incorporates the policies to improve productivity, creating financial institutions and assisting the self-employed to adopt modern technologies. These policies only increase the size of the self-employed sector without increasing worker income, due to the possibility of separation of the mode of production from the ownership of resources. This indicates that influencing the decision to retain resources in absentee mode for renting out should be the key element of a policy framework to improve income distribution that should alleviate poverty.

Influencing Income Distribution

The cost of owning capital resources in absentee form can be increased by imposing a tax on rent income. The results of implementing this policy, together with the policies of Fig. 9 are shown in Fig. 10. In the face of a tax on rent income, resources which cannot be employed efficiently under the commercial system are offered for sale to the self-employed instead of being leased out to them. Purchase of these resources by the self-employed raises the entitlement of the workers to the income of the economy, which increases the opportunity cost of supplying wage-labor to the commercial sector. This raises wage rate, which makes the commercial mode of production even more uneconomical, unless it is able to apply a superior technology. Such changes spiral in the long run into a transfer of a substantial amount of resources to the self-employed sector. Concomitant efforts to decrease the financial fragmentation of households and the technological differentiation between the two modes of production, along with improving productivity, further accelerate these changes.

Facilitation of Innovation and Productivity Improvement

Macroeconomic analyses concerning the industrialized countries show that technological innovation is one of the most important sources of growth [9,61]. Studies conducted at the organizational level in the industrialized countries also show that innovations creating technological progress originate largely from small entrepreneurs or from large companies structured in a way to encourage small independent working groups [38,42]. Thus, entrepreneurial activity is often credited with raising productivity through creation of technical and business-related innovations [6]. The high and rising cost of labor in the developed countries, possibly also forces the wage-employers into finding innovative ways of maintaining high labor productivity and continuously striving to improve it.

On the other hand, economic growth has been dominated in the developing countries by relatively large and often highly centralized and vertically integrated companies. Technologies of production have mostly been replanted from the industrialized countries and indigenous innovation and technological development have had a poor track record [1]. These technologies often do not perform as well as at their respective sources, but due to the availability of cheap labor, their inefficient performance still yields comfortable profits; hence little effort is made to improve productivity. There also exist serious limitations on the number of small players as large cross-section of the households in the developing countries lack the resources to effectively participate in any form of entrepreneurial activity [53,58]. Innovation being a probabilistic process, limited participation drastically limits its scope.

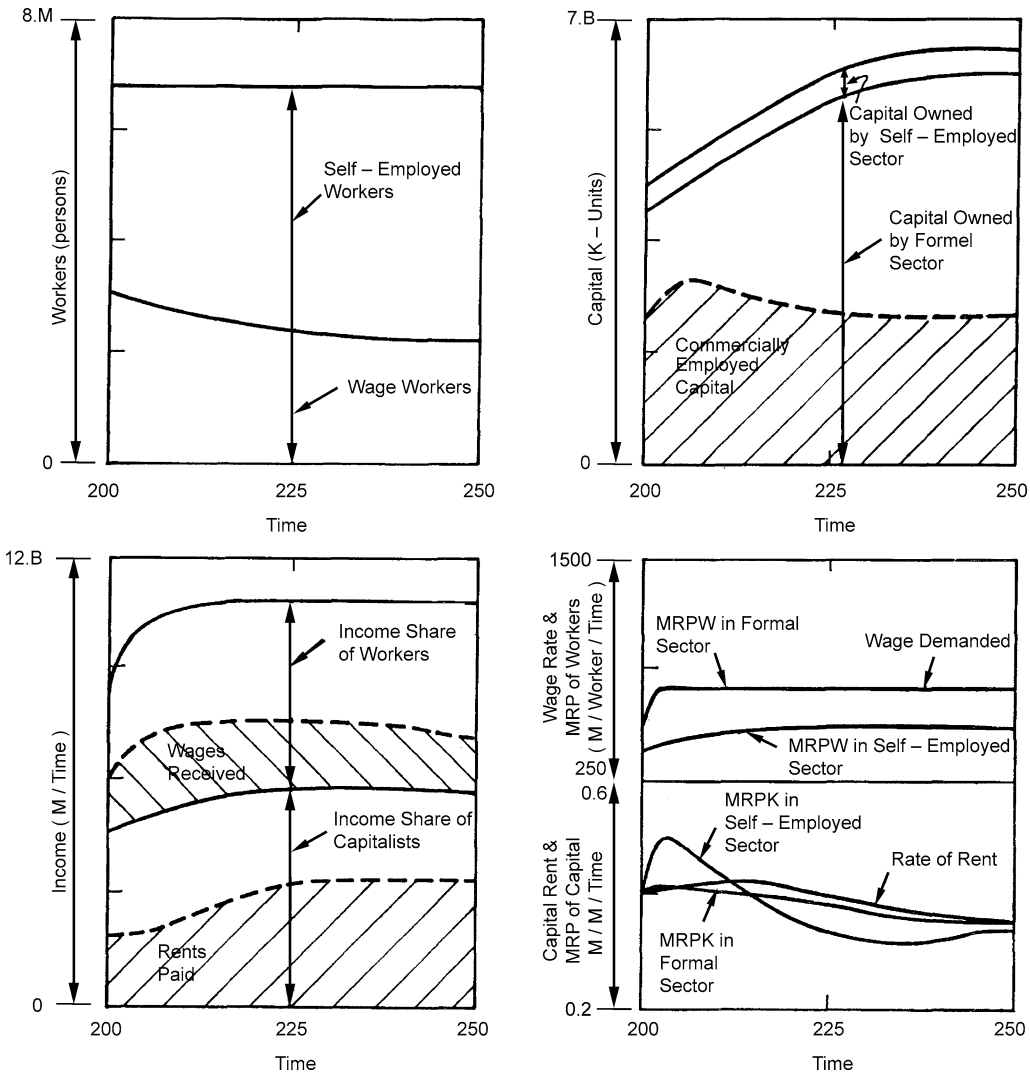

Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 9
Perpetuation of low wage and unequal income distribution resulting from widely used economic development policies

There exists a promising institution in most developing countries, however, which has great potential as a focal point of entrepreneurial activity, which has remained dormant for lack of empowerment. This institution is the small family enterprise in the self-employed sector, which may take the form of a shop-house or an artisan manufacturing firm in the urban sector or a peasant farm in the rural sector. It allows participation from all members of the family while also providing the informal small-group organization considered conducive to innovation in many studies. Its members are highly motivated to work hard and assume the risk of enterprise because of their commitment to support the extended family. This enterprise is somewhat similar to the small manufacturing units that created the industrial revolution in England in the early nineteenth century. It has also been observed that the small family enterprise tends to maximize consumption; hence its income significantly affects demand, which creates new marketing opportunities [1,4]. Unfortunately, this enterprise has been systematically suppressed and discriminated against in favor of the large-scale capitalist sector. Even its output remains largely unaccounted for in the national accounting systems of most countries [11,20].

The small family enterprise, variously described as the informal, labor-intensive, traditional, peasant, peripheral and sometimes inefficient sector in the developing countries has been stifled in the first instance by a set of social and legal norms through which the wealth has become

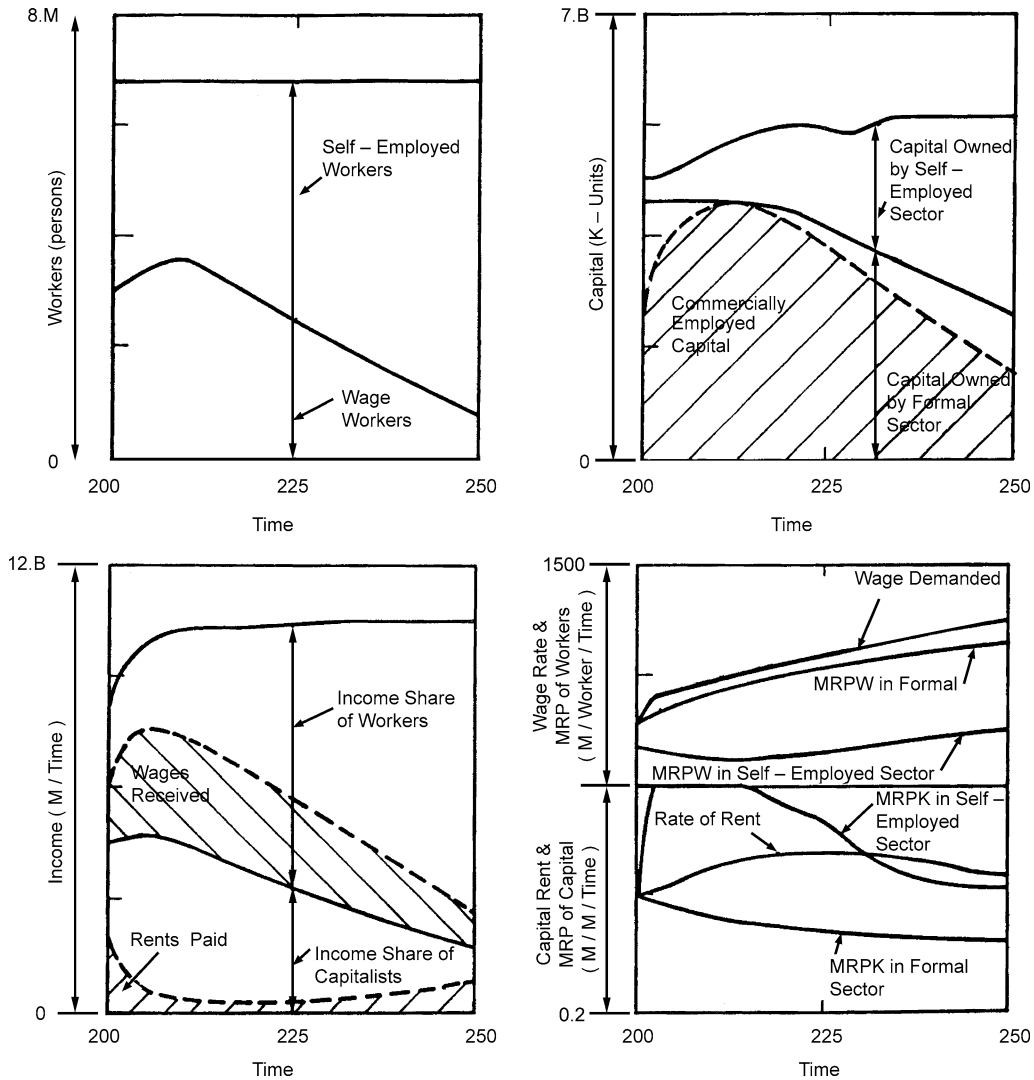

Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 10
Changes in wage and income distribution resulting from adding taxation of rent income to the policy package

concentrated in an absentee ownership mode. Working households are mostly poor and own few assets [19]. The prosperity of these households will not only provide the much-needed financial resources for entrepreneurial activity, their capacity to spend will also create many marketing opportunities for the potential entrepreneur. Thus, influencing income distribution, through the policy framework proposed in the last section, ranks first on the agenda also for developing entrepreneurship and encouraging innovation. Once a significant cross-section of the populace becomes a potential participant in the economic activity, the development of infrastructure and facilitation to manage risk will also appear to be effective instruments to support entrepreneurship and innovation [51]. The rise in the wage rates due to the possibility of alternative self-employment opportunities would, at the same time, force the large commercial enterprise to invest in technological innovation for productivity improvement, which should further improve the efficacy of the overall system.

Conclusion

Both neoclassical and Marxist models of economic growth seem to make restricting assumptions about ownership and mechanisms of wage determination, which are linked with specific time- and geography- related historical ev-

idence. These restricting assumptions give internal consistency and a semblance of sustainability to each model, although they remove both from reality. A failure in the free-market system based on the no-classical model occurs when the invisible hand concentrates ownership of resources in a small minority, suppressing wage rate and creating social conflict due to income inequalities. On the other hand, a failure in the socialist system based on the Marxist model occurs, when the visible hand empowered to act in the public interest stifles entrepreneurial energy while also ignoring public interest in favor of its power interests [48,50].

A behavioral model underlying wage and income distribution has been proposed in this paper, in which the opportunity cost of supplying a unit of labor to the capitalist sector is used as a basis for negotiating a wage. Neither this opportunity cost nor the ownership pattern are taken as given, while the dynamic interaction between the two creates a tendency in the system to generate numerous wage and income distribution patterns, subsuming those postulated in the neo-classical and Marxist theories of economics. The realization of a specific wage and income distribution pattern depends on legal and social norms concerning ownership, renting, the financing of investment and the state of technology.

Private ownership seems to have three forms, commercial, artisan and absentee. Predominance of artisan ownership creates an egalitarian wage and income distribution pattern while a healthy competition between the commercial and artisan firms may release considerable entrepreneurial energy. These functional forms can grow only if the renting of resources can be discouraged. On the other hand, absentee ownership creates a low wage rate and an unequal income distribution, while the growth of this form of ownership is facilitated through the renting mechanism. Potentially, all three ownership forms can exist in an economic system. The problem, therefore, is not to favor or condemn private ownership *per se* as the alternative theories of economics have often advocated, but to understand the reasons behind the development of a particular ownership pattern and identify human motivational factors that would change an existing pattern into a desired one.

The most important reform needed at government level to alleviate poverty is the discouragement of the absentee ownership of capital assets, which would create a wider distribution of wealth. Widespread artisan ownership resulting from this would increase participation in entrepreneurial activity, which would allow adequate performance from the human actors in the system. Such reforms may however not be possible in the authoritarian systems of government pervasive in the developing countries since they must often limit civil rights and public freedoms to sustain power. Hence, the creation of a democratic political system may be a pre-condition to any interventions aimed at poverty alleviation. This, I have discussed elsewhere [50,53,54].

Future Directions

While the market system has often been blamed by the proponents of central planning for leading to concentration of wealth and income among few, it in fact offers a powerful means for redistributing income if the process of concentration is carefully understood and an intervention designed on the basis of this understanding. In fact, all economic systems can be reformed to alleviate the dysfunctional tendencies they are believed to have, provided the circular relationships creating such dysfunctions can be understood which should be the first objective of policy design for economic development.

Contrary to this position, economic development has often viewed developmental problems as pre-existing conditions, which must be changed through external intervention. Poverty, Food shortage, poor social services and human resources development infrastructure, technological backwardness, low productivity, resource depletion, environmental degradation and poor governance are cases in point. In all such cases, the starting point for a policy search is the acceptance of a snapshot of the existing conditions. A developmental policy is then constructed as a well-intended measure that should improve existing conditions. Experience shows, however, that policies implemented with such a perspective not only give unreliable performance, they also create unintended consequences. This happens because the causes leading to the existing conditions and their future projections are not adequately understood. The well-intentioned policies addressing problem symptoms only create ad hoc changes, which are often overcome by the system's reactions.

Table 1 collects three key developmental problems, poverty, food shortage and social unrest, and the broad policies implemented over the past several decades to address them. These problems have, however, continued to persist or even become worse.

The policy response for overcoming poverty was to foster economic growth so aggregate income could be increased; that for creating food security was intensive agriculture so more food could be produced; and for containing social unrest, the broad prescription is to strengthen internal security and defense infrastructure so public could be protected from social unrest. The unintended

Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Table 1
Developmental problems, policies implemented to address them and unintended consequences experienced

Initially perceived problems	Policies implemented	Unintended consequences
Poverty	Economic growth capital formation sectoral development technology transfer external trade	Low productivity indebtedness natural resources depletion environmental degradation continuing/increased poverty
Food shortage	Intensive agriculture land development irrigation fertilizer application use of new seeds	Land degradation depletion of water aquifers vulnerability to crop failure population growth continuing/increased vulnerability to food shortage
Social unrest	Spending on internal security and defense infrastructure limiting civil rights	Poor social services poor economic infrastructure authoritarian governance insurgence continuing/increased social unrest

consequences of these policies are many, but in most instances, they include a continuation or worsening of the existing problems.

Thus, poverty and income differentials between rich and poor have in fact shown a steady rise, which is also accompanied by unprecedented debt burdens and extensive depletion of natural resources and degradation of environment. Food shortages have continued but are now accompanied also by land degradation, depletion of water aquifers, the threat of large-scale crop failure due to a reduction in crop diversity and a tremendous growth in population. Social unrest has often intensified together with appearance of organized insurgence burgeoning expenditures on internal security and defense, which has stifled development of social services and human resources and have created authoritarian governments with little commitment to public welfare.

The unintended consequences are often more complex than the initial problems and have lately drawn concerns at the global level, but whether an outside hand at the global level would alleviate them is questionable. This is evident from the failure to formulate and enforce global public policy in spite of active participation by national governments, global agencies like the UN, the World Bank, the World Trade Organization, and advocacy networks sometimes referred to as the civil society. This failure can largely be attributed to the lack of a clear understanding of the roles of the actors who precipitated those problems and whose motivations must be influenced to turn the tide.

Thus, development planning must adopt a problem solving approach in a mathematical sense if it is to achieve sustainable solutions. In this approach, a problem must be defined as an internal behavioral tendency and not as a snap shot of existing conditions. It may represent a set of patterns, a series of trends or a set of existing conditions that appear either to characterize a system or to be resilient to policy intervention. In other words, an end condition by itself must not be seen as a problem definition. The complex pattern of change implicit in the time paths preceding this end condition would, on the other hand, represent a problem. The solution to a recognized problem should be a solution in a mathematical sense, which is analogous to creating an understanding of the underlying causes of a delineated pattern. A development policy should then be perceived as a change in the decision rules that would change a problematic pattern to an acceptable one. Such a problem solving approach can be implemented with advantage using system dynamics modeling process that entails building and experimenting with computer models of problems, provided of course a succinct problem definition has first been created.

Appendix

Model Description

Wage rate WR is assumed to adjust over period $WRAT$ towards indicated wage rate IWR.

$$\mathrm{d}/\mathrm{d}t[WR] = (IWR - WR)/WRAT \tag{1}$$

IWR depends on the wage-bargaining position of the workers, which is determined by their opportunity cost of accepting wage-employment. It is assumed that the opportunity cost of transferring a self-employed worker to wage-work is zero when wage offered is equal to the current con-

sumption expenditure per worker averaged over the whole workforce.

$$IWR = [(R_s * (1 - SP_s) + (AS_s/LAS))/TW] , \quad (2)$$

where R_s, SP_s and AS_s are, respectively, income share, saving propensity and accumulated unspent savings of the self-employed sector. LAS and TW are, respectively, life of accumulated unspent savings and total workforce. Subscripts s and f designate, respectively, self-employed and capitalist sectors.

Ownership of land and capital as well as contribution to labor are the bases for claim to income while absentee ownership is possible through leasing arrangements. Thus, R_s is computed by adding together the value of output produced by the self-employed sector VQ_s and the wage payments received by the wage-workers W_f, and subtracting from the sum the rent payments made to the absentee owners. R_f is given by adding together the value of output produced by the capitalist sector VQ_f and the rent payments it receives from the self-employed sector, and subtracting from the sum the wage-payments it makes.

$$R_s = VQ_s + WR * W_f - LR * RL - KR * RK , \quad (3)$$

$$R_f = VQ_f - WR * W_f + LR * RL + KR * RK , \quad (4)$$

where LR, RL, KR, and RK, are, respectively, land rent, rented land, capital rent, and rented capital.

KR and LR depend, respectively, on the long-term averages of the marginal revenue products of capital and land ($AMRPK$ and $AMRPL$) in the economy, and the demand for renting capital and land (RKD and RLD) as compared with the supply of rentable assets (RK and RL). The demand for renting, in turn, depends on the lack of ownership of adequate resources for productively employing the workers in the self-employed sector.

$$KR = AMRPK * f_1[RKD/RK] ; \quad f_1' > 0 \quad (5)$$

$$RKD = DKE_s - KO_s . \quad (6)$$

Where DKE_s is desired capital to be employed in the self-employed sector and KO_s is capital owned by it. Land rent LR and demand for renting land RLD are determined similarly.

The saving propensity of all households in not uniform. Since capitalist households associated with the capitalist sector receive incomes which are much above subsistence, their saving propensity is stable. On the other hand, the saving propensity of the worker households depends on their need to save for supporting investment for self-employment and on how their absolute level of income compares with their inflexible consumption. Thus, SP_s in the model is determined by the utility of investment in the self-employed sector arising from a comparison of worker productivity in the sector with the wage rate in the capitalist sector, and the rent burden of this sector compared with the factor contribution to its income from land and capital.

$$SP_s = \mu * f_2[MRPW_s/WR] * f_3[(LR * RL + KR * RK) /(VQ_s - MRPW_s * W_s)] , \quad (7)$$

$$SP_f = \mu , \quad (8)$$

where $f_2' > 0$, $f_3' < 0$, μ is a constant, and MRPW is marginal revenue product of workers.

AS represent the balance of unspent savings, which determine the availability of liquid cash resources for purchase of assets. AS are consumed over their life LAS whether or not any investment expenditure occurs.

$$\mathrm{d}/\mathrm{d}t[AS_i] = R_i * SP_i - AS_i/LAS - LA_i * PL - \sum_j KA_i^j * GPL ; \quad i = \mathrm{s,f} ; \quad j = \mathrm{m,t} , \quad (9)$$

where LA, PL, KA, and GPL are, respectively, land acquisitions, price of land, capital acquisitions, and general price level. Subscript i refers to any of the two sectors, self-employed (s) and capitalist (f), and superscript j to the type of capital, modern (m) or traditional (t).

W_f is assumed to adjust towards indicated workers IW_f given by desired workers DW_f and total workforce TW. TW is assumed to be fixed, although, relaxing this assumption does not alter the conclusions of this paper. All workers who are not wage-employed must be accommodated in self-employment. Thus W_s represents the remaining workers in the economy.

$$\mathrm{d}/\mathrm{d}t[W_f] = (IW_f - W_f)/WAT \quad (10)$$

$$IW_f = TW * f_4(DW_f/TW) \quad (11)$$

$$W_s = TW - W_f \quad (12)$$

where $1 \geq f_4 \geq 0$, and $f_4' > 0$. WAT is worker adjustment time.

The desired workers in each sector DW_i is determined by equating wage rate with the marginal revenue product of workers. A modified Cobb–Douglas type production function is used.

$$DW_i = E_i^w * VQ_i/WR , \quad (13)$$

where E_i^w is the elasticity of production of workers in a sector.

Land and capital owned by the capitalist sector (LO_f and KO_f) are allocated to commercial production (KE_f and LE_f) and renting (RK and RL) activities depending on the desired levels of these factors in each activity. Thus,

$$RK = (DRK/(DRK + DKE_f)) * KO_f \tag{14}$$

$$RL = (DRL/(DRL + DLE_f)) * LO_f \tag{15}$$

$$KE_f = KO_f - RK \tag{16}$$

$$LE_f = LO_f - RL \tag{17}$$

Capital and land employed by the self-employed sector consist of these production factors owned by them and those rented from the capitalist sector.

$$KE_s = Ko_s + RK \tag{18}$$

$$LE_s = Lo_s + RL\,. \tag{19}$$

Desired capital and land to be employed in any sector (DKE_i and DLE_i) are determined on the basis of economic criteria.

$$\mathrm{d}/\mathrm{d}t(DKE_i)/KE_i = f_6[MRPK_i/MFCK] \tag{20}$$

$$\mathrm{d}/\mathrm{d}t(DLE_i)/LE_i = f_5[MRPL_i/MFCL]\,, \tag{21}$$

where f_5' and $f_6' > 0$. $MRPL_i$ and $MRPK_i$ are respectively marginal revenue products of land and capital in a sector, and $MFCL$ AND $MFCK$ are respectively marginal factor costs of land and capital.

$$MRPL_i = (E_i^{\mathrm{l}} * VQ_i/LE_i) \tag{22}$$

$$MRPK_i = (E_i^{\mathrm{k}} * VQ_i/KE_i) \tag{23}$$

$$MFCL = PL * IR \tag{24}$$

$$MFCK = IR + (1/LK) * GPL\,, \tag{25}$$

where E_i^{l} and E_i^{k} are, respectively, elasticities of production of land and capital in a sector. PL is price of land, IR is exogenously defined interest rate, LK is life of capital and GPL is general price level.

Changes in the quantities of capital and land desired to be rented out (DRK and DRL) depend on their respective rents KR and LR compared with their marginal factor costs $MFCK$ and $MFCL$.

$$\mathrm{d}/\mathrm{d}t[DRK]/RK = f_7[KR/MFCK]\,; \quad f_7' > 0 \tag{26}$$

$$\mathrm{d}/\mathrm{d}t[DRL]/RL = f_8[LR/MFCL]\,; \quad f_8' > 0\,. \tag{27}$$

The value of output produced by each sector is given by the product of the quantity it produces Q_i and the general price level GPL.

$$VQ_i = Q_i * GPL \tag{28}$$

$$Q_i = A_i * K_i^{E^{\mathrm{ki}}} * L_i^{E^{\mathrm{li}}} * W_i^{E^{\mathrm{wi}}}\,, \tag{29}$$

where K_i, L_i, and W_i represent capital, land and workers employed by a sector. A_i represent technology constants, which increase with the use of modern capital.

$$A_i = \mathring{A} * f_9\left[K_i^m/(K_i^t + K_i^m)\right]\,, \tag{30}$$

where $f_9' > 0$ and $\mathring{A}$ is a scaling factor based on initial conditions of inputs and output of the production process.

Ownership is legally protected and the financial market is fragmented by households. Thus, purchase of any productive assets must be self-financed by each sector through cash payments. Land ownership LO_i of each sector changes through acquisitions LA_i from each other. Each sector bids for the available land on the basis of economic criteria, its current holdings, and the sector's liquidity.

$$LA_i = \mathrm{d}/\mathrm{d}t[LO_i] \tag{31}$$

$$LO_i = \left(DLO_i/\sum_i DLO_i\right) * TL\,, \tag{32}$$

where DLO_i is desired land ownership in a sector and TL is total land which is fixed,

$$DLO_i = LO_i * f_6[MRPL_i/MFCL] * f_{11}[CA_i]\,, \tag{33}$$

where $f_{11}'[CA_i]$ is > 0, and CA_i is cash adequacy of a sector.

Cash adequacy of a sector CA_i is given by the ratio of its accumulated unspent savings to the desired savings. The latter is computed by multiplying cash needed to finance investment and the traditional rate of consumption of savings in the sector by cash coverage CC.

$$CA_i = AS_i/\left(\left((AS_i/LAS) + (LA_i * PL) + \left(\sum_j KA_{ij} * GPL\right)\right) * CC\right)\,. \tag{34}$$

Capital ownership in a sector $KO_i = KO_i^t + KO_i^m$ changes through acquisitions KA_i^j and decay. Although there is a preference for modern capital, its acquisition KA_i^m de-

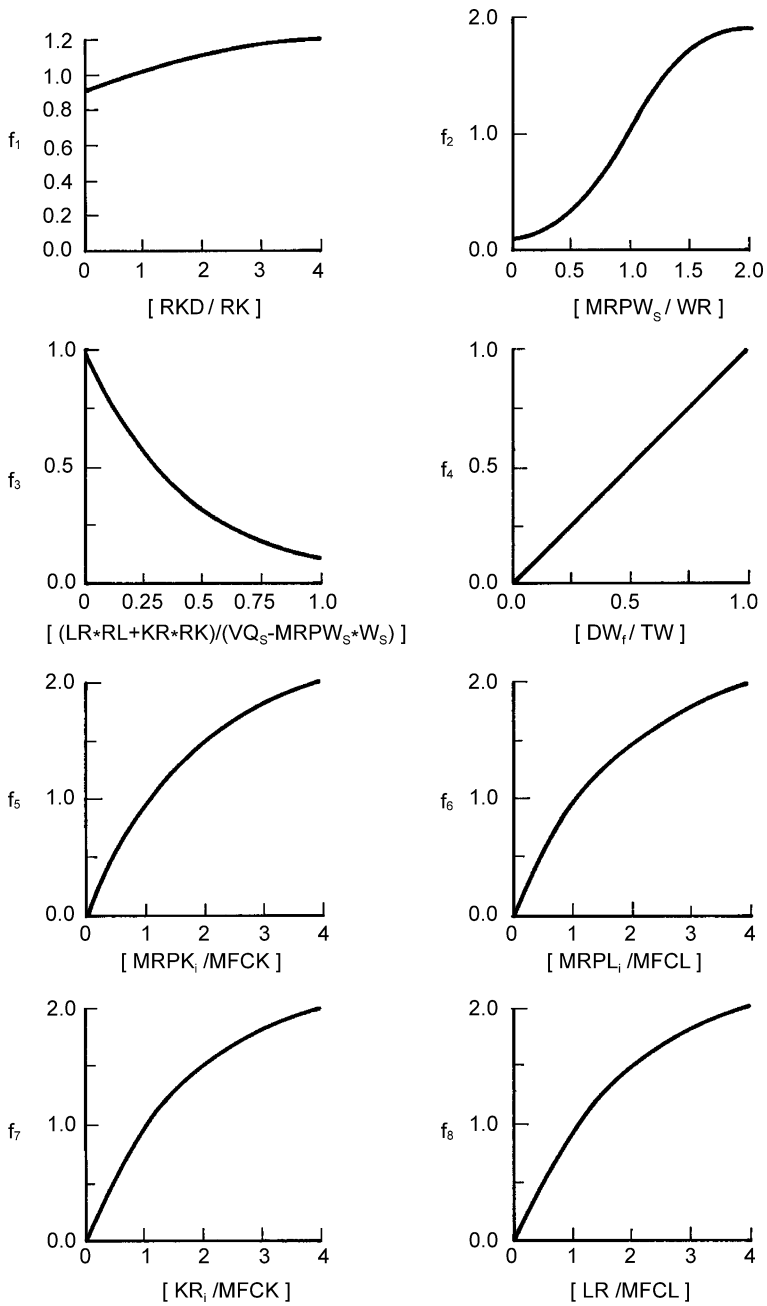

Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 11
Behavioral relationships f_1 through f_8

pends on the ability to accommodate the technology represented by it. Inventory availability of each type of capital KIA^j also limits its purchases.

$$\mathrm{d}/\mathrm{d}t[KO_i] = \sum_j KA_i^j - KO_i/LK\,, \tag{35}$$

$$KA_i^j = DKA_i^j * KIA^j\,, \tag{36}$$

$$DKA_i^m = (KO_i/LK) * f_5[MRPK_i/MFCK] * f_{11}[CA_i] * TCF_i\,, \tag{37}$$

$$DKA_i^t = (KO_i/LK) * f_5[MRPK_i/MFCK] * f_{11}[CA_i] * (1 - TCF_i)\,, \tag{38}$$

where DKA_i are desired capital acquisitions, $f'_{11} \geq 0$, and

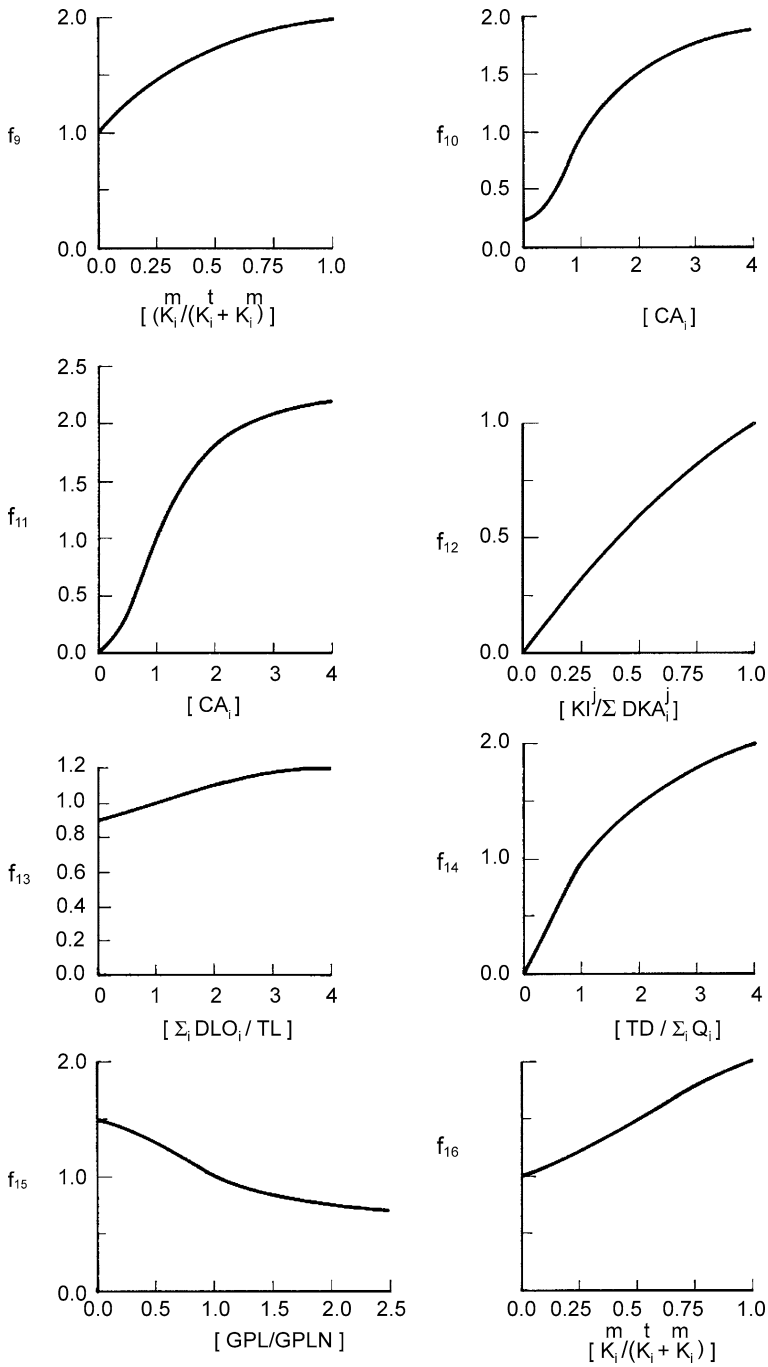

Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation, Figure 12
Behavioral relationships f_9 through f_{16}

LK is life of capital. TCF_i represent exogenously defined technological capability. $0 < TCF_i < 1$.

$$KIA^j = f_{12}\left[KI^j / \left(\sum_i DKA_i^j\right) * KIC\right], \tag{39}$$

where $0 \leqslant f_{12} \leqslant 1$, $f'_{12} > 0$, and KIC is capital inventory coverage

$$\mathrm{d}/\mathrm{d}t[KI^j] = KQ^j - \sum_i KA_i^j, \tag{40}$$

where KQ^j represent supply of capital. KQ^m is imported, while KQ^t is created within the economy by allocating a part of the capacity to its production.

$$KQ^t = \sum Q_i * \left(\sum_i DKA_i^t / TD \right) . \tag{41}$$

The price of land PL is assumed to adjust towards indicated price of land IPL which is given by the economy-wide average of the marginal revenue product of land $AMRPL$, interest rate IR and the desired land ownership in each sector DLO_i

$$\mathrm{d}/\mathrm{d}t[PL] = (IPL - PL)/LPAT \,, \tag{42}$$

$$IPL = (AMRPL/IR) * f_{13} \left[\sum_i DLO_i / TL \right] ; \quad \text{where} \quad f'_{13} > 0 \,. \tag{43}$$

General price level GPL is determined by supply and demand considerations.

$$\mathrm{d}/\mathrm{d}t[GPL] = GPLN * f_{14}[TD/\sum_i Q_i] \tag{44}$$

where $f'_{14} > 0$. $GPLN$ is normal value of GPL and TD is total demand for goods and services to be produced within the economy. TD is given by adding up non-food consumption C_i, traditional capital acquisition KA_i^t and production of traditional capital for inventory, food demand FD and government spending G which is equal to taxes, if any, collected.

$$TD = \sum C_i + \sum_i DKA_i^t + \left(\left(KIC * \sum_i DKA_i^t - KI^j \right) / IAT \right) + FD + G \,, \tag{45}$$

$$\mathrm{d}/\mathrm{d}t(C_i) = \frac{[(((R_i * (1 - SP_i) + AS_i/LAS)/GPL) * FNFC_i) - C_i]}{CAT} \,, \tag{46}$$

where IAT is inventory adjustment time, $FNFC_i$ fraction non-food consumption, and CAT is consumption adjustment time. Food demand FD is given by multiplying population P with normal per capita food demand $NFPCD$ and a function f_{15} representing a weak influence of price.

$$FD = P * NFPCD * f_{15}[GPL/GPLN] \,, \tag{47}$$

where $f'_{15} < 0$ and P bears a fixed proportion with total workforce TW.

The elasticity of production of land E_i^{l} is assumed to be constant as is suggested by empirical evidence concerning agricultural economies [Strout 1978, Heady and Dillon 1961]. Elasticity of production of capital E_i^{k} depends on the technology of production, which is determined by the proportions of traditional and modern capital employed. Since constant returns to scale are assumed, E_i^{w} is given by (47).

$$E_i^{\mathrm{k}} = f_{16}[K_i^m/(K_i^t + K_i^m)] \,; \quad f'_{16} > 0 \,, \tag{48}$$

$$E_i^{\mathrm{w}} = 1 - E_i^{\mathrm{k}} - E_i^{\mathrm{l}} \,. \tag{49}$$

Behavioral Relationships

Sixteen behavioral relationships $[f_1 \cdots f_{16}]$ have been incorporated into the model. The slope characteristics of these relationships have already been described in above equations. The graphical forms of the functions representing these relationships are shown in Figs. 11 and 12 placed below. General considerations for specifying such relationships are discussed in [63].

Bibliography

Primary Literature

1. APO (1985) Improving productivity through macro-micro linkage. Survey and Symposium Report. Tokyo: Asian Productivity Organization
2. Applebaum E (1979) The labor market. In: Eichner A (ed) A Guide to Post-Keynesian Economics. ME Sharpe, White Plains, New York
3. Averitt RT (1968) The dual economy: the dynamics of american industry structure. Norton, New York
4. Bardhan PK (1973) A model of growth in a dual agrarian economy. In: Bhagwati G, Eckus R (eds) Development and planning: essays in honor of Paul Rosenstein-Roden. George Allen and Unwin Ltd, New York
5. Barro RJ (1997) Macroeconomics, 5th edn. MIT Press, Cambridge
6. Baumol WJ (1988) Is entrepreneurship always productive? J Dev Plan 18:85–94
7. Boeke JE (1947) Dualist economics. Oriental economics. Institute of Pacific Relations, New York
8. Cornwall J (1978) Growth and stability in a mature economy. John Wiley, London
9. Denison EF (1974) Accounting for United States economic growth 1929–1969. Brookings Institution, Washington
10. Eichner A, Kregel J (1975) An essay on post-Keynesian theory: a new paradigm in economics. J Econ Lit 13(4):1293–1314
11. Eisner R (1989) Divergences of measurement theory and some implications for economic policy. Am Econ Rev 79(1):1–13

12. Fie JC, Ranis G (1966) Agrarianism, dualism, and economic development. In: Adelman I, Thorbecke E (eds) The Theory and Design of Economic Development. Johns Hopkins Press, Baltimore
13. Forrester JW (1979) Macrobehavior from microstructure, In: Karmany NM, Day R (eds) Economic issues of the eighties. Johns Hopkins University Press, Baltimore
14. Forrester JW (1987) Lessons from system dynamics modelling. Syst Dyn Rev 3(2):136–149
15. Galbraith JK (1979) The nature of mass poverty. Harvard University Press, Cambridge
16. Gordon DM (1972) Economic theories of poverty and underemployment. DC Heath, Lexington
17. Graham AK (1977) Principles of the relationships between structure and behavior of dynamic systems. Ph D Thesis. MIT, Cambridge
18. Griffin K, Ghose AK (1979) Growth and impoverishment in rural areas of Asia. World Dev 7(4/5):361–384
19. Griffin K, Khan AR (1978) Poverty in the third world: ugly facts and fancy models. World Dev 6(3):295–304
20. Hicks J (1940) The valuation of the social income. Economica 7(May):163–172
21. Higgins B (1959) Economic development. Norton, New York
22. Hirshliefer J (1976) Price theory and applications. Prentice Hall, Englewood Cliffs
23. Kaldor N (1966) Marginal productivity and the macro-economic theories of distribution. Rev Econ Stud 33:309–319
24. Kaldor N (1969) Alternative theories of distribution. In: Stiglitz J, Ozawa H (eds) Readings in modern theories of economic growth. MIT Press, Cambridge
25. Kalecki M (1965) Theory of economic dynamics, revised edn. Allen and Unwin, London
26. Kalecki M (1971) Selected essays on dynamics of capitalist economy. Cambridge Univ Press, London
27. Kindelberger C, Herrick B (1977) Economic development, 3rd edn. McGraw Hill, New York
28. Leontief W (1977) Theoretical assumptions and non-observable facts. In: Leontief W (ed) Essays in Economics, vol II. ME Sharpe, White Plains
29. Lewis WA (1958) Economic development with unlimited supply of labor. In: Agarwala I, Singh SP (eds) The Economics of Underdevelopment. Oxford University Press, London
30. Lipton M (1977) Why poor people stay poor. Harvard University Press, Cambridge
31. Marglin SA (1984) Growth, distribution and prices. Harvard Univ Press, Cambridge
32. Marx K (1891) Capital. International Publishers, New York (Reprinted)
33. McKinnon RI (1973) Money and capital in economic development. The Brookings Institution, New York
34. Minsky H (1975) John Maynard Keynes. Columbia University Press, New York
35. Mukhia H (1981) Was there feudalism in indian history. J Peasant Stud 8(3):273–310
36. Myrdal G (1957) Economic theory and under-developed regions. Gerald Duckworth Ltd, London
37. Pack SJ (1985) Reconstructing Marxian economics. Praeger, New York
38. Quinn JB (1985) Managing innovation: controlled chaos. Harvard Bus Rev 85(3):73–84
39. Ricardo D (1817) Principles of political economy and taxation, Reprint 1926. Everyman, London
40. Richardson GP (1991) Feedback thought in social science and systems theory. University of Pennsylvania Press, Philadelphia
41. Riech M, Gordon D, Edwards R (1973) A theory of labor market segmentation. Am Econ Rev 63:359–365
42. Roberts EB (1991) Entrepreneurs in high technology, lessons from MIT and beyond. Oxford University Press, New York
43. Robinson J (1955) Marx, Marshal and Keynes: three views of capitalism. Occasional Paper No. 9. Delhi School of Economics, Delhi
44. Robinson J (1969) The theory of value reconsidered. Aust Econ Pap June 8:13–19
45. Robinson J (1978) Contributions to modern economics. Basil Blackwell, Oxford
46. Robinson J (1979) Aspects of development and underdevelopment. Cambridge University Press, London
47. Roulet HM (1976) The Historical context of Pakistan's rural agriculture. In: Stevens RD et al (eds) Rural development in Bangladesh and Pakistan. Hawaii University Press, Honolulu
48. Rydenfelt S (1983) A pattern for failure. Socialist economies in crisis. Harcourt Brace Jovanavich, New York
49. Saeed K (1988) Wage determination, income distribution and the design of change. Behav Sci 33(3):161–186
50. Saeed K (1990) Government support for economic agendas in developing countries. World Dev 18(6):758–801
51. Saeed K (1991) Entrepreneurship and innovation in developing countries: basic stimulants, organizational factors and hygienes. Proceedings of Academy of International Business Conference. Singapore, National University of Singapore
52. Saeed K (1992) Slicing a complex problem for system dynamics modelling. Syst Dyn Rev 8(3):251–261
53. Saeed K (1994) Development planning and policy design: a system dynamics approach. Foreword by Meadows DL. Ashgate/Avebury Books, Aldershot
54. Saeed K (2002) A pervasive duality in economic systems: implications for development planning. In: "Systems dynamics: systemic feedback modeling for policy analysis". In: Encyclopedia of life support systems (EOLSS). EOLSS Publishers, Oxford
55. Saeed K (2005) Limits to growth in classical economics. 23rd International Conference of System Dynamics Society, Boston
56. Saeed K, Prankprakma P (1997) Technological development in a dual economy: alternative policy levers for economic development. World Dev 25(5):695–712
57. Sen AK (1966) Peasants and dualism with or without surplus labor. J Polit Econ 74(5):425–450
58. Sen AK (1999) Development as freedom. Oxford University Press, Oxford
59. Simon HA (1982) Models of bounded rationality. MIT Press, Cambridge
60. Skinner A (ed) (1974) Adam Smith: The wealth of nations. Pelican Books, Baltimore
61. Solow R (1988) Growth theory and after. Am Econ Rev 78(3):307–317
62. Sraffa P (1960) Production of commodities by means of commodities. Cambridge University Press, Cambridge

63. Sterman J (2000) Business dynamics. McGraw Hill, Irwin
64. Streeten P (1975) The limits of development studies. 32nd Montague Burton Lecture on International Relations. Leads University Press, Leeds
65. Takahashi Y et al (1970) Control and dynamic systems. Addison-Wesley, Reading
66. Weintraub S (1956) A macro-economic approach to theory of wages. Am Econ Rev. 46(Dec):835–856

Books and Reviews

Atkinson G (2004) Common ground for institutional economics and system dynamics modeling. Syst Dyn Rev 20(4):275–286
Ford A (1999) Modeling the environment. Island Press, Washington DC
Forrester JW (1989) The system dynamics national model: macrobehavior from microstructure. In: Milling PM, Zahn EOK (eds) Computer-based management of complex systems: International System Dynamics Conference. Springer, Berlin
Forrester N (1982) A Dynamic synthesis of basic macroeconomic theory: implications for stabilization policy analysis. Ph D Dissertation, MIT
Forrester N (1982) The life cycle of economic development. Pegasus Communications, Walthum
Hines J (1987) Essays in behavioral economic modeling. Ph D Dissertation, MIT, Cambridge
Mass NJ (1976) Economic cycles, an analysis of the underlying causes. Pegasus Communications, Walthum
Radzicki M (2003) Mr. Hamilton, Mr. Forrester and a foundation for evolutionary economics. J Econ Issues 37(1):133–173
Randers J (1980) Elements of system dynamics method. Pegasus Communications, Walthum
Richardson GP (1996) Modeling for management: simulation in support of systems thinking. In: Richardson GP (ed) The international library of management. Dartmouth Publishing Company, Aldershot
Saeed K (1980) Rural development and income distribution, the case of pakistan. Ph D Dissertation. MIT, Cambridge
Saeed K (1994) Development planning and policy design: a system dynamics approach. Ashgate/Avebury Books, Aldershot
Saeed K (1998) Towards sustainable development, 2nd edn: Essays on System Analysis of National Policy. Ashgate Publishing Company, Aldershot
Saeed K (2002) System dynamics: a learning and problem solving approach to development policy. Glob Bus Econ Rev 4(1):81–105
Saeed K (2003) Articulating developmental problems for policy intervention: A system dynamics modeling approach. Simul Gaming 34(3):409–436
Saeed K (2003) Land Use and Food Security – The green revolution and beyond. In: Najam A (ed) Environment, development and human security, perspectives from south asia. University Press of America, Lanham
Saeed K (2004) Designing an environmental mitigation banking institution for linking the size of economic activity to environmental capacity. J Econ Issues 38(4):909–937
Sterman JD (2000) Business dynamics. Systems thinking and modeling for a complex world. Irwin McGraw Hill, Boston
Wolstenholme E (1990) System enquiry, a system dynamics approach. John Wiley, Chichester

Dynamics of Language

Robert Port
Linguistics/Cognitive Science, Indiana University, Bloomington, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Speech and Meters
Further Problems with an Alphabetical Model of Language
Two Levels of Complex System
Language as a Social Institution
Realtime Language Processing
Future Directions
Bibliography

Glossary

Phone A 'minimal' speech sound whether consonant or vowel, the unit that is represented by a single letters of the International Phonetic Alphabet. A phone is invariant across syllable positions, neighboring context, speaking rate, speaker, etc.

Phoneme An abstract speech sound unit in a particular language, typically consisting of several phone types that are treated as the same (despite any differences) by speakers of the language. Thus, English /t/ (slashes indicate use of a symbol as a phoneme, not a phone) includes both an **allophone** (a particular phoneme variant) that is aspirated (as in the word *take*), another allophone that is a glottal stop (in the usual American pronunciation of *cotton*) and another that is a flap or tap (as in *butter*).

Phonology The branch of linguistics that studies the nature of the speech sound patterns of particular languages, such as the inventory of phonemes, the patterns of syllable construction, stress patterns, etc. Thus for the English word *stops*, the *st-* is the onset of the syllable, the vowel is the nucleus and the coda is the *–ps*. Phonology should be concerned with intonation and speech timing as well, although these are not traditional interests.

Meter An underlying temporal pattern of a fixed number of 'beats' within a larger cycle. Thus, there are meters consisting of 2 beats per measure, 3 beats or 4, 5 or 6. Speech often aligns itself with such patterns, e. g., in

chant or song, by locating vowel onsets close in time to the pulse of each cycle.

Orthography The set of conventions about how to write a language. This includes conventions about which letters are used and what sounds they represent in the ideal case. The conventions also include standard spellings for every word, plus conventions of capitalization, punctuation, the definition of a sentence and so forth.

Definition of the Subject

Language has been a topic of great interest at least since the Greek philosophers. Clearly speech is one of the most sophisticated and complex of human capabilities and has remained a major puzzle through the centuries. Complex systems promise to provide analytical tools that will help break through some of the mystery. From this perspective, it appears that there are three complex systems relevant to speech. The first is natural selection, the slow influences on the gene pool of the species that give us whatever innate skills we need to learn at least one language. However, we will have nothing further to say about this issue here since the time scale involved is so long relative to human history. The remaining two complex systems are highly relevant. They are, first, the language itself, viewed here as a particular type of social institution, a set of cultural practices that present the language learner with a broad range of linguistic materials with which to communicate with others. This system clearly evolves as history and the cultural practices of the community change. It sometimes requires only a couple hundred years (a dozen or so generations) for a language to split into mutually incomprehensible dialects.

The final complex system is the one that the individual speaker develops as he/she gradually becomes competent in the use of the language. This system begins before birth and continues to learn and adapt right up until death as the language environment of an individual changes. It involves controlling the many muscle systems of the vocal tract and respiratory system as well as specializations of the auditory system that customize it for hearing and understanding speech in the ambient language or languages.

There is one historical development of central importance that greatly complicates our ability to apply our intuitions to language. This is the 3 thousand-year-old tradition of training our children to read and write using an alphabet. This training, involving hundreds of hours over many years, typically results in our hearing speech as though it consists of a sequence of discrete letter-sized sounds, the consonants and vowels. The lifelong training and practice interpreting letter strings as continuous speech and interpreting speech as a string of letters has, in this author's opinion, encouraged unwarranted confidence in the reliability of our intuitions about the structure of language. Thus, one goal of this essay is to show that our intuitions about the segmental structure of speech are untrustworthy.

Introduction

What kind of structures are exhibited by a natural spoken language? Lay people think first of *words*. Words are the archetypal units of language. It seems every language displays sequences of words and must have a dictionary, its foundational list of words. Linguists agree with this intuition and add another unit that may be somewhat less obvious to laymen: the speech sound, the *phone* or phoneme, that is, the fragment of speech that is represented by a single *letter* from some alphabetical orthography or from the International Phonetic Alphabet [23]. To begin the study of language by using letters and a list of words seems very natural but turns out to be problematic if our goal is to understand linguistic cognition because letters (and phonemes) provide inadequate models of speech. The main problem is that letters (and phonemes) provide too few bits of information to represent what speakers need to know about speech. One unfortunate consequence of this supposed efficiency is that only the ordering of letters can represent the temporal layout of speech events. Let us begin then by making observations, not about our intuitive description of language structure in terms of an alphabet, but rather by making observations of some specific realtime linguistic behavior.

Demonstration

In order to provide an intuitive example of realtime linguistic behavior, the reader is urged to participate in a tiny experiment on oneself. The reader is encouraged to repeat the following phrase out loud: *Take a pack of cards*. Please repeat this phrase over and over – aloud (a whisper will work fine), between 5 and 10 times. Now. After cycling through the phrase, notice the rhythm pattern that you adopted unconsciously. When a person repeats a short phrase like this over and over, it seems the easiest way to do it is to slip into some rhythmic timing pattern that is the same from repetition to repetition. A remarkable fact is that there is only a small number of ways that most people will arrange the timing of a cycled phrase like this one. To discover which rhythm you used, repeat the phrase again just as you did before but tap a finger in time with the stressed syllables of the phrase. You should find your-

self tapping with only some of the syllables of the phrase, probably at least the first and last words, *Take* and *cards*. Now there appear to be only 3 rhythmic patterns for reading phrases like this that are readily available to English speakers, and probably speakers of some other languages as well [12]. The three preferred patterns are shown in Example 1 where underline and bold face mark stressed syllables that should align with finger taps. The reader is urged to produce a few cycles of the three discrete patterns in Example 1. The notations |: and :| are borrowed from music notation to indicate cycling of the pattern between them.

Example 1

(a) |: **Take** a pack of **cards**, **Take** a pack of **cards**, :|
(b) |: **Take** a pack of **cards** [***rest***],
Take a pack of **cards** [***rest***], :|
(c) |: **Take** a **pack** of **cards**, **Take** a **pack** of **cards**, :|

Pattern (1a) has 2 beats per cycle, a beat on *Take* and another on *cards* with the next beat on *Take* again. There is a finger tap for each beat (but none on *pack)*. Pattern (1b) has 3 beats (and taps) per cycle: the first on *Take*, the second on *cards* and the third is a beat that has no syllable, just a longer pause than (1a) has – equivalent to a musical 'rest'. Patterns (a) and (b) must have a long period so that *Take a pack of* can be spoken in a single beat-cycle. Example (1c) also has 3 beats, on *Take, pack* and *cards*, and, as in the waltz rhythm, the repetition cycle begins immediately with another *Take*. Most Americans first discover pattern (a) of Example 1, but most people can produce all three without much difficulty. To those familiar with music notation, we could say Example (1a) has a 2-beat measure with *Take* and *cards* beginning at each beat and dividing the measure into equal halves. Both of the others have a 3-beat measure but (1b) has a rest on beat 3. The unstressed syllables seem to be squeezed or stretched to assure that the salient target syllable onsets begin at the proper time, that is, as close as possible to phase zero of one or both of the coupled oscillators generating the metrical time structure.

Other timing patterns can be achieved with a little training. Try pronouncing the sample phrase with equal stress on all syllables. Be sure to repeat the phrase several times as in Example 2.

Example 2 |: **Take** **a** **pack** **of** **cards**, [***rest***] :|
(1 2) (3 4) (5 6)

The easiest way (for English speakers) to repeat the phrase so that you put equal stress on all 5 syllables is to use a 6-beat meter (probably with emphasis on beats 1, 3, and 5) and with a rest on beat 6. For another attractor, one can leave out the rest on beat 6 but to immediately return to Take after *cards*. This creates a 5-beat pattern. (If leaving out the rest on beat 6 proves difficult for the reader, then count, **1**, 2, **1**, 2, 3, **1**, 2, **1**, 2, 3 a few times, then switch to the test phrase.) English speakers can learn to produce these more exotic 5- and 6-beat meters, but it takes some effort and practice to achieve them. There is some anecdotal evidence, however, that both Japanese and Finnish speakers, for example, are able to effortlessly produce the 5-beat pattern. A few experimental tasks like these have been looked at with speakers of several languages – at least Arabic, Japanese, German and Finnish [53,57]. Everyone tested so far seems to find at least one of the temporal patterns of Example 1 to be very natural and almost unavoidable for many short phrases that are cycled in any language.

Since it is likely that differences between languages in preferred rhythm patterns will continue to hold up, it is likely that rhythmic patterns are an intrinsic aspect of any language, and thus that research on the phonology of a language must begin to rely on continuous time data. These language-specific constraints on speech timing (as well as many subtle phonetic differences between dialects and languages) show that it is reckless to begin the study of language by first encoding all speech data into a small set of uniform letter-like tokens that have only serial order instead of real time. The problem is that, even though our intuitions about language tell us letters are obvious units of language, letters are actually a culturally-transmitted technology that humans have learned to apply to (and, if necessary, impose upon) spoken language. Letters and their psychological counterparts, phones and phonemes, have very weak experimental support as psychological units used for cognitive "spelling" of words [42,43,45,47].

Speech and Meters

Given the intuitive descriptions of the timing of repeated short phrases, an experimental study is needed to verify that these three patterns are the primary ones that English speakers can reliably produce. But first, for rhythmically produced speech, an important question arises: just what in the acoustics are subjects aligning with the finger taps? That is, what is the physical definition of a "beat" or auditory pulse? Most scientists working in this area of speech agree that it is approximately the onset of a vowel [2,12,30,49]. This is usually located by low-pass filtering of speech (at about 800 Hz) and integrating the remaining audio signal over about a 250 ms time window and looking for positive peaks in the second derivative that indicate a rapid increase in the energy in the low frequen-

cies. That is where people tap their finger in time to a syllable.

In the experiment, speakers were asked to cycle short phrases like *Dig for a dime* and *King for a day* and to put the final noun (e. g., *dime*) at specific phase lags relative to the *Dig-Dig* interval [12]. The target phase angle was indicated by a metronomic pattern of 2 tones repeated 10–15 times per trial. Can speakers begin the word *dime* anywhere we ask within the cycle? It was expected the participants would find it easier to locate the onset at the simple harmonic fractions discovered above, i. e., one half, one third and two-thirds of the cycle. These phase angles might be accurately reproduced and other phase lags would be less accurately reproduced. Subjects were played a metronomic pattern of 2 tones, one High and one Low which they listened to for several cycles. Then they were to repeat the indicated phrase 10 or 12 times in a row aligning the Low tone with *Dig* and the High tone with *dime*. Now in the stimulus patterns, the phase angle φ by which the High tone (i. e., for *dime)* divided the Low-Low interval (i. e., *Dig-Dig)* was varied randomly and uniformly over the range from $\varphi = 0.20$ to 0.70. The overall period of *Dig-Dig* was also varied inversely with the ratio. The earlier the target phase angle, the slower the tempo of the repetition cycle, so as to leave a constant amount of time for pronouncing the phrase.

If the American participants were able to imitate those phase angles without bias, as they were asked to do, then they should produce a flat, uniform distribution of the onset of the vowel of *dime* relative to *Dig* onsets over that range – thereby replicating the pattern of the stimulus metronome. But Fig. 1 shows the observed frequency histogram of the median phase angle of the onset of *dime* for each trial across 8 speakers. The histogram represents about 1400 trials where each trial consisted of 12–15 cycles of a test phrase said to a single metronome pattern.

Clearly the participants exhibited a strong tendency to locate their syllable onsets near just the 3 points: 1/3 of the cycle, 1/2 of the cycle or 2/3 of the cycle corresponding to the readings of Example (1b, 1a) and (1c) respectively. But they do not show evidence of any preference for fifths of the repetition cycle – no hint of peaks at 0.4 or 0.6. The subjects were trying to do as they were told, but when the high tone fell at, e. g., $\varphi = 0.4$, the speakers tended to produce φ closer either to 0.5 or 0.33.

Dynamical Model of Meter

These timing phenomena can be accounted for with dynamical models for neural oscillations correlated with auditory patterns [27,32,47]. A natural interpretation for the three peaks is that they represent attractors of a cyclical metric pattern specified by 2 coupled oscillators locked into one of 2 patterns, either a 2-beats-per-cycle pattern or 3-beats-per-cycle pattern. An important point about meter is that it is audition, the auditory sense, that has the

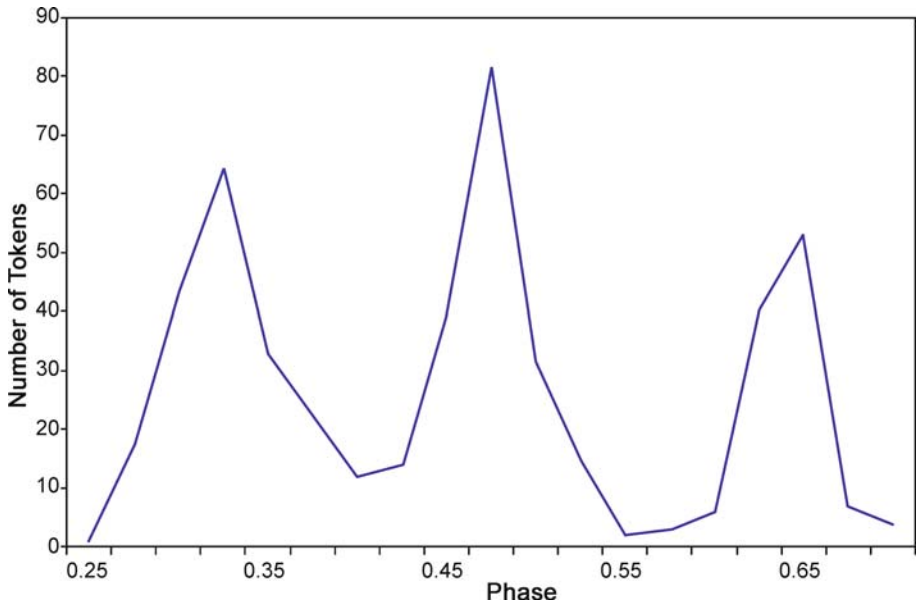

Dynamics of Language, Figure 1
The distribution of the vowel onsets for the last stressed syllable by 8 speakers saying phrases resembling *Dig for a dime* with target phase angles specified by a metronomic sequence of 2 tones. Figure from [44]

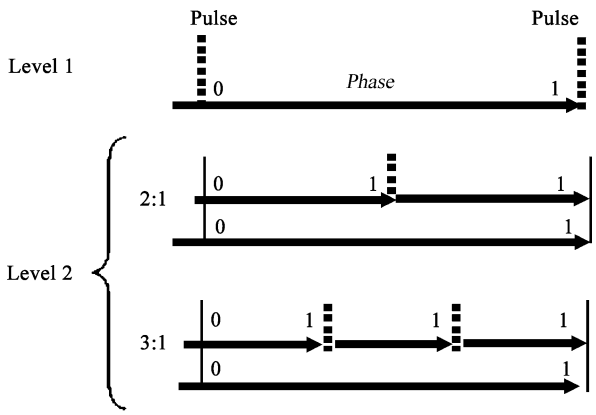

Dynamics of Language, Figure 2
Simple meters represented as circles opened out into a line. Phase 0 and Phase 1 are, of course, the same point on the circle. The upper panel shows a single oscillator with periodic phase zero serving as an attractor for perceived or produced energy onsets. The second and third panels show two ways that a second oscillator can be coupled with the first. Each produces either 2 cycles or 3 cycles before coinciding with the slower cycle. It is assumed that all phase zeros attract energy onsets both in speech and nonspeech sound. The 3 patterns of Example 1 can be accounted for by the 2:1 and 3:1 coupled oscillator or meter systems. Figure from [44]

strongest link to meter. Of course, we cannot tell from these data whether this bias toward the 3 attractor phase angles comes primarily from the participants' perception systems (so, e. g., $\varphi = 0.4$ simply *sounds* like $\varphi = 0.5$) or if the distortion reflects the influence of their motor system which might find it easier to consistently produce the auditorily salient pulse at one of the attractors.

Notice that these patterns involve cycles at two time scales nested within each other as shown in Fig. 2. The slower cycle is the repetition cycle of the whole phrase, represented in the upper panel in the figure. Its phase zero coincides with the onset of *Dig*. For all the phrases in this experiment, there is also a second cycle (thereby implying a state space that is a torus although a torus is inconvenient for display). The second one is faster and tightly coupled with the repetition-cycle oscillator. It cycles either 2 or 3 times before its phase zero aligns with phase zero of the slower oscillator [44]. If we assume that the temporal location of any phase zero is an attractor for a prominent auditory onset, then such a system of 2 coupled oscillators could define several musical meters and account for the 3 powerful patterns of Example 1.

Of course, subject behavior in "speech cycling" experiments is not the only evidence for a preference by humans for periodic speech. Seemingly all human communities have developed traditions of religious, martial or commercial chants and songs as well as other genres of vocal music that build on the metrical principles exhibited by people performing basic speech repetition. So the phenomena reviewed here point toward several important lessons about how to analyze human speech behavior.

The first lesson is that human speech performance, like so many other kinds of dynamical systems, finds periodic cycling easy and natural to do even though there is no mass-bearing object (such as a leg or finger) oscillating at the observed frequencies. The speech periodicities reflect oscillations that must be occurring in various regions of the brain stem, cortex and cerebellum see [38,40,56]. Humans, in just about any language, can warp their speech into singsong or chant very easily. We seem to enjoy entraining our speech to each other see [11] and to external auditory patterns [41]. Just above, the reader was asked simply to repeat a phrase and within a couple repetitions, the performance slipped into one of a small number of periodic attractors that this particular fragment of English can settle into. Once there, the speech production system tends to stay there until you force yourself out of the attractor (e. g., by trying to obey impertinent instructions from the author about assigning stress to bold-face syllables and tapping a finger).

The most important lesson from the phenomena reviewed so far is that we must be very careful about taking alphabetical forms of language as the basic data for a study of human speech. We may do so (as most linguistic research has done for a century) for convenience but must keep in mind that alphabetical representation makes constrained timing relationships like those revealed in the demonstration in the "Introduction" completely invisible, since letters and words in their written form exhibit only serial order, not continuous time. It seems like for this speech cycling task the speech is riding on the dynamics of low frequency oscillations. Examples 1a, 1b and 1c are, of course, identical in terms of alphabetical description, but they are far from identical in timing. Speakers of different languages exhibit different rhythmic effects in speech cycling and in other aspects of timing. Clearly it makes little sense to define language as somehow completely independent of these powerful speech behaviors.

Further Problems with an Alphabetical Model of Language

The invisibility of real time is one reason to reject representations of language using an alphabet. There are many more reasons. For example, the results of the demonstration and the laboratory experiment show that many properties of human speech resemble those of familiar dynam-

ical systems in which oscillatory patterns are frequently observed [1]. But within linguistics and much of psychology, language has been defined for the past century or so as a 'symbol system' or a 'code' shared by the speaker and hearer [3,7,13,19,37]. These timing patterns cannot be symbolically coded since they are defined by measurements in continuous time. The tendency of people to warp speech timing to fit patterns of integer-ratio time intervals is inexplicable by any model consisting only of serially ordered symbols – like words or the consonant and vowel segments. There are many kinds of temporal patterns in various languages that are inexpressible with any alphabet. So before approaching the study of speech and language in dynamical terms, we need to clarify that we will address a domain that is much broader than 'language' as most linguists and psychologists have considered it.

The domain of phenomena for linguistic research needs to be ***speech events in continuous time*** of people in contextualized linguistic performance. This kind of raw data can easily be recorded, thanks to recent technologies on audio and video or partially displayed as a sound spectrogram (where time is mapped onto one spatial axis). But transcription or reduction of raw speech sound into any alphabet cannot be presumed, whether a hypothetical psychological alphabet such as of (language specific) phonemes, or an orthographic alphabet (such as the form you see printed on this page), or a technical phonetic alphabet (e. g., the International Phonetic Alphabet used by linguists for all languages for over a century now). Letters throw away far too much of what is critical for the dynamics of language as used by speakers.

In addition to speech rhythm, a second major problem with the traditional view is that it predicts invariant acoustic correlates for each segment. But it has been known for almost 50 years, that speech perception relies on formant trajectories and details of spectral shape and spectral change that cannot be captured by any invariant, context-free acoustic cues (e. g., [34,35]). Speech cues for phonetic segments simply cannot be defined so as to do the job of specifying phones or phonemes independently of the identity of neighboring vowels and consonants, speaker identity, speaking rate, etc. despite many attempts to do so see [28,52]. Thus, the shape of the second formant that specifies [d] before an [u] (as in English *do*) shows a falling trajectory whereas the second formant rises for [d] before an [e^I] (as in *day*). Such examples are found throughout the domain of speech cues. The acoustic cues for each phonetic segment tend to be different for each context, different enough so that a general definition can not be formulated that works across a range of different contexts. One consequence of these failures is that engineers working on speech recognition long ago abandoned attempts to recognize words by first recognizing segments or segmental features as a way of specifying words. Instead, the engineers try to specify whole words and phrases directly in terms of acoustic spectra rather than try to detect the kind of segmental units that linguists continue to insist are essential for word specification. By representing words using a codebook of spectrum slice descriptions, successful artificial speech perception systems effectively store context-specific versions of the phones and phonemes [22,25].

A third difficulty with the view that spoken words have a phonemic spelling for memory is that it makes the wrong prediction about the results of recognition memory experiments, where a subject indicates whether each word in a list is repeated (that is, has occurred earlier in the list) or not. The data show that words repeated by the same voice have a significant advantage in recognition accuracy over words that are repeated in a different voice. This shows that subjects do retain speaker-specific information (often called "indexical information") in memory. This is difficult to reconcile with the claim that words are stored using only representations like phones or phonemes that are abstracted away from speaker-specific properties indicating the speaker's voice [14,39,43,45].

Finally, if words are perceived, produced and stored using a discrete list of distinct, abstract phones or phonemes, then linguistic variation of all kinds – dialect variation, historical changes in pronunciation, speaker idiosyncrasies, etc. – should all show discrete jumps in pronunciation between these discrete sound types. Speech should not show continuous variation in parameters like vowel height or backness, degrees of lip rounding, place of articulation, voice-onset time, segment durations, etc. All these parameters should exhibit audible jumps when the transcription changes, e. g., from [t] to [d] or from [i] to [I] (except for possible noisy variation in production). However, generations of phonetics research reveals no evidence of discrete changes in any of these parameters as far as I can tell [6,32,36,45,46]. The apparent target values of vowel quality show no signs of discrete jumps in target in a plot of the first formant against the second formant. And for voice-onset time, the time lag between the release of a stop and the onset of voicing exhibits many apparently continuous values of VOT target depending on the degree of stress, the following vowel or consonant identity, etc. [45]. Nowhere is there evidence that speech sounds are discretely different from each other in the same way that letters are. Nor is there evidence that the gestures are nonoverlapping in time (quite the opposite, they overlap a great deal, [4]. So, the hypothesis that words are cognitively spelled, that is, represented in memory, in terms of

discrete, alphabet-like units simply finds no experimental support. It is extremely unlikely that the standard view of lexical representations used in linguistics will turn out to be correct.

Could there be some reason we have been overlooking for why the nervous system might not rely on these abstract units like phones and phonemes? One problem is probably that a vast amount of essential information about the speech would be lost in the transcription process. In addition to all temporal properties of speech (such as segment durations, rhythmic patterns, etc.), information about coordination details of the speaker motor activity, details of voice quality and speaker identifying properties, etc. are all discarded when speech is reduced to alphabetical form. Many linguists have suggested that an alphabetical transcription is a "more efficient" representation than an audio or video recording [3,10,24]. But more efficient for what? Is it more efficient as a form of word memory that is useful to a human speaker? This is, of course, an empirical question. The answer depends on how memory for words actually works. For a memory system optimized for dealing with a non-discrete world where non-discrete sound patterns are the norm, perhaps storing as much information as possible works better and might actually be "easier".

Like most modern linguists, Chomsky [7,8] followed his predecessors in the "structural linguistics" movement (e. g., [3,13,21]) in presuming that words have a representation in a psychological alphabet that is small enough that each segment type can be assumed to be distinct from all the others. Every language is expected to exhibit this kind of static 'representation'. Words are presumed to be distinguished from each other by letter-like segments (unless they happen to be homonyms, like *two, too, to,* and *steal, steel, stele*). Yet the evidence reviewed above shows that speakers **do** encode much phonetic detail in their speech perception and in speech productions. Consider all that a good actor encodes into speech: subtle, constantly changing moods, and the region of the speaker's origin, their social class, etc. as well as the text of the screenplay. Listeners, too, are able to interpret the speaker's mood, feelings, region, etc. The traditional picture of language as represented in memory in alphabet-like form finds no more support from these familiar phenomena about speech than it found in experimental studies.

Origin of Alphabetical Intuitions

Where might these intuitions come from if, in fact, words are not stored in the abstract way that we supposed? Why are we so comfortable and satisfied with alphabet-like descriptions of speech? The answer should have been obvious. For roughly the past 3k years, the educated minority in the Western cultural tradition have had a powerful technological scaffold for thinking about language [9,10]. They have had an alphabetical orthography for their language. Writing captures a few important and relatively distinctive properties of speech that can be encoded (by someone with appropriate training) into letters and individual words, as in either an orthographic or a phonetic transcription. Please consider that every reader of this page has spent many hours a week since age 4 or 5, practicing the conversion of letters into words and words into letters. These hundreds of hours of training with an alphabetical orthography surely play a major role in shaping the intuitions that all of us share about language, and surely biased the very definition of language, and thus of linguistics. In the linguistic tradition since Saussure, including the Chomskyan tradition, knowledge of a language is assumed to consist in large part, of linguistic representations of words in memory. Since we are talking about 'knowledge', each word is stored just once – just as in a published dictionary of a language. In addition to the set of lexical tokens, there are supposed to be rules in the "grammar" that manipulate the linguistic representations as simple tokens (moving segments or segment strings around, deleting or inserting them). This symbol-processing mechanism constitutes a powerful assumption about highly abstract representations of words. But why is it so easy for people to make such elaborate assumptions about language? The traditional linguistic representations for phonemes and words, etc., turn out to get many of their properties by generalization from the technology of conventional orthographic letters and words. Because of our experience using the technology of writing, it has come to seem natural to assume that something like this would be used by speakers mentally.

An orthographic system is a collection of customs about how to write a language. Development of a useable orthography requires deciding exactly what the alphabet will be and its application to sounds spoken in the language, which fragments of speech constitute a 'word' and how each word is to be spelled. Other orthographic conventions address what patterns count as a sentence and what words a sentence may or may not begin with. For a learner of English orthography (including linguists and foreign language teachers), one problem is that the relation between letters and sounds is far from consistent [48]. Thus, we no longer pronounce either the **k** or the **gh** in *knight*. The letter **i,** which in most languages sounds like the vowel in *pea*, often sounds like *my* in English, and the vowel [ɛ] can be spelled either **ea** (in *head*), **e** (in *bed*) or

ai (in *said*). Because of these well-known inconsistencies in many languages of Europe, scientists of speech together with second-language pedagogues developed the first version of the IPA phonetic alphabet about 1890. This system employs primarily Roman letters plus many others to describe speech sounds as pronounced with a certain level of detail. But it takes some training to be able to accurately hear some sounds that speakers use. This method pays no attention to 'intended' speech sounds or to orthographic spellings. It can be seen that now the 3k year-old technology of an alphabet was employed for two very different purposes: It plays its traditional role of providing the basis for a useful orthography for various languages but also very recently now provides a consistent notation system for scientific recording of speech gestures and speech sounds in many languages.

It seems likely that the development of the IPA alphabet influenced theories of language. It was immediately recognized that one could transcribe to varying levels of detail, so the distinction between *broad* vs. *narrow* transcription was discussed. Within 20 years of the release of the first IPA alphabet, some language scientists (in particular, a Polish linguist named Beaudoin de Courtenay) began speaking of '*phonemes*' (see [53] for the early history of the term). These were to be a minimal set of abstract speech sounds, the smallest set that seems adequate to write a language. So the phoneme appeared as a kind of *idealized letter* that is hypothesized to describe something in the head of speakers. To see how the phoneme is an idealized letter, consider first the properties of letters. Letters in an alphabetical orthography:

1. *are discretely* different from each other.
2. There is a *fixed small set* of them available in any orthography.
3. They are *serially ordered* and graphically *nonoverlapping*.
4. Each letter has an *invariant shape* across different writers. (Although handwriting and font may differ, the letters are still discretely identifiable.)
5. Each word has a *single canonical spelling* in the 'dictionary' (that is, in the mental inventory of the linguistic representations of the language).

Notice that because in orthographies words are spelled only from letters, words will always be discretely distinct as well. In fact, even syntax is assured of inheriting the discreteness of the orthographic alphabet. The usefulness and efficiency of letters (plus their intimate familiarity to literate scientists) makes it very plausible to hypothesize psychological counterparts to letters. It seems quite natural to imagine a *phone* or *phoneme* as the psychological analogue of a letter. Thus linguists routinely assume that there are:

1. *discrete* differences between *phones* or *phonemes* which are
2. *drawn from a small set* of phoneme types for each language,
3. and are *serially ordered* and *nonoverlapping in time*.
4. Each phone or phoneme has an *invariant physical form* across contexts, speakers, rates of speech, etc.
5. The words of a language have a *single "canonical" representation* in memory, that is, some prototype version to which incoming tokens are compared.

From this traditional view, linguistic memory is primarily a dictionary with every word having its distinctive spelling in a small, abstract alphabet so it can be differentiated discretely from all non-homophones in the language. A major problem has been Assumption 4 about an invariant physical form for each phoneme. A century of phonetics research on phones and phonemes has shown unambiguously that ***there is no way to give these abstract linguistic segments concrete specification***. Successful acoustic definitions for all linguistic segments and features have not been found – in fact, there are no definitions for *any* of the segments or features that are effective across neighboring contexts, speaking rate, etc. Thus there is an incommensurability between letter representations and the actual representations in memory. Very simply, there are no "acoustic letters" for speech. We phoneticians spent a generation or two trying to provide physical acoustic definitions of linguistic units like particular features and segments, but we have not been able to. It almost certainly cannot be done.

So before starting to analyze the dynamics of human speech, we must accomplish some mental self-cleansing. It is important to really understand that letters in the phonetic alphabet are just one kind of model for speech. This model is very useful – especially to those of us with literacy education. It is convenient for a writing system (because it requires learning only a small number of letters). However, such a representation is not, it turns out, closely related to the way speakers, literate or illiterate, store words in memory, recognize them or produce them.

A century ago, the phonetic alphabet provided the best technology available for representing speech for scientific study. But for up to half a century now continuous-time technical displays like audio waveforms, sound spectrograms and smoothed electromyographic plots have been available to show us vividly that letter-like units of speech sound are ***not*** what people use for representing speech. We have been making excuses for this counter-evidence for

over half a century. It is time to give up the illusion that words have a cognitive spelling into units that resemble letters. There are only two places where they really have such a spelling: on paper when people use orthographic letters to write words down, and in the conscious conceptualization of language by literate people.

The conclusion is that we must reject letter-like phones or phonemes as actual units that define spoken language. But then how can we begin to analyze any particular language or human language in general? We must begin afresh with a clean slate, looking at language without the biases that come from our cultural history. This literate culture demands that we develop great skill at interpreting speech sound as letters and reading letters as speech sound.

Two Levels of Complex System

A very different basic cut of the phenomena surrounding human speech is required, one that distinguishes the two primary complex systems that support human speech communication. It is proposed, following Smith, Brighton and Kirby [50], that we separate:

(A) the properties of the ***language patterns of a community of speakers***, that is, what we might call the social institution of language, on one hand, from
(B) the skills of ***real-time speech processing***, on the other.

The "grammar of a language" should be understood as a description of patterns across a community of speakers and contexts summarized over some historically brief time window (of, say, a generation). Linguistics is (or should be) studying the patterns and the cultural categories that are found to define the unit-like chunks in any language. Linguists should describe the most common patterns and the categories of sounds, and make whatever generalizations they are able to, stated in terms of whatever descriptive vocabulary for timing and articulatory state seem to work. But, if we look into the details, it will be clear that all speakers in a community discover their own pattern definitions.

This implies that approximation across speakers is the best that can be done. There is an unavoidable uncertainty about the precise definitions of the auditory micro-features that "spell" words (or word-like units) for any individual. In addition, each speaker has been exposed to a different subset of the corpus of speech in that language (differing in regional and social dialects, foreign accents, etc.). Thus any given language may have a gross description, the kind linguists and phoneticians are trained to produce using the IPA alphabet or one that is provided by an orthographic representation. But the closer we look, the more variation will be found. Spoken languages do not have a consistent or uniform finely detailed description. Very simply, there exists no universal vocabulary with which to describe it. The reason for this is, again, that the fine description in each speaker is created independently by developmental processes based on the language-exposure history of each speaker. This is why phonological observations about the sound patterns of any language exist only at the group level and are not necessarily "in" the individual speaker (nor even in *any* speaker). The idea that phonology only exists in the corpus of a community and not in the individual speaker's representations simply acknowledges that speakers can and will differ from each other to varying degrees.

The acquisition of the skills of real-time speech processing can now be understood in outline. The child, well before birth, begins learning the auditory patterns (human generated and otherwise) [26]. During the first year, the child learns to recognize the typical speech sound-gestures of his ambient language and classifies them into categories [31,55]. On this theory, the child increasingly develops his auditory analysis system to perceive speech in his ambient language and then to produce it. Notice that this system has to work without any a priori description of what the speech patterns of the community actually are. The learner has no idea in advance what "the grammar" really is, so the perceptual system learns as best it can to predict the patterns based on very concrete descriptions. It appears the learner makes progress by storing lots of detail about specific utterances that have been heard. Counter-intuitively (at least to literates like us), the real-time speech system makes no use whatever of the abstract segments or letter-like units that are so prominent in our conscious experience of language. Similarly, speech perception takes place with no necessity of recovering any low-dimensional alphabet-like description along the way.

The story told here implies several observations: (1) discrete categories of linguistic patterns are not the same as discrete symbol structures, (2) the abstract and partly discrete structure of phonology exists in a speaker only as a structure of categories. But phonological categories are sets of speech chunks specified in a rich code that are taken to be 'the same' as other members of some category. Their discreteness is never guaranteed because they are fairly high dimensional. Plus some events will be found where category assignment is difficult to impossible (or ambiguous). Finally note that (3) speakers can exhibit patterns in their behavior that they have no explicit representation of (and thus the patterns are not appropriately described as 'knowledge' of any kind). The child learns to perceive and produce speech that is compatible with the

way others talk in his or her community but each does so using idiosyncratic components (because each is controlling only his own vocal tract and exploiting his own auditory system whose detailed structure reflects the speaker's personal history).

So then, the two complex systems of human speech are, first, the cultural institution, the set of linguistic habits, patterns and category types exhibited by the community of speakers (what we usually call the phonology, grammar and lexicon but also cultural style, etc.), and, second, the individual's working system for speech production (controlling a vast number of articulatory degrees of freedom) and speech perception (employing the speaker's partly idiosyncratic speech sound analysis system). This realtime system seeks sufficient compatibility with the community's usage patterns that the speaker can speak and be understood. The two systems each have their own development processes on very different time scales. The social institution is shaped slowly over time by changes in the statistical patterns of a community of speakers, but the speaker continues to develop new linguistic skills throughout life. We now look closer at each of these systems.

Language as a Social Institution

The first level of complex structure is in the **language as a system of shared speech patterns produced by speakers in a community**. When coming into the world, the child begins to have exposure to a corpus of the language of his community in various social contexts, first from the parents and later from siblings, peers, teachers and pop singers. The system of regularities in the corpus of language a child is exposed to has been self-organized by the speech behavior of the community of speakers of the language over many generations. The result is the existence of patterns that seem to underlie most word shapes, including certain "speech sound" types (typical consonant and vowels of the language), syllable-structure patterns, lexical items and "grammatical patterns." The child learns to perceive and to produce adequate versions of this system of composite patterns as he becomes more skilled at listening to and producing speech [31,55].

There is plenty of evidence of discreteness in these phonological categories. For example, the vowels in the series *bead, bid, bade, bed* seem to be the same as the vowels in *mean, min, Maine, men* and in *peal, pill, pail, Pell* (at least in many English dialects). These similar vowel contrasts suggest a set of categories of words in English (e. g., the category that includes *bead, mean* and *peal* versus a different category that includes *bed, men* and *Pell*). But the claim that these are best understood psychologically as categories and not as symbol types implies greatly weakening the predictions that can be made about them. There are many ways to be assigned to a mere category, some rule-based and easily brought to awareness and others quite arbitrary and explicable only historically. Linguistic categories (such as /i/ or /t/, and what pronunciations we recognize as productions of the word *and*, etc.) are just aspects of the culture of a community. We cannot expect that *bead, mean* and *peel* will be identical in respect to any particular acoustic or auditory property. Speakers do not necessarily represent these three words as sharing any specific feature. They share only the category that we represent with the IPA symbol /i/. We English speakers (or probably primarily educated English speakers) think of them as "sharing the same vowel." It may be alright to speak this way, but it is important to keep in mind that they are not represented by or spelled in memory with any specific symbol token. The same analysis applies to the variants of English /t/ as found in *teach* [t^h], *stop* [t], *butter* [ɾ] and *cat* [$t^ʔ$]. They are considered by all of us to be /t/s at the same time that we know they are not the same but differ greatly. Ultimately, the main reason to call them all varieties of /t/ is that orthographically we conventionally spell them all with a **t** (and, of course, they were once pronounced with more similar **t**-like sounds a few centuries ago).

Small changes in statistical distributions by many individuals gradually have the effect of facilitating a continual self-organization process in the community. For reasons that are not yet clear, this process tends to lead eventually to such typical phonological patterns as (a) the use of a similar vowel inventory in different consonantal contexts, (b) the symmetrical patterns of voiced and voiceless stop and nasal consonants (e. g., [b d g, p t k, m n ŋ]), that appear in many languages, (c) the constraints or preferences on the form of syllables in each language, and so on.

Although languages have been known for almost two centuries now to change gradually over time, research has yet to provide a convincing explanation of how or why these changes occur. Since it is very difficult to study such slow changes in real human communities, various attempts have been made to simulate socially defined language-like systems in computational systems employing various kinds of linguistic agents [5,29,50,51]. This work, much of it based on the techniques of artificial life, seems very likely to contribute to our understanding of the process of language evolution.

Realtime Language Processing

The second level of complex system that is central to human speech is the **psychological system for speech pro-**

duction and perception – the neural control system that manages an individual's physiological speech system, the lips, tongue, velum, larynx and respiratory system plus the cognitive mechanisms that accomplish speech perception and understanding in realtime. The individual speaker acquires these skills in developmental time, that is, within the speaker's lifetime. However, the units of the social institution of language are not very useful for representation in memory or linguistic processing of speakers. Speech perceivers apparently rely on detailed episodic representations – much more like an auditory description of a spectrotemporal pattern than like something spelled using a small, abstract alphabet.

It is now clear that dynamical systems play a central role in speech production and perception skills [54]. The individual speaker-hearer must discover ways to coordinate all their muscular systems for the production of speech and also somehow to simulate the auditory-phonetic patterns heard in the speech of others in a realtime perceptual system. These skills require several years to achieve normal competence at basic speaking and listening. The more successful scientific efforts to model the real-time processing of speech production and speech perception have focused in recent years on dynamical system models, most often implemented as "neural networks" [15,16,17]. These models simulate the realtime processing of language but do not rely much on the letter and word based description.

Speech Production Modeling

There has been a great deal of research and modeling effort in the study of speech production. The challenge here is that speech production is a formidable motor problem that seems to tax the limits of human motor capabilities. It demands impressive feats of coordination and skilled control of a great many articulatory variables. There are many domains for modeling aspects of human speech performance. But, in order to give some feeling for the kind of dynamical models that appear to be successful, the DIVA model of Frank Guenther will be sketched here.

This model addresses both the problem of skilled speech production as it occurs in realtime, as well as the problem of how such skills could be acquired by the child language learner. In addition, this is one of the first models of speech production for which specific neural sites can be proposed for each functional component of the model (although the neurophysiology will not be discussed here). The model gains inspiration from much previous work in neuroscience demonstrating, for example, that the brain contains many 'maps' of cells arranged in two or three-dimensional arrays that are topographically arranged, meaning that specific regions of the map are devoted to specific content (or actions) and that neighboring regions are devoted to similar content (or action). Thus, in vision, neighboring regions of visual cortex typically correspond to neighboring regions of the visual field while in audition, neighboring regions of auditory cortex correspond to similar frequencies of sound. Typically some other features (such as frequency rise vs. fall will be represented by the axis perpendicular to the frequency axis).

In the DIVA model, we can begin the overview with the Speech Sound Map (SSM, near the upper left region of Fig. 3) of an adult speaker, where each speech sound type (which might be a specific gesture, syllable or even phrase) is postulated to have a small area whose activation (and its inhibition of competing regions/gestures) sets off the pattern of neural events that give rise to the skilled production of the sound type. There are three output mappings from the Speech Sound Map. Going downward in the figure is the Feedforward Control Subsystem. Neural paths from the SSM to the motor cortex produce the gesture sequence that the speaker learned for producing this sound type. In addition, there are 2 sets of sensory expectations generated: what the sound of this speech gesture should be like in auditory terms and what sensory patterns in the vocal tract should be like. Thus, there is a mapping from the SSM to a somatosensory target region where a pattern of expectations is generated for a specific temporal sequence of sensations from the vocal tract (lips, tongue, palate, glottis, etc.) as the gesture is produced. (These expectations were learned during an earlier babbling stage and from previous speech production experience.) A similar set of predictions is produced about the expected sequence of auditory patterns. Then as the speech gestures are carried out (due to the Feedforward Control Subsystem), they produced actual auditory and somatosensory patterns which are compared to the expectations in the Auditory and Somatosensory Error Maps. If the inhibition from the realtime sensory feedback matches the pattern of excitations from the SSM, then the error maps register zero. If there are deviations (because something expected did not occur or if something unexpected occurs in either stream), then the error is detected and some form of motor correction is applied.

During the acquisition of speech skills, beginning with random babbling, there are many errors and the speaker learns gradually to produce better versions of feedforward gestural control, but during normal, skilled speech, there will be very few errors or deviations. It can be seen that this system gathers whatever rich sensory information it can, both for audition and for proprioception and pre-

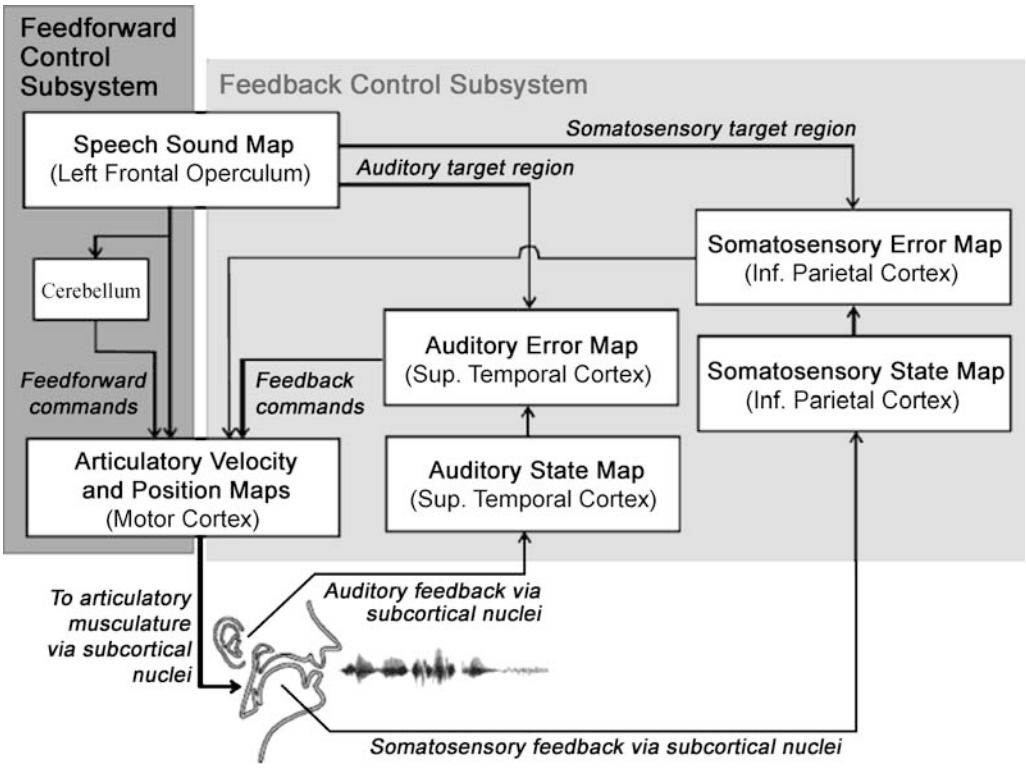

Dynamics of Language, Figure 3
Schematic diagram of the DIVA model of speech production (from [18])

dicts the sensory patterning in continuous time. As long as predictions are fulfilled, all proceeds as an approximately 'open loop' system. But clearly speech production is not simply a matter of issuing commands. There is much real-time control keeping the gesture on track at every instant.

Speech Perception Models

For audition of speech, the primary problem, rather than the control of many motor variables, it is recognition of the vast number of variants that can arise in ordinary speech. Audition skills begin early since hearing is possible in the womb. Infants at birth are already familiar with much about the auditory environment their mother lives in and even with her voice [26]. Infants gradually learn the auditory patterns of their language and the many variants that speech exhibits. Increasing perception skills lead to recognition of "chunks" of speech that are statistically predominant [16] so speech perception results in a kind of parsing of the speech signal into recognizable chunks of variable size.

Literacy has a dramatic influence on us. Somewhere between about 12 months and 5 years, children in many communities are taught the names of the letters and proceed over the next few years (often for many years) to become more and more literate and familiar with the conventions of the written language styles of the culture as well. But the written orthography has much that is unknown in the spoken language. The written language presumes discrete words, separated by spaces, enforces fixed conventional spellings, imposes sentence structure on texts, marks proper names and paragraphs, and can even support tabular layouts of text [20]. Literate people become very skilled at interpreting letter strings as speech and interpreting speech as letter sequences. These skills are the reason for our powerful intuitions that speech comes automatically and uniformly in letter-like form. But the description of language offered by alphabetical orthographies, despite its enormous influence on how linguists conceptualize their domain, plays almost no role whatever beyond its role in our conscious experience of language.

Future Directions

The discipline of linguistics has been severely hampered by the inability of linguists to escape from their intuitive understanding of language, an understanding that has been shaped by the literacy training of linguists. Because of our

lifelong experience using alphabetical representations of language, it has been nearly impossible to look beyond our intuitions to see what the experimental data show. The conceptualization of the problem of language presented here says that a language is a social product and not a cognitive one. Speakers cobble together methods that are suitable for talking and hearing within the conventions and customs of their linguistic culture, but this does not imply (as we thought earlier) that the apparent units of language (the sound types, lexical entries, etc.) are themselves cognitive units used and manipulated by speakers.

Simulation work on the social aspects of language, i. e., how a language comes into being in a community and how it evolves over time, is well under way. And the study of "sociolinguistics", studying language variation in its social context has recently become important. Similarly, there has been much work done on speech production and perception. But the view that intuitive linguistic units, like phones, phonemes, syllables, morphemes, words, etc., must comprise stages in the realtime processing of language has proven to be a red herring that has led us astray for at least a century. It is time to let this go and to focus on the two complex systems, one communal and one personal, that underlie human speech.

Bibliography

Primary Literature

1. Abraham R, Shaw C (1983) Dynamics: The Geometry of Behavior, Part 1. Aerial Press, Santa Cruz
2. Allen G (1972) The location of rhythmic stress beats in English: An experimental study I. Lang Speech 15:72–100
3. Bloomfield L (1933) Language. Holt Reinhart Winston, New York
4. Browman C, Goldstein L (1992) Articulatory phonology: An overview. Phonetica 49:155–180
5. Cangelosi A, Parisi D (2002) Simulating the Evolution of Language. Springer, New York
6. Cho T, Ladefoged P (1999) Variations and universals in VOT: Evidence from 18 languages. J Phonetics 27:207–229
7. Chomsky N (1965) Aspects of the Theory of Syntax. MIT Press, Cambridge
8. Chomsky N, Halle M (1968) The Sound Pattern of English. Harper and Row, New York
9. Clark A (1997) Being There: Putting Brain, Body, and World Together Again. Bradford Books/MIT Press, Cambridge
10. Clark A (2006) Language, embodiment and the cognitive niche. Trends Cogn Sci 10:370–374
11. Cummins F (2003) Rhythmic grouping in word lists: Competing roles of syllables, words and stress feet. In: Proceedings of the 15th International Conference on Spoken Language Processing, Barcelona, 2003, pp 325–328
12. Cummins F, Port R (1998) Rhythmic constraints on stress timing in English. J Phonetics 26:145–171
13. de Saussure F (1916) Course in General Linguistics. Philosophical Library, New York
14. Goldinger SD (1996) Words and voices: Episodic traces in spoken word identification and recognition memory. J Exp Psychol: Learn Mem Cogn 22:1166–1183
15. Grossberg S (1995) The neurodynamics of motion perception, recognition learning and spatial atttention. In: Port R, v Gelder T (eds) Mind as Motion: Explorations in the Dynamics of Cognition. MIT Press, Cambridge
16. Grossberg S (2003) The resonant dynamics of speech perception. J Phonetics 31:423–445
17. Guenther F (1995) Speech sound acquisition, coarticulation and rate effects in a neural network model of speech production. Psychol Rev 102:594–621
18. Guenther F, Perkell J (2004) A neural model of speech production and supporting experiments. In: From Sound to Sense: Fifty+ Years of Discoveries in Speech Communication. Cambridge, Massachusetts
19. Harris R (1981) The Language Myth. Duckworth, London
20. Harris R (2000) Rethinking Writing. Continuum, London
21. Hockett C (1968) The State of the Art. Mouton, The Hague
22. Huckvale M (1997) 10 things engineers have discovered about speech recognition. In: NATO ASI Speech Patterning Conference Jersey, 1997
23. IPA (1999) Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet. Cambridge University Press, Cambridge
24. Jakobson R, Fant G, Halle M (1952) Preliminaries to Speech Analysis: The Distinctive Features. MIT, Cambridge
25. Jelinek F (1988) Applying information theoretic methods: Evaluation of grammar quality. In: Workshop on Evaluation of Natural Language Processing Systems, 7-9 Dec 1988. Wayne
26. Jusczyk P (1997) The Discovery of Spoken Language. MIT Press, Cambridge
27. Kelso JAS (1995) Dynamic Patterns: Self-organization of Brain and Behavior. MIT Press, Cambridge
28. Kewley-Port D (1983) Time-varying features as correlates of place of articulation in stop consonants. J Acoust Soc Am 73:322–335
29. Kirby S (2001) Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. IEEE J Evol Comput 5:102–110
30. Kochanski G, Orphanidou C (2008) What marks the beat of speech? J Acoust Soc Am 123:2780–2791;
31. Kuhl P, Iverson P (1995) Linguistic experience and the perceptual magnet effect. In: Strange W (ed) Speech Perception and Linguistic Experience: Issues in Cross-Language Research. York Press, Timonium, pp 121–154
32. Ladefoged P, Maddieson I (1996) Sounds of the World's Languages. Blackwell, Oxford
33. Large E, Jones M (1999) The dynamics of attending: How we track time-varying events. Psychol Rev 106:119–159
34. Liberman AM, Harris KS, Hoffman H, Griffith B (1957) The discrimination of speech sounds within and across phoneme boundaries. J Exp Psychol 54:358–368
35. Liberman AM, Delattre P, Gerstman L, Cooper F (1968) Perception of the speech code. Psychol Rev 74:431–461
36. Lisker L, Abramson A (1964) A cross-language study of voicing in initial stops: acoustical measurements. Word 20:384–422
37. Love N (2004) Cognition and the language myth. Lang Sci 26:525–544
38. Mayville J, Jantzen K, Fuchs A, Steinberg F, Kelso JS (2002) Cor-

tical and subcortical networks underlying syncopated an synchronized coordination reveals using fMRI. Hum Brain Mapp 17:214–229
39. Palmeri TJ, Goldinger SD, Pisoni DB (1993) Episodic encoding of voice attributes and recognition memory for spoken words. J Exp Psychol, Learn Mem Cogn 19:309–328
40. Patel A (2003) Language, music, syntax and the brain. Nat Neurosci 6:674–681
41. Patel A, Iverson J, Chen Y, Repp B (2005) The influence of metricality and modality on synchronization with a beat. Exp Brain Res 163:226–238
42. Pisoni D, Levi S (2006) Some observations on representation and representational specificity in spoken word processing. In: Gaskell G (ed) Oxford Encyclopedia of Psycholinguistics. Oxford University Press, Oxford
43. Pisoni DB (1997) Some thoughts on 'normalization' in speech perception, in Talker variability. In: Johnson K, Mullennix J (eds) Speech processing. Academic Press, San Diego, pp 9–32
44. Port R (2003) Meter and speech. J Phonetics 31:599–611
45. Port R (2007) What are words made of?: Beyond phones and phonemes. New Ideas in Psychology 25:143–170
46. Port R (2008) All is prosody: Phones and phonemes are the ghosts of letters. Prosody 2008, Campinas
47. Port RF, Leary A (2005) Against formal phonology. Language 81:927–964
48. Rayner K, Foorman B, Perfetti C, Pesetsky D, Seidenberg M (2001) How psychological science informs the teaching of reading. Psychol Sci Public Interest 2:31–74
49. Scott S (1998) The point of P-centres. Psychol Res 61:4–11
50. Smith K, Brighton H, Kirby S (2003) Complex systems in language evolution: The cultural emergence of compositional structure. Adv Complex Syst 6:537–558
51. Steels L (2006) Experiments on the emergence of human communication. Trends Cogn Sci 10:347–434
52. Stevens K, Blumstein S (1978) Invariant cues for place of articulation in stop consonants. J Acoust Soc Am 64:1358–1368
53. Twaddell WF (1935) On defining the phoneme. Language, vol Language Monograph 16
54. van Gelder T, Port R (1995) Its about time. In: Port R, v Gelder T (eds) Mind as Motion: Explorations in the Dynamics of Cognition. MIT Press, Cambridge, pp 1–44
55. Werker J, Tees RC (1984) Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. Infant Behav Dev 7:49–63
56. Zanto T, Snyder J, Large E (2006) Neural correlates of rhythmic expectancy. Adv Cogn Psychol 2:221–231
57. Zawaydeh B, Tajima K, Kitahara M (2002) Discovering Arabic rhythm through a speech cycling task. In: Parkinson D, Benmamoun E (eds) Perspectives on Arabic Linguistics. John Benjamins, pp 39–58

Books and Reviews

Abler W (1989) On the particulate principle of self-diversifying systems. J Soc Biol Struct 12:1–13
Guenther F, Ghosh S, Tourville JA (2006) Neural modeling and imaging of the cortical interactions underlying syllable production. Brain Lang 96:280–301
Guenther FH, Gjaja M (1996) The perceptual magnet effect as an emergent property of neural map formation. J Acoust Soc Am 100:1111–1121
Haken H, Kelso JAS, Bunz H (1985) A theoretical model of phase transitions in human hand movements. Biol Cybern 51:347–356
Patel A, Lofquist A, Naito W (1999) The acoustics and kinematics of regularly timed speech: A database and method for the study of the P-center problem. In: 14th International Congress of Phonetic Sciences, San Francisco, 1-7 Aug
Smith E, Medin D (1981) Categories and Concepts. Harvard University Press, Cambridge
Tajima K, Port R (2003) Speech rhythm in English and Japanese. In: Local J, Ogden R, Temple R (eds) Phonetic Interpretation: Papers in Laboratory Phonology, vol 6. Cambridge University Press, Cambridge, pp 317–334

Dynamics of Parametric Excitation

Alan Champneys
Department of Engineering Mathematics, University of Bristol, Bristol, United Kingdom

Article Outline

Glossary
Definition of the Subject
Introduction
Linear Resonance or Nonlinear Instability?
Multibody Systems
Continuous Systems
Future Directions
Bibliography

Glossary

Parametric excitation Explicit time-dependent variation of a parameter of a dynamical system.

Parametric resonance An instability that is caused by a rational relationship between the frequency of parametric excitation and the natural frequency of free oscillation in the absence of the excitation. If ω is the excitation frequency, and ω_0 the natural frequency, then parametric resonance occurs when $\omega = (n/2)\omega_0$ for any positive integer n. The case $n = 1$ is usually the most prominent form of parametric resonance, and is sometimes called the principle subharmonic resonance.

Autoparametric resonance A virtual parametric resonance that occurs due to the coupling between two independent degrees of freedom within a system. The output of one degree of freedom acts like the parametric excitation of the other.

Ince–Strutt diagram A two-parameter bifurcation diagram indicating stable and unstable regions, specifically plotting the instability boundaries as the required

amplitude of parametric excitation against the square of the ratio of natural to excitation frequency.

Floquet theory The determination of the eigenvalue spectrum that governs the stability of periodic solutions to systems of ordinary differential equations.

Bifurcation A qualitative change in a system's dynamics as a parameter is varied. One-parameter bifurcation diagrams often depict invariant sets of the dynamics against a single parameter, indicating stability and any bifurcation points. Two-parameter bifurcation diagrams depict curves in a parameter plane on which one-parameter bifurcations occur.

Modal analysis The study of continuum models by decomposing their spatial parts into eigenmodes of the dynamic operator. The projection of the full system onto each mode gives an infinite system of differential equations in time, one for each mode.

Monodromy matrix The matrix used in Floquet theory, whose eigenvalues (also known as Floquet multipliers) determine stability of a periodic solution.

Definition of the Subject

Parametric excitation of a system differs from direct forcing in that fluctuations appear as temporal modulation of a parameter rather than as a direct additive term. A common paradigm is that of a pendulum hanging under gravity whose support is subjected to a vertical sinusoidal displacement. In the absence of any dissipative effects, instabilities occur to the trivial equilibrium whenever the natural frequency is a multiple of *half* the excitation frequency. At amplitude levels beyond the instability, further *bifurcations* (dynamical transitions) can lead to more complex quasi-periodic or chaotic dynamics. In multibody mechanical systems, one mode of vibration can effectively act as the parametric excitation of another mode through the presence of multiplicative nonlinearity. Thus *autoparametric* resonance occurs when one mode's frequency is a multiple of half of the other. Other effects include combination resonance, where the excitation is at a sum or difference of two modal frequencies. Parametric excitation is important in continuous systems too and can lead to counterintuitive effects such as stabilization "upside-down" of slender structures, complex wave patterns on a fluid free surface, stable pulses of light that overcome optical loss, and fluid-induced motion from parallel rather than transverse flow.

Introduction

Have you observed a child on a playground swing? In either a standing or a sitting position, generations of children have learned how to "pump" with their legs to set the swing in motion without making contact with the ground, and without external forcing. Note that pumping occurs at *twice* the natural frequency of swing of the pendulum, since the child extends her legs maximally at the two extremities of the swing's cycle, that is, two times per period. What is happening here? The simplest case to analyze is where the child is standing. Then, it has been argued by Curry [17] that the child plus swing is effectively a simple pendulum system, where the child's pumping has the effect of periodic variation of the position of the center of the mass along the arm of the pendulum (see Fig. 1a,b). We shall return to the self-propelled swing problem in Sect. "Future Directions" below, where we shall find out that all is not what it seems with this motivating example, but for the time being lets make the simple hypothesis that the child's pumping causes the center of mass to move up and down.

Using the standard model of a pendulum that consists of a lumped mass suspended by a rigid massless rod of length l from a pivot, the equations of motion for the angular displacement $\theta(t)$ of a pendulum hanging under gravity can be written in the form

$$\ddot{\theta}(t) + \frac{g}{l}\sin[\theta(t)] = 0 \,.$$

Here g is the acceleration due to gravity, a dot denotes differentiation with respect to time t and, for the time being, we have ignored any damping that must inevitably be present. Now, the moving center of mass means that the effective length of the rigid rod oscillates, and let us suppose for simplicity that this oscillation is sinusoidal about some mean length l_0:

$$l(t) = l_0[1 - \varepsilon\cos(\omega t)] \,.$$

Here $\varepsilon > 0$ assumed to be small and as yet we haven't specified the frequency ω. Then, upon Taylor expansion in ε, we find to leading order

$$\ddot{\theta} + \omega_0^2[1 + \varepsilon\cos(\omega t)]\sin\theta = 0 \,,$$

where $\omega_0 = \sqrt{(g/l_0)}$ is the natural frequency of small amplitude swings in the absence of the pump. Finally, upon rescaling time $\tilde{t} = \omega t$ and dropping the tildes we find

$$\ddot{\theta} + [\alpha + \beta\cos t]\sin\theta = 0 \,, \tag{1}$$

where

$$\alpha = \frac{\omega_0^2}{\omega^2} \quad \text{and} \quad \beta = \varepsilon\frac{\omega_0^2}{\omega^2} \tag{2}$$

are dimensionless parameters that respectively describe

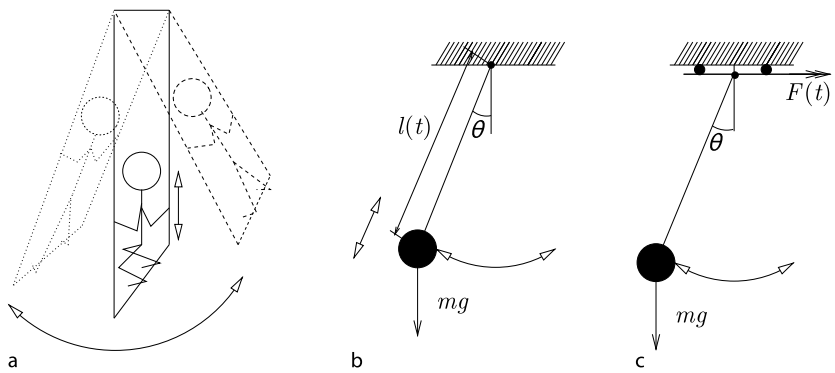

Dynamics of Parametric Excitation, Figure 1
a A child standing on a swing modeled as b a parametrically excited pendulum c a directly forced pendulum

the square of the ratio of natural frequency to excitation frequency, and the amplitude of the excitation.

Equation (1) is a canonical example of a *parametrically excited* system. Note that (1) also describes, at least to leading order, the case of a pendulum hanging under gravity that is excited by a vertical force of size $\varepsilon\omega_0^2 \cos t$ which we can think of as a periodic modulation of the gravity. This would be in contrast to a *directly forced* simple pendulum where the forcing would act in the direction of swing for small amplitude motion; see Fig. 1c. For the forced pendulum, the equivalent equations of motion would look more like

$$\ddot{\theta} + \alpha \sin\theta = \beta \cos t \,, \tag{3}$$

where the periodic forcing occurs as an additive input into the differential equation. In contrast, for (1) the periodic excitation occurs as a modulation of a *parameter* (which you can think of as gravity), hence the name 'parametric'. Another feature of (1) that makes it a good paradigm for this article is that it is nonlinear and, as we shall see, a full appreciation of the dynamics resulting from parametric excitation requires treatment of nonlinear terms.

An interesting contrast between parametrically excited and directly forced systems is in the nature of the trivial response to a non-zero input. In particular, for any α and β, Eq. (1) admits the trivial solution $\theta(t) \equiv 0$, whereas there is no such solution to (3) for non-zero β. The simplest "steady" solution to the directly forced system for small β is a small-amplitude periodic oscillation of period 2π (in our dimensionless time units). So any analysis of a parametrically excited system should start with an analysis of the steady state $\theta = 0$, which corresponds to the pendulum just hanging under gravity. To analyze stability of this equilibrium state let $\theta = 0 + x$ where x is small, then x satisfies linearization (1)

$$\ddot{x} + (\alpha + \beta \cos t)x = 0 \,. \tag{4}$$

The equilibrium position is stable if all solutions $x(t)$ of this equation, remain bounded as $t \to \infty$ and it is unstable if there are solutions $x(t) \to \infty$ as $t \to \infty$.

Equation (4) is known as the *Mathieu equation* [42], and is a specific example of Hill's equation [29,40]

$$\ddot{x} + [\alpha + V(t)]x = 0 \,, \tag{5}$$

where $V(t)$ is any periodic function with period 2π. Hill's equation can be solved using classical methods for linear time-dependent ordinary differential equations (ODEs), see e.g. [31,33]. In particular, by writing $y_1 = x$ and $y_2 = \dot{x}$, Hill's equation (5) is just a two-dimensional example of a linear first-order time-periodic equation

$$\dot{\mathbf{y}} = P(t)\mathbf{y} \,, \qquad \mathbf{y} \in \mathbb{R}^n \,. \tag{6}$$

The general solutions to (6) for any initial condition can be expressed in terms of a fundamental matrix $\Phi(t)$, with $\Phi(0) = I_n$ (the identity matrix of dimension n):

$$\mathbf{y}(t) = \Phi(t)\mathbf{y}(0) \,.$$

Note that just because $P(t)$ has period 2π it does not follow that $\Phi(t)$ does. Also $\Phi(t)$ cannot be computed exactly, except in highly specialized cases, and one has to rely on approximate methods as described in the next Sect. "Floquet Analysis" and in the companion article to this one ▶ Perturbation Analysis of Parametric Resonance. However, in order to study stability we don't actually need to construct the full solution $\Phi(t)$, it is sufficient to consider $M = \Phi(2\pi)$, which is called the *monodromy matrix* associated with Hill's equation. Stability is determined by studying the eigenvalues of M: solutions to (5) with small initial conditions remain small for all time if all eigenvalues of M lie on or within the unit circle in the complex plane; whereas a general initial condition will lead to un-

bounded motion if there are eigenvalues outside this circle. See Fig. 2a. Eigenvalues of the Monodromy matrix are also known as *Floquet multipliers* of the system, and the general study of the stability of periodic systems is called *Floquet theory*.

In the case of the Mathieu equation, where $V(t) = \beta\cos(t)$, we have

$$P(t) = \begin{bmatrix} 0 & 1 \\ 0 & \alpha + \beta\cos(t) \end{bmatrix}. \tag{7}$$

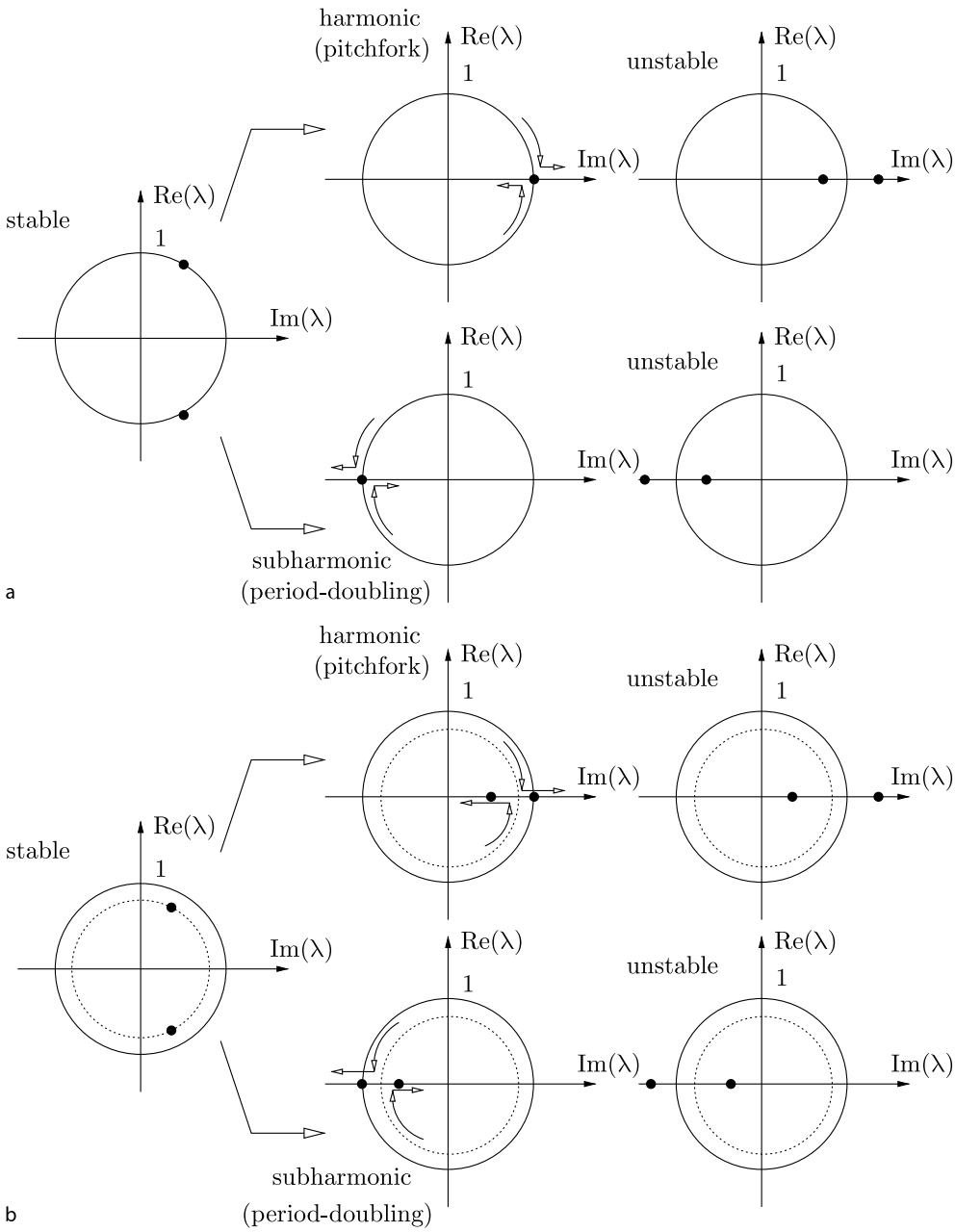

Dynamics of Parametric Excitation, Figure 2
Possible behavior of Floquet multipliers λ (eigenvalues of the monodromy matrix M) for a a conservative Hill equation and b in the presence of small damping

Figure 3 depicts graphically the behavior of Floquet multipliers for this case. Here, in what is known as an *Ince–Strutt diagram* [31,63], the shaded regions represent the values of parameters at which the origin $x = 0$ is stable (with Floquet multipliers on the unit circle) and the white regions are where the origin is unstable (at least one multiplier outside the unit circle). We describe how to calculate such a diagram in the next section. For the time being we note that there are *resonance tongues* that emanate from $\beta = 0$ whenever α is the square of a half-integer: $\alpha = 1/4, 1, 9/4, 4, 25/4$, etc. The width (in α) of the tongue for small β turns out to be proportional to β to the power of twice the half-integer that is being squared. That is, the $\alpha = 1/4$ tongue has width $O(\beta)$, the $\alpha = 1$ tongue has width $O(\beta)^2$, the $\alpha = 9/4$ tongue has width $O(\beta^3)$ etc.

On the tongue boundaries (between the shaded and the unshaded regions of Fig. 3), the Mathieu equation admits non-trivial periodic solutions. For the tongues that are rooted at square integers, $\alpha = 1, 9, 16$, etc., these solutions have period 2π (the *same* period as that of the excitation) and the instability that occurs upon crossing such a tongue boundary is referred to as a *harmonic instability*. On the other boundaries, those of tongues rooted at $\alpha = 1/4, 9/4, 25/4$ etc., the periodic solution has period 4π (*twice* that of the excitation). Hence such instabilities are characterized by frequencies that are *half* that of the excitation and are hence referred to as *subharmonic instabilities*.

So far, we have ignored any dissipative effects, so that Eq. (5) can be regarded as a conservative or *Hamiltonian* system. Damping can be introduced via a dissipative force that is proportional to velocity, hence (4) becomes

$$\ddot{x} + \delta\dot{x} + (\alpha + \beta\cos t)x = 0\,, \tag{8}$$

where δ is a dimensionless damping coefficient. One can also apply Floquet theory to this damped Mathieu equation; see Fig. 2b. The results for Eq. (8) are plotted in Fig. 4, where we can see the intuitive effect that damping increases the areas of stability. In particular, in comparison with Fig. 3, note that the resonance tongues have been lifted from the α-axis, with the amount of the raise being inversely proportional to the width of the undamped tongue. Thus, the resonance tongue corresponding to

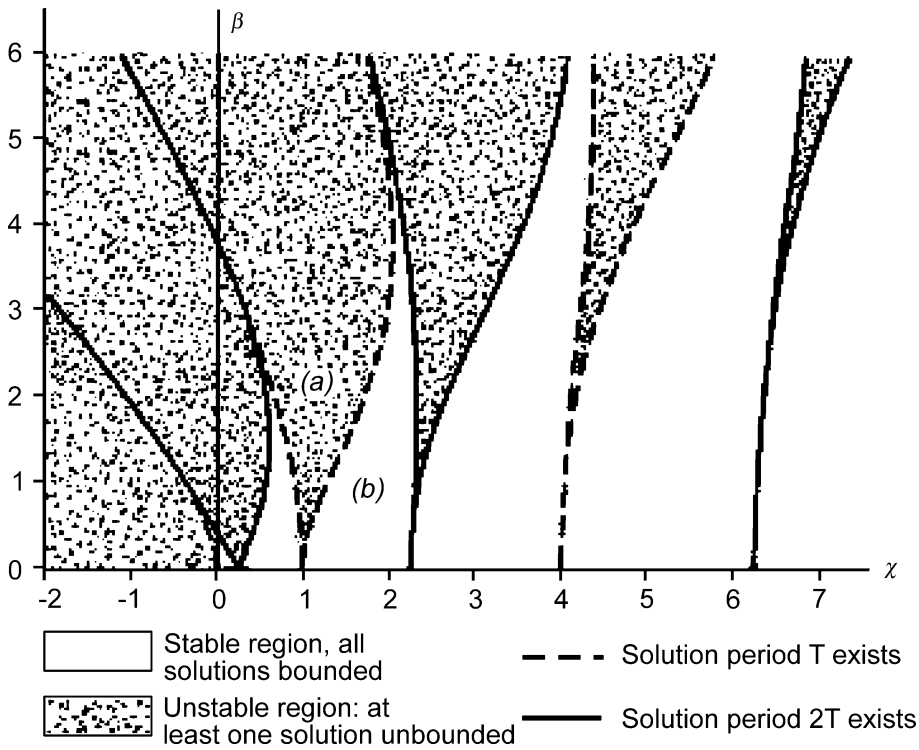

Dynamics of Parametric Excitation, Figure 3

Ince–Strutt diagram showing resonance tongues of the Mathieu equation (4). Regions of instability of the trivial solution $x = 0$ are *shaded* and a distinction is drawn between instability curves corresponding to a period-T ($T = 2\pi$) and to a period-$2T$ orbit. After (2nd edn. in [33]), reproduced with permission from Oxford University Press

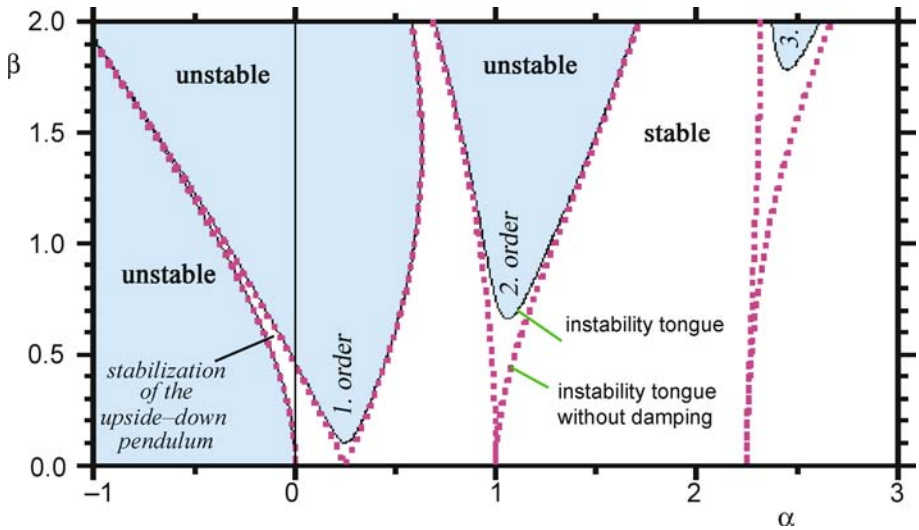

Dynamics of Parametric Excitation, Figure 4
Similar to Fig. 3 but for the damped Mathieu equation (8) with $\delta = 0.1$. After http://monet.unibas.ch/~elmer/pendulum/parres.htm, reproduced with permission

$\alpha = 1/4$ is easily the most prominent. Not only for fixed forcing amplitude β does this tongue occupy for the greatest interval of square frequency ratio α, but it is accessible with the least forcing amplitude β. Thus, practicing engineers often think of this particular *subharmonic instability* close to $\alpha = 1/4$ (i. e. where the forcing frequency is approximately half the natural frequency) as being *the* hallmark of parametric resonance. Such an instability is sometimes called the *principal* parametric resonance. Nevertheless, we shall see that other instability tongues, in particular the *fundamental* or *first harmonic* parametric resonance near to $\alpha = 1$ can also sometimes be of significance.

Returning briefly to the problem of how children swing, it would seem that the choice $\alpha \approx 1/4$ is in some sense preferable. This is the largest instability tongue for small β. That is, the smallest effort is required in order to cause an exponential instability of the downwards hanging solution, and set the swing in motion. And, according to (2), $\alpha = 1/4$ corresponds to $\omega = 2\omega_0$. Thus generations of children seem to have hit upon an optimal strategy [48], by pumping their legs at precisely twice the frequency of the swing.

Finally, looking at Fig. 4, note the curious feature that we have continued the diagrams into negative α, despite the fact that α was defined as a square term. In the case of a pendulum though, $\alpha = l/(g\omega^2)$, and so we can think of negative α as representing negative g. That is, a pendulum with negative α is defying gravity by 'hanging upwards'. Clearly without any excitation, such a situation would be violently unstable. The big surprise then is that there is a thin wedge of stability (which is present either with or without damping) for $\alpha < 0$. Note though that the wedge is thickest for small β and small $|\alpha|$, which limit corresponds to small amplitude, high-frequency oscillation. This remarkable stabilization of an upside-down pendulum by tiny, rapid vertical vibration was first described by Stephenson [53] in 1908, re-discovered by Kapitza [34] in 1951 (and hence often called the Kapitza pendulum problem) and has been experimentally verified many times; see e. g. [3]. We shall return to this upside down stability in Sect. "Upside-Down Stability".

The rest of this article is outlined as follows. In Sect. "Linear Resonance or Nonlinear Instability?" we shall examine how the Ince–Strutt diagram is constructed and see that this is just an example of the linearized analysis around any periodic state of a nonlinear system. Further quasi-periodic or chaotic motion can ensue as we push beyond this initial instability. Section "Multibody Systems" concerns mechanical systems with multiple degrees of freedom. We consider *autoparametric resonance* where the motion of one degree of freedom acts as the parametric excitation of another. We also look at conditions in multibody systems where a combination of two natural frequencies can be excited parametrically, focusing in particular on the somewhat illusive phenomenon of *difference* combination resonances. Section "Continuous Systems" then looks at parametric excitation in continuous systems, looking at phenomena as diverse as pattern formation on the surface of a liquid, flow-induced oscillation of pipes, a clever way of overcoming loss in opti-

cal fibers and the oscillation-induced stiffening of structures. Finally, Sect. "Future Directions" draws conclusions by looking at future and emerging ideas.

Linear Resonance or Nonlinear Instability?

Let us now consider a more detailed analysis of parametric resonance of the Mathieu equation and of the implications for a nonlinear system, using the parametrically excited pendulum as a canonical example.

Floquet Analysis

We begin by giving a reasoning as to why the Ince–Strutt diagram Fig. 3 for the Mathieu equation (ignoring damping for the time being) looks the way it does, by making a more careful consideration of how Floquet multipliers behave as parameters are varied. Essentially, we explain the behavior of Fig. 2a. Our treatment follows the elementary account in [33], to which the interested reader is referred for more details. For a general linear system of the form (6), a lot of insight can be gained by considering the so-called *Wronskian* $W(t) = \det(\Phi(t))$. From the elementary theory of differential equations (e. g. [33]) we have

$$W(t) = W(t_0) \exp\left(\int_{t_0}^{t} \operatorname{trace}[P(s)]\right) \mathrm{d}s \,, \tag{9}$$

where the *trace* of a matrix is the sum of its diagonal elements. Note that the specific form of $P(t)$ for the Mathieu equation (7) has $\operatorname{trace}[P(t)] = 0$, hence, from (9) that that $\det(M) = \det[W(2\pi)] = 1$. But from elementary algebra we know that the trace of the matrix is the product of its eigenvalues. So we know that the two Floquet multipliers $\lambda_{1,2}$ satisfy $\lambda_1\lambda_2 = 1$ and must therefore solve a characteristic equation of the form

$$\lambda^2 - 2\phi(\alpha, \beta)\lambda + 1 = 0 \,,$$

for some unknown real function ϕ. There are two generic cases $|\phi| < 1$ and $|\phi| > 1$. If $\phi > 1$, then the Floquet multipliers are complex and lie on the unit circle. This corresponds to stability of the zero-solution of the Mathieu equation. Conversely if $|\phi| < 1$, then the Floquet multipliers are real and satisfy $|\lambda_1| < 1$ and $|\lambda_2| = 1/|\lambda_1| > 1$. The larger multiplier represents an exponentially growing solution and hence corresponds to an instability. The two boundary cases are: $\phi = 1$, in which case we have a double Floquet multiplier at $\lambda = 1$; or $\phi = -1$, which corresponds to a double multiplier at $\lambda = -1$. These represent respectively the harmonic and the subharmonic stability boundaries in the Ince–Strutt diagram.

But, how do we compute ϕ? The answer is that we don't. We simply recognize that $\phi = 1$ corresponds to the existence of a pure 2π-periodic solution of the fundamental matrix $M = \Phi(2\pi)$, and that $\phi = -1$ represents a periodic solution of $\Phi(2\pi)^2 = \Phi(4\pi)$ and hence a 4π-periodic solution of the fundamental matrix. To look for a 2π-periodic solution of the Mathieu equation, it is convenient to use Fourier series. That is, we seek a solution (4) in the form

$$x(t) = \sum_{n=-\infty}^{n=+\infty} c_n \mathrm{e}^{int} \,.$$

Substitution into (4), using $\cos(t) = (1/2)(\mathrm{e}^{it} + \mathrm{e}^{-it})$ leads to infinitely many equations of the form

$$\begin{aligned} &(1/2)\beta c_{n+1} + (\alpha - n^2)c_n + (1/2)\beta c_{n-1} = 0 \,, \\ &\qquad n = -\infty, \ldots, \infty \,, \quad \text{with } c_{-n} = c_n^* \,, \end{aligned} \tag{10}$$

where an asterisk represents complex conjugation. Note that the final condition implies that x is real. Clearly, for $\beta = 0$, a nontrivial solution can be found for any integer p which has $c_n = 0$ for all $n \neq p$, provided $\alpha^2 = p^2$. That is, solution curves bifurcate from $\beta = 0$ for each α-value that is the square of a whole integer. Furthermore, nonlinear analysis can be used to show that for small positive β, there are *two* solution branches that emerge from each such point (provided $p \neq 0$); one corresponding to c_p being real, which corresponds to motion $x(t) \sim \cos(pt)$ in phase with the excitation; and another corresponding to c_p imaginary, which corresponds to motion $x(t) \sim \sin(pt)$ out of phase with the excitation. Similarly, if we look for solutions which are 4π-periodic by making the substitution

$$x(t) = \sum_{m=-\infty}^{m=+\infty} d_m \mathrm{e}^{imt/2} \,, \tag{11}$$

we arrive at an infinite system of equations

$$\begin{aligned} &(1/2)\beta d_{m+2} + (\alpha - (1/4)m^2)d_m + (1/2)\beta d_{m-1} = 0 \,, \\ &\qquad m = -\infty, \ldots, \infty \,, \quad \text{with } d_{-m} = d_m^* \,. \end{aligned} \tag{12}$$

Note that the equations for m even and m odd in (12) are decoupled. The equations for even m become, on setting $m = 2n$, precisely the system (10) with $d_{2m} = c_n$. For $m = 2n + 1$, however, we arrive at new points of bifurcation, by exactly the same logic as we used to analyze (10). Namely whenever $\alpha^2 = (p/2)^2$ for any odd integer p we have the bifurcation from $\beta = 0$ of two 4π-periodic solutions (one like $\cos(pt/2)$ and one like $\sin(pt/2)$).

Perturbation analysis can be used to obtain asymptotic expressions for the transition curves that bifurcate from

$\beta = 0$ at these special α-values; see the companion article ▶ Perturbation Analysis of Parametric Resonance for the details. In particular it can be shown that the width of the stability tongue (separation in α-values from the sine and the cosine instability curves) originating at $\alpha = (p/2)^2$ (for p being odd or even) scales like

$$\text{tongue width}_p \sim \beta^p + \text{ h.o.t} .$$

Hence the principal subharmonic instability arising from $\alpha = 1/4$ has width $O(\beta)$ (actually the transition curves are given by $\alpha = 1/4 \pm (1/2)\beta + O(\beta^2)$), and the fundamental harmonic instability arising from $\alpha = 1$ has width $O(\beta^2)$, etc. Hence the principal subharmonic instability is by far the most prevalent as it occupies the greatest frequency width for small β, and is often thought to be the hallmark of parametric resonance in applications.

Note that we have not as yet mentioned the transition curve emerging from $\alpha = \beta = 0$, which corresponds to $n = 0$ in the above analysis. This case can also be analyzed asymptotically, and the single transition curve can be shown to scale like $\alpha = -(1/2)\beta^2 + O(\beta^3)$ for small β. It is this bending back into $\alpha < 0$ that leads to the Kapitza pendulum effect, which we shall return to in Sect. "Upside-Down Stability" below.

Finally, the addition of small damping can be shown to narrow the tongues and lift them off from the $\beta = 0$ axis as shown in Fig. 4. The amount of lift is in proportion to the width of the tongue. This property further enhances the pre-eminence of the principal subharmonic instability for parametrically excited systems in practice, as they thus require the least amplitude of vibration in order to overcome the damping forces.

Bifurcation Theory

We have just analyzed the Mathieu equation, which is purely linear. But, for example as shown in the Introduction, such an equation is often derived in the course of analyzing the stability of a trivial solution to a nonlinear parametrically excited system. It is worthwhile therefore to consider the nonlinear implications of the harmonic and subharmonic instabilities. The time-periodic Mathieu equation can be written as an example of a three-dimensional autonomous system of ODEs

$$\dot{x} = y , \quad \dot{y} = f(x, y, s) , \quad \dot{s} = 1 \,(\text{mod}\, 2\pi) ,$$

for $(x, y, s) \in \mathbb{R}^2 \times S^1$, where $f(x, y, s) = -\delta y - (\alpha + \beta \cos(s))x$. As such, the trivial solution $x = 0$ should actually be seen as the periodic solution $x = y = 0$, $s = t$ (mod 2π). Hence we can understand instability of the trivial solution as points of bifurcation of periodic solutions, for which there is a mature theory, see for example [27,37,65,67]. In particular, the subharmonic instability with Floquet multiplier -1 is a *period doubling bifurcation*, and the harmonic bifurcation corresponding to multiplier $+1$ is a *pitchfork* bifurcation. Note that both these bifurcations come in super- and sub-critical forms. Hence, depending on the nature of the nonlinearity we may find either stable bounded motion close to the instability inside the resonant tongue close to the boundary edge, or motion that is not of small amplitude. See Fig. 5.

The case of the Mathieu or Hill equations without damping have received special attention in the literature. This is because of the extra mathematical structure they posses, namely that they can be expressed as *reversible* and *Hamiltonian* systems. Specifically, Hill's equation of the form (5) can be shown to conserve the energy function

$$H(x, y, t) = (1/2)y^2 + (1/2)(\alpha + V(t))x^2 ,$$

$y = \dot{x}$, and to be expressible in Hamiltonian form

$$\dot{x} = \frac{\partial H}{\partial y} , \qquad \dot{y} = -\frac{\partial H}{\partial x} .$$

Reversibility of Hill's equation is expressed in the fact that the model is invariant under the transformation $t \to -t$, $y \to -y$. Broer and Levi [7] used Hamiltonian and reversible theory to look at Hill's equations for general periodic functions $V(t)$. They showed that under certain conditions, the resonance tongues may become of zero width for nonzero β. Effectively, the left and right-hand boundaries of the tongue pass through each other. This is a generic phenomenon, in that a small perturbation within the class of Hill's equations will lead not destroy such an intersection. However, the Mathieu equation for which $V(t) = \cos(t)$ does not posses this property.

Many nonlinear models that undergo parametric resonance also preserve the properties of being Hamiltonian and reversible. Special tools from reversible or Hamiltonian systems theory can sometimes be used to analyze the bifurcations that occur under parametric resonance in nonlinear systems, see for example [8,9]. Points outside the resonance tongues correspond to where the trivial solution is an *elliptic* (weakly stable) fixed point of the time-2π map, and those inside the resonance tongue to where it is a *hyperbolic* (unstable) fixed point. One therefore does not generally see asymptotically stable non-trivial solutions being born at a resonance tongue boundary, rather one sees elliptic nontrivial dynamics. In fact, for a general periodic solution in a Hamiltonian system, pitchfork and period-doubling bifurcations are not the only kind of bifurcations that produce non-trivial elliptic periodic or-

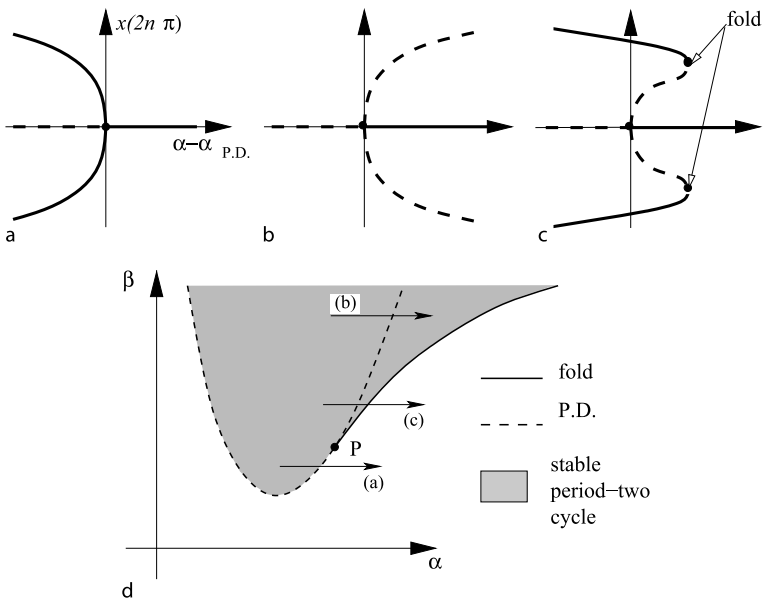

Dynamics of Parametric Excitation, Figure 5

Contrasting a super-critical and b sub-critical bifurcations close to the resonance tongue boundary of a subharmonic (period-doubling) instability, here represented by the abbreviation P.D. *Solid lines* represent stable solutions, *dashed lines* represent unstable solutions. c A one-parameter bifurcation diagram close to the super/sub-critical boundary. d A two parameter unfolding of the codimension-two point *P* where the super/sub-critical transition occurs. *Shading* is used to represent where a stable period-two cycle exists. Note in this case that for higher β-values the right-hand boundary of this existence region is therefore the fold curve and not the period-doubling point. Between the period-doubling and the fold there is hysteresis (coexistence) between the stable trivial solution and the stable period-two cycle

bits. If the Floquet multipliers pass through any nth root of unity, there is a *period multiplying* (subharmonic) bifurcation at which a periodic orbit of period $2n\pi$ is born (see, e. g. [62]). A period-doubling is just the case $n = 2$. Such bifurcations can also occur for the non-trivial elliptic periodic orbits within a resonance tongue, but they are all destroyed by the presence of even a small amount of damping.

Beyond the Initial Instability

For a nonlinear system, the pitchfork or period-doubling bifurcation that occurs upon crossing a resonance tongue might just be the first in a sequence of bifurcations that is encountered as a parameter is varied. As a general rule, the more β is increased for fixed α, the more irregular the stable dynamics becomes. For example, the period-doubling bifurcation that is encountered upon increasing β for α close to 1/4 is typically just the first in a sequence of period-doublings, that accumulate in the creation of a chaotic attractor, obeying the famous Feigenbaum scaling of this route to chaos (see e. g. [18]).

Van Noort [64] produced a comprehensive numerical and analytical study into the nonlinear dynamics of the parametrically excited pendulum (1) without damping. Of interest in that study was to complete the nonlinear dynamics of one of the simplest non-integrable Hamiltonian and reversible dynamical systems. The parameter β acts like a perturbation to integrability. Certain structures that can be argued to exist in the integrable system, like closed curves of quasiperiodic motion in a Poincaré section around an elliptic fixed point, serve as approximate organizers of the dynamics for non-zero β. However, KAM theory (see e. g. [55]) shows that these closed curves become invariant tori that typically breakup as β is increased. At the heart of the regions of chaotic dynamics lie the resonance tongues associated with a period-multiplying bifurcation. Inside the resonance tongues are nontrivial fixed points, whose surrounding invariant tori break up into heteroclinic chains, which necessarily imply chaos.

Acheson [2] studied the dynamics of the parametrically forced pendulum in the presence of damping (i. e. Eq. (1) with an additional term $\Delta\dot{x}$ on the right-hand side) via computer simulation. He found parameter regions inside the principal resonance tongue of the pendulum where it undergoes what he referred to as *multiple nodding* oscillations. The subharmonic motion inside the tongue should be such that during one cycle of the excita-

tion, the pendulum swings to the left and back, and during the next cycle it swings to the right and back. That is, one complete oscillation per two cycles. Multiple nodding motion by contrast is asymmetric and involves two swings to the left, say, (one swing per cycle of the excitation), followed by one swing to the right, with the motion repeating not every two but every three cycles. He also found similar four-cycle and five-cycle asymmetric motion. This behavior can be explained as pockets of stable behavior that survives from the period-multiplying bifurcations that are only present in the case of zero damping [21].

Clearly, beyond the initial instability the universality of the Ince–Strutt diagram for the Mathieu equation, or for any other Hill equation for that matter, disappears and the precise features of the nonlinear dynamics depends crucially on the details of the particular nonlinear system being investigated.

Multibody Systems

Of course, there is little in the real world that can be accurately described by deterministic, parametrically excited single degree-of-freedom ODEs. In this section we shall consider slightly more realistic situations that are governed by higher-dimensional systems, albeit still deterministic, and still described by finitely many dynamic variables. For simplicity, we shall restrict attention to rigid-body mechanisms that can be described by a system of (finitely many) coupled modes of vibration. Furthermore, we shall adopt a Lagrangian framework where the dynamics of mode i can be represented by a generalized coordinate q_i and its associated generalized velocity $\dot{q}_i$, such that we arrive at a system of (generally nonlinear) coupled second-order ODEs; one equation for each acceleration $\ddot{q}_i$.

Autoparametric Resonance

Autoparametric resonance occurs when there is asymmetric nonlinear multiplicative coupling between modes of a multibody system. Specifically, suppose that a mode $(q_1, \dot{q}_1)$ is coupled to a mode $(q_2, \dot{q}_2)$ through a term proportional to $q_1 q_2$ in the equation of motion for the 2nd mode (i. e. the $\ddot{q}_2$ equation), yet there is no corresponding $q_2 q_1$ term in the $\ddot{q}_1$ equation. For example, consider

$$\ddot{q}_1 + \delta_1 \dot{q}_1 + \alpha_1 q_1 = f_1(q_1, t) \tag{13}$$

$$\ddot{q}_2 + \delta_2 \dot{q}_2 + \alpha_2 q_2 + \beta q_1 q_2 = f_2(q_2)\,, \tag{14}$$

where f_1 and f_2 are nonlinear functions of their arguments and the dependence of f_1 on t is periodic. Furthermore suppose that mode 1 undergoes a periodic motion (either as a result of direct forcing, its own parametric resonance, or self-excitation). For example, if $f_2 = 0$ and f_1 were simply $\cos(\omega t)$ for some ω^2 sufficiently detuned from α_1, then the solution $q_1(t)$ to (13) would, after allowing sufficient time for transients to decay, be simply proportional to $\cos(\omega t + \phi)$ for some phase ϕ. Then, the $q_1 q_2$ term in (14) would be effectively act like $\beta \cos(\omega t + \phi) q_2$ and so (14) would be precisely the Mathieu equation. Values of $\omega^2 = 4\alpha_2/n$ for any positive integer n correspond to the root points of the resonance tongues of the Ince–Strutt diagram. Such points are referred to as *internal* or *autoparametric* resonances.

A canonical example of an autoparametrically resonant system is the simple spring-mass-pendulum device depicted in Fig. 6, which we shall use to discuss the some key generic features of autoparametric resonance; see the book by Tondl et al. [58] for more details. In the figure, a mass M is mounted on a linear spring with stiffness k, is forced by an applied displacement $P\cos(\omega\tau)$ and is attached to the pivot of a pendulum of length l and lumped mass m that hangs under gravity. The motion of the mass and the pendulum are assumed to be subject to proportional viscous damping with coefficients b and c respectively. Using standard techniques, the equations of motion can of such a device can be written in the dimensionless form

$$\ddot{x} + \delta_1 \dot{x} + \alpha_1 + \mu(\ddot{\phi}\sin(\phi) + \dot{\phi}^2 \cos\phi) = A\cos t \tag{15}$$

$$\ddot{\phi} + \delta_2 \dot{\phi} + \alpha_2 \sin\phi + \ddot{x}\sin\phi = 0\,, \tag{16}$$

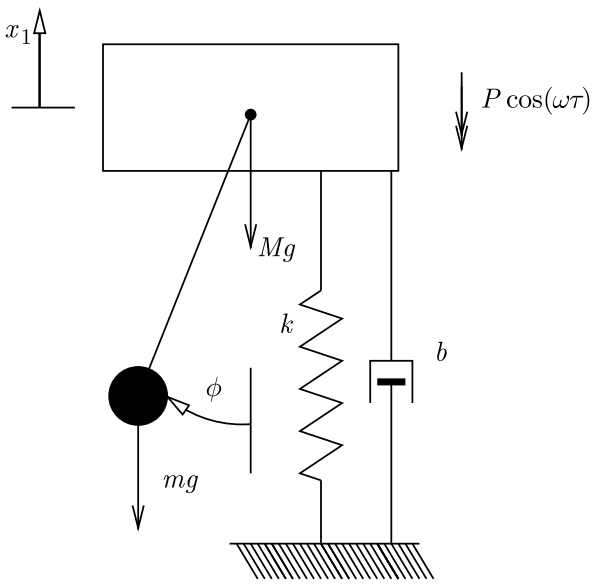

Dynamics of Parametric Excitation, Figure 6
A mass-spring-pendulum system

where $x = y/l$, time t has been scaled so that $t = \omega\tau$, and the dimensionless parameters are

$$\delta_1 = \frac{b}{\omega(M+m)}\,, \quad \alpha_1 = \frac{k}{\omega^2(M+m)}\,, \quad \mu = \frac{m}{M+m}\,,$$
$$A = \frac{P}{l\omega^2(M+m)}\,, \quad \delta_2 = \frac{c}{\omega m l^2}\,, \quad \alpha_2 = \frac{g}{l\omega^2}\,.$$

The simplest form of output of such a model is the so-called *semi-trivial* solution in which the pendulum is at rest, $\dot\phi = \phi \equiv 0$, and the mass-spring system alone satisfies the simple equation

$$\ddot x + \delta_1 \dot x + \kappa_1 x = A\cos t\,.$$

After transients have died out, this single equation has the attracting solution

$$x(t) = R_0\cos(t+\psi_0)\,,$$
$$\text{where}\quad R_0 = \frac{A}{\sqrt{\delta_1^2 + (\alpha_1 - 1)^2}}$$
$$\text{and}\quad \tan\psi_0 = \frac{\delta_1}{\alpha_1 - 1}\,. \tag{17}$$

Note that the solution (17) exists for all values of the physical parameters, in particular for all positive α_1 provided $\delta_1 > 0$.

Substitution of the semi-trivial solution (17) into the pendulum Eq. (16), leads to

$$\ddot\phi + \delta_2\dot\phi + [\alpha_2 + R_0\cos(t+\psi_0)]\sin\phi = 0\,,$$

which, after a time shift $t \to t - \psi_0$ is precisely the Eq. (1) for a parametrically excited pendulum in the presence of damping. Thus, the Ince–Strutt diagram of the damped Mathieu equation, with $\beta = A/(\sqrt{\delta_1^2 + (\alpha_1 - 1)^2})$, $\delta = \delta_2$ and $\alpha = \alpha_2$ tell us precisely the parameter values for which the semi-trivial solution is stable. Inside the resonance tongues we have an instability of the semi-trivial solution, which leads to stable non-trivial motion with the pendulum swinging. To determine precisely what happens away from the boundaries of the stability tongue, one again has to perform fully nonlinear analysis, as outlined in Sect. "Linear Resonance or Nonlinear Instability?". Note now though that for nonzero pendulum motion, $\phi \not\equiv 0$, the Eqs. (15) and (16) become a fully coupled two-degree-of-freedom system, whose dynamics can be even more complex than that of the one-degree-of-freedom parametrically excited pendulum. Indeed in [58], a further destabilizing secondary-Hopf (Neimark–Sacker) bifurcation is identified within the principal ($\alpha_2 = 1/4$) subharmonic resonance tongue, and numerical evidence is found for several different kinds of quasiperiodic motion.

There is one difference though between autoparametric resonance and pure parametric resonance. The paradigm for parametric resonance is that the external parametric excitation is held constant (its amplitude and frequency are *parameters* of the system). Here, however, for nonzero ϕ, the system is fully coupled, so that the parametric term (17) can no longer be thought of as being independent of ϕ, effectively the excitation becomes a *dynamic variable* itself. Thus, motion of the pendulum can affect the motion of the mass-spring component of the system. In particular if there is an autoparametric instability where the pendulum is set in motion, then energy must be transferred from the mass-spring component to the pendulum, a phenomenon known as *quenching*. But the pendulum is only set in motion close to certain resonant frequency ratios α_2. Thus, we have a *tuned damper* that is designed to take energy out of the motion of the mass at certain input frequencies ω frequencies.

To see how such a damper might work in practice, suppose we build a civil engineering structure such as a bridge or a building, that has a natural frequency ω_1 in some mode, whose amplitude we call $x(t)$. Suppose the system is subjected to an external dynamic loading (perhaps from traffic, wind or earthquakes) that is rich in a frequency ω which is close to ω_1. Then, if the damping is small (as is typical in large structures) using (15) with $\mu = 0$, we find that the response given by (17) is large since α_1 is close to 1 and δ_1 is small. This is the fundamental *resonant response* of the structure close to its natural frequency. Now suppose we add the pendulum as a tuned damper. In particular we design the pendulum so that $\alpha_2 \approx \alpha_1$. The pendulum now sees a large amplitude parametric excitation within its main subharmonic resonance tongue. It becomes violently unstable, thus sucking energy from the structure.

To see just how effective this quenching effect can be, consider the following asymptotic calculation [58]. Suppose the external forcing is very small, $A = \varepsilon^2\hat A$, for some small parameter ε, and that the damping and frequency detunings scale with ε: $\delta_1 = \varepsilon\hat\delta_1$, $\delta_2 = \varepsilon\hat\delta_2$, $\alpha_1 = 1 + \varepsilon\hat\sigma_1$, $\alpha_2 = (1/4) + \varepsilon\hat\sigma_2$, where the hatted quantities are all assumed to be $O(1)$ as $\varepsilon \to 0$. Upon making these substitutions, and dropping all hats, Eqs. (15) and (16) become

$$\ddot x + x + \varepsilon[\delta_1\dot x + \sigma_1 - (1/4)\mu\phi^2 + \mu\dot\phi^2] = \varepsilon A\cos(t) + O(\varepsilon^2)\,, \tag{18}$$

$$\ddot\phi + \frac{1}{4}\phi + \varepsilon[\delta_2\dot\phi + (1/2)\sigma_2\phi - x\phi] = O(\varepsilon^2)\,. \tag{19}$$

We then look for solutions of the form

$$x = R_1(t)\cos[t+\psi_1(t)], \quad \phi = R_2(t)\cos[(t/2)+\psi_2(t)].$$

Making this substitution into (18), (19), and keeping only $O(\varepsilon)$ terms we find

$$\begin{aligned} 2\dot{R}_1 &= \varepsilon[-\delta_1 R_1 - (1/4)R_2^2\sin(\psi_1 - 2\psi_2) - A\sin(\psi_1)]\,, \\ 2R_1\dot{\psi}_1 &= \varepsilon[\sigma_1 R_1 - (1/4)\mu R_2^2\cos(\psi_1 - 2\psi_2) - A\cos\psi_1]\,, \\ 2\dot{R}_2 &= \varepsilon[-\delta_2 R_2 + R_1 R_2\sin(\psi_1 - 2\psi_2)]\,, \\ 2R_2\dot{\psi}_2 &= \varepsilon[\sigma_2 R_2 - R_1 R_2\cos(\psi_1 - 2\psi_2)]\,. \end{aligned}$$

Equilibrium solutions of these equations correspond to 4π-periodic solutions that are valid close to the instability tongue. Upon setting the left-hand sides to zero, a straightforward combination of the final two equations yields

$$R_1 = \sqrt{\sigma_2^2 + \delta_2^2}\,, \tag{20}$$

and a slightly longer calculation reveals that this leads to a unique (stable) solution for R_2 provided $A^2 > (\delta_1^2 + \sigma_1^2)(\delta_2^2 + \sigma_2^2)$.

Equation (20) gives the amplitude of the quenched solution for the structure. Compare this with the amplitude R_0 of the semi-trivial solution (17), which is the forced response of the structure without the pendulum-tuned-damper added. Written in the rescaled coordinates, this solution is

$$R_0 = \frac{A^2}{\sigma_1^2 + \delta_1^2}\,.$$

Here we make the important observation that, unlike the simple forced response R_0, the quenched amplitude R_1 is independent of the forcing amplitude A, and does not blow up at the fundamental resonance $\sigma_1 \to 0$. In fact, at least to leading order in this asymptotic approximation, the amplitude of the quenched solution is *independent of the frequency detuning and damping constant of the structure, and depends only on the frequency detuning and proportional damping of the pendulum*!

Tondl et al. [58] give a number of applications of autoparametric resonance. First, the above quenching effect means that analogues of the mass-spring-pendulum system can be used as tuned dampers, that are capable of suppressing certain undesirable response frequencies from structures or machines. Another important example of parametric resonance they cite is in flow-induced oscillation, where periodic flow features such as vortex shedding can cause a parametric response of a structure. The galloping of cables in strong winds is an example of just such an effect. Cable-stayed bridges are engineering structures that are particularly susceptible to autoparametric effects. Not only is there potential for resonant interaction between the cables and the wind, but, since the cables are all of different lengths, there is a good chance that there is at least one cable whose natural frequency is either half or twice that of a global vibration mode of the entire bridge. Models of deck-cable interaction have shown the propensity for undesirable effects, see e. g. [23,24]. For example, vibrations caused by traffic could cause particular cables to undergo large-amplitude vibration leading to potential problems with wear. Alternatively, large-amplitude cable vibrations (caused for example by fluid-structure interaction) could cause significant deck vibration leading to an uncomfortable ride for those crossing the bridge.

Combination Parametric Resonances

Mailybayev and Seyranian [41] consider general systems of second-order equations of the form

$$M\ddot{y} + \delta D\dot{y} + (A + \beta B(\Omega t))y = 0\,, \quad y \in \mathbb{R}^n\,, \tag{21}$$

where M, D and A are positive definite symmetric matrices, and B is a periodic matrix with period $2\pi/\Omega$. They derive conditions under which parametric resonance may occur. Suppose that when $\delta = 0 = \beta = 0$, the linear system has non-degenerate natural frequencies, $\omega_1, \omega_2, \ldots \omega_n$. Then they showed that the only kinds of parametric resonance that are possible are when

$$\begin{aligned} &\Omega = \frac{2\omega_j}{k}\,, \quad j = 1, \ldots m\,, \\ &\qquad\qquad k = 1, 2, \ldots, \quad \text{(simple resonance)}\,, \\ \text{or} \quad &\Omega = \frac{\omega_i \pm \omega_j}{k}\,, \quad i, j = 1, \ldots m\,,\ i > j\,, \\ &\qquad\qquad k = 1, 2, \ldots, \quad \text{(combination resonance)}\,. \end{aligned} \tag{22}$$

The sign '+' or '−' in (22) gives what are called *sum* or *difference* combination resonances respectively.

There are many examples of such combination resonances in the literature. For example, the book by Nayfeh [45] considers many such cases, especially of mechanical systems where the two modes are derived from a Galerkin reduction of a continuum system such as a plate or beam. In all these examples, though, it is a sum combination resonance that is excited. That is, by forcing the system at the sum of the two individual frequencies $\Omega = \omega_1 + \omega_2$, individual responses are found at frequency ω_1 or ω_2.

However, there do not seem to be any concrete examples of *difference* resonances in the literature. In fact, such a mechanical device might be somewhat strange. If we had $\omega_1 \approx \omega_2$, then $\Omega = \omega_1 - \omega_2$ would be small, perhaps several orders of magnitude smaller. So a difference combination resonance would give a response at a high frequency from low frequency excitation. This would be like making a drum vibrate by sending it up and down in an elevator!

In fact, for the case that $B = B_0\phi(t)$ is a constant matrix times a function of time, then it is shown in (Theorem 1 in [41]) that a given system can only excite *either* sum *or* difference combination resonances, but not both. To explain this result, it is sufficient to consider a simplification of (21) to the case $n = 2$. Consider (21) in the case $\delta = 0$ and where $B(t)$ is a constant matrix times a pure function of time with period Ω. Without loss of generality, let us suppose this function to be a pure cosine $B = B_0 \cos(\Omega t)$. Suppose that we change coordinates so that the system with $\beta = 0$ is written in diagonal form, and finally that time has been rescaled so that $\Omega = 1$. Hence we obtain a system of the form

$$\begin{pmatrix} \ddot{x}_1 \\ \ddot{x}_2 \end{pmatrix} + \left[\begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix} + \beta \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \cos(t) \right] \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0 , \tag{23}$$

where $\alpha_1 = \omega_1^2/\Omega^2$, $\alpha_2 = \omega_2^2/\Omega^2$ and the matrix $\hat{B} = \{b_{ij}\}$ is the original constant matrix within B_0 written in the transformed coordinates. Recalling, how in Subsect. "Floquet Analysis" we found the transition curves for the Mathieu equation using Floquet theory, we can look for solutions to (23) in the form

$$x_1 = c_n \sum_{n=-\infty}^{\infty} e^{int/\sqrt{2}} , \qquad x_2 = \pm c_n \sum_{n=-\infty}^{\infty} e^{sint/\sqrt{2}} , \tag{24}$$

where $s = \pm 1$. We find an infinite system of algebraic equations analogous to (10). It is straightforward to see that in the case $\beta = 0$, there is a non-trivial solution with $c_n = 0$ for all $n \neq k$ and $c_k \neq 0$, whenever $\alpha_1 + s\alpha_2 = k^2$. This would suggest that both sum and difference combination resonances are possible. However, when looking at the conditions for the bifurcation equations to have a real solution for nonzero β, one finds the following condition

$$\mathrm{sign}(b_{12} b_{21}) = s . \tag{25}$$

That is, to excite a sum resonance, the diagonal entries of B_0 must be of the same sign, and to excite a difference resonance these diagonal entries must be of opposite sign. In particular, if B_0 is a symmetric matrix (as is the case in many mechanical applications) then only sum combinations can be excited. In fact, it is argued in [68] that pure Hamiltonian systems cannot ever excite difference combination resonance; in other words, if difference parametric resonance is possible at all, then the matrix $B(\Omega t)$ must contain dissipative terms.

One might wonder whether difference combination resonance can ever be excited in any physically derived system. To show that they can, consider the following example with a non-conservative parametric excitation force, a detailed analysis of which will appear elsewhere [59].

Figure 7 depicts a double pendulum (which we think of as lying in a horizontal plane so that gravity does not affect it) with stiff, damped joints and is loaded by an end force – a so-called *follower load* – that is maintained in the direction of the axis of the second pendulum. Note that such a load is by its nature non-conservative because work (either positive or negative) has to be done to maintain the direction of the follower. In practice, such a force might be produced by a jet of fluid emerging from the end of the outer pendulum, for example if the pendula were actually pipes, although the equations of motion (and the consequent nonlinear dynamics) are rather different for

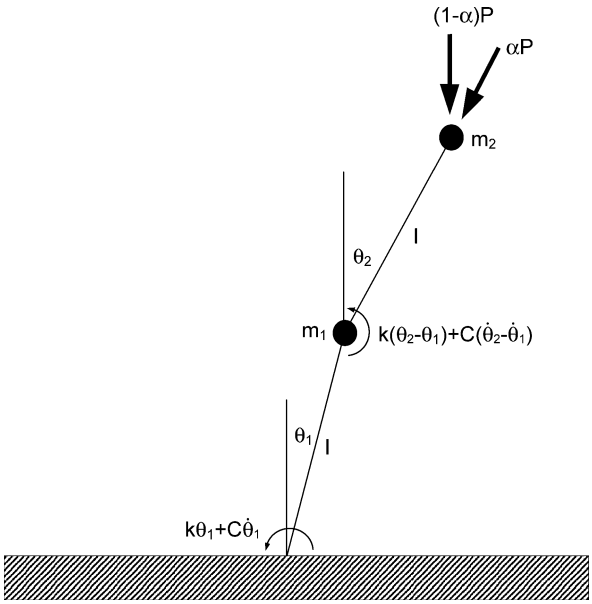

Dynamics of Parametric Excitation, Figure 7
The partially follower-loaded elastic double pendulum, depicting a definition of the various physical quantities. In this study we take $m_1 = m_2 = m$ and $\alpha = 1$

that case [5,12,13]. Non-dimensional equations of motion for this system are derived in [56], where it was shown that the system can undergo quite violent chaotic dynamics for large deflections. We shall restrict ourselves here to small deflection theory and allow the follower load to be a harmonic function of time $P = A\cos\Omega t$. After nondimensionalization using $\beta = |A|l/k$, $\delta = C/\sqrt{kml^2}$, and rescaling time according to $t_{\text{new}} = \sqrt{k/m_2 l^2)}\, t_{\text{old}}$, we obtain the equations

$$\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} + \begin{pmatrix} \ddot{\theta}_1 \\ \ddot{\theta}_2 \end{pmatrix} + \delta \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \dot{\theta}_1 \\ \dot{\theta}_2 \end{pmatrix} \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} + p\cos\Omega t \begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = 0\,. \quad (26)$$

Now, (26) is not in diagonal form, but a simple coordinate transformation can it in the form (21) with M the identity matrix,

$$D = A = \begin{pmatrix} 0.1492 & 0 \\ 0 & 3.3508 \end{pmatrix}$$

$$\text{and} \quad B_0 = \begin{pmatrix} -0.0466 & 0.3788 \\ -0.1288 & 1.0466 \end{pmatrix}.$$

Hence the system has two natural frequencies $\omega_1 = \sqrt{0.1492} = 0.3863$ and $\omega_2 = \sqrt{3.3508} = 1.8305$, and the quantities $b_{12}b_{21} = 0.3788 \times -0.1288 = -0.0488 < 0$. Hence, according to (25), a difference combination resonance should be possible by exciting the system at $\Omega = \omega_2 - \omega_1 = 1.4442$, as verified in [59].

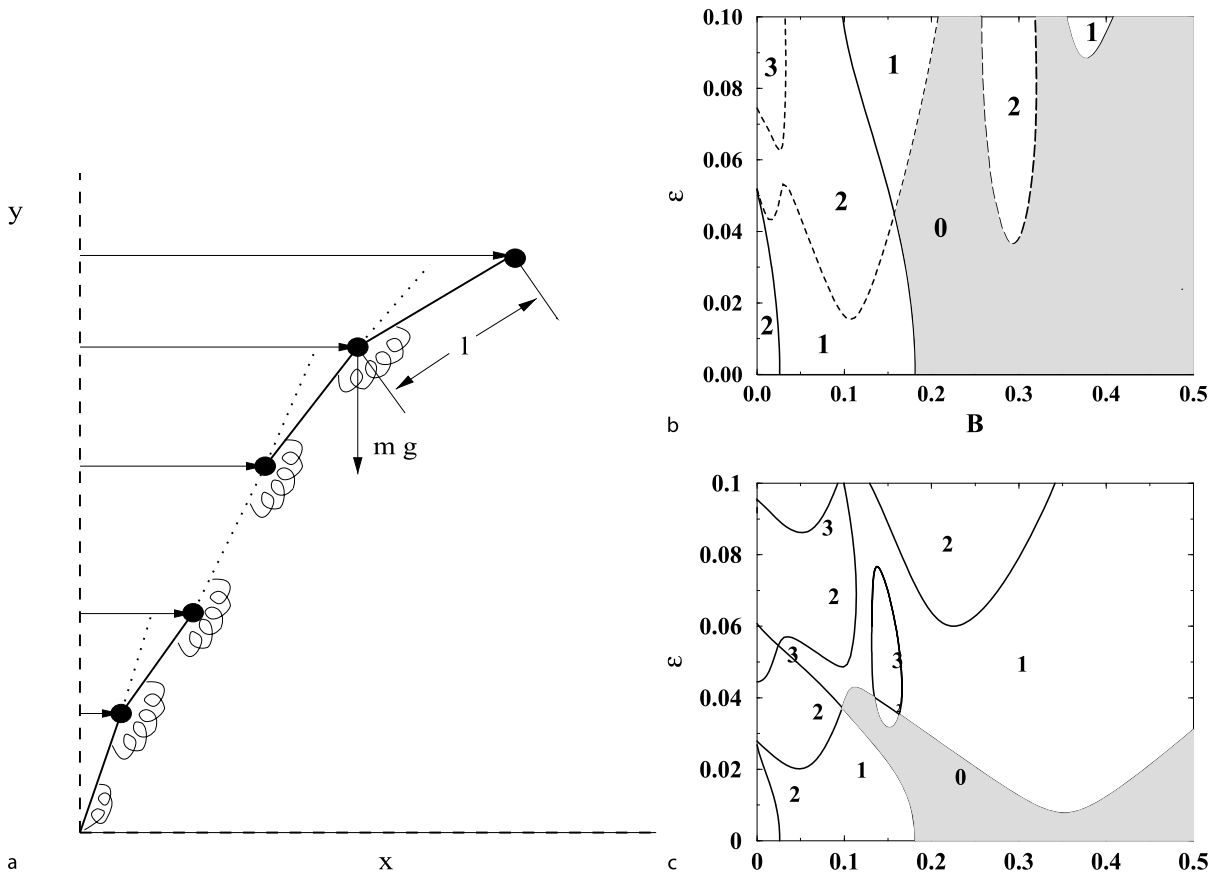

Dynamics of Parametric Excitation, Figure 8

a A multiple pendulum with stiff joints. **b,c** Ince–Strutt diagrams of dimensionless amplitude of vibration ε against scaled elastic stiffness B for eight identical jointed pendulums with weak damping. **b** For dimensionless applied frequency $\omega = 10$; **c** $\omega = 20$. Shading represents the stability region of the upright equilibrium and digits in other regions give the number of unstable Floquet multipliers. After [22], to which the reader is referred to for the precise details; reproduced with permission from Elsevier

Upside-Down Stability

As mentioned in the Introduction, it has been known for 100 years [53] that the 'negative gravity' ($\alpha < 0$) region of the Ince–Strutt diagram for the Mathieu equation implies that a simple pendulum can be stabilized in the inverted position by application of parametric excitation. More recently Acheson and Mullin [3] demonstrated experimentally that a double and even a triple pendulum can be similarly stabilized when started in the inverted position. Moreover, the stabilized equilibrium is quite robust, such that the system relaxes asymptotically back to its inverted state when given a moderate tap.

At first sight, these observations seem remarkable. A multiple pendulum has several independent normal modes of oscillation and naively one might expect that single frequency excitation could at best stabilize only one of these modes at a time. However, Acheson [1] (see also the earlier analysis of Otterbein [46]) provided a simple one-line proof that, in fact, for sufficiently high frequency and sufficiently small amplitude sinusoidal excitation, in theory *any* ideal finite chain of N pendulum can be stabilized in the inverted position. To see why this is true, consider the multiple pendulum system represented in Fig. 8 but without the springs at the joints. Then, upon performing a modal analysis, we find that the equations of motion reduce to the study of N uncoupled Mathieu equations each with different parameters $\alpha_i > 0$ and $\beta_i > 0$, for $i = 1, \ldots, N$. Hence the frequency can be chosen to be sufficiently high and the amplitude chosen to be sufficiently small for each α_i, β_i to lie in the thin wedge of stability for $\alpha < 0$ in Fig. 4. Thus, each mode is stable, and so the whole finite chain has been stabilized by parametric excitation.

It has been suggested that the limit of such a multiple pendulum, in which the total length and mass of the system stays constant while the number of pendulums in the chain becomes infinite, could be used as a possible explanation of the so called '*Indian rope trick*', a classic conjuring trick in which a rope is 'charmed' by a magician to appear vertically out of a basket like a snake. Unfortunately, as noted [1], in this limit, which is that of a piece of string, the stability region becomes vanishingly small and so this potential scientific explanation of the magic trick fails.

However, a further experiment by Mullin announced in [4], demonstrated that a piece of 'bendy curtain wire', clamped at the bottom and free at the top, that is just too long to support its own weight can be stabilized by parametric excitation (see Fig. 9). In order to explain this result, Galán et al. [22] introduced the model depicted in Fig. 8 and studied the stability of the vertical equilibrium position in the presence of both damping and stiffness in each of the joints. The normal modes of this problem lead to fully coupled equations and so Acheson's argument does not apply. Using numerical bifurcation analysis Galán et al. were able to find stability boundaries and also to find an appropriate continuum limit; providing a good motivation for us to turn our attention to parametric resonance in continuous systems.

Continuous Systems

Parametric resonance is also of importance in continuum systems. Canonical effects we shall discuss here include the

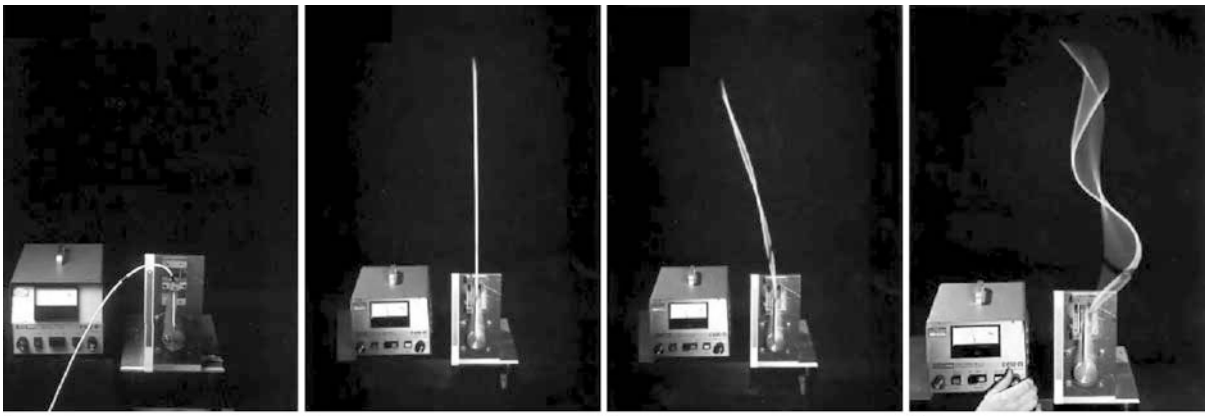

Dynamics of Parametric Excitation, Figure 9

Experimental results on a piece of 'bendy curtain wire' which is just long enough to buckle under its own weight (a). For tiny vertical oscillations between about 15 and 35 Hz, the wire can be made to stand upright (b), and to be stable against even quite large perturbations (c). Beyond the upper frequency threshold, a violent dynamic stability sets in which involves the third vibration mode excited harmonically (d). After [44], reproduced with permission from the Royal Society

wire stiffening we have just introduced, the excitation of patterns at fluid interfaces, more general theories of flow-induced oscillations of structures, and the stabilizing effects of periodic modulation of optical fibers.

Structural Stiffening; the 'Indian Rod Trick'

Thomsen [57] has shown that parametric excitation generally speaking has a stiffening effect on a structure. This is essentially because of the same effect that causes the upside down stability of the simple pendulum for high-enough frequencies and small enough amplitude. However, detailed experiments by Mullin and co-workers [44], reproduced in Fig. 9, showed that there can also be a destabilizing effect of parametric resonance. In particular, a piece of curtain wire (a thin helical spring surrounded by plastic) that was just too long (about 50 cm) to support its own weight, was clamped vertically upright and subjected to rapid, small-amplitude oscillations of about 1 cm peak-to-peak. At around 15 Hz the rod was found to be stabilized, such that when released it remains erect (see Fig. 9b). This effect continues for higher frequencies, until about 30 Hz at which point a violent dynamic instability sets in that is commensurate with the driving frequency (harmonic, rather than subharmonic). The experimentally determined parameter values in which the upside-down stabilization can be achieved are represented in terms of dimensionless parameters in Fig. 10b.

Champneys and Fraser [14,15,20], introduced a continuum theory to explain these experimental results, which produced the shaded region in Fig. 10b, that we now explain. Consider an intrinsically straight column of length ℓ, with a uniform circular cross-section of radius $a \ll \ell$ and mass linear density m per unit length (see Fig. 10a). The column is assumed to be inextensible, unshearable and linearly elastic with bending stiffness $\hat{B}$ and material damping coefficient $\hat{\gamma}$ (where hatted variables are dimensional). The lower end is clamped upright and is to vertical displacement $\Delta \cos \omega \hat{t}$, whereas the upper end is free. In [15], it is shown that the stability of the vertical equilibrium position can be studied by analyzing solutions to the PDE

$$Bu_{ssss} + [(1-s)u_s]_s + \eta\,(u_{tt} + \gamma u_{sssst} - \varepsilon[(1-s)u_s]_s \cos t) = 0\,, \quad (27)$$

for the scaled lateral displacement $u(s,t)$ assumed without loss of generality to be in a fixed coordinate direction, where the scaled arc length $s \in (0,1)$, the scaled time variable $t = \omega \hat{t}$, and the boundary conditions are

$$u = u_s = 0\,, \text{ at } s = 0\,; \quad u_{ss} = u_{ssss} = 0 \text{ at } s = 1\,. \quad (28)$$

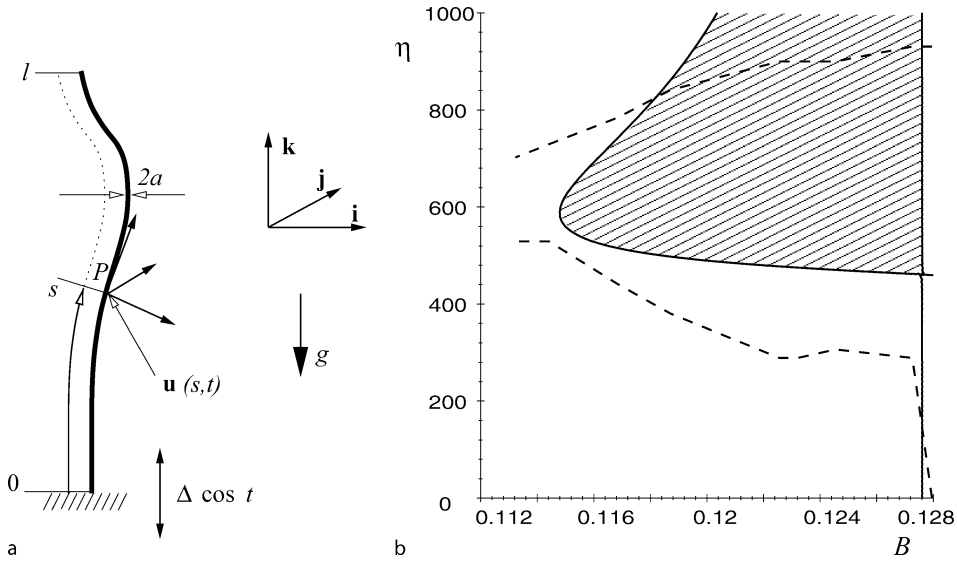

Dynamics of Parametric Excitation, Figure 10

a Definition sketch of the parametrically excited flexible rod model. b Comparison between theory (*solid line* and *shading*) and the experimental results of [44] (*dashed line*), for the *stability* region as a function of the dimensionless parameters B and η using the experimental value $\varepsilon = 0.02$, and the representative damping value $\gamma_1 = 0.01$ (which was not measured in the experiment) The theory is based on calculations close to the resonance tongue interaction between when $B_{0,1} = B_{1,3}$ when $\eta = \eta_c = 460.7$. After [15], reproduced with permission from SIAM

The dimensionless parameters appearing in (27) are

$$\gamma = \frac{\Gamma \ell \omega}{mg}, \quad B = \frac{\hat{B}}{mg\ell^3}, \quad \eta = \frac{\omega^2 \ell}{g}, \quad \varepsilon = \frac{\Delta}{\ell}, \quad (29)$$

which represent respectively material damping, bending stiffness, the ratio of forcing frequency to the equivalent pendulum natural frequency, and forcing amplitude.

If we momentarily consider the problem with no forcing or damping, $\varepsilon = \gamma = 0$, then the eigenmodes ϕ_n and corresponding natural frequencies $\sqrt{\lambda_n/\eta}$ of free vibration of the rod satisfy

$$B(\phi_n)_{ssss} + [(1-s)(\phi_n)_s]_s \phi_n - \lambda_n \phi_n = 0 ,$$

and form a complete basis of functions that satisfy the boundary conditions (28). Now the eigenmodes $\phi_n(s; B)$ are in general not expressible in closed form except at the special values of B at which $\lambda_n(B) = 0$, in which case [20] there is a solution in terms of the Bessel function $J_{-1/3}$. The same analysis shows that there are infinitely many such B-values, $B_{0,n}$, for $n = 1, 2, 3, 4, \ldots$, accumulating at $B = 0$, the first few values of which are $B_{0,1} = 0.127594$, $B_{0,2} = 0.017864$, $B_{0,3} = 0.0067336$ and $B_{0,4} = 0.0003503$. These correspond respectively to where each eigenvalue locus $\lambda_n(B)$, for $n = 1, 2, 3, 4$, crosses the B-axis, with the corresponding eigenmodes there having $n-1$ internal zeros. The lowering of B through each $B_{0,n}$-value implies that the nth mode becomes linearly unstable. Hence for $B > B_c := B_{0,1}$ the unforced rod is stable to self-weight buckling (a result known to Greenhill [26]). Moreover, it can be shown [14,15] that each eigenmode $\phi_n(s)$ retains a qualitatively similar mode shape for $B > B_{0,n}$ with the nth mode having $n = 1$ internal zeros, and that the corresponding loci $\lambda_n(B)$ are approximately straight lines as shown in the upper part of Fig. 11.

The lower part of Fig. 11 shows the results of a parametric analysis of the Eq. (27) for $\gamma = 0$. Basically, each time one of the eigenvalue loci passes through η times the square of a half-integer $(j/2)^2$ we reach a B-value for which there is the root point of an instability tongue. Essentially we have a Mathieu-equation-like Ince–Strutt diagram in the (B, ε) plane for *each* mode n. These diagrams are overlaid on top of each other to produce the overall stability plot. However, note from Fig. 11 that the slope of each eigenvalue locus varies with the frequency ratio η. Hence as η increases, the values of all the $B_{j/2,n}$, $j = 0, \ldots, \infty$ for a given n increase with the same fixed speed. But this speed varies with the mode number n. Hence the instability tongues slide over each other as the frequency varies.

Note the schematic diagram in the lower panel of Fig. 11 indicates a region of the stability for $B < B_{0,1}$ for

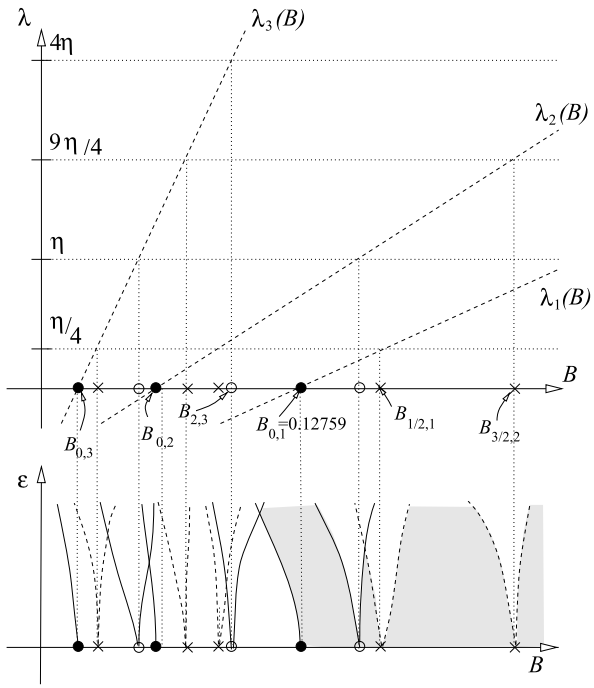

Dynamics of Parametric Excitation, Figure 11
Summarizing schematically the results of the parametric resonance analysis of the parametric excited vertical column in the absence of damping. The *upper part* shows eigenvalue loci $\lambda_n(B)$ and the definition of the points $B_{j/2,n}$, $j = 0, 1, 2, 3, \ldots$ In the *lower part*, instability tongues in the (B, ε)-plane are shown to occur with root points $B_{j/2,n}$ and to have width ε^j. The *shaded regions* correspond to the where the vertical solution is stable. *Solid lines* represent neutral stability curves with Floquet multiplier +1 (where α is an integer) and *dashed lines* to multipliers −1 (where α is half an odd integer). Reproduced from [14] with permission from the Royal Society

small positive ε. Such a region would indicate the existence of rod of (marginally) lower bending stiffness than that required for static stability, that can be stabilized by the introduction of the parametric excitation. Essentially, the critical bending stiffness for the buckling instability is reduced (or, equivalently since $B \propto \ell^{-3}$, the critical length for buckling is increased). This effect has been dubbed the '*Indian rod trick*' because the need for bending stiffness in the wire means that mathematically this is a rod, not a rope (or string); see, e. g. the account in [4].

To determine in theory the precise parameter values at which the rod trick works turns out to be an involved process. That the instability boundary emerging from $B_{0,1}$ in the lower panel of Fig. 11 should *usually* bend back to the left can be argued geometrically [7]. However, for η-values at which $B_{0,1} \approx B_{1,n}$ for some n, it has been shown that there is actually a singularity in the coefficient of the quadratic part of this buckling curve causing the instabil-

ity boundary to bend to the right [20]. A complete unfolding of this so-called *resonance tongue interaction* was performed in [15], also allowing for the presence of material damping. The resulting asymptotic analysis needs to be carried out in the presence of *four* small parameters; ε, γ, $B - B_{0,1}$ and $\eta - \eta_c$ where η_c is the frequency-ratio value at which the resonance-tongue interaction occurs. The results produce the shaded region shown in Fig. 10b, which can be seen to agree well with the experimental results.

Faraday Waves and Pattern Formation

When a vessel containing a liquid is vibrated vertically, it is well known that certain standing wave patterns form on the free surface. It was Faraday in 1831 [19] who first noted that the frequency of the liquid motion is half that of the vessel. These results were confirmed by Rayleigh [50] who suggested that what we now know as parametric resonance was the cause. This explanation was shown to be correct by Benjamin and Ursell [6] who derived a system of uncoupled Mathieu equations from the Navier–Stokes equations describing the motion of an ideal (non-viscous) fluid. There is one Mathieu equation (4), for each wave number i that fits into the domain, each with its own parameters α_i and β_i. They argue that viscous effects would effectively introduce a linear damping term, as in (8) and they found good agreement between the broad principal (subharmonic) instability tongue and the results of a simple experiment. Later, Kumer and Tuckerman [36] produced a more accurate Ince–Strutt diagram (also known as a dispersion relation in this context) for the instabilities of free surfaces between two viscous fluids. Through numerical solution of appropriate *coupled* Mathieu equations that correctly take into account the boundary conditions, they showed that viscosity acts in a more complicated way than the simple velocity proportional damping term in (8). They produced the diagram shown in Fig. 12b, which showed a good match with the experimental results. Note that, compared with the idealized theory, Fig. 12a, the correct treatment of viscosity causes a small shift in the critical wavelength of each instability, a broadening of the higher-harmonic resonance tongues and a lowering of the amplitude required for the onset of the principal subharmonic resonance if damping is introduced to the idealized theory (see the inset to Fig. 12b).

More interest comes when Faraday waves are excited in a vessel where two different wave modes have their subharmonic resonance at nearly the same applied frequency, or when the excitation contains two (or more) frequencies each being close to twice the natural frequency of a standing wave. See for example the book by Hoyle [30] and references therein, for a glimpse at the amazing complexity of the patterns of vibration that can result and for their explanation in terms of dynamical systems with symmetry. Recently a new phenomenon has also been established, namely that of localized spots of pattern that can form under parametric resonance, so-called *oscillons* see e. g. [61] for a review and [39] for the first steps at an explanation.

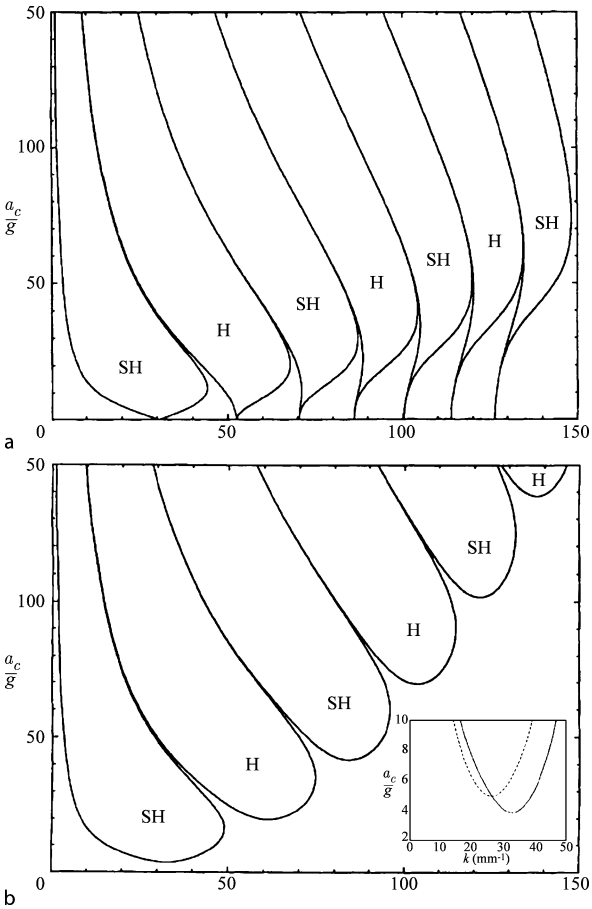

Dynamics of Parametric Excitation, Figure 12
Stability boundaries for Faraday waves from: a Benjamin and Ursell's theory for ideal fluids; b the results of full hydrodynamic stability analysis by Kumar and Tuckerman. The horizontal axis is the frequency and the vertical scale the amplitude of the excitation. The *inset* to panel b shows the difference between the true instability boundary for the principal subharmonic resonance (*lower curve*) and the idealized boundary in the presence of simple proportional damping. After [36], to which the reader is referred for the relevant parameter values; reproduced with permission from Cambridge University Press

Fluid-Structure Interaction

Faraday waves are parametric effects induced by the gross mechanical excitation of a fluid. In Subsect. "Autopara-

metric Resonance" we briefly mentioned that fluid-structure interaction can go the other way, namely that external fluid flow effects such as vortex shedding can excite (auto)parametric resonances in simple mechanical systems. Such flow-induced parametric resonance can also occur for continuous structures, in which waves can be excited in the structure due to external parametric loading of the fluid; see for example the extensive review by Païdousis and Li [47]. They also treat cases where the flow is internal to the structure: we are all familiar with the noisy ducts and pipes, for example in central heating or rumbling digestive systems, and it could well be that many of these effects are a result of parametric resonance.

Semler and Païdousis [52] consider in detail parametric resonance effects in the simple paradigm of a cantilevered (free at one end, clamped at the other) flexible pipe conveying fluid. Think of this as a hosepipe resting on a garden lawn. Many authors have considered such an apparatus as a paradigm for internal flow-induced oscillation problems (see the extensive references in [47,52]), starting from Benjamin's [5] pioneering derivation of equations of motion from the limit of a segmented rigid pipe. The equations of motion for such a situation are similar to those of the rod (27), but with significant nonlinear terms arising from inertial and fluid-structure interaction effects. In the absence of periodic fluctuations in the flow, then it is well known that there is a critical flow rate beyond which the pipe will become oscillatory, through a Hopf bifurcation (flutter instability). This effect can be seen in the way that hosepipes writhe around on a lawn, and indeed forms the basis for a popular garden toy. Another form of instability can occur when the mean flow rate is beneath the critical value for flutter, but the flow rate is subject to a periodic variation, see [52] and references therein. There, a parametric instability can occur when the pulse rate is twice the natural frequency of small amplitude vibrations of the pipe. In [52] this effect was observed in theory, in numerical simulation and in a careful experimental study, with good agreement between the three. Parametric resonance could also be found for super-critical flow rates. In this case, the effect was that the periodic oscillation born at the Hopf bifurcation was excited into a quasi-periodic response.

Another form of fluid structure interaction can occur when the properties of an elastic body immersed in a fluid are subject to a temporal periodic variation. With a view to potential biological applications such as active hearing processes and heart muscles immersed in blood, Cortez et al. [16] consider a two-dimensional patch of viscous fluid containing a ring filament that is excited via periodic variation of its elastic modulus. Using a mixture of immersed boundary numerical methods and analytic reductions, they are able to find an effective Ince–Strutt diagram for this system and demonstrate the existence of parametric instability in which the excitation of the structure causes oscillations in the fluid. Taking a fixed wave number p of the ring in space, they find a sequence of harmonic and subharmonic instabilities as a parameter effectively like α in the Mathieu equation is increased. Interestingly though, even for low viscosity, the principal subharmonic resonance tongue (corresponding to $\alpha = 1/4$) does not occur for low amplitude forcing, and the most easily excited instability is actually the fundamental, harmonic resonance (corresponding to $\alpha = 1$).

Dispersion Managed Optical Solitons

The stabilizing effect of high-frequency parametric excitation has in recent years seen a completely different application, in an area of high technological importance, namely optical communications. Pulses of light sent along optical fibers are a crucial part of global communication systems. A pulse represents a packet of linear waves, with a continuum of different frequencies. However, optical fibers are fundamentally dispersive, which means that waves with different wavelengths travel at different group velocities. In addition, nonlinearity, such as the Kerr effect of intensity-dependent refractive index change, and linear losses mean that there is a limit to how far a single optical signal can be sent without the need for some kind of amplifier. Depending on the construction of the fiber, dispersion can be either *normal* which means that lower frequencies travel faster, or *anomalous* in which case higher frequencies travel faster.

One idea to overcome dispersive effects, to enable pulses to be travel over longer distances without breaking up, is to periodically vary the medium with which the fiber is constructed so that the dispersion is alternately normally and anomalously dispersive [38]. Another idea, first proposed by Hasegawa and Tappert [28], is to balance dispersion with nonlinearity. This works because the complex wave envelope A of the electric field of the pulse can be shown to leading order to satisfy the nonlinear Schrödinger (NLS) equation

$$i\frac{\mathrm{d}A}{\mathrm{d}z} + \frac{1}{2}\frac{\mathrm{d}^2A}{\mathrm{d}t^2} + |A|^2A = 0\,, \tag{30}$$

in which time and space have been rescaled so that the anomalous dispersion coefficient (multiplying the second-derivative term) and the coefficient of the Kerr nonlinearity are both unity. Note the peculiarity in optics compared to other wave-bearing system in that time t plays the role

of the spatial coordinate, and the propagation distance z is the evolution variable. Thus we think of pulses as being time-traces that propagate along the fiber. Now, the NLS equation is completely integrable and has the well-known soliton solution

$$A(z,t) = \eta \operatorname{sech}(\eta t - cz) e^{i\eta^2 z} , \tag{31}$$

which represents a smooth pulse with amplitude η, which propagates at 'speed' c. Moreover, Eq. (30) is completely integrable [69] and so arbitrary initial conditions break up into solitons of different speed and amplitude. Optical solitons were first realized experimentally by Mollenauer et al. [43]. However, true optical fibers still suffer linear losses, and the necessary periodic amplification of the optical solitons to enable them to survive over long distances can cause them to *jitter* [25] and to interact with each other as they move about in time.

The idea of combining nonlinearity with dispersion management in order to stabilize soliton propagation was first introduced by Knox, Forysiak and Doran [35], was demonstrated experimentally by Suzuki et al. [54] and has shown surprisingly advantageous performance that is now being exploited technologically (see [60] for a review). Essentially, dispersion management introduces a periodic variation of the dispersion coefficient $d(z)$, that can be used to compensate the effects of periodic amplification, which with the advent of modern erbium-doped optical amplifiers can be modeled as a periodic variation of the nonlinearity. As an envelope equation we arrive at the modified form of (30),

$$i\frac{\mathrm{d}A}{\mathrm{d}z} + \frac{d(z)}{2}\frac{\mathrm{d}^2 A}{\mathrm{d}t^2} + \sigma(z)|A|^2 A = 0 , \tag{32}$$

where the period L of the dispersion map $d(z)$ and the amplifier spacing (the period of $\sigma(z)$) Z may not necessarily be the same. Two particular regimes are of interest. One, of application to long-haul, sub-sea transmission systems where the cable is already laid with fixed amplifiers, is to have $L \gg Z$. Here, one modifies an existing cable by applying bits of extra fiber with large dispersion of the opposite sign only occasionally. In this case, one can effectively look only at the long scale and average out the effect of the variation of σ. Then the parametric term only multiplies the second derivative. Another interesting regime, of relevance perhaps when designing new terrestrial communication lines, is to have $L = Z$ so that one adds loops of extra opposite-dispersion fiber at each amplification site. In both of these limits we have an evolution equation whose coefficients are periodic functions of the evolution variable z with period L.

In either regime, the periodic (in z) parameter variation has a profound stabilizing influence, such that the effects of linear losses and jitter are substantially reduced. A particular observation is that soliton solutions appear to (32) *breathe* during each period; that is, their amplitude and phase undergo significant periodic modulation. Also, with increase in the amplitude of the oscillatory part of d, so the average shape of the breathing soliton becomes less like a sech-pulse and more like a Gaussian. An advantage is that the Gaussian has much greater energy the regular NLS soliton (31) for the same average values of d and σ, hence there is less of a requirement for amplification. Surprisingly, it has been found that the dispersion managed solitons can propagate stably even if the average of d is slightly positive, i. e. in the normal dispersion region where there is no corresponding NLS soliton. This has a hugely advantageous consequence. The fact that the local dispersion can be chosen to be high, but the average dispersion can be chosen to be close to zero, is the key to enabling the jitter introduced by the amplifiers to be greatly suppressed compared with the regular solitons. Finally the large local modulation of phase that is a consequence of the breathing, means that neighboring solitons are less likely to interact with each other. A mathematical analysis of why these surprising effects arise is beyond the scope of this article, see e. g. [32,60] for reviews, but broadly-speaking the stable propagation of dispersion-managed solitons is due to the stabilizing effects of high frequency parametric excitation.

Future Directions

This article has represented a whistle-stop tour of a wide areas of science. One of the main simplifications we have taken is that we have assumed the applied excitation to the system to be periodic; in fact, in most examples it has been pure sinusoidal. There is a growing literature on quasiperiodic forcing, see e. g. [51], and it would be interesting to look at those effects that are unique to parametrically excited systems that contain two or more frequencies. The inclusion of noise in the excitation also needs to be investigated. What effect does small additive or multiplicative noise have on the shape of the resonance tongues? And what is the connection with the phenomenon of stochastic resonance [66]?

There is a mature theory of nonlinear resonance in systems with two degrees of freedom which, when in an integrable limit, undergoes quasiperiodic motion with two independent frequencies. Delicate results, such as the KAM theorem, see e. g. [55], determine precisely which quasiperiodic motions (invariant tori) survive as a func-

tion of two parameters, the size of the perturbation from integrability and the frequency detuning between the two uncoupled periodic motions. For the tori that break up (inside so-called Arnol'd tongues) we get heteroclinic tangles and bands of chaotic motion. Parametric resonance tends to be slightly different though, since in the uncoupled limit, one of the two modes typically has zero amplitude (what we referred to earlier as a semi-trivial solution). It would seem that there is scope for a deeper connection to be established between the Arnol'd tongues of KAM theory and the resonance tongues that occur under parametric excitation.

In terms of applications, a large area that remains to be explored further is the potential exploitation of parametric effects in fluid-structure interaction problems. As I write, the heating pipes in my office are buzzing at an extremely annoying frequency that seems to arise on warm days when many radiators have been switched off. Is this a parametric effect? It would also seem likely that nature uses parametric resonance in motion. Muscles contract longitudinally, but motion can be in a transverse direction. Could phenomena such as the swishing of bacterial flagella, fish swimming mechanisms and the pumping of fluids through vessels, be understood in terms of exploitation of natural parametric resonances? How about motion of vocal chords, the Basilar membrane in the inner ear, or the global scale pumping of the heart? Does nature naturally try to tune tissue to find the primary subharmonic resonance tongue? Perhaps other effects of parametric excitation that we have uncovered here, such as the quenching of resonant vibrations, structural stiffening, and frequency tuned motion, are already being usefully exploited in nature. If so, there could be interesting consequences for nature inspired design in the engineered world.

Finally, let us return to our opening paradigm of how children effectively excite a parametric instability in order to induce motion in a playground swing. In fact, it was carefully shown by Case and Swanson [11] that the mechanism the child uses when in the sitting position is far more accurately described as an example of direct forcing rather than parametric excitation. By pushing on the ropes of the swing during the backwards flight, and pulling on them during the forward flight, the child is predominately shifting his center of gravity to and fro, rather than up and down; effectively as in Fig. 1c rather than Fig. 1b. Later, Case [10] argued that, even in the standing position, the energy gained by direct forcing from leaning backwards and forwards is likely to greatly outweigh any gained from parametric up-and-down motion of the child. The popular myth that children swinging is a suitable playground science demonstrator of parametric resonance was finally put to bed by Post et al. [49]. They carried out experiments on human patients and analyzed the mechanisms they use to move, reaching the conclusion that even in the standing position, typically 95% of the energy comes from direct forcing, although parametric effects do make a slightly more significant contribution for larger amplitudes of swing.

Bibliography

1. Acheson D (1993) A pendulum theorem. Proc Roy Soc Lond A 443:239–245
2. Acheson D (1995) Multiple-nodding oscillations of a driven inverted pendulum. Proc Roy Soc Lond A 448:89–95
3. Acheson D, Mullin T (1993) Upside-down pendulums. Nature 366:215–216
4. Acheson D, Mullin T (1998) Ropy magic. New Scientist February 157:32–33
5. Benjamin T (1961) Dynamics of a system of articulated pipes conveying fluid. 1. Theory. Proc Roy Soc Lond A 261:457–486
6. Benjamin T, Ursell F (1954) The stability of a plane free surface of a liquid in vertical periodic motion. Proc Roy Soc Lond A 255:505–515
7. Broer H, Levi M (1995) Geometrical aspects of stability theory for Hills equations. Arch Ration Mech Anal 131:225–240
8. Broer H, Vegter G (1992) Bifurcation aspects of parametric resonance. Dynamics Reported (new series) 1:1–53
9. Broer H, Hoveijn I, van Noort M (1998) A reversible bifurcation analysis of the inverted pendulum. Physica D 112:50–63
10. Case W (1996) The pumping of a swing from the standing position. Am J Phys 64:215–220
11. Case W, Swanson M (1990) The pumping of a swing from the seated position. Am J Phys 58:463–467
12. Champneys A (1991) Homoclinic orbits in the dynamics of articulated pipes conveying fluid. Nonlinearity 4:747–774
13. Champneys A (1993) Homoclinic tangencies in the dynamics of articulated pipes conveying fluid. Physica D 62:347–359
14. Champneys A, Fraser W (2000) The 'Indian rope trick' for a continuously flexible rod; linearized analysis. Proc Roy Soc Lond A 456:553–570
15. Champneys A, Fraser W (2004) Resonance tongue interaction in the parametrically excited column. SIAM J Appl Math 65:267–298
16. Cortez R, Peskin C, Stockie J, Varela D (2004) Parametric resonance in immersed elastic boundaries. SIAM J Appl Math 65:494–520
17. Curry S (1976) How children swing. Am J Phys 44:924–926
18. Cvitanovic P (1984) Universality in Chaos. Adam Hilger, Bristol
19. Faraday M (1831) On the forms of states of fluids on vibrating elastic surfaces. Phil Trans Roy Soc Lond 52:319–340
20. Fraser W, Champneys A (2002) The 'Indian rope trick' for a parametrically excited flexible rod: nonlinear and subharmonic analysis. Proc Roy Soc Lond A 458:1353–1373
21. Galan J (2002) Personal communication
22. Galán J, Fraser W, Acheson D, Champneys A (2005) The parametrically excited upside-down rod: an elastic jointed pendulum model. J Sound Vibration 280:359–377
23. Gattulli V, Lepidi M (2003) Nonlinear interactions in the

planar dynamics of cable-stayed beam. Int J Solids Struct 40:4729–4748
24. Gattulli V, Lepidi M, Macdonald J, Taylor C (2005) One-to-two global-local interaction in a cable-stayed beam observed through analytical, finite element and experimental models. Int J Nonlin Mech 40:571–588
25. Gordon J, Haus H (1986) Random walk of coherently amplified solitons in optical fiber transmission. Opt Lett 11:665–666
26. Greenhill A (1881) Determination of the greatest height consistent with stability that a pole or mast can be made. Proceedings of the Cambridge Philosophical Society IV, Oct 1880 – May 1883, pp 65–73
27. Guckenheimer J, Holmes P (1983) Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields. Springer, New York
28. Hasegawa A, Tappert F (1991) Transmission of stationary nonlinear optical pulses in dispersive dielectric fibres. i. anomolous dispersion. Appl Phys Lett 23:142–144
29. Hill G (1886) On the part of the motion of lunar perigee which is a function of the mean motions of the sun and moon. Acta Math 8:1–36
30. Hoyle R (2005) Pattern Formation; An Introduction to Methods. CUP, Cambridge
31. Ince E (1956) Ordinary Differential Equations. Dover, New York
32. Jones C (2003) Creating stability from instability. In: Nonlinear dynamics and chaos; where do we go from here? IOP Publishing, Bristol, chap 4, pp 73–90
33. Jordan DW, Smith P (1998) Nonlinear Ordinary Differential Equations, 3rd edn. Oxford University Press, Oxford
34. Kapitza P (1951) Dinamicheskaya ustoichivost mayatnika pri koleblyushcheisya tochke podvesa. Z Eksperimentalnoi I Teoreticheskoi Fiziki 21:588–597, in Russian
35. Knoz F, Forysiak W, Doran N (1995) 10 gbit/*s* soliton communication systems over standard fiber at $1.55\mu m$ and the use of dispersion compensation. IEEE J Lightwave Technol 13:1960–1995
36. Kumar K, Tuckerman L (1994) Parametric instability of the interface between two fluids. J Fluid Mech 279:49–68
37. Kuznetsov YA (2004) Elements of Applied Bifurcation Theory, 3rd edn. Springer, New York
38. Lin C, Kogelnik H, Cohen L (1980) Optical-pulse equilization of low-dispersion transmission in single fibers in the 1.3–$1.7\mu m$ spectral region. Opt Lett 5:476–480
39. Lloyd D, Sandstede B, Avitabilie D, Champneys A (2008) Localized hexagon patterns in the planar swift-hohenberg equation. To appear in SIAM Appl Dyn Sys
40. Magnus W, Winkler S (1979) Hill's Equation. Dover, New York
41. Mailybayev A, Seyranian AP (2001) Parametric resonance in systems with small dissipation. PMM J App Math Mech 65:755–767
42. Mathieu E (1868) Mémoire sur le mouvement vibratoire d'une membrane de forme elliptique. J Math Pure Appl 13:137–203
43. Mollenauer L, Stolen R, Gordon J (1980) Experimental observation of picosecond pulse narrowing and solitons in optical fibers. Phys Rev Lett 45:1095–1097
44. Mullin A, Champneys T, Fraser W, Galan J, Acheson D (2003) The 'Indian wire trick' via parametric excitation: a comparison between theory and experiment. Proc Roy Soc Lond A 459:539–546
45. Nayfeh A (2000) Nonlinear Interactions: Analytical, Computational and Experimental Methods. Wiley Interscience, New York
46. Otterbein S (1982) Stabilization of the *N*-pendulum and the Indian link trick. Arch Rat Mech Anal 78:381–393
47. Païdoussis MP, Li G (1993) Pipes conveying fluid: a model dynamical problem. J Fluids Structures 7:137–204
48. Piccoli B, Kulkarni J (2005) Pumping a swing by standing and squatting: Do children pump time optimally? IEEE Control Systems Magazine 25:48–56
49. Post A, de Groot G, Daffertshofer A, Beek P (2007) Pumping a playground swing. Motor Control 11:136–150
50. Rayliegh L (1883) On the crispations of fluid resting upon a vibrating support. Phil Mag 16:50
51. Romeiras F, Romeiras F, Bondeson A, Ott E, Antonsen TM, Grebogi C (1989) Quasiperiodic forcing and the observability of strange nonchaotic attractors. Phys Scr 40:442–444
52. Semler C, Païdoussis M (1996) Nonlinear analysis of the parametric resonances of a planar fluid-conveying cantilevered pipe. J Fluids Structures 10:787–825
53. Stephenson A (1908) On a new type of dynamical stability. Mem Proc Manch Lit Phil Soc 52:1–10
54. Suzuki M, Morita N, Edagawa I, Yamamoto S, Taga H, Akiba S (1995) Reduction of gordon-haas timing jitter by periodic dispersion compensation in soliton transmission. Electron Lett 31:2027–2035
55. Tabor M (1989) Chaos and Integrability in Nonlinear Dynamics: An Introduction. Wiley, New York
56. Thomsen J (1995) Chaotic dynamics of the partially follower-loaded elastic double pendulum. J Sound Vibr 188:385–405
57. Thomsen J (2003) Vibrations and stability: advanced theory, analysis, and tools, 2nd edn. Springer, New York
58. Tondl A, Ruijgrok T, Verhulst F, Nabergoj R (2000) Autoparametric Resonance in Mechanical Systems. Cambridge University Press, Cambridge
59. Truman M, Galán J, Champneys A (2008) An example of difference combination resonance. In preparation
60. Turytsyn S, Shapiro E, Medvedev S, Fedoruk M, Mezentsev V (2003) Physics and mathematics of dispersion-managed optical solitons. CR Physique 4:145–161
61. Umbanhowar PB, Melo F, Swinney HL (1996) Localized excitations in a vertically vibrated granular layer. Nature 382:793–796
62. Vanderbauwhede A (1990) Subharmonic branching in reversible-systems. SIAM J Math Anal 21:954–979
63. van der Pol B, Strutt M (1928) On the stability of the solutions of mathieu's equation. Phil Mag 5:18–38, sp. Iss. 7th Series
64. van Noort M (2001) The parametrically forced pendulum. a case study in 1 1/2 degree of freedom. Ph D thesis, RU Groningen
65. Verhulst F (2000) Nonlinear Differential Equations and Dynamical Systems. Springer, New York
66. Wiesenfeld K, Moss F (1995) Stochastic resonance and the benefits of noise: from ice ages to crayfish and squids. Nature pp 33–36
67. Wiggins S (2003) Introduction to Applied Nonlinear Dynamical Systems and Chaos, 2nd edn. Springer, New York
68. Yakubovich VA, Starzhinskii VM (1987) Pammetric Resonance in Linear Systems. Nauka, Moscow, in Russian
69. Zakharov V, Shabat A (1971) Exact theory of two-dimensional self focussing and one-dimensional modulation of waves in nonlinear media. Sov Phys JETP 33:77–83

E

Earthquake Clusters over Multi-dimensional Space, Visualization of

David A. Yuen[1], Witold Dzwinel[2], Yehuda Ben-Zion[3], Ben Kadlec[4]
[1] Dept. of Geology and Geophysics, University of Minnesota, Minneapolis, USA
[2] Dept. of Computer Science, AGH University of Sci. and Technol., Kraków, Poland
[3] Department of Earth Sciences, University of Southern California, Los Angeles, USA
[4] Department of Computer Science, University of Colorado, Boulder, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Earthquakes Clustering
Multidimensional Feature Space
The Detection and Visualization of Clusters in Multi-Dimensional Feature Space
Description of the Data
Earthquake Visualization by Using Clustering in Feature Space
Remote Problem Solving Environment (PSE) for Analyzing Earthquake Clusters
Future Directions
Acknowledgments
Bibliography

Glossary

Grid Virtual metacomputer, which uses a network of geographically distributed local networks, computers and computational resources and services. *Grid Computing* focuses on distributed computing technologies, which are not in the traditional dedicated clusters. *Data Grids* – represent controlled sharing and management of large amounts of distributed data.

Problem solving environment (PSE) A specialized computer software for solving one class of problems. They use the language of the respective field and often employ modern graphical user interfaces. The goal is to make the software easy to use for specialists in fields other than computer science. PSEs are available for generic problems like data visualization or large systems of equations and for narrow fields of science or engineering.

Global seismographic network (GSN) The goal of the GSN is to deploy permanent seismic recording stations uniformly over the earth's surface. The GSN stations continuously record seismic data from very broad band seismometers at 20 samples per second, and to provide for high-frequency (40 sps) and strong-motion (1 and 100 sps) sensors where scientifically warranted. It is also the goal of the GSN to provide for real-time access to its data via Internet or satellite. Over 75% of the over 128 GSN stations meet this goal as of 2003.

WEB-IS A software tool that allows remote, interactive visualization and analysis of large-scale 3-D earthquake clusters over the Internet through the interaction between client and server.

Scientific visualization is branch of computer graphics and user interface design that are dealing with presenting data to users, by means of patterns and images. The goal of scientific visualization is to improve understanding of the data being presented.

Interactive visualization is a branch of graphic visualization that studies how humans interact with computers to create graphic illustrations of information and how this process can be made more efficient. **Remote-visualization** – the tools for interactive visualization of high-resolution images on remote client machine, rendered and preprocessed on the server.

OpenGL A standard specification defining a cross-language cross-platform API for writing applications that produce 2D and 3D computer graphics.

Sumatra-Andaman earthquake An undersea earthquake that occurred at 00:58:53 UTC (07:58:53 local time) December 26, 2004, with an epicenter off the west coast of Sumatra, Indonesia. The earthquake triggered a series of devastating tsunamis along the coasts of most landmasses bordering the Indian Ocean, killing large numbers of people and inundating coastal communities across South and Southeast Asia, including parts of Indonesia, Sri Lanka, India, and Thailand.

Earthquake catalog Data set consisting of earthquake hypocenters, origin times, and magnitudes. Additional information may include phase and amplitude readings, as well as first-motion mechanisms and moment tensors.

Pattern recognition The methods, algorithms and tools to analyze data based on either statistical information or on a priori knowledge extracted from the patterns. The patterns for classification are groups of observations, measurements, objects, defining feature vectors in an appropriate multidimensional feature space.

Data mining Algorithms, tools, methods and systems used in extraction of knowledge hidden in a large amount of data.

Features denoted f_i or F_j (i, j – feature indices) – a set of variables which carry discriminating and characterizing information about the objects under consideration. The features can represent raw measurements (data) f_i or can be generated in a non-linear way from the data F_j (features).

Feature space The multidimensional space in which the $\mathbf{F}_k$ vectors are defined. Data and feature vectors represent vectors in respective spaces.

Feature vector A collection of features ordered in some meaningful way into multi-dimensional feature vectors $\mathbf{F}_l$ ($\mathbf{F}_l$ where l – feature vector index) that represents the signature of the object to be identified represented by the generated features F_l.

Feature extraction The procedure of mapping source feature space into output feature space of lower dimensionality, retaining the minimal value of error cost function.

Multidimensional scaling The nonlinear procedure of feature extraction, which minimizes the value of the "stress" being the function of differences of all the distances between feature vectors in the source space and corresponding distances in the resulting space of lower dimensionality.

Data space The multi-dimensional space in which the data vectors $\boldsymbol{f}_k$ exist.

Data vector A collection of features ordered in some meaningful way into multi-dimensional vectors $\boldsymbol{f}_k$ ($\boldsymbol{f}_k$, k – data vector index) and $\boldsymbol{f}_k = [m_k, z_k, \boldsymbol{x}_k, t_k]$ where m_k is the magnitude and $\boldsymbol{x}_k$, z_k, t_k – its epicentral coordinates, depth and the time of occurrence, respectively.

Cluster Isolated set of feature (or data) vectors in data and feature spaces.

Clustering The computational procedure extracting clusters in multidimensional feature spaces.

Agglomerative (hierarchical) clustering algorithm The clustering algorithm in which at the start the feature vectors represent separate clusters and the larger clusters are built-up in a hierarchical way. The procedure repeats the process of gluing-up the closest clusters up to the stage when a desired number of clusters is achieved.

k-Means clustering Non-hierarchical clustering algorithm in which the randomly generated centers of clusters are improved iteratively.

Multi-resolutional clustering analysis Due to clustering a hierarchy of clusters can be obtained. The analysis of the results of clustering in various resolution levels allows for extraction of knowledge hidden in both local (small clusters) and global (large clusters) similarity of multidimensional feature vectors.

N-body solver The algorithm exploiting the concept of time evolution of an ensemble of mutually interacting particles.

Non-hierarchical clustering algorithm The clustering algorithm in which the clusters are searched for by using global optimization algorithms. The most representative algorithms of this type is **k-means** procedure.

Definition of the Subject

Earthquakes have a direct societal relevance because of their tremendous impact on human community [59]. The genesis of earthquakes is an unsolved problem in the earth sciences, because of the still unknown underlying physical mechanisms. Unlike the weather, which can be predicted for several days in advance by numerically integrating non-linear partial differential equations on massively parallel systems, earthquake forecasting remains an elusive goal, because of the lack of direct observations and the fact that the governing equations are still unknown. Instead one must employ statistical approaches (e. g., [61,72,82]) and data-assimilation techniques (e. g., [6,53,81]). The nature of the spatio-temporal evolution of earthquakes has

to be assessed from the observed seismicity and geodetic measurements. Problems of this nature can be analyzed by recognizing non-linear patterns hidden in the vast amount of seemingly unrelated information. With the proliferation of large-scale computations, data mining [77], which is a time-honored and well-understood process, has come into its own for extracting useful patterns from large incoherent data sets found in diverse fields, such as astronomy, medical imaging, combinatorial chemistry, bio-informatics, seismology, remote sensing and stock markets [75]. Recent advances in information technology, high performance computing, and satellite imagery have led to the availability of extremely large data sets, exceeding Terabytes at each turn, that are coming regularly to physical scientists who need to analyze them quickly. These data sets are non-trivial to analyze without the use of new computer science algorithms that find solutions with a minimal computing complexity. With the imminent arrival of petascale computing by 2011 in USA, we can expect some breakthrough results from clustering analysis. Indeed, clustering has become a widely successful approach for revealing features and patterns in the data-mining process. We describe the method of using clustering as a tool for analyzing complex seismic data sets and the visualization techniques necessary for interpreting the results. Petascale computing will also spur visualization techniques, which are sorely needed to understand the vast amounts of data compressed in many different kinds of spaces, with spatial, temporal and other types of dimensions [78]. Examples of clusters abound in nature include stars in galaxies, hubs in airline routes and centers of various human relationships [5]. Clustering comes from multi-scale, nonlinear interactions due to the rock rheology and earthquakes.

Introduction

Earthquake clustering is automatically implicated by the classical Gutenberg–Richter relationship [40], which specifies the frequency of earthquakes between some small magnitude cutoff and a certain large magnitude around 8 [39]. This empirical finding with a broad magnitude range implies that the largest seismic events are surrounded by a large number of smaller events. This clustering may have both spatial and temporal dependences. One of the goals of earthquake clustering studies is to find these special points in a high-dimensional space related to the nature of the dimensional space associated with earthquake dynamics [24,25]. One major goal of this chapter is to introduce the reader to the notion of searching for clustering points in dimensional spaces higher than the 3D physical space we are used to. This concept is crucial to our understanding of the clustering points of earthquakes in these higher-dimensional spaces, which may enable progress in forecasting earthquakes. Information in seismicity data sets can be both relevant and irrelevant from the point of view of deterministic earthquake dynamics. It can be also "entangled" and impossible to be interpreted with normal human perception. The role of data mining is to have a mathematically rigorous algorithm for extracting relevant information from this deluge of data, and make it understandable. Clustering techniques, which are commonly used today in many fields, ranging from biology (e. g., [26]) to astrophysics, allows us to produce specially crafted data models that can be employed for predicting the nature of future events. In more complex cases, these special data models can work in concert with formal mathematical and physical paradigms to give us deeper physical insight.

The concept of clustering has been used for many years in pattern recognition [2,50,78]. The clustering can use more (e. g. [54]) or less mathematically rigorous principles (e. g. [33]). Nowadays clustering and other feature extraction algorithms are recognized as important tools for revealing coherent features in the earth sciences [32,65,66,67], bioinformatics [51] and in data mining [37,43,44,57]. Depending on the data structures and goals of classification, different clustering schemes must be applied [36,55].

In this chapter we emphasize the role of clustering in the understanding of earthquake dynamics and the way to visualize and interpret the computed results from clustering. All the seismic events occurring over a certain region during a given time period can be viewed as a single cluster of correlated events. The strength of mutual correlations between events, such as correlations in spatial and time positions along with magnitude, cause this single cluster to have very complex internal structure. The correlations – the measures of similarity between events – divide the global cluster into variety of small clusters of multi-scale nature, i. e., small clusters may consist of a cascade of smaller ones. Coming down the scale we record clusters of more and more tightly correlated events. Exploring the nature of events belonging to a single cluster, we can extract common features they posses. Having more information about events belonging to the same cluster we can derive hidden dependences between them. Moreover, we can anticipate the type of an unknown event belonging to a certain cluster from the character of the other events of this cluster.

In the following sections we describe the idea of clustering and the new idea of higher dimensions associated with data sets. We also demonstrate the results of cluster-

ing analysis of both synthetic and real data. Long synthetic data were derived by using a model for a segmented strike-slip fault zone in a 3D elastic half-space [7]. The real data represent short time (5 years interval) seismic activities of the Changbaishan volcano (the north-east frontier of the North China craton) and the Japanese Archipelago. Lastly, we also highlight the role of visualization of clusters as an important tool for understanding this type of new data arrangement, and we describe the role played by remote visualization environment specially devised for visualization of earthquake clusters.

Earthquakes Clustering

Statistical Laws as Elementary Building Bricks of Earthquake Models

The earthquake prediction problem is of fundamental importance to society and also geosciences. Progress in this field is hampered, mainly because many important dynamic variables – such as stress – are not accessible for direct observations. Moreover, instrumental observations of seismicity are possible only for a fraction of a single large earthquake cycle. Overcoming these difficulties will require combining analyzes of model and observed data by using knowledge extraction instruments. The fundamental process of knowledge extraction is finding dependences between data and/or between model parameters. They can be revealed as patterns (clusters) in time, spatial and feature (parameter) space domains. The most elementary dependences can be expressed in the form of semi-empirical functional laws.

There are a few basic statistical laws which represent the basis for earthquake models development. The frequency-size statistics of regular tectonic earthquakes (excluding swarms and deep focus earthquakes) follow the Gutenberg–Richter relation [39,80,84]:

$$\log N(M) = a - bM \tag{1}$$

where N is the number of events with magnitude larger than M and a, b are constants giving, respectively, the overall seismicity rate and relative rates of events in different magnitude ranges. Observed b-values of regional seismicity typically fall in the range 0.7–1.3.

Aftershock decay rates are usually be described by the Omori–Utsu law [71,79]:

$$\Delta N/\Delta t = K(t + c)^{-p} \tag{2}$$

where N is the cumulative number of events, t is the time after the mainshock, and K, c, and p are empirical constants. The epidemic-type aftershock-sequences (ETAS) model combines the Omori–Utsu law with the Gutenberg–Richter frequency-magnitude relation for a history-dependent occurrence rate of a point process in the form (e. g., [61])

$$\lambda(t|H_t) = \mu + \sum_{t_i<t} \frac{K_0 \exp[\alpha(M_i - M_c)]}{(t - t_i + c)^p} \tag{3}$$

where α is a constant background rate, M_i is the magnitude of earthquake at time t_i, M_c is a lower magnitude cut-off, H_t denotes the history, and the productivity factor $K_0 \exp[\alpha(M_i - M_c)]$ gives the number of events triggered by a parent earthquake with magnitude M_i. The ETAS model is used widely in analysis of seismic data, owing to its built-in clustering associated with the incorporation of the Gutenberg–Richter and Omori–Utsu laws. Examples of recent applications can be found in [45,62,68].

These results can be used to derive additional properties such as average recurrence times (e. g., [4,18,19,20,68,89]). It is usually defined as the number of years between occurrences of an earthquake of a given magnitude in a particular area. For example, the probability of a devastating earthquake striking the greater San Francisco Bay Region over the following 25 years (2007–2031) is 0.62 [68]. Corral [18,19,20] proposed the existence of a universal scaling law for the probability density function $H(\tau)$ of recurrence times (or interevent times) τ between earthquakes in a given region:

$$H(\tau) \cong \lambda \times f(\lambda\tau) \ . \tag{4}$$

The function $f(x)$ appears to be similar for many different seismic regions, which suggests some universal properties. The average rate λ represent the region specific constant, whose reciprocal is the only relevant characteristic time for the recurrence times. Molchan [58] showed that under general conditions, the only universal distribution of inter-event times in a stationary point process is exponential. Hainzl et al. [42] and Saichev and Sornette [68] discussed relations between statistics of interevent times, the ETAS model of triggered seismicity, and the Corral [18,19] distribution of Eq. (4).

In the context of earthquake prediction it is important to analyze earthquake cycles with repeating sequences of events such as foreshocks, mainshocks and aftershocks (e. g., [9,74,80]). Apart from qualitative tendencies reflected by statistical laws, the earthquakes exhibit various types of more subtle spatio-temporal clustering, i. e., grouping of events of the same type both in time and in spatial coordinates. The recognition of these patterns followed by the analysis of the reasons of their appearance may lead to the development of improved prediction algorithms.

In the following section we present a closer look of clustering as a knowledge extraction technique and a possible way of its application to earthquake data analysis.

Basic Concepts of Clustering

Clustering analysis is a mathematical concept whose main useful role is to extract the most similar (or dissimilar) separated sets of objects according to a given similarity (or dissimilarity) measure [2]. Clustering is one of the most fundamental processes generated by nature. For example, people gathering in groups, tribes, demonstrations, parties, cities, produce clusters. Similarly, towns and cities are clusters of buildings while galaxies are clusters of stars. The local computer networks and bacterial colonies are also clusters. The objects forming clusters can be the clusters of smaller objects, which in turn, are clusters of even smaller and smaller building bricks. The complexity of cluster structure reflects the complexity of the real world. The clusters of various shapes, densities and sizes, with additional attributes as colors, transparency etc. built up patterns, which are the fingerprints of all multi-scale processes and phenomena. The clusters are the primitives of the patterns.

The same notion of clustering concerns geographical locations and other properties of earthquakes. In Fig. 1a we present a spatial distribution of earthquake epicenters in the western hemisphere of the Earth (data from http://quake.geo.berkeley.edu/cnss/maps/cnss-map.html). One can see with the naked eye that their distribution is far from being uniform. We observe both elongated and oblate structures – the earthquake clusters – separated at this resolution by large holes of seismically quiescent area.

Properties of the clusters result from properties of the generating processes. The shape and structure of clusters are visual representation of information on these processes. Therefore, detection of clusters and their analysis is the first step for knowledge extraction from this information. For example, as shown in Fig. 1a, the earthquake clusters on Earth are located in geologically active regions, mainly, on the edges of colliding tectonic plates. The distribution and shape of the earthquake clusters follow the borders between the plates. In Fig. 1b we show the large earthquake cluster from Fig. 1a located at the US western coast. One can distinguish here many smaller clusters of different density separated by geologically inactive area. A similar pattern (see Fig. 1c) is observed by zooming-in one of denser clusters from Fig. 1b. This multi-resolutional and self-similar system is characteristic for many critical phenomena ► Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of [3,9,16,74]. The worldwide fault network has a fractal structure (or multifractal) [22,27,79]. Wavelet-based multi-fractal anal-

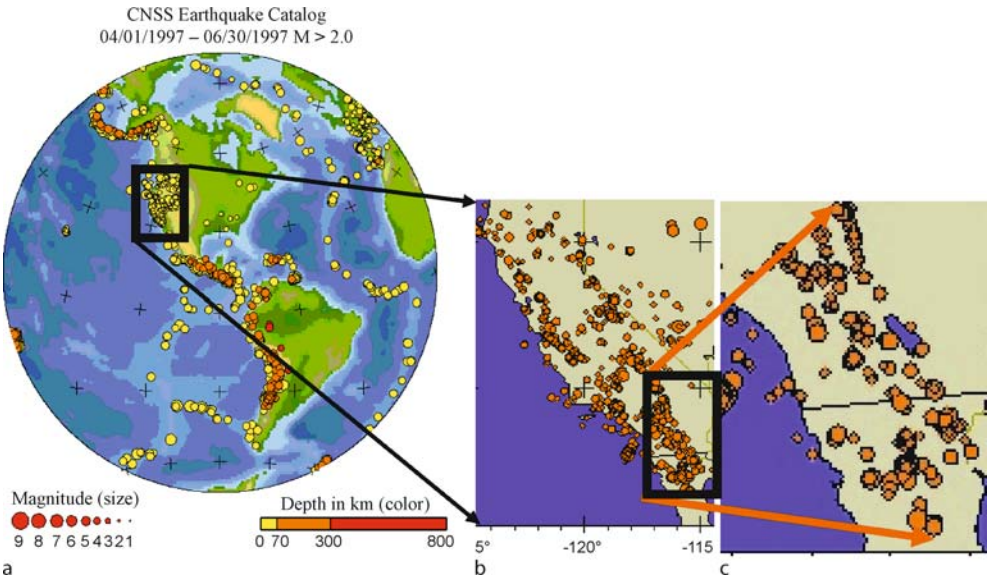

Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 1
Multiscale character of the earthquake clusters. The epicenters of earthquakes of various depth and magnitude are displayed. The data come from the CNSS Earthquake Catalog (http://quake.geo.berkeley.edu/cnss/maps/cnss-map.html). a the western hemisphere, b the US western coast c California and Nevada

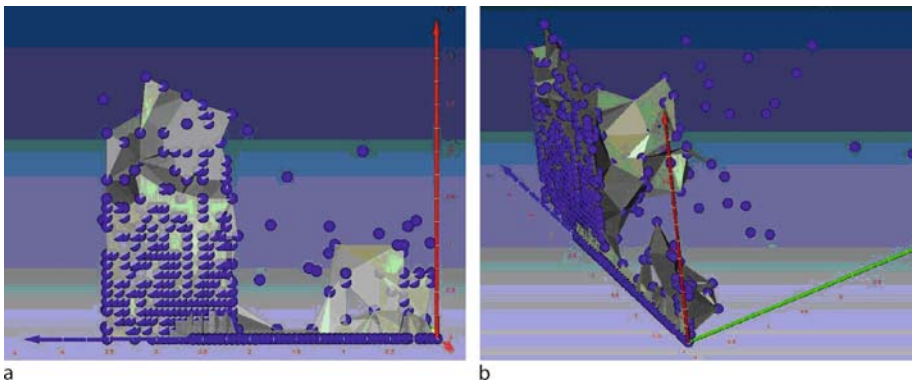

Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 2
Seismic activity of the Changbaishan volcano during 5 years time span from 07.1999 to 05.2004 (the north-east frontier of the North China craton) [47]. The plates represent the seismic events in 3-D feature space attributed by eruption time (*blue axis*), magnitude (*red axis*) and distance to the epicenter (*green axis*) coordinates. The clusters are rendered using the *wrap point* technique (the Amira visualization package www.amiravis.com). Two different positions of coordinates are shown. The *large cluster* representing the earthquake swarm is preceded by the *small precursory cluster* of seismic activity and quiescent time period

ysis [27] shows clearly several distinct scaling domains in earthquake catalogs revealing rich self-similar multiscale structure. However, the spatial structure of earthquake clusters alone is inadequate to formulate plausible hypotheses about earthquake dynamics. More information is required.

As shown in Fig. 1, besides the geographical location, earthquakes have additional features such as the time and depth of occurrence and the amount of energy released (proportional to $10^{\alpha m}$ with $\alpha \sim 1.5$ and m the magnitude). These attributes can be used as additional coordinates of, so called, feature space (e. g., [78]). In Fig. 2 we display the earthquake clusters representing the seismic activity nearby the Changbaishan volcano in an abstract 3-D feature space. Apart from geographical location – represented by the distance from the epicenter – other coordinates (features) are employed: the time of occurrence and the magnitude of the earthquake. As shown in Fig. 2, the large cluster of seismic activity is preceded by the small precursory cluster and low activity region. The larger cluster is characterized by the seismic events from broader interval of magnitudes and with satellite earthquakes more distant from the epicenter than in the preceding smaller cluster.

The dynamics of the volcanic earthquakes covers only a period of 5 years. The time is too short to conclude about the long-time earthquake dynamics. To obtain data covering much longer time period we used synthetic data generated by numerical simulations of seismicity on a heterogeneous fault governed by 3-D elastic dislocation theory, power-law creep and boundary conditions corresponding to the central San Andreas Fault [7,28,29]. In Fig. 3 we represent seismic activity during 150 years. This period contains $M_f \sim 1 - 3 \times 10^4$ events (represented in Fig. 3 by colored dots) in the magnitude interval [3.3–6.8]. Unlike in the Changbaishan case, the seismic events have one more feature – the earthquake depth. Thus the feature space has now four dimensions. In Fig. 3 we display the data distribution in time-depth-position 3-D space. The fourth dimension – the magnitude – is displayed in Fig. 3 by the size of dot. To make the situation clearer only the large earthquakes with magnitudes $m > 6$ (large dots) and the smallest ones $m < 4$ (small dots) are distinguished in Fig. 3. As shown in Fig. 3 and in [24], the synthetic seismic events with magnitudes $m < 4$ produce stripe-like clusters in the data space. They precede large earthquakes ($m > 6$) and are separated in time by the regions of mixed type of events (i. e., with $4 < m < 5$).

Another system of earthquake clusters are shown in Fig. 4 The synthetic data ($M_f \sim 10^5$ events) corresponding to the seismic activity during 1500 years were generated by the same model [7] for similar geological and boundary conditions. Only medium size events with $4.5 < m < 6$ were taken for clustering. In addition to the local strip like clusters of smaller events ($m < 4$) detected for 150-years data, one can observe in Fig. 4a distinct spatio-temporal patchwork structure of clusters of medium sized events ($4.5 < m < 6$). These clusters follow spatio-temporal changes in strength-stress properties of the fault in the region simulated.

In summary, we can highlight very fundamental properties of earthquakes, multi-resolutional clusters are built up by the earthquake epicenters. The clustering is a dynamical process involving many spatio-temporal scales.

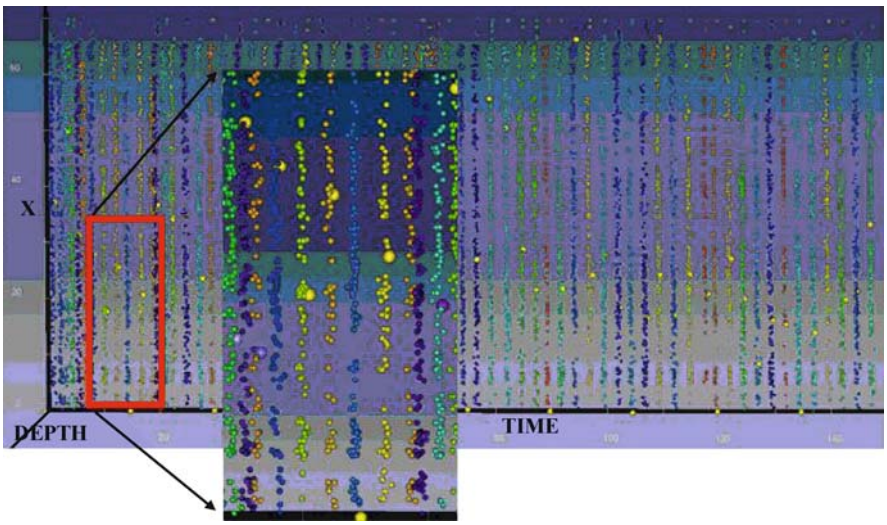

Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 3
The plot reconstructing seismic activity during 150 years from synthetic data [7] (horizontal distance – X, depth – z; visualized by using the Amira visualization package [1]). Large events (with magnitude $m > 6$) are shown as distinctly larger dots on the background of the lowest magnitude events ($m < 4$). There are visualized patches of low magnitude events preceding larger events [24]. The separate clusters are marked in colors

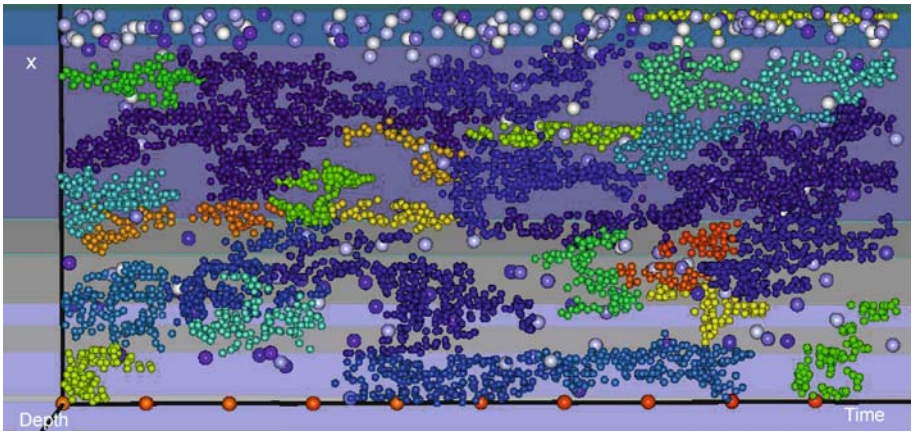

Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 4
The plot reconstructing seismic activity during 1500 years from synthetic data [7]. The largest clusters obtained for events with magnitudes $4.5 < m < 6$. Large events ($m > 6$) are shown as distinctly larger plates. The separate clusters are marked in colors

The dynamic nature of earthquake clusters in a very long time horizon is obvious because everyone can expect that tectonic plates will change dynamically the geo-mechanical properties of the Earth crust. In a long time period covering thousands years, the patterns from Fig. 1 will evolve following the changes in the fault network. More mysterious is the character of earthquake dynamics in spatio-temporal scales allowing for making realistic predictions. We show that in the medium-time period lasting more than a hundred of years the seismic events may produce periodic system of clusters in approximately equal time intervals with increasing and decreasing seismic activity. The large earthquakes, preceded by the quiescent time periods, appear. The short-time dynamics reveal additionally, that the earthquake swarms are signaled by the smaller precursory cluster of seismic activity. The earthquake attributes such as the magnitude, and the epicenter depth, allow for better interpretation of emerging clusters and exploration of hypotheses space. Therefore, for studying various aspects of earthquake dynamics, in-

cluding their prediction, we have to analyze the cluster structures in multi-dimensional feature space, to be sure that none of important information will be lost or neglected.

Multidimensional Feature Space

In Fig. 1 every point i representing one out of M_f earthquakes has two dimensions – the geographical coordinates $\boldsymbol{x}_i = [x_1, x_2]$ of the epicenter. The point can be treated as 2-D vector $\boldsymbol{f}_i$ in the feature space where $\boldsymbol{f}_i = \boldsymbol{x}_i$. Assuming additional coordinates, at the highest level of resolution, a single seismic event i can be represented as a five-dimensional data vector $\boldsymbol{f}_i = [m_i, z_i, \boldsymbol{x}_i, t_i]$ where m_i is the magnitude and $\boldsymbol{x}_i, z_i, t_i$ – its epicentral coordinates, depth and the time of occurrence, respectively. The spatio-temporal clusters can be extracted by 3-D visualization similar as those of Figs. 3, 4 distinguishing extra dimension by the size of dots and colors. Only clusters in the three spatially visualized dimensions can be extracted, while the other attributes associated with the earthquake characteristics are used for discriminating among the different types of clusters.

As shown in Figs. 3, 4 and in [24], at the lowest resolution level we can analyze the data locally by looking for clusters with similar events. However, considering a single event on a given area as a feature vector [78] cannot be a good approach from a generalization point of view. The number of events is usually large. There are many noisy background events, which destroy the relevant clusters or produce artificial ones. Moreover, the clustering of raw data neglects the important statistical information, which concerns the entire inspected area. An alternative approach exists in which the entire seismic area can be described as a multidimensional feature vector evolving in time. In the following these features will represent descriptors a_k (seismicity parameters) corresponding to different statistical properties of all the events measured in a given time interval. The number of descriptors N defines the dimensionality of the feature vector $\boldsymbol{F}_i = [a_1, a_2, \ldots, a_N]$, $i = 1, 2, \ldots, M$. The vector represents not a single seismic event but it corresponds to seismic situation on the whole controlled area in the subsequent time interval indexed by i. The number of feature vectors M is equal to the number of time intervals in which the descriptors are computed. The index i is a discrete equivalent of time. We expect that the features vectors representing different moments of time also have the tendency to produce clusters in the abstract N-dimensional feature space. Monitoring changes of these time-series in abstract N-dimensional space may be used as a proxy for the evolution of stress and a large earthquake cycle on a heterogeneous fault [9].

To explain this approach better, let us assume that we have to analyze the client behaviors in a hypermarket. We can watch every client separately assuming that it can be defined as a feature vector consisting of only two coordinates: the time he entered the shop, money spent. Then we can try to find clusters emerging with time during a shopping day. This cannot be easy due to both a large number of feature vectors (clients) producing statistical noise and lack of correlations between them. Another approach consists in treating as a feature vector not a single client but every subsequent time interval $t_i = i \times \Delta T (i = 0, 1, 2, \ldots, M_F;\ t_i < t_e;\ M_F = (t_e - t_b)/\Delta T$ and t_b – beginning of the working day and t_e – closing time). Let the coordinates of the subsequent feature vector define the following descriptors averaged in Δt: the number of people inside the shop (crowding), the flow, items bought, money spent per person. We note that now the number of feature vectors will be substantially smaller than in the previous approach but the dimensionality of feature space is larger. Let us assume that as a result of clustering we extract two distinct clusters. The first one consists of feature vectors (time intervals) from between 10.00–11.00 and 13.00–14.00. The cluster is characterized by very small values of the first three descriptors (crowding, flow, numbers of items sold) and relatively large expenses. The second cluster consists of time intervals from between 8.00–9.00 and 16.00–18.00 with all descriptors large. We could conclude that the first cluster consists of time intervals from shopping hours that are the favorite for wealthy retired people from the rich village in the neighborhood, while the second cluster is associated with the rush in shopping just before before the beginning and after the end of working hours.

In the same way, the clusters of feature vectors (time intervals) consisting of seismicity parameters should reflect the similarity between seismic activities in various time intervals. As we show before the large seismic events are preceded by precursory events, reflected by an abnormal seismic activity in the whole area. We suppose that these moments of time are similar in the context of the set of seismicity parameters selected. Thus the feature vectors corresponding to precursory events should belong to the same cluster. The idea of predictive system based on clustering consists in detecting the clusters of former foreshocks and signal if the current feature vector – which represents current seismic situation over the area – is or is not the member of this clusters.

The seismicity parameters are computed as time and space averages in a given time and space intervals within

Earthquake Clusters over Multi-dimensional Space, Visualization of, Table 1
Definition of seismicity parameters

NS	Degree of spatial non-randomness at short distances. The differences between distributions of event distances and distances between randomly distributed points.
NL	Degree of spatial non-randomness at short distances.
CD	Spatial correlation dimension calculated on the basis of correlation integrals and on interevent distances.
SR	Degree of spatial repetitiveness represents the tendency of events with similar magnitudes to have nearly the same locations of hypocenters.
AZ	Average depth of the earthquake occurrence.
TI	Inverse of seismicity rate – time interval in which a given (constant) number of events occurs.
MR	Ratio of the numbers of events falling into two different magnitude ranges $= M_f(m \geq M_0)/M_f(m < M_0)$.

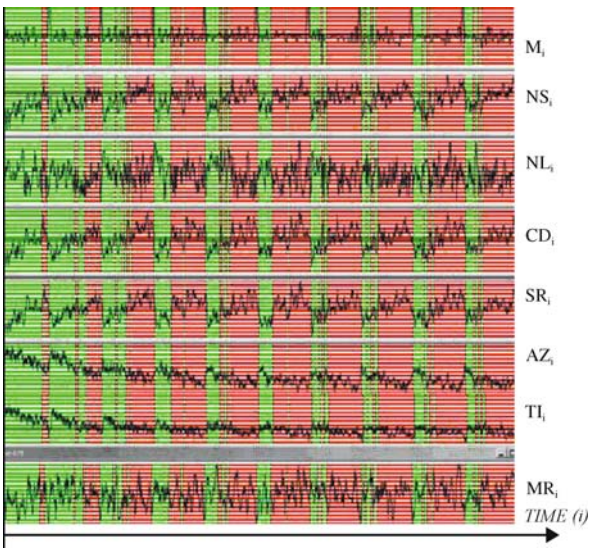

Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 5
The exemplary set of seismicity parameters {M, *NS, NL, CD, SR, AZ, TI, MR*} in time (*i* – subsequent number of the feature vector) for a file from the 1500-years synthetic data catalog (from [25]). The *green* and *red strips* show the time moments belonging to the two different clusters. The *green cluster* corresponds to the time intervals of lower while the *red cluster* of higher seismic activities. The time series represent about $M_F = 10^3$ feature vectors F_i

a sliding time window with a length ΔT and time step $\mathrm{d}t$. The values of a_k represents one of the following seismicity parameters: **NS, NL, CD, SR, AZ, TI, MR**. The value of $\mathrm{d}t$ was assumed to be equal to the average time difference between two recorded consecutive events while ΔT is equal to about 1/10 of the average time distance between two successive large events ($m > 6$ or $m > 5$). By increasing the values of $\mathrm{d}t$ and ΔT one can obtain smoother time series due to better statistics. On the other hand, poorer prediction characteristics can be expected then. We define the seismicity parameters as shown in Table 1 [28,29].

The seismicity parameters produces seven time series and create the abstract 7-dimensional feature space of time events $\boldsymbol{F}_i = (NS_i, NL_i, CD_i, SR_i, AZ_i, TI_i, MR_i)$ where i are discretized values of time $t = t_b + i\Delta T$. In Fig. 5 we display an example set of seismicity parameters (with average magnitude **M**) for synthetic data [25]. The precise location of the clusters and the visualization of the clustering results are significant challenges in clustering over multi-dimensional space. In the following section we present briefly the basics of clustering and algorithms needed in this venture.

The Detection and Visualization of Clusters in Multi-Dimensional Feature Space

Our main challenge is to devise a clustering scheme which can divide the M feature vectors $\boldsymbol{x}_i$ $i = 1, 2, \ldots M$ into k separate groups (clusters). More formally, assuming that $\mathrm{X} = \{\boldsymbol{x}_i\}_{i=1,\ldots,M}$ and $\boldsymbol{x} \in \mathbf{R}^N$; $\boldsymbol{x}_i = \{x_{i1}, x_{i2}, \ldots, x_{iN}\}$ we define as an *k-clustering* of X, i. e., the partition of X into k clusters $C_1, \ldots, C_k$ provided three conditions are met:

- $C_i \neq 0, i = 1, \ldots, k$ – the clusters are non empty sets,
- $\cup_{i=1,\ldots,m} C_i = \mathrm{X}$ – the sum of elements inside clusters is equal to the total number of feature vectors,
- $C_i \cap C_j = 0, i \neq j, j = 1, \ldots, k$ – each feature vector belongs to only one cluster.

The computational problem with clustering is that the number of possible clustering of M vectors into k groups is given by the Stirling numbers (very large numbers) of the second kind:

$$S(M, k) = \frac{1}{k!} \sum_{i=1}^{k} (-1)^{k-1} \binom{k}{i} \cdot i^M . \tag{5}$$

Some values of $S(N, k)$ are: $S(15, 3) \approx 2 \times 10^6$, $S(20, 4) \approx 45 \times 10^9$, $S(25, 8) \approx 7 \times 10^{17}$, $S(100, 5) \approx 2 \times 10^{68}$. Knowing that the value of N in typical clustering problems can be 10^2 to 10^9 and more we see that the clustering problem is intrinsically hard and exhaustive search – looking through all possible clusterings – cannot be considered. The special clustering schemes based on the proximity measures between feature vectors have to be exploited. The basic steps must be followed in order to develop a clustering task are the following:

1. *Feature selection* – Features must be properly selected to encode as much information as possible. Parsimony

and minimum redundancy among the features is a major goal.
2. *Proximity measure* – This is the measure how "similar" (or "dissimilar") two features vectors are.
3. *Clustering criterion*, which depends on the interpretation of the term "sensible", depending on the type of clusters expected in the data set e. g., oblate, elongated, "bridged", circular etc.
4. *Clustering algorithms*. Choose a specific algorithmic scheme that unravels the clustering structure of the data set.
5. *Validation and interpretation of results* are the final processes of clustering.

There are two principal types of clustering algorithms: non-hierarchical and agglomerative schemes [2,50,78].

Clustering Techniques

The non-hierarchical clustering algorithms are used mainly for extracting compact clusters by using global knowledge about the data structure. The well known *k-means* based schemes [78], consist in finding the global minimum of the following goal function:

$$J(w,z) = \sum_j \sum_{i \in C_j} \left| x_i - z_j \right|^2 , \tag{6}$$

where: $\mathbf{z}_j$ is the position of the center of mass of the cluster j, while $\boldsymbol{x}_i$ are the feature vectors closest to $\mathbf{z}_j$. To find a global minimum of function $J()$, one repeats many times the clustering procedures for different initial conditions [48]. Each new initial configuration is constructed in a special way from the previous results by using the methods from [48,87]. The cluster structure with the lowest $J(w,z)$ minimum is selected.

Agglomerative clustering schemes consist in the subsequent merging of smaller clusters into the larger clusters, basing on proximity and clustering criteria. Depending on the definition of these criteria, there exist many agglomerative schemes such as: average link, complete link, centroid, median, minimum variance and nearest neighbor algorithm. The hierarchical schemes are very fast for extracting localized clusters with non-spherical shapes. The proper choice of proximity and clustering criteria depend on many aspects such as dimensionality of data. For example, a smart clustering criterion based on linked-list scheme for finding neighbors used for molecules clustering is completely worthless for clustering N-dimensional data for which it has extremely high computational complexity. All of agglomerative algorithms suffer from the problem of not having properly defined control parameters, which can be matched for the data of interest and hence can be regarded as invariants for other similar data structures.

Majority of the classical clustering algorithms require knowledge on the number of clusters. However this number is usually unknown a priori. Furthermore, these methods do not perform well in the presence of heavy noise or outliers. Recently, new methods have been proposed that can: deal with noisy data, discover non-spherical clusters and allow for automatic assessment of number of clusters. Some important examples are the Chameleon [55], DBSCAN [70] and CURE [38] algorithms. Unfortunately, these methods are suited only for low dimensional data and are rather inefficient limiting their use for data mining of large-scale sets. For clustering of large data sets of multidimensional data other approaches are in great demand. In the innovative work by Frey and Dueck [33] the authors use the concept of "affinity propagation," which takes as input measures of similarity between pairs of data points. Real-valued messages are exchanged between data points until a high-quality set of exemplars and corresponding clusters gradually emerges. Affinity propagation promises to find clusters with much lower error than other methods, and it can do this in less than one-hundredth the amount of time.

Clustering schemes do not produce univocal results. For low dimensional 2-3-D spaces human eye can decide whether the clustering result is optimal or not. However, it becomes hopeless for higher dimensions. There exist many techniques for visualization multidimensional clusters. One of them is the multi-dimensional scaling (MDS) (see overview of mapping techniques in [73]) – the most powerful non-linear mapping technique. This method allows for visualization of the multidimensional data in 2-D or 3-D and for interactive extraction of clusters.

Multidimensional Scaling

Multi-dimensional scaling or MDS is mathematically a non-linear transformation of N-dimensional data onto n-dimensional space, where $n \geq N$ [23,73,78]. The MDS algorithm bases on the "stress function" criterion. The goal is to maintain all the distances between points $\mathbf{R}_i \in \omega \subset \Re^N$ in the Euclidean 3-D (or 2-D) space with a minimum error. The "stress function" can be written as follows:

$$E\left(\omega, \omega'\right) = \sum_{j<i} s_{i,j}^{w \cdot mi} \cdot \left(s_{i,j} - s'_{i,j}\right)^{mi} = \min$$
$$\text{where: } s'_{i,j} = \left(\boldsymbol{y}_i - \boldsymbol{y}_j\right) \cdot \left(\boldsymbol{y}_i - \boldsymbol{y}_j\right), i, j = 1, \ldots, M, \tag{7}$$

and $D_{i,j}$ – is a squared distance between points $\mathbf{R}_i$, $\mathbf{R}_j \in \omega \subset \Re^N$ and $\mathbf{r}_i$, $\mathbf{r}_j \in \omega' \subset \mathrm{E}^3$ – coordinates of the respective points in 3-D Euclidean space. The values of w and mi are the parameters of transformation.

The result of mapping depends on the quality of the minimum obtained for the "stress function". Usually the dimensionality of the "stress function" domain is very high and is equal to $N \cdot M$, i. e., thousands, in the smallest and billions in large problems. For more than $M = 10^3$ feature vectors, the high dimensionality of source space and data complexity may cause the resulting low dimensional patterns to be completely illegible. The application of standard numerical algorithms for finding global minimum of this multimodal, non-linear and complex criterion becomes hopeless. Therefore, for visualization of $M > 10^3$ multidimensional data samples, more reliable minimization techniques extracting global minimum of the "stress function" are required. In [23] we proposed N-body solver by ODE's as a heuristic means. The algorithm is as follows:

1. The initial configuration of M interacting "particles" is generated in $\mathbf{E}^3$,
2. Every "particle" corresponds to the respective N-dimensional point from $\Re^N$,
3. The "particles" interact with each other with $\Phi_{i,j}$ particle-particle potential:

$$V_{i,j} = \frac{1}{4} \cdot k \left(r_{i,j}^2 - a_{i,j}^2 \right)^2 \tag{8}$$

(k – is the stiffness factor) and the energy produced is dissipated by the friction force proportional to the velocity of the particles.
4. The system of particles evolves according to the Newtonian equations of motion.

In this way the interactions between each pair of particles are described by various spring like potentials, dependent on the separation distance between particles r_{ij} and the distance D_{ij} between respective multidimensional points in $\Re^N$. If the distance between particles i and j in the output 2(3)-D space is smaller than the distance between respective i and j feature vectors in the source N-D space these points repel one another. Otherwise, i. e. the distance is larger, the particles attract one another. By using the *leap-frog* numerical scheme for time-integration [41] the following formula for velocities and positions of "particles" can be derived from the Newtonian equations:

$$\begin{aligned} \boldsymbol{v}_i^{n+1/2} &= \frac{(1-\varphi)}{(1+\varphi)} \cdot \boldsymbol{v}_i^{n-1/2} \\ &\quad + \frac{\alpha \Delta t}{(1+\varphi)} \left\{ \sum_{j=1}^{K} (r_{i,j}^n 2 - a_{i,j}^2) \boldsymbol{r}_{i,j}^n + \frac{g}{\alpha} \boldsymbol{i}_z \right\} \\ \boldsymbol{r}_i^{n+1} &= \boldsymbol{r}_i^n + \boldsymbol{v}_i^{n+1/2} \cdot \Delta t \\ \alpha &= \frac{k}{m}, \quad \varphi = \frac{\lambda}{2m} \cdot \Delta t, \end{aligned} \tag{9}$$

where $\mathbf{v}_i^n$, — the particle i, n – the time-step number, $m = 1$ – particle mass.

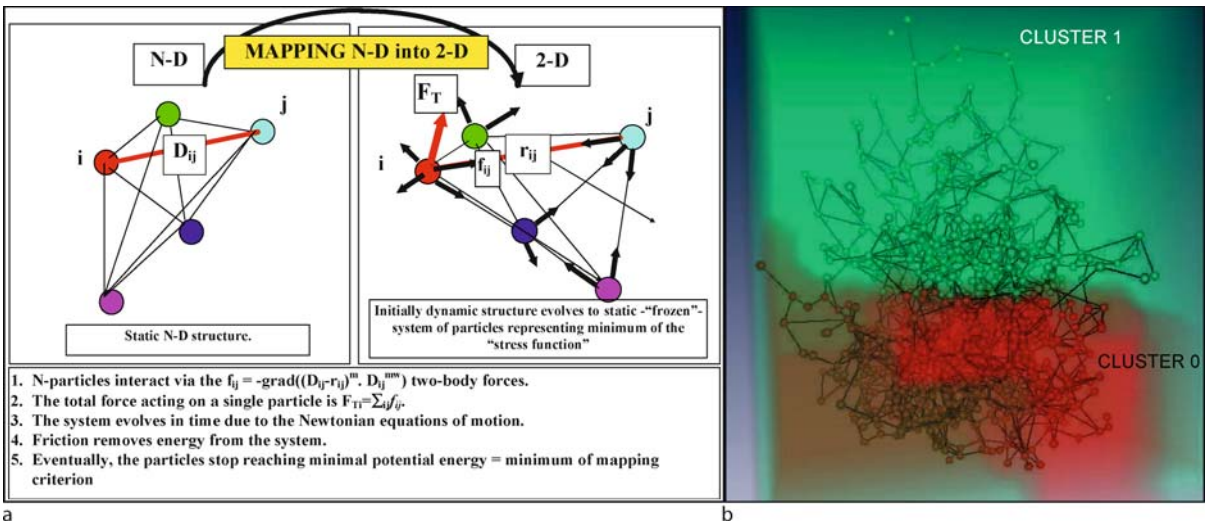

Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 6
a The conceptual diagram of MDS transformation, b The clusters from Fig. 6 mapped by using multidimensional scaling into 3D space for synthetic seismic data catalog A covering 1500 years

As it is common in molecular dynamics [41], the system of "particles" evolves in time until the global (or close to the global) minimum of Eq. (8) (the total potential energy of the particle system) is gained. Two free parameters, λ and k, have to be fit to obtain the stable state, where the final positions of frozen "particles" reflect the result of N–D to 3-D mapping. The conceptual scheme of MDS exploiting N-body solver is shown in Fig. 6A. In Fig. 6B we present the feature vectors shown in Fig. 5, using the 7-dimensional feature space which has been transformed by using the MDS procedure and mapping onto the 3-D space. Take a look on the movies (Movie 1 and 2 in Supplementary Materials), which shows how rotation in the 3-D space can help in cluster recognition.

Description of the Data

Natural Datasets

We analyze the observed and synthetic earthquake catalogs for three time intervals of 5, 150 and 1500 years respectively. The observed data (Fig. 7) represents seismic activities of the Japanese islands collected by the Japan Meteorological Agency (JMA).

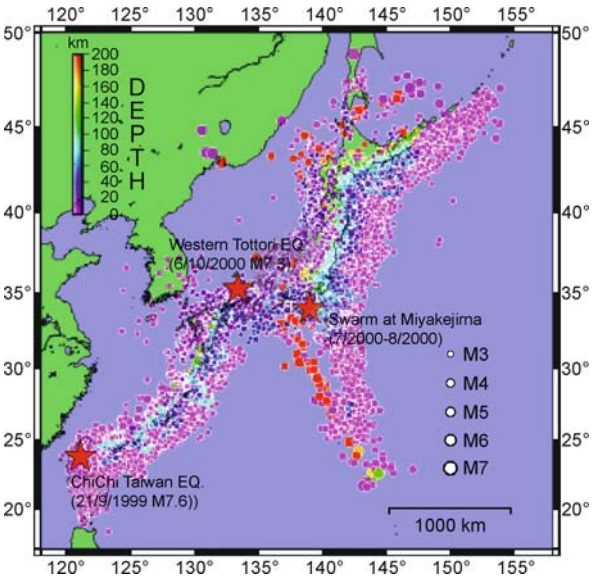

Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 7
Seismic activities around the Japanese Archipelago with a time period of 5 years. We use the hypocentral data provided by the Japan Meteorological Agency (JMA). The magnitude of the earthquakes (JMA magnitude) and their depths are represented by differences of the radius of the circle and colors, respectively. The *red stars* symbolize large events such as: Chi-Chi Taiwan earthquake (21/9/1999 M7.6 latitude 23.8 longitude 121.1), Swarm at Miyakejima (7/2000-8/2000 latitude 34.0 longitude 139.0), Western Tottori earthquake (6/10/2000 M7.3 latitude 35.3 longitude 133.4) (from [25])

The JMA Catalog consists of 915,829 events detected in Japan Islands between 1923 and January 31, 2003. The original catalog includes also events with magnitudes less then 1.0. The lowest magnitudes were determined by using a detection level, estimated from the Gutenberg–Richter frequency-size distribution. We have assumed that the cutoff magnitude of earthquake is equal to 3 ($m > 3$). We do not use any cutoff depth of hypocenter events. The seismic events shown in Fig. 7, were recorded during the 5 years time interval from October 1, 1997 to January 31, 2003. The data set processed consists of $M = 42\,370$ seismic events with magnitudes m, position in space (latitude X, longitude Y, depth z) and occurrence time t.

To analyze the seismic activity in longer time periods, we use data from synthetic catalogs generated by numerical earthquake models [7].

Physical Model of Earthquake Dynamics

The synthetic catalogs are generated by the model of Ben-Zion [7] for a segmented strike-slip fault zone in a 3D elastic half-space, based on earlier developments of Ben-Zion and Rice [10,11]. The model attempts to account for statistical properties of earthquake ruptures on long and narrow fault zones with bends, offsets, etc (Fig. 8a), represented by a cellular structure in a 2D plane with discrete cells and spatial variations of frictional parameters (Fig. 8b). The model contains a computational grid (region II of Fig. 8b) where evolving stress and seismicity are generated in response to ongoing loading imposed as slip boundary conditions on the other fault regions. Regions III and V creep at constant plate velocity of 35 mm/yr, while regions I and IV follow staircase slip histories with recurrence times of 150 yr. The stress transfer due to the imposed boundary conditions and failing grid cells is calculated by using a discretized form of a boundary integral equation and employing the static solution for dislocations in a 3D elastic half-space [10,63].

Deformation at each computational cell is the sum of slip contributions from brittle and creep processes. The brittle process (Fig. 8c) is governed by distributions of static friction τ_s, dynamic friction τ_d, and arrest stress τ_a. The static friction characterizes the brittle strength of a cell until its initial failure in a given model earthquake. When stress τ at a cell reaches the static friction, the strength drops to the dynamic friction for the remaining duration of the event. The stress at a failing cell drops to the arrest level τ_a, which may be lower than τ_d to ac-

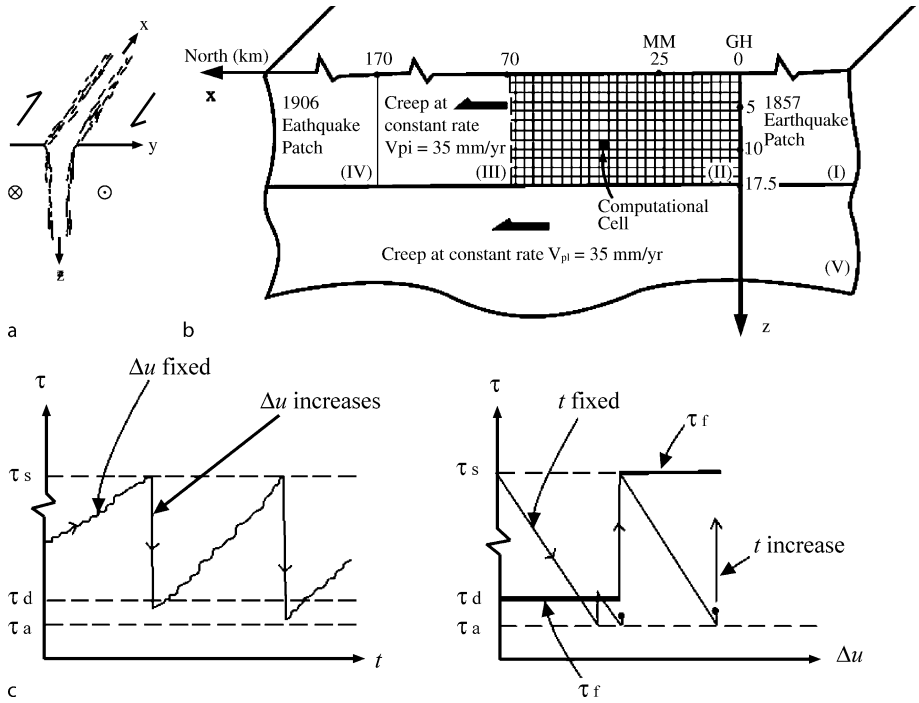

Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 8
The schematics of the model of Ben-Zion [7] for a segmented strike-slip fault zone in a 3D elastic half-space

commodate dynamic overshoot, producing local slip governed by dislocation theory [17,63]. The static friction, dynamic friction, and arrest stress are connected via a dynamic overshoot coefficient $D = (\tau_s - \tau_a)/(\tau_s - \tau_d)$. If the stress transfer from failing regions increases the stress at other cells to their static or dynamic strength thresholds, as appropriate, these cells fail and the event grows. When the stress at all cells is below the brittle failure thresholds, the model earthquake ends and the strength at all failing cells recovers back to τ_s. The creep process is governed by a power-law dependence of creep-velocity on the local stress and space-dependent coefficients that increase exponentially with depth and with distance from the southern edge of the computational grid. The chosen parameters produce an overall "pine-tree" stress-depth profile with a "brittle-ductile" transition at a depth of about 12.5 km, and variable stress-along-strike profiles with a gradual "brittle-creep" transition near the boundary between regions II and III (see Ben-Zion [7] for additional details). The model generates many realistic features of seismicity compatible with observations, including frequency-size and temporal event statistics, hypocenter distribution with depth and along strike, intermittent criticality, accelerated seismic release, scaling of source time functions and more (e. g., [9,29,56,88]).

Synthetic Catalogs

Synthetic data generated by computational models can comprise many events covering large spatial areas and extremely long time spans. Moreover, the synthetic data retain the statistical reliability of the results. The data are free of measurement errors, which occur in estimating earthquake magnitudes and hypocentral locations, and do not suffer from incomplete recording of small events, which exist in natural catalogs. These are significant advantages for our study, which attempts to illustrate clearly the performance of clustering analysis and visualization techniques.

In Sect. "Description of the Data" we analyze synthetic catalogs generated by two model realizations (A and M) of Ben-Zion [7]. The catalogs contain the time, location and magnitude of earthquakes calculated by the model for 150 and 1500 years. Extensive numerical simulations with several different classes of models, summarized by Ben-Zion [8] and Zöller et al. [9], suggest that the degree of disorder in fault heterogeneities is a tuning parameter of the earthquake dynamics. Catalog A is generated by a model realization tailored to the Parkfield section of the San Andreas fault. Catalog M is generated by a realization of a more-disordered system like the San Jac-

into fault or the Eastern California Shear Zone in Southern California. In both data sets the time interval covers all events ($M \sim 1 - 3 \times 10^4$) that have occurred in the last 150 years of simulated fault activity. These simulations were repeated for ten times larger time scale i. e. 1500 year interval (the number of events $M \sim 10^5$) covering hundreds of large earthquakes ($m > 6$) and correspondingly wider time window.

The seismicity parameters were obtained by averaging the data using a sliding time window of constant width ΔT and shift $\mathrm{d}t$. We employ $\Delta T = 10$ days and $\mathrm{d}t = 2$ days for the Japanese data, $\Delta T = 10$ months and $\mathrm{d}t = 2$ months for the 150-years synthetic data and $\Delta T = 30$ months and $\mathrm{d}t = 6$ months for the data covering 1500 years time period. Each parameter in the clustering was normalized with respect to the standard deviation.

Earthquake Visualization by Using Clustering in Feature Space

Short-Time Period

Results of clustering of the observed Japanese seismic catalogs (see Fig. 7) both in raw data and in feature spaces are shown in Fig. 9. At the data resolution level a single seismic event i can be represented as a multi-dimensional data vector $\boldsymbol{f}_i = [m_i, z_i, X_i, Y_i, t_i]$ where: m_i is the magnitude, X_i – the latitude, Y_i – the longitude, z_i and t_i – the depth and the time of occurrence, respectively. The seismic events are visualized with the Amira package in Fig. 9a,b as irregular clouds of colored dots with (z, x, t) coordinates.

In accordance with the Gutenberg–Richter relationship, we find that the number of events from various

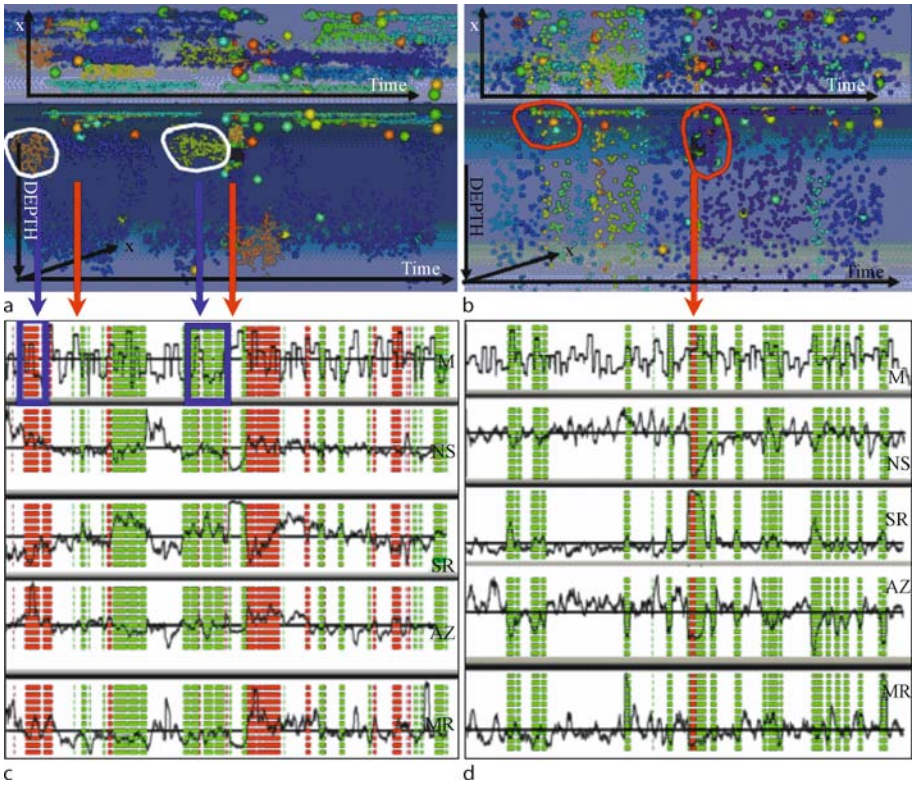

Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 9

Real seismic data [49] analyzed by using clustering in both the data a,b and the feature (c,d) spaces. In panels a and b one can see the results of clustering in the data space (from two different perspectives, X-Time and Depth-Time) for small ($3 < m < 4$) and medium magnitude ($4 < m < 6$) events, respectively, represented by small dots. The different colors of the dots denote different clusters. Large events are visualized by the larger spheres. Their colors show the difference in magnitudes m (*red* – the largest, *green* – the smallest). The clusters in panels a,b encircled in *red* display the places with the largest seismic activity, while those in *white* represent the clusters of small precursory events. The *red*, *white* and *green* stripes in panel c and d representing 4 (out of 7) seismic parameters and maximum magnitude M show the clusters of similar time events for situations corresponding to panels a and b, respectively. The Amira visualization package was used (http://www.amiravis.com)

ranges of magnitudes differs considerably, and divide the entire set of data onto three subsets. The first one comprises the small, the second medium and the last one represents the largest earthquakes displayed in Fig. 9a,b as big dots. The deepest earthquakes $z > 150$ km are not displayed in the Fig. 9. The various shades represent the magnitudes of earthquakes from $m = 6$ (green) to $m = 7$ (red). In Fig. 9 we present the clustering results in both the data $\boldsymbol{f}_i$ and the feature $\boldsymbol{F}_i$ spaces. We look for clusters of similar seismic events (data space) and time events (feature space). The dots (data vectors), belonging to the same clusters, have the same color. The Fig. 9a,b is very rich in cluster-like forms, some of them hard to interpret. Correspondence of the cluster structure of data $\boldsymbol{f}_j (j = 1, \ldots, M_f)$ with the clusters of averaged events $\boldsymbol{F}_i (i = 1, \ldots, M_F)$ in the feature space can reveal interesting information. As one can see from the panel C, only three clusters are obtained in the feature space consisting of small data events ($3 < m < 4$). The green cluster corresponds to two relatively large time intervals of small events preceding Miyakejima earthquake and many smaller post shock periods. The time events $\boldsymbol{F}_i$ from this cluster represent averaged data events $\boldsymbol{f}_j$, mainly shallow (**AZ**) of high degree of spatial repetitiveness **SR** and small diversity of magnitudes (**MR**). The red cluster consists of deeper events of smaller repetitiveness, and more diversified in magnitudes. The larger time interval of this type of behavior is recognized just after Miyakejima shock. The white cluster is not interesting in this scale of small events and includes all other events (including the earthquake swarm).

In panel D we display the seismicity parameters, which form three clusters of time events obtained for seismic events of larger magnitude $4 < m < 6$. Clusters of these events have different structure than in the previous case. They are parallel to X-depth plane. The borders between clusters roughly correspond to the borders of successive showers of the earthquakes. The red cluster comprises only the earthquakes corresponding to the Miyakejima swarm encircled in red in Fig. 9b. As we can see by the MDS visualization displayed in Fig. 10a, this cluster is made up from a needle of time events sprouting away from the two remaining and oval clusters. The green cluster in Fig. 9d represents the deep events, diversified in magnitude of high repetitiveness and rather high degree of spatial nonrandomness at short distances (**NS**). As shown in Fig. 9, these time events represent mainly the post-swarm series of shocks. The white cluster, as before, includes all the other events.

Time Period of 150 Years

In Fig. 11 we display the time series of seismicity parameters computed for the complete synthetic data catalog A. These time series follow the situation from Fig. 3 where dots represent separate data events. The green, red and white strips in Fig. 11 separate 3 clusters of similar time events represented by 7-dimensional feature vectors. In Fig. 10b these clusters are visualized due to the MDS transformation of 7-dimensional feature space into 3-D space. In Fig. 10b each dot represents a 7-dimensional feature vector mapped into 3-D space by MDS transformation.

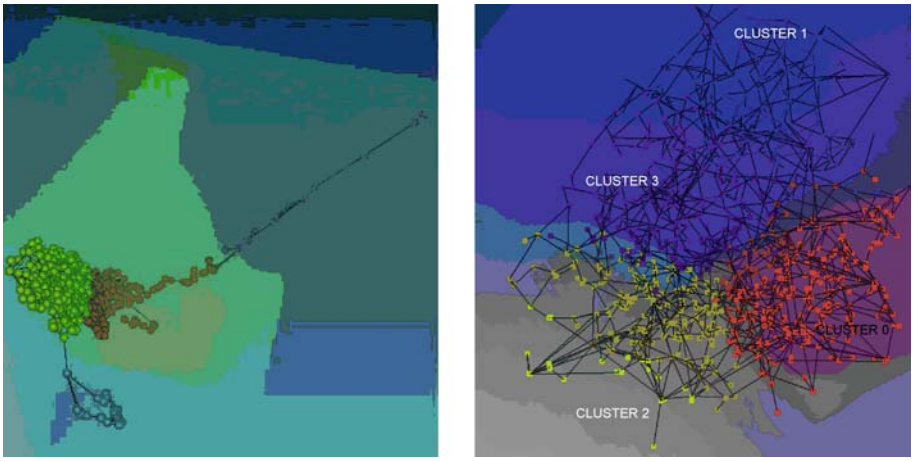

Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 10
The clusters from feature space mapped into 3D space for a realistic short-time interval seismic data. The small *blue* cluster at the bottom represents the events at the end of the time interval, which are averaged within a shrinking time window. b The synthetic seismic data catalog A covering 150 years

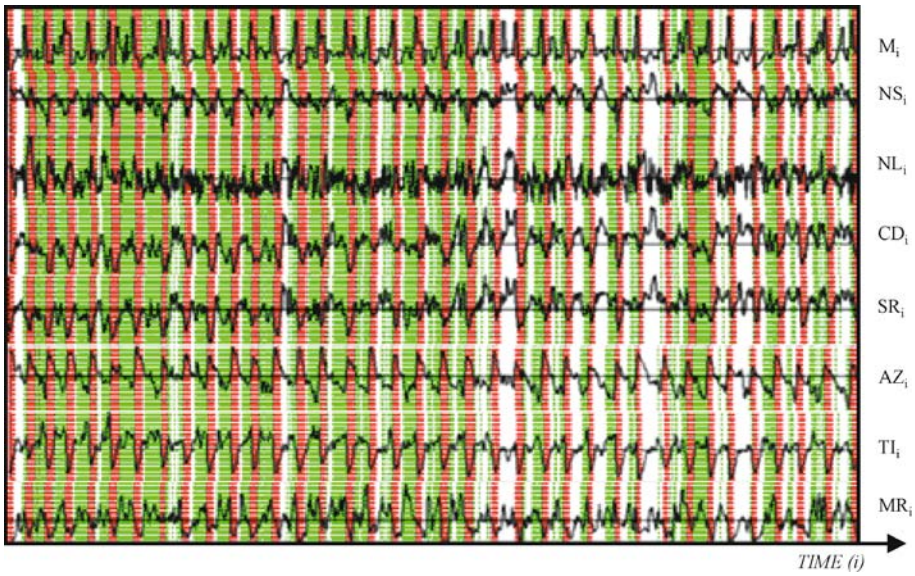

Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 11
The seismicity parameters {M,NS,NL,CD,SR,AZ,TI,MR} in time for synthetic data catalog A (from [25])

From the top panel of Fig. 11 displaying the largest events **M** in the sliding time window, we may conclude that the white (blue in Fig. 10b) and red clusters from Figs. 10, 11b comprise time events, which correspond to the aftershock effects. The white cluster represents the net aftershock events, while the red one includes the earthquake effects averaged in sliding time window. Conversely, the green cluster (yellow in Fig. 10b) contains the time events preceding the earthquakes.

The selectivity in time of the seismicity parameters depends on the width ΔT and shift dt of the sliding time window. Due to space and time averaging, it is impossible to correlate precisely the appearance of an earthquake with the rest of the seismicity parameters when two earthquakes are too close to each other. Therefore, the sequence of green-red-white cluster events can be broken (Fig. 11) into time domains with many large earthquakes. As shown in Fig. 11 the occurrence of the largest events correlates well with the minima of **NS**, **CD, SR, TI**, and maxima of **AZ, MR** parameters. This means that the occurrence of large earthquakes is preceded by increasing spatial diffusion of events and increasing seismicity rate. Moreover, the results confirm the some findings from the real data in a shorter time-scale:

1. The events preceding large earthquakes are shallow and have small magnitudes. They have also higher degree of spatial repetitiveness than events from different clusters.
2. The earthquakes accompanying and following the mainshock are rather large in magnitude, deep, have high seismicity rates and low spatial correlation dimension (this drops off rapidly at the onset of large events),

The analysis of synthetic data shows clearly that the clusters in the feature space reflect well the periodicity of increasing and decreasing seismic activity in a given area. For this scale, however, the fine scale characteristics of precursory and after-shock effects become fuzzy.

Time Period of 1500 Years

In Figs. 4, 5, 6b and Figs. 12, 13 we visualize the feature vectors for data covering 1500 year period for two models: the A model with a Parkfield-type asperity and the M model with multi-size-heterogeneities. In Fig. 4 and Fig. 12 one can recognize two types of clusters with different sizes. The larger cluster comprises feature vectors forming approximately 150-year long periodic time intervals, which are represented by red strips in Fig. 4 and by green strips in Fig. 12. The second cluster consists of feature vectors from periodic gaps colored in green in Fig. 4 and in white in Fig. 12. This anomalous cluster corresponds to the periodic changes in the character of seismic activities. The third cluster (see Fig. 13), marked in red for M type of data in Fig. 12, consists of periodic and short time intervals representing rapid bursts of seismic activity within every 150-year interval.

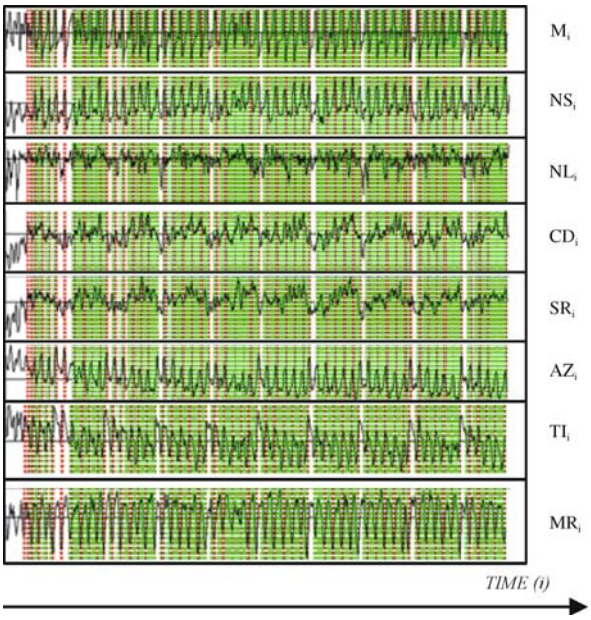

Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 12
The seismicity parameters with time for synthetic (catalog M) for seismic data representing time interval of 1500 years. The *red* and *green* strips depict the events belonging to *red* and *green* clusters from Fig. 7b, respectively

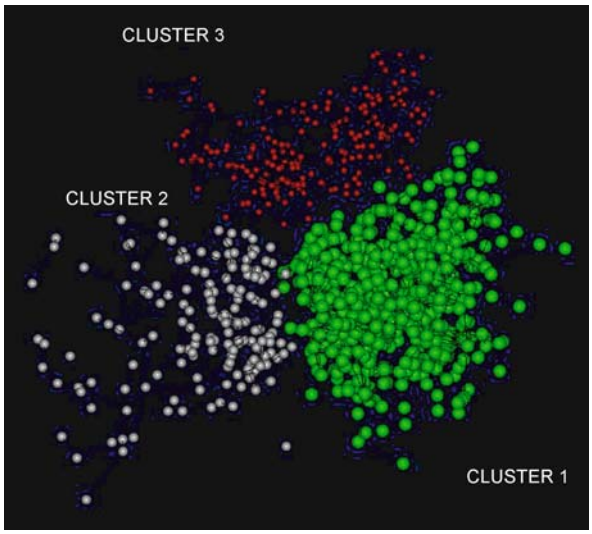

Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 13
The clusters from Fig. 12 mapped by using multidimensional scaling from a 7-dimensional feature space into 3D space for synthetic seismic data catalog M covering 1500 years

In both A and M models the gaps between 150-year long intervals are correlated with decrease of: the correlation dimension (**CD**), degree of spatial repetitiveness (**SR**) and seismicity rate. These gaps are preceded by large earthquakes. The simulations used for generating the datasets incorporate imposed large earthquakes on regions (I) and (IV) of Fig. 8b that bound the computational grid (region II), as staircase boundary conditions with a step at every 150 years. The analysis detected the effects of these boundary conditions on the seismicity that is calculated in region II.

There are also evident differences between the A and M data in the time intervals belonging to the second cluster. For A environment the gaps between 150-year intervals are greater and the secondary periodicity within them is less clear. Moreover, within time intervals from the second cluster, the degree of spatial non-randomness decreases at long distances (**NL**) for M model while for A data it decreases at short distances (**NS**). In addition, the average depth of earthquakes (**AZ**) is then clearly larger for A model, while for M data it remains on the average level.

In sum, by means of analyzing earthquake clusters in feature space over long time-scale, we can investigate important characteristics of seismic activity such as:

1. The occurrence of hierarchical time-periodicity in seismic activity caused by increase of short-time correlations and their destruction, respectively. Correlations can be broken both due to short-wave and long-wave resonances of the Poincare type (e. g. during largest earthquakes) [Sornette, 2004].
2. The dependence of seismic activities on the ambient rheological and geological properties of the environment, which strongly modify the cluster structure of the feature vectors.

Remote Problem Solving Environment (PSE) for Analyzing Earthquake Clusters

Need for Remote Visualization and Analysis

We need fast access to large databases in order to forecast earthquakes by observation of similarities between thousands and millions of seismic events by visualization of earthquake clusters. The largest earthquake catalogs comprise TBytes of data. Taking into account also the data from tsunami earthquakes and micro-earthquakes in mines, the total amount of data collected by seismic centers spread all over the world is humongous. Moreover, knowledge extraction of earthquake precursors may demand exploration of cross-correlation relationships among many different catalogs. Therefore, both

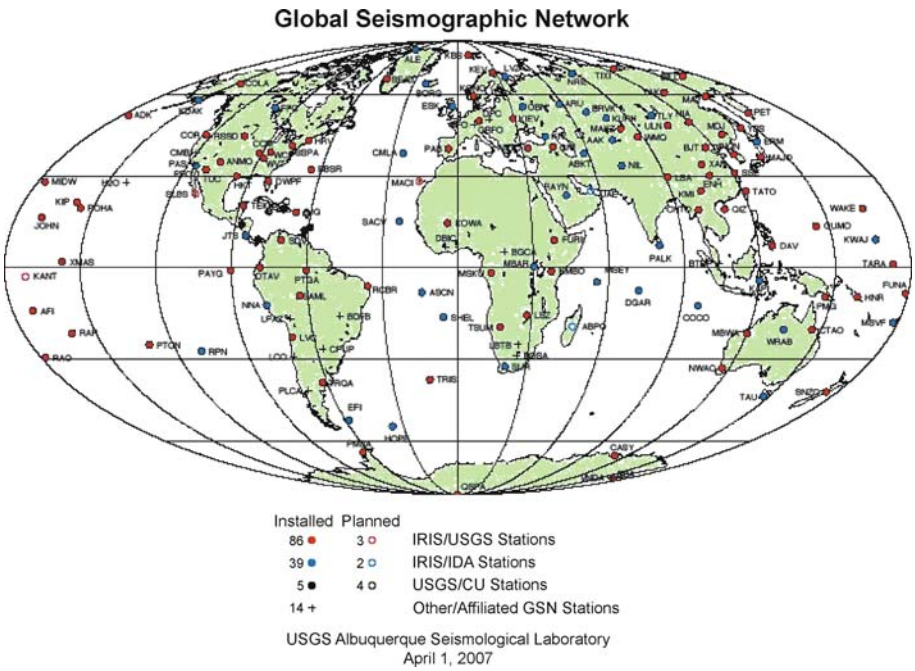

Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 14
Worldwide distribution of earthquake seismographic stations (© USGS)

fast communication between data centers and large disk spaces are sorely needed.

As shown in Fig. 14, earthquake seismograph stations, which collect earthquake data from regions with high seismic activity, are distributed worldwide. Therefore, the unprocessed data needs to be stored and then transferred to a dedicated remote server for data processing. After processing, the results must be returned to data acquisition centers and/or other clients. Broadband access to remote facilities dedicated specifically to pattern recognition and visualization allows for scrutinizing local data catalogs by using peer-to-peer connections of data acquisition centers to data preprocessing servers. Clients in the network can automatically compare various types of earthquake catalogs, data measured in distant geological regions, and the results from various theoretical and computational models. By comparing data accessible in the network we have a chance to eliminate the environmental factors and to extract the resultant earthquake precursory effects.

Integration of a variety of hardware, operating systems, and their proper configuration results in many communication problems between data centers. Efficient, reliable, and secure integration of distributed data and software resources, such as pattern recognition and visualization packages, is possible only within the GRID paradigm of computing [13,31]. The GRID mode of computing has flourished rapidly in recent years and has facilitated collaboration and accessibility to many types of resources, such as large data sets, visualization servers and computing engines. Scientific teams have developed easy-to-use, flexible, generic and modular middleware, enabling today's applications to make innovative use of global computing resources. Remote access tools were also produced to visualize huge datasets and monitor performance and data analysis properties, effectively steering the data processing procedures interactively [21]. The TeraGrid project (http://www.teragrid.org) is a successful high-performance implementation of such a GRID infrastructure and is being used as an integrated, persistent computational resource at universities and laboratories across the USA. The Teragrid development impacts also the earthquake science. The National Science Foundation has awarded the Southern California Earthquake Center 15 million service units of computer processing time on supercomputers nationwide [*Grid Today*, August 2007]. These computational resources will be used for simulating thousands of possible earthquakes scenarios in Southern California, including the largest breaks on the San Andreas fault (www.scec.org/cybershake). SCEC will be able to simulate the most disastrous earthquakes ($M > 7$), such as events that could produce Katrina-scale disasters.

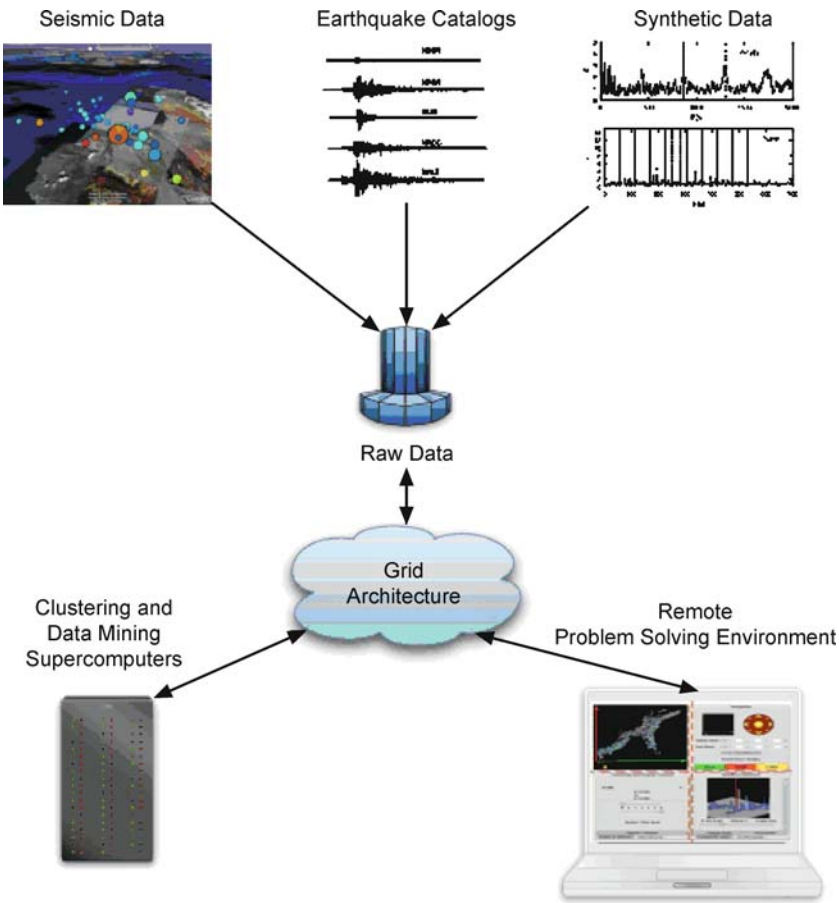

Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 15
Data acquisition, storage, processing, and remote problem solving environments

We discuss the idea of an integrated problem-solving environment (PSE) intended for the analysis of earthquake clusters for the prediction of earthquakes. A simplified scheme for data acquisition and visualization of earthquake clusters is displayed in Fig. 15. This system promotes portability, dynamic results on-demand, and collaboration among researchers separated by long distances by using a client server paradigm. This is provided through a lightweight front-end interface for users to run locally while the a remote server takes care of intensive processing tasks on large databases, off-screen rendering, and data visualization.

Grid Environment

In general, large datasets and high-performance computing resources are distributed across the world. When collaboration and sharing of resources are required, a computational GRID infrastructure needs to be in place to connect these servers (see, e. g., [15]). There must exist protocols available to allow clients to tap into these resources and harness their power. The computational grid can be seen as a distributed system of "clients", which consists of either "users" or "resources" and proxies. A GRID can be implemented using an event brokering system designed to run on a large network of brokering nodes. Individually, these brokering nodes are competent servers, but when connected to the brokering system, they are able to share the weight of client requests in a powerful and efficient manner. Examples of this include GRID Resource Brokering [30] and NaradaBrokering.

These GRID architectures are well suited to the functionality of a PSE for earthquake cluster analysis and as an integrated computational environment for data exchange and common ventures. The seismic data centers from the networking point of view represent a complex hierarchical cluster structure. They are located geographically in the regions of high seismic activity within heavily popu-

lated areas of economic importance. Therefore, the seismic data centers create distant superclusters of various "density" of computational resources corresponding to the size and importance of the regions. These superclusters are sparse in the sense of computational resources devoted for earthquake detection and data acquisition. However, these same structures contain important computational, scientific and visualization facilities with strong interest in the analysis of earthquake data and earthquake modeling. The efficient interconnection of these sites is of principal interest. Due to the "small world network" structure of GRID architectures it is possible to select the most efficient routing schemes, considerably shortening the average communication path length between brokers. GRID architectures are appropriate to link the clients, both users and resources, together. Construction of efficient and user friendly Problem Solving Environments requires integration of data analysis and visualization software within the GRID environment, in such a way that it can be easily accessed via the Internet. We created an integrated data interrogation toolkit to act as a PSE for visualization and clustering of seismic data, which we call WEB-IS.

Example of Remote PSE

WEB-IS is a software tool that allows remote, interactive visualization and analysis of large-scale 3-D earthquake clusters over the Internet [85] through the interaction between client and server. WEB-IS acts as a PSE through a web portal used to solve problems by visualizing and analyzing geophysical datasets, without requiring a full understanding of the underlying details in software, hardware and communication [34,52]. As shown in Fig. 16, the primary goal of WEB-IS in the geosciences is to provide middleware that sits between the modeling, data analysis tools and the display systems that local or remote users access. In the case of large and physically distributed datasets, it is necessary to perform some preprocessing and then transmit a subset of the data to one or more processes or visualization servers to display. The details of where and how the data migrates should be transparent to the user. WEB-IS makes available to the end users the capability of interactively exploring their data, even though they may not have the necessary resources such as sufficient software, hardware or datasets at their local sites. This method of visualization allows users to navigate through their rendered 3-D data and analyze for statistics or apply earthquake cluster analysis. To the client, the process of accessing and manipulating the data appears simple and robust, while the middleware takes care of the network communication, security and data preparation.

Complete realization of an earthquake clustering PSE consists of:

1. Data analysis tools to implement earthquake clustering techniques;
2. High performance visualization techniques using OpenGL or Amira;
3. The Grid environment;
4. Integration toolkit, such as WEB-IS.

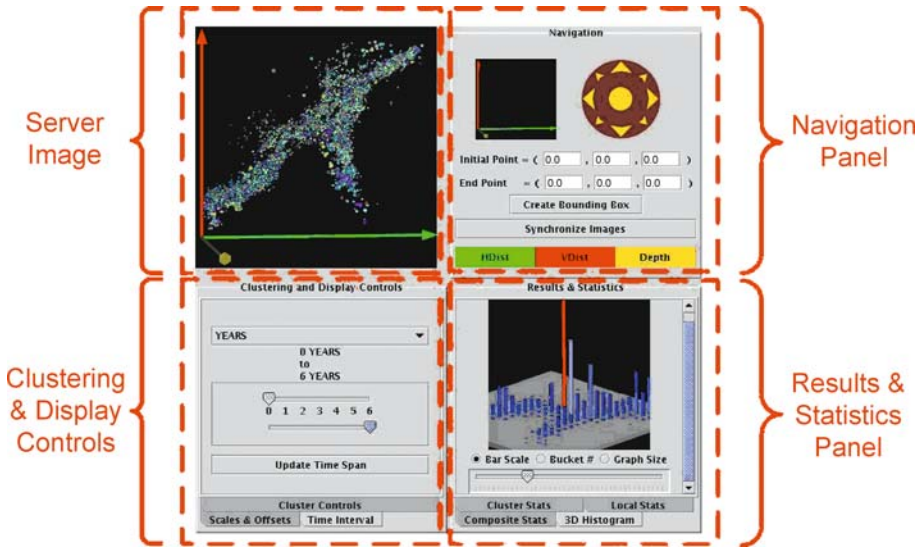

Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 16
WEB-IS is an example of a remote earthquake clustering PSE

These exist and can work both independently and coupled in a single special purpose system. This system can be developed creating the backbone of the sophisticated computational data acquisition environment, which can be devised specifically for earthquake clustering or for general needs of the geophysical community. Equipped with only PDAs or laptops, and working on location in unreachable desert terrains with remote data acquisition centers or perhaps just analyzing data in one of the many computation facilities located around the globe, geophysicists will be enabled unlimited access to data resources spread all over the world.

We see the principal goal of our work in contributing to the construction of a global warning system, which can be used for prediction of catastrophes such as various types of earthquakes along the circum Pacific belt, where there is a great concentration of people. For example, similar methodology can be used for tsunami earthquake alerting. Theoretical models of faulting and seismic wave propagation used for the computation of radiated seismic energy from broad-band records at teleseismic distances [14] can be adapted to the real-time situation when neither the depth nor the focal geometry of the source is known accurately. The distance-dependent approximation was used in [60]. By analyzing some singular geophysical parameters such as the energy-to moment ratio H [60] for regular earthquakes, the results obtained from the theoretical models agree well with values computed from available source parameters (e. g., as published by the National Earthquake Information Center). It appears however that the so called "tsunami earthquakes" – characterized by the significant deficiency of moment release at high frequencies – yield the values of H considerably different the regular earthquakes. Thus H value can be used as a suitable criterion for discriminating various types of earthquakes in a short duration of time, like an hour. However, this hypothesis holds only for a few cases. For, so called, "tsunamigenic earthquakes" this difference is not so clear. Moreover, the value of the moment computed on the base of long-period seismic waves can be underestimated. For example, analysis of the longest period normal modes of the Earth, 0S2 and 0S3, excited by the December 26, 2004 Sumatra earthquake [76], yields an earthquake moment of $1.3 \cdot 1030$ dyn-cm, approximately three times larger than the $4 \cdot 1029$ dyn-cm measured from long-period surface waves. Therefore, instead of a single-value discrimination we recommend using more parameters (dimensions) for detecting tsunami earthquakes. As shown in [64] and [83], one could employ other T-phase characteristics such as its duration, seismic moment, and spectral strength or even similar features associated with the S-phase. We believe that the lack of success in predicting earthquakes still comes from the lack of communications between researchers and difficulties in free and fast access to the various types of data. Therefore, we hope that globalization of computation, data acquisition and visualization resources, together with fast access through a scale-free network, will provide a triumphant solution to this problem.

Future Directions

In this chapter we endeavor to bring across the basic concept of clustering and its role in earthquake forecasting. Indeed we find that the clustering of seismic activities reflects both the similarity among them and their correlation properties. As discussed in, e. g., Ben-Zion et al. [9], Saichev and Sornette [68] and Zöller et al. ► Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, there exists an evolutionary process or memory between successive earthquakes, which impact the distribution of the inter-event times. We believe that by means of earthquake clustering we can capture the essence of this predictive information [27]. Therefore, in order to carry out real-time earthquake forecasting for short-time scales, it is necessary to derive a thorough understanding of all families of earthquake clusters produced over an earthquake-prone region.

We stress here that in obtaining this type of information one must first be able to detect the precise location of the significant clusters, by filtering out simultaneously the noise and the outliers. While the existence of spatial-temporal clusters is important, they do not reveal the subtle information hidden behind the relations among the data events, such as: spatial-temporal correlation dimensions, correspondence between the numbers of small and large magnitude events, degree of spatial randomness, repetitiveness at different distances and other factors. The features – "descriptors" or seismicity parameters – constructed from the empirical knowledge of the researcher should be largely independent and should represent aptly distinctive features, which are useful for the purpose of pattern recognition. Unlike single events described only by spatio-temporal features (and magnitude), the N-dimensional feature vectors can represent better the dynamics of the seismically active area in different moments of time. By following the basic rules of learning theory, we may be able to arrive at the number N and quality of features, which can assure the generalization power of the data and allow us to construct reliable data-models or classifiers.

We have shown that clustering, as a well-honed tool in data mining and pattern recognition, represents the clas-

sifier without the teacher, which means that the nature of the clustering is unknown and its exact background must be guessed at from expert knowledge and analysis of the cluster properties. Clustering is a process based on a priori knowledge extraction for constructing the hypothesis space needed for reliable classifiers that can be taught and used for forecasting [25]. However, the quality of these data models depends strongly on the quality of hypothesis space constructed. Consequently, it depends on the quality of clusters extraction. The major problem comes from the lack of a universal clustering scheme, thus making the clustering process somewhat subjective. In this case we must visualize the multidimensional feature space. Visual confirmation gives one a confidence concerning the validity of the clusters and we can then adjust for the optimal clustering procedures by removing the noise and outliers. Among the major goals of earthquake clustering, we can include the following salient points:

- classification of the chaotic properties of seismicity patterns [35], for example to recognize the three main groups of shocks: foreshocks, mainshocks and aftershocks or to remove the temporary clustering to estimate the background seismicity;
- understanding the correlations between observed properties of earthquakes in different domains (e. g., space, time, number, size);
- understanding the relations between various physical parameters of the models and properties of the generated earthquakes;
- investigating the multi-scale nature of the cluster structure and reconstructing the important and hidden information associated with the stress characteristics.

Classification of type of shocks seems to be an unresolved problem because there are no observable differences between foreshocks, main shocks and aftershocks [68]. Each earthquake is able of triggering other earthquakes according to the basic laws from [46,69]. Despite this difficulty, as shown in [9], it is possible to construct some sort of stochastic classifiers based on theoretical footing. The method proposed here closely related to the epidemic-type aftershock sequence (ETAS) model [61]. It is important that the principal characteristics of ETAS-based models correspond to experimental verifications, i. e., they treat all earthquakes on the same footing and there is not distinction between foreshocks, main shocks and aftershocks. The key points of the method are the probabilities of one event being triggered by a previous event (e. g., [82]). Making use of these probabilities, we can reconstruct the functions associated with the characteristics of earthquake clusters to test a number of plausible hypotheses about the earthquake clustering phenomena.

As shown above by our results on seismicity clustering for the three different time epochs, clustering can be truly regarded as a coarse-graining procedure. We can see details from the smaller scales are erased, thereby exposing the general trends associated with the long correlation length. For large data bases covering long time intervals we can unveil the shorter timescale characteristics by removing the background events, using successive clustering. Eventually, we can build up the strong classifiers. In the case where the long-time data catalogs are missing, we can employ the stochastic classifiers advocated Ben-Zion et al. [9] for prior thresholding of the background data or what is sometimes called "fuzzification" [86]. By this procedure we can construct the hypothesis space for data models by clustering (or fuzzy clustering) procedures.

The results discussed in this paper contribute to the development of improved software infrastructure for analysis of seismicity. A combined clustering analysis of observed and synthetic data, aided by state-of-the-art visualization of multidimensional clusters, undoubtedly lead to improved earthquake forecasting algorithms with shorter time windows of increased probability of large seismic events.

Acknowledgments

This research was supported by NSF ITR and Math-Geo grants. WD acknowledges support from the Polish Committee for Scientific Research (KBN) Grant No. 3T11C05926. YBZ acknowledges support from the NSF, USGS and SCEC.

Bibliography

Primary Literature

1. Amira visualization package. http://www.amiravis.com
2. Andenberg MR (1973) Clusters analysis for applications. Academic Press, New York
3. Bak P, Tang C (1989) Earthquakes as a self-organized critical phenomena. J Geophys Res 94(B11):15635–15637
4. Bak P, Christensen K, Danon L, Scanlon T (2002) Unified scaling law for earthquakes. Phys Rev Lett 88:178501
5. Barabasi AL, Jeong H, Neda Z, Ravasz E, Schubert A, Vicsek T (2002) Evolution of the social network of scientific collaborations. Physica A 311(3–4):590–614
6. Bennett AF (1992) Inverse methods in physical oceanography. Cambridge University Press, Cambridge, pp 346
7. Ben-Zion Y (1996) Stress, slip and earthquakes in models of complex single-fault systems incorporating brittle and creep deformations, J Geophys Res 101:5677–5706
8. Ben-Zion Y (2001) Dynamic rupture in recent models of earthquake faults. J Mech Phys Solids 49:2209–2244

9. Ben-Zion Y (2003) Appendix 2, key formulas in earthquake seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International handbook of earthquake and engineering seismology, Part B. Academic Press, pp 1857–1875
10. Ben-Zion Y, Rice JR (1993) Earthquake failure sequences along a cellular fault zone in a three-dimensional elastic solid containing asperity and nonasperity regions. J Geophys Res 98:14109–14131
11. Ben-Zion Y, Rice JR (1995) Slip patterns and earthquake populations along different classes of faults in elastic solids. J Geophys Res 100:12959–12983
12. Ben-Zion Y, Eneva M, Liu Y (2003) Large earthquake cycles and intermittent criticallity on heterogeneous faults due to evolving stress and seismicity. J Geophys Res 108(B6):2307–27
13. Berman F, Fox GC, Hey AJG (2003) Grid computing – making the global infrastructure a reality. Wiley Series in Communications Networking and Distributed Systems, pp 1007
14. Boatwright J, Choy GL (1986) Teleseismic estimates of the energy radiated by shallow earthquakes. J Geophys Res 91:2095–2112
15. Bollig EF, Jensen PA, Lyness MD, Nacar MA, da Silveira PR, Erlebacher G, Pierce M, Yuen DA (2007) VLAB: Web services, portlets, and workflows for enabling cyber infrastructure in computational mineral physics. J Phys Earth Planet Inter 163:333–346
16. Chen C-C, Rundle JB, Li H-C, Holliday JR, Turcotte DL, Tiampo KF (2006) Critical point theory of earthquakes: Observations of correlated and cooperative behavior on earthquake fault systems. Geophys Res Lett L18302
17. Chinnery M (1963) The stress changes that accompany strike-slip faulting. Bull Seismol Soc Am 53:921–932
18. Corral A (2005) Mixing of rescaled data and Bayesian inference for earthquake recurrence times. Nonlinear Process Geophys 12:89–100
19. Corral A (2005) Renormalization-group transformations and correlations of seismicity. Phys Rev Lett 95:028501
20. Corral A, Christensen K (2006) Comment on earthquakes descaled: On waiting time distributions and scaling laws. Phys Rev Lett 96:109801
21. da Silva CRS, da Silveira PRC, Karki B, Wentzcovitch RM, Jensen PA, Bollig EF, Pierce M, Erlebacher G, Yuen DA (2007) Virtual laboratory for planetary materials: System service architecture overview. Phys Earth Planet Inter 163:323–332
22. Davy P, Sornette A, Sornette D (1990) Some consequences of a proposed fractal nature of continental faulting. Nature 348:56–58
23. Dzwinel W, Blasiak J (1999) Method of particles in visual clustering of multi-dimensional and large data sets. Future Gener Comput Syst 15:365–379
24. Dzwinel W, Yuen DA, Kaneko Y, Boryczko K, Ben-Zion Y (2003) Multi-resolution clustering analysis and 3-D visualization of multitudinous synthetic earthquakes. Vis Geosci 8:12–25
25. Dzwinel W, Yuen DA, Boryczko K, Ben-Zion Y, Yoshioka S, Ito T (2005) Nonlinear multidimensional scaling and visualization of earthquake clusters over space, time and feature space. Nonlinear Process Geophys 12:117–128
26. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95:14863–14868
27. Enescu B, Ito K, Struzik ZR (2006) Wavelet-based multiscale analysis of real and simulated time-series of earthquakes. Geophys J Int 164:63–74
28. Eneva M, Ben-Zion Y (1997) Techniques and parameters to analyze seismicity patterns associated with large earthquakes. J Geophys Res 102(B8):785–795
29. Eneva M, Ben-Zion Y (1997) Application of pattern recognition techniques to earthquake catalogs generated by models of segmented fault systems in three-dimensional elastic solids. J Geophys Res 102:24513–24528
30. Ferreira L (2002) Introduction to grid computing with Globus IBM Redbook series. IBM Corporation http://ibm.com/redbooks
31. Foster I, Kesselman C (eds) (1998) Building a computational grid: state-of-the art and future directions in high-performance distributed computing. Morgan-Kaufmann, San Francisco
32. Freed AM, Lin J (2001) Delayed triggering of the 1999 Hector Mine earthquake by viscoelastic stress transfer. Nature 411:180–183
33. Frey BJ, Dueck D (2007) Clustering by Passing Messages Between Data Points, Science 315(5814):972–976
34. Garbow ZA, Erlebacher G, Yuen DA, Sevre EO, Nagle AR, Kaneko Y (2002) Web-based interrogation of large-scale geophysical datasets and clustering analysis of many earthquake events from desktop and handheld devices. American Geophysical Union Fall Meeting, Abstract
35. Goltz C (1997) Fractal and chaotic properties of earthquakes. In: Goltz C (ed) Lecture notes in earth sciences, vol. 77. Springer, Berlin, p 3–164
36. Gowda CK, Krishna G (1978) Agglomerative clustering using the concept of nearest neighborhood. Pattern Recognit 10:105
37. Grossman RL, Karnath Ch, Kegelmeyer P, Kumar V, Namburu RR (2001) Data mining for scientific and engineering applications. Kluwer, Dordrecht
38. Guha S, Rastogi R, Shim K (1998) CURE: An efficient algorithm for large databases. In: Proceedings of SIGMOD '98, Seattle, June 1998. pp 73–84
39. Gutenberg B (1942) Earthquake magnitude, intensity, energy and acceleration. Bull Seismol Soc Am 32:163–191
40. Gutenberg B, Richter CF (1954) Seismicity of the earth and associated phenomena. Princeton University Press, Princeton
41. Haile PM (1992) Molecular Dynamics Simulation. Wiley, New York
42. Hainzl S, Scherbaum F, Beauval C (2006) Estimating background activity based on interevent-time distribution. Bull Seismol Soc Am 96:313–320. doi:10.1785/0120050053
43. Hand D, Mannila H, Smyth P (2001) Principles of data mining. MIT Press, Cambridge
44. Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning: Data mining, inference and prediction. Springer, New York, pp 533
45. Helmstetter A, Sornette D, Grasso J-R (2003) Mainshocks are aftershocks of conditional foreshocks: How do foreshock statistical properties emerge from aftershock laws. J Geophys Res 108:2046
46. Helmstetter A, Kagan Y, Jackson D (2005) Importance of small earthquakes for stress transfers and earthquake triggering. J Geophys Res 110:B05S08

47. Hong H, Kadlec DJ, Yuen DA, Zheng Y, Zhang H, Liu G, Dzwinel W (2004) Fast timescale phenomena at Changbaishan volcano as inferred from recent seismic activity. Eos Trans AGU Fall Meet. 85(47) http://www.agu.org
48. Ismail MA, Kamel MS (1989) Multi-dimensional data clustering utilizing hybrid search strategies. Pattern Recognit 22(1):77–89
49. Ito T, Yoshioka S (2002) A dike intrusion model in and around Miyakejima, Niijima and Kozushima. Tectonophysics 359:171–187
50. Jajuga K, Sokolowski A, Hermann H (eds) (2002) Classification, clustering and data analysis. Springer, Berlin, pp 497
51. Jones NC, Pevzner P (2004) An introduction to bioinformatics algorithms. MIT Press, Cambridge
52. Kadlec BJ, Yang XL, Wang Y, Bollig EF, Garbow ZA, Yuen DA, Erlebacher G (2003) WEB-IS (Integrated System): An overall view. Eos Trans AGU 84(46), Fall Meet. Suppl., Abstract NG11A-0163
53. Kalnay E (2003) Atmospheric modeling, data assimilation and predictability. Cambridge University Press, Cambridge, pp 341
54. Karypis G, Han E, Kumar V (1999) Chameleon: A hierarchical clustering algorithms using dynamic modeling. IEEE Computer 32(8):68–75
55. Karypis G, Aggarwal R, Kumar V, Shekhar S (1999) Multilevel hypergraph partitioning: applications in VLSI domain. IEEE Trans on Very Large Scale Systems Integration (VLSI) 7(1):69–79
56. Mehta AP, Dahmen KA, Ben-Zion Y (2006) Universal mean moment rate profiles of earthquake ruptures. Phys Rev E 73:056104
57. Mitra S, Acharya T (2003) Data mining: multimedia, soft computing and bioinformatics. Wiley, New Jersey
58. Molchan GM (2005) Interevent time distribution of seismicity: A theoretical approach. Pure Appl Geophys 162:1135–1150
59. National Research Council (2003) Living on an active earth, perspectives on earthquake sciences. The National Academies Press, Washington DC
60. Newman AV, Okal EA (1998) Teleseismic estimates of radiated seismic energy: the S/M0 discriminant for tsunami earthquakes. J Geophys Res 103(B11):23885–23898
61. Ogata Y (1999) Seismicity analysis through point-process modeling: A review. Pure Appl Geophys 155:471–507
62. Ogata Y, Zhuang J (2006) Space-time ETAS models and an improved extension. Tectonophysics 413:13–23
63. Okada Y (1992) Internal deformation due to shear and tensile faults in a half space. Bull Seismol Soc Am 82:1018–1040
64. Okal EA, Alasset P-J, Hyvernaud O, Schindele F (2003) The deficient T waves of tsunami earthquakes. Geophys J Int 152:416–432
65. Rundle JB, Gross S, Klein W, Ferguson C, Turcotte DL (1997) The statistical mechanics of earthquakes. Tectonophysics 277:147–164
66. Rundle JB, Klein W, Tiampo K, Gross S (2000) Linear pattern dynamics in nonlinear threshold systems. Phys Rev E 61(3):2418–2143
67. Rundle JB, Turcotte DL, Klein W (eds) (2000) GeoComplexity and the physics of earthquakes. American Geophysical Union, Washington, pp 284
68. Saichev A, Sornette D (2007) Theory of earthquake recurrence times. J Geophys Res 112(B4):1–26
69. Saichev A, Helmstetter A, Sornette D (2005) Power law distributions of offspring and generation numbers in branching models of earthquake triggering. Pure Appl Geophys 162:1113–1134
70. Sander J, Ester M, Krieger H (1998) Density based clustering in spatial databases, The algorithm DBSCAN and its applications. Data Min Knowl Discov 2(2):169–194
71. Shcherbakov R, Turcotte DL (2004) A modified form of Bath's law. Bull Seismol Soc Am 94:1968–1975
72. Shcherbakov R, Turcotte DL, Rundle JB (2005) Aftershock statistics. Pure Appl Geophys 162(6–7):1051–1076
73. Siedlecki W, Siedlecka K, Sklanski J (1988) An overview of mapping for exploratory pattern analysis. Pattern Recognit 21(5):411–430
74. Sornette D (2006) Critical phenomena in natural sciences. Springer Series in Synergetics, Berlin, pp 528
75. Sornette D, Johansen A, Bauchaud J-P (1996) Stock market crashes, precursors and replicas. J Phys Int Finance 5:167–175
76. Stein S, Okal E (2004) Ultra-long period seismic moment of the great December 26, 2004 Sumatra earthquake and implications for the slip process. http://www.earth.nwu.edu/people/emile/research/Sumatra.pdf
77. Tan P-N, Steinbach M, Kumar V (2005) Introduction to data mining. Addison Wesley, Boston, pp 769
78. Theodoris S, Koutroumbas K (1998) Pattern recognition. Academic Press, San Diego
79. Turcotte DL (1997) Fractals and chaos in geology and geophysics, 2nd Edn. Cambridge University Press, New York
80. Utsu T (2002) Statistical Features of Seismicity. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part A. Academic Press, pp 719–732
81. Van Aalsburg J, Grant LB, Yakovlev G, Rundle PB, Rundle JB, Turcotte DL, Donnellan A (2007) A feasibility study of data assimilation in numerical simulations of earthquake fault systems. Phys Earth Planet Inter 163:149–162
82. Vere-Jones D (1976) A branching model for crack propagation. Pure Appl Geophys 114(4):711–726
83. Walker DA, Mc Creery CS, Hiyoshi Y (1992) T-phase spectra, seismic moment and tsunamigenesis. Bull Seismol Soc Am 82:1275–1305
84. Wesnousky SG (1994) The Gutenberg-Richter or characteristic earthquake distribution, which is it? Bull Seismol Soc Amer 84:1940–1959
85. Yuen DA, Garbow ZA, Erlebacher G (2004) Remote data analysis, Visualization and Problem Solving Environment (PSE) based on wavelet analysis in the geosciences. Vis Geosci 8:83–92. doi:10.1007/x10069-003-0012-z
86. Zadeh LA (1996) Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers by Lotfi A Zadeh, World Scientific Series In Advances In Fuzzy Systems. World Scientific Publishing, River Edge, pp 826
87. Zhang Q, Boyle R (1991) A new clustering algorithm with multiple runs of iterative procedures. Pattern Recognit 24(9):835–848
88. Zöller G, Hainzl S, Ben-Zion Y, Holschneider M (2006) Earthquake activity related to seismic cycles in a model for a heterogeneous strike-slip fault. Tectonophysics 423:137–145. doi:10.1016/j.tecto.2006.03.007
89. Zöller G, Ben-Zion Y, Holschneider M (2007) Estimating recurrence times and seismic hazard of large earthquakes on an

individual fault. Geophys J Int 170:1300–1310. doi:10.1111/j.1365-246X.2007.03480.x

Books and Reviews

Ertoz L, Steinbach M, Kumar V (2003) Finding clusters of different size, shapes and densities in noisy, high-dimensional data. Army High Performance Center, technical report, April 2003

Yuen DA, Kadlec BJ, Bollig EF, Dzwinel W, Garbow ZA, da Silva C (2005) Clustering and visualization of earthquake data in a grid environment, vol 10/1. Vis Geosci http://www.springerlink.com/content/n60423820556/

Earthquake Damage: Detection and Early Warning in Man-Made Structures

Maria I. Todorovska
Department of Civil Engineering,
University of Southern California, Los Angeles, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Literature Review
Damage and Damage-Sensitive Features
Structural Models and Identification
Examples
Future Directions
Bibliography

Glossary

Structural health monitoring Is the process of determining and tracking the structural integrity and assessing the nature of damage in a structure. It is often used interchangeably with *structural damage detection*.

Inter-story drift Is the ratio between the relative horizontal displacements at two levels of the structure and the distance between them. It is important to distinguish between drift resulting from *deformation* of the structure, which is directly related to damage, and drift resulting from the deformation of the soil and rocking of the structure as a rigid body. It is also important to estimate reliably the drift due to permanent displacement (its "DC" component), which cannot be done reliably using data from vibrational sensors unless six degrees of freedom of motion (three translations and three rotations) are recorded.

Soil-structure interaction (SSI) Is a process occurring during vibration of structures founded on flexible soil, in which the structure and soil interact, and their motions are modified. *Kinematic interaction* refers to the effects of scattering and diffraction of the incident seismic waves from the soil excavation for the foundation. *Dynamic interaction* refers to the effects caused by the inertia forces of the structure and foundation, which lead to deformation of the soil, and results in modification of the *resonant frequencies* and *damping* of the response of the structure, foundation and soil acting as a system.

Resonant frequencies of vibration Of a structure on flexible soil are those of the *soil-structure system*, and the energy of the vibrational response is concentrated around these frequencies. They depend on the stiffness of the building and that of the soil. *Fixed-base frequencies of vibration* are the resonant frequencies of the structure on rigid soil, and depend only on the stiffness of the structure. Loss of stiffness of the structure due to damage results in reduction of the fixed-base frequencies, and indirectly of the system frequencies. Monitoring changes in the fixed-base frequencies is most reliable because it eliminates the effects of the soil, which can exhibit (recoverable) nonlinear behavior during strong shaking.

Definition of the Subject

Structural health monitoring and structural damage detection refers to the process of determining and tracking the structural integrity and assessing the nature of damage in a structure. Being able to detect the principal components of damage in structures as they occur during an earthquake or soon after the earthquake, or the absence of it, before physical inspection is possible, is an important and challenging problem. Considering the challenges faced and the potential benefits for safety and for minimizing disruption of productivity, structural health monitoring has the elements of a *grand challenge* problem in civil engineering [12].

Structural damage can be described by the following five attributes: existence, location, type, extent, and prognosis for the remaining useful life. Structural damage is a complex state, which can occur on different time scales, suddenly during some catastrophic event such as earthquake or explosion, or gradually over the life of the structure, due to deterioration of the structural materials by aging, service, and exposure to environmental influences. This article is concerned primarily with identification of the most significant components in the space of complex

patterns of damage caused by earthquakes. Damage in structures also can be described on different spatial scales, e. g. from small defects and localized damage in a component, to global state of damage of the structural system. Hence the damage detection methods are classified as *local* and *global*. The local methods are those for nondestructive testing (NDT) of materials, which can determine the location of the damage in a structural component. They involve use of actuators (radiating ultrasonic waves into the structural element), and require access to the element. The global methods assess the overall state of damage of the structural system (as it reflects on its overall performance during an extreme event). The focus of this review is on the global methods, and intermediate scale methods, which can point to the part of the structure that has been damaged.

Structural damage detection and early warning involve: (1) *recording* some sensory data, (2) *identification* of some structural parameter(s) sensitive to damage (e. g. natural frequencies of vibration, or wave travel times), some characteristic of response (e. g. levels of inter-story drift) that can be correlated with damage, or some other patterns (e. g. abrupt changes in the response detected as novelties), (3) *comparison* of the result of the identification with some knowledge base of correlation of such patterns with levels of damage, and (4) *decision making* (e. g. whether to evacuate or continue occupancy). Because of various uncertainties, the answer can be only expressed probabilistically, and the decision will also depend on the nature of the use of the structure and level of tolerance of the user.

The earliest and most wide-spread methods of structural damage detection are those based on data from vibrational sensors. In fact, the hope to eventually be able to detect hidden damage has been one of the motivations for the development and deployment of seismic sensors in structures. The first strong motion recordings in a building are those during the $M = 5.4$ Southern California earthquake of October 2, 1933, obtained in the Hollywood Storage Building, the instrumented structure in US with the longest history of recording earthquakes [63]. The earliest identification methods consisted of estimation of the building resonant frequencies and damping, from energy distributions of small amplitude ambient noise and forced vibration tests [3], as well as from earthquake records [63]. These studies identified the resonant frequencies and damping of the soil-structure system, which depend on the properties of the soil, and can change significantly even when there is no damage. Detailed system identification studies from full-scale test vibration data that separate the effects of the soil-structure interaction appeared in the 1970s, following theoretical developments that helped understanding the phenomenon of soil-structure interaction [13,32,33,34,71]. Thirty years later, such studies are still rare, due to a combination of factors, one of which is the inadequate coverage of this topic in the graduate curricula, and the other is the emphasis of earthquake engineering research on laboratory experimentation and numerical simulations, rather than on the full-scale testing of structures [63].

Despite the progress made to date in instrumentation of structures as well as in development of theoretical methods, structural health monitoring systems are deployed in structures only on an experimental basis. The main obstacles to the routine practical deployment of such systems are: (1) the high cost of sensors and monitoring systems, which limits the number of structures that are instrumented and the detail of the measurements (spatial resolution, e. g.), (2) the low sensitivity and robustness of the methods, and ability to discriminate between changes in the damage sensitive feature caused by damage from changes caused by other factors (e. g. age, level of excitation, and weather), and (3) the paucity of data recorded in damaged structures necessary to calibrate the health monitoring methods. Consequently, the main challenges for future research are: (1) to design low cost but high performance sensors and monitoring systems, making it possible to densely instrument many structures, (2) to develop methods that are robust and sensitive enough to detect also light damage (in particular one that is not visible), and (3) to build a knowledge base that can help reliably relate observed patterns in the data with actual observations of damage.

Recently, structural identification and health monitoring of buildings by detecting changes in wave travel time through the structure has received revived attention and has proven to be very promising [20,24,25,37,39,41,42,53,54,65]. Exploratory applications to data from damaged buildings [53,54] showed that the method (1) is robust when applied to damaging levels of earthquake response data, (2) is not sensitive to the effects of soil-structure interaction, and (3) is local in nature (i. e. gives results consistent with the spatial distribution and degree of the observed damage).

Introduction

This volume would not be complete without addressing the catastrophic consequences of earthquakes, and damage in soil-structure systems, which is a complex, multidimensional, and highly interrelated set of phenomena.

Since the early days, the mathematical formulation of practical earthquake engineering problems has been dominated by *linear* differential equations [58], which *cannot* lead to chaos. Nevertheless, cost and the increasing needs of society have pushed the design into the nonlinear regimes of large deformations increasing the possibility of encountering chaotic dynamic phenomena in structural response, and have increased the complexity of the possible damage outcomes. However, working with parameters that produce chaotic output reduces the ability to predict the outcome. The chaotic behavior of nonlinear systems does not completely exclude the possibility to predict the response, but introduces an upper bound (prediction horizons) [30]. Then the remaining question is over what time-scales can the predictions still be reliable. Also, the prediction of response requires a realistic physical *model*, while the practical outcome of most work in engineering remains *empirical*. Consequently, there is a conflict in the classical engineering description of the world. This conflict is in part due to the assumption that nature is moving forward, according to a deterministic law, and in part due to the fact that engineers model the world based on incomplete data, and thus working with unverifiable representation. This leads to the question what models are good for. The problem is further aggravated by the fact that the art of dynamical modeling tends to be neglected in discussions of nonlinear and chaotic systems, in spite of its crucial importance [2]. In the following review of structural health monitoring in earthquake engineering, it is accepted that there is a modeling problem, and the success of a method is gauged by the degree to which its predictions match the observed outcomes.

During the last several decades, stochastic processes have been used to help analyze the irregular behavior of deterministic systems with too many variables to be described in detail. Stochastic processes have been used also to model the deterministic response of structures to earthquake and wind forces, and as an approximate description of deterministic systems sensitive to their initial conditions. In some analyses, random noise is added to the model to account for the differences between the behaviors of model and prototype. This noise represents no more than lack of knowledge of the system structure or inadequacy of the identification procedure [23].

Following a damaging earthquake, buildings, bridges, dams and other structures are physically inspected for damage, and their safety is assessed. To assess the safety of buildings, the city departments of public safety (or their equivalents) dispatch inspectors to the field to "walk through" each building and write a report on the observed damage and safety concerns to its occupants. On the basis of such assessments, a color tag can be assigned to the building: (1) green if the structure is safe, (2) yellow if it has been damaged and needs to be evacuated, but is safe for the occupants to return to retrieve their belongings, and (3) red if it has been damaged to a degree that it is unsafe for the occupants to return to the structure [1]. When the affected area is relatively large, such inspection takes time (several weeks or longer), and the tagging is often first preliminary, to be revised at a later time after a preliminary inspection of all buildings has been completed. Such walk-in inspections can detect only damage that is visible, and there is always considerable subjectivity in the assessments. The major problem with such inspections is however the timeliness, as aftershocks following the earthquakes can further damage a structure that has survived the main event but is weakened, and endanger the occupants. Another problem is the loss of function of a structure that may be safe, until a more detailed inspection and assessment is possible. This is particularly important for critical facilities, such as hospitals, as well as for major businesses, such as banks, for which interruption of work can cause major financial losses. Without a doubt, the ability to detect damage in structures early, as it occurs or soon after the earthquake, using some structural health monitoring system, and assess the state of safety of the structure before physical inspection is possible, can benefit society immensely. Ideally, based on instrumental data, such systems would be able to detect also hidden damage that is not visible to the naked eye. There would be benefit even when the damage is obvious, if that information is available immediately after the earthquake. To be effective, however, such systems must be sensitive enough to detect at least the significant damage, and also be accurate enough, to avoid false alarms and unnecessary and costly service interruption.

The objective of this article is to review the basic principles on which such systems operate, and to present some illustrative examples of several robust methods applied to full-scale buildings. This is followed by a discussion of remaining critical issues and directions for future research, in the view of the author.

Literature Review

Earthquake Damage Detection in Structural Health Monitoring Research

Earthquake damage detection in civil structures, such as buildings and bridges, is closely related to structural health monitoring of structures such as light aerospace structures, rotating machinery and offshore platforms, for example, that are of concern to other disciplines. A review

of recent developments in this broader field, as applied to civil and mechanical systems, can be found in Chang et al. [7] and Liu et al. [31]. The earliest, and still the most popular methods for civil structures are those that use data from vibrational sensors, and detect changes in the vibrational characteristics of the structure – frequencies of vibration and mode shapes. Detailed reviews of vibrational methods in the general area of structural health monitoring can be found in a report by Doebling et al. [11], its shorter version as a journal paper [10], and a follow up report by Sohn et al. [40]. Another recent review of the vibrational methods can be found in Carden and Fanning [4].

These detailed reviews conclude that the currently available vibrational methods can determine if the structure has been damaged, but cannot indicate precisely the location of the damage, and are therefore referred to as *global*. Most vibrational methods monitor changes in the *modal* properties of the structures (modal frequencies and mode shapes). The stated difficulties associated with these methods include: (1) the presence of other factors than damage that produce similar effects on the monitored parameters not easy to isolate (e. g. the effects of soil-structure interaction on the measured frequencies of vibration, as well as environmental influences such as temperature and rain; [8,46,47]); (2) the redundancy of the civil engineering structures, which results in low sensitivity of the method (i. e. small change of the overall stiffness and consequently of the measured frequencies) when the damage is localized; and (3) dependence on detailed prior analytical models and/or prior test data for the detection and location of damage (supervised learning), which may not be readily available for a structure, may be outdated, and even when available represent only an idealization of the real structure [7,11]. Further critical issues identified are (4) the scarcity of objective comparisons of different procedures applied to a common data set, and (5) the number and location of sensors (techniques to be seriously considered for implementation in the field should demonstrate that they can perform well for small numbers of measurements). Finally, Doebling et al. [11] conclude that "while sufficient evidence exists to promote the use of measured vibration data for the detection of damage in structures, using both forced-response testing and long-term monitoring of ambient signals, the research needs to be more focused on the specific applications and industries that would benefit from this technology… Additionally, research should be focused more on testing of real structures in their operating environment, rather than laboratory tests of representative structures."

In the follow up review, Sohn et al. [40] mention as outstanding problems: The reliance on analytical models to obtain the structural parameters from the data, not only in methods involving direct inversion, but also in those that use neural networks; and that the damage sensitive features are also sensitive to changes of the environmental and operational conditions of the structures. They mention as one of the most significant improvements since the previous review [11] the signal processing methods that do not rely on detailed analytic models, such as novelty/outlier analysis, statistical process control charts, and simple hypothesis testing (unsupervised learning), shown to be very effective to identify the onset of damage growth, and the presence of damage but not the damage type. In this article, one such method – based on detection of novelties using wavelets – is reviewed and illustrated. Another significant advancement is the availability of more affordable MEMS sensors, as well as fiber optics, and piezoceramic sensors, and of wireless data communication technology.

In structural health monitoring literature, the vibrational methods are referred to as *global*, due to the relatively small number of sensors typically installed in structures, and can detect only significant damage [11,40]. The cost of seismic monitoring systems is still high, and tradeoffs have to be made between the detail of the instrumentation of a particular structure and the number of structures that are instrumented. The truly *local* methods are those for nondestructive testing (NDT) of materials, which can detect the location of cracks or some other defects in a structural member. These methods typically use: (1) ultrasonic waves, which are attenuated quickly along the wave path, (2) need an actuator to create such waves, and (3) require direct access to the structural member, usually not readily available. Consequently, they are used to detect the location of the damage in a particular structural member, known or suspected to have been damaged, but are too costly and impractical for structural health monitoring of an entire structure [7]. To make a difference for society, structural health monitoring and early warning systems have to be reasonably priced so that they can be installed in many structures.

Earthquake Damage Detection in Earthquake Engineering Research

In the earthquake engineering research, earthquake damage detection emerges from system identification studies of full-scale structures (typically involving identification of their frequencies of vibration and damping) from ambient and forced vibration test data, or earthquake records. Consequently, it is *data driven*, in contrast to the structural health monitoring research, which focuses on methodolo-

gies, validated mostly on "clean" numerically simulated data, and sometimes on laboratory data or small amplitude full-scale data. In the US, the earliest system identification studies from full-scale data follow the first deployment of strong motion instruments in structures [3], and continue through the 1960s [9,19,70]. More sophisticated studies from the system identification point-of-view using earthquake response data appear in the 1970s, following the San Fernando, California earthquake of 1971, which produced strong motion records in many buildings in the Los Angeles metropolitan area [66,67,68,69]. A significant finding of these studies is that the building frequencies of actual structures vary significantly as a function of the level of the response. The variation is such that the fundamental frequency decreases during the largest shaking, but recovers afterwards during the remaining smaller amplitude shaking, or during subsequent shaking from aftershocks or small amplitude tests. The recovery may be partial or complete, and a large reduction of frequency of vibration during the earthquake is not always associated with visible damage. This is an important fact, as the decrease of the fundamental frequency of vibration is used as one of the global indicators of damage in structural health monitoring research, and also because many sophisticated structural identification methods are based on the assumption of stationarity and time invariance of the response.

Further, system identification studies of structures using earthquake records, considering the effects of the interaction of the structural vibrations with the vibration of the surrounding soil, appear in the 1960 and 1970s. The most detailed such full-scale studies are probably those of the Millikan library in Pasadena [33,34,71]. Understanding and consideration of the effects of soil-structure interaction in system identification and health monitoring of structures is of *fundamental importance* for the development of reliable methodologies, as this phenomenon is an integral part of the seismic response, and affects the estimation of both the frequencies of vibration and the interstory drift, both used to infer about the state of damage. Nevertheless, these effects are typically ignored in structural health monitoring research. A detailed literature review on full-scale studies of soil-structure interaction can be found in Trifunac et al. [63], and a discussion of critical issues in recording and interpreting earthquake response of full-scale structures can be found in Trifunac and Todorovska [60,61].

Damage and Damage-Sensitive Features

The damage of a structure can be described by the following five states: (1) no damage, (2) repairable (light and moderate) damage, (3) irreparable damage, (4) extreme damage, and (5) collapse [14].

Damage is associated with large deformations of the structural elements (usually expressed via the inter-story drift), which cause yielding of the structural steel or steel reinforcement and cracking of the structural concrete. Also, damage causes changes of the structural vibrational characteristics (frequencies of vibration), and wave propagation characteristics (wave velocities/travel times). This section presents the rationale for damage detection algorithms based on monitoring such changes. The concepts are illustrated on a simple soil-structure interaction model.

Structural Models and Identification

Structure as an Oscillator

From an elementary vibrational viewpoint, a structure responds to earthquake shaking as an oscillator characterized by its frequencies of vibration. The fixed-base frequencies are those of free vibration of the structure on *rigid* ground. They are the eigenvalues of a boundary value problem, and the associated eigenfunctions are referred to as mode shapes in structural engineering. The fixed-base frequencies depend only on the properties of the structure, i. e. on the structural stiffness and mass, while their dependence on the structural damping is small for most structures, which are lightly damped. In the linear range, the response of an n-degree-of-freedom system to earthquake shaking is a superposition of the modal responses. The contribution of the fundamental mode is usually the largest, and in engineering design structures are often represented by an equivalent single degree-of-freedom oscillator. For a single degree-of-freedom oscillator, the natural frequency of vibration is

$$\omega_1 = \sqrt{k/m}\,, \tag{1}$$

where k is its stiffness and m its mass. The frequency of such an oscillator is affected little by typical fluctuations of the mass due to variations in the life load of the structure, and is mostly affected by changes in the stiffness. Damage would cause loss of stiffness, and consequently reduction of the fixed-base frequency of vibration. If $\omega_{1,\mathrm{ref}}$ is a reference frequency corresponding to reference stiffness k_{ref}, then for the damaged structure

$$(\omega_1/\omega_{1,\mathrm{ref}})^2 = k/k_{\mathrm{ref}}\,. \tag{2}$$

As the fixed-base frequency depends on the *overall* stiffness of the structure, it is by definition a *global* property, and would not change much due to localized damage of civil structures, which are designed to be highly redundant. One advantage of detecting damage by monitor-

ing changes in fixed-base frequency of vibration is that, in the ideal case when the ground is practically rigid (as compared to the structure), and the excitation is relatively broadband, the fixed-base frequency can be determined using only one sensor, on the roof, as the frequency of the peak of the Fourier transform of the roof response. The availability of recorded response at ground level would produce a more accurate estimate, as a transfer-function can be computed between the roof and ground level response motion. Changes in the frequency versus time can be estimated from the Fourier transform in moving windows in time.

Buildings are founded on soil, which is flexible and deforms under the action of forces from the incident waves and from the vibrating structure. Even if rigid, a structure founded on soft soil will vibrate, with the soil acting as a spring. The soil adds both *flexibility* and *dissipation mechanism* to the vibrations of the structure and soil, which act as a coupled system. Two sources of dissipation are (1) scattering of the incident waves from the foundation and (2) radiation of energy into the soil (through vibration of the foundation, which acts as a source of waves radiated in semi-infinite medium). The third source of dissipation is in the structure, and includes a distribution of frictional sources, and hysteretic damping during nonlinear response. The soil-structure system has its own resonant frequencies and "damping", which is a combination of the contributions from the structure and from the soil. The fundamental frequency of the system is always lower than the fundamental fixed-base frequency of the structure, but the associated system damping can be larger or smaller than the damping of the structure alone, depending on the radiation damping and relative stiffness of the structure with respect to the soil.

In conclusion, the difficulties with Fourier-type analyses for identification of the building frequencies are that these give the resonant frequencies and equivalent damping of the *system*, which depend on the soil, and that they are global properties. Also, there is no knowledge base of changes in such frequencies (for different types of structures and different types of soils) related to different degrees of damage.

Structure as a Wave Guide

Alternatively, the seismic response can be represented as a superposition of waves that propagate through the structure, reflect from its exterior and interior boundaries and interfere [22,39,41,48,49,50,56,57]. Loss of stiffness due to local damage would cause delays in the wave propagation through the damaged part, which could be detected using seismic response data recorded on each side of the damaged area, along the wave path. A change in wave travel time would depend *only on the changes of the physical properties between the sensors*. Hence, the wave methods are more sensitive to local damage than the modal methods, and should be able to point out the location of damage with a relatively small number of sensors. Additionally, the local changes in travel time are not sensitive to the effects of soil-structure interaction (as demonstrated in [44,45]), which is a major obstacle for the modal methods based on detecting changes in the structural frequencies.

The basic principles of the method are as follows. It is based on D'Alambert's solution of the wave equation, and representation of the structural response as a superposition of waves traveling through the structure. In contrast, the modal methods are based on representation in the Fourier domain, as superposition of modes of vibration.

The wave travel time between two pints

$$\tau = d/V_s \,, \tag{3}$$

where d is the distance traveled and V_s is the equivalent shear wave velocity in the part of the building between the two sensors. The latter is related to the rigidity via

$$V_s = \sqrt{\mu/\rho} \,, \tag{4}$$

where μ is the shear modulus and ρ is the density. Hence, reduction of rigidity due to damage will produce a reduction of the equivalent shear wave velocity, which will produce an increase in the pulse travel time, relative to the travel time for the undamaged state. Let μ_{ref} be reference rigidity, and $V_{s,ref}$ and τ_{ref} be the corresponding shear wave velocity and wave travel time. Then their changes are related as follows

$$\frac{\tau}{\tau_{ref}} = \frac{1}{V_s/V_{s,ref}} = \frac{1}{\sqrt{\mu/\mu_{ref}}} \,. \tag{5}$$

Global changes can also be detected by monitoring the *total* wave travel time from the base to the roof of a building. Let τ_{tot} be the travel time of seismic waves from the point of fixity (ground level) to the roof. Then the building fundamental fixed-base frequency $f_1 = 1/(4\tau_{tot})$ assuming that the building as a whole deforms like a shear beam. Based on this relation, f_1 can be estimated using data from only two horizontal sensors. While the goodness of this approximation of f_1 may vary from one building to another, the changes in $f_1 = 1/(4\tau_{tot})$ will still depend *only* on changes in the building itself, and not on changes in the soil, and monitoring of changes in such an estimate of f_1 can be used as a global indicator of damage in a building [44,45].

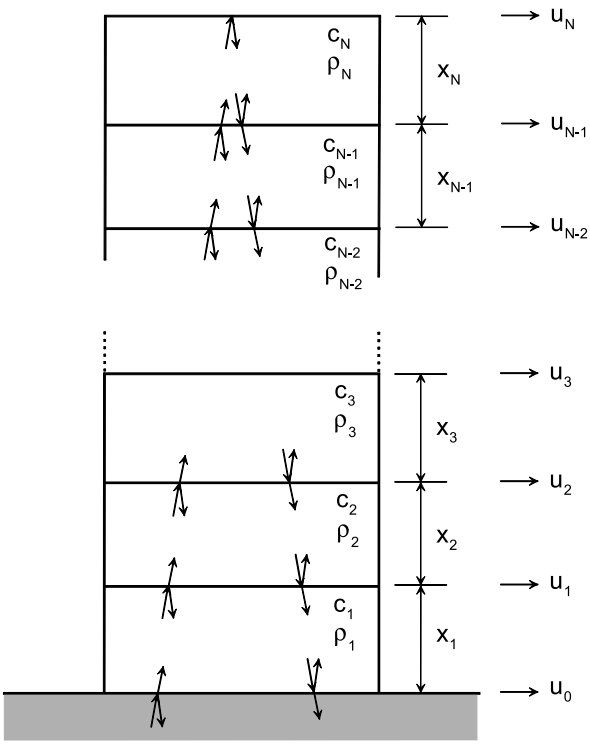

Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 1
Layered building model

Figure 1 shows a conceptual model for the analysis, in which the building is a horizontally layered medium, with the interfaces between layers at the floor slabs. For vertically incident waves, or/and a narrow building, the layered medium will be traversed by waves propagating upward and downward. Let the excitation be a pulse. At each interface, an incident pulse will be split into a reflected pulse, and a transmitted pulse, and at the roof total reflection will occur. The transmission and reflection coefficients will depend on the impedance contrast between the layers, in particular on the shear wave velocities, which will change due to loss of stiffness caused by damage. Because of reflections and material damping, an incident wave pulse will attenuate as it propagates through the structure, and will be also modified in a dispersive medium. This is schematically illustrated in Fig. 1. The total wave motion propagating upward in a layer will be a superposition of all the pulses, those from direct incidence and those from different generations of reflections. The same applies for the pulses propagating downward. The downward propagating pulses that are reflected back into the building, from the interface with the soil, will interfere with the newly incident pulses just transmitted into the building. Eventually, constructive interference will occur, and the standing waves will be formed, which are the fixed-base modes of vibration of the building.

The wave travel times can be detected by tracing the propagation of a pulse. Such a pulse can be created by signal processing of recorded earthquake response data, i. e. by deconvolution of the recorded response, which results in the *system impulse response functions*. These can be obtained by computing the transfer-functions between the motion at a particular level and the reference motion, and then computing inverse Fourier transform. The location of the virtual source would coincide with the location of the sensor that recorded the reference motion. Let $u_{\mathrm{ref}}(t)$ be the reference motion, $u_i(t)$ be the motion at level i. Then the impulse response at that level, $h_i(t)$, can be computed as

$$h_i(t) = \mathrm{FT}^{-1}\left\{\frac{\hat{u}_i(\omega)\bar{\hat{u}}_{\mathrm{ref}}(\omega)}{|\hat{u}_{\mathrm{ref}}(\omega)|^2 + \varepsilon}\right\}, \tag{6}$$

where the hat symbol indicates Fourier transform, the bar indicates complex conjugate, and ε is a regularization parameter, used to avoid dividing by a very small number [39]. At the reference level, the transfer-function is unity, and its inverse is a Dirac delta function.

Proof-of-concept applications to two buildings damaged by earthquakes, and to an analytical model of a building-foundation-soil system showed that the method (1) is robust when applied to damaging levels of earthquake response data, (2) is not sensitive to the effects of soil-structure interaction, and (3) is local in nature (i. e. gave results consistent with the spatial distribution and degree of the observed damage) [44,45,53,54]. The damaged buildings are the former Imperial County Services Building – a 6-story RC structure in El Centro, California, damaged by the 1979 Imperial County earthquake and later demolished [52,54], and the 7-story RC building in Van Nuys, damaged by both the 1971 San Fernando and the 1994 Northridge earthquakes [53,62]. Another application is to a building in Banja Luka in former Yugoslavia, using records of 20 earthquakes, one of which led to levels of response that might have caused structural damage, but no damage was reported following a detailed inspection [65]. This study was aimed at learning about the *threshold change* in the building fixed-base frequency, estimated from wave travel time, associated with damage.

While this method is local, its spatial resolution is limited by the number of sensors. A minimum of two sensors (at the base and at the roof) are required to determine if the structure has been damaged, and additional sensors at the intermediate floors would help point out the part of

the structure that has been damaged. For example, one additional sensor between these two would help identify if the damage has been in the part of the structure above or beyond that sensor.

There have been only a few publications in the literature on wave propagation methods for structural health monitoring and damage detection in civil structures other than the NDT methods [20,35,37,41,53,54,64]. Similar wave travel time analyses (using deconvolution or the NIOM method) of buildings that have not been damaged include Kawakami and Oyunchimeg [24,25], Snieder and Şafak [39], Kohler et al. [26], and Todorovska [44,45]. These studies show that the wave travel times reflect well the characteristics of the buildings studied. A recent review can be found in [53,54].

In conclusion, the advantages of this wave method are its local nature achieved with a relatively small number of sensors, its insensitivity to the effects of soil-structure interaction, and the ability to estimate the structural fixed-base frequency using data from only two sensors (one at the base and one at the roof), which will extend the usability of old data. An outstanding issue to its implementation is the lack of a knowledge base relating changes in wave travel times (and fixed-base frequency) with different levels of damage for different types of structures.

Inter-Story Drift

Structural damage of a building under seismic loads occurs primarily due to large *lateral* deformations of its columns and shear walls, as they are by design much stiffer in the vertical direction to carry the static gravity loads. A measure of the lateral deformations is the inter-story drift. The inter-story drift is also a good indicator of the damage to the architectural (nonstructural) components (partition walls, facade, windows, etc.), which can be costly. As the value of the structure is only about 10–25% of the total construction cost of a building, the damage to the nonstructural components represents a significant portion of the total repair cost following an earthquake. For these reasons, the inter-story drift is one of the performance parameters considered in design. It is important to note that the structural and nonstructural damage are related only to the drift caused by *deformation* of the structure, and not by the drift caused by *rigid body motion*.

The level of structural damage (to a particular element and to the structure as a whole) associated with a particular level of inter-story drift varies depending on the type of structure, height and ductility, among other factors, and is still *not a completely resolved issue in structural engineering* [14]. To illustrate this correlation, Table 1 shows some values of drift associated with different levels of damage (simplified from [14]) for ductile and nonductile moment resisting frames (MRF), and based on experimental data, field observations and measurements and theoretical analyses. (Ductile are those structures that can undergo large nonlinear deformations before failure as opposed to the nonductile ones, which experience quick brittle failure soon after exceeding the linear range of response). It can be seen from Table 1 that, roughly, inter-story drift $> 1\%$ for ductile and $> 0.5\%$ for nonductile moment resisting frames causes damage beyond repair, and drift $> 3\%$ and $> 1\%$ for the same type of frames is significant for life safety.

Earthquake Damage: Detection and Early Warning in Man-Made Structures, Table 1
Drift ratios (in %) associated with various damage levels (based on [14])

State of damage	Ductile MRF	Nonductile MRF
No damage	< 0.2	< 0.1
Repairable damage		
Light	0.4	0.2
Moderate	< 1.0	< 0.5
Irreparable damage	> 1.0	> 0.5
Severe damage, life safe, partial collapse	1.8	0.8
Collapse	> 3.0	> 1.0

Drift-based assessment of the state of damage of a building following an earthquake would require: (1) measurement of the drift during the earthquake shaking, and (2) knowledge base of values of drift associated with different states of damage for the particular structure. The accuracy of the assessment would depend on the accuracy of both the measurements and knowledge base, as discussed in the following.

The drift is commonly estimated from the difference of displacements obtained by double integration of recorded velocities in the structure [28]. While in the past these calculations were performed by specialists, after the data had been manually collected, at present, such calculations can be done in near real time either using telemetry or at the site by "client" software supplied by the instrument manufacturer. Such estimates of drift however are limited by: (1) the inability to estimate reliably the *static* component of the drift associated with permanent deformations (i. e. the drift at $\omega \to 0$), which is not negligible for structures experiencing large deformations in the nonlinear range of response, when damage occurs, and (2) the inability to separate the drift due to deformation of the structure

(which is directly related to damage) from the drift due to rigid body rocking because of inadequate instrumentation.

The inability to estimate reliably the static part of the displacement (and drift) is due to the fact that the traditional (translational) sensors are sensitive also to rotational motions of their support [16,59], which produce errors in the recorded translations and the integrated displacements mimicking permanent displacement [16]. This problem can be solved, by deploying sensors recording all six components of motion (three translations and three rotations) and performing appropriate instrument correction. Such future deployments and their assessment are of interest to and have been advocated by the International Working Group on Rotational Seismology [29].

The *dynamic* (at $\omega > 0$) drift due to deformation of the structure only is not simple to estimate, especially for structures on soft soil, with significant rocking response of their foundation. The rocking motions of the foundation are due to the wave nature of the incident seismic waves, and also due to feedback forces from the structure acting on the soil. The foundation rocking results in relative horizontal displacement between two floors and is not related to damage. Such excessive relative displacements, can affect the stability of the structure, which may collapse before yielding occurs in its members, but that is out of the scope of this article. The *average dynamic* floor rocking can be calculated from the difference of vertical motions recorded by two sensors on that floor, assuming the floor slab is rigid, but such sensor configurations are not routinely installed even in recent denser deployments in buildings. If the building foundation is fairly rigid, the rigid body rocking of the structure can be estimated from two vertical sensors at foundation level. Unfortunately, even such data is lacking for most of the significant earthquake records in buildings, and even in recent dense deployments (e. g. in [6]). It is noted that vertical sensors are also less sensitive to rotation of their support and to cross-axis motion [15,43].

It should be noted here that permanent displacements can be measured directly using GPS (Global Positioning System), and there have been such deployments in long period structures [5]. While GPS measurements are not contaminated by rotation, they are limited by the fact that what is measured are only the roof *absolute* displacements, which makes it impossible to separate the displacement due to deformation of the structure from the rigid body horizontal translation and rocking. The other two limitations in the presently available systems are the small sampling rate (10–20 Hz) and the limited resolution of GPS for civilian applications (± 1 cm horizontally and ± 2 cm vertically; [5]).

Damage estimation algorithms based on published damage versus drift relationships (e. g. in [1]) started to be implemented by manufacturers of strong motion instruments in structural seismic monitoring systems but there is no data yet of their performance. Despite errors in the assessment resulting from the mentioned difficulties, such algorithms are robust when applied to earthquake data and can be useful within a suite of methods.

Matrices like the one in Table 1 [14] can serve as a knowledge base in assessing the class of damage state for a given maximum drift reached. Such matrices are associated with scatter, due to the variability from one structure to another within the same class. Another source of scatter is the source of the data. Because of the limited amount of full-scale earthquake response data, information for such relationships is complemented by laboratory data (e. g. pushover tests). While the drift in the former is the total drift, which includes the drift due to rigid body motion, the drift in the latter is only due to deformation of the structural elements.

In conclusion, outstanding issues in measuring the drifts are: (1) separation of the drift due to deformation of the structure only, and (2) estimation of the static component of the drift. It may be possible to resolve these issues by deploying six degrees-of-freedom sensors. An outstanding issue in the knowledge base is more accurate drift versus damage state relations for specific buildings.

System Identification Considering the Effects of Soil-Structure Interaction – Example

As mentioned earlier, both for frequency-based identification and for damage assessment based on drift, the effects of soil-structure interaction have a significant effect on the reliability of the estimation. This section presents a simple soil-structure interaction model, in which the building is represented as a shear beam. It illustrates the different contributions to the inter-story drift, the difference between fixed-base and apparent building frequencies and their relationship, and the relationship between the model fixed-base frequencies and wave travel times. More detailed analysis can be found in [44,45].

The model is shown in Fig. 2. It consists of a shear beam of height H and fundamental fixed-base frequency of vibration f_1, representing the building, and a rigid foundation of width $2a$ embedded in elastic half-space. The excitation, in general, is an incident wave (plane P and SV or a Rayleigh wave). The motion on the surface of the half-space in the absence of any structures and excavations, acting as scatterers, is commonly referred to as "free-field." The effective motion at the base of the

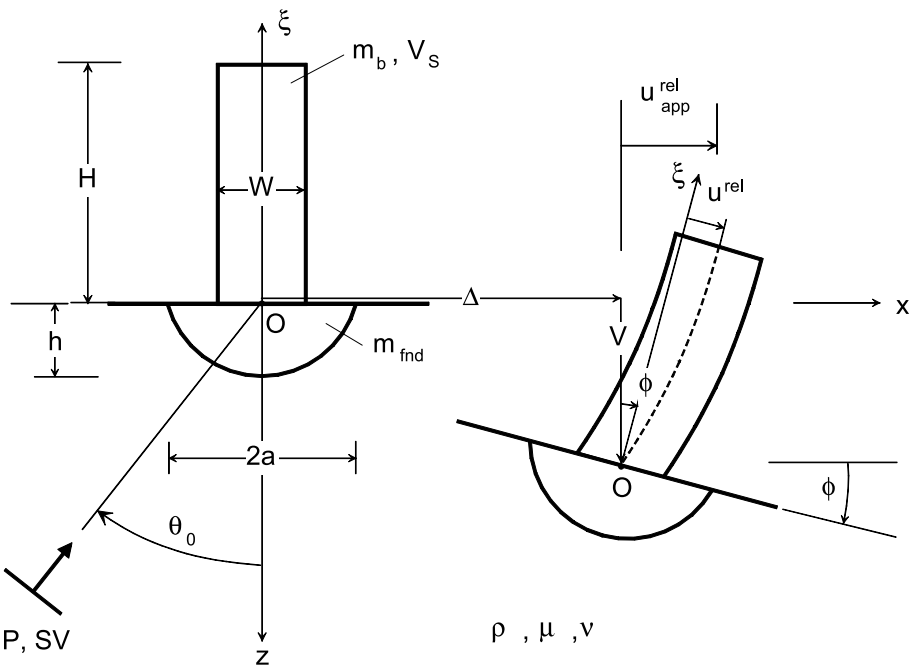

Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 2
Soil-structure interaction model

building differs from the free field motion at the half-space surface, because of two phenomena: (1) scattering and diffraction of the incident waves from the excavation for the foundation, (2) differential displacements due to feedback forces from the building and foundation acting on the half-space through the contact with the foundation. The former phenomenon is referred to as kinematic and the latter as dynamic or inertial interaction. For the linear problem and a rigid foundation, the two problems can be solved separately and their effects superimposed. The apparent frequency, which is the one estimated from peaks of energy distributions of the response, is affected mostly by the dynamic interaction, while the apparent drift is also affected by the type of incident waves, and the point rotation they produce on the ground surface, and the size of the foundation relative to the wavelength of the incident waves. To consider only the effects of the dynamic interaction, it suffices to take excitation consisting of only horizontal foundation driving motion.

The building foundation has three degrees of freedom: horizontal translation Δ, vertical translation V, and rocking angle φ. In the *linear approximation* of the solution, only the horizontal and rocking motions are coupled, while the vertical motions are independent of the other degrees of freedom. Let the excitation be horizontal driving motion Δ_{inp}. Then, the foundation response will be

$$\Delta = \Delta_{\text{inp}} + \Delta_{\text{fb}} \tag{7}$$

$$\varphi = \varphi_{\text{fb}}, \quad \varphi_{\text{inp}} = 0\,, \tag{8}$$

where Δ_{fb} and φ_{fb} are the feedback horizontal displacement and rocking angle, which depend on the stiffness of the foundation and on the forces with which the structure and foundation interact with the soil, and are the solution of the dynamic equilibrium equations of the foundation. The building horizontal displacement $u(\xi)$, as a function of the height ξ measured from ground level, is a sum of three terms

$$u(\xi) = \Delta + \varphi\xi + u^{\text{rel}}(\xi)\,, \tag{9}$$

where the first two terms are from the translation and rotation as a rigid body, and the third term is relative displacement from *deformation* during the vibration. The damage in the building will depend only on $u^{\text{rel}}(\xi)$. It is noted here that including the coupling between horizontal and vertical motions turns a linear elastic system into a nonlinear elastic system [21].

For the shear beam, $u(\xi)$ can be computed as a solution of the wave equation for moving boundary conditions. It can be represented as a sum of motions $u_\Delta(\xi)$, which is

due to translation of the base only, and $u_\varphi(\xi)$, which is due to rotation of the base only, where

$$u_\Delta(\xi) = \Delta \frac{\cos k_S(H-\xi)}{\cos k_S H} \tag{10}$$

$$u_\varphi(\xi) = \frac{\varphi}{k_S} \frac{\sin k_S \xi}{\cos k_S H}\,, \tag{11}$$

where $k_S = \omega/V_S$ and $V_S = \sqrt{\mu_b/\rho_b}$ is the shear wave velocity in the building. Equations (10) and (11), reflecting the interference conditions in the building, imply fundamental fixed-base frequency of the structure $f_1 = V_S/(4H)$ and overtones at $f_n = (2n-1)\,V_S/(4H)$, $n > 1$. If τ is the time it takes for a wave to propagate from the base (at $\xi = 0$) to the top (at $\xi = H$), the interference conditions in the shear beam imply

$$f_1 = 1/(4\tau)\,. \tag{12}$$

Let us now consider the frequencies of vibration. If the building did not deform, the foundation and the building would oscillate freely as a rigid body with frequency f_{RB} such that

$$\frac{1}{f_{RB}^2} = \frac{1}{f_H^2} + \frac{1}{f_R^2}\,, \tag{13}$$

where f_H and f_R, referred to as the horizontal and rocking foundation frequency, depend on the stiffness of the foundation and on the system mass [33]. If the building is flexible and would freely vibrate on a fixed base with fundamental frequency f_1, on flexible soil it would freely vibrate with fundamental frequency f_{sys}, which is the soil-structure *system* frequency, and is a result of the coupling between the vibration of the building and the vibrations of the foundation. The following relationship holds approximately

$$\frac{1}{f_{sys}^2} = \frac{1}{f_{RB}^2} + \frac{1}{f_1^2}\,. \tag{14}$$

This relationship implies that $f_{sys} < \min\left(f_1, f_{RB}\right)$, i. e. f_{sys} is always lower than both f_1 and f_{RB}, and that if f_1 and f_{RB} differ significantly, then f_{sys} would be closer to the smaller one of them. How much f_1 would differ from f_{sys} would depend on the *relative* stiffness of the soil compared to the building. The energy of the response of vibrating systems is concentrated around their resonant frequencies, which are measured from the frequency of the peaks of the corresponding transfer-functions. The energy of the building roof response (absolute and relative) will be concentrated around $f = f_{sys}$.

Of interest is how to estimate the relevant quantities from recorded response during an earthquake. If the building foundation is fairly rigid, and there are at least two appropriately located vertical sensors to compute the foundation rocking φ (average value), then $u^{rel}(\xi)$ can be computed. To measure f_{sys}, the driving motion Δ_{inp} is also needed, so that the transfer function between the building response and Δ_{inp} can be computed. Motion from a nearby free-field site can be used for that purpose, but such sites are often not available, and truly free-field sites practically do not exist in urban areas. Also, for most instrumented buildings, the foundation rocking cannot be estimated because of the lack of two vertical sensors even under the ideal conditions that the foundation behaves as rigid.

Consequently, in reality, for most instrumented buildings, the true relative roof displacement cannot be estimated from the recorded data, but only the *apparent* relative displacement

$$\begin{aligned} u_{app}^{rel}(H) &= u(H) - \Delta \\ &= u^{rel}(H) + \varphi H \end{aligned} \tag{15}$$

which includes the contribution of the roof displacement due to rigid body rotation, and only the transfer-function $|u_{app}^{rel}(H)/\Delta|$ can be computed, the peak of which gives the *apparent* building frequency f_{app}, which is different from both the fixed base frequency and the system frequency.

What is of interest for structural health monitoring is that the energy of the roof response will be concentrated around $f = f_{sys}$, not around $f = f_1$. It is also significant that the damage will depend on $u^{rel}(\xi)$, while what is usually measured is $u^{rel}(\xi) + \varphi H$.

Figure 3 (redrawn from [45]) shows a comparison of model and measured transfer-functions for a model of the NS response of Millikan library. The model has $f_1 = 2.5$ Hz, height $H = 44$ m, shear wave velocity in the soil 300 m/s, and Poison ratio 0.333, while the data are from the Yorba Linda earthquake of 2002. Figure 4 shows the corresponding impulse response function for a virtual source at the ground floor. It can be seen that there is a very good qualitative agreement despite the model simplicity and roughly chosen parameters.

Novelty Detection in the Recorded Response

Novelty detection is used in data mining to detect unusual events in data. The unusual events are *outliers* deviating from the *trend*. Within the framework of multi-resolution analysis, the trends and novelties are determined by

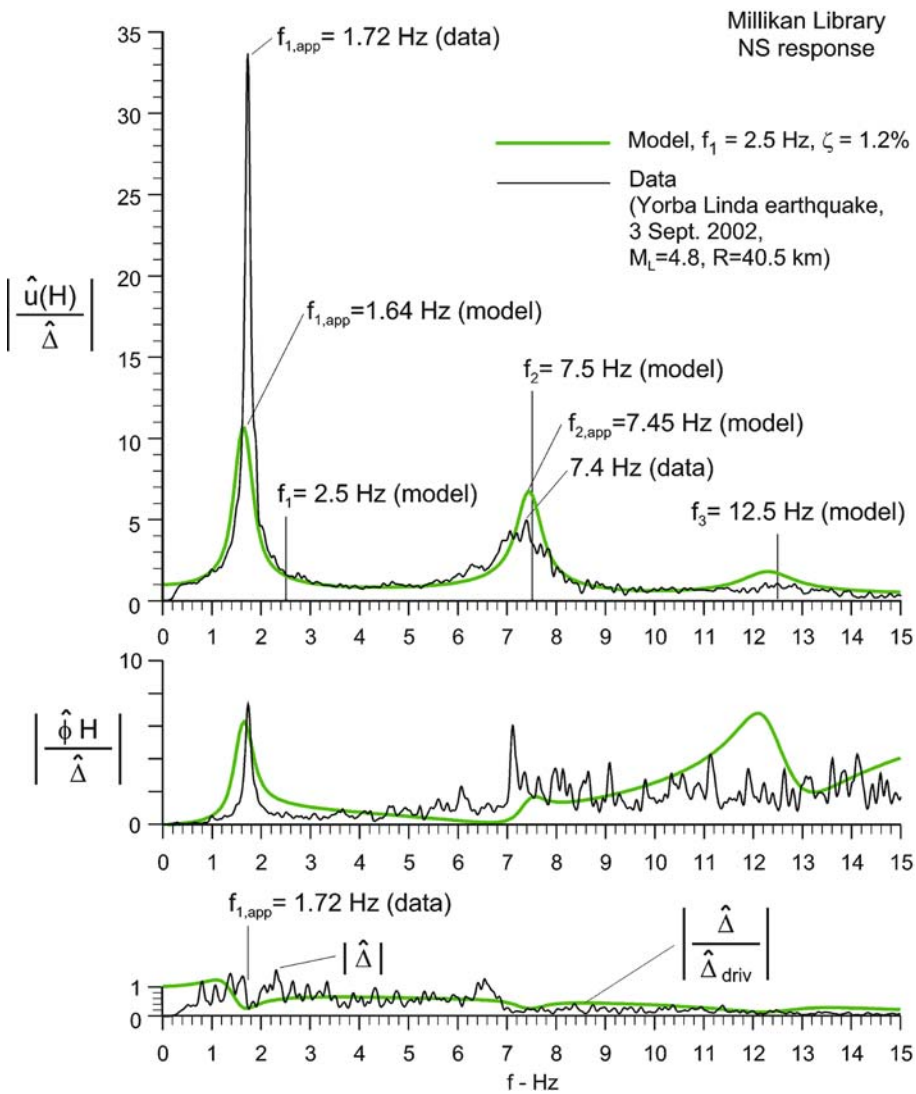

Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 3
Model (*thick line*) versus Yorba Linda, 2002, earthquake (*thin line*) NS response: transfer-functions of roof response (*top*), and base rocking response (*middle*) with respect to horizontal response of ground level. The plot in the *bottom* shows the model horizontal response at ground level for unit driving motion (*thick line*), and the Fourier spectrum of the earthquake response at ground level (*thin line*) on a relative scale

splitting the signal in two subbands, one smooth (low frequency) and the other one containing the detail (high frequency). By consecutively splitting the smooth subband, trends and detail are obtained at different resolution levels. If J is the last level, then there will be J detail subbands D_i, $i = 1, \ldots, J$ and one smooth subband S_J. The last smooth subband can be expanded in a basis of scaling functions $\varphi_{J,k}(t)$, and each of the detail subbands – in a basis of wavelet functions $\psi_{j,k}(t)$, leading to the representation of a discrete time signal $s[n]$, $n = 1, \ldots, N$

$$s[n] = \sum_{j=1}^{J} D_j[n] + S_J[n] = \sum_{j=1}^{J} \sum_{k=1}^{N/2^j} d_{j,k}\psi_{j,k}[n] + \sum_{k=1}^{N/2^J} s_{J,k}\varphi_{J,k}[n] \,. \quad (16)$$

The coefficients of the expansion, $d_{j,k}$ and $s_{J,k}$, can be computed using the fast wavelet transform. The pyramid algorithm on which it is based [36] is shown in Fig. 5.

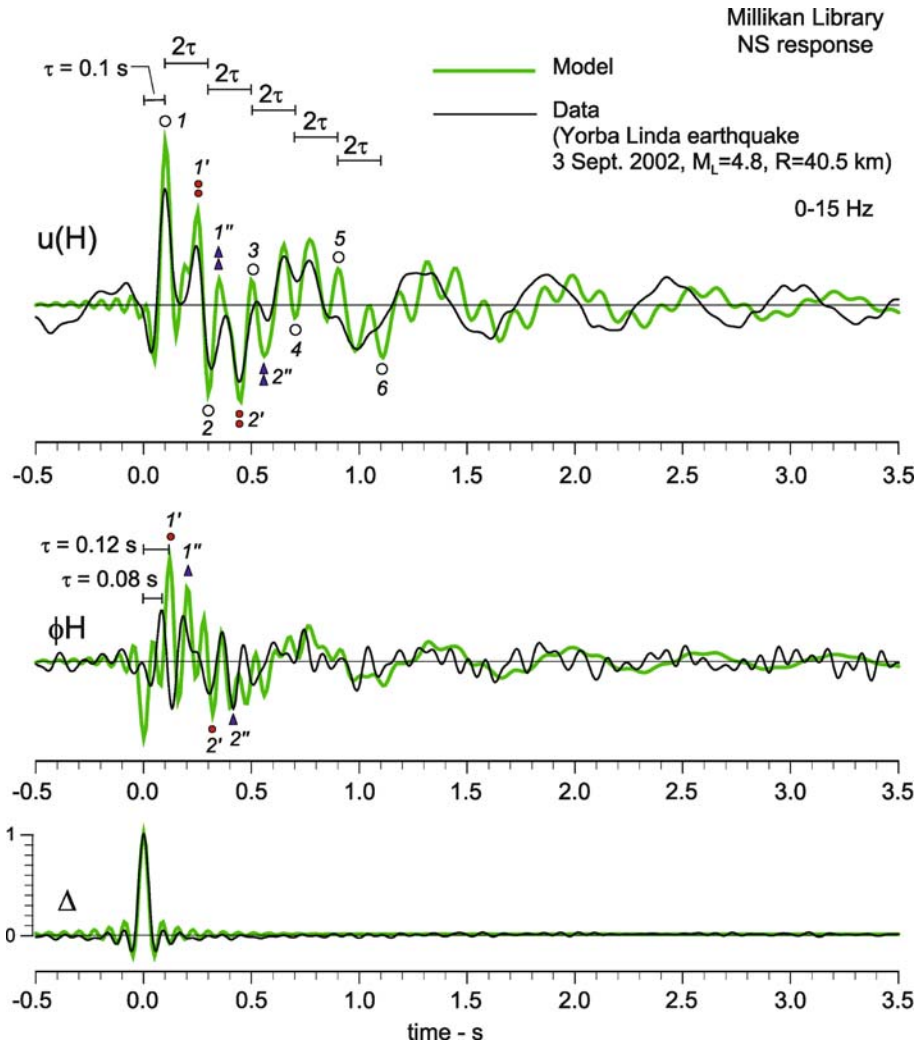

Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 4
Model (*thick line*) versus Yorba Linda, 2002, earthquake (*thin line*) NS response: impulse responses of roof (*top*), and base rocking (*middle*) to input impulse at ground level (*bottom*), i. e. roof horizontal motion, foundation rocking and ground level horizontal motions deconvolved with the resultant horizontal motion at ground level

The wavelet functions $\psi_{j,k}[n]$, where j is a level and k is the time shift, are localized both in frequency and in time, and each wavelet is a projection of the signal onto the corresponding tile of the phase plane. For a wavelet basis that is orthonormal, the square of a wavelet coefficient represents the energy of the signal in the corresponding tile of the phase plane.

The damage detection method is based on the assumption that, when damage occurs and there is a sudden loss of stiffness, *there will be some abrupt change in the response that would produce novelties*. These would be seen as spikes in the time series of the square of the detail coefficients (e. g. $d^2_{1,k}$, $k = 1, \ldots, N/2$ for the highest detail coefficients) plotted versus the central time of the corresponding wavelet. These spikes indicate high frequency energy in the response. For data with Nyquist frequency 25 Hz, the novelties can be best seen in the highest detail subband (12.5 to 25 Hz), which is away from the frequency of the first few modes of typical buildings, where the response is amplified by the structure.

Applications to numerically simulated response of simple models with postulated damage [17,18] have shown that this method *can point out very precisely the time of damage*, but the changes are detectable only if the spikes

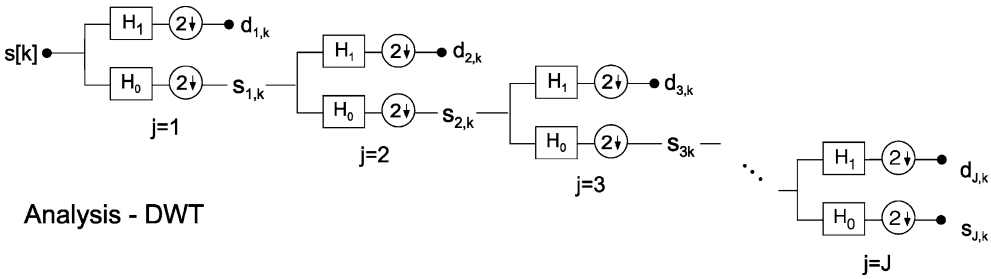

Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 5
The pyramid algorithm for the fast wavelet transform

Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 6
Imperial County Services (ICS) building: a view (towards north); b photographs of damage: columns F1 and F2 at the ground floor; and c column F1

in the wavelet coefficients are above the noise. Further, the magnitude of the novelties is larger if the sensor is closer to the location of the damaged member, and may be difficult to detect if the sensor is far from the location of damage. There have been only few applications to earthquake response records in buildings. Rezai et al. [38] and Hou et al. [18] have shown that there *are* novelties (spikes) in earthquake records of damaged buildings, but have not discussed and extracted other possible causes. Todorovska and Trifunac [51,55] presented a detailed analysis of the correspondence between the spatial distribution and amplitudes of the detected novelties and the observed damage for the Imperial County Services building (see illustrations in Sect."Examples"), and also analyzed the "noise." Their study shows that: (1) the spatial distribution and magnitudes of the novelties were generally consistent with the spatial distribution and degree of the observed damage, (2) the timing of those suggesting major damage agreed with the time of significant drops in frequency and of large inter-story drifts, and (3) were much larger in the transverse response, in which the building was stiffer.

In summary, the method of novelties is very effective in determining the time of occurrence of damage, and can reveal the spatial distribution and degree of damage if there is sufficiently dense instrumentation. Unresolved issues are how to distinguish novelties that are not caused by damage, and small novelties due to larger damage far from the sensor from those due to small damage close to the sensor.

Examples

In this section, the methods previously described are illustrated for the former Imperial County Services (ICS) building – a rare example of an instrumented building damaged by an earthquake, for which description of damage and the strong motion data are available. The building is first described and the strong motion data of the Imperial Valley earthquake, which severely damaged the building.

The ICS building was a 6-story reinforced concrete structure located in El Centro, California (Fig. 6a). It was designed in compliance with the 1967 Uniform Building Code, and its construction was completed in 1969. It had plan dimensions 41.70 × 26.02 m, height 25.48 m, and pile foundation. Up to depth of 9 m, the underlying soil consisted of soft to medium-stiff damp sandy clay with organic materials, with inter-layers of medium dense moist sand, and beneath 9 m it consisted of stiff, moist sandy clay and silty clay [27].

The building was severely damaged by the Imperial Valley earthquake of October 15, 1979 ($M = 6.6$), and was

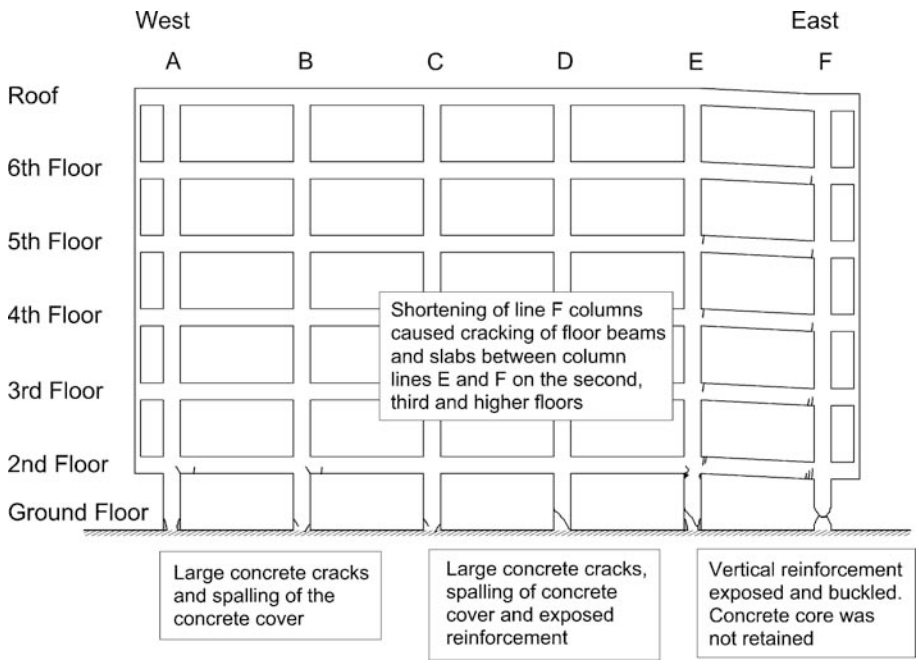

Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 7
ICS building: schematic representation of the damage following the 1979 Imperial Valley earthquake (reproduced from [27])

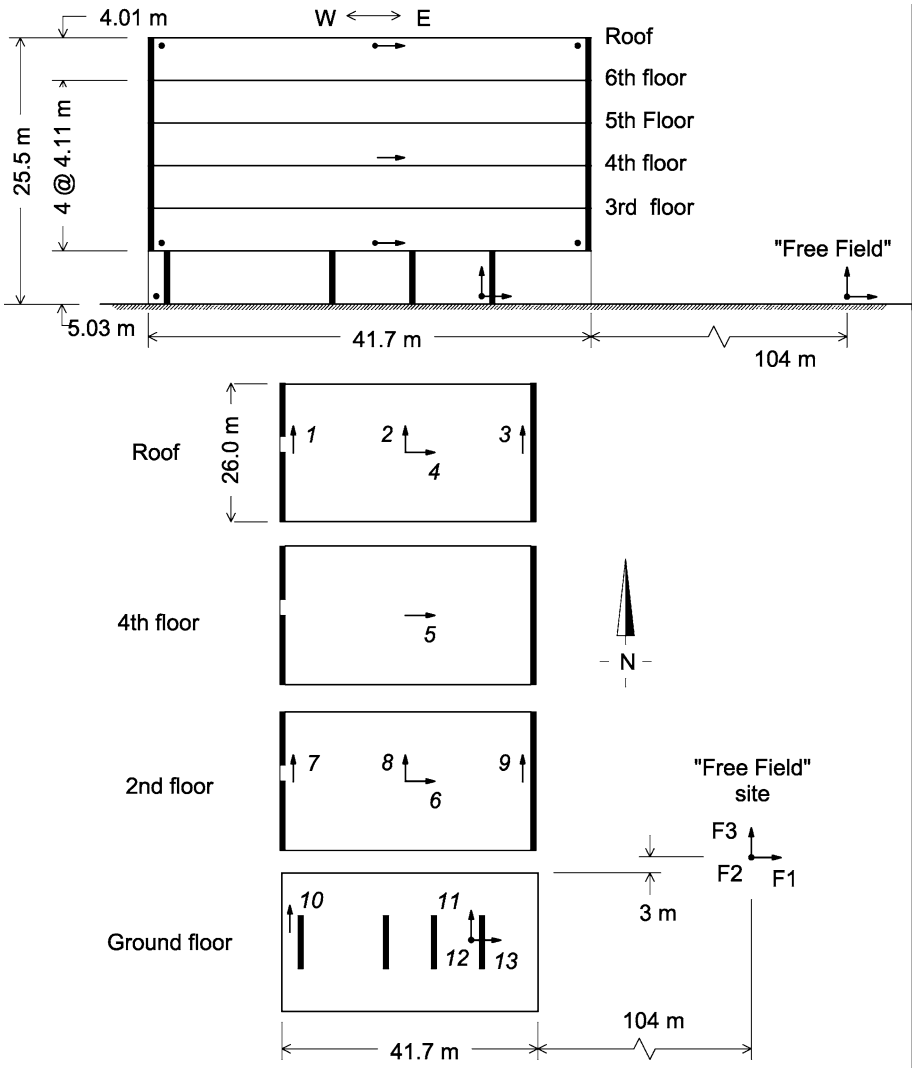

Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 8
ICS building: layout of the seismic monitoring array

later demolished (Fig. 6b,c). Figure 7 shows a schematic representation of the observed damage. The major failure occurred in the columns of frame F (at the east end of the building) at the ground floor. The vertical reinforcement was exposed and buckled, and the core concrete could not be contained, resulting in sudden failure and shortening of the columns subjected to excessive axial loads. This in turn caused an incipient vertical fall of the eastern end of the building, causing cracking of the floor beams and slabs near column line F on the second, third and higher floors. Columns in lines A, B, D, and E also suffered damage. Columns in frames A and E did not suffer as extensive damage as shortening and buckling of the reinforcement in line F at the east side, but large concrete cracks and exposed reinforcement could be seen near the base. In the columns in interior frames B through E, visible cracks and spalling of the concrete cover were also observed [27].

The building was instrumented by a 16-channel seismic monitoring array (installed by the California Geological Survey, formerly the California Division of Mines and Geology) consisting of a 13-channel structural array of force balance accelerometers (FBA-1), with a central analog recording system, and a tri-axial SMA-1 accelerometer in the "free field," approximately 104 m east from the northeast corner of the building (Fig. 8). Figure 9 shows the accelerations (corrected) during the Impe-

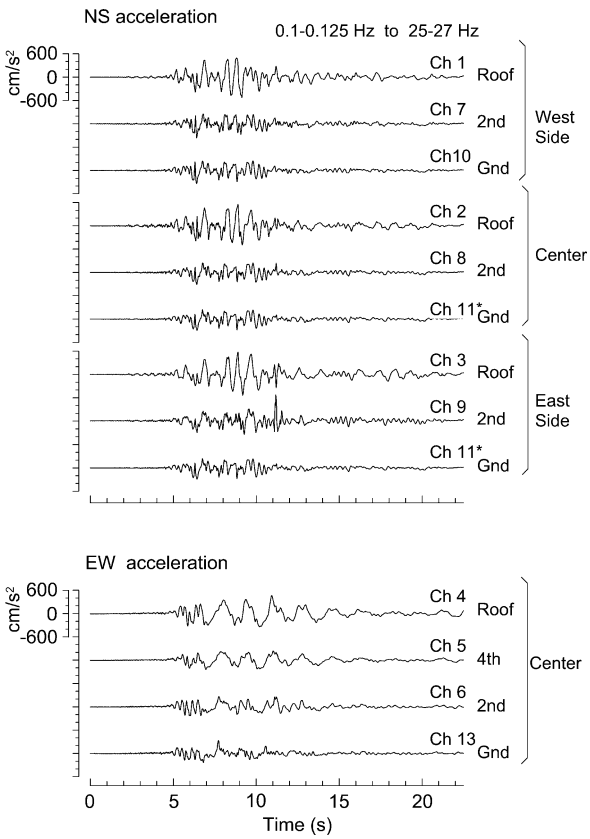

Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 9
ICS building: accelerations (NS and EW components) recorded during the 1979 Imperial Valley earthquake

rial Valley earthquake. The peak accelerations at the roof and ground floor were 571 cm/s^2 and 339 cm/s^2 in the NS direction and 461 cm/s^2 and 331 cm/s^2 in the EW direction.

Figure 10 shows the NS (top) and EW (bottom) inter-story drifts computed from band-pass filtered displacements (between 0.1–0.125 Hz and 25–27 Hz) (redrawn from [52]). Hence, they represent only a limited view of the actual drifts – through a tapered window in the frequency domain, and a combination of the drift due to *rigid body rocking* (one of the effects of soil-structure interaction) and drift due to relative *deformation* of the building. The horizontal lines show 0.5%, 1%, and 1.5% drift levels. The plotted drifts suggest: (1) "soft" first story in both NS and EW directions, (2) larger flexibility in the EW direction, and (3) significant torsional response, probably amplified by the wave passage, and by the asymmetric distribution of stiffness in the NS direction at the soft first story (see Fig. 8). It can be seen that during the most severe shaking, the inter-story drifts exceeded 0.5% for NS and 1.5% for EW motions, consistent with irreparable to severe damage (Table 1).

Figure 11 shows results of time frequency analysis (using Gabor transform) for the EW response (redrawn from [52]). Parts a and b show the ground floor accelerations, and the roof relative displacements (at the center of the building), both included as background information. Part c shows the skeleton (the thicker line), which is a smoothed estimate of the amplitude envelope of the estimated signal, which is the relative roof response near the first system frequency. The thin line is the actual amplitude envelope (that for the broad-band signal), determined by Hilbert transform. This plot is included to help monitor rapid changes in the amplitude of the signal and artifacts in the estimate of instantaneous frequency caused by violations of the asymptoticity condition. Part d shows the Fourier spectra of the relative roof displacement (the solid line), and of the ground floor acceleration (the dashed line, on a relative scale), both included as background information. Part e shows the variations of the system frequency as a function of amplitude of response (estimated from the ridge and skeleton of the Gabor transform), with the arrows indicating the direction of increasing time. Part f shows the variations of the system frequency versus time, estimated from the ridge of the Gabor transform. The missing segments and the dashed lines in parts e and f correspond to time intervals where the estimates cannot be obtained or are not believed to be reliable, due to rapid variations of the envelope of the amplitude, and/or very weak "signal." The rectangle in part f with sides $2\sigma_t = 1.42$ s and $2\sigma_\nu = 0.22$ Hz illustrates the theoretical uncertainty of the estimates due to the finite resolution of the Gabor transform. In practice, the uncertainty is larger due to violations of the asymptoticity assumption. Finally, the numbered open dots (occurring at different times in parts b, c, e, and f correspond to some characteristic points in time associated with changes in amplitude or frequency, as well as a few other points in-between. It can be seen that the EW frequency dropped rapidly from $\nu \approx 0.88$ Hz at $t \approx 3.5$ s to $\nu \approx 0.67$ Hz at $t \approx 7$ s ($\Delta\nu \approx 0.21$ Hz $> \sigma_\nu$; $\Delta\nu/\nu \approx 24\%$), and then continued to drop gradually to $\nu \approx 0.53$ Hz at $t \approx 17$ s ($\Delta\nu \approx 0.14$ Hz $\approx \sigma_\nu$; $\Delta\nu/\nu \approx 20.9\%$).

Figure 12 shows results of impulse response analysis for the EW response (redrawn from [53]). The different types of lines correspond to different time intervals of the recorded motion, before, during, and after the major damage occurred: $t < 7$ s, $7 < t < 13$ s, and $t > 13$ s (based on novelty analysis, discussed below). The plots on the left correspond to an input impulse at the ground floor, and those on the right – to an input impulse at the top. The

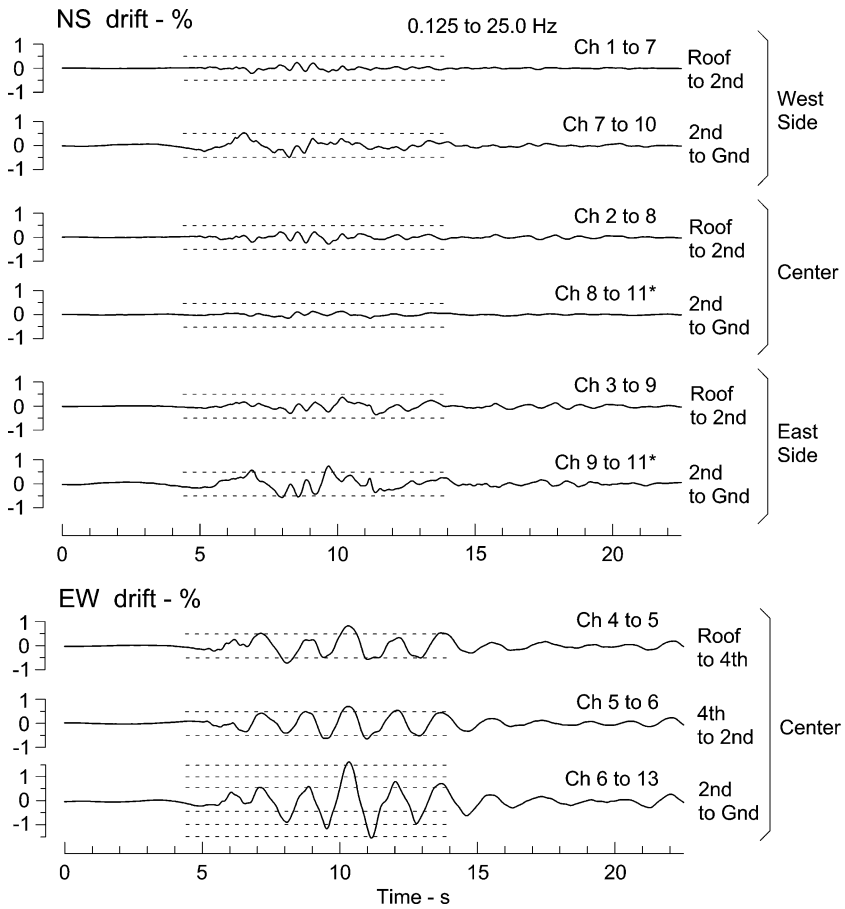

Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 10
ICS building: inter-story drifts during the Imperial Valley earthquake

latter plots show two waves propagating downwards, one acausal (in negative time, representing the wave going up) and one causal (in positive time). The delays in the pulse arrival during the second and third time interval are obvious, and are consistent with the occurrence of damage, as determined using other methods. The wave travel times suggest, for EW motions initial wave velocities of 201 m/s through the first floor, 183 m/s between the 2nd and 4th floors, and 111 m/s between the 4th floor and roof. The velocity of an equivalent uniform shear beam is 142 m/s. Figure 13 shows the corresponding reduction of stiffness. It can be seen that, for EW motions, the reduction was the largest in the first story (80% during the second time window), but was also large in the upper stories (72% between the 2nd and 4th floors, and 60% between the 4th floor and roof). This is consistent with the spatial distribution of the observed damage (Fig. 6), which was the largest in the first story.

Figure 14 shows the results of novelty analysis, for the EW accelerations (part a) and for the NS accelerations at the east side of the building, where the most severe damage occurred (part b) (redrawn from [55]). The inter-story drifts (in %) between the corresponding stories are also shown, by a solid line for NS and by a dashed line for EW motions. Selected novelties are identifies by letters. Novelties T1–T3 are believed to be caused by damage, and are seen in all channels. Novelties G1–G3 and g1–g4 originate in the ground motion, and L1–L6 are possibly caused by local damage close to the sensor, or by other causes. By far the largest novelty is T3, which has amplitude more than an order of magnitude larger than all other novelties in the NS acceleration at the 2nd floor at the east side of the building, where the most severe damage (failure of the first story columns of frame F) occurred. The timing of T3 suggests that the collapse of the columns of the first story occurred at about 11.2 s after trigger. The other two large

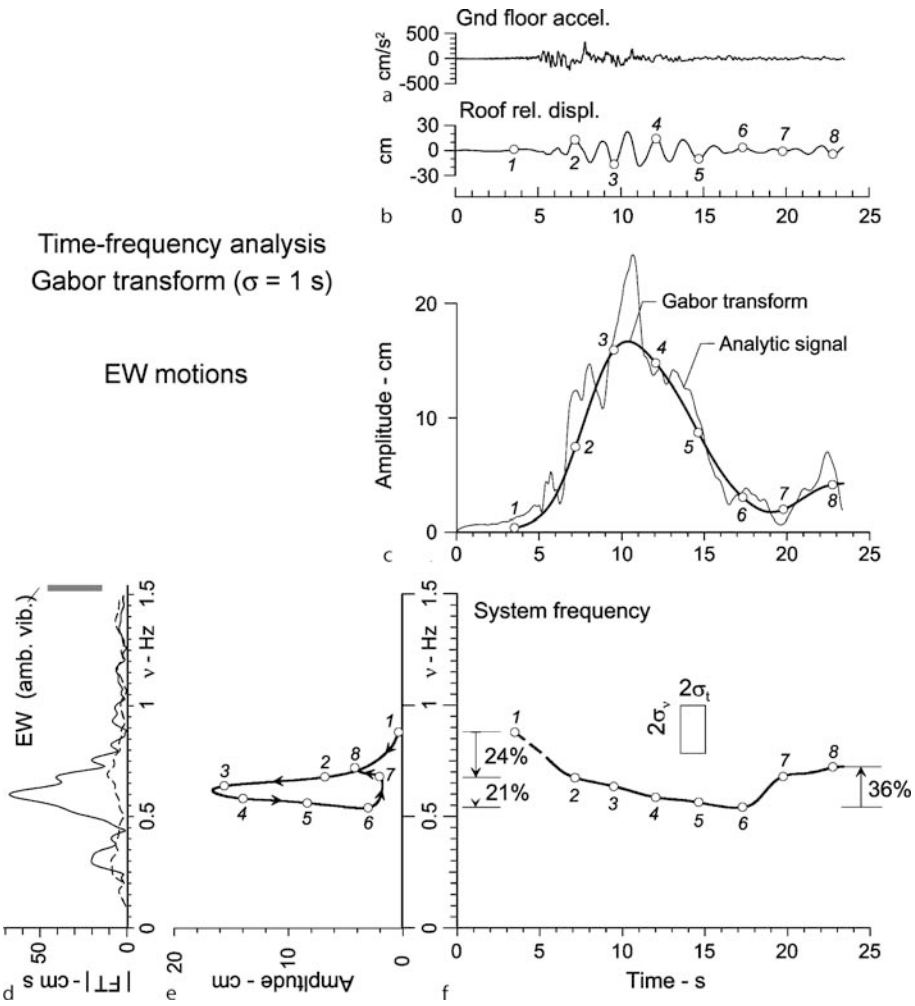

Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 11
ICS building: time frequency analysis for EW response

novelties consistent with the observed damage, T1 and T2, occurring at about 8.2 s and 9.2 s after trigger, indicating damage that weakened the structure, before the collapse of the first story columns.

Figure 15 (redrawn from [54]) shows a comparison of different values of frequency for EW motions: f_1 from wave travel times (the gray line), system frequency f_{sys} estimated from time-frequency analysis (the red line; [52], and f_1 using ETABS models [27]. T1, T2, and T3 mark the times of occurrence of major damage, as indicated by novelties in the response [51,51]. It can be seen that f_1 from wave travel times is consistent with the results of other independent studies.

Finally, Fig. 16 shows results for another building, the Van Nuys 7-story hotel, which has been damaged by earthquakes [53]. It shows a comparison of fixed-base frequency f_1 during 11 earthquakes estimated from wave travel times, and system frequency f_{sys} during the same earthquakes estimated by time frequency analysis (Gabor transform), as well as estimates of f_{sys} during ambient vibration tests. The analysis shows that, during the San Fernando earthquake, f_1 decreased by about 40% (relative to its value within the first 5 s from trigger), which corresponds to a decrease in the global rigidity of about 63%. During the Northridge earthquake, f_1 decreased by about 22% (relative to its value within the first 3 s from trigger), which corresponds to a decrease in the global rigidity of about 40%. The analysis also showed that, although f_{sys} was always smaller than f_1, their difference varied, contrary to what one could expect from a linear soil-structure

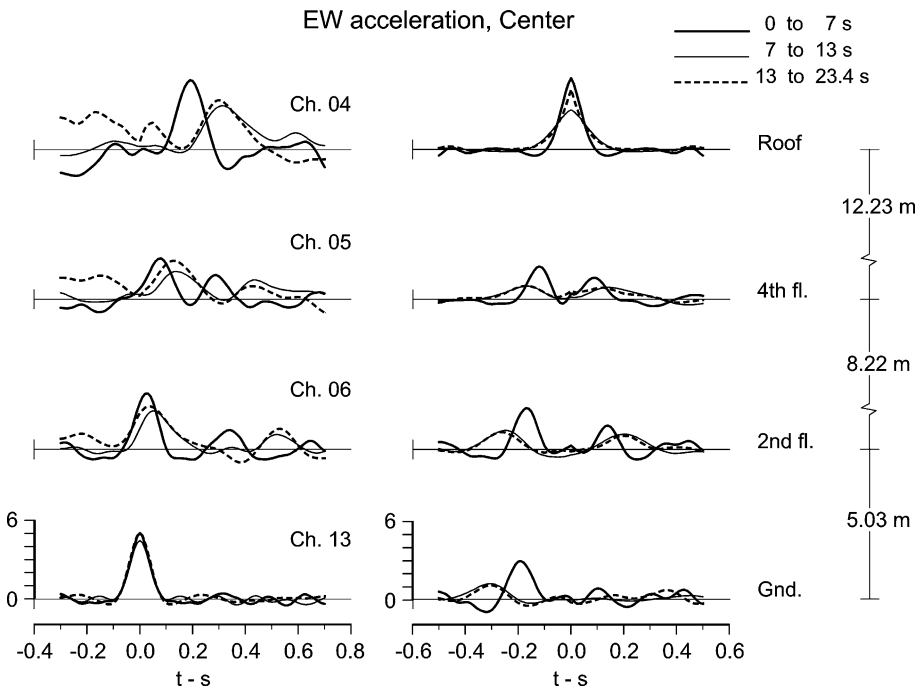

Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 12
ICS building: Impulse response analysis and wave travel times for EW response, and for a virtual source at the ground floor (*left*) and at the roof (*right*)

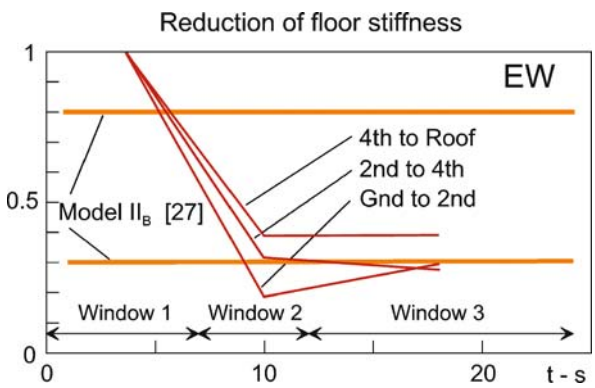

Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 13
ICS building: reduction of floor stiffness versus time [54]

interaction model. It also showed that while f_{sys} was significantly lower during the Landers and Big Bear earthquakes, compared to the previous earthquakes, f_1 did not change much, with is consistent with the fact that these earthquakes (which occurred about 200 km away from the building) did not cause any damage. The study concluded that monitoring changes in f_{sys} can lead to false alarms about the occurrence of damage, and that f_1, as estimated from wave travel times by the proposed method, is a much more reliable estimator of damage.

Future Directions

A successful system for earthquake damage detection and early warning would involve applications of technologies in fields other than structural mechanics and engineering, such as sensing, data communication, signal processing, artificial intelligence, and decision analysis. The end of the 20th and the beginning of the 21st centuries have been marked by a revolution in the development and affordability of the technologies in these other fields. Much research in structural health monitoring for civil structures has been directed towards *adaptation* of these technologies to civil structures. The remaining challenge is to develop a system that is robust, redundant and well calibrated, which will neither miss significant damage nor produce many false alarms. Achieving this would require focusing the efforts and resources to further develop those methodologies that are robust when applied to real structures and data, and to calibrate them using documented full-scale data. Further enhancement of the spatial resolution of such methods would benefit from inexpensive and reliable new sensors.

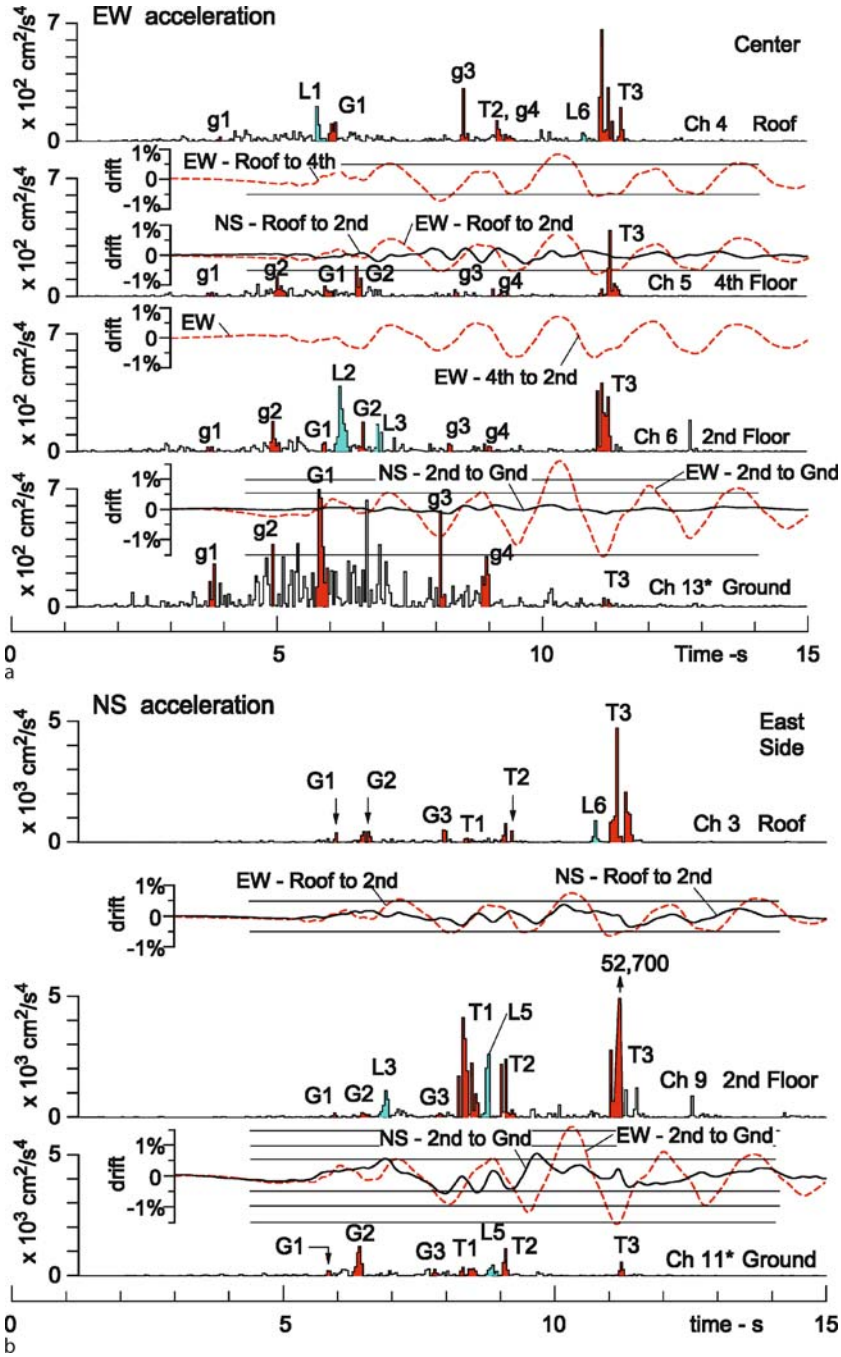

Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 14

a ICS building: novelties analysis of the EW accelerations at the center of the building. **b** Same as Fig. 13a but for the NS accelerations at the east end of the building

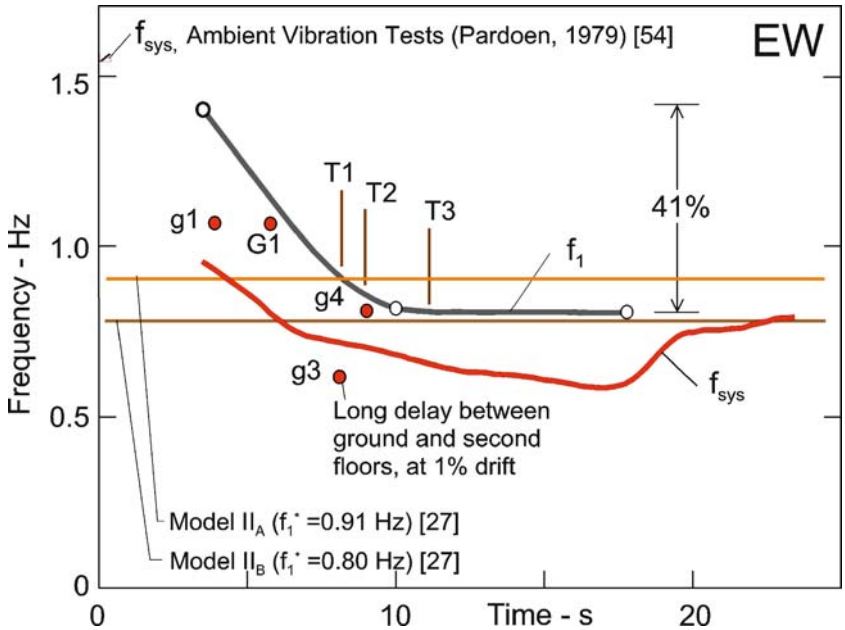

Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 15
ICS building: comparison of results for EW motions from different methods. System frequency f_{sys} from time frequency analysis, fixed base frequency f_1 from wave travel time analysis, and times of occurrence of novelties, T1, T2, and T3

All this will have to be accomplished by continuously expanding our experience in dealing with the complexities of metastable damage states of engineering structures, which will gradually become more feasible with the formulation of realistic physical models. Nevertheless, the practical outcome of most approaches in engineering will probably remain empirical. Also, the art of dynamical modeling will have to be further developed, especially for the assessment of the damaged states of engineering structures that are highly nonlinear and chaotic. In the end, in structural health monitoring, and in design of earthquake resistant structures, the fact that some modeling problems will remain will have to be accepted. However, considerable progress will be achieved if the success is gauged by the degree to which the predictions match observations in the full-scale structures, contributing towards safety and minimization of disruption and productivity of society in seismically active regions.

Bibliography

1. Applied Technology Council (1989) Procedures for post-earthquake safety evaluation of buildings. Report ATC-20. Redwood City
2. Beltrami E (1987) Mathematics for Dynamic Modeling. Wiley, New York
3. Carder DS (1936) Vibration observations. In: Earthquake Investigations in California 1934–1935. US Dept. of Commerce, Coast and Geological Survey, Special Publication No 201. Washington DC, pp 49–106
4. Carden EP, Fanning P (2004) Vibration Based Condition Monitoring: a Review. Struct Health Monit 3(4):355–377. doi:10.1177/1475921704047500
5. Celebi M, Sanli A (2002) GPS in pioneering dynamic monitoring of long-period structures. Earthq Spectr 18(1):47–61
6. Celebi M, Sanli A, Sinclair M, Gallant S, Radulescu D (2004) Real-time seismic monitoring needs of a building owner–and the solution: a cooperative effort. Earthq Spectr 20(2):333–346
7. Chang PC, Flatau A, Liu SC (2003) Review paper: health monitoring of civil infrastructure. Struct Health Monit 2(3):257–267
8. Clinton JF, Bradford SK, Heaton TH, Favela J (2006) The observed wander of the natural frequencies in a structure. Bull Seism Soc Am 96(1):237–57
9. Crawford R, Ward HS (1968) Determination of the natural periods of building. Bull Seism Soc Am 54(6A):1743–1756
10. Doebling SW, Farrar CR, Prime MB (1998) A summary review of vibration-based damage identification methods. Shock Vib Dig 30(2):91–105. doi:10.1177/058310249803000201
11. Doebling SW, Farrar CR, Prime MB, Shevitz DW (1996) Damage identification and health monitoring of structural and mechanical systems from changes in their vibration characteristics: a literature review. Report LA-13070-MS. Los Alamos National Laboratory, Los Alamos
12. Farrar CR, Worden K (2007) An introduction to structural health monitoring. Phil Trans R Soc A 365:303–315. doi:10.1098/rsta.2006.1928

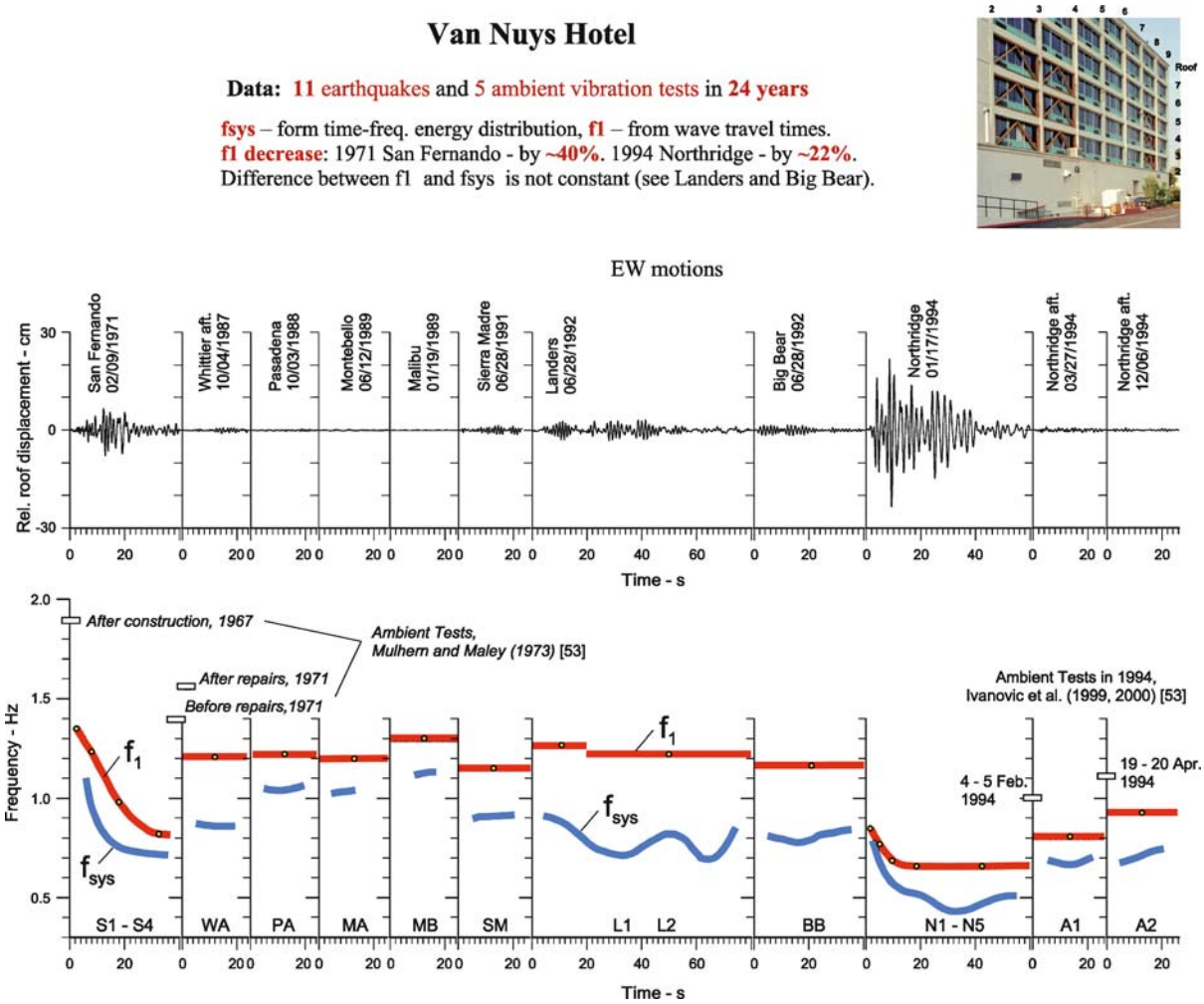

Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 16

Variations of f_1 and f_{sys} in the Van Nuys building during the 11 earthquakes, between February of 1971 and December of 1994. Measured values of f_{sys} during five ambient vibration tests: (1) in 1967, following construction, (2) in 1971, after San Fernando earthquake and before repairs, (3) in 1971 after the repairs, (4) in January of 1994, 18 days after the Northridge earthquake, and (5) in April of 1994, after the building was restrained by wooden braces

13. Foutch DA, Luco JE, Trifunac MD, Udwadia FE (1975) Full-scale three-dimensional tests of structural deformations during forced excitation of a nine-story reinforced concrete building. Proc. of the US National Conference on Earthquake Engineering. Ann Arbor, pp 206–215
14. Ghobarah A (2004) On drift limits associated with different damage levels. Proc. of the International Workshop on Performance-Based Design, 28 June–1 July 2004, Bled, Slovenia, pp 4321–332
15. Graizer VM (1991) Inertial seismometry methods, Izvestiya. Earth Phys Akad Nauk SSSR 27(1):51–61
16. Graizer VM (2005) Effect of tilt on strong motion data processing. Soil Dyn Earthq Eng 25:197–204
17. Hera A, Hou Z (2004) Application of wavelet approach for ASCE structural health monitoring benchmark studies. J Eng Mech ASCE 130(1):96–104
18. Hou Z, Noori M, Amand R (2000) Wavelet-based approach for structural damage detection. J Eng Mech ASCE 126(7): 677–683
19. Hudson DE (1970) Dynamic tests of full scale structures. In: Wiegel RL (ed) Earthquake Engineering. Prentice Hall, pp 127–149
20. Ivanović SS, Trifunac MD, Todorovska MI (2001) On identification of damage in structures via wave travel times. In: Erdik M, Celebi M, Mihailov V, Apaydin N (eds) Proc. of the NATO Advanced Research Workshop on Strong-Motion Instrumenta-

tion for Civil Engineering Structures, 2–5 June, 1999. Kluwer, Istanbul, pp 447–468
21. Kalkan E, Graizer V (2007) Multi-component ground motion response spectra for coupled horizontal, vertical, angular accelerations and tilt. Indian J Earthq Technol (special issue on Response Spectra) 44(1):259–284
22. Kanai K (1965) Some new problems of seismic vibrations of a structure. Proc. of the Third World Conf. Earthquake Eng, 22 January, 1 February, 1965. Auckland and Wellington, New Zealand, pp II-260–II-275
23. Kapitaniak T (1991) Chaotic Oscillations in Mechanical Systems. Manchester Univ. Press, Manchester
24. Kawakami H, Oyunchimeg M (2003) Normalized input-output minimization analysis of wave propagation in buildings. Eng Struct 25(11):1429–1442
25. Kawakami H, Oyunchimeg M (2004) Wave propagation modeling analysis of earthquake records for buildings. J Asian Archit Build Eng 3(1):33–40
26. Kohler MD, Heaton T, Bradford SC (2007) Propagating waves in the steel, moment-frame Factor building recorded during earthquakes. Bull Seism Soc Am 97(4):1334–1345
27. Kojić S, Trifunac MD, Anderson JC (1984) A post earthquake response analysis of the Imperial County Services building in El Centro. Report CE 84-02. University of Southern California, Department of Civil Engineering, Los Angeles
28. Lee VW, Trifunac MD (1990) Automatic digitization and processing of accelerograms using PC, Dept. of Civil Eng. Report CE 90-03. Univ. Southern California, Los Angeles
29. Lee WHK, Celebi M, Todorovska MI, Diggles MF (eds) (2007) Rotational Seismology and Engineering Applications. Online Proceedings for the First International Workshop, September 18–19 September, Menlo Park. US Geological Survey, Open-File Report 2007-1144 http://pubs.usgs.gov/of/2007/1144
30. Lighthill J (1994) Chaos: A historical perspective. In: Newman WI, Gabrielov A, Turcotte D (eds) Nonlinear Dynamics and Predictability of Geophysical Phenomena. Geophysical Monograph 83, IUGG, vol 18 pp 1–5
31. Liu SC, Tomizuka M, Ulsoy G (2006) Strategic issues in sensors and smart structures. Struct Control Health Monit 13:946–957
32. Luco JE, Trifunac MD, Udwadia FE (1975) An experimental study of ground deformations caused by soil-structure interaction. Proc. US National Conf. on Earthq. Eng. Ann Arbor, MI, pp 136–145
33. Luco JE, Trifunac MD, Wong HL (1987) On the apparent change in the dynamic behavior of a nine-story reinforced concrete building. Bull Seism Soc Am 77(6):1961–1983
34. Luco JE, Trifunac MD, Wong HL (1988) Isolation of soil-structure interaction effects by full-scale forced vibration tests. Earthq Eng Struct Dyn 16:1–21
35. Ma J, Pines DJ (2003) Damage detection in a building structure model under seismic excitation using dereverberated wave machines. Eng Struct 25:385–396
36. Mallat SG (1989) Multiresolution approximations and wavelet orthonormal bases of $L_2(R)$. Trans Am Math Soc 315:69–87
37. Oyunchimeg M, Kawakami H (2003) A new method for propagation analysis of earthquake waves in damaged buildings: Evolutionary Normalized Input-Output Minimization (NIOM). J Asian Archit Build Eng 2(1):9–16
38. Rezai M, Rahmatian P, Ventura C (1998) Seismic data analysis of a seven-storey building using frequency response function and wavelet transform. Proc. of the NEHRP Conference and Workshop on Research on the Northridge, California Earthquake, 17 January, 1994. CUREe, Oakland, pp 421–428
39. Snieder R, Şafak E (2006) Extracting the building response using interferometry: theory and applications to the Millikan Library in Pasadena, California. Bull Seism Soc Am 96(2):586–598
40. Sohn H, Farrar CR, Hemez FM, Shunk DD, Stinemates DW, Nadler BR (2003) A Review of Structural Health Monitoring Literature: 1996–2001, Report LA-13976-MS. Los Alamos National Laboratory
41. Şafak E (1998) Detection of seismic damage in multi-story buildings by using wave propagation analysis. Proc. of the Sixth US National Conf. on Earthquake Eng. EERI, Oakland, Paper No 171, pp 12
42. Şafak E (1999) Wave propagation formulation of seismic response of multi-story buildings. J Struct Eng ASCE 125(4):426–437
43. Todorovska MI (1998) Cross-axis sensitivity of accelerographs with pendulum like transducers: mathematical model and the inverse problem. Earthq Eng Struct Dyn 27:1031–1051
44. Todorovska MI (2009) Seismic interferometry of a soil-structure interaction model with coupled horizontal and rocking response. Bull Seism Soc Am 99-2A (in press)
45. Todorovska MI (2008) Soil-structure system indetification of Millikan Library north-south response during four earthquakes (1970–2002): what caused the observed wandering of the system frequencies? Bull Seism Soc Am 99-2A (in press)
46. Todorovska MI, Al Rjoub Y (2006) Effects of rainfall on soil-structure system frequency: examples based on poroelasticity and a comparison with full-scale measurements. Soil Dyn Earthq Eng 26(6–7):708–717
47. Todorovska MI, Al Rjoub Y (2008) Environmental effects on measured structural frequencies – model prediction of short term shift during heavy rainfall and comparison with full-scale observations. Struct Control Health Monit (in press)doi:10.1002/stc.260
48. Todorovska MI, Lee VW (1989) Seismic waves in buildings with shear walls or central core. J Eng Mech ASCE 115(12):2669–2686
49. Todorovska MI, Trifunac MD (1989) Antiplane earthquake waves in long structures. J Eng Mech ASCE 115(12):2687–2708
50. Todorovska MI, Trifunac MD (1990) A note on the propagation of earthquake waves in buildings with soft first floor. J Eng Mech ASCE 116(4):892–900
51. Todorovska MI, Trifunac MD (2005) Structural Health Monitoring by Detection of Abrupt Changes in Response Using Wavelets: Application to a 6-story RC Building Damaged by an Earthquake. Proc. of the 37th Joint Panel Meeting on Wind and Seismic Effects, 16–21 May, 2005. Tsukuba, Japan. US Japan Natural Resources Program (UJNR), pp 20
52. Todorovska MI, Trifunac MD (2007) Earthquake damage detection in the Imperial County Services Building I: the data and time-frequency analysis. Soil Dyn Earthq Eng 27(6):564–576
53. Todorovska MI, Trifunac MD (2008) Impulse response analysis of the Van Nuys 7-storey hotel during 11 earthquakes and earthquake damage detection. Struct Control Health Monit 15(1):90–116. doi:10.1002/stc.208
54. Todorovska MI, Trifunac MD (2008) Earthquake damage detection in the Imperial County Services Building III: analysis

of wave travel times via impulse response functions. Soil Dyn Earthq Eng 21(5):387–404. doi:10.1016/j.soildyn.2007.07.001
55. Todorovska MI, Trifunac MD (2008) Earthquake damage detection in the Imperial County Services Building II: analysis of novelties via wavelets. Struct Control Health Monit (submitted for publication)
56. Todorovska MI, Trifunac MD, Ivanović SS (2001) Wave propagation in a seven-story reinforced concrete building, Part I: theoretical models. Soil Dyn Earthq Eng 21(3):211–223
57. Todorovska MI, Trifunac MD, Ivanović SS (2001) Wave propagation in a seven-story reinforced concrete building, Part II: observed wave numbers. Soil Dyn Earthq Eng 21(3):225–236
58. Trifunac MD (2007) Early History of the Response Spectrum Method, Dept. of Civil Engineering, Report CE 07-01. Univ. Southern California, Los Angeles, California
59. Trifunac MD, Todorovska MI (2001) A note on the useable dynamic range of accelerographs recording translation. Soil Dyn Earthq Eng 21(4):275–286
60. Trifunac MD, Todorovska MI (2001) Evolution of accelerographs, data processing, strong motion arrays and amplitude and spatial resolution in recording strong earthquake motion. Soil Dyn Earthq Eng 21(6):537–555
61. Trifunac MD, Todorovska MI (2001) Recording and interpreting earthquake response of full-scale structures: In Erdik M, Celebi M, Mihailov V, Apaydin N (eds) Proc. of the NATO Advanced Research Workshop on Strong-Motion Instrumentation for Civil Engineering Structures, 2–5 June, 1999. Kluwer, Istanbul, p 24
62. Trifunac MD, Ivanovic SS, Todorovska MI (1999) Experimental evidence for flexibility of a building foundation supported by concrete friction piles. Soil Dyn Earthq Eng 18(3):169–187
63. Trifunac MD, Todorovska MI, Hao TY (2001) Full-scale experimental studies of soil-structure interaction – a review. Proc. of the 2nd US Japan Workshop on Soil-Structure Interaction, 6–8 March, 2001. Tsukuba City, Japan, pp 52
64. Trifunac MD, Ivanović SS, Todorovska MI (2003). Wave propagation in a seven-story reinforced concrete building, Part III: damage detection via changes in wave numbers. Soil Dyn Earthq Eng 23(1):65–75
65. Trifunac MD, Todorovska MI, Manić MI, Bulajić BĐ (2008) Variability of the fixed-base and soil-structure system frequencies of a building – the case of Borik-2 building. Struct Control Health Monit (in press). doi:10.1002/stc.277
66. Udwadia FE, Jerath N (1980) Time variations of structural properties during strong ground shaking. J Eng Mech Div ASCE 106(EM1):111–121
67. Udwadia FE, Marmarelis PZ (1976) The identification of building structural systems I. The linear case. Bull Seism Soc Am 66(1):125–151
68. Udwadia FE, Marmarelis PZ (1976) The identification of building structural systems II. The nonlinear case. Bull Seism Soc Am 66(1):153–171
69. Udwadia FE, Trifunac MD (1974) Time and amplitude dependent response of structures. Earthq Eng Struct Dyn 2:359–378
70. Ward HS, Crawford R (1966) Wind induced vibrations and building modes. Bull Seism Soc Am 56(4):793–813
71. Wong HL, Trifunac MD, Luco JE (1988) A comparison of soil-structure interaction calculations with results of full-scale forced vibration tests. Soil Dyn Earthq Eng 7(1):22–31

Earthquake Early Warning System in Southern Italy

Aldo Zollo[1], Giovanni Iannaccone[2], Vincenzo Convertito[2], Luca Elia[2], Iunio Iervolino[3], Maria Lancieri[2], Anthony Lomax[4], Claudio Martino[1], Claudio Satriano[1], Emanuel Weber[2], Paolo Gasparini[1]

[1] Dipartimento di Scienze Fisiche, Università di Napoli "Federico II" (RISSC-Lab), Napoli, Italy
[2] Osservatorio Vesuviano, Istituto Nazionale di Geofisica e Vulcanologia (RISSC-Lab), Napoli, Italy
[3] Dipartimento di Ingegneria Strutturale, Università di Napoli "Federico II", Napoli, Italy
[4] Alomax Scientific, Mouans-Sartoux, France

Article Outline

Glossary
Definition of the Subject
Introduction
Earthquake Potential and Seismic Risk in the Campania Region
Seismic Network Architecture and Components
Real-Time Data Transmission System
Network Management and Data Archiving
Real-Time Earthquake Location and Magnitude Estimation
Real-Time Hazard Analysis for Earthquake Early Warning
Future Directions
Bibliography

Glossary

Data transmission system A multi-component device aimed at the transmission of seismic signals over a distance, also denoted as a telecommunication system. Each data transmission system consists of two basic elements: a transmitter that takes information and converts it to an electromagnetic signal and a receiver that receives the signal and converts it back into usable information.

Modern telecommunication systems are two-way and a single device, a transceiver, acts as both a transmitter and receiver. Transmitted signals can either be analogue or digital. In an analogue signal, the signal is varied continuously with respect to the information. In a digital signal, the information is encoded as a set of discrete, binary values. During transmission, the in-

formation contained in analogue signals will be degraded by noise, while, unless the noise exceeds a certain threshold, the information contained in digital signals will remain intact. This represents a key advantage of digital signals over analogue signals. A collection of transmitters, receivers or transceivers that communicate with each other is a telecommunication network. Digital networks may consist of one or more routers that route data to the correct user.

Earthquake early warning system (EEWS) A real-time, modern information system that is able to provide rapid notification of the potential damaging effects of an impending earthquake, through rapid telemetry and processing of data from dense instrument arrays deployed in the source region of the event of concern (regional EEWS) or surrounding the target infrastructure (site-specific EEWS). A "regional" EEWS is based on a dense sensor network covering a portion or the entirety of an area that is threatened by earthquakes. The relevant source parameters (event location and magnitude) are estimated from the early portion of recorded signals and are used to predict, with a quantified confidence, a ground motion intensity measure at a distant site where a target structure of interest is located. On the other hand, a "site-specific" EEWS consists of a single sensor or an array of sensors deployed in the proximity of the target structure that is to be alerted, and whose measurements of amplitude and predominant period on the initial *P*-wave motion are used to predict the ensuing peak ground motion (mainly related to the arrival of *S* and surface waves) at the same site.

Earthquake location An earthquake location specifies the spatial position and time of occurrence for an earthquake. The location may refer to the earthquake hypocenter and corresponding origin time, a mean or centroid of some spatial or temporal characteristic of the earthquake, or another property of the earthquake that can be spatially and temporally localized.

Earthquake magnitude The magnitude is a parameter used by seismologists to quantify the earthquake size. The Richter magnitude scale, or more correctly, local magnitude ML scale, assigns a single number to quantify the amount of seismic energy released by an earthquake. It is a base-10 logarithmic scale obtained by calculating the logarithm of the combined horizontal amplitude of the largest displacement from zero on a seismometer output. Measurements have no limits and can be either positive or negative.

Introduced by the Japanese seismologist Aki in 1962, the seismic moment is the present-day physical parameter used to characterize the earthquake strength. It represents the scalar moment of one the couples of forces producing the dislocation at an earthquake fault and it is measured from the asymptotic DC level on displacement Fourier spectra of recorded seismic signals.

Probability density function – PDF A function in one or more dimensional space **X** that (i) when integrated over some interval Δx in **X** gives a probability of occurrence of any event within Δx, and (ii) has unit integral over space **X**, where **X** represents a space of possible events.

Seismic data-logger A core element of a digital seismic station, whose aim is to record the analogue signals from seismic sensors and convert them in digital form with an assigned sampling frequency. Ground motion signals acquired by seismic sensors are pre-amplified and anti-aliasing filtered in a data-logger before they are digitalized through an AD (analog-to-digital) converter. The main technical features of a modern data-logger are the number of available channels, the allowed sampling frequencies, the dynamic range, the digitizer clock type, the storage capacity (PCMCIA, internal flash and/or hard disk, USB, ...), network interfaces (ethernet, wireless lan, or ppp) and power consumption.

Seismic hazard The probability that at a given site, a strong motion parameter (generally the peak ground acceleration) exceeds an assigned value in a fixed time period. When the seismic hazard is computed for an extended region it is generally represented as a map. The hazard map is commonly computed for a constant probability level (10%, 5% or 2%) and a given time window (50 years). It represents the spatial variation of the peak ground acceleration (expressed in percentage of gravity *g*) to be exceeded in the given period with the chosen probability level.

Earthquake early warning systems can provide a mean for the evaluation of real-time hazard maps which evolve with time, as new information about source location, magnitude and predicted peak ground motion parameters are available soon after the earthquake occurrence.

Seismic sensors Instruments used to record the ground vibration produced by natural and artificial sources, generally denoted as seismometers. A seismometer measures the relative motion between its frame and a suspended mass. Early seismometers used optics, or motion-amplifying mechanical linkages. The motion was recorded as scratches on smoked glass, or exposures of light beams on photographic paper. In modern

instruments the proof mass is held motionless by an electronic negative feedback loop that drives a coil. The distance moved, speed and acceleration of the mass are directly measured. Most modern seismometers are broadband, working on a wide range of frequencies (0.01–100 Hz). Another type of seismometer is a digital strong-motion seismometer, or accelerometer, which measures soil acceleration. Due to its relatively high dynamic range, the accelerometer can record unsaturated strong amplitude signals at close distances from a large earthquake. This data is essential to understand how an earthquake affects human structures.

Definition of the Subject

The origin of the term "early warning" probably goes back to the first decades of the last century. However, the first practical use of an "early warning" strategy was military and it was developed during the "cold war" years as a countermeasure to the potential threat from intercontinental ballistic missiles. The objective of these systems was to give an alert to target areas as soon as a missile was detected by a radar system or a launch was detected by a satellite system. In this context the term "lead time" was defined as the time elapsing between the detection of the missile and the estimated impact on the target.

In the last decades the use of the term "early warning" greatly expanded. It is used with small, but significant, variations in various types of risks, from epidemiological, to economic, social, and of course all the types of natural and environmental risks.

In fact, in these contexts, including some natural risks such as hydro-geological and volcanic, the warning is not given at the onset of the catastrophic phenomenon, but after the occurrence of some precursory phenomena which can trigger a catastrophic event (for instance intensive rainfall for hydrological risk, earthquakes and/or ground deformation for volcanic risk). The main consequence of this difference is an increase in the probability of issuing false alarms.

The case of earthquake early warning is similar to missile early warning. The alert is given after an earthquake is detected by a network of seismometers. An earthquake early warning is based on the fact that most of the radiated energy is contained in the slower traveling phases (*S*- and surface waves traveling at about 3.5 km/s or less) which arrive at any location with a delay with respect to small amplitude higher velocity phases (*P*-waves, travelling at about 6–7 km/s) or to an electromagnetically transmitted (EM) signal giving the warning.

Introduction

Many regions in the world are affected by natural hazards such as earthquakes, tsunamis, volcanoes, floods, storms, landslides, etc., each of which can have devastating socio-economic impacts. Among these natural events, earthquakes, have been among the most recurrent and damaging hazards during last few decades, resulting in large numbers of casualties, and massive economic losses [30].

The problem of earthquake risk mitigation is faced using different approaches, depending upon the time scale being considered. Whilst over time scales of decades it is of utmost importance that land use regulations and building/infrastructure codes are continuously updated and improved, for time scales of a few years, the main risk mitigation actions are at the level of information and education in order to increase individual and social community awareness about potentially damaging hazards. Over shorter time scales (months to hours), it would naturally be of great benefit to society as a whole if the capability to accurately predict the time, location and size of a potentially catastrophic natural event were available. However, due to the great complexity of the natural processes of concern, such predictions are currently not possible.

On the other hand, on very short time scales (seconds to minutes), new strategies for earthquake risk mitigation are being conceived and are under development worldwide, based on real-time information about natural events that is provided by advanced monitoring infrastructures, denoted as "early warning systems".

Regional and On-site Early Warning Systems

Earthquake Early Warning Systems (EEWS) are modern, real-time information systems that are able to provide rapid notification of the potential damaging effects of an impending earthquake through the rapid telemetry and processing of data from dense instrument arrays deployed in the source region of the event of concern. Such systems allow mitigating actions to be taken before strong shaking and can significantly shorten the time necessary for emergency response and the recovery of critical facilities such as roads and communication lines.

Advances have been made towards the implementation of operational systems in Japan, Taiwan, and Mexico using two different approaches, i. e., "regional warning" and "onsite warning" [25]. A regional warning system is based on a dense sensor network covering a portion or the entire area that is threatened by earthquakes. The relevant source parameters (earthquake location and magnitude) are estimated from the early portion of recorded signals and are used to predict, with a quantified confidence,

a ground motion intensity measure at a distant site where a target structure of interest is located. Alternatively, "on-site warning" systems consist of a single sensor or an array of sensors deployed in the proximity of the target structure that is to be alerted, and whose measurements on the initial *P*-wave motion are used to predict the ensuing peak ground motion (mainly related to the arrival of *S* and surface waves) at the same site.

Implementation of Early Warning Systems Worldwide

In Japan, since the 1965, the JNR (Japanese National Railway) has developed and operated the Urgent Earthquake Detection and Alarm System (UrEDAS), which is an on-site warning system along the Shinkansen (bullet train) railway. UrEDAS is based on seismic stations deployed along the Japanese Railway with an average distance of 20 km. An alert is issued if the horizontal ground acceleration exceeds 40 cm/s^2. In the 1996, the UrEDAS was combined with a new seismometer called "compact UrEDAS" [31,32,33].

On the other hand, for about one decade the Japanese Meteorological Agency (JMA) has been developing and experimenting with a mixed single station and network based early warning system to generate immediate alerts after earthquakes with JMA Intensity greater than "lower 5" (approximately $M > 6$) [24]. During a testing period from February 2004 to July 2006, the JMA sent out 855 earthquake early warnings, only 26 of which were recognized as false alarms [40]. On October 1, 2007 the broadcast early warning system developed by the Japanese Meteorological Agency (JMA) became operative. In this system, the first warning is issued 2 s after the first *P* phase detection, if the maximum acceleration amplitude exceeds the threshold of 100 cm/s^2.

In the United States the first prototype of an early warning system was proposed by Bakun et al. [4] and developed for mitigating earthquake effects in California. It was designed to rapidly detect the Loma Prieta aftershocks and send an alert when the estimated magnitude was greater than 3.7, in order to reduce the risk of the crews working in the damaged area. The system is composed of four components: ground motion sensors deployed in the epicentral area, a central receiver, radio repeaters and radio receivers. The prototypical system worked for 6 months, during which time 19 events with $M > 3.5$ occurred, 12 alerts were issued with only 2 missed triggers and 1 false alarm.

Based on pioneering work by Allen and Kanamori [2] seismologists across California are currently planning real-time testing of earthquake early warning across the state using the ElarmS (Earthquake Alarms Systems) methodology [1]. The approach uses a network of seismic instruments to detect the first-arriving energy at the surface, the *P*-waves, and translate the information contained in these low amplitude waves into a prediction of the peak ground shaking that follows. Wurman et al. [47] illustrated the first implementation of ElarmS in an automated, non-interactive setting, and the results of 8 months of non-interactive operation in northern California.

Since 1989, in Mexico, the civil association CIRES (Centro de Instrumentacion y REgistro Sismico) with the support of Mexico City Government Authorities, developed and implemented the Mexican Seismic Alert System (SAS) [15]. The SAS is composed of (a) a seismic detection network, 12 digital strong motion stations deployed along 300 km of the Guerrero coast, (b) a dual communication system: a VHF central radio relay station and three UHF radio relay stations, (c) a central control system which continuously controls the operational status of the seismic detection and communication system and, when an event is detected, automatically determines the magnitude and issues the alarm, and (d) a radio warning system for broadcast dissemination of the alarm to end users. After 11 years, the SAS system recorded 1373 events in the Guerrero coast, it issued 12 alerts in Mexico city, with only one false alarm.

In Taiwan, the Taiwan Central Weather Bureau (CWB) developed an early warning system based on a seismic network consisting of 79 strong motion stations installed across Taiwan and covering an area of 100×300 km^2 [44]. Since 1995 the network has been able to report event information (location, size, strong motion map) within 1 min after an earthquake occurrence [39]. To reduce the report time, Wu and Teng [44] introduced the concept of a virtual sub-network: as soon as an event is triggered by at least seven stations, the signals coming from the stations less distant than 60 km from the estimated epicenter are used to characterize the event. This system successfully characterized all the 54 events occurrd during a test period of 7 months (December 2000 – June 2001), with an average reporting time of 22 s.

In Europe, the development and testing of EEWS is being carried out in several active seismic regions. Europe is covered by numerous high-quality seismic networks, managed by national and European agencies, including some local networks specifically designed for seismic early warning around, for example, Bucharest, Cairo, Istanbul and Naples.

In Turkey, an EEWS is operative, called PreSEIS (pre-seismic shaking), to provide rapid alert for Istanbul and surrounding areas. It consists of 10 strong motion sta-

tions located along the border of the Marmara sea along an arc of about 100 km, close to the seismogenetic zone of the Great Marmara Fault Zone with real time data transmission to Kandilli-Oservatory [7,14]. An alarm is issued when a threshold amplitude level is exceeded.

In Romania, the EEWS is based on three tri-axial strong motion sensors deployed in the Vrancea area with a satellite communication link to the Romanian Data Center at NIEP in Bucharest [7,42]. The system is based on first *P* wave detection and prediction of the peak horizontal acceleration recorded in Bucharest, allowing for a warning time of about 25 s.

On 2006 the European Union launched the 3-year project SAFER (Seismic Early Warning for Europe), which is a cooperative scientific program aimed at developing technological and methodological tools that exploit the possibilities offered by real-time analysis of signals coming from these networks for a wide range of actions, performed over time intervals of a few seconds to some tens of minutes. The project includes the participation of 23 research groups from several countries of Europe. The primary aim of SAFER is to develop tools that can be used by disaster management authorities for effective earthquake early warning in Europe and, in particular, its densely populated cities.

The Development of an Early Warning System in Campania Region, Southern Italy

The present article is focused on the description of technologies and methodologies developed for the EEWS under construction in southern Italy.

With about 6 million inhabitants, and a large number of industrial plants, the Campania region (southern Italy), is a zone of high seismic risk, due to a moderate to large magnitude earthquake on active fault systems in the Apenninic belt. The 1980, $M = 6.9$ Irpinia earthquake, the most recent destructive earthquake to occur in the region, caused more than 3000 causalities and major, widespread damage to buildings and infrastructure throughout the region.

In the framework of an ongoing project financed by the Regional Department of Civil Protection, a prototype system for seismic early and post-event warning is being developed and tested, based on a dense, wide dynamic seismic network under installation in the Apenninic belt region (ISNet, Irpinia Seismic Network).

Considering an earthquake warning window ranging from tens of seconds before to hundred of seconds after an earthquake, many public infrastructures and buildings of strategic relevance (hospitals, gas pipelines, railways, railroads, ...) in the Campania region can be considered as potential EEWS target-sites for experimenting with innovative technologies for data acquisition, processing and transmission based on ISNet. The expected time delay to these targets for the first energetic *S* wave train is around 30 s at about 100 km from a crustal earthquake occurring in the source region. The latter is the typical time window available for mitigating earthquake effects through early warning in the city of Naples (about 2 million inhabitants, including suburbs).

This article illustrates the system architecture and operating principles of the EEWS in the Campania region, focusing on its innovative technological and methodological aspects. These are relevant for a reliable real-time estimation of earthquake location and magnitude which are used to predict, with quantified confidence, ground motion intensity at a distant target site.

The system that we describe in this article uses an integrated approach from real time determination of source parameters to estimation of expected losses.

This problem must be dealt in an evolutionary (i. e., time-dependent) and probabilistic framework where probability density functions (PDFs) for earthquake location, magnitude and attenuation parameters are combined to perform a real-time probabilistic seismic hazard analysis.

Earthquake Potential and Seismic Risk in the Campania Region

The southern Apennines are an active tectonic region of Italy that accommodates the differential motions between the Adria and Tyrrhenian microplates [23]. The majority of the seismicity in this region can be ascribed to this motion. These earthquakes mainly occur in a narrow belt along the Apennine chain and are associated with young faults, with lengths ranging from 30 to 50 km, and mainly confined to the upper 20 km of the crust [28,41].

Recent stress and seismic data analyzed by [29] using earthquake locations and fault mechanisms show that the southern Apennines are characterized by an extensional stress regime and normal-fault earthquakes. However, the occurrence of recent (e. g., 5 May, 1990, Potenza, *M* 5.4; 31 October – 1 November, 2002, Molise, M 5.4) and historic (e. g., 5 December, 1456, *M* 6.5) earthquakes do not exclude other mechanisms such as strike-slip faulting.

There have been numerous large and disastrous events in the southern Apennines, including those which occurred in 1694, 1851, 1857 and 1930. The location of historical earthquakes retrieved from the CFTI (Catalogo dei Forti Terremoti in Italia, Catalogue of Strong Earthquakes

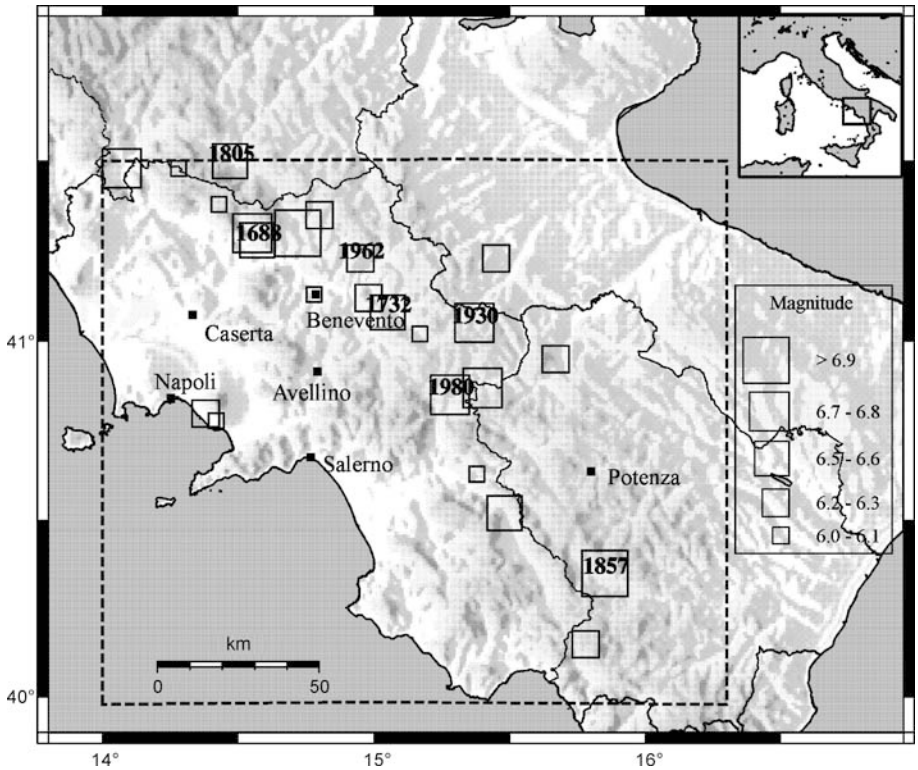

Earthquake Early Warning System in Southern Italy, Figure 1

Location of the main historic earthquakes retrieved from the CFTI database using as region of interest that defined by the external rectangle. The box dimensions are proportional to magnitude. The best constrained historic earthquakes are reported along with their date of occurrence

in Italy) database [6] is shown in Fig. 1. The most recent and well documented event is the complex normal-faulting M 6.9 Irpinia earthquake of 23 November, 1980 [5,43].

As recently indicated in the study by Cinti et al. [9], the southern Apennines has a high earthquake potential with an increasing probability of occurrence for $M \geq 5.5$ earthquakes in the next decade. The new national hazard map (Gruppo di lavoro MPS, 2004), indicates that the main towns of the region fall in a high seismic hazard area, where it is expected that a peak ground acceleration value ranging between 0.15 and 0.25 g will be exceeded in 475 years.

These aspects make the Campania region a suitable experimental site for the implementation and testing of an early warning system. A potential application of an early warning system in the Campania region should consider an expected time delay to the first energetic S wave train varying between 14–20 s at 40–60 km distance to 26–30 s at about 80–100 km, from a crustal earthquake occurring along the Apenninic fault system. Based on those delay times, a large number of civil and strategic infrastructures located in the Campania region are eligible for early waning applications, as shown in Fig. 2.

Seismic Network Architecture and Components

The Irpinia Seismic Network (ISNet) is a local network of strong motion, short period and broadband seismic stations deployed along the southern Apenninic chain covering the seismogenic areas of the main earthquakes that occurred in the region in the last centuries, including the $Ms = 6.9$, 23 November 1980 event.

The seismic network is composed of 29 stations organized in six sub-nets, each of them composed of a maximum of 6–7 stations (Fig. 3). The stations of a given sub-net are connected with real-time communications to a central data-collector site called the Local Control Center (LCC).

The different LCCs are linked to each other and to a Network Control Center (NCC) with different types of transmission systems. The whole data transmission system is fully digital over TCP/IP, from the data-loggers, through

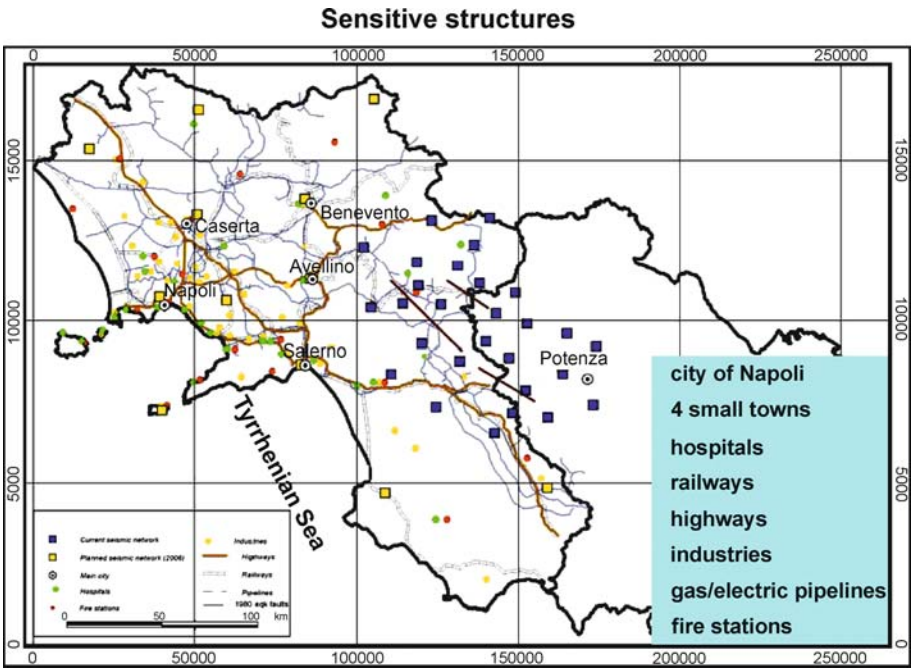

Earthquake Early Warning System in Southern Italy, Figure 2
Distribution of the sensitive structures, potential candidates for an early warning system in the Campania-Lucania region

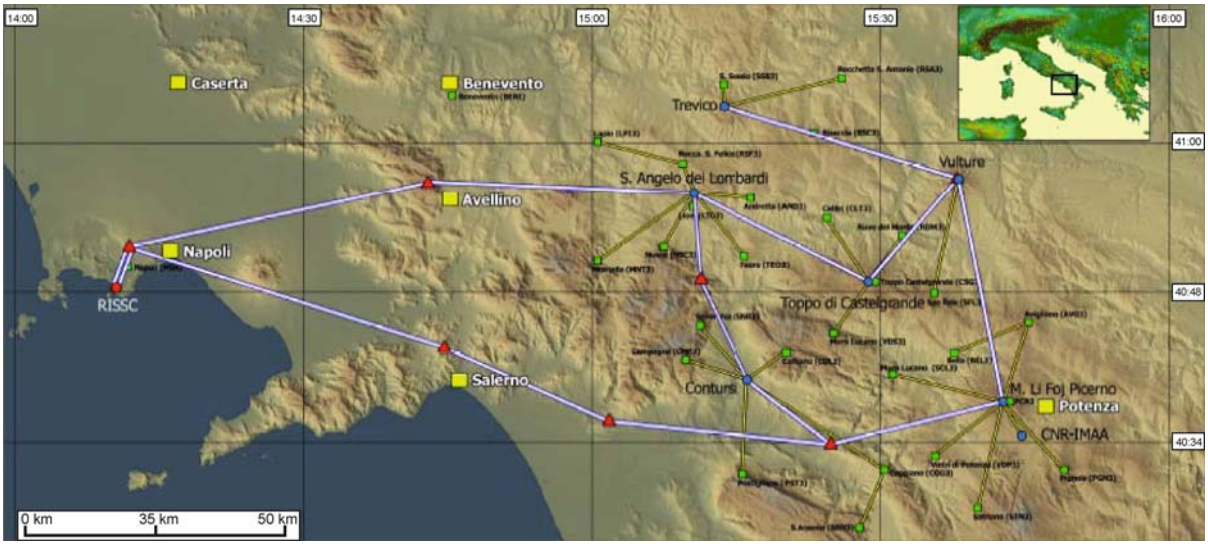

Earthquake Early Warning System in Southern Italy, Figure 3
Topology of the communication system of ISNet showing the extended-star configuration of the seismic network. Symbols explanation: *green squares* – seismic stations; *blue squares* – Local Control Centres (LCC); *yellow lines* – WLAN radio linkconnecting seismic stations and LCC; *white segments* – SDH carrier-class radio; *red triangles* – radio link repeaters; *red circle* – Network Control Centre RISSC in Naples; *yellow squares* – main cities

the LCC, to the NCC, located in the city of Naples, 100 km away from the network center.

To ensure a high dynamic recording range, each seismic station is equipped with a strong-motion accelerometer and a three-component velocity meter (natural period = 1 s). In five station locations the seismometers are replaced by broadband (0.025–50 Hz) sensors to guarantee good-quality recording of teleseismic events. Data ac-

quisition at the seismic stations is performed by the innovative data-logger Osiris-6, produced by Agecodagis sarl. The hardware/software characteristics of the system allow it to install self-developed routines to perform real time specific analysis.

The data-loggers are remotely controlled through a configuration tool accessible via TCP/IP, managing sampling rate, gain, application of calibration signal to the resets of disks, GPS, etc. Furthermore, a complete station health status is available, which helps in the diagnosis of component failure or data-logger malfunction. The data-loggers store the data locally or send it to each LCC where the real-time data management system Earthworm (developed at USGS-United State Geological Survey) is operating.

A calibration unit is installed at each seismic station to automatically provide a periodic calibration signal to seismic sensors in order to verify the correct response curve of the overall acquisition chain.

The power supply of the seismic station is provided by two solar panels (120 W peak, with 480 Wh/day), two 130 Ah gel cell batteries, and a custom switching circuit board between the batteries. With this configuration, 72-h autonomy is ensured for the seismic and radio communication equipment. Each site is also equipped with a GSM/GPRS programmable control/alarm system connected to several environmental sensors and through which the site status is known in real time. With SMS (Short Message Service) and through the programmable GSM controller, the seismic equipment can be completely reset remotely with a power shutdown/restart. The GSM also controls the device start/stop release procedure when the battery goes over/under a predefined voltage level.

Unlike the seismic stations, LCCs, which host the data server and transmission system instruments, are AC power supplied with back-up gel batteries guaranteeing 72-h stand-by power.

Real-Time Data Transmission System

ISNet has a distributed star topology that uses different types of data transmission systems.

The seismic stations are connected via spread-spectrum radio bridges to the LCCs. Data transmission between LCCs from the local control center to the network control center in Naples is performed through different technologies and media types as shown in Table 1.

To transmit waveforms in real time from the seismic stations to the LCCs, a pair of outdoor Wireless LAN bridges operating in the 2.4 GHz ISM band are used. Our tests have shown that these instruments operate continuously without any radio link failure due to adverse weather conditions (snow, heavy rain).

The two primary backbone data communication systems of the central site use Symmetrical High-speed Digital Subscriber Line (SHDSL) technology over a frame-relay protocol. Frame relay offers a number of significant

Earthquake Early Warning System in Southern Italy, Table 1
Specification of the ISNet data communication links

Type	Frequency (GHz)	Bandwidth (Mbps)	# Number of		Comments
			Stations	LCCs	
Spread spectrum Radio	2.45	54	27[3]	–	Throughput around 20–24 Mbps for links between 10–15 km (based on ethernet packets with an average size of 512 bytes).
Ethernet	–	100	2[3]		Stations connected with ethernet cable to LCC infrastructure.
Wireline SHDSL over Frame Relay	–	2.048	–	2	At the central site (RISSC) the CIR[1] is maximum 1.6 Mbps depending upon number of PVCs[2]. At the remote (LCC) site the bandwidth is 640/256 kbps with CIR of 64 kbps in up and download, over ADSL with ATM ABR service class.
Microwave Radio SDH	7	155	–	6	Carrier-class microwave link. Connect six LCC with 155 Mbps (STM-1) truly full bandwidth available. First link constructed for early warning applications.
Microwave Radio HyperLAN/2	5.7	54	–	2	The true usable maximum throughput of HyperLAN/2 is 42 Mbps.

[1] CIR Commited Information Rate.
[2] PVC permanent virtual circuit.
[3] Not included stations hosted by LCCs.

benefits over analogue and digital point-to-point leased lines. With the latter, each LCC requires a dedicated circuit between the LCCs and NCC. Instead, the SHDSL frame relay is a packet-switched network, which allows a site to use a single frame-relay phone circuit to communicate with multiple remote sites through the use of permanent virtual circuits. With virtual circuits, each remote site is seen as part of a single private LAN, simplifying IP address scheme maintenance and station monitoring.

Each seismic site has a real-time data flow of 18.0 kbps (at 125 Hz sampling rate for each physical channel), and the overall data communication bandwidth that is needed is around 540 kbps for 30 stations. ISNet supports this throughput under the worst conditions seen and it has been designed to guarantee further developments, such as the addition of further seismic or environmental sensors, without the need for larger economic and technological investment.

Network Management and Data Archiving

The Network Manager Application and Implementation Overview

As seen in the previous paragraphs, ISNet is a complex infrastructure, and thus needs a suitable software application in order to be effectively managed: a front-end to users and administrators with an interface that is simple to use. To this aim we developed a server-client database-driven application, dubbed *SeismNet Manager*, to keep track of the several components that comprise or are produced by the network, such as stations, devices and recorded data. This application, whose front page is shown in Fig. 4, lets the administrators manage (insert, edit, view and search) the details of (a) seismic stations and Local Control Centers (sites), (b) data communication links between sites (wired or wireless), instruments and devices (sensors, loggers, network hardware), and (c) recorded and computed data (waveforms, events). SeismNet Manager also keeps an

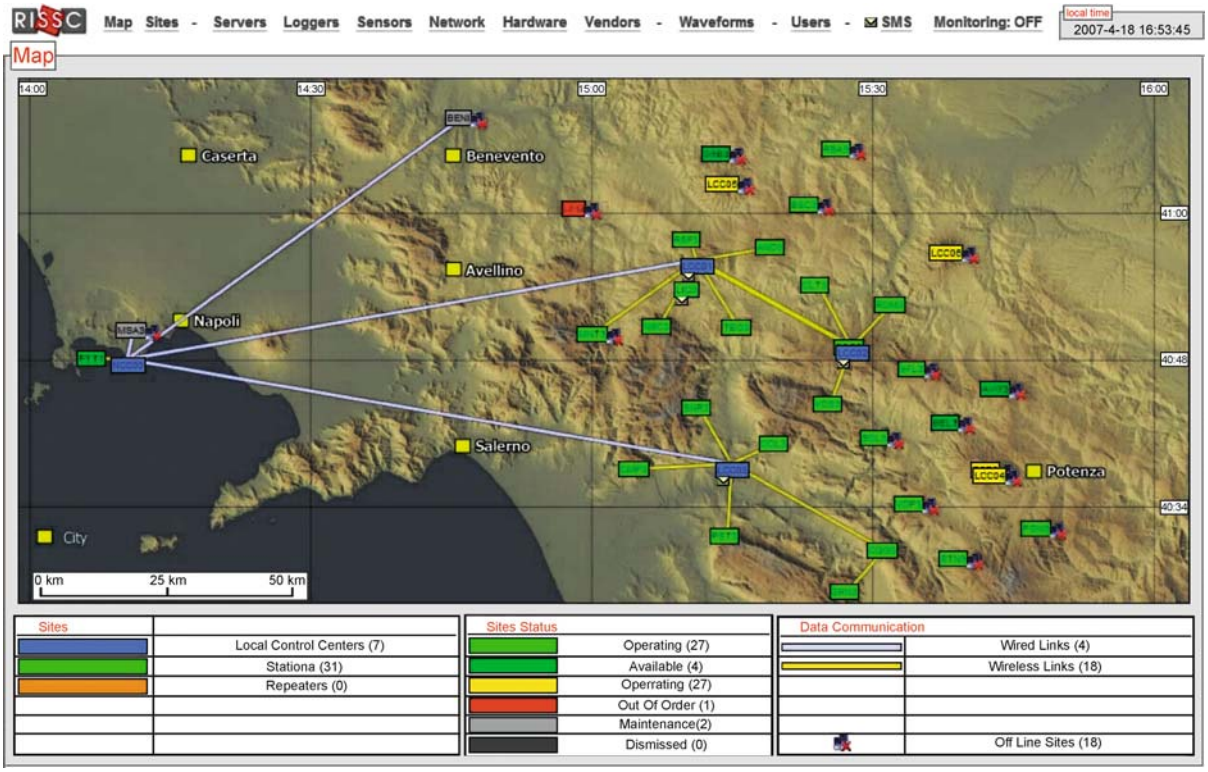

Earthquake Early Warning System in Southern Italy, Figure 4
The front page of SeismNet Devices Manager. This page is meant to convey the state of the whole network at a glance. Each node (station or LCCs) is shown along with its operating state, data links of different types to nearby nodes, whether it's currently on-line or not, along with eventual alarms still pending

historical record of the installations and configurations of the above elements.

All of the mentioned components are handled by leveraging an instrumental database, a flexible repository of information that was implemented by using PostgreSQL, a robust and feature-rich Database Management System available as open source.

The Instrumental Database

The instrumental database is a web-oriented application tool where, at the top level, the network is modeled as a set of sites, with installed loggers, sensors, data acquisition servers, network hardware and generic hardware, in a given configuration. Each of the mentioned entities is mirrored by a different class of objects in the database, where the relevant details are stored and then presented to the users as interactive web pages. As an example, see the page for a typical seismic station in Fig. 5.

The instrumental database was implemented with a layer of abstraction that lets one easily to perform complex queries and hides the actual implementation details of the underlying structure to a possible client. There are both *stored procedures*, i. e., functions that perform complex tasks given simple inputs, and *views*, i. e., virtual database tables that collect the most important pieces of information about an abject, physically scattered in many tables, in a single place and make it possible to easily query,

Earthquake Early Warning System in Southern Italy, Figure 5

The page relative to a seismic station. This page is a collection of all the pieces of information linked to a particular site: location details and map; some pictures and notes; recently received warning messages; currently installed devices and their configurations and mutual connections; data links to other stations; most recent waveforms recorded. Every device at a site also has an associated installation object, that records the configuration parameters and the physical connections to other nearby devices, valid over a period of time. Some elements, such as the data storage servers and the loggers, also need some further configuration parameters, that are independent of their actual physical installation, for things like firmware release and versions of the software packages run

for example, for all the details of the correctly operating sensors installed one year ago at stations with a working wireless link to a given LCC server.

This abstract interface makes the devices database a central repository for effectively cross-correlating the seismic data recorded at any given place and time with the details of the instrument(s) that recorded them, and the configuration details of the systems that ultimately made them available. The interface approach also makes it easier to change the implementation details without the need to update the web application, or any external client procedures that need to interact with the instrumental database.

Automatic Monitoring of the Devices and Automatic Data Retrieval

All of the details about the network described so far are provided by the administrators of the system and are manually updated every time the configuration of something in the network changes, e. g., after installing a new sensor or replacing some faulty hardware at a station. This manual input is needed for "dumb" devices, such as sensors. "Smart" devices, i. e., computers with an IP address (loggers, bridges and Earthworm servers), on the other hand, can be queried about their actual configuration from time to time. The web application can plot the temporal evolution of some selected parameters as graphs, spanning a period ranging from hours to years (Fig. 6). This is useful to correlate issues spotted on the recorded seismic data (typically, "holes" in the stream of data) to hardware problems. It is possible to inspect the whole chain of data transfers to pinpoint the source of the problem (e. g., low batteries on a logger due to a faulty inverter, low signal of a wireless connection due to harsh weather conditions).

There are both automatic and manual procedures to insert new events and data files in the system. The automatic procedures make use of several sources of events to process, such as: INGV (Istituto Nazionale di Geofisica e Vulcanologia, Italy) bi-weekly bulletins; INGV real time alerts; our early warning system. Likewise, they exploit several sources of recorded seismic data in order to provide a SAC file (SAC – seismic analysis code from

Earthquake Early Warning System in Southern Italy, Figure 6

Health graphs of the devices. A device can be marked for monitoring and its internal state, or "health", gets polled at regular intervals. Several of its internal variables are then retrieved and stored into the database, and their temporal evolution can be plotted as a graph. In this case both the internal temperature and CPU load of an OSISRIS data logger are shown, over periods ranging from one hour to one year

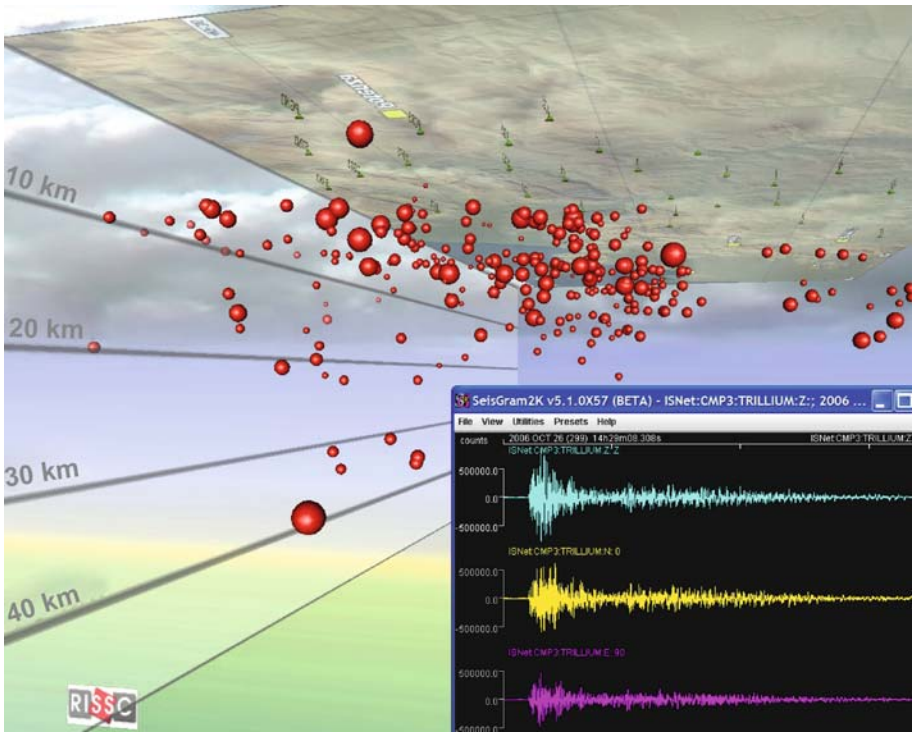

Earthquake Early Warning System in Southern Italy, Figure 7
Visualization of the seismic data. This is the graphical presentation of data recorded by ISNet. The waveforms matching the user's search criteria can be viewed on-line via Seisgram2K (where they can also be processed), while the events are rendered via VRML as a fully interactive 3D scene in the browser itself

Lawrence Livermore National Laboratory) spanning the period from just before the arrival time, up to the end of the event.

Sensor data are retrieved from: (1) a repository of files from the internal mass storage of the loggers; (2) a local Earthworm Wave Server that caches older data collected from all the LCCs; (3) the most recent real time recording from the remote Wave Servers. The instrumental database is used to determine which sites/sensors/configurations recorded each event and to fill the headers of the files using the standard SAC format. The waveforms and events database, on the other end, is used by the automatic procedures to know which pieces of data are still missing for already recorded events (due to e. g., the temporary unavailability of one or more seismic data sources) and need to be collected.

The Waveforms and Events Database: Searching and Visualizing the Seismic Data

We also built a waveform and event database, the natural complement to the instrumental database. It keeps track of the events detected by the network and the relative waveforms recorded by the sensors. This database stores objects for events, origin estimations (time and location), magnitude estimations and waveforms. Several origins can be attached to a single event, as different algorithms and different institutions provide different estimations. Likewise, several magnitude types and estimations are attached to each origin. A waveform object for each sensor that recorded the earthquake is also linked to the event object, and stores a pointer to a SAC file, and its source (site and channel). The latter records are then used to gather, from the instrumental database, the actual details of the instruments that recorded the data.

An interface for searching both events and waveforms is provided, as pictured in Fig. 7. Events can be filtered on origin time and location, magnitude, and distance to the stations. Waveforms can be filtered on station, component, instrument and quality.

Real-Time Earthquake Location and Magnitude Estimation

Real-Time Earthquake Location

Previous Related Studies There are many methodologies for standard earthquake location, performed when

most or all the phase arrival times for an event are available. Standard analysis techniques are generally not suited for early warning applications, since they typically need the seismic event to be fully recorded at several stations, leaving little or no lead time for the warning [25]. For this reason, a different strategy is required, where the computation starts when a few seconds of data and a small number of recording stations are available, and the results are updated with time.

Previous work on earthquake location for early warning includes several approaches to gain constraints on the location at an earlier time and with fewer observations than for standard earthquake location.

In the ElarmS methodology [47], when the first station triggers, the event is temporarily located beneath that station; after a second station trigger the location moves to a point between the two stations, based on the timing of the arrivals; with three or more triggered arrivals, the event location and origin time is estimated using trilateration and a grid search algorithm.

Horiuchi et al. [18] combine standard L2–norm event location, equal differential-time (EDT) location on quasi-hyperbolic surfaces, and the information from not-yet arrived data to constrain the event location beginning when there are triggered arrivals from two stations. The two arrivals times define a hyperbolic surface, which contains the event location. This solution is further constrained by EDT surfaces constructed using the current time (t_{now}) as a substitute for future, unknown arrival times at the stations, which have not yet recorded arrivals. The constraint increases as t_{now} progresses, even if no further stations record an arrival.

Rydelek and Pujol [36], applying the approach of Horiuchi et al. [18], show that useful constraints on an event location can be obtained with only two triggered stations. Cua and Heaton [12], generalized the approach by Rydelek and Pujol in order to start the location with one single triggering station.

The real-time location technique described in this paper is based on the equal differential-time (EDT) formulation [16,27] for standard earthquake location. The EDT location is given by the point traversed by the maximum number of quasi-hyperbolic surfaces, on each of which the difference in calculated travel-time to a pair of stations is equal to the difference in observed arrival times for the two stations. The EDT location determination is independent of origin time and reduces to a 3D search over latitude, longitude and depth. Furthermore, EDT is highly robust in the presence of outliers in the data [27]. This robustness is critical for the problem of earthquake location for seismic early warning, since we will often work with small numbers of data and may have outlier data such as false triggers, picks from other events, and misidentified picks from energetic, secondary phases.

Assuming that a dense seismic network is deployed around the fault zone, we define as the "*evolutionary approach*" a type of analysis where the estimates of earthquake location and size, and their associated uncertainty, evolve with time as a function of the number of recording stations and of the length of the portion of signal recorded at each station.

A direct implication of the evolutionary strategy is that each algorithm must be capable of *real-time* operation, i. e., its computational time must be smaller than the rate at which data enters the system.

Furthermore, since each algorithm starts processing a limited amount of information, the estimated earthquake parameter must be provided, at each time step, as a *probability density function* (PDF) which incorporates in its definition the uncertainties related both to the model employed and to the available data.

The Real-Time Earthquake Location Method The methodology is related to that of Horiuchi et al. [18], which has been extended and generalized by (a) starting the location procedure after only one station has triggered, (b) using the equal differential-time approach proposed by Font [16] to incorporate the triggered arrivals and the not-yet-triggered stations, (c) estimating the hypocenter probabilistically as a PDF instead of as a point, and (d) applying a full, non-linearized, global-search for each update of the location estimate.

We assume that a seismic network has known sets of operational and non-operational stations (Fig. 8a), that when an earthquake occurs, triggers (first *P*-wave arrival picks) will become available from some of the operational stations, and that there may be outlier triggers which are not due to *P* arrivals from the earthquake of interest.

Let's denote the operational stations as $(S_0, \ldots, S_N)$, and consider a gridded search volume V containing the network and target earthquake source regions, and the travel times from each station to each grid point (i, j, k) in V computed for a given velocity model.

The standard EDT approach states that, if the hypocenter (i, j, k) is exactly determined, then the difference between the observed arrival times t_n and t_m at two stations S_n and S_m is equal to the difference between calculated travel times tt_n and tt_m at the hypocentral position, since the observed arrival times share the common earthquake origin time. In other words, the hypocenter must

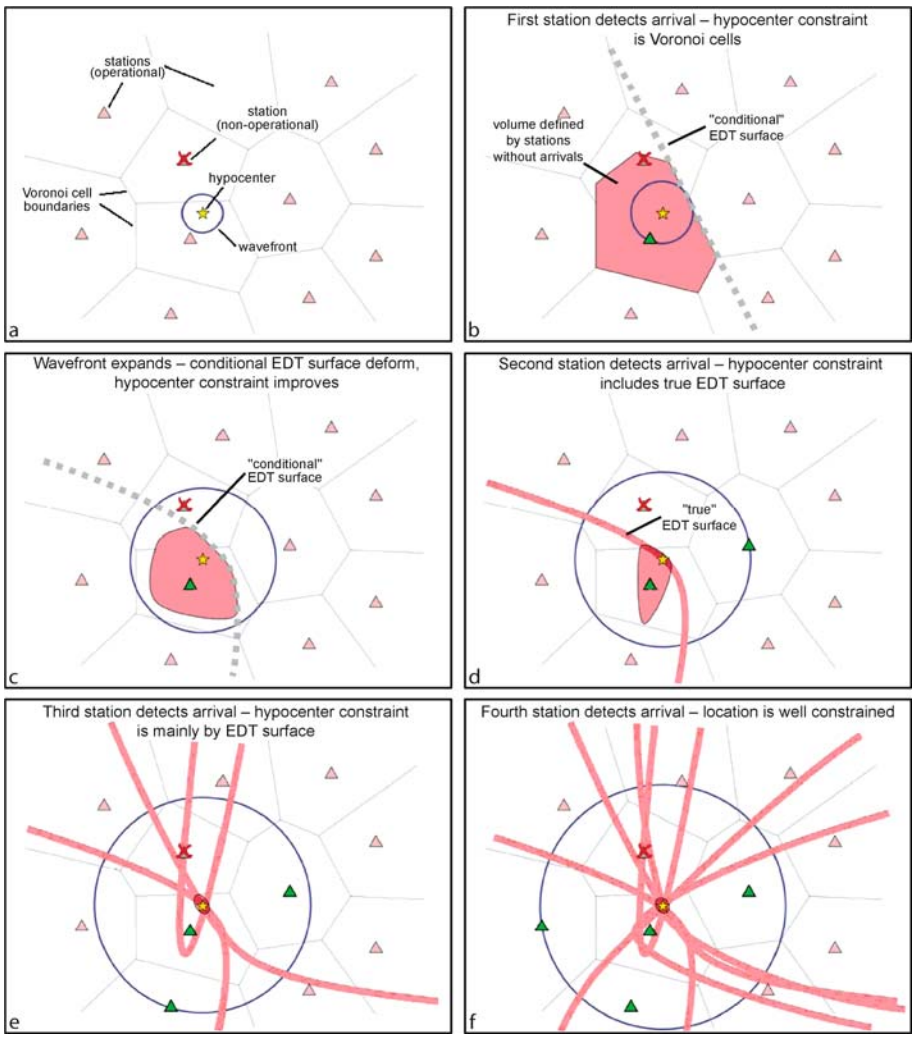

Earthquake Early Warning System in Southern Italy, Figure 8

Schematic illustration of the evolutionary earthquake location algorithm. For clarity, only a map view with the epicentral location is represented. a Given a seismic network with known sets of operational and non-operational stations, we can define a priori the Voronoi cell associated to each station. b When the first station triggers, we can define a volume that is likely to contain the location, this volume is limited by conditional EDT surfaces on which the *P* travel time to the first triggering station is equal to the travel-time to each of the operational but not-yet-triggered stations. c As time progresses, we gain additional information from the stations that have not yet triggered, the EDT surfaces move towards and bend around the first triggering station, and the likely-location volume decreases in size. d When the second station triggers, we can define a true EDT surface; the hypocenter is on the intersection between this surface and the volume defined by the conditional EDT surfaces, which continues decreasing in size. e When a third station triggers, we can define two more true EDT surfaces, further increasing the constraint on hypocenter position. f As more stations trigger, the location converges to the standard EDT location composed entirely of true EDT surfaces

satisfy the equality:

$$(tt_m - tt_n)_{i,j,k} = t_m - t_n \; ; \quad m \neq n \tag{1}$$

for each pair of triggering stations S_n and S_m. For a constant velocity model, this equation defines a 3D hyperbolic surface whose symmetry axis goes through the two stations. Given N triggering stations, $N(N-1)/2$ surfaces can be drawn; the hypocenter is defined as the point crossed by the maximum number of EDT surfaces.

Following an evolutionary approach, the method evaluates, at each time step, the EDT equations considering not only each pair of triggered stations, but also those pairs where only one station has triggered.

Therefore, when the first station, S_n, triggers with an arrival at $t_n = t_{\text{now}}$ (t_{now} is the current clock time), we can already place some limit on the hypocenter position (Fig. 8b). These limits are given by EDT surfaces defined by the condition that each operational but not-yet-triggered station S_l will trigger in the next time instant, $t_l \geq t_n$. That is:

$$(tt_l - tt_n)_{i,j,k} = t_l - t_n \geq 0\,; \quad l \neq n\,. \tag{2}$$

On these conditional EDT surfaces, the P travel time to the first triggering station tt_n is equal to the travel-time to each of the not-yet-triggered stations, tt_l, $l \neq n$. These surfaces bound a volume (defined by the system of inequalities) which must contain the hypocenter. In the case of a homogeneous medium with constant P-wave speed, this hypocentral volume is the Voronoi cell around the first recording station, defined by the perpendicular bisector surfaces with each of the immediate neighboring stations.

As the current time t_{now} progresses, we gain the additional information that the not-yet-triggered stations can only trigger with $t_l > t_{\text{now}}$. Thus the hypocentral volume is bounded by conditional EDT surfaces that satisfy the inequality:

$$(tt_l - tt_n)_{i,j,k} \geq \delta t_{n,l}\,; \quad l \neq n\,. \tag{3}$$

δt is the time interval between the arrival time at station S_n and the latest time for which we have information from station S_l,

$$\delta t_{n,l} = t_{\text{now}} - t_n\,, \tag{4}$$

where t_n is the observed arrival time at station S_n.

The system (3) defines the volume, bounded by the conditional EDT surfaces, in which the hypocenter may be located given that, at current time t_{now}, only the station S_n has triggered. When $\delta t = 0$ the system (3) reduces to the system (2); for $\delta t > 0$, the hypocentral volume will be smaller than the previous one, since the updated, conditional EDT surfaces tend to fold towards and around the first triggered station (Fig. 8c).

We interpret the hypocentral volume in a probabilistic way by defining, for each inequality in (3), a value $p_{n,l}(i,j,k)$ which is 1 if the inequality is satisfied and 0 if not. Then we sum the $p_{n,l}(i,j,k)$ over stations l at each grid point, obtaining a non-normalized probability density $P(i,j,k)$, where $P(i,j,k) = N-1$ for grid points where all the inequalities are satisfied and a value less than $N-1$ elsewhere.

When the second and later stations trigger, we first re-evaluate the system (3) for all pairs of triggered stations S_n and all not-yet-triggered stations S_l. Secondly, we construct standard, true EDT surfaces (see Eq. 2) between each pair S_n, S_m of the triggered stations, by evaluating for each grid point the quantity:

$$q_{n,m}(i,j,k) = \exp\left\{-\frac{[(tt_n - tt_m)_{i,j,k} - (t_n - t_m)]^2}{2\sigma^2}\right\}\,; \quad n \neq m\,. \tag{5}$$

The expression between square brackets at the exponent is the standard EDT Eq. 2 whose solutions are quasi-hyperbolic surfaces; in practice all true EDT surfaces are given a finite width by including the uncertainty σ in the arrival time picking and the travel-time calculation.

The quantity $q_{n,m}(i,j,k)$ has values between 0 and 1. We sum the $q_{n,m}(i,j,k)$ with the $p_{n,l}(i,j,k)$ obtained from the re-evaluation of (4) to obtain a new $P(i,j,k)$.

Starting from P, we define a value:

$$Q(i,j,k) = \left(\frac{P(i,j,k)}{P_{\max}}\right)^N\,, \tag{6}$$

which forms a relative probability density function (PDF, with values between 0 and 1) for the hypocenter location within the grid cell (i,j,k). The function $Q(i,j,k)$ may be arbitrarily irregular and may have multiple maxima.

At predetermined time intervals, we evaluate (3) and (5) to obtain $Q(i,j,k)$ in the search volume, using the Oct-tree importance sampling algorithm ([13,27], http://www.alomax.net/nlloc/octtree). This algorithm uses recursive subdivision and sampling of rectangular cells in 3D space to generate a cascade structure of sampled cells, such that the spatial density of sampled cells follows the target function values. The Oct-tree search is much faster than a simple or nested grid search (factor 10–100 faster) and more global and complete than stochastic search methods algorithms such as simulated annealing and genetic algorithms [13]. For each grid point, an origin time estimate can be obtained from the observed arrival times and the calculated travel times.

As more stations trigger, the number of not-yet-triggered stations becomes small, and the location converges towards the hypocentral volume that is obtained with standard EDT location using the full set of data from all operational stations (Fig. 8d–f).

If there are uncorrelated outlier data (i. e., triggers that are not compatible with P arrivals from a hypocenter within or near the network), then the final hypocentral volume will usually give an unbiased estimate of the hypocentral location, as with standard EDT location.

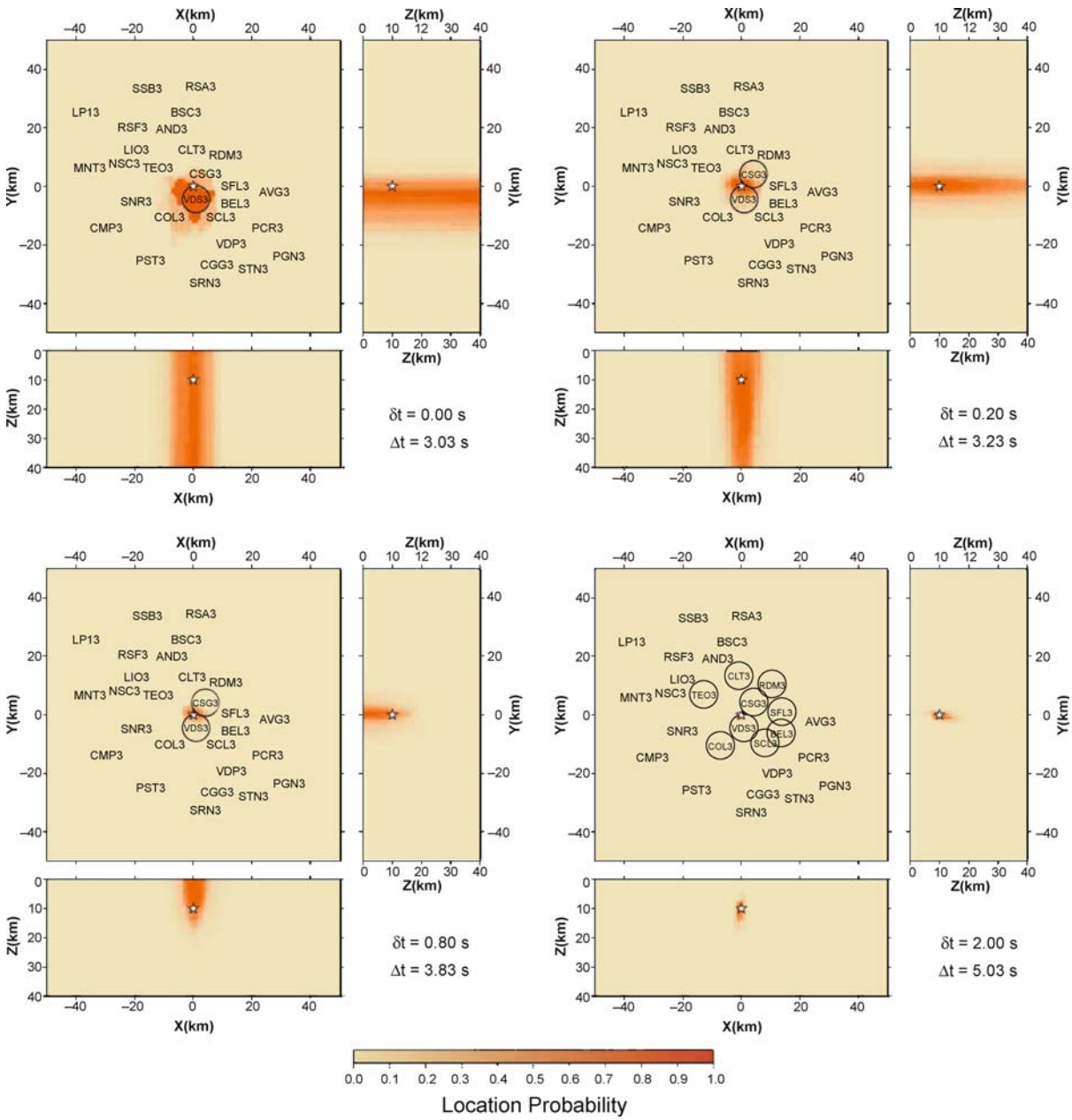

Earthquake Early Warning System in Southern Italy, Figure 9

Location test for a synthetic event occurring at the center of the Irpinia Seismic Network (ISNet). The three orthogonal views show marginal values of the probability function $Q(i, j, k)$. The true hypocenter is identified by a *star*. δt is the time from the first trigger, Δt is the time from event origin. For each snapshot, stations that have triggered are marked with a *circle*

However, if one or more of the first arrival times is an outlier, then the earliest estimates of the hypocentral volume may be biased. Synthetic tests have shown that, if N_{out} is the number of outlier data, the bias reduces significantly after about $4 + N_{out}$ arrivals have been obtained, and then decreases further with further arrivals, as the solution converges towards a standard EDT location [37].

We performed several synthetic tests using the geometry of the ISNet network. For each simulated event, we computed theoretical arrival picks using travel times ob-

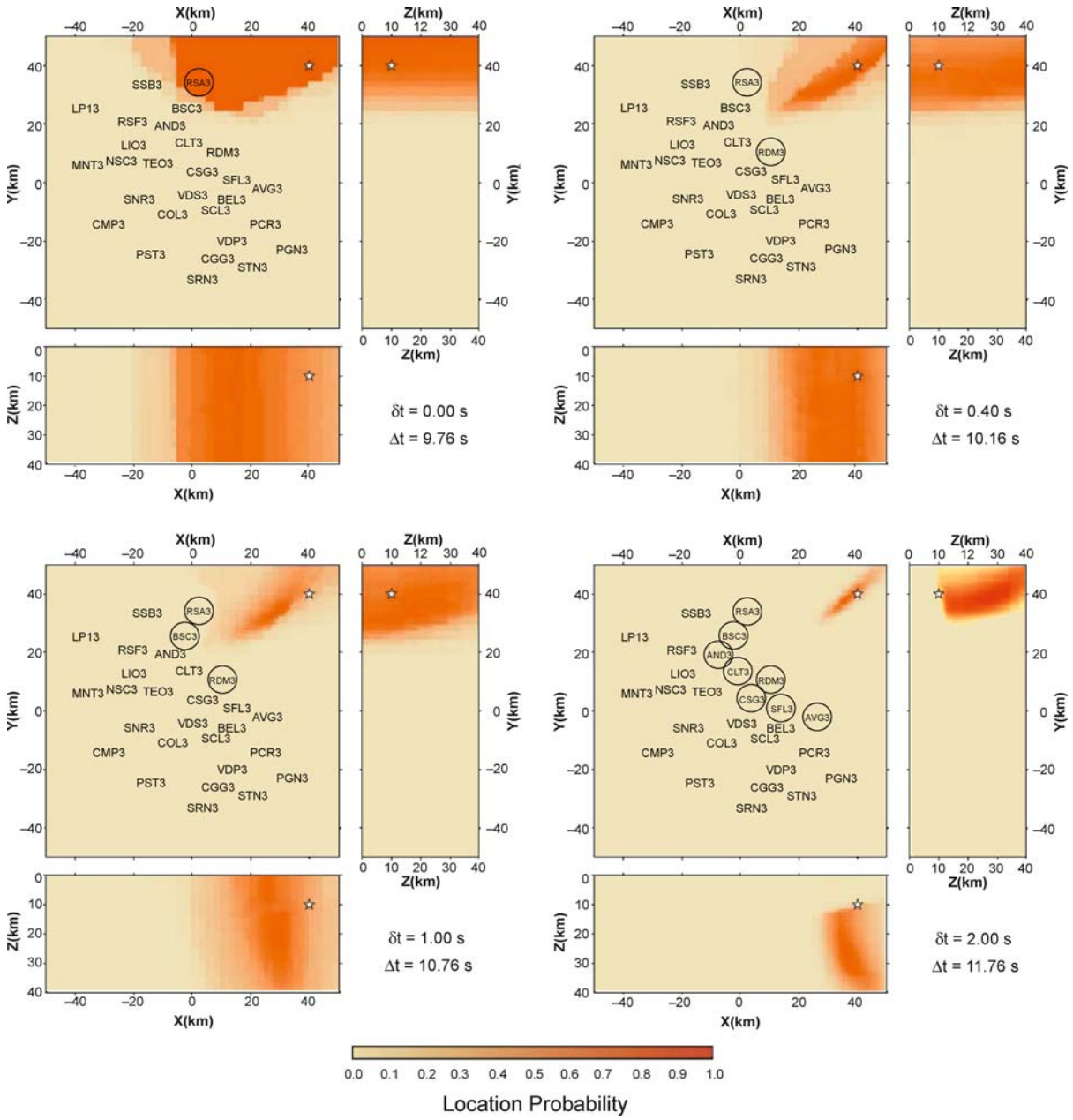

Earthquake Early Warning System in Southern Italy, Figure 10
Location test for a synthetic event occurring outside the ISNet network (see Fig. 9 for explanation)

tained by the finite difference solution of the eikonal equation [35] for a 1D, *P*-wave velocity model. To reproduce uncertainties introduced by the picking algorithm, we add to each arrival time a random error following a Gaussian distribution with a variance of 0.02 s.

Here we use only *P* picks since currently most networks have poor capability to perform real-time *S* picking. Our tests consider an earthquake occurring at the center of the network at a depth of 10 km (Fig. 9) and an earthquake occurring outside the network at a depth of 10 km (Fig. 10). Each panel in Figs. 9 and 10 is a snapshot at a given time showing the marginal map (i. e., summed over i, j or k) for $Q(i, j, k)$ along the horizontal (x, y) and the two vertical (x, z and y, z) planes. The star shows the

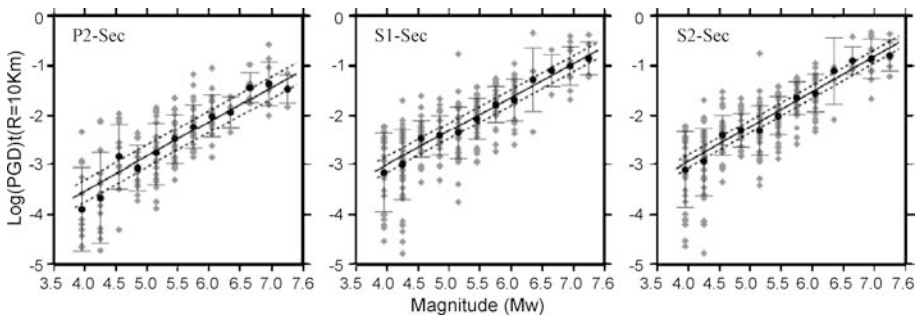

Earthquake Early Warning System in Southern Italy, Figure 11

Correlation between low-pass filtered peak ground motion value and moment-magnitude for earthquakes occurred in the Euro-Mediterranean region (after [49]). The panels show the logarithm of peak ground displacement normalized at a reference distance of 10 km as a function of Mw in time windows of (*left*) 2 s length from the first *P*-arrival and (*middle*) 1- and (*right*) 2-s from the first *S*-arrivals. *P*- and *S*-data are measured on vertical and root-squared sum of horizontal components, respectively. Each panel shows the best fit regression line (*solid line*) along with 1-WSE limits (*dashed lines*)

known, synthetic hypocentral location. In the first case, two seconds after the first trigger (5.03 s from the event origin), 9 stations have triggered and the location is already well constrained for early warning purposes.

In the second case, at $\Delta t = 11.76$ s, 2 s after the first event detection, the constraint on the location PDF improves further, but the PDF retains an elongated shape because of the poor azimuthal coverage of the network for this event. The event depth is only constrained by an upper bound, but the depth range includes the true value.

Real-Time Magnitude Estimation Using a Bayesian, Evolutionary Approach

Previous Related Studies The problem of magnitude estimation from early seismic signal has been previously approached and analyzed by different authors.

Nakamura [31] first proposed the correlation between the event magnitude and the characteristic period of *P*-phase defined as the ratio between the energy of the signal and its first derivative.

Allen and Kanamori [2] modified the original Nakamura method and described the correlation between the predominant period and the event magnitude for Southern California events. Lockman and Allen (2007) studied the predominant period – magnitude relations for the Pacific Northwest and Japan. They also investigated the sensitivity of such relations using different frequency bands.

Using a complementary approach, Wu and Kanamori [45] investigated the feasibility of an on-site EEWS for Taiwan region based on prediction of earthquake damage, based on measurements of the predominant period and peak displacement on early *P*-wave signals detected at the network.

Odaka et al. (2003) proposed a single station approach for the real-time magnitude estimation. The authors fit the initial part of waveform envelope and showed a relation between the final event magnitude, the envelope shape coefficient and the maximum *P* amplitude measured in a 3 s time window.

Wu and Zhao [46] and Zollo et al. [49] (Fig. 11) demonstrated the existence of a correlation between the event magnitude and the peak displacement measured a few seconds after the *P* arrival based on massive analysis of Southern Californian and Euro–Mediterranean earthquake records. In particular, Zollo et al. showed that both *P* and *S* wave early phases have the potential for real time estimation of magnitude up to about *M* 7. Zollo et al. [50] and Lancieri and Zollo [26] extended this observation to Japanese earthquake records, showing that a possible saturation effect may exist at about *M* 6.5 for *P* measurements in 2 s windows while it vanishes when a larger, 4 s window is considered. The scaling of displacement peak with magnitude, instead, appears at even shorter (1 s) time lapses after the first *S*-arrival.

Using an alternative method, Simmons [38] proposed a new algorithm based on discrete wavelet transforms able to detect first *P* arrival and to estimate final magnitude analyzing first seconds of *P*-wave.

The Real-Time Magnitude Estimation Method The real time and evolutionary algorithm for magnitude estimation presented in this paper is based on a magnitude predictive model and a Bayesian formulation. It is aimed at evaluating the conditional probability density function of magnitude as a function of ground motion quantities measured on the early part of the acquired signals [19].

The predictive models are empirical relationships which correlate the final event magnitude with the logarithm of quantities measured on first 2–4 s of record.

The first prediction model, based on the predominant period of P-phase (τ_P), has been introduced by Allen and Kanamori [2]. Recently, Wu and Zhao [46] showed the existence of a correlation between magnitude, distance and peak displacement measured in a 2–4 s window after P-phase.

Zollo et al. [49,50] refined this correlation and extended the observation on the peaks measured in 2 s after the S-phase arrival through the analysis of the European and Japanese strong motion data-bases (Ambraseys et al. [3], K-NET www service of NIED – National Research Institute for Earth Science and Disaster Prevention, Japan).

The method therefore assumes that the linear relationship between the logarithm of the observed quantity and magnitude is known, along with standard errors of the predictive models.

At each time step t from the first station trigger, the conditional PDF of magnitude M given the observed data vector $\underline{d} = \{d_1, d_2, \ldots, d_n\}$ is expressed via the Bayes theorem as:

$$f(m|\underline{d}) = \frac{f(\underline{d}|m)f(m)}{\int_{M_{\mathrm{MIN}}}^{M_{\mathrm{MAX}}} f(\underline{d}|m)f(m)\mathrm{d}M} , \tag{7}$$

where $f(m)$ is the a priori distribution which incorporates the information available before the experimental data are collected through a truncated exponential functional form, derived by the Gutenberg–Richter recurrence relationship,

$$f(m) : \begin{cases} \frac{\beta \mathrm{e}^{-\beta m}}{\mathrm{e}^{-\beta M_{\min}} - \mathrm{e}^{-\beta M_{\max}}} & M_{\min} \le m \le M_{\max} \\ 0 & m \notin [M_{\min}, M_{\max}] \end{cases} , \tag{8}$$

where $\{\beta, M_{\min}, M_{\max}\}$ depend on the seismic features and on the detection threshold of the seismic network of the considered region.

The conditional probability $f(\underline{d}|m)$ contains all the information concerning the magnitude as retrievable from the data acquired at time t.

Assuming that components of the observed data vector $\underline{d}$ have a lognormal distribution, and that they are stochastically independent and identically distributed random variables of parameters $\mu_{\log(d)}$ and $\sigma_{\log(d)}$, then the likelihood is written as:

$$f(\underline{d}|m) = \prod_{i=1}^{\nu} \frac{1}{\sqrt{2\pi}\sigma_{\log(d)} d_i} \mathrm{e}^{-\frac{1}{2}\left(\frac{\log(d_i)-\mu_{\log(d)}}{\sigma_{\log(d)}}\right)^2} , \tag{9}$$

where ν is the number of stations acquiring at the instant t; $\mu_{\log(d)}$ and $\sigma_{\log(d)}$ are the mean and the standard deviation of the logs of d_i, respectively..

Substituting Eq. 8 and Eq. 9 into Eq. 7, $f(m|\underline{d})$ results as in Eq. 10 where it depends on data only trough $\sum_{i=1}^{\nu} \log(d_i)$ and ν, which therefore are jointly sufficient statistics for the estimation of magnitude [21]:

$$f(m|\underline{d}) = f\left(m| \sum_{i=1}^{\nu} \log(d_i)\right) = \frac{\mathrm{e}^{\left(2\mu_{\log(d)}\left(\sum_{i=1}^{\nu}\log(d_i)\right) - \nu\mu_{\log(d)}^2\right)\Big/ 2\sigma_{\log(d)}^2} \mathrm{e}^{-\beta m}}{\int_{M_{\mathrm{MIN}}}^{M_{\mathrm{MAX}}} \mathrm{e}^{\left(2\mu_{\log(d)}\left(\sum_{i=1}^{\nu}\log(d_i)\right) - \nu\mu_{\log(d)}^2\right)\Big/ 2\sigma_{\log(d)}^2} \mathrm{e}^{-\beta m} \mathrm{d}M} . \tag{10}$$

As just outlined, $f(m|\underline{d})$ depends on $\sum_{i=1}^{\nu} \log(d_i)$ and on the number of stations triggered, ν, at the time of the estimation and, consequently, on the amount of information available. As more stations are triggered, and provide more measures of d, the estimation improves.

The described technique is evolutionary in the sense that $f(m|\underline{d})$ depends on time, i. e., as time passes, additional stations provide new observations (predominant period and/or P-, S-peaks), which are used to refine the probabilistic estimation of magnitude.

Magnitude Estimation from Peak Displacement Measurements

The empirical relationships between low-pass filtered, initial P- and S-peak displacement amplitudes and moment magnitude (e. g. [49]) can be used as predictive models for the real-time estimation of magnitude using the Bayesian approach described above.

While the P-wave onset is identified by an automatic picking procedure, the S-onset can be estimated from an automatic S-picking or from a theoretical prediction based on the hypocentral distance given by the actual earthquake location. At a given time step after the first P-wave detection at the network, progressively refined estimates of magnitude are obtained from P- and S-peak displacement data. These are preliminarily corrected for distance amplitude effects through an empirical attenuation relationship obtained from available strong motion records [46,49]:

$$f(M, R) = A_{\mathrm{phase}} + B_{\mathrm{phase}} M + C_{\mathrm{phase}} \log(R) , \tag{11}$$

where the constants A_{phase}, B_{phase} and C_{phase} are determined through a best-fit regression with a retrieved standard error of $SE_{\mathrm{phase}}^{\mathrm{PMR}}$ and R is the hypocentral distance.

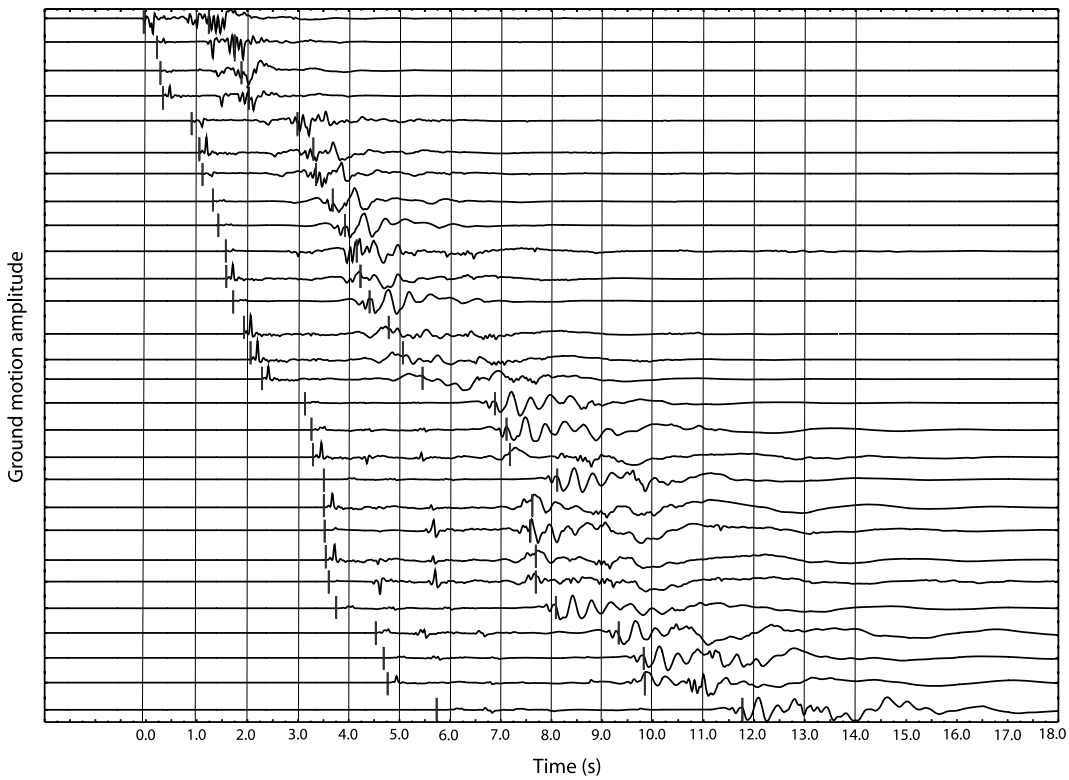

Earthquake Early Warning System in Southern Italy, Figure 12

Synthetic seismograms for a *M* 7.0 earthquake at the center of the network (see Fig. 9). The seismograms are computed using a line source, rupture model (constant rupture velocity) while complete wavefield green's functions in a flat-layered model are computed by using the discrete wavenumber summation method of Bouchon [8]. Each *vertical line* indicates the 1 s signal packets examined at each time step. This plot allows us to understand seconds after seconds which stations are acquiring and what sort of input (*P* or *S* peak) they are giving to the real time system. For example after three seconds to the first *P* phase picking thirteen stations are acquiring, the 2 s *S*-phase peak is available at the nearest stations. This observation motivates the use of the *S* phase information in a real time information. If a dense network is deployed in the epicentral area the nearest station will record the *S*-phase before the *P* phase arrives to the far ones, as seen in previous example, and this is perfectly compatible with the real time analysis

Following the procedure described in [49], the relationship (11) is used to correct observed peaks for the distance effect, by normalizing them to a reference distance (e. g., $R = 10\,\mathrm{km}$) and to determine a new best fit regression between the distance corrected peak value $(PD_{\mathrm{phase}})^{10\,\mathrm{km}}$ and the final magnitude:

$$\log\left(PD_{\mathrm{phase}}^{10\,\mathrm{km}}\right) = \log\left(PD_{\mathrm{phase}}^{R}\right) - C_{\mathrm{phase}}\log\left(\frac{R}{10}\right) \quad (12)$$

$$\log\left(PD_{\mathrm{phase}}^{10\,\mathrm{km}}\right) = A'_{\mathrm{phase}} + B'_{\mathrm{phase}}M\,. \quad (13)$$

Assuming a standard error of SE_{phase}^{PM} on peak displacements retrieved from (13) and combining the Eqs. (11) and (13), the mean values and standard deviation of quantity $\log(PD_{\mathrm{phase}})$, can be written as:

$$\begin{aligned}\mu_{\log(PD_{\mathrm{phase}})} &= B'_{\mathrm{phase}}M + A'_{\mathrm{phase}} + C_{\mathrm{phase}}\log\left(\frac{R}{10}\right)\\ \sigma_{\log(PD_{\mathrm{phase}})} &= SE_{\mathrm{phase}}^{PM} + \log\left(\frac{R}{10}\right)\Delta C_{\mathrm{phase}}\\ &\quad + C_{\mathrm{phase}}\frac{1}{R}\Delta R\,,\end{aligned} \quad (14)$$

where R is estimated with an error of ΔR and $\Delta C_{\mathrm{phase}}$ is the error on the C_{phase} coefficient in Eq. (12).

The values of coefficients in (14) used for real time magnitude estimates at ISNet are obtained from the regression analysis based on records from the European Strong Motion Database [49] and given in Table 2.

Figure 12 illustrates an example of real time magnitude estimation on a simulated event with $M = 7.0$, whose epicenter is located along the 1980 Irpinia earthquake faulting

Earthquake Early Warning System in Southern Italy, Table 2
Coefficients of the empirical regression relationships between low-pass filtered *P* and *S* displacement peaks and magnitude

Phase	A'_{phase}	B'_{phase}	C_{phase}	SE^{PM}_{phase}	ΔC_{phase}
2P	−6.31	0.70	−1.05	0.22	0.30
2S	−5.77	0.71	−0.71	0.13	0.16

system. Synthetic seismograms have been computed by using the discrete wave-number method of Bouchon [8] and Coutant (1989) for a extended source model in a flat-layered velocity model.

Figure 13a shows the probability density function defined in Eq. (8) evaluated at each time step. Time zero is assigned to the first P detection at the network. As time evolves the PDF tightens around the predicted magnitude value, indicating a more refined, probabilistic estimate of magnitude.

By defining $F_t(m)$ as the cumulative PDF at time t, it is possible to estimate a magnitude range of variation $[M_{\min}; M_{\max}]$ whose limits are defined based on the shape of the $F_t(m)$ function:

$$M_{\min} : \int_{-\infty}^{M_{\min}} f_t(m|\underline{d})\mathrm{d}m = \alpha \,, \qquad M_{\max} : \int_{-\infty}^{M_{\max}} f_t(m|\underline{d})\mathrm{d}m = 1-\alpha \,. \tag{15}$$

For example, if we assume $\alpha = 1\%$, then $M_{\min}$ and $M_{\max}$ will be, respectively, the $F_t(m)$ evaluated at 0.01 and 0.99.

In Fig. 13b the estimates of magnitude uncertainty bounds are reported as a function of time. After three iterations (corresponding to a time of 9 s from the event origin time and 4 s after the first P-phase arrival at the network) the magnitude estimation converges to the true magnitude value. In fact, due to the high density of seismic station in the epicentral area, at that time most of seismic station contributes to the magnitude estimation with peaks read on P-phase windows (Fig. 14), while a further refinement of magnitude estimate is due to the near source S-wave arrivals.

Real-Time Hazard Analysis for Earthquake Early Warning

The Real-Time Hazard Determination

Using the methods previously described for estimating in real-time the event magnitude and location, it is possible to perform a real-time hazard analysis [19]. This analysis is based on the extension of classical Probabilistic Seismic Hazard Analysis (PSHA) proposed by Cornell [11] that is generally used for long-term probabilistic hazard assessment. Classical PSHA integrates data from existing seismic catalogs both in terms of magnitude, location and recorded strong ground motion values in addition to the information concerning seismogenic areas of interest (ex-

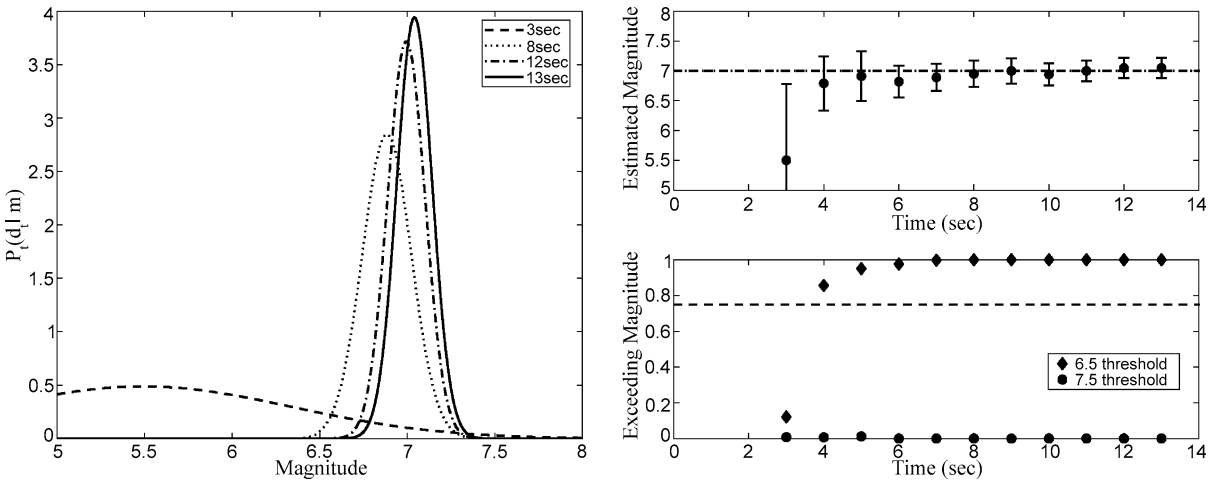

Earthquake Early Warning System in Southern Italy, Figure 13
Application of the method for real time magnitude estimate to a *M* 7 simulated event occurring within the area covered by the ISNet network. *Left panel*. PDF distribution at several time steps measured from the first *P*-phase picking. *Right top*, magnitude estimation with uncertainties as a function of time. The *dashed line* refers to the actual magnitude value, the errors represent the 95% of confidence bound evaluated as cumulative PDF integral in the 5–95% range. *Right bottom*, probability to exceed magnitude 6.5 and magnitude 7.5 thresholds in function of time. The *dashed line* is the 75% probability level

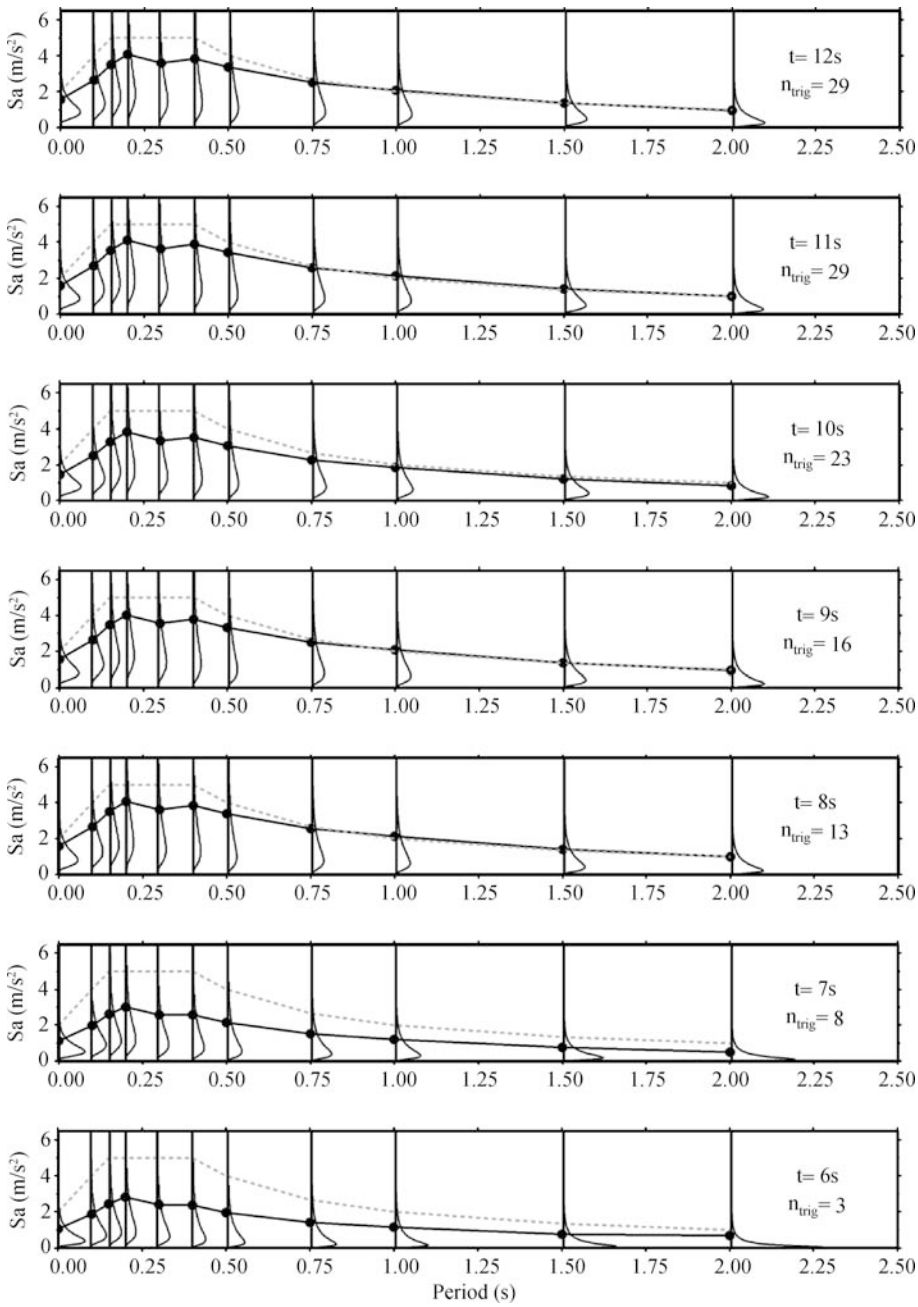

Earthquake Early Warning System in Southern Italy, Figure 14

Real-time estimation of spectral ordinates' distributions as function of the number of stations triggered for a *M* 7.0 event with an epicentral distance of 50 km from the early warning target site. The parameter n_{trig} in the figure is equivalent to the number of stations ν in the text. The acceleration spectrum (*black curve*) was obtained by choosing at each period the spectral value with 20% exceedance probability according to the corresponding distribution, so it is analogous to a uniform hazard spectrum with the exception that it is computed in real-time. The *grey dashed line* is the Italian code spectrum assigned for building design in the target location at the town of Avellino, 40 km distant from the earthquake epicenter, and is reported for comparison purposes (after [10])

pected maximum magnitude, b-value of the Gutenberg Richter relationship, etc.) to provide the hazard curve as the final outcome. Each point on that curve corresponds to the value of a ground motion intensity measure (IM) (e. g., peak ground acceleration, PGA, peak ground velocity, PGV or the spectral acceleration, Sa), having a given probability or frequency of exceedance in a fixed period of time for a site of interest.

The probabilistic framework of the PSHA, specifically the hazard integral, can be used for real-time hazard if the PDFs of magnitude and source-to-site distance are replaced with those depending on the data gathered by the EEWS during the occurrence of a specific earthquake.

This is the case, for example, of the PDF on the source-to-site distance whose statistical moments evolve with real-time earthquake location. As a consequence, this PDF does not depend on the seismic potential of the area of interest (as in the case of the classical PSHA, which accounts for the occurrence of all the earthquake in a fixed range of magnitude), but rather depends on the time evolving event location provided by the EEWS. The same considerations apply to the PDF on the magnitude as described in the following sections whose statistical moment, at a given time, depends on the number of triggered stations at that time.

In this theoretical framework the real-time hazard integral can be written as:

$$f(\mathrm{IM}|\underline{d},\underline{s}) = \int_M \int_R f(\mathrm{IM}|m,r) f(m|\underline{d}) f(r|\underline{s}) \mathrm{d}M \mathrm{d}R, \quad (16)$$

where $f(r|\underline{s})$ is the PDF of distance r, which eventually depends only on the triggering sequence of the stations in the network, where $\underline{s} = \{s_1, \ldots, s_\nu\}$ is such a sequence. This renders also the PDF of r time dependent.

Given that for each point in a volume containing the earthquake hypocenter, the probability of that point being coincident with the true hypocenter is calculated via a rapid location technique, a simple geometrical transformation allows one to obtain the probabilistic distribution of the source-to-site distance.

The PDF $f(\mathrm{IM}|m,r)$, is given, for example, by an ordinary attenuation relationship. It is worth to recall that the computed hazard refers to a particular set of triggered stations and, consequently, it depends on the information available at time t from the first detection of the event.

Figure 14 illustrates, as an example, the estimation of spectral acceleration ordinates for different periods, for a M 7.0 event located at an epicentral distance of 50 km from the early warning target site [10].

We note the evolution of Sa predictions via the corresponding PDFs. The different panels correspond to increasing times from the earthquake origin and, therefore, to different numbers of stations triggered.

The False Alarm Issue

Once the EEWS provides a probability distribution of the ground motion intensity measure (IM) at the target site (e. g., peak ground acceleration or velocity), a decisional condition has to be checked in order to decide whether to alert or not.

Several options are available to formulate a decisional rule, for example the alarm may be issued if the probability of the predicted IM exceeding a critical threshold ($\mathrm{IM_C}$) is greater than a reference value (P_c):

$$\text{Alarm if}: \int_0^{\mathrm{IM_C}} f(\mathrm{IM}|\underline{d},\underline{s}) d(\mathrm{IM}) = P[\mathrm{IM} > \mathrm{IM_C}] > P_c\,. \quad (17)$$

The efficiency of the decisional rule may be evaluated in terms of false and missed alarms probabilities (known as the "cry wolf" issue, e. g., [20]). The false alarm occurs when, on the basis of the information processed by the EEWS, the alarm is issued while the intensity measure at the site $\mathrm{IM_T}$ (T subscript means "true", indicating the realization of the IM to be distinguished from the prediction $\mathrm{IM_C}$) is smaller than the threshold $\mathrm{IM_C}$. A missed alarm corresponds to not launching the alarm if needed,

$$\begin{aligned} &\text{Missed Alarm}: \{\text{No Alarm} \cap \mathrm{IM_T} > \mathrm{IM_C}\} \\ &\text{False Alarm}: \{\text{Alarm} \cap \mathrm{IM_T} \leq \mathrm{IM_C}\}\,. \end{aligned} \quad (18)$$

It has been discussed above how the information and the uncertainties on earthquake location and magnitude are dependent on the number of stations triggered at a certain time.

Therefore, in principle, the decisional rule may be checked at any time after the first station has triggered and, consequently, the false and missed alarm probabilities are also time dependent.

Using the decisional rule of (18) and considering PGA as IM, the time evolution of false/missed alarm probabilities has been simulated for the Campania EEWS, given the occurrence of a M 7 earthquake, and a target site at an epicentral distance of 110 km.

Figure 15 reports the missed and false alarm probabilities as a function of time from the first trigger at the ISNet network.

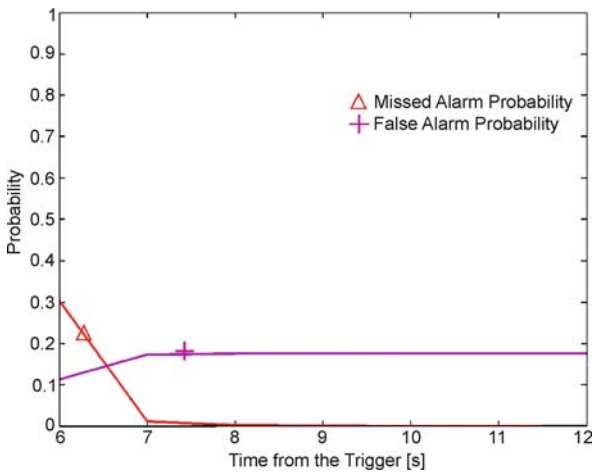

Earthquake Early Warning System in Southern Italy, Figure 15
Example of estimation of false and missed alarm probabilities as function of the time from the first trigger for a *M* 7.0 event with an epicentral distance of 110 km from the early warning target site. For the decisional rule adopted in this case the threshold is PGAc = 0.3 m/s^2 and the limit probability is $P_c = 0.2$ (after [19])

A Loss Estimation Approach to Early Warning

Magnitude and distance distributions conditioned to the measurements of the seismic network can also be used for a real-time estimation of risk, which includes losses produced by the earthquake [21]. Based on the real-time risk assessment, a security action aimed at risk mitigation is undertaken if the alarm is issued.

For example, some critical system could shut down or people in buildings may shelter themselves if the warning time is not sufficient to evacuate the dangerous buildings. More complex security measures may be related to the semi-active control of buildings [22].

Therefore, if an EEWS exists, it may trigger a security procedure in case of warning. The estimation of the expected losses for a specific building may be computed, for the case of warning issued and not issued respectively:

$$E^W[L|\underline{d},\underline{s}] = \int_L \int_{\mathrm{DM}} \int_{\mathrm{EDP}} \int_{\mathrm{IM}} l f^W(l|dm) f(dm|edp) \times f(edp|im) f(im|\underline{d},\underline{s}) \mathrm{d}L\mathrm{dDMdEDPdIM}\,, \quad (19)$$

where $f^W(l|dm)$ is the PDF of the loss (L) given the structural and non-structural damage vector (DM) reflecting the risk reduction in the case of warning; and $f^W(l|dm)$ is the loss function if no alarm is issued (no security action is undertaken); $f(dm|edp)$ is the joint PDF of damages given the Engineering Demand Parameters (EDP), proxy for the structural response; $f(edp|im)$ is the joint PDF of the EDPs conditioned to a vector of ground motion intensity measures (IM); $f(im|\underline{d},\underline{s})$ is the real-time hazard expressed by (16) in the case of a scalar IM.

Being able to compute, before the ground motion hits the site, the expected losses in case of warning *(W)* or not ($\bar{W}$), is relevant for taking the optimal decision, i. e., to alarm if this reduces the expected losses and to not issue any warning otherwise:

$$\text{Optimal decision:} \quad \begin{array}{l} \text{to alarm if } E^W[L|\underline{d},\underline{s}] \le E^{\bar{W}}[L|\underline{d},\underline{s}] \\ \text{to not alarm if } E^W[L|\underline{d},\underline{s}] > E^{\bar{W}}[L|\underline{d},\underline{s}] \end{array} \quad (20)$$

which is a better decisional rule in respect to that of (18).

Computing and comparing expected losses, conditioned to the real-time information coming from the EEWS, in the case of alarming or not, allows the determination of the alarm threshold above which it is convenient to issue the warning according to the optimally maximum criterion.

Assessment of average loss reduction determined by issuing an Early Warning provides a quantitative tool to evaluate the efficiency and feasibility of an EEWS.

Other potential advantages given by this approach are that: (a) the threshold may be set on a statistic (i. e., the summation of the logs) inferred from seismic network measurements, dramatically reducing the required computational effort for real-time decision making; (b) it minimizes the cry wolf issue reducing the probability of false and missed alarms thanks to threshold optimization. In fact, although the number of MA and FA depend on the decisional rule adopted to issued the alarm, the approach developed in Iervolino et al. [20,21,22] avoids explicitly considering the missed and false alarm rates associated with the decision, as the choice to alarm or not is taken based on the expected economic loss (not on the estimation of peak ground motion). In other words, if in computing the expected loss one accounts for the costs of false and missed alarms, there is no need to optimize the *cry wolf* issue, and MA and FA rates are at their values determined by the respective costs, and in this sense are optimal.

Future Directions

We have analyzed and illustrated the main scientific and technological issues related to the implementation and management of an earthquake early warning system under development in the Campania region of southern Italy.

The system is designed for early warning alert notification at distant coastal targets based on a dense, wide-dynamic seismic network (accelerometers, seismometers

and broadband sensors) deployed in the Apenninic belt region (ISNet – Irpinia Seismic Network). It can therefore be classified as a regional Early Warning System consisting of a wide seismic sensor network covering a portion or the entire area which is threatened by a quake's strike.

According to [25], real-time estimates of earthquake location and magnitude are needed for regional warning systems (EEWS), i. e., dense seismic networks covering all or a portion of an area of interest. However the alarm decision in an early warning system is based, rather, on the prediction, with quantified confidence, of a ground motion intensity at a distant target site (where a sensitive structure is located). This problem needs an evolutionary (i. e., time-dependent) and probabilistic frame where pdfs for earthquake location, magnitude and attenuation parameters are combined to perform a real-time probabilistic seismic hazard analysis (e. g., [19]).

Considering the peak displacement amplitude and/or predominant frequency measured in the early portion of *P*-waves, we have shown that suitable probability density functions for the earthquake location and magnitude parameters can be constructed and used for real-time probabilistic assessment of false alarms and loss estimation, which are the key elements based on which automatic actions can be undertaken to mitigate earthquake effects.

Based on the analysis of acceleration records of Euro-Mediterranean and Japanese earthquakes, Zollo et al. [49,50] have shown the advantages of using near source strong motion records for real time estimation of earthquake magnitude. In fact they provide unsaturated recordings of moderate to large earthquakes and, in case of dense station coverage of the source area, the combination of both *P*- and *S*-wave amplitude information can be used to get fast and robust earthquake location and magnitude estimates.

We support the use of *S*-waves recorded in the near-source of an impending earthquake for earthquake early warning, especially in view of the excellent correlation that *S*-peaks show with magnitude up to about $M = 7$ for Euro-Mediterranean and Japanese earthquakes [49,50]. Dense accelerometric networks now operating in Europe, USA, Taiwan, Japan and other seismic regions in the world can provide a sufficient number of records at distances smaller than 20–30 km from potentially damaging crustal earthquakes so that *S*–*P* times are expected to be smaller than 2–3 s. A magnitude estimation using *S*-waves could be therefore available 4–5 s after the first *P*-wave is recorded, which is still useful for sending an alert to distant target sites.

Although relatively few magnitude 7 and larger earthquakes have hit the Apenninic belt, and generally the Mediterranean region, during the last century, there have been many instances of damaging quakes in the magnitude 6 range.

Earthquake early warning systems have the potential to mitigate the effects of moderate size earthquakes ($M = 6$–7), which can produce severe damage in densely urbanized areas and places where old structures were not built to current standards. This has been the case for a significant number of earthquakes occurred in the Mediterranean basin during last decades: the 1976 Friuli ($M = 6$–6.5) and 1997 Colfiorito ($M = 6$) in Italy, 1999 Athens ($M = 5.9$) in Greece, 2002 Nahrin, in Afghanistan ($M = 6.1$), 2003 in Algeria ($M = 6.7$), 2003 Bam ($M = 6.3$) in Iran, 2004 in Morocco ($M = 6.4$).

An earthquake early warning system can be effective for mitigating the effects of moderate earthquakes. For moderate size events, early warning systems could also mitigate earthquake effects in terms of infrastructure operability (e. g., hospitals, firehouses, telecommunication hubs, ...) during the post-event emergency phase and rescue operations. For instance, in tall buildings, the higher floors generally sway much more than those near ground level, so that even a moderate earthquake could cause severe damage to a high rise. Therefore, even at 70–80 km distance from its epicenter, a magnitude 6 quake could affect hospital operating rooms and other critical installations.

Installations as close as 50 km from the epicenter could receive an earthquake warning 10 s prior to the arrival of the more energetic waves (*S* and surface waves) of an earthquake. To take advantage of this brief warning period, automated systems would have to be created that respond instantly to notification alert signals, and they would have to be carefully calibrated to avoid false or missed alarms. Closer to the epicenter, a magnitude 6 or higher earthquake can damage critical infrastructures, such as telephone lines, gas pipelines, highways, and railroads, as well as airport runways and navigation systems. These disruptions would have a domino effect in more distant areas, which could be mitigated by an early warning alert system, based on the earliest primary wave data to arrive at recording stations close to the epicenter.

Finally, we note that earthquake early warning systems can also help mitigate the effects of such earthquake-induced disasters as fires, explosions, landslides, and tsunamis, which can in many cases be more devastating than the earthquake itself. Systems could be installed at relatively low cost in developing countries, where moderate sized earthquakes can cause damage comparable to that caused by much larger earthquakes in developed countries.

Bibliography

Primary Literature

1. Allen RM (2007) The ElarmS earthquake early warning methodology and its application across California. In: Gasparini P, Manfredi G, Zschau J (eds) Earthquake early warning systems. Springer, Berlin, pp 21–44. ISBN-13 978-3-540-72240-3
2. Allen RM, Kanamori H (2003) The potential for earthquake early warning in Southern California. Science 300:786–789. doi:10.1126/science.1080912
3. Ambraseys N, Smit P, Douglas J, Margaris B, Sigbjornsson R, Olafsson S, Suhadolc P, Costa G (2004) Internet site for European strong-motion data. Boll Geofis Teor Appl 45(3):113–129
4. Bakun W, Fischer HF, Jensen E, VanSchaack J (1994) Early warning system for aftershocks. Bull Seismol Soc Am 84(2):359–365
5. Bernard P, Zollo A (1989) The Irpinia (Italy) 1980 earthquake: detailed analysis of a complex normal fault. J Geophys Res 94:1631–1648
6. Guidoboni E, Ferrari G, Mariotti D, Comastri A, Tarabusi G, Valensise G (2007) CFTI4Med, Catalogue of Strong Earthquakes in Italy (461 B.C.-1997) and Mediterranean Area (760 B.C.-1500). INGV-SGA. Available from http://storing.ingv.it/cfti4med/
7. Bose M, Ionescu C, Wenzel F (2007) Earthquake early warning for Bucharest, Romania: Novel and revised scaling relations. Geophys Res Lett 34:L07302. doi:10.1029/2007GL029396
8. Bouchon M (1979) Discrete wave number representation of elastic wave fields in three-space dimensions, J Geophys Res 84:3609–3614
9. Cinti FR, Faenza L, Marzocchi W, Montone P (2004) Probability map of the next $M \geq 5.5$ earthquakes in Italy. Geochem Geophys Geosyst 5:Q1103. doi:10.1029/2004GC000724.
10. Convertito V, Iervolino I, Giorgio M, Manfredi G, Zollo A (2008) Prediction of response spectra via real-time earthquake measurements. Soil Dyn Earthq Eng 28(6):492–505. doi:10.1016/j.soildyn.2007.07.006
11. Cornell CA (1968) Engineering seismic hazard analysis. Bull Seismol Soc Am 59(5):1583–1606
12. Cua G, Heaton T (2007) The virtual seismologist (VS) method: A Bayesian approach to earthquake early warning. In: Gasparini P, Manfredi G, Zschau J (eds) Earthquake early warning systems. Springer, Berlin. doi:10.1007/978-3-540-72241-0_7
13. Curtis A, Lomax A (2001) Prior information, sampling distributions and the curse of dimensionality. Geophysics 66:372–378. doi:10.1190/1.1444928
14. Erdik M, Fahjan Y, Ozel O, Alcik H, Mert A, Gul M (2003) Istanbul earthquake rapid response and the early warning system. Bull Earthquake Eng 1(1):157–163. doi:10.1023/A:1024813612271
15. Espinosa-Aranda JM, Jimenez A, Ibarrola G, Alcantar F, Aguilar A, Inostroza M, Maldonado S (1995) Mexico City seismic alert system. Seismol Res Lett 66:42–53
16. Font Y, Kao H, Lallemand S, Liu C-S, Chiao L-Y (2004) Hypocentral determination offshore Eastern Taiwan using the maximum intersection method. Geophys J Int 158(2):655–675. doi:10.1111/j.1365-246X.2004.02317.x
17. Grasso V, Allen RM (2005) Earthquake warning systems: Characterizing prediction uncertainty. Eos Trans AGU 86(52), Fall Meet. Suppl., Abstract S44B-03
18. Horiuchi S, Negishi H, Abe K, Kamimura A, Fujinawa Y (2005) An automatic processing system for broadcasting earthquake alarms. Bull Seism Soc Am 95(2):708–718. doi:10.1785/0120030133
19. Iervolino I, Convertito V, Giorgio M, Manfredi G, Zollo A (2006) Real time risk analysis for hybrid earthquake early warning systems. J Earthq Eng 10(6):867–885
20. Iervolino I, Convertito V, Giorgio M, Manfredi G, Zollo A (2007a) The cry wolf issue in seismic early warning applications for the campania region. In: Gasparini P et al (eds) Earthquake early warning systems. Springer, Berlin. doi:10.1007/978-3-540-72241-0_11
21. Iervolino I, Giorgio M, Manfredi G (2007b) Expected loss-based alarm threshold set for earthquake early warning systems. Earthq Eng Struc Dyn 36(9):1151–1168. doi:10.1002/eqe.675
22. Iervolino I, Manfredi G, Cosenza E (2007c) Earthquake early warning and engineering applications prospects. In: Gasparini P, Manfredi G, Szchau J (eds) Earthquake early warning systems. Springer, Berlin, doi:10.1007/978-3-540-72241-0_12
23. Jenny S, Goes S, Giardini D, Kahle H-G (2006) Seismic potential of Southern Italy. Tectonophysics 415:81–101. doi:10.1016/j.tecto.2005.12.003.
24. Kamigaichi O (2004) JMA earthquake early warning. J Japan Assoc Earthq Eng 3:134—137
25. Kanamori H (2005) Real-time seismology and earthquake damage mitigation. Ann Rev Earth Planet Sci 33:195–214. doi:10.1146/annurev.earth.33.092203.122626
26. Lancieri M, Zollo A (2008) A bayesian approach to the real-time estimation of magnitude from the early *P*- and *S*-wave displacement peaks. J Geophys Res. doi:10.1029/2007JB005386, in press
27. Lomax A (2005) A Reanalysis of the hypocentral location and related observations for the great 1906 California earthquake. Bull Seism Soc Am 95(3):861–877. doi:10.1785/0120040141
28. Meletti C, Patacca E, Scandone P (2000) Construction of a seismotectonic model: The case of Italy. Pure Appl Geophys 157:11–35
29. Montone P, Mariucci MT, Pondrelli S, Amato A (2004) An improved stress map for Italy and surrounding regions (central Mediterranean). J Geophys Res 109:B10410. doi:10.1029/2003JB002703
30. Munich Re (eds) (2005) Environmental report – perspectives – Today's ideas for tomorrow's world, WKD-Offsetdruck GmbH, München
31. Nakamura Y (1988) On the urgent earthquake detection and alarm system (UrEDAS). Proc 9th World Conf Earthquake Eng VII, Toyko, 673–678
32. Nakamura Y (1989) Earthquake alarm system for Japan railways. Japan Railway Eng 109:1–7
33. Nakamura Y (2004) Uredas, urgent earthquake detection and alarm system, now and future. 13th World Conference on Earthquake Engineering 908
34. Okada T et al (2003) A new method of quickly estimating epicentral distance and magnitude from a single seismic record. Bull Seismol Soc Am 93(1):526–532. doi:10.1785/0120020008
35. Podvin P, Lecomte I (1991) Finite difference computations of traveltimes in very contrasted velocity models: a massively parallel approach and its associated tools. Geophys. J Int. 105:271–284

36. Rydelek P, Pujol J (2004) Real-time seismic warning with a 2-station subarray, Bull Seism Soc Am 94(4):1546–1550. doi:10.1785/012003197
37. Satriano C, Lomax A, Zollo A (2008) Real-time evolutionary earthquake location for seismic early warning. Bull Seism Soc Am 98(3):1482–1494. doi:10.1785/0120060159
38. Simons F, Dando JB, Allen R (2006) Automatic detection and rapid determination of earthquake magnitude by wavelet multiscale analysis of the primary arrival. Earth Planet Sci Lett 250:214–223. doi:10.1016/j.epsl.2006.07.039
39. Teng TL, Wu Y-M, Shin TC, Tsai YB, Lee WHK (1997) One minute after: strong-motion map, effective epicenter, and effective magnitude. Bull Seism Soc Am 87(5):1209–1219
40. Tsukada S (2006) Earthquake early warning system in Japan. Proc 6th Joint Meeting UJNR Panel on Earthquake Research, Tokushima, Japan
41. Valensise G, Amato A, Montone P, Pantosti D (2003) Earthquakes in Italy: Past, present and future. Episodes 26(3):245–249
42. Wenzel FM et al (1999) An early warning system for Bucharest. Seismol Res Lett 70(2):161–169
43. Westaway R, Jackson J (1987) The earthquake of 1980 November 23 in Campania-Basilicata (southern Italy), Geophys J R Astron Soc 90:375–443. doi:10.1111/j.1467-6435.1999.tb00581.x
44. Wu Y-M, Teng T (2002) A virtual subnetwork approach to earthquake early warning. Bull Seismol Soc Am 92(5):2008–2018. doi:10.1785/0120040097
45. Wu Y-M, Kanamori H (2005) Experiment on onsite early warning method for the Taiwan early warning system. Bull Seismol Soc Am 95(1):347–353. doi:10.1785/0120040097
46. Wu YM, Zhao L (2006) Magnitude estimation using the first three seconds of *p*-wave amplitude in earthquake early warning. Geophys Res Lett 33:L16312. doi:10.1029/2006GL026871
47. Wurman G, Allen RM, Lombard P (2007) Toward earthquake early warning in Northern California. J Geophys Res 112:B08311. doi:10.1029/2006JB004830
48. Zhou H (1994) Rapid 3-D hypocentral determination using a master station method, J Geophys Res 99(B8):15439–15455
49. Zollo A, Lancieri M, Nielsen S (2006) Earthquake magnitude estimation from peak amplitudes of very early seismic signals on strong motion records. Geophys Res Lett 33:L23312. doi:10.1029/2006GL027795
50. Zollo A, Lancieri M, Nielsen S (2007) Reply to comment by P. Rydelek et al on "Earthquake magnitude estimation from peak amplitudes of very early seismic signals on strong motion records". Geophys Res Lett 34:L20303. doi:10.1029/2007GL030560.

Books and Reviews

Berger JO (1985) Statistical decision theory and Bayesian analysis. Springer, New York
Coutant O (1989) Program de simulation numerique AXITRA. Rapport LGIT, Grenoble, France
Gruppo di Lavoro MPS (2004) Redazione della mappa di pericolosità sismica prevista dall'Ordinanza PCM 3274 del 20 marzo (2003) Rapporto Conclusivo per il Dipartimento della Protezione Civile, INGV, Milano-Roma, aprile (2004) 65 pp + 5 appendici
Milne J (1886) Earthquakes and other earth movements. Appelton, New York, p 361

Earthquake Engineering, Non-linear Problems in

Mihailo D. Trifunac
Department of Civil Engineering,
University of Southern California, Los Angeles, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Vibrational Representation of Response
Response in Terms of Wave Propagation – An Example
Observations of Nonlinear Response
Future Directions
Bibliography

Glossary

Meta-stability of man-made structures is the consequence of their upright construction above ground. For excessive dynamic (earthquake) loads, when the lateral deflection exceeds some critical value (this is normally accompanied by softening nonlinear behavior of the structural members), the overturning moment of the gravity forces becomes larger than the restoring moment, and the structure becomes unstable and moves exponentially toward collapse.

Complex and evolving structural systems are structures with a large number of degrees of freedom and many structural members, which for given loads experience softening nonlinear deformations. During strong excitation, continuous changes (typically decreases) in effective stiffness and time-dependent changes in boundary conditions result in a system whose properties are changing with time.

Soil–structure interaction is a process in which the soil and the structure contribute to mutual deformations while undergoing dynamic response. In time, with continuously changing contact area between the foundation and the soil (opening and closing of gaps), when the deformations are large, soil–structure interaction is characterized by nonlinear geometry and nonlinear material properties in both the soil and in the structure.

Definition of the Subject

Nonlinear problems in structural earthquake engineering deal with the dynamic response of meta-stable, man-

made buildings subjected to strong earthquake shaking. During earthquakes, structures constructed on soft sediments and soils deform together with the underlying soil in the dynamic process called soil–structure interaction. Strong shaking forces the soil–structure systems to evolve through different levels of nonlinear response, with continuously changing properties that depend upon the time history of excitation and on the progression and degree of damage. Thus far, the analyses of this response have used the vibrational approach and lumped mass discrete models to represent real structures. Loss of life and property, however, continue to be high during strong shaking in the vicinity of the faults responsible for earthquakes. This calls for new, more physically refined methods of analysis, which can be based on nonlinear wave propagation, and for balancing of the structural capacities with the power carried by the earthquake waves.

After a brief discussion of the literature on the complex and chaotic dynamics of simple mechanical oscillators, the dynamic characteristics and governing equations in the meta-stable structural dynamics of earthquake engineering are introduced. The nature of the solutions of the governing equations in terms of both the vibrational and the wave representations is discussed, and the dynamic instability, material and geometric nonlinearities, and complexities of the governing equations associated with nonlinear soil–structure interaction are described. Collectively, the examples presented reflect the complex physical nature of meta-stable structural systems that experience nonlinear dynamic response, the characteristics of which change and evolve during earthquake excitation.

Introduction

Earthquake engineering, through a cooperation of structural and geotechnical engineers with seismologists and geologists, aims to develop methods for safer design of man-made structures to withstand shaking near intermediate and large earthquakes. This requires addressing the problems of predictability of the response of complicated nonlinear systems, which is one of the important subjects of modern nonlinear science. Through the studies of the dynamic response, earthquake engineers address complex physical problems and issues with important social implications.

The completeness and beauty of the linear differential equations appear to have led to their dominance in the mathematical training of engineers and scientists during most of the 20th century. The recognition that chaotic dynamics is inherent in all nonlinear physical phenomena, which has created a sense of revolution in applied mechanics and physics today, so far has had little if any effect on the research and design of earthquake-resistant structures. In the past, the designs in structural engineering and control systems were kept within the realm of linear system dynamics. However, the needs of modern technology have pushed the design into the nonlinear regimes of large deformations, which has increased the possibility of encountering chaotic dynamic phenomena in structural response. Even a cursory review of papers on chaotic vibrations in mechanical systems leads to the conclusion that chaotic dynamics is not a small, insignificant class of motions and that chaotic oscillations occur in many nonlinear systems and for a wide range of values of the parameters.

If an engineer chooses parameters that produce chaotic output, then he or she loses predictability. However, the chaotic behavior of nonlinear systems does not exclude predictability of the response but rather introduces upper bounds (prediction horizons) [31] and renders the predictions probabilistic. The important question is then over what time-scale are the forecasts reliable, given the current state and knowledge of the system. Another key ingredient for prediction is an adequate physical model. At present, because of the multitude of interacting phenomena and the absence of physically complete equations of motion, there exists no adequate general model of the complete earthquake response process. While the practical outcome of most work in earthquake engineering remains empirically based, the nonlinear methods are gaining popularity, aiming to decipher the governing phenomena and to assess the reliability of the models. It appears now that the broad-based revolution in the worldview of science that begun in the twentieth century will be associated with chaotic dynamics [43]. This revolution should eventually also contribute to better understanding and more complete representation of the response analyses in earthquake engineering.

It has been argued that major changes in science occur not so much when new theories are advanced but when the simple models with which scientists conceptualize a theory are changed [24]. In vibrations, such a conceptual model that embodies the major features of a whole class of problems is the spring-mass system. Lessons emerging from studies of the spring-mass model and several other relevant models can serve as conceptual starting points for generalizations and also as a guide to further studies of more complex models in earthquake engineering and structural dynamics.

Studies of forced vibrations of a pendulum have revealed complex dynamics and chaotic vibrations [14,15]. A simply supported beam with sub-buckling axial com-

pression modeled by a single mode approximation yields a Mathieu type equation and for certain values of the parameters leads to unstable solutions. When nonlinearities are added, these vibrations result in a limit cycle. A related problem is a classical pendulum with a vibrating pivot support, which also leads to chaotic vibrations [29,34]. Chaotic motions in a double pendulum have been studied by Richter and Scholz [45], and the complex dynamics and chaotic solutions for a spherical pendulum with two degrees of freedom have been described by Miles [35].

Impact-type problems result in explicit difference equations or maps, which can yield chaotic vibrations for certain values of the governing parameters [30]. A mass vibrating in a gap between two stiff springs on either side [17,46,47] is a simple related model, which suggests a starting points for research in nonlinear vibration of piles, and for impact-type interaction of adjacent buildings, excited by strong earthquake ground motion. The reader can find examples of such problems in the description of damage in Mexico City, for example, during several earthquakes [32].

Chaotic motions of an elasto-plastic arch have been studied by Poddar et al. [42]. Forced vibrations of a buckled beam, modeled by the Duffing equation, showed that chaotic vibrations are possible [16]. Forced vibrations described by a Duffing equation with viscous damping and nonlinear (cubic) elastic (stiffening) spring were studied by Ueda [67]. Figure 1 summarizes his results and describes the regions of chaotic, periodic (I, II, etc.), and subharmonic (m/n) motions as functions of the damping and forcing amplitudes. This simple equation, representing a hardening spring system, has direct analogues in the dynamics of piles and in the rocking of buildings, both following the strong-motion phase of earthquake shaking after horizontal gaps have been created between the pile (foundation walls) and the soil [63].

A mechanical system with a nonlinear restoring force and with a control force added to move the system according to some prescribed signal has been studied by Holmes and Moon [19] and Holmes [18]. It was shown that such a system exhibits both periodic limit-cycle oscillation and chaotic motions. Chaotic vibrations in continuous beams have been studied for nonlinear body forces and nonlinear boundary conditions (that depend on the motion), and for motions large enough for the nonlinear terms in the equations of motion to be significant [37,38,39,40,41]. Forced planar vibrations of nonlinear elastica [35,36], were shown to become unstable and exhibit chaotic motions under certain conditions.

The above-mentioned studies imply that there is a conflict in the classical engineering description of the world.

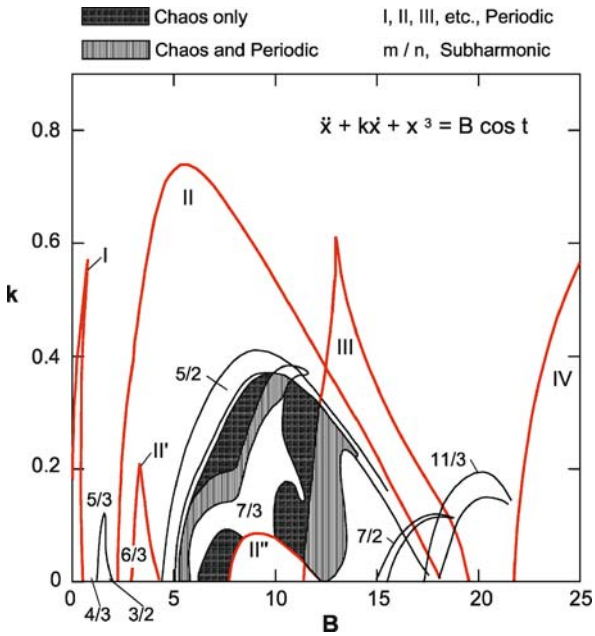

Earthquake Engineering, Non-linear Problems in, Figure 1
Chaos diagram showing regions of chaotic, chaotic and periodic, periodic (I, II, III, etc.), and sub-harmonic (4/3, 3/2, 5/3, etc.) motions for a nonlinear equation as functions of non-dimensionalized damping and forcing amplitude (from [67])

One aspect of this conflict is the assumption that nature is a deductive system, moving forward in time according to deterministic laws. Another aspect is that a scientist attempting to model portions of the world from finite data projects unverifiable structure onto the local environment. The conflict is that these two views do not match, leaving us with a question: what are models good for? There are many systems in nature that are observed to be chaotic, and for which no adequate physical model exists. Whether a model is adequate or not depends, of course, on the questions asked [7]. Unfortunately, the art of dynamical modeling is often neglected in discussions of nonlinear and chaotic systems, in spite of its crucial importance [1]. In the following, the modeling problem in earthquake engineering will be illustrated using two common approaches to the solution, one based on an equivalent oscillator and the other one using wave representation.

Stochastic processes have been developed to describe irregular phenomena in deterministic systems that are too complicated or have too many variables to be fully described in detail. For example, stochastic processes have been used to model the response of structures to earthquake and wind forces, which are deterministic, and in principle could be completely described. In practice, the

stochastic modeling has been used also as an approximate description of a deterministic system that has unknown initial conditions and may be highly sensitive to the initial conditions. In trying to model real systems, as a result of the modeling process, we sometimes obtain a model that shows very regular behavior, while the real system has very irregular behavior. In that case, random noise is added to the model, but this represents no more than our lack of knowledge of the system structure or the inadequacy of the identification procedure [22].

In earthquake engineering, the complexity of the multi-dimensional real world is reduced to a sub-space, which is defined by (1) the dimensions and properties of the adopted mathematical models, (2) the nature of the adopted boundary conditions, and (3) the method of solution. A linear mechanical system cannot exhibit chaotic vibrations, and for periodic inputs it produces periodic outputs. The chaotic system must have nonlinear elements or properties, which can include, for example, (1) nonlinear elastic or spring elements; (2) nonlinear damping (such as stick-slip friction); (3) backlash, play, or bilinear springs; and (4) nonlinear boundary conditions. The nonlinear effects can be associated with the material properties, with the geometric effects, or both. In the following, the consequences of unorthodox boundary conditions and nonlinear waves in a building will be used to illustrate the extensions and complexities associated with evolving systems. The utility of this complexity can be viewed as the arbiter of the order and randomness.

Vibrational Representation of Response

The first modern uses of mechanics in problems of earthquake engineering appeared during the early 1900s, following the earthquake disasters in San Francisco (1906), Messina-Reggio (1908), and Tokyo (1923) and the realization that something needed to be done to prevent such losses of life and property during future events. The first practical steps consisted of introducing the *seismic coefficient* (*shindo* in Japan, and *rapporto sismico* in Italy). This was followed by earthquake-resistant design codes, first adopted in Japan in 1923, and then in California in 1934 [44]. During the same period, there also appeared the first studies of the effects of earthquake shaking on structures in terms of simple mechanical oscillators [48], and in the early 1930s the modern theory based on the response spectrum method was introduced [2,3,4]. These early developments follow the deterministic formulations of Newtonian mechanics and employ linear models and equations of motion.

Elementary Vibrational Representation of Response

The basic model employed to describe the response of a simple structure to only horizontal earthquake ground acceleration, $\ddot{\Delta}_x$, is a single-degree-of-freedom system (SDOF) that experiences rocking ψ_r relative to the normal to the ground surface. The model also assumes that the ground does not deform in the vicinity of the foundation—that is, it neglects the soil–structure interaction (Fig. 2). The rotation ψ_r is restrained by a spring with stiffness K_r and by a dashpot with rocking damping constant C_r, providing the fraction of critical damping ς_r. The natural frequency of this system is $\omega_r = \left(K_r/h^2 m_b\right)^{1/2}$, and for small rocking angles it is governed by the linear ordinary differential equation

$$\ddot{\psi}_r + 2\omega_r\varsigma_r\dot{\psi}_r + \omega_r^2\psi_r = -\ddot{\Delta}_x/h\,. \tag{1}$$

For any initial conditions, and for arbitrary excitation, this system always leads to a deterministic and predictable response. Equation (1) was used originally to develop the concept of relative response spectrum and continues to this day as the main vehicle in formulation of most earthquake engineering analyses of response [56]. If the gravity force is considered, ω_r in Eq. (1) has to be reduced [5]. The system described by Eq. (1) is meta-stable for ψ_r smaller than its critical value. At the critical value of ψ_r, the over-

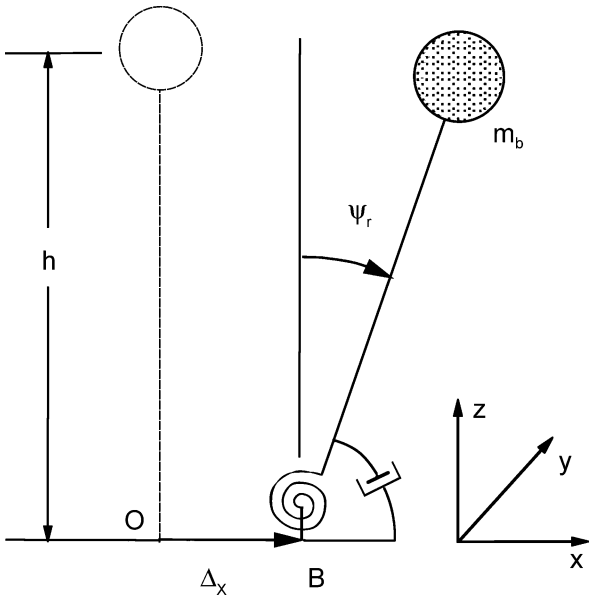

Earthquake Engineering, Non-linear Problems in, Figure 2
Single-degree-of-freedom system (SDOF) representation of a building (inverted pendulum) with equivalent mass m_b and mass-less column of height h, experiencing rocking ψ_r due to horizontal motion of its base Δ_x

turning moment of the gravity force is just balanced by the elastic moment in the restraining spring, and for values greater than the critical value the system becomes unstable.

Advanced Vibrational Representation of Response

In more advanced vibrational representations of the response, additional components of the earthquake excitation, structural dynamic instability, soil–structure interaction, spatial and temporal variations of the excitation, differential motions at different support points, and nonlinear behavior of the stiffness K_r can be considered, but the structure usually continues to be modeled by mass-less columns, springs, and dashpots, and with a rigid mass m_b. In the following, we illustrate some of the above-mentioned cases.

Dynamic Instability An example of a simple model that includes instability is shown in Fig. 3. It experiences horizontal, vertical, and rocking excitations, which can result, for example, from incident P and SV waves. The structure

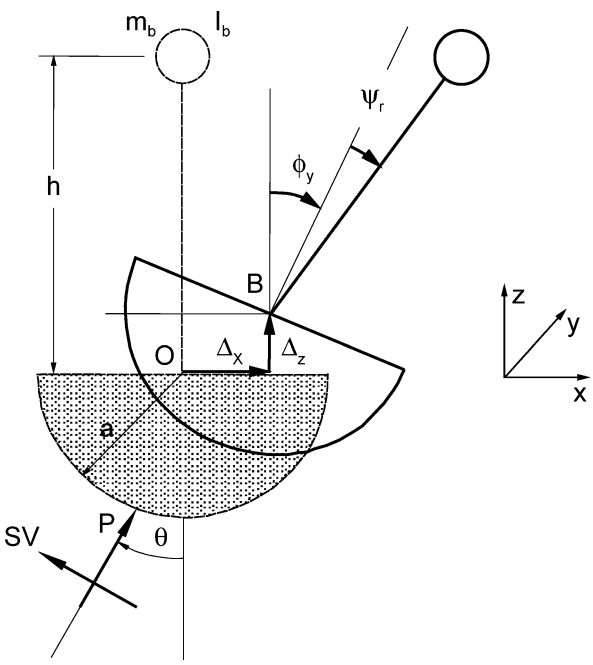

Earthquake Engineering, Non-linear Problems in, Figure 3
Single-degree-of-freedom system (SDOF) representation of a building (inverted pendulum), with equivalent mass m_b, moment of inertia (about O) I_b, and a mass-less column of height h, experiencing relative rocking ψ_r due to horizontal, vertical, and rocking motions of its foundation (Δ_x, Δ_z, and ϕ_y), which result from soil–structure interaction when excited by incident wave motion

is represented by an equivalent single-degree-of-freedom system, with a concentrated mass m_b at height h above the foundation. It has a radius of gyration r_b and a moment of inertia $I_b = m_b r_b^2$ about point O. The degree-of-freedom in the model is chosen to correspond to the relative rocking angle ψ_r. This rotation is restrained by a spring with rocking stiffness K_r and by a dashpot with rocking damping C_r (both not shown in Fig. 3), and the gravitational force $m_b g$ is considered. Taking moments about B results in the equation of motion

$$\ddot{\phi}_y + \ddot{\psi}_r + 2\omega_r \varsigma_r \dot{\psi}_r + \omega_r^2 \psi_r = \left\{ -\left(\ddot{\Delta}_x/a\right)\cos\left(\phi_y + \psi_r\right) + \left(\omega_r^2 \varepsilon_g + \ddot{\Delta}_z/a\right)\sin\left(\phi_y + \psi_r\right)\right\}/\varepsilon\,, \quad (2)$$

where $\varepsilon = h\left(1 + (r_b/h)^2\right)/a, \omega_r^2 = K_r/\left[m\left(h^2 + r_b^2\right)\right], \omega_r$ is the natural frequency of rocking, ζ_r is a fraction of critical damping in $2\omega_r\varsigma_r = C_r/\left[m\left(h^2 + r_b^2\right)\right]$, and $\varepsilon_g = 2/\omega_r^2 a$. Equation (2) is a differential equation coupling the rocking of the foundation, ϕ_y, and of the structure, ψ_r, with the horizontal and vertical motions of the foundation. It is a nonlinear equation the solution to which requires numerical analysis. In this example, we will discuss only the case in which $\phi_y + \psi_r$ is small. Then,

$$\ddot{\psi}_r + 2\omega_r\varsigma_r\dot{\psi}_r + \left\{\omega_r^2\left(1 - \varepsilon_g/\varepsilon\right) - \ddot{\Delta}_z/\varepsilon a\right\}\psi_r = -\ddot{\phi}_y + \left\{-\ddot{\Delta}_x/a + \left(\omega_r^2\varepsilon_g + \ddot{\Delta}_z/a\right)\phi_y\right\}/\varepsilon\,. \quad (3)$$

For steady-state excitation by incident P and SV waves with frequency ω, Δ_x, ϕ_y, and Δ_z, and therefore the forcing function of Eq. (3), will be periodic. Equation (3) is then a special form of the Hill's equation. Analysis of the stability of this equation can be found in the work of Lee [25]. For general earthquake excitation, Δ_x, ϕ_y, and Δ_z will be determined by the recorded components of motion, and in predictive analyses by simulated ground motions [27,28,70].

In Eq. (3), ϕ_y describes rocking of the foundation to which the structure is attached. In analyses that do not consider soil–structure interaction, ϕ_y will be determined directly by the rocking component of strong ground motion [21,28], and in studies that consider soil–structure interaction ϕ_y will be one of the variables to be determined by the analysis [25].

Soil–Structure Interaction The problem of linear soil–structure interaction embodies the phenomena that result from (1) the presence of an inclusion (foundation, Fig. 4) in the soil [26], and (2) the vibration of the structure supported by the foundation, which exerts dynamic forces on the foundation [25]. Examples and a discussion of the nonlinear aspects of soil–structure interaction can be found

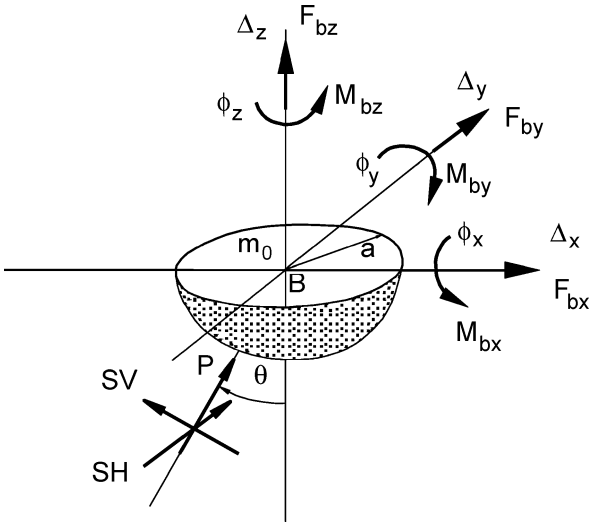

Earthquake Engineering, Non-linear Problems in, Figure 4
Six components of motion (three translations and three rotations) $\{\Delta_x, \Delta_y, \Delta_z, \phi_x, \phi_y, \phi_z\}$ of point *B*, and six components of force (three forces and three moments) $\{F_{ext}\} = \{F_{bx}, F_{by}, F_{bz}, M_{bx}, M_{by}, M_{bz}\}$, that the structure exerts on the foundation at *B*

in Gicev [9] and in a review of observations of response to earthquake shaking in full-scale structures in Trifunac et al. [63,64,65].

The dynamic response of a rigid, embedded foundation to seismic waves can be separated into two parts. The first part corresponds to the determination of the restraining forces due to the motion of the inclusion, usually assumed to be a rigid body. The second part deals with the evaluation of the driving forces due to scattering of the incident waves by the inclusion, which is presumed to be immobile. This can be illustrated by considering a foundation embedded in an elastic medium and supporting an elastic superstructure. The steady-state harmonic motion of the foundation having frequency ω can be described by a vector $\{\Delta_x, \Delta_y, \Delta_z, \phi_x, \phi_y, \phi_z\}^T$ (Fig. 4), where Δ_x and Δ_y are horizontal translations, Δ_z is vertical translation, ϕ_x and ϕ_y are rotations about horizontal axes, and ϕ_z is torsion about the vertical axis. Using superposition, displacement of the foundation is the sum of two displacements:

$$\{U\} = \{U^*\} + \{U_0\} \ , \tag{4}$$

where $\{U^*\}$ is the foundation input motion corresponding to the displacement of the foundation under the action of the incident waves in the absence of external forces, and $\{U_0\}$ is the relative displacement corresponding to the displacement of the foundation under the action of the external forces in the absence of incident wave excitation.

The interaction force $\{F_s\}$ generates the relative displacement $\{U_0\}$, which corresponds to the force that the foundation exerts on the soil and that is related to $\{U_0\}$ by $\{F_s\} = [K_s(\omega)]\{U_0\}$, where $[K_s(\omega)]$ is the 6×6 complex stiffness matrix of the embedded foundation. It depends upon the material properties of the soil medium, the characteristics and shape of the foundation, and the frequency of the harmonic motion, and it describes the force-displacement relationship between the rigid foundation and the soil medium.

The driving force of the incident waves is equal to $\{F_s^*\} = [K_s]\{U^*\}$, where the input motion $\{U^*\}$ is measured relative to an inertial frame. The "driving force" is the force that the ground exerts on the foundation when the rigid foundation is kept fixed under the action of the incident waves. It depends upon the properties of the foundation and the soil and on the nature of excitation.

The displacement $\{U\}$ is related to the interaction and driving forces via $[K_s]\{U\} = \{F_s\} + \{F_s^*\}$. For a rigid foundation having a mass matrix $[M_0]$ and subjected to a periodic external force, $\{F_{ext}\}$, the dynamic equilibrium equation is

$$[M_0]\{\ddot{U}\} = -\{F_s\} + \{F_{ext}\} \ , \tag{5}$$

where $\{F_{ext}\} = \{F_{bx}, F_{by}, F_{bz}, M_{bx}, M_{by}, M_{bz}\}$ is the force the structure exerts on the foundation (Fig. 4). Then, Eq. (5) becomes

$$[M_0]\{\ddot{U}\} + [K_s]\{U\} = \{F_s^*\} + \{F_{ext}\} \ . \tag{6}$$

The solution of $\{U\}$ requires the determination of the mass matrix, the impedance matrix, the driving forces, and the external forces [25].

After the mass matrix $[M_0]$, the stiffness matrix $[K_s]$, and the force $\{F_s^*\}$ have all been evaluated, they can be used to determine the foundation displacement $\{U\}$. For in-plane response excited by P and SV waves, for example, the relative response ψ_r is then given by Eq. (3).

Differential Motions Common use of the response spectrum method [56] and many dynamic analyses in earthquake engineering implicitly assume that all points of building foundations move synchronously and with the same amplitudes. This, in effect, implies that the wave propagation in the soil is neglected. Unless the structure is long (e. g., a bridge with long spans, a dam, a tunnel) or "stiff" relative to the underlying soil, these simplifications are justified and can lead to a selection of approximate design forces if the effects of soil-foundation interaction in the presence of differential ground motions can be

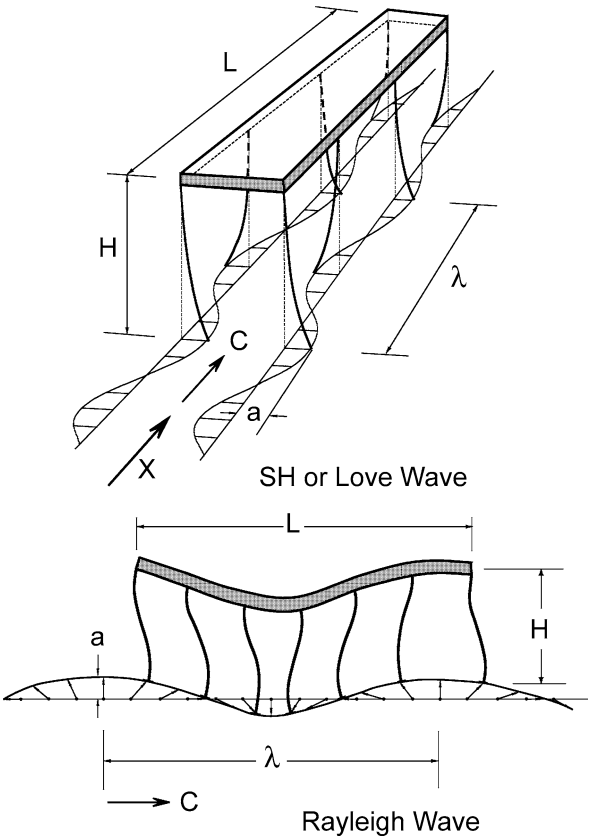

Earthquake Engineering, Non-linear Problems in, Figure 5
Schematic representation of the deformation of columns accompanying differential wave excitation of long structures for out-of-plane response (*top*) and in-plane response (*bottom*) when SH or Lave waves (*top*) or Rayleigh waves (*bottom*) propagate along the longitudinal axis of a structure

neglected [6]. Simple analyses of two-dimensional models of long buildings suggest that when $a/\lambda < 10^{-4}$, where a is wave amplitude and λ is the corresponding wavelength, the wave propagation effects on the response of simple structures can be neglected [50].

Figure 5 illustrates the "short" waves propagating along the longitudinal axis of a long building or a multiple-span bridge. For simplicity, the incident wave motion has been separated into out-of-plane motion (Fig. 5, top), consisting of SH and Love waves, and in-plane motion (Fig. 5, bottom) consisting of P, SV, and Rayleigh waves. The in-plane motion can further be separated into horizontal (longitudinal), vertical, and rocking components, while out-of-plane motion consists of horizontal motion in the transverse direction and torsion along the vertical axis. Trifunac and Todorovska [61] analyzed the effects of the horizontal in-plane components of differential motion for buildings with models that are analogous to the sketch in Fig. 5 (bottom), and they showed how the response spectrum method can be modified to include the first-order effects of differential motions. Trifunac and Gicev [59] showed how to modify the spectra of translational motions, into a spectrum that approximates the total (translational and torsional) responses, and how this approximation is valid for strong motion waves an order of magnitude longer than the structure ($\lambda \gg L$).

As can be seen from the above examples the differential motions lead to complex excitation and deformation of the structural members (columns, shear walls, beams, braces), increase the dimensions of the governing differential equations, lead to thee-dimensional dynamic instability problems, and can lead to nonlinear boundary conditions. These are all conditions that create an environment in which, even with the most detailed numerical simulations, it is difficult to predict all of the complexities of the possible responses.

Nonlinear Vibrational Analyses of Response

For engineering estimation of the maximum nonlinear response of a SDOF system, u_m, in terms of the maximum linear response, u_0, it is customary to specify a relation between u_m and u_0 (Fig. 6). By defining the yield-strength reduction factor as $R_y = u_0/u_y$, where u_y is the yielding displacement of the SDOF system equivalent spring, and ductility as $\mu = u_m/u_y$, for the same ground motion the ratio u_m/u_0 is then equal to μ/R_y. Veletsos and New-

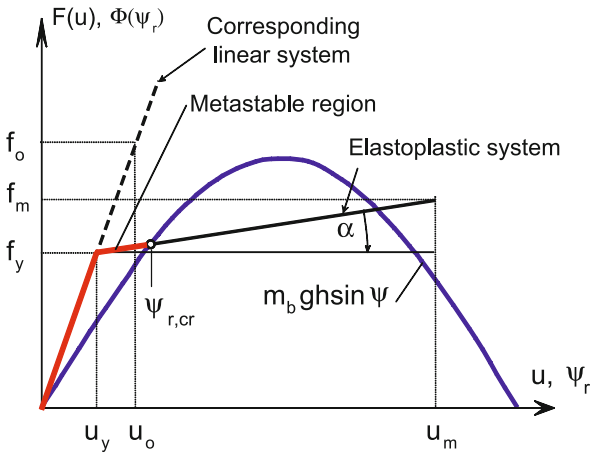

Earthquake Engineering, Non-linear Problems in, Figure 6
Bi-linear representation of stiffness (yielding at (u_y, f_y)), overturning moment of gravity force ($m_b g \sin \psi$), critical rocking angle $\psi_{r,cr}$, and meta-stable region ($0 < \psi_r < \psi_{r,cr}$) for an SDOF system

mark [68,69] showed that (1) for a long-period SDOF system when its natural period $T_n = 2\pi/\omega_n$ becomes very long, u_m / u_0 tends toward 1 and R_y approaches μ (equal deformation rule); (2) for the response amplitudes governed mainly by the peak excitation velocities, u_m / u_0 can be approximated by $\mu/\sqrt{2\mu - 1}$ and R_y by $\sqrt{2\mu - 1}$ (equal strain energy rule); and (3) for a high-frequency (stiff) system when $T_n \sim 0$, $R_y \sim 1$.

Complexities of Simultaneous Action of Dynamic Instability, Nonlinearity, and Kinematic Boundary Conditions – Example The model we illustrate next is an SDOF when it is excited by synchronous horizontal ground motion at its two supports (1 and 2 in Fig. 7), but it behaves like a three-degree-of-freedom (3DOF) system when excited by propagating horizontal, vertical, and rocking ground motions. For such a system, the above classical equal energy and equal displacement rules for SDOF system will not apply.

The goals here are to describe the effects of differential motion on strength-reduction factors R_y of the simple structure shown in Fig. 7 when it is subjected to all of the components of near-source ground motions, and to illustrate the resulting complexities of nonlinear response. Analyses of the consequences of the differences in ground motion at structural supports, caused by non-uniform soil properties, soil–structure interaction, and lateral spreading, for example, will further contribute to the complexities of the response, but these factors will not be discussed here.

The original response spectrum method was formulated using a vibrational solution of the differential equation of an SDOF system excited by synchronous, and only horizontal (one component), ground motion. The consequences of simultaneous action of all six components of ground motion (three translations and three rotations) on the relative response of an SDOF system are still rarely considered in modern engineering design [58], even though it has been 75 years since the original response spectrum method was formulated and about 40 years since it became the principal tool in engineering design [56]. Because the response spectrum method has become an essential part of the design process and of the description of how strong motion should be specified for a broad range of design applications [52], we hope that the present examples will help to further understanding of the complexities of response in more realistic models of structures.

The nature of the relative motion of individual column foundations or of the entire foundation system will depend upon the type of foundation, the characteristics of the soil surrounding the foundation, the type of incident

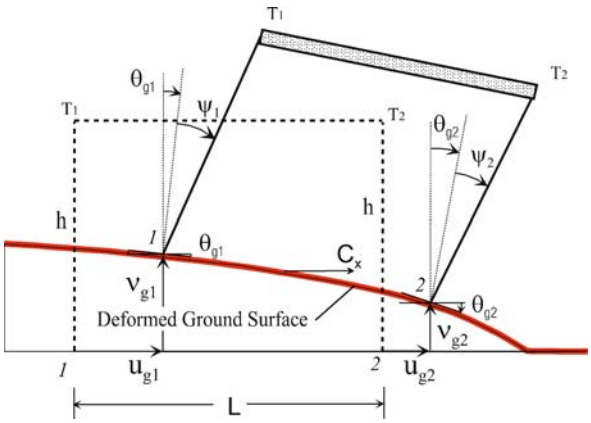

Earthquake Engineering, Non-linear Problems in, Figure 7
The structure deformed by the wave, propagating from left to right, with phase velocity C_x, for the case of $+v_{g_i}$ ("up" motion). Different column rotations ψ_1 and ψ_2 result from different translations and rotations at supports 1 and 2 (from [21])

waves, and the direction of wave arrival, with the motion at the base of each column having six degrees of freedom. In the following example, we assume that the effects of soil–structure interaction are negligible; consider only the in-plane horizontal, vertical, and rocking components of the motion of column foundations; and show selected results of the analysis for a structure on only two separate foundations. We assume that the structure is near the fault and that the longitudinal axis of the structure (X axis) coincides with the radial direction (r axis) of the propagation of waves from the earthquake source, so that the displacements at the base of columns are different as a result of the wave passage alone. We suppose that the excitations at the piers have the same amplitude but different phases and that the phase difference (or time delay) will depend upon the distance between the piers and the horizontal phase velocity of the incident waves.

The simple model we consider, which is described in Fig. 7, represents a one-story structure consisting of a rigid mass, m, with length L, supported by two rigid, mass-less columns with height h, which are connected at the top to the mass and at the bottom to the ground by rotational springs (not shown in Fig. 7). The stiffness of the springs, k_ϕ, is assumed to be elastic-plastic, as in Fig. 6, without hardening ($\alpha = 0$). The mass-less columns are connected to the ground and to the rigid mass by rotational dashpots, c_ϕ, providing a fraction of critical damping equal to 5 percent. Rotation of the columns, $\phi_i = \theta_{g_i} + \psi_i$ for $i = 1, 2$, which is assumed to be not small, leads us to consider the geometric nonlinearity. The mass is acted upon by the acceleration of gravity, g, and is excited by differential hori-

zontal, vertical, and rocking ground motions, u_{g_i}, v_{g_i}, and θ_{g_i}, $i = 1, 2$ (Fig. 7) at the two bases, so that

$$u_{g_2}(t) = u_{g_1}(t-\tau)\,; \quad v_{g_2}(t) = v_{g_1}(t-\tau)\,;$$
$$\theta_{g_2}(t) = \theta_{g_1}(t-\tau)\,; \quad \tau = L/C_x\,,$$

with τ being the time delay between the motions at the two piers and C_x the horizontal phase velocity of the incident waves. The functional forms of u_{g_i}, v_{g_i}, and θ_{g_i} are defined by the near-source ground motions [21], and the rocking component of the ground motion is approximated by [28] $\theta_{g_i}(t) = -\dot{v}_{g_i}(t)/C_x$, where $\dot{v}_{g_i}(t)$ is the vertical velocity of the ground motion at the ith column. Of course, in a more accurate modeling, the ratio of the v_{g_i} to u_{g_i} amplitudes will depend upon the incident angle and the character of incident waves, while the associated rocking θ_{g_i} will be described by a superposition of the rocking angles associated with incident body and dispersed surface waves [28].

The yield-strength reduction factor for the system subjected to synchronous ground motion is $R_y = f_0/f_y = u_0/u_y$, where all of the quantities are defined in Fig. 6. In this example, for the assumed model and because of the differential ground motions and rotation of the beams, the relative rotation for the two columns at their top and bottom will be different. Therefore, it is necessary to define the R-factor and ductility for each corner of the system, instead of one factor for the entire system. In all calculations here, we consider the actions of the horizontal, vertical, and rocking components of the ground motion, the effects of gravity force, dynamic instability, and geometric nonlinearity. For the structure in Fig. 7, we calculate maximum linear and nonlinear relative rotations at four corners of the system under downward ($-v_{g_i}$), radial, and rocking, and upward ($+v_{g_i}$), radial and rocking near-source differential ground motions corresponding to a given earthquake magnitude, ductility μ, and for different time delays, τ. Then we plot R_y versus T_n for the four corners of the system.

Figure 8 illustrates typical results for R_y versus the oscillator period for near-source, fault-parallel displacement $d_N(t) = A_N(1 - e^{-t/\tau_N})/2$ [21], with downward vertical ground displacement, magnitude $M = 8$, for a ductility ratio of 8 and a time delay of $\tau = 0.05$ s. It shows the results for the top-left, top-right, bottom-left, and bottom-right corners of the system, assuming wave propagation from left to right (see Fig. 7). For reference and easier comparison with the previously published results, we also plot one of the oldest estimates of R_y versus period, using piecewise straight lines [21]. The curve $(R_y)_{min}$ shows the minimum values of R_y for $d_N(t)$ motion with $-v_{g_i}$, and for $M = 8$, $\mu = 8$, and $\tau = 0.05$ s.

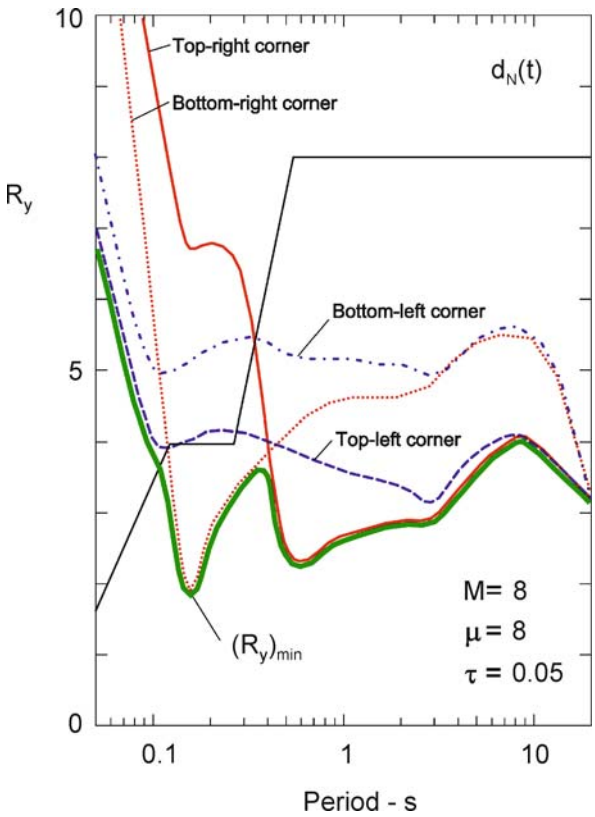

Earthquake Engineering, Non-linear Problems in, Figure 8
Example of the effects of the differential ground motion on the strength-reduction factors R_y at the four corners of the structure in Fig. 7, subjected to horizontal, vertical, and rocking components of the fault-parallel displacement, for downward vertical motion ($-v_{g_i}$) for earthquake magnitude $M = 8$, ductility $\mu = 8$, and delay at the right support $\tau = 0.05$ s. The amplitudes of the piecewise straight representation of the classical R_y are shown for comparison [21]. $(R_y)_{min}$ shows the smallest values of the R-factors, which for the set of conditions in this example are determined by the response at the *top left corner* (for periods shorter than 0.1 s), at the *bottom right corner* (for periods between 0.1 and 0.35 s), and at the *top right corner* (for periods longer than 0.35 s)

For periods longer than 5 to 10 s, R_y curves approach "collapse boundaries" [21]. This is implied in Fig. 8 by the rapid decrease of R_y versus period for periods longer than about 7 s. At or beyond these boundaries, the nonlinear system collapses due to the action of gravity loads and dynamic instability.

The complex results illustrated in Fig. 8 can be simplified by keeping only $(R_y)_{min}$, since it is only the minimum value of R_y that is needed for engineering design. By mapping $(R_y)_{min}$ versus period of the oscillator for different earthquake magnitudes, M, different ductilities, μ, and different delay times, τ, design criteria can be formulated

for design of simple structures to withstand near-fault differential ground motions [21]. Nevertheless, the above shows how complicated the response becomes even for as simple a structure as the one shown by the model in Fig. 7, when differential ground motion with all of the components of motion is considered. In this example, this complexity results from simultaneous consideration of material and geometric nonlinearities, dynamic instability, and kinematic boundary conditions.

Response in Terms of Wave Propagation – An Example

The vibrational representation of the solution of response of a multi-degree-of-freedom system subjected to earthquake shaking is frequently simplified by considering only the fundamental and, occasionally, a few of the lowest frequencies of the system. Doing so is analogous to low-pass filtering of the complete solution [56,57], but it can work well when the excitation amplitudes are small and the motions are associated with long waves. However, during strong earthquakes, the ground motion contains large displacement pulses, the duration of which can be shorter than the fundamental period of the structure. For this type of excitation, the vibrational representation of response and the response spectrum superposition method cease to be suitable and should be replaced by a solution in terms of propagating waves. For short impulsive ground motions, the damage can occur before the wave entering the structure completes its travel up and down the structure, and well before the wave interference can occur—that is, well before the physical conditions can lead to the interference of waves and creation of the mode shapes.

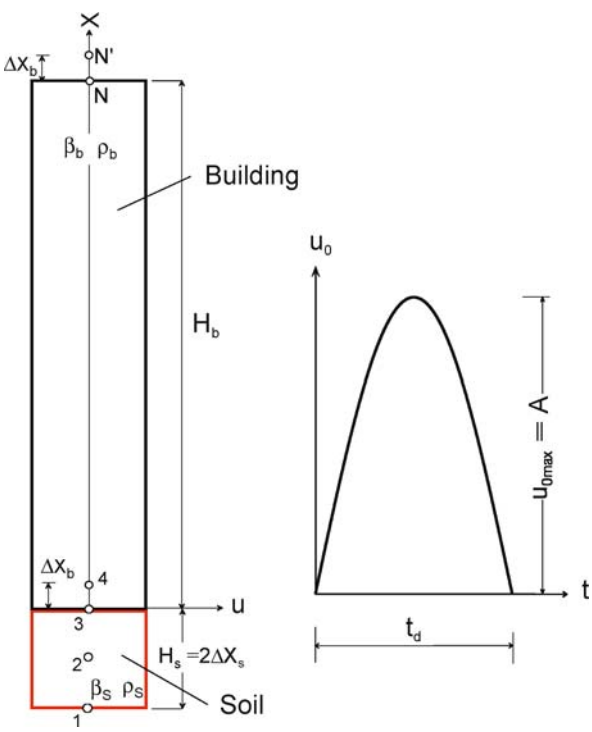

Earthquake Engineering, Non-linear Problems in, Figure 9
Shear beam (building) (*left*) and incoming strong-motion displacement pulse (*right*) in the soil

To illustrate the phenomena that can occur during nonlinear wave propagation in a building, we describe horizontal motions, u, in a one-dimensional shear beam, supported by one-dimensional half space and excited by a vertically propagating shear wave described by a half-sine-pulse (Fig. 9). A finite-difference scheme for solution of this problem with accuracy, $O(\Delta t^2, \Delta x^2)$, where Δx and Δt are the space and time increments, leads to the exact solution for $\beta\Delta t/\Delta x = 1$, where β is the velocity of shear waves. For simplicity, the incident displacement in the soil is chosen to be a sinusoidal pulse with the characteristics shown on Fig. 9.

A mesh with different spatial intervals in the soil and in the building will be used. The equation of motion is

$$v_t = (\sigma)_x/\rho \,, \tag{7a}$$

and the relation between the derivative of the strain and the velocity is

$$\varepsilon_t = v_x \,, \tag{7b}$$

where v, ρ, σ, and ε are particle velocity, density, shear stress, and shear strain, respectively, and the subscripts t and x represent derivatives with regard to time and space.

The domain consists of two materials (Fig. 9): (1) $-2\Delta x_s \leqslant x < 0$ with physical properties ρ_s and μ_s, representing foundation soil, and (2) $0 < x \leqslant H_b$ with physical properties ρ_b and μ_b for linear response, where ρ_i is the density and μ_i is the shear modulus in the soil ($i = s$) or in the building ($i = b$). $v = \partial u/\partial t$ and $\varepsilon = \partial u/\partial x$ are the velocity and the strain of a particle, and u is out-of-plane displacement of a particle perpendicular to the propagation ray.

It is assumed that the incoming wave is known and that its displacement as a function of time is prescribed at the point 1 in the soil ($x = -2\Delta x_s$). Also, it is assumed that the soil is always in the linear elastic state. The finite difference method for a set of simultaneous equations is used to solve the problem, and spatial intervals are defined by $\Delta x_i = \beta_i \cdot \Delta t$, where β_i is the velocity of shear waves in the soil ($i = s$) or in the building ($i = b$) and Δt is the time step. The transparent boundary adopted for this study, which is described in Fujino and Hakuno [8],

is a perfect, transparent boundary for one-dimensional waves when $\beta\Delta x/\Delta t = 1$. Point 1 is where the prescribed displacement is applied, and we assume that this displacement travels upward in each time step. Point 2 is the boundary point of the model, where the quantities of motion are updated in each time step, and point 3 is the first spatial point, where the motion is computed using finite differences.

For the linear case at the contact (see point 3 in Fig. 9), one part of the incoming wave is transmitted into the other medium and one is reflected back into the same medium. The corresponding coefficients are obtained from the boundary conditions of continuity of the displacements and stresses at the contact. For a transmitted wave from medium B to medium A, the transmission coefficient is equal to $k_{\mathrm{tr\,B\to A}} = 2/\left[1 + \rho_a\beta_a/\left(\rho_b\beta_b\right)\right]$. For a reflected wave from medium A back into medium B, this coefficient is $k_{\mathrm{ref\,B\to B}} = \left[1 - \rho_a\beta_a/\left(\rho_b\beta_b\right)\right]/\left[1 + \rho_a\beta_a/\left(\rho_b\beta_b\right)\right]$. For the opposite direction of propagation, the numerators and the denominators in these fractions exchange places.

Numerical Examples

We consider a shear beam supported by elastic soil, as shown in Fig. 9. The densities of the soil and of the beam are assumed to be the same: $\rho_b = \rho_s = \rho = 2000\,\mathrm{kg/m^3}$. The velocity of the shear waves in the soil is taken as $\beta_s = 250\,\mathrm{m/s}$, and in the building as $\beta_b = 100\,\mathrm{m/s}$.

To describe nonlinear response and the development of permanent deformations in the beam, we introduce two dimensionless parameters: (1) dimensionless amplitude $\alpha = A/\left(H_b\varepsilon_{yb}\right)$, where A is the amplitude of the pulse (Fig. 9), H_b is the height of the building, and ε_{yb} is the yielding strain in the building, and (2) dimensionless frequency $\eta = H_b/\left(\beta_b t_d\right)$, where $\beta_b t_d$ is one half of the wavelength of the wave in the building, β_b is the shear-wave velocity in the building, and t_d is the duration of the half-sine pulse.

To understand the development of the permanent strain in the nonlinear beam, we describe first the solution for the linear beam. The displacement and the strain for the linear beam are:

$$\begin{aligned} u(x,t) &= A\sum_{j=1}^{\infty} k_j \left\{ \sin\frac{\pi}{t_d}\left(t - t_{j-1} - \frac{x}{\beta_b}\right)\left[H\left(t - t_{j-1} - \frac{x}{\beta_b}\right) - H\left(t - t_{j-1} - \frac{x}{\beta_b} - t_d\right)\right] + \sin\frac{\pi}{t_d}\left(t - t_j + \frac{x}{\beta_b}\right) \right. \\ &\quad \left. \cdot\left[H\left(t - t_j + \frac{x}{\beta_b}\right) - H\left(t - t_j + \frac{x}{\beta_b} - t_d\right)\right]\right\} \end{aligned} \tag{8}$$

and

$$\begin{aligned} \varepsilon(x,t) &= A\frac{\pi}{\beta_b t_d}\sum_{j=1}^{\infty} k_j \left\{ -\cos\frac{\pi}{t_d}\left(t - t_{j-1} - \frac{x}{\beta_b}\right) \right. \\ &\quad \cdot\left[H\left(t - t_{j-1} - \frac{x}{\beta_b}\right) - H\left(t - t_{j-1} - \frac{x}{\beta_b} - t_d\right)\right] \\ &\quad + \cos\frac{\pi}{t_d}\left(t - t_j + \frac{x}{\beta_b}\right) \\ &\quad \left. \cdot\left[H\left(t - t_j + \frac{x}{\beta_b}\right) - H\left(t - t_j + \frac{x}{\beta_b} - t_d\right)\right]\right\} \end{aligned} \tag{9}$$

where j is the order number of the passage of the wave on the path bottom-top-bottom in the building, $t_j = 2jH_b/\beta_b (j = 0, 1, 2, 3, \ldots)$, is the time required for the wave to pass j times over the path bottom-top-bottom (two heights), $k_j = k_t k_r^{j-1}$ is the amplitude factor of the pulse in the soil in its jth passage along the path bottom-top-bottom through the building, and k_t and k_r are coefficients defined by $k_{\mathrm{tr\,B\to A}}$ and $k_{\mathrm{ref\,B\to B}}$ above.

The odd terms in Eq. (8) and Eq. (9) describe the response to the pulse coming from below, while the even terms describe the response to the pulse arriving from above. For the shear-wave velocities in our example, $k_t = 10/7$ and $k_r = -3/7$. In Eq. (8) the displacement is positive for odd passages and negative for even passages. The displacement and velocity change sign after reflection from the soil-building interface and do not change sign after reflection from the top of the building. The strain changes sign after reflection from the top of the building and does not change sign after reflection from the building-soil interface. The constant that multiplies the series in Eq. (8) in terms of dimensionless amplitude and dimensionless frequency is $A\pi/(\beta_b t_d) = A_\varepsilon = \pi\alpha\eta\varepsilon_{yb}$.

To describe the occurrence of permanent strain, we consider two characteristic points in the building: (1) Point B ($x = 0$) at the soil-building interface (point 3 in the grid, see Fig. 9), and (2) point T ($x = H_b - \beta_b t_b/2$), where the amplitudes of the strain with the same sign meet after reflection from the top of the building. The location of this point is dependent upon the duration (wavelength) of the pulse. The first term in Eq. (8) is one if the argument of the cosine function is equal to $t_d(t - t_0 - x/\beta_b = t_d)$, and the second term is one if the argument of the second cosine function is equal to 0 ($t - t_1 + x/\beta_b = 0$). The position of point T, where the strain amplitude is two times larger than the strain entering the beam, is at $x = H_b - \beta_b t_d/2$, and the time when this occurs is $t = H_b/\beta_b + t_d/2$. From Eq. (9) in the first passage of the pulse, $t < 2H_b/\beta_b$, and only the first term in the series ex-

ists. The strain at point B reaches its absolute maximum at the very beginning, during the entrance of the pulse into the building, and its value is $|\varepsilon^1_{\mathrm{B\,max}}| = \pi\alpha\eta\varepsilon_{\mathrm{yb}}k_{\mathrm{t}}$. If this strain is greater than the yielding strain in the building, $\varepsilon_{\mathrm{yb}}$, a permanent strain at the interface will develop, and the condition for occurrence of permanent strain at this point is $|\varepsilon^1_{\mathrm{B\,max}}| > \varepsilon_{\mathrm{yb}}$, or, in terms of the dimensionless parameters,

$$\alpha\eta > (\pi k_{\mathrm{t}})^{-1} = (\beta_{\mathrm{b}} + \beta_{\mathrm{s}})/(2\pi\beta_{\mathrm{s}}) = C_{\mathrm{B}}\,. \qquad (10\mathrm{B})$$

At point T (this point does not exist if $t_{\mathrm{d}} > 2H_{\mathrm{b}}/\beta_{\mathrm{b}}$, and it coincides with point B if $t_{\mathrm{d}} = 2H_{\mathrm{b}}\beta_{\mathrm{b}}$), from Eq. (9), the maximum strain during the first passage occurs at $t = H_{\mathrm{b}}/\beta_{\mathrm{b}} + t_{\mathrm{d}}/2$, and its amplitude is $2A_{\varepsilon}\cdot k_{\mathrm{t}}$. The condition for occurrence of the permanent strain is

$$\alpha\eta > (2\pi k_{\mathrm{t}})^{-1} = (\beta_{\mathrm{b}}+\beta_{\mathrm{s}})/(4\pi\beta_{\mathrm{s}}) = C_{\mathrm{B}}/2 = C_{\mathrm{T}}. \qquad (10\mathrm{T})$$

For the shear-wave velocities in our example $C_{\mathrm{B}} = 0.2228$ and $C_{\mathrm{T}} = 0.1114$.

For the above simple model, the occurrence, development, and amplitudes of permanent strains and displacements have been studied by Gicev and Trifunac [10,11]. They found that for large ground-displacement pulses (large α) the maximum permanent strains occur mainly at the interface of the building with the soil, while for smaller amplitudes of pulses permanent strains occur closer to the top of the building. They distinguished three zones of the permanently deformed beam: (1) a permanently deformed zone at the bottom; (2) an intermediate zone, which is not deformed at its bottom part and is deformed in the top part; and (3) a non-deformed zone at the top of the beam. The occurrence and development of these zones depends upon the dimensionless excitation amplitudes and the dimensionless frequencies, and in particular on the conditions that lead to the occurrence of the first permanent strain (see Eqs. (10B) and (10T)). For large and long strong-motion pulses ($\eta \leqslant 0.5$; first, the condition in Eq. (10B) is relevant), only zones 1 and 3 are present in the beam. For large amplitudes and short strong-motion pulses, all three zones develop and are present. For smaller excitation amplitudes (when the condition in Eq. (10B) cannot be satisfied for long pulses, and when the condition in Eq. (10T) is satisfied), only zones 2 and 3 exist in the beam. For larger values of η (when the condition in Eq. (10B) is satisfied) all three zones exist.

Gicev and Trifunac [10,11] found a similar situation for the occurrence of the maximum strains. For large and long pulses, maximum strain is located at the bottom of the building, and, as the pulses become shorter, peak strains occur at higher positions in the building. For some high frequencies of excitation, the maximum strain again appears at the bottom of the building because the loss of energy due to the development of the permanent strain at the bottom overcomes the effects of the wave reflections from the top of the building (Fig. 10).

Creation of large permanent deformation zones in the building by the incident waves absorbs some or most of the incident wave energy and can reduce or eliminate further wave propagation and the associated energy transport (Figs. 11 and 12). To the extent that the locations of the plastic deformation zones can be controlled by the design process, absorption of the incident-wave energy by structural members may become a new and powerful tool for performance-based design. To take advantage of such possibilities, the governing differential equations must be solved by the wave-propagation method.

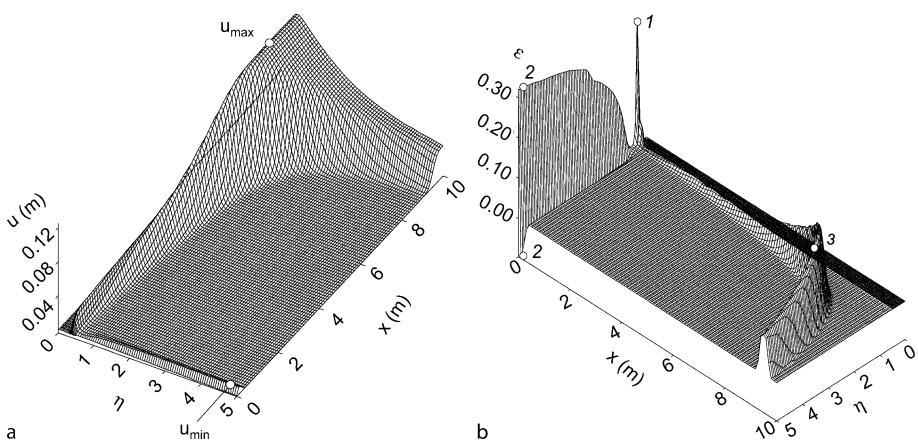

Earthquake Engineering, Non-linear Problems in, Figure 10

Permanent displacements ($u_{\max} = 0.126$ m) (*left*), and permanent strains ($\varepsilon_1 = 0.31$, $\varepsilon_2 = 0.32$, $\varepsilon_3 = 0.20$) (*right*), along the building versus dimensionless frequency η and for dimensionless amplitude $\alpha = 0.3$

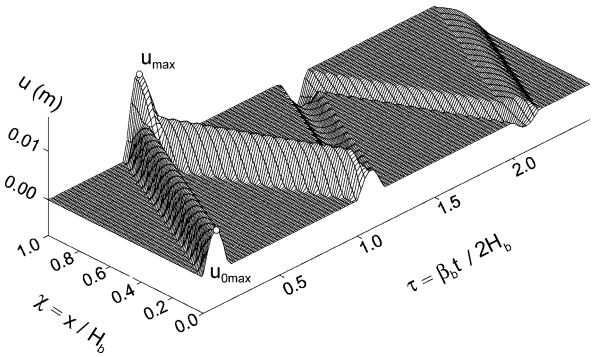

Earthquake Engineering, Non-linear Problems in, Figure 11
Linear displacements along the normalized length of the beam, $\chi = x/H_b$, versus normalized time $\tau = \beta_b t/2H_b$, for dimensionless pulse amplitude $\alpha = 0.03$ and dimensionless frequency $\eta = 3$

Examples illustrated here show that for excitation of structures by large, near-field displacement pulses failure can occur anywhere in the building before the incident wave has completed its first travel from the foundation to the top of the building and back to the foundation ($2H_b/\beta_b$). Because this travel time is shorter (by 1/2) than the natural period of the structure on the fixed base, it is seen that the common response spectrum method of analysis (based on the vibrational formulation of the solution) cannot provide the required details for the design of structures for such excitation. The complexity of the outcome increases with amplitudes of excitation and depends upon the pulse duration. Because actual strong ground motion in the near field has at least several strong pulses, it can be seen that the complexity in real structures responding to strong earthquake motions will be even greater. In engineering approximation based on the vibrational solution of the problem and on the SDOF models, where the location of ductile response is predetermined by the simple modeling assumptions, this complexity cannot be included because of the modeling constraints. The outcome is that it is virtually impossible for simplified models to identify or to predict the location of damage. In contrast, for properly chosen wave propagation models, prediction and identification of damage is a natural and logical outcome of interaction between excitation and model properties. A good example of this can be found in Gicev and Trifunac [12], who showed how a simple wave-propagation model can predict the actually observed location of damage.

Observations of Nonlinear Response

Invaluable for understanding and proper treatment of the actual nonlinear response, and for validation of vibration monitoring and analysis methods for real-life problems, are earthquake response data from well-instrumented, full-scale structures that have been damaged by an earthquake. Such data are rare and are not always freely available. An example of an instrumented building that has been damaged by an earthquake, and for which information about the damage and strong-motion data on the causative earthquake are available, is the former Imperial County Services Building in El Centro, California, which was severely damaged by the magnitude 6.6 Imperial Valley earthquake of October 15, 1979, and later demolished [23,51]. Its transverse (NS) response was recorded by three vertical arrays (recording channels 1, 3, 7, 9, 10, and 11; see Fig. 13), and its longitudinal (EW) response was recorded by one vertical array (recording channels 4, 5, 6, and 13, also shown in Fig. 13).

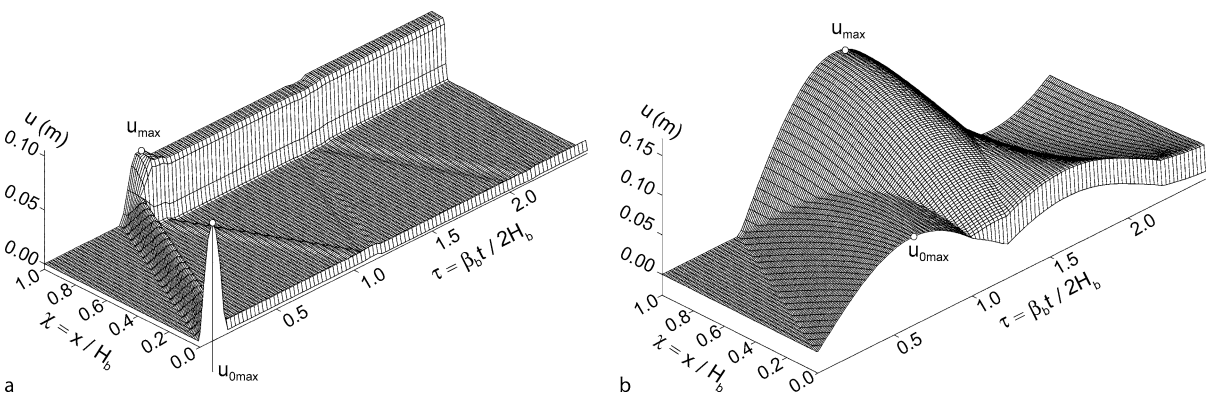

Earthquake Engineering, Non-linear Problems in, Figure 12
Nonlinear displacements along the normalized length of the beam, $\chi = x/H_b$, versus normalized time $\tau = \beta_b t/2H_b$ for dimensionless pulse amplitude $\alpha = 0.3$ and dimensionless frequencies $\eta = 3$ (*left*) and $\eta = 0.41$ (*right*)

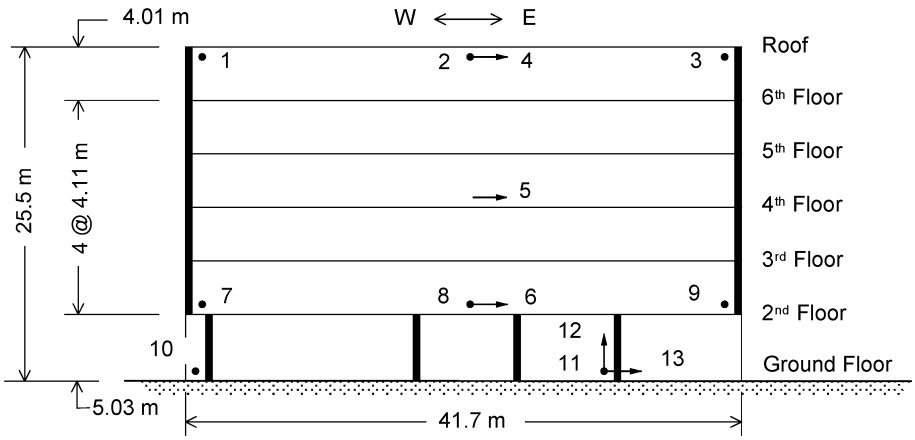

Earthquake Engineering, Non-linear Problems in, Figure 13
Layout of the seismic monitoring array in the ICS building (*dots*, without *arrows*, show the NS recording channels)

For a simplified soil–structure interaction model of a building supported by a rigid foundation, the difference between the roof and base horizontal displacements during earthquake shaking is the sum of the horizontal displacements due to (1) horizontal deformation of the soil, (2) rigid-body rocking of the foundation, and (3) deformation of the structure. The estimated frequency from such data is referred to as system or "apparent" frequency, which differs from the fixed-base frequency of the building. While the fixed-base frequency depends only upon the properties of the structure, the apparent frequency depends also upon the stiffness of the foundation soil. The following relationship holds:

$$\frac{1}{\omega_{\text{sys}}^2} = \frac{1}{\omega_1^2} + \frac{1}{\omega_{\text{H}}^2} + \frac{1}{\omega_{\text{R}}^2}\,, \tag{13}$$

where $\omega_{\text{sys}} = 2\pi\nu_{\text{sys}}$ is the soil–structure system frequency, ω_1 is the fundamental fixed-base frequency of the structure, and ω_{H} and ω_{R} are the horizontal and rocking frequencies, respectively, of a rigid structure on flexible soil [33].

Figure 14c shows that during earthquake shaking (Fig. 14a) the NS frequency of relative system response (Fig. 14b) dropped from $\nu \approx 2.12\,\text{Hz}$ in the early stage of response (at $t \approx 2\,\text{s}$) to $\nu \approx 1.52\,\text{Hz}$ at $t \approx 6.8\,\text{s}$ ($\Delta\nu \approx 0.6\,\text{Hz}$, $\Delta\nu/\nu \approx 28\%$), that it was constant during the interval $t \approx 6.8 - 8.5\,\text{s}$, and that it dropped further to $\nu \approx 0.85\,\text{Hz}$ at $t \approx 12\,\text{s}$ ($\Delta\nu \approx 0.67\,\text{Hz}$, $\Delta\nu/\nu \approx 44\%$). Then, toward the end of the recorded shaking, the frequency increased to $\nu \approx 1.15\,\text{Hz}$ ($\Delta\nu \approx 0.3\,\text{Hz}$; $\Delta\nu/\nu \approx 35\%$). Early in the response ($t < 7\,\text{s}$), the amplitudes of the first story drifts in the building were relatively small (< 0.5%), and the observed decrease of system frequency is believed to be due to changes in the soil and bonding between the soil and foundation. This was followed by a further decrease in the system frequency of

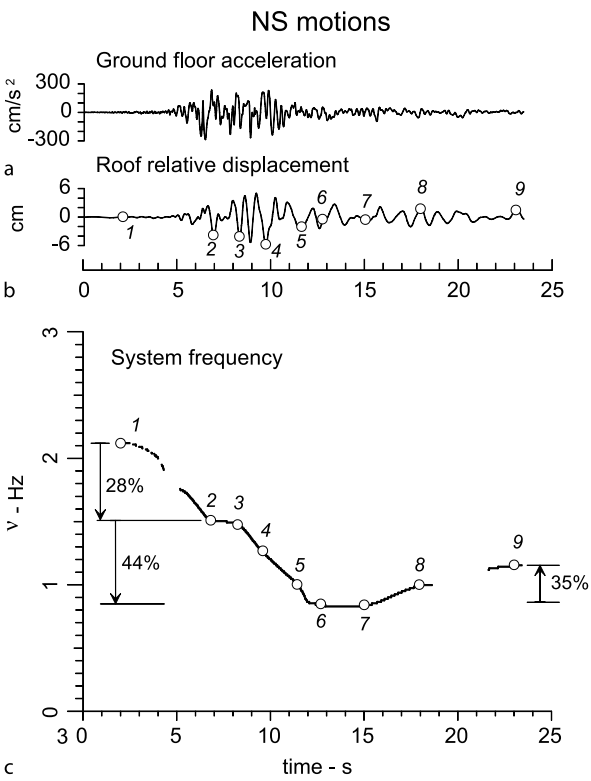

Earthquake Engineering, Non-linear Problems in, Figure 14
Time-frequency analysis for the NS response of the ICS building: a ground acceleration, b relative roof response, and c system frequency versus time

about 44% (between 8 and 12 s). The first-story drifts in the building were large when this occurred (> 0.5% for NS), and the principal cause for this change is believed to be the damage, with the most severe damage occurring between 8 and 12 s after trigger. Near the end of the shaking, a 35% increase in system frequency was observed, suggesting system hardening, which is believed to be due to changes in the soil [51].

Changes similar to what is shown in Fig. 14c were first observed following the San Fernando earthquake in California in 1971 [66] and then during many subsequent earthquakes. It is known at present that many different factors can contribute to fluctuations of the system frequency, including rainfall, temperature fluctuations, changes in occupancy, remodeling and strengthening of buildings, wind, and earthquakes [49]. The simultaneous action of some of these factors and the associated time-dependent changes in the physical model contribute to complex and evolving system changes that make predictions of the dynamic response difficult.

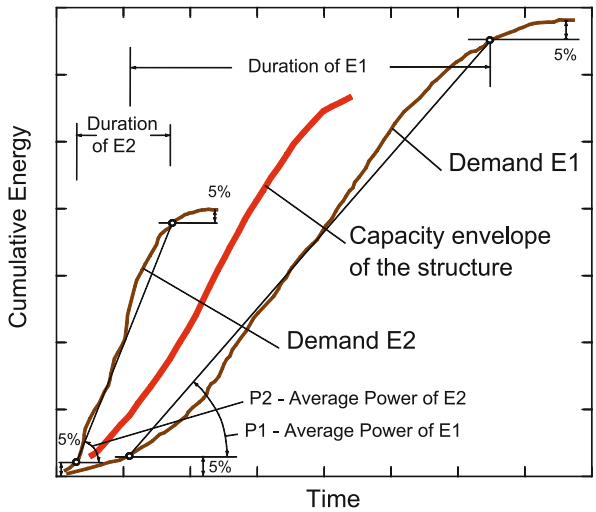

Earthquake Engineering, Non-linear Problems in, Figure 15
Schematic comparison of strong-motion power demands E1 and E2 with an envelope of structural power capacity

Future Directions

Well-designed structures are expected to have ductile behavior during the largest credible shaking, and a large energy reserve to at least delay failure if it cannot be avoided. As the structure finally enters large nonlinear levels of response, it absorbs the excess of the input energy through ductile deformation of its components. Thus, it is logical to formulate future earthquake-resistant design procedures in terms of the energy driving this process. From the mechanics point of view, this introduces nothing new, because the energy equations can be derived directly from the dynamic equilibrium equations. The advantage of using energy is that the duration of strong motion, the number of cycles to failure, and dynamic instability all can be addressed directly and explicitly. This, of course, requires scaling of the earthquake source and of the attenuation of strong motion described in terms of its wave energy. Trifunac et al. [65] reviewed the seismological aspects of empirical scaling of seismic wave energy, E_s, and showed how the radiated energy can be represented by the functionals of strong ground motion [53,54,55]. They described the energy propagation and attenuation with distance and illustrated it for the three-dimensional geological structure of the Los Angeles basin during the 1994 Northridge, CA earthquake, then they described the seismic energy flow through the response of soil-foundation-structure systems, analyzed the energy available to excite the structure, and finally examined the relative response of the structure.

Power Design

Figure 15 illustrates the cumulative wave energies recorded at a building site during two hypothetical earthquakes, E1 an E2, and presents a conceptual framework that can be used for development of the power design method. E1 results in a larger total shaking energy at the site and has a long duration of shaking, leading to relatively small average power, P1. E2 leads to smaller total shaking energy at the site but has short duration and thus greater power, P2. The power capacity of a structure cannot be described by one unique cumulative curve, as this depends upon the time history of shaking. For the purposes of this illustration, the line labeled "capacity envelope of the structure" can be thought of as an envelope of all possible cumulative energy paths for the response of this structure. Figure 15 implies that E1 will not damage this structure, but E2 will. Hence, *for a given structure, it is not the total energy of an earthquake event (and the equivalent energy-compatible relative velocity spectrum) but the rate with which this energy arrives and shakes the structure that is essential for the design of the required power capacity of the structure to withstand this shaking and to control the level of damage.*

Trifunac [57] outlined the elementary aspects of such design based on the power of the incident wave pulses. He showed how this power can be compared with the capacity of the structure to absorb the incident wave energy and described the advantages of using the computed power of incident strong motion for design. Power (amplitude

and duration) of the strong near-field pulses will determine whether the wave entering the structure will continue to propagate through the structure as a linear wave or will begin to create nonlinear zones (at first near the top and/or near the base of the structure; Gicev and Trifunac [10,11,12]). For high-frequency pulses, the nonlinear zone, with permanent strains, can be created before the wave motion reaches the top of the structure—that is, before the interference of waves has even started to occur and lead to formation of mode shapes. Overall duration of strong motion [60] will determine the number of times the structure may be able to complete full cycles of response and the associated number of "minor" excursions into the nonlinear response range when the response is weakly non-linear [13], while the presence of powerful pulses of strong motion will determine the extent to which the one-directional quarter period responses [57] may lead to excessive ductility demand, leading to dynamic instability and failure, precipitated by the gravity loads [20]. All of these possibilities can be examined and quantified deterministically by computation of the associated power capacities and power demands for different scenarios, for given recorded or synthesized strong-motion accelerograms, or probabilistically by using the methods developed for Uniform Hazard Analysis [52].

Bibliography

1. Beltrami E (1987) Mathematics for Dynamic Modeling. Wiley, New York
2. Biot MA (1932) Vibrations of buildings during earthquakes. In: Transient Oscillations in Elastic System. Ph D Thesis No. 259, Chapter II. Aeronautics Department, California Institute of Technology, Pasadena
3. Biot MA (1933) Theory of elastic systems vibrating under transient impulse with an application to earthquake-proof buildings. Proc Natl Acad Sci 19(2):262–268
4. Biot MA (1934) Theory of vibration of buildings during earthquakes. Z Angew Math Mech 14(4):213–223
5. Biot MA (2006) Influence of foundation on motion of blocks. Soil Dyn Earthq Eng 26(6–7):486–490
6. Bycroft GN (1980) Soil-foundation interaction and differential ground motions. Earthq Eng Struct Dyn 8(5):397–404
7. Crutchfield JP (1992) Knowledge and meaning. In: Lam L, Naroditsky V (eds) Modeling Complex Phenomena. Springer, New York, pp 66–101
8. Fujino Y, Hakuno M (1978) Characteristics of elasto-plastic ground motion during an earthquake. Bull Earthq Res Inst Tokyo Univ 53:359–378
9. Gicev V (2005) Investigation of soil-flexible foundation-structure interaction for incident plane SH waves. Ph D Dissertation, Department of Civil Engineering. University Southern California, Los Angeles
10. Gicev V, Tifunac MD (2006) Rotations in the transient response of nonlinear shear beam. Department of Civil Engineering Report, CE 06–02. University Southern California, Los Angeles
11. Gicev V, Tifunac MD (2006) Non-linear earthquake waves in seven-story reinforced concrete hotel. Dept. of Civil Engineering, Report, CE 06–03. University Southern California, Los Angeles
12. Gicev V, Tifunac MD (2007) Permanent deformations and strains in a shear building excited by a strong motion pulse. Soil Dyn Earthq Eng 27(8):774–792
13. Gupta ID, Trifunac MD (1996) Investigation of nonstationarity in stochastic seismic response of structures. Dept. of Civil Eng. Report, CE 96–01. University of Southern California, Los Angeles
14. Gwinn EG, Westervelt RM (1985) Intermittent chaos and low-frequency noise in the driven damped pendulum. Phys Rev Lett 54(15):1613–1616
15. Hackett K, Holmes PJ (1985) Josephson Junction, annulus maps, Birkhoff Attractors, horsehoes and rotation sets. Center for Applied Math Report, Cornell University, Ithaca
16. Holmes PJ (1979) A nonlinear oscillator with a strange attractor. Philos Trans R Soc London A 292:419–448
17. Holmes PJ (1982) The dynamics of repeated impacts with a sinusoidally vibrating table. J Sound Vib 84:173–189
18. Holmes PJ (1985) Dynamics of a nonlinear oscillator with feedback control. J Dyn Syst Meas Control 107:159–165
19. Holmes PJ, Moon FC (1983) Strange Attractors and Chaos in Nonlinear Mechanics. J Appl Mech 50:1021–1032
20. Husid R (1967) Gravity effects on the earthquake response of yielding structures. Ph D Thesis, California Institute of Technology, Pasadena
21. Jalali R, Trifunac MD (2007) Strength-reduction factors for structures subjected to differential near-source ground motion. Indian Soc Earthq Technol J 44(1):285–304
22. Kapitaniak T (1991) Chaotic Oscillations in Mechanical Systems. Manchester University Press, Manchester
23. Kojic S, Trifunac MD, Anderson JC (1984) A post-earthquake response analysis of the Imperial County Services building in El Centro. Report, CE 84–02. University of Southern California, Department of Civil Engineering, Los Angeles
24. Kuhn T (1962) The Structure of Scientific Revolutions. The University of Chicago Press, Chicago
25. Lee VW (1979) Investigation of three-dimensional soil–structure interaction. Department of Civil Engineering Report, CE 79–11. University of Southern California, Los Angeles
26. Lee VW, Trifunac MD (1982) Body wave excitation of embedded hemisphere. ASCE, EMD 108(3):546–563
27. Lee VW, Trifunac MD (1985) Torsional accelerograms. Int J Soil Dyn Earthq Eng 4(3):132–139
28. Lee VW, Trifunac MD (1987) Rocking strong earthquake accelerations. Int J Soil Dyn Earthq Eng 6(2):75–89
29. Levin PW, Koch BP (1981) Chaotic behavior of a parametrically excited damped pendulum. Phys Lett A 86(2):71–74
30. Lichtenberg AJ, Lieberman MA (1983) Regular and Stochastic Motion. Springer, New York
31. Lighthill J (1994) Chaos: A historical perspective. In: Newman WI, Gabrielov A, Turcotte D (eds) Nonlinear Dynamics and Predictability of Geophysical Phenomena. Geophysical Monograph 83, vol 18. IUGG, American Geophysical Union, Washington DC, pp 1–5

32. Lomnitz C, Castanos H (2006) Earthquake hazard in the valley of mexico: entropy, structure, complexity, Chapter 27. In: Teisseyre R, Takeo M, Majewski E (eds) Earthquake Source Asymmetry, Structural Media and Rotation Effects. Springer, Heidelberg
33. Luco JE, Wong HL, Trifunac MD (1986) Soil–structure interaction effects on forced vibration tests. Department of Civil Engineering Report, No. 86–05. University of Southern California, Los Angeles
34. McLaughlin JB (1981) Period-doubling bifurcations and chaotic motion for a parametrically forced pendulum. J Stat Phys 24(2):375–388
35. Miles J (1984) Resonant motion of spherical pendulum. Physica 11D:309–323
36. Miles J (1984) Resonantly forced motion of two quadratically coupled oscillators. Physica 13D:247–260
37. Moon FC (1980) Experiments on chaotic motions of a forced nonlinear oscillator: Strange attractors. ASME J Appl Mech 47:638–644
38. Moon FC (1980) Experimantal models for strange attractor vibration in elastic systems. In: Holmes PJ (ed) New Approaches to Nonlinear Problems in Dynamics. SIAM, Philadelphia, pp 487–495
39. Moon FC, Holmes PJ (1979) A magnetoelastic strange attractor. J Sound Vib 65(2):275–296; A magneto-elastic strange attractor. J Sound Vib 69(2):339
40. Moon FC, Holmes WT (1985) Double Poincare sections of a quasi-periodically forced, chaotic attractor. Phys Lett A 111(4):157–160
41. Moon FC, Shaw SW (1983) Chaotic vibration of beams with nonlinear boundary conditions. J Nonlinear Mech 18:465–477
42. Poddar B, Moon FC, Mukherjee S (1986) Chaotic motion of an elastic-plastic beam. J Appl Mech ASME 55(1):185–189
43. Rasband SN (1990) Chaotic Dynamics of Nonlinear Systems. Wiley, New York
44. Reitherman R (2006) The effects of the 1906 earthquake in california on research and education. Earthq Spectr S2(22):S207–S236
45. Richter PH, Scholtz HJ (1984) Chaos and classical mechanics: The double pendulum. In: Schuster P (ed) Stochastic Phenomena and Chaotic Behavior in Complex Systems. Springer, Berlin, pp 86–97
46. Shaw SW (1985) The dynamics of a harmonically excited system having rigid amplitude constraints, parts 1, 2. J Appl Mech 52(2):453–464
47. Shaw S, Holmes PJ (1983) A periodically forced piecewise linear oscillator. J Sound Vib 90(1):129–155
48. Sorrentino L (2007) The early entrance of dynamics in earthquake engineering: Arturo Danusso's contribution. ISET J 44(1):1–24
49. Todorovska MI, Al Rjoub Y (2006) Effects of rainfall on soil–structure system frequency: Examples based on poroelasticity and comparison with full-scale measurements. Soil Dyn Earthq Eng 26(6–7):708–717
50. Todorovska MI, Trifunac MD (1990) Note on excitation of long structures by ground waves. ASCE, EMD 116(4):952–964 (Errata in 116:1671)
51. Todorovska MI, Trifunac MD (2007) Earthquake Damage Detection in the Imperial County Services Building I: the Data and Time-Frequency Analysis. Soil Dyn Earthq Eng 27(6):564–576
52. Todorovska MI, Gupta ID, Gupta VK, Lee VW, Trifunac MD (1995) Selected topics in probabilistic seismic hazard analysis. Department of Civil Engineering Report, No. CE 95–08. University of Southern California, Los Angeles
53. Trifunac MD (1989) Dependence of Fourier spectrum amplitudes of recorded strong earthquake accelerations on magnitude, local soil conditions and on depth of sediments. Earthq Eng Struct Dyn 18(7):999–1016
54. Trifunac MD (1993) Long-period Fourier amplitude spectra of strong motion acceleration. Soil Dyn Earthq Eng 12(6):363–382
55. Trifunac MD (1994) Q and High-Frequency Strong-Motion Spectra. Soil Dyn Earthq Eng 13(3):149–161
56. Trifunac MD (2003) 70th Anniversary of Biot Spectrum, 23rd Annual ISET Lecture. Indian Soc Earthq Technol 1(40):19–50
57. Trifunac MD (2005) Power design method. Proc. of Earthquake Engineering in the 21st Century to Mark 40th Anniversary of IZIIS-Skopje, 28 Aug–1 Sept. Skopje and Ohrid, Macedonia
58. Trifunac MD (2006) Effects of torsional and rocking excitations on the response of structures, Ch 39. In: Teisseyre R, Takeo M, Majewski E (eds) Earthquake Source Asymmetry, Structural Media, and Rotation Effects. Springer, Heidelberg
59. Trifunac MD, Gicev V (2006) Response Spectra for Differential Motion of Columns, Paper II: Out of-Plane Response. Soil Dyn Earthq Eng 26(12):1149–1160
60. Trifunac MD, Novikova EI (1994) State-of-the-art review on strong motion duration. 10th European Conf on Earthquake Eng, Vienna, 28 Aug – 2 Sep 1994 vol I. AA Balkema, Rotterdam, pp 131–140
61. Trifunac MD, Todorovska MI (1997) Response spectra and differential motion of columns. Earthq Eng Struct Dyn 26(2): 251–268
62. Trifunac MD, Ivanovic SS, Todorovska MI (2001) Apparent periods of a building I: Fourier analysis. J Struct Eng ASCE 127(5):517–526
63. Trifunac MD, Ivanovic SS, Todorovska MI (2001) Apparent periods of a building II: Time-frequency analysis. J Struct Eng ASCE 127(5):527–537
64. Trifunac MD, Hao TY, Todorovska MI (2001) Response of a 14-story reinforced concrete structure to nine earthquakes: 61 years of observation in the hollywood storage building. Department of Civil Engineering Report, CE 01–02. University of Southern California, Los Angeles
65. Trifunac MD, Hao TY, Todorovska MI (2001) On energy flow in earthquake response, Department of Civil Engineering Report, No. CE 01–03. University of Southern California, Los Angeles
66. Udwadia FE, MD Trifunac (1974) Time and amplitude dependent response of structures. Earthq Eng Struct Dyn 2:359–378
67. Ueda Y (1980) Steady motions exhibited by Duffing's equation. In: Holmes PJ (ed) A picture book of regular and chaotic motions. New Approaches to Nonlinear Problems in Dynamics. SIAM, Philadelphia
68. Veletsos AS, Newmark NM (1960) Effect of inelastic behavior on the response of simple systems to earthquake motions. Proc 2nd World Conf on Earthquake Engineering, Jul 1960, vol II. Science Council of Japan, Tokoyo, pp 859–912
69. Veletsos AS, Newmark NM (1964) Response spectra for single-degree-of-freedom elastic and inelastic systems. Report No. RTD-TDR-63-3096, vol III. Air Force Weapons Lab, Albuquerque
70. Wong HL, Trifunac MD (1979) Generation of artificial strong motion accelerograms. Int J Earthq Eng Struct Dyn 7(6): 509–527

Earthquake Forecasting and Verification

James R. Holliday, John B. Rundle, Donald L. Turcotte
University of California, Davis, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Earthquake Forecasting
Forecast Verification
Future Directions
Bibliography

Glossary

Binary forecast A type of forecast where earthquakes are forecast to occur in certain regions and forecast not to occur in other regions.

Continuum forecast A type of forecast where the likelihood of an earthquake throughout an entire region is specified.

Failure to predict Earthquake event that occurs where no earthquake are forecasted to occur.

False alarm Earthquake event that is forecasted to occur at a specific location at a specific time but does not occur.

PDF Probability Density Function – A probability density function is any function $f(x)$ that describes the probability density in terms of the input variable x such that $f(x)$ is greater than or equal to zero for all values of x and the total area of the function is 1.

Definition of the Subject

Forecasts of likely future events are used in almost every field: from forecasting tomorrow's weather to modeling the rise and fall of financial indices to predicting the growth of cancerous cells in human tissue. Generally, these forecasts are created under the belief that having a forecast – regardless of the level of complexity of the underlying models – is more desirable than not having a forecast. That is, human nature prefers the foreseeable over the unexpected. It is therefore important to verify the forecast and measure its skill, or "goodness", and its value, or "usefulness". The process of testing a given forecast with past trend data is the study of forecast verification. Forecast verification allows for a precise and repeatable (as opposed to relative or subjective) judgment of a forecasting model.

Introduction

Earthquakes are the most feared of natural hazards because they generally occur without warning. Hurricanes can be tracked, floods develop gradually, tornados are caused by measurable atmospheric conditions, and volcanic eruptions are preceded by a variety of precursory phenomena. Earthquakes, however, occur suddenly and often without precursory indicators. There have been a wide variety of approaches applied to the forecasting of earthquakes [30,31,37,40,56,65]. These approaches can be divided into two general classes. The first approach is based on empirical observations of precursory changes. Examples include precursory seismic activity, precursory ground motions, and many others. The second approach is based on statistical patterns of seismicity. Neither approach has been able to provide reliable short-term forecasts (days to months) on a consistent basis.

Although short-term predictions are not available (see Table 1), long-term seismic-hazard assessments can be made. It is also possible to assess the long-term probability of having an earthquake of a given magnitude in a given region. These assessments are primarily based on the hypothesis that future earthquakes will occur in regions where past earthquakes have occurred [14,35]. Specifically, the rate of occurrence of small earthquakes in a region can be analyzed to assess the probability of occurrence of much larger earthquakes. While some earthquakes occur in plate interiors – a specific example is the three large (magnitude ~7.7) earthquakes that occurred near New Madrid, Missouri in 1810 and 1811 – the large majority of all earthquakes occur in the vicinity of plate boundaries. A number of large cities are located very close to plate boundaries. Examples include Tokyo, Los Angeles, San Francisco, Seattle, Lima, Jakarta, and Santiago. Much of China is a diffuse plate boundary, and major earthquakes have caused devastating losses of life throughout this region. A recent example was in the 1976 Tangshan earthquake with some 500,000 deaths.

A major goal for earthquake forecasting is to quantify the risk of occurrence of an earthquake of a specified magnitude, in a specified area, and in a specified time window. This is routinely done and results in the creation of hazard maps. Another goal is to specifically forecast or predict earthquakes. The fundamental question is whether forecasts of the time and location of future earthquakes can be accurately made. It is accepted that long term hazard maps of the expected rate of occurrence of earthquakes are rea-

Earthquake Forecasting and Verification, Table 1
Warning times, scientific bases, and scientific feasibility for various types of earthquake predictions and estimates of long-term potential [62]

Term	Warning Time	Scientific Basis	Feasibility
Immediate alert	0 to 20 seconds	Speed of electro -magnetic waves $\gg$ speed of seismic waves	Good
Short-term prediction	Hours to weeks	Accelerating aseismic slip, foreshocks for some events	Unknown
Mid-term prediction	1 month to 10 years	Changes in seismicity, strain, chemistry, and fluid pressure	Fair
Long-term prediction	10 to 30 years	Time remaining in cycle of large shocks, increase in regional shocks	Good
Long-term potential	$>$ 30 years	Long-term rate of activity, plate tectonic setting	Very good

sonably accurate. But is it possible to do better? Are there precursory phenomena that will allow earthquakes to be forecast?

Earthquake Forecasting

Chaos and Forecasting

One of the reasons earthquakes are difficult to accurately forecast is the underlying complexity of the fault system. Earthquakes are caused by displacements on preexisting faults. Most earthquakes occur at or near the boundaries between the near-rigid plates of plate tectonics. Earthquakes in California are associated with the relative motion between the Pacific plate and the North American plate. Much of this motion is taken up by displacements on the San Andreas fault, but deformation and earthquakes extend from the Rocky Mountains on the east into the Pacific Ocean on the west. Clearly this deformation and the associated earthquakes are extremely complex.

It is now generally accepted that earthquakes are examples of deterministic chaos [66]. Some authors [16,17] have argued that this chaotic behavior precludes the prediction of earthquakes. Weather systems, however, are also chaotic, yet short-term forecasts are routinely made. Weather forecasts are probabilistic in the sense that weather cannot be predicted exactly. One such example is the track of a hurricane. Probabilistic forecasts of hurricane tracks are made every year; sometimes they are extremely accurate while at other times they are not. Another example of weather forecasting is the forecast of El Niño events. Forecasting techniques based on pattern recognition and principle components of the sea surface temperature fluctuation time series have been developed that are quite successful in forecasting future El Niños, but again they are probabilistic in nature [11]. It has also been argued [62] that chaotic behavior does not preclude the probabilistic forecasting of future earthquakes. The belief is that the chaos and nonlinearity in earthquakes arise mainly during unstable sliding in large events. Thus, predictions are possible before large earthquakes, but take a finite amount of time for the system to recover after large earthquakes.

Unobservable Dynamics

Another reason earthquakes are difficult to accurately forecast is that the true dynamics driving the system are simply unobservable and unmeasurable. As discussed above, earthquake faults occur in topologically complex, multi-scale networks that are driven to failure by external forces arising from plate tectonic motions [66]. The basic problem is that the details of the true space-time, force-displacement dynamics are in general unobservable, except in a few selected locations such as deep drill holes [52] or in a very crude, time-averaged sense such as the World Stress Map [81]. In order to completely describe the system, the true dynamics would have to be observable for all space and at all times. In fault systems these unobservable dynamics are usually encoded [59] in the time evolution of the Coulomb failure function, $\mathrm{CFF}(x, t)$:

$$\mathrm{CFF}(x, t) = \tau(x, t) - \mu_s \sigma_N(x, t) \,, \tag{1}$$

where $\tau(x, t)$ is the shear stress at point x and at time t, μ_s is the coefficient of static friction, and $\sigma_N(x, t)$ is normal stress at point x and at time t. The space-time patterns associated with the time, location, and magnitude of the earthquakes, however, are observable. This leads to a focus on understanding the observable, multi-scale, apparent dynamics [52] of earthquakes in an attempt to infer the underlying dynamics.

Empirical Approaches

Empirical approaches to earthquake prediction rely on local observations of precursory phenomena in the vicinity of the earthquake to be predicted. It has been suggested that one or more of the following phenomena may indicate a future earthquake [30,31,37,40,56,65]:

1. precursory increase or decrease in seismicity in the vicinity of the origin of a future earthquake rupture,

2. precursory fault slip that leads to surface tilt and/or displacements,
3. electromagnetic signals,
4. chemical emissions, and
5. changes in animal behavior.

Examples of successful near-term predictions of future earthquakes based solely on empirical observations have been rare. A notable exception was the prediction of the $M = 7.3$ Haicheng earthquake in northeast China that occurred on 4 February 1975. This prediction led to the evacuation of the city which undoubtedly saved many lives. The Chinese reported that the successful prediction was based on foreshocks, groundwater anomalies, and animal behavior. Unfortunately, a similar prediction was not made prior to the magnitude $M = 7.8$ Tangshan earthquake that occurred on 28 July 1976 [68]. Official reports placed the death toll in this earthquake at 242,000, although unofficial reports placed it as high as 655,000.

In order to thoroughly test for the occurrence of direct precursors the United States Geological Survey (USGS) initiated the Parkfield (California) Earthquake Prediction Experiment in 1985 [1,30]. Earthquakes on this section of the San Andreas had occurred in 1857, 1881, 1901, 1922, 1934, and 1966. It was expected that the next earthquake in this sequence would occur by the early 1990s, and an extensive range of instrumentation was installed. The next earthquake in the sequence finally occurred on 28 September 2004. No precursory phenomena were observed that were significantly above the background noise level. Although the use of empirical precursors cannot be ruled out, the future of those approaches does not appear to be promising at this time.

Statistical Approaches

A variety of studies have utilized variations in seismicity over relatively large distances to forecast future earthquakes. The distances are large relative to the rupture dimension of the subsequent earthquake. These approaches are based on the concept that the earth's crust is an activated, or driven, thermodynamic system [52]. Among the evidence for this behavior is the continuous level of background seismicity in all seismographic areas. About a million magnitude two earthquakes occur each year on our planet. In southern California about a thousand magnitude two earthquakes occur each year. Except for the aftershocks of large earthquakes, such as the 1992 $M = 7.3$ Landers earthquake, this seismic activity is essentially constant over time. If the level of background seismicity varied systematically with the occurrence of large earthquakes, earthquake forecasting would be relatively easy. This, however, is not the case.

While there is yet no indication of a universal earthquake indicator, there is increasing evidence that there are systematic precursory variations in some aspects of regional seismicity at least some of the time. For example, it has been observed that there is a systematic variation in the number of magnitude $M = 3$ and larger earthquakes prior to at least some magnitude $M = 5$ and larger earthquakes, and a systematic variation in the number of magnitude $M = 5$ and larger earthquakes prior to some magnitude $M = 7$ and larger earthquakes. The spatial regions associated with this phenomena tend to be relatively large, suggesting that an earthquake may resemble a phase change with an increase in the "correlation length" prior to an earthquake [5,26]. A specific example is the sequence of earthquakes that preceded the 1906 San Francisco earthquake [61]. This seismic activation has been quantified as a power law increase in seismicity prior to earthquakes [4,5,6,7,8,9,10,26,34,38,46,55,79]. There have also been reports of anomalous quiescence in the source region prior to a large earthquake, a pattern that is often called a "Mogi Donut" [30,40,76,77]. Unfortunately, these studies have all been performed retrospectively and their successes have depended on knowing the location of the subsequent earthquake.

There are two fundamentally different approaches to assessing the probabilistic risk of earthquake occurrence using statistical methods. The first of these is fault based, where the statistical occurrence of earthquakes is determined for mapped faults. The applicable models are known as renewal models and a tectonic loading of faults is included. The second approach is seismicity based, where the risk of future earthquakes is based on the past seismicity in the region. These are also known as cluster models and include the epidemic type aftershock sequence (ETAS) model and the branching aftershock sequence (BASS) model.

Fault Based Models Fault based models consider the earthquakes that occur on recognized (i. e., previously known) active faults. These models are also known as renewal models. Renewal models assume that the stress on an individual fault is "renewed" by the tectonic drive of plate tectonics. The simplest renewal model would be that of a single planar strike-slip fault subjected to a uniform rate of strain accumulation (plate motion). In this case, "characteristic" earthquakes would occur periodically. Clearly the earth's crust is much more complex with faults present at all scales and orientations. This complexity leads to chaotic behavior and statistical variability.

An important question is whether the concept of quasi-periodic "characteristic" earthquakes is applicable to tectonically active areas. There is extensive evidence that characteristic earthquakes do occur quasi-periodically on major faults. Many studies have been carried out to quantify the recurrence time statistics of these characteristic earthquakes [43,45,67]. Recurrence time statistics can be characterized by a mean value, μ, and a coefficient of variation, C_v. The coefficient of variation is the ratio of the standard deviation to the mean. Mathematically, $C_v = 0$ for periodic characteristic earthquakes and $C_v = 1$ for a random distribution of recurrence times. Ellsworth et al. [13] reviewed many examples of recurrence time statistics and concluded that $C_v \approx 0.5$ for characteristic earthquakes. Many probability distribution functions have been proposed for recurrence times, including the Weibull, lognormal, Brownian passage time, and gamma distributions.

Two major renewal simulation models have been developed. The first is "Virtual California" [49,50,53]. This is a geometrically realistic numerical simulation of earthquakes occurring on the San Andreas fault system and includes all major strike-slip faults in California. The second model is the "Standard Physical Earth Model" (SPEM) developed by Ward [69] and applied to characteristic earthquakes associated with subduction at the Middle American trench. This model was further developed and applied to the entire San Andreas fault system by Goes and Ward [18], to the San Andreas system in southern California by Ward [70], and to the San Andreas system in northern California by Ward [71].

Both simulation models utilize backslip, with the accumulation of a slip deficit on each fault segment prescribed using available data. The backslip represents the tectonic drive. Both models "tune" the prescribed static friction to give recurrence times that are consistent with available data. In both models fault segments are treated as dislocations when characteristic earthquakes occur, and all fault segments interact with each other elastically utilizing dislocation theory. These chaotic interactions result in statistical distributions of recurrence times on each fault. The resulting coefficients of variation are measures of this interaction.

Yakovlev et al. [78] utilized the Virtual California model to test alternative distributions of recurrence times. They concluded that the Weibull distribution is preferable and based its use on its scale invariance. The hazard rate is the probability that a characteristic earthquake will occur at a given time after the last characteristic earthquake. The Weibull distribution is the only distribution that has a power-law (scale-invariant) hazard function. In the same study, Yakovlev et al. [78] found that the coefficient of variation of the recurrence times of 4606 simulated great earthquakes on the northern San Andreas fault is $C_v = 0.528$. Goes and Ward [18] using the SPEM simulator found that $C_v = 0.50 - 0.55$ on this fault. The two simulations are quite different, so the statistical variability appears to be a robust feature of characteristic earthquakes. A similar simulation model for New Zealand has been given by Robinson and Benites [47,48].

Renewal models have also formed the basis for three formal assessments of future earthquake probabilities in California. These assessments were carried out by the United States Geological Survey [72,73,74,75]. A major problem with renewal models is that large earthquakes in nature often occur on faults that were not previously recognized. Recent examples in California include the 1952 Kern County earthquake, the 1971 San Fernando Valley earthquake, the 1992 Landers earthquake, the 1994 Northridge earthquake, and the 1999 Hector Mine earthquake. At the times when these earthquakes occurred, the associated faults were either not mapped or were considered too small to have such large earthquakes. To compensate for this problem, renewal models often include a random level of background seismicity unrelated to recognized faults.

Seismicity Based Models An alternative approach to probabilistic seismic hazard assessment and earthquake forecasting is to use observed seismicity. The universal applicability of Gutenberg–Richter frequency-magnitude scaling allows the rate of occurrence of small earthquakes to be extrapolated to estimate the rate of occurrence and location of large earthquakes. This type of extrapolation played an important role in creating the national seismic hazard map for the United States [15].

A more formalistic application of this extrapolation methodology is known as a relative intensity (RI) forecast. This type of forecast was made on a world wide basis by Kossobokov et al. [35] and to California by Holliday et al. [23]. A related forecasting methodology is the pattern informatics (PI) method [22,24,25,51,63,64]. This method was used by Rundle et al. [51] to forecast $m = 5$ and larger earthquakes in California for the time period 2000–2010. This forecast successfully predicted the locations of 16 of the 18 large earthquakes that have subsequently occurred.

Keilis-Borok [31,33] and colleagues utilized patterns of seismicity to make formal intermediate term earthquake predictions. The most widely used algorithm, M8, has been moderately successful in predicting the times and locations of large earthquakes. More recently, this group has used chains of premonitory earthquakes to make interme-

diate term predictions [32,58]. Again, moderate success was achieved.

It has also been proposed that there is an increase in the number of intermediate sized earthquakes prior to a large earthquake [26]. This phenomenon is known as accelerating moment release (AMR) and is due primarily to an increase in the number of intermediate-size events that occur within a characteristic distance of the main shock and that scale with magnitude. AMR is characterized by a decrease in the rate of regional seismicity followed by a rapid rebound back to historic levels. Sammis and Bowman [54] have proposed a number of physical models to explain AMR. These include:

1. an analogy with critical phase transitions where the correlation length of the stress field rapidly increases as the system nears the critical point,
2. an erosion of a stress shadow from some previous, large event, and
3. a slow, silent earthquake propagating upward on a ductile extension loading the seismogenic crust above.

The existence of such a seismicity pattern does, however, appear to require a certain regional fault system structure and density. Simulation models using a hierarchical distribution of fault sizes match this pattern well, but other types of fault distributions may also support AMR [26]. Conversely, some real-world fault distributions may not support AMR as a predictive tool. The AMR approach has shown considerable success retrospectively [5,10,55] but has not evolved into a successful prediction algorithm as of yet.

Seismicity based models are often referred to as clustering models. That is, clusters of small earthquakes indicate the future occurrence of larger earthquakes. The RI, PI, and AMR models clearly belong to this class. Other approaches in this class are the epidemic type aftershock sequence (ETAS) model [21,29,42,44] and the branching aftershock sequence (BASS) model. These are statistical models based on applicable scaling laws: Gutenberg–Richter scaling relation and the modified Båth's law for the scaling relation of magnitudes, Omori's law for the distribution of earthquake times, and a modified form of Omori's law for the distribution of earthquake locations. Clustering by definition is not a random process. A rational for the application of clustering models is that the clustering is related to families of foreshocks, main shocks, and aftershocks.

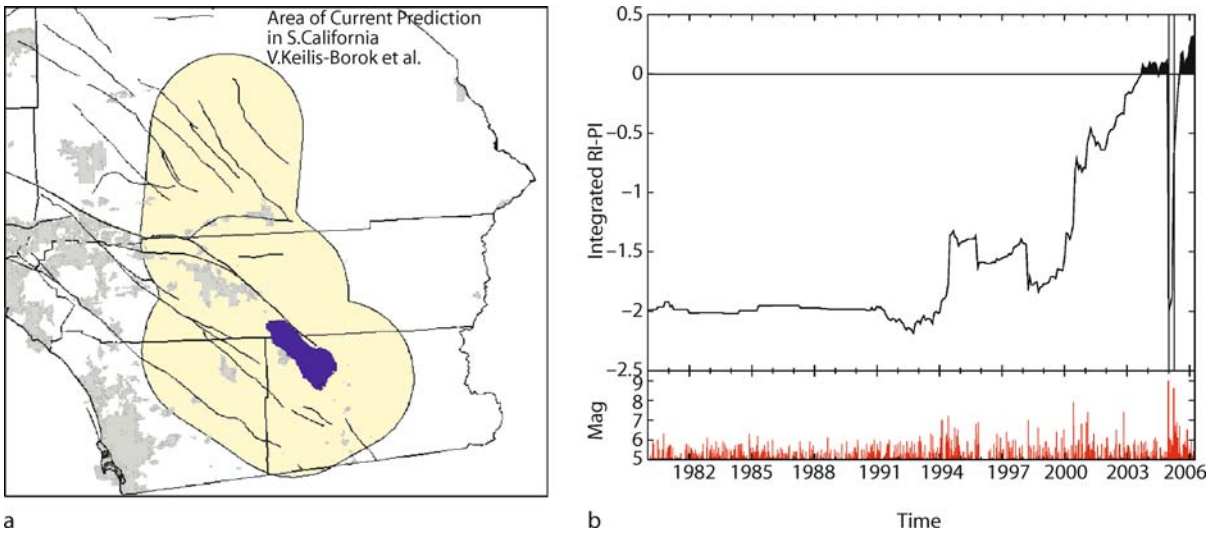

Earthquake Forecasting and Verification, Figure 1

Plots of sample binary earthquake forecasts. a Prediction for a magnitude 6.4 or greater earthquake to occur between 5 January 2004 and 4 September 2004, within a 12,440 sq. miles area of southern California using the M8 algorithm (image courtesy Dr. Vladimir Keilis-Borok). This is a binary forecast since it forecasts an earthquake to occur within the *shaded* area during the time period and not to occur in the *non-shaded* region. Ultimately no magnitude 6.4 or greater earthquake occurred in the test region during the forecast interval. b Retrospective prediction for a magnitude 8.0 or greater earthquake to occur in the Sumatra region between 1 June 2003 and 1 June 2005 (image courtesy Dr. James Holliday). This is a time-dependent binary forecast since large earthquakes are forecasted to occur in the test region within a two year window once the time series becomes positive. Note that the magnitude 9.0 Sumatra-Andaman earthquake occurred 18 months after the time series became positive

Types of Forecasts

Binary Forecasts The simplest type of earthquake forecast is a binary forecast. An earthquake is forecast to occur in a certain regions and forecast not to occur in other regions. This is analogous to the issuance of tornado warnings. Examples of two binary forecast maps are presented in Fig. 1. The plot on the left is for a prediction for a magnitude 6.4 or greater earthquake to occur between 5 January 2004 and 4 September 2004, within a 12,440 sq. miles area of southern California using the M8 algorithm. This map is a binary forecast since it forecasts an earthquake to occur within the shaded area during the time period and not to occur in the non-shaded region. The plot on the right is a retrospective prediction for a magnitude 8.0 or greater earthquake to occur in the Sumatra region between 1 June 2003 and 1 June 2005. This is a time-dependent binary forecast since large earthquakes are forecasted to occur in the test region within a two year window once the time series becomes positive and not to occur when the time series is negative.

Continuum Forecasts The alternative to binary forecasts is a continuum forecast. The likelihood of an earthquake throughout the entire region is specified. This would be analogous to temperature forecasts in the atmospheric sciences. Examples of two continuum forecast maps are presented in Fig. 2. The plot on the left is a time-dependent map produced in real time by the USGS Earthquake Hazards Program giving the probability of strong shaking at any location in California within a given 24-hour period. The plot on the right is a forecast map giving the probability for large (magnitude greater than 5.0) earthquakes in southern California using the Pattern Informatics method.

Any continuum forecast can be converted into a binary forecast through the use of a hard threshold. Spatial regions where the likelihood value is greater than the threshold are taken to be regions where earthquakes are forecasted to occur. Spatial regions where the likelihood value is less than the threshold value are taken to be regions where earthquakes are forecasted not to occur.

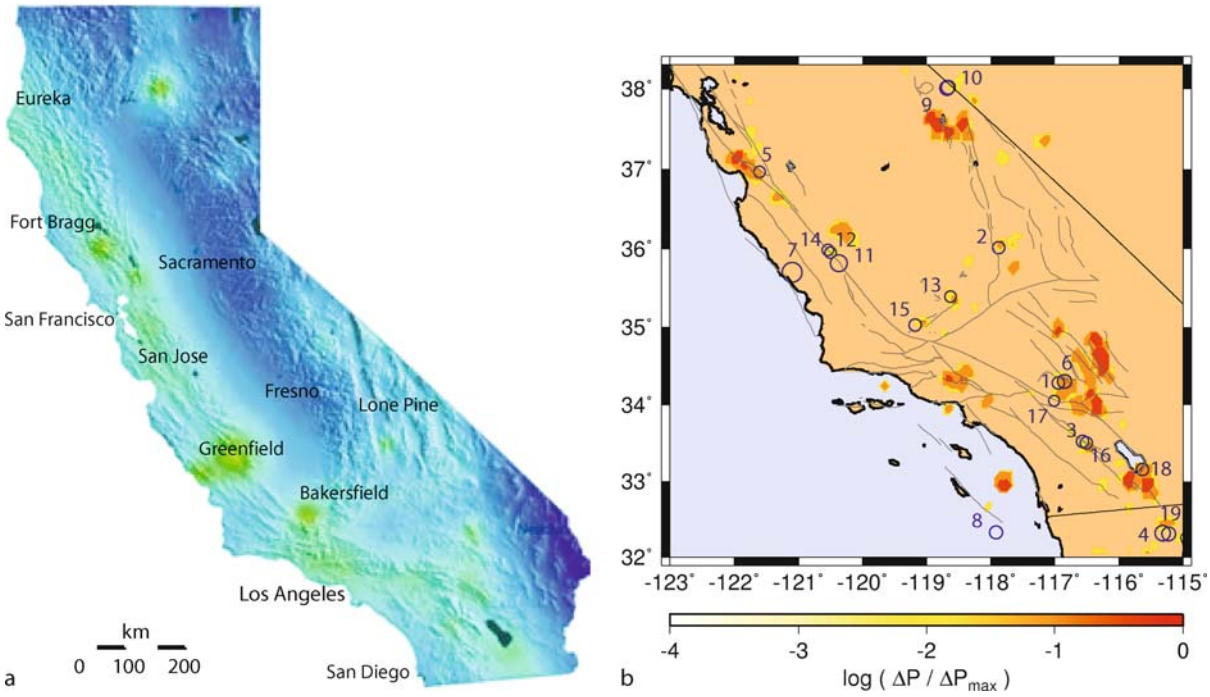

Earthquake Forecasting and Verification, Figure 2
Plots of sample continuum earthquake forecasts. a Time-dependent map giving the probability of strong shaking at any location in California within a given 24-hour period (image courtesy USGS Earthquake Hazards Program). b Forecast map giving the probability for large ($m > 5$) earthquakes in southern California using the Pattern Informatics method (image courtesy Dr. Kristy Tiampo). Circles mark the locations of large earthquakes which occurred after the forecast creation. Both of these are continuum forecasts since they present a continuous likelihood for earthquakes to occur throughout the entire test region

Forecast Verification

Continuum Forecasts

Likelihood Tests The standard approach for testing the hypothesis that a probability measure can forecast future earthquakes is the maximum likelihood test [2,19,24,28,51,57,63]. The likelihood $\mathbf{L}$ is a probability measure that can be used to assess the quality of one forecast measure over another. Typically, one computes the log-likelihood $\mathcal{L} \equiv \log(\mathbf{L})$ for the proposed forecast measure $\mathbf{L}$. Models with higher (less negative) log-likelihood values are said to perform better than models with lower (more negative) log-likelihood values. In these types of likelihood tests, a probability density function (PDF) is required. Two different PDFs are commonly used: a global, Gaussian model and a local, Poissonian model.

Tiampo et al. [63] calculated likelihood values by defining $P[\mathbf{x}]$ to be the union of a set of N Gaussian density functions $p_G(|\mathbf{x}-\mathbf{x}_i|)$ [2] centered at each location $\mathbf{x}_i$. Each individual Gaussian density has a standard deviation σ equal to the width of their coarse-grained lattice cell and a peak value equal to the calculated probability divided by σ^2. $P[\mathbf{x}(e_j)]$ was then interpreted as a probability measure that a future large event e_j would occur at location $\mathbf{x}(e_j)$:

$$P[\mathbf{x}(e_j)] = \sum_i \frac{P_i}{\sigma^2} \mathrm{e}^{-\frac{|\mathbf{x}(e_j)-\mathbf{x}_i|^2}{\sigma^2}} \,. \tag{2}$$

If there are J future events, the normalized likelihood $\mathbf{L}$ that all J events are forecast is

$$\mathbf{L} = \prod_j \frac{P[\mathbf{x}(e_j)]}{\sum_i P[\mathbf{x}_i]} \,. \tag{3}$$

Furthermore, the log-likelihood value $\mathcal{L}$ for a given calculation can be calculated and used in ratio comparison tests:

$$\mathcal{L} = \sum_j \log \frac{P[\mathbf{x}(e_j)]}{\sum_i P[\mathbf{x}_i]} \,. \tag{4}$$

The second model commonly used is based on work performed by the Regional Earthquake Likelihood Models (RELM) group [57]. For each coarse-grained lattice cell i an expectation value λ_i is calculated by scaling the local probability value P_i by the number of earthquakes that occurred over all space during the forecasted time period:

$$\lambda_i = n \cdot P_i \,, \tag{5}$$

where n is the number of future events. Note that for any future time interval (t_2, t_3), n could in principle be estimated by using the Gutenberg–Richter relation. For each bin an observation value ω_i is also calculated such that ω_i contains the number of future earthquakes that actually occurred in cell i. Note that $\sum_i \omega_i = n$. For the RELM model, it is assumed that earthquakes are independent of each other. Thus, the probability of observing ω_i events in cell i with expectation λ_i is the Poissonian probability

$$p_i(\omega_i|\lambda_i) = \frac{\lambda_i^{\omega_i}}{\omega_i!} \mathrm{e}^{-\lambda_i} \,. \tag{6}$$

The log-likelihood $\mathcal{L}$ for observing ω earthquakes at a given expectation λ is defined as the logarithm of the probability $p(\omega|\lambda)$, thus

$$\mathcal{L}(\omega|\lambda) = \log p(\omega|\lambda) = -\lambda + \omega \log \lambda - \log(\omega!) \,. \tag{7}$$

Since the joint probability is the product of the individual cell probabilities, the log-likelihood value for a given calculation is the sum of $\mathcal{L}(\omega_i|\lambda_i)$ over all cells i:

$$\mathcal{L} = \sum_i \mathcal{L}(\omega_i|\lambda_i) = \sum_i (-\lambda_i + \omega_i \log \lambda_i - \log(\omega_i!)) \,. \tag{8}$$

Most tests of earthquake forecasts have emphasized the likelihood test [24,28,51,64]. These tests have the significant disadvantage that they are overly sensitive to the least probable events. For example, consider two forecasts. The first perfectly forecasts 99 out of 100 events but assigns zero probability to the last event. The second assigns zero probability to all 100 events. Under a log-likelihood test, both forecasts will have the same skill score of $-\infty$. Furthermore, a naive forecast that assigns uniform probability to all possible sites will always score higher than a forecast that misses only a single event but is otherwise superior. For this reason, likelihood tests are more subject to unconscious bias. Other methods of evaluating earthquake forecasts are suggested by Vere-Jones [20] and Holliday et al. [25].

Information Metrics One such alternative is the use of information metrics. Using methods from information theory [12], it is possible to calculate the entropy, H, of a forecast map. Entropy can be considered a measure of disorder (e. g. randomness) or "surprise", hence maps with lower entropy contain more useful information than maps with higher entropy. We define entropy as

$$H(z) = -\sum_{i=1}^{N} p(\mathbf{x}_i; z) \log p(\mathbf{x}_i; z) \,, \tag{9}$$

where

$$p(\mathbf{x}_i; z) = \begin{cases} P(\mathbf{x}_i) & P(\mathbf{x}_i) \geq z \\ 0 & P(\mathbf{x}_i) < z \end{cases} \,, \tag{10}$$

and the probabilities are scaled such that $\sum_{i=1}^{N} p(\mathbf{x}_i) = 1$. This definitions allows a measurement of entropy as a function of some lower threshold z. A small, non-zero value for z allows for the measurement of the entropy above and relative to background noise.

Binary Forecasts

ROC Analysis The standard approach to the evaluation of a binary forecast is the use of a relative operating characteristic (ROC) diagram [39,60]. This method evaluates the performance of the forecast method relative to random guessing by constructing a plot of the fraction of failures to predict (events that occur where no event is forecast) against the fraction of false alarms (events that are forecasted to occur at a location but do not occur) for an ensemble of forecasts. Molchan [41] has used a modification of this method to evaluate the success of intermediate term earthquake forecasts.

The binary approach has a long history, over 100 years, in the verification of tornado forecasts [39]. These forecasts take the form of a tornado forecast for a specific location and time interval, each forecast having a binary set of possible outcomes. For example, during a given time window of several hours duration, a forecast is issued in which a list of counties is given with a statement that one or more tornadoes will or will not occur. A 2×2 contingency table is then constructed, the top row contains the counties in which tornadoes are forecast to occur and the bottom row contains counties in which tornadoes are forecast to not occur. Similarly, the left column represents counties in which tornadoes were actually observed, and the right column represents counties in which no tornadoes were observed.

With respect to earthquakes, binary forecasts take exactly this form. A time window is proposed during which the forecast of large earthquakes having a magnitude above some minimum threshold is considered valid. An example might be a forecast of earthquakes larger than $M = 5$ during a period of five or ten years duration. A map of the seismically active region is then completely covered ("tiled") with boxes of two types: boxes in which the epicenters of at least one large earthquake are forecast to occur and boxes in which large earthquakes are forecast to not occur. In other types of forecasts, large earthquakes are given some continuous probability of occurrence from 0% to 100% in each box [28]. These forecasts can be converted to the binary type by the application of a threshold value. Boxes having a probability below the threshold are assigned a forecast rating of non-occurrence during the time window, while boxes having a probability above the threshold are assigned a forecast rating of occurrence. A high threshold value may lead to many failures to predict, but few false alarms. The level at which the threshold is set is then a matter of public policy specified by emergency planners, representing a balance between the prevalence of failures to predict and false alarms.

Earthquake Forecasting and Verification, Table 2
Schematic contingency table for categorical forecasts of a binary event

Forecast	Observed		
	Yes	No	Total
Yes	a	b	$a+b$
No	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d=N$

Contingency Tables An extensive review on forecast verification in the atmospheric sciences has been given by Jolliffe and Stephenson [27]. The wide variety of approaches that they consider are directly applicable to earthquake forecasts as well. Verification of earthquake forecasts proceeds in exactly the same was as for, say, tornado forecasts when using these approaches. For a given forecast, the contingency table (see Table 2) is constructed. Values for the table elements a (Forecast=yes, Observed=yes), b (Forecast=yes, Observed=no), c (Forecast=no, Observed=yes), and d (Forecast=no, Observed=no) are obtained from the forecast map. The fraction of alarm space, also called the probability of forecast of occurrence, is $r = (a+b)/N$, where the total number of boxes is $N = a+b+c+d$. The hit rate is $H = a/(a+c)$ and the false alarm rate is $F = b/(b+d)$. From these quantities a number of descriptive, performance, and skill measures can be constructed [27]. Table 3 lists a few possible measures.

ROC Curves The standard ROC diagram [23,27] is a plot of the points $\{H, F\}$ calculated for a binary forecast (see Fig. 3). If the forecast was converted from continuum map, H and F are plotted as the lower (conversion) threshold is varied. A perfect forecast of occurrence (perfect order, no fluctuations) would consist of two line segments, the first connecting the points $(H, F) = (0, 0)$ to $(H, F) = (1, 0)$, and the second connecting $(H, F) = (1, 0)$ to $(H, F) = (1, 1)$. A curve of this type can be described as maximum possible hits ($H = 1$) with minimum possible false alarms ($F = 0$). Another type of perfect forecast consists of two lines connecting the points (0,0) to (0,1) and (0,1) to (1,1), a perfect forecast of non-occurrence.

Earthquake Forecasting and Verification, Table 3

Table of various descriptive and performance measures that can be calculated from the binary contingency table for an earthquake forecast

Name	Definition	Definition (H, F, and τ)	Range
Fraction of alarm space τ	$\tau = (a + c)/N$	τ	[0, 1]
Hit rate H	$H = a/(a + c)$	H	[0, 1]
False alarm rate F	$F = b/(b + d)$	F	[0, 1]
False alarm ratio FAR	$FAR = b/(a + b)$	$FAR = (1 + \frac{\tau}{1-\tau}\frac{H}{F})^{-1}$	[0, 1]
Miss rate ν	$\nu = c/(a + c)$	$\nu = 1 - H$	[0, 1]
Peirce's skill score PSS	$PSS = \frac{ad-bc}{(b+d)(a+c)}$	$PSS = H - F$	[−1, 1]
Yule's Q	$Q = \frac{ad-bc}{ad+bc}$	$Q = \frac{H-F}{H(1-f)+F(1-H)}$	[−1, 1]
Peirce Area A	$A = \int PSS$	$A = \int HdF - \frac{1}{2}$	$[-\frac{1}{2}, \frac{1}{2}]$

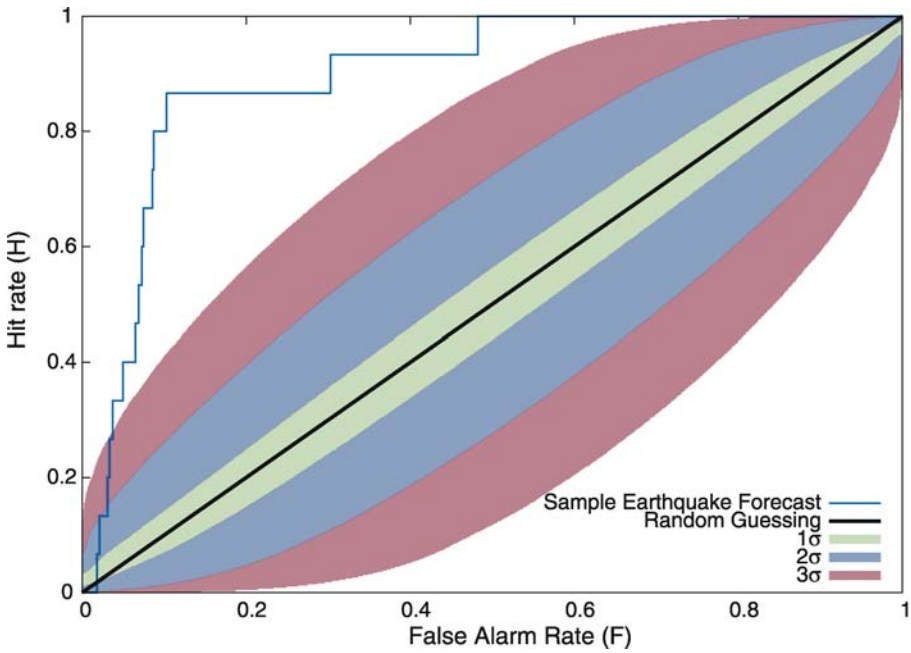

Earthquake Forecasting and Verification, Figure 3

Sample relative operating characteristic (ROC) diagram. Shown is a plot of hit rates, *H*, as a function of false alarm rates, *F*, for a sample earthquake forecast (*blue*) and random guessing (*black*). Confidence intervals for the one-, two- and three-σ levels are shown as well [23,80]

The line $H = F$ occupies a special status, and corresponds to a completely random forecast [23,27] (maximum disorder, maximum fluctuations) where the false alarm rate is the same as the hit rate and no information is produced by the forecast. Points above this line are said to have performed better than simple random guessing. If competing forecasts are plotted on the same graph, the forecast whose H–F curves lies the highest is said to outperform the others. Often, however, competing forecasts will have intersecting curves. In this case, forecasts are said outperform each other only in specific ranges and only for specific choices of the lower (conversion) threshold value.

ν–τ Curves An alternative diagram [41] is a plot of the points $\{\nu, \tau\}$ for a binary forecast map. In this case a perfect forecast of occurrence would consist of two line segments, the first connecting the points $(\nu, \tau) = (0, 0)$ to $(\nu, \tau) = (0, 1)$, and the second connecting $(\nu, \tau) = (0, 1)$ to $(\nu, \tau) = (1, 1)$. A curve of this type can be described as

minimum possible missed events ($\nu = 0$) with minimum possible alarm space ($\tau = 0$).

As with the H–F curve, the line $\nu = \tau$ corresponds to a completely random forecast. Points below this line are said to have performed better than simple random guessing. If competing forecasts are plotted on the same graph, the forecast whose ν–τ curves lies the lowest is said to outperform the others. As can be verified from Table 3, ν–τ curves offer the same information as H–F curves and are identical in the range $a \ll d$.

Future Directions

It is actually quite surprising that immediate local precursory phenomena are not seen. Prior to a volcanic eruption, increases in regional seismicity and surface movements are generally observed. For a fault system, the stress gradually increases until it reaches the frictional strength of the fault and a rupture is initiated. It is certainly reasonable to hypothesize that the stress increase would cause increases in background seismicity and aseismic slip. In order to test this hypothesis the Parkfield Earthquake Prediction Experiment was initiated in 1985. The expected Parkfield earthquake occurred beneath the heavily instrumented region on 28 September 2004. No local precursory changes were observed [3,36]. In the absence of local precursory signals, the next question is whether broader anomalies develop, and in particular whether there is anomalous seismic activity.

Assuming precursors do exist and can be exploited to create earthquake forecasts, testing and verification of the usefulness of the forecasts is the necessary next step. A forecast method that predicts all earthquakes but carries a high false alarm rate is likely to be useless as a public warning tool. Similarly, a forecast method that issues warnings for only a small fraction of actual earthquakes but never issues false alarms is likely to be a poor tool for catastrophe preparation. Forecast verification techniques can be used to find the middle ground that is most useful.

As a final warning, researchers must be careful not to create artificial skill in their forecasts. Since nature provides us only with one earthquake record (the actual history of a given test region), the quality of a new forecast system is often assessed on the same data set used to create it. This can potentially lead to an optimistic bias in any skill scores. This is a particular problem if the score itself is used to calibrate the method, either directly or indirectly. Development of realistic earthquake simulators and cross-testing against test regions with similar fault structures can help protect against this.

Bibliography

Primary Literature

1. Bakun WH, Lindh AG (1985) The Parkfield, California, earthquake prediction experiment. Science 229:619–624
2. Bevington PR, Robinson DK (1992) Data Reduction and Error Analysis for the Physical Sciences. McGraw-Hill, New York
3. Borcherdt RD, Johnston MJS, Glassmoyer G, Dietel C (2006) Recordings of the 2004 parkfield earthquake on the general earthquake observation system array: Implications for earthquake precursors, fault rupture, and coseismic strain changes. Bull Seismol Soc Am 96(4b):73–89
4. Bowman DD, King GCP (2001) Accelerating seismicity and stress accumulation before large earthquakes. Geophys Res Lett 28:4039–4042
5. Bowman DD, Ouillon G, Sammis CG, Sornette A, Sornette D (1998) An observational test of the critical earthquake concept. J Geophys Res 103:24359–24372
6. Bowman DD, Sammis CG (2004) Intermittent criticality and the Gutenberg–Richter distribution. Pure Appl Geophys 161:1945–1956
7. Brehm DJ, Braile LW (1998) Intermediate-term earthquake prediction using precursory events in the New Madrid seismic zone. Bull Seismol Soc Am 88:564–580
8. Brehm DJ, Braile LW (1999) Intermediate-term earthquake prediction using the modified time-to-failure method in southern California. Bull Seismol Soc Am 89:275–293
9. Buffe CG, Nishenko SP, Varnes DJ (1994) Seismicity trends and potential for large earthquakes in the Alaska–Aleutian region. Pure Appl Geophys 142:83–99
10. Buffe CG, Varnes DJ (1993) Predictive modeling of the seismic cycle of the greater San Francisco Bay region. J Geophys Res 98:9871–9883
11. Chen D, Cane MA, Kaplan A, Zebian SE, Huang D (2004) Predictability of El Niño in the past 148 years. Nature 428:733–736
12. Cover TM, Thomas JA (1991) Elements of Information Theory. Wiley-Interscience, New York
13. Ellsworth WL, Mathews MV, Nadeau RM, Nishenko SP, Reasenberg PA, Simpson RW (1999) A physically-based earthquake recurrence model for estimation of long-term earthquake probabilities. Open-File Report 99-522, US Geological Survey
14. Frankel AF (1995) Mapping seismic hazard in the central and eastern United States. Seismol Res Let 60:8–21
15. Frankel AF, Mueller C, Barnhard T, Perkins D, Leyendecker EV, Dickman N, Hanson S, Hopper M (1996) National seismic hazard maps. Open-File Report 96-532, US Geological Survey
16. Geller RJ (1997) Earthquake prediction: A critical review. Geophys J Int 131:425–450
17. Geller RJ, Jackson DD, Kagen YY, Mulargia F (1997) Earthquakes cannot be predicted. Science 275:1616–1617
18. Goes SDB, Ward SN (1994) Synthetic seismicity for the San Andreas fault. Annali Geofis 37:1495–1513
19. Gross S, Rundle JB (1998) A systematic test of time-to-failure analysis. Geophys J Int 133:57–64
20. Harte D, Vere-Jones D (2005) The entropy score and its uses in earthquake forecasting. Pure Appl Geophys 162:1229–1253. doi:10.1007/s00024-004-2667-2
21. Helmstetter A Is earthquake triggering driven by small earthquakes? Phys Rev Let 91:0585014

22. Holliday JR, Chen CC, Tiampo KF, Rundle JB, Turcotte DL, Donnellan A (2007) A RELM earthquake forcast based on pattern informatics. Seis Res Lett 78(1):87–93
23. Holliday JR, Nanjo KZ, Tiampo KF, Rundle JB, Turcotte DL (2005) Earthquake forecasting and its verification. Nonlinear Process Geophys 12:965–977
24. Holliday JR, Rundle JB, Tiampo KF, Klein W, Donnellan A (2006) Modification of the pattern informatics method for forecasting large earthquake events using complex eigenvectors. Tectonophysics 413:87–91. doi:10.1016/j.tecto.2005.10.008
25. Holliday JR, Rundle JB, Tiampo KF, Klein W, Donnellan A (2006) Systematic procedural and sensitivity analysis of the pattern informatics method for forecasting large ($M \geq 5$) earthquake events in southern California. Pure Appl Geophys 163:2433–2454. doi:10.1007/s00024-006-0131-1
26. Jaumé SC, Sykes LR (1999) Evolving towards a critical point: A review of accelerating seismic moment/energy release prior to large and great earthquakes. Pure Appl Geophys 155:279–306
27. Jolliffe IT, Stephenson DB (2003) Forecast Verification. Wiley, Chichester
28. Kagan YY, Jackson DD (2000) Probabilistic forecasting of earthquakes. Geophys J Int 143:438–453
29. Kagan YY, Knopoff L (1981) Stochastic synthesis of earthquake catalogs. J Geophys Res 86(4):2853–2862
30. Kanamori H (2003) Earthquake prediction: An overview. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake & Engineering Seismology. Academic Press, Amsterdam, pp 1205–1216
31. Keilis-Borok V (2002) Earthquake predictions: State-ofthe-art and emerging possibilities. Ann Rev Earth Planet Sci 30:1–33
32. Keilis-Borok V, Shebalin P, Gabrielov A, Turcotte D (2004) Reverse tracing of short-term earthquake precursors. Phys Earth Planet Int 145:75–85
33. Keilis-Borok VI (1990) The lithosphere of the earth as a nonlinear system with implications for earthquake prediction. Rev Geophys 28:19–34
34. King GCP, Bowman DD (2003) The evolution of regional seismicity between large earthquakes. J Geophys Res 108:2096
35. Kossobokov VG, Keilis-Borok VI, Turcotte DL, Malamud BD (2000) Implications of a statistical physics approach for earthquake hazard assessment and forecasting. Pure Appl Geophys 157:2323–2349
36. Lindh AG (2005) Success and failure at Parkfield. Seis Res Let 76:3–6
37. Lomnitz C (1994) Fundamentals of Earthquake Prediction. Wiley, New York
38. Main IG (1999) Applicability of time-to-failure analysis to accelerated strain before earthquakes and volcanic eruptions. Geophys J Int 139:F1–F6
39. Mason IB (2003) Binary events. In: Joliffe IT, Stephenson DB (eds) Forecast Verification. Wiley, Chichester, pp 37–76
40. Mogi K (1985) Earthquake Prediction. Academic Press, Tokyo
41. Molchan GM (1997) Earthquake predictions as a decision-making problem. Pure Appl Geophys 149:233–247
42. Ogata Y (1988) Statistical models for earthquake occurrences and residual analysis for point processes. J Am Stat Assoc 83:9–27
43. Ogata Y (1999) Seismicity analysis through point-process modeling: a review. Pure Appl Geophys 155:471–507
44. Ogata Y, Zhuang J (2006) Space-time ETAS models and an improved extension. Tectonophysics 413:13–23. doi:10.1016/j.tecto.2005.10.016
45. Rikitake T (1982) Earthquake forecasting and warning. D. Reidel Publishing Co, Dordrecht
46. Robinson R (2000) A test of the precursory accelerating moment release model on some recent New Zealand earthquakes. Geophys J Int 140:568–576
47. Robinson R, Benites R (1995) Synthetic seismicity models of multiple interacting faults. J Geophys Res 100:18229–18238
48. Robinson R, Benites R (1996) Synthetic seismicity models for the Wellington Region, New Zealand: implications for the temporal distribution of large events. J Geophys Res 101:27833–27844
49. Rundle JB, Rundle PB, Donnellan A (2005) A simulation-based approach to forecasting the next great San Francisco earthquake. Proc Natl Acad Sci 102(43):15363–15367
50. Rundle JB, Rundle PB, Donnellan A, Fox G (2004) Gutenberg–Richter statistics in topologically realistic system-level earthquake stress-evolution simulations. Earth Planets Space 55(8):761–771
51. Rundle JB, Tiampo KF, Klein W, Martins JSS (2002) Self-organization in leaky threshold systems: The influence of near-mean field dynamics and its implications for earthquakes, neurobiology, and forecasting. Proc Natl Acad Sci USA 99(Suppl 1):2514–2521
52. Rundle JB, Turcotte DL, Shcherbakov R, Klein W, Sammis C (2003) Statistical physics approach to understanding the multiscale dynamics of earthquake fault systems. Rev Geophys 41(4):1019. doi:10.1029/2003RG000135
53. Rundle PB, Rundle JB, Tiampo KF, Donnellan A, Turcotte DL (2006) Virtual California: fault model, frictional parameters, applications. Pure Appl Geophys 163:1819–1846
54. Sammis CG, Bowman DD (2006) Competing models for accelerating moment release before large earthquakes. 5th Annual ACES International Workshop, Mauii, Hawaii, USA
55. Sammis CG, Bowman DD, King G (2004) Anomalous seismicity and accelerating moment release preceding the 2001–2002 earthquakes in northern Baha California, Mexico. Pure Appl Geophys 161:2369–2378
56. Scholz CH (2002) The Mechanics of Earthquakes & Faulting, 2nd edn. Cambridg University Press, Cambridge
57. Schorlemmer D, Jackson DD, Gerstenberger M (2003) Earthquake likelihood model testing. http://moho.ess.ucla.edu/~kagan/sjg.pdf. Accessed 8 Oct 2004
58. Shebalin P, Keilis-Borok V, Zaliapin I, Uyeda S, Nagao T, Tsybin N (2004) Advance short-term prediction of the large Tokachi-oki earthquake, September 25, $M = 8.1$: A case history. Earth Planets Space 56:715–724
59. Stein RS (1999) The role of stress transfer in earthquake occurence. Nature 402:605–609
60. Swets JA (1973) The relative operating characteristic in psychology. Science 182:990–1000
61. Sykes LR, Jaumé SC (1990) Seismic activity on neighboring faults as a long-term precursor to large earthquakes in the San Francisco Bay area. Nature 348:595–599
62. Sykes LR, Shaw BE, Scholz CH (1999) Rethinking earthquake prediction. Pure Appl Geophys 155:207–232
63. Tiampo KF, Rundle JB, McGinnis S, Gross SJ, Klein W (2002) Eigenpatterns in southern California seismicity. J Geophys Res 107(B12):2354. doi:10.1029/2001JB000562

64. Tiampo KF, Rundle JB, McGinnis S, Gross SJ, Klein W (2002) Pattern dynamics and forecast methods in seismically active regions. Pure Appl Geophys 159:2429–2467
65. Turcotte DL (1991) Earthquake prediction. Ann Rev Earth Planet Sci 19:263–281
66. Turcotte DL (1997) Fractals & Chaos in Geology & Geophysics, 2nd edn. Cambridge University Press, Cambridge
67. Utsu T (1984) Estimation of parameters for recurrence models of earthquakes. Earthquake Res Insti-Univ Tokyo 59:53–66
68. Utsu T (2003) A list of deadly earthquakes in the world: 1500–2000. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake & Engineering Seismology. Academic Press, Amsterdam, pp 691–717
69. Ward SN (1992) An application of synthetic seismicity in earthquake statistics: the Middle America trench. J Geophys Res 97(B5):6675–6682
70. Ward SN (1996) A synthetic seismicity model for southern California: cycles, probabilities, and hazard. J Geophys Res 101(B10):22393–22418
71. Ward SN (2000) San Francisco Bay Area earthquake simulations: a step toward a standard physical earthquake model. Bull Seismo Soc Am 90(2):370–386
72. Working Group on California Earthquake Probabilities (1988) Probabilities of large earthquakes occurring in California on the San Andreas fault. Open-File Report 88-398, US Geological Survey
73. Working Group on California Earthquake Probabilities (1990) Probabilities of large earthquakes in the San Francisco Bay region, California. Circular 1053, US Geological Survey
74. Working Group on California Earthquake Probabilities (1995) Seismic hazards in southern California: probable earthquakes, 1994–2024. Seis Soc Am Bull 85:379–439
75. Working Group on California Earthquake Probabilities (2003) Earthquake probabilities in the San Francisco Bay Region, 2002–2031. Open-File Report 2003-214, US Geological Survey
76. Wyss M (1997) Nomination of precursory seismic quiescence as a significant precursor. Pure Appl Geophys 149:79–114
77. Wyss M, Habermann RE (1988) Precursory seismic quiescence. Pure Appl Geophys 126:319–332
78. Yakovlev G, Turcotte DL, Rundle JB, Rundle PB (2006) Simulation-based distributions of earthquake recurrence times on the San Andreas fault system. Bull Seis Soc Am 96:1995–2007
79. Yang W, Vere-Jones D, Li M (2001) A proposed method for locating the critical region of a future earthquake using the critical earthquake concept. J Geophys Res 106:4151–4128
80. Zechar JD, Jordan TH (2005) Evaluation techniques for alarm-based forecasts. EOS Trans. AGU. Fall meeting
81. Zoback ML (1992) First- and second-order patterns of stress in the lighosphere: The world stress map project. J Geophys Res 97:11703–11728

Books and Reviews

Jolliffe IT, Stephenson DB (2003) Forecast Verification. Wiley, Chichester
Turcotte DL, Schubert G (2002) Geodynamics. Cambridge University Press, Cambridge
Wilks DS (1995) Statistical Methods in the Atmospheric Sciences. Academic Press, San Diego

Earthquake Location, Direct, Global-Search Methods

Anthony Lomax[1], Alberto Michelini[2], Andrew Curtis[3]
[1] ALomax Scientific, Mouans-Sartoux, France
[2] Istituto Nazionale di Geofisica e Vulcanologia, Roma, Italy
[3] ECOSSE (Edinburgh Collaborative of Subsurface Science and Engineering), Grant Institute of GeoSciences, The University of Edinburgh, Edinburgh, United Kingdom

Article Outline

Glossary
Definition of the Subject
Introduction
The Earthquake Location Problem
Location Methods
Illustrative Examples
Future Directions
Bibliography

Glossary

Arrival time The time of the first measurable energy of a seismic phase on a seismogram.

Centroid The coordinates of the spatial or temporal average of some characteristic of an earthquake, such as surface shaking intensity or moment release.

Data space If the data are described by a vector **d**, then the data space $\boldsymbol{D}$ is the set of all possible values of **d**.

Direct search A search or inversion technique that does not explicitly use derivatives.

Earthquake early-warning The goal of earthquake early-warning is to estimate the shaking hazard of a large earthquake at a nearby population center or other critical site before destructive *S* and surface waves have reached the site. This requires that useful, probabilistic constraint on the location and size of an earthquake is obtained very rapidly.

Earthquake location An earthquake location specifies a spatial position and time of occurrence for an earthquake. The location may refer to the earthquake hypocenter and corresponding origin time, a mean or centroid of some spatial or temporal characteristic of the earthquake, or another property of the earthquake that can be spatially and temporally localized. This term also refers to the process of locating an earthquake.

Epicenter The point on the Earth's surface directly above a hypocenter.

Error A specified variation in the value assumed by a variable. See also *uncertainty*.

Global search A search or inversion that samples throughout the prior *pdf* of the unknown parameters.

Hypocenter The point in three-dimensional space of initial energy release of an earthquake rupture or other seismic event.

Importance sampling A sampling procedure that draws samples following the posterior *pdf* of an inverse, optimization or other search problem. Since these problems involve initially unknown, posterior *pdf* functions, importance sampling can only be performed approximately, usually through some adaptive or learning procedure as sampling progresses.

Inverse problem, inversion The problem of determining the parameters of a physical system given some data. The solution of an inverse problem requires measurements of observable quantities of the physical system, and the mathematical expression (the forward problem) that relates the parameters defining the physical system (model space) to the data (data space). In inverse problems, estimates of the unknown parameters in the model space and of their uncertainties are sought from the combination of the available information on the model parameters (prior *pdf*), the data and the forward problem.

Likelihood function A non-normalized *pdf*.

Misfit function A function that quantifies the disagreement between observed and calculated values of one or more quantities. See *objective function*.

Model space If the model parameters are described by a vector $\mathbf{m}$, then model space $\boldsymbol{M}$ is the set of all possible values of $\mathbf{m}$.

Objective function A function expressing the quality of any point in the model space. Inversion and optimization procedures use an objective function to rank and select models. Usually objective functions are defined in terms of misfit functions, and for probabilistic inversion the objective function must be a *pdf* or likelihood function.

Origin time The time of occurrence of initial energy release of an earthquake rupture or other seismic event.

Prior pdf A *pdf* that expresses the information on the unknown parameters available before an inverse problem is solved. For an earthquake location, the prior *pdf* is often a simple function (e. g., boxcar) of three spatial dimensions and time. See also *Inverse problem*.

Probability density function – pdf A function in one or more dimensional space $\boldsymbol{X}$ that (i) when integrated over some interval $\Delta\mathbf{x}$ in $\boldsymbol{X}$ gives a probability of occurrence of any event within $\Delta\mathbf{x}$, and (ii) has unit integral over space $\boldsymbol{X}$, where $\boldsymbol{X}$ represents a space of possible events. An earthquake location *pdf* is often a 3-dimensional probability density function over all possible spatial locations or a 4-dimensional probability density function over all possible spatial locations and times of occurrence.

Posterior pdf A *pdf* that expresses the information about the unknown parameters available after inversion. The posterior *pdf* for an earthquake location is often a function of the three spatial dimensions and the origin time of the hypocenter parameters; this function may be complicated. See also *Inverse problem*.

Ray path A local minimum-time path between a source and receiver of idealized, infinite frequency wave energy of a specified wave type (e. g., *P* or *S*).

Receiver or station Synonyms for an observation point where ground motion is detected and a seismogram recorded.

Seismic phase A distinct packet of energy from a seismic source. Usually refers to a specified wave type (e. g. *P* or *S*) satisfying a particular physics of wave propagation.

Seismicity The distribution in space and time of seismic event locations.

Seismogram An analogue or digital recording of the ground motion at a point (receiver or station) in the Earth. Also called a waveform.

Source A general term referring to an earthquake, explosion or other release of seismic energy as a physical phenomenon localized in space and time.

Station See *receiver*.

Travel time The time that a signal, e. g. elastic wave energy of a seismic phase, takes to propagate along a ray path between two points in a medium.

Uncertainty Random variation in the values assumed by a variable. See also *error*.

Definition of the Subject

An earthquake location specifies the place and time of occurrence of energy release from a seismic event. A location together with a measure of size forms a concise description of the most important characteristics of an earthquake. The location may refer to the earthquake's epicenter, hypocenter, or centroid, or to another observed or calculated property of the earthquake that can be spatially and temporally localized. A location is called *absolute* if it is determined or specified within a fixed, geographic coordinate system and a fixed time base (e. g., Coordinated

Universal Time, UTC); a location is called *relative* if it is determined or specified with respect to another spatio-temporal object (e. g., an earthquake or explosion) which may have unknown or uncertain absolute location.

For rapid hazard assessment and emergency response, an earthquake location provides information such as the locality of potential damage or the source region of a possible tsunami, and a location is required to calculate most measures of the size of an earthquake, such as magnitude or moment. Locations are required for further analysis and characterization of the event, for studies of general patterns of seismicity, to calculate distributions of stress and strain changes around the earthquake, for assessing future earthquake hazard, and for basic and applied seismological research.

Since earthquakes occur deep in the Earth, their source locations must be inferred indirectly from distant observations, and earthquake location is thus a remote-sensing problem. Most commonly an earthquake location is determined by the match or misfit between observed arrival times of seismic wave-energy at seismic stations, and predictions of these arrival times for different source locations using a given elastic-wave speed model; this is an inverse problem. Essentially, many potential locations (place and time) are examined and those for which some measure of misfit between predicted and measured arrival times is smallest are retained as best estimates of the true location.

Many numerical location methods involve linearization of the equations relating the predicted arrival times to the location through Taylor expansion involving partial derivatives; these are called *linearized* methods. Methods that do not involve linearization are called *nonlinearized* or *direct-search* methods. The term *nonlinear* is used ambiguously in geophysics to refer to linearized-iterated and to nonlinearized methods. In this chapter we focus on nonlinearized, direct-search methods, and to avoid ambiguity we identify them with the term *direct-search*.

Direct-search location can be performed through graphical analysis, regular or stochastic searches over a space of possible locations, and other algorithms. Direct-search earthquake location is important because, relative to linearized methods, it is easy to apply with realistic earth models which may have abrupt and complicated velocity variations in three-dimensions, it places little restriction on the form of the measure of misfit, it is stable (i. e., does not suffer numerical convergence problems) when the observations are insufficient to fully constrain the spatial location or origin time, and it can produce comprehensive, probabilistic solutions which indicate the full location uncertainty, often a complex function of space and time. Conversely, the primary advantage of linearized location methods is that they are much less demanding computationally than direct-search methods.

Introduction

Most commonly, an earthquake location is determined using observed seismic-phase arrival-times and associated uncertainties, and predicted travel times in a given wave-speed model. Ideally, the location procedure will determine a 4-dimensional, posterior probability density function, or location *pdf*, over all possible solutions (spatial locations and origin times). This location *pdf* quantifies the agreement between predicted and observed arrival times in relation to all uncertainties, and forms a complete, probabilistic solution. In practice, however, an earthquake location is often specified as some optimal solution (a point in space and time) with associated uncertainties.

The earliest, formal earthquake locations using seismic-phase arrival-time observations employed direct-search procedures such as graphical methods (e. g., [37]) or simple grid searches (e. g., [52]). The advent of digital computers in the 1960's lead to the use of iterated linearized approaches based mainly on Geiger's method [17]. Since the 1980's, the increasing power of digital computers has made large-scale, grid and stochastic direct searches practical for routine earthquake location. Direct-search methods are now used routinely in research and earthquake monitoring (e. g., [22,23,31,32,46,48,61]).

In principle, direct-search methods can be applied to locate the relative positions of ensembles of events, and for joint epicentral determination (e. g., [47]) to simultaneously determine multiple earthquake locations and station corrections related to errors in the velocity model. However, the high-dimensionality of such problems makes direct-search solution difficult and computationally demanding; at the present time these problems are usually performed through large scale, linearized procedures. For these reasons, we mainly consider here absolute location of individual events.

In this article we describe the earthquake location problem and direct-search methods used to perform this location, and we present a number of examples of direct-search location. We do not compare different direct-search location methods or compare direct-search to linearized algorithms, but instead focus on illustrating important features and complexity in earthquake location results. For this reason we emphasize direct, global-search, probabilistic location, which produces general and complete solutions that best illuminate these features and complexity.

The Earthquake Location Problem

An Inherently Nonlinear Problem

In a homogeneous medium with wave speed v and *slowness* defined to be $u = 1/v$, the arrival time, t_{obs}, at an observation point x_{obs}, y_{obs}, z_{obs} of a signal emitted at origin time t_0 from a source location at x_0, y_0, z_0 is,

$$t_{\text{obs}} = t_0 + u\left[(x_{\text{obs}} - x_0)^2 + (y_{\text{obs}} - y_0)^2 + (z_{\text{obs}} - z_0)^2\right]^{1/2}, \tag{1}$$

This expression shows that a change in the spatial position of the source introduces a nonlinear change in t_{obs}, even in the simplest possible medium. When the speed v and hence slowness u are inhomogeneous in space, the arrival time at the observation point becomes,

$$t_{\text{obs}} = t_0 + \int_{\mathbf{r}_0(s)} u(\mathbf{r}_0)\, \mathrm{d}s\,, \tag{2}$$

where $\mathbf{r}_0(s)$ denotes a point at distance s along ray path $\mathbf{r}_0$ between source and receiver locations. Equation (1) is a special case of (2) that has straight source-receiver ray paths. Equation (2) is nonlinear since a change in the source location changes the ray path over which the integral is calculated. Thus, earthquake location, which maps arrival times into spatial location and origin time, is inherently a nonlinear problem.

The Observed Data

Data used to constrain earthquake locations are usually derived from seismograms recorded at seismic stations distributed around the earthquake source area, usually at or near the surface of the Earth. The derived data for earthquake location include arrival times, polarization angles, or array slownesses and azimuths. For earthquake location there are three important aspects of this data determination:

1) choosing locations for the stations (before data have been collected),
2) deriving data and associated uncertainties from the seismograms, and
3) association of the derived data into subsets of data corresponding to unique events.

The first important aspect of data determination is choosing station locations with the goal of constraining as tightly as possible event locations for a given source area; this is classified as a problem of "experimental design" in the field of statistics. The design problem must be resolved prior to data collection and so is posed in terms of *expected* data, and *expected* location results. We describe experimental design techniques in more detail later, after introducing and discussing the location solution on which such designs depend.

Once stations are installed and have recorded seismograms from earthquakes of interest, a data set must be extracted that is sensitive to the event source location, and which we can associate with some physics (e. g., of P or S waves) and paths of wave propagation. Most commonly for earthquake location the data set will be phase arrival times and associated uncertainties picked manually or automatically from seismograms (Fig. 1). It is often easy to detect and pick arrivals manually since the human eye can identify a change in amplitude or frequency in the signal even in the presence of significant noise. The picking of the S phase is sometimes more difficult because it arrives in the P coda and can be preceded by S to P or other converted phases; this is a common problem with recordings at local (e. g., up to about 100 km) and near-regional (e. g., up to about 300 km) distances, especially if horizontal component seismograms are not available. The automatic detection, identification and picking of P and S arrivals is much more difficult, especially in the presence of high noise levels. However, automatic detection and picking is faster and, for the case of initial P phases or other phases with characteristic forms, can produce a more consistent data set than manual processing. Automatic arrival detection relies on identifying temporal variations in energy, frequency content, polarization or other characteristics of the signal which are anomalous relative to their background or noise level. Often the detection and picking algorithms are applied to filtered and processed time-series in order either to reduce noise, or to augment the signal in pre-set or dynamically-determined frequency bands or polarization directions. See [18] for an approach that exploits neural networks for phase identification and picking, and [82] for a review and systematic comparison of several approaches to automatic detection and picking.

The data used for earthquake location (e. g., arrival times) must have associated uncertainty estimates otherwise the location uncertainty and a probabilistic solution (i. e., location *pdf*) can not be calculated. Most generally, a vector $\mathbf{d}$ that describes the data takes values from a data space $\boldsymbol{D}$, and $p(\mathbf{d})$ denotes the *pdf* representing uncertainty in $\mathbf{d}$. The uncertainty in arrival time data should include not only an estimate of the uncertainty in the picked phase arrival time, but also uncertainty in which phase (e. g., P or S) is associated with the pick. If there are multiple expected phase arrivals close to the picked arrival time of a phase, then ideally these should all be taken as can-

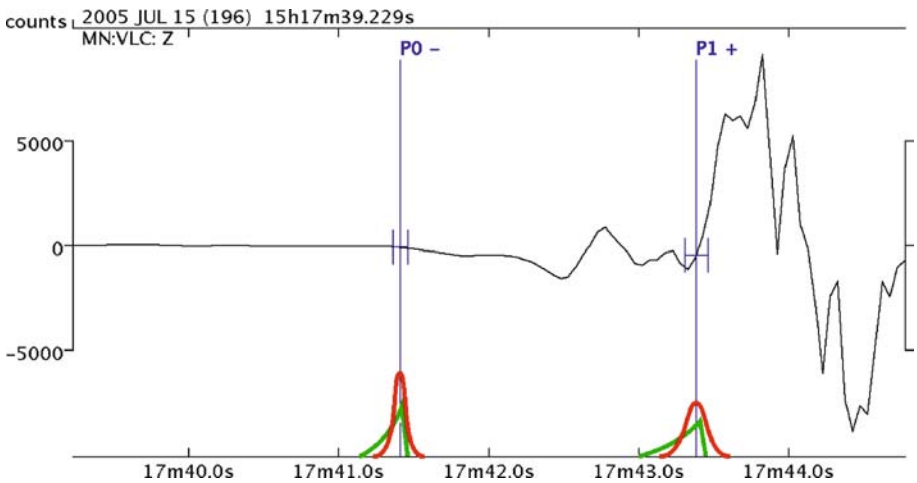

Earthquake Location, Direct, Global-Search Methods, Figure 1

A short waveform segment (∼ 5 sec) showing the first *P* wave arrivals from a small earthquake in Northern Italy recorded on a vertical component seismogram at a nearby station. Automatic arrival pick times (*vertical blue lines*) and uncertainty estimates (*blue error-bars*) are shown for two phases, a first arriving *P* phase (P0) and secondary *P* arrival (P1). The red curves show the data *pdf* functions representing these arrival pick times and uncertainties for an event location procedure where the data *P*(d) is approximated by a normal distribution. The *green curves* show irregular, asymmetric *pdf* functions that may more accurately represent the uncertainty in the phase arrival times; if such *pdf* functions were routinely estimated during arrival picking, they could be used for direct-search location without major difficulty

didate phase types for the arrival. Also, the pick uncertainty of each phase may be best described by a *pdf* that is asymmetric in time, since usually a latest-possible time for a pick is much easier to define than the earliest time (Fig. 1). True data uncertainty *pdf*'s are therefore generally multi-modal, and can be quite complex to calculate and parametrize. In practice, an enumerated quality indication or, at best, a simple normal distribution (Gaussian uncertainty) is used to describe the picking error, and the phase association is usually fixed (e. g. to *P* or *S*) so corresponding uncertainties are ignored. In many cases these simplified data uncertainty estimates will lead to bias or increased error in the resulting event locations.

The third important aspect of data determination is the association of the derived data into sets of data for unique events. For example, this association may entail the assignment of each observed arrival time within a specified time window to a unique event, forming the minimum possible number of events and corresponding arrival time sets required to explain observed data. This association procedure can be very difficult, especially with automatic systems and when there are signals from multiple seismic events that are close or overlapping in time (e. g., [24]), and we do not address this issue further here. In the following, except for an examination of outlier data, we implicitly assume that location is performed with a data set that is already associated to a unique event.

The Velocity or Slowness Model

The velocity or slowness model specifies seismic wave-speeds in the region of the Earth containing the sources, the receivers and the ray paths between the sources and receivers. Equation (2) is nonlinear with respect to source location, but also with respect to slowness since a change in the slowness distribution of the medium changes the ray path. The velocity structure is sometimes estimated through coupled, simultaneous inversions for velocity structure and event locations (commonly called seismic tomography), but these are very large inverse problems solved mainly with linearized methods. Usually for earthquake location the velocity model is taken as known and fixed.

Often, for computational convenience or due to lack of information, the velocity model is parametrized with velocity varying only with depth. This is commonly called a laterally homogeneous or 1-dimensional (1D) model. Such a model may consist of one or more layers of constant or vertical-gradient wave-speeds. For work at a local or near-regional scale the layers may be horizontal and flat; for larger, regional or global scale problems the layers should be spherically symmetric shells to represent the curvature of the Earth. When more information on the velocity structure is available, a 3D model may be used in order to increase the accuracy of the ray paths and

travel times, and hence of the locations, relative to a 1D model. All models, whether 1D or 3D, are described by a limited number of parameters and include some form of spatial averaging or interpolation with respect to the true Earth. Although 3D models can potentially represent velocity variations in the Earth more accurately than 1D models, in practice the velocities in 3D models can locally be poorly constrained and have large errors. It is therefore often important to consider several different possible 1D and 3D velocity structures in a location study, either to test the sensitivity of the locations to errors in velocity, or to better estimate the travel-time uncertainties and produce a more meaningful location *pdf*. In principle the use of diverse velocity models poses no difficulties with direct-search location methods.

The Travel-Time Calculation

The theoretical seismic wave travel-times through a given velocity model between any particular source and receiver locations are required by most location algorithms. The calculation of travel times is commonly referred to as forward modeling, because inverse theory need not be invoked. There are three basic classes of methods to calculate the travel times: full-waveform methods, ray methods, and Huygens wavefront or eikonal methods.

Full-waveform methods (e. g., [1]) produce complete synthetic seismograms from which predicted travel times can be extracted. These methods include frequency–wavenumber or modal-summation techniques which are valid for a broad range of frequencies and can produce exact waveforms, but which are only applicable for relatively simple velocity structures. Numerical techniques such as finite elements and finite differences can accurately model full wave phenomena in complicated structures, but these methods typically require large computing resources and computing time. Currently, full-waveform methods are rarely used to determine predicted travel times for earthquake location because these times can be obtained directly and more efficiently with ray and eikonal methods.

Ray methods (e. g., [1,7,75]) provide travel times and the path, or ray, traveled by high-frequency waves, and can be applied to complicated and 3D velocity structures. With simple model parametrizations such as flat layers with constant or gradient velocity, ray paths and travel times can be determined very rapidly with analytical or semi-numerical algorithms. For these and more complicated models, shooting, or ray tracing techniques generate rays by iterativly solving of a set of ray-tracing equations starting in a specified direction at the source or receiver location. The ray that passes through a specified end point is found by a search over the direction at the starting point; this search can be time consuming or unstable. In addition, shooting methods do not produce diffracted arrivals (e. g., "head waves" from the Mohorovičić discontinuity) which are often the first arriving signal at near-regional distances and are thus critical for earthquake location. Two-point, ray bending and perturbation techniques rely on Fermat's principle of least time: an initial guess at the ray between two points is perturbed repeatedly to attain a minimum travel time and corresponding ray between the points. These techniques perform best with smooth models, but do produce diffracted arrivals. In general, except for analytical or semi-numerical algorithms in simple models, ray methods are too computationally expensive for direct-search location, which usually requires evaluation of travel times between a very large number of source and receiver positions. However, some ray bending methods (e. g., [39,77]) are efficient enough to be used in direct-search location when a relatively small number of source and receiver positions position need to be examined.

Wavefront, eikonal and graph-based methods [75] provide travel-times of the first arriving, high frequency waves including diffracted arrivals, and are efficient and applicable with complicated, 3D velocity structures. In effect, these methods propagate wavefronts through a velocity model with repeated application of Huygen's principle, by considering a large number of virtual sources (*Huygens sources*) along each wavefront. At time t these sources emit circular wavelets which expand for a small time Δt through the (constant) local, medium velocity. The locus of the first arriving circular wavelets defines the new wavefront location at time $t + \Delta t$. The synthetic travel time of the first-arriving energy at the receiver is the time at which a wavefront first touches the receiver. In practice, this problem is solved on a computer either by replicating this "wavefront marching" process (e. g., [50,66]), or by finding a numerical solution to the eikonal equation (e. g., [44,79]), or by graphical analysis (e. g., [38]). Though wavefront, eikonal and graph-based methods produce directly only the travel time of the first-arriving signal, information about the path traveled by the signal can be derived numerically from the travel-time field or from ray-tracing, and travel-times of secondary arrivals can be obtained through multi-stage calculations (e. g., [44,51]).

Wavefront, eikonal and graph-based methods can efficiently generate the travel-times from one point in a gridded velocity model to all other points in the model. This makes these methods particularly useful for direct-search location, which may test a large number of possible source positions widely distributed in space. For this purpose, the travel times from each seismic station to all points in the

model can be pre-calculated and stored in computer disk files or in memory; obtaining the travel time from a station to any other point then reduces to a simple lookup (e. g., [35,38]).

A Complete Solution – Probabilistic Location

Consider a vector $\mathbf{d}_{\text{obs}}$ of observed data (e. g., arrival times) that takes values in a data space $\boldsymbol{D}$, and let $p(\mathbf{d})$ be the *pdf* over $\boldsymbol{D}$ describing the data uncertainty in $\mathbf{d}_{\text{obs}}$ due to measurement and processing uncertainties. Similarly, let $\mathbf{m}$ denote the vector of source location parameters (spatial coordinates and origin time) which take values from parameter space $\boldsymbol{M}$. Let $p(\mathbf{m})$ be the prior *pdf* representing all information available about the location before (prior to) using the data $\mathbf{d}_{\text{obs}}$; $p(\mathbf{m})$ might include knowledge of the known, active fault zones in the area, or might specify the bounds of a region within which we know the event occurred from damage reports, or of a region containing the network of stations that recorded the event. Also consider the forward problem (e. g., travel time calculation) relating $\mathbf{m}$ to a vector of predicted data (e. g., arrival times), $\mathbf{d}_{\text{calc}}$. In general the forward problem may also be uncertain, for example due to uncertainties in velocity structure, so we use $F(\mathbf{d}, \mathbf{m})$ to denote the *pdf* of the relationship between $\mathbf{d}_{\text{calc}}$ and $\mathbf{m}$ as constrained by the forward problem.

As an example of F, it is commonly assumed that for each particular $\mathbf{m}$, the corresponding predicted data $\mathbf{d}_{\text{calc}}$ are given by a function $\mathbf{f}(\mathbf{m})$ with negligible errors. Then the conditional *pdf* $F(\mathbf{d}|\mathbf{m})$ (the probability distribution of $\mathbf{d}$ when $\mathbf{m}$ is fixed at a particular value) is described by $F(\mathbf{d}|\mathbf{m}) = \delta[\mathbf{d} - \mathbf{f}(\mathbf{m})]$ where δ is the Dirac delta-function. Also, the forward problem is often assumed to place minimum possible constraint on parameters $\mathbf{m}$; the *pdf* describing this state of information about $\mathbf{m}$ is called the homogeneous distribution, represented by $\mu(\mathbf{m})$. No *pdf* exists that describes zero information, but *some* information about $\mathbf{m}$ always exists in practice (the positivity of parameter values, for example); $\mu(\mathbf{m})$ describes that minimum state of information. In that case the forward problem is given by [73,74],

$$F(\mathbf{d}, \mathbf{m}) = \delta[\mathbf{d} - \mathbf{f}(\mathbf{m})]\, \mu\,(\mathbf{m}) \ . \tag{3}$$

A solution to the earthquake location problem is found by combining the information in the observed data, $p(\mathbf{d})$, the prior *pdf*, $p(\mathbf{m})$, and the ability of the forward problem to predict the observed data, $F(\mathbf{d}, \mathbf{m})$ [73,74]. This is achieved in a probabilistic framework by constructing a *pdf* Q describing the state of posterior (post-experimental) information by:

$$Q(\mathbf{d}, \mathbf{m}) = k\, \frac{p(\mathbf{d})F(\mathbf{d}, \mathbf{m})p(\mathbf{m})}{\mu(\mathbf{d}, \mathbf{m})}, \tag{4}$$

where the constant k normalizes Q to unit integral over $\boldsymbol{D} \times \boldsymbol{M}$ and $\mu(\mathbf{d}, \mathbf{m})$ is the homogeneous distribution over data $\mathbf{d}$ and parameters $\mathbf{m}$. Equation (4) contains all information (from the prior knowledge, data and physics) that could have a bearing on location $\mathbf{m}$, and defines a joint *pdf* between parameters $\mathbf{m}$ and data $\mathbf{d}$. The final, posterior state of information about location parameters $\mathbf{m}$ is given by integrating over the data $\mathbf{d}$ to obtain the marginal posterior *pdf*,

$$Q(\mathbf{m}) = k\, p(\mathbf{m}) \int_{\mathbf{D}} \frac{p(\mathbf{d})F(\mathbf{d}, \mathbf{m})}{\mu(\mathbf{d}, \mathbf{m})} \mathrm{d}\mathbf{d}\ . \tag{5}$$

Equation (5) is the general, probabilistic solution to the inverse problem of event location from the available data since it describes the uncertainty in event location $\mathbf{m}$ given all available information. It is usual to call the integral in (5) the *likelihood* function $L(\mathbf{m})$, which gives a (non-normalized) measure of how good any model $\mathbf{m}$ is in explaining the observed data $p(\mathbf{d})$.

As mentioned earlier it is often the case that $p(\mathbf{d})$ for the observed data is approximated by a Gaussian distribution, described by mean $\mathbf{d}_0$ and covariance matrix $\mathbf{C}_d$. Assuming that the uncertainties in the forward problem F relating $\mathbf{d}$ and $\mathbf{m}$ are negligible results in the form of F in Eq. (3). It is also usually assumed that $\mathbf{d}$ and $\mathbf{m}$ are independent and hence that $\mu(\mathbf{d}, \mathbf{m})$ can be written $\mu(\mathbf{d})\mu(\mathbf{m})$; $\mu(\mathbf{d})$ is usually taken to be constant. With these simplifications, used by many current direct-search location procedures, the (non-normalized) likelihood function is given by,

$$L(\mathbf{m}) = \exp\left\{-\frac{1}{2}[\mathbf{d}_0 - \mathbf{f}(\mathbf{m})]^{\mathrm{T}}\, \mathbf{C}_{\mathbf{d}}^{-1}[\mathbf{d}_0 - \mathbf{f}(\mathbf{m})]\right\}\ . \tag{6}$$

With the above simplifications a maximum likelihood origin time, t_0, can be determined analytically from weighted means of the observed arrival times and the predicted travel times (e. g., [74]), and if the observed and predicted times are uncorrelated we arrive at a likelihood function,

$$L\,(\mathbf{x}) = \exp\left\{-\frac{1}{2}\sum_{\mathrm{i}} \frac{\left[T_i^{\mathrm{o}} - T_i^{\mathrm{c}}\,(\mathbf{x})\right]^2}{\sigma_i^2}\right\}\ , \tag{7}$$

where $\mathbf{x}$ is the spatial part of $\mathbf{m}$, T_i^{o} are observed travel times, T_i^{c} are the calculated travel times for observation i (i. e., T_i^{c} represents the travel time, rather than arrival time,

part of $\mathbf{f}(\mathbf{m})$), and σ_i summarizes the associated standard deviation of uncertainty in T_i^{o} and T_i^{c}.

Though not normalized, $L(\mathbf{x})$ is sufficient to provide the *relative* probability of any location $\mathbf{m}$ being the best estimate of the event location given the available data measurements. Since in practice integrating over all of $\boldsymbol{D} \times \boldsymbol{M}$ to find normalizing constant k in Eq. (5) is often computationally intractable, the product of the prior, spatial location information $p(\mathbf{x})$ (i. e., the spatial part of $p(\mathbf{m})$) and the non-normalized likelihood $L(\mathbf{x})$ is usually taken as the objective function for inversion and searching in direct-search location algorithms. If $L(\mathbf{x})$ is determined throughout the prior *pdf* $p(\mathbf{x})$ through a global-search, then Eq. (5) can be normalized approximately after location. In the following text and examples, we refer to such an approximately normalized function, $p(\mathbf{x})\, L(\mathbf{x})$, as a location *pdf*.

The likelihood function in Eq. (5) is entirely defined by the probabilistic error processes involved. However, often it is desirable to change the approximations employed in deriving Eqs. (6) and (7) from Eq. (5), in order to remove biases or instability in the solution. The approximation in Eq. (6) uses the exponential of an L2-norm misfit function (the term in braces $\{\}$ in Eq. (6) or (7)) to represent the *pdf* of the data error variation, but because data used for location often contain outliers it is often considered that an L1 norm or other Lp norm ($p < 2.0$) is more appropriate (e. g., [69]), where Lp—norm $|\mathbf{x}| = \sqrt[1/p]{\sum |x_i|^p}$. Earthquake location problems formulated with an Lp norm (or indeed other kinds of likelihood functions – see Eq. (8) below), can be solved relatively easily with direct-search methods, which, unlike linearized methods, do not require determination of partial derivatives of the likelihood or objective function with respect to event location.

An alternative to Lp-likelihood functions that is very robust in the presence of outliers is given by the equal differential-time (EDT) formulation [16,31,84]. For the EDT case, the location likelihood is given by,

$$L(\mathbf{x}) = \left[\sum_{a,b} \frac{1}{\sqrt{\sigma_a^2 + \sigma_b^2}} \cdot \exp\left(-\frac{\left\{ [T_a^{\mathrm{o}} - T_b^{\mathrm{o}}] - [TT_a^{\mathrm{c}}(\mathbf{x}) - TT_b^{\mathrm{c}}(\mathbf{x})] \right\}^2}{\sigma_a^2 + \sigma_b^2} \right) \right]^N , \tag{8}$$

where $\mathbf{x}$ is the spatial part of $\mathbf{m}$, T_a^{o} and T_b^{o} are the observed arrival times and TT_a^{c} and TT_b^{c} are the calculated travel times for two observations a and b; the sum is taken over all pairs of observations, and N is the total number of observations. Standard deviations σ_a and σ_b summarize the assigned uncertainties on the observed arrival times and calculated travel times, where it is assumed that the observed and the calculated times are uncorrelated.

In Eq. (8), the first and second terms in brackets in the exponent are, respectively, the differences between the observed arrival times and the differences between the calculated travel times. The exponent is the difference between these two terms, and thus the exponential has a maximum value of 1 which occurs at points $\mathbf{x}$ where the two differences are equal (hence, the name "equal differential time"). Such points $\mathbf{x}$ best satisfy the two observations a and b together, and, in general, the set of $\mathbf{x}$ where the exponential is nonzero forms a "fat," curved surface in 3D space. Because the summation over observations is outside the exponential, the EDT location *pdf* has its largest values for those points $\mathbf{x}$ where the most pairs of observations are satisfied and thus is far less sensitive to outlier data than Lp norms which seek to best satisfy all of the observations simultaneously. Note that the EDT likelihood function $L(\mathbf{x})$ does not require calculation of an origin time t_0; this reduces the hypocenter search to a purely 3-parameter problem and contributes to the robustness of the EDT method. Nevertheless, a compatible estimate of t_0 can be calculated for any hypocenter point $\mathbf{x}$.

Ultimately, the full solution to the probabilistic location problem is a posterior *pdf* which includes as comprehensive as possible uncertainty information over parameters $\mathbf{m}$. This may include multiple "locally-optimal" solutions, e. g., $Q(\mathbf{m})$ or $p(\mathbf{x})\, L(\mathbf{x})$ may have multiple maxima, and may have a highly irregular form. Some studies of seismicity and seismotectonics make explicit use of a probabilistic representation of seismic event locations (e. g., [22,31,46]).

Experimental Design Methods – Choosing Receiver Locations

As noted earlier, it is important to position stations so as to constrain as tightly as possible the event locations for a given source area. The location inverse problem solution in Eqs. (5), (7) or (8) is constrained by prior information on location $p(\mathbf{m})$, by observed data $p(\mathbf{d})$, and by forward-problem physics relating $\mathbf{d}$ and $\mathbf{m}$. One way to significantly influence the form of this inverse problem, and hence uncertainty in its solution, is to change the data we record. Thus, we alter both $p(\mathbf{d})$ and the forward-problem physics, $F(\mathbf{d}, \mathbf{m})$.

For seismic location problems we may change the data by employing experimental design methods to choose or change the locations of seismic receivers. The goal of the design procedure is to place receivers such that the loca-

tion information described by solution $Q(\mathbf{m})$ is expected to be maximized. This is a "macro-optimization" problem where, prior to the occurrence of an earthquake, we optimize the design of the inverse problem that we expect to solve after an earthquake has occurred.

The design is varied such that it maximizes an objective function. This is usually taken to be the expected value of some approximation to the unique measure of information that was discovered by Shannon [67],

$$J(\mathbf{R}) = E_{\mathbf{m}_t}\left\{I\left[Q(\mathbf{m});\ \mathbf{R}, \mathbf{m}_t\right]\right\} \tag{9}$$

where $\mathbf{R}$ is a vector describing the design (e. g., receiver locations), $I[Q(\mathbf{m});\ \mathbf{R}, \mathbf{m}_t]$ is the information contained in the resulting posterior *pdf* $Q(\mathbf{m})$ for design $\mathbf{R}$ when the true parameters (e. g., event location) is $\mathbf{m}_t$, and the statistical expectation $E_{\mathbf{m}_t}$ is taken over all possible $\mathbf{m}_t$ which (according to our prior knowledge) are expected to be distributed according to the prior distribution $p(\mathbf{m})$. $J(\mathbf{R})$ should be maximized.

Within the expectation in Eq. (9), the design criterion $J(\mathbf{R})$ takes account of all possible potential true event locations $\mathbf{m}_t$, their prior probability of occurrence $p(\mathbf{m})$, and the corresponding data (including their uncertainties) that are expected to be recorded for each location (the latter are included within $Q(\mathbf{m})$). To calculate the expectation usually requires integration over a far greater proportion of the model and data spaces, $\boldsymbol{M}$ and $\boldsymbol{D}$ respectively, than need be considered when solving the inverse problem after a particular event has occurred (since then $p(\mathbf{d})$ and hence $Q(\mathbf{m})$ are fixed, and $p(\mathbf{m})$ is more tightly constrained). Consequently, experimental design is generally far more computationally costly than solving any particular inverse problem post-event.

For this reason, design methods invoking linearized approximations to the model-data relationship $F(\mathbf{m}, \mathbf{d})$ (e. g., Eq. (10) below) have been employed by necessity in the past [11,13,49,70], or indeed non-probabilistic methods have been employed (e. g., [9,10,36,71]). Truly nonlinearized design methods have been developed for location problems only relatively recently [12,78,80]. Historically, however, station network geometry has been defined more by heuristics (rules of thumb) and geographical, logistical and financial constraints, with design theory only recently being deployed.

Location Methods

Once data have been recorded and prior *pdf*'s defined, a solution such as that of Eqs. (5), (7) or (8) must be evaluated throughout the prior *pdf*, $p(\mathbf{x})$, to identify one or more "locally-optimal" solutions, or, preferably, to obtain a full probabilistic location *pdf*. This evaluation generally requires direct-search optimization and search techniques, which we discuss below. We first digress and summarize linearized location procedures, which typically determine a single optimal hypocenter along with a simplified and approximate representation of the location *pdf* (e. g., a confidence ellipsoidal centered on the estimated hypocenter and origin time).

Linearized Location Methods

With linearized methods the arrival time expression (2), which is nonlinear with respect to the spatial location $\mathbf{m} = (x, y, z)$, is approximated by a Taylor series expansion around some prior estimate $\mathbf{m}_0 = (x_0, y_0, z_0)$ of the spatial location:

$$f(\mathbf{m}) = f(\mathbf{m}_0) + (\mathbf{m} - \mathbf{m}_0) f'(\mathbf{m}_0) + \frac{(\mathbf{m} - \mathbf{m}_0)^2}{2} f''(\mathbf{m}_0) + O\left[(\mathbf{m} - \mathbf{m}_0)^3\right] \tag{10}$$

where $f(\mathbf{m})$ is the forward problem that calculates an arrival time d_{calc} given a location $\mathbf{m}$ (e. g., $f(\mathbf{m})$ might represent the right hand side of Eq. (2) directly). A linear vector-matrix inverse problem is obtained if we approximate the forward problem for all d_{calc} by using only the first two terms of the Taylor series. The resulting vector-matrix equation may be solved using linear algebraic methods. This process is called *linearized inversion*.

Usually, this linearized inversion is iterated: the prior estimate $\mathbf{m}_0$ is set equal to the newly-found, best-fit location, the problem is re-linearized around this new estimate using Eq. (10), and the new linear problem solved again. This method may be repeated (iterated) many times, as needed to attain some convergence criteria.

Linearized methods produce a single, best-fit (e. g., maximum likelihood) hypocenter and origin time location, and associated, linearly-estimated uncertainties, such as a multi-dimensional, normal-distribution centered on the best-fit hypocenter and origin time. However, this linearized solution is often a poor representation of the complete solution *pdf* (Fig. 2 and see examples), and it may be unstable when the *pdf* is irregular or has multiple peaks due to insufficient or outlier data, velocity model complexities, and other causes (e. g., [5,34]).

Direct-Search Location Methods

The earliest, formal earthquake locations from phase arrival time observations used nonlinearized procedures. Milne [37] describes and applies several graphical and algebraic methods to determine earthquake locations. These

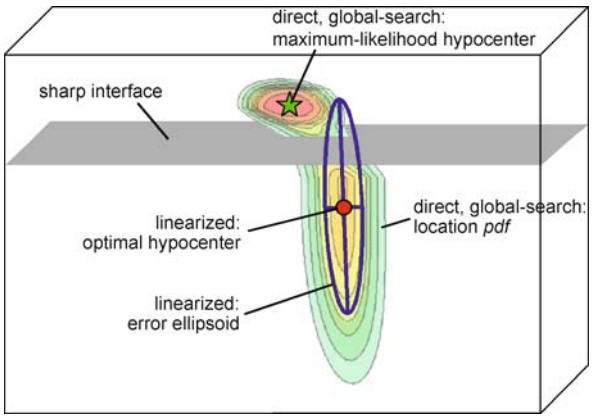

Earthquake Location, Direct, Global-Search Methods, Figure 2
Schematic diagram comparing linearized and direct-search locations for the case where the complete location *pdf* is moderately complicated, with two maxima. This example arises from the case of a location at the limits of the recording network and near a sharp, horizontal interface in the velocity model between lower velocities above and higher velocities below. The *colored, contoured form* shows the true location *pdf*, as should be determined by a complete, probabilistic, direct-search location procedure. A linearized location that iterates from an initial trial location below the sharp interface will find an optimal hypocenter near the secondary, local maximum of the location *pdf*, below the interface. The linearized error ellipsoid, based on the curvature of the misfit function at this optimal hypocenter, reflects the form of this secondary maximum only. The linearized location procedure never identifies or explores the primary maximum of the *pdf* above the sharp interface, and produces incorrect error information above this interface (i. e. the uppermost part of the error ellipsoid). A probabilistic, direct, global-search procedure can determine the complete location *pdf* and identify correctly the maximum likelihood hypocenter located above the sharp interface

include a perpendicular bisector method for the case of 3 or more simultaneous arrival time observations (related to the modern arrival order or bisector method), a method of hyperbolae based on the differences in arrival times at pairs of stations (related to the modern EDT method) and a method using the differences in arrival times of different wave types at individual stations. The latter is a generalization of the method of circles using *S*-*P* times, in which the distance from a station to the source is, for given *P* and *S* velocity models, a function of the difference of the *S* and *P* arrival times; an epicenter can be constrained with such *S*-*P* based distances from 3 stations. Reid [52] determined a hypocenter location for the great 1906 California earthquake through a coarse, systematic grid search over velocity, position along the causative fault and depth, solving for the origin time and wave velocity by least-squares at each grid point.

The arrival order or bisector method [2,42] is a nonlinear, geometrical approach that uses the constraint that if a phase arrival is earlier at station A than at station B, then the event is closer to A than to B (assuming the velocity model is such that arrival order implies distance order). Applying this constraint to all pairs of stations defines a convex region containing the event location. This method is useful for obtaining some constraint on the location of events far outside of an observing station network, and for rapidly and robustly obtaining starting locations for linearized methods.

Most other modern, direct-search earthquake location methods (excluding graphical methods that are now mainly used for illustrative and educational purposes) are based on deterministic or stochastic searches which may be exhaustive or directed and evolutionary. These searches are used to explore or map likelihood functions such as those given in Eqs. (5), (7) or (8). When these searches gather and retain information globally, throughout the prior *pdf* $p(\mathbf{x})$, they can produce a complete, probabilistic location *pdf*. Otherwise, searches may determine a global or local maximum of the location *pdf*, or may explore the neighborhood around these optimal points to locally estimate the *pdf* and obtain uncertainty information.

Regular, Deterministic Search Regular and deterministic searches, such as grid-searches, nested grid-searches and stochastic, "crude" Monte–Carlo searches (e. g., [20,62]) use global and well-distributed sampling of the model space and thus can estimate the complete location *pdf*. All of these approaches are computationally demanding for problems with many unknown parameters, large parameter spaces, or time consuming forward calculations, because the number of models that must be tested can be very large. These methods have been successfully applied to the determination of optimal hypocenters (i. e., [14,26,61,69]), and to probabilistic location (i. e., [6,34,38,83]), but their inefficiency may impose unacceptable limitations on the number of events that can be considered, or on the size of the search volume.

Directed Search Directed, stochastic search techniques include evolutionary, adaptive global search methods such as the genetic algorithm [19,59] and simulated annealing [28,53,72]. The simplex method is a directed, deterministic search technique that is nonlinearized and can be used for earthquake location (e. g., [48]). Most of these methods were developed for optimization or the identification of some very good solutions, which is equivalent to identifying a global or local maximum of the location *pdf*. In general, these methods do not explore the prior

pdf $p(\mathbf{x})$ in a manner that can produce complete, probabilistic solutions to inverse problems. For example, the genetic algorithm performs global searching and may be one of the most efficient stochastic methods for optimization, but it does not use well distributed sampling (the sampling tends to converge rapidly to the region of a locally optimum solution). Similarly, in the simulated annealing, random-walk method the interaction of its variable "temperature" parameter and step size with the local structure of the misfit function can lead to convergence and stalling near a locally optimum solution, and a sample distribution that is neither well nor globally distributed. Both the genetic algorithm and simulated annealing can be tuned to sample more broadly and in the limit become crude Monte Carlo searches, but this removes the main advantage of these methods – that of rapid stochastic optimization.

Though not directly applicable to complete, probabilistic location, directed search algorithms are useful for direct-search, earthquake hypocenter estimation because of their efficiency (e. g., [4,48,60,61]).

Importance sampling The efficiency of a Monte Carlo algorithm used to estimate properties of a target (misfit or likelihood) function can be increased by choosing a sampling density which follows the target function as closely as possible [20,30,45]. Techniques that follow this rule are referred to as importance sampling methods, and were originally developed in physics for fast and accurate numerical integration of multi-dimensional functions. The target function is unknown, however, and consequently the optimum importance sampling distribution cannot be determined a priori. Instead, improved efficiency is attained by adjusting (or adapting or evolving) the sampling by using information gained from previous samples so that the sampling density tends towards the target function as the search progresses [30,40,45,65]. For example, importance sampling to determine an earthquake location *pdf* or likelihood function (e. g., Eqs. (5), (7) or (8)), can be obtained by beginning with a sampling that follows the prior *pdf*, $p(\mathbf{m})$, and then adjusting the sampling as the search progresses so that the sampling density approaches the posterior, location *pdf*.

Importance sampling techniques that can be used to find complete, probabilistic solutions to inverse problem include the VEGAS algorithm [30], the Metropolis algorithm [40], the neighborhood algorithm [55] and, for three-dimensional problems, oct-tree [33]. Other importance sampling methods are discussed in Hammersley and Handscomb [20] and in Press et al. [45] in the context of numerical integration.

The VEGAS algorithm [30,45] performs importance sampling by accumulating appropriate sampling distributions independently for each parameter as the sampling proceeds. This method can give very good estimates of an individual or a joint marginal *pdf*, but it looses efficiency if the target function includes strong correlation between parameters or if it is independent of some parameters [45]. In addition, the VEGAS algorithm may be difficult or impossible to implement with prior information, such as smoothness constraints, that introduces correlation between parameters. Consequently, this algorithm may not be appropriate for some geophysical problems, including earthquake location, when the location parameters are often correlated or poorly resolved.

The Metropolis or Metropolis-Hastings algorithm (e. g., [40]) is similar to simulated annealing but with a constant temperature parameter. The Metropolis algorithm performs a random walk in the model space, testing at each step nearby trial samples which are accepted or rejected after evaluation of the forward problem according to a likelihood $L(\mathbf{m})$. In [40] it is shown that this algorithm samples from the posterior *pdf* of the problem and is therefore an importance sampling method. They show that, in the limit of a very large number of trials, it will not become permanently "trapped" near local maxima and consequently will produce global sampling. Also, because it is a random walk technique, the Metropolis algorithm can perform well even if the volume of the significant regions of the posterior *pdf* is small relative to the volume of the prior *pdf*. However in practical application, with a finite number of samples, this algorithm can become trapped in strong local maxima of the posterior *pdf* if this function is complicated. The Metropolis algorithm has been applied to earthquake location in 3D structures [34,35].

Another recently developed importance sampling technique used in geophysics is the neighborhood algorithm [55,56,57], applicable to high dimensional model spaces. Given an existing set of samples of the objective function, the neighborhood algorithm forms a conditional *pdf* using an approximate Voronoi cell partition of the space around each sample. The algorithm generates new samples through a uniform random walk within the Voronoi cells of the best fitting models determined so far. This algorithm is applied to the 4D hypocenter location problem in [27,58].

The oct-tree importance-sampling method [33] uses recursive subdivision and sampling of rectangular cells in three-dimensional space to generate a cascade structure of sampled cells, such that the spatial density of sampled cells follows the target *pdf* values. The relative probability

that an earthquake location lies within any given cell i is approximately,

$$P_i = V_i L(\mathbf{x}_i) \ , \tag{11}$$

where V_i is the cell volume and $\mathbf{x}_i$ is the vector of coordinates of the cell center. Oct-tree importance-sampling is used to determine a location *pdf* by first taking a set of samples on a coarse, regular grid of cells throughout the search volume. This is followed by a recursive process which takes the cell k that has the highest probability P_k of containing the event location, and subdividing this cell into 8 child cells (hence the name oct-tree), from which 8 new samples of the *pdf* are obtained. These samples are added to a list of all previous samples, from which the highest probability cell is again identified according to Eq. (11). This recursive process is continued until a predetermined number of samples are obtained, or until another termination criterion is reached.

For most location problems, including those with a complicated location *pdf*, the oct-tree recursive subdivision procedure converges rapidly and robustly, producing an oct-tree structure of cells specifying location *pdf* values in 3D space. This oct-tree structure will have a larger number of smaller cells in areas of higher probability (lower misfit) relative to areas of lower *pdf* value and thus the oct-tree method produces approximate importance-sampling without the need for complex geometrical constructs such as Voronoi cells. Oct-tree sampling can be used with the L2-norm likelihood function in Eq. (7) or the EDT likelihood function in Eq. (8), since both require searching over three-dimensional spatial locations only. Oct-tree sampling has been applied to earthquake location in 3D structures [22,23,31,32]; we use this sampling method to determine locations in the examples presented below. Though limited to determination of the 3D, spatial location, this recursive sampling procedure can be extended to 4D to allow determination of the origin time.

Illustrative Examples

We illustrate the concepts described in the previous sections using an M3.3 earthquake that occurred in the Garfagnana area of Northern Tuscany, Italy, on March 5, 2007 at 20:16 GMT. The earthquake was recorded by stations of the Italian National Seismic Network (INSN) at distances from less than 10 km to more than 300 km. We use manually picked P and S phase arrival times from the INSN bulletin with Gaussian uncertainties (standard deviations from 0.01 to 0.1 s), and a 1-D velocity model similar to the standard model used by INSN for routine earthquake location in Italy. We perform all event locations with the probabilistic location program NonLinLoc [31,34,35] (http://www.alomax.net/nlloc; NLL hereafter), using the oct-tree sampling algorithm (Sect. "Location Methods") to perform a global-search within a parameter space $\boldsymbol{M}$ formed by a rectangular volume 360 km on each side and from the Earth's surface to 35 km depth (except as noted in figure captions). We use the L2-norm (Eq. (7)) or EDT (Eq. (8)) likelihood functions to obtain location *pdf*'s in 3D space and corresponding maximum likelihood origin times.

In order to describe the location problem and the solution quality for each of the examples presented below we focus on geometrical properties of the location *pdf*, which represents most completely the results of probabilistic, direct, global-search methodologies. We also consider the maximum likelihood hypocenter, defined as the point in space of the maximum value of the location *pdf*, and the corresponding origin time. We examine statistics of the quality of the solutions using the half-lengths of three principal axes of a 68% confidence error ellipsoid approximation to the location *pdf*, $\boldsymbol{l}_{\text{ell}}$, the weighted, root-mean-square of the arrival residual (observed – calculated) times, *rms*, and a relative measure of the volume of the high likelihood region of the location *pdf*, V_{pdf}, given by,

$$V_{\text{pdf}} = \int_{\boldsymbol{M}} \frac{pdf(\mathbf{x})}{pdf^{\max}} \mathrm{d}V \ , \tag{12}$$

where $pdf^{\max}$ is the maximum value of the location *pdf* in $\boldsymbol{M}$. We also make use of standard measures of the experimental design quality (i.e, stations coverage) including the *gap* – the largest angle between the epicenter and two azimuthally adjacent stations used for location, and the distance Δ_0 from the hypocenter to the closest station. These indicators are summarized in Table 1 for the examples presented here.

These examples are meant to show important features and complexity in earthquake location results, not to compare different direct-search location methods or to compare direct-search to linearized algorithms. However, because linearized earthquake location has been and remains an important and widely used tool, we indicate for each example the location results obtained with a linearized algorithm, Hypoellipse [29]. Hypoellipse uses a least-squares, L2-norm and produces a 68% confidence ellipsoid for the hypocenter location. For well constrained locations this ellipsoid should closely match the *pdf* of our probabilistic, L2-norm locations; we plot the Hypoellipse ellipsoid for cases where it differs notably from the probabilistic location, L2-norm *pdf*.

Earthquake Location, Direct, Global-Search Methods, Table 1

Summary of results and quality indicators for the example locations. R_{pdf} is the radius of a sphere with volume V_{pdf}; l^1_{ell}, l^2_{ell}, l^3_{ell} are the half-lengths of the error ellipsoid axes; N_P is the number of phases used for the location; Δ_0 is the distance to the closest station

		Example	Lat (°)	Lon (°)	Depth (km)	*rms* (s)	*gap* (°)	Δ_0 (km)	N_P	V_{pdf} (km^3)	R_{pdf} (km)	l^1_{ell} (km)	l^2_{ell} (km)	l^3_{ell} (km)
	Ideal	**1a**	44.208	10.295	10.98	0.399	89	9.1	20	26	1.8	1.6	2.2	3.2
		1b	44.208	10.295	10.98	0.000	63	9.1	50	7	1.2	1.1	1.3	2.1
Station distribution	**Few stations**	**2a**	44.163	12.267	47.44	0.000	335	232	2	742001	82.3	30.9	53.9	223
		2b	44.172	9.565	77.60	0.001	192	25.8	4	21600	25.3	5.9	43.2	81.9
		2c	44.208	10.295	14.40	0.000	173	64.5	3	2011	7.8	4.8	7.3	20.2
		2d	44.208	10.295	10.98	0.000	99	9.1	8	66	2.5	2.1	2.7	4.9
	Side	**3**	44.207	10.296	11.03	0.004	251	29.0	19	444	4.7	3.4	4.7	11.0
	Far	**4a**	44.207	10.296	11.03	0.006	103	106	46	234	3.8	2.0	2.4	17.8
		4b	44.208	10.295	10.98	0.000	103	106	50	66	2.5	1.7	2.2	8.1
	Experimental design	**5a**	44.215	10.290	15.13	0.014	229	9.9	6	1806	13.4	8.3	13.7	59.1
		5b	44.208	10.295	7.83	0.007	89	36.3	6	381	6.6	3.1	3.7	11.5
Outlier	**L2-norm**	**6a**	44.207	10.296	11.03	0.007	135	9.00	10	172	3.5	2.9	3.7	6.7
		6b	44.120	10.267	9.02	0.813	156	10.5	10	172	3.5	3.0	4.2	9.0
	EDT	**6c**	44.221	10.305	9.94	0.006	132	9.4	10	167	3.4	4.5	7.7	15.2
		6d	44.215	10.304	10.02	0.540	133	9.0	10	275	4.0	10.7	30.8	42.9
	Early warning	**7a**	44.139	10.194	24.79	0.012	307	15.6	3	628090	53.1	18.7	81.6	104
		7b	44.210	10.302	10.57	0.015	250	8.8	4	33908	20.1	18.4	30.0	105
		7c	44.208	10.295	11.12	0.008	227	9.1	5	1704	7.4	5.3	13.9	30.2
		7d	44.207	10.296	11.03	0.007	135	9.0	10	172	3.5	2.9	3.7	6.7
Incorrect velocity model	**L2-norm**	**8a**	44.208	10.295	10.98	0.000	63	9.1	50	15	1.5	1.3	1.7	2.9
		8b	44.160	10.244	2.69	0.808	66	11.4	50	17	1.6	1.3	1.7	3.0
	EDT	**8c**	44.220	10.305	9.98	0.000	63	9.4	50	11	1.4	1.2	1.5	2.6
		8d	44.192	10.280	7.20	0.795	64	9.3	50	167	3.4	2.8	3.9	6.7

Example 1: An Ideal Location

To construct an ideal, reference location and synthetic data set for the 2007 Italian earthquake we first locate the event using the earliest 20 observed *P* or *S* arrival times (Fig. 3; Table 1; Example 1a). Next, we subtract the arrival residuals for this location from the corresponding times for all observations and relocate the event with the earliest 50 of these "corrected" times (Fig. 3; Table 1; Example 1b). This procedure results in an ideal, synthetic data set and a location problem that are equivalent to the case of no "a posteriori" picking error and no travel-time error (i. e., no velocity model error). For this problem the quality of the solution and the shape of the resulting location *pdf* reflect primarily the station geometry and corresponding ray take-off angles about the source.

The reference location (Fig. 3; Table 1; Example 1b) has $\text{rms} = 0\,\text{s}$, $\text{gap} = 63°$, $\Delta_0 \sim 9\,\text{km}$, $V_{\text{pdf}} \sim 7.0\,\text{km}^3$ and $\boldsymbol{l}_{\text{ell}} = 1.05$, 1.32 and 2.05 km. The *rms* is necessarily zero because we used residuals as time corrections, while the other indicators and the near-ellipsoidal form of the location *pdf* show a well constrained location. The location is well constrained by the data because stations are available at a wide range of distances and azimuths. In particular, the presence of a station nearly above the event, and of both *P* and *S*-wave arrival times for the closer stations, give good depth constraint.

Examples 2–5: Station Distribution

In the next examples we show locations for three cases with poor station distribution about the source:

1) few available stations;
2) stations all to one side of the event; and
3) no data for stations near or above the source. In addition we illustrate the application of experimental design techniques to improve the station distribution.

Example 2: Few Available Stations We first examine relocations of the 2007 Italian earthquake obtained with different numbers of *P* and *S* arrival times selected from the ideal, synthetic data set (Fig. 4; Table 1; Example 2a–d). With only two stations and 2 arrivals (2 *P* phases) the location *pdf* is a fat, near-vertical, planar surface with an elon-

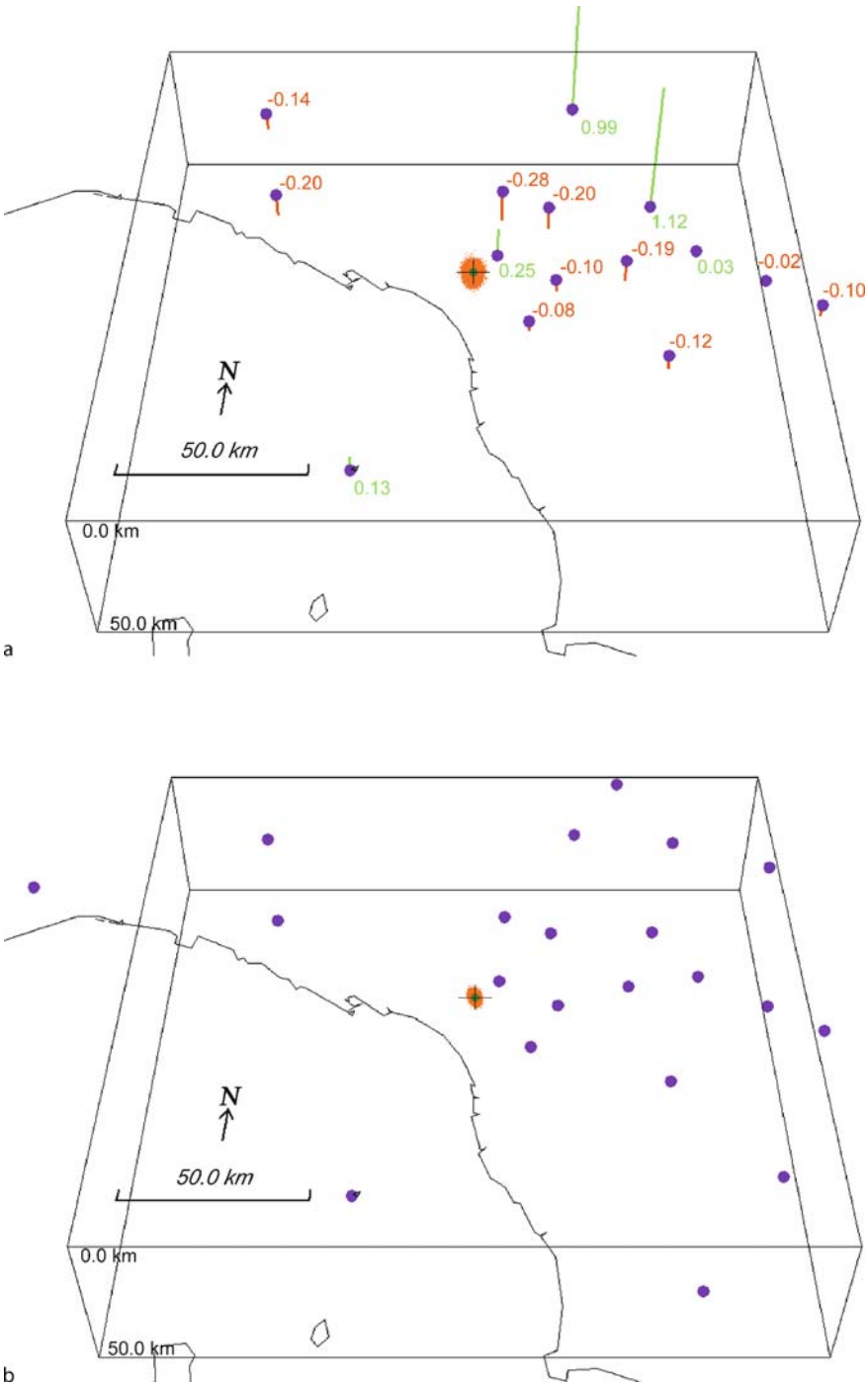

Earthquake Location, Direct, Global-Search Methods, Figure 3

Example 1: An ideal location. a Location obtained using the first 20, observed *P* or *S* arrival times; **b** location obtained using the first 50, *P* or *S* corrected arrival times from the ideal, synthetic data set. The elements shown in these and the following figures are: stations used for location (*blue dots*, in some cases stations fall outside the plotted region); location *pdf* (*red cloud of points* showing an importance sample drawn from the pdf); maximum likelihood hypocenter (*green dot*); ideal, synthetic location (*black cross*); *P* arrival residuals at each station: positive (*green, up-going bars*) and negative (*red, down-going bars*), numbers indicate residual value in sec. The Hypoellipse linearized locations and ellipsoids do not differ significantly from the direct-search locations shown in this figure

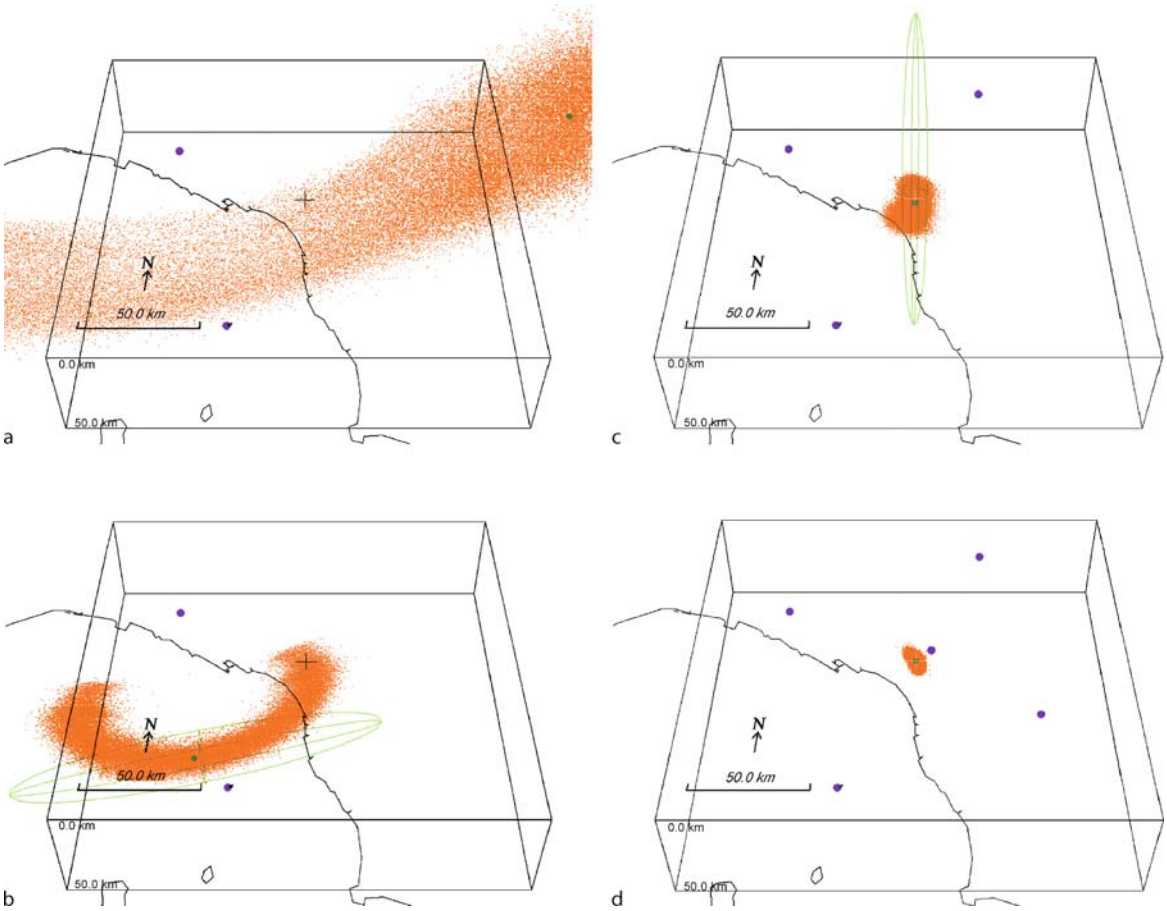

Earthquake Location, Direct, Global-Search Methods, Figure 4

Example 2: Few available stations. Locations obtained using progressively (**a**–**d**) a larger number of arrival observations. **a** 2P phases (2 stations); **b** 2P and 2 *S* phases (2 stations); **c** 3P phases (3 stations); **d** 5P and 3S phases (5 stations). For the locations in **a** and **b** the oct-tree search is performed to 100 km depth. In this and the following figures the 68% Hypoellipse ellipsoid is shown with *green lines*. Hypoellipse linearized location: does not converge for the location in panel **a**; ellipsoid differs markedly from the direct-search location *pdf* in panels **b** and **c**; and does not differ markedly from the direct-search location in panel **d**

gated, boomerang shape trending perpendicular to the line connecting the two receivers (Fig. 4a). With the addition of *S* arrivals from the same stations (4 arrivals – 2 *P* and 2 *S* phases) the location *pdf* is greatly reduced in volume, and has the form of an annulus oriented roughly perpendicular to the line connecting the two receivers (Fig. 4b). The annular form of this *pdf* results from the intersection of the boomerang shape *pdf* produced by the 2 *P* phases (Fig. 4a) and two hemispherical *pdf*'s centered on each station. Each of these hemispherical *pdf*'s would be produced by location using only the *P* and the *S* reading from either station; this is the probabilistic analogue to the method of circles using *S*-*P* times.

With three stations (3 arrivals – 3P phases) the location *pdf* forms one mass and its volume is further reduced. This location *pdf* retains an irregular, curved shape resulting from poor constraint of one spatial dimension that trades off with origin time (Fig. 4c). For all of these locations the problem is effectively underdetermined – the data cannot constrain all three hypocentral coordinates and origin time. In these cases a linearized location algorithm may not converge and would be unable to represent properly the effective location uncertainties. As more data are added, the location *pdf* progressively reduces in size and complexity, and with the addition of a station close to and above the source (8 arrivals – 5 *P* and 3 *S* phases), the location *pdf* has a compact, near ellipsoidal form indicating some constraint on all hypocentral coordinates and origin time (Fig. 4d). This location is similar to that obtained with the complete, ideal data set (Fig. 3b), though the location *pdf* remains much larger than that of the ideal case which has arrival times from many more stations.

Example 3: Stations to One Side of the Event – Large Gap Next, we examine the case of earthquakes occurring outside of the recording network with an example using *P* arrival times from stations only to the southeast of the earthquake (Fig. 5; Table 1; Example 3). The location *pdf* is large and elongated in a northwest-southeast direction oriented towards the centroid of the available stations because the lack of stations to the northwest (and use of *P* times only) allows a strong trade-off between potential hypocenter locations along this direction and origin time. In contrast, there is some constraint of the *pdf* to the northeast and southwest due to the aperture of the available stations. The poor station distribution and potential lack of constraint is clearly indicated by the large gap value for this location, gap $= 251°$. One or more good quality *S* readings can reduce the elongation of the *pdf*.

Example 4: Stations Far From the Event – Vertically Elongated PDF We next show an example where the nearest recording stations are far from the earthquake, relative to its depth, and either *P* arrival times only or both *P* and *S* arrival times are available (Fig. 6; Table 1; Examples 4a–b). With this station geometry the seismic rays leave the source region with approximately the same dip-direction to all stations. Consequently a change in source depth gives about the same change in predicted travel times to all stations. This change in travel time is indistinguishable from a change in origin time (c.f., Eqs. (1) or (2), leading to a strong trade-off between origin time and depth. Consequently the location *pdf* has a vertically elongated shape which, for the case of *P* arrivals only (Fig. 6a), extends throughout the entire search range in depth indicating no depth constraint. For a linearized location algorithm this location problem can be effectively underdetermined, though most linearized algorithms can fix the hypocenter depth artificially in order to obtain a stable epicentral location. The addition of *S* arrival times (Fig. 6b) improves the depth constraint to some extent, although the location *pdf* remains highly elongated in the vertical direction. The lack of close stations and potential lack of constraint is clearly indicated by the large Δ_0 value for this location, $\Delta_0 \approx 106$ km.

This case is common with sparse networks and with shallow sources. Reducing the vertical extent of the *pdf* requires stations at distances of the order of the source depth or less. The addition of one or more good quality *S* readings, especially at the closest stations, would further improve the depth constraint.

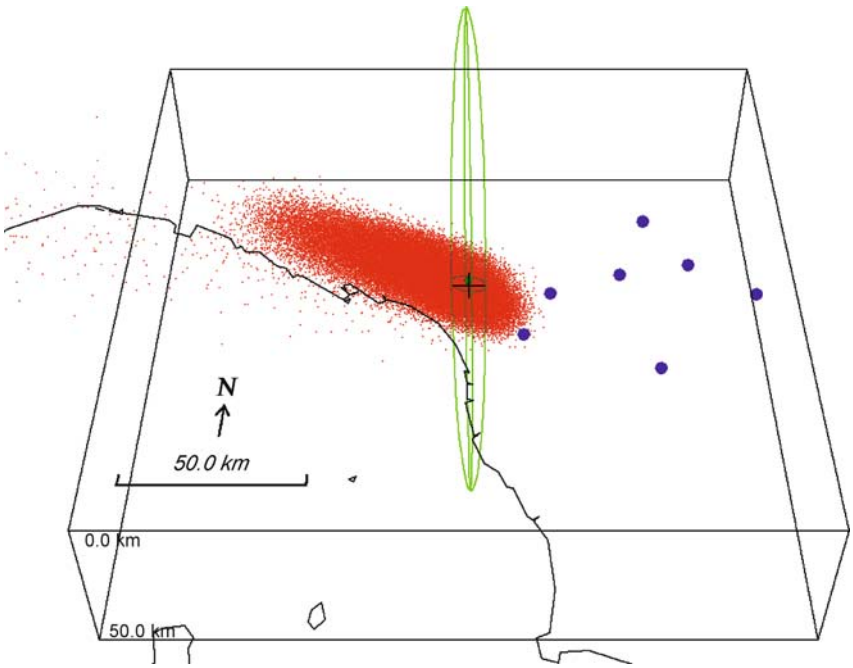

Earthquake Location, Direct, Global-Search Methods, Figure 5
Example 3: Stations to one side of the event. A location example with *P*-wave arrival times at 7 stations only to the southeast of the event. The Hypoellipse ellipsoid differs markedly from the direct-search location *pdf* in this figure

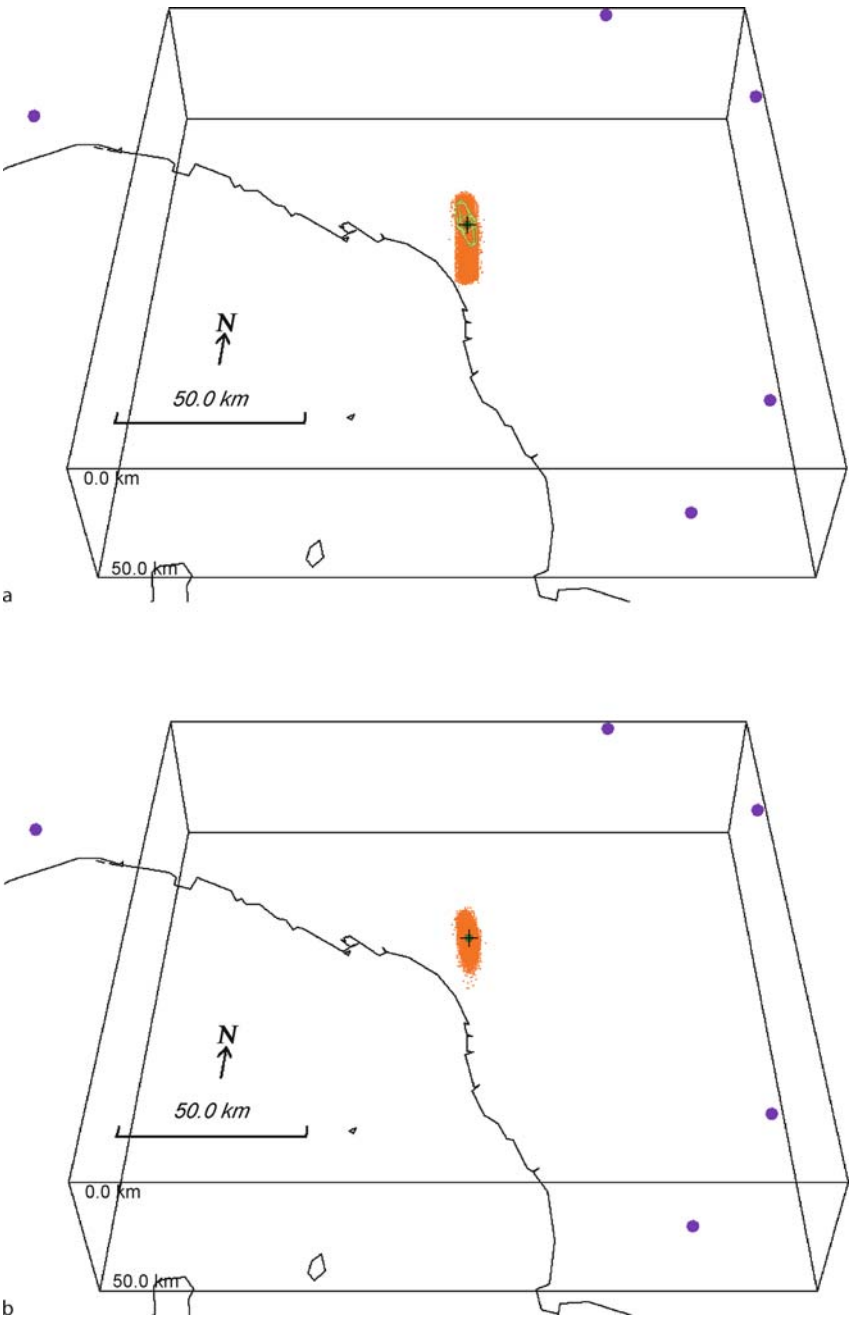

Earthquake Location, Direct, Global-Search Methods, Figure 6
Example 4: Stations far from the event. A location example using stations far from the epicenter, with **a** *P* arrival times only, **b** both *P* and *S* arrival times. Hypoellipse linearized location: ellipsoid differs markedly from the direct-search location *pdf* in panel **a**; and does not differ markedly from the direct-search location in panel **b**

Example 5: Stations Selection with Experimental Design
Next, we illustrate the application of experimental design techniques to station selection (Fig. 7; Table 1; Example 5). Considering a case similar to Example 3, which has 6 stations to one side of the event giving poor constraint on the location, we determine an optimal set of 6 stations to best constrain the location. To do this we apply a linearized design method [13] to select an optimal subset of 6

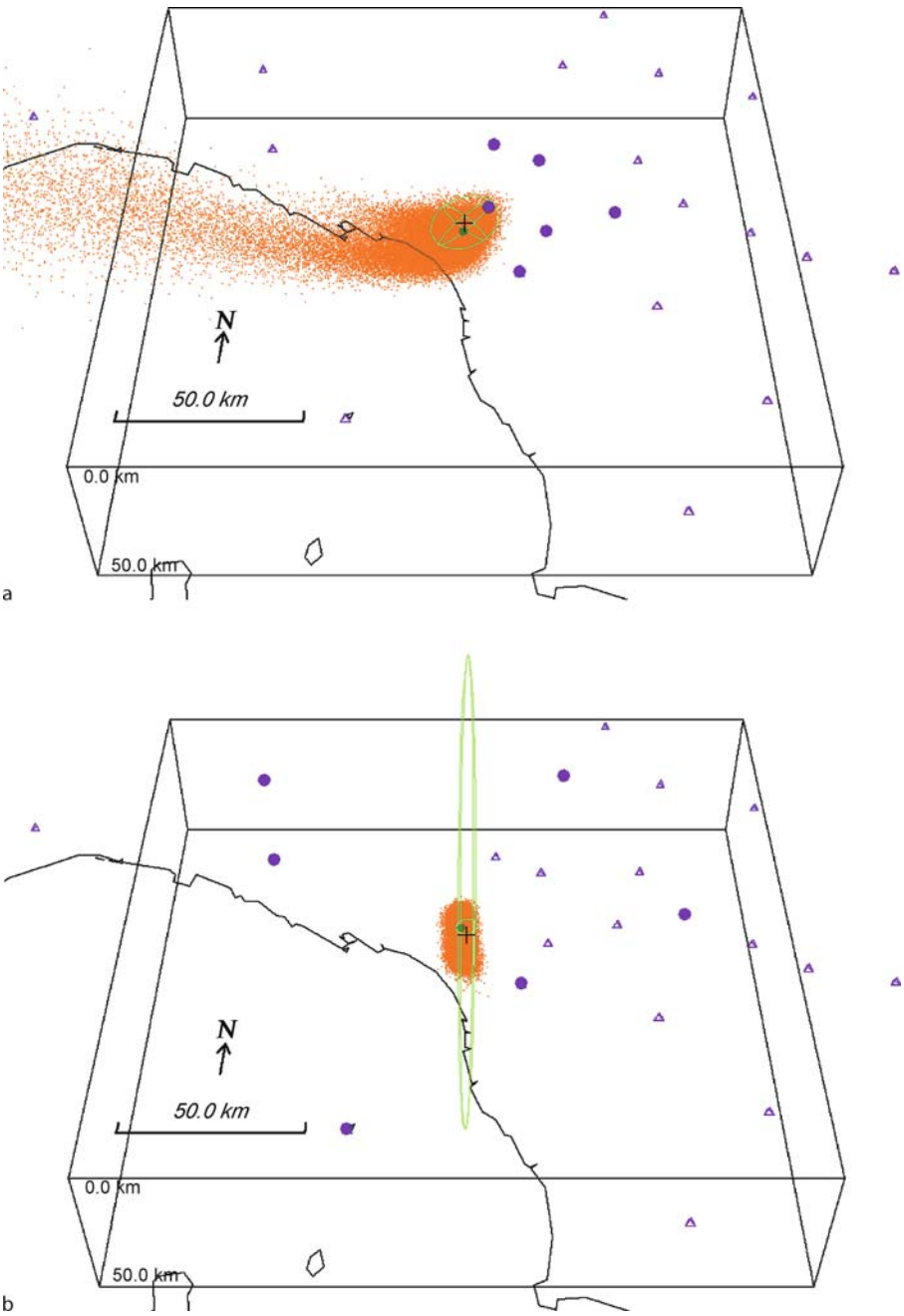

Earthquake Location, Direct, Global-Search Methods, Figure 7

Example 5: Stations selection with experimental design. A location example showing a Location using the stations with the first 6 available arrival times, b location using an optimal set of 6 stations as determined with a linearized experimental-design method. Available stations not used or selected are shown with *open triangle symbols*. Hypoellipse linearized location: ellipsoids differ markedly from the direct-search location *pdf*'s in panels a and b

of the available INSN stations to best constrain an event at the (known) location produced by the ideal, synthetic data (Example 1b).

The design procedure does not simply select the 6 closest receivers to the source (i. e. first 6 available arrival times, Fig. 7a), but instead selects receivers distributed

around, and to a large distance away from the source (Fig. 7b). This choice can be understood as balancing the distribution of directions (azimuth and inclination) that the rays leave the source to the selected receivers, a direct result of the use of the linearized approximations to the model-data relationship Eq. (10) in the linearized design method [13]. This method is based on selecting stations based on the similarity between the rows of the location kernel matrix of the linearized problem; the approach does not differ significantly from that of Uhrhammer [76] based on the condition number of the same matrix. The improvement in station distribution in azimuth is indicated by the small gap value for this location, gap = 89°. The resulting location *pdf* (Fig. 7b) is compact and symmetric relative to the location *pdf* obtained from the first 6 stations recording the *P* phases (Fig. 7a), and the maximum likelihood hypocenter is close to the ideal location hypocenter.

Example 6: Incorrect Picks and Phase Identification – Outlier Data

For a given hypocenter location, an outlier arrival time has a residual that is much greater than its nominal error. Data outliers are common with automatic phase arrival picking algorithms, with *S* arrival picks, for small events, distant stations, or other cases where the signal to noise ratio is low, and for early instrumental data where large timing errors are common. In many cases, such as automatic earthquake monitoring and early warning systems, it is important to have robust location procedures that are influenced as little as possible by the presence of outliers. One way to achieve this is to use robust likelihood functions such as EDT Eq. (8). In the example below, we compare the performance of EDT and the more commonly used L2-norm likelihood functions.

This example uses only stations near the source, and arrival times from ideal, synthetic data sets for both the L2-norm and the EDT likelihood functions. We add 3 s to the *P* arrival time at two stations to generate outlier data, and examine L2-norm and EDT locations without and with the outlier data (Fig. 8; Table 1; Examples 6a–d). The L2-norm location with the outlier data (Fig. 8b) does not identify and isolate the two outlier *P*-arrivals but instead mixes information from these arrivals with the other data resulting in relatively large, non-zero residuals for all arrivals. This results in a bias of about 10 km in the maximum likelihood hypocenter location relative to the ideal location hypocenter, while the location *pdf* for the L2-norm locations with and without outlier data have about the same size and form, but have little overlap (Figs. 8a and 8b). Thus the L2-norm solution gives no clear indication of the presence of outlier data, or that the solution may be biased. In contrast, the EDT location for the data set containing the outliers (Fig. 8d) correctly identifies the two outlier arrivals (the EDT residuals for these two outlier data are both about 2.9 s) and strongly downweights them (from 1.2 to 0.17 posterior weight), while producing small residuals (< 0.08 s) for the remaining arrival, as would be the case without outlier data. The maximum likelihood hypocenters for the EDT locations with and without outlier data are almost identical, but the location *pdf*'s are very different (Fig. 8c and 8d). With outlier data, the *pdf* has an irregular shape and several distinct parts, reflecting the inconsistency of the data set to constrain a unique event location. For the outlier locations, a potential problem with the data set is indicated by the large *rms* values with both L2-norm and EDT, and with EDT alone, by the asymmetry in residuals, the irregular *pdf* shape, and the large V_{pdf} and $\boldsymbol{l}_{\mathrm{ell}}$ values.

This result shows that location in the presence of outlier data can be remarkably stable with the EDT likelihood function, which is easy to implement with direct-search location techniques. In contrast, the same location with the commonly used, L2-norm likelihood function is biased, while presenting few indicators of this bias.

Example 7: Earthquake Early-Warning Scenario

Location for earthquake early-warning must be performed rapidly and in an evolutionary manner starting with the first available phase arrivals. In this example we examine the ability of direct-search location to obtain robust and useful location information using *P* arrivals from the first stations that record the Northern Italian event (Fig. 9; Table 1; Examples 7a–d).

Within about 6 seconds after the origin time, t_0, three *P* reading are available. Location with these readings produces an extensive location *pdf* that fills the southwest quadrant of the search region (Fig. 9a); this *pdf* does not provide useful constraint on the location, but is robust in that it includes the true location. Progressive addition of more arrival time data (Fig. 9b and 9c) reduces the size of the location *pdf*. With 5 arrivals, at about 7 s after t_0 (Fig. 9c), the maximum likelihood location is close to that of the ideal, synthetic location and the location *pdf* is well delimited, although elongated towards the west because no arrivals are yet available from stations in that direction. By 13 s after t_0 (Fig. 9d), 10 *P* arrivals are available and the location *pdf* is now compact and symmetrical, primarily because a station to the northwest is included. This *pdf* has small enough V_{pdf} and $\boldsymbol{l}_{\mathrm{ell}}$ values to provide useful, prob-

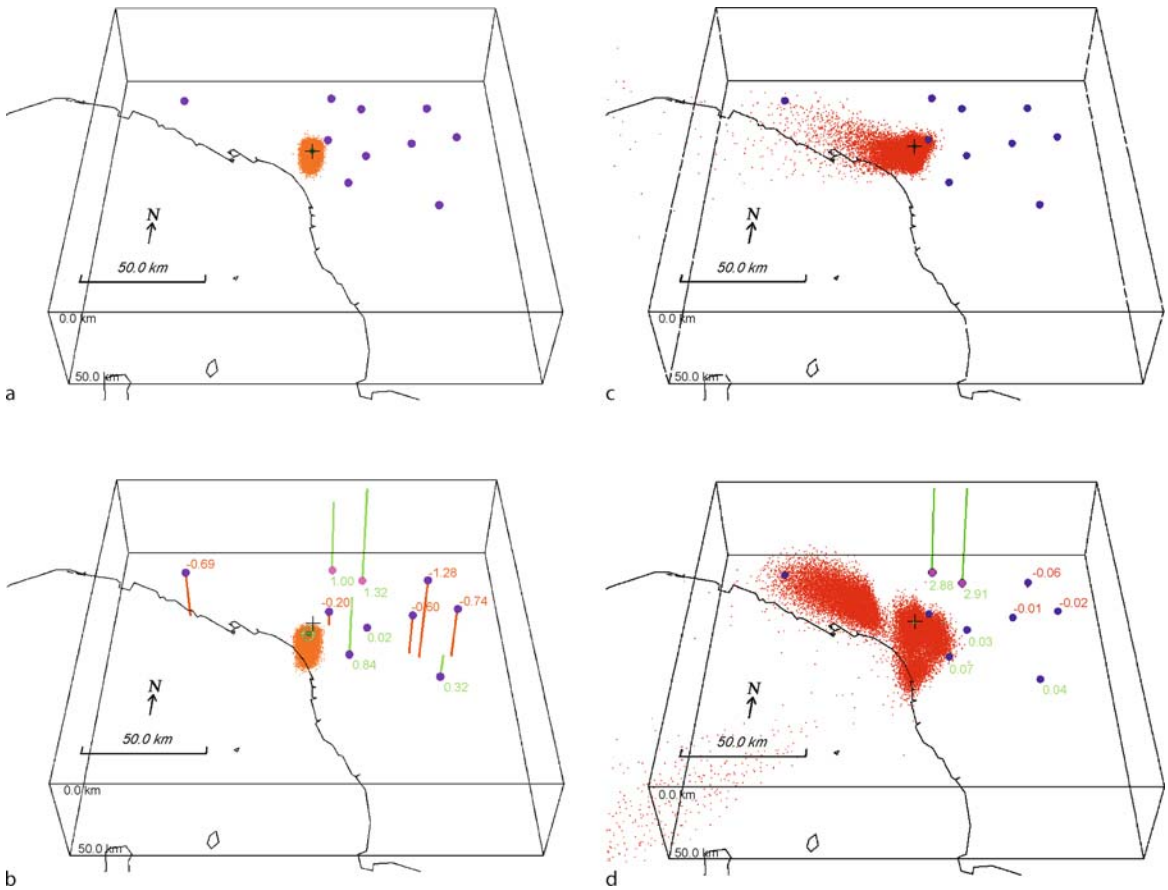

Earthquake Location, Direct, Global-Search Methods, Figure 8

Example 6: Incorrect picks and phase identification – outlier data. Locations using ten *P*-wave arrival times with L2-norm and a no outliers, b two arrival-time outliers, and with EDT and c no outliers, d two arrival-time outliers. The stations with outlier arrivals are shown with *violet dots*. Note the small *pdf* of L2-norm regardless of the outliers and, in contrast, the ability of EDT to detect the outliers (see text). The Hypoellipse ellipsoid differs markedly from the direct-search location *pdf* in panel b. Hypoellipse not compared to EDT locations in panels c and d

abilistic constraint on the location for early-warning purposes at a regional scale, while the maximum likelihood hypocenter is effectively the same as that of the ideal location. In practical application, direct-search location results similar to those illustrated here can be obtained within a delay of less than 1 sec after the readings are available (e. g. [64]).

Example 8: Incorrect Velocity Model

Any velocity model used for earthquake location is an approximation to the true Earth and thus will in general produce erroneous predicted travel times. The magnitude of error in the travel times depends on many factors, but will in general be larger for more distant stations and with increased complexity in the true Earth structure. We examine the effect of an incorrect velocity models by repeating the ideal location (Example 1a and b) with and without the "corrected" times, and using 50 *P* arrivals (the ideal location was determined using the first 20 *P* or *S* arrivals). We examine locations using the L2-norm and EDT likelihood functions (Fig. 10; Table 1; Examples 8a–d).

The locations with time corrections (Fig. 10a and 10c) simulate the unrealizable case of perfect knowledge of the velocity structure. With both the L2-norm and EDT the location results show zero residuals, compact location *pdf*'s and a maximum likelihood hypocenter that necessarily matches exactly the corresponding ideal location. We note, however, that the L2-norm and EDT "ideal" locations differ slightly because they are derived from noisy, real data, and they use different likelihood functions.

The locations without time corrections (Fig. 10b and 10d) use the true observed data (i. e., travel times through the true Earth) and thus show the effect of an in-

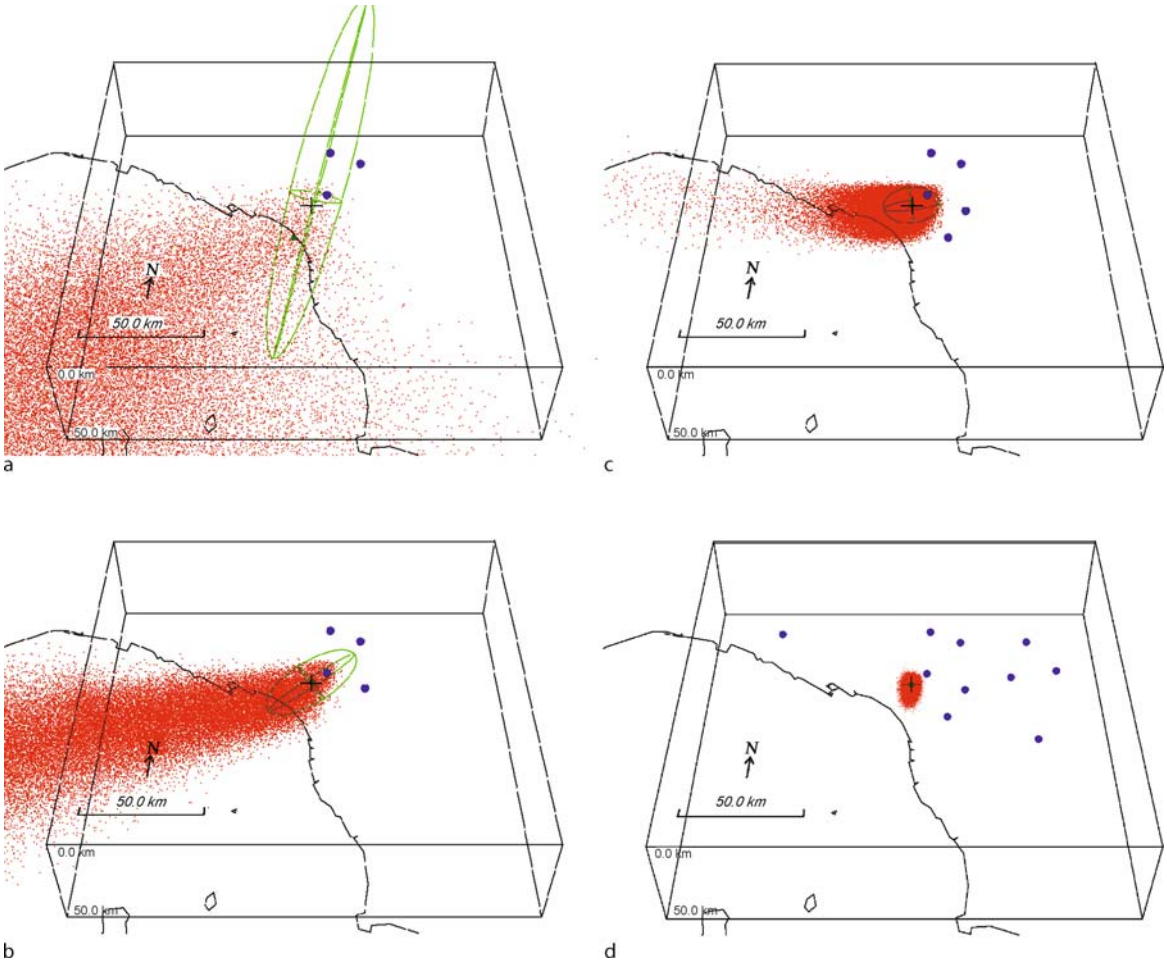

Earthquake Location, Direct, Global-Search Methods, Figure 9

Example 7: Earthquake early-warning scenario. Progressive location using a 3, b 4, c 5 and d 10 stations. Hypoellipse linearized location: ellipsoid differs markedly from the direct-search location *pdf*'s in panel a, b and c; and does not differ markedly from the direct-search location in panel d

correct velocity model (i. e., the 1-D velocity model used for location). This is shown by the pattern of positive and negative residuals obtained with both the L2-norm and EDT. The L2-norm location without time corrections has a balanced distribution of positive and negative residuals and, relative to the L2-norm location with corrections, a similar size location *pdf* and a biased maximum likelihood hypocenter. In contrast, the EDT location without corrections has more positive than negative residuals and, relative to the EDT location with corrections, a larger location *pdf* and nearly identical, unbiased maximum likelihood hypocenter. For these locations, a potential problem with the velocity model is indicated by the large residuals and *rms* values with both L2-norm and EDT, and, with EDT, by the asymmetry in residuals, the irregular *pdf* shape, and the large V_{pdf} and $\boldsymbol{l}_{\mathrm{ell}}$ values, as with the outlier data example (Example 6).

In effect, locations with an incorrect velocity model and with outlier data are mathematically similar, though in the former case all or most residuals may be large while in the latter case only a few will be large. It is difficult to distinguish between the two cases with the L2-norm because this algorithm seeks to best satisfy *all* of the observations simultaneously (cf., Eq. (7)) by balancing the distribution of positive and negative residual (cf., Fig. 8b and Fig. 10b). Thus, relative to the residuals corresponding to the correct location, the L2-norm solution damps and hides larger residuals at the expense of increasing small residuals. In contrast, EDT seeks to best satisfy the *most* pairs of observations (cf., Eq. (8)) and imposes no inher-

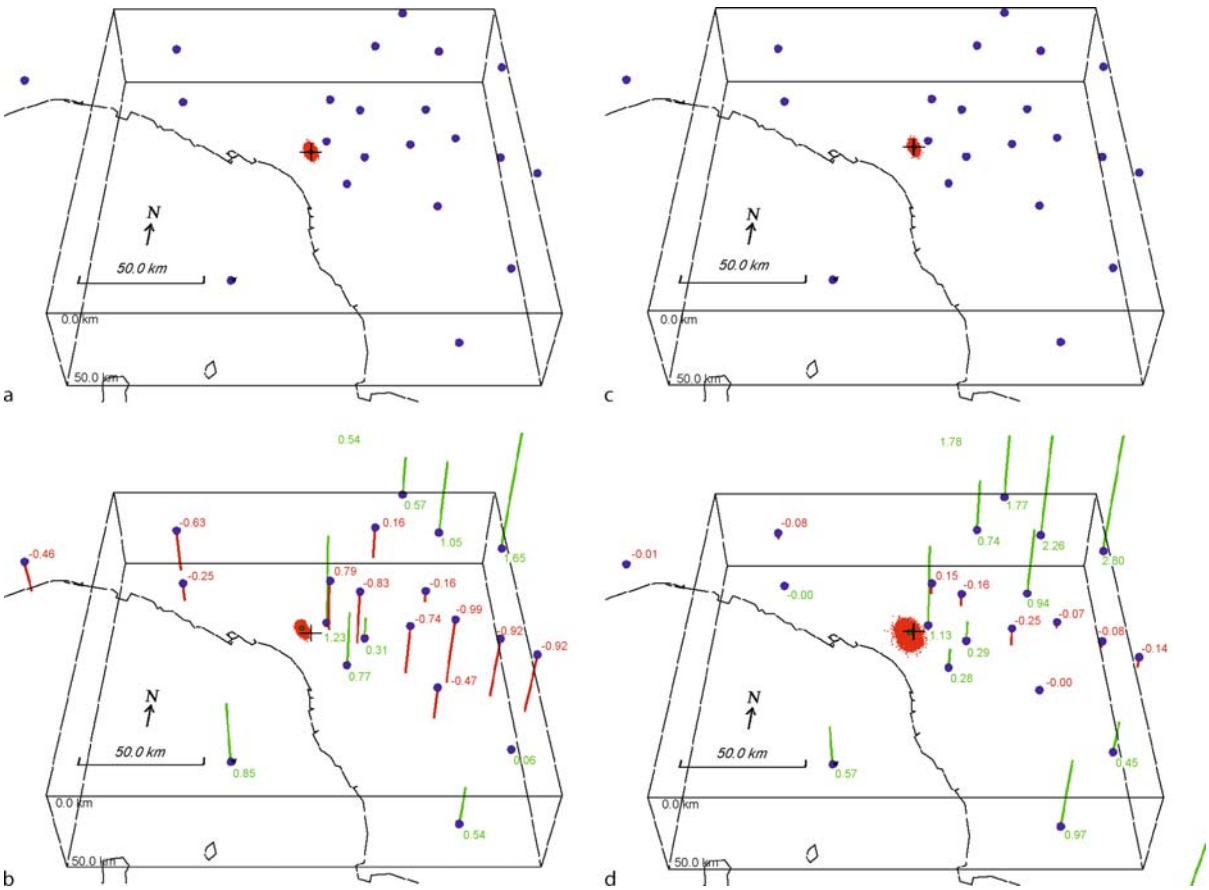

Earthquake Location, Direct, Global-Search Methods, Figure 10

Example 8: Incorrect velocity model. Locations using 50 *P* arrivals with the L2-norm and **a** time corrections, **b** no time corrections, and with EDT and **c** time corrections, **d** no time corrections. The locations without the ideal time corrections show the effect of an incorrect velocity model. The Hypoellipse linearized locations and ellipsoids do not differ markedly from the direct-search locations shown in this figure

ent constraint on the distribution of residuals. Thus with EDT the difference in number, magnitude and distribution of large residuals – few and large for the outlier case, many of similar magnitude and spatially correlated for the incorrect velocity model case – allows one, in principle, to distinguish between the two cases (cf., Fig. 8d and Fig. 10d). In addition, the size and complexity of the location *pdf*'s generally increases more rapidly with EDT than with the L2-norm as the solution quality decreases. Thus, with both the outlier and incorrect velocity model cases, the location results with the EDT likelihood function are more informative than with the L2-norm. However, location with the EDT likelihood function can become unstable (e. g. define only a local maximum of the *pdf*) for cases where the outlier data or velocity model errors lead to extreme complexity in the topology of the EDT location *pdf*.

Future Directions

There are various ways that direct, global-search location methodologies may evolve in the future. For example, the stability and completeness of the location and location *pdf* could be improved with the use of more complete data uncertainties, expressed as a *pdf*. These *pdf*'s may typically be irregular and asymmetric, and difficult to determine and parametrize. Currently, enumerated quality indications or, at best, simple normal distributions (describing Gaussian uncertainty) are used to describe the picking error.

Similarly, we have shown that earthquake location depends inherently on the velocity model adopted, but that no realistic uncertainties are associated with this model. Differences between the velocity model and the true Earth can result in complicated differences in ray-paths and

travel-times, which will depend strongly on the source and receiver positions. These complications, combined with the lack of knowledge about the true Earth, makes estimating true travel time uncertainties effectively impossible. However, it can be assumed that changes become progressively larger with increasing ray-path length. This effect could be accounted for approximately by travel-time uncertainties that increase with the ray-length or travel time. Instead of using a velocity model to generate travel times, another approach is to derive the required times from tables of empirically determined or corrected travel times (e. g. [41,43]). With this approach the travel-time uncertainties are estimated from timing information, with little or no direct use of velocity structures or ray paths.

We have described and illustrated the importance of the source-receiver geometry for locating earthquakes, notably with regards to constraining a compact and symmetric location *pdf*. Thus, improved constraint on event locations can be achieved through prior use of survey design techniques to select station sites. In a related manner, after an event occurs, these techniques could be employed dynamically to weight the available arrival times used for location with respect to the geometry of the available stations around the likely source region.

The demand for rapid, real-time location and earthquake early warning requires improvements in the integration, speed, quality and robustness of the phase arrival picking, phase association and event location procedures. Currently, development is progressing on integrated procedures which are evolutionary and probabilistic, using, for example, robust likelihood functions such as EDT and information from not-yet-triggered stations (e. g., [8,21,54,63,64]).

A current problem in direct-search location is how to describe in a standardized and compact way the sometimes topologically-complex location *pdf*. For example, such a description is needed if the *pdf* is to be included in standard earthquake catalogs and for rapid dissemination of probabilistic location information for earthquake early-warning. More generally, making full use of the extensive information in direct-search location solutions will require new methods and procedures to store, distribute and analyze the location *pdf*, maximum likelihood hypocenter, arrival residuals and weights, and other statistics and quality indicators of the solutions.

The continuing increase in computer speed will allow application of direct-search inversion methods to relative location of ensembles of events and for joint epicentral determination in the near future. The use of these methods will be important to explore more completely the vast solution space and better determine the error and resolution for such high-dimensional inverse problems.

The continuing increase in computer speed will also make practical earthquake location techniques using waveform recordings directly, without the intermediate stage of extracting phase arrival times. In these techniques, continuous waveform data streams are matched to synthetic Green's functions within a global-search over possible source locations and source parameters. This type of approach is used to locate previously unidentified earthquakes using low amplitude surface waves on off-line, continuous, broadband waveforms [15,68], and for automatic, real-time estimation of moment tensors and location from continuous broadband data streams (e. g., [25]). Waveform methods will likely be applied to earthquake location on local and regional scales as faster computers and more accurate 3D velocity models become available [81]; related applications using simple ray or acoustic theories to generate the Green's functions show promising results (e. g., [3]).

Bibliography

Primary Literature

1. Aki K, Richards PG (1980) Quantitative Seismology. Freeman, New York
2. Anderson K (1981) Epicentral location using arrival time order. Bull Seism Soc Am 71:541–545
3. Baker T, Granat R, Clayton RW (2005) Real-time Earthquake Location Using Kirchhoff Reconstruction. Bull Seism Soc Am 95:699–707
4. Billings SD (1994) Simulated annealing for earthquake location. Geophys J Int 118:680–692
5. Buland R (1976) The mechanics of locating earthquakes. Bull Seism Soc Am 66:173–187
6. Calvert A, Gomez F, Seber D, Barazangi M, Jabour N, Ibenbrahim A, Demnati A (1997) An integrated geophysical investigation of recent seismicity in the Al-Hoceima region of North Morocco. Bull Seism Soc Am 87:637–651
7. Červený V (2001) Seismic Ray Theory. Cambridge University Press, Cambridge
8. Cua G, Heaton T (2007) The Virtual Seismologist (VS) Method: a Bayesian Approach to Earthquake Early Warning. In: Gasparini P, Gaetano M, Jochen Z (eds) Earthquake Early Warning Systems. Springer, Berlin
9. Curtis A (1999) Optimal experiment design: Cross-borehole tomographic examples. Geophys J Int 136:637–650
10. Curtis A (1999) Optimal design of focussed experiments and surveys. Geophys J Int 139:205–215
11. Curtis A (2004) Theory of model-based geophysical survey and experimental design Part A – Linear Problems. Lead Edge 23(10):997–1004
12. Curtis A (2004) Theory of model-based geophysical survey and experimental design Part B – Nonlinear Problems. Lead Edge 23(10):1112–1117

13. Curtis A, Michelini A, Leslie D, Lomax A (2004) A deterministic algorithm for experimental design applied to tomographic and microseismic monitoring surveys. Geophys J Int 157:595–606
14. Dreger D, Uhrhammer R, Pasyanos M, Frank J, Romanowicz B (1998) Regional an far-regional earthqauke locations and source parameters using sparse broadband networks: A test on the Ridgecrest sequence. Bull Seism Soc Am 88:1353–1362
15. Ekström G (2006) Global Detection and Location of Seismic Sources by Using Surface Waves. Bull Seism Soc Am 96:1201–1212. doi:10.1785/0120050175
16. Font Y, Kao H, Lallemand S, Liu CS, Chiao LY (2004) Hypocentral determination offshore Eastern Taiwan using the Maximum Intersection method. Geophys J Int 158:655–675
17. Geiger L (1912) Probability method for the determination of earthquake epicenters from the arrival time only (translated from Geiger's 1910 German article). Bull St Louis Univ 8:56–71
18. Gentili S, Michelini A (2006) Automatic picking of P and S phases using a neural tree. J Seism 10:39–63. doi:10.1007/s10950-006-2296-6
19. Goldberg DE (1989) Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading
20. Hammersley JM, Handscomb DC (1967) Monte Carlo Methods. Methuen, London
21. Horiuchi S, Negishi H, Abe K, Kamimura A, Fujinawa Y (2005) An Automatic Processing System for Broadcasting Earthquake Alarms. Bull Seism Soc Am 95:708–718
22. Husen S, Smith RB (2004) Probabilistic Earthquake Relocation in Three-Dimensional Velocity Models for the Yellowstone National Park Region, Wyoming. Bull Seism Soc Am 94:880–896
23. Husen S, Kissling E, Deichmann N, Wiemer S, Giardini D, Baer M (2003) Probabilistic earthquake location in complex three-dimensional velocity models: Application to Switzerland. J Geophys Res 108:2077–2102
24. Johnson CE, Lindh A, Hirshorn B (1994) Robust regional phase association. US Geol Surv Open-File Rep pp 94–621
25. Kawakatsu H (1998) On the real-time monitoring of the long-period seismic wavefield. Bull Earthq Res Inst 73:267–274
26. Kennett BLN (1992) Locating oceanic earthquakes – the influence of regional models and location criteria. Geophys J Int 108:848–854
27. Kennett BLN (2006) Non-linear methods for event location in a global context. Phys Earth Planet Inter 158:46–54
28. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220:671–680
29. Lahr JC (1999) HYPOELLIPSE: A Computer Program for Determining Local Earthquake Hypocentral Parameters, Magnitude, and First-Motion Pattern (Y2K Compliant Version) 1999 Version 1.0. US Geological Survey Open-File Report, pp 99–23. http://jclahr.com/science/software/hypoellipse/hypoel/hypoman/hypomst_pdf.pdf
30. Lepage GP (1978) A new algorithm for adaptive multidimensional integration. J Comput Phys 27:192–203
31. Lomax A (2005) A Reanalysis of the Hypocentral Location and Related Observations for the Great 1906 California Earthquake. Bull Seism Soc Am 91:861–877
32. Lomax A (2008) Location of the Focus and Tectonics of the Focal Region of the California Earthquake of 18 April 1906. Bull Seism Soc Am 98:846–860
33. Lomax A, Curtis A (2001) Fast, probabilistic earthquake location in 3D models using oct-tree importance sampling. Geophys Res Abstr 3:955. www.alomax.net/nlloc/octtree
34. Lomax A, Virieux J, Volant P, Berge C (2000) Probabilistic earthquake location in 3D and layered models: Introduction of a Metropolis-Gibbs method and comparison with linear locations. In: Thurber CH, Rabinowitz N (eds) Advances in Seismic Event Location. Kluwer, Amsterdam
35. Lomax A, Zollo A, Capuano P, Virieux J (2001) Precise, absolute earthquake location under Somma-Vesuvius volcano using a new 3D velocity model. Geophys J Int 146:313–331
36. Maurer HR, Boerner DE (1998) Optimized and robust experimental design. Geoph J Int 132:458–468
37. Milne J (1886) Earthquakes and Other Earth Movements. Appelton, New York
38. Moser TJ, van Eck T, Nolet G (1992) Hypocenter determination in strongly heterogeneous earth models using the shortest path method. J Geophys Res 97:6563–6572
39. Moser TJ, Nolet G, Snieder R (1992) Ray bending revisited. Bull Seism Soc Am 82:259–288
40. Mosegaard K, Tarantola A (1995) Monte Carlo sampling of solutions to inverse problems. J Geophys Res 100:12431–12447
41. Myers SC, Schultz CA (2000) Improving Sparse Network Seismic Location with Bayesian Kriging and Teleseismically Constrained Calibration Events. Bull Seism Soc Am 90:199–211
42. Nicholson T, Gudmundsson Ó, Sambridge M (2004) Constraints on earthquake epicentres independent of seismic velocity models. Geophys J Int 156:648–654
43. Nicholson T, Sambridge M, Gudmundsson Ó (2004) Three-dimensional empirical traveltimes: construction and applications. Geophys J Int 156:307–328
44. Podvin P, Lecomte I (1991) Finite difference computations of traveltimes in very contrasted velocity models: a massively parallel approach and its associated tools. Geophys J Int 105:271–284
45. Press WH, Flannery BP, Saul AT, Vetterling WT (1992) Numerical Recipes, 2nd edn. Cambridge Univ Press, New York
46. Presti D, Troise C, De Natale G (2004) Probabilistic Location of Seismic Sequences in Heterogeneous Media. Bull Seism Soc Am 94:2239–2253
47. Pujol J (2000) Joint event location – The JHD technique and applications to data from local seismic networks. In: Thurber CH, Rabinowitz N (eds) Advances in Seismic Event Location. Kluwer, Amsterdam
48. Rabinowitz N (2000) Hypocenter location using a constrained nonlinear simplex minimization method. In: Thurber CH, Rabinowitz N (eds) Advances in Seismic Event Location. Kluwer, Amsterdam
49. Rabinowitz N, Steinberg DM (2000) A statistical outlook on the problem of seismic network configuration. In: Thurber CH, Rabinowitz N (eds) Advances in Seismic Event Location. Kluwer, Amsterdam
50. Rawlinson N, Sambridge M (2004) Wave front evolution in strongly heterogeneous layered media using the fast marching method. Geophys J Int 156:631–647
51. Rawlinson N, Sambridge M (2004) Multiple reflection and transmission phases in complex layered media using a multistage fast marching method. Geophys 69:1338–1350
52. Reid HF (1910) The Mechanics of the Earthquake. Vol II of: The California Earthquake of 18 April 1906. Report of the State Earthquake Investigation Commission, Lawson AC

(Chairman). Carnegie Institution of Washington Publication, vol 87 (reprinted 1969)
53. Rothman DH (1985) Nonlinear inversion, statistical mechanics, and residual statics estimation. Geophysics 50:2784–2796
54. Rydelek P, Pujol J (2004) Real-Time Seismic Warning with a Two-Station Subarray. Bull Seism Soc Am 94:1546–1550
55. Sambridge M (1998) Exploring multi-dimensional landscapes without a map. Inverse Probl 14:427–440
56. Sambridge M (1999) Geophysical inversion with a Neighbourhood algorithm, vol I. Searching a parameter space. Geophys J Int 138:479–494
57. Sambridge M (1999) Geophysical inversion with a neighbourhood algorithm, vol II. Appraising the ensemble. Geophys J Int 138:727–746
58. Sambridge M (2003) Nonlinear inversion by direct search using the neighbourhood algorithm. In: International Handbook of Earthquake and Engineering Seismology, vol 81B. Academic Press, Amsterdam, pp 1635–1637
59. Sambridge M, Drijkoningen G (1992) Genetic algorithms in seismic waveform inversion. Geophys J Int 109:323–342
60. Sambridge M, Gallagher K (1993) Earthquake hypocenter location using genetic algorithms. Bull Seism Soc Am 83:1467–1491
61. Sambridge M, Kennett BLN (1986) A novel method of hypocentre location. Geophys J R Astron Soc 87:679–697
62. Sambridge M, Mosegaard K (2002) Monte Carlo Methods In Geophysical Inverse Problems. Rev Geophys 40:1009–1038
63. Satriano C, Lomax A, Zollo A (2007) Optimal, Real-time Earthquake Location for Early Warning. In: Gasparini P, Gaetano M, Jochen Z (eds) Earthquake Early Warning Systems. Springer, Berlin
64. Satriano C, Lomax A, Zollo A (2007) Real-time evolutionary earthquake location for seismic early warning. Bull Seism Soc Am 98:1482–1494
65. Sen M, Stoffa PL (1995) Global optimization methods in geophysical inversion. Elsevier, Amsterdam, p 281
66. Sethian JA (1999) Level set methods and fast marching methods. Cambridge University Press, Cambridge
67. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423
68. Shearer PM (1994) Global seismic event detection using a matched filter on long-period seismograms. J Geophys Res 99:13,713–13,735
69. Shearer PM (1997) Improving local earthquake locations using the L1 norm and waveform cross correlation: Application to the Whittier Narrows, California, aftershock sequence. J Geophys Res 102:8269–8283
70. Steinberg DM, Rabinowitz N, Shimshoni Y, Mizrachi D (1995) Configuring a seismographic network for optimal monitoring of fault lines and multiple sources. Bull Seism Soc Am 85:1847–1857
71. Stummer P, Maurer HR, Green AG (2004) Experimental Design: Electrical resistivity data sets that provide optimum subsurface information. Geophysics 69:120–139
72. Tarantola A (1987) Inverse problem theory: Methods for data fitting and model parameter estimation. Elsevier, Amsterdam
73. Tarantola A (2005) Inverse Problem Theory and Methods for Model Parameter Estimation. SIAM, Philadephia
74. Tarantola A, Valette B (1982) Inverse problems = quest for information. J Geophys Res 50:159–170
75. Thurber CH, Kissling E (2000) Advances in travel-time calculations for three-dimensional strucutres. In: Thurber CH, Rabinowitz N (eds) Advances in Seismic Event Location. Kluwer, Amsterdam
76. Uhrhammer RA (1980) Analysis of small seismographic station networks. Bull Seism Soc Am 70:1369–1379
77. Um J, Thurber C (1987) A fast algorithm for two-point seismic ray tracing. Bull Seism Soc Am 77:972–986
78. van den Berg J, Curtis A, Trampert J (2003) Bayesian, nonlinear experimental design applied to simple, geophysical examples. Geophys J Int 55(2):411–421. Erratum: 2005. Geophys J Int 161(2):265
79. Vidale JE (1988) Finite-difference calculation of travel times. Bull Seism Soc Am 78:2062–2078
80. Winterfors E, Curtis A (2007) Survey and experimental design for nonlinear problems. Inverse Problems (submitted)
81. Wither M, Aster R, Young C (1999) An automated local and regional seismic event detection and location system using waveform correlation. Bull Seism Soc Am 8:657–669
82. Withers M, Aster R, Young C, Beiriger J, Harris M, Moore S, Trujillo J (1998) A comparison of select trigger algorithms for automated global seismic phase and event detection. Bull Seism Soc Am 88:95–106
83. Wittlinger G, Herquel G, Nakache T (1993) Earthquake location in strongly heterogeneous media. Geophys J Int 115:759–777
84. Zhou H (1994) Rapid 3-D hypocentral determination using a master station method. J Geophys Res 99:15439–15455

Books and Reviews

Gasparini P, Gaetano M, Jochen Z (eds) (2007) Earthquake Early Warning Systems. Springer, Berlin
Lee WHK, Stewart SW (1981) Principles and applications of microearthquake networks. Academic Press, New York
Thurber CH, Rabinowitz N (eds) (2000) Advances in Seismic Event Location. Kluwer, Amsterdam

Earthquake Magnitude

PETER BORMANN, JOACHIM SAUL
GeoForschungsZentrum Potsdam, Potsdam, Germany

Article Outline

Glossary
Definition of the Subject
Introduction to Common Magnitude Scales: Potential and Limitations
Common Magnitude Estimates for the Sumatra 2004 M_w 9.3 Earthquake
Magnitude Saturation and Biases Due to Earthquake Complexity
Proposals for Faster Magnitude Estimates of Strong Earthquakes
Future Requirements and Developments
Bibliography

Glossary

Technical terms that are written in the text in italics are explained in the Glossary.

Corner frequency The frequency f_c at which the curve that represents the Fourier amplitude spectrum of a recorded seismic signal abruptly changes its slope (see Fig. 5). For earthquakes, this frequency is related to the fault size, rupture velocity, rupture duration and stress drop at the source. Also the frequency at which the magnification curve of a recording system (e. g., Fig. 3) changes its slope.

Dispersion Frequency-dependence of the wave propagation velocity. Whereas seismic body-waves show virtually no dispersion, it is pronounced for seismic surface waves. It causes a significant stretching of the length of the surface-wave record and the rather late arrival of its largest amplitudes (Airy phases) from which the surface-wave magnitude M_S and the mantle magnitude M_m, respectively, are determined.

Earthquake size A frequently used, but not uniquely defined term. It may be related – more or less directly – to either the geometric-kinematic size of an earthquake in terms of area and slip of the fault or to the *seismic energy* radiated from a seismic source and its potential to cause damage and casualty (moment or energy *magnitude*).

Earthquake source In general terms, the whole area or volume of an *earthquake* rupture where seismic body waves are generated and radiated outwards. More specifically, one speaks either of the *source mechanism* or the source location. The latter is commonly given as earthquake hypocenter (i. e. the location at the source depth h from where the seismic rupture, collapse or explosion begins) or as the point on the Earth's surface vertically above the hypocenter, called the epicenter. Earthquakes at $h < 70$ km are shallow, those at larger depth either intermediate (up to $h = 300$ km) or deep earthquakes ($h = 300$–700 km). The determination of the geographical coordinates latitude φ, longitude λ, and focal depth h, is the prime task of seismic source location. However, for extended seismic sources, fault ruptures of great earthquakes in particular, the hypocenter is generally not the location of largest fault slip and/or seismic moment/energy release and the epicenter is then also not the location where the strongest ground shaking is felt. The locations of largest effects may be dozens of kilometers in space and many seconds to minutes in time away from the hypocenter or epicenter, respectively.

Fundamental modes The longest period oscillations of the whole Earth with periods of about 20 min (spheroidal mode), 44 min. (toroidal mode) and some 54 min ("rugby" mode), excited by great earthquakes.

Magnitude A number that characterizes the relative *earthquake size*. It is usually based on measurement of the maximum motion recorded by a seismograph (sometimes for waves of a particular type and frequency) and corrected for the decay of amplitudes with epicenter distance and source depth due to geometric spreading and attenuation during wave propagation. Several magnitude scales have been defined. Some of them show *saturation*. In contrast, the moment magnitude (M_w), based on the concept of *seismic moment*, is uniformly applicable to all earthquake sizes but is more difficult to compute than the other types, similarly the energy magnitude, M_e, which is based on direct calculation of the *seismic energy* E_s from broadband seismic records.

Saturation (of magnitudes) Underestimation of *magnitude* when the duration of the earthquake rupture significantly exceeds the seismic wave period at which the magnitude is measured. The shorter this period, the earlier respective magnitudes will saturate (see relation (13) and Figs. 4 and 5).

Seismic energy Elastic energy E_s (in joule) generated by, and radiated from, a seismic source in the form of seismic waves. The amount of E_s is generally much smaller than the energy associated with the non-elastic deformation in the seismic source (see *seismic moment* M_o). The ratio $E_s/M_o = (\Delta\sigma/2\mu) = \tau_a/\mu$, i. e., the seismic energy released per unit of M_o, varies for earthquakes in a very wide range between some 10^{-6} and 10^{-3}, depending on the geologic-tectonic environment, type of *source mechanism* and related stress drop $\Delta\sigma$ or apparent stress τ_a.

Seismic moment M_o A special measure of earthquake size. The moment tensor of a shear rupture (see *earthquake source*) has two non-zero eigenvalues of the amount $M_o = \mu \bar{D} F_a$ with μ-shear modulus of the ruptured medium, $\bar{D}$-average source dislocation and F_a-area of the ruptured fault plane. M_o is called the scalar seismic moment. It has the dimension of Newton meter (Nm) and describes the total non-elastic (i. e., ruptural and plastic) deformation in the seismic source volume. Knowing M_o, the moment *magnitude* M_w can be determined via Eq. (11).

Source mechanism Depending on the orientation of the earthquake fault plane and slip direction in space, one discerns different source mechanisms. Strike-slip faults are vertical (or nearly vertical) fractures along

which rock masses have mostly shifted horizontally. Dip-slip faults are inclined fractures. If the rock mass above an inclined fault moves down (due to lateral extension) the fault is termed normal, whereas, if the rock above the fault moves up (due to lateral compression), the fault is termed reverse (or thrust). Oblique-slip faults have significant components of both slip styles (i. e., strike-slip and dip-slip). The greatest earthquakes with the largest release of seismic moment and the greatest potential for generating tsunamis are thrust faults in subduction zones where two of Earth's lithosphere plates (e. g., ocean–continent or continent–continent) collide and one of the two plates is subducted underneath the overriding plate down into the Earth's mantle. Different source mechanisms are characterized by different radiation patterns of seismic wave energy.

Transfer function The transfer function of a seismic sensor-recorder system (or of the Earth medium through which seismic waves propagate) describes the frequency-dependent amplification, damping and phase distortion of seismic signals by a specific sensor-recorder (or medium). The modulus (absolute value) of the transfer function is termed the amplitude-frequency response or, in the case of seismographs, also magnification curve (see Fig. 3).

Definition of the Subject

Besides earthquake location (i. e., the determination of the geographical coordinates of the epicenter, the hypocenter depth and the origin time; for definition of these terms see *earthquake source* in the Glossary), the *magnitude* is the most frequently determined and commonly used parameter to characterize an earthquake. Despite its various imperfections, it provides important information concerning the earthquake source spectrum at the period where the magnitude is measured and current source theories (cf. [3]) allow one to understand differences in the source spectra of different earthquakes in terms of source dimension and stress drop, i. e., the difference between the stress level before and after the earthquake. Via various empirical relations, magnitudes enable estimates of the *seismic moment* and the *seismic energy* released by the earthquake. These parameters are important in the discussion of various global problems such as the seismic slip rates between lithosphere plates and the excitation of Chandler Wobble [25]. Besides these more academic issues, magnitude values have an immense practical value in providing:

a) Rapid simple parameter estimates of the strength of an earthquake that can help to realistically assess the related ground shaking or tsunami potential and thus assist efficient disaster management response;
b) Mass data in earthquake catalogs and data banks, covering long time periods over many decades – and hopefully centuries in future, which allows one to assess the seismic activity and related hazards of Earth's regions and their possible variability in space and time. This is not only of high scientific interest, but also the very basis for realistic long-term disaster preparedness and risk mitigation efforts.

The term magnitude and the basic method of its determination were introduced by Charles F. Richter in 1935 [71]. He intended to compare the relative *earthquake size* in southern California in terms of differences in the maximum amplitudes A recorded at a network of seismic stations that were equipped with standard short-period Wood–Anderson (WA) torsion seismometers.

The WA seismometer response is depicted in Fig. 3 and Fig. 1 shows a WA record and magnitude measurement example. In order to make amplitudes recorded by stations at different epicentral distances D from the earthquake comparable, Richter had to compensate for the amplitude decay with D using an appropriate correction term $-A_o(D)$. Since the strength and thus the radiated amplitudes of earthquakes vary in a wide range Richter defined his local magnitude scale M_L, determined from records at source distances up to 600 km, as follows:

> "*The magnitude of any shock is taken as the logarithm of the maximum trace amplitude, expressed in microns, with which the standard short-period torsion seismometer . . . would register that shock at an epicentral distance of 100 km.*"

Thus:

$$M_L = \log A_{max} - \log A_o(D) \,. \tag{1}$$

According to the above definition, an amplitude of 1 μm in a WA record at a distance $D = 100$ km from the epicenter would correspond to $M_L = 0$. Amplitude means in (1) and the following text either the center-to-peak or half of the peak-to-trough amplitude.

Wood–Anderson (WA) seismographs record horizontal short-period ground motions with an amplification of only about 2080 times [82]. Modern electronic seismographs may achieve magnifications larger than 10^6 and thus are able to record local earthquakes with even negative magnitudes, down to about -2. The largest values determined with the M_L scale are around seven. Later it was found that all magnitudes derived from short-period waves (typically with periods $T < 3$ s) show *saturation*

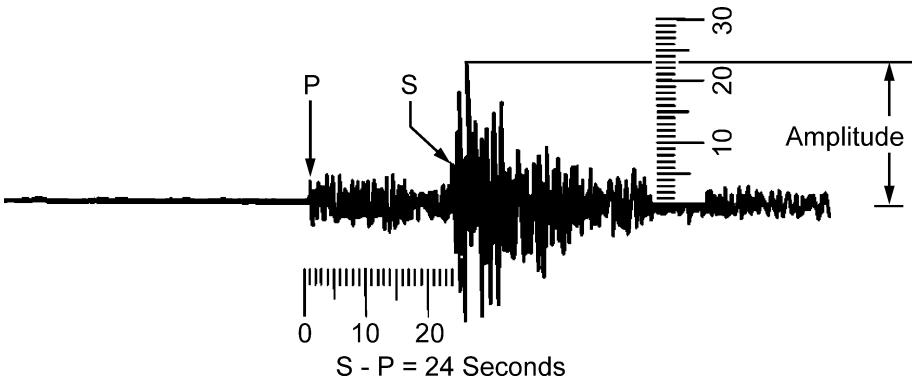

Earthquake Magnitude, Figure 1
Record of a short-period Wood–Anderson seismograph (frequency-magnification curve see Fig. 3) of a local earthquake. *P* marks the onset of the first arriving longitudinal *P* wave, and *S* the onset of the much stronger secondary, transverse polarized shear wave. Note the long tail of coda-waves following *S*. From the time difference $S - P = 24$ s follows a hypocentral distance $R = 190$ km. The maximum record amplitude is $A_{max} = 23$ mm. Applying the amplitude-distance correction $-\log A_o(190\,\text{km}) = 3.45$ according to Richter [72] results in a magnitude $M_L = 4.8$

(see Glossary, Fig. 4 and Sect. "Magnitude Saturation and Biases Due to Earthquake Complexity"). Therefore, it was necessary to develop complementary magnitude scales that use medium to long-period ($T \approx 5\,\text{s} - 30\,\text{s}$) as well as very long-period waves ($T \approx 50\,\text{s} - 3000\,\text{s}$) in order to enable less or non-saturating magnitude estimates (see Sect. "Introduction to Common Magnitude Scales: Potential and Limitations"). For the so far strongest instrumentally recorded earthquake (Chile 1960) a value of $M = 9.5$ was determined that way. Accordingly, instrumental seismic monitoring currently covers the magnitude range of about $-2 \leqslant M < 10$. This roughly corresponds to ruptures of some millimeters to more than 1000 km long. They radiate approximately the same amount of seismic wave energy E_s as well-contained underground explosions with yields ranging from a few milligrams (10^{-9} t) to several 10 to 100 Gt (1 Gt $= 10^9$ t) Trinitrotoluol (TNT) equivalent, thus covering about 20 orders in energy. Earthquakes with magnitudes around four may cause only minor local damage, those with magnitudes > 6 heavy damage, and those with magnitudes > 7 already widespread devastating damage. Shallow submarine earthquakes with magnitudes > 7 may generate significant local tsunamis with damage potential to nearby shores whereas those with magnitudes > 8.5 may stimulate ocean-wide tsunamis causing destruction and casualties even at shores thousands of kilometers away from such earthquakes.

In order to measure and classify *earthquake size* in the wide range of magnitudes from about -2 to < 10 and satisfy specific requirements in research and application which are based on magnitude data, it was indispensable to develop different magnitude scales that are complementary, but properly scaled to the original Richter M_L. Thus, there exists today a host of magnitude scales applicable in a wide range of source distances from less than 1 km up to more than 10,000 km. These scales, their specifics, potential and limitations are discussed in detail (with many reference given) in Chapter 3 of the IASPEI New Manual of Seismological Observatory Practice [6]. The early pioneers of magnitude scales, Beno Gutenberg and Charles Richter, had hoped that different magnitude scales could be cross-calibrated to yield a unique value for any given earthquake (cf. [25,30]. In their joint book [29] "Seismicity of the Earth" (1954; first edition 1949) and later in Richter's [72] famous text book "Elementary Seismology" as well as in Duda [22] only one magnitude value M was given per earthquake. However, this approach proved only partially realistic under certain conditions and within limited magnitude ranges because of the often significant differences in measurement procedures as well as period and bandwidth ranges used in later magnitudes scales. Decades later it took significant efforts (cf. [1,2,25]) to reconvert these M values, which turned out to be not even compatible (cf. [25]) into their original body or surface wave magnitudes in order to get values that agree with the original definition of these specific magnitude scales and can be compared with current data of the same type.

In general, such magnitude conversion relations strongly depend on initial data errors and the type of least-square regression procedure applied [11,14]. Moreover, the latter have often not been interpreted and used in a correct way. This may result in the case of noisy mag-

nitude data for events at the upper and lower end of the investigated magnitude range, in conversion errors of more than 0.5 magnitude units (m.u.) with serious consequences on seismic hazard estimates based on such converted magnitudes (cf. [7,11,14,15]). Moreover, magnitude values determined within the *saturation* range of a given scale cannot reliably be converted via empirical regression relations into the equivalent magnitude values of another less or non-saturating magnitude scale (see Fig. 4 and [44]). Furthermore, some magnitudes relate best to the released *seismic energy* while others are scaled to the static *seismic moment*, i. e., they measure equally important but fundamentally different physical aspects of the source and the radiated seismic waves and may differ by sometimes more than 1 m.u. Thus there is no way to characterize *earthquake size* in all its different aspects by just a single magnitude value. Proper interpretation and use of different types of magnitude data, however, requires one to understand the physics behind such values and how these may be affected by the complexity and duration of the earthquake rupture process. Further, this necessitates one to discriminate unambiguously the different types of magnitude values by using a unique nomenclature and to assure that magnitude values published with a given nomenclature have been determined with an internationally agreed standard procedure. With this in mind, the most important magnitude scales and related problems are summarized in Sects. "Introduction to Common Magnitude Scales: Potential and Limitations" and "Common Magnitude Estimates for the Sumatra 2004 M_w 9.3 Earthquake".

Introduction to Common Magnitude Scales: Potential and Limitations

Magnitude Scales Used in the Local and Regional Distance Range ($D < 2000$ km)

The original Richter local magnitude scale for Southern California [71] has been further developed since its invention [38]. In its expanded form (with the nomenclature M_L common in the United States), the following relation now holds:

$$M_L = \log_{10}(A_{max}) + 1.11 \log_{10} R + 0.00189\, R - 2.09 \quad (2)$$

with R = distance from the station to the hypocenter in kilometers and A_{max} = maximum trace amplitude in nanometers (instead of μm in a WA record). This amplitude is measured on the output from a horizontal-component seismograph that is filtered so that the response of the seismograph/filter system replicates that of a WA standard seismograph but with a static magnification of one. The underlying procedure of M_L determination according to relation (2) was adopted by the International Association of Seismology and Physics of the Earth's Interior (IASPEI) in 2004 as the standard procedure for determining local magnitudes in the distance range up to typically less than 1000 km [42]. For earthquakes in the Earth's crust of regions with attenuation properties that differ from those of coastal California, and for measuring M_L with vertical component seismographs, the standard equation takes the form:

$$M_L = \log_{10}(A_{max}) + F(R) + G \quad (3)$$

where $F(R)$ is an R-dependent calibration function and G a constant which have to compensate for different regional attenuation and/or for any systematic biases of amplitudes measured on vertical instead on horizontal seismographs. Examples of regional M_L calibration functions developed for different parts of the world have been compiled by Bormann (Chap. 3, p. 26, and DS 3.1 in [6]).

A few decades ago, analog seismic records prevailed. They had a rather limited dynamic range of only some 40 dB. This caused record traces often to go off-scale when stronger seismic events were recorded at local or regional distances. Then A_{max} could not be measured. Yet, it was found that the duration **d** of the coda that follows A_{max} with exponentially decaying amplitudes (see Fig. 1) increases with magnitude and distance D. On this basis, local duration magnitude formulas of the following general form

$$M_d = a + b \log \mathbf{d} + cD \quad (4)$$

have been developed with a, b and c being coefficients to be determined locally. When using only recordings at distances $D < 100$ km the distance term cD is not even needed. However, crustal structure, scattering and attenuation conditions vary from region to region. Moreover, the resulting specific equations will also depend on the chosen definition for **d**, the local signal-to-noise (SNR) conditions and the sensor sensitivity at the considered seismic station(s) of a network. Therefore, M_d scales have to be determined locally for a given source-network configuration and scaled to the best available amplitude-based M_L scale.

Nowadays digital recorders with large usable dynamic range of about 140 dB are common. Thus even sensitive modern broadband seismographs remain on scale when recording local or regional earthquakes up to $M \approx 7$. This reduces the need for M_d scales. Moreover, the increasing availability of modern strong-motion (SM) recorders with comparably large dynamic range, which will not clip even in the case of very strong nearby earthquakes, have led to

the development of (partially) frequency-dependent M_L^{SM} scales. They are usually based on the calculation of synthetic WA seismograph outputs from strong-motion accelerograms [35,54].

Also, amplitudes of short-period L_g waves with periods around 1 s are sometimes used to determine magnitudes, termed $m_b(L_g)$. L_g waves travel with group velocities of 3.6 to 3.2 km/s and arrive after the (secondary, shear) S wave onset (Fig. 1). They propagate well in continental platform areas. Recently, the IASPEI [42] adopted a measurement procedure for $m_b(L_g)$ as international standard, which had been developed for eastern North America [62] with the aim to improve yield estimates of Nevada Test Site explosions. However, as for all other local or regional magnitude scales, the calibration term is strongly influenced by the local/regional geologic-tectonic conditions in the Earth's crust and requires a proper scaling to this standard, when applied to other areas than eastern North America.

Tsuboi developed for the Japan Meteorological Agency (JMA) in 1954 [79] a magnitude formula for shallow earthquakes (depth $h < 60$ km) that have been recorded at epicentral distances D up to 2000 km:

$$M_{JMA} = \log_{10} A_{max} + 1.73 \log_{10} D - 0.83 \,. \qquad (5)$$

A_{max} is the largest ground motion amplitude (in μm) in the total event record of a seismograph with an eigenperiod of 5 s. If horizontal seismographs are used then $A_{max} = (A_{NS}^2 + A_{EW}^2)^{1/2}$ with A_{NS} and A_{EW} being half the maximum peak-to-trough amplitudes measured in the two horizontal components. This formula was devised to be equivalent to the medium to long-period Gutenberg–Richter [29] magnitude M. Therefore, M_{JMA} agrees rather well with the *seismic moment* magnitude M_w. The average difference is less than 0.1 in the magnitude range between 4.5 and 7.5 but becomes > 0.5 for $M_w > 8.5$ (see Fig. 4). Katsumata [49,50] has later modified the M_{JMA} formula for earthquakes deeper than 60 km.

Another, more long-period regional moment magnitude scale, termed M_{wp}, has been developed in Japan as well [80]. It provides quick and less saturating magnitude estimates for tsunami early warning. Velocity-proportional records are twice integrated and approximately corrected for geometrical spreading and an average P-wave radiation pattern (see *source mechanism*) to obtain estimates of the scalar *seismic moment* M_o at each station. Usually the first maximum in the integrated displacement trace, called "moment history" $M_o(t)$, is assumed to represent M_o. From these M_o values moment magnitudes M_w are then calculated for each station according to Eq. (11) and averaged. M_{wp} results from adding an empirically derived correction of 0.2 m.u. to the averaged station M_w [80]. Finally, a magnitude-dependent correction is applied to M_{wp} [86] in order to get an even better estimate of the recognized "authoritative" Global Centroid Moment Tensor magnitude M_w (GCMT) which is calculated according to the Harvard procedure [23] and now published under [41].

The M_{wp} concept was originally developed for earthquakes at $5° \leqslant D° \leqslant 15°$, but can be applied for $M_w < 7.5$ (down to about $M_w \approx 5$) even to shorter local distances as long as this distance is significantly larger than the rupture length. Later the M_{wp} procedure has been adopted for application to records of deep and teleseismic earthquakes as well [81]. M_{wp} estimates are standard routine in Japan, at the Alaska and the Pacific Tsunami Warning Centers (ATWC and PTWC), and the National Earthquake Information Center (NEIC) of the United States Geological Survey (USGS). However, each of these centers use slightly different procedures. Values for most strong earthquakes are usually available some 10 to 15 min after the origin time (OT). On average M_{wp} data scale well with M_w. Exceptions, however, are extremely slow or very large complex earthquakes. Then M_{wp} is usually too small, up to about 1 m.u.

In recent years great attention is paid to the development of even more rapid earthquake early warning systems (EWS). They aim at event location and magnitude estimates from the very first few seconds of broadband acceleration, velocity or displacement records and within about 10 to 30 s after origin time (OT) of strong damaging earthquakes on land. These data are to be used for instantaneous public alarms and/or automatically triggered risk mitigation actions after strong earthquakes with damage potential. The goal is to minimize the area of "blind zones" which are left without advanced warning before the arrival of the S waves which have usually the largest strong-motion amplitudes (see Fig. 1). This necessitates very dense and robust local seismic sensor networks within a few tens of kilometers from potentially strong earthquake sources. Such networks are at present available only in very few countries, e. g. in Japan, Taiwan, Turkey, and Italy.

Their principles of rapid magnitude estimates differ from those mentioned above and below and the data analysis from such systems is largely based on still much debated concepts such as the hypothesis of the deterministic nature of earthquake rupture [66,73]. Data presented in [66] seem to suggest that in the range $3.0 < M$ (not specified) < 8.4 the magnitude can be estimated with an average absolute deviation of 0.54 m.u. from the maximum period within the initial 4 s of the first arriving

(primary, longitudinal) P wave when many low-pass filtered velocity records within 100 km from the epicenter are available. However, for $M > 6$ the systematic increase of these greatly scattering periods becomes rather questionable. When analyzing waveforms of the Japanese Hi-net seismic network [73], it could not be confirmed that such a dominant frequency scaling with magnitude exists. Also Kanamori [46], together with Nakamura [60,61], one of the fathers of this idea, expressed much more caution about the prospects of this method after he had run, together with Wu [89], an experiment with the Taiwan EWS. For each event they analyzed the first 3 s of at least eight P-wave records at epicentral distances < 30 km. They knew that: "… the slip motion is in general complex and even a large event often begins with a small short-period motion, followed by a long-period motion. Consequently, it is important to define the average period during the first motion." (termed τ_c in [46,89]). However, after applying the τ_c concept to the Taiwan EWS they concluded: "For EWS applications, if $\tau_c < 1$ s, the event has already ended or is not likely to grow beyond $M > 6$. If $\tau_c > 1$ s, it is likely to grow, but how large it will eventually become, cannot be determined. In this sense, the method provides a threshold warning". Thus it seems that these new concepts work reasonably well only for earthquakes with $M < 6.5$ and thus total rupture durations that are according to Eq. (13) on average not more than about 2–3 times the measurement time windows of 3 s or 4 s used in [46,66,89]. Nakamura and Saita [61] reported data from a much smaller set of events ($N = 26$) recorded at local distances in the range $4.6 < M < 6.9$. We calculated the average absolute deviation of their rapid UrEDAS system magnitudes (0.47 m.u.) from the official magnitudes M_{JMA} published later by the Japan Meteorological Agency. This error decreases to 0.32 m.u. when only earthquakes with magnitudes up to $M_{JMA} = 6.0$ are considered. This seems to support our assessment that the reliability of real-time EMS magnitudes decreases rapidly if the analyzed time window is much shorter than the rupture duration.

Magnitude Scales Used in the Teleseismic Distance Range ($D > 2000$ km)

Ten years after the introduction of the local magnitude M_L, Beno Gutenberg [26,27,28] extended the concept of magnitude determination to teleseismic distances larger than about 1000–2000 km. He used both records of seismic waves that propagate along the Earth's surface (or near to it with a period-dependent penetration depth) and waves which travel through the Earth. Accordingly, the former are termed surface waves and the latter body waves. For the surface-wave magnitude Gutenberg [28] gave the following relation:

$$M_S = \log_{10} A_{Hmax} + 1.656 \log D^\circ + 1.818 \qquad (6)$$

with A_{Hmax} = maximum "total" horizontal displacement amplitude of surface-waves in μm for periods around 20 ± 2 s measured in the distance range $15^\circ < D^\circ < 130^\circ$ ($1^\circ = 111{,}195$ km).

While the original Richter M_L and Gutenberg M_S magnitudes were calculated from the maximum ground displacement amplitudes, Gutenberg [26,27,30] proposed to determine the body-wave magnitudes m_B from the relation:

$$m_B = \log_{10} (A/T)_{max} + Q(D^\circ, h) \,, \qquad (7)$$

i. e., by measuring the maximum ratio of ground displacement amplitude A (in μm) divided by the related period T (in s). A/T is equivalent to measuring the maximum ground motion velocity $A_{vmax}/2\pi$ which is proportional to the square root of *seismic energy*, i. e. $\sqrt{E_s}$. Thus the magnitude becomes a measure of the elastic kinetic wave energy radiated by an earthquake. Only in this way comparable magnitude data could be obtained for different types of body waves and measurements at different sites. Another great advantage of m_B is that it permits magnitude estimates also from intermediate and deep *earthquake sources*, which produce only weak or no surface waves at all. Empirical relationships permit estimating E_s (in units of Joule) from body-wave magnitude m_B [30]

$$\log_{10} E_s = 2.4\, m_B - 1.2 \qquad (8)$$

or surface-wave magnitude M_S [72]

$$\log_{10} E_s = 1.5\, M_S + 4.8 \,. \qquad (9)$$

Accordingly, an increase by 1 m.u. in m_B and M_S corresponds to an increase of radiated seismic energy by about 250 and 30 times, respectively.

Revised empirical distance-depth corrections for the calibration of body-wave magnitudes, so-called Q-functions, were published in 1956 by Gutenberg and Richter [30]. They are given as separate tables and charts for the body-wave phases P, PP (a P wave reflected at the surface of the Earth about the half way between source and station) and S. They are still in use, especially Q_{PV} for calibrating amplitude measurements made on vertical component P-wave records (Fig. 2). However, for epicenter distances between 5° and 20° these calibration values are not reliable enough for global application. In this range

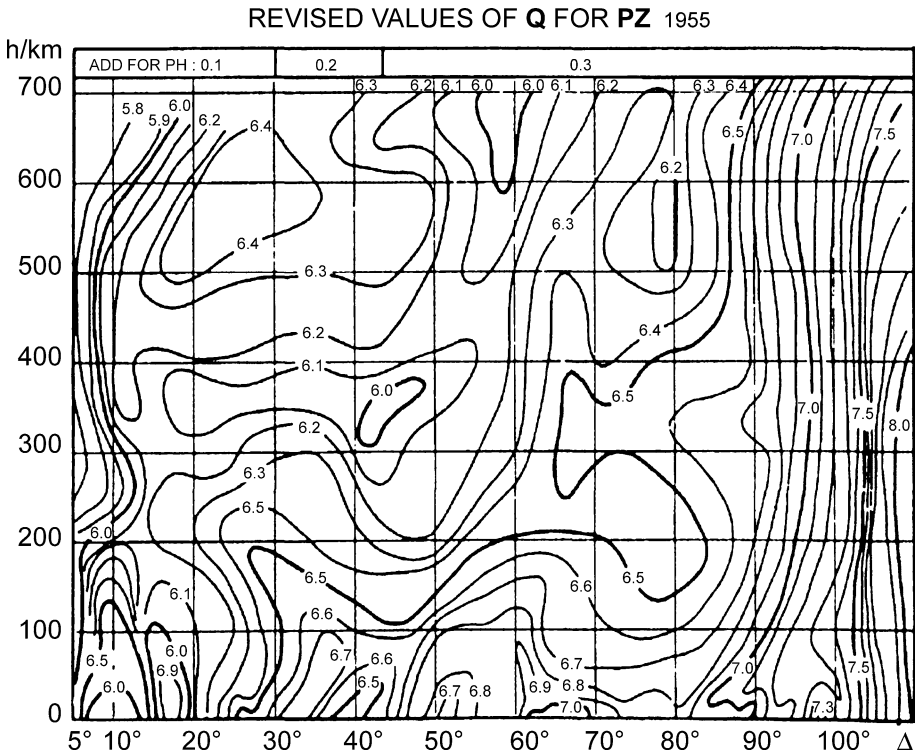

Earthquake Magnitude, Figure 2
Calibration values $Q(D°, h)$ for vertical (Z) component P-wave amplitudes depending on epicentral distance $D° = \Delta$ and source depth h as used in the calculation of body-wave magnitudes m_b and m_B according to Gutenberg and Richter, 1956 [30]

the wave propagation is strongly affected by regional variations of the structure and properties of the Earth's crust and upper mantle. And for $D > 100°$ the P-wave amplitudes decay rapidly because of the propagation of P waves is influenced by the Earth's core (so-called core shadow). Therefore, in agreement with current IASPEI recommendations [42], m_B and its short-period complement m_b (see below), should be determined by using Q_{PV} only between $21° \leqslant D \leqslant 100°$.

These body-wave magnitude calibration functions had been derived from amplitude measurements made mostly on medium-period broadband displacement records which dominated during the first half of the 20th Century at seismological stations. Their period-dependent magnification curve resembled more or less that of the classical standard seismograph type C shown in Fig. 3, although for some of these instruments the roll-off of the amplification occurred already at periods $T > 10$ s.

Another, so-called Prague–Moscow formula for surface-wave magnitudes was proposed in 1962 by Vaněk et al. [84]. It is based on the measurement of $(A/T)_{max}$ in records of shallow earthquakes ($h < 60$ km) in wide period and distance ranges ($3\,\text{s} < T < 30\,\text{s}$; $2° \leqslant D° \leqslant 160°$):

$$M_S = \log_{10}(A/T)_{max} + 1.66 \log_{10} D° + 3.3 \,. \qquad (10)$$

This relationship, which is – as Eq. (7) – more directly related to E_s, was adopted by the IASPEI in 1967 as international standard.

The NEIC adopted Eq. (10), but continues to limit the range of application to distances between $20° \leqslant D° \leqslant 160°$ and displacement amplitudes in the very limited period range as in formula (6) although Soloviev [74] had shown already in 1955 that $(A/T)_{max}$ is a stable quantitative feature of surface waves whatever the period of their maximum at all epicentral distances. Also theory has confirmed [63] that using the ratio (A/T) is a partial and ad hoc compensation for a large number of frequency-dependent terms ignored in (10). In fact, the periods at the surface-wave maximum used for M_S determination vary in a wide range between some 3 s and 25 s and show – despite large scatter – a clear distance dependence [84,87]. Therefore, several authors [36,70] showed that using Eq. (10) only for amplitude readings around 20 s results in system-

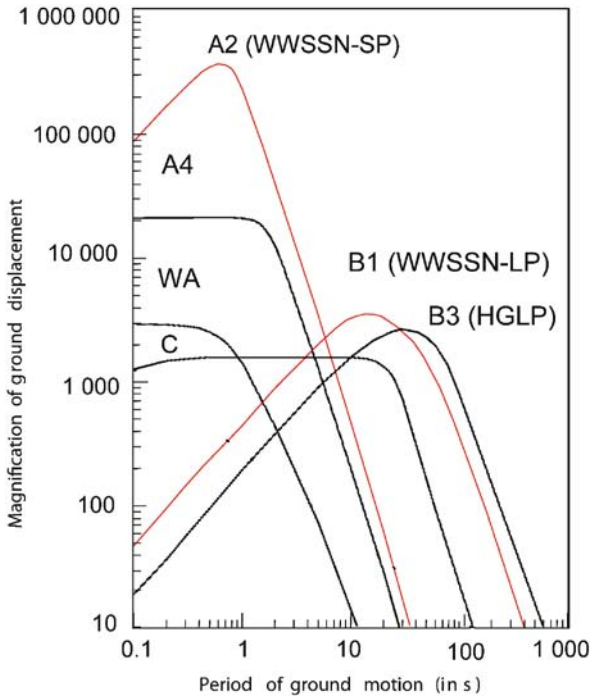

Earthquake Magnitude, Figure 3
Magnification of ground displacement amplitudes by common standard types of seismographs. WA = Wood–Anderson seismograph; WWSSN-SP and WWSSN-LP = short-period and long-period seismographs used in the former United States World-Wide Seismograph Standard Network; HGLP = US type of High Gain Long Period seismographs; A2, A3, B1, B3 and C = standard types of seismographs according to Willmore [87]. Reprint from [6] with © granted by IASPEI

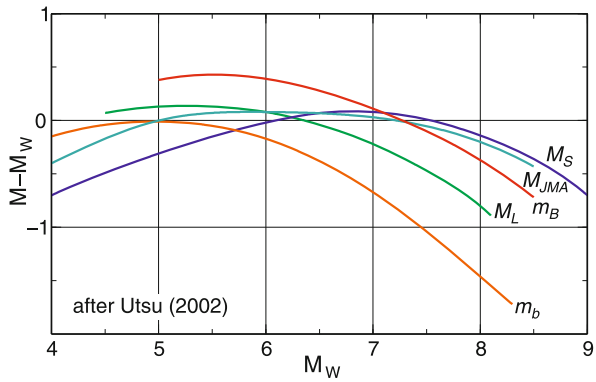

Earthquake Magnitude, Figure 4
Average relationships between different common types of magnitudes and the moment magnitude M_W. Modified from Fig. 1 in [83]

atic distance-dependent biases. However, their proposed revised calibration functions for 20 s waves are not yet used in routine practice at international seismological data centers.

The formulas (6) and (10) had originally been developed for horizontal component amplitude readings. Beginning in the 1960s, however, more and more long-period and broadband vertical component instruments became available and are now commonly used for magnitude determination from surface waves. This procedure is easier and better defined than measuring and combining the amplitude measurements made in two horizontal components, yields on average values that are largely comparable with the Gutenberg M_S [25] and has recently been adopted as IASPEI [42] standard. Herak et al. [37] published theoretical and observed depth corrections for $M_S(20)$ when determined according to (10). These corrections allow determination of more reliable surface-wave magnitudes for earthquakes in all depth ranges and improve significantly the relationship between M_S and the *seismic moment* M_o.

In the 1960s, the United States deployed a World-Wide Standard Seismograph Network (WWSSN) equipped with short-period (SP) and long-period (LP) seismographs of limited bandwidth (cf. Fig. 3). This network had two priority tasks. Firstly, to significantly increase the signal-to-noise ratio of the seismic records by narrow-band short-period filtering, thus improving the global detection threshold for teleseismic events down to magnitudes around 4–4.5 and the location accuracy for seismic events. Secondly, to realize an effective discriminator between underground nuclear explosions (UNE) and natural earthquakes based on the ratio of a short-period body-wave and a long-period surface-wave magnitude. Natural earthquakes have a much longer source duration (seconds to minutes) than explosions of comparable size (typically milliseconds). Also, at comparable *seismic moment* magnitude, UNEs radiate significantly more high-frequency energy (see dotted curve in Fig. 5) than earthquakes. Therefore, a better discrimination of the two types of events was achieved by measuring the *P*-wave amplitude only at periods < 3 s (typically around 1 s) and calculating a short-period *P*-wave magnitude termed m_b. In contrast, the Gutenberg m_B is based on measuring A_{max} at periods T usually between 2 s and 30 s. Further, during the first two decades of the WWSSN, the *P*-wave amplitude was not – as required by Gutenberg's procedure for m_B determination – always measured at the maximum of the whole *P*-wave train (whose length depends on the source duration and thus on the magnitude itself) but initially within the first five half-cycles only and later by USGS requirement in the first 5 s of the record.

Because of the short source duration of explosions, their *P*-waves will always reach maximum amplitudes

within such a short time-interval. However, P waves radiated by large earthquakes of much longer source duration will reach their maximum amplitude usually much later in the rupture process. For magnitude 6 the average rupture duration is on average already 6 s, and may increase to about 600 s for the strongest earthquakes (cf. relation (13)). Both effects together, plus the fact, that m_b was still computed using the Q_{PV} function derived for mainly medium-period P waves, which are much less affected by frequency-dependent attenuation than 1 Hz P waves, resulted in a systematic underestimation of the earthquake size for magnitudes larger than 5 and a *saturation* of m_b at around 6.5.

In the late 1970s, the NEIC switched back to a longer time window of about 15 s and more recently, with an automatic procedure, to a window covering the first 10 cycles of short-period teleseismic P waves. In the case of strong earthquakes this window may later be extended interactively up to 60 s. This mitigates to some extend the saturation of m_b. However, no m_b-values larger than 7.2 have ever been measured with this procedure. On the other hand m_b yields rather reliable values for magnitudes < 5 when the corner frequency of the average source spectra falls within the passband of the short-period seismograph or is even more high frequency (cf. Figs. 3, 4, 5). For magnitudes < 5 m_B can usually no longer be determined because of too small signal-to-noise ratio (SNR) in broadband records. Then m_b is often the only available teleseismic estimator of earthquake size for small earthquakes.

Most seismic stations and networks worldwide adopted the US procedure for m_b measurement and – with the exception of Russia, China and their former allies – completely abandoned measuring m_B as originally defined. This change in attitude was stimulated by the fact that the NEIC, which serves in fact as one of the leading international data centers for seismology, did not accept reported P-wave amplitudes other than those obtained from short-period measurements. Some stations, national and global data centers continue (at least up to 2008) to measure for m_b the maximum amplitude of P exclusively within the first 5 s after the P-wave *first arrival*, such as the China Earthquake Network Center and the International Data Center (IDC) of the Comprehensive Test-Ban Treaty Organization (CTBTO) in Vienna.

Because of these inconsistencies in m_b and M_S determination and the proven merits of both broadband m_B and M_S (see also [11]) the IASPEI Working Group on Magnitude Measurements recommended that in future:

a) m_b is always determined from A_{max} at periods $T < 3$ s within the whole P-wave train;
b) The band-limited magnitudes m_b and $M_S(20)$ be complemented by true broadband magnitudes m_B and $M_S(BB)$. The latter two will be obtained by measuring A_{vmax} on unfiltered velocity broadband records and thus always include the maximum velocity amplitudes of the source spectrum in the magnitude range of interest (cf. Fig. 5). This will link these two broadband magnitudes to the seismic energy released by an earthquake, more closely than the common band-limited magnitudes.

These recommendations have been adopted by the IASPEI Commission on Seismic Observation and Interpretation (CoSOI) in 2005 as new magnitude measurement standards. More details about the new measurement procedures for m_b, m_B, $M_S(20)$ and $M_S(BB)$ are given on the CoSOI web site [42]. Beginning in 2007 they are gradually implemented at the main seismological data centers and networks.

Since all magnitudes discussed so far show more or less pronounced *saturation* for large earthquakes (cf. Fig. 4 and [44]) a non-saturating magnitude, termed M_w, has been proposed [31,43,69]. The moment magnitude M_w is derived from the scalar *seismic moment* M_o via the relation

$$M_w = (2/3)(\log_{10} M_o - 9.1) \,. \tag{11}$$

M_o has the dimension of Newton meter (Nm) and expresses the total inelastic "work" required for rupturing and displacing the considered earthquake fault. It can be determined either by waveform analysis and inversion in the time domain or by measuring the spectral amplitude $u_{0p,s}$ of the low-frequency level (plateau) of the displacement spectrum of P or S waves (cf. Fig. 5) via the relationship

$$M_o = 4\pi r \rho \, v_{p,s}^3 u_{0p,s} \big/ R_{\theta,\phi}^{p,s} \tag{12}$$

with r = hypocenter distance, ρ = average density of rocks in the source and receiver area, $v_{p,s}$ = average velocity of the P or S waves from the source to the receiver area and $R_{\theta,\phi}^{p,s}$ = a factor correcting the observed seismic amplitudes for the influence of the radiation pattern of the given *source mechanism*, which is different for P and S waves.

M_o is expected to show no *saturation*, provided that the amplitude level is measured only at periods significantly larger than the magnitude-dependent *corner period* of the seismic source spectrum (cf. Fig. 5). In Sects. "Common Magnitude Estimates for the Sumatra 2004 M_w 9.3 Earthquake" and "Magnitude Saturation and Biases Due to Earthquake Complexity" we will show, however, that in-

correct determination of M_o may still result in an underestimation of the earthquake size. Since M_w is derived from M_o it is related to the tectonic effect of earthquakes, i. e., to the product of rupture area and average fault slip and thus also relevant to assess the tsunami potential of strong shallow marine earthquakes. An example is the off-shore Nicaragua earthquake of 2 September 1992. Its $m_b = 5.3$ was too weak to alert the people ashore, some 70–120 km away from the source area. However, its $M_w = 7.6$ was much larger and caused a damaging local tsunami with almost 200 casualties.

Yet, M_o and thus M_w do not carry any direct information about the dominant frequency content and thus of the seismic energy released by the earthquake (cf. Sect. "Magnitude Saturation and Biases Due to Earthquake Complexity"). In fact, relation (11) was derived by assuming constant stress drop and an average ratio of $E_s/M_o = 5 \times 10^{-5}$ on the basis of elastostatic considerations and empirical data [43] and then replacing in Eq. (9) M_S by M_w.

As source theory has advanced and broadband digital data have became readily available, the radiated *seismic energy* E_s could be computed explicitly rather than from an empirical formula. Boatwright and Choy (cf. [5,16]) developed such an algorithm for computing E_s as well as a related energy magnitude M_e which agrees with M_w for $E_s/M_o = 2 \times 10^{-5}$. E_s is computed by integrating squared velocity-proportional broadband records over the duration of the *P*-wave train, corrected for effects of geometrical spreading, frequency-dependent attenuation during wave propagation and source radiation pattern. According to [16], the radiated seismic energy may vary for a given seismic moment by two to three orders of magnitude. Further, it was found that a list of the largest events is dominated by earthquakes with thrust mechanisms when size is ranked by moment, but dominated by strike-slip earthquakes when ranked by radiated seismic energy. Choy and Kirby [18] gave a striking example for differences between M_e and M_w for two Chile earthquakes in 1997 which occurred in the same area but with different *source mechanisms*. One was interplate-thrust with $M_w = 6.9$ and relatively low $M_e = 6.1$, whereas the other was intraslab-normal with $M_w = 7.1$ and rather large $M_e = 7.6$. The first earthquake had a low potential to cause shaking damage and was felt only weakly in a few towns. In contrast, the second one caused widespread damage, land- and rockslides, killed 300 people and injured 5000. Thus, M_w, although it theoretically does not saturate, may strongly underestimate or overestimate the size of an earthquake in terms of its potential to cause damage and casualties. Shaking damage is mainly controlled by the relative amount of released high-frequency energy at $f > 0.1$ Hz which is better measured by M_e.

The quantity $\tau_a = \mu E_s/M_o$ is termed apparent stress [90]. It represents the dynamic component of stress acting on the fault during slip, which is responsible for the generation of radiated kinetic seismic wave energy E_s. On average it holds that $\tau_a \approx 2\Delta\sigma$ (with $\Delta\sigma$ = stress drop = difference between the stress in the source area before and after the earthquake rupture). Both τ_a and $\Delta\sigma$ depend strongly on the seismotectonic environment, i. e., the geologic-tectonic conditions, fault maturity and type of earthquake *source mechanisms* prevailing in seismically active regions [16,17,18,19]. However, $M_e \approx M_w$ holds only for $\tau_a \approx 0.6$ MPa.

Another important teleseismic magnitude is called mantle magnitude M_m. It uses surface waves with periods between about 60 s and 410 s that penetrate into the Earth's mantle. The concept has been introduced by Brune and Engen [13] and further developed by Okal and Talandier [64,65]. M_m is firmly related to the seismic moment M_o. Best results are achieved for $M_w > 6$ at distances > 15–$20°$ although the M_m procedure has been tested down to distances of 1.5° [77]. However, at $D < 3°$ the seismic sensors may be saturated in the case of big events. Also, at short distances one may not record the very long periods required for unsaturated magnitude estimates of very strong earthquakes, and for $M_w < 6$, the records may become too noisy at very long-periods. A signal-to-noise ratio larger than 3 is recommended for reliable magnitude estimates. M_m determinations have been automated at the PTWC and the CPPT [39,85] so that estimates are available in near real-time within about 10 min after OT from near stations, however typically within about half an hour, plus another few minutes for great earthquakes measured at the longest periods. Since M_m is determined at variable very long periods this magnitude does not – or only marginally – saturate even for very great, slow or complex earthquakes.

Common Magnitude Estimates for the Sumatra 2004 M_w 9.3 Earthquake

On 26 December 2004, the great Sumatra–Andaman Island earthquake with a rupture length of more than 1000 km occurred. It caused extensive damage in Northern Sumatra due to strong earthquake shaking. Moreover, it generated an Indian Ocean-wide tsunami with maximum run-up heights of more than 10 m. In total, this event claimed more than 200,000 victims and caused widespread damage on the shores of Sumatra, Thailand, India and Sri Lanka that were reached by the tsunami wave

within some 15 min to about two hour's time. This earthquake put the current procedures for magnitude determination to a hard test both in terms of the reliability and compatibility of calculated values and the timeliness of their availability to guide early warning and disaster management activities. Here we address only seismological aspects, not the additional major problems of inadequate global monitoring and insufficient regional communication and disaster response infrastructure. The earliest magnitudes reported by or made available to the Pacific Tsunami Warning Center (PTWC) were:

- $m_b > 7$, about 8 min after origin time (OT);
- $M_{wp} = 8.0$, available after some 12 minutes at the PTWC (including a magnitude-dependent correction [86]);
- somewhat later in Japan $M_{wp} = 8.2$ after magnitude-dependent correction [48];
- $M_m \geqslant 8.5$ at the PTWC about 45 min after OT, hours later upgraded to $M_m = 8.9$ by using mantle surface waves with longer periods ($T \approx 410$ s);
- a first surface-wave magnitude estimate $M_S = 8.5$, some 65 min after OT;
- $M_w = 8.9$ (later revised to 9.0) released by Harvard Seismology more than 6 h after OT.

Other available measurements were: $m_b = 5.7$ and $M_S = 8.3$ by the IDC of the CTBTO, $m_b = 7.0$, $M_S = 8.8$, $M_e = 8.5$ and another long-period P-wave based $M_w = 8.2$ by the NEIC. All these values were too small and mostly available only after several hours or days (e. g., IDC data). Weeks later, after the analysis of Earth's *fundamental modes* with periods up to 54 min and wavelength of several 1000 km, the now generally accepted value $M_w = 9.3$ was published [76]. Why were the other magnitude values all too low and/or too late?:

- m_b NEIC suffers from the combined effect of both spectral and time-window dependent saturation that we will discuss in more detail in Sect. "Magnitude Saturation and Biases Due to Earthquake Complexity";
- m_b IDC is even more affected by these saturation effects, because of the very short measurement time window of only 5 s after the first P-wave onset. In the case of the Sumatra 2004 earthquake, the first P-wave maximum occurred after some 80 s and another, with comparable amplitude, after about 330 s (cf. Fig. 8). Further, prior to m_b measurement, the IDC broadband data are filtered with a more narrow-band response peaked at even higher frequencies (3–4 Hz) than at NEIC (≈ 2.5 Hz) [11];
- The reported surface-wave magnitudes ranged between $M_S = 8.3$ (IDC), 8.8 (NEIC and Japan Meteorological Agency) and 8.9 (Beijing), i. e., some of them are close to the moment magnitudes. However, because of the late arrival of long-period teleseismic surface waves, good estimates are usually not available within 1–2 h after OT. This leaves a sufficient tsunami warning lead time only for shores more than 1000–2000 km away from the source.
- The NEIC P-wave moment magnitude $M_w = 8.2$ was to small because its procedure is, similar as for M_{wp} determinations, based on relatively short-period (typically $T < 25$ s) P-wave recordings and a single-source model (cf. Sect. "Magnitude Saturation and Biases Due to Earthquake Complexity").
- The preliminary $M_e = 8.5$, computed a few hours after the December 2004 earthquake agreed with the final M_e computed later using a more robust method [20]. Another algorithm simulating a near-real-time computation would have yielded $M_e = 8.3$. Yet M_e, by its very nature as an energy magnitude and because of the relation $E_s = \Delta\sigma/2\mu\, M_e$, will generally be smaller than M_w for slow, long duration earthquakes with low stress drop. This is often the case for shallow thrust earthquakes in subduction zones. Extreme examples are four well-known slow tsunami earthquakes of 1992 (Nicaragua; $M_w = 7.6, \Delta M_e = -0.9$), 1994 (Java; $M_w = 7.8, \Delta M_e = -1.3$), 2000 (New Britain Region; $\Delta M_w = 7.8, \Delta M_e = -1.0$) and 2006 (Java; $M_w = 7.7, \Delta M_e = -0.9$) [40].

Magnitude Saturation and Biases Due to Earthquake Complexity

Currently, the most common magnitude scales, especially those based on band-limited short-period data, still suffer *saturation*, e. g., the magnitudes m_b, M_L, m_B and M_S, which are typically measured at periods around 1 s, 2 s, 5–15 s and 20 s, respectively begin to saturate for moment magnitudes M_w larger than about 5.5, 6.5, 7.5 and 8.0. Earthquakes with $m_b > 6.5$, $M_L > 7.0$, $m_B > 8.0$ and $M_S > 8.5$ are rare been found due to saturation (cf. Fig. 4 and [44]). Magnitude saturation has two causes: spectral saturation and saturation due to insufficient time-window length for the amplitude measurements. Source complexity may cause additional biases between different magnitude scales.

Spectral Saturation of Magnitudes

Spectral saturation occurs when the magnitude-dependent *corner frequency* f_c (for energy-related magnitudes) or

the low-frequency plateau of displacement amplitudes (for moment magnitude) fall outside of the passband range of the seismographs *transfer function* or magnification curve (Fig. 3) of the recording seismograph or of the filter applied to broadband data before magnitude measurements are made. The reason for spectral saturation can best be explained by way of idealized average "source spectra" of ground displacement $u(f)$ and ground velocity $v(f)$ that have been corrected for the instrument response, for the decay of the wave amplitudes due to attenuation that is caused by internal friction and scattering of the seismic waves at heterogeneities of the Earth and for amplification effects at the surface of the receiver site. For better understanding such source spectra have been multiplied in Fig. 5 by the factor $4\pi r\rho\, v_{\mathrm{p,s}}^3/R_{\theta,\phi}^{p,s}$ given in Eq. (12) in order to get for the displacement amplitudes $u_0 =$ constant at $f < f_{\mathrm{c}}$ the related scalar seismic moment M_{o} (Fig. 5, left) and its time-derivative, the so-called moment rate (Fig. 5, right).

The shape of the source spectrum can be interpreted as follows: The critical wavelength, which corresponds to f_{c}, is $\lambda_{\mathrm{c}} = v_{\mathrm{p,s}}/f_{\mathrm{c}} = c_{\mathrm{m1}}\pi R = c_{\mathrm{m2}}\,(\mathrm{L}\times\mathrm{W})^{1/2}$ with $v_{p,s}$ – velocity of the P or S waves in the source region, depending on whether f_{c} relates to a P-wave or an S-wave spectrum, R-radius of a circular fault rupture model, L– length and W– width of a rectangular fault rupture model; c_{m1} and c_{m2} are model dependent constants. For very long fault ruptures, i. e., $L \gg W$, one can even write $\lambda_{\mathrm{c}} = c_{\mathrm{m3}}L$. Thus, λ_{c} is proportional to the linear dimension of the fault. For $f < f_{\mathrm{c}}$, λ_{c} becomes larger than the fault. Rupture details along the fault can then no longer be resolved and the fault is "seen" by these long wavelengths just as a point source. Therefore, all frequencies $f < f_{\mathrm{c}}$ have the same displacement amplitudes. Accordingly, M_{o}, which is proportional to the fault area and the average slip over the fault, has to be determined either in the spectral domain from the low-frequency asymptote u_{o} to the displacement spectrum or in the time domain by fitting synthetic long-period waves with $f < f_{\mathrm{c}}$ to observed ones that have been low-pass filtered in the same frequency range.

For radiated frequencies $f > f_{\mathrm{c}}$ with $\lambda < \lambda_{\mathrm{c}}$, the shape of the spectrum changes drastically. Related displacement amplitudes are then excited by successively smaller patches of the rupture plane. The area of the rupture elements decreases with the second order of their linear dimension. Accordingly, the generated displacement amplitudes are $A_{\mathrm{d}} \sim f^{-2}$, while the related velocity amplitudes $A_{\mathrm{v}} = A_{\mathrm{d}} 2\pi f$ decay only $\sim f^{-1}$. In the seismological literature this is usually called the ω^{-2} rupture model [3], based on the concept of similarity, which implies a constant stress drop independent of source size. More complicated rupture models yield a high-frequency amplitude decay $\sim \omega^{-3}$ [33,34] and even more rapid decays have sometimes been found in empirical data (up to 5th order). Steeper than ω^{-2} amplitude decay would further amplify the spectral saturation of magnitude data discussed below.

The Harvard standard procedure for M_{o} determination assumes a single point source model with a prescribed, triangular moment-rate function in the time domain (as an approximation to moment-rate curves such the ones shown in Fig. 7) as well a minimum period of 200 s for strong earthquakes with magnitudes > 8. Assuming an average rupture velocity of 2.5 km/s, this period would correspond to a wavelength of 500 km. This is much shorter than the total rupture length of more than 1100 km for the great Sumatra 2004 earthquake and explains why $M_{\mathrm{w}}(\mathrm{HRV}) = 9.0$ was smaller than the moment magnitude $M_{\mathrm{w}} = 9.3$ determined by using fundamental Earth's modes with periods of 1000 s and more [76].

The relationship between the two currently most common magnitudes, m_{b} and $M_{\mathrm{S}}(20)$, can be understood with reference to Fig. 5. m_{b} is measured in the period range $0.5 < T < 3$ s, typically around 1 s. This corresponds approximately to the *corner frequencies* of earthquakes with $M_{\mathrm{S}} \approx 3$ to 4.5. According to Utsu [83] this is equivalent to an m_{b} between about 3.5 and 5.0. For $M_{\mathrm{S}} < 4.5$ or $m_{\mathrm{b}} < 5$, m_{b} is thus likely to be determined from amplitude measurements near or below the *corner frequency* of the source spectrum. In that case m_{b} is a good measure of seismic moment. However, for larger magnitudes m_{b} samples spectral amplitudes well above f_{c}, resulting in systematically too small m_{b} values as compared to M_{S} and M_{w}. For great earthquakes this difference may reach 2 m.u. (Fig. 4). In contrast, $M_{\mathrm{S}}(20)$ is measured at periods around 20 s and thus saturates much later at values between about 8.5 to 9.

However, these arguments only hold on average. The stress drop $\Delta\sigma$ of individual events may vary by about 2 to 3 orders, as apparent stress τ_{a}, especially for earthquakes with $M_{\mathrm{w}} < 7.5$ [16,17]. According to the relation $M_{\mathrm{o}} = (16/7)\Delta\sigma R^3$ given by Keilis–Borok [52] this may change source radii R and associated f_{c} by about one order. As an example, the dotted curve in Fig. 5 shows the approximate seismic source spectrum for a well contained underground nuclear explosion (UNE) of an equivalent yield of 1 kt TNT which corresponds to a magnitude $m_{\mathrm{b}} \approx 4$. Its source volume is much smaller than that of an earthquake with same seismic moment. Hence the *corner frequency* of its source spectrum is not around 1 Hz but around 10 Hz. This is the reason why m_{b} determined from UNE records does not saturate, even for the strongest UNE ever tested with $m_{\mathrm{b}} \approx 7$. Moreover, Fig. 5 also illustrates that an earthquake and an UNE

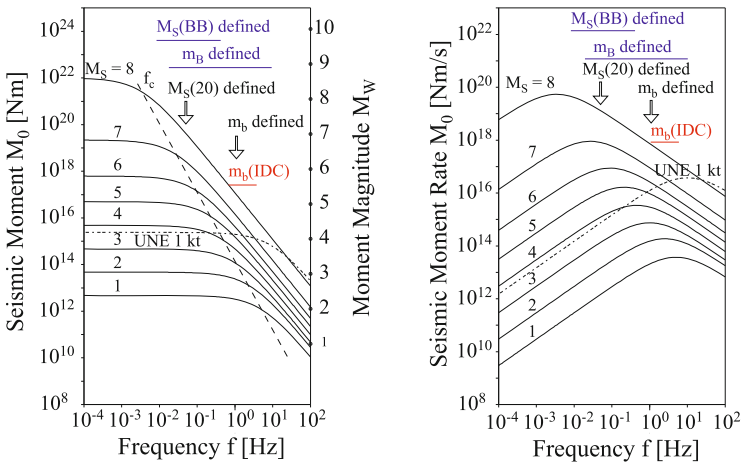

Earthquake Magnitude, Figure 5

"Source spectra" of ground displacement (*left*) and velocity (*right*) for an average single rupture seismic shear source, scaled on the left ordinates to *seismic moment* M_o (*left diagram*) and moment rate (*right diagram*), respectively. The black spectral lines have been scaled according to Aki [3] to integer surface-wave magnitudes M_S between 1 and 8. For reference the respective integer moment magnitude values M_w between 1 and 10, calculated according to Eq. (11), have been marked with equidistant dots on the right-side ordinate of the *left diagram*. The broken line shows the increase of the *corner frequency* f_c with decreasing seismic moment of the event, the dotted curve gives the approximate "source spectrum" for a well contained underground nuclear explosion (UNE) of an equivalent yield of 1 kt TNT. Note the plateau in the displacement spectrum towards low frequencies (corresponding to u_o = constant for $f < f_c$), from which M_o is determined according to Eq. (11) when using the frequency-domain approach. For $f > f_c$ the amplitudes decay $\sim f^{-2}$. The open arrows point to the center frequencies on the abscissa at which the 1Hz body-wave magnitude m_b and the 20 s surface-wave magnitude $M_s(20)$, respectively, are determined and the blue horizontal interval bars mark the range of frequencies within which the maximum P-wave and Rayleigh-wave amplitudes for m_b and M_s(BB) should be measured according to the new IASPEI standards [37]. In contrast, the red bar marks the frequency range of maximum velocity-proportional magnification of the bandpass filter between 1 Hz and 4 Hz which is used for m_b determination at the IDC.

with seismic moment around 4×10^{15} Nm and $M_w \approx 4$ have different maximum seismic moment-rate release at about 4×10^{15} and 4×10^{16} Nm/s, respectively. The latter corresponds to 100 times higher seismic energy release or to an energy magnitude M_e that is 1.3 m.u. larger. Large differences have also been observed amongst earthquakes, e. g., the Balleny Island earthquake of 25.03.1998 had M_w(HRV) = 8.1 and M_e(NEIC) = 8.8. The opposite will happen in the case of low stress drop earthquakes propagating with very low rupture velocity [38]. The Java tsunami earthquake of 17 July 2006 was a striking example with $M_e = 6.8$, $m_B = 7.0$ and $M_w = 7.7$.

Similar observations had already been made in the 1970s when comparing m_b and M_S values of identical events. This prompted the Russian scientist Prozorov to propose a "creepex" parameter $c = M_S - a \times m_b$ (with a = constant to be determined empirically for different source types and stress drop conditions). It aims at discriminating between normal, very slow (creeping) and explosion-like (fast rupture, high stress drop) earthquakes. World-wide determination of this parameter for earthquakes in different regions revealed interesting relations of c to source-geometry and tectonic origin [51]. Similar systematic regional differences were also reported for $M_S - M_w$ [24,67] and $M_e - M_w$ [16,19], suggesting systematic regional differences in stress drop.

Magnitude Saturation Due to Insufficient Time-Window Length for Amplitude Measurement

The second reason for magnitude saturation is insufficient time-window length for measuring $(A/T)_{max}$ in seismic records. It is most relevant when determining body-wave magnitudes, but it has been a subject of controversy, misconceptions and disregard of earlier recommendations for decades. The reason is that in teleseismic seismograms the P-wave group does not always appear sufficiently well separated in time from later phase arrivals such as the depth phases pP and sP. These do not directly travel from the seismic source at depth h to the recording station but travel first to the Earth's surface above the source and from there, after reflection or conversion from S to P, propagate back into the Earth. Depending on h, which may vary from a few kilometers up to 700 km, and the type of depth phase recorded, they may arrive from a few seconds up to about 4.5 min after the onset of direct P. Depending on the radi-

ation pattern of the *source mechanism*, some stations may even record the depth phases with larger amplitudes than the direct *P* wave. This is one of the concerns that led many researchers to propose measuring the *P*-wave amplitudes for magnitude measurements within a short time window after the *P* onset. On average, however, the depth phases have smaller amplitudes than *P* and will not bias m_b estimates at all. If, however, a seismic station is situated near to the nodal line of the so-called focal sphere, corresponding to strongly reduced *P*-wave radiation in these directions, the amplitude of the depth phase is a better estimator for the body-wave energy radiated by this seismic source and thus of its corresponding magnitude.

Two or three more phases of longitudinal waves may arrive close to the direct *P* at teleseismic distances between 20° and 100°. These include *PcP*, which results from *P*-wave energy reflected back from the surface of the Earth's core at 2900 km depth, and the phases *PP* and *PPP*, which are *P* waves that have been reflected back from the Earth's surface once at half-way or twice at 1/3- and 2/3-way between the seismic source and the recording station, respectively. However, in short-period records the amplitudes of *PP* and *PPP* are generally smaller and those of *PcP* even much smaller than the amplitudes of direct *P* waves. These later arrivals will therefore, never bias m_b estimates. Yet on broadband records *PP* may sometimes have equal or even slightly larger amplitudes than primary *P*. However, *P* and *PP* phases are usually well separated by more than 1 min (up to 4 min) and not likely misinterpreted. Only for rare large earthquakes with $M > 7.5$ the rupture duration and related *P*-wave radiation may extend into the time window where *PP* should arrive. But even then, wrongly taking PP_{max} for P_{max}, the bias in m_B estimate will not exceed 0.2 m.u. and usually be much smaller.

This experience from extensive seismogram analysis practice led Bormann and Khalturin [7] to state in 1974:

> … "*that the extension of the time interval for the measurement of* $(A/T)_{max}$ up to 15 or 25 sec., as proposed …in the Report of the first meeting of the IASPEI Commission on Practice (1972) … is not sufficient in all practical cases, especially not for the strongest earthquakes with $M > 7.5$ …".

This was taken into account in the Manual of Seismological Observatory Practice edited by Willmore [87]. It includes the recommendation to extend the measurement time window for *P*-wave magnitudes up to 60 s for very large earthquakes. But still, this has not yet become common practice (see Sect. "Introduction to Common Magnitude Scales: Potential and Limitations") although even a limit of 60 s may not be sufficient for extreme events such as the Sumatra M_w 9.3 earthquake when the first $P1_{max}$ appeared around 80 s and a second $P2_{max}$ of comparable amplitude at about 330 s after the first *P*-wave onset (cf. Fig. 8).

To allow a quick rough estimate of earthquake rupture duration τ_d as a function of magnitude we derived from extrapolation of data published in [66] the average relation

$$\log \tau_d \approx 0.6M - 2.8 \,. \tag{13}$$

It yields for $M = 6, 7, 8$ and 9 $\tau_d \approx 6\,\text{s}, 25\,\text{s}, 100\,\text{s}$ and 400 s, respectively. Measurement time windows of 5 s, 25 s or 60 s may therefore underestimate the magnitude of earthquakes with $M_w > 6, > 7$ or > 8, respectively. We call this effect the time-window component of magnitude saturation. It aggravates the pure spectral saturation component. To avoid this in future, the new IASPEI standards of amplitude measurements for m_b and m_B (cf. [42]) recommend to measure $(A/T)_{max} = A_{vmax}/2\pi$ in the entire *P*-phase train (time span *including P*, *pP*, *sP*, and possibly *PcP* and their codas but ending *preferably* before PP).

In fact the pioneers of the magnitude scales, Richter and Gutenberg, knew this, because they were still very familiar with the daily analysis of real seismic records and their complexity. Regrettably, they never wrote this down, with respect to magnitude measurements, in detail for easy reference. In the current era of automation and scientific depreciation of alleged "routine processes" the younger generation of seismologists usually had no chance to gather this experience themselves and occasionally introduced technologically comfortable but seismologically questionable practices. In an interview given in 1980 [75] Prof. Richter remembered that Gutenberg favored the body-wave scale in preference to the surface-wave scale because it is theoretically better founded. However, he said:

> "… it gives results comparable with Gutenberg's only if his procedure is closely followed. Experience has shown that misunderstanding and oversimplified misapplications can occur. For instance, magnitude is sometimes assigned on the first few waves of the *P* group rather than the largest *P* waves as Gutenberg did."

In order to avoid too-short measurement time windows when searching for the largest *P* amplitude one can estimate the rupture duration independently from the duration of large *P*-wave amplitudes in high-frequency filtered BB records because the generation of high-frequency waves is directly related to the propagation of the rupture

front. Thus one may find P_{max} for great earthquakes even beyond the theoretically expected *PP* arrival (cf. Fig. 8).

Magnitude Biases Due to Neglecting Multiple Source Complexity

Realizing that strong earthquakes usually consist of multiple ruptures Bormann and Khalturin [7] also wrote:

> "*In such cases we should determine the onset times and magnitudes of all clear successive P-wave onsets separately, as they give a first rough impression of the temporal and energetic development of the complex rupture process. … The magnitude* $MP = \log \Sigma_n(A_i/T_i) + Q(D, h)$ *(n is the number of successive P-wave onsets) could be considered as a more realistic measure of the P-wave energy released by such a multiple seismic event than the* m_b*-values from …* (single amplitude) $(A/T)_{max}$ *within the first five half cycles or within the whole P-wave group.*"

This magnitude, which is based on summed amplitudes in broadband records, is now called m_{Bc} [10], which stands for cumulative body-wave magnitude.

The 1985 $M_w = 8.1$ Mexico earthquake was a striking example for the development of such a multiple rupture process in space and time ([58], Fig. 6). A dense network of high-frequency strong-motion recordings near to the source revealed that the earthquake had a total rupture duration of about 60 s and consisted of two main sub-ruptures with significantly increased slip-velocities (12–32 cm/s). These two fault segments were separated in space by roughly 100 km in strike direction and ruptured between 10–22 s and 34–50 s after rupture start. Such a complicated rupture is not well represented by calculating the average slip and rupture velocity for a single point-source model. Also the *corner frequencies* related to these smaller sub-ruptures will be higher and not correspond to $(L \times W)^{-1/2}$ of the total rupture area.

Such multiple ruptures are not an exception but rather the rule for earthquakes with magnitudes above 7.5 (and often also for smaller ones, even down to events with magnitudes around 5.0). The detailed patterns of the respective moment-rate curves differ from event to event (Fig. 7). Often they can not be approximated by a single-source triangular moment-rate function, as commonly assumed in the standard procedure for moment tensor solutions practiced at Harvard [78] and other centers.

Therefore, the Harvard group [78] re-analyzed the data of the great Sumatra 2004 earthquake for which $M_w = 9.0$ had been calculated with the standard procedure. Interactively fitting synthetic records for five successive point sources to the observed mantle surface-wave data in the 200–500 s period range yielded the same value of $M_w = 9.3$ as derived by [76] for a single-source model but using much longer periods between 20 min and 54 min. In fact, the multiple Centroid Moment Tensor (CMT) source analysis applied in [78] resembles the concept proposed in [7] for *P*-wave magnitudes of strong earthquakes, but applied to long-period surface waves. Presently, a multiple CMT source analysis still requires human interaction and takes too much time for early warning applications. Possible alternative procedures such a the automatic line source inversion [21] have been developed but demonstrated so far only for earthquakes with magnitudes < 7 for which classical m_B, M_S or M_w do not saturate due to source complexity.

Proposals for Faster Magnitude Estimates of Strong Earthquakes

Soon after the great Sumatra earthquake of 2004 several authors suggested improvements to obtain more reliable and faster magnitude estimates of strong earthquakes. Menke and Levin [59] proposed to use a representative selection of 25 globally distributed high quality stations of the IRIS (Incorporated Research Institutions for Seismology) Global Seismic Network as a reference data base of available strong long-period master-event records with known M_w. In case of a new strong earthquake, a search for the nearest (within a few hundred kilometers) reference event in the data base is performed and waveforms are compared for a time window of about 30 min. By adding the $\log_{10}$ of the average amplitude ratio of the two events to the M_w of the master event, a moment magnitude estimate of the actual event is obtained. This procedure is based on the assumption of similarity of *source mechanisms* and radiation patterns, slip rates and stress drops, at least within the reference regions. The authors expect reasonably good magnitude estimates, with only small underestimation for events with $M_w > 8.6$. Thus warnings could be issued within about 40 min after OT (measurement time window plus travel-time to stations of a global network). This would still be relevant for distant coasts that might be affected by a tsunami. However, no data have been published until now that demonstrate the near-real-time operational capability of this procedure for a representative set of strong events.

Another approach by Lomax et al. [55,56,57] uses high-frequency seismograms ($f \geqslant 1$ Hz) that contain predominantly *P* signals radiated directly from the propagating rupture front and show little interference with later

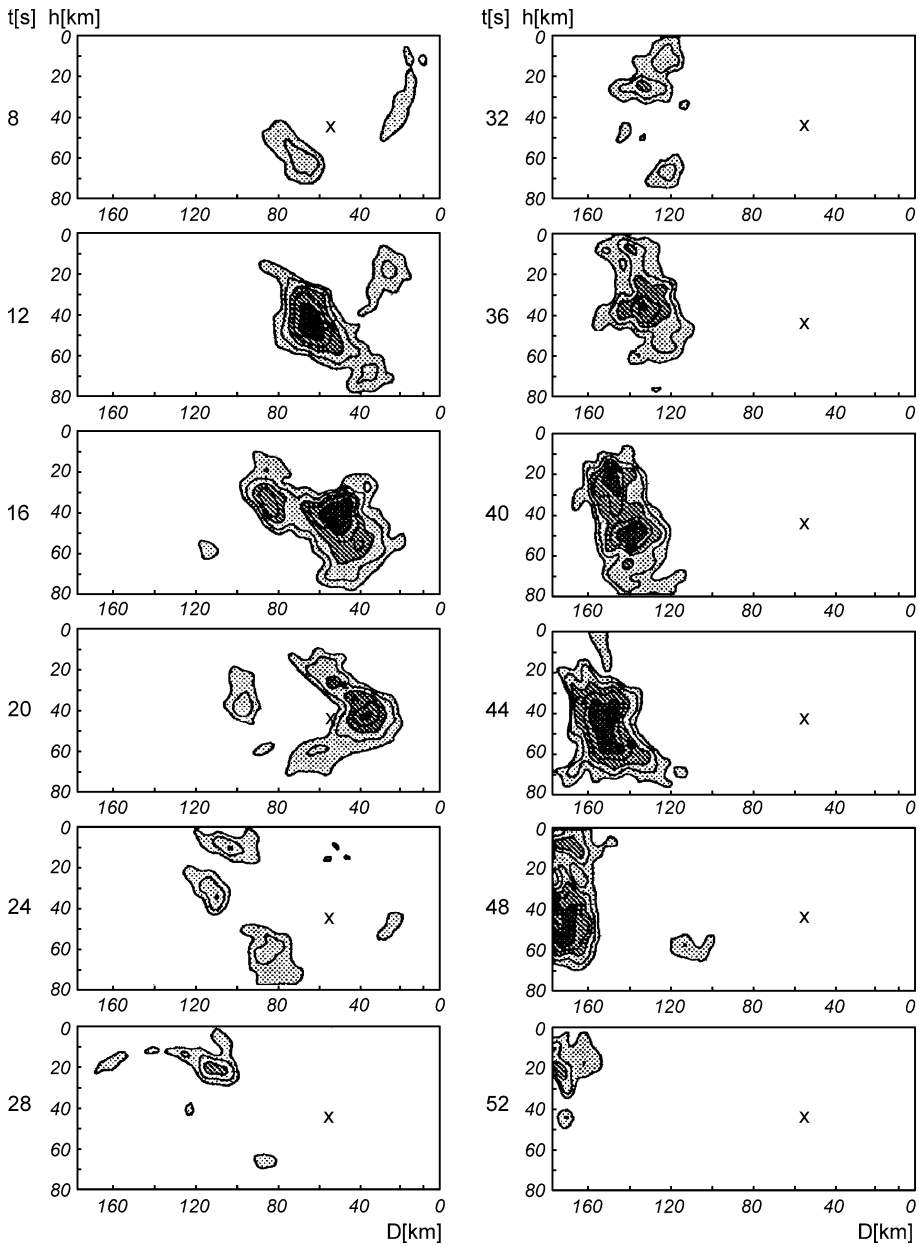

Earthquake Magnitude, Figure 6

Snapshots of the development in space and time of the inferred rupture process of the 1985 Michoácan, Mexico earthquake. The cross denotes the NEIC hypocenter position, the shading of the patches (from the outer part inwards *dotted*, *hatched* and *black*) relate to areas with velocities of dip slip (see *source mechanism*) in the ranges between 12 and 22 cm/s, 22 and 32 cm/s and greater than 32 cm/s. Redrawn and modified from Fig. 6 in [58]; taken from Fig. 3.8 in Vol. 1 of [6], © Seismological Society of America and IASPEI; with permission of the authors

secondary waves such as *PP* or *S*, thus providing a direct estimate of the rupture duration. Such recordings are available at teleseismic distances (30°–90°) within about 20 min after OT, even after strong events with long durations and provide an early picture of the total rupture process. When assuming constant rupture velocity and mean slip for stronger and weaker earthquakes, the seismic moment M_o and thus moment magnitude M_w could be estimated by comparing the actual rupture duration (averaged from observations at several seismic stations) with that of

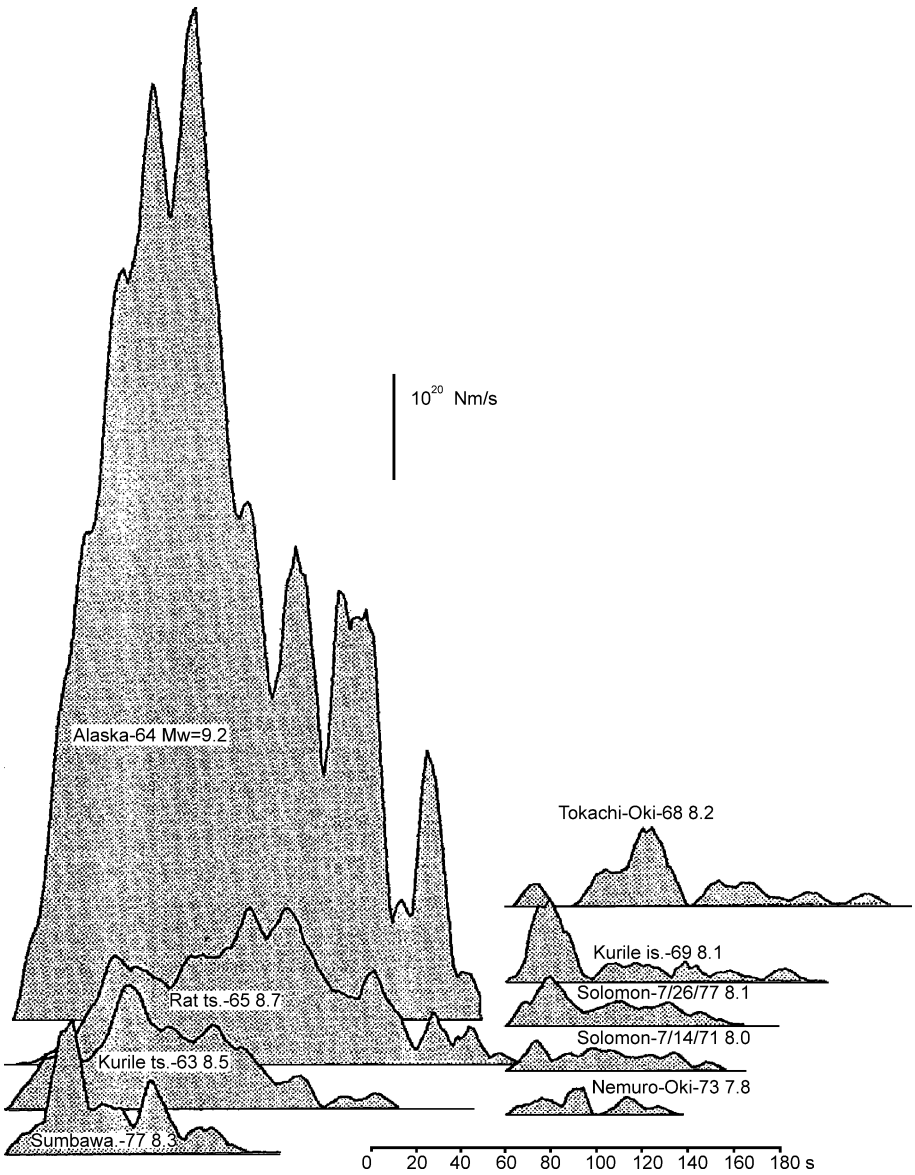

Earthquake Magnitude, Figure 7
Moment-rate functions for the largest earthquakes in the 1960 and 1970s (modified from Fig. 9, p. 1868 in [53]), taken from Fig. 3.7 in Vol. 1 of [6], © Seismological Society of America and IASPEI; with permission of the authors

a reference event with known M_o and rupture duration. This is conceptually similar to the approach in [59] but with high-frequency observations and the ratio of rupture duration instead of amplitudes.

However, Hara [32] demonstrated with a large data set of strong earthquakes that it is difficult to estimate earthquake size reliably only from durations t of high-frequency radiation. Therefore, he measured duration t in combination with the maximum displacement amplitude A_{dmax} within this time interval and derived the following empirical relation:

$$M = 0.79 \log A_{dmax} + 0.83 \log D + 0.69 \log t + 6.47 \quad (14)$$

with A_{dmax}, D and t in units of m, km and s, respectively. He applied Eq. (14) to 69 shallow earthquakes in the magnitude range $7.2 \leqslant M_w(\text{HRV}) \leqslant 9.0$ at distances between 30° and 85° and on average got a 1:1 relation between M_w(HRV) and his magnitude with a standard deviation

of 0.18 m.u. All event estimates were within ±0.5 m.u. of M_w(HRV), with the exception of the heavily underestimated Denali/Alaska earthquake of 3 November 2002 (7.1 instead of 7.8). This is a promising and simple procedure.

Bormann and Wylegalla [9] applied the earlier proposal in [7] to recordings with a velocity passband between 40 Hz and 125 s. They interactively summed up the maximum amplitudes of all visually discernible sub-ruptures in the recordings of several recent great earthquakes with $M_w \geqslant 8.3$, amongst them the tsunamigenic $M_w = 9.3$ Sumatra earthquake of 2004. For the latter they obtained a cumulative broadband body-wave magnitude $m_{Bc} = 9.3$ in records of just a single German station (RUE; $D = 82.5°$) at the time of the second major amplitude maximum, some 330 s after the first P onset and 18 min after OT. For three more events with magnitudes M_w 8.3, 8.4 and 8.6 they calculated m_{Bc} values of 8.4, 8.4 and 8.6, respectively, i. e., excellent agreement. Subsequently, 50 more earthquakes in the magnitude range 6 to 9 were analyzed interactively [10] with the following results:

- Average difference $m_B - M_w(\text{HRV}) = 0.00 \pm 0.27$ in the range $6.0 \leqslant M_w(\text{HRV}) < 8$. For magnitudes > 7.8–8, however, m_B tends to underestimate M_w, e. g., $m_B = 8.3$ for the Sumatra earthquake of 26 December 2004 based on the BB record of station RUE. Remarkably this m_B value is still very close to M_w, M_{wp} and M_e of the NEIC, which ranged between 8.2 and 8.5.
- The average difference $m_{Bc} - M_w(\text{HRV}) = +0.18 \pm 0.26$ in the range $6.0 \leqslant M_w(\text{HRV}) \leqslant 9.0$, i. e., m_{Bc} has a tendency to slightly overestimate M_w(HRV) on average, but not for $M_w > 8$ (see the four values above).

In [10] also first results of a fully automatic determination of m_B and m_{Bc} have been presented. The algorithm has been improved by incorporating automatic estimates of the rupture duration calculated from the envelope of the high-frequency P-wave radiation from filtered broadband records of globally distributed stations in a wide range of azimuths and source distances. In the case of strong earthquakes with long rupture duration this justifies the search for broadband P_{max} even beyond the onset of PP and to sum-up the amplitudes of major sub-ruptures over the whole rupture duration as defined above. Figure 8 gives an example for a BB record of the Sumatra earthquake of 26 December 2004. The largest P-wave amplitudes at about 80 s, 280 s and 330 s after the P onset each yield a single amplitude $m_B = 8.2$, whereas the cumulative magnitude $m_{Bc} = 9.3$ is in perfect agreement with the best moment magnitude estimates for this event.

The automatic algorithm for m_B and m_{Bc} determination has been in use since spring 2007 in the operational Indonesian prototype tsunami early warning system and yields online estimates of m_B. Before the implementation it had been tested whether the automatic procedure produces results that are comparable with those determined earlier interactively by two experienced seismogram analysts. Identical broadband records of 54 earthquakes in the magnitude range $6 \leqslant M_w(\text{HRV}) \leqslant 9$ were used for this comparison based on 138 m_B and 134 m_{Bc} values. The average difference between the interactively and automatically determined magnitudes was 0.03 and 0.02 m.u. with standard deviations of ±0.13 and ±0.19 m.u., respectively. This is in the range of other high-quality magnitude measurements. Even single station m_B and m_{Bc} estimates differed on average < 0.08 m.u. from average global network estimates based on up to hundreds of stations. Their standard deviations were $< \pm 0.25$ m.u. and decreased to ±0.10 m.u. for m_B and ±0.14 m.u. for m_{Bc} when just a few stations (between two and seven) were used to estimate the m_B and m_{Bc} event magnitudes. This documents both the reliability of the automatic procedure as well as the reliability of m_B and m_{Bc} estimates, even if derived from a few records of globally distributed high-fidelity stations. Thus, the automatic procedure is suitable for reproducibly determining the IASPEI recommended standard magnitude m_B and its proposed non-saturating extension m_{Bc} in near real-time. When using only observations in the distance range $21° \leqslant D° \leqslant 100°$ saturation-free teleseismic magnitude estimates of earthquakes with potential for strong shaking damage and tsunami generation could be made available in near real-time within about 4 to 18 min after OT, depending on epicentral distance and rupture duration.

Compared to other more theoretically based methods such as M_{wp} and M_w, the empirical $m_B - m_{Bc}$ method is free of any hypothesis or model assumptions about the rupture process (single or multiple source), type of rupture mechanism, rupture velocity, average slip and/or stress drop, complexity or simplicity of the moment-release function, etc. It just measures velocity amplitudes on the unfiltered broadband record, complex or not, sums them up over the duration of the rupture process and calibrates them with the classical empirical broadband Q_{PV} function (Fig. 2 and [30]). However, one has to consider that – in contrast to all types of moment magnitudes – m_B and m_{Bc} are not determined from the maximum long-period displacement amplitudes, but from the maximum velocity amplitudes. Therefore, m_B (for earthquakes with $M_w < 8.0$) and m_{Bc} (for earthquakes with $M_w > 7.8$) are better estimators than M_w for the seismic energy released

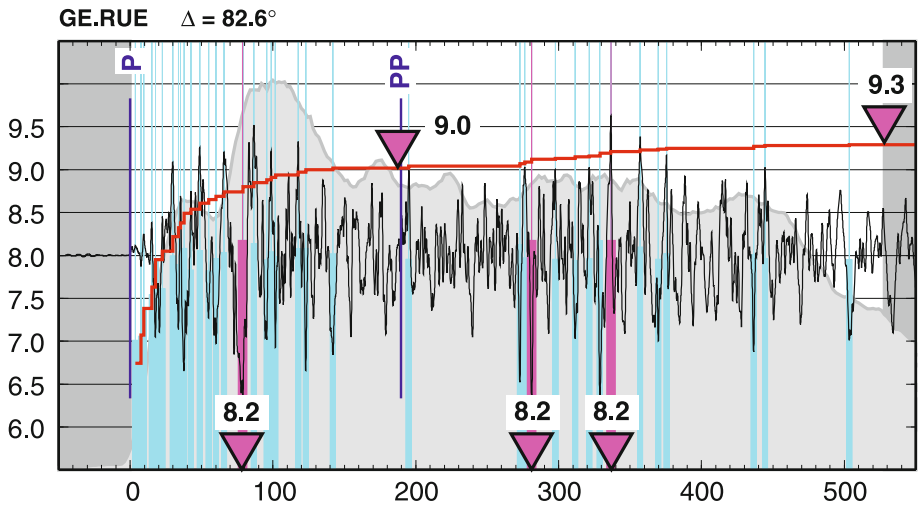

Earthquake Magnitude, Figure 8

Velocity broadband record at the Berlin station RUE in $D° = 82.6°$ epicentral distance of the great M_w9.3 tsunamigenic Sumatra earthquake of Dec. 26, 2004. The record is projected into a time-magnitude diagram as plotted by the automatic $m_B - m_{Bc}$ algorithm. The *red inverted triangles* mark the times and give the values of m_B for the three largest sub-ruptures. The *red step curve* shows the development of the cumulative magnitude m_{Bc} as a function of time. The *inverted red triangles* on this curve give the m_{Bc} before the onset of *PP* and at the end of the rupture process, about 530 s after the first *P*-wave onset, as estimated from the decay of the amplitude envelope of short-period filtered *P*-waves (see text)

by the earthquake and thus of its shaking-damage potential. Figure 9 compares m_B and m_{Bc} with M_w(HRV) for 76 earthquakes in the range $6 \leqslant M_w \leqslant 9.3$. The respective standard regression relations are:

$$M_w(\text{HRV}) = 1.22\, m_B - 1.54 \pm 0.29 \tag{15}$$

and

$$M_w(\text{HRV}) = 1.16\, m_{Bc} - 1.59 \pm 0.25 \tag{16}$$

These scaling relations allow much faster estimates of M_w than current routine standard M_w procedures. The rough moment estimates derived from m_B and m_{Bc} data, $M_w(m_B)$ or $M_w(m_{Bc})$, are sufficiently reliable for initial earthquake and tsunami alarms with standard deviations of the M_w estimates of about ± 0.29 and ± 0.25 m.u., respectively. However, looking into details of the somewhat irregular data scatter one realizes how *seismic source* complexity may "spoil" such regression relations. The five data points marked red in Fig. 9(left and right) are distinct outliers in the left diagram, i. e., the respective m_B values are 0.5 to 0.75 m.u. smaller than M_w(HRV) although usually m_B scales rather well with M_w(HRV) between $6.5 < m_B < 8.0$. These points correspond to slow earthquakes, one in Peru (1996) and four are tsunami earthquakes as mentioned at the end of Sect. "Common Magnitude Estimates for the Sumatra 2004 M_w 9.3 Earthquake". Their rupture durations ranged from about 100 s to 200 s, i.e, according to relationship (13) about 2–3 times longer than expected on average for their M_w value. Both m_B and M_e are usually much smaller than M_w for such events. In contrast, when calculating m_{Bc}, then these five data points all move close to the (not marked) 1:1 line in the m_{Bc}–M_w(HRV) diagram Fig. 9(right). Thus m_{Bc} becomes a good direct estimator of M_w for typical slow earthquakes, much better than via relation (16), which compensates for the usually too large m_{Bc} values of shallow depth and "normal" rupture earthquakes with $M_w < 8$. Thus, by determining rupture duration independently and treating very slow events separately, the standard deviation in relations (15) and (16) can be reduced. Moreover, the blue dots in Fig. 9 belong to very deep earthquakes with $h = 525$ km, 583 km and 631 km, respectively. Such deep earthquakes are "explosion-like" with comparably short rupture durations. Both m_B (for $M_w < 8$) and m_{Bc} yield values very close to M_w. In the m_{Bc}–M_w diagram, which is dominated by shallow and normal rupture earthquakes, such deep events appear as outliers. However, rapid event locations with good depth estimates allow one to identify such events and m_{Bc} (or m_B) should then be taken directly as estimator of M_w and not via relation (16).

M_{wp} has so far been the fastest operationally determined estimator of M_w. Comparably fast automatic m_B determination is now implemented, complementary to

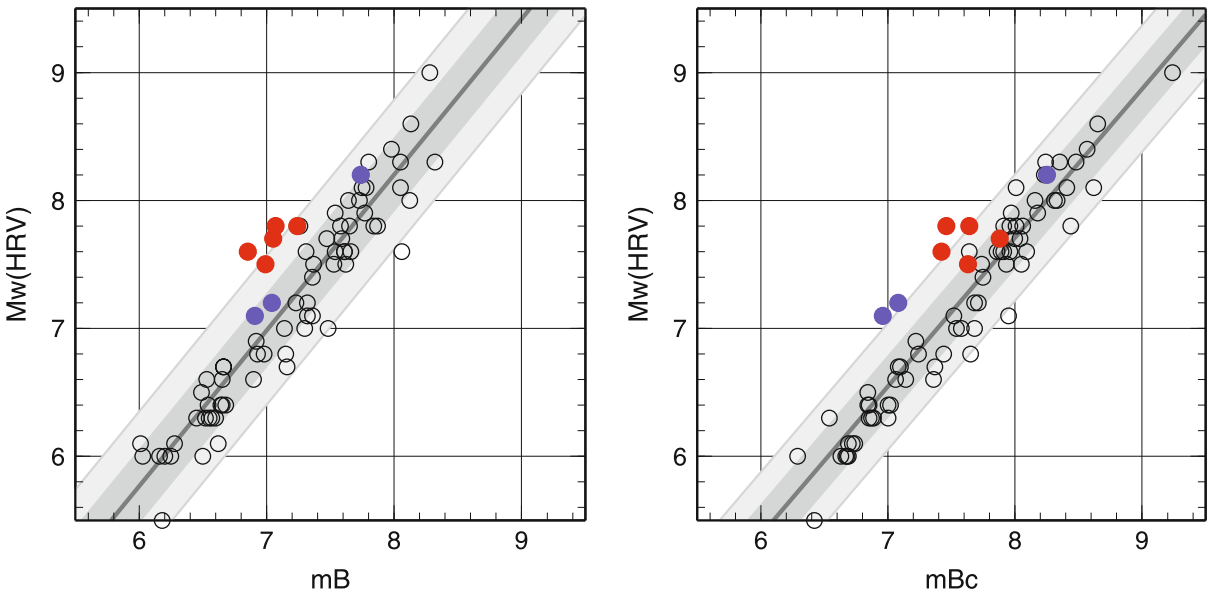

Earthquake Magnitude, Figure 9
Standard regression relationships of M_w(HRV) over m_B (*left*) and m_{Bc} (*right*). *Red dots* correspond to very slow earthquakes (Nicaragua 1992, Java 1994, New Britain Region 2000, Peru 2001 and Java 2006) and the *blue dots* belong to very deep earthquakes (Bolivia 1994, Philippines 2005 and Fiji Island 2006) with source depths $h = 631$ km, 525 km and 583 km, respectively. The *gray band* and the two *white bands* around the average *straight line* correspond to the width of one and two standard deviations in *y*-direction

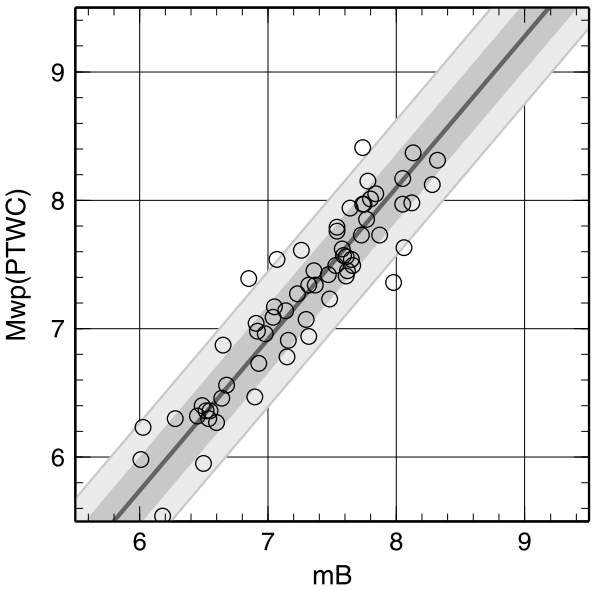

Earthquake Magnitude, Figure 10
Standard regression of M_{wp}(PTWC) over m_B. The standard deviations in *y*-direction are marked as in Fig. 9. The M_{wp} data have been kindly provided by the PTWC (courtesy of B. Hirshorn)

M_{wp}, in the German Indonesian Tsunami Early Warning System (GITEWS). Figure 10 compares the relation between m_B and M_{wp} for our test data set. These two magnitudes scale almost 1:1, following the standard regression relation:

$$M_{wp} = 1.08\, m_B - 0.638 \pm 0.24\,. \tag{17}$$

Future Requirements and Developments

Few national seismological data centers and stations report amplitude, period and/or magnitude data to the international data centers. The main reason is usually the lack of manpower to make competent measurements of these parameters interactively for the large amount of data recorded nowadays. Instrument responses of the seismographs used are sometimes not known accurately enough. There is, however, a growing practical and research need for such parameter data that have been determined according to international standards. Therefore, the most urgent requirements in the field of magnitudes are:

- Training of station and network operators to understand and practice proper magnitude measurements, instrument calibration and control;
- Implementation of the IASPEI magnitude standards [42];
- Making the tested and calibrated automatic algorithms available worldwide to data producers so that lack of

manpower is no longer a hindrance to mass-produce such data;

- Use of such standardized mass data with significantly reduced procedure-dependent errors for improved research into the attenuation properties of the Earth and deriving better magnitude calibration functions for all distance ranges;
- Comparison of magnitude data derived from identical record sets by applying both traditional and new standard measurement procedures and to derive standardized conversion relationships. This is a precondition for assuring long-term compatibility of magnitude data in national and international data catalogs and their usefulness for seismic hazard assessment and research;
- Improvement of current procedures for direct determination of seismic moment and energy in a wider magnitude range than currently possible, down to small magnitudes that are at present well covered only by M_L and m_b;
- Development of regional calibration functions for m_b and m_B, which will permit more reliable and much faster body-wave magnitude estimates from records at distances down to about 5°
- Development and consequent use of standard procedures for M_o and E_s measurements that assure non-saturating and globally compatible estimates of seismic moment and energy and of their related magnitude scales M_w and M_e;
- Use of these data for in-depth studies in the regional variability of apparent stress conditions and their relevance for improving (time-variable) regional earthquake and tsunami hazard and risk assessment;
- Comprehensive testing of speed and reliability of the various methods recently proposed for more rapid (near) real-time magnitude estimates (e. g. [9,10,32,46,55,56,57,59,61,66]) under operational EWS conditions;
- Development of faster automated procedures for direct non-saturating M_w and M_e determination for improving quick and realistic disaster response;
- Development of alternative automatic (near) real-time procedures of magnitude determination such as the rapid finite-source analysis [21], their scaling to both seismic energy and moment and operational testing also for very large earthquakes.

Bibliography

Primary Literature

1. Abe K (1981) Magnitudes of large shallow earthquakes from 1904 to 1980. Phys Earth Planet Int 27:72–92
2. Abe K (1984) Complements to Magnitudes of large shallow earthquakes from 1904 to 1980. Phys Earth Planet Int 34:17–23
3. Aki K (1967) Scaling law of seismic spectrum. J Geophys Res 72(4):1217–1231
4. Båth M (1981) Earthquake magnitude – recent research and current trends. Earth Sci Rev 17:315–398
5. Boatwright J, Choy GL (1986) Teleseismic estimates of the energy radiated by shallow earthquakes. J Geophys Res 91(B2):2095–2112
6. Bormann P (ed) (2002) IASPEI New manual of seismological observatory practice, vol 1 and 2. GeoForschungsZentrum, Potsdam
7. Bormann P, Khalturin V (1975) Relations between different kinds of magnitude determinations and their regional variations. In: Proceed XIVth General Ass European Seism. Comm Trieste Sept pp 16–22, 1974. Academy of Sciences of DDR, Berlin, pp 27–39
8. Bormann P, Baumbach M, Bock G, Grosser H, Choy GL, Boatwright J (2002) Seismic sources and source parameters. In: Bormann P (ed) IASPEI New manual seismological observatory practice. GeoForschungsZentrum Potsdam, chap 3, pp 1–94
9. Bormann P, Wylegalla K (2005) Quick estimator of the size of great earthquakes. EOS 86(46):464
10. Bormann P, Wylegalla K, Saul J (2006) Broadband body-wave magnitudes m_B and m_{Bc} for quick reliable estimation of the size of great earthquakes. http://spring.msi.umn.edu/USGS/Posters/
11. Bormann P, Liu R, Ren X, Gutdeutsch R. Kaiser D, Castellaro S (2007) Chinese national network magnitudes, their relation to NEIC magnitudes, and recommendations for new IASPEI magnitude standards. Bull Seism Soc Am 97(1B):114–127
12. Brune JN (1970) Tectonic stress and the spectra of shear waves from earthquakes. J Geophys Res 75:4997–5009
13. Brune JN, Engen GR (1969) Excitation of mantle Love waves and definition of mantle wave magnitude. Bull Seism Soc Am 49:349–353
14. Castellaro S, Mulargia F, Kagan YY (2006) Regression problems for magnitudes. Geophys J Int 165:913–930
15. Castellaro S, Bormann P (2007) Performance of different regression procedures on the magnitude conversion problem. Bull Seism Soc Am 97:1167–1175
16. Choy GL, Boatwright J (1995) Global patterns of radiated seismic energy and apparent stress. J Geophys Res 100, B9:18,205–18,228
17. Choy GL, Boatwright J, Kirby SH (2001) The radiated seismic energy and apparent stress of interplate and intraslab earthquakes at subduction zone environments: Implications for seismic hazard estimation. US Geological Survey Open-File Report 01–0005:18
18. Choy GL, Kirby S (2004) Apparent stress, fault maturity and seismic hazard for normal-fault earthquakes at subduction zones. Geophys J Int 159:991–1012
19. Choy GL, McGarr A, Kirby SH, Boatwright J (2006) An overview of the global variability in radiated energy and apparent stress. In: Abercrombie R, McGarr A, Kanamori H (eds) Radiated energy and the physics of earthquake faulting, AGU. Geophys Monogr Ser 170:43–57
20. Choy GL, Boatwright J (2007) The energy radiated by the 26 December 2004 Sumatra-Andaman earthquake estimated from 10-minute *P*-wave windows. Bull Seism Soc Am 97:18–24

21. Dregers DS, Gee L, Lombard P, Murray MH, Romanowicz B (2005) Rapid finite-source analysis and near-fault strong ground motions: Application to the 2003 M_w6.5 San Simeon and 2004 M_w6.0 Parkfield earthquakes. Seism Res Lett 76:40–48
22. Duda SJ (1965) Secular seismic energy release in the circum-Pacific belt. Tectonophysics 2:409–452
23. Dziewonski AM, Chou TA, Woodhous JH (1981) Determination of earthquake source parameters from waveform data for studies of global and regional seismicity. J Geophys Res 86:2825–2852
24. Ekström G, Dziewonski AM (1988) Evidence of bias in estimations of earthquake size. Nature 332:319–323
25. Geller RJ, Kanamori H (1977) Magnitudes of great shallow earthquakes from 1904 to 1952. Bull Seism Soc Am 67:587–598
26. Gutenberg B (1945) Amplitudes of P, PP, and S and magnitude of shallow earthquakes. Bull Seism Soc Am 35:57–69
27. Gutenberg B (1945) Magnitude determination of deep-focus earthquakes. Bull Seism Soc Am 35:117–130
28. Gutenberg B (1945) Amplitude of surface waves and magnitude of shallow earthquakes. Bull Seism Soc Am 35(3):3–12
29. Gutenberg B, Richter CF (1954) Seismicity of the earth and associated phenomena, 2nd Edn. Princeton University Press, Princeton
30. Gutenberg B, Richter CF (1956) Magnitude and energy of earthquakes. Ann Geofis 9:1–15
31. Hanks TC, Kanamori H (1979) A moment magnitude scale. J Geophys Res 84(B5):2348–2350
32. Hara T (2007) Measurement of duration of high-frequency energy radiation and its application to determination of magnitudes of large shallow earthquakes. Earth Planet Space 59:227–231
33. Haskell N (1964) Total energy and energy spectral density of elastic wave radiation from propagating faults, vol 1. Bull Seismol Soc Am 54:1811–1842
34. Haskell N (1964) Total energy and energy spectral density of elastic wave radiation from propagating faults, vol 2. Bull Seismol Soc Am 56:125–140
35. Hatzidimitriou P, Papazachos C, Kiratzi A, Theodulidis N (1993) Estimation of attenuation structure and local earthquake magnitude based on acceleration records in Greece. Tectonophysics 217:243–253
36. Herak M, Herak D (1993) Distance dependence of M_S and calibrating function for 20 second Rayleigh waves. Bull Seism Soc Am 83:1881–1892
37. Herak M, Panza GF, Costa G (2001) Theoretical and observed depth corrections for M_S. Pure Appl Geophys 158:1517–1530
38. Hutton LK, Boore DM (1987) The M_L scale in Southern California. Bull Seism Soc Am 77:2074–2094
39. Hyvernaud O, Reymond D, Talandier J, Okal EA (1993) Four years of automated measurements of seismic moments at Papeete using the mantle magnitude M_m: 1987–1991. In: Duda SJ, Yanovskaya TB (eds) Special section: Estimation of earthquake size. Tectonophysics 217:175–193. Elsevier Science
40. http://neic.usgs.gov./neis/sopar
41. http://www.globalcmt.org/CMTsearch.html
42. http://www.iaspei.org/commissions/CSOI/Summary_of_WG_recommendations.pdf
43. Kanamori H (1977) The energy release in great earthquakes. J Geophys Res 82:2981–2987
44. Kanamori H (1983) Magnitude scale and quantification of earthquakes. Tectonophysics 93:185–199
45. Kanamori H (1988) The importance of historical seismograms for geophysical research. In: Lee WHK (ed) Historical seismograms and earthquakes of the world. Academic Press, New York, pp 16–33
46. Kanamori H (2005) Real-time seismology and earthquake damage prediction. Ann Rev Earth Planet Sci 33:195–214
47. Kanamori H, Hauksson E, Heaton T (1997) Real-time seismology and earthquake hazard mitigation. Nature 390:461–464
48. Kanjo K, Furudate T, Tsuboi S (2006) Application of M_{wp} to the great December 26, 2004 Sumatra earthquake. Earth Planet Space 58:121–126
49. Katsumata M (1964) A method of determination of magnitude for near and deep-focus earthquakes (in Japanese with English abstract). A J Seism 22:173–177
50. Katsumata M (1996) Comparison of magnitudes estimated by the Japan Meteorological Agency with moment magnitudes for intermediate and deep earthquakes. Bull Seism Soc Am 86:832–842
51. Kaverina AN, Lander AV, Prozorov AG (1996) Global creepex distribution and its relation to earthquake – source geometry and tectonic origin. Geophys J Int 135:249–265
52. Keilis-Borok VI (1959) On the estimation of displacement in an earthquake source and of source dimension. Ann Geofis 12:205–214
53. Kikuchi M, Ishida M (1993) Source retrieval for deep local earthquakes with broadband records. Bull Seism Soc Am 83:1855–1870
54. Lee V, Trifunac M, Herak M, Živèic M, Herak D (1990) M_L^{SM} computed from strong motion accelerograms recorded in Yugoslavia. Earthq Eng Struct Dyn 19:1167–1179
55. Lomax A (2005) Rapid estimation of rupture extent for large earthquakes: Application to the 2004, M9 Sumatra-Andaman mega-thrust. Geophys Res Lett 32:L10314
56. Lomax A, Michelini A (2005) Rapid determination of earthquake size for hazard warning. EOS 86(21):202
57. Lomax A, Michelini A, Piatanesi A (2007) An energy-duration procedure for rapid and accurate determination of earthquake magnitude and tsunamigenic potential. Geophys J Int 170:1195–1209
58. Mendez AJ, Anderson JG (1991) The temporal and spatial evolution of the 19 September 1985 Michoacan earthquake as inferred from near-source ground-motion records. Bull Seism Soc Am 81:1655–1673
59. Menke W, Levin R (2005) A strategy to rapidly determine the magnitude of great earthquakes. EOS 86(19):185–189
60. Nakamura Y (1989) Earthquake alarm system for Japan railways. Jpn Railw Eng 109:1–7
61. Nakamura Y, Saita J (2007) UrEDAS, the earthquake warning system: today and tomorrow. In: Gasperini P, Manfredi G, Zschau J (eds) Earthquake early warning systems. Springer, Berlin, pp 249–281
62. Nuttli OW (1986) Yield estimates of Nevada test site explosions obtained from seismic Lg waves. J Geophys Res 91:2137–2151
63. Okal EA (1989) A theoretical discussion of time domain magnitudes: The Prague formula for M_S and the mantle magnitude M_m. J Geophys Res 94:4194–4204
64. Okal EA, Talandier J (1989) M_m: A variable-period mantle magnitude. J Geophys Res 94:4169–4193

65. Okal EA, Talandier J (1990) M_m: Extension to Love waves of the concept of a variable-period mantle magnitude. Pure Appl Geophys 134:355–384
66. Olson EL, Allen R (2005) The deterministic nature of earthquake rupture. Nature 438:212–215
67. Patton HJ (1998) Bias in the centroid moment tensor for central Asian earthquakes: Evidence from regional surface wave data. J Geophys Res 103(26):885–898
68. Polet J, Kanamori H (2000) Shallow subduction zone earthquakes and their tsunamigenic potential. Geophys J Int 142:684–702
69. Purcaru G, Berckhemer H (1978) A magnitude scale for very large earthquakes. Tectonophysics 49:189–198
70. Rezapour M, Pearce RG (1998) Bias in surface-wave magnitude M_S due to inadequate distance corrections. Bull Seism Soc Am 88:43–61
71. Richter CF (1935) An instrumental earthquake magnitude scale. Bull Seism Soc Am 25:1–32
72. Richter CF (1958) Elementary seismology. W.H. Freeman, San Francisco
73. Rydelek P, Horiuchi S (2006) Is earthquake rupture deterministic? Nature 444:E5–E6
74. Soloviev SL (1955) Classification of earthquakes by energy value (in Russian). Trudy Geophys Inst Acad Sci USSR 39(157):3–31
75. Spall H (1980) Charles F. Richter – an interview. Earthq Inf Bull 12(1):5–8
76. Stein S, Okal E (2005) Speed and size of the Sumatra earthquake. Nature 434:581–582
77. Talandier J, Okal EA (1992) One-station estimates of seismic moments from the mantle magnitude M_m: The case of the regional field ($1.5° \leqslant \Delta \leqslant 15°$). Pure Appl Geophys 138:43–60
78. Tsai VC, Nettles M, Ekström G, Dziewonski AM (2005) Multiple CMT source analysis of the 2004 Sumatra earthquake. Geophys Res Lett 32(L17304):1–4
79. Tsuboi C (1954) Determination of the Gutenberg-Richter's magnitude of earthquakes occurring in and near Japan (in Japanese with English abstract). Zisin Second Ser 7:185–193
80. Tsuboi S, Abe K, Takano K, Yamanaka Y (1995) Rapid determination of M_w from broadband P waveforms. Bull Seism Soc Am 85:606–613
81. Tsuboi S, Whitmore PM, Sokolovski TJ (1999) Application of M_{wp} to deep and teleseismic earthquakes. Bull Seism Soc Am 89:1345–1351
82. Uhrhammer RA, Collins ER (1990) Synthesis of Wood-Anderson seismograms from broadband digital records. Bull Seism Soc Am 80:702–716
83. Utsu T (2002) Relationships between magnitude scales. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of earthquake and engineering seismology, Part A. Academic Press, Amsterdam, pp 733–746
84. Vanek J, Zátopek A, Kárnik V, Kondorskaya N, Riznichenko Y, Savarenski S, Solovév S, Shebalin N (1962) Standardization of magnitude scales. Izv Acad Sci USSR, Geophys Ser, pp 108–111 (English translation)
85. Weinstein S, Okal E (2005) The mantle magnitude M_m and the slowness parameter Θ: Five years of real-time use in the context of tsunami warning. Bull Seism Soc Am 95:779–799
86. Whitmore PM, Tsuboi S, Hirshorn B, Sokolowski TJ (2002) Magnitude-dependent correction for M_{wp}. Sci Tsunami Hazard 20(4):187–192
87. Willmore PL (ed) (1979) Manual of seismological observatory practice, World data center A for solid earth geophysics. Report SE–20. Boulder, Colorado
88. Wu Z (2001) Scaling of apparent stress from broadband radiated energy catalogue and seismic moment catalogue and its focal mechanism dependence. Earth Planets Space 53:943–948
89. Wu KM, Kanamori H (2005) Experiment on an onsite early warning method for the Taiwan early warning system. Bull Seism Soc Am 95:347–353
90. Wyss M, Brune JN (1968) Seismic moment, stress, and source dimensions for earthquakes in the California-Nevada regions. J Geophys Res 73:4681–4694

Books and Reviews

Båth M (1979) Introduction to seismology. Birkhäuser, Basel

Bolt BA (1999) Earthquakes, 4th edn. W.H. Freeman, San Francisco

Duda S, Aki K (Eds) (1983) Quantification of earthquakes. Tectonophysics 93(3/4):183–356

Gasperini P, Manfredi G, Zschau J (eds) (2007) Earthquake early warning systems. Springer, Berlin

Kulhánek O (1990) Anatomy of seismograms. Developments in solid earth geophysics, vol 18. Elsevier, Amsterdam

Lay T, Wallace TC (1995) Modern global seismology. Academic Press, San Diego

Lee WHK (ed) (1988) Historical seismograms and earthquakes of the world. Academic Press, New York

Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) (2002) International handbook of earthquake and engineering seismology, part A and B. Academic Press, London (an imprint of Elsevier Science)

Scholz CH (2002) The mechanics of earthquake faulting, 2nd edn. Cambridge University Press, Cambridge

Shearer PM (1999) Introduction to seismology. Cambridge University Press, Cambridge

Stein S, Wysession M (2002) Introduction to seismology, Earthquakes and earth structure. Blackwell Publishing, Malden

Earthquake Monitoring and Early Warning Systems

William H. K. Lee[1], Yih-Min Wu[2]

[1] US Geological Survey, Menlo Park, USA

[2] Department of Geosciences, National Taiwan University, Taipei, Taiwan

Article Outline

Glossary

Definition of the Subject

Introduction

Earthquake Monitoring: Instrumentation

Earthquake Monitoring:
Regional and Local Networks
Seismograms and Derived Products
Earthquake Early Warning (EEW) Systems
Future Directions
Acknowledgments
Appendix: A Progress Report on Rotational Seismology
Bibliography

Glossary

Active fault A *fault* (q.v.) that has moved in historic (e. g., past 10,000 years) or recent geological time (e. g., past 500,000 years).

Body waves Waves which propagate through the interior of a body. For the Earth, there are two types of seismic body waves: (1) compressional or longitudinal (*P* wave), and (2) shear or transverse (*S* wave).

Coda waves Waves which are recorded on a *seismogram* (q.v.) after the passage of *body waves* (q.v.) and *surface waves* (q.v.). They are thought to be back-scattered waves due to the Earth's inhomogeneities.

Earthquake early warning system (EEWS) An earthquake monitoring system that is capable of issuing warning message after an earthquake occurred and before strong ground shaking begins.

Earthquake precursor Anomalous phenomenon preceding an earthquake.

Earthquake prediction A statement, in advance of the event, of the time, location, and *magnitude* (q.v.) of a future earthquake.

Epicenter The point on the Earth's surface vertically above the *hypocenter* (q.v.).

Far-field Observations made at large distances from the *hypocenter* (q.v.), compared to the wave-length and/or the source dimension.

Fault A fracture or fracture zone in the Earth along which the two sides have been displaced relative to one another parallel to the fracture.

Fault slip The relative displacement of points on opposite sides of a *fault* (q.v.), measured on the fault surface.

Focal mechanism A description of the orientation and sense of slip on the causative fault plane derived from analysis of *seismic waves* (q.v.).

Hypocenter Point in the Earth where the rupture of the rocks originates during an earthquake and *seismic waves* (q.v.) begin to radiate. Its position is usually determined from arrival times of *seismic waves* (q.v.) recorded by *seismographs* (q.v.).

Intensity, earthquake Rating of the effects of earthquake vibrations at a specific place. Intensity can be estimated from instrumental measurements, however, it is formally a rating assigned by an observer of these effects using a descriptive scale. Intensity grades are commonly given in Roman numerals (in the case of the Modified Mercalli Intensity Scale, from I for "not perceptible" to XII for "total destruction").

Magnitude, earthquake Quantity intended to measure the size of earthquake at its source, independent of the place of observation. *Richter magnitude* (M_L) was originally defined in 1935 as the logarithm of the maximum amplitude of seismic waves in a seismogram written by a Wood–Anderson seismograph (corrected to) a distance of 100 km from the epicenter. Many types of magnitudes exist, such as *body-wave magnitude* (m_b), *surface-wave magnitude* (M_S), and *moment magnitude* (M_W).

Moment tensor A symmetric second-order tensor that characterizes an internal seismic point source completely. For a finite source, it represents a point source approximation and can be determined from the analysis of *seismic waves* (q.v.) whose wavelengths are much greater than the source dimensions.

Near-field A term for the area near the causative rupture of an earthquake, often taken as extending a distance from the rupture equal to its length. It is also used to specify a distance to a seismic source comparable or shorter than the wavelength concerned. In engineering applications, near-field is often defined as the area within 25 km of the fault rupture.

Plate tectonics A theory of global *tectonics* (q.v.) in which the Earth's lithosphere is divided into a number of essentially rigid plates. These plates are in relative motion, causing earthquakes and deformation along the plate boundaries and adjacent regions.

Probabilistic seismic hazard analysis Available information on earthquake sources in a given region is combined with theoretical and empirical relations among earthquake *magnitude* (q.v.), distance from the source, and local site conditions to evaluate the exceedance probability of a certain ground motion parameter, such as the peak ground acceleration, at a given site during a prescribed time period.

Seismic hazard Any physical phenomena associated with an earthquake (e. g., ground motion, ground failure, liquefaction, and tsunami) and their effects on land use, man-made structure, and socio-economic systems that have the potential to produce a loss.

Seismic hazard analysis The calculation of the *seismic hazard* (q.v.), expressed in probabilistic terms (See *probabilistic seismic hazard analysis*, q.v.). The result is usually displayed in a *seismic hazard map* (q.v.).

Seismic hazard map A map showing contours of a specified ground-motion parameter or response spectrum ordinate for a given *probabilistic seismic hazard* analysis (q.v.) or return period.

Seismic moment The magnitude of the component couple of the double couple that is the point force system equivalent to a *fault slip* (q.v.) in an isotropic elastic body. It is equal to rigidity times the fault slip integrated over the fault plane. It can be estimated from the far-field seismic spectrum at wave lengths much longer than the source size. It can also be estimated from the near-field seismic, geologic and geodetic data. Also called "scalar seismic moment" to distinguish it from *moment tensor* (q.v.).

Seismic risk The risk to life and property from earthquakes.

Seismic wave A general term for waves generated by earthquakes or explosions. There are many types of seismic waves. The principle ones are *body waves* (q.v.), *surface waves* (q.v.), and *coda waves* (q.v.).

Seismograph Instrument which detects and records ground motion (and especially vibrations due to earthquakes) along with timing information. It consists of a *seismometer* (q.v.) a precise timing device, and a recording unit (often including telemetry).

Seismogram Record of ground motions made by a *seismograph* (q.v.).

Seismometer Inertial sensor which responds to ground motions and produces a signal that can be recorded.

Source parameters of an earthquake The parameters specified for an earthquake source depends on the assumed earthquake model. They are origin time, *hypocenter* (q.v.), *magnitude* (q.v.), *focal mechanism* (q.v.), and *moment tensor* (q.v.) for a point source model. They include fault geometry, rupture velocity, stress drop, slip distribution, etc. for a finite fault model.

Surface waves Waves which propagate along the surface of a body or along a subsurface interface. For the Earth, there are two common types of seismic surface waves: Rayleigh waves and Love waves (both named after their discoverers).

Tectonics Branch of Earth science which deals with the structure, evolution, and relative motion of the outer part of the Earth, the lithosphere. The lithosphere includes the Earth's crust and part of the Earth's upper mantle and averages about 100 km thick. See *plate tectonics* (q.v.).

Teleseism An earthquake at an epicentral distance greater than about 20° or 2000 km from the place of observation.

Definition of the Subject

When a sudden rupture occurs in the Earth, elastic (seismic) waves are generated. When these waves reach the Earth's surface, we may feel them as a series of vibrations, which we call an earthquake. *Seismology* is derived from the Greek word σεισμός (seismos or earthquake) and λόγος (logos or discourse); thus, it is the science of earthquakes and related phenomena. Seismic waves can be generated naturally by earthquakes or artificially by explosions or other means. We define earthquake monitoring as a branch of seismology, which systematically observes earthquakes with instruments over a long period of time.

Instrumental recordings of earthquakes have been made since the later part of the 19th century by seismographic stations and networks of various sizes from local to global scales. The observed data have been used, for example, (1) to compute the source parameters of earthquakes, (2) to determine the physical properties of the Earth's interior, (3) to test the theory of plate tectonics, (4) to map active faults, (5) to infer the nature of damaging ground shaking, and (6) to carry out seismic hazard analyzes. Constructing a satisfactory theory of the complex earthquake process has not yet been achieved within the context of physical laws, e. g., realistic equations for modeling earthquakes do not exist at present. Good progress, however, has been made in building a physical foundation for the earthquake source process [62], partly as a result of research directed toward earthquake prediction.

Earthquakes release large amounts of energy that potentially can cause significant damage and human deaths. During an earthquake, potential energy (mainly elastic strain energy and some gravitational energy) that has accumulated in the hypocentral region over decades to centuries or longer is released suddenly [63]. This energy is partitioned into (1) radiated energy in the form of propagating seismic waves, (2) energy consumed in overcoming fault friction, (3) the energy which expands the rupture surface area or changes its properties (e. g., by pulverizing rock), and (4) heat. The radiated seismic energy is a small fraction (about 7%) of the total energy budget, and it can be estimated using the recorded seismograms. Take, for example, the 1971 San Fernando earthquake ($M_W = 6.6$) in southern California. Its radiated energy was about 5×10^{21} ergs, or about 120 kilotons of TNT explosives, or the energy released by six atomic bombs of the size used in World War II. The largest earthquake recorded instrumentally (so far) is the 1960 Chilean earthquake ($M_W = 9.5$). Its radiated energy was about 1.1×10^{26} ergs, an equivalent of about 2,600 megatons of TNT explosives, the energy released by about 130,000

atomic bombs. It is, therefore, no surprise that an earthquake can cause up to hundreds of thousands of human deaths, and produce economic losses of up to hundreds of billions of dollars.

Monitoring earthquakes is essential for providing scientific data to investigate complex earthquake phenomena, and to mitigate seismic hazards. The present article is a brief overview of earthquake monitoring and early warning systems; it is intended for a general scientific audience, and technical details can be found in the cited references. Earthquakes are complex natural phenomena and their monitoring requires an interdisciplinary approach, including using tools from computer science, electrical and electronic engineering, mathematics, physics, and others. Earthquake early warning systems (which are based on earthquake monitoring) offer practical information for reducing seismic hazards in earthquake-prone regions.

After the "Introduction", we will present a summary of earthquake monitoring, a description of the products derived from the analysis of seismograms, and a discussion of the limitations of these products. Earthquake early warning systems are then presented briefly, and we conclude with a section on future directions, including a progress report on rotational seismology (Appendix). We present overviews of most topics in earthquake monitoring, and an extensive bibliography is provided for additional reading and technical details.

Introduction

Earthquakes, both directly and indirectly, have caused much suffering to mankind. During the 20th century alone about two million people were killed as a result of earthquakes. A list of deadly earthquakes (death tolls ≥ 25) of the world during the past five centuries was compiled by Utsu [115]. It shows that earthquakes of magnitude ≥ 6 (~ 150 per year worldwide) can be damaging and deadly if they occur in populated areas, and if their focal depths are shallow (< 50 km). Seismic risk can be illustrated by plotting the most deadly earthquakes of the past five centuries (1500–2000) over a map of current population density. This approach was used by Utsu [115], and his result is shown in Fig. 1. Most of these deadly earthquakes are concentrated (1) along the coasts of Central America, the Caribbean, western South America, and Indonesia, and (2) along a belt that extends from southern Europe,

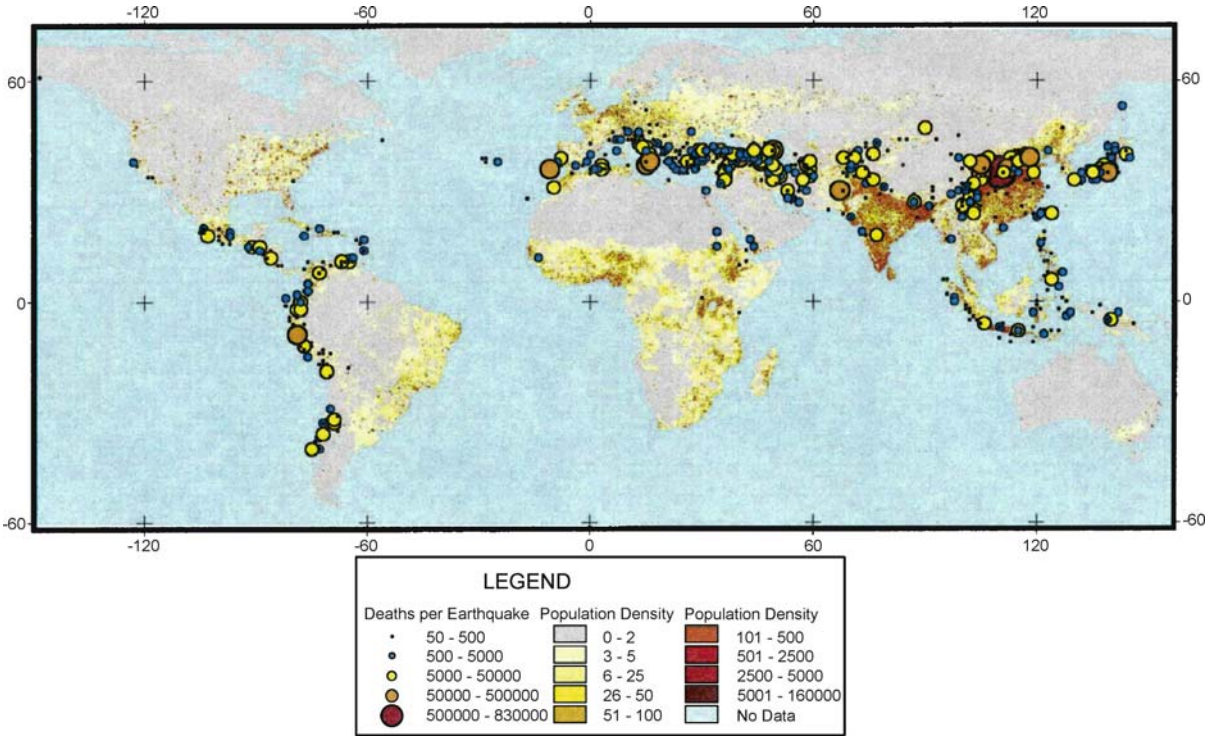

Earthquake Monitoring and Early Warning Systems, Figure 1
Location of deadly earthquakes around the world, 1500–2000. Population density is shown by the *background colors*. See [115] for details

Earthquake Monitoring and Early Warning Systems, Table 1
Deadly Earthquakes/Tsunamis from 1896–2005 ([115] and recent sources)

Origin Time Year MM/DD Hr:Min (UTC, except L=local)	Hypocenter			Magnitude	Location	Deaths (Approximate)
	Lat. (deg)	Lon. (deg)	Depth (km)			
2005 10/08 3:50	34.432	73.573	10	7.6	Pakistan, Kashmir	80,361+
2004 12/26 0:58	3.298	95.778	7	9.2	Indonesia, Sumatra	283,106+
2003 12/26 1:56	29.004	58.337	15	6.6	Iran, Bam	26,000
2001 01/26 3:16	23.420	70.230	16	7.7	India, Gujarat, Bhuj	20,000+
1990 06/20 21:00	37.008	49.213	18	7.4	Iran, western	∼40,000
1988 12/07 7:41	40.919	44.119	7	6.8	Armenia, Spitak	∼40,000
1976 07/27 19:42	39.605	117.889	17	7.6	China, Tangshan	∼242,000
1976 02/04 9:01	15.298	−89.145	13	7.5	Guatemala	23,000
1970 05/31 20:23	−9.248	−78.842	73	7.5	Peru	67,000
1948 10/05 20:12	37.500	58.000	0	7.2	USSR, Ashgabat	∼65,000
1939 12/26 23:57	39.770	39.533	35	7.7	Turkey, Erzincan	33,000
1939 01/25 3:32	−36.200	−72.200	0	7.7	Chile, Chillian	28,000
1935 05/30 21:32	28.894	66.176	35	8.1	Pakistan, Quetta	60,000
1932 12/25 2:04	39.771	96.690	25	7.6	China, Gansu	∼70,000
1927 05/22 22:32	37.386	102.311	25	7.7	China, Tsinghai	∼100,000
1923 09/01 2:58	35.405	139.084	35	7.9	Japan, Kanto	143,000
1920 12/16 12:05	36.601	105.317	25	8.6	China, Gansu	∼240,000
1915 01/13 6:52	42.000	13.500	0	6.9	Italy, Avezzano	33,000
1908 12/28 4:20	38.000	15.500	0	7.0	Italy, Messina	∼ 82,000
1906 08/17 0:40	−33.000	−72.000	0	8.2	Chile, Valparaiso	20,000
1905 04/04 0:50	33.000	76.000	0	8.1	India, Kangra	20,000
1896 06/15 19:32L	39.500	144.000	0	8.2	Japan, Sanriku-oki	22,000

"∼" denotes large uncertainties because a range of deaths had been reported.
"+" denotes a minimum value.

the Middle East, Iran, Pakistan and India, to China and Japan.

Table 1 lists the most deadly earthquakes (death toll $> 20,000$) of the past 110 years based on official estimates (often under-estimated for political reasons, or lack of accurate census data in many areas of the world). In the first 5 years of the 21st century, four disastrous earthquakes occurred in India, Indonesia, Iran, and Pakistan. In the 20th century, the average death toll caused by earthquakes (and tsunamis they triggered) was about 16,000 per year. For the past seven years the yearly death toll was about 60,000 – four times higher than the average in the previous century. In Fig. 2 we extracted a portion of Fig. 1 to illustrate the relationship between past earthquakes and population in India, Pakistan, northern Indonesia, and adjoining regions. We numbered the four most recent disastrous earthquakes in Fig. 2. It is obvious that the large populations in India, Indonesia, Iran, Pakistan, and their adjoining regions (over 1.5 billion people) has been and will continue to be adversely affected by earthquakes. Fatalities depend largely on resistance of building construction to shaking, in addition to population density and earthquake occurrence.

In recent decades, population increases, accelerated urbanization, and population concentration along coastal areas prone to earthquakes suggest that many more earthquake-related fatalities will occur unless effective steps are taken to minimize earthquake and tsunami hazards.

Earthquake Monitoring: Instrumentation

Besides geodetic data [28], the primary instrumental data for the quantitative study of earthquakes are *seismograms*, records of ground motion caused by the passage of seismic waves. Seismograms are written by *seismographs*, instruments which detect and record ground motion along with timing information. A seismograph consists of three basic components: (1) a seismometer, which responds to ground motion and produces a signal proportional to acceleration, velocity, or displacement over a range of amplitudes and frequencies; (2) a timing device; (3) either a local recording unit which writes seismograms on paper, film, or elec-

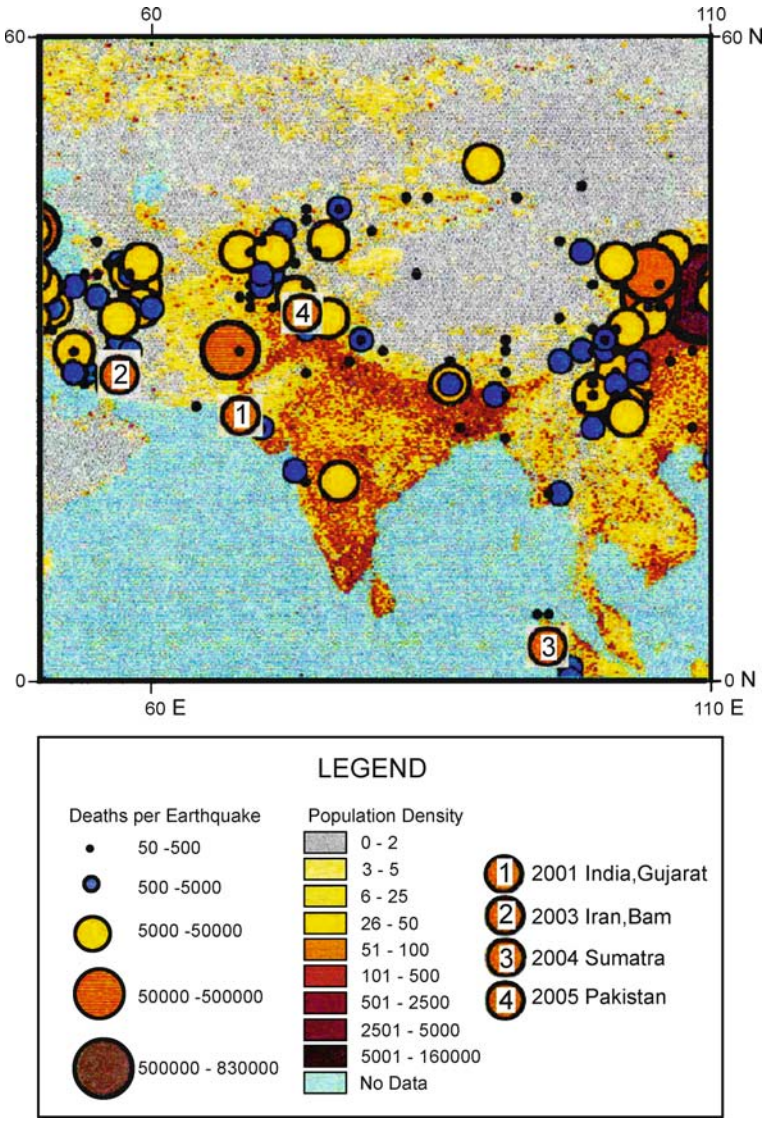

Earthquake Monitoring and Early Warning Systems, Figure 2
Location of the 4 most deadly earthquakes of the 21st century (up to the end of 2007) on a map showing the location of the deadly earthquakes from 16th to 20th centuries (after [115] and Table 1)

tronic storage media, or more recently, a telemetry system for delivering the seismograms to a central laboratory for recording. Technical discussions of seismometry may be found, for example, in Wielandt [122], and of seismic instruments in Havskov and Alguacil [48]. An overview of challenges in observational earthquake seismology is given by Lee [71], and a useful manual of seismological observatory practice is provided by Bormann [12].

An *accelerograph* is a seismograph designed to record, on scale, the acceleration time history of strong ground motions. Measuring acceleration is important for studying response of buildings to strong ground motions close to earthquakes. Many modern sensitive seismographs are *velocigraphs* recording the time history of ground velocity. They are designed to measure seismic waves of small amplitudes (because seismic waves attenuate quickly from their sources) either from small earthquakes nearby, or from large earthquakes that are far away.

A seismic network (or an "array") is a group of seismographs "linked" to a central headquarters. Nowadays the

link is by various methods of telemetry, but in early days the links were by mail or telegrams, or simply by manual collection of the records. When we speak of a seismic *station*, we may mean an observatory with multiple instruments in special vaults or a small instrument package at a remote site.

Seismographs were first developed in the late 19th century, and individual seismographic observatories (often a part of astronomical or meteorological observatories) began earthquake monitoring by issuing earthquake information in their station bulletins and other publications. However, in order to accurately locate an earthquake, data from several seismographic stations are necessary. It was then natural for many governments to assume responsibility for monitoring earthquakes within their territories. However, because seismic waves from earthquakes do not

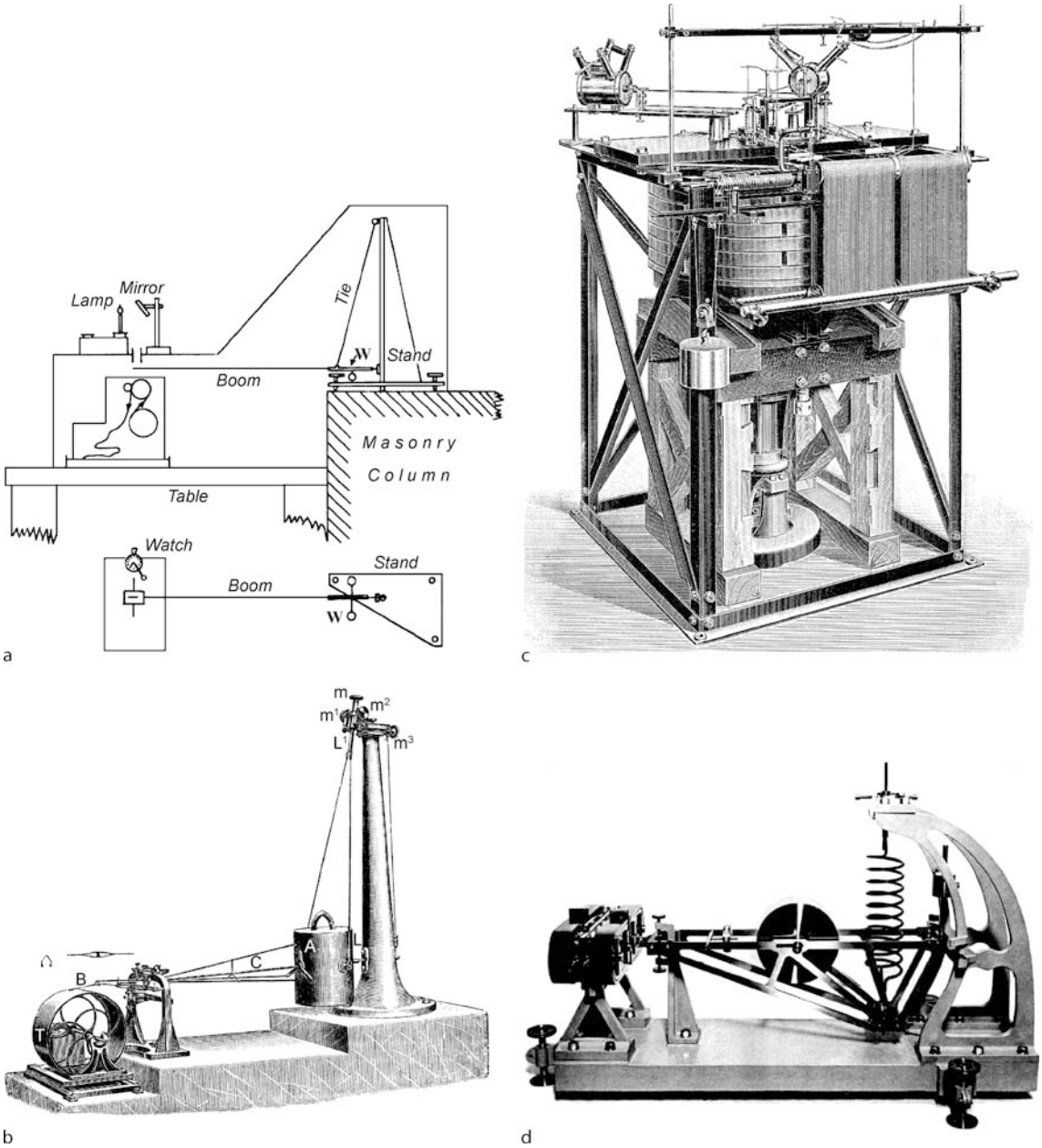

Earthquake Monitoring and Early Warning Systems, Figure 3
Some classical seismographs: a Milne, b Bosch–Omori, c Wiechert, and d Galitzin (after [101])

recognize national boundaries, the need for international cooperation became clear. In the following subsections, we present an overview of the history and results of earthquake monitoring.

Historical Developments

In 1897, John Milne designed the first inexpensive seismograph, which was capable of recording very large earthquakes anywhere in the world. With a small grant from the British Association for the Advancement of Science (BAAS), a few other donations, and his own money, Milne managed to deploy about 30 of his instruments around the world, forming the first worldwide seismographic network. At the same time, seismogram readings were reported voluntarily to Milne's observatory at Shide on the Isle of Wight, England. A global earthquake summary with these seismogram readings was issued by Milne beginning in 1899. These summaries are now known as the "Shide Circulars". Milne also published progress and results in the "Reports of the BAAS Seismological Committee" from 1895 to 1913. A review of Milne's work and a reproduction of his publications as computer readable files were given by Schweitzer and Lee [101] and its attached CD-ROM. After Milne's death in 1913, Herbert H. Turner continued Milne's efforts, and in 1918 established publication of the International Seismological Summary (ISS).

The shortcomings of the Milne seismograph (low magnification, no damping, and poor time resolution) were soon recognized. Several improved seismographs (notably the Omori, Bosch–Omori, Wiechert, Galitzin, and Milne–Shaw) were developed and deployed in the first three decades of the 20th century. Figure 3 shows several of these classical seismographs (see Schweitzer and Lee [101] for further explanation). Although the ISS provided an authoritative compilation arrival-time data of seismic waves and determinations of earthquake hypocenters beginning in 1918, its shortcomings were also evident. These include difficulties in collecting the available arrival-time data around the world (which were submitted on a voluntary basis), and in the processing and analysis of data from many different types of seismographs. Revolutions and wars during the first half of the 20th century frequently disrupted progress, particularly impacting collection and distribution earthquake information.

In the late 1950s, attempts to negotiate a comprehensive nuclear test ban treaty failed, in part because of perceptions that seismic methods were inadequate for monitoring underground nuclear tests [95]. The influential Berkner report of 1959 therefore advocated major support for seismology [66]. As a result, the Worldwide Standardized Seismograph Network (WWSSN) was created in the early 1960s with about 120 continuously recording stations located across much of the world, except China and the USSR [91]. Each WWSSN station was equipped with *identical* sets of short-period and long-period three-component seismographs and accurate chronometers. Figure 4 shows some of the equipment at a WWSSN station, including three-components of long-period seismometers, long-period recording and test instruments, and the time and power console. A similar set of three-component short-period seismometers and recording and test instruments, nearly identical in appearance, was also deployed at each station. Seismograms from the WWSSN were sent to the United States to be photographed on 70 mm film chips for distribution (about US$ 1 per chip as then sold to any interested person).

The WWSSN network is credited with making possible rapid progress in global seismology, and with helping to spark the plate tectonics revolution of the late 1960s [117]. At about the same time, the Unified System of Seismic Observations (ESSN) of the former USSR and its allied countries was established, consisting of almost 100 stations equipped with Kirnos short-period, 1–20 s displacement sensors, and long-period seismographs.

Samples of seismograms recorded on smoked paper and photographic paper or film by analog seismographs are shown in Figs. 5 and 6. Two efforts to preserve and make such records available online are now underway: the *SeismoArchives* (www.iris.edu/seismo/ [72]), and *Sismos* (sismos.rm.ingv.it [82]).

With the establishment of the WWSSN, the United States also assumed the task of monitoring earthquakes on a global scale beginning in the early 1960s. The mission of the US National Earthquake Information Center (NEIC, now part of the US, Geological Survey) is "to determine rapidly the location and size of all destructive earthquakes worldwide and to immediately disseminate this information to concerned national and international agencies, scientists, and the general public" (http://earthquake.usgs.gov/regional/neic/).

In 1964, the ISS was reorganized as the International Seismological Centre (ISC). Since then, the ISC (http://www.isc.ac.uk/) has issued annual global earthquake catalogs with a time lag of about two years [123].

Technical Considerations

To record seismic waves, we must consider both the available technology for designing seismographs, and the nature of the Earth's background noise [121]. The Earth is constantly in motion. This "background" noise is usu-

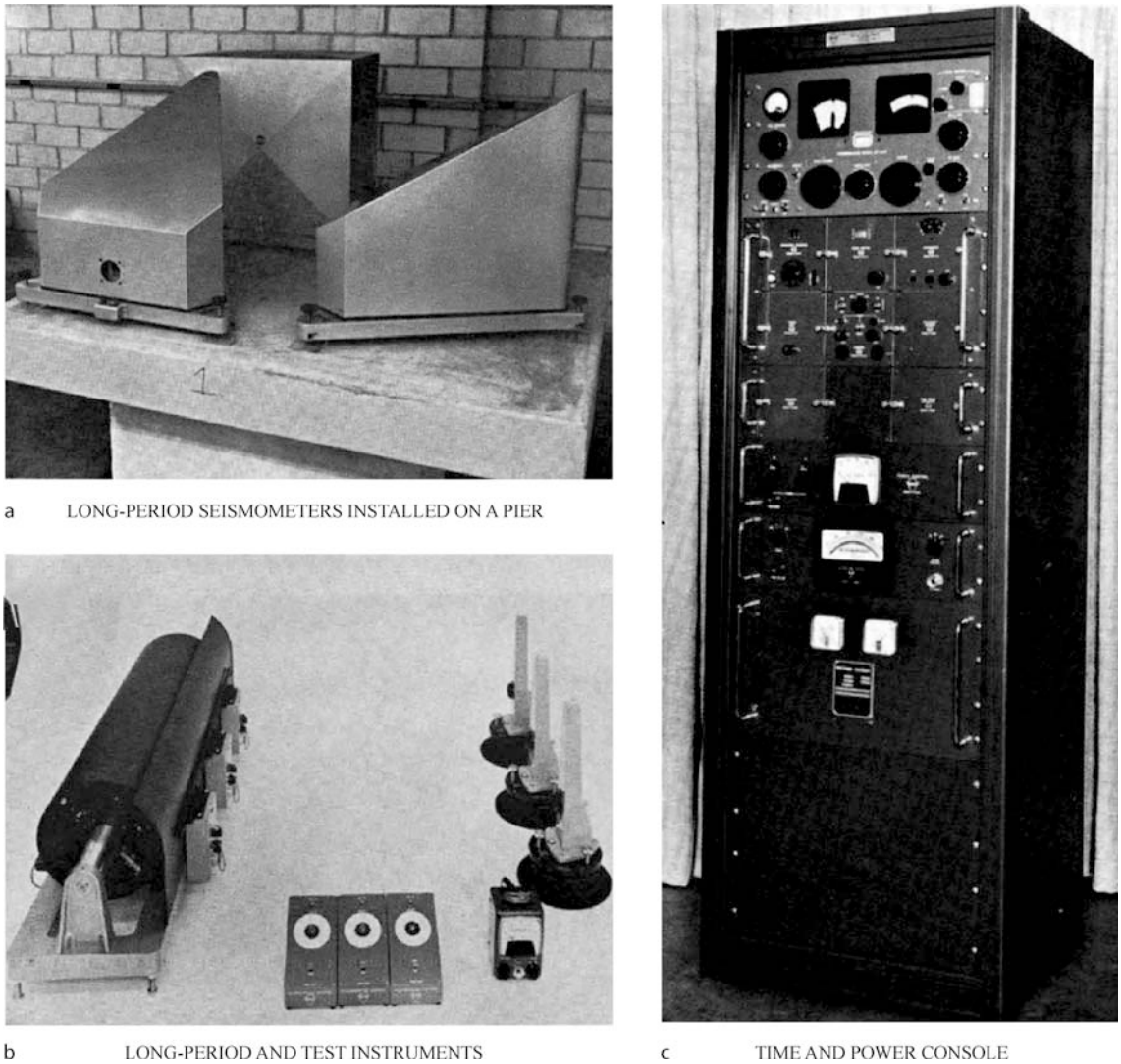

Earthquake Monitoring and Early Warning Systems, Figure 4
Some WWSSN station equipment: a Three-component, long-period seismometers installed on a seismic pier, b Long-period recording and test instruments, and c Time and power console. A similar set of three-component, short-period seismometers and recording/test instruments is not shown

ally classified as either (1) *microseisms*, which typically have frequencies below about 1 Hz, are often the largest background signals, and are usually caused by natural disturbances (largely caused by ocean waves near shorelines); or (2) *microtremors*, which have frequencies higher than about 1 Hz, and are due to human activities (such as traffic and machinery) and local natural sources (such as wind and moving vegetation). Ground motions from earthquakes vary more than ten orders of magnitude in amplitude and six orders of magnitude in frequency, depending on the size of the earthquake and the distance at which it is recorded. Figure 7 illustrates the relative dynamic range of some common seismometers for global earthquake monitoring. A "low Earth noise" model [10,92] is the lower limit of Earth's natural noise in its quietest locations – it is desirable to have instruments that are sensitive enough to detect this minimal background Earth signal. In the analog instrument era (i. e., prior to about 1980), short-period and long-period seismometers were designed separately to avoid microseisms, which have predominant periods of about 6 s. Short-period seismometers were designed to detect tiny ground motions from

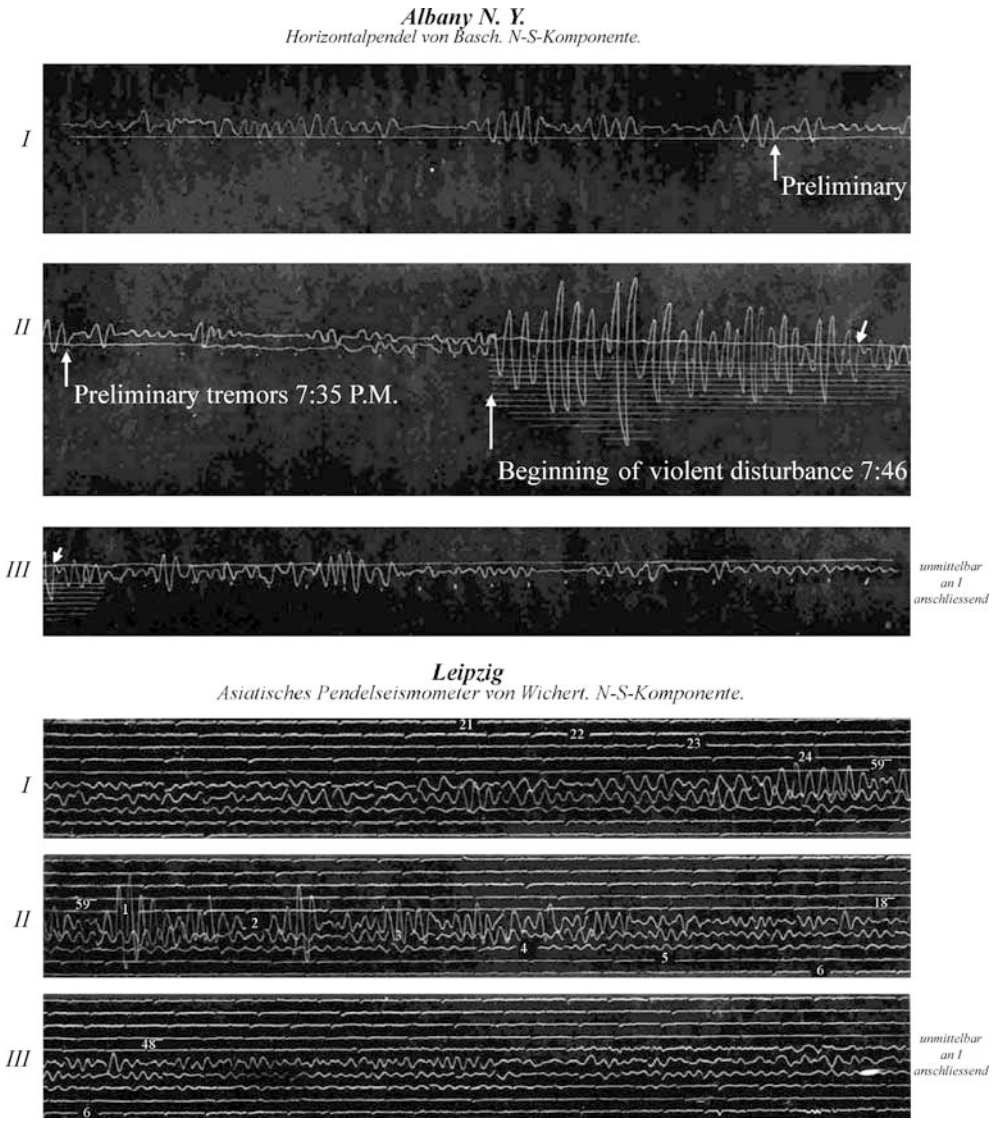

Earthquake Monitoring and Early Warning Systems, Figure 5
Some sample analog seismograms recorded on smoked paper

smaller, nearby earthquakes, while long-period instruments were designed to recover the motions of distant, larger earthquakes ("teleseisms"). Additionally, strong-motion accelerometers, generally recording directly onto 70 mm-wide film strips, were used to measure large motions from nearby earthquakes. In today's much more capable digital instrumentation, two major types of instruments are deployed: (1) "broadband" seismometers, which replace and improve upon both short-period and long-period seismometers, and (2) strong-motion accelerometers for high-amplitude, high-frequency, seismic waves from local earthquakes, which often drive broadband seismometers off scale. While rare examples of the old analog instruments are still in use, the vast majority of instruments presently operating are digital.

In addition to having large variations in amplitudes and frequencies, seismic waves from earthquakes also attenuate rapidly with distance, that is, they lose energy as they travel, particularly at higher frequencies. We must consider these effects in order to monitor seismic waves effectively.

In 1935, C.F. Richter introduced the concept of *magnitude* to classify local earthquakes by their "size", effectively the amount of energy radiated at the actual rup-

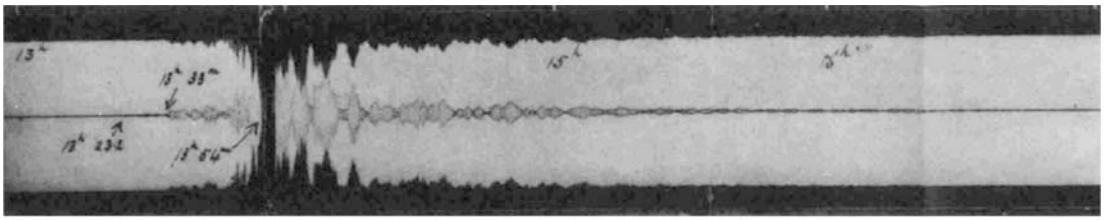

PAISLEY, SCOTLAND. Milne Seismograph. (*From photographic copy.*)

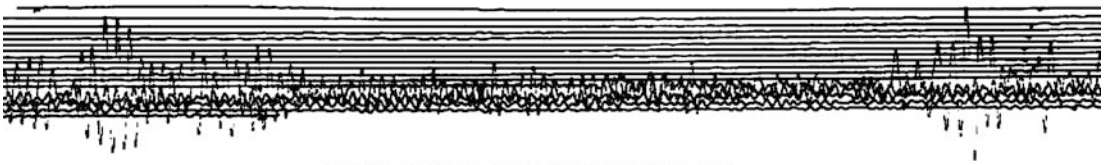

De Bilt, the Netherlands. Galitzin Seismograph.

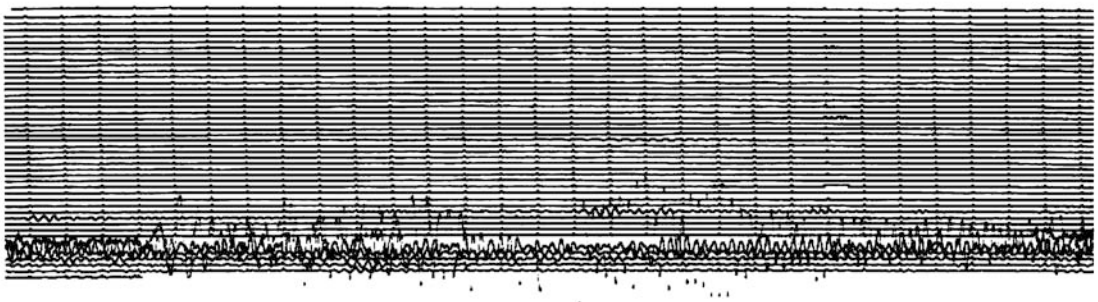

Weston Observatory, USA. Benioff Seismograph

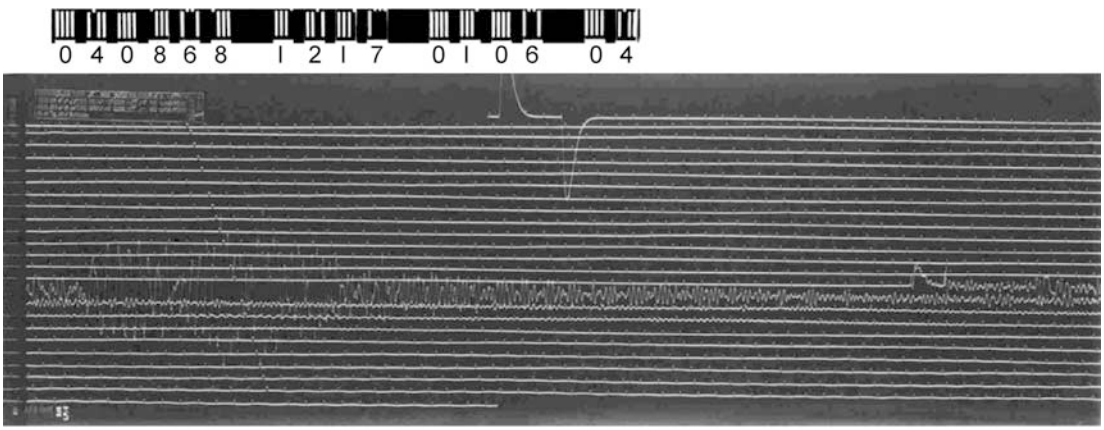

San Juan, Puerto Rico. WWSSN Long-Period Seismograph.

Earthquake Monitoring and Early Warning Systems, Figure 6
Some sample analog seismograms recorded on photographic paper or film

ture surface within the Earth. See the entry by Bormann and Saul ► Earthquake Magnitude for a discussion of the various magnitude scales in use. While every effort is made to make these different scales overlap cleanly, each has strengths and weaknesses that make one or another preferable in a given situation. Probably the most general and robust of these methods is called a "moment magnitude", symbolized as M_W. Existing instruments and environments are such that the smallest natural earthquakes we routinely observe close by are about magnitude = 1. The largest earthquake so far recorded by instrumentals is the $M_W = 9.5$ Chilean earthquake in 1960. In 1941, B.

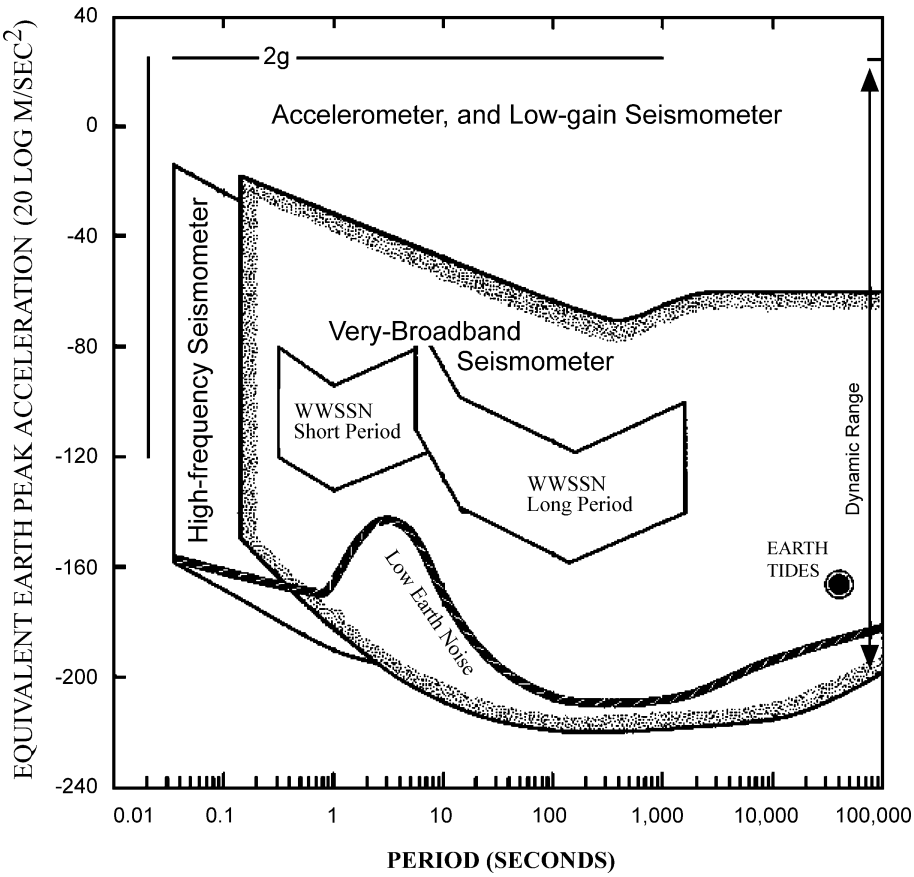

Earthquake Monitoring and Early Warning Systems, Figure 7
Relative dynamic range of some common seismometers for global earthquake monitoring (modified from Fig. 1 in [54]). The *Y*-axis is marked in decibel (dB) where dB = 20 log(A/A_0); *A* is the signal amplitude, and A_0 is the reference signal amplitude

Gutenberg and C.F. Richter discovered that over large geographic regions the rate of earthquake occurrence is empirically related to their magnitudes by:

$$\log N = a - bM \tag{1}$$

where N is the number of earthquakes of magnitude M or greater, and a and b are numerical constants. It turns out that b is usually about 1, which implies that $M = 6$ earthquakes are about ten times more frequent than $M = 7$ earthquakes. Engdahl and Villasenor [24] show that there has been an *average* of about 15 major ($M \geqslant 7$) earthquakes per year over the past 100 years, and about 150 large ($M \geq 6$) earthquakes per year during this same time interval. Strong ground motions (above 0.1 g in acceleration) over sizeable areas are generated by $M \geq 6$ earthquakes; these are potentially damaging levels of ground shaking.

Earthquakes are classified by magnitude (M) as *major* if $M \geq 7$, as *moderate* to *large* if M ranges from 5 to 7, as *small* if M ranges from 3 to 5, as *micro* if $M < 3$, and as *nano* if $M < 0$. An earthquake with $M \geq 7\,3/4$ is often called *great*, and if $M \geq 9$, *mega*.

Earthquake Monitoring in the Digital Era

Figure 8 shows the expected amplitudes of seismic waves by earthquake magnitude. The top frame is a plot of the equivalent peak ground acceleration versus frequency. The two heavy curves denote the "minimum Earth noise", and the "maximum Earth noise" (i. e., for seismographic station located in the continental interior versus near the coast).

The two domains of the WWSSN equipment, short-period long-period seismometers are shown as gray shading. The domains for two other instruments, SRO (Seismic Research Observatories Seismograph) and IDA (International Deployment of Accelerometers), are also shown; these were the early models of the current instruments

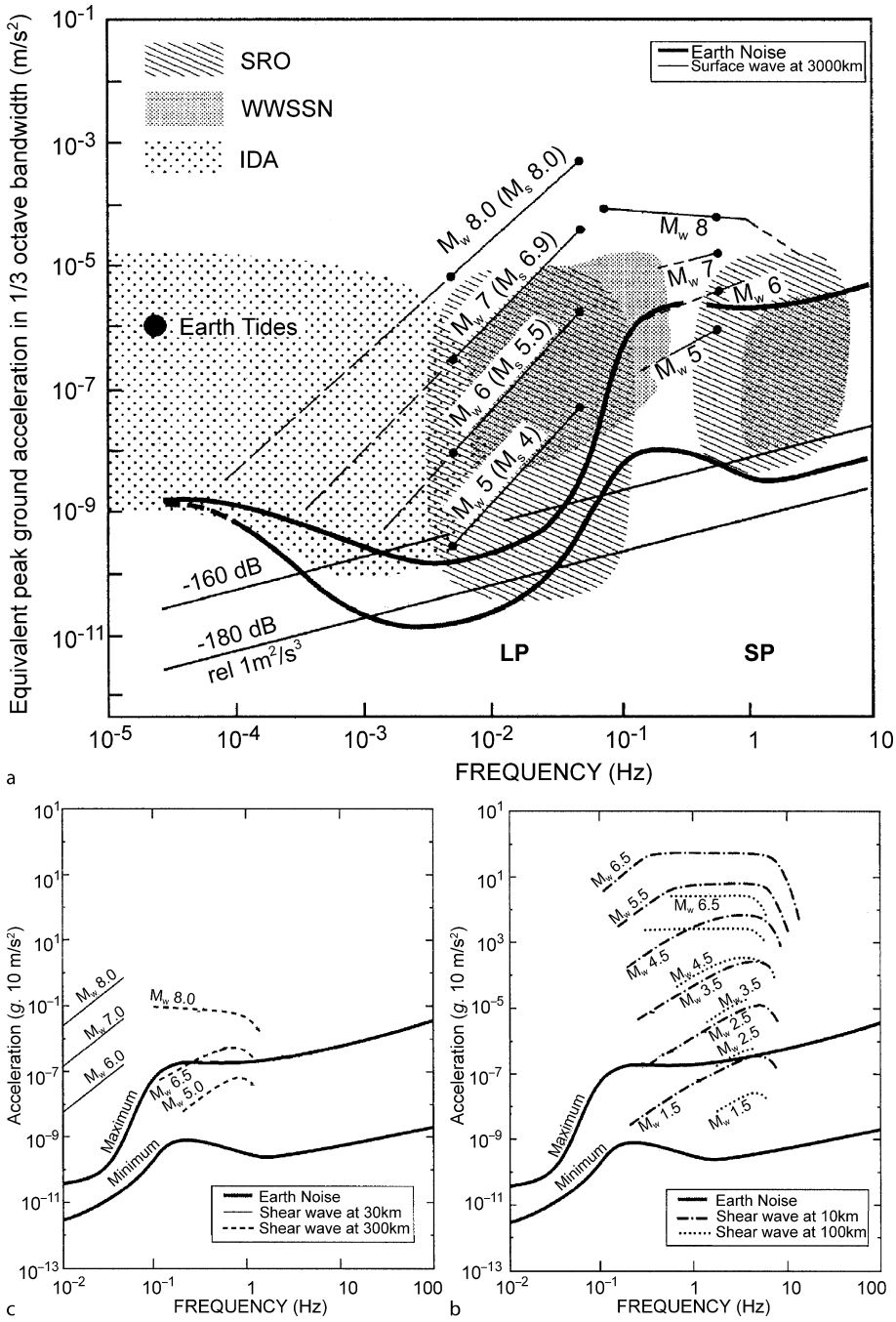

Earthquake Monitoring and Early Warning Systems, Figure 8
Expected amplitudes of seismic waves by earthquake magnitude. See text for explanations

now in operation in the *Global Seismographic Network* (GSN). The bottom two frames indicate expected amplitudes of seismic waves from earthquakes of a range of magnitudes (we use the moment magnitude, M_W). For simplicity, we consider two cases: (bottom left) global earthquakes recorded at a large distance with a seismographic network spaced at intervals of about 1000 km, and (bottom right) local earthquakes recorded at short dis-

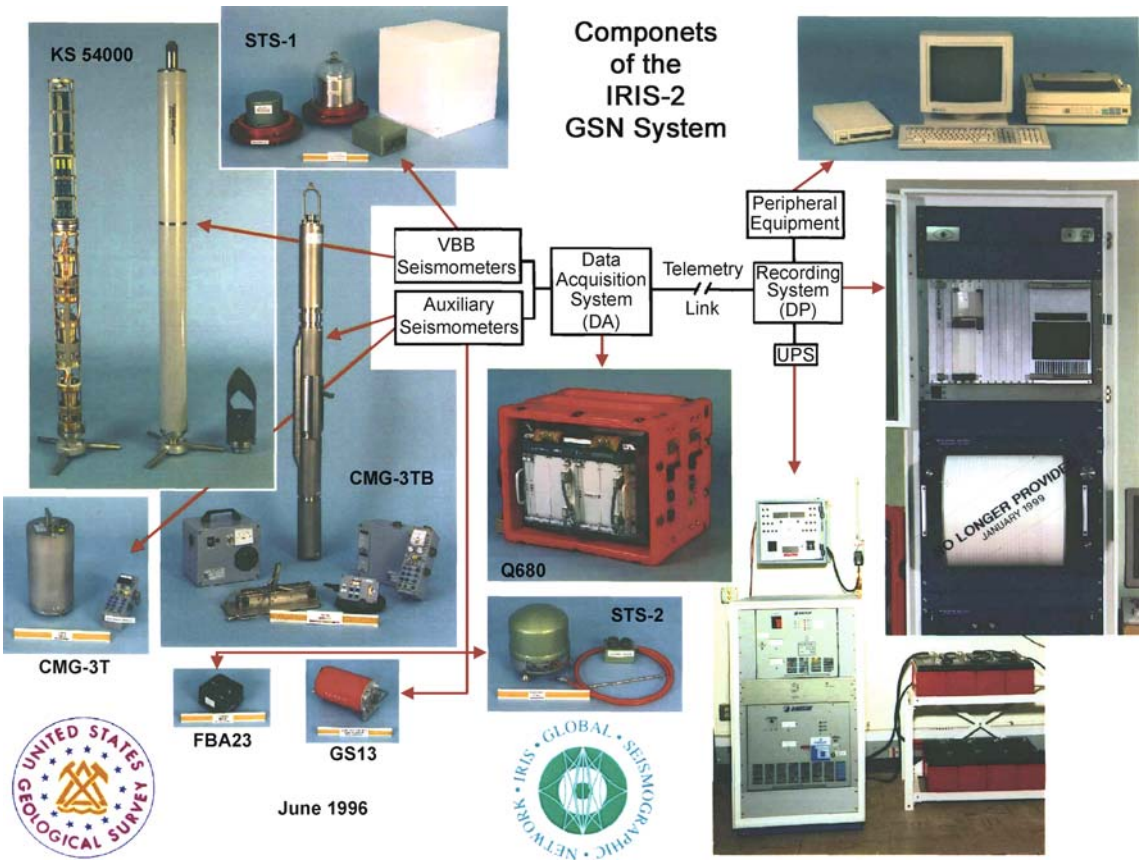

Earthquake Monitoring and Early Warning Systems, Figure 9
Components of the IRIS-2 GSN System: broadband seismometers, accelerometers and recording equipment

tances with a seismic array spaced at intervals of about 50 km. In the bottom left plot, the global-scale network, the expected amplitudes of *P*-wave and surface wave at 3000 km from the earthquake source are shown; for the bottom right plot, a local seismic array, the expected amplitudes of *S*-wave at 10 km and 100 km from the earthquake source are shown. Seismologists use this and similar figures in planning seismographic networks. Local noise surveys are usually conducted as well when designing specific seismographic networks.

With advances in digital technology, earthquake monitoring entered the digital era in the 1980s. Older analog equipment was gradually phased out as modern digital equipment replaced it [54]. The WWSSN was replaced by the *Global Seismographic Network* (GSN), a collaboration of several institutions under the IRIS consortium (http://www.iris.edu/). The goal of the GSN (http://www.iris.edu/about/GSN/index.htm) is "to deploy over 128 permanent seismic recording stations uniformly over the Earth's surface". The GSN project provides funding for two network operators: (1) the IRIS/ASL Network Operations Center, in Albuquerque, New Mexico (operated by the US Geological Survey), and (2) the IRIS/IDA Network Operations Center in La Jolla, California (operated by personnel from the Scripps Institution of Oceanography). Components of a modern IRIS GSN seismograph system, which include broadband seismometers, accelerometers, and recording equipment, are shown in Fig. 9.

Figure 10 shows the station map of the Global Seismographic Network as of 2007. IRIS GSN stations continuously record seismic data from very broad band seismometers at 20 samples per second (sps), and also include high-frequency (40 sps) and strong-motion (1 and 100 sps) sensors where scientifically warranted. It is the goal of the GSN project to provide real-time access to its data via Internet or satellite. Since 1991, the IRIS Data Management Center has been providing easy access to comprehensive seismic data from the GSN and elsewhere [1].

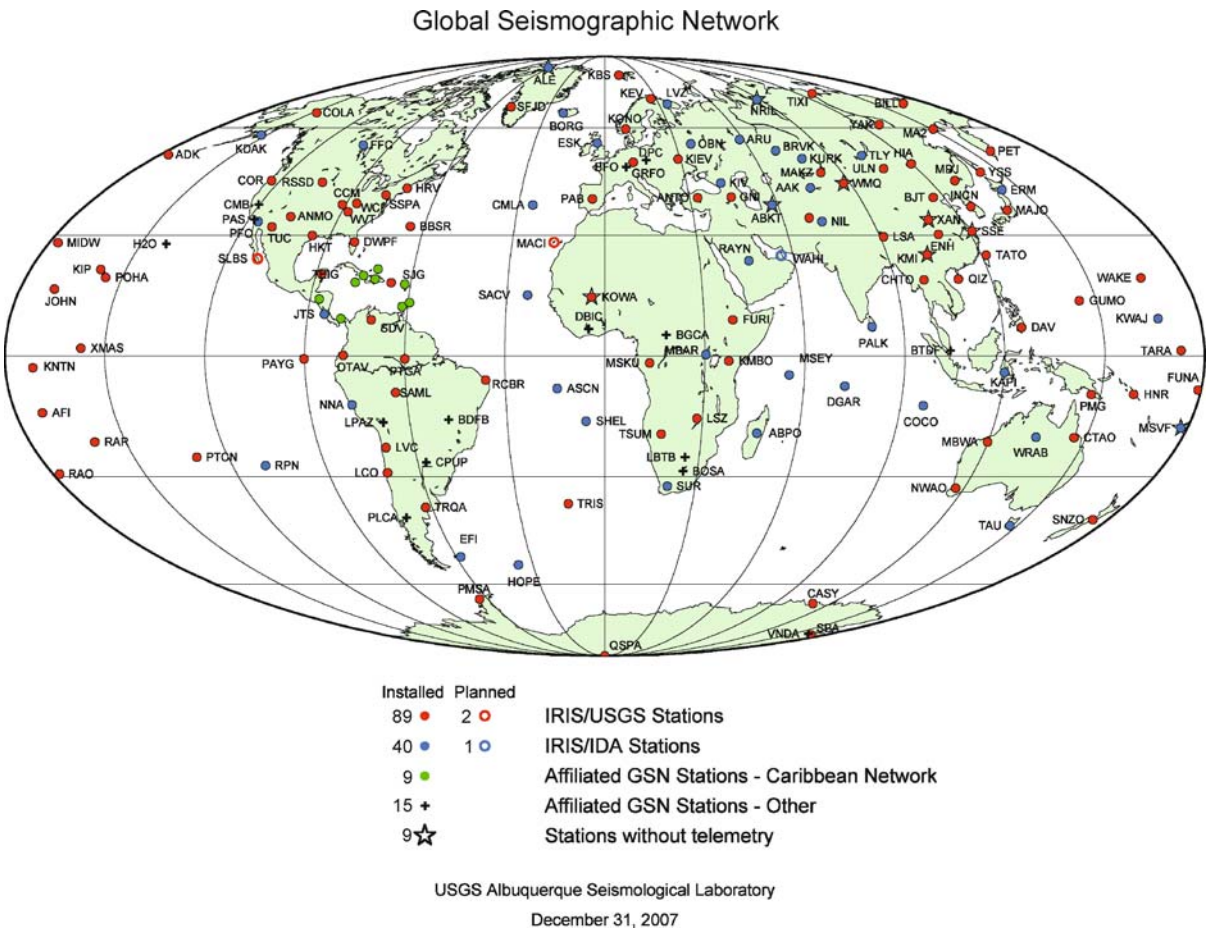

Earthquake Monitoring and Early Warning Systems, Figure 10
Station map of the Global Seismographic Network (GSN) as of 2007

Earthquake Monitoring: Regional and Local Networks

A major development in earthquake monitoring was the establishment of seismographic networks optimized to record the many *frequent* but *smaller* regional and local earthquakes occurring in many locations. To observe as many of these nearby earthquakes as possible, inexpensive seismographs with high magnifications and low dynamic-range telemetry are used to record the smallest earthquakes feasible with current technology and local background noise. As a result, the recorded amplitudes often overdrive the instruments for earthquakes with $M \gtrsim 3$ within about 50 km of such seismographs. This is not a serious defect, since the emphasis for these networks is to obtain as many first arrival times as possible, and to detect and to locate the maximum number of earthquakes. Because seismic waves from small earthquakes are quickly attenuated with increasing distance, it is also necessary to deploy many instruments at small station spacing (generally from a few to a few tens of kilometers), and to cover as large a territory as possible in order to record at least a few earthquakes every week. Since funding often is limited, these local and regional seismic networks are commonly optimized for the largest number of stations rather than for the highest quality data.

A Brief History

In the 1910s, the Carnegie Institution of Washington D.C. (CIW) was spending a great deal of money building the world's then largest telescope (100 inch) at Mount Wilson Observatory, southern California [38]. Since astronomers were concerned about earthquakes that might disturb their

telescopes, Harry O. Wood was able to persuade CIW to support earthquake investigations, and as a result, a regional network of about a dozen Wood–Anderson seismographs was established in southern California in the 1920s. See Goodstein [38] for the early history leading to the establishment of the California Institute of Technology (Caltech) and its Seismological Laboratory. Astronomers played important roles in getting seismic monitoring established in various other regions of the world as well.

Regional networks using different types of seismographs were established in many countries about this time, such as in Japan, New Zealand, and the USSR and its allies. In the 1960s, high-gain, short-period, telemetered networks were developed to study microearthquakes. To support detailed studies of local earthquakes and especially for the purpose of earthquake prediction, over 100 microearthquake networks were established in various parts of the world by the end of the 1970s [74]. These microearthquake networks comprised from tens to hundreds of short-period seismometers, generally with their signals telemetered into central recording sites for processing and analysis. High magnification was achieved through electronic amplification, permitting recording of very small earthquakes (down to $M = 0$), though this came at the expense of saturated records for earthquakes of $M \gtrsim 3$ within about 50 km. Unfortunately, the hope of discovering some sort of earthquake precursor from the data obtained by these microearthquake networks did not work out. For a review of the earthquake prediction efforts, please read Kanamori [60].

Some Recent Advances

Because of recent advances in electronics, communications, and microcomputers, it is now possible to deploy sophisticated digital seismograph stations at global, national, regional, and local scales for *real-time* seismology [64]. Many such networks, including temporary portable networks, have been implemented in many countries. In particular, various real-time and near real-time seismic systems began operation in the 1990s: for example, in Mexico [25], California [32,47], and Taiwan [110]. The Real-Time Data (RTD) system operated by the Central Weather Bureau (CWB) of Taiwan is based on a network of telemetered digital accelerographs [102]; since 1995, this system has used pagers, e-mail, and other techniques to automatically and rapidly disseminate information about the hypocenter, magnitude, and shaking amplitude of felt earthquakes ($M \gtrsim 4$) in the Taiwan region. The disastrous Chi-Chi earthquake ($M_W = 7.6$) of 20 September 1999 caused 2,471 deaths and total economic losses of US$ 11.5 billion. For this earthquake sequence, the RTD system delivered accurate information to government officials 102 seconds after the origin time of the main shock (about 50 seconds for most aftershocks), and proved to be useful in the emergency response of the Taiwan government [37,131].

Recording Damaging Ground Shaking

Observing teleseisms on a global scale with station spacing of several hundreds of kilometers does not yield critical information about near-source strong ground shaking required for earthquake structural engineering purposes and seismic hazard reduction. Broadband seismometers, which are optimized to record earthquakes at great distances, do not perform well in the near-field of a major earthquake. For example, during the 1999 Chi–Chi earthquake the nearest broadband station in Taiwan (epicentral distance of about 20 km) was badly overdriven, recorded no useful data beyond the arrival time of the initial *P*-wave, and finally failed about one minute into the shock.

A regional seismic network with station spacing of a few tens of kilometers cannot do the job either: the station spacing is still too large and the records are typically overdriven for earthquakes of $M \gtrsim 3$ (any large earthquake would certainly overdrive these sensitive instruments in the entire network). In his account of early earthquake engineering, Housner [51] credited John R. Freeman, an eminent engineer, with persuading the then US Secretary of Commerce to authorize a strong-motion program, and, in 1930, the design of an accelerograph for engineering purposes. In a letter to R.R. Martel, Housner's professor at Caltech, Freeman wrote:

> *I stated that the data which had been given to structural engineers on acceleration and limits of motion in earthquakes as a basis for their designs were all based on guesswork, that there had never yet been a precise measurement of acceleration made. That of the five seismographs around San Francisco Bay which tried to record the earthquake of 1906 not one was able to tell the truth.*

Strong-motion recordings useful to engineers must be on-scale for damaging earthquakes and, in particular, from instruments located on or near built structures in densely urbanized environments, within about 25 km of the earthquake-rupture zone for sites on rock, or within about 100 km for sites on soft soils. Recordings of motions sufficient to cause damage at sites at greater distances are also of interest for earthquake engineering in areas likely to be affected by major subduction-zone earth-

quakes and in areas with exceptionally low attenuation rates [11]. In addition, densely-spaced networks of strong-motion recorders are needed to study the large variations in these motions over short distances [26,29].

Although several interesting accelerograms were recorded in southern California in the 1930s and 1940s, most seismologists did not pursue strong-motion monitoring until much later. The 1971 San Fernando earthquake emphatically demonstrated the need for more strong-motion data [9]. Two important programs emerged in the United States – the National Strong-Motion Program (http://nsmp.wr.usgs.gov/), and the California Strong Motion Instrumentation Program (http://docinet3.consrv.ca.gov/csmip/). However, the budgets for these programs were and continue to be small in comparison to other earthquake programs. High levels of funding for strong-motion monitoring, comparable to that of the GSN and the regional seismic networks, became available in Taiwan in the early 1990s, and in Japan in the mid-1990s. The Consortium of Organizations for Strong-Motion Observation Systems (http://www.cosmos-eq.org/) was established recently to promote the acquisition and application of strong-motion data.

Seismograms and Derived Products

Even before instruments were developed to record seismic waves from earthquakes, many scholars compiled catalogs of earthquake events noted in historical and other documents. Robert Mallet in 1852–1854 published the first extensive earthquake catalog of the world (1606 B.C.–A.D. 1842) totaling 6831 events [79]. Based on this compilation, Mallet prepared the first significant seismicity map of the Earth in 1858. Mallet's map is remarkable in that it correctly identifies the major earthquake zones of the Earth excepting for parts of the oceans. Although Mallet's earthquake catalog and similar compilations contain a wealth of information about earthquakes, they were made without the aid of instruments, and thus were subject to the biases of the observers as well as to population distributions. These non-instrumental earthquake catalogs also contain errors because the source materials were commonly incomplete and inconsistent regarding date, time, place names, and reported damage. Ambraseys et al. [8] discusses these difficulties for a regional case and Guidoboni [42] addresses the matter in general.

Today, seismograms are the fundamental data produced by earthquake monitoring. An analyst's first task is to find out when and where the earthquakes occurred, its size, and other characteristics. The accuracy of determining earthquake parameters, as well as the number of parameters used to characterize and earthquake, has progressed along with the availability of seismograms and computers, as well as advances in seismology. In the analog era, earthquake parameters were primarily the origin time, geographical location (epicenter), focal depth, and magnitude. A list of these parameters for earthquakes occurring over some time interval is called an *earthquake catalog*. A useful and common illustration of such results is a map showing the locations of earthquakes by magnitude (a seismicity map). Figure 11 is such a seismicity map for 1900–1999 as prepared by Engdahl and Villasensor [24]. The map shows that moderate and large earthquakes are concentrated in tectonic active areas while most areas of the Earth are aseismic.

Earthquake Location

Several methods have been developed to locate earthquakes (i. e., determine origin time, latitude and longitude of the epicenter, and focal depth). Common to most of these methods is the use of arrivals times of initial *P*- and *S*-waves. In particular, Geiger [33] applied the Gauss–Newton method to solve for earthquake location, which is a nonlinear problem, by formulating it as an inverse problem. However, since Geiger's method is computational intensive, it was not practical to apply it for the routine determinations of earthquake hypocenters until the advance of modern computers in the early 1960s.

Before computers became widely available starting in the 1960s, earthquakes were usually located by a manual, graphical method. In any location method, we assume that an earthquake is a point source and its sole parameters are origin time (time of occurrence, t_o) and hypocenter position (x_o, y_o, z_o). If both *P*- and *S*-arrival times are available, one may use the time intervals between *P*- and *S*-waves at each station (*S*-*P* times) and estimates of seismic wave velocities in the Earth to obtain a rough estimate of the epicentral distance, *D*, from that station:

$$D = [V_P V_S/(V_P - V_S)](T_S - T_P) \quad (2)$$

where V_P is the *P*-wave velocity, V_S the *S*-wave velocity, T_S the *S*-wave arrival time, and T_P the *P*-wave arrival time. For a typical crustal *P*-wave velocity of 6 km/s, and $V_P/V_S \approx 1.8$, the distance *D* in kilometers is about 7.5 times the *S*-*P* interval measured in seconds. If three or more epicentral distances are available, the epicenter may be placed at the intersection of circles with the stations as centers and the appropriate *D* as radii. The intersection

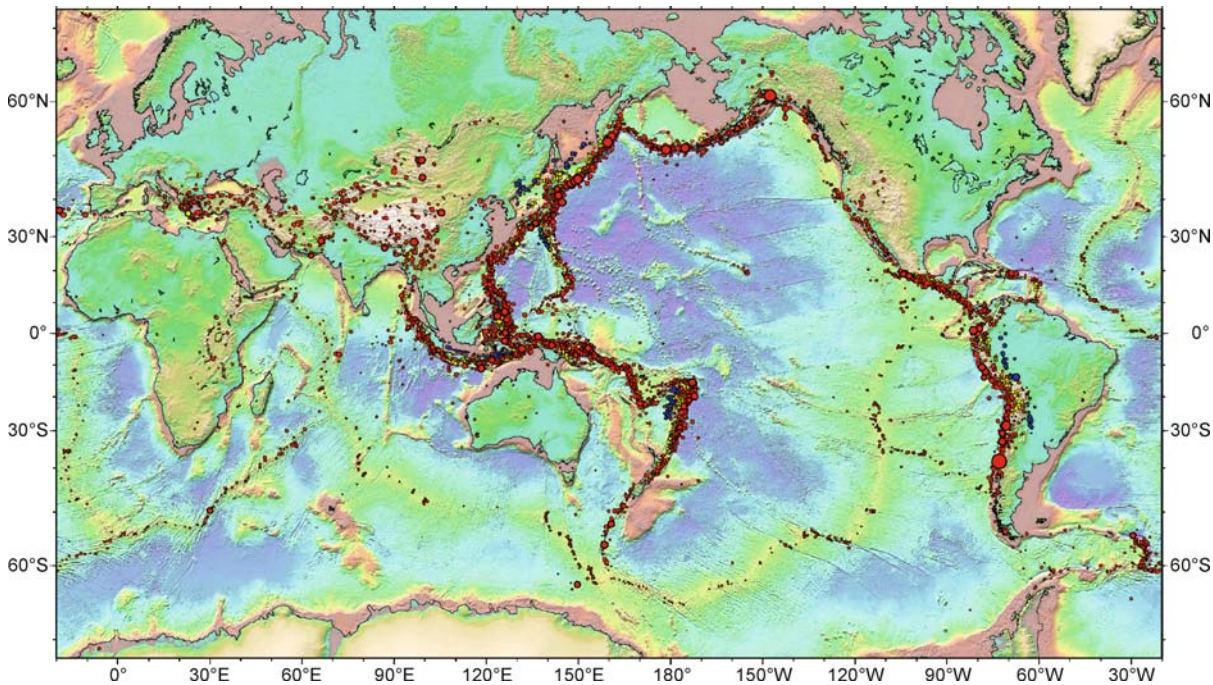

Earthquake Monitoring and Early Warning Systems, Figure 11
Seismicity of the Earth: 1900–1999 (see [24] for details)

will seldom be a point, and its areal extent gives a rough estimate of the uncertainty of the epicenter and hypocentral (focal) depth. In the early days, the focal depth was usually assumed or occasionally determined using a "depth phase" (generally, a ray that travels upward from the hypocenter and reflects back from the Earth's surface, then arcs through the Earth to reach a distant seismograph).

Although Geiger [33] presented a method for determining the origin time and epicenter, the method can be extended easily to include focal depth. To locate an earthquake using a set of arrival times, τ_k, from stations at positions (x_k, y_k, z_k), $k = 1, 2, \ldots, m$, we must assume a model of seismic velocities from which theoretical travel times, T_k for a trial hypocenter at (x^*, y^*, z^*) to the stations can be computed. Let us consider a given trial origin time and hypocenter as the trial vector $\boldsymbol{\chi}^*$ in a four-dimensional Euclidean space:

$$\boldsymbol{\chi}^* = (t^*, x^*, y^*, z^*)^{\mathrm{T}} \tag{3}$$

where the superscript T (^{T}) denotes the vector transpose. Theoretical arrival time, t_k, from $\boldsymbol{\chi}^*$ to the k-th station is the theoretical travel time, T_k, plus the trial origin time, t^*. We now define the arrival time residual at the k-th station, r_k, as the difference between the observed and the theoretical arrival times. We may consider this set of station residuals as a vector in an m-dimensional Euclidean space and write:

$$\mathbf{r} = (r_1(\boldsymbol{\chi}^*), \mathrm{r}_2(\boldsymbol{\chi}^*), \ldots, \mathrm{r}_m(\boldsymbol{\chi}^*))^{\mathrm{T}} . \tag{4}$$

We now apply the least squares method to obtain a set of linear equations solving for an adjustment vector, $\boldsymbol{\delta\chi}$:

$$\mathbf{A}^{\mathrm{T}}\mathbf{A}\boldsymbol{\delta\chi} = -\mathbf{A}^{\mathrm{T}}\mathbf{r} , \tag{5}$$

where A is the Jacobian matrix consisting of partial derivatives of travel time with respect to t, x, y, and z. A detailed derivation of the Geiger method is given by Lee and Stewart (see, pp 132–134 in [74]). There are many practical difficulties in implementing Geiger's method for locating earthquakes, as discussed by Lee and Stewart (see, pp 134–139 in [74]). Although standard errors for these earthquake locations can be computed, they are often not meaningful because errors in the measurement of arrival times usually do not obey a Gaussian probability distribution. In recent years, many authors applied various nonlinear methods to locate earthquakes; a review of these methods is given by Lomax et al. ► Earthquake Location, Direct, Global-Search Methods.

Earthquake Magnitude

After an earthquake is located, the next question is: how big was it? The Richter magnitude scale was originally devised to measure the "size" of an earthquake in southern California. Richter [96] defined the local (earthquake) magnitude, M_L, of an earthquake observed at any particular station to be:

$$M_L = \log A - \log A_0(\Delta) \tag{6}$$

where A is the maximum amplitude in millimeters as recorded by a Wood–Anderson seismograph for an earthquake at epicentral distance of Δ km. The correction factor, $\log A_0(\Delta)$, is the maximum amplitude at Δ km for a "standard" earthquake. Thus, three arbitrary choices enter into the definition of local magnitude: (1) the use of the Wood–Anderson seismographs, (2) the use of the common logarithm (i. e., the logarithm to the base 10), and (3) the selection of the standard earthquake, whose amplitudes as a function of distance Δ are represented by $A_0(\Delta)$.

In the 1940s, B. Gutenberg and C.F. Richter extended the local magnitude scale to include more distant earthquakes. Gutenberg [43] defined the surface-wave magnitude, M_S, as

$$M_S = \log(A/T) - \log A_0(\Delta^\circ) \tag{7}$$

where A is the maximum combined horizontal ground displacement in micrometers (μm) for surface waves with a period of 20 s, and $-\log A_0$ is tabulated as a function of epicentral distance Δ in degrees, in a similar manner to that for the local magnitude's A_0 (Δ). Specifically, surface-wave magnitude is calculated from

$$M_S = \log A + 1.656 \log \Delta + 1.818 \tag{8}$$

using the prominent 20 s period surface waves commonly observed on the two horizontal-component seismograms from earthquakes of shallow focal depth.

Both magnitude scales were derived empirically and have scale-saturation problems, e. g., for very large earthquakes above a certain size the computed magnitudes of a particular type are all about the same After the pioneering work of Charles F. Richter and Beno Gutenberg, numerous authors have developed alternative magnitude scales, as reviewed recently by Utsu [116] and by Bormann and Saul ► Earthquake Magnitude. A current magnitude scale widely accepted as "best" (as having the least saturation problem and being a close match to an earthquake's total release of stress and strain) is the "moment magnitude", M_W, computed from an earthquake's "moment tensor".

Quantification of the Earthquake Source

As pointed out by Kanamori [59], it is not a simple matter to find a single measure of the "size" of an earthquake, simply because earthquakes result from complex physical processes. The elastic rebound theory of Harry F. Reid suggests that earthquakes originate from spontaneous slippage on active faults after a long period of elastic strain accumulation [94]. Faults may be considered the slip surfaces across which discontinuous displacement occurs in the Earth, while the faulting process may be modeled mathematically as a shear dislocation in an elastic medium (see [100], for a review). A shear dislocation (or slip) is equivalent to a double-couple body force [15,81]. The scaling parameter of each component couple of a double-couple body force is its *moment*. Using the equivalence between slip and body forces, Aki [2] introduced the *seismic moment*, M_0, as:

$$M_0 = \mu \int D(A)\mathrm{d}A = \mu s A \tag{9}$$

where μ is the shear modulus of the medium, A is the area of the slipped surface or source area, and s is the slip $D(A)$ averaged over the area A. If an earthquake produces surface faulting, we may estimate its rupture length, L, and its average slip, s, from measurement of that faulting. The area A may be approximated by Lh, where h is the focal depth (it is often, but not always, found that the hypocenter is near the bottom of the rupture surface). A reasonable estimate for μ is 3×10^{11} dynes/cm^2. With these quantities, we can estimate the seismic moment from Eq. (9).

Seismic moment also can be estimated independently from seismograms. From dislocation theory, the seismic moment can be related to the far-field seismic displacement recorded by seismographs. For example, Hanks and Wyss [46] showed that

$$M_0 = (\Omega_0/\psi_{\theta\phi})4\pi\rho R v^3 \tag{10}$$

where Ω_0 is the long-period limit of the displacement spectrum of either P or S waves, $\psi_{\theta\phi}$ is a function accounting for the body-wave radiation pattern, ρ is the density of the medium, R is a function accounting for the geometric spreading of body waves, and v is the body-wave velocity. Similarly, seismic moment can be determined from surface waves or coda waves [2,3].

In 1977, Hiroo Kanamori recognized that a new magnitude scale can be developed using seismic moment (M_0) by comparing the earthquake energy and seismic moment relation

$$E_S = (\Delta\sigma/2\mu)M_0\ , \tag{11}$$

where $\Delta\sigma$ is the stress drop and μ is the shear modulus, with the surface-wave magnitude and energy relation [45]

$$\log E_S = 1.5M_S + 11.8\,, \tag{12}$$

where E_S and M_0 are expressed in ergs and dyne-cm, respectively. The average value of $(\Delta\sigma/2\mu)$ is approximately equal to 1.0×10^{-4}. If we use this value in Eq. (11), we obtain

$$\log M_0 = 1.5M_S + 16.1\,. \tag{13}$$

It is known that M_S values saturate for great earthquakes (M_0 about 10^{29} dyne − cm or more) and, therefore, that Eqs. (12) and (13) do not hold for such great earthquakes. If a new moment-magnitude scale using the notation M_W is defined by

$$\log M_0 = 1.5M_W + 16.1 \tag{14}$$

then M_W is equivalent to M_S below saturation and provides a reasonable estimate for great earthquakes without the saturation problem [58]. The subscript letter W stands for the work at an earthquake fault, but soon M_W became known as the *moment magnitude*. Determining earthquake magnitude using seismic moment is clearly a better approach because it has a physical basis.

The concept of seismic moment led to the development of moment tensor solutions for quantifying the earthquake source, including its focal mechanism [35,36]; the seismic moment is just the scalar value of the moment tensor. Since the 1980s, Centroid-Moment-Tensor (CMT) solutions have been produced routinely for events with moment magnitudes (M_W) greater than about 5.5. The CMT methodology is described by Dziewonski et al. [22] and Dziewonski and Woodhouse [20]; a comprehensive review is given in Dziewonski and Woodhouse [21]. These CMT solutions are published yearly in the journal *Physics of the Earth and Planetary Interiors*, and the entire database is accessible online. This useful service is now performed by the Global CMT Project (http://www.globalcmt.org/), and more than 25,000 moment tensors have been determined for large earthquakes from 1976 to 2007. In the most recent decade, Quick CMT solutions [23] determined in near-real time have been added and are distributed widely via e-mail (http://www.seismology.harvard.edu/projects/CMT/QuickCMTs/).

Limitations of Earthquake Catalogs

In addition to international efforts to catalog earthquakes on a global scale, observatories and government agencies issue more-detailed earthquake catalogs at local, regional, and national scales. However, earthquake catalogs from local to global scales vary greatly in spatial and temporal coverage and in quality, with respect to completeness and accuracy, because of the ongoing evolution of instrumentation, data processing procedures, and agency staff. An earthquake catalog, to be used for research, should have at least the following source parameters: origin time, epicenter (latitude and longitude), focal depth, and magnitude.

The International Seismological Summary and its predecessors provided compilations of arrival times and locations of earthquakes determined manually from about 1900 to 1963. Despite their limitations (notably the lack of magnitude estimates), these materials remain valuable. The first global earthquake catalog that contains both locations and magnitudes was published by Gutenberg and Richter in 1949, and was followed by a second edition in 1954 [44]. This catalog contains over 4,000 earthquakes from 1904 to 1951. Unfortunately, its temporal and spatial coverage is uneven as a result of rapid changes in seismic instrumentation, and of the interference of both World Wars. Nevertheless, the procedures used for earthquake location and magnitude estimation were the same throughout, using the arrival-time and amplitude data available to Gutenberg and Richter during the 1940s and early 1950s.

Since 1964, the International Seismological Centre has performed systematic cataloging of earthquakes worldwide by using computers and more modern seismograph networks. The spatial coverage of this catalog is not complete for some areas of the Earth (especially the oceans) because of the paucity of seismographic stations in such areas. By plotting the cumulative numbers of earthquakes above a certain magnitude versus magnitude, and using Eq. (1), the lower limit of completeness of an earthquake catalog may be estimated – it is the magnitude below which the data deviate below a linear fit to Eq. (1).

A *Centennial Earthquake Catalog* covering ISS- and ISC-reported global earthquakes from 1900–1999 was generated using an improved Earth model that takes into account regional variations in seismic wave velocities in the Earth's crust and upper mantle [24,118]. Engdahl and Villasenor [24] also compiled existing magnitude data from various authors and suggested preferred values. However, these "preferred magnitudes" were not determined by the same procedures. At present, the Global CMT Project (http://www.globalcmt.org/) provides the most complete online source parameters for global earthquakes (with $M_W > 5.5$), including Centroid-Moment-Tensor solutions. Although the CMT catalog starts in 1976, the improved global coverage of modern broadband digital seismographs began only in about 1990.

In summary, earthquake catalogs have been used extensively for earthquake prediction research and seismic hazard assessment since the first such catalog was produced. Reservations have been expressed about the reliability of the results and interpretations from these studies because the catalogs cover too little time and have limitations in completeness and accuracy (both random and systematic). Nevertheless, advances have been made in using earthquake catalogs to (1) study the nature of seismicity (e. g., ► Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space), (2) investigate earthquake statistics (e. g., ► Earthquake Occurrence and Mechanisms, Stochastic Models for), (3) forecast earthquakes (e. g., ► Earthquake Forecasting and Verification), (4) predict earthquakes (e. g., ► Geo-complexity and Earthquake Prediction), (5) assess seismic hazards and risk, and so forth.

Earthquake Early Warning (EEW) Systems

With increasing urbanization worldwide, earthquake hazards pose ever greater threats to lives, property, and livelihoods in populated areas near major active faults on land or near offshore subduction zones. Earthquake early-warning systems can be useful tools for reducing the impact of earthquakes, provided that cities are favorably located with respect to earthquake sources and their citizens are properly trained to respond to the warning messages. Recent reviews of earthquake early warning systems may be found in Lee and Espinosa-Aranda [73], Kanamori [61], and Allen [6], as well as a monograph on the subject by Gasparini et al. [31].

Under favorable conditions, an EEW system can forewarn an urban area of impending strong shaking with lead times that range from a few seconds to a few tens of seconds. A lead time is the time interval between issuing a warning and the arrival of the *S*-waves, which are the most destructive seismic waves. Even a few seconds of advanced warning is useful for pre-programmed emergency measures at various critical facilities, such as the deceleration of rapid-transit vehicles and high-speed trains, the orderly shutoff of gas pipelines, the controlled shutdown of some high-technological manufacturing operations, the safe-guarding of computer facilities (e. g., disk-head parking), and bringing elevators to a stop at the nearest floor.

Physical Basis and Limitations of EEW Systems

The physical basis for earthquake early warning is simple: damaging strong ground shaking is caused primarily by shear (*S*) and subsequent surface waves, both of which travel more slowly that the primary (*P*) waves, and seismic waves travel much more slowly than electromagnetic signals transmitted by telephone or radio. However, certain physical limitations must be considered, as shown by Fig. 12.

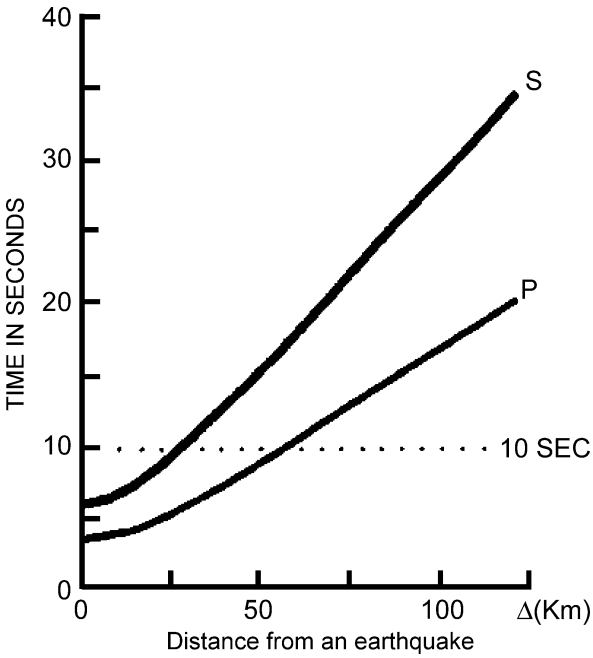

Earthquake Monitoring and Early Warning Systems, Figure 12
Travel time of *P*-waves and of *S*-waves versus distance for a typical earthquake

Figure 12 is a plot of the travel time for the *P*-wave and *S*-wave as a function of distance from an earthquake. We make the following assumptions about a typical destructive earthquake: (1) focal depth at ~ 20 km, (2) *P*-wave velocity ~ 8 km/s, and (3) *S*-wave velocity ~ 4.5 km/s. If an earthquake is located 100 km from a city, the *P*-wave arrives at the city after about 13 s, and the *S*-waves in about 22 s (Fig. 12). If we deploy a dense seismic network near the earthquake source area (capable of locating and determining the size of the event in about 10 s), we will have about 3 s to issue the warning before the *P*-wave arrives, and about 12 s before the more destructive *S*-waves and surface waves arrive at the city. We have assumed that it takes negligible time to send a signal from the seismic network to the city via electromagnetic waves, which travel at about one-third the velocity of light or faster (between about 100,000 and 300,000 km/s depending on the method of transmission).

From Fig. 12 it is clear that this strategy may work for earthquakes located at least about 60 km from the urban area. For earthquakes at shorter distances (~ 20 to

$\sim$ 60 km), we must reduce the time needed to detect the event and issue a warning to well under 10 s. This requirement implies that we must deploy a very dense seismic network very close to the fault and estimate the necessary parameters very fast. However, such dense networks are not economical to deploy using existing seismic instruments.

For earthquakes within 20 km of a city, there is little one can do other than installing motion-sensitive automatic shut-off devices at critical facilities (natural gas, for example) and hope that they are either very quick when responding to *S*-waves or are triggered by the onset of the *P*-wave. Normally an earthquake rupture more than $\sim$ 100 km from an urban area does not commonly pose a large threat (seismic waves would be attenuated and spread out farther). There are exceptions caused either by unusual local site conditions, such as Mexico City, or by earthquakes with large rupture zones which therefore radiate efficiently to greater distances.

Design Considerations for EEW Systems

In the above discussion, we have assumed that one implements an earthquake early warning system with a traditional seismic network. Such EEW systems have limitation as illustrated by Fig. 13, which shows the expected early warning times for a repeat of the 1999 Chi–Chi earthquake. However, Nakamura and his colleagues have been successful in applying a single-station approach [84,99], where seismic signals are recorded and processed locally by the seismograph and an earthquake warning is issued whenever ground motions there exceed some trigger threshold. We will next discuss these two basic approaches, regional versus on-site in designing an earthquake early warning system.

Earthquake early warning capability can be implemented through a rapid reporting system (RRS) from a traditional network, assuming real-time telemetry into the network's central laboratory. This type of system provides, to populated areas and other sensitive locations, primary event information (hypocenter, magnitude, ground shaking intensities, and potential damage) about one minute after the earthquake begins. The RRS transmits this critical information electronically to emergency response agencies and other interested organizations and to individuals. Each recipient can then take action (some of which may be pre-programmed) shortly after the earthquake begins. Response measures can include the timely dispatch of rescue equipment and emergency supplies to the likely areas of damage.

California's ShakeMap [119,120], Taiwan's CWB, and Japan's JMA systems are typical examples of RSS. In the case of the Taiwan RRS, the CWB has, since 1995, provided intensity maps, hypocenters, and magnitudes within one minute of the occurrence of $M > 4$ earthquakes [110,128]. This system's reliability, documented by electronic messages to government agencies and scientists, has been close to perfect, particularly for large, damaging earthquakes. Figure 14 shows a block diagram of the Taiwan RRS, and details may be found in [128].

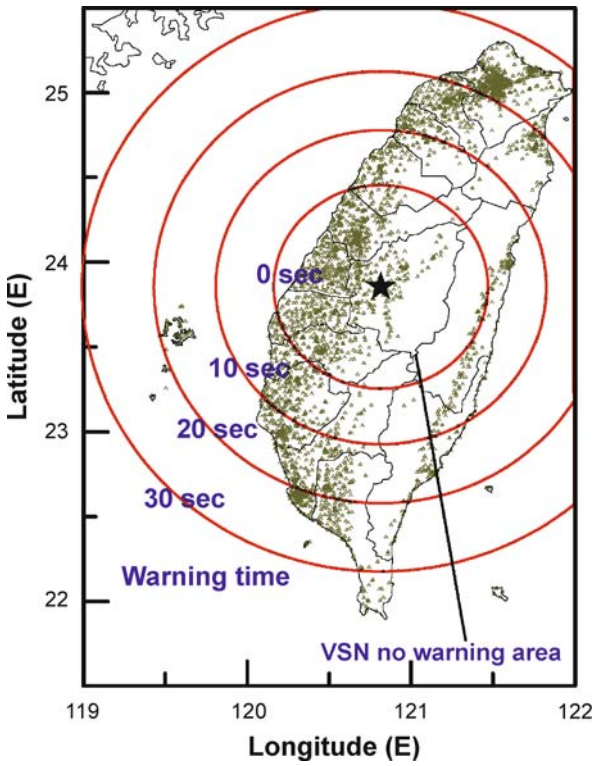

Earthquake Monitoring and Early Warning Systems, Figure 13
Expected EWS early warning times (indicated by *circles*) in Taiwan with respect to the occurrence of an event similar to the Chi–Chi earthquake of 20 September 1999. *Triangles* are locations of elementary schools, which can be regarded as a good indicator for the population density of Taiwan

Using a set of empirical relationships derived from the large data set collected during the 1999 Chi–Chi earthquake, CWB now releases, within a few minutes of an event, the estimated distributions of PGA and PGV, refined magnitudes, and damage estimates [129]. This near-real-time damage assessment is useful for rapid post-disaster emergency response and rescue missions.

Regional Warning Versus Onsite Warning

Two approaches have been adopted for earthquake early warning systems: (1) regional warning, and (2) on-site

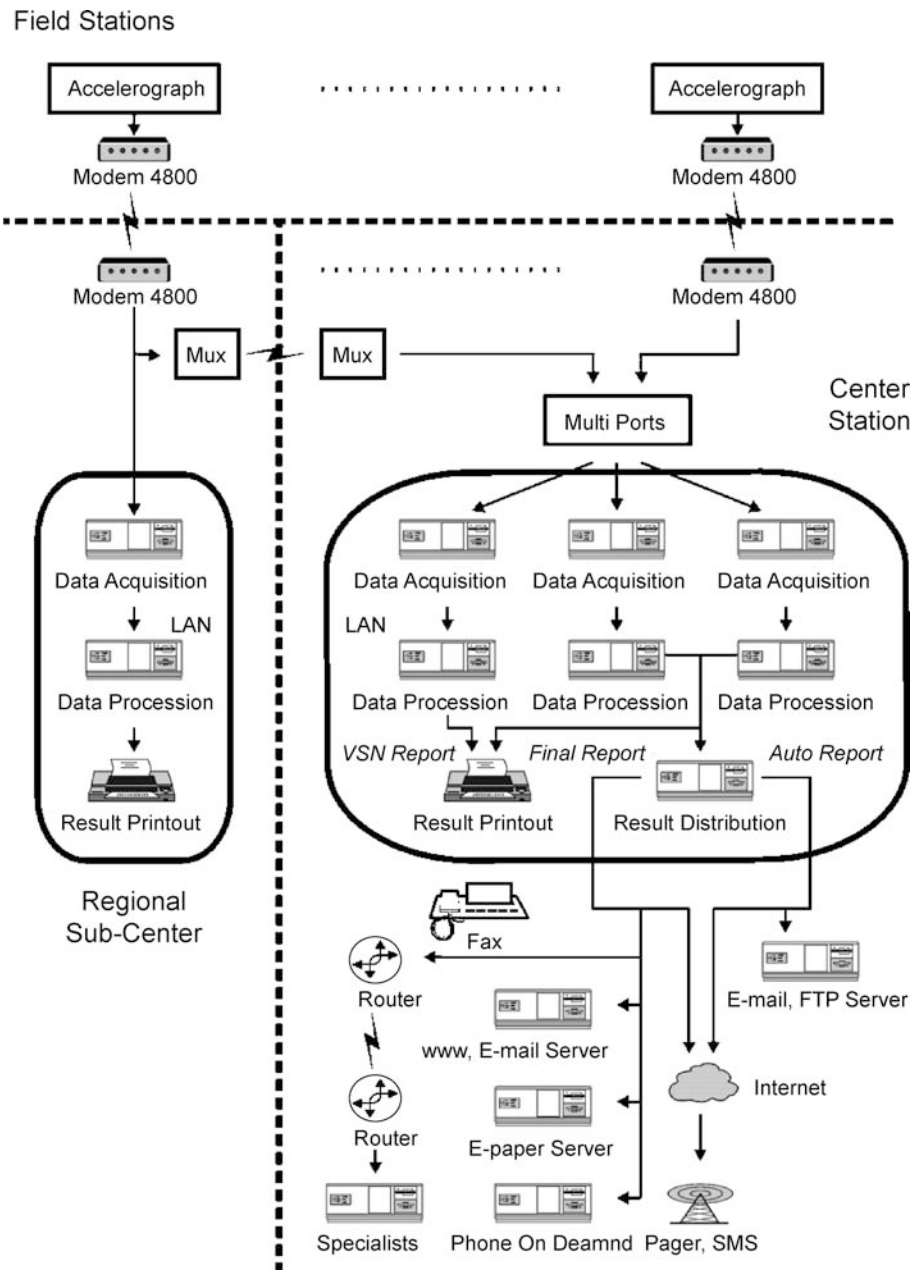

Earthquake Monitoring and Early Warning Systems, Figure 14
A block diagram showing the hardware of the Taiwan Earthquake Rapid Reporting System

warning. The first approach relies on traditional seismological methods in which data from a seismic network are used to locate an earthquake, determine the magnitude, and estimate the ground motion in the region involved. In the second approach, the initial ground motions (mainly *P* wave) observed at a site are used to predict the ensuing ground motions (mainly *S* and surface waves) at the same site.

The regional approach is more comprehensive, but takes a longer time to issue an earthquake warning. An advantage of this approach is that estimates of the timing of expected strong motions throughout the affected region can be predicted more reliably. The early warning system in Taiwan is a typical example and it uses a regional warning system called virtual sub-network approach (VSN) that requires an average of 22 s to determine earthquake

parameters with magnitude uncertainties of ±0.25. It provides a warning for areas beyond about 70 km from the epicenter (Fig. 13). This system has been in operation since 2002 with almost no false alarms [129]. With the advancement of new methodology and more dense seismic networks, regional systems are beginning to be able to provide early warnings to areas closer to the earthquake epicenter.

The regional approach has also been used in other areas. The method used in Mexico [25] is slightly different from the traditional seismological method. It is a special case of EEW system due to the relatively large distance (about 300 km in this case) between the earthquake source region (west coast of Central America) and the warning site (Mexico City). However, the warning is conceptually "regional".

In Japan, various EEW techniques have been developed and deployed by the National Research Institute for Earth Science and Disaster Prevention (NIED) and Japan Meteorological Agency (JMA) since 2000 [49,57,89], ► Tsunami Forecasting and Warning. In particular, JMA has started sending early warning messages to potential users responsible for emergency responses [50]. The potential users include railway systems, construction companies, and others; and they are familiar with the implications of early warning messages, as well as the technical limitations of EEW [57].

Some Recent EEW Advances

Allen and Kanamori [7] proposed the Earthquake Alarm System (ElarmS) to issue an earthquake warning based on information determined from the *P*-wave arrival only. Kanamori [61] extended the method of Nakamura [84] and Allen and Kanamori [7] to determine a period parameter, τ_c, from the initial 3 s of the *P* wave. τ_c is defined as

$$\tau_c = 2\pi/\sqrt{r} \tag{15}$$

where

$$r = \frac{\int_0^{\tau_0} \dot{u}^2(t)\,\mathrm{d}t}{\int_0^{\tau_0} u^2(t)\,\mathrm{d}t} \tag{16}$$

$u(t)$ is the ground-motion displacement; τ_0 is the duration of record used (usually 3 s), and τ_c, which represents the size of an earthquake, can be computed from the incoming data sequentially.

The τ_c method was used for earthquake early warning in southern California, Taiwan, and Japan by Wu and Kanamori [124,125,126] and Wu et al. [130]. At a given site, the magnitude of an event is estimated from τ_c and the peak ground-motion velocity (*PGV*) from P_d (the peak amplitude of displacement in the first 3 s after the arrival of the *P* wave). The incoming 3-component signals are recursively converted to ground acceleration, velocity and displacement. The displacements are recursively filtered using an accusal Butterworth high-pass filter with a cut-off frequency of 0.075 Hz, and a *P*-wave threshold trigger is constantly monitored. When a trigger occurs, τ_c and P_d are computed. The relationships between τ_c and magnitude (*M*), and P_d and peak ground velocity (*PGV*) for southern California, Taiwan, and Japan were investigated. Figure 15 shows a good correlation between τ_c and M_W from the K-NET records in Japan, and Fig. 16 shows the P_d versus PGV plot for southern California, Taiwan, and Japan. These relationships may be used to detect the occurrence of a large earthquake and provide onsite warning in the area immediately around the station where the onset of strong ground motion is expected within a few seconds after the arrival of the *P*-wave. When the station density is high, the onsite warning methods may be applied to data from multiple stations to increase the robustness of an onsite early warning, and to complement the regional warning approach. In an ideal situation, such warnings would be available within 10 s of the origin time of a large earthquake whose subsequent ground motion may last for tens of seconds.

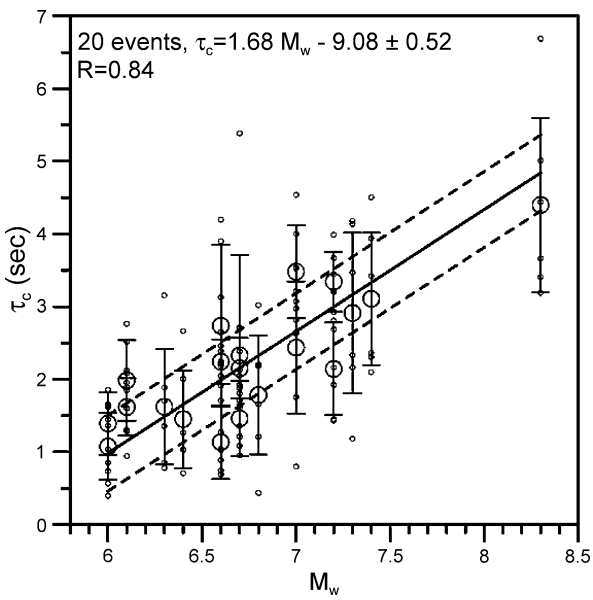

Earthquake Monitoring and Early Warning Systems, Figure 15
τ_c estimates from 20 events using the nearest 6 stations of the K-NET. *Small open circles* show single-record results, and large circles show event-average values with one standard deviation bars. *Solid line* shows the least squares fit to the event-average values, and the *two dashed lines* show the range of one standard deviation

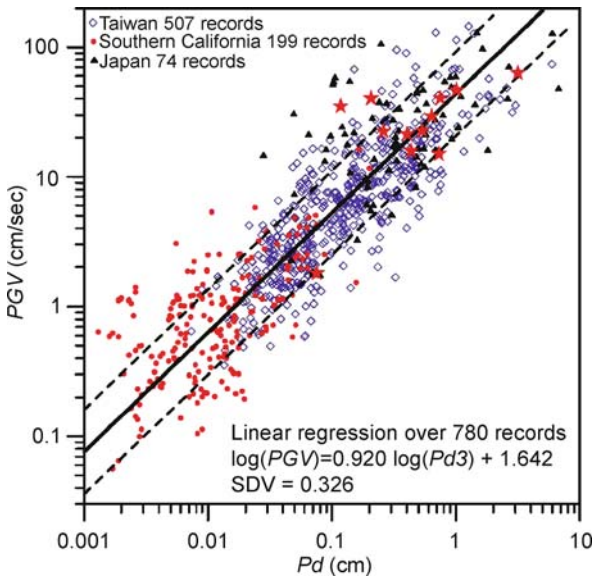

Earthquake Monitoring and Early Warning Systems, Figure 16
Relationship between peak initial displacement amplitude (P_d) measurements and peak ground velocity (*PGV*) for the records with epicentral distances less than 30 km from the epicenter in Southern California (*red solid circles*), Taiwan (*blue diamonds*) and Japan (*black solid triangles*). *Solid line* shows the least squares fit and the two *dashed lines* show the range of one standard deviation

Wu and Zhao [127] investigated the attenuation of P_d with the hypocentral distance R in southern California as a function of magnitude M, and obtained the following relationships:

$$M_{P_d} = 4.748 + 1.371 \times \log(P_d) + 1.883 \times \log(R) \quad (17)$$

and

$$\log(P_d) = -3.463 + 0.729 \times M - 1.374 \times \log(R)\,. \quad (18)$$

For the regional warning approach, when an earthquake location is determined by the P-wave arrival times at stations close to the epicenter, this relationship can be used to estimate the earthquake magnitude. Their result shows that for earthquakes in southern California the P_d magnitudes agree with the catalog magnitudes with a standard deviation of 0.18 for events less than magnitude 6.5. They concluded that P_d is a robust measurement for estimating the magnitudes of earthquakes for regional early warning purposes in southern California. This method has also applied to Italian region by Zollo et al. [132] with a very good performance.

Because the on-site approach is faster than the regional approach, it can provide useful early warning to sites at short distances from the earthquake epicenter where early warning is most needed. Onsite early warning can be generated by either a single station or by a dense array. For a single station operation, signals from P-waves are used for magnitude and hypocenter determination to predict strong ground shaking. Nakamura [83] first proposed this concept, developed the Urgent Earthquake Detection and Alarm System or UrEDAS [86], and introduced a simple strong-motion index for onsite EEW [85]. However, the reliability of on-site earthquake information is generally less than that obtained with the regional warning system. There currently is a trade-off between warning time and the reliability of the earthquake information. Generally, an information updating procedure is necessary for any EEW system. On-site warning methods can be especially useful in regions where a dense seismic network is deployed.

The Japan Meteorological Agency (JMA) began distribution of earthquake early warning information to the public in October 1, 2007 through several means, such as TV and radio [50] (http://www.jma.go.jp/jma/en/Activities/eew.html). The JMA system was successfully activated during the recent Noto Hanto and Niigata Chuetsu–Oki earthquakes in 2007, and provided accurate information of hypocenter, magnitude, and intensity about 3.8 s after the arrival of P-waves at nearby stations. The warning message reached sites further than about 30 km from the epicenter as an early warning alert (i. e., information arrived before shaking started at the site). This is a remarkable performance of the system for damaging earthquakes and gives promise of an early warning system as a practical means for earthquake damage mitigation. Although warning alert is most needed within 30 km of the epicenter, it is not feasible with the current density and configuration of the JMA network.

Lawrence and Cochran [68] proposed a collaborative project for rapid earthquake response and early warning by using the accelerometers that are already installed inside many laptop computers. Their Quake-Catcher Network (QCN) will employ existing laptops, which have accelerometers already installed, and desktops outfitted with inexpensive (under $ 50) USB accelerometers to form the world's largest high-density, distributed computing seismic network for monitoring strong ground motions (http://qcn.stanford.edu/). By freely distributing the necessary software, anyone having a computer with an Internet connection can join the project as a collaborative member. The Quake-Catcher Network also has the potential to provide better understanding of earthquakes, and the client-based software is also intended to be educational, with instructive material displaying the current seismic signal and/or recent earthquakes in the region. It

is an effective way to bring earthquake awareness to students and the general public.

Future Directions

To be successful, monitoring earthquakes requires large, stable funding over a long period of time. The most direct argument for governments to support long-term earthquake monitoring is to collect scientific data for hazard mitigation. In the past two decades about half a million of human lives have been lost due to earthquakes, and economic losses from earthquake damage total about $ 200 billion. Future losses will be even greater as rapid urbanization is taking place worldwide. For example, the recent Japanese Fundamental Seismic Survey and Observation Plan (costing several hundred million US dollars) is a direct response to the economic losses of about $ 100 billion due to the 1995 Kobe earthquake. In addition to scientific and technological challenges in monitoring earthquakes, seismologists must pay attention to achieve (1) stable long-term funding, (2) effective management and execution, and (3) delivery of useful products to the users.

Seismologists benefit greatly from scientific and technological advances in other fields. For example, Global Positioning Systems (GPS) open a new window for monitoring crustal deformation which is important to understand the driving forces that generate earthquakes (► GPS: Applications in Crustal Deformation Monitoring, ► Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of). Under the US Earth Scope Program (http://www.earthscope.org/) the Plate Boundary Observatory (PBO) is covering the western Northern America and Alaska with a network of high precision GPS and strain-meter stations in order to measure deformation across the active boundary between the Pacific and North America plates (http://www.earthscope.org/observatories/pbo). As the sampling rate of GPS data increases, they can provide time histories of displacement during an earthquake. Monitoring earthquakes with multiple types of instruments and sensors is now increasingly popular, and "integrated" or "super" stations are increasingly common. Figure 17 shows an example of an integrated station (HGSD) in eastern Taiwan. Instruments deployed at the HGSD station in eastern Taiwan include: a broadband seismometer, a continuous GPS instrument, a strain-meter, and a 6-channel accelerograph (Model K2 by Kinemetrics) with an internal accelerometer and a rotational sensor (Model R-1 by eentec). A digital seismogram recorded at the HGSD station from an earthquake ($M_W = 5.1$) of July 23, 2007 at a distance of 34 km is

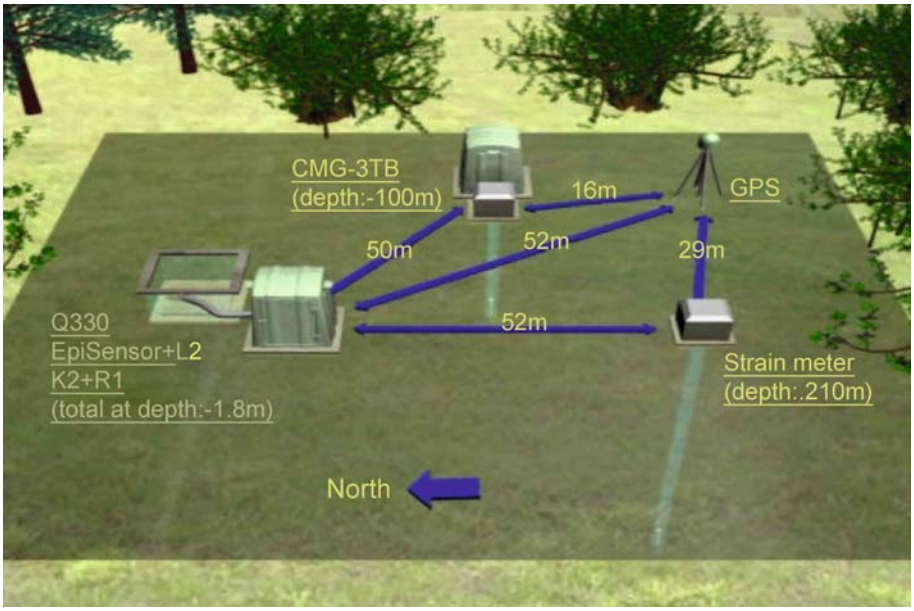

Earthquake Monitoring and Early Warning Systems, Figure 17
Instruments deployed at the HGSD station in eastern Taiwan. Clockwise from the *top*: (1) A broadband seismometer (Model CMG-3TB) installed at a depth of 100 m), (2) A continuous GPS instrument, (3) A strain-meter installed at a depth of 210 m), (4) A Model Q330 6-channel recorder with an accelerometer (Model EpiSensor) and a short-period seismometer (Model L2), and (5) A Model K2 6-channel accelerograph with an internal accelerometer and a rotational sensor (Model R-1)

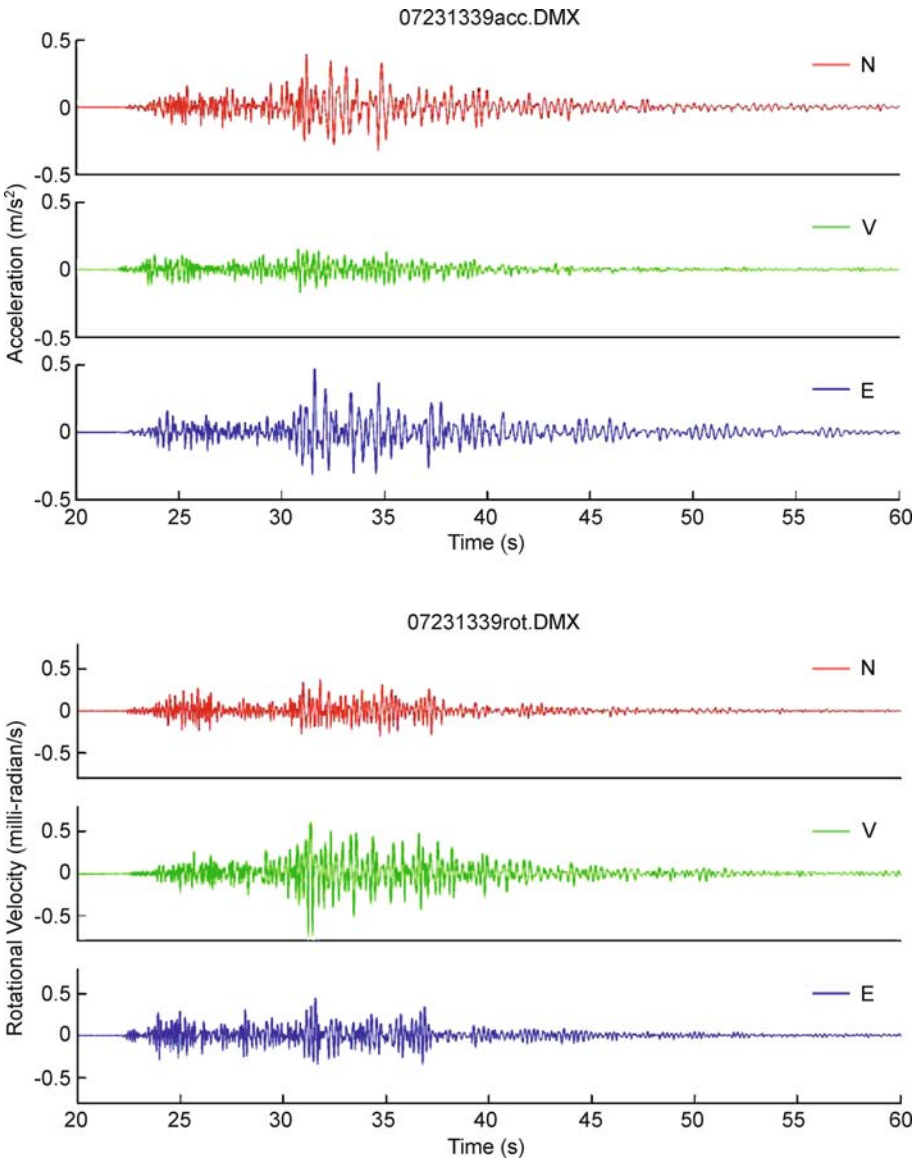

Earthquake Monitoring and Early Warning Systems, Figure 18

A digital seismogram recorded at the HGSD station from an earthquake ($M_W = 5.1$) of July 23, 2007 at a distance of 34 km. *Top frame*: 3-component translational accelerations. *Bottom frame*: 3-component rotation velocity motions. N = North-South; V = Vertical, and E = East-West

shown in Fig. 18. The importance of rotational seismology and its current status are given in the Appendix.

A radically different design of seismographic networks (and earthquake early warning system in particular) is now possible using the "Sensor Network" developed by Intel Research. Intel is working with the academic community and industry collaborators to actively explore the potential of wireless sensor networks. This research is already demonstrating the potential for this new technology to enhance public safety, reduce the cost of doing business, and bring a host of other benefits to business and society (http://www.intel.com/research/exploratory/wireless_sensors.htm).

It has been very difficult historically to obtain adequate and stable funding for long-term earthquake monitoring, largely because disastrous earthquakes occur infrequently. Since there are many pressing problems facing modern societies, almost all governments react to earthquake (and

tsunami) disasters only after the fact, and even then for relatively short periods of time. To advance earthquake prediction research and to develop effective earthquake warning systems will require continuous earthquake monitoring with extensive instrumentations in the near-field for decades and even centuries. Therefore, innovative approaches must be developed and perseverance is needed.

Acknowledgments

We thank John Evans, Fred Klein, Woody Savage, and Chris Stephens for reviewing the manuscript, their comments and suggestions greatly improved it. We are grateful to Lind Gee and Bob Hutt for information about the Global Seismographic Network (GSN) and for providing a high-resolution graphic file of an up-to-date GSN station map.

Appendix: A Progress Report on Rotational Seismology

Seismology is based primarily on the observation and modeling of three orthogonal components of translational ground motions. Although effects of rotational motions due to earthquakes have long been observed (e. g., [80]), Richter (see, p. 213 in [97]) stated that:

> *Perfectly general motion would also involve rotations about three perpendicular axes, and three more instruments for these. Theory indicates, and observation confirms, that such rotations are negligible.*

However, Richter provided no references for this claim, and the available instruments at that time did not have the sensitivity to measure the very small rotation motions that the classical elasticity theory predicts.

Some theoretical seismologists (e. g., [4,5]) and earthquake engineers have argued for decades that the rotational part of ground motions should also be recorded. It is well known that standard seismometers and accelerometers are profoundly sensitive to rotations, particularly tilt, and therefore subject to rotation-induced errors (see e. g., [39,40,41,93]). The paucity of instrumental observations of rotational ground motions is mainly the result of the fact that, until recently, the rotational sensors did not have sufficient resolution to measure small rotational motions due to earthquakes.

Measurement of rotational motions has implications for: (1) recovering the complete ground-displacement history from seismometer recordings; (2) further constraining earthquake rupture properties; (3) extracting information about subsurface properties; and (4) providing additional ground motion information to engineers for seismic design.

In this Appendix, we will first briefly review elastic wave propagation that is based on the linear elasticity theory of simple homogeneous materials under infinitesimal strain. This theory was developed mostly in the early nineteenth century: the differential equations of the linear elastic theory were first derived by Louis Navier in 1821, and Augustin Cauchy gave his formulation in 1822 that remains virtually unchanged to the present day [103]. From this theory, Simeon Poisson demonstrated in 1828 the existence of longitudinal and transverse elastic waves, and in 1885, Lord Rayleigh confirmed the existence of elastic surface waves. George Green put this theory on a physical basis by introducing the concept of strain energy, and, in 1837, derived the basic equations of elasticity from the principle of energy conservation. In 1897, Richard Oldham first identified these three types of waves in seismograms, and linear elasticity theory has been embedded in seismology ever since.

In the following we summarize recent progress in rotational seismology and the need to include measurements of rotational ground motions in earthquake monitoring. The monograph by Teisseyre et al. [109] provides a useful summary of rotational seismology.

Elastic Wave Propagation

The equations of motion for a homogeneous, isotropic, and initially unstressed elastic body may be obtained using the conservation principles of continuum mechanics (e. g., [30]) as

$$\rho \frac{\partial^2 u_i}{\partial t^2} = (\lambda + \mu)\frac{\partial \theta}{\partial x_i} + \mu \nabla^2 u_i, \quad i = 1, 2, 3 \tag{A1}$$

and

$$\theta = \sum\nolimits_j \partial u_j / \partial x_j \tag{A2}$$

where θ is the dilatation, ρ is the density, u_i is the ith component of the displacement vector $\vec{u}$, t is the time, and λ and μ are the elastic constants of the media. Eq. (A1) may be rewritten in vector form as

$$\rho(\partial^2 \vec{u}/\partial t^2) = (\lambda + \mu)\nabla(\nabla \bullet \vec{u}) + \mu \nabla^2 \vec{u}\,. \tag{A3}$$

If we differentiate both sides of Eq. (A1) with respect to x_i, sum over the three components, and bring ρ to the right-hand side, we obtain

$$\partial^2 \theta / \partial t^2 = [(\lambda + 2\mu)/\rho]\nabla^2 \theta\,. \tag{A4}$$

If we apply the curl operator ($\nabla\times$) to both sides of Eq. (A3), and note that

$$\nabla \bullet (\nabla \times \vec{u}) = 0 \tag{A5}$$

we obtain

$$\partial^2(\nabla \times \vec{u})/\partial t^2 = (\mu/\rho)\nabla^2(\nabla \times \vec{u}) \,. \tag{A6}$$

Now Eqs. (A4) and (A6) are in the form of the classical wave equation

$$\partial^2\Psi/\partial t^2 = v^2\nabla^2\Psi \,, \tag{A7}$$

where Ψ is the wave potential, and v is the wave-propagation velocity (a pseudovector; wave slowness is a proper vector). Thus a dilatational disturbance θ (or a compressional wave) may be transmitted through a homogenous elastic body with a velocity V_P where

$$V_P = \sqrt{[(\lambda + 2\mu)/\rho]} \tag{A8}$$

according to Eq. (A4), and a rotational disturbance $\nabla \times \vec{u}$ (or a shear wave) may be transmitted with a wave velocity V_S where

$$V_S = \sqrt{\mu/\rho} \tag{A9}$$

according to Eq. (A6). In seismology, and for historical reasons, these two types of waves are called the primary (P) and the secondary (S) waves, respectively.

For a heterogeneous, isotropic, and elastic medium, the equation of motion is more complex than Eq. (A3), and is given by Karal and Keller [65] as

$$\begin{aligned}\rho(\partial^2\vec{u}/\partial t^2) &= (\lambda + \mu)\nabla(\nabla \bullet \vec{u}) + \mu\nabla^2\vec{u} \\ &\quad + \nabla\lambda(\nabla \bullet \vec{u}) + \nabla\mu \times (\nabla \times \vec{u}) + 2(\nabla\mu \bullet \nabla)\vec{u} \,.\end{aligned} \tag{A10}$$

Furthermore, the compressional wave motion is no longer purely longitudinal, and the shear wave motion is no longer purely transverse. A review of seismic wave propagation and imaging in complex media may be found in the entry by Igel et al. ▶ Seismic Wave Propagation in Media with Complex Geometries, Simulation of.

A significant portion of seismological research is based on the solution of the elastic wave equations with the appropriate initial and boundary conditions. However, explicit and unique solutions are rare, except for a few simple problems. One approach is to transform the wave equation to the eikonal equation and seek solutions in terms of wave fronts and rays that are valid at high frequencies. Another approach is to develop through specific boundary conditions a solution in terms of normal modes [77]. Although ray theory is only an approximation [17], the classic work of Jeffreys and Bullen, and Gutenberg used it to determine Earth structure and locate earthquakes that occurred in the first half of the 20th century. It remains a principal tool used by seismologists even today. Impressive developments in normal mode and surface wave studies (in both theory and observation) started in the second half of the 20th century, leading to realistic quantification of earthquakes using moment tensor methodology [21].

Rotational Ground Motions

Rotations in ground motion and in structural responses have been deduced indirectly from accelerometer arrays, but such estimates are valid only for long wavelengths compared to the distances between sensors (e. g., [16,34,52,88,90,104]). The rotational components of ground motion have also been estimated theoretically using kinematic source models and linear elastodynamic theory of wave propagation in elastic solids [14,69,70,111].

In the past decade, rotational motions from teleseismic and small local earthquakes were also successfully recorded by sensitive rotational sensors, in Japan, Poland, Germany, New Zealand, and Taiwan (e. g., [53,55,56,105,106,107,108]). The observations in Japan and Taiwan show that the amplitudes of rotations can be *one to two orders of magnitude greater than expected* from the classical linear theory. Theoretical work has also suggested that, in granular materials or cracked continua, asymmetries of the stress and strain fields can create rotations in addition to those predicted by the classical elastodynamic theory for a perfect continuum (▶ Earthquake Source: Asymmetry and Rotation Effects).

Because of lack of instrumentation, rotational motions have not yet been recorded in the near-field (within ~ 25 km of fault ruptures) of strong earthquakes (magnitude > 6.5), where the discrepancy between observations and theoretical predictions may be the largest. Recording such ground motions will require extensive seismic instrumentation along some well-chosen active faults and luck. To this end, several seismologists have been advocating such measurements, and a current deployment in southwestern Taiwan by its Central Weather Bureau is designed to "capture" a repeat of the 1906 Meishan earthquake (magnitude 7.1) with both translational and rotational instruments.

Rotations in structural response, and the contributions to the response from the rotational components of the ground motion, have also been of interest for many decades (e. g., [78,87,98]. Recent reviews on rotational motions in seismology and on the effects of the rota-

tional components of ground motion on structures can be found, for examples, in Cochard et al. [18] and Pillet and Virieux [93], and Trifunac [112], respectively.

Growing Interest – The IWGoRS

Various factors have led to spontaneous organization within the scientific and engineering communities interested in rotational motions. Such factors include: the growing number of successful direct measurements of rotational ground motions (e. g., by ring laser gyros, fiber optic gyros, and sensors based on electro-chemical technology); increasing awareness about the usefulness of the information they provide (e. g., in constraining the earthquake rupture properties, extracting information about subsurface properties, and about deformation of structures during seismic and other excitation); and a greater appreciation for the limitations on information that can be extracted from the translational sensors due to their sensitivity to rotational motions e. g., computation of permanent displacements from accelerograms (e. g., [13,39,40,41,93,113]).

A small workshop on Rotational Seismology was organized by W.H.K. Lee, K. Hudnut, and J.R. Evans of the USGS on 16 February 2006 in response to grassroots interest. It was held at the USGS offices in Menlo Park and in Pasadena, California, with about 30 participants from about a dozen institutions participating via teleconferencing and telephone [27]. This event led to the formation of the *International Working Group on Rotational Seismology* in 2006, inaugurated at a luncheon during the AGU 2006 Fall Meeting in San Francisco.

The *International Working Group on Rotational Seismology* (IWGoRS) aims to promote investigations of rotational motions and their implications, and the sharing of experience, data, software and results in an open web-based environment (http://www.rotational-seismology.org). It consists of volunteers and has no official status. H. Igel and W.H.K. Lee currently serve as "co-organizers". Its charter is accessible on the IWGoRS web site. The Working Group has a number of active members leading task groups that focus on the organization of workshops and scientific projects, including: testing and verifying rotational sensors, broadband observations with ring laser systems, and developing a field laboratory for rotational motions. The IWGoRS web site also contains the presentations and posters from related meetings, and eventually will provide access to rotational data from many sources.

The IWGoRS organized a special session on *Rotational Motions in Seismology*, convened by H. Igel, W.H.K. Lee, and M. Todorovska during the 2006 AGU Fall Meeting [76]. The goal of that session was to discuss rotational sensors, observations, modeling, theoretical aspects, and potential applications of rotational ground motions. A total of 21 papers were submitted for this session, and over 100 individuals attended the oral session.

The large attendance at this session reflected common interests in rotational motions from a wide range of geophysical disciplines, including strong-motion seismology, exploration geophysics, broadband seismology, earthquake engineering, earthquake physics, seismic instrumentation, seismic hazards, geodesy, and astrophysics, thus confirming the timeliness of IWGoRS. It became apparent that to establish an effective international collaboration within the IWGoRS, a larger workshop was needed to allow sufficient time to discuss the many issues of interest, and to draft research plans for rotational seismology and engineering applications.

First International Workshop

The *First International Workshop on Rotational Seismology and Engineering Applications* was held in Menlo Park, California, on 18–19 September 2007. This workshop was hosted by the US Geological Survey (USGS), which recognized this topic as a new research frontier for enabling a better understanding of the earthquake process and for the reduction of seismic hazards. The technical program consisted of three presentation sessions: plenary (4 papers) and oral (6 papers) held during the first day, and poster (30 papers) held during the morning of the second day. A post-workshop session was held on the morning of September 20, in which scientists of the Laser Interferometer Gravitational-wave Observatory (LIGO) presented their work on seismic isolation of their ultra-high precision facility, which requires very accurate recording of translational and rotational components of ground motions (3 papers). Proceedings of this Workshop were released in Lee et al. [75] with a DVD disc that contains all the presentation files and supplementary information.

One afternoon of the workshop was devoted to in-depth discussions on the key outstanding issues and future directions. The participants could join one of five panels on the following topics: (1) theoretical studies of rotational motions (chaired by L. Knopoff), (2) measuring far-field rotational motions (chaired by H. Igel), (3) measuring near-field rotational motions (chaired by T.L. Teng), (4) engineering applications of rotational motions (chaired by M.D. Trifunac), and (5) instrument design and testing (chaired by J.R. Evans). The panel reports on key issues and unsolved problems, and on research strategies and plans, can be found in Appendices

2.1 through 2.5 in Lee et al. [75]. Following the in-depth group discussions, the panel chairs reported on the group discussions in a common session, with further discussions among all the participants.

Discussions

Since rotational ground motions may play a significant role in the near-field of earthquakes, rotational seismology has emerged as a new frontier of research. During the Workshop discussions, L. Knopoff asked: Is there a quadratic rotation-energy relation, in the spirit of Green's strain-energy relation, coupled to it or independent of it? Can we write a rotation-torque formula analogous to Hooke's law for linear elasticity in the form

$$L_{ij} = d_{ijkl}\omega_{kl} \tag{A11}$$

where ω_{kl} is the rotation,

$$\omega_{kl} = \tfrac{1}{2}(u_{k,l} - u_{l,k})\,. \tag{A12}$$

L_{ij} is the torque density; and d_{ijkl} are the coefficients of rotational elasticity? How are the d's related to the usual c's of elasticity? If we define the rotation vector as

$$\vec{\Omega} = \tfrac{1}{2}(\nabla \times \vec{u}) \tag{A13}$$

we obtain

$$-V_s^2 \nabla \times (\nabla \times \vec{\Omega}) = \partial^2 \vec{\Omega}/\partial t^2 - \tfrac{1}{2}\rho^{-1}(\nabla \times \vec{f}) \tag{A14}$$

where the torque density is $\nabla \times \vec{f}$, $\vec{f}$ is the body force density, and ρ is density of the medium. This shows that rotational waves propagate with S-wave velocity and that it may be possible to store torques. Eq. (15) is essentially an extension using the classical elasticity theory.

Lakes [67] pointed out that the behavior of solids can be represented by a variety of continuum theories. In particular, the elasticity theory of the Cosserat brothers [19] incorporates (1) a local rotation of points as well as the translation motion assumed in the classical theory, and (2) a couple stress (a torque per unit area) as well as the force stress (force per unit area). In the constitutive equation for the classical elasticity theory, there are two independent elastic constants, whereas for the Cosserat elastic theory there are six. Lakes (personal communication, 2007) advocates that there is substantial potential for using generalized continuum theories in geo-mechanics, and any theory must have a strong link with experiment (to determine the constants in the constitutive equation) and with physical reality.

Indeed some steps towards better understandings of rotational motions have taken place. For example, Twiss et al. [114] argued that brittle deformation of the Earth's crust (► Brittle Tectonics: A Non-linear Dynamical System) involving block rotations is comparable to the deformation of a granular material, with fault blocks acting like the grains. They realized the inadequacy of classical continuum mechanics and applied the Cosserat or micropolar continuum theory to take into account two separate scales of motions: macro-motion (large-scale average motion composed of macrostrain rate and macrospin), and micro-motion (local motion composed of microspin). A theoretical link is then established between the kinematics of crustal deformation involving block rotations and the effects on the seismic moment tensor and focal mechanism solutions.

Recognizing that rotational seismology is an emerging field, the *Bulletin of Seismological Society of America* will be publishing in 2009 a special issue under the guest editorship of W.H.K. Lee, M. Çelebi, M.I. Todorovska, and H. Igel.

Bibliography

Primary Literature

1. Ahern TK (2003) The FDSN and IRIS Data Management System: providing easy access to terabytes of information. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part B. Academic Press, Amsterdam pp 1645–1655
2. Aki K (1966) Generation and propagation of G waves from the Niigata Earthquake of June 16, 1964: Part 1. A statistical analysis. Bull Earthq Res Inst 44:23–72; Part 2. Estimation of earthquake moment, released energy, and stress-strain drop from the G wave spectrum. Bull Earthq Res Inst 44:73–88
3. Aki K (1969) Analysis of the seismic coda of local earthquakes as scattered waves. J Geophys Res 74:6215–6231
4. Aki K, Richards PG (1980) Quantitative Seismology. W.H. Freeman, San Francisco
5. Aki K, Richards PG (2002) Quantitative Seismology: Theory and Methods, 2nd edn. University Science Books, Sausalito
6. Allen RM (2007) Earthquake hazard mitigation: New directions and opportunities. In: Kanamori H (ed) Earthquake Seismology. Treatise on Geophysics, vol 4. Elsevier, Amsterdam, pp 607–648
7. Allen RM, Kanamori H (2003) The potential for earthquake early warning in Southern California. Science 300:786–789
8. Ambraseys NN, Jackson JA, Melville CP (2002) Historical seismicity and tectonics: The case of the Eastern Mediterranean and the Middle East. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part A. Academic Press, Amsterdam, pp 747–763
9. Anderson JG (2003) Strong-motion seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International

Handbook of Earthquake and Engineering Seismology, Part B. Academic Press, Amsterdam pp 937–965
10. Berger J, Davis P, Ekström G (2004) Ambient earth noise: A survey of the Global Seismographic Network. J Geophys Res 109:B11307
11. Borcherdt RD (ed) (1997) Vision for the future of the US National Strong-Motion Program, The committee for the future of the US National Strong Motion Program. US Geol Surv Open-File Rept B97530
12. Bormann P (ed) (2002) New Manual of Seismological Observatory Practice. GeoForschungsZentrum Potsdam http://www.gfz-potsdam.de/bib/nmsop_formular.html
13. Boroschek R, Legrand D (2006) Tilt motion effects on the double-time integration of linear accelerometers: an experimental approach. Bull Seism Soc Am 96:2072–2089
14. Bouchon M, Aki K (1982) Strain, tilt, and rotation associated with strong ground motion in the vicinity of earthquake faults. Bull Seism Soc Am 72:1717–1738
15. Burridge R, Knopoff L (1964) Body force equivalents for seismic dislocations. Bull Seism Soc Am 54:1875–1888
16. Castellani A, Boffi G (1986) Rotational components of the surface ground motion during an earthquake. Earthq Eng Struct Dyn 14:751–767
17. Chapman CH (2002) Seismic ray theory and finite frequency extensions. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part A. Academic Press, Amsterdam, pp 103–123
18. Cochard A, Igel H, Schuberth B, Suryanto W, Velikoseltsev A, Schreiber U, Wassermann J, Scherbaum F, Vollmer D (2006) Rotational motions in seismology: theory, observation, simulation. In: Teisseyre R, Takeo M, Majewski M (eds) Earthquake Source Asymmetry, Structural Media and Rotation Effects. Springer, Heidelberg, pp 391–411
19. Cosserat E, Cosserat F (1909) Theorie des Corps Deformables. Hermann, Paris
20. Dziewonski AM, Woodhouse JH (1983) An experiment in the systematic study of global seismicity: centroid-moment tensor solutions for 201 moderate and large earthquakes of 1981. J Geophys Res 88:3247–3271
21. Dziewonski AM, Woodhouse JH (1983) Studies of the seismic source using normal-mode theory. In: Kanamori H, Boschi B (eds) Earthquakes: Observation, Theory, and Interpretation. North-Holland, Amsterdam, pp 45–137
22. Dziewonski AM, Chou TA, Woodhouse JH (1981) Determination of earthquake source parameters from waveform data for studies of global and regional seismicity. J Geophys Res 86:2825–2852
23. Ekström G (1994) Rapid earthquake analysis utilizes the Internet. Comput Phys 8:632–638
24. Engdahl ER, Villasenor A (2002) Global seismicity: 1900–1999. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part A. Academic Press, Amsterdam, pp 665–690
25. Espinosa-Aranda JM, Jimenez A, Ibarrola G, Alcantar F, Aguilar A, Inostroza M, Maldonado S (1995) Mexico City Seismic Alert System. Seism Res Lett 66(6):42–53
26. Evans JR, Hamstra RH, Kündig C, Camina P, Rogers JA (2005) TREMOR: A wireless MEMS accelerograph for dense arrays. Earthq Spectr 21(1):91–124
27. Evans JR, Cochard A, Graizer V, Huang B-S, Hudnut KW, Hutt CR, Igel H, Lee WHK, Liu C-C, Majewski E, Nigbor R, Safak E, Savage WU, Schreiber U, Teisseyre R, Trifunac M, Wassermann J, Wu C-F(2007) Report of a workshop on rotational ground motion. US Geol Surv Open File Rep 20:2007–1145 http://pubs.usgs.gov/of/2007/1145/
28. Feigl KL (2002) Estimating earthquake source parameters from geodetic measurements. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part A. Academic Press, Amsterdam, pp 607–620
29. Field EH, Hough SE (1997) The variability of PSV response spectra across a dense array deployed during the Northridge aftershock sequence. Earthq Spectr 13:243–257
30. Fung YC (1965) Foundations of Solid Mechanics. Prentice-Hall, Englewood Cliffs
31. Gasparini P, Manfredi G, Zschau J (eds) (2007) Seismic Early Warning Systems. Springer, Berlin
32. Gee L, Neuhauser D, Dreger D, Pasyanos M, Uhrhammer R, Romanowicz B (2003) The Rapid Earthquake Data Integration Project. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part B. Academic Press, Amsterdam, pp 1261–1273
33. Geiger LC (1912) Probability method for the determination of earthquake epicenters from the arrival time only. Bull St Louis Univ 8:60–71
34. Ghayamghamian MR, Nouri GR (2007) On the characteristics of ground motion rotational components using Chiba dense array data. Earthq Eng Struct Dyn 36(10):1407–1429
35. Gilbert F (1971) Excitation of the normal modes of the Earth by earthquake sources. Geophys J R Astron Soc 22:223–226
36. Gilbert F, Dziewonski AM (1975) Application of normal mode theory to the retrieval of structural parameters and source mechanisms from seismic spectra. Phil Trans Roy Soc Lond A 278:187–269
37. Goltz JD, Flores PJ, Chang SE, Atsumi T (2001) Emergency response and early recovery. In: 1999 Chi-Chi, Taiwan, Earthquake Reconnaissance Report. Earthq Spectra Suppl A 17:173–183
38. Goodstein JR (1991) Millikan's School: A History of the California Institute of Technology. Norton, New York
39. Graizer VM (1991) Inertial seismometry methods. Izv Earth Phys Akad Nauk SSSR 27(1):51–61
40. Graizer VM (2005) Effect of tilt on strong motion data processing. Soil Dyn Earthq Eng 25:197–204
41. Graizer VM (2006) Tilts in strong ground motion. Bull Seis Soc Am 96:2090–2106
42. Guidoboni E (2002) Historical seismology: the long memory of the inhabited world. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part A. Academic Press, Amsterdam, pp 775–790
43. Gutenberg B (1945) Amplitudes of surface waves and magnitudes of shallow earthquakes. Bull Seism Soc Am 35:3–12
44. Gutenberg B, Richter CF (1954) Seismicity of the Earth, 2nd edn. Princeton University Press, Princeton
45. Gutenberg B, Richter CF (1956) Magnitude and energy of earthquakes. Ann Geofis 9:1–15

46. Hanks TC, Wyss M (1972) The use of body wave spectra in the determination of seismic source parameters. Bull Seism Soc Am 62:561–589
47. Hauksson E, Jones LM, Shakal AF (2003) TriNet: a modern ground motion seismic network. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part B. Academic Press, Amsterdam, pp 1275–1284
48. Havskov J, Alguacil G (2004) Instrumentation in Earthquake Seismology. Springer, Berlin
49. Horiuchi S, Negishi H, Abe K, Kamimura A, Fujinawa Y (2005) An automatic processing system for broadcasting earthquake alarms. Bull Seism Soc Am 95:708–718
50. Hoshiba M, Kamigaichi O, Saito M, Tsukada S, Hamada N (2008) Earthquake early warning starts nationwide in Japan, EOS. Trans Am Geophys Un 89(8):73–74
51. Housner GW (2002) Historical view of earthquake engineering. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part A. Academic Press, Amsterdam, pp 13–18
52. Huang BS (2003) Ground rotational motions of the 1991 Chi-Chi, Taiwan earthquake as inferred from dense array observations. Geophys Res Lett 30(6):1307–1310
53. Huang BS, Liu CC, Lin CR, Wu CF, Lee WHK (2006) Measuring mid- and near-field rotational ground motions in Taiwan. Poster, presented at 2006 Fall AGU Meeting. San Francisco
54. Hutt CR, Bolton HF, Holcomb LG (2002) US contribution to digital global seismograph networks. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part A. Academic Press, Amsterdam, pp 319–322
55. Igel H, Cochard A, Wassermann J, Schreiber U, Velikoseltsev A, Dinh NP (2007) Broadband observations of rotational ground motions. Geophys J Int 168(1):182–197
56. Igel H, Schreiber U, Flaws A, Schuberth B, Velikoseltsev A, Cochard A (2005) Rotational motions induced by the M8.1 Tokachi-oki earthquake, September 25, 2003. Geophys Res Lett 32:L08309. doi:10.1029/2004GL022336
57. Kamigaichi O (2004) JMA Earthquake Early Warning. J Japan Assoc Earthq Eng 4:134–137
58. Kanamori H (1977) The energy release in great earthquakes. J Geophys Res 82:2921–2987
59. Kanamori H (1978) Quantification of earthquakes. Nature 271:411–414
60. Kanamori H (2003) Earthquake prediction: an overview. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part B. Academic Press, Amsterdam, pp 1205–1216
61. Kanamori H (2005) Real-time seismology and earthquake damage mitigation. Annual Rev Earth Planet Sci 33:195–214
62. Kanamori H, Brodsky EE (2000) The physics of earthquakes. Phys Today 54(6):34–40
63. Kanamori H, Rivera L (2006) Energy partitioning during an earthquake. In: Abercrombie R, McGarr A, Kanamori H, Di Toro G (eds) Earthquakes: Radiated Energy and the Physics of Faulting. Geophysical Monograph, vol 170. Am Geophys Union, Washington DC, pp 3–13
64. Kanamori H, Hauksson E, Heaton T (1997) Real-time seismology and earthquake hazard mitigation. Nature 390:461–464
65. Karal FC, Keller JB (1959) Elastic wave propagation in homogeneous and inhomogeneous media. J Acoust Soc Am 31:694–705
66. Kisslinger C, Howell BF (2003) Seismology and physics of the Earth's interior in the USA, 1900–1960. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part B. Academic Press, Amsterdam, p 1453
67. Lakes RS (1995) Experimental methods for study of Cosserat elastic solids and other generalized continua. In: Mühlhaus H (ed) Continuum Models for Materials with Micro-structure. Wiley, New York, pp 1–22
68. Lawrence JF, Cochran ES (2007) The Quake Catcher Network: Cyberinfrastructure bringing seismology into schools and homes. American Geophysical Union, Fall Meeting 2007, abstract #ED11C-0633
69. Lee VW, Trifunac MD (1985) Torsional accelerograms. Int J Soil Dyn Earthq Eng 4(3):132–139
70. Lee VW, Trifunac MD (1987) Rocking strong earthquake accelerations. Int J Soil Dyn Earthq Eng 6(2):75–89
71. Lee WHK (2002) Challenges in observational seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part A. Academic Press, Amsterdam, 269–281
72. Lee WHK, Benson RB (2008) Making non-digitally-recorded seismograms accessible online for studying earthquakes. In: Fréchet J, Meghraoui M, Stucchi M (eds) Modern Approach in Historical Seismology: Interdisciplinary studies of past and recent earthquakes. Springer, Berlin, pp 403–427
73. Lee WHK, Espinosa-Aranda JM (2003) Earthquake early warning systems: Current status and perspectives. In: Zschau J, Kuppers AN (eds) Early Warning Systems for Natural Disaster, Reduction. Springer, Berlin, pp 409–423
74. Lee WHK, Stewart SW (1981) Principles and Applications of Microearthquake Networks. Academic Press, New York
75. Lee WHK, Celebi M, Todorovska MI, Diggles MF (2007) Rotational seismology and engineering applications: Proceedings for the First International Workshop, Menlo Park, California, USA, 18–19 September. US Geol Surv Open File Rep 2007–1144. http://pubs.usgs.gov/of/2007/1144/
76. Lee WHK, Igel H, Todorovska MI, Evans JR (2007) Rotational Seismology: AGU Session, Working Group, and Website. US Geol Surv Open File Rep 2007–1263. http://pubs.usgs.gov/of/2007/1263/
77. Lognonne P, Clevede E (2002) Normal modes of the Earth and planets. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part A. Academic Press, Amsterdam, pp 125–147
78. Luco JE (1976) Torsional response of structures to obliquely incident seismic SH waves. Earthq Eng Struct Dyn 4:207–219
79. Mallet R (1858) Fourth report on the facts of earthquake phenomena. Ann Rep Brit Assn Adv Sci 28:1–136
80. Mallet R (1862) Great Neapolitan Earthquake of 1857, vol I, II. Chapman and Hall, London
81. Maruyama T (1963) On the force equivalents of dynamical elastic dislocations with reference to the earthquake mechanism. Bull Earthq Res Inst 41:467–486
82. Michelini A, De Simoni B, Amato A, Boschi E (2005) Collecting, digitizing and distributing historical seismological data. EOS 86(28) 12 July 2005

83. Nakamura Y (1984) Development of the earthquake early-warning system for the Shinkansen, some recent earthquake engineering research and practical in Japan. The Japanese National Committee of the International Association for Earthquake Engineering, pp 224–238
84. Nakamura Y (1988) On the urgent earthquake detection and alarm system, UrEDAS. Proc Ninth World Conf Earthq Eng 7:673–678
85. Nakamura Y (2004) On a rational strong motion index compared with other various indices. 13th World Conf Earthq Eng, Paper No 910
86. Nakamura Y, Saita J (2007) UrEDAS, The Earthquake Warning System: today and tomorrow. In: Gasparini P, Manfredi G, Zschau J (eds) Earthquake Early Warning Systems. Springer, Berlin, pp 249–281
87. Newmark NM (1969) Torsion in symmetrical buildings. Proc. Fourth World Conference on Earthquake Eng, vol II. pp A3/19–A3/32
88. Niazi M (1987) Inferred displacements, velocities and rotations of a long rigid foundation located at El-Centro differential array site during the 1979 Imperial Valley, California, earthquake. Earthq Eng Struct Dyn 14:531–542
89. Odaka T, Ashiya K, Tsukada S, Sato S, Ohtake K, Nozaka D (2003) A new method of quickly estimating epicentral distance and magnitude from a single seismic record. Bull Seism Soc Am 93:526–532
90. Oliveira CS, Bolt BA (1989) Rotational components of surface strong ground motion. Earthq Eng Struct Dyn 18:517–526
91. Oliver J, Murphy L (1971) WWNSS: Seismology's global network of observing stations. Science 174:254–261
92. Peterson J (1993) Observations and modeling of seismic background noise. US Geol Surv Open File Rep, 93–322
93. Pillet R, Virieux J (2007) The effects of seismic rotations on inertial sensors. Geophys J Int. doi:10.1111/j.1365-246X.2007.03617.x
94. Reid HF (1910) The California Earthquake of 18 April 1906, vol 2. The Mechanics of the Earthquake. Carnegie Inst, Washington DC
95. Richards PG (2002) Seismological methods of monitoring compliance with the Comprehensive Nuclear Test-Ban Treaty. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part A. Academic Press, Amsterdam, pp 369–382
96. Richter CF (1935) An instrumental earthquake magnitude scale. Bull Seis Soc Am 25:1–32
97. Richter CF (1958) Elementary Seismology. Freeman, San Francisco
98. Rutenberg A, Heidebrecht AC (1985) Rotational ground motion and seismic codes. Can J Civ Eng 12(3):583–592
99. Saita J, Nakamura Y (2003) UrEDAS: the early warning system for mitigation of disasters caused by earthquakes and tsunamis. In: Zschau J, Kuppers AN (eds) Early Warning Systems for Natural Disaster, Reduction. Springer, Berlin, pp 453–460
100. Savage JC (1978) Dislocation in seismology. In: Nabarro FRN (ed) Dislocation in Solids. North-Holland, Amsterdam, 251–339
101. Schweitzer J, Lee WHK (2003) Old seismic bulletins: a collective heritage from early seismologists. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part B. Academic Press, Amsterdam, pp 1665–1717 (with CD-ROM)
102. Shin TC, Tsai YB, Yeh YT, Liu CC, Wu YM (2003) Strong-motion instrumentation programs in Taiwan. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part B. Academic Press, Amsterdam, pp 1057–1062
103. Sokolnikoff IS (1956) Mathematical Theory of Elasticity, 2nd edn. McGraw-Hill, New York
104. Spudich P, Steck LK, Hellweg M, Fletcher JB, Baker LM (1995) Transient stresses at Park-field, California, produced by the m 7.4 Landers earthquake of 28 June 1992: Observations from the UPSAR dense seismograph array. J Geophys Res 100:675–690
105. Suryanto W, Igel H, Wassermann J, Cochard A, Schubert B, Vollmer D, Scherbaum F (2006) Comparison of seismic array-derived rotational motions with direct ring laser measurements. Bull Seism Soc Am 96(6):2059–2071
106. Takeo M (1998) Ground rotational motions recorded in near-source region. Geophy Res Lett 25(6):789–792
107. Takeo M, Ito HM (1997) What can be learned from rotational motions excited by earthquakes? Geophys J Int 129:319–329
108. Teisseyre R, Suchcicki J, Teisseyre KP, Wiszniowski J, Palangio P (2003) Seismic rotation waves: basic elements of theory and recording. Annali Geofis 46:671–685
109. Teisseyre R, Takeo M, Majewski E (eds) (2006) Earthquake Source Asymmetry, Structural Media and Rotation Effects. Springer, Berlin
110. Teng TL, Wu L, Shin TC, Tsai YB, Lee WHK (1997) One minute after: strong motion map, effective epicenter, and effective magnitude. Bull Seism Soc Am 87:1209–1219
111. Trifunac MD (1982) A note on rotational components of earthquake motions on ground surface for incident body waves. Soil Dyn Earthq Eng 1:11–19
112. Trifunac MD (2006) Effects of torsional and rocking excitations on the response of structures. In: Teisseyre R, Takeo M, Majewski M (eds) Earthquake Source Asymmetry, Structural Media and Rotation Effects. Springer, Heidelberg, pp 569–582
113. Trifunac MD, Todorovska MI (2001) A note on the useable dynamic range of accelerographs recording translation. Soil Dyn Earthq Eng 21(4):275–286
114. Twiss R, Souter B, Unruh J (1993) The effect of block rotations on the global seismic moment tensor and the patterns of seismic P and T axes. J Geophys Res 98(B1):645–674
115. Utsu T (2002) A list of deadly earthquakes in the world (1500–2000). In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part A. Academic Press, Amsterdam, pp 691–717
116. Utsu T (2002) Relationships between magnitude scales. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part A. Academic Press, Amsterdam, pp 733–746
117. Uyeda S (2002) Continental drift, sea-floor spreading, and plate/plume tectonics. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part A. Academic Press, Amsterdam, pp 51–67
118. Villasenor A, Engdahl ER (2007) Systematic relocation of early instrumental seismicity: Earthquakes in the International Seismological Summary for 1960–1963. Bull Seism Soc Am 97:1820–1832

119. Wald DJ, Quitoriano V, Heaton TH, Kanamori H (1999) Relationships between peak ground acceleration, peak ground velocity, and modified Mercalli intensity in California. Earthq Spectr 15:557–564
120. Wald DJ, Quitoriano V, Heaton TH, Kanamori H, Scrivner CW, Worden CB (1999) TriNet "ShakeMaps": Rapid generation of peak ground motion and intensity maps for earthquakes in Southern California. Earthq Spectr 15:537–555
121. Webb SC (2002) Seismic noise on land and on the seafloor. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part A. Academic Press, Amsterdam, pp 305–318
122. Wielandt E (2002) Seismometry. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part A. Academic Press, Amsterdam, pp 283–304
123. Willemann RJ, Storchak DA (2001) Data collection at the International Seismological Centre. Seism Res Lett 72:440–453
124. Wu YM, Kanamori H (2005) Experiment on an onsite early warning method for the Taiwan early warning system. Bull Seism Soc Am 95:347–353
125. Wu YM, Kanamori H (2005) Rapid assessment of damaging potential of earthquakes in Taiwan from the beginning of P Waves. Bull Seism Soc Am 95:1181–1185
126. Wu YM, Kanamori H (2008) Exploring the feasibility of on-site earthquake early warning using close-in records of the 2007 Noto Hanto earthquake. Earth Planets Space 60:155–160
127. Wu YM, Zhao L (2006) Magnitude estimation using the first three seconds P-wave amplitude in earthquake early warning. Geophys Res Lett 33:L16312. doi:10.1029/2006GL026871
128. Wu YM, Chen CC, Shin TC, Tsai YB, Lee WHK,Teng TL (1997) Taiwan Rapid Earthquake Information Release System. Seism Res Lett 68:931–943
129. Wu YM, Hsiao NC, Lee WHK, Teng TL, Shin TC (2007) State of the art and progresses of early warning system in Taiwan. In: Gasparini P, Manfredi G, Zschau J (eds) Earthquake Early Warning Systems. Springer, Berlin, pp 283–306
130. Wu YM, Kanamori H, Allen R, Hauksson E (2007) Determination of earthquake early warning parameters, τ_c and P_d, for southern California. Geophys J Int 169:667–674
131. Wu YM, Lee WHK, Chen CC, Shin TC, Teng TL, Tsai YB (2000) Performance of the Taiwan Rapid Earthquake Information Release System (RTD) during the 1999 Chi-Chi (Taiwan) earthquake. Seism Res Lett 71:338–343
132. Zollo A, Lancieri M, Nielsen S (2006) Earthquake magnitude estimation from peak amplitudes of very early seismic signals on strong motion records. Geophys Res Lett 33:L23312. doi:10.1029/2006GL027795

Books and Reviews

Abercrombie R, McGarr A, Kanamori H, Di Toro G (2006) Earthquakes: Radiated Energy and the Physics of Faulting. Geophysical Monograph, vol 170. American Geophysical Union, Washington DC
Bolt BA (1993) Earthquakes. W.H. Freeman, New York
Chen YT, Panza GF, Wu ZL (2004) Earthquake Hazard, Risk, and Strong Ground Motion. Seismological Press, Beijing
Kanamori H (ed) (2007) Earthquake Seismology, Treatise on Geophysics, vol 4. Elsevier, Amsterdam
Kanamori H, Boschi E (1983) Earthquakes: Observation, Theory and Interpretation. North-Holland, Amsterdam
Keilis-Borok VI, Soloviev AA (eds) (2003) Nonlinear Dynamics of the Lithosphere and Earthquake Prediction. Springer, Berlin
Lee WHK, Meyers H, Shimazaki K (eds) (1988) Historical Seismograms and Earthquakes of the World. Academic Press, San Diego
Lee WHK, Kanamori H, Jennings JC, Kisslinger C (eds) (2002) International Handbook of Earthquake and Engineering Seismology, Part A. Academic Press, San Diego, pp 933 (and 1 CD-ROM)
Lee WHK, Kanamori H, Jennings JC, Kisslinger C (eds) (2003) International Handbook of Earthquake and Engineering Seismology, Part B. Academic Press, San Diego, pp 1009 (and 2 CD-ROMs)
Pujol J (2003) Elastic Wave Propagation and Generation in Seismology. Cambridge Univ Press, Cambridge
Zschau J, Kuppers AN (eds) (2003) Early Warning Systems for Natural Disaster Reduction. Springer, Berlin

Earthquake Networks, Complex

Sumiyoshi Abe[1,2], Norikazu Suzuki[3]
[1] Department of Physical Engineering, Mie University, Tsu, Japan
[2] Institut Supérieur des Matériaux et Mécaniques, Le Mans, France
[3] College of Science and Technology, Nihon University, Chiba, Japan

Article Outline

Glossary
Definition of the Subject
Introduction
Construction of an Earthquake Network
Scale-free Nature of Earthquake Network
Small-World Nature of Earthquake Network
Hierarchical Structure
Mixing Property
Period Distribution
Future Directions
Addendum
Bibliography

Glossary

Network or graph A network (or a graph) [28] consists of vertices (or nodes) and edges (or links) connecting them. In general, a network contains loops (i. e., edges with both ends attached to the same vertices) and multiple edges (i. e., edges more than one that connect two different vertices). If edges have their directions,

such a network is called directed. A simple graph is a network, in which loops are removed and each multiple edge is replaced by a single edge. In a stochastic network, each connection is inherently probabilistic. A classical random graph is a simple example, in which each two vertices are connected by an edge with probability p and unconnected with probability $1-p$ $(0 < p < 1)$.

Connectivity distribution or degree distribution The connectivity distribution (or the degree distribution), $P(k)$, is the probability of finding vertices with k edges in a stochastic network. In a directed network, the number of incoming/outgoing edges is called the in-degree/out-degree. Connectivity of a classical random graph obeys the Poissonian distribution in the limit of the large number of vertice [11,14,20], $P(k) = \mathrm{e}^{-\lambda}\lambda^k/k!$ (λ: a positive parameter, $k = 0, 1, 2, \ldots$), whereas a scale-free network [11,12,14,20] has a power-law shape, $P(k) \sim k^{-\gamma}$ (γ: a positive exponent), for large k.

Preferential attachment rule This is a concept relevant to a growing network, in which the number of vertices increases. Preferential attachment [11,12,14,20] implies that a newly created vertex tends to link to pre-existing vertices with the probability $\Pi(k_i) = k_i/\sum_j k_j$, where k_i stands for the connectivity of the ith vertex. That is, the larger the connectivity of a vertex is, the higher the probability of getting linked to a new vertex is.

Clustering coefficient The clustering coefficient [27] is a quantity characterizing an undirected simple graph. It quantifies the adjacency of two neighboring vertices of a given vertex, i. e., the tendency of two neighboring vertices of a given vertex to be connected to each other. Mathematically, it is defined as follows. Assume the ith vertex to have k_i neighboring vertices. There can exist at most $k_i(k_i-1)/2$ edges between the neighbors. Define c_i as the ratio

$$c_i = \frac{\text{actual number of edges between the neighbors of the } i\text{th vertex}}{k_i(k_i-1)/2} . \quad (1)$$

Then, the clustering coefficient is given by the average of this quantity over the network:

$$C = \frac{1}{N}\sum_{i=1}^{N} c_i , \quad (2)$$

where N is the total number of vertices contained in the network. The value of the clustering coefficient of a random graph, C_{random}, is much smaller than unity, whereas a small-world network has a large value of C which is much larger than C_{random}.

Hierarchical organization Many complex networks are structurally modular, that is, they are composed of groups of vertices that are highly interconnected to each other but weakly connected to outside groups. This hierarchical structure [22] can conveniently be characterized by the clustering coefficient at each value of connectivity, $c(k)$, which is defined by

$$c(k) = \frac{1}{NP_{\text{SG}}(k)}\sum_{i=1}^{N} c_i \delta_{k_i,k} , \quad (3)$$

where c_i is given by (1), N the total number of vertices, and $P_{\text{SG}}(k)$ the connectivity distribution of an undirected simple graph. Its average is the clustering coefficient in (2): $C = \sum_k c(k)P_{\text{SG}}(k)$. A network is said to be hierarchically organized if $c(k)$ varies with respect to k, typically due to a power law, $c(k) \sim k^{-\beta}$, with a positive exponent β.

Assortative mixing and disassortative mixing Consider the conditional probability, $P(k'|k)$, of finding a vertex with connectivity k' linked to a given vertex with connectivity k. Then, the nearest-neighbor average connectivity of vertices with connectivity k is defined by [20,21,26]

$$\bar{k}_{nn}(k) = \sum_{k'} k' P(k'|k) . \quad (4)$$

If $\bar{k}_{nn}(k)$ increases/decreases with respect to k, mixing is termed assortative/disassortative. A simple model of growth with preferential attachment is known to possess no mixing. That is, $\bar{k}_{nn}(k)$ does not depend on k. The above-mentioned linking tendency can be quantified by the correlation coefficient [17] defined as follows. Let $e_{kl} (= e_{lk})$ be the joint probability distribution for an edge to link with a vertex with connectivity k at one end and a vertex with connectivity l at the other. Calculate its marginal, $q_k = \sum_l e_{kl}$. Then, the correlation coefficient is given by

$$r = \frac{1}{\sigma_q^2}\sum_{k,l} kl\left(e_{kl} - q_k q_l\right) , \quad (5)$$

where $\sigma_q^2 = \sum_k k^2 q_k - (\sum_k kq_k)^2$ stands for the variance of q_k. $r \in [-1, 1]$, and if r is positive/negative, mixing is assortative/disassortative [17,20].

Definition of the Subject

Complexity is an emergent collective property, which is hardly understood by the traditional approach in natural

science based on reductionism. Correlation between elements in a complex system is strong, no matter how largely they are separated both spatially and temporally, therefore it is essential to treat such a system in a holistic manner, in general.

Although it is generally assumed that seismicity is an example of complex phenomena, it is actually nontrivial to see how and in what sense it is complex. This point may also be related to the question of primary importance of why it is so difficult to predict earthquakes.

Development of the theory of complex networks turns out to offer a peculiar perspective on this point. Construction of a complex earthquake network proposed here consists of mapping seismic data to a growing stochastic graph. This graph, or network, turns out to exhibit a number of remarkable behaviors both physically and mathematically, which are in common with many other complex systems. The scale-free and small-world natures are typical examples. In this way, one will be able to obtain a novel viewpoint of seismicity.

Introduction

Seismicity is a field-theoretical phenomenon. Released energy of each earthquake may be regarded as a field amplitude defined at a discrete spacetime point. However, in contrast to a familiar field theory such as the electromagnetic theory, both amplitudes and locations are intrinsically probabilistic. The fault distribution may geometrically be fractal [18], and the stress distribution superposed upon it often has a complex landscape. Accordingly, seismicity is characterized by extremely rich phenomenology, which attracts physicists' attention from the viewpoint of science of complex systems.

There are at least two celebrated empirical laws known in seismology. One is the Gutenberg–Richter law [16], which states that the frequency of earthquakes obeys a power law with respect to released energy. This power-law nature makes it difficult or even meaningless to statistically distinguish earthquakes by their values of magnitude because of the absence of typical energy scales. The other is the Omori law [19], which states that the rate of the frequency of aftershocks following a main shock algebraically decays with respect to time elapsed from the main shock. This slow relaxation reminds one of complex glassy dynamics [13]. Such a viewpoint is supported by the discovery of the aging phenomenon and the scaling law for aftershocks [2].

Another point, which seems less noticed, is that correlation of two successive events is strong, no matter how large their spatial separation is. There is, in fact, an observation [23] that an earthquake can be triggered by a foregoing one more than 1000 km away. The reason why two successive events are indivisibly related can also be found in another observation [3,6] that both spatial distance and time interval between two successive events obey the q-exponential distributions in nonextensive statistics [1,15,24], which offers a statistical-mechanical framework for describing complex systems. Thus, the correlation length can be enormously large and long-wave-length modes of seismic waves play an important role. This has a strong similarity to phase transitions and critical phenomena. Accordingly, it may not be appropriate to use spatial windows in analysis of seismicity. Furthermore, all of the data in a relevant area (ideally the whole globe, though still not satisfactorily available) should be treated based on the nonreductionistic standpoint.

The network approach is a powerful tool for analyzing kinematical and dynamical structures of complex systems in a holistic manner. Such a concept was introduced to seismology by the present authors in 2004 [4] in order to represent complexity of seismicity. The procedure described in Sect. "Construction of an Earthquake Network" allows one to map a seismic time series to a growing stochastic network in an unambiguous way. Vertices and edges of such a network correspond to coarse-grained events and event-event correlations, respectively. Yet unknown microscopic dynamics governing event-event correlations and fault-fault interactions are replaced by these edges. Global physical properties of seismicity can then be explored by examining its geometric (e. g., topological etc.) and dynamical features. It turns out that earthquake networks have a number of intriguing properties, some of which are shared by many other natural as well as artificial systems including metabolic networks, food webs, the Internet, the world-wide web, and so on [11,14,20]. This, in turn, enables seismologists to study seismicity in analogy with such relatively better understood complex systems. Thus, the network approach offers a novel way of analyzing seismic time series and casts fresh light on the physics of earthquakes.

In this article, only the data taken from California is utilized. However, it has been ascertained that the laws and trends discussed here are universal and hold also in other geographical regions including Japan.

Construction of an Earthquake Network

An earthquake network is constructed as follows [4]. A geographical region under consideration is divided into small cubic cells. A cell is regarded as a vertex if earthquakes with any values of magnitude above a certain de-

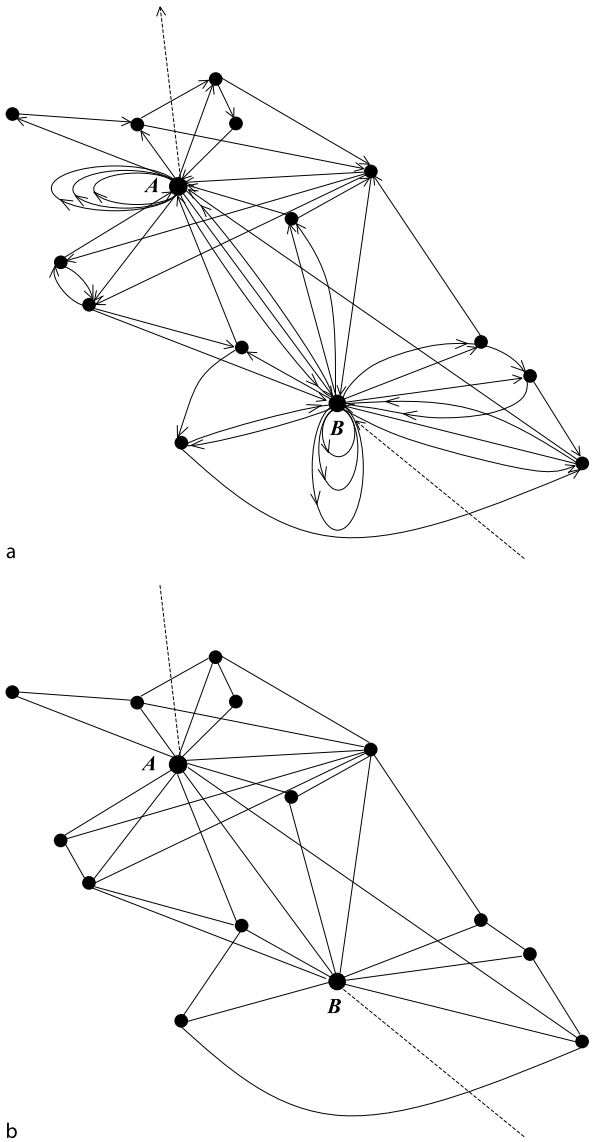

Earthquake Networks, Complex, Figure 1
a A schematic description of earthquake network. The *dashed lines* correspond to the initial and final events. The vertices, *A* and *B*, contain main shocks and play roles of hubs of the network. b The undirected simple graph reduced from the network in a.

tection threshold occurred therein. Two successive events define an edge between two vertices. If they occur in the same cell, a loop is attached to that vertex. This procedure enables one to map a given interval of the seismic data to a growing probabilistic graph, which is referred to as an earthquake network (see Fig. 1a).

Several comments are in order. Firstly, this construction contains a single parameter: cell size, which is a scale of coarse graining. Once cell size is fixed, an earthquake network is unambiguously defined. However, since there exist no a priori operational rule to determine cell size, it is important to notice how the properties of an earthquake network depend on this parameter. Secondly, as mentioned in Sect. "Introduction", edges and loops efficiently represent event-event correlation. Thirdly, an earthquake network is a directed graph in its nature. Directedness does not bring any difficulties to statistical analysis of connectivity (degree, i. e., the number of edges attached to the vertex under consideration) since, by construction, the in-degree and out-degree are identical for each vertex except the initial and final vertices in analysis. Therefore, the in-degree and out-degree are not distinguished from each other in the analysis of the connectivity distribution (see Sections "Scale-free Nature of Earthquake Network" and "Mixing Property"). However, directedness becomes essential when the path length (i. e., the number of edges) between a pair of connected vertices, i. e., the degree of separation between the pair, is considered. This point is explicitly discussed in the analysis of the period distribution in Sect. "Period Distribution". Finally, a full directed earthquake network has to be reduced to a simple undirected graph, when its small-worldness and hierarchical structure are examined (see Sections "Small-World Nature of Earthquake Network" and "Hierarchical Structure"). There, loops are removed and each multiple edge is replaced by a single edge (see Fig. 1b). The path length in this case is the smallest value among the possible numbers of edges connecting a pair of vertices.

Scale-free Nature of Earthquake Network

An earthquake network contains some special vertices which have large values of connectivity. Such "hubs" turn out to correspond to cells with main shocks. This is due to a striking fact discovered from real data analysis that aftershocks associated with a main shock tend to return to the locus of the main shock, geographically. This is the primary reason why a vertex containing a main shock becomes a hub. The situation is analogous to the preferential attachment rule for a growing network [11,14,20]. According to this rule, a newly created vertex tends to be connected to the (already existing) ith vertex with connectivity k_i with probability, $\Pi(k_i) = k_i / \sum_j k_j$. It can generate a scale-free network characterized by the power-law connectivity distribution [11,12]:

$$P(k) \sim k^{-\gamma} , \tag{6}$$

where γ is a positive exponent.

In Fig. 2, the connectivity distribution of the full earthquake network with loops and multiple edges is pre-

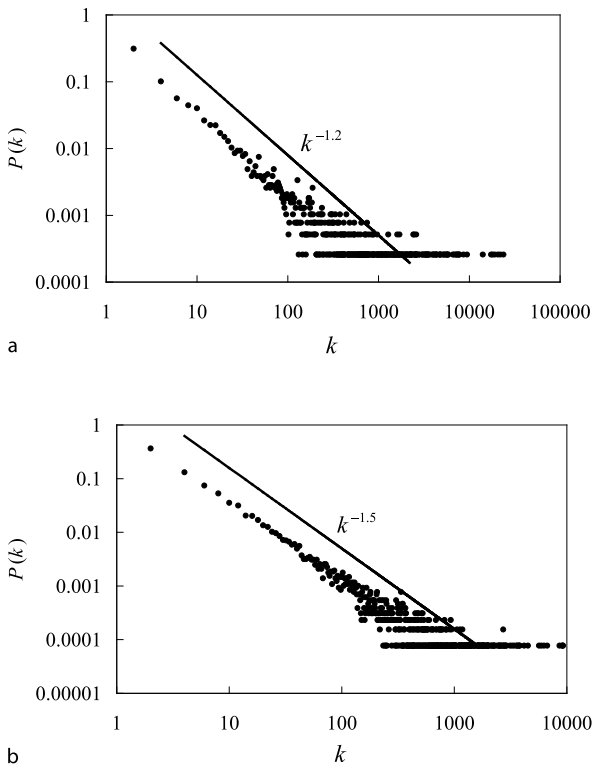

Earthquake Networks, Complex, Figure 2

The log-log plots of the connectivity distributions of the earthquake network constructed from the seismic data taken in California [the Southern California Earthquake Data Center (http://www.data.scec.org/)]. The time interval analyzed is between 00:25:8.58 on January 1, 1984 and 22:21:52.09 on December 31, 2003. The region covered is 29°06.00′*N*–38°59.76′*N* latitude and 113°06.00′*W*–122°55.59′*W* longitude with the maximal depth 175.99 km. The total number of events is 367 613. The data contains no threshold for magnitude (but "quarry blasts" are excluded from the analysis). Two different values of cell size are examined: a 10 km × 10 km × 10 km and b 5 km × 5 km × 5 km. All quantities are dimensionless.

sented [4]. From it, one appreciates that the earthquake network in fact possesses connectivity of the form in (6) and is therefore scale-free. The smaller the cell size is, the larger the exponent, γ, is, since the number of vertices with large values of connectivity decreases as cell size becomes smaller. The scale-free nature may be interpreted as follows. As mentioned above, aftershocks associated with a main shock tend to be connected to the vertex of the main shock, satisfying the preferential attachment rule. On the other hand, the Gutenberg–Richter law states that frequency of earthquakes decays slowly as a power law with respect to released energy. This implies that there appear to be quite a few giant components, and accordingly the network becomes highly inhomogeneous.

Earthquake Networks, Complex, Table 1

The small-world properties of the undirected simple earthquake network. The values of the number of vertices, *N*, the clustering coefficient, *C*, (compared with those of the classical random graphs, C_{random}) and the average path length, *L* are presented. The data employed is the same as that in Fig. 2.

Cell size	10 km × 10 km × 10 km	5 km × 5 km × 5 km
Number of vertices	$N = 3869$	$N = 12913$
Clustering coefficient	$C = 0.630$ ($C_{random} = 0.014$)	$C = 0.317$ ($C_{random} = 0.003$)
Average path length	$L = 2.526$	$L = 2.905$

Small-World Nature of Earthquake Network

The small-world nature is an important aspect of complex networks. It shows how a complex network is different from both regular and classical random graphs [27]. A small-world network resides in-between regularity and randomness, analogous to the edge of chaos in nonlinear dynamics.

To study the small-world nature of an earthquake network, a full network has to be reduced to a simple undirected graph: that is, loops are removed and each multiple edge is replaced by a single edge (see Fig. 1b). This is because in the small-world picture one is concerned only with simple linking pattern of vertices.

A small-world network is characterized by a large value of the clustering coefficient in (2) and a small value of the average path length [27]. The clustering coefficient quantifies the tendency of two neighboring vertices of a given vertex to be connected to each other. A small-world network has a large value of the clustering coefficient, whereas the value for the classical random graph is very small [11,14,20,27]: $C_{random} = \langle k \rangle / N \ll 1$, where N and $\langle k \rangle$ are the total number of vertices and the average value of connectivity, respectively.

In Table 1, the results are presented for the clustering coefficient and the average path length [5,9]. One finds that the values of the clustering coefficient are in fact much larger than those of the classical random graphs and the average path length is short. Thus, the earthquake networksportant features of small-world network.

Hierarchical Structure

As seen above, seismicity generates a scale-freeexity of an earthquake network further, one may examine if it is hierarchically organized [8]. The hierarchical structure can be revealed by analyzing the clustering coefficient as a function of connectivity. The connectivity-dependent cluster-

ing coefficient, $c(k)$, is defined in (3). This quantifies the adjacency of two vertices connected to a vertex with connectivity, k, and gives information on hierarchical organization of a network.

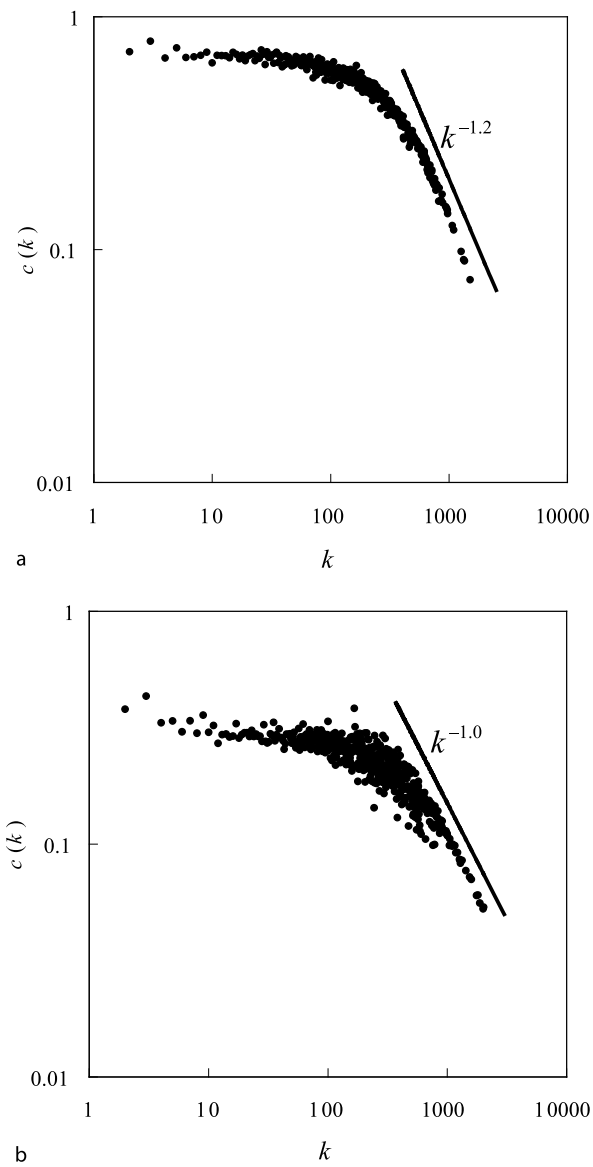

Earthquake Networks, Complex, Figure 3
The log-log plots of the connectivity-dependent clustering coefficient for two different values of cell size: a 10 km × 10 km × 10 km and b 5 km × 5 km × 5 km. The analyzed period is between 00:25:8.58 on January 1, 1984 and 22:50:49.29 on December 31, 2004, which is taken from the same catalog as in Fig. 2. The region covered is 28°36.00′*N*–38°59.76′*N* latitude and 112°42.00′*W*–123°37.41′*W* longitude with the maximal depth 175.99 km. The total number of the events is 379728. All quantities are dimensionless.

In Fig. 3, the plots of $c(k)$ are presented [8]. As can be clearly seen, the clustering coefficient of the undirected simple earthquake network asymptotically follows the scaling law

$$c(k) \sim k^{-\beta} \tag{7}$$

with a positive exponent β. This highlights hierarchical organization of the earthquake network.

Existence of the hierarchical structure is of physical importance. The earthquake networkment [11,12,14,20]. However, the standard preferential-attachment-model is known to fail at generating hierarchical organization [22]. To mediate between growth with preferential attachment and the presence of hierarchical organization, the concept of vertex deactivation has been introduced in the literature [25]. According to this concept, in the process of network growth, some vertices deactivate and cannot acquire new edges any more. This has a natural physical implication for an earthquake network: active faults may be deactivated through the process of stress release. In addition, the fitness model [26] is also known to generate hierarchical organization. This model generalizes the preferential attachment rule in such a way that not only connectivity but also "charm" of vertices (i. e., attracting a lot of edges) are taken into account. Seismologically, fitness is considered to describe intrinsic properties of faults such as geometric configuration and stiffness. Both of these two mechanisms can explain a possible origin of the complex hierarchical structure, by which relatively new vertices have chances to become hubs of the network. In the case of an earthquake network, it seems plausible to suppose that the hierarchical structure may be due to both deactivation and fitness.

A point of particular interest is that the hierarchical structure disappears if weak earthquakes are removed. For example, setting a lower threshold for earthquake magnitude, say $M_{\mathrm{th}} = 3$, makes it difficult to observe the power-law decay of the clustering coefficient hierarchical structure of an earthquake network is largely supported by weak shocks.

Mixing Property

The scale-free nature, small-worldness, growth with preferential attachment, and hierarchical organization all indicate that earthquake networks are very similar to other known networks, for example, the Internet. However, there is at least one point which shows an essential difference between the two. It is concerned with the mixing property, which is relevant to the concept of the nearest-

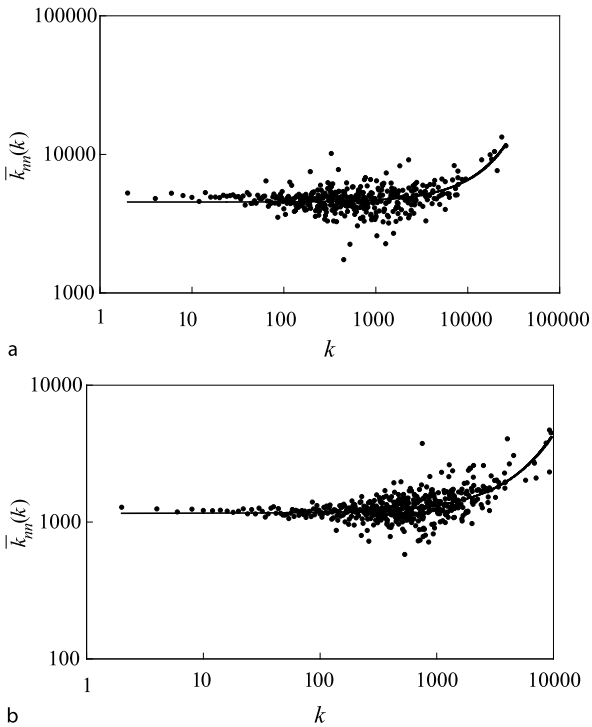

Earthquake Networks, Complex, Figure 4
The log-log plots of the nearest-neighbor average connectivity for two different values of cell size: a 10 km × 10 km × 10 km and b 5 km × 5 km × 5 km. The data employed is the same as that in Fig. 3. The solid lines show the trends depicted by the exponentially increasing functions. All quantities are dimensionless.

Earthquake Networks, Complex, Table 2
The values of the dimensionless correlation coefficient. The data employed is the same as that in Fig. 3. Positivity of the values implies that mixing is assortative.

10 km × 10 km × 10 km	5 km × 5 km × 5 km
$r = 0.285$	$r = 0.268$

neighbor average connectivity $\bar{k}_{nn}(k)$ [in (4)], of a full network with loops and multiple edges.

The plots of this quantity are presented in Fig. 4. There, the feature of assortative mixing [8] is observed, since $\bar{k}_{nn}(k)$ increases with respect to connectivity k. Therefore, vertices with large values of connectivity tend to be linked to each other. That is, vertices containing stronger shocks tend to be connected among themselves with higher probabilities.

To quantify this property, the correlation coefficient in (5) is evaluated [8]. The result is summarized in Table 2. The value of the correlation coefficient is in fact positive, confirming that the earthquake network has assortative mixing. On the other hand, the Internet is of disassortative mixing [17,20,21,26]. That is, the mixing properties of the earthquake network and the Internet are opposite to each other. It is noticed however that the loops and multiple edges play essential roles for the assortative mixing: an undirected simple graph obtained by reducing a full earthquake network turns out to exhibit disassortative mixing. These are purely the phenomenological results, and their physical origins still have yet to be clarified.

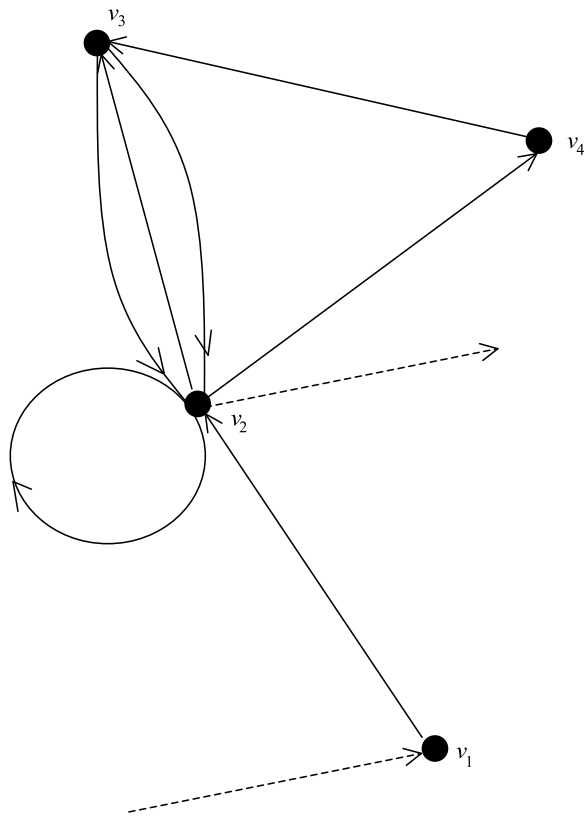

Earthquake Networks, Complex, Figure 5
A full directed network: $\cdots \to v_1 \to v_2 \to v_3 \to v_2 \to v_2 \to v_4 \to v_3 \to v_2 \to \cdots$. The period associated with v_3 is 4, whereas v_2 has 1, 2 and 3.

Period Distribution

So far, directedness of an earthquake network has been ignored. The full directed network picture is radically different from the small-world picture for a simple undirected graph. It enables one to consider interesting dynamical features of an earthquake network. As an example, here the concept of period [7] is discussed. This is relevant to the question "after how many earthquakes does an event return to the initial cell, statistically?" It is therefore of obvious interest for earthquake prediction.

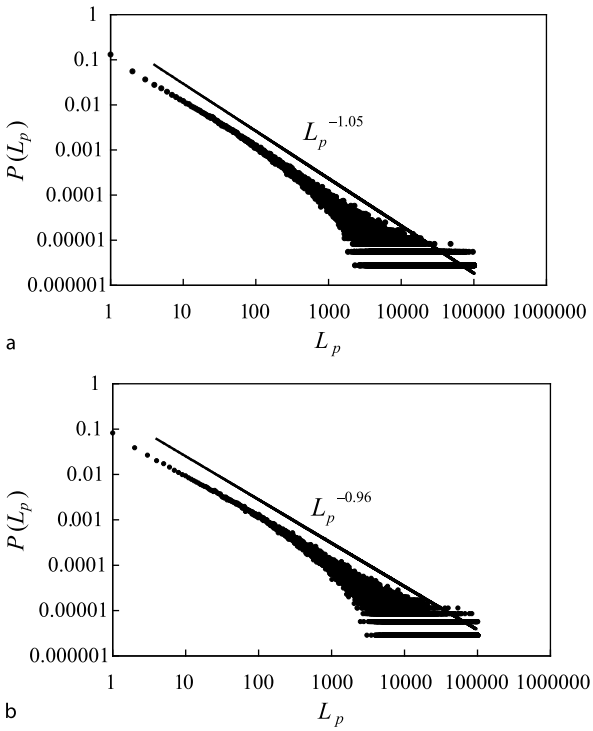

Earthquake Networks, Complex, Figure 6
The log-log plots of the period distribution for two different values of cell size: a 10 km × 10 km × 10 km and b 5 km × 5 km × 5 km. The data employed is the same as that in Fig. 2. All quantities are dimensionless.

Period in a directed network is defined as follow. Given a vertex of a network, there are various closed routes starting from and ending at this vertex. The period, L_p, of a chosen closed route is simply the number of edges forming the route (see Fig. 5).

The period distribution, $P(L_\mathrm{p})$, is defined as the number of closed routes. The result is presented in Fig. 6 [7]. As can be seen there, $P(L_\mathrm{p})$ obeys a power law

$$P(L_\mathrm{p}) \sim (L_\mathrm{p})^{-\alpha} , \tag{8}$$

where α is a positive exponent. This implies that there exist a number of closed routes with significantly long periods in the network. This fact makes it highly nontrivial to statistically estimate the value of period.

Future Directions

In the above, the long-time statistical properties of an earthquake network have mainly been considered. On the other hand, given the cell size, an earthquake network represents all the dynamical information contained in a seismic time series, and therefore the study of its time evolution may give a new insight into seismicity. This, in turn, implies that it may offer a novel way of monitoring seismicity.

For example, it is of interest to investigate how the clustering coefficient changes in time as earthquake network dynamically evolves. According to the work in [10], the clustering coefficient remains stationary before a main shock, suddenly jumps up at the main shock, and then slowly decays to become stationary again following the power-law relaxation. In this way, the clustering coefficient successfully characterizes aftershocks in association with main shocks.

A question of extreme importance is if precursors of a main shock can be detected through monitoring dynamical evolution of earthquake network. Clearly, further developments are needed in science of complex networks to address to this question.

Addendum

Some authors (e. g., Sornette and Werner) raised questions about the applicability of complex earthquake network using online accessible earthquake catalog data and on the validity of our results. A preprint of an article by the authors discussing these questions can be found in the e-print available at http://arxiv.org/abs/0708.2203.

Bibliography

1. Abe S, Okamoto Y (eds) (2001) Nonextensive statistical mechanics and its applications. Springer, Heidelberg
2. Abe S, Suzuki N (2003) Aging and scaling of earthquake aftershocks. Physica A 332:533–538
3. Abe S, Suzuki N (2003) Law for the distance between successive earthquakes. J Geophys Res 108(B2):2113 ESE 19-1-4
4. Abe S, Suzuki N (2004) Scale-free network of earthquakes. Europhys Lett 65:581–586
5. Abe S, Suzuki N (2004) Small-world structure of earthquake network. Physica A 337:357–362
6. Abe S, Suzuki N (2005) Scale-free statistics of time interval between successive earthquakes. Physica A 350:588–596
7. Abe S, Suzuki N (2005) Scale-invariant statistics of period in directed earthquake network. Eur Phys J B 44:115–117
8. Abe S, Suzuki N (2006) Complex earthquake networks: Hierarchical organization and assortative mixing. Phys Rev E 74:026113-1-5
9. Abe S, Suzuki N (2006) Complex-network description of seismicity. Nonlin Processes Geophys 13:145–150
10. Abe S, Suzuki N (2007) Dynamical evolution of clustering in complex network of earthquakes. Eur Phys J B 59:93–97
11. Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
12. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512
13. Debenedetti PG, Stillinger FH (2001) Supercooled liquids and the glass transition. Nature 410:259–267

14. Dorogovtsev SN, Mendes JFF (2003) Evolution of networks: from biological nets to the Internet and WWW. Oxford University Press, Oxford
15. Gell-Mann M, Tsallis C (eds) (2004) Nonextensive Entropy: Interdisciplinary Applications. Oxford University Press, Oxford
16. Gutenberg B, Richter CF (1944) Frequency of earthquakes in California. Bull Seismol Soc Am 34:185–188
17. Newman MEJ (2002) Assortative mixing in networks. Phys Rev Lett 89:208701-1-4
18. Okubo PG, Aki K (1987) Fractal geometry in the San Andreas fault system. J Geophys Res 92(B1):345–355
19. Omori F (1894) On the after-shocks of earthquakes. J Coll Sci Imper Univ Tokyo 7:111–200
20. Pastor-Satorras R, Vespignani A (2004) Evolution and structure of the Internet: a statistical physics approach. Cambridge University Press, Cambridge
21. Pastor-Satorras R, Vázquez A, Vespignani A (2001) Dynamical and correlation properties of the Internet. Phys Rev Lett 87:258701-1-4
22. Ravasz E, Barabási A-L (2003) Hierarchical organization in complex networks. Phys Rev E 67:026112-1-7
23. Steeples DW, Steeples DD (1996) Far-field aftershocks of the 1906 earthquake. Bull Seismol Soc Am 86:921–924
24. Tsallis C (1988) Possible generalization of Boltzmann–Gibbs statistics. J Stat Phys 52:479–487
25. Vázquez A, Boguña M, Moreno Y, Pastor-Satorras R, Vespignani A (2003) Topology and correlations in structured scale-free networks. Phys Rev E 67:046111-1-10
26. Vázquez A, Pastor-Satorras R, Vespignani A (2002) Large-scale topological and dynamical properties of the Internet. Phys Rev E 65:066130-1-12
27. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442
28. Wilson RJ (1996) Introduction to graph theory, 4th edn. Prentice Hall, Harlow

Earthquake Nucleation Process

Yoshihisa Iio
Disaster Prevention Research Institute,
Kyoto University, Kyoto, Japan

Article Outline

Glossary
Definition of the Subject
Introduction
Contribution from the Development of Earthquake Early-Warning Systems
Observations of Initial Rupture Processes
Discussion
Future Directions
Acknowledgments
Bibliography

Glossary

Nucleation process The process in which rupture velocity accelerates from quasi-static to dynamic. The dynamic rupture velocity almost equals the shear wave velocity.

Nucleation zone The portion of the fault where rupture velocity accelerates from quasi-static to dynamic.

Initial rupture process The rupture process that precedes the largest slip. This term is used when the acceleration of rupture velocity is not clear. This is a wider concept that includes the earthquake nucleation process. The area where the initial rupture process occurs is called the initial rupture fault.

Slip velocity The dislocation velocity at a point on the fault. The rupture velocity is the velocity at which the rupture front is expanding.

Preslip model An earthquake source model having a detectable size nucleation zone.

Cascade model An earthquake source model in which smaller sub-events successively trigger larger sub-events. A sub-event is the same as a small earthquake if it does not trigger a successive sub-event.

Stress drop (static stress drop) The amount of shear stress change at a point on the fault before and after an earthquake. It is proportional to the strain released on the fault.

Dynamic stress drop The difference between the initial shear stress and the minimum frictional stress at a point on the fault during fault slip.

Fault strength The shear stress level necessary to initiate slip at a point on the fault.

M Magnitude. Earthquake size computed basically from waveform amplitudes and focal distances.

Seismic moment The most reliable measure of earthquake size which is determined from the products of the rigidity near the fault, the amount of slip, and the area of the fault surface.

M_w Moment magnitude. Earthquake magnitude derived from the seismic moment.

Definition of the Subject

Earthquake prediction in the long, intermediate, and short terms is essential for the reduction of earthquake disasters. However, it is not practical at present, in particular, for the intermediate and short time scales of a few days to years. This is mainly because we do not know exactly how and why earthquakes begin and grow larger or stop. Theoretical and laboratory studies have confirmed that earthquakes do not begin abruptly with dynamic rupture propagation. They show that a quasi-static rupture

growth precedes dynamic rupture. Thus, if we can detect the quasi-static rupture growth, we could forecast the following dynamic rupture. A key issue is how natural earthquakes initiate. To solve this issue, a first approach would be to investigate the very beginning parts of observed waveforms of earthquakes, since they can reflect the earthquake nucleation process from a quasi-static to a dynamic rupture. This paper reviews the studies analyzing the beginning parts of observed waveforms, and shows what we presently understand about the earthquake nucleation process.

Introduction

Earthquakes initiate at a small portion on a fault. Then, their rupture fronts expand outward until they stop. Some large earthquakes have a rupture extent greater than 1000 km (e. g., the 2004 Sumatra Earthquake), while fault lengths of very small microearthquakes ($M = 0$) are estimated to be a few meters [13]. Fault lengths of earthquakes range at least over 6 orders of magnitude. Surprisingly, the concept that earthquakes are self-similar is widely accepted in spite of the difference in fault length (e. g., [40]). One example of the similarity is that the average fault slip is proportional to the fault length. In other words, the static stress drop is constant independent of earthquake size.

The similarity law raises fundamental questions: Why do large earthquakes grow larger? What is the difference between large and small earthquakes? An end member model represents earthquakes as ruptures that grow randomly, and terminate in an earlier stage for smaller earthquakes while continuing longer for larger earthquakes. This type of model has been proposed mainly to explain the frequency–size distribution of earthquakes (e. g., [8]) and implies that it is impossible to forecast the final size of earthquakes at the time of the rupture initiations. However, another end member model predicts that larger earthquakes have a larger "seed" than smaller earthquakes and that large and small earthquakes are different even at their beginnings (e. g., [66]).

The key issue is how earthquakes initiate and grow larger.

All the theoretical models of earthquake sources predict that earthquake shear failures begin with a quasi-static rupture growth on a small portion on the fault (e. g., [5]). The shear failures become unstable and the rupture growth begins to accelerate after the energy released by the rupture growth equals to or becomes larger than the work necessary for producing new fracture surfaces. Finally, the rupture velocity, the velocity at which the rupture front is expanding, reaches a constant value comparable to the shear wave velocity. We call the process from the rupture initiation to rupture growth acceleration the earthquake nucleation process, and the portion of the fault where the rupture growth accelerates the nucleation zone. The important point is that the rupture velocity (and slip velocity) accelerates during the nucleation process, since it is a transient process from quasi-static to dynamic. One of the major differences between various models is whether the size of the nucleation zone is much smaller than the final fault length and is not detected by observations, and whether the size of the nucleation zone is different between large and small earthquakes.

If the size of the nucleation zone is extremely small and can be approximated as a point, the waveforms radiated from the earthquake are shown in Fig. 1a. Figure 1a displays the initial rise of the P-wave velocity pulse at a far-field (distant relative to the fault length) station, when the waveform propagates in a purely elastic homogenous medium with no attenuation and scattering under the following assumption: the rupture initiates at a point, the circular rupture front expands at a constant velocity, the stress on the fault abruptly drops from an initial level to a final level, and the stress drop is constant over the fault, as modeled by Sato and Hirasawa [60]. In this case, the initial portion of the P-wave velocity pulse is characterized by a linear rise (and the initial rise of the displacement pulse is quadratic with time, due to increasing circular fault area with time). The tangent of the linear rise is proportional to the stress drop and rupture velocity.

On the other hand, if the size of the nucleation zone is relatively large, the velocity pulse can be approximated as shown in Fig. 1b. It is seen that the linear rise is delayed and the slope of the initial portion gradually increases. Theoretical models explain such a slow rise as follows. Shibazaki and Matsu'ura [64] demonstrated that the slow rise could be explained by gradually accelerating rupture and slip velocities in a relatively large nucleation zone. The size of the nucleation zone is controlled by the critical slip distance, D_c, that is the amount of slip necessary to drop the peak shear strength down to the dynamic frictional level (e. g., [19]). This frictional behavior is called slip-weakening, which is theoretically predicted (e. g., [15]) and is observed in laboratory experiments (e. g., [52,56]). According to their theory, since large slips are necessary to decrease the friction on faults with a large D_c, the rupture and slip velocities are not accelerated in the beginning on the faults. Sato and Kanamori [61] modeled the slow rise by Griffith's fracture criterion based on the energy balance indicated above, as the rupture velocity grad-

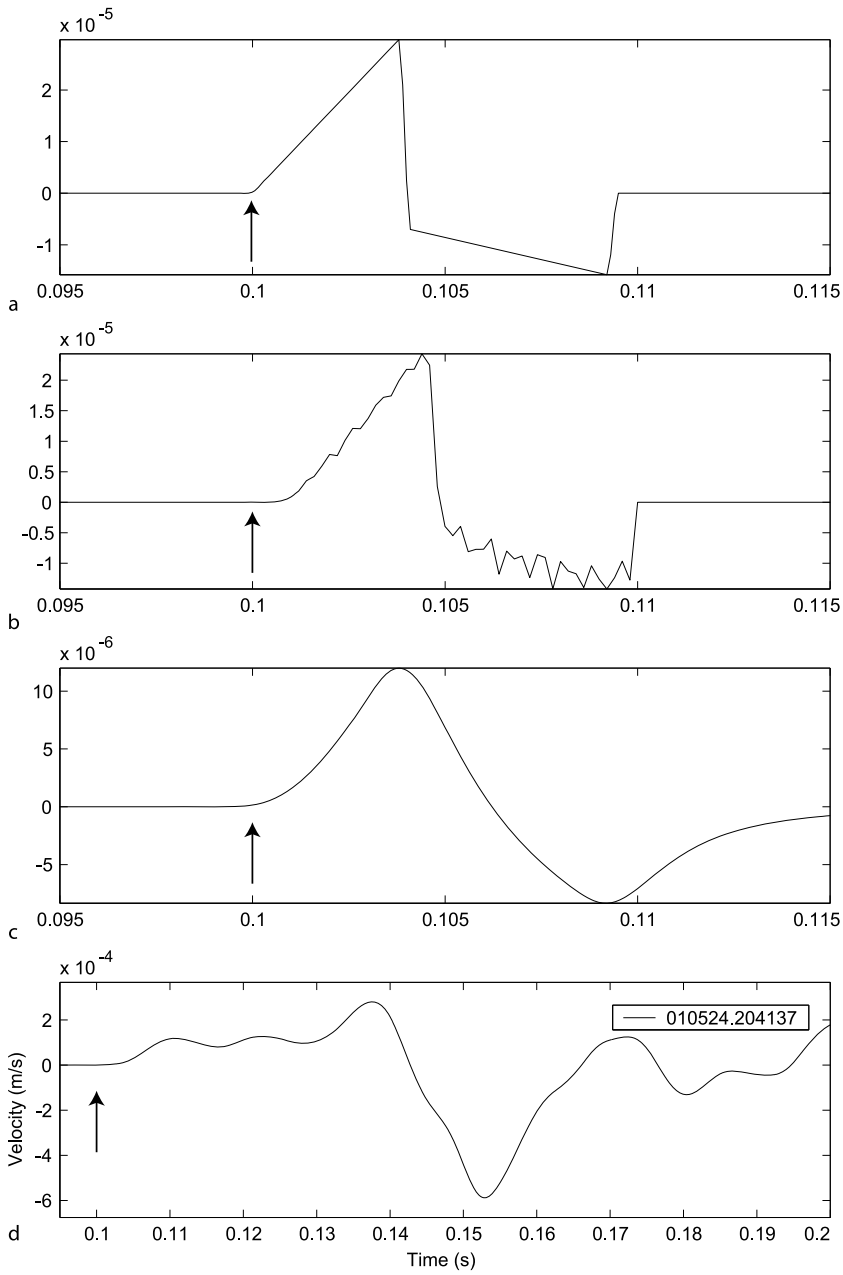

Earthquake Nucleation Process, Figure 1

Examples of P-wave velocity pulses. Arrows indicate onsets of P-waves. a Source pulse from the circular fault model [60]. Stress drop: 1 MPa, fault radius: 17 m, rupture velocity: 0.8 Vs, take-off angle: 61 degrees. b Source pulse from a circular fault model with a rupture velocity accelerating with time for first 20% of the total duration time. The other parameters are the same as a. c *Q* convolved velocity pulse at a distance of 3.58 km from the source. *Q* is set as 300. The other parameters are the same as a. d Velocity waveform of a complicated shape observed in the Western Nagano Prefecture region

ually increases under the assumptions of a large pre-existing fault and small trigger factor (instantaneous small stress increment) on the fault. The pre-existing fault is often called the initial crack. In this model, large initial cracks result from large fracture energies on the fault and the rupture and slip velocities are not accelerated right after the rupture onset owing to the large fracture energies. In this paper, the model having a detectable size of the nu-

cleation zone is called the preslip model, following Beroza and Ellsworth [9].

It is important to examine whether initial rises of observed velocity pulses are linear or gradually increasing, since the above two models predict different initial rises. Although the examination seems to be very easy at first glance, it is actually very difficult. The path effect is an inevitable obstacle to the examination, in particular for small earthquakes. The effects of anelastic attenuation and/or scattering contaminate observed waveforms and can reproduce a slow rise of observed waveforms even though the waveforms show a linear rise at the source. Figure 1c shows the initial rise of velocity pulses calculated by considering anelastic attenuation by a convolution of a Q operator [7]. It is found that the linear rise is delayed as shown in Fig. 1b.

Another serious obstacle is the complexity of observed waveforms. They are not often as simple and smooth as those shown in Figs. 1b and c, but complicated as shown in Fig. 1d. If several small patches in a relatively large nucleation zone break during the nucleation process [9,65] as shown in Fig. 2a, it is difficult to detect an acceleration of the rupture velocity from the complicated waveforms that include several small phases radiated from the breaks of the small patches. In other words, it is difficult to infer the nucleation process from the complicated waveforms. On the other hand, the model with small nucleation zones claims that the complicated waveforms are radiated only from successive breaks of small sub-events and a delayed break of the largest sub-event, as shown in Fig. 2b. This concept is known as the cascade model (e. g., [1,9,23,78]). In this model, a former sub-event triggers successive larger sub-events and the final event is the largest among detectable sub-events. Although it is not always clear how larger sub-events are delayed from former smaller sub-events, the growth of the rupture can raise the possibility that larger sub-events are generated. For small earthquakes also, it is possible that very small sub-events trigger successive sub-events. In this case, their waveforms produced by the source are complicated, but the observed waveforms are likely to show a smooth initial rise as shown in Figs. 2b and c, due to the path effects. Thus, in this paper, both smooth initial rises of small earthquakes and complicated initial portions of large earthquakes are discussed in the same manner.

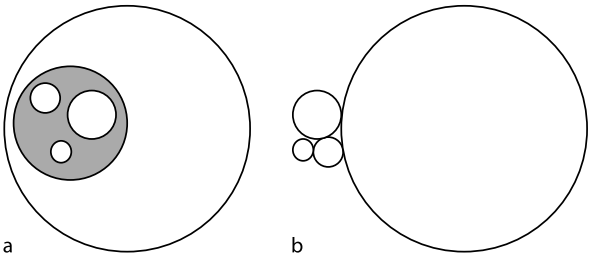

Earthquake Nucleation Process, Figure 2
Schematic source models in which small amplitude phases are generated by breaking of small fault patches within the nucleation zone (*shaded portion*) a, and as successive small subevents b. *Large* and *small circles* are the mainshock fault and small subevent fault patches, respectively

By the way, if the size of the nucleation zone is relatively large, it could be possible to detect a quasi-static rupture growth at the very beginning of the earthquake nucleation process by near-field broadband instruments, such as strain meters. However, it has not, to present, succeeded except for a few very large earthquakes, such as the 1964 Chile earthquake (e. g., [41]). This fact suggests the possibility that the size of the nucleation zone is too small to be recorded by strain meters or the duration of the quasi-static rupture process is much longer than a practical frequency range of strain meters. Furthermore, the observations for very large earthquakes were not explained by the nucleation process that occurs near the hypocenter but by slow slips on the downward extension of the seismogenic faults (e. g., [33,41]). Thus, we will examine data obtained by seismometers.

It is very important to carefully analyze observed waveforms recorded at a short focal distance by a wide-dynamic range and frequency range. This paper reviews various studies related to the earthquake nucleation process, indicates the problems about these studies, and summarizes the current observations. First, in Sect. "Introduction", contributions from earthquake early-warning systems are reviewed, since they are directly influenced by the problems cited above. In Sect. "Contribution from the Development of Earthquake Early-Warning Systems", important studies about the initial rupture process are reviewed basically in the order in which they were published. In Sect. "Observations of Initial Rupture Processes", a probable model for the process will be proposed based on the reviews in the former sections.

Contribution from the Development of Earthquake Early-Warning Systems

To solve the problem of how earthquakes initiate and grow larger, the most straightforward approach is a comparison between initial rises of observed waveforms of large and small earthquakes. Important studies were done for the development of earthquake early-warning systems which hope to forecast, as early as possible, the final size of earth-

quakes from observed waveforms, from the very beginning parts of seismograms. Olson and Allen [57] applied the method by Allen and Kanamori [3] to a new dataset including many large earthquakes ($M_w > 6$) and claimed that the final size of earthquakes ($3 < M_w < 8$) can be estimated from waveforms of the first several seconds. The method is based on the scaling relationship between the predominant period of waveforms and the final size of earthquakes, which was first derived by Aki [2]. To implement the scaling relationship in earthquake early-warning systems, Allen and Kanamori [3] adopted an algorithm to estimate the period from a short waveform, computing regressively the ratio of velocity and displacement amplitudes (modified after Nakamura [49]).

For smaller earthquakes, it is quite natural to be able to estimate the final size from the first several seconds of the waveforms, since the waveforms cover the entire source duration. The problem is estimating large earthquakes ($M_w > 5.5$). Rydelek and Horiuchi [58] evaluated the statistical significance of the results by Olson and Allen [57] and reported that the results are not clear. Rydelek and Horiuchi [58] analyzed waveforms observed by Hi-net, a nation-wide high gain seismometer network operated by NIED in Japan, and found no trend between the period and earthquake size for larger earthquakes. On the other hand, Wu et al. [77] analyzed many waveforms recorded at the southern California Seismic Network stations and showed that the period increases with magnitude ($4 < M < 7.5$). For earthquake early-warning system applications, where a quick response is essential, a precise waveform analysis is not required. A more careful analysis is needed to answer the question raised by Rydelek and Horiuchi [58].

Observations of Initial Rupture Processes

Measurements of Initial Parts of Observed Seismograms

How are Initial Portions of the Waveforms Measured? Observed waveforms of large to small earthquakes are measured in various manners in the studies reviewed in the following sections. Sometimes special names are given to a characteristic phase of the waveforms. First, the problems concerning the measurement of observed waveforms will be discussed.

Many studies try to find which phase is radiated from the source process with a constant rupture velocity and slip velocity, since the nucleation process is basically an accelerating rupture process. Since the nucleation process is a beginning rupture process, it is likely that the rupture front just after the nucleation process expands nearly circularly. Consequently, the phase following the nucleation phase is thought to show a linear rise in velocity waveforms. For this reason, the first linear rise in velocity waveforms is extensively investigated in various studies.

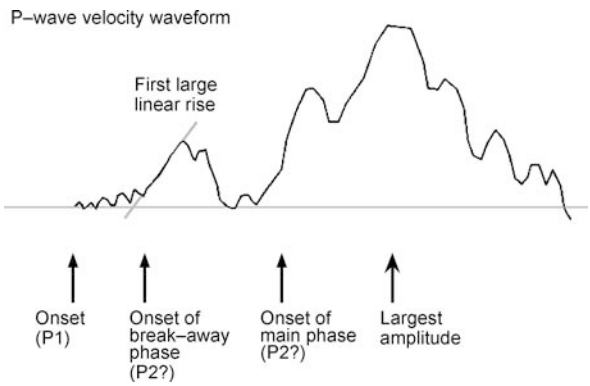

Earthquake Nucleation Process, Figure 3
Schematic illustration of large and small earthquakes to show the first linear rise and the main phase

Various studies also measure a pulse of the largest amplitude or a group of pulses including the largest amplitude, which is called the main phase in this paper. Amplitudes of waveforms are thought to show the maximum value after they reach terminal velocity, which is the maximum rupture velocity determined by the shear wave velocity. However, the largest amplitude does not necessarily occur just after rupture velocities reach the terminal velocity; all the waveforms preceding the maximum phase do not necessarily reflect the nucleation phase.

The problem is determining which phase is radiated just after rupture velocities reach terminal velocity. Furthermore, since slip velocities can cover a wide range of magnitude, it is important to investigate whether slip velocities are similar to the average values of earthquakes. In the following, when it is necessary to explicitly indicate the phase generated by faulting with a nearly constant rupture velocity and dynamic stress drop that are representative of average values for earthquakes, the phase is called the ordinary phase.

As schematically shown in Fig. 3, for large earthquakes, the first linear rise is not always a part of the main phase, which has the largest amplitude. For small earthquakes, the main phase often shows a linear rise. However, it is possible that the main phase is not the ordinary phase, as we discuss in a later section.

Large Earthquakes One of the first studies that indicated the importance of small amplitude waveforms preceding the main pulse is Furumoto and Nakanishi [24],

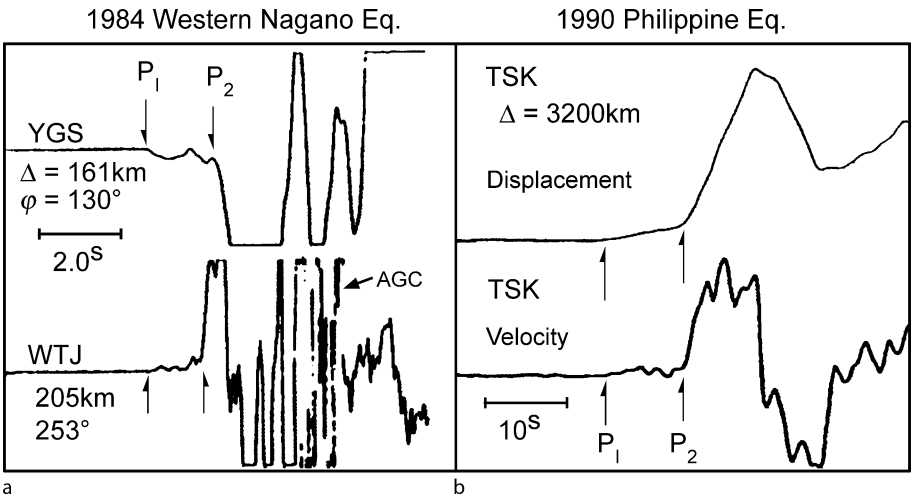

Earthquake Nucleation Process, Figure 4

Initial parts of observed waveforms of large earthquakes [73]. P1 and P2 indicate the onsets of P-waves and the large amplitude phases, respectively. Velocity waveforms are shown in a and the lower panel of b. Figure 7 in [73], copyright 1992 by Elsevier Science Publishers B.V.

which analyzed long-period waveforms from a worldwide network; this was followed by many similar studies (e. g., [4,14,68]). Umeda [72,73] analyzed broadband seismograms at local, regional and teleseismic distances and found that small amplitude waveforms preceded large amplitude waveforms, as shown in Fig. 3. Umeda [72] called the onsets of the P-waves and large amplitude waveforms P1 and P2, respectively. He originally thought that the large amplitude waveforms were generated by somewhat special processes that characterized large earthquakes, e. g., breaks of many small faults triggered by both static and dynamic stress concentrations due to the preceding fault motion and/or a dynamic connection of separated faults [74], while the smaller waveforms were radiated from ordinary smooth rupture propagation. His important result is the relationship between the duration of P2-P1 and magnitude ($3 < M < 7$), suggesting earthquake sizes are scaled with P2-P1, the duration of small amplitude waveforms.

The determinations of P2 seem to be reasonable from his figures. However, the criterion of P2 determination is not necessarily clear and quantitative compared to the studies cited in the following sections. In many cases it seems that P2 is the onset of the main phase that includes the largest amplitude, in particular for smaller earthquakes, but this point is not clearly shown in his papers. Furthermore, it is seen that absolute velocity amplitudes after P1 do not show an accelerating increase with time, although both the preslip model and the cascade model basically indicate that velocity amplitudes should have a growth that is faster than linear with time.

One of the important problems raised by Umeda [72,73] is identification of the ordinary phase, the main phase or preceding smaller amplitude waveforms. The ordinary phase is the phase generated by faulting with a nearly constant rupture velocity and a dynamic stress drop, which are representative of average values for earthquakes. Furumoto and Nakanishi [24] regarded the main phase as the ordinary phase with a special rupture process occurring before the main phase, while in Umeda [72,73] the opposite is the case.

Small Earthquakes Iio [30,31] analyzed waveforms of microearthquakes recorded at very short focal distances (down to a few hundreds of meters) using instruments with a wide-frequency response and found that the initial rises of velocity pulses did not show a linear increase but, rather, a convex downward (or upward) shape as approximated by t^n (t is the time measured from the onset and $2 < n < 4$). He termed the initial rise "the slow initial phase", and measured the duration of the slow initial phase relative to the main phase that shows a linear increase, as shown in Fig. 5. He found that the duration of the slow initial phase is proportional to the earthquake size. Although the slow initial phase of microearthquakes could be the effect of anelastic attenuation as shown in Sect. "Introduction", Iio et al. [32] analyzed velocity pulses recorded at a 10 kHz sampling frequency at numerous stations at

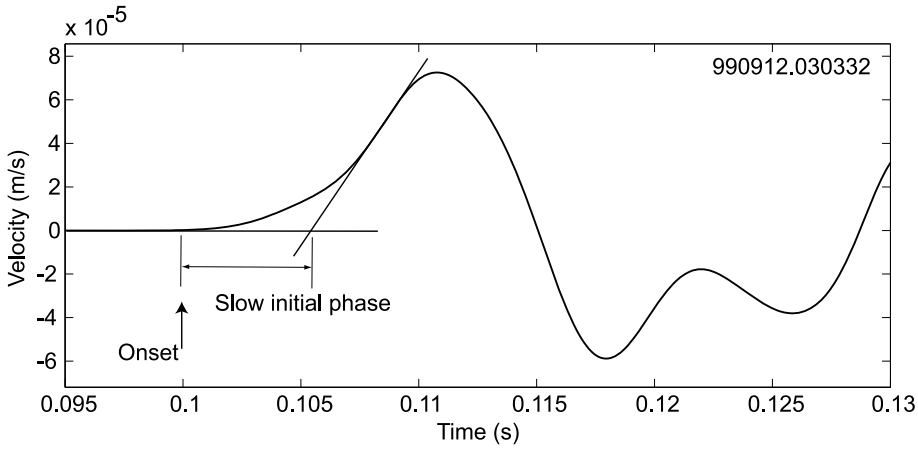

Earthquake Nucleation Process, Figure 5

Slow initial phase of the velocity pulse observed in the Western Nagano Prefecture region and the measurement of the duration of the slow initial phase relative to the main phase that shows a linear increase of velocity amplitudes. The vertical arrow indicates the onset of P-waves, determined by their amplitudes. The portion indicated by the *horizontal arrow* is defined as the slow initial phase. The inclined line is the tangent at the maximum slope. The same pulse is shown in Fig. 7 by the *black line*

short focal distances using an instrument with a wide dynamic range and frequency range, and concluded that the slow initial phase mainly reflects the source process for M2 events.

These observations are unique and important, but their interpretations and implications contain several fundamental problems and might cause some misunderstandings. First, it was intuitively regarded that the main phase showing the linear rise was the ordinary phase radiated from a circular fault with constant rupture and slip velocities of ordinary magnitudes. This was derived from the interpretation that the slow initial phase reflected the nucleation process in which the rupture and/or slip velocities gradually accelerated. However, the interpretation has not been thoroughly examined. As emphasized by Ellsworth and Beroza [21] and Beroza and Ellsworth [9], waveforms recorded by seismometers basically reflect slips on the fault that have accelerated above a certain level.

The second problem is the meaning of the scaling relationship indicating that larger earthquakes have a longer slow initial phase. Although the possibility was suggested from this relationship that larger earthquakes had a larger nucleation zone, this is true only if the main phase is the ordinary phase and the slow initial phase reflects the nucleation process. The relationship could be also misunderstood as larger earthquakes begin more slowly. The observational results mean only that the portion of velocity pulse approximated by t^n $(2 < n < 4)$ is longer for larger earthquakes. Furthermore, Iio [30,31] mentioned nothing about the difference between the slopes right after the onsets of larger and smaller earthquakes. Thus, another possible explanation is that for larger earthquakes, the slope of the velocity pulse increases with time for a longer period and the maximum slope becomes larger, even though the initial slope is the same as that for smaller earthquakes. This possibility implies that the final size of earthquakes is not estimated only from the initial part of velocity pulses, as pointed by Mori and Kanamori [48], although larger earthquakes have a greater dynamic stress drop. Several studies have indicated that the initial rises of large earthquakes $(M > 6)$ are similar to small events occurring in the vicinity of their hypocenters (e. g., [12]), while other studies have indicated that larger earthquakes display a larger initial rise (e. g., [35,50]). However, it is possible that the initial parts of these waveforms mainly reflect the initial rise radiated from a small patch within the initial rupture fault, as shown in Sect. "Introduction". In that case, the results did not negate the existence of an observable earthquake nucleation process for large events. Further, these results were obtained only for a few earthquake sequences and need to be systematically examined for a greater data set.

The third problem is the propagation effects from the source to the observation stations. Iio [30,31] thought that observed waveforms basically reflect the characteristics of the source process in which the rupture and slip velocities gradually increase during the slow initial phase, assuming that source pulses of small earthquakes are simple and consist of a single event. However, it is possible that anelastic attenuation modifies two connected linear trends

of small and large slopes from two subevents, creating a smoothed waveform of convex downward shape. This point is qualitatively examined in a later section. Furthermore, it might be possible that waveforms near the source are complex; several very short duration small pulses may precede the main phase, since short duration pulses can be smoothed by the path effect to produce a slow initial phase.

Estimate of Source Time Functions

Ellsworth and Beroza [21] and Beroza and Ellsworth [9] focused on initial parts of seismograms of a very wide range of earthquake sizes ($1 < M_w < 8$) and estimated source time functions. They paid attention to the first large linear rise of velocity pulses and deduced that the signal before the linear rise reflected the earthquake nucleation process, since a linear rise of velocity pulses is characteristic of the waveforms radiated from an ordinary circular fault model. They termed the linear rise "the breakaway phase" and the portion before "the seismic nucleation phase". The term "seismic" is added, since waveforms detected by seismometers are radiated from dynamic slip on the fault. The criterion for the detection of the breakaway phase is not necessarily quantitative, but it seems from their figures that their classifications are reasonable. As shown above, they regarded the breakaway phase as the ordinary phase, since they roughly estimated that dynamic stress drops of the breakaway phases were several or several tens of MPa, that is comparable to average values of earthquakes. However, it is possible that the seismic nucleation phase is the ordinary phase, since the rupture velocity and slip velocity during the seismic nucleation phase is not quantitatively estimated.

Kilb and Gomberg [43] analyzed initial portions of the waveforms of the Northridge earthquake, which were also used in Ellsworth and Beroza [21] as a typical example, and claimed that the initial portions are very similar to those from nearby small earthquakes. They inferred from these results that the cascade model is plausible. However, their results did not exclude the preslip model in which the rupture and slip velocities are accelerating in the nucleation zone, since they analyzed only the very beginning portions and not the following portions that might result from slow slips in the nucleation zone. Shibazaki et al. [67] discussed the possibility that large earthquakes begin by a breaking of a small patch in the nucleation zone. As discussed in the Introduction, the breaking of small patches can mask the nucleation process.

Ellsworth and Beroza [21] demonstrated the relationship between the duration of the seismic nucleation phase and the seismic moment, together with the data of the slow initial phases by Iio [30,31] and P2-P1 by Umeda [72,73], as shown in Fig. 6, and concluded that the duration of the seismic nucleation phase scaled with the seismic moment. However, it is seen in Fig. 6 that the data of Iio [30,31] and Umeda [72,73] are shifted upward from the regression line. It seems that their data for small earthquakes also shifted upward from the regression line. The offsets of these shifts are about 1 order of magnitude. For larger earthquakes, this may be because Umeda [72,73] determined P2 as the onset of the main phase, while Ellsworth and Beroza [21] determined onsets of the first large linear rises. For smaller earthquakes, Iio [30,31] also detected the onsets of the main phases. They regarded the phases as the first large linear rises, but it is possible that the onsets of the first large linear rises are earlier than the main phases, as discussed in the following section.

Estimate of Source Processes

Small Earthquakes In order to solve the problems cited in the previous sections, several studies tried to estimate the source process that produces the initial phases. Hiramatsu et al. [26] analyzed rising parts of seismograms using the model of Sato and Kanamori [61], in which larger initial cracks generate a slower initial phase under a small triggering factor.

It was inferred from deep borehole (1800 m) seismograms by Hiramatsu et al. [26] that initial rises of velocity pulses of 5 events were explained by the ordinary circular fault model [60], namely the initial crack is too small to be detected, while the other 7 events needed a large initial crack. They first estimated source processes of the beginning of such small earthquakes; however, some basic problems remained. The first is that they modeled the first half cycle of velocity pulses. Since even portions of the first half cycle of the waveform can be affected by rupture arrest, it is not reasonable to fit the first half cycle of the waveform by using a model that does not include a reasonable rupture stopping process. Furthermore, a closer inspection of their results (Fig. 5 of [26]) reveals that some of the waveforms modeled with an initial crack do not display a smooth increase of the rising slope but appear to consist of two linear phases with different slopes. It is possible that these earthquakes result from a cascade rupture of a first and second sub-events with smaller and larger dynamic stress drops, respectively. For the data analyzed by Iio [30,31], waveforms with a predominantly longer slow initial phase appear to show a similar feature. Furthermore, it is not clear which is the ordinary phase, the first or second linear phase.

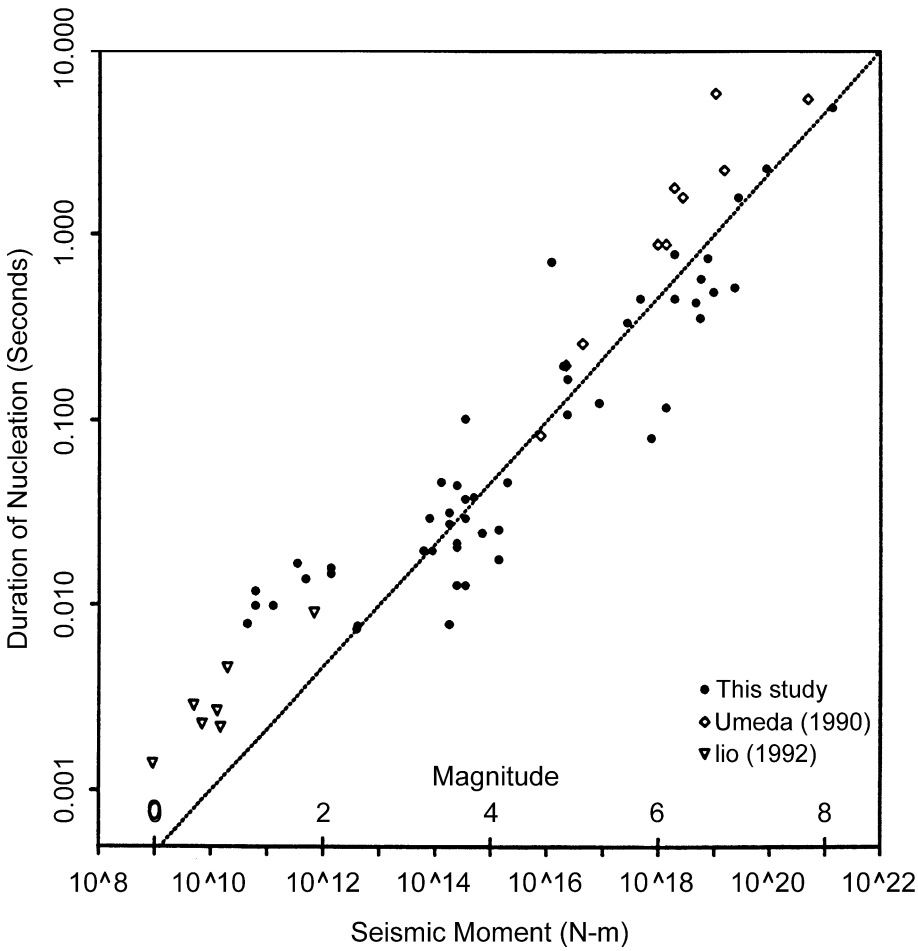

Earthquake Nucleation Process, Figure 6

Relationship between the duration of the seismic nucleation phase and the seismic moment, together with the data of the slow initial phases by Iio [30,31] and P2-P1 by Umeda [72,73](Beroza and Ellsworth [9])

The problem of identifying the "ordinary phase", the initial or main phase, is again pointed out. Recently, Iio et al. [34] analyzed the 10 kHz sampling data [32,36] and obtained results that suggest the initial rise is the ordinary phase.

Figure 7 shows velocity pulses of microearthquakes ($0.5 < M < 2.0$) for which relative locations are estimated within 100 m and fault plane solutions are similar [34]. These pulses were recorded at a borehole station at a depth of 800 m in the Western Nagano prefecture region [37,76]. The focal distances are about 3.6 km. It is seen that the smallest event (shown in red) shows a linear rise except for the first 1 ms, while slopes of the other events increase with time. Thus, if the linear rise is explained by a circular fault model of the Sato and Hirasawa type with an ordinary rupture velocity and stress drop, the increasing slopes reflect increasing slip velocities accelerated beyond ordinary values.

The linear rise of the waveforms was modeled by various kinematic fault models. Figure 8 displays the comparison of synthesized and observed velocity pulses for three borehole stations at depths of 800, 150, and 100 m [37]. The fault plane was determined from focal mechanisms and hypocentral distributions around these events. Since all three waveforms were not explained simultaneously by a Sato and Hirasawa type of circular fault model, they used a fan fault of various angles. Further, they did not assume that the slip terminated simultaneously over the fault, but that it begins to stop at a point on the fault edge and the stopping phase propagates circularly at a constant velocity. The stress drop, rupture and stopping phase expanding velocities, fault radius, fan angle and fan fault geometry were

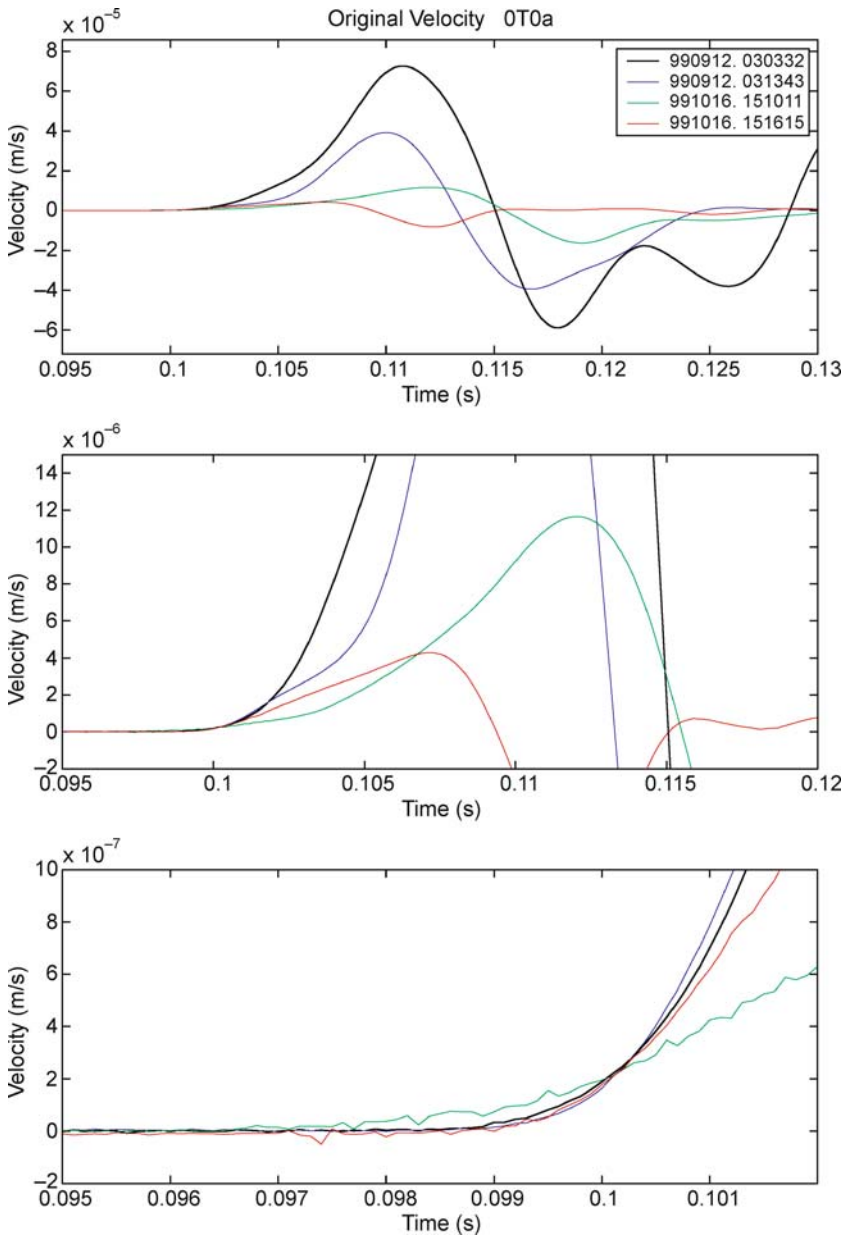

Earthquake Nucleation Process, Figure 7

Comparison of velocity pulses of microearthquakes ($0.5 < M < 2.0$) of which relative locations are estimated within 100 m and fault plane solutions are similar, recorded at a borehole station at a depth of 800 m in the Western Nagano prefecture region [34]. Focal distances are about 3.6 km

determined by a grid search technique. Q values are set as 300 by trial and error, to fit the very beginning initial rises with duration of about 1 ms.

It is seen that the observed waveforms in the first half cycle are well explained by the fan fault model, except for the middle trace, which might be contaminated by a surface reflection. The rupture velocity was estimated as 0.8 Vs (the shear wave velocity), which is similar to that for large earthquakes. The stress drop was estimated as about 5 MPa by the formula for the circular fault of the equivalent fault area. Although they modeled only two events, they found that a few percent of events that oc-

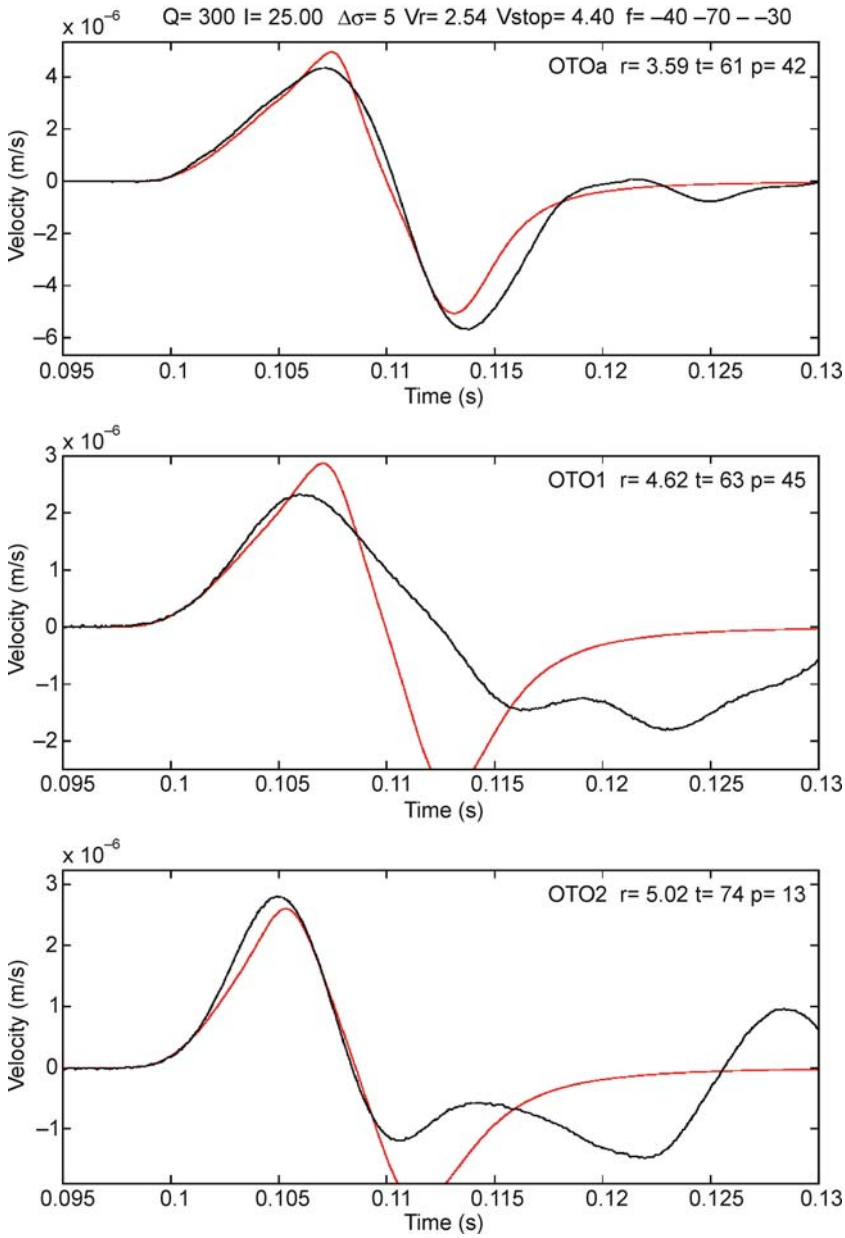

Earthquake Nucleation Process, Figure 8

Comparison of observed and synthesized waveforms for an event that displays a linear rise, shown in Fig. 7 by the *red line* [34]. Waveforms observed at three borehole stations of depths of 800, 150, and 100 m are shown in the *top*, *middle* and *bottom* panels, respectively. The synthesized waveforms are calculated by a fan fault model in the homogenous half space using the following parameters: The stress drop is estimated as about 5 MPa. The rupture and stopping phase expanding velocities are 0.8 Vs and 0.8 Vp (the P-wave velocity), respectively. Fault radius: 25 m, fan angle: 40. A *Q* value is 300

curred within about 5 km from the 800 m borehole station also showed such a linear rise. Furthermore, the other events generally show a steeper initial rise than the linear rise events. Consequently, it is likely that small earthquakes occurring in the Western Nagano prefecture region have a rupture velocity equivalent to those of large earthquakes even in the initial rupture process. This suggests that the slow initial phase of small earthquakes does

not reflect the nucleation process that is characterized by accelerating rupture velocity, since the rupture velocity has already accelerated to a shear wave velocity in an early stage of rupture growth.

What do steeper main phases of larger earthquakes in the Western Nagano region reflect? In order to clarify this problem, Miura et al. [46] analyzed waveforms of earthquakes ($0.0 < M < 4.0$) occurring from 1996 to 2003 in the Western Nagano prefecture region. They selected M3 events ($3.0 < M < 4.0$) and identified 21 earthquake clusters that consist of M3 events and earthquakes occurring within 500 m of the hypocenters of M3 events. They investigated P-wave velocity pulses observed at three borehole stations for each cluster, in particular the difference in pulse shapes of large and small earthquakes. They found that waveforms of half of the clusters displayed complicated shapes, as shown in Fig. 1d, which are characterized by more inflection points than simple pulses [63]. The other half showed simple waveforms that enabled them to clarify the difference between large and small earthquakes. One example of waveforms is shown in Fig. 9. The observed P-wave velocity pulses of $0.5 < M < 3.1$ are displayed in different magnifications. It is seen that initial rises are similar for the first 1 ms but that slopes of larger events increase with time. The theoretical pulses from the circular fault model [60] are synthesized at the same focal distances for a similar magnitude range in Fig. 9b, assuming a constant stress drop independent of earthquake size, and a Q value of 230, which was obtained by modeling the waveform of the smallest event. Responses of seismometers are also included. The synthesized pulse shapes of larger events do not display a linear rise but a slight convex upward shape. It is found that a distinct difference between theoretical and observed waveforms is seen in the maximum slope of larger events. This figure clearly demonstrates that larger earthquakes have larger dynamic stress drops than smaller events. Similar features are also seen in the other clusters that show simple waveforms. These results imply that the similarity law does not hold for a small range of magnitudes in the Western Nagano prefecture region, as indicated by Venkataraman et al. [58]. The similarity law predicts pulses as shown in the right hand side of the figure.

Large Earthquakes Sato and Mori [62] used the same method as Hiramatsu et al. [26] and analyzed the very beginnings of waveforms for a very wide range of magnitudes ($3 < M < 8$) recorded by high gain short-period seismometers at local distances. They showed that large initial cracks (a few tens of meters) are necessary to explain the initial rises and that initial crack lengths are almost constant, independent of the eventual earthquake size. Al-

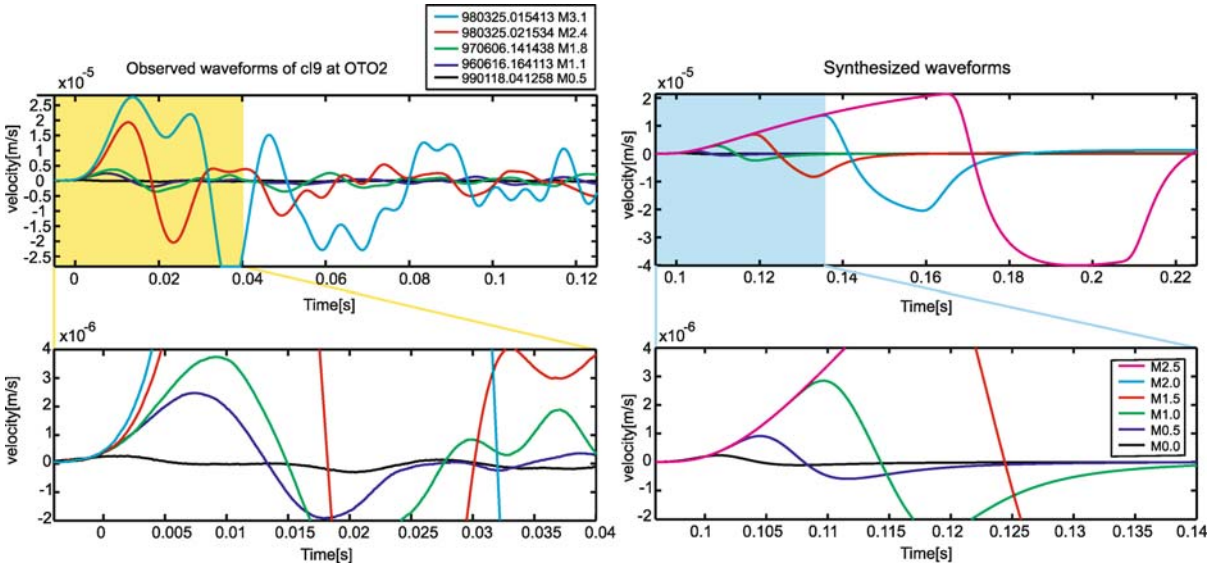

Earthquake Nucleation Process, Figure 9

Comparison of observed (*left*) and synthesized (*right*) waveforms for events of an earthquake cluster that shows smooth waveform shapes [46]. The cluster consists of the M3 class events and earthquakes occurring within 500 m from their hypocenters. The synthesized waveforms are calculated by an ordinary circular fault model [60] at the same focal distance. A Q value of 230 was obtained by modeling the waveform of the smallest event. Responses of the instruments are also convolved

though they might have analyzed only the beginning of the first sub-event in the nucleation zone, it is important that they found that very beginning portions of large earthquakes show an accelerating rise over the first 0.1 s.

For large earthquakes, initial rupture processes can be estimated by near-source broadband seismograms. Shibazaki et al. [67] determined a P2 phase 0.5 s after the P-wave onset (P1) for the Kobe earthquake and performed a waveform inversion using the initial portion of seismograms. They estimated that the slip velocity slowly accelerated for 0.5 s. The average rupture velocity within 0.5 s after the onset of dynamic rupture was about 2 km/s, which is not very low. However, their results possibly detect the earthquake nucleation process, since it is difficult to estimate accelerating rupture velocity in the small nucleation zone of about 500 m, which was inferred from the waveform inversion. It is noted that their estimate of P2-P1 of 0.5 s is consistent with the scaling relationship of Ellsworth and Beroza [21], but is not consistent with that of Umeda [72,73]. The initial phase of the Kobe earthquake shows an accelerated increase of velocity amplitude with time [67], while a significant part of the data analyzed by the other studies do not show the accelerated increase. These results suggest the possibility that a major part of the initial rises analyzed in the other studies do not reflect the earthquake nucleation process.

Recently, longer durations of P2-P1 were observed for two intraplate earthquakes in Japan, the 2000 Western Tottori Prefecture earthquake (M7.3) and the 2004 West-off Fukuoka earthquake (M7.0). The initial phases of long durations are suitable for analyzing the rupture processes. For the Western Tottori Prefecture earthquake, Hirata [27] measured a duration of 2.5 s for P2-P1, determined the location of the P2 source and estimated the initial rupture process by a waveform inversion. They estimated that during the first 2.5 s, a small slip occurred within the limited area around the hypocenter and later, a large slip began near the location of P2 about 5 km southeast of the hypocenter. The average rupture velocity during the first 2.5 s was estimated to be 1.8 km/s from the waveform inversion of the initial phase. The stress drop in the limited portion was estimated to be small compared with the large slip area.

Since the duration of P2-P1 is long, the above features are also obtained in ordinary waveform inversions. Iwata and Sekiguchi [39] estimated the slip distribution using strong ground motion seismograms and geodetic data, and found a large slip area located about 5 km southeast of the hypocenter. They regarded the large slip area as an "asperity" on the fault. It has been well known that slip distributions of large earthquakes are heterogeneous

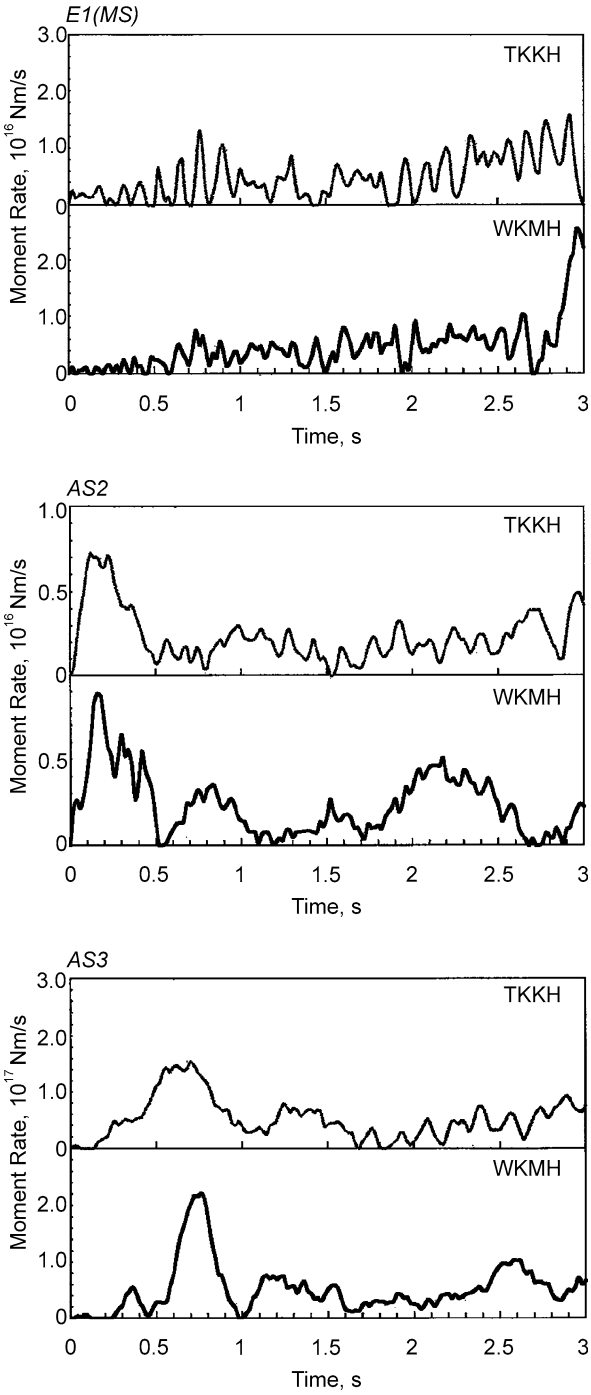

Earthquake Nucleation Process, Figure 10
Source time functions for the initial part of the 2004 West-Off Fukuoka earthquake (M7.0) The waveforms after the onsets are shown. The mainshock (*top*), a M4.5 aftershock (*middle*), and a M5.4 aftershock (*bottom*) estimated by an empirical green function method [75]. Zisin, 59:250, Figure 11, copyright 2007 by the Seismological Society of Japan

and large slips occur within a limited portion on the fault (e. g., [44]). Although the origin of large slip areas has not yet been clarified, one interpretation is that fault strength is higher in the areas of large slip than in the surrounding areas. This concept is called the asperity model. The observations about P1 and P2 shown above can be explained by the asperity model: The earthquake rupture begins at the weakest portion on the fault, then propagates to stronger portions where large slips occur.

For the 2004 West-off Fukuoka earthquake (M7.0), similar results were obtained. A P2 phase was detected by Yamaguchi et al. [75] and the duration of P2-P1 was long, estimated as 3.38 s. They determined the location of the P2 source as 3.44 km southeast of the hypocenter. The average rupture velocity during the 3.38 s period was computed to be about 1.02 km/s from the distance between the locations of P1 and P2. These results are consistent with those from ordinary waveform inversion studies (e. g., [6]). They also estimated the source time functions of the mainshock and a few large aftershocks by an empirical Green function method, as shown in Fig. 10. It was found that the source time function of the mainshock is not impulsive but has a gradual onset and long duration, while those of the aftershocks are impulsive. Only one aftershock shows a small initial phase, as shown in Fig. 10c. The seismic moment released before the P2 phase is comparable for those of M4 aftershocks. These results clearly show that the initial rise of the West-off Fukuoka Prefecture earthquake is different from those of the small aftershocks and is characterized by a small stress drop. It is suggested that large earthquakes do not accidentally grow larger.

For these two earthquakes with long initial phases, the average rupture velocities and stress drops on the initial rupture faults are small; however, the initial phases do not show a accelerating increase of velocity amplitude with time. These facts suggest the possibility that long initial phases do not reflect the nucleation process.

Discussion

Summary of the Observations

The reviews in the previous sections have revealed several important characteristics concerning the initial phases that are commonly seen in many studies. In this section, first, these characteristics are summarized, and then, an inferred initial rupture process will be discussed.

The almost linear initial rises were observed in the Western Nagano region and were well explained by a fan fault model with a constant rupture velocity comparable to the shear wave velocity. On the other hand, nearby larger earthquakes showed an initial rise in which the slope is of the same order of magnitude as that of the linear rise, plus increasing slopes with times of a few to several ms after the onset [34]. Thus, it is thought that the rupture velocity of these larger earthquakes is also comparable to the shear wave velocity during the initial phase. Furthermore, observed waveforms do not necessarily display a gradual increase of rising slope but sometimes show a discrete change in the slope, suggesting that the initial rise of the velocity pulse at the source is not smooth. Consequently, successive sub-events with larger dynamic stress drops possibly produce larger earthquakes for these events occurring in the Western Nagano Prefecture. These inferences are consistent with the results obtained by Hiramatsu et al. [26] that about half of their data are explained by an ordinary circular fault model and longer initial phases are not always smooth. For small earthquakes ($M < 4$), the slow initial phase probably does not reflect the earthquake nucleation process. The nucleation size of small earthquakes is probably small and thus, the very beginning of observed waveforms should be analyzed very carefully considering path effects. This matter is beyond the scope of this paper and is left to future studies.

Larger earthquakes showed a variety of initial phases, as pointed out by Ellsworth and Beroza [21]. It is likely that the observed initial phase of the Kobe earthquakes reflects the earthquake nucleation process [67]. Also, that of the Northridge earthquake probably reflects the earthquake nucleation process, since the very weak initial phase shows an accelerated increase of velocity amplitudes with time [21], although the very beginning part of the initial phase is similar to the waveform of nearby small aftershocks [43]. Further, several data of Ellsworth and Beroza [21], in particular, those shifted downward from the regression line shown in Fig. 6 possibly reflect the earthquake nucleation process. However, it is likely that those of the other large earthquakes do not reflect the earthquake nucleation process, since they do not show an accelerated increase with time but rather are roughly flat. These facts are clear for the two intraplate earthquakes that have a relatively long initial phase (e. g., [27,75]). For these two earthquakes, the average rupture velocity during the initial phases are estimated as 1.02 to 1.8 km/s, slightly smaller than ordinary values, and the stress drops on these large initial rupture faults are estimated to be small (e. g., [27,75]). Ordinary waveform inversions for these earthquakes showed that the main phases were generated by the breaking of asperities (e. g., [6,28,39]).

As summarized above, a major part of the initial phases of large earthquakes do not show an accelerating increase of velocity amplitudes with time and it is thought that the rupture and slip velocities of these earthquakes are

not accelerating during this time period. Consequently, it is thought that the initial portions of the observed waveforms of these large earthquakes do not reflect the nucleation process that is characterized by the transition from a quasi-static to a dynamic state; instead they represent a part of the dynamic rupture process characterized by a smaller stress drop, before the breaking of large asperities. Small earthquakes possibly show similar characteristics with an initial rupture process that is characterized by an ordinary rupture velocity and much smaller dynamic stress drop than the main phase. Although the nucleation process is probably seen in the very beginning stage of the large earthquakes analyzed by Sato and Mori [62] and Shibazaki et al. [67], it is likely that the scaling relationships between the duration of the initial phases and the rupture size shown by Umeda [72,73], Iio [30,31], and a major part of the data Ellsworth and Beroza [21] do not reflect the nucleation process but a part of the dynamic rupture process before the breaks of relatively large asperities.

A Possible Model

As summarized in the previous section, since a major part of the observed data do not reflect the earthquake nucleation process, in particular the data from small earthquakes, we cannot examine the preslip model here.

The observed data do not match the cascade model, since the cascade model basically predicts self-similar fault breaks for successive events, not increasing dynamic stress drop events. A variation on the cascade model, the hierarchy fault model explains increasing slopes of initial rises by the abrupt increase of the moment rate function due to a longer fault edge at a higher hierarchy level [23]. However, as shown in Fig. 6, the difference between the slopes of larger and smaller earthquakes can be greater than one order of magnitude, so it is difficult to explain the observed data only by the change in rupture front geometry. It may be necessary to consider differences in dynamic fault parameters.

So the questions to be answered are, why do larger earthquakes have a longer slow initial rupture process? In the first place, why do earthquakes need an initial rupture process to break a stronger portion on the fault? Why don't earthquakes initiate as a breaking of an asperity without the initial phase?

The key to solving these problems may lie in the studies of the two intraplate earthquakes with longer initial phases. The important observations about these earthquakes are the geometries of the initial rupture fault and mainshock fault. For the 2000 Tottori earthquake, it is found that azimuths of the initial rupture fault and mainshock fault are estimated as N135°E and N150°E, respectively [51]. Since the direction of the maximum compressional stress is estimated as N90°–100°E [42], it is found that the initial rupture fault is favorably oriented, while the mainshock fault is unfavorably oriented. For the 2004 West-off Fukuoka prefecture earthquake, the azimuth of the initial rupture fault is different from that of the mainshock fault by 20 degrees (e. g., [38,69,71]). A more obvious example is obtained for the Landers earthquake, where the Emerson fault with the largest slip is unfavorably oriented, compared with the faults that had broken before (e. g., [25]). These observations indicate that a larger slip occurred on an unfavorably oriented fault. It is possible that the mainshock fault, which generates the main phase, generally has a higher strength than the initial rupture fault. This may be the reason why larger dynamic stress drops occurred on faults of successive larger events.

If this proposed strength profile on the initial and mainshock faults holds for all the earthquakes, the above questions can be re-written: Why is a larger weak initial rupture fault necessary to break a larger strong mainshock fault?

A similar problem has been discussed by Ohnaka [55]. He derived the relationship between the critical slip distance D_c and the asperity size from laboratory experiments and the physics of contacts on faults. Since D_c is related to the size of the nucleation zone (e. g., [52]) and it is inferred from the results of waveform inversions that the asperity size is proportional to the total fault length (e. g., [47]), his relationship might be regarded as the relationship between the lengths of initial rupture fault and the mainshock fault. However, the D_c used in his relationship is that of the asperity, not of the initial rupture fault. Further, it is inferred from the above discussion that a major part of observed initial phases probably do not reflect the nucleation process (namely, D_c).

It is important to clarify strength profiles along faults; however, we do not presently have enough information. In the following, we will assume the fault strength is controlled by the geometry of the fault surface, as estimated for the Tottori earthquake. More concretely, it is assumed that the angle between the tangent of the local fault surface and the direction of the uniform principal stresses controls the fault strength at each point. In this case, it is deduced that the fault strength shows a fractal-like distribution along the fault as shown in Fig. 11, since the geometry of fault surface is thought to be fractal [45]). The fault strength profile should include various wavelengths, but here we will consider the longest wavelength to investigate the interaction between the asperity and initial

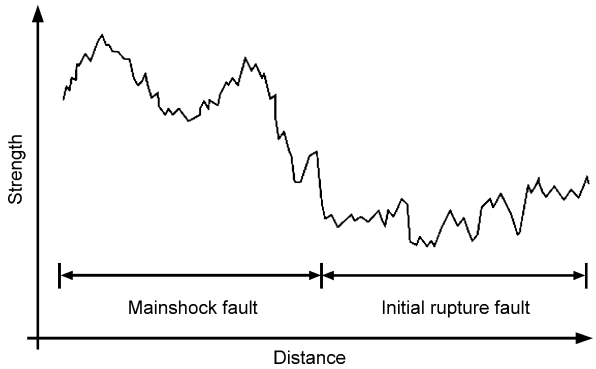

Earthquake Nucleation Process, Figure 11
Schematic illustration of the strength profile along the fault deduced from a fractal geometry of fault surface. The rupture initiates at the weakest portion on the initial rupture fault and then propagates to a portion of higher strengths

rupture fault, since amplitudes of a longer wavelength are thought to be larger on fractal fault surfaces (e. g., [11]).

Under the above assumptions, the asperity and initial rupture fault are attributed to the portions of higher and lower strengths on the strength profile, respectively. Strength profiles of faults of larger earthquakes are thought to have a longer wavelength. Actually, it is empirically derived that the asperity size is proportional to the total fault length [47]. Consequently, it is inferred from these results that larger earthquakes have larger initial rupture faults. On initial rupture faults, the fault strength increases with distance from the hypocenter, since the hypocenter is the weakest point on the fault. In other words, the strength at a rupture front increases with rupture growth. Thus, it is thought that the initial rupture does not expand smoothly and the slip velocity does not significantly accelerate, but a large asperity can break after a breaking of the initial rupture fault. This is only a possible qualitative model for the initial rupture process and it suggests a possibility for explaining the observed data. It should be examined by extensive studies.

We do not presently have any clear answers to the subtitle question, 'Does the initiation of earthquake rupture knows about its termination?'. Even if the above model is correct, we also have to know the geometry of the fault surface and the time, slip and slip velocity dependent stress on the fault for large and small earthquakes, in order to simulate the rupture propagation and must understand the factors that control the earthquake size.

Future Directions

This paper reviewed studies analyzing the very beginning portions of observed waveforms of earthquakes and showed what we have presently clarified about the earthquake nucleation process. To make further progress, the most straightforward path is to investigate the initial rupture process by precise waveform inversions for large earthquakes. In this case, new inversion methods, as proposed by Uchide and Ide [70], are useful. Furthermore, it is necessary to use broadband near field data, since it is possible that very slow slips occur on the initial rupture fault in association with the initial rupture process or for a very long time before the initial rupture process. For smaller earthquakes, high resolution data, as those in the Western Nagano Prefecture region, are necessary for investigating their rupture processes. These extensive studies about the initial rupture process can clarify the true feature of the earthquake nucleation process.

Acknowledgments

The project in the Western Nagano Prefecture is a co-operative study with Shigeki Horiuchi, Shiro Ohmi, Hisao Ito, Yasuto Kuwahara, Eiji Yamamoto, Kentaro Omura, Koichi Miura, Bun'ichiro Shibazaki, and Haruo Sato. We thank James Mori and Masumi Yamada for their critical reviews of the manuscript. This work is partly supported by JSPS.KAKENHI (19204043), Japan. We are grateful for two anonymous reviewers for their critical and thoughtful comments.

Bibliography

1. Abercrombie R, Mori J (1994) Local observations of the onset of a large earthquake, 28 June 1992. Landers, California. Bull Seismol Soc Am 84:725–734
2. Aki (1967) Scaling law of seismic spectrum. J Geophys Res 72:1217–1231
3. Allen RM, Kanamori H (2003) The potential for earthquake early warning in Southern California. Science 300(5620):786–789
4. Anderson JG, Bodin P, Brune JN, Pince J, Singh SK, Quaas R, Ohnate M (1986) Strong ground motion from the Michoacan, Mexico, earthquake. Science 233:1043–1049
5. Andrews DJ (1976) Rupture velocity of plane strain shear cracks. J Geophys Res 81:5679–5687
6. Asano K, Iwata T (2006) Source process and near-source ground motions of the 2005 West Off Fukuoka Prefecture earthquake. Earth Planet Space 58:93–98
7. Azimi SA, Kalinin AV, Kalinin VV, Pivovarov BL (1968) Impulse and transient characteristics of media with linear quadratic absorption laws, Izv. Earth Phys 1968(2):88–93
8. Bak P, Teng C (1989) Earthquakes as self-organized critical phenomenon. J Geophys Res 94:635-15–637-15
9. Beroza GC, Ellsworth WL (1996) Properties of the seismic nucleation phase. Tectonophysics 261:209–227
10. Boatwright J (1978) Detailed spectral analysis of two small New York State earthquakes. Bull Seismol Soc Am 68:1117–1131

11. Brown SR, Scholz CH (1985) Broad bandwidth study of the topography of natural rock surfaces. J Geophys Res 90:12575–12582
12. Brune JN (1979) Implications of earthquake triggering and rupture propagation for earthquake prediction based on premonitory phenomena. J Geophys Res 84:2195–2198
13. Cheng X, Fenglin Niu, Silver PG, Horiuchi S, Takai K, Ito H, Iio Y Similar microearthquakes observed in western Nagano, Japan and implications to rupture mechanics. J Geophys Res 112:B04306. doi:10.1029/2006JB004416
14. Christensen DH, Ruff LJ (1986) Rupture process of the Chilean earthquake, 3 March 1985. Geophys Res Lett 13:721–724
15. Das S, Scholz CH (1982) Theory of time-dependent rupture in the Earth. J Geophys Res 86:6039–6051
16. Deichman N (1997) Far-field pulse shapes from circular sources with variable rupture velocities. Bull Seismol Soc Am 87:1288–1296
17. Dieterich JH (1978) Preseismic fault slip and earthquake prediction. J Geophys Res 83:3940–3948
18. Dieterich JH (1979) Modelling of rock friction: 1 Experimental results and constitutive equations. J Geophys Res 84:2161–2168
19. Dieterich JH (1986) A model for the nucleation of earthquake slip. In: Earthquake source mechanics. Geophysical Monograph. Maurice Ewing Series, vol 6. Am Geophys Union, Washington DC, pp 37,36–47
20. Dodge DA, Beroza GC, Ellsworth WL (1996) Detailed observations of California foreshock sequences: Implications for the earthquake initiation process. J Geophys Res 101(22):371–392
21. Ellsworth WL, Beroza GC (1995) Seismic evidence for an earthquake nucleation phase. Science 268:851–855
22. Ellsworth WL, Beroza GC (1998) Observation of the seismic nucleation phase in the 1995 Ridgecrest, California sequence. Geophys Res Lett 25:401–404
23. Fukao Y, Furumoto M (1985) Hierarchy in earthquake size distribution. Phys Earth Planet Inter 37:149–168
24. Furumoto M, Nakanishi I (1983) Source times and scaling relations of large earthquakes. J Geophys Res 88:2191–2198
25. Hardebeck JL, Hauksson E (2001) The crustal stress field in southern California and its implications for fault mechanics. J Geophys Res 106(21):859–882
26. Hiramatsu Y, Furumoto M, Nishigami K, Ohmi S (2002) Initial rupture process of microearthquakes recorded by high sampling borehole seismographs at the Nojima fault, central Japan. Phys Earth Planet Inter 132:269–279
27. Hirata M (2003) The initial rupture process of the 2000 Western Tottori Earthquake. Master Thesis, Kyoto University
28. Horikawa H (2006) Rupture process of the 2005 West Off Fukuoka Prefecture, Japan, earthquake. Earth Planet Space 58:87–92
29. Ide S, Beroza GC, Prejean SG, Ellsworth WL (2003) Apparent break in earthquake scaling due to path and site effects on deep borehole recordings. J Geophys Res 108(B5):2271; doi:10.1029/2001JB001617
30. Iio Y (1992) Slow initial phase of the P-wave velocity pulse generated by microearthquakes. Geophys Res Lett 19:477–480
31. Iio Y (1995) Observation of the slow initial phase generated by microearthquakes: Implications for earthquake nucleation and propagation. J Geophys Res 100:15333–15349
32. Iio Y, Ohmi S, Ikeda R, Yamamoto E, Ito H, Sato H, Kuwahara Y, Ohminato T, Shibazaki B, Ando M (1999) Slow initial phase generated by microearthquakes occurred in the Western Nagano prefecture, Japan -the source effect-. Geopys Res Lett 26(13):1969–1972
33. Iio Y, Kobayashi Y, Tada T (2002) Large earthquakes initiate by the acceleration of slips on the downward extensions of seismogenic faults, Earth Planet. Sci Lett 202:337–343
34. Iio Y, Horiuchi S, Ohmi S, Ito H, Kuwahara Y, Yamamoto E, Omura K, Miura K, Shibazaki B, Sato H (2006) Slow initial phase of microearthquakes. Program and abstracts of 2006 fall meeting of the Seismological Society of Japan, A48 (in Japanese)
35. Ishihara Y, Fukao Y, Yamada I, Aoki H (1992) Rising slope of moment rate functions: the 1989 earthquakes off east coast of Honshu. Geophys Res Lett 19:873–876
36. Ito S (2003) Study for the initial rupture process of microearthquakes in western Nagano, central Japan, estimated from seismograms recorded in three boreholes. Ph D Thesis, Tohoku University (in Japanese)
37. Ito S, Ito H, Horiuchi S, Iio Y (2004) Local attenuation in western Nagano, central Japan, estimated from seismograms recorded in three boreholes. Geophys Res Lett 31:L20604; doi:10.1029/2004GL020745
38. Ito Y, Obara K, Takeda T, Shiomi K, Matsumoto T, Sekiguchi S, Hori S (2006) Initial-rupture fault, main-shock fault, and aftershock faults: Fault geometry and bends inferred from centroid moment tensor inversion of the 2005 West Off Fukuoka Prefecture earthquake. Earth Planet Space 58:69–74
39. Iwata T, Sekiguchi H (2002) Source process and near-source ground motion during the 2000 Tottori-ken Seibu earthquake (M_w 6.8). Reports on Assessments of Seismic local-site effects at plural test sites. MEXT, pp 231–241
40. Kanamori H, Anderson DL (1975) Theoretical bases for some empirical relations in seismology. Bull Seism Soc Am 65:1073–1095
41. Kanamori H (1996) Initiation process of earthquakes and its implications for seismic hazard reduction strategy. Proc Natl Acad Sci 93:3726–3731
42. Kawanishi R, Iio Y, Yukutake Y, Katao H, Shibutani T (2006) Estimate of the stress field in the region of the 2000 Western Tottori earthquake. Program and abstracts of 2006 fall meeting of the Seismological Society of Japan, P099 (in Japanese)
43. Kilb D, Gomberg J (1999) The initial subevent of the 1994 Northridge, California, Earthquake – is earthquake size predictable? J Seismol 3:409–420
44. Lay T, Kanamori H, Ruff L (1982) The asperity model and the nature of large subduction zone earthquakes. Earthq Predict Res 1:3–71
45. Mandelbrot BB (1982) The fractal geometry of nature. W.H. Freeman, New York
46. Miura K, Iio Y, Yukutake Y, Takai K, Horiuchi S (2005) The feature of initial motion for waveforms of microearthquakes in Western Nagano, Japan. Program and abstracts of 2005 fall meeting of the Seismological Society of Japan, P103. (in Japanese)
47. Miyake H, Iwata T, Irikura K (2003) Source characterization for broadband ground motion simulation: Kinematic heterogeneous source model and strong motion generation area. Bull Seism Soc Am 93:2531–2545
48. Mori J, Kanamori H (1996) Initial rupture of earthquake in the 1995 Ridgecrest, California sequence. Geophys Res Lett 23:2437–2440
49. Nakamura Y (1988) Proc World Conference on Earthquake Engineering, VII, 6763

50. Nakatani M, Kaneshima S, Fukao Y (2000) Size-dependent microearthquake initiation inferred from high-gain and low-noise observations at Nikko district, Japan. J Geophys Res 105(B12):28095–28110; doi:10.1029/2000JB900255
51. Ohmi S, Watanabe K, Shibutani T, Hirano N, Nakao S (2002) The 2000 Western Tottori Earthquake—Seismic activity revealed by the regional seismic networks. Earth Planet Space 54:819–830
52. Ohnaka M, Kuwahara Y, Yamamoto K, Hirasawa T (1986) Dynamic breakdown processes and the generating mechanism for high-frequency elastic radiation during stick-slip instabilities. In: Das S, Boatwright J, Scholz CH, AGU (eds) Earthquake source mechanics. Geophysical Monograph, vol 37. Maurice Ewing Series, vol 6. American Geophysical Union, Washington DC, pp 13–24
53. Ohnaka M, Kuwahara Y (1990) Characteristic features of local breakdown near a crack-tip in the transition zone from nucleation to unstable rupture during stick-slip shear failure. Tectonophysics 175:197–220
54. Ohnaka M, Shen L (1999) Scaling of the shear rupture process from nucleation to dynamic propagation: Implications of geometric irregularity of the rupture surfaces. J Geophys Res 104:817–844
55. Ohnaka M (2000) A physical scaling relation between the size of an earthquake and its nucleation zone size. Pure Appl Geophys 157:2259–2282
56. Okubo PG, Dieterich JH (1984) Effects of physical fault properties on frictional instabilities produced on a simulated faults. J Geophys Res 89:5817–5827
57. Olson EL, Allen RM (2006) Is earthquake rupture deterministic? Nature 442:E5–E6; doi:10.1038/nature04963
58. Rydelek P, Horiuchi S (2006) Is earthquake rupture deterministic? (Reply). Nature 442:E6; doi:10.1038/nature04964
59. Sato T (1994) Seismic radiation from circular cracks growing at variable rupture velocity. Bull Seismol Soc Am 84:1199–1215
60. Sato T, Hirasawa T (1973) Body wave spectra from propagating shear cracks. J Phys Earth 21:415–431
61. Sato T, Kanamori H (1999) Beginning of earthquakes modeled with the Griffith's fracture criterion. Bull Seismol Soc Am 89:80–93
62. Sato K, Mori J (2006) Scaling relationship of initiations for moderate to large earthquakes. J Geophys Res 111:B05306; doi:10.1029/2005JB003613
63. Sato K, Mori J (2006) Relation between rupture complexity and earthquake size for two shallow earthquake sequences in Japan. J Geophys Res 10.1029/2005JB003613
64. Shibazaki B, Matsu'ura M (1992) Spontaneous processes for nucleation, dynamic propagation, and stop of earthquake rupture. Geophys Res Lett 19:1189–1192
65. Shibazaki B, Matsu'ura M (1995) Foreshocks and pre-events associated with the nucleation of large earthquakes. Geophys Res Lett 22(10):1305–1308; doi:10.1029/95GL01196
66. Shibazaki B, Matsu'ura M (1998) Transition process from nucleation to high-speed rupture propagation: Scaling from stick-slip experiments to natural earthquakes. Geophys J Int 132:14–30
67. Shibazaki B, Yoshida Y, Nakamura M, Nakamura M, Katao H (2002) Rupture nucleations in the 1995 Hyogo-ken Nanbu earthquake and its large aftershocks. Geophys J Int 149:572–588
68. Spudich P, Cranswick E (1984) Direct observation of rupture propagation during the 1979 Imperial Valley earthquake using a short-baseline accelerometer array. Bull Seismol Soc Am 74:2083–2114
69. Takenaka H, Nakamura T, Yamamoto Y, Toyokuni G, Kawase H (2006) Precise location of the fault plane and the onset of the main rupture of the 2005 West Off Fukuoka Prefecture earthquake. Earth Planets Space 58:75–80
70. Uchide T, Ide S (2007) Development of multiscale slip inversion method and its application to the 2004 Mid-Niigata Prefecture earthquake. J Geophys Res doi:10.1029/2006JB004528
71. Uehira K, Yamada T, Shinohara M, Nakahigashi K, Miyamachi H, Iio Y, Okada T, Takahashi H, Matsuwo N, Uchida K, Kanazawa T, Shimizu H (2006) Precise aftershock distribution of the 2005 West Off Fukuoka Prefecture Earthquake ($Mj = 7.0$) using a dense onshore and offshore seismic network. Earth Planet Space 58:1605–1610
72. Umeda Y (1990) High-amplitude seismic waves radiated from the bright spot of an earthquake. Tectonophysics 175:81–92
73. Umeda Y (1992) The bright spot of an earthquake. Tectonophysics 211:13–22
74. Umeda Y, Yamashita T, Tada T, Kame N (1996) Possible mechanisms of dynamic nucleation and arresting of shallow earthquake faulting. Tectonophysics 261:179–192
75. Yamaguchi S, H Kawakata, T Adachi, Y Umeda (2007) Features of initial process of rupture for the 2005 West off Fukuoka Prefecture Eathquake. Zisin Ser 2:241–252 (in Japanese)
76. Venkataraman A, Beroza GC, Ide S, Imanishi K, Ito H, Iio Y (2006) Measurements of spectral similarity for microearthquakes in western Nagano, Japan. J Geophys Res 111:B03303; doi:10.1029/2005JB003834
77. Wu Y, Kanamori H, Allen R, Hauksson E (2007) Determination of earthquake early warning parameters, τ_c and P_d, for southern California. Geophys J Int (OnlineEarly Articles) doi:10.1111/j.1365-246X.2007.03430.x
78. Wyss M, Brune J (1967) The Alaska earthquake of 28 March 1964—a complex multiple rupture. Bull Seismol Soc Am 57:1017–1023

Earthquake Occurrence and Mechanisms, Stochastic Models for

David Vere-Jones
Statistical Research Associates and Victoria University, Wellington, New Zealand

Article Outline

Glossary
Definition of the Subject
Introduction
Historical Overview
Stochastic Models for Earthquake Mechanisms
Models for Paleoseismological and Historical Earthquakes
Point Process Models for Regional Catalogues

Stochastic Models with Precursors
Further Topics
Future Directions
Acknowledgments
Bibliography

Glossary

Stochastic occurring by chance;

Stochastic process physical or other process evolving in time governed in part by chance.

Earthquake mechanism physical processes causing the occurrence of an earthquake.

Independent events events not affecting each other's probability of occurrence.

Branching process process of ancestors and offspring, as in the model of nuclear fission.

Point process stochastic process of point-events in time or space.

Probability forecast prediction of the probability distribution of the time and other features of some future event, as distinct from a forecast for the time (etc.) of the event itself.

Model test a statistical test for the extent to which a stochastic model is supported by the relevant data.

Precursory signal observed quantity which affects the occurrence probability of a future event (earthquake).

Definition of the Subject

Stochastic models for earthquake mechanism and occurrence combine a model for the physical processes generating the observable data (catalog data) with a model for the errors, or uncertainties, in our ability to predict those observables. Such models are essential to properly quantify the uncertainties in the model, and to develop probability forecasts. They also help to isolate those features of earthquake mechanism and occurrence which can be attributed to mass action effects of a statistical mechanical character. We do not consider in this paper applications of the models to earthquake engineering and insurance.

Introduction

The complexity of earthquake phenomena, the difficulty of understanding and monitoring the processes involved in their occurrence, and the consequent difficulty of accurately predicting them, are now widely accepted points of view. What are stochastic models, and what role do they play in aiding our understanding of such phenomena?

The present article is an attempt to address these questions. We start from the beginnings, the distinction between stochastic and deterministic models, and the first attempts to model earthquake phenomena in stochastic or statistical terms. We then follow through with a systematic account of some of the main classes of stochastic models that are currently in use, discussing in turn earthquake mechanisms, historical earthquakes, regional catalogs, descriptive patterns, and earthquake precursors.

The focus throughout is on the stochastic modeling aspects, rather than on statistical procedures or associated algorithms. As a result we have given only brief mention to pattern-recognition techniques, or to descriptive procedures such as the estimation of fractal dimensions or of second order properties, which do not lead to fully defined models. Again, although a primary use of stochastic models is in developing probability forecasts, we have limited ourselves to the briefest account of how such forecasts can be produced and assessed. Nor do we directly consider applications to engineering and insurance problems.

The fundamental difference between a physical model and a stochastic model, in broad terms, is that while the physical model attempts to fully describe and predict the process under study, the stochastic model treats some aspects of the physical process as out of range of exact modeling, at least for the time being, and replaces it by some unpredictable and hence random process. The resulting stochastic model should reproduce those aspects of the physical phenomenon which are important and accessible to measurement, but may relegate the rest to dice-tossing or one of its more contemporary avatars such as Brownian motion or the Poisson process.

Across their many different fields of application, two broad roles for stochastic models may be distinguished. The first is epitomized by statistical mechanics. Here the stochastic model plays an integral role in understanding the physical processes themselves. The macroscopic phenomena that we are able to observe directly – temperature, pressure and the like – are shown to be a consequence, not of the details of the collision processes at the microscopic level, but of their mass interactions, which are governed largely by laws of an essentially statistical character such as the law of large numbers or the central limit theorem. For predicting the macroscopic behavior, it is not necessary to know the details of the complex interactions between individual molecules; it is sufficient to replace them by a simple random process that nonetheless preserves the crucial physical aspects such as mean velocities and angular distributions.

Within seismology such a role is implicit when the fracture processes within the earth's crust are compared to

'frozen turbulence', or in applications of branching process or percolation theory to explain energy distributions, or in discussions of the fracture strength of materials as functions of the density and size distribution of microcracks, or in the use of cellular automata and similar models to explain the appearance of long-range correlations and power-law distributions in the approach to criticality of certain types of complex systems.

In the other, by far more common, type of application, the stochastic model is used as a basis for planning and prediction. In such situations it is vital to know, not just a forecast value, but also something about the reliability of that value. It is also vitally important that the models can be fully fitted to the observable data. Most branches of applied statistics have evolved in response to such requirements. Within seismology, applied models of this type are needed in discussions of earthquake risk for insurance or building codes, in many parts of engineering seismology, and in the development of decision rules for earthquake response and emergency planning. Probability forecasts of any kind, including all forecasts with some associated estimate of precision, necessarily rely on stochastic models of this kind.

Many decades ago, the famous geophysicist and seismologist Sir Harold Jeffreys, who is also regarded as a pioneer in inferential statistics, argued that, to be worthy of its name, every physical theory should contain within itself the means not only of predicting the relevant quantities, but also of predicting their uncertainties [45]. In our terminology, he was arguing that every physical theory should be based on a stochastic model. In the classical studies of physics and astronomy, the uncertainties in the model are assumed to be due to nothing deeper than observational errors. In a subject such as seismology, however, the uncertainties are much more fundamental.

While general patterns of earthquake behavior may be predicted from physical theories, the predictions do not extend to the times and locations of individual earthquakes. Moreover, the available observational data are rarely more than indirectly relevant to the physical processes controlling the details of earthquake occurrence, as these usually take place many kilometers beneath the surface of the earth, and out of range of direct observation.

Stochastic models of earthquake occurrence that can be used for earthquake prediction must somehow marry the limited physical theory to the limited available data. Attempts to grapple with this central problem have intensified in recent years. They form one factor in the emergence of 'Statistical Seismology' as a new sub-discipline. Another, perhaps dominating, factor, is the enormous improvement in both the quantity and quality of the data that are available, whether from earthquake catalogs, from GPS measurements of ground deformation, or, less commonly, from data on auxiliary quantities such as well levels, electrical signals, ionospheric depression and others thought to have a potential bearing on earthquake occurrence. The high quality data demand a comparable quality in the statistical modeling and analysis.

Historical Overview

The forerunner of any serious statistical modeling is the availability of reliable and relevant data. For models of earthquake occurrence this means the availability of good quality earthquake catalogs. Broadly speaking, such catalogs had to wait, not only until around the turn of the 20th century, when the first instrumental records became available, but until the development of modern instrumentation and the extensive station networks which came into being after the second World War. Before then, the lack of any consistent measure of the size of an earthquake, and the general unevenness of network coverage, made the records of limited value for statistical purposes.

A turning point was the appearance of the first edition of the classic text [28] by Gutenberg and Richter (1949). For the first time it gave a comprehensive overview of the major features of the seismicity of the earth, and of the key empirical relations governing earthquake occurrence. From that time onwards, the way was open for serious statistical analysis, although recent data is far more comprehensive and detailed. Modern instrumental catalogs, prepared from digital records telemetered to a local center from a dense network of stations, may contain hundreds of thousands of events down to very small magnitudes. Typically such catalogs list for each event the origin time (initiation of rupture), epicenter (latitude and longitude of place of first motion), depth, magnitude or seismic moment (alternative measures of earthquake size), and often other parameters relating to the fault mechanism (orientation of the fault and direction of first motion).

The availability of these high-quality catalogs, alongside the increasing availability of data from GPS measurements and other earthquake-related phenomena, is a key reason for the recent upsurge of interest and research in statistical seismology. The broad aims of this emerging field may be described as finding statistical models to describe and make use of such data, and to marry it with the existing physical theory.

Even preinstrumental catalogs inspired the investigation of two statistical issues at least: does the occurrence

of major earthquakes exhibit some form of periodicity in time? do the numbers of events in time intervals of fixed length follow a Poisson distribution? We comment briefly on these two questions before looking at statistical models more generally.

Periodicity of Earthquakes

Periodicity of earthquakes, in some more or less regular sense, was the earliest issue to be investigated, and inspired many early studies, including one of Schuster's classic papers on the periodogram [112]. Until the 1930s, however, neither the data nor the statistical techniques were sufficiently developed to allow the question to be properly resolved. On the statistical side, for instance, the periodogram was a new concept, and statistical tests based on it were still in development. Schuster's paper, applied to a special case, contains within itself all the basic elements of point process spectral theory, starting from the finite Fourier transform, calculating the equivalent of rough significance levels using Rayleigh's random flights, and briefly treating the problems caused by binning the data and by clustering. Jeffreys [44] was one of the first to use a modern statistical approach to tackle the question, while Davison [21] reviewed many of the earlier studies and came to the conclusion that most of those studies were inconclusive.

The topic remains controversial, although it is now clear that no obvious periodicities exist. The most important current contenders for small-scale periodicities are in relation to earth tides. It is suggested that the small fluctuations in crustal stress due to the relative movements of the moon and sun around the earth may be large enough to trigger earthquake activity under favorable conditions, for example in regions already under high stress. A careful recent study of lunar tides on microearthquakes, with further references, is given in [40]. The possibility of using the response of small-scale seismicity to lunar tides as a possible indicator of regions in some near-critical state, and hence as a precursor for larger events, has been suggested in [139]; a statistical analysis is given in [142].

The possibility of long-term periodicities, of the order of decades or centuries, is unclear because of the shortage of data; substantial fluctuations certainly exist. The problem is still full of potential traps. In testing for periodic effects, for example, it is essential to take into account earthquake clustering, and whether or not the period being tested for is preassigned (as for the lunar cycle) or suggested by the data. Both of these issues are illustrated in the discussion in [134] of Kawasumi's historical data for large earthquakes in the Kanto region of Japan.

The Poisson Distribution and Process

The distribution

$$p_n = (\mu^n/n!)\mathrm{e}^{-\mu}\,, \quad \mu > 0\,, \quad n \geq 0\,,$$

was introduced by Poisson as an approximation to the binomial distribution when the number of trials N becomes very large but the probability p of success becomes very small, the two balancing in such a way in such a way that the expected number $\mu = Np$ of successes remains moderate in size.

Earthquakes were included among the examples studied by von Bortkiewicz [136] in his 1898 compilation of phenomena to which he could apply 'the law of small numbers', the name he gave to the Poisson approximation to the binomial. The question was studied in greater depth by later writers, including Gutenberg and Richter [28], and several important qualifications were noted. In the first instance, the disturbing effect of aftershocks was pointed out, and so the Poisson distribution was supposed to apply just to main shocks. Then other disturbing effects, such as trends and longer-term fluctuations in activity, were noted. In fact almost no catalog fits the Poisson description exactly, and for research purposes its role as a base-line model for 'standard seismicity' has been replaced by the ETAS model (see Sect. "The ETAS Model"), which provides a much better approximation to the clustering properties of smaller earthquakes.

Nevertheless the simple Poisson form is still the principal basis for determining earthquake risk and for earthquake insurance practices. Underlying its continued relevance is the same idea underlying Poisson's original approximation to the binomial: when the data under examination consists of rare 'successes' from many different and essentially unrelated sources, the Poisson distribution generally emerges as a good approximation.

It is necessary to distinguish between the *Poisson distribution* and the *Poisson process*. The Poisson process refers to an evolutionary model for the occurrence of events in time or space or both. Its principal characteristics, at least in the stationary case, are

1. The number of events within any bounded region (interval in time; area or volume in space) follows a Poisson distribution with parameter μ proportional to the size (length, area etc) of the region selected for study;
2. The numbers of events in disjoint regions are independent random variables.

The second condition embodies the famous 'lack of memory' property of the Poisson process: the temporal version asserts that the occurrence of one or more events before

a certain time has no effect on the occurrence probabilities of subsequent events. It dictates the exponential form of the distribution of the time interval between events, and under simple conditions even dictates the form of the Poisson distribution itself; see, for example, the discussion in Chap. 2 of [20].

The Empirical Laws of Seismology

The advent of more complete and reliable catalogs saw the recognition of a number of statistical regularities in the occurrence of earthquakes. Two of these are central features of seismicity studies today.

Omori's Law Already by the end of the 19th century, the Japanese pioneer seismologist Omori had made detailed studies of some large Japanese aftershock sequences [92], and suggested that the frequency of aftershock occurrence, say $\lambda(\tau)$, decayed approximately hyperbolically with the time τ after the main event:

$$\lambda(\tau) \approx A/\tau$$

where A is a constant characteristic of the particular mainshock and associated with its size. His own and subsequent studies suggested the need for refinements, and the most widely accepted form today is the the Omori–Utsu formula

$$\lambda(\tau) = A/(c+\tau)^p \tag{1}$$

where the parameters A, c, p are again peculiar to the individual aftershock sequence, c is generally small (of the order of seconds to days) and p is close to 1. A detailed study of the history and other issues associated with the Omori law over the 100 years 1894–1994 is given in [124].

The simplest stochastic model for aftershocks is that suggested by Jeffreys in [44], namely an inhomogeneous Poisson process in which $\lambda(\tau)$ is interpreted as the current value of the time-varying Poisson intensity; the independence property (2) of the Poisson process of Sect. "The Poisson Distribution and Process" is retained, but the mean parameter for the number of events in (s, t) is now $\mu = \int_s^t \lambda(\tau)\mathrm{d}\tau$.

The Gutenberg–Richter (GR) Law The law was formulated after the definition of earthquake magnitude gave an objective method of quantifying the size of an earthquake. It is a basic component of [28], although a similar relationship, based on the more qualitative maximum intensity concept, had been formulated somewhat earlier by Ishimoto and Iida for Japanese earthquakes [39].

The GR law provides a summary of the magnitude data in a catalog of earthquakes with magnitudes complete above a certain threshold, say M_0. It is commonly written in the form

$$\{\text{Number of events above magnitude M}\} \approx 10^{a-b(M-M_0)} \tag{2}$$

or equivalently

$$\{\text{Proportion of events above Magnitude M}\} = 10^{-b(M-M_0)}\,. \tag{3}$$

It is a pity in our view that the former rather than the latter of these two forms has become traditional. The danger then is that a becomes treated as a separate parameter instead of as a normalizing constant, $a = \log_{10} N$, where N is the total number of events under consideration. One reason for this tradition may have been the common (and incorrect) use of ordinary least squares methods to compute the line of best fit from a graph of the binned numbers. Such an approach will certainly produce a slope as one of the parameters, but the estimate is unstable and distorts the interpretation unless it is especially modified to fit the distribution function context.

The second form makes it clear that what we are looking at is an empirical probability distribution, and that the right hand side could equally and more appropriately be written in the form

$$10^{-b(M-M_0)} = \mathrm{e}^{-\beta(M-M_0)}$$

where $\beta = b\log_{\mathrm{e}} 10 \approx 2.3b$. Then it is clear that the G-R law asserts that, under suitable conditions, the empirical distribution of magnitudes is approximately exponential.

In principle, a could be regarded as a parameter in an extended model for the space-time-magnitude distribution of events in a given space-time window, but such an interpretation is rarely given.

How valid the exponential distribution remains when examined in greater detail, and whether, and if so by what, it should be replaced for general modeling purposes, is still a subject of debate. The main reservation relates to the possibility of extremely large events, which is physically unreasonable and can lead to misleading conclusions if used in simulation studies of long-term behavior.

Of the many alternatives offered, which include truncated and multi-parameter versions (see [123] for a listing and software), perhaps the most plausible is the 'tapered Pareto distribution', or 'Kagan distribution' (e. g. [48, 135]), with distribution function written out in terms of

seismic moments or energies as (4) below. This distribution arises in branching and similar models for crack propagation (see [123]), and is derived from maximum-entropy considerations in [73]; a further derivation from a critical phase transition in a finite elastic solid is given in [26].

The use of magnitude itself as a basic variable is also open to question. It is not a uniquely or tightly defined quantity. In terms of quantities such as energies or seismic moments with a more direct physical interpretation, the exponential distribution for magnitudes becomes a Pareto (inverse power-law) distribution,

$$\Pr\{E > x\} = (x/x_0)^{-\alpha} , \quad (x > x_0)$$

through a transformation of the form

$$\log_{10}(E) \approx 1.5M + \text{const}\, n ,$$

where E is the energy. This illustrates the fact that the magnitude scale is essentially a decibel scale, ultimately a consequence of its initial definition in terms of the logarithm of the maximum amplitude of the trace on a seismograph. The tapered Pareto form mentioned earlier is

$$\Pr\{X > x\} \approx Cx^{-\alpha} \mathrm{e}^{-\gamma x} , \quad (x \geq x_0) , \tag{4}$$

where C is a normalizing constant, or the variant with a similar form for the density.

Båth's Law The so-called Båth's Law asserts, loosely, that in an aftershock sequence, the difference between the magnitude of the mainshock and that of the largest aftershock is around 1.2 magnitude units.

Although noted by Båth in 1965, and even earlier by Utsu, this regularity has never enjoyed quite the same status as the other two laws. The question, still an active topic of debate, is whether it represents a physical phenomenon in its own right, or is merely a consequence of the more general properties of earthquake clustering. It was suggested in [125] (see also the reviews and more extensive studies in [17,63]) that it might be simply a consequence of the statistical properties of the largest and second largest in a sequence of events following the G-R law. More recently, it has been linked to the 'productivity function' of earthquake clustering: the expected number of aftershocks increases typically as an exponential function $K\mathrm{e}^{\alpha M}$ of the magnitude of the main shock, with Båth's Law resulting when the exponent α equals the exponent β in the GR law (see [25]). This suggestion is supported by the appearance of a Båth's law phenomenon in the ETAS model, where there is certainly no explicit model feature relating to Båth's law, but there is an exponential productivity function [38].

Stochastic Models for Earthquake Mechanisms

General Considerations

The earliest model for earthquake mechanism is Reid's elastic rebound model [98]. It was inspired by studies of large-scale earthquakes, in particular the famous San Francisco earthquake of 1906. The upper part of the crust is deformed elastically by large scale tectonic motions, then ruptures and rebounds when its breaking strength is reached, resulting in an earthquake.

The many attempts to to marry this physical picture with simple stochastic ideas lead typically to models based broadly on the renewal process, and will be picked up in the discussion in Sect. "Background and Data".

In this section we look rather at models which describe the behavior at the microscopic level, the evolution of the fracture itself, and can be used to explain the basic empirical laws, among other features.

There are strong links with theories on the strength of materials, starting from the classic studies of Griffiths (e. g. [27]) on crack extension in brittle materials, and the role of microfractures in controlling the fracture strength of glass. Griffiths' crack theory is basic to models of fracture in brittle materials, whether at the scale of rock fracture in laboratory specimens or fault propagation in the earth's crust (see e. g. [42,82,109,110]).

Griffiths' ideas were later developed by Weibull [136] into a model explaining the variations in strength of otherwise similar specimens of rock and many other substances. Weibull supposed that the underlying cause was the random distribution of microfracture lengths in the specimen, and used an argument based on the distribution of the length of the largest such microfracture to deduce a form for the distribution of strengths. Indeed it is from these studies that the 'Weibull distribution' takes its name.

The branching process, percolation, and cellular automata interpretations of the earthquake process start from similar general premises. The underlying idea is that, instead of progressing smoothly, as might a fault or fracture in a homogeneous elastic medium, the progress of the fault in a medium containing many weaknesses is controlled by the essentially random locations of these weaknesses. The various models which have been proposed differ mainly in the assumptions governing these random locations.

In Otsuka's original 'go-game' model [93], points were laid down on a lattice in much the same way as in the game of 'Go', but at random, using a simulation technique, with the interpretation that the enclosed pieces determined a rupture area. In [108] this was idealized into

a model linking weaknesses located on a Bethe lattice, where every node has one input link and the same fixed number of outward links, and each node may or may not be a point of weakness.

Otsuka's model has both branching process and percolation model interpretations, with a considerable literature surrounding extensions of both. There also links to other, apparently more deterministic, approaches to the generation of the empirical laws, for example through block slider models or in the general class of complex systems. Although the models differ in approach and detail, the size distributions and the like often turn out to be very similar to those derived from the branching models. As in statistical mechanics, the properties have their origin in the mass interactions of many small components, and are relatively insensitive to the details at the microscopic level. The simpler statistical models, such as the branching models, allow these distributions to be explored directly by analytical means. In complex system theory the aim is rather to show how similar results arise from approximating the deterministic equations governing large families of interacting bodies.

Branching Models

The conceptual framework here is that the crack initiates from an initial weakness (dislocation or microfracture) and spreads to one or more others, or terminates, the 'others' being interpreted as 'offspring' and the initial weakness as the 'ancestor'. Each 'other' then acts as an ancestor in its own right, and the process continues until either all branches have died out (subcritical and critical cases) or the process explodes (supercricital state). The behavior is controlled by a 'criticality parameter' ρ, effectively the mean number of offspring per ancestor. The subcritical, critical, and supercritical cases correspond respectively to $\rho < 1$, $\rho = 1$ and $\rho > 1$.

This model was developed in general form in [123], following [109] and the earlier work on the 'go-game' model in Japan. Related ideas occur in many places papers by Kagan and Kagan and Knopoff, see especially the extended branching model described in [54].

The distribution of the size or energy release of the rupture is then obtained by counting the total number of offspring before the process dies out (critical or subcritical cases). The remarkable feature here is that even when the individual offspring distributions are very regular, the total size distribution approaches a power-law (Pareto) form whose basic parameters are independent of the details of the offspring distribution. In the limiting critical case, the power-law distribution for sizes has $\Pr\{N > n\} \sim Cn^{-1/2}$, corresponding roughly, assuming equal energies/event on average, to a G-R law with $b \approx 0.75$.

When the process is just subcritical the Pareto distribution becomes a tapered Pareto distribution, with power-law behavior for moderate to large events, and an exponential tail-off at high magnitudes which cuts in at a point determined by the distance from criticality, $\delta = 1 - \rho$. Again the behavior is otherwise largely independent of the details of the offspring distribution.

Many further developments and ramifications of this underlying model have been proposed. One of the deepest is the simulation model for earthquakes developed in [54], starting from the scale of dislocations or other defects in the rock fabric, and incorporating temporal, directional, and distance factors into the model evolution to develop an impressive array of properties akin to those of real earthquakes. The model still awaits a full analytical treatment.

It is also remarkable that a branching model underlies one of the most successful models for earthquake occurrence at the regional level, namely the ETAS model described in Sect. "The ETAS Model". The fact that the same mechanism seems implicated at both levels lends plausibility to Kagan's conjecture that the physical process is one and the same at all scales, and that our attempts to decompose it into elements at the fracture formation and inter-fracture stages are more a result of our perceptions and measuring instruments than they are of the underlying physical processes.

Percolation Models

The classical percolation model starts from a two- or three-dimensional lattice, the sites (or alternatively the bonds between lattice points) being randomly and independently labeled 'open' or 'closed' with a fixed probability p and its complement $1 - p$. A crack initiated at an open site links up all contiguous open sites until it can spread no further. In both cases, a critical regime, characterized by a critical value of the probability p, marks the transition between subcritical (small finite events only) and supercritical (infinite or explosive events) regimes. As with the branching models, it is assumed that the crust is generally in or just below the critical state.

An underlying difficulty is that the available observational data are insufficient to provide any easy control over the best interpretation. As with the branching process model again, the percolation models lead to forms of the G-R law, and with additional features can often be extended to cover aftershock phenomena. [15], and [65,66,67] are among the many papers which discuss

and develop these ideas. [5] highlights some of the difficulties of interpretation.

Percolation processes are extensively used in statistical physics to model phase transitions, and their appearance here invites an interpretation of fracture as a phenomenon analogous in some ways to a phase transition. Ideas from the phase transition context that have been transferred to both earthquakes and fracture mechanics include especially features characteristic of the approach to the critical conditions required for the occurrence of a phase transition: the development of long-range correlations, the appearance of power-law or fractal distributions, and approximate self-similarity. Many authors have sought to develop these analogies, often using analogue or simulation models, and attempted to use the appearance of different interaction ranges to identify the approach to near-critical stress conditions in the crust. See [9,10,26,121], as well as the papers cited above, for further references and discussion of such ideas.

Cellular Automata and Self-Organizing Criticality

A third type of model with a similar general role is the cellular automaton, with the distinction that the application here is not to a single faulting or fracture episode, but to a whole network of interacting faults. The simple basic form, first applied to the earthquake context by Bak and Tang [3], again relates to a two-dimensional lattice model. With each point of the lattice is associated a certain integer stress or force, say Z_{ij} for points on a 2-dimensional lattice $\{i, j\}$. The external force (the 'immigrants' in this context) is manifested through the addition of unit force to a site chosen at random through the lattice or on its boundary. When the force exceeds a certain critical value Z_c on a given site, a 'microfracture' occurs, and single units of force are transferred to each of the four directly adjacent sites, while four units are subtracted from the force at the original site. Such transfers may overload one or more of the adjacent sites, which then in turn transfer units of stress to their neighbors (including possibly the initial site), and so on until the system is at rest. Then another unit is added in and a further redistribution of stress takes place. The whole episode is interpreted as an earthquake, and the total number of steps in the episode is taken as proportional to the energy of the earthquake.

The process as a whole is said to exhibit 'self-organizing (or 'self-organized') criticality'. Even if the process is started from a situation where the forces are set to zero at all sites, they will gradually build up, first to the stage where small individual episodes take place, then, as more and more sites approach the the critical value of stress, the episodes become larger, until a stochastically stationary state is reached where the input of stress units is just balanced by the loss of stress units from points on the boundary of the region. So long as the region under consideration is sufficiently large, a process reaching the critical regime exhibits many of the features already indicated as characteristic of the approach to a phase-change: a G-R relation, long-range correlation effects, and (with some elaborations) an Omori-type phenomenon for aftershock sequences.

Models for Paleoseismological and Historical Earthquakes

Background and Data

We move now to models designed for use with data on earthquake occurrences. These generally belong to the second class of stochastic models referred to in the introduction. They should be able to be fitted to real catalog data; simulations from them should mimic real catalogs; and they should be useful in real applications, capable in particular of generating probability forecasts.

We have grouped the models into two main types, those developed to model large earthquakes on historical or even geological time scales, and those developed for use with modern instrumental catalogs, where smaller events are included. The main difference between the two types of model is their treatment of clustering; this is largely ignored in models of the first type, but plays a central role for models of the second type. Models of the first type are considered in the present section, models of the second type in Sect. "Point Process Models for Regional Catalogs".

The distinction between the groups is associated with one of the longest and still unresolved debates over earthquake mechanism, namely the validity of the *characteristic earthquake hypothesis*. Crudely stated, this asserts that, for any given fault or fault segment, there exists an earthquake of approximately fixed magnitude, which is determined by the physical attributes of the fault, and repeats itself after time intervals of approximately fixed length. Since faults occur over a very wide range of sizes (themselves having a power law or Pareto distribution), this does not contradict, but rather suggests a different origin for, the GR distribution.

The empirical evidence for such a hypothesis is equivocal. Its main support comes from the paleoseismological studies on repeated events along a single fault (e. g. [113]), but the data from such studies is usually so limited that it is hard to accept the evidence as conclusive. Other supporting evidence, again observed sometimes but not always, is the occurrence of a hump, corresponding to repeating

earthquakes with similar magnitudes, in the frequency-magnitude distribution for selected regions. For large regions (scale of major faults), this may occur around magnitudes 6–7, suggesting that the larger events from that region occur more regularly (with higher relative frequency) than would be expected from the GR model. One difficulty with the hypothesis is that several of the best-known sequences, such as the Parkfield earthquake sequence, ultimately deviate from the prescribed regularity. Studies of microearthquakes suggest that in some circumstances similar-sized small events may repeat themselves several times in almost identical locations.

Such results suggest that no simple, universal mode of behavior is likely to be found in earthquakes from particular fault structures. Indeed, recent studies by Ben-Zion and colleagues (see [7,8,9,10] and further references therein) have emphasized the possible role played by evolving heterogeneities and damage rheology in the occurrence patterns on a fault system, and have suggested that, according to their ages and past histories, some faults may exhibit characteristic earthquake behavior while others exhibit GR behavior and others again may alternate between the two.

For paleoseismic studies, each data point is extracted with effort from trenching along fault traces or similar exercises. Moreover, only the largest events leave traces identifiable over thousands of years or longer, while estimating magnitudes and other characteristics is at best informed guess-work. Even the dates, usually determined from some form of radio-carbon or other isotope-based method, can be subject to substantial errors.

Similarly in historical studies, such as Ambraseys' history of Persian earthquakes [1], only the largest events affecting a given region or territory are likely to appear sufficiently prominently in the historical records to allow the size and epicenter of the earthquake to be estimated even roughly. Periods of civil war, famine, foreign invasion and bureaucratic neglect leave gaps and further uncertainties which are difficult if not impossible to resolve.

Despite such difficulties, these data provide the only records we have of seismic activity over periods stretching backwards in time beyond the last hundred years or so, and are worthy of the most serious attempts to collect and interpret.

We consider a sequence of three models, starting from the simple renewal model, then considering variants more closely linked to the elastic rebound model.

Renewal Models

With paleological data in particular, attention is generally focused on major events along a single fault, where magnitudes are poorly constrained, and the stochastic elements are introduced primarily to describe, and if possible predict, the time intervals between events.

For a renewal process, magnitudes are neglected, and it is assumed that the successive intervals are independent, both of each other and of other processes, with a common distribution. The independence assumptions are questionable, but with no further information available, this is at least a reasonable starting model.

Let $f(x)$ denote the density and $F(x)$ the distribution function of the common interval distribution. If the observation record, over $(0, T)$, say, comprises an interval of length ℓ_0 to the first recorded event, then n complete intervals $\ell_1, \ell_2 \ldots, \ell_n$, and finally an unfinished interval ℓ_{n+1}, the likelihood is given by

$$L(\ell_0, \ell_1, \ldots, \ell_n, \ell_{n+1}) = a(\ell_0)\left[\prod_1^n f(\ell_i)\right] b(\ell_{n+1}), \quad (5)$$

where ℓ_0 and ℓ_{n+1} are the incomplete intervals from the commencement of study to the first event, and from the last event to the end of the study, respectively, $a(x) = [1 - F(x)]/\mu$, $b(x) = 1 - F(x)$, $\mu = \int_0^\infty u f(u)\mathrm{d}u$.

The term $a(x)$ at the beginning of the sequence is the appropriate form to use if the process can be supposed stationary, but nothing is known about events before the commencement of the observation period. The term $b(x) = 1 - F(x)$ at the end of the sequence merely acknowledges the fact that the final interval has begun but not yet concluded.

The main uses of the model are in estimating long-term average or static hazards, in which case the mean of the interevent times plays the crucial role and the form of the distribution is largely irrelevant (so that a Poisson approximation would generally be adequate). The other application is to estimating the residual time to the next event, which is governed by the extended hazard function

$$h(y|x) = f(x + y)/[1 - F(x)] \quad (y \geq 0;\ x > 0)\,, \quad (6)$$

giving the density of the distribution of the time y from the present to the next event, given that time x has elapsed since the last event.

Distributions commonly used in these situations include the Weibull, gamma, log-normal and Brownian first passage time (inverse normal). Recent work has tended to favor the last of these: see [79]. Two further recent studies which look carefully at the statistical issues, including those relating to errors in the occurrence times, are in [64] and [121]. In long period studies, care needs to be taken that consistent procedures have been used over the

whole period, particularly in the determination of magnitude thresholds, and rules for the exclusion of aftershocks (which must be removed here since otherwise they would contradict the assumption of i.i.d. intervals).

Time- and Slip-Predictable Models

The time-predictable model, introduced by Shimazaki and Nakata in [116], is a widely used alternative for major events along a given fault, when magnitudes as well as interoccurrence times are available.

As in the elastic rebound model, it is supposed that stress along a particular fault builds up linearly until a critical value is reached, representing in some sense the strength of the fault. The size (magnitude) of the resulting event is not known beforehand, but is supposed to be selected randomly either from the standard G-R form or a variant suggested by the characteristic earthquake model.

Once it has occurred, the stress along the fault is instantaneously reduced by an amount determined by the magnitude of the event. Then stress build-up continues until the critical stress level is reached again. The time needed for this to occur is determined by the stress released by the previous event, and so is known, whence the 'time-predictable' title.

The major unknown in the model is the rate of stress build-up between events. If only observation times and magnitudes are available, this can be estimated, albeit crudely, by regressing the observed time intervals onto the magnitudes of the preceding events. If the magnitudes are determined up to a normal error term, with variance independent of the magnitude, this will result in a log-normal distribution for the length of the time interval following an event of given magnitude, and indeed this is commonly used. Another approach would be to use geological data to provide an initial ('prior') distribution for the slip rate, and put the further analysis into a Bayesian framework.

A simple model used for predictive purposes in some of the papers by the Working Group on Californian Earthquakes (see [137] for example), can be represented as

$$\log T_i = A + M_i + \epsilon_i \tag{7}$$

where the ϵ_i are independent, normally distributed errors with zero mean and constant variance, the M_i are the observed magnitudes of the events, and $A = -\log V$ is the logarithm of the slip rate, estimated from geological and GPS studies. Given the time x since the last major event on the fault, the distribution of the remaining time y until the next event is governed by the extended hazard function of the log-normal distribution, as in the discussion (6) of the renewal model, but with the mean adjusted to take into account the extra information provided by the magnitude of the previous event.

An underlying but subtle logical difficulty with the model is that if applied on a long time basis, the assumption of i.i.d. lognormal errors leads to unbounded fluctuations in the accumulated sums

$$V \sum_1^n T_i - \sum_1^N S_i = V \sum_1^N S_i(\epsilon_i - 1)\,,$$

where S_i is an estimate of the slip from an event with magnitude M_i. The cumulative sum on the right hand side can oscillate without bound, implying the unphysical possibility of indefinitely large fluctuations in the accumulated stress.

Shimazaki and Nakata also suggested a dual version, the *slip-predictable* model, characterized by a return after each event to a constant resting stress. The time at which the next event will occur is unknown, but given the time since the last event, the minimum expected size is determined by the stress accumulated since the previous event. [57] develops a more detailed version for use in earthquake engineering applications.

The Stress-Release Model

The stress-release model is an attempt to address similar issues from within a stochastic point process framework (see for example Chap. 7 of [20]), incorporating both occurrence times and magnitudes. As in the previous case, it is assumed that the rate of stress build-up is constant (say ρ), and that sizes of successive events are i.i.d. and independent of the stress level at the time of occurrence. Most commonly they are assumed to follow the exponential form associated with the GR law, but this is not inherent in the model.

The crucial difference with the time-predictable model is that, instead of assuming that the strength of the crust is fixed, it is assumed to be variable with distribution function say $\Phi(s)$ with density $\phi(s)$. The probability that the next earthquake occurs when the stress passes through $s, s + \mathrm{d}s$, but not before, is then given by the hazard function $\Psi(s) = \phi(s)/[1 - \Phi(s)]$. This hazard function $\Psi(s)$ determines the pattern of occurrence probabilities. Most commonly, it is taken to have an exponential form $\Psi(s) = A\mathrm{e}^{\lambda s}$, corresponding to the double exponential distribution function $\Phi(S) = 1 - \mathrm{e}^{-A[\mathrm{e}^{\lambda S}-1]}$ for the breaking strength itself. This has a well-marked mode at $(-\log A)/\lambda$ if A is rather small.

In stochastic point process terms, the quantity $\lambda^*(t) = \Psi[X(t)]$ can be interpreted as the *conditional intensity* of the model, meaning approximately the instantaneous occurrence rate, given the history of the process up to time t:

$$\lambda^*(t)\mathrm{d}t \approx \mathrm{E}[\mathrm{d}N(t) \mid \mathcal{H}(t)] \approx \Pr\{\mathrm{d}N(t) > 0 | \mathcal{H}(t)\}. \quad (8)$$

Roughly speaking, the process behaves locally like a Poisson process with instantaneous rate $\lambda^*(t)$, which in the stress-release model can be written more explicitly as

$$\lambda^*(t) = \Psi[X(t)] = \Psi\left[X(0) + \rho t - \sum_1^{N(t)} S_n\right]. \quad (9)$$

This model has several useful features. First, the fact that a simple explicit form exists for the conditional intensity means that it can be readily incorporated into standard procedures for maximum likelihood estimation, simulation, and prediction (see again Chap. 7 of [20]). In particular, the likelihood ratio for a set of observed events (t_i, M_i) over the interval $[0, T]$ can be written in the form

$$\log L/L_0 = \left[\sum \log[\lambda^*(t_i)/\lambda] - \int_0^T [\lambda^*(u) - \lambda]\mathrm{d}u\right] + \sum_1^{N(T)} \log[g(M_i)/g_0(M_i)] \quad (10)$$

where λ is the rate of the background (null) model, assumed constant rate Poisson, $g(x)$ and $g_0(x)$ are the densities of the proposed and background magnitude distributions, and the magnitudes are assumed independent.

Second, as in earlier, related work by Knopoff [58], the current stress level, say $X(t)$, is Markovian, for the current value of $X(t)$ determines the probability of the next jump occurring, while the remaining components (size of jump, rate of build-up between jumps) are independent of the past history of the process. Hence the extensive knowledge of Markov processes can be brought to bear on the properties of $X(t)$ (e. g. [11]).

A third point is that as the stress level increases, the rate of occurrence of new events will remain relatively high until a large enough event occurs to reduce the stress level to substantially lower values. The model therefore embodies a modest form of accelerated moment release [43].

The model assumes only a simple scalar concept for regional stress, much as in the early chapters of [83], and does not allow for stress interactions between regions. To address the latter point, the *coupled stress release model* was introduced by Shi Yaolin and students [68], to allow stress transfers between regions as well as simple stress drops. Further discussions and examples are in [4] and [71].

Point Process Models for Regional Catalogues

Data Consistency and Declustering

Regional catalogs, based on instrumental data from the last century or so, present a very different picture, but one with its own problems also. Of these, the two most important are the maintenance of consistency and the problem of clustering (or declustering).

It is characteristic of such catalogs that the networks supplying the data undergo many changes with the passing of the years. Although it is just these changes that have made possible the more serious statistical studies of recent years, they create their own problems in terms of lack of data consistency. Using such data for any form of long-term study requires continual vigilance over questions such as improvements and other changes in the individual network stations and their instruments, shifts in magnitude definitions or thresholds, changes in the routines used in determining epicenter locations, policy decisions over the events to be listed in the catalog, etc. Unless such factors are carefully listed and properly allowed for, they can easily lead to misinterpretation of statistical features observed in the data. As just one illustration, [18] gives some vivid examples of features of apparent physical interest which in fact have their origins in catalog artefacts induced by changes in magnitude registration.

An even more vexed question is whether, and if so how, to remove major clusters from (i. e. 'decluster') the catalog. Large aftershock sequences look simple to identify and remove, but the process is considerably more difficult than it might appear.

The possible justifications for doing so are two-fold. If it is believed that the large events are different in kind from the smaller events, then declustering is simply a procedure to isolate the events of primary importance. This assumption was once standard, and in any case the large events appear to be responsible for the major part of the large-scale tectonic motion. But with data on small events becoming ever more plentiful and increasingly reliable, their role is undergoing reassessment.

The second justification for removing aftershocks and other clusters is that they are a nuisance. They negate the assumptions of independence which lie at the basis of most standard statistical tests (e. g. for trends or periodic effects), they greatly complicate analysis and interpretation, and they require elaborate and difficult techniques to deal with explicitly.

Nevertheless, most statisticians, myself included, would tend to look askance at throwing away a substantial portion of the data on the basis of what are inevitably somewhat ad-hoc rules. Many procedures for removing aftershocks have been proposed, and the fact that none has gained general acceptance is evidence of this underlying problem. Moreover, while declustering removes the most obvious earthquake clusters, it rarely removes the clustering completely. The interpretation of results based on the remaining data remains equivocal, partly physical and partly man-induced.

For such reasons we do not discuss declustering in detail in the present article, but concentrate rather on procedures for modeling the data without removing the aftershocks.

A general caution in handling clustered data is not to presume that standard statistical procedures, especially tests, can be applied without modification. In general, the presence of clustering severely affects significance levels. For example, attention is drawn in [81] to the dangers of assessing the significance of precursory effects without properly allowing for clusters. A similar point occurs in assessing the significance of periodic effects, as was pointed out by Schuster in [112], as well as more recently in [129] and no doubt in other places.

To allow for the effects of clustering, and to examine the structuring features themselves, some form of explicit modeling is generally desirable. For example, one possible approach to highly clustered data is to remove as much of the gross clustering as possible with a basic cluster model, and then examine the residuals from fitting the model. Ogata, Zhuang and colleagues have recently developed various techniques, described in [86] and [143] for example, for examining the residuals from catalog data initially fitted by the ETAS model. Alternatively the cluster model can be fitted locally (i. e. with parameters allowed to vary in time or in space and time), and the parameter variations examined to shed more light on the features of interest: [89] contains a compelling example of such an analysis.

We proceed to describe three types of cluster model, starting from the ETAS model itself. All three models are defined through the form of the conditional intensity function, as outlined in the discussion of the stress-release model. In all three models again, magnitudes are allocated independently and randomly, either according to the GR law, or some variant such as the tapered Pareto distribution for seismic moments. The final feature in common is that in all three models the main component in the conditional intensity is a linear combination of contributions from past events.

The ETAS Model

The ETAS model (the initials standing for Epidemic Type Aftershock Sequence) first appeared in Ogata's paper [86], but was preceded by a series of studies by Ogata and colleagues in Tokyo on processes which, like the ETAS model itself, have conditional intensities of the linear, Hawkes type, following [34,35]. Earlier cluster models included the Neyman–Scott process, reincarnated as a 'trigger model' in [133] and [126].

In its basic time-magnitude form, the ETAS model has conditional intensity

$$\lambda^*(t, M) = \beta e^{-\beta(M-M_0)} \cdot \left\{ \mu + A \sum_{i:t_i<t} e^{\alpha(M_i-M_0)} f(t-t_i) \right\} , \quad (11)$$

where the first term on the RHS is the GR density for magnitudes, μ is an arrival rate for background (ancestor) events, the constant A is related to the criticality of the process, the 'productivity function' $e^{\alpha(M-M_0)}$ describes how the number of first-generating offspring increases with magnitude of the parent event, and $f(u) = p\, c^p/(c+u)^{1+p}$, $c > 0$, $p > 0$ is the density (here a Pareto form) for the distribution of the temporal lag between the arrival or birth of the parent and that of its offspring.

In the full space-time-magnitude version

$$\lambda^\dagger(t, x, M) = \beta e^{-\beta(M-M_0)} \cdot \left\{ \mu h(x) + A \sum_{i:t_i<t} e^{\alpha(M_i-M_0)} f(t-t_i) g(x-x_i) \right\} . \quad (12)$$

The new terms are the density h of new arrivals over the spatial region, and the density g in space for the location of an 'offspring' event about its parent. We suppose that f, g and h are all normalized to form probability densities.

One of the main attractions of the Hawkes' processes, including the ETAS model, is that they have a branching process interpretation, first pointed out in [36] and implicit already in the description of the conditional intensity. For example, the criticality parameter (mean number of offspring per ancestor, averaged over the magnitude distribution for the ancestor) is given by $A/(1-\alpha/\beta)$ for both the above forms. Thus a stable version of the process can exist only if $\alpha < \beta$, and then only if A is small enough. Of course, branching process ideas appear in many earthquake occurrence models, notably in Kagan's work (e. g. [36]).

While the branching process interpretation gives much insight into the structure of the ETAS model, statistical analysis depends crucially on the representations in (11) and (12), since they lead to the relatively tractable form (10) for the likelihood.

For computational purposes, the likelihood of a general marked point process, of which the ETAS models are examples, is often written most conveniently in terms of the conditional intensity $\lambda_g^*(t)$ for the *ground process*, the overall occurrence of points, irrespective of location or mark, and the conditional mark (in our case space and magnitude coordinates) distribution $f^*(x, M|t)$, so that $\lambda^*(t, x, M) = \lambda_g^*(t)\, f^*(x, M|t)$. The star indicates that the quantities so labeled are in general conditional on (and hence functions of) the histories up to time t. Provided f^* is normalized to a probability density for any given past history, we can write the likelihood ratio in the form

$$\log L_1/L_0 = \left[\sum_{i=1}^{N(T)} \log[\lambda_g^*(t_i)/\lambda] - \int_0^T [\lambda_g(t) - \lambda]\mathrm{d}t\right] + \sum_{1}^{N_g(T)} \log[f^*(x_i, M_i \mid t_i)/f(x_i, M_i)]\,, \quad (13)$$

where the terms λ and $f(x, M)$ relate to the rate and mark distribution for the background process (null model), here taken to be a constant rate Poisson process with independent (and usually GR) magnitudes.

This form represents the likelihood ratio as the sum of two terms, the first involving the time points only, and the second involving the marks (spatial locations) given the time points. In many models, the parameters appearing in the two terms have no common variables, in which case optimization can be carried out for the two terms separately.

Because the ETAS model fits well to catalogue data over a wide range of scales and contexts, in recent years its properties have been examined in detail, with the aim of verifying its ability, or otherwise, to reproduce specific features of the real process, such as Båth's law or the occurrence of foreshocks; see, for example, [37,38].

Moreover, the procedures developed by Ogata and colleagues for fitting versions of the model in which the parameters can vary in both location and time, and for detecting local departures from a good fit of the model, have made the ETAS model a powerful diagnostic tool. In this way it has been used to estimate local variations in the stress field (e.g. [89,91]), or changes in seismicity due to the intrusion of ground water [30,58].

For long it was believed that an immigration component, coupled to a subcritical branching structure for the offspring, was the only way to produce a stable process with branching structure. However, it was shown recently in [12] that when the temporal lag distribution $f(.)$ of (11) is very long-tailed, a critical Hawkes process can sustain itself indefinitely as a 'process without ancestors'. Another somewhat unexpected extension, described in [132], is to a self-similar version over an infinite range of magnitudes.

Perhaps the one serious limitation of the ETAS model is its rather poor performance as an intermediate-term predictor. The reason for this is that its predictie power is basically dependent on its ability to fit aftershock sequences. Hence it does not show significant gains, even against the Poison model, until the time intervals between forecasts are of similar order of magnitude to the time intervals between the larger events in an aftershock sequence.

The Kagan–Jackson Models

Kagan and Jackson have proposed a number of forms of which we refer to two, the long-term and short-term versions of [41]. The long-term version has its origins in [51], while the short-term version has its origins in [55].

In the long-term model, the current value of the conditional intensity, within a spatial region A and based on observations within $(0, t)$, has the form

$$\lambda^*(t, x, M) = f_t(x)g(M)h(t) \quad (14)$$

where $h(t)$ is the overall current risk (ground process intensity), $g(M)$ is a (fixed) magnitude distribution, commonly that corresponding to the tapered Pareto form for seismic moments, and

$$f_t(x) = \sum_{0<t_i<t} k(x - x_i) \Big/ \sum_{0<t_i<t} \int_A k(x' - x_i)\mathrm{d}x'$$

k being a spatial kernel function. Thus f_t is a normalized sum of contributions from previous events within the observation period and spatial region. It is time-independent except insofar as the advent of additional events requires additional renormalization.

Although the model is well-defined by its conditional intensity and an initial condition at $t = 0$, and can be fitted by likelihood methods much as for the ETAS model, the renormalization introduces a non-linear component into the model which makes its properties more difficult to analyze than those of the ETAS model. Moreover, like other forms of moving average model, it is non-ergodic: there is no unique stationary form to which it will converge from different initial conditions. Nevertheless, it serves the principal purpose of providing a baseline comparison for other putative prediction models for the same region.

The short-term form is very close to the spatial ETAS model. As in the ETAS model, each past event is associated with both a spatial and a temporal decay function, the temporal decay following the Omori law. Again, however, it involves a renormalization rather than the introduction of an explicit immigration term into the model, although some versions of the model allow a small quota of 'surprises' in parts of the region with no previous earthquakes.

An important role for both models has been to focus attention on the need for systematic, long-term evaluation and comparison of forecasting models, and to provide a more relevant null model than the constant-rate Poisson process.

The EEPAS Model

This model grew out of several decades of experimentation with precursory swarm models by Evison and Rhoades; see [24] and [104] for more of the history and underlying concepts. The precursory swarm models identify groups of moderate-sized earthquakes as precursory swarms and use these as possible precursors of large earthquakes. Since both precursory phenomena and forecast phenomena belong ultimately to the same process of earthquake formation, a more satisfactory approach is to try and develop a joint model for both phenomena. This, in effect, is what the EEPAS model achieves. It has its own rationale, based on a theory of growth and development of crustal fractures outlined in the papers cited above, and has been successfully applied in several major seismic regions (e. g. [100] and [105]).

Much as in the ETAS model, the conditional intensity has the general form

$$\lambda^*(t,m,x) = \mu\lambda_0(t,m,x) + \sum_{t_i<t} w_i\,\eta(m_i)\,r(M \mid M_i)\,f(t-t_i|M_i)\,g(x-x_i|M_i), \tag{15}$$

but the details are significantly different.

First, the conditional intensity in (15) is not taken to apply to the whole catalogue from which the events on the right side are derived, but only for events above a higher threshold. Thus, the model might be used for modeling (and predicting) events over magnitude 5.8, but would take explanatory data from the catalogue of events with $M \geq 4$.

Second, the functions f and g are not based on Omori-type decay formula, but on logarithmic regressions for the time and space delays between an initiating event and the event it may precede, and the magnitudes of the two events. Thus for example

$$f(u|M_i) = \frac{1}{u\sigma_T\sqrt{2\pi}}\exp\left[-\frac{(\log u - a_T - b_T M_i)^2}{2\sigma_T^2}\right],$$

with an analogous expression for $g(w)$, while $r(m)$ takes the form

$$r(m|M_i) = \frac{1}{\sigma_M\sqrt{2\pi}}\exp\left[-\frac{(m - a_M - b_M M_i)^2}{2\sigma_M^2}\right].$$

They differ considerably from the functional forms used in the ETAS model, but are similar to relations used in the precursory swarm models.

The weight factors w_i are commonly set to unity, but in more refined analyzes may be down-weighted when the triggering event has been identified as an aftershock. One way of finding suitable weights is to carry out an initial ETAS stochastic declustering, as in [143], and base the weights on the probability that a given event is independent.

The further normalizing factor $\eta(m_i)$ in (15) is introduced, much as in the Kagan–Jackson models, to offset the absence of any immigration term, and to compensate for the input from earthquakes below the magnitude threshold. It is adjusted for each magnitude class m_i so that the overall rates follow the G-R law.

In practice the contribution from the baseline rate density is often so small as to be negligible.

Our impression is that the EEPAS model is currently the best-performing of the general seismicity models in the sense of producing the highest average probability gains or entropy scores (see Sect. "Assessing Probability Forecasts") for predicting moderate to large events on intermediate time scales.

Stochastic Models with Precursors

General Considerations

The search for reliable earthquake precursors has a long and troubled history. High hopes in the 1970s met many disappointments, some at least arising from an inadequate appreciation of the many statistical pitfalls. These difficulties are now much better appreciated, but even so there are relatively few studies based on a satisfactory statistical model, incorporating a proper assessment of the uncertainties, and showing a significant precursory effect.

From a modeling point of view it is important to distinguish between *complete* and *partial* models. In a complete model, both the earthquakes and the precursors are included as components of an overarching joint process, the earthquakes forming one marginal process and the precursors another. In principle such a complete model

should be the aim, but in most situations either the background physics is insufficiently understood, or observations on the precursors are inadequate, or the statistical analysis is too difficult, to allow such a joint analysis to proceed. In many situations also, the modeling and analysis lie outside the realms of conventional statistical models (in dealing with self-similarity, for example), raising further procedural problems.

In a partial model, no attempt is made to model the precursors as a stochastic process. They are treated as given, and their data used in regression-like procedures to modify the probabilities of earthquake occurrence. They can be used retrospectively to examine the performance of a suggested predictive relation, but their use in developing probability forecasts is limited to the short term because there is no model to forecast the future behavior of the precursors.

In the later parts of this section we illustrate some of the modeling approaches that can be used with precursors, using examples drawn mainly from our own experience.

One issue that commonly arises is that the precursory signal data are derived from observations taken at fixed sampling intervals. To match such data to the point process data for the earthquakes, it is generally easiest to switch the whole analysis to a discrete-time study. In this case the continuous time point process models of the previous sections need to be replaced by approximating discrete time models. The two most common of these are the logistic, or binary data, models, and the discrete Poisson process models, illustrated in the first two examples below. In these two classes of models, the probabilities p_n that an event occurs in the nth interval (respectively, the means μ_n of the Poisson distribution for that interval) play the role of the conditional intensity function $\lambda(t)$ of the continuous time models of the previous section.

Example 1: Logistic Regression Analysis of M8 Series

Logistic regressions are used to directly assess the effect of precursor observations on the event probabilities p_n. Suppose that at the nth interval, observations $(U_1^n, U_2^n, \ldots, U_k^n)$ are available on k precursors. Dropping the n for brevity, the logistic regression takes the general form

$$\log\left[\frac{p}{1-p}\right] = \alpha_0 + \sum_{k=1}^{K} \alpha_k U_k \tag{16}$$

where the left side is the logit (log-odds) transform of the event probability, and the right side is the regression term. This representation corresponds to the canonical form for a binomial distribution as a member of the exponential family. Standard routines exist for estimating the parameters α_k by maximum likelihood or closely related methods, and form part of the generalized linear model procedures.

As an example we consider the model used in [32,33] on the output from the M8 algorithm on New Zealand data.

The M8 algorithm itself is not a stochastic process model, but a decision procedure for calling an earthquake alert based on the analysis of 7 contributing series from earthquake data within a specified 'region (circle) of investigation'. It is the best known of a number of pattern-recognition algorithms developed by the Russian group headed by Keilis-Borok during the 1970s and subsequently. The basic form of the algorithm is described in [56] and [59], with recent reviews in [60,61].

The heart of the algorithm consists of a set of decision rules, based on the joint behavior of the 7 time series, for calling an alert, or more specifically the announcement of a 'TIP' (time of increased probability of an event above a given magnitude threshold) over the region of investigation. The time series are updated every six months, and a TIP extends for three years in the first instance.

A key feature of the analysis in [32,33] is that, in each six-month interval, the values of the seven series are combined by a non-linear formula (linear methods seem less effective) into the value U_1^n of a 'critical series' which is then used as the single precursor in a logistic regression model. The non-linear formula mimics the structure of the decision rules used in declaring a TIP.

The logistic regression analysis then provides the probability for the occurrence of an event over the specified magnitude threshold within the region of investigation for the current 6-monthly period. Much of the further discussion in [32,33] is concerned with combining the outputs from overlapping regions of investigation.

Note that the analysis is typical of that for a partial model; a complete model for M8 would model the joint distribution of the M8 series and the target events.

Example 2: Discrete-Time Poisson-Type Model for ULF Electric Signals

In a discrete-time Poisson-type model, the number of events Z_n in the nth time interval is modeled as a Poisson variable with mean μ_n^* that is treated as a function of the past history in much the same way as the conditional intensity in the continuous-time model. The likelihood ratio against a constant mean Poisson process takes the form

$$\log L/L_0 = \sum_{1}^{N} Z_n \log(\mu_n^*/\mu) - \sum_{1}^{N} (\mu_n^* - \mu)\,. \tag{17}$$

The analogy with the point process form (10) is very obvious, particularly as the time intervals become small so that with high probability the Z_n are either 0 or 1.

The way in which μ_n^* depends on the past can be very general, subject only to the constraint $\mu_n^* \geq 0$. In particular, μ_n^* can depend on prior observations both of the process itself and of auxiliary (precursory) variables. If a linear conditioning of the type (16) is required, the multiplicative form

$$\log \frac{\mu_n^*}{\mu} = \sum_{k=1}^{K} \alpha_k^{(n)} U_k^{(n)} \tag{18}$$

can be used, ensuring that μ_n is positive, and slotting into the canonical form for the Poisson distribution as a further member of the exponential family, so that the generalized linear model procedures become available. However, the form of the likelihood (17) is usually simple enough to be maximized directly even in more general cases.

An example is the analysis of ULF (ultra low frequency) electric field data in [144].

The ULF signal referred to here is made up of small fluctuations in electric potential measured some meters below the ground surface by sensitive and well shielded electrodes. Its role as a precursor, and the physical explanation of the phenomenon, if it exists, are still unclear. The data analyzed in the paper cited come from some thirty years of recordings from stations around Beijing in China.

Here the base-line (reference) model is a self-exciting (Hawkes type) process in discrete time for the daily earthquake numbers in a wider region around Beijing. Such a model is needed as a reference model because it takes into the inherent clustering effects of the earthquakes themselves. Otherwise there is a temptation to interpret (wrongly) all improvements over the constant rate Poisson process as due to the signals and not to the inherent clustering of the earthquakes. The regressands were the daily readings of ULF anomalies at a set of some half-dozen recording stations around Beijing, simplified to form 0-1 series of observations above a threshold. In fact the daily numbers of earthquakes were small enough for the corresponding continuous and discrete time models to be essentially identical.

Two analyzes were carried out, first with a linear Hawkes-type representation incorporating the effects of past earthquakes alone (self-exciting model), none of the ULF data being used, and second with a double (mutually exciting) linear Hawkes-type representation for the effects of both past earthquakes and ULF signals on the current rate. Likelihood ratios were taken first with respect to a constant rate Poisson process, optimizing parameters in both cases, then (as a ratio of ratios) for the first and second models against each other, to allow the improvement due to adding in the information from the ULF signals to be assessed.

In this study the model was also tested in reverse mode, to see whether the earthquakes improved the likelihood performance of a Hawkes-type model for the ULF signals alone. The results were positive in the direct mode and negative in the reverse mode. Either way, the models were partial, not complete, as no attempt was made to provide a joint model for the earthquakes and the ULF anomalies together.

Discrete-time Poisson-type models are also used as the basis for model testing within the RELM testing center in Southern California [111]. The modeler supplies the Poisson rates $\mu(n, r, m)$ not only for each time interval (n) but also for each spatial bin (r) and magnitude bin (m). The (approximative) assumption is then made that all the Poisson variables relating to a given time interval are conditionally independent given the current Poisson rates, so that a likelihood ratio of the form (17) can still be used, and made the basis of comparing different proposed models.

Example 3: Point Process Regression Models

As already hinted at, continuous time (point process) procedures can be developed along similar lines to those in the previous example, if the past history includes information on auxiliary (precursory) variables as well as the history of the point process itself.

The form of such dependences can be very general, but if a linear form is wanted it can be incorporated through expressions of the form

$$\log\left[\lambda^*(t)/\lambda_0^*(t)\right] = \sum_{k=1}^{K} \alpha_k U_k(t) \tag{19}$$

for the ratio of intensities, in a similar way to the multiplicative form for the discrete Poisson model. In the special case that $\lambda_0(t)$ corresponds to a renewal process, this is the well-known *Cox regression model* [18]. A comprehensive treatment of models of this kind, with mainly medical and social science applications, is in [2].

Few examples of this type are known to us in the seismological literature, and this may be one area with scope for further development. For example, it is possible that the time-predictable model could be reformulated as a Cox regression model, by building in the dependence on the size of the previous event as a regressand. The closest to a model of this kind that already considered in this article is perhaps the EEPAS model, where the events used in

developing the conditional intensities are mainly smaller than the events being modeled, and so could be described as precursors. Thus the conditional intensity for the larger events being modeled is regressed onto the magnitudes, times and locations of the smaller events in the catalogue. If the renormalization and immigration terms were omitted, the conditional intensity (15) of the EEPAS model would have a similar basic form to (19) above.

Example 4: Foreshocks

Foreshocks, with their potential for earthquake prediction, have been of interest to seismologists since the days of Omori and other earthquake pioneers. The hope that increased quantity and quality of catalogue data would lead to a more definitive picture of how, when and why foreshocks occur has not yet been realized, however. Features which would discriminate foreshocks from other earthquakes or earthquake clusters have proved hard to identify, although careful studies by Ogata and colleagues (e. g. [90]) do suggest that limited opportunities for discrimination may exist.

As precursors, foreshocks retain a specific but limited role. Some degree of forecasting power is available simply from the fact that any newly-observed event, outside those in clearly defined aftershock sequences, has the potentiality of being a foreshock. Recent studies of foreshock occurrence from this point of view are given (among others) in [46] for Southern California, [96] for large global events, [80] for New Zealand earthquakes.

The studies suggest that, leaving aside events in an obvious aftershock sequence, between 5 and 10% of earthquakes with magnitudes 4 and over are likely to be followed by a larger event within time and space windows of the order of 4–5 days and 20–30 km radius respectively. No very sophisticated stochastic model is required to describe such a feature: within the defined window, the probability that a larger event will occur is simply increased from its background value to about 5%, and multiplied by a standard GR factor to take into account magnitude variation; a further separate factor can be used to take into account sub-regional variations of foreshock probabilities.

In the last few decades foreshocks have been studied from a more general point of view, as evidence or otherwise of the self-similarity of earthquake occurrence. In a branching model such as the ETAS model, no foreshock feature is explicitly built into the model, but from among a given parent's offspring, one will occasionally appear with a magnitude larger than that of its parent – with frequencies again in the vicinity of 5%, approximately irrespective of the magnitude of the parent.

That just such an interpretation may apply also to real earthquakes is suggested in [25]. Foreshocks then are not a specific physical phenomenon, but just parent events which happen to have offspring larger than themselves. An interesting connection between the probability that an initial event is a foreshock (followed by a larger event as offspring), and the distribution of the Bath's law magnitude gap, is put forward in [25] and [38].

Further Topics

There are many further topics where stochastic modeling ideas are relevant, even if they do not necessarily involve the development and fitting of a full model. For example, the last few decades have seen considerable work on the development of procedures for producing and assessing probability forecasts, and for quantitatively describing spatial or space-time point patterns. In this section we list what seem to us to be some of the more important topics of this kind, space precluding more than very cursory accounts.

Generating Probability Forecasts

There is no longer any very hard and fast line between predictions and probability forecasts. It has long been recognized that any prediction of the time and place of a forthcoming event must be accompanied by some statement of the uncertainties in the prediction. But this, as discussed in the introduction, is precisely the motivation for introducing stochastic models. Using such a model, the uncertainties can be rephrased in terms of the probabilities of occurrence within specified time intervals and spatial regions, i. e. by probability forecasts. In our view such probability forecasts represent the most useful way of summarizing information about the uncertainties regarding future events. In statistical jargon, they represent the 'predictive distributions' for future events, and form the basis not only for the probability forecasts themselves, but also for any associated cost-benefit analyzes.

Even within a probabilistic framework, the idea that precise, medium-term or even short-term forecasts may be possible has looked increasingly like a pipe dream over the last few decades. Nevertheless, it is not yet entirely ruled out. While the emphasis in the last few decades has been on increased surveillance, improved knowledge and understanding of long-term hazards, and the reduction of risks from earthquake hazards, medium term (months to years) probability forecasts are winning a new role as refinements of more traditional long-term, static hazards for both building and insurance purposes. In addition very short term forecasts play a useful role in connection with

the progression of aftershock sequences, and in developing real-time warnings for trains, gas supplies, and other facilities at high risk from a serious earthquake.

All of the models outlined in the previous sections can be used to develop probability forecasts. Typically, such forecasts are derived from simulations rather than from analytical studies. For example, simulation schemes for use with conditional intensity models are described, with further references, in Chap. 7 of [20].

For complete models, forecasting schemes making use of such procedures involve first fitting a model on all the data up to the current time, then simulating the model as far as is desired into the future, and using these simulations to estimate any required probabilities or expected values as in a Monte Carlo study.

Forecasts for partial models can proceed along similar general lines, although in the first instance the dependence on auxiliary variables restricts the probability forecasts to just the next forecasting period. To obtain forecasts beyond this first step, and in the absence of any updated precursor data, it is necessary to develop a further set of 2-step forecasts, using only the current precursor data, and so on successively, the forecasts gradually reducing in power.

Assessing Probability Forecasts

Two main approaches to assessing probability forecasts are through likelihoods and probability gains, and through their performance in decision schemes based on the forecast probability exceeding some threshold. Since most procedures are reduced in practice to forecasts for finite forecasting periods, we outline assessment procedures for this case only.

The *probability gain* for each forecast (i. e. for each time, time-space or other interval for which a probability is forecast) is the ratio of the forecast probability for the observed occurrence in that interval to the corresponding probability from a standard reference model, such as a simple or compound Poisson process, for that interval.

The sample averages of the log-probability-gains, or *entropy scores* in the terminology of [34], over all or certain classes of intervals or of observed outcomes, provide useful insights into the performance of the forecasts. From them one can quickly perceive the outcomes which the scheme is forecasting well, and those which it is forecasting badly.

The overall average, or more properly the expected value, of these entropy scores is called in [34] the *information gain* relative to the reference model. In a complete model, the information gain is a numerical characteristic of the model, giving an upper bound for the improvement in performance that can be expected for the proposed model, relative to the reference model, when the proposed model is the true model. In many cases it reduces to the expected value of the mean likelihood ratio. For example, if successive forecasts, derived from the use of a particular model, give probabilities $p_1^*, p_2^*, \ldots, p_N^*$ for the successive observed outcomes, and $\bar{p}$ is a constant probability used as reference, then the sum

$$(1/N)\sum_{1}^{N}\log[p_n^*/\bar{p}]$$

is the empirical entropy score for that model as well as the mean log-likelihood ratio. It approximates the corresponding expected value (the information gain) if the model used is the true model. If the true model is unknown, as will usually be the case in practice, the above average approximates the difference in the Kullback–Leibler distances between the given model and the true model, and the reference model and the true model. For a set of models, that giving the largest entropy score should be that closest to the true model in the sense of Kullback–Leibler distance. In a partial model, expectations cannot be taken over values of the auxiliary variable, since its distribution is not included in the model specification, but at least empirical averages based on past observations can be developed. [34] gives further background and examples; the basic idea of using log-probabilities (i. e. loglikelihoods) as an indicator of forecasting performance goes back at least to [52].

We noted in Subsect. "Example 2: Discrete-Time Poisson-Type Model for ULF Electric Signals" that such comparisons of likelihood ratios form the basis of the assessment procedures used in the RELM testing center for probability forecasting schemes; see further [111].

The more traditional procedures for assessing probability forecasts suppose that the schemes are first turned into prediction schemes by predicting that an event will occur whenever the forecast probability of that event exceeds a certain threshold value. The results can then be put into a 2 x 2 table (occurrences or non-occurrences, versus predictions or non-predictions). If the entries in the table are labeled as

(a) number of successful forecasts of occurrence,
(b) number of failures to predict,
(c) number of successful forecasts of non-occurrence,
(d) number of false alarms,

the commonly used *R-score*, or *Hanssen–Kuiper skill score*, can be defined as

$$R = \frac{ac - bd}{(a+b)(c+d)} = \frac{a}{a+b} - \frac{d}{c+d} = \frac{c}{c+d} - \frac{b}{a+b}. \tag{20}$$

The R-score varies between -1 and $+1$, with 0 denoting forecasts independent of outcomes, and the two extreme values perfect non-prediction and perfect prediction, respectively. [91] illustrates the use of the R-score to evaluate Chinese yearly forecasts, even though these are based on expert opinion rather than any probability model.

The usefulness of this approach was greatly extended in [84,85], by allowing the threshold probability p_c to vary and calculating for each such value the ratios

$$\nu(p_c) = \frac{c}{(a+c)}$$

and

$$\tau(p_c) = \frac{(a+b)}{(a+b+c+d)}$$

The resulting $\nu - \tau$ diagram, obtained by plotting these quantities against each other, provides a comprehensive summary of the behavior of the probability forecasting scheme. The diagram typically consists of a convex (downwards) curve reducing to the diagonal joining the points $(0,1)$ and $(1,0)$ in the case of purely random forecasts, and to the two $(0,1)$ segments of the axes in the case of perfect prediction. It is essentially a Q-Q (quantile-quantile) plot of the distribution of the proportion of time on trial, and the the proportion of failures to predict. See the papers quoted or [34] for further details.

It should be emphasized that the procedures described are concerned with the scientific issues of assessing the quality of proposed models rather than the practical issues of issuing and using probability forecasts. The latter raise just as many difficulties, if not more, than the former. Among these are the need to take into account errors in the model as well as the uncertainties described by the model itself, the need to develop decision-making frameworks that can take advantage of the information within probability forecasts, and the need to address the social, political and economic consequences of issuing forecasts.

Change-Point Models

Change-point models are used retrospectively to identify time-points at which a change occurs in quantities such as a mean value or rate. They are not themselves precursors, but are used rather to indicate the onset of a period which is anomalous in some sense, and may therefore have some precursory significance. In practice, one of the biggest difficulties in using such methods with seismicity data lies not in identifying the change point, but determining the region from which the data for the change-point analysis should taken.

The procedure consists essentially of dividing an observation time period, say $(0, t)$, into two segment $(0, t_1)$ and (t_1, t), and finding the value of t_1 which maximizes the discrepancy in the values of the quantity being studied. The value of this maximized discrepancy can then be tested for significance using the null hypothesis that the values in both periods are equal.

Such a change-point technique is developed in [86] to detect the onset of precursory quiescence within a selected observation region. The data is first fitted to the ETAS model. The time axis is then transformed by the random time transformation

$$\tau = \int_0^t \lambda^*(u)\mathrm{d}u$$

which has the effect of transforming the original point process, with conditional intensity $\lambda^*(t)$, into a unit rate Poisson process (see, for example, Sect. 7.4 in [20]). If, however, there is a change in the parameters of the original ETAS model, this will show up as a change in the rate or other perturbation in the unit rate Poisson process, for which many tests are available.

This technique can be applied both to background events in a specified region (e. g. [140]), or to events within the course of an aftershock sequence, anticipating the occurrence of a large aftershock (e. g. [77]).

Ogata, Toda and colleagues make use of similar but more sophisticated procedures for identifying the regions and onset times of stress-shadowing through careful monitoring of activity of small events; see in particular [89,91].

A different type of change-point analysis has been developed by W. Smith for detecting changes in b-value. Although one of the earliest possibilities suggested, the precursory role of b-value changes as a precursor has never been unambiguously identified. Early papers such as [82] or [109] suggested a link to physical features such as heterogeneity of the crustal material, or changing levels of stress. The latter interpretation is at least partially supported by the branching and similar models for earthquake mechanism outlined in Sect. "Branching Models ".

Smith's approach is based on modifications of the CUSUM procedure widely used in quality control contexts for detecting departures from normal behavior in a production process. They are essentially methods for detecting a change in slope of cumulative occurrences or other similar sums. For further details and reviews of earlier work on b-value changes, see [117,118].

Finally, there is the possibility of a Bayesian approach to change-point problems in seismology; some discussion and an example are given in [94].

Change point models are closely related to *hidden Markov models* in which the rates and other characteristics of the observed process change with the state of a Markov process which is hidden or at best only partly observable. Such models have been widely used in speech recognition and modeling of IT traffic (see, e. g. [72]). Their use in seismology is relatively new (see [23] for a recent example) but their potential seems worthy of further exploration.

Moment Measures and Correlation Functions

Moment structure plays an important role in most stochastic processes, and the same is true for point processes. The main attention is on second order or correlation properties. Indeed, where distributions are more or less Gaussian, the distinction between models and second order properties is largely nominal, since a Gaussian model is fully described by its means and covariances. The same is not true for count data, for which the second order properties form an important but not in general a definitive aspect of the overall model structure, since a range of different models can be developed to fit particular second order characteristics.

Early papers in the seismological context include [16, 53,128]. There is also a considerable literature in astrophysics, relating especially to the distribution of galaxies and the role of 2-point, 3-point and higher order correlation functions (e. g. [76]).

Second-order properties for point processes mean properties of point-pairs. Their behavior is described through the second moment measure, whose density (when it exists) is given for $x_1 \neq x_2$ by

$$m_2(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 = E[N(\mathrm{d}x_1)N(\mathrm{d}x_2)] \,.$$

Apart from a renormalization, this is also the 2-point correlation function of [53]. It is also the basis of the covariance measure, with density

$$c(x_1, x_2) = m_2(x_1, x_2) - m_1(x_1)m_1(x_2) \quad (x \neq y) \,.$$

This also exists in various renormalized forms, for example the radial correlation function defined, when the process is isotropic, by

$$\rho(r) = \mathrm{d}K(r)/\mathrm{d}A(r)$$

where $A(r) = \pi r^2$ is the area of a circle centered on an arbitrary point of the process, and $K(r)$ (often referred to as 'Ripley's K-function') is the expected number of additional points (i. e. apart from the point at the center) in the same circle (see e. g. [22,104]).

In a general treatment, the coordinates x_1, x_2 may combine both space and time components. The second moment densities then give information about the expected density of occurrence of additional points at given time or space intervals about a given point taken as origin. In this way they can display distance variations in the strength of the clustering tendency.

The second order techniques are particularly valuable when the process is stationary in time, so that spectral methods can be used. In this situation the covariance measure becomes a function of time through the difference $t_1 - t_2$ in the time coordinates of the two points being considered:

$$c_2(x_1, x_2) = c_2(t_1 - t_2; y_1, y_2) \,,$$

where y_1, y_2 may represent locations or magnitudes. Taking Fourier transforms with respect to time leads to a spectral density, multivariate if the dependence on space or magnitude is retained, which can be used for the analysis of periodic effects in point processes. [133] is an early example of the spectral analysis of earthquake occurrence data. More general discussions can be found in [13] or Chap. 8 of [20].

Principal Component Analysis

Principal components are the names given to the eigenvectors in the orthogonal decomposition of a model-derived or empirical covariance matrix. Being symmetric and positive definite, the diagonal representation of such a matrix has non-negative eigenvalues, which measure the proportion of variation associated with the given eigenvector. Thus the principal components associated with the largest eigenvalues define those linear combinations of the observation vector components which explain the greatest amount of variability.

In geophysical applications, notably in meteorology, the observations typically arise as time sequences of observations from a set of recording stations. Each time point gives a vector observation. The principal components associated with the largest eigenvalues then often describe recognizable weather occurrence patterns, and are used as a means for identifying and classifying such patterns.

This idea admits many specializations and extensions. For stationary processes in time, for example, the covariance matrix has a special structure ($c_{i.j} = c_{i-j}$) which for infinite series, or a finite series with periodic structure, causes the eigenvectors to reduce to complex exponentials $\mathrm{e}^{\mathrm{i}\omega_r n}$. These periodic terms form the principal components of the sequence, while the corresponding eigenvalues are amplitudes associated with the given frequen-

cies. Thus standard spectral analysis can be interpreted as a special form of principal component analysis. The so-called Karhunen–Loeve theory (see e. g. [69]) is an extension to very general covariance operators admitting a diagonal decomposition.

Adaptations of these ideas for earthquake data have recently been considered by Rundle, Tiampo and colleagues (see e. g. [107,120]). Here the data is typically of point process form (0-1 functions) on a sequence of observation regions, centered either on a lattice of space or space-time points, or on the centers of a family of small fault segments over an interacting fault system. Their papers combine the Karhunen–Loeve theory with ideas originating in dynamical systems theory, and seek to extend the analysis to provide short or medium term probability forecasts.

Fractals and Fractal Dimensions

Many of the models so far considered show long-range dependence, power-law distributions, and some form of self-similarity (similar appearances on different scales). The term 'fractal' is commonly used in this context, with a variety of interrelated interpretations. A wide-ranging review of such properties for earthquakes is given in [121].

In the basic text [74] of Mandelbrot, fractals are introduced as a class of geometric objects (sets) with features which repeat on a family of reducing scales. The Cantor set in one dimension and the snowflake curve in 2 dimensions are well-known examples. Their characteristic feature is that the Hausdorff dimension of the fractal set may differ from the dimension of the space in which it is embedded. Thus the Cantor set has dimension $\log_3(2) < 1$ despite being embedded in a 1-dimensional space. As in this example, these aberrant dimensions often have a non-integer value, whence the name.

Various empirical procedures have been devised for determining these fractional dimensions. For example, for a set in 2 dimensions, we may count the proportion of cells from a 2-dimensional lattice which have a non-empty intersection with the set, and consider the behavior as the cell size in the lattice approaches zero. Typically this results in a log-log relation between number of affected cells and the cell dimension, corresponding in fact to power-law behavior. Such methods can be applied also to physical sets such as the traces of faults on the earth's surface (see [121] and [45] for examples). If a suitably linear portion of the log-log plot can be found, the slope is used as an estimate of their fractal dimension.

Because power-laws arise inevitably in the study of fractal sets, the term *fractal behavior* has come to be used loosely to describe the behavior of any probability distribution for which

$$1 - F(x) = \Pr(X > x) \sim cx^{-\alpha} ,$$

or even when power-law forms arise in features such as time or space correlation functions.

A different fractal concept derives from the work of Renyi [99]. Renyi's dimension estimates (also called multifractal dimensions) reflect the irregularities of a measure (distribution) rather than a set. Thus, if a probability distribution on a square is used as an example, and the square is divided into a lattice of small squares Δ_i each of side Δ, we may consider the limit as $\Delta \to 0$ of the ratios

$$d_q(\Delta) = \left[\log \sum_i p(\Delta_i)^q \right] \Big/ \log \Delta \quad (q \neq 1) \tag{21}$$

with

$$d_1(\Delta) = \left[\sum_i p(\Delta_i) \log p(\Delta_i) \right] \Big/ \log \Delta \, . \tag{22}$$

For each q, d_q defines a *multifractal dimension* associated with the distribution, and for varying q the family of such dimensions defines a type of transform of the underlying probability distribution.

The term multifractal is used because the distribution may exhibit different forms of singularity, associated with different power-law behavior, at different points of the space. In applications, for example in turbulence and meteorology as well as seismology, the process is visualized as a descending sequence of random eddies ('random cascades') with dimensions shrinking to zero. It is in this sense that Kagan [49] coined the phrase 'frozen turbulence' to describe earthquake processes in the crust.

The computations in the equations above can be applied empirically to quantities such as the counts or the energy release from earthquakes and used to estimate the multifractal dimensions of some underlying spatial distribution. For general discussions of multifractals, see [95] and [31]; the latter gives an extended discussion of applications of the theory to earthquake data.

With count data, there is a link between the multifractal dimension d_q for integer q and the q-point correlation function. In particular d_2 describes the growth rate of the two-point correlation function for vanishingly small separations; see [53,131].

Self-Similarity

Self-similarity is used in a broad sense to describe processes in space or time where, much as in a fractal set, key structural features are preserved on a descending sequence

of scales. Plots of fault traces or epicenters do indeed suggest such scale invariance properties.

In stochastic process theory, self-similarity (automodeling) is used to denote a precisely defined property of random processes or random measures. In particular, a random measure $\xi(A)$ defined on sets in the plane (or other Euclidean space) is said to be *self-similar*, with similarity index H, if a change of scale in the space can be compensated for by a change of scale in the quantity being measured:

$$\xi(rA) =^{\mathcal{D}} r^{-H}\xi(rA)\,, \quad r > 0 \tag{23}$$

where $=^{\mathcal{D}}$ means the two sides have equal probability distributions.

Self-similar random measures of this form arise in turbulence theory, seismology, finance, and elsewhere. In one dimension, the concept can be reinterpreted as implying that a stochastic process has *self-similar increments*. Brownian and fractional Brownian motions both have increments of this type. However, the Brownian motions take both positive and negative values, whereas any random measures associated with earthquakes (for example, by way of energy release) should be non-negative.

Thus the self-similar processes arising in seismology have a rather special character, since they must be non-negative as well as self-similar. The only known examples are purely atomic random measures: $\xi(A)$ is obtained by summing the contributions (energies, for example) from point-events in A.

One such example is well-known: the *stable* random measures, which are characterized by a strong independence property, and a power law distribution for event sizes equivalent to a G-R relation for magnitudes. Superficially, this example matches the earthquake data rather well, but it fails to incorporate one key feature of earthquake occurrence: the lack of independence inherent in earthquake clustering. Recently, however, it was shown in [132] that it is possible to define a variant on the ETAS model which combines self-similarity with a non-trivial dependence (cluster) structure. It is not known at present how wide this class of models may be.

Block-Slider and Related Mechanical Models

The mechanical models attempt to illuminate the processes of earthquake occurrence by devising complex mechanical systems that will reproduce many observed features of earthquake catalogs. The reason for including a brief section on such models in the present article is that one of their underlying purposes is to demonstrate that deterministic models of sufficient complexity can exhibit just that random-appearing behavior that characterizes the appearance of earthquake data. Two important issues then arise. The first is whether the mechanical models can be matched closely enough to any real seismic system to produce forecasts similar to or better than those produced by the stochastic models. The second is whether those stochastic models currently fitted to earthquake catalogs also provide good descriptions of the catalogs produced by the mechanical models. Such studies may then suggest ways in which the stochastic models themselves can be improved.

In the pioneering article [14], the mechanical system comprises a series of blocks, linearly connected by springs, which are pulled over a rough plane. The resultant movement is not smooth, but jerky, as the tension in one of the springs builds up to the point where it overcomes the frictional forces opposing motion and a slip occurs. Often movement of one block will initiate movement in a whole set of blocks.

The population of such movement sequences is then compared to a population of earthquakes. Although the motion of the system is entirely deterministic, analytic solutions are too complex to obtain in explicit form, and the behavior is irregular, with features of (pseudo)randomness. In particular a crude form of GR law generally results.

Many further mechanical models have been invented and studied, both in the laboratory and via numerical solution of the appropriate dynamical equations. Often the masses of the blocks or the spring parameters are selected at random from some plausible overall population. It is not this randomness which is the cause of the random-looking behavior, however, but rather the complexity of the movements (trajectory) of any particular system of blocks and springs or other similar elements.

One of the difficulties in matching such models to particular fault systems is to match the initial conditions, which are not easy to determine for real processes, even if they can be reproduced in the artificial system. It is precisely the relative simplicity of the stochastic model, which comes at the expense of describing many important physical details, which allows it to be fitted to catalogue data and then used to produce rough probabilistic forecasts.

In [70], one of the relatively few studies to fit point process models to the output from such mechanical systems, outputs were taken from the four different models described in [6] (see also [10]) and fitted by the simple stress release model. The four models illustrated different patterns of behavior, varying from highly random behavior with typical GR distributions, to regular, characteristic earthquake behavior. After suitable adjustment of the fre-

quency-magnitude law in the stress-release model, it was found that the stress release model could be fitted to all four versions, and used to describe the overall energy state of the mechanical system; as such it operated as a crude predictor of the next major event in the system.

Future Directions

There are many possible directions in which the applications of stochastic models in seismology could be extended and deepened. The fundamental limitations in the past have most commonly been limitations in data, and consequent limitations in the physical understanding of the processes. It has been the great improvement in data which has allowed the development of better and more insightful seismicity models in the last two decades; this tendency is continuing strongly and may well lead to unexpected new developments.

Two directions in particular appear to me to hold out scope for further development of stochastic models for earthquake occurrence. The first is related to the collection and integration of data on earth deformation. The extensive data now becoming available from GPS measurements have already led to new discoveries, for example 'slow earthquakes' which relieve strain without being registered on conventional seismometers. The underlying problem is how to link the data on strain to data on seismicity. This is not easy, and is likely to require new ideas on both the physical and statistical sides.

The second relates to the systematic collection and evaluation of data on potential precursor events. Many of the ideas initiated in the 1970's have been abandoned, with the result that for most potential phenomena there exists no substantial body of data by which their effectiveness can be adequately tested, or the underlying physical processes modeled. Many controversial and so far inadequately explained phenomena fall into this category. They have the potential to generate projects of interest and importance from the physical as well as the hazard estimation viewpoints. Unfortunately collecting and archiving such data is a long-term process with uncertain future outcomes, and therefore difficult to fund under current funding criteria. The balance of scientific, public and even government opinion may change, however, and further work in these fields may be anticipated.

Fundamental work on earthquake mechanism – still a largely unsolved problem – is also likely to attract attention during the next decades, and to require a combination of physical and statistical modeling.

If these are rather long-term developments, there are many smaller scale, more immediate problems that require further investigation. On the theoretical side, one important issue is the development of improved models and procedures for analyzing data with self-similar characteristics. Another area where further research is needed, particularly in subduction regions, is in the development of better physical and statistical models for deep earthquakes, including their possible links to different forms of activity (seismic, volcanic etc) closer to the surface.

The current interest in developing testing centers for probability forecasting is likely to promote more strenuous efforts to develop improved forecasting models, whether based just on catalogue data, or allied with deformation data, or with data on other precursory phenomena. Even the existing models are capable of producing time-varying forecasts which in principle could lead to significant reductions in earthquake risk. The problem here is how to realize these reductions in practice, for example through improved insurance or disaster mitigation activities. Many questions arise, of an operations research as much as a statistical character, which warrant further study and effort. The probability forecasts themselves, if they are to be useful in such contexts, need to incorporate the uncertainties in the underlying models and hence to take on a more explicitly Bayesian character, and to be presented in such a way that they can be related to the many additional factors that have to be borne in mind when making real-life decisions.

Acknowledgments

I am very grateful to friends and colleagues, especially David Harte, Mark Bebbington, David Rhoades and Yehuda Ben-Zion, for helpful discussions, correcting errors and plugging gaps.

Bibliography

1. Ambraseys NN, Melville CP (1982) A History of Persian Earthquakes. Cambridge University Press, Cambridge
2. Andersen PK, Borgan Ø, Gill RD, Keiding N (1993) Statistical Models Based on Counting Processes. Springer, New York
3. Bak P, Tang C (1989) Earthquakes as a self-organized critical phenomenon. J Geophys Res 94:15635–15637
4. Bebbington M, Harte DS (2003) The linked stress release model for spatio-temporal seismicity: formulations, procedures and applications. Geophys J Int 154:925–946
5. Bebbington M, Vere-Jones D, Zheng X (1990) Percolation theory: a model for earthquake faulting? Geophys J Int 100:215–220
6. Ben-Zion Y (1996) Stress, Slip and earthquakes in models of complex single-fault systems incorporating brittle and creep deformations. J Geophys Res 101:5677–5706
7. Ben-Zion Y, Dahmen K, Lyakhowsky V, Ertas D, Agnon A (1999) Self-driven mode-switching of earthquake activity on a fault system. Earth Planet. Sci Lett 172:11–21

8. Ben-Zion Y, Eneva M, Liu Y (2003) Large earthquake cycles and intermittent criticality on heterogeneous faults due to evolving stress and seismicity. J Geophys Res 108:2307V. doi:10.1029/2002JB002121
9. Ben-Zion Y, Lyakhovsky V (2002) Accelerated seismic release and related aspects of seismicity patterns on earthquake faults. Pure Appl Geophys 159:2385–2412
10. Ben-Zion Y, Rice J (1995) Slip patterns and earthquake populations along different classes of faults on elastic solids. J Geophys Res 100:12959–12983
11. Borovkov K, Vere-Jones D (2000) Explicit formulae for stationary distributions of stress release processes. J Appl Prob 37:315–321
12. Brémaud P, Massoulié L (2001) Hawkes branching processes without ancestors. J App Prob 38:122–135
13. Brillinger DR (1981) Time Series: Data Analysis and Theory, 2nd edn. Holden Day, San Francisco
14. Burridge R, Knopoff L (1967) Model and theoretical seismicity. Bull Seismol Soc Am 57:341–371
15. Chelidze TL, Kolesnikov YM (1983) Modelling and forecasting the failure process in the framework of percolation theory. Izvestiya Earth Phys 19:347–354
16. Chong FS (1983) Time-space-magnitude interdependence of upper crustal earthquakes in the main seismic region of New Zealand. J Geol Geophys 26:7–24, New Zealand
17. Console R, Lombardi AM, Murru M, Rhoades DA (2003) Båth's Law and the self-similarity of earthquakes. J Geophys Res 108(B2):2128V. doi:10.1029/2001JB001651
18. Cox DR (1972) Regression models and life tables (with discussion). Roy J Stat Soc Ser B 34:187–220
19. Dahmen K, Ertas D, Ben-Zion Y (1998) Gutenberg-Richter and characteristic earthquake behavior in simple mean-field models of heterogeneous faults. Phys Rev E 58:1494–1501
20. Daley DJ, Vere-Jones D (2003) An Introduction to the Theory of Point Processes, 2nd edn, vol I. Springer, New York
21. Davison C (1938) Studies on the Periodicity of Earthquakes. Murthy, London
22. Diggle PJ (2003) Statistical Analysis of Spatial Point Patterns. 2nd edn. University Press, Oxford
23. Ebel JB, Chambers DW, Kafka AL, Baglivo JA (2007) Non-Poissonian earthquake clustering and the hidden Markov model as bases for earthquake forecasting in California. Seismol Res Lett 78:57–65
24. Evison F, Rhoades D (2001) Model of long-term seismogenesis. Annali Geofisica 44:81–93
25. Felzer KR, Abercrombie RE, Ekström G (2004) A common origin for aftershocks, foreshocks and multiplets. Bull Amer Seismol Soc 94:88–98
26. Fisher RL, Dahmen K, Ramanathan S, Ben-Zion Y (1997) Statistics of earthquakes in simple models of heterogeneous faults. Phys Rev Lett 97:4885–4888
27. Griffiths AA (1924) Theory of rupture. In: Proceedings 1st Int Congress in Applied Mech, Delft, pp 55–63
28. Gutenberg B, Richter C (1949) Seismicity of the Earth and Associated Phenomena, 2nd edn. University Press, Princeton
29. Habermann RE (1987) Man-made changes of seismicity rates. Bull Seismol Soc Am 77(1):141–159
30. Hainzl S, Ogata Y (2005) Detecting fluid signals in seismicity data through statistical earthquake modelling. J Geophys Res 110. doi:10.1029/2004JB003247
31. Harte D (2001) Multifractals: Theory and Applications. Chapman and Hall/CRC, Boca Raton
32. Harte D, Li DF, Vreede M, Vere-Jones D (2003) Quantifying the M8 prediction algorithm: reduction to a single critical variable and stability results. NZ J Geol Geophys 46:141–152
33. Harte D, Li D-F, Vere-Jones D, Vreede M, Wang Q (2007) Quantifying the M8 prediction algorithm II: model, forecast and evaluation. NZ J Geol Geophys 50:117–130
34. Harte D, Vere-Jones D (2005) The entropy score and its uses in earthquake forecasting. Pure Appl Geophys 162:1229–1253
35. Hawkes AG (1971) Spectra of some self-exciting and mutually exciting point processes. Biometrika 58:83–90
36. Hawkes AG, Oakes D (1974) A cluster representation of a self-exciting process. J Appl Prob 11:493–503
37. Helmstetter A, Sornette D (2002) Subcritical and supercritical regimes in epidemic models of earthquake aftershocks. J Geophys Res 107:2237. doi:10.1029/2001JB001580
38. Helmstetter A, Sornette D (2003) Båth's law derived from the Gutenberg-Richter law and from aftershock properties. Geophys Res Lett 103(20):2069. doi:10.1029/2003GL018186
39. Ishimoto M, Iida K (1939) Bull Earthq Res Inst Univ Tokyo 17:443–478
40. Iwata T, Young RP (2005) Tidal stress/strain and the *b*-values of acoustic emissions at the Underground Research Laboratory. Canada. Pure Appl Geophys 162:(6–7):1291–1308. doi:10.1007/s00024-005-2670-2 (P*1357)
41. Jackson DD, Kagan YY (1999) Testable earthquake forecasts for 1999. Seismol Res Lett 70:393–403
42. Jaeger JC, Cook NGW (1969) Fundamentals of Rock Mechanics. Methuen, London
43. Jaume SC, Bebbington MS (2004) Accelerating seismic moment release from a self-correcting stochastic model. J Geophys Res 109:B12301. doi:10.1029/2003JB002867
44. Jeffreys H (1938) Aftershocks and periodicity in earthquakes. Beitr Geophys 53:111–139
45. Jeffreys H (1939) Theory of Probability, 1st edn (1939), 3rd edn (1961). University Press, Cambridge
46. Jones LM, Molnar P (1979) Some characteristics of foreshocks and their possible relationship to earthquake prediction and premonitory slip on a fault. J Geophys Res 84:3596–3608
47. Kagan Y (1973) Statistical methods in the study of the seismic process. Bull Int Stat Inst 45(3):437–453
48. Kagan Y (1991) Seismic moment distribution. Geophys J Int 106:121–134
49. Kagan Y (1991) Fractal dimension of brittle fracture. J Nonlinear Sci 1:1–16
50. Kagan Y (1994) Statistics of characteristic earthquakes. Bull Seismol Soc Am 83:7–24
51. Kagan Y, Jackson DD (1994) Probabilistic forecasting of earthquakes. Geophys J Int 143:438–453
52. Kagan Y, Knopoff L (1977) Earthquake risk prediction as a stochastic process. Phys Earth Planet Inter 14:97–108
53. Kagan Y, Knopoff L (1980) Spatial distribution of earthquakes: the two-point correlation function. Geophys J Roy Astronom Soc 62:303–320
54. Kagan Y, Knopoff L (1981) Stochastic synthesis of earthquake catalogues. J Geophys Res 86:2853–2862
55. Kagan Y, Knopoff L (1987) Statistical short-term earthquake prediction. Sci 236:1563–1567

56. Keilis-Borok VI, Kossobokov VG (1990) Premonitory activation of the earthquake flow: algorithm M8. Phys Earth Planet Inter 61:73–83
57. Kiremidjian AS, Anagnos T (1984) Stochastic slip predictable models for earthquake occurrences. Bull Seismol Soc Am 74:739–755
58. Knopoff L (1971) A stochastic model for the occurrence of main sequence earthquakes. Rev Geophys Space Phys 9:175–188
59. Kossobokov VG (1997) User manual for M8. In: Algorithms for Earthquake Statistics and Prediction. IASPEI Softw Ser 6:167–221
60. Kossobokov VG (2005) Earthquake prediction: principles, implementation, perspectives. Part I of Computational Seismology 36, "Earthquake Prediction and Geodynamic Processes." (In Russian)
61. Kossobokov VG (2006) Testing earthquake prediction methods: The West Pacific short-term forecast of earthquakes with magnitude MwHRV $\geq$ 5.8. Tectonophysics 413:25–31
62. Libicki E, Ben-Zion Y (2005) Stochastic branching models of fault surfaces and estimated fractal dimensions. Pure Appl Geophys 162:1077–1111
63. Lombardi A (2002) Probabilistic interpretation of Båth's law. Ann Geophys 45:455–472
64. Lomnitz CA (1974) Plate Tectonics and Earthquake Risk. Elsevier, Amsterdam
65. Lomnitz-Adler J (1985) Asperity models and characteristic earthquakes Geophys. J Roy Astron Soc 83:435–450
66. Lomnitz-Adler J (1985) Automaton models of seismic fracture: constraints imposed by the frequency-magnitude relation. J Geophys Res 95:491–501
67. Lomnitz-Adler J (1988) The theoretical seismicity of asperity models; an application to the coast of Oaxaca. Geophys J 95:491–501
68. Liu J, Chen Y, Shi Y, Vere-Jones D (1999) Coupled stress release model for time dependent earthquakes. Pure Appl Geophys 155:649–667
69. Loève M (1977) Probability Theory I, 4th edn. Springer, New York
70. Lu C, Vere-Jones D (2001) Statistical analysis of synthetic earthquake catalogs generated by models with various levels of fault zone disorder. J Geophys Res 106:11115–11125
71. Lu C, Harte D, Bebbington M (1999) A linked stress release model for Japanese historical earthquakes: coupling among major seismic regions. Earth Planet. Science 51:907–916
72. Macdonald II, Zucchini W (1997) Hidden Markov and Other Models for Discrete-Valued Time Series. Chapman and Hall, London
73. Main IG, Burton PW (1984) Information theory and the earthquake frequency-magnitude distribution. Bull Seismol Soc Am 74:1409–1426
74. Mandelbrot BB (1977) Fractals: Form, Chance and Dimension. Freeman, San Francisco
75. Mandelbrot BB (1989) Multifractal measures, especially for the geophysicist. Pure Appl Geophys 131:5–42
76. Martínez VJ, Saar E (2002) Statistics of the Galaxy Distribution. Chapman & Hall/CRC, Boca Raton
77. Matsu'ura RS (1986) Precursory quiescence and recovery of aftershock activities before some large aftershocks. Bull Earthq Res Inst Tokyo 61:1–65
78. Matsu'ura RS, Karakama I (2005) A point process analysis of the Matsushiro earthquake swarm sequence: the effect of water on earthquake occurrence. Pure Appl Geophys 162 1319–1345. doi:10.1007/s00024-005-2762-0
79. Matthews MV, Ellsworth WL, Reasenberg PA (2002) A Brownian model for recurrent earthquakes. Bull Seism Soc Amer 92:2232–2250
80. Merrifield A, Savage MK, Vere-Jones D (2004) Geographical distributions of prospective foreshock probabilities in New Zealand. J Geol Geophys 47:327–339, New Zealand
81. Michael A (1997) Test prediction methods: earthquake clustering versus the Poisson model. Geophys Res Lett 24:1891–1894
82. Mogi K (1962) Study of elastic shocks caused by the fracture of heterogeneous materials and its relation to earthquake phenomena. Bull Earthq Res Inst Tokyo Univ 40:125–173
83. Mogi K (1985) Earthquake Prediction. Academic Press, Tokyo
84. Molchan GM (1990) Strategies in strong earthquake prediction. Phys Earth Plan Int 61:84–98
85. Molchan GM, Kagan YY (1992) Earthquake prediction and its optimization. J Geophys Res 106:4823–4838
86. Ogata Y (1988) Statistical models for earthquake occurrence and residual analysis for point processes. J Amer Stat Soc 83:9–27
87. Ogata Y (1998) Space-time point process models for earthquake occurrences. Annals Inst Stat Math 50:379–402
88. Ogata Y (1999) Estimating the hazard of rupture using uncertain occurrence times of paleoearthquakes. J Geophys Res 104:17995–18014
89. Ogata Y (2005) Detection of anomalous seismicity as a stress change sensor. J Geophys Res 110(B5):B05S06. doi:10.1029/2004JB003245
90. Ogata Y, Utsu T, Katsura K (1996) Statistical discrimination of foreshocks from other earthquake clusters. Geophys J Int 127:17–30
91. Ogata Y, Jones L, Toda S (2003) When and where the aftershock activity was depressed: Contrasting decay patterns of the proximate large earthquakes in southern California. J Geophys Res 108(B6):2318. doi:10.1029/2002JB002009
92. Omori F (1894) On aftershocks of earthquakes. J Coll Sci Imp Acad Tokyo 7:111–200
93. Otsuka M (1972) A chain reaction type source model as a tool to interpret the magnitude-frequency relation of earthquakes. J Phys Earth 20:35–45
94. Pietavolo A, Rotondi R (2000) Analyzing the interevent time distribution to identify seismicity patterns: a Bayesian nonparametric approach to the multiple change-point problem. Appl Stat 49:543–562
95. Pisarenko DV, Pisarenko VF (1995) Statistical estimation of the correlation dimension. Phys Lett A 197:31–39
96. Reasenberg PA (1999) Foreshock occurrence before large earthquakes. J Geophys Res 104:4755–4768
97. Reasenberg PA, Jones LM (1989) Earthquake hazard after a mainshock in California. Sci 243:1173–1176
98. Reid HF (1911) The elastic-rebound theory of earthquakes. Bull Dept Geol Univ Calif 6:413–444
99. Renyi A (1959) On the dimension and entropy of probability distributions. Acta Math 10:193–215
100. Rhoades DA (2007) Application of the EEPAS model to forecasting earthquakes of moderate magnitude in Southern California. Seismol Res Lett 78:110–115

101. Rhoades DA, Evison FF (2004) Long-range earthquake forecasting with every event a precursor according to scale. Pure Appl Geophys 161:147–171
102. Rhoades DA, Evison FF (2005) Test of the EEPAS forecasting model on the Japan earthquake catalogue. Pure Appl Geophys 162:1271–1290
103. Rhoades DA, Van Dissen RJ (2003) Estimation of the time-varying hazard of rupture of the Alpine Fault of New Zealand, allowing for uncertainties. NZ J Geol Geophys 40:479–488
104. Ripley BD (1988) Statistical Inference for Spatial Processes. University Press, Cambridge
105. Robinson R (2000) A test of the precursory accelerating moment release model on some recent New Zealand earthquakes. Geophys J Int 140:568–576. doi:10.1046/j.1365-246X2000.00054.x
106. Robinson R, Benites R (1995) Synthetic seismicity models for the Wellington region of New Zealand: implications for the temporal distribution of large events. J Geophys Res 100:18229–18238. doi:10.1029/95JB01569
107. Rundle JB, Klein W, Tiampo K, Gross S (2000) Dynamics of seismicity patterns in systems of earthquake faults. In: Geocomplexity and the Physics of Earthquakes. Geophysical Monograph 120, American Geophysical Union
108. Saito M, Kikuchi M, Kudo M (1973) An analytical solution of: Go-game model of earthquakes. Zishin 26:19–25
109. Scholz CH (1968) The frequency-magnitude relation of microfaulting in rock and its relation to earthquakes. Bull Seism Soc Am 58:399–415
110. Scholz CH (1990) The Mechanics of Earthquakes and Faulting. Cambridge University Press, New York
111. Schorlemmer D, Gerstenberger MC, Wiemer S, Jackson DD, Rhoades DA (2007) Earthquake likelihood model testing. Seismol Res Lett 78:17–29
112. Schuster A (1897) On lunar and solar periodicities of earthquakes. Proc Roy Soc London 61:455–465
113. Schwartz DP, Coppersmith K (1984) Fault behavior and characteristic earthquakes: examples from the Wasatch and San Andreas Faults. J Geophys Res 89:5681–5698
114. Shi YL, Liu J, Chen Y, Vere-Jones D (1999) Coupled stress release models for time-dependent seismicity. J Pure Appl Geophys 155:649–667
115. Shi Y, Liu J, Zhang G (2001) An evaluation of Chinese annual earthquake predictions, 1990–1998. J Appl Prob 38A:222–231
116. Shimazaki K, Nakata T (1980) Time-predictable recurrence model for large earthquakes. Geophys Res Lett 7:179–282
117. Smith WD (1986) Evidence for precursory changes in the frequency-magnitude *b*-value. Geophys J Roy Astron Soc 86:815–838
118. Smith WD (1998) Resolution and significance assessment of precursory changes in mean earthquake magnitude. Geophys J Int 135:515–522
119. Stoyan D, Stoyan H (1994) Fractals, Random Shapes and Point Fields. Wiley, Chichester
120. Tiampo KF, Rundle JB, Klein W, Ben-Zion Y, McGinnis SA (2004) Using eigenpattern analysis to constrain seasonal signals in Southern California. Pure Appl Geophys 16:19–10, 1991 V2003. doi:10.1007/s00024-004-2545-y
121. Turcotte DL (1992) Fractals and Chaos in Geology and Geophysics. Cambridge University Press, Cambridge
122. Utsu T (1961) A statistical study on the properties of aftershocks. Geophys Mag 30:521–605
123. Utsu T, Ogata Y (1997) IASPEI Softw Libr 6:13–94
124. Utsu T, Ogata Y, Matu'ura RS (1995) The centenary of the Omori formula for a decay law of aftershock activity. J Phys Earth 43:1–33
125. Vere-Jones D (1969) A note on the statistical interpretation of Båth's law. Bull Seismol Soc Amer 59:1535–1541
126. Vere-Jones D (1970) Stochastic models for earthquake occurrence. J Roy Stat Soc B 32:1–62
127. Vere-Jones D (1977) Statistical theories for crack propagation. Pure Appl Geophys 114:711–726
128. Vere-Jones D (1978) Space-time correlations of microearthquakes – a pilot study. Adv App Prob 10:73–87, supplement
129. Vere-Jones D (1978) Earthquake prediction: a statistician's view. J Phys Earth 26:129–146
130. Vere-Jones D (1995) Forecasting earthquakes and earthquake risk. Int J Forecast 11:503–538
131. Vere-Jones D (1999) On the fractal dimension of point patterns. Adv Appl Prob 31:643–663
132. Vere-Jones D (2003) A class of self-similar random measures. Adv Appl Prob 37:908–914
133. Vere-Jones D, Davies RB (1966) A statistical analysis of earthquakes in the main seismic region of New Zealand. J Geol Geophys 9:251–284
134. Vere-Jones D, Ozaki T (1982) Some examples of statistical inference applied to earthquake data. Ann Inst Stat Math 34:189–207
135. Vere-Jones D, Robinson R, Yang W (2001) Remarks on the accelerated moment release model for earthquake forecasting: problems of simulation and estimation. Geophys J Int 144:515–531
136. von Bortkiewicz L (1898) Das Gesetz der kleinen Zahlen. Teubner, Leipzig
137. Weibull W (1939) A statistical theory of the strength of materials. Ingvetensk Akad Handl no 151
138. Working Group on Californian Earthquake Probabilities (1990) Probabilities of earthquakes in the San Francisco Bay region of California. US Geological Survey Circular 153
139. Yin X, Yin C (1994) The precursor of instability for non-linear systems and its application to the case of earthquake prediction – the load-unload response ratio theory. In: Newman WI, Gabrielov AM (eds) Nonlinear dynamics and Predictability of Natural Phenomena. AGU Geophysical Monograph 85:55–66
140. Zheng X, Vere-Jones D (1994) Further applications of the stress release model to historical earthquake data. Tectonophysics 229:101–121
141. Zhuang J (2000) Statistical modelling of seismicity patterns before and after the 1990 Oct 5 Cape Palliser earthquake, New Zealand. NZ J Geol Geophys 43:447–460
142. Zhuang J, Yin X (2000) The random distribution of the loading and unloading response ratio under the assumptions of the Poisson model. Earthq Res China 14:38–48
143. Zhuang J, Ogata Y, Vere-Jones D (2004) Analyzing earthquake clustering features by using stochastic reconstruction. J Geophys Res 109(B5):B05301. doi:10.1029/2003JB002879
144. Zhuang J, Vere-Jones D, Guan H, Ogata Y, Ma L (2005) Preliminary analysis of precursory information in the observations on the ultra low frequency electric field in the Beijing region. Pure Appl Geophys 162:1367–1396. doi:10.10007/s00024-004-2674-3

Earthquake Scaling Laws

Raul Madariaga
Ecole Normale Superieure, Laboratoire de Géologie, Paris, France

Article Outline

Glossary
Definition of the Subject
Introduction
Earthquakes and Seismic Radiation
Earthquake Fault Models:
The Scaling of Geometry and Stress
Earthquake Dynamics and the Scaling of Energy
Kinematics and Statistical Models for Fault Slip
Future Directions
Acknowledgments
Bibliography

Glossary

Seismic moment The most fundamental measure of the size of an earthquake. In the simplest situation it represents the moment of on of the couples of forces that make up a dipolar source. In more general cases it is a 3 by 3 symmetric tensor of elementary force couples.

Seismic radiation The seismic waves emitted by a seismic source. For point sources these are spherical P and S waves emitted by the point tensor source.

Seismic spectrum The absolute value of the Fourier transform of the displacement field radiated by an earthquake in the far field. For almost all earthquakes it has a common shape: flat at low frequencies and decays like the inverse squared power at high very high values of frequency.

Corner frequency The low and high frequency asymptotes of the earthquake spectrum intersect at a characteristic frequency, called the corner frequency. The corner frequency scales with the size of the earthquake measured by the seismic moment.

Radiated or seismic energy Total energy of the seismic waves radiated by a seismic source. It can be computed from the energy flow relatively far from the source of the earthquake.

Apparent stress Originally defined as the product of seismic efficiency times the average stress during earthquake slip. In practice, it is computed from the ratio of radiated energy to moment release of the earthquake multiplied by the shear modulus.

Energy release rate Amount of energy per unit surface used to make a rupture advance by a unit distance.

Static stress drop The static change in shear traction between the sides of the fault occurs during an earthquake. In principle, it could be determined by measuring stress before and after the earthquake. In practice stress drop is computed using very specific source models, like a circular crack.

Dynamic stress drop The stress change in shear traction as a function of time while the rupture is still growing. It can only be estimated from seismic records obtained in the near field by elaborate inversion schemes. The relation between static and dynamic stress drop can only be estimated once the friction law between the sides of the fault has been defined.

Definition of the Subject

Earthquake scaling laws provide some of the most basic knowledge about seismic sources. Since the end of the 70s, a very successful model for earthquakes was developed by seismologists. In this model earthquakes are due to rapid slip on pre-existing faults driven by steady load due to plate motion and resisted by friction between the fault walls. This model may be used to predict many of the general properties of seismic radiation that can be derived from a simple spectral shape of type omega-squared. In this article I derive general expressions for energy, moment and stress in terms of measured spectral parameters. The available data shows that earthquakes can be reduced to a single family in terms of three parameters: moment, corner frequency and radiated energy. Using specific models of rupture these three parameters can be reinterpreted in terms of moment, size and stress drop. Although details differ between the models proposed by seismologists, both seismic spectra and the wave-number spectra of slip distributions can be explained with a simple circular crack model. This does not mean that a circular crack is the best earthquake model; it means that the ensemble average of seismic sources has properties that are similar to those of simple circular shear cracks. A direct result of scaling laws is that total fracture energy must scale like radiated and strain release energy, so that fracture energy should scale with fault size as observed for many earthquakes and in certain laboratory experiments.

Introduction

It has been 40 years since Aki [4] published his seminal paper on the scaling law of earthquakes that established that to first order seismic moment scales like the third power of the fault size. This paper came just a year after Aki [3]

made the very first measurement of seismic moment, the torque of one of the couples that make up a basic source mechanism. Not much later, in 1970, Brune introduced a very simple source model and established a generalization of Aki's [4] spectral model of earthquakes that is now known as the omega-squared model of earthquakes. Almost simultaneously, Kostrov [43], Savage [66], Sato and Hirasawa [65] and Madariaga [46] developed models of a circular faults. Radiation from a circular crack explained well the omega-squared model. Digital data was not available at the time when these models were introduced. It is nowadays an almost standard observatory practice to study the scaling law, but as many recent studies have shown, proper estimation of spectral parameters like moment and corner frequency and high frequency decay for a large range of earthquake sizes is not trivial and it is often difficult to obtain from a single instrument. In the 1990s good quality digital data, both broad band seismograms and accelerograms, became available not just from surface instruments, but also from boreholes opening the way to a new appraisal of the scaling law.

We will briefly review some of these observations and we will try to establish general properties of earthquake radiation based on the recent work by McGarr [53], Abercrombie and Rice [2], Ide et al. [35], Prieto et al. [59], etc. Our purpose is not to review the very extensive literature on the determination of seismic source parameters; most of those papers assume specific scaling laws like circular cracks, Brune's relation between corner frequency and earthquake size, etc. Our purpose here is to derive the scaling law from same basic physical concepts and to test the validity of some common assumptions in seismology. Because different authors are interested in certain specific aspects of earthquake sources, often the data from different authors is hard to combine. Some authors use the Brune's [17] empirical model as a basis for the scaling law; others use the radiation from quasidynamic circular cracks as a model leading to substantial variations. Even more serious differences come from data processing, some authors using specific attenuation corrections in order to correct for Earth's Q, others use small events as empirical Green functions, etc. These corrections are very important, but they are beyond our goal which is to try to extract information about scaling and its inferences for earthquake physics. Two aspects will be particularly discussed: how to make model independent estimates of source parameters and how to establish model independent scaling laws. Some recent evidence indicates, for instance, that stress change during earthquakes as measured by apparent stress varies independently of Moment and size, so that earthquakes are probably quantified by three indendent parameters that need to be carefully chosen. The other will be an aspect that is often overlooked in the literature: Brune's [17] model, as well as Madariaga's [46] circular crack model make very specific statements about the partition of seismic energy at the source. In particular, the scaling law implies that fracture energy is not a constant but that it scales with earthquake size in a manner that was predicted on the basis of fracture dynamics [46]; friction experiments by Ohnaka and Shen [56] and Ohnaka [55]; arguments about scaling by McGarr and Fletcher [54], and dynamic seismic source inversions [34,57].

Earthquakes and Seismic Radiation

It is now well established that earthquakes are due to faulting and that the simplest way to measure them is to use the seismic moment M_0, introduced by Aki [3], and given by

$$M_0 = \mu \, \overline{D} \, S \tag{1}$$

where μ is the shear or rigidity modulus of the material surrounding the fault, $\overline{D}$ is the mean value of the final slip on the fault and S is the area of the fault rupture. M_0 is the moment of one of the couples that constitute a double couple, the simplest possible model of a point-like earthquake. Radiation from a point double couple source has been completely solved by seismologists and is thus the natural starting point for the development of earthquake scaling laws.

The far-field displacement $\boldsymbol{u}_c$ radiated by a point double couple source can be written in the following form:

$$\boldsymbol{u}_c(r,t) = \frac{1}{4\pi\rho c^3}\frac{1}{R}\,\boldsymbol{e}_c^{\mathrm{T}} \cdot \dot{\mathbf{M}}_0 \cdot \boldsymbol{e}_{\mathrm{R}}(t - R/c)\,, \tag{2}$$

(see, e. g. [5]) where the subscript c stands for P or S waves, i. e. c is either α, the P wave speed, or β, the S wave speed; ρ is the density; R is the distance of the observation point from the source. $\boldsymbol{e}_{\mathrm{R}}$ is the radial unit vector in the direction of radiation. The unit vector $\boldsymbol{e}_c$ is the polarization of the wave, that is $\boldsymbol{e}_c = \boldsymbol{e}_{\mathrm{R}}$ for P waves, or $\boldsymbol{e}_{\mathrm{T}}$ the appropriate transverse unit vector for SH or SV waves. $\mathbf{M}_0(t)$ is the moment tensor of the source, a symmetric tensor of order three that describes the geometry and amplitude of the seismic source (see [5] for details).

Very often in seismology it is assumed that the geometry of the source can be separated from its time variation, so that the moment tensor can be written in the simpler form:

$$\mathbf{M}_0(t) = \mathbf{I}_0 M_0 s(t) \tag{3}$$

where $\mathbf{I}_0$ is a time-invariant tensor that describes the orientation of the source, M_0 is the scalar moment tensor of

the source, and $s(t)$ is the time variation of the moment, the source time function determined by seismologists. In the following we assume that $s(t)$ is causal, i. e. it is zero up to $t = 0$, and normalized so that

$$\int_0^\infty s(t)\mathrm{d}t = 1 \,.$$

Using (3) we can now write a simpler form of (2):

$$u_c(r,t) = \frac{1}{4\pi\rho c^3}\frac{\mathcal{R}_c}{R}\,\Omega(t - R/c)\,. \tag{4}$$

For P waves, u_c is the radial component; for S waves, it is the appropriate transverse component for SH or SV waves. In (4) we have introduced the standard notation $\Omega(t) = M_0\mathrm{d}s(t)/\mathrm{d}t$ for the source time function, the signal emitted by the source as seen in the far field. The term $\mathcal{R}_c(\theta,\phi)$ is the radiation pattern, a function of the takeoff direction of the ray from the source. In a spherical coordinate system (R,θ,ϕ) centered at the source, the radiation patterns are given by Aki and Richards [5]. We simply quote here the case of a so-called strike-slip earthquake where the fault is vertical, normal to the y axis, and slip is parallel to the x axis, so that only the $M_{xy} = M_{yx}$ components of the moment tensor are different from zero. In this case

$$\mathcal{R}_\mathrm{P} = \sin^2\theta\,\sin 2\phi\,,$$
$$\mathcal{R}_\mathrm{SV} = \frac{1}{2}\sin 2\theta\,\sin 2\phi\,,$$
$$\mathcal{R}_\mathrm{SH} = \sin\theta\,\cos 2\phi\,.$$

On the $z = 0$ plane ($\theta = \pi/2$), there are no SV waves. On the other hand, on this plane, the radiation patterns of P and SH waves have typical quadrupole distributions proportional to $\sin 2\phi$ and $\cos 2\phi$ respectively.

Spectral Domain Approach

At high frequencies, the signals radiated by earthquakes may become quite complex because of multipathing, scattering, etc., so that the actually observed seismogram $u(t)$ resembles the source time function $\Omega(t)$ only at long periods. It is usually verified that complexities in the wave propagation affect much more the phase of seismic waves than the spectral amplitudes in the Fourier transformed domain. The spectral domain approach was introduced by Aki [4], Wyss and Brune [73] and Brune [17]. Radiation from a simple point moment-tensor source can be obtained from (4) by Fourier transformation. Displacement pulses radiated from a point moment tensor in the Fourier transformed domain is then

$$u_c(r,\omega) = \frac{1}{4\pi\rho c^3}\,\frac{\mathcal{R}_c}{R}\,\tilde{\Omega}(\omega)\mathrm{e}^{-i\omega R/c}\,. \tag{5}$$

Where $\tilde{\Omega}(\omega)$ is the Fourier transform of the source time function $\Omega(t)$. A well-known property of the Fourier transform is that

$$\lim_{\omega\to 0}\tilde{\Omega}(\omega) = M_0 * \int_0^\infty \dot{s}(t)\mathrm{d}t = M_0 \tag{6}$$

so that the in the low-frequency limit of the source time function spectrum approaches the scalar moment.

From the observation of many earthquake spectra, and from the computation of magnitudes in different frequency bands, Aki [4] and Brune [17] concluded that the seismic spectra had a universal shape with a flat low frequency asymptote given by (6), a certain characteristic frequency called **corner frequency** by seismologists, and a decay at high frequencies that tends asymptotically to ω^{-2}. I will not repeat here the arguments that led to this model: Aki [4] proposed this spectral shape in order to explain the differences in magnitude determined from seismic waves of different frequencies. Brune's [17] argument was based on simple concepts about the high frequency radiation and the amount of energy radiated by a seismic event. The omega-squared model has now been confirmed by numerous observations and detailed studies of seismic radiation. As an example, Fig. 1 shows the displacement record of a M5 earthquake that occurred inside the Nazca plate, 99 km under Santiago de Chile on 7 January 2003. The spectrum shown at the bottom presents the typical low frequency asymptote proportional to the seismic moment, and high frequency decay proportional to ω^{-2}. The corner frequency is approximately 1.2 Hz for this event.

In its simpler form the ω^{-2} spectrum is [17]:

$$\tilde{\Omega}(\omega) = \frac{M_0}{1 + \omega^2/\omega_0^2} \tag{7}$$

where ω_0 is the corner frequency. Based on considerations about the spectrum of random signals, Boatwright [12] proposed the alternative model:

$$\tilde{\Omega}(\omega) = \frac{M_0}{\left[1 + \omega^4/\omega_0^4\right]^{1/2}}\,. \tag{8}$$

In these simple omega-squared models, seismic sources are characterized by only two independent scalar parameters: the seismic moment M_0 and the corner frequency ω_0.

Brune's model (7) explains well the spectrum of S waves, it describes also that of P waves, but the corner frequencies are different for P and S. For a long time observations were not able to distinguish between the P and S waves spectra. The fist clear observations of the corner frequency ratio were made by Hanks [26]. Recent work by Abercrombie and Rice [2], Prieto et al. [59], Ide et al. [35]

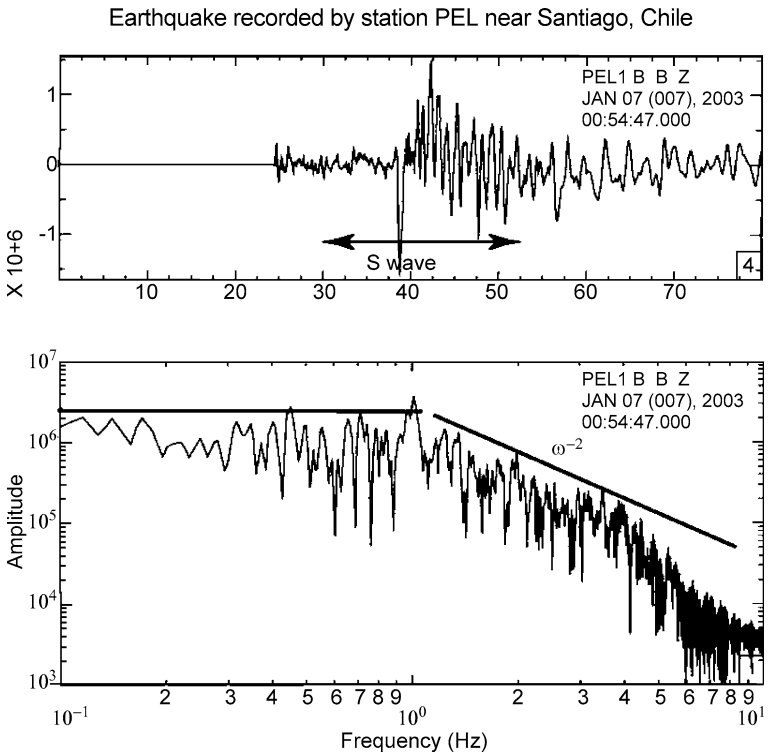

Earthquake Scaling Laws, Figure 1

Example of a seismic signal and its spectrum. Recording at the PEL broad-band station of the GEOSCOPE network of a $M = 5$ intermediate depth earthquake. The event occurred inside the subducted Nazca plate under Santiago, Chile at 99 km depth. The top panel shows the displacement, integrated from the broad band velocity record. The shear wave section used for computing the spectrum is shown with arrows. The bottom panel shows the amplitude Fourier spectrum. The low frequency and high frequency trends are indicated by the thick straight lines. The corner frequency is located at the intersection of the two asymptotes

has confirmed that P spectra have a higher corner frequency than S waves and that very roughly

$$\omega_0^{\mathrm{P}} \cong 1.6\omega_0^{\mathrm{S}} \,. \tag{9}$$

This is very close to $\sqrt{3}$, the ratio of P to S wave speed. This ratio is similar to that predicted for the quasi-dynamic circular crack model [46].

As mentioned earlier, not all earthquakes have displacement spectra as simple as (7), but the omega-squared model is a simple starting point for understanding seismic radiation.

From (7), it is possible to compute the spectra predicted for ground velocity:

$$\dot{\tilde{\Omega}}(\omega) = \frac{i\omega M_0}{1 + \omega^2/\omega_0^2} \,. \tag{10}$$

Ground velocity spectra are characterized by a peak situated roughly at the corner frequency ω_0. In actual earthquake ground velocity spectra, this peak is usually broadened and contains oscillations and secondary peaks, but (10) is a good approximation to the spectra of ground velocity for frequencies lower than a certain cut-off frequency called $f_{\max}$ by Hanks [27], and is close to 6–7 Hz in many areas. At frequencies higher than $f_{\max}$, attenuation, propagation and scattering modify the velocity spectrum.

Seismic Energy Radiated by Point Moment-Tensor Sources

In order to establish the most basic scaling relationship for seismic sources we have to compute the energy radiated by an omega-squared source like (8). This was actually the way Brune originally calibrated his source time function and Aki established the variation of corner frequency with moment. Assuming that the source is embedded in a homogeneous medium, and that the observation point is sufficiently far from the source, the energy flow per unit solid angle, e_{r} is proportional to the square of the particle velocity v_c (see [11,12,17,18,31]), so that the total flow per unit

solid angle is:

$$e_r^c = \rho c R^2 \int_0^\infty v_c^2(t)\mathrm{d}t \tag{11}$$

where ρc is the P or S wave impedance, $v_c(t)$ is the ground velocity and R is again the distance of the observation point from the source. We can compute the radiated energy density replacing $v_c(t)$ by the time derivative of the far field displacement (4), and get:

$$e_r^c = \frac{1}{16\pi^2\rho c^5}\mathcal{R}_c^2 \int_0^\infty \dot{\Omega}^2(t)\mathrm{d}t\,. \tag{12}$$

As expected the energy flow per unit solid angle does not depend on the distance from the source R. We can now apply Parseval's theorem to express the energy in terms of the source spectral amplitude

$$\int_0^\infty \dot{\Omega}^2(t)\mathrm{d}t = \frac{1}{\pi}\int_0^\infty \omega^2 \left|\tilde{\Omega}(\omega)\right|^2 \mathrm{d}\omega$$

and compute the total radiated energy, E_r, for each type of wave. Integrating (12) over the angles θ and ϕ we get

$$E_r^c = \frac{1}{4\pi^2\rho c^5}\langle\mathcal{R}_c^2\rangle \int_0^\infty \omega^2 \left|\tilde{\Omega}(\omega)\right|^2 \mathrm{d}\omega \tag{13}$$

where

$$\langle R_c^2\rangle = \frac{1}{4\pi}\iint_\Omega \mathcal{R}_c^2(\theta,\phi)\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi$$

is the mean-squared radiation pattern. This expression for the total radiated energy is very interesting because it does not depend on any assumption about earthquake dynamics, just on the shape of the spectra.

For the ω-squared model (7) the integral over circular frequency in (13) can be evaluated exactly to $(\pi/4M_0^2\omega_0^3)$, so that the radiated energy is simply

$$E_r^c = \frac{1}{16\pi}\langle\mathcal{R}_c^2\rangle\frac{M_0^2\omega_0^3}{\rho c^5} \tag{14}$$

where we grouped in the last term all the dimensional variables. Let us remark that the numerical factor $1/16\pi$ depends on the particular model assumed for the spectrum near the corner frequency. Thus, for the Boatwright model (8), the coefficient is slightly larger ($\sqrt{2}/16\pi$).

Since radiated energy and moment have the same dimensional units it is customary to rewrite this expression in the non-dimensional form:

$$\frac{E_r^c}{M_0} = \frac{\langle\mathcal{R}c\rangle^2}{16\pi}\frac{M_0}{\rho}\frac{\omega_0^3}{c^5}\,. \tag{15}$$

Very often this expression is written in terms of frequency f_0 instead of the circular frequency ($\omega_0 = 2\pi f_0$), so that

$$\frac{E_r^S}{M_0} = \frac{\pi^2\langle\mathcal{R}_c\rangle^2}{2}\frac{M_0}{\rho}\frac{f_0^3}{c^5}\,. \tag{16}$$

The average radiation patterns are well known, $\langle R_p\rangle^2 = 4/15$ and $\langle R_s\rangle^2 = 6/15$, see, e. g. Haskell [31], so that for S waves most authors use the following expression to quantify the ratio between the S wave radiated energy and the seismic moment

$$\frac{E_r^S}{M_0} = 1.9739\frac{M_0}{\mu}\frac{f_0^3}{\beta^3}\,. \tag{17}$$

Where we used the definition of S wave speed ($\mu = \rho\beta^2$). This non-dimensional relation makes no assumptions about the rupture process at the source except that the spectrum decays like ω^{-2} at high frequencies. Note that the numerical coefficient is smaller by a factor of four from that computed by Singh and Ordaz [69]. The factor of four seems to be a misprint.

Apparent Stress

A very important parameter of seismic sources that can be computed independently of any particular source geometry is the apparent stress. Originally, apparent stress was defined as the product of the seismic efficiency η by the average stress $\overline{\sigma}$ that acts across the fault during the earthquake

$$\sigma_a = \eta\overline{\sigma}\,,$$

efficiency η was in turn defined as the ratio between the radiated energy E_r and the total released energy W. Unfortunately, neither W nor the average stress can be directly inverted from seismic observations because seismic waves have no information about the average stress that acts on the fault. Wyss and Brune [73] proved that for uniform average stress, W could be written as

$$W = \frac{\overline{\sigma}}{\mu}M_0$$

so that apparent stress can be defined as

$$\sigma_a = \frac{\mu E_r^S}{M_0}\,.$$

This expression, originally derived for uniform average stress has become one of the most useful measures of stress

on seismic sources (see, e. g. [53]). Using the expression for radiated energy (17) we get:

$$\sigma_a = \frac{\mu E_r^S}{M_0} = 1.9739 M_0 \frac{f_0^3}{\beta^3} \tag{18}$$

an expression that depends only on three measurable quantities: total S wave radiated energy, seismic moment and corner frequency. Because the energy flow can usually be computed only for those directions where stations are available, (18) can never be evaluated very accurately. This problem still persists at present; in spite of the deployment of increasingly denser instrumental networks, there will always be large areas of the focal sphere that remain outside the domain of seismic observations because the waves in those directions are refracted away from the station networks, energy is dissipated due to attenuation, etc.

Equation (18) shows that energy moment ratio is a non-dimensional number that depends only on observable quantities (E_r, M_0 and f_0) and wave speed. No assumption is made in Eq. (18) about particular models to convert corner frequencies into source dimensions. Let us finally remark that if σ_a is computed from the estimated radiated energy E_r, then the apparent stress may be considered as an independent parameter that may be used to test the relation (18).

Time Domain Approach

In the previous section we approached seismic radiation from the spectral point of view. An alternative way to understand radiation is to approach it from the time domain. The obvious question is what is the time domain signal associated with ω^{-2} spectrum (17)? Brune [17] proposed one of them:

$$\Omega(t) = M_0 \omega_0^2 t e^{-\omega_0 t} \quad \text{for} \quad t > 0 \tag{19}$$

which is a causal function (i. e., it is zero for $t < 0$). It is also normalized so that

$$\int_0^\infty \Omega(t) \mathrm{d}t = M_0 \,.$$

The high frequency content for this function is controlled by the slope discontinuity at the origin. This is however only one of many functions sharing the same spectral amplitude described by (7) and (8), leaving sufficient freedom for variations in the source time function shape.

Kanamori and Rivera [40] defined a function of finite duration T, finite moment and minimum radiated energy. They found such signal by solving a variational problem for fixed signal duration T. The time signal is the parabola defined by

$$\Omega(t) = \frac{6M_0}{T^3} t(T - t) \tag{20}$$

for $0 < t < T$ and 0 elsewhere. The radiated energy for this signal can be computed using the time domain expression (12). Using the integral

$$I_v = \int_0^\infty \dot{\Omega}^2(t) \mathrm{d}t = \frac{12}{T^3} M_0^2$$

we can compute the total radiated energy as a function of T. In order to write the energy in the same form as the frequency domain expressions we determine the corner frequency from the Fourier transform of (20):

$$\tilde{\Omega}(\omega) = \frac{6M_0}{\omega^3 T^3} \left[(\omega T - 2i) + (\omega T + 2i) e^{i\omega T} \right] . \tag{21}$$

At low frequencies this expression tends to M_0, as expected, while at high frequencies it behaves like

$$\lim_{\omega \to \infty} \tilde{\Omega}(\omega) = \frac{12 M_0}{\omega^2 T^2} \frac{\left[1 + e^{i\omega T}\right]}{2} \,.$$

So that its envelope decreases asymptotically as $12 M_0 (\omega T)^{-2}$ that is, it has the same inverse omega-squared behavior as Brune's model (7). The corner frequency for this signal computed from the intersection of the asymptotes is $\omega_0 = \sqrt{12}/T$. Inserting into (17) we get the following energy moment ratio:

$$\frac{E_r^c}{M_0} = \frac{\langle \mathcal{R}_c \rangle^2}{4\sqrt{12}\pi} \frac{M_0}{\rho} \frac{\omega_0^3}{c^5} \,. \tag{22}$$

which has a numerical coefficient that is slightly larger than that of (17). This is apparent contradiction to the method used by Kanamori and Rivera, who derived (20) from the condition that E_r be a minimum for a given moment. The reason these two results are not contradictory is that Brune's signal has infinite duration and therefore the variational principle posed by Kanamori and Rivera does not apply to it. Thus the two expressions give very similar answers, but clearly the signal (20) is not the only one that minimizes radiated energy. It is interesting to observe that all these signals decay like ω^{-2} at high frequencies. The reason is that high frequency radiation in these models is controlled by the slope discontinuities in displacement. In Brune's signal (19) the slope discontinuity is at the origin, while in (20) there are two slope discontinuities at the origin and at $t = T$.

In conclusion, the radiation models proposed by seismologists share the following properties: (1) the amplitude is controlled by the seismic moment, (2) the displacement spectrum decreases like ω^{-2} at high frequencies, (3) the spectral shape has a corner frequency f_0 and (4) the radiated energy to moment ratio satisfies the relation

$$\frac{E_r^c}{M_0} = C_r \frac{M_0}{\rho} \frac{f_0^3}{c^5} , \tag{23}$$

where C_r is a numerical constant on the order of two. Its consequences for energy balance are quite interesting as we shall promptly discuss.

Aki's Scaling Law

From observation of seismic data most authors (for recent data see, e. g., Abercrombie [1], Ide and Beroza [33] Ide et al. [35], McGarr [53]). Abercrombie and Rice [2], Prieto et al. [68] have concluded that apparent stress σ_a is almost independent of moment for most earthquakes. If that is correct, its immediate consequence is that moment scales like the inverse third power of the corner frequency:

$$M_0 \propto f_0^{-3} . \tag{24}$$

If apparent stress is constant, seismic moment is inversely proportional to the cube of the corner frequency. This result

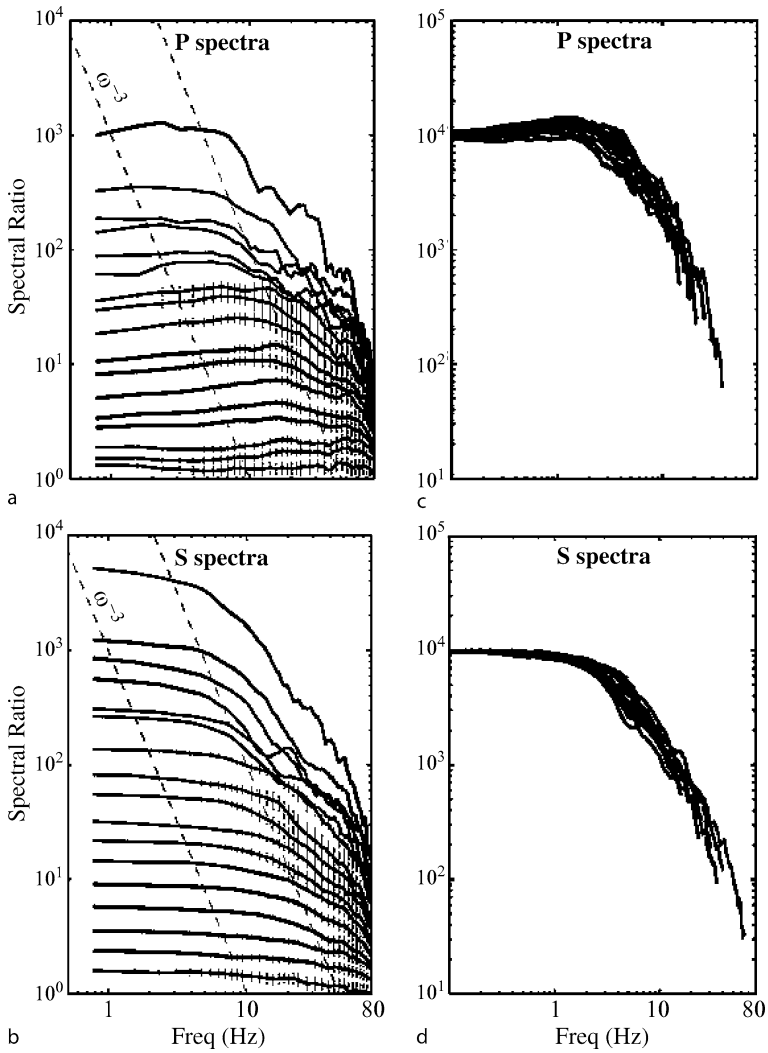

Earthquake Scaling Laws, Figure 2

Scaling of the displacement spectrum in the far field as illustrated by several spectra for different size earthquakes reported by Prieto et al. [59]. Body wave spectra have the typical Brune [17] spectrum and can be collapsed into a single scaling figure by gliding the corner frequencies along an ω^{-3} line

is independent of the particular source model used and can be tested directly from seismic observations. Figure 2 from Prieto et al. [59] illustrates this scaling law. On the left-hand side a series of spectra for different size earthquakes are shown. The corner frequencies align along a line with slope ω^{-3}. Letting all the spectra glide along this line they computed the spectra stack shown at the right-hand side. The same properties are shared by P and S wave spectra. The P wave corner frequencies are higher by a factor of 1.6 than, those of shear waves (9).

Originally, the scaling law (24) was proposed by Aki [4] following a study of spectral data from several collocated earthquakes of different magnitude. Comparing their spectra, he concluded that corner frequencies scaled with seismic moment like (24). This is the so-called **scaling law of seismic spectrum**. This law has been tested by numerous authors with increasingly reliable digital data. Figure 3 shows an example derived from data published by Ide et al. [35] and Abercrombie and Rice [2]. In this Figure I plotted the corner frequency of S waves as measured by the authors as a function of seismic moment. The moment vs. corner frequency plot clearly follows the trend of Eq. (24). The fact that in Fig. 3 moment and corner frequency scale like (24) is often taken as a proof that earthquakes scale with a single parameter: the seismic moment. This is however not sufficient to prove scaling because seismic sources require at least three independent parameters for their quantification. It has been traditional to add an additional model dependent relation in order to derive length, time and stresses from (23). The most common assumption is that the corner frequency is related to the radius of an equivalent circular crack using the frequency radius relation proposed by Brune [17]. Some authors use other similar relations derived from quasidynamic models (e. g. [46,66]. With that assumption, corner frequency can be converted into fault size and stress drop can be derived from moment and source radius.

In order to test whether earthquakes scale with a single parameter, it is necessary to obtain an additional objective measure of seismic sources. The best current candidate is the radiated energy of S waves defined by (13). Originally proposed by Boatwright [11,12,13], estimates of seismic energy have become quite common but they are still difficult to obtain as discussed by many authors including Boatwright and Fletcher [15], Abercrombie [1], McGarr [53], Ide et al. [32] Singh and Ordaz [69], etc. In the following I will test the scaling law of earthquakes by computing the non dimensional ratio $C_r = \mu E_r \beta^3/(M_0 f_0^3)$ from published data.

In Fig. 4 I test expression (18) for the data studied by Ide et al. [35] and Abercrombie and Rice [2]. The figure plots the non dimensional ratio $C_r = \mu E_r \beta^3/M_0 f_0^3$ as a function of seismic moment. According to (23), for strict scaling of seismic sources with a single parameter, this non-dimensional ratio should be a constant, indepen-

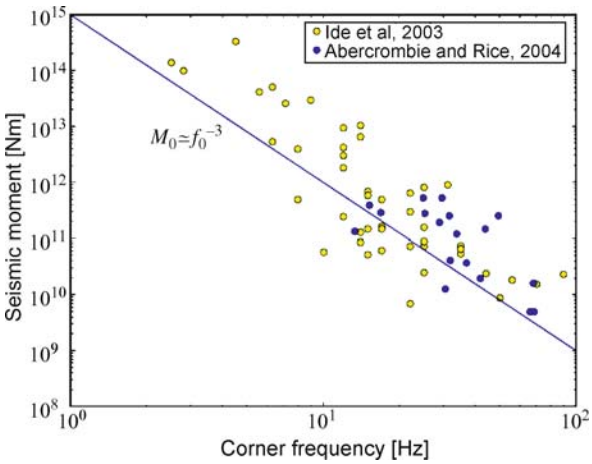

Earthquake Scaling Laws, Figure 3
Self-similarity of earthquake spectra. The figure includes data from two different studies by Ide et al. [32] and by Abercrombie and Rice [2]. These authors measured all the quantities needed for testing the energy/moment scaling relation (23). The line labeled $M_0 \cong f_0^{-3}$ indicates the trend of variation of moment with corner frequency predicted by Aki's scaling law

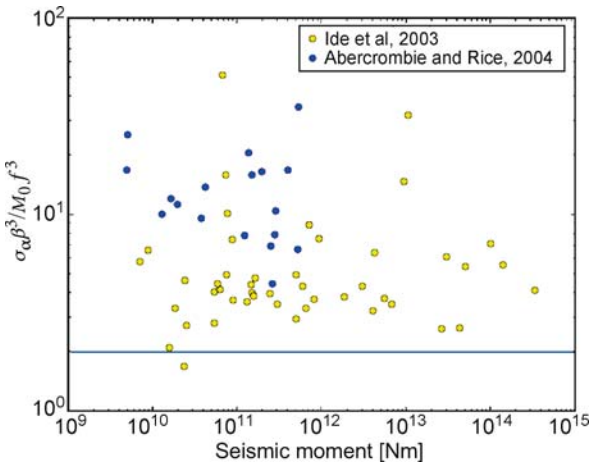

Earthquake Scaling Laws, Figure 4
The nondimensional coefficient C_r defined by expression (23) plotted as a function of seismic moment. Data from Abercrombie [2] and Ide et al. [32]. The non dimensional coefficient predicted by Brune's spectral model should be equal to 1.9739 as indicated by the horizontal line. Actually it varies over almost two orders of magnitude in this data set. This may be due to errors in the measurement of radiated energy, moment and corner frequencies or it may reflect an important departure from single parameter scaling of seismic sources

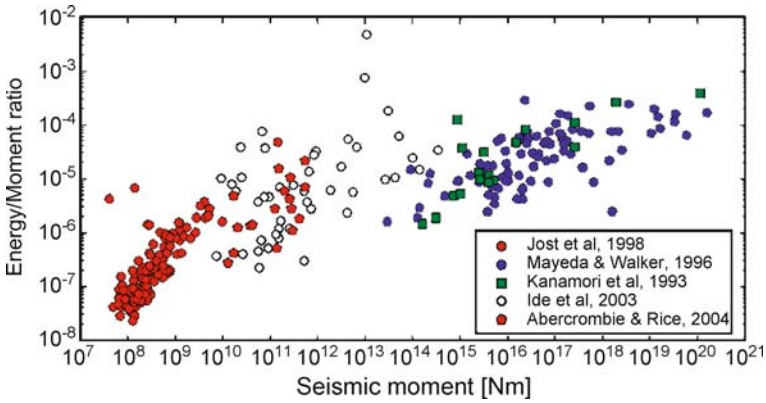

Earthquake Scaling Laws, Figure 5

Energy moment ratio as a function of moment for 14 orders of magnitude of moment. This figure is inspired on similar plots by McGarr [53] and Ide and Beroza [33]. We used data from Jost et al. [36], Mayeda and Walker [52], Kanamori et al. [41], Ide et al. [35] and Abercrombie and Rice [2]. The data of Jost et al. [36] was not corrected for attenuation as suggested by Ide and Beroza [33]. At this scale it is obvious that moment is the most broadly variable parameter, but there is also a large spread of apparent stress, over at least three orders of magnitude

dent of the seismic moment. Figure 4 shows that there is substantial scattering of the values of the non-dimensional ratio. There are many reasons for this variation; the most obvious one is experimental error due to uncorrected path or site effects, like attenuation, scattering and site amplification. I believe that even if those errors were corrected there would remain some variation due to source complexity that is not fully explained by the assumption that all earthquakes scale with a single parameter. In the literature authors generally assume that the scaling law applies and proceed to compute model-dependent quantities like static or dynamic stress drop using very specific models of rupture (see also p. 36 in [10]).

Clearly moment, energy/moment ratio and apparent stress are broadly distributed. This is illustrated in Fig. 5, inspired by previous figures of the same kind by McGarr [53] and Ide and Beroza [33]. The main difference between the results reported by these two authors is that Ide and Beroza [33] introduced a correction for attenuation at high frequencies. The figure shows the energy/moment ratio as a function of moment over 15 orders of magnitude of seismic moment. The sources of data [2,35,36,41,52] are not the same as those used by McGarr [53] and Ide and Beroza [33], because not all the data they used is published. I did not include small mine earthquakes from South Africa because it is not clear whether those events are due to frictional slip of pre-existing surfaces (see, e. g. [63]). For such a broad range of moments, the variation in energy-moment ratio is bounded, but it ranges over close to three orders of magnitude. It is clear that for the group of earthquakes reported in Fig. 5, apparent stress changes were important and deserve further work. This is not completely unexpected: a multitude of observations point out that stresses in the seismic zones are highly variable as well as the geometry of faulting.

Earthquake Fault Models: The Scaling of Geometry and Stress

The previous discussion focused on the properties of seismic radiation from moment tensor sources and the time and frequency dependence of the moment rate function. Actually, the Brune spectrum and Aki's scaling law can be retrieved from seismic waves without any reference to a particular fault model. In order to understand how the moment is related to source dimensions and the origin of omega-squared radiation, we have to introduce a specific fault model. We will proceed in two steps: first we will study a simple source model that explains most of the observations and, in a second step, we will discuss how this model can be generalized.

A fault is defined as a rupture in the earth crust with a relative displacement of its two sides. The relative displacement between the two sides of the fault (or fault slip) will be denoted by $\boldsymbol{D}(\boldsymbol{x}, t)$, a vector function of position on the fault ($\boldsymbol{x}$) and time (t). Thus, in general, $\boldsymbol{D}$ may vary in amplitude and direction over the fault plane and at each point is a function of time. The scalar seismic moment of an earthquake defined in (1) depends on source area and slip on the fault. It is essentially a static measurement of earthquake size. Corner frequency, on the other hand, is a measure of the duration of the earthquake signal which

is controlled by the time it takes for the rupture front to propagate across the fault. Thus, corner frequencies depend clearly on fault size, but one can expect this relation to be complex and very dependent on the details of the rupture process. This is often ignored in practical work, and simple source models are adopted in order to express the scaling law (23) in terms of source dimension. The argument in favor of this approach is that in order to build scaling laws that extend over several orders of magnitude of moment we may ignore the details of slip and geometry. Let us start by studying a simple circular crack, probably the simplest realistic fault model one can consider.

In a simplified model of fracture, the relative slip D of the two sides of a fault is produced by the relaxation of the shear stress transmitted across the fault. Shear stress changes with time due to the slow motion of plates, orogeny and a number of other processes that transfer stresses in the Earth's crust. When shear stress exceeds the strength of the material or the friction that maintains the fault locked, slip on the fault starts and, simultaneously, shear stress relaxes to a lower value until all motion on the fault ceases. This process is obviously very complex; detailed studies of even the simplest models of faulting show that stress relaxation at any point on the fault is complex function of not just local stress release on the fault but of other slipping points and of wave propagation on the fault. Solving such a complex problem is only worth the effort for very special, very well recorded events. For most other earthquakes we want to make global estimates of stress relaxation, and related them to the simple spectral model we studied in the previous section.

Let the shear stress acting on the fault plane before and after the occurrence of an earthquake be T_0 and T_f, respectively. We define stress drop $\Delta\sigma$, as the difference

$$\Delta\sigma = T_0 - T_f \,. \tag{25}$$

The stress drop represents the part of the acting stress which is used to produce the slip of the fault so that $\Delta\sigma$ is related to slip $\boldsymbol{D}$. The relation between stress drop and slip will be in general very complex: it will depend on the geometry of the fault, but also on certain fundamental assumptions about the stress field in the earth. In general, stress will be much more heterogeneous than slip because, at least in the static case, stress is a generalized derivative of stress. We will discuss these properties briefly in Sect. "Earthquake Dynamics and the Scaling of Energy". We will assume here that the stress distribution has to be such that a finite amount of energy is released during faulting. This assumption is the basis of fracture mechanics, leading to the condition that stress may have at most inverse square root singularities on the fault surface (see, e. g. [22,45,60]). In the early work on faults, many authors assumed that earthquakes were due to dislocations, slip distributions that present slip discontinuities at their borders. The best known example of such model is the Haskell [30] rectangular dislocation model. This model produces non-integrable stress concentrations around the edges of the fault that store an infinite amount of strain energy [48]. Thus, even if this model radiates a finite amount of energy it can not be used to estimate energy balance during seismic rupture. This paradox was well known in mechanics, where dislocations and cracks are treated as very different phenomena.

The Static Circular Crack

A simple model that may be used to explain many of the scaling laws observed in earthquake seismology is that of a static circular ("penny shaped") crack of radius a lying on the x, y plane. We assume that the fault is loaded by a uniform initial shear stress T_0, and that T_f, the final stress, is also uniform inside the fault. The slip on the fault produced by a constant static stress drop $\Delta\sigma$ inside a circular crack was computed by Eshelby [20] and Keilis-Borok [42]

$$D(r) = \frac{24}{7\pi}\frac{\Delta\sigma}{\mu}\sqrt{a^2 - r^2} \,, \tag{26}$$

where r is the radial distance from the center of the crack on the (x, y) plane, a is the radius of the crack, and μ is the elastic rigidity of the medium surrounding the crack. Slip has the typical elliptical shape associated with cracks. The distribution of slip and the stress change for a circular crack are schematically shown in Fig. 6. Using the definition of the seismic moment (1) we can determine the scalar seismic moment for this circular fault:

$$M_0 = \frac{16}{7}\Delta\sigma a^3 \tag{27}$$

so that the moment is the product of the stress drop times the cube of the fault size. This simple relation will be used to explain the seismic scaling law in terms of fault radius and stress drop. Other fault geometries produce somewhat different scaling laws, including products of fault length and fault width. Unfortunately, as far as I know, no other geometry can be solved in such a simple closed form as the circular crack.

We can also compute the static strain energy change in the elastic medium surrounding the circular fault. This is defined as

$$\Delta W = \frac{1}{2}\int_S \Delta\sigma D \mathrm{d}S \,. \tag{28}$$

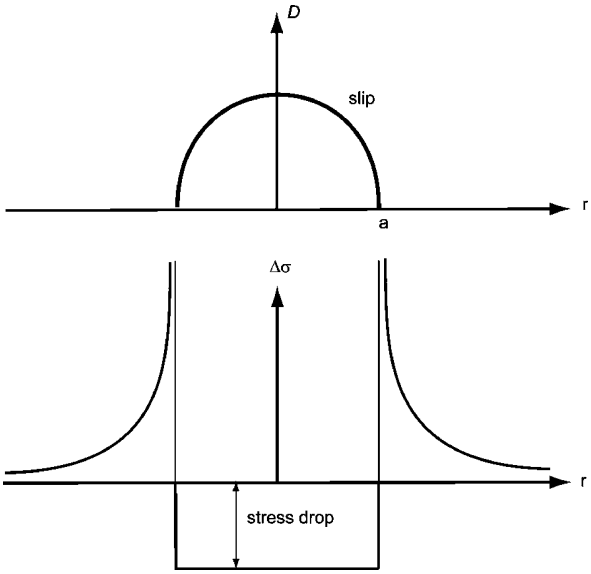

Earthquake Scaling Laws, Figure 6
The simple static circular crack. The upper panel shows the slip distribution as a function of radius. The bottom panel shows the stress change produced by the slip at the top

From simple thermodynamic considerations, ΔW should be negative so that stress drop and slip should have opposite signs. For the circular crack this can be easily computed replacing the slip distribution (26) in this integral. Integrating, we find

$$\Delta W = \frac{8}{7}\frac{\Delta\sigma^2}{\mu}a^3 \tag{29}$$

where we have omitted the negative sign, so that ΔW should be interpreted as the reduction of strain energy from the elastic body caused by the earthquake. Dividing (27) into (29) we find that the strain energy to moment ratio for the circular crack is just.

$$\frac{\Delta W}{M_0} = \frac{1}{2}\frac{\Delta\sigma}{\mu}\,. \tag{30}$$

In the very early studies of seismic rupture it was sometimes assumed that radiated energy was equal to strain energy change, i. e $\Delta W = E_r$. In that case apparent stress $\sigma_\alpha 1/2\Delta\sigma$ his assumption is sometimes referred to as the Orowan [58] model. This model is very unlikely to hold for real earthquakes: if all the available energy were radiated, there would be no energy left for producing rupture propagation and consequently rupture should propagate exactly at the P wave or the S wave speeds or should stop immediately in front of any obstacle (see [21,44]).

The circular crack model has been used to quantify numerous earthquakes for which the moment was estimated from the amplitude of seismic waves, and the source radius was estimated from corner frequencies, surface deformation, etc. The result is that for shallow earthquakes in the seismogenic zones like the San Andreas Fault, or the North Anatolian Fault in Turkey, average stress drops are of the order of 1–10 MPa. For deeper events in subduction zones, stress drops can reach several tens of MPa. Thus, average stresses do not vary much, compared to the variation of moment over more than 15 orders of magnitude.

Brune's [17] Model of Seismic Radiation

Brune developed a model of seismic radiation based on the observation that seismic spectra had the omega-squared spectral shape (7). Brune [17] proposed a model only for shear waves; although, as we already mentioned, P waves have a similar spectral shape with a corner frequency that is different from that of S waves. In his 1970 paper proposed also a relation between the corner frequency f_0 and the source radius a of a circular fault:

$$f_0 = 0.3724\frac{\beta}{a}\,. \tag{31}$$

That is, the corner frequency is inversely proportional to the size of the fault. The origin of the coefficient 0.3724 that appears in (31) is very important because it has major consequences for the partitioning of elastic energy. Although the steps followed by Brune were different, we can obtain his results in the following way.

Replacing the expression for moment of a circular crack (27) into the expression for energy moment ratio (17) we get

$$\frac{E_r^S}{M_0} = 4.5118\frac{\Delta\sigma}{\mu}\frac{a^3 f_0^3}{\beta^3} \tag{32}$$

inserting Brune's expression for the corner frequency (31) into (32) we get the apparent stress

$$\sigma_a = \frac{\mu E_r^S}{M_0} = 0.2331\Delta\sigma\,. \tag{33}$$

So that apparent stress drop in Brune's model is proportional to the static stress drop $\Delta\sigma$. This result, derived by Singh and Ordaz [69], has a very interesting consequence for the energy balance in earthquakes. Indeed, inserting the relation (30) between moment and strain energy change during an earthquake into (34), we get

$$E_r^S = 0.466\,\Delta W\,. \tag{34}$$

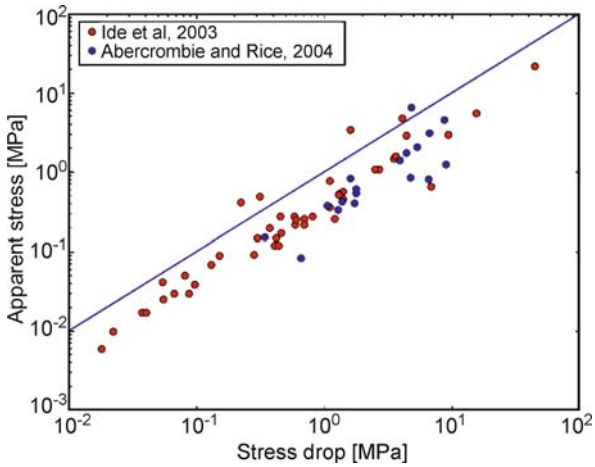

Earthquake Scaling Laws, Figure 7

Apparent stress versus stress drop for the data reported by Ide et al. [32] and Abercrombie and Rice [2]. This figure suggests that apparent stresses computed directly from radiated energy and moment are not independent of stress drop computed for a static circular crack model

That is Brune's model makes the implicit assumption that the energy radiated in the form of S waves is 46.6% of the available strain energy. This was computed by Brune [17]; he found that S waves carried 44% of the available energy in his Equation (40). The number cited here, 46.6% is due to Brune's [18] correction of (31). I think that this is a very important consequence of the corner frequency source radius relationship (31), that is not often cited in the literature.

We can now test whether stress drops computed using the static circular crack model produces are or not independent of the apparent stresses computed from the energy moment ratio E_r/M_0.

In Fig. 7 we plot the ratio of apparent stress to stress drop for the data of Ide et al. [35] and Abercrombie and Rice [2]. The relation is roughly linear but the ratio is not well defined, although it is clearly less than 1 for most of the events.

Earthquake Dynamics and the Scaling of Energy

Earthquakes are dynamic processes in which rupture propagates under the control of friction that acts between the two sides of the fault as they slip. The study of the friction law that actually operates on seismic faults is a major problem in seismology and fracture mechanics. Laboratory experiments, seismic observations and field studies are needed to solve this complex problem. In this section we will attempt to establish some general properties of seismic sources without getting involved with fine details about friction and rupture propagation (see the contribution by Ampuero in the present volume for a fuller discussion). The main question in this context is: can we establish some general properties of seismic ruptures that are independent of the details of friction? Do the observations of seismic spectra and scaling laws constrain in any way the overall properties of seismic sources? This approach has been taken in recent years by many authors, some have tried to convert slip models inverted from near field seismic observations to determine energy balance [16,34]; others have tried to do the same by remarking that dynamic ruptures only propagate at reasonable rupture speeds for a very limited range of seismic parameters [57]; or, very recently, have tried to derive general properties of the friction law from the scaling of seismic spectra [2]. We will follow the latter approach because I believe that it is very promising.

The Dynamic Circular Crack Model

Perhaps, the simplest fault model that can be imagined is a circular crack that grows from a point at a constant or variable rupture speed and then stops at the rim of the fault, arrested by the presence of unbreakable barriers. This model is the natural dynamic equivalent to the static circular crack discussed in the previous section. The circular crack problem is posed in terms of stresses not of slip, but the rupture process is fixed in advance so that rupture does not develop spontaneously. This is the only unrealistic feature of this model, hence it is considered as quasidynamic, that is, rupture is kinematically defined, but slip is computed solving the elastodynamic equations. This model was carefully studied by a number of authors in the 1970s [43,46,47,62,65].

Let us consider a rupture that starts form a point and then spreads self-similarly at constant rupture speed v_r without ever stopping. Slip on this model is driven by stress drop inside the fault. The solution of this problem is somewhat difficult to obtain because it requires very advanced use of self-similar solutions to the wave equation and its complete solution for displacements and stresses must be computed using the Cagniard de Hoop method [62]. Fortunately, the solution for slip across the fault found by Kostrov [43] is surprisingly simple. Slip in the circular fault is parallel to the direction of stress drop on the fault and it has the typical elliptical shape:

$$D(r,t) = C(v_r)\frac{\Delta\sigma}{\mu}\sqrt{v_r^2 t^2 - r^2} \tag{35}$$

where r is the radius in a cylindrical coordinate system centered on the point of rupture initiation. $v_r t$ is the instantaneous radius of the rupture at time t. $\Delta\sigma$ is the dynamic stress drop assumed to be constant inside the rupture zone, μ is the elastic rigidity, and $C(v_r)$ is a slowly varying function of the rupture velocity v_r. For most practical purposes $C \sim 1$. This simple solution constitutes a key result containing one of the most important properties of circular cracks. Slip inside the fault scales with the ratio of stress drop over rigidity times the instantaneous radius of the fault. As rupture develops, slip increases with the size of the rupture zone.

Energy Release Rate for a Dynamic Circular Crack

We can determine the energy release rate for Kostrov's model (35) from the behavior of stresses near the edge of the crack. At time t, the fault radius is $r = v_r t$, the slip velocity field derived from (35) has the form

$$V(r,t) = C(v_r)\frac{\Delta\sigma}{\mu}\frac{v_r^2 t}{\sqrt{v_r^2 t^2 - r^2}} \tag{36}$$

so that, near the rupture front, the velocity field presents the well known inverse-squared root singularity predicted by dynamic crack theory [22,45]. We can then approximate the singularity in slip rate in the general form

$$V(r,t) = \frac{V_d}{\sqrt{2\pi}}\frac{1}{\sqrt{v_r t - r}} \tag{37}$$

where V_d is the velocity intensity factor, a measure of the amplitude of the square root singularity in slip velocity that moves with the rupture front. This velocity singularity is associated with a dynamic stress concentration ahead of the rupture front

$$\Delta\sigma(r,t) = \frac{K_d}{\sqrt{2\pi}}\frac{1}{\sqrt{r - v_r t}} \tag{38}$$

which is also of the inverse square root type. K_d is the dynamic stress intensity factor. The amplitudes V_d and K_d are linearly related to each other with a coefficient that is different for fracture modes II and III. Avoiding details that are discussed by Freund [22], we can write

$$V_d = a(v_r)\frac{K_d}{\mu}v_r \tag{39}$$

where $a(v_r)$ is a coefficient that depends on the instantaneous rupture velocity v_r.

We can compute the energy release rate near the border of the fault directly from these expressions, using several results from Kostrov [43] and Madariaga [47]:

$$G_c(v_r) = \frac{K_d V_d}{v_r} = a(v_r)\frac{K_d^2}{2\mu}, \tag{40}$$

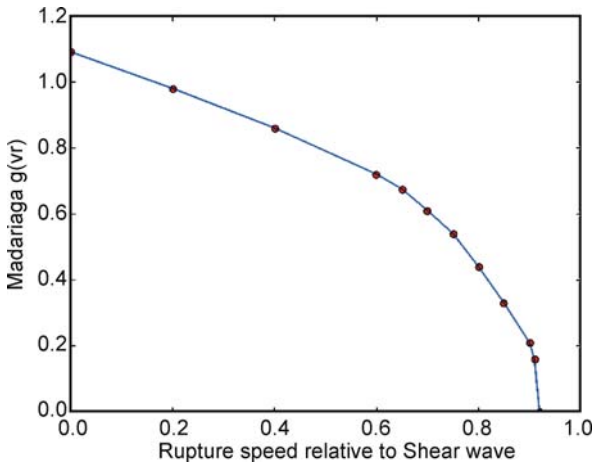

Earthquake Scaling Laws, Figure 8
The $g(v_r)$ function of rupture velocity. This function controls the fraction of strain energy that is used as fracture energy in a quasi-dynamic circular shear crack that ruptures at a constant rupture speed

that is, the energy release rate G_c near the crack front is proportional to the square of the dynamic stress intensity factor. Since the dynamic stress intensity factor K_d tends to zero at high rupture speeds, the energy release rate G_c also decreases at high speed rates. Thus, the faster the rupture, the less energy is spent in making the rupture advance.

For the circular crack, the energy release rate is not uniform around the perimeter of the circular fault because it is different in mode II (in-plane) and mode III (anti-plane). From results by Madariaga [46] we get for a circular crack

$$G_c(v_r, r) = \frac{g(v_r)}{3}\frac{\Delta\sigma^2}{\mu}r\,. \tag{41}$$

Where $g(v_r)$ is a monotonically decreasing function of rupture speed shown in Fig. 8. An exact expression for $g(v_r)$ was proposed by Madariaga [46], but his Eq. (30) contains a misprint that was corrected by Ide [32].

The Scaling of Energy Release Rate with Earthquake Size

We can now estimate the energy used for the propagation of the seismic rupture using the previous estimate for Brune's [17] model (34). We can establish the following global energy balance for an earthquake that is well modeled by Brune calibration of the shear wave spectra.

As rupture propagates, strain energy released by faulting ΔW is used in part to produce seismic waves, and in part to make fracture advance (measured by the energy re-

lease rate G_c). Assuming that G_c is constant over the fault surface we find the following earthquake energy balance

$$\Delta W = E_r + G_c S \tag{42}$$

where S is the area of the fault. In order to estimate G_c, we need to estimate the total radiated energy E_r. Brune's model is only suitable to compute S wave radiated energy which as shown by Eq. (34) is 46.6% of the available strain energy. Energy transported by P waves can be computed using either theoretical arguments or observations of P to S energy rations. This question was examined by Boatwright and Fletcher [15] who concluded that, theoretically at least, a crack-like source should produce about 15 times more energy carried by S waves than P wave energy. This number is confirmed by observations reported by Abercrombie and Rice [2], Prieto et al. [59]. We can thus estimate that radiated energy in the Brune model is roughly 50% of the strain energy released by the earthquake; the other 50% goes into rupture work, i. e.

$$G_c S \approx \frac{1}{2}\Delta W\,. \tag{43}$$

Using the expression (29) for ΔW we get

$$G_c \approx \frac{1}{2}\frac{\Delta W}{\pi a^2} = \frac{4}{7\pi}\frac{\Delta\sigma^2}{\mu}\,a\,. \tag{44}$$

That is, if we adopt Brune's model, *energy release rate scales with fault size*.

In the previous calculation we assumed that G_c was uniform inside the fault. A better assumption would be that, as the rupture propagates, G_c grows with the radius as in the simple quasi-dynamic model of Madariaga [46],

$$G_c(r) = \frac{\Delta\sigma^2}{3\mu}\,r\,g(v_r) \tag{45}$$

where $g(v_r)$ will be determined for Brune's model. The total energy release during rupture is then

$$\int_S G_c(r)\mathrm{d}S = \frac{2\pi}{9}\frac{\Delta\sigma^2}{\mu}g(v_r)a^3\,.$$

Using (43) again we get

$$g(v_r) = \frac{18}{7\pi} = 0.818\,. \tag{46}$$

Thus, Brune's model is equivalent to a circular quasidynamic shear crack propagating such that the energy release rate grows with fault radius like

$$G_c = \frac{6}{7\pi}\frac{\Delta\sigma^2}{\mu}\,r\,.$$

Thus, whether we use a constant energy release rate on the fault (44), or a more realistic model where energy release grows with the radius of the fault, we find that energy release rate grows with fault radius in Brune's model. This result confirms that energy release rate scales with the fault radius and that it adjusts as the fault grows [55,56].

The relation between energy release rate and fault size was studied by Abercrombie and Rice [2] using their own data and data from a number of previous studies. They reached the conclusion that G_c grows with radius roughly like $r^{0.4}$, not like the radius as in (41) and (47). Abercrombie and Rice estimated G_c from the expression

$$G_c = \frac{1}{2}(\Delta\sigma - \sigma_a)D$$

where D is slip. This expression is entirely compatible with ours, so that the reason G_c scales with the power 0.4 of the radius is that stress drops scale with earthquake size, specially if $\Delta\sigma$ is not linearly related to σ_a as shown in Fig. 7. After checking Fig. 8 of Abercrombie and Rice [2], I conclude that they obtained scaling with a power of 0.4 using the full data set, including much larger earthquakes. For the Cajon pass earthquakes studied by Abercrombie and Rice the power of radius in scaling is much closer to one. Seismic data have now reached the quality necessary to test different hypothesis about the scaling of rupture energy in order to see whether larger earthquakes are really different from smaller ones.

Scaling of Energy, Magnitude and Moment

Kanamori [38] introduced the so-called moment magnitude, Mw, assuming that all available strain energy was converted into seismic waves, i. e. that $E_r \approx \Delta W$. This assumption means that no energy is used to propagate fracture so that $G_c = 0$. Using the definition of strain energy change, (28), we get Kanks and Kanamori [28]:

$$E_r = \frac{1}{2}\Delta\sigma\overline{D}S\,, \tag{47}$$

where $\overline{D}$ is the average slip of the earthquake and S its source area.

This expression shows that from the radiated energy we only have information about the stress drop, not about the absolute stress level acting on the fault during faulting. Rewriting expression (27)

$$M_0 = \frac{16}{7\pi^{3/2}}\Delta\sigma S^{3/2} \tag{48}$$

and taking logarithms

$$\log M_0 = \frac{3}{2}\log S + \log\left(\frac{16\Delta\sigma}{7\pi^{3/2}}\right) . \tag{49}$$

From this equation it follows that, for constant stress drop scaling, $\log S$ should be proportional to $2/3 \log M_0$. This hypothesis had been shown empirically to be valid for a large range of values of M_0 [39]. We noticed however that more recent data (Fig. 6) shows that stress drop varies over 3 orders of magnitudes, at least in the data reported by Ide and Beroza [33].

Assuming constant stress drop, Kanamori [38] defined the moment magnitude M_w based on the empirical relation of Gutenberg and Richter [23,24] between surface wave magnitude and seismic energy (in Joules): $\log E_r = 1.5M_S + 4.8$. Hanks and Kanamori [28] proposed the moment magnitude scale

$$M_W = \frac{2}{3}\log M_0 - 6.07 \tag{50}$$

where M_0 is measured in Nm. The moment magnitude has become the standard way to measure the size of earthquakes. Both M_w and seismic moment M_0 (Nm) can be related to other magnitude measurements by a number of empirical relations. (see, e. g. [39]).

Radiated seismic energy can not be computed directly from (47) because it needs to be corrected for the part of the strain energy that is used to propagate the fracture (see, (42)).

More General Scaling Relations Derived from the Scaling Law of Earthquake Spectra

In the two previous sections several scaling relations have been established which relate the parameters involved in the fracture process of earthquakes. It has been shown that slips and slip velocities scale linearly with stress drop, which is the most fundamental scaling parameter (pp. 202–211 in [67]). If the average stress drop measured over the whole fault plane is roughly constant for all earthquakes, the slip on the fault should scale with the dimensions of the fault (L) which for small earthquakes represents the length (L) or radius of the fault and for large earthquakes the width (W). An unsolved issue, due mostly to lack of data for very long strike slip earthquakes, is that of a possible difference in the scaling of the seismic moment with fault length, between large and small earthquakes. Strike slip earthquakes with seismic moment less than 10^{21} Nm ($M_w < 8$) should scale with L^3, while larger ones with L^2.

Many other relations can be established starting from the basic scaling laws discussed earlier in this chapter. For instance, maximum and average slip for earthquakes scale like the cubic root of moment for most earthquakes [53,54], but these scaling relations can be derived from the basic relations discussed earlier.

More Realistic Radiation Model

In reality earthquakes occur in a complex medium that is usually heterogeneous and dissipative. Seismic waves become diffracted, reflected, and in general suffer from multipathing in those structures. Accurate seismic modeling would require perfect knowledge of Earth's structure. It is well known and understood that structural complexities dominate signals at certain frequency bands. For this reason the simple model presented here can be used to understand the main features of earthquakes at long wavelengths, while the more sophisticated approaches that attempt to model every detail of the wave form are reserved only for more advanced studies. Here, like in many other areas of geophysics, a balance between simplicity and concepts must be kept against numerical complexity that may not always be warranted by lack of knowledge of the details of Earth's structure. If the simple approach were not possible, then many standard methods to study earthquakes would be impossible to use. A good balance between simple, but robust concepts and the sophisticated reproduction of the complex details of real wave propagation is a permanent challenge for seismologists.

Why Does the Spectrum Decay Like ω Squared?

We have seen that seismic data is in very broad agreement with the general features of Brune's spectral model (7). We have explained the scaling of low frequencies in terms of simple static source models, the corner frequency and the high frequency decay are explained by the energy balance Eq. (32). Seismic energy must be finite and a well defined fraction of the available strain energy. These conditions require that the spectrum of seismic energy is integrable in expression (13). General properties of Fourier transform can be invoked to demonstrate that in the time domain displacement signals are continuous functions of time with discontinuous derivatives. That is, the velocity field emitted by a seismic source in the far field contains jumps in particle velocity as is the case with the seismic signals proposed by Brune (19) or by Kanamori and Rivera (20). For the dynamic circular crack studied in this section, the velocity jumps are emitted when rupture stops abruptly at the rim of the circle of radius a. The

nature of these stopping phases has been carefully studied by a number of authors, including Madariaga [46,47], Boatwright [12], Spudich and Frazer [71] and Bernard and Madariaga [8]. Their study is very complex and beyond the purpose of the present article, we will use a simpler approach based on a scaling argument.

Let us now consider how the stopping phases scale with earthquake size and stress drop. In the omega squared model, the high frequency decay of the far-field displacement produced by shear waves can be written in the very general form

$$u_s(r,\omega) = C(\theta,\varphi)\frac{M_0\omega_0^2}{\mu\beta}\frac{1}{R}\omega^{-2}\,,$$

where we lumped the numerical coefficients and the radiation pattern in the single non-dimensional coefficient $C(\theta,\phi)$. R,θ and ϕ are spherical coordinates at a reference point on the fault. Using the expressions for M_0 and corner frequency in terms of fault radius and stress drop we obtain

$$u(r,\omega) = C(\theta,\varphi)\frac{\Delta\sigma\, a}{\mu}\frac{\beta}{R}\omega^{-2} \tag{51}$$

with a slightly different dimension-less coefficient. The far field waves scale at high frequencies like the product of stress drop times the radius of the fault. That is exactly what was predicted by Madariaga [47] for a circular crack. It is interesting to remark that the factor $\Delta\sigma a$ actually comes from the product of the stress intensity factor K and the square root of the radius a. Although the scaling of high frequencies was derived here for a very special circular crack model, it can be easily generalized to ruptures of any shape, splitting the factor $\Delta\sigma a$ into a stress intensity factor and the square root of the local radius of curvature of the wave front of the stopping phase.

The previous model for the radiation of omega squared high frequency waves can be extended to more complex source models, in particular to source models that contain a number of subfaults (see, e. g., [14]). The high frequency seismic waves emitted by such a model are due to stopping phases emitted all along their propagation process. Each such stopping phase contributes to enriching the high frequency contents of the seismic waves. The incoherent sum of those phases produces a total spectrum that scales with fault radius as in (51). In this sense, omega squared decay is the signature of the presence of cracks on the fault. Recent work has shown that omega squared waves are emitted every time the rupture front changes rupture velocity, or that the rupture is deviated from a plane by the presence of fault kinks or discontinuities.

Kinematics and Statistical Models for Fault Slip

So far we have discussed a dynamic crack approach to understanding earthquake scaling. Another method to describe seismic sources radiation was introduced by Haskell [30]. He assumed that earthquakes could be described by simple propagating dislocations leaving a constant slip in their trail. The most common such model is that of a flat rectangular fault with constant slip in it. Such model is mechanically impossible, because it needs infinite amount of energy to be created. Curiously, though, the Haskell model produces a finite amount of radiated energy and the radiated field can be computed exactly both in the far [31] and near field [48]. Because the strain energy change produced by Haskell's model is infinite, an energy balance equation like (42) can not be established. In Haskell model all of the energy released from the medium is absorbed by the dislocation motion and seismic radiation is just a secondary feature of the source processes.

Dislocation models have appeared in a different form, derived from statistical considerations about the distribution of slip on faults. In his seminal paper, Andrews [6] established some basic statistical properties of slip distributions that are actually based on Haskell's [31] original study of the power spectra of slip and the correlation functions of slip. His analysis is based on some general features of fractal surfaces and distributions; here we will look at these scaling relationships from the point of view of the circular crack model that we discussed in previous sections.

The slip function of a circular crack was defined in (26). This is a function of radius only so that its Fourier transform is very easy to compute using the Hankel transform:

$$\tilde{D}(k) = 2\pi\int_0^\infty D(r)J_0(kr)r\mathrm{d}r \tag{52}$$

where k is the radial wave-number, J_0 is the Bessel function of degree zero. Inserting the expression (26) in (52) and integrating we get

$$\tilde{D}(k) = \frac{48}{7}\frac{\Delta\sigma}{\mu}a^3\frac{\sin(ak)-ak\cos(ak))}{k^3a^3}\,. \tag{53}$$

At low wave-number, when k tends to 0, the spectrum (53) tends to the value

$$\tilde{D}(0) = \frac{16}{7}\frac{\Delta\sigma}{\mu}a^3 = \frac{M_0}{\mu}$$

that is, the low wave number limit of the slip spectrum is the seismic moment, just as the low frequency limit of Brune's spectrum is the seismic moment. This is of course

not a coincidence but a consequence of the fact that the seismic moment is the source of low frequency waves. At high wave numbers, when $ka \gg 1$, the spectrum $D(k)$ behaves like

$$\lim_{k\to\infty} \tilde{D}(k) = -\frac{48}{7}\frac{\Delta\sigma}{\mu} a^3 \frac{\cos(ak)}{a^2k^2} \,. \tag{54}$$

The slip spectrum decays like k^{-2} at high wave numbers, a property that seems to be as universal as the high frequency decay of seismic spectra with omega-squared.

The important issue is why is it k^{-2}? The origin of the high frequency behavior of the slip spectrum can be determined with some simple properties of the Hankel transform (52). Take a circular fault of finite radius a. Then for different types of slip discontinuity we get the following asymptotic behavior:

	$\lim_{r\to a} D(r)$	$\lim_{k\to\infty} D(k)$
Constant	1	$k^{-3/2}$
Crack-like	$(a-r)^{1/2}$	k^{-2}
Conical	$(a-r)$	$k^{-5/2}$
Smooth	$(a-r)^{3/2}$	k^{-3}

Thus the high wave number behavior is a reflection of the discontinuity of slip at the border of the fault. Andrews [6] studied a slip distribution that behaves like $(a-r)^{3/2}$ near the edge of the fault. In his case the high wave number decay is k^{-3}. Thus the high wave number behavior of slip distributions is controlled by the discontinuities of slip, a crack like discontinuity producing a k-squared distribution. Note that the spectral behavior for two-dimensional distributions is quite different than for two dimensional slip distributions. Haskell [30] used the properties of 2D Fourier transforms to derive several conclusions about earthquake spectra that do not apply to circular cracks. For a two-dimensional plane or anti-plane crack the spectrum decays like $k^{-3/2}$, such a spectrum is inadmissible in 3D because it would imply non-integrable stress distributions as we will show now.

The stress field associated with the circular shear crack slip can be computed in a straightforward way using the expressions provided by Eshelby [20] or Sneddon [70]. Let the fault be located on the plane (x,y) and slip be parallel to the axis x, (i. e. $D(x) = \Delta u_x(x)$. In the spectral domain the associated stress drop can be computed from slip by

$$\Delta\sigma_{xz}(k) = -\frac{\mu}{2k}\left(\frac{2(\lambda+\mu)}{\lambda+2\mu}k_x^2 + k_z^2\right) D(k) \tag{55}$$

(see [64]). This is a relatively simple expression, but it is not easy to compute analytically because stress drop for the circular crack does not have cylindrical symmetry. Andrews [6] provided a simplification that we will use here: assuming that the elastic constant $\lambda = 0$ i. e. that is the elastic medium is incompressible), we get in Fourier domain:

$$\Delta\sigma_{xz}(k) = -\mu k D(k) \,. \tag{56}$$

Multiplying (53) by $-k$ and doing the inverse Hankel transform (see [70]) we get, approximately

$$\begin{aligned} \Delta\sigma_{xz}(r) &= -\Delta\sigma \quad \text{for} \quad r < a \,, \\ \Delta\sigma_{xz}(r) &\cong K/\sqrt{r-a} \quad \text{for} \quad r > a \,. \end{aligned} \tag{57}$$

That is, inside the crack, stress drop is constant while outside stress drop exhibits an inverse square root singularity typical of cracks as discussed in (38). This explains why the spectrum of a circular crack decays as k^{-2} in the wavenumber domain. The k^{-2} spectrum is the signature of the presence of a crack.

Mai and Beroza [51] computed the correlation lengths, fractal dimensions, Hurst exponents and wave number spectra of 42 earthquakes for which the slip distribution on the fault planes were available from inversion of seismic and geodetic data. They reached the conclusion that the high frequency decay had an average fractal dimension of $2.29 + / - 0.23$ that implies a high wave number asymptotic decay of $k^{1.71}$, that is slip distributions determined from slip inversions tend to be rougher than the spectra of circular cracks. The origin of this exponent needs to be carefully scrutinized in terms of fault segmentation.

Bernard and Herrero [7] and Bernard et al. [9] proposed a model linking seismic radiation to the spectral properties of the distribution of slip on the fault. In their model rupture propagates in a single space direction at constant speed. In order to obtain an omega squared far field spectrum they assumed that rise time is essentially a delta function or that it scales with wave number, which is equivalent. This result needs to be confronted with dynamic simulations propagating at finite speeds and finite energy release rate. In the circular crack model, the high frequency spectrum was controlled by stopping phases, which do not exist in the Bernard et al. model. This problem needs careful attention, specially in the presence of geometrical heterogeneity that may produce stopping phases.

Future Directions

Most properties of ensemble averaged seismic spectra and slip distributions can be explained by a simple circular crack model. Seismic waves as well as slip distributions determined from seismic and geodetic inversions carry

the signature of the crack models that are at the base of earthquake ruptures. Whether the earthquake can be modeled as a simple circular crack or as the complex sum of a distribution of such cracks the result is the same: slip is of k-squared type and seismic radiation is of omega-squared type. Departure from these models can be expected if stress drop scales with fault size. There is no clear-cut evidence for such behavior because of the difficulties in estimating radiated energy mentioned several times in this review. The variations of stress drop required to explain observations may be an intrinsic variation of stresses depending on fault maturity, the position of the fault in the seismic cycle, etc.

In recent years the quality and quantity of seismic data has improved very significantly with the deployment of digital instruments in many active areas of the earth. This is a unique opportunity to test the self-similarity of earthquakes. If tests like those of Fig. 4 are applied to new data, the problem of the variability of radiated energy/moment ratio (or equivalent of apparent stress and stress drop) will be addressed more carefully. I would not be surprised if we finally concluded that apparent stress varies significantly among different earthquakes as suggested by Fig. 4.

Acknowledgments

This research was partially funded by the SEISMULATORS contract with ANR under program Catastrophes Telluriques et Tsunamis, and by the Research training network SPICE of the 7th PCRD of the European Union. I am deeply indebted to Luis Rivera, Martin Mai and Art McGarr for their careful and patient review of an initial version of this paper.

Bibliography

1. Abercrombie RE (1995) Earthquake source scaling relationships from −1 to 5 ML using seismograms recorded at 2.5 km depth. J Geophys Res 100:24015–24036
2. Abercrombie RE, Rice JR (2005) Can observations of earthquake scaling constrain slip weakening? Geophys J Int 162:406–424
3. Aki K (1966) Generation and propagation of G waves from the Niigata earthquake of June 16, (1964) part 2. Estimation of earthquake movement, released energy, and stress-strain drop from G wave spectrum. Bull Earthq Res Inst Univ Tokyo 44:23–88
4. Aki K (1967) Scaling law of seismic spectrums. J Geophys Res 72:1217–1231
5. Aki K, Richards PG (2002) Quantitative Seismology, 2nd edn. University Science Books, Sausalito
6. Andrews DJ (1980) A stochastic fault model: 1. Static case. J Geophys Res 85:3867–3877
7. Bernard P, Herrero A (1994) A kinematic self-similar rupture process for earthquakes. Bull Seismol Soc Am 84:1216–1228
8. Bernard P, Madariaga R (1984) A new asymptotic method for the modeling of near-field accelerograms. Bull Seismol Soc Am 74:539–557
9. Bernard P, Herrero A, Berge C (1996) Modeling directivity of heterogeneous earthquake ruptures. Bull Seismol Soc Am 86:1149–1160
10. Beroza G, Kanamori H (2007) Comprehensive overview in Treatise on geophysics. In: Schubert G (ed) Earthquake Seismology. Elsevier, Amsterdam
11. Boatwright J (1978) Detailed spectral analysis of two small New York State earthquakes. Bull Seismol Soc Am 68:1177–1131
12. Boatwright J (1980) A spectral theory for circular seismic sources: Simple estimates of source dimension, dynamic stress drop, and radiated energy. Bull Seism Soc Am 70:1–26
13. Boatwright J (1982) A dynamic model for far field acceleration. Bull Seism Soc Am 72:1049–1068
14. Boatwright J (1988) The seismic radiation from composite models of faulting. Bull Seism Soc Am 78:489–508
15. Boatwright J, Fletcher JB (1985) The partition of radiated energy between P and S waves. Bull Seismol Soc Am 75:361–376
16. Bouchon M (1997) The state of stress on some faults of the San Andreas Fault system as inferred from near-field strong-motion data. Bull Seismol Soc Am 58:367–398
17. Brune JN (1970) Tectonic stress and spectra of seismic shear waves from earthquakes. J Geophys Res 75:4997–5009
18. Brune JN (1971) Correction. J Geophy Res 76:5002
19. Brune JN, Archuleta RJ, Hartzell S (1979) Far-field S wave spectra, corner frequencies, and pulse shapes. J Geophys Res 84:2262–2272
20. Eshelby JD (1957) The elastic field of an ellipsoid inclusion and related Problems. Proc R Soc Lond A 241:376–396
21. Freund LB (1972) Energy flow into the tip of an extending crack in an elastic solid. J Elast 2:341–348
22. Freund LB (1989) Fracture Dynamics. Cambridge University Press, Cambridge
23. Gutenberg B, Richter CF (1942) Earthquake magnitude, intensity, energy, and acceleration. Bull Seism Soc Am 32:163–191
24. Gutenberg B, Richter CF (1956) Earthquake magnitude, intensity, energy, and acceleration (second paper). Bull Seism Soc Am 46:105–145
25. Hanks TC (1979) *b* values and w^{-g} seismic source models: Implications for tectonic stress variations along active crustal fault zones and the estimation of high-frequency ground motion. J Geophs Res 84:2235–2242
26. Hanks TC (1981) The corner frequency shift, earthquake source models, and *Q*. Bull Seismol Soc Am 71:597–612
27. Hanks TC (1982) f_{max}. Bull Seismol Soc Am 72:1867–1879
28. Hanks TC, Kanamori H (1979) A moment magnitude scale. J Geophys Res 84:2348–2350
29. Hanks T, Thatcher W (1972) A graphical representation of seismic source parameters. J Geophys Res 77:4393–4405
30. Haskell NA (1964) Total energy spectral density of elastic wave radiation from propagating faults. Bull Seism Soc Am 54:1811–1841
31. Haskell NA (1966) Total energy spectral density of elastic wave radiation from propagating faults: Part II. A statistical source model. Bull Seism Soc Am 56:125–140
32. Ide S (2002) Estimation of Radiated energy of finite-source earthquake models. Bull Seism Soc Am 92:2294–3005

33. Ide S, Beroza GC (2001) Does apparent stress vary with earthquake size? Geophys Res Letters 28:3349–3352
34. Ide S, Takeo M (1997) Determination of constitutive relations of fault slip based on seismic wave analysis. J Geophys Res 102(27)379–391
35. Ide S, Beroza GC, Prejean SG, Ellsworth WL (2003) Apparent Break in Earthquake Scaling Due to Path and Site Effects on Deep Borehole Recordings. J Geophys Res 108(B5)2271. doi:10.1029/2001JB001617
36. Jost M, Busselberg LT, Jost O, Harjes HP (1998) Source parameters of injection-induced microearthquakes at 9 km depth at the KTB deep drilling site, Germany. Bull Seismol Soc Am 88:815–832
37. Joyner WB (1984) A scaling law for the spectra of large earthquakes. Bull Seism Soc Am 74:1167–1188
38. Kanamori H (1977) The energy release in great earthquakes. J Geophys Res 82:2921–2987
39. Kanamori H, Anderson DL (1975) Theoretical basis of some empirical relations in seismology. Bull Seismol Soc Am 65:2981–2987
40. Kanamori H, Rivera L (2004) Static and Dynamic Scaling Relations for Earthquakes and their implications for Rupture Speed and Stress Drop. Bull Seismol Soc Am 94:314–319. http://www.gps.caltech.edu/faculty/kanamori/static_dynamic_scaling_relations.pdf
41. Kanamori H, Mori J, Hauksson E, Heaton TH, Hutton LK, Jones LM (1993) Determination of earthquake energy release and M_L using Terrascope. Bull Seismol Soc Am 83:330–346
42. Keilis-Borok VI (1957) Investigation of the mechanism of earthquakes. Tr Inst Geofis Akad Nauk, SSSR 40 (in Russian). (1960) Sov Res Geophys Ser 4 (Engl transl)
43. Kostrov BV (1964) Self-similar problems of propagation of shear cracks. J Appl Math Mech 28:1077–1087
44. Kostrov BV (1974) Seismic moment and energy of earthquakes and seismic flow of rock. Izv Earth Phys 1:23–40
45. Kostrov B, Das S (1988) Principles of Earthquake Source Mechanics. Cambridge University Press, Cambridge
46. Madariaga R (1976) Dynamics of an expanding circular fault. Bull Seism Soc Am 66:639–666
47. Madariaga R (1977) High frequency radiation from crack (stress drop) models of earthquake fauting. Geophys J R Astr Soc 51:625–651
48. Madariaga R (1978) The dynamic field of Kaskell's rectangular dislocation fault model. Bull Seismol Soc Am 68:869–887
49. Madariaga R (1979) On the relation between seismic moment and stress drop in the presence of stress and strength heterogeneity. J Geophys Res 84:2243–2250
50. Mai PM, Beroza GC (2000) Source-scaling properties from finite-fault rupture models. Bull Seis Soc Am 90(3): 604–615. http://www.seismo.ethz.ch/staff/martin/papers/BSSA00scalingRP.pdf
51. Mai PM, Beroza GC (2002) A spatial random-field model to characterize complexity in earthquake slip. J Geophys Res 107:2308. doi:10.1029/2001JB000588. http://www.seismo.ethz.ch/staff/martin/papers/JGR02MaiSlipComplex.pdf
52. Mayeda K, Walker WR (1996) Moment, energy, stress drop, and source spectra of western United States earthquakes from regional coda envelopes. J Geophys Res 101:11195–11208
53. McGarr A (1999) On relating apparent stress to the stress causing earthquake fault slip. J Geophys Res 104:3003–3011
54. McGarr A, Fletcher JB (2003) Maximum Slip in Earthquake Fault Zones, Apparent Stress, and Stick-Slip Friction. Bull Seismol Soc Am 93:2355–2362
55. Ohnaka M (2003) A constitutive scaling law and a unified comprehension for frictional slip failure, shear fracture of intact rock, and earthquake rupture. J Geophys Res 108:B2080. doi:10.1029/2000JB000123
56. Ohnaka M, Shen L-F (1999) Scaling of the shear rupture process from nucleation to dynamic propagation: Implications of geometric irregularity of the rupturing surfaces. J Geophys Res 104:817–844
57. Olsen KB, Madariaga R, Archuleta RJ (1997) Three-dimensional dynamic simulation of the 1992 Landers Earthquake. Science 278:834–838
58. Orowan E (1960) Mechanism of seismic faulting. Geol Soc Am Memoir 79:323–345
59. Prieto GA, Shearer PM, Vernon FL, Kilb D (2004) Earthquake source scaling and self-similarity estimation from stacking P and S spectra. J Geophys Res 109:B08310. doi 10.1029/2004JB003084. http://pangea.stanford.edu/~gprieto/publication/scaling.pdf
60. Rice JR (1980) The mechanics of earthquake rupture. In: Dziewonski AM, Boschi E (eds) Physics of the Earth's Interior. Proceedings of the International School of Physics "Enrico Fermi", Course 78, 1979, pp 555–569. North Holland, Amsterdam
61. Rivera L, Kanamori H (2005) Representations of the radiated energy in earthquakes. Geophy J Int 162:148–155
62. Richards PG (1973) The dynamic field of a growing plane elliptical shear crack. Int J Solids Struct 9:843–861
63. Richardson E, Jordan TH (2002) Seismicity in deep gold mines of South Africa: Implications for tectonic earthquakes. Bull Seismol Soc Am 92:1766–1782
64. Ripperger J, Mai PM (2004) Fast computation of static stress changes on 2D faults from final slip distributions. Geophys Res Let 31:L18610. doi:10.1029/2004GL0594
65. Sato T, Hirasawa T (1973) Body wave spectra from propagating shear cracks. J Phys Earth 21:415–431
66. Savage JC (1974) Relation of corner frequency to fault dimensions. J Geophys Res 77:3788–3795
67. Scholz CH (2002) The mechanics of earthquakes and faulting. Cambridge University Press, Cambridge
68. Shearer PM, Prieto GA, Hauksson E (2006) Comprehensive analysis of earthquake source spectra in southern California. J Geophys Res, 111, B06303, doi:10.1029/2005JB003979. http://pangea.stanford.edu/~gprieto/publication/scsn_spectra.pdf
69. Singh SK, Ordaz M (1994) Seismic energy release in Mexican subduction zone earthquakes. Bull Seismol Soc Am 84:1533–1550
70. Sneddon IN (1951) Fourier Transforms. McGraw-Hill, New York
71. Spudich P, Frazer LN (1984) Use of ray theory to calculate high-frequency radiation from earthquake sources having spatially variable rupture velocity and stress drop. Bull Seismol Soc Am 74:2061–2082
72. Vassiliou MS, Kanamori H (1982) The energy released in earthquakes. Bull Seism Soc Am 72:371–387
73. Wyss M, Brune JN (1968) Seismic moment, stress, and source dimensions for earthquakes in the California-Nevada region. J Geophys Res 73:4681–4684

Earthquakes, Dynamic Triggering of

Stephanie G. Prejean[1], David P. Hill[2]
[1] US Geological Survey, Alaska Science Center, Anchorage, USA
[2] US Geological Survey, Volcano Hazards Program, Menlo Park, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Review of Dynamic Triggering Observations
Characteristics of Dynamic Triggering
Physical Models of Dynamic Triggering
Future Directions
Bibliography

Glossary

Dynamic stress change a transient, often oscillatory change in the Earth's stress field.

Static stress change a permanent, step-like change in the Earth's stress field.

Definition of the Subject

Geoscientists have long sought understanding of how earthquakes interact. Can earthquakes trigger other earthquakes? The answer is clearly yes over short time and distance scales, as in the case of mainshock – aftershock sequences. Over increasing time and distance scales however, this question becomes more difficult to answer. The study of dynamically triggered earthquakes explores the most distant boundaries over which earthquakes trigger other earthquakes.

Dynamic triggeringe to temporary and oscillatory fluctuations in the stress/strain regime in a volume of the Earth's crust. Dynamic stress fluctuations are associated with ground shaking resulting from either anthropogenic activities or natural sources. Dynamic triggering occurs as seismic waves from an initial earthquake propagate through the Earth's crust, triggering secondary earthquakes. Once the seismic wave train has passed and ground shaking ends in a given locale, the crust returns to its previous stress state modified by the combined stress drops associated with any locally triggered earthquakes.

Dynamic triggering includes wide ranging phenomenon, both geographically and in its characteristics. It has been observed across the globe in a variety of geologic and tectonic environments. It has been shown to occur at distances from the initial earthquake rupture varying from meters [21,27,52] to over 11,000 km [103]. In the most distant cases, earthquake triggering results from dynamic stress perturbations as low as 0.01 MPa. Earthquakes have also been shown to trigger other earthquakes at a variety of time scales. In many cases triggering of earthquakes occurs during or within minutes to hours following the responsible seismic waves (e. g. [26,39,74,103]). In other cases, earthquakes occurring weeks to months after the initial earthquake have been interpreted as a delayed response to dynamic triggering (e. g. [41,89,102]). Delayed triggered responses may reflect a more complex series of physical processes. For example, dynamic waves may trigger an aseismic process such as fault creep or changes in a volcanic system, which subsequently triggers earthquakes secondarily [2,4,37,50].

This field has been an area of extensive research in the past twenty five years. It offers a potentially important key to improving our understanding of earthquake nucleation in that, in principle, we can determine in-situ perturbations in the local stress field that lead to earthquake nucleation and rupture. In particular, the availability of broadband seismic data near sites of triggered seismicity allows us to calculate the time history of stress field fluctuations responsible for earthquake nucleation given adequate knowledge of the local seismic velocity structure [28,38,103].

The study of dynamically triggered earthquakes can also help better characterize the physical condition of the Earth's crust at seismogenic depths. Many researchers were surprised that earthquakes could be triggered by stress perturbations as small as 0.01 MPa. This observation indicates the Earth's crust is on the verge of failure in areas with triggered responses to distant earthquakes. This field of research may also provide clues to the hydrologic regime at depth. It has long been recognized that water tables change in response to earthquakes thousands of km distant [17]. The link between dynamically triggered earthquakes and dynamically triggered hydrological changes is an active area of research [8,19,78].

Within the context of complexity and system science, [82] suggest that remotely triggered seismicity may reflect large activation correlation lengths (ACL) in fault systems and stress fields that have reached a state of self-organized criticality. This statistical physics approach to earthquake occurrence focuses on the exploration of both analog and computational models that can mimic observed dynamical space-time patterns spanning a wide range of spatial-temporal scales. It is not concerned with inferred (or "non-observable") physical models for the lo-

cal processes linking dynamic stresses and brittle failure (triggered earthquakes) on faults (see [82]). In this review, however, we focus on these physical models together with a description of documented patterns of remote dynamic triggering.

Introduction

Introduction to Stress Triggering of Earthquakes

Earthquake triggering refers to a process by which any change in fault properties or the processes acting on a fault leads to rupture initiation. More specifically, stress triggering occurs when a change in the stress field acting on a fault leads to rupture. Stress triggering of earthquakes can result from stresses applied over a variety of time scales and with a variety of frequencies, which generally fall into three partially overlapping categories, 1) static stress triggering, 2) quasi-static stress triggering, and 3) dynamic stress triggering. In the case of static stress changes, the state of stress acting across a fault is permanently perturbed. This form of stress triggering is important in the near field of an earthquake where fault displacement significantly alters the stress field in the surrounding crust. Static stress triggering is commonly regarded as the dominant factor controlling aftershock generation (e. g. [54]). Because static stress changes decay rapidly with distance from the earthquake rupture (as d^{-3}, where d is distance from the earthquake epicenter), they are generally thought to be significant only within two to three fault lengths of the earthquake rupture. The role of static stress changes in triggering aftershocks and other earthquakes has been a vigorous and productive area of research in the past two decades (for reviews see [34,53,92,93]). The relative importance of static vs. dynamic triggering of aftershocks in the near field, however, has recently become an actively debated topic (e. g. [21,71]).

Because static stress changes in the near field develop essentially simultaneously with earthquake rupture, simple static stress triggering must appeal to other mechanisms to explain the time delay associated with many aftershocks and subsequently triggered earthquakes. In contrast, quasi-static stress triggering results from viscoelastic relaxation of the crust after an earthquake. Because quasi-static stress changes decay as d^{-2} and because viscoelastic relation is a time dependent process, these stress changes may explain triggered earthquakes more distant from an initial earthquake and triggered earthquakes with delay times from years to decades [72].

Dynamic stress changes decay more slowly with distance than either static stress changes or quasi – static stress changes (as $d^{-1.5}$ for surface waves). Thus dynamic stress changes become increasingly dominant with increasing distance from the fault rupture. In this review we discuss dynamic triggering of earthquakes resulting from ground shaking due to the passing seismic wavetrain of other earthquakes. We focus on dynamic triggering due to remote earthquakes (greater than two fault lengths distance from the earthquake rupture), though we will briefly discuss the active research topic of dynamic triggering in an earthquake's aftershock zone as well. We also limit discussion to frequencies of ground shaking above ~ 0.01 Hz (periods less than ~ 100 s), though some work has been done on earthquakes triggered by longer wavelength fluctuations. For example, [16,95] recently found evidence that solid Earth tides can modulate background seismicity rates. Other reviews of dynamic stress triggering can be found in [23,37,92] and [38].

Brief History of Dynamic Stress Triggering Research

The ability of earthquakes to trigger other earthquakes at great distances has been discussed in scientific literature throughout the latter half of the 20th century (see [38] for review). However, making a credible case for a causal link between two earthquakes remains a major challenge in this field. Beyond the realm of aftershock zones, it was difficult to justify statistically that one earthquake triggered another until the 1980s and 1990s. By then continuously recording telemetered seismic networks and automated processing of data became commonplace, providing reliable spatial and temporal records of earthquake occurrence at $M \geq 1-2$ and the statistical leverage associated with large numbers of small earthquakes.

Dynamic triggering of earthquakes was widely accepted in the scientific community following the 1992 M 7.3 Landers earthquake in southern California. In the minutes to days following the Landers earthquake, earthquake rates increased dramatically across the western United States at distances well beyond the aftershock zone [39]. Earthquakes were triggered throughout California, Nevada, Utah, Wyoming, and Idaho at distances of up to 1250 km (Fig. 1). Although time delays of triggered events ranged from seconds to 33 hours after the arrival of the Landers earthquake wavetrain, the sudden increase in seismicity across the Western United States could not be ignored. This earthquake spawned a plethora of studies into the nature of earthquake triggering and remains one of the best studied triggering episodes to date.

The geophysical community had a unique research opportunity when the 1999 M 7.1 Hector Mine earthquake occurred. Because it was an earthquake with similar magnitude to the Landers earthquake occurring in a similar

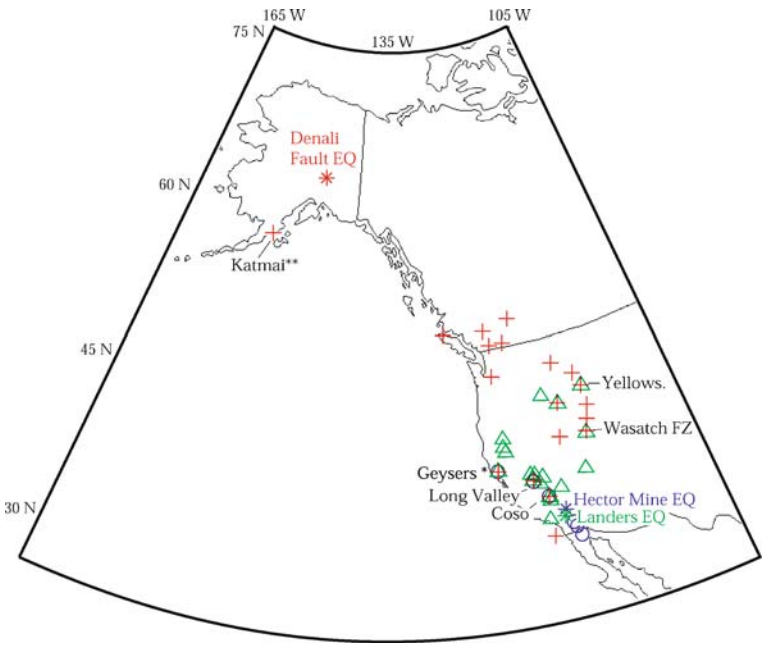

Earthquakes, Dynamic Triggering of, Figure 1
Map showing sites of triggered seismicity in western North America from the Landers (*green triangles*), Hector Mine (*blue circles*), and Denali Fault (*red crosses*) earthquakes. Modified from [38], Treatise on Geophysics

location, it provided leverage to tune ideas about dynamically triggered seismicity and the underlying physical processes. Although the Hector Mine earthquake triggered seismicity at some of the same locations as the Landers earthquake, the Hector Mine earthquake had a much more limited triggered response (Fig. 1) [30]. The difference is likely due in part to differences in seismic radiation patterns between the two earthquakes [30,43]. The Landers earthquake ruptured unilaterally to the north, whereas the Hector Mine earthquake ruptured bilaterally, primarily to the south.

Following the Landers and Hector Mine earthquakes, the search for dynamic triggering began in earnest. Researchers across the globe began scanning earthquake catalogs and waveform data searching for dynamic triggering in a wide variety of environments (see "Sect. Review of Dynamic Triggering Observations"). At the Geysers, CA alone [91] identified 7 episodes of dynamic triggering between 1988 and 1994, making this among the most frequently triggered locations known.

Like the Landers earthquake, the 2002 M 7.9 Denali Fault earthquake triggered a widespread response across western Canada and the United States (Fig. 1) [27,33,45,64,68,74]. The increase in the number of broadband and high-dynamic range seismometers by 2002 allowed scientists to visually scan on-scale seismic data during the earthquake's wavetrain. Thus, many triggered events were detected which were absent from earthquake catalogs. As an example of how instrumentation improvements increase our ability to detect dynamic triggering, [45] point out that the triggered response of the Yellowstone caldera to the Denali Fault earthquake could not have been detected at the time of the Hector Mine earthquake only three years earlier.

Review of Dynamic Triggering Observations

Detection of Dynamic Triggering

Dynamic triggering has been observed in a variety of locations around the globe. Some suggest that dynamic triggering of earthquakes is a ubiquitous process in the Earth's crust (e. g. [26,41]). Others suggest that some areas are more likely to experience dynamic triggering of earthquakes than others (e. g. [39,64]). Observations of dynamic triggering are limited geographically due to uneven seismic network coverage and the effort applied to examining seismic data.

Following the 1992 M 7.3 Landers earthquake, dynamic triggering was recognized by the sudden increase in the number of earthquakes located through standard network processing across the western US in the days to weeks after the large earthquake. Searching earthquake

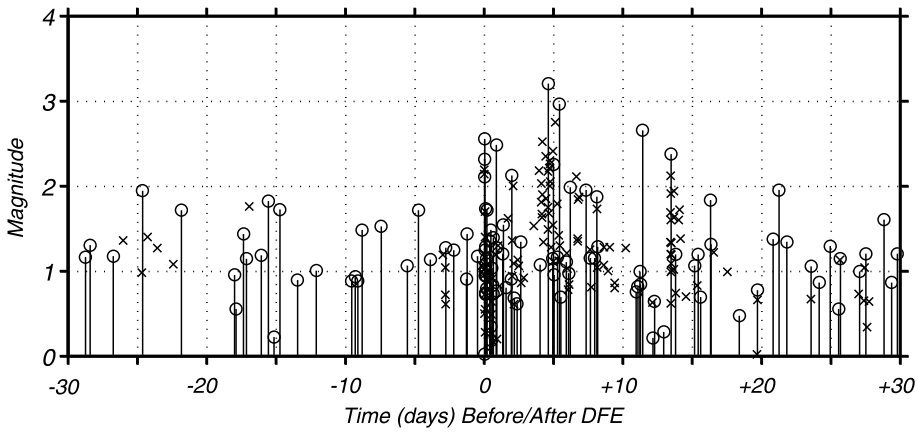

Earthquakes, Dynamic Triggering of, Figure 2
Plot of earthquake magnitude versus time in the area of the Wasatch Front, Utah, 30 days before and after the Denali Fault earthquake. Circles represent independent events. *Crosses* indicate secondary events determined by declustering the earthquake catalog. Figure reprinted from [68], BSSA

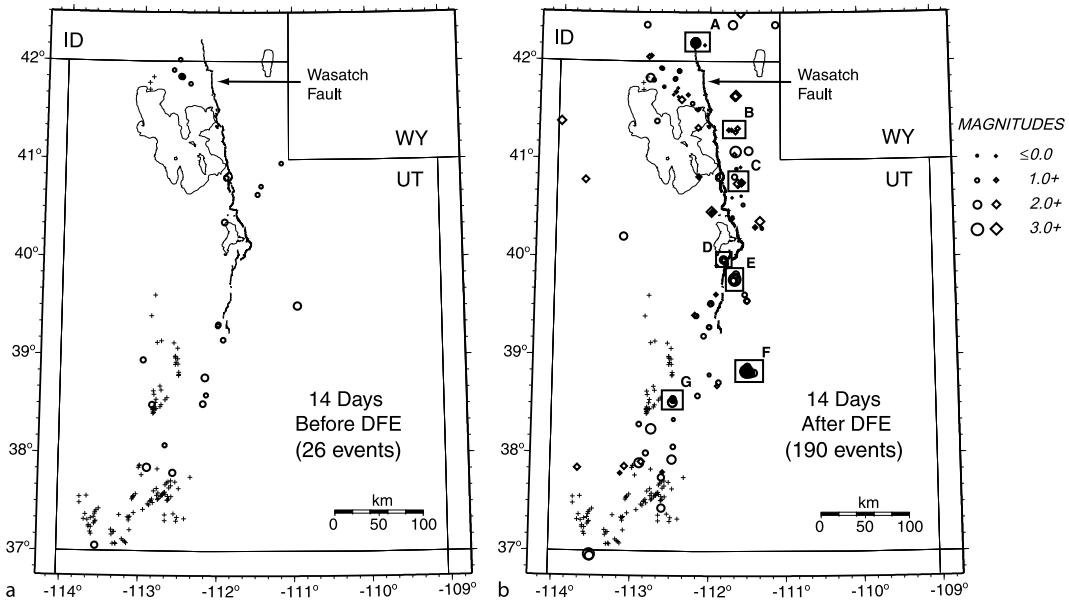

Earthquakes, Dynamic Triggering of, Figure 3
Seismicity in Utah 14 days before and after the Denali Fault earthquake. *Diamonds* in b show earthquakes occurring in the first 24 hours after the arrival of the wave train from the Denali Fault earthquake. *Cross* are locations of quaternary volcanic vents [3]. Figure reprinted from [68], BSSA

catalogs for sudden increases in seismicity after a large earthquake is one common method of identifying dynamic triggering (Figs. 2 and 3) (e.g. [6,30,39,41,68]). Identifying triggered seismicity using earthquake catalogs simplifies interpretation of triggered response with respect to background seismicity rates, as earthquake catalogs often provide stable long-term records of earthquake occurrence at a consistent threshold.

A second commonly used method of detecting dynamic triggering involves visually scanning continuous seismic data shortly before and after a large earthquake to identify a sudden increase in earthquakes too small to be detected and located through standard network processing (Fig. 4). This latter method is effective at identifying triggered seismicity in sparsely instrumented areas, identifying very small triggered earthquakes, and identifying

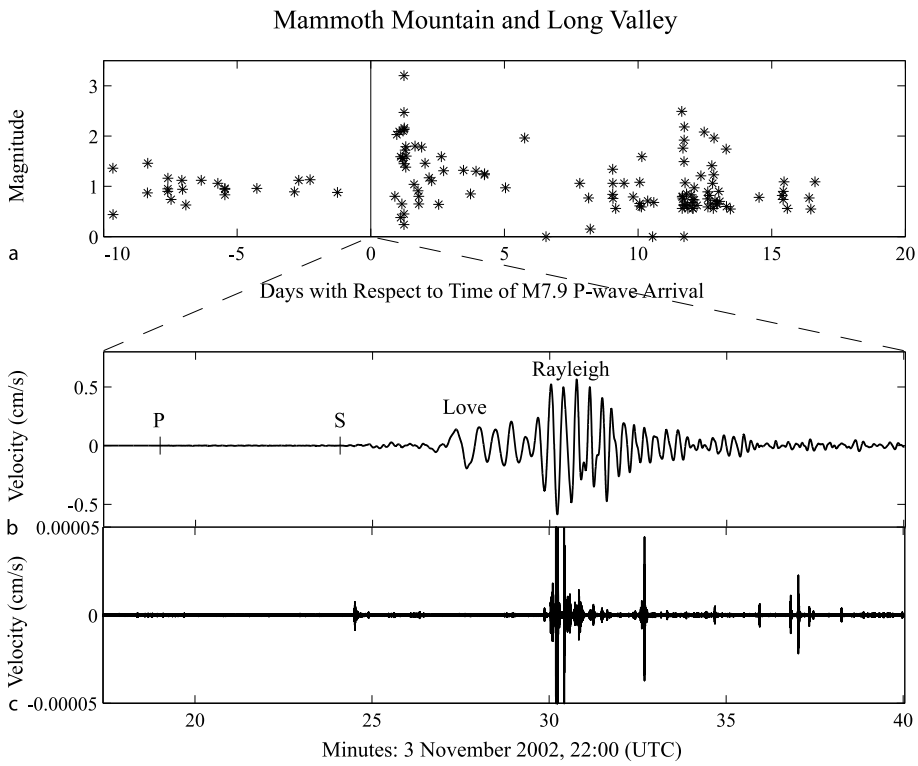

Earthquakes, Dynamic Triggering of, Figure 4

Seismicity triggered at Mammoth Mt. b-c and within the Long Valley caldera, California, a following the Denali Fault earthquake. a Catalog from NCEDC showing two swarms following the Denali fault earthquake in the caldera's south moat. The two lower panels show data from this very small earthquake swarm recorded on the broadband UNR/USGS station OMM from rotated to transverse direction, showing Denali earthquake wavetrain at Long Valley. Major arrivals are labeled. c Record from b high pass-filtered, showing small local earthquakes occurring during Denali wavetrain. Modified from [74], BSSA

earthquakes that occurred during the wavetrain from the initial earthquake (e. g. [26,47,64,74,103]). This method of detecting triggering has become more common with increasing availability of continuously recorded high-dynamic-range seismic data.

Possible instances of dynamic triggering have also been proposed based on historical accounts [40,42,44,59]. Because these studies rely on felt reports, they are generally limited to moderate to large sized triggered earthquakes that are separated in time by days to months.

With any method of detecting dynamic triggering, it must be shown that one earthquake is likely causally linked to the dynamic waves radiating from a previous earthquake, rather than by coincidence. Earthquakes near each other in time are more likely to be related than earthquakes separated in time. Additionally, earthquakes unlikely to occur randomly (e. g. large events in seismically quiet areas) are more likely to be related than earthquakes occurring commonly as background seismicity (e. g. small earthquakes in a seismically active area). To calculate the probability that two earthquakes are related, one must first calculate the probability of each occurring randomly. This is usually done using patterns of earthquake occurrence based on local earthquake catalogs. The most commonly used statistical test to identify whether an increase in number of earthquakes is statistically significant is the Beta statistic [58]. Pankow et al. [68] also employ a binomial distribution analysis to this end. These techniques have potential pitfalls however, as they rely on assumptions about earthquake distributions and compare snapshots of seismicity in time in regions where seismicity rates fluctuate regularly [58]. Objectively determining whether one earthquake is genetically related to another remains a challenge.

Because spatial-temporal clusters of earthquakes are less common than isolated events, clusters of earthquakes temporally coincident with dynamic stresses are more easily identified as being triggered than isolated earthquakes. In the case of earthquake clusters, however, it may be difficult to discriminate between earthquakes directly trig-

gered by dynamic stresses from a remote earthquake and secondary aftershocks to directly triggered events [9]. To address this, earthquake catalogs are frequently 'declustered' (e. g. [45,68]). This process involves decomposing a catalog into primary and secondary earthquakes based on statistical patterns of aftershock sequences (e. g. [76]). [68] and [9] show that in some cases triggered seismicity is modeled well as an aftershock sequence. In other cases, however, triggered seismicity swarms cannot be dismissed as secondary aftershock sequences (e. g. [74,103]). In such cases it is likely that most events in a swarm were triggered directly by the dynamic waves radiated from a distant earthquake or perhaps as a secondary response to some aseismic process (e. g. fluid flow, or local deformation associated with fault creep) triggered by the dynamic waves.

Dynamic Triggering in Volcanic and Geothermal Regimes

Although dynamic triggering has been observed in a variety of environments, many of these observations are from areas with active volcanic and hydrothermal systems (Table 1) [6,73,74]. These areas typically have high background seismicity rates indicating that the crust habitually hovers near failure and thus is particularly susceptible to dynamic triggering. Furthermore, because these areas tend to be well instrumented, dynamically triggered earthquakes may be unusually easy to detect. Here we summarize observations of triggered seismicity in volcanic and hydrothermal areas.

The Geysers geothermal field in northern California is among the most frequently triggered sites known with 9 cases of dynamic triggering documented in the past 20 years [28,74,91]. Earthquakes that have caused triggering at the Geysers range in magnitude from 6.9 to 7.9 and in distance from 212 km to 3120 km. The Coso geothermal field in southern California has also experienced repeated episodes of dynamic triggering following the M 7.3 Landers, the M 7.1 Hector Mine, and the M 7.9 Denali Fault earthquakes (Fig. 1) [39,74]. Following the Hector Mine earthquake, dynamically triggered earthquakes and ground deformation were observed near a third geothermal field – Cerro Prieto, Baja California [24,30].

Yellowstone, Wyoming is a larger and more complicated system than the geothermal fields mentioned above, as it is a caldera system characterized by ongoing magmatic and tectonic activity, in addition to hydrothermal activity. Yellowstone had a triggered response to the M 7.3 Landers earthquake [39] and the M 7.9 Denali Fault earthquake [45,47]. The triggered response to the Denali Fault earthquake was particularly dramatic (Figs. 5 and 6). Seismicity increased immediately following the arrival of surface waves from the Denali Fault earthquake and remained unusually high for 30 days with magnitudes ranging from < 0.0 to M 3.2 [45]. The time scale of the triggered response was spatially variable (Figs. 5 and 6). Because the Denali Fault earthquake led to immediate triggering in the area of geysers and affected periodicity of geysers at Yellowstone, it is likely that changes in the hydrothermal regime induced a triggered response in some areas ([45,46]. In other areas however, the development of triggered earthquake sequences was delayed and similar to commonly observed tectonic activity [45].

Like the Yellowstone caldera, the Long Valley caldera experiences dynamic triggering with complex characteristics. The area responded to the Landers earthquake [39], the Hector Mine earthquake [50], and the Denali Fault earthquake [50,74] both seismically and geodetically, although each response varied in location and intensity (Fig. 7). The Landers earthquake produced the largest triggered response with 340 earthquakes in seven days up to M 3.4 throughout the south moat of the caldera [39]. The seismic response to the Hector Mine earthquake was comparatively short lived and limited to the region of Mammoth Mountain. After the Denali fault earthquake, the caldera area experienced two phases of triggered seismicity. A burst of ~ 60M ≤ 0.8 earthquakes occurred beneath Mammoth Mountain during and shortly after the arrival of the surface waves from the Denali Fault earthquake [74]. Twenty-four hours later a larger swarm of earthquakes of M ≤ 3.4 occurred in the Long Valley caldera's south moat [74]. All three episodes of dynamic triggering in the Long Valley caldera were accompanied by deformation transients with geodetic moments an order of magnitude larger than the cumulative seismic moment of the triggered seismicity [36,50], though the time history and magnitude of each deformation response varied.

Iwo Jima, Japan, a volcanic island hosting a Holocene eruption, geothermal activity, and historic phreatic (steam) eruptions is a third complex caldera system which has experienced dynamic triggering of local earthquakes. [98] examined continuous waveform data of 21M > 7 earthquakes < 3000 km distance from Iwo Jima, and identified 4 cases of resulting increased local seismicity. In all cases earthquakes were triggered locally during surface wave arrivals and persisted for 6–15 minutes.

Dynamic triggering of earthquakes has been observed at shallow depths in volcanic edifices at a variety of locales (Table 1). In the Pacific Northwest, Mt. Rainier experienced 6–8 M < 0 earthquakes during the wavetrain of the Denali Fault earthquake and 8M ≤ 0.9 earthquakes in

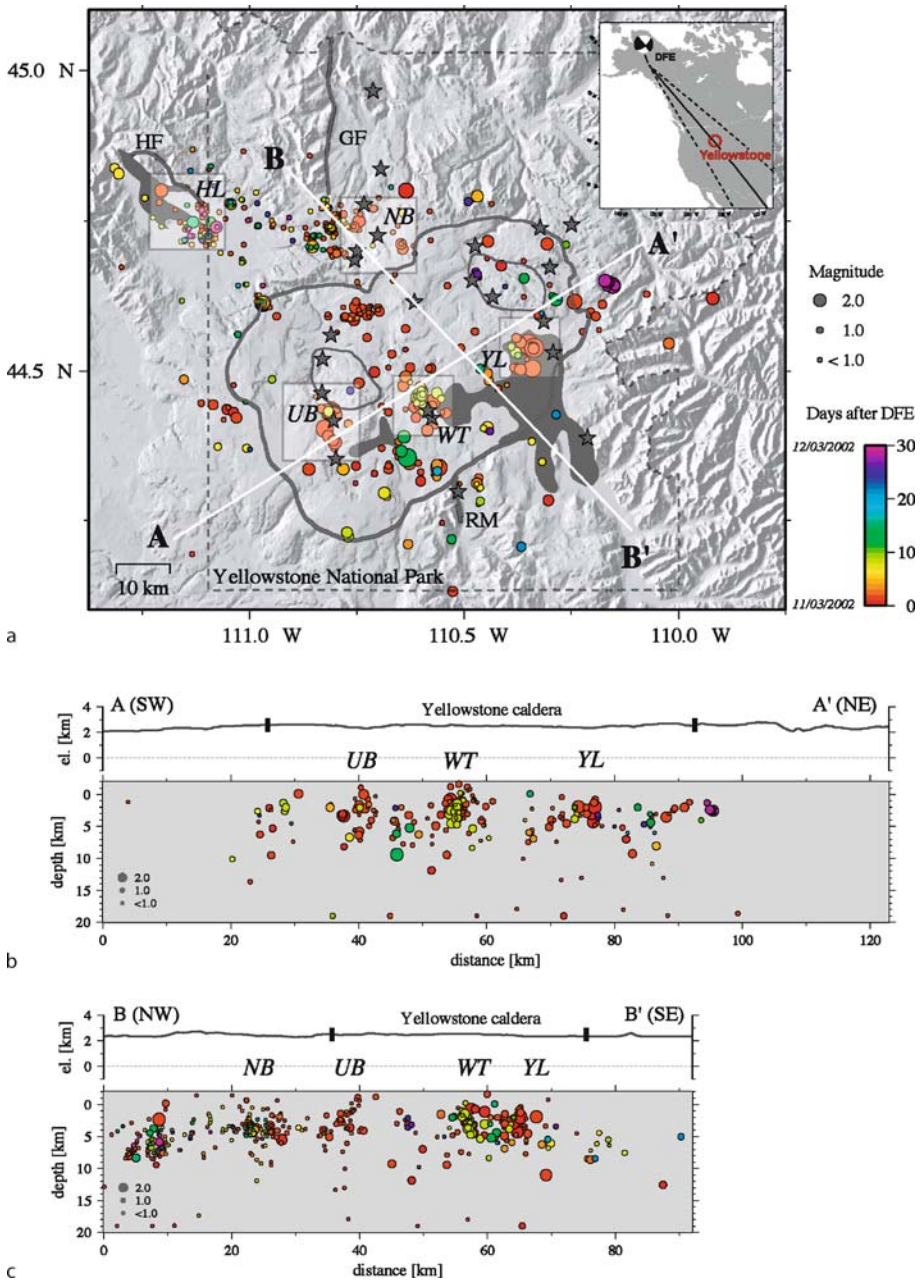

Earthquakes, Dynamic Triggering of, Figure 5

Seismicity within one month of the Denali Fault earthquake at Yellowstone caldera, *color* coded with time: a earthquake locations, b cross sections along AA′, c cross section along BB′. Specific areas labeled: HL, Hebgen Lake area; NB, Norris geyser basin; UB, Upper geyser basin; WT, West Thumb geyser basin; YL, northern end of Yellowstone Lake. Large normal faults are represented with *thick black lines*: RM, Red Mountain fault zone; GF, Gallatin fault; HF, Hebgen and Red Canyon faults. *Inset* shows location of Denali Fault earthquake and Yellowstone. *Solid and dashed lines* in *inset* show the great circle path +/—10 degrees along the strike of the Denali Fault earthquake. Figure reprinted from [45], BSSA

Earthquakes, Dynamic Triggering of, Table 1
Published occurrences of discrete remotely triggered earthquakes

Locations	Citation	Distance (km)	Triggering Earthquake	M	Onset	M_max	Reg	Env
Aso, Japan	[61]	1400	Chi-Chi, 1999	7.7	During P waves	–	E	V
British Columbia	[26]	1800–2200	Denali Fault, 2002	7.9	During surface and coda waves	–	–	N
Burney, CA	[39]	900	Landers, 1992	7.3	23 hour	2.8	E	N
Central and Southern, CA	[41]	variable	15 Central and Southern CA earthquakes, 1988–2004	5.3–7.1	Within 1 month	–	–	–
Cerro Prieto, Mexico	[24]	260	Hector Mine, 1999	7.1	–	4.1	E	V,G
Coso, CA	[39]	165–205	Landers, 1992	7.3	~ 3 hour	4.4	E	G
Coso, CA	[74]		Hector Mine, 1999	7.1	–	–	E	G
Coso, CA	[74]	3,660	Denali Fault, 2002	7.9	15 min	2.3	E	G
Geysers, CA	[28]	2,500	Gulf of Alaska, 1988	7.6	–	0.2–2.5	E	G
Geysers, CA	[28]	212	Loma Prieta, 1989	7.1	–	0.2–2.5	E	G
Geysers, CA	[28]	443	Off Oregon Coast, 1991	6.9	–	0.2–2.5	E	G
Geysers, CA	[28]	390	Gorda Plate, CA, 1991	7.1	–	0.2–2.5	E	G
Geysers, CA	[28]	202	Petrolia, CA, 1992	7.0	–	0.2–2.5	E	G
Geysers, CA	[39]	740	Landers, 1992	7.3	3 min	1.6	E	G
Geysers, CA	[91]	635	Northridge, 1994	6.6	–	–	E	G
Geysers, CA	[28]	308	Cape Mendoceno, CA, 1994	6.9	–	0.2–2.5	E	G
Geysers, CA	[74]	3,120	Denali Fault, 2002	7.9	12 min	2.5	E	G
Greece	[6]	400–1000	Izmit, 1999	7.4	After surface waves	3.5	E	G
Iceland	[2]	64–78	South Iceland Seismic Zone, 2000	6.5	< 5 min	5	E	V, G
Idaho, Cascade	[39]	1100	Landers, 1992	7.3	33 hour	1.7	E	G
Idaho, Cascade	[26,47]	2300	Denali Fault, 2002	7.9	During Rayleigh waves	4.6	E	G
Iwo Jima, Japan	[98]	≤ 2009 km	4 earthquakes, 1983–1993	7.1–8.0	During surface waves	< 2	–	V,G
Katmai volcanoes	[73]	115	1999	7.0	< 3 min	2.3	–	V, G
Katmai volcanoes	[64]	122	2000	6.8	–	0.9	–	V, G
Katmai volcanoes	[64]	161	2001	7.0	< 2 min	1.5	–	V, G
Katmai volcanoes	[64]	161	2001	6.8	–	–	–	V, G
Katmai volcanoes	[64]	740	Denali Fault, 2002	7.9	3.9 min	2.0	–	V, G
Katmai volcanoes	This paper	3620	Kurile, 2007	8.2	During surface waves	–	–	V, G
Lassen, CA	[39]	840	Landers, 1992	7.3	12 min	2.8	E	V, G
Little Skull Mt., NV	[39]	240	Landers, 1992	7.3	1.5 hour	5.6	E	N
Long Valley, CA	[39]	415	Landers, 1992	7.3	9 min.	3.4	E	V, G
Long Valley, CA	[74]	3,454	Denali Fault, 2002	7.9	23.5 hour	3.0	E	V, G
Mammoth, CA	[50]	450	Hector Mine, 1999	7.1	20 min.	–	E	V, G
Mammoth, CA	[74]	3,454	Denali Fault, 2002	7.9	17 min.	0.8	E	V, G
Mono Basin, CA	[39]	450	Landers, 1992	7.3	19 hour	3.1	E	N
Mt. Rainier, WA	[74]	3,108	Denali Fault, 2002	7.9	12 min.	0.0	E	V
Mt. Rainier, WA	[74]	3,108	Denali Fault, 2002	7.9	2.5 hour	0.9	E	V
Nanki Trough, Japan	[60]	900–1400	Tokachi-oki, 2003	8.1	After surface waves	–	S	–

the following days. In response to the Landers earthquake, Mt. Lassen in northern California hosted 14 earthquakes of $M \leq 2.8$. Volcanoes in the Katmai Volcanic Cluster, Alaska, have experienced triggered seismicity on at least seven occasions since 1999 ([64,65,73], this chapter). The largest of these triggered responses included 17 earth-

Earthquakes, Dynamic Triggering of, Table 1
(continued)

Locations	Citation	Distance (km)	Triggering Earthquake	M	Onset	$M_m$ax	Reg	Env
The Netherlands Roer Valley	[12]	40	Roermond, 1992	5.4	–	3.7	E	N
New Madrid, MO	[41]	1000	1811–1812 New Madrid	∼ 7.8	–	–	C	N
Offshore Southern CA	[74]	4,003	Denali Fault, 2002	7.9	Mainshock coda	2.5	E	N
Salton Sea, CA	[43]	120–150	Hector Mine, 1999	7.1	–	4.7	E	V,G
Syria – Lebanon border	[62]	500	Gulf of Aqaba, 1995	7.3	2 hr 47 min	3.7	C	N
Taiwan	[102]	variable	9 earthquakes, 1978–1994	6.5–7.1	≤ 15 days	≥ 4.0	–	V,G
Tonga Trench	[96]	290–313	Tonga Region, 2002	7.6	2 min.	7.7	S	–
Utah, Cedar City	[39]	490	Landers, 1993	7.3	39 min.	4.1	E	G
Utah, Wasatch Front	[68]	3000–3500	Denali Fault, 2002	7.9	During surface waves	3.2	E	G
Valley of Mexico	[89]	303–588	7 earthquakes	7.6–8.0	–	∼ 4.0	E	V,G
Western Nevada	[1,39]	450–650	Landers, 1993	7.3	9 min.	4.0	E	G
White Mts., CA	[39]	380–420	Landers, 1993	7.3	11.6 hour	3.7	E	N
Mt. Wrangell, AK	[45]	11,000	Denali Fault, 2002	7.9	During Rayleigh waves	1.0	–	V, G
Yellowstone	[39]	1250	Landers, 1993	7.3	1.8 hour	2.1	E	V, G
Yellowstone	[45]	120	Hector Mine, 1999	7.1	During surface waves	–	E	V,G
Yellowstone	[45]	3150	Denali Fault, 2002	7.9	During Love waves	3.2	E	V, G

Location is location of triggered seismicity. Distance is distance from location of triggering to triggering earthquake (mainshock) epicenter. M is magnitude of mainshock. Onset is onset time of triggered activity with respect to arrival of waves from mainshock. M_{max} is magnitude of the largest triggered earthquake. Reg describes stress regime: E-extensional or transtensional, C-compressional or transpressional, S-subduction zone. Env describes if the area is volcanic (V), geothermally active (G), or neither (N). – indicates data not available or inconclusive.

quakes of M ≤ 2.3. During the wavetrain of the 2006 M 8.7 Sumatra–Andaman Islands earthquake, Mt. Wrangell, Alaska had triggered 14 earthquakes [103]. With the exception of the Mt. Lassen response following the Landers earthquake and the delayed events at Mt. Rainier following the Denali Fault earthquake, these earthquakes triggered in volcanic edifaces were too small to be detected and located by automatic processing systems.

The Valley of Mexico is a large volcanically and geothermally active area located in the Trans Mexican Volcanic Belt. [89] searched for dynamically triggered earthquakes in the Valley of Mexico following 18 M ≥ 7.0 Mexican earthquakes between 1920 and 1998. In seven cases, they found evidence for dynamic triggering of earthquakes within 2 days of a large earthquake. In four additional cases, seismicity increased after a large earthquake, but was delayed by up to one month. Because this study used only one station however, the potentially triggered events can only be located to within some ill-defined region surrounding the station.

The South Iceland Seismic Zone is a transform zone in a volcanically and geothermally active area. In 2000, a Mw = 6.5 earthquake in the South Iceland Seismic Zone triggered widespread seismicity, including three Mw ∼ 5.0 earthquakes within 5 minutes of its occurrence. Coulomb failure stress calculations indicate that the two M > 5 earthquakes located ∼ 100 km to the west on the Reykjanes Peninsula are beyond the range where static stress changes are significant [2], and thus appear to have been dynamically triggered. Furthermore, one of these M > 5 earthquakes had a geodetic moment significantly larger than it seismic moment suggesting that deformation associated with aseismic fault creep may have indirectly triggered many of the smaller earthquakes in the area [2].

Dynamic Triggering in Regimes with Limited Volcanic and Geothermal Activity

Extensional and Transtensional Environments The majority of occurrences of triggered seismicity documented to date have been in extensional or transtensional tectonic regimes (Table 1). In the western United States dynamic triggering following the M 7.3 Landers earthquake occurred exclusively in transtensional tectonic regimes, many of which were also volcanically or geother-

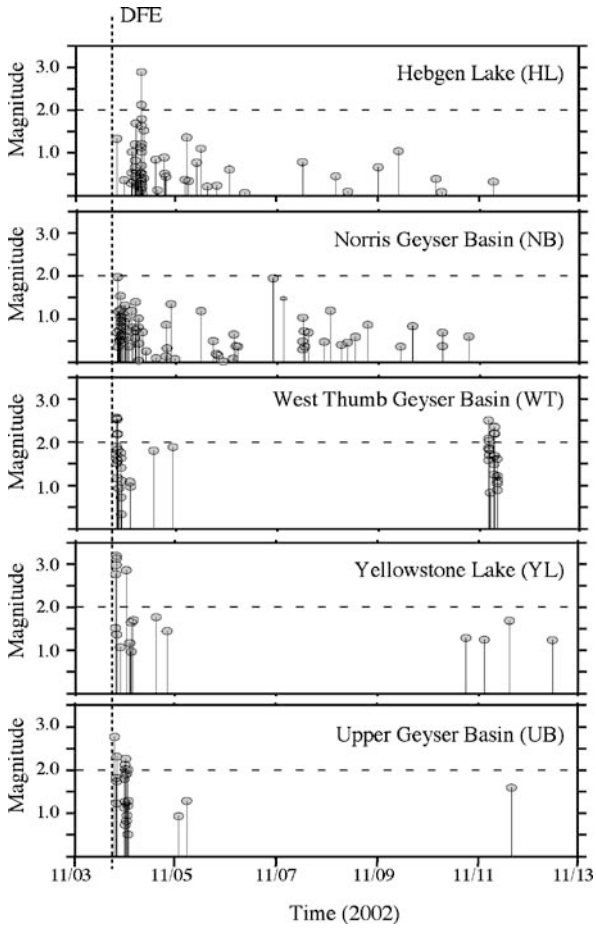

Earthquakes, Dynamic Triggering of, Figure 6
Plot of earthquake magnitude versus time of seismicity for selected areas in Yellowstone caldera. See Fig. 5 for locations of these areas. Dashed line DFE as the origin time of the Denali Fault Earthquake. Figure reprinted from [45], BSSA

mally active, to distances of up to 1250 km [1,39]. These locations included Little Skull Mountain, Nevada; western Nevada, White Mountains, California, Mono Basin California, Cedar City Utah, the Wastach Front in central Utah, Burney, California, and Cascade, Idaho. The onset of triggering ranged from during the passage of the Landers wavetrain to 33 hours after the Landers earthquake. The largest of these earthquakes was a M = 5.6 earthquake triggered beneath Little Skull Mountain, Nevada. Otherwise, triggered earthquakes had M =< 3.0. The most vigorous responses containing tens to hundreds of triggered earthquakes occurred near Cedar City Utah, Western Nevada, and Cascade Idaho.

The Mw = 7.1 Hector Mine earthquake also led to an impressive display of triggered earthquakes in exclusively extensional, transtensional, and geothermal environments in the western United States. Triggered earthquakes began during the passage of the wavetrain near the Salton Trough in Indio and at the southern end of the Salton Sea [30,43]. In general, the triggered response to Hector Mine was less extensive and energetic than that of the Landers earthquake [30].

Following the Denali Fault earthquake seismicity was triggered in several extensional and transtensional areas in the western United States. [47] detected a M 4.6 earthquake triggered during the Denali Fault earthquake wavetrain in Cascade, Idaho. Seismicity remained elevated for 25 days along a 500 km stretch of the Intermountain seismic belt in Utah, on the border of the Basin and Range province (Figs. 2 and 3) [68].

Through examining historical documents [59] identify several earthquakes in extensional/transtensional environments that may have been dynamically triggered by the Mw = 7.8 1906 San Francisco earthquake, including a M 3.5 and M 4.5 earthquake in western Nevada and a M 6.1 earthquake in the Brawley Seismic Zone near the Salton Sea in Southern California. These events are within 400–700 km from the fault rupture, thus beyond the aftershock zone of the San Francisco earthquake.

In the day following the arrival of surface waves from the Mw = 7.4 Izmit, Turkey earthquake, catalog seismicity rates throughout continental Greece, 400–1000 km from the epicenter, increased significantly [6]. Greece is an area of active extension and hosts significant hydrothermal activity. Although [6] did not address a possible correlation to hydrothermal activity systematically, at least some clusters of dynamically triggered seismicity occurred in areas with active hot springs.

A second report of dynamically triggered seismicity in Europe comes the Roer Valley, the Netherlands. This area is an actively extending northern branch of the Rhine Graben System. Following a Mw = 5.4 earthquake in 1992, [12] determine that a large cluster of aftershocks occurred at distance of 40 km from the mainshock. They conclude that these events are dynamically triggered because they are located beyond the zone where static stress changes are significant.

Transpressional and Compressional Environments
Although dynamic triggering is not commonly observed in compressional environments, several studies suggest it does occur. Less than three hours after a Ms = 7.3 earthquake in the Gulf of Aqaba 1995, an earthquake swarm began 500 km distant from the mainshock epicenter in a restraining bend of the Dead Sea transform fault on the Syria–Lebanon border [62]. The swarm consisted of 21 earthquakes of Md $\leq$ 3.7.

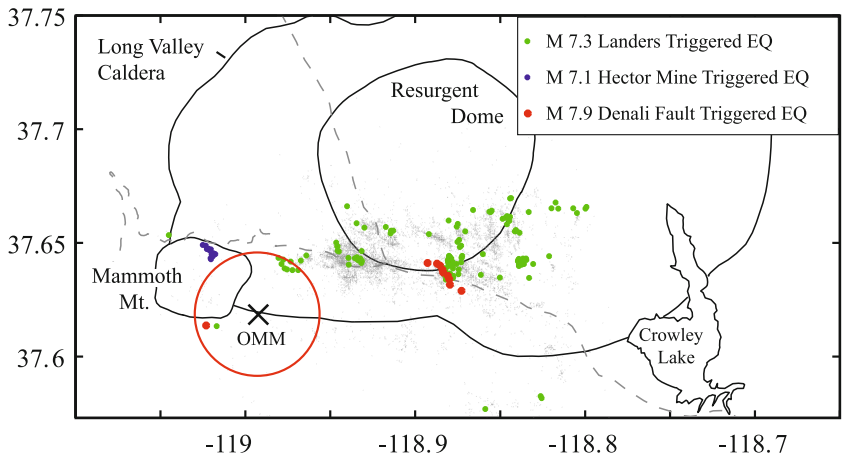

Earthquakes, Dynamic Triggering of, Figure 7

Map of triggered seismicity beneath Long Valley caldera and Mammoth Mountain, California, for the Landers (*green*), Hector Mine (*blue*), and Denali Fault (*red*) earthquakes. *Gray dots* show background seismicity from 1997–1998. The *red circle* centered on station OMM indicates area within which the earthquakes triggered by the Denali Fault earthquake must be located based on S-P phase arrival times. The *single red dot* was large enough to be located [74]. Modified from [38], Treatise on Geophysics

The central United States is a transpressional environment with low strain rates. Dynamic triggering in the central US has not yet been observed instrumentally. However, [40,44,66] suggest that dynamic triggering occurred during the 1886 Charleston, South Carolina earthquake and 1811–1812 New Madrid earthquakes based on examination of historical felt reports. Similarly, [42] describe historical evidence for dynamic triggering of a M $\sim$ 7 earthquake following the 1905 Kangra earthquake in India.

The stress state in Taiwan is variable, but generally transpressional [104]. [102] searched for dynamically triggered seismicity in the Taiwan region following 12 regional M 6.5+ earthquakes occurring between 1973 and 1994. They identify 9–10 cases of increased seismicity following a large event, although the increase is small in all cases, with 1–7 M 4–4.5 earthquakes more in the 15 days following the large earthquake than in the 15 days before.

Dynamic Triggering in Subcrustal Environments

The occurrence of dynamically triggered earthquakes in subcrustal environments has been investigated in subduction zones in South America and Japan. [96] found that a M 7.6 earthquake at 598 km depth in the Tonga trench in 2002 was followed by M 5.9 and M 7.7 earthquakes at 647 and 664 km depth within 2 and 7 minutes of the initial earthquake, respectively. By investigating the rupture history and Coulomb stress change resulting from the initial event, they conclude that the secondary events were triggered dynamically. They highlight 4 additional earthquakes of > 450 km depth which have similar large aftershocks that may be dynamically triggered. During the surface waves of the M 8.1 Tokachi-oki earthquake, [60] identified deep low frequency earthquakes triggered in the Nankai subduction zone through analyzing Hi-Net borehole seismic data and use of the Beta statistic. The triggering occurred during a slow slip event in a region of the subduction zone which was active with deep low frequency tremor.

Triggered Tremor

With the exception of triggered subduction zone seismicity described above, the majority of dynamically triggered earthquakes described in this review are typical brittle failure earthquakes. For example, [43] show that the earthquakes triggered by the Hector Mine earthquake near the Salton Sea had typical spectra and stress drops, consistent with standard brittle failure source mechanism. In the last few years however researchers have demonstrated that volcanic tremor and deep non-volcanic tremor respond to dynamic waves from regional and teleseismic earthquakes as well as typical crustal earthquakes (Table 2) ([32,60,61,81]). These findings emphasize that dynamic triggering can occur in a wide variety of environments and affect multiple seismic processes in addition to brittle failure of crustal rock. They provide an intriguing new perspective on the triggering processes.

At Aso volcano, Japan, [61] identify dynamically triggered earthquakes and volcanic tremor following the 1999

Earthquakes, Dynamic Triggering of, Table 2
Published occurrences of dynamically triggered tremor

Site	Citation	Triggering Earthquake	M**	Type of tremor	Responsible phase
Aso volcano, Japan	[61]	Chi-Chi, 1999	7.7	Shallow volcanic	P waves
Cascadia subduction zone, Canada	[81]	Denali Fault, 2002	7.9	Non-volcanic subduction zone	Love waves
7 sites throughout California	[32]	Denali Fault, 2002	7.9	Non-volcanic	Surface waves

** M is the magnitude of the triggering earthquake.

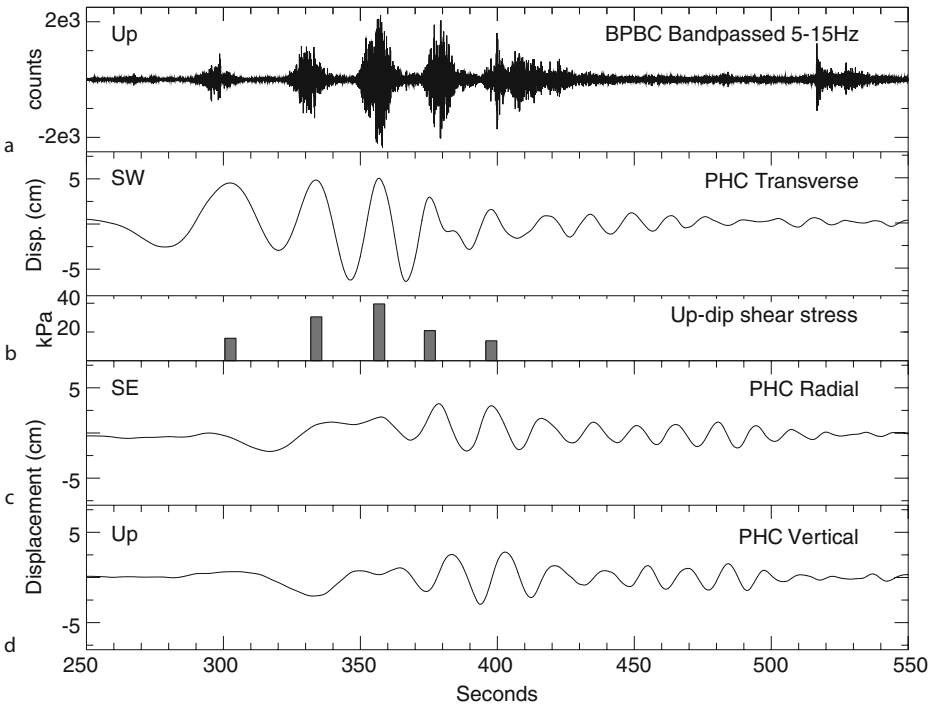

Earthquakes, Dynamic Triggering of, Figure 8
Time series showing tremor triggered by Love waves from the Denali Fault Earthquake in the Cascadia subduction zone: a Tremor at station BPBC, time adjust to correct for travel time from source to seismometer, b–d Displacement seismograms for transverse, radial, and vertical components at station PCH, the closest 3 component broadband station to the tremor, time adjust to correct for travel time from source to seismometer. Tremor occurs when the Love wave displacement is to the SW. Figure reprinted from [81], Nature

M 7.7 Chi-Chi earthquake. To test the uniqueness of these observations, they searched for triggered tremor at Aso following 20 other Mw $\geq$ 7 earthquakes occurring within 3000 km distance between 1995 and 2002. Five of these earthquakes triggered tremor following P wave arrivals at Aso. All occurred between 1998 and 1999, a time with usually high heat supply to the volcano's crater. As yet, this is the only documented episode of dynamically triggered volcanic tremor.

On the other side of the Pacific Ocean and a different tectonic environment, [81] identified episodes of deep non-volcanic tremor in the Cascadia subduction zone, Canada, which were triggered by the Love waves of the M 7.9 Denali Fault earthquake. In this case tremor amplitude modulates perfectly with strain amplitude from the incident Love waves (Fig. 8).

More recently, [32] identified triggered non-volcanic tremor in seven locations in California following the Denali Fault earthquake. In all cases tremor amplitude modulates with strain amplitude from incident surface waves. Five of these are strike-slip faulting regimes. These observations are the first reported cases of non-volcanic tremor beyond subduction zones (e. g. [80]) and the San Andreas fault [67].

Lack of Triggering Observations

Interestingly, some areas of high ambient seismicity show a notable lack of dynamically triggered seismicity. For example, the San Andreas fault near Parkfield, California showed no triggered response to the Landers earthquake [90]. Japan boasts high rates of shallow background seismicity, frequent large earthquakes from the subduction zone, high seismic network density, and a variety of crustal stress environments and volcanic and geothermal regions. However, through examining both earthquake catalogs and waveform data from individual seismic stations before and after nine large remote events, [33], show that dynamic triggering in Japan is not common, as it is in extensional regimes of the Western United States.

Similarly, Alaska abounds with crustal and subduction zone seismicity and active volcanic and geothermal systems, although network density is far lower than in Japan. Though the Katmai Volcanic cluster appears to be particularly susceptible to triggering [64,73] and dynamic triggering has been observed at Mt. Wrangell [103], dynamic triggering is rare compared to the western United States. [64] suggest that this results from unknown differences in the magmatic and hydrothermal systems of the volcanoes. [83] document a decrease in seismicity at Mt. Wrangell and Veniaminof volcanoes following the M 7.9 Denali Fault earthquake. To date these are the only documented examples of seismicity repression from a large distant earthquake.

Characteristics of Dynamic Triggering

Environmental Controls on Dynamic Triggering

Extensional and transtensional tectonic regimes with high levels of background seismicity are highly susceptible to dynamic triggering [38]. This may reflect the ease with which fluids can migrate upwards in these stress environments [32,38]. Because such fluids are often hot with high concentrations of dissolved solids, rapid precipitation may form high pressure compartments over rapid time scales, further enhancing a tendency toward failure. Faults in extensional stress regimes are also inherently weak compared to those in compressional environments [43,88]. [26,41] suggest that dynamic triggering is a ubiquitous process in the crust which is detected more commonly in certain areas due to high instrumentation and scrutiny levels. Only one study to date has carefully addressed this question. By comparing seismicity rates on the San Andreas fault in California and the Western United States Basin and Range Province, [90] show that the San Andreas fault is less likely to experience dynamic triggering than similarly instrumented areas with similar levels of background seismicity in the Western United States Basin and Range province. More studies like [90] are necessary to resolve whether triggering is truly ubiquitous or favored in specific tectonic environments.

Triggering Thresholds and Recharge Times

In most reports of remote dynamic triggering, seismicity is triggered by earthquakes of M 6.5 or greater (Table 1). Dynamic triggering responses are strongest in areas that experience strong directivity [26,30,39]. These first order observations suggest that strength of triggered response is a function of ground shaking amplitude. Although amplitude-based triggering thresholds have been suggested for some areas [28,29,31,64], a consistent triggering threshold that applies throughout the crust has not been established [38]. Large earthquakes regularly occur without dynamically triggering seismicity beyond their aftershock zones.

Lack of triggering reports below M 6.5 may reflect subtle triggered responses. [41] uses the Beta statistic to give evidence of small seismicity increases at distances of 70–110 km in the month following 14 moderate (M 5.5–7) earthquakes in California. Because this distance corresponds with where a large SMS phase should arrive, [41] suggests that the SMS phase is responsible for the triggered response in these cases.

If a simple amplitude-of-shaking threshold is required to dynamically trigger earthquakes, we would expect that even moderate earthquakes trigger seismicity near their epicenters. [21,27,52,71] give strong evidence that dynamic triggering occurs in the near field. Because it is difficult to distinguish the influence of static and dynamic stress changes in the near field, many studies of dynamic triggering have limited their investigation to the realm beyond the aftershock zone.

Because many aftershocks in the near field are likely dynamically triggered, [31] include aftershocks in a search for an amplitude-based triggering threshold. They find that peak dynamic stress distributions correlate well with aftershock and remotely triggered seismicity distributions, except in the Long Valley caldera, CA. The result of [31] is consistent with failure thresholds found in laboratory studies [49] and independent of frequency of shaking. [64] also find evidence for a ground shaking amplitude-based triggering threshold at the Katmai Volcanic Cluster, Alaska by comparing magnitude and distance of mainshock with triggered response (Fig. 9). Their magnitude-distance relationship is similar to that proposed by [28] for the Geysers, CA. However, the triggering

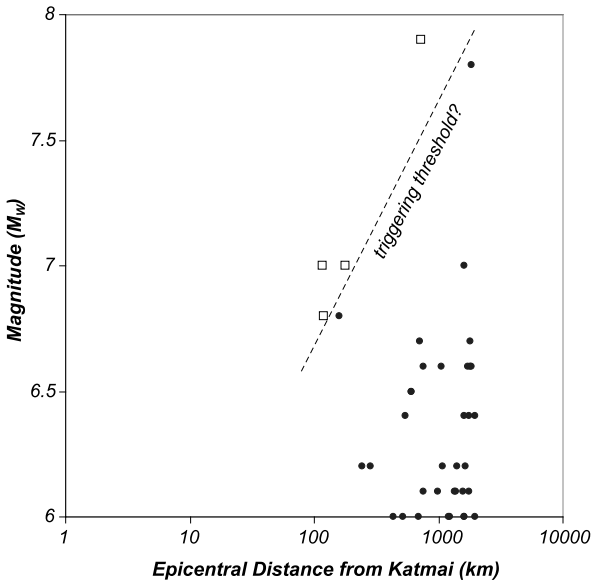

Earthquakes, Dynamic Triggering of, Figure 9
Plot of magnitude vs. distance from Mageik volcano in the Katmai Volcanic Cluster for all Mw > 6 earthquakes between 1996 and 2003 located within 2000 km of Katmai. *Hollow squares* triggered seismicity in the KVC. *Solid circles* did not. *Dashed line* represents possible triggering threshold. Figure reprinted from [64], BSSA

threshold at Katmai appears to be higher than that suggested for the Geysers.

In other cases, a simple amplitude-of-shaking threshold is not consistent with data, and large amplitude ground shaking is neither a necessary nor sufficient condition to cause dynamically triggered earthquakes. [31] show that the Long Valley caldera appears to be more susceptible to triggering than other areas they studied. Because their study was based on catalog seismicity, it did not include triggered earthquakes that were too small to appear in earthquake catalogs, such as those at the Coso geothermal field in response to the Denali Fault earthquake. These events were triggered by dynamic stresses of < 0.01 MPa [74] and, like Long Valley, would not fit the thresholds proposed by [29] and [31].

By comparing spectra of all earthquakes with high amplitude ground shaking in the Long Valley caldera [7] find that in this area, high-amplitude low-frequency shaking is more likely to trigger seismicity than high-amplitude high-frequency shaking. [1] come to the same conclusion after examining strong ground shaking spectra of earthquakes which did and did not trigger seismicity in the Western Great Basin. Longer wave lengths associated with low frequency ground shaking favor triggering by larger earthquakes in at least some locales. Whether remote dynamic triggering in both the near and far field results from the same physical process or processes remains an open question.

One parameter that may complicate the search for a triggering threshold in amplitude and/or frequency is recharge time. Because the occurrence of earthquakes releases stored strain energy, it may take time for an area to re-accumulate strain energy sufficiently to be primed for failure again following local earthquake activity or previous episodes of remotely triggered seismicity [38]. However some areas, such as the Geysers geothermal field require little to no time to recharge, as triggered seismicity episodes have been separated by time intervals of months or less [28]. Recharge times are dependent on many parameters including earthquake history, regional tectonic strain rates, and mass and heat advection rates in areas of hydrothermal and volcanic activity.

Time Scales of Dynamic Triggering and Responsible Phases

Remote dynamic triggering of earthquakes occurs over a variety of time scales following the onset of dynamic stressing. At Aso Volcano, Japan [61] triggered tremor begins with the P-wave arrival from distant large earthquake. In the case of discrete earthquakes however, it is most common for triggering to begin during surface wave arrivals (Table 1), leading many to suggest that the specific low-frequency large-amplitude ground motions associated with surface waves initiate the failure process [1,7,38,103]. Although the onset of dynamic triggering at remote locations is most commonly observed during Rayleigh wave arrivals [38], clear cases of remote triggering of tremor on the Love wave exist as well [81].

In some cases, dynamic triggering begins hours to days after an initial stress perturbation (e. g. [39,89,102]), hinting that the physical process responsible for initiating earthquake failure evolve with time. For example the largest triggered event following the M 7.3 Landers earthquake, a M 5.6 at the Little Skull Mountain, Nevada, occurred 33 hours after the mainshock[39]. In the case of Long Valley caldera's south moat and Mt. Rainier after the Denali Fault earthquake, delayed earthquake swarms began 24 hours and 2 hours respectively after the passage of the dynamic waves from the mainshock (Fig. 4) [74]. Both of these areas also had much smaller triggered swarms during the mainshock's wavetrain.

Determining the duration and decay time of triggered swarms is more difficult than detecting their onsets, particularly in areas of high ambient seismicity. Many triggered earthquakes may be triggered secondarily as aftershocks

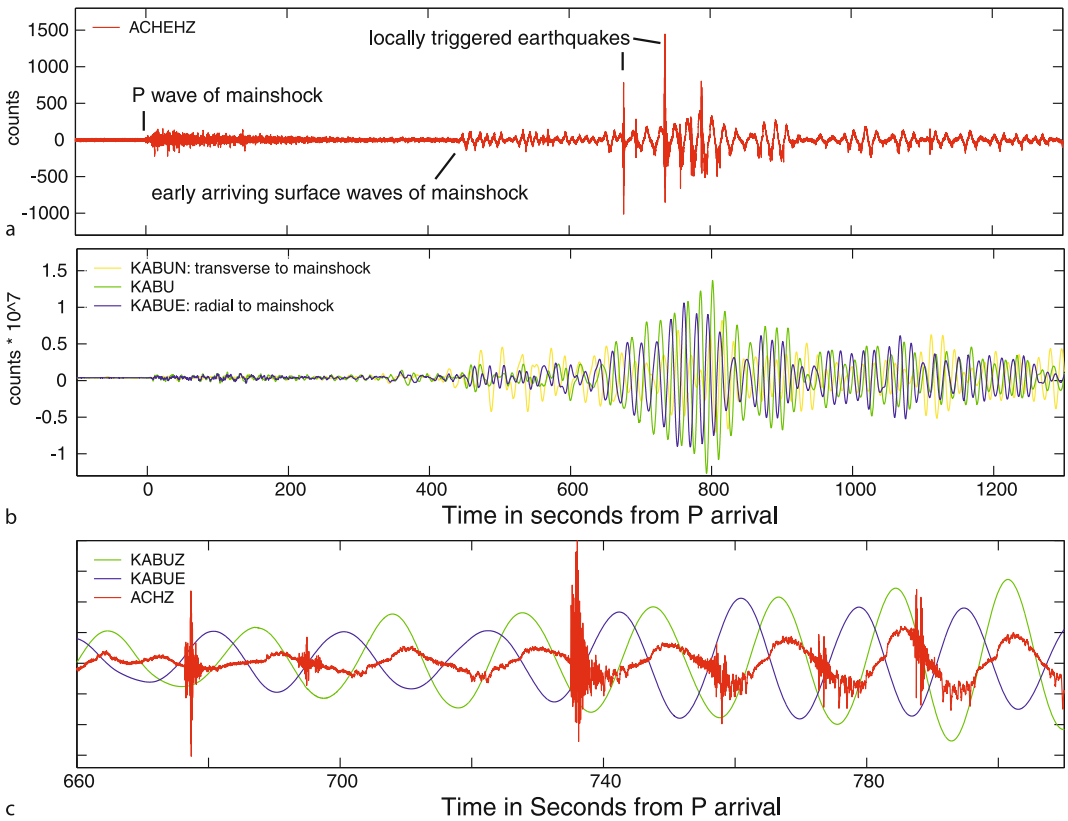

Earthquakes, Dynamic Triggering of, Figure 10

Phase modulated dynamically triggered earthquakes in the Katmai Volcanic Cluster following the 2007 M 8.2 Kurile earthquake: a short period record from station ACH showing both wavetrains for the Kurile earthquake and the larger amplitude, locally triggered earthquakes, b broadband record from station KABU showing wave motion of the Kurile earthquake, c time series from ACH and KABU zoomed in to show how local earthquakes seen clearly in *red* are occurring on a specific phase of the wavetrain from the Kurile earthquake

to earlier triggered earthquakes [9]. The Yellowstone response to the Denali Fault earthquake and the Long Valley caldera response to Landers are fit well with an Omori-type law decay [45]. In some cases however, triggered swarms end abruptly after the dynamic stress perturbation stops (e. g. [103]). Although our understanding of decay rates of triggered swarms is incomplete emphasizing that the subject deserves further investigation, decay rates give strong constraints on physical processes responsible for triggering.

Phase Modulated Triggering

Recent findings show that earthquakes can be triggered during specific phases of the wavetrain. At Mt. Wrangell, Alaska, triggered earthquakes occurred preferentially during phases of the largest positive vertical ground displacement from the 2004 M 9.0 Sumatra earthquake [103]. Similarly at Katmai Volcanic Cluster, Alaska, triggered earthquakes occurred only during specific phases of Rayleigh waves from the 2007 M 8.2 Kurile earthquake (Fig. 10). Such observations will allow us to resolve the precise dynamic stress field perturbations at the moment of earthquake nucleation on specific failure planes (e. g. [38]).

Physical Models of Dynamic Triggering

The wide variations in the characteristics of dynamic triggering and the limited data for individual response instances admit a spectrum of competing models for the physical processes linking dynamic stresses from a large, distant earthquake to the locally triggered response. Broadly considered, published models fall into three partially overlapping categories: 1) those involving some form of stress-driven brittle failure across local fractures, 2) those involving the activation of hydrous or mag-

matic fluids, and 3) those involving some form of localized aseismic relaxation (deformation). The brittle failure models are generally consistent with the onset of locally triggered seismicity during dynamic stressing (rapid-onset triggering), including the possibility that seismicity may persist as aftershocks for some time after the dynamic stressing has stopped [21]. Under the latter two categories, the onset of local seismicity represents a second-order phenomenon driven by a first order response to dynamic stressing in the form of fluid activation or transient deformation. In principle, models under these two categories admit a significant delay in the onset of the triggered seismicity with respect to the dynamic stresses generated by a distant earthquake. Because the dynamic stress amplitudes that trigger a response at remote distances are typically an order of magnitude or more below background tectonic stress levels, all models carry the implicit assumption that a crustal volume susceptible to dynamic stress triggering must be in a near-critical stress state prior to a triggered response.

Brittle Failure

Brittle failure models are based on the premise that the dynamic stresses propagating with the seismic waves from a distant earthquake are sufficient to nudge the local stress acting on a pre-existing dislocation beyond the threshold for the particular failure mode. This threshold may be the Griffith criteria for the tensile strength of a partially healed crack or the Coulomb criteria for frictional strength of a fault [86,87]. Crustal fluids play an important passive role in all brittle failure models by counteracting the rock matrix stress acting on a fracture through pore pressure, p, according to

$$\boldsymbol{\sigma}' = \boldsymbol{\sigma} - \boldsymbol{I}p$$

where $\boldsymbol{\sigma}'$ and $\boldsymbol{\sigma}$ are the effective and rock matrix stress tensors, respectively, and $\mathbf{I}$ is the identity tensor. Thus, pore pressure reduces the effective normal stress, σ_n', acting on a fracture by opposing the rock matrix normal stress as $\sigma_n' = \sigma_n - p$. Alternatively, for pressure-sensitive friction models the role of pore pressure can be expressed in terms of an effective coefficient of friction as $\mu' = \mu(1 - \lambda_p)$, where $\lambda_p = p/\sigma_n$. Elevated pore pressures lower the effectively strength by moving the background stress state closer to extensional or shear failure thresholds thereby increasing vulnerability for failure by imposition of small dynamic stress perturbations.

In the simplest frictional failure model, a triggered earthquake occurs when the stress acting on a fault exceeds the Coulomb threshold for static friction, or CFF(t) = 0, and friction abruptly drops from static to dynamic values with $\mu_s > \mu_d$, respectively. Here, CFF(t) is the Coulomb Failure Function defined as

$$\begin{aligned} \mathrm{CFF(t)} &= |\tau(t)| - \mu_s \sigma_n'(t) - \mathrm{C}\,, \quad \text{or its equivalent} \\ &= |\tau(t)| - \mu_s' \sigma_n(t) - \mathrm{C} \end{aligned}$$

where σ_n, σ_n', μ_s, μ_s' are defined in the preceding paragraph, τ is the shear, and C is the cohesive strength ([34], and references therein). This simple case implies rapid-onset triggering with the triggered seismicity beginning promptly when CFF(t) first becomes positive for a fault optimally oriented for failure in the background stress field. The combination of dynamic stress components $\Delta\tau$ and $\Delta\sigma_n$ for which CFF > 0 will depend on the wave type (e. g. Love or Rayleigh wave) and its incidence angle on the optimally oriented fault [35]. Although details vary, Love waves will generally have a greater triggering potential than Rayleigh waves when incident on vertical, strike-slip faults while the opposite is the case for incidence on inclined, dip-slip faults.

The Coulomb failure criterion applies to more elaborate non-linear friction models as well (see [18,25,31,70,100]). Because the behavior of non-linear models depend on factors such as slip history and slip rate, however, the failure threshold for static friction may vary with time, and the triggered earthquake may be delayed with respect to the time the failure criterion was first exceeded (e. g. [69]). Susceptibility to dynamic triggering may result when a dynamic stress is imposed on quasi-static loading under a conditionally stable regime (e. g. [84]). Based on their analysis of the dynamic triggering observed at Long Valley caldera, [7] conclude that this mechanism requires near-lithostatic pore pressures to be effective.

Models based on the non-linear response of granular media to dynamic stresses may apply to dynamic triggering of mature faults with a well-developed core of fault gouge. [49] document an abrupt decrease in the modulus of fault gouge under low effective normal stress (σ_n' 0.1 MPa) when excited by dynamic strains $> 10^{-6}$ in the laboratory. Thus, this model also requires near-lithostatic pore pressures to be effective.

Sub-critical crack growth, or stress corrosion, is another non-linear form of brittle failure that has a potential role in dynamic triggering. Under this model, a sudden increase in differential stress or an oscillatory stress applied to a pre-existing crack can lead to crack growth due to weakening of the crack tip by chemical corrosion. This can shorten the time to earthquake rupture. This process will be enhanced in an environment with fluids at elevated temperatures. It turns out that the equations governing sub-critical crack growth have the same mathemat-

ical form as rate-state friction equations above [51]. Thus near-lithostatic pore pressure appears to be a requirement for each of these non-linear brittle-failure models, at least as they apply to dynamic triggering at remote distances.

Fluid Activation Models

In addition to their passive role in reducing the effective strength of a rock volume through ambient pore pressure, fluids may play an active role in the dynamic triggering process. Fluid activation models generally appeal to either 1) pore-pressure re-distribution associated with changes in permeability and fluid transport, or 2) state changes induced in multi-phase fluids.

Dynamic stressing may be capable of physically disrupting permeability barriers separating volumes of differing pore pressure. For example, dynamic stress may shake accumulated detritus from clogged fractures or opening partially healed fractures by extensional failure. In either case, fluid diffusion down the pressure gradient will result in a re-distribution of pore pressure with the potential for triggering seismicity in previously under-pressured volumes in a near-critically stressed state. The evolution of triggered seismicity in this case will be governed by the diffusion length for a given permeability and the proximity of the pre-existing stress state to brittle failure. [8] proposed the clogged fracture model as an explanation for the hydrologic response of water wells in southern Oregon to surface waves from $M > 7$ earthquakes at distances of 300 km and 3850 km.

Geothermal areas may be particularly susceptible to dynamic triggering through pore pressure re-distribution. In these areas fractures are rapidly sealed by precipitation from circulating, solute-rich geothermal fluids and plastic deformation of quartz-rich rocks under elevated temperatures tend to isolate pockets of elevated pore pressure. Most active geothermal systems are located in areas of extensional tectonism. In these areas normal stresses induced by Rayleigh waves on vertical planes may open vertical fractures, allowing high-pore-pressure fluids access to shallower crustal volumes with lower pore pressure [35]. The hydraulic surge model described by [22] for volcanic and geothermal systems is a version of this process in which the brittle-plastic transition at the base of the seismogenic crust serves as a low-permeability barrier separating near-lithostatic pore pressures in the plastic regime from a hydrostatic regime in the overlying seismogenic crust. Rupturing the permeability seal by dynamic stresses would release near-lithostatic pore pressures into the brittle, seismogenic crust thereby inducing a surge in triggered seismicity.

Models for bubble excitation by dynamic stresses in a two-phase fluid (multi-phase in a partially crystallized magma) offers interesting possibilities for remotely triggered responses in geothermal and volcanic systems. This is a particularly intriguing concept for remote triggering in volcanic systems because of the importance of bubbles in eruption dynamics [57] and the source mechanisms of long-period volcanic earthquakes [13], ► Volcanoes, Non-linear Processes in. Advective overpressure and rectified diffusion were the first bubble models proposed as explanations for remotely triggered seismicity [10,55,94]; although subsequent work has shown that both hold less promise as viable explanations than initially thought [56].

Under the advective overpressure model, the pressure in a gas-saturated, incompressible fluid confined in a rigid container increases as $\rho g \Delta h$ as a pre-existing bubble adhering to the wall of the container is shaken loose by passing seismic waves. The bubble ascends buoyantly a distance Δh through a fluid of density ρ where g is the acceleration of gravity [55]. The resulting pressure increase in the container (magma body) deforms the surrounding rock inducing small earthquakes. This model is criticized on the basis that assumptions of a ridged container and an incompressible fluid seriously violate realistic conditions in the earth [75].

Under rectified diffusion, pressure oscillations imposed on a gas-saturated fluid with pre-existing bubbles pump gas into the bubbles over multiple cycles. Gas exolves from the fluid into the bubble during the dilatational phase, when bubble surface area is maximal, and out of the bubble back into solution during compressional phase, when the bubble surface area is minimal [94]. The implied pressure gain integrated over multiple cycles is then transmitted to the surrounding rock inducing small earthquakes. [48] point out, however, that the effectiveness of this model is limited by reasonable gas diffusion rates in hydrous fluids or magma with respect to the frequencies of seismic waves driving the pressure oscillations.

More promising bubble models appeal to the strong sensitivity of bubble nucleation rate to the supersaturation pressure [56] and the results of numerical models by ► Volcanoes, Non-linear Processes in and [14,85], indicating that a small pressure drop imposed on a densely packed matrix of tiny bubbles can lead to rapid, diffusion-driven bubble growth. The implications of these models, however, have yet to be more fully explored in the context of dynamic triggering.

Two more speculative models involve magma instabilities triggered by dynamic stresses. In one, a loosely held crystal mush accumulated on the walls of a crystal-

lizing magma body may be dislodged by dynamic shaking. The sinking crystal mush would induce a convective plume as it displaced hotter, less dense magma. In the case of volatile-rich magma, buoyant convection would be enhanced by bubble nucleation and growth as confining pressure drops with decreasing depth [37]. Under suitable conditions, the resulting pressure increase within the magma body could evolve over days [56]. If the magma chamber was already in a near critical state, the culmination could be magma intrusion into the overlying crust or the onset of an eruption. Whether this process culminates in a simple pressure increase, an intrusion, or an eruption, the sensible onset of locally triggered seismicity and deformation might be delayed by hours to perhaps days with respect to the passing seismic waves from the distant earthquake. A second, even more speculative model appeals to dynamic stresses disrupting the solid matrix of a partially crystallized magma body thereby releasing any differential tectonic stress sustained by the solid matrix [36,37]. As the magma body relaxes with a time constant governed by the effective viscosity of the disrupted crystal mush, stress would be transferred to the surround crust inducing deformation and local seismicity. In essence, this model corresponds to the relaxation of an Eshelby inclusion in an elastic medium [20].

Aseismic Deformation

The relaxing magma body of the previous paragraph is one example of aseismic deformation with the potential of triggering local deformation and the onset of secondary seismicity. A less speculative example involves aseismic creep on faults triggered by dynamic stressing. Deformation associated with fault creep transfers stress to the adjacent crust, which in turn triggers local seismicity, as in the example involving seismicity triggered on the Reykjanes Peninsula following the $M = 6.5$ earthquake in the South Iceland Seismic Zone in 2000 [2]. [4] document aseismic fault slip (creep) on faults in the Salton Trough of southern California triggered by the three $M > 6$ earthquakes in the Landers, California sequence of 1992 (the $M = 6.1$ Joshua Tree, $M = 7.3$ Landers, and the $M = 6.2$ Big Bear earthquakes). In this case, all instances of triggered slip were on faults with within 150 km of the $M > 6$ earthquakes. In these examples and observations from triggering in Long Valley caldera and Cierra Prieto geothermal field in Baja California [38,50] the geodetic moment for triggered aseismic deformation exceeds the cumulative seismic moment for the triggered earthquakes by a factor of two or more. This emphasizes the importance of high-resolution deformation monitoring in areas susceptible to dynamic triggering for resolving the role of aseismic deformation in the dynamic triggering process.

Future Directions

In the last 25 years, in the wake of the Landers earthquake, the study of dynamically triggered seismicity has given us new insight into earthquake initiation and the failure regime in the Earth's crust. Some argue that the state of stress in the crust is highly spatially variable [77]. Given this, the likelihood of triggering seismicity would also be spatially variable. [97] and [105], however, argue that the Earth's crust is critically stressed and on the verge of failure nearly everywhere. If this were the case, one might expect triggering due to small dynamic stress perturbations to be a ubiquitous phenomenon. In either case, the study of remotely triggered seismicity provides clues to spatial distribution of critically stressed crustal volumes.

Unfortunately, although we can measure stress field perturbations from dynamic waves from earthquakes, we rarely have a detailed understanding of the background stress field these perturbations are modulating. In addition, dynamically triggered earthquakes are often too small or occur in too sparsely instrumented areas to resolve reliable focal mechanisms. Because of these limitations, our understanding of how dynamic stresses from remote earthquake wavetrains induce a given crustal volume to respond with triggered seismicity remains incomplete. Advances will require more cases of dynamically triggered seismicity captured by both spatially dense seismic networks and continuous, high-resolution deformation monitoring networks.

Recent observations of phase modulated dynamic triggering offer powerful datasets of the precise time history of dynamic stress triggering. Because similarities exist between phase modulated dynamic triggering of seismicity in the shallow crust of a volcano's edifice [103] and deep in a subduction zone [81], the emerging study of non-volcanic tremor may provide new leverage on understanding how dynamic stresses influence seismic slip.

As new observations of dynamically triggered seismicity are reported, one conclusion is becoming increasingly evident: multiple causative processes exist. The wide variety in time scales over which triggering occurs and the spatial and temporal characteristics of triggered seismicity sequences and associated deformation responses cannot be fit with any one model yet proposed. Rather, different models are consistent with different episodes of triggering. For example, fluid activation and stress corrosion models

are most applicable in volcanically and geothermally active environments. In some cases, such as the complex triggered response of Yellowstone, Mt. Rainier, and the Long Valley caldera areas to the Denali Fault earthquake, multiple processes may be occurring simultaneously in the same locale, yet on different time scales. Of the physical models described above, seismicity triggered instantaneously or within seconds of the dynamic stress perturbation is consistent with models based on simple brittle failure, brittle failure with nonlinear friction effects, stress corrosion, unclogging of fractures, or rectified diffusion, whereas triggered seismicity delayed by hours to days is more consistent with models involving aseismic deformation, advective overpressure, sinking crystal plumes, or a relaxing magma body.

In the few cases where hydrologic and high-sample rate strain data are available, dynamically triggered seismicity is accompanied by changes in water levels in wells [79] and significant deformation signals [2,25,36,50]. A complete understanding of dynamic triggering will require research approaches that integrate seismic, deformation, and hydrologic datastreams. To this end, we challenge Earth scientists to broaden their thinking and tap these observations to better understand the initiation of earthquake failure.

Bibliography

Primary Literature

1. Anderson JG, Brune JN, Louie JN, Zeng Y, Savage M, Yu G, Chen Q, de Polo D (1994) Seismicity in the western Great Basin apparently triggered by the Landers, California, earthquake, 28 June 1992. Bul Seismo Soc Am 84:863–891
2. Arnadottir T, Geirsson H, Einarsson P (2004) Coseismic stress changes and crustal deformation on the Reykjanes Peninsula due to triggered earthquakes on 17 June 2000. J Geophys Res. doi:109:10.1029/2004JB003130
3. Blacket RE, Wakefield S (2002) Geothermal resources of utah, Utah geological survey open file report 397, ISBN 1-55791-677-2
4. Bodin P, Bilham R, Behr J, Gomberg J, Hudnut K (1994) Slip triggered on southern California faults by the 1992 Joshua Tree, Landers, and Big Bear earthquakes. Bul Seismo Soc Am 84:806–816
5. Bodin P, Gomberg J (1995) Triggered seismicity and deformation between the Landers, California and Little Skull Mountain Nevada earthquakes. Bul Seismo Soc Am 84:835–843
6. Brodsky EE, Karakostas V, Kanamori H (2000) A new observation of dynamically triggered regional seismicity: earthquakes in Greece following the August, 1999, Izmit, Turkey earthquake. Geophys Res Let 27:2741–2744
7. Brodsky EE, Prejean SG (2005) New constraints on mechanisms of remotely triggered seismicity at Long Valley Caldera. J Geophys Res. doi:110:10.1029/2004JB003211
8. Brodsky EE, Roeloffs E, Woodcock D, Gall I, Manga M (2003) A mechanism for sustained groundwater pressure changes induced by distant earthquakes. J Geophys Res. doi:108:10.1029/2002JB002321
9. Brodsky EE (2006) http://www.pmc.ucsc.edu/~brodsky/reprints/Sus5_merged.pdf. Long-range triggered earthquakes that continue after the wavetrain passes. Geophys Res Let 33:L15313
10. Brodsky EE, Sturtevant B, Kanamori H (1998) Earthquakes, volcanoes, and rectified diffusion. J Geophys Res 103:23827–23838
11. Brune J (1970) Tectonic stress and the spectra of seismic shear waves from earthquakes. J Geophys Res 75:4997–5009
12. Camelbeeck T, van Eck T, Pelzing R, Ahorner L, Loohuis J, Haak HW, Hoang–Trong P, Hollnack D (1994) The 1992 Roermond earthquake, the Netherlands, and its aftershocks. Geologie en Mijnbouw 73:181–197
13. Chouet B (1992) A seismic model for the source of long-period events and harmonic tremor. In: Gasparini P, Scarpa R, Aki K (eds) Volcanic seismology. IAVCEI Proceedings in Volcanology. Springer, Berlin, pp 133–156
14. Chouet B, Dawson P, Nakano M (2006) Dynamics of diffusive bubble growth and pressure recovery in a bubbly rhyolitic melt embedded in an elastic solid. J Geophys Res. doi:111:10.1029/2005JB04174
15. Cocco M, Rice JR (2002) Pore pressure and poroelasticity effects in Coulomb stress analysis of earthquake interactions. J Geophys Res doi:107:10.1029/2002JB002319
16. Cochran ES, Vidale JE, Tanaka S (2004) Earth tides can trigger shallow thrust fault earthquakes. Science 306:1164–1166
17. Cooper HH, Bredehoeft JD, Papdopulos S, Bennett RR (1965) The response of well-aquifer systems to seismic waves. J Geophys Res 70:3915–3926
18. Dieterich JH (1979) Modeling of rock friction 1, Experimental results and constitutive equations. J Geophys Res 84:2161–2168
19. Elkhoury JE, Brodsky EE, Agnew DC (2006) Seismic waves increase permeability. Nature 441:1135–1138
20. Eshelby JD (1957) The determination of the elastic field of an ellipsoidal inclusion, and related problems. Proceedings of the Royal Society of London A241:376–396
21. Felzer KR, Brodsky EE (2006) http://www.nature.com/nature/journal/v441/n7094/full/nature04799.html. Decay of aftershock density with distance indicates triggering by dynamic stress. Nature 441:735–738
22. Fournier RO (1999) Hydrothermal processes related to movement of fluid from plastic to brittle rock in the magmatic-epithermal environment. Economic Geology 94:1193–1211
23. Freed AM (2005) Earthquake triggering by static, dynamic, and postseismic stress transfer. Annual Rev Earth and Planet Sci 33:1255–1256
24. Glowacka E, Nava AF, Cossio DD, Wong V, Farfan F (2002) Fault slip, seismicity, and deformation in the Mexicali Valley, Baja California, Mexico, after the M 7.1 Hector Mine earthquake. Bul Seismo Soc Am 92:1290–1299
25. Gomberg J, Blanpied ML, Beeler NM (1997) Transient triggering of near and distant earthquakes. Bul Seismol Soc Am 87:294–309
26. Gomberg J, Bodin P, Larson K, Dragert H (2004) Earthquakes nucleated by transient deformations caused by the M = 7.9 Denali, Alaska, earthquake. Nature 427:621–624

27. Gomberg J, Bodin P, Reasenberg PA (2003) Observing earthquakes triggered in the near field by dynamic deformations. Bul Seismo Soc Am 93:118–138
28. Gomberg J, Davis S (1996) Stress/strain changes and triggered seismicity at The Geysers, California. J Geophys Res 101:733–749
29. Gomberg J, Johnson P (2005) Dynamic triggering of earthquakes. Nature 437:830
30. Gomberg J, Reasenberg PA, Bodin P, Harris R (2001) Earthquakes triggering by seismic waves following the Landera and Hector Mine earthquakes. Nature 411:462–465
31. Gomberg J, Reasenberg PA, Cocco M, Belardinelli ME (2005) A frictional population model of seismcity rate change. J Geophys Res. doi:110:10.1029/2004JB003404
32. Gomberg J, Rubinstein JL, Peng Z, Creager KC, Vidale JE, Bodin P (in press) Widespread triggering of non-volcanic tremor in California. Science 319:117
33. Harrington RM, Brodsky EE (2006) The absence of remotely triggered seismicity in Japan. Bul Seismo Soc Am 96:871–878
34. Harris RA (1998) Introduction to a special section: Stress triggers, stress shadows, and implications for seismic hazards. J Geophys Res 103:24347–24358
35. Hill DP (2008) Dynamic stresses, Coulomb failure, and remote triggering. Bul Seismo Soc Am 98:66-92
36. Hill DP, Johnston MJS, Langbein JO (1995) Response of Long Valley caldera to the Mw = 7.3 Landers, California, earthquake. J Geophys Res. doi:100:12985-13005
37. Hill DP, Pollitz F, Newhall C (2002) Earthquake-volcano interactions. Physics Today 55:41–47
38. Hill DP, Prejean SG (2007) Dynamic triggering. In: Kanamori H (ed) Geophysical treatise, earthquake seismology. Elsevier, Amsterdam
39. Hill DP, Reasenberg PA, Michael A, Arabaz WJ, Beroza G, Brumbaugh D, Brune JN, Castro R, Davis S, dePolo D, Ellsworth WL, Gomberg J, Harmsen S, House L, Jackson SM, Johnston MJS, Jones L, Keller R, Malone S, Munguia L, Nava S, Pechmann JC, Sanford A, Simpson RW, Smith RB, Stark M, Stickney M, Vidal A, Walter A, Wong A, Zollweg J (1993) Seismicity remotely triggered by the magnitude 7.3 Landers, California, earthquake. Science 260:1617–1622
40. Hough SE (2001) Triggered earthquakes and the 1811–1812 New Madrid, central United States, earthquake sequence. Bul Seismo Soc Am 91:1547–1581
41. Hough SE (2005) Remotely triggered earthquakes following moderate mainshocks (or why California is not falling into the ocean). Seismological Research Letters 76:58–66
42. Hough SE, Billham R, Ambraseys N, Field N (2005) Revisiting the 1897 Shillong and 1905 Kangra earthquakes in northern India: site response, Moho reflections and a triggered earthquake. Current Science 88:1632–1638
43. Hough SE, Kanamori H (2002) Source properties of earthquakes near the Salton Sea triggered by the 16 October 1999 Mw 7.1 Hector Mine, California, earthquake. Bul Seismo Soc Am 92:1281–1289
44. Hough SE, Seeber L, Armbruster JG (2003) Intraplate triggered earthquakes: observations and interpretation. Bul Seismo Soc Am 93:2212–2221
45. Husen S, Taylor R, Smith RB, Healser H (2004) Changes in geyser eruption behavior and remotely triggered seismicity in Yellowstone National Park produced by the 2002 M 7.9 Denali fault earthquake, Alaska. Geology 32:537–540
46. Husen S, Wiemer S, Smith RB (2004) Remotely triggered seismicity in the Yellowstone National Park region by the 2002 Mw 7.9 Denali Fault earthquake, Alaska. Bul Seismo Soc Am 94:S317–S331
47. Husker AL, Brodsky EE (2004) Seismcity in Idaho and Montana triggered by the Denali Fault earthquake: a window into the geologic context for seismic triggering. Bul Seismo Soc Am 94:S310–S316
48. Ichihara M, Brodsky EE (2006) http://www.pmc.ucsc.edu/~brodsky/reprints/2005GL024753.pdf. A limit on the effect of rectified diffusion in volcanic systems. Geophys Res Let. doi:33:10.1029/2005GL024753
49. Johnson P, Jia X (2005) Nonlinear dynamic, granular media and dynamic earthquake triggering. Nature 437:871–874
50. Johnston MJS, Prejean SG, Hill DP (2004) Triggered deformation and seismic activity under Mammoth Mountain in Long Valley caldera by the 3 November 2002 Mw 7.9 Denali Fault earthquake. Bul Seismo Soc Am 94:S360–S369
51. Kanamori H, Brodsky EE (2004) http://www.pmc.ucsc.edu/~brodsky/reprints/rpp4_8_R03.pdf. The physics of earthquakes, Reports on Progress in Physics 67:1429–1496
52. Kilb D, Gomberg J, Bodin P (2000) Triggering of earthquake aftershocks by dynamic stresses. Nature 408:570–574
53. King GCP, Cocco M (2001) Fault interactions by elastic stress changes: new clues from earthquake sequences. Advances in Geophysics 44:1–38
54. King GCP, Stein RS, Lin J (1994) Static stress changes and the triggering of earthquakes. Bul Seismol Soc Am 84:935–953
55. Linde AT, Sacks IS, Johnston MJS, Hill DP, Billham RG (1994) Increased pressure from rising bubbles as a mechanism for remotely triggered seismicity. Nature 371:408–410
56. Manga M, Brodsky EE (2006) Seismic triggering of eruptions in the far field: volcanoes and geysers. Annual Rev Earth and Planet Sci 34:263–291
57. Mangan M, Sisson T (2000) Delayed, disequilibrium degassing in rhyolite magma: decompression experiments and implications for explosive volcanism. Earth and Planetary Science Letters 183:441–455
58. Matthews MV, Reasenberg PA (1988) Statistical methods for investigating quiescence and other temporal seismicity patterns. Pure Appl Geophys 126:357–372
59. Meltzner AJ, Wald DJ (2003) Aftershocks and triggered events of the great 1906 California earthquake. Bul Seismo Soc Am 93:2160–2186
60. Miyazawa M, Mori J (2005) Detection of triggered deep low-frequency events from the 2003 Takachi-oki earthquake. Geophys Res Let 32:L10307
61. Miyazawa M, Nakanishi I, Sudo Y, Ohkura T (2005) Dynamic response of frequent tremors at Aso volcano to teleseismic waves from the 1999 Chi-Chi, Taiwan earthquake. J Vol Geotherm Res 147:173–186
62. Mohamad RA, Darkal N, Seber D, Sandoval E, Gomez F, Barazangi M (2000) Remote earthquake triggering along the Dead Sea Fault in Syria following the 1995 Gulf of Aqaba earthquake (Ms = 7.3). Seismol Res Let 71:47–52
63. Moran SC (2003) Multiple seismogenic processes for high-frequency earthquakes at Katmai National Park, Alaska: evidence from stress tensor inversions of fault plane solutions. Bul Seismo Soc Am 93:94–108
64. Moran SC, Power JA, Stihler SD, Sanchez JJ, Caplin-Auerbach J (2004) Earthquake triggering at Alaskan volcanoes follow-

ing the 3 November 2002 Denali Fault earthquake. Bul Seismo Soc Am 94:S300–S309
65. Moran SC, Zimbelman DR, Malone SD (2003) A model for the magmatic-hydrothermal system at Mount Rainier, Washington, from seismic and geochemical observations. Bull Volcanol 61:425–436
66. Mueller K, Hough SE, Bilham R (2004) Analysing the 1811–1812 New Madrid earthquakes with recent instrumentally recorded aftershocks. Nature 429:284–288
67. Nadeau R, Dolenc D (2005) Nonvolcanic tremors deep beneath the San Andreas Fault. Science 300:1942–1943
68. Pankow KL, Arabasz WJ, Pechmann JC, Nava SJ (2004) Triggered seismicity in Utah from the 3 November 2002 Denali Fault earthquake. Bul Seismo Soc Am 94:S332–S347
69. Parsons T (2005) A hypothesis for delayed dynamic earthquake triggering. Geophys Res Let 32:L04302
70. Perfettini HJ, Schmittbuhl J, Cochard A (2003) Shear and normal load perturbations on a two-dimensional continuous fault: 2. dynamic triggering. J Geophys Res. doi:108:10.1029/2002JB001805
71. Pollitz FF, Johnston MJS (2006) Direct test of static-stress versus dynamic-stress triggering of aftershocks. Geophys Res Let 33:L15318
72. Pollitz FF, Sacks IS (2002) Stress triggering of the 1999 Hector Mine earthquake by transient deformation following the 1992 Landers earthquake. Bul Seismo Soc Am 92:1487–1496
73. Power JA, Moran SC, McNutt SR, Stihler SD, Sanchez JJ (2001) Seismic response of the Katmai volcanoes to the 6 December 1999 magnitude 7.0 Karluk Lake earthquake, Alaska. Bul Seismo Soc Am 91:57–63
74. Prejean SG, Hill DP, Brodsky EE, Hough SE, Johnston MJS, Malone SD, Oppenheimer DH, Pitt AM, Richards-Dinger KB (2004) Remotely triggered seismicity on the United States west coast following the Mw 7.9 Denali Fault earthquake. Bul Seismo Soc Am 94:S348–S359
75. Pyle DM, Pyle DL (1995) Bubble migration and the initiation of volcanic eruptions. J Vol Geotherm Res 67:227–232
76. Reasenberg P (1985) Second-order moment of central California seismcitiy, 1969–1982. J Geophys Res 90:5479–5495
77. Rivera L, Kanamori H (2002) Spatial heterogeneity of tectonic stress and friction in the crust. Geophys Res Let 29:10.1029/2001GLO13803
78. Roeloffs E (1998) Poroelastic techniques in the study of earthquake-related hydrologic phenomena. Adv Geophys 37:135–195
79. Roeloffs E, Sneed M, Galloway DL, Sorey ML, Farrar CD, Howle JF, Hughes J (2003) Water level changes induced by local and distant earthquakes at Long Valley Caldera, California. J Vol Geotherm Res 127:269–303
80. Rogers G, Dragert H (2003) Episodic tremor and slip on the Cascadia subduction zone: The chatter of silent slip. Science 296:1679–1681
81. Rubinstein JL, Vidale JE, Gomberg J, Bodin P, Creager KC, Malone S (2007) http://www.nature.com/nature/journal/v448/n7153/full/nature06017.html. Non-Volcanic Tremor Driven by Large Transient Shear Stresses. Nature 448:579–582
82. Rundle JB, Turcotte D, Shcherbakov R, Klein W, Sammis C (2003) Statistical physics approach to understanding the multiscale dynamics of earthquake fault systems. Rev Geophys 41:4/1019
83. Sanchez JJ, McNutt SR (2004) Intermediate-term declines in seismicity at Mt. Wrangell and Mt. Veniaminof volcanoes, Alaska, following the 3 November 2002 Mw 7.9 Denali Fault earthquake. Bul Seismo Soc Am 94:S370–S383
84. Scholz CH (1998) Earthquakes and friction laws. Nature 391:37–42
85. Shimomura Y, Nishimura T, Sato H (2006) Bubble growth processes in magma surrounded by an elastic medium. J Vol Geotherm Res 155:307–322
86. Sibson R (2000a) A brittle failure mode plot defining conditions for high-flux flow. Economic Geology 95:41–48
87. Sibson R (2000b) Fluid involvement in normal faulting. Geodynamics 29:469–499
88. Sibson RH (1982) Fault zone models, heat flow, and the depth distribution of earthquakes in the continental crust of the United States. Bul Seismo Soc Am 72:151–163
89. Singh SK, Anderson JG, Rodriguez M (1998) Triggered seismicity in the Valley of Mexico from major Mexican earthquakes. Geofiscia International 37:3–15
90. Spudich P, Steck LK, Hellweg M, Fletcher JB, Baker LM (1995) Transient stresses at Parkfield, California, produced by the M 7.4 Landers earthquake of June 28, 1992: observations from the UPSAR dense seismograph array. J Geophys Res 100:675–690
91. Stark MA, Davis SD (1996) Remotely triggered microearthquakes at The Geysers geothermal field, California. Geophys Res Let 23:945–948
92. Steacy S, Gomberg J, Cocco M (2005) Introduction to special section: Stress transfer, earthquake triggering, and time-dependent seismic hazard. J Geophys Res. doi:110:10.1029/2005JB003692
93. Stein RS (1999) The role of stress transfer in earthquake occurrence. Nature 402:605–609
94. Sturtevant B, Kanamori H, Brodsky E (1996) Seismic triggering by rectified diffusion in geothermal systems. J Geophys Res 101:25269–25282
95. Tanaka S, Ohtake M, Sato H (2003) Tidal triggering of earthquakes in Japan related to the regional tectonic stress. Earth Planets and Space 56:511–515
96. Tibi R, Wiens DA, Inoue H (2003) Remote triggering of deep earthquakes in the 2002 Tonga sequence. Nature 424:921–925
97. Townend J, Zoback MD (2000) How faulting keeps the crust strong. Geology 28:399–402
98. Ukawa M, Fujita E, Kumagai T (2002) Remote triggering of microearthquakes at the Iwo-Jima volcano. J Geography 111:277–286
99. Unruh JR, Hauksson E, Monastero FC, Twiss RJ, Lewis JC (2002) Seismotectonics of the Coso Range – Indian Wells Valley region, California: Transtensional deformation along the southeastern margin of the Sierran microplate. Geol Soc Am Mem 195:277–294
100. Voisin C (2002) Dynamic triggering of earthquakes: the nonlinear slip-dependent friction case. J Geophys Res 107(B12):10.1–10.11
101. Weaver CS, Hill DP (1978/79) Earthquake swarms and local crustal spreading along major strike-slip faults in California. Pageoph 117:51–64
102. Wen KL, Beresnev IA, Cheng S (1996) Moderate-magnitude seismicity remotely triggered in the Taiwan Region by large

earthquakes around the Philippine Sea Plate. Bul Seismo Soc Am 86:843–847
103. West M, Sanchez JJ, McNutt SR (2005) Periodically triggered seismicity at Mount Wrangell, Alaska, after the Sumatra earthquake. Science 308:1144–1146
104. Yabe Y, Song S, Wang C (2005) Stress state around Chelungpu Fault, Taiwan, Estimated from boring core samples. EOS Trans 86:T51A–1316
105. Zoback MD, Zoback ML (2002) State of stress in the Earth's lithosphere In: Lee WH, Kanamori H, Jennings PC, Kisslinger C (eds) International handbook of earthquake and engineering seismology, Part A. Academic Press, Amsterdam, pp 559–568

Books and Reviews

Freed AM (2005) Earthquake triggering by static, dynamic, and postseismic stress transfer. Annual Rev Earth and Planet Sci 33:1255–1256
Harris RA (1998) Introduction to a special section: Stress triggers, stress shadows, and implications for seismic hazards. J Geophys Res 103:24347–24358
Hill DP, Pollitz F, Newhall C (2002) Earthquake-volcano interactions. Physics Today 55:41–47
Hill DP, Prejean SG (2007) Dynamic triggering. In: Kanamori H (ed) Geophysical treatise, earthquake seismology. Elsevier, Amsterdam
Manga M, Brodsky EE (2005) Seismic triggering of eruptions in the far field: volcanoes and geysers. Annual Rev Earth and Planet Sci 34:263–291
Steacy S, Gomberg J, Cocco M (2005) Introduction to special section: Stress transfer, earthquake triggering, and time-dependent seismic hazard. J Geophys Res. doi:110:10.1029/2005JB003692

Earthquakes, Electromagnetic Signals of

Seiya Uyeda[1], Masashi Kamogawa[2], Toshiyasu Nagao[1]
[1] Earthquake Prediction Research Center, Tokai University, Shizuoka, Japan
[2] Department of Physics, Tokyo Gakugei University, Koganei-shi, Japan

Article Outline

Glossary
Definition of the Subject
Introduction
Telluric Current Anomalies and Natural Time
Ultra Low Frequency (ULF) Anomalies
Higher Frequency Electromagnetic Emission and Earthquake Light
Lithosphere-Atmosphere-Ionosphere (LAI) Coupling
Mechanism of Pre-Seismic EM Phenomena
Future Directions
Bibliography

Glossary

Earthquake prediction Place of epicenter, time of occurrence, and magnitude are the three main items of earthquake prediction. Occurrence time is the most difficult to predict. Depending on the concerned time scales, prediction is usually classified as long term ($\sim$ tens of years), intermediate term ($\sim$ a few years), and short term (months to days) predictions. Electromagnetic signals of earthquakes are mainly concerned with the short term prediction.

Piezo-electric effect Piezo-electricity is the electric polarization produced in certain crystals and ceramics by the application of mechanical stress. Among rock-forming minerals, quartz is most strongly piezo-electric, but its effect is much reduced because quartz crystals are usually randomly oriented. Moreover, stress-induced piezo-electric polarization in rocks is kept canceled by compensating charges. At rapid stress drop, bulk polarization appears as the compensating charge cannot disappear instantly and decays with a time constant $\tau = \varepsilon/\sigma$, where ε is dielectric constant and σ electric conductivity.

Electro-kinetic effect Electro-kinetic effect, also called streaming potential, is caused by the presence of the solid-liquid interface. The double layer consists of ions (anions in most cases of rock-water system) that are firmly anchored to the solid phase and ions of the opposite sign (cations) in the liquid phase attracted to them near the boundary. The liquid phase is in surplus of cations so that when the liquid flows due to a pressure gradient, an electric potential gradient is formed. It is expressed as grad $V = -(\varepsilon\zeta/\eta\sigma)$ grad P, where ε, σ and η are the dielectric constant, electric conductivity, and viscosity of the fluid, whereas ζ is a constant called zeta potential. Thus, the streaming potential is small for high conductive and viscous liquid.

Telluric current Electric current flowing in the surface layer of the earth's crust is called telluric current. Mainly it consists of the current induced by extra-terrestrial geomagnetic field variations (called magneto-telluric or MT current) and the current as a part of the global circuit between ionosphere and ground. MT current carries information on the electrical structure of the earth's interior: higher (lower) frequency for shallower (deeper) structure. Telluric current can also be of man-made origin leaking from such electric sources as factories and trains. Telluric current is mea-

sured by dipoles of electrodes inserted into the ground at separate points. It has been postulated that transient anomalous telluric currents are observed before earthquakes.

Frequency bands of electromagnetic waves Electromagnetic waves are classified by frequency bands as follows:
ULF (< a few Hz), ELF (a few Hz $\sim$ 3 kHz), VLF (3–30 kHz), LF (30–300 kHz), MF (300–3000 kHz), HF(3–30 MHz), VHF (30–300 MHz), UHF (300–3000 MHz), SHF (3–30 GHz). Not only ULF to VHF bands, but also infrared ($\sim 10^{13}$ Hz) and visible ($\sim 10^{14}$ Hz) bands are considered to be involved in earthquake-related electromagnetic waves.

Skin effect The intensity of electromagnetic wave decreases exponentially with distance in a conductive medium. In a simple case, the distance where the intensity becomes $1/e$, called the skin depth δ, is expressed as $\delta = \sqrt{2/\mu\sigma\omega}$, where μ and σ are magnetic permeability, and electric conductivity of the medium and ω is the angular frequency of the wave.

Ionosphere The upper atmosphere, where electrons are stripped off from oxygen and nitrogen atoms by solar radiation, is called the ionosphere. It consists of a D-layer (60–90 km), E-layer (90–130 km), F_1-layer (130–210 km), and F_2-layer (210–1000 km). Electron density is highest in the F_2-layer. The electron density of the ionospheric lower layer can be measured by ground-based ionosonde, whereas total electron content (TEC) of the whole ionosphere is estimated by global position system (GPS). Electric currents in the ionosphere produce transient variations of geomagnetic field. The suggestion has been made that the ionosphere is affected before earthquakes.

Definition of the Subject

Throughout most of human history, electromagnetic phenomena associated with earthquakes have been repeatedly told. A typical one is earthquake light. Until rather recently, however, most records were in the realm of folklore [31,71]. Since earthquakes are understood as a catastrophic event to occur when slowly increasing tectonic stress in the earth's crust reaches a critical level, it may well be expected that the same stress may give rise to some electric, magnetic, or electromagnetic phenomena (EM phenomena hereafter) and some persistent research on them was initiated more or less simultaneously in varied parts of the world in the 1980s in two main streams. One was monitoring of possible emissions from focal regions in a wide range of frequency from DC to VHF, whereas the other was to monitor the anomalous transmission of man-made EM waves of varied frequencies over focal regions. Theoretical and experimental studies on the mechanism of EM phenomena have also been made. This relatively new branch of science is now called Seismo-Electromagnetics.

These EM phenomena attract high attention for their possible usefulness in earthquake prediction, which is of immense societal importance and considered as one of the last frontiers in earth sciences. Because many of the EM phenomena are observed prior to earthquakes, they may serve as their precursors, which have been difficult to find by usual seismological and geodetic methods. The extremely interdisciplinary nature of the subject matter is the distinct feature of Seismo-Electromagnetics and the backgrounds of many research fore-runners are neither seismology nor geodesy, but other fields, e. g., general geophysics, solid state, statistical, and ionospheric physics, radio, space, and even biological sciences. This situation in turn tends to make their accomplishments difficult to be understood and accepted by the conventional earthquake community. Of course, EM phenomena do not cause earthquakes. Both EM phenomena and earthquakes are considered to be caused by regional or local tectonic stresses, but some EM phenomena seem to appear shortly before the occurrence of earthquakes. That EM phenomena do not cause earthquakes may be another reason why few seismologists are interested in them.

Introduction

Earthquake (EQ) related EM signals may be classified into two major groups, each covering wide frequency ranges. One is EM signals supposedly emitted from the focal zone and the other is anomalous transmission of EM waves over the epicentral region.

The emission type signals are reported for geo-electrical potential (telluric current) and geomagnetic field and for EM waves. The best known example of the former is the Seismic Electric Signals (SES) in the VAN-method which has been developed in Greece since the early 1980s and applied also in Japan since the 1990s [76,78,82]. SES are transient DC geo-electrical potential variations observed before EQs by dipoles of buried electrodes. Experimentally and theoretically, the VAN method is by far the best equipped method in this category and has survived debates [22,41]. Along with the current views that earthquakes are catastrophic events at a critical state of complex systems, a new time domain called Natural Time has been introduced to integrate SES with seismicity for short-term EQ prediction (e. g., [82]). Late in the 1980s, pre-seismic magnetic signals began to be reported. They were

ultra low frequency (ULF) anomalous changes observed before $M7.1$ Loma–Prieta EQ, 1988 [16] and $M7.1$ Spitak EQ, 1989 [39] followed by $M8.0$ Guam EQ, 1993 [29]. Among them the case of Loma–Prieta is considered as most convincing. For the higher frequency range, there have been reports of VLF signals received by aerial antennas [3,23,95]. There have been reports from Greece that even VHF signals have also been received [9]. For high frequency signals, considering the electric conductivity of the crust, their emergence to the surface from the source regions at depth presents problems to be solved.

Pre-seismic appearance of EM signals makes them useful for EQ prediction. It has often been questioned, however, why they appear only pre-seismically and not co-seismically. This point makes the scientific community dubious about EQ-related EM signals in general. Actually, co-seismic signals are routinely observed but not reported each time for the obvious reason that they are not useful for EQ prediction. However, all the co-seismic signals observed so far were found to occur at the time of the arrival of seismic waves. They are, therefore, co-seismic wave signals and not "true" co-seismic. The fact that no true co-seismic signals are observed may be an important clue in exploring the physical mechanism of signal generation.

The generation mechanism of EM signal emission may be different for different frequencies. They may involve the electro-kinetic effects and pressure-stimulated polarization effects [79] for DC to low frequency signals, and piezo-electric effects and exo-electron emission [10] for higher frequency ones. For the optical range like EQ light, a very different mechanism such as de-excitation of air molecules may be involved. Apart from these, a mechanism involving so-called positive holes (p-holes) in rock forming minerals under stress has also been proposed recently [18].

The second class of signals, i. e., the anomalous transmission of EM waves, began to be actively discussed in the late 1980s [24]. One of the best documented early cases may be at the 1995 Kobe earthquake in which received VLF radio waves for navigation purposes showed anomalies in both phase and amplitude a few days before the main shock [52]. Moreover, at the same Kobe EQ, it was first found that FM radio waves from stations beyond the line of sight can be received before main shocks [40].

The anomalous transmission of EM waves means that there are some anomalies in the path, i. e., ionosphere or atmosphere, over the epicentral region, which may be verified independently. Investigation to detect such changes has been vigorously conducted both by ionosonde [45] and by topside observation from satellites [63]. On the basis of these observations, the concept of Lithosphere-Atmosphere-Ionosphere (LAI) coupling through which pre-seismic changes in the earth's crust may be transferred to the upper atmosphere became one of the central issues of Seismo-EM studies.

This article does not deal with the aspects of EM studies of the earth devoted for elucidating the subterranean electrical structures. Although their EQ related structural time changes, if observed, would be of great interest, there has practically been no significant reported progress.

Telluric Current Anomalies and Natural Time

The VAN Method

The best example of modern research on DC electric signals is that of the VAN method [18,80,82], named after the initials of the founding Greek scientists, P. Varotsos, K. Alexopoulos, and K. Nomikos. The VAN group has been making actual short-term predictions of $M \geq 5$ Greek EQs during well over a couple of decades. The criteria for successful prediction imposed by themselves are: < a few weeks in time, < 0.7 units in magnitude (M, hereafter), and < 100 km in epicentral distance. The length of time window depends on the type of signals.

The changes in the geo-electrical potential differences between buried electrodes, called Seismic Electric Signals (SES), are continuously monitored at many stations (Fig. 1a). At each station, several short (50–200 m) dipoles in both EW and NS directions and a few long dipoles (2–20 km) in appropriate directions are installed. Compared with all earlier works using only one or two dipoles, adoption of the multiple dipole system was a distinct progress in noise rejection.

Amplitude of SES is of the order of 1 mV/100 m. There are four types of signals, i. e., single SES, SES Activity, Gradual Variation of Electric Field (GVEF), and short duration pulse. Single SES, having duration 1/2 min ∼ several hours, precedes single EQ, whereas SES Activity, which consists of a number of SES in a short-time, is followed by a series of EQs before the main shock (Fig. 1b). As will be explained later, SES Activity has been playing a major role in the recent VAN work related to Natural Time analysis. GVEF has amplitude an order of magnitude larger than usual SES, but is only rarely observed for large EQ. The last type, i. e., short duration pulses appear shortly (some minutes) before EQs. These pulses, with amplitude sometimes amounting orders stronger than SES, have received rather little attention mainly because their lead time of minutes has been considered too short for useful EQ prediction.

In the VAN type of observation, noise discrimination is critically important. To eliminate noise, they have developed a set of rules as follows:

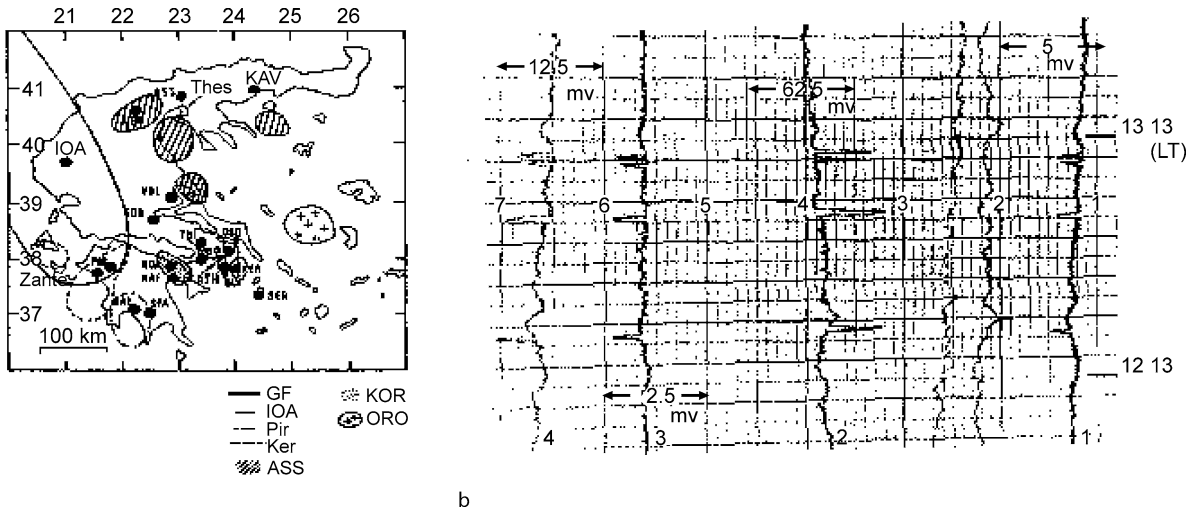

Earthquakes, Electromagnetic Signals of, Figure 1

a Distribution of VAN stations and "Selectivity map" of several stations as of 1996 (after [74]). For Zante, see text. b An example of SES Activity recorded on three short dipoles (labeled 1, 3, and 4) and long dipole (2) at Ioannina (IOA in a) station on August 31, 1988. Note that intensity scales in mV are different for different dipoles. The Killini–Varthelomio EQs were predicted based on these data [82]

1. Changes with magneto-telluric origin can be eliminated because they appear at all the stations simultaneously.
2. SES must appear simultaneously on all of short and long dipoles, but only at the concerned station.
3. SES must satisfy the $\Delta V/L =$ constant relation for short parallel dipoles, where ΔV is the amplitude of SES and L the dipole length.
4. The polarity and amplitude of SES of short and long dipoles must be compatible with the assumption that the source is distant compared with the dipole lengths.

The VAN group made two major discoveries. One is the so-called "Selectivity" and the other is the so-called "VAN-relation". The Selectivity has two aspects. (1) There are only selected sites which are sensitive to SES (sensitive sites). They were found only through testing at many sites: Almost 90% of sites were insensitive. This fact gives another strong reason why earlier efforts to catch precursory electric signals failed. (2) A sensitive site is sensitive only to SES from some specific focal area(s), which are not always in close proximity. A map identifying those focal area(s), SES from which are sensed by a site, is called the "Selectivity map" of that site (Fig. 1a), which provides information on the epicentral location of the impending EQ when a SES is observed at the site. The Selectivity is considered to originate from the inhomogeneity of the subterranean electrical structures, i. e., SES goes only through conductive channels. The VAN group has presented many model studies of channels [82]. So far, however, the real existence of such subterranean channels has not been verified by usual MT or other electric exploration techniques, possibly because the scales of the proposed channel structures are too small for the presently available resolving power.

The other discovery from the VAN research, i. e., the "VAN relation", is the following relationship among the focal distance, r, EQ magnitude, M and the observed intensity of SES, $\Delta V/L$.

$$\log(\Delta V/L \times r) = aM + b\,, \tag{1}$$

where a is a constant 0.34–0.37 and b is a site-dependent constant. Once the epicentral location is estimated from the Selectivity map mentioned above, M of the impending EQ can be assessed since both $\Delta V/L$ and r are known.

The VAN method has been a contentious subject (e. g., [22,41]). It is difficult, in principle, to prove the causal relationship between SES and EQ occurring at separate times and the only reasonable way to make the relationship credible would be to accumulate as many case studies as possible on one hand and to build plausible physical models on the other. Both endeavors have been published in papers too many to quote here and they are summarized in [82]. According to an independent evaluation (Uyeda et al. [75] and later additional check), out of 16 mb $\geq$ 5.5 EQs which occurred in the Greek region during Jan. 1, 1984–Jan. 1, 2004, 13 were successfully predicted. After the predictions of three large EQs in 1995,

there were no large EQs until another three mb $\geq$ 5.5 occurred in late 1997. It is remarkable that during the 2.5-year of quiescence no prediction was issued for the area and two out of the three 1997 events were predicted remarkably well. The score of the VAN method has been assessed by many authors. Most of the evaluations were in favor, but some were not. The low scores often resulted either when the assessors did not follow what the VAN group designated on the items such as allowable lead times for different type of SES, or the magnitude scales to use. Mulargia and Gasperini [53] claimed that "the apparent success of VAN predictions can be confidently ascribed to chance" and ignited heated debates (e. g., [22,41]). In the present authors' view, however, VAN has well survived them. In Greece, VAN-type SES has been observed by other groups also.

In Japan, VAN-type monitoring was initiated on a trial basis in the late 1980s and was expanded in 1996 [76]. Despite the serious problems caused by the high level of artificial noise, in particular from DC-driven electric trains, the existence of the VAN-type SES has been confirmed for $M > 5$ EQs occurring within $\sim$20 km or so of a station. Moreover, phenomena attesting to "Selectivity" were discovered. In the year 2000, a two-month long seismic swarm, with $\sim$7,000 $M \geq 3$ shocks and five $M \geq 6$ shocks, occurred in Izu Island region: See Fig. 2a. For this swarm activity, significant pre-seismic electric disturbances were observed [77]. From about 2 months before the swarm onset on June 26, innumerable clear, unusual geo-electrical potential changes started on Niijima Island (Fig. 2). These anomalous changes appeared only in the northern part of the island, possibly reflecting the extremely heterogeneous underground structure of the volcanic region.

Co-seismic signals have been observed for many EQs. However, they always started with the arrival of seismic waves and not at the origin time of EQs. The changes are probably local effects of passing seismic waves. There may be many reasons why no true co-seismic signals are observed. One is that, as laboratory experiments show, signals generated at ruptures are in much higher frequency range, so that they cannot be registered by usual high-cut measurement (0.1–1 Hz sampling) and the second is that, even when a higher sampling rate is employed, the high frequency signals attenuate before reaching the receiver. In fact, the pre-seismic stress accumulating process giving rise to SES and the instantaneous stress releasing event are physically very different processes and there seems to be no compelling reason why they generate similar signals.

Natural Time

Seismicity as a critical phenomenon has been actively discussed by many authors (e. g., [4,38,65,68,73]. It has been shown that SES and EQs reveal dynamic evolution characteristics to the critical stage when their time series is analyzed in the framework of natural time χ, which was introduced by the Varotsos' group (e. g. [82,83]). The symbol χ stands for the ancient Greek word $\chi\rho o\upsilon o\sigma$, which means "time". The possible usefulness of natural time analysis in predicting catastrophic events has been demonstrated not only for the subjects of our immediate concern, but also for other critical phenomena, including sudden cardiac death [84,85].

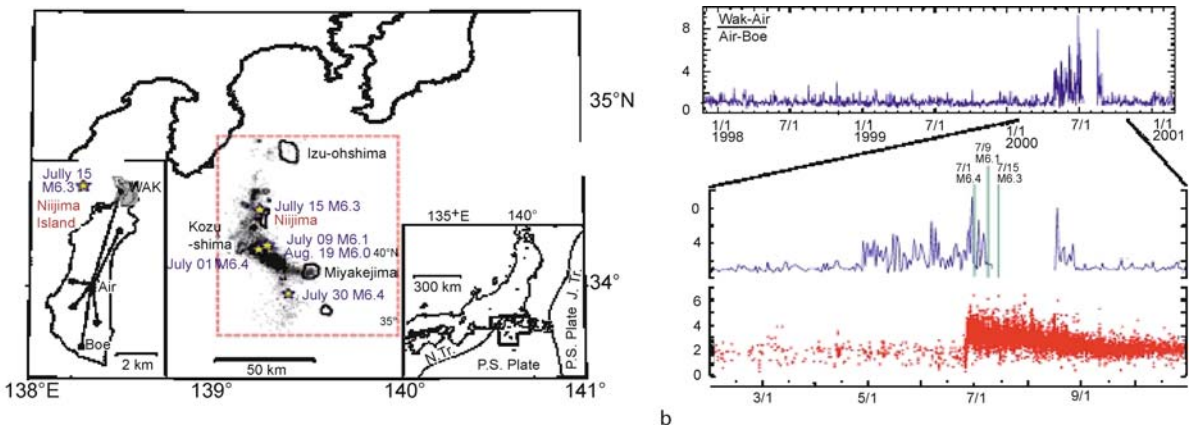

Earthquakes, Electromagnetic Signals of, Figure 2

a Seismic swarm activity in 2000 in Izu island region. *Inset in the left* shows the dipole configuration in Niijima Island. Each end of the long dipole had short dipoles. Only Wak-Air dipole showed the pre-swarm signals shown in b. The *bottom panel* in b shows seismicity (modified from [77])

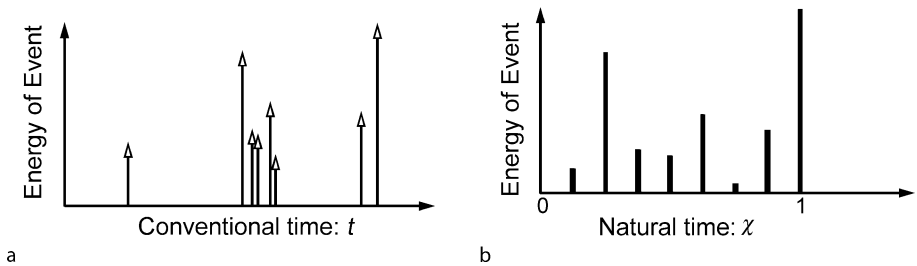

Earthquakes, Electromagnetic Signals of, Figure 3
Time series of events, a in conventional time *t*, and b in the natural time χ

In a time series comprised of N events, the natural time $\chi_k = k/N$ serves as the index for the occurrence of the k th event. In natural time analysis, the evolution of the pair of two quantities (χ_k, Q_k) is investigated where Q_k denotes the quantity proportional to the energy of the kth event. The time series of events as shown in Fig. 3a is expressed in natural time as in Fig. 3b. For the purpose of analysis, the following function $\Phi(\omega)$ was introduced.

$$\Phi(\omega) = \sum_{k=1}^{N} p_k e^{i\omega \frac{k}{N}} , \tag{2}$$

where $p_k = Q_k / \sum_{n=1}^{N} Q_n$ and $\omega = 2\pi\phi$, ϕ standing for the frequency in natural time (natural frequency). This $\Phi(\omega)$ should not be confused with the discrete Fourier Transform because ω is a continuous variable. If we regard p_k as the probability density function of χ, in analogy with probability theory, its Fourier transform $\Phi(\omega)$ may be regarded as the characteristic function of ω. The power spectrum of $\Phi(\omega)$, $\Pi(\omega) = |\Phi(\omega)|^2$, for the dynamical system approaching critical state with infinitely long-range temporal correlation was calculated to be as follows (see p.259–260 in [82]):

$$\Pi(\omega) = |\Phi(\omega)|^2 = \frac{18}{5\omega^2} - \frac{6\cos\omega}{2\omega^2} - \frac{12\sin\omega}{5\omega^3} . \tag{3}$$

Taylor expansion of Eq. (3) gives, for small values of $\omega(\omega \to 0)$,

$$\begin{aligned}\Pi(\omega) &= 1 - \kappa_1\omega^2 + \kappa_2\omega^4 + \kappa_3\omega^6 + \kappa_4\omega^8 + \cdots \\ &= 1 - 0.07\omega^2 + \cdots .\end{aligned} \tag{4}$$

Thus, a time series should show $\kappa_1 \approx 0.07$ when approaching the critical stage. The reason why the natural time domain is useful when information on intervals between events is lost while retaining only information on the order and relative importance of events is an intriguing question. As to this point, Abe et al. [1] have shown that this time domain in fact is optimal for enhancing the signals in time-frequency space.

In Greece, $\kappa_1 \approx 0.07$ was experimentally ascertained first for SES activities preceding four large EQs; 1995 $M6.6$ Kozani–Grevena EQ, 1995 $M6.5$ Eratini–Egio EQ, 1997 $M6.4$ Strofades EQ, 2001 $M6.6$ Aegean Sea EQ and later also for other major EQs, supporting this view [82,86]. Infinitely long-range temporal correlations of Greek SES were independently confirmed by Weron et al. [89].

In the case of seismicity, to investigate its time evolution, the power spectrum $\Pi(\omega)$ of the seismicity in natural time subsequent to associated SES activity was calculated as each consecutive EQ occurred. It was, then, shown that, for major Greek EQs, $\Pi(\omega)$ approached that of the critical state ($\kappa_1 = 0.07$) a few days before the main shocks [82]. This indicated that the seismicity approached the critical state at that time. This unexpected discovery may shed new light on the EQ generation mechanism itself. At the same time, this suggests the possibility of narrowing the time window of predicting EQs to a few days, when SES data are available. It may be added, albeit different from the Natural Time defined here, that an attempt was made to identify seismic quiescence with the viewpoint that the seismic process proceeds with its internal clock called "events time scale" [66].

Ultra Low Frequency (ULF) Anomalies

ULF generally means lower than several Hz. Research in this frequency range was started late in the 1980s. ULF signals are advantageous over those in higher frequencies because of their large skin depth. The best-known example is the case of the $M7.1$ Loma-Prieta (California) EQ in 1989 [16]. Observation was made at a site which happened to be at 7 km from the epicenter. The amplitude of the horizontal component started anomalous enhancement at about 2 weeks prior to and a sharp increase a few hours before the EQ. Figure 4 shows the records at 0.01 Hz band. The disturbance lasted for about 3 months after the

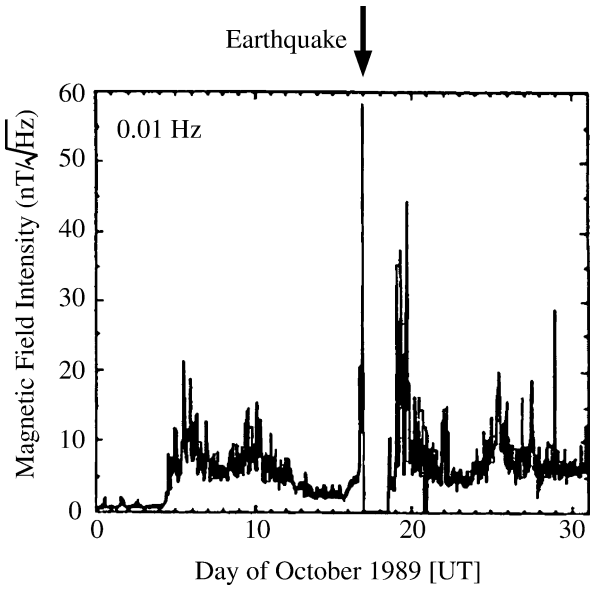

Earthquakes, Electromagnetic Signals of, Figure 4
The amplitude of the geomagnetic horizontal component at 0.01 Hz band [16]

EQ. These anomalous changes were not of solar terrestrial origins because they were not observed at other distant stations. Moreover, these have never been observed at any other time during the whole period of observation of more than 15 years. It was, thus, concluded that the anomalies were related to the EQ. Reports of observing pre-seismic ULF geomagnetic anomalies have been made also for M6.9 Spitak (Armenia) EQ in 1988 [39] and M8.0 Guam (Marianas) EQ in 1993 [29]. Further efforts in Japan and elsewhere have been summarized by Hattori [28]. It seems, however, that a more rigorous approach is needed to make the ULF studies sufficiently credible to the scientific community.

Higher Frequency Electromagnetic Emission and Earthquake Light

Pre-seismic electromagnetic wave emission in the VLF–LF-range has been reported since the 1980s. Gokhberg et al. [23] reported pioneering observations as shown in Fig. 5. Emissions at 81 kHz increased one or two hours before M6.1 and M5.3 earthquakes took place and decreased after the second shock.

Asada and his group started investigation of EQ-related VLF emissions in the early 1990s [3]. They monitored the wave forms of two horizontal magnetic components of VLF waves, through which the apparent incoming direction of VLF pulses was determined. They found that, before M5 class land EQs within 100 km of their stations, some pulses with a fixed incoming direction appeared and the EQs actually occurred in that direction, whereas the sources of overwhelmingly numerous and stronger noises were moving along with lightning sources (Fig. 6). Moreover, there is a well-documented report of undeniable noise in commercial MF radio bands, experienced by an automobile driver approaching Kobe, some minutes before the Kobe EQ of 1995 (see [55]).

Enomoto et al. [12] recorded anomalous pulses of geoelectrical current (HF-band) at Erimo station, Hokkaido, Japan, from February 2000 to March 2001 and from August to September 2003. The former anomalies occurred before and during the volcanic activity of Mt. Usu (200 km away), while the latter started one month before the 2003 September 26 M8.0 Tokachi-Oki EQ (80 km away). These were the only anomalies during their 10-year observation period.

For the Kobe EQ, while measuring sporadic Jovian decametric emissions with a radio interferometer at an observatory at about 80 km from the epicenter, unusual pulsed emissions at 22.2 MHz were detected tens of minutes both before and after the main shock [47]. Such unusual pulses have never been observed at other times and the possible source direction was estimated to be that of the main surface exposure of the EQ fault. There was no clear co-seismic radiation. Warwick et al. [88] reported a similar observation related to the 1960 Great M9.5 Chilean EQ.

Also in the high frequency range, Eftaxias et al. [9] have reported results obtained in Greece. Since 1994, they have been running a station on Zante Island in the Ionian Sea (see Fig. 1a), where installation was performed of (i) six loop antennas in EW, NS and vertical magnetic field at both 3 kHz and 10 kHz; (ii) $\lambda/2$ electric dipoles for 41, 54, and 135 MHz and (iii) two Short Thin Wire Antennas for ULF (< 1 Hz) anomalies. Sampling rate is 1 Hz. They report that MHz–kHz EM anomalies have been detected during a few days to a few hours prior to near-surface land EQ with $M > 6$, i. e., the 1995 Kozani-Grevena EQ and 1999 Athens EQ. The MHz radiation appeared earlier than the kHz. They interpret the phenomena as due to small-scale cracking and assume that the more grown-up cracks generate the lower frequency anomalies. They also look to the EQ as a critical phenomenon and suggest that the shift from MHz to kHz activity corresponds to an anti-persistency to persistency shift. In their observations, there was no co-seismic anomaly.

As mentioned earlier, however, transmission of EM waves in the conducting earth beyond the skin depth distance is an important unresolved problem common to all the topics reported in this section.

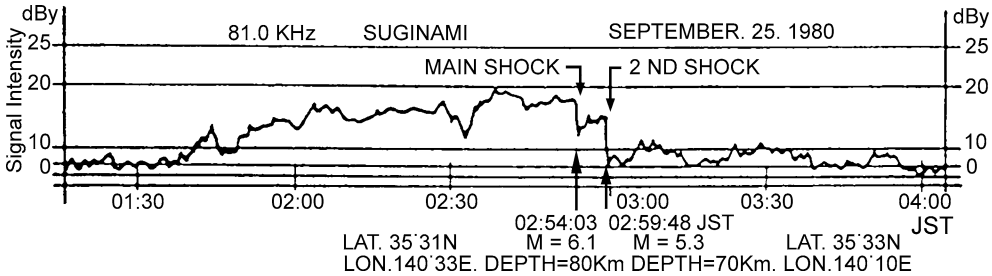

Earthquakes, Electromagnetic Signals of, Figure 5
Change of 81 kHz electromagnetic wave observed in Tokyo [23]

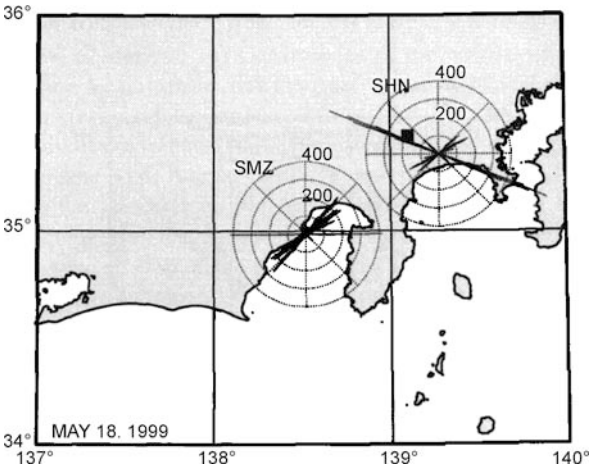

Earthquakes, Electromagnetic Signals of, Figure 6
Rose diagram of incoming VLF signals observed on May 18, 1999 at two sites. *M*4.1 EQ occurred at *black square point* on May 22 [3]

Earthquakes, Electromagnetic Signals of, Figure 7
Photograph of EQ light at Matsushiro seismic swarm taken by Kuribayashi (after [91])

EQ Light

Earthquake light (mostly co-seismic) has been reported all over the world from ancient Greek, Roman, and Chinese times. There is no doubt that the phenomena exist. Light emanates from the whole sky, or locally from the ground. The shape reported is like aurora, pole, flash, ball lightning, and so on, while the color widely ranged from blue and blue-white to red-yellow and orange. We, however, note that all reports may not be on natural phenomena but on some artificial effects such as sparks at power lines.

Galli [21] collected 148 eyewitness reports in late 19th Century Europe (see [71]). In Japan, Musha [54] collected about 2,000 eyewitness reports for 65 EQs, while Terada [69] discussed the theoretical aspects and suggested that the electro-kinetic effect may be a possible cause. From 1965 to 1967, there was a large EQ swarm at Matsushiro area in central Japan and numerous luminous phenomena were seen and photographs were taken as shown in Fig. 7 [91].

For the 1995, *M*7.3 Kobe EQ, Enomoto and Zheng [11] examined the trace of gas emission in the Awaji fault where the rupture started. They suggested that the gas plasma emission might have emitted the light. Kamogawa et al. [37] reported some independent witnesses that a luminous object moved a long distance a few seconds before the main shock in the direction of the rupture. Ikeya and Takaki [32] numerically showed that the screening charges neutralizing the polarized piezoelectrical rock may generate a strong co-seismic electric field, and the de-excitation of nitrogen molecules excited by collision of electrons accelerated by the electric field produce the blue EQ-light.

Lithosphere-Atmosphere-Ionosphere (LAI) Coupling

Pre-seismic atmospheric-ionospheric anomalies before EQs have been reported since the 1970s [2,20,24,27] and the concept of pre-seismic lithosphere-atmosphere-ionosphere coupling arose. Historical reviews and important works that are not introduced in this article may be tracked from references of Pulinets and Boyarchuk [61] and Kamogawa [33].

Liu et al. [43] found in Taiwan that the ionosonde measured critical plasma frequency, foF2, corresponding to the electron density of the ionospheric F2 layer, significantly decreased during afternoons within a few days before $M \geq 6$ EQs. For example, such ionospheric anomalies appeared 3 and 4 days before the 1999 M7.7 Chi-Chi EQ. Similar EQ-related electron density depression occurring above Taiwan Island was observed in the GPS total electron density (TEC) [44]. From such observations, Liu et al. [45] demonstrated that the appearance of the anomalies within 5 days was statistically significant at 5% level for the $M \geq 5.4$ EQs occurring within 150 km.

Sub-ionospheric anomalies before large EQs were reported by Gokhberg et al. [24] and Gufeld et al. [27]. They used VLF ship-navigation waves (10–20 kHz) and observed pre-seismic anomalies between the transmitter and the receiver during mid-night. Marenko [49] statistically supported the results of Gokhberg et al. [24], while Michael [50] obtained a less optimistic conclusion. Meanwhile, the studies have been further developed mainly in Russia, Japan, and Italy. For example, pre-seismic variations of terminator-times, i. e. the sunrise and sunset for VLF waves, were demonstrated [52]. Clilverd et al. [6], on the other hand, did not obtain similar positive results when they applied the terminator-time method to their 5-year data of reception at Faraday, Antarctica (receiver) of VLF waves transmitted from the northern United States. Maekawa et al. [48], measuring LF waves, statistically investigated the correlation between sub-ionospheric anomalies and $M \geq 6$ EQs in Japan and found that the amplitude and dispersion of received signals significantly decreased 2–6 days before the EQs. Thus, this issue is still controversial [34,64].

In the VHF range, Kushida and Kushida [40], while monitoring meteorites plunging into the high atmosphere by reflection of FM radio waves, detected anomalous reception, just a few days before the Kobe EQ, of the FM radio waves from distant (beyond the line-of-sight) stations. This was a new discovery and these authors consequently began extensive measurements on other EQs. With regard to the anomalous reception of VHF waves from transmitters beyond the line-of-sight, Fujiwara et al. [20] statistically showed significant enhancement of atmospheric anomalies lasting for a few minutes–several hours within 5 days before $M \geq 4.8$ EQs.

Mechanism of Pre-Seismic EM Phenomena

Generation Mechanism of EM Signals

Electro-Kinetic Effect The electro-kinetic effect can be a plausible source for SES (DC) and ULF emission. Mizutani et al. [51] first proposed a model in which, during the dilatancy stage, pore pressure in the dilatant region decreases and water flows into this region from the surrounding area, generating electric and magnetic precursors of EQs. Since then, many models have been proposed (e. g., [15,93]). Fedorov et al. [13], however, suggested that the expected magnitude of seismo-EM signals in the ULF-VAN range from an electro-kinetic source may reach the detection level only for a favorable set of crustal parameters.

Models Related to Defects in Solids A SES-generation model by pressure-stimulated currents (PSC) was proposed by Varotsos and Alexopoulos [79]. Their model is based on the physics of the point defects in solids. The impurities and vacancies have excessive and opposite-sign effective charges and form local electric dipoles. The directions of the local electric dipoles usually distribute randomly. Under an electric field, dipoles will align in its direction. The alignment is an activation process in which the time constant is an Arrhenius-type function of stress as well as temperature. Therefore, an avalanche of alignment takes place when stress approaches a critical level (Fig. 8). It has been later suggested that, instead of electric field, inhomogeneous deformation mentioned in the next paragraph may work to align the dipoles in the direction of the stress gradient (see [14]). This model is unique among other models in that SES is generated spontaneously during gradual increase of stress without requiring any sudden change of stress such as micro-fracturing. For the SES to work as a precursor, it is assumed that the critical level

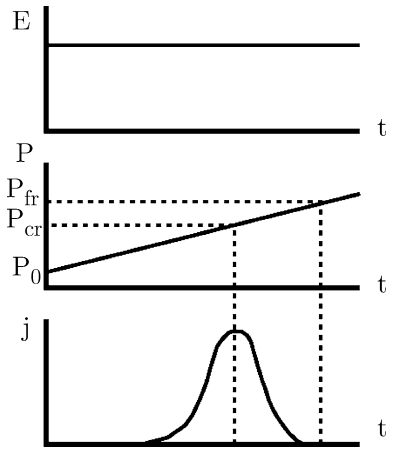

Earthquakes, Electromagnetic Signals of, Figure 8
Pressure-stimulated current j occurs at a critical pressure P_{cr} under the external electric field E. P_{fr} is the fracture pressure (after [79])

of stress for SES generation is lower than that of mechanical failure causing EQ. A thorough verification of the PSC model by laboratory pressure experiments is fatally lacking up to this stage.

In relation to SES generation, deformation-induced charged flow is an interesting possibility [56]. This flow was observed to take place as a result of inhomogeneous plastic deformation of ionic crystals, such as NaCl, in the direction of the stress gradient without applying electric field. It was interpreted that charge carriers are charged dislocations. Some experiments were conducted on rocks with similar results (see [82]). Independent of these, Freund and his colleagues have recently been proposing a unique mechanism for ULF electric signals ([18] and Ref. therein). They have discovered in the laboratory that when a block of igneous rock is put under stress locally, the rock turns into a battery without any external electric field (Fig. 9).

This striking phenomenon is interpreted as follows: A fraction of the oxygen anions in the rock-forming silicate minerals is not in their usual 2-valence state (O^{2-}) but in the 1-valence state (O^{1-}), which represent defect electrons, i. e., positive holes (*p*-holes). They are unstable and form more stable positive hole pairs (PHP), chemically equivalent to peroxy links, $O_3X/^{\circ\circ}\backslash XO_3$, which are electrically inactive. These dormant PHPs, however, are awoken by deviatory stress, and make the insulating host material a *p*-type semi-conductor. The *p*-holes flow out of the stressed volume because of mutual electrostatic repulsion. If scaling to earthquake size is allowed, the current thus produced may attain 10^3–10^5 A/km^3.

Other Models For generation of high frequency signals, models related to micro-cracking have been proposed from laboratory experiments. They are (1) Discharge of screening charge of piezo-electric polarization [31,94], (2) electrification of fresh crack surfaces [90], (3) exo-electron [10]. As to the occurrence of pre-main shock micro-cracking, there have been only a few reliable field reports. Furthermore, it might be pointed out that in these models much stronger co-seismic signals would be expected. Some ad hoc mechanism, therefore, would be needed to explain that no co-seismic signals have been observed so far in the field.

Transmission Mechanism of EM Signals

Even if EM signals are generated around a seismic focal region, signals except in the ULF range cannot be transmitted long distance in the conductive crust due to the decay caused by the skin effect, as long as the displacement current component is negligible. Even for DC signals, geometric decay would prohibit their long distance reception in a homogeneous or simple layered earth [5]. To overcome this difficulty, Varotsos et al. [81] proposed a conductive channel model, in which electric signals are transmitted through the conductive channel to a surface point close to the upper end of the channel. Freund [17] reported that, in the laboratory experiment, mobile positively charged holes (*p*-holes) appeared on the rock surface when a stress gradient was given to the rock sample. His results implied the possibility of appearance of a positively charged area at long distance on the surface before EQs. Kamogawa and Ohtsuki [35] proposed a model to explain how the higher frequency EM waves can be observed before EQs, i. e., longitudinal plasma waves excited by exo-electrons [10] may be transformed into EM waves by the surface roughness.

LAI Coupling Mechanism

If the pre-seismic atmospheric-ionospheric anomalies are real, some causative factors may be detected on the ground surface. Possible mechanisms of pre-seismic lithosphere-atmosphere-ionosphere coupling have been proposed by

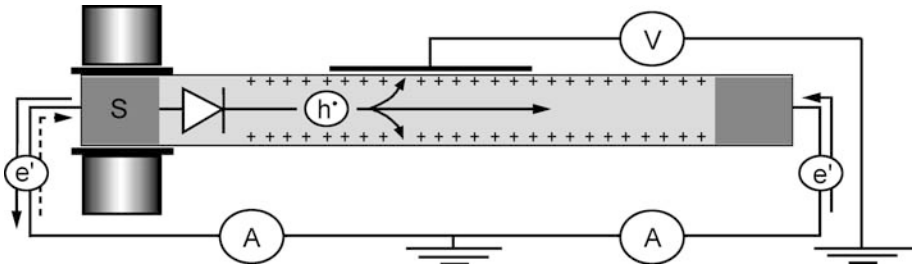

Earthquakes, Electromagnetic Signals of, Figure 9
Conceptual diagram of the battery current carried by electrons which flow out of the stressed portion *S* (*left*) through the outer circuit and by *p*-holes which close the circuit by flowing through the unstressed portion and meeting the electrons at the far end (*right*), flowing out from the stressed portion *S*, (after [18])

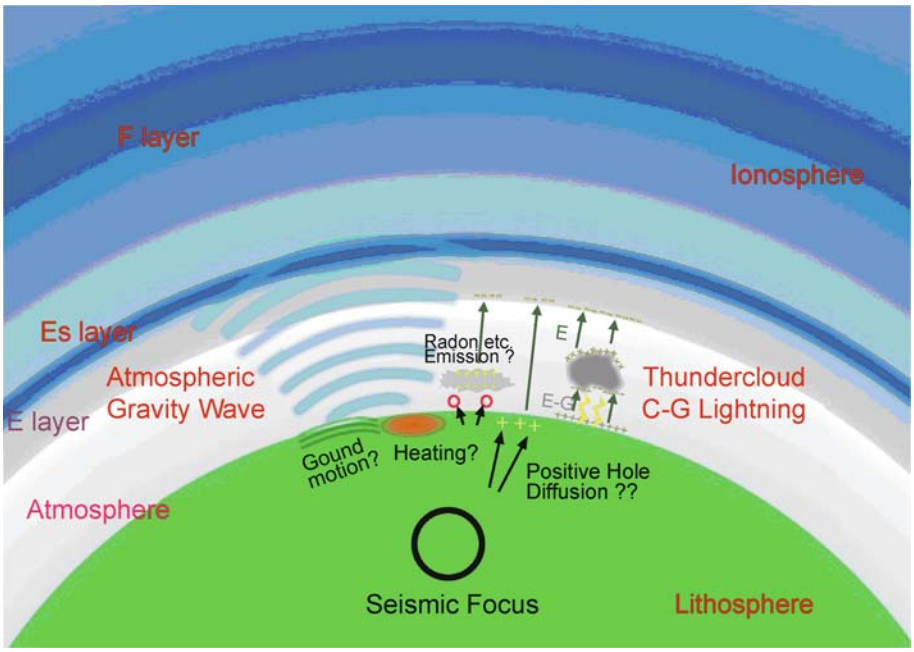

Earthquakes, Electromagnetic Signals of, Figure 10
Concept of LAI coupling (modified from [33])

many researchers. They may be categorized in two groups as shown in Fig. 10.

First, some atmospheric electric field $\vec{E}$ is generated on/near the ground surface during the pre-seismic period and it will cause the ionospheric anomalies [25,26,62]. Pulinets et al. [62] proposed that such an atmospheric electric field is caused by radon emission (see [30,87,92]). Alternatively, it is proposed that positively charged holes diffused from the seismic focal area to the ground surface generate the electric field [17]. However, such an electric field on the ground has not yet been observed even when pre-seismic ionospheric anomalies were detected [36].

Second, some researchers proposed that atmospheric gravity waves (AGW) propagate into the ionosphere, and disturb it before EQs [46,52,59]. The proposed source of AGW is the long-period ground oscillation or appearance of thermal anomalies on the ground. The former was inferred from some observations that co-seismic ground vibration actually excited AGW which propagated into the ionosphere (e. g. [8]). However, there is no report that long-period ground oscillations were detected at the pre-seismic stage even by high-sensitive superconducting gravimeter observations so far. The latter proposed source of AGW is the pre-seismic temperature rise, "thermal anomalies", reaching 2–4°C or higher in a wide area around impending EQs, based on satellite observation of enhanced infrared (IR) emission from the ground surface [7,70,72]. Many models have been put forward to explain the origin of the "thermal anomalies", including latent heat release at condensing water vapor due to enhanced radon emission. Pulinets [60] develops a scenario where the "thermal anomalies" give rise to ionospheric anomalies. Freund et al. [19] cast another interpretation on the enhanced IR emission based on their *p*-hole model mentioned above. When *p*-holes appear on the surface of the unstressed area, they form a positive charge layer and recombine to form the more stable O^{-}–O^{-} bond, emitting IR as de-excitation energy.

Future Directions

It seems that, despite much circumstantial evidence, earthquake related electromagnetic signals, in particular those at the pre-seismic stage, have not yet been completely accepted as real physical quantities. Putting the common indifference and prejudice of the conventional scientific community against new science aside, it seems appropriate at this stage to recognize that there are legitimate reasons for the critical views. In fact, most of the problems of fundamental importance in seismo-electromagnetics are still unresolved.

To name a few, propagation of high-frequency EM signals in conductive earth has been proven unequivocally enough neither empirically nor theoretically. Techniques of direction finding of EM signals at various frequency ranges and atmospheric-ionospheric anomalies have not shown sufficiently credible results yet. Solving these problems will be important issues in near future investigation.

The mechanisms of signal generation are still far from established. The majority of proposed mechanisms attribute pre-seismic signals to effects such as piezo-electric, electro-kinetic, charged dislocations, *p*-holes, and exo-electrons, all induced by stress release at micro-fracturing in the last stage of EQ preparation. In such cases, critics demand that by far the largest signals should be observed at the instance of the main shocks when the largest stress drop takes place. However, as described above, all the "co-seismically" observed electric or ULF signals are associated with the arrivals of seismic waves (to be called co-seismic wave) and are not co-seismic in the true sense. For higher frequency signals, even co-seismic wave signals have not been confirmed. This fact is a popular basis for negating the EQ-related signals in general. However, this very fact, i. e., the non-observation of true co-seismic signals in any frequency range may present some important clues with regard to the mechanism of both the signal generation and earthquakes as follows.

Numerous lab-experiments show strong co-fracturing signals in the form of high frequency EM waves. They are very different from the low frequency signals observed during pre-seismic stages. Thus, one explanation for non-observation of co-seismic signals in the ULF range is (1) field observation uses a low-pass recording system to avoid high frequency noise, and (2) high frequency signals are attenuated in short distance in the earth. However, this explanation seems to suffer from a weak point as follows: Since it takes some seconds for the fault motion of a large EQ to terminate, the overall signals should contain a low frequency component powerful enough to be caught by the low-pass recording system. Moreover, even higher frequency wave monitoring systems have not captured any co-seismic signals. All these seem to speak for non-generation of co-seismic signals of any frequency in the field and researchers had to devise some ad hoc scenarios as to how to reconcile with lab results, often invoking overgrowth of micro-faults by the time of main shocks to produce signals. For pre-seismic signal generation, these suggested mechanisms assume pre-seismic micro-fractures, which in fact, are micro-EQs that may be observed by high sensitivity seismic networks. Although depending on the required size, there has been no such observation, which constitutes another objection to pre-seismic EM signals.

The pressure-stimulated currents cited above regards the SES generation as a critical phenomenon. SES is supposed to be spontaneously generated when the gradually increasing tectonic stress level reaches a critical value. This seems to be the only proposed mechanism which needs no stress release by micro-fracturing or any special events, although it makes the causal relationship between SES and EQ less apparent. For the same reason, this mechanism does not need to generate any strong signals at EQ itself. Some of the other mechanisms, such as electro-kinetic, deformation-induced charged flow or *p*-holes flow mentioned earlier might be modified to fit the observation by incorporating the concept of critical state since they only need a development of stress gradient for signal generation.

Experimental verification of these mechanisms is urgently needed as it has been decisively inadequate. In any case, it should be kept in mind that the EQ preparation and EQ itself are different physical processes, the former being a gradual stress increasing process, whereas the latter is an instantaneous stress drop.

No true co-seismic signal in contrast to lab fracture experiments presents a question if the EQ is a fracture or not. It is now well-known that the EQ is a sudden sliding of pre-existing faults. However, according to Yoshida et al. [94], even stick-slip experiments on dry granite revealed strong signals at the time of slip which were understood as due to the piezo-electric effect. Non-generation of co-seismic EM signals still remains an important problem requiring further investigation.

Important unsolved questions are by no means confined to pre-seismic signal emissions. On the contrary, even more fundamental unsolved problems lie in LAI coupling. Here, the very origin on the ground to cause any of the suggested elementary agents, such as anomalous atmospheric electric field, atmospheric gravity wave, and thermal anomaly, is unknown. So far, observations on these features have been carried out by various researchers independently, so that the integration of fragmentary results for constructing a unified physical scenario of the whole process has been difficult. Since very active multi-national as well as multi-disciplinary cooperative research has been underway recently, involving GPS-TEC estimation and even topside measurement of the ionosphere by satellites of several nations, substantial progress in the upper end of LAI coupling is expected in the near future. Lately, active pursuit of EQ-related ionospheric anomalies has been made through topside and in-situ observation by satellites such as the French micro-satellite DEMTER [58]. How-

ever, the lower initiating side of the LAI coupling appears much more difficult to elucidate. It would require long sustained pre-seismic ground-based network observations on such phenomena as long-period ground motion and radon emission in as many earthquake prone areas as possible. But these tedious efforts should be enhanced on a global scale at all cost. Finally, it may be added that Kamogawa [33] pointed out that reported atmospheric-ionospheric anomalies might be caused by some EM phenomena which also trigger seismicity. For instance, suggestions have been made that geomagnetic storms [67] and cloud-to-ground lightning [42,57] may trigger EQs. It may be worthwhile to keep such a possibility of difference in cause and effect in mind in the future studies.

Bibliography

1. Abe S, Sarlis NV, Skordas ES, Tanaka HK, Varotsos PA (2005) Origin of the usefulness of the natural-time representation of complex time series. Phys Rev Lett 94:170601
2. Antselevich MG (1971) The influence of Tashkent earthquake on the Earth's magnetic field and the ionosphere. In: Tashkent earthquake 26 April 1966. FAN, Tashkent, pp 187–188 (in Russian)
3. Asada T, Baba H, Kawazoe K, Sugiura M (2001) An attempt to delineate very low frequency electromagnetic signals associated with earthquakes. Earth Planets Space 53:55–62
4. Bak P, Tang C (1989) Earthquakes as a self-organized critical phenomenon. J Geophys Res 94(15):15637–15639
5. Bernard P (1992) Plausibility of long distance electrotelluric precursors to earthquakes. J Geophys Res 97:17531–17536
6. Clilverd MA, Rodger CJ, Thomson NR (1999) Investigating seismo-ionospheric effects on a long subionospheric path. J Geophys Res 104(A12):28171–28179
7. Dey S, Singh RP (2003) Surface latent heat flux as an earthquake precursor. Nat Haz Earth Syst Sci 3:749–755
8. Ducic V, Artru J, Lognonne P (2003) Ionospheric remote sensing of the Denali Earthquake Rayleigh surface waves. Geophys Res Lett 30(18):1951. doi:10.1029/2003GL017812
9. Eftaxias K, Kapiris P, Polygiannakis J, Peratzakis A, Kopanas J, Antonopoulos G, Rigas D (2002) Experience of short term earthquake precursors with VLF-VHF electromagnetic emissions. Nat Hazards Earth Syst Sci 20:1–12
10. Enomoto Y, Hashimoto H (1990) Emission of charged particles from indentation fracture of rocks. Nature 346:641–643
11. Enomoto Y, Zheng Z (1998) Possible evidences of earthquake lightning accompanying the 1995 Kobe earthquake inferred from the Nojima fault gouge. Geophys Res Lett 25:2721–2724
12. Enomoto Y, Hashimoto H, Shirai N, Murakami Y, Mogi T, Takada M, Kasahara M (2006) Anomalous geoelectric signals possibly related to the 2000 Mt. Usu eruption and 2003 Tokachi-oki earthquake. Phys Chem Earth 31:319–324
13. Fedorov E, Pilipenko V, Uyeda S (2001) Electric and magnetic fields generated by electrokinetic processes in a conductive crust. Phys Chem Earth C 26:793–799
14. Fischbach DB, Nowick AS (1958) Some transient electrical effects of plastic deformation in NaCl crystals. J Phys Chem Solids 5:302–315
15. Fitterman DV (1978) Electrokinetic and magnetic anomalies associated with dilatant regions in a layered earth. J Geophys Res 83:5923–5928
16. Fraser-Smith AC, Bernardi A, McGill PR, Ladd ME, Helliwell RA, Villard OG Jr (1990) Low-frequency magnetic field measurements near the epicenter of the Ms 7.1 Loma Prieta earthquake. Geophys Res Lett 17:1465–1468
17. Freund F (2000) Time-resolved study of charge generation and propagation in igneous rocks. J Geophys Res 105:11001–11019
18. Freund F, Takeuchi A, Lau BES (2006) Electric currents streaming out of stressed igneous rocks – a step towards understanding pre-earthquake low frequency EM emissions. Phys Chem Earth 31:389–396
19. Freund FT, Takeuchi A, Lau BWS, Al-Manaseer A, Fu CC, Bryant NA, Ouzounov D (2007) Stimulated infrared emission from rocks: Assessing a stress indicator. eEarth 2:7–16
20. Fujiwara H, Kamogawa M, Ikeda M, Liu JY, Sakata H, Chen YI, Ofuruton H, Muramatsu S, Chuo YJ, Ohtsuki YH (2004) Atmospheric anomalies observed during earthquake occurrences. Geophys Res Lett 31:L17110. doi:10.1029/2004GL019865
21. Galli I (1910) Raccolta e classificazione de fenomeni luminosi osservati nei terremoti. Bull Soc Sis Ital 14:221–447 (in Italian)
22. Geller R (ed) (1996) Debate on "VAN". Geophys Res Lett 23:1291–1452
23. Gokhberg MB, Morgounov VA, Yoshino T, Tomizawa I (1982) Experimental measurement of electromagnetic emissions possibly related to earthquakes in Japan. J Geophys Res 87(B9):7824–7828
24. Gokhberg MB, Gufeld IL, Rozhnoy AA, Marenko VF, Yampolsky VS, Ponomarev EA (1989) Study of seismic influence on the ionosphere by super long-wave probing of the Earth ionosphere wave-guide. Phys Earth Planet Inter 57:64–67
25. Gokhberg MB, Morgounov VA, Pokhotelov OA (1995) Earthquake prediction, seismo-electromagnetic phenomena. Gordon and Breach, Reading, p 289
26. Grimalsky VV, Hayakawa M, Ivchenko VN, Rapoport YG, Zadorozhnii VI (2003) Penetration of an electrostatic field from the lithosphere into the ionosphere and its effect on the D-region before earthquakes. J Atmos Solar-Terr Phys 65:391–407
27. Gufeld IL, Rozhnoi AA, Tyumensev SN, Sherstuk SV, Yampolsky VS (1992) Radiowave disturbances in period to Rudber and Rachinsk earthquakes. Phys Solid Earth 28:267–270
28. Hattori K (2004) ULF geomagnetic changes associated with large earthquakes. Terr Atmos Ocean Sci 15:329–360
29. Hayakawa M, Kawate R, Molchanov OA, Yumoto K (1996) Results of ultra-low frequency magnetic field measurements during Guam earthquake of 8 August 1993. Geophys Res Lett 23:241–244
30. Igarashi G, Saeki S, Takahata N, Sumikawa K, Tasaka S, Sasaki Y, Takahashi M, Sano Y (1995) Ground-water radon anomaly before the Kobe earthquake in Japan. Science 269:60–61
31. Ikeya M (2004) Earthquakes and Animals. World Scientific, Singapore, 294 pp
32. Ikeya M, Takaki S (1996) Electromagnetic fault for earthquake lightning. Jpn Jour Appl Phys Part 2 35(3A):355–357
33. Kamagowa M (2006) Preseismic lithosphere-atmosphere-ionosphere coupling. Eos 87:417, 424
34. Kamogawa M (2007) Reply to comment on preseismic lithosphere-atmosphere-ionosphere coupling. Eos 88:248

35. Kamogawa M, Ohtsuki YH (1999) Plasmon model for origin of earthquake related electromagnetic wave noises. Proc Japan Acad 75(Ser. B):186–189
36. Kamogawa M, Liu JY, Fujiwara H, Chuo YJ, Tsai YB, Hattori K, Nagao T, Uyeda S, Ohtsuki YH (2004) Atmospheric field variations before the March 31 2002 *M*6.8 Earthquake in Taiwan. Terr Atmos Ocean Sci 15:445–461
37. Kamogawa M, Ofuruton H, Ohtsuki YH (2005) Earthquake light: 1995 Kobe earthquake in Japan. Atmos Res 76:438–444
38. Keilis-Borok VI, Soloviev AA (eds) (2003) Nonlinear dynamics of the lithosphere and earthquake prediction. Springer, Heidelberg, 335 pp
39. Kopytenko YA, Matishvili TG, Voronov PM, Kopytenko EA, Molchanov OA (1993) Detection of ultra-low-frequency emissions connected with the Spitak earthquake and its aftershock activity, based on geomagnetic pulsation data at Dusheti and Vardzia observatories. Phys Earth Planet Inter 77:85–95
40. Kushida Y, Kushida R (2002) Possibility of earthquake forecast by radio observations in the VHF band. J Atmos Electr 22:239–255
41. Lighthill J Sir (ed) (1996) A critical review of VAN. World Scientific, Singapore, 376 pp
42. Liu J, Chen Y, Ho Y (2004) A study of lightning activities and $M \geq 5.0$ Earthquakes in Taiwan during 1993–2002. Eos Trans AGU 85(47):T51B-0456 (Fall Meet. Suppl., Abstract)
43. Liu JY, Chen YI, Pulinets SA, Tsai YB, Chuo YJ (2000) Seismo-ionospheric signatures prior to $M \geq 6.0$ Taiwan earthquakes. Geophys Res Lett 27:3113–3116
44. Liu JY, Chen YI, Chuo YJ, Tsai HF (2001) Variations of ionospheric total electron content during the Chi-Chi earthquake. Geophys Res Lett 28:1383–1386
45. Liu JY, Chen YI, Chuo YJ (2006) A statistical investigation of pre-earthquake ionospheric anomaly. J Geophys Res 111:A05304. doi:10.1029/2005JA011333
46. Lizunov G, Hayakawa M (2004) Atmospheric gravity waves and their role in the lithosphere-troposphere-ionosphere interaction 1109. IEEJ Trans Fundam Mater 124-A:1109–1120
47. Maeda K, Tomisaka T (1996) Decametric radiation at the time of the Hyogo-ken Nanbu earthquake near Kobe in 1995. Geophys Res Lett 23:2433–2436
48. Maekawa S, Horie T, Yamauchi T, Sawaya T, Ishikawa M, Hayakawa M, Sasaki H (2006) A statistical study on the effect of earthquakes on the ionosphere, based on the subionospheric LF propagation data in Japan. Ann Geophys 24:2219–2225
49. Marenko VF (1989) Investigation of the relationship between seismic processes and disturbances to the lower ionosphere by means of VLF radio transmissions. Ph.D. Dissertation, USSR Academy of Sciences, Siberian Department, Irkutsk, 160 pp
50. Michael AJ (1997) Testing prediction methods: Earthquake clustering versus the Poisson model. Geophys Res Lett 24:1891–1894
51. Mizutani H, Ishido T, Yokokura T, Ohnishi S (1976) Electrokinetic phenomena associated with earthquakes. Geophys Res Lett 3:365–368
52. Molchanov OA, Hayakawa M (1998) Subionospheric VLF signal perturbations possibly related to earthquakes. J Geophys Res 100:1691–1712
53. Mulargia F, Gasperini P (1992) Evaluating the statistical validity beyond chance of VAN earthquake precursors. Geophys J Int 111:32–44
54. Musha K (1932) Investigations into the luminous phenomena accompanying earthquakes. Bull Earthquake Res Inst Tokyo Univ 10:666–673
55. Nagao T, Enomoto Y, Fujinawa Y, Hata M, Hayakawa M, Huang Q, Izutsu J, Kushida Y, Maeda K, Oike K, Uyeda S, Yoshino T (2002) Electromagnetic anomalies associated with 1995 Kobe earthquake. J Geodynamics 33:349–359
56. Norwick AS (1996) The golden age of crystal defects. Ann Rev Mater Sci 26:1–19
57. Ouzounov DP, Williams RG, Wohlman R (2000) A joint analysis of earthquake and lightning activity in the Southern California (1995–1999). Eos Trans AGU 81(19):S41B-08 (Spring Meet. Suppl. Abstract)
58. Parrot M (ed) (2007) First results of the DEMETER micro-satellite. Planet Space Sci 54(5):411–558
59. Pilipenko V, Shamimov S, Uyeda S, Tanaka H (2001) Possible mechanism of the over-horizon reception of FM radio waves during earthquake preparation period. Proc Japan Acad 77(Ser. B):125–130
60. Pulinets S (2007) Natural radioactivity, earthquakes, and the ionosphere. Eos 88:217–218
61. Pulinets S, Boyarchuk K (2005) Ionospheric precursors of earthquakes. Springer, p 315
62. Pulinets SA, Boyarchuk KA, Hegai VV, Kim VP, Lomonosov AM (2000) Quasielectrostatic model of atmosphere-thermosphere-ionosphere coupling. Adv Space Res 26:1209–1218
63. Pulinets SA, Legen'ka AD, Gaivoronskaya TV, Depuev VK (2003) Main phenomenological features of ionospheric precursors of strong earthquakes. J Atmos Solar Terr Phys 65:1337–1347
64. Rodger CJ, Clilverd MA (2007) Comment on preseismic lithosphere-atmosphere-ionosphere coupling. Eos 88:248
65. Rundle JB, Turcotte DL, Sammis C, Klein W, Shcherbakov R (2003) Statistical physics approach to understanding the multiscale dynamics of earthquake fault systems. Rev Geophys 41(4). doi:10.1029/2003RG000135
66. Schreider SY (1990) Formal definition of premonitory seismic quiescence. Phys Earth Planet Interi 61:113–127
67. Sobolev GA, Zakrzhevskaya NA, Kharin EP (2001) On a relation between seismicity and magnetic storms. Phys Earth 11:66–72
68. Sornette D (2000) Critical phenomena in natural sciences. Springer, Berlin, 434 pp
69. Terada T (1931) On luminous phenomena accompanying earthquakes. Bull Earthq Res Inst Tokyo Univ 9:225–255
70. Tramutoli V, Di Bello G, Pergola N, Piscitelli S (2001) Robust satellite, techniques for remote sensing of seismically active areas. Annali di Geofisica 44:295–312
71. Tributsch H (1982) When the snakes awake. MIT Press, Cambridge, 248 pp
72. Tronin AA (1996) Satellite thermal survey – a new tool for the study of seismoactive regions. Int J Remote Sens 41:1439–1455
73. Turcotte DL (1997) Fractals and chaos in geology and geophysis. Cambridge University Press, Cambridge, 398 pp
74. Uyeda S (1996) Introduction to the VAN method of earthquake prediction, a critical review of VAN. World Scientific, London, Singapore, pp 3–28
75. Uyeda S, Al-Damegh K, Dologlou E, Nagao T (1999) Some relationship between VAN seismic electric signals (SES) and earthquake parameters. Tectonophysics 304:41–55
76. Uyeda S, Nagao T, Orihara Y, Yamaguchi Y, Takahashi T (2000) Geoelectric potential changes: Possible precursors to earthquakes in Japan. Proc Nat Acad Sci USA (PNAS) 97:4561–4566

77. Uyeda S, Hayakawa M, Nagao T, Molchanov O, Hattori K, Orihara Y, Gotoh K, Akinaga Y, Tanaka H (2002) Electric and magnetic phenomena observed before the volcano-seismic activity 2000 in the Izu Island Region, Japan. Proc Nat Acad Sci USA (PNAS) 99(11):7352–7355
78. Varotsos P, Alexopoulos K (1984) Physical properties of the variations of the electric field of the earth preceding earthquakes. Tectonophysics 110:73–125
79. Varotsos P, Alexopoulos K (1986) Stimulated current emission in the earth and related geophysical aspects. In: Amelinckx S, Gevers R, Nihoul J (eds) Thermodynamics of point defects and their relation with bulk properties. North Holland, Amsterdam
80. Varotsos P, Kulhanek O (eds) (1993) Measurements and theoretical models of the Earth's electric field variations related to earthquakes. Tectonophysics 224:1–288
81. Varotsos P, Sarlis N, Lazaridou M, Kapiris P (1998) Transmission of stress induced electric signals in dielectric media. J Appl Phys 83:60–70
82. Varotsos PA (2005) The physics of seismic electric signals. TerraPub, Tokyo, 338 pp
83. Varotsos PA, Sarlis N, Skordas E (2002) Long-range correlations in the electric signals that precede rupture. Phys Rev E 66:011902
84. Varotsos PA, Sarlis NV, Skordas ES, Lazaridou MS (2004) Entropy in the natural time domain. Phys Rev E 70:011106
85. Varotsos PA, Sarlis NV, Skordas ES, Lazaridou MS (2005) Natural entropy fluctuations discriminate similar-looking electric signals emitted from systems of different dynamics. Phys Rev E 71:011110
86. Varotsos PA, Sarlis NV, Skordas ES, Tanaka HK, Lazaridou MS (2006) Entropy of seismic electric signals: Analysis in natural time under time reversal. Phys Rev E 73:031114
87. Wakita H, Nakamura Y, Notsu K, Noguchi M, Asada T (1980) Radon anomaly: A possible precursor of the 1978 Izu-Oshima-Kinkai Earthquake. Science 207:882–883
88. Warwick JW, Stoker C, Meyer TR (1982) Radio emission associated with rock fracture: Possible application to the great Chilean earthquake of May 22 1960. J Geophys Res 87:2851–2859
89. Weron A, Burnecki K, Mercik S, Weron K (2005) Complete description of all self-similar models driven by Lévy stable noise. Phys Rev E 71:016113
90. Yamada I, Masuda K, Mizutani H (1989) Electromagnetic and acoustic emission associated with rock fracture. Phys Earth Planet Int 57:157–168
91. Yasui Y (1968) A study on the luminous phenomena accompanied with earthquakes (part 1). Mem Kakioka Mag Obs 13:25–61
92. Yasuoka Y, Igarashi G, Ishikawa T, Tokonami S, Shinogi M (2006) Evidence of precursor phenomena in the Kobe earthquake obtained from atmospheric radon concentration. Appl Geochem 21:1064–1072
93. Yoshida S (2001) Convection current generated prior to rupture in saturated rocks. J Geophys Res 106(B2):2103–2120
94. Yoshida S, Uyeshima M, Nakatani M (1997) Electric potential changes associated with slip failure of granite: Preseismic and coseismic signals. J Geophys Res 102:14883–14897
95. Yoshino T, Tomozawa I, Sugimoto T (1993) Results of statistical analysis of low-frequency seismogenic EM emissions as precursors to earthquakes and volcanic eruptions. Phys Earth Planet Interi 77:21–31

Earthquake Source: Asymmetry and Rotation Effects

Roman Teisseyre
Institute of Geophysics, Polish Academy of Sciences, Warsaw, Poland

Article Outline

Glossary
Definition of the Subject
Asymmetric Continuum and Rotation Effects
Earthquake Source: Fracture Processes
Final Remarks
Acknowledgments
Bibliography

Glossary

We use the tensor notation and the summation convention for the repeating indices:

$$T_{ss} = \sum_s T_{ss} \,, \quad A_k B_k = \sum_k A_k B_k \,.$$

In some places we underline the symmetric and antisymmetric properties of the tensors using the (..), [..] brackets for indices:

$$S_{(ik)} = S_{(ki)} \,, \quad S_{[ik]} = -S_{[ki]} \,.$$

The deviatoric part of a symmetric tensor, $T^D_{(ik)}$, having zero value of trace, is defined as

$$T^D_{(ik)} = T_{(ik)} - \frac{1}{3}\delta_{ik} T_{(ss)} \,, \quad T^{\mathrm{A}}_{(ik)} = \frac{1}{3}\delta_{ik} T_{(ss)} \,,$$
$$T^D_{(ss)} = \sum_s T^D_{(ss)} = 0 \,,$$

where $T_{(ik)}$ is any symmetric tensor, while $T^{\mathrm{A}}_{(ik)}$ is the axial tensor.

In some places we use for the partial differentiations the following notations equivalently:

$$\frac{\partial u_n}{\partial x_k} \leftrightarrow u_{n,k} \,, \quad \frac{\partial U}{\partial x_\alpha} \leftrightarrow U_{,\alpha} \,,$$

where the indices with Roman characters run from 1 to 3, while those with Greek characters run from 1 to 4.

We use the fully antisymmetric tensor

$$\varepsilon_{lps} = \begin{Bmatrix} 1 \\ 0 \\ -1 \end{Bmatrix} \text{ for } \begin{Bmatrix} \text{even permutation of} \\ \text{repeating} \\ \text{odd permutation of} \end{Bmatrix} \text{ indices ;}$$

this fits to the tensor notation and helps to express some operations, e. g. curl:

$$\text{curl}\,\psi \;\Leftrightarrow\; \varepsilon_{lps}\frac{\partial}{\partial x_p}\psi_s\,.$$

Definition of the Subject

The problem of rotation waves becomes actual again due to the recent observations based on very precise instruments able to measure very small rotation time rates, and due to development and new approaches to the continuum theories.

Our aim is to present a consistent theory describing a continuum subjected to complex internal processes. First, we consider all possible kinds of point-related motions and deformations (strictly speaking, a single couple is not a point-source, as a displacement derivative only tends to a point, but a double couple enters into a family of point-deformations as it can be given by the string–string deformation, that is as a shear point-nucleus).

We define the complex rotation field which includes the spin and twist; the latter describes the angular oscillations of the shear axes and related amplitude variations. Twist point deformation can be represented by the string–string and string-membrane motions. A twist vector is defined as a vector perpendicular to a string–string plane; it becomes an important counterpart to spin and a key to presented theory, in which we shall also include the axial point deformation (e. g., the thermal one).

We believe that all point motions, displacement and rotation, and point deformations, axial and twist, shall be governed by some fundamental laws, and we intend to find the invariant forms of such relations in a frame of a modified continuum theory. Such a continuum theory may give us a new insight into the complexity of processes which can be included in a continual material description; we will demonstrate that interaction of these motions and deformations can lead us to a rich variety of internal processes.

We may mention also that the theory of continuum containing all these deformations with rotational motions (spin and twist), with the inner central motion and with defects may be projected on the intrinsic properties of non-Euclidean space. We confine ourselves only to a remark that application of differential geometry is extremely enlightening for the fundamental understanding of the nonlinear processes.

The independent rotation field can be related to the additional constitutive law joining the antisymmetric stresses and rotations, as proposed by Shimbo [1,2] in his considerations on the friction and fracturing processes; the introduced motion equations are equivalent to the stress moment – angular velocity relation.

For a reference to our consideration, we recall also the Kröner theory and its modifications as introduced by Teisseyre and Boratyński [3,4].

The possible dual description of any motions by means of displacement and rotation fields are being discussed and their formal equivalence concerning the propagation processes is demonstrated, but not valid for the source phenomena. Thus, we point out the importance of physically independent rotation motions in the inner granulation and fracture processes, and different types of rotational deformations are analyzed.

Our considerations bring several results for the following subjects.

For seismology:

- New description of the source processes including the role of rotational processes, and explanation of co-action of the slip and rotation motions
- Theory of the seismic rotation waves
- Thermodynamical conditions for a seismic energy release

For continuum and fracture mechanics:

- Theory of asymmetric continuum with the balance equations for the symmetric and antisymmetric stresses
- The relations between the asymmetric stresses and dislocation field
- A new approach to fracture processes with the hypothesis of the twist-shear release following the extreme angular deformation related to the internal particles, or grains constituting a continuum
- A synchronization role of the specific wave fields in the fracture processes
- Formation of a mylonite zone adjacent to fracturing and its constitutive description

For fluid mechanics:

- Theory of the asymmetric fluid continuum with the non-vanishing rates of the asymmetric stresses
- Explanation of the extreme wave phenomena (solitons)

Asymmetric Continuum and Rotation Effects

Introduction

Earthquake rotation effects were observed and discussed at the time when the 19th century seismological science was formed. Some eminent scientists, e. g., Charles Lyell (1797–1875), Charles Darwin (1809–1882), Robert Mallet

(1810–1881), and Alexander von Humboldt (1769–1859), raised the problem of the vortical movements, or vortex motion, induced by earthquakes. After the Lisbon earthquakes (1755) and those of Calabria (1783), many scientists focused their attention on the effects induced by such "vortical" waves. Robert Mallet was the first who precisely explained the observed rotation effects of some surface objects, pointing out the roles of the center of adherence of these objects and their inertia moment in relation to forces twisting the objects (see: Kozak [5], and Ferrari [6]). Many scholars tried to design the instruments to record the "vortical motions", but the first instrument prepared especially to record such motions was that constructed by Filippo Cecchi, the director of the Ximeniano Observatory of Florence, in 1875. Cecchi's electrical seismograph used sliding smoked paper. However, at that time it was too early to construct an instrument sensitive enough to obtain any traces of such wave motions.

The problem of seismic rotation waves was apparently closed after the Gutenberg [7] statement (1926) that such waves cannot propagate as they will be immediately attenuated, even when generated at the source. Of course the rotation effects remained, with the related explanation by Mallet, as objects of studies especially in the domain of the macroseismic observations.

From the contemporary point of view, two groups of achievements shall be mentioned; first, related to continuum theories, and second, related to development of the modern very precise instruments, able to record extremely small rotation time rates.

The continuum elastic theory bears from its very origin the serious limitation that the angular motions and related moments are not included. In such a situation, there was no place for a constitutive law describing the reaction between the stress moments and rotation processes. The lack of such a law automatically denies the existence of the rotation waves. We will return to these problems further on.

Experimental Evidence

The modern instrumentation techniques and the obtained results presenting the rotation wave seismograms need more attention. First, we can mention that maybe the first rotation seismogram (see Teisseyre [8]) was achieved in an indirect way: the azimuth array of horizontal seismographs, installed in one of the coal mines in Upper Silesia, Poland, to record the nearby tremors permitted one to deduce the rotational component of motions. However, the first, fully documented, rotation seismograms were obtained at the two geodetic fundamental stations in Germany (Wetzell) and in Australia (Cochard et al. [9]; Schreiber et al. [10]), equipped with ring-laser interferometers based on the Sagnac principle. These stations were established, primarily, to record very small deviations and disturbances of the Earth's rotational motion. However, the instruments having sensitivity up to 10^{-9} rad/s were able to record the rotation motions related to many distant earthquakes.

Latter, sensors of another type – the fiber-optic interferometers – were used by Takeo [11] especially for seismic observations; one version of his sensors included the tri-axial system. Jaroszewicz et al. [12] followed this system of rotation seismographs for the study of Silesian seismic events.

In a more traditional way, Moriya (see Moriya and Teisseyre [13]) has constructed the first rotation seismograph system consisting of a pair of anti-parallel seismographs; such a system, with the common suspension of the anti-parallel pendulums was repeated in latter constructions (Wiszniowski [14]).

Data collected by the recording systems mentioned above brought at least two important results:

- Records of different events in the very near field indicate that some events, e. g., shallow volcanic and those of explosion type, differ from the common characteristics by the extremely small rotation components (Teisseyre et al. [15]).
- Correlations between the rotation seismograms obtained from the ring-laser system and the rotation motions, curl u, derived from the array of seismometers (located at the same site) show almost perfect fit (Cochard et al. [9]).

Following the Cosserat theory and the micropolar and micromorphic theories (see Subsect. "Asymmetric Continuum Theory"), the independent rotation field, e. g., rotation related to grains or points of a continuum, were considered by Shimbo [1,2] in relation to the friction and fracture processes; the related constitutive law, we will call it the Shimbo law, joins the antisymmetric stresses with rotations and leads us towards the asymmetric continuum theory. The Shimbo law was latter generalized for the spin and twist rotation motions (Teisseyre et al. [16]; twist motion is introduced as the equivalent to oscillations of the shear axes).

We shall remind the reader that Gutenberg [7], in the frame of the classical elasticity theory, has proved that the independent rotation waves must be immediately attenuated. Now we know that this statement is due only to the fact that in the classical theory the rotations are not related to the stress or stress moment response; we cannot intro-

duce the constitutive law joining the symmetric stresses with the antisymmetric rotations.

However, in the asymmetric continuum theory such a constitutive law is required and appears as a natural element of the theory.

Displacements and Rotations

We cannot deny that independent rotation waves do not exist in a continuum built by the point-particles; however, this question is reduced to the magnitude of the independent rotations (that is, independent of rot **u**) generated in the seismic sources; the rotation wave motion is assured by the Shimbo constitutive law; this constitutive law joining the antisymmetric stresses and rotation relates directly to the friction as kind of material resistance:

$$S_{[ik]} = 2\mu^* \omega_{[ik]} \,, \quad S^D_{(ik)} = 2\mu E^D_{(ik)} \,, \tag{1}$$

where we have added the constitutive law for pure shear (further on we define the pure shear oscillations as twist), the tensors $S^D_{(ik)}$ and $E^D_{(ik)}$ are the deviatoric stress and strain tensors, and where μ is the rigidity modulus and the constant μ^* is defined as rotation rigidity, the constant entering in the antisymmetric stress – angular velocity relation (this constant is not equal to the rotation modulus in the stress moment – angular velocity relation); further on, we assume that both constants are equal, $\mu^* = \mu$, as it follows from the seismic wave observations.

We shall notice that both motions, displacements and rotations, are interrelated, which follows also from the fact that pure rotation, $\omega_{[s]}$, can be presented by means of the potentials represented by some displacement field, $\mathbf{u}^{\text{micro}}$; conversely, the displacements U (excluding those related to the scalar potential, e. g., those of thermal origin or that related to explosion process, that is, we put $\frac{\partial}{\partial x_i} U_i = 0$) can be described by the vector potentials represented by some rotation field $\Omega_{[.]}$:

$$\omega_{[.]} = \text{curl}\, U \,, \quad U = l^2 \text{curl}\, \Omega_{[.]} \tag{2}$$

and we arrive at the possible dual approach to the continuum mechanics.

When applying such equivalent approaches twice, e. g., from rotation field to displacement and again to rotations we obtain

$$\left\{ \omega_{[s]} = \varepsilon_{smn} \frac{\partial U_n}{\partial x_m}, \; u_i = l^2 \varepsilon_{iks} \frac{\partial \omega_s}{\partial x_k} \right\} \\ \to u_i = -l^2 \Delta U_i \quad \text{at} \quad \frac{\partial}{\partial x_i} U_i = 0 \,. \tag{3a}$$

Or otherwise:

$$\left\{ \omega_{[i]} = \varepsilon_{iks} \frac{\partial U_s}{\partial x_k}, \; U_s = l^2 \varepsilon_{smn} \frac{\partial \Omega_n}{\partial x_m} \right\} \\ \to \omega_{[i]} = -l^2 \Delta \Omega_i \quad \text{at} \quad \frac{\partial}{\partial x_i} \Omega_i = 0 \,, \tag{3b}$$

where l represents the basic intrinsic length measure.

The intrinsic length (Cosserat characteristic length) plays an important role in material properties; there is extensive literature related to this subject, however we limit ourselves to the remark that the displacement and rotation motions could be completely independent only for the case with $l = 0$, but such a case is excluded by quantum mechanics with the minimal Planck's length of the order of 10^{-34} m.

Excluding the axial motions, our consideration leads us to an apparent equivalence of the two descriptions, those by means of displacements and rotations. However, there remains the problem of the scales of these motions generated at the different fracture modes in the seismic source and also the scales of these motions observed at the Earth's surface.

The observed rotation fields and effects are usually much smaller than those related to the displacement field. There are, however, some exceptions leading to situations in which rotations play an important role, e. g.: the tilt motions, the rocking and tilting components related to building structures hit by strong ground motions and, as it will be discussed further on, the rotations generated by fracture which occurred under the prevailing compression load.

These equivalent descriptions of the motions in a continuum can be combined in the asymmetric theory, but we shall note that it could be constructed as well, the continuum theory neglecting completely the displacement fields and using only the rotation motions – such a case can be called a degenerated continuum.

Therefore, keeping in mind the above remarks that the rotations contribute to displacement derivatives and that displacements may contribute to rotations, we can state that these motions are interrelated; this statement is empirically supported by the above mentioned almost perfect fit between the derived rotations, curl u, and the rotations obtained from the ring-laser system data. Therefore, we think that the problem related to existence of rotation waves appears as an irrelevant question. However, we shall stress that the displacements and rotation motions differ in general, especially when considering their physical origins and effects. Instead of that problem, we propose to consider the classification of rotation motions from the point of view of their origins, scales and effects produced.

We propose the following classification:

- The micro-rotations or rotations, ω, as related to the wave motions based on the internal friction processes (rotation rigidity), as well as to slip motions with friction/fracture processes
- The meso-rotations related to material granulation and formation of the mylonite zones under the shear load fracturing processes
- The total rotations, ω^{T}, (the nomenclature introduced by Kröner [17]) related to the displacement field, **u**;
- The macro-rotations as related to fragmentation of material at the fracturing under compression load
- The mega-rotation effects related to the ground tilts and tilting of high objects on the ground

The important counterpart of rotational processes in the mechanics of fracturing and the related energy release shall be underlined (Teisseyre et al. [18]). Both under confining pressure and under external shears, the role of micro-fracturing in the bond breaking process is similar; however, we observe here the essential differences for rotations in larger scales.

The confining load condition leads to formation of the induced opposite arrays of dislocations, resulting in fragmentation processes and related macro-rotations. On the other hand, shear load leads to more concentrated fracturing along some planes. In the thermodynamical fracture band theory, see Subsect. "Earthquake Thermodynamics", we consider the additional super-lattice formed by dislocations and the properly defined vacant dislocations; with this advanced approach we express an effective role of dislocation band structure in the shear fracture thermodynamics. Similarly, the fragmentation and macro-rotation processes become more effective for the fracturing under confining pressure. Thus, we try to find a continuum description for a number of processes leading us from an elastic solid to that undergoing successive deterioration by crushing, granulation and fragmentation.

These considerations give us ground for the classification of the basic motions.

Basic Deformations and Simple Motions in Asymmetric Continuum

Basic and simple motions could be defined as those which may be reduced to the 3D point motion in the Cartesian or Riemann spaces, or those deformations conceived as the respective curvatures.

Considering basic motions, we can distinguish the simple motions. First the translation described by vector **u** and the independent rotation, called spin; in a non-homogeneous continuum, the grains having different material parameters, can rotate due to an interaction of a displacement field with the related moment of inertia of the grains. Then, we shall pass to the tensorial motions/deformations:

Any antisymmetric tensor can be related to the vectorial field, e. g., to spin motion; thus, we come again to the equivalent vector field. However, this simple spin motion, $\omega_{[.]}$, shall be treated basically as independent of the displacement rotation, however, both contribute to the total spin field, $\omega_{[.]} + \mathrm{curl}\,\mathbf{u}$, observed, e. g. in seismology. We have already mentioned that any symmetric tensor can be split into the axial and deviatoric tensors. The axial deformation tensor relates to the point deformations representing compression/dilatation nuclei, e. g., related to thermal anomaly.

These total axial oscillation motions include the equal translation motions along all three axes and relate to that part of the displacement field which can be derived from a scalar potential (see Fig. 1).

For the point-like continuum with the axial deformations, we would obtain either the Riemannian curvature or, for more complicated cases, the Riemannian torsion tensor.

There remains the deviatoric field to consider. This field relates to pure shear deformations; it is possible to

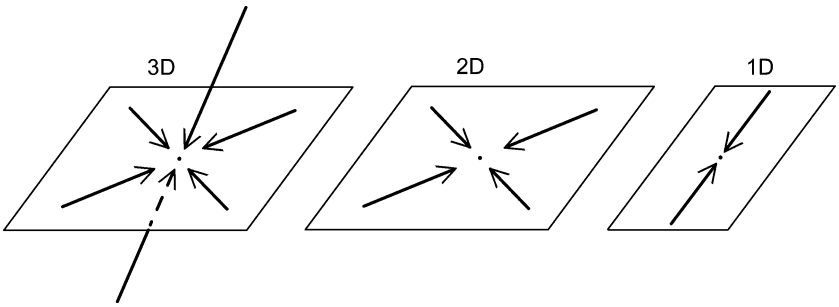

Earthquake Source: Asymmetry and Rotation Effects, Figure 1
Axial basic deformations (3D, 2D and 1D)

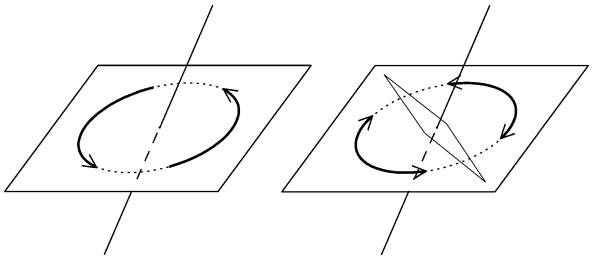

Earthquake Source: Asymmetry and Rotation Effects, Figure 2
Rotational motions: spin and twist

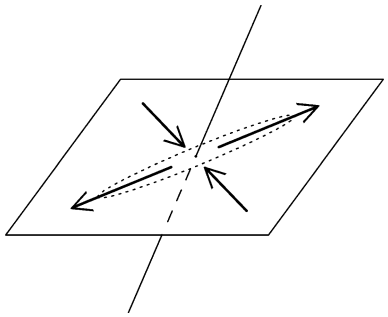

Earthquake Source: Asymmetry and Rotation Effects, Figure 3
String-string nucleus

show that the deviatoric field may be used to define the new antisymmetric tensor related to the simple deformations representing another kind of rotation motion – the twist deformation. These deviatoric deformations, for continua formed by particles/grains relate to pure shear oscillations (see Fig. 2; the left side shows a spin motion and the right side a twist one).

However, considering the point motions, it is better to relate these shear deformations to the equivalent twist tensor, $\omega_{(ks)}$; such motions contribute to the observed, e. g. in seismology, total shear/twist field. This field is directly related to the S-waves ($\mathbf{u}^S = l\omega$, where l is the effective radius of grains/particles forming the continuum and ω is the related rotation).

In the limit related to the point-like deformations, we would arrive at the string–string type motions (see Fig. 3) leading to the another representation of twist as vector, $\omega_{(k)}$, perpendicular to the string–string plane and having an appropriate magnitude of a string–string deformation (invariant representation of the string–string vector is discussed in Subsect. "Spin and Twist Motions", while determination of its amplitude Subsect. "Recording Spin and Twist Angle Variation").

A combination of the axial oscillation motions and those related to twist deformations leads to different forms

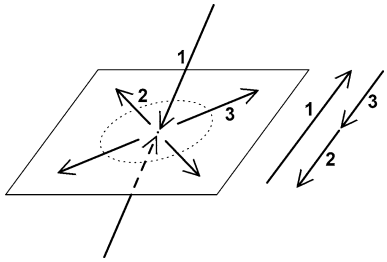

Earthquake Source: Asymmetry and Rotation Effects, Figure 4
String-membrane

of the point deformations: Fig. 4 relates to the string-membrane oscillations.

We shall repeat that all motions, except axial deformation, may be described by the displacement field, but this is not true for the point related spin and string–string deformation originated as the independent source motions.

Of course, there remains a number of the first, and higher order moments of these basic motions and deformations; such source models tend in a limit to a point source, but cannot be treated as exact point nuclei.

Asymmetric Continuum Theory

We shall add to the above considerations some arguments forming the fundamentals of the asymmetric continuum theory which is based on the asymmetric stresses:

- When studying the elastic field of an edge dislocation, we find some asymmetry in relation to its components in the plane perpendicular to its line (wedge direction); in confrontation with the symmetry of shears, this fact results in the asymmetry of stresses for a continuous distribution of dislocations (for screw dislocations such a contradiction does not exist). Therefore, a direct differential relation between any density of dislocations and the related stresses cannot be adequately found in a symmetric continuum.
- Fracture usually reveals its asymmetric pattern with a main slip plane; we shall believe that the premonitory processes, as described by deformations in a continuum with defects, develop also in an asymmetric way.
- In the classical continuum, the balance of angular momentum holds only if the stresses are symmetric; here the angular motions can be introduced only artificially with the help of a length element and a reference rotation point. This classical theory has also many other limitations and therefore many trials have been undertaken to generalize it. The asymmetric theory of elasticity with asymmetric stresses and couple-stresses

was founded by Nowacki [19]. However, a first generalization to include the moments in a continuum is due to Voigt in 1887 and a complete theory, including the asymmetry of stress and strain, is that known as the Cosserat theory of elasticity with the displacement vector and rotation vector [20]. Micropolar and micromorphic elastic theories were developed by Eringen and his co-workers and Mindlin (see: Eringen and Suhubi [21] and Mindlin [23]). Teisseyre [22] proposed a simpler version of the asymmetric theory which includes asymmetric stresses, strains and rotations, but in which the equations for the antisymmetric stresses differ from those of the couple moments in the Nowacki theory; their roles are interchanged, but both systems remain almost equivalent.

- Usually, when searching the fault slip solutions, we rely on classical elasticity with the friction constitutive laws introduced additionally in accordance with the experimental data. The obtained results well explain the observational data. Instead, we consider the consistent elastic continuum with asymmetric stresses and defects; such an approach enables one to study the defect interactions and elastodynamic solutions describing a slip propagation along a fault, including friction effects and related seismic radiation.
- The asymmetry of fields follows also from the notion of antisymmetric stresses considered by Shimbo [1,2] in relation to the friction processes and rotation of grains. Fracture processes develop usually along the main fault plane; hence there appears the initial asymmetry of the fracture pattern [24]; because of friction, the rotation of grains adjacent to the main slip plane causes an appearance of the antisymmetric part of the stresses and twist deformations. Following Shimbo [1,2], we introduced the constitutive law joining the antisymmetric stresses with the rotation nuclei (self-rotation field); without such a constitutive law any theory reduces both the rotation motions (except the rotation of displacements) and the related rotation waves to zero.
- In the asymmetric continuum, defined as that including both the symmetric stresses and the antisymmetric stresses, there appear also the rotational motions/deformations which split into pure spin and twist motions, the latter relate to the shear deformations of the grains; when considering the point-like nuclei, the twist deformation passes into 3D space torsion (Riemannian space).
- Experimental evidence for an appearance of spin and twist motions in a seismic field is based on the records of seismic rotation fields. For spin motion we shall be aware that the recorded rotation contains two elements: a rotation of displacements and an independent spin motion. Both these elements co-act in motion propagation and represent its dual description, but differ in their origin, depending on the source processes and material properties. At fracturing under a confining load, we deal rather with a high rotation release process and therefore the spin motion for the very near seismic events usually distinctly overpasses the displacement rotation. For some events of an explosive nature, or for some near-surface volcanic events, both the pure spin motion and rotation of displacement almost disappear; some observed effects might be related to a nearby $P \Leftrightarrow S$ conversion. For the strong motions which include a tilting component, the rotation of displacements exceeds a spin motion. In engineering seismology we observe that the rotation of displacements may exceed the pure spin motion; such an effect is due to magnification of a horizontal rotation of displacements and to the appearance of a rocking/tilting component of displacement rotation caused by the geometry of construction, especially for high buildings.

Self-Field Nuclei: Deviations from Classical Elasticity

Any continuum could be described using the Kröner approach [17] based on a concept of internal fields excited by a density of defects and internal nuclei; stresses and strains are related by the unique constitutive law for the ideal elasticity. This approach is equivalent to another approach in which we change the constitutive law in a way appropriate to describe the plastic, viscous and relaxation effects. In the Kröner continuum with a density of the internal point-like nuclei, the elastic strains, rotations and stresses can be expressed as differences between total and self-fields.

Following the Kröner approach, we can keep the ideal elastic relation for the stresses and strains, supplemented by the constitutive law joining the antisymmetric stresses with rotations, and we introduce the self/inner stresses, strains and rotations as related to the internal nuclei or defects: $\mathbf{S}^{\mathrm{S}}$, $\mathbf{E}^{\mathrm{S}}$, ω^{S}. We distinguish between the total stresses, strains and rotations related to the displacement field: $\mathbf{S}^{\mathrm{T}}$, $\mathbf{E}^{\mathrm{T}}$, ω^{T}, and the asymmetric elastic stresses, strains and rotations $\mathbf{S}$, $\mathbf{E}$, ω:

$$\mathbf{E} = \mathbf{E}^{\mathrm{T}} - \mathbf{E}^{\mathrm{S}}\,, \quad \omega = \omega^{\mathrm{T}} - \omega^{\mathrm{S}}\,, \quad \beta = \beta^{\mathrm{T}} - \beta^{\mathrm{S}}\,,$$
$$\mathbf{S} = \mathbf{S}^{\mathrm{T}} - \mathbf{S}^{\mathrm{S}}\,, \tag{4a}$$

$$\mathbf{E}^{\mathrm{T}}_{ki} = \mathbf{u}_{(i,k)}\,, \quad \omega^{\mathrm{T}}_{i,k} = \mathbf{u}_{[ik]}\,. \tag{4b}$$

The elastic and self-deformations, strains and rotations, and stresses can be, in general, asymmetric ones (see: Teis-

seyre and Boratyński [4]); under the conditions that the antisymmetric parts of the stresses and strains, as well as the symmetric parts for elastic and self-rotations, be mutually compensated for:

$$\mathbf{E}_{[ik]}+\mathbf{E}^{\mathrm{S}}_{[ik]}=0,\quad \mathbf{S}_{[ik]}+\mathbf{S}^{\mathrm{S}}_{[ik]}=0,\quad \omega_{(ik)}+\omega^{\mathrm{S}}_{(ik)}=0. \tag{5}$$

However, referring to our earlier papers (see: Teisseyre and Boratyński [4]) we shall then assume that the respective self-parts of the asymmetric strain and rotation are equal to each other:

$$E^{\mathrm{S}}_{[ik]}=\omega^{\mathrm{S}}_{[ik]}\,,\quad \omega^{\mathrm{S}}_{(ik)}=E^{\mathrm{S}}_{(ik)}\,, \tag{6}$$

where symmetric rotation is related to the shear axes oscillations (comp.: twist definition and Figs. 2 and 3). The elastic fields $\mathbf{S}, \mathbf{E}, \omega$ represent the physical fields, while the total fields $\mathbf{S}^{\mathrm{T}}, \mathbf{E}^{\mathrm{T}}, \omega^{\mathrm{T}}$ relate, according to the compatibility condition, to the displacement motions u_i, and the self-fields relate to the internal nuclei, defect densities and continuum structure.

Any deviations from the symmetry properties of fields, and any deviations from the ideal elasticity, can be described by suitable forms of the self-field, represented by the internal nuclei for both the defects and interaction fields.

The defects, dislocation and disclination densities can be defined, following Kossecka and De Witt [25], by considering the total disclosure and twist along a closed circuit (the Burgers vector and the Frank vector) and the appropriate form of the twist-bend tensor:

$$\begin{aligned} B_l &= -\oint\left[E_{(kl)}-\varepsilon_{lqr}\chi^{\mathrm{S}}_{kq}x_r\right]\mathrm{d}l_k\,,\\ \Omega_q &= -\oint\chi^{\mathrm{S}}_{kq}\mathrm{d}l_k=\theta_{pq}\mathrm{d}s_p \end{aligned} \tag{7}$$

and the definitions of the dislocation and disclination densities, α and θ, become based on the self-fields $E^{\mathrm{S}}_{(kl)}$ and χ^{S}_{kq}:

$$\begin{aligned} \alpha_{pl} &= -\varepsilon_{pmk}\left(\frac{\partial E^{\mathrm{S}}_{(kl)}}{\partial x_m}+\varepsilon_{klq}\chi^{\mathrm{S}}_{mq}\right)\,,\\ \theta_{pq} &= -\varepsilon_{pmk}\frac{\partial\chi^{\mathrm{S}}_{kq}}{\partial x_m}\,. \end{aligned} \tag{8}$$

After Teisseyre [26], the total twist-bend tensor can be defined as follows:

$$\chi^{\mathrm{T}}_{mq}=\varepsilon_{ksq}\frac{\partial\omega^{\mathrm{T}}_{mk}}{\partial x_s}\,,\quad \chi^{\mathrm{T}}_{mq}=\chi_{mq}+\chi^{\mathrm{S}}_{mq} \tag{9}$$

where, for the continuum with the asymmetric part of the stresses, we are not restricted to the compatibility condition for the twist-bend tensor (Kleman [27]).

The compatibility conditions for the asymmetric stresses and strains lead us to the physical equations for the dislocation and disclination densities in relation to the elastic fields of strain $E_{(kl)}$ and twist-bend χ_{kq}:

$$\alpha_{pl}=\varepsilon_{pmk}\left(\frac{\partial E_{(kl)}}{\partial x_m}+\varepsilon_{klq}\chi_{mq}\right)\,,\quad \theta_{pq}=\varepsilon_{pmk}\frac{\partial\chi_{kq}}{\partial x_m}\,. \tag{10}$$

Further on we will not rely on the Kröner approach, instead we will confine ourselves to a simpler approach given by the standard asymmetric continuum theory.

Asymmetric Continuum: Standard Theory

In opposition to the Kröner approach presented above, we may construct the asymmetric standard theory entirely related to the displacement field. Such a theory shall be based both on the symmetric and asymmetric stresses and on the related constitutive laws and motion equations. The asymmetric deformations contain the symmetric strain and antisymmetric rotation. Thus, our theory is based on two groups of relations; for the symmetric and antisymmetric fields:

$$\begin{aligned} &S_{kl}=S_{(kl)}+S_{[kl]}\,,\quad E_{kl}=E_{(kl)}\,,\quad \omega_{kl}=\omega_{[kl]}\,,\\ &D_{ks}=E_{ks}+\omega_{ks}\,, \end{aligned} \tag{11a}$$

where D_{ks} means the asymmetric deformation tensor.

However, when introducing the new material parameters (material structure indices): e^0, χ^0, we may join these deformation fields in an independent way, with some reference displacement motion:

$$E_{kl}=e^0\frac{1}{2}\left(\frac{\partial u_l}{\partial x_k}+\frac{\partial u_k}{\partial x_l}\right),\quad \omega_{kl}=\chi^0\frac{1}{2}\left(\frac{\partial u_l}{\partial x_k}-\frac{\partial u_k}{\partial x_l}\right). \tag{11b}$$

For an internal energy stored in such a medium we obtain:

$$E=S_{(ks)}E_{ks}+S_{[ks]}\omega_{ks}\,.$$

The indices e^0, χ^0 are not new constitutive constants, but they define families of solutions describing the complexity of deformation processes in continua; their ratio determines the phase shift between strain and rotation tensors. In this sense, the strain and rotation can be shifted in phase as follows from the particular deformations considered.

For the particular cases of these index values, e^0, χ^0, we have:

- The classic elasticity, obtained for $\chi^0 = 0$;
- For $e^0 = 0$ we obtain a granular/crushed medium filled with rigid spheres interacting by friction; when applying a torque load on its surface boundary, e. g., a cylindrical one, we would obtain only some angular deformation, and torque energy stored given as $E = S_{[ks]}\omega_{ks}$;
- The cases with $e^0 = \chi^0$ relate to the elastic continua with friction and different kinds of internal defects – different kinds of dislocation densities and the granulated materials;
- A continuum densely filled with the edge dislocations is described by the case $e^0 = -1$, $\chi^0 = -1$; while that of a partial content of that density $\alpha^E = \{0, 1\}$ by $e^0 = 1 - 2\alpha^E$, $\chi^0 = \alpha^E$;
- A continuum filled densely with the screw dislocations would be given by $e^0 = 2$, $\chi^0 = 2$; while that of a partial content of that density $\alpha^S = \{0, 1\}$ by $e^0 = 1 + \alpha^S$, $\chi^0 = 2\alpha^S$.

Further on, we will consider a more general continuum with the constitutive laws, including also the time rates processes; for such cases we might discuss in a similar way the different particular cases of the material structure indices including dynamic objects.

For the symmetric part of stresses we can assume the classical constitutive relation:

$$S_{(kl)} = \lambda\delta_{kl}E_{ss} + 2\mu E_{kl} \,. \tag{12}$$

But there is no problem to include in it the appropriate linear deviations related to visco-plastic effects.

To construct the asymmetric theory, we assume, after Shimbo [1,2], the appropriate constitutive law for the antisymmetric part of stresses. It joins the friction/fracture rotations with the antisymmetric stresses:

$$S_{[kl]} = 2\mu\omega_{kl} \,, \tag{13}$$

where rigidity constant μ plays the role of rotation rigidity entering in the antisymmetric stress-angular velocity relation (while the rotation modulus enters in the stress moment-angular velocity relation and may be considered as product of the rigidity μ and the Cosserat characteristic length l).

The motion equation for antisymmetric stresses $S_{[ni]}$ shall replace the balance law for the stress moments. To this end, we take the divergence of the rotation force moment acting on a body element due to the antisymmetric stresses (rotational moment of forces per infinitesimal arm length corresponding to stress moments), and, on the other hand, the balancing term – the acceleration related to angular momentum [3]:

$$\varepsilon_{lki}\frac{\partial^2}{\partial x_k \partial x_n}S_{[ni]} = \rho\frac{1}{2}\varepsilon_{lki}\frac{\partial^2}{\partial t^2}\left(\frac{\partial u_i}{\partial x_k} - \frac{\partial u_k}{\partial x_i}\right) + \varepsilon_{lki}\rho K_{[ki]} \,, \tag{14}$$

where we have put $e^0 = 1$ and we have introduced the body couple $K_{[ki]}$ equivalent to body moment $K_{[l]} = \varepsilon_{lki}\rho K_{[ki]}$.

With the compatibility condition introduced in a similar way as for the symmetric strains:

$$I_{[ij]} = \varepsilon_{imk}\varepsilon_{jns}\frac{\partial^2}{\partial x_m \partial x_n}\omega_{ks} = 0$$

we obtain from Eqs. (11b), (13) and (14):

$$\begin{aligned} &\frac{\partial^2 S_{[ki]}}{\partial x_s \partial x_s} = 2\rho\frac{\partial^2\omega_{ki}}{\partial t^2} + 2\rho K_{[ki]} \,, \quad \text{or} \\ &\mu\frac{\partial^2\omega_{ki}}{\partial x_s \partial x_s} - \rho\frac{\partial^2\omega_{ki}}{\partial t^2} = \rho K_{[ki]} \,, \end{aligned} \tag{15a}$$

where we have introduced also the body couple $K_{[ki]}$ or body moment $K_{[l]} = \varepsilon_{lki}\rho K_{[ki]}$.

Otherwise, we can write:

$$\mu\frac{\partial^2}{\partial x_k \partial x_k}\omega_{[l]} - \rho\frac{\partial^2}{\partial t^2}\omega_{[l]} = \rho K_{[l]} \,, \tag{15b}$$

where the left-hand side of this form presents the basic expression for the resulting stress moment divergence.

These relations are equivalent to the following ones:

$$\begin{aligned} &\frac{1}{l^2}\frac{\partial}{\partial x_k}M_{lk} = \varepsilon_{lki}\frac{\partial^2}{\partial x_k \partial x_n}S_{[ni]} = \varepsilon_{lki}\frac{\partial}{\partial x_n}\frac{\partial}{\partial x_n}S_{[ki]} \,, \\ &\frac{1}{l^2}M_{lk} = \varepsilon_{lki}\frac{\partial}{\partial x_n}S_{[ni]} \end{aligned}$$

or defining the angular moment Ξ_i, we obtain:

$$\frac{\partial}{\partial x_k}M_{ik} = 2\mu\Xi_i \,, \quad \Xi_i = l^2\varepsilon_{iks}\frac{\partial}{\partial x_n}\frac{\partial}{\partial x_s}\omega_{[kn]} \,.$$

From the motion equation for the symmetric part of stresses

$$\frac{\partial}{\partial x_k}S_{(kl)} = \rho\frac{\partial^2}{\partial t^2}u_l + F_l$$

and using the scalar and vector potentials

$$\begin{aligned} u_l &= l^2\frac{\partial}{\partial x_l}\varphi + l^2\varepsilon_{lps}\frac{\partial}{\partial x_p}\psi_s \,, \\ F_l &= l^2\frac{\partial}{\partial x_l}\Phi + l^2\varepsilon_{lps}\frac{\partial}{\partial x_p}\Psi_s \end{aligned}$$

we obtain:

$$(\lambda + 2\mu)\frac{\partial^2}{\partial x_k \partial x_k}\varphi = \rho\ddot{\varphi} + \Phi\,,$$
$$\mu\frac{\partial^2}{\partial x_k \partial x_k}\psi_s = \rho\ddot{\psi}_s + \Psi_s\,, \tag{16}$$

where according to Eq. (11b) we have introduced the index e^0 and we assume $\frac{\partial}{\partial x_s}\psi_s = 0$, $\frac{\partial}{\partial x_s}\Psi_s = 0$ and where we have introduced the intrinsic length unit l.

Here the potential ψ_s may be interpreted as rotation vector motions in another scale than that defined by relation curl $\upsilon = \omega$:

$$\varepsilon_{mql}\frac{\partial}{\partial x_q}u_l = -l^2\frac{\partial^2}{\partial x_k \partial x_k}\psi_m\,,$$

where ω means the micro-rotations, and ψ – the meso-rotations related to the granulated material (mylonite) and shear processes.

The strain tensor and its trace can be presented with the help of the introduced potentials as follows:

$$E_{lq} = l^2\frac{\partial^2}{\partial x_l \partial x_q}\varphi + \frac{1}{2}l^2\varepsilon_{lps}\frac{\partial^2}{\partial x_p \partial x_q}\psi_s + \frac{1}{2}l^2\varepsilon_{qps}\frac{\partial^2}{\partial x_p \partial x_l}\psi_s\,.$$

We can divide this expression into the axial and deviatoric parts:

$$E_{kk} = l^2\frac{\partial^2\varphi}{\partial x_s \partial x_s}\,,$$
$$E^D_{lq} = l^2\left(\frac{\partial^2\varphi}{\partial x_l \partial x_q} - \frac{\delta_{lq}}{3}\frac{\partial^2\varphi}{\partial x_s \partial x_s} + \frac{1}{2}\frac{\partial}{\partial x_p}\left(\varepsilon_{lps}\frac{\partial}{\partial x_q} + \varepsilon_{qps}\frac{\partial}{\partial x_l}\right)\psi_s\right). \tag{17}$$

Returning to our wave Eqs. (16) we arrive at the wave equations for the axial and deviatoric strain parts:

$$(\lambda + 2\mu)\Delta E_{kk} - \rho\frac{\partial^2 E_{kk}}{\partial t^2} = l^2\Delta\Phi\,, \tag{18a}$$

$$(\lambda + \mu)\left(\frac{\partial^2 E_{ss}}{\partial x_l \partial x_q} - \frac{\delta_{lq}}{3}\frac{\partial^2 E_{ss}}{\partial x_k \partial x_k}\right) + \mu\frac{\partial^2 E^D_{lq}}{\partial x_k \partial x_k} - \rho\frac{\partial^2 E^D_{lq}}{\partial t^2} = l^2\left(\frac{\partial^2\Phi}{\partial x_l \partial x_q} - \frac{\delta_{lq}\Delta\Phi}{3} + \frac{\varepsilon_{lps}}{2}\frac{\partial^2\Psi}{\partial x_p \partial x_q} + \frac{\varepsilon_{qps}}{2}\frac{\partial^2\Psi}{\partial x_p \partial x_l}\right). \tag{18b}$$

In terms of the potentials we obtain

$$\left(\mu\Delta - \rho\frac{\partial^2}{\partial t^2}\right)\left(\left(\frac{\partial^2}{\partial x_l \partial x_q} - \frac{\delta_{lq}}{3}\Delta\right)\varphi + \frac{\partial}{2\partial x_p}\left(\varepsilon_{lps}\frac{\partial}{\partial x_q} + \varepsilon_{qps}\frac{\partial}{\partial x_l}\right)\psi_s\right) = \left(\left(\frac{\partial^2}{\partial x_l \partial x_q} - \frac{\delta_{lq}}{3}\Delta\right)\Phi + \frac{\partial}{2\partial x_p}\left(\varepsilon_{lps}\frac{\partial}{\partial x_q} + \varepsilon_{qps}\frac{\partial}{\partial x_l}\right)\Psi_s\right)$$

Spin and Twist Motions

The spin motion is governed by Eq. (15a), or equivalently by its vector form Eq. (15b). We may consider the system related to the main shear axes or that related only to the off-diagonal components; in the latter case the motion equation for the deviatoric strains Eq. (18b) can be presented in the form of the rotation vector motion – the twist, $\omega_{(s)}$:

$$\{\omega_{(s)}\} = \{E^D_{23}, E^D_{31}, E^D_{12}\}\,. \tag{19a}$$

The defined twist motion, $\omega_{(s)}$, means the rotational oscillation of the off-diagonal shear axes of the deviatoric tensor (corresponding to oscillation of the main shear axes), E^D_{lq}, accompanied by the changes of the shear magnitude; such perturbation of the shear load may be caused by the internal fracturing processes (see Fig. 5).

Once having defined the twist vector field we can maintain its form due to the invariant properties of the

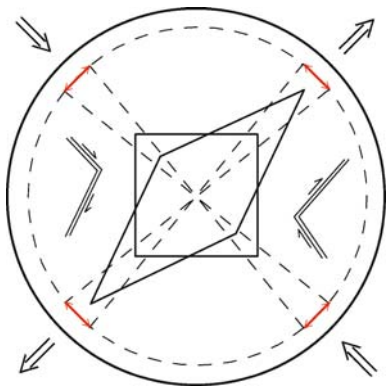

Earthquake Source: Asymmetry and Rotation Effects, Figure 5
Twist motion: rotational oscillations of the off-diagonal shear axes and internal fractures as the sources of perturbations; in the *center* we present an external shear deformation, while *arrows* along the *circle* give possible oscillations of the shear axes as influenced by some intrinsic processes, e. g. the fractures marked inside

Dirac tensors applied to the symmetric off-diagonal tensor $\omega_{(ik)}$ in its 4D form:

$$\omega_{(\lambda\kappa)} = \omega_{(1)}\gamma^1 + \omega_{(2)}\gamma^2 + \omega_{(3)}\gamma^4\gamma^2\gamma^3 = \begin{bmatrix} 0 & -\omega_{(3)} & -\omega_{(2)} & -\omega_{(1)} \\ -\omega_{(3)} & 0 & \omega_{(1)} & -\omega_{(2)} \\ -\omega_{(2)} & \omega_{(1)} & 0 & -\omega_{(3)} \\ -\omega_{(1)} & -\omega_{(2)} & -\omega_{(3)} & 0 \end{bmatrix}, \tag{19b}$$

where

$$\gamma^1 = \begin{bmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix}, \quad \gamma^2 = \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}, \quad \gamma^3 = \mathrm{i}\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix}, \quad \gamma^4 = \mathrm{i}\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \tag{19c}$$

and

$$\gamma^4\gamma^2\gamma^3 = \begin{bmatrix} 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \end{bmatrix}.$$

In a similar way, we may define the external off-diagonal part of the right-side expression of Eq. (18b):

$$\mathbf{Y}_{(lq)} = l^2\left(\frac{\partial^2}{\partial x_l \partial x_q}\Phi + \frac{\partial}{2\partial x_p}\left(\varepsilon_{lps}\frac{\partial}{\partial x_q} + \varepsilon_{qps}\frac{\partial}{\partial x_l}\right)\Psi_s\right).$$

For its 4D form we can write:

$$\mathbf{Y}_{(\lambda\kappa)} = \mathbf{Y}_{(12)}\gamma^1 + \mathbf{Y}_{(13)}\gamma^2 + \mathbf{Y}_{(23)}\gamma^4\gamma^2\gamma^3 = \begin{bmatrix} 0 & -\mathbf{Y}_{(12)} & -\mathbf{Y}_{(13)} & -\mathbf{Y}_{(23)} \\ -\mathbf{Y}_{(12)} & 0 & \mathbf{Y}_{(23)} & -\mathbf{Y}_{(13)} \\ -\mathbf{Y}_{(13)} & \mathbf{Y}_{(23)} & 0 & -\mathbf{Y}_{(12)} \\ -\mathbf{Y}_{(23)} & -\mathbf{Y}_{(13)} & -\mathbf{Y}_{(12)} & 0 \end{bmatrix}.$$

Using these definitions for the off-diagonal form Eq. (19b) we obtain

$$\mu\frac{\partial^2\omega_{(\lambda\kappa)}}{\partial x_k \partial x_k} - \rho\frac{\partial^2\omega_{(\lambda\kappa)}}{\partial t^2} = \mathbf{Y}_{(\lambda\kappa)}. \tag{20}$$

The defined 4D twist motion, $\omega_{(\lambda\kappa)}$, means the rotational oscillation of the off-diagonal shear axes of the deviatoric tensor, E^D_{lq}, accompanied by the changes of the shear magnitude; such perturbation of the shear load may be caused by internal fracturing processes (Fig. 5).

The spin and twist motions form the complex rotation field defined as:

$$\omega_s = \omega_{[s]} + \mathrm{i}\omega_{(s)} \tag{21}$$

From the balance relation (see: Subsect. "Recording Spin and Twist Angle Variation") we obtain the relations joining the spin and twist motions.

Defects: Dislocation and Disclination Densities

The classical approach to the dislocation and disclination densities is based on the Kröner description of continuum with the self-fields (compare Subsect. "Self-field Nuclei: Deviations from Classical Elasticity"). In the asymmetric homogeneous continuum, the defect density can be introduced using the modified definition of disclosure, B_l, and the following definition of the twist-bend vector (compare Eqs. (7–9)) we define:

$$B_l = \oint [E_{kl} - \omega_{kl}]\mathrm{d}l_k, \quad \Omega_q = \oint \chi^{\mathrm{T}}_{kq}\mathrm{d}l_k = \iint \theta_{kq}\mathrm{d}s_k, \tag{22a}$$

where for

$$\chi^{\mathrm{T}}_{mq} = \varepsilon_{ksq}\frac{\partial\omega_{mk}}{\partial x_s} \tag{22b}$$

the disclination density vanishes due to the compatibility conditions:

$$\theta_{pq} = \varepsilon_{pmk}\frac{\partial\chi^{\mathrm{T}}_{kq}}{\partial x_m} = \chi^0\varepsilon_{pmk}\varepsilon_{qns}\frac{\partial^2\omega_{ks}}{\partial x_m \partial x_n} = 0.$$

For the dislocation field we obtain (compare Eq. (8)):

$$\alpha_{pl} = \varepsilon_{pmk}\left(\frac{\partial E_{kl}}{\partial x_m} - \frac{\partial\omega_{kl}}{\partial x_m}\right) = \varepsilon_{pmk}\frac{\partial}{\partial x_m}\left(\frac{e^0}{2}\left(\frac{\partial u_l}{\partial x_k} + \frac{\partial u_k}{\partial x_l}\right) - \frac{\chi^0}{2}\left(\frac{\partial u_l}{\partial x_k} - \frac{\partial u_k}{\partial x_l}\right)\right). \tag{23}$$

With the help of the constitutive relations (12) and (13) we arrive at the relation between the dislocation density and asymmetric stresses:

$$\alpha_{pl} = \frac{\varepsilon_{pmk}}{2\mu}\frac{\partial}{\partial x_m}\left((S_{(kl)} - \frac{\nu}{1+\nu}\delta_{kl}S_{ii}) - S_{[kl]}\right). \tag{24}$$

We shall note that the material constants, e^0 and χ^0, define the types of defects and types of rotation nuclei; the complex constants will mean the constant phase shift between the fields.

Note that in the classic theory with defects, we distinguish also the different definitions for a dislocation field, e. g., the Burgers and Nye dislocations (comp: [28]).

For some particular case, $e^0 = 1$, $\chi^0 = -1$: we obtain a vanishing of defects, like:

$$B_l = \oint [E_{kl} - \omega_{kl}]\, \mathrm{d}l_k = \oint \frac{\partial u_l}{\partial x_k} \mathrm{d}l_k = 0\,, \quad \alpha_{pl} = 0\,. \tag{25}$$

This case may represent an extreme shear deformation.

We give also relations for another simple case, $e^0 = \chi^0$, which leads to dislocation density:

$$\begin{aligned} B_l &= \oint [E_{kl} - \omega_{kl}] \mathrm{d}l_k = \chi^0 \oint \frac{\partial u_k}{\partial x_l} \mathrm{d}l_k\,, \\ \alpha_{pl} &= \chi^0 \varepsilon_{pmk} \frac{\partial^2 u_k}{\partial x_m \partial x_l}\,. \end{aligned} \tag{26}$$

We can consider two particular cases, the first giving a relation between the asymmetric stresses and the edge type dislocations ($e^0 = -1$, $\chi^0 = -1$):

$$\alpha_{pl} = -\varepsilon_{pmk} \frac{\partial^2 u_k}{\partial x_m \partial x_l} \tag{27}$$

and the other which may describe the relation between the screw-type dislocations and asymmetric stresses ($e^0 = 2$, $\chi^0 = 2$):

$$\alpha_{pl} = 2\varepsilon_{pmk} \frac{\partial^2 u_k}{\partial x_m \partial x_l}\,, \tag{28}$$

where $p = l = s$, no summation over indices p and l.

Both of the considered cases could relate to the formation of the respective slip-discontinuities – Fig. 6.

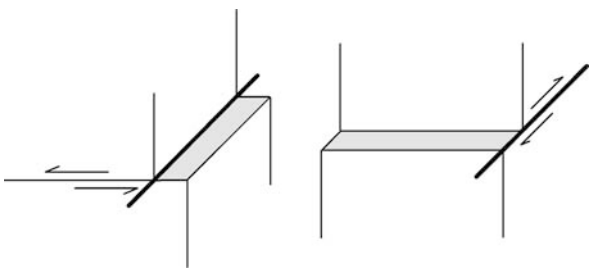

Earthquake Source: Asymmetry and Rotation Effects, Figure 6
The edge and screw dislocation types

We find that for a suitable choice of the disclosure definition, we may arrive at different definitions of the dislocation and disclination densities; in particular, we note that for the edge and screw dislocations we shall consider different rotation nuclei.

The case (Eq. (25)) presents the extreme deformation while the further cases present the standard source models and related rotations.

Balance Laws for the Rotation Field and the EM Analogy

The complex rotation field (21), $\omega_s = \omega_{[s]} + \mathrm{i}\omega_{(s)}$, may be presented in the tensor form:

$$\begin{aligned} \varepsilon_{kis}\omega_s &= \omega_{ki} \\ &= \begin{pmatrix} 0 & \omega_{[3]} + \mathrm{i}\omega_{(3)} & -\omega_{[2]} - \mathrm{i}\omega_{(2)} \\ -\omega_{[3]} - \mathrm{i}\omega_{(3)} & 0 & \omega_{[1]} + \mathrm{i}\omega_{(1)} \\ \omega_{[2]} + \mathrm{i}\omega_{(2)} & -\omega_{[1]} - \mathrm{i}\omega_{(1)} & 0 \end{pmatrix}. \end{aligned} \tag{29}$$

We can write the balance condition as

$$\iint \varepsilon_{kps} \frac{\partial}{\partial x_p} \omega_s \mathrm{d}s_k = \iint \left(\frac{\partial}{\mathrm{i}V\partial t} \omega_k + \frac{4\pi}{\upsilon} J_k \right) \mathrm{d}s_k\,, \tag{30}$$

where we introduce the current field J_k and velocity V.

Hence, we obtain the field equations for the complex rotation motions:

$$\varepsilon_{kps} \frac{\partial}{\partial x_p} \omega_s - \frac{\partial}{\mathrm{i}V\partial t} \omega_k = \frac{4\pi}{\upsilon} J_k \tag{31a}$$

or for spin and twist motions explicitly:

$$\varepsilon_{kps} \frac{\partial \omega_{[s]}}{\partial x_p} - \frac{1}{V} \dot{\omega}_{(k)} = \frac{4\pi}{V} J_k\,, \quad \varepsilon_{kps} \frac{\partial \omega_{(s)}}{\partial x_p} + \frac{1}{V} \dot{\omega}_{[k]} = 0\,. \tag{31b}$$

These equations lead us to the related wave forms:

$$\begin{aligned} \Delta\omega_{[n]} - \frac{1}{V^2} \ddot{\omega}_{[n]} &= \frac{4\pi}{V} \varepsilon_{npk} \frac{\partial}{\partial x_p} J_k\,, \\ \Delta\omega_{(n)} - \frac{1}{V^2} \ddot{\omega}_{(n)} &= 4\pi \frac{\partial}{\partial x_n} \varepsilon - \frac{4\pi}{V^2} \dot{J}_n\,, \end{aligned} \tag{32}$$

where $\omega_{[s],s} = 0$ and $\omega_{(s),s} = 4\pi\varepsilon$, and under the condition that the velocity, V, is simultaneously transformed according to relativistic rules for a sum of velocities.

The obtained wave equations coincide with those derived previously (compare Eqs. (15, and 18b) with the definition for twist – Eq. (19)).

We shall note that in the asymmetric elastic continuum, the bonds related to rotational deformations are considered as comparable to those related to elastic rigidity moduli. More complicated situations with the material constant appears in the micropolar and micromorphic theories with the infinitesimally small nuclei (Eringen [21]). In the asymmetric continuum theory, presented in our treatise, the displacements and rotations appear as equally and similarly treated independent fields. Here enter also in a natural way, the axial deformation fields, with a structure similar to that of a thermal field (comp. Eq. (18a)).

Finally, we shall note that these wave fields correspond with $\upsilon \to c$ to the EM fields, $\omega_{[n]} \to B_n$, $\omega_{(n)} \to E_n$. The form of the rotation complex tensor (Eq. (29)) is fully analogous to the definition of the complex electromagnetic field $F_s = B_s + \mathrm{i}E_s$.

Recording Spin and Twist Angle Variation

We shall find the suitable links between the defined fields and experimental data.

The spin motion can be precisely recorded by means of the Sagnac type interferometers (up to 10^{-9} rad/s); there are different types of such systems, e. g., ring laser and fiber optic, to record spin motion.

The angular twist oscillations and shear-twist motions we can record using a system of rotation seismometers. Such a system is based on the rotation seismographs that can record simultaneously the spin and twist angular motions [14,15,29].

In order to obtain the rotation motions, e. g., spin and twist, around the vertical axis we need the data from two parallel horizontal pendulums of opposite orientation. The observations collected clearly indicate that both the mean values of the spin and those of twist angular motions show the seismic oscillations with the same order of magnitudes [13,15,29].

We stress that the twist field, measured in this way gives only the angular variations of the off-diagonal axes of shears (19a); however, we may note that both the spin motion and the twist variation are mutually joint (see Eq. (31)) and therefore, we might theoretically derive knowledge of the shear state from the spin observations.

We shall add that when measuring the shear deformations with the help of a system of strainmeters, we can achieve more reliable and independent data on the shear-twist variations. Moreover, the strainmeter system can measure also the axial deformations.

Finally, we shall reply to the question of how we could compare the invariant twist field (19a and 19b) with the observed shear variations. An exact procedure requires the following: the 6 components of the shear strain determined in an observation site system shall be transformed, at each time moment, into the off-diagonal system:

$$\{E_{11},\ E_{22},\ E_{33}\ ,\ E_{23},\ E_{31},\ E_{12}\} \to \{E^D_{23},\ E^D_{31},\ E^D_{12}\} = \{\omega_{(s)}\}\ .$$

Conclusions

In the standard asymmetric continuum theory the defects defined are not the material defects, but only those related to the structural deformations. This standard asymmetric theory permits one to find the differential relation between the dislocation density and the asymmetric stress field. Moreover, in this theory we may consider also other deviations related to other defects or interaction fields; to this end we could apply the Kröner approach with the elastic, self- and total fields [4].

The other important conclusion is that the influence of rotational processes in earthquake sources spreads outward, because these waves are not attenuated strongly, as it was believed according to classical ideal elasticity.

Earthquake Source: Fracture Processes

Introduction

We start our considerations with the thermodynamical conditions related to seismic energy release, and then we consider the rotation counterpart in the fracturing.

We shall be aware that the rotation processes of different nature and scale take part in such extremely complicated fracture phenomena, in which the dynamic processes proceed together with the simultaneous changes of material properties (see Teisseyre [30]). We shall recall the special role of rotations in the energy release effectiveness under different load conditions, and further on we shall include the rotation impact on the granulation processes accompanying the material crushing.

The constitutive laws must undergo simultaneously considerable changes, from the rigid elastic to plastic, and further, to mylonite-type material (in tectonics the mylonite means the crushed, granulated and even partly melted material in zone adjacent to fracture plane). In the narrow zones adjacent to fracturing, the shear stresses break the molecular bonds and in the crashed rock material the stresses immediately drop to a much lower level, while together with the advancing material granulation, we shall include a rapid increase of the stress and strain rates. Finally, in that narrow zone adjacent to fracturing, the stresses and strains may be gradually neglected

and progressively replaced by their time-rates. To describe these processes we shall simultaneously introduce the changes into the related constitutive relations. In result, the rock properties in this zone may even approach those characteristic for fluid. Such conditions may permit one to include in the fracture description the transport Navier–Stokes relations. The fracturing transport process, the bond breaking and granulation processes force us to include in the fracturing description, the hypothesis that the twist-shear deformations leading to the bond breaking precede the rebound rotation motion by $\pi/2$ in phase; this means that the difference between the shear motion and spin motion shall reach minimum when the latter is shifted by $\pi/2$ in phase.

We shall underline that the considered conditions in the mylonite zone can serve as the basis to formulate the asymmetric fluid theory with the extreme motion phenomena and dynamic defect objects.

A counterpart to the rotations and rotation energy release at fracture processes (e. g., in an earthquake source) explains fragmentation and spall processes and makes it possible to estimate the efficiency of different fracturing modes. Again we shall underline that in any theoretical approach, the elastic rotation energy can be considered only when assuming the constitutive law joining rotations with the antisymmetric stresses or stress moments.

Teisseyre et al. [18] have reexamined Dietrich's compression experiments [31], coming to the conclusion that under the compression load, there arise at some centers in the source region the induced precursory shear stresses; at a fracturing event we would arrive at the coseismic rebound compensation leading to a release of the induced stresses by the rebound process. Similarly, the precursory rotations associated with the newly formed dislocations or cracks shall have an opposite orientation to that related to the coseismic process. At the precursory stage these repeated processes lead to micro-fracturings, while during the seismic event there will occur under compression load, the fracturing with the rock fragmentation and the rebound macro-rotations at the inner centers where the precursory induced stresses accumulate.

Earthquake Thermodynamics

Basic thermodynamic relations for line defects (dislocations and vacant dislocations) are derived under the assumption of a dense network of defects forming a kind of super-lattice [32,33,34,35]. The thermodynamic functions of line defects can be associated with defects in the super-lattice. Let us confine our considerations to the irreversible (plastic) deformations of solids.

To distinguish the thermodynamic functions used here from those used under pressure conditions, we will use the symbols with hat and we will consider only a pure shear work under shear load $S_{(..)}$ and under induced friction stress moment $S_{[..]}$ (see: Eqs. (13) and (14)) – we consider deformations at a constant volume; the work $\mathrm{d}\hat{W}$ done on a body (per unit volume), the internal energy change $\mathrm{d}\hat{U}$ and the heat received in an exchange with the surrounding $\mathrm{d}Q$ are related:

$$\begin{aligned} \mathrm{d}\hat{W} &= S\mathrm{d}E = S_{(..)}\mathrm{d}E_{(..)} + S_{[..]}\,\mathrm{d}\omega_{[..]} \geq 0\,, \\ \mathrm{d}\hat{U} &= \mathrm{d}Q + S\mathrm{d}E\,, \end{aligned} \tag{33}$$

where $\mathrm{d}E_{(..)}$ and $\mathrm{d}\omega_{[..]}$ are increments of strain and spin.

For the Helmholtz free energy $\hat{F}$ and Gibbs free energy $\hat{G}$ we have:

$$\begin{aligned} \hat{F} &= \hat{U} - T\tilde{S}\,, \qquad \hat{G} = U - SE - T\,; \\ T\mathrm{d}\tilde{S} &\geq \mathrm{d}Q\,; \qquad\qquad \mathrm{d}\tilde{S} \geq 0\,, \end{aligned} \tag{34}$$

where T is the absolute temperature and $\tilde{S}$ is the entropy; $\mathrm{d}\tilde{S}$ would be the entropy production due to the irreversible processes occurring inside the system.

The local formulation of the second law of thermodynamics requires that the entropy production be positive wherever an irreversible process occurs [34]. It is postulated that even outside equilibrium, the entropy depends only on the same variables as at equilibrium. In order to derive the expression for the entropy production, Prigogine [36] introduced some additional assumptions. Namely, he assumed that the entropy production can be determined for conditions near equilibrium.

Formation of a dislocation gives negative contribution to the Gibbs energy [37] and therefore it is not possible to find a minimum of the Gibbs function with respect to the number of dislocations. Thus, the dislocation distribution cannot exist as a thermodynamically stable system, since the Gibbs free energy has no minimum of any equilibrium concentration of dislocations.

However, for a dense dislocation distribution there enter the repulsive interactions between dislocations, and a kind of dislocation super-lattice can be considered [32,34,35]. The "ideal super-lattice" can be treated as a reference state and the real super-lattice, in the case of dense distribution of dislocations, can be in the equilibrium state. We define the vacant dislocations in the following way: to the randomly formed network of dislocations we shall add a number of line vacancies – the vacant dislocations – in such a way that as a result we obtain the super-lattice filled by dislocations and vacant dislocations. In this situation, a real distribution of dislocations can be

described as a departure from the state of ideal super-lattice, given by the amount of vacant dislocations.

These processes are accompanied by an internal friction related to displacement formed by dislocations and hence a spin motion appears as inherently present there.

The Gibbs energy minimum can now exist as the equilibrium number of the vacant dislocations. We can consider the structure of a cross-zone consisting of bands of layerlets; such a structure favors the appearance of some macroscopic dislocations under conditions of shearing deformation. The particular values of the Burgers vector become related to particular layer thicknesses. In this sense we suppose that a fine band structure could play the role of a quantization kind factor; this problem is related to the earthquake shear band model.

Consider a continuum that contains a regular (cubic) super-lattice of dislocation lines with a certain super-lattice parameter Λ ($\Lambda \gg \lambda$; λ relates to a basic rock lattice). The notion of the super-lattice is directly related to the shear band model of fracturing [34], see Fig. 7.

We associate the thermodynamic functions of line defects with the defects in a super-lattice; the Gibbs free energy may have a minimum corresponding to the equilibrium concentration of the vacant dislocations in the super-lattice. Many results can be now transferred from the thermodynamics of point defects (Varotsos and Alexopoulos [38]). The regular super-lattice, which includes the dislocations and vacant dislocations, may be described in a very rough approximation by a characteristic distance Λ (super-lattice constant). For the ideal super-lattice (no vacant dislocations), the mean value of distances following from distribution of dislocations defines the reference dislocation density $\alpha^0 = \lambda/\Lambda^2$, while for a real body with n dislocations we may add to it other $\hat{n}$ vacant dislocations in such a way that the whole set $n + \hat{n} = N$ (dislocations and vacant dislocations) fits to a regular super-lattice with the smallest error. For the density of dislocations α, and vacant dislocations $\hat{\alpha}$, we can write [34]:

$$\alpha = \left(1 - \frac{\hat{n}}{N}\right)\frac{\lambda}{\Lambda^2}, \quad \hat{\alpha} = \frac{\hat{n}\lambda}{N\Lambda^2}\frac{\lambda}{\Lambda^2}\exp\left(-\frac{\hat{g}^f}{kT}\right), \quad (35)$$

where the number $\hat{n}$ can be identified with an equilibrium value in relation to the formation energy of vacant dislocation $\hat{g}^f$ per length of the crystal lattice λ.

The stress field and the resistance stress (e. g., the drag resistance in a dislocation motion and the friction stress in a crack motion) are defined as [37]:

$$S = \frac{\partial \hat{W}}{\partial E}, \quad S_F \equiv \frac{\partial \hat{F}}{\partial E},$$

while the Gibbs function for a crystal containing the vacant dislocations can be written as

$$\hat{G} = \hat{G}^0 + \hat{n}\hat{g}^f - T\tilde{S}_c,$$

where $\tilde{S}_c$ is the configuration entropy.

Near the equilibrium state under a constant local shear S and temperature T the Gibbs energy is close to its minimum and the equilibrium values could be found as follows:

$$\left.\frac{\partial \hat{G}}{\partial \hat{n}}\right|_{S,T} = 0, \quad \hat{n}^{eq} = N\exp\left(-\frac{\hat{g}^f}{kT}\right),$$
$$\hat{\alpha} = \frac{\lambda}{\Lambda^2}\exp\left(-\frac{\hat{g}^f}{kT}\right), \quad \tilde{S}_c = \hat{n}\left(k + \frac{\hat{g}^f}{T}\right), \quad (36a)$$

while the Gibbs energy function becomes

$$\hat{G} = \hat{G}^0 - \hat{n}kT. \quad (36b)$$

The equilibrium free energy is less than that for an ideal super-lattice $\hat{G}^0$; the difference is kT per line vacancy, per length of crystal lattice.

For the point defect thermodynamics, Varotsos and Alexopoulos [38] have introduced the so-called $CB\Omega$ theory approximating the contribution to the Gibbs energy from the formation of point defects.

For the line vacancies, a change of the Gibbs energy depends on the stress level and resistance stress. Therefore, we postulate for the approximative value of such change per unit element (formation energy of vacant dislocation) the following expression defining the $C\mu b\lambda^2$ model:

$$\hat{g}^f = C\mu b\lambda^2, \quad \hat{n}^{eq} = N\exp\left(-\frac{C\mu b\lambda^2}{kT}\right), \quad (37)$$

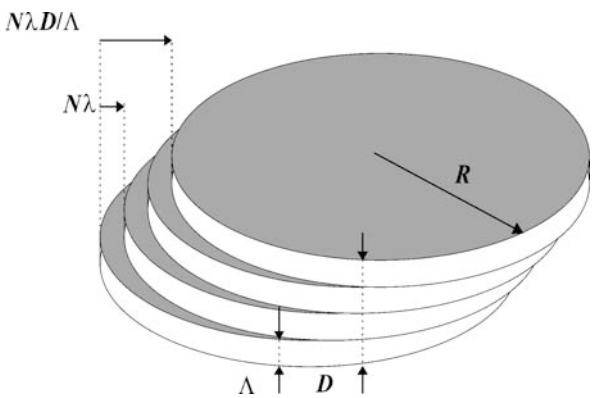

Earthquake Source: Asymmetry and Rotation Effects, Figure 7
Shear band model

where C is constant; $\hat{g}^f$ becomes here independent on stress load and resistance, μ is the rigidity, b is the Burgers vector of dislocation.

Concluding, a body containing some number of dislocations cannot be in a state of equilibrium; there is no minimum of the Gibbs function, because when reducing the number of dislocations we always get a smaller value of the free energy. For a dense distribution of dislocations we can assume, due to their interaction, that there exists a certain super-lattice composed of dislocations.

The equilibrium density of the vacant dislocations may be written now with help of Eq. (37)

$$\hat{\alpha} = \frac{\lambda}{\Lambda^2} \exp\left(-\frac{C\mu b\lambda^2}{kT}\right) \tag{38}$$

and becomes useful, when looking for the most probable density value of defects after the energy release in a fracturing process. The density $\alpha^0 = \lambda/\Lambda^2$ may be identified here with the reference density.

We can assume that before an earthquake a super-lattice is almost completely filled in by dislocations ($n \approx N$ and $\hat{n} \approx 0$). The maximum number of dislocations in arrays could reach the value $(\Lambda/\lambda)^2$ per area Λ^2. The total moment for an area $\Delta s = N\Lambda^2$ affected by the arrays of dislocation along the slip planes becomes:

$$\bar{M} = \mu\lambda\Delta s = \mu\lambda N\Lambda^2\left(\frac{\Lambda}{\lambda}\right) = \mu N\Lambda^3 .$$

After an earthquake, the number of vacant dislocations $\hat{n}$ shall increase, probably to the equilibrium value (37) and hence we can express the seismic moment by the number of coalescence processes related to surface element Λ^2 as equal to $\Delta\hat{n} = \frac{\Lambda}{\lambda}\hat{n}^{\mathrm{eq}}$; the factor Λ/λ expresses a maximum concentration of dislocations in the arrays.

We obtain for the seismic moment

$$\begin{aligned}\bar{M}_0 &= \bar{M}\Delta\hat{n} = \mu N\Lambda^3\left(\frac{\Lambda}{\lambda}\right)\Delta\hat{n}\\ &= \mu N\Lambda^3\left(\frac{\Lambda}{\lambda}\right)\exp\left(-\frac{C\mu\lambda\Lambda^2}{kT}\right),\end{aligned}$$

where C is constant for given structure.

Using the expression for a change of the free energy values we may include the formation of dislocation arrays along the glide planes and we put

$$G = G^0 + \Delta\hat{n}\left(\frac{\Lambda}{\lambda}\right)kT$$

According to these results, the total energy release ΔE and seismic moment are:

$$\begin{aligned}\Delta E &= G - G^0 = \Delta\hat{n}\left(\frac{\Lambda}{\lambda}\right)kT\\ &= \left(\frac{\Lambda}{\lambda}\right)NkT\exp\left(-\frac{C\mu\lambda\Lambda^2}{kT}\right) \text{ and}\\ \bar{M}_0 &= \mu\Lambda^3\frac{\Delta E}{kT} .\end{aligned} \tag{39}$$

This formula is an important relation between the energy release density and seismic moment density; for instance, for a given ΔE the elementary seismic moment $\bar{M}_0$ decreases with temperature. Free energy related to defect formation, $\hat{g}^f$, is proportional to $\mu\lambda\Lambda^2$ being constant for a given structure; with growing value of Λ the seismic moment becomes greater.

Neglecting the term related to the formation entropy, we can write for entropy density change:

$$\Delta\tilde{S} = kN\left(\frac{\Lambda}{\lambda}\right)\left(1 + \frac{C\mu b\lambda^2}{kT}\right)\exp\left(-\frac{C\mu\lambda\Lambda^2}{kT}\right) .$$

All of these relations concern the quantities referred to the multiple of the cubic volume $N\Lambda^3$ thus, we can correct these quantities to that related to a given source volume by introducing the factor $\pi R^2 D/N\Lambda^3$:

$$\begin{aligned}M_0 &= \mu\pi R^2 D\left(\frac{\Lambda}{\lambda}\right)\exp\left(-\frac{C\mu\lambda\Lambda^2}{kT}\right),\\ \Delta E &= \pi R^2 D\frac{kT}{\Lambda^2\lambda}\exp\left(-\frac{C\mu\lambda\Lambda^2}{kT}\right) \text{ and}\\ \frac{M_0}{E^{\mathrm{rad}}} &= \frac{\mu\Lambda^3}{\eta kT},\\ \Delta\tilde{S} &= \pi R^2 D\frac{k}{\Lambda^2\lambda}\left(1 + \frac{C\mu\lambda\Lambda^2}{kT}\right)\exp\left(-\frac{C\mu\lambda\Lambda^2}{kT}\right),\end{aligned} \tag{40}$$

where $\eta\Delta E = E^{\mathrm{rad}}$, η is the seismic efficiency; E^{rad} is radiated energy.

In the above consideration we took into account both energies related to slip and friction processes, and thus, the total released energy includes that related to stress drop and that related to heat caused by friction processes.

Further on, we will consider the fracturing processes in an earthquake source.

Synchronization and Fracturing

The inner stress accumulation relates to the formation of defect densities; due to the interaction between dislocations, we arrive at stress concentration at the first blocking dislocations of the formed dislocation arrays.

In the compression case with no initial shear field and due to the lower value of shear resistance, we have to assume that inside a body there appear regions with induced shear stresses of opposite signs, and induced antisymmetric stresses. The earthquake process and its energy release relate to a coalescence of dislocation arrays of opposite signs and related rotation release motion. Of course, we shall consider a fracture process as a chain of events; let us consider the micro-fracture centers formed on two perpendicular plane fragments; the induced shear stresses will be opposite on those plane fragments, but will have the common orientation of a spin motion – see Fig. 8; hence, the shears will be almost compensated for, while the spin field will remain unchanged.

The spin field, $\omega_{[s]}$, propagates and influences the processes in the adjacent regions; we believe that this propagation synchronizes the spin motions in the adjacent centers, in such a way that the sense of spin motion becomes the same over the whole fracture region. That means, the spin propagation assures a synchronization of fracture processes, especially under compression load where the energy release relates to the fragmentation revealed by rotation and granulation processes.

Reversely, under the shear load, while common shear deformation, E^D_{kn} (or expressed as twist $\omega_{(s)}$), progresses, the spins on the main fracture differ from those on the adjacent perpendicular fragments and attenuate fracture progress on those fragments.

Accordingly, we can believe that at the compression load, the total shear stress drop will be relatively small, while the rebound rotations will release an important amount of rotation energy. At the shear load the release of shear stresses will prevail.

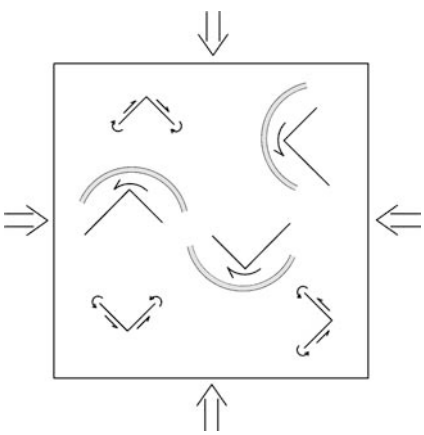

Earthquake Source: Asymmetry and Rotation Effects, Figure 8
Compression load: induced shear centers and formation of fragments with related rotations

Concluding, the rotation processes in fragmentation and fracturing under compression load play an essential role. Under prevailing shear load the rebound process releases shear load with the regional stress drop, while the rotation processes play a minor role.

Further on, we will discuss the importance of the granulation processes related to rotations in meso-scale, which we can place between the bond breaking processes in the micro-scale and material fragmentation in the macro-scale.

Granulation and Formation of Mylonite Zones

The fracturing process, especially under the action of shearing load, is accompanied by material granulation adjacent to shear fracture planes; thus it becomes spectacular at the formation of narrow, long mylonite zones. In this process we shall take into account a special role of rotations – the meso-rotations of different scales; these rotations are related to bond breaking and friction processes.

Co-action of the spin and twist-shear motions in bond breaking, granulation and formation of mylonite material can effectively help us to explain the fracture process; the simultaneous formation of the adjacent mylonite zone appears due to such a co-action of spin and shears and of the fracture transport phenomena.

Based on the standard asymmetric continuum theory, as presented in the first part, we would like to consider the material undergoing a progressive crashing process. We may arrive even at the conditions more similar to fluid material, and thus finally shall enter into our considerations the Navier–Stokes transport equations.

Starting with the description of the rock continuum following from the standard asymmetric theory of continuum $(S_{ik} = S_{(ik)} + S_{[ik]},\ E_{ik} = E_{(ik)},\ \omega_{ik} = \omega_{[ik]})$, we approach the final stage of the crushing/granulation process in zones adjacent to fracture planes. In these zones, simultaneously with dynamic processes, there occur changes of material properties from hard rocks to mylonite granulated material.

Approaching the final stage, the stresses, strains and rotations presented in the description of the standard asymmetric continuum become gradually neglected and progressively replaced by the constitutive relations for time-rates of stresses and strains.

The constitutive laws for rock asymmetric continuum – the relations (1) written for the deviatoric fields and for the antisymmetric fields

$$S^D_{(kl)} = 2\mu E^D_{kl}\,, \quad S_{[kl]} = 2\mu\omega_{kl} \tag{41}$$

– will gradually change during fracturing to those including the time dependent processes:

$$\begin{aligned} \sigma S^D_{(ik)} + \tau \dot{S}^D_{(ik)} &= 2\mu E^D_{ik} + 2\eta \dot{E}^D_{ik} \,, \\ \sigma S_{[ik]} + \tau \dot{S}_{[ik]} &= 2\mu E_{ik} + 2\eta \dot{\omega}_{ik} \,. \end{aligned} \tag{42}$$

The introduced material constants are related to magnitudes of the slip, u, and slip rate, υ.

When in a narrow zone the huge shear stresses break the molecular bonds, the stresses crushing rock material immediately drop down to the low values and in the crushed mylonite material we observe the immediate increase of the stress and strain rates to such degree that the stresses and strains may be neglected in the respective constitutive relations for that narrow zone. Finally, these changes will lead to the constitutive laws for the melt and granulated parts of mylonite material in which, practically, will remain only the field time rates:

$$\dot{S}_{(ik)} = 2\eta \dot{E}_{ik} \,, \quad \dot{S}_{[ik]} = 2\eta \dot{\omega}_{ik} \,. \tag{43}$$

The direct observation of the gauge zone of the Kobe (Japan, 1995) earthquake at the Avaji island suggests that the size of an inner completely melted part of the mylonite zone ranges around couple of centimeters (private communication W. Debski).

Further on, we will assume for the sake of simplicity, that during the fracturing the mylonite material remains incompressible.

In such a way, the nucleation progresses and fracture propagates simultaneously with the granulation process in the intact material (or in the compact zone previously crushed) – see Fig. 9.

In this new description, the shear rates create the dynamic angular deformations then lead to the bond breaking processes, and finally to the fracturing transport process.

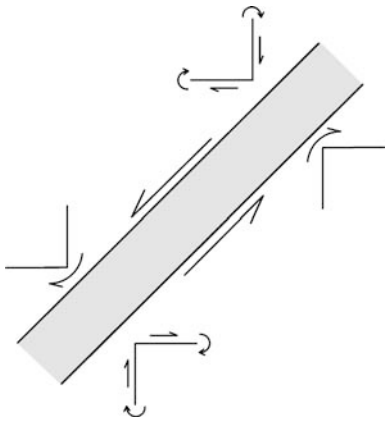

Earthquake Source: Asymmetry and Rotation Effects, Figure 9
Mylonite zone and neighboring deformations

We pass to the final stage; for the crushed incompressible mylonite or sand, similarly to incompressible fluids, where tensor $\dot{\omega}^{\mathrm{T}}_{ik}$ is related to spin motion (not to rotation of displacement), the mylonite viscosity is η, and the mylonite relaxation time is denoted by τ.

The relations (42) define the ideal quasi-viscous mylonite for an incompressible crushed material.

Further, we assume the coincidence/identity of rotation of velocity field $\dot{\mathbf{u}}$ with the point rotation field ω. So, we assume that rotations of particles (micro-rotation ω) coincide with macro-rotations (rot $\mathbf{u}$). For mylonite, such a coincidence between the micro-rotation and macro-rotation seems reasonable, and thus, our assumption that viscosity η coincides with rotation viscosity η^* may be correct.

For our narrow mylonite zone, existing already near the pre-slip planes or just simultaneously formed, we may, further on, apply the Navier–Stokes transport equation. Referring to our former considerations on the asymmetric continuum theory (see Subsect. "Spin and Twist Motions") we may add to the spin rotational motions the oscillations of the strain shear rates (called twist motion). Such motions, especially in an earthquake source zone, are due to the friction processes.

We may note that when including these complex rotational motions in the theory we may replace the friction constitutive laws, as based on the experimental data, by the constitutive law joining the asymmetric stresses with spin and shears field oscillations or otherwise with spin and twist.

Slip Propagation and Spin Release Hypothesis

While searching the fault slip solutions, we use the classical elasticity tools with an additional friction constitutive law based on experimental data. When instead of it we consider the asymmetric elastic continuum, we are able to include the defect interaction and we can derive the elastodynamic fault solution describing slip propagation with fracturing process and related seismic radiation.

The angular deformations preceding the bond breaking process lead to the efficient rise of the angular moments around material grains. In the narrow mylonite zone, we arrive at the equivalence between this expression and the laws introduced in the considerations on the friction resistance and slip.

The co-action of the rotation (rot $\mathbf{u}$, or spin $\omega_{[.]}$) and shear ($\mathbf{E}^D(\mathbf{u})$, or twist – $\omega_{(.)}$) motions can lead further to the slip fracturing motion. We assume that the bond

breaking process and granulation of material precede the slip movement: just after the bond breaking micro-process there, we would have the released rebound spin motion retarded in phase.

This hypothesis is supported by the following solution of the homogeneous wave equations for the twist and spin in a mylonite zone (see Eq. (31)):

$$\begin{aligned} &\omega_{(s)} = \mathrm{i}\omega_{[s]}\,; \quad \omega_{[s]} = \omega^0_{[s]} \exp[\mathrm{i}(k_i x_i - \omega t)\,, \\ &\omega_{(s)} = \omega^0_{(s)} \exp[\mathrm{i}(k_i x_i - \omega t)]\,, \end{aligned} \tag{44}$$

where the six constants in $\omega^0_{[s]} = \mathrm{abs}(\omega^0_{[s]}) \exp(\mathrm{i}\psi_s)$, $\omega^0_{(s)} = \mathrm{abs}(\omega^0_{(s)}) \exp(\mathrm{i}\varphi_s)$ shall fulfill the six conditions.

We may consider the following 2D solution of Eq. (44) in the systems $\{r, \varphi, z\}$:

$$\omega_{(\varphi)}(r) = \mathrm{i}\omega_{[\varphi]}(r)\,, \quad \Box\omega_{[\varphi]}(r) = 0\,.$$

The related solution corresponds to a turbulence structure. Thus, from a dislocation-slip structure formed in the earthquake premonitory domain (see: Subsect "Earthquake Thermodynamics"), gradually destroyed during a fracture process by a spin release motion, we can arrive at a turbulence structure appearing in a melted or fully granulated material.

With the introduced waves, $\omega_{(s)} = \mathrm{i}\omega_{[s]}$ we arrive at the possibility to study the dynamic defect objects and to explain the synchronization of the micro-fracturing processes due to an influence of the propagating waves. For the fracture processes under compression such a synchronization will assure the common sense of the induced twist and spin motions, while under shear load – the formation of a long shearing fracturing. In the last case, the spin waves related to a given slip on the main fracture plane attenuate those with the opposite spins generated at the perpendicular fragments, and due to the conjugate solution (see Eq. (44) reduce the slip motions on those fragments.

The presented conjugate solution Eq. (44) suggests that the spin rebound motion is delayed in phase by $\pi/2$ (as we have $\exp[\mathrm{i}(k_i x_i - \omega t)] = \exp[\mathrm{i}k_i x_i - \mathrm{i}(\omega t - \pi/2)]$); when slip starts due to breaking of bonds, the micro-spin motions are released.

Following this assumption we expect that such a correlation between the recorded twist motions and spin motions shifted by $\pi/2$ in phase can exist in some wavelets.

Now we can propose the following description of the fracture process.

- First, according to external load conditions the stresses rise while the disclosure and dislocation field can be neglected – the case given by relations (25).
- Next, approaching the fracture process we may observe the "accumulation" phase with the co-action of the twist and spin – the case given by relations (26).
- Finally, fracturing processes start and when entering into the time rate domain we can describe the "release" phase of the process as follows (compare: Eqs. (21–26 and 43):

$$\dot{B}_l = \oint [E_{(kl)} + \mathrm{i}\omega_{[kl]}]\mathrm{d}l_k\,, \tag{45}$$

$$\begin{aligned} \dot{\alpha}_{pl} &= \varepsilon_{pmk}\left(\frac{\partial \dot{E}_{kl}}{\partial x_m} + \mathrm{i}\frac{\partial \omega_{kl}}{\partial x_m}\right) \\ \dot{\alpha}_{pl} &= \frac{\varepsilon_{pmk}}{2\mu}\frac{\partial}{\partial x_m}\left((\dot{S}_{(kl)} - \frac{\nu}{1+\nu}\delta_{kl}\dot{S}_{ii}) + \mathrm{i}\dot{S}_{[kl]}\right)\,, \end{aligned} \tag{46}$$

  where we have the dynamic disclosure and v-dislocation density under the conditions formed by solution (see Eq. (44)), supplemented with the relations between the asymmetric stress rates and the dynamic dislocation objects (ν-dislocations).
- This case presents a formation of dynamic discontinuities and the related dynamic processes in which the slip and bond breaking leads to the rebound spin motions delayed in phase by $\pi/2$.

The co-action of the spin and twist motions leads to the "accumulation" phase, while the conjugate solution (see Eq. (44)) presents a fracture process – "release" phase.

We might suppose that the fracture process could proceed with the consecutive accumulation and release micro-processes; in such a situation the related twist and spin motions will appear consecutively as pairs of wavelets in phase (or anti-phase) and those differing in phase by $\pi/2$. Figure 10 presents an example of the coincidence of the spin and twist motions after the Hilbert transformation shifting the angular twist record ahead in phase by $\pi/2$.

The presented theory, due to its simplicity, could be very useful for some problems, among others those in which macro-rotation takes an important role in the asymmetric fluid dynamics.

We can call the solution (44) as the fracture synchronization waves.

Finally, we shall notice that similar solutions may exist for the electric and magnetic induction vectors:

$$D_s = \mathrm{i}B_s \leftrightarrow D^0_s = \mathrm{i}B^0_{[s]}\,, \tag{47}$$

where this relation shall be assured by the appropriate material constants.

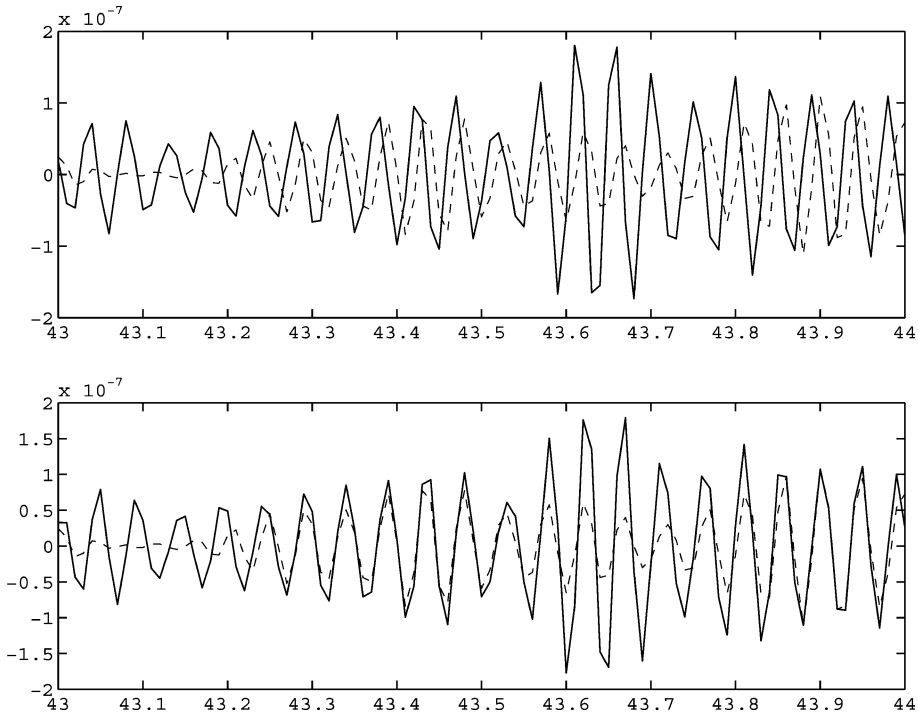

Earthquake Source: Asymmetry and Rotation Effects, Figure 10

Example of coincidence of the spin and twist angular motions (rad/s versus s) after the Hilbert transformation shifting the twist record ahead in phase by $\pi/2$; *upper part* – the original records, *lower part* – the twist record transformed (from the original seismic record obtained by the system of the rotation seismometers; L'Aquila Observatory, 17.02.2006; the *continuous line* – twist, the *broken line* – spin)

Towards Asymmetric Fluid Theory and Extreme Phenomena

Approaching the conclusions, we shall specify how we could formulate the asymmetric theory of the fluid continuum in which the stress, strain and rotation fields vanish, but their rates exist as related to the velocity field υ:

$$\begin{aligned} \dot{E}_{ik} &= e^0 \frac{1}{2}\left(\frac{\partial \upsilon_k}{\partial x_i} + \frac{\partial \upsilon_i}{\partial x_k}\right) , \\ \dot{\omega}_{ik} &= \chi^0 \frac{1}{2}\left(\frac{\partial \upsilon_k}{\partial x_i} - \frac{\partial \upsilon_i}{\partial x_k}\right) . \end{aligned} \tag{48}$$

The velocity field υ shall obey the Navier–Stokes **transport equation**.

For the sake of simplicity, let us consider the incompressible fluid ($\dot{E}_{ss} = 0$, $\dot{E}_{ik} = \dot{E}^D_{ik}$), the basic constitutive laws are assumed similarly to those for the asymmetric continuum:

$$\dot{S}_{(ik)} = \eta \dot{E}_{ik} , \quad \dot{S}_{[ik]} = \eta \dot{\omega}_{ik} , \tag{49}$$

where η is viscosity.

In a similar manner, the structural dynamic objects can then be defined as defects in the standard asymmetric continuum.

The extreme motion phenomena, related to the shear rate and spin, $\dot{E}_{(kl)} + \dot{\omega}_{(kl)}$, could be expected for the following case (compare: Eqs. (22–26 and 43):

$$\begin{aligned} \dot{B}_l &= \oint [\dot{E}_{kl} + \dot{\omega}_{kl}] \mathrm{d}l_k = \oint \frac{\partial \upsilon_l}{\partial x_k} \mathrm{d}l_k = 0 , \\ \dot{\alpha}_{pl} &= 0 , \quad \dot{\Omega}_q = 0 , \end{aligned} \tag{50}$$

while formation of the dynamic defect objects can be described as:

$$\begin{aligned} \dot{B}_l &= \oint [\dot{E}_{kl} - \dot{\omega}_{kl}] \mathrm{d}l_k = , \\ \dot{\alpha}_{pl} &= \frac{1}{2\mu} \varepsilon_{pmk} \frac{\partial}{\partial x_m} [\dot{S}_{(kl)} - \dot{S}_{[kl]}] . \end{aligned} \tag{51}$$

The former case (see Eq. (50)) presents an extreme shear rate deformation, like soliton waves, while this case (see Eq. (51)) would relate to a formation of the ν-slip-discontinuity.

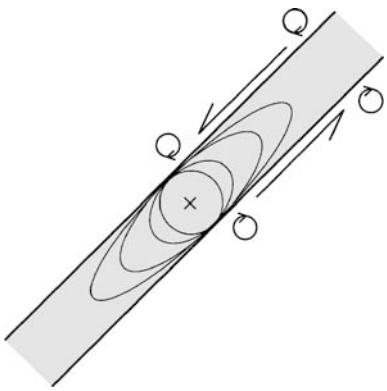

Earthquake Source: Asymmetry and Rotation Effects, Figure 11
Extreme motions: soliton wave will be related to a given deformation of the circle as follows for the parameter $C = e^0$; Amplitude plot (Mathematica 5.0): with *AspectRatio* → *Automatic*: $A = \text{Plot}[y = -Cx/2 \pm 0.5\sqrt{[(Cx) \wedge 2 - 4x \wedge 2 + 4]}],$ $\{x, -\sqrt{[4/(4 - C \wedge 2)]}, \sqrt{[4/(4 - C \wedge 2)]}\}$

In 2D we can show an effect of the co-action of the macro spin and twist motions in the following way; let us put $C = e^0$ and $\chi^0 = 1$

$$\dot{E}_{12} + \dot{\omega}_{12} = C\upsilon_{(2,1)} + \upsilon_{(2,1)} = C\frac{1}{2}\left(\frac{\partial \upsilon_k}{\partial x_i} + \frac{\partial \upsilon_i}{\partial x_k}\right) + \frac{1}{2}\left(\frac{\partial \upsilon_k}{\partial x_i} - \frac{\partial \upsilon_i}{\partial x_k}\right) .$$

For $C = 0$ we would have only the spin motion, while a full coincidence will occur at $C = 1$.

The effect of such a superposition of the spin and twist motions is presented on Fig. 11.

We believe that this approach might explain some extreme fluid phenomena related to atmosphere and oceans.

The balance equations for field rates (48) at $e^0 = \chi^0$ may lead us to the wave equations for the related spin and twist rate fields, $\dot{\omega}_{[s]}$ and $\dot{\omega}_{(k)}$, as defined similarly to the relations derived in Subsect. "Balance Laws for the Rotation Field and the EM Analogy" (see: Eqs. (29 and 31):

$$\begin{aligned} \varepsilon_{kps}\frac{\partial}{\partial x_p}\dot{\omega}_{[s]} - \frac{1}{V}\frac{\partial}{\partial t}\dot{\omega}_{(k)} &= \frac{4\pi}{V}\dot{J}_k , \\ \varepsilon_{kps}\frac{\partial}{\partial x_p}\dot{\omega}_{(s)} + \frac{1}{V}\frac{\partial}{\partial t}\dot{\omega}_{[k]} &= 0 . \end{aligned} \tag{52}$$

The appearance of such coupled waves transversal to the transport motion brings physical background for diffraction in fluids as explained usually by the Huygens principle.

Conclusions

Under both the confining pressure and external shear, the role of micro-fracturing in the bond breaking process is similar; however, we observe the essential differences for rotations in larger scales.

The confining condition leads to formation of induced opposite arrays of dislocations, resulting in fragmentation processes and chaotically oriented macro-rotations, leading therefore to a rotation release process.

The shear condition leads to more concentrated fracturing along some planes, high shear strain release and correlated rotations.

Both cases include formation of narrow mylonite zones adjacent to the fracturing planes or their fragments, but these processes prevail rather under shear conditions.

We draw attention to the importance of the rotations in meso-scales – between the micro-scale bond breaking process and that related to macro-rotation at material fragmentation. The meso-scale rotations are related to material granulation and become observed in any fracturing process; such motions may be revealed in the spectacular formation of the narrow, long mylonite zones under shear load conditions. Coincidence and co-action of the spin and twist-shear motions in bond breaking and formation of mylonite material help one to understand the fracture motion; the simultaneous formation of the mylonite zones appears due to common action of these motions and to fracture transport phenomena.

Rotations at source zones help to understand geometry of fracturing and releases of stress and rotation counterparts as a result of precursory and rebound processes.

We have presented also a new idea how to construct the asymmetric fluid theory with the asymmetric stress rate field.

Final Remarks

With the additional constitutive law joining the rotations with antisymmetric part of stresses, we have proved that the rotation waves exist, even in a homogeneous elastic continuum. We have defined the twist motion as the rotational oscillation of the main shear axes including the shear magnitude variations. The derived wave equations for the twist and spin motions have been considered in relation to the processes in seismic source.

Only in the presented approach with the standard asymmetric theory may we study the co-action of the independent motions and deformations; this is due to the new relations joining the spin and slip motions. A possible phase shift between these motions leads to different families of deformation and related solutions.

We have shown how the rotations at source zones help us to understand physics and geometry of fracturing

and release of stresses in the precursory and rebound processes.

The derived wave equations for spin and twist motions are similar to the EM wave equations.

Our considerations show the importance of the simultaneous recording of the translational and rotational earthquake motions, and also the strains (at least the deviatoric strains).

Finally, we have shown how to construct the asymmetric fluid theory in which, with a help of the asymmetric stress rates, the various extreme phenomena, including soliton waves, can be theoretically explained.

Acknowledgments

Work done under the support of the INTAS Project 05-1000008-7889.

Bibliography

Primary Literature

1. Shimbo M (1975) A geometrical formulation of asymmetric features in plasticity. Bull Fac Eng Hokkaido Univ 77:155–159
2. Shimbo M (1995) Non-Riemannian geometrical approach to deformation and friction. In: Teisseyre R (ed) Theory of earthquake premonitory and fracture processes. PWN, Polish Scientific Publishers, Warszawa, pp 520–528
3. Teisseyre R, Boratyński W (2003) Continua with self-rotation nuclei: evolution of asymmetric fields. Mech Res Commun 30:235–240
4. Teisseyre R, Boratyński W (2006) Deviations from symmetry and elasticity: Asymmetric continuum mechanics. In: Teisseyre R, Takeo M, Majewski E (eds) Earthquake source asymmetry, structural media and rotation effects. Springer, Berlin, pp 31–42
5. Kozak JT (2006) Development of earthquake rotational effect study. In: Teisseyre R, Takeo M, Majewski E (eds) Earthquake source asymmetry, structural media and rotation effects. Springer, Berlin, pp 3–10
6. Ferrari G (2006) Note on the historical rotation seismographs. In: Teisseyre R, Takeo M, Majewski E (eds) Earthquake source asymmetry, structural media and rotation effects. Springer, Berlin, pp 367–376
7. Gutenberg B (1926) Grundlagender Erdbebenkunde. Univ. Frankfurt a/M, Franfurt
8. Teisseyre R (1973) Earthquake processes in a micromorphic continuum. Pure Appl Geophys 102:15–28
9. Cochard A, Igel H, Schuberth B, Suryanto W, Velikoseltsev A, Schreiber U, Wassermann J, Scherbaum F, Vollmer D (2006) Rotational motions in seismology: Theory, observation, simulation. In: Teisseyre R, Takeo M, Majewski E (eds) Earthquake source asymmetry, structural media and rotation effects. Springer, Berlin, pp 391–411
10. Schreiber KU, Stedman GE, Igel H, Flaws A (2006) Ring laser gyroscopes as rotation sensors for seismic wave studies. In: Teisseyre et al (eds) Earthquake source asymmetry, structural media and rotation effects. Springer, Berlin, pp 377–389
11. Takeo M (2006) Rotational motions excited by earthquakes. In: Teisseyre R, Takeo M, Majewski E (eds) Earthquake source asymmetry, structural media and rotation effects. Springer, Berlin, pp 131–156
12. Leszek R, Jaroszewicz LR, Krajewski Z, Solarz L (2006) Absolute rotation measurement based on the Sagnac effect. In: Teisseyre R, Takeo M, Majewski E (eds) Earthquake source asymmetry, structural media and rotation effects. Springer, Berlin pp 413–438
13. Moriya T, Teisseyre R (2006) Design of rotation seismometer and non-linear behaviour of rotation components of earthquakes. In: Teisseyre R, Takeo M, Majewski E (eds) Earthquake source asymmetry, structural media and rotation effects. Springer, Berlin, pp 439–450
14. Wiszniowski J (2006) Rotation and twist motion recording – couple pendulum and rigid seismometers system. In: Teisseyre R, Takeo M, Majewski E (eds) Earthquake source asymmetry, structural media and rotation effects. Springer, Berlin, pp 451–470
15. Teisseyre R, Suchcicki J, Teisseyre KP, Wiszniowski J, Palangio P (2003) Seismic rotation waves: basic elements of the theory and recordings. Ann Geophys 46:671–685
16. Teisseyre R, Białecki M, Górski M (2006) Degenerated asymmetric continuum theory. In: Teisseyre R, Takeo M, Majewski E (eds) Earthquake source asymmetry, structural media and rotation effects. Springer, Berlin, pp 43–56
17. Kröner E (1981) Continuum theory of defects. In: Balian R, Kleman M, Poirer JP (eds) Physique des defauts/physics of defects (Les Houches, Session XXXV, (1980)). North Holland Publ Com, Dordrecht
18. Teisseyre R, Górski M, Teisseyre KP (2006) Fracture–band geometry and rotation energy release. In: Teisseyre R, Takeo M, Majewski E (eds) Earthquake source asymmetry, structural media and rotation effects. Springer, Berlin, pp 169–184
19. Nowacki W (1986) Theory of asymmetric elasticity. PWN, Warszawa and Pergamon Press, Oxford, New York, Toronto, Sydney, Paris, Frankfurt, p 383
20. Cosserat E, Cosserat F (1909) Theorie des Corps Déformables. A. Hermann, Paris
21. Eringen AC, Suhubi ES (1964) Non-linear theory of simple micro-elastic solids. I Int J Eng Sci 2:189–203
22. Mindlin RD (1965) On the equations of elastic materials with microstructure. Int J Solids Struct 1:73
23. Teisseyre R (2005) Asymmetric continuum mechanics: Deviations from elasticity and symmetry. Acta Geophys Polon 53:115–126
24. Teisseyre R, Kozak JT (2003) Considerations on the seismic rotation effects. Acta Geophys Polon 51:243–256
25. Kossecka E, DeWitt R (1977) Disclination kinematic. Arch Mech 29:633–651
26. Teisseyre R (2001) Evolution, propagation and diffusion of dislocation fields. In: Teisseyre R, Majewski E (eds) Earthquake thermodynamics and phase transformations in the earth's interior. Academic Press (Vol. 76 of International Geophysical Series), San Diego,San Francisco, NewYork, Boston, London, Sydney, Tokyo, pp 167–198
27. Kleman M (1980) The general theory of disclinations. In: Nabarro FRN (ed) Dislocations of solids, vol 5. Other effects of dislocations: Disclinations. North-Holland Publ. Comp., Amsterdam, pp 243–297

28. Nabarro FRN (1967) Theory of crystal dislocations. Clarendon Press, Oxford
29. Moriya T, Marumo T (1998) Design for rotation seismometers and their calibration. Geophys Bull Hokkaido Univ 61:99–106
30. Teisseyre R (1996) Shear band thermodynamical earthquake model. Acta Geophys Polon 44:219–236
31. Dietrich JHJ (1978) Preseismic fault slip and earthquakes prediction. J Geophys Res 83(B8):3940–3954
32. Teisseyre R, Majewski E (1990) Thermodynamics of line defects and earthquake processes. Acta Geophys Polon 38:355–373
33. Teisseyre R, Majewski E (1995) Earthquake thermodynamics. In: Teisseyre R (ed) Theory of earthquake premonitory and fracture processes. PWN, Warszawa, pp 586–590
34. Teisseyre R, Majewski E (2001) Thermodynamics of line defects and earthquake thermodynamics. In: Teisseyre R, Majewski E (eds) Earthquake thermodynamics and phase transformations in the earth's interior. Academic Press (Vol. 76 of International Geophysical Series), San Diego, San Francisco, NewYork, Boston, London, Sydney, Tokyo, pp 261–278
35. Majewski E, Teisseyre R (1997) Earthquake thermodynamics. Tectonophysics 227:219–233
36. Prigogine I (1979) Irreversibility and randomness. Astron Phys Space Sci 65:371–381
37. Kocks UF, Argon AS, Ashby MF (1975) Thermodynamics and kinetics of slip. Pergamon Press, Oxford, New York, p 288
38. Varotsos PA, Alexopoulos KD (1986) Thermodynamics of point defects and their relation with bulk properties. North–Holland, Amsterdam, New York, p 474

Books and Reviews

Bridgman W (1950) The thermodynamics of plastic deformation and generalized entropy. Rev Mod Phys 22:56–63
Eringen AC (1999) Microcontinuum field theories I: Foundations and solids. Springer, Berlin, p 325
Eringen AC (2001) Microcontinuum field theories II: Fluent media. Springer, Berlin, p 340
Drazin PG (1983) Solitons. Cambridge University Press, Cambridge
Infeld E, Rowlands G (2000) Nonlinear waves, solitons and chaos. Cambridge University Press, Cambridge
Muskhelishvili NT (1953) Some basic problems of the elasticity. Noordhoff, Groningen
Newell A (1985) Solitons in mathematics and physics. Society for Industrial and Applied Mathematics, Philadelphia
Prigogine I (1978) Thermodynamics of irreversible processes, 3rd edn. Wiley, New York
Teisseyre R (1974) Symmetric micromorphic continuum: wave propagation, point source solutions and some applications to earthquake processes. In: Thoft-Christensen P (ed) Continuum mechanics aspects of geodynamics and rock fracture mechanics. D. Reidel Publ. Comp., Dordrecht-Holland/Boston, pp 201–244
Teisseyre R (ed) (1995) Theory of earthquake premonitory and fracture processes. PWN, Warszawa, p 648
Teisseyre R, Czechowski L, Leliwa-Kopystynski J (eds) (1993) Dynamics of the earth's interior. PWN, Warszawa, p 469
Teisseyre R, Majewski E (eds) (2001) Earthquake thermodynamics and phase transformations in the earth's interior. Academic Press (Vol. 76 of International Geophysical Series), San Diego, NewYork, Boston, London, Sydney, Tokyo, p 670
Teisseyre R, Majewski E (2002) Physics of earthquakes. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International handbook of earthquake and engineering seismology, Part A. Academic Press, Amsterdam, pp 229–235
Teisseyre R, Takeo M, Majewski E (eds) (2006) Earthquake source asymmetry, structural media and rotation effects. Springer, Berlin, p 582
Thoft-Christensen P (ed) (1974) Continuum mechanics aspects of geodynamics and rock fracture mechanics, NATO Advance Study Institutes Series C, vol 12. D. Reidel Publ. Comp., Dordrecht-Holland/Boston, p 273

Earthquake Source Parameters, Rapid Estimates for Tsunami Warning

Barry Hirshorn, Stuart Weinstein
NOAA/NWS/Pacific Tsunami Warning Center, Ewa Beach, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Tsunami Warning Center Operations
Seismic Methods
Earthquake Source Parameters
Future Directions
Acknowledgments
Bibliography

Glossary

CMT centroid moment tensor The CMT represents the displacement of the Earth's crust that best reproduces the observed wave-field generated by an earthquake and gives the average location in time and space of the earthquake energy release. The seismic moment can be determined from the CMT.

Convolution Convolution is a type of integral transform combining two signals to form a third signal or output. It is the single most important technique in Digital Signal Processing. In the case of Seismology, the two signals can be e. g., the ground motion as a function of time and the response of the seismometer, and the output is the seismogram.

Deconvolution Does the reverse of convolution. In the case of Seismology, one uses deconvolution to remove the instrument response from the seismogram to recover the actual ground motion.

Deep earthquake An earthquake characterized by a hypocenter located more than 100 km below the Earth's surface.

Hypocenter The point within the Earth where the earthquake rupture starts. The epicenter is the projection of the hypocenter onto the Earth's surface.

Local tsunami A tsunami that has little effect beyond 100 km from its source.

Magnitude m_B: The "broad-band" body-wave magnitude, generally based on measurements of the amplitude of P-waves with periods in the 2 to 20 s range.

M_S: The surface-wave magnitude. M_S is generally based on measurements of the amplitude of the surface (Love or Rayleigh) waves with periods of about 20 s. The US tsunami warning centers have applied a correction to the IASPEI formula that allows the estimation of M_S closer to the epicenter at a period of 20 s.

M_E: The "energy magnitude" scale, derived from velocity power spectra.

M_m: The mantle wave magnitude, based on the measurement of the amplitude of surface waves with periods of 50–400 s.

M_W: The moment magnitude or the "work magnitude" is based on the estimation of the scalar seismic moment, M_0.

M_{wp}: The moment magnitude based on the initial long period P-waves.

M_L: The Local magnitude scale, based on the measurement of the maximum peak-to-peak amplitude observed on a Wood-Anderson seismogram, corrected for the decrease in amplitude with increasing epicentral distance. Generally based on the analysis of Sg, Lg or Rg surface waves oscillating with periods observed out to about 600 km. from the earthquake's epicenter.

pMag: A magnitude scale based on the average of the absolute values of the first three half cycles of the P-waves recorded at local distances.

Marogram A recording of sea-level variations obtained by tide gauges.

Regional tsunami A tsunami that has observable effects up to 1000 km from its source.

Seismic body waves Waves that propagate through the interior of an unbounded continuum. Primary waves (P-waves) are longitudinal body waves that shake the ground in a direction parallel with the direction of travel. Secondary body waves (S-waves) are shear waves that shake the ground in a direction perpendicular to the direction of travel. There are other types of arrivals (also known as phases) visible on seismographs corresponding to reflections of P- and S-waves from the earth's surface: The pP phase is a P-wave that travels upwards from the hypocenter and reflects once off the surface and the PP phase is a P-wave that travels downwards from the epicenter and reflects once off of the surface. The definitions of the S-wave phases follow in the same manner.

Seismic moment The seismic moment M_0, (expressed in units of force times distance; e. g. Newton-meters, or dyne-cm) is the moment of either couple of an equivalent double couple point source representation of the slip across the fault area during the earthquake. Mathematically, the Seismic Moment, $M_0 = \mu A d$, where μ denotes the shear rigidity, or resistance of the faulting material to shearing forces, A represents the area of the fault plane over which the slip occurs, and d represents the average co-seismic slip across A.

Seismic waves Elastic waves generated by movements of the earth's crust that propagate as radiated seismic energy, E_R.

Seismic surface waves Waves that propagate along the surface boundary of a medium, e. g. along the surface of the earth.

Shallow earthquake An earthquake characterized by a hypocenter located within 100 km of the Earth's surface.

Teletsunami A tsunami that has observable effects on coastlines more than 1000 km away from its source.

Tsunami A series of water waves generated by any rapid, large-scale disturbance of the sea. Most are generated by sea floor displacements from large undersea earthquakes, but they can also be caused by large submarine landslides, volcanic eruptions, calving of glaciers and even by meteorite impacts into the ocean.

Tsunami earthquake An earthquake that generates a much larger tsunami than expected given its seismic moment.

Tsunami warning system A tsunami warning system consists of a tsunami warning center such as the Pacific Tsunami Warning Center (PTWC), a formal response structure that includes Civil Defense authorities and Government Officials, and an education program that brings a minimum level of awareness and education to the coastal populations at risk.

Definition of the Subject

Tsunamis are among nature's most destructive natural hazards. Typically generated by large, underwater earthquakes near the Earth's surface, tsunamis can cross an ocean basin in a matter of hours. Although difficult to detect, and not dangerous while propagating in deep water, tsunamis can unleash awesome destructive power when

they reach coastal areas. With advance warning, populations dwelling in coastal areas can be alerted to move to higher ground and away from the coast saving many lives. Unfortunately, due to the lack of a tsunami warning system in the Indian Ocean, the Sumatra earthquake of Dec. 26, 2004 killed over 250 000 people with thousands of lives lost as far as away as East Africa many hours after the earthquake occurred. Had a tsunami warning system been in place many lives could have been saved [77].

As fast as tsunami waves are, seismic waves can travel at speeds more than 40 times greater. Because of this large disparity in speed, scientists rely on seismic methods to detect the possibility of tsunami generation and to warn coastal populations of an approaching tsunami well in advance of its arrival. The seismic P-wave for example, travels from Alaska to Hawaii in about 7 min, whereas a tsunami will take about 5½ h to travel the same distance. Although over 200 sea-level stations reporting in near-real time are operating in the Pacific it may take an hour or more, depending on the location of the epicenter, before the existence (or not) of an actual tsunami is confirmed. In other ocean basins where the density of sea-level instruments reporting data in near real-time is less, the delay in tsunami detection is correspondingly longer. In addition, global, regional, and local seismic networks, and the infrastructure needed to process the large amounts of seismic data that they record, are widespread. For these reasons, tsunami warning centers provide initial tsunami warnings to coastal populations based entirely on seismic data.

Introduction

A tsunami can be produced by any mechanism that causes a sudden displacement of the ocean's surface affecting a significant volume of water. Tsunamis can be generated by undersea earthquakes, landslides and volcanic explosions, calving of icebergs, and even meteorite impacts. However, the majority of tsunamis are generated by earthquakes. Not uncommon are earthquakes that trigger landslides so that both the displacement of the crust due to the earthquake, and the landslide, each contribute to the generation and size of the tsunami. Tsunamis are a devastating, natural, high fatality hazard [18]. In the absence of a proper tsunami warning system, a destructive tsunami will cause death and destruction as it encounters coastal areas while propagating across an ocean basin as it did in the case of the Sumatra tsunami of December 2004.

Although tsunamis propagate in deep water with speeds exceeding 900 km/h they are hard to detect in the open ocean. For instance, the first wave of the great Sumatra tsunami had a wave height of only one meter in deep water (> 500 m) [28], and a wavelength on the order of several hundred kilometers. Consequently, people aboard ocean vessels did not feel the accelerations caused by the Sumatra tsunami as it passed under them. However, as the speed, v, of a tsunami is governed by the simple relation $v = \sqrt{gh}$ where g is the acceleration of gravity (in m/s), and h is the thickness of the water column (in m), the tsunami will slow down as it propagates into shallow waters. At this point, the wave speed and wavelength decrease, causing the wave height to increase. Depending on the nature of the tsunami, and the shape and bathymetry of the coastal area, the tsunami wave height can be greatly amplified, thus magnifying its destructive power.

Because most tsunamis are generated by earthquakes, and seismic waves travel more than 40 times faster than tsunamis, the first indication that a tsunami may have been generated is the earthquake itself. Depending on the earthquake's location (undersea or inland), depth (shallow or deep) in the Earth's crust, and magnitude, a warning center may be required to issue an official message product. If the earthquake is a shallow, under sea earthquake, the severity of the message will depend upon the magnitude of the earthquake. The more rapidly and accurately the tsunami warning center can characterize the earthquake source, the quicker the initial evaluation of the tsunamigenic potential of the earthquake can be disseminated.

While some tsunamis are destructive, most are rather small, producing few if any casualties and little or no damage, although they are easily observable on marograms (Fig. 1). On the basis of how widespread their effects are, tsunamis can be classified as local (within 100 km of the epicenter), regional (up to 1000 km from the epicenter) or teletsunamis (greater than 1000 km from the epicenter). In the Pacific Basin there are warning centers designed to respond to tsunami threats on each of these scales.

The Richard H. Hagemeyer Pacific Tsunami Warning Center (PTWC) provides basin-wide warnings to the coastal areas of the Pacific basin. PTWC also functions as a local tsunami warning center for the Hawaii region. Other local warning centers include CPPT (French Polynesia Tsunami Warning Center) which is based in Tahiti, GFZ-Indonesia which is based in Jakarta provides local warnings for Indonesia, and Japan's JMA (Japan Meteorological Agency) which operates Japan's tsunami warning centers provides local tsunami warnings to Japan. Examples of regional warning centers include JMA which provides regional tsunami warnings to the Northwest Pacific and the West Coast and Alaska Tsunami Warning Center (WC/ATWC) which provides regional and local warnings to the US mainland coasts and the west coast of Canada.

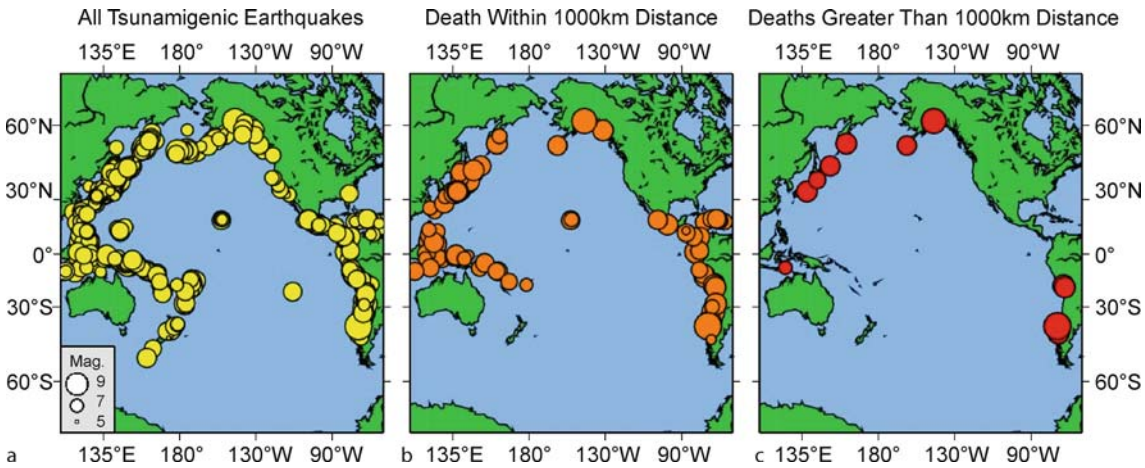

Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 1

Epicenters of tsunamigenic earthquakes occurring in the Pacific since 1 A.D. Of those earthquakes that do produce a tsunami a, most tsunamis cause no damage. Most events that cause casualties and/or damage do so within 1000 km of the epicenter b, leaving only a few great earthquake sources that generated tsunamis which caused casualties and/or damage more than a 1000 km from the epicenter, c. Data provided by the NOAA National Geophysical Data Center (NGDC), (www.ngdc.noaa.gov/hazard/tsu.html)

The tsunami warning centers themselves are not a complete tsunami warning system, they are simply the first of line of defense within the warning system. The warning system consists of three main components a) the tsunami warning centers, b) emergency management/civil defense authorities who receive tsunami warning center message products and c) a public in coastal areas that is educated in how to respond to tsunami emergencies. If any of these three components are lacking, the tsunami warning system can fail. Unfortunately, none of these components existed in the Indian Ocean at the time of the December 2004 Sumatra earthquake.

The greatest challenge for a tsunami warning system, particularly in the near field, is the slow (in terms of rupture speed) or "tsunami" earthquake. Tsunami earthquakes are so-called because they generate much larger than expected tsunamis given the size of the seismic moment of the earthquake [47]. In a well functioning tsunami warning system, residents in coastal areas are educated to immediately move inland and onto higher ground if they feel strong ground shaking and not wait for an official tsunami alert [24]. However, because a tsunami earthquake produces much less radiated high frequency body-wave energy than normal, even a large (in terms of moment magnitude) tsunami earthquake may not be strongly felt in the near field so that this strategy of having people self-evacuate upon feeling strong ground shaking will not work. This was, unfortunately, made dramatically clear by the Java earthquake of July 17, 2006. The tsunami generated by the Java earthquake killed $\sim$ 500 people as many residents in coastal areas near the earthquake did not feel strong shaking [83]. Tsunami warning centers need to be able to properly detect the occurrence of these tsunami earthquakes.

Tsunami Warning Center Operations

Tsunami warning center functions are much like those of a seismic observatory, i. e.: detecting, locating and characterizing the source of major earthquakes occurring around the world as fast as possible. Depending on the earthquake's location (underwater vs. inland), depth below the surface, and magnitude, tsunami warning centers may issue an official message product to advise Civil Defense/Emergency Management authorities within the warning centers AOR (area of responsibility), of the occurrence of a large earthquake and its potential for generating a tsunami. The PTWC, located in Ewa Beach, Hawaii, provides advance warning of the generation of a destructive tsunami for the Pacific Ocean Basin, and on an interim basis, for the Indian, and Caribbean ocean basins.

After consultation with the member states of the Pacific Tsunami Warning System (PTWS), the PTWC has agreed to issue tsunami bulletins for the Pacific Ocean Basin according to the criteria shown in Fig. 2.

An Observatory (Earthquake) Message: The PTWC sends observatory messages to certain seismological observatories and organizations for any earthquake in or near the vicinity of the Pacific, Indian, or Caribbean ocean basins for seismic events when the magnitude is larger

Mw less than 6.5 (Mw: Moment Magnitude)	Earthquake Message Only
Mw 6.5 to 7.5	Tsunami Information Bulletin
Mw 7.6 to 7.8	Regional Tsunami Warning
Mw > 7.8	Expanding Warning / Watch
Confirmed Teletsunami	Pacific-Wide Warning

Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 2
PTWC Bulletin Criteria for the Pacific Basin

than about 5.5. This unofficial message contains only the earthquake's epicentral location, origin time, depth, magnitude, and a list of stations used in computing these parameters. These messages contain no evaluations regarding seismic or tsunami hazard, as the magnitude of the earthquake is far too small to have a significant tsunami generation potential.

A Tsunami Information Bulletin (TIB): The PTWC issues this message product for any earthquake in or near the vicinity of the Pacific Basin with a magnitude in the range $6.5 \leq M_W \leq 7.5$. A TIB states that a destructive tsunami is not expected outside the area of the epicenter. However, it does warn of the possibility of a destructive tsunami along coastlines within 100 km of the epicenter.

A Fixed Regional Warning Bulletin: The PTWC issues this message product for shallow underwater earthquakes (depth < 100 km) with a magnitude in the range $7.6 \leq M_W \leq 7.8$. This bulletin warns of the possibility of a regionally destructive tsunami within 1000 km of the epicenter. All regions within 1000 km of the epicenter are thus initially placed in a warning.

An Expanding Watch/Warning Bulletin: The PTWC issues this message for shallow underwater earthquakes with magnitude $M_W \geq 7.9$. This criteria is similar to the fixed watch/warning with the exception that this bulletin also warns of the possibility of a destructive tsunami traveling greater than 1000 km away from its source area. The use of the term "expanding" stems from the fact that the watch and warning regions expand across the Pacific as time progresses until the watch/warning is canceled. The extensions of the watch/warning area are referenced to the leading edge of the tsunami waves at the time the bulletin is issued. Areas within 3 to 6 h tsunami travel-time from the predicted current leading edge of the tsunami are placed in a watch. Areas within 3 h tsunami travel-time are placed in a warning. All other areas are placed in an advisory. Because of the expanding nature of the watch/warning, areas that were initially only placed in an advisory may eventually come to fall into the watch or warning region. If no potentially destructive tsunami is detected by sea-level stations, the watch/warning is canceled. On the other hand, if the data provided by sea-level stations provide evidence that a destructive tsunami is moving across the Pacific, the PTWC may upgrade to its most severe message, the Pacific-Wide Warning. A Pacific-Wide Warning is a tsunami warning for all coasts in the Pacific Basin. Before issuing a Pacific-Wide Warning the scientist on duty must confirm the presence of a potentially dangerous tsunami on sea-level instruments.

Figure 3 summarizes our response in retrospect, had these criteria been in place over the 20th century, after applying them to the earthquakes and tsunami that occurred during this period. The application of these criteria would have resulted in the issuance of a TIB, but no warning for 23 destructive locally generated tsunamis occurring over the last century in the Pacific Basin. At larger distances from the earthquake, the PTWC would have issued a *Fixed Regional Warning Bulletin* ahead of 20 destructive tsunamis generated in the last century, an *Expanding Watch/Warning Bulletin* for thirty-eight destructive tsunamis, as well as eight major Pacific Basin wide tsunami warnings within the same time period.

Coastlines close to the earthquake epicenter can experience tsunami waves within two to fifteen minutes after the earthquake; hence a local tsunami warning needs to be issued within a few minutes to be effective. This requires access to real-time data provided by a dense local network of seismic stations near the epicenter that allows both, the rapid location, and source characterization of the earthquake. In the case of the Hawaii region, the PTWC uses data from its own seismic network, and from the dense seismic network maintained by the USGS Hawaii Volcano Observatory (HVO) to rapidly detect Hawaii earthquakes. These data enabled PTWC to issue an information bulletin to the state of Hawaii, and to the Pacific Basin for the Kiholo Bay earthquake (M_W 6.7) within 3 min of the origin time of the earthquake [42]. However, without access to dense local seismic networks around the Pacific rim, the PTWC is unable to issue timely warnings for populations in the immediate vicinity of large earthquakes outside of Hawaii. As a result, the PTWC does not have the capability of functioning as a local warning center for areas outside of Hawaii.

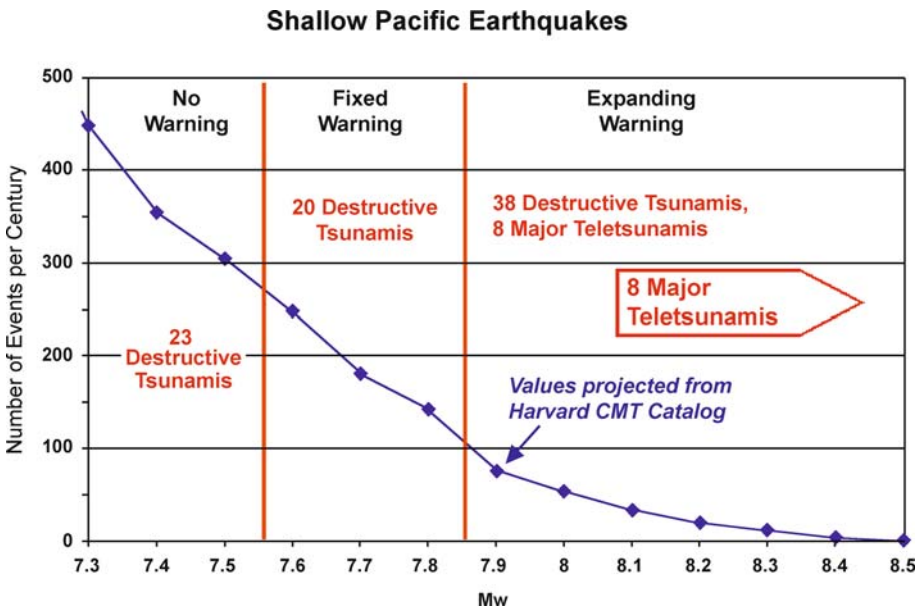

Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 3
Retroactive performance based on current bulletin criteria

The WC/ATWC of the US National Weather Service has access to dense local seismic networks on the US mainland, Puerto Rico, and Canada, and can therefore provide rapid warnings to the US and Canadian West Coast as well as to Puerto Rico and the Virgin Islands. Japan has a similar capability for its coastlines, and spurred on by the December 2004 Sumatra earthquake, several other nations such as Indonesia and New Zealand, for example, are rapidly developing and improving their seismic networks in an effort to improve their tsunami warning capabilities.

Seismic Methods

To rapidly detect, locate, and characterize the source of earthquakes occurring around the world, tsunami warning centers rely on the Global Seismic Network (GSN USGS/IRIS) which has many contributors in the US and worldwide. It is this unfettered access to real time seismic data supplied by a number of different networks that makes a basin-wide tsunami warning center possible. To rapidly deal with the threat posed by locally generated tsunamis to the state of Hawaii, PTWC processes seismic data from about 70 stations located in the Hawaiian Islands. The USGS HVO's dense network supplies most of this data. The US tsunami warning centers use the Earthworm software developed by the USGS to import and export seismic data [46].

PTWC duty scientists can receive automatic pages at any time, for any earthquake with magnitude M_W above ~ 5.5. The system generating these pages combines Evan's and Allen's [23] teleseismic event detection algorithm, adapted for broadband data by Wither's [84], and Whitmore's [81] teleseismic picker and associator. In the Hawaiian Islands, the application of Hirshorn and Lindh's [40] algorithm notifies duty scientists for earthquakes with magnitudes larger than about 3.5 within 10 to 20 s of the earthquakes origin time. Other software automatically locates the event, and provides a first estimate of the earthquake's magnitude, and other source parameters in real time [3,39,40,45]. PTWC duty scientists then refine and supplement the software's automated real-time hypocenter location and magnitude estimates. Determining the earthquake's depth is particularly important as earthquakes occurring at depths greater than about 100 km generally don't cause tsunamis.

Earthquakes are located using P-wave arrival times recorded at a number of seismic instruments. As both the locations of seismic instruments and P-wave travel-times as a function of distance are well known, a process analogous to triangulation is used to locate the earthquake. The pickers and associators perform these functions automatically on a continuous basis.

While the depth of the earthquake can be estimated on the basis of P-times alone, a more robust result often requires the addition of depth phases such as pP which is

a P-wave that travels directly up to the earth's surface from the earthquake source and reflects once off of the Earth's surface before arriving at the seismometer. The duty scientists use pP arrival times to refine hypocentral depths of distant earthquakes (teleseisms). For earthquakes in Hawaii, observed at local distances, the S-wave arrival time would be useful for constraining earthquake depth. However, automatic picking of S-wave arrival times is not yet robust, and manual picking and incorporation of S-wave arrival times into the analyzes by a duty scientist would take too long in the local earthquake case. In addition, as the deepest earthquakes in Hawaii have a hypocenter located about 50 km deep, (Fig. 6) which is comparable to the rupture length of an earthquake over about magnitude seven, the PTWC's local tsunami warning criteria are currently based on magnitude only.

Seismologists use a panoply of different magnitudes to characterize the seismic source. These different methods examine different parts of the seismic wave train, such as short and long period body waves (seismic waves that

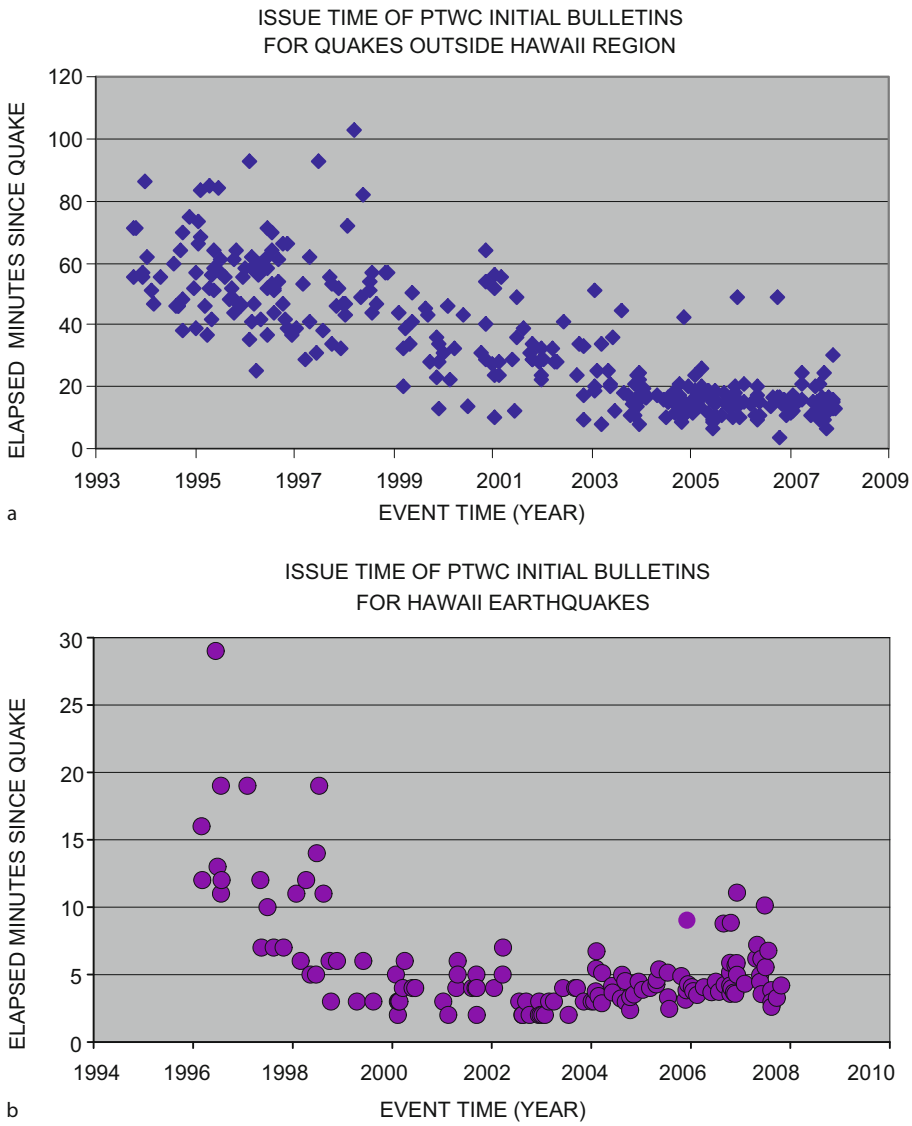

Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 4
Elapsed time from earthquake origin time to issuance of first official message product for a earthquakes outside the Hawaii Region and b for earthquakes within the Hawaii Region

travel through the earth's interior like the P- and S-waves) and long period surface waves (slower seismic waves that are constrained to travel along the earth's surface). Most of these magnitude scales were developed to estimate the same quantity; the energy released by the earthquake as radiated seismic wave energy, E_R. Traditional magnitude measures such as M_L [66], and m_b (a shorter period variant of m_B Gutenberg [30,31] that examines high frequency body waves). The Gutenberg surface wave magnitude M_S [9], modified by Vanek [75] as well as the newer M_m Okal and Talandier [63], or mantle magnitude are derived from the surface waves. A relatively new and quick method, M_{wp} analyzes long period P-waves [73,74,82]. The M_{wp} magnitude is now the magnitude used in the decision process for deciding which if any official message product to issue, supplanting the M_S method which had been used for over 50 years. For large earthquakes, duty scientists also routinely estimate M_m, a very long period surface wave magnitude based on mantle waves with periods in the range 50–410 s [63]. The relationship between these magnitudes, each looking at different parts of the seismic wave spectrum of an earthquake, can be used to characterize the earthquake source [2,16,17].

When evaluating the tsunamigenic potential of an event, PTWC duty scientists also compute the quantity $\log_{10}(E_R/M_0)$, known as "Theta", Θ, where M_0 is the seismic moment [1]. Newman and Okal [62] showed that Θ is anomalously small for tsunami earthquakes.

Since about the mid 1990's the two US Tsunami Warning Centers response times to potentially tsunamigenic teleseisms has decreased dramatically due to the much larger amounts of seismic data that they now receive, and to the switch from the slower M_S magnitude method to the faster M_{wp} moment magnitude method for their initial messages (Fig. 4a). For local events, using the real-time associator binder_agl [45], and the very fast *pMag* scale [40], has brought the PTWC's response time down to less than 3 min (Fig. 4b).

Earthquake Source Parameters

A fundamental problem with traditional magnitude estimates such as M_L, m_b, and M_S, is that they are based on the amplitudes of relatively short period seismic waves with periods usually less than 3 s for m_b and M_L, and about 20 s for M_S. When the largest rupture dimension of the earthquake exceeds the wavelength of these seismic waves, which is about 50 km for the 20 s period surface waves used for M_S [48,49], these magnitude values will start to "saturate". Saturation in this case means that these magnitude measures will underestimate the true size of the

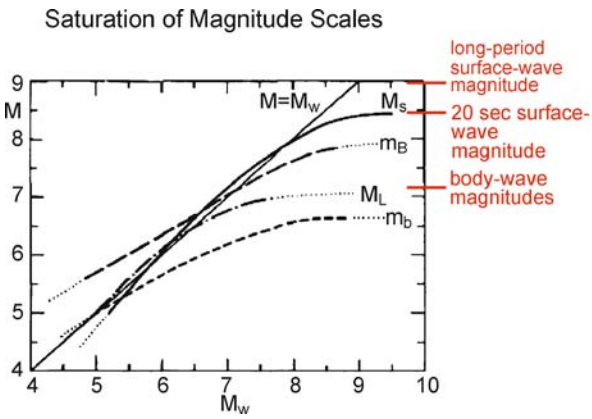

Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 5
Saturation of different classical magnitude scales with respect to non-saturating moment magnitude according to Kanamori [50]. Note that m_B refers to the original Gutenberg–Richter [32,33] body-wave magnitude scale based on amplitude measurements made on medium-period broadband instruments. It saturates at larger magnitudes when compared to the short-period based m_b

earthquake when the periods of the amplitudes on which they are based are shorter than the corner period of the earthquake's seismic wave spectrum [2,16,17] (see Fig. 5). Another, equivalent explanation is that these magnitude methods, which look at waves with periods of a fraction of a second to a few tens of seconds, cannot sample enough of the energy released by an earthquake whose source duration (the length of time over which the rupture occurs) is many times larger than the periods used by these methods. As the earthquake becomes very large, one needs to examine longer period waves to avoid saturation.

K. Aki used a spectral representation to establish that earthquakes of varying size had spectra of similar shape, differing primarily in the low frequency amplitude, proportional to seismic moment, and the location of the "characteristic frequency" (corner frequency of the source spectrum) which he related to the characteristic length scale of an earthquake [2]. Subsequent studies by Brune [16,17], and Savage [68] also related the corner frequency to the dimensions of the fault plane.

To circumvent the saturation problem, Kanamori defined a new magnitude scale, the moment magnitude M_W, [48], in terms of a minimum estimate of the total co-seismic strain energy drop, W_0, via Gutenberg and Richter's energy-magnitude relationship [33]. The M_W scale, more properly a magnitude that describes the total "work" required to rupture the fault, is computed from the seismic moment, M_0, assuming 1) that the stress changes associated with large earthquakes are approximately con-

stant, and 2) that the stress release during an earthquake is about the same as the kinetic frictional stress during faulting. M_W and its agreement with the M_L and M_S magnitude scales in their unsaturated ranges was discussed by T.C. Hanks and H. Kanamori [34] while Kanamori [50] discusses additionally the average relationship of M_W with m_b and m_B, also in the range where these magnitudes saturate (see Fig. 5).

Traditional Amplitude Based Magnitudes at the PTWC

Local Earthquake Magnitude Methods Hirshorn and Lindh [40] developed a short period P-wave magnitude scale, called *pMag*, which is based on the average of the absolute values of the amplitudes of the first three half-cycles of the initial p-waves recorded, at local distances, on short period seismometers [40,45]. The *pMag* scale is based on the assumption that that the decrease of locally recorded initial P-wave amplitudes with increasing hypocentral distance shares a common decay curve in a given geographic area, independent of the magnitude of the earthquake. Lindh and Hirshorn incorporated *pMag* into Carl Johnson's [46] local p-wave associator, binder_agl, enabling automatic pages, containing the hypocentral parameters and a the lower bound magnitude estimate provided by *pMag*, within about 10 to 20 s of an events origin time. At the PTWC, the System for Processing Local Earthquakes in Real Time (SPLERT) is based on this software.

PTWC also uses a very band-limited M_L scale, based on the maximum amplitudes measured on the horizontal components of short period seismograms recorded at local hypocentral distances from earthquakes that occur in Hawaii. These short period waves attenuate about 6 times less along the path between the hypocenter and recording stations for the "deeper" population of Hawaiian earthquakes (with hypocenters (Fig. 6) located between 20 and 50 km depth), then they do along the path from source to receiver for the "shallower" event population in the lower oceanic crust (about 10–20 km below the sea surface). For this reason, PTWC adds 0.8 magnitude units to the $M_L$values obtained for events with hypocentral depths $\geq$ 25 km.

Because of the bimodal depth distribution of Hawaii earthquakes (Fig. 6.) our M_L calculation requires only

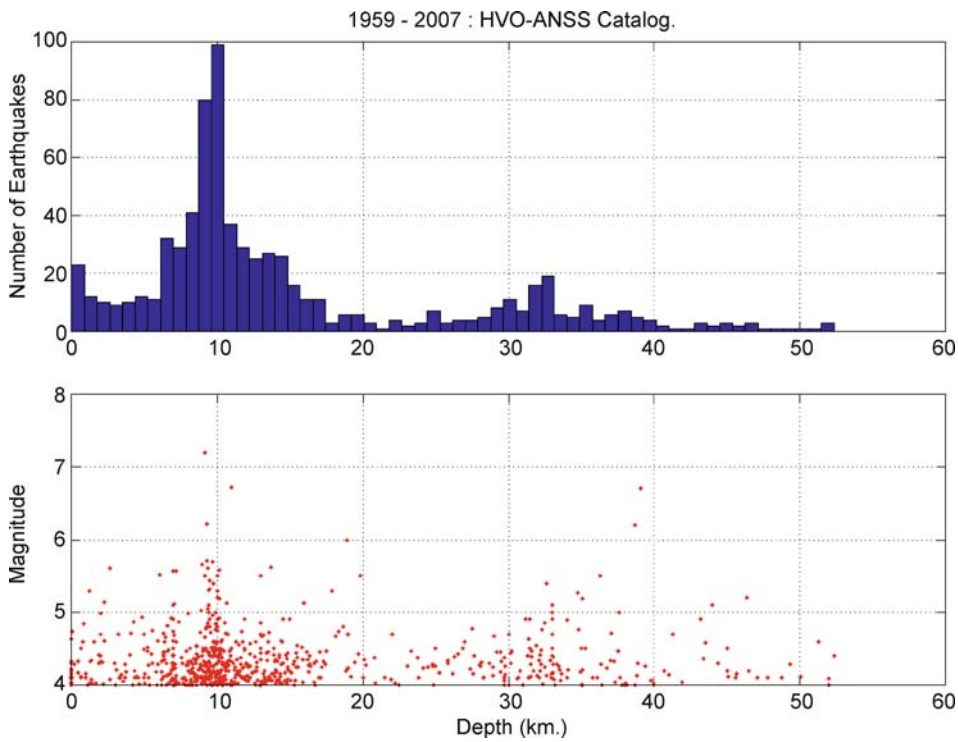

Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 6
This figure shows the bimodal depth distribution of Hawaii earthquakes of $M \geq 4$, taken from the USGS Hawaiian Volcano Observatory (HVO) ANSS Catalog. The *top figure* is a histogram of all events binned by hypocentral depth. The *bottom figure* shows the magnitude vs. hypocentral depth for each event

a good enough depth estimate to discriminate between these two populations.

Teleseismic Magnitude Methods PTWC's body wave magnitude method is called *bMag* and has similarities to the intermediate period broadband body wave m_B magnitude as defined by IASPEI. The IASPEI m_B [11,16] is based on Gutenberg's [30,31] and Gutenberg and Richter's m_B [32,33]. *bMag* uses a 90 s window of broadband vertical component seismogram starting 30 s prior to the arrival of the P-wave. This window is band-passed filtered between .3 s and 5 s. The largest amplitude and its period found in the 60 s after the first P-wave arrival are chosen for use in the magnitude formula. In PTWC's implementation, the formula used is the same as Gutenberg and Richter's [32,33] relation, adopted by IASPEI for m_B:

$$bMag = \log(A_{\max}/T_{\max}) + Q(\Delta, z) \ ,$$

where Δ is the epicentral distance ($15° \leq \Delta \leq 90°$), z is the hypocentral depth, $A_{\max}$ is the maximum wave amplitude obtained from the band-pass filtered record, and $T_{\max}$ is the period of the wave with the maximum amplitude. Gutenberg and Richter's [33] table of $Q(\Delta, z)$ is used to provide the distance and depth correction. The largest amplitude found in the 30 s prior to the P-wave arrival time is used as the basis for the signal to noise ratio. *bMag* differs from the IASPEI m_B in three respects,

1. m_B uses the largest amplitude wave in the P-wave coda up to the arrival of the PP phase,
2. m_B uses a slightly different distance range ($15° \leq \Delta \leq 90°$), and
3. for m_B, the seismogram is band-pass filtered using the band .3 s to 5 s.

bMag will saturate at lower magnitudes than M_S does, so it is of limited use for large earthquakes. However, *bMag*, is still useful for three main reasons

1. unlike M_S, *bMag* has a correction for the depth of the event's hypocenter,
2. it is useful for determining the magnitude of moderate earthquakes that occur as aftershocks of much larger earthquakes, e. g. when longer period energy is still present in the signal from an earlier, larger event, that can adversely affect magnitude methods based on longer periods, and
3. by comparison with magnitudes based on longer periods, such as M_w, it can also provide a way to detect slow or tsunami earthquakes.

In computing M_S (first proposed by Gutenberg [9] and later revised by Vanek et al. [75]) at the PTWC, we band-pass filter 14 min of the broadband velocity seismogram with a 7 s band, from 16 to 23 s, starting 3 min before the expected arrival time of the surface waves. We then apply the following equation, similar to the IASPEI [11,75] formula:

$$M_S = \log_{10}(A_{\max}/T_{\max}) + 1.66 \log_{10}(\Delta) + 3.3 + \text{correction}.$$

The *correction* term is 0 for epicentral distances, Δ, greater than 16°, and 0.53–0.033Δ for Δ less than 16°. Note that the IASPEI implementation considers a much greater range in periods from 3 to 60 s for $T_{\max}$ and has no need for the *correction* term. The *correction* term allows the US TWC's to compute the M_S magnitudes from stations as close as 600 km to the epicenter at a period of 20 s. This method is susceptible to saturation effects as the magnitude reaches the high 7's.

Although M_S is no longer used as the basis for issuing bulletins, it is still helpful in diagnosing deep earthquakes, and for comparing the amount of 20 s radiated energy to the amounts of radiated energy at other periods. Deep earthquakes do not excite large surface waves. Hence if $bMag > M_S$ the hypocenter is likely to be deep.

The M_{wp} Method

The broadband P-wave moment magnitude, M_{wp}, has replaced M_s as the magnitude upon which the US TWCs initial tsunami messages are based [73,74,82]. This is because M_{wp} uses P-waves, recorded at any epicentral distance, up to about 90°, when the observed initial P-waves are affected by refraction due to the earth's outer core. M_{wp} is obtained much quicker than M_s, which is based on the slower traveling surface waves, and because M_{wp} examines much longer period waves than the 20 s surface used by M_s making it less susceptible to the saturation effects discussed above. M_{wp}, as implemented at the PTWC, uses the first 120 s of the vertical component, broadband velocity seismogram, beginning at the P-wave arrival time (gray trace in Fig. 7).

The derivation of M_{wp} assumes that we can obtain the seismic moment, M_0, from the far-field P- and/or pP-wave portion of the vertical broadband displacement waveform, $u_Z(x_r, t)$, using

$$M_0 = (4\pi\rho\alpha^3 r/F^P)\text{Max}\left|\int u_Z(x_r, t)\mathrm{d}t\right| \ ,$$

where ρ and α are the density and P-wave velocity averaged along the propagation path, r is the epicentral

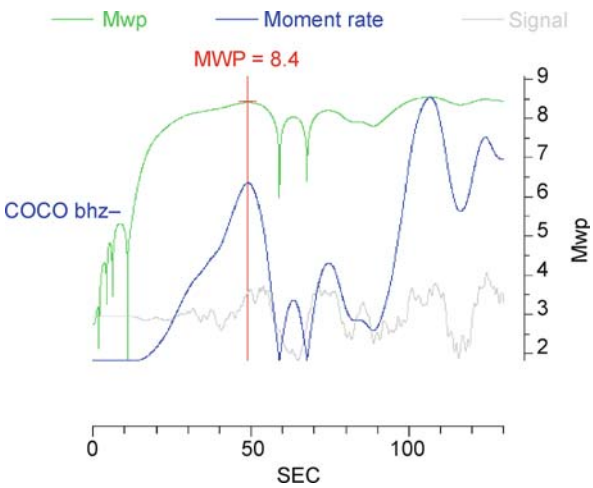

Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 7

The first 2 min of the broadband vertical velocity seismogram (*gray*) recorded by the GSN USGS/IRIS broadband station COCO, on Cocos Island, about 15 degrees south of the epicenter of the M_W 9.2 Sumatra Earthquake of Dec. 26, 2004. Note that this portion of the broadband velocity seismogram is not clipped. This instrument, a KS54000, has a flat frequency response to velocity to a period of about 350 s. The blue trace is the integrated displacement record (doubly integrated velocity), and the *green* trace is M_{wp} as a function of time

distance, and F^P is the earthquake source radiation pattern [73,74,82]. At the PTWC, we follow Tsuboi [73], approximating Max $\left|\int u_Z(x_r, t)\mathrm{d}t\right|$ by the first significant or "big" peak in the absolute value of the integrated displacement record. We prefer to use velocity seismograms, $v(t)$, from STS-1 or KS54000 broadband seismometers as there is then no need to deconvolve the instrument response from the data. We simply scale the data by a gain factor, because we can assume that the instrument response function is flat in the frequency band of interest. For the STS-1, or the KS54000, which both have a flat velocity up to about 350 s, this works for all but the very largest or slowest earthquakes.

We first remove any pre-event offset from $v(t)$, ending before the P-waves from the earthquake arrive, integrate $v(t)$ twice, and then multiply the absolute value of each data point by $4\rho r\alpha^3$ to obtain $M_0(t)$ in N-m (the blue trace in Fig. 7). We then apply the standard IASPEI moment magnitude formula [34]:

$$M_W = (\log_{10} M_0 - 9.1)/1.5$$

to $M_0(t)$, to calculate $M_W(t)$ (the green trace in Fig. 7). To correct for the radiation pattern, F^P, we then add 0.2 to the average of the individual M_{wp} values, each obtained at different azimuths and distances from the epicenter. This is because $\int (F^P)^2 \mathrm{d}\Omega = 4/15$, where Ω is the azimuthal angle of the observation around the epicenter, and $\sqrt{4/15} = 0.52$. Therefore, we multiply the averaged M_0 by 2, which is equivalent to adding 0.2 to M_{wp}. Finally, we apply the Whitmore et al. [82] magnitude dependent correction, $M_{wp} = (M_{wp} - 1.3)/0.843$, to get a final value for M_{wp}.

Figure 8 compares these final M_{wp} values resulting from this procedure with the Harvard moment magnitude M_W estimated from their CMT [19] solutions.

For some complex earthquakes, such as the M_W 8.4 [19] Peru earthquake of June 21, 2001, or the great M_W 9.2 [5,65,71] Sumatra earthquake of December 2004, M_{wp} (7.4 and 8.1, respectively) will underestimate M_W, when the first moment release is not the largest. In contrast, the PTWC's final estimate of M_{wp} 8.4 for the M_W 8.6 [19] Nias event of March 28, 2005 was acceptable, as it was for 9 other earthquakes in the range $8.0 \le M_W \le 8.4$ (Fig. 8).

The PTWC also uses M_{wp} for large local earthquakes, occurring in the Hawaiian Islands [41,42]. M_{wp} is based on the far field formulation for P-wave displacements [73]. For earthquakes whose largest source dimension is small compared to the distances at which the P-waves are observed, this assumption is satisfied. Up to approximately M_W 7.5, M_{wp} calculated from these locally recorded, far field P-waves agrees well with the Harvard M_W values for the same events [41,42]. For example the PTWC calculated a value of M_{wp} 6.5 for the M_W 6.7 [19] Kiholo Bay event within two minutes of the initiation of rupture at the hypocenter [41,42].

The Mantle Magnitude (M_m) Method

Emile Okal and J. Talandier developed the M_m method in 1989 [63]. The mantle magnitude is related to the moment magnitude via the simple expression $M_W = M_m/1.5+2.6$. This work was inspired by the need to develop a magnitude method for tsunami warning centers that would not suffer the saturation problem of M_S [64]. Not only may M_S saturate as the magnitude becomes large (> 8) but slow earthquakes can cause M_S to be seriously deficient and *bMag* even more so. Severely underestimating the magnitude of an earthquake can lead to a failure to warn. PTWC's implementation of the M_m method is based on analyzing Rayleigh waves obtained on vertical component seismograms.

M_m being based on slow traveling long period surface waves, is available too late to be used in the decision process for issuing an initial bulletin. Notwithstanding, it does

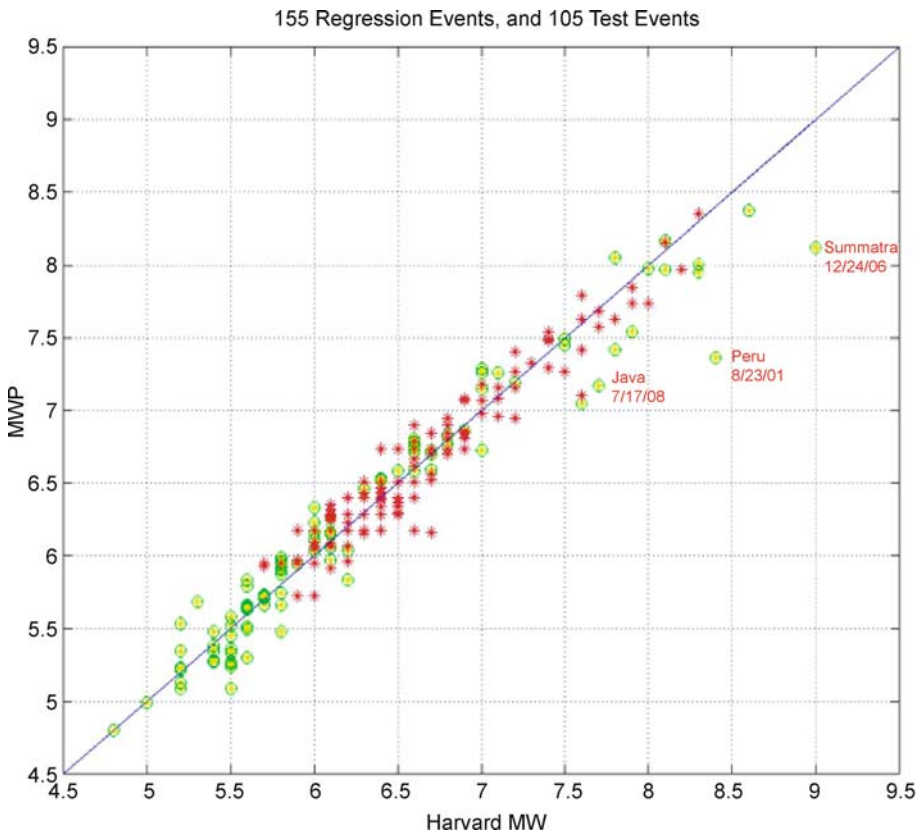

Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 8

A Scatter plot of average M_{wp} versus the M_W [19] for a set of 260 earthquakes with magnitudes in the interval $4.8 \leq M_W \leq 9.2$ occurring from 1994 through October of 2007. Whitmore et al. [82] found that with the application of an empirical correction made to the results, satisfactory results could be obtained. Our linear, least squares fit to the 155 earthquakes (*red stars*) yielded a slope of 0.83, close to the Whitmore et al. [82] slope of 0.84. Since April of 2002, we have used this corrected relationship (*green circles filled in yellow*) to calculate M_{wp}

provide a useful check on the magnitude obtained from the M_{wp} method and if there is a discrepancy between M_{wp} and $M_W(M_m)$ on the order of 2–3 tenths or more in the 7+ magnitude range, the duty scientist may instead use the results of the M_m method in subsequent bulletins. The M_m method overcomes the limitation of saturation because it is a variable period magnitude. Multiple values of M_m are routinely computed for a number of fixed periods ranging from 50 to 270 s for each station. Because M_m may saturate at the smaller periods for great earthquakes, while at longer periods M_m will be unsaturated, Okal and Talandier's [63] procedure was to choose the largest M_m mitigating the effects of saturation.

M_m is more complicated than the other methods described here as it uses frequency domain deconvolution. This can cause problems due to deconvolution noise at low magnitudes, where the amplification of noise by the deconvolution process at long periods may result in spurious magnitudes. Thus M_m works best with very long period broadband seismometers such as the KS54000's, KS36000's and STS-1's. While STS-2 seismometers tend to do well, however, the shorter period broadband seismometers tend to behave poorly at the longest periods [79]. Using the maximum M_m obtained for each station proved to be suboptimal due the heterogeneous distribution of instruments coupled with the total automation of the procedure at the PTWC. Weinstein and Okal [79] devised a sampling method that alleviates most of these difficulties in PTWC's implementation of M_m.

The December 2004 Sumatra earthquake showed that for earthquakes with an unusually long source duration (in this case ~ 600 s), even M_m at 270 s will saturate. Hence PTWC's M_m implementation will now automatically extend the period range to 410 s when the magnitude exceeds 8.0. At 410 s $M_W(M_m)$ is 8.9 [79] for the December 2004, Sumatra earthquake, still deficient, but a marked improve-

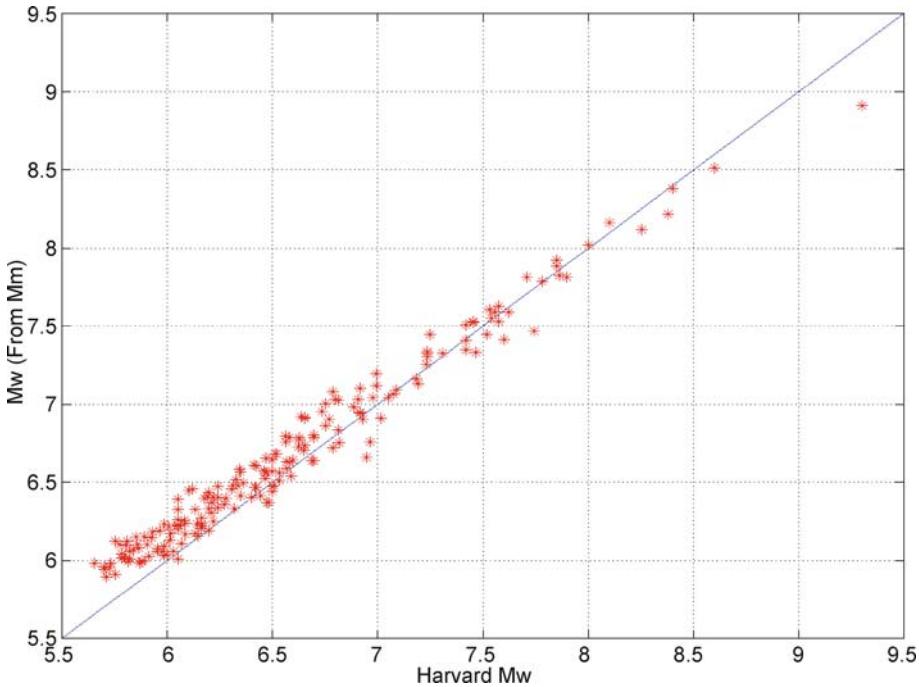

Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 9
Scatter plot of $M_W(M_m)$ vs. Harvard/GCMT [19] M_W for over 200 Earthquakes

ment over the moment magnitude 8.5 obtained by PTWC and 8.2 obtained by the USGS (NEIC Fast Moment Tensor) on Dec. 26, 2004. M_m normally uses a 660 s window of the surface wave train, but when M_m exceeds 8.0, the window expands to 910 s. Given the mix of instruments and their distribution used at PTWC, and the effects of broadband deconvolution noise, Weinstein and Okal [79], found that the M_m method was not useful for $M_W < 6.0$.

Figure 9 compares $M_W(M_m)$ values obtained for more than 200 recent earthquakes with the respective M_W values of HRV/GCMT [19] for the same events. PTWC's implementation of M_m of still tends to over estimate M_W by about .15 magnitude units for $M_W < 7.0$.

Rupture Slowness Estimation (Theta Program)

One way in which the occurrence of a tsunami earthquake may be indicated is if *bMag* and/or M_S are significantly smaller than M_W obtained from the longer period P-waves, or longer period mantle waves. This is made clear in Fig. 10. Note the population of 4 tsunami earthquakes that fall well off the trend. The short period magnitudes may also simply be deficient simply due to the size of the earthquake. Measuring the "rupture slowness" of the earthquake can further aid the warning centers in deciding between the two possibilities. As can be determined from Fig. 10, the body wave magnitude for July 2006, Java earthquake was deficient by nearly 1.5 magnitude units.

As mentioned earlier, a fundamental characteristic of a tsunami earthquake is the slowness of the rupture speed. Newman and Okal [62] showed that the log ratio of the radiated energy E_R [9,10], to the seismic moment M_0, $\mathrm{Log}_{10}(E_R/M_0)$ (also denoted by Theta, or "Θ") is anomalously small for tsunami earthquakes. A number of factors can affect this ratio such as rupture velocity, stress-drop/apparent stress, fault plane geometry, maximum strain at rupturing, and directivity (bi-lateral vs. unilateral rupture). However, for shallow thrust, low stress-drop subduction zone earthquakes, unusually slow rupture velocity may have the largest influence on the value of Θ.

Newman and Okal [62] showed that for tsunami quakes, the value of Θ is usually about -6.0 or less. For an earthquake with a unilateral rupture with nominal speed (~ 3 km/s), theory suggests that Θ is about -4.9 [26,69,76]. Weinstein and Okal [79] extended the original dataset of Newman and Okal [62] by including an additional 118 earthquakes. The mean value of all Θ values is approximately -5.1. However, when averaged by event, the distribution of Θ's peaks precisely at -4.9, in accordance with theoretical expectations. Given the standard deviation of 0.39 for all Θ's (for the 118 earthquakes), Weinstein and Okal [79] found that values of Θ around

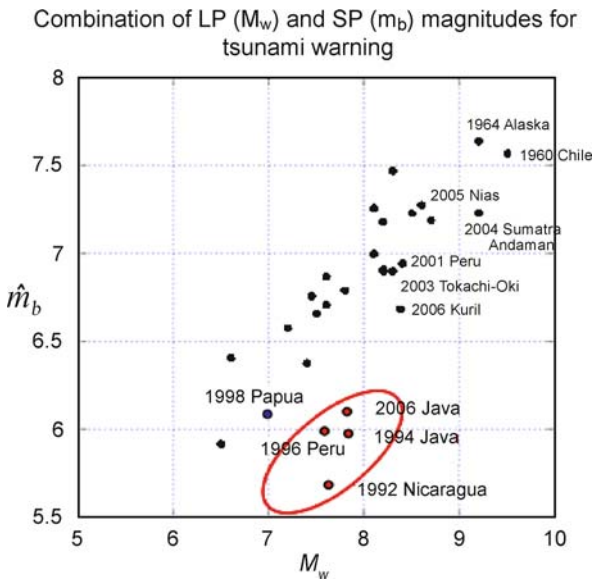

Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 10

Comparison between short-period $\hat{m}_b$ [43] and M_W for earthquakes with $M_W > 6$. Note the cluster of *red dots* representing tsunami earthquakes. This illustrates the diagnostic potential of short-period/long-period magnitude ratios to identify unusually slow earthquakes with high tsunami potential (Kanamori 2007, Talk at the PTWC in April of 2007)

-6.0 or below are more than 2σ off the mean and hence clearly anomalous.

The PTWC uses broadband vertical component seismograms, obtained in the distance interval $25° \leq \Delta \leq 90°$, to compute Θ. A window of 75 s is used starting approximately 5 s before the P-wave arrival. This is done to insure that the first arrivals are not missed by the integration. This window is deconvolved with the instrument response and the radiated energy contained between .1Hz and 2 Hz is computed.

In general it is thought that anomalously slow rupture speed is due to either low rigidity sediments in the fault or faulting through an accretionary prism [6,8,25,47,52,67]. In either case, the small shear rigidity associated with weak materials retards the rupture speed. As to why "slow" earthquakes produce more destructive than expected tsunamis, one can look at the well-known relation for moment magnitude:

$$M_0 = \mu A d \,,$$

where μ is the shear rigidity, A is the fault plane area and d is the average slip over the fault plane. Thus for two quakes with the same M_0 and all other properties equal except for μ and hence rupture speed, the slow quake requires a correspondingly larger slip, d, in order to achieve the same moment as the earthquake with a nominal value of μ and hence normal rupture speed.

One problem with Θ, is that it can be misleading and occasionally yield false indications of rupture slowness. This was made apparent by the Peru earthquake of June 23, 2001. This earthquake began with a initial event that had a moment magnitude of approximately 7.4, followed almost 60 s later by a much larger event, which had a moment magnitude of almost 8.4. [7,27,55]. Due to the 60 s delay, the Θ computation used mainly P-wave coda from the first shock, and little if any energy from the main shock. As a result the PTWC initially obtained a Θ of -6.1, using a moment based on M_m, making this earthquake appear very slow indeed. However, this result is spurious and was due to the complexity of the earthquake itself and not to actual slowness of the rupture.

Weinstein and Okal [79] found that by sliding the window over which Θ is computed forward in time, Θ would increase as the Θ window overlapped with the occurrence of the main event of the Peru 2001 earthquake. Indeed for a window offset of 70 s, Θ increases to -5.6, which is a strong trend to slowness, but not a slow or tsunami quake. This was further borne out by the size of the tsunami, which while detected on sea-level instruments around the Pacific (more than 2 m peak-to-peak in Chile), was not destructive outside of Peru.

Weinstein and Okal [79] explored the windowing technique and found that in actuality, it was a more comprehensive method than the single determination of Θ (zero offset). Computing theta in a succession of windows separated in time by 10 s (each window spanning 70 s) up to 100 s post P-wave arrival yields a better method of detecting slowness (see Fig. 11). What Weinstein and Okal [79] found is that for true tsunami earthquakes, the variation of Θ with offset time was small, generally no more than 0.1 log units over the entire 100 s. It is this flat trend that is probably the best discriminant for tsunami earthquakes.

In effect, the curve resulting from the window-offset technique tells us something about the source duration of the earthquake. Gigantic earthquakes have long source durations, and slow earthquakes have anomalously long source durations for their seismic moment. Therefore Θ can be viewed as a measurement of how anomalous the source duration is in terms of whether the earthquake is anomalously slow, or simply anomalously large. It turns out that in the case of the Sumatra earthquake of December 2004, Θ has little variation, even when the integration window is increased to 200 s and the offset carried out to 300 s. The magnitude of Θ based on PTWC's $M_W(M_m)$ of

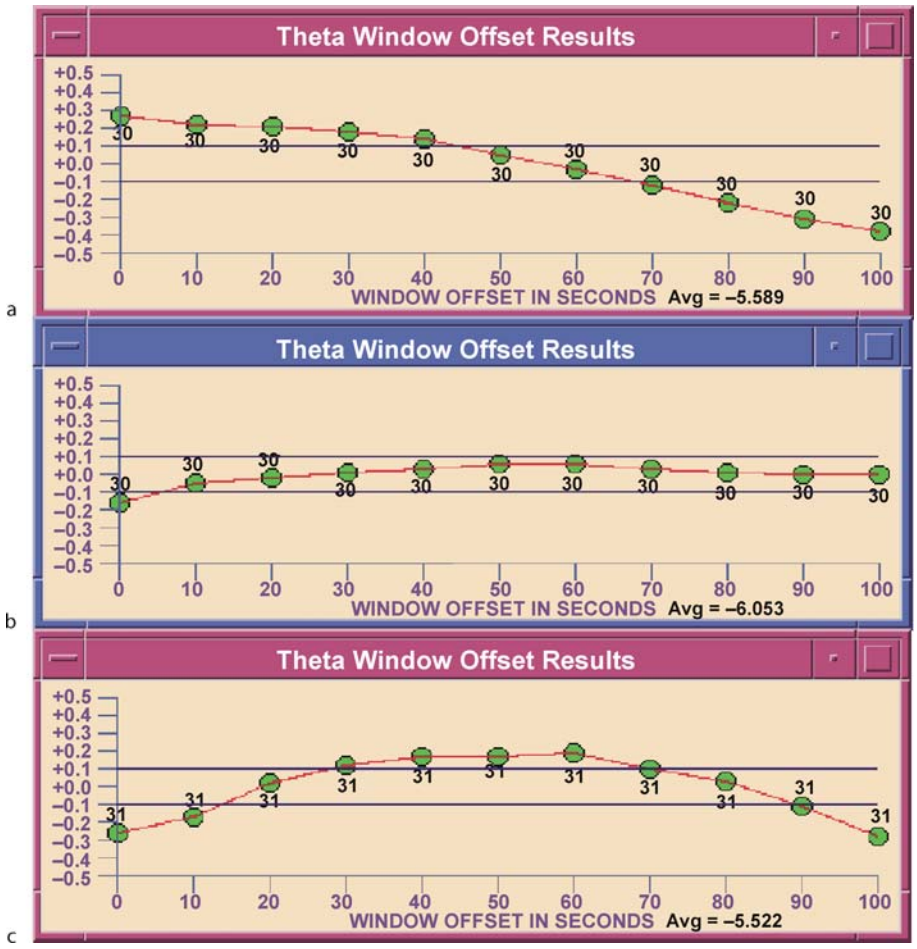

Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 11
The variation of Θ with offset for a a "normal" earthquake, b a "slow earthquake" (Java, 2006), and c a complex earthquake (Peru, 2008) respectively. In these plots Θ is de-meaned (the mean is found on the *bottom right* of the plot in *black*) and the number next to the dots indicates how many stations were used in computing that value. These plots are taken from PTWC's operational software

8.5 was ~ -5.6, a trend to slowness, but not slow. Using the $M_W$based on normal mode studies, Θ is ~ -6.1 (with a 200 s integration window!) and discussion continues to the current day as to whether or not the Sumatra earthquake of 2004 was slow, simply had aspects of a tsunami earthquake, or none at all [5,20,53,57,70].

Future Directions

Given the availability of high quality broadband seismic data, the tsunami warning centers can determine basic earthquake source parameters rapidly. However, the source characterization at the warning centers has rested largely on scalar measures of earthquake magnitude and slowness. The reasons for this are historical and practical. The warning centers have not always received the quantity of seismic data they do now, and in the interest of speed, the calculation of scalar measures can be accomplished with the data at hand in a small amount of time. One issue the PTWC faced during the 2004 Sumatra earthquake was that no near real-time magnitude method existed at the time that would correctly estimate the size of the Sumatra earthquake. Since then, new techniques have been developed to determine the magnitude of great earthquakes. Among these are techniques based on P-wave broadband signals [12,13,15,22,35,36,59,60] and the W-phase [21,52,55,58].

Hara [35,36] and Lomax et al. [59], both use techniques that involve estimating the source duration from the P-wave coda. This estimate is obtained by applying a high pass filter to the velocity seismogram, squaring the result and smoothing it. This procedure results in a rela-

tively smooth curve or envelope function that tracks the variation in the velocity-squared time-series. The source duration estimate is then obtained by measuring the time from the beginning of the P-wave to the point when the envelope function falls below a certain percentage of its maximum value. However, these studies also show that the use of source duration alone is not a completely satisfactory basis for a moment magnitude estimate.

Lomax et al. [59] also measures the radiated energy of the P-wave in the interval between the P- and S-arrivals, and uses both the radiated energy and source duration estimate to formulate a magnitude scale based on the relation $M_0 \approx E^{1/2} \bullet T^{3/2}$ [76] where E is the high frequency radiated energy and T is the source duration estimate. Hara [35] uses the estimated source duration and the maximum displacement measured in the interval of the estimated source duration from a number of earthquakes to construct an empirical formula for the magnitude. Hara [36] showed that this technique also works well for tsunami earthquakes.

Lomax et al. [60] has derived a duration-amplitude procedure for determination of a moment magnitude, M_{wpd}, for large earthquakes within 20 min of the event origin time using teleseismic P-wave recordings. Their procedure determines apparent source durations, T_0, from high frequency, P-wave records, and estimates seismic moments via integration of broadband seismograms over the interval t_p to $t_p + T_0$, where t_p is the P-wave arrival time. The characteristics of this method make it an extension of M_{wp}.

De Kool et al. [22] present a variation of the M_{wp} method which estimates the asymptotic behavior of the integrated displacement seismogram caused by the P-wave arrivals. Their results for M_{wp} show less scatter than do the PTWC's M_{wp} values, described above. In addition, they have automated their method.

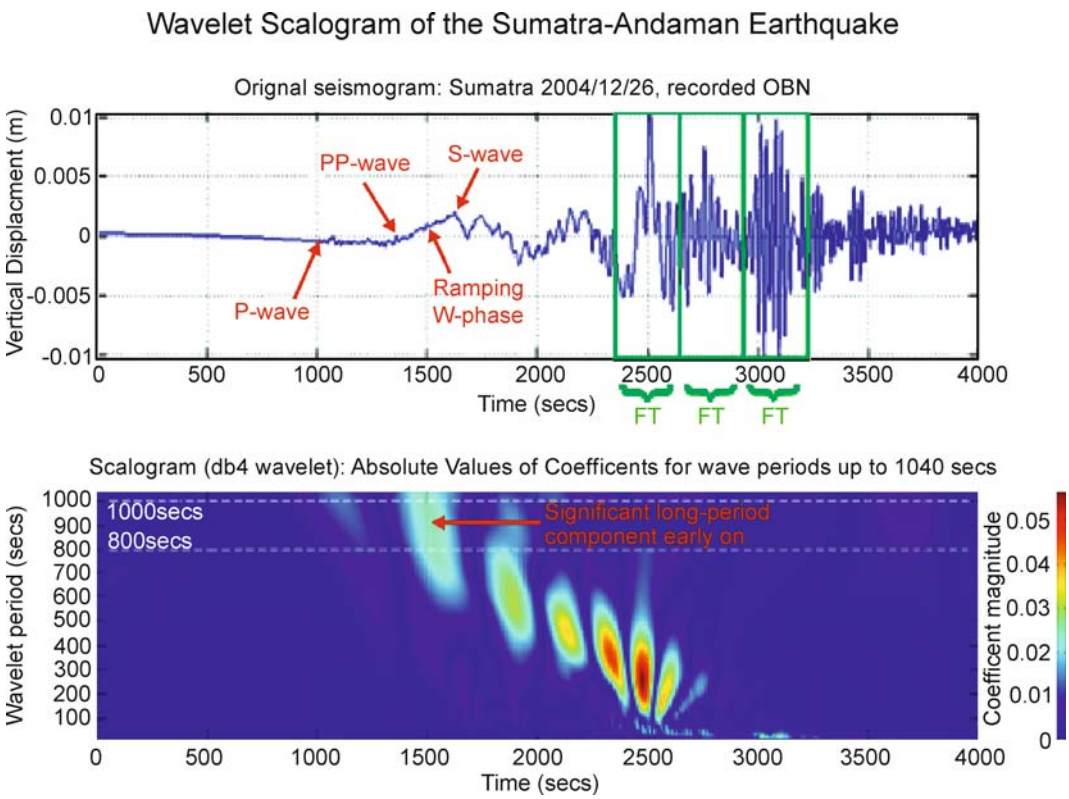

Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 12

***Top*: (Fig. 2a and b from [58]). Displacement seismogram of the 2004 Sumatra-Andaman earthquake recorded at OBN. *Bottom*: Scalogram of top seismogram. A diagram which displays the wavelet scale as a function of time is called a "scalogram". *Bottom figure* shows the scalogram for the 2004 event. Color intensity at any point in the picture corresponds to the coefficient magnitude of a wavelet with a particular period at a particular point of the time series. The *y*-axis has been translated from wavelet scale into corresponding wavelet time period. The long-period component arrived at about 1500 s. The wavelet transform can simultaneously achieve: (1) Accurate frequency representation for low frequencies, and (2) Good time resolution for high frequencies [58]**

The method of Bormann and Wylegalla [13] and Bormann and Saul [14] calculate what they refer to as a cumulative body wave magnitude. They do this by summing up all of the peak velocity amplitudes for all pulses, which represent the rupturing of sub-faults, over the P-wave coda. For the December 2004 Sumatra earthquake they obtained an M_W of 9.3 in agreement with the estimate of Stein and Okal [71].

Since these methods are based on analyzing the P-waves in the P-S interval, they provide accurate moment magnitude estimates for great earthquakes within 20–25 min of the earthquake origin time. Therefore these estimates will come before estimates obtained by other methods like M_m.

The W-phase is a long period, up to 1000 s, wave that arrives before the S-wave (Fig. 12). It can be interpreted as a superposition of the fundamental, 1st, 2nd, and 3rd overtones of spheroidal modes or Rayleigh waves and has a group velocity of 8 km/s at 1000 s period, and 8.6 km/s at 100 s period [51,54]. Kanamori [54] has devised a magnitude scale based on the amplitude of the W-phase observed on deconvolved displacement records.

In addition to size and source duration, the warning centers are interested in more detailed properties of the source than can be obtained from the scalar measures we have just discussed, such as direction of rupture and the distribution of slip along the fault. This information is important as it can be used by tsunami wave-height forecast models to better their predictions.

In the near future, the tsunami warning centers will incorporate the results of centroid moment tensors and finite fault modeling. Finite Fault modeling involves the inversion of seismic waveforms to recover more detailed information about the source process including the slip distribution, rupture propagation speed and moment release history [7,27,37,38,62,77]. Weinstein and Lundgren [80] explored the potential of a simple teleseismic P-wave inverse method for the rupture history of an earthquake for use in a tsunami warning center context. The calculations proceed quickly enough that a slip distribution may be available just a few minutes after a suitable set of P-waveforms are obtained. Hence finite fault modeling results can be used in tsunami wave height forecast models to provide a timely initial estimate of tsunami wave heights.

The warning centers are also actively investigating the use of seismic arrays. Seismic arrays can be used to determine the direction along which the P-waves have propagated to the array. As the rupture propagates, this direction will change. By analyzing the seismic array data, this change in direction can be measured (by computing back azimuths) and the history of energy/moment release as well as the extent and direction of the rupture propagations can be determined [44,56].

Acknowledgments

The authors greatly appreciate very thorough reviews by Kenji Satake, Anthony Lomax, Peter Bormann and our colleague Victor Sardina. Their comments greatly improved this manuscript. We also thank Paula Dunbar of the National Geophysical Data Center for helping us obtain the data we needed and Nathan Becker's GMT artistry for Fig. 1. We also thank the PTWC for use of its facilities in preparing this manuscript.

Bibliography

Primary Literature

1. Aki K (1966) Generation and propagation of G waves from the Niigata earthquake of June 16, 1964, Part 2. Estimation of earthquake moment, from the G wave spectrum. Bull Earthquake Res Inst Tokyo Univ 44:73–88
2. Aki K (1967) Scaling law of seismic spectrum. J Geophys Res 72:1217–1231
3. Allen RV (1978) Automatic earthquake recognition and timing from single traces. Bull Seism Soc Am 68:1521–1532
4. Allen RV (1982) Automatic phase pickers: their present use and future prospects. Bull Seism Soc Am 72:225–242
5. Ammon CJ, Ji C, Thio HK, Robinson D, Sidao N, Hjorleifsdottir V, Kanamori H, Lay T, Das S, Helmberger D, Ichinose G, Polet J, Wald D (2005) Rupture process of the 2004 Sumatra–Andaman Earthquake. Science 6:113–1139
6. Bilek SL, Lay T (1999) Rigidity variations with depth along interplate megathrust faults in subduction zones. Nature 400: 443–446
7. Bilek SL, Ruff LJ (2002) Analysis or the 23 June 2001 $M_W = 8.4$ Peru underthrusting earthquake and its aftershocks. Geophys Res Lett 8:21-1–21-4
8. Bilek SL, Lay T, Ruff LJ (2004) Radiated seismic energy and earthquake source duration variations from teleseismic source time functions for shallow subduction zone thrust earthquakes. J Geophys Res 109:B09308
9. Boatwright J, Choy GL (1986) Teleseismic estimates of the energy radiated by shallow earthquakes J Geophys Res 91: 2095–2112
10. Boatwright J, Choy GL, Seekins LC (2002) Regional estimates of radiated seismic energy. Bull Seism Soc Am 92:1241–1255
11. Bormann P, Baumbach M, Bock G, Grosser H, Choy GL, Boatwright J (2002) Seismic sources and source parameters. In: Bormann P (ed) IASPEI New Manual Seismological Observatory Practice, vol 1, Chapter 3. GeoForschungsZentrum Potsdam, pp 1–94. (This can also go as in the text cited review paper)
12. Bormann P, Wylegalla K (2005) Quick estimator of the size of great earthquakes. Eos Trans AGU 86(46):464
13. Bormann P, Wylegalla K, Saul J (2006) Broadband body-wave magnitudes m_B and m_{BC} for quick reliable estimation of the size of great earthquakes. USGS Tsunami Sources Work-

shop 2006, poster, http://spring.msi.umn.edu/USGS/Posters/Bormann_etal_poster.pdf
14. Bormann P, Saul J (2008) Earthquake magnitude. In: Meyers A (ed) Encyclopedia of complexity and systems science. Springer, Heidelberg
15. Bormann P, Saul J (2008) The new IASPEI standard broadband magnitude m_B. Seism Res Lett 79:698–705
16. Brune J (1970) Tectonic stress and seismic shear waves from earthquakes. J Geophys Res 75:4997–5009
17. Brune J (1971) Tectonic stress and seismic shear waves from earthquakes; Correction. J Geophys Res 76:5002
18. Bryant E (2001) Distribution and fatalities. In: Bryant E (ed) Tsunami: The underrated hazard. School of Science and the Environment, Coventry University, Coventry. Cambridge University Press, Cambridge, pp 15–24
19. Centroid Moment Tensor (2008) Catalog. http://www.globalcmt.org, accessed June 2008
20. Choy GL, Boatwright J (2007) The energy radiated by the 26 December 2004 Sumatra–Andaman earthquake estimated from 10-minute P-wave windows. Bull Seism Soc Am 97: S18–S24
21. Cummins PR (1997) Earthquake near field and W phase observations at teleseismic distances. Geophys Res Lett 24: 2857–2860
22. De Kool M, Jepsen D, Purss, Matthew (2007) Rapid moment estimation of large earthquakes using a variation of the Mwp method. In press
23. Evans JR, Allen S (1983) A teleseism-specific detection algorithm for single short-period traces. Bull Seism Soc Am 73:1173–1186
24. Fryer G, Hirshorn B, McCreery S, Cessaro RK, Weinstein S (2005) Tsunami warning in the near field: The approach in Hawaii. EOS 86:S44B-04
25. Fukao Y (1979) Tsunami earthquakes and subduction processes near deep-sea trenches. J Geophys Res 84:2303–2314
26. Geller RJ, Kanamori H (1977) Magnitudes of great shallow earthquakes from 1904 to 1952. Bull Seismol Soc Am 67: 587–598
27. Giovanni MK, Beck SL, Wagner L (2002) The June 23, 2001 Peru earthquake and the sourthern Peru subduction zone. Geophys Res Lett 29:14-1–14-4
28. Gower J (2005) Jason 1 detects the 26 December 2004 tsunami. Eos Trans AGU 86(4):37–38
29. Gutenberg B (1945) Amplitudes of surface waves and magnitudes of shallow earthquakes. Bull Seism Soc Am 35:3–12
30. Gutenberg B (1945) Amplitudes of P, PP, and S, and magnitudes of shallow earthquakes. Bull Seism Soc Am 35:57–69
31. Gutenberg B (1945) Magnitude determinations of deep-focus earthquakes. Bull Seism Soc Am 35:117–130
32. Gutenberg B, Richter CF (1956) Earthquake magnitude, intensity, energy and acceleration. Bull Seism Soc Am 46:105–145
33. Gutenberg B, Richter CF (1956) Magnitude and energy of earthquakes. Annali di Geofisica 9:1–15
34. Hanks T, Kanamori H (1979) A moment magnitude scale. J Geophys Res 84:2348–2350
35. Hara T (2007) Measurement of the duration of high-frequency radiation and its application to determination of the magnitudes of large shallow earthquakes. Earth Planets Space 59:227–231
36. Hara T (2007) Magnitude determination using duration of high frequency energy radiation and displacement amplitude: application to tsunami earthquakes. Earth Planets Space 59: 561–565
37. Hartzell S, Heaton T (1986) Rupture history of the 1984 Morgan Hill, California, Earthquake from the inversion of strong motion records. Bull Seism Soc Am 76:649–674
38. Hartzell S, Mendoza C (1991) Application of an iterative least-squares waveform inversion of strong-motion and teleseismic records to the 1978 Tabas, Iran earthquake. Bull Seism Soc Am 811:305–331
39. Hirshorn B, Lindh A, Allen R (1987) Real Time Signal Duration Magnitudes from Low-gain Short Period Seismometers. USGS OFR 87:630
40. Hirshorn B, Lindh G, Allen RV, Johnson C (1993) Real time magnitude estimation for a prototype early warning system (EWS) from the P-wave, and for earthquake hazards monitoring from the coda envelope. Seis Res Lett 64:48
41. Hirshorn B (2004) Moment magnitudes from the initial P-wave for local tsunami warnings. Seism Res Lett 74:272–273
42. Hirshorn B (2007) The Pacific Tsunami Warning Center Response to the Mw6.7 Kiholo Bay Earthquake and Lessons for the Future. Seism Res Lett 78:299
43. Houston H, Kanamori H (1986) Source spectra of great earthquakes, teleseismic constraints on rupture process and strong motion. Bull Seism Soc Am 76:19–42
44. Ishii M, Shearer PM, Houston H, Vidale JE (2005) Extent, duration and speed of the 2004 Sumatra–Andaman earthquake imaged by the Hi-Net array. Nature 435:933–936
45. Johnson CE, Lindh A, Hirshorn B (1994) Robust regional phase association. US Geol Surv Open-File Rept 94:621
46. Johnson CE, Bittenbinder A, Bogaert B, Dietz L, Kohler W (1995) Earthworm: a flexible approach to seismic network processing. IRIS Newslett XIV 2:1–4
47. Kanamori H (1972) Mechanism of tsunami earthquakes. Phys Earth Planet Inter 6:246–259
48. Kanamori H (1977) The energy release in great earthquakes: J Geophys Res 82:2981–2987
49. Kanamori H (1978) Quantification of earthquakes. Nature 271:411–414
50. Kanamori H (1983) Magnitude scale and quantification of earthquakes. Tectonophysics 93:185–199
51. Kanamori H (1993) W Phase. Geophys Res Lett 20:1691–1694
52. Kanamori H, Kikuchi M (1993) The 1992 Nicaragua earthquake: a slow tsunami earthquake associated with subducted sediment. Nature 361:714–715
53. Kanamori H (2006) Seismological Aspects of the December 2004 great Sumatra Andaman Earthquake. Earthquake Spectra 22:S1–S12
54. Kanamori H, Rivera L (2008) Source inversion of W phase – Speeding up tsunami warning. Geophys J Int 175:222–238
55. Kikuchi M, Yamanaka Y (2001) EIC Seismological Note Number 105, www.eic.eri-u-tokyo.ac.jp/EIC/EIC_news/105E.html
56. Krüger F, Ohrnberger M (2005) Tracking the rupture of the Mw9.3 Sumatra earthquake over 1150 km at teleseismic distance. Nature 435:937–939
57. Lay T, Kanamori H, Ammon C et al (2005) The Great Sumatra – Andaman Earthquake of 26 December 2004. Science 308:1127–1133
58. Lockwood OG, Kanamori H (2006) Wavelet analysis of the seismograms of the 2004 Sumatra–Andaman earthquake and its application to tsunami early warning. Geochem Geophys Geosyst 7, Q09013, doi:10.1029/2006GC001272

59. Lomax A, Michelini A, Piatanesi A (2007) An energy-duration procedure for rapid determination of earthquake magnitude and tsunamigenic potential. Geophys J Int 170:1195–1209
60. Lomax A, Michelini A (2008) Mwpd: A duration-amplitude procedure for rapid determination of earthquake magnitude and tsunamigenic potential from P waveforms. Geophys J Int (accepted, in press)
61. Mendoza C (1996) Rapid derivation of rupture history for large earthquakes. Seismol Res Lett 67:19–26
62. Newman AV, Okal EA (1998) Teleseismic estimates of radiated seismic energy: The E/M0 discriminant for tsunami earthquakes. J Geophys Res 103:26885–26898
63. Okal EA, Talandier J (1989) Mm: a variable-period mantle magnitude. J Geophys Res 94:4169–4193
64. Okal EA (1992) Use of mantle magnitude M_m for reassessment of the moment of historical earthquakes-I: Shallow events. PAGEOPH 139:17–57
65. Park J, Song TA, Tromp J, Okal E, Stein S, Roult G, Clevede E, Laske G, Kanamori H, Davis P, Berger J, Braitenberg C, Camp MV, Xiang'e L, Heping S, Houze X, Rosat S (2005) Earth's free oscillations excited by the 26 December 2004 Sumatra–Andaman earthquake. Science 20:1139–1144
66. Richter CF (1935) An instrumental earthquake magnitude scale. Bull Seism Soc Am 25:1–32
67. Satake K (1994) Mechanism of the 1992 Nicaragua tsunami earthquake. Geophys Res Lett 21:2519–2522
68. Savage JC (1972) Relation of corner frequency to fault dimensions. J Geophys Res 77:3788–3795
69. Scholz C (1982) Scaling laws for large earthquakes: Consequences for physical models. Bull Seism Soc Am 72:1–14
70. Seno T, Hirata K (2007) Did the 2004 Sumatra–Andaman earthquake involve a component of tsunami earthquakes? Bull Seism Soc Am 97:S296–S306
71. Stein S, Okal EA (2007) Ultralong period seismic study of the December 2004 Indian Ocean earthquake and implications for regional tectonics and the subdaction process. Bull Seism Soc Am 97:279–295
72. Indian Ocean Earthquake and implications for regional tectonics and the subduction process. Bull Seism Soc Am 97: S279–S295
73. Tsuboi SK, Abe K, Takano K, Yamanaka Y (1995) Rapid determination of Mw from broadband P waveforms. Bull Seism Soc Am 83:606–613
74. Tsuboi S, Whitmore PM, Sokolowski TJ (1999) Application of Mwp to deep and teleseismic earthquakes. Bull Seism Soc Am 89:1345–1351
75. Vanek J, Zatopek A, Karnik V, Kondorskaya N, Riznichenko Y, Savarenski S, Solovev S, Shebalin N (1962) Standardization of magnitude scales. Izv Acad Sci USSR Geophys Ser, pp 108–111 (English translation)
76. Vassiliou MS, Kanamori H (1982) The energy release in earthquakes. Bull Seism Soc Am 72:371–387
77. Wald DJ, Helmberger DV, Hartzell S (1990) Rupture process of the 1987 Superstition Hills earthquake from the inversion of strong-motion data. Bull Seism Soc Am 80:1079–1098
78. Weinstein SA, McCreery C, Hirshorn B, Whitmore P (2005) Comment on "A strategy to rapidly determine the magnitude of great earthquakes" by Menke W, Levin V. Eos 86:263
79. Weinstein SA, Okal EA (2005) The mantle magnitude M_m and the slowness parameter Θ: Five years of real-time use in the context of tsunami warning. Bull Seism Soc Am 95:779–799
80. Weinstein SA, Lundgren PL (2008) Finite fault modeling in a tsunami warning center context. In: Tiampo KF, Weatherley DK, Weinstein SA (eds) Earthquakes: Simulations, sources and tsunamis. Birkhauser, Basel
81. Whitmore PM, Sokolowski TJ (2002) Automatic earthquake processing developments at the US West Coast/Alaska tsunami warning center. Recent Research Developments in Seismology, 1–13. Kervala, India: Transworld Research Network. ISBN 81-7895, 072-3
82. Whitmore PM, Tsuboi S, Hirshorn B, Sokolowski TJ (2002) Magnitude-dependent correction for M_{wp}. Sci Tsunami Hazards J 20:187–192
83. Widjo K et al (2006) Rapid survey on tsunami Java 17 July, 2006. http://nctr.pmel.noaa.gov/java20060717/tsunami-java170706_e.pdf. Accessed July 2008
84. Withers M, Aster R, Young C, Beiriger J, Harris M, Trujillo J (1998) A comparison of select trigger algorithms for automated global seismic phase and event detection. Bull Seism Soc Am 88:95–106

Books and Reviews

Abercrombie R, McGarr A, Di Toro G, Kanamori H (eds) Earthquakes: Radiated energy and the physics of faulting. In: Geopysical Monographs 170. American Geophysical Union, Washington DC

Aki K, Richards PG (1980) Quantitative Seismology Theory and Methods, 2 vol. W.H. Freeman Co, San Francisco

Aki K, Richards PG (2002) Quantatative Seismology, 2nd edn. University Science Books, Sausalito

Båth M (1981) Earthquake magnitude – recent research and current trends. Earth Sci Rev 17:315–398

Bormann P (ed)(2002) IASPEI new manual of seismological observatory practice, vol 1 and 2. GeoForschungsZentrum Potsdam, p 1250

Duda S, Aki K (eds) (1983) Quantification of earthquakes. Tectonophysics 93, Special issue 3/4:183–356

Kanamori H, Anderson DL (1975) Theoretical basis of some empirical relations in seismology. Bull Seismol Soc Am 65(5): 1073–1095

Kanamori H (1994) Mechanics of Earthquakes. Annu Rev Earth Planet Sci 22:307–237

Kanamori H, Rivera L (2006) Energy partitioning during an earthquake. In: Abercrombie R, McGarr A, Kanamori H, Di Toro G (eds) Earthquakes: Radiated energy and the physics of faulting. Geophysical Monograph 170, American Geophysical Union, Washington, DC, pp 3–13

Kanamori H, Brodsky E (2004) The Physics of earthquakes. Rep Prog Phys 67:1429–1496

Kanamori H, The diversity of the physics of earthquakes. Proc Japan Acad Ser B 80:297–316

Lay T, Wallace TC (1995) Modern global seismology. Academic Press

Okal EA (1992) A student's guide to teleseismic body wave amplitudes. Seism Res Lett 63(N2)169–180

Richter CF (1958) Elementary seismology. WH Freeman Co, San Francisco

Stein S, Wysession M (2003) An introduction to seismology, earthquakes, and earth structure. Blackwell Publishing

Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in

YASUKO TAKEI
Earthquake Research Institute, University of Tokyo, Tokyo, Japan

Article Outline

Glossary
Definition of the Subject
Introduction
General Theoretical Framework to Describe the Dynamics of Solid–Liquid Composite Systems
Overview of Applications
Elastic Wave Propagation in a Solid–Liquid Composite System
Future Directions
Acknowledgments
Bibliography

Glossary

Partially molten rock The partially molten state is a thermodynamic state between solidus and liquidus temperatures, where both solid and liquid phases co-exist. In the Earth's interior, partial melting of rocks occurs in the upper mantle and/or crust beneath volcanic areas.

Melt Liquid phase in partially molten rocks or completely molten rock above the liquidus temperature is called melt. Density of melt is about 10% lower than solid. Hence, melt phase in the partially molten rocks tend to ascend toward the Earth's surface.

Aqueous fluid H_2O-rich fluid. In a subducting oceanic plate, at the depths of several tens of km, aqueous fluids are released by the dehydration of hydrated minerals. Aqueous fluids, having much lower density and viscosity than melts, tend to ascend due to the buoyancy force.

Seismic tomographic image A number of seismometer networks have been placed on the surface of the Earth to record the seismic wave propagation from seismic sources at depths to the surface. Using the traveltime data obtained from these observations, three-dimensional seismic velocity structures in the Earth can be obtained, with a process called seismic tomographic imaging. By using P and S wave traveltimes, V_P and V_S structures, respectively, can be obtained.

Definition of the Subject

The dynamics of solid-liquid composite systems are of great relevance to many problems in the earth sciences, including how melts or aqueous fluids generated by partial melting or dehydration migrate through the mantle and crust toward the surface, how deformation and fracture in these regions are influenced by the existence of fluids, and also how these fluids can be observed in the seismic tomographic images. The mechanical and transport properties of the solid-liquid composite systems strongly depend on liquid volume fraction and pore geometry, such as pore shape, pore size, and a detailed porosity distribution. Therefore, the microstructural processes that control pore geometry influence macroscopic dynamics, and vice versa. This article introduces a general continuum mechanical theory to treat the macroscopic dynamics of solid-liquid composite systems with a special emphasis on how such interactions with pore geometry can be described. Although intensive experimental and modeling approaches have been performed to investigate the interactive evolution of pore geometry and matrix deformation or fluid flow, many problems still remain unsolved and the actual liquid content and pore geometries in the crust and mantle are poorly understood. Therefore, in the latter part of this article, by applying the general theoretical framework introduced in the former part to the seismic wave propagation, the determinability of porosity and pore geometry from seismic tomographic images is discussed in detail. Due to the recent advances in seismic tomography, we can obtain three-dimensional and highly resolved images of both V_P and V_S structures. From the V_P or V_S structure alone, neither porosity nor pore geometry can be determined independently. However, if both V_P and V_S structures are available, porosity and pore geometry can be determined independently, thus providing valuable information complementary to experimental and modeling approaches. A practical method to determine porosity and pore geometry from the V_P and V_S images is presented.

Introduction

Melt segregation from partially molten mantle or crust to the surface is the fundamental process of volcanism. In arc volcanism, aqueous fluids derived from the dehydrating subducting slab migrate through the mantle wedge and play an important role in the melting process. Modeling approaches to these fluid migration processes have been developed to explain various chemical and petrological observations, e. g. [13,20,35,37]. To make in situ observations of these fluids, seismic and/or electromagnetic tomographic images of the partially molten regions are

produced. These images are then interpreted using methods built on experimental and theoretical studies of the effects of fluids on seismic and electromagnetic properties, e. g. [2,23,24,44]. The effects of fluids on the tectonic and/or volcanic earthquake source process have also been of great interests; recently, observations of deep low-frequency earthquakes and tremors have been considered to be indicators of the presence or active migration of fluids in the crust and mantle, e. g. [9,26,29].

Fluids in the crust and mantle exist as solid-liquid composite systems in which fluid-filled pores are included in the solid matrix. Solid-liquid composite systems are characterized by high structural sensitivity. When the liquid volume fraction increases from zero to a few tens of %, the mechanical and transport properties of the system change greatly from those of a solid to those of a liquid. These properties are not simply determined by the liquid volume fraction but also strongly depend on the geometry of the liquid-filled pores. Here, the term pore geometry represents not only pore shape, but also pore size, orientation, and homogeneous or heterogeneous porosity distribution. Therefore, liquid content and pore geometry strongly influence the melt segregation and/or matrix deformation dynamics. Pore geometry is not constant but can change interactively with the macroscopic dynamics. Experimental and modeling approaches have been performed to investigate the interactive evolution of pore geometry and matrix deformation or fluid flow, e. g. [11,35,39]. However, many problems, including a poor understanding of the underlying physics, discrepancies between the experimental and modeling results, and the issue of scaling laboratory results to km scale, remain unsolved and detailed forward approaches to the interactions are the subjects of future studies.

Remarkable progress has been made in the seismic observations of the solid-liquid composite systems in the crust and mantle. Due to the recent advances in seismic tomography, we can now obtain three-dimensional and highly-resolved images of both V_P and V_S structures. Because the seismic velocity depends on both liquid volume fraction and pore geometry, neither can be estimated independently from V_P or V_S alone. However, when both V_P and V_S are available, the liquid volume fraction and pore geometry can be estimated independently. By using V_P and V_S images, an inverse approach to estimate the actual pore geometry and fluid content in the crust and mantle can be performed, yielding complementary information to the experimental and modeling approaches mentioned above. Therefore, in the present introduction of the dynamics of solid-liquid composite systems, elastic wave propagation is discussed in detail with a special focus on the determinability of fluid content and pore geometry from seismic tomographic data.

First, in Sect. "General Theoretical Framework to Describe the Dynamics of Solid–Liquid Composite Systems", I introduce a general continuum mechanical formulation of the macroscopic dynamics of solid-liquid composite systems, with a special emphasis on how structural sensitivity is described. The elastic version of this theory, called "linear poroelasticity", was developed by Biot and coworkers, e. g. [5,6,31,48]. This theory, which assumes infinitesimal strain in the solid phase, is applicable to the propagation of elastic waves. A more general version applicable to large deformations was developed based on the fluid dynamic theory, and applied to the melt segregation dynamics in partially molten mantle, e. g. [7,20]. On the one hand, in most reviews of the theory of linear poroelasticity, the governing equations are introduced empirically and are difficult to compare to the mass and momentum conservation equations used in the usual continuum mechanical formulation, e. g. [48]. However, in those reviews, detailed explanations are given for the meaning of the macroscopic constitutive relation that describes the structural sensitive character of the solid-liquid composite systems. On the other hand, in most reviews of the general fluid-dynamical formulation the mass and momentum conservations can be easily confirmed, but the physical meanings of the macroscopic constitutive relation and the structural-sensitive character are difficult to understand. Therefore, by taking advantage of both fields of study, I introduce here a general fluid-dynamical formulation with a detailed explanation of the macroscopic constitutive relation. The general formulation includes as a special case the theory of linear poroelasticity. The governing equations introduced in Sect. "General Theoretical Framework to Describe the Dynamics of Solid–Liquid Composite Systems" include several structural-sensitive parameters that are given as functions of liquid volume fraction and pore geometry. Provided an evolution equation of pore geometry is obtained in future studies, it will be possible to investigate the interaction between pore geometry and macroscopic dynamics by solving the governing equations including the constitutive relation together with the evolution equation.

In Sect. "Overview of Applications", a brief summary of the various applications of the general theory introduced in Sect. "General Theoretical Framework to Describe the Dynamics of Solid–Liquid Composite Systems" is presented. In Sects. "Derivation of Wave Equations" to "Dispersion and Attenuation of Waves in Solid–Liquid Composite Systems", the theory is applied to the elastic wave propagation in a solid-liquid composite system; we linearize the general formulation and derive the wave

equations (Sect. "Derivation of Wave Equations"). Based on the detailed descriptions of V_P and V_S obtained from the wave equations, the effects of liquid volume fraction, pore geometry, and liquid compressibility on the velocities are summarized systematically (Sects. "Porosity and Pore Shape" to "Determinability of Porosity and Pore Shape from Elastic Wave Velocities"). To assess the determinability of pore geometry, the usual forward modeling based on a priori assumed pore geometries, e. g. [15,17,18,27,41] are not enough, and a systematic treatment of general pore geometries is required, which is enabled by the introduction of the concept of equivalent aspect ratio (Sect. "Porosity and Pore Shape"). The determinability of porosity and pore geometry from seismic tomographic data is discussed in Sect. "Application to Seismic Tomographic Images". In Sect. "Dispersion and Attenuation of Waves in Solid–Liquid Composite Systems", so as to clarify the limitation in the application of the theoretical results, I briefly discuss the attenuation and dispersion of elastic waves. In this article, the term 'liquid' is used with the same meaning as 'fluid'. Hence, 'liquid' in this paper includes 'gas'.

General Theoretical Framework to Describe the Dynamics of Solid–Liquid Composite Systems

A schematic illustration of a solid-liquid composite system considered in this article is shown in Fig. 1a. Length scales less than the grain size are referred to as "microscopic"; length scales greater than the grain size are referred to as "macroscopic". Solid-liquid composite systems are characterized by large differences in mechanical properties between solid and liquid phases, and hence the stress and velocity fields that develop under external forces are usually highly heterogeneous at the microscopic scale. However, when studying macroscopic dynamics such as mantle-scale melt segregation and propagation of seismic waves with wavelengths much larger than the grain size, it is not practical to solve both the microscopic and macroscopic processes simultaneously. In this section, I review the theoretical framework to treat the macroscopic dynamics separately from the microscopic processes. In this theory, macroscopic dynamics of solid-liquid composite systems are described within the framework of continuum mechanics, using the macroscopic variables obtained by averaging the microscopic fields. The averages within the solid and liquid phases are taken separately, so that the theory can be applied to the phenomena involving a relative motion between the two phases. Although microscopic variables do not explicitly appear in the governing equations, several parameters included in these equations are sensitive to the microstructures and thus the microscopic processes do affect the macroscopic dynamics through these parameters. In this review, a special emphasis is given to these structurally sensitive parameters.

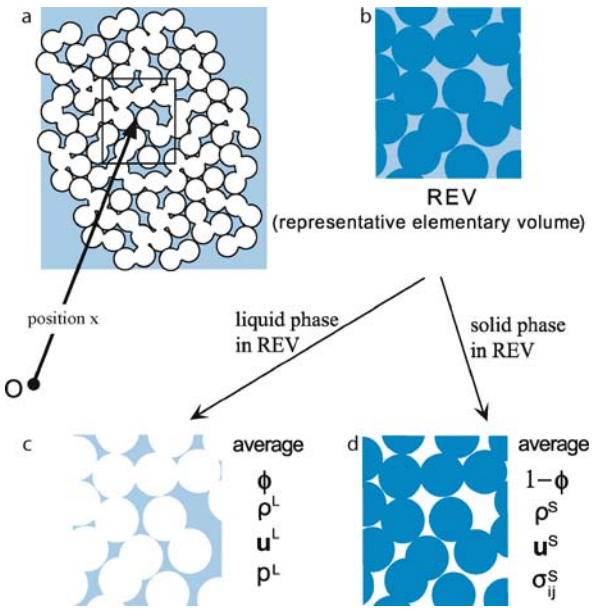

Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in, Figure 1

a A solid-liquid composite system with a representative elementary volume (REV) at position x. b REV consisting of solid (*thick blue*) and liquid (*light blue*) phases. c Liquid phase in REV and the averaged quantities representing the macroscopic mechanical state of the liquid. d Solid phase in REV and the averaged quantities representing the macroscopic mechanical state of the solid

Macroscopic Variables

We consider a region called REV (representative elementary volume), which is small enough to consider as a point in the macroscopic scale but large enough to contain a number of solid grains. Macroscopic quantities are defined by the average of the corresponding quantities in REV, where the averages within the liquid phase and those within the solid phase are taken separately. The averaging procedure can be defined by using a window function W and phase function A. By considering the REV as a cuboid with edge length L_i, the window function $W(\mathbf{x})$ takes the value 1 at $-L_i/2 \le x_i \le L_i/2$ $(i = x, y, z)$ and value 0 otherwise. The phase function $A(\mathbf{x})$ takes the value 1 if the position $\mathbf{x}$ is in the liquid phase, and value 0 in the solid phase; A and $1 - A$ quantify the properties of the liquid and solid phases, respectively. Let $a(\mathbf{x}, t)$ be a microscopic

field of the physical quantity a, which is a function of the position $\mathbf{x}$ and time t. The phasic average of a in the liquid phase (a^{L}) or solid phase (a^{S}) is defined by

$$\begin{aligned} a^{\mathrm{L}}(\mathbf{x},t) &= \frac{1}{V\phi}\int W(\boldsymbol{\xi}-\mathbf{x})\,A(\boldsymbol{\xi},t)\,a(\boldsymbol{\xi},t)\mathrm{d}V_{\xi} \\ a^{\mathrm{S}}(\mathbf{x},t) &= \frac{1}{V(1-\phi)}\int W(\boldsymbol{\xi}-\mathbf{x})\,(1-A(\boldsymbol{\xi},t)) \\ &\quad \cdot a(\boldsymbol{\xi},t)\mathrm{d}V_{\xi}\,, \end{aligned} \tag{1}$$

where $V(=L_xL_yL_z)$ represents the volume of REV and ϕ represents the liquid volume fraction, defined as

$$\phi(\mathbf{x},t) = \frac{1}{V}\int W(\boldsymbol{\xi}-\mathbf{x})\,A(\boldsymbol{\xi},t)\mathrm{d}V_{\xi}\,. \tag{2}$$

The volume integrals in Eqs. (1)–(2) are taken over whole of the solid-liquid system.

The dynamic state of the solid-liquid composite system at a spatially-fixed position $\mathbf{x}$ is described by the following 7 macroscopic variables defined by Eqs. (1) and (2):

$\phi(\mathbf{x},t)$....liquid volume fraction (nondimensional)

$\rho^{\mathrm{L}}(\mathbf{x},t)$....density of liquid (kg/m^3)

$\rho^{\mathrm{S}}(\mathbf{x},t)$....density of solid (kg/m^3)

$\mathbf{u}^{\mathrm{L}}(\mathbf{x},t)$....displacement of liquid (m)

$\mathbf{u}^{\mathrm{S}}(\mathbf{x},t)$....displacement of solid (m)

$p^{\mathrm{L}}(\mathbf{x},t)$....liquid pressure (Pa), with compression positive

$\sigma^{\mathrm{S}}_{ij}(\mathbf{x},t)$....solid stress (Pa), with tension positive.

It is rigorous to define $\mathbf{u}^{\mathrm{L}}$ and $\mathbf{u}^{\mathrm{S}}$ by using mass-weighted average [7]. However, for simplicity, the density heterogeneity within each phase is assumed to be small and the mass-weighted average is approximated by the phasic average.

Governing Equations

The seven variables introduced in Sect. "Macroscopic Variables" are governed by the following seven equations:

mass conservation of liquid

$$\frac{\partial(\phi\rho^{\mathrm{L}})}{\partial t} + \nabla\cdot(\phi\rho^{\mathrm{L}}\dot{\mathbf{u}}^{\mathrm{L}}) = \Gamma \tag{A}$$

mass conservation of solid

$$\frac{\partial\left\{(1-\phi)\rho^{\mathrm{S}}\right\}}{\partial t} + \nabla\cdot\left\{(1-\phi)\rho^{\mathrm{S}}\dot{\mathbf{u}}^{\mathrm{S}}\right\} = -\Gamma \tag{B}$$

intrinsic constitutive relation of liquid

$$\frac{\delta\rho^{\mathrm{L}}}{\rho^{\mathrm{L}}} = \frac{1}{k_{\mathrm{L}}}\delta p^{\mathrm{L}} \tag{C}$$

intrinsic constitutive relation of solid

$$\frac{\delta\rho^{\mathrm{S}}}{\rho^{\mathrm{S}}} = \frac{1}{k_{\mathrm{S}}}\delta p^{\mathrm{S}} \tag{D}$$

macroscopic constitutive relation of solid framework

$$\epsilon_{ij} = S_{ijkl}\left(\sigma^{\mathrm{S}}_{kl} + p^{\mathrm{L}}\delta_{kl}\right) - \frac{1}{3k_{\mathrm{S}}}p^{\mathrm{L}}\delta_{ij} \tag{E}$$

linear momentum conservation of liquid

$$\phi\rho^{\mathrm{L}}\ddot{\mathbf{u}}^{\mathrm{L}} = -\nabla(\phi p^{\mathrm{L}}) + \phi\rho^{\mathrm{L}}\mathbf{g} + \mathbf{I} \tag{F}$$

linear momentum conservation of solid

$$(1-\phi)\rho^{\mathrm{S}}\ddot{\mathbf{u}}^{\mathrm{S}} = \nabla\cdot\{(1-\phi)\sigma^{\mathrm{S}}\} + (1-\phi)\rho^{\mathrm{S}}\mathbf{g} - \mathbf{I}\,, \tag{G}$$

where $\mathbf{I}$ (N/m^3) in Eqs. (F) and (G) represents the interaction between the solid and liquid phases (the force applied to the liquid from the solid is taken positive), and is explicitly written as

$$\mathbf{I} = -\frac{\eta_{\mathrm{L}}\phi^2}{k_\phi}\left(\dot{\mathbf{u}}^{\mathrm{L}} - \dot{\mathbf{u}}^{\mathrm{S}}\right) + p^{\mathrm{L}}\nabla\phi\,. \tag{3}$$

The Γ (kg/s/m^3) in Eqs. (A) and (B) represents the net mass flux from the solid to the liquid, k_{L} and k_{S} (Pa) in Eqs. (C), (D), and (E) represent the intrinsic bulk moduli of the liquid and solid, respectively, $p^{\mathrm{S}} = -(\sigma^{\mathrm{S}}_{xx} + \sigma^{\mathrm{S}}_{yy} + \sigma^{\mathrm{S}}_{zz})/3$ (Pa) in Eq. (D) represents the solid pressure (compression positive), ϵ_{ij} in the left hand side of Eq. (E) represents the framework strain (extension positive), whose definition is given in Sect. "Equation (E)", S_{ijkl} (Pa^{-1}) in Eq. (E) represents the elastic compliance tensor of the solid framework, $\mathbf{g}$ (N/kg) in Eqs. (F) and (G) represents the gravitational acceleration vector, and η_{L} (Pa s) and k_ϕ (m^2) in Eq. (3) represent the liquid viscosity and permeability, respectively. For $\alpha = S, L$, $\dot{\mathbf{u}}^{\alpha} = \mathrm{D}\,\mathbf{u}^{\alpha}/\mathrm{D}\,t$ and $\ddot{\mathbf{u}}^{\alpha} = \mathrm{D}\,\dot{\mathbf{u}}^{\alpha}/\mathrm{D}\,t$ represent velocity and acceleration vectors, respectively, where $\mathrm{D}/\mathrm{D}\,t = \partial/\partial t + \dot{\mathbf{u}}^{\alpha}\cdot\nabla$. Also, $\nabla\cdot\sigma = \partial\sigma_{ij}/\partial x_j = \sigma_{ij,j}$. The summation convention for repeated subscripts is employed.

These equations, except for Eq. (E), are rigorously derived by averaging the mass and linear momentum conservation equations and constitutive relations required in the microscopic scale [7]. The physical meanings of Eqs. (A)–(D) are made clear by analogy with the standard fluid dynamic equations. The term Γ included in Eqs. (A)–(B) is zero unless melting/solidification or dissolution/precipitation occurs. The parameters that are sensitive to the microstructures are S_{ijkl} and k_ϕ included in Eqs. (E)–(G). Therefore, in the following part of this

section, detailed discussions of these three equations are given. In Sect. "Semi-Intuitive Derivations of Equations (E), (F), and (G)", derivations of Eqs. (E)–(G) are given in a semi-intuitive manner to clarify the physical meaning of these equations and the structural-sensitive parameters. Several important aspects of these equations are also discussed in Sects. "Relation to the Effective Medium Theory"–"Fundamental Assumption for Stress Heterogeneity".

Semi-Intuitive Derivations of Equations (E), (F), and (G)

Equations (F) and (G) Equations (F) and (G) describe the linear momentum conservations for the liquid and solid phases, respectively, in the REV. The left hand sides (LHS) of these equations represent the acceleration terms. The right hand sides (RHS) represent the total forces applied on each system, including the body force and the surface force. The surface force is applied through the boundary surface of each system; this boundary is made of the boundary on the surface of REV (black boundary in Figs. 2a and 2b) and the boundary with the other phase (red boundary in Figs. 2a and 2b). The former boundary (black) exists within each phase and the latter boundary (red) exists on the phase boundary. In the RHS of Eq. (F) or (G), the second term represents the body force due to gravity, and the first and the third terms represent the surface forces applied through the black and red boundaries, respectively. The surface force through the red boundary is the interaction between solid and liquid, and hence the third term in the RHS of Eq. (F), **I**, and that of Eq. (G), −**I**, are of the same magnitude and of opposite sign.

Equation (3) shows that interaction **I** consists of two terms, corresponding to the contributions from the traction components tangential and normal to the phase boundary. The first term, corresponding to the tangential component, represents the viscous drag force proportional to the relative velocity of the two phases. The permeability k_ϕ included in the proportionality constant depends on the detailed geometry of the liquid-filled pores. The second term in the RHS of Eq. (3) represents the contribution from the normal component of traction, which is determined by the liquid pressure. An intuitive explanation for the proportional dependence of this force on the porosity gradient is given in Fig. 3, in which a simple pore geometry is assumed. A more rigorous derivation of this term for general pore geometries is presented in Sect. "Fundamental Assumption for Stress Heterogeneity", where the assumption needed to derive this term is clarified and its validity is discussed. In the dynamics of solid-liquid composite systems, the motion of each phase is significantly affected by the interaction with the other phase through the phase boundary.

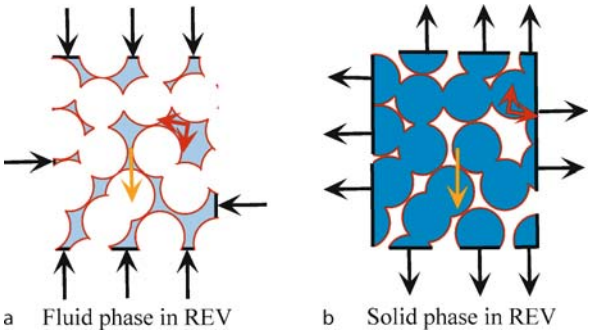

Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in, Figure 2
a Forces applied on the liquid phase in REV. b Forces applied on the solid phase in REV. Forces applied on each system consist of body force (*orange arrow*) and surface force (*black* and *red arrows*). The body force is due to gravity. The surface force is applied through the boundary surface of each system, which is divided into the boundary on the surface of REV (*black boundary*) and the boundary with the other phase (*red boundary*). The surface force applied through the *red boundary* is shown by dividing into normal and tangential components to the interface (*red arrows*). Solid stress is taken to be tension positive and liquid pressure is taken to be compression positive

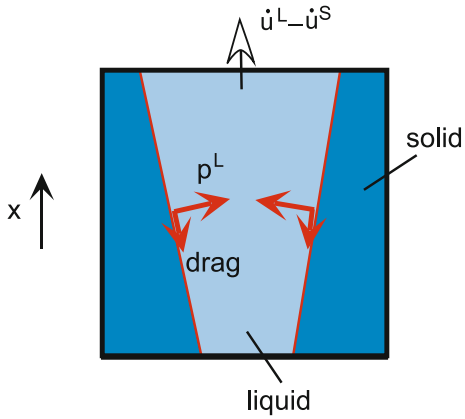

Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in, Figure 3
Interaction between the solid and liquid phases through the phase boundary (*red boundary*) is schematically shown for a simple pore geometry with a porosity gradient. The forces applied from solid to liquid are shown. The total force of the force component normal to the phase boundary does not vanish if porosity gradient is not zero, and that tangential to the phase boundary does not vanish if average velocity of liquid relative to solid is not zero

Note that Eqs. (F) and (G) are derived by implicitly assuming a connectivity of each phase (black boundary in

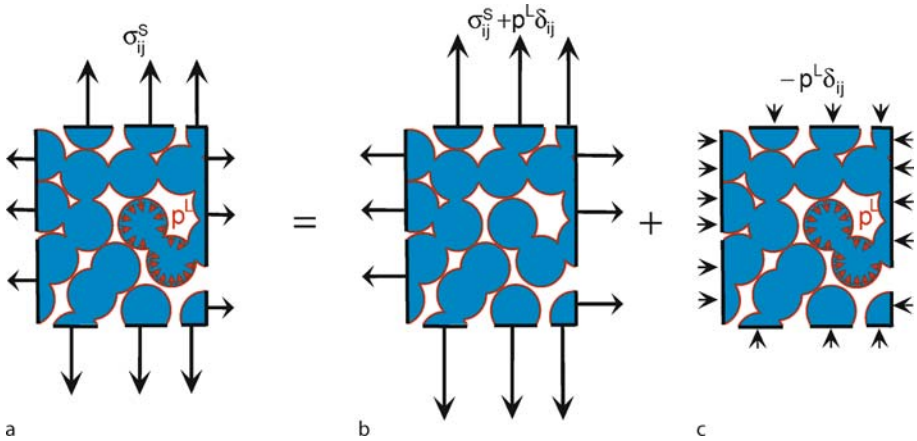

Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in, Figure 4
a A stress state of the solid phase in REV, generally represented by the solid stress σ^S_{ij} (tension positive) and liquid pressure p^L (compression positive), is expressed by the superposition of b effective stress state, defined by solid stress $\sigma^S_{ij} + p^L\delta_{ij}$ and liquid pressure 0, and c uniform stress state, defined by solid stress $-p^L\delta_{ij}$ and liquid pressure p^L. Equation (E) in the text states that the framework strain under a given stress state is obtained by the superposition of the framework strain under the effective stress state and that under the uniform stress state

Figs. 2a and 2b). Therefore, we need to be careful in applying the present theory to end-member systems of suspensions of solid or isolated inclusions of liquid, in which one phase is dispersed in the other phase without connectivity.

Equation (E) The framework strain ϵ_{ij} in the LHS of Eq. (E) is defined by using the macroscopic displacement of the solid $\mathbf{u}^S$ as

$$\epsilon_{ij} = \frac{1}{2}\left(\frac{\partial u^S_i}{\partial x_j} + \frac{\partial u^S_j}{\partial x_i}\right), \tag{4}$$

and hence ϵ_{ij} represents the macroscopic deformation of the solid framework. Equation (E) provides a macroscopic constitutive relationship between the framework strain and the macroscopic (averaged) stress state of the framework. The coefficient S_{ijkl} included in this equation is not only determined by the solid intrinsic properties, but shows a large structural sensitivity, which plays a significant role in the dynamics of solid-liquid composite systems. Equation (E) is one of the key equations characterizing the two-phase dynamics. I present here a semi-intuitive derivation of Eq. (E).

As shown in Fig. 4a, a macroscopic stress state of the solid framework is generally described by the solid stress σ^S_{ij} applied through the boundary on the surface of REV (black boundary) and liquid pressure p^L applied through the boundary with the liquid phase (red boundary). The framework strain under this stress state is calculated by the superposition of the framework strain under stress $\sigma^S_{ij} + p^L\delta_{ij}$ applied through the black boundary while leaving the red boundary as a free surface (Fig. 4b) plus the framework strain under uniform pressure p^L applied to all boundaries (Fig. 4c). The assumption of the superposition is valid when the response of the solid framework to the applied stresses is linear. The term $\sigma^S_{ij} + p^L\delta_{ij}$ is called the effective stress. Figure 4b shows that the elastic compliance tensor S_{ijkl} specifying the effect of effective stress on the framework strain describes the mechanical properties of the 'skeleton' (solid framework obtained by replacing the regions occupied by the liquid phase with a vacuum), which is not only determined by the intrinsic properties of solid but also strongly depends on the porosity and pore geometry. For a given solid-liquid composite system, S_{ijkl} is estimated by using experimental and/or modeling approaches, some of which are presented in a later section. The framework strain under a given pressure p^L applied to all boundaries (Fig. 4c) is identical to the strain of REV completely filled with solid ($\phi = 0$) under uniform pressure p^L, and hence is determined only by the intrinsic bulk modulus of the solid phase (the last term in the RHS of Eq. (E)).

In the derivation mentioned above, the solid phase is assumed to deform elastically. For an isotropic system, Eq. (E) is written as

$$\epsilon_{ij} = \frac{1}{3k_{sk}}\left(\frac{\sigma^S_{kk}}{3} + p^L\right)\delta_{ij} + \frac{1}{2\mu_{sk}}\left(\sigma^S_{ij} - \frac{\sigma^S_{kk}}{3}\delta_{ij}\right) - \frac{1}{3k_S}p^L\delta_{ij}, \tag{E$_e$}$$

where k_{sk} and μ_{sk} represent the bulk and shear moduli, respectively, of the skeleton. Similarly, when the solid phase deforms viscously, the macroscopic constitutive relation for an isotropic system is written as

$$\dot{\epsilon}_{ij} = \frac{1}{3\xi_{sk}}\left(\frac{\sigma^{\mathrm{S}}_{kk}}{3} + p^{\mathrm{L}}\right)\delta_{ij} + \frac{1}{2\eta_{sk}}\left(\sigma^{\mathrm{S}}_{ij} - \frac{\sigma^{\mathrm{S}}_{kk}}{3}\delta_{ij}\right), \tag{E$_{\mathrm{v}}$}$$

where ξ_{sk} and η_{sk} represent the bulk and shear viscosities, respectively, of the skeleton. Because viscous deformation is usually large in amplitude, the volumetric deformation due to the intrinsic compressibility of the solid phase, which corresponds to the 3rd term in the RHS of Eq. (E$_{\mathrm{e}}$), is neglected in Eq. (E$_{\mathrm{v}}$). Similar to k_{sk} and μ_{sk}, ξ_{sk} and η_{sk} are not only determined by the intrinsic property of the solid phase but also strongly depend on the porosity and pore geometry. Although the intrinsic compressibility of the solid phase is neglected in Eq. (E$_{\mathrm{v}}$), the volumetric component of the framework strain rate controlled by ξ_{sk} cannot be neglected. This is because even when the solid phase is made of incompressible material, the solid framework can change its volume by changing the porosity. This demonstrates the essential difference between the skeleton property and the intrinsic property of the solid. A more general description of the viscous constitutive relation is given by neglecting the last term in the RHS of Eq. (E) and replacing ϵ_{ij} and elastic compliance tensor S_{ijkl} by $\dot{\epsilon}_{ij}$ and viscous compliance tensor S^{v}_{ijkl}, respectively.

Relation to the Effective Medium Theory

The structural sensitivity of the skeleton properties S_{ijkl} plays an important role in the two-phase dynamics. To predict quantitatively the microstructural effects on the skeleton properties, an effective medium theory has been developed. However, the applicability of these theoretical results to Eq. (E) is not self-evident, because the definition of the macroscopic strain given by Eq. (4) is different from that of the average strain used in the effective medium theory. In the effective medium theory, it is well-known that the stress and strain fields of the solid phase are highly heterogeneous at the microscopic scale, so that the local stress can be largely different from the macroscopic stress. However, in the fluid dynamical formulation of the two-phase dynamics, this point is not emphasized and a confusion between microscopic and macroscopic stresses sometimes occurs (Sect. "Fundamental Assumption for Stress Heterogeneity"). Therefore, it is important to establish a connection between the fluid dynamical formulation and the effective medium theory.

When the skeleton properties are calculated in the effective medium theory, the effective properties of the solid-liquid composites are calculated under either drained conditions, in which the liquid pressure is kept constant, or under dry conditions, in which the liquid phase with zero bulk and shear moduli is kept under undrained conditions. In both cases, the space which remains after subtracting the solid phase from REV is considered to be filled with the pore phase. The microscopic displacement field in the pore phase can be obtained from the undrained solution by setting the bulk and shear moduli of the liquid to zero. The displacement field in the solid phase is continuously connected with that in the pore phase and there is no relative motion between the solid and pore phases. The effective properties under the dry (or drained) condition is given by

$$\varepsilon^{\mathrm{B}}_{ij} = S^{\mathrm{dry}}_{ijkl}\,\sigma^{\mathrm{B}}_{kl}\,, \tag{5}$$

where $\varepsilon^{\mathrm{B}}_{ij}$ and σ^{B}_{ij} represent the average strain and stress, respectively, over both solid and pore phases,

$$\begin{cases} \sigma^{\mathrm{B}}_{ij} &= (1-\phi)\sigma^{\mathrm{S}}_{ij} \\ \varepsilon^{\mathrm{B}}_{ij} &= (1-\phi)\varepsilon^{\mathrm{S}}_{ij} + \phi\varepsilon^{P}_{ij}\,, \end{cases} \tag{6}$$

the superscript "B" means bulk, and

$$\varepsilon^{\alpha}_{ij} = \frac{1}{2}\left\langle \frac{\partial u_i}{\partial \xi_j} + \frac{\partial u_j}{\partial \xi_i}\right\rangle_{\alpha} \quad (\alpha = S, P) \tag{7}$$

represents the phasic average of strain in the solid or pore phase, e. g. [21,50]. To clarify the relationship between S_{ijkl} and S^{dry}_{ijkl}, the relationship between ϵ_{ij} and $\varepsilon^{\mathrm{B}}_{ij}$ needs to be specified.

Let $\langle a\rangle_{\alpha}$ ($\alpha = S, L$) be the phasic average of the quantity a defined in Eqs. (1). Let $\Sigma^{\alpha}_{\mathrm{black}}$ ($\alpha = S, L$) be the boundary of the phase α on the surface of REV (black boundary in Figs. 2 and 4) and let Σ_{red} be the phase boundary in REV (red boundary in Figs. 2 and 4). Let n_i be the outward unit normal to these boundaries, where the positive direction at Σ_{red} is outward to the liquid phase. From the definition of phasic average $\langle\ \rangle_{\mathrm{S}}$, we obtain

$$\frac{\partial\langle u_i\rangle_{\mathrm{S}}}{\partial x_j} = \frac{\partial}{\partial x_j}\left[\frac{1}{V(1-\phi)}\int W(\boldsymbol{\xi}-\mathbf{x})(1-A(\boldsymbol{\xi}))u_i(\boldsymbol{\xi})\mathrm{d}V_{\xi}\right]$$

$$
\begin{aligned}
&= \frac{\partial \phi}{\partial x_j} \frac{1}{V(1-\phi)^2} \int W(\boldsymbol{\xi}-\mathbf{x})(1-A(\boldsymbol{\xi}))u_i(\boldsymbol{\xi})\mathrm{d}V_\xi \\
&\quad + \frac{1}{V(1-\phi)} \frac{\partial}{\partial x_j}\left[\int W(\boldsymbol{\xi}-\mathbf{x})(1-A(\boldsymbol{\xi}))u_i(\boldsymbol{\xi})\mathrm{d}V_\xi\right] \\
&= \frac{\langle u_i\rangle_\mathrm{S}}{1-\phi}\frac{\partial\phi}{\partial x_j} \\
&\quad + \frac{1}{V(1-\phi)} \int -\frac{\partial W(\boldsymbol{\xi}-\mathbf{x})}{\partial \xi_j}(1-A(\boldsymbol{\xi}))u_i(\boldsymbol{\xi})\mathrm{d}V_\xi \\
&= \frac{\langle u_i\rangle_\mathrm{S}}{1-\phi}\frac{\partial\phi}{\partial x_j} \\
&\quad - \frac{1}{V(1-\phi)} \int \frac{\partial\{W(\boldsymbol{\xi}-\mathbf{x})\cdot(1-A(\boldsymbol{\xi}))u_i(\boldsymbol{\xi})\}}{\partial \xi_j}\mathrm{d}V_\xi \\
&\quad + \frac{1}{V(1-\phi)} \int W(\boldsymbol{\xi}-\mathbf{x})(1-A(\boldsymbol{\xi}))\frac{\partial u_i(\boldsymbol{\xi})}{\partial \xi_j}\mathrm{d}V_\xi \\
&\quad + \frac{1}{V(1-\phi)} \int W(\boldsymbol{\xi}-\mathbf{x})\frac{\partial(1-A(\boldsymbol{\xi}))}{\partial \xi_j}u_i(\boldsymbol{\xi})\mathrm{d}V_\xi \\
&= \frac{\langle u_i\rangle_\mathrm{S}}{1-\phi}\frac{\partial\phi}{\partial x_j} + \left\langle\frac{\partial u_i}{\partial \xi_j}\right\rangle_\mathrm{S} \\
&\quad + \frac{1}{V(1-\phi)} \int W(\boldsymbol{\xi}-\mathbf{x})\frac{\partial(1-A(\boldsymbol{\xi}))}{\partial \xi_j}u_i(\boldsymbol{\xi})\mathrm{d}V_\xi \\
&= \frac{\langle u_i\rangle_\mathrm{S}}{1-\phi}\frac{\partial\phi}{\partial x_j} + \left\langle\frac{\partial u_i}{\partial \xi_j}\right\rangle_\mathrm{S} \\
&\quad + \frac{1}{V(1-\phi)} \int_{\Sigma_\mathrm{red}} u_i(\boldsymbol{\xi})n_j(\boldsymbol{\xi})\mathrm{d}S_\xi\,, \qquad (8)
\end{aligned}
$$

where the volume integral of factor $\partial\{(W(\boldsymbol{\xi}-\mathbf{x})(1-A(\boldsymbol{\xi}))u_i(\boldsymbol{\xi})\}/\partial\xi_i$ in the RHS of the 4th equation is converted to the surface integral of $W(\boldsymbol{\xi}-\mathbf{x})(1-A(\boldsymbol{\xi}))u_i(\boldsymbol{\xi})$ on the outermost boundary of the solid-liquid system, which is zero because $W(\boldsymbol{\xi}-\mathbf{x})$ is zero outside the REV.

Because the displacement field u_i of the solid phase can be continuously connected to u_i of the pore phase, by using Gauss's theorem, the integral over Σ_red in the third term on the RHS of the last equation of (8) can be rewritten in terms of the volume integral in the pore phase. Because the boundary of the pore phase is given by Σ_red and $\Sigma^\mathrm{L}_\mathrm{black}$, we thus obtain

$$
\begin{aligned}
&\frac{\partial\langle u_i\rangle_\mathrm{S}}{\partial x_j} \\
&= \left\langle\frac{\partial u_i}{\partial \xi_j}\right\rangle_\mathrm{S} + \frac{\langle u_i\rangle_\mathrm{S}}{1-\phi}\frac{\partial\phi}{\partial x_j} \\
&\quad + \frac{1}{V(1-\phi)} \int_{\Sigma_\mathrm{red}+\Sigma^\mathrm{L}_\mathrm{black}} u_i(\boldsymbol{\xi})n_j(\boldsymbol{\xi})\mathrm{d}S_\xi \\
&\quad - \frac{1}{V(1-\phi)} \int_{\Sigma^\mathrm{L}_\mathrm{black}} u_i(\boldsymbol{\xi})n_j(\boldsymbol{\xi})\mathrm{d}S_\xi \\
&= \left\langle\frac{\partial u_i}{\partial \xi_j}\right\rangle_\mathrm{S} + \frac{\langle u_i\rangle_\mathrm{S}}{1-\phi}\frac{\partial\phi}{\partial x_j} \\
&\quad + \frac{\phi}{(1-\phi)}\frac{1}{V\phi} \int W(\boldsymbol{\xi}-\mathbf{x})A(\boldsymbol{\xi})\frac{\partial u_i}{\partial \xi_j}\mathrm{d}V_\xi \\
&\quad - \frac{1}{(1-\phi)}\frac{\partial(\phi\langle u_i\rangle_\mathrm{S})}{\partial x_j} \\
&= \left\langle\frac{\partial u_i}{\partial \xi_j}\right\rangle_\mathrm{S} + \frac{\phi}{1-\phi}\left\langle\frac{\partial u_i}{\partial \xi_j}\right\rangle_\mathrm{P} - \frac{\phi}{1-\phi}\frac{\partial\langle u_i\rangle_\mathrm{S}}{\partial x_j}\,, \qquad (9)
\end{aligned}
$$

where u_i at the surface of REV ($\Sigma^\mathrm{L}_\mathrm{black}$ and $\Sigma^\mathrm{S}_\mathrm{black}$) is considered to be the same between the solid and pore phases, because there is no relative motion between these two phases. From Eqs. (4) and (9), we obtain

$$
\begin{aligned}
\epsilon_{ij} &= (1-\phi)\varepsilon^\mathrm{S}_{ij} + \phi\varepsilon^P_{ij} \\
&= \varepsilon^\mathrm{B}_{ij}\,. \qquad (10)
\end{aligned}
$$

Therefore, a simple conversion formula can be obtained linking S^dry_{ijkl} and S_{ijkl}

$$
S^\mathrm{dry}_{ijkl} = (1-\phi)S_{ijkl}\,. \qquad (11)
$$

We further discuss the peculiarity of Eq. (E) that it cannot be derived by the phasic average of the microscopic constitutive relationship. Because the microscopic stress and strain fields in the solid phase are related to each other by the intrinsic constitutive relationship of the solid phase, we may take the average of this relationship over the solid phase in REV, to obtain

$$
\varepsilon^\mathrm{S}_{ij} = S^\mathrm{intrinsic}_{ijkl}\sigma^\mathrm{S}_{kl}\,, \qquad (12)
$$

where $S^\mathrm{intrinsic}_{ijkl}$ represents the solid intrinsic properties. By comparing Eq. (E) to Eq. (12), it is apparent that the significant difference between S_{ijkl} and $S^\mathrm{intrinsic}_{ijkl}$ comes from the significant difference between ϵ_{ij} and $\varepsilon^\mathrm{S}_{ij}$ in that ϵ_{ij} represents not only $\varepsilon^\mathrm{S}_{ij}$ but also ε^P_{ij}. From the definition of ϕ,

$$
\begin{aligned}
\frac{\partial\phi}{\partial x_i} &= \frac{\partial}{\partial x_i}\left[\frac{1}{V}\int W(\boldsymbol{\xi}-\mathbf{x})A(\boldsymbol{\xi})\mathrm{d}V_\xi\right] \\
&= \frac{-1}{V}\int W(\boldsymbol{\xi}-\mathbf{x})n_i(\boldsymbol{\xi})\mathrm{d}S_\xi\,. \qquad (13)
\end{aligned}
$$

Then, from Eqs. (4), (8), and (13), ϵ_{ij} can be rewritten as

$$
\epsilon_{ij} = \varepsilon^\mathrm{S}_{ij} + \frac{\int_{\Sigma_\mathrm{red}}\left((u_i(\boldsymbol{\xi})-\langle u_i\rangle_\mathrm{S})\,n_j(\boldsymbol{\xi}) + \left(u_j(\boldsymbol{\xi})-\langle u_j\rangle_\mathrm{S}\right)n_i(\boldsymbol{\xi})\right)\mathrm{d}S_\xi}{2V(1-\phi)}\,. \qquad (14)
$$

Thus, a significant difference between ϵ_{ij} and $\varepsilon^\mathrm{S}_{ij}$ implies a significant contribution from the second term in the RHS

of (14). This means that the microscopic displacement field of the solid phase at the boundary with the liquid can be systematically different from the average displacement of the solid phase. An example of such systematic deviation can be seen in the compaction of the solid framework, where a contraction of the solid phase observed macroscopically is compensated by a displacement of the solid-liquid phase boundary into the pore space.

Fundamental Assumption for Stress Heterogeneity

A fundamental assumption implicitly used in formulating the dynamics of solid-liquid composite systems is that the microscopic stress field is homogeneous in the liquid phase but can be heterogeneous in the solid phase. This point is rarely stated explicitly. Here, I discuss this assumption first with respect to the liquid phase and then with respect to the solid phase.

The second term in the RHS of Eq. (3) represents the total force due to the normal traction component applied to the liquid phase through the phase boundary (red boundary). This term is derived as

$$\begin{aligned}
&\frac{-1}{V}\int_{\Sigma_{\mathrm{red}}} p(\boldsymbol{\xi})n_i(\boldsymbol{\xi})\mathrm{d}S_\xi \\
&= \frac{-1}{V}\int_{\Sigma_{\mathrm{red}}+\Sigma^{\mathrm{L}}_{\mathrm{black}}} p(\boldsymbol{\xi})n_i(\boldsymbol{\xi})\mathrm{d}S_\xi \\
&\quad + \frac{1}{V}\int_{\Sigma^{\mathrm{L}}_{\mathrm{black}}} p(\boldsymbol{\xi})n_i(\boldsymbol{\xi})\mathrm{d}S_\xi \\
&= \frac{-1}{V}\int_{\mathrm{REV}} A(\boldsymbol{\xi})\frac{\partial p(\xi)}{\partial \xi_i}\mathrm{d}V_\xi + \frac{1}{V}\int_{\Sigma^{\mathrm{L}}_{\mathrm{black}}} p(\boldsymbol{\xi})n_i(\boldsymbol{\xi})\mathrm{d}S_\xi \\
&= -\phi\left\langle\frac{\partial p}{\partial \xi_i}\right\rangle_{\mathrm{L}} + \frac{\partial(\langle p\rangle_{\mathrm{L}}\phi)}{\partial x_i} \\
&= -\phi\frac{\partial\langle p\rangle_{\mathrm{L}}}{\partial x_i} + \frac{\partial(\langle p\rangle_{\mathrm{L}}\phi)}{\partial x_i} \\
&= p^{\mathrm{L}}\frac{\partial\phi}{\partial x_i}\,,
\end{aligned} \tag{15}$$

where the relationship

$$\left\langle\frac{\partial p}{\partial \xi_i}\right\rangle_{\mathrm{L}} = \frac{\partial\langle p\rangle_{\mathrm{L}}}{\partial x_i} \tag{16}$$

is assumed to obtain the 4th equation. The validity of this assumption is checked as follows. In the same manner as Eq. (8), we obtain

$$\frac{\partial\langle p\rangle_{\mathrm{L}}}{\partial x_i} = \frac{-\langle p\rangle_{\mathrm{L}}}{\phi}\frac{\partial\phi}{\partial x_i} + \left\langle\frac{\partial p}{\partial \xi_i}\right\rangle_{\mathrm{L}} - \frac{1}{V\phi}\int_{\Sigma_{\mathrm{red}}} p(\boldsymbol{\xi})n_i(\boldsymbol{\xi})\mathrm{d}S_\xi\,. \tag{17}$$

From Eqs. (13) and (17),

$$\frac{\partial\langle p\rangle_{\mathrm{L}}}{\partial x_i} = \left\langle\frac{\partial p}{\partial \xi_i}\right\rangle_{\mathrm{L}} - \frac{1}{V\phi}\int_{\Sigma_{\mathrm{red}}} \left(p(\boldsymbol{\xi}) - \langle p\rangle_{\mathrm{L}}\right) n_i(\boldsymbol{\xi})\mathrm{d}S_\xi \tag{18}$$

is obtained. Equation (18) shows that the phasic average $\langle\ \rangle_L$ is not exchangeable with the differential operator, and that Eq. (16) is not valid if the 2nd term in the RHS of Eq. (18) is non-negligible, that is, if the liquid pressure at the boundary with the solid phase is systematically different from the average. In solid-liquid composite systems, the stress heterogeneity within the liquid phase relaxes quickly and the liquid pressure is usually considered to be uniform in REV. Therefore, the 2nd term in the RHS of (18) is considered to be negligible. This also confirms the validity of an assumption implicitly used in Figs. 3 and 4: because the liquid pressure is homogeneous at the microscopic scale, then from the continuity of stress, the normal compressive stress of the solid phase at the solid-liquid interface is equal to the macroscopic liquid pressure p^L.

The inexchangeability between the differential operator and phasic average is described by Eq. (14) for the solid phase and by Eq. (18) for the liquid phase. Although the second term in the RHS of Eq. (14) was considered to be non-negligible, the corresponding term in Eq. (18) was considered to be negligible. The different treatments applied to the two phases are based on the fact that the stress heterogeneity within the liquid phase relaxes quickly but that the stress heterogeneity within the solid phase does not relax (elastic solid phase) or relaxes much more slowly (viscous solid phase). As can be seen from Fig. 4b, under nonzero effective stress, the traction applied to the surface of each solid grain is significantly different between the areas in contact with the liquid phase and the areas in contact with the neighboring grains, indicating that the microscopic stress field in each grain is highly heterogeneous. Such a heterogeneous stress field causes a systematic deviation of the microscopic displacement at the boundary with the liquid, making the second term in the RHS of Eq. (14) non-negligible.

For the solid phase, it is therefore important to recognize a possible difference between local and macroscopic stresses. However, there seems to be a confusion in some studies considering the effect of the solid-liquid interfacial tension γ_{sl}. When γ_{sl} is taken into account, the stress continuity condition required at the solid-liquid interface is replaced by the Laplace condition, e. g. [38]. Hence, the solid stress used in the Laplace condition is the local stress, which is locally determined by p^L, γ_{sl}, and interfacial mean curvature, regardless of the macroscopic solid stress σ^{S}_{ij}. I emphasize this point because in the previous studies,

several questionable results were obtained due to the improper use of σ_{ij}^{S} in the Laplace condition, e. g. [38,46].

Overview of Applications

The theoretical framework introduced in Sect. "General Theoretical Framework to Describe the Dynamics of Solid–Liquid Composite Systems" is applicable to various geophysical and geological phenomena occurring in solid-liquid two-phase systems. Typical applications are summarized in Fig. 5 by classifying processes into three categories based on their time scales. The propagation of elastic waves is investigated using the wave equations derived from Eqs. (A)–(G). The application to elastic wave propagation is discussed in detail in the following sections. The time-dependent evolution of a perturbation in liquid pressure in the porous media, and the interaction between liquid pressure and matrix deformation or fracture, have long been of interest in investigations of the possible occurrence of earthquakes due to dehydration, e. g. [51] and earthquake triggering, e. g. [16]. These processes have much longer time scales than the periods of elastic waves, and the acceleration terms in Eqs. (F) and (G), which play an essential role in the wave equations, are negligible compared to the viscous drag force included in the interaction term **I**. Then, as is well known, the evolution equation for liquid pressure is derived in the form of a diffusion equation, where the diffusivity is given by $D = k_\phi k_L/(\eta_L \phi)$ (hydraulic diffusivity).

Melt segregation from a partially molten mantle has been of great interest in volcanology, petrology, and geochemistry. This process occurs over much longer time scales than the processes mentioned above, and involves large viscous deformations. Accordingly, the intrinsic compressibilities of the constituent materials are neglected. By applying the governing equations (A)–(D), (E_v), and (F)–(G) to a one-dimensional column of partially molten mantle with a homogeneous porosity distribution at $t = 0$ ($\phi = \phi_0$ at $z \geq -H$ and $\phi = 0$ at $z < -H$), the initial stage of melt segregation was solved analytically [20]. In most of the column, the buoyancy force $(1-\phi)(\rho^S - \rho^L)g$ is in equilibrium with the viscous drag force $\eta_L \phi(\dot{u}_z^L - \dot{u}_z^S)/k_\phi$, while within the compaction length δ_c from the bottom ($-H \leq z < -H + \delta_c$), the buoyancy force balances with the compaction resistance of the solid framework. Therefore, when the compaction length $\delta_c = \sqrt{k_\phi(1-\phi)(\xi_{sk} + 4\eta_{sk}/3)/\eta_L}$ is smaller than H, which is the case for the mantle, the segregation velocity is determined by the permeability k_ϕ. The permeability control is more clear in the steady-state model, in which the steady-state porosity structure develops to satisfy the balance between melt production rate and melt segregation rate, which is often assumed for the decompressional melting of the upwelling mantle below a ridge [30]. However, several instabilities inherent to the solid-liquid systems that result in a time-dependent evolution of the porosity distribution have also been reported. These include the propagation of melt as solitary waves or porosity waves, e. g. [3,34], and the unstable evolution of the perturbation in melt fraction under pure shear deformation of the solid matrix [39], in which both melt migration and matrix deformation are involved. These phe-

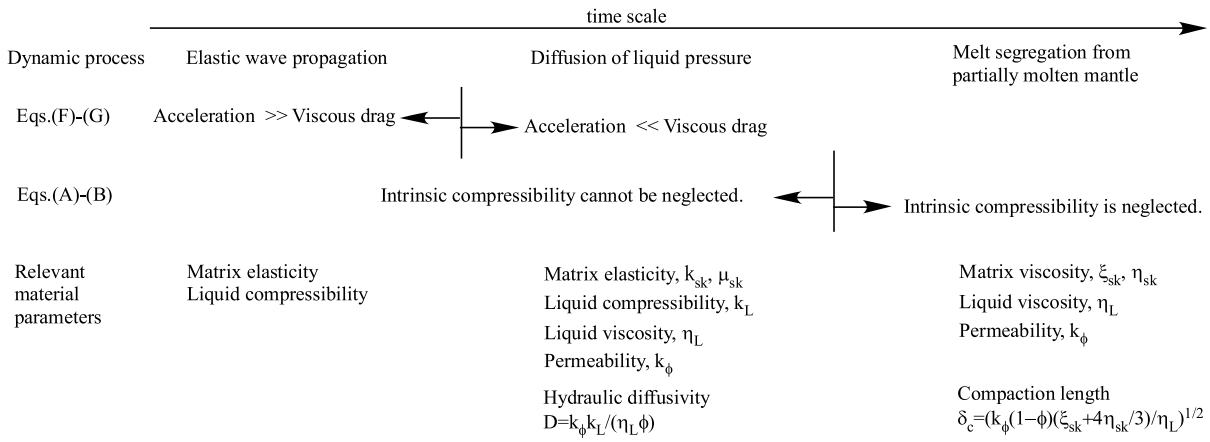

Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in, Figure 5
Dynamic processes in solid-liquid composite systems. Typical assumptions adopted in applying the governing equations (A)–(G) in the text to these processes are shown, with the material properties relevant to the processes

nomena are described by Eqs. (A)–(D), (E_v), and (F)–(G), where, in the solitary or porosity waves, the nonlinearity caused by the dependence of permeability on melt fraction ($k_\phi \propto \phi^{n>1}$) plays an essential role and, in the latter instability, the dependence of η_{sk} on ϕ plays an essential role. The possible occurrence of these instabilities has been of great interest, because the melt ascending velocity is significantly affected by the spatial distribution of porosity. The melt velocity and its spatial distribution determine the degree of chemical interaction between the melt and host rocks, and thus influence the major and trace element compositions. The microstructural dependences of permeability and viscous properties of the partially molten rocks affecting these instabilities are poorly understood and are the subject of future studies. The basic framework introduced in Sect. "General Theoretical Framework to Describe the Dynamics of Solid–Liquid Composite Systems" can be further extended to take into account additional factors, such as the chemical interaction between the fluid and host rocks, e. g. [1,36] or interfacial tension e. g. [32].

Elastic Wave Propagation in a Solid–Liquid Composite System

Derivation of Wave Equations

In this section, the general theoretical framework introduced in Sect. "General Theoretical Framework to Describe the Dynamics of Solid–Liquid Composite Systems" is applied to the elastic wave propagation in a solid-liquid composite system. The governing equations are linearized by considering the infinitesimal strain and displacement of a macroscopically homogeneous medium, and the wave equations are derived. The linearized equations are shown to be equivalent to the basic equations used in the theory of linear poroelasticity.

When we consider a macroscopically homogeneous solid-liquid composite system, the spatial and temporal variations in ϕ, ρ^L, and ρ^S are caused by the displacements $\mathbf{u}^L$ and $\mathbf{u}^S$. Therefore, if $\mathbf{u}^L$ and $\mathbf{u}^S$ are infinitesimally small, such terms as $\dot{\mathbf{u}}^L \cdot \nabla(\phi\rho^L)$ and $\dot{\mathbf{u}}^S \cdot \nabla((1-\phi)\rho^S)$ are negligible as higher-order terms. Under these approximations, by substituting Eqs. (C) and (D) into (A)/ρ^L+(B)/ρ^S, and taking the time integration, we obtain

$$-\phi\,\nabla\cdot(\mathbf{u}^L-\mathbf{u}^S) = \frac{\phi}{k_L}p^L + \frac{1-\phi}{k_S}p^S + \nabla\cdot\mathbf{u}^S. \quad (19)$$

By using Eq. (E_e), $\nabla\cdot\mathbf{u}^S = \epsilon_{kk}$ in Eq. (19) can be expressed in terms of stresses. Then, Eq. (19) and Eq. (E_e) are written as

$$\phi\,\nabla\cdot(\mathbf{u}^L-\mathbf{u}^S) = \phi\left(\frac{1}{k_S}-\frac{1}{k_L}\right)p^L + (1-\phi)\left(\frac{1}{K_b}-\frac{1}{k_S}\right)(p^S-p^L) \quad (20)$$

$$\epsilon_{ij} = \frac{(1-\phi)}{2N}\left(\sigma^S_{ij} - \frac{\sigma^S_{kk}}{3}\delta_{ij}\right) - \frac{(1-\phi)}{3K_b}(p^S-p^L)\delta_{ij} - \frac{1}{3k_S}p^L\delta_{ij}, \quad (21)$$

where K_b and N represent the bulk and shear moduli, respectively, of the skeleton and are related to k_{sk} and μ_{sk} introduced in Eq. (E_e) as $K_b = (1-\phi)k_{sk}$ and $N = (1-\phi)\mu_{sk}$. In the theory of linear poroelasticity, K_b and N, rather than k_{sk} and μ_{sk}, are commonly used. Similarly, by substituting Eq. (3) into Eqs. (F) and (G), and neglecting the effect of gravity, we obtain

$$\phi\rho^L\ddot{\mathbf{u}}^L = -\phi\nabla\cdot p^L - \frac{\eta_L\phi^2}{k_\phi}(\dot{\mathbf{u}}^L-\dot{\mathbf{u}}^S) \quad (22)$$

$$(1-\phi)\rho^S\ddot{\mathbf{u}}^S = (1-\phi)\nabla\cdot\sigma^S + \frac{\eta_L\phi^2}{k_\phi}(\dot{\mathbf{u}}^L-\dot{\mathbf{u}}^S)\,. \quad (23)$$

Equations (20)–(23) are equivalent to the basic equations used in the theory of linear poroelasticity. Compared to the framework introduced in Sect. "General Theoretical Framework to Describe the Dynamics of Solid–Liquid Composite Systems", the number of governing equations is reduced from 7 to 4, because the variables ρ^L and ρ^S are eliminated and ϕ in Eqs. (20)–(23) can be treated as constant. In the theory of linear poroelasticity, Eq. (20) is called the constitutive relation for the relative motion between the two phases, and this relation is usually introduced empirically, e. g. [48]. The present derivation from the more general framework shows that Eq. (20) is based on the requirement of mass conservation and intrinsic constitutive relations. When the acceleration terms are negligible, Eqs. (22) and (23) are further rewritten as

$$\phi(\dot{\mathbf{u}}^L-\dot{\mathbf{u}}^S) = -\frac{k_\phi}{\eta_L}\nabla p^L \quad (24)$$

$$\nabla\cdot\sigma^B = 0, \quad (25)$$

where Eq. (24) represents Darcy's law, and the bulk stress σ^B_{ij} represents $\sigma^B_{ij} = (1-\phi)\sigma^S_{ij} - \phi p^L\delta_{ij}$.

By eliminating pressures and stresses from Eqs. (20)–(23), we obtain

$$(1-\phi)\rho^S\ddot{\mathbf{u}}^S = P\,\nabla(\nabla\cdot\mathbf{u}^S) - N\,\nabla\times\nabla\times\mathbf{u}^S + Q\,\nabla(\nabla\cdot\mathbf{u}^L) + \frac{\eta_L\phi^2}{k_\phi}(\dot{\mathbf{u}}^L-\dot{\mathbf{u}}^S) \quad (26)$$

$$\phi\rho^{\mathrm{L}}\ddot{\mathbf{u}}^{\mathrm{L}} = Q\,\nabla(\nabla\cdot\mathbf{u}^{\mathrm{S}}) + R\,\nabla(\nabla\cdot\mathbf{u}^{\mathrm{L}}) - \frac{\eta_{\mathrm{L}}\phi^2}{k_\phi}(\dot{\mathbf{u}}^{\mathrm{L}} - \dot{\mathbf{u}}^{\mathrm{S}}), \quad (27)$$

where P, Q, and R are given by

$$\begin{cases} P = K_{\mathrm{b}} + \frac{4}{3}N + \dfrac{(1-\phi-\frac{K_{\mathrm{b}}}{k_{\mathrm{S}}})^2 k_{\mathrm{S}}}{1-\phi-\frac{K_{\mathrm{b}}}{k_{\mathrm{S}}}+\phi\frac{k_{\mathrm{S}}}{k_{\mathrm{L}}}} \\ Q = \dfrac{\phi(1-\phi-\frac{K_{\mathrm{b}}}{k_{\mathrm{S}}})k_{\mathrm{S}}}{1-\phi-\frac{K_{\mathrm{b}}}{k_{\mathrm{S}}}+\phi\frac{k_{\mathrm{S}}}{k_{\mathrm{L}}}} \\ R = \dfrac{\phi^2 k_{\mathrm{S}}}{1-\phi-\frac{K_{\mathrm{b}}}{k_{\mathrm{S}}}+\phi\frac{k_{\mathrm{S}}}{k_{\mathrm{L}}}} \,. \end{cases} \quad (28)$$

By taking the curl of Eqs. (26) and (27), and using the expressions $\Omega_{\mathrm{S}} = \nabla\times\mathbf{u}^{\mathrm{S}}$ and $\Omega_{\mathrm{L}} = \nabla\times\mathbf{u}^{\mathrm{L}}$, we obtain wave equations for the shear component;

$$\begin{cases} (1-\phi)\rho^{\mathrm{S}}\ddot{\Omega}_{\mathrm{S}} = N\nabla^2\Omega_{\mathrm{S}} + \dfrac{\eta_{\mathrm{L}}\phi^2}{k_\phi}(\dot{\Omega}_{\mathrm{L}} - \dot{\Omega}_{\mathrm{S}}) \\ \phi\rho^{\mathrm{L}}\ddot{\Omega}_{\mathrm{L}} = -\dfrac{\eta_{\mathrm{L}}\phi^2}{k_\phi}(\dot{\Omega}_{\mathrm{L}} - \dot{\Omega}_{\mathrm{S}}) \,. \end{cases} \quad (29)$$

By taking the divergence of Eqs. (26) and (27), and using the expressions $e_{\mathrm{S}} = \nabla\cdot\mathbf{u}^{\mathrm{S}}$ and $e_{\mathrm{L}} = \nabla\cdot\mathbf{u}^{\mathrm{L}}$, we obtain wave equations for the longitudinal component;

$$\begin{cases} (1-\phi)\rho^{\mathrm{S}}\ddot{e}_{\mathrm{S}} = P\nabla^2 e_{\mathrm{S}} + Q\nabla^2 e_{\mathrm{L}} + \dfrac{\eta_{\mathrm{L}}\phi^2}{k_\phi}(\dot{e}_{\mathrm{L}} - \dot{e}_{\mathrm{S}}) \\ \phi\rho^{\mathrm{L}}\ddot{e}_{\mathrm{L}} = Q\nabla^2 e_{\mathrm{S}} + R\nabla^2 e_{\mathrm{L}} - \dfrac{\eta_{\mathrm{L}}\phi^2}{k_\phi}(\dot{e}_{\mathrm{L}} - \dot{e}_{\mathrm{S}}). \end{cases} \quad (30)$$

The elastic wave propagation in a solid-liquid composite system was first formulated by Biot [5,6]. The wave equations (29)–(30) are almost the same as those obtained by Biot [5,6], except for the acceleration terms, which are slightly different. This is because the interaction **I** given in Eq. (3) does not take into account the effect of the relative acceleration between the solid and liquid phases. If an additional term proportional to the relative acceleration is added to the RHS of Eq. (3), then the same equations as Biot [5,6] can be derived. The proportionality constant attached to the relative acceleration is called tortuosity; tortuositym represents the deviation from the straight pore channel.

The elastic waves obtained by solving Eqs. (29)–(30) are dispersive and dissipative due to the relative motion between the solid and liquid phases. However, as shown below in Sect. "Dispersion and Attenuation of Waves in Solid–Liquid Composite Systems", the characteristic frequency for the dispersion and attenuation is much higher than the seismic frequency range. Therefore, in predicting the seismic wave velocities, the solutions obtained at the low-frequency limit are of importance. The wave solutions at the low-frequency limit do not involve relative motions, because the velocity terms, $\dot{\Omega}_{\mathrm{L}} - \dot{\Omega}_{\mathrm{S}}$ and $\dot{e}_{\mathrm{L}} - \dot{e}_{\mathrm{S}}$, if any, dominate the acceleration terms, and hence are not dispersive nor dissipative. The longitudinal and shear wave velocities at the low-frequency limit are obtained as

$$V_P = \sqrt{\frac{K_{\mathrm{b}} + \frac{4}{3}N + \frac{k_{\mathrm{S}}(1-K_{\mathrm{b}}/k_{\mathrm{S}})^2}{1-\phi-K_{\mathrm{b}}/k_{\mathrm{S}}+\phi k_{\mathrm{S}}/k_{\mathrm{L}}}}{\bar{\rho}}} \quad (31)$$

$$V_{\mathrm{S}} = \sqrt{\frac{N}{\bar{\rho}}}\,, \quad (32)$$

where $\bar{\rho} = (1-\phi)\rho^{\mathrm{S}} + \phi\rho^{\mathrm{L}}$ represents the average density of the medium. Without relative motion, neither permeability nor tortuosity affect the velocities. Therefore, Eqs. (31)–(32) are exactly the same as the results of Biot [5,6].

Porosity and Pore Shape

Because the bulk and shear moduli of the skeleton, K_b and N, included in Eqs. (31)–(32) depend not only on porosity but also on pore geometry, various models assuming various pore geometries have been developed to predict K_b and N quantitatively, e. g. [15,17,18,27,41]. The oblate spheroid model (Fig. 6a), tube model (Fig. 6b),

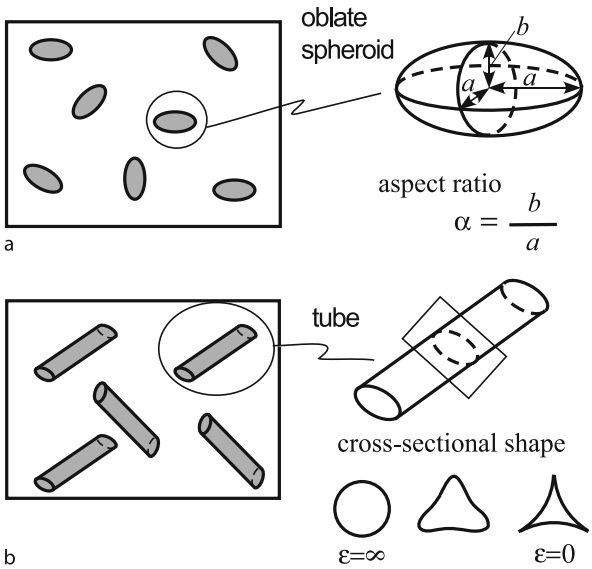

Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in, Figure 6
Inclusion models of a solid-liquid composite system. a Oblate spheroid model. b Tube model

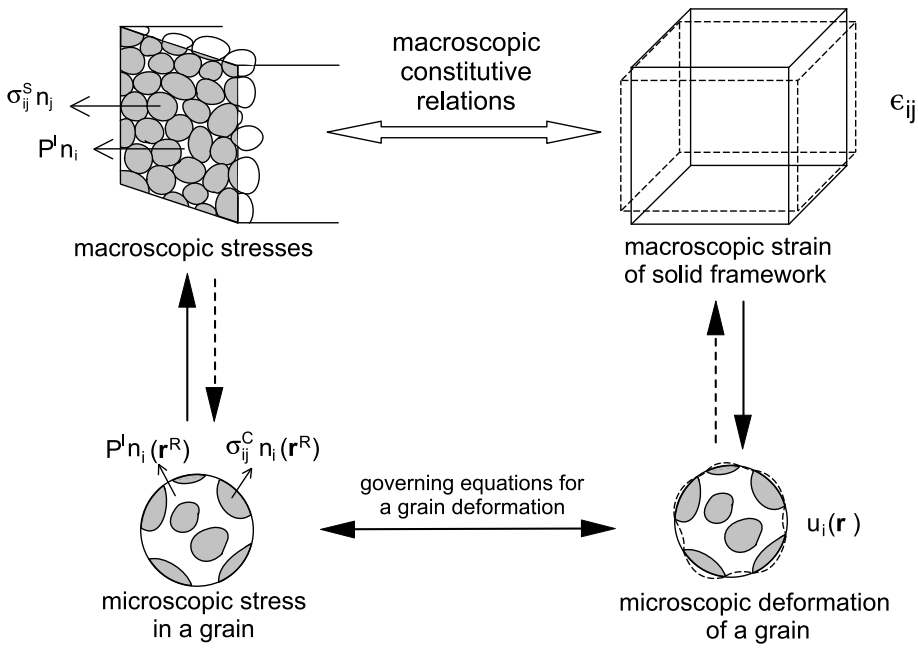

Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in, Figure 7
Granular model of a solid-liquid composite system, showing the procedures to derive macroscopic constitutive relation, Eq. (E), based on the microscopic deformation of each grain [41]

granular model (Fig. 7), and crack model are four representative models in which analytical results can be obtained for K_b and N (Table 1). All but the granular model are inclusion models in which the liquid phase is modeled by inclusions contained in a continuum solid phase, and K_b and N are derived based on the effective medium theories. In the granular model, the constitutive equation (E) is derived directly. The results from the different theories can be compared based on Eq. (11). Although the connectivity of the liquid phase is not guaranteed in the inclusion models, when considering waves in the low-frequency regime where relative motion between solid and liquid does not occur, connectivity of the liquid phase is not important.

Here, by assuming a random orientation and homogeneous distribution of the pores, the macroscopic properties are assumed to be isotropic. Then, K_b and N are derived as functions of the porosity ϕ and aspect ratio α (α=short radius/long radius) for the oblate spheroid model, as functions of the porosity ϕ and parameter ε for the tube model, where ε represents the cross-sectional tube shape (Fig. 6b), as functions of the contiguity φ for the granular model, and as functions of the crack density parameter κ for the crack model. The contiguity φ used in the granular model is defined by the ratio of the grain-to-grain contact area relative to the total surface area of each grain; thus, $\varphi = 0$ when there is no grain-to-grain contact, and $\varphi = 1$ when there is no liquid or pore phase. The crack density parameter κ used in the crack model is defined by $\kappa = n_\kappa a_\kappa^3$, where a_κ represents the radius of the circular crack and n_κ represents number density. Walsh [47] showed that in the limit of small aspect ratio, K_b and N obtained from the oblate spheroid model depend only on the crack density parameter $\kappa = 3\phi/(4\pi\alpha)$, and the results of the oblate spheroid model and crack model become equivalent. Therefore, the crack model can be included in the oblate spheroid model as a special case of small aspect ratio.

In the granular model, the dependence of contiguity φ on porosity ϕ first needs to be assessed in order to specify the dependences of K_b and N on the porosity ϕ. For a random packing of elastic spheres, which is often used as a model of soil, the total area of the elastic contacts of the spheres increases with raising confining pressure, and φ and ϕ are derived as functions of confining pressure [8]. However, when the temperature is higher than a few hundreds °C, such an elastic model is not realistic. In the deep crust and mantle, solid grains are single crystals and the contiguity is related to the area of (liquid-free) grain boundaries. Because the grain boundary and crystal-liquid interface both have interfacial energies, there ex-

Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in, Table 1
Microstructual Models for Solid-Liquid Composites

Model	Structural parameters		References
Oblate spheroid	Porosity ϕ,	Aspect ratio α	e. g., [4][1]
Tube	Porosity ϕ,	Tube geometry ε	[17]
Granular	Contiguity φ		[41]
Equilibrium geometry[2]	Porosity ϕ,	Dihedral angle θ	[41]
Crack	Crack density parameter κ		[27]

1: Typographical errors in the former studies were corrected.
2: Equilibrium geometry model is a special case of granular model.

ists an equilibrium shape of the liquid phase which minimizes the total interfacial energy of the system. The equilibrium shapes were actually observed in the high T and high P experiments for various rock + melt and rock + aqueous fluid systems e. g. [10]. Therefore, the relationship between contiguity φ and porosity ϕ is derived by assuming the equilibrium shape of the liquid phase. The granular model under this assumption is called the equilibrium geometry model (Table 1). Under a given liquid volume fraction ϕ, the equilibrium shape is controlled by the dihedral angle θ, which is determined by the grain boundary energy γ_{ss} and crystal-liquid interfacial energy γ_{sl} as $\gamma_{ss}/\gamma_{\mathrm{sl}} = 2\cos(\theta/2)$ (Fig. 8a). The theoretical results of von Bargen and Waff [45] show that under a given ϕ, the equilibrium contiguity is smaller for smaller θ (Fig. 8b). By substituting φ obtained as functions of ϕ and θ into the results of the granular model, K_b and N in the equilibrium geometry model can be derived as functions of ϕ and θ. Most rock + melt systems have θ between 20–40° and most rock + aqueous fluid systems have θ between 40–100° [10]. When $\theta \leq 60°$, a connected liquid network develops along the grain edges at $\phi > 0$. Although the tube model considers such grain edge tubules, the parameter ε in the tube model cannot be quantitatively related to the dihedral angle θ.

Now, solid-liquid composite systems are characterized in terms of the two parameters: porosity ϕ and pore shape x ($= \alpha, \varepsilon$, or θ). Because a random orientation and homogeneous distribution of the pores are assumed for simplicity, and since K_b and N do not depend on pore size, pore geometry is parameterized only by the shape. Porosity ϕ and pore shape x are generally not dependent but can vary independently governed by different physics. For example, in a texturally equilibrated system, the dihedral angle θ is determined by thermodynamic conditions such as temperature, pressure, and chemical compositions, whereas ϕ can vary mechanically through flow-in or flow-out of the liquid. In a rock + water system stressed under undrained condition, the pore shape can vary by fracture, while ϕ is kept constant. In most solid-liquid systems in the Earth, neither the porosity nor the pore shape are known.

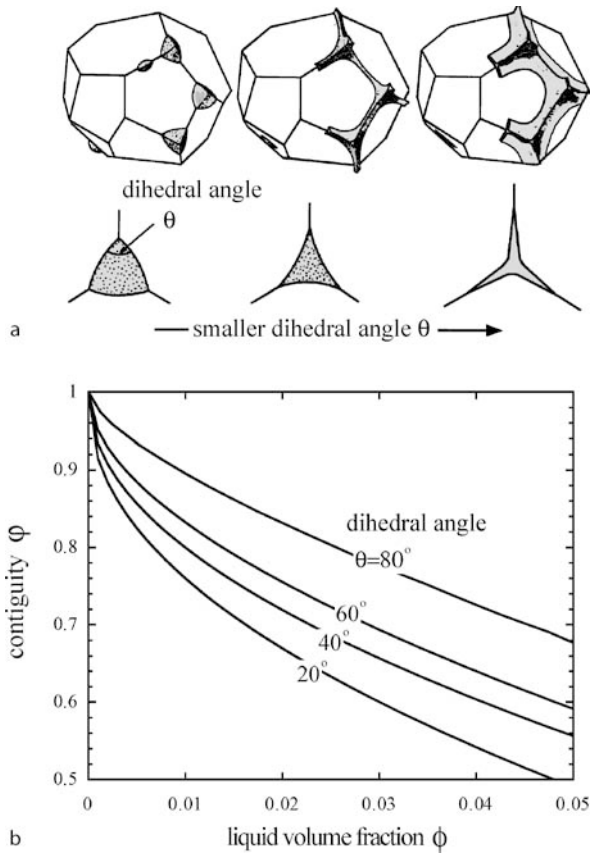

Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in, Figure 8
a Equilibrium geometry of the liquid phase characterized by dihedral angle θ. b Contiguity φ versus liquid volume fraction ϕ calculated theoretically for the equilibrium geometry with dihedral angle θ

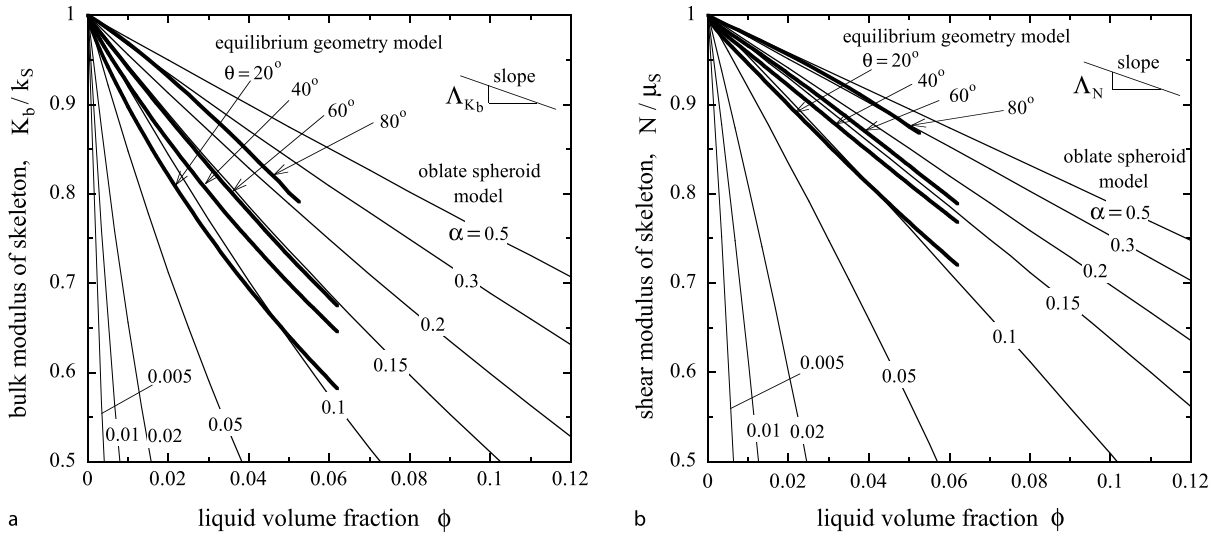

Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in, Figure 9
a Bulk modulus K_b and b shear modulus N of skeleton versus porosity ϕ, for the oblate spheroid model with various aspect ratios α and for the equilibrium geometry model with various dihedral angles θ

Because the parameter describing the pore shape is different in different models, it seems difficult to investigate the effects of pore shapes systematically. However, introduction of the concept of "equivalent aspect ratio" enables us to treat various pore shapes systematically. When the porosity ϕ is small, K_b and N are closely approximated by linear functions of ϕ,

$$\begin{cases} \dfrac{K_b}{k_S}(\phi, x) = 1 - \phi\, \Lambda_{K_b}(x) \\ \dfrac{N}{\mu_S}(\phi, x) = 1 - \phi\, \Lambda_N(x) \end{cases} \quad (x = \alpha, \varepsilon, \theta)\,, \qquad (33)$$

where k_S and μ_S represent the intrinsic bulk and shear moduli, respectively, of the solid, and the proportionality coefficients Λ_{K_b} and Λ_N are functions of pore shape x (Fig. 9). In other words, the effects of porosity and pore shape on K_b and N can be separated in such simple forms as given in Eq. (33), in which the pore shape given by x is characterized in terms of two parameters Λ_{K_b} and Λ_N. If a tube model with ε (or an equilibrium geometry model with θ) has almost the same values of Λ_{K_b} and Λ_N as the oblate spheroid model with α,

$$\begin{cases} \Lambda_{K_b}(x) \simeq \Lambda_{K_b}(\alpha) \\ \Lambda_N(x) \simeq \Lambda_N(\alpha) \end{cases} \quad (x = \varepsilon, \theta)\,, \qquad (34)$$

these two models yield almost the same values of K_b and N for a given ϕ. This also means that V_P and V_S calculated with Eqs. (31)–(32) are almost the same in these two models. Therefore, the value of α satisfying Eqs. (34) is called the equivalent aspect ratio of x; this aspect ratio guarantees the equivalence between the different pore shapes in predicting V_P and V_S.

Figure 10a shows the relationships between the different models in terms of equivalent aspect ratio. The solid and open symbols represent the equivalent aspect ratios determined from Λ_{K_b} and Λ_N, respectively. The difference between these symbols is small, indicating that, in a practical sense, one value of equivalent aspect ratio satisfying both equations in (34) can be determined. Figure 10a shows that the tube model with $\varepsilon = 0$, equilibrium geometry model with $\theta = 30°$, and oblate spheroid model with $\alpha = 0.1$ are all equivalent. Rigorously speaking, Λ_{K_b} and Λ_N depend on the intrinsic Poisson's ratio ν of the solid phase. The results shown in Fig. 10a are calculated for $\nu = 0.25$. Fortunately, however, the effects of ν are almost the same in all models, and the equivalent aspect ratio can be determined almost independently of ν. The present method to determine the equivalent aspect ratio from Eqs. (33)–(34) is applicable to general isotropic solid-liquid systems. By using the equivalent aspect ratio, general pore geometries can be treated systematically.

Determinability of Porosity and Pore Shape from Elastic Wave Velocities

It can be shown that when the porosity ϕ is small, the effects of ϕ on the skeleton properties K_b and N can be

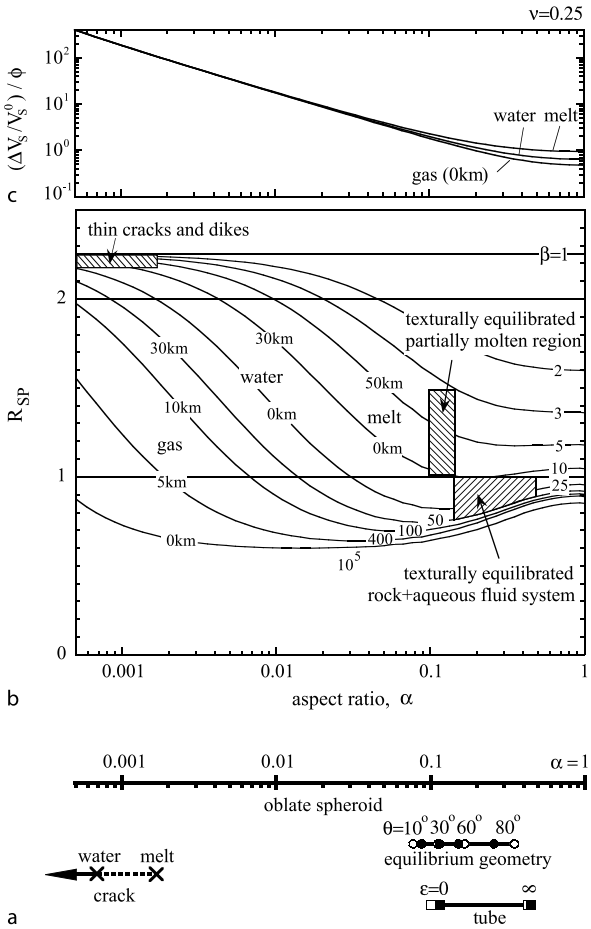

Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in, Figure 10

a Equivalence of the equilibrium geometry model, tube model, and crack model to the oblate spheroid model is shown by the equivalent aspect ratio α. Solid and open symbols are the equivalent aspect ratios determined from Λ_{K_b} and Λ_N, respectively. b R_{SP}, representing the ratio between V_S and V_P perturbations, $(\Delta V_S/V_S^0)/(\Delta V_P/V_P^0)$, versus pore aspect ratio α, for various fluid compressibilities $\beta = k_S/k_L$. c Proportionality constant between V_S perturbation $\Delta V_S/V_S^0$ and porosity ϕ versus pore aspect ratio α, for various liquid types (gas, water, and melt)

closely approximated by linear functions of ϕ. In the same manner, when ϕ is small, the effects of ϕ on the velocities V_P and V_S can be closely approximated by linear functions of ϕ. Let $\Delta V_P = V_P^0 - V_P$ and $\Delta V_S = V_S^0 - V_S$ be reductions in V_P and V_S, respectively, caused by liquid-filled pores, where $V_P^0 = \sqrt{(k_S + 4\mu_S/3)/\rho^S}$ and $V_S^0 = \sqrt{\mu_S/\rho^S}$ represent the intrinsic elastic wave velocities of the solid phase. Let $\Delta V_P/V_P^0$ and $\Delta V_S/V_S^0$ be perturbations in V_P and V_S, respectively. By substituting Eqs. (33) into Eqs. (31)–(32) and neglecting higher-order terms in ϕ (ϕ^n with $n \geq 2$), we obtain

$$\begin{cases} \dfrac{\Delta V_P}{V_P^0} = \left[\dfrac{\frac{(\beta-1)\Lambda_{K_b}}{(\beta-1)+\Lambda_{K_b}} + \frac{4}{3}\gamma\Lambda_N}{1+\frac{4}{3}\gamma} - \left(1 - \dfrac{\rho^L}{\rho^S}\right) \right] \dfrac{\phi}{2} \\ \dfrac{\Delta V_S}{V_S^0} = \left[\Lambda_N - \left(1 - \dfrac{\rho^L}{\rho^S}\right) \right] \dfrac{\phi}{2} , \end{cases} \tag{35}$$

where $\beta = k_S/k_L$ and $\gamma = \mu_S/k_S$. Without loss of generality, Λ_{K_b} and Λ_N can be treated as functions of the equivalent aspect ratio α. Equations (35) demonstrate that the velocity perturbations are affected by the five non-dimensional parameters ϕ, α, β, γ, and ρ^L/ρ^S. Because a possible variation in γ in response to a variation in the intrinsic Poisson's ratio ν of the solid phase is small, γ can be fixed to 0.6 ($\nu = 0.25$). Also, as shown below, the effect of ρ^L/ρ^S on the perturbations is small. Therefore, in a practical sense, the velocity perturbations are controlled by the three non-dimensional factors: liquid volume fraction ϕ, pore aspect ratio α, and liquid compressibility β. If only one of $\Delta V_P/V_P^0$ and $\Delta V_S/V_S^0$ is known, ϕ cannot be determined without knowing α (and β). However, if both $\Delta V_P/V_P^0$ and $\Delta V_S/V_S^0$ are known, significant constraints can be placed on ϕ, α, and/or β. A practical method to obtain these constraints is presented below.

First, we introduce R_{SP}, representing the ratio of the perturbations in V_S and V_P. From Eq. (35), R_{SP} is written as

$$\begin{aligned} R_{SP} &= \frac{\Delta V_S/V_S^0}{\Delta V_P/V_P^0} \\ &= \frac{\Lambda_N - \left(1 - \frac{\rho^L}{\rho^S}\right)}{\dfrac{\frac{(\beta-1)\Lambda_{K_b}}{(\beta-1)+\Lambda_{K_b}} + \frac{4}{3}\gamma\Lambda_N}{1+\frac{4}{3}\gamma} - \left(1 - \dfrac{\rho^L}{\rho^S}\right)} . \end{aligned} \tag{36}$$

R_{SP} can be closely related to the V_P/V_S ratio frequently used in seismology: when $R_{SP} < 1$, the perturbation (reduction positive) in V_P is larger than that in V_S and hence the V_P/V_S ratio decreases; when $R_{SP} > 1$, the perturbation in V_S is larger than that in V_P and thus the V_P/V_S ratio increases. Because R_{SP} is independent of the liquid volume fraction ϕ, this factor provides a useful insight into the effects of pore shape α and liquid compressibility β on the velocity perturbations. R_{SP} is sometimes written as $d \ln V_S / d \ln V_P$. Figure 10b shows R_{SP} versus pore aspect ratio α for various compressibility β of pore fluids. Values of β are estimated as 5–10 for rock + melt systems, 10–40 for rock + water systems, and 50–10^5 for rock + ideal

Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in, Table 2
Liquid bulk modulus k_L

depth, km	P, GPa	T, °C	k_L, GPa gas[1]	water[2]	melt[3]
0	10^{-4}	20	1.3×10^{-4} ($\beta=10^5$–10^6)[4]	2.2 (18–50)	7–25 (4–10)
5	0.15	75	0.2 (200–600)	3.1 (13–40)	
10	0.3	150	0.4 (100–300)	1.8 (22–66)	
35	1	500	1.3 (30–100)	4.5 (9–25)	
70	2				20–40 (3–6)

1: Adiabatic bulk modulus estimated by 1.3 P.
2: Isothermal bulk modulus estimated at each (P, T) condition. Data from *Schäfer* [33].
3: Data from Stolper et al. [40]. Data at $P = 2$ GPa are estimated from $\partial k_L/\partial P = 6$–$7$.
4: Numerals in the parentheses show $\beta = k_S/k_L$ evaluated for $k_S \simeq 40$–120 GPa

gas systems in the 0–50km depth range (Table 2). Values of ρ^L/ρ^S are estimated as 0.92, 0.33, and 0 for rock + melt, rock + water, and rock + ideal gas systems, respectively. The effect of ρ^L/ρ^S on R_{SP} is small, and practically the same figure as Fig. 10b can be obtained by simply assuming $\rho^L/\rho^S = 1$ ([42] Fig. 4). Figure 10b shows that for a given pore shape α, R_{SP} increases with decreasing liquid compressibility β. Figure 10b also shows that for a fixed liquid compressibility β, R_{SP} varies significantly with the variation of pore shape α. When β is fixed to 25, for example, R_{SP} is smaller than 1 for moderate values of pore aspect ratio ($\alpha > 0.03$), larger than 1 for small aspect ratio (< 0.03), and larger than 2 for very small aspect ratio (< 0.0016). Therefore, R_{SP} provides a good seismological indicator of pore shape.

When $\Delta V_P/V_P^0$ and $\Delta V_S/V_S^0$ are obtained from seismological observations or laboratory experiments, R_{SP} is calculated by taking the ratio of these two. If we know whether the liquid phase is melt, water, or gas, Fig. 10b can be used for estimating the equivalent aspect ratio α from R_{SP} under known β. Without any additional information about the liquid phase, α is estimated from R_{SP} under an assumed β. Figure 10c shows the proportionality coefficient between $\Delta V_S/V_S^0$ and ϕ (the 2nd equation of 35) versus α for various liquid types. By applying α estimated from R_{SP} to Fig. 10c, the liquid volume fraction ϕ can be determined from $\Delta V_S/V_S^0$.

Figures 10a–10c are a complete summary of the effects of liquid volume fraction ϕ, pore shape α, and liquid compressibility β on V_P and V_S. Using these figures, we can combine and confirm our understandings from the previous forward approaches, such as the different effects of gas, water, and melt, on the V_P/V_S ratio [27,49]. Also, Figs. 10b–10c represent a practical method in the inverse approach to constrain ϕ, α, and/or β from the observation of V_P and V_S. Figures 10a–10c are based on Eqs. (33), (35), and (36) in which the higher-order terms in ϕ ($\phi^{n\geq2}$) are neglected. A possible variation of R_{SP} with ϕ caused by the higher-order terms in ϕ can be obtained by calculating $(\Delta V_S/V_S^0)/(\Delta V_P/V_P^0)$ directly from Eqs. (31) and (32) without using Eq. (33). Using results from the latter calculations, it is confirmed that the present method based on the linearized equations is valid for $\Delta V_S/V_S^0 \leq 0.35$ for the equilibrium geometry model, which corresponds to $\phi \leq 0.15$, and valid for $\Delta V_S/V_S^0 \leq 0.2$ for the oblate spheroid, tube, and crack models. The applicable range is large for the equilibrium geometry model because the higher-order effects neglected in obtaining Eq. (33) cancel those neglected in obtaining Eq. (35). Compared to the equilibrium geometry model, the applicable range is smaller for the oblate spheroid, tube, and crack models. To use the present method outside the applicable range, a small modification is required ([42], Appendix B).

Application to Seismic Tomographic Images

The information on pore geometry obtainable from seismic tomographic data is the equivalent aspect ratio. Whether the actual geometry is an oblate spheroid or tube, for example, cannot be determined from seismological data. A reasonable interpretation of the derived equivalent aspect ratio requires a knowledge of probable pore geometries in the Earth. Two end-member images of pore geometry have been inferred based on experimental and field observations: one is the equilibrium geometry characterized by a dihedral angle, and the other is thin dikes and veins. In the equilibrium geometry, the pore size is much smaller than the grain size and the permeability is small. For rock + water systems, because the dihedral angle is usually larger than 60°, the equilibrium geometry may have nearly zero permeability. Dikes and veins can develop on much larger scales than the grain size and hence may have much larger permeability than the equilibrium geometry. Therefore, the particular geometry realized between these two end-members significantly affects the liquid migration velocity of the buoyancy ascent. As discussed in Sect. "Overview of Applications", whether the liquid phase is in the equilibrium shape or not can provide us with valuable information about the degree of interaction between pore geometry and macroscopic dynam-

ics. It is therefore desirable to distinguish between these two end-members using seismic tomographic data. Because the equivalent aspect ratio for the equilibrium geometry is about a factor of 100 larger than that for the thin cracks and dikes (Fig. 10a), the expected values of R_{SP} are significantly different between these two end-member geometries (hatched regions in Fig. 10b); for rock + water systems, R_{SP} is <1 for the equilibrium geometry and >2 for thin cracks and dikes; for rock + melt systems, R_{SP} is 1–1.5 for the equilibrium geometry and >2 for thin cracks and dikes . Therefore, R_{SP} can be used as a seismological indicator of the textural equilibrium, and the information on pore geometry obtained from this indicator can provide us a valuable constraint on actual fluid migration processes in the Earth. Low-velocity regions observed in the mantle wedge beneath Northeastern Japan subduction zone show a systematic change in R_{SP} with depth. Nakajima et al. [24] applied the method introduced in Sect. "Determinability of Porosity and Pore Shape from Elastic Wave Velocities" to seismic tomographic data and inferred a systematic change in pore geometry from an equilibrium geometry at a depth of ∼90 km to thin cracks and dikes at a depth of ∼65 km.

Beneath volcanic areas, low-velocity regions with lower V_P/V_S ratio than the surrounding regions are sometimes observed at depths of several km [22]. Rock + melt systems usually have β smaller than 10. This means that R_{SP} is larger than 1 regardless of α (Fig. 10b) so that the observed low V_P/V_S ratio cannot be explained by the melt-filled pores. At these shallow depths, however, the melt phase can be a mixture of melt and H_2O vapor, because H_2O initially dissolved in the melt in the deeper reaches of the subduction zone starts to exsolve. Here, we briefly discuss such situation, which is not considered in Fig. 10b and Table 2. Because of the high temperature (900–1000°C) of melt, the water phase in the melt is much more compressible than the estimates in Table 2. Figure 11 shows the fraction of water (wt%) above which β of the mixture exceeds 10, 20, or 50. It is shown that at a depth of 3–4 km, $\beta \geq 20$ occurs for the melt containing 0.2–0.5 wt% water as vapor phase, which is realistic in the subduction zone. Therefore, from Fig. 10b, $R_{SP} < 1$ can occur at large α and can explain the observed reduction in the V_P/V_S ratio. If the H_2O vapor in the melt phase can be detected by the low V_P/V_S ratio, we can obtain valuable constraints on the water content of melts and on the evolution of a magma chamber in the crust. However, because melt viscosity is considered to be high in shallow magma chambers, it is important to be careful about wave dispersion, as discussed below.

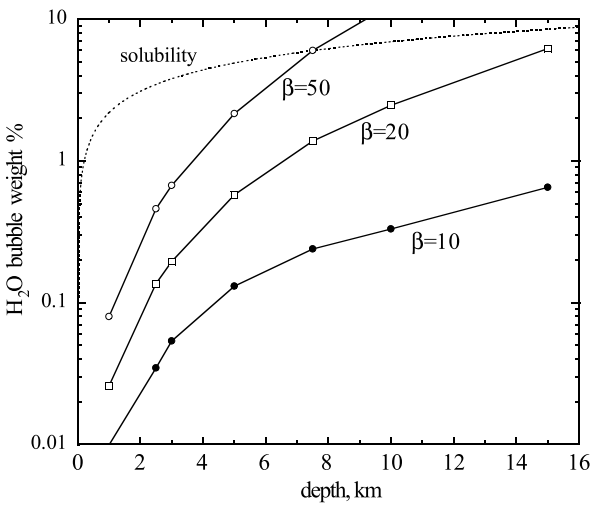

Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in, Figure 11
The amount of H_2O bubble (weight %) in melt at which the compressibility of the mixture is equal to $\beta = 10$, 20, or 50, is shown as a function of depth

Dispersion and Attenuation of Waves in Solid–Liquid Composite Systems

In deriving the elastic wave velocities in Sect. "Derivation of Wave Equations", it was implicitly assumed that the frequencies of the seismic waves are lower than the characteristic frequencies of several relaxation processes inherent to solid-liquid composite systems. If this assumption is not valid, the relaxation processes affect the wave velocities due to dispersion. To assess the applicability of the theoretical results presented in Sect. "Determinability of Porosity and Pore Shape from Elastic Wave Velocities", I present here a brief discussion of such relaxation processes. To show the mutual relationship between dispersion and attenuation, I consider the relaxation mechanism predicted from Eqs. (29)–(30). This mechanism was first studied by Biot [5,6] and is hereafter referred to as the Biot mechanism. The behavior obtained for the Biot mechanism describes the fundamental characteristic of relaxation. Several other relaxation mechanisms inherent to the solid-liquid composites also affect wave propagation and these are further summarized below.

Let ω and k be the angular frequency and wave number, respectively. By substituting the traveling wave solutions $\Omega_S = \Omega_S^0 e^{-i(\omega t - kx)}$ and $\Omega_L = \Omega_L^0 e^{-i(\omega t - kx)}$ into Eq. (29), the dispersion relation, under which non-trivial solutions exist, is obtained as

$$\left(\frac{k}{\omega}\right)^2 = \frac{\rho_U}{N} \cdot f(\omega)\,, \tag{37}$$

where the complex function $f(\omega) = f_1(\omega) + if_2(\omega)$ is explicitly written as

$$\begin{cases} f_1(\omega) = 1 + \dfrac{\Delta}{1 + \left(\frac{\omega}{\omega_c}\right)^2} \\ f_2(\omega) = \dfrac{\Delta \cdot \left(\frac{\omega}{\omega_c}\right)}{1 + \left(\frac{\omega}{\omega_c}\right)^2}, \end{cases} \tag{38}$$

and $\rho_U = (1-\phi)\rho^S$, $\rho_R = (1-\phi)\rho^S + \phi\rho^L$, $\Delta = (\rho_R - \rho_U)/\rho_U$, and $\omega_c = \eta_L\phi/(k_\phi\rho^L)$. The phase velocity V and attenuation Q^{-1} are defined by $k/\omega = V^{-1}(1 + i/(2Q))$. By assuming Q^{-1} to be small ($f_2 \ll f_1$), we obtain

$$\begin{cases} V = \sqrt{\dfrac{N}{\rho_U \cdot f_1(\omega)}} \\ Q^{-1} = \dfrac{f_2(\omega)}{f_1(\omega)}. \end{cases} \tag{39}$$

The phase velocity V and attenuation Q^{-1} given by Eqs. (39) are shown in Fig. 12 as functions of the normalized frequency ω/ω_c. Both dispersion and attenuation occur near $\omega/\omega_c = 1$ and the total amplitude of dispersion and peak value of Q^{-1} are equal;

$$Q^{-1}_{\mathrm{MAX}} = \frac{\Delta}{2} = \frac{V_U - V_R}{V_U}, \tag{40}$$

where V_U (unrelaxed velocity) represents V at $\omega/\omega_c \gg 1$ and V_R (relaxed velocity) represents V at $\omega/\omega_c \ll 1$.

Although obtained for the Biot mechanism, Eqs. (38) and (39) describe the fundamental characteristics of dispersion and attenuation regardless of the individual mechanism. These equations simply mean that the dispersion and attenuation are caused by a relaxation process. When a response $J(t)$ (e. g., strain) of a medium to a constant unit force (in the form of a Heaviside function $H(t)$) applied at $t \geq 0$ is not instantaneous but shows a time delay expressed in the form of

$$J(t) = J_U\left[1 + \Delta \cdot (1 - e^{-\omega_c t})\right] \cdot H(t), \tag{41}$$

this phenomenon is called relaxation. In other words, Eq. (41) gives a phenomenological model of relaxation. The relaxation process is characterized by relaxation strength Δ and relaxation time scale ω_c^{-1}. The time derivative of Eq. (41), $\dot{J} = J_U[\delta(t) + \omega_c\Delta\, e^{-\omega_c t}H(t)]$, yields the impulse response, where the Fourier transform of $\dot{J}/J_U$ is equal to $f(\omega) = f_1(\omega) + if_2(\omega)$ with $f_1(\omega)$ and $f_2(\omega)$ given by Eq. (38). Therefore, Eq. (38), which is called Debye equation [25], is a phenomenological model of relaxation in the frequency domain.

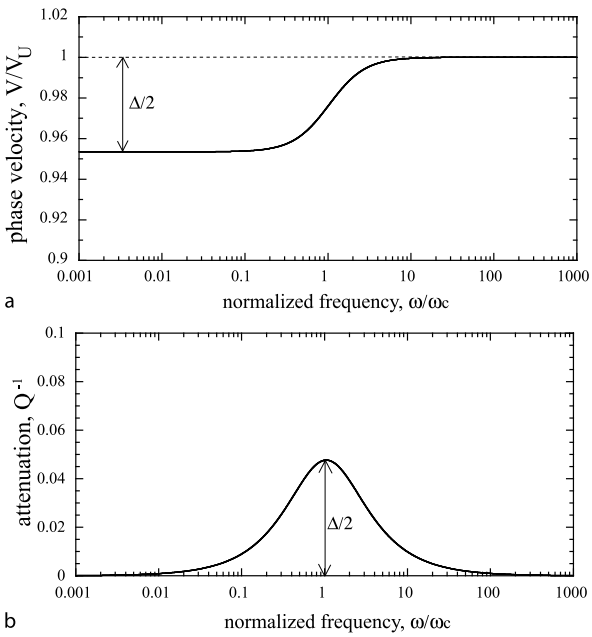

Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in, Figure 12
a Frequency-dependent phase velocity (dispersion) and b Q^{-1} (attenuation) described by Debye equations (38)

Solid-liquid composite systems show several relaxation mechanisms caused by the liquid phase. One example is the Biot mechanism exemplified above, in which the density of the system relaxes from ρ_U to ρ_R. At $\omega/\omega_c \gg 1$, due to a dominant effect of liquid inertia, a wave field cannot cause motion in the liquid phase. Hence, ρ_U is associated only with the solid mass. At $\omega/\omega_c \ll 1$, due to a dominant effect of viscous drag force, relative motion does not occur between the solid and liquid phases. Hence, ρ_R is associated with the total mass. The characteristic frequency ω_c is estimated as 200 kHz for water-saturated sandstone ($k_\phi = 10^{-12}\,\mathrm{m}^2$, $\eta_L = 10^{-3}$ Pa s, $\rho^L = 10^3\,\mathrm{kg/m}^3$, and $\phi = 0.2$). Because k_ϕ in the mantle is usually smaller than the value for sandstone, ω_c is usually much higher than the seismic frequency. The Biot mechanism for the longitudinal waves also has the value of ω_c much higher than the seismic frequency. Therefore, the assumption of the relaxed state is valid for the Biot mechanism.

Another relaxation mechanism is squirt flow [19]. When the pore shape is not spherical, changes in pore pressure induced by the elastic waves depend on the aspect ratio and orientation of each pore. Therefore, liquid pressure becomes heterogeneous at the microscopic scale, and this pressure heterogeneity is relaxed by viscous flow of the liquid between pores (squirt flow). This causes a relax-

ation of the skeleton moduli from N_U and K_{bU} to N_R and K_{bR}. The relaxation strength of N, $(N_U - N_R)/N_R$, is generally larger than that of K_b, $(K_{bU} - K_{bR})/K_{bR}$, and therefore the effect of squirt flow is larger on V_S than on V_P. In Sect. "General Theoretical Framework to Describe the Dynamics of Solid–Liquid Composite Systems", the liquid pressure was assumed to be homogeneous within REV and hence the velocities obtained in Sect. "Derivation of Wave Equations" represent relaxed velocities. When α is large, the relaxation strength is small and the difference between the relaxed and unrelaxed velocities is not significant. As α becomes smaller, the relaxation strength becomes larger, and hence the relaxed velocities can be used only when liquid pores are not isolated and the frequency of the waves are much lower than $\omega_c = k_S\alpha^3/\eta_L$ [28]. For water ($\eta_L = 10^{-3}$Pa s) and basaltic melt ($\eta_L = 1 - 10^3$Pa s), ω_c approaches the seismic frequency when α is smaller than $10^{-2} - 10^{-3}$. If the pore orientation is not random, even under the relaxed state of squirt flow, the pore pressure is different from that estimated for the random orientation. Therefore, when V_P and V_S derived in Sect. "Determinability of Porosity and Pore Shape from Elastic Wave Velocities" are applied to $\alpha < 10^{-2}$–10^{-3}, the connectivity of the pores, characteristic frequency of squirt flow, and the randomness of pore orientation should all be checked.

As exemplified in the previous section, the presence of H_2O vapor in the melt phase can relax the liquid compressibility from that of a pure melt (β_U) to that of a water–melt mixture (β_R). Figure 11 is obtained by assuming that the characteristic frequency ω_c is much higher than the seismic frequency. However, ω_c decreases with increasing viscosity of the melt. For andesitic and rhyolitic melts, the characteristic frequency is close to or lower than the seismic frequency range [12].

Future Directions

One practical problem limiting the determinability of porosity and pore geometry from seismological data lies in the difficulty of accurately estimating ΔV_P and ΔV_S. For example, a low-velocity anomaly in the upper mantle is generally caused by the superposition of high-temperature anomaly and partial melting. In order to determine porosity and pore geometry from these data, ΔV_P and ΔV_S associated only to the existence of liquid-filled pores (hereafter referred to as poroelastic effect) should be estimated by accurately correcting the data for the temperature effect. Recent experimental studies on the elastic properties of melt–free olivine polycrystals performed in the seismic-frequency range have demonstrated that at T > 1000°C and at such low frequency, the temperature effect consists of both anharmonic and anelastic effects [14]. Unlike the anharmonicity, anelasticity cannot be measured by the usual experimental methods using ultrasonic waves. Because experimental data on anelasticity are still limited and the detailed mechanism of anelasticity with or without melt is poorly understood, correction of the data for the temperature effect is difficult. Also, in the crust, accurate estimation of ΔV_P and ΔV_S for the poroelastic effect is difficult because the effect of anelasticity has not been assessed under crustal conditions and also because the lithological heterogeneity is considered to be larger than in the mantle. Therefore, the separation of poroelastic, temperature, and lithological effects affecting the velocity perturbations is an important subject of future study. Recently, not only the V_P and V_S structures but also the three-dimensional Q_P and/or Q_S (seismic attenuation) structures and two-dimensional or three-dimensional electrical conductivity structures have become available. These structures provide additional information on liquid-filled pores, temperature anomaly, and/or lithological heterogeneity. The Q_P and/or Q_S structures, for example, are important in constraining the magnitude of the anelastic effect [24]. Although the separation of individual factors is difficult to determine from velocity structures alone, additional information from Q and/or electrical conductivity structures will be very valuable.

Acknowledgments

The original and more simplified form of this article was published in Japanese [43]. I especially thank Tokyo Geographical Society, for the permission to use a modified version of figures and limited text. I thank S. Nagumo for helpful discussions. I also thank B. K. Holtzman and B. Chouet for reading the manuscript and providing helpful comments.

Bibliography

1. Aharonov E, Whitehead JA, Kelemen PB, Spiegelman M (1995) Channeling instability of upwelling melt in the mantle. J Geophys Res 100:20433–20450
2. Baba K, Chave AD, Evans RL, Hirth G, Mackie RL (2006) Mantle dynamics beneath the East Pacific Rise at 17°S: Insights from the mantle electromagnetic and tomography (MELT) experiments. J Geophys Res 111:B02101. doi:10.1029/2004JB003598
3. Barcilon V, Richter FM (1986) Nonlinear waves in compacting media. J Fluid Mech 164:429–448
4. Berryman JG (1980) Long-wavelength propagation in composite elastic media 2: Ellipsoidal inclusions. J Acoust Soc Am 68:1820–1831
5. Biot MA (1956) Theory of propagation of elastic waves in a fluid-saturated porous solid, 1, Low-frequency range. J Acoust Soc Am 28:168–178

6. Biot MA (1956) Theory of propagation of elastic waves in a fluid-saturated porous solid, 2, Higher frequency range. J Acoust Soc Am 28:179–191
7. Drew DA (1983) Mathematical modeling of two-phase flow. Annu Rev Fluid Mech 15:261–291
8. Duffy J, Mindlin RD (1957) Stress-strain relations and vibrations of a granular medium. J Appl Mech 24:585–593
9. Hasegawa A, Yamamoto A (1994) Deep low-frequency micro-earthquakes in or around seismic low-velocity zones beneath active volcanoes in northeastern Japan. Tectonophysics 233:233–252
10. Holness MB (1997) Surface chemical controls on pore-fluid connectivity in texturally equilibrated materials. In: Jamtveit B, Yardley B (eds) Fluid flow and transport in rocks. Chapman and Hall, London, pp 149–169
11. Holtzman BK, Groebner NJ Zimmerman ME, Ginsberg SB, Kohlstedt DL (2003) Stress-driven melt segregation in partially molten rocks. Geochem Geophys Geosyst 4:8607, doi:10.1029/2001GC000258
12. Ichihara M (1997) Mechanics of viscoelastic liquid containing bubbles; implications to the dynamics of magma. Ph D thesis, Univ. of Tokyo (in Japanese)
13. Iwamori H (1994) ^{238}U-^{230}Th-^{226}Ra- and ^{235}U-^{231}Pa disequilibria produced by mantle melting with porous and channel flows. Earth Planet Sci Lett 125:1–16
14. Jackson I, Fitz Gerald JD, Faul UH, Tan BH (2002) Grain-size-sensitive seismic wave attenuation in polycrystalline olivine. J Geophys Res 107(B12):2360, doi:10.1029/2001JB001225
15. Kuster GT, Toksöz MN (1974) Velocity and attenuation of seismic waves in two-phase media, 1, Theoretical formulations. Geophysics 39:587–606
16. Masterlark T, Wang HF (2002) Transient stress-coupling between the 1992 landers and 1999 Hoctor Mine, California, earthquakes. Bull Seism Soc Am 92:1470–1486
17. Mavko GM (1980) Velocity and attenuation in partially molten rocks. J Geophys Res 85:5173–5189
18. Mavko G, Mukerji T, Dvorkin J (1998) The Rock Physics Handbook. Cambridge University Press, New York
19. Mavko GM, Nur A (1975) Melt squirt in the asthenosphere. J Geophys Res 80:1444–1448
20. McKenzie D (1984) The generation and compaction of partially molten rock. J Petrol 25:713–765
21. Mura T (1987) Micromechanics of defects in solids, 2nd edn. Martinus Nijhoff Publishers, Dordrecht
22. Nakajima J, Hasegawa A (2003) Tomographic imaging of seismic velocity structure in and around the Onikobe volcanic area, northeastern Japan: implications for fluid distribution. J Vol Geotherm Res 127:1–18
23. Nakajima J, Matsuzawa T, Hasegawa A, Zhao D (2001) Three-dimensional structure of Vp, Vs, and Vp/Vs beneath the northeastern Japan arc: Implications for arc magmatism and fluids. J Geophys Res 106:21843–21857
24. Nakajima J, Takei Y, Hasegawa A (2005) Quantitative analysis of the inclined low-velocity zone in the mantle wedge of northeastern Japan: A systematic change of melt-filled pore shape with depth and its implications for melt migration. Earth Planet Sci Lett 234:59–70
25. Nowick AS, Berry BS (1972) Anelastic relaxation in crystalline solids. Academic Press, New York
26. Obara K (2002) Nonvolcanic deep tremor associated with subduction in Southwest Japan. Science 296:1679–1681
27. O'Connell RJ, Budiansky B (1974) Seismic velocities in dry and saturated cracked solids. J Geophys Res 79:5412–5426
28. O'Connell RJ, Budiansky B (1977) Viscoelastic properties of fluid-saturated cracked solids. J Geophys Res 82:5719–5735
29. Ohmi S, Obara K (2002) Deep low-frequency earthquakes beneath the focal region of the Mw 6.7 2000 Western Tottori Earthquake. Geophys Res Lett 29:1807, doi:10.1029/2001GL014469
30. Ribe N (1985) The deformation and compaction of partially molten zone. Geophys J R astr Soc 83:487–501
31. Rice JR, Cleary MP (1976) Some basic stress diffusion solutions for fluid-saturated elastic porous media with compressible constituents. Rev Geophys 14:227–241
32. Riley GN, Kohlstedt DL (1991) Kinetics of melt migration in upper mantle-type rocks. Earth Planet Sci Lett 105:500–521
33. Schäfer K (ed) (1980) Landolt–Börnstein Numerical Data and Functional Relationships in Science and Technology, New Series IV, vol 4, High-Pressure Properties of Matter. Springer, Berlin
34. Scott DR, Stevenson DJ (1984) Magma solitons. Geophys Res Lett 11:1161–1164
35. Spiegelman M, Kelemen PB (2003) Extreme chemical variability as a consequence of channelized melt transport. Geochem Geophys Geosyst 4:1055, doi:10.1029/2002GC000336
36. Spiegelman M, Kelemen PB, Aharonov E (2001) Causes and consequences of flow organization during melt transport: The reaction infiltration instability in compactible media. J Geophys Res 106:2061–2077
37. Spiegelman M, McKenzie D (1987) Simple 2-D models for melt extraction at mid-ocean ridges and island arcs. Earth Planet Sci Lett 83:137–152
38. Stevenson DJ (1986) On the role of surface tension in the migration of melts and fluids. Geophys Res Lett 13:1149–1152
39. Stevenson DJ (1989) Spontaneous small-scale melt segregation in partial melts undergoing deformation. Geophys Res Lett 16:1067–1070
40. Stolper E, Walker D, Hager BH, Hays JH (1981) Melt segregation from partially molten source regions: The importance of melt density and source region size. J Geophys Res 86:6261–6271
41. Takei Y (1998) Constitutive mechanical relations of solid-liquid composites in terms of grain-boundary contiguity. J Geophys Res 103:18183–18203
42. Takei Y (2002) Effect of pore geometry on Vp/Vs: From equilibrium geometry to crack. J Geophys Res 107(B2):2043, doi:10.1029/2001JB000522
43. Takei Y (2005) A review of the mechanical properties of solid-liquid composites. in Japanese, J Geography 114(6):901–920
44. Tsumura N, Matsumoto S, Horiuchi S, Hasegawa A (2000) Three-dimensional attenuation structure beneath the northeastern Japan arc estimated from spectra of small earthquakes. Tectonophysics 319:241–260
45. von Bargen N, Waff HS (1986) Permeabilities, interfacial areas and curvatures of partially molten systems: Results of numerical computations of equilibrium microstructures. J Geophys Res 91:9261–9276
46. Waff HS (1980) Effects of the gravitational field on liquid distribution in partial melts within the upper mantle. J Geophys Res 85:1815–1825
47. Walsh JB (1969) New analysis of attenuation in partially melted rock. J Geophys Res 74:4333–4337

48. Wang HF (2000) Theory of linear poroelasticity with applications to geomechanics and hydrogeology. Princeton University Press, Princeton, New Jersey
49. Watanabe T (1993) Effects of water and melt on seismic velocities and their application to characterization of seismic reflectors. Geophys Res Lett 20:2933–2936
50. Watt JP, Davies GF, O'Connell RJ (1976) The elastic properties of composite materials. Rev Geophys 14:541–563
51. Wong T-F, Ko S, Olgaard DL (1997) Generation and maintenance of pore pressure excess in a dehydration system 2, Theoretical analysis. J Geophys Res 102:481–852

Ecological Complexity

Brian A. Maurer
Department of Fisheries and Wildlife, Graduate Program in Ecology, Evolutionary Biology, and Behavior, Michigan State University, East Lansing, USA

Article Outline

Glossary
Definition of the Subject
Introduction: The Nature of Ecological Complexity
Describing Ecological Complexity
Survey of Different Solutions to the Problem of Describing Ecological Complexity
Future Directions: Complexity and Complementarity
Cross References
Bibliography

Glossary

Ascendancy The tendency, in the absence of disturbances, for an ecosystem to increase in size or total throughput and to have more constrained pathways for within system flows.

Bioinformatics The storage, processing, and analysis of very large arrays of biological data.

Dispersal limitation Limitation of the number of species within an ecological community due to decreased probabilities of some species entering a local ecological. community by dispersal.

Ecological community The collection of individual organisms of different species that are found within the boundaries of an ecosystem.

Ecological drift Random changes in the relative abundances of species within a community due to stochastic population processes.

Ecosystem an arbitrary ensemble of macroscopic matter that captures, stores, and uses energy to circulate and rearrange matter within the system.

Emobodied energy (emergy) Potential energy stored in chemicals bonds within an ecological entity (organism, population, community, etc.).

Food web A network describing the flows of energy and matter within an ecosystem.

General metabolic equation Phenomenological description of how mass, temperature, and resource concentration affect the metabolic rate of an organism or an ensemble of organisms.

Metabolic scaling The exponential relationship of average body mass with the rates of many metabolic processes.

Metacommunity A collection of many local communities aggregated are larger spatial scales.

Tranformity The total amount of solar energy required to form a unit of biological material.

Definition of the Subject

Living systems are collections of entities at multiple scales (e. g., cells, organisms, populations) that undergo a wide variety of interactive processes. Simply by the sheer magnitude of the different possible behaviors of such systems, the problem of describing and understanding ecological processes is daunting. In addition to the size problem, ecological systems have large numbers of unique parts, the behavior of which can vary considerably in space and time. Because of the complexity inherent in ecological ensembles, devising adequate methods to describe, analyze, and predict their behavior is a major challenge to science. Ecological systems combine both idiosyncratic, unpredictable outcomes with strong constraints on system structure that makes them paradoxically both deterministic and unpredictable at the same time. Because of this, has been no universal theory to guide research on ecological phenomena. What is needed is the development of complementary approaches that emphasize some important aspect of the greater whole of an ecosystem, but are amendable to synthesis with other approaches that emphasize other aspects. Thus, ecological complexity presents a unique challenge to science that will require a wide variety of approaches and conceptual infrastructures.

Introduction: The Nature of Ecological Complexity

Ecological systems are large ensembles of macroscopic matter that capture energy, store it, and use it to circulate and rearrange matter within the system. The term "use energy" aptly describes the material flows within an ecological system because information-based processes determine both the material structure of the system and its function. Because of the vast number of possible config-

urations ecological systems can take on, determining fundamental properties and cause-effect relationships within such systems is fraught with many conceptual and empirical challenges.

Energy flow through ecological systems occurs via a hierarchical arrangement of matter starting at microscopic scales and continuing up to encompass all life on earth. At the microscopic scale, complex molecules located within the bodies of plants and microbes intercept high-energy photons. The captured energy drives a variety of molecular pathways that ultimately produce potential energy stored in chemical bonds of complex polymers of glucose. The energy capture system is located in small organelles called choroplasts, which are part of large, highly organized molecular ensembles called cells. Chemical work done by the transformation of potential energy in chemical bonds maintains cellular cohesion and reproduction. The information regarding the configuration of cellular structure and execution of chemical work is stored on large highly organized polymers of nucleotides and ribose (DNA and RNA).

Chemical work generated by cells is expended in a variety of ways beyond maintenance of intracellular cohesion and information processing. In prokaryotes, individual cells form networks of interactions among themselves and their environments. These interactions involve secretions of chemicals produced within cells, energy-driven movement of cells within complex media, and complex interactions with other cells of the same or different species. In mutlicellular organisms, cellular interactions proceed along canonized sequences of cell division and proliferation determined by intracellular information contained on DNA. Cell ensembles follow developmental pathways driven by energy derived from stored cellular potential energy. Cell proliferation and diversification ultimately results in organisms, which obtain energy either through photosynthesis or by ingestion of potential energy stored in the cells of other organisms.

Organisms form various types of ensembles that give rise to a variety of potential ways of describing them. Because the same organism can be part of several different types of ensembles, a great deal of confusion in terminology and conceptualization ensues. Furthermore, the spatial and temporal structure of these ensembles makes most attempts at identifying system boundaries for them at best arbitrary. Terms like population, species, community, ecosystem, and biome lack rigorous, consistent definitions, and often mean different things in different situations. The necessity for arbitrary system definitions makes the study of ecological complexity above the organismal level particularly challenging.

Describing Ecological Complexity

Three general types of phenomena constitute the complex spatial and temporal structure of ecological systems: (1) energetic phenomena resulting from energy transfer and storage; (2) kinetic phenomena describing changes in amounts of substances or entities; and (3) informational phenomena describing information transfer and storage within and among ecological entities. These phenomena are at the same time complementary and interacting. Explicit description of one aspect often requires implicit assumptions regarding other aspects. Presumably, the same sequence of events in an arbitrarily defined ecological system could be described from the perspective of any one of these three aspects, although it may be more straightforward to develop one type of description for a specific phenomenon.

Ecological Energetics

Because of the energy gradient necessitated by the second law of thermodynamics, maintenance of any ecological system requires a constant influx of energy. Thus, any change in the structure or function of a system requires an exchange of energy, requiring the eventual replacement of expended potential energy. Failure to replenish stored energy results in dissolution of the system.

Although simple in concept, it is difficult in practice to describe an arbitrary ecological system as a purely energetic phenomenon. Because most of the energy contained in such systems is stored potential energy, the mass of the compounds storing that energy must be measured. It is very difficult to separate out compounds that do not store energy from those that do, so typically, either the total mass of the system or the mass of some important element (most often carbon) is measured. Since energy exchanges occur on the microscopic scale, this necessary simplification is justifiable, but also creates further difficulties. Describing the energetic aspects of the system as mass storages and flows makes concurrent definition of energy losses difficult. Energy is lost as heat, so in order to resolve the measurement scales, either mass must be expressed in energetic units, or heat loss must become an implicit assumption of the description. This creates a mismatch between the units of measure used to describe the system (g) and the units that measure fluxes (kcal).

Despite the conceptual difficulty between relating the microscopic description of system energetics (units of energy) with the macroscopic description (units of mass), for many practical problems the macroscopic description is sufficient. For example, net primary productivity (usually expressed as grams of carbon per unit time per unit area)

shows striking geographical patterns that match the geophysical properties of the earth closely. With recent developments in remote sensing, however, the thermal properties of ecosystems can be measured directly. Furthermore, many important organismal activities are also measurable directly using thermal units. It is often difficult to reconcile mass-based descriptions of the energetic properties of ecological systems with thermal-based descriptions.

Ecological Kinetics

As implied by the preceding discussion, ecological systems exhibit a kind of duality in that they are both energetic systems as well as material systems. When the focus of study is on the material properties of ecological systems, it is often convenient to express state changes of the system in material units. Often the units of interest are concentrations. In a kinetic system description, the energetic mechanisms underlying material flows are not explicitly included.

One of the most useful types of kinetic descriptions models the spatial and temporal properties of ensembles of organisms. Because organisms are fundamental entities in the organization of living matter, many practical applications focus on the kinetics of organisms belonging arbitrary ensembles such as populations and communities. These descriptions, however, require the energetic basis of organismal kinetics to be simplified or only implicitly expressed. Most often, this difficulty arises when incorporating mass transfer from organisms of one species (say a prey organism) to an individual of a different species (e. g., a predator). The actual mechanism of the mass transfer involves significant and complex mechanical and chemical processing, which in turn, requires expenditure of large amounts of stored energy. This complexity cannot be captured by a kinetic description in which the units of measure are organisms.

Further complications for kinetic descriptions arise because, most of the time, the fundamental process of change is discrete, involving addition and subtraction of individual units (organisms). Times between additions and subtractions vary in length. In addition, a large amount of uncertainty may exist regarding the sequence of events leading to these discrete changes. All of these complications create challenges for particular formalisms.

Ecological Information Content and Exchange

There are two very different ways of describing the information content of an ecological system. First, the information contained in cellular DNA profoundly shapes both kinetics and energetics in ecosystems. Fundamental biochemical pathways may differ profoundly among different kinds of organisms. For example, there are at least three different photosynthetic pathways exhibited by plants, each of which has profound consequences for energy capture efficiency under different ecological conditions. Secondly, ecological systems often exhibit configurations far from thermodynamic equilibrium. The difference between the system state and its thermodynamic equilibrium indicates ecological systems exist in very low probability states resulting from continual energetic expenditures. These low probability states imply a high degree of "organization", that is, system configurations that are perpetuated by the self-maintenance and self-replicating nature of the constituent energetic and kinetic processes.

With the emergence of molecular biology, description of genetic information content in ecological systems became possible. Although it is sometimes possible to establish direct links between genetic information and specific protein products, it is difficult to connect the spatial and temporal productions of multiple proteins across taxa and relate them to basic ecosystem functions such as matter cycling and energy flow. Furthermore, it is not apparent how to relate genetic information to kinetic descriptions of ecosystems. Since most kinetic descriptions focus on organisms, the link between genetic information content of an organism and its performance in specific environments is difficult to establish and incorporate into modeling efforts.

Genetic information exchange occurs within sexually reproducing species or by other mechanisms such as conjugation in bacteria. However, not all information contained in an ecosystem is genetic. There is a wide variety of chemical and physical signals used by organisms to exchange information about their proximate environment. Pheromones, sounds, visual stimuli, and toxins are examples of such extra-genetic information exchange. This specialized information profoundly influences ecosystem kinetics and energetics.

Another quite different aspect of ecological information is the degree to which an ecological system departs from some low-information content state. The most extreme of such states is thermodynamic equilibrium, where system components have dispensed all of their kinetic and potential energy. Such a state is problematic and of limited utility, since the entire earth is far from thermodynamic equilibrium. Alternatively, a low information state of an ecosystem would exist if its components were apportioned equally among N possible states or configurations. In such a situation, the information about any single state provides information on all other states. Borrowing from information theory, the system would be in a state of maximum

"entropy", where entropy is a measure of the homogeneity of the system. The more homogeneous a system is, the less information is needed to describe it. Therefore, in this information theoretic sense, entropy is the opposite of information. N equiprobable states correspond to maximum entropy (S), or low information content. Using a maximum entropy approach [1], the lowest information state has an entropy of

$$S_1 = K \log(N) ,$$

where K is an arbitrary constant. If the probabilities (p_i) of each state differ, then maximum entropy is less than $\log N$, and is given as

$$S_2 = -K \Sigma_i p_i \log p_i .$$

In ecological applications, K is usually given an arbitrary value of 1. The difference between S_1 and S_2 represents the difference between the actual information content of the system and the lowest possible information content (equiprobability). Difficulties arise when attempting to define a useful set of states that can be related to ecological kinetics or energetics. States that are amendable to kinetic analysis, such as population abundance, are not readily definable in terms of entropy or information. For example, it is not clear how to establish the lowest information configuration of a population.

Survey of Different Solutions to the Problem of Describing Ecological Complexity

It was widely accepted by the end of the nineteenth century that assemblages of plants and animals presented a unique challenge to science. A simple Newtonian mechanics solution to the problem of describing the huge variety of life forms on earth was not possible. At a time when physics was being revolutionized, biologists were often left to enumerating ecological phenomena without reference to any attempts to develop a comprehensive theory. Applying methods of the preceding generations of nineteenth century naturalists, ecology became a field that simply cataloged ecological phenomena without reference to testing or developing new theory. The only guiding theory was Darwin's poorly understood ideas regarding evolution. Ecologists did not test hypotheses about evolution; rather, they provided verbal descriptions of natural systems that were at best confirmatory.

The practice of ecology as descriptive natural history began to change early in the twentieth century. Mathematical probability and statistical physics were maturing, and new tools became available for application to the problem of ecological complexity. By the end of the twentieth century, a myriad of approaches had developed, each of which made major strides towards finding a comprehensive solution to ecological complexity. The advantages and drawbacks of these different solutions are examined in what follows.

Physical Biology

One of the first attempts to develop a comprehensive theoretical structure for ecology was Lotka's concept of "physical biology" [2]. Lotka's ideas had far-reaching implications that several scientific generations of ecologists used in developing specific types of solutions to the problem of complexity. Many of his ideas were co-opted for specific problems, and after passing through several generations of ecologists, the original context of his ideas were lost. Lotka's vision, however, persists as a fundamental attempt to integrate ecological complexity with the broader framework of theoretical physics. Although both physics and ecology have undergone radical changes since his time, Lotka's insights permeate many of the theoretical approaches to ecological complexity developed over the decades since he formulated them.

Using fundamental insights from thermodynamics and physical chemistry, Lotka recognized that kinetic and energetic descriptions of biological systems were complementary. He viewed the kinetic problem as the formulation and solution of a set of differential equations representing the different material states that comprised an evolving biological system. The parameters for these equations described both the interactions among the material states as well as any influences from outside the material system. Most of his examples dealt with populations of organisms, but he also envisioned kinetic descriptions of element cycling and storage. To make the kinetic equations amendable to analysis, Lotka studied the behavior of the material system in the neighborhood of its multidimensional equilibrium. This simplification allowed the linearization of the system of differential equations around the equilibrium. Descriptions of transient dynamics in the region of equilibrium were limited to smooth approaches towards or away from equilibrium, periodic oscillations dampening towards equilibrium or increasing in amplitude away from equilibrium, or "stable limit cycles", where oscillations persist in the system indefinitely.

The energetic description of biological systems Lotka found more difficult to specify. He envisioned the need for a "statistical mechanics" to describe the macroscopic consequences of the myriad of interactions that material entities in biological systems undergo. Here again he consid-

ered the problem from the perspective of a population of individual organisms encountering one another and undergoing interactions that caused the organisms to undergo changes. In these interactions, organisms could be viewed as having "kinetic energy" that was exchanged via the type of interaction the organisms engaged in (e. g., a prey organism transfers its kinetic energy to a predator). Ultimately, each ensemble of organisms (e. g., populations of predators and prey) increased or decreased in size. The rate of this increase, Lotka thought, could be related to general boundary constraints imposed by the environment to create macroscopic laws isomorphic with physical laws like the ideal gas law.

Neither information theory nor genetics were available at the time Lotka was conducting his early work on physical biology. Although he was aware, for example, of Fisher's early work in genetics, he was unable to elaborate much on the nature of ecological information. The closest he came to addressing that problem was in his comments regarding what he called "intra-species" evolution. Lotka appreciated that within any given species, individuals were not identical and could not be treated as simple particles with only a few properties relevant to the kinetics and energetics of the systems that contained them. He understood the significance of this variability to defining system level properties, but was unable to move much beyond outlining the general nature of this problem.

These early insights provided the conceptual basis for a number of different approaches to ecological complexity that emerged in the later decades of the twentieth century. Ultimately, bringing these differing approaches back together into a unifying framework, such as Lotka envisioned, would fuse a strong theoretical foundation for understanding ecological complexity.

Kinetics of Steady State Systems

Because many ecological systems are persistent as recognizable entities over the time periods typically studied by ecologists, the relatively stable configuration lends itself to description as a steady state ensemble. Borrowing from the ecological idea of an environmentally determined "carrying capacity" for each component of the ensemble, a kinetic description of the system's deviations away from stable states (i. e., equilibria) is derivable by linearization of the unknown underlying kinetics. That is, for each state variable X_i representing a component of the ecological ensemble, a steady state value for that variable (X_i^*) is assumed. Given a description of the (unknown) kinetics for component i as $\mathrm{d}x_i/\mathrm{d}t = f_i(\mathbf{X})$, where $\mathbf{X}$ is a vector containing variables for each component in the ensemble, kinetics around the collective equilibria for all variables ($\mathbf{X}^*$) is obtained by a Taylor series expansion as

$$f_i(\mathbf{X}) = f_i(\mathbf{X}^*) + \sum_j \partial f_i \big/ \partial X_j (X_j - X_j^*) + \mathbf{O}_i(\mathbf{X} - \mathbf{X}^*),$$

where j indexes over all system components, and $\mathbf{O}_i$ contains higher order terms of deviations away from equilibria. Since $f_i(\mathbf{X}^*) = 0$, and the higher order terms are assumed to be close to zero, evaluating the partial derivatives at $\mathbf{X}^*$ leads to a linearization of the near equilibrium kinetics [3,4,5,6]. Eigenanalysis of the resulting linear system indicates a limited range of potential behaviors for the system. If all eigenvalues are real, the system converges towards equilibrium or expands away from it. Imaginary eigenvalues indicate either convergent or divergent oscillations, or, if the real parts of all imaginary eigenvalues are zero, sustained periodic oscillations.

The linearized solution to ecological dynamics is most appropriate over relatively short time spans. Because organisms are somewhat resistant to change, kinetics of ecological systems are expected to maintain inertia in the face of changing environments. However, the interpretation of this inertia is somewhat different from standard physical models. In physical models, a force applied to a mass must overcome the inertia of that mass for the system state to change. In a sense, the system is pushed into a new state by expenditure of energy. Ecological systems generally change from within. The environment acts as a rate control on internal processes within such systems, so that changes in the state of the system are accomplished by shifting the balance among competing internal processes. The inertia of the system corresponds to the finite time lag over which internal processes reconfigure. Over appropriately scaled time spans, the system can be thought of as fluctuating around a relatively static state. However, over longer time spans, the static state disappears, limiting the usefulness of the linearized description.

Systems Analysis

The linearization of ecological kinetics has an implicit assumption of closure. The state variables, $\mathbf{X}$, are assumed to completely specify the system. Unmeasured quantities that may impinge on that description do so primarily through the constants obtained through the Taylor expansion. However, ecological systems are almost never perfectly closed, and it is therefore necessary to recognize how inputs to and outputs from them influence the internal workings of the system. Following a general trend in science to consider phenomena as coherent systems [7], ap-

plication of systems analysis to ecological complexity extends the range of describable behaviors.

Systems theory recognizes that complex behavior arises both from the internal workings of the system and its connections to other systems with which it is associated. Given a set of state variables, the effects of variables on one another can generate "feedback" loops. That is, any given state variable can affect its own kinetics by increasing effects of other variables on itself (positive feedback) or decreasing them (negative feedback). Each state variable is connected either directly or indirectly with many other variables, so the balance between positive and negative feedbacks determines the overall impact of the variable on itself. Steady state is achieved when positive and negative feedbacks are nearly equivalent across all state variables. Because the balance between positive and negative feedbacks can never be precise, the expectation is that a system that maintains integrity over time will show bounded oscillations.

Ecologists typically choose state variables as the amount of a particular element (say nitrogen) or as the amount of potential energy (usually expressed as the amount of organic carbon). Each state variable is represented as the amount of matter or potential energy located in different intermediate locations in the transfer of mass from one group of organisms to another. Different types of organisms obtain matter/energy in different ways. Producers generate potential energy from sunlight and absorb matter (hereafter called nutrients) from their surroundings. Consumers obtain energy and nutrients from other organisms. Nutrients are recycled through the physical environment by decomposition of dying and dead organisms or their parts. By tracing the route by which nutrients and energy are exchanged in an ecosystem, it is possible to describe a complex network, where nodes represent organisms obtaining nutrients and energy from the same sources. There are several approaches to describing these networks, or food webs.

Kinetic descriptions use differential or difference equations to describe changes in nutrients or energy within nodes (i. e. the state variables **X**). Including terms for other state variables in the functional form of the differential equation for a given state variable is typically used to represent feedbacks. Information or hypotheses regarding specific interactions among organisms determine the functional form of equations. For example, the amount

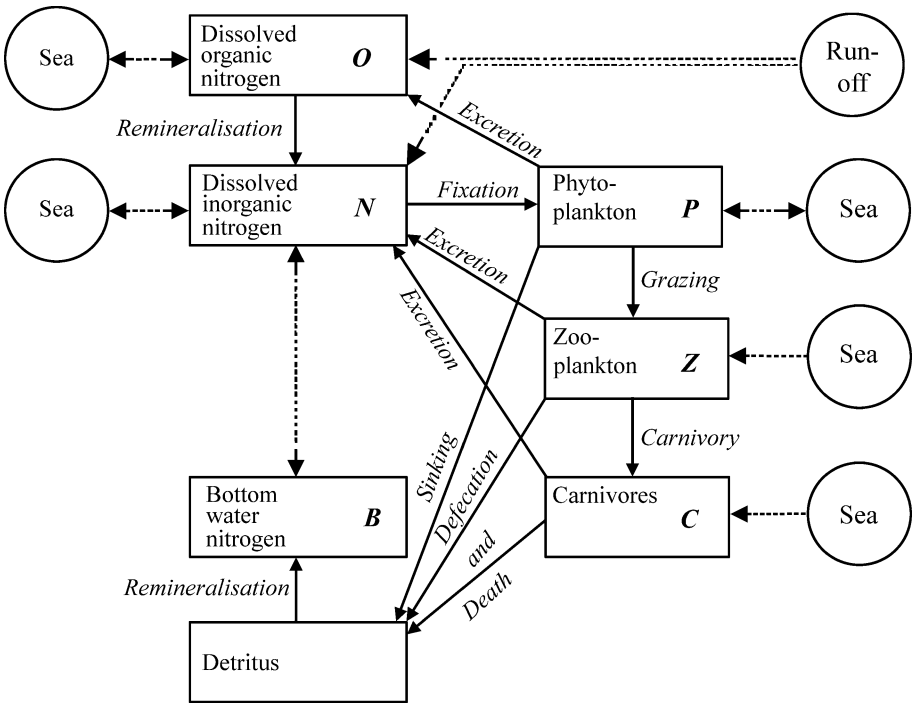

Ecological Complexity, Figure 1

Schematic representation of nitrogen fluxes in sea lochs in Scotland. *Circles* represent influxes into the system. Rectangles represent the various state variables describing nitrogen accumulation in different populations. *Arrows* represent nitrogen movements among components. System kinetics are modeled using a variety of functional relationships among inputs and state variables. From [8]

of nitrogen obtained by a consumer may saturate beyond a certain nitrogen concentration. Solutions to these sets of equations are most often obtained by numerical methods. Sensitivity analyzes that vary key parameters or functional forms gauge the dependence of numerical solutions on different assumptions made about the system. The final step in system modeling is statistical comparison of model output to data collected from the system.

The following example highlights the details of the systems modeling approach. Scottish sea-loch ecosystems are formed along the coast by variation in tidal cycles, which isolate bodies of seawater behind topographic features [8]. Freshwater enters the sea-lochs through runoff from the land and is mixed with saline water at high tide. The network of nutrient flows in these ecosystems is described by a schematic representation of the major nutrient storages and fluxes (Fig. 1). The kinetics of the system are represented by a system of differential equations such as

$$\mathrm{d}X_i/\mathrm{d}t = f_i(\mathbf{X}) \quad i = 1, 2, \ldots 7\,,$$

where $\mathbf{X}$ is a vector containing the X_i's. Rather than linearize the system, each function is determined by the nature of the processes responsible for nutrient fluxes. The system of equations includes terms that connect the state variables $\mathbf{X}$ to influxes from outside the ecosystem. These influxes are "driving functions", that is, they are inputs into the system that are responsible for system change. The system of equations are then numerically solved by dividing time into discrete intervals, and computing the resulting difference equations. The discrete intervals are made to be small relative to the span of time over which the system is being modeled, so the result is somewhat like a numerical integration. At each time step, input from driving functions is determined by data or functions thought to describe temporal variation in those inputs (Fig. 2). Validation of model output is done by comparison to data from specific systems (Fig. 3). Sensitivity analysis is often conducted by varying different parameters or functional forms in the model and determining how such changes influence model output. Once the model has been validated, it is used to forecast system behavior in response to changes in the driving functions.

Another systems-based approach to ecological complexity focuses on the topology of the network describing material flows in ecosystems (see ► Ecological Topology

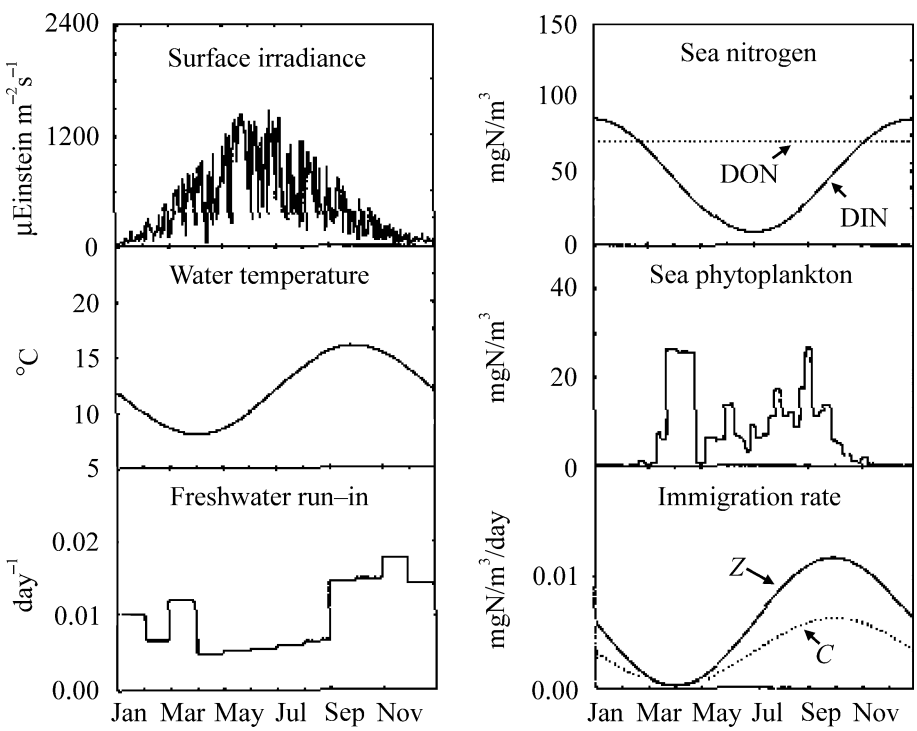

Ecological Complexity, Figure 2
Driving functions used to model exogenous inputs of nitrogen into sea lochs in Scotland. These inputs are incorporated into a variety of functions that relate inputs with nitrogen uptake rates in different organisms. From [8]

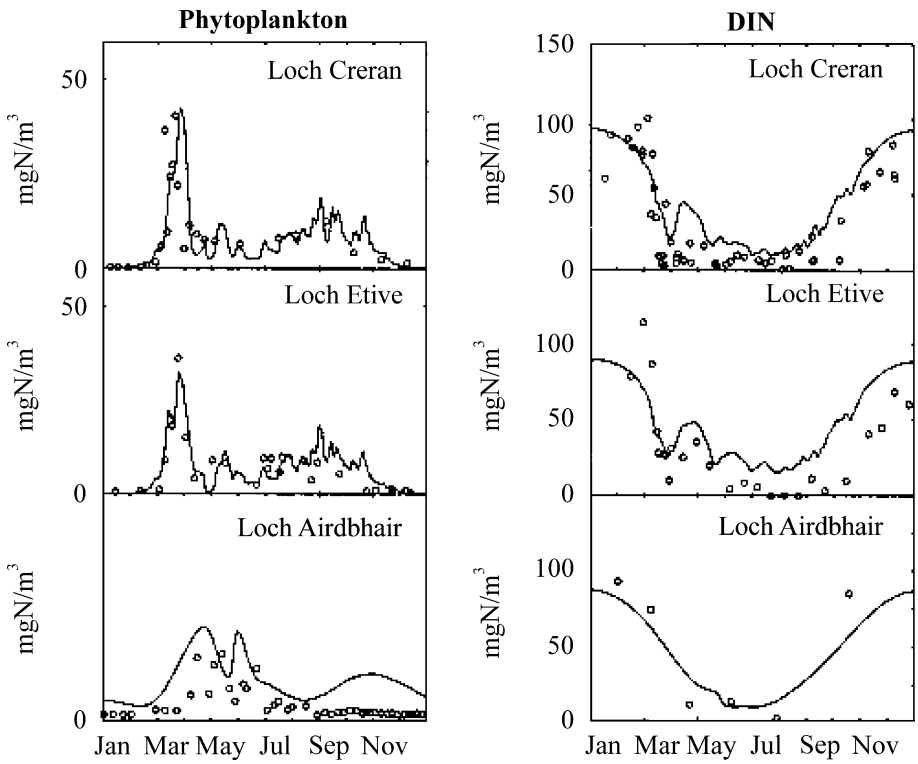

Ecological Complexity, Figure 3
Model output for monthly concentrations of nitrogen in two state variables (phytoplankton and dissolved inorganic nitrogen) for three different sea lochs in Scotland compared with field measurements. Note the general agreement between model output and field measurements. From [8]

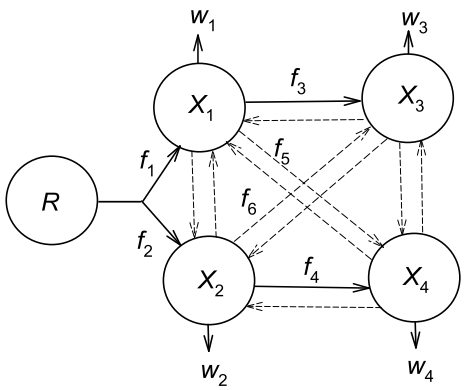

Ecological Complexity, Figure 4
Network diagram for a relatively simple food web with four nodes. *R* represents exogenous input into the web, w_i's represent outfluxes. Under conservation of matter and energy, $R = \Sigma w_i$. *Solid arrows* denote flows within the network, while *dashed lines* represent possible flows. As is typical with most food webs, the network is not over-connected, with flows to higher trophic levels (X_3 and X_4) dominating the network. Adding flows f_5 and f_6 to the web would retain these properties while making the web more complex

and Networks. A food web comprises the network of connections among different components of an ecosystem. These links can be represented as a directed graph, with edges identifying one-way flows of energy and biomass from one node (usually a species population) to others (Fig. 4). Rather than representing the amount of materials flowing through the network, analysis of food webs focuses on the nature of connectedness among food web components. Food webs are not maximally connected, that is, not all nodes in a food web are connected to all others. There is a definite hierarchical arrangement of nodes, with producers (photosynthetic/chemosynthetic organisms) at the basal level and predators at the apices. Edges rarely describe flows moving opposite the direction from base to apex. Many insights from general networking theory can be applied to the analysis of food webs. Simple food webs often have unexpected properties [9]. Furthermore, relatively simple models can describe many of the properties observed in data accumulated on trophic connections among species in nature [10,11,12]. These simplifications, however, are not fundamentally derived from the ther-

modynamics of ecosystems, but rather represent heuristic tools for describing ecosystem organization and the response of that organization to changing conditions.

The most wide-ranging attempt to deal with ecosystems as thermodynamic systems has been H.T. Odum's concept of systems ecology [13]. Beginning from fundamental principles of energy exchange co-opted from Lotka [2], Odum constructed an elaborate framework capable of expressing nearly every ecological phenomenon, from energy exchanges among molecules to the flow of goods and services through economies. The key concept uniting such diverse systems was the idea that all entities in a complex system must ultimately be composed of "embodied" energy, that is, stored energy available to do work. With few exceptions, energy in ecological systems originates from solar radiation captured by photosynthesis, hence, at least in concept, any quantity in an ecological system could be expressed in terms of the amount of solar energy expended to get the quantity to its current state [14,15]. This amount, referred to as the "transformity" of an entity, requires knowledge of the amount of solar energy that flows from its source through a complex ecosystem to its current storage location. For example, Odum [15] calculated that a unit of biological matter containing one joule (J) of energy, had a transformity of approximately 4000 solar emjoules/J. The dimension "emjoule" refers to the embodied energy contained in an entity. Odum's sweeping vision of ecological organization and its consequences for human energy transforming systems (e. g., economies) has yet to be fully realized.

Nonlinear Dynamics

In the late 1970's, the search for general solutions to the dynamics of ecological systems took an unexpected turn. Beginning with May's [16,17] demonstration that simple population models produced unexpectedly complicated behavior under certain conditions, it became apparent that departures from the "linearized" version of ecological kinetics could produce behavior that mimicked persistent fluctuations observed in ecological communities [18]. Further exploration in both ecology and other applications indicated that solutions to arbitrary sets of non-linear kinetic equations under certain conditions could generate a complex topology in the multidimensional state space representing the set of state variables describing a system. The solution set to such equations, often called a "strange attractor", has the paradoxical properties of being a deterministic attractor that produces a unique set of solutions for any arbitrary starting point in the state space. The attractor represents a complicated "folding" and "compression" of the state space onto a fractal-like solution manifold. An original high dimensional system is mapped onto a surface with a smaller fractal dimension.

Dynamics resulting from systems whose solution is a strange attractor are often referred to using the misleading label of "chaos". Chaos as defined in this way results in microscopic (i. e., the original state variables) uncertainty due to dependence on initial conditions coupled with macroscopic (i. e., the attractor surface) certainty. The nature of the macroscopic attractor can be derived from general properties of the ensemble of local state change vectors. Chaotic attractors have the property that correlations in the time evolution of points that begin in the same region of state space decline as the trajectories resulting from those points follow their unique paths on the attractor surface. This type of behavior departs from the expectation of a linearized system, where points close to one another converge along similar trajectories to the final steady state (which is often a point attractor).

A significant challenge in identifying the existence of deterministic uncertainty is the fact that any ecological system also contains other types of uncertainty, or "randomness". Uncertainty can arise from the inability to specify completely all factors responsible for system change in the kinetic description of the system. Uncertainty can also arise from the indeterminate nature of the behavior of the state variables. In an open ecological system, both types of uncertainty are likely to occur because the specified components of the system are embedded in a more inclusive system that impinges on events occurring within the specified system boundaries. For example, a community of organisms at the same trophic level may be represented by a set of differential equations, yet there may be other species not included explicitly in the kinetics represented by those equations that nonetheless cause changes in the species being modeled. There are a number of approaches that attempt to differentiate between kinetics due to "random motion" in state space from those governed by the "pull" of a strange attractor [19,20].

The importance of uncertainty in the face of arbitrary kinetic specifications of an ecological system has lead to many attempts to develop stochastic process models that might account for this uncertainty. In addition, the data themselves may have some degree of measurement error associated with them. Estimating parameters for such models using existing types of data is complicated by the existence of "measurement error", that is, uncertainty due to the relationship between measured quantities and the underlying state variables [21].

Statistical Mechanics

As the size and complexity of a system being modeled increases, the more uncertainty enters into the resulting kinetic descriptions. This is reminiscent of Lotka's [2] vision of a statistical mechanics for ecological systems. During the middle of the twentieth century, a number of attempts to define statistical descriptions of ecological systems that might arise from their complicated dynamics arose [22,23,24,25,26,27]. In the late 1970's and early 1980's, these attempts were abandoned in favor of a "mechanistic" approach to ecology, shored up by increasingly sophisticated experimental approaches [28,29,30]. This necessary phase in the development of ecology produced a wealth of evidence for the importance of local interactions among species. However, most studies were limited to a few years in length, and few lasted long enough to observe the long-term outcomes of manipulations. Those that lasted longer showed complicated behaviors, suggesting that events occurring outside the experimental system modified the results of experimental manipulations in unexpected ways [5].

Renewed interest in the statistical mechanical approach arose when it was determined that relatively simple assumptions applied to stochastic population change could reproduce a number of statistical patterns observed in large ensembles of species in space and time [31,32,33]. The basic approach begins with the description of the random variable X_{ij}, defined as

$$\begin{aligned} X_{ij} &= 1 \quad \text{if species } j \text{ has exactly } i \text{ individuals,} \\ &\qquad \text{if species probability } P_{ij} \\ X_{ij} &= 0 \quad \text{otherwise, with probability } (1 - P_{ij})\,. \end{aligned}$$

If the occurrence of species in a community of S species is independent of other species, the number of species that have i individuals at a given steady state is the sum over $j = 1, 2, \ldots S$ of the individual X_{ij}. The expected value of this sum, S_i^* is

$$S_i^* = \sum_j P_{ij}$$

and the variance is

$$\sigma_i^2 = \sum_j P_{ij}(1 - P_{ij})$$

The set of expected values S_i^*, under the assumption that species are approximately equivalent in their demography and migration rates, produces a species relative abundance distribution (SAD).

There are a wide variety of methods to obtain estimates of the probabilities P_{ij} [34,35,36,37,38,39,40,41]. The simplest formulation assumes that species are ecologically equivalent (ecological symmetry) and that the environment in which they interact is homogeneous. Under these assumptions, a stochastic birth-death process ensues which leads to fixation of a single species in a completely isolated local community [42]. That is, $P_{ik} \to 1$ for species k and $P_{ij} \to 0$ for all $j \neq k$. The random fluctuations that result from this stochastic birth-death process is referred to as *ecological drift* [31].

For species richness to be greater than unity, the community must be open to influx of individuals of other species [42]. The source of individuals available for immigration into the community is called a *metacommunity*. There are a variety of ways to define a metacommunity, and different definitions result in different expressions for the expected relative frequencies of species (P_{ij}). The fundamental result under ecological symmetry is that the SAD becomes a function of the relative frequencies of species in the metacommunity and the rate of migration into local communities [31]. The form of this function depends on the underlying structure of the metacommunity (e. g., [41]). If migration into a local community is relatively low, the community is *dispersal limited*. A balance between ecological drift and dispersal limitation then determines local species diversity. The challenge then becomes to obtain appropriate descriptions of metacommunities for an ecosystem of interest [43] and deduce the sampling distribution for local communities under the appropriate metacommunity description [41].

Solutions for SAD's become more complicated under ecological asymmetry. For example, consider the situation where species are ecologically symmetric accept for their immigration probabilities. For a sample in a local community containing J individuals, rescaling the P_{ij}'s so that they sum to one, the average probability of a species being found in a community of size J is $1/S$. The rescaled probability for species j (dropping the redundant index i) is p_j, and the variance of its limiting distribution under migration rate m_j is

$$\begin{aligned} \mathrm{var}(p_i) = [(p_i/J)(J\{1 - p_i\} - m_i\{1 - p_i\})]/ \\ [m_i(J - 1) + 1 - m_i] \end{aligned}$$

The average variance in relative frequency across species is

$$\sigma_S^2 = \Sigma_j p_j(1 - p_j)/S$$

Although the migration rate of individual species may not be known, it is possible to find the average migration rate

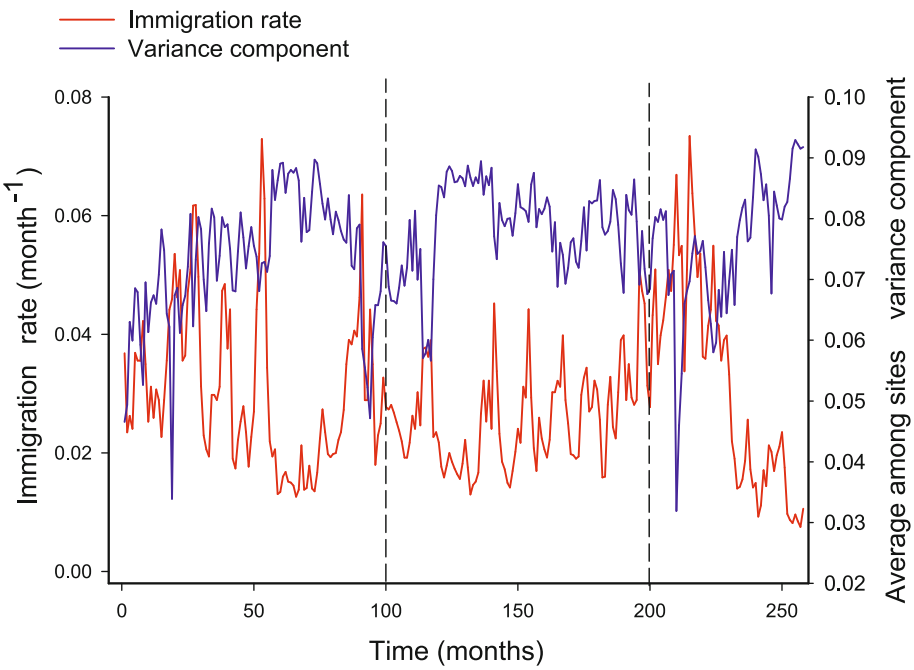

Ecological Complexity, Figure 5

Temporal patterns of variation in spatial variation in abundance and average immigration rate in a community of desert rodents living in a Chihuahuan desert ecosystem. *Vertical dashed lines* represent periods of transition among relatively distinct sets of species in response to changes in the local environment. Data provided by S.K. M. Ernest, Utah State University

across species (m^*) as a function of the p_j's, which is

$$m^* = 2J\sigma_S^2/(J^2\sigma_S^2 - 2J\sigma_S^2 + 1 - 1/S)$$

For an open community constantly undergoing fluctuations, the average migration rate provides an index of the degree to which the community fluctuations are influenced by dispersal limitation (Fig. 5).

Information Theory

Recall that there are two different types of information that can be used to describe an arbitrary complex ecological system: genetic information and what we might refer to as "configurational" information. This second type of information describes the departure of an aspect of the ecosystem, such as nutrient flows, from some hypothetical "most probable" state. Thermodynamic equilibrium is ultimately most probable, but in the face of import of energy and matter into an ecosystem, transient kinetics can maintain the system far from thermodynamic equilibrium for indefinite lengths of time, causing it to occupy a large range of low probability states. To distinguish among these alternative, low probability states, some baseline is needed for comparison. As mentioned above, if equiprobability of individual states holds, then there is very little information needed to describe the entire system. This condition implies symmetry exists among components. More accurately, if there are N possible states that the system can occupy, each of which has a relative frequency of p_i, the "entropy" of the system (see ► Entropy Maximization and Species Abundance) is

$$S = -k\Sigma_i p_i \log p_i$$

which is maximized when $S = k \log N$, in other words, all relative frequencies are equal. MacArthur [44] introduced S as a measure of ecological system "stability". Using heuristic arguments, he reasoned that stability in this sense increased either by increasing the number of system states (N) or more evenly distributing relative frequencies among system states. For many reasons, MacArthur's interpretation has been questioned, but the basic application of information theory in this context is sound. The question has become, what exactly does S represent in ecological systems? The answer hinges on how one defines the N states being considered. In a kinetic system description, the amounts of the state variables X_i are summed and used to calculate S. N represents the number of state variables (e. g., the number of species in the ecosystem or some ar-

bitrary part of it). For a kinetic description, when k is set to unity, S is the *ecological diversity* of the ecosystem. If the state variables are counts of organisms (or other ecological units), setting $k = \Sigma_i X_i$ gives the likelihood function for a multinomial distribution describing the probability density of the counts. This fact can be used to test hypotheses about the distribution of counts among state variables (assuming stationarity). In an energetic ecosystem description, the states of interest are the flow rates between state variables [45,46]. The p_i's are the fraction of total system throughput flowing through a specified edge connecting two state variables.

Based on phenomenological considerations, Ulanowickz [45,46] argued that because S represented the degree to which flows in an ecosystem are organized, that is, the degree to which they depart from symmetry of flows among all ecosystem components, it expressed a fundamental quality that changes directionally in an ecosystem over time. Letting k = total system throughput (a measure of system size), S was called the *ascendancy* of an ecosystem [45,46]. Barring any outside disturbances, according to Ulanowickz's phenomenology, ascendancy increases because over time, an ecosystem evolves to maximize the rate of energy processing [2,13], which involves adding additional energy pathways to the ecosystem (effectively increasing N, the number of system units, i. e., energy pathways and/or species). However, there is also a tendency for the pathways that maximize energy flow to increase their share of energy at the expense of other pathways, preventing the system from reaching the maximum possible ascendancy, namely $(\log N)\Sigma_i X_i$. Ulanowickz argued that all of E.P. Odum's [47] principles of ecosystem succession where specific realizations of the overall tendency of ecosystems to maximize ascendancy.

It is interesting to note that in these sense Ulanowickz used ascendancy, it plays a descriptive role not unlike fitness in Fisher's fundamental theorem of natural selection [48]. In both cases, biological processes operating among a large number of entities lead to increasing levels of organization through competitive dominance among entities balanced by inevitable tradeoffs within the entities. In most cases, this leads to a breaking of symmetry among entities, increasing the amount of information needed to describe the system.

Bioinformatics

Because the amount of information that ecosystems store is vast, unique challenges face scientists attempting to describe that information. This stored information exists as sequences of cellular DNA and RNA, distributions of gene products in space and time, and higher-level chemical and biophysical transmissions among organisms. Because of the sheer magnitude of this information, processing, storing, and describing data presents unique technological challenges. Solutions to these problems require very large computational capacity, algorithmic efficiency, and a mathematical infrastructure capable of extracting desired information from large databases. *Genomics* is the application of these technologies to large databases on gene sequences, while *ecoinformatics* does the same with large databases on geographic variation of ecologically relevant information. Much of this information involves remote sensors deployed on satellite platforms in orbit around the earth.

The ability to access such vast amounts of biological information opens the door to ask questions that have been unanswerable or even unthinkable in the past. For example, human capacity to modify the earth's ecological systems (see ► Human–Environment Interactions, Complex Systems Approaches for Dynamic Sustainable Development) has increased rapidly as both human populations and resource use have increased nearly exponentially in the past century. These global changes have induced widespread changes in the chemistry of the earth's atmosphere and oceans. Understanding how and why such changes occur is impossible without the ability to process the vast amounts of data required to describe and monitor global ecosystems. Obtaining such an understanding is critical to the long-term ability of humans to persist while maintaining an acceptable quality of life for every person.

Bioinformatics in many ways is a science in its formative stages. Although technologies are evolving rapidly, there are yet to emerge accompanying theoretical constructs that will allow the formulation and testing of hypotheses about how large-scale ecosystems interact and function. These interactions and functions operate on a vast range of scales from molecular to planetary. It is not entirely clear that present mathematical and technological developments are adequate to describe such multiscaled complexity. Because of this, bioinformatics represents a fundamental conceptual frontier into which science is just beginning to explore.

Metabolic Scaling

There has been a long tradition in biology to examine constraints on biological form and function that result from physical limitations [49,50]. Towards the end of the twentieth century, it was discovered that size, in particular, played a fundamental constraining role in the transport of energy through individual organisms [51,52,53]. The scal-

ing is allometric, that is, if M is a measure of size (usually mass) and E is the flow of energy through an individual organism, then

$$E = aM^b.$$

The empirically derived exponent, b, is usually close to 0.75. The coefficient of this scaling can be considered a function of other effects, such as temperature [54] and resource concentration [55,56,57], rendering a *general metabolic equation*. Although phenomenological in its derivation, the equation can be thought of as a macroscopic, statistical-mechanical description of energetic processes involving organisms. The mass effect represents a fundamental design constraint on fractal-like networks that distribute mass, energy, and information within cells and organisms [58]. In organisms, those networks have a minimum size corresponding to scales below which only passive diffusion can move molecular products carrying energy and information. Below this minimum threshold, movement of energy and information is random (Brownian) in nature. Here, temperature exerts its effect on metabolism. Molecular collision is required for almost any cellular action, and under Brownian motion, collisions will occur more rapidly as temperature increases. Many enzymatic-regulated cellular processes operate optimally under a relatively narrow range of temperatures. Within that range of temperatures, the range of energy states that characterize a particular enzymatic reaction can be modeled using a Boltzmann-like function, where the energy state needed to produce the enzyme product is the "activation energy" of the reaction. Metabolism, however, involves many such molecular processes, but at the macroscopic scale of the organism, one can envision an "activation energy of metabolism" that results from the various cellular processes. Hence, the macroscopic description of a metabolic process, E, can be written as

$$E = a'M^b \exp(-E_a/kT)$$

where E_a is the activation energy of the metabolic process, T is temperature (in Kelvin) and k is Boltzmann's constant. The exponential term, in statistical mechanics, represents the fraction of molecules that reach energy level E_a. However, at the macroscopic scale, this formulation can only be approximate, since so many individual reactions are involved in metabolism. Finally, the rate of supply of resources to an organism saturates at high resource density, hence, a Michaelis–Menten type function has been used to represent this final component of metabolism [55,56,59].

The application of these considerations to ecological processes has yielded a number of important insights into how energy processing in living systems constrains the form and function of organisms and the ensembles they form [60,61]. The metabolic theory in some ways brings both energetic and kinetic descriptions of ecological complexity together in a synthesis that sheds light on both types of processes. Although in its infancy, the metabolic approach to ecology is a promising approach to ecological complexity that is solidly grounded in basic physical principles and links a wide range of ecological phenomena into a coherent theoretical construct.

Future Directions: Complexity and Complementarity

Summing together was is know about ecological complexity and how it is described underscores some fundamental insights into how living systems produce order along energy gradients on the earth. The most basic principle is that the order observed in nature is a direct consequence of thermodynamic principles in open systems. The abundance of high-energy solar radiation on the earth's surface is captured as potential energy that is degraded in relatively small steps to assemble highly complicated physical networks of material and information. These networks perform recursive physical operations that open a vast number of possible routes by which solar energy can travel as it is transformed into biological work and heat.

Because of the huge number of possibilities that life presents, it is difficult, or even impossible to provide a complete description of ecological systems. To this end, scientific approaches to ecological complexity must be necessarily incomplete. The fact that there are at least three ways of describing complexity, as discussed above, is a result of this incompleteness. However, since the same set of ecological phenomena can, at least in principle, be described in several different ways, it may be possible to envision progress by looking for theoretical and empirical complementarities among different descriptions of the same system. The metabolic approach to ecology described in the last section is one such approach. For example, by linking body mass with individual production, it is possible to link kinetic descriptions of population dynamics with energy processing by the constituent organisms [60,62]. This in turn can be used to examine geographic patterns in biological diversity [63].

Ecology is a science on the verge of discovering the fundamental conceptual framework that will unite it with other sciences to portray a more complete picture of how life develops in the one system currently available to scientific study. The principles that will be formulated in the future will provide the necessary paradigms and empirical

approaches that will be needed when, if ever, living systems other than the earth are discovered.

Cross References

This article serves as the introduction to the Ecological Complexity Section. The three additional articles in the section elaborate on portions of this article. They are:

- ► Ecological Topology and Networks
- ► Entropy Maximization and Species Abundance
- ► Human–Environment Interactions, Complex Systems Approaches for Dynamic Sustainable Development.

Bibliography

Primary Literature

1. Jaynes ET, Rosenkrantz RD (1983) E.T. Jaynes: Papers on probability, statistics, and statistical physics. Kluwer, Boston, Dordrecht, Holland; Boston Hingham
2. Lotka AJ (1925) Elements of physical biology. Williams and Wilkins, Baltimore
3. Pielou EC (1977) Mathematical ecology. Wiley, New York
4. Puccia CJ, Levins R (1985) Qualitative modeling of complex systems: An introduction to loop analysis and time averaging. Harvard University Press, Cambridge
5. Maurer BA (1999) Untangling ecological complexity: The macroscopic perspective. University of Chicago Press, Chicago
6. Schaffer WM (1981) Ecological abstraction – the consequences of reduced dimensionality in ecological models. Ecol Monogr 51:383–401
7. von Bertalanffy L (1969) General systems theory. George Braziller, New York
8. Gurney WSC, Nisbet RM (1998) Ecological dynamics. Oxford University Press, New York
9. Holt RD, Lawton JH (1994) The ecological consequences of shared natural enemies. Annu Rev Ecol Syst 25:495–520
10. Pimm SL (1991) The balance of nature? Ecological issues in the conservation of species and communities. University of Chicago Press, Chicago
11. Cohen JE, Briand F, Newman CM (1990) Community food webs: Data and theory. Springer, Berlin
12. Jonsson T, Cohen JE, Carpenter SR (2005) Food webs, body size, and species abundance in ecological community description. Adv Ecol Res 36:1–84
13. Odum HT (1994) Ecological and general systems: An introduction to systems ecology. University Press of Colorado, Niwot
14. Odum HT (1988) Self organization, transformity, and information. Science 242:1132–1139
15. Odum HT (1995) Self organization and maximum empower. In: Hall CAS (ed) Maximum power. University Press of Colorado, Niwot, pp 311–330
16. May RM (1975) Biological populations obeying difference equations: Stable points, stable cycles, and chaos. J Theor Biol 51:511–524
17. May RM (1976) Simple mathematical models with very complicated dynamics. Nature 261:459–467
18. Turchin P (2003) Complex population dynamics: A theoretical/empirical synthesis. Princeton University Press, Princeton
19. Parker TS, Chua LO (1989) Practical numerical algorithms for chaotic systems. Springer, Berlin
20. Barahona M, Poon CS (1996) Detection of nonlinear dynamics in short, noisy time series. Nature 381:215–217
21. Dennis B, Ponciano JM, Lele SR, Taper ML, Staples DF (2006) Estimating density dependence, process noise, and observation error. Ecol Monogr 76:323–341
22. Fisher RA, Corbet AS, Williams CB (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. J Anim Ecol 12:42–58
23. Kendall DG (1948) On some modes of population growth leading to R.A.Fisher's logarithmic series distribution. Biometrika 35:6–15
24. Kerner EH (1957) A statistical mechanics of interacting biological species. Bull Math Biophys 19:121–146
25. Kerner EH (1959) Further considerations on the statistical mechanics of biological associations. Bull Math Biophys 21:217–255
26. Preston FW (1948) The commonness and rarity of species. Ecology 29:254–283
27. Preston FW (1962) The canonical distribution of commonness and rarity. Ecology 43:185–215, 410–432
28. Hairston NG (1989) Ecological experiments: Purpose, design, and execution. Cambridge University Press, Cambridge
29. Resetarits WJ, Bernardo J (1998) Experimental ecology: Issues and perspectives. Oxford University Press, New York
30. Underwood AJ (1997) Experiments in ecology: Logical design and interpretation using analysis of variance. Cambridge University Press, New York
31. Hubbell SP (2001) The unified neutral theory of biodiversity and biogeography. Princeton University Press, Princeton
32. Bell G (2001) Neutral macroecology. Science 293:2413–2418
33. Bell G (2000) The distribution of abundance in neutral communities. Am Nat 155:606–617
34. Etienne RS (2005) A new sampling formula for neutral biodiversity. Ecol Lett 8:253–260
35. Etienne RS (2007) A neutral sampling formula for multiple samples and an 'exact' test of neutrality. Ecol Lett 10:608–618
36. Etienne RS, Alonso D (2007) Neutral community theory: How stochasticity and dispersal-limitation can explain species coexistence. J Stat Phys 128:485–510
37. Etienne RS, Olff H (2005) Confronting different models of community structure to species-abundance data: A bayesian model comparison. Ecol Lett 8:493–504
38. Alonso D, Etienne RS, McKane AJ (2006) The merits of neutral theory. Trends Ecol Evol 21:451–457
39. Alonso D, McKane AJ (2004) Sampling hubbell's neutral theory of biodiversity. Ecol Lett 7:901–910
40. Volkov I, Banavar JR, Hubbell SP, Maritan A (2003) Neutral theory and relative species abundance in ecology. Nature 424:1035–1037
41. Volkov I, Banavar JR, Hubbell SP, Maritan A (2007) Patterns of relative species abundance in rainforests and coral reefs. Nature 450:45–49
42. Maurer BA, McGill BJ (2004) Neutral and non-neutral macroecology. Basic Appl Ecol 5:413–422
43. Leibold MA, Holyoak M, Mouquet N, Amarasekare P, Chase JM, Hoopes MF, Holt RD, Shurin JB, Law R, Tilman D, Loreau M, Gonzalez A (2004) The metacommunity concept: A framework for multi-scale community ecology. Ecol Lett 7:601–613

44. MacArthur RH (1955) Fluctuations of animal populations and a measure of community stability. Ecology 36:533–536
45. Ulanowicz RE (1986) Growth and development: Ecosystems phenomenology. Springer, New York
46. Ulanowicz RE (1997) Ecology, the ascendent perspective. Columbia University Press, New York
47. Odum EP (1969) The strategy of ecosystem development. Science 164:262–270
48. Fisher RA (1958) The genetical theory of natural selection. Dover Publications, New York
49. Thompson DAW (1942) On growth and form. The Universtiy Press, Cambridge
50. Pennycuick CJ (1992) Newton rules biology: A physical approach to biological problems. Oxford University Press, Oxford
51. Calder WA (1984) Size, function, and life history. Harvard University Press, Cambridge
52. Peters RH (1983) The ecological implications of body size. Press Syndicate of the University of Cambridge, Cambridge
53. Schmidt-Nielsen K (1984) Scaling, why is animal size so important? Cambridge University Press, Cambridge, New York
54. Gillooly JF, Brown JH, West GB, Savage VM, Charnov EL (2001) Effects of size and temperature on metabolic rate. Science 293:2248–2251
55. Real L (1977) The kinetics of functional response. Am Nat 111:239–300
56. Real L (1979) Ecological determinants of functional response. Ecology 60:481–485
57. Marquet PA, Labra FA, Maurer BA (2004) Metabolic ecology: Linking individuals to ecosystems. Ecology 85:1794–1796
58. West GB, Brown JH, Enquist BJ (1999) The fourth dimension of life: Fractal geometry and allometric scaling of organisms. Science 284:1677–1679
59. Maurer BA (1990) Dipodomys populations as energy-processing systems – regulation, competition, and hierarchical organization. Ecol Model 50:157–176
60. Brown JH, Gillooly JF, Allen AP, Savage VM, West GB (2004) Toward a metabolic theory of ecology. Ecology 85:1771–1789
61. Brown JH, Gillooly JF, West GB, Savage VM (2003) The next step in macroecology: From general empirical patterns to universal ecological laws. In: Blackburn TM, Gaston KJ (eds) Macroecology: Concepts and consequences. Blackwell, Oxford, pp 408–423
62. Ernest SKM, Enquist BJ, Brown JH, Charnov EL, Gillooly JF, Savage VM, White EP, Smith FA, Hadly EA, Haskell JP, Lyons SK, Maurer BA, Niklas KJ, Tiffney B (2003) Thermodynamic and metabolic effects on the scaling of production and population energy use. Ecol Lett 6:990–995
63. Allen AP, Gillooly JF (2006) Assessing latitudinal gradients in speciation rates and biodiversity at the global scale. Ecol Lett 9:947–954

Books and Reviews

Allen TFH, Starr TB (1982) Hierarchy: perspectives for ecological complexity. University of Chicago Press, Chicago
Botkin DB (1990) Discordant harmonies: a new ecology for the twenty-first century. Oxford University Press, New York
Brown JH (1999) Macroecology: progress and prospect. Oikos 87:3–14
Brown JH, West GB (2000) Scaling in biology. Oxford University Press, Oxford, New York
Gaston KJ, Blackburn TM (2000) Pattern and process in macroecology. Blackwell Science, Oxford, Malden
Gotelli NJ, Graves GR (1996) Null models in ecology. Smithsonian Institution Press, Washington
Hastings A, Hom CL, Ellner S, Turchin P, Godfray HCJ (1993) Chaos in ecology – Is Mother-Nature a strange attractor. Annu Rev Ecol Syst 24:1–33
Holling CS (1992) Cross-scale morphology, geometry, and dynamics of ecosystems. Ecol Monogr 62:447–502
Levin SA (1999) Fragile dominion: complexity and the commons. Perseus Books, Reading
Milne BT (1998) Motivation and benefits of complex systems approaches in ecology. Ecosystems 1:449–456
Peterson DL, Parker VT (1998) Ecological scale: theory and applications. Columbia University Press, New York
Roughgarden J (1979) Theory of population genetics and evolutionary ecology: An introduction. Macmillan, New York
Weber BH, Depew DJ, Smith JD (1988) Entropy, information, and evolution: New perspectives on physical and biological evolution. MIT Press, Cambridge
Yodzis P (1989) Introduction to theoretical ecology. Harper and Row, New York

Ecological Systems

JORDI BASCOMPTE
Integrative Ecology Group, Estación Biológica de Doñana, CSIC, Seville, Spain

Article Outline

Glossary
Definition of the Subject
Introduction
Information Theory and Diversity
Networks
Complex Dynamics
Spatiotemporal Dynamics
Thresholds
Future Directions
Bibliography

Glossary

Food webs Networks depicting who eats whom in an ecological community.

Compartments Groups of highly interacting nodes with few connections to nodes from other groups.

Scale-free networks Very heterogeneous networks in which the bulk of nodes have a few links, but a few nodes have a very large number of links.

Mutualistic networks Two-mode networks depicting the mutually beneficial interactions between plants and their pollinators or seed dispersers.

Connectivity correlation A measure of network structure that represents the correlation between the number of interactions of a node and the average number of interactions of the nodes it interacts with. A negative connectivity correlation would represent a modular network.

Species strength A measure of the importance of a species in terms of the total weight of its connections.

Network motifs Patterns of interconnections significantly over-represented in complex networks. These may be regarded as the simple building blocks of complex networks.

Trophic cascades Changes in population abundance that propagate through more than one trophic link in the food chain.

Ecosystem shifts Sudden qualitative changes in the state of an ecosystem (i. e., from clear to turbid waters in a lake) following a continuous tuning of a variable such as nutrient load.

Deterministic chaos A periodic, random-like time series generated by low dimensional, non-linear, deterministic models.

Lyapunov exponent A measure of the degree of divergence of initially close trajectories in the phase space that is characteristic of deterministic chaos.

Coupled map lattice Dynamical system with discrete time, discrete space, and continuous state. It was first used by the physicist Kunihiko Kaneko in relation to spatiotemporal chaos and later on used in ecology as a model of spatiotemporal systems.

Interacting particle system Stochastic spatial models with discrete time, discrete space, and finite states. They have been used as spatially extended models of populations and epidemics, and have been widely analyzed by Richard Durrett and Simon Levin.

Metapopulation A population of populations maintained in a dynamical balance between local extinctions and recolonizations from nearby local populations.

Extinction thresholds Critical values in the amount of habitat destroyed at which a metapopulation goes extinct.

Definition of the Subject

Ecological systems are paradigmatic examples of complex systems. Just think about the thousands of species interacting in complex ways within rich communities such as tropical rainforests or coral reefs. The most pressing questions ecologists face deal with concepts such as stability, resilience, thresholds and non-linearities which are at the core of the sciences of complexity. How robust are these cathedrals of biodiversity? At which rate will they disassemble as a consequence of global change? For example, one of the long-standing questions in ecology is the relationship between complexity and stability. This contribution will present a brief review of some of the applications of the complexity sciences into the realm of ecological systems and discuss the implications for our understanding of ecosystems. Predicting the consequences of global change on biodiversity and the services it provides will need an interdisciplinary approach in which concepts from the sciences of the complexity may be very useful. Not only complexity sciences are important for ecology, but ecological research has also provided concepts and ideas to the science of complexity, for example in the context of deterministic chaos.

Introduction

Ecology is a relatively new science. It focuses on the relationships of species among themselves and with their environments. Because of the huge number of entities and their multiple interactions and feedbacks, the study of ecosystems is amenable to some macroscopic approaches.

Although ecology has been an eminently descriptive science, important theoretical contributions were made almost from the beginning starting with the pioneering work by Lotka [68], Volterra [125], and Nicholson and Bailey [89]. These first contributions analyzed the dynamics of simple models describing two coupled populations such as a predator and its prey or two competing species. This early theoretical work defined the steady state solutions of these systems and their stability. The lessons from this exercise were to understand the possible dynamic outcomes from species interactions. For example, predators and their prey may become engaged in cycles. Some of these cycles were quite similar to cycles observed in nature such as the textbook example of the Canadian lynx and its main prey, the snowshoe hare. The competition models, on the other hand, were used to understand under what circumstances two species will coexist. These type of models are usually more useful when they do not describe appropriately the reality pointing towards important missing variables. It is the case with the Nicholson–Bailey model [89] of a host-parasitoid interaction, a type of specific predator-prey interaction in which an insect such as a wasp lays its eggs in, at or near the body of an arthropod such as a caterpillar. It is nature's own version of the celebrated movie Alien. Nicholson was fascinated by the coexistence of these insects whose abundances tend to oscillate in the field. However, the model was unstable and

led to the extinction of one of the species. Thus, something else, such as the consideration of space, was needed as we will see below.

Models of one or a few species were later on replaced by other type of models representing entire communities. The dominant question revolved around the relationship between the complexity and stability of ecological communities [74]. Also, single population models were analyzed in the context of nonlinearities, as for example in relation to deterministic chaos [75]. Another emphasis was in stochastic models where ecologists explored how time to extinction scales with population size [31,43,62], and how species coexistence depends on fluctuating environments [23,24]

Similarly, another extension of single-population, or two coupled population models were in the direction of addressing spatial degrees of freedom, that is, incorporating a spatial dimension and exploring how this new dimension made species coexistence easier.

Field ecologists, on the other hand, took other avenues but with similar goals. What regulates populations? What shapes the structure of communities? The first type of question emphasized the role of density-dependence versus external variables in explaining population change through time. The interest of this work is twofold since it may guide a biologically-informed pest control as William Murdoch and colleagues have advocated [83,84]. At the community level, the question was to understand the suite of mechanisms allowing the high levels of biodiversity that can be found in coral reefs and tropical rainforests. Joseph Connell, for example, analyzed the role of competitive interaction in structuring the marine intertidal [26]. In particular, he analyzed how patterns of recruitment, mortality and competition affected the distribution of barnacles [26]. More generally, he addressed how the high diversity in coral reefs and tropical forests is related to external perturbations in his famous intermediate disturbance hypothesis [27]. This states that the highest diversity levels are found neither in the absence of perturbations (competitive exclusion eliminates some species), nor with perturbations too frequent or intense (the bulk of species can not survive). Highest diversity levels are found at intermediate levels of perturbation.

Robert Paine emphasized the role of predation in controlling biodiversity in the intertidal [95]. His seminal work led to the concept of keystone species. He experimentally excluded the starfish in plots of the intertidal. The starfish is an important predator. It mainly preys on a species of algae which is competitively superior, keeping a control on its abundance and allowing the coexistence of several algae species. When Paine removed the starfish, the competitively superior algae out-competed the other algae species and the system became quite simplified. The predator had a strong interaction with its main prey and that had implications for the whole ecosystem. The importance of some species is much higher than what one would have predicted based on their abundance. The keystone concept is extremely important in ecology. It has shown beyond any doubt the potential ecological impact of a single species, and thus that we need to consider the roles of individual species in order to manage ecological communities [95].

What precedes is a very simplified and biased review of milestones in ecology and does not claim to be representative of the wonderful work that has been achieved in its century of history. It just tries to provide some general background on the ideas of diversity, complexity, nonlinear dynamics and threshold behavior that will be illustrated in the following sections as examples of applications of the paradigm of complex systems to the problems of ecology and the preservation of natural resources in the face of human-induced perturbations. Next, I will explore this suite of studies and how they have shed light on our understanding of ecological processes.

Information Theory and Diversity

A great contribution in ecology was made from the perspectives of general systems. Ramon Margalef was pioneering the use of Information Theory as a way to describe ecological systems [70]. He was inspired by the work of Norbert Wiener, who introduced the concept of cybernetics [4,126]. The key idea was to emphasize the feedbacks between components of the ecosystem as a way to understand the control of one system by another. A classic example of negative feed-back is that between a predator and its prey. Predator and prey regulate each other's population as a thermostat would regulate a room's temperature. Margalef's book *Perspective in Theoretical Ecology* [71] was a classic in that regard. Margalef felt completely comfortable in the context of cybernetics because it perfectly described his view of ecology, a view defined as "the study of systems at a level in which individuals or whole organisms may be considered elements of interaction, either among themselves, or with a loosely organized environmental matrix" [70]. Information Theory was applied to ecology mainly as a way to characterize the diversity of an ecosystem measured as the number of different species and their relative abundances. Diversity would be maximum when each individual was from a different species, and minimum when all individuals were from the same species. Margalef used to talk about the museum

and the field of agriculture to refer to these two extreme cases.

In the context of the theory of information, an ecosystem is like a channel that transfers information. The amplitude of this channel is measured by the Shannon Entropy, which is a measure of disorder or uncertainty. In our context, let's say that we randomly pick up an individual. What is the uncertainty that this individual belongs to a specific species? Let's assume that an ecosystem has s different species, each one with an abundance $n_1, n_2, \ldots, n_s$, so that the total number of individuals is $N = \sum_{i=1}^{s} n_i$. The probability that the randomly picked individual belongs to species i is then $p_i = n_i/N$, and one can define the diversity of the community as

$$H = -\sum_{i=1}^{s} p_i \log_2 p_i \,. \tag{1}$$

MacArthur [69] and many others advocated the use of this type of measure to describe diversity and many different uses of these indices have been applied since then, for example in trophic studies of animals' diets, or in the quantification of energy flows in food webs [123].

Networks

Another significant contribution to a general system view of ecology was the concept of food webs, networks that represent who eats whom in ecological communities. These graphical representations of communities were first drawn by ecologists such as Lindenman and Odum [32,67,90]. Odum [90] used his engineering training to represent the interrelationships of ecological systems. As in the case of Margalef, he was emphasizing the interrelationships more than the nodes. He also had a broad and rich background that allowed him to think about ecosystems with fresh views. And he insisted on the concept of energy as one of the most important currencies in ecology. Odum used energy diagrams in the hope of seeing general patterns across systems regardless of taxonomic differences [118].

Food webs have constituted one of the classic subjects in ecology, with changing emphasis through the years. In the 1970s, and as a consequence of the seminal paper by Robert M. May [74], people started looking at food web structure due to the evidence that structure greatly affects food web dynamics.

Stability and Complexity

May [74] used Gardner and Ashby's previous result to determine under what circumstances a random food web will be stable. This work was based on matrix algebra and was very successful at starting a rich research agenda. Roughly speaking, May was using Lotka–Volterra models with random interactions among species, and analyzed the probability of this model to be linearly stable. Given a certain connectance C measuring the fraction of non-zero interactions among species, May used previous results on random matrices to show that the system will be stable if

$$\alpha < SC^{-\frac{1}{2}} \,, \tag{2}$$

where S is the number of species and α is the average interaction strength among species. As noted from the previous inequality, the probability of a community to remain stable decreases as either the number of species or connections increases. This result essentially tells us that there are some constraints to randomly built communities in order to remain stable. Complexity begets instability, which contrasted with classical arguments by MacArthur, Elton and Margalef that suggested that complex ecosystems are more stable than simple ones. The question, thus, was to explore what properties of food webs counterbalance this tendency towards instability.

As a consequence of May's [74] paper, ecologists become interested in the modularity or compartmentalization of these ecological networks. The reason, at least in part, was the discussion at the end of that influential paper. May, after showing that complexity begets instability, performed some numerical experiments with non-random networks. He concluded his paper by noting that "*such examples suggest that our model multispecies communities, for a given average interaction strength and web connectance, will do better if the interactions tend to be arranged in "blocks" – again a feature observed in many natural ecosystems*". Thus a whole research program was set on compartments. Part of the research focused on exploring the theoretical implications of compartments [98]; part was trying to explore whether real food webs are compartmentalized [58,100,103]. For a review on the studies on food webs see [25,94,99,117]. This body of work emphasized invariant properties of food webs, their structure, the frequency distribution of interaction strengths, and simple models able to generate food webs with a similar structure as the observed in nature. I will not review this interesting literature here, instead I will emphasize the latest round of research in food webs that echoes similar work in complex networks. Recently, tools from the study of complex networks have been successfully applied to food webs. Food webs are now seen as another example of a complex network, with several papers comparing their structure with that shown by other types of networks such as the Internet, protein networks, or social networks [2].

Scaling in Ecological Networks

A first descriptor of network structure is the connectivity distribution, defined as the probability distribution of the number of interactions per node. The idea is to pick randomly a node in the network and represent the probability of this node interacting with one, two, . . . , n other nodes. The relevance of this descriptor of network structure stems from two facts. First, from its relationship with early graph theory by Paul Erdös and Alfred Rényi [34] and recent build up models that generated a good correspondence between several models of network formation and their consequent connectivity distribution. Second, because the paper by Albert et al. [1] clearly related the shape of the connectivity correlation with the network robustness to error and attack. Albert et al. [1] found that the Internet has a connectivity distribution that follows a scale-free distribution defined by a power-law of the type:

$$p(k) \propto k^{-\gamma} , \tag{3}$$

where $p(k)$ is the probability of a node having k links and γ is a critical exponent. In a log-log plot, this relationship is defined by a straight line of slope $-\gamma$ for all the range of k values. That is, Eq. (3) is a relationship not defined on a particular scale. This would not be the case, for example, for an exponential distribution that has a specific scale, the average number of links per node [108].

Barabasi and Albert [6], building on a previous result by Simon [110] showed that a process of network build up where new nodes link preferentially with already well-connected nodes (a type of "rich gets richer process") is a simple recipe to generate scale-free networks. On the contrary, a random model such as the classical random graph by Erdös and Rényi [34] generates distributions with Poisson distributions or exponential distributions if the number of nodes keeps growing. The important point is that in the latter case, the resulting network is much more homogeneous in the sense that all nodes have a similar number of interactions.

A randomly built network with connectivity distributions with thin tails, such as the Erdös–Rény random graphs, are very fragile to the random deletion of nodes. After a certain fraction has been removed, the networks fragment. This fragmentation threshold will be revisited later on in the context of spatial processes, where space is represented as a regular (or irregular) network of points. Thus, random networks are very fragile [1]. On the other hand, scale-free networks are much more robust to the random deletion of nodes. One has to remove a high fraction before the network gets fragmented. The reason is that the few highly connected nodes (the hubs) play a major role in keeping the entire network together. Since these hubs are quite rare, it is very unlikely to remove them by chance. However, as shown by Albert et al. [1], these hubs are the Achilles' heel of the network. If one now starts removing the most connected nodes, the whole network collapses. Thus, scale-free networks are very robust to the random loss of nodes but very fragile to the loss of the hubs.

Ecological Systems, Figure 1
The mutually beneficial interactions between plants and their animal pollinators (picture) and seed dispersers have played a major role in the generation of earths' biodiversity. Picture courtesy of Mark Chappell

The work by Albert et al. inspired ecologists who turned to food webs in search of their connectivity distributions. Solé and Montoya [112] first analyzed a few food webs and found evidence for a scale-free distribution, while Camacho et al. [20] compared different distributions and found the best fit to be to an exponential distribution. Dunne et al. [30] generalized these previous results by using a broader data set and testing several functions. Their conclusion was that even when there were a few food webs described by fat tail distributions, the bulk of the food webs had tails following an exponential distribution.

Jordano et al. [53] extended the argument by focusing on a different type of ecological network, the one describing the mutually beneficial interactions between plants and their animal pollinators or seed dispersers (Fig. 1). These are two-mode networks with a much higher level of resolution than traditional food webs. While food webs have a high level of lumping so that a node contains several taxonomic species, mutualistic networks have a level of resolution almost always corresponding to a taxonomic species. These networks describe the coevolutionary process in species-rich communities [8]. This study analyzed 53 communities and concluded that in the bulk of cases,

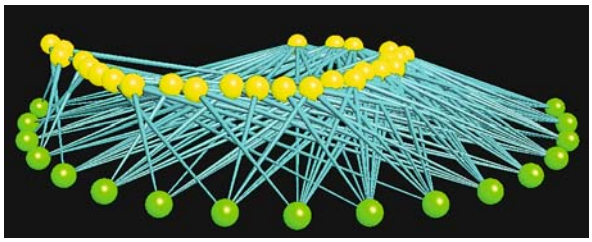

Ecological Systems, Figure 2
The mutualistic interactions such as the one depicted in Fig. 1 form complex networks of species interdependence. The architecture of these networks greatly affect their robustness to the extinction of one of the species. The picture represents a plant-pollinator network in the Arctic. Plants and insects are represented as green and yellow nodes, respectively

connectivity distribution for both plants and animals was best fitted by a truncated power law, a distribution of the following form:

$$p(k) \propto k^{-\gamma} e^{-k/k_c} . \quad (4)$$

The main difference in relation to Eq. (3) is the existence of a critical connectivity level k_c beyond which the connectivity distribution decays faster than expected for a power-law. These mutualistic networks are still very heterogeneous but not as heterogeneous as predicted for a scale-free distribution (Fig. 2) [53].

There are several non-exclusive factors that may account for the existence of these truncated power-law distributions. Jordano et al. [53] focused on what they termed *forbidden links*, that is, the existence of interactions that are not possible due to size or phenology uncoupling. For example, an insect cannot pollinate a plant species if it is a migrant that arrives after the flowering period of the plant. Or a bird species will not disperse a tree species if their seeds are larger than the width of the bird's beak. By combining analytic thinking and natural history, Jordano et al. [53] were able to account for a large fraction of the non-observed interactions in two well-studied communities.

Of course the fact that forbidden links exist and that their existence can lead to a truncation of an otherwise power-law does not exclude additional mechanisms. Forbidden links and similar mechanisms such as filtering information (i. e., a new node can sample only a subset of the network) constrain the preferential attachment mechanism. Other processes also lead to truncated power-law distributions without any constraint on such a process. For example, the same preferential attachment process taking place on a bipartite network leads to a truncated power-law distribution if there is any asymmetry between the two sets such as one set (e. g., plants) growing faster than the other set (e. g., animals) [39,41].

From the point of view of the robustness to species extinction of these mutualistic networks, the truncated power-law distribution confers more robustness than an exponential distribution to the random extinction of species but less dependence to the extinction of the hubs than for a power-law distribution.

Network Structure: Modules

The connectivity distribution is just a first description of network structure. In the general field of complex networks, scientists looked at deeper measures of network structure such as connectivity correlation or modularity. This was mainly analyzed for genetic networks and the Internet. For example, the connectivity correlation measures the average correlation between the number of links of a node and the average number of links of the nodes it interacts with. Maslov and Sneppen [73] found that both the internet and protein networks had a negative connectivity correlation, which means that hubs tend to interact with poorly connected nodes. This corresponds to an organization in compartments, which may buffer from the propagation of mutations or other perturbations [73]. Melián and Bascompte [78] applied this idea to food webs and found that they are more cohesive than the Internet or protein networks. Generalist species tend to interact among themselves. This may make these communities less robust to the propagation of a perturbation such as a contaminant, but more resistant to the extinction of a species. There is more than a single way to be robust [78]. This description of food webs is complementary to a cohesive modular organization where several k-subwebs, that is, groups of species with at least k interactions among other species in the subweb, are linked to a densest central subweb which induces cohesion to the entire food web [79].

This work on the structure of food webs links to the early attempts to characterize compartments mentioned above [74,100,103]. In this regard, research on food webs [74,98] pioneered the search for network structure that 30 years later would be so important in complex networks. The search for compartmentalization in food webs has not found too much evidence, partly because a lack of high quality data, partly because a lack of appropriate statistical tools to unambiguously define and characterize modules. More recently, Krause et al. [58] used software available to sociologists and found the strongest evidence of compartmentalization in three out of five food webs studied. In the context of the physics of complex networks, recent work has addressed the role of compartments in the

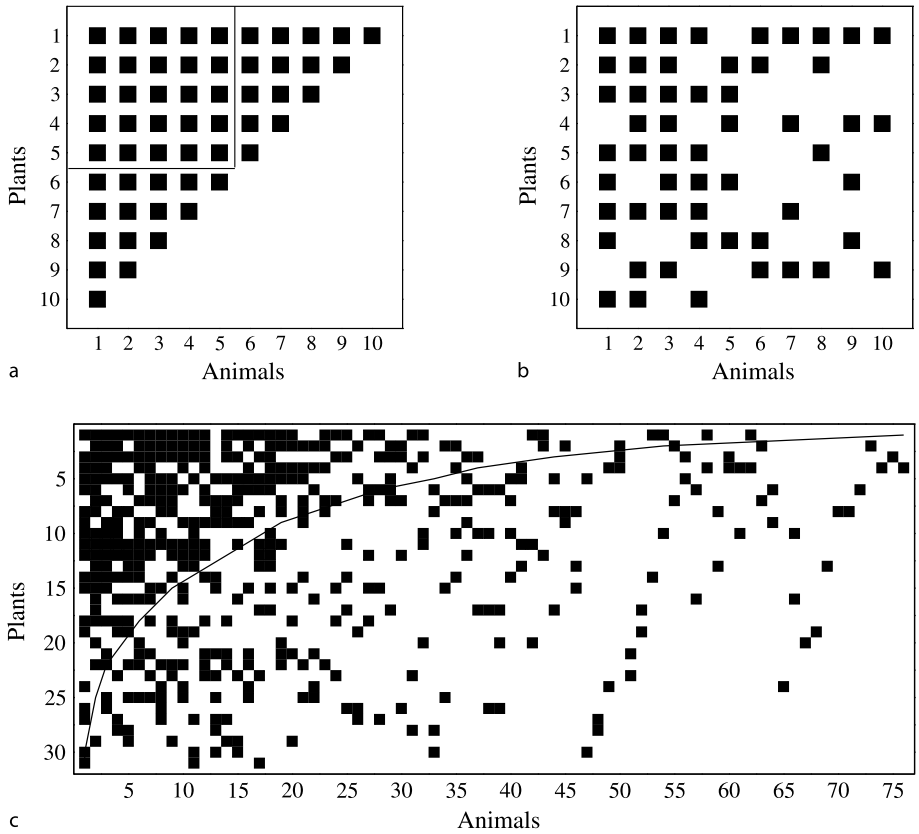

Ecological Systems, Figure 3

The nested assembly of mutualistic networks. Two-mode interacting networks are represented as a matrix with plants in rows and animals in columns. A square indicates that the plant in this row and the animal in this column interact. Panels a, b and c represent a totally nested, random and real network. The line in c represents the isocline of perfect nestedness. Modified from [14]

structure of complex networks such as the world-wide air traffic [42]. Several algorithms to quantify modularity are now available. Olesen et al. [92] have used these algorithms to detect modularity in pollination networks. These modules are interpreted as the basic units of coevolution, that is, small groups of highly interacting plants and animals. The modularity analysis is useful in this context in showing the denser areas of the network. These denser areas have the potential to be coevolutionary hotspots or vortices [119].

Another concept ecologists have used to further describe the structure of mutualistic networks provides from island biogeography, nestedness. In the mutualistic context, a matrix of plant-animal interactions is nested if specialists interact with species that form perfect subsets of the species with which generalists interact (Fig. 3) [14]. This is a pervasive community organization that has been described for other ecological interaction such as those between cleaning fish and their hosts [41], or parasites [60]. Nestedness implies a central core of interactions where generalist plants and generalist animals interact among themselves. This originates a dense core of interactions with a high level of redundancy and the possibility for the system to respond to perturbations. This is somehow in agreement with the cohesive organization of food webs found through the connectivity correlation and *k*-subweb distribution seen above. On the other hand, a nested mutualistic pattern implies an asymmetric pattern of specialization since specialists tend to interact with generalists. The latter tend to be more abundant and less fluctuating, and thus these community patterns confer mechanisms for the persistence of rare species [14]. Ecologists are now starting to explore the implications of these universal community patterns from the point of view of community responses to perturbations such as habitat loss [5,37] or the invasions of foreign species [19,80,82,91].

Weighted Ecological Networks

These heterogeneous, asymmetric network patterns in mutualistic networks are also observed in weighted networks. In this case, species strength, the weighted equivalent of species degree, grows faster than linear with species degree [7]. This pattern had been previously found for the world-wide airport network, but not for the scientific collaboration network [7]. The strength of highly connected species is even higher than expected based on their degree because specialists tend to interact exclusively with the most generalized species [14], and so depend completely on them. Thus, specialists contribute disproportionately to increase the overall strength of the generalists they depend upon. The nested structure of these mutualistic networks accounts for this pattern. The predominance of weak interactions and the asymmetry in pairwise interaction when a plant, for example, depends highly on an animal, tends to increase the conditions for the persistence and stability of species-rich communities as indicated by analytical results of a simple community model [16].

The role of weak interaction strengths on community stability has also been analyzed in studies of weighted food webs. This research agenda may be traced back to the seminal work by Robert Paine [95,96] who in his classic experiments on the intertidal noted in the introduction, found that the strength of interactions between predators and their prey are defined by a few strong interactions in a matrix of weak interactions. This pattern has then been observed over and over in other food webs and using other measures of interaction strength [15,35,103,123,127]. Simple dynamic models have shown that this frequency distribution of interaction strengths increases the stability of communities [57,77]. However, the frequency distribution of interaction strength is only a first descriptor that does not tell us how these interaction strengths are combined in the basic components of the food web. Thus, Neutel et al. [88] found that weak interactions tend to be distributed in long loops. This avoidance of strong interactions in long loops induces the stability of the whole community [88]. Similarly, Bascompte et al. [15] found that the co-occurrence of two strong interactions in a tri-trophic food chain occurs less often than expected by chance, and that in the few cases in which this occurs, it tends to be accompanied by strong omnivory (predator preying on two consecutive levels of the food chain) more often than expected by chance. These results have implications for the likelihood of trophic cascades, that is, changes in species abundance that transmit at least through two consecutive levels of a food chain. An example of a tropic cascade would be a decrease in sharks through overfishing, a subsequent increase in abundance of big fish that constitute their prey, and a concomitant decrease in the abundance of smaller, herbivorous fish the former prey on. Two strong interaction strengths have the potential to induce trophic cascades after the overfishing of top predators, but when accompanied by strong omnivory the magnitude of this cascade is severely reduced. Thus, in the light of the dynamic results of a biologically parametrized bioenergetic model, one can conclude that, other things being equal, the reduced frequency of two consecutive interaction strengths and their association to strong omnivory reduce the likelihood of trophic cascades [15]. However, this global pattern does not assure that the marine food web is buffered from the effects of overfishing, since overfishing does not affect randomly picked species but tends to focus on large, top predators. In the Caribbean food web, fishing selectively targets a biased sample of species belonging to upper trophic levels [86,97]. These species include ten heavily fished shark species from seven families that account for almost one-half of the strongly interacting food chains in the Caribbean food web [15]. The likelihood of trophic cascades after the overfishing of these predators is thus high. These cascades can contribute to the depletion of herbivorous fishes at the base of the chain such as parrotfishes that are important grazers of the algae. The reduction of these herbivorous fishes can accelerate the transition from corals to algae, an example of bistable steady state that will be considered later on in the context of ecosystem shifts as examples of phase transitions (see below).

First quantifications of interaction strength through energy fluxes was using information theory, the same framework we have described in the previous section in the context of species diversity [123]. New metrics to characterize weighted food webs build on this preliminary study [18].

Network Motifs and Trophic Modules

A final parallelism between complex networks and ecological food webs has to do with network motifs, patterns of interconnections that are over-represented in complex networks and that can be considered as the simple building blocks of complex networks [81]. Interestingly enough, there is a significant difference between research on network motifs and its equivalent research in ecology. The approach in complex networks is eminently structural, while that of their ecological counterpart is eminently dynamical. For example, the first papers on network motifs quantified their representation in entire networks and compared their frequency with that predicted by appropriate

null models. Only later on there were some studies exploring the dynamics of these different motifs [101]. On the other hand, ecology has been studying the dynamics of simple trophic modules such a tri-trophic food chain without looking at how frequent are these simple modules in entire food webs [9]. What remains to be done now is to scale-up from these isolated modules to the entire food web.

Complex Dynamics

Let us now move from structural complexity to dynamical complexity. There is a strong relationship between ecology and complexity sciences in the context of population dynamics. As a matter of fact, one of the more seminal contributions to deterministic chaos came from theoretical ecology. Once more, the great talent of Robert May was behind this contribution [79]. May was looking at one of the simplest models one could think of in theoretical ecology. It describes the dynamics of populations with non-overlapping generations. Let's assume that the density of insects in a generation t is N_t. Let's first normalize this value by dividing it by the highest density ever observed $N_{\max}$ so that $x_t = N_t/N_{\max}$. If we imagine a deterministic model with density dependence, one can write what the density of insects at the next generation will be

$$x_{t+1} = \mu x_t(1 - x_t) \,, \tag{5}$$

where μ represents the population rate of increase. Robert May analyzed the temporal dynamics of system (5) as he was increasing the growth rate μ. For very low values, the population goes extinct. When $\mu > 1$ the population reaches a steady state. If μ is further increased, at $\mu > 3$ the steady state becomes unstable and a cycle of period two becomes stable. For even higher μ-values, there are other period-doubling bifurcations and so the population oscillates with cycles of higher frequency. Finally, when μ reaches a critical value the dynamics never repeats itself, the system shows deterministic chaos [79].

Finding this period-doubling route to chaos in an ecological model opened a research agenda that found that this scenario has universal properties, e. g., the relationship between the successive critical μ_k values at which a new bifurcation k appears are independent of the model. More than that, even in experimental systems one could observe the same universal laws [37]. Specifically, Feigenbaum showed that for a large enough μ-value, the following relationship takes place:

$$\delta = \lim_{k\to\infty} \frac{\mu_k - \mu_{k-1}}{\mu_{k+1} - \mu_k} = 4.6692\ldots \,, \tag{6}$$

The co-discovery of deterministic chaos in the logistic equation (together with parallel work in meteorology and mathematics) had a huge importance, not only in the field of ecology, but beyond. It is one of the few examples in which the flow of ideas has gone from ecology to physics. Since this important discovery, a rich research program of research in ecology revolved around the role of deterministic chaos in ecological systems, both theoretically [76] as well as empirically [51,106].

Chaos in the Real World

William Schaffer and Mark Kot [106] were among the pioneers in the search for deterministic chaos in real ecological systems such as the cycles of the Canadian lynx or the monthly records of measles in big cities. Ecology was facing the possibility that complex temporal series were not the result of hundreds of stochastic variables, but of a few variables in deterministic, yet non-linear dynamical systems [52]. This is not just a technical issue. If complex dynamics in the populations of diseases or pests were deterministic we could understand the underlying rules. However, the evidence for chaos has been more evasive due to the shortness of temporal series and their high amounts of noise.

Arguably, the first serious attempt to quantify chaos in nature was the paper by Hassell et al. [46], who analyzed the temporal series of 28 arthropod insects from both the lab and the field. Detecting chaos depends very much on the way it is attempted, and ecologists have been very imaginative in their search for chaos in a noisy world. In this first study, Hassell et al. fitted their temporal series to a previously studied non-linear population model. Twenty-six populations had temporal series which best fit model parameters within the parameter region corresponding to steady states, and only one example corresponded to the region of deterministic chaos. This was thought to be of little empirical support for chaos to begin with.

Schaffer and Kott [106] used a different approach and different data sets, and their work supported the notion that chaos may be common in ecology. They used the same techniques physicists were using, such as attractor reconstruction and estimation of its fractal dimension. Several time series such as the Canadian lynx had attractors reminiscent of the strange attractors that are the hallmark of deterministic chaos. Similar results were obtained for the measles records in Baltimore [3,106].

Sugihara and May [116] used non-linear forecasting techniques to distinguish deterministic chaos from noise (both correlated and uncorrelated). This clever approach

is based on dividing the temporal series in two halves and using the first half as a source of known data, and considering the second half as the unknown future. The correlation coefficient between predicted and observed values is plotted versus prediction time. Deterministic systems show an exponential decay in correlation, while this is constant for noisy systems. Ellner and Turchin [32] used non-linear techniques to estimate the largest Lyapunov exponent, i. e., the parameter telling at which rate two nearby trajectories in the phase space will diverge. Their strategy to reduce the high levels of noise present in the original time series was to first estimate the map that best fits their temporal series, and then use it to calculate the Lyapunov exponent from the dynamics of the map. Tilman and Wedin [120] built a Poincaré map by plotting the biomass of annual plants one year versus the same biomass the next year. The slope of this map for a vegetation model previously used dictates whether the system is or is not chaotic. In summary, there was a serious effort to find chaos in real ecosystems by using a broad spectrum of techniques. There was evidence of chaos, but also evidence against its presence. This line of research is almost extinguished, but a recent paper by Sibly et al. [109] touched on it by estimating return rates after a perturbation for a very large number of temporal series of groups of mammals, birds, fish, and insects. They found that in the bulk of cases the return rates were quite below the threshold for chaos, which corresponds to stable populations. However, there is a clear case for the potential of chaos in population dynamics, a beautiful example that comes from an interaction between analytical and lab work.

Costantino and colleagues [28,29] combined an experimental setting where a population of the flour beetle *Tribolium* was growing in milk bottles, and a population dynamic model that, although simple, incorporated the basic dynamics of the species life cycle. This was a discrete time model that described the three phases of the beetle life cycle, namely feeding larvae, non-feeding larvae, and adults. Cannibalism is very common in this species and mathematically induces a strong nonlinear term that is partly responsible for the presence of deterministic chaos. This team first proceeded by parametrizing their model with the temporal data they obtained in the lab. From the experimental point of view, they manipulated the recruitment rate into the adult stage by adding or removing adults at the time of the census. As this recruitment rate was increased, there was a sequence of period-doubling route to chaos shown both by their lab census as well as by their model with the adequate parameter values.

In sum, maybe chaos is not common in nature, but nature certainly has the potential to show chaos.

Spatiotemporal Dynamics

Deterministic chaos was perhaps the first pedagogical example of the potential for non-linear dynamics in ecology. The lesson was that other dynamical behaviors beyond steady states and cycles are compatible with a deterministic, density-dependent model. The next finding of the potential of non-linear dynamics to generate complex phenomena was provided by the study of dynamic systems extended in space. Imagine a discrete lattice of sites simulating the patchy distribution of some available habitat. Within each one of these habitat patches a local population can be described by a dynamic model such as the logistic map. However, we allow now for the fact that a fraction of the individuals born in a patch disperse to neighboring sites. The resulting spatiotemporal dynamics can be described by a coupled map lattice (CML), a dynamical system with discrete time, discrete space and continuous state, first used by the Japanese physicist Kunihiko Kaneko working with problems of diffusion and spatiotemporal chaos [54,55]. This approach allows one to easily study the combined action of two processes: local dynamics (described by an appropriate *map* or discrete time model such as the logistic Eq. (5), and the coupling through dispersal of these local maps. A CML can be written in the following way:

$$x_{t+1}(i) = (1-D)F\{x_t(i)\} + \frac{D}{k}\sum_{j=1}^{k} F\{x_t(j)\}\,, \qquad (7)$$

where $x_t(i)$ is the density of population at site i and time t, D is the fraction of individuals leaving its patch and k is a certain neighborhood around a local patch to where individuals can move. Coupled map lattices such as (7) have been extensively used in ecology [10,112]. For example, coupled map lattices have shown how dispersal may affect the temporal dynamics [49], and the length of the transients [50]. Particularly relevant is the finding of spatial self-organizing patterns such as spiral waves in the abundance of populations [47,48,115]. This phenomenon is qualitatively similar to the one found for excitable media where symmetry breaking takes place around a pacemaker [85]. From the point of view of ecology this suggests that simple rules can create long-range spatial patterns. Similar spatial models belong to the class of interacting particle systems, where not only space and time are discrete, but also the state of a cell is on one of a few discrete values. As opposed to CMLs, the latter are stochastic models for which there is a rich body of mathematical work addressing, among others, how spatial pattern arises in ecology [31]

Coupled map lattices are useful to link patterns at different spatial scales. The problem of pattern and scale is at the core of ecology [63]. For example, one can describe the population dynamics at a lattice site. In the case of a chaotic map, this dynamics will be quite unstable, with strong fluctuations. This unstable characteristic was one of the arguments by which some field ecologists argued that chaos would not be common in nature [17]. If populations oscillate so heavily, at some point population density will be low enough for stochastic events to lead the population to extinction. However we can now see the same unstable dynamics from a larger spatial scale, let's say that we plot the total abundance in 2×2 lattice sites, 4×4 lattice sites, and so on. How unstable will the chaotic dynamics look? Surprisingly, the dynamics will now appear very constant. Solé and Bascompte [12,111] coined the term *chaotic stability* to refer to the fact that chaos and its instability at a local scale can induce a strong stability at a global scale. Thus, the criticism of chaos based on its instability does not apply when space is considered.

One way to understand the previous result is to remember the strong dependence on initial conditions of a chaotic system. Because of this, two nearby local populations will start to oscillate out of phase, and so ups and downs will soon cancel each other. Technically, this can be analytically proved following reference [105]. These authors found a relationship between the largest Lyapunov exponent and the spatial coherence length. The coherence length describes how far away two points oscillate in a correlated way. As a matter of fact, the inverse Lyapunov exponent can be regarded as a correlation time, i. e., the time horizon beyond which two initial trajectories fluctuate totally independent from each other. In the vicinity of the onset of chaos, the following relationship between the largest Lyapunov exponent (λ) and the spatial coherence length (ξ) holds: $\xi \approx \lambda^{-1}$. Thus, the more chaotic a system is, the faster spatial correlation decays with distance. An application of this idea in ecology was proposed by [10], providing a clear mechanism for an early suggestion by [120] on the difficulty of detecting chaos at larger spatial scales even when present at local scales.

Another interesting application of coupled map lattices is to the problem of pattern formation, which is a celebrated one in several fields such as developmental biology and excitable media [85]. In the context of ecology, pattern formation in space is related to the problem of species coexistence. This is another example where introducing spatial degrees of freedom changes entirely our picture of ecological systems. At the beginning of this contribution I mentioned that Lotka and Volterra had derived a mathematical model for two competing species. The lesson from that model was that the coexistence of the two species is only compatible with low values of interspecific competition. One interesting avenue has been analyzing the mathematical conditions under which species coexist with different environmental fluctuations [23,24]. Another aspect has been considering the spatial component. If one considers the spatial extension of a competitive model, local exclusion is compatible with global coexistence if stochasticity plays a role in the dynamics. From a spatially homogeneous setting where both species were initially present, one ends up with clusters of patches where species one is the survivor and alternative clusters where the second species wins [114]. These steady state spatial patterns are equivalent to Turing patterns in models of development.

An even more striking example of spatial self-organization is the existence of spiral waves such as the ones observed in excitable media as some chemical reactions or electric activity in the heart [38,85]. If one extends the host parasitoid model by Nicholson and Bailey into a coupled map lattice, one can observe the spontaneous emergence of traveling waves in the density of one of the species (Fig. 4) [47,113,115]. This is relevant from several points of view. From the point of view of complex systems, this shows how the interplay between local non-linear dynamics and short-range dispersal can generate large scale self-organized spatial patterns as first shown by the great mathematician Alan Turing [63,121]. This opens a new way to interpret large scale patterns in ecology, traditionally adduced to reflect environmental causes [15]. Interestingly, these spiral waves are related to the persistence of interacting species: once more, despite the local instability, the system is globally stable and all species coexist more easily than predicted by non-spatial models.

Coupled models have also shown how dynamics can be affected by dispersal. For example, reaction-diffusion models of population dynamics have illustrated the dispersal-induced route to chaos, that is, the change in the type of dynamics from steady states to cycles and to chaos as the dispersal rate is increased [59,93]. Similarly, coupled map lattices have shown how the transient time in non-spatial models becomes now extremely large [50], which had also been noted by Kaneko from a physical perspective. This has important implications in ecology. We implicitly assume that the steady state is the relevant dynamics, but if transients are as long as thousands of years, transient dynamics may be much more relevant for ecology than long-term steady states [50].

As for the case of deterministic chaos, a myriad of papers looked for these self-organizing patterns in nature, and good evidence for pattern formation come from ex-

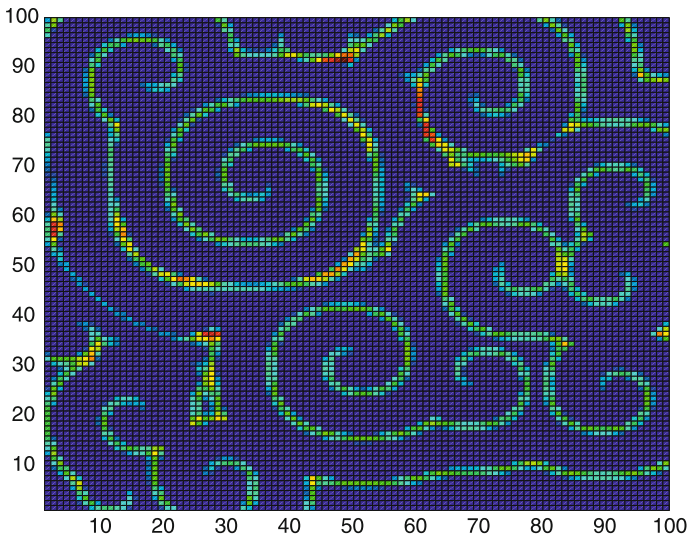

Ecological Systems, Figure 4
Spiral waves in a coupled map lattice model of interacting populations. The figure corresponds to an iteration and each lattice size codes the abundance of hosts in a host-parasitoid system. These self-organizing spatial patterns are very much related to the persistence of populations

amples of rodents in northern Europe [104], host-parasitoid interactions [72], and outbreaks of the moth *Zeiraphera diniana* in the Alps [19]. Once more, theory was ahead and lead field ecologists to search for examples of complexity in real nature. This was expanding our horizons and moving from a classical view of ecosystems where all complex processes were associated with external variables, to another scenario where internal processes were able to account for much of the complexity observed in real nature.

Thresholds

One important application of the sciences of complexity is the concept of phase transition from statistical mechanics. This is very important because in ecology we are used to thinking in terms of linear relationships between a cause and its effects. Oftentimes, as we tune a parameter we find the occurrence of a critical point in which a sudden qualitative change takes place. A previously stable solution becomes unstable and two new stable solutions emerge, as we have seen in the period-doubling route to chaos. A symmetry-breaking process takes place and the system chooses one of two possible solutions. A mechanical analogy would be a ball rolling on a surface with two minima. This describes the dynamics of a phase transition.

The paradigmatic example of a second-order phase transition in physics is the Ising model. This model describes the behavior of a set of magnets on a square lattice of length side N. The state of each lattice site i at time t is defined by the spin $S_t(i)$. Each spin can be in the states upwards (1) or downwards (-1) and interacts with its four nearest neighbors to minimize energy, that is, to have parallel alignment. The global magnetization is $M = \sum_K S_i$, and the idea is to plot this measure as a function of the temperature. For high temperatures, noise dominates, and the distribution of spins is random, i. e., $M = 0$. At very low temperatures, the system is ordered and all spins point towards the same direction (either upwards or downwards). M becomes maximum. As we progressively decrease the temperature, a sudden transition takes place at a critical temperature T_c. The magnetization per spin $m = M/N$ behaves close to T_c (for $T < T_c$) as $m \approx |\tau|^\alpha$ where $\tau = (T - T_c)/T_c$ [107].

A relevant parameter to characterize spatially distributed systems is the correlation length ξ. For $T > T_c$ we already said the system is random and correlation lengths are small. Close to T_c, ξ scales as $\xi \approx |\tau|^{-\nu}$, ν being another critical exponent. Below the critical point, the model exhibits long-range order. Clusters exist on every length scale. That is, the system is scale-free. The correlation length ν, the size of the maximum cluster, and the variance in sizes diverges to infinity as we approach the critical point.

Percolation theory has had a nice application as a null model in landscape ecology [122]. A useful example is its

application to the problem of habitat fragmentation. This is an extraordinarily complex problem due to the accelerating rates of habitat destruction everywhere and the well-known fact that habitat transformation is the number one cause of biodiversity decline. Imagine a spatial lattice as the one described above. A direct application of percolation consists in envisioning a situation in which each site is originally pristine, i. e., occupied by vegetation, and one proceeds by destroying an increasing fraction of randomly placed sites. As for the Ising example, the size of the largest patch starts declining smoothly at the beginning. A new destruction event just reduces the size of the single large patch by one site. But close to the percolation threshold, an additional destruction implies that the previously continuous cluster breaks down in small pieces. To separate the effects of habitat loss from those of habitat fragmentation, one can use the following order parameter [13]:

$$\Omega = \frac{S_{\max}}{\sum_{k=1}^{N} \Theta(k)} , \tag{8}$$

where $S_{\max}$ is the size of the largest cluster, and $\Theta(k)$ is one if site k is available, and zero if it is destroyed. As can be easily seen, when all available sites belong to the same cluster, the previous equation is one. From a biological point of view, it means that we are only facing habitat loss. However, when habitat loss induces habitat fragmentation, the value of the order parameter drops suddenly. This is because we have now several disjoint clusters of vegetation, and thus only a small fraction of the available sites belong to the largest cluster. Interestingly enough, the order parameter drops really fast near the percolation threshold, so its value is one below a critical level of habitat destruction, and becomes almost zero after that threshold.

Extinction Thresholds

The above non-linear changes in landscape structure as more habitat is destroyed have implications for the persistence of a species inhabiting such a landscape. Species inhabiting heterogeneous landscapes living in a dynamical balance between local extinctions and recolonizations from neighborhood patches are called metapopulations [44,66]. If we plot the regional abundance of a metapopulation (i. e., the fraction of sites occupied) versus the fraction of sites destroyed, one observes the presence of an extinction threshold defined as a critical destruction value at which the metapopulation goes extinct despite a fraction of the habitat is still available [61]. In spatially explicit systems with local dispersal, the rate at which a metapopulation's regional abundance decreases is faster than in the case of spatially implicit models. That is, the effects of habitat loss are higher as more habitat has already been destroyed [13]. The reason has to do with the previously reported non-linear changes in the landscape. Essentially thus, habitat destruction models are equivalent to models of phase transitions in statistical mechanics. This theory has served to better understand the consequences of habitat destruction on metapopulations. It has been very pedagogical in suggesting how changes in ecological systems are not smooth, but rather non-linear.

The presence of extinction thresholds in metacommunity dynamics is equivalent to the eradication thresholds in epidemiological models. Nee [87] already noted the equivalence between these two types of models and their critical points. These critical points can be simplified to expressions where species-specific parameters such as colonization or transmission rates cancel out. Eradication thresholds can be phrased as the points at which the destructive process reaches the amount of resource used when all the resource is available. For example, consider the following metapopulation model originally proposed by Levins [66]:

$$\frac{\mathrm{d}v}{\mathrm{d}t} = cv(1 - v - D) - ev . \tag{9}$$

The previous equation assumes an infinite number of habitat sites and describes the temporal dynamics in the fraction of sites occupied by a metapopulation (v). It contains a positive term describing the increase in occupied sites due to the colonization of empty sites. c is the species-specific colonization rate, and D is the fraction of sites permanently destroyed. The second term in Eq. (9) refers to the loss of previously occupied sites due to local extinction, e being the extinction rate. Note that the previous model can also be used as a toy model of an infectious disease, in which case v would be the fraction of hosts infected, c would be the transmission rate, D would be the fraction of hosts vaccinated and e would be the clearance rate. Model (9) has a positive steady state as long as the fraction of sites destroyed (hosts vaccinated) is lower than the threshold $D_c = 1 - e/c$. As noted by Nee [87], this threshold is equivalent to the steady state in the absence of destruction: $v^* = 1 - e/c$. Thus, one can predict D_c without knowing the parameters c and e by measuring the amount of habitat occupied when all habitat is pristine. This simple rule was coined as the Levins rule by Hanski et al. [45]. In epidemiological models, the infectious disease disappears when the fraction of hosts vaccinated is equal to the fraction of non-infected hosts when all hosts are suscep-

tible [11,87]. Similar ideas can be applied to the dynamics of transposable elements, a type of intragenomic parasites [11].

Stochastic spatial models such as the contact process have allowed us to derive mathematical conditions under which a species will persist or die out with almost certainty, that is, similar thresholds for species persistence [31]. In this review we have encountered both spatially explicit models such as the CMLs and the interacting particle system, as well as spatially implicit models such as Model (9). There are analytical approaches such as moment-closure that allow to bridge between these two extremes [64].

Ecosystem Shifts

Thresholds such as the one here illustrated for the case of habitat loss are common in ecology. One can find several examples in which a variable is smoothly changed with no apparent consequence in the macroscopic properties of the system until a threshold in which an abrupt transition in the state of the system takes place. These are known as ecosystem shifts [107,111]. One classical example of an ecosystem showing two alternative steady states is that provided by shallow lakes [22,107]. There are documented examples where a lake has shifted between an initial state characterized by clear water and submerged vegetation to a state characterized by turbulent water, a high concentration of phytoplankton, and an absence of submerged vegetation. This lack of vegetation is associated with a reduction in diversity, since several species of fish and other taxa use vegetation as food and refugia [107]. This transition in shallow lakes occurs as a consequence of human-induced eutrophication, and constitutes a global problem affecting also small seas such as the Baltic. The tuning parameter in this example would be the amount of fertilizers dumped into the lake. As one starts increasing this parameter nothing seems to occur for a while. Until the critical point is reached and the new state suddenly takes place. Once more, there is no apparent correlation between the last push and its amplified consequence. Shallow lakes show a profound hysteresis in response to nutrient load [107]. This means that the system is irreversible, and that environmental actions to recover the pristine state may be costly. Now one needs to almost clean the lake completely to revert the change to the pristine state.

The case of a lake is by no means the only example. Other examples involve the transitions from corals to macroalgae, or from herbaceous vegetation to bare desert. In the first case, extensive areas of coral reef have been replaced by a system dominated by algae. Corals hosts countless numbers of other species. Currently, ecologists have started documenting these transitions and looked for their explanations. Several non-exclusive explanations involve increased nutrient loading and overfishing that have decreased the abundance of herbivorous fish, thus freeing algae from their control and allowing them to take over and replace corals [52]. Overfishing of sharks may have also contributed to the depletion of herbivorous fish through trophic cascades [15].

The transition from a vegetated state to a desert one is also one of concern. Vegetated and desert seem to be two alternative stable states. This example has further implications in the context of global change and can feed-back into further increases of temperature without the layer of vegetation. Knowing that these transitions are irreversible due to the hysteresis cycle is worrisome as it has also profound implications in the context of human migration in search of available water.

Future Directions

The previous cases of ecosystem shifts suggest the role for non-linearities in conservation biology. The take home message is that we can not think anymore in linear terms. There is not necessarily a proportional relationship between cause and consequence. This calls for caution when assessing the consequences of global change and other types of human-induced perturbations. For example, the consequences of habitat destruction may even be worse than expected and further destruction values may cause the system to cross a threshold where extinctions may take place at a much higher rate. The existence of ecosystem shifts suggests that ecological systems may behave in qualitatively similar terms as other simpler physical systems. This is good news in the sense of being able to use a well-developed theoretical framework to make predictions of ecological systems. Near the critical points the system's macroscopic properties may be described by simple models [111].

On the other hand, due to the abrupt changes that take place in the critical points, it is very important to develop early indicators of the proximity of a system to such thresholds. For example, Kleinen et al. [56] analyzed changes in the power spectrum of temporal series and concluded that there is a reddening of the signal in the vicinity of the critical point. Similarly, there is increasing evidence that a clear early-warning signal of an ecosystem shift is an increase in the variance of the temporal series [21,124]. Further studies will be very useful in developing new and easy to measure early-warning signs. This may be very im-

portant in predicting major shifts in the state of ecosystems and the services they provide.

To sum up, ecological systems are wonderful examples of complex systems with multiple states, phase transitions, and non-linear dynamics. They provide opportunities to further apply concepts and tools from the physics of complex systems. And in the face of the multiple risks from global change, there is an urgent need to do so.

Funding was provided by the European Heads of Research Councils, the European Science Foundation, and the EC Sixth Framework Programme through a EURYI (European Young Investigator) Award.

Bibliography

Primary Literature

1. Albert R, Jeong H, Barabasi AL (2000) Error and attack tolerance of complex networks. Nature 406:378–382
2. Amaral LA, Scala A, Barthelemy M, Stanley HE (2000) Classes of small-world networks. Proceedings of the National Academy of Science of the United States of America 97:11149–11152
3. Anderson RM, May RM (1991) Infectious diseases of humans. Dynamics and control. Oxford University Press, Oxford
4. Ashby WR (1954) Design for a Brain. Chapman and Hall, London
5. Ashworth L, Aguilar R, Galetto L, Aizen MA (2004) Why do pollinator generalist and specialist plant species show similar reproductive susceptibility to habitat fragmentation? J Ecol 92:717–719
6. Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512
7. Barrat A, Barthélemy M, Pastor-Satorras V, Vespignani A (2004) The architecture of weighted complex networks. Proceedings of the National Academy of Sciences of the United States of America 101:3747–3752
8. Bascompte J, Jordano P (2007) Plant-animal mutualistic networks: the architecture of biodiversity. Annual Rev Ecol Evol Syst 38:567–593
9. Bascompte J, Melián CJ (2005) Simple trophic modules for complex food webs. Ecology 86:2868–2873
10. Bascompte J, Rodríguez MA (2000) Self-disturbance as a source of spatiotemporal heterogeneity: the case of the tallgrass prairie. J Theor Biol 204:153–164
11. Bascompte J, Rodríguez-Trelles F (1998) Eradication thresholds in epidemiology, conservation biology and genetics. J Theor Biol 192:415–418
12. Bascompte J, Solé RV (1995) Rethinking complexity: modelling spatiotemporal dynamics in ecology. Trends Ecol Evol 10:361–366
13. Bascompte J, Solé RV (1996) Habitat fragmentation and extinction thresholds in spatially explicit metapopulation models. J Anim Ecol 65:465–473
14. Bascompte J, Jordano P, Melián CJ, Olesen JM (2003) The nested assembly of plant-animal mutualistic networks. Proceedings of the National Academy of Sciences of the United States of America 100:9383–9387
15. Bascompte J, Melián CJ, Sala E (2005) Interaction strength combinations and the overfishing of a marine food web. Proceedings of the National Academy of Sciences of the United States of America 102:5443–5447
16. Bascompte J, Jordano P, Olesen JM (2006) Asymmetric coevolutionary networks facilitate biodiversity persistence. Science 312:431–433
17. Berryman AA, Millstein JA (1989) Are ecological systems chaotic: and if not why not? Trends Ecol Evol 4:26–28
18. Bersier L-F, Banašek-Richter C, Cattin M-F (2002) Quantitative descriptors of food-web matrices. Ecology 83:2394–2407
19. Bjørnstad ON, Peltonen M, Liebhold AM, Baltensweiler W (2002) Waves of larch Budmoth outbreaks in the European Alps. Science 298:1020–1023
20. Camacho J, Guimerà R, Amaral LAN (2002) Robust patterns in food web structure. Physical Review Letters 88:228102
21. Carpenter SR, Brock WA (2006) Rising variance: a leading indicator of ecological transition. Ecology Letters 9:308–315
22. Carpenter SR, Kitchell JF (1996) The trophic cascade in lakes. Cambridge University Press, Cambridge
23. Chesson P (1994) Multispecies competition in variable environments. Theor Popul Biol 45:227–276
24. Chesson PL, Ellner S (1989) Invasibility and stochastic boundedness in monotonic competition models. J Math Biol 27:117–138
25. Cohen JE (1978) Food Webs and Niche Space. Princeton University Press, Princeton
26. Connell JH (1961) Influence of interspecific competition and other factors on distribution of barnacle Chthamalus stellatus. Ecology 42:710–723
27. Connell JH (1978) Diversity in tropical rain forests and coral reefs- high diversity of trees and corals is maintained only in a non-equilibrium state. Science 199:1302–1310
28. Costantino RF, Cushing JM, Dennis B, Desharnais RA (1995) Experimentally induced transitions in the dynamic behaviour of insect populations. Nature 375:227–230
29. Costantino RF, Desharnais RA, Cushing JM, Dennis B (1997) Chaotic dynamics in an insect population. Science 275:389–391
30. Dunne JA, Williams RJ, Martinez ND (2002) Food-web structure and network theory: the role of connectance and size. Proceedings of the National Academy of Sciences, USA 99:12917–12922
31. Durrett R, Levin SA (1994) Stochastic spatial models: a user's guide to ecological applications. Philos Trans R Soc London B 343:329–350
32. Egerton FN (2007) Understanding food chains and food webs, 1700–1979. Bull Ecol Soc Am 88:50–69
33. Ellner SP, Turchin P (1995) Chaos in a nosiy world: new methods and evidence from time-series analysis. Am Nat 145:343–375
34. Erdös P, Rényi A (1959) On Random Graphs. Pub Math 6:290–297
35. Fagan WF, Hurd LE (1994) Hatch density variation of a generalist arthropod predator: population consequences and community impact. Ecology 75:2022–2032
36. Feigenbaum M (1978) Quantitative universality for a class of non-linear transformations. J Stat Phys 19:25–52
37. Fortuna MA, Bascompte J (2006) Habitat loss and the structure of plant-animal mutualistic networks. Ecol Lett 9:281–286

38. Glass L, Mackey MC (1990) From Clocks to Chaos. The Rhytms of Life. Princeton University Press, Princeton
39. Guimarães PR Jr, de Aguiar MAM, Bascompte J, Jordano P, dos Reis SF (2005) Random initial conditions in small Barabasi–Albert networks and deviations from the scale-free behavior. Phys Rev E 71:037101
40. Guimarrães PR Jr, Machado G, de Aguiar MAM, Jordano P, Bascompte J, Pineiro A, dos Reis SF (2007) Build-up mechanisms determining the topology of mutualistic networks. J Theor Biol 249:181–189
41. Guimarrães PR Jr, Sazima C, Furtado dos Reis S, Sazima I (2007) The nested structure of marine cleaning symbiosis: is it like flowers and bees? Biol Lett 3:51–54
42. Guimerà R, Amaral LAN (2005) Functional cartography of complex metabolic networks. Nature 433:895–900
43. Gurney WSC, Nisbet RM (1998) Ecological Dynamics. Oxford University Press, Oxford
44. Hanski I (1999) Metapopulation Ecology. Oxford University Press, Oxford
45. Hanski I, Moilanen A, Gyllenberg M (1996) Minimum viable metapopulation size. Am Nat 147:527–541
46. Hassell MP, Lawton JN, May RM (1976) Patterns of dynamics behaviour in saingle-speceis populations. J Anim Ecol 45:471–486
47. Hassell MP, Comins HN, May RM (1991) Spatial structure and chaos in insect population dynamics. Nature 353:255–258
48. Hassell MP, Comins HN, May RM (1994) Species coexistence and self-organizing spatial dynamics. Nature 370:290–292
49. Hastings A (1993) Complex interactions between dispersal and dynamics: lessons from coupled logistic equations. Ecology 74:1362–1372
50. Hastings A, Higgins K (1994) Persistence of transients in spatially-structured ecological models. Science 263:1133–1136
51. Hastings A, Hom CL, Ellner S, Turchin P, Godfray HCJ (1993) Chaos in ecology: is mother nature a strange attractor? Annual Rev Ecol Syst 24:1–33
52. Hughes TP (1994) Catastrophes, phase-shifts, and large-scale degradation of a Caribbean coral reef. Science 265:1547–1551
53. Jordano P, Bascompte J, Olesen JM (2003) Invariant properties in coevolutionary networks of plant-animal interactions. Ecol Lett 6:69–81
54. Kaneko K (1984) Period-doubling of Kink-antiking patterns, quasi-periodicity in Antiferro-like structures and spatial intermittency in coupled map lattices: towards a prelude to a Field Theory of Chaos. Prog Theor Phys 72:480–486
55. Kaneko K (ed) (1992) Special issue on Coupled Map Lattices. Chaos 2
56. Kleinen T, Held H, Petschel-Held G (2003) The potential role of spectral properties in detecting thresholds in the Earth system: application to the thermohaline circulation. Ocean Dyn 53:53–63
57. Kokkoris GD, Troumbis AY, Lawton JH (1999) Patterns of species interaction strength in assembled theoretical competition communities. Ecol Lett 2:70–74
58. Krause AE, Frank KA, Mason DM, Ulanowicz RE, Tayor WW (2003) Compartments revealed in food-web structure. Nature 426:282–285
59. Kuramoto Y (1984) Chemical Oscillations, Chaos and Turbulence. Springer, Berlin
60. Lafferty K, Dobson A, Kurtis A (2006) Parasites dominate food web links. Proceedings of the National Academy of Sciences of the United States of America 103:11211–11216
61. Lande R (1987) Extinction thresholds in demographic models of territorial populations. Am Nat 130:624–635
62. Lande R, Engen S, Saether BE (2003) Stochastic Population Dynamics in Ecology and Conservation. Oxford University Press, Oxford
63. Levin SA (1992) The problem of pattern and scale in ecology. Ecology 73:1943–1967
64. Levin SA, Durrett R (1996) From individuals to epidemics. Philos Trans R Soc London B 351:1615–1621
65. Levin SA, Segel LA (1976) Hypothesis for the origin of planktonic patchiness. Nature 259:659
66. Levins R (1969) Some demographic and genetic consequences of environmental heterogeneity for biological control. Bull Entomol Soc Am 15:237–240
67. Lindenman RL (1942) The trophic dynamic aspect of ecology. Ecology 23:399–418
68. Lotka AJ (1925) Elements of Physical Biology. Williams and Wilkins, Baltimore
69. MacArthur RM (1964) Environmental factors affecting bird species diversity. Am Nat 98:387–397
70. Margalef R (1958) Information theory in ecology. Gen Syst 3:36–71
71. Margalef R (1968) Perspectives in theoretical ecology. Chicago University Press, Chicago
72. Maron JL, Harrison S (1997) Spatial pattern formation in an insect host-parasotoid system. Science 278:1619–1621
73. Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. Science 296:910–913
74. May RM (1972) Will a large complex system be stable? Nature 238:413–414
75. May RM (1974) Biological populations with non-overlapping generations: stable points, stable cycles, and chaos. Science 186:645–647
76. May RM, Oster GF (1976) Bifurcations and dynamics complexity in simple ecological models. Am Nat 110:573–599
77. McCann K, Hastings A, Huxel GR (1998) Weak trophic interactions and the balance of nature. Nature 395:794–798
78. Melián CJ, Bascompte J (2002) Complex networks: two ways to be robust? Ecol Lett 5:705–708
79. Melián CJ, Bascompte J (2004) Food web cohesion. Ecology 85:352–358
80. Memmott J, Waser NM (2002) Integration of alien plants into a native flower-pollination visitation web. Proceedings of the Royal Society Series B 269:2395–2399
81. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alton U (2002) Network motifs: simple building blocks of complex networks. Science 298:824–827
82. Morales CL, Aizen MA (2006) Invasive mutualisms and the structure of plant-pollinator interactions in the temperate forests of the north-west Patagonia, Argentina. J Ecol 94:171–180
83. Murdoch WW (1994) Population regulation in theory and practice. Ecology 75:271–287
84. Murdoch WW, Chesson J, Chesson PL (1985) Biological control in theory and practice. Am Nat 125:344–366

85. Murray JD (1989) Mathematical Biology. Springer, Heidelberg
86. Myers RA, Worm B (2003) Rapid worldwide depletion of predatory fish communities. Nature 423:280–283
87. Nee S (1994) How populations persist. Nature 367:123–124
88. Neutel A, Heesterbeek JAP, Ruiter PC (2002) Stability in real food webs: Weal links in long loops. Science 296:1120–1123
89. Nicholson AJ, Bailey VA (1935) The balance of animal populations. 1st Proceedings of the Zoological Society of London 3:551–598
90. Odum HT (1956) Primary production in flowing waters. Limnol Oceanogr 1:102–117
91. Olesen JM, Eskildsen L, Venkatasamy S (2002) Invasion of pollination networks on oceanic islands: importance of invader complexes and epidemic super generalists. Divers Distrib 8:181–192
92. Olesen JM, Bascompte J, Dupont YL, Jordano P (2007) The modularity of pollination networks. submitted to Proceedings of the National Academy of Sciences of the United States of America
93. Pascual M (1993) Diffusion-induced chaos in a spatial predator-prey system. Proceedings of the Royal Society of London B 251:1–7
94. Pascual M, Dunne JA (eds) (2006) Ecological Networks: Linking Structure to Dynamics in Food Webs. Oxford University Press, Oxford
95. Paine RT (1969) A note on trophic complexity and community stability. Am Nat 103:91–93
96. Paine RT (1992) Food-web analysis through field measurements of per capita interaction strength. Nature 355:73–75
97. Pauly D, Christensen V, Dalsgaard J, Froese R, Torres F Jr (1998) Fishing down marine food. webs. Science 279:860–863
98. Pimm SL (1979) The structure of food webs. Theor Popul Biol 16:144–158
99. Pimm SL (1982) Food Webs. University of Chicago Press, Chicago
100. Pimm SL, Lawton JH (1980) Are food webs divided into compartments? J Anim Ecol 49:879–898
101. Prill RJ, Iglesias PA, Levchenko A (2005) Dynamic properties of network motifs contribute to biological network organization. PLoS Biology 3811:e343
102. Raffaelli D, Hall SJ (1992) Compartments and predation in an esturine food web. J Anim Ecol 61:551–560
103. Raffaelli D, Hall S (1995) Integration of Patterns and Dynamics. In: Polis G, Winemiller K (eds) Food Webs. Chapman and Hall, New York, pp 185–191
104. Ranta E, Kaitala V (1997) Travelling waves in vole population dynamics. Nature 390:456–456
105. Rasmussen DR, Bohr T (1987) Temporal chaos and spatial disorder. Phys Lett A 125:107, 125:107–110
106. Schaffer WM, Kot M (1986) Chaos in ecological systems: the coals that Newcastle forgot. Trends Ecol Evol 1:58–63
107. Scheffer M, Carpenter SR, Foley J, Folke C, Walker B (2001) Catastrophic shifts in ecosystems. Nature 413:591–596
108. Schroeder MM (1991) Fractals, Chaos, Power Laws. Freeman, New York
109. Sibly RM, Barker D, Hone J, Pagel M (2007) On the stability of populations of mammals, birds, fish and insects. Ecol Lett 10:970–976
110. Simon H (1955) On a class of skewed distribution functions. Biometrika 42:425–440
111. Solé RV, Bascompte J (2006) Self-organization in complex ecosystems. Princeton University Press, Princeton
112. Solé RV, Montoya JM (2001) Complexity and fragility in ecological networks. Proceedings of the Royal Society of London Series B 268:2039–2045
113. Solé RV, Valls J (1991) Order and chaos in a two-dimensional Lotka–Volterra coupled map lattice. Phys Lett A 153:330–336
114. Solé RV, Bascompte J, Valls J (1992) Stability and complexity of spatially extended two-species competition. J Theor Biol 159:469–480
115. Solé RV, Valls J, Bascompte J (1992) Spiral waves, chaos and multiple attractors in lattice models of interacting populations. Phys Lett A 166:123–128
116. Sugihara G, May RM (1990) Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. Nature 344:734–741
117. Sugihara G, Schoenly K, Trombla A (1989) Scale invariance in food web properties. Science 245:48–52
118. Taylor PJ (2005) Unruly Complexity: Ecology, Interpretation, Engagement. University of Chicago Press, Chicago
119. Thompson JN (1994) The Coevolutionary Process. University of Chicago Press, Chicago
120. Tilman D, Wedin D (1991) Oscillations and chaos in the dynamics of a perennial grass. Nature 353:653–655
121. Turing A (1952) On the chemical basis of morphogenesis. Philos Trans R Soc Series B 237:37–72
122. Turner MG, Gardner RH, Dale VH, O'Neill RV (1989) Predicting the spread of disturbances across heterogeneous landscapes. Oikos 55:121–129
123. Ulanowicz RE, Wolf WF (1991) Ecosystem flow networks: loaded dice. Math Biosci 103:45–68
124. van Nes EH, Scheefer M (2003) Alternative attractors may boost uncertainty and sensitivity in ecological models. Ecol Model 159:117–124
125. Volterra V (1926) Variations and fluctuations of the number of individuals in animal species living together. Translation In: Chapman RN (1931) Animal Ecology. Wiley, New York, pp 409–448
126. Wiener N (1948) Cybernetics or control and communication in the animal and the machine. Wiley, New York
127. Wootton JT (1997) Estimates and tests of per-capita interaction strength: diet, abundance, and impact of intertidally foraging birds. Ecol Monogr 67:45

Books and Reviews

Hastings A, Hom CL, Ellner S, Turchin P, Godfray HCJ (1993) Chaos in ecology: is mother nature a strange attractor? Annual Rev Ecol Syst 24:1–33

Levin SA (1999) Fragile Dominion: Complexity and the Commons. Perseus Publishing, Reading

May RM (1973) Stability and Complexity in Model Ecosystems. Princeton University Press, Princeton

Prigogine I, Stangers I (1984) Order Out of Chaos. Bantam Doubleday Dell Publishing Group, Toronto

Stewart I (1989) Does God play dice? The mathematics of chaos. Basil Blackwell, Oxford

Ecological Topology and Networks

Örjan Bodin
Stockholm Resilience Centre, Stockholm University, Stockholm, Sweden

Article Outline

Glossary
Definition of the Subject
Introduction
Network Analysis
Food Webs
Network Perspective of Fragmented Landscapes
Summary
Future Directions
Bibliography

Glossary

Network A network is (1) a set of system entities (nodes), which may be interconnected through (2) links.

Graph A graph is the mathematical notation of a network, i. e., it is an abstraction of a network.

Nodes Nodes are the separate entities of a network, i. e., the system components that may be interconnected. Sometimes they are called vertices.

Links Links are the realizations of the possible connections among the nodes of a network. Sometimes they are called edges.

Topology Topology, in this context, represents the pattern of interconnections, i. e., the links, of a network. A topological analysis is hence concerned with analyzing the structural characteristics of the set of links defining the network.

Network analysis Network analysis refers to all kinds of quantitative approaches used in analyzing the topology of networked systems. In doing so, methods from the mathematical branch of graph theory are often applied.

Definition of the Subject

Ecology is, simply speaking, the science of ecosystems, i. e., sets of interacting species constrained by the physical environment. Due to earth's enormous species diversity, species' patterns of interactions quickly become very complex and thus difficult to oversee, although it is clear that the interaction patterns themselves often have profound effects on the behaviors and functioning of the ecosystems. To enable systematic pattern analyses, it is often favorable to represent these patterns of interactions as *networks* where the nodes are some sort of biological entities and the links represent some sort of interaction between these entities. The entities can e. g., represent species, but they could also represent individuals or groups of organisms. The links could e. g., represent trophic interactions ("who eats whom"). By representing interacting species as a network, analytical focus is set on the actual pattern, or *topology*, of the interactions themselves. Topological analysis of ecological systems has a relatively long history in ecology which can be exemplified by the long-lasting scientific debate, spurred by Sir Robert May in the 1970s when he, against prevailing interpretation, suggested that an increased number of species interaction would actually lead to decreased ecosystem stability. More recently, an increased interest among various scientific disciplines on network approaches in complex systems research has re-energized the topological perspective of ecosystem research.

Introduction

In nature, species interact in many different ways; no species exists in isolation. Understanding these interactions and how these affect individual organisms, species and whole ecosystems are, therefore, key to a systemic understanding of the natural environment. In the early twentieth century, Lotka and Voltera [1,2] paved the way for theoretical and mathematical approaches in understanding predator-prey interaction and the resulting dynamics of species populations. In the 1950s, the Odum brothers [3] revolutionized ecology by emphasizing the need for a systemic perspective of the natural environment [4]. In order to take the field of ecology beyond a mainly descriptive science, and to find solution to challenges facing the natural environment, they argued that better understanding of the large-scale properties of the environment is needed [4].

Ecosystems, although over time defined and/or perceived in many different ways, are basically systems of interacting species limited by constraints arising from the physical environment. The Odum brothers originally modeled ecosystems as sets of components (e. g., species) and flows of energy (the common denominator) cycling through these components. Thus, they essentially laid the foundation for seeing the environment as a networked system consisting of nodes and links. What constitutes a node depends on the question at hand; it could be a species, a group of similar species, an individual organism, groups of organisms, physical objects, etc. The links, i. e., the relations between the system components of interest, also depends on the chosen question; they could e. g., be flows

of energy going from prey to predators in food webs (e. g., [5]), or flows of genes spread through species dispersals among localized populations (e. g., [6]). Networks can be viewed as maps that outline these local interactions and preserve their importance at the system level [7]. The perspective of networks carries the advantage of simultaneously addressing the members of the systems as well as the patterns of their interactions. This modeling approach, emphasizing localized interaction between separated parts of the system, captures some of the fundamental characteristics of a complex adaptive system [8].

As mentioned earlier, the ecological impact of various characteristics of patterns of interactions has been studied and debated for a considerable period of time. Of focal interest here is, however, research approaches paying special attention to the actual pattern, or *topology*, of the interactions themselves. For example, how may the distribution of links among the nodes in food webs affect ecosystem stability in terms of risk of species extinctions? Alternatively, how may the average topological distance between pairs of habitat patches in a fragmented landscape affect metapopulation dynamics? These, and other topological aspects of ecological systems, are the subject of this review.

In order to analyze networked systems, formal methods addressing the topology are needed. Such methods have, fortunately, been developed in various scientific disciplines such as sociology, computer sciences and mathematics over a considerable amount of time (e. g., [9]). These methods aim at quantifying different topological aspects of concern, some of which are briefly described in the following section. Recently, the interest for topological analyses of various kinds of systems has grown tremendously; a development that e. g., has generated an abundance of new insights and approaches in topological analyses. This development has, among other things, contributed to re-energize the topological perspective in ecological research.

Network Analysis

A system of interconnected entities, i. e., a network, is mathematically represented as a graph. Graphs consist of nodes and links. Nodes are the terminal points or intersection points of the graph (sometimes called vertices). Links represent connections between nodes and represent the structure of the network over which interaction occurs. There is an entire branch of mathematics called graph theory that deals with the analysis of such graphs. Furthermore, network-oriented analyses are undertaken in several other disciplines; thus methodological, technical and theoretical developments of relevance for networked systems are taking place across disciplines. An example of an interdisciplinary endeavor is the fast-growing organization INSNA (International Network for Social Network Analysis, see http://www.insna.org) consisting of sociologists, mathematicians, physicists, computer scientists and others that are mainly occupied with studying patterns of social interactions (i. e., social networks). Thus, the term network analysis will be used here when referring to all kinds of quantitative approaches in analyzing the patterns of interconnections in networked systems.

Modeling a Networked System

In conducting topological analyses, the first step would be to represent the system under study as a network, i. e., to define what entities will constitute the nodes, and what kind of relations among the nodes will constitute the links. Modeling a system as a network is in some cases straight forward, because it is more or less obvious what entities will make up the nodes, and what kind of relations would constitute the links. For example, in studying pattern of friendships among students in a classroom each student would constitute a node, and the reported friendship between any two students would be represented by a link. In other cases, it might be less obvious how to define a node. In studying the dynamic interactions in a ecosystem, should the nodes be represented by individual species, groups of similar species, or even by individual organisms? The method by which one chooses a suitable level of abstraction will depend on the research questions. The problem is, as always in modeling, to choose a level of abstraction that is as simple and aggregated as possible, but still fine-grained enough to capture the essential characteristics of the system in order to help answer the questions at hand.

The issue of choosing an appropriate level of abstraction likewise applies when defining the links. Often many different kinds of relations exist between the different entities (nodes) of the system under study; thus the question is which type of relations should be modeled? Furthermore, in many cases the strengths of the relations are of interest. The strength of a relation could be though of as the intensity, and/or the frequency of interactions, the flow of energy, material, information etc between the nodes in question. Although the focus here is primarily on analysis of the topological properties of a networked system (where differences in the strength of the links are not considered), the issue of link strength is important when e. g., constructing the network. If the strength, or intensity, of the type of relation under consideration varies, it might be useful to define a threshold; and relations between any pair of nodes with strengths below such threshold would

thus be ignored. Therefore, when defining the threshold value, the resulting topology is affected since links below the threshold are omitted. Also, in this context it is important to mention that many commonly used measures defining various network characteristics can, in addition to purely topological characteristics, also take links strengths into account.

Finally, the links of the network representing the studied system could be bidirectional or unidirectional. For example, in food webs the predator consumes the prey, which is an example of a unidirectional flow of energy (which in a network context translates to a unidirectional link). If, however, two species prey on each other, the flow of energy is bidirectional.

Topological Characteristics

In this section, three important characteristics of networks, and some of the associated and commonly used measures defining these characteristics, are briefly reviewed. These are degree distribution, modularity and centrality. How these are of relevance in ecology will be discussed and exemplified in coming sections. Naturally, there are many characteristics that define a network other than those presented here. The applicability of various existing network measures in network-oriented ecological studies is almost a research topic in itself e. g., [10,11,12].

Degree Distribution and Small-World Properties In most real-world networks, links are very unevenly distributed among the nodes (see e. g., [13] for a review). If, for example, all links were distributed randomly, the degree distribution would follow a Poisson distribution. There are, however, lots of examples of real networks not following this distribution; thus there must be processes other than chance alone that are in play when networks are formed and shaped over time (e. g., [13]). Of recent interest are the so-called scale free networks following a power-law degree distribution, meaning that most nodes have few links, but that some rare nodes posses very many links (these nodes are often called hubs). Such networks are quite robust (in terms of the risk for severe network fragmentation) to random node removals (since most of the nodes have very few links), but they would be very vulnerable to a targeted removal of the hubs [14].

Somewhat related to the degree distribution is the concept of *small world* networks [15]. A small world network displays a high level of clustering, meaning that two nodes that both possess links to a common third are much more likely to be directly linked to each other as compared to the likelihood that any other arbitrary pair of nodes should be directly linked. In spite of this high degree of clustering, in a small world network the average topological distance between any arbitrary chosen pair of node remains relatively short, thus implying that there are still many links that cross boundaries and therefore link together different clusters. A small world network does not necessarily follow any particular degree distribution, but it has been shown that scale-free networks often display small-world characteristics see (e. g., [13]).

Modularity Within a network there may be groups (or *modules*) of nodes that, from a topological perspective, distinguish themselves from the rest. For example, it could be that these groups of nodes are more internally than externally interlinked, i. e., the distinction of groups is based on a high density of interconnecting links among each group's members. In this way, a group would have a relatively high frequency of direct (or indirect) relations within the group compared to outside the group. Examples of such group-assessment methods are LS-sets and lambda sets (see [16] and references therein).

A specific example of another type of group definition is the clique ([17] and references therein). In a clique every member is connected to every other member; thus this definition of a group does not define members based on their relative cohesion versus non-member – instead it uses an absolute criterion for defining a group. The definition of a clique can be extended to account for directional and weighted links as well.

Generally, methods of assessing groups can be classified as either generating hierarchically nested groups, or as generating groups that can partially overlap. In the former case, an individual node cannot be a member of more than one hierarchal branch. Furthermore, a decomposition algorithm used to identify hierarchical branches can be either divisive or agglomerative, where agglomerative algorithms have been more commonly used [18]. If the focus is to find cores of strongly and/or intensely interlinked groups, agglomerative methods are preferred since they will identify groups using a bottom-up approach, starting with the most densely connected subgroups. Divisive algorithms, on the other hand, uses a top-down approach and iteratively divide the network in smaller and smaller branches.

A different way of distinguishing groups within a network is to group nodes according to the set of relations they have with one another (called equivalence, see e. g., [17]). Here, two members of a group would have similar sets of relations to others, i. e., they occupy similar positions in the network. This kind of group can e. g., be useful to locate functionally similar groups of species in a food web (as will be described below).

Centrality A fundamental topological characteristic of a node in a network is its level of *centrality*. The concept of centrality is devoted to analyzing the position of nodes' in the network. The underlying assumption is that some positions are more favorable than others in terms of the influence the nodes occupying them can exert on others (for an introduction and review of the literature, see e. g., [17]). There are, however, numerous ways to exert influence, and accordingly many different measures of centrality have been developed – each focusing on different topological aspects. Here some of these are presented:

1. Degree centrality. This is the number of links a node possesses. In a network with directed links, one could distinguish between in-degree and out-degree centrality.
2. Betweenness centrality (see [19]). This measure assesses how much "in-between" a particular node is, based on how many shortest pathways (connecting other nodes) that goes through this particular node.
3. Closeness centrality [20]. This measure assesses how close (from a topological perspective) a particular node is to the rest of the nodes in the network.

Food Webs

In nature, species interact in many ways. One of these types of interactions is predation, i. e., when a species eat another species. Predation can also occur within species (cannibalism) when larger individuals eats smaller one. In a food web, the nodes represent a set of interacting species and the links represent the structure of energy flows, i. e., the links represent who eats who. Typically, the links in a food web are directed, meaning that a specific predator species consumes one or several specific prey species, and not vice versa (even if bi-directed size-dependent predation does occur, which may e. g., pose challenges in modeling food webs, see e. g., [21]).

The strength of the link between species A and species B can be assessed in various ways. First, it can be modeled as binary, i. e., either species A eats species B or not, and the analytical focus is thus set on the food-web's topology alone. The strength could, however, be weighted according to the change in the abundance of species A following the removal of species B (functional edges, see [5]). It could also be assessed based on the flow of energy between species A and species B. Other assessment methods are also available (for an overview, see [22]).

When assessing the link strength as in the last two examples, research focus is not set solely on the topology of the food-web. This increases the complexity of the analysis, since the researcher has to (1) somehow assess the link strength, and (2) take the assessed strength into account in studying the food-web structures. In some cases the increased level of complexity may be redundant (i. e., a topological analysis would suffice given the particular research question at hand), but in other cases the more complex approach might be needed to find structural network characteristics of interest (e. g., [5]). An example of the latter is a study by Krause et al. [23], where it was shown that the modular structure for some of the studied food webs was only detectable when taking link strength into account. Another example is given by Scotti and colleagues [24], where they showed that the ranking order of the most central nodes were highly affected when link weights were taken into account.

Much of early research, up until the late 1980s, in food web topology was in fact highly affected by problems associated with assessing link strength. Much of the empirical data available at the time were lacking information of the weakest links, i. e., the links representing low levels of energy flows (corresponding to gut content $<5\%$, see e. g., [25] for a review). However, weak links abound in empirical food webs (e. g., [26,27,28]). Later research has subsequently showed that when weak links are taken into account, structural patterns of food webs change drastically (e. g., [29]). Up to this point, theoretical development was in many cases actually based on incomplete data, often leading to false conclusions on the structural properties of food webs. One example was the false assumption of the independence between a species average number of links and the total number of species in the food web. Instead, the number of links per species increases as the number of available species increase [30].

Finally, of significant interest with respect to the topological research approach of complex ecosystem as described here, is the insight that the actual pattern of linkages, i. e., the food web topology alone (regardless of link strengths) appears to be extremely important in defining ecosystem characteristics (e. g., [31]). In this context, weak interactions representing relatively small rates of energy flows can have a large impact on systemic properties whereas interactions with large flows can reversely have a relatively small impact (see [32] for a review of a number of important contributions in this research).

Degree Distribution

In terms of degree distributions, it appears that food webs, like many other types of networks, often experience a skewed link distribution where a few nodes pos-

sess many links [33]. This is completely different from what would be expected if the links were distributed randomly. This makes food webs, in accordance with other systems experiencing, a skewed link distribution, quite robust with respect to a random removal of nodes (i. e., species), but quite vulnerable to a targeted removal of the most connected nodes [33]. The distribution of links in food webs is generally different from the previously mentioned scale-free degree distribution in that the tail of the distribution towards very high degree centralities is truncated (see e. g., [7] and references therein). Furthermore, since food webs often tend to experience a relatively high density of links as compared to many other kind of networks, the degree distribution accordingly gets more uniform.

In using a method designed to find subgroups (identical to a method named k-core used to find cohesive groups in social network, see [34]), Melián and Bascompte [35] found that of the five studied food webs, all had one dense and cohesive core subgroup. In these dense subgroups all species have a least k links to other group members. Thus, it is not only the distribution of links that is skewed; the species with most links appear to be connected to other highly connected species. Such a positive correlation between a species number of links and its neighbors average number of links (connectivity correlation, see e. g., [36]), has been reported in other studies as well (e. g., [37]).

Sparsely connected food webs do also experience the small world effect [15], but this gets much less clearly articulated as the number of links increases [38,39,40]. Furthermore, it appears that the average distance between any two species in a food web is even smaller than in a small world network [33,41]. Thus, food webs could be considered as being rather tight when, on average, only two intermediate links are needed to jump between any two species in the network, and that >95% of the species are within three degrees of separation from each other [41].

In food webs with a high number of links, the small world effect often disappears. If, however, a food web's minimal spanning tree is analyzed, some interesting insights can be found. A minimal spanning tree is essentially a simplified version of the original food web where only the links that are minimizing the distances from all species to the basal species are kept, the rest are removed (e. g., [25]). Garlaschelli and colleagues [38] found that the minimal spanning trees experienced both the small-world effect and adhered to a scale-free degree distribution irrespective of the size or the link density of the set of seven different food webs studied. This suggests that there may exist some core structural principles upon which all food webs are built upon.

Modularity

With the significantly increased resolution in empirical food web data collection starting in the early 1990s, it has become increasingly feasible to look at the possibilities for modular structures in food webs (see [25] for a review). Modularity can, however, be defined – topologically – in (at least) two different ways. Modules can be defined as groups of species that are structural similar to each other, or it can be defined as groups of species that are interacting more intensively and/or frequently among themselves than with other species outside the group (hereafter called compartments). The former definition is exemplified in the concepts of trophic role or trophic level, although neither of these (and other similar) terms has been fully defined and collectively accepted in the ecological research community (see e. g., [10]). From a topological perspective, an interesting approach is to define, or single out, groups of species with similar trophic roles by analyzing the structure of the food web. Luczkovich and colleagues [10] used the network analytical approach based on regular equivalence, developed within the field of social network analysis (e. g., [17]), to partition species that play the same structural roles into different isotrophic classes (see Fig. 1). The approach, as described earlier, is based on finding positionally equivalent species in the food web, and provides a way to mathematically formalize trophic position, trophic group and trophic niche.

Research studying compartmentalization in food webs, according to the latter definition of modularity based on the frequency and/or intensity of interactions, has gained interest in recent years. One reason for the interest in compartmentalization lies in its potential effect on ecosystem stability and robustness (or using different terminology, resilience) (see e. g., [42]). A food web consisting of just one single coherent group of species would likely help to propagate both harmful and beneficial effects throughout the ecological community, whereas the presence of internally coherent but externally isolated or only weakly connected subgroups would reduce the possibilities for such large-scale propagations. This has lead to the hypothesis that intermediate modularity may be the most beneficial structural composition of food webs (see e. g., [43]). Such structures are formed in networks where coherent groups of species with strong and/or frequent internal interactions are only weakly coupled to other groups. If this hypothesis holds true, one would expect to find modular structures in food webs since, over time, only the most robust food webs would be expected to prevail despite disturbances in the past.

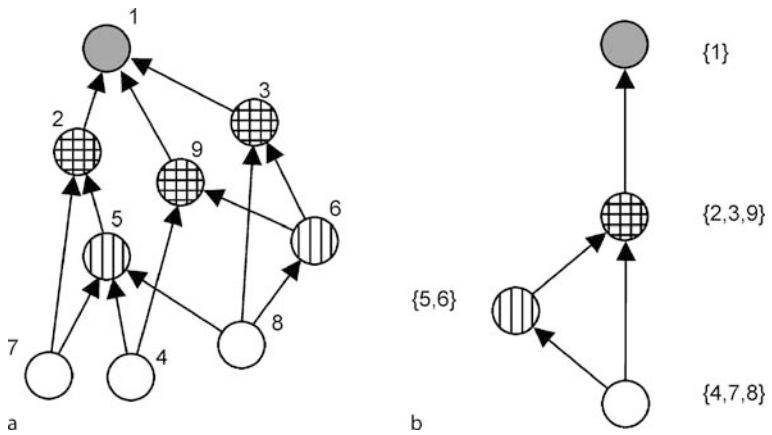

Ecological Topology and Networks, Figure 1
a A food web where nodes (species) are patterned according to their regular equivalence, and b the same food web, but here are regular equivalent species grouped together. Thus, a simplified image of the food web is created where the relations among groups of species occupying analogous positions is presented. (Source: [10])

Food webs appear to be more coherent (i. e., less compartmentalized) than many other kind of biological networks (see e. g., [37,44]). This tendency seems to be derived, at least partly, from the relatively high link density that many food webs exhibit (e. g., [39]). It is obviously harder to distinguish subgroups when every species is connected to many others. In spite of this, recent studies have been able to find evidence for compartmentalization in food webs. Girvan and Newman [45] used e. g., a divisive hierarchical decomposition algorithm to detect ecologically relevant subgroups in a coastal food web. In addition, Krause and colleagues [23] were able to detect modularity in three out of five complex food webs (see Fig. 2 for an example) using a group assessment methods developed for analysis of groups in social networks [46].

Interestingly, Krause and colleagues barely found any compartments when the weakest links were left unconsidered, or when the food web consisted of very few species. In addition, the ability to detect compartments increased by taking link strengths into account. However, in another recent study by Dunne and colleagues [39], the presence of compartmentalization was not very pronounced. They found only five of sixteen food webs exhibiting a clear tendency for compartmentalization. Interestingly, but perhaps not surprisingly considering the previous argument, these were the sparsest food webs among the studied set.

Hence, although it seems clear that food webs often experience compartmentalization, existing evidence of a clear trend of intermediate level of compartmentalization in respect of stability is mainly from anecdotal evidence and models [7]. In the previously mentioned study by Krause and colleagues [23], they simulated the effect of selectively removing species, one at the time, in one of the two distinct compartments in the Chesapeake Bay food web (each compartment corresponded to benthic and pelagic species respectively) alternatively. The resulting loss of links was found to be significantly higher in the compartment from where the species were removed. Thus this simulation supported the idea that compartmentalization reduces the spread of harmful effects. Direct evaluation of the impact of intermediate levels of compartmentalization for food web stability will, however, require long term food web data, for many different systems, with varying levels of compartmentalization which have experienced disturbance at some point during their monitoring.

Similar arguments as the ones proposing the beneficial effect of intermediate modularity are to be found in the discussion of the importance of weak links. Not only did the increased resolution of food web data reveal the abundance of weak links leading to the realization that food webs are much more complex then previously thought. Weak links may also help to stabilize oscillatory dynamics resulting from strongly interacting species [47], but see also May [48] for a different view on stability. In coupling oscillatory pairs, or groups, of species to other more stable groups of species, weak links act to reduce the amplitude of variations, thus helping to create a more stable system. Furthermore, if hypothesizing that interaction strength among species within coherent compartment is high, it may seem natural to assume that links bridging different compartments would be mostly weak. The presence of weak links thus suggests that the food web exhibits a modular structure (see e. g., [7]), although support for such statement is scant.

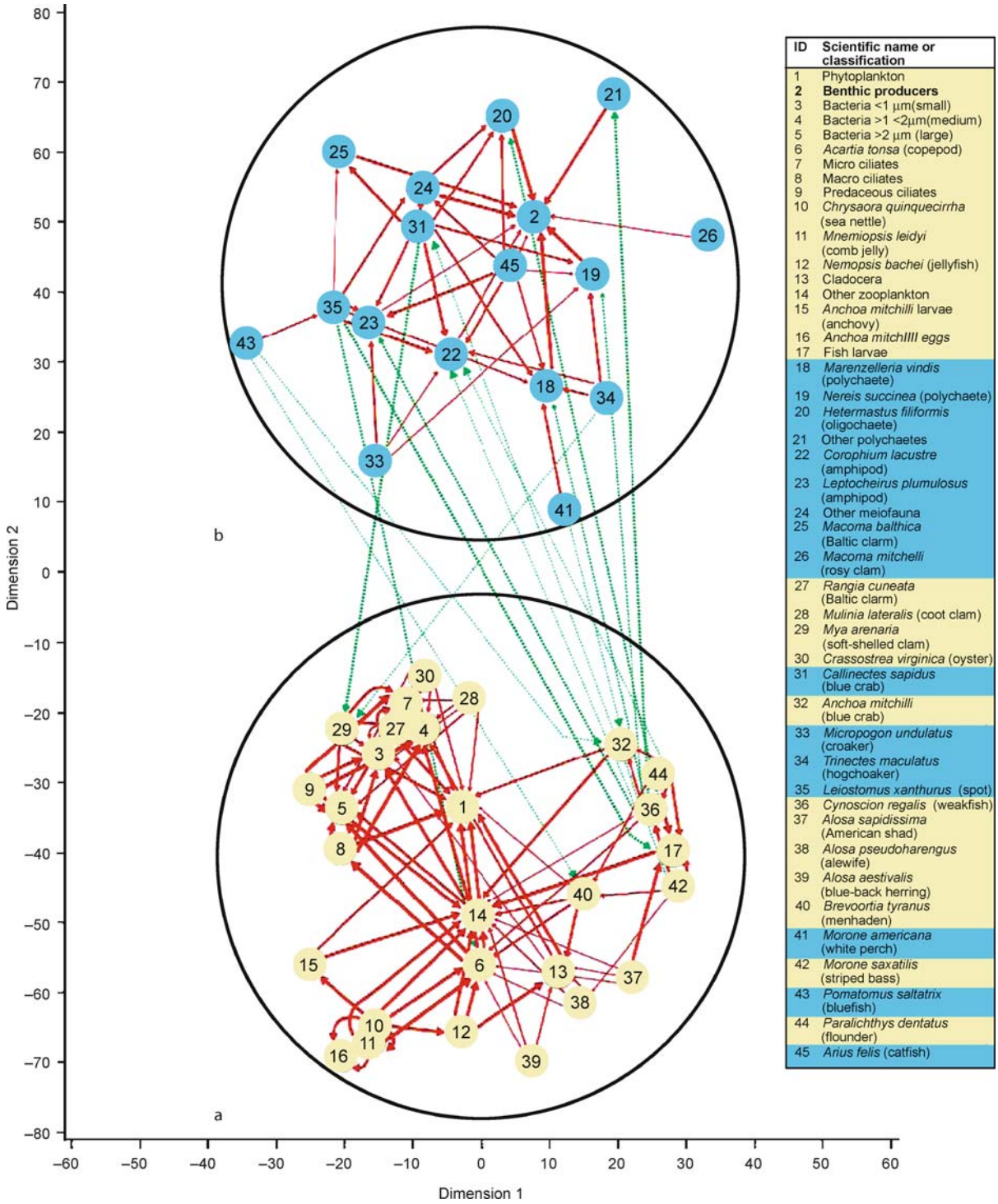

Ecological Topology and Networks, Figure 2

Graphical display of the results for the Chesapeake Bay food web with 45 taxa and weighted by interaction strength. Units are relative distances based on the inverse of the density of interactions. Within-compartment distances were decreased by a factor of 6.2 for aesthetic purposes. *Circles* indicate compartment boundaries and numbers identify taxa (*yellow*, within compartment a; *blue*, within compartment b). *Arrows* indicate interactions between taxa (*solid red*, within compartment; *dashed green*, between compartments; thickness indicates rank of associated interaction strength) and point from predator to prey (Source: [23])

Network Centrality and Keystone Species

The concept of keystone species in ecology is different, but related to, the concept of dominant species. A dominant species is a species that is high in abundance, and exert a large impact on the ecosystem where they are situated (see e. g., [33]). Keystone species, on the other hand, differ from dominant species in that their effects on the ecosystem are much larger than would be predicted from their abundance alone (e. g., [49]). It has been suggested that a keystone species is a species with a disproportionably high number of links in the food web [50]. Dunne et al. [33] have also put forward the idea of a structural keystone, i. e., a species that exerts influence on the basis of its structural position within the food web, and not only on the basis of the number of links it has to others.

The recent interest in using network analytical approaches in studying food webs have caused a renewed interest in the debate on whether increased ecosystem complexity (levels of diversity and connectivity) leads to less stability. Although many food webs do not experience a truly scale-free degree distribution, they nonetheless appear to be much more sensitive to a targeted removal of highly connected species than to a random removal (as described above). This is an effect of the skewed degree distribution, where the removal of the most connected species (i. e., the species with most links – highest degree centrality) will increase the average path length (i. e., increase the distance between consumer species and basal resources), and may also lead to network fragmentation. If a food web fragments into smaller isolated pieces, species might become excluded from their energy sources (their prey and/or basal resources); leading to secondary extinctions. From a purely topological perspective, a secondary extinction occurs when a non-basal species loses all of if prey species (e. g., [33]). Hence, in considering the risk of network fragmentation leading to secondary extinctions, and degree centrality is obviously of great relevance in assessing individual species' importance. Furthermore, the number of secondary extinctions following random species removal is often highly non-linear. Dense food webs are relatively insensitive to species loss up to a certain point. Beyond this point, the number of secondary extinctions sharply increases [33].

Dunne and colleagues [33] have, however, demonstrated that degree centrality alone cannot always predict the number of secondary species extinctions following species loss. In some cases the removal of species with a relatively few links led to a disproportional large number of secondary extinctions (hence leading to their notion of structural keystone species). Such structural keystone species might play an important role in maintaining the minimum spanning tree, i. e., they may possess exclusive, non-redundant links that, more so than other links, help keep the food web together (see discussion below on non-redundant links and structural holes).

The insight that a species structural position in a food web may have a strong impact on the ecosystem is not entirely new. There are several examples in the literature of cascading effects taking place on the level of entire ecosystems following the extinction of certain keystone species. One of the most famous examples is the large scale destruction of underwater kelp forests following the extinction of sea otters due to intensive hunting for the fur trade [51,52,53]. Sea otters predate sea urchins, which in turn predate algae such as kelp. When sea otters disappeared, urchin abundance increased drastically leading to the destruction of the kelp forests. The destruction of kelp forests, in turn, produced cascading effects on other species dependant on the kelp resulting in a species-poor ecological community. Later, conservation measures designed to protect sea otters were implemented, which resulted in a recovery of the original ecological community composition [54]. Hence, as seen from a topological perspective, this example shows how a species (the sea otter), in occupying a structurally influential position in a food web (here, the Pacific coastal marine food web), become a keystone species.

Related to the discussion on structural keystone species is the insight that not only direct links, but also indirect links, are important in food webs because, in combination, they determine a species contribution in maintaining the overall structure of the food web. This was e. g., clearly demonstrated by the sea otter example presented above. The influence of indirect links is caused by the fact that these connect otherwise disconnected species. For example, a species that in itself possesses only very few links, but has network neighbors that possess many links, can still exert a very large impact if it is the only species that, indirectly, links its neighbor's neighboring species. This example illustrates the relevance of betweenness centrality [19], described earlier, in studying aspects of species' structural influence in food web. A similar example of a centrality measure is the concept of structural holes in social network analysis [55], where a distinction is made based on whether a particular link is redundant or not. A redundant link (or path) has an equivalent counterpart in the sense that there are other links that also connect its corresponding pair of nodes, whereas removing a non-redundant link would disconnect the pair of nodes. Species possessing non-redundant links clearly exert a negative influence on a food webs ability to withstand species loss or

extinctions since, if removed, they would disconnect consumer species from basal and fundamental energy sources. Non-redundant links can be particular important when they connect otherwise disconnected groups of species.

Finally, there are numerous other measures of centrality available; each designed to assess a particular characteristic of a node's structural importance in a network. Research exploring the applicability of such, and other, topological centrality measures in advancing the concept of keystone species in food webs has just started (see e. g., [56,57,58]). In all, species abundance, link distributions, network position, and the strength of the links, clearly play a role in determining the impact a certain species exerts on the ecosystem. Thus, the relationships among network positions, keystone species, and highly interactive species (i. e., species with many connections) are not entirely straightforward but characterized by complex dependencies and dynamics. Some of these will be discussed below.

Dynamics of Secondary Extinctions

The topological approach to analysis of food web stability, with respect to their vulnerability to secondary extinctions following species removals, clearly lacks a dynamic part. I. e., the analysis is based solely on the topological changes of the food web following species removals. Hence, secondary extinctions are assessed only on the assumption that species that become isolated from all their prey species go extinct, otherwise they remain. Although this assumption is reasonable in many cases, it does not take into account indirect dynamical effects resulting from species loss (see e. g., [59]). For example, two competing prey species may coexist due to predator mediation. If the predator goes extinct, one of the prey species may then out-compete the other. This leads to the counter-intuitive result that a prey species may in fact depend on its predator for its prevalence. Such effects are very difficult to assess in purely topological analyses, as are effects of varying link strengths (see e. g., [60]). Also, the time it takes for a ghost species (a species doomed to become extinct, see [61]) to go extinct may be important. If, for example, the time it takes for a certain species to go extinct is significantly longer than the rate upon which new individuals are moving into the area, the extinction might never take place [62].

However, the strength in using topological analyses lies in its relative simplicity. A dynamical perspective typically involves solving N coupled differential equations (N-species Lotka-Volterra models) numerically in studying food webs with N species. This requires access to large computer processing capacity even for rather small food webs. Secondly, in order to correctly parameterize the set of differential equations, detailed empirical data on the strength of all species interactions are needed.

Plant/Pollinator (Mutualistic) Networks

Until now, focus has been on patterns of predation among species. In nature, species relations are, however, often mutualistic. In such cases both species benefit from their interactions. A typical example of mutualism is given by the relations between plants and pollinators (e. g., insects, birds, mammals etc.). Such mutualism may be just as ecologically important as trophic interactions [63]. As for food webs, the set of mutualistic interactions among a given set of species, such as plants and pollinators, can favorably be described as an ecological network. Thus, the topological characteristics of mutualistic networks have spurred interest among scholars. Jordano and colleagues [64] have e. g., studied a large set of mutualistic network of plants and the animals that pollinate them or disperse their seeds. They found that the degree distribution follows a truncated power law. That is, up to a certain degree, the relation between the number of links per species versus the number of species follows a power law (i. e., the scale free characteristic as discussed previously). Beyond that critical degree, the number of species decline much faster. Thus, these networks are, as for many food webs, vulnerable to a targeted removal of highly connected species, but not as vulnerable as other kinds of networks that strictly follows the power law.

Furthermore, Bascompte and colleagues [65] have studied 27 plant-frugivore networks and 25 plant-pollinator networks, and they found that most of these network consisted of a highly connected core and asymmetrically connected peripherals. This topological characteristic was denoted *nestedness*, which is similar to the concept of core-peripheral structures in social network analysis (see e. g., [66]). The core consists of generalist species where, ideally, every species is connected to every other generalist species; whereas the peripherals are specialized species that, ideally, only link to generalist species. This topological structure clearly departs from the compartmentalization that is often found in food webs as discussed previously.

Network Perspective of Fragmented Landscapes

In the early days of theoretical ecology, space was largely omitted from the analysis. The only dimension was time, and two-species interaction described by continuous time equations made up the bulk of the work ([32], which present a comprehensive review of the development in

theoretical ecology up to present time). As time went on, observations from the real world began to challenge theory development; it seemed something was missing in the equations. It was later shown that this missing ingredient was space. For example, Alan Turing showed, in the early 1950s, that an initially homogenous system could emerge into a heterogeneous and ordered spatial structure – a finding that influenced e. g., theoretical ecology (see also [32,67]). Following the introduction of space into ecology, species movements were hereafter acknowledged as being of crucial importance.

In the late 60s, Levin [68] introduced the concept of metapopulation. A metapopulation is a set of spatially separated local populations which may, individually, undergo local extinctions, but where the aggregated population is maintained by re-colonization events resulting from dispersing organisms. Thus, even though a local population (confined to a spatially distinct patch in the landscape) may go extinct, dispersing organisms from other local populations can re-colonize the empty patch. The fairly simplistic model of metapopulation dynamics introduced by Levin was later extended by Lande [69] who incorporated more details. He also applied his metapopulation models to study the dynamics of the northern spotted owl in northern California. By assessing the fraction of occupied patches, and the fraction of suitable patches, he could define an extinction threshold. If the amount of suitable patches drops below a certain threshold, the whole metapopulation would go extinct.

Levin's and Lande's models are spatially implicit, i. e., they take space into account, but they do not explicitly incorporate geographic location in the models. They assume global mixing, i. e., a mean field approximation of dispersals where all patches are equally likely to receive dispersing organisms. Thus, these models are not able to model the effect of habitat fragmentation, i. e., when habitat patches gets more and more separated, or even disconnected, from each other (from the perspective of a dispersing organism). To enable that kind of analysis, spatially explicit metapopulations models are needed (see e. g., [70,71,72]).

Hanski and Ovaskainen [73] used a matrix notation to describe the flux of dispersing organisms between individual patches at the landscape level in their spatially explicit metapopulation model. They then defined the leading eigenvalue of the matrix as the metapopulation capacity, and that single parameter summarized how the spatial structure of the landscape influenced the metapopulation's dynamics. Of particular interest here is the matrix representation. This matrix can be interpreted as a graph where each matrix element represents the link strength between each pair of nodes (patches). Thus, it describes a network of interconnected patches. Keitt and colleagues [6] explicitly applied the graph-theoretical perspective in studying the level of connectivity of fragmented patches of ponderosa pine and mixed-conifer forest at the landscape level. They modeled how these patches might be experienced in terms of connectivity, or isolation, by dispersing juvenile Mexican spotted owls. Connectivity is, in this context, defined as the degree to which the scattered habitat patches in the landscape facilitate the dispersals of organisms [74].

The landscape network modeled by Keitt and colleagues consisted of nodes (representing individually and spatially distinct habitat patches in the landscape) and links (representing the possibility for dispersal between individual patches, see Fig. 3). Hence, such a network represents the landscape's spatial structure of connectivity (from an organism's point of view). It encapsulates the potential of an organism to traverse the landscape by moving from patch to patch [11]. Using this network perspective on fragmented landscapes, topological analysis of interacting local species populations is made possible. The network approach merges recruitment and dispersal processes with spatial patterns of habitat patches, thus enabling process-oriented analyses of landscape connectivity [75]. In comparison with food web analysis, where focus was on different interacting species, the network approach is here applied in studying how patterns of dispersals between separated habitat patches in the landscape may affect a target species metapopulation dynamics.

Modeling a Landscape as a Network

The starting point in studying topology of fragmented landscapes is to construct the network representation of the patches and their dispersal linkages. The first step in this process is to define what type of land is habitat, and what is not. In it simplest form, a binary image representing the landscape is then created, where all kinds of land cover assessed as being hospitable by the target species is marked, and the rest left unmarked and thus considered as inhospitable (called the landscape matrix, not to be confused with the matrix defined by Hanski and Ovaskainen [73]). The input data in constructing such binary image could be a satellite images or aerial photographs of the landscape under study. This binary simplification may, however, be too simplistic in analyzing heterogeneous landscapes (as will be discussed below), but there are available methods that can incorporate this heterogeneity into the model. Using the binary landscape image, a patch is defined as a coherent area of hospitable land

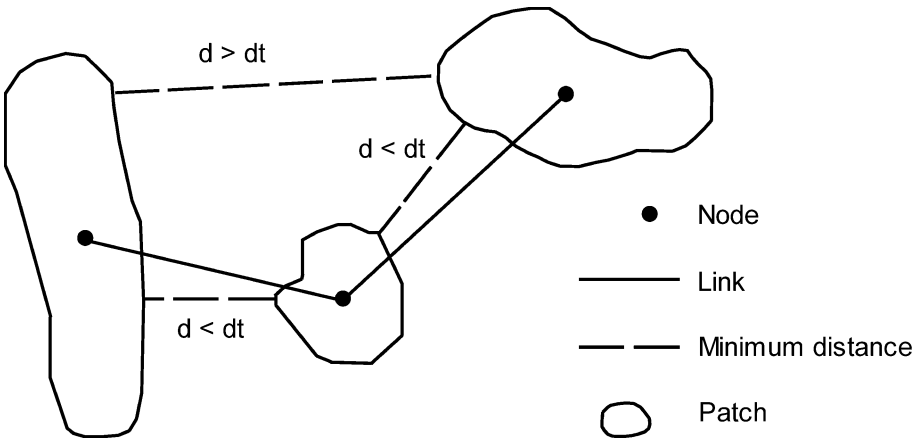

Ecological Topology and Networks, Figure 3
Construction of a landscape graph (*d* = distance, *dt* = threshold distance). This example landscape contains three separate habitat patches and a single component. The three patches belong to a single component because there exists a path along the links (*solid lines*) that connects all three patches. If either one of the links shown were removed, there would be two components. If all links were removed, then there would be three components, each consisting of a single patch (Adapted from: [6])

(in practice, all adjacent pixels that are marked as hospitable make up a single patch). From a network perspective, a patch is from now on seen as a node.

Emanating from the binary image, the dispersal links of the patches are then assessed. The simplest way to assess the links is to define a threshold distance [6]. Two habitat patches that are separated by a distance less than the threshold distance are considered connected, otherwise they are considered mutually disconnected and thus no link exists between these. The threshold distance should be defined based on the target species dispersal ability, i. e., how far the species is able to move in the inhospitable matrix. The threshold distance for a species depends on a number of different factors such as e. g., body mass, movement strategy, visual ability, and the type of land between the habitat patches.

Refinement of the dichotomous distance-based link assessment procedure is often preferred. Instead of the binary link/no link assessment; the link strength could be weighted proportionally, taking more factors than distance into account. The area of the patches could e. g., be factored in using a negative dispersal kernel (cf. [71], see also [76]). Thus, two large patches could be considered as connected, while another pair of smaller patches (although separated by the same distance) could be considered as disconnected (Eq. 1). Furthermore, directionality of dispersals could be accounted for. For example, consider a pair of patches where one is large and the other is small. Here, it might be reasonable to assume that organisms' movement towards the large patch is more probable than the opposite. Although link strengths are taken into consideration, it would simplify the analysis if the weakest links are disregarded (by introducing a cut-off value C, see Eq. (1)).

$$\begin{aligned} S_{ij} &= b\sqrt{A_j}e^{\alpha D_{ij}} \\ S_{ij} &= 0 \quad (\text{if } S_{ij} \text{ was less than } C) \end{aligned} \tag{1}$$

(see further details in [76])

S_{ij} = Metric proportional to the dispersal flux rate from patch i to patch j
A_j = Area of patch j (m^2)
D_{ij} = Distance between patch i and patch j(m)
α = Constant ($1/m$)
C = Cut-off value
b = Constant ($1/m$)

Until now, a binary habitat/matrix landscape has been assumed. This approximation may be too simplistic in heterogeneous landscapes where the permeability of different types of land covers differs considerably. If the matrix consists of an array of different land cover types, the ability to move between any arbitrary pair of patches depends on the exact composition of land types in-between, and not only on the geographic distance separating these patches (e. g., [77]). Fortunately, there are methods available that take the differing permeability of a heterogeneous matrix into account when calculating the least cost paths (or effective distance) of moving between any two arbitrary locations in the landscape [78,79,80,81]. Thus, more sophisticated link-assessment methods than just applying a threshold distance could favorably be used in analyzing heterogeneous landscapes. The key point is, however, that after all links have been assessed, the resulting network

representation will incorporate the topological characteristics of the fragmented landscape, and further analyses of the network do not depend on the chosen link-assessment method.

Assessing Fragmentation and Modularity

In spite of being fragmented, a landscape may very well be perceived as connected by species being able to move between patches of habitats [74]. In studying landscape connectivity, the overall research question is therefore to what extent the landscape supports species movement, i. e., how well connected are the scattered habitat patches as perceived by an organism on the move? Of central importance is the level of traversability, i. e., how far can dispersing organisms reach by moving from patch to patch in the landscape?

In terms of network terminology, traversability can be expressed as if the network itself is divided into separate subnets, or whether all nodes are confined within one single coherent network [6]. A subset of nodes where a path between all member nodes exists is called a network *component*. Thus, by definition, no paths exist between nodes of different components. Therefore, if a network is made up of two, or more, components, it is not possible to traverse the whole network by moving from node to node. Ecologically, an organism located in an arbitrary habitat patch can, directly or indirectly, move to any other patch within the same component, but it can never disperse to a patch belonging to another component (the landscape is compartmentalized). Thus, a component can host a set of connected but localized populations – a metapopulation – but the metapopulation itself would be isolated from other metapopulations confined to other components [11,75].

The number of components in a network representing the landscape's spatial structure of connectivity for a dispersing species would, therefore, directly relate to the actual level of fragmentation (or connectivity) in the landscape. A large number of components correspond to a heavily fragmented landscape, whereas a network with just one component implies a well-connected landscape where a large and persistent metapopulation may thrive. A second metric of connectivity is the percentage of patches confined to the largest component in the network. If the fraction of patches that belong to the largest component is high, the fraction of the landscape that is within reach for a single metapopulation would be high.

In recalling the concept of threshold distance, a question that arises is how the number of components relates to the threshold distance. Obviously, a very large threshold distance would imply that all patches would be reachable from each other, whereas a very small threshold distance would imply that virtually no patch is connected to any other patch (see Fig. 4). Thus, by measuring the number of components as a function of the threshold distance (edge thinning, see [78]), one can assess at what

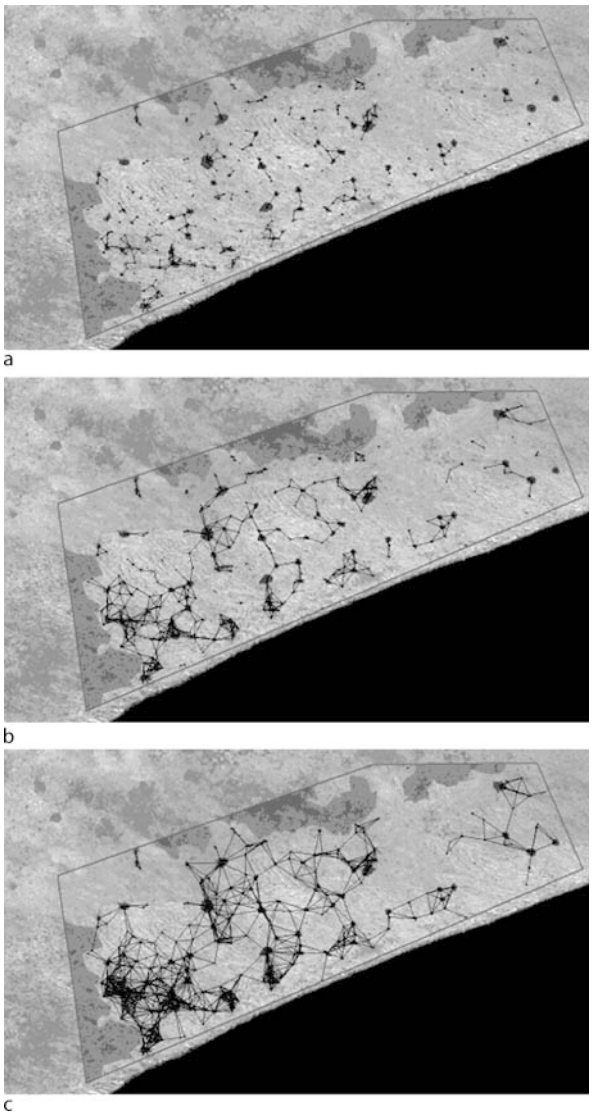

Ecological Topology and Networks, Figure 4
The figures show an agricultural landscape in southern Madagascar with three different overlaid networks representing the connectivity among scattered forest patches assuming dispersal abilities set to a 500 m, b 1000 m, and c 1500 m. The forest patches are identifiable by the distinct dark spots, situated within a matrix consisting of cultivated land (*light grey*). Patches range in size from <1–95 ha and are fairly evenly distributed in the landscape. In the western and northern part of the studied area, the *shaded/darker grey* zones indicate areas classified as potential source areas (Source: [76])

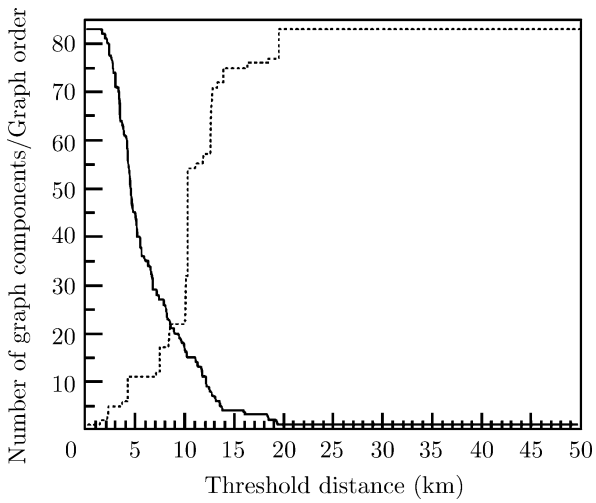

Ecological Topology and Networks, Figure 5
Number of components (*solid line*) and size of largest component (*dotted line*) as a function of threshold distance (Source: [78])

threshold distance the landscape starts to be perceived as connected (see Fig. 5). The transition interval, i. e., where the number of components rapidly decreases as a function of an increased threshold distance, has been suggested as a scalar measure of the landscape's structural scale of connectivity [82]. Species with dispersal capabilities below this transition interval would experience the landscape as severely fragmented, whereas species with dispersal capabilities above the interval would experience the landscape as connected.

Measuring the number of component seems to be a useful method in assessing spatial compartmentalization of the landscape. In some circumstances, it may however be too crude as it lacks the ability to estimate a more continual degree of landscape compartmentalization (i. e., where compartments are not necessarily isolated from each other, but merely separated to a certain degree) [11]. These types of loosely detached compartments of internally well-connected habitat patches can affect age and sex structures of populations in the landscape [83], and it could potentially contribute to genetic differentiations even at small temporal and spatial scales (cf. [84,85]). The inability to assess such structures is a consequence of the binary perspective of the component-based analysis; either two patches are being considered as connected, or else they are considered as being completely isolated. But what if a set of patches are very well connected internally, but rather weakly connected to the outside landscape? Using component-based analysis, such set of patches would not be identifiable unless they are completely isolated from the remaining landscape.

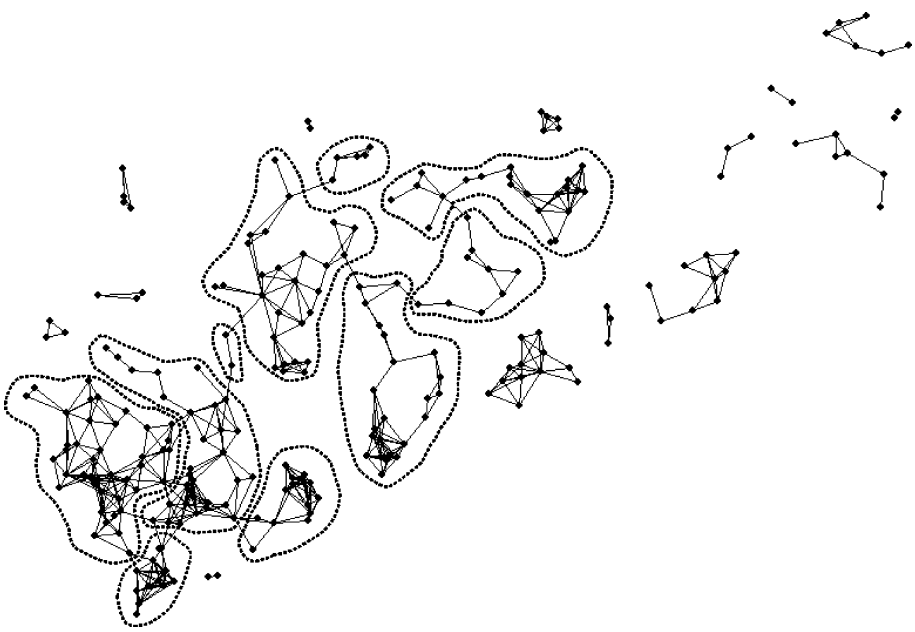

Ecological Topology and Networks, Figure 6
Decomposition of the largest component in the landscape (from Fig. 4b), using the Community Structure method [45], taken at the hierarchical level with the best "fit" ("Modularity" equal to 0.84, see [18]). Note that the largest component is decomposed into ten smaller compartments

Bodin and Norberg [11] addressed this possible limitation of the component-based analysis by using a divisive hierarchical decomposition algorithm called *Community Structure* recently proposed by Girvan and Newman [45]. This method fulfilled three potentially important topological criteria: (1) groups of coherent sets of patches could be assessed even though not necessarily being completely isolated from each other, (2) the algorithm is hierarchical and would therefore create non-overlapping groups of patches which means that it would help to divide the landscape into spatially separated compartments, and (3) it did not discriminate against peripheral patches (e. g., patches with just one link would be assigned the same group membership as their neighbor instead of being treated as isolated). Figure 6 illustrates how this method can be applied to decompose a single large component into a set of spatially distinguishable groups of patches.

Pascual–Hortal and Saura [86] have systematically studied how a set of ten different network-based measures, including the size of the largest component, respond to a number of different spatially explicit fragmentation scenarios. They suggest a new measure called *Integral index of connectivity* (IIC) which extends the previous size-of-largest-components analysis by also incorporating the area of the patches, as well as the actual topological distances between patches, into a single normalized measure of connectivity. They showed that the only measure that responded in a desirable and consistent way to the full set of fragmentation scenarios was the proposed IIS. However, the size of the largest component did also respond reasonably well.

Assessing Critical Habitat Patches

The concept of stepping stones is similar to the topological importance assigned to certain keystone species in food webs. A stepping stone patch can be seen as a bridge connecting otherwise separated groups of internally interconnected patches. A particular patch may also, although not being the very last remaining bridging patch in a otherwise largely fragmented landscape, exert influence since it could make it much more difficult for dispersing species to reach over the whole landscape if it is removed (i. e., its presence in the landscape provide shortcuts, thus reducing the overall topological distance between patches, see also [16]). Accordingly, assessing the impact individual patches exert on the overall landscape connectivity would be of uttermost importance in management targeted to preserve functional and well-connected landscapes (e. g., [11,75,86]).

A reasonable approach to assess individual patches' contribution to the overall connectivity of the landscape is to remove them, one by one, from the modeled network and then measure the effect on some relevant measures of landscape connectivity [75,78,86]. If, for example, the issue of interest is the topological traversability of the landscape, it is relevant to measure the size, or diameter, of the largest component following the removal of each and every patch, and then rank the patches according to the induced change their removal inflicted on the largest component [75,78]. Similarly, in the previously mentioned study by Pascual–Hortal and Saura [86], the effect on the chosen set of 10 different connectivity measures were assessed for a number of spatially explicit fragmentation scenarios wherein individual patches occupying different kinds of topological positions were removed.

Large changes in the size of the largest component following the removal of a certain patch would, however, only occur if the removal would split the largest component into two or more non-trivial components. From a management perspective, it would be desirable if such patches were identified before they become that critical (as a result of continuous fragmentation). If such early identification was possible, it would enable managers to more proactively target conservation efforts in order to protect patches that could become critical. The trial-and-error method of removing nodes, one-by-one, lacks this ability.

To enable such proactive topological analysis of individual patches' contribution to the landscapes connectivity, various measures of network centrality may be plausible. Following this line of though, the betweenness centrality index [19] has been suggested as a potentially relevant measure in the context of landscape fragmentation [11,87]. Patches with a high score of betweenness centrality are e. g., expected to significantly contribute to: (1) Reducing the overall topological distances between all pair of patches, and (2) Linking otherwise separated groups of patches [11]. As opposed to the trial-and-error approach described earlier, it appears that by using the betweenness centrality as the criterion for ranking the topological importance of individual patches, potentially critical patches could be identified even in cases where their removal would not split the network into isolated components. Furthermore, even in cases where the overall connectivity of the landscape is fairly high, the betweenness centrality index could be of use in identifying patches that may be critical in the future as a consequence of further fragmentation ([11], but see also [87]).

Another approach to identifying critical habitat patches is to study the minimal spanning tree (MST) of the

network representation of the landscape. The MST would visually represent the "backbone" of the connected patches in the landscape, and from there important patches can be identified [75]. Furthermore, if the strength of the links are accounted for (using e. g., Eq. 1), a link-weighted MST can be constructed where the specific patches which contribute, more than others, to species dispersals stand out by their high number of links [78]. Furthermore, Fall and colleagues [81] have developed a modeling approach, which they call spatial graphs, where they explicitly incorporate geographical details in a network representation of a fragmented landscape. A spatial graph is e. g., spatially explicit in describing the geometrics of the least-cost paths a dispersing organism would follow when moving from patch to patch. Using this representation of the landscape, they suggest that critical passages, barriers and stepping stone patches can be visually identified and targeted for management and/or restoration efforts. This modeling approach has also been tested using empirical data on the movement of woodland caribou (*Rangifer tarandus caribou*), where a strong relationship between the distribution of caribou and larger clusters of high-quality habitat (identified using spatial graphs) was shown [88].

Summary

Three topological characteristics, degree distribution, modularity and centrality, all appear to be of relevance in studying different aspects of ecological systems. In food webs, the degree distribution affects the stability in terms of the risk for secondary extinctions following species loss. In addition, the density of links is believed to have effects on the stability of ecosystems, although how, and in what direction, is not entirely clear. Many food webs have a modular structure, i. e., the web is divided into several groups that are only weakly connected to each other. This may effect how far disturbances spread throughout the food web. In a web with many dense but separated groups of species, disturbances may very well be confined within groups. On the other hand, a highly modular structure implies there are fewer opportunities for species to compensate if some of their prey species would decline.

Furthermore, some species may be more influential than others, and that influence may be attributed to their structural position in the food web (structural keystone species). Influence may result from having many links to others, but it could also derive from the possession of structurally important links that for example connects otherwise disconnected groups of species.

In order to analyze the spatial structure of connectivity of fragmented landscapes, a network representation where nodes are patches and links are dispersal possibilities paves the way for topological analyses at the level of landscapes. It appears that simple network characteristics, such as the number of components, can help in assessing how connected different species, with varying dispersal capabilities, actually experience the landscape. Furthermore, network analysis targeted at finding compartments of internally well-connected habitat patches can help in identifying population and metapopulation boundaries. Finally, by assessing the structural importance of individual habitat patches (similar to structural keystone species), land management could be made more efficient by letting different network centrality measures provide guidance in prioritizing conservation efforts.

Future Directions

Topological analyses alone can provide important insights into complex ecological systems. However, in many cases, it is desirable to also include the strengths of the links in the analyses. Many network analytical approaches support the inclusion of link strengths, but there are many others that do not. Furthermore, the network perspective has traditionally been rather static, and the analytical focus has been the topological structure currently at hand. But networks evolve and develop dynamically; thus the topology given at one moment in time may be outdated later on. Better understanding of processes that shape network evolution are needed, as are methods capable of longitudinal network analyses applicable in an ecological context. The last statement touches upon the fact that a topological perspective will only be of any significant value when we possess knowledge and/or well-grounded assumptions of important processes. Thus topological approaches need to develop alongside theoretical advances of understanding key ecological processes.

Bibliography

Primary Literature

1. Volterra V (1926) Fluctuations in the abundance of a species considered mathematically. Nature 118:558
2. Lotka AJ (1925) Elements of physical biology. Williams and Wilkins, Baltimore
3. Odum EP, Odum HT (1953) Fundamentals of ecology. Saunders, Philadelphia
4. Odum EP (1977) The emergence of ecology as a new integrative discipline. Science 195:1289
5. Paine RT (1980) Food Webs: Linkage, interaction strength and community infrastructure. J Anim Ecol 49:666
6. Keitt TH, Urban DL, Milne BT (1997) Detecting critical scales in fragmented landscapes. Conserv Ecol 1

7. Webb C, Bodin Ö (2008) A network perspective on modularity and control of flow in robust systems. In: Norberg J, Cumming G (eds) Complexity theory for a sustainable future. Columbia, New York Chichester, pp 85–118
8. Levin SA (1998) Ecosystems and the biosphere as complex adaptive systems. Ecosystems 1:431
9. Freeman LC (2004) The development of social network analysis – A study in the sociology of science. Empirical Press, Vancouver
10. Luczkovich JJ, Borgatti SP, Johnson JC, Everett MG (2003) Defining and measuring trophic role similarity in food webs using regular equivalence. J Theor Biol 220:303
11. Bodin Ö, Norberg J (2007) A network approach for analyzing spatially structured populations in fragmented landscape. Landsc Ecol 22:31
12. Benedek Z, Jordán F, Báldi A (2007) Topological keystone species complexes in ecological interaction networks. Community Ecol 8:1
13. Barabási AL (2002) Linked: The new science of networks. Perseus, Cambridge
14. Albert R, Jeong H, Barabasi AL (2000) Error and attack tolerance of complex networks. Nature 406:378
15. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440
16. Borgatti SP, Everett MG, Shirey PR (1990) LS sets, lambda sets and other cohesive subsets. Soc Netw 12:337
17. Wasserman S, Faust K (1994) Social network analysis – methods and applications. Cambride University Press, Cambridge
18. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69:026113
19. Freeman L (1979) Centrality in social networks. Conceptual clarifications. Soc Netw 1:215
20. Sabidussi G (1966) The centrality index of a graph. Psychometrika 31:581
21. Woodward G et al (2005) Body size in ecological networks. Trends Ecol Evol 20:402
22. Berlow EL et al (2004) Interaction strengths in food webs: issues and opportunities. J Anim Ecol 73:585–598
23. Krause AE, Frank KA, Mason DM, Ulanowicz RE, Taylor WW (2003) Compartments revealed in food-web structure. Nature 426:282
24. Scotti M, Podani J, Jordán F (2007) Weighting, scale dependence and indirect effects in ecological networks: A comparative study. Ecol Complex 4:148–159
25. Dunne JA (2006) The network structure of food webs. In: Pascual M, Dunne JA (eds) Ecological networks: Linking structure to dynamics in food webs. Oxford University Press, Oxford, pp 27–86
26. Wootton JT (1997) Estimates and tests of per capita interaction strength: diet, abundance, and impact of intertidally foraging birds. Ecol Monogr 67:45
27. Paine RT (1992) Food-web analysis through field measurement of per capita interaction strength. Nature 355:73
28. Goldwasser L, Roughgarden J (1993) Construction and analysis of a large caribbean food web. Ecology 74:1216
29. Winemiller KO (1990) Spatial and temporal variation in tropical fish trophic networks. Ecol Monogr 60:331–367
30. Warren PH (1990) Variation in food-web structure—the determinants of connectance. Am Nat 136:689–700
31. de Ruiter PC, Neutel AM, Moore JC (1995) Energetics, patterns of interaction strengths, and stability in real ecosystems. Science 269:1257
32. Solé RV, Bascompte J (2006) Self-organization in complex ecosystems. Princeton University Press, New Jersey
33. Dunne JA, Williams RJ, Martinez ND (2002) Network structure and biodiversity loss in food webs: robustness increases with connectance. Ecol Lett 5:558–567
34. Seidman SB (1983) LS sets and cohesive subsets of graphs and hypergraphs. Soc Netw 5:92
35. Melián CJ, Bascompte J (2004) Food web cohesion. Ecology 85:352
36. Pastor-Satorras R, Vázquez A, Vespignani A (2001) Dynamical and correlation properties of the internet. Phys Rev Lett 87:258701
37. Melián CJ, Bascompte J (2002) Complex networks: Two ways to be robust? Ecol Lett 5:705
38. Garlaschelli D, Caldarelli G, Pietronero L (2003) Universal scaling relations in food webs. Nature 423:165–168
39. Dunne JA, Williams RJ, Martinez ND (2002) Food-web structure and network theory: The role of connectance and size. Proc Natl Acad Sci USA 99:12917–12922
40. Camacho J, Guimerá R, Amaral LAN (2002) Robust patterns in food web structure. Phys Rev Lett 88:228102
41. Williams RJ, Berlow EL, Dunne JA, Barabási AL, Martinez ND (2002) Two degrees of separation in complex food webs. Proc Natl Acad Sci USA 99:12913–12916
42. Pimm SL, Lawton JH (1980) Are food webs divided into compartments? J Anim Ecol 49:879
43. Webb C, Bodin Levin SA (2005) Cross-system perspectives on the ecology and evolution of resilience. In: Jen E (ed) Robust design: A repertoire of biological, ecological, and engineering case studies. SFI Lecture Note Series. Oxford University Press, Oxford, pp 151–172
44. Amaral LAN, Scala A, Barthélémy M, Stanley HE (2000) Classes of small-world networks. Proc Natl Acad Sci USA 97:11149–11152
45. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA 99:7821
46. Frank KA (1995) Identifying Cohesive Subgroups. Soc Netw 17:27
47. McCann K, Hastings A, Huxel GR (1998) Weak trophic interactions and the balance of nature. Nature 395:794
48. May RM (1974) Biological populations with nonoverlapping generations: stable points, stable cycles, and chaos. Science 186:645
49. Power ME et al (1996) Challenges in the quest for keystones. Bioscience 46:609
50. Solé RV, Montoya JM (2001) Complexity and fragility in ecological networks. Proc Biol Sci 268:2039
51. Steneck RS et al (2002) Kelp forest ecosystem: biodiversity, stability, resilience and their future. Environ Conserv 29:436
52. Estes JA, Palmisano JF (1974) Sea otters: their role in structuring nearshore communities. Science 185:1058
53. Levin SA (2000) Fragile Dominion: Complexity and the Commons. Perseus Publishing, Cambridge
54. Estes JA, Duggins DO (1995) Sea otters and kelp forests in Alaska-generality and variation in a community ecological paradigm. Ecol Monogr 65:75
55. Burt R (1992) Structural holes: The social structure of competiton. Harvard University Press, Cambridge

56. Jordán F, Liu W-C, Davis AJ (2006) Topological keystone species: measures of positional importance in food webs. Oikos 112:535
57. Jordán F, Benedek Z, Podani J (2007) Quantifying positional importance in food webs: A comparison of centrality indices. Ecol Model 205:270–275
58. Estrada E (2007) Characterization of topological keystone species – Local, global and "meso-scale" centralities in food webs. Ecol Complex 4:48
59. Yodzis P (1988) The indeterminacy of ecological interactions, as perceived through perturbation experiments. Ecology 69:508
60. Eklöf A, Ebenman B (2006) Species loss and secondary extinctions in simple and complex model communities. J Anim Ecol 75:239
61. Tilman D, May RM, Lehman CL, Nowak MA (1994) Habitat destruction and the extinction debt. Nature 371:65
62. Borrvall C, Ebenman B (2006) Early onset of secondary extinctions in ecological communities following the loss of top predators. Ecol Lett 9:435–442
63. Vasas V, Jordán F (2006) Topological keystone species in ecological interaction networks: Considering link quality and nontrophic effects. Ecol Model 196:365–378
64. Jordano P, Bascompte J, Olesen JM (2003) Invariant properties in coevolutionary networks of plant-animal interactions. Ecol Lett 6:69
65. Bascompte J, Jordano P, Melián CJ, Olesen JM (2003) The nested assembly of plant–animal mutualistic networks. Proc Natl Acad Sci USA 100:9383
66. Borgatti SP, Everett MG (1999) Models of core/periphery structures. Soc Netw 21:375
67. Turing A (1952) On the chemical basis of morphogenesis. Philos Trans R Soc Lond B Biol Sci 237:37
68. Levin R (1969) Some demographic and genetic consequences of environmental heterogeneity for biological. Control Bull Entomol Soc Am 15:237
69. Lande R (1987) Extinction thresholds in demographic models of territorial populations. Am Nat 130:624
70. Bascompte J, Solé RV (1996) Habitat fragmentation and extinction thresholds in spatially explicit metapopulation models. J Anim Ecol 65:465
71. Hanski I (1994) A practical model of metapopulation dynamics. J Anim Ecol 63:151
72. Hanski I (2001) Spatially realistic theory of metapopulation ecology. Naturwissenschaften 88:372
73. Hanski I, Ovaskainen O (2000) The metapopulation capacity of a fragmented landscape. Nature 404:755–758
74. Taylor PD, Fahrig L, Henein K, Merriam G (1993) Connectivity is a vital element of landscape structure. Oikos 68:571–573
75. Urban D, Keitt T (2001) Landscape connectivity: A graph-theoretic perspective. Ecology 82:1205
76. Bodin Ö, Tengö M, Norman A, Lundberg J, Elmqvist T (2006) The value of small size: Loss of forest patches and ecological thresholds in southern Madagascar. Ecol Appl 16:440
77. Baum KA, Haynes KJ, Dillemuth FP, Cronin JT (2004) The matrix enhances the effectiveness of corridors and stepping stones. Ecology 85:2671
78. Bunn AG, Urban DL, Keitt TH (2000) Landscape connectivity: A conservation application of graph theory. J Environ Manage 59:265
79. Ricketts TH (2001) The matrix matters: Effective isolation in fragmented landscapes. Am Nat 158:87
80. Verbeylen G, Bruyn LD, Adriaensen F, Matthysen E (2003) Does matrix resistance influence Red squirrel (Sciurus vulgaris L. 1758) distribution in an urban landscape? Landscape Ecol 18:791
81. Fall A, Fortin MJ, Manseau M, O'Brien D (2006) Spatial graphs: Principles and applications for habitat connectivity. Ecosystem 10:448–461
82. Brooks CP (2003) A scalar analysis of landscape connectivity. Oikos 102:433
83. Sutherland GD, Harestad AS, Price K, Lertzman KP (2000) Scaling of natal dispersal distances in terrestrial birds and mammals. Conserv Ecol 4
84. Garant D, Kruuk LEB, Wilkin TA, McCleery RH, Sheldon BC (2005) Evolution driven by differential dispersal within a wild bird population. Nature 433:60
85. Postma E, Van Noordwijk AJ (2005) Gene flow maintains a large genetic difference in clutch size at a small spatial scale. Nature 433:65
86. Pascual-Hortal L, Saura S (2006) Comparison and development of new graph-based landscape connectivity indices: towards the priorization of habitat patches and corridors for conservation. Landscape Ecol 21:959
87. Minor ES, Urban DL (2007) Graph theory as a proxy for spatially explicit population models in conservation planning. Ecol Appl 17:1771
88. O'Brien D, Manseau M, Fall A, Fortin MJ (2006) Testing the importance of spatial configuration of winter habitat for woodland caribou: An application of graph theory. Biol Conserv 130:70–83

Books and Reviews

Buchanan M (2002) Small world: Uncovering nature's hidden networks. Weidenfeld & Nicolson, London
Levin SA (2000) Fragile dominion: Complexity and the commons. Perseus Publishing, Cambridge
Pascual M, Dunne JA (eds) (2006) Ecological networks: Linking structure to dynamics in food webs. Oxford University Press, Oxord
Pimm SL (1982) Food webs. Chapman and Hall, London

Econometrics: Models of Regime Changes

Jeremy Piger[1]
Department of Economics,
University of Oregon, Eugene, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Threshold and Markov-Switching Models of Regime Change

Estimation of a Basic Markov-Switching Model
Extensions of the Basic Markov-Switching Model
Specification Testing for Markov-Switching Models
Empirical Example: Identifying Business Cycle Turning Points
Future Directions
Bibliography

Glossary

Filtered probability of a regime The probability that the unobserved Markov chain for a Markov-switching model is in a particular regime in period t, conditional on observing sample information up to period t.

Gibbs sampler An algorithm to generate a sequence of samples from the joint probability distribution of a group of random variables by repeatedly sampling from the full set of conditional distributions for the random variables.

Markov chain A process that consists of a finite number of states, or regimes, where the probability of moving to a future state conditional on the present state is independent of past states.

Markov-switching model A regime-switching model in which the shifts between regimes evolve according to an unobserved Markov chain.

Regime-Switching Model A parametric model of a time series in which parameters are allowed to take on different values in each of some fixed number of regimes.

Smooth transition threshold model A threshold model in which the effect of a regime shift on model parameters is phased in gradually, rather than occurring abruptly.

Smoothed probability of a regime The probability that the unobserved Markov chain for a Markov-switching model is in a particular regime in period t, conditional on observing all sample information.

Threshold model A regime-switching model in which the shifts between regimes are triggered by the level of an observed economic variable in relation to an unobserved threshold.

Time-varying transition probability A transition probability for a Markov chain that is allowed to vary depending on the outcome of observed information.

Transition probability The probability that a Markov chain will move from state j to state i.

[1] I am grateful to Jim Hamilton and Bruce Mizrach for comments on an earlier draft.

Definition of the Subject

Regime-switching models are time-series models in which parameters are allowed to take on different values in each of some fixed number of "regimes." A stochastic process assumed to have generated the regime shifts is included as part of the model, which allows for model-based forecasts that incorporate the possibility of future regime shifts. In certain special situations the regime in operation at any point in time is directly observable. More generally the regime is unobserved, and the researcher must conduct inference about which regime the process was in at past points in time. The primary use of these models in the applied econometrics literature has been to describe changes in the dynamic behavior of macroeconomic and financial time series.

Regime-switching models can be usefully divided into two categories: "threshold" models and "Markov-switching" models. The primary difference between these approaches is in how the evolution of the state process is modeled. Threshold models, introduced by Tong [91], assume that regime shifts are triggered by the level of observed variables in relation to an unobserved threshold. Markov-switching models, introduced to econometrics by [16,39,41], assume that the regime shifts evolve according to a Markov chain.

Regime-switching models have become an enormously popular modeling tool for applied work. Of particular note are regime-switching models of measures of economic output, such as real Gross Domestic Product (GDP), which have been used to model and identify the phases of the business cycle. Examples of such models include [3,7,41,57,60,61,73,75,77,90,93]. A sampling of other applications include modeling regime shifts in inflation and interest rates [2,25,34], high and low volatility regimes in equity returns [23,46,48,92], shifts in the Federal Reserve's policy "rule" [55,83], and time variation in the response of economic output to monetary policy actions [35,53,69,81].

Introduction

There is substantial interest in modeling the dynamic behavior of macroeconomic and financial quantities observed over time. A challenge for this analysis is that these time series likely undergo changes in their behavior over reasonably long sample periods. This change may occur in the form of a "structural break", in which there is a shift in the behavior of the time series due to some permanent change in the economy's structure. Alternatively, the change in behavior might be temporary, as in the case of wars or "pathological" macroeconomic episodes such as

economic depressions, hyperinflations, or financial crises. Finally, such shifts might be both temporary and recurrent, in that the behavior of the time series might cycle between regimes. For example, early students of the business cycle argued that the behavior of economic variables changed dramatically in business cycle expansions vs. recessions.

The potential for shifts in the behavior of economic time series means that constant parameter time series models might be inadequate for describing their evolution. As a result, recent decades have seen extensive interest in econometric models designed to incorporate parameter variation. One approach to describing this variation, denoted a "regime-switching" model in the following, is to allow the parameters of the model to take on different values in each of some fixed number of regimes, where, in general, the regime in operation at any point in time is unobserved by the econometrician. However, the process that determines the arrival of new regimes is assumed known, and is incorporated into the stochastic structure of the model. This allows the econometrician to draw inference about the regime that is in operation at any point in time, as well as form forecasts of which regimes are most likely in the future.

Applications of regime-switching models are usually motivated by economic phenomena that appear to involve cycling between recurrent regimes. For example, regime-switching models have been used to investigate the cycling of the economy between business cycle phases (expansion and recession), "bull" and "bear" markets in equity returns, and high and low volatility regimes in asset prices. However, regime switching models need not be restricted to parameter movement across recurrent regimes. In particular, the regimes might be non-recurrent, in which case the models can capture permanent "structural breaks" in model parameters.

There are a number of formulations of regime-switching time-series models in the recent literature, which can be usefully divided into two broad approaches. The first models regime change as arising from the observed behavior of the level of an economic variable in relation to some threshold value. These "threshold" models were first introduced by Tong [91], and are surveyed by [78]. The second models regime change as arising from the outcome of an unobserved, discrete, random variable, which is assumed to follow a Markov process. These models, commonly referred to as "Markov-switching" models, were introduced in econometrics by [16,39], and became popular for applied work following the seminal contribution of Hamilton [41]. Hamilton and Raj [47] and Hamilton [44] provide surveys of Markov-switching models, while Hamilton [43] and Kim and Nelson [62] provide textbook treatments.

There are by now a number of empirical applications of regime-switching models that establish their empirical relevance over constant parameter alternatives. In particular, a large amount of literature has evaluated the statistical significance of regime-switching autoregressive models of measures of US economic activity. While the early literature did not find strong evidence for simple regime-switching models over the alternative of a constant parameter autoregression for US real GDP (e. g. [33]), later researchers have found stronger evidence using more complicated models of real GDP [57], alternative measures of economic activity [45], and multivariate techniques [63]. Examples of other studies finding statistical evidence in favor of regime-switching models include Garcia and Perron [34], who document regime switching in the conditional mean of an autoregression for the US real interest rate, and Guidolin and Timmermann [40], who find evidence of regime-switching in the conditional mean and volatility of UK equity returns.

This article surveys the literature surrounding regime-switching models, focusing primarily on Markov-switching models. The organization of the article is as follows. Section "Threshold and Markov-Switching Models of Regime Change" describes both threshold and Markov-switching models using a simple example. The article then focuses on Markov-switching models, with Sect. "Estimation of a Basic Markov-Switching Model" discussing estimation techniques for a basic model, Sect. "Extensions of the Basic Markov-Switching Model" surveying a number of primary extensions of the basic model, and Sect. "Specification Testing for Markov-Switching Models" surveying issues related to specification analysis. Section "Empirical Example: Identifying Business Cycle Turning Points" gives an empirical example, discussing how Markov-switching models can be used to identify turning points in the US business cycle. The article concludes by highlighting some particular avenues for future research.

Threshold and Markov-Switching Models of Regime Change

This section describes the threshold and Markov-switching approaches to modeling regime-switching using a specific example. In particular, suppose we are interested in modeling the sample path of a time series, $\{y_t\}_{t=1}^T$, where y_t is a scalar, stationary, random variable. A popular choice is an autoregressive (AR) model of order k:

$$y_t = \alpha + \sum_{j=1}^{k} \phi_j y_{t-j} + \varepsilon_t \,, \tag{1}$$

where the disturbance term, ε_t, is assumed to be normally distributed, so that $\varepsilon_t \sim N(0, \sigma^2)$. The AR($k$) model in (1) is a parsimonious description of the data, and has a long history as a tool for establishing stylized facts about the dynamic behavior of the time series, as well as an impressive record in forecasting.

In many cases however, we might be interested in whether the behavior of the time series changes across different periods of time, or regimes. In particular, we may be interested in the following regime-switching version of (1):

$$y_t = \alpha_{S_t} + \sum_{j=1}^{k} \phi_{j,S_t} y_{t-j} + \varepsilon_t \,, \tag{2}$$

where $\varepsilon_t \sim N(0, \sigma^2_{S_t})$. In (2), the parameters of the AR(k) depend on the value of a discrete-valued state variable, $S_t = i, i = 1, \ldots, N$, which denotes the regime in operation at time t. Put simply, the parameters of the AR(k) model are allowed to vary among one of N different values over the sample period.

There are several items worth emphasizing about the model in (2). First, conditional on being inside of any particular regime, (2) is simply a constant parameter linear regression. Such models, which are commonly referred to as "piecewise linear", make up the vast majority of the applications of regime-switching models. Second, if the state variable were observed, the model in (2) is simply a linear regression model with dummy variables, a fact that will prove important in our discussion of how the parameters of (2) might be estimated. Third, although the specification in (2) allows for all parameters to switch across all regimes, more restrictive models are certainly possible, and indeed are common in applied work. For example, a popular model for time series of asset prices is one in which only the variance of the disturbance term is allowed to vary across regimes. Finally, the shifts in the parameters of (2) are modeled as occurring abruptly. An example of an alternative approach, in which parameter shifts are phased in gradually, can be found in the literature investigating "smooth transition" threshold models. Such models will not be described further here, but are discussed in detail in [93].

Threshold and Markov-switching models differ in the assumptions made about the state variable, S_t. Threshold models assume that S_t is a deterministic function of an observed variable. In most applications this variable is taken to be a particular lagged value of the process itself, in which case regime shifts are said to be "self-exciting". In particular, define $N-1$ "thresholds" as $\tau_1 < \tau_2 < \ldots < \tau_{N-1}$. Then, for a self-exciting threshold model, S_t is defined as follows:

$$\begin{aligned} S_t &= 1 \quad & y_{t-d} &< \tau_1 \,, \\ S_t &= 2 \quad & \tau_1 \le y_{t-d} &< \tau_2 \,, \\ &\vdots \quad & &\vdots \\ S_t &= N \quad & \tau_{N-1} &\le y_{t-d} \,. \end{aligned} \tag{3}$$

In (3), d is known as the "delay" parameter. In most cases S_t is unobserved by the econometrician, because the delay and thresholds, d and τ_i, are generally not observable. However, d and τ_i can be estimated along with other model parameters. [78] surveys classical and Bayesian approaches to estimation of the parameters of threshold models.

Markov-switching models also assume that S_t is unobserved. In contrast to threshold models however, S_t is assumed to follow a particular stochastic process, namely an N-state Markov chain. The evolution of Markov chains are described by their transition probabilities, given by:

$$\begin{aligned} P(S_t = i | S_{t-1} = j, S_{t-2} = q, \ldots) \\ = P(S_t = i | S_{t-1} = j) = p_{ij} \,, \end{aligned} \tag{4}$$

where, conditional on a value of j, we assume $\sum_{i=1}^{N} p_{ij} = 1$. That is, the process in (4) specifies a complete probability distribution for S_t. In the general case, the Markov process allows regimes to be visited in any order and for regimes to be visited more than once. However, restrictions can be placed on the p_{ij} to restrict the order of regime shifts. For example, [12] notes that the transition probabilities can be restricted in such a way so that the model in (2) becomes a "changepoint" model in which there are $N-1$ structural breaks in the model parameters. Finally, the vast majority of the applied literature has assumed that the transition probabilities in (4) evolve independently of lagged values of the series itself, so that

$$\begin{aligned} P(S_t = i | S_{t-1} = j, S_{t-2} = q, \ldots, y_{t-1}, y_{t-2}, \ldots) \\ = P(S_t = i | S_{t-1} = j) = p_{ij} \,, \end{aligned} \tag{5}$$

which is the polar opposite of the threshold process described in (3). For this reason, Markov-switching models are often described as having regimes that evolve "exogenously" of the series, while threshold models are said to have "endogenous" regimes. However, while popular in practice, the restriction in (5) is not necessary for estimation of the parameters of the Markov-switching model. Section "Extensions of the Basic Markov-Switching Model" of this article discusses models in which the transition probabilities of the Markov process are allowed to be partially determined by lagged values of the series.

The threshold and Markov-switching approaches are best viewed as complementary, with the "best" model likely to be application specific. Certain applications appear tailor-made for the threshold assumption. For example, we might have good reason to think that the behavior of time series such as an exchange rate or inflation will exhibit regime shifts when the series moves outside of certain thresholds, as this will trigger government intervention. The Markov-switching model might instead be the obvious choice when one does not wish to tie the regime shifts to the behavior of a particular observed variable, but instead wishes to let the data speak freely as to when regime shifts have occurred.

In the remainder of this article I will survey various aspects regarding the econometrics of Markov-switching models. For readers interested in learning more about threshold models, the survey article of Potter [78] is an excellent starting point.

Estimation of a Basic Markov-Switching Model

This section discusses estimation of the parameters of Markov-switching models. The existing literature has focused almost exclusively on likelihood-based methods for estimation. I retain this focus here, and discuss both maximum likelihood and Bayesian approaches to estimation. An alternative approach based on semi-parametric estimation is discussed in [4].

To aid understanding, we focus on a specific baseline case, which is the Markov-switching autoregression given in (2) and (5). We simplify further by allowing for $N = 2$ regimes, so that $S_t = 1$ or 2. It is worth noting that in many cases two regimes is a reasonable assumption. For example, in the literature using Markov-switching models to study business cycles phases, a two regime model, meant to capture an expansion and recession phase, is an obvious starting point that has been used extensively.

Estimation of Markov-switching models necessitates two additional restrictions over constant parameter models. First of all, the labeling of S_t is arbitrary, in that switching the vector of parameters associated with $S_t = 1$ and $S_t = 2$ will yield an identical model. A commonly used approach to normalize the model is to restrict the value of one of the parameters when $S_t = 1$ relative to its value when $S_t = 2$. For example, for the model in (2) we could restrict $\alpha_2 < \alpha_1$. For further details on the choice of normalization, see [49]. Second, the transition probabilities in (5) must be constrained to lie in [0,1]. One approach to implement this constraint, which will be useful in later discussion, is to use a probit specification for S_t. In particular, the value of S_t is assumed to be determined by the realization of a random variable, η_t, as follows:

$$S_t = \left\{ \begin{array}{lll} 1 & \text{if} & \eta_t < \gamma_{S_{t-1}} \\ 2 & \text{if} & \eta_t \geq \gamma_{S_{t-1}} \end{array} \right\} , \tag{6}$$

where $\eta_t \sim i.i.d.N(0, 1)$. The specification in (6) depends on two parameters, γ_1 and γ_2, which determine the transition probabilities of the Markov process as follows:

$$\begin{array}{l} p_{1j} = P(\eta_t < \gamma_j) = \Phi(\gamma_j) \\ p_{2j} = 1 - p_{1j} \end{array} , \tag{7}$$

where $j = 1$, 2 and Φ is the standard normal cumulative distribution function.

There are two main items of interest on which to conduct statistical inference for Markov-switching models. The first are the parameters of the model, of which there are $2(k + 3)$ for the two-regime Markov-switching autoregression. In the following we collect these parameters in the vector

$$\theta = (\alpha_1, \phi_{1,1}, \phi_{2,1}, \ldots, \phi_{k,1}, \sigma_1, \alpha_2, \phi_{1,2}, \phi_{2,2}, \ldots, \phi_{k,2}, \sigma_2, \gamma_1, \gamma_2)' . \tag{8}$$

The second item of interest is the regime indicator variable, S_t. In particular, as S_t is unobserved, we will be interested in constructing estimates of which regime was in operation at each point in time. These estimates will take the form of posterior probabilities that $S_t = i, i = 1, 2$. We assume that the econometrician has a sample of $T + k$ observations, $(y_T, y_{T-1}, y_{T-2}, \ldots, y_{-(k-1)})$. The series of observations available up to time t is denoted as $\Omega_t = (y_t, y_{t-1}, y_{t-2}, \ldots, y_{-(k-1)})$.

We begin with maximum likelihood estimation of θ. Maximum likelihood estimation techniques for various versions of Markov-switching regressions can be found in the existing literature of multiple disciplines, for example [52,76,79] in the speech recognition literature, and [16,41] in the econometrics literature. Here we focus on the presentation of the problem given in [41], who presents a simple iterative algorithm that can be used to construct the likelihood function of a Markov-switching autoregression, as well as compute posterior probabilities for S_t.

For a given value of θ, the conditional log likelihood function is given by:

$$L(\theta) = \sum_{t=1}^{T} \log f(y_t | \Omega_{t-1}; \theta) . \tag{9}$$

Construction of the conditional log likelihood function then requires construction of the conditional density function, $f(y_t | \Omega_{t-1}; \theta)$, for $t = 1, \ldots, T$. The "Hamilton

Filter" computes these conditional densities recursively as follows: Suppose for the moment that we are given $P(S_{t-1} = j|\Omega_{t-1};\theta)$, which is the posterior probability that $S_{t-1} = j$ based on information observed through period $t-1$. Equations (10) and (11) can then be used to construct $f(y_t|\Omega_{t-1};\theta)$:

$$P(S_t = i|\Omega_{t-1};\theta) = \sum_{j=1}^{2} P\left(S_t = i|S_{t-1} = j, \Omega_{t-1};\theta\right) \cdot P\left(S_{t-1} = j|\Omega_{t-1};\theta\right) , \quad (10)$$

$$f\left(y_t|\Omega_{t-1};\theta\right) = \sum_{i=1}^{2} f\left(y_t|S_t = i, \Omega_{t-1};\theta\right) \cdot P\left(S_t = i|\Omega_{t-1};\theta\right) . \quad (11)$$

From (5), the first term in the summation in (10) is simply the transition probability, p_{ij}, which is known for any particular value of θ. The first term in (11) is the conditional density of y_t assuming that $S_t = i$, which, given the within-regime normality assumption for ε_t, is:

$$f(y_t|S_1 = i, \Omega_0;\theta) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(\frac{-\left(y_t - \alpha_i - \sum_{j=1}^{k}\phi_{j,i}y_{t-j}\right)^2}{2\sigma_i^2}\right) . \quad (12)$$

With $f(y_t|\Omega_{t-1};\theta)$ in hand, the next step is then to update (10) and (11) to compute $f(y_{t+1}|\Omega_t;\theta)$. To do so requires $P(S_t = i|\Omega_t;\theta)$ as an input, meaning we must update $P(S_t = i|\Omega_{t-1};\theta)$ to reflect the information contained in y_t. This updating is done using Bayes' rule:

$$P(S_t = i|\Omega_t;\theta) = \frac{f(y_t|S_t = i, \Omega_{t-1};\theta)P\left(S_t = i|\Omega_{t-1}\right)}{f\left(y_t|\Omega_{t-1};\theta\right)} , \quad (13)$$

where each of the three elements on the right-hand side of (13) are computable from the elements of (10) and (11). Given a value for $P(S_0 = i|\Omega_0;\theta)$ to initialize the filter, Eqs. (10) through (13) can then be iterated to construct $f(y_t|\Omega_{t-1};\theta)$, $t = 1,\ldots,T$, and therefore the log likelihood function, $L(\theta)$. The *maximum likelihood estimate* $\hat{\theta}_{\text{MLE}}$, is then the value of θ that maximizes $L(\theta)$, and can be obtained using standard numerical optimization techniques.

How do we set $P(S_0 = i|\Omega_0;\theta)$ to initialize the filter? As is discussed in [41], exact evaluation of this probability is rather involved. The usual practice, which is possible when S_t is an ergodic Markov chain, is to simply set $P(S_0 = i|\Omega_0;\theta)$ equal to the unconditional probability, $P(S_0 = i)$. For the two-regime case considered here, these unconditional probabilities are given by:

$$P(S_0 = 1) = \frac{1 - p_{22}}{2 - p_{11} - p_{22}}$$
$$P(S_0 = 2) = 1 - P(S_0 = 1) . \quad (14)$$

Alternatively, $P(S_0 = i|\Omega_0;\theta)$ could be treated as an additional parameter to be estimated. See Hamilton [43] and Kim and Nelson [62] for further details.

An appealing feature of the Hamilton filter is that, in addition to the likelihood function, the procedure also directly evaluates $P(S_t = i|\Omega_t;\theta)$, which is commonly referred to as a "filtered" probability. Inference regarding the value of S_t is then sometimes based on $P(S_t = i|\Omega_t;\hat{\theta}_{\text{MLE}})$, which is obtained by running the Hamilton filter with $\theta = \hat{\theta}_{\text{MLE}}$. In many circumstances, we might also be interested in the so-called "smoothed" probability of a regime computed using all available data, or $P(S_t = i|\Omega_T;\theta)$. [54] presents an efficient recursive algorithm that can be applied to compute these smoothed probabilities.

We now turn to Bayesian estimation of Markov-switching models. In the Bayesian approach, the parameters θ are themselves assumed to be random variables, and the goal is to construct the posterior density for these parameters given the observed data, denoted $f(\theta|\Omega_T)$. In all but the simplest of models, this posterior density does not take the form of any well known density whose properties can be analyzed analytically. In this case, modern Bayesian inference usually proceeds by sampling the posterior density repeatedly to form estimates of posterior moments and other objects of interest. These estimates can be made arbitrarily accurate by increasing the number of samples taken from the posterior. In the case of Markov-switching models, Albert and Chib [1] demonstrate that samples from $f(\theta|\Omega_T)$ can be obtained using a simulation-based approach known as the Gibbs Sampler. The Gibbs Sampler, introduced by [37,38,89], is an algorithm that produces random samples from the joint density of a group of random variables by repeatedly sampling from the full set of conditional densities for the random variables.

We will sketch out the main ideas of the Gibbs Sampler in the context of the two-regime Markov-switching autoregression. It will prove useful to divide the parameter space into $\theta = (\theta_1', \theta_2')'$, where $\theta_1 = (\alpha_1, \phi_{1,1}, \phi_{2,1}, \ldots, \phi_{k,1}, \sigma_1, \alpha_2, \phi_{1,2}, \phi_{2,2}, \ldots, \phi_{k,2}, \sigma_2)'$ and $\theta_2 = (\gamma_1, \gamma_2)'$. Suppose it is feasible to simulate draws from the three conditional distributions, $f(\theta_1|\theta_2, \tilde{S}, \Omega_T)$, $f(\theta_2|\theta_1, \tilde{S}, \Omega_T)$, and $P(\tilde{S}|\theta_1, \theta_2, \Omega_T)$, where $\tilde{S} = (S_1, S_2, \ldots, S_T)'$. Then, conditional on arbitrary initial values, $\theta_2^{(0)}$ and

$\tilde{S}^{(0)}$, we can obtain a draw of θ_1, denoted $\theta_1^{(1)}$, from $f(\theta_1|\theta_2^{(0)}, \tilde{S}^{(0)}, \Omega_T)$, a draw of θ_2, denoted $\theta_2^{(1)}$, from $f(\theta_2|\theta_1^{(1)}, \tilde{S}^{(0)}, \Omega_T)$, and a draw of $\tilde{S}$, denoted $\tilde{S}^{(1)}$, from $P(\tilde{S}|\theta_1^{(1)}, \theta_2^{(1)}, \Omega_T)$. This procedure can be iterated to obtain $\theta_1^{(j)}, \theta_2^{(j)}$, and $\tilde{S}^{(j)}$, for $j = 1, \ldots, J$. For large enough J, and assuming weak regularity conditions, these draws will converge to draws from $f(\theta|\Omega_T)$ and $P(\tilde{S}|\Omega_T)$. Then, by taking a large number of such draws beyond J, one can estimate any feature of $f(\theta|\Omega_T)$ and $P(\tilde{S}|\Omega_T)$, such as moments of interest, with an arbitrary degree of accuracy. For example, an estimate of $P(S_t = i|\Omega_T)$ can be obtained by computing the proportion of draws of $\tilde{S}$ for which $S_t = i$.

Why is the Gibbs Sampler useful for a Markov-switching model? It turns out that although $f(\theta|\Omega_t)$ and $P(\tilde{S}|\Omega_T)$ cannot be sampled directly, it is straightforward, assuming natural conjugate prior distributions, to obtain samples from $f(\theta_1|\theta_2, \tilde{S}, \Omega_T)$, $f(\theta_2|\theta_1, \tilde{S}, \Omega_T)$, and $P(\tilde{S}|\theta_1, \theta_2, \Omega_T)$. This is most easily seen for the case of θ_1, which, when $\tilde{S}$ is conditioning information, represents the parameters of a linear regression with dummy variables, a case for which techniques to sample the parameter posterior distribution are well established (Zellner 96). An algorithm for obtaining draws of $\tilde{S}$ from $P(\tilde{S}|\theta_1, \theta_2, \Omega_T)$ was first given in Albert and Chib [1], while Kim and Nelson [59] develop an alternative, efficient, algorithm based on the notion of "multi-move" Gibbs Sampling introduced in [6]. For further details regarding the implementation of the Gibbs Sampler in the context of Markov-switching models, see Kim and Nelson [62].

The Bayesian approach has a number of features that make it particularly attractive for estimation of Markov-switching models. First of all, the requirement of prior density functions for model parameters, considered by many to be a weakness of the Bayesian approach in general, is often an advantage for Bayesian analysis of Markov-switching models [42]. For example, priors can be used to push the model toward capturing one type of regime-switching vs. another. The value of this can be seen for Markov-switching models of the business cycle, for which the econometrician might wish to focus on portions of the likelihood surface related to business cycle switching, rather than those related to longer term regime shifts in productivity growth. Another advantage of the Bayesian approach is with regards to the inference drawn on S_t. In the maximum likelihood approach, the methods of [54] can be applied to obtain $P(S_t = i|\Omega_T; \hat{\theta}_{\mathrm{MLE}})$. As these probabilities are conditioned on the maximum likelihood parameter estimates, uncertainty regarding the unknown values of the parameters has not been taken into account. By contrast, the Bayesian approach yields $P(S_t = i|\Omega_T)$, which is not conditional on a particular value of θ and thus incorporates uncertainty regarding the value of θ that generated the observed data.

Extensions of the Basic Markov-Switching Model

The basic, two-regime Markov-switching autoregression in (2) and (5) has been used extensively in the literature, and remains a popular specification in applied work. However, it has been extended in a number of directions in the substantial literature that follows [41]. This section surveys a number of these extensions.

The estimation techniques discussed in Sect. "Estimation of a Basic Markov-Switching Model" can be adapted in a straightforward manner to include several extensions to the basic Markov-switching model. For example, the filter used in (10) through (13) can be modified in obvious ways to incorporate the case of $N > 2$ regimes, as well as to allow y_t to be a vector of random variables, so that the model in (2) becomes a Markov-switching vector autoregression (MS-VAR). Hamilton [43] discusses both of these cases, while Krolzig [68] provides an extensive discussion of MS-VARs. [83] is a recent example of applied work using an MS-VAR with a large number of regimes. In addition, the (known) within-regime distribution of the disturbance term, ε_t, could be non-Gaussian, as in [23] or [45]. Further, the parameters of (2) could be extended to depend not just on S_t, but also on a finite number of lagged values of S_t, or even a second state variable possibly correlated with S_t. Indeed, such processes can generally be rewritten in terms of the current value of a single, suitably redefined, state variable. [58,66] provide examples of such a redefinition. For further discussion of all of these cases, see [43].

The specification for the transition probabilities in (5) restricted the probability $S_t = i$ to depend only on the value of S_{t-1}. However, in some applications we might think that these transition probabilities are driven in part by observed variables, such as the past evolution of the process. To this end, [21,28] develop Markov-switching models with time-varying transition probabilities (TVTP), in which the transition probabilities are allowed to vary depending on conditioning information. Suppose that z_t represents a vector of observed variables that are thought to influence the realization of the regime. The probit representation for the state process in (6) and (7) can then be extended as follows:

$$S_t = \left\{ \begin{array}{lll} 1 & \text{if} & \eta_t < \left(\gamma_{S_{t-1}} + z_t' \lambda_{S_{t-1}}\right) \\ 2 & \text{if} & \eta_t \geq \left(\gamma_{S_{t-1}} + z_t' \lambda_{S_{t-1}}\right) \end{array} \right\} , \qquad (15)$$

with associated transition probabilities:

$$\begin{aligned} p_{1j}(z_t) &= P\big(\eta_t < (\gamma_j + z_t'\lambda_j)\big) = \Phi\big(\gamma_j + z_t'\lambda_j\big) \\ p_{2j}(z_t) &= 1 - p_{1j}(z_t)\,, \end{aligned} \tag{16}$$

where $j = 1$, 2 and Φ is again the standard normal cumulative distribution function. Estimation of the Markov-switching autoregression with TVTP is then straightforward. In particular, assuming that z_t contains lagged values of y_t or exogenous random variables, a maximum likelihood estimation proceeds by simply replacing p_{ij} with $p_{ij}(z_t)$ in the filter given in (10) through (13). Bayesian estimation of TVTP models via the Gibbs Sampler is also straightforward, and is discussed in [29]. Despite its intuitive appeal, the literature contains relatively few applications of the TVTP model. A notable example of the TVTP framework is found in Durland and McCurdy [24], Filardo and Gordon [29] and Kim and Nelson [59], who study business cycle "duration dependence", or whether the probability of a business cycle phase shift depends on how long the economy has been in the current phase. Other applications include Ang and Bekaert [2], who model regime-switches in interest rates, and Lo and Piger [69], who investigate sources of time-variation in the response of output to monetary policy actions.

The TVTP model is capable of relaxing the restriction that the state variable, S_t, is independent of the lagged values of the series, y_t, and thus of lagged values of the disturbance term, ε_t. Kim, Piger and Startz [65] consider a Markov-switching model in which S_t is also correlated with the contemporaneous value of ε_t, and is thus "endogenous". They model this endogenous switching by assuming that the shock to the probit process in (6), η_t, and ε_t are jointly normally distributed as follows:

$$\begin{bmatrix} \varepsilon_t \\ \eta_t \end{bmatrix} \sim N(0, \Sigma),\ \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}. \tag{17}$$

Kim, Piger and Startz [65] show that when $\rho \neq 0$, the conditional density in (12) is no longer Gaussian, but can be evaluated analytically. Thus, the likelihood function for the endogenous switching model can be evaluated with simple modifications to the recursive filter in (10) through (13). Tests of the null hypothesis that S_t is exogenous can also be implemented in a straightforward manner. Chib and Dueker [13] consider endogenous switching as in (17) from a Bayesian perspective.

The extensions listed above are primarily modifications to the stochastic process assumed to drive S_t. A more fundamental extension of (2) is to consider Markov-switching in time series models that are more complicated than simple autoregressions. An important example of this is a state-space model with Markov-switching parameters. Allowing for Markov-switching in the state-space representation for a time series is particularly interesting because a large number of popular time-series models can be given a state-space representation. Thus, incorporating Markov-switching into a general state-space representation immediately extends the Markov-switching framework to these models.

To aid discussion, consider the following Markov-switching state-space representation for a vector of R random variables, $Y_t = (y_{1t}, y_{2t}, \ldots, y_{Rt})'$, given as follows:

$$\begin{aligned} Y_t &= H_{S_t}' X_t + W_t \\ X_t &= A_{S_t} + F_{S_t} X_{t-1} + V_t \end{aligned}, \tag{18}$$

where $X_t = (x_{1t}, x_{2t}, \ldots, x_{Dt})'$, $W_t \sim N(0, B_{S_t})$ and $V_t \sim N(0, Q_{S_t})$. The parameters of the model undergo Markov switching, and are contained in the matrices $H_{S_t}, B_{S_t}, A_{S_t}, F_{S_t}, Q_{S_t}$. A case of primary interest is when some or all of the elements of X_t are unobserved. This is the case for a wide range of important models in practice, including models with moving average (MA) dynamics, unobserved components (UC) models, and dynamic factor models. However, in the presence of Markov-switching parameters, the fact that X_t is unobserved introduces substantial complications for construction of the likelihood function. In particular, as is discussed in detail in [54] and Kim and Nelson [62], exact construction of the conditional density $f(y_t|\Omega_{t-1};\theta)$ requires that one consider all possible permutations of the entire history of the state variable, $S_t, S_{t-1}, S_{t-2}, \ldots, S_1$. For even moderately sized values of t, this quickly becomes computationally infeasible.

To make inference via maximum likelihood estimation feasible, [54] develops a recursive filter that constructs an approximation to the likelihood function. This filter "collapses" the number of lagged regimes that are necessary to keep track of by approximating a nonlinear expectation with a linear projection. Kim and Nelson [62] provide a detailed description of the Kim [54] filter, as well as a number of examples of its practical use.

If one is willing to take a Bayesian approach to the problem, Kim and Nelson [59] show that inference can be conducted via the Gibbs Sampler without resorting to approximations. As before, the conditioning features of the Gibbs sampler greatly simplifies the analysis. For example, by conditioning on $\tilde{S} = (S_1, S_2, \ldots, S_T)'$, the model in (18) is simply a linear, Gaussian, state-space model with dummy variables, for which techniques to sample the posterior distribution of model parameters and the unobserved elements of X_t are well established [6]. Kim and

Nelson [62] provide detailed descriptions of how the Gibbs Sampler can be implemented for a state-space model with Markov switching.

There are many applications of state space models with Markov switching. For example, a large literature uses UC models to decompose measures of economic output into trend and cyclical components, with the cyclical component often interpreted as a measure of the business cycle. Until recently, this literature focused on linear representations for the trend and cyclical components [14,51,72,94]. However, one might think that the processes used to describe the trend and cyclical components might display regime switching in a number of directions, such as that related to the phase of the business cycle or to longer-run structural breaks in productivity growth or volatility. A UC model with Markov switching in the trend and cyclical components can be cast as a Markov-switching state-space model as in (18). Applications of such regime-switching UC models can be found in [58,60,64,71,84]. Another primary example of a Markov-switching state-space model is a dynamic factor model with Markov-switching parameters, examples of which are given in [7,59]. Section "Empirical Example: Identifying Business Cycle Turning Points" presents a detailed empirical example of such a model.

Specification Testing for Markov-Switching Models

Our discussion so far has assumed that key elements in the specification of regime-switching models are known to the researcher. Chief among these is the number of regimes, N. However, in practice there is likely uncertainty about the appropriate number of regimes. This section discusses data-based techniques that can be used to select the value of N.

To fix ideas, consider a simple version of the Markov-switching model in (2):

$$y_t = \alpha_{S_t} + \varepsilon_t \,, \tag{19}$$

where $\varepsilon_t \sim N(0, \sigma^2)$. Consider the problem of trying to decide between a model with $N = 2$ regimes vs. the simpler model with $N = 1$ regimes. The model with one regime is a constant parameter model, and thus this problem can be interpreted as a decision between a model with regime-switching parameters vs. one without. An obvious choice for making this decision is to construct a test of the null hypothesis of $N = 1$ vs. the alternative of $N = 2$. For example, one might construct the likelihood ratio statistic:

$$LR = 2\big(L(\hat{\theta}_{\mathrm{MLE}(2)}) - L(\hat{\theta}_{\mathrm{MLE}(1)})\big) \,, \tag{20}$$

where $\hat{\theta}_{\mathrm{MLE}(1)}$ and $\hat{\theta}_{\mathrm{MLE}(2)}$ are the maximum likelihood estimates under the assumptions of $N = 1$ and $N = 2$ respectively. Under the null hypothesis there are three fewer parameters to estimate, α_2, γ_1 and γ_2, than under the alternative hypothesis. Then, to test the null hypothesis, one might be tempted to proceed by constructing a p-value for LR using the standard χ^2 (3) distribution.

However, this final step is not justified, and can lead to very misleading results in practice. In particular, the standard conditions for LR to have an asymptotic χ^2 distribution include that all parameters are identified under the null hypothesis [17]. In the case of the model in (19), the parameters γ_1 and γ_2, which determine the transition probabilities p_{ij}, are not identified assuming the null hypothesis is true. In particular, if $\alpha_1 = \alpha_2$, then p_{ij} can take on any values without altering the likelihood function for the observed data. A similar problem exists when testing the general case of N vs. $N + 1$ regimes.

Fortunately, a number of contributions in recent years have produced asymptotically justified tests of the null hypothesis of N regimes vs. the alternative of $N + 1$ regimes. In particular, [33,50] provide techniques to compute asymptotically valid critical values for LR. Recently Carrasco, Hu and Ploberger [5] have developed an asymptotically optimal test for the null hypothesis of parameter constancy against the general alternative of Markov-switching parameters. Their test is particularly appealing because it does not require estimation of the model under the alternative hypothesis, as is the case with LR.

If one is willing to take a Bayesian approach, the comparison of models with N vs. $N + 1$ regimes creates no special considerations. In particular, one can proceed by computing standard Bayesian model comparison metrics, such as Bayes Factors or posterior odds ratios. Examples of such comparisons can be found in [11,63,78].

Empirical Example: Identifying Business Cycle Turning Points

This section presents an empirical example demonstrating how the Markov-switching framework can be used to model shifts between expansion and recession phases in the US business cycle. This example is of particular interest for two reasons. First, although Markov-switching models have been used to study a wide variety of topics, their most common application has been as formal statistical models of business cycle phase shifts. Second, the particular model we focus on here, a dynamic factor model with Markov-switching parameters, is of interest in its own right, with a number of potential applications.

The first presentation of a Markov-switching model of the business cycle is found in [41]. In particular, [41] showed that US real GDP growth could be characterized as an autoregressive model with a mean that switched between low and high growth regimes, where the estimated timing of the low growth regime corresponded closely to the dates of US recessions as established by the Business Cycle Dating Committee of the National Bureau of Economic Research (NBER). This suggested that Markov-switching models could be used as tools to identify the timing of shifts between business cycle phases, and a great amount of subsequent analysis has been devoted toward refining and using the Markov-switching model for this task.

The model used in [41] was univariate, considering only real GDP. However, as is discussed in [22], a long emphasized feature of the business cycle is comovement, or the tendency for business cycle fluctuations to be observed simultaneously in a large number of economic sectors and indicators. This suggests that, by using information from many economic indicators, the identification of business cycle phase shifts might be sharpened. One appealing way of capturing comovement in a number of economic indicators is through the use of dynamic factor models, as popularized by [85,86]. However, these models assumed constant parameters, and thus do not model business cycle phase shifts explicitly.

To simultaneously capture comovement and business cycle phase shifts, [7] introduces Markov-switching parameters into the dynamic factor model of [85,86]. Specifically, defining $y^*_{rt} = y_{rt} - \bar{y}_r$ as the demeaned growth rate of the rth economic indicator, the dynamic factor Markov-switching (DFMS) model has the form:

$$y^*_{rt} = \beta_r c_t + e_{rt} \,. \tag{21}$$

In (21), the demeaned first difference of each series is made up of a component common to each series, given by the dynamic factor c_t, and a component idiosyncratic to each series, given by e_{rt}. The common component is assumed to follow a stationary autoregressive process:

$$\phi(L)(c_t - \mu_{S_t}) = \varepsilon_t \,, \tag{22}$$

where $\varepsilon_t \sim i.i.d.N(0,1)$. The unit variance for ε_t is imposed to identify the parameters of the model, as the factor loading coefficients, β_r, and the variance of ε_t are not separately identified. The lag polynomial $\phi(L)$ is assumed to have all roots outside of the unit circle. Regime switching is introduced by allowing the common component to have a Markov-switching mean, given by μ_{S_t}, where $S_t = \{1, 2\}$. The regime is normalized by setting $\mu_2 < \mu_1$. Finally, each idiosyncratic component is assumed to follow a stationary autoregressive process:

$$\theta_r(L) e_{rt} = \omega_{rt} \,. \tag{23}$$

where $\theta_r(L)$ is a lag polynomial with all roots outside the unit circle and $\omega_{rt} \sim N(0, \sigma^2_{\omega,r})$.

[7] estimates the DFMS model for US monthly data on non-farm payroll employment, industrial production, real manufacturing and trade sales, and real personal income excluding transfer payments, which are the four monthly variables highlighted by the NBER in their analysis of business cycles. The DFMS model can be cast as a state-space model with Markov switching of the type discussed in Sect. "Extensions of the Basic Markov-Switching Model". Chauvet estimates the parameters of the model via maximum likelihood, using the approximation to the likelihood function given in [54]. Kim and Nelson [59] instead use Bayesian estimation via the Gibbs Sampler to estimate the DFMS model.

Here I update the estimation of the DFMS model presented in [59] to a sample period extending from February 1967 through February 2007. For estimation, I use the Bayesian Gibbs Sampling approach, with prior distributions and specification details identical to those given in [59]. Figure 1 displays $P(S_t = 2|\Psi_T)$ obtained from the Gibbs Sampler, which is the estimated probability that the low growth regime is active. For comparison, Fig. 1 also indicates NBER recession dates with shading.

There are two items of particular interest in Fig. 1. First of all, the estimated probability of the low growth regime is very clearly defined, with $P(S_t = 2|\Psi_T)$ generally close to either zero or one. Indeed, of the 481 months

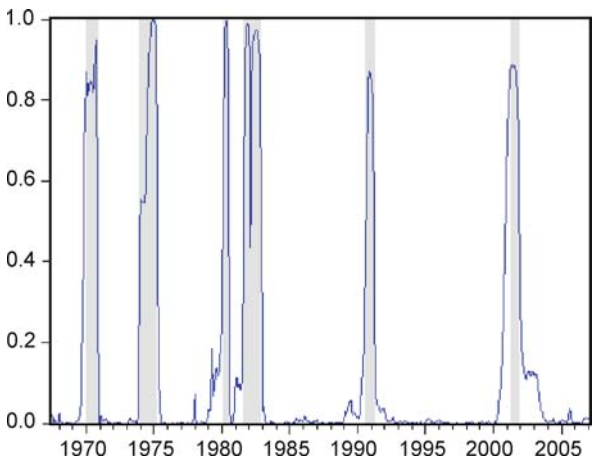

Econometrics: Models of Regime Changes, Figure 1
Probability of US Recession from Dynamic Factor Markov-Switching Model

in the sample, only 32 had $P(S_t = 2|\Psi_T)$ fall between 0.2 and 0.8. Second, $P(S_t = 2|\Psi_T)$ is very closely aligned with NBER expansion and recession dates. In particular, $P(S_t = 2|\Psi_T)$ tends to be very low during NBER expansion phases and very high during NBER recession phases.

Figure 1 demonstrates the added value of employing the DFMS model, which considers the comovement between multiple economic indicators, over models considering only a single measure of economic activity. In particular, results for the Markov-switching autoregressive model of real GDP presented in [41] were based on a data sample ending in 1984, and it is well documented that Hamilton's original model does not perform well for capturing the two NBER recessions since 1984. Subsequent research has found that allowing for structural change in the residual variance parameter [61,70] or omitting all linear dynamics in the model [1,9] improves the Hamilton model's performance. By contrast, the results presented here suggest that the DFMS model accurately identifies the NBER recession dates without a need for structural breaks or the omission of linear dynamics.

In some cases, we might be interested in converting $P(S_t = 2|\Psi_T)$ into a specific set of dates establishing the timing of shifts between business cycle phases. To do so requires a rule for establishing whether a particular month was an expansion month or a recession month. Here we consider a simple rule, which categorizes any particular month as an expansion month if $P(S_t = 2|\Psi_T) \leq 0.5$ and a recession month if $P(S_t = 2|\Psi_T) > 0.5$. Table 1 displays the dates of turning points between expansion and recession phases (business cycle peaks), and the dates of turning points between recession and expansion phases (business cycle troughs) that are established by this rule. For comparison, Table 1 also lists the NBER peak and trough dates.

Table 1 demonstrates that the simple rule applied to $P(S_t = 2|\Psi_T)$ does a very good job of matching the NBER peak and trough dates. Of the twelve turning points in the sample, the DFMS model establishes eleven within two months of the NBER date. The exception is the peak of the 2001 recession, for which the peak date from the DFMS model is four months prior to that established by the NBER. In comparing peak and trough dates, the DFMS model appears to do especially well at matching NBER trough dates, for which the date established by the DFMS model matches the NBER date exactly in five of six cases.

Why has the ability of Markov-switching models to identify business cycle turning points generated so much attention? There are at least four reasons. First, it is sometimes argued that recession and expansion phases may not be of any intrinsic interest, as they need not reflect any real differences in the economy's structure. In particular, as noted by [95], simulated data from simple, constant parameter, time-series models, for which the notion of separate regimes is meaningless, will contain episodes that look to the eye like "recession" and "expansion" phases. By capturing the notion of a business cycle phase formally inside of a statistical model, the Markov-switching model is then able to provide statistical evidence as to the extent to which business cycle phases are a meaningful concept. Second, although the dates of business cycle phases and their associated turning points are of interest to many economic researchers, they are not compiled in a systematic fashion for many economies. Markov-switching models could then be applied to obtain business cycle turning point dates for these economies. An example of this is given in [74], who use Markov-switching models to establish business cycle phase dates for US states. Third, if economic time-series do display different behavior over business cycle phases, then Markov-switching models designed to capture such differences might be exploited to obtain more accurate forecasts of economic activity. Finally, the current probability of a new economic turning point is likely of substantial interest to economic policymakers. To this end, Markov-switching models can be used for "real-time" monitoring of new business cycle phase shifts. Indeed, Chauvet and Piger [10] provide evidence that Markov-switching mod-

Econometrics: Models of Regime Changes, Table 1
Dates of Business Cycle Turning Points Produced by NBER and Dynamic Factor Markov-Switching Model

Peaks			Troughs		
DFMS	NBER	Discrepancy	DFMS	NBER	Discrepancy
Oct 1969	Dec 1969	2M	Nov 1970	Nov 1970	0M
Dec 1973	Nov 1973	−1M	Mar 1975	Mar 1975	0M
Jan 1980	Jan 1980	0M	Jun 1980	Jul 1980	1M
Jul 1981	Jul 1981	0M	Nov 1982	Nov 1982	0M
Aug 1990	Jul 1990	−1M	Mar 1991	Mar 1991	0M
Nov 2000	Mar 2001	4M	Nov 2001	Nov 2001	0M

els are often quicker to establish US business cycle turning points, particularly at business cycle troughs, than is the NBER. For additional analysis of the ability of regime-switching models to establish turning points in real time, see [8,9].

Future Directions

Research investigating applied and theoretical aspects of regime-switching models should be an important component of the future research agenda in macroeconomics and econometrics. In this section I highlight three directions for future research which are of particular interest.

To begin, additional research oriented toward improving the forecasting ability of regime-switching models is needed. In particular, given that regime-switching models of economic data contain important deviations from traditional, constant parameter, alternatives, we might expect that they could also provide improved out-of-sample forecasts. However, as surveyed in [15], the forecasting improvements generated by regime-switching models over simpler alternatives is spotty at best. That this is true is perhaps not completely surprising. For example, the ability of a Markov-switching model to identify regime shifts in past data does not guarantee that the model will do well at detecting regime shifts quickly enough in real time to generate improved forecasts. This is particularly problematic when regimes are short lived. Successful efforts to improve the forecasting ability of Markov-switching models are likely to come in the form of multivariate models, which can utilize additional information for quickly identifying regime shifts.

A second potentially important direction for future research is the extension of the Markov-switching dynamic factor model discussed in Sects."Extensions of the Basic Markov-Switching Model" and "Empirical Example: Identifying Business Cycle Turning Points" to settings with a large cross-section of data series. Indeed, applications of the DFMS model have been largely restricted to a relatively small number of variables, such as in the model of the US business cycle considered in Sect. "Empirical Example: Identifying Business Cycle Turning Points". However, in recent years there have been substantial developments in the analysis of dynamic factor models comprising a large number of variables, as in [31,32,87,88,92]. Research extending the regime-switching framework to such "big data" factor models will be of substantial interest.

Finally, much remains to be done incorporating regime-switching behavior into structural macroeconomic models. A number of recent studies have begun this synthesis by considering the implications of regime-switches in the behavior of a fiscal or monetary policymaker for the dynamics and equilibrium behavior of model economies [18,19,20,26,27]. This literature has already yielded a number of new and interesting results, and is likely to continue to do so as it expands. Less attention has been paid to reconciling structural models with a list of new "stylized facts" generated by the application of regime-switching models in reduced-form settings. As one example, there is now a substantial list of studies, including [3,45,57,58,82], and Kim and Nelson [60] finding evidence that the persistence of shocks to key macroeconomic variables varies dramatically over business cycle phases. However, such an asymmetry is absent from most modern structural macroeconomic models, which generally possess a symmetric propagation structure for shocks. Research designed to incorporate and explain business cycle asymmetries and other types of regime-switching behavior inside of structural macroeconomic models will be particularly welcome.

Bibliography

1. Albert J, Chib S (1993) Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. J Bus Econ Stat 11:1–15
2. Ang A, Bekaert G (2002) Regime switches in interest rates. J Bus Econ Stat 20:163–182
3. Beaudry P, Koop G (1993) Do recessions permanently change output? J Monet Econ 31:149–163
4. Campbell SD (2002) Specification testing and semiparametric estimation of regime switching models: An examination of the us short term interest rate. Brown University Department of Economics Working Paper #2002–26, Providence
5. Carrasco M, Hu L, Ploberger W (2004) Optimal test for Markov switching. Working paper, University of Rochester, Rochester
6. Carter CK, Kohn P (1994) On Gibbs sampling for state space models. Biometrica 81:541–553
7. Chauvet M (1998) An Econometric Characterization of Business Cycle Dynamics with Factor Structure and Regime Switching. Int Econ Rev 39:969–996
8. Chauvet M, Hamilton J (2006) Dating Business Cycle Turning Points. In: Milas C, Rothman P, van Dijk D (eds) Nonlinear Time Series Analysis of Business Cycles. Elsevier, North Holland
9. Chauvet M, Piger J (2003) Identifying Business Cycle Turning Points in Real Time. Fed Res Bank St. Louis Rev 85:47–61
10. Chauvet M, Piger J (2004) A Comparison of the Real-Time Performance of Business Cycle Dating Methods. J Bus Econ Stat 26:42–49
11. Chib S (1995) Marginal Likelihood from the Gibbs Output. J Am Stat Assoc 90:1313–1321
12. Chib S (1998) Estimation and Comparison of Multiple Change-Point Models. J Econ 86:221–241
13. Chib S, Dueker M (2004) Non-Markovian Regime Switching with Endogenous States and Time Varying State Strengths. Federal Reserve Bank of St. Louis Working Paper #2004–030A, St. Louis

14. Clark PK (1987) The Cyclical Component of US Economic Activity. Quart J Econ 102:797–814
15. Clements MP, Franses PH, Swanson NR (2004) Forecasting Economic and Financial Time-Series with Non-Linear Models. Int J Forecast 20:169–183
16. Cosslett SR, Lee LF (1985) Serial Correlation in Discrete Variable Models. J Econ 27:79–97
17. Davies RB (1977) Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 64:247–254
18. Davig T, Leeper E (2005) Generalizing the Taylor Principle. Am Econ Rev 97:607–635
19. Davig T, Leeper E (2006) Endogenous Monetary Policy Regime Change. In: Reichlin L, West KD (eds) International Seminar on Macroeconomics. MIT Press, Cambridge
20. Davig T, Leeper E, Chung H (2004) Monetary and Fiscal Policy Switching. J Mon Credit Bank 39:809–842
21. Diebold FX, Lee JH, Weinbach G (1994) Regime Switching with Time-Varying Transition Probabilities. In: Hargreaves C (ed) Non-stationary Time Series Analysis and Cointegration. Oxford University Press, Oxford UK
22. Diebold FX, Rudebusch GD (1996) Measuring business cycles: A modern perspective. Rev Econ Stat 78:67–77
23. Dueker M (1997) Markov Switching in GARCH Processes and Mean-Reverting Stock-Market Volatility. J Bus Econ Stat 15:26–34
24. Durland JM, McCurdy TH (1994) Duration Dependent Transitions in a Markov Model of USGNP Growth. J Bus Econ Stat 12:279–288
25. Evans M, Wachtel P (1993) Inflation Regimes and the Sources of Inflation Uncertainty. J Mon Credit Bank 25:475–511
26. Farmer REA, Waggoner DF, Zha T (2006) Indeterminacy in a Forward Looking Regime Switching Model. NBER working paper no. 12540, Cambridge
27. Farmer REA, Waggoner DF, Zha T (2007) Understanding the New-Keynesian Model when Monetary Policy Switches Regimes. NBER working paper no. 12965, Cambridge
28. Filardo AJ (1994) Business-Cycle Phases and Their Transitional Dynamics. J Bus Econ Stat 12:299–308
29. Filardo AJ, Gordon SF (1998) Business Cycle Durations. J Econ 85:99–123
30. Forni M, Hallin M, Lippi F, Reichlin L (2000) The Generalized Dynamic Factor Model: Identification and Estimation. Rev Econ Stat 82:540–554
31. Forni M, Hallin M, Lippi F, Reichlin L (2002) The generalized dynamic factor model: consistency and convergence rates. J Econ 82:540–554
32. Forni M, Hallin M, Lippi F, Reichlin L (2005) The generalized dynamic factor model: one-sided estimation and forecasting. J Am Stat Assoc 100:830–840
33. Garcia R (1998) Asymptotic null distribution of the likelihood ratio test in Markov switching models, Int Econ Rev 39:763–788
34. Garcia R, Perron P (1996) An Analysis of the Real Interest Rate under Regime Shifts. Rev Econ Stat 78:111–125
35. Garcia R, Schaller H (2002) Are the Effects of Monetary Policy Asymmetric? Econ Inq 40:102–119
36. Granger CWJ, Teräsvirta T (1993) Modelling Nonlinear Economic Relationships. Oxford University Press, Oxford
37. Gelfand AE, Smith AFM (1990) Sampling-Based Approaches to Calculating Marginal Densities. J Am Stat Assoc 85:398–409
38. Geman S, Geman D (1984) Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. IEEE Trans Patt Anal Machine Int 6:721–741
39. Goldfeld SM, Quandt RE (1973) A Markov Model for Switching Regressions. J Econ 1:3–16
40. Guidolin M, Timmermann A (2005) Economic Implications of Bull and Bear Regimes in UK Stock and Bond Returns. Econ J 115:111–143
41. Hamilton JD (1989) A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. Econometrica 57:357–384
42. Hamilton JD (1991) A Quasi-Bayesian Approach to Estimating Parameters for Mixtures of Normal Distributions. J Bus Econ Statistics 9:27–39
43. Hamilton JD (1994) Time Series Analysis. Princeton University Press, Princeton NJ
44. Hamilton JD (2005) Regime-Switching Models. In: Durlauf S, Blume L (eds) New Palgrave Dictionary of Economics, 2nd edn. Palgrave McMillan Ltd, Hampshire
45. Hamilton JD (2005) What's Real About the Business Cycle? Fed Res Bank St. Louis Rev 87:435–452
46. Hamilton JD, Lin G (1996) Stock Market Volatility and the Business Cycle. J Appl Econ 11:573–593
47. Hamilton JD, Raj B (2002) New Directions in Business Cycle Research and Financial Analysis. Empir Econ 27:149–162
48. Hamilton JD, Susmel R (1994) Autoregressive Conditional Heteroskedasticity and Changes in Regime. J Econ 64:307–333
49. Hamilton, Waggoner JD DF, Zha T (2004) Normalization in Econometrics. Econ Rev 26:221–252
50. Hansen BE (1992) The likelihood ratio test under nonstandard conditions: Testing the Markov switching model of GNP. J Appl Econ 7:S61–S82
51. Harvey AC (1985) Trends and Cycles in Macroeconomic Time Series. J Bus Econ Stat 3:216–227
52. Juang BH, Rabiner LR (1985) Mixture Autoregressive Hidden Markov Models for Speech Signals. IEEE Trans Acoust Speech Signal Proc ASSP-30:1404–1413
53. Kaufmann S (2002) Is there an Asymmetric Effect of Monetary Policy Over Time? A Bayesian Analysis using Austrian Data. Empir Econ 27:277–297
54. Kim CJ (1994) Dynamic linear models with Markov-switching. J Econ 60:1–22
55. Kim CJ (2004) Markov-Switching Models with Endogenous Explanatory Variables. J Econ 122:127–136
56. Kim CJ, Morley J, Nelson C (2002) Is there a Positive Relationship between Stock Market Volatility and the Equity Premium? J Money Cred Bank 36:339–360
57. Kim CJ, Morley J, Piger J (2005) Nonlinearity and the Permanent Effects of Recessions. J Appl Econ 20:291–309
58. Kim CJ, Murray CJ (2002) Permanent and transitory components of recessions. Empir Econ 27:163–183
59. Kim CJ, Nelson CR (1998) Business Cycle Turning Points, a New Coincident Index, and Tests of Duration Dependence Based on a Dynamic Factor Model with Regime Switching. Rev Econ Stat 80:188–201
60. Kim CJ, Nelson CR (1999b) Friedman's Plucking Model of Business Fluctuations: Tests and Estimates of Permanent and Transitory Components. J Money Cred Bank 31:317–34
61. Kim CJ, Nelson CR (1999c) Has the US economy become more stable? A Bayesian approach based on a Markov-switching model of the business cycle. Rev Econ Stat 81:608–616

62. Kim CJ, Nelson C (1999a) State-Space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications. MIT Press, Cambridge
63. Kim CJ, Nelson CR (2001) A Bayesian approach to testing for Markov-switching in univariate and dynamic factor models. Int Econ Rev 42:989–1013
64. Kim CJ, Piger J (2002) Common stochastic trends, common cycles, and asymmetry in economic fluctuations. J Monet Econ 49:1189–1211
65. Kim CJ, Piger J, Startz R (2003) Estimation of Markov Regime-Switching Regression Models with Endogenous Switching. J Econom, (in press)
66. Kim CJ, Piger J, Startz R (2007) The Dynamic Relationship Between Permanent and Transitory Components of US Business Cycles. J Money Cred Bank 39:187–204
67. Koop G, Potter SM (1999) Bayes Factors and Nonlinearity: Evidence from Economic Time Series. J Econ 88:251–281
68. Krolzig HM (1997) Markov-Switching Vector Autoregressions: Modelling, Statistical Inference, and Application to Business Cycle Analysis. Springer, Berlin
69. Lo M, Piger J (2005) Is the Response of Output to Monetary Policy Asymmetric? Evidence from a Regime-Switching Coefficients Model. J Money Cred Bank 37:865–887
70. McConnell MM, Quiros GP (2000) Output Fluctuations in the United States: What has Changed Since the Early (1980s)? Am Econ Rev 90:1464–1476
71. Mills TC, Wang P (2002) Plucking Models of Business Cycle Fluctuations: Evidence from the G-7 Countries. Empir Econ 27:255–276
72. Morley JC, Nelson CR, Zivot E (2003) Why Are the Beveridge-Nelson and Unobserved-Components Decompositions of GDP So Different? Review Econ Stat 85:235–243
73. Öcal N, Osborn DR (2000) Business cycle non-linearities in UK consumption and production. J Appl Econ 15:27–44
74. Owyang MT, Piger J, Wall HJ (2005) Business Cycle Phases in US States. Rev Econ Stat 87:604–616
75. Pesaran MH, Potter SM (1997) A floor and ceiling model of US output. J Econ Dyn Control 21:661–695
76. Poritz AB (1982) Linear Predictive Hidden Markov Models and the Speech Signal. Acoustics, Speech and Signal Processing IEEE Conference on ICASSP '82, vol 7:1291–1294
77. Potter SM (1995) A Nonlinear Approach to US GNP. J Appl Econ 10:109–125
78. Potter SM (1999) Nonlinear Time Series Modelling: An Introduction. J Econ Surv 13:505–528
79. Rabiner LR (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proc IEEE 77:257–286
80. Rapach DE, Wohar ME (2002) Regime Changes in International Real Interest Rates: Are They a Monetary Phenomenon? J Mon Cred Bank 37:887–906
81. Ravn M, Sola M (2004) Asymmetric Effects of Monetary Policy in the United States. Fed Res Bank St. Louis Rev 86:41–60
82. Sichel DE (1994) Inventories and the three phases of the business cycle. J Bus Econ Stat 12:269–277
83. Sims C, Zha T (2006) Were there Regime Changes in US Monetary Policy? Am Econ Rev 96:54–81
84. Sinclair T (2007) Asymmetry in the Business Cycle: A New Unobserved Components Model. George Washington University working paper, Washington
85. Stock JH, Watson MW (1989) New Indexes of Coincident and Leading Economic Indicators. NBER Macroeconomics Annual 4:351–393
86. Stock JH, Watson MW (1991) A Probability Model of the Coincident Economic Indicators. In: Lahiri K, Moore GH (eds) Leading Economic Indicators: New Approaches and Forecasting Records. Cambridge University Press, Cambridge
87. Stock JH, Watson MW (2002a) Forecasting using principal components from a large number of predictors. J Am Stat Assoc 97:1167–1179
88. Stock JH, Watson MW (2002b) Macroeconomic Forecasting Using Diffusion Indexes (with James H Stock). J Bus Econ Stat 20(2):147–162
89. Tanner M, Wong W (1987) The Calculation of Posterior Distributions by Data Augmentation. J Am Stat Assoc 82:528–550
90. Tiao GC, Tsay RS (1994) Some advances in non-linear and adaptive modeling in time-series analysis. J Forecast 13:109–131
91. Tong H (1983) Threshold models in non-linear time series analysis, Lecture Notes in Statistics, No. 21. Springer, Heidelberg
92. Turner CM, Startz R, Nelson CR (1989) A Markov Model of Heteroskedasticity, Risk, and Learning in the Stock Market. J Financ Econ 25:3–22
93. van Dijk D, Franses PH (1999) Modeling multiple regimes in the business cycle. Macroecon Dyn 3:311–340
94. Watson MW (1986) Univariate Detrending Methods with Stochastic Trends. J Monet Econ 18:49–75
95. Watson MW (2005) Comment on James D Hamilton's 'What's Real about the Business Cycle?' Fed Res Bank St. Louis Rev 87:453–458
96. Zellner A (1971) An Introduction to Bayesian Inference in Econometrics. Wiley, New York

Econometrics: Non-linear Cointegration

Juan-Carlos Escanciano[1], Alvaro Escribano[2]
[1] Department of Economics, Indiana University, Bloomington, USA
[2] Department of Economics, Universidad Carlos III de Madrid, Madrid, Spain

Article Outline

Glossary
Definition of the Subject
Introduction
Linear Measures of Memory and Linear Error Correction Models
Nonlinear Error Correction (NEC) Models
Nonlinear Cointegration
Future Directions
Bibliography

Glossary

Cointegration Cointegration is an econometric property relating time series variables. If two or more series are themselves nonstationary, but a linear combination of them is stationary, then the series are said to be cointegrated.

Short memory A time series is said to be short memory if its information decays through time. In particular, we say that a variable is short memory in mean (in distribution), if the conditional mean (distribution) of the variable at time t given the information at time $t-h$ converges to a constant (to an unconditional distribution) as h diverges to infinity. Shocks in short memory time series have transitory effects.

Extended memory A time series is said to be extended memory in mean (in distribution), if it is not short memory in mean (distribution). Shocks in extended memory time series have permanent effects.

Nonlinear cointegration If two or more series are of extended memory, but a nonlinear transformation of them is short memory, then the series are said to be nonlinearly cointegrated.

Error correction model An Error Correction Model is a dynamic model in which the rate of growth of the variables in any period is related to the previous period's gap from long-run equilibrium.

Definition of the Subject

This paper is a selective review of the literature on nonlinear cointegration and nonlinear error correction models. The concept of cointegration plays a major role in macroeconomics, finance and econometrics. It was introduced by Granger in [42] and since then, it has achieved immense popularity among econometricians and applied economists. In fact in 2003 the Royal Swedish Academy of Science gave the Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel to C. W. J. Granger for his contribution to the analysis of economic relationships based on cointegrated variables. In this paper we discuss the nonlinear extensions of the linear cointegration theory. Some authors consider nonlinear cointegration as a particular case of nonlinear error correction models. Although both concepts are related, we believe that it is useful to distinguish between them. After making this point clear, by relating linear and nonlinear error correction models, we discuss alternative measures of temporal dependence (memory) and co-dependence that are useful to characterize the usual notion of integration of order zero, $I(0)$, and cointegration in nonlinear contexts. We discuss parametric and nonparametric notions of nonlinear cointegration. Finally, we conclude pointing out several lines of research that we think are promising in nonlinear and nonstationary contexts and therefore deserve further analysis.

Introduction

Granger in [42] introduced the concept of *cointegration* in a linear context; for further development see [20,64,65,85]. The alternative ways to deal with integrated and cointegrated series are now clear only in the linear context; see for example [43,52,57,59,67,77,105].

In macroeconomic and financial applications there are many cases where *nonlinearities* have been found in nonstationary contexts and therefore, there is a need for a theoretical justification of those empirical results. To reach this goal is not an easy target since the usual difficulties analyzing nonlinear time series models within a stationary and ergodic framework are enhanced in nonstationary contexts.

The purpose of this survey on *nonlinear cointegration* is to give a selected overview on the state of the art of econometrics that simultaneously analyzes nonstationarites and nonlinearities. The structure of this paper is the following: Sect. "Linear Measures of Memory and Linear Error Correction Models" discusses linear concepts of memory and dependence, cointegrated and error correction models. Section "Nonlinear Error Correction (NEC) Models" introduces nonlinear error correction models. Section "Nonlinear Cointegration" investigates nonlinear measures of memory and dependence and nonlinear cointegration. Finally, Sect. "Future Directions" concludes and mentions some open questions for future research.

Linear Measures of Memory and Linear Error Correction Models

The time series x_t is *integrated of order d*, denoted $x_t \sim I(d)$, if $\Delta^d x_t = (1-L)^d x_t \sim I(0)$, where L is the lag operator such that $L^k x_t = x_{t-k}$ and d is an integer number. Here $I(0)$ denotes a covariance stationary short memory process with positive and bounded spectral density. We can extrapolate the concepts of integration to the *fractional case* where now d is not an integer but a real number. However, in this paper we will not cover fractional integration nor fractional cointegration; see Chapter 9.4.1 in [103] for a review.

Following the ideas of the seminal paper of [42], the most simple *definition of cointegration* could be the following; we say that two $I(1)$ series, y_t and x_t, are *cointegrated* if there is a linear combination $(1, -\beta)(y_t, x_t)'$ that is $I(0)$; $z_t = y_t - \beta x_t$ is $I(0)$, but any other linear combination, $z_t = y_t - \alpha x_t$, is $I(1)$ where $\alpha \neq \beta$. For sim-

plicity, through all this paper we will assume a bivariate system with a single cointegrating vector. Notice that the second condition of the above definition of cointegration ($z_t = y_t - \alpha x_t$, is $I(1)$ for any $\alpha \neq \beta$) is redundant in the linear context. However, it will be useful for the identification of the cointegrating vector in nonlinear cointegration, see Definition 3.

A nonparametric characterization of cointegration was introduced by [1] and [78]. Let x_t, y_t be the two $I(d)$ time series of interest, $d = 1$, and let $\gamma_{yx}(\tau, t)$ represent the *cross-covariance function* (*CCF*) of x_t, y_t, defined by $\gamma_{yx}(\tau, t) = \operatorname{cov}(y_t, x_{t-\tau})$, where we make explicit the time dependence in $\gamma_{yx}(\tau, t)$ to allow for some degree of heterogeneity in the series. Similarly, define $\gamma_x(\tau, t) = \operatorname{cov}(x_t, x_{t-\tau})$. Cointegration implies that the rates of convergence of $\gamma_{yx}(\tau, t)$ and $\gamma_x(\tau, t)$ should be the same as τ increases without bound and $\tau = o(t)$. Intuitively, under cointegration, the remote past of x_t should be as useful as the remote past of y_t in terms of the long-run linear forecast of y_t. For example, suppose $x_t, y_t \sim I(1)$ and $z_t = y_t - \beta x_t$ a sequence of independent and identically distributed (i.i.d.) random variables independent of x_t. In this case, $\gamma_{yx}(\tau, t)/\gamma_x(\tau, t) = \beta$ for all τ, t, $\tau \leqslant t$. In more general cases, z_t might have serial correlation and might not be independent of x_t and therefore, the constancy of this ratio will only take place for τ's beyond some value, see the Monte Carlo simulation results in [78]. On the other hand, in the spurious cointegration case where x_t, y_t are stochastically independent, $\lim_{\tau\to\infty} \gamma_{yx}(\tau, t)/\gamma_x(\tau, t) = 0$ for all τ, t, $\tau \leqslant t$, therefore the ratio $\gamma_{yx}(\tau, t)/\gamma_x(\tau, t)$ is consistent against this type of spurious alternative hypothesis. As we will see later on, this notion of cointegration accepts nonlinear generalizations (nonlinear cointegration).

The most simple version of *Granger's Representation Theorem*, see [20], states that two series y_t and x_t are cointegrated if and only if they have an error correction representation, see Eqs (1a)–(1c) below. Therefore, either x_t Granger-causes y_t or y_t Granger-causes x_t or both.

Let y_t and x_t be two $I(1)$ series, where x_t is a pure random walk and y_t is generated by the following linear error correction (EC) model (1a) with linear cointegration (1b),

$$\Delta y_t = \psi_1 \Delta x_t + \gamma z_{t-1} + \upsilon_t \tag{1a}$$

$$y_t = \beta x_t + z_t \tag{1b}$$

$$\Delta x_t = \varepsilon_t \tag{1c}$$

where all the random error terms (υ_t, z_t, ε_t) are $I(0)$. The errors z_t of (1b) form the error correction terms of (1a) and have usually more persistence (longer memory) than the other two random error terms (υ_t, ε_t). Therefore, in the system of Eqs. (1a)–(1c), x_t Granger-causes y_t but not the other way around.

Notice that we can write Eq. (1a), with $\phi_1 = \psi_1 - \beta$, in an equivalent way that will be very useful to introduce later on nonlinear error correction (NEC) models,

$$z_t = z_{t-1} + \phi_1 \Delta x_t + \gamma z_{t-1} + \upsilon_t \,. \tag{2}$$

Several alternative estimation procedures have been discussed in the literature to estimate the cointegrating parameter β:

i) Maximum likelihood approach of [66] and [7]. Assumes that the conditional distribution of y given x and the lagged values of x and y is *Normal* and that the bivariate data generating process (DGP) of y and x is a *VAR* of *finite autoregressive order k*, VAR(k) in error correction form. Furthermore, if the contemporaneous x-variable is *weakly exogenous*, then the partial maximum likelihood estimators is obtained by nonlinear least squares (NLS) on the error correction model obtained substituting (1b) in (1a). [98] and [38], derived the asymptotic properties of the NLS estimator of the error correction model (1a) and (1b), without the Normality assumption.
ii) *Two-step approach* of Engle and Granger, see [20]. In the first step, Eq. (1b) is estimated by ordinary least squares (OLS) to get the residuals (z). In the second step, Eq. (1a) is estimated by OLS after substituting z_{t-1} by the corresponding lagged residuals from the first step. The OLS estimator of the first step is *superconsistent but biased*, and the limiting distribution depends on nuisance parameters. However, if z_t is serially uncorrelated and x_t is strictly exogenous, then the OLS estimator in (1b) coincides with the fully modified estimator and therefore it is asymptotically efficient, see [38].
iii) *Fully modified OLS*, FM-OLS. This is a 2-step procedure developed by [79,80,86], and [81]. In the *first step*, Eq. (1b) is estimated by OLS. In the second step, semiparametric corrections are made for the *serial correlation* of the residuals z_t and for the *endogeneity* of the x-regressors. Under general conditions the fully modified estimator, is *asymptotically efficient*. The small sample behavior of these estimators was analyzed by Monte Carlo simulations by [53,56,57,58,71,86].
iv) *Fully modified instrumental variable* estimator, FM-IV, of [71,86]. [78] showed that their nonparametric notion of cointegration has an instrumental variable (IV) interpretation if the instruments are the lagged

values of x. Furthermore, they showed that choosing those instruments has an extra advantage; we do not need to make the usual two corrections (endogeneity and serial correlation) to obtain a fully modified estimator. This particular IV-estimator has important advantages (bias reductions) over OLS in small samples.

v) Recently [89] also studied the asymptotic properties of instrumental variables estimators (IV) in a fractional cointegration context, as in [78]. They propose to use IV estimates based on single equations estimation like (1b) employing exclusion and normalization restrictions, without correcting for the serial correlation of z_t.

vi) [100] and [91], suggested a parametric correction for the endogeneity of the regressor (x_t) when estimating (1b) by OLS. The idea, based on the work of [97] about testing for causality, is to include additional future and past values of the Δx_t in Eq. (1b) when estimating it by OLS.

vii) [87] proposed to add integral error correction terms (lagged values of the EC terms), to the procedure described in *vi)* in order to parsimoniously correct for serial correlation.

[63], using Monte Carlo simulations, compares some of these parametric and semi parametric estimators of the cointegrating vector. In the context of normally distributed errors, [63] recommends to model explicitly the dynamics instead of using nonparametric corrections based on fully modified estimators.

Nonlinear Error Correction (NEC) Models

There are interesting macroeconomic applications where nonlinearities have been found in nonstationary contexts. The first example of a nonlinear error correction (NEC) model is the UK money demand from 1878 to 1970 of [25,27]. Later on [61] used this nonlinear error correction strategy in their money demand estimation as an improvement over the usual linear money demands equations suggested by [37,60,76].

The variables of the usual money demand are: $m =$ log money stock (millions), $y =$ log real net national product Y, $p =$ log of the price deflator of Y, $rs =$ log of short term interest rate, $rl =$ log of long-term interest rate, and $RS =$ short term interest rate (without logs). Let V be the velocity of circulation of money, a version of the *quantity theory of money* says that $MV(RS) = PY$ or in logs $m + v(RS) = p + y$. Rearranging terms we can write $(m - p - y) = -v(RS)$ as a *long run money demand*.

[27] applied the *2-step approach* of [20] obtaining the following results:

1st Step:

$$\left(m - p - y\right)_t = -0.31 - 7RS_t + \hat{u}_t \tag{3a}$$

where $\hat{u}_t$ are the residuals from 1878 to 2000 of the cointegrating relationship estimated by the super-consistent ordinary least squares (OLS) estimator. The inverse of the log of velocity of circulation of money, $(m - p - y) = \log\left(\frac{M}{PY}\right) = -v(RS)$, is $I(1)$ and the short run interest rate (RS) is also $I(1)$. Therefore, since the error term $\hat{u}_t$ is *stable and significant* it is $I(0)$, see conditions (e) and (f) of Theorem 1 below. Equation (3a), or (3b), is the first example of *nonlinear cointegration* given by;

$$\frac{M}{PY} = \exp(-0.31 - 7RS + \hat{u})\,. \tag{3b}$$

Similar nonlinear cointegrating relationships based on long run money demand equations are recently estimated by [3].

[22] and [27] showed that, even if OLS might not be a consistent estimator (see [95]) when the errors of (3a), (3b) are nonlinear, the OLS estimates of (3a) and the NLS estimates of (4) in 1-step are very similar.

2nd Step:

$$\begin{aligned}(1-L)\left(m-p\right)_t &= 0.45\,(1-L)\left(m-p\right)_{t-1} \\ &- (1-L)^2\left(m-p\right)_{t-2} - 0.60\,(1-L)\,p_t \\ &+ 0.39\,(1-L)\,p_{t-1} - 0.021\,(1-L)\,rs_t \\ &- 0.062\left(1-L^2\right)rl_t - 2.55\,(\hat{u}_{t-1} - 0.2)\,\hat{u}_{t-1}^2 \\ &+ 0.005 + 3.7\,(D1 + D3) + \hat{\varepsilon}_t\end{aligned} \tag{4}$$

where D_1 and D_3 are dummy variables for the two world wars. The second nonlinear characteristic of model (4), apart from the nonlinear cointegration relationship, comes from the fact that the $\hat{u}_{t-1}$ term enters in a *cubic polynomial form* as a particular nonlinear error correction (NEC) model, see also Sect. 3.3 in [17,18,95] discuss the inconsistencies derived from the 2-step approach OLS estimator in the context of nonlinear smooth transition error correction model. However, Monte Carlo simulations should be done to identify the type of nonlinearities that create series biases and inconsistencies using the 2-step estimation of NEC models.

A Nonlinear Version of Granger Representation Theorem

To justify this type of nonlinear models, we need to generalize the linear notions of temporal memory based on

the linear ARIMA concepts of integration, usually $I(1)$ and $I(0)$, to nonlinear measures of dependence. Several generalizations have been proposed in the literature, as we will see later on. Our first definition is motivated from asymptotic theory, more concretely from functional central limit theorems (FCLT). See Subsect. "A Nonlinear Version of Granger Representation Theorem" for alternative definitions.

FCLT-Based Definition of I(0): *A sequence $\{m_t\}$ is $I(0)$ if the "high level" condition that m_t verifies a FCLT is satisfied, i. e.*

$$T^{-1/2}\sum_{t=1}^{[Tr]} m_t \xrightarrow{d} B(r)$$

where $B(r)$ is a Brownian motion, see [73,74,99].

Definition 1 (Strong mixing) Let $\{v_t\}$ be a sequence of random variables. Let $\Im_s^t \equiv \sigma(v_s, \ldots, v_t)$ be the generated sigma-algebra. Define the α-mixing coefficients

$$\alpha_m \equiv \sup_t \sup_{\left\{F\in\Im_{-\infty}^t, G\in\Im_{t+m}^{\infty}\right\}} |\mathrm{P}(G\cap F) - \mathrm{P}(G)\,\mathrm{P}(F)|$$

The process $\{v_t\}$ is said to be strong mixing (also α-mixing) if $\alpha_m \to 0$ as $m \to \infty$. If $\alpha_m \leqslant m^{-a}$ we say that $\{v_t\}$ is strong mixing of size $-a$.

Definition 2 (NED) Let $\{w_t\}$ be a sequence of random variables with $E\left\{w_t^2\right\} < \infty$ for all t. It is said that $\{w_t\}$ is NED on the underlying sequence $\{v_t\}$ of size $-a$ if $\phi(n)$ is of size $-a$, where $\phi(n)$ given by

$$\sup \left\| w_t - E_{t-n}^{t+n}(w_t)\right\|_2 \equiv \phi(n)$$

where $E_{t-n}^{t+n}(w_t) = E(w_t|v_{t-n}, \ldots, v_{t+n})$ and $\|\cdot\|_2$ is the L_2 norm of a random vector, defined as $E^{1/2}|\cdot|^2$ where $|\cdot|$ denotes the Euclidean norm.

Weak-Dependence-Based Definition of I(0): *A sequence is $I(0)$ if it is NED on an underlying α-mixing sequence $\{v_t\}$ but the sequence $\{x_t\}$ given by $x_t = \sum_{s=1}^t w_s$ is not NED on$\{v_t\}$. In this case, we will say that x_t is $I(1)$.*

Notice that if x_t is $I(1)$ then Δx_t is $I(0)$. This definition excludes $I(-1)$ series as $I(0)$, like $z_t = e_t - e_{t-1}$ for α-mixing sequences e_t, since in this case $\sum_{s=1}^t z_s$ is α-mixing.

Definition 3 Two $I(1)$ sequences $\{y_t\}$ and $\{x_t\}$ are (linearly) cointegrated with cointegrating vector $[1, -\beta]'$, if $y_t - \beta x_t$ is NED on a particular α-mixing sequence but $y_t - \delta_{12}x_t$ is not NED for $\delta_{12} \neq \beta$.

Theorem 1 (Granger's Representation Theorem, see [31]) *Consider the nonlinear correction model (NEC) for the (2×1) vector $X_t = (y_t, x_t)'$, given by*

$$\Delta X_t = \Psi 1 \Delta X_{t-1} + F(X_{t-1}) + \upsilon_t \quad . \tag{5}$$

Assume that:

(a) $\upsilon_t = (\upsilon_{y_t}, \upsilon_{x_t})'$ is α-mixing of size $-s/(s-2)$ for $s > 2$
(b) $\sum_t v_t$ is not NED on α-mixing sequence
(c) $E\|v_t\|^2 \leqslant \Delta_v$
(d) $F(X_{t-1}) = J(Z_{t-1})$, where $Zt \equiv y_t - \beta x_t$ and $J(\cdot)$ is a continuously differentiable function, which satisfies a generalized Lipschitz condition of Lemma 2 of [31].
(e) Let $SR(\Psi 1) < 1$, where SR(M) is the spectral radius of the matrix M, and
(f) for some fixed $\delta \in (0,1)$

$$SR\begin{pmatrix} \Psi 1 & \nabla_z J(Z) \\ \beta'\Psi 1 & I_r + \beta'\nabla_z J(z) \end{pmatrix} \leqslant 1-\delta\,.$$

The above conditions ensure that;

(i) ΔX_t and Z_t are simultaneously NED on the α-mixing sequence (v_t, u_t), where $u_t = v_{y,t} - \beta' v_{x,t}$; and
(ii) X_t is $I(1)$.

This theorem gives sufficient conditions for cointegrated variables to be generated by a nonlinear error correction model.

Single-Equation Parametric NEC Models with Linear Cointegration

Consider the following NEC with linear cointegration

$$\Delta y_t = \psi_1 \Delta x_t + f(z_{t-1};\gamma) + \upsilon_t$$
$$y_t = \beta x_t + z_t\,.$$

As we said in the previous section, it is not difficult to generalize this model to include other variables, lags and cointegrating relations. Consider two independent α-mixing sequences $\{a_t\}$ and $\{\upsilon_t\}$ with a zero mean. Then the following three equations represent the DGP,

$$x_t = x_{t-1} + a_t \tag{6a}$$

$$z_t = z_{t-1} + \phi_1 \Delta x_t + f(z_{t-1},\gamma) + v_t \tag{6b}$$

$$y_t = \beta x_t + z_t \tag{6c}$$

where the nonlinear function $f(z_{t-1},\gamma)$ form the nonlinear error correction term and $\beta \neq 0$. Notice that x_t is $I(1)$ by construction from (6a). Notice the similarity between Eqs. (6b) and the linear error correction of Eq. (2).

If we can ensure that z_t is NED then y_t is also $I(1)$ and linearly cointegrated with x_t, where the cointegration relationship, $y_t - \beta x_t$, is linear. If we apply the difference operator to (6c) and substitute in (6b) we obtain (7),

$$\Delta y_t = (\beta + \phi_1)\,\Delta x_t + f(z_{t-1}, \gamma) + v_t \tag{7}$$

which is a nonlinear error correction model with linear cointegration, with $\psi_1 = \beta + \phi_1$. For the sake of simplicity, and without loss of generality, we impose a *common factor restriction* so that $\phi_1 = 0$ on (7) obtaining,

$$z_t = z_{t-1} + f(z_{t-1}, \gamma) + v_t \tag{8}$$

and then ψ_1 equals β, the cointegration parameter, and therefore (8) is a nonlinear extension of the Dickey–Fuller equation used in unit root testing. The errors of the cointegration relation are given by $z_{t-1} = y_{t-1} - \beta x_{t-1}$, and the OLS residuals are given by $\hat{z}_{t-1} = y_{t-1} - \hat{\beta} x_{t-1}$, where $\hat{\beta}$ is the value of β estimated in the OLS regression (6c). Substituting z_t by $\hat{z}_{t-1}$ in (8) we obtain a nonlinear version of Engle and Granger's cointegration test (cf. [20]).

Differentiating (8) with respect to z_{t-1} we obtain

$$\frac{\mathrm{d}}{\mathrm{d}z_{t-1}} z_t = 1 + \frac{\mathrm{d}}{\mathrm{d}z_{t-1}} f(z_{t-1}, \gamma)$$

and therefore, our boundedness condition (see assumptions (e) and (f) of Theorem 1) is $-1 < \frac{\mathrm{d}}{\mathrm{d}z_{t-1}} z_t < 1$, or $-2 < \frac{\mathrm{d}f(z_{t-1},\gamma)}{\mathrm{d}z_{t-1}} < 0$ (models (6b), (7) and (8) are error correcting), which is sufficient to ensure that the series z_t is near epoch dependent (NED) and therefore y_t and x_t are cointegrated, see [26,27] and [31].

We discuss now few alternative nonlinear error correction (or equilibrium correction) functions $f(.)$ that could generate the series z_t from the system (6a) to (6c).

NEC Model 1: Arctan, [32]

$$f(z, \delta_1, \delta_2, \gamma_2) = -\gamma_2 \arctan(\delta_1 z + \delta_2) \quad \text{for } \gamma_2 > 0\,.$$

NEC Model 2: Rational Polynomial, [27]

$$f(z, \delta_1, \delta_2, \delta_3, \delta_4, \gamma_2) = -\gamma_2 \left((z + \delta_1)^3 + \delta_2\right) / \left((z + \delta_3)^2 + \delta_4\right) \quad \text{for } \gamma_2 > 0\,.$$

In the first two models, the derivatives are in the desired region (satisfy assumptions (e) and (f)) for appropriate values of some of the parameters but not for all. However, within the class of rational polynomials the model considered can satisfy the condition on the absolute value of the derivative, see [27] and [30,32]. Other empirical examples of nonlinear error correction models are [11,22,29,33,49,61,72].

An important body of the literature has focused on *threshold models*, see [5,6,39,48,54,75,94], among others.

NEC Model 3: Switching Exponential, [30,32]

$$f(z, \delta_1, \delta_2, \delta_3, \delta_4, \gamma_2) = \gamma_2 (\exp(-\delta_1 z) - \delta_2)\, I_{\{z \geqslant 0\}} + \gamma_2 (\delta_4 - \exp(\delta_3 z))\, I_{\{z<0\}}\,,$$

where $I_{\{S\}}$ is the characteristic function of the set S, $\gamma_2 > 0, \delta_1 > 0$ and δ_3.

NEC Model 4: Regime Switching Error Correction Models, [92]

$$f(z, \delta_1, \delta_2, \gamma_2) = \sum_{s=1}^{3} 1(z \in R_s)\gamma_s z\,,$$

where $1(.)$ is the indicator function selecting the three regimes, $R_1 = (-\infty, c_1]$, $R_2 = (c_1, c_2]$ and $R_3 = (c_2, \infty)$.

NEC Model 5: Random Regime Switching Error Correction Models, [6] and [92]

$$f(z, \delta_1, \delta_2, \gamma_2) = \sum_{s=1}^{3} 1(z + \eta \in R_s)\gamma_s z\,,$$

where $1(.)$ is the indicator function selecting the three regimes, $R_1 = (-\infty, c_1]$, $R_2 = (c_1, c_2]$ and $R_3 = (c_2, \infty)$.

Another important literature is related to *smooth transition error correction models*, see [50,90,92].

NEC Model 6: Smooth Transition Error Correction Models, [102] and [92]

$$f(z, \delta_1, \delta_2, \gamma_2) = \sum_{s=1}^{3} h(z)\gamma_s z$$

$$h(z) = \left\{ \begin{array}{ll} 1 - L_1(z), & s = 1 \\ L_1(z) - L_2(z), & s = 2 \\ L_3(z), & s = 3 \end{array} \right\}$$

$$\text{where} \quad L_s(z) = (1 + \exp\{-\gamma(z - c_s)\} \quad \text{for} \quad \gamma \succ 0 \quad \text{and} \quad s = 1, 2\,.$$

Notice, that many smooth transition error correction models allow the nonlinear error correction function to affect all the parameters of the model, and not only the error correction term. However, in this paper we do not discuss them since they belong to a more general class of time varying models which is out of the context of this survey, see for example [50,101,102].

Nonparametric NEC Models with Linear Cointegration

[25,27] applied the semi-parametric *smoothing splines* estimation procedure of [21,96,104] to the estimation of the unknown nonlinear error correction function of Eq. (4). They found that in the long run money demand of the UK, see Eq. (3a), there are either multiple equilibria, or a threshold error correction with two attraction points (two equilibria); one negative and equal to -0.05 and one positive and equal to 0.2. They suggest estimating those unknown thresholds using a cubic polynomial parametric functional form, see Eq. (4). Notice that the corresponding cubic polynomial error correction term, $-2.55\,(\hat{u}_{t-1} - 0.2)\,\hat{u}_{t-1}^2$, identifies perfectly one of the thresholds, the one that is equal to 0.2. The second threshold could be obtained from the roots of the polynomial. Other empirical examples of threshold error correction models are [5,10,48,54]. In fact [10] used a similar nonparametric approach to estimate the nonlinear error correction function, but instead of using smoothing splines they used the Nadaraya–Watson kernel estimator discussed in [55].

Nonlinear Cointegration

In the recent years several proposals have been considered to extend linear cointegration and linear error correction of Granger [42] to a nonlinear framework. One possibility is to allow for a NEC model in the Granger's representation. We have discussed such an approach in the previous section. Alternatively, one may consider a nonlinear cointegration relation.

Despite the fact that many macroeconomic and financial time series dynamics are nonlinear, there are still today relatively few useful analytical tools capable of assessing the dependence and persistence behavior of nonlinear time series appropriately (cf. [50]). This problem is even more accentuated by the fact that traditional measures of dependence, such as autocorrelations and periodograms, may be inappropriate when the underlying time series is nonlinear and/or non-Gaussian. Then, it is generally accepted that new measures of dependence have to be defined in order to develop a new concept of nonlinear cointegration. We have already discussed measures based on FCLT and on NED concepts. We shall explore several alternative measures in this section. All the measures considered can be grouped in measures of conditional mean dependence or in distributional dependence. Higher order conditional moments, other than the mean, can of course be considered. In any case, we shall use the general terminology *extended memory* and *short memory* to indicate a nonlinear persistence and non-persistence process, respectively (cf. [44]).

Once a concept of nonlinear persistence is introduced, a general definition of nonlinear cointegration is as follows. We say that two "extended memory" series y_t and x_t are nonlinear cointegrated if there exist a function f such that $z_t = f(y_t, x_t)$ is short memory. This definition is more appropriate when dealing with distributional persistence, and it is perhaps too general to be fully operative. Identification problems arise in this general context, as noted by many authors, so one should restrict the class of functions f to avoid such identification problems. [46] considered functions of the form $z_t = g(y_t) - h(x_t)$, and estimate g and h nonparametrically by means of the Alternating Conditional Expectations (ACE) algorithm. See also [40] for a related approach. It is still an open problem the theoretical justification of these nonparametric estimation procedures.

A less ambitious approach is to consider transformations of the form $z_t = y_t - f(x_t)$. This framework is especially convenient with conditional mean persistence measures. We review the existing measures in the next section.

Nonlinear Measures of Memory

As already discussed by [46], a generalization of linear cointegration to a nonlinear set-up goes through proper extensions of the linear concepts of $I(0)$ and $I(1)$. We introduce in this section alternative definitions of nonlinear $I(0)$ and $I(1)$ processes. We first focus on conditional mean persistence, we shall discuss distributional dependence at the end of this section. Define the conditional mean function $E(y_{t+h}|I_t)$, where $I_t = (x_t, x_{t-1}, \dots)$ is the conditioning set at time t. [44] defines the Short Memory in Mean (SMM) and Extended Memory in Mean (EMM) concepts as follows.

Definition 4 (SMM and EMM) $\{y_t\}$ is said to be SMM if for all t, $M(t,h) = E(y_{t+h}|I_t)$, $h > 0$, tends to a constant μ as h becomes large. More precisely, $E\,|M(t,h) - \mu|^2 < c(h)$, where $c(h) \equiv c(h,t)$ is some positive sequence that tends to zero as h increases to infinity, for all t. If $\{y_t\}$ does not satisfy the previous condition is called EMM.

Note that to be mathematically precise in Definition 4, we should specify that $\{y_t\}$ is SMM or EMM with respect to $\{x_t\}$. Referring to this definition, [44] considered the case $x_t = y_t$. [46] replaced the name of EMM by long memory in mean, and [40] denoted EMM and SMM by nonlinear integrated (NLI) and nonlinear integrated of order zero (NLI(0)), respectively. As noted by [44] the concepts

of SMM and EMM are related to a kind of "mixing in mean" property, more precisely to the concept of mixingale, see [16].

[23] introduced the pairwise equivalent measures of the previous concepts, which, although weaker, are more operative because they only involve finite-dimensional random variables.

Definition 5 (PSMM and PEMM) $\{y_t\}$ is said to be Pairwise SMM (PSMM) if for all t, $m(t,h) = E(y_{t+h}|x_t)$, $h > 0$, tends to a constant μ as h becomes large. More precisely, $E|m(t,h) - \mu|^2 < c(h)$, where $c(h) \equiv c(h,t)$ is some positive sequence that tends to zero as h increases to infinity, for all t. If $\{y_t\}$ does not satisfy the previous condition is called Pairwise EMM (PEMM).

From the previous definitions and the law of iterated expectations, we easily observe that a process SMM is PSMM. The reciprocal is false. There exist processes which are PSMM but not SMM, although they are rare in practice.

We now discuss generalizations of the usual autocovariances and crosscovariances to a nonlinear framework. These generalizations were introduced by [23]. It is well-known that in the presence of nonlinearity (or non-Gaussianity) the autocovariances do not characterize the dependence and the practitioner needs more reliable measures such as the pairwise regression functions $m(t,h)$. In general, inference on these functions involves nonparametric estimation with bandwidth choices, hampering their application to practical situations. By a measure-theoretic argument, the regression function $m(t,h)$ can be characterized by the integrated regression function $\gamma_{t,h}(x)$ given by

$$\begin{aligned}\gamma_{t,h}(x) &= E\left[(y_{t+h} - \mu_t)1(x_t \leqslant x)\right] \\ &= E\left[m(t,h)1(x_t \leqslant x)\right] ,\end{aligned}$$

where the second equality follows by the law of iterated expectations. The measures $\gamma_{t,h}(x)$ are called the Integrated Pairwise Regression Functions (IPRF), see [24]. Extensions to other weight functions different from the indicator weight $1(x_t \leqslant x)$ are possible. The integrated measures of dependence $\gamma_{t,h}(x)$ are useful for testing interesting hypotheses in a nonlinear time series framework and, unlike $m(t,h)$, they do not need of smoothing estimation and are easily estimated by the sample analogue. Moreover, they characterize the pairwise versions of the concepts introduced by [44], making these concepts more operative. First we need a definition, a norm $\|\|$ is nondecreasing if for all f and g with $|f(x)| \leqslant |g(x)|$ for all x, it holds that $\|f\| \leqslant \|g\|$. Associated to the norm $\|\|$, we define the distance $d(f,g) = \|f - g\|$. Usual nondecreasing norms are the L_2 norm and the supremum norm.

Definition 6 (PSMM$_d$ and PEMM$_d$) $\{y_t\}$ is said to be Pairwise SMM relative to d (PSMM$_d$) if for all t, $\left\|\gamma_{t,h}(x)\right\|$, $h > 0$, tends to zero as h becomes large for any t. More precisely, $\left\|\gamma_{t,h}(x)\right\| < c(h)$, where $c(h) \equiv c(h,t)$ is some positive sequence that tends to zero as h increases to infinity for all t. If $\{y_t\}$ does not satisfy the previous condition is called Pairwise EMM relative to d (PEMM$_d$).

Theorem 2 (Relationship between PSMM and PSMM$_d$ [23]) *If the norm $\|\|$ is non-decreasing and $E\left\{y_t^2\right\} < \infty$ for all t, then $\{y_t\}$ is PSMM if and only if $\{y_t\}$ is PSMM$_d$.*

Based on these concepts we define nonlinear cointegration as follows. We say that two PEMM$_d$ series y_t and x_t are nonlinear cointegrated if there exist a function f such that $z_t = y_t - f(x_t)$ is PSMM$_d$.

In analogy with the linear world and based on the results in [1,78], a possible nonparametric characterization of nonlinear cointegration can be based on the rates of convergence of $\gamma_{t,h}(x)$ and $\gamma^x_{t,h}(x) = E\left[(x_{t+h} - \mu_t)1(x_t \leqslant x)\right]$ as h diverges to infinity. Intuitively, under cointegration, the remote past of x_t should be as useful as the remote past of y_t in long-run non-linearly forecasting y_t.

Similarly, we can define distributional measures of persistence and nonlinear cointegration. [45] defines a persistent process in distribution using the bivariate and marginal densities at different lags. [41] considered parametric and nonparametric methods for studying serial distributional dependence. In the nonparametric case [45] consider series expansions estimators for the nonlinear canonical analysis of the series. These authors apply their results to study the dynamics of the inter-trade durations of the Alcatel-stock on the Paris Borse and find evidence of nonlinear strong persistence in distribution.

Regarding distributional dependence, we formalize a definition given in [45]. Let $f_{t,h}(y,x)$, $k_{t+h}(y)$ and $g_t(x)$ be, respectively, the bivariate and marginal densities of y_{t+h} and x_t. To define persistence in distribution one can use the Hellinger distance

$$H_{t,h} = \iint \left|f^{1/2}_{t,h}(y,x) - k^{1/2}_{t+h}(y)g^{1/2}_t(x)\right|^2 \mathrm{d}y\mathrm{d}x ,$$

and define Pairwise Short Memory in Distribution (PSMD) according to the decay of $H_{t,h}$ to zero as h diverges to infinity. Alternative definitions can be given in terms of other divergence measures or distances, see [51] and references therein. This approach is explored in [1], who define nonlinear cointegration using mutual information measures.

Persistence in distribution is related to mixing concepts. In fact, uniformly in t, $H_{t,h} \leqslant 2\alpha(h)$, where $\alpha(h)$ is certain α-mixing coefficient, see [23] for details.

The aforementioned measures of nonlinear distributional dependence need of smoothing estimation, e. g. kernel estimators. Similarly to the case of conditional mean measures, we can avoid smoothing by means of the integrated measures of dependence

$$\begin{aligned}\eta_{t,h}(y,x) &= \operatorname{cov}(1(y_{t+h} \leqslant y), 1(x_t \leqslant x)) \\ &= F_{t,h}(y,x) - K_{t+h}(y)G_t(x)\,,\end{aligned}$$

where $F_{t,h}(y,x)$, $K_{t+h}(y)$ and $G_t(x)$ are, respectively, the bivariate and marginal cumulative distribution functions (cdf) of y_{t+h} and x_t. The measures $\eta_{t,h}(y,x)$ can bt estimated at different lags by using the sample analogue, i. e. the empirical distribution functions. Similar definitions to Definition 6 can be given for distributional persistence based on $\eta_{t,h}(y,x)$. See [23] for further generalizations and definitions. Definitions of nonlinear cointegration can be formulated along the lines in [1]. For instance, we can say that two persistent (in distribution) series y_t and x_t are nonlinear cointegrated (in distribution) if

$$\lim_{\tau\to\infty} \left\| \frac{\eta_{t,\tau}(\cdot)}{\eta^x_{t,\tau}(\cdot)} - \beta \right\| = 0$$

for all τ, t, $\tau \leqslant t$, where $\eta^x_{t,h}(y,x)$ is defined as $\eta_{t,h}(y,x)$ but replacing y_t by x_t there, β is a real number and $\|\|$ a suitable norm.

Integration and Cointegration Based on Order Statistics

[9,36,47] have considered rank based unit roots test to avoid the extreme sensitivity of usual test like the Dickey–Fuller type test to presence of nonlinearities and outliers, see [19] for an overview of the problems of unit root tests in general contexts. [2] suggested a range unit root test (RUR) based on the first differences of the ranges of a series. The range is the difference between the maximum and the minimum taken by a series at any given point in time. Therefore, the difference of the ranges is a measure of records. Counting the number of new records is an interesting way of distinguishing between stationary and nonstationary series since the frequency of new records vanishes faster for stationary series than for series containing unit roots. They have shown that this RUR test is robust to monotonic transformations, distributions of the errors and many structural breaks and additive outliers.

[8] suggests using the differences between the sequences of ranks. If there is no cointegration the sequence of ranks tends to diverge, while under cointegration they evolve similarly. [34] consider a record counting cointegration test (RCC) based on the synchronicity of the jumps between cointegrated series. They suggest a test statistic based on counting the number of jumps (records) that simultaneously occur in both series. Certainly, those series that are cointegrated have a much larger number of simultaneous jumps in the ranges of the series. They show that the cointegration test based on RCC is robust to monotonic nonlinearities and many structural breaks and does not require a prior estimation of the nonlinear or linear cointegrating function. There is a large literature on the effects of structural breaks and outliers on unit root and on cointegration testing but it is out of the scope of this paper, see for example the references in [62].

Another cointegration test robust to nonlinearities and structural breaks is based on induced order statistics. In particular [35] consider that two series y_t and x_t are cointegrated (either linear or nonlinear) if the corresponding induced order series are plotted around the 45° line. Their test-statistic compares the two induced order series by comparing their corresponding empirical distributions, the empirical distribution of y_t and the empirical distribution of y_t induced by x_t, using the Kolmogorov–Smirnov type statistic.

Parametric Nonlinear Cointegration

As previously discussed, an important body of research has focused on nonlinear cointegration relations. The model used being a nonlinear cointegration regression or a nonlinear regression model with integrated regressors. References on this line of research include [12,13,82,84]. [93] study smooth transition regressions with integrated regressors. An example considered by these authors is the nonlinear extension of (1b)

$$y_t = \alpha x_t + \delta x_t g(x_t - c, \gamma) + v_t\,,$$

where $g(x-c,\gamma) = -1/(1+\mathrm{e}^{-\gamma(x-c)})$ is the logistic function. Under the restriction $\delta = 0$ the latter model reduces to the linear model in (1b), see [14] for linearity tests under this framework. Other examples of parametric nonlinear cointegration are given by Eqs. (3a), (3b) and by the exponential model of [68], see also the references given in [19]. Those models are a particular case of the more general *nonlinear parametric cointegration* model of the form,

$$y_t = f(x_t, \beta) + v_t\,,$$

where x_t is $px1$ vector of $I(1)$ regressors, v_t is zero-mean stationary error term, and $f(x_t,\beta)$ an smooth function of

the process x_t, known up to the finite-dimensional parameter vector β.

At least, there are two different asymptotic justifications within these nonlinear cointegration models. In the classical asymptotic theory (e. g. [82,84]) all the existing literature has been confined to the case $p = 1$, although some extensions to single-index type regressions have been studied, see [83]. The main reason for the restriction to the univariate case is that commonly used asymptotic techniques are not appropriate for the case $p > 1$, e. g. asymptotics based on local times are not available for $p > 2$. Intrinsic to this problem is the non-recurrent property of the p-variate Brownian motion when $p > 2$. On the other hand, the triangular array asymptotics used in [93] allow for a general $p > 0$.

In the classical asymptotic theory, the properties of estimators of β, e. g. the nonlinear least squares estimator (NLSE), depend on the specific class of functions where $f(x_t, \beta)$ belongs. Commonly used classes are integrable functions, asymptotic homogeneous functions or exponential functions, see [82]. The rate of convergence of the NLSE is class-specific and, in some cases, involves random scaling. In the triangular array asymptotic theory of [93] the distribution theory of estimators of β, e. g. rates of convergence, does not depend on the specific class of functions.

Several authors exploit the previous asymptotic results on β to develop tests of nonlinear cointegration in this parametric framework. In [15] the so-called KPSS test is applied to the parametric residuals. Since the resulting limiting distribution depends on nuisance parameters, these authors implement the test with the assistance of a subsampling procedure as a smoothing device.

Nonparametric Nonlinear Cointegration

Nonparametric estimates of nonlinear cointegration relations were already computed by [46], but it has not been until the recent works by [69,70,88] that a nonparametric estimation theory for nonstationary processes has been developed. [88] considered a theory based on local time arguments, whereas [69,70] used the theory of null recurrent Markov processes. A comparison of both methodologies is discussed in [4], where near-integrated nonparametric asymptotics are studied.

More specifically, these authors estimate the transfer function $f(x_t)$ in the nonlinear regression model

$$y_t = f(x_t) + v_t \,,$$

where the series y_t and x_t are univariate observed nonstationary processes and v_t is a non-observed stationary process. [70] study the nonparametric kernel estimation of $f(x_t)$ as

$$\hat{f}(x) = \frac{\sum_{t=0}^{n} y_t K_{x,h}(x_t)}{\sum_{t=0}^{n} K_{x,h}(x_t)} \,,$$

where $K_{x,h}(x_t) = h^{-1}K((y-x)/h)$, K is a kernel function and h is a bandwidth parameter. These authors investigate the asymptotic theory for $\hat{f}(x)$ under some regularity conditions and different assumptions on the dependence relation between x_t and v_t. Especially convenient for the nonlinear cointegration framework are those assumptions that allow for dependence between x_t and v_t. The family of nonstationary processes considered by these authors is the class of the so-called β-null recurrent Markov processes satisfying a restriction on the tail distribution of the recurrence time. The class is large enough to contain the random walk, unit-root processes as well as other nonlinear nonstationary processes. It is shown that the nonparametric estimation theory is different to that in the stationary case, with slower rates of convergences, as expected. This new nonparametric asymptotic theory opens the door for future developments in inferences in nonlinear cointegration models.

Future Directions

This chapter has provided a selected overview of the available nonlinear extensions of this concept. While in the linear set-up there exists a complete theory and set of tools for studying the cointegration problem, it has been made clear that a nonlinear version of this theory possesses nontrivial challenges. A first natural nonlinear extension is to allow for a NEC model but still a linear cointegration regression. On the other hand, one can consider a nonlinear regression cointegration equation. It has been recognized that an extension of the concept of linear cointegration to a nonlinear set-up needs of appropriate extensions of the linear concepts of $I(0)$ and $I(1)$ to nonlinear time series (cf. [46]). Several extensions have been provided and discussed. We recommend operative integrated measures of dependence, since they are simple to estimate and avoid smoothing of the data, which can be a challenging problem when dealing with nonstationary variables (cf. [69]). An important line of future research is the development of inferential procedures for nonlinear cointegration based on these new integrated measures. This is currently investigated by the authors.

There is a large evidence of empirical applications in economics and finance where nonlinearities are found in nonstationary contexts. However, given the difficulty of

the theory involved, only few papers provide a sound justification of the empirical use of cointegration regressions (nonlinear cointegration) in nonlinear frameworks. The difficulties analyzing nonlinear time series models within a stationary and ergodic frameworks are substantially enhanced in nonstationary contexts. In particular, the classical asymptotic theory for nonlinear transformations of nonstationary variables becomes case-dependent (i. e. depends on the specific class of functions), and the available results are confined to the univariate case. A challenging and important line of research deals with the extension of this theory to multivariate frameworks.

Recently, an important step towards the development of a nonlinear cointegration theory has been accomplished by the nonparametric estimation theory of [69,70]. The application of this theory to inference in nonlinear cointegrated models is not fully explored yet. Residual-based tests for testing nonlinear cointegration, such as the so-called KPSS test (cf. [73]), can be constructed using nonparametric residuals. Moreover, model specification tests for nonlinear parametric cointegration can be based on the comparison between parametric and nonparametric fits. Finally, in a recent unpublished lecture, Clive Granger suggested to extend the concept of cointegration to quantiles. These are promising lines of future research which deserve serious attention in the economics literature.

Bibliography

Primary Literature

1. Aparicio F, Escribano A (1999) Information-theoretic analysis of serial correlation and cointegration. Stud Nonlinear Dyn Econ 3:119–140
2. Aparicio F, Escribano A, Sipols AE (2006) Range unit root (RUR) tests: Robust against nonlinearities, error distributions, structural breaks and outliers. J Time Ser Anal 27:545–576
3. Bae Y, de Jong RM (2005) Money demand function estimation by nonlinear cointegration. Working Paper, Department of Economics, Ohio State University
4. Bandi FM (2004) On persistente and nonparametric estimation (with an application to stock return predictability). Unpublished manuscript
5. Balke NS, Fomby TB (1997) Threshold cointegration. Int Econ Rev 38:627–645
6. Bec F, Rahbek A (2004) Vector equilibrium correction models with nonlinear discontinuous adjustments. Econ J 7:628–651
7. Boswijk HP (1995) Efficient inference on cointegration parameters in structural error correction models. J Econ 69:113–158
8. Breitung J (2001) Rank tests for nonlinear cointegration. J Bus Econ Stat 19:331–340
9. Breitung J, Goriéroux C (1997) Rank tests for unit roots. J Econ 81:7–27
10. Breitung J, Wulff C (2001) Nonlinear error correction and the efficient market hypothesis: The case of german dual-class shares. Ger Econ Rev 2:419–434
11. Burgess SM (1992) Nonlinear dynamics in a structural model of employment. J Appl Econ 7:101–118
12. Chang Y, Park JY (2003) Index models with integrated time series. J Econ 114:73–16
13. Chang Y, Park JY, Phillips PCB (2001) Nonlinear econometric models with cointegrated and deterministically trending regressors. Econ J 4:1–36
14. Choi I, Saikkonen P (2004) Testing linearity in cointegrating smooth transition regressions. Econ J 7:341–365
15. Choi I, Saikkonen P (2005) Tests for nonlinear cointegration. Unpublished manuscript
16. Davidson J (1994) Stochastic limit theory. Oxford U. P., New York
17. de Jong RM (2001) Nonlinear estimation using estimated cointegrating relations. J Econ 101:109–122
18. de Jong RM (2002) Nonlinear minimization estimators in the presence of cointegrating relations. J Econ 110:241–259
19. Dufrénot G, Mignon V (2002) Recent developments in nonlinear cointegration with applications to macroeconomics and finance. Kluwer, Boston
20. Engle RF, Granger CWJ (1987) Co-integration and error correction: Representation, estimation, and testing. Econetrica 55:251–276
21. Engle RF, Granger CWJ, Rice J, Weiss A (1986) Semiparametric estimates of the relationship between weather and electricity sales. J Am Stat Assoc 81:310–320
22. Ericsson NR, Hendry DF, Prestwich KM (1998) The demand for broad money in the United Kingdom, 1878–1993. Scand J Econ 100:289–324
23. Escanciano JC, Hualde J (2005) Persistence and long memory in nonlinear time series. Unpublished manuscript
24. Escanciano JC, Velasco C (2006) Testing the martingale difference hypothesis using integrated regression functions. Comput Stat Data Anal 51:2278–2294
25. Escribano A (1986) Non-Linear error-correction: The case of money demand in the UK (1878–1970), ch IV. Ph.D. Dissertation, University of California, San Diego
26. Escribano A (1987) Error-Correction systems: Nonlinear adjustment to linear long-run relationships. Core Discussion Paper 8730, C.O.R.E
27. Escribano A (2004) Nonlinear error correction: The case of money demand in the UK (1878–2000). Macroecon Dyn 8:76–116
28. Escribano A, Aparicio F (1999) Cointegration: Linearity, nonlinearity, outliers and structural breaks. In: Dahiya SB (ed) The current state of economic science. Spellbound Publications, pp 383–408
29. Escribano A, Granger CWJ (1998) Investigating the relationship between gold and silver prices. J Forecast 17:81–107
30. Escribano A, Mira S (1997) Nonlinear error correction models. Working Paper 97-26. Universidad Carlos III de Madrid
31. Escribano A, Mira S (2002) Nonlinear error correction models. J Time Ser Anal 23:509–522
32. Escribano A, Mira S (2007) Specification of nonlinear error correction models: a simulation study. Mimeo, Universidad Carlos III de Madrid
33. Escribano A, Pfann GA (1998) Non-linear error correction, asymmetric adjustment and cointegration. Econ Model 15:197–216
34. Escribano A, Sipols AE, Aparicio F (2006) Nonlinear cointegration and nonlinear error correction: Applications of tests

based on first differences of ranges. Commun Stat Simul Comput 35:939–956
35. Escribano A, Santos MT, Sipols AE (2008) Testing for cointegration using induced order statistics. Comput Stat 23:131–151
36. Fotopoulos SB J, Ahn SK (2001) Rank based Dickey-Fuller tests statistics. J Time Ser Anal 24:647–662
37. Friedman M, Schwartz A (1982) Monetary trends in the united sates and the United Kingdom: Their relation to income, prices, and interest rates, 1867–1975. University of Chicago Press, Chicago
38. Gonzalo J (1994) Five alternative methods of estimating long run equilibrium relationships. J Econ 60:1–31
39. Gonzalo J, Pitarakis J-Y (2006) Threshold cointegrating relationships. Oxford Bull Econ Stat 68:813–833
40. Gouriéroux C, Jasiak J (1999) Nonlinear persistence and copersistence. Unpublished manuscript
41. Gouriéroux C, Jasiak J (2002) Nonlinear autocorrelograms: An application to inter-trade durations. J Time Ser Anal 23:127–154
42. Granger CWJ (1981) Some properties of time series data and their use in econometric model specification. J Econ 16:121–130
43. Granger CWJ (1986) Developments in the study of cointegrated variables. Oxford Bull Econ Stat 48:213–228
44. Granger CWJ (1995) Modelling nonlinear relationships between extended-memory variables. Econetrica 63:265–279
45. Granger CWJ (2002) Long memory, volatility, risk and distribution. Unpublished manuscript
46. Granger CWJ, Hallman J (1991) Long memory series with attractors. Oxford Bull Econ Stat 53:11–26
47. Granger CWJ, Hallman J (1991) Nonlinear transformations of integrated time series. J Time Ser Anal 12:207–224
48. Granger CWJ, Lee TH (1989) Investigation of production, sales and inventory relationships using multicointegration and non-symmetric error correction models. J Appl Econ 4:145–159
49. Granger CWJ, Swanson N (1996) Further developments in the study of cointegrated variables. Oxford Bull Econ Stat 58:537–553
50. Granger CWJ, Teräsvirta T (1993) Modelling nonlinear economic relationships. Oxford University Press, New York
51. Granger CWJ, Maasoumi E, Racine J (2004) A dependence metric for possibly nonlinear processes. J Time Ser Anal 25:649–669
52. Hamilton JD (1994) Time series analysis. Princeton University Press, Princeton
53. Hansen B, Phillips PCB (1990) Estimation and inference in models of cointegration: A simulation study. Adv Econ 8:225–248
54. Hansen B, Seo B (2002) Testing for two-regime threshold cointegration in vector error correction models. J Econ 110:293–318
55. Härdle W (1990) Applied nonparametric regression. Econetric Society Monographs, vol 19. Cambridge University Press, Cambridge
56. Hargreaves CP (1994) Nonstationary time series analysis and cointegration. Oxford University Press, Oxford
57. Hargreaves CP (1994) A review of methods of estimating cointegrating relationships. In: Hargreaves CP (ed) Nonstationary time series analysis and cointegration. Oxford University Press, Oxford
58. Haug AA (2002) Testing linear restrictions on cointegrating vectors: Sizes and powers of wald and likelihood ratio tests in finite samples. Econ Theory 18:505–524
59. Hendry DF (1995) Dynamic econometrics. Oxford University Press, Oxford
60. Hendry DF, Ericsson N (1983) Assertion without empirical basis: An econometric appraisal of 'monetary trends in the … the United Kingdom' by Friedman M, Schwartz AJ. In: Monetary trends in the United Kingdom. Bank of England Panel of Academic Consultants, Paper 22, 45–101
61. Hendry DF, Ericsson NR (1991) An econometric analysis of the uk money demand in 'monetary trends in the united states and the United Kingdom' by Friedman M, Schwartz AJ. Am Econ Rev 81:8–38
62. Hendry DF, Massmann M (2007) Co-breaking: Recent advances and a synopsis of the literature. J Bus Econ Stat 25:33–51
63. Inder B (1993) Estimating long-run relationships in economics: A comparison of different approaches. J Econ 57:53–68
64. Johansen S (1988) Statistical analysis of cointegration vectors. J Econ Dyn Control 12:231–54
65. Johansen S (1991) Estimation and hypothesis testing of cointegration vectors, in gaussian vector autoregressive models. Econetrica 59:1551–1580
66. Johansen S (1992) Cointegration in partial systems and the efficiency of single-equation analysis. J Econ 52:389–402
67. Juselius K (2006) The cointegrated VAR model: Methodology and applications. Oxford University Press, Oxford
68. Jusmah A, Kunst RM (2008) Modelling macroeconomic subaggregates: An application of non-linear cointegration. Economics Series, Institute for Advanced Studies. Macroecon Dyn 12:151–171
69. Karlsen HA, Tjøstheim D (2001) Nonparametric estimation in null recurrent time series. Ann Stat 29:372–416
70. Karlsen HA, Myklebust T, Tjøstheim D (2007) Nonparametric estimation in a nonlinear cointegration type model. Ann Stat 35(1):252–299
71. Kitamura Y, Phillips PCB (1995) Efficient IV estimation in nonstationary regression: An overview and simulation study. Econ Theory 11:1095–1130
72. Kunst RM (1992) Threshold cointegration in interest rates. Working Paper, Institute for Advanced Studies, Vienna
73. Kwiatkowski D, Phillips PCB, Schmidt P, Shin Y (1992) Testing the null hypothesis of stationarity against the alternative of a unit root. J Econ 54:159–178
74. Lo AW (1991) Long-term memory in stock market prices. Econometrica 59:1279–1313
75. Lo M, Zivot E (2001) Threshold cointegration and nonlinear adjustment to the law of one price. Macroecon Dyn 5:533–576
76. Longbottom A, Holly S (1985) Econometric methodology and monetarism: Professor Friedman and Professor Hendry on the demand for money, Discussion Paper No. 131. London Business School
77. Lütkephol H (2005) New introduction to multiple time series analysis. Springer
78. Marmol F, Escribano A, Aparicio FM (2002) Instrumental variable interpretation of cointegration with inference results for fractional cointegration. Econ Theory 18:646–672

79. Park JY (1992) Canonical cointegrating regressions. Econometrica 60:119–143
80. Park JY, Phillips PCB (1988) Statistical inference in regressions with integrated process: Part 1. Econ Theory 4:468–498
81. Park JY, Phillips PCB (1989) Statistical inference in regressions with integrated process: Part 2. Econ Theory 4:95–132
82. Park JY, Phillips PCB (1989) Asymptotics for nonlinear transformations of integrated time series. Econ Theory 15:269–298
83. Park JY, Phillips PCB (2000) Nonstationary binary choice. Econometrica 68:1249–1280
84. Park JY, Phillips PCB (2001) Nonlinear regressions with integrated time series. Econometrica 69:117–161
85. Phillips PCB (1991) Optimal inference in cointegrated systems. Econometrica 59:283–306
86. Phillips PCB, Hansen BE (1990) Statistical inference in instrumental variable regression with *I*(1) processes. Rev Econ Stud 57:99–125
87. Phillips PCB, Loretan M (1991) Estimating long-run economic equilibria. Rev Econ Stud 59:407–436
88. Phillips PCB, Park JY (1998) Nonstationary density estimation and kernel autoregression. Unpublished manuscript
89. Robinson PM, Gerolimetto M (2006) Instrumental variables estimation of stationary and nonstationary cointegrating regressions. Econ J 9:291–306
90. Rothman P, van Dijk D, Franses PH (2001) A multivariate star analysis of the relationship between money and output. Macroecon Dyn 5:506–532
91. Saikkonen P (1991) Asymptotically efficient estimation of cointegration regressions. Econ Theory 7:1–21
92. Saikkonen P (2005) Stability results for nonlinear error correction models. J Econ 127:69–81
93. Saikkonen P, Choi I (2004) Cointegrating smooth transition regressions. Econ Theory 20:301–340
94. Seo MH (2006) Bootstrap testing for the null of no cointegration in a threshold vector error correction model. J Econ 134:129–150
95. Seo MH (2007) Estimation of nonlinear error correction models. Unpublished manuscript
96. Silverman BW (1985) Some aspects of the spline smoothing approach to nonparametric regression curve fitting. J Roy Stat Soc Ser B 47:1–52
97. Sims CA (1972) Money, income, and causality. Am Econ Rev 62:540–552
98. Stock JH (1987) Asymptotic properties of least squares estimation of cointegrating vectors. Econometrica 55:1035–1056
99. Stock JH (1994) Deciding between *I*(1) and *I*(0). J Econ 63:105–131
100. Stock JH, Watson MW (1993) A simple estimator of cointegration vectors in higher order integrated systems. Econometrica 61:783–820
101. Teräsvirta T (1998) Modelling economic relationships with smooth transition regressions. In: Ullah A, Giles DEA (eds) Handbook of applied economic statistics. Marcel Dekker, New York, pp 507–552
102. Teräsvirta T, Eliasson AC (2001) Non-linear error correction and the uk demand for broad money, 1878–1993. J Appl Econ 16:277–288
103. Velasco C (2006) Semiparametric estimation of long-memory models. In: Patterson K, Mills TC (eds) Palgrave handbook of econometrics, vol 1. Econometric Theory. MacMillan, Palgrave, pp 353–395
104. Wahba G (1975) Smoothing noisy data with spline functions. Num Math 24:309-63–75
105. Watson M (1994) Large sample estimation and hypothesis testing, vector autoregression and cointegration. In: Engle RF, McFadden DL (eds) Handbook of econometrics, vol IV. Elsevier Science, Amsterdam

Books and Reviews

Franses PH, Terasvirta T (eds)(2001) Special issue on nonlinear modeling of multivariate macroeconomic relations. Macroecon Dyn 5(4)
Banerjee A, Dolado JJ, Galbraith JW, Hendry DF (1993) Co-integration, error correction and the econometric analysis of non-stationary data. Oxford University Press, Oxford
Bierens HJ (1981) Robust methods and asymptotic theory in nonlinear econometrics. Lecture Notes in Economics and Mathematical Systems, vol 192. Springer, Berlin
Clements MP, Hendry DF (1999) Forecasting non-stationary economic time series. MIT Press, Cambridge
Dhrymes P (1998) Time series, unit roots, and cointegration. Academic Press, San Diego
Enders W (1995) Applied econometric time series. Wiley
Engle RF, Granger CWJ (eds) (1991) Long-run economic relationships: Readings in cointegration. Oxford University Press, Oxford
Franses PH, van Dijk D (2000) Nonlinear time series models in empirical finance. Cambridge University Press, Cambridge
Gonzalo J, Dolado JJ, Marmol F (2001) A primer in cointegration. In: Baltagui BH (ed) A companion to theoretical econometrics. Blackwell, New York, ch 30
Granger CWJ (2001) Overview of nonlinear macroeconometric empirical models. Macroecon Dyn 5:466–481
Granger CWJ (2007) Some thoughts on the past and future of cointegration. Mimeo
Hatanaka M (1996) Time series-based econometrics. Oxford University Press, Oxford
Maddala GS, Kim I-M (1998) Unit roots, cointegration and structural change. Cambridge University Press, Cambridge
Phillips PCB (1986) Understanding spurious regressions in econometrics. J Econ 33:311–340

Econometrics: Panel Data Methods

Jeffrey M. Wooldridge
Department of Economics, Michigan State University, East Lansing, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Overview of Linear Panel Data Models
Sequentially Exogenous Regressors and Dynamic Models
Unbalanced Panel Data Sets
Nonlinear Models

Future Directions
Bibliography

Glossary

Panel data Data on a set of cross-sectional units followed over time.

Unobserved effects Unobserved variables that affect the outcome which are constant over time.

Fixed effects estimation An estimation method that removes the unobserved effects, implying that the unobserved effects can be arbitrarily related to the observed covariates.

Correlated random effects An approach to modeling where the dependence between the unobserved effects and the history of the covariates is parametrically modeled. The traditional random effects approach is a special case under the assumption that he unobserved effects are independent of the covariates.

Average partial effect The partial effect of a covariate averaged across the distribution of the unobserved effects.

Definition of the Subject

Panel data consist of repeated observations over time on the same set of cross-sectional units. These units can be individuals, firms, schools, cities, or any collection of units one can follow over time. Special econometric methods have been developed to recognize and exploit the rich information available in panel data sets. Because the time dimension is a key feature of panel data sets, issues of serial correlation and dynamic effects need to be considered. Further, unlike the analysis of cross-sectional data, panel data sets allow the presence of systematic, unobserved differences across units that can be correlated with observed factors whose effects are to be measured. Distinguishing between persistence due to unobserved heterogeneity and that due to dynamics in the underlying process is a leading challenge for interpreting estimates from panel data models.

Panel data methods are the econometric tools used to estimate parameters compute partial effects of interest in nonlinear models, quantify dynamic linkages, and perform valid inference when data are available on repeated cross sections. For linear models, the basis for many panel data methods is ordinary least squares applied to suitably transformed data. The challenge is to develop estimators assumptions with good properties under reasonable assumptions, and to ensure that statistical inference is valid. Maximum likelihood estimation plays a key role in the estimation of nonlinear panel data models.

Introduction

Many questions in economics, especially those with foundations in the behavior of relatively small units, can be empirically studied with the help of panel data. Even when detailed cross-sectional surveys are available, collecting enough information on units to account for systematic differences is often unrealistic. For example, in evaluating the effects of a job training program on labor market outcomes, unobserved factors might affect both participation in the program and outcomes such as labor earnings. Unless participation in the job training program is randomly assigned, or assigned on the basis of observed covariates, cross-sectional regression analysis is usually unconvincing. Nevertheless, one can control for this individual heterogeneity – including unobserved, time-constant human capital – by collecting a panel data set that includes data points both before and after the training program.

Some of the earliest econometric applications of panel data methods were to the estimation of agricultural production functions, where the worry was that unobserved inputs – such as soil quality, technical efficiency, or managerial skill of the farmer – would generally be correlated with observed inputs such as capital, labor, and amount of land. Classic examples are [31,45].

The nature of unobserved heterogeneity was discussed early in the development of panel data models. An important contribution is [46], which argued persuasively that in applications with many cross-sectional units and few time periods, it always makes sense to treat unit-specific heterogeneity as outcomes of random variables, rather than parameters to estimate. As Mundlak made clear, for economic applications the key issue is whether the unobserved heterogeneity can be assumed to be independent, or at least uncorrelated, with the observed covariates. [25] developed a testing framework that can be used, and often is, to test whether unobserved heterogeneity is correlated with observed covariates. Mundlak's perspective has had a lasting impact on panel data methods, and his insights have been applied to a variety of dynamic panel data models with unobserved heterogeneity.

The 1980s witnessed an explosion in both methodological developments and applications of panel data methods. Following the approach in [15,16,46], and [17] provided a unified approach to linear and nonlinear panel data models, and explicitly dealt with issues of inference in cases where full distributions were not specified. Dynamic linear models, and the problems they pose for estimation

and inference, were considered in [4]. Dynamic discrete response models were analyzed in [29,30]. The hope in estimating dynamic models that explicitly contain unobserved heterogeneity is that researchers can measure the importance of two causes for persistence in observed outcomes: unobserved, time-constant heterogeneity and so-called *state dependence*, which describes the idea that, conditional on observed and unobserved factors, the probability of being in a state in the current time period is affected by last period's state.

In the late 1980s and early 1990s, researchers began using panel data methods to test economic theories such as rational expectations models of consumption. Unlike macro-level data, data at the individual or family level allows one to control for different preferences, and perhaps different discount rates, in testing the implications of rational expectations. To avoid making distributional assumptions on unobserved shocks and heterogeneity, researchers often based estimation on conditions on expected values that are implied by rational expectations, as in [40].

Other developments in the 1990s include studying standard estimators under fewer assumptions – such as the analysis in [53] of the fixed effects Poisson estimator under distributional misspecification and unrestricted serial dependence – and the development of estimators in nonlinear models that are consistent for parameters under no distributional assumptions – such as the new estimator proposed in [33] for the panel data censored regression model.

The past 15 years has seen continued development of both linear and nonlinear models, with and without dynamics. For example, on the linear model front, methods have been proposed for estimating models where the effects of time-invariant heterogeneity can change over time – as in [2]. Semiparametric methods for estimating production functions, as in [48], and dynamic models, as in the dynamic censored regression model in [34], have been developed. Flexible parametric models, estimated by maximum likelihood, have also been proposed (see [57]).

Many researchers are paying closer attention to estimation of partial effects, and not just parameters, in nonlinear models – with or without dynamics. Results in [3] show how partial effects, with the unobserved heterogeneity appropriately averaged out, can be identified under weak assumptions.

The next several sections outline a modern approach to panel data methods. Section "Future Directions" provides an account of more recent advances, and discusses where those advances might head in the future.

Overview of Linear Panel Data Models

In panel data applications, linear models are still the most widely used. When drawing data from a large population, random sampling is often a realistic assumption; therefore, we can treat the observations as independent and identically distributed outcomes. For a random draw i from the population, the linear panel data model with additive heterogeneity can be written as

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}\,, \quad t = 1, \ldots, T\,, \tag{1}$$

where T is the number of time periods available for each unit and t indexes time periods. The time periods are often years, but the span between periods can be longer or shorter than a year. The distance between any two time periods need not be the same, although different spans can make it tricky to estimate certain dynamic models. As written, Eq. (1) assumes that we have the same time periods available for each cross-sectional unit. In other words, the panel data set is *balanced*.

As in any regression analysis, the left-hand-side variable is the dependent variable or the response variable. The terms η_t, which depend only only time, are treated here as parameters. In most microeconometric applications, the cross-sectional sample size, denoted N, is large – often very large – compared with T. Therefore, the η_t can be estimated precisely in most cases. Almost all applications should allow for aggregate time effects as captured by η_t. Including such time effects allows for secular changes in the economic environment that affect all units in the same way (such as inflation or aggregate productivity). For example, in studying the effects of school inputs on performance using school-level panel data for a particular state, including η_t allows for trends in statewide spending along with separate, unrelated trends in statewide test performance. It could be that, say, real spending rose at the same time that the statewide standardized tests were made easier; a failure to account for such aggregate trends could lead to a spurious association between performance and spending. Only occasionally are the η_t the focus of a panel data analysis, but it is sometimes interesting to study the pattern of aggregate changes once the covariates contained in the $1 \times K$ vector $\mathbf{x}_{it}$ are netted out.

The parameters of primary interest are contained in the $K \times 1$ vector $\boldsymbol{\beta}$, which contains the coefficients on the set of explanatory variables. With the presence of η_t in (1), $\mathbf{x}_{it}$ cannot include variables that change only over time. For example, if y_{it} is a measure of labor earnings for individual i in year t for a particular state in the US, $\mathbf{x}_{it}$ cannot contain the state-level unemployment rate. Unless interest centers on how individual earnings depend on the

state-level unemployment rate, it is better to allow for different time intercepts in an unrestricted fashion: this way, any aggregate variables that affect each individual in the same way are accounted for without even collecting data on them. If the η_t are restricted to be functions of time – for example, a linear time trend – then aggregate variables can be included, but this is always more restrictive than allowing the η_t to be unrestricted.

The composite error term in (1), $c_i + u_{it}$, is an important feature of panel data models. With panel data, it makes sense to view the unobservables that affect y_{it} as consisting of two parts: the first is the time-constant variable, c_i, which is often called an *unobserved effect* or *unit-specific heterogeneity*. This term aggregates all factors that are important for unit i's response that do not change over time. In panel data applications to individuals, c_i is often interpreted as containing cognitive ability, family background, and other factors that are essentially determined prior to the time periods under consideration. Or, if i indexes different schools across a state, and (1) is an equation to see if school inputs affect student performance, c_i includes historical factors that can affect student performance and also might be correlated with observed school inputs (such as class sizes, teacher competence, and so on). The word "heterogeneity" is often combined with a qualifier that indicates the unit of observation. For example, c_i might be "individual-specific heterogeneity" or "school-specific heterogeneity". Often in the literature c_i is called a "random effect" or "fixed effect", but these labels are not ideal. Traditionally, c_i was considered a random effect if it was treated as a random variable, and it was considered a fixed effect if it was treated as a parameter to estimate (for each i). The flaws with this way of thinking are revealed in [46]: the important issue is not whether c_i is random, but whether it is correlated with $\mathbf{x}_{it}$.

The sequence of errors $\{u_{it}: t = 1, \ldots, T\}$ are specific to unit i, but they are allowed to change over time. Thus, these are the time-varying unobserved factors that affect y_{it}, and they are often called the *idiosyncratic errors*. Because u_{it} is in the error term at time t, it is important to know whether these unobserved, time-varying factors are uncorrelated with the covariates. It is also important to recognize that these idiosyncratic errors can be serially correlated, and often are.

Before treating the various assumptions more formally in the next subsection, it is important to recognize the asymmetry in the treatment of the time-specific effects, η_t, and the unit-specific effects, c_i. Language such as "both time and school fixed effects are included in the equation" is common in empirical work. While the language itself is harmless, with large N and small T it is best to view the time effects, η_t, as parameters to estimate because they can be estimated precisely. As already mentioned earlier, viewing c_i as random draws is the most general, and natural, perspective.

Assumptions and Estimators for the Basic Model

The assumptions discussed in this subsection are best suited to cases where random sampling from a (large) population is realistic. In this setting, it is most natural to describe large-sample statistical properties as the cross-sectional sample size, N, grows, with the number of time periods, T, fixed.

In describing assumptions in the model (1), it probably makes more sense to drop the i subscript in (1) to emphasize that the equation holds for an entire population. Nevertheless, (1) is useful for emphasizing which factors change i, or t, or both. It is sometimes convenient to subsume the time dummies in $\mathbf{x}_{it}$, so that the separate intercepts η_t need not be displayed.

The traditional starting point for studying (1) is to rule out correlation between the idiosyncratic errors, u_{it}, and the covariates, $\mathbf{x}_{it}$. A useful assumption is that the sequence of explanatory variables $\{\mathbf{x}_{it}: t = 1, \ldots, T\}$ is *contemporaneously exogenous conditional on* c_i:

$$E(u_{it}|\mathbf{x}_{it}, c_i) = 0\,, \quad t = 1, \ldots, T\,. \tag{2}$$

This assumption essentially defines $\boldsymbol{\beta}$ in the sense that, under (1) and (2),

$$E(y_{it}|\mathbf{x}_{it}, c_i) = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i\,, \tag{3}$$

so the β_j are partial effects holding fixed the unobserved heterogeneity (and covariates other than x_{tj}). Strictly speaking, c_i need not be included in the conditioning set in (2), but including it leads to the useful Eq. (3). Plus, for purposes of stating assumptions for inference, it is convenient to express the contemporaneous exogeneity assumption as in (2).

Unfortunately, with a small number of time periods, $\boldsymbol{\beta}$ is not identified by (2), or by the weaker assumption $\mathrm{Cov}(\mathbf{x}_{it}, u_{it}) = \mathbf{0}$. Of course, if c_i is assumed to be uncorrelated with the covariates, that is $\mathrm{Cov}(\mathbf{x}_{it}, c_i) = \mathbf{0}$ for any t, then the composite error, $v_{it} = c_i + u_{it}$ is uncorrelated with $\mathbf{x}_{it}$, and then $\boldsymbol{\beta}$ is identified and can be consistently estimated by a cross section regression using a single time period t, or by using pooled regression across t. (See Chaps. 7 and 10 in [55] for further discussion.) But one of the main purposes in using panel data is to allow the unobserved effect to be correlated with time-varying $\mathbf{x}_{it}$.

Arbitrary correlation between c_i and $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT})$ is allowed if the sequence of explanatory variables

is *strictly exogenous conditional on* c_i,

$$E(u_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}, c_i) = 0\,, \quad t = 1, \ldots, T\,, \tag{4}$$

which can be expressed as

$$E(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, c_i) = E(y_{it}|\mathbf{x}_{it}, c_i) = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i\,. \tag{5}$$

Clearly, assumption (4) implies (2). Because the entire history of the covariates is in (4) for all t, (4) implies that $\mathbf{x}_{ir}$ and u_{it} are uncorrelated for all r and t, including $r = t$. By contrast, (2) allows arbitrary correlation between $\mathbf{x}_{ir}$ and u_{it} for any $r \neq t$. The strict exogeneity assumption (4) can place serious restrictions on the nature of the model and dynamic economic behavior. For example, (4) can never be true if $\mathbf{x}_{it}$ contains lags of the dependent variable. Of course, (4) would be false under standard econometric problems, such as omitted time-varying variables, just as would (2). But there are important cases where (2) can hold but (4) might not. If, say, a change in u_{it} causes reactions in future values of the explanatory variables, then (4) is generally false. In applications to the social sciences, the potential for these kind of "feedback effects" is important. For example, in using panel data to estimate a firm-level production function, a shock to production today (captured by changes in u_{it}) might affect the amount of capital and labor inputs in the next time period. In other words, u_{it} and $\mathbf{x}_{i,t+1}$ would be correlated, violating (4).

How does assumption (4) (or (5)) identify the parameters? In fact, it only allows estimation of coefficients on time-varying elements of $\mathbf{x}_{it}$. Intuitively, because (4) puts no restrictions on the dependence between c_i and $\mathbf{x}_i$, it is not possible to distinguish between the effect of a time-constant observable covariate and that of the unobserved effect, c_i. For example, in an equation to describe the amount of pension savings invested in the stock market, c_i might include innate of tolerance for risk, assumed to be fixed over time. Once c_i is allowed to be correlated with any observable covariate – including, say, gender – the effects of gender on stock market investing cannot be identified because gender, like c_i, is constant over time. Mechanically, common estimation methods eliminate c_i along with any time-constant explanatory variables. (What is meant by "time-varying" x_{itj} is that for at least some i, x_{itj} changes over time. For some units i, x_{itj} might be constant). When a full set of year intercepts – or even just a linear time trend – is included, the effects of variables that increase by the same amount in each period – such as a person's age – cannot be included in $\mathbf{x}_{it}$. The reason is that the beginning age of each person is indistinguishable from c_i, and then, once the initial age is know, each subsequent age is a deterministic – in fact, linear – function of time.

Perhaps the most common method of estimating $\boldsymbol{\beta}$ (and the η_t) is so-called *fixed effects (FE)* or *within* estimation. The FE estimator is obtained as a pooled OLS regression on variables that have had the unit-specific means removed. More precisely, let $\ddot{y}_{it} = y_{it} - T^{-1}\sum_{r=1}^{T} y_{ir} = y_{it} - \bar{y}_i$ be the deviation of y_{it} from the average over time for unit i, $\bar{y}_i$ and similarly for $\ddot{\mathbf{x}}_{it}$ (which is a vector). Then,

$$\ddot{y}_{it} = \ddot{\eta}_t + \ddot{\mathbf{x}}_{it}\boldsymbol{\beta} + \ddot{u}_{it}\,, \quad t = 1, \ldots, T\,, \tag{6}$$

where the year intercepts and idiosyncratic errors are, of course, also demeaned. Consistency of pooled OLS (for fixed T and $N \to \infty$) applied to (6) essentially requires rests on $\sum_{t=1}^{T} E(\ddot{\mathbf{x}}_{it}'\ddot{u}_{it}) = \sum_{t=1}^{T} E(\ddot{\mathbf{x}}_{it}'u_{it}) = \mathbf{0}$, which means the error u_{it} should be uncorrelated with $\mathbf{x}_{ir}$ for all r and t. This assumption is implied by (4). A rank condition on the demeaned explanatory variables is also needed. If $\ddot{\eta}_t$ is absorbed into $\ddot{\mathbf{x}}_{it}$, the condition is *rank* $\sum_{t=1}^{T} E(\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}) = K$, which rules out time constant variables and other variables that increase by the same value for all units in each time period (such as age).

A different estimation method is based on an equation in first differences. For $t > 1$, define $\Delta y_{it} = y_{it} - y_{i,t-1}$, and similarly for the other quantities. The first-differenced equation is

$$\Delta y_{it} = \delta_t + \Delta\mathbf{x}_{it}\boldsymbol{\beta} + \Delta u_{it}\,, \quad t = 2, \ldots, T\,, \tag{7}$$

where $\delta_t = \eta_t - \eta_{t-1}$ is the change in the intercepts. The *first-difference* (FD) estimator is pooled OLS applied to (7). Any element $x_{i\text{th}}$ of $\mathbf{x}_{it}$ such that Δx_{ith} is constant for all i and t (most often zero) drops out, just as in FE estimation. Assuming suitable time variation in the covariates, $E(\Delta\mathbf{x}_{it}'\Delta u_{it}) = \mathbf{0}$ is sufficient for consistency. Naturally, this assumption is also implied by assumption (4).

Whether FE or FD estimation is used – and it is often prudent to try both approaches – inference about $\boldsymbol{\beta}$ can and generally should be be made fully robust to heteroksedasticity and serial dependence. The robust asymptotic variance of both FE and FD estimators has the so-called "sandwich" form, which allows the vector of idiosyncratic errors, $\mathbf{u}_i = (u_{i1}, \ldots, u_{iT})'$, to contain arbitrary serial correlation and heteroskedasticity, where the conditional covariances and variances can depend on $\mathbf{x}_i$ in an unknown way. For notational simplicity, absorb dummy variables for the different time periods into $\mathbf{x}_{it}$. Let $\hat{\boldsymbol{\beta}}_{\text{FE}}$ denote the fixed effects estimator and $\hat{\ddot{\mathbf{u}}}_i = \ddot{\mathbf{y}}_i - \ddot{\mathbf{X}}_i\hat{\boldsymbol{\beta}}_{\text{FE}}$ the $T \times 1$ vector of fixed effects residuals for unit i. Here, $\ddot{\mathbf{X}}_i$ is the $T \times K$ matrix with t^{th} row $\ddot{\mathbf{x}}_{it}$.

Then a fully robust estimator of the asymptotic variance of $\hat{\boldsymbol{\beta}}_{\mathrm{FE}}$ is

$$\widehat{\mathrm{Avar}}(\hat{\boldsymbol{\beta}}_{\mathrm{FE}}) = \left(\sum_{i=1}^{N} \ddot{\mathbf{X}}_i' \ddot{\mathbf{X}}_i\right)^{-1} \left(\sum_{i=1}^{N} \ddot{\mathbf{X}}_i' \widehat{\ddot{\mathbf{u}}}_i \widehat{\ddot{\mathbf{u}}}_i' \ddot{\mathbf{X}}_i\right) \cdot \left(\sum_{i=1}^{N} \ddot{\mathbf{X}}_i' \ddot{\mathbf{X}}_i\right)^{-1}, \quad (8)$$

where it is easily seen that $\sum_{i=1}^{N} \ddot{\mathbf{X}}_i' \ddot{\mathbf{X}}_i = \sum_{i=1}^{N} \sum_{t=1}^{T} \ddot{\mathbf{x}}_{it} \ddot{\mathbf{x}}_{it}$ and the middle part of the sandwich consists of terms $\widehat{\ddot{u}}_{ir} \widehat{\ddot{u}}_{it} \ddot{\mathbf{x}}_{ir}' \ddot{\mathbf{x}}_{it}$ for all $r, t = 1, \ldots, T$. See Chap. 10 in [55] for further discussion. A similar expression holds for $\hat{\boldsymbol{\beta}}_{\mathrm{FD}}$ but where the demeaned quantities are replaced by first differences.

When $T = 2$, it can be shown that the FE and FD estimation and inference about $\boldsymbol{\beta}$ are identical. If $T > 2$, the procedures generally differ. If (4) holds and $T > 2$, how does one choose between the FE and FD approaches? Because both are consistent and $\sqrt{N}$-asymptotically normal, the only way to choose is from efficiency considerations. Efficiency of the FE and FD estimators hinges on second moment assumptions concerning the idiosyncratic errors. Briefly, if $E(\mathbf{u}_i \mathbf{u}_i' | \mathbf{x}_i) = E(\mathbf{u}_i \mathbf{u}_i') = \sigma_u^2 \mathbf{I}_T$, then the FE estimator is efficient. Practically, the most important implication of this assumption is that the idiosyncratic errors are serially uncorrelated. But they should also be homoskedastic, which means the variances can neither depend on the covariates nor change over time. The FD estimator is efficient if the errors in (7) are serially uncorrelated and homoskedasticity, which can be stated as $E(\Delta \mathbf{u}_i \Delta \mathbf{u}_i' | \mathbf{x}_i) = E(\Delta \mathbf{u}_i \Delta \mathbf{u}_i') = \sigma_e^2 \mathbf{I}_{T-1}$, where $e_{it} = u_{it} - u_{i,t-1}$ and $\Delta \mathbf{u}_i$ is the $T-1$ vector of first-differenced errors. These two sets of conditions – that $\{u_{it}: t = 1, \ldots, T\}$ is a serially uncorrelated sequence (for FE to be efficient) versus $\{u_{it}: t = 1, \ldots, T\}$ is a random walk (for FD to be efficient) – represent extreme cases. Of course, there is much in between. In fact, probably neither condition should be assumed to be true, which is a good argument for robust inference. More efficient estimation can be based on generalized method of moments (GMM – see Chap. 8 in [55] – or minimum distance estimation, as in [16]).

It is good practice to compute both FE and FD estimates to see if they differ in substantive ways. It is also helpful to have a formal test of the strict exogeneity assumption that is easily computable and that maintains only strict exogeneity under the null – in particular, that takes no stand on whether the FE or FD estimator is asymptotically efficient. Because lags of covariates can always be included in a model, the primary violation of (4) that is of interest is due to feedback. Therefore, it makes sense to test that $\mathbf{x}_{i,t+1}$ is uncorrelated with u_{it}. Actually, let $\mathbf{w}_{it}$ be a subset of $\mathbf{x}_{it}$ that is suspected of failing the strict exogeneity assumption, and consider the augmented model

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{w}_{i,t+1}\boldsymbol{\delta} + c_i + u_{it}, \quad t = 1, \ldots, T-1. \quad (9)$$

Under the null hypothesis that $\{\mathbf{x}_{it}: t = 1, \ldots, T\}$ is strictly exogenous, $H_0: \boldsymbol{\delta} = \mathbf{0}$, and this is easily tested using fixed effects (using all but the last time period) or first differencing (where, again, the last time period is lost). It makes sense, as always, to make the test fully robust to serial correlation and heteroskedasticity. This test may probably has little power for detecting contemporaneous endogeneity, that is, correlation between $\mathbf{w}_{it}$ and u_{it}.

A third common approach to estimation of unobserved effects models is so-called *random effects* estimation. RE estimation differs from FE and FD by leaving c_i in the error term and then accounting for its presence via generalized least squares (GLS). Therefore, the exogeneity requirements of the covariates must be strengthened. The most convenient way of stating the key random effects (RE) assumption is

$$E(c_i | \mathbf{x}_i) = E(c_i), \quad (10)$$

which ensures that every element of $\mathbf{x}_i$ – that is, all explanatory variables in all time periods – is uncorrelated with c_i. Together with (4), (10) implies

$$E(v_{it} | \mathbf{x}_i) = 0, \quad t = 1, \ldots, T, \quad (11)$$

where $v_{it} = c_i + u_{it}$ is the composite error. Condition (11) is the key condition for general least squares methods that exploit serial correlation in v_{it} to be consistent (although zero correlation would be sufficient). The random effects estimator uses a special structure for the variance-covariate matrix of $\mathbf{v}_i$, the $T \times 1$ vector of composite errors. If $E(\mathbf{u}_i \mathbf{u}_i') = \sigma_u^2 \mathbf{I}_T$ and c_i is uncorrelated with each u_{it} (as implied by assumption (4)), then

$$\mathrm{Var}(\mathbf{v}_i) = \begin{pmatrix} \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & \cdots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sigma_c^2 \\ \sigma_c^2 & \cdots & \sigma_c^2 & \sigma_c^2 + \sigma_u^2 \end{pmatrix}. \quad (12)$$

Both σ_c^2 and σ_u^2 can be estimated after, say, preliminary estimation by pooled OLS (which is consistent under (11)) –

see, for example, Chap. 10 in [55] – and then a feasible GLS is possible. If (12) holds, along with the system homoskedasticity assumption $\mathrm{Var}(\mathbf{v}_i|\mathbf{x}_i) = \mathrm{Var}(\mathbf{v}_i)$, then feasible GLS is efficient, and the inference is standard. Even if $\mathrm{Var}(\mathbf{v}_i|\mathbf{x}_i)$ is not constant, or $\mathrm{Var}(\mathbf{v}_i)$ does not have the random effects structure in (12), the RE estimator is consistent provided (11) holds (Again, this is with N growing and T fixed). Therefore, although it is still not common, a good case can be made for using robust inference – that is, inference that allows an unknown form of $\mathrm{Var}(\mathbf{v}_i|\mathbf{x}_i)$ – in the context of random effects. The idea is that the RE estimator can be more efficient than pooled OLS even if (12) fails, yet inference should not rest on (12). Chapter 10 in [55] contains the sandwich form of the estimator.

Under the key RE assumption (11), $\mathbf{x}_{it}$ can contain time-constant variables. In fact, one way to ensure that the omitted factors are uncorrelated with the key covariates is to include a rich set of time-constant controls in $\mathbf{x}_{it}$. RE estimation is most convincing when many good time-constant controls are available. In some applications of RE, the key variable of interest does not change over time, which is why FE and FD cannot be used. (Methods proposed in [26] can be used when some covariates are correlated with c_i, but enough others are assumed to be uncorrelated with c_i).

Rather than eliminate c_i using the FE or FD transformation, or assuming (10) and using GLS, a different approach is to explicitly model the correlation between c_i and $\mathbf{x}_i$. A general approach is to write

$$c_i = \psi + \mathbf{x}_i\boldsymbol{\lambda} + a_i \,, \tag{13}$$

$$E(a_i) = 0 \quad \text{and} \quad E(\mathbf{x}_i' a_i) = \mathbf{0} \,, \tag{14}$$

where $\boldsymbol{\lambda}$ is a $TK \times 1$, vector of parameters. Equations (13) and (14) are definitional, and simply define the population linear regression of c_i on the entire set of covariates, $\mathbf{x}_i$. This representation is due to [16], and is an example of a *correlated random effects* (CRE) model. The uncorrelated random effects model occurs when $\boldsymbol{\lambda} = \mathbf{0}$.

A special case of (13) was used in [46], assuming that each $\mathbf{x}_{ir}$ has the same set of coefficients. Plus, [46] actually used conditional expectations (which is unnecessary but somewhat easier to work with):

$$c_i = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i \tag{15}$$

$$E(a_i|\mathbf{x}_i) = 0 \,, \tag{16}$$

where recall that $\bar{\mathbf{x}}_i = T^{-1}\sum_{t=1}^{T}\mathbf{x}_{it}$. This formulation conserves on degrees of freedom, and extensions are useful for nonlinear models.

Plugging (15) into the original equation gives

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i + u_{it} \,, \tag{17}$$

where ψ is absorbed into the time intercepts. The composite error $a_i + u_{it}$ satisfies $E(a_i + u_{it}|\mathbf{x}_i) = 0$, and so pooled OLS or random effects applied to (17) produces consistent, $\sqrt{N}$-asymptotically normal estimators of all parameters, including $\boldsymbol{\xi}$. In fact, if the original model satisfies the second moments ideal for random effects, then so does (17). Interesting, both pooled OLS and RE applied to (17) produce the fixed effects estimate of $\boldsymbol{\beta}$ (and the η_t). Therefore, the FE estimator can be derived from a correlated random effects model. (Somewhat surprisingly, the same algebraic equivalence holds using Chamberlain's more flexible device. Of course, the pooled OLS estimator is not generally efficient, and [16] shows how to obtain the efficient minimum distance estimator. See also Chap. 11 in [55]).

One advantage of Eq. (17) is that it provides another interpretation of the FE estimate: it is obtained by holding fixed the time averages when obtaining the partial effects of each x_{itj}. This results in a more convincing analysis than not controlling for systematic differences in the levels of the covariates across i.

Equation (17) has other advantages over just using the time-demeaned data in pooled OLS: time-constant variables can be included in (17), and the resulting equation gives a simple, robust way of testing whether the time-varying covariates are uncorrelated with It is helpful to write the original equation as

$$y_{it} = \mathbf{g}_t\boldsymbol{\eta} + \mathbf{z}_i\boldsymbol{\gamma} + \mathbf{w}_{it}\boldsymbol{\delta} + c_i + u_{it} \,, \quad t = 1, \ldots, T, \tag{18}$$

where $\mathbf{g}_t$ is typically a vector of time period dummies but could instead include other variables that change only over time, including linear or quadratic trends, $\mathbf{z}_i$ is a vector of time-constant variables, and $\mathbf{w}_{it}$ contains elements that vary across i and t. It is clear that, in comparing FE to RE estimation, $\boldsymbol{\gamma}$ can play no role because it cannot be estimated by FE. What is less clear, but also true, is that the coefficients on the aggregate time variables, $\boldsymbol{\eta}$, cannot be included in any comparison, either. Only the $M \times 1$ estimates of $\boldsymbol{\delta}$, say $\hat{\boldsymbol{\delta}}_{\mathrm{FE}}$ and $\hat{\boldsymbol{\delta}}_{\mathrm{RE}}$, can be compared. If $\hat{\boldsymbol{\eta}}_{\mathrm{FE}}$ and $\hat{\boldsymbol{\eta}}_{\mathrm{RE}}$ are included, the asymptotic variance matrix of the difference in estimators has a nonsingularity in the asymptotic variance matrix. (In fact, RE and FE estimation only with aggregate time variables are identical.) The Mundlak equation is now

$$y_{it} = \mathbf{g}_t\boldsymbol{\eta} + \mathbf{z}_i\boldsymbol{\gamma} + \mathbf{w}_{it}\boldsymbol{\delta} + \bar{\mathbf{w}}_i\boldsymbol{\xi} + a_i + u_{it} \,, \quad t = 1, \ldots, T, \tag{19}$$

where the intercept is absorbed into $\mathbf{g}_t$. A test of the key RE assumption is $H_0\colon \boldsymbol{\xi} = \mathbf{0}$ is obtained by estimating (19) by RE, and this equation makes it clear there M restrictions to test. This test was described in [5,46] proposed the robust version. The original test based directly on comparing the RE and FE estimators, as proposed in [25], it more difficult to compute and not robust because it maintains that the RE estimator is efficient under the null.

The model in (19) gives estimates of the coefficients on the time-constant variables $\mathbf{z}_i$. Generally, these can be given a causal interpretation only if

$$E(c_i|\mathbf{w}_i, \mathbf{z}_i) = E(c_i|\mathbf{w}_i) = \psi + \bar{\mathbf{w}}_i\boldsymbol{\xi}\ , \tag{20}$$

where the first equality is the important one. In other words, $\mathbf{z}_i$ is uncorrelated with c_i once the time-varying covariates are controlled for. This assumption is too strong in many applications, but one still might want to include time-constant covariates.

Before leaving this subsection, it is worth point out that generalized least squares methods with an unrestricted variance-covariance matrix can be applied to every estimating equation just presented. For example, after eliminating c_i by removing the time averages, the resulting vector of errors, $\ddot{\mathbf{u}}_i$, can have an unrestricted variance matrix. (Of course, there is no guarantee that this matrix is the same as the variance matrix conditional on the matrix of time-demeaned regressors, $\ddot{\mathbf{X}}_i$.) The only glitch in practice is that $\mathrm{Var}(\ddot{\mathbf{u}}_i)$ has rank $T-1$, not T. As it turns out, GLS with an unrestricted variance matrix for the original error vector, $\mathbf{u}_i$, can be implemented on the time-demeaned equation with any of the T time periods dropped. The so-called *fixed effects GLS* estimates are invariant to whichever equation is dropped. See [41] or [37] for further discussion. The initial estimator used to estimate the variance covariance matrix would probably be the usual FE estimator (applied to all time periods).

Feasible GLS can be applied directly the first differenced equation, too. It can also be applied to (19), allowing the composite errors $a_i + u_{it}$, $t = 1, \ldots, T$, to have an unrestricted variance-covariance matrix. In all cases, the assumption that the conditional variance matrix equals the unconditional variance can fail, and so one should use fully robust inference even after using FGLS. Chapter 10 in [55] provides further discussion. Such options are widely available in software, sometimes under the rubric of *generalized estimating equations* (GEE). See, for example, [43].

Models with Heterogeneous Slopes

The basic model described in the previous subsection introduces a single source of heterogeneity in the additive effect, c_i. The form of the model implies that the partial effects of the covariates depend on a fixed set of population values (and possibly other unobserved covariates if interactions are included in $\mathbf{x}_{it}$). It seems natural to extend the model to allow interactions between the observed covariates and time-constant, unobserved heterogeneity:

$$y_{it} = c_i + \mathbf{x}_{it}\mathbf{b}_i + u_{it} \tag{21}$$

$$E(u_{it}|\mathbf{x}_i, c_i, \mathbf{b}_i) = 0\ , \quad t = 1, \ldots, T\ , \tag{22}$$

where $\mathbf{b}_i$ is $K \times 1$. With small T, one cannot precisely estimate $\mathbf{b}_i$. Instead, attention usually focuses on the *average partial effect* (APE) or *population averaged effect* (PAE). In (21), the vector of APEs is $\boldsymbol{\beta} \equiv E(\mathbf{b}_i)$, the $K \times 1$ vector of means. In this formulation, aggregate time effects are in $\mathbf{x}_{it}$. This model is sometimes called a *correlated random slopes* model – which means the slopes are allowed to be correlated with the covariates.

Generally, allowing $(c_i, \mathbf{b}_i)$ and $\mathbf{x}_i$ to be arbitrarily correlated requires $T > K + 1$ – see [56]. With a small number of time periods and even a modest number of regressors, this condition often fails in practice. Chapter 11 in [55] discusses how to allow only a subset of coefficients to be unit specific. Of interest here is the question: if the usual FE estimator is applied – that is, ignoring the unit-specific slopes $\mathbf{b}_i$ – does this ever consistently estimate the APEs in $\boldsymbol{\beta}$? In addition to the usual rank condition and the strict exogeneity assumption (22), [56] shows that a simple sufficient condition is

$$\mathrm{E}(\mathbf{b}_i|\ddot{\mathbf{x}}_{it}) = \mathrm{E}(\mathbf{b}_i) = \boldsymbol{\beta}\ , \quad t = 1, \ldots, T\ . \tag{23}$$

Importantly, condition (23) allows the slopes, $\mathbf{b}_i$, to be correlated with the regressors $\mathbf{x}_{it}$ through permanent components. It rules out correlation between idiosyncratic movements in $\mathbf{x}_{it}$ and $\mathbf{b}_i$. For example, suppose the covariates can be decomposed as $\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{r}_{it}$, $t = 1, \ldots, T$. Then (23) holds if $\mathrm{E}(\mathbf{b}_i|\mathbf{r}_{i1}, \mathbf{r}_{i2}, \ldots, \mathbf{r}_{iT}) = \mathrm{E}(\mathbf{b}_i)$. In other words, $\mathbf{b}_i$ is allowed to be arbitrarily correlated with the permanent component, $\mathbf{f}_i$. Condition (23) is similar in spirit to the key assumption in [46] for the intercept c_i: the correlation between the slopes b_{ij} and the entire history $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$ is through the time averages, and not through deviations from the time averages. If $\mathbf{b}_i$ changes across i, ignoring it by using the usual FE estimator effectively puts $\ddot{\mathbf{x}}_{it}(\mathbf{b}_i - \boldsymbol{\beta})$ in the error term, which induces heteroskedasticity and serial correlation in the composite error even if the $\{u_{it}\}$ are homoskedastic and serially

independent. The possible presence of this term provides another argument for making inference with FE fully robust to arbitrary conditional and unconditional second moments.

The (partial) robustness of FE to the presence of correlated random slopes extends to a more general class of estimators that includes the usual fixed effects estimator. Write an extension of the basic model as

$$y_{it} = \mathbf{g}_t \mathbf{a}_i + \mathbf{x}_{it}\mathbf{b}_i + u_{it}\,, \quad t = 1, \ldots, T\,, \tag{24}$$

where $\mathbf{g}_t$ is a set of deterministic functions of time. A leading case is $\mathbf{g}_t = (1, t)$, so that each unit has its own time trend along with a level effect. (The resulting model is sometimes called a *random trend model*). Now, assume that the random coefficients, $\mathbf{a}_i$, are swept away be regressing y_{it} and $\mathbf{x}_{it}$ each on $\mathbf{g}_t$ for each i. The residuals, $\ddot{y}_{it}$ and $\ddot{\mathbf{x}}_{it}$, have had unit-specific trends removed, but the $\mathbf{b}_i$ are treated as constant in the estimation. The key condition for consistently estimating $\boldsymbol{\beta}$ can still be written as in (23), but now $\ddot{\mathbf{x}}_{it}$ has had more features removed at unit-specific level. When $\mathbf{g}_t = (1, t)$, each covariate has been demeaned within each unit. Therefore, if $\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{h}_i t + \mathbf{r}_{it}$, then $\mathbf{b}_i$ can be arbitrarily correlated with $(\mathbf{f}_i, \mathbf{h}_i)$. Of course, individually detrending the $\mathbf{x}_{it}$ requires at least three time periods, and it decreases the variation in $\ddot{\mathbf{x}}_{it}$ compared with the usual FE estimator. Not surprisingly, increasing the dimension of $\mathbf{g}_t$ (subject to the restriction $\dim(\mathbf{g}_t) < T$), generally leads to less precision of the estimator. See [56] for further discussion.

Sequentially Exogenous Regressors and Dynamic Models

The summary of models and estimators from Sect. "Overview of Linear Panel Data Models" used the strict exogeneity assumption $E(u_{it}|\mathbf{x}_i, c_i) = 0$ for all t, and added an additional assumption for models with correlated random slopes. As discussed in Sect. "Overview of Linear Panel Data Models", strict exogeneity is not an especially natural assumption. The contemporaneous exogeneity assumption $E(u_{it}|\mathbf{x}_{it}, c_i) = 0$ is attractive, but the parameters are not identified. In this section, a middle ground between these assumptions, which has been called a *sequential exogeneity assumption*, is used. But first, it is helpful to understand properties of the FE and FD estimators when strict exogeneity fails.

Behavior of Estimators Without Strict Exogeneity

Both the FE and FD estimators are inconsistent (with fixed T, $N \to \infty$) without the strict exogeneity assumption stated in Eq. (4). But it is also pretty well known that, at least under certain assumptions, the FE estimator can be expected to have less "bias" for larger T. Under the contemporaneous exogeneity assumption (2) and the assumption that the data series $\{(\mathbf{x}_{it}, u_{it}) \colon t = 1, \ldots, T\}$ is "weakly dependent" – in time series parlance, "integrated of order zero", or $I(0)$ – then it can be shown that

$$\operatorname{plim} \hat{\boldsymbol{\beta}}_{\mathrm{FE}} = \boldsymbol{\beta} + O(T^{-1}) \tag{25}$$

$$\operatorname{plim} \hat{\boldsymbol{\beta}}_{\mathrm{FD}} = \boldsymbol{\beta} + O(1)\,; \tag{26}$$

see Chap. 11 in [55]. In some very special cases, such as the simple AR(1) model discussed below, the "bias" terms can be calculated, but not generally.

Interestingly, the same results can be shown if $\{\mathbf{x}_{it} \colon t = 1, \ldots, T\}$ has unit roots as long as $\{u_{it}\}$ is $I(0)$ and contemporaneous exogeneity holds. However, there is a catch: if $\{u_{it}\}$ is $I(1)$ – so that the time series version of the "model" would be a spurious regression (y_{it} and $\mathbf{x}_{it}$ are not "cointegrated"), then (25) is no longer true. On the other hand, first differencing means any unit roots are eliminated and so there is little possibility of a spurious regression. The bottom line is that using "large T" approximations such as those in (25) and (26) to choose between FE over FD obligates one to take the time series properties of the panel data seriously; one must recognize the possibility that the FE estimation is essentially a spurious regression.

Consistent Estimation Under Sequential Exogeneity

Because both the FE and FD estimators are inconsistent for fixed T, it makes sense to search for estimators that are consistent for fixed T. A natural specification for dynamic panel data models, and one that allows consistent estimation under certain assumptions, is

$$E(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}, c_i) = E(y_{it}|\mathbf{x}_{it}, c_i) = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i\,, \tag{27}$$

which says that $\mathbf{x}_{it}$ contains enough lags so that further lags of variables are not needed. When the model is written in error form, (27) is the same as

$$E(u_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}, c_i) = 0\,, \quad t = 1, 2, \ldots, T\,. \tag{28}$$

Under (28), the covariates $\{\mathbf{x}_{it} \colon t = 1, \ldots, T\}$ are said to be *sequentially exogenous conditional on* c_i. Some estimation methods are motivated by a weaker version of (28), namely,

$$E(\mathbf{x}_{is}' u_{it}) = \mathbf{0}\,, \quad s = 1, \ldots, t\,, \quad t = 1, \ldots, T\,, \tag{29}$$

but (28) is natural in most applications.

Assumption (28) is appealing in that it allows for finite distributed lag models as well as models with lagged dependent variables. For example, the finite distributed lag model

$$y_{it} = \eta_t + \mathbf{z}_{it}\boldsymbol{\delta}_0 + \mathbf{z}_{i,t-1}\boldsymbol{\delta}_1 + \cdots + \mathbf{z}_{i,t-L}\boldsymbol{\delta}_L + c_i + u_{it} \quad (30)$$

allows the elements of $\mathbf{z}_{it}$ to have effects up to L time periods after a change. With $\mathbf{x}_{it} = (\mathbf{z}_{it}, \mathbf{z}_{i,t-1}, \ldots, \mathbf{z}_{i,t-L})$, Assumption (28) implies

$$\begin{aligned} &E(y_{it}|\mathbf{z}_{it}, \mathbf{z}_{i,t-1}, \mathbf{z}_{i,t-2}, \ldots, c_i) \\ &= E(y_{it}|\mathbf{z}_{it}, \mathbf{z}_{i,t-1}, \mathbf{z}_{i,t-2}, c_i) \\ &= \eta_t + \mathbf{z}_{it}\boldsymbol{\delta}_0 + \mathbf{z}_{i,t-1}\boldsymbol{\delta}_1 + \cdots + \mathbf{z}_{i,t-L}\boldsymbol{\delta}_L + c_i\,, \end{aligned} \quad (31)$$

which means that the distributed lag dynamics are captured by L lags. The important difference with the strict exogeneity assumption is that sequential exogeneity allows feedback from u_{it} to $\mathbf{z}_{ir}$ for $r > t$.

How can (28) be used for estimation? The FD transformation is natural because of the sequential nature of the restrictions. In particular, write the FD equation as

$$\Delta y_{it} = \Delta\mathbf{x}_{it}\boldsymbol{\beta} + \Delta u_{it}\,, \quad t = 2, \ldots, T\,. \quad (32)$$

Then, under (29),

$$E(\mathbf{x}_{is}'\Delta u_{it}) = \mathbf{0}\,, \quad s = 1, \ldots, t-1\,; \quad t = 2, \ldots, T\,, \quad (33)$$

which means any $\mathbf{x}_{is}$ with $s < t$ can be used as an instrument for the time t FD equation. An efficient estimator that uses (33) is obtained by stacking the FD equations as

$$\Delta\mathbf{y}_i = \Delta\mathbf{X}_i\boldsymbol{\beta} + \Delta\mathbf{u}_i\,, \quad (34)$$

where $\Delta\mathbf{y}_i = (\Delta y_{i2}, \Delta y_{i3}, \ldots, \Delta y_{iT})'$ is the $(T-1)\times 1$ vector of first differences and $\Delta\mathbf{X}_i$ is the $(T-1)\times K$ matrix of differences on the regressors. (Time period dummies are absorbed into $\mathbf{x}_{it}$ for notational simplicity.) To apply a system estimation method to (34), define

$$\mathbf{x}_{it}^o \equiv (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{it})\,, \quad (35)$$

which means the valid instruments at time t are in $\mathbf{x}_{i,t-1}^o$ (minus redundancies, of course). The matrix of instruments to apply to (34) is

$$\mathbf{W}_i = \mathrm{diag}(\mathbf{x}_{i1}^o, \mathbf{x}_{i2}^o, \ldots, \mathbf{x}_{i,T-1}^o)\,, \quad (36)$$

which has $T-1$ rows and a large number of columns. Because of sequential exogeneity, the number of valid instruments increases with t.

Given $\mathbf{W}_i$, it is routine to apply generalized method of moments estimation, as summarized in [27,55]. A simpler strategy is available that can be used for comparison or as the first-stage estimator in computing the optimal weighting matrix. First, estimate a reduced form for $\Delta\mathbf{x}_{it}$ separately for each t. In other words, at time t, run the regression $\Delta\mathbf{x}_{it}$ on $\mathbf{x}_{i,t-1}^o$, $i = 1, \ldots, N$, and obtain the fitted values, $\widehat{\Delta\mathbf{x}}_{it}$. Of course, the fitted values are all $1\times K$ vectors for each t, even though the number of available instruments grows with t. Then, estimate the FD Eq. (32) by pooled IV using $\widehat{\Delta\mathbf{x}}_{it}$ as instruments (not regressors). It is simple to obtain robust standard errors and test statistics from such a procedure because the first stage estimation to obtain the instruments can be ignored (asymptotically, of course).

One potential problem with estimating the FD equation using IVs that are simply lags of $\mathbf{x}_{it}$ is that changes in variables over time are often difficult to predict. In other words, $\Delta\mathbf{x}_{it}$ might have little correlation with $\mathbf{x}_{i,t-1}^o$. This is an example of the so-called "weak instruments" problem, which can cause the statistical properties of the IV estimators to be poor and the usual asymptotic inference misleading. Identification is lost entirely if $\mathbf{x}_{it} = \boldsymbol{\lambda}_t + \mathbf{x}_{i,t-1} + \mathbf{q}_{it}$, where $\mathrm{E}(\mathbf{q}_{it}|\mathbf{x}_{i,t-1}, \ldots, \mathbf{x}_{i1}) = \mathbf{0}$ – that is, the elements of $\mathbf{x}_{it}$ are random walks with drift. Then, then $\mathrm{E}(\Delta\mathbf{x}_{it}|\mathbf{x}_{i,t-1}, \ldots, \mathbf{x}_{i1}) = \mathbf{0}$, and the rank condition for IV estimation fails. Of course, if some elements of $\mathbf{x}_{it}$ are strictly exogenous, then their changes act as their own instruments. Nevertheless, typically at least one element of $\mathbf{x}_{it}$ is suspected of failing strict exogeneity, otherwise standard FE or FD would be used.

In situations where simple estimators that impose few assumptions are too imprecise to be useful, sometimes one is willing to improve estimation of $\boldsymbol{\beta}$ by adding more assumptions. How can this be done in the panel data case under sequential exogeneity? There are two common approaches. First, the sequential exogeneity condition can be strengthened to the assumption that the conditional mean model is *dynamically complete*, which can be written in terms of the errors as

$$E(u_{it}|\mathbf{x}_{it}, y_{i,t-1}\mathbf{x}_{i,t-1}, \ldots, y_{i1}, \mathbf{x}_{i1}, c_i) = 0\,, \quad t = 1, \ldots, T\,. \quad (37)$$

Clearly, (37) implies (28). Dynamic completeness is neither stronger nor weaker than strict exogeneity, because the latter includes the entire history of the covariates while (37) conditions only on current and past $\mathbf{x}_{it}$. Dynamic completeness is natural when $\mathbf{x}_{it}$ contains lagged dependent variables, because it basically means enough lags have been included to capture all of the dynamics. It

is often too restrictive in finite distributed lag models such as (30), where (37) would imply

$$E(y_{it}|\mathbf{z}_{it}, y_{i,t-1}\mathbf{z}_{i,t-1}, \ldots, y_{i1}, \mathbf{z}_{i1}, c_i) = E(y_{it}|\mathbf{z}_{it}, \mathbf{z}_{i,t-1}, \ldots, \mathbf{z}_{i-L}, c_i), \quad t = 1, \ldots, T, \tag{38}$$

which puts strong restrictions on the fully dynamic conditional mean: values y_{ir}, $r \leq t-1$, do not help to predict y_{it} once $(\mathbf{z}_{it}, \mathbf{z}_{i,t-1}, \ldots)$ are controlled for. FDLs are of interest even if (38) does not hold. Imposing (37) in FDLs implies that the idiosyncratic errors must be serially uncorrelated, something that is often violated in FDLs.

Dynamic completeness is natural in a model such as

$$y_{it} = \rho y_{i,t-1} + \mathbf{z}_{it}\boldsymbol{\delta}_0 + \mathbf{z}_{i,t-1}\boldsymbol{\delta}_1 + c_i + u_{it}. \tag{39}$$

Usually – although there are exceptions – (39) is supposed to represent the conditional mean $E(y_{it}|\mathbf{z}_{it}, y_{i,t-1}\mathbf{z}_{i,t-1}, \ldots, y_{i1}, \mathbf{z}_{i1}, c_i)$, and then the issue is whether one lag of y_{it} and $\mathbf{z}_{it}$ suffice to capture the dynamics.

Regardless of what is contained in $\mathbf{x}_{it}$, assumption (37) implies some additional moment conditions that can be used to estimate $\boldsymbol{\beta}$. The extra moment conditions, first proposed in [1] in the context of the AR(1) unobserved effects model, can be written as

$$E[(\Delta y_{i,t-1} - \Delta\mathbf{x}_{i,t-1}\boldsymbol{\beta})'(y_{it} - \mathbf{x}_{it}\boldsymbol{\beta})] = \mathbf{0}, \quad t = 3, \ldots, T; \tag{40}$$

see also [9]. The conditions can be used in conjunction with those in Eq. (33) in a method of moments estimation method. In addition to imposing dynamic completeness, the moment conditions in (40) are nonlinear in parameters, which makes them more difficult to implement than just using (33). Nevertheless, the simulation evidence in [1] for the AR(1) model shows that (40) can help considerably when the coefficient ρ is large.

[7] suggested a different set of restrictions,

$$\mathrm{Cov}(\Delta\mathbf{x}_{it}', c_i) = 0, \quad t = 2, \ldots, T. \tag{41}$$

Interestingly, this assumption is very similar in spirit to assumption (23), except that it is in terms of the first difference of the covariates, not the time-demeaned covariates. Condition (41) generates moment conditions in the levels of equation,

$$E\big[\Delta\mathbf{x}_{it}'(y_{it} - \alpha - \mathbf{x}_{it}\boldsymbol{\beta})\big] = \mathbf{0}, \quad t = 2, \ldots, T, \tag{42}$$

where α allows for a nonzero mean for c_i. [10] applies these moment conditions, along with the usual conditions in (33), to estimate firm-level production functions. Because of persistence in the data, they find the moments in (33) are not especially informative for estimating the parameters, whereas (42) along with (33) are. Of course, (42) is an extra set of assumptions.

The previous discussion can be applied to the AR(1) model, which has received much attention. In its simplest form the model is

$$y_{it} = \rho y_{i,t-1} + c_i + u_{it}, \quad t = 1, \ldots, T, \tag{43}$$

so that, by convention, the first observation on y is at $t = 0$. The minimal assumptions imposed are

$$E(y_{is}u_{it}) = 0, \quad s = 0, \ldots, t-1, \quad t = 1, \ldots, T, \tag{44}$$

in which case the available instruments at time t are $\mathbf{w}_{it} = (y_{i0}, \ldots, y_{i,t-2})$ in the FD equation

$$\Delta y_{it} = \rho\Delta y_{i,t-1} + \Delta u_{it}, \quad t = 2, \ldots, T. \tag{45}$$

Written in terms of the parameters and observed data, the moment conditions are

$$E[y_{is}(\Delta y_{it} - \rho\Delta y_{i,t-1}) = 0, \quad s = 0, \ldots, t-2, \quad t = 2, \ldots, T. \tag{46}$$

[4] proposed pooled IV estimation of the FD equation with the single instrument $y_{i,t-2}$ (in which case all $T-1$ periods can be used) or $\Delta y_{i,t-2}$ (in which case only $T-2$ periods can be used). A better approach is pooled IV where $T-1$ separate reduced forms are estimated for $\Delta y_{i,t-1}$ as a linear function of $(y_{i0}, \ldots, y_{i,t-2})$. The fitted values $\widehat{\Delta y}_{i,t-1}$, can be used as the instruments in (45) in a pooled IV estimation. Of course, standard errors and inference should be made robust to the MA(1) serial correlation in Δu_{it}. [6] suggested full GMM estimation using all of the available instruments $(y_{i0}, \ldots, y_{i,t-2})$, and this estimator uses the conditions in (44) efficiently.

Under the dynamic completeness assumption

$$E(u_{it}|y_{i,t-1}, y_{i,t-2}, \ldots, y_{i0}, c_i) = 0, \tag{47}$$

the extra moment conditions in [1] become

$$E[(\Delta y_{i,t-1} - \rho\Delta y_{i,t-2})(y_{it} - \rho y_{i,t-1})] = 0, \quad t = 3, \ldots, T. \tag{48}$$

[10] noted that if the condition

$$\mathrm{Cov}(\Delta y_{i1}, c_i) = \mathrm{Cov}(y_{i1} - y_{i0}, c_i) = 0 \tag{49}$$

is added to (47) then the combined set of moment conditions becomes

$$E[\Delta y_{i,t-1}(y_{it}-\alpha-\rho y_{i,t-1})] = 0\,, \quad t = 2,\ldots,T\,, \quad (50)$$

which can be added to the usual moment conditions (46). Conditions (46) and (50) combined are attractive because they are linear in the parameters, and they can produce much more precise estimates than just using (46).

As discussed in [10], condition (49) can be interpreted as a restriction on the initial condition, y_{i0}, and the steady state. When $|\rho| < 1$, the steady state of the process is $c_i/(1-\rho)$. Then, it can be shown that (49) holds if the deviation of y_{i0} from its steady state is uncorrelated with c_i. Statistically, this condition becomes more useful as ρ approaches one, but this is when the existence of a steady state is most in doubt. [22] shows theoretically that such restrictions can greatly increase the information about ρ.

Other approaches to dynamic models are based on maximum likelihood estimation. Approaches that condition on the initial condition y_{i0}, suggested by [10,13,15], seem especially attractive. Under normality assumptions, maximum likelihood conditional on y_{i0} is tractable.

If some strictly exogenous variables are added to the AR(1) model, then it is easiest to use IV methods on the FD equation, namely,

$$\Delta y_{it} = \rho\Delta y_{i,t-1} + \Delta\mathbf{z}_{it}\boldsymbol{\gamma} + \Delta u_{it}\,, \quad t = 1,\ldots,T\,. \quad (51)$$

The available instruments (in addition to time period dummies) are $(\mathbf{z}_i, y_{i,t-2},\ldots,y_{i0})$, and the extra conditions (42) can be used, too. If sequentially exogenous variables, say $\mathbf{h}_{it}$, are added, then $(\mathbf{h}_{i,t-1},\ldots,\mathbf{h}_{i1})$ would be added to the list of instruments (and $\Delta\mathbf{h}_{it}$ would appear in the equation).

Unbalanced Panel Data Sets

The previous sections considered estimation of models using balanced panel data sets, where each unit is observed in each time period. Often, especially with data at the individual, family, or firm level, data are missing in some time periods – that is, the panel data set is *unbalanced*. Standard methods, such as fixed effects, can often be applied to produce consistent estimators, and most software packages that have built-in panel data routines typically allow unbalanced panels. However, determining whether applying standard methods to the unbalanced panel produces consistent estimators requires knowing something about the mechanism generating the missing data.

Methods based on removing the unobserved effect warrant special attention, as they allow some nonrandomness in the sample selection. Let $t = 1,\ldots,T$ denote the time periods for which data can exist for each unit from the population, and again consider the model

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}\,, \quad t = 1,\ldots,T\,. \quad (52)$$

It is helpful to have, for each i and t, a binary selection variable, s_{it}, equal to one of the data for unit i in time t can be used, and zero otherwise. For concreteness, consider the case where time averages are removed to eliminate c_i, but where the averages necessarily only include the $s_{it} = 1$ observations. Let $\ddot{y}_{it} = y_{it} - T_i^{-1}\sum_{r=1}^{T} s_{ir}y_{ir}$ and $\ddot{\mathbf{x}}_{it} = \mathbf{x}_{it} - T_i^{-1}\sum_{r=1}^{T} s_{ir}\mathbf{x}_{ir}$ be the time-demeaned quantities using the observed time periods for unit i, where $T_i = \sum_{t=1}^{T} s_{it}$ is the number of time periods observed for unit i – properly viewed as a random variable. The fixed effects estimator on the unbalanced panel can be expressed as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\mathrm{FE}} &= \left(N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}\right)^{-1} \cdot \left(N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}_{it}'\ddot{y}_{it}\right) \\ &= \boldsymbol{\beta} + \left(N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}\right)^{-1} \cdot \left(N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}_{it}'u_{it}\right)\,.\end{aligned} \quad (53)$$

With fixed T and $N \to \infty$ asymptotics, the key condition for consistency is

$$\sum_{t=1}^{T} E(s_{it}\ddot{\mathbf{x}}_{it}'u_{it}) = \mathbf{0}\,. \quad (54)$$

In evaluating (54), it is important to remember that $\ddot{\mathbf{x}}_{it}$ depends on $(\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT}, s_{i1},\ldots,s_{iT})$, and in a nonlinear way. Therefore, it is not sufficient to assume $(\mathbf{x}_{ir}, s_{ir})$ are uncorrelated with u_{it} for all r and t. A condition that is sufficient for (54) is

$$E(u_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT}, s_{i1},\ldots,s_{iT}, c_i) = 0\,, \quad t = 1,\ldots,T\,. \quad (55)$$

Importantly, (55) allows arbitrary correlation between the heterogeneity, c_i, and selection, s_{it}, in any time period t. In other words, some units are allowed to be more likely to

be in or out of the sample in any time period, and these probabilities can change across t. But (55) rules out some important kinds of sample selection. For example, selection at time t, s_{it}, cannot be correlated with the idiosyncratic error at time t, u_{it}. Further, feedback is not allowed: in affect, like the covariates, selection must be strictly exogenous conditional on c_i.

Testing for no feedback into selection is easy in the context of FE estimation. Under (55), $s_{i,t+1}$ and u_{it} should be uncorrelated. Therefore, $s_{i,t+1}$ can be added to the FE estimation on the unbalanced panel – where the last time period is lost for all observations – and a t test can be used to determine significance. A rejection means (55) is false. Because serial correlation and heteroskedasticity are always a possibility, the t test should be made fully robust.

Contemporaneous selection bias – that is, correlation between s_{it} and u_{it} – is more difficult to test. Chapter 17 in [55] summarizes how to derive tests and corrections by extending the corrections in [28] (so-called "Heckman corrections") to panel data.

First differencing can be used on unbalanced panels, too, although straight first differencing can result in many lost observations: a time period is used only if it is observed along with the previous or next time period. FD is more useful in the case of attrition in panel data, where a unit is observed until it drops out of the sample and never reappears. Then, if a data point is observed at time t, it is also observed at time $t-1$. Differencing can be combined with the approach in [28] to solve bias due to attrition – at least under certain assumptions. See Chap. 17 in [55].

Random effects methods can also be applied with unbalanced panels, but the assumptions under which the RE estimator is consistent are stronger than for FE. In addition to (55), one must assume selection is unrelated to c_i. A natural assumption, that also imposes exogeneity on the covariates with respect to c_i, is

$$E(c_i|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT},s_{i1},\ldots,s_{iT}) = E(c_i)\,. \tag{56}$$

The only case beside randomly determined sample selection where (56) holds is when s_{it} is essentially a function of the observed covariates. Even in this case, (56) requires that the unobserved heterogeneity is mean independent of the observed covariates – as in the typical RE analysis on balanced panel.

Nonlinear Models

Nonlinear panel data models are considerably more difficult to interpret and estimate than linear models. Key issues concern how the unobserved heterogeneity appears in the model and how one accounts for that heterogeneity in summarizing the effects of the explanatory variables on the response. Also, in some cases, conditional independence of the response is used to identify certain parameters and quantities.

Basic Issues and Quantities of Interest

As in the linear case, the setup here is best suited for situations with small T and large N. In particular, the asymptotic analysis underlying the discussion of estimation is with fixed T and $N \to \infty$. Sampling is assumed to be random from the population. Unbalanced panels are generally difficult to deal with because, except in special cases, the unobserved heterogeneity cannot be completely eliminated in obtaining estimating equations. Consequently, methods that model the conditional distribution of the heterogeneity conditional on the entire history of the covariates – as we saw with the Chamberlain–Mundlak approach – are relied on heavily, and such approaches are difficult when data are missing on the covariates for some time periods. Therefore, this section considers only balanced panels. The discussion here takes the response variable, y_{it}, as a scalar for simplicity.

The starting point for nonlinear panel data models with unobserved heterogeneity is the conditional distribution

$$D(y_{it}|\mathbf{x}_{it},\mathbf{c}_i)\,, \tag{57}$$

where $\mathbf{c}_i$ is the unobserved heterogeneity for observation i drawn along with the observables. Often there is a particular feature of this distribution, such as $E(y_{it}|\mathbf{x}_{it},\mathbf{c}_i)$, or a conditional median, that is of primary interest. Even focusing on the conditional mean raises some tricky issues in models where $\mathbf{c}_i$ does not appear in an additive or linear form. To be precise, let $E(y_{it}|\mathbf{x}_{it} = \mathbf{x}_t, \mathbf{c}_i = \mathbf{c}) = m_t(\mathbf{x}_t,\mathbf{c})$ be the mean function. If x_{tj} is continuous, then the partial effect can be defined as

$$\theta_j(\mathbf{x}_t,\mathbf{c}) \equiv \frac{\partial m_t(\mathbf{x}_t,\mathbf{c})}{\partial x_{tj}}\,. \tag{58}$$

For discrete (or continuous) variables, (58) can be replaced with discrete changes. Either way, a key question is: How can one account for the unobserved $\mathbf{c}$ in (58)? In order to estimate magnitudes of effects, sensible values of $\mathbf{c}$ need to be plugged into (58), which means knowledge of at least some distributional features of $\mathbf{c}_i$ is needed. For example, suppose $\boldsymbol{\mu}_\mathbf{c} = E(\mathbf{c}_i)$ is identified. Then the *partial effect at the average* (PEA),

$$\theta_j(\mathbf{x}_t,\boldsymbol{\mu}_\mathbf{c})\,, \tag{59}$$

can be identified if the regression function m_t is identified. Given more information about the distribution of $\mathbf{c}_i$, different quantiles can be inserted into (59), or a certain number of standard deviations from the mean.

An alternative to plugging in specific values for $\mathbf{c}$ is to average the partial effects across the distribution of $\mathbf{c}_i$:

$$\mathrm{APE}(\mathbf{x}_t) = E_{\mathbf{c}_i}[\theta_j(\mathbf{x}_t, \mathbf{c}_i)] \,, \tag{60}$$

the so-called *average partial effect* (APE). The difference between (59) and (60) can be nontrivial for nonlinear mean functions. The definition in (60) dates back at least to [17], and is closely related to the notion of the *average structural function* (ASF), as introduced in [12]. The ASF is defined as

$$\mathrm{ASF}(\mathbf{x}_t) = E_{\mathbf{c}_i}[m_t(\mathbf{x}_t, \mathbf{c}_i)] \,. \tag{61}$$

Assuming the derivative passes through the expectation results in (60); computing a discrete change in the ASF always gives the corresponding APE. A useful feature of APEs is that they can be compared across models, where the functional form of the mean or the distribution of the heterogeneity can be different. In particular, APEs in general nonlinear models are comparable to the estimated coefficients in a standard linear model.

Average partial effects are not always identified, even when parameters are. Semi-parametric panel data methods that are silent about the distribution of $\mathbf{c}_i$, unconditionally or conditional on $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$, cannot generally deliver estimates of APEs, essentially by design. Instead, an index structure is usually imposed so that parameters can be consistently estimated. A common setup with scalar heterogeneity is

$$m_t(\mathbf{x}_t, c) = G(\mathbf{x}_t\boldsymbol{\beta} + c) \,, \tag{62}$$

where, say, $G(\cdot)$ is strictly increasing and continuously differentiable. The partial effects are proportional to the parameters:

$$\theta_j(\mathbf{x}_t, c) = \beta_j g(\mathbf{x}_t\boldsymbol{\beta} + c) \,, \tag{63}$$

where $g(\cdot)$ is the derivative of $G(\cdot)$. Therefore, if β_j is identified, then so is the sign of the partial effect, and even the relative effects of any two continuous variables: the ratio of partial effects for x_{tj} and x_{th} is β_j/β_h. However, even if $G(\cdot)$ is specified (the common case), the magnitude of the effect evidently cannot be estimated without making assumptions about the distribution of c_i; otherwise, the term $E[g(\mathbf{x}_t\boldsymbol{\beta} + c_i)]$ cannot generally be estimated. The probit example below shows how the APEs can be estimated in index models under distributional assumptions for c_i.

The previous discussion holds regardless of the exogeneity assumptions on the covariates. For example, the definition of the APE for a continuous variable holds whether $\mathbf{x}_t$ contains lagged dependent variables or only contemporaneous variables. However, approaches for estimating the parameters and the APEs depend critically on exogeneity assumptions.

Exogeneity Assumptions on the Covariates

As in the case of linear models, it is not nearly enough to simply specify a model for the conditional distribution of interest, $D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$, or some feature of it, in order to estimate parameters and partial effects. This section offers two exogeneity assumptions on the covariates that are more restrictive versions of the linear model assumptions.

It is easiest to deal with estimation under a strict exogeneity assumption. The most useful definition of strict exogeneity for nonlinear panel data models is

$$D(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \mathbf{c}_i) = D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i) \,, \tag{64}$$

which means that $\mathbf{x}_{ir}$, $r \neq t$, does not appear in the conditional distribution of y_{it} once $\mathbf{x}_{it}$ and $\mathbf{c}_i$ have been counted for. [17] labeled (64) *strict exogeneity conditional on the unobserved effects* $\mathbf{c}_i$. Sometimes, a conditional mean version is sufficient:

$$E(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, \mathbf{c}_i) = E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i) \,, \tag{65}$$

which already played a role in linear models. Assumption (64), or its conditional mean version, are less restrictive than if $\mathbf{c}_i$ is not in the conditioning set, as discussed in [17]. Indeed, it is easy to see that, if (64) holds and $D(\mathbf{c}_i|\mathbf{x}_i)$ depends on $\mathbf{x}_i$, then strict exogeneity without conditioning on $\mathbf{c}_i$, $D(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}) = D(y_{it}|\mathbf{x}_{it})$, cannot hold. Unfortunately, both (64) and (65) rule out lagged dependent variables, as well as other situations where there may be feedback from idiosyncratic changes in y_{it} to future movements in $\mathbf{x}_{ir}$, $r > t$. Nevertheless, the conditional strict exogeneity assumption underlies the most common estimation methods for nonlinear models.

More natural is *sequential exogeneity conditional on the unobserved effects*, which, in terms of conditional distributions, is

$$D(y_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}, \mathbf{c}_i) = D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i) \,. \tag{66}$$

Assumption (66) allows for lagged dependent variables and does not restrict feedback. Unfortunately, (66) is substantially more difficult to work with than (64) for general nonlinear models.

Because $\mathbf{x}_{it}$ is conditioned on, neither (64) nor (66) allows for contemporaneous endogeneity of $\mathbf{x}_{it}$ as would arise with measurement error, time-varying omitted variables, or simultaneous equations. This chapter does not treat such cases. See [38] for a recent summary.

Conditional Independence Assumption

The exogeneity conditions stated in Subsect. "Exogeneity Assumptions on the Covariates" generally do not restrict the dependence in the responses, $\{y_{it}: t = 1, \ldots, T\}$. Often, a *conditional independence* assumption is explicitly imposed, which can be written generally as

$$D(y_{i1}, \ldots, y_{iT}|\mathbf{x}_i, \mathbf{c}_i) = \prod_{t=1}^{T} D(y_{it}|\mathbf{x}_i, \mathbf{c}_i) . \tag{67}$$

Equation (67) simply means that, conditional on the entire history $\{\mathbf{x}_{it}: t = 1, \ldots, T\}$ and the unobserved heterogeneity $\mathbf{c}_i$, the responses are independent across time. One way to think about (67) is that time-varying unobservables are independent over time. Because (67) conditions on $\mathbf{x}_i$, it is useful only in the context of the strict exogeneity assumption (64). Then, conditional independence can be written as

$$D(y_{i1}, \ldots, y_{iT}|\mathbf{x}_i, \mathbf{c}_i) = \prod_{t=1}^{T} D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i) . \tag{68}$$

Therefore, under strict exogeneity and conditional independence, the panel data modeling exercise reduces to specifying a model for $D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$, and then determining how to treat the unobserved heterogeneity, $\mathbf{c}_i$. In random effects and correlated RE frameworks, conditional independence can play a critical role in being able to estimate the parameters and the distribution of $\mathbf{c}_i$. As it turns out, conditional independence plays no role in estimating APEs for a broad class of models. Before explaining how that works, the key issue of dependence between the heterogeneity and covariates needs to be addressed.

Assumptions About the Unobserved Heterogeneity

For general nonlinear models, the *random effects assumption* is independence between $\mathbf{c}_i$ and $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$:

$$D(\mathbf{c}_i|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}) = D(\mathbf{c}_i) . \tag{69}$$

Assumption (69) is very strong. To illustrate how strong it is, suppose that (69) is combined with a model for the conditional mean, $E(y_{it}|\mathbf{x}_{it} = \mathbf{x}_t, \mathbf{c}_i = \mathbf{c}) = m_t(\mathbf{x}_t, \mathbf{c})$. Without any additional assumptions, the average partial effects are nonparametrically identified. In particular, the APEs can be obtained directly from the conditional mean

$$r_t(\mathbf{x}_t) \equiv E(y_{it}|\mathbf{x}_{it} = \mathbf{x}_t) . \tag{70}$$

(The argument is a simple application of the law of iterated expectations; it is discussed in [56]). Nevertheless, (69) is still common in many applications, especially when the explanatory variables of interest do not change over time.

As in the linear case, a *correlated random effects* (CRE) framework allows dependence between $\mathbf{c}_i$ and $\mathbf{x}_i$, but the dependence in restricted in some way. In a parametric setting, a CRE approach involves specifying a distribution for $D(\mathbf{c}_i|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$, as in [15,17,46], and many subsequent authors; see, for example, [55] and [14]. For many models – see, for example, Subsect. "Binary Response Models" – one can allow $D(\mathbf{c}_i|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$ to depend on $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$ in a "nonexchangeable" manner, that is, the distribution need not be symmetric on its conditioning arguments. However, allowing nonexchangeability usually comes at the expense of potentially restrictive distributional assumptions, such as homoskedastic normal with a linear conditional mean. For estimating APEs, it is sufficient to assume, along with strict exogeneity,

$$D(\mathbf{c}_i|\mathbf{x}_i) = D(\mathbf{c}_i|\bar{\mathbf{x}}_i) , \tag{71}$$

without specifying $D(\mathbf{c}_i|\bar{\mathbf{x}}_i)$ or restricting any feature of this distribution. (See, for example, [3,56].) As a practical matter, it makes sense to adopt (71) – or perhaps allow other features of $\{\mathbf{x}_{it}: t = 1, \ldots, T\}$ – in a flexible parametric analysis.

Condition (71) still imposes restrictions on $D(\mathbf{c}_i|\mathbf{x}_i)$. Ideally, as in the linear model, one could estimate at least some features of interest without making any assumption about $D(\mathbf{c}_i|\mathbf{x}_i)$. Unfortunately, the scope for allowing unrestricted $D(\mathbf{c}_i|\mathbf{x}_i)$ is limited to special nonlinear models, at least with small T. Allowing $D(\mathbf{c}_i|\mathbf{x}_i)$ to be unspecified is the hallmark of a "fixed effects" analysis, but the label has not been used consistently. Often, fixed effects has been used to describe a situation where the $\mathbf{c}_i$ are treated as parameters to be estimated, along with parameters that do not vary across i. Except in special cases or with large T, estimating the unobserved heterogeneity is prone to an *incidental parameters problem*. Namely, using a fixed T, $N \to \infty$ framework, one cannot get consistent estimators of the $\mathbf{c}_i$, and the inconsistency in, say, $\hat{\mathbf{c}}_i$, generally transmits itself to the parameters that do not vary with i. The incidental parameters problem does not arise in estimating the coefficients $\boldsymbol{\beta}$ in a linear model because the estimator obtained by treating the c_i as parameters to estimate is equivalent to pooled OLS on the time-demeaned data –

that is, the fixed effects estimator can be obtained by eliminating the c_i using the within transformation or estimating the c_i along with $\boldsymbol{\beta}$. This occurrence is rare in nonlinear models. Section "Future Directions" further discusses this issue, as there is much ongoing research that attempts to reduce the asymptotic bias in nonlinear models.

The "fixed effects" label has also been applied to settings where the $\mathbf{c}_i$ are not treated as parameters to estimate; rather, the $\mathbf{c}_i$ can be eliminated by conditioning on a sufficient statistic. Let $\mathbf{w}_i$ be a function of the observed data, $(\mathbf{x}_i, \mathbf{y}_i)$, such that

$$D(y_{i1}, \ldots, y_{it}|\mathbf{x}_i, \mathbf{c}_i, \mathbf{w}_i) = D(y_{i1}, \ldots, y_{it}|\mathbf{x}_i, \mathbf{w}_i) . \quad (72)$$

Then, provided the latter conditional distribution depends on the parameters of interest, and can be derived or approximated from the original specification of $D(y_{i1}, \ldots, y_{it}|\mathbf{x}_i, \mathbf{c}_i)$, maximum likelihood methods can be used. Such an approach is also called *conditional maximum likelihood estimation* (CMLE), where "conditional" refers to conditioning on a function of $\mathbf{y}_i$. (In traditional treatments of MLE, conditioning on so-called "exogenous" variables is usually implicit.) In most cases where the CMLE approach applies, the conditional independence assumption (67) is maintained, although one conditional MLE is known to have robustness properties: the so-called "fixed effects" Poisson estimator (see [53]).

Maximum Likelihood Estimation and Partial MLE

There are two common approaches to estimating the parameters in nonlinear, unobserved effects panel data models when the explanatory variables are strictly exogenous. (A third approach, generalized method of moments, is available in special cases but is not treated here. See, for example, Chap. 19 in [55].) The first approach is full maximum likelihood (conditional on the entire history of covariates). Most commonly, full MLE is applied under the conditional independence assumption, although sometimes models are used that explicitly allow dependence in $D(y_{i1}, \ldots, y_{iT}|\mathbf{x}_i, \mathbf{c}_i)$. Assuming strict exogeneity, conditional independence, a model for the density of y_{it} given $(\mathbf{x}_{it}, \mathbf{c}_i)$ (say, $f_t(y_t|\mathbf{x}_t, \mathbf{c}; \theta)$), and a model for the density of $\mathbf{c}_i$ given $\mathbf{x}_i$ (say, $h(\mathbf{c}|\mathbf{x}; \boldsymbol{\delta})$), the log likelihood for random draw i from the cross section is

$$\log\left\{\left[\int \prod_{t=1}^{T} f_t(y_{it}|\mathbf{x}_{it}, \mathbf{c}; \theta)\right] h(\mathbf{c}|\mathbf{x}_i; \boldsymbol{\delta})\mathrm{d}\mathbf{c}\right\} . \quad (73)$$

This log-likelihood function "integrates out" the unobserved heterogeneity to obtain the joint density of $(y_{i1}, \ldots, y_{iT})$ conditional on $\mathbf{x}_i$. In the most commonly applied models, including logit, probit, Tobit, and various count models (such as the Poisson model), the log likelihood in (73) identifies all of the parameters. Computation can be expensive but is typically tractable. The main methodological drawback to the full MLE approach is that it is not robust to violations of the conditional independence assumption, except for the linear model where normal conditional distributions are used for y_{it} and c_i.

The *partial* MLE ignores temporal dependence in the responses when estimating the parameters – at least when the parameters are identified. In particular, obtain the density of y_{it} given $\mathbf{x}_i$ by integrating the marginal density for y_{it} against the density for the heterogeneity:

$$g_t(y_t|\mathbf{x}; \theta, \boldsymbol{\delta}) = \int f_t(y_t|\mathbf{x}_t, \mathbf{c}; \theta) h(\mathbf{c}|\mathbf{x}; \boldsymbol{\delta})\,\mathrm{d}\mathbf{c} . \quad (74)$$

The *partial* MLE (PMLE) (or pooled MLE) uses, for each i, the partial log likelihood

$$\sum_{t=1}^{T} \log[g_t(y_{it}|\mathbf{x}_i; \theta, \boldsymbol{\delta}) . \quad (75)$$

Because the partial MLE ignores the serial dependence caused by the presence of $\mathbf{c}_i$, it is essentially never efficient. But in leading cases, such as probit, Tobit, and Poisson models, $g_t(y_t|\mathbf{x}; \theta, \boldsymbol{\delta})$ has a simple form when $h(\mathbf{c}|\mathbf{x}; \boldsymbol{\delta})$ is chosen judiciously. Further, the PMLE is fully robust to violations of (67). Inference is complicated by the neglected serial dependence, but an appropriate adjustment to the asymptotic variance is easily obtained; see Chap. 13 in [55].

One complication with PMLE is that in the cases where it leads to a simple analysis (probit, ordered probit, and Tobit, to name a few), not all of the parameters in θ and $\boldsymbol{\delta}$ are separately identified. The conditional independence assumption and the use of full MLE serves to identify all parameters. Fortunately, the PMLE does identify the parameters that index the average partial effects, a claim that will be verified for the probit model in Subsect. "Binary Response Models".

Dynamic Models

General models with only sequentially exogenous variables are difficult to estimate. [8] considered binary response models and [54] suggested a general strategy that requires modeling the dynamic distribution of the variables that are not strictly exogenous.

Much more is known about the specific case where the model contains lagged dependent variables along with strictly exogenous variables. The starting point is a model for the dynamic distribution,

$$D(y_{it}|\mathbf{z}_{it}, y_{i,t-1}, \mathbf{z}_{i,t-1} \ldots, y_{i1}, \mathbf{z}_{i1}, y_{i0}, \mathbf{c}_i)\,, \quad t = 1, \ldots, T\,, \tag{76}$$

where $\mathbf{z}_{it}$ are variables strictly exogenous (conditional on $\mathbf{c}_i$) in the sense that

$$D(y_{it}|\mathbf{z}_i, y_{i,t-1}, \mathbf{z}_{i,t-1} \ldots, y_{i1}, \mathbf{z}_{i1}, y_{i0}, \mathbf{c}_i) = D(\mathbf{y}_{it}|\mathbf{z}_{it}, y_{i,t-1}, \mathbf{z}_{i,t-1} \ldots, y_{i1}, \mathbf{z}_{i1}, y_{i0}, \mathbf{c}_i)\,, \tag{77}$$

where $\mathbf{z}_i$ is the entire history $\{\mathbf{z}_{it}\colon t = 1, \ldots, T\}$.

In the leading case, (76) depends only on $(\mathbf{z}_{it}, y_{i,t-1}, \mathbf{c}_i)$ (although putting lags of strictly exogenous variables only slightly changes the notation). Let $f_t(y_t|\mathbf{z}_t, y_{t-1}, \mathbf{c}; \theta)$ denote a model for the conditional density, which depends on parameters θ. The joint density of $(y_{i1}, \ldots, y_{iT})$ given $(y_{i0}, \mathbf{z}_i, \mathbf{c}_i)$ is

$$\prod_{t=1}^{T} f_t(y_t|\mathbf{z}_t, y_{t-1}, \mathbf{c}; \theta)\,. \tag{78}$$

The problem with using (78) for estimation is that, when it is turned into a log likelihood by plugging in the "data", $\mathbf{c}_i$ must be inserted. Plus, the log likelihood depends on the initial condition, y_{i0}. Several approaches have been suggested to address these problems: (i) Treat the $\mathbf{c}_i$ as parameters to estimate (which results in an incidental parameters problem). (ii) Try to estimate the parameters without specifying conditional or unconditional distributions for c_i. (This approach is available for very limited situations, and other restrictions are needed. And, generally, one cannot estimate average partial effects.) (iii) Find, or, more practically, approximate $D(y_{i0}|\mathbf{c}_i, z_i)$ and then model $D(\mathbf{c}_i|\mathbf{z}_i)$. Integrating out $\mathbf{c}_i$ gives the density for $D(y_{i0}, y_{i1}, \ldots, y_{iT}|\mathbf{z}_i)$, which can be used in an MLE analysis (conditional on $\mathbf{z}_i$), (iv) Model $D(\mathbf{c}_i|y_{i0}, \mathbf{z}_i)$. Then, integrate out $\mathbf{c}_i$ conditional on $(y_{i0}, \mathbf{z}_i)$ to obtain the density for $D(y_{i1}, \ldots, y_{iT}|y_{i0}, \mathbf{z}_i)$. Now, MLE is conditional on $(y_{i0}, \mathbf{z}_i)$. As shown by [57], in some leading cases – probit, ordered probit, Tobit, Poisson regression – there is a density $h(\mathbf{c}|y_0, \mathbf{z})$ that mixes with the density $f(y_1, \ldots, y_T|y_0, \mathbf{z}, \mathbf{c})$ to produce a log-likelihood that is in a common family and programmed in standard software packages.

If $m_t(\mathbf{x}_t, \mathbf{c}, \theta)$ is the mean function $E(y_t|\mathbf{x}_t, \mathbf{c})$, with $\mathbf{x}_t = (\mathbf{z}_t, y_{t-1})$, then APEs are easy to obtain. The average structural function is

$$\begin{aligned}\mathrm{ASF}(\mathbf{x}_t) &= E_{\mathbf{c}_i}[m_t(\mathbf{x}_t, \mathbf{c}_i, \theta)] \\ &= E\left\{\left[\int m_t(\mathbf{x}_t, \mathbf{c}, \theta) h(\mathbf{c}|y_{i0}, \mathbf{z}_i, \boldsymbol{\gamma}) \mathrm{d}\mathbf{c}\right] |y_{i0}, \mathbf{z}_i\right\}.\end{aligned} \tag{79}$$

The term inside the brackets, say $r_t(\mathbf{x}_t, y_{i0}, \mathbf{z}_i, \theta, \boldsymbol{\gamma})$ is available, at least in principle, because $m_t()$ and $h()$ have been specified. Often, they have simple forms, or they can be simulated. A consistent estimator of the ASF is obtained by averaging out $(y_{i0}, \mathbf{z}_i)$:

$$\widehat{\mathrm{ASF}}(\mathbf{x}_t) = N^{-1} \sum_{t=1}^{T} r_t(\mathbf{x}_t, y_{i0}, \mathbf{z}_i, \hat{\theta}, \hat{\boldsymbol{\gamma}})\,. \tag{80}$$

Partial derivatives and differences with respect to elements of $\mathbf{x}_t$ (which, remember, includes functions of y_{t-1}) can be computed. With large N and small T, the panel data bootstrap – where resampling is carried out in the cross section so that every time period is kept when a unit i is resampled – can be used for standard errors and inference. The properties of the nonparametric bootstrap are standard in this setting because the resampling is carried out in the cross section.

Binary Response Models

Unobserved effects models – static and dynamic – have been estimated for various kinds of response variables, including binary responses, ordered responses, count data, and corner solutions. Most of the issues outlined above can be illustrated by binary response models, which is the topic of this subsection.

The standard specification for the unobserved effects (UE) probit model is

$$P(y_{it} = 1|\mathbf{x}_{it}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i)\,, \quad t = 1, \ldots, T, \tag{81}$$

where $\mathbf{x}_{it}$ does not contain an overall intercept but would usually include time dummies, and c_i is the scalar heterogeneity. Without further assumptions, neither $\boldsymbol{\beta}$ nor the APEs are identified. The traditional RE probit model imposes a strong set of assumptions: strict exogeneity, conditional independence, and independence between c_i and $\mathbf{x}_i$ with $c_i \sim \mathrm{Normal}(\mu_c, \sigma_c^2)$. Under these assumptions, $\boldsymbol{\beta}$ and the parameters in the distribution of c_i are identified and are consistently estimated by full MLE (conditional on $\mathbf{x}_i$).

Under the strict exogeneity assumption (64), a correlated random effects version of the model is obtained from

the Chamberlain–Mundlak device under conditional normality:

$$c_i = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i, a_i|\mathbf{x}_i \sim \text{Normal}(0, \sigma_a^2)\,. \tag{82}$$

The less restrictive version $c_i = \psi + \mathbf{x}_i\boldsymbol{\xi} + a_i = \psi + \mathbf{x}_{i1}\boldsymbol{\xi}_1 + \cdots + \mathbf{x}_{iT}\boldsymbol{\xi}_T + a_i$ can be used, but the time average conserves on degrees of freedom.

As an example, suppose that y_{it} is a binary variable indicating whether firm i in year t was awarded at least one patent, and the key explanatory variable in $\mathbf{x}_{it}$ is current and past spending on research and development (R&D). It makes sense that R&D spending is correlated, at least on average, with unobserved firm heterogeneity, and so a correlated random effects model seems natural. Unfortunately, the strict exogeneity assumption might be problematical: it could be that being awarded a patent in year t might affect future values of spending on R&D. Most studies assume this is not the case, but one should be aware that, as in the linear case, the strict exogeneity assumption imposes restrictions on economic behavior.

When the conditional independence assumption (67) is added to (81), strict exogeneity, and (82), all parameters in (81) and (82) are identified (assuming that all elements of $\mathbf{x}_{it}$ are time-varying) and the parameters can be efficiently estimated by maximum likelihood (conditional on $\mathbf{x}_i$). Afterwards, the mean of c_i can be consistently estimated as $\hat{\mu}_c = \hat{\psi} + \left(N^{-1}\sum_{i=1}^N \bar{\mathbf{x}}_i\right)\hat{\boldsymbol{\xi}}$ and the variance as $\hat{\sigma}_c^2 = \hat{\boldsymbol{\xi}}'\left(N^{-1}\sum_{i=1}^N \bar{\mathbf{x}}_i'\bar{\mathbf{x}}_i\right)\hat{\boldsymbol{\xi}} + \hat{\sigma}_a^2$. Because a_i is normally distributed, c_i is not normally distributed unless $\bar{\mathbf{x}}_i\boldsymbol{\xi}$ is. A normal approximation for $D(c_i)$ gets better as T gets large. In any case, the estimated mean and standard deviation can be used to plug in values of c that are a certain number of estimated standard deviations from $\hat{\mu}_c$, say $\hat{\mu}_c \pm \hat{\sigma}_c$ or $\hat{\mu}_c \pm 2\hat{\sigma}_c$.

The APEs are identified from the ASF, which is consistently estimated by

$$\widehat{\text{ASF}}(\mathbf{x}_t) = N^{-1}\sum_{i=1}^N \Phi(\mathbf{x}_t\hat{\boldsymbol{\beta}}_a + \hat{\psi}_a + \bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}_a) \tag{83}$$

where the "a" subscript means that a coefficient has been divided by $(1+\hat{\sigma}_a^2)^{1/2}$, for example, $\hat{\boldsymbol{\beta}}_a = \hat{\boldsymbol{\beta}}/(1+\hat{\sigma}_a^2)^{1/2}$. The derivatives or changes of $\widehat{\text{ASF}}(\mathbf{x}_t)$ with respect to elements of $\mathbf{x}_t$ can be compared with fixed effects estimates from a linear model. Often, to obtain a single scale factor, a further averaging across $\mathbf{x}_{it}$ is done. The APEs computed from such averaging can be compared to linear FE estimates.

The CRE probit model is an example of a model where the APEs are identified without the conditional independence assumption. Without (67) – or any restriction on the joint distribution – it can still be shown that

$$P(y_{it} = 1|\mathbf{x}_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta}_a + \psi_a + \bar{\mathbf{x}}_i\boldsymbol{\xi}_a)\,, \tag{84}$$

which means a number of estimation approaches identify the scaled coefficients $\boldsymbol{\beta}_a$, ψ_a, and $\boldsymbol{\xi}_a$. The estimates of these scaled coefficients can be inserted directly into (83). The unscaled parameters and σ_a^2 are not separately identified, but in most cases this is a small price to pay for relaxing the conditional independence assumption. Note that for determining directions of effects and relative effects, $\boldsymbol{\beta}_a$ is just as useful as $\boldsymbol{\beta}$. Plus, it is $\boldsymbol{\beta}_a$ that appears in the APEs. The partial effects at the mean value of c_i are not identified.

Using pooled probit can be inefficient for estimating the scaled parameters. Full MLE, with a specified correlation matrix for the $T \times 1$ vector $\mathbf{u}_i$, is possible in principle but difficult in practice. An alternative approach, the *generalized estimating equations* (GEE) approach, can be more efficient than pooled probit but just as robust in that only (84) is needed for consistency. See [38] for a summary of how GEE – which is essentially the same as multivariate weighted nonlinear least squares – applies to the CRE probit model.

A simple test of the strict exogeneity assumption is to add selected elements of $\mathbf{x}_{i,t+1}$, say $\mathbf{w}_{i,t+1}$, to the model and computing a test of joint significance. Unless the full MLE is used, the test should be made robust to serial dependence of unknown form. For example, as a test of strict exogeneity of R&D spending when y_{it} is a patent indicator, one can just include next year's value of R&D spending and compute a t test. In carrying out the test, the last time period is lost for all firms.

Because there is nothing sacred about the standard model (81) under (82) – indeed, these assumptions are potentially quite restrictive – it is natural to pursue other models and assumptions. Even with (81) as the starting point, and under strict exogeneity, there are no known ways of identifying parameters or partial effects without restricting $D(c_i|\mathbf{x}_i)$. Nevertheless, as mentioned in Subsect. "Assumptions About the Unobserved Heterogeneity", there are nonparametric restrictions on $D(c_i|\mathbf{x}_i)$ that do identify the APEs under strict exogeneity – even if (81) is dropped entirely. As shown in [3], the restriction $D(c_i|\mathbf{x}_i) = D(c_i|\bar{\mathbf{x}}_i)$ identifies the APEs. While fully nonparametric methods can be used, some simple strategies are evident. For example, because the APEs can be obtained from $D(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$, it makes sense to apply flexible parametric models directly to this distribution – without worrying about the original models for $D(y_{it}|\mathbf{x}_{it}, c_i)$ and $D(c_i|\mathbf{x}_i)$.

As an example of this approach, a flexible parametric model, such as

$$P(y_{it} = 1|\mathbf{x}_{it}, \bar{\mathbf{x}}_i) = \Phi[\theta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\gamma} + (\bar{\mathbf{x}}_i \otimes \bar{\mathbf{x}}_i)\boldsymbol{\delta} + (\mathbf{x}_{it} \otimes \bar{\mathbf{x}}_i)\boldsymbol{\eta}], \tag{85}$$

might provide a reasonable approximation. The average structural function is estimated as

$$\widehat{\mathrm{ASF}}(\mathbf{x}_t) = N^{-1}\sum_{i=1}^{N} \Phi[\hat{\theta}_t + \mathbf{x}_t\hat{\boldsymbol{\beta}} + \bar{\mathbf{x}}_i\hat{\boldsymbol{\gamma}} + (\bar{\mathbf{x}}_i \otimes \bar{\mathbf{x}}_i)\hat{\boldsymbol{\delta}} + (\mathbf{x}_t \otimes \bar{\mathbf{x}}_i)\hat{\boldsymbol{\eta}}], \tag{86}$$

where the estimates can come from pooled MLE, GEE, or a method of moments procedure. The point is that extensions of the basic probit model such as (85) can provide considerable flexibility and deliver good estimators of the APEs. The drawback is that one has to be willing to abandon standard underlying models for $P(y_{it} = 1|\mathbf{x}_{it}, c_i)$ and $D(c_i|\mathbf{x}_i)$; in fact, it seems very difficult to characterize models for these two features that would lead to an expression such as (85).

An alternative model for the response probability is the logit model

$$P(y_{it} = 1|\mathbf{x}_{it}, c_i) = \Lambda(\mathbf{x}_{it}\boldsymbol{\beta} + c_i), \tag{87}$$

where $\Lambda(z) = \exp(z)/[1 + \exp(z)]$. In cross section applications, researchers often find few practical differences between (81) and (87). But when unobserved heterogeneity is added in a panel data context, the logit formulation has an advantage: under the conditional independence assumption (and strict exogeneity), the parameters $\boldsymbol{\beta}$ can be consistently estimated, with a $\sqrt{N}$-asymptotic normal distribution, without restricting $D(c_i|\mathbf{x}_i)$. The method works by conditioning on the number of "successes" for each unit, that is, $n_i = \sum_{t=1}^{T} y_{it}$. [17] shows that $D(y_{i1}, \ldots, y_{iT}|\mathbf{x}_i, c_i, n_i) = D(y_{i1}, \ldots, y_{iT}|\mathbf{x}_i, n_i)$, and the latter depends on $\boldsymbol{\beta}$ (at least when all elements of $\mathbf{x}_{it}$ are time varying). The conditional MLE – sometimes called the "fixed effects logit" estimator – is asymptotically efficient in the class of estimators putting no assumptions on $D(c_i|\mathbf{x}_i)$. While this feature of the logit CMLE is attractive, the method has two drawbacks. First, it does not appear to be robust to violations of the conditional independence assumption, and little is known about the practical effects of serial dependence in $D(y_{i1}, \ldots, y_{iT}|\mathbf{x}_i, c_i)$. Second, and perhaps more importantly, because $D(c_i|\mathbf{x}_i)$ and $D(c_i)$ are not restricted, it is not clear how one estimates magnitudes of the effects of the covariates on the response probability. The logit CMLE is intended to estimate the parameters, which means the effects of the covariates on the log-odds ratio, $\log\{[P(y_{it} = 1|\mathbf{x}_{it}, c_i)]/[1 - P(y_{it} = 1|\mathbf{x}_{it}, c_i)]\} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i$, can be estimated. But the magnitudes of the effects of covariates on the response probability are not available. Therefore, there are tradeoffs when choosing between CRE probit and "fixed effects" logit: the CRE probit identifies average partial effects with or without the conditional independence assumptions, at the cost of specifying $D(c_i|\mathbf{x}_i)$, while the FE logit estimates parameters without specifying $D(c_i|\mathbf{x}_i)$, but requires conditional independence and still does not deliver estimates of partial effects. As often is the case in econometrics, there are tradeoffs between assumptions between the logit and probit approaches, and also tradeoffs. See [38] for further discussion.

Estimation of parameters and APEs is more difficult in simple dynamic probit models. Consider

$$P(y_{it} = 1|\mathbf{z}_{it}, y_{i,t-1}, c_i) = \Phi(\mathbf{z}_{it}\boldsymbol{\delta} + \rho y_{i,t-1} + c_i), \tag{88}$$

which assumes first-order dynamics and strict exogeneity of $\{\mathbf{z}_{it}: t = 1, \ldots, T\}$. Treating the c_i as parameters to estimate causes inconsistency in $\boldsymbol{\delta}$ and ρ because of the incidental parameters problem. A simple analysis is available under the assumption

$$c_i|y_{i0}, \mathbf{z}_i \sim \mathrm{Normal}(\psi + \xi_0 y_{i0} + \mathbf{z}_i\boldsymbol{\xi}, \sigma_a^2). \tag{89}$$

Then,

$$P(y_{it} = 1|\mathbf{z}_i, y_{i,t-1}, \ldots, y_{i0}, a_i) = \Phi(\mathbf{z}_{it}\boldsymbol{\delta} + \rho y_{i,t-1} + \psi + \xi_0 y_{i0} + \mathbf{z}_i\boldsymbol{\xi} + a_i), \tag{90}$$

where $a_i \equiv c_i - \psi - \xi_0 y_{i0} - \mathbf{z}_i\boldsymbol{\xi}$. Because a_i is independent of $(y_{i0}, \mathbf{z}_i)$, it turns out that standard random effects probit software can be used, with explanatory variables $(1, \mathbf{z}_{it}, y_{i,t-1}, y_{i0}, \mathbf{z}_i)$ in time period t. All parameters, including σ_a^2, are consistently estimated, and the ASF is estimated by averaging out $(y_{i0}, \mathbf{z}_i)$:

$$\widehat{\mathrm{ASF}}(\mathbf{z}_t, y_{t-1}) = N^{-1}\sum_{i=1}^{N} \Phi(\mathbf{z}_t\hat{\boldsymbol{\delta}}_a + \hat{\rho}_a y_{t-1} + \hat{\psi}_a + \hat{\xi}_{a0} y_{i0} + \mathbf{z}_i\hat{\boldsymbol{\xi}}_a), \tag{91}$$

where the coefficients are multiplied by $(1 + \hat{\sigma}_a^2)^{-1/2}$. APEs are gotten, as usual, by taking differences or derivatives with respect to elements of $(\mathbf{z}_t, y_{t-1})$. Both (88) and

the model for $D(c_i|y_{i0}, \mathbf{z}_i)$ can be made more flexible (such as including interactions, or letting $\mathrm{Var}(c_i|\mathbf{z}_i, y_{i0})$ be heteroskedastic). See [57] for further discussion.

Similar analyses hold for other nonlinear models, although the particulars differ. For count data, maximum likelihood methods are available – based on correlated random effects or conditioning on a sufficient statistic. In this case, the CMLE based on the Poisson distribution has very satisfying robustness properties, requiring only the conditional mean in the unobserved effects model to be correctly specified along with strict exogeneity (Conditional independence is not needed). These and dynamic count models are discussed in Chap. 19 in [55,57].

Correlated random effects Tobit models are specified and estimated in a manner very similar to CRE probit models; see Chap. 16 in [55]. Unfortunately, there are no known conditional MLEs that eliminate the unobserved heterogeneity in Tobit models. Nevertheless, [33,34] show how the parameters in models for corner solutions can be estimated without distributional assumptions on $D(c_i|\mathbf{x}_i)$. Such methods do place exchangeability restrictions on $D(y_{i1}, \ldots, y_{iT}|\mathbf{x}_i, c_i)$, but they are not as strong as conditional independence with identical distributions.

Future Directions

Research in panel data methods continues unabated. Dynamic linear models are a subject of ongoing interest. The problem of feedback in linear models when the covariates are persistent – and the weak instrument problem that it entails – is important for panels with small T. For example, with firm-level panel data, the number of time periods is typically small and inputs into a production function would often be well-approximated as random walks with perhaps additional short-term dependence. The estimators described in Sect. "Sequentially Exogenous Regressors and Dynamic Models" that impose additional assumptions should be studied when those assumptions fail. Perhaps the lower variance of the estimators from the misspecified model is worth the additional bias.

Models with random coefficients, especially when those random coefficients are on non-strictly exogenous variables (such as lagged dependent variables), have received some attention, but many of the proposed solutions require large T. (See, for example, [49,50]). An alternative approach is flexible MLE, as in [57], where one models the distribution of heterogeneity conditional on the initial condition and the history of covariates. See [19] for any application to dynamic product choice.

When T is large enough so that it makes sense to use large-sample approximations with large T, as well as large N, one must make explicit assumptions about the time series dependence in the data. Such frameworks are sensible for modeling large geographical units, such as states, provinces, or countries, where long stretches of time are observed. The same estimators that are attractive for the fixed T case, particularly fixed effects, can have good properties when T grows with N, but the properties depend on whether unit-specific effects, time-specific effects, or both are included. The rates at which T and N are assumed to grow also affect the large-sample approximations. See [52] for a survey of linear model methods with T and N are both assumed to grow in the asymptotic analysis. A recent study that considers estimation when the data have unit roots is [44]. Unlike the fixed T case, a unified theory for linear models, let alone nonlinear models, remains elusive when T grows with N and is an important area for future research.

In the models surveyed here, a single coefficient is assumed for the unobserved heterogeneity, whereas the effect might change over time. In the linear model, the additive c_i can be replaced with $\psi_t c_i$ (with $\psi_1 = 1$ as a normalization). For example, the return to unobserved managerial talent in a firm production function can change over time. Conditions under which ignoring the time-varying loads, ψ_t, and using the usual fixed effects estimator, consistently estimates the coefficients on $\mathbf{x}_{it}$ are given in [47]. But one can also estimate the ψ_t along with $\boldsymbol{\beta}$ using method of moments frameworks. Examples are [2,32]. An area for future research is to allow heterogeneous slopes on observed covariates along with time-varying loads on the unobserved heterogeneity. Allowing for time-varying loads and heterogeneous slopes in nonlinear models can allow for significant flexibility, but only parametric approaches to estimation have been studied.

There is considerable interest in estimating production functions using proxy variables, such as investment, for time-varying, unobserved productivity. The pioneering work is [48]; see also [42]. Estimation in this case does not rely on differencing or time-demeaning to remove unobserved heterogeneity, and so the estimates can be considerably more precise than the FE or FD estimators. But the assumption that a deterministic function of investment can proxy for unobserved productivity is strong. [11] provides an analysis that explicitly allows for unobserved heterogeneity and non-strictly exogenous inputs using the methods described in Sect. "Sequentially Exogenous Regressors and Dynamic Models". An interesting challenge for future researchers is to unify the two approaches to exploit the attractive features of each.

The parametric correlated random effects approach for both static and dynamic nonlinear models is now fairly

well understood in the balanced case. Much less attention has been paid to the unbalanced case, and missing data, especially for fully dynamic models, is a serious challenge. [57] discusses the assumptions under which using a balanced subset produces consistent estimates.

Identification of average partial effects (equivalently, the average structural function) has recently received the attention that it deserves, although little is known about how robust are the estimated APEs under various misspecifications of parametric models. One might hope that using flexible models for nonlinear responses might provide good approximations, but evidence on this issue is lacking.

As mentioned earlier, recent research in [3] has shown how to identify and estimate partial effects without making parametric assumptions about $E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$ or $D(\mathbf{c}_i|\mathbf{x}_i)$. The setup in [3] allows for $D(\mathbf{c}_i|\mathbf{x}_i)$ to depend on $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$ in an exchangeable way. The simplest case is the one given in (71), $D(\mathbf{c}_i|\mathbf{x}_i) = D(\mathbf{c}_i|\bar{\mathbf{x}}_i)$. Under (71) and the strict exogeneity assumption $E(y_{it}|\mathbf{x}_i, \mathbf{c}_i) = E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$, the average structural function is identified as

$$\mathrm{ASF}_t(\mathbf{x}_t) = E_{\bar{\mathbf{x}}_i}[r_t(\mathbf{x}_t, \bar{\mathbf{x}}_i)] \,, \tag{92}$$

where $r_t(\mathbf{x}_{it}, \bar{\mathbf{x}}_i) = E(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$. Because $r_t(\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ can be estimated very generally – even using nonparametric regression of y_{it} on $(\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ for each t – the average partial effects can be estimated without any parametric assumptions. Research in [3] shows how $D(\mathbf{c}_i|\mathbf{x}_i)$ can depend on other exchangeable functions of $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$, such as sample variances and covariances. As discussed in [38], nonexchangeable functions, such as trends and growth rates, can be accommodated, provided these functions are known. For example, for each i, let $(\hat{\mathbf{f}}_i, \hat{\mathbf{g}}_i)$ be the vectors of intercepts and slopes from the regression $\mathbf{x}_{it}$ on $1, t$, $t = 1, \ldots, T$. Then, an extension of (71) is $D(c_i|\mathbf{x}_i) = D(c_i|\hat{\mathbf{f}}_i, \hat{\mathbf{g}}_i)$. It appears these kinds of assumptions have not yet been applied, but they are a fertile area for future research because they extend the typical CRE setup.

Future research on nonlinear models will likely consider the issue of the kinds of partial effects that are of most interest. [3] studies identification and estimation of the *local average response* (LAR). The LAR at $\mathbf{x}_t$ for a continuous variable x_{tj} is

$$\int \frac{\partial m_t(\mathbf{x}_t, \mathbf{c})}{\partial x_{tj}} \mathrm{d}H_t(\mathbf{c}|\mathbf{x}_t) \,, \tag{93}$$

where $m_t(\mathbf{x}_t, \mathbf{c})$ is the conditional mean of the response and $H_t(\mathbf{c}|\mathbf{x}_t)$ denotes the cdf of $D(\mathbf{c}_i|\mathbf{x}_{it} = \mathbf{x}_t)$. This is a "local" partial effect because it averages out the heterogeneity for the slice of the population described by the vector of observed covariates, $\mathbf{x}_t$. The APE averages out over the entire distribution of $\mathbf{c}_i$, and therefore can be called a "global average response". See also [21]. The results in [3] include general identification results for the LAR, and future empirical researchers using nonlinear panel data models may find the local nature of the LAR more appealing (although more difficult to estimate) than APEs.

A different branch of the panel data literature has studied identification of coefficients or, more often, scaled coefficients, in nonlinear models. For example, [35] shows how to estimate $\boldsymbol{\beta}$ in the model

$$y_{it} = 1[w_{it} + \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it} \geq 0] \tag{94}$$

without distributional assumptions on the composite error, $c_i + u_{it}$. In this model, w_{it} is a special continuous explanatory variable (which need not be time varying). Because its coefficient is normalized to unity, w_{it} necessarily affects the response, y_{it}. More importantly, w_{it} is assumed to satisfy the distributional restriction $D(c_i + u_{it}|w_{it}, \mathbf{x}_{it}, \mathbf{z}_i) = D(c_i + u_{it}|\mathbf{x}_{it}, \mathbf{z}_i)$, which is a conditional independence assumption. The vector $\mathbf{z}_i$ is assumed to be independent of u_{it} in all time periods. (So, if two time periods are used, $\mathbf{z}_i$ could be functions of variables determined prior to the earliest time period). The most likely scenario where the framework in [35] applies is when w_{it} is randomized and therefore independent of the entire vector $(\mathbf{x}_{it}, \mathbf{z}_i, c_i + u_{it})$. The key condition seems unlikely to hold if w_{it} is related to past outcomes on y_{it}. The estimator of $\boldsymbol{\beta}$ derived in [35] is $\sqrt{N}$-asymptotically normal, and fairly easy to compute; it requires estimation of the density of w_{it} given $(\mathbf{x}_{it}, \mathbf{z}_i)$ and then a simple IV estimation. Essentially by construction, estimation of partial effects on the response probability is not possible.

Recently, [36] shows how to obtain bounds on parameters and APEs in dynamic models, including the dynamic probit model in Eq. (85) under the strict exogeneity assumption on $\{\mathbf{z}_{it}\colon t = 1, \ldots, T\}$. A further assumption is that c_i and $\mathbf{z}_i$ are independent. By putting restrictions on $D(c_i)$ – which nevertheless allow flexibility – [36] explains how to estimate bounds for the unknown ρ. The bounds allow one to determine how much information are in the data when few assumptions are made. Similar calculations can be made for average partial effects, so that the size of so-called state dependence – the difference between $E_{c_i}[\Phi(\mathbf{z}_t\boldsymbol{\delta} + \rho + c_i) - \Phi(\mathbf{z}_t\boldsymbol{\delta} + c_i)]$ – can be bounded.

Because CRE methods require some restriction on the distribution of heterogeneity, and estimation of scaled coefficients leaves partial effects unidentified, the theoretical literature has returned to the properties of parameter estimates and partial effects when the heterogeneity

is treated as unit-specific parameters to estimate. Recent work has focused on adjusting the "fixed effects" estimates (of the common population parameters) so that they have reduced bias.

An emerging question is whether the average partial effects might be estimated well even though the parameters themselves are biased. In other words, suppose that for a nonlinear model one obtains $\{\hat{\theta}, \hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2, \ldots, \hat{\mathbf{c}}_N\}$, typically by maximizing a pooled log-likelihood function across all i and t. If $m_t(\mathbf{x}_t, \mathbf{c}, \theta,) = E(y_t|\mathbf{x}_t, \mathbf{c})$ is the conditional mean function, the average partial effects can be estimated as

$$N^{-1} \sum_{i=1}^{N} \frac{\partial m_t(\mathbf{x}_t, \hat{\mathbf{c}}_i, \hat{\theta})}{\partial x_{tj}} . \tag{95}$$

In the unobserved effects probit model, (95) becomes

$$N^{-1} \sum_{i=1}^{N} \hat{\beta}_j \phi(\mathbf{x}_t \hat{\boldsymbol{\beta}} + \hat{c}_i) . \tag{96}$$

[20] studied the properties of (96) with strictly exogenous regressors under conditional independence, assuming that the covariates are weakly dependent over time. Interestingly, the bias in (96) is of order T^{-2} when there is no heterogeneity, which suggests that estimating the unobserved effects might not be especially harmful when the amount of heterogeneity is small. Unfortunately, these findings do not carry over to models with time heterogeneity or lagged dependent variables. While bias corrections are available, they are difficult to implement.

[24] proposes both jackknife and analytical bias corrections and show that they work well for the probit case. Generally, the jackknife procedure to remove the bias in $\hat{\theta}$ is simple but can be computationally intensive. The idea is this. The estimator based on T time periods has probability limit (as $N \to \infty$) that can be written as

$$\theta_T = \theta + \mathbf{b}_1/T + \mathbf{b}_2/T^2 + O(T^{-3}) \tag{97}$$

for vectors $\mathbf{b}_1$ and $\mathbf{b}_2$. Now, let $\hat{\theta}_{(t)}$ denote the estimator that drops time period t. Then, assuming stability across t, it can be shown that the jackknife estimator,

$$\tilde{\theta} = T\hat{\theta} - (T-1)T^{-1} \sum_{t=1}^{T} \hat{\theta}_{(t)} \tag{98}$$

has asymptotic bias of $\tilde{\theta}$ on the order of T^{-2}.

Unfortunately, there are currently some practical limitations to the jackknife procedure, as well as to the analytical corrections derived in [24]. First, aggregate time effects are not allowed, and they would be very difficult to include because the analysis is with $T \to \infty$. (In other words, time effects would introduce an incidental parameters problem in the time dimension, in addition to the incidental parameters problem in the cross section). Plus, heterogeneity in the distribution of the response y_{it} across t changes the bias terms $\mathbf{b}_1$ and $\mathbf{b}_2$ when a time period is dropped, and so the adjustment in (98) does not remove the bias terms. Second, [24] assumes independence across t conditional on c_i. It is a traditional assumption, but in static models it is often violated, and it must be violated in dynamic models. Plus, even without time heterogeneity, the jackknife does not apply to dynamic models; see [23].

Another area that has seen a resurgence is so-called pseudo panel data, as initially exposited in [18]. A pseudo-panel data set is constructed from repeated cross sections across time, where the units appearing in each cross section are not repeated (or, if they are, it is a coincidence and is ignored). If there is a natural grouping of the cross-sectional units – for example, for individuals, birth year cohorts – one can create a pseudo-panel data set by constructing group or cohort averages in each time period. With relatively few cohorts and large cross sections, one can identify pseudo panels in the context of minimum distance estimation. With a large number of groups, a different large-sample analysis might be warranted. A recent contribution is [39] and [38] includes a recent survey. Open questions include the most efficient way to use the full set of restrictions in the underlying individual-level model.

As mentioned earlier, this chapter did not consider panel data model with explanatory variables that are endogenous in the sense that they are correlated with time-varying unobservables. For linear models, the usual fixed effects and first differencing transformations can be combined with instrumental variables methods. In nonlinear models, the Chamberlain–Mundlak approach can be combined with so-called "control function" methods, provided the endogenous explanatory variables are continuous. [38] includes a discussion of some recent advances for complicated models such as multinomial response models; see also [51]. Generally, structural estimation in discrete response models with unobserved heterogeneity and endogenous explanatory variables is an area of great interest.

Bibliography

Primary Literature

1. Ahn SC, Schmidt P (1995) Efficient estimation of models for dynamic panel data. J Econom 68:5–27

2. Ahn SC, Lee YH, Schmidt P (2001) GMM estimation of linear panel data models with time-varying individual effects. J Econom 101:219–255
3. Altonji JG, Matzkin RL (2005) Cross section and panel data estimators for nonseparable models with endogenous regressors. Econometrica 73:1053–1102
4. Anderson TW, Hsiao C (1982) Formulation and estimation of dynamic models using panel data. J Econom 18:47–82
5. Arellano M (1993) On the testing of correlated effects with panel data. J Econom 59:87–97
6. Arellano M, Bond SR (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. Rev Econ Stud 58:277–297
7. Arellano M, Bover O (1995) Another look at the instrumental variable estimation of error components models. J Econom 68:29–51
8. Arellano M, Carrasco R (2003) Binary choice panel data models with predetermined variables. J Econom 115:125–157
9. Arellano M, Honoré B (2001) Panel data models: Some recent developments. In: Heckman JJ, Leamer E (eds) Handbook of econometrics, vol 5. North Holland, Amsterdam, pp 3229–3296
10. Blundell R, Bond SR (1998) Initial conditions and moment restrictions in dynamic panel data models. J Econom 87:115–143
11. Blundell R, Bond SR (2000) GMM estimation with persistent panel data: An application to production functions. Econom Rev 19:321–340
12. Blundell R, Powell JL (2003) Endogeneity in nonparametric and semiparametric regression models. In: Dewatripont M, Hansen LP, Turnovsky SJ (eds) Advances in economics and econonometrics: Theory and applications, 8th World Congress, vol 2. Cambridge University Press, Cambridge, pp 312–357
13. Blundell R, Smith RJ (1991) Initial conditions and efficient estimation in dynamic panel data models – an application to company investment behaviours. Ann d'écon stat 20–21:109–124
14. Cameron AC, Trivedi PK (2005) Microeconometrics: Methods and applications. Cambridge University Press, Cambridge
15. Chamberlain G (1980) Analysis of covariance with qualitative data. Rev Econ Stud 47:225–238
16. Chamberlain G (1982) Multivariate regression models for panel data. J Econom 1:5–46
17. Chamberlain G (1984) Panel data. In: Griliches Z, Intriligator MD (eds) Handbook of econometrics, vol 2. North Holland, Amsterdam, pp 1248–1318
18. Deaton A (1985) Panel Data from time series of cross-sections. J Econom 30:109–126
19. Erdem T, Sun B (2001) Testing for choice dynamics in panel data. J Bus Econ Stat 19:142–152
20. Fernández-Val I (2007) Fixed effects estimation of structural parameters and marginal effects in panel probit models. Mimeo. Boston University, Department of Economics
21. Florens JP, Heckman JJ, Meghir C, Vytlacil E (2007) Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. Mimeo. Columbia University, Department of Economics
22. Hahn J (1999) How informative is the initial condition in the dynamic panel model with fixed effects? J Econom 93:309–326
23. Hahn J, Kuersteiner G (2002) Asymptotically unbiased inference for a dynamic panel model with fixed effects when both *n* and *T* are large. Econometrica 70:1639–1657
24. Hahn J, Newey WK (2004) Jackknife and analytical bias reduction for nonlinear panel models. Econometrica 72:1295–1319
25. Hausman JA (1978) Specification tests in econometrics. Econometrica 46:1251–1271
26. Hausman JA, Taylor WE (1981) Panel data and unobservable individual effects. Econometrica 49:1377–1398
27. Hayashi F (2000) Econometrics. Princeton University Press, Princeton
28. Heckman JJ (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Ann Econ Soc Meas 5:475–492
29. Heckman JJ (1981) Statistical models for discrete panel data. In: Manski CF, McFadden DL (eds) Structural analysis of discrete data and econometric applications. MIT Press, Cambridge, pp 114–178
30. Heckman JJ (1981) The incidental parameters problem and the problem of initial condition in estimating a discrete time-discrete data stochastic process. In: Manski CF, McFadden DL (eds) Structural analysis of discrete data and econometric applications. MIT Press, Cambridge, pp 179–195
31. Hoch I (1962) Estimation of production function parameters combining time-series and cross-section data. Econometrica 30:34–53
32. Holtz-Eakin D, Newey W, Rosen HS (1988) Estimating vector autoregressions with panel data. Econometrica 56:1371–1395
33. Honoré BE (1992) Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects. Econometrica 60:533–565
34. Honoré BE, Hu L (2004) Estimation of cross sectional and panel data censored regression models with endogeneity. J Econom 122:293–316
35. Honoré BE, Lewbel A (2002) Semiparametric binary choice panel data models without strictly exogeneous regressors. Econometrica 70:2053–2063
36. Honoré BE, Tamer E (2006) Bounds on parameters in panel dynamic discrete choice models. Econometrica 74:611–629
37. Im KS, Ahn SC, Schmidt P, Wooldridge JM (1999) Efficient estimation of panel data models with strictly exogenous explanatory variables. J Econom 93:177–201
38. Imbens GW, Wooldridge JM (2007) What's new in econometrics? Lecture Notes. National Bureau of Economic Research, Summer Institute
39. Inoue A (2008) Efficient estimation and inference in linear pseudo-panel data models. J Econom 142:449–466
40. Keane MP, Runkle DE (1992) On the estimation of panel-data models with serial correlation when instruments are not strictly exogenous. J Bus Econ Stat 10:1–9
41. Kiefer NM (1980) Estimation of fixed effect models for time series of cross-sections with arbitrary intertemporal covariance. J Econom 14:195–202
42. Levinshohn J, Petrin A (2003) Estimating production functions using inputs to control for unobservables. Rev Econ Stud 70:317–341
43. Liang K-Y, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22
44. Moon HR, Phillips PCB (2004) GMM estimation of autoregressive roots near unity with panel data. Econometrica 72:467–522
45. Mundlak Y (1961) Empirical production function free of management bias. Farm J Econ 43:44–56

46. Mundlak Y (1978) On the pooling of time series and cross section data. Econometrica 46:69–85
47. Murtazashvili I, Wooldridge JM (2007) Fixed effects instrumental variables estimation in correlated random coefficient panel data models. J Econom 142:539–552
48. Olley S, Pakes A (1996) The dynamics of productivity in the telecommunications equipment industry. Econometrica 64:1263–1298
49. Pesaran MH, Takashi Y (2008) Testing slope homogeneity in large panels. J Econom 142:50–93
50. Pesaran MH, Smith RP, Im KS (1996) Dynamic linear models for heterogeneous panels. In: Mátáyas L, Sevestre P (eds) The econometrics of panel data. Kluwer, Dordrecht, pp 145–195
51. Petrin A, Train KE (2005) Tests for omitted attributes in differentiated product models. Mimeo. University of Minnesota, Department of Economics
52. Phillips PCB, Moon HR (2000) Nonstationary panel data analysis: An overview of some recent developments. Econom Rev 19:263–286
53. Wooldridge JM (1999) Distribution-free estimation of some nonlinear panel data models. J Econom 90:77–97
54. Wooldridge JM (2000) A framework for estimating dynamic, unobserved effects panel data models with possible feedback to future explanatory variables. Econ Lett 68:245–250
55. Wooldridge JM (2002) Econometric analysis of cross section and panel data. MIT Press, Cambridge
56. Wooldridge JM (2005) Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models. Rev Econ Stat 87:385–390
57. Wooldridge JM (2005) Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. J Appl Econom 20:39–54

Books and Reviews

Arellano M (2003) Panel data econometrics. Oxford University Press, Oxford
Baltagi BH (2001) Econometric analysis of panel data, vol 2e. Wiley, New York
Hsiao C (2003) Analysis of panel data, vol 2e. Cambridge University Press, Cambridge

Econophysics, Observational

Bertrand M. Roehner
Institute for Theoretical and High Energy Physics, University of Paris 7, Paris, France

Article Outline

Glossary
Definition of the Subject
Introduction
The Primacy of Observation
Investigating One Effect at a Time
What Guidance Can Physics Provide?
How Cross-National Observations Can Be Used to Test the Role of Different Factors
How Vested Interests May Affect the Accessibility and Reliability of Social Data
How Can Exogenous Factors be Taken into Account?
Future Directions
Bibliography

Glossary

Cross national comparisons Comparing cross-national data for a specific phenomenon, e. g. a surge in housing prices, is the key to distinguishing between essential factors which are common to all episodes and those which are accessory and context dependent.

Economathematicians Mathematicians or theoretical physicists who develop mathematical tools, models or simulations for social phenomena but do not try to confront these models to actual observations.

Econophysics A field of physics which originated in the mid-1990s. Throughout this article, we use the term in a broad sense which includes econophysics, sociophysics and historiophysics. As a matter of fact, these fields can hardly be studied separately in the sense that economic effects depend upon social reactions (e. g. reactions of consumers to advertising campaigns); furthermore, economic investigations crucially rely on statistics which typically must combine present-day data with data from former historical episodes.

Econophysicists Physicists who study social, economic or political issues.

Endogenous mechanisms Models usually describe endogenous mechanisms. For instance a population model would describe how people get married and have children.

Exogenous factors Exogenous factors are more or less unexpected external forces which act on the system. Thus, for a population wars or epidemics may bring about sudden population changes. It is only when exogenous factors are recurrent and fairly repetitive that they can be taken into account in models.

Experiment Apart from its standard meaning in physics or biology we also use this term to designate the process of (i) defining the phenomenon that one wants to study (ii) locating and collecting the data which are best suited for the investigation (iii) analyzing the data and deriving *regularity rules* or testing a model.

Model testing Before confronting the predictions of a model to statistical evidence it is necessary to ensure that the system was not subject to unexpected exogenous shocks. The impact of exogenous factors which

are not accounted for in the model must in some way be removed, that is to say the data must be corrected in a way which takes these shocks out of the picture. Usually, such corrections are very tricky to implement.

Definition of the Subject

> "No science thrives in the atmosphere of direct practical aim. We should still be without most of the conveniences of modern life if physicists had been as eager for immediate applications as most economists are and always have been." (J. Schumpeter p.6 in [11])

> "The free fall is a very trivial physical phenomenon, but it was the study of this exceedingly simple fact and its comparison with the astronomical material which brought forth mechanics. The sound procedure [in every science] is to obtain first utmost precision and mastery in a limited field, and then to proceed to another, somewhat wider one and so on." (J. von Neumann and O. Morgenstern [5])

These two quotes define fairly well the path that econophysics tries to follow. They both insist on the fact that one should begin by focusing on simple phenomena even if at first sight they have little practical implications. In what follows we will develop this point but first of all we must address a question which comes to the mind of all persons who hear about econophysics for the first time, namely:

> "Why should physicists have something to say about economic and social phenomena. Admittedly, biology can benefit from physics because of the means of observation [e. g. exploration of protein molecules by X-ray scattering] that it provides, but there are no similar needs in economics."

I have heard this question asked repeatedly by many of my colleagues. In my answer I usually emphasize that what matters is more the method of investigation than the phenomena by themselves. I stress that applying to the social sciences the experimental methodology invented by physicists and chemists would mark a great progress. However, with the benefit of insight, I realize that these answers may have appeared far fetched and unconvincing to many of my listeners. A better and more factual claim is to observe that over the past century several of the most renowned economists and sociologists were in fact econophysicists in the sense defined in the glossary. Indeed, back in the nineteenth century, the only way to get a decent mathematical training was to study astronomy, engineering, mathematics or physics. When such people entered the social sciences this lead to two kinds of approaches which we may designate as econophysics and economathematics (see Sect. "Glossary"). In the first category one may mention the astronomer Adolphe Quételet (1796–1874), Clément Juglar (1819–1905) educated as a medical doctor, Vilfredo Pareto (1848–1923) educated as an engineer, the mathematician Louis Bachelier (1870–1946), the physicist Elliott Montroll (1916–1983), the mathematician Benoît Mandelbrot (1924-). In the second category one may mention Léon Walras (1834–1910) who was educated as an engineer, the astronomer Simon Newcomb (1835–1905), the physicist Maurice Allais (1911–).

Of course, if the economic discipline had been highly successful there would be little need for an alternative approach. However, great doubts have been expressed by some of the most renowned economists about the attainments of their discipline. We have already cited Schumpeter's opinion on this matter. In addition one may mention the judgments of Vassily Leontief, Anna Schwartz, Lawrence Summers or the thesis developed in a recent book by Masanao Aoki and Hiroshi Yoshikawa.

- Leontief and Schwartz emphasized that the present organization of economic research discourages observational research. In Schwartz's words [12][1]

  > "The main disincentive to improve the handling and use of data is that the profession withholds recognition to those who devote their energies to measurement. Someone who introduces an innovation in econometrics, by contrast, will win plaudits."

- The assessment made by Summers in a paper published in 1991 is well summarized by its title: "The scientific illusion of empirical macroeconomics".
- In their book, Aoki and Yoshikawa (p. 25 in [1]) point out that the representative agent assumption which is supposed to provide a connection between micro- and macroeconomics is fundamentally flawed because it neglects both social variability and stochastic fluctuations. It may be true that in recent years a greater emphasis has been put on the issue of heterogeneity. Yet, is this the right way to takle the problem? A model is a simplification of reality anyway, so if it is not tenable to use loosely defined representative agents, an alternative solution may be to focus on sharply defined agent's

[1] Leontief (p. xi in [3]) has even stronger words: "The methods used to maintain intellectual discipline in this country's most influential economics departments can occasionally remind one of those employed by the Marines to maintain discipline on Parris Island [a training camp of US Marines]."

attitudes. For instance, whereas without further specification home buyers may not be well defined as a useful category, the behavior of investors during the final phases of speculative price peaks may be sufficiently recurrent to make up for a well defined category.

Introduction

What are the main characteristics of econophysics? In what follows we will try to summarize some basic principles. Each of them will be illustrated by one or several studies performed by econophysicists over the past decade. Although the wording that we use is fairly personal, we believe that fundamentally these principles are shared by many econophysicists. In the course of more than a decade, econophysics has become a big tree with many branches. Obviously it is impossible to describe all of them if only because the knowledge and understanding of the present author is limited. He apologizes in advance for his limitations and for the fact that the present selection is by necessity fairly subjective.

The Primacy of Observation

Econophysics started around 1995 in sync with the creation of huge computerized databases giving minute by minute transactions on financial markets such as the New York stock market, the dollar-yen exchange rate, the forward interest rates or providing individual income data for millions of people. It may be estimated that between 1995 and 2005 about two thirds of the papers published by econophysicists aimed at deriving *regularity rules* from such databases. Let us illustrate this point by the case of income data. Since Pareto's work we know that the distribution of high incomes can be described by a power law with an exponent α comprised between 1 and 1.5. With databases comprising millions of income data one can get high accuracy estimates for α and observe how α changes as the result of economic booms or stock market crashes. It turns out that α decreases during booms and increases in the wake of stock market collapses [6].

Other empirical investigations were carried out in the past decades. We list some of them below. The list is arranged by topic and by research teams.

- Stock transactions, (i) Boston University: see publications involving G. Stanley. (ii) CEA (i. e. Commissariat à l'Energie Atomique which means Institute for Atomic Research) and "Science-Finance": see publications involving J.P. Bouchaud. (iii) Nice University and UCLA: see publications involving D. Sornette. (iv) University of Warsaw: see publications involving J. Kertesz.
- Forward interest rates, Singapore University: see publications involving B. Baaquie.
- Exchange rates, Zurich: see publications involving M. Dacorogna.

To many physicists the statement that observation is supreme could seem self evident. In economics, however, such a statement represents a revolution. We already mentioned the fact that observation is a neglected topic in economics. As a matter of fact, before econophysics started it was impossible to publish a paper which would identify *regularity rules* without at the same time providing a model[2].

Investigating One Effect at a Time

In most natural phenomena different effects occur simultaneously. For instance, if one leaves a glass of cold water in the sun, the water will of course get warmer but if one looks at the mechanisms which are implied this involves many different effects: interaction of light and water, interaction of light and glass, conduction of heat, creation of convection currents between layers of water which are at different temperatures, and so on. One of the main challenges of physics was to identify these effects and to study them separately. Similarly, most social phenomena involve different effects; thus, one of the main tasks of the social sciences should be to disentangle and decompose complex phenomena into simple effects. In principle this is easier to do in physics than in the social sciences because one can change experimental conditions fairly easily. However, history shows that the main obstacle are conceptual. The previous phenomenon involves the transformation of one form of energy (light) into other forms of energy and it is well know that it took centuries for a clear understanding of these processes to emerge. In order to convince the reader that the same approach can be used in the social sciences we briefly describe a specific case.

Suicide is commonly considered as a phenomenon which is due to many factors. One of them is the strength of the marital bond. How can we isolate that factor? Of course, it is impossible to isolate it completely but one can at least make it so predominant that other factors become negligible. To achieve that objective, we consider a population in which the number of males is much larger than

[2] In what economists call "empirical econometrics" the researcher necessarily must provide a multivariate econometric model which means that even before he analyses the data he already knows the theory which rules the phenomenon. Moreover, all factors whether they have a weak or a strong impact are treated on the same footing. As we will see in the next point this has important implications.

the number of females. Such a population will necessarily have a large proportion of bachelors and therefore will be an ideal testing ground to study the role of the marital bond. Where can we find populations with a large excess of men? Almost all populations of immigrants are characterized by an excess of males. It turns out that due to specific circumstances, this imbalance was particularly large in the population of Chinese people living in the United States. By the end of the 19th century there were about 27 Chinese men for one Chinese woman[3].

What makes the present principle important? Unless one is able to estimate the impact of each factor separately, one will never gain a *lasting* understanding. It is important to understand why. Let us for a moment return to the previous experiment. In the econometric approach one would conduct multivariate regressions of the temperature as a function of various (pre-conceived) parameters such as the volume of the liquid, the thickness of the glass and so on. Now suppose we wish to predict what happens when water is replaced by black ink. As a result of greater light absorption temperature differentials will be larger and convection currents will be stronger. The fact that many effects change at the same time will make the multivariate estimates irrelevant. Unless one has an understanding of the various individual effects it will be impossible to make any sound prediction. To sum up, any major change in business and social conditions will invalidate the previously accepted econometric models. This explains why the econometric approach fails to ensure that knowledge grows in a cumulative way.

What Guidance Can Physics Provide?

One can recall that the experimental methodology pioneered by researchers such as Tycho Brahe (1546–1601), Johannes Kepler (1571–1630) or Galileo (1564–1642) marked the beginning of modern physics. Two centuries later, that methodology was adapted to the exploration of the living world by people such as Claude Bernard (1813–1878), Louis Pasteur (1822–1895) and Gregor Mendel (1822–1884). In a sense it is a paradox that this method has been used successfully for the understanding of living organisms but has not yet gained broad acceptance in the social sciences for it can be argued with good reason that living organisms are more complex systems than are states or societies[4]. In short, applying the experimental methodology to the social sciences is a move which seems both natural and long overdue. Actually, serious efforts were made in this direction by social scientists such as Emile Durkheim (1858–1917) or Vilfredo Pareto (1848–1923) but this route seems to have been sidetracked in the second half of the 20th century.

Can we use the mathematical framework of physics in the investigation of social phenomena? This approach has been tried with some success by renowned econophysicists such as Belal Baaquie and coworkers (2004, 2007) and Jean-Philippe Bouchaud and coworkers [2,4]. In those cases the success must probably be attributed to the fact that the methods of theoretical physics which were used could be formulated in a purely mathematical way which did not rely on any physical concepts such as energy, momentum or temperature. As we do not yet know how these notions should be transposed to social systems, it seems impossible to apply the formalism of statistical mechanics to social phenomena[5].

Our claim that the experimental methodology of physics can be used to explore social phenomena must be substantiated by explaining how it is possible to carry out "experiments" in social phenomena. This is the purpose of the next section.

How Cross-National Observations Can Be Used to Test the Role of Different Factors

Nowadays when a solid state physicist wants to measure, say, the interaction between ultraviolet light and a crystal of germanium, the experiment involves little uncertainties. That is so because this field of physics is already well understood. On the contrary, in the case of new and not well understood phenomena there is considerable uncertainty about the specific conditions of the experimental set up. In the two years after Léon Foucault demonstrated the Foucault pendulum experiment, at least twenty physicists tried to repeat it. Some succeeded while others did not. Indeed the experimental conditions, e. g. the length of the pendulum or the nature of the suspension wire, ensuring that the Foucault effect will be observed were not well understood.

[3] For more details about this case, see [9].

[4] We will not develop this point here but it can be observed that a bacteria or a cell contains thousands of different proteins which interact in various ways. In the same line of thought one may recall that living organisms have been around for several billions years whereas societies appeared less than 100,000 years ago and states less than 10,000 years ago.

[5] It could be argued that one is free to define "social energy" in the way which one wishes. However, one should remember that the notion of energy is pivotal in physics only because it is ruled by (experimentally proved) conservation laws, such as the equivalence between heat and mechanical energy demonstrated by James Joule. Naturally, prior to defining a "social temperature", it would seem natural to define a herd- or swarm-temperature describing aggregated populations of bacteria, insects or animals. As far as we know, no operational definition of this kind has yet been proposed.

It is only through various attempts with different settings that a better understanding progressively emerged. For instance it was realized that by using a pendulum of great length one would be able to reduce two undesirable effects (i) the sensitivity of the pendulum to exogenous noise[6] (ii) the Puiseux effect which generates a rotation of the oscillation plane which interferes with the Foucault effect.

Few (if any) sociological phenomena are well understood which means that social researchers are basically in the same situation as those physicists in the years 1851–1852 who tried to observe the Foucault effect[7]. As an illustration suppose we wish to know if the publication of a specific type of news has an effect on the number of suicides[8]. Such an observation depends upon many parameters: the nature of the news and the amount of attention that it receives, the time interval (days, weeks or months?) between the publication of the news and the occurrence of the suicides. In addition one does not know if there will be an increase or a decrease in the number of suicides, if men will be more or less affected than women, and so on. All these questions can in principle be answered by conducting many observations in different countries and in different periods of time. In other words, if we are sufficiently determined, patient and tenacious and if we can get access to the statistical data that are needed, we should be able to disentangle and unravel the phenomenon under consideration in the same way as experimenters have been able to determine how the Foucault effect can be observed.

How Vested Interests May Affect the Accessibility and Reliability of Social Data

So far we have emphasized the similarities between natural and social phenomena but there are also some stumbling blocks which are specific to the social sciences. One of them is the fact that some data may have been altered or swept under the carpet by some sort of ideological, political or social bias, pressure or interference. Needless to say, extreme care must be exercised in such cases before making use of the data.

As an illustration, suppose that an econophysicist or a sociologist wants to study episodes of military occupation of one country by another. Such episodes are of particular interest from a sociological perspective because they bring about strong interactions and can serve to probe the characteristics of a society. Moreover, because armies display many similarities no matter their country of origin, such episodes offer a set of *controlled experiments*. Naturally, in order to be meaningful the comparison must rely on trustworthy accounts for each of the episodes. Unfortunately, it turns out that in many cases only scant and fairly unreliable information is available . Consider for instance the occupation of Iceland by British and American forces during World War II. Among all occupation episodes this one was particularly massive with troops representing 50% of the population of Iceland prior to the occupation. The same proportion in a country such as Japan would have meant 30 million occupation troops that is 60 times more than the peak number of 500,000 reached at the end of 1945. Quite understandably for such a high density of troops, there were many incidents with the population of Iceland[9]; yet, is is difficult to find detailed evidence. Due to the paucity of data a superficial investigation would easily lead to the conclusion that there were in fact only few incidents. It does not require much imagination to understand why this information has not been released. The fact that in a general way all countries whatsoever are reluctant to recognize possible misconduct of their military personnel explains why the information is still classified in British and American archives. Because Iceland and the United States became close allies after 1945, one can also understand that the Icelandic National Archive is reluctant to release information about these incidents. The same observation also applies (and for the same reasons) to the occupation of Japan, 1946–1951; for more details see Roehner pp. 90–98 in [9] and [10]. Naturally, similar cases abound. Due to a variety of reasons well-meaning governments, archivists and statistical offices keep sensitive files closed to social scientists. Most often it is in fact sufficient to catalog sensitive file units in a fairly obscure way. The plain effect is that the information will not be found except perhaps by pure luck, a fairly unlikely prospect in big archives.

How Can Exogenous Factors be Taken into Account?

This question is not specific to social phenomena, it is also of importance in physics. As a matter of fact, in astronomy

[6] Indeed, it is when the speed of the pendulum goes through zero that it is particularly sensitive to external perturbations; increasing the length of the pendulum reduces the number of oscillations in a given time interval and therefore the drift due to noise.

[7] As a more recent and even less understood case, one can mention the physicists who keep on trying to observe the cold fusion effect.

[8] This question is connected to what is known in sociology as the Werther effect; for more details see the papers written by Phillips (in particular [7]) and Chap. 3 in [9].

[9] According to a report that Prime Minister Hermann Jonasson sent to the American Headquarters, there were 136 incidents between troops and Icelanders during the period between July 1941 (arrival of the American troops) and April 1942 (Hunt 1966) in Reykjavik alone. Unfortunately, no copy of this report seems to be available at the National Archives of Iceland.

it provides a powerful method for observing objects that cannot be observed directly. Thus, we know the existence of exoplanets only from the perturbing effect which they have on the position of the star around which they move. However, for social phenomena the problem of exogenous factors is much more serious because (i) they may not be known to observers (ii) even once they are identified it is very difficult to correct the data in a reliable way. One of the main pitfalls in the modeling of socio-economic phenomena is to explain them through endogenous mechanisms while they are in fact due to exogenous factors. The following examples make clear that this difficulty exists for many phenomena, whether they belong to the financial, economic or social sphere.

- In their paper of 2005 about consensus formation and shifts in opinion Michard and Bouchaud confront their theory to two classes of social phenomena: (i) the diffusion of cell phones (ii) the diffusion of birth rate patterns. In the first case it is clear that advertising campaigns may have played an important role. Of course, one could argue that these campaigns were part of the endogenous diffusion process. However, this argument does not hold for big telecom companies (e. g. Vodafone) which operate in many countries. In such cases the decision about the magnitude of the advertising campaigns are taken by the board of the company which means that such campaigns can hardly be considered as endogenous effects. Similarly, birth rates depend upon exogenous factors. For instance the length of time spent in higher education has an effect on the average age of marriage and the later has an effect on birth rates.
- On 21 July 2004 the share price of Converium, a Swiss reinsurance company listed on the New York Stock Exchange dropped 50%. Was this fall the result of an avalanche effect due to a movement of panic among investors? In fact, the most likely explanation is that it was the consequence of a decision taken by the board of Fidelity International, a major investment fund and one of the main shareholders of Converium. Indeed in a statement issued by Converium on August 3, 2004 it was announced that Fidelity had reduced its holdings from 9.87% to 3.81%. In other words, it would be completely irrelevant to explain such a fall through a herd effect model or through any other endogenous mechanism (more details can be found in [8]). Similar conclusions apply to corporate stock buybacks, as well as to mergers, acquisitions, buyouts and takeovers; in all these cases decisions taken by a few persons (the average board of directors has nine members) may trigger substantial changes in share prices. How should such effects be taken into account by stock market models?
- At the end of 2004 and in the first months of 2005 British housing prices began to decline after having risen rapidly during several years. Yet after May 2005, they suddenly began to pick up again at an annual rate of about 10%. This resurgence was particularly intriguing because at the same time US housing prices began to decline. To what factor should this unexpected rise be attributed? Most certainly this was the market response to a plan introduced by the Chancellor of the Exchequer Gordon Brown in late May (The Economist May 28, 2005). Under this plan which aimed at propping up house prices new buyers would benefit from a zero-interest loan for 12% of the price. In addition, the government would cover all losses incurred by banks as a result of possible bankruptcies of borrowers (at least so long as prices did not fall by more than 12%). It appears that the plan indeed propped up the market. Consequently, in order to confront the predictions of any model (e. g. see Richmond's paper which was published in 2007) with observation the impact of this plan effect must first be taken out of the picture.
- The same difficulty is also encountered in socio-political phenomena. Here is an illustration. On 5 October 2000, in protest against the publication of the results of the presidential election there was a huge mass demonstration in Belgrade which involved thousands of people from the provinces who were transported to the federal capital by hundreds of buses. It clearly showed that president Milosevic was no longer in control of the police and army and lead to his retirement from the political scene. Thus, what NATO air strikes (24 March-11 June 1999[10]) had not been able to achieve was accomplished by one night of street demonstrations. What was the part of exogenous factors in this event? Although in many similar cases it is very difficult to know what really happened, in this specific case a partial understanding is provided by a long article published in the New York Times[11]. In this article

[10] It can be noted that similarly to what would happen in 2003 for the invasion of Iraq, these air strikes were carried out without the authorization of the United Nations Security Council.

[11] New York Times, Sunday 26 November 2000, Magazine Section, p. 43, 7705 words; the article by Roger Cohen is entitled: "Who really brought down Milosevic". What makes this account particularly convincing is the fact that it was preceded by another article entitled: "US anti-Milosevic plan faces major test at polls" which appeared on September 23, 2000 (p. 6, 1150 words); this article described the way Milosevic would be removed from power two weeks *before* the events. The article makes clear that the course of events would be the same no matter what the results of the election would be.

we learn that several American organizations belonging to the intelligence network supported, financed and trained Serbian opposition groups. For instance the article mentions the Albert Einstein Foundation, the International Republican Institute, the National Endowment for Democracy, the US Agency for International Development. Although the amount of the total financial support is not known, the New York Times article says that it exceeded $ 28 million. The plan comprised two facets: the organization of demonstrations on the one hand and the infiltration of the army and police on the other hand in order to undermine their loyalty and convince them to remain passive during the demonstrations. According to the article this second facet remains classified. With an exogenous interference of such a magnitude, it would clearly be meaningless to describe this upheaval as a purely endogenous process. Moreover, the fact that we have only partial knowledge about the exogenous forces makes it very difficult (if not altogether impossible) to come up with a satisfactory description. It should also be noted that the influence of these groups did not disappear overnight after October 4, which means that the subsequent history of Serbia must also take them into account at least to some extent.

Future Directions

In this article we have described the challenges and obstacles to which one is confronted in trying to understand socio-economic phenomena. In parallel we have shown that the econophysics approach has many assets. One of them which has not yet been mentioned is the fact that econophysicists are not subject to the rigid barriers which exist between various fields and subfields of the human sciences. Thus, if it turns out that in order to explain an economic phenomena one needs to understand a social effect, econophysicists would have no problem in shifting from one field to another. There is another historical chance that we have not mentioned so far, namely the development of the Internet. In the past decade 1997–2007 the amount of information to which one has access has increased tremendously. Electronic catalogs of major libraries or of national archives, indexes of newspaper, search engines on the Internet, searchable databases of books, all these innovations contributed to give the researcher easy access to information sources that have never been available before. In particular it has become fairly easy to find cross-national data. Thus, social scientists and econophysicists are in a better position than ever for carrying out the kind of comparative studies that we called for in this article.

Bibliography

Primary Literature

1. Aoki M, Yoshikawa H (2007) Reconstructing macroeconomics. Cambridge University Press, Cambridge
2. Bouchaud J-P, Potters M (2003) Theory of financial risk and derivative pricing. Cambridge University Press, Cambridge
3. Leontief W (1983) Foreword. In: Eichner AS (ed) Why economics is not yet a science. M.E. Sharpe, Armonk(New York)
4. Michard Q, Bouchaud J-P (2005) Theory of collective opinion shifts: from smooth trends to abrupt swings. Eur. Phys. J. B 47:151–159
5. Neumann J von, Morgenstern O (1953) Theory of games and economic behavior. Princeton University Press, Princeton
6. Nirei M, Souma W (2007) Two factor model of income distribution dynamics. Review of Income and Wealth 53(3):440–459
7. Phillips DP (1974) The influence of suggestion on suicide: substantive and theoretical implications of the Werther effect. Am. Sociol. Rev. 39:340–354
8. Roehner BM (2006) Macroplayers in stock markets. In: Takayasu H (ed) Proceedings of the 3rd Nikkei Economics Symposium, Tokyo. Springer, Tokyo, pp 262–271
9. Roehner BM (2007) Driving forces in physical, biological and socio-economic phenomena. Cambridge University Press, Cambridge
10. Roehner BM (2008) Relations between Allied forces and the population of Japan, Working Report UPMC, Paris
11. Schumpeter J (1933) The common sense of econometrics. Econometrica 1:5–12
12. Schwartz AJ (1995) An interview with Anna J. Schwartz. Newsl. Cliometric Soc. 10(2):3–7

Books and Reviews

Two observations are in order about this reference section:

Many of these references are not mentioned in the text; the objective is to give readers a starting point for further readings on various aspects of econophysics.

There is a fairly complete list of publications of the present author; it is given for the purpose of illustrating through one specific case the "trajectory" of an econophysicist in the course of time (1995–2007).

Amaral LAN, Buldyrev SV, Havlin S, Leschhorn H, Maass P, Salinger A, Stanley HE, Stanley MHR (1997) Scaling behavior in economics: I. Empirical results for company growth. J Phys. I Fr. 7:621–633

Amaral LAN, Buldyrev SV, Havlin S, Salinger MA, Stanley HE (1998) Power law scaling for a system of interacting units with complex internal structure. Phys. Rev. Lett. 80(7):1385–1388

Aoki M, Yoshikawa H (2007) Reconstructing macroeconomics. Cambridge University Press, Cambridge

Baaquie BE (2004) Quantum finance. Cambridge University Press, Cambridge

Baaquie BE (2007) Feynman perturbation expansion for the price of coupon bond options and swaptions in quantum finance. I. Theory Phys. Rev. E 75, 016703

Baaquie BE, Liang C (2007) Feynman perturbation expansion for the price of coupon bond options and swaptions in quantum finance. II. Empirical Phys. Rev. E 75, 016704

Baaquie BE, Srikant M (2004) Comparison of field theory models of interest rates with market data. Phys. Rev. E 69, 036129
Borghesi C, Bouchaud J-P (2007) On songs and men. Quality and Quantity 41(4):557–568
Bouchaud J-P, Marsili M, Roehner BM, Slanina F (eds) (2001) Application of physics in economic modelling. Proceedings of the NATO Advanced Research Workshop held in Prague, Czech Republic, 8–10 February 2001. Physica A 299(1–2):1–355
Bouchaud J-P, Potters M (1997) Théorie des risques financiers. Aléa, Saclay
Bouchaud J-P, Potters M (2003) Theory of financial risk and derivative pricing. Cambridge University Press, Cambridge
Buldyrev SV, Amaral LAN, Havlin S, Leschhorn H, Maass P, Salinger MA, Stanley HE, Stanley MHR (1997) Scaling behavior in economics: II. Modeling of company growth. J Phys. I Fr. 7:635–650
Chakraborti A, Chakrabarti BK (2000) Statistical mechanics of money: how saving propensity affects its distribution. Eur. Phys. J. B 17:167–170
de Oliveira SM,de Oliveira PMC, Stauffer D (1999) Evolution, money, war and computer. Teubner, Leipzig
Deschâtres F, Sornette D (2005) The dynamics of book sales: endogenous versus exogenous shocks in complex networks. Phys. Rev. E 72, 016112
Dragulescu A, Yakovenko VM (2000) Statistical mechanics of money. Eur. Phys. J. B 17(4):723–729
Farmer JD (1999) Physicists attempt to scale the ivory towers of finance. Comput. Sci. Eng. Nov-Dec 1999, 26–39
Farmer JD, Lillo F (2004) On the origin of power law tails in price fuctuations. Quant. Finance 4(1):7–11
Feigenbaum JA, Freund PGO (1998) Discrete scale invariance and the second Black Monday. Mod. Phys. Lett. B 12(2–3):57–60
Fu Y-Q, Zhang H, Cao Z, Zheng B, Hu G (2005) Removal of pinned spiral by generating target waves with a localized stimulus. Phys. Rev. E 72, 046206
Galam S (2006) Opinion dynamics, minority spreading and heterogenous beliefs. In: Chakrabarti BK, Chakraborti A, Chatterjee A (eds) Econophysics and Sociophysics. Wiley-VCH, Weinheim
Ghashghaie S, Breymann W, Peinke J, Talkner P, Dodge Y (1996) Turbulent cascades in foreign exchange markets. Nature 381:767–770
Guillaume DM, Dacorogna MM, Davé R, Müller UA, Olsen RB, Pictet OV (1997) From the bird's eye to the microscope: a survey of new stylized facts of the intra-daily foreign exchange markets. Finance Stoch. 1:95–129
Hunt JJ (1966) The United States occupation of Iceland, 1941–1946. Thesis. Georgetown University, Washington DC
Johansen A, Sornette D (1999) Financial anti-bubbles: Log-periodicity in gold and Nikkei collapses. Int. J. Mod Phys C 10(4):563–575
Johansen A, Sornette D (2001) Bubbles and anti-bubbles in Latin-American, Asian and Western stock markets: An empirical study. Int. J. Theor. Appl. Finance 4(6):853–920
Juglar C (1862): Des crises commerciales et de leur retour périodique en France, en Angleterre et aux Etats-Unis. English translation (1893, 1966) A brief history of panics and their periodical occurrence in the United States. A.M. Kelley, New York
Lai KK, Leung FKN, Tao B, Wang S (2000) Practices of preventive maintenance and replacement for engines: a case study. Eur J Oper Res 124:2
Leontief W (1983) Foreword. In: Eichner AS (ed) Why economics is not yet a science. M.E. Sharpe, Armonk New York
Li M, Wu J, Wang D, Zhou T, Di Z, Fan Y (2006) Evolving model of weighted networks inspired by scientific collaboration networks. Physica A 375(1):355–364
Lillo F, Mike S, Farmer JD (2005) Theory for Long Memory in supply and semand. Physical Review E 7106 (6 pt 2) 287–297
Lux T (1996) The stable Paretian hypothesis and the frequency of large returns: an examination of major German stocks. Appl Financial Econ 6:463–475
Mandelbrot B (1997) Les fractales et la Bourse. Pour Sci 242:16–17
Mantegna RN (1999) Hierarchical structure in financial markets. Eur Phys J B 11:193–197
Mantegna RN, Stanley HE (1995) Scaling behavior in the dynamics of an economic index. Nature 376:46–49
Mantegna RN, Stanley HE (1999) Introduction to econophysics. Cambridge University Press, Cambridge
McCauley JL (2004) Dynamics of markets. Cambridge University Press, Cambridge
Michard Q, Bouchaud J-P (2005) Theory of collective opinion shifts: from smooth trends to abrupt swings. Eur Phys J B 47:151–159
Mimkes J (2006) A thermodynamic formulation of social science. In: Chakrabarti BK, Chaterjee A (eds) Econophysics and sociophysics: trends and perspectives. Wiley-VCH, Weinheim, pp 279–310
Müller UA, Dacorogna MM, Davé R, Olsen RB, Pictet OV, Weizsäcker J von (1997) Volatilities of different time resolutions. Analysing the dynamics of market components. J Empir Finance 4(2–3):213–240
Müller UA, Dacorogna MM, Olsen RB, Pictet OV, Schwarz M (1990) Statistical study of foreign exchange rates, empirical evidence of a price scaling law, and intraday analysis. J Bank Finance 14:1189–1208
Neumann J von, Morgenstern O (1953) Theory of games and economic behavior. Princeton University Press, Princeton
Phillips DP (1974) The influence of suggestion on suicide: substantive and theoretical implications of the Werther effect. Am Sociol Rev 39:340–354
Plerou V, Amaral LAN, Gopikrishnan P (1999) Similarities between the growth dynamics of university research and of competitive economic activities. Nature 400(6743):433–437
Plerou V, Gopikrishnan P, Rosenow B, Amaral LA, Stanley H (1999) Universal and nonuniveral properties of cross-correlation in financial time series. Phys Rev Lett 83(7):1471–1474
Richmond P (2007) A roof over your head; house price peaks in the UK and Ireland. Physica A 375(1,15):281–287
Roehner BM (1995) Theory of markets. Trade and space-time patterns of price fluctuations: a study in analytical economics. Springer, Berlin
Roehner BM (1997) Jesuits and the state. A comparative study of their expulsions (1500–1990). Religion 27:165–182
Roehner BM (1997) The comparative way in economics: a reappraisal. Econom Appl 50(4):7–32
Roehner BM (1999) Spatial analysis of real estate price bubbles: Paris 1984–1993. Reg Sci Urban Econ 29:73–88
Roehner BM (1999) The space-time pattern of price waves. Eur Phys J B 8:151–159
Roehner BM (2000) Determining bottom price-levels after a speculative peak. Eur Phys J B 17:341–345
Roehner BM (2000) Identifying the bottom line after a stock market crash. Int J Mod Phys C 11(1):91–100

Roehner BM (2000) Speculative trading: the price multiplier effect. Eur Phys J B 14:395–399
Roehner BM (2000) The correlation length of commodity markets: 1. Empirical evidence. Eur Phys J B 13:175–187
Roehner BM (2000) The correlation length of commodity markets: 2. Theoretical framework. Eur Phys J B 13:189–200
Roehner BM (2001) Hidden collective factors in speculative trading: a study in analytical economics. Springer, Berlin
Roehner BM (2001) To sell or not to sell? Behavior of shareholders during price collapses. Int J Mod Phys C 12(1):43–53
Roehner BM (2001) Two classes of speculative peaks. Physica A 299:71–83
Roehner BM (2002) Patterns of speculation: a study in observational econophysics. Cambridge University Press, Cambridge
Roehner BM (2002) Patterns and repertoire. Harvard University Press, Cambridge Massachussets
Roehner BM (2002) Separatism and integration. Rowman and Littlefield, Lanham Maryland
Roehner BM (2004) Patterns of speculation in real estate and stocks. In: Takayasu H (ed) Proceedings of the 2nd Nikkei Economics Symposium, Tokyo. Springer, Tokyo, pp 103–116
Roehner BM (2005) A bridge between liquids and socio-economic systems: the key-role of interaction strengths. Physica A 348:659–682
Roehner BM (2005) Cohésion sociale. Odile Jacob, Paris
Roehner BM (2005) Stock markets are not what we think they are: the key roles of cross-ownership and corporate treasury stock. Physica A 347:613–626
Roehner BM (2006) Macroplayers in stock markets. In: Takayasu H (ed) Proceedings of the 3rd Nikkei Economics Symposium, Tokyo. Springer, Tokyo, pp 262–271
Roehner BM (2006) Real estate price peaks: a comparative perspective. Evol Institutional Econ Rev 2(2):167–182
Roehner BM (2007) Driving forces in physical, biological and socio-economic phenomena. Cambridge University Press, Cambridge
Roehner BM (2008) Econophysics: Challenges and promises. Evolutionary abd Institutional Economics Review 4(2):251–266
Roehner BM, Jego C (2006) White flight or flight from poverty? J Econ Intearct Coord 1:75–87
Roehner BM, Maslov S (2003) Does the price multiplier effect also hold for stocks? Int J Mod Phys C 14(10):1439–1451
Roehner BM, Maslov S (2003) The conundrum of stock versus bond prices. Physica A 335:164–182 (2004)
Roehner BM, Rahilly LJ (2002) Separatism and integration: a study in analytical history. Rowman and Littlefield, Lanham Maryland
Roehner BM, Shiue C (2001) Comparing the correlation length of grain markets in China and France. Int J Mod Phys C 11(7):1383–1410
Roehner BM, Sornette D (1998) The sharp peak – flat trough pattern and critical speculation. Eur Phys J B 4:387–399
Roehner BM, Sornette D (1999) Analysis of the phenomenon of speculative trading in one of its basic manifestations: postage stamp bubbles. Int J Mod Phys C 10(6):1099–1116
Roehner BM, Sornette D (2000) Thermometers of speculative frenzy. Eur Phys J B 16:729–739
Roehner BM, Sornette D, Andersen J (2004) Response functions to critical shocks in social sciences: an empirical and numerical study. Int J Mod Phys C 15(6):809–834
Roehner BM, Syme T (2002) Pattern and repertoire in history: an introduction to analytical history. Harvard University Press, Cambridge Massachusetts
Schumpeter J (1933) The common sense of econometrics. Econometrica 1:5–12
Schwartz AJ (1995) An interview with Anna J Schwartz. Newsl Cliometric Soc 10(2):3–7
Sornette D (2003) Why stock markets crash. Critical events in complex financial systems. Princeton University Press, Princeton
Stauffer D, Sornette D (1999) Self-organized percolation model for stock market fluctuations. Physica A 271(3–4):496–506
Summers LH (1991) The scientific illusion in empirical macroeconomics. Scand J Econ 93(2):129–148
Takayasu H (ed) (2004) The application of econophysics. In: Proceedings of the 2nd Nikkei Econophysics Symposium. Springer, Tokyo
Takayasu H (ed) (2006) Practical fruits of econophysics. In: Proceedings of the 3rd Nikkei Econophysics Symposium. Springer, Tokyo
Turchin P (2003) Historical dynamics. Why states rise and fall. Princeton University Press, Princeton
Wyatt M, Bouchaud J-P (2003) Self referential behaviour, overreaction and conventions in financial markets. Cond-mat/03033584
Zhou W-X, Sornette D (2003) 2000–2003 real estate bubble in the UK and not in the USA. Physica A 329(1–2):249–263
Zhou W-X, Sornette D (2003) Evidence of a worldwide stock market log-periodic anti-bubble since mid-2000. Physica A 330:543–583
Zhou W-X, Sornette D (2004) Antibubble and Prediction of China's stock market and Real-Estate. Physica A 337(1–2):243–268

Econophysics, Statistical Mechanics Approach to

Victor M. Yakovenko
Department of Physics, University of Maryland, College Park, USA

"Money, it's a gas." Pink Floyd

Article Outline

Glossary
Definition of the Subject
Historical Introduction
Statistical Mechanics of Money Distribution
Statistical Mechanics of Wealth Distribution
Data and Models for Income Distribution
Other Applications of Statistical Physics
Future Directions, Criticism, and Conclusions
Bibliography

Glossary

Probability density $P(x)$ is defined so that the probability of finding a random variable x in the interval from x to $x + \mathrm{d}x$ is equal to $P(x)\,\mathrm{d}x$.

Cumulative probability $C(x)$ is defined as the integral $C(x) = \int_x^\infty P(x)\mathrm{d}x$. It gives the probability that the random variable exceeds a given value x.

The Boltzmann–Gibbs distribution gives the probability of finding a physical system in a state with the energy ε. Its probability density is given by the exponential function (1).

The Gamma distribution has the probability density given by a product of an exponential function and a power-law function, as in (9).

The Pareto distribution has the probability density $P(x) \propto 1/x^{1+\alpha}$ and the cumulative probability $C(x) \propto 1/x^{\alpha}$ given by a power law. These expressions apply only for high enough values of x and do not apply for $x \to 0$.

The Lorenz curve was introduced by American economist Max Lorenz to describe income and wealth inequality. It is defined in terms of two coordinates $x(r)$ and $y(r)$ given by (19). The horizontal coordinate $x(r)$ is the fraction of the population with income below r, and the vertical coordinate $y(r)$ is the fraction of income this population accounts for. As r changes from 0 to ∞, x and y change from 0 to 1, parametrically defining a curve in the (x, y)-plane.

The Gini coefficient G was introduced by the Italian statistician Corrado Gini as a measure of inequality in a society. It is defined as the area between the Lorenz curve and the straight diagonal line, divided by the area of the triangle beneath the diagonal line. For perfect equality (everybody has the same income or wealth) $G = 0$, and for total inequality (one person has all income or wealth, and the rest have nothing) $G = 1$.

The Fokker–Planck equation is the partial differential equation (22) that describes evolution in time t of the probability density $P(r, t)$ of a random variable r experiencing small random changes Δr during short time intervals Δt. It is also known in mathematical literature as the Kolmogorov forward equation. The diffusion equation is an example of the Fokker–Planck equation.

Definition of the Subject

Econophysics is an interdisciplinary research field applying methods of statistical physics to problems in economics and finance. The term "econophysics" was first introduced by the prominent theoretical physicist Eugene Stanley in 1995 at the conference *Dynamics of Complex Systems*, which was held in Calcutta (now known as Kolkata) as a satellite meeting to the STATPHYS-19 conference in China [1,2]. The term appeared in print for the first time in the paper by Stanley et al. [3] in the proceedings of the Calcutta conference. The paper presented a manifesto of the new field, arguing that "behavior of large numbers of humans (as measured, e. g., by economic indices) might conform to analogs of the scaling laws that have proved useful in describing systems composed of large numbers of inanimate objects" [3]. Soon the first econophysics conferences were organized: *International Workshop on Econophysics*, Budapest, 1997 and *International Workshop on Econophysics and Statistical Finance*, Palermo, 1998 [2], and the book *An Introduction to Econophysics* [4] was published.

The term "econophysics" was introduced by analogy with similar terms, such as "astrophysics", "geophysics", and "biophysics", which describe applications of physics to different fields. Particularly important is the parallel with biophysics, which studies living creatures, which still obey the laws of physics. It should be emphasized that econophysics does not literally apply the laws of physics, such as Newton's laws or quantum mechanics, to humans, but rather uses mathematical methods developed in statistical physics to study statistical properties of complex economic systems consisting of a large number of humans. So, it may be considered as a branch of applied theory of probabilities. However, statistical physics is distinctly different from mathematical statistics in its focus, methods, and results.

Originating from physics as a quantitative science, econophysics emphasizes quantitative analysis of large amounts of economic and financial data, which became increasingly available with the massive introduction of computers and the Internet. Econophysics distances itself from the verbose, narrative, and ideological style of political economy and is closer to **econometrics** in its focus. Studying mathematical models of a large number of interacting economic agents, econophysics has much common ground with the **agent-based modeling and simulation**. Correspondingly, it distances itself from the representative-agent approach of traditional economics, which, by definition, ignores statistical and heterogeneous aspects of the economy.

Two major directions in econophysics are applications to finance and economics. Observational aspects are covered in the article ► Econophysics, Observational. The present article, ► Econophysics, Statistical Mechanics Approach to, concentrates primarily on statistical distributions of money, wealth, and income among interacting economic agents.

Another direction related to econophysics has been advocated by the theoretical physicist Serge Galam since the early 1980s under the name "**sociophysics**" [5], with the first appearance of the term in print in [6]. It echoes the term *physique sociale* proposed in the nineteenth century by Auguste Comte, the founder of sociology. Unlike econophysics, the term "sociophysics" did not catch on when first introduced, but it is coming back with the popularity of econophysics and active promotion by some physicists [7,8,9]. While the principles of both fields have a lot in common, econophysics focuses on the narrower subject of economic behavior of humans, where more quantitative data are available, whereas sociophysics studies a broader range of social issues. The boundary between econophysics and sociophysics is not sharp, and the two fields enjoy a good rapport [10]. A more detailed description of the historical development in presented in Sect. "Historical Introduction".

Historical Introduction

Statistical mechanics was developed in the second half of the nineteenth century by James Clerk Maxwell, Ludwig Boltzmann, and Josiah Willard Gibbs. These physicists believed in the existence of atoms and developed mathematical methods for describing their statistical properties, such as the probability distribution of velocities of molecules in a gas (the Maxwell–Boltzmann distribution) and the general probability distribution of states with different energies (the Boltzmann–Gibbs distribution). There are interesting connections between the development of statistical physics and statistics of social phenomena, which were recently brought up by the science journalist Philip Ball [11,12].

Collection and study of "social numbers", such as the rates of death, birth, and marriage, has been growing progressively since the seventeenth century (see Chap. 3 in [12]). The term "statistics" was introduced in the eighteenth century to denote these studies dealing with the civil "states", and its practitioners were called "statists". Popularization of social statistics in the nineteenth century is particularly accredited to the Belgian astronomer Adolphe Quetelet. Before the 1850s, statistics was considered an empirical arm of political economy, but then it started to transform into a general method of quantitative analysis suitable for all disciplines. It stimulated physicists to develop statistical mechanics in the second half of the nineteenth century.

Rudolf Clausius started development of the kinetic theory of gases, but it was James Clerk Maxwell who made a decisive step of deriving the probability distribution of velocities of molecules in a gas. Historical studies show (see Chap. 3 in [12]) that, in developing statistical mechanics, Maxwell was strongly influenced and encouraged by the widespread popularity of social statistics at the time. This approach was further developed by Ludwig Boltzmann, who was very explicit about its origins (see p. 69 in [12]):

> "The molecules are like individuals, … and the properties of gases only remain unaltered, because the number of these molecules, which on the average have a given state, is constant."

In his book *Populäre Schriften* from 1905 [13], Boltzmann praises Josiah Willard Gibbs for systematic development of statistical mechanics. Then, Boltzmann says (cited from [14]):

> "This opens a broad perspective if we do not only think of mechanical objects. Let's consider to apply this method to the statistics of living beings, society, sociology and so forth."

(The author is grateful to Michael E. Fisher for bringing this quote to his attention.)

It is worth noting that many now-famous economists were originally educated in physics and engineering. Vilfredo Pareto earned a degree in mathematical sciences and a doctorate in engineering. Working as a civil engineer, he collected statistics demonstrating that distributions of income and wealth in a society follow a power law [15]. He later became a professor of economics at Lausanne, where he replaced Léon Walras, also an engineer by education. The influential American economist Irving Fisher was a student of Gibbs. However, most of the mathematical apparatus transferred to economics from physics was that of Newtonian mechanics and classical thermodynamics [16]. It culminated in the neoclassical concept of mechanistic equilibrium where the "forces" of supply and demand balance each other. The more general concept of statistical equilibrium largely eluded mainstream economics.

With time, both physics and economics became more formal and rigid in their specializations, and the social origin of statistical physics was forgotten. The situation is well summarized by Philip Ball (see p. 69 in [12]):

> "Today physicists regard the application of statistical mechanics to social phenomena as a new and risky venture. Few, it seems, recall how the process originated the other way around, in the days when physical science and social science were the twin siblings of a mechanistic philosophy and when it was

not in the least disreputable to invoke the habits of people to explain the habits of inanimate particles".

Some physicists and economists attempted to connect the two disciplines during the twentieth century. The theoretical physicist Ettore Majorana argued in favor of applying the laws of statistical physics to social phenomena in a paper published after his mysterious disappearance [17]. The statistical physicist Elliott Montroll coauthored the book *Introduction to Quantitative Aspects of Social Phenomena* [18]. Several economists applied statistical physics to economic problems [19,20,21,22]. An early attempt to bring together the leading theoretical physicists and economists at the Santa Fe Institute was not entirely successful [23]. However, by the late 1990s, the attempts to apply statistical physics to social phenomena finally coalesced into the robust movements of econophysics and sociophysics, as described in Sect. "Definition of the Subject".

The current standing of econophysics within the physics and economics communities is mixed. Although an entry on econophysics has appeared in the *New Palgrave Dictionary of Economics* [24], it is fair to say that econophysics is not accepted yet by mainstream economics. Nevertheless, a number of open-minded, nontraditional economists have joined this movement, and the number is growing. Under these circumstances, econophysicists have most of their papers published in physics journals. The journal *Physica A: Statistical Mechanics and Its Applications* emerged as the leader in econophysics publications and has even attracted submissions from some bona fide economists. The mainstream physics community is generally sympathetic to econophysics, although it is not uncommon for econophysics papers to be rejected by *Physical Review Letters* on the grounds that "it is not physics". There are regular conferences on econophysics, such as *Applications of Physics in Financial Analysis* (sponsored by the European Physical Society), *Nikkei Econophysics Symposium*, and *Econophysics Colloquium*. Econophysics sessions are included in the annual meetings of physical societies and statistical physics conferences. The overlap with economics is the strongest in the field of agent-based simulation. Not surprisingly, the conference series WEHIA/ESHIA, which deals with heterogeneous interacting agents, regularly includes sessions on econophysics.

Statistical Mechanics of Money Distribution

When modern econophysics started in the middle of the 1990s, its attention was primarily focused on analysis of financial markets. However, three influential papers [25,26,27], dealing with the subject of money and wealth distributions, were published in 2000. They started a new direction that is closer to economics than finance and created an expanding wave of follow-up publications. We start reviewing this subject with [25], whose results are the most closely related to the traditional statistical mechanics in physics.

The Boltzmann–Gibbs Distribution of Energy

The fundamental law of equilibrium statistical mechanics is the Boltzmann–Gibbs distribution. It states that the probability $P(\varepsilon)$ of finding a physical system or subsystem in a state with the energy ε is given by the exponential function

$$P(\varepsilon) = c\mathrm{e}^{\frac{-\varepsilon}{T}} , \tag{1}$$

where T is the temperature, and c is a normalizing constant [28]. Here we set the Boltzmann constant k_B to unity by choosing the energy units for measuring the physical temperature T. Then, the expectation value of any physical variable x can be obtained as

$$\langle x \rangle = \frac{\sum_k x_k \mathrm{e}^{\frac{-\varepsilon_k}{T}}}{\sum_k \mathrm{e}^{\frac{-\varepsilon_k}{T}}} , \tag{2}$$

where the sum is taken over all states of the system. Temperature is equal to the average energy per particle: $T \sim \langle \varepsilon \rangle$, up to a numerical coefficient of the order of 1.

Equation (1) can be derived in different ways [28]. All derivations involve the two main ingredients: statistical character of the system and conservation of energy ε. One of the shortest derivations can be summarized as follows. Let us divide the system into two (generally unequal) parts. Then, the total energy is the sum of the parts, $\varepsilon = \varepsilon_1 + \varepsilon_2$, whereas the probability is the product of probabilities, $P(\varepsilon) = P(\varepsilon_1)P(\varepsilon_2)$. The only solution of these two equations is the exponential function (1).

A more sophisticated derivation, proposed by Boltzmann himself, uses the concept of entropy. Let us consider N particles with total energy E. Let us divide the energy axis into small intervals (bins) of width $\Delta\varepsilon$ and count the number of particles N_k having energies from ε_k to $\varepsilon_k + \Delta\varepsilon$. The ratio $N_k/N = P_k$ gives the probability for a particle having the energy ε_k. Let us now calculate the multiplicity W, which is the number of permutations of the particles between different energy bins such that the occupation numbers of the bins do not change. This quantity is given by the combinatorial formula in terms of the factorials

$$W = \frac{N!}{N_1! N_2! N_3! \ldots} . \tag{3}$$

The logarithm of multiplicity of called the entropy $S = \ln W$. In the limit of large numbers, the entropy per particle can be written in the following form using the Stirling approximation for the factorials:

$$\frac{S}{N} = -\sum_k \frac{N_k}{N} \ln\left(\frac{N_k}{N}\right) = -\sum_k P_k \ln P_k \,. \tag{4}$$

Now we would like to find what distribution of particles between different energy states has the highest entropy, i. e. the highest multiplicity, provided that the total energy of the system, $E = \sum_k N_k \varepsilon_k$, has a fixed value. Solution of this problem can be easily obtained using the method of Lagrange multipliers [28], and the answer gives the exponential distribution (1).

The same result can be derived from the **ergodic theory**, which says that the many-body system occupies all possible states of a given total energy with equal probabilities. Then it is straightforward to show [29,30] that the probability distribution of the energy of an individual particle is given by (1).

Conservation of Money

The derivations outlined in Sect. "The Boltzmann–Gibbs Distribution of Energy" are very general and use only the statistical character of the system and the conservation of energy. So, one may expect that the exponential Boltzmann–Gibbs distribution (1) may apply to other statistical systems with a conserved quantity.

The economy is a big statistical system with millions of participating agents, so it is a promising target for applications of statistical mechanics. Is there a conserved quantity in economy? Drăgulescu and Yakovenko [25] argue that such a conserved quantity is money m. Indeed, the ordinary economic agents can only receive money from and give money to other agents. They are not permitted to "manufacture" money, e. g., to print dollar bills. When one agent i pays money Δm to another agent j for some goods or services, the money balances of the agents change as follows:

$$\begin{aligned} m_i &\to m_i' = m_i - \Delta m \,, \\ m_j &\to m_j' = m_j + \Delta m \,. \end{aligned} \tag{5}$$

The total amount of money of the two agents before and after the transaction remains the same,

$$m_i + m_j = m_i' + m_j' \,, \tag{6}$$

i. e., there is a local conservation law for money. The rule (5) for the transfer of money is analogous to the transfer of energy from one molecule to another in molecular collisions in a gas, and (6) is analogous to conservation of energy in such collisions.

Addressing some misunderstandings developed in economic literature [31,32,33,34], we should emphasize that, in the model of [25], the transfer of money from one agent to another happens voluntarily, as a payment for goods and services in a market economy. However, the model only keeps track of money flow, and does not keep track of what kinds of goods and service are delivered. One reason for this is that many goods, e. g., food and other supplies, and most services, e. g., getting a haircut or going to a movie, are not tangible and disappear after consumption. Because they are not conserved and also because they are measured in different physical units, it is not very practical to keep track of them. In contrast, money is measured in the same unit (within a given country with a single currency) and is conserved in transactions, so it is straightforward to keep track of money flow.

Unlike ordinary economic agents, a central bank or a central government can inject money into the economy. This process is analogous to an influx of energy into a system from external sources, e. g., the Earth receives energy from the Sun. Dealing with these situations, physicists start with an idealization of a closed system in thermal equilibrium and then generalize to an open system subject to an energy flux. As long as the rate of money influx from central sources is slow compared with relaxation processes in the economy and does not cause hyperinflation, the system is in quasi-stationary statistical equilibrium with slowly changing parameters. This situation is analogous to heating a kettle on a gas stove slowly, where the kettle has a well-defined, but slowly increasing temperature at any moment of time.

Another potential problem with conservation of money is debt. This issue is discussed in more detail in Sect. "Models with Debt". As a starting point, Drăgulescu and Yakovenko [25] first considered simple models, where debt is not permitted. This means that money balances of agents cannot go below zero: $m_i \geq 0$ for all i. Transaction (5) takes place only when an agent has enough money to pay the price, $m_i \geq \Delta m$, otherwise the transaction does not take place. If an agent spends all the money, the balance drops to zero $m_i = 0$, so the agent cannot buy any goods from other agents. However, this agent can still produce goods or services, sell them to other agents, and receive money for them. In real life, cash balance dropping to zero is not at all unusual for people who live from paycheck to paycheck.

The conservation law is the key feature for the successful functioning of money. If the agents were permitted to "manufacture" money, they would be printing money

and buying all goods for nothing, which would be a disaster. The physical medium of money is not essential, as long as the conservation law is enforced. Money may be in the form of paper cash, but today it is more often represented by digits in computerized bank accounts. The conservation law is the fundamental principle of accounting, whether in the single-entry or in the double-entry form. More discussion of banks and debt is given in Sect. "Models with Debt".

The Boltzmann–Gibbs Distribution of Money

Having recognized the principle of money conservation, Drăgulescu and Yakovenko [25] argued that the stationary distribution of money should be given by the exponential Boltzmann–Gibbs function analogous to (1):

$$P(m) = c e^{\frac{-m}{T_m}} \ . \tag{7}$$

Here c is a normalizing constant, and T_m is the "money temperature", which is equal to the average amount of money per agent: $T = \langle m \rangle = M/N$, where M is the total money, and N is the number of agents.

To verify this conjecture, Drăgulescu and Yakovenko [25] performed agent-based computer simulations of money transfers between agents. Initially all agents were given the same amount of money, say, \$ 1000. Then, a pair of agents (i, j) were randomly selected, the amount Δm was transferred from one agent to another, and the process was repeated many times. Time evolution of the probability distribution of money $P(m)$ can be seen in computer animation videos at the Web pages [35,36]. After a transitory period, money distribution converges to the stationary form shown in Fig. 1. As expected, the distribution is very well fitted by the exponential function (7).

Several different rules for Δm were considered in [25]. In one model, the amount transferred was fixed to a constant $\Delta m = 1\$$. Economically, it means that all agents were selling their products for the same price $\Delta m = 1\$$. Computer animation [35] shows that the initial distribution of money first broadens to a symmetric, Gaussian curve, characteristic for a diffusion process. Then, the distribution starts to pile up around the $m = 0$ state, which acts as the impenetrable boundary, because of the imposed condition $m \geq 0$. As a result, $P(m)$ becomes skewed (asymmetric) and eventually reaches the stationary exponential shape, as shown in Fig. 1. The boundary at $m = 0$ is analogous to the ground-state energy in statistical physics. Without this boundary condition, the probability distribution of money would not reach a stationary state. Computer animation [35,36] also shows how the entropy of money distribution, defined as $S/N = -\sum_k P(m_k) \ln P(m_k)$, grows from the initial value $S = 0$, when all agents have the same money, to the maximal value at the statistical equilibrium.

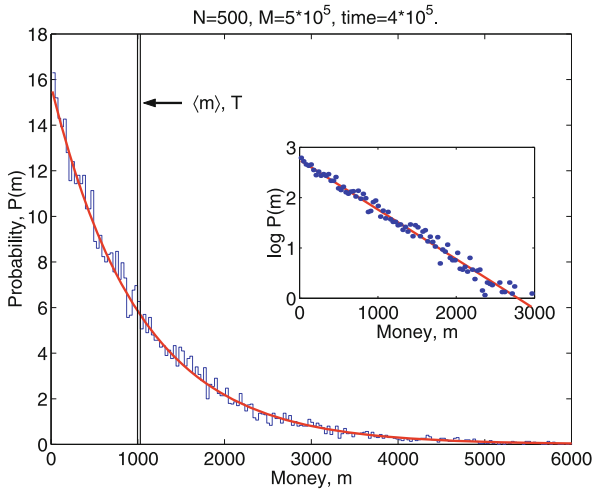

Econophysics, Statistical Mechanics Approach to, Figure 1
Stationary probability distribution of money $P(m)$ obtained in agent-based computer simulations. *Solid curves*: fits to the Boltzmann–Gibbs law (7). *Vertical lines*: the initial distribution of money. (Reproduced from [25])

While the model with $\Delta m = 1$ is very simple and instructive, it is not very realistic, because all prices are taken to be the same. In another model considered in [25], Δm in each transaction is taken to be a random fraction of the average amount of money per agent, i. e., $\Delta m = \nu(M/N)$, where ν is a uniformly distributed random number between 0 and 1. The random distribution of Δm is supposed to represent the wide variety of prices for different products in the real economy. It reflects the fact that agents buy and consume many different types of products, some of them simple and cheap, some sophisticated and expensive. Moreover, different agents like to consume these products in different quantities, so there is variation in the amounts Δm paid, even though the unit price of the same product is constant. Computer simulation of this model produces exactly the same stationary distribution (7) as in the first model. Computer animation for this model is also available on the Web page [35].

The final distribution is universal despite different rules for Δm. To amplify this point further, Drăgulescu and Yakovenko [25] also considered a toy model, where Δm was taken to be a random fraction of the average amount of money of the two agents: $\Delta m = \nu(m_i + m_j)/2$. This model produced the same stationary distribution (7) as the other two models.

The pairwise models of money transfer are attractive in their simplicity, but they represent a rather prim-

itive market. The modern economy is dominated by big firms, which consist of many agents, so Drăgulescu and Yakovenko [25] also studied a model with firms. One agent at a time is appointed to become a "firm". The firm borrows capital K from another agent and returns it with interest hK, hires L agents and pays them wages W, manufactures Q items of a product, sells them to Q agents at price R, and receives profit $F = RQ - LW - hK$. All of these agents are randomly selected. The parameters of the model are optimized following a procedure from economics textbooks [37]. The aggregate demand–supply curve for the product is used in the form $R(Q) = v/Q^{\eta}$, where Q is the quantity consumers would buy at price R, and η and v are some parameters. The production function of the firm has the traditional Cobb–Douglas form: $Q(L, K) = L^{\chi} K^{1-\chi}$, where χ is a parameter. Then the profit of the firm F is maximized with respect to K and L. The net result of the firm activity is a many-body transfer of money, which still satisfies the conservation law. Computer simulation of this model generates the same exponential distribution (7), independently of the model parameters. The reasons for the universality of the Boltzmann–Gibbs distribution and its limitations are discussed from a different perspective in Sect. "Additive Versus Multiplicative Models".

Well after paper [25] appeared, Italian econophysicists [38] found that similar ideas had been published earlier in obscure journals in Italian by Eleonora Bennati [39,40]. They proposed calling these models the Bennati–Drăgulescu–Yakovenko game [41]. The Boltzmann distribution was independently applied to social sciences by Jürgen Mimkes using the Lagrange principle of maximization with constraints [42,43]. The exponential distribution of money was also found in [44] using a Markov chain approach to strategic market games. A long time ago, Benoit Mandelbrot observed (see p. 83 in [45]):

> "There is a great temptation to consider the exchanges of money which occur in economic interaction as analogous to the exchanges of energy which occur in physical shocks between gas molecules".

He realized that this process should result in the exponential distribution, by analogy with the barometric distribution of density in the atmosphere. However, he discarded this idea, because it does not produce the Pareto power law, and proceeded to study the stable Lévy distributions. Ironically, the actual economic data, discussed in Sect. "Empirical Data on Money and Wealth Distributions" and "Empirical Data on Income Distribution", do show the exponential distribution for the majority of the population. Moreover, the data have finite variance, so the stable Lévy distributions are not applicable because of their infinite variance.

Models with Debt

Now let us discuss how the results change when debt is permitted. Debt may be considered as negative money. When an agent borrows money from a bank (considered here as a big reservoir of money), the cash balance of the agent (positive money) increases, but the agent also acquires a debt obligation (negative money), so the total balance (net worth) of the agent remains the same, and the conservation law of total money (positive and negative) is satisfied. After spending some cash, the agent still has the debt obligation, so the money balance of the agent becomes negative. Any stable economic system must have a mechanism preventing unlimited borrowing and unlimited debt. Otherwise, agents can buy any products without producing anything in exchange by simply going into unlimited debt. The exact mechanisms of limiting debt in the real economy are complicated and obscured. Drăgulescu and Yakovenko [25] considered a simple model where the maximal debt of any agent is limited by a certain amount m_{d}. This means that the boundary condition $m_i \geq 0$ is now replaced by the condition $m_i \geq -m_{\mathrm{d}}$ for all agents i. Setting interest rates on borrowed money to be zero for simplicity, Drăgulescu and Yakovenko [25] performed computer simulations of the models described in Sect. "The Boltzmann–Gibbs Distribution of Money" with the new boundary condition. The results are shown in Fig. 2. Not surprisingly, the stationary money distribution again has an exponential shape, but now with the new boundary condition at $m = -m_{\mathrm{d}}$ and the higher money temperature $T_{\mathrm{d}} = m_{\mathrm{d}} + M/N$. By allowing agents to go into debt up to m_{d}, we effectively increase the amount of money available to each agent by m_{d}. So, the money temperature, which is equal to the average amount of effectively available money per agent, increases. A model with nonzero interest rates was also studied in [25].

We see that debt does not violate the conservation law of money, but rather modifies boundary conditions for $P(m)$. When economics textbooks describe how "banks create money" or "debt creates money" [37], they count only positive money (cash) as money, but do not count liabilities (debt obligations) as negative money. With such a definition, money is not conserved. However, if we include debt obligations in the definition of money, then the conservation law is restored. This approach is in agreement with the principles of double-entry accounting, which records both assets and debts. Debt obligations are as real as positive cash, as many borrowers painfully real-

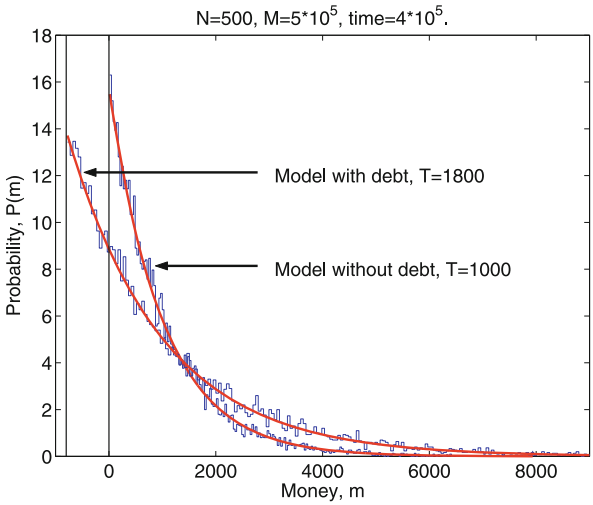

Econophysics, Statistical Mechanics Approach to, Figure 2
Stationary distributions of money with and without debt. The debt is limited to $m_d = 800$. *Solid curves*: fits to the Boltzmann–Gibbs laws with the "temperatures" $T = 1800$ and $T = 1000$. (Reproduced from [25])

ized in their experience. A more detailed study of positive and negative money and bookkeeping from the point of view of econophysics is presented in a series of papers by the physicist Dieter Braun [46,47,48].

One way of limiting the total debt in the economy is the so-called required reserve ratio r [37]. Every bank is required by law to set aside a fraction r of the money deposited with the bank, and this reserved money cannot be loaned further. If the initial amount of money in the system (the money base) is M_0, then with loans and borrowing the total amount of positive money available to the agents increases to $M = M_0/r$, where the factor $1/r$ is called the money multiplier [37]. This is how "banks create money". Where does this extra money come from? It comes from the increase of the total debt in the system. The maximal total debt is equal to $D = M_0/r - M_0$ and is limited by the factor r. When the debt is maximal, the total amounts of positive, M_0/r, and negative, $M_0(1-r)/r$, money circulate between the agents in the system, so there are effectively two conservation laws for each of them [49]. Thus, we expect to see the exponential distributions of positive and negative money characterized by two different temperatures: $T_+ = M_0/rN$ and $T_- = M_0(1-r)/rN$. This is exactly what was found in computer simulations in [49], shown in Fig. 3. Similar two-sided distributions were also found in [47].

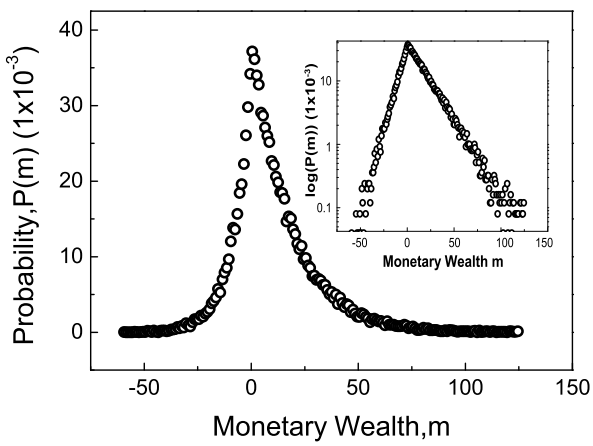

Econophysics, Statistical Mechanics Approach to, Figure 3
The stationary distribution of money for the required reserve ratio $r = 0.8$. The distribution is exponential for positive and negative money with different "temperatures" T_+ and T_-, as illustrated by the *inset* on log–linear scale. (Reproduced from [49])

Proportional Money Transfers and Saving Propensity

In the models of money transfer considered thus far, the transferred amount Δm is typically independent of the money balances of agents. A different model was introduced in the physics literature earlier [50] under the name multiplicative asset exchange model. This model also satisfies the conservation law, but the amount of money transferred is a fixed fraction γ of the payer's money in (5):

$$\Delta m = \gamma m_i \,. \tag{8}$$

The stationary distribution of money in this model, shown in Fig. 4 with an exponential function, is close, but not exactly equal, to the Gamma distribution:

$$P(m) = c m^{\beta} \mathrm{e}^{\frac{-m}{T}} \,. \tag{9}$$

Equation (9) differs from (7) by the power-law prefactor m^{β}. From the Boltzmann kinetic equation (discussed in more detail in Sect. "Additive Versus Multiplicative Models"), Ispolatov et al. [50] derived a formula relating the parameters γ and β in (8) and (9):

$$\beta = \frac{-1 - \ln 2}{\ln(1-\gamma)} \,. \tag{10}$$

When payers spend a relatively small fraction of their money $\gamma < 1/2$, (10) gives $\beta > 0$, so the low-money population is reduced and $P(m \to 0) = 0$, as shown in Fig. 4.

Later, Thomas Lux brought to the attention of physicists [32] that essentially the same model, called the inequality process, had been introduced and studied much earlier by the sociologist John Angle [51,52,53,54,55], see also the review [56] for additional references. While Ispo-

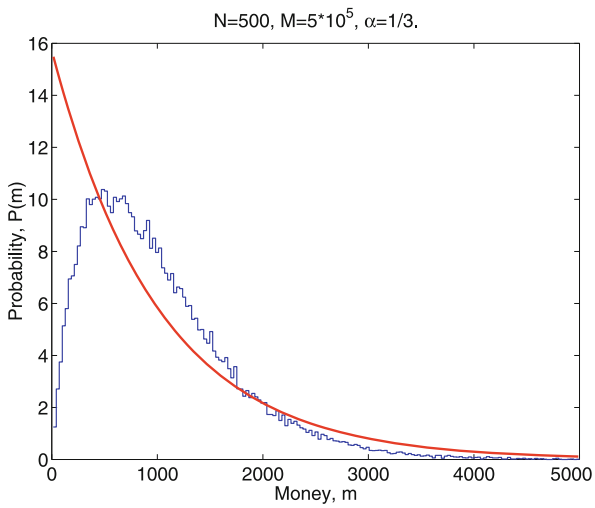

Econophysics, Statistical Mechanics Approach to, Figure 4
Stationary probability distribution of money in the multiplicative random exchange model (8) for $\gamma = 1/3$. *Solid curve*: the exponential Boltzmann–Gibbs law. (Reproduced from [25])

latov et al. [50] did not give much justification for the proportionality law (8), Angle [51] connected this rule with the surplus theory of social stratification [57], which argues that inequality in human society develops when people can produce more than necessary for minimal subsistence. This additional wealth (surplus) can be transferred from original producers to other people, thus generating inequality. In the first paper by Angle [51], the parameter γ was randomly distributed, and another parameter, δ, gave a higher probability of winning to the agent with a higher money balance in (5). However, in the following papers, he simplified the model to a fixed γ (denoted as ω by Angle) and equal probabilities of winning for higher- and lower-balance agents, which makes it completely equivalent to the model of [50]. Angle also considered a model [55,56] where groups of agents have different values of γ, simulating the effect of education and other "human capital". All of these models generate a Gamma-like distribution, well approximated by (9).

Another model with an element of proportionality was proposed in [26]. (This paper originally appeared as a follow-up preprint cond-mat/0004256 to the preprint cond-mat/0001432 of [25].) In this model, the agents set aside (save) some fraction of their money λm_i, whereas the rest of their money balance $(1-\lambda)m_i$ becomes available for random exchanges. Thus, the rule of exchange (5) becomes

$$\begin{aligned} m_i' &= \lambda m_i + \xi(1-\lambda)(m_i + m_j)\,, \\ m_j' &= \lambda m_j + (1-\xi)(1-\lambda)(m_i + m_j)\,. \end{aligned} \tag{11}$$

Here the coefficient λ is called the saving propensity, and the random variable ξ is uniformly distributed between 0 and 1. It was pointed out in [56] that, by the change of notation $\lambda \to (1-\gamma)$, (11) can be transformed to the same form as (8), if the random variable ξ takes only discrete values 0 and 1. Computer simulations [26] of the model (11) found a stationary distribution close to the Gamma distribution (9). It was shown that the parameter β is related to the saving propensity λ by the formula $\beta = 3\lambda/(1-\lambda)$ [38,58,59,60]. For $\lambda \neq 0$, agents always keep some money, so their balances never go to zero and $P(m \to 0) = 0$, whereas for $\lambda = 0$ the distribution becomes exponential.

In the subsequent papers by the Kolkata school [1] and related papers, the case of random saving propensity was studied. In these models, the agents are assigned random parameters λ drawn from a uniform distribution between 0 and 1 [61]. It was found that this model produces a power-law tail $P(m) \propto 1/m^2$ at high m. The reasons for stability of this law were understood using the Boltzmann kinetic equation [60,62,63], but most elegantly in the mean-field theory [64]. The fat tail originates from the agents whose saving propensity is close to 1, who hoard money and do not give it back [38,65]. An interesting matrix formulation of the problem was presented in [66]. Patriarca et al. [67] studied the relaxation rate in the money transfer models. Drăgulescu and Yakovenko [25] studied a model with taxation, which also has an element of proportionality. The Gamma distribution was also studied for conservative models within a simple Boltzmann approach in [68] and using much more complicated rules of exchange in [69,70].

Additive Versus Multiplicative Models

The stationary distribution of money (9) for the models of Sect. "Proportional Money Transfers and Saving Propensity" is different from the simple exponential formula (7) found for the models of Sect. "The Boltzmann–Gibbs Distribution of Money". The origin of this difference can be understood from the Boltzmann kinetic equation [28,71]. This equation describes time evolution of the distribution function $P(m)$ due to pairwise interactions:

$$\begin{aligned} \frac{\mathrm{d}P(m)}{\mathrm{d}t} = \iint \big\{ & -f_{[m,m']\to[m-\Delta,m'+\Delta]} P(m)P(m') \\ & + f_{[m-\Delta,m'+\Delta]\to[m,m']} P(m-\Delta)\cdot P(m'+\Delta) \big\}\, \mathrm{d}m'\mathrm{d}\Delta\,. \end{aligned} \tag{12}$$

Here $f_{[m,m']\to[m-\Delta,m'+\Delta]}$ is the probability of transferring money Δ from an agent with money m to an agent

with money m' per unit time. This probability, multiplied by the occupation numbers $P(m)$ and $P(m')$, gives the rate of transitions from the state $[m, m']$ to the state $[m-\Delta, m'+\Delta]$. The first term in (12) gives the depopulation rate of the state m. The second term in (12) describes the reverse process, where the occupation number $P(m)$ increases. When the two terms are equal, the direct and reverse transitions cancel each other statistically, and the probability distribution is stationary: $\mathrm{d}P(m)/\mathrm{d}t = 0$. This is the principle of detailed balance.

In physics, the fundamental microscopic equations of motion of particles obey time-reversal symmetry. This means that the probabilities of the direct and reverse processes are exactly equal:

$$f_{[m,m']\to[m-\Delta,m'+\Delta]} = f_{[m-\Delta,m'+\Delta]\to[m,m']} \,. \tag{13}$$

When (13) is satisfied, the detailed balance condition for (12) reduces to the equation $P(m)P(m') = P(m-\Delta)P(m'+\Delta)$, because the factors f cancel out. The only solution of this equation is the exponential function $P(m) = c\exp(-m/T_m)$, so the Boltzmann–Gibbs distribution is the stationary solution of the Boltzmann kinetic equation (12). Notice that the transition probabilities (13) are determined by the dynamical rules of the model, but the equilibrium Boltzmann–Gibbs distribution does not depend on the dynamical rules at all. This is the origin of the universality of the Boltzmann–Gibbs distribution. It shows that it may be possible to find out the stationary distribution without knowing details of the dynamical rules (which are rarely known very well), as long as the symmetry condition (13) is satisfied.

The models considered in Sect. "The Boltzmann–Gibbs Distribution of Money" have the time-reversal symmetry. The model with the fixed money transfer Δ has equal probabilities (13) of transferring money from an agent with balance m to an agent with balance m' and vice versa. This is also true when Δ is random, as long as the probability distribution of Δ is independent of m and m'. Thus, the stationary distribution $P(m)$ is always exponential in these models.

However, there is no fundamental reason to expect time-reversal symmetry in economics, so (13) may be not valid. In this case, the system may have a nonexponential stationary distribution or no stationary distribution at all. In model (8), the time-reversal symmetry is broken. Indeed, when an agent i gives a fixed fraction γ of his money m_i to an agent with balance m_j, their balances become $(1-\gamma)m_i$ and $m_j + \gamma m_i$. If we try to reverse this process and appoint an agent j to be the payer and to give the fraction γ of her money, $\gamma(m_j + \gamma m_i)$, to agent i, the system does not return to the original configuration $[m_i, m_j]$. As emphasized by Angle [56], the payer pays a deterministic fraction of his money, but the receiver receives a random amount from a random agent, so their roles are not interchangeable. Because the proportional rule typically violates the time-reversal symmetry, the stationary distribution $P(m)$ in multiplicative models is typically not exactly exponential.[1] Making the transfer dependent on the money balance of the payer effectively introduces a Maxwell's demon into the model. That is why the stationary distribution is not exponential, and, thus, does not maximize entropy (4). Another view on the time-reversal symmetry in economic dynamics is presented in [72].

These examples show that the Boltzmann–Gibbs distribution does not hold for any conservative model. However, it is universal in a limited sense. For a broad class of models that have time-reversal symmetry, the stationary distribution is exponential and does not depend on the details of the model. Conversely, when the time-reversal symmetry is broken, the distribution may depend on the details of the model. The difference between these two classes of models may be rather subtle. Deviations from the Boltzmann–Gibbs law may occur only if the transition rates f in (13) explicitly depend on the agent's money m or m' in an asymmetric manner. Drăgulescu and Yakovenko [25] performed a computer simulation where the direction of payment was randomly selected in advance for every pair of agents (i, j). In this case, money flows along directed links between the agents: $i \to j \to k$, and the time-reversal symmetry is strongly violated. This model is closer to the real economy, where one typically receives money from an employer and pays it to a grocery store. Nevertheless, the Boltzmann–Gibbs distribution was found in this model, because the transition rates f do not explicitly depend on m and m' and do not violate (13).

In the absence of detailed knowledge of real microscopic dynamics of economic exchanges, the semiuniversal Boltzmann–Gibbs distribution (7) is a natural starting point. Moreover, the assumption of [25] that agents pay the same prices Δm for the same products, independent of their money balances m, seems very appropriate for the modern anonymous economy, especially for purchases over the Internet. There is no particular empirical evidence for the proportional rules (8) or (11). However, the differ-

[1] However, when Δm is a fraction of the total money $m_i + m_j$ of the two agents, the model is time-reversible and has an exponential distribution, as discussed in Sect. "The Boltzmann–Gibbs Distribution of Money".

ence between the additive (7) and multiplicative (9) distributions may be not so crucial after all. From the mathematical point of view, the difference is in the implementation of the boundary condition at $m = 0$. In the additive models of Sect. "The Boltzmann–Gibbs Distribution of Money", there is a sharp cutoff of $P(m)$ at $m = 0$. In the multiplicative models of Sect. "Proportional Money Transfers and Saving Propensity", the balance of an agent never reaches $m = 0$, so $P(m)$ vanishes at $m \to 0$ in a power-law manner. At the same time, $P(m)$ decreases exponentially for large m for both models.

By further modifying the rules of money transfer and introducing more parameters in the models, one can obtain even more complicated distributions [73]. However, one can argue that parsimony is the virtue of a good mathematical model, not the abundance of additional assumptions and parameters, whose correspondence to reality is hard to verify.

Statistical Mechanics of Wealth Distribution

In the econophysics literature on exchange models, the terms "money" and "wealth" are often used interchangeably; however, economists emphasize the difference between these two concepts. In this section, we review the models of wealth distribution, as opposed to money distribution.

Models with a Conserved Commodity

What is the difference between money and wealth? One can argue [25] that wealth w_i is equal to money m_i plus the other property that an agent i has. The latter may include durable material property, such as houses and cars, and financial instruments, such as stocks, bonds, and options. Money (paper cash, bank accounts) is generally liquid and countable. However, the other property is not immediately liquid and has to be sold first (converted into money) to be used for other purchases. In order to estimate the monetary value of property, one needs to know the price p. In the simplest model, let us consider just one type of property, say, stocks s. Then the wealth of an agent i is given by the formula

$$w_i = m_i + ps_i \,. \tag{14}$$

It is assumed that the price p is common for all agents and is established by some kind of market process, such as an auction, and may change in time.

It is reasonable to start with a model where both the total money $M = \sum_i m_i$ and the total stock $S = \sum_i s_i$ are conserved [74,75,76]. The agents pay money to buy stock and sell stock to get money, and so on. Although M and S are conserved, the total wealth $W = \sum_i w_i$ is generally not conserved, because of the price fluctuation [75] in (14). This is an important difference from the money transfer models of Sect. "Statistical Mechanics of Money Distribution". Here the wealth w_i of an agent i, not participating in any transactions, may change when transactions between other agents establish a new price p. Moreover, the wealth w_i of an agent i does not change after a transaction with an agent j. Indeed, in exchange for paying money Δm, agent i receives the stock $\Delta s = \Delta m/p$, so her total wealth (14) remains the same. In principle, the agent can instantaneously sell the stock back at the same price and recover the money paid. If the price p never changes, then the wealth w_i of each agent remains constant, despite transfers of money and stock between agents.

We see that redistribution of wealth in this model is directly related to price fluctuations. One mathematical model of this process was studied in [77]. In this model, the agents randomly change preferences for the fraction of their wealth invested in stocks. As a result, some agents offer stock for sale and some want to buy it. The price p is determined from the market-clearing auction matching supply and demand. Silver et al. [77] demonstrated in computer simulations and proved analytically using the theory of Markov processes that the stationary distribution $P(w)$ of wealth w in this model is given by the Gamma distribution, as in (9). Various modifications of this model [32], such as introducing monopolistic coalitions, do not change this result significantly, which shows the robustness of the Gamma distribution. For models with a conserved commodity, Chatterjee and Chakrabarti [75] found the Gamma distribution for a fixed saving propensity and a power law tail for a distributed saving propensity.

Another model with conserved money and stock was studied in [78] for an artificial stock market where traders follow different investment strategies: random, momentum, contrarian, and fundamentalist. Wealth distribution in the model with random traders was found have a power-law tail $P(w) \sim 1/w^2$ for large w. However, unlike in most other simulation, where all agents initially have equal balances, here the initial money and stock balances of the agents were randomly populated according to a power law with the same exponent. This raises the question whether the observed power-law distribution of wealth is an artifact of the initial conditions, because equilibrization of the upper tail may take a very long simulation time.

Models with Stochastic Growth of Wealth

Although the total wealth W is not exactly conserved in the models considered in Sect. "Models with a Conserved Commodity", W nevertheless remains constant on average, because the total money M and stock S are conserved. A different model for wealth distribution was proposed in [27]. In this model, time evolution of the wealth w_i of an agent i is given by the stochastic differential equation

$$\frac{\mathrm{d}w_i}{\mathrm{d}t} = \eta_i(t)w_i + \sum_{j(\neq i)} J_{ij}w_j - \sum_{j(\neq i)} J_{ji}w_i \,, \tag{15}$$

where $\eta_i(t)$ is a Gaussian random variable with mean $\langle\eta\rangle$ and variance $2\sigma^2$. This variable represents growth or loss of wealth of an agent due to investment in stock market. The last two terms describe transfer of wealth between different agents, which is taken to be proportional to the wealth of the payers with the coefficients J_{ij}. So, the model (15) is multiplicative and invariant under the scale transformation $w_i \to Zw_i$. For simplicity, the exchange fractions are taken to be the same for all agents: $J_{ij} = J/N$ for all $i \neq j$, where N is the total number of agents. In this case, the last two terms in (15) can be written as $J(\langle w\rangle - w_i)$, where $\langle w\rangle = \sum_i w_i/N$ is the average wealth per agent. This case represents a "mean-field" model, where all agents feel the same environment. It can be easily shown that the average wealth increases in time as $\langle w\rangle_t = \langle w\rangle_0 \mathrm{e}^{(\langle\eta\rangle+\sigma^2)t}$. Then, it makes more sense to consider the relative wealth $\tilde{w}_i = w_i/\langle w\rangle_t$. Equation (15) for this variable becomes

$$\frac{\mathrm{d}\tilde{w}_i}{\mathrm{d}t} = (\eta_i(t) - \langle\eta\rangle - \sigma^2)\tilde{w}_i + J(1-\tilde{w}_i) \,. \tag{16}$$

The probability distribution $P(\tilde{w}, t)$ for the stochastic differential equation (16) is governed by the Fokker–Planck equation:

$$\frac{\partial P}{\partial t} = \frac{\partial[J(\tilde{w}-1)+\sigma^2\tilde{w}]P}{\partial\tilde{w}} + \sigma^2\frac{\partial}{\partial\tilde{w}}\left(\tilde{w}\frac{\partial(\tilde{w}P)}{\partial\tilde{w}}\right) \,. \tag{17}$$

The stationary solution ($\partial P/\partial t = 0$) of this equation is given by the following formula:

$$P(\tilde{w}) = c\frac{\mathrm{e}^{\frac{-J}{\sigma^2\tilde{w}}}}{\tilde{w}^{\frac{2+J}{\sigma^2}}} \,. \tag{18}$$

The distribution (18) is quite different from the Boltzmann–Gibbs (7) and Gamma (9) distributions. Equation (18) has a power-law tail at large $\tilde{w}$ and a sharp cutoff at small $\tilde{w}$. Equation (15) is a version of the generalized Lotka–Volterra model, and the stationary distribution (18) was also obtained in [79,80]. The model was generalized to include negative wealth in [81].

Bouchaud and Mézard [27] used the mean-field approach. A similar result was found for a model with pairwise interaction between agents in [82]. In this model, wealth is transferred between the agents using the proportional rule (8). In addition, the wealth of the agents increases by the factor $1+\zeta$ in each transaction. This factor is supposed to reflect creation of wealth in economic interactions. Because the total wealth in the system increases, it makes sense to consider the distribution of relative wealth $P(\tilde{w})$. In the limit of continuous trading, Slanina [82] found the same stationary distribution (18). This result was reproduced using a mathematically more involved treatment of this model in [83]. Numerical simulations of the models with stochastic noise η in [69,70] also found a power-law tail for large w.

Let us contrast the models discussed in Sect. "Models with a Conserved Commodity" and "Models with Stochastic Growth of Wealth". In the former case, where money and commodity are conserved and wealth does not grow, the distribution of wealth is given by the Gamma distribution with an exponential tail for large w. In the latter models, wealth grows in time exponentially, and the distribution of relative wealth has a power-law tail for large $\tilde{w}$. These results suggest that the presence of a power-law tail is a nonequilibrium effect that requires constant growth or inflation of the economy, but disappears for a closed system with conservation laws.

Reviews of the models discussed were also given in [84,85]. Because of lack of space, we omit discussion of models with wealth condensation [27,50,86,87,88], where a few agents accumulate a finite fraction of the total wealth, and studies of wealth distribution on networks [89,90,91,92]. Here we discuss the models with long-range interaction, where any agent can exchange money and wealth with any other agent. A local model, where agents trade only with the nearest neighbors, was studied in [93].

Empirical Data on Money and Wealth Distributions

It would be very interesting to compare theoretical results for money and wealth distributions in various models with empirical data. Unfortunately, such empirical data are difficult to find. Unlike income, which is discussed in Sect. "Data and Models for Income Distribution", wealth is not routinely reported by the majority of individuals to the government. However, in many countries, when a person dies, all assets must be reported for the purpose of inheritance tax. So, in principle, there exist good statis-

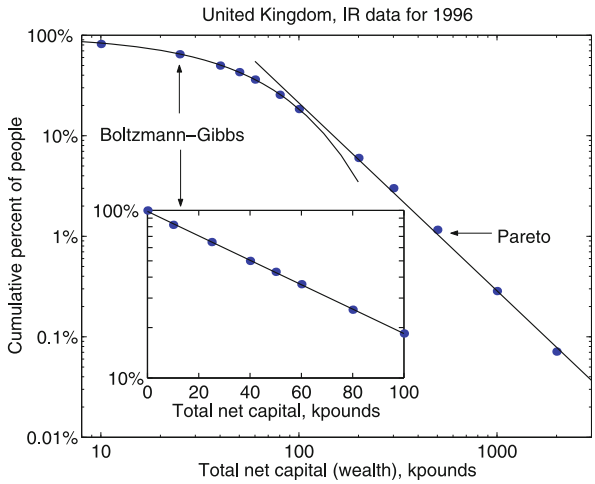

Econophysics, Statistical Mechanics Approach to, Figure 5
Cumulative probability distribution of net wealth in the UK shown on log–log and log–linear (*inset*) scales. *Points* represent the data from the Inland Revenue, and *solid lines* are fits to the exponential (Boltzmann–Gibbs) and power (Pareto) laws. (Reproduced from [95])

tics of wealth distribution among dead people, which, of course, is different from the wealth distribution among living people. Using an adjustment procedure based on the age, gender, and other characteristics of the deceased, the UK tax agency, the Inland Revenue, reconstructed the wealth distribution of the whole population of the UK [94]. Figure 5 shows the UK data for 1996 reproduced from [95]. The figure shows the cumulative probability $C(w) = \int_w^\infty P(w')\mathrm{d}w'$ as a function of the personal net wealth w, which is composed of assets (cash, stocks, property, household goods, etc.) and liabilities (mortgages and other debts). Because statistical data are usually reported at nonuniform intervals of w, it is more practical to plot the cumulative probability distribution $C(w)$ rather than its derivative, the probability density $P(w)$. Fortunately, when $P(w)$ is an exponential or a power-law function, then $C(w)$ is also an exponential or a power-law function.

The cumulative probability distribution in Fig. 5 is plotted on a log–log scale, where a straight line represents a power-law dependence. The figure shows that the distribution follows a power law $C(w) \propto 1/w^\alpha$ with exponent $\alpha = 1.9$ for wealth greater than about £100,000. The inset in Fig. 5 shows the data on log–linear scale, where the straight line represents an exponential dependence. We observe that below £100,000 the data are well fitted by the exponential distribution $C(w) \propto \exp(-w/T_w)$ with the effective "wealth temperature" $T_w =$ £60,000, (which corresponds to the median wealth of £41,000). So, the distribution of wealth is characterized by the Pareto power law in the upper tail of the distribution and the exponential Boltzmann–Gibbs law in the lower part of the distribution for the great majority (about 90%) of the population. Similar results are found for the distribution of income, as discussed in Sect. "Data and Models for Income Distribution". One may speculate that the wealth distribution in the lower part is dominated by distribution of money, because the corresponding people do not have other significant assets, so the results of Sect. "Statistical Mechanics of Money Distribution" give the Boltzmann–Gibbs law. On the other hand, the upper tail of the wealth distribution is dominated by investment assess, where the results of Sect. "Models with Stochastic Growth of Wealth" give the Pareto law. The power law was studied by many researchers for the upper-tail data, such as the Forbes list of the 400 richest people [96,97], but much less attention was paid to the lower part of the wealth distribution. Curiously, Abdul-Magd [98] found that the wealth distribution in ancient Egyptian society was consistent with (18).

For direct comparison with the results of Sect. "Statistical Mechanics of Money Distribution", it would be very interesting to find data on the distribution of money, as opposed to the distribution of wealth. Making a reasonable assumption that most people keep most of their money in banks, one can approximate the distribution of money by the distribution of balances on bank accounts. (Balances on all types of bank accounts, such as checking, saving, and money manager, associated with the same person should be added up.) Despite imperfections (people may have accounts with different banks or not keep all their money in banks), the distribution of balances on bank accounts would give valuable information about the distribution of money. The data for a big enough bank would be representative of the distribution in the whole economy. Unfortunately, it has not been possible to obtain such data thus far, even though it would be completely anonymous and not compromise the privacy of bank clients.

Measuring the probability distribution of money would be very useful, because it determines how much people can, in principle, spend on purchases without going into debt. This is different from the distribution of wealth, where the property component, such as house, car, or retirement investment, is effectively locked up and, in most cases, is not easily available for consumer spending. So, although wealth distribution may reflect the distribution of economic power, the distribution of money is more relevant for consumption. Money distribution can be useful for determining prices that maximize revenue of a manufacturer [25]. If a price p is set too high, few people can afford it, and, if a price is too low, the sales revenue is

small, so the optimal price must be in-between. The fraction of the population who can afford to pay the price p is given by the cumulative probability $C(p)$, so the total sales revenue is proportional to $pC(p)$. For the exponential distribution $C(p) = \exp(-p/T_m)$, the maximal revenue is achieved at $p = T_m$, i. e., at the optimal price equal to the average amount of money per person [25]. Indeed, the prices of mass-market consumer products, such as computers, electronics goods, and appliances, remain stable for many years at a level determined by their affordability to the population, whereas the technical parameters of these products at the same price level improve dramatically owing to technological progress.

Data and Models for Income Distribution

In contrast to money and wealth distributions, a lot more empirical data are available for the distribution of income r from tax agencies and population surveys. In this section, we first present empirical data on income distribution and then discuss theoretical models.

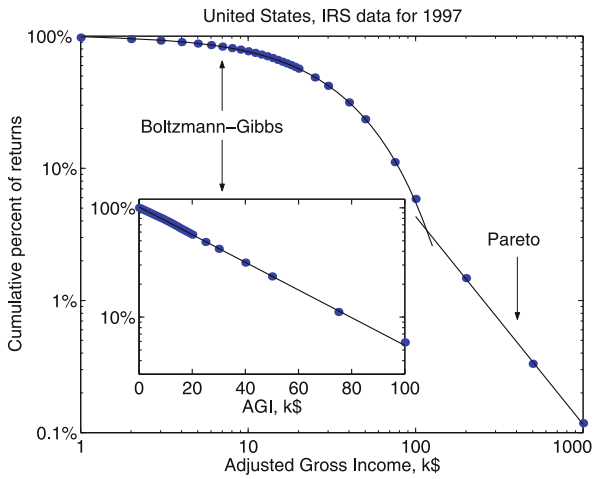

Econophysics, Statistical Mechanics Approach to, Figure 6
Cumulative probability distribution of tax returns for USA in 1997 shown on log–log and log–linear (*inset*) scales. *Points* represent the Internal Revenue Service (IRS) data, and *solid lines* are fits to the exponential and power-law functions. (Reproduced from [103])

Empirical Data on Income Distribution

Empirical studies of income distribution have a long history in the economics literature [99,100,101]. Following the work by Pareto [15], much attention was focused on the power-law upper tail of the income distribution and less on the lower part. In contrast to more complicated functions discussed in the literature, Drăgulescu and Yakovenko [102] introduced a new idea by demonstrating that the lower part of income distribution can be well fitted with a simple exponential function $P(r) = c\exp(-r/T_r)$ characterized by just one parameter, the "income temperature" T_r. Then it was recognized that the whole income distribution can be fitted by an exponential function in the lower part and a power-law function in the upper part [95,103], as shown in Fig. 6. The straight line on the log–linear scale in the inset of Fig. 6 demonstrates the exponential Boltzmann–Gibbs law, and the straight line on the log–log scale in the main panel illustrates the Pareto power law. The fact that income distribution consists of two distinct parts reveals the two-class structure of American society [104,105]. Coexistence of the exponential and power-law distributions is known in plasma physics and astrophysics, where they are called the "thermal" and "superthermal" parts [106,107,108]. The boundary between the lower and upper classes can be defined as the intersection point of the exponential and power-law fits in Fig. 6. For 1997, the annual income separating the two classes was about $120,000. About 3% of the population belonged to the upper class, and 97% belonged to the lower class.

Silva and Yakovenko [105] studied time evolution of income distribution in the USA during 1983–2001 on the basis of data from the Internal Revenue Service (IRS), the government tax agency. The structure of the income distribution was found to be qualitatively the same for all years, as shown in Fig. 7. The average income in nominal dollars approximately doubled during this time interval. So, the horizonal axis in Fig. 7 shows the normalized income r/T_r, where the "income temperature" T_r was obtained by fitting of the exponential part of the distribution for each year. The values of T_r are shown in Fig. 7. The plots for the 1980s and 1990s are shifted vertically for clarity. We observe that the data points in the lower-income part of the distribution collapse on the same exponential curve for all years. This demonstrates that the shape of the income distribution for the lower class is extremely stable and does not change in time, despite the gradual increase of the average income in nominal dollars. This observation suggests that the lower-class distribution is in statistical, "thermal" equilibrium.

On the other hand, Fig. 7 shows that the income distribution in the upper class does not rescale and significantly changes in time. Silva and Yakovenko [105] found that the exponent α of the power law $C(r) \propto 1/r^{\alpha}$ decreased from 1.8 in 1983 to 1.4 in 2000. This means that the upper tail became "fatter". Another useful parameter is the total in-

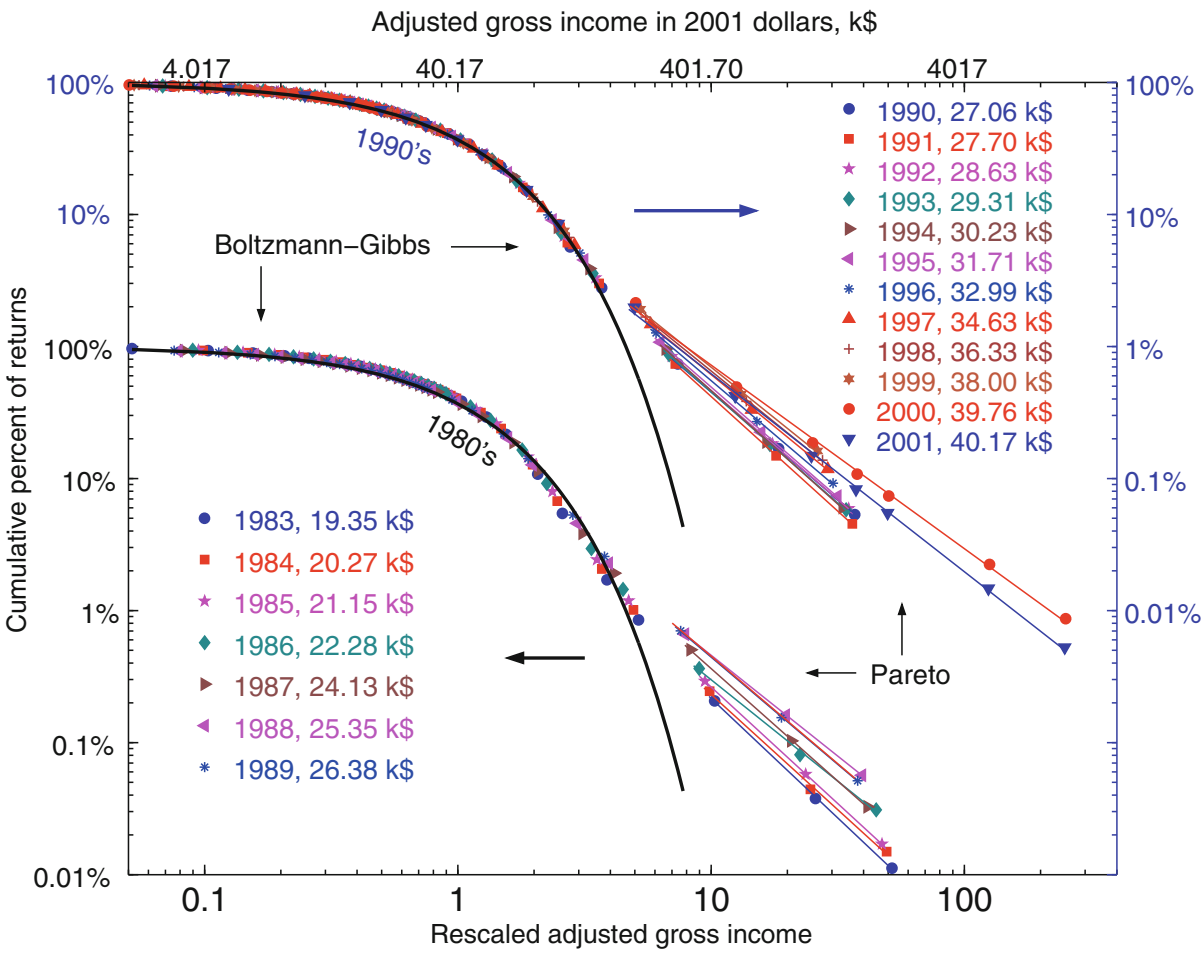

Econophysics, Statistical Mechanics Approach to, Figure 7

Cumulative probability distribution of tax returns plotted on log–log scale versus r/T_r (the annual income r normalized by the average income T_r in the exponential part of the distribution). The IRS data points are for 1983–2001, and the *columns of numbers* give the values of T_r for the corresponding years. (Reproduced from [105])

come of the upper class as the fraction f of the total income in the system. The fraction f increased from 4% in 1983 to 20% in 2000 [105]. However, in 2001, α increased and f decreased, indicating that the upper tail was reduced after the stock market crash at that time. These results indicate that the upper tail is highly dynamical and not stationary. It tends to swell during the stock market boom and shrink during the bust. Similar results were found for Japan [109,110,111,112].

Although relative income inequality within the lower class remains stable, the overall income inequality in the USA has increased significantly as a result of the tremendous growth of the income of the upper class. This is illustrated by the Lorenz curve and the Gini coefficient shown in Fig. 8. The Lorenz curve [99] is a standard way of representing income distribution in the economics literature. It is defined in terms of two coordinates $x(r)$ and $y(r)$ depending on a parameter r:

$$x(r) = \int_0^r P(r')\mathrm{d}r' \,, \qquad y(r) = \frac{\int_0^r r' P(r')\mathrm{d}r'}{\int_0^\infty r' P(r')\mathrm{d}r'} \,. \tag{19}$$

The horizontal coordinate $x(r)$ is the fraction of the population with income below r, and the vertical coordinate $y(r)$ is the fraction of the income this population accounts for. As r changes from 0 to ∞, x and y change from 0 to 1 and parametrically define a curve in the (x, y)-plane.

Figure 8 shows the data points for the Lorenz curves in 1983 and 2000, as computed by the IRS [113]. Drăgulescu and Yakovenko [102] analytically derived the Lorenz curve

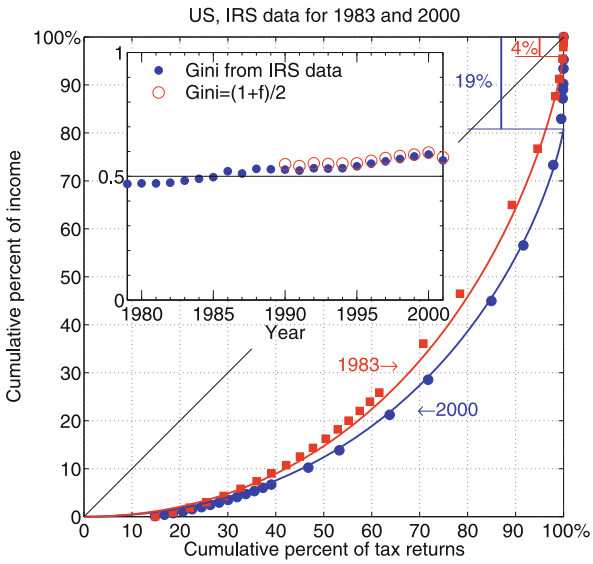

Econophysics, Statistical Mechanics Approach to, Figure 8
Lorenz plots for income distribution in 1983 and 2000. The data points are from the IRS [113], and the theoretical curves represent (20) with *f* from Fig. 7. *Inset*: The *closed circles* are the IRS data 113 for the Gini coefficient *G*, and the *open circles* show the theoretical formula $G = (1 + f)/2$. (Reproduced from [105])

formula $y = x + (1-x)\ln(1-x)$ for a purely exponential distribution $P(r) = c\exp(-r/T_r)$. This formula is shown by the red line in Fig. 8 and describes the 1983 data reasonably well. However, for 2000, it is essential to take into account the fraction f of income in the upper tail, which modifies the Lorenz formula as follows [103,104,105]:

$$y = (1-f)[x + (1-x)\ln(1-x)] + f\Theta(x-1)\,. \quad (20)$$

The last term in (20) represent the vertical jump of the Lorenz curve at $x = 1$, where a very small percentage of the population in the upper class accounts for a substantial fraction f of the total income. The blue curve representing (20) fits the 2000 data in Fig. 8 very well.

The deviation of the Lorenz curve from the straight diagonal line in Fig. 8 is a certain measure of income inequality. Indeed, if everybody had the same income, the Lorenz curve would be a diagonal line, because the fraction of income would be proportional to the fraction of the population. The standard measure of income inequality is the so-called Gini coefficient $0 \le G \le 1$, which is defined as the area between the Lorenz curve and the diagonal line, divided by the area of the triangle beneath the diagonal line [99]. Time evolution of the Gini coefficient, as computed by the IRS [113], is shown in the inset of Fig. 8. Drăgulescu and Yakovenko [102] derived analytically the result that $G = 1/2$ for a purely exponential distribution. In the first approximation, the values of G shown in the inset of Fig. 8 are indeed close to the theoretical value 1/2. If we take into account the upper tail using (20), the formula for the Gini coefficient becomes $G = (1+f)/2$ [105]. The inset in Fig. 8 shows that this formula is a very good fit to the IRS data for the 1990s using the values of f deduced from Fig. 7. The values $G < 1/2$ for the 1980s cannot be captured by this formula, because the Lorenz data points are slightly above the theoretical curve for 1983 in Fig. 8. Overall, we observe that income inequality has been increasing for the last 20 years, because of swelling of the Pareto tail, but decreased in 2001 after the stock market crash.

Thus far we have discussed the distribution of individual income. An interesting related question is the distribution $P_2(r)$ of family income $r = r_1 + r_2$, where r_1 and r_2 are the incomes of spouses. If individual incomes are distributed exponentially $P(r) \propto \exp(-r/T_r)$, then

$$P_2(r) = \int_0^r \mathrm{d}r' P(r')P(r-r') = cr\exp(-r/T_r)\,, \quad (21)$$

where c is a normalization constant. Figure 9 shows that (21) is in good agreement with the family income distribution data from the US Census Bureau [102]. In (21), we assumed that incomes of spouses are uncorrelated. This simple approximation is indeed supported by the scatter plot of incomes of spouses shown in Fig. 10. Each family is represented in this plot by two points (r_1, r_2) and (r_2, r_1) for symmetry. We observe that the density of points is approximately constant along the lines of constant family in-

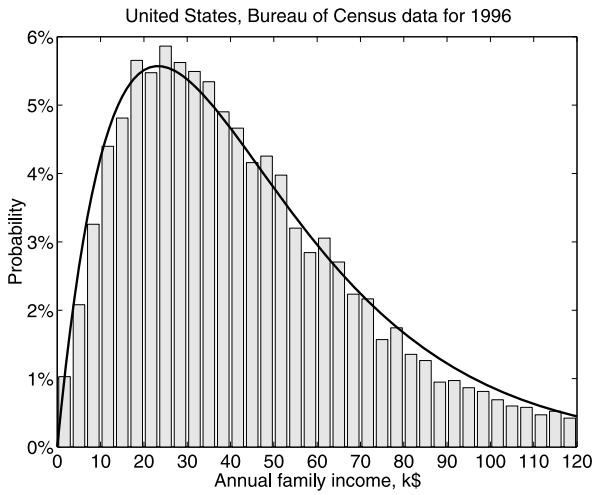

Econophysics, Statistical Mechanics Approach to, Figure 9
Probability distribution of family income for families with two adults (US Census Bureau data). *Solid line*: fit to (21). (Reproduced from [102])

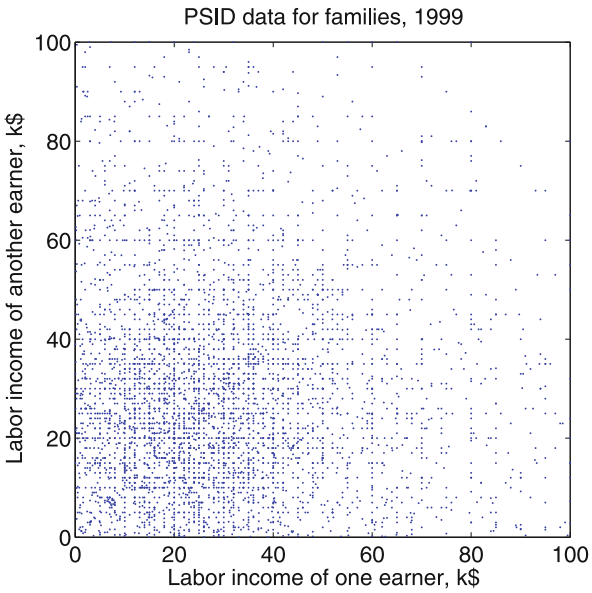

Econophysics, Statistical Mechanics Approach to, Figure 10
Scatter plot of the spouses' incomes (r_1, r_2) and (r_2, r_1) based on the data from the Panel Study of Income Dynamics (PSID). (Reproduced from [103])

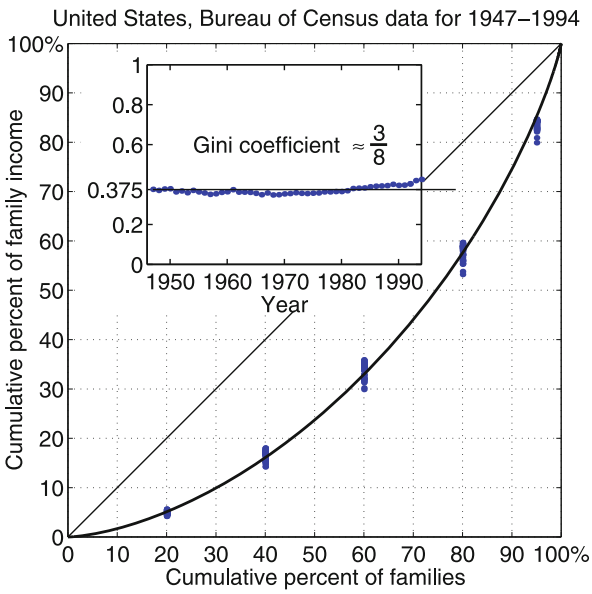

Econophysics, Statistical Mechanics Approach to, Figure 11
Lorenz plot for family income calculated from (21), compared with the US Census data points. *Inset*: The US Census data points for the Gini coefficient for families, compared with the theoretically calculated value 3/8=37.5%. (Reproduced from [102])

come $r_1 + r_2 = \text{const}$, which indicates that incomes of spouses are approximately uncorrelated. There is no significant clustering of points along the diagonal $r_1 = r_2$, i. e., no strong positive correlation of spouses' incomes.

The Gini coefficient for the family income distribution (21) was calculated in [102] as $G = 3/8 = 37.5\%$. Figure 11 shows the Lorenz quintiles and the Gini coefficient for 1947–1994 plotted from the US Census Bureau data. The solid line, representing the Lorenz curve calculated from (21), is in good agreement with the data. The systematic deviation for the top 5% of earners results from the upper tail, which has a less pronounced effect on family income than on individual income, because of income averaging in the family. The Gini coefficient, shown in the inset of Fig. 11, is close to the calculated value of 37.5%. Moreover, the average G for the developed capitalist countries of North America and western Europe, as determined by the World Bank [103], is also close to the calculated value of 37.5%.

Income distribution has been examined in econophysics papers for different countries: Japan [68,109,110,111,112,114,115,116], Germany [117,118], the UK [68,85,116,117,118], Italy [118,119,120], the USA [117,121], India [97], Australia [91,120,122], and New Zealand [68,116]. The distributions are qualitatively similar to the results presented in this section. The upper tail follows a power law and comprises a small fraction of the population. To fit the lower part of the distribution, the use of exponential, Gamma, and log-normal distributions was reported in different papers. Unfortunately, income distribution is often reported by statistical agencies for households, so it is difficult to differentiate between one-earner and two-earner income distributions. Some papers reported the use of interpolating functions with different asymptotic behavior for low and high incomes, such as the Tsallis function [116] and the Kaniadakis function [118]. However, the transition between the lower and upper classes is not smooth for the US data shown in Figs. 6 and 7, so such functions would not be useful in this case. The special case is income distribution in Argentina during the economic crisis, which shows a time-dependent bimodal shape with two peaks [116].

Theoretical Models of Income Distribution

Having examined the empirical data on income distribution, let us now discuss theoretical models. Income r_i is the influx of money per unit time to an agent i. If the money balance m_i is analogous to energy, then the income r_i would be analogous to power, which is the energy flux per unit time. So, one should conceptually distinguish between the distributions of money and income. While money is regularly transferred from one agent to another in pairwise

transactions, it is not typical for agents to trade portions of their income. Nevertheless, indirect transfer of income may occur when one employee is promoted and another demoted, while the total annual budget is fixed, or when one company gets a contract, whereas another one loses it, etc. A reasonable approach, which has a long tradition in the economics literature [123,124,125], is to treat individual income r as a stochastic process and study its probability distribution. In general, one can study a Markov process generated by a matrix of transitions from one income to another. In the case where income r changes by a small amount Δr over a time period Δt, the Markov process can be treated as income diffusion. Then one can apply the general Fokker–Planck equation [71] to describe evolution in time t of the income distribution function $P(r,t)$ [105]:

$$\begin{aligned} \frac{\partial P}{\partial t} &= \frac{\partial}{\partial r}\left[AP + \frac{\partial(BP)}{\partial r}\right] , \\ A &= -\frac{\langle \Delta r\rangle}{\Delta t} , \; B = \frac{\langle (\Delta r)^2\rangle}{2\Delta t} . \end{aligned} \tag{22}$$

The coefficients A and B in (22) are determined by the first and second moments of income changes per unit time. The stationary solution $\partial_t P = 0$ of (22) obeys the following equation with the general solution:

$$\begin{aligned} \frac{\partial(BP)}{\partial r} &= -AP , \\ P(r) &= \frac{c}{B(r)} \exp\left(-\int^r \frac{A(r')}{B(r')} \mathrm{d}r'\right) . \end{aligned} \tag{23}$$

For the lower part of the distribution, it is reasonable to assume that Δr is independent of r, i. e., the changes of income are independent of income itself. This process is called additive diffusion [105]. In this case, the coefficients in (22) are constants A_0 and B_0. Then (23) gives the exponential distribution $P(r) \propto \exp(-r/T_r)$, with the effective income temperature $T_r = B_0/A_0$. Notice that a meaningful stationary solution (23) requires that $A > 0$, i. e., $\langle \Delta r\rangle < 0$. The coincidence of this result with the Boltzmann–Gibbs exponential law (1) and (7) is not accidental. Indeed, instead of considering pairwise interaction between particles, one can derive (1) by considering energy transfers between a particle and a big reservoir, as long as the transfer process is "additive" and does not involve a Maxwell-demon-like discrimination. Stochastic income fluctuations are described by a similar process. So, although money and income are different concepts, they may have similar distributions, because they are governed by similar mathematical principles. It was shown explicitly in [25,82,83] that the models of pairwise money transfer can be described in a certain limit by the Fokker–Planck equation.

On the other hand, for the upper tail of the income distribution, it is reasonable to expect that $\Delta r \propto r$, i. e., income changes are proportional to income itself. This is known as the proportionality principle of Gibrat [123], and the process is called multiplicative diffusion [105]. In this case, $A = ar$ and $B = br^2$, and (23) gives the power-law distribution $P(r) \propto 1/r^{\alpha+1}$, with $\alpha = 1 + a/b$.

Generally, lower-class income comes from wages and salaries, where the additive process is appropriate, whereas upper-class income comes from bonuses, investments, and capital gains, calculated in percentages, where the multiplicative process applies [126]. However, the additive and multiplicative processes may coexist. An employee may receive a cost-of-living rise calculated in percentages (the multiplicative process) and a merit rise calculated in dollars (the additive process). In this case, we have $A = A_0 + ar$ and $B = B_0 + br^2 = b(r_0^2 + r^2)$, where $r_0^2 = B_0/b$. Substituting these expressions into (23), we find

$$P(r) = c\frac{\mathrm{e}^{-(\frac{r_0}{T_r})\arctan(\frac{r}{r_0})}}{[1+(\frac{r}{r_0})^2]^{\frac{1+a}{2b}}} . \tag{24}$$

The distribution (24) interpolates between the exponential law for low r and the power law for high r, because either the additive or the multiplicative process dominates in the corresponding limit. The crossover between the two regimes takes place at $r \sim r_0$, where the additive and multiplicative contributions to B are equal. The distribution (24) has three parameters: the "income temperature" $T_r = A_0/B_0$, the Pareto exponent $\alpha = 1 + a/b$, and the crossover income r_0. It is a minimal model that captures the salient features of the empirical income distribution shown in Fig. 6. A mathematically similar, but more economically oriented model was proposed in [114,115], where labor income and asset accumulation are described by the additive and multiplicative processes correspondingly. A general stochastic process with additive and multiplicative noise was studied numerically in [127], but the stationary distribution was not derived analytically. A similar process with discrete time increments was studied by Kesten [128]. Recently, a formula similar to (24) was obtained in [129].

To verify the multiplicative and additive hypotheses empirically, it is necessary to have data on income mobility, i. e., the income changes Δr of the same people from one year to another. The distribution of income changes $P(\Delta r|r)$ conditional on income r is generally not available publicly, although it can be reconstructed by researchers at the tax agencies. Nevertheless, the multiplicative hypothesis for the upper class was quantitatively verified

in [111,112] for Japan, where tax identification data is published for the top taxpayers.

The power-law distribution is meaningful only when it is limited to high enough incomes $r > r_0$. If all incomes r from 0 to ∞ follow a purely multiplicative process, then one can change to a logarithmic variable $x = \ln(r/r_*)$ in (22) and show that it gives a Gaussian time-dependent distribution $P_t(x) \propto \exp(-x^2/2\sigma^2 t)$ for x, i. e., the log-normal distribution for r, also known as the Gibrat distribution [123]. However, the width of this distribution increases linearly in time, so the distribution is not stationary. This was pointed out by Kalecki [124] a long time ago, but the log-normal distribution is still widely used for fitting income distribution, despite this fundamental logical flaw in its justification. In a classic paper, Champernowne [125] showed that a multiplicative process gives a stationary power-law distribution when a boundary condition is imposed at $r_0 \neq 0$. Later, this result was rediscovered by econophysicists [130,131]. In our (24), the exponential distribution of the lower class effectively provides such a boundary condition for the power law of the upper class. Notice also that (24) reduces to (18) in the limit $r_0 \to 0$, which corresponds to purely multiplicative noise $B = br^2$.

There are alternative approaches to income distribution in the economics literature. One of them, proposed by Lydall [132], involves social hierarchy. Groups of people have leaders, who have leaders of a higher order, and so on. The number of people decreases geometrically (exponentially) with the increase of the hierarchical level. If individual income increases by a certain factor (i. e., multiplicatively) when moving to the next hierarchical level, then income distribution follows a power law [132]. However, the original argument of Lydall can be easily modified to produce an exponential distribution. If individual income increases by a certain amount, i. e., income increases linearly with the hierarchical level, then income distribution is exponential. The latter process seems to be more realistic for moderate incomes below \$ 100,000. A similar scenario is the Bernoulli trials [133], where individuals have a constant probability of increasing their income by a fixed amount. We see that the deterministic hierarchical models and the stochastic models of additive and multiplicative income mobility represent essentially the same ideas.

Other Applications of Statistical Physics

Statistical physics was applied to a number of other subjects in economics. Because of lack of space, only two such topics are briefly discussed in this section.

Economic Temperatures in Different Countries

As discussed in Sect. "Empirical Data on Money and Wealth Distributions" and "Empirical Data on Income Distribution", the distributions of money, wealth, and income are often described by exponential functions for the majority of the population. These exponential distributions are characterized by the parameters T_m, T_w, and T_r, which are mathematically analogous to temperature in the Boltzmann–Gibbs distribution (1). The values of these parameters, extracted from the fits of the empirical data, are generally different for different countries, i. e., different countries have different economic "temperatures". For example, Drăgulescu and Yakovenko [95] found that the US income temperature was 1.9 times higher than the UK income temperature in 1998 (using the exchange rate of dollars to pounds at that time). Also, there was $\pm 25\%$ variation between income temperatures of different states within the USA. [95].

In physics, a difference of temperatures allows one to set up a thermal machine. In was argued in [25] that the difference of money or income temperatures between different countries allows one to extract profit in international trade. Indeed, as discussed at the end of Sect. "Empirical Data on Money and Wealth Distributions", the prices of goods should be commensurate with money or income temperature, because otherwise people cannot afford to buy those goods. So, an international trading company can buy goods at a low price T_1 in a "low-temperature" country and sell them at a high price T_2 in a "high-temperature" country. The difference of prices $T_2 - T_1$ would be the profit of the trading company. In this process, money (the analog of energy) flows from the "high-temperature" to the "low-temperature" country, in agreement with the second law of thermodynamics, whereas products flow in the opposite direction. This process very much resembles what is going on in the global economy now. In this framework, the perpetual trade deficit of the USA is the consequence of the second law of thermodynamics and the difference of temperatures between the USA and "low-temperature" countries, such as China. Similar ideas were developed in more detail in [134,135], including a formal Carnot cycle for international trade.

The statistical physics approach demonstrates that profit originates from statistical nonequilibrium (the difference of temperatures), which exists in the global economy. However, it does not answer the question what is the origin of this difference. By analogy with physics, one would expect that the money flow should reduce the temperature difference and, eventually, lead to equilibrization

of temperatures. In physics, this situation is known as the "thermal death of the universe". In a completely equilibrated global economy, it would be impossible to make profit by exploiting differences of economic temperatures between different countries. Although globalization of the modern economy does show a tendency toward equilibrization of living standards in different countries, this process is far from straightforward, and there are many examples contrary to equilibrization. This interesting and timely subject certainly requires further study.

Society as a Binary Alloy

In 1971, Thomas Schelling [136] proposed the now-famous mathematical model of segregation. He considered a lattice, where the sites can be occupied by agents of two types, e. g., blacks and whites in the problem of racial segregation. He showed that if the agents have some probabilistic preference for the neighbors of the same type, the system spontaneously segregates into black and white neighborhoods. This mathematical model is similar to the so-called Ising model, which is a popular model for studying phase transitions in physics. In this model, each lattice site is occupied by a magnetic atom, whose magnetic moment has only two possible orientations, up or down. The interaction energy between two neighboring atoms depends on whether their magnetic moments point in the same or in the opposite directions. In physics language, the segregation found by Schelling represents a phase transition in this system.

Another similar model is the binary alloy, a mixture of two elements which attract or repel each other. It was noticed in [137] that the behavior of actual binary alloys is strikingly similar to social segregation. In the following papers [42,138], this mathematical analogy was developed further and compared with social data. Interesting concepts, such as the coexistence curve between two phases and the solubility limit, were discussed in this work. The latter concept means that a small amount of one substance dissolves in another up to some limit, but phase separation (segregation) develops for higher concentrations. Recently, similar ideas were rediscovered in [139,140,141]. The vast experience of physicists in dealing with phase transitions and alloys may be helpful for practical applications of such models [142].

Future Directions, Criticism, and Conclusions

The statistical models described in this review are quite simple. It is commonly accepted in physics that theoretical models are not intended to be photographic copies of reality, but rather to be caricatures, capturing the most essential features of a phenomenon with a minimal number of details. With only few rules and parameters, the models discussed in Sect. "Statistical Mechanics of Money Distribution", "Statistical Mechanics of Wealth Distribution", and "Data and Models for Income Distribution" reproduce spontaneous development of stable inequality, which is present in virtually all societies. It is amazing that the calculated Gini coefficients, $G = 1/2$ for individuals and $G = 3/8$ for families, are actually very close to the US income data, as shown in Figs. 8 and 11. These simple models establish a baseline and a reference point for development of more sophisticated and more realistic models. Some of these future directions are outlined below.

Future Directions

Agents with a Finite Lifespan The models discussed in this review consider immortal agents who live forever, like atoms. However, humans have a finite lifespan. They enter the economy as young people and exit at an old age. Evolution of income and wealth as functions of age is studied in economics using the so-called overlapping-generations model. The absence of the age variable was one of the criticisms of econophysics by the economist Paul Anglin [31]. However, the drawback of the standard overlapping-generations model is that there is no variation of income and wealth between agents of the same age, because it is a representative-agent model. It would be best to combine stochastic models with the age variable. Also, to take into account inflation of average income, (22) should be rewritten for relative income, in the spirit of (17). These modifications would allow one to study the effects of demographic waves, such as baby boomers, on the distributions of income and wealth.

Agent-Based Simulations of the Two-Class Society The empirical data presented in Sect. "Empirical Data on Income Distribution" show quite convincingly that the US population consists of two very distinct classes characterized by different distribution functions. However, the theoretical models discussed in Sect. "Statistical Mechanics of Money Distribution" and "Statistical Mechanics of Wealth Distribution" do not produce two classes, although they do produce broad distributions. Generally, not much attention has been paid in the agent-based literature to simulation of two classes. One exception is [143], in which spontaneous development of employers and employees classes from initially equal agents was simulated [36]. More work in this direction would be certainly desirable.

Access to Detailed Empirical Data A great amount of statistical information is publicly available on the Internet, but not for all types of data. As discussed in Sect. "Empirical Data on Money and Wealth Distributions", it would be very interesting to obtain data on the distribution of balances on bank accounts, which would give information about the distribution of money (as opposed to wealth). As discussed in Sect. "Theoretical Models of Income Distribution", it would be useful to obtain detailed data on income mobility, to verify the additive and multiplicative hypotheses for income dynamics. Income distribution is often reported as a mix of data on individual income and family income, when the counting unit is a tax return (joint or single) or a household. To have a meaningful comparison with theoretical models, it is desirable to obtain clean data where the counting unit is an individual. Direct collaboration with statistical agencies would be very useful.

Economies in Transition Inequality in developed capitalist countries is generally quite stable. The situation is very different for former socialist countries making a transition to a market economy. According to the World Bank data [103], the average Gini coefficient for family income in eastern Europe and the former Soviet Union jumped from 25% in 1988 to 47% in 1993. The Gini coefficient in the socialist countries before the transition was well below the equilibrium value of 37.5% for market economies. However, the fast collapse of socialism left these countries out of market equilibrium and generated a much higher inequality. One may expect that, with time, their inequality will decrease to the equilibrium value of 37.5%. It would be very interesting to trace how fast this relaxation takes place. Such a study would also verify whether the equilibrium level of inequality is universal for all market economies.

Relation to Physical Energy The analogy between energy and money discussed in Sect. "Conservation of Money" is a formal mathematical analogy. However, actual physical energy with low entropy (typically in the form of fossil fuel) also plays a very important role in the modern economy, as the basis of current human technology. In view of the looming energy and climate crisis, it is imperative to find realistic ways for making a transition from the current "disposable" economy based on "cheap" and "unlimited" energy and natural resources to a sustainable one. Heterogeneity of human society is one of the important factors affecting such a transition. Econophysics, at the intersection of energy, entropy, economy, and statistical physics, may play a useful role in this quest [144].

Criticism from Economists

As econophysics is gaining popularity, some criticism has appeared from economists [31], including those who are closely involved with the econophysics movement [32,33,34]. This reflects a long-standing tradition in economic and social sciences of writing critiques on different schools of thought. Much of the criticism is useful and constructive and is already being accommodated in econophysics work. However, some criticism results from misunderstanding or miscommunication between the two fields and some from significant differences in scientific philosophy. Several insightful responses to the criticism have been published [145,146,147]; see also [7,148]. In this section, we briefly address the issues that are directly related to the material discussed in this review.

Awareness of Previous Economics Literature One complaint of [31,32,33,34] is that physicists are not well aware of the previous economics literature and either rediscover known results or ignore well-established approaches. To address this issue, it is useful to keep in mind that science itself is a complex system, and scientific progress is an evolutionary process with natural selection. The sea of scientific literature is enormous, and nobody knows it all. Recurrent rediscovery of regularities in the natural and social world only confirms their validity. Independent rediscovery usually brings a different perspective, broader applicability range, higher accuracy, and better mathematical treatment, so there is progress even when some overlap with previous results exists. Physicists are grateful to economists for bringing relevant and specific references to their attention. Since the beginning of modern econophysics, many old references have been uncovered and are now routinely cited.

However, not all old references are relevant to the new development. For example, Gallegati et al. [33] complained that the econophysics literature on income distribution ignores the so-called Kuznets hypothesis [149]. The Kuznets hypothesis postulates that income inequality first rises during an industrial revolution and then decreases, producing an inverted-U-shaped curve. Gallegati et al. [33] admitted that, to date, the large amount of literature on the Kuznets hypothesis is inconclusive. Kuznets [149] mentioned that this hypothesis applies to the period from colonial times to the 1970s; however, the empirical data for this period are sparse and not very reliable. The econophysics literature deals with reliable volumes of data for the second half of the twentieth century, collected with the introduction of computers. It is not clear what is the definition of industrial revolution

and when exactly it starts and ends. The chain of technological progress seems to be continuous (steam engine, internal combustion engine, cars, plastics, computers, Internet), so it is not clear where the purported U-curve is supposed to be placed in time. Thus, the Kuznets hypothesis appears to be, in principle, unverifiable and unfalsifiable. The original paper by Kuznets [149] actually does not contain any curves, but it has one table filled with made-up, imaginary data! Kuznets admits that he has "neither the necessary data nor a reasonably complete theoretical model" (p. 12 in [149]). So, this paper is understandably ignored by the econophysics community. In fact, the data analysis for 1947–1984 shows amazing stability of income distribution [150], consistent with Fig. 11. The increase of inequality in the 1990s resulted from growth of the upper class relative to the lower class, but the relative inequality within the lower class remains very stable, as shown in Fig. 7.

Reliance on Visual Data Analysis Another complaint of [33] is that econophysicists favor graphic analysis of data over the formal and "rigorous" testing prescribed by mathematical statistics, as favored by economists. This complaint goes against the trend of all sciences to use increasingly sophisticated data visualization for uncovering regularities in complex systems. The thick IRS publication 1304 [151] is filled with data tables, but has virtually no graphs. Despite the abundance of data, it gives a reader no idea about income distribution, whereas plotting the data immediately gives insight. However, intelligent plotting is the art with many tools, which not many researchers have mastered. The author completely agrees with Gallegati et al. [33] that too many papers mindlessly plot any kind of data on a log–log scale, pick a finite interval, where any smooth curved line can be approximated by a straight line, and claim that there is a power law. In many cases, replotting the same data on a log–linear scale converts a curved line into a straight line, which means that the law is actually exponential.

Good visualization is extremely helpful in identifying trends in complex data, which can then be fitted to a mathematical function; however, for a complex system, such a fit should not be expected with infinite precision. The fundamental laws of physics, such as Newton's law of gravity or Maxwell's equations, are valid with enormous precision. However, the laws in condensed matter physics, uncovered by experimentalists with a combination of visual analysis and fitting, usually have much lower precision, at best 10% or so. Most of these laws would fail the formal criteria of mathematical statistics. Nevertheless these approximate laws are enormously useful in practice, and everyday devices engineered on the basis of these laws work very well for all of us.

Because of the finite accuracy, different functions may produce equally good fits. Discrimination between the exponential, Gamma, and log-normal functions may not be always possible [122]. However, the exponential function has fewer fitting parameters, so it is preferable on the basis of simplicity. The other two functions can simply mimic the exponential function with a particular choice of the additional parameters [122]. Unfortunately, many papers in mathematical statistics introduce too many fitting parameters into complicated functions, such as the generalized beta distribution mentioned in [33]. Such overparameterization is more misleading than insightful for data fitting.

Quest for Universality Gallegati et al. [33] criticized physicists for trying to find universality in economics data. They also seemed to equate the concepts of power law, scaling, and universality. These are three different, albeit overlapping, concepts. Power laws usually apply only to a small fraction of data at the high ends of various distributions. Moreover, the exponents of these power laws are usually nonuniversal and vary from case to case. Scaling means that the shape of a function remains the same when its scale changes. However, the scaling function does not have to be a power-law function. A good example of scaling is shown in Fig. 7, where income distributions for the lower class collapse on the same exponential line for about 20 years of data. We observe amazing universality of income distribution, unrelated to a power law. In a general sense, the diffusion equation is universal, because it describes a wide range of systems, from dissolution of sugar in water to a random walk in the stock market.

Universalities are not easy to uncover, but they form the backbone of regularities in the world around us. This is why physicists are so interested in them. Universalities establish the first-order effect, and deviations represent the second-order effect. Different countries may have somewhat different distributions, and economists often tend to focus on these differences. However, this focus on details misses the big picture that, in the first approximation, the distributions are quite similar and universal.

Theoretical Modeling of Money, Wealth, and Income It was pointed out in [31,33,34] that many econophysics papers confuse or misuse the terms for money, wealth, and income. It is true that terminology is sloppy in many papers and should be refined. However, the terms in [25,26] are quite precise, and this review clearly distinguishes between these concepts in Sect. "Statistical Mechanics of Money Distribution", "Statistical Mechanics of Wealth

Distribution", and "Data and Models for Income Distribution".

One contentious issue is about conservation of money. Gallegati et al. [33] agree that "transactions are a key economic process, and they are necessarily conservative", i. e., money is indeed conserved in transactions between agents. However, Anglin [31], Gallegati et al. [33], and Lux [34] complain that the models of conservative exchange do not consider production of goods, which is the core economic process and the source of economic growth. Material production is indeed the ultimate goal of an economy, but it does not violate conservation of money by itself. One can grow coffee beans, but nobody can grow money on a money tree. Money is an artificial economic device that is designed to be conserved. As explained in Sect. "Statistical Mechanics of Money Distribution", the money transfer models implicitly assume that money in transactions is voluntarily paid for goods and services generated by production for the mutual benefit of the parties. In principle, one can introduce a billion variables to keep track of every coffee bean and other product of the economy. What difference would it make for the distribution of *money*? Despite the claims in [31,33], there is no contradiction between models of conservative exchange and the classic work of Adam Smith and David Ricardo. The difference is only in the focus: We keep track of money, whereas they keep track of coffee beans, from production to consumption. These approaches address different questions, but do not contradict each other. Because money constantly circulates in the system as payment for production and consumption, the resulting statistical distribution of money may very well not depend on what exactly is produced and in what quantities.

In principle, the models with random transfers of money should be considered as a reference point for developing more sophisticated models. Despite the totally random rules and "zero intelligence" of the agents, these models develop well-characterized, stable, and stationary distributions of money. One can modify the rules to make the agents more intelligent and realistic and see how much the resulting distribution changes relative to the reference one. Such an attempt was made in [32] by modifying the model of [77] with various more realistic economic ingredients. However, despite the modifications, the resulting distributions were essentially the same as in the original model. This example illustrates the typical robustness and universality of statistical models: Modifying details of microscopic rules does not necessarily change the statistical outcome.

Another misconception, elaborated in [32,34], is that the money transfer models discussed in Sect. "Statistical Mechanics of Money Distribution" imply that money is transferred by fraud, theft, and violence, rather than voluntarily. One should keep in mind that the catchy labels "theft-and-fraud", "marriage-and-divorce", and "yard-sale" were given to the money transfer models by the journalist Brian Hayes [152] in a popular article. Econophysicists who originally introduced and studied these models do not subscribe to this terminology, although the early work of Angle [51] did mention violence as one source of redistribution. In the opinion of the author, it is indeed difficult to justify the proportionality rule (8), which implies that agents with high balances pay proportionally greater amounts in transactions than agents with low balances. However, the additive model of [25], where money transfers Δm are independent of money balances m_i of the agents, does not have this problem. As explained in Sect. "The Boltzmann–Gibbs Distribution of Money", this model simply means that all agents pay the same prices for the same product, although prices may be different for different products. So, this model is consistent with voluntary transactions in a free market.

McCauley [145] argued that conservation of money is violated by credit. As explained in Sect. "Models with Debt", credit does not violate conservation law, but creates positive and negative money without changing net worth. Negative money (debt) is as real as positive money. McCauley [145] claimed that money can be easily created with the tap of a computer key via credit. Then why would an employer not tap the key and double salaries, or a funding agency double research grants? Because budget constraints are real. Credit may provide a temporary relief, but sooner or later it has to be paid back. Allowing debt may produce a double-exponential distribution as shown in Fig. 3, but it does not change the distribution fundamentally.

As discussed in Sect. "Conservation of Money", a central bank or a central government can inject new money into the economy. As discussed in Sect. "Statistical Mechanics of Wealth Distribution", wealth is generally not conserved. As discussed in Sect. "Data and Models for Income Distribution", income is different from money and is described by a different model (22). However, the empirical distribution of income shown in Fig. 6 is qualitatively similar to the distribution of wealth shown in Fig. 5, and we do not have data on money distribution.

Conclusions

The "invasion" of physicists into economics and finance at the turn of the millennium is a fascinating phenomenon. The physicist Joseph McCauley proclaims that "Econo-

physics will displace economics in both the universities and boardrooms, simply because what is taught in economics classes doesn't work" [153]. Although there is some truth in his arguments [145], one may consider a less radical scenario. Econophysics may become a branch of economics, in the same way as games theory, psychological economics, and now agent-based modeling became branches of economics. These branches have their own interests, methods, philosophy, and journals. The main contribution from the infusion of new ideas from a different field is not in answering old questions, but in raising new questions. Much of the misunderstanding between economists and physicists happens not because they are getting different answers, but because they are answering different questions.

The subject of income and wealth distributions and social inequality was very popular at the turn of another century and is associated with the names of Pareto, Lorenz, Gini, Gibrat, and Champernowne, among others. Following the work by Pareto, attention of researchers was primarily focused on the power laws. However, when physicists took a fresh, unbiased look at the empirical data, they found a different, exponential law for the lower part of the distribution. The motivation for looking at the exponential law, of course, came from the Boltzmann–Gibbs distribution in physics. Further studies provided a more detailed picture of the two-class distribution in a society. Although social classes have been known in political economy since Karl Marx, the realization that they are described by simple mathematical distributions is quite new. Demonstration of the ubiquitous nature of the exponential distribution for money, wealth, and income is one of the new contributions produced by econophysics.

Bibliography

Primary Literature

1. Chakrabarti BK (2005) Econophys-Kolkata: a short story. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) Econophysics of Wealth Distributions. Springer, Milan, pp 225–228
2. Carbone A, Kaniadakis G, Scarfone AM (2007) Where do we stand on econophysics? Phys A 382:xi–xiv
3. Stanley HE et al (1996) Anomalous fluctuations in the dynamics of complex systems: from DNA and physiology to econophysics. Phys A 224:302–321
4. Mantegna RN, Stanley HE (1999) An introduction to econophysics: correlations and complexity in finance. Cambridge University Press, Cambridge
5. Galam S (2004) Sociophysics: a personal testimony. Phys A 336:49–55
6. Galan S, Gefen Y, Shapir Y (1982) Sociophysics: a new approach of sociological collective behaviour. I. Mean-behaviour description of a strike. J Math Soc 9:1–13
7. Stauffer D (2004) Introduction to statistical physics outside physics. Phys A 336:1–5
8. Schweitzer F (2003) Brownian agents and active particles: collective dynamics in the natural and social sciences. Springer, Berlin
9. Weidlich W (2000) Sociodynamics: a systematic approach to mathematical modeling in the social sciences. Harwood Academic Publishers, Amsterdam
10. Chakrabarti BK, Chakraborti A, Chatterjee A (eds) (2006) Econophysics and sociophysics: trends and perspectives. Wiley-VCH, Berlin
11. Ball P (2002) The physical modelling of society: a historical perspective. Phys A 314:1–14
12. Ball P (2004) Critical mass: how one thing leads to another. Farrar, Straus and Giroux, New York
13. Boltzmann L (1905) Populäre Schriften. Barth, Leipzig, p 360
14. Austrian Central Library for Physics (2006) Ludwig Boltzmann 1844–1906. ISBN 3-900490-11-2. Vienna
15. Pareto V (1897) Cours d'Économie Politique. L'Université de Lausanne
16. Mirowski P (1989) More heat than light: economics as social physics, physics as nature's economics. Cambridge University Press, Cambridge
17. Majorana E (1942) Il valore delle leggi statistiche nella fisica e nelle scienze sociali. Scientia 36:58–66 (English translation by Mantegna RN in: Bassani GF (ed) (2006) Ettore Majorana Scientific Papers. Springer, Berlin, pp 250–260)
18. Montroll EW, Badger WW (1974) Introduction to quantitative aspects of social phenomena. Gordon and Breach, New York
19. Föllmer H (1974) Random economies with many interacting agents. J Math Econ 1:51–62
20. Blume LE (1993) The statistical mechanics of strategic interaction. Games Econ Behav 5:387–424
21. Foley DK (1994) A statistical equilibrium theory of markets. J Econ Theory 62:321–345
22. Durlauf SN (1997) Statistical mechanics approaches to socioeconomic behavior. In: Arthur WB, Durlauf SN, Lane DA (eds) The Economy as a Complex Evolving System II. Addison-Wesley, Redwood City, pp 81–104
23. Anderson PW, Arrow KJ, Pines D (eds) (1988) The economy as an evolving complex system. Addison-Wesley, Reading
24. Rosser JB (2008) Econophysics. In: Blume LE, Durlauf SN (eds) New Palgrave Dictionary of Economics, 2nd edn. Macmillan, London (in press)
25. Drăgulescu AA, Yakovenko VM (2000) Statistical mechanics of money. Europ Phys J B 17:723–729
26. Chakraborti A, Chakrabarti BK (2000) Statistical mechanics of money: how saving propensity affects its distribution. Europ Phys J B 17:167–170
27. Bouchaud JP, Mézard M (2000) Wealth condensation in a simple model of economy. Phys A 282:536–545
28. Wannier GH (1987) Statistical physics. Dover, New York
29. Lopez-Ruiz R, Sanudo J, Calbet X (2007) Geometrical derivation of the Boltzmann factor. Available via DIALOG. http://arxiv.org/abs/0707.4081. Accessed 1 Jul 2008
30. Lopez-Ruiz R, Sanudo J, Calbet X (2007) On the equivalence of the microcanonical and the canonical ensembles: a geometrical approach. Available via DIALOG. http://arxiv.org/abs/0708.1866. Accessed 1 Jul 2008
31. Anglin P (2005) Econophysics of wealth distributions: a comment. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds)

Econophysics of wealth distributions. Springer, New York, pp 229–238
32. Lux T (2005) Emergent statistical wealth distributions in simple monetary exchange models: a critical review. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) Econophysics of wealth distributions. Springer, Milan, pp 51–60
33. Gallegati M, Keen S, Lux T, Ormerod P (2006) Worrying trends in econophysics. Phys A 370:1–6
34. Lux T (2008) Applications of statistical physics in finance and economics. In: Rosser JB (ed) Handbook of complexity research. Edward Elgar, Cheltenham, UK and Northampton, MA (in press)
35. Computer animation videos of money-transfer models. http://www2.physics.umd.edu/~yakovenk/econophysics/animation.html. Accessed 1 Jul 2008
36. Wright I (2007) Computer simulations of statistical mechanics of money in mathematica. Available via DIALOG. http://demonstrations.wolfram.com/StatisticalMechanicsOfMoney. Accessed 1 Jul 2008
37. McConnell CR, Brue SL (1996) Economics: principles, problems, and policies. McGraw-Hill, New York
38. Patriarca M, Chakraborti A, Kaski K, Germano G (2005) Kinetic theory models for the distribution of wealth: Power law from overlap of exponentials. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) Econophysics of wealth distributions. Springer, Milan, pp 93–110
39. Bennati E (1988) Un metodo di simulazione statistica per l'analisi della distribuzione del reddito. Rivista Internazionale di Scienze Economiche e Commerciali 35:735–756
40. Bennati E (1993) Il metodo di Montecarlo nell'analisi economica. Rassegna di Lavori dell'ISCO (Istituto Nazionale per lo Studio della Congiuntura), Anno X 4:31–79
41. Scalas E, Garibaldi U, Donadio S (2006) Statistical equilibrium in simple exchange games I: methods of solution and application to the Bennati–Drăgulescu–Yakovenko (BDY) game. Europ Phys J B 53:267–272
42. Mimkes J (2000) Society as a many-particle system. J Therm Anal Calorim 60:1055–1069
43. Mimkes J (2005) Lagrange principle of wealth distribution. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) Econophysics of wealth distributions. Springer, Milan, pp 61–69
44. Shubik M (1999) The theory of money and financial institutions, vol 2. MIT Press, Cambridge, p 192
45. Mandelbrot B (1960) The Pareto-Lévy law and the distribution of income. Int Econ Rev 1:79–106
46. Braun D (2001) Assets and liabilities are the momentum of particles and antiparticles displayed in Feynman-graphs. Phys A 290:491–500
47. Fischer R, Braun D (2003) Transfer potentials shape and equilibrate monetary systems. Phys A 321:605–618
48. Fischer R, Braun D (2003) Nontrivial bookkeeping: a mechanical perspective. Phys A 324:266–271
49. Xi N, Ding N, Wang Y (2005) How required reserve ratio affects distribution and velocity of money. Phys A 357:543–555
50. Ispolatov S, Krapivsky PL, Redner S (1998) Wealth distributions in asset exchange models. Europ Phys J B 2:267–276
51. Angle J (1986) The surplus theory of social stratification and the size distribution of personal wealth. Soc Forces 65:293–326
52. Angle J (1992) The inequality process and the distribution of income to blacks and whites. J Math Soc 17:77–98
53. Angle J (1992) Deriving the size distribution of personal wealth from 'the rich get richer, the poor get poorer'. J Math Soc 18:27–46
54. Angle J (1996) How the Gamma Law of income distribution appears invariant under aggregation. J Math Soc 21:325–358
55. Angle J (2002) The statistical signature of pervasive competition on wage and salary incomes. J Math Soc 26:217–270
56. Angle J (2006) The Inequality Process as a wealth maximizing process. Phys A 367:388–414
57. Engels F (1972) The origin of the family, private property and the state, in the light of the researches of Lewis H. Morgan. International Publishers, New York
58. Patriarca M, Chakraborti A, Kaski K (2004) Gibbs versus non-Gibbs distributions in money dynamics. Phys A 340:334–339
59. Patriarca M, Chakraborti A, Kaski K (2004) Statistical model with a standard Gamma distribution. Phys Rev E 70:016104
60. Repetowicz P, Hutzler S, Richmond P (2005) Dynamics of money and income distributions. Phys A 356:641–654
61. Chatterjee A, Chakrabarti BK, Manna SS (2004) Pareto law in a kinetic model of market with random saving propensity. Phys A 335:155-163
62. Das A, Yarlagadda S (2005) An analytic treatment of the Gibbs-Pareto behavior in wealth distribution. Phys A 353:529–538
63. Chatterjee S, Chakrabarti BK, Stinchcombe RB (2005) Master equation for a kinetic model of a trading market and its analytic solution. Phys Rev E 72:026126
64. Mohanty PK (2006) Generic features of the wealth distribution in ideal-gas-like markets. Phys Rev E 74:011117
65. Patriarca M, Chakraborti A, Germano G (2006) Influence of saving propensity on the power-law tail of the wealth distribution. Phys A 369:723–736
66. Gupta AK (2006) Money exchange model and a general outlook. Phys A 359:634–640
67. Patriarca M, Chakraborti A, Heinsalu E, Germano G (2007) Relaxation in statistical many-agent economy models. Europ Phys J B 57:219–224
68. Ferrero JC (2004) The statistical distribution of money and the rate of money transference. Phys A 341:575–585
69. Scafetta N, Picozzi S, West BJ (2004) An out-of-equilibrium model of the distributions of wealth. Quant Financ 4:353–364
70. Scafetta N, Picozzi S, West BJ (2004) A trade-investment model for distribution of wealth. Physica D 193:338–352
71. Lifshitz EM, Pitaevskii LP (1981) Physical kinetics. Pergamon Press, Oxford
72. Ao P (2007) Boltzmann–Gibbs distribution of fortune and broken time reversible symmetry in econodynamics. Commun Nonlinear Sci Numer Simul 12:619–626
73. Scafetta N, West BJ (2007) Probability distributions in conservative energy exchange models of multiple interacting agents. J Phys Condens Matter 19:065138
74. Chakraborti A, Pradhan S, Chakrabarti BK (2001) A self-organising model of market with single commodity. Phys A 297:253–259
75. Chatterjee A, Chakrabarti BK (2006) Kinetic market models with single commodity having price fluctuations. Europ Phys J B 54:399–404
76. Ausloos M, Pekalski A (2007) Model of wealth and goods dynamics in a closed market. Phys A 373:560–568

77. Silver J, Slud E, Takamoto K (2002) Statistical equilibrium wealth distributions in an exchange economy with stochastic preferences. J Econ Theory 106:417–435
78. Raberto M, Cincotti S, Focardi SM, Marchesi M (2003) Traders' long-run wealth in an artificial financial market. Comput Econ 22:255–272
79. Solomon S, Richmond P (2001) Power laws of wealth, market order volumes and market returns. Phys A 299:188–197
80. Solomon S, Richmond P (2002) Stable power laws in variable economies; Lotka-Volterra implies Pareto-Zipf. Europ Phys J B 27:257–261
81. Huang DW (2004) Wealth accumulation with random redistribution. Phys Rev E 69:057103
82. Slanina F (2004) Inelastically scattering particles and wealth distribution in an open economy. Phys Rev E 69:046102
83. Cordier S, Pareschi L, Toscani G (2005) On a kinetic model for a simple market economy. J Statist Phys 120:253–277
84. Richmond P, Repetowicz P, Hutzler S, Coelho R (2006) Comments on recent studies of the dynamics and distribution of money. Phys A 370:43–48
85. Richmond P, Hutzler S, Coelho R, Repetowicz P (2006) A review of empirical studies and models of income distributions in society. In: Chakrabarti BK, Chakraborti A Chatterjee A (eds) Econophysics and sociophysics: trends and perspectives. Wiley-VCH, Berlin
86. Burda Z, Johnston D, Jurkiewicz J, Kaminski M, Nowak MA, Papp G, Zahed I (2002) Wealth condensation in Pareto macroeconomies. Phys Rev E 65:026102
87. Pianegonda S, Iglesias JR, Abramson G, Vega JL (2003) Wealth redistribution with conservative exchanges. Phys A 322:667–675
88. Braun D (2006) Nonequilibrium thermodynamics of wealth condensation. Phys A 369:714–722
89. Coelho R, Néda Z, Ramasco JJ, Santos MA (2005) A family-network model for wealth distribution in societies. Phys A 353:515–528
90. Iglesias JR, Gonçalves S, Pianegonda S, Vega JL, Abramson G (2003) Wealth redistribution in our small world. Phys A 327:12–17
91. Di Matteo T, Aste T, Hyde ST (2004) Exchanges in complex networks: income and wealth distributions. In: Mallamace F, Stanley HE (eds) The physics of complex systems (New advances and perspectives). IOS Press, Amsterdam, p 435
92. Hu MB, Jiang R, Wu QS, Wu YH (2007) Simulating the wealth distribution with a Richest-Following strategy on scale-free network. Phys A 381:467–472
93. Bak P, Nørrelykke SF, Shubik M (1999) Dynamics of money. Phys Rev E 60:2528–2532
94. Her Majesty Revenue and Customs (2003) Distribution of personal wealth. Available via DIALOG. http://www.hmrc.gov.uk/stats/personal_wealth/wealth_oct03.pdf. Accessed 1 Jul 2008
95. Drăgulescu AA, Yakovenko VM (2001) Exponential and power-law probability distributions of wealth and income in the United Kingdom and the United States. Phys A 299:213–221
96. Klass OS, Biham O, Levy M, Malcai O, Solomon S (2007) The Forbes 400, the Pareto power-law and efficient markets. Europ Phys J B 55:143–147
97. Sinha S (2006) Evidence for power-law tail of the wealth distribution in India. Phys A 359:555–562
98. Abul-Magd AY (2002) Wealth distribution in an ancient Egyptian society. Phys Rev E 66:057104
99. Kakwani N (1980) Income Inequality and Poverty. Oxford University Press, Oxford
100. Champernowne DG, Cowell FA (1998) Economic inequality and income distribution. Cambridge University Press, Cambridge
101. Atkinson AB, Bourguignon F (eds) (2000) Handbook of income distribution. Elsevier, Amsterdam
102. Drăgulescu AA, Yakovenko VM (2001) Evidence for the exponential distribution of income in the USA. Europ Phys J B 20:585–589
103. Drăgulescu AA, Yakovenko VM (2003) Statistical mechanics of money, income, and wealth: a short survey. In: Garrido PL, Marro J (eds) Modeling of complex systems: seventh granada lectures, Conference Proceedings 661. American Institute of Physics, New York, pp 180–183
104. Yakovenko VM, Silva AC (2005) Two-class structure of income distribution in the USA: Exponential bulk and power-law tail. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) Econophysics of wealth distributions. Springer, Milan, pp 15–23
105. Silva AC, Yakovenko VM (2005) Temporal evolution of the 'thermal' and 'superthermal' income classes in the USA during 1983-2001. Europhys Lett 69:304–310
106. Hasegawa A, Mima K, Duong-van M (1985) Plasma distribution function in a superthermal radiation field. Phys Rev Lett 54:2608–2610
107. Desai MI, Mason GM, Dwyer JR, Mazur JE, Gold RE, Krimigis SM, Smith CW, Skoug RM (2003) Evidence for a suprathermal seed population of heavy ions accelerated by interplanetary shocks near 1 AU. Astrophys J 588:1149–1162
108. Collier MR (2004) Are magnetospheric suprathermal particle distributions (κ functions) inconsistent with maximum entropy considerations? Adv Space Res 33:2108–2112
109. Souma W (2001) Universal structure of the personal income distribution. Fractals 9:463–470
110. Souma W (2002) Physics of personal income. In: Takayasu H (ed) Empirical science of financial fluctuations: the advent of econophysics. Springer, Tokyo, pp 343–352
111. Fujiwara Y, Souma W, Aoyama H, Kaizoji T, Aoki M (2003) Growth and fluctuations of personal income. Phys A 321:598–604
112. Aoyama H, Souma W, Fujiwara Y (2003) Growth and fluctuations of personal and company's income. Phys A 324:352–358
113. Strudler M, Petska T, Petska R (2003) An analysis of the distribution of individual income and taxes, 1979–2001. The Internal Revenue Service, Washington DC. Available via DIALOG. http://www.irs.gov/pub/irs-soi/03strudl.pdf. Accessed 1 Jul 2008
114. Souma W, Nirei M (2005) Empirical study and model of personal income. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) Econophysics of wealth distributions. Springer, Milan, pp 34–42
115. Nirei M, Souma W (2007) A two factor model of income distribution dynamics. Rev Income Wealth 53:440–459
116. Ferrero JC (2005) The monomodal, polymodal, equilibrium and nonequilibrium distribution of money. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) Econophysics of wealth distributions. Springer, Milan, pp 159–167
117. Clementi F, Gallegati M (2005) Pareto's law of income distribution: evidence for Germany, the United Kingdom, the

United States. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) Econophysics of wealth distributions. Springer, Milan, pp 3–14
118. Clementi F, Gallegati M, Kaniadakis G (2007) κ-generalized statistics in personal income distribution. Europ Phys J B 57:187–193
119. Clementi F, Gallegati M (2005) Power law tails in the Italian personal income distribution. Phys A 350:427–438
120. Clementi F, Di Matteo T, Gallegati M (2006) The power-law tail exponent of income distributions. Phys A 370:49–53
121. Rawlings PK, Reguera D, Reiss H (2004) Entropic basis of the Pareto law. Phys A 343:643–652
122. Banerjee A, Yakovenko VM, Di Matteo T (2006) A study of the personal income distribution in Australia. Phys A 370:54–59
123. Gibrat R (1931) Les Inégalités Economiques. Sirely, Paris
124. Kalecki M (1945) On the Gibrat distribution. Econometrica 13:161–170
125. Champernowne DG (1953) A model of income distribution. Econ J 63:318–351
126. Milaković M (2005) Do we all face the same constraints? In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) Econophysics of wealth distributions. Springer, Milan, pp 184–191
127. Takayasu H, Sato AH, Takayasu M (1997) Stable infinite variance fluctuations in randomly amplified Langevin systems. Phys Rev Lett 79:966–969
128. Kesten H (1973) Random difference equations and renewal theory for products of random matrices. Acta Math 131:207–248
129. Fiaschi D, Marsili M (2007) Distribution of wealth: theoretical microfoundations and empirical evidence. Working paper. Avialable via DIALOG. http://www.dse.ec.unipi.it/persone/docenti/fiaschi/Lavori/distributionWealthMicrofoundations.pdf. Accessed 1 Jul 2008
130. Levy M, Solomon S (1996) Power laws are logarithmic Boltzmann laws. Int J Mod Phys C 7:595–751
131. Sornette D, Cont R (1997) Convergent multiplicative processes repelled from zero: power laws and truncated power laws. J Phys I (France) 7:431–444
132. Lydall HF (1959) The distribution of employment incomes. Econometrica 27:110–115
133. Feller W (1966) An Introduction to Probability Theory and Its Applications, vol 2. Wiley, New York, p 10
134. Mimkes J, Aruka Y (2005) Carnot process of wealth distribution. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) Econophysics of wealth distributions. Springer, Milan, pp 70–78
135. Mimkes J (2006) A thermodynamic formulation of economics. In: Chakrabarti BK, Chakraborti A Chatterjee A (eds) Econophysics and sociophysics: trends and perspectives. Wiley-VCH, Berlin, pp 1–33
136. Schelling TC (1971) Dynamic models of segregation. J Math Soc 1:143–186
137. Mimkes J (1995) Binary alloys as a model for the multicultural society. J Therm Anal 43:521–537
138. Mimkes J (2006) A thermodynamic formulation of social science. In: Chakrabarti BK, Chakraborti A, Chatterjee A (eds) Econophysics and sociophysics: trends and perspectives. Wiley-VCH, Berlin
139. Jego C, Roehner BM (2007) A physicist's view of the notion of "racism". Available via DIALOG. http://arxiv.org/abs/0704.2883. Accessed 1 Jul 2008
140. Stauffer D, Schulze C (2007) Urban and scientific segregation: the Schelling-Ising model. Avialable via DIALOG. http://arxiv.org/abs/0710.5237. Accessed 1 Jul 2008
141. Dall'Asta L, Castellano C, Marsili M (2007) Statistical physics of the Schelling model of segregation. Available via DIALOG. http://arxiv.org/abs/0707.1681. Accessed 1 Jul 2008
142. Lim M, Metzler R, Bar-Yam Y (2007) Global pattern formation and ethnic/cultural violence. Science 317:1540–1544
143. Wright I (2005) The social architecture of capitalism. Phys A 346:589–620
144. Defilla S (2007) A natural value unit – Econophysics as arbiter between finance and economics. Phys A 382:42–51
145. McCauley JL (2006) Response to 'Worrying Trends in Econophysics'. Phys A 371:601–609
146. Richmond P, Chakrabarti BK, Chatterjee A, Angle J (2006) Comments on 'Worrying Trends in Econophysics': income distribution models. In: Chatterjee A, Chakrabarti BK (eds) Econophysics of stock and other markets. Springer, Milan, pp 244–253
147. Rosser JB (2006) Debating the Role of Econophysics. Working paper. Available via DIALOG. http://cob.jmu.edu/rosserjb/. Accessed 1 Jul 2008
148. Rosser JB (2006) The nature and future of econophysics. In: Chatterjee A, Chakrabarti BK (eds) Econophysics of stock and other markets. Springer, Milan, pp 225–234
149. Kuznets S (1955) Economic growth and income inequality. Am Econ Rev 45:1–28
150. Levy F (1987) Changes in the distribution of American family incomes, 1947 to 1984. Science 236:923–927
151. Internal Revenue Service (1999) Statistics of Income–1997, Individual Income Tax Returns. Publication 1304, Revision 12-99, Washington DC
152. Hayes B (2002) Follow the money. Am Sci 90:400–405
153. Ball P (2006) Econophysics: culture crash. Nature 441:686–688

Books and Reviews

McCauley J (2004) Dynamics of markets: econophysics and finance. Cambridge University Press, Cambridge
Farmer JD, Shubik M, Smith E (2005) Is economics the next physical science? Phys Today 58(9):37–42
Samanidou E, Zschischang E, Stauffer D, Lux T (2007) Agent-based models of financial markets. Rep Prog Phys 70:409–450
Econophysics forum. Avialable via DIALOG. http://www.unifr.ch/econophysics/. Accessed 1 Jul 2008

Elastic Percolation Networks

Phillip M. Duxbury
Michigan State University, East Lansing, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Basic Theoretical Concepts

Idealized Experiments
Chalcogenide Glasses
Gels and Semiflexible Rod Networks
Granular Media
Exact Solution on Bethe Lattices
Exact Algorithms and Percolative Geometries
Elastic Critical Behavior
Final Remarks and Future Directions
Bibliography

Glossary

Boson peak The Boson peak is an excess of low frequency modes observed in glasses, as manifested, for example, in inelastic neutron scattering data. In rigidity percolation the Boson peak is related to the number of floppy modes.

Constraint Edges in a graph constrain the degrees of freedom of the nodes in the graph. If edges are independent, each edge acts as one constraint.

Degrees of freedom In d dimensions a point object has d degrees of freedom, while a body has $d(d+1)/2$ degrees of freedom due to rotations and translations.

Floppy mode A floppy mode is a deformation of a structure which is soft and in ideal models is treated as a zero energy deformation.

Generic rigidity A network is generic if none of its edges are dependent due to the particular geometric arrangement of the nodes in the network. With high probability random networks are generic, while regular lattices are non-generic.

Isostatic network An isostatic network is rigid but has no redundant bonds. Isostatic networks are marginally rigid as removal of any edge induces a floppy mode. Ideal generic granular media are isostatic at jamming.

Redundant bond A redundant bond is not essential to the rigidity of a structure. Generic networks which contain redundant bonds are overconstrained and internally stressed.

Rigidity percolation threshold The rigidity threshold marks the transition from floppy networks which have zero elastic moduli to rigid networks with finite elastic moduli.

Definition of the Subject

Materials or structures with sufficiently low connectivity are floppy and have very low elastic moduli, while at high connectivity they are rigid and have relatively high elastic moduli. Elastic percolation networks describe the transition from floppy to rigid that occurs as the network connectivity increases. The percolative geometry and elastic behavior near percolation are of particular interest. Conventional percolative geometries describe some experimental systems however the elastic critical behavior falls into several different universality classes. Moreover, distinct percolative geometries occur in systems with only central forces or which have soft torsional forces and in these cases both the geometry and elastic behavior may be distinct from conventional percolation. Granular media manifest a further distinct elastic percolation network, with the concept of an isostatic network underlying elastic behavior near jamming. This rich fundamental research framework is relevant to an enormous range of materials of scientific and technical interest [79], including physical and chemical gels [84,87], semiflexible networks in biology [6,42], proteins [79], chalcogenide glasses [10,79] and granular media [85]. This brief review outlines the broad underlying principles common to these diverse systems.

Introduction

Many materials exhibit a sharp change in elastic behavior as the degree of interconnection in the material is increased. One well studied example is the gel transition, in which a liquid becomes a gel (solid) as the number of short-range crosslinks increases [18,28,73,84,87]. A similar behavior occurs in epoxies where a polymer melt is irreversibly crosslinked by addition of crosslinking agents to form a stiff, hard solid [18,28,72,73,84]. Crosslinking can be invoked using a variety of stimuli, including chemical, microwave or thermal methods, given the appropriate precursors. Vulcanization in rubber formation is another example. Following the work by de Gennes [17], there has been considerable debate about the elastic scaling behavior occurring near the gelation point. Semiflexible rod networks in biology, including actin filaments in the cell and F filaments in blood clots, [6,42] are also gels. Moreover, paper is composed of cellulose rods and their crosslinking is critical to their mechanical performance [1]. As elucidated in Sects. "Idealized Experiments"–"Granular Media" the ideas of floppy modes and network rigidity are useful in understanding the elastic behavior of these materials, particularly near the gel point.

Elastic percolation is richer than conductivity percolation in that transmission of stress is a vector process while transmission of current only requires simple connectivity. Nevertheless, in some cases simple connectivity is sufficient to enable transmission of stress, while in others a more highly connected structure is required [67,79]. At the engineering level, the reasons for these distinctions have been known at least since the time of Maxwell [48],

though explicit models to elucidate the various types of elastic percolation processes in disordered media were only developed in the early 1980's [24,41]. Chalcogenide glasses, for example $Se_{1-x}Ge_x$, are miscible and provide a unique system in which to study the effect of increasing the crosslinking of low coordination networks. Selenium is relatively floppy as it is two-fold coordinated, while germanium is tetrahedrally bonded and rigid [59,78]. Though computational studies of these glasses indicate critical behavior near average co-ordination $r_c = 2.4$ [32], the experimental consensus is for rather smooth elastic moduli near threshold, presumably due to rounding effects such as dihedral forces or entropic elasticity. Nevertheless quantities such as the number of floppy modes in the Boson peak [40] and Raman scattering [26], do have a clear experimental signature near r_c.

The jamming transition of hard particles is different than the examples listed above as hard core repulsion provides a strong resistance to compression while there is no resistance to extension [2,30,55,85]. However in colloidal gels where there is a weak attractive potential, the behavior may be restored to that of rigidity percolation. Even in the absence of these attractive terms, some of the ideas of rigidity percolation are very useful in granular media, particularly the idea of an isostatic network [21,55]. In this picture, the onset of rigidity in a granular packing occurs at a stress free isostatic critical point. By considering systems with soft repulsive potentials, the elastic behavior in compression can be described by deviations from this critical point, as evident in recent large scale computational studies [57]. Experiments generally support the idea of an isostatic critical point [30,46,85].

The onset of elastic percolation can be determined by making realistic models and by studying their response to an applied stress. This strategy is important, however a deeper understanding has emerged through combinatorial methods which identify constraints that reduce the number of floppy modes [38,52,59,78]. These constraint counting methods enable determination of the ability of a structure or graph to transmit stress, as developed by Maxwell for engineering structures [48] (see Fig. 1). Understanding of stress bearing geometries enables a broader view of elasticity percolation, unifying a broad range of experimental examples and elucidating the aspects which are universal and those that are not. It also enables unification of perspectives from several disciplines (as illustrated in Fig. 1), including engineering, mathematics, material science, physics and more recently biological physics. Algorithms and concepts developed in the graph theory and topology communities have proved particularly rich [34,43,83]. These methods and their relation to elastic percolation are surveyed in Sect. "Exact Algorithms and Percolative Geometries". Constraint counting methods also enable an exact analysis on Bethe lattices [22] which demonstrate that rigidity percolation is often first order (see Sect. "Exact Solution on Bethe Lattices"). From

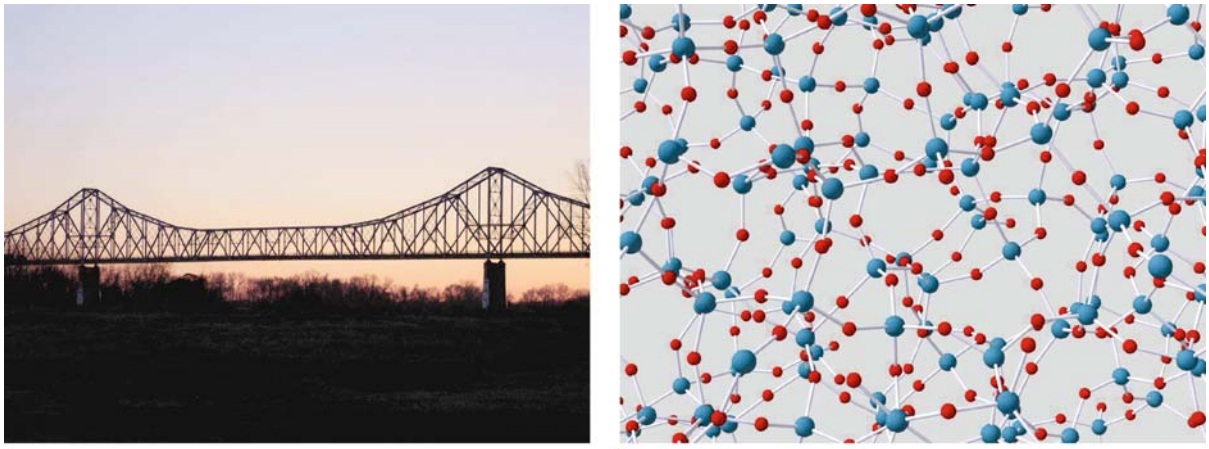

a b

Elastic Percolation Networks, Figure 1

Constraint counting methods are similar at the scale of bridges or atoms. a Maxwell asked how many beams are required to make engineering structures rigid. Triangulation is the standard method for ensuring enough central force bonds (rods or beams) are available to support a load, as illustrated by the Cairo Mississippi River bridge from www.bridgehunter.com. b Phillips [59] asked how many higher co-ordination atoms are required to make a random network rigid, for example a structure such as silica from www.phys.uu.nl/~Barkema. Dohler et al. [20] showed that Maxwell counting implies that in bond-bending networks rigidity sets in when the average co-ordination is $r_c = 2.4$

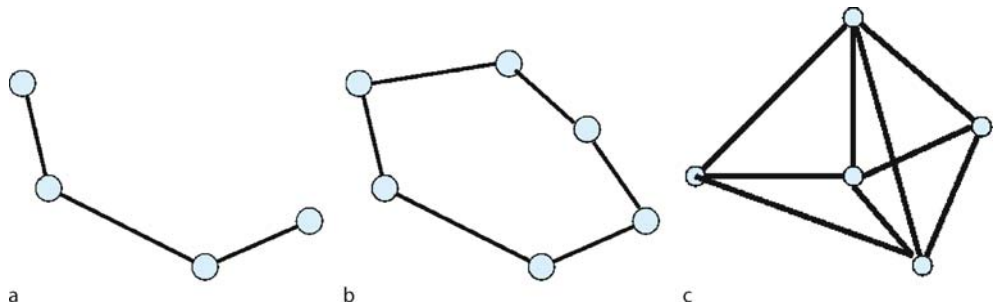

Elastic Percolation Networks, Figure 2
Examples of networks which have just enough connectivity to be rigid. These structures are called isostatic and are stress free. They have a finite elastic constant and no floppy modes. a When central forces, bond angle forces and dihedral terms are all important (all three terms in Eq. (1) or (2)). b When dihedral terms are not important (only the first two terms in Eq. (1) or (2) are important). c When only central forces are important (only the first term in (1) or (2))

a theoretical perspective, replica symmetry breaking is not required in solving the rigidity percolation problem on Bethe lattices, though the system is frustrated [22,64].

There are several different types of elastic percolation, depending on the interaction energy, as illustrated by considering standard force fields, such as the CHARMM [12], AMBER [82] and the Dreiding [49] potentials. They contain three short range terms in their potential energy expressions, with a common form being,

$$E = K_b \sum_{\text{bonds}} (b - b_0)^2 + K_\theta \sum_{\text{angles}} (\theta - \theta_0)^2 + K_\phi \sum_{\text{torsions}} (1 + \cos(n\phi - \phi_0)) \,. \tag{1}$$

The first term in this expression describes the energy required to change the length of a nearest neighbor bond and is a central force term, while the second and third terms provide a restoring force when bond angles or dihedral (torsion) angles are deformed. For the arguments that follow we only need to be able to divide the energy into three terms so that,

$$E = E_{\text{bonds}} + E_{\text{angles}} + E_{\text{torsions}} \,. \tag{2}$$

Even long-range non-bonded interaction terms can be represented using this type of model, by adding further neighbor bonds. The theoretical problem is then: Given a graph or network where the nodes interact with a potential such as Eq. (1), what is the elastic response? Where is the rigidity percolation threshold located as we increase the average connectivity of the network? How many floppy modes are there?

The most interesting geometric aspect of elastic percolation is that the geometry which can support a stress depends on the forces which are important, so that three different cases are illustrated by considering Eq. (1): (i) If all terms in Eq. (1) are important so that, $K_b \sim K_\theta \sim K_\phi$, the underlying geometry is the same as connectivity percolation as any connected network transmits stress (see Fig. 2a); (ii) If only the first two terms are important (i. e. $K_b \sim K_\theta \gg K_\phi$) then the underlying geometry is a special sort of rigidity percolation which can be solved using efficient combinatorial methods (see Fig. 2b). This applies to many Chalcogenide glasses as well as to many polymers and proteins, and is often called the bond-bending case; (iii) The hardest case is when only central force terms are present (i. e. $K_b \gg K_\theta, K_\phi$) in three dimensions (see Fig. 2c). This problem is called the central force rigidity problem and has no rigorous combinatorial characterization, so that we have to resort to direct simulations or approximations. Note that in two dimensions the torsional term is absent and the central force rigidity percolation problem can be solved using combinatorial methods [34,37,50].

It is intuitively evident that as the number of connections in a network increases, the system becomes more rigid or constrained. This intuitive observation is the basis of constraint counting methods introduced by James Clerk Maxwell in 1864 [48], who asked the question: "how many edges are required to make a graph internally rigid, so that it can support an applied stress?" He made the following simple constraint counting argument for three dimensional systems. Consider the case of central forces (only the first term in Eq. (1)) and a set of "n" nodes connected by "b" bonds. Each node has three degrees of freedom, its three translations. If a graph is internally rigid, it still has six degrees of freedom due to its global translations and rotations. Maxwell then stated that the minimal

number of bonds required to make a central force system rigid is given by the relation,

$$b = 3n - 6\,. \tag{3}$$

The example in Fig. 2c has $n = 5$ and $b = 9$ and so it satisfies this "Maxwell counting" condition. Maxwell counting is not exact but it provides a useful mean field model for most rigidity problems. It can also be applied to connectivity percolation, but it is a much poorer approximation in that case.

In 1970, Laman [43] proved a theorem showing that for a restricted class of rigidity problems, namely planar graphs, a Maxwell counting calculation on subgraphs can be used to infer the rigidity of a graph. This revived interest in combinatorial rigidity and several methods for implementing his theorem were developed, with Hendrickson's [34] bond-testing procedure yielding the highly efficient algorithms in use today. Extensions of Laman's theorem to body-bar networks and the conjecture of its extension to molecular frameworks [36,83] have yielded practical algorithms for a wide range of macromolecular systems and other networks. Maxwell's work is also important in engineering where truss networks are analogous to central force systems and have a variety of applications (see e. g. Fig. 1a).

Basic Theoretical Concepts

Maxwell Counting in Random Networks

Maxwell counting [59,78] provides a quick, and frequently quite accurate, estimate of the number of floppy modes and the rigidity threshold. For example, consider bond-diluted triangular lattices where a fraction p of bonds are present. Following Thorpe [78], an unconstrained or flexible degree of freedom is called a floppy mode so that the number of floppy modes, F, remaining in a diluted triangular lattice is,

$$F = 2N - \tfrac{1}{2}pzN + R\,, \tag{4}$$

where N is the number of nodes in the lattice $z = 6$ is the coordination number of the lattice and R is the number of redundant bonds. In this expression $2N$ is the number of degrees of freedom in the absence of any bonds, while $pzN/2$ is the the number of bonds in the network. Maxwell counting sets $R = 0$ so that every bond added to the lattice reduces the number of floppy modes by one. Redundant bonds are not essential for the rigidity of a structure, though they do increase the elastic moduli and they may also induce internal stresses in rigidity percolation problems. A comparison of the simplest examples of redundant bonds occurring in connected and rigid clusters on a triangular lattice is presented in Fig. 3.

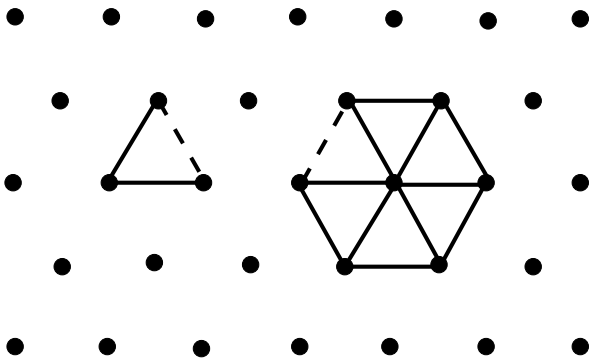

Elastic Percolation Networks, Figure 3
The simplest subgraphs on a triangular lattice which contain a redundant bond *(dashed)*. Connectivity case *(left)*, $g = 2$ rigidity case *(right)*. From article by C. Moukarzel and P. M. Duxbury in [79]

In Maxwell counting we make the mean field approximation that the number of redundant bonds is zero, and that the rigidity transition occurs when the number of floppy modes goes to zero. In that case, Eq. (4) predicts that the rigidity transition for bond diluted triangular lattices occurs at, $p_r = 2/3$. This is surprisingly accurate as large scale simulations using methods which count the number of redundant bonds exactly, determine the threshold $p_r = 0.6602(3)$ [38] which is within 1% of the Maxwell counting estimate. The percolation threshold for bond bending systems is the same as that for connectivity percolation which is known exactly for bond diluted triangular lattices, $p_c = 2\sin(\pi/18) = .347\ldots$. It is then clear that the central-force rigidity threshold occurs at a much higher bond concentration than the connectivity threshold [24]. Note that the Maxwell estimate of the connectivity percolation threshold is found from $F = N - pzN/2$, yielding $p_c = 0.33$, which is a much poorer approximation than for the central force rigidity case. In general Maxwell counting is a good approximation to the percolation threshold for all rigidity problems studied so far, except for cases where the connectivity percolation geometry applies!

Though Maxwell counting is very useful, the fact that it ignores redundant bond makes it incomplete. A Bethe lattice approach [22,53], and field theory methods [56,64] provide more complete theories. These approaches and also simulations on certain two [51] and three dimensional lattices [15], e. g. body-centered-cubic lattices, have demonstrated that the rigidity percolation transition is frequently first order with a large jump in the infinite cluster probability at the rigidity threshold. In fully random

networks, the onset of an infinite rigid cluster and the onset of an internally stressed cluster occur at the same threshold. However, in several important cases, the infinite cluster is unstressed or isostatic which is believed to occur in granular media [55,57] due to repulsive terms and in chalcogenide glasses due to self-organization [9,80]. In these cases, a mixed transition may occur where a first order jump and a continuous singularity occur at the same threshold. This behavior was first observed in Bethe lattice models of rigidity percolation [53].

Elastic Behavior

In all random connectivity and rigidity percolation problems studied so far the elastic behavior is continuous near the rigidity threshold, even in cases where the infinite rigid cluster undergoes a first order jump [15,51,56]. However, the question remains open in models of granular media and in self-organizing glass systems where an isostatic critical point plays a role. Direct numerical studies of the elastic constants of bond diluted central force systems indicate that they are quite linear for $p > p_r$, except very near the rigidity threshold, and that mean field estimates of the elastic behavior are surprisingly accurate in these cases [24,25,70]. In contrast the elastic behavior of bond bending networks in three dimensions is non-linear over most of the concentration range (see Fig. 4). Nevertheless, in all cases, the number of floppy modes as a function of $r < r_c$ is close to linear as is consistent with the simple Maxwell estimate.

In ideal elastic percolation networks, the elastic behavior is singular with a behavior typical of a continuous singularity, $(p - p_r)^T$. The value of the critical exponent T has been a matter of debate. There is general agreement that in cases where bond-bending terms are dominant, the critical exponent T is larger than the conductivity exponent t [41]. For example in two dimensional lattices with bond-bending terms, $t = 1.31(1)$ while $T = 3.96(3)$ [90]. It is also well accepted that continuum systems may exhibit non-universal critical behavior due to the occurrence of necks of varying size in continuum systems [23]. Moreover, even in the absence of continuum effects there are several different elasticity models, with three well studied cases being: gel models where entropic effects are important [17,62] and following de Gennes, $T \approx t$: bonding bending networks where $T \approx t + 2\nu$ [8,24,69,90] and: central force networks where the elastic critical behavior is close to the bond bending case [31,68], at least on regular lattices.

However, in connecting network geometry to elastic behavior in central force networks and in three dimensional networks without torsional forces, it is important to distinguish between generic networks and non-generic networks. This distinction is noticed in granular media where perfectly monodisperse spheres can form regular packings in two dimensions with average contact number six, while random packings, through jamming of regular or polydisperse spheres, have average contact number near four. The latter case is the generic case, while the former is the non-generic case. Non-generic systems are characterized by constraints which are degenerate and occur in the dynamical matrix as dependent equations. In network structures they can occur as special sets of parallel bonds or series combinations of bonds pointing in the same direction (see Sect. "Exact Algorithms and Percolative Geometries"). These configurations do not occur on random networks which are therefore generic. Regular lattices such as the triangular lattice are thus non-generic and the consequences of this on elasticity is still poorly understood.

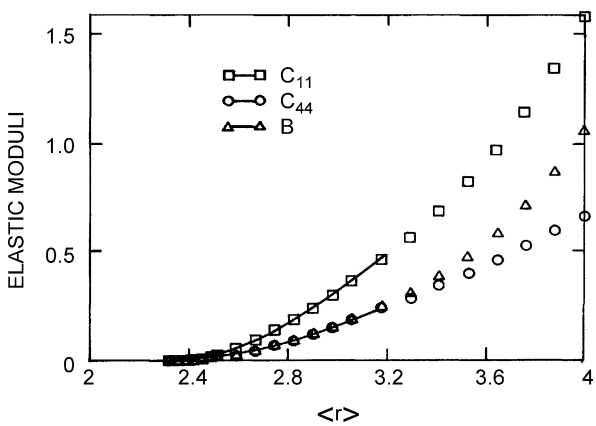

Elastic Percolation Networks, Figure 4
Elastic behavior of glass models as a function of average site coordination using continuous random network models like that of Fig. 1b, from [32]

Idealized Experiments

Idealized experiments have been carried out to test the prediction that, for a given geometric structure, the elastic moduli have different critical exponents than the those which apply to conductivity. A simple tabletop experiment to test this prediction was devised by Benguigui [7]. In his experiment, holes are drilled on a square grid in a metal sheet. The conductivity, σ and Young's modulus, C, were measured as a function of the remaining metal, ϕ. From these experiments the elastic exponent $T \approx 3.5$, and conductivity exponent $t \approx 1.3$ were extracted, which are in agreement with theoretical expectations in two dimen-

sions, within the experimental uncertainty. Conductivity and elasticity measurements on sintered sub micron silver powder aggregates [19] yielded $t = 2.15(25)$ and $T = 3.8(5)$, which agree with theoretical calculations on bond-bending networks in three dimensions which predict $T = t + 2\nu$ [8] and with simulations [69]. To our knowledge there have been no idealized experiments on central force systems, or on systems without torsional forces in three dimensions.

Chalcogenide Glasses

Following Phillips and Thorpe [20,59,78], Maxwell counting for the number of floppy modes in the $Se_{1-x}Ge_x$ system proceeds as follows. Each atom has three degrees of freedom while each r-fold co-ordinated atom imposes $r/2$ central-force constraints and $2r - 3$ bond bending constraints on the network. A two fold Se atom then imposes two constraints while a four fold Ge atom imposes seven constraints. Here we assume that the torsional forces are negligible and the network is continuous with all atoms in one connected cluster. The Maxwell counting estimate for the number of floppy modes is then,

$$F = 3N - 2(1 - x)N - 7xN \tag{5}$$

where N is the number of atoms in the network and $r = 2$ for Se while $r = 4$ for Ge so that the average co-ordination of a glass is $r = 4x + 2(1 - x)$. Equation (5) indicates that the number of floppy modes is linear in x and goes to zero at $x_c = 0.2$, which corresponds to the critical average co-ordination $r_c = 2.4$ [20]. This critical co-ordination also applies when a fraction y of three-fold coordinated atoms are added to the model, as is relevant to the ternary $Se_{1-x-y}As_yGe_x$. In that case $F = 3N - 2(1 - x - y)N - 9yN/2 - 7xN$, and the average site coordination, $r = 2(1 - x - y) + 3y + 4x$. Of course torsional forces are non-zero in these materials and have to be considered in comparisons with experiment.

The prediction of critical coordination $r_c = 2.4$ in chalcogenide glasses stimulated a search for experimental signatures of rigidity percolation. Though early experiments indicated a singular behavior in the elastic constants near r_c, these results were later found to be due to experimental artifacts. Later experiments indicated that the elastic moduli of chalcogenides are smooth near r_c. However the number of floppy modes [40] and an appropriately defined asymptotic glass transition temperature [3] do clearly indicate a rigidity threshold. The number of floppy modes is manifested in inelastic neutron scattering data [40] where a strong low frequency peak, called the Boson peak, exists for $r < r_c$ (see Fig. 5). The number of modes in the Boson peak gives a measure of the number of floppy modes and the location of the peak provides an estimate of the dihedral forces in the material. Theo-

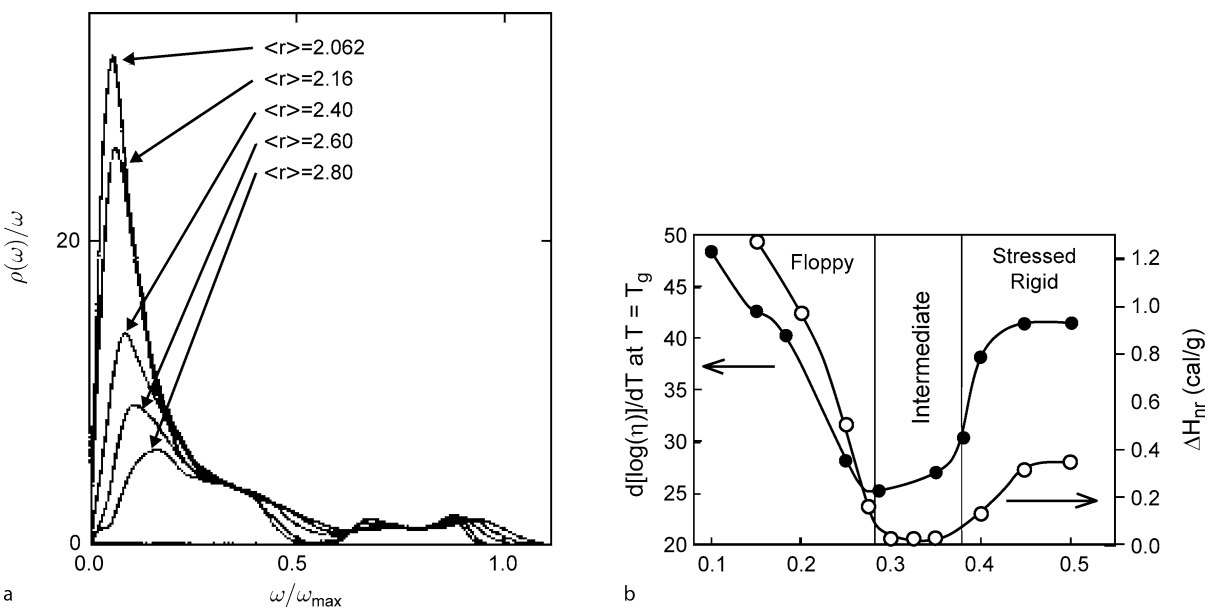

Elastic Percolation Networks, Figure 5

Evidence for rigidity percolation in chalcogenides. a The boson peak near the critical coordination $r_c \approx 2.4$. $\rho(\omega)$ is the density of inelastic neutron scattering modes at frequency ω [40]. b Evidence for an intermediate phase in the chalcogenide glasses $Se_{1-x}As_x$. The intermediate phase observed in the derivative of the viscosity $d\eta/dT$ at T_g and in the non-reversible enthalpy ΔH_{nr}, from [9]

retical analysis of the dynamical response of glassy networks yields good agreement with this data [13]. Several other measurements, including Mössbauer spectra [11], Raman scattering [26], vibrational lifetimes [81], also support a critical value of r_c which is close to 2.4.

A stimulating, though still controversial, concept is the intermediate phase in chalcogenide glasses [9,10,80]. The intermediate phase lies between the traditional floppy and rigid phases and occurs by self-organization to avoid internal stress. There are then two co-ordination thresholds r_1 and r_2, with the first heralding the onset of a rigid but stress-free network and the second the onset of a rigid but internally-stressed network [80]. The extent of the intermediate regime between r_1 and r_2 has been explored in a variety of models [80] and experimental support has come from studies of chalcogenide glasses over narrow composition regimes [9,10], as illustrated in Fig. 5b.

Constraint counting and rigidity concepts developed in chalcogenide glasses have been applied to a wide variety of covalently bonded materials, ranging from amorphous carbons to complex ternaries. Thermal effects along with local structural and chemical ordering often confound a simple interpretation of the data as a purely random rigidity percolation process, nevertheless rigidity and elastic percolation concepts are fundamental to much of the literature in this area [2,3,79,85].

Gels and Semiflexible Rod Networks

Soon after the development of percolation theory, gelation was recognized as a related process and the geometry of gels has been compared to percolation in many different systems [6,42,47,66,73,84,87,89]. Though percolation provides a useful framework for the description of the onset of rigidity in gels, it is often difficult to access the critical regime [65].

Gels are ubiquitous in science, nature and technology though there is no consensus definition of what constitutes a gel. Nevertheless we are all familiar with gels, ranging from jelly to clay dispersions and to chemically crosslinked systems such as rubber and epoxies. Gelation may occur through physical or chemical crosslinking, or through a combination of both processes. Crosslinking is a short range concept and long range forces can sometimes be critical. Nevertheless a broad range of physical and chemical gels can be understood using rigidity concepts based on the change in structure as the crosslinking in a network increases. Even at the single molecule level it is sometimes reasonable to use rigidity percolation ideas to evaluate flexibility as a function of internal crosslinking, for example in proteins [63]. Colloidal gels with strong repulsive terms and weak short-range attractive interactions have many features in common with granular media, as described in the next section.

The original work on gels related the gelation point to connectivity percolation [73], while recent work relates the gel point in rod networks to rigidity percolation [33,35,44,45], as illustrated in Fig. 6. Maxwell counting in two dimensional rod networks [44] states that the number of floppy modes in a network with N rods connected by P pivots is given by,

$$F = 3N - 2P - 3 \,. \tag{6}$$

There are three degrees of freedom per rod, two translations and one rotation, while each pivot removes two translational degrees of freedom from the two rods which it connects. The Maxwell counting estimate, found by set-

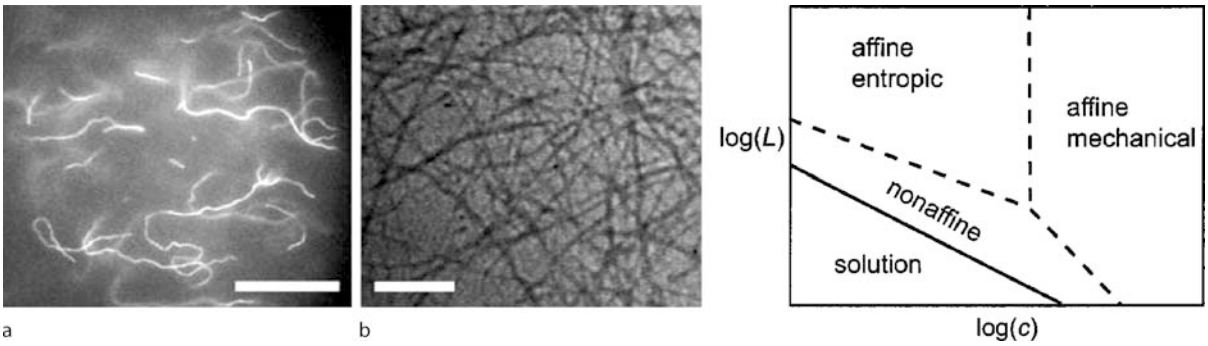

Elastic Percolation Networks, Figure 6

Elastic percolation of biofibers. a, b Two views of the actin filament network in a human cell, from [77]. The right panel is a schematic phase diagram of semiflexible rod networks, from [33]. The solid line is associated with a rigidity threshold in a random rod network. The mean distance between crosslinks or entanglements is l_c, and $c \approx 1/l_c$. L is the molecular weight of the semiflexible polymers and the rigidity threshold occurs at $L \approx 1/c$ which defines the sol-gel transition in this model

ting $F = 0$ in Eq. (6), for the onset of rigidity in these two dimensional rod networks is then $P_r/N = 3/2$, so that there are three pivot points on each rod. Connectivity percolation occurs at $P_c/N = 1$. Elastic models have been simulated and good agreement with Eq. (6) was found. However experimental gels exhibit correlations in local structure [87], and in the case of fibers orientational ordering and bundling [6,42]. Moreover the elastic behavior of crosslinks is difficult to quantify. These factors are particularly important for non-universal parameters such as the percolation or gelation threshold, and they strongly affect the non-linear response of gels.

The elastic behavior near gelation is expected to be more universal provided long-range correlations in structure are absent. In a well known paper, de Gennes predicted that the elastic critical exponent, T, of gels should be the same as the conductivity exponent t [17]. However, as discussed in Sect. "Basic Theoretical Concepts" and elaborated upon in Sect. "Elastic Critical Behavior", purely geometric models predict that the elasticity exponent is significantly larger than $t_{3D} = 2.01(1)$, and is well described by $T_{3D} = t_{3D} + 2\nu_{3D} = 3.77(3)$, where $\nu_{3D} = 0.88$ is the percolation correlation length exponent [8,24,41,69,90]. Over the years many experiments have measured T at the gelation threshold, for example [4,29,47,65,84,87], with the consensus that $T = t$ in most gels as predicted by de Gennes. This issue will be discussed further in Sect. "Elastic Critical Behavior".

Granular Media

Jamming occurs as the packing fraction or density of a colloidal or granular system increases. Though the relations between rigidity and jamming have been discussed for many years [30], it is only recently that the relationship has yielded to a clear quantitative analysis [2,55,57,85]. Alexander [2] discussed the presence of floppy modes or rattlers in granular systems and recent work has noted their relation to the Boson peak in glasses [86].

The presence of complex stress bearing networks has been imaged [30] and studied analytically and computationally and the relation to an isostatic network was noted by Moukarzel [55] who developed the idea of jamming as an isostatic critical point. Molecular dynamics simulations in systems with purely repulsive potentials provide a beautiful synthesis of the relations between rigidity, colloidal glasses and granular media [57] (see Fig. 7). These simulations demonstrate that the zero temperature jamming

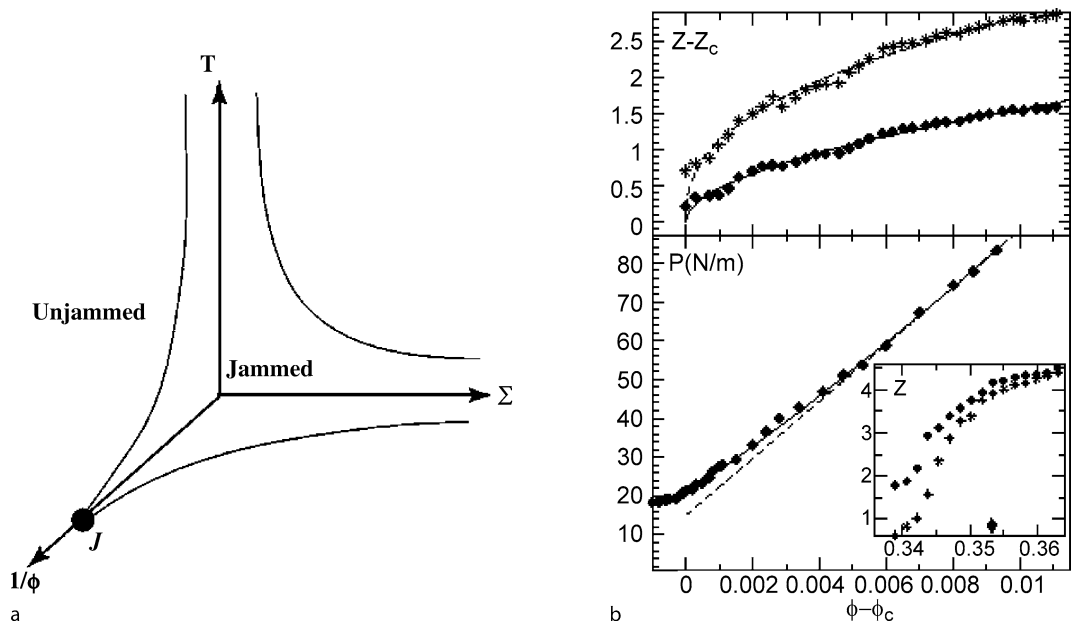

Elastic Percolation Networks, Figure 7

Behavior of a theoretical and b experimental dense packings near jamming. a schematic of phase behavior found from simulations of purely repulsive models, from [57]. The jamming point J is an isostatic critical point. Σ is the applied shear stress, T is temperature and ϕ is packing fraction. b Behavior of the pressure, P and average co-ordination number $Z - Z_c$ on approach to jamming of photoelastic spheres. $\phi - \phi_c$ is the deviation from close packing, from [46]. In the *top panel* and in the *inset*, the diamonds exclude rattlers while the stars include them

point provides insight into the whole phase diagram and enables calculation of the way in which various properties approach zero on approach to jamming from above. These simulations avoid the non-generic critical point which is characteristic of the crystal phase of monodisperse packings, and instead focused upon non-crystalline random packing states which are characteristic of generic rigidity. The computed critical exponents for the shear modulus depend on the form of the repulsive potential but not on the spatial dimension. Later simulations suggest that the stress bearing backbone has a mixed behavior on approach to the jamming point, and its behavior is argued to be essentially the same as that observed in the Bethe lattice theory near an isostatic critical point [71]. Experiments on photoelastic spheres [46] are in general supportive of the physical picture emerging from these theoretical insights (see Fig. 7), including the idea of a jump in average coordination at the jamming point.

Exact Solution on Bethe Lattices

Maxwell counting provides a very useful mean field theory for the rigidity transition, nevertheless, it is not a complete theory as it ignores redundant bonds and does not provide a geometry for the rigidity percolation process. Moreover Maxwell counting predicts that the onset of rigidity occurs when all floppy modes are removed so that the whole network is in one giant isostatic cluster.

Bethe lattice theory is a more complete theory which resolves most of the problems with Maxwell counting methods, moreover it is simple and exactly solvable. In the following we describe the simplest Bethe lattice theory to illustrate the way in which the classic theory of connectivity percolation compares to rigidity percolation [27] and how Maxwell counting should be modified in light of the Bethe analysis. The extent of the intermediate phase [80] can also be simply calculated from the results of the Bethe lattice rigidity theory, as outlined at the end of this section.

The rigidity problem on Bethe lattices encompasses a wide range of different models [53], though here we focus on one generic case [22]. In d dimensions a point object has d degrees of freedom (d translations), while extended objects or bodies also have rotational modes and a total of $d(d+1)/2$ degrees of freedom. We define g as the number of degrees of freedom of a free or unbonded site, so that $g = 1$ corresponds to connectivity percolation, $g = d$ to point objects and $g = d(d+1)/2$ to bodies. A network with N sites (and no edges) has a total of $F = Ng$ degrees of freedom, or floppy (zero frequency) modes. Constraint counting notes that each time an *independent* edge is added to the network, the number floppy modes is reduced by one, so that

$$F = Ng - E + R \tag{7}$$

where E is the number edges in the graph and R is the number of redundant edges. Note the additional term R on the right hand side. This term is key in understanding the relation between constraint counting and percolation, and in finding algorithms for rigidity percolation. An edge does not reduce the number of floppy modes if it is placed between two sites which are already mutually rigid, in which case this edge is redundant. The simplest examples of subgraphs containing a redundant bond on a triangular lattice are illustrated in Fig. 3 for the connectivity ($g = 1$) and $g = 2$ rigidity cases. Note that any one of the bonds in these structures could be labeled as the redundant one. However once any one of them is removed, all of the others are necessary to ensure the mutual rigidity of the structure. The set of all bonds which are mutually redundant form an overconstrained or stressed cluster. In rigidity theory each edge can be considered to be a central force spring, which means that there is a restoring force only in tension and compression. Then an overconstrained cluster of such springs (with random natural lengths) is internally stressed due to a redundant bond. In the connectivity case each bond is like a wire which can carry current or fluid flow. The simplest overconstrained cluster is then a loop which can support an internal eddy current. Rigid structures which contain no redundant bonds are minimally rigid or isostatic. In connectivity percolation isostatic structures are trees, whereas in ($g > 1$) rigidity percolation isostatic structures always contain many loops.

In percolation problems, we are interested in the asymptotic limit of very large graphs ($N \to \infty$), and it is more convenient to work with intensive quantities, so we define $f(p) = F/gN$ and $r(p) = 2R/gzN$, so that

$$f(p) = 1 - \frac{z}{2g}(p - r(p)) \tag{8}$$

where, on average, the number of edges is $E/N = zp/2$. $g * r(p)$ is the probability that a bond is redundant [22,38]. The number of floppy modes, $f(p)$, acts as a free energy for both connectivity and rigidity problems, so that if $\partial f(p)/\partial p$ undergoes a jump discontinuity, the transition is first order [22]. The behavior of this quantity is directly related to the probability that a bond is overconstrained P_{ov} via the relation,

$$\frac{\partial f}{\partial p} = -\frac{z}{2g}(1 - P_{\text{ov}}) \,. \tag{9}$$

If the transition is second order, the second derivative $\partial^2 f/\partial p^2 \sim (p - p_c)^{-\alpha}$, where α is the specific heat exponent [38].

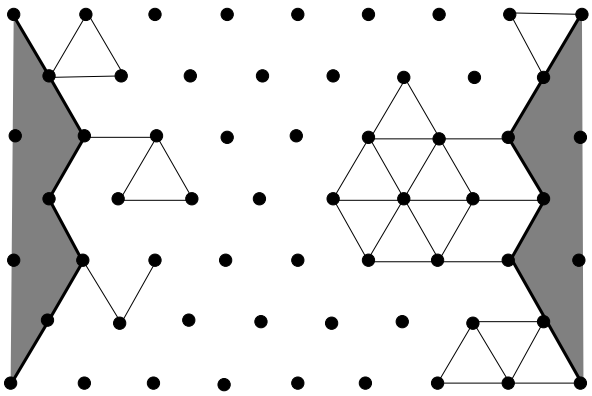

Elastic Percolation Networks, Figure 8
Dangling ends connected to a backbone (*shaded*). Connectivity case (*left*) and $g = 2$ rigidity case (*right*). Dangling ends don't contribute to current (*left*) or stress (*right*) which is applied in the vertical direction

The infinite cluster probability in rigidity problems is composed of the stress-bearing backbone plus dangling ends. Dangling ends are rigidly connected to the backbone but are not part of the stress bearing backbone. Examples of dangling ends in the connectivity and rigidity cases are illustrated in Fig. 8 for central force problems on the triangular lattice.

The quantities, $f(p), \partial f(p)/\partial p, r(p), P_\infty, P_{\rm ov}$ can be calculated exactly on Bethe lattices of general coordination z, where p is the bond probability of the Bethe lattice. The key quantity from which all the others is derived is the probability, T, that a site on a branch of a Bethe lattice is part of the infinite rigid cluster. This quantity has a steady state solution which is found from the equation,

$$T = \sum_{l=g}^{z-1} \binom{z-1}{l} (pT)^l (1-pT)^{z-1-l} \,, \tag{10}$$

where T is the probability that a site is connected to the infinite rigid cluster through one branch of the Bethe lattice. The case $g = 1$ recovers the well known Bethe lattice equation for connectivity percolation [27] while for general g, these equations are the same as those for *k-core* percolation on Bethe lattices [14], where $k = g + 1$. As illustrated in Fig. 9, the solutions to Eq. (10) depend sensitively on the value of g. The case $g = 1$ corresponds to connectivity percolation and shows a typical continuous behavior, while the case $g = 2$ is first order and is typical of all cases where $g > 1$. This means that central force percolation on random graphs is typically first order [22,54].

It is clearly seen from these figures that the rigidity transition is first order on Bethe lattices, while the connectivity transition is second order. We identify the point at which the stable solution becomes nonzero as $p_{\rm s}$, the spinodal point, and in the connectivity case $p_{\rm s} = p_{\rm c}$ because the transition is second order.

Combining the stable solutions to branch probabilities, yields the infinite cluster probability of the Bethe lattice,

$$P_\infty = \sum_{l=g}^{z} \binom{z}{l} (pT)^l (1-pT)^{z-l} \,. \tag{11}$$

The other quantities of interest are found from the stable solution for T as follows. The probability that a bond is overconstrained is [22],

$$P_{\rm ov} = T^2 \,, \tag{12}$$

so that Eq. (9) yields,

$$\frac{\partial f}{\partial p} = -\frac{z}{2g}(1-T^2) \,. \tag{13}$$

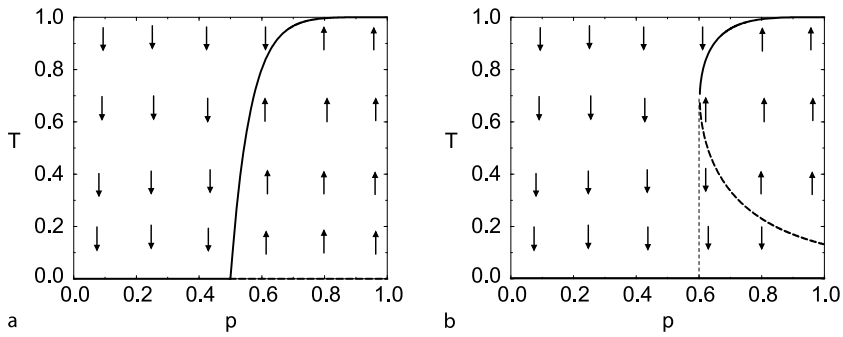

Elastic Percolation Networks, Figure 9
Domains of attraction of the mean field equations, Eq. (10). a Typical connectivity percolation behavior (this example is $z = 3$); b a typical rigidity percolation behavior (this example is $z = 6$, $g = 2$). Dark lines are stable (attractive) solutions and the dashed line is unstable, from article by C. Moukarzel and P. Duxbury in [79]

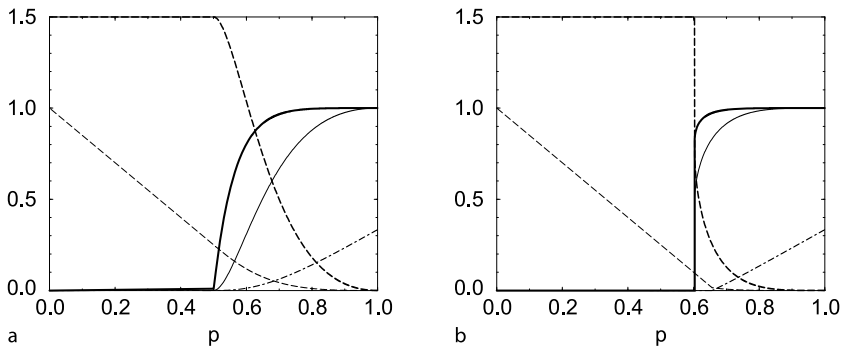

Elastic Percolation Networks, Figure 10

The order parameters, floppy modes and its derivative as a function of *p* for a a typical connectivity percolation case $z = 3, g = 1$, and b a typical rigidity percolation case, $z = 6, g = 2$. The quantities plotted are $f(p)$ (*dashed*), $r(p)$ (*dot dashed*), $-df/dp$ (*heavy dashed*), P_{ov} (*thin solid*), P_{inf} (*heavy solid*). In the connectivity case we find $p_c = p_s$, while in rigidity cases $p_s < p_c$. For the rigidity case, $p_s = 0.605$ and $p_c = 0.655$, from article by C. Moukarzel and P. M. Duxbury in [79]

The probability that a bond is on the infinite cluster is,

$$P_{\text{inf}} = T^2 + 2TT_1 \,. \tag{14}$$

where T_1 is the probability that a site has one degree of freedom and can also be found straightforwardly [22]. We also want to find the total number of redundant bonds $r(p)$ and the total number of floppy modes $f(p)$. In order to find these quantities, we integrate Eq. (13) and then use Eq. (7). However, the integration of Eq. (13) leads to one free constant. To determine this constant, for a lattice of co-ordination z, we impose the constraint [22],

$$r(1) = 1 - \frac{2g}{z} \,. \tag{15}$$

We then find that r(p) approaches zero at a critical point p_c, which lies above p_s, for any $g > 1$, and that p_c is close to the Maxwell estimate $p_c = 2g/z$. Results for five key quantities found from Bethe lattice theory are presented in Fig. 10 for typical conductivity and rigidity percolation cases.

The Bethe lattice theory outlined above has been carefully compared with exact simulations of random graphs which demonstrate that it is exact [22]. This is a symmetric theory as the order parameter does not need to be described by a non-trivial distribution. The fact that this symmetric theory is exact [64] and yet applies to glassy systems such as granular media and rod networks is intriguing as replica symmetry breaking would be expected in those cases, as occurs in problems such as KSAT, spin glasses, and lattice gases. It is worth noting that the Bethe lattice Eqs. (10)–(11) are the same as the Bethe lattice equations for k-core percolation [14] which is also being used as a simple model for granular media [71]. Extensions of the Bethe lattice theory to chalcogenide glasses have been carried out and provide a more complete theory than Maxwell counting [79].

Extension to the Intermediate Phase

In the intermediate phase [80], the network is dominated by a giant rigid but isostatic cluster. To generate such a cluster, we consider adding bonds to a graph, with the constraint that any redundant bond is not added to the network. As described in the next section this is how the exact constraint counting algorithms proceed, so computation of geometric structures which are stress free is straightforward using Hendrickson's bond testing procedure [34].

In the Bethe lattice solution, the onset of a metastable solution occurs at p_s, however this solution contains a number, $r(p_s)$, of redundant bonds. Since bonds are added randomly, the lowest threshold at which a stress-free percolating cluster can occur is $p_1 = p_s - r(p_s)$. In first order rigidity cases $r(p_s)$ is small, so to a good approximation $p_1 \approx p_s$. The upper limit of the intermediate phase is an isostatic network to which no more isostatic bonds can be added. This is just the Maxwell counting estimate, so $p_2 \approx p_m$. Calculation of p_1 and p_2 for other graphs and lattices can be carried out in a similar way, using the exact algorithms outlined in the next subsection. In second order cases however, $p_s = p_c$, so that $p_1 = p_c - r(p_c)$ and $p_2 = p_m$. In glasses, it is unlikely that all stress inducing bonds can be avoided so we expect the observed values of the intermediate regime to begin at $p > p_1$ and to end at $p < p_2$.

Exact Algorithms and Percolative Geometries

Exact constraint counting provides a procedure for figuring out if a structure can carry stress, without having to solve the elastic equations for the structure. The procedure is based on determining whether all internal degrees of freedom in the structure are fixed by the constraints. A small activity in this area has been ongoing for many years in the graph theory and topology branches of mathematics. The problem is call the graph rigidity problem and is a subset of the problem of graphic matroids. Laman's theorem in 1970 [43] provides the basis for practical algorithms and states that: A bar-joint graph in the plane is rigid *iff* it has no redundant bonds and $b = 2n - 3$, where b is the number of edges (bars) in the graph and n is the number of nodes. The is basically Maxwell counting (see Eq. (3)) in two dimensions, modified by the key requirement that there be no redundant bonds. Hendrickson [34] provided a practical algorithm which tests if a bond in a graph is redundant. In his algorithm, edges are added to the graph one at a time, and tested to see if they are redundant. If they are redundant, then they are not placed in the structure but are noted and stored elsewhere. In this way an isostatic graph obeying Maxwell counting is generated. A structure with no redundant bonds has the Maxwell counting value, $F = 2n - b - 3$, of floppy modes. The number of redundant bonds are also counted enabling a check of the relation Eq. (7). It is also possible to use this method to find the number of stressed bonds, as when a redundant bonds is added to a network, an internally stressed cluster of bonds is generated. By cleverly adding a redundant bond across the whole network, this can also be used to identify the stressed backbone [37,50].

Laman's theorem has been extended to body-bar systems in arbitrary dimensions [75,76]. There is also a conjecture that it is exact for molecular frameworks which is important in applications to chalcogenide glasses, polymers and proteins [36,83]. Molecular frameworks are systems where central forces and bond-bending terms are strong but torsional terms are neglected. However a key system where no combinatorial characterization is available is the central force rigidity problem in three dimensions where exceptions to Laman's theorem are known to exist, such as the famous double banana configuration of Fig. 11a.

A second issue, of significance to a variety of applications including granular media, is the issue of generic rigidity as compared to non-generic rigidity. The issue is illustrated in Fig. 11 where two examples of systems with two bodies and three bars in two dimensions are compared. In the generic rigidity case of Fig. 11c, the three bars have different lengths and are at different angles. Maxwell counting indicates that this system is rigid as there are three degrees of freedom per body and there are three bars. However the non-generic example in Fig. 11b is not rigid as the three bars are parallel enabling a shear distortion of the two bodies with respect to each other. The three bars, which are the constraints in this system, are not independent and this dependence is due to the particular way in which the bars are arranged. In this non-generic case, it is not sufficient to know only the connectivity of the graph as the particular geometry of the structure plays a key role. Clearly regular lattices and regular packings of granular media are non-generic while random lattices, polydisperse packings and random packings are typically generic. Combinatorial rigidity applies to generic cases so that disorder in some sense makes the generic problem more tractable.

To illustrate Hendrickson's redundant bond testing procedure, we consider percolation on diluted triangular lattices. The procedure applies to both rigidity and connectivity cases and it is useful to compare the way in which bond testing compares in the two cases. In the connectivity

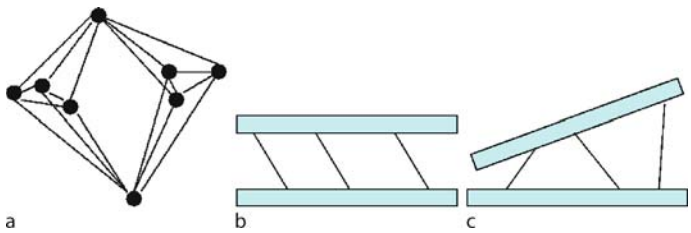

Elastic Percolation Networks, Figure 11

Illustration of two key issues in combinatorial rigidity. a The double banana configuration which violates Laman's theorem and illustrates a difficulty in applying combinatorial rigidity to three dimensional central force networks. b Non-generic and c generic geometric structures. Laman's theorem and its extensions apply only to generic networks. The non-generic case depends on geometry and so a theory depending only on connectivity is insufficient. The probability of finding non-generic configurations in large random structures is negligible so generic rigidity applies

case Hendrickson's algorithm checks for loops. His redundant bond testing procedure is based on bipartite matching which is a well known problem in combinatorial optimization. It is implemented as follows [34,37,50]:

Start with an empty triangular lattice (no bonds) and assign to each node g degrees of freedom.
Then:

1. Randomly add a bond to the lattice.
2. Test whether this bond is redundant with respect to the bonds which are currently in the lattice.
3. If the bond is redundant *do not add it to the lattice*, but instead store its location in a different array.
4. Return to 1.

End

Step 2 is the key one and is implemented by matching constraints (bonds) to degrees of freedom, with the restriction that the number of constraints assigned to each node must be less than or equal to the number of degrees of freedom at a node ($g = 2$). It is natural to represent the assignment of constraints to nodes using arrows as illustrated in Fig. 12. When a new bond is added, we add $G + 1$ bonds to take account of the global translations and rotations of a body. For a triangular lattice a body has three degrees of freedom, so we add four bonds as indicated in the figure. Hendrickson noted that if the $G + 1$ added bonds can be matched to the degrees of freedom in the graph, then the bond is not redundant. However, if the arrows cannot be matched, added edge is *redundant*. Another useful way to think about the matching procedure is to consider associating pebbles with the degrees of freedom of the nodes in a graph. Then these pebbles may be placed on the edges of the graph with the constraint that they may only cover edges which enter the node. The matching of arrows to nodes then corresponds to placing pebbles on edges, with the constraint that pebbles can only sit on adjacent edges. The matching fails if it is not possible to cover all edges with pebbles [38]. A successful and a failed match are illustrated in Fig. 12 for a connectivity ($g = 1$) case and a rigidity ($g = 2$) case. Note that the bond that is being tested carries with it G additional copies which account for global degrees of freedom of a rigid cluster. In the connectivity case $G = 1$, while on central-force bar-joint networks in two dimensions $G = 3$.

When the bond test fails, the algorithm identifies all bonds which are overconstrained or stressed with respect to the redundant bond. This set of bonds is called a *Laman subgraph*. Note that if a redundant bond is already in a graph, it is not possible to add a new bond and test its redundancy with this method. This is the reason that the algorithm proceeds by adding bonds one at a time starting with an empty lattice. Any error in testing the redundancy of a bond invalidates the rest of the addition sequence. However since this algorithm is an integer method there is no problem with roundoff. It is easy to see that the matching algorithm is quite efficient, however it requires quite a bit of effort to fully optimize these methods. A key step in this optimization procedure is to identify rigid clusters and to describe them in a more condensed way. Moukarzel [50] condenses rigid clusters to a body and hence renormalizes an original bar-joint network to an effective body-joint-bar network. Jacobs [37] instead uses a reduced bar-joint representation of rigid subgraphs. In either case the computational complexity is reduced significantly especially for $p > p_r$.

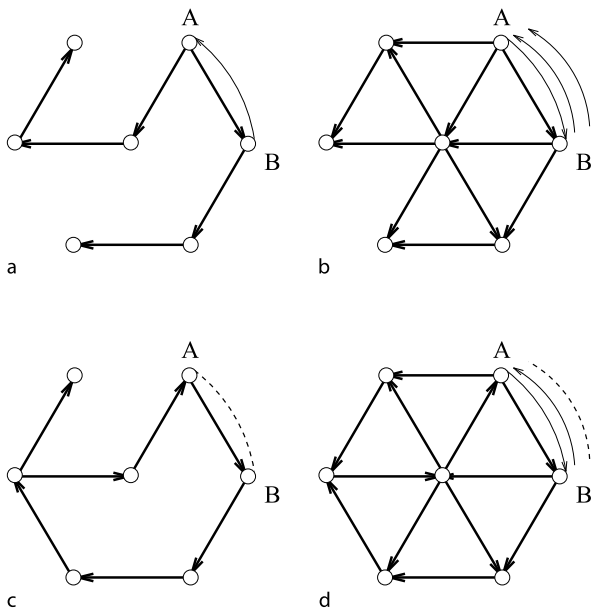

Elastic Percolation Networks, Figure 12
Examples of successful (*top* figures a and b) and failed (*bottom* figures, c and d) matches in the connectivity ($g = G = 1$, *left* figures, a and c) and joint-bar rigidity ($g = 2$, $G = 3$, *right* figures, b and d) cases on triangular lattices. AB is the new bond that is being tested. Each site has g degrees of freedom and therefore accepts at most g incoming arrows. Each new bond carries with it G auxiliary arrows, which must also be matched. If any of the arrows cannot be matched (*dashed*), the new bond is redundant. From article by C. Moukarzel and P. M. Duxbury in [79]

The geometry of rigidity percolation on a triangular lattice is presented in Fig. 13 [51]. Because of the fact that we add bonds one at a time until the percolation point is reached, it is possible to identify p_c or p_r *exactly* for

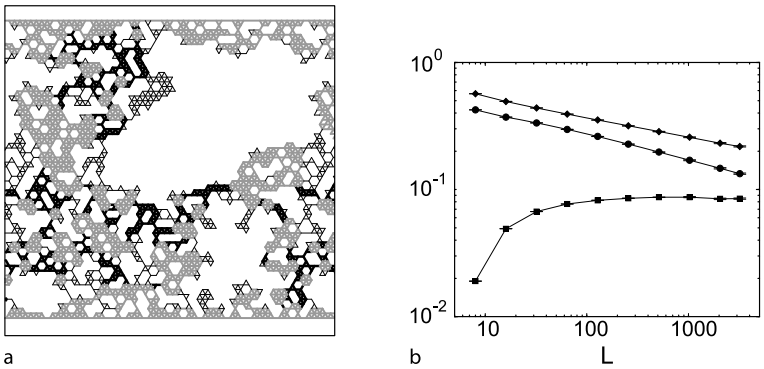

Elastic Percolation Networks, Figure 13

The geometry of rigidity percolation on site diluted triangular lattices. a The infinite-cluster geometry at threshold. *Dark wide lines* are cutting bonds, *wide lines* are non-critical backbone bonds (blobs) and *thin lines* are dangling ends. b The density of backbone bonds (*circles*), infinite cluster bonds (*diamonds*) and dangling bonds (*squares*) at p_r. A fit to the backbone density yields the exponent $P_B \approx L^{-\beta'/\nu}$, with $\beta' = 0.25(2)$. The dangling ends have not reached the asymptotic regime so it is not clear if the infinite cluster probability has either. Nevertheless, the transition is second order with correlation length exponent $\nu = 1.16(2)$ found from the finite size scaling of the number of cutting bonds and also from the scaling of finite size corrections to p_c [39,51,52], from [51]

each sample, and therefore measure the components of the spanning cluster exactly at p_c or p_r. This eliminates the error associated with measurements at fixed values of p, since estimated exponents are known to depend very sensitively on p. Just as in connectivity percolation, at p_r we identify three different types of bonds: *backbone bonds*, *dangling ends* and *cutting bonds*, as illustrated in Fig. 13a. These together form the *infinite cluster*. The cutting bonds are stressed (belong to the backbone), but they are "critical" because if one of them is removed, load is no longer transmitted across the infinite cluster.

In order to find the correlation length exponent, two relations were used [38,39,51,52]: the size dependence of the threshold behaves as $\delta p_c \sim L^{-1/\nu}$ and secondly, the number of cutting bonds varies as $n_c \sim L^{1/\nu}$, yielding $\nu = 1.16(2)$ [52] or $\nu = 1.20(3)$ [38]. Calculations on the rigidity of rod networks [44] find $\nu = 1.18(2)$. From data such as Fig. 13b the backbone density is found to decrease algebraically $P_B \sim L^{-\beta'/\nu}$, with $\beta' = 0.25 \pm 0.02$ [51,52,54]. It also appears that the infinite-cluster probability is decreasing algebraically, however that is difficult to reconcile with the behavior of the dangling ends which must also eventually decrease for this trend to be asymptotic.

Since it has a diverging correlation length, rigidity percolation on triangular lattices is second order, however first order cases do occur in two dimensions, for example a square network to which diagonals are added at random locations exhibits a strong first order transition, even though the elastic behavior remains continuous [51,56]. Moreover, simulations of central force networks in three dimensions [15] demonstrate a strong first order transition, as illustrated in Fig. 14.

Elastic Critical Behavior

After a flurry of activity in the 1980's and early 1990's study of elastic critical exponents has diminished, nevertheless many issues remain unresolved. The distinction between bond-bending networks and central force networks was brought into sharp focus in the early 1980's [24], moreover the distinction between the critical exponents describing conductivity and elasticity was emphasized [41]. Many studies were purely energetic, so that thermal effects were ignored. In contrast, de Gennes [17] early argument concerning the equivalence of the conductivity and elasticity exponents of gels relies on entropic springs. The effect of entropic terms has been simulated more recently, supporting the early analysis by de Gennes [60,61].

In two dimensional systems with bond-bending terms or in three dimensional systems with both bond-bending and torsion or dihedral terms (see Eq. (1)), elastic percolation occurs at the conductivity percolation threshold. In the absence of entropic terms, it is well established that in this case the elastic critical behavior, $E \sim (\delta p)^T$ is different than the conductivity critical behavior, $\sigma \sim (\delta p)^t$. E is the Young's modulus while σ is the conductivity, and $\delta p = p - p_c$ with p_c the conductivity percolation threshold. T is the elastic critical exponent and t is the conductivity critical exponent. A variety of methods have been applied to the calculation of t and T in two and three dimensions, with the currently

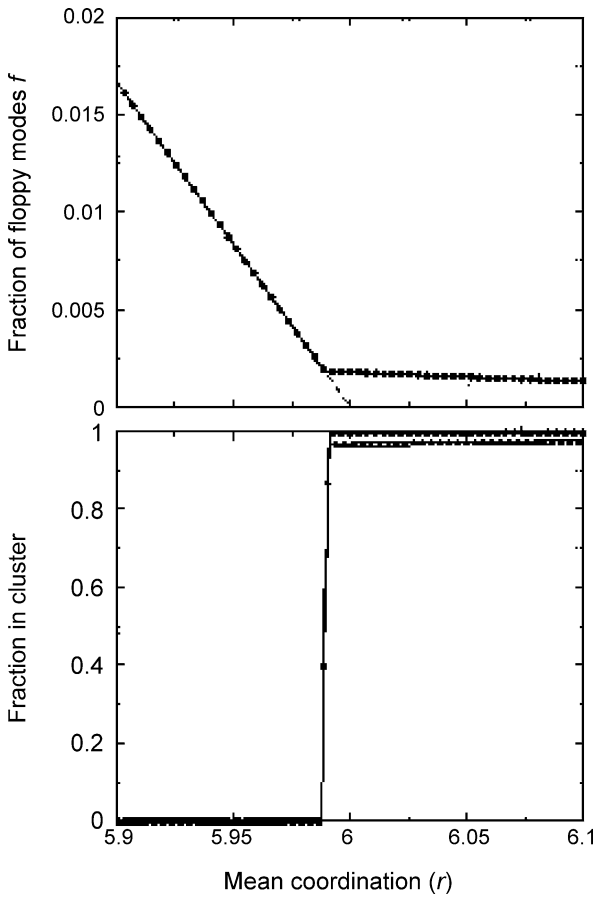

Elastic Percolation Networks, Figure 14
The number of floppy modes (*top panel*). The infinite-cluster probability and stressed backbone probability (*bottom panel*) on bond diluted bcc lattices. The transition is strongly first order, though the elastic behavior appears to be continuous, from [15]

accepted values being $t_{2D} = 1.30(1)$, $t_{3D} = 2.01(2)$ and $T_{2D} = 3.96(4)$ [90]. In three dimensional bond-bending networks $T_{3D} \approx 3.75$ [69], and both of these elastic critical exponents are consistent with the relation $T = t + 2\nu$, which is claimed to be exact [8]. Here ν is the percolation critical exponent and $\nu_{2D} = 4/3$ and $\nu_{3D} = 0.88(1)$. It is worth noting that a model where hard inclusions are placed in a soft elastic background has a different behavior. The critical point is the point at which an infinite hard cluster percolates. On approach to this point from below, the elastic constant diverges $E \sim |\Delta p|^{-S}$. The analogous electrical problem consists of perfectly conducting inclusions in a finite conductivity matrix where $\sigma \sim |\Delta p|^{-s}$. In this case Bergman [8] argues that $s = S$. The lattice results quoted above require modification in the case of continuum systems, for example in cases where holes are punched randomly in a sheet so that narrow necks of material can exist [23]. This case is non-universal in both the conductivity and elasticity cases, though the geometric exponents, such as ν, remain universal.

Elastic percolation on central force networks occurs at different thresholds than conductivity percolation on the same networks. Maxwell counting gives a good first approximation to the elastic percolation thresholds, however precise values can only be found by direct simulation. Direct simulations of central force networks in two and three dimensions have been quite controversial, however the consensus seems to be that the behavior in two dimensions close to criticality is in the bond-bending universality class described in the paragraph above [31]. However central force systems in three dimensions [68] exhibit a different behavior, with $T/\nu \approx 2.1$ for body centered cubic lattices, which is much smaller than the bond bending value $T/\nu \approx 4.4$. Moreover, the geometry of these systems is now known to exhibit a strong first order jump (see Fig. 14) in three dimensions [15].

The elastic behavior of bond-diluted central force systems is remarkably simple away from p_r and is well described by a simple linear relation found using effective medium theory [25]. The critical regime appears to be quite narrow in these central force models. It is also important to note that the results for p_r found from combinatorial methods apply to triangular lattices where the sites are randomly displaced from their regular locations to remove degeneracies. This may be the reason that geometric calculations of p_r on triangular lattices are higher 0.6603 [38] than values found from direct solution of elastic response on regular lattices where $p_r \sim 0.64$ [16]. A further complication is that elastic calculations on displaced lattices lead to a non-linear stiffening of the lattice as bonds which are nearly co-linear are brought into alignment by the applied stress and lead to a stiffening of the network [52].

In geometric models of rigidity, as described above and in Sect. "Exact Solution on Bethe Lattices", the rigidity percolation threshold, p_r, lies above the connectivity percolation threshold, p_c. However in real materials there is usually a finite elastic modulus even in the regime $p_c < p < p_r$, which may arise through a variety of different effects. In polymers, proteins and chalcogenide glasses, the torsional forces are smaller but significant leading to the onset of elasticity at p_c rather than p_r. Moreover entropic effects may be significant even in ideal central force systems, so that thermal fluctuations lead to a finite elastic modulus for all $p > p_c$. This has been demonstrated in simple central force systems such as diluted square, triangular and cubic lattices [62]. This effect is even more pronounced in crosslinked polymeric systems, such as rubber,

where entropic rubber elasticity dominates [5,17,88]. In these systems the regime $p_c < p < p_r$ is extremely broad so that even heavily crosslinked flexible polymeric systems are in this regime. A further mechanism which leads to a reduction of the percolation threshold from that predicted by geometric rigidity is the presence of tension in the network [74], or if the natural length of the springs in the network is taken to zero. Moreover the elastic critical exponent observed in entropic systems and in systems in tension is usually quite close to the conductivity value, as first predicted by de Gennes in 1976 [5,17,88] and as observed in many experiments on gels (see Sect. "Gels and Semiflexible Rod Networks").

Final Remarks and Future Directions

The relevance of rigidity percolation to granular media and to semiflexible polymer networks is now well established, though many issues remain unresolved. Ideal experiments to test the recent theoretical predictions will greatly clarify the extent to which current rigidity concepts are sufficient to describe these systems. The intermediate phase in chalcogenide glasses is a stimulating, and physically very natural, concept, that perhaps could be observed in large scale molecular dynamics simulations with realistic potentials of systems such as $Se_{1-x}Ge_x$. Presumably there is a dynamical lengthscale at which self-organization is possible, which limits the extent of the intermediate phase.

Graph combinatorial methods and direct elastic model simulations provide complementary information about the mechanical response of complex materials. The combinatorial methods however remain limited due to the absence of an exact combinatorial characterization of rigid structures in three dimensional central force networks. Moreover, the physical significance of banana configurations, as illustrated in Fig. 11a, remains unexplored and is an intriguing outstanding problem. The common occurrence of a strong first order jump in the infinite rigid cluster along with a continuous elastic behavior [15,51,56] is surprising and also largely unexplored.

Rigidity percolation is emerging as a geometric model for structural glasses [85]. Clearly thermal effects are critical in many applications of rigidity percolation so that models which treat both geometric and thermal effects are valuable. Molecular dynamics models provide insight into the interplay of geometry and temperature and warrant further study. In particular the evolution of the boson peak in model glasses as a function of geometry and temperature remains unexplored and would help resolve contrasting models which emphasize either temperature effects [58] or geometric effects [85].

The role of generic as compared to non-generic configurations is critical to granular media and to the elasticity of lattices. Even the most basic questions remain unresolved, for example are the elastic exponents of regular lattices the same as the elastic exponents of random lattices with the same co-ordination? Is the elasticity of random generic networks inherently non-linear due to alignment of bonds under stress?

Nevertheless, rigidity percolation provides a useful model for a broad range of physical phenomena, significantly generalizing the connectivity percolation model. Combinatorial methods drawn from graph theory have greatly extended our understanding of rigidity, providing both practical tools as well as conceptual novelties including the distinction between generic and non-generic rigidity. The applications of related geometric methods to granular media and to semiflexible gels is extremely rich and are harbingers of a broader use of rigidity concepts in the analysis of the mechanical behavior of complex materials.

Bibliography

1. Alava MJ, Niskanen KJ (2006) The physics of paper. Rep Prog Phys 69:669–724
2. Alexander S (1998) Amorphous materials: Their structure, lattice dynamics and elasticity. Phys Rep 296:65–236
3. Angell CA (2004) Boson peaks and floppy modes: Some relations between constraint and excitions phenomenology, and interpretation, of glasses and glass transition. J Phys Cond Mat 16:S5153–S5164
4. Axelos MAV, Kolb M (1990) Crosslinked biopolymers: Experimental evidence for scalar percolation theory. Phys Rev Lett 64:1457–1460
5. Barsky SJ, Plischke M (1996) Order and localization in randomly cross-linked polymer networks. Phys Rev E 53:871–876
6. Bausch AR, Kroy K (2006) A bottom-up appoach to cell mechanics. Nat Phys 2:231–238
7. Benguigui L (1984) Experimental study of the elastic properties of a percolating system. Phys Rev Lett 53:2028–2030
8. Bergman DJ (2003) Exact relations between critical exponents for elastic stiffness and electrical conductivity of percolating networks. Physica B 338:240–246
9. Boolchand P, Georgiev DG, Goodman B (2001) Discovery of the intermediate phase in chalcogenide glasses. J Optoelectron Ad Mat 3:703–720
10. Boolchand P, Lucovsky G, Phillips JC, Thorpe MF (2005) Self-organization and the physics of glassy networks. Phil Mag 85:3823–3838
11. Bresser W, Boolchand P, Suranyi P (1986) Rigidity percolation and molecular clustering in network glasses. Phys Rev Lett 56:2493–2496
12. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S et al (2004) Charmm: A program for macromolecular energy, minimization and dynamics calculations. J Comp Chem 4:187–217

13. Cai Y (1989) Floppy modes in network glasses. Phys Rev B 40:10535–10542
14. Chalupa J, Leath PL, Reich GR (1979) Bootstrap percolation on a bethe lattice. J Phys C 12:L31–L35
15. Chubynsky MV, Thorpe MV (2007) Algorithms for 3d-rigidity analysis and a first-order percolation transition. Phys Rev E 76:41135
16. Day AR, Tremblay RR, Tremblay AMS (1986) Rigid backbone: A new geometry for percolation. Phys Rev Lett 56:2501–2504
17. de Gennes PG (1976) On the relation between percolation theory and the elasticity of gels. J Physique 37:L1–L2
18. de Gennes PG (1979) Scaling concepts in polymer physics. Cornell University Press, New York
19. Deptuck D, Harrison JP, Zawadzki P (1985) Measurement of elasticity and conductivity of a three-dimensional percolation system. Phys Rev Lett 54:913–916
20. Dohler GH, Dandoloff R, Bilz H (1980) A topological-dynamical model of amorphycity. J Non-Cryst Sol 42:87–95
21. Donev A, Torquato S, Stillinger FH (2005) Pair correlation function characteristics of nearly jammed disordered and ordered hard-sphere packings. Phys Rev E 71:11105
22. Duxbury PM, Jacobs DJ, Thorpe MF, Moukarzel C (1999) Floppy modes and the free energy: Rigidity and connectivity percolation on bethe lattices. Phys Rev E 59:2084–2092
23. Feng S, Halperin B, Sen PN (1987) Transport properties of continuum systems near the percolation threshold. Phys Rev B 35:197–214
24. Feng S, Sen PN (1984) Percolation on elastic networks: New exponent and threshold. Phys Rev Lett 52:216–219
25. Feng S, Thorpe MF, Garboczi E (1985) Effective-medium theory of percolation on central-force networks. Phys Rev B 31:276–280
26. Feng X, Bresser WJ, Boolchand P (1997) Direct evidence for stiffness threshold in chalcogenide glasses. Phys Rev Lett 78:4422–4425
27. Fisher ME, Essam JW (1961) Some cluster size and percolation problems. J Math Phys 2:609–619
28. Flory PJ (1953) Principles of polymer chemistry. Cornell University Press, New York
29. Grant MC, Russell WB (1993) Volume fraction dependence of elastic moduli and transition temperatures for colloidal silica gels. Phys Rev E 47:2606–2614
30. Guyon E, Roux S, Hansen A, Bideau D, Troadec JP et al (1990) Non-local and non-linear problems in the mechanics of disordered systems: Application to granular media and rigidity problems. Rep Prog Phys 53:373–419
31. Hansen A, Roux S (1989) Universality class of central-force percolation. Phys Rev B 40:749–752
32. He H, Thorpe MF (1985) Elastic properties of glasses. Phys Rev Lett 54:2107–2110
33. Head DA, Levine AJ, MacKintosh FC (2003) Distinct regimes of elastic response and deformation modes of crosslinked cytoskeletal and semiflexible polymer networks. Phys Rev E 68:61907
34. Hendrickson B (1992) Conditions for unique graph realizations. SIAM J Comput 21:65–84
35. Heussinger C, Frey E (2006) Stiff polymers, foams and fiber networks. Phys Rev Lett 96:1–4
36. Jacobs DJ (1998) Generic rigidity in three-dimensional bond-bending networks. J Phys A: Math Gen 31:6653–6668
37. Jacobs DJ, Hendrickson B (1997) An algorithm for two-dimensional rigidity percolation: The pebble game. J Comp Phys 137:346–365
38. Jacobs DJ, Thorpe MF (1995) Generic rigidity percolation: The pebble game. Phys Rev Lett 75:4051–4054
39. Jacobs DJ, Thorpe MF (1996) Generic rigidity percolation in two dimensions. Phys Rev E 53:3682–3693
40. Kamitakahara WA, Cappelletti RL, Boolchand P, Halfpap B, Gompf F et al (1991) Vibrational densities of states and network rigidity in chalcogenide glasses. Phys Rev B 44:94–100
41. Kantor Y, Webman I (1984) Elastic properties of random percolating systems. Phys Rev Lett 52:1891–1894
42. Kasza KE, Rowat AC, Liu J, Angelini RE, Brangwynne CP et al (2007) The cell as a material. Curr Opin Cell Bio 19:101–107
43. Laman G (1970) On graphs and rigidity of plane skeletal structures. J Eng Math 4:331–340
44. LatvaKokko M, Makinen J, Timonen J (2001) Rigidity transition in two-dimensional random fiber networks. Phys Rev E 63:46113
45. LatvaKokko M, Timonen J (2001) Rigidity of random networks of stiff fibers in the low density limit. Phys Rev E 64:66117
46. Majmudar TS, Sperl M, Luding S, Behinger RP (2007) Jamming transition in granular systems. Phys Rev Lett 98:58001
47. Martin JE, Adolf D, Wilcox JP (1988) Viscoelasticity of near-critical gels. Phys Rev Lett 61:2620–2623
48. Maxwell JC (1864) On the calculation of the equilibrium stiffness of frames. Phil Mag 27:294–301
49. Mayo SL, Olafson BD, Goddard WA (1990) Dreiding: A generic force field for molecular simulations. J Phys Chem 94:8897–8909
50. Moukarzel C (1996) An efficient algorithm for testing the generic rigidity of graphs in the plane. J Phys A: Math Gen 29:8079–8098
51. Moukarzel C, Duxbury PM (1990) Comparison of rigidity and connectivity percolation in two dimensions. Phys Rev E 59:2614–2622
52. Moukarzel C, Duxbury PM (1995) Stressed backbone and elasticity of random central-force system. Phys Rev Lett 75:4055–4058
53. Moukarzel C, Duxbury PM, Leath PL (1997) First-order rigidity on cayley trees. Phys Rev E 55:5800–5811
54. Moukarzel C, Duxbury PM, Leath PL (1997) Infinite-cluster geometry in central-force networks. Phys Rev Lett 78:1480–1483
55. Moukarzel CF (1998) Isostatic phase transition and instability in stiff granular materials. Phys Rev Lett 81:1634–1637
56. Obukhov SP (1995) First order rigidity transition in random rod networks. Phys Rev Lett 74:4472–4475
57. Ohern CS, Silbert LE, Liu AJ, Nagel SR (2003) Jamming at zero temperature and zero applied stress: The epitome of disorder. Phys Rev E 68:11306
58. Parisi G (2003) On the origin of the boson peak. J Phys Cond Mat 15:S765–S774
59. Phillips JC (1979) Topology of covalent non-crystalline solids i: Short-range order in chalcogenide alloys. J Non-Cryst Sol 34:153–181
60. Plischke M (2006) Critical behavior of entropic shear rigidity. Phys Rev E 73:61406
61. Plischke M, Joos B (1998) Entropic elasticity of diluted central force networks. Phys Rev Lett 80:4907–4910
62. Plischke M, Vernon DC, Joos B, Zhou Z (1998) Entropic elasticity of diluted central force networks. Phys Rev E 60:3129–3135

63. Rader AJ, Hespenheide BM, Kuhn LA, Thorpe MF (2002) Protein unfolding: Rigidity lost. PNAS 99:3540–3545
64. Rivoire O, Barré J (2006) Exactly solvable models of adaptive networks. Phys Rev Lett 97:148701
65. Ross-Murphy SB (2007) Biopolymer gelation-exponents and critical exponents. Polym Bull 58:119–126
66. Rueb CJ, Zukoski CF (1997) Viscoelastic properties of colloidal gels. J Rheol 41:197–218
67. Sahimi M (1998) Non-linear and non-local transport processes in heterogeneous media: From long-range correlated percolation to fracture and materials breakdown. Phys Rep 306:214–395
68. Sahimi M, Arbabi S (1993) Mechanics of disordered solids. i. percolation on elastic networks with central forces. Phys Rev B 47:695–702
69. Sahimi M, Arbabi S (1993) Mechanics of disordered solids. ii. percolation on elastic networks with bond-bending forces. Phys Rev B 47:703–712
70. Schwartz LM, Feng S, Thorpe MF, Sen PN (1985) Behavior of depleted elastic networks: Comparison of effective medium theory and numerical simulations. Phys Rev B 32:4607–4617
71. Schwarz JH, Liu AJ, Chayes LQ (2006) The onset of jamming as the sudden emergence of an infinite k-core cluster. Europhys Lett 73:560–566
72. Serrano D, Peyrelasse J, Boned C, Harran D, Monge P (1990) Application of the percolation model to gelation of an epoxy resin. J Appl Poly Sci 39:670–693
73. Stauffer D, Coniglio A, Adam M (1982) Gelation and critical phenomena. Adv Poly Sci 44:103–158
74. Tang W, Thorpe MF (1988) Percolation of elastic networks under tension. Phys Rev B 37:5539–5551
75. Tay TS (1984) Rigidity of multi-graphs. i. linking rigid bodies in n-space. J Comb Theory B 36:95–112
76. Tay TS, Whiteley W (1985) Generating isostatic frameworks. Struct Topol 11:21–68
77. Tharmann R, Claessens MMAE, Bausch AR (2007) Viscoelasticity of isotropically crosslinked actin networks. Phys Rev Lett 98:88103
78. Thorpe MF (1983) Continuous deformations in random networks. J Non-Cryst Sol 57:355–370
79. Thorpe MF, Duxbury PM (eds) (1999) Rigidity theory and applications. Plenum, New York
80. Thorpe MF, Jacobs DJ, Chubynsky MV, Phillips JC (2000) Self-organization in network glasses. J Non-Cryst Sol 266-269:859–866
81. Uebbing B, Sievers AJ (1996) Role of network topology on the vibrational lifetime of an h_2o molecule in the ge-as-se glass series. Phys Rev Lett 76:932–935
82. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general amber force field. J Comp Chem 25:1157–1174
83. Whiteley W (2005) Counting out to the flexibility of molecules. Phys Biol 2:S116–S126
84. Winter HH, Mours M (1997) Rheology of polymers near the liquid-solid transition. Adv Poly Sci 134:165–234
85. Wyart M (2005) On the rigidity of amorphous solids. Ann Phys Fr 30:1–96
86. Wyart M, Nagel SR, Witten TA (2005) Geometric origin of excess low-frequency vibrational modes in weakly connected amorphous solids. Europhys Lett 72:486–492
87. Wyss HM, Tervoort EV, Gauckler LJ (2005) Mechanics and microstructures of concentrated particle gels. J Am Ceram Soc 88:2337–2348
88. Xing X, Mukhopadhyay S, Goldbart PM (2004) Scaling of entropic shear rigidity. Phys Rev Lett 93:225701
89. Xu J, Bohnsack DA, Mackay ME, Wooley KL (2007) Unusual mechanical performance of amphiphilic crosslinked polymer networks. J Amer Chem Soc Communications 129:506–507
90. Zabolitzky JG, Bergman DJ, Stauffer D (1985) Precision calculation of elasticity for percolation. J Stat Phys 54:913–916

Embodied and Situated Agents, Adaptive Behavior in

Stefano Nolfi
Institute of Cognitive Sciences and Technologies, National Research Council (CNR), Rome, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Embodiment and Situatedness
Behavior and Cognition as Complex Adaptive Systems
Adaptive Methods
Evolutionary Robotics Methods
Discussion and Conclusion
Bibliography

Glossary

Phylogenesis Indicates the variations of the genetic characteristics of a population of artificial agents throughout generations.

Ontogenesys Indicates the variations which occur in the phenotypical characteristics of an artificial agent (i. e. in the characteristics of the control system or of the body of the agent) while it interacts with the environment.

Embodied agent Indicates an artificial system (simulated or physical) which has a body (characterized by physical properties such us shape, dimension, weight, etc), actuators (e. g. motorized wheels, motorized articulated joints), and sensors (e. g. touch sensors or vision sensors). For a more restricted definition see the concluding section of the paper.

Situated agent Indicates an artificial system which is located in a physical environment (simulated or real) with which it interacts on the basis of the law of physics. For a more restricted definition see the concluding section of the paper.

Morphological computation Indicates the ability of the body of an agent (with certain specific characteristics) to control its interaction with the environment so to produce a given desired behavior.

Definition of the Subject

Adaptive behavior concerns the study of how organisms develop their behavioral and cognitive skills through a synthetic methodology which consists in designing artificial agents which are able to adapt to their environment autonomously. These studies are important both from a modeling point of view (i. e. for making progress in our understanding of intelligence and adaptation in natural beings) and from an engineering point of view (i. e. for making progresses in our ability to develop artefacts displaying effective behavioral and cognitive skills).

Introduction

Adaptive behavior research concerns the study of how organisms can develop behavioral and cognitive skills by adapting to the environment and to the task they have to fulfill autonomously (i. e. without human intervention). This goal is achieved through a synthetic methodology, i. e. through the synthesis of artificial creatures which: (i) have a body, (ii) are situated in an environment with which they interact, and (iii) have characteristics which vary during an adaptation process. In the rest of the paper we will use the term "agent" to indicate artificial creatures which posses the first two features described above and the term "adaptive agent" to indicate artificial creatures which also posses the third feature.

The agents and the environment might be simulated or real. In the former case the characteristics of agents' body, motor, and sensory system, the characteristics of the environment, and the rules that regulate the interactions between all the elements are simulated on a computer. In the latter case, the agents consist of physical entities (mobile robots) situated in a physical environment with which they interact on the basis of the physical laws.

The adaptive process which regulates how the characteristics of the agents (and eventually of the environment change) change might consist of a population-based evolutionary process and/or of a developmental/learning process. In the former case, the characteristics of the agents do not vary during their "lifetime" (i. e. during the time in which the agents interact with the environment) but phylogenetically, while individual agents "reproduce". In the latter case, the characteristics of the agents vary ontogenetically, while they interact with the environment. The criteria which determine how variations are generated and/or whether or not variations are retained can be task-dependent and/or task-independent, i. e. might be based on an evaluation of whether the variation increase or decrease agents' ability to display a behavior which is adapted to the task/environment or might be based on task-independent criteria (i. e. general criteria which do not reward directly the exhibition of the requested skill).

The paper is organized as follows. In Sect. "Introduction" we briefly introduce the notion of *embodiment* and *situatedness* and their implications. In Sect. "Embodiment and Situatedness" we claim that behavior and cognition in embodied and situated adaptive agents should be characterized as a complex adaptive system. In Sect. "Behavior and Cognition as Complex Adaptive Systems" we briefly describe the methods which can be used to synthesize embodied and situated adaptive agents. Finally in Sect. "Adaptive Methods", we draw our conclusions.

Embodiment and Situatedness

The notion of *embodiment* and *situatedness* has been introduced [8,9,12,34,48] to characterize systems (e. g. natural organism and robots) which have a physical body and which are situated in a physical environment with which they interact. In this and in the following sections we will briefly discuss the general implications of these two fundamental properties. This analysis will be further extended in the concluding section where we will claim on the necessity to distinguish between a weak and a strong notion of embodiment and situatedness.

One first important implication of being embodied and situated consists in the fact that these agents and their parts are characterized by their physical properties (e. g. weight, dimension, shape, elasticity etc.), are subjected to the laws of physics (e. g. inertia, friction, gravity, energy consumption, deterioration etc.), and interact with the environment through the exchange of energy and physical material (e. g. forces, sound waves, light waves etc.). Their physical nature also implies that they are quantitative in state and time [49]. The fact that these agents are quantitative in time implies, for example, that the joints which connect the parts of a robotic arm can assume any possible position within a given range. The fact that these agents are quantitative in time implies, for example, that the effects of the application of a force to a joint depend from the time duration of its application.

One second important implication is that the information measured by the sensors is not only a function of the environment but also of the relative position of the agent in the environment. This implies that the motor actions performed by an agent, by modifying the

agent/environmental relation or the environment, co-determine the agent sensory experiences.

One third important implication is that the information measured by the sensors provide information about the external environment which is egocentric (depends from the current position and the orientation of the agent in the environment), local (only provide information related to the local observable portion of the environment), incomplete (due to visual occlusion, for example), and subjected to noise. Similar characteristics apply to the motor actions produced by the agent's effectors.

It is important to notice that these characteristics do not only represent constraints but also opportunities to be exploited. Indeed, as we will see in the next section, the exploitation of some of these characteristics might allow embodied and situated agents to solve their adaptive problems through solutions which are robust and parsimonious (i. e. minimal) with respect to the complexity of the agent's body and control system.

Behavior and Cognition as Complex Adaptive Systems

In embodied and situated agents, behavioral and cognitive skills are dynamical properties which unfold in time and which arise from the interaction between agents' nervous system, body, and the environment [3,11,19,29,31] and from the interaction between dynamical processes occurring within the agents' control system, the agents' body, and within the environment [4,15,45]). Moreover, behavioral and cognitive skills typically display a multi-level and multi-scale organization involving bottom-up and top-down influences between entities at different levels of organization. These properties imply that behavioral and cognitive skills in embodied and situated agents can be properly characterized as complex adaptive systems [29].

These aspects and the complex system nature of behavior and cognition will be illustrated in more details in the next subsections also with the help of examples. The theoretical and practical implication of these aspects for developing artificial agents able to exhibit effective behavioral and cognitive skills will be discussed in the forthcoming sections.

Behavior and Cognition as Emergent Dynamical Properties

Behavior and cognition are dynamical properties which unfold in time and which emerge from high-frequent non-linear interactions between the agent, its body, and the external environment [11].

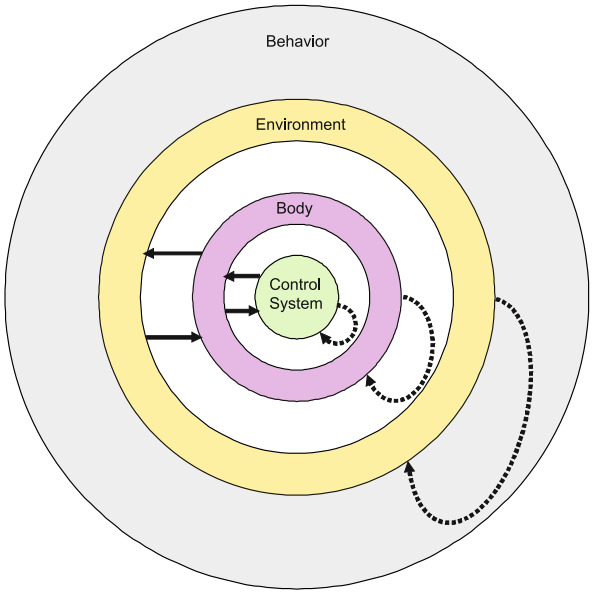

Embodied and Situated Agents, Adaptive Behavior in, Figure 1
A schematic representation of the relation between agent's control system, agent's body, and the environment. The behavioral and cognitive skills displayed by the agent are the emergent result of the bi-directional interactions (represented with *full arrows*) between the three constituting elements – agent's control system, agent's body, and environment. The *dotted arrows* indicate that the three constituting elements might be dynamical systems on their own. In this case, agents' behavioral and cognitive skills result of the dynamics originating from the agent/body/environmental interactions but also from the combination and the interaction between dynamical processes occurring within the agent's body, within the agent's control system, and within the environment (see Sect. "Embodiment and Situatedness")

At any time step, the environmental and the agent/environmental relation co-determine the body and the motor reaction of the agent which, in turn, co-determines how the environment and/or the agent/environmental relation vary. Sequences of these interactions, occurring at a fast time rate, lead to a dynamical process – behavior – which extends over significant larger time span than the interactions (Fig. 1).

Since interactions between the agent's control system, the agent's body, and the external environment are non-linear (i. e. small variations in sensory states might lead to significantly different motor actions) and dynamical (i. e. small variations in the action performed at time t might significantly impact later interactions at time t_{+x}) the relation between the rules that govern the interactions and the behavioral and cognitive skills originating from the interactions tend to be very indirect. Behavioral and cogni-

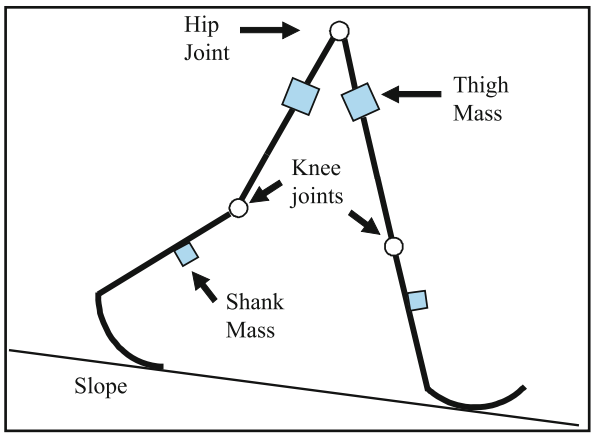

Embodied and Situated Agents, Adaptive Behavior in, Figure 2
A schematization of the passive walking machine developed by McGeer [24]. The machine includes two passive knee joints and a passive hip joint

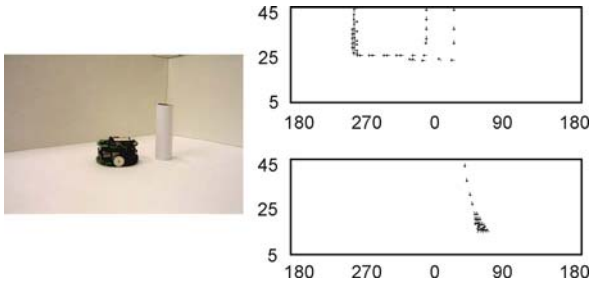

Embodied and Situated Agents, Adaptive Behavior in, Figure 3
***Left*: The agent situated in the environment. The agent is a Khepera robot [26]. The environment consists of an arena of 60 × 35 cm containing cylindrical objects placed in a randomly selected location. *Right:* Angular trajectories of an evolved robot close to a wall (*top graph*) and to a cylinder (*bottom graph*). The picture was obtained by placing the robot at a random position in the environment, leaving it free to move for 500 time steps each lasting 100 ms, and recording its relative movements with respect to the two types of objects for distances smaller than 45 mm. The *x*-axis and the *y*-axis indicate the relative angle (in degrees) and distance (in mm) between the robot and the corresponding object. For sake of clarity, *arrows* are used to indicate the relative direction, but not the amplitude of movements**

tive skills thus emerge from the interactions between the three foundational elements and cannot be traced back to any of the three elements taken in isolation. Indeed, the behavior displayed by an embodied and situated agent can hardly be predicted or inferred from an external observer even on the basis of a complete knowledge of the interacting elements and of the rules governing the interactions.

A clear example of how behavioral skill might emerge from the interaction between the agents' body and the environment is constituted by the passive walking machines developed in simulation by McGeer [24] – a two-dimensional bipedal machines able to walk down a four-degree slope with no motors and no control system (Fig. 2). The walking behavior arises from the fact that the physical forces resulting from gravity and from the collision between the machine and the slope produce a movement of the robot and the fact that robot's movements produce a variation of the agent-environmental relation which in turn produce a modification of the physical forces to which the machine will be subjected in the next time step. The sequence of by-directional effects between the robot's body and the environment can lead to a stable dynamical process – the walking behavior.

The type of behavior which arises from the robot/environmental interaction depends from the characteristics of the environment, the physics law which regulate the interaction between the body and the environment, and the characteristics of the body. The first two factors can be considered as fixed but the third factor, the body structure, can be adapted to achieve a given function. Indeed, in the case of this biped robot, the author carefully selected the leg length, the leg mass, and the foot size to obtain the desired walking behavior. In more general term, this example shows how the role of regulating the interaction between the robot and the environment in the appropriate way can be played not only but the control system but also from the body itself providing that the characteristics of the body has been shaped so to favor the exhibition of the desired behavior. This property, i. e. the ability of the body to control its interaction with the environment, has been named with the term "morphological computation" [35]. For related work which demonstrate how effective walking machines can be obtained by integrating passive walking techniques with simple control mechanisms, see [6,13,50]. For related works which show the role of elastic material and elastic actuators for morphological computing see [23,40].

To illustrate of how behavioral and cognitive skills might emerge from agent's body, agent's control system, and environmental interactions we describe a simple experiment in which a small wheeled robot situated in an arena surrounded by walls has been evolved to find and to remain close to a cylindrical object. The Khepera robot [26] is provided with eight infrared sensors and two motors controlling the two corresponding wheels (Fig. 3).

From the point of view of an external observer, solving this problem requires robots able to: (a) explore the environment until an obstacle is detected, (b) discriminate whether the obstacle detected is a wall or a cylindrical ob-

ject, and (c) approach or avoid objects depending on the object type. Some of these behaviors (e. g. the wall-avoidance behavior) can be obtained through simple control mechanisms but others require non trivial control mechanisms. Indeed, a detailed analysis of the sensory patterns experienced by the robot indicated that the task of discriminating the two objects is far from trivial since the two classes of sensory patterns experienced by robots close to a wall and close to cylindrical objects largely overlap.

The attempt to solve this problem through an evolutionary adaptive method (see Sect. "Behavior and Cognition as Complex Adaptive Systems") in which the free parameters (i. e. the parameters which regulate the fine-grained interaction between the robot and the environment) are varied randomly and in which variations are retained or discarded on the basis on an evaluation of the overall ability of the robot (i. e. on the basis of the time spent by the robot close to the cylindrical object) demonstrated how adaptive robots can find solutions which are robust and parsimonious in term of control mechanisms [28]. Indeed, in all replications of these experiment, evolved robot solve the problem by moving forward, by avoiding walls, and by oscillating back and fourth and left and right close to cylindrical objects (Fig. 3, right). All these behaviors result from sequences of interactions between the robot and the environment mediated by four types of simple control rules which consist in: turning left when the right infrared sensors are activated, turning right when the left infrared sensors are activated, moving back when the frontal infrared sensors are activated, and moving forward when the frontal infrared sensors are not activated.

To understand how these simple control rules can produce the required behaviors and the required arbitration between behaviors we should consider that the same motor responses produce different effects on different agent/environmental situations. For example, the execution of a left-turning action close to a cylindrical object and the subsequent modification of the robot/object relative position produce a new sensory state which triggers a right-turning action. Then, the execution of the latter action and the subsequent modification of the robot/object relative position produce a new sensory state which triggers a left-turning action. The combination and the alternation of these left and right-turning actions over time produce an attractor in the agent/environmental dynamics (Fig. 3, right, bottom graph) which allows the robot to remain close to the cylindrical object. On the other hand the execution of a left-turning behavior close to a wall object and the subsequent modification of the robot/wall position produce a new sensory state which triggers the reiteration of the same motor action. The execution of a sequence of left-turning action then leads to the avoidance of the object and to a modification of the robot/environmental relation which finally lead to a perception of a sensory state which trigger a move-forward behavior (Fig. 4, right, top graph).

Before concluding the description of this experiment, it is important to notice that, although the rough classification of the robot motor responses into four different types of actions is useful to describe the strategy with which these robots solve the problem qualitatively, the quantitative aspects which characterize the robot motor reactions (e. g. how sharply a robot turns given a certain pattern of activation of the infrared sensors) are crucial for determining whether the robot will be able to solve the problem or not. Indeed, small differences in the robot's motor response tend to cumulate in time and might prevent the robot for producing successful behavior (e. g. might prevent the robot to produce a behavioral attractor close to cylindrical objects).

This experiment clearly exemplifies some important aspects which characterize all adaptive behavioral system, i. e. systems which are embodied and situated and which have been designed or adapted so to exploit the properties that emerge from the interaction between their control system, their body, and the external environment. In particular, it demonstrates how required behavioral and cognitive skills (i. e. object categorization skills) might emerge from the fine-grained interaction between the robot's control system, body, and the external environment without the need of dedicated control mechanisms. Moreover, it demonstrates how the relation between the control rules which mediate the interaction between the robot body and the environment and the behavioral skills exhibited by the agents are rather indirect. This means, for example, that an external human observer can hardly predict the behaviors which will be produced by the robot, before observing the robot interacting with the environment, even on the basis of a complete description of the characteristics of the body, of the control rules, and of the environment.

Behavior and Cognition as Phenomena Originating from the Interaction Between Coupled Dynamical Processes

Up to this point we restricted our analysis to the dynamics originating from the agent's control system, agents' body, and environmental interactions. However, the body of an agent, its control system, and the environment might have their own dynamics (dotted arrows in Fig. 1). For the sake of clarity, we will refer to the dynami-

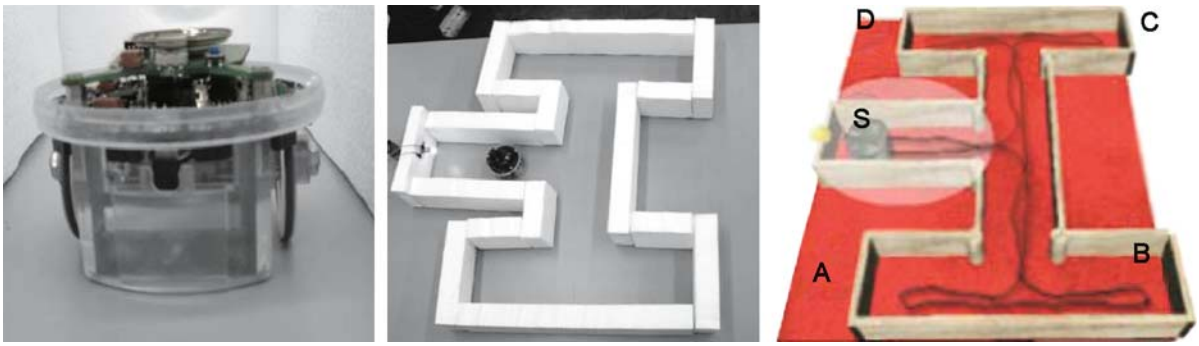

Embodied and Situated Agents, Adaptive Behavior in, Figure 4
***Left*: The e-puck robot developed at EPFL, Switzerland http://www.e-puck.org/. *Center*: The environment which have a size of 52 cm by 60 cm. The light produced by the light bulb located on the *left side* of the central corridor cannot be perceived from the other two corridors. *Right*: The motor trajectory produced by the robot during a complete lap of the environment**

cal processes occurring within the agent control system, within the agent body, or within the environment as *internal dynamics* and to the dynamics originating from the agent/body/environmental interaction as *external dynamics*. In cases in which agents' body, agents' control system, or the environment have their own dynamics, behavior should be characterized as a property emerging from the combination of several coupled dynamical processes.

The existence of several concurrent dynamical processes represents an important opportunity for the possibility to exploit emergent features. Indeed, behavioral and cognitive skills might emerge not only from the external dynamics, as we showed in the previous section, but also from the internal dynamical processes or from the interaction between different dynamical processes.

As an example which illustrates how complex cognitive skills can emerge from the interaction between a simple agent/body/environmental dynamic and a simple agent's internal dynamic consider the case of a wheeled robot placed in a maze environment (Fig. 4) which has been trained to: (a) produce a wall-following behavior which allows the robot to periodically visit and re-visit all environmental areas, (b) identify a target object constituted by a black disk which is placed in a randomly selected position of the environment for a limited time duration, and (c) recognize the location in which the target object was previously found every time the robot re-visit the corresponding location [15].

The robot has infrared sensors (which provide information about nearby obstacles), light sensors (which provide information about the light gradient generated by the light bulb placed in the central corridor), ground sensors (which detect the color of the ground), two motors (which control the desired speed of the two corresponding wheels), and one additional output units which should be turned on when the robot re-visit the environmental area in which the black disk was previously found. The robot's controller consists of a three layers neural network which includes a layer of sensory neurons (which encode the state of the corresponding sensors), a layer of motor neurons which encode the state of the actuators, and a layer of internal neurons which consist of leaky integrators operating at tuneable time scale [3,15]. The free parameters of the robot's neural controllers (i. e. the connection weights, and the time constant of the internal neurons which regulate the time rate at which this neurons change their state over time) were adapted through an evolutionary technique [31].

By analyzing the evolved robot the authors observed how they are able to generate a spatial representation of the environment and of their location in the environment while they are situated in the environment itself. Indeed, while the robot travel by performing different laps of the environment (see Fig. 4, right), the states of the two internal neurons converge on a periodic limit cycle dynamic in which different states correspond to different locations of the robot in the environment (Fig. 5).

As we mentioned above, the ability to generate this form of representation which allow the robot to solve its adaptive problem originate from the coupling between a simple robot's internal dynamics and a simple robot/body/environmental dynamics. The former dynamics is characterized by the fact that the state of the two internal neurons tends to move slowly toward different fixed point attractors, in the robot's internal dynamics, which correspond to different type of sensory states exemplified in Fig. 5. The latter dynamics originate from the

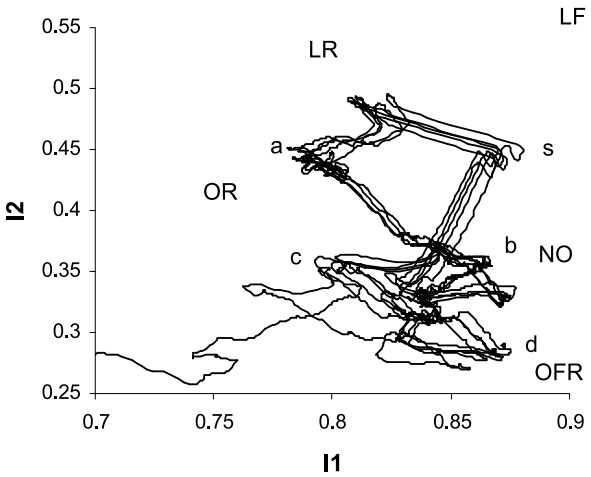

Embodied and Situated Agents, Adaptive Behavior in, Figure 5
The state of the two internal neurons (i1 and i2) of the robot recorded for 330 s while the robot performs about 5 laps of the environment. The s, a, b, c, and d labels indicate the internal states corresponding to five different positions of the robot in the environment shown in Fig. 4. The other labels indicate the position of the fixed point attractors in the robot's internal dynamics corresponding to five types of sensory states experienced by the robot when it detects: a light in its frontal side (LF), a light on its rear side (LR), an obstacle on its right and frontal side (OFR), an obstacle on its right side (OR), no obstacles and no lights (NO)

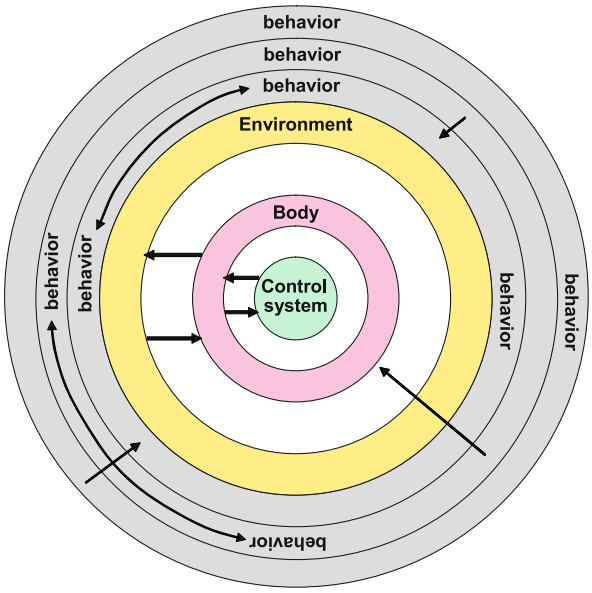

Embodied and Situated Agents, Adaptive Behavior in, Figure 6
A schematic representation of multi-level and multi-scale organization of behavior. The behaviors represented in the *inner circles* represent elementary behaviors which arise from fine-grained interactions between the control system, the body, and the environment, and which extend over limited time spans. The behaviors represented in the *external circles* represent higher level behaviors which arise from the combination and interaction between lower-level behaviors and which extend over longer time spans. The *arrows* which go from higher level behavior toward lower levels indicate the fact that the behaviors currently exhibited by the agents later affect the lower level behaviors and/or the fine-grained interaction between the constituting elements (agent's control system, agent's body, and the environment)

fact that different types of sensory states last for different time durations and alternate with a given order while the robot move in the environment. The interaction between these two dynamical processes leads to a transient dynamics of agents' internal state which moves slowly toward the current fixed point attractor without never fully reaching it (thus preserving information about previously experienced sensory states, the time duration of these states, and the order with which they have been experienced). The coupling between the two dynamical processes originates from the fact that the free parameters which regulate the agent/environmental dynamics (e. g. the trajectory and the speed with which the robot moves in the environment) and the agent internal dynamics (e. g. the direction and the speed with which the internal neurons change their state) have been co-adapted and co-shaped during the adaptive process.

For related works which show how navigation and localization skills might emerge from the coupling between agent's internal and external dynamics, see [45]. For other works addressing other behavioral/cognitive capabilities see [4] for what concerns categorization, [16,41] for what concerns selective attention and [44] for what concern language and compositionality.

Behavior and Cognition as Phenomena with a Multi-Level and Multi-Scale Organization

Another fundamental feature that characterizes behavior is the fact that it is a multi-layer system with different levels of organizations extending at different time scales [2,19]. More precisely, as exemplified in Fig. 6, the behavior of an agent or of a group of agents involve both lower and higher level behaviors which extend for shorter or longer time spans, respectively. Lower level behaviors arise from few agent/environmental interactions and short term internal dynamical processes. Higher level behaviors, instead, arise from the combination and interaction of lower level behaviors and/or from long term internal dynamical processes.

The multi-level and multi-scale organization of agents' behavior play important roles: it is one of the factors which allow agents to produce functionally useful be-

Embodied and Situated Agents, Adaptive Behavior in, Figure 7
Left: **Four robots assembled into a linear structure.** ***Right*****: A simulation of the robots shown in the** ***left part*** **of the figure**

havior without necessarily developing dedicated control mechanisms [8,9,29], it might favor the development of new behavioral and/or cognitive skills thanks to the recruitment of pre-existing capabilities [22], it allow agents to generalize their skills in new task/environmental conditions [29].

An exemplification of how the multi-level and multi-scale organization of behavior allow agents to generalize their skill in new environmental conditions is represented by the experiments carried our by Baldassarre et al. [2] in which the authors evolved the control system of a group of robots assembled into a linear structure (Fig. 7) for the ability to move in a coordinated manner and for the ability to display a coordinated light approaching behavior.

Each robot [27] consists of a mobile base (chassis) and a main body (turret) that can rotate with respect to the chassis along the vertical axis. The chassis has two drive mechanisms that control the two corresponding tracks and teethed wheels. The turret has one gripper, which allows robots to assemble together and to grasp objects, and a motor controlling the rotation of the turret with respect to the chassis. Robots are provided with a traction sensor, placed at the turret-chassis junction, that detects the intensity and the direction of the force that the turret exerts on the chassis (along the plane orthogonal to the vertical axis) and light sensors. Given that the orientations of individual robots might vary and given that the target light might be out of sight, robots need to coordinate to choose a common direction of movement and to change their direction as soon as one or few robots start to detect a light gradient.

Evolved individuals show the ability to negotiate a common direction of movement and by approaching light targets as soon as a light gradient is detected. By testing evolved robots in different conditions the authors observed that they are able to generalize their skills in new conditions and also to spontaneously produce new behaviors which have not been rewarded during the evolutionary process. More precisely, groups of assembled robots display a capacity to generalize their skills with respect to the number of robots which are assembled together and to the shape formed by the assembled robots. Moreover, when the evolved controllers are embodied in eight robots assembled so to form a circular structure and situated in the maze environment shown in Fig. 8, the robots display an ability to collectively avoid obstacles, to rearrange their shape so to pass through narrow passages, and to explore the environment. The ability to display all these behavioral skills allow the robots to reach the light target even in large maze environments, i. e. even in environmental conditions which are rather different from the conditions that they experienced during the training process (Fig. 8).

By analyzing the behavior displayed by the evolved robots tested in the maze environment, a complex multi-level organization can be observed. The simpler behaviors that can be identified consist of low level individual behaviors which extend over short time spans:

1. A *move-forward behavior* which consists of the individuals' ability to move forward when the robot is coordinated with the rest of the team, is oriented toward the direction of the light gradient (if any), and does not collide with obstacles. This behavior results from the combination of: (a) a control rule which produces a move forward action when the perceived traction has a low intensity and when difference between the intensity of the light perceived on the left and the right side of the robot is low, and (b) the sensory effects of the execution of the move forward action selected mediated by the external environment which does not produce a variation of the state of the sensors until the conditions that should be satisfied to produce this behaviors hold.

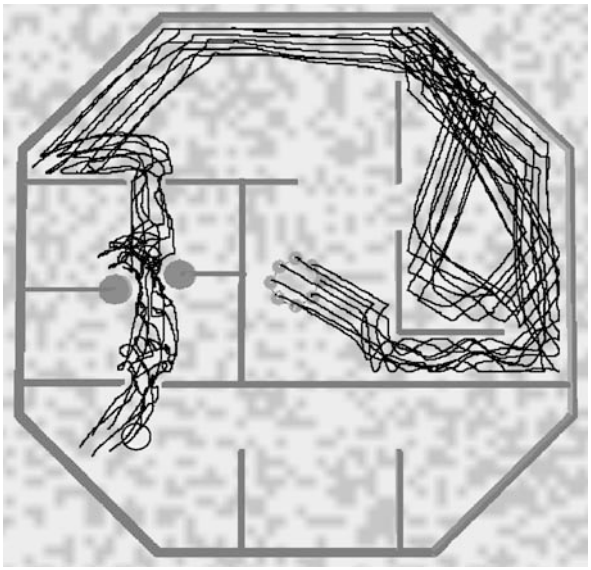

Embodied and Situated Agents, Adaptive Behavior in, Figure 8
The behavior produced by eight robots assembled into a circular structure in a maze environment including walls and cylindrical objects (represented with *gray lines* and *circles*). The robots start in the central portion of the maze and reach the light target located in the bottom-left side of the environment (represented with an *empty circle*) by exhibiting a combination of co-ordinated-movement behaviors, collective obstacle-avoidance, and collective light-approaching behaviors. The *irregular lines*, that represent the trajectories of the individual robots, show how the shape of the assembled robots changes during motion by adapting to the local structure of the environment

2. A *conformistic behavior* which consists of the individuals' ability to conform its orientation with that of the rest of the team when the two orientations differ significantly. This behavior results from the combination of: (a) a control rule that makes the robot turns toward the direction of the traction when its intensity is significant, and (b) the sensory effects produced by the execution of this action mediated by the external environment that lead to a progressive reduction of the intensity of the traction until the orientation of the robot conform with the orientation of the rest of the group.
3. A *phototaxis behavior* which consists of the individuals' ability to orient toward the direction of the light target. This behavior results from the combination of: (a) a control rule that makes the robot turns toward the direction in which the intensity of the light gradient is higher, and (b) the sensory effects produced by the execution of this action mediated by the external environment that lead to a progressive reduction of the difference in the light intensity detected on the two side of the robot until the orientation of the robot conforms with the direction of the light gradient.
4. An *obstacle-avoidance behavior* which consists of the individuals' ability to change direction of motion when the execution of a motor action produced a collision with an obstacle. This behavior results from the combination of: (a) the same control rule which lead to behavior #2 which make the robot turns toward the direction of the perceived traction (which in this case is caused by the collision with the obstacle while in the case of behavior #2 is caused by the forces exhorted by the other assembled robots), and (b) the sensory effects produced by the execution of the turning action mediated by the external environment which make the robot turns until collisions do not prevent anymore the execution of a moving forward behavior.

The combination and the interaction between these three behaviors produce the following higher levels collective behaviors that extend over a longer time span:

5. A *coordinated-motion behavior* which consists in the ability of the robots to negotiate a common direction of movement and to keep moving along such direction by compensating further misalignments originating during motion. This behavior emerges from the combination and the interaction of the conformistic behavior (which plays the main role when robots are misaligned) and the move-forward behavior (which plays the main role when robots are aligned).
6. A *coordinated-light-approaching behavior* which consists in the ability of the robots to co-ordinately move toward a light target. This behavior emerges from the combination of the conformistic, the move-forward, and the phototaxis behaviors (which is triggered when the robots detect a light gradient). The relative importance of the three control rules which lead to the three corresponding behaviors depends both on the strength of the corresponding triggering condition (i. e. the extent of lack of traction forces, the intensity of traction forces, and the intensity of the light gradient, respectively) and on a priority relations among behaviors (i. e. the fact that the conformistic behavior tends to play a stronger role than the phototaxis behavior).
7. A *coordinated-obstacle-avoidance behavior* which consists in the ability of the robots to co-ordinately turn to avoid nearby obstacles. This behavior arises as the result of the combination of the obstacle avoidance-, the conformistic and the move-forward behaviors.

The combination and the interaction between these behaviors lead to the following higher levels collective behaviors that extend over still longer time spans:

8. A *collective-exploration-behavior* which consists in the ability of the robots to visit different area on the environment when the light target cannot be detected. This behavior emerges from the combination of the coordinated-motion behavior and the coordinate obstacle-avoidance behavior which ensures that the assembled robots can move in the environment without getting stuck and without entering into limit cycle trajectories.
9. A *shape-re-arrangement behavior* which consists in the ability of the assembled robots to dynamically adapt their shape to the current structure of the environment so to pass through narrow passages especially when the passages to be negotiated are in the direction of the light gradient. This behavior emerges from the combination and the interaction between coordinated motion and coordinated-light-approaching behaviors mediated by the effects produced by relative differences in motion between robots resulting from the execution of different motor actions and/or from differences in the collisions. The fact that the shape of the assembled robots adapt to the current environmental structure so to facilitate the overcoming of narrow passages can be explained by considering that collisions produce a modification of the shape which affect on particular the relative position of the colliding robots.

The combination and the interaction of all these behavior leads to a still higher level behavior:

10. A *collective-navigation-behavior* which consists in the ability of the assembled robots to navigate toward the light target by producing coordinated movements, exploring the environment, passing through narrow passages, and producing a coordinated-light-approaching behavior (Fig. 8).

This analysis illustrates two important mechanisms which explain the remarkable generalization abilities of these robots. The first mechanism consists in the fact that the control rules which regulate the interaction between the agents' and the environment so to produce certain behavioral skills in certain environmental conditions will produce different but related behavioral skills in other environmental conditions. In particular, the control rules which generate the behaviors #5 and #6 for which evolving robots have been evolved in an environment without obstacles also produce behavior #7 in an environment with obstacles. The second mechanism consists in the fact that the development of certain behaviors at a given level of organization which extend for a given time span will automatically lead to the exhibition of related higher-level behaviors extending at longer time spans which originate form the interactions from the former behaviors (even if these higher level behaviors have not being rewarded during the adaptation process). In particular, the combination and the interaction of behaviors #5, #6, and #7 (which have been rewarded during the evolutionary process or which arise from the same control rules which lead to the generation of rewarded behaviors) automatically lead to the production of behaviors #8, #9, and #10 (which have not been rewarded). Obviously, there no warranty that the new behaviors obtained as a result of these generalization processes will play useful functions. However, the fact that these behaviors are related to the other functional behavioral skills implies that the probabilities that these new behavior will play useful functions is significant.

In principle, these generalization mechanisms can also be exploited by agents during their adaptive process to generate behavioral skills which play new functionalities and which emerge from the combination and the interaction between pre-existing behavioral skills playing different functions.

On the Top-Down Effect from Higher to Lower Levels of Organization

In the previous sections we have discussed how the interactions between the agents' body, the agents' control system, and the environment lead to behavioral and cognitive skills and how such skills have a multi-level and multi-scale organization in which the interaction between lower-level skills lead to the emergence of higher-level skills. However, higher level skills also affect lower level skills up to the fine-grained interaction between the constituting elements (agents' body, agents' control system, and environment). More precisely, the behaviors which originate from the interaction between the agent and the environment and from the interaction between lower levels behaviors, later affect the lower levels behaviors and the interaction from which they originate. These bi-directional influences between different levels of organization can lead to circular causality [20] where high level processes act as independent entities which constraint the lower level processes from which they originate.

One of the most important effects of this top-down influences consists in the fact that the behavior exhibited by an agent constraint the type of sensory patterns that the agent will experience later on (i. e. constraint the fine-grained agent/environmental interactions which determine the behavior that will be later exhibited by the agent). Since the complexity of the problem faced by an

agent depends on the sensory information experienced by the agent itself, this top down influences can be exploited in order to turn hard problems into simple ones.

One neat demonstration of this type of phenomena is given by the experiments conducted by Marocco and Nolfi [32] in which a simulated finger robot with six degree of freedom provided with sensors of its joint positions and with rough touch sensors is asked to discriminate between cubic and spherical objects varying in size. The problem is not trivial since, in general terms, the sensory patterns experienced by the robot do not provide clear regularities for discriminating between the two types of objects. However, the type of sensory states which are experienced by the agent also depend on the behavior previously exhibited by the agent itself – agents exhibiting different behavior might face simpler or harder problems. By evolving the robots in simulation for the ability to solve this problem and by analyzing the complexity of the problem faced by robots of successive generations, the authors observed that the evolved robot manage to solve their adaptive problem on the basis of simple control rules which allow the robot to approach the object and to move following the surface of the object from left to right, independently from the object shape. The exhibition of this behavior in interaction with objects characterized by a smooth or irregular surface (in the case of spherical or cubic objects, respectively) ensures that the same control rules lead to two types of behaviors depending on the type of the object. These behaviors consist in following the surface of the object and then moving away from the object in the case of spherical objects, and in following the surface of the object by getting stuck in a corner in the case of cubic objects. The exhibition of these two behaviors allows the agent to experience rather different proprioceptors states as a consequence of having had interacted with spherical or cubic object which nicely encode the regularities which are necessary to differentiate the two types of objects.

For other examples which shows how adaptive agents can exploit the fact that behavioral and cognitive processes which arise from the interaction between lower-level behaviors or between the constituting elements later affect these lower level processes see [4,28,39].

Adaptive Methods

In this section we briefly review the methods through which artificial embodied and situated agents can develop their skill autonomously while they interact at different levels of organization with the environment and eventually with other agents. These methods are inspired by the adaptive process observed in nature: evolution, maturation, development, and learning.

We will focus in particular on self-organized adaptive methodologies in which the role of the experimenter/designer is reduced to the minimum and in which the agents are free to develop their strategy to solve their adaptive problems within a large number of potentially alternative solutions. This choice is motivated by the following considerations:

(a) These methods allow agents to identify the behavioral and cognitive skills which should be possessed, combined, and integrated so to solve the given problem. In other words, these methods can come up with effective ways of decomposing the overall required skill into a collection of simpler lower levels skills. Indeed, as we showed in the previous section, evolutionary adaptive techniques can discover ways of decomposing the high-level requested skill into lower-levels behavioral and cognitive skills so find solutions which are effective and parsimonious thanks to the exploitation of properties emerging from the interaction between lower-levels processes and skills and thanks to the recruitment of previously developed skills for performing new functions. In other words, these methods release the designer from the burden of deciding how the overall skill should be divided into a set of simpler skills and how these skills should be integrated. More importantly, these methods can come up with solutions exploiting emergent properties which would be hard to design [17,31].
(b) These methods allow agents to identify how a given behavioral and cognitive skill can be produced, i. e. the appropriate fine-grained characteristics of agents' body structure and control rules regulating the agent/environmental interaction. As for the previous aspect, the advantage of using adaptive techniques lies not only in the fact that the experimenter is released from the burden of designing the fine-grained characteristics of the agents but also in the fact that adaptation might prove more effective than human design due to the inability of an external observer to foresee the effects of large number of non-linear interactions occurring at different levels of organization.
(c) These methods allow agents to adapt to variations of the task, of the environment, and of the social conditions.

Current approaches, in this respect, can be grouped into two families which will be illustrated in the following subsections and which include Evolutionary Robotics methods and Developmental Robotics methods.

Evolutionary Robotics Methods

Evolutionary Robotics [14,31] is a method which allow to create embodied and situated agents able to adapt to their task/environment autonomously through an adaptive process inspired by natural evolution [18] and, eventually, through the combination of evolutionary, developmental, and learning processes.

The basic idea goes as follows (Fig. 9). An initial population of different artificial genotypes, each encoding the control system (and possibly the morphology) of an agent, is randomly created. Each genotype is translated into a corresponding phenotype (i. e. a corresponding agent) which is then left free to act (move, look around, manipulate the environment etc.) while its performance (fitness) with respect to a given task is automatically evaluated. In cases in which this methodology is applied to collective behaviors, agents are evaluated in groups which might be heterogeneous or homogeneous (i. e. might consist of agents which differ not with respect to their genetic and phenotypic characteristics). The fittest individuals (those having higher fitness) are allowed to reproduce by generating copies of their genotype with the addition of changes introduced by some genetic operators (e. g., mutations, exchange of genetic material). This process is repeated for a number of generations until an individual or a group of individuals is born which satisfies the performance level set by the user.

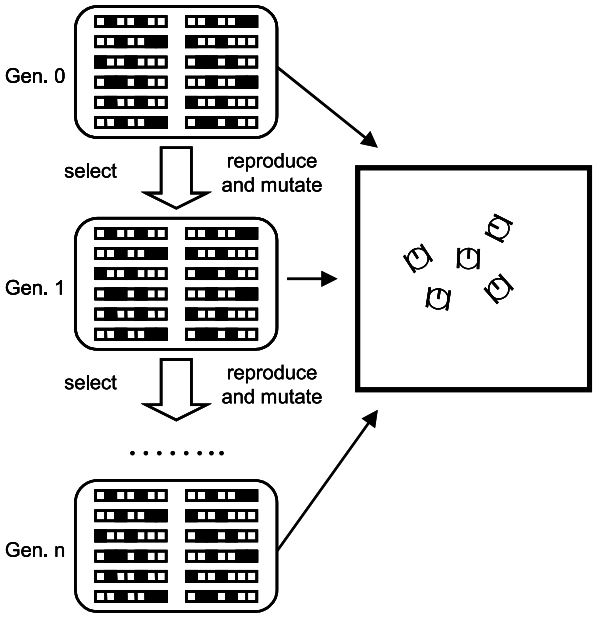

Embodied and Situated Agents, Adaptive Behavior in, Figure 9
A schematic representation of the evolutionary process. The *stripes* with *black and white squares* represent individual genotypes. The *rectangular boxes* indicate the genome of a population of a certain generation. The small robots placed inside the *square* on the *right part* of the figure represent a group of robots situated in an environment which interact with the environment and between themselves

The process that determines how a genotype (i. e. typically a string of binary values) is turned into a corresponding phenotype (i. e. a robot with a given morphology and control system) might consist of a simple one-to-one mapping or of a complex developmental process. In the former case, many of the characteristics of the phenotypical individual (e. g. the shape of the body, the number and position of the sensors and of the actuators, and in some case the architecture of the neural controller) are pre-determined and fixed and the genotype encodes a vector of free parameters (e. g. the connection weights of the neural controller [31]). In the latter case, the genotype might encode a set of rules that determine how the body structure and the control system of the individual growth during an artificial developmental processes. Through these type of indirect developmental mappings most of the characteristics of the phenotypical robot can be encoded in the genotype and subjected to the evolutionary adaptive process [31,36]. Finally, in some cases the adaptation process might involve both an evolutionary process that regulates how the characteristics of the robots vary phylogenetically (i. e. throughout generations) and a developmental/learning process which regulates how the characteristics of the robots vary ontogenetically (i. e. during the phase in which the robots act in the environment [30]).

Evolutionary methods can be used to allow agents to develop the requested behavioral and cognitive skills from scratch (i. e. starting from agents which do not have any behavioral or cognitive capability) or in an incremental manner (i. e. starting from pre-evolved robots which already have some behavioral capability which consists, for example, in the ability to solve a simplified version of the adaptive problem).

The fitness function which determines whether an individual will be reproduced or not might also include, in the addition to a component that score the performance of the agent with respect to a given task, additional task-independent components. These additional components, in fact, can lead to the development of behavioral skills which are not necessarily functional but which can favor the development of functional skills later on [37].

Evolutionary methods can allow agents to develop low-levels behavioral and cognitive skills which have been previously identified by the designer/experimenter, which might later be combined and integrated in order to realize the high-level requested skill, or directly to develop the

high-level requested skill. In the former case the adaptive process leads to the identification of the fine-grained features of the agent (e. g. number and type of sensors, body shape, architecture and connection weights of the neural controller) which by interacting between themselves and with the environment will produce the required skill. In the latter case, the adaptive process leads to the identification of the lower-levels skills (at different levels of organization) which are necessary to produce the required high-level skill, the identification of the way in which these lower levels skills should be combined and integrated, and (as for the formed case) the identification of the fine grained features of the agent which, in interaction with the physical and social environment, will produce the required behavioral or cognitive skills.

Developmental Robotics Methods

Developmental Robotics [1,10,21], also known as *epigenetic robotics*, is a method for developing embodied and situated agents that adapt to their task/environment autonomously through processes inspired by biological developmental and learning processes.

Evolutionary and developmental robotics methods share the same fundamental assumptions but also present differences for what concerns the way in which they are realized and the type of situations in which they are typically applied. For what concerns the former aspect, unlike evolutionary robotics methods which operate on 'long' phylogenetic time scales, developmental methods typically operate on 'short' ontogenetic time scales. For what concerns the latter aspects, unlike evolutionary methods which are usually used to develop behavioral and cognitive skills from scratch, developmental methods are typically adopted to model the development of complex developmental and cognitive skills from simpler pre-existing skills which represents pre-requisites for the development of the required skills.

At the present stage, developmental robotics does not consist of a well defined methodology [1,21] but rather of a collection of approaches and methods often addressing complementary aspects which hopefully would be integrated in a single methodology in the future. Below briefly summarize some of the most important methodological aspects of the developmental robotics approach.

The Incremental Nature of the Developmental Process Development should be characterized as an incremental process in which pre-existing structures and behavioral skills constitute important prerequisites and constraints for the development of more complex structures and behavioral skills and in which the complexity of the internal and external characteristics increases during developmental. One crucial aspect of developmental approach therefore consists in the identification of the initial characteristics and skills which should enable the bootstrapping of the developmental process: the layering of new skills on top of existing ones [10,25,38]. Another important aspect consists in shaping the developmental process so to ensure that the progressive increase in the complexity of the task matches the current competency of the system and so to drive the developmental process toward the progressive acquisition of the skills which represent the prerequisites for further developments. The progressive increase in complexity might concern not only the complexity of the task or of the required skills but also the complexity of single components of the robot/environmental interaction such us, for example, the number of freeze/unfreeze degrees of freedom [5].

The Social Nature of the Developmental Process Development should involve social interaction with human subjects and with other developing robots. Social interactions (e. g. scaffolding, tutelage, mimicry, emulation, and imitation), in fact, play an important role not only for the development of social skills [7] but also as facilitators for the development of individual cognitive and behavioral skills [47]. Moreover, other types of social interactions (i. e. alignment processes or social games) might lead to the development of cognitive and/or behavioral skills which are generated by a collection of individuals and which could not be developed by a single individual robot [43].

Exploitation of the Interaction Between Concurrent Developmental Processes Development should involve the exploitation of properties originating from the interaction and the integration of several co-occurring processes. Indeed, the co-development of different skills at the same time can favor the acquisition of the corresponding skills and of additional abilities arising from the combination and the integration of the developed skills. For example, the development of an ability to anticipate the sensory consequences of our own actions might facilitate the concurrent development of other skills such us categorical perception skills [46]. The development of an ability to pay attention to new situations (curiosity) and to look for new experiences after some time (boredom) might improve the learning of a given functional skill [33,42]. The co-development of behavioral and linguistic skills might favor the acquisition of the corresponding skills and the development of semantic combinatoriality skills Sugita and Tani [44].

Discussion and Conclusion

In this paper we described how artificial agents which are embodied and situated can develop behavioral and cognitive skills autonomously while they interact with their physical and social environment.

After having introduced the notion of embodiment and situatedness, we illustrated how the behavioral and cognitive skills displayed by adaptive agents can be properly characterized as complex system with multi-level and multi-scale properties resulting from a large number of interaction at different levels of organization and involving both bottom-up processes (in which the interaction between elements at lower levels of organization lead to higher levels properties) and top-down processes (in which properties at a certain level of organization later affect lower level properties or processes).

Finally, we briefly introduced the methods which can be used to synthesize adaptive embodied and situated agents.

The complex system nature of adaptive agents which are embodied and situated has important implications which constraint the organization of these systems and the dynamics of the adaptive process through which they develop their skills.

For what concerns the organization of these systems, it implies that agents' behavioral and/or cognitive skills (at any stage of the adaptive process) cannot be traced back to anyone of the three foundational elements (i. e. the body of the agents, the control system of the agents, and the environment) in isolation but should rather be characterized as properties which emerge from the interactions between these three elements and the interaction between behavioral and cognitive properties emerging from the former interactions at different levels of organizations. Moreover, it implies that 'complex' behavioral or cognitive skills might emerge from the interaction between simple properties or processes.

For what concerns agents' adaptive process, it implies that the development of new 'complex' skills does not necessarily require the development of new 'complex' morphological features or new 'complex' control mechanisms. Indeed, new 'complex' skills might arise from the addition of new 'simple' features or new 'simple' control rules which, in interaction with the pre-existing features and processes, might produce the required new behavioral or cognitive skills.

The study of adaptive behavior in artificial agents which has been reviewed in this paper has important implication both from an engineering point of view (i. e. for progressing in our ability to develop effective machines) and from a modeling point of view (i. e. for understanding the characteristics of biological organisms).

In particular, from an engineering point of view, progresses in our ability to develop adaptive embodied and situated agents can lead to development of machines playing useful functionalities.

From a modeling point of view, progresses in our ability to model and analyze artificial adaptive agents can improve our understanding of the general mechanisms behind animal and human intelligence. For example, the comprehension of the complex system nature of behavioral and cognitive skills illustrated in this paper can allow us to better define the notion of embodiment and situatedness which represent two foundational concepts in the study of natural and artificial intelligence. Indeed, although possessing a body and being in a physical environment certainly represent a pre-requisite for considering an agent embodied and situated, a more useful definition of embodiment (or of truly embodiment) can be given in term of the extent to which a given agent exploits its body characteristics to solve its adaptive problem (i. e. the extent to which its body structure is adapted to the problem to be solved, or in other words, the extent to which its body performs morphological computation). Similarly, a more useful definition of situatedness (or truly situatedness) can be given in terms of the extent to which an agent exploits its interaction with the physical and social environment and the properties originating from this interaction to solve its adaptive problem. For sake of clarity we can refer to the former definition of the terms (i. e. possessing a physical body and being situated in a physical environment) as embodiment and situatedness as weak sense, and to the latter definition as embodiment and situatedness in a strong sense.

Bibliography

1. Asada M, MacDorman K, Ishiguro H, Kuniyoshi Y (2001) Cognitive developmental robotics as a new paradigm for the design of humanoid robots. Robot Auton Syst 37:185–193
2. Baldassarre G, Parisi D, Nolfi S (2006) Distributed coordination of simulated robots based on self-organisation. Artif Life 3(12):289–311
3. Beer RD (1995) A dynamical systems perspective on agent-environment interaction. Artif Intell 72:173–215
4. Beer RD (2003) The dynamics of active categorical perception in an evolved model agent. Adapt Behav 11:209–243
5. Berthouze L, Lungarella M (2004) Motor skill acquisition under environmental perturbations: on the necessity of alternate freezing and freeing. Adapt Behav 12(1):47–63
6. Bongard JC, Paul C (2001) Making evolution an offer it can't refuse: Morphology and the extradimensional bypass. In: Keleman J, Sosik P (eds) Proceedings of the Sixth European Con-

ference on Artificial Life. Lecture Notes in Artificial Intelligence, vol 2159. Springer, Berlin
7. Breazeal C (2003) Towards sociable robots. Robotics Auton Syst 42(3–4):167–175
8. Brooks RA (1991) Intelligence without reason. In: Mylopoulos J, Reiter R (eds) Proceedings of 12th International Joint Conference on Artificial Intelligence. Morgan Kaufmann, San Mateo
9. Brooks RA (1991) Intelligence without reason. In: Proceedings of 12th International Joint Conference on Artificial Intelligence. Sydney, Australia, pp 569–595
10. Brooks RA, Breazeal C, Irie R, Kemp C, Marjanovic M, Scassellati B, Williamson M (1998) Alternate essences of intelligence. In: Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98), Madison, Wisconsin, pp 961–976
11. Chiel HJ, Beer RD (1997) The brain has a body: Adaptive behavior emerges from interactions of nervous system, body and environment. Trends Neurosci 20:553–557
12. Clark A (1997) Being there: Putting brain, body and world together again. MIT Press, Cambridge
13. Endo I, Yamasaki F, Maeno T, Kitano H (2002) A method for co-evolving morphology and walking patterns of biped humanoid robot. In: Proceedings of the IEEE Conference on Robotics and Automation, Washington, D.C.
14. Floreano D, Husband P, Nolfi S (2008) Evolutionary Robotics. In: Siciliano B, Oussama Khatib (eds) Handbook of Robotics. Springer, Berlin
15. Gigliotta O, Nolfi S (2008) On the coupling between agent internal and agent/environmental dynamics: Development of spatial representations in evolving autonomous robots. Adapt Behav 16:148–165
16. Goldenberg E, Garcowski J, Beer RD (2004) May we have your attention: Analysis of a selective attention task. In: Schaal S, Ijspeert A, Billard A, Vijayakumar S, Hallam J, Meyer J-A (eds) From Animals to Animats 8: Proceedings of the Eighth International Conference on the Simulation of Adaptive Behavior. MIT Press, Cambridge
17. Harvey I (2000) Robotics: Philosophy of mind using a screwdriver. In: Gomi T (ed) Evolutionary Robotics: From Intelligent Robots to Artificial Life, vol III. AAI Books, Ontario
18. Holland J (1975) Adaptation in natural and artificial systems. University of Michigan Press, Ann Arbor
19. Keijzer F (2001) Representation and behavior. MIT Press, London
20. Kelso JAS (1995) Dynamics patterns: The self-organization of brain and behaviour. MIT Press, Cambridge
21. Lungarella M, Metta G, Pfeifer R, Sandini G (2003) Developmental robotics: a survey. Connect Sci 15:151–190
22. Marocco D, Nolfi S (2007) Emergence of communication in embodied agents evolved for the ability to solve a collective navigation problem. Connect Sci 19(1):53–74
23. Massera G, Cangelosi A, Nolfi S (2007) Evolution of prehension ability in an anthropomorphic neurorobotic arm. Front Neurorobot 1(4):1–9
24. McGeer T (1990) Passive walking with knees. In: Proceedings of the IEEE Conference on Robotics and Automation, vol 2, pp 1640–1645
25. Metta G, Sandini G, Natale L, Panerai F (2001) Development and Q30 robotics. In: Proceedings of IEEE-RAS International Conference on Humanoid Robots, pp 33–42
26. Mondada F, Franzi E, Ienne P (1993) Mobile robot miniaturisation: A tool for investigation in control algorithms. In: Proceedings of the Third International Symposium on Experimental Robotics, Kyoto, Japan
27. Mondada F, Pettinaro G, Guigrard A, Kwee I, Floreano D, Denebourg J-L, Nolfi S, Gambardella LM, Dorigo M (2004) Swarm-bot: A new distributed robotic concept. Auton Robots 17(2–3):193–221
28. Nolfi S (2002) Power and limits of reactive agents. Neurocomputing 49:119–145
29. Nolfi S (2005) Behaviour as a complex adaptive system: On the role of self-organization in the development of individual and collective behaviour. Complexus 2(3–4):195–203
30. Nolfi S, Floreano D (1999) Learning and Evolution. Auton Robots 1:89–113
31. Nolfi S, Floreano D (2000) Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organizing Machines. MIT Press/Bradford Books, Cambridge
32. Nolfi S, Marocco D (2002) Active perception: A sensorimotor account of object categorization. In: Hallam B, Floreano D, Hallam J, Hayes G, Meyer J-A (eds) From Animals to Animats 7, Proceedings of the VII International Conference on Simulation of Adaptive Behavior. MIT Press, Cambridge, pp 266–271
33. Oudeyer P-Y, Kaplan F, Hafner V (2007) Intrinsic motivation systems for autonomous mental development. IEEE Trans Evol Comput 11(2):265–286
34. Pfeifer R, Bongard J (2007) How the body shape the way we think. MIT Press, Cambridge
35. Pfeifer R, Iida F, Gómez G (2006) Morphological computation for adaptive behavior and cognition. In: International Congress Series, vol 1291, pp 22–29
36. Pollack JB, Lipson H, Funes P, Hornby G (2001) Three generations of coevolutionary robotics. Artif Life 7:215–223
37. Prokopenko M, Gerasimov V, Tanev I (2006) Evolving spatiotemporal coordination in a modular robotic system. In: Rocha LM, Yaeger LS, Bedau MA, Floreano D, Goldstone RL, Vespignani A (eds) Artificial Life X: Proceedings of the Tenth International Conference on the Simulation and Synthesis of Living Systems. MIT Press, Boston
38. Scassellati B (2001) Foundations for a Theory of Mind for a Humanoid Robot. Ph D thesis, Department of Electrical Engineering and Computer Science, MIT, Boston
39. Scheier C, Pfeifer R, Kunyioshi Y (1998) Embedded neural networks: exploiting constraints. Neural Netw 11:1551–1596
40. Schmitz A, Gómez G, Iida F, Pfeifer R (2007) On the robustness of simple speed control for a quadruped robot. In: Proceeding of the International Conference on Morphological Computation, Venice, Italy
41. Slocum AC, Downey DC, Beer RD (2000) Further experiments in the evolution of minimally cognitive behavior: From perceiving affordances to selective attention. In: Meyer J, Berthoz A, Floreano D, Roitblat H, Wilson S (eds) From Animals to Animats 6. Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior. MIT Press, Cambridge
42. Schmidhuber J (2006) Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. Connect Sci 18(2):173–187
43. Steels L (2003) Evolving grounded communication for robots. Trends Cogn Sci 7(7):308–312
44. Sugita Y, Tani J (2005) Learning semantic combinatoriality from the interaction between linguistic and behavioral processes. Adapt Behav 13(1):33–52

45. Tani J, Fukumura N (1997) Self-organizing internal representation in learning of navigation: A physical experiment by the mobile robot Yamabico. Neural Netw 10(1):153–159
46. Tani J, Nolfi S (1999) Learning to perceive the world as articulated: An approach for hierarchical learning in sensory-motor systems. Neural Netw 12:1131–1141
47. Tani J, Nishimoto R, Namikawa J, Ito M (2008) Co-developmental learning between human and humanoid robot using a dynamic neural network model. IEEE Trans Syst Man Cybern B. Cybern 38:1
48. Varela FJ, Thompson E, Rosch E (1991) The Embodied mind: Cognitive science and human experience. MIT Press, Cambridge
49. van Gelder TJ (1998) The dynamical hypothesis in cognitive science. Behav Brain Sci 21:615–628
50. Vaughan E, Di Paolo EA, Harvey I (2004) The evolution of control and adaptation in a 3D powered passive dynamic walker. In: Pollack J, Bedau M, Husband P, Ikegami T, Watson R (eds) Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems. MIT Press, Cambridge

Entropy

Constantino Tsallis[1,2]
[1] Centro Brasileiro de Pesquisas Físicas, Rio de Janeiro, Brazil
[2] Santa Fe Institute, Santa Fe, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Some Basic Properties
Boltzmann–Gibbs Statistical Mechanics
On the Limitations of Boltzmann–Gibbs Entropy and Statistical Mechanics
The Nonadditive Entropy S_q
A Connection Between Entropy and Diffusion
Standard and q-Generalized Central Limit Theorems
Future Directions
Acknowledgments
Bibliography

Glossary

Absolute temperature Denoted T.

Clausius entropy Also called *thermodynamic entropy*. Denoted S.

Boltzmann–Gibbs entropy Basis of Boltzmann–Gibbs statistical mechanics. This entropy, denoted S_{BG}, is *additive*. Indeed, for two probabilistically independent subsystems A and B, it satisfies $S_{\mathrm{BG}}(A+B) = S_{\mathrm{BG}}(A) + S_{\mathrm{BG}}(B)$.

Nonadditive entropy It usually refers to the basis of nonextensive statistical mechanics. This entropy, denoted S_q, is *nonadditive* for $q \neq 1$. Indeed, for two probabilistically independent subsystems A and B, it satisfies $S_q(A+B) \neq S_q(A) + S_q(B)$ $(q \neq 1)$. For historical reasons, it is frequently (but inadequately) referred to as *nonextensive entropy*.

q-logarithmic and q-exponential functions Denoted $\ln_q x$ $(\ln_1 x = \ln x)$, and e_q^x $(\mathrm{e}_1^x = \mathrm{e}^x)$, respectively.

Extensive system So called for historical reasons. A more appropriate name would be *additive system*. It is a system which, in one way or another, relies on or is connected to the (additive) Boltzmann–Gibbs entropy. Its basic dynamical and/or structural quantities are expected to be of the exponential form. In the sense of complexity, it may be considered a *simple system*.

Nonextensive system So called for historical reasons. A more appropriate name would be *nonadditive system*. It is a system which, in one way or another, relies on or is connected to a (nonadditive) entropy such as $S_q(q \neq 1)$. Its basic dynamical and/or structural quantities are expected to asymptotically be of the power-law form. In the sense of complexity, it may be considered a *complex system*.

Definition of the Subject

Thermodynamics and statistical mechanics are among the most important formalisms in contemporary physics. They have overwhelming and intertwined applications in science and technology. They essentially rely on two basic concepts, namely *energy* and *entropy*. The mathematical expression that is used for the first one is well known to be *nonuniversal*; indeed, it depends on whether we are say in classical, quantum, or relativistic regimes. The second concept, and very specifically its connection with the microscopic world, has been considered during well over one century as essentially unique and *universal* as a physical concept. Although some mathematical generalizations of the entropy have been proposed during the last forty years, they have frequently been considered as mere practical expressions for disciplines such as cybernetics and control theory, with no particular physical interpretation. What we have witnessed during the last two decades is the growth, among physicists, of the belief that it is not necessarily so. In other words, the physical entropy would basically rely on the microscopic dynamical and structural properties of the system under study. For example, for systems microscopically evolving with strongly chaotic dynamics, the connection between the thermodynamical entropy and the thermostatistical entropy would be the one

found in standard textbooks. But, for more complex systems (e. g., for weakly chaotic dynamics), it becomes either necessary, or convenient, or both, to extend the traditional connection. The present article presents the ubiquitous concept of entropy, useful even for systems for which no energy can be defined at all, within a standpoint reflecting a *nonuniversal* conception for the connection between the thermodynamic and the thermostatistical entropies. Consequently, both the standard entropy and its recent generalizations, as well as the corresponding statistical mechanics, are here presented on equal footing.

Introduction

The concept of *entropy* (from the Greek $\epsilon\nu\ \tau\rho\epsilon\pi\omega$, *en trepo*, at turn, at transformation) was first introduced in 1865 by the German physicist and mathematician Rudolf Julius Emanuel Clausius, Rudolf Julius Emanuel in order to mathematically complete the formalism of classical thermodynamics [55], one of the most important theoretical achievements of contemporary physics. The term was so coined to make a parallel to *energy* (from the Greek $\epsilon\nu\epsilon\rho\gamma o\varsigma$, *energos*, at work), the other fundamental concept of thermodynamics. Clausius connection was given by

$$\mathrm{d}S = \frac{\delta Q}{T}\,, \tag{1}$$

where δQ denotes an infinitesimal transfer of heat. In other words, $1/T$ acts as an integrating factor for δQ. In fact, it was only in 1909 that thermodynamics was finally given, by the Greek mathematician Constantin Caratheodory, a logically consistent axiomatic formulation.

In 1872, some years after Clausius proposal, the Austrian physicist Ludwig Eduard Boltzmann introduced a quantity, that he noted H, which was defined in terms of microscopic quantities:

$$H \equiv \iiint f(\mathbf{v})\,\ln[f(\mathbf{v})]\,\mathrm{d}\mathbf{v}\,, \tag{2}$$

where $f(\mathbf{v})\,\mathrm{d}\mathbf{v}$ is the number of molecules in the velocity space interval $\mathrm{d}\mathbf{v}$. Using Newtonian mechanics, Boltzmann showed that, under some intuitive assumptions (*Stoßzahlansatz* or *molecular chaos hypothesis*) regarding the nature of molecular collisions, H does not increase with time. Five years later, in 1877, he identified this quantity with Clausius entropy through $-kH \equiv S$, where k is a constant. In other words, he established that

$$S = -k \iiint f(\mathbf{v})\,\ln[f(\mathbf{v})]\,\mathrm{d}\mathbf{v}\,, \tag{3}$$

later on generalized into

$$S = -k \iint f(\mathbf{q},\mathbf{p})\,\ln[f(\mathbf{q},\mathbf{p})]\,\mathrm{d}\mathbf{q}\,\mathrm{d}\mathbf{p}\,, \tag{4}$$

where $(\mathbf{q},\mathbf{p})$ is called the *μ-space* and constitutes the phase space (coordinate $\mathbf{q}$ and momentum $\mathbf{p}$) corresponding to *one* particle.

Boltzmann's genius insight – the first ever mathematical connection of the macroscopic world with the microscopic one – was, during well over three decades, highly controversial since it was based on the hypothesis of the existence of atoms. Only a few selected scientists, like the English chemist and physicist John Dalton, the Scottish physicist and mathematician James Clerk Maxwell, and the American physicist, chemist and mathematician Josiah Willard Gibbs, believed in the reality of atoms and molecules. A large part of the scientific establishment was, at the time, strongly against such an idea. The intricate evolution of Boltzmann's lifelong epistemological struggle, which ended tragically with his suicide in 1906, may be considered as a neat illustration of Thomas Kuhn's *paradigm shift*, and the corresponding reaction of the scientific community, as described in *The Structure of Scientific Revolutions*. There are in fact two important formalisms in contemporary physics where the mathematical theory of probabilities enters as a central ingredient. These are statistical mechanics (with the concept of entropy as a functional of probability distributions) and quantum mechanics (with the physical interpretation of wave functions and measurements). In both cases, contrasting viewpoints and passionate debates have taken place along more than one century, and continue still today. This is no surprise after all. If it is undeniable that energy is a very deep and subtle concept, entropy is even more. Indeed, energy concerns the world of (microscopic) *possibilities*, whereas entropy concerns the world of the *probabilities* of those possibilities, a step further in epistemological difficulty.

In his 1902 celebrated book *Elementary Principles of Statistical Mechanics*, Gibbs introduced the modern form of the entropy for classical systems, namely

$$S = -k \int \mathrm{d}\Gamma\, f(\mathbf{q},\mathbf{p})\,\ln[C f(\mathbf{q},\mathbf{p})]\,, \tag{5}$$

where Γ represents the *full phase space* of the system, thus containing all coordinates and all momenta of its elementary particles, and C is introduced to take into account the finite size and the physical dimensions of the smallest admissible cell in Γ-space. The constant k is known today to be a universal one, called *Boltzmann constant*, and given by $k = 1.3806505(24) \times 10^{-23}$ Joule/Kelvin. The studies

of the German physicist Max Planck along Boltzmann and Gibbs lines after the appearance of quantum mechanical concepts, eventually led to the expression

$$S = k \ln W \,, \tag{6}$$

which he coined as *Boltzmann entropy*. This expression is carved on the stone of Boltzmann's grave at the Central Cemetery of Vienna. The quantity W is the total number of microstates of the system that are compatible with our macroscopic knowledge of it. It is obtained from Eq. (5) under the hypothesis of an uniform distribution or equal probabilities.

The Hungarian–American mathematician and physicist Johann von Neumann extended the concept of BG entropy in two steps – in 1927 and 1932 respectively –, in order to also cover quantum systems. The following expression, frequently referred to as the *von Neumann entropy*, resulted:

$$S = -k \,\mathrm{Tr}\, \rho \ln \rho \,, \tag{7}$$

ρ being the *density operator* (with $\mathrm{Tr}\, \rho = 1$).

Another important step was given in 1948 by the American electrical engineer and mathematician Claude Elwood Shannon. Having in mind the theory of digital communications he explored the properties of the discrete form

$$S = -k \sum_{i=1}^{W} p_i \ln p_i \,, \tag{8}$$

frequently referred to as *Shannon entropy* (with $\sum_{i=1}^{W} p_i = 1$). This form can be recovered from Eq. (5) for the particular case for which the phase space density $f(\mathbf{q}, \mathbf{p}) = \sum_{i=1}^{W} p_i \, \delta(\mathbf{q} - \mathbf{q_i}) \, \delta(\mathbf{p} - \mathbf{p_i})$. It can also be recovered from Eq. (7) when ρ is diagonal. We may generically refer to Eqs. (5), (6), (7) and (8) as the *BG entropy*, noted S_{BG}. It is a measure of the disorder of the system or, equivalently, of our degree of ignorance or lack of information about its state. To illustrate a variety of properties, the discrete form (8) is particularly convenient.

Some Basic Properties

Non-negativity It can be easily verified that, in all cases, $S_{\mathrm{BG}} \geq 0$, the zero value corresponding to *certainty*, i. e., $p_i = 1$ for one of the W possibilities, and zero for all the others. To be more precise, it is exactly so whenever S_{BG} is expressed either in the form (7) or in the form (8). However, this property of non-negativity may be no longer true if it is expressed in the form (5). This violation is one of the mathematical manifestations that, at the microscopic level, the state of any physical system exhibits its quantum nature.

Expansibility Also $S_{\mathrm{BG}}(p_1, p_2, \ldots, p_W, 0) = S_{\mathrm{BG}}(p_1, p_2, \ldots, p_W)$, i. e., zero-probability events do not modify our information about the system.

Maximal value S_{BG} is maximized at equal probabilities, i. e., for $p_i = 1/W \,, \forall i$. Its value is that of Eq. (6). This corresponds to the Laplace *principle of indifference* or *principle of insufficient reason*.

Concavity If we have two arbitrary probability distributions $\{p_i\}$ and $\{p'_i\}$ for the same set of W possibilities, we can define the *intermediate* probability distribution $p''_i = \mu \, p_i + (1 - \mu) \, p'_i \quad (0 < \mu < 1)$. It straightforwardly follows that $S_{\mathrm{BG}}(\{p''_i\}) \geq \mu \, S_{\mathrm{BG}}(\{p_i\}) + (1 - \mu) \, S_{\mathrm{BG}}(\{p'_i\})$. This property is essential for thermodynamics since it eventually leads to *thermodynamic stability*, i. e., to *robustness* with regard to energy fluctuations. It also leads to the tendency of the entropy to attain, as time evolves, its maximal value compatible with our macroscopic knowledge of the system, i. e., with the possibly known values for the macroscopic constraints.

Lesche stability or experimental robustness B. Lesche introduced in 1982 [107] the definition of an interesting property, which he called *stability*. It reflects the *experimental robustness* that a physical quantity is expected to exhibit. In other words, similar experiments should yield similar numerical results for the physical quantities. Let us consider two probability distributions $\{p_i\}$ and $\{p'_i\}$, assumed to be close, in the sense that $\sum_{i=1}^{W} |p_i - p'_i| < \delta \,, \delta > 0$ being a small number. An entropic functional $S(\{p_i\})$ is said *stable* or *experimentally robust* if, for any given $\epsilon > 0$, a $\delta > 0$ exists such that $|S(\{p_i\}) - S(\{p'_i\})|/S_{\max} < \epsilon$, where $S_{\max}$ is the maximal value that the functional can attain ($\ln W$ in the case of S_{BG}). This implies that $\lim_{\delta \to 0} \lim_{W \to \infty} (S(\{p_i\}) - S(\{p'_i\}))/S_{\max} = 0$. As we shall see soon, this property is much stronger than it seems at first sight. Indeed, it provides a (necessary but not sufficient) criterion for classifying entropic functionals into physically admissible or not. It can be shown that S_{BG} is *Lesche-stable* (or *experimentally robust*).

Entropy production If we start the (deterministic) time evolution of a generic classical system from an arbitrarily chosen point in its Γ phase space, it typically follows a quite erratic trajectory which, in many cases, gradually visits the entire (or almost) phase space. By making partitions of this Γ-space, and counting the frequency of visits to the various cells (and related symbolic quantities), it is possible to define probabil-

ity sets. Through them, we can calculate a sort of time evolution of $S_{BG}(t)$. If the system is *chaotic* (sometimes called *strongly chaotic*), i. e., if its sensitivity to the initial conditions increases exponentially with time, then $S_{BG}(t)$ increases *linearly* with t in the appropriate asymptotic limits. This rate of increase of the entropy is called *Kolmogorov–Sinai entropy rate*, and, for a large class of systems, it coincides (*Pesin identity* or *Pesin theorem*) with the sum of the positive Lyapunov exponents. These exponents characterize the exponential divergences, along various directions in the Γ-space, of a small discrepancy in the initial condition of a trajectory.

It turns out, however, that the Kolmogorov–Sinai entropy rate is, in general, quite inconvenient for computational calculations for arbitrary nonlinear dynamical systems. In practice, another quantity is used instead [102], usually referred to as *entropy production per unit time*, which we note K_{BG}. Its definition is as follows. We first make a partition of the Γ-space into many W cells ($i = 1, 2, \ldots, W$). In one of them, arbitrarily chosen, we randomly place M initial conditions (i. e., an *ensemble*). As time evolves, the occupancy of the W cells determines the set $\{M_i(t)\}$, with $\sum_{i=1}^{W} M_i(t) = M$. This set enables the definition of a probability set with $p_i(t) \equiv M_i(t)/M$, which in turn determines $S_{BG}(t)$. We then define the *entropy production per unit time* as follows:

$$K_{BG} \equiv \lim_{t\to\infty} \lim_{W\to\infty} \lim_{M\to\infty} \frac{S_{BG}(t)}{t} . \tag{9}$$

Up to date, no theorem guarantees that this quantity coincides with the Kolmogorov–Sinai entropy rate. However, many numerical approaches of various chaotic systems strongly suggest so. The same turns out to occur with what is frequently referred in the literature as a Pesin-like identity. For instance, if we have a one-dimensional dynamical system, its sensitivity to the initial conditions $\xi \equiv \lim_{\Delta x(0)\to 0} \Delta x(t)/\Delta x(0)$ is typically given by

$$\xi(t) = e^{\lambda t} , \tag{10}$$

where $\Delta x(t)$ is the discrepancy in the one-dimensional phase space of two trajectories initially differing by $\Delta x(0)$, and λ is the Lyapunov exponent ($\lambda > 0$ corresponds to *strongly sensitive to the initial conditions*, or *strongly chaotic*, and $\lambda < 0$ corresponds to *strongly insensitive to the initial conditions*). The so-called Pesin-like identity amounts, if $\lambda \geq 0$, to

$$K_{BG} = \lambda . \tag{11}$$

Additivity and extensivity If we consider a system $A + B$ constituted by two *probabilistically independent* subsystems A and B, i. e., if we consider $p_{ij}^{A+B} = p_i^A p_j^B$, we immediately obtain from Eq. (8) that

$$S_{BG}(A + B) = S_{BG}(A) + S_{BG}(B) . \tag{12}$$

In other words, the BG entropy is *additive* [130]. If our system is constituted by N probabilistically independent identical subsystems (or *elements*), we clearly have $S_{BG}(N) \propto N$. It frequently happens, however, that the N elements are not exactly independent but only asymptotically so in the $N \to \infty$ limit. This is the usual case of many-body Hamiltonian systems involving only *short-range* interactions, where the concept of short-range will be focused in detail later on. For such systems, S_{BG} is only asymptotically additive, i. e.,

$$0 < \lim_{N\to\infty} \frac{S_{BG}(N)}{N} < \infty . \tag{13}$$

An entropy $S(\{p_i\})$ of a specific systems is said *extensive* if it satisfies

$$0 < \lim_{N\to\infty} \frac{S(N)}{N} < \infty , \tag{14}$$

where *no hypothesis at all* is made about the possible independence or weak or strong correlations between the elements of the system whose entropy S we are considering. Equation (13) amounts to say that the additive entropy S_{BG} is extensive for weakly correlated systems such as the already mentioned many-body short-range-interacting Hamiltonian ones. It is important to clearly realize that *additivity* and *extensivity* are independent properties. An additive entropy such as S_{BG} is extensive for simple systems such as the ones just mentioned, but it turns out to be nonextensive for other, more complex, systems that will be focused on later on. For many of these more complex systems, it is the nonadditive entropy S_q (to be analyzed later on) which turns out to be extensive for a non standard value of q (i. e., $q \neq 1$).

Boltzmann–Gibbs Statistical Mechanics

Physical systems (classical, quantum, relativistic) can be theoretically described in very many ways, through microscopic, mesoscopic, macroscopic equations, reflecting either stochastic or deterministic time evolutions, or even both types simultaneously. Those systems whose time evolution is completely determined by a well defined Hamiltonian with appropriate boundary conditions and ad-

missible initial conditions are the main purpose of an important branch of contemporary physics, named *statistical mechanics*. This remarkable *theory* (or *formalism*, as sometimes called), which for large systems satisfactorily matches classical thermodynamics, was primarily introduced by Boltzmann and Gibbs. The physical system can be in all types of situations. Two paradigmatic such situations correspond to *isolation*, and *thermal contact* with a large reservoir called *thermostat*. Their *stationary state* ($t \to \infty$) is usually referred to as *thermal equilibrium*. Both situations have been formally considered by Gibbs within his mathematical formulation of statistical mechanics, and they respectively correspond to the so-called *micro-canonical* and *canonical* ensembles (other ensembles do exist, such as the *grand-canonical* ensemble, appropriate for those situations in which the total number of elements of the system is not fixed; this is however out of the scope of the present article).

The stationary state of the micro-canonical ensemble is determined by $p_i = 1/W$ ($\forall i$, where i runs over all possible microscopic states), which corresponds to the extremization of S_{BG} with a single (and trivial) constraint, namely

$$\sum_{i=1}^{W} p_i = 1 \,. \tag{15}$$

To obtain the stationary state for the canonical ensemble, the thermostat being at temperature T, we must (typically) add one more constraint, namely

$$\sum_{i=1}^{W} p_i E_i = U \,, \tag{16}$$

where $\{E_i\}$ are the energies of all the possible states of the system (i. e., eigenvalues of the Hamiltonian with the appropriate boundary conditions). The extremization of S_{BG} with the two constraints above straightforwardly yields

$$p_i = \frac{\mathrm{e}^{-\beta E_i}}{Z} \tag{17}$$

$$Z \equiv \sum_{j=1}^{W} \mathrm{e}^{-\beta E_j} \tag{18}$$

with the *partition function* Z, and the Lagrange parameter $\beta = 1/kT$. This is the celebrated *BG distribution* for thermal equilibrium (or *Boltzmann weight*, or *Gibbs state*, as also called), which has been at the basis of an enormous amount of successes (in fluids, magnets, superconductors, superfluids, Bose–Einstein condensation, conductors, chemical reactions, percolation, among many other important situations). The connection with classical thermodynamics, and its Legendre-transform structure, occurs through relations such as

$$\frac{1}{T} = \frac{\partial S}{\partial U} \tag{19}$$

$$F \equiv U - TS = -\frac{1}{\beta} \ln Z \tag{20}$$

$$U = -\frac{\partial}{\partial \beta} \ln Z \tag{21}$$

$$C \equiv T \frac{\partial S}{\partial T} = \frac{\partial U}{\partial T} = -T \frac{\partial^2 F}{\partial T^2} \,, \tag{22}$$

where F, U and C are the *Helmholtz free energy*, the *internal energy*, and the *specific heat* respectively. The BG statistical mechanics historically appeared as the first connection between the microscopic and the macroscopic descriptions of the world, and it constitutes one of the cornerstones of contemporary physics. The Establishment resisted heavily before accepting the validity and power of Boltzmann' s revolutionary ideas. In 1906 Boltzmann dramatically committed suicide, after 34 years that he had first proposed the deep ideas that we are summarizing here. At that early 20th century, few people believed in Boltzmann's proposal (among those few, we must certainly mention Albert Einstein), and most physicists were simply unaware of the existence of Gibbs and of his profound contributions. It was only half a dozen years later that the emerging new generation of physicists recognized their respective genius (thanks in part to various clarifications produced by Paul Ehrenfest, and also to the experimental successes related with Brownian motion, photoelectric effect, specific heat of solids, and black-body radiation).

On the Limitations of Boltzmann–Gibbs Entropy and Statistical Mechanics

Historical Background

As any other human intellectual construct, the applicability of the BG entropy, and of the statistical mechanics to which it is associated, naturally has restrictions. The understanding of present developments of both the concept of entropy, and its corresponding statistical mechanics, demand some knowledge of the historical background.

Boltzmann was aware of the relevance of the range of the microscopic interactions between atoms and molecules. He wrote, in his 1896 *Lectures on Gas Theory* [41], the following words:

> *When the distance at which two gas molecules interact with each other noticeably is vanishingly small relative to the average distance between a molecule*

and its nearest neighbor—or, as one can also say, when the space occupied by the molecules (or their spheres of action) is negligible compared to the space filled by the gas—then the fraction of the path of each molecule during which it is affected by its interaction with other molecules is vanishingly small compared to the fraction that is rectilinear, or simply determined by external forces. […] The gas is "ideal" in all these cases.

Also Gibbs was aware. In his 1902 book [88], he wrote:

In treating of the canonical distribution, we shall always suppose the multiple integral in equation (92) [the partition function, as we call it nowadays] *to have a finite value, as otherwise the coefficient of probability vanishes, and the law of distribution becomes illusory. This will exclude certain cases, but not such apparently, as will affect the value of our results with respect to their bearing on thermodynamics. It will exclude, for instance, cases in which the system or parts of it can be distributed in unlimited space […]. It also excludes many cases in which the energy can decrease without limit, as when the system contains material points which attract one another inversely as the squares of their distances. […]. For the purposes of a general discussion, it is sufficient to call attention to the assumption implicitly involved in the formula (92).*

The extensivity/additivity of S_{BG} has been challenged, along the last century, by many physicists. Let us mention just a few. In his 1936 *Thermodynamics* [82], Enrico Fermi wrote:

The entropy of a system composed of several parts is very often equal to the sum of the entropies of all the parts. This is true if the energy of the system is the sum of the energies of all the parts and if the work performed by the system during a transformation is equal to the sum of the amounts of work performed by all the parts. Notice that these conditions are not quite obvious and that in some cases they may not be fulfilled. Thus, for example, in the case of a system composed of two homogeneous substances, it will be possible to express the energy as the sum of the energies of the two substances only if we can neglect the surface energy of the two substances where they are in contact. The surface energy can generally be neglected only if the two substances are not very finely subdivided; otherwise, it can play a considerable role.

Laszlo Tisza wrote, in his *Generalized Thermodynamics* [178]:

The situation is different for the additivity postulate P a2, the validity of which cannot be inferred from general principles. We have to require that the interaction energy between thermodynamic systems be negligible. This assumption is closely related to the homogeneity postulate P d1. From the molecular point of view, additivity and homogeneity can be expected to be reasonable approximations for systems containing many particles, provided that the intermolecular forces have a short range character.

Corroborating the above, virtually all textbooks of quantum mechanics contain the mechanical calculations corresponding to a particle in a square well, the harmonic oscillator, the rigid rotator, a spin 1/2 in the presence of a magnetic field, and the Hydrogen atom. In the textbooks of statistical mechanics we can find the thermostatistical calculations of all these systems … excepting the Hydrogen atom! Why? Because the long-range electron-proton interaction produces an energy spectrum which leads to a *divergent* partition function. This is but a neat illustration of the above Gibbs' alert.

A Remark on the Thermodynamics of Short- and Long-Range Interacting Systems

We consider here a simple d-dimensional classical fluid, constituted by many N point particles, governed by the Hamiltonian

$$\mathcal{H} = K + V = \sum_{i=1}^{N} \frac{p_i^2}{2m} + \sum_{i \neq j} V(r_{ij}) \,, \tag{23}$$

where the potential $V(r)$ has, if it is attractive at short distances, no singularity at the origin, or an integrable singularity, and whose asymptotic behavior at infinity is given by $V(r) \sim -B/r^{\alpha}$ with $B > 0$ and $\alpha \geq 0$. One such example is the $d = 3$ Lennard–Jones fluid, for which $V(r) = A/r^{12} - B/r^{6} (A > 0)$, i. e., repulsive at short distances and attractive at long distances. In this case $\alpha = 6$. Another example could be Newtonian gravitation with a phenomenological short-distance cutoff (i. e., $V(r) \to \infty$ for $r \leq r_0$ with $r_0 > 0$. In this case, $\alpha = 1$. The full Γ-space of such a system has $2dN$ dimensions. The total potential energy is expected to scale (assuming a roughly homogeneous distribution of the particles) as

$$\frac{U_{\mathrm{pot}}(N)}{N} \propto -B \int_1^{\infty} \mathrm{d}r \, r^{d-1} \, r^{-\alpha} \,, \tag{24}$$

where the integral starts appreciably contributing above a typical cutoff, here taken to be unity. This integral is finite

$[\,=-B/(\alpha-d)\,]$ for $\alpha/d>1$ (*short-range interactions*), and diverges for $0\le\alpha/d\le 1$ (*long-range interactions*). In other words, the energy cannot be *generically* characterized by Eq. (24), and we must turn onto a different and more powerful estimation. Given the *finiteness* of the size of the system, an appropriate one is, *in all cases*, given by

$$\frac{U_{\rm pot}(N)}{N}\propto -B\int_1^{N^{1/d}}{\rm d}r\,r^{d-1}\,r^{-\alpha}=-\frac{B}{d}N^\star\,, \qquad (25)$$

where

$$N^\star\equiv\frac{N^{1-\alpha/d}-1}{1-\alpha/d}\sim\begin{cases}\dfrac{1}{\alpha/d-1} & \text{if}\quad \alpha/d>1\,;\\ \ln N & \text{if}\quad \alpha/d=1\,;\\ \dfrac{N^{1-\alpha/d}}{1-\alpha/d} & \text{if}\quad 0<\alpha/d<1\,.\end{cases} \qquad (26)$$

Notice that $N^\star=\ln_{\alpha/d}N$ where the q-log function $\ln_q x\equiv(x^{1-q}-1)/(1-q)(x>0;\ln_1 x=\ln x)$ will be shown to play an important role later on. Satisfactorily enough, Eqs. (26) recover the characterization with Eq. (24) in the limit $N\to\infty$, but they have the great advantage of providing, for *finite* N, a *finite* value. This fact will be now shown to enable to properly scale the macroscopic quantities in the thermodynamic limit ($N\to\infty$), *for all values of* $\alpha/d\ge 0$.

Let us address the thermodynamical consequences of the microscopic interactions being short- or long-ranged. To present a slightly more general illustration, we shall assume from now on that our homogeneous and isotropic classical fluid is made by magnetic particles. Its Gibbs free energy is then given by

$$\begin{aligned}G(N,T,p,H)&=U(N,T,p,H)-TS(N,T,p,H)\\&\quad+pV(N,T,p,H)-HM(N,T,p,H)\,,\end{aligned} \qquad (27)$$

where (T,p,H) correspond respectively to the temperature, pressure and external magnetic field, V is the volume and M the magnetization. If the interactions are short-ranged (i. e., if $\alpha/d>1$), we can divide this equation by N and then take the $N\to\infty$ limit. We obtain

$$\begin{aligned}g(T,p,H)&=u(T,p,H)-Ts(T,p,H)\\&\quad+pv(T,p,H)-Hm(T,p,H)\,,\end{aligned} \qquad (28)$$

where $g(T,p,H)\equiv\lim_{N\to\infty}G(N,T,p,H)/N$, and analogously for the other variables of the equation. If the interactions were instead long-ranged (i. e., if $0\le\alpha/d\le 1$), all these quantities would be *divergent*, hence thermodynamically nonsense. Consequently, the generically correct procedure, i. e. $\forall\,\alpha/d\ge 0$, must conform to the following lines:

$$\begin{aligned}\lim_{N\to\infty}\frac{G(N,T,p,H)}{NN^\star}&=\lim_{N\to\infty}\frac{U(N,T,p,H)}{NN^\star}\\&\quad-\lim_{N\to\infty}\frac{T}{N^\star}\frac{S(N,T,p,H)}{N}\\&\quad+\lim_{N\to\infty}\frac{p}{N^\star}\frac{V(N,T,p,H)}{N}\\&\quad-\lim_{N\to\infty}\frac{H}{N^\star}\frac{M(N,T,p,H)}{N}\end{aligned} \qquad (29)$$

hence

$$\begin{aligned}g(T^\star,p^\star,H^\star)&=u(T^\star,p^\star,H^\star)-T^\star s(T^\star,p^\star,H^\star)\\&\quad+p^\star v(T^\star,p^\star,H^\star)-H^\star m(T^\star,p^\star,H^\star)\,,\end{aligned} \qquad (30)$$

where the definitions of $T^\star$ and all the other variables are self-explanatory (e. g., $T^\star\equiv T/N^\star$). In other words, in order to have *finite* thermodynamic equations of states, we must in general express them in the $(T^\star,p^\star,H^\star)$ variables. If $\alpha/d>1$, this procedure recovers the usual equations of states, and the usual *extensive* (G,U,S,V,M) and *intensive* (T,p,H) thermodynamic variables. But, if $0\le\alpha/d\le 1$, the situation is more complex, and we realize that *three*, instead of the traditional *two*, classes of thermodynamic variables emerge. We may call them *extensive* (S,V,M,N), *pseudo-extensive* (G,U) and *pseudo-intensive* (T,p,H) variables. All the energy-type thermodynamical variables (G,F,U) give rise to pseudo-extensive ones, whereas those which appear in the usual Legendre thermodynamical pairs give rise to pseudo-intensive ones (T,p,H,μ) and extensive ones (S,V,M,N). See Figs. 1 and 2.

The possibly long-range interactions within Hamiltonian (23) refer to the *dynamical variables themselves*. There is another important class of Hamiltonians, where the possibly long-range interactions refer to the *coupling constants between localized dynamical variables*. Such is, for instance, the case of the following classical Hamiltonian:

$$\begin{aligned}\mathcal{H}&=K+V=\sum_{i=1}^{N}\frac{L_i^2}{2I}\\&\quad-\sum_{i\neq j}\frac{J_x s_i^x s_j^x+J_y s_i^y s_j^y+J_z s_i^z s_j^z}{r_{ij}^\alpha}\qquad(\alpha\ge 0)\,,\end{aligned} \qquad (31)$$

where $\{L_i\}$ are the angular momenta, I the moment of inertia, $\{(s_i^x,s_i^y,s_i^z)\}$ are the components of classical rotators, (J_x,J_y,J_z) are coupling constants, and r_{ij} runs

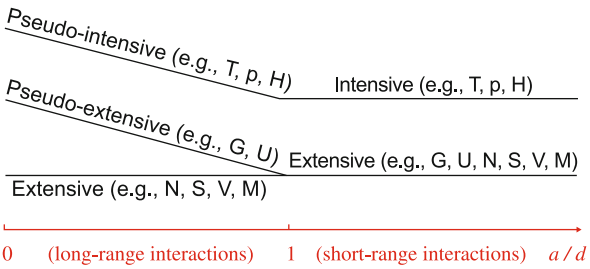

Entropy, Figure 1
For long-range interactions ($0 \leq \alpha/d \leq 1$) we have *three* classes of *thermodynamic variables*, namely the *pseudo-intensive* (scaling with $N^\star$), *pseudo-extensive* (scaling with $NN^\star$) and *extensive* (scaling with N) ones. For short range interactions ($\alpha/d > 1$) the pseudo-intensive variables become *intensive* (independent from N), and the pseudo-extensive merge with the extensive ones, all being now *extensive* (scaling with N), thus recovering the traditional *two* textbook classes of thermodynamical variables

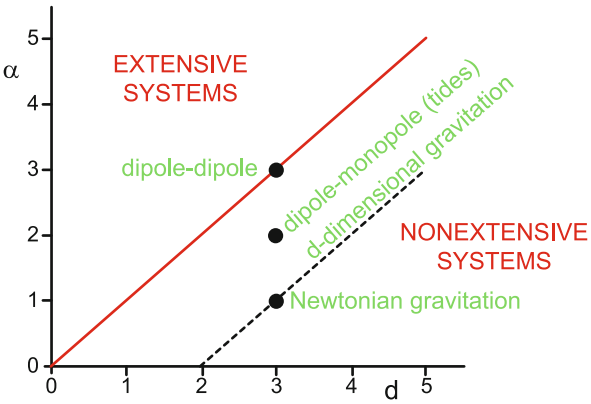

Entropy, Figure 2
The so-called *extensive systems* ($\alpha/d > 1$ for the classical ones) typically involve *absolutely convergent series*, whereas the so-called *nonextensive systems* ($0 \leq \alpha/d < 1$ for the classical ones) typically involve *divergent series*. The marginal systems ($\alpha/d = 1$ here) typically involve *conditionally convergent series*, which therefore depend on the boundary conditions, i. e., typically on the external shape of the system. Capacitors constitute a notorious example of the $\alpha/d = 1$ case. The model usually referred to in the literature as the *Hamiltonian–Mean–Field* (HMF) one lies on the $\alpha = 0$ axis ($\forall d > 0$). The model usually referred to as the d-dimensional α-XY model [19] lies on the vertical axis at abscissa d ($\forall \alpha \geq 0$)

over all distances between sites i and j of a d-dimensional lattice. For example, for a simple hypercubic lattice with unit crystalline parameter we have $r_{ij} = 1, 2, 3, \ldots$ if $d = 1$, $r_{ij} = 1, \sqrt{2}, 2, \ldots$ if $d = 2$, $r_{ij} = 1, \sqrt{2}, \sqrt{3}, 2, \ldots$ if $d = 3$, and so on. For such a case, we have that

$$N^\star \equiv \sum_{i=2}^{N} r_{1i}^{-\alpha} , \tag{32}$$

which has in fact the same asymptotic behaviors as indicated in Eq. (26). In other words, here again $\alpha/d > 1$ corresponds to short-range interactions, and $0 \leq \alpha/d \leq 1$ corresponds to long-range ones.

The correctness of the present generalized thermodynamical scalings has already been specifically checked in many physical systems, such as a ferrofluid-like model [97], Lennard–Jones-like fluids [90], magnetic systems [16,19,59,158], anomalous diffusion [66], percolation [85,144].

Let us mention that, for the $\alpha = 0$ models (i. e., mean field models), it is largely spread in the literature to divide by N the potential term of the Hamiltonian in order to make it extensive *by force*. Although mathematically admissible (see [19]), this is obviously very unsatisfactory in principle since it implies a microscopic coupling constant which depends on N. What we have described here is the thermodynamically proper way of eliminating the mathematical difficulties emerging in the models in the presence of long-range interactions.

Last but not least, we verify a point which is crucial for the developments here below, namely that the entropy S is expected to be extensive *no matter the range of the interactions*.

The Nonadditive Entropy S_q

Introduction and Basic Properties

The possibility was introduced in 1988 [183] (see also [42,112,157,182]) to generalize the BG statistical mechanics on the basis of an entropy S_q which generalizes S_{BG}. This entropy is defined as follows:

$$S_q \equiv k \frac{1 - \sum_{i=1}^{W} p_i^q}{q-1} \quad (q \in \mathcal{R};\ S_1 = S_{BG}). \tag{33}$$

For equal probabilities, this entropy takes the form

$$S_q = k \ln_q W \quad (S_1 = k \ln W) , \tag{34}$$

where the q-logarithmic function has already been defined.

Remark With the same or different prefactor, this entropic form has been successively and independently introduced in many occasions during the last decades. J. Havrda and F. Charvat [92] were apparently the first to ever introduce this form, though with a different prefactor (adapted to binary variables) in the context of cybernetics and information theory. I. Vajda [207], further studied this form, quoting Havrda and Charvat. Z. Daroczy [74] rediscovered this form (he quotes neither Havrda–Charvat

nor Vajda). J. Lindhard and V. Nielsen [108] rediscovered this form (they quote none of the predecessors) through the property of entropic composability. B.D. Sharma and D.P. Mittal [163] introduced a two-parameter form which reproduces both S_q and Renyi entropy [145] as particular cases. A. Wehrl [209] mentions the form of S_q in p. 247, quotes Daroczy, but ignores Havrda–Charvat, Vajda, Lindhard–Nielsen, and Sharma–Mittal. Myself I rediscovered this form in 1985 with the aim of generalizing Boltzmann–Gibbs statistical mechanics, but quote none of the predecessors in the 1988 paper [183]. In fact, I started knowing the whole story quite a few years later thanks to S.R.A. Salinas and R.N. Silver, who were the first to provide me with the corresponding informations. Such rediscoveries can by no means be considered as particularly surprising. Indeed, this happens in science more frequently than usually realized. This point is lengthily and colorfully developed by S.M. Stigler [167]. In p. 284, a most interesting example is described, namely that of the celebrated *normal distribution*. It was first introduced by Abraham De Moivre in 1733, then by Pierre Simon de Laplace in 1774, then by Robert Adrain in 1808, and finally by Carl Friedrich Gauss in 1809, nothing less than 76 years after its first publication! This distribution is universally called *Gaussian* because of the remarkable insights of Gauss concerning the theory of errors, applicable in all experimental sciences. A less glamorous illustration of the same phenomenon, but nevertheless interesting in the present context, is that of Renyi entropy [145]. According to I. Csiszar [64], p. 73, the Renyi entropy had already been essentially introduced by Paul-Marcel Schutzenberger [161].

The entropy defined in Eq. (33) has the following main properties:

(i) S_q is nonnegative ($\forall q$);
(ii) S_q is expansible ($\forall q > 0$);
(iii) S_q attains its maximal (minimal) value $k \ln_q W$ for $q > 0$ (for $q < 0$);
(iv) S_q is concave (convex) for $q > 0$ (for $q < 0$);
(v) S_q is Lesche-stable ($\forall q > 0$) [2];
(vi) S_q yields a finite upper bound of the entropy production per unit time for a special value of q, whenever the sensitivity to the initial conditions exhibits an upper bound which asymptotically increases as a *power* of time. For example, many $D = 1$ nonlinear dynamical systems have a vanishing maximal Lyapunov exponent λ_1 and exhibit a sensitivity to the initial conditions which is (upper) bounded by

$$\xi = \mathrm{e}_q^{\lambda_q t}\,, \tag{35}$$

with $\lambda_q > 0$, $q < 1$, the q-exponential function e_q^x being the inverse of $\ln_q x$. More explicitly (see Fig. 3)

$$\mathrm{e}_q^x \equiv \begin{cases} [1 + (1-q)\,x]^{\frac{1}{1-q}} & \text{if} \quad 1 + (1-q)x > 0\,; \\ 0 & \text{otherwise}\,. \end{cases} \tag{36}$$

Such systems have a finite entropy production per unit time, which satisfies a q-generalized Pesin-like identity, namely, for the construction described in Sect. "Introduction",

$$K_q \equiv \lim_{t\to\infty} \lim_{W\to\infty} \lim_{M\to\infty} \frac{S_q(t)}{t} = \lambda_q\,. \tag{37}$$

The situation is in fact sensibly much richer than briefly described here. For further details, see [27,28,29,30,93,116,117,146,147,148,149,150,151,152].

(vii) S_q is nonadditive for $q \neq 1$. Indeed, for independent subsystems A and B, it can be straightforwardly proved

$$\frac{S_q(A+B)}{k} = \frac{S_q(A)}{k} + \frac{S_q(B)}{k} + (1-q)\frac{S_q(A)}{k}\frac{S_q(B)}{k}\,, \tag{38}$$

or, equivalently,

$$S_q(A+B) = S_q(A) + S_q(B) + \frac{(1-q)}{k}\,S_q(A)\,S_q(B)\,, \tag{39}$$

which makes explicit that $(1-q) \to 0$ plays the same role as $k \to \infty$. Property (38), occasionally referred to in the literature as *pseudo-additivity*, can be called *subadditivity* (*superadditivity*) for $q > 1$ ($q < 1$).

(viii) $S_q = -k\,D_q \sum_{i=1}^{W} p_i^x|_{x=1}$, where the 1909 Jackson differential operator is defined as follows:

$$D_q f(x) \equiv \frac{f(qx) - f(x)}{qx - x} \qquad (D_1 f(x) = \mathrm{d}f(x)/\mathrm{d}x)\,. \tag{40}$$

(ix) An uniqueness theorem has been proved by Santos [159], which generalizes, for arbitrary q, that of Shannon [162].
Let us assume that an entropic form $S(\{p_i\})$ satisfies the following properties:
(a)

$$S(\{p_i\}) \quad \text{is a continuous function of } \{p_i\}; \tag{41}$$

(b)

$$S(p_i = 1/W, \forall i) \text{ monotonically increases with the total number of possibilities } W; \tag{42}$$

(c)

$$\frac{S(A+B)}{k} = \frac{S(A)}{k} + \frac{S(B)}{k} + (1-q)\frac{S(A)}{k}\frac{S(B)}{k} \quad \text{if} \quad p_{ij}^{A+B} = p_i^A p_j^B \; \forall (i,j)\,, \quad \text{with } k > 0; \tag{43}$$

(d)

$$S(\{p_i\}) = S(p_L, p_M) + p_L^q S(\{p_i/p_L\}) + p_M^q S(\{p_i/p_M\}) \quad \text{with } p_L \equiv \sum_{L \text{ terms}} p_i\,, p_M \equiv \sum_{M \text{ terms}} p_i\,(L+M=W)\,, \quad \text{and} \quad p_L + p_M = 1\,. \tag{44}$$

Then and only then [159] $S(\{p_i\}) = S_q(\{p_i\})$.

(x) Another (equivalent) uniqueness theorem was proved by Abe [1], which generalizes, for arbitrary q, that of Khinchin [100].
Let us assume that an entropic form $S(\{p_i\})$ satisfies the following properties:

(a)

$$S(\{p_i\}) \quad \text{is a continuous function of } \{p_i\}; \tag{45}$$

(b)

$$S(p_i = 1/W, \forall i) \quad \text{monotonically increases with the total number of possibilities } W; \tag{46}$$

(c)

$$S(p_1, p_2, \ldots, p_W, 0) = S(p_1, p_2, \ldots, p_W)\,; \tag{47}$$

(d)

$$\frac{S(A+B)}{k} = \frac{S(A)}{k} + \frac{S(B|A)}{k} + (1-q)\frac{S(A)}{k}\frac{S(B|A)}{k}$$

where $S(A+B) \equiv S\left(\left\{p_{ij}^{A+B}\right\}\right)$,

$S(A) \equiv S\left(\left\{\sum_{j=1}^{W_B} p_{ij}^{A+B}\right\}\right)$, and the conditional entropy

$$S(B|A) \equiv \frac{\sum_{i=1}^{W_A} \left(p_i^A\right)^q S\left(\left\{p_{ij}^{A+B}/p_i^A\right\}\right)}{\sum_{i=1}^{W_A} \left(p_i^A\right)^q} \quad (k > 0) \tag{48}$$

Then and only then [1] $S(\{p_i\}) = S_q(\{p_i\})$.

Additivity Versus Extensivity of the Entropy

It is of great importance to distinguish *additivity* from *extensivity*. An entropy S is additive [130] if its value for a system composed by two independent subsystems A and B satisfies $S(A+B) = S(A) + S(B)$ (hence, for N independent equal subsystems or elements, we have $S(N) = NS(1)$). Therefore, $S_{\rm BG}$ is additive, and $S_q (q \neq 1)$ is nonadditive. A substantially different matter is whether a given entropy S is extensive *for a given system*. An entropy is extensive if and only if $0 < \lim_{N\to\infty} S(N)/N < \infty$. What matters for satisfactorily matching thermodynamics is extensivity *not* additivity. For systems whose elements are nearly independent (i. e., essentially weakly correlated), $S_{\rm BG}$ is extensive and S_q is nonextensive. For systems whose elements are strongly correlated in a special manner, $S_{\rm BG}$ is nonextensive, whereas S_q is extensive *for a special value of* $q \neq 1$ (and nonextensive for all the others).

Let us illustrate these facts for some simple examples of equal probabilities. If $W(N) \sim A\mu^N (A > 0, \mu > 1$, and $N \to \infty)$, the entropy which is extensive is $S_{\rm BG}$. Indeed, $S_{\rm BG}(N) = k \ln W(N) \sim (\ln \mu)N \propto N$ (it is equally trivial to verify that $S_q(N)$ is nonextensive for any $q \neq 1$). If $W(N) \sim BN^\rho (B > 0, \rho > 0$, and $N \to \infty)$, the entropy which is extensive is $S_{1-(1/\rho)}$. Indeed, $S_{1-(1/\rho)}(N) \sim k\rho B^{1/\rho} N \propto N$ (it is equally trivial to verify that $S_{\rm BG}(N) \propto \ln N$, hence nonextensive). If $W(N) \sim C\mu^{N^\gamma} (C > 0, \mu > 1, \gamma \neq 1$, and $N \to \infty)$, then $S_q(N)$ is nonextensive for *any* value of q. Therefore, in such a complex case, one must in principle refer to some other kind of entropic functional in order to match the extensivity required by classical thermodynamics.

Various nontrivial abstract mathematical models can be found in [113,160,186,198,199] for which $S_q (q \neq 1)$ is extensive. Moreover, a physical realization is also available now [60,61] for a many-body quantum Hamiltonian, namely the ground state of the following one:

$$\mathcal{H} = -\sum_{i=1}^{N-1} \left[(1+\gamma) S_i^x S_{i+1}^x + (1-\gamma) S_i^y S_{i+1}^y\right] - 2\lambda \sum_{i=1}^{N} S_i^z\,, \tag{49}$$

where λ is a transverse magnetic field, and (S_i^x, S_i^y, S_i^z) are Pauli matrices; for $|\gamma| = 1$ we have the Ising model, for $0 < |\gamma| < 1$, we have the anisotropic XY model, and, for $\gamma = 0$, we have the isotropic XY model. The two former share the same symmetry and consequently belong to the same critical universality class (the Ising uni-

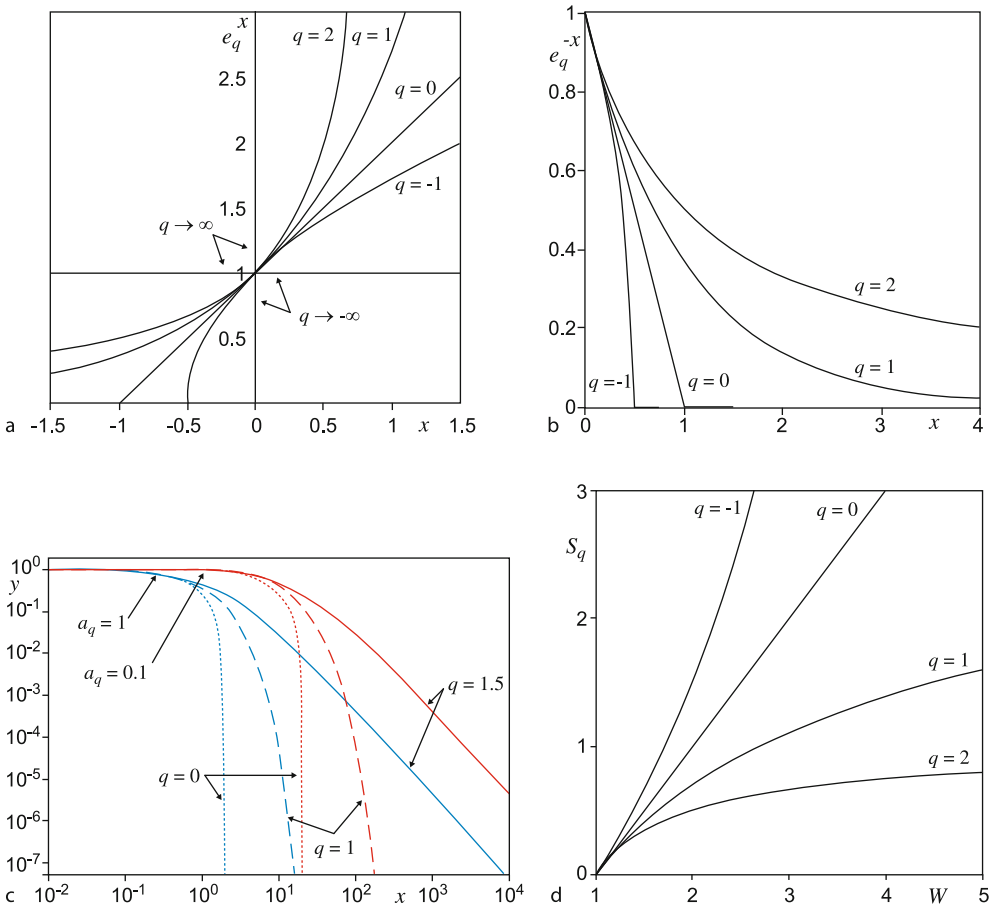

Entropy, Figure 3

The q-exponential and q-logarithm functions in typical representations: a Linear-linear representation of e_q^x; b Linear-linear representation of e_q^{-x}; c Log-log representation of $y(x) = e_q^{-a_q x}$, solution of $dy/dx = -a_q y^q$ with $y(0) = 1$; d Linear-linear representation of $S_q = \ln_q W$ (value of the entropy for equal probabilities)

versality class, which corresponds to a so-called *central charge* $c = 1/2$), whereas the latter one belongs to a different universality class (the XX one, which corresponds to a central charge $c = 1$). At temperature $T = 0$ and $N \to \infty$, this model exhibits a second-order phase transition as a function of λ. For the Ising model, the critical value is $\lambda = 1$, whereas, for the XX model, the entire line $0 \le \lambda \le 1$ is critical. Since the system is at its ground state (assuming a vanishingly small magnetic field component in the $x - y$ plane), it is a *pure state* (i. e., its density matrix ρ_N is such that $\mathrm{Tr}\,\rho_N^2 = 1, \forall N$), hence the entropy $S_q(N)(\forall q > 0)$ is strictly zero. However, the situation is drastically different for any L-sized block of the infinite chain. Indeed, $\rho_L \equiv \mathrm{Tr}_{N-L}\rho_N$ is such that $\mathrm{Tr}\,\rho_L^2 < 1$, i. e., it is a *mixed state*, hence it has a nonzero entropy. The block entropy $S_q(L) \equiv \lim_{N\to\infty} S_q(N, L)$ monotonically increases with L for all values of q. And it does so *linearly* for

$$q = \frac{\sqrt{9 + c^2} - 3}{c}\,, \tag{50}$$

where c is the central charge which emerges in quantum field theory [54]. In other words, $0 < \lim_{L\to\infty} S_{(\sqrt{9+c^2}-3)/c}(L)/L < \infty$. Notice that q increases from zero to unity when c increases from zero to infinity; $q = \sqrt{37} - 6 \simeq 0.083$ for $c = 1/2$ (Ising model), $q = \sqrt{10} - 3 \simeq 0.16$ for $c = 1$ (isotropic XY model), $q = 1/2$ for $c = 4$ (dimension of space-time), and $q = (\sqrt{685} - 3)/26 \simeq 0.89$ for $c = 26$, related to string theory [89]. The possible physical interpretation of the limit $c \to \infty$ is still unknown, although it could correspond to some sort of mean field approach.

Nonextensive Statistical Mechanics

To generalize BG statistical mechanics for the canonical ensemble, we optimize S_q with constraint (15) and also

$$\sum_{i=1}^{W} P_i E_i = U_q \,, \tag{51}$$

where

$$P_i \equiv \frac{p_i^q}{\sum_{j=1}^{W} p_i^q} \qquad \left(\sum_{i=1}^{W} P_i = 1\right) \tag{52}$$

is the so-called *escort distribution* [33]. It follows that $p_i = P_i^{1/q}/\sum_{j=1}^{W} P_j^{1/q}$. There are various converging reasons for being appropriate to impose the energy constraint with the $\{P_i\}$ instead of with the original $\{p_i\}$. The full discussion of this delicate point is beyond the present scope. However, some of these intertwined reasons are explored in [184]. By imposing Eq. (51), we follow [193], which in turn reformulates the results presented in [71,183]. The passage from one to the other of the various existing formulations of the above optimization problem are discussed in detail in [83,193].

The entropy optimization yields, for the stationary state,

$$p_i = \frac{\mathrm{e}_q^{-\beta_q(E_i - U_q)}}{\bar{Z}_q} \,, \tag{53}$$

with

$$\beta_q \equiv \frac{\beta}{\sum_{j=1}^{W} p_j^q} \,, \tag{54}$$

and

$$\bar{Z}_q \equiv \sum_{i}^{W} \mathrm{e}_q^{-\beta_q(E_i - U_q)} \,, \tag{55}$$

β being the Lagrange parameter associated with the constraint (51). Equation (53) makes explicit that the probability distribution is, for fixed β_q, invariant with regard to the arbitrary choice of the zero of energies. The stationary state (or (meta)equilibrium) distribution (53) can be rewritten as follows:

$$p_i = \frac{\mathrm{e}_q^{-\beta'_q E_i}}{Z'_q} \,, \tag{56}$$

with

$$Z'_q \equiv \sum_{j=1}^{W} \mathrm{e}_q^{-\beta'_q E_j} \,, \tag{57}$$

and

$$\beta'_q \equiv \frac{\beta_q}{1 + (1-q)\beta_q U_q} \,. \tag{58}$$

The form (56) is particularly convenient for many applications where comparison with experimental or computational data is involved. Also, it makes clear that p_i asymptotically decays like $1/E_i^{1/(q-1)}$ for $q > 1$, and has a cutoff for $q < 1$, instead of the exponential decay with E_i for $q = 1$.

The connection to thermodynamics is established in what follows. It can be proved that

$$\frac{1}{T} = \frac{\partial S_q}{\partial U_q} \,, \tag{59}$$

with $T \equiv 1/(k\beta)$. Also we prove, for the free energy,

$$F_q \equiv U_q - TS_q = -\frac{1}{\beta} \ln_q Z_q \,, \tag{60}$$

where

$$\ln_q Z_q = \ln_q \bar{Z}_q - \beta U_q \,. \tag{61}$$

This relation takes into account the trivial fact that, in contrast with what is usually done in BG statistics, the energies $\{E_i\}$ are here referred to U_q in (53). It can also be proved

$$U_q = -\frac{\partial}{\partial \beta} \ln_q Z_q \,, \tag{62}$$

as well as relations such as

$$C_q \equiv T\frac{\partial S_q}{\partial T} = \frac{\partial U_q}{\partial T} = -T\frac{\partial^2 F_q}{\partial T^2} \,. \tag{63}$$

In fact the entire Legendre transformation structure of thermodynamics is q-invariant, which is both remarkable and welcome.

A Connection Between Entropy and Diffusion

We review here one of the main common aspects of entropy and diffusion. We shall present on equal footing both the *BG* and the nonextensive cases [13,138,192,216]. Let us extremize the entropy

$$S_q = k\frac{1 - \int_{-\infty}^{\infty} \mathrm{d}(x/\sigma)\,[\sigma\, p(x)]^q}{q-1} \tag{64}$$

with the constraints

$$\int_{-\infty}^{\infty} \mathrm{d}x\, p(x) = 1 \tag{65}$$

and

$$\langle x^2 \rangle_q \equiv \frac{\int_{-\infty}^{\infty} \mathrm{d}x\, x^2 [p(x)]^q}{\int_{-\infty}^{\infty} \mathrm{d}x\, [p(x)]^q} = \sigma^2 \,, \tag{66}$$

$\sigma > 0$ being some fixed value having the same physical dimensions of the variable x. We straightforwardly obtain the following distribution:

$$p_q(x) = \begin{cases} \frac{1}{\sigma}\left[\frac{q-1}{\pi(3-q)}\right]^{1/2} \frac{\Gamma\left(\frac{1}{q-1}\right)}{\Gamma\left(\frac{3-q}{2(q-1)}\right)} \frac{1}{\left[1+\frac{q-1}{3-q}\frac{x^2}{\sigma^2}\right]^{1/(q-1)}} & \text{if} \quad 1 < q < 3\,; \\ \frac{1}{\sigma}\frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-x^2/2\sigma^2} & \text{if} \quad q = 1\,; \\ \frac{1}{\sigma}\left[\frac{1-q}{\pi(3-q)}\right]^{1/2} \frac{\Gamma\left(\frac{5-3q}{2(1-q)}\right)}{\Gamma\left(\frac{2-q}{1-q}\right)} \left[1-\frac{1-q}{3-q}\frac{x^2}{\sigma^2}\right]^{1/(1-q)} & \text{for } |x| < \sigma\,[(3-q)/(1-q)]^{1/2}\,, \text{ and zero otherwise}\,, \\ & \text{if} \quad q < 1\,. \end{cases} \tag{67}$$

These distributions are frequently referred to as *q-Gaussians*. For $q > 1$, they asymptotically have a power-law tail ($q \geq 3$ is not admissible because the norm (65) cannot be satisfied); for $q < 1$, they have a compact support. For $q = 1$, the celebrated Gaussian is recovered; for $q = 2$, the Cauchy–Lorentz distribution is recovered; finally, for $q \to -\infty$, the uniform distribution within the interval $[-1, 1]$ is recovered. For $q = \frac{3+m}{1+m}$, m being an integer ($m = 1, 2, 3, \ldots$), we recover the Student's t-distributions with m degrees of freedom [79]. For $q = \frac{n-4}{n-2}$, n being an integer ($n = 3, 4, 5, \ldots$), we recover the so-called r-distributions with n degrees of freedom [79]. In other words, q-Gaussians are analytical extensions of Student's t- and r-distributions. In some communities they are also referred to as the Barenblatt form. For $q < 5/3$, they have a finite variance which monotonically increases for q varying from $-\infty$ to 5/3; for $5/3 \leq q < 3$, the variance diverges.

Let us now make a connection of the above optimization problem with diffusion. We focus on the following quite general diffusion equation:

$$\frac{\partial^\delta p(x,t)}{\partial t^\delta} = \frac{\partial}{\partial x}\left[\frac{\partial U(x)}{\partial x} p(x,t)\right] + D\frac{\partial^\alpha [p(x,t)]^{2-q}}{\partial |x|^\alpha} \quad (0 < \delta \leq 1;\ 0 < \alpha \leq 2;\ q < 3;\ t \geq 0)\,, \tag{68}$$

with a generic nonsingular *potential* $U(x)$, and a *generalized diffusion coefficient* D which is positive (negative) for $q < 2$ ($2 < q < 3$). Several particular instances of this equation have been discussed in the literature (see [40,86,106,131,188] and references therein).

For example, the stationary state for $\alpha = 2$, $\forall\delta$, and any confining potential (i. e., $\lim_{|x|\to\infty} U(x) = \infty$) is given by [106]

$$p(x,\infty)_q = \frac{\mathrm{e}_q^{-\beta\,[U(x)-U(0)]}}{Z}\,, \tag{69}$$

$$Z \equiv \int_{-\infty}^{\infty} \mathrm{d}x\, \mathrm{e}_q^{-\beta\,[U(x)-U(0)]}\,, \tag{70}$$

$$1/\beta \equiv kT \propto |D|\,, \tag{71}$$

which precisely is the distribution obtained within nonextensive statistical mechanics through extremization of S_q.

Also, the solution for $\alpha = 2$, $\delta = 1$, $U(x) = -k_1 x + \frac{k_2}{2} x^2 (\forall\, k_1,$ and $k_2 \geq 0)$, and $p(x,0) = \delta(x)$ is given by [188]

$$p_q(x,t) = \frac{\mathrm{e}_q^{-\beta(t)[x-x_M(t)]^2}}{Z_q(t)}\,, \tag{72}$$

$$\frac{\beta(t)}{\beta(0)} = \left[\frac{Z_q(0)}{Z_q(t)}\right]^2 = \left[\left(1-\frac{1}{K_2}\,\mathrm{e}^{-t/\tau}\right)+\frac{1}{K_2}\right]^{-2/(3-q)}\,, \tag{73}$$

$$K_2 \equiv \frac{k_2}{2(2-q)D\beta(0)[Z_q(0)]^{q-1}}\,, \tag{74}$$

$$\tau \equiv \frac{1}{k_2(3-q)}\,, \tag{75}$$

$$x_M(t) \equiv \frac{k_1}{k_2} + \left[x_M(0) - \frac{k_1}{k_2}\right] \mathrm{e}^{-k_2 t}\,. \tag{76}$$

In the limit $k_2 \to 0$, Eq. (73) becomes

$$Z_q(t) = \left\{[Z_q(0)]^{3-q} + 2(2-q)(3-q)D\beta(0)\,[Z_q(0)]^2\, t\right\}^{1/(3-q)}\,, \tag{77}$$

which, in the $t \to \infty$ limit, yields

$$\frac{1}{\beta(t)} \propto [Z_q(t)]^2 \propto t^{2/(3-q)}\,. \tag{78}$$

In other words, x^2 scales like t^γ, with

$$\gamma = \frac{2}{3-q}\,, \tag{79}$$

hence, for $q > 1$ we have $\gamma > 1$ (i. e., *superdiffusion*; in particular, $q = 2$ yields $\gamma = 2$, i. e., ballistic diffusion),

for $q < 1$ we have $\gamma < 1$ (i. e., *subdiffusion*; in particular, $q \to -\infty$ yields $\gamma = 0$, i. e., localization), and naturally, for $q = 1$, we obtain *normal diffusion*. Four systems are known for which results have been found that are consistent with prediction (79). These are the motion of *Hydra viridissima* [206], defect turbulence [73], simulation of a silo drainage [22], and molecular dynamics of a many-body long-range-interacting classical system of rotators ($\alpha - XY$ model) [143]. For the first three, it has been found $(q, \gamma) \simeq (3/2, 4/3)$. For the latter one, relation (79) has been verified for various situations corresponding to $\gamma > 1$.

Finally, for the particular case $\delta = 1$ and $U(x) = 0$, Eq. (68) becomes

$$\frac{\partial p(x,t)}{\partial t} = D \frac{\partial^{\alpha} [p(x,t)]^{2-q}}{\partial |x|^{\alpha}} \quad (0 < \alpha \leq 2;\ q < 3)\,. \tag{80}$$

The diffusion constant D just rescales time t. Only two parameters are therefore left, namely α and q.

The linear case (i. e., $q = 1$) has two types of solutions: Gaussians for $\alpha = 2$, and *Lévy-* (or α*-stable*) distributions for $0 < \alpha < 2$. The case $\alpha = 2$ corresponds to the *Central Limit Theorem*, where the $N \to \infty$ attractor of the sums of N *independent* random variables with *finite* variance precisely is a Gaussian. The case $0 < \alpha < 2$ corresponds to the sometimes called *Levy–Gnedenko Central Limit Theorem*, where the $N \to \infty$ attractor of the sums of N *independent* random variables with *infinite* variance (and appropriate asymptotics) precisely is a Lévy distribution with index α.

The nonlinear case (i. e., $q \neq 1$) has solutions that are q-Gaussians for $\alpha = 2$, and one might conjecture that, similarly, interesting solutions exist for $0 < \alpha < 2$. Furthermore, in analogy with the $q = 1$ case, one expects corresponding q-generalized Central Limit Theorems to exist [187]. This is precisely what we present in the next Section.

Standard and q-Generalized Central Limit Theorems

The q-Product

It has been recently introduced (independently and virtually simultaneously) [43,125] a generalization of the product, which is called *q-product*. It is defined, for $x \geq 0$ and $y \geq 0$, as follows:

$$x \otimes_q y \equiv \begin{cases} [x^{1-q} + y^{1-q} - 1]^{1/(1-q)} & \text{if} \quad x^{1-q} + y^{1-q} > 1\,; \\ 0 & \text{otherwise}\,. \end{cases} \tag{81}$$

It has, among others, the following properties:
it recovers the *standard product* as a particular instance, i. e.,

$$x \otimes_1 y = xy\,; \tag{82}$$

it is *commutative*, i. e.,

$$x \otimes_q y = y \otimes_q x\,; \tag{83}$$

it is *additive under q-logarithm*, i. e.,

$$\ln_q(x \otimes_q y) = \ln_q x + \ln_q y \tag{84}$$

(whereas we remind that $\ln_q(x\,y) = \ln_q x + \ln_q y + (1-q)(\ln_q x)(\ln_q y)$;
it has a $(2-q)$*-duality/inverse* property, i. e.,

$$1/(x \otimes_q y) = (1/x) \otimes_{2-q} (1/y)\,; \tag{85}$$

it is *associative*, i. e.,

$$\begin{aligned} x \otimes_q (y \otimes_q z) &= (x \otimes_q y) \otimes_q z = x \otimes_q y \otimes_q z \\ &= (x^{1-q} + y^{1-q} + z^{1-q} - 2)^{1/(1-q)}\,; \end{aligned} \tag{86}$$

it admits *unity*, i. e.,

$$x \otimes_q 1 = x\,. \tag{87}$$

and, for $q \geq 1$, also a *zero*, i. e.,

$$x \otimes_q 0 = 0 \quad (q \geq 1)\,. \tag{88}$$

The q-Fourier Transform

We shall introduce the q-Fourier transform of a quite generic function $f(x)$ $(x \in \mathcal{R})$ as follows [140,189,202,203,204,205]:

$$\begin{aligned} F_q[f](\xi) &\equiv \int_{-\infty}^{\infty} \mathrm{d}x\, \mathrm{e}_q^{\mathrm{i}x\xi} \otimes_q f(x) \\ &= \int_{-\infty}^{\infty} \mathrm{d}x\, \mathrm{e}_q^{\mathrm{i}x\xi [f(x)]^{q-1}} f(x) \end{aligned}\,, \tag{89}$$

where we have primarily focused on the case $q \geq 1$. In contrast with the $q = 1$ case (standard Fourier transform), this integral transformation is *nonlinear* for $q \neq 1$. It has a remarkable property, namely that the q-Fourier transform of a q-Gaussian is another q-Gaussian:

$$F_q\left[N_q \sqrt{\beta}\, \mathrm{e}_q^{-\beta x^2}\right](\xi) = \mathrm{e}_{q_1}^{-\beta_1 \xi^2}\,, \tag{90}$$

with

$$N_q \equiv \begin{cases} \left[\frac{q-1}{\pi}\right]^{1/2} \frac{\Gamma\left(\frac{1}{q-1}\right)}{\Gamma\left(\frac{3-q}{2(q-1)}\right)} & \text{if} \quad 1 < q < 3\,, \\ \frac{1}{\sqrt{\pi}} & \text{if} \quad q = 1\,, \\ \frac{3-q}{2}\left[\frac{1-q}{\pi}\right]^{1/2} \frac{\Gamma\left(\frac{3-q}{2(1-q)}\right)}{\Gamma\left(\frac{1}{1-q}\right)} & \text{if} \quad q < 1\,, \end{cases} \tag{91}$$

and

$$q_1 = z(q) \equiv \frac{1+q}{3-q}\,, \tag{92}$$

$$\beta_1 = \frac{1}{\beta^{2-q}} \frac{N_q^{2(1-q)}(3-q)}{8}\,. \tag{93}$$

Equation (93) can be re-written as $\beta^{\sqrt{2-q}}\beta_1^{1/\sqrt{2-q}} = [(N_q^{2(1-q)}(3-q))/8]^{1/\sqrt{2-q}} \equiv K(q)$, which, for $q = 1$, recovers the well known Heisenberg-uncertainty-principle-like relation $\beta\beta_1 = 1/4$.

If we iterate n times the relation $z(q)$ in Eq. (92), we obtain the following algebra:

$$q_n(q) = \frac{2q + n(1-q)}{2 + n(1-q)} \quad (n = 0, \pm 1, \pm 2, \ldots)\,, \tag{94}$$

which can be conveniently re-written as

$$\frac{2}{1-q_n(q)} = \frac{2}{1-q} + n \quad (n = 0, \pm 1, \pm 2, \ldots)\,. \tag{95}$$

(See Fig. 4). We easily verify that $q_n(1) = 1\,(\forall n)$, $q_{\pm\infty}(q) = 1\,(\forall q)$, as well as

$$\frac{1}{q_{n+1}} = 2 - q_{n-1}\,. \tag{96}$$

This relation connects the so called *additive duality* $q \to (2-q)$ and *multiplicative duality* $q \to 1/q$, frequently emerging in all types of calculations in the literature. Moreover, we see from Eq. (95) that multiple values of q are expected to emerge in connection with diverse properties of nonextensive systems, i. e., in systems whose basic entropy is the nonadditive one S_q. Such is the case of the so called *q-triplet* [185], observed for the first time in the magnetic field fluctuations of the solar wind, as it has been revealed by the analysis of the data sent to NASA by the spacecraft Voyager 1 [48].

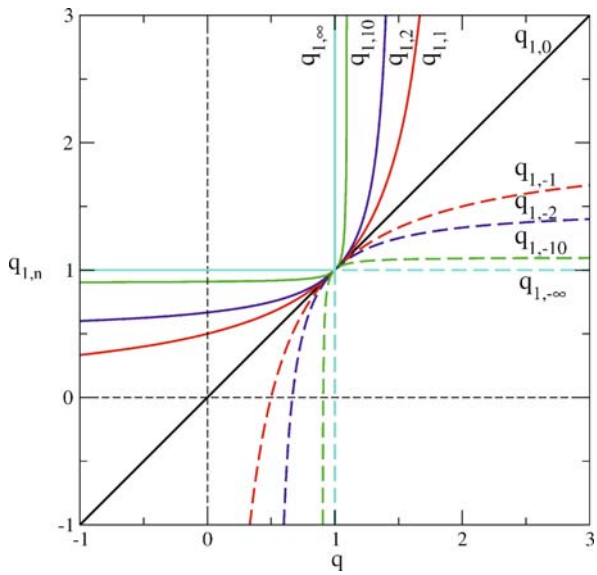

Entropy, Figure 4
The q-dependence of $q_n(q) \equiv q_{2,n}(q)$

q-Independent Random Variables

Two random variables X [with density $f_X(x)$] and Y [with density $f_Y(y)$] having zero q-mean values (e. g., if $f_X(x)$ and $f_Y(y)$ are even functions) are said *q-independent*, with q_1 given by Eq. (92), if

$$F_q[X+Y](\xi) = F_q[X](\xi) \otimes_{q_1} F_q[Y](\xi)\,, \tag{97}$$

i. e., if

$$\int_{-\infty}^{\infty} \mathrm{d}z\, \mathrm{e}_q^{\mathrm{i}z\xi} \otimes_q f_{X+Y}(z) = \left[\int_{-\infty}^{\infty} \mathrm{d}x\, \mathrm{e}_q^{\mathrm{i}x\xi} \otimes_q f_X(x)\right] \otimes_{(1+q)/(3-q)} \left[\int_{-\infty}^{\infty} \mathrm{d}y\, \mathrm{e}_q^{iy\xi} \otimes_q f_X(y)\right]\,, \tag{98}$$

with

$$\begin{aligned} f_{X+Y}(z) &= \int_{-\infty}^{\infty} \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{d}y\, h(x,y)\, \delta(x+y-z) \\ &= \int_{-\infty}^{\infty} \mathrm{d}x\, h(x, z-x) \\ &= \int_{-\infty}^{\infty} \mathrm{d}y\, h(z-y, y) \end{aligned} \tag{99}$$

where $h(x,y)$ is the joint density.

Clearly, q-independence means *independence* for $q = 1$ (i. e., $h(x,y) = f_X(x) f_Y(y)$), and implies a special correlation for $q \neq 1$. Although the full understanding of this correlation is still under progress, q-independence appears to be consistent with scale-invariance.

Entropy, Table 1

The attractors corresponding to the four basic cases, where the N variables that are being summed are q-independent (i. e., globally correlated) with $q_1 = (1+q)/(3-q)$; $\sigma_Q \equiv (\int_{-\infty}^{\infty} dx\, x^2\, [f(x)]^Q)/(\int_{-\infty}^{\infty} dx\, [f(x)]^Q)$ with $Q \equiv 2q-1$. The attractor for $(q,\alpha) = (1,2)$ is a Gaussian $G(x) \equiv L_{1,2}$ (standard Central Limit Theorem); for $q = 1$ and $0 < \alpha < 2$, it is a Lévy distribution $L_\alpha \equiv L_{1,\alpha}$ (the so called Lévy-Gnedenko limit theorem); for $\alpha = 2$ and $q \neq 1$, it is a q-Gaussian $G_q \equiv L_{q,2}$ (the q-Central Limit Theorem; [203]); finally, for $q \neq 1$ and $0 < \alpha < 2$, it is a generic (q,α)-stable distribution $L_{q,\alpha}$ ([204,205]). See [140,189] for typical illustrations of the four types of attractors. The distribution $L_\alpha(x)$ remains, for $1 < \alpha < 2$, close to a Gaussian for $|x|$ up to about $x_c(1,\alpha)$, where it makes a crossover to a power-law. The distribution $G_q(x)$ remains, for $q > 1$, close to a Gaussian for $|x|$ up to about $x_c(q,2)$, where it makes a crossover to a power-law. The distribution $L_{q,\alpha}(x)$ remains, for $q > 1$ and $\alpha < 2$, close to a Gaussian for $|x|$ up to about $x_c^{(1)}(q,\alpha)$, where it makes a crossover to a power-law (*intermediate regime*), which lasts further up to about $x_c^{(2)}(q,\alpha)$, where it makes a second crossover to another power-law (*distant regime*)

	$q = 1$ [independent]	$q \neq 1$ (i. e., $Q \neq 1$) [globally correlated]
$\sigma_Q < \infty$ ($\alpha = 2$)	$G(x)$ [with same σ_1 of $f(x)$]	$G_q(x) = G_{(3q_1-1)/(1+q_1)}(x)$ [with same σ_Q of $f(x)$] $G_q(x) \sim G(x)$ if $\lvert x\rvert \ll x_c(q,2)$ $G_q(x) \sim C_{q,2}/\lvert x\rvert^{2/(q-1)}$ if $\lvert x\rvert \gg x_c(q,2)$ for $q > 1$, with $\lim_{q\to 1} x_c(q,2) = \infty$
$\sigma_Q \to \infty$ ($\alpha < 2$)	$L_\alpha(x)$ [with same $\lvert x\rvert \to \infty$ behavior of $f(x)$] $L_\alpha(x) \sim G(x)$ if $\lvert x\rvert \ll x_c(1,\alpha)$ $L_\alpha(x) \sim C_{1,\alpha}/\lvert x\rvert^{1+\alpha}$ if $\lvert x\rvert \gg x_c(1,\alpha)$ with $\lim_{\alpha\to 2} q_c(1,\alpha) = \infty$	$L_{q,\alpha}(x)$ [with same $\lvert x\rvert \to \infty$ behavior of $f(x)$] $L_{q,\alpha} \sim C_{q,\alpha}^{(\text{intermediate})}/\lvert x\rvert^{\frac{2(1-q)-\alpha(3-q)}{2(q-1)}}$ if $x_c^{(1)}(q,\alpha) \ll \lvert x\rvert \ll x_c^{(2)}(q,\alpha)$ $L_{q,\alpha} \sim C_{q,\alpha}^{(\text{distant})}/\lvert x\rvert^{\frac{1+\alpha}{1+\alpha(q-1)}}$ if $\lvert x\rvert \gg x_c^{(2)}(q,\alpha)$

q-Generalized Central Limit Theorems

It is out of the scope of the present survey to provide the details of the complex proofs of the q-generalized central limit theorems. We shall restrict to the presentation of their structure. Let us start by introducing a notation which is important for what follows. A distribution is said (q,α)-*stable distribution* $L_{q,\alpha}(x)$ if its q-Fourier transform $\mathcal{L}_{q,\alpha}(\xi)$ is of the form

$$\mathcal{L}_{q,\alpha}(\xi) = a\, e_{q_1}^{-b\,|\xi|^\alpha} \quad [a > 0,\ b > 0,\ 0 < \alpha \le 2,\ q_1 = (q+1)/(3-q)]\,. \tag{100}$$

Consistently, $L_{1,2}$ are Gaussians, $L_{1,\alpha}$ are Lévy distributions, and $L_{q,2}$ are q-Gaussians.

We are seeking for the $N \to \infty$ *attractor* associated with the sum of N identical and distinguishable random variables each of them associated with one and the same arbitrary symmetric distribution $f(x)$. The random variables are independent for $q = 1$, and correlated in a special manner for $q \neq 1$. To obtain the $N \to \infty$ invariant distribution, i. e. the attractor, the sum must be rescaled, i. e., divided by N^δ, where

$$\delta = \frac{1}{\alpha(2-q)}\,. \tag{101}$$

For $(\alpha, q) = (2,1)$, we recover the traditional $1/\sqrt{N}$ rescaling of Brownian motion. At the present stage, the theorems have been established for $q \ge 1$ and are summarized in Table 1. The case $q < 1$ is still open at the time at which these lines are being written. Two $q < 1$ cases have been preliminarily explored numerically in [124] and in [171]. The numerics seemed to indicate that the $N \to \infty$ limits would be q-Gaussians for both models. However, it has been analytically shown [94] that it is not exactly so. The limiting distributions numerically are amazingly close to q-Gaussians, but they are in fact different. Very recently, another simple scale-invariant model has been introduced [153], whose attractor has been analytically shown to be a q-Gaussian.

These $q \neq 1$ theorems play for the nonadditive entropy S_q and nonextensive statistical mechanics the same grounding role that the well known $q = 1$ theorems play for the additive entropy S_{BG} and BG statistical mechanics. In particular, interestingly enough, the ubiquity of Gaussians and of q-Gaussians in natural, artificial and social systems may be understood on equal footing.

Future Directions

The concept of entropy permeates into virtually all quantitative sciences. The future directions could therefore be very varied. If we restrict, however, to the evidence

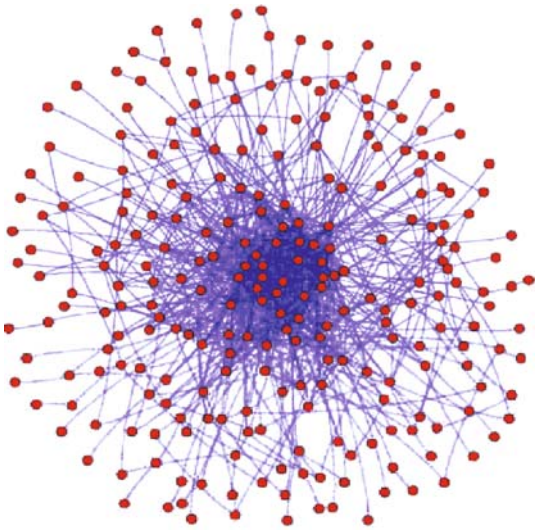

Entropy, Figure 5
Snapshot of a nongrowing dynamic network with $N = 256$ nodes (see details in [172], by courtesy of the author)

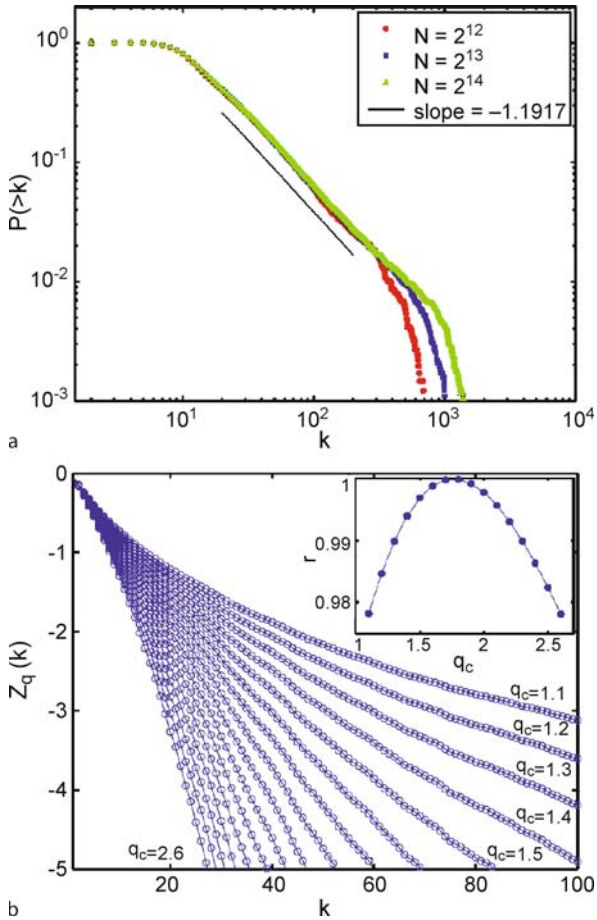

Entropy, Figure 6
Nongrowing dynamic network: a Cumulative degree distribution for typical values for the number N of nodes; b Same data of a in the convenient representation *linear q-log versus linear* with $Z_q(k) \equiv \ln_q[P_q(> k)] \equiv ([P_q(> k)]^{1-q} - 1)/(1-q)$ (the optimal fitting with a q-exponential is obtained for the value of q which has the highest value of the linear correlation r as indicated in the *inset*; here this is $q_c = 1.84$, which corresponds to the slope − 1.19 in a). See details in [172,173]

presently available, the main lines along which evolution occurs are:

Networks Many of the so-called scale-free networks, among others, systematically exhibit a *degree distribution* $p(k)$ (probability of a node having k links) which is of the form

$$p(k) \propto \frac{1}{(k_0 + k)^{\gamma}} \quad (\gamma > 0;\ k_0 > 0)\,, \tag{102}$$

or, equivalently,

$$p(k) \propto e_q^{-k/\kappa} \quad (q \geq 1;\ \kappa > 0)\,, \tag{103}$$

with $\gamma = 1/(q-1)$ and $k_0 = \kappa/(q-1)$ (see Figs. 5 and 6). This is not surprising since, if we associate to each link an "energy" (or *cost*) and to each node half of the "energy" carried by its links (the other half being associated with the other nodes to which any specific node is linked), the distribution of energies optimizing S_q precisely coincides with the degree distribution. If, for any reason, we consider k as the modulus of a d-dimensional vector $\mathbf{k}$, the optimization of the functional $S_q[p(\mathbf{k})]$ may lead to $p(k) \propto k^{\eta}\, e_q^{-k/\kappa}$, where k^{η} plays the role of a density of states, $\eta(d)$ being either zero (which reproduces Eq. (103)) or positive or negative. Several examples [12,39,76,91,165,172,173,212,213] already exist in the literature; in particular, the Barabasi–Albert universality class $\gamma = 3$ corresponds to $q = 4/3$. A deeper understanding of this connection might enable the systematic calculation of several meaningful properties of networks.

Nonlinear dynamical systems, self-organized criticality, and cellular automata Various interesting phenomena emerge in both low- and high-dimensional weakly chaotic deterministic dynamical systems, either dissipative or conservative. Among these phenomena we have the sensitivity to the initial conditions and the entropy production, which have been briefly addressed in Eq. (37) and related papers. But there is much more, such as relaxation, escape, glassy states, and distributions associated with the station-

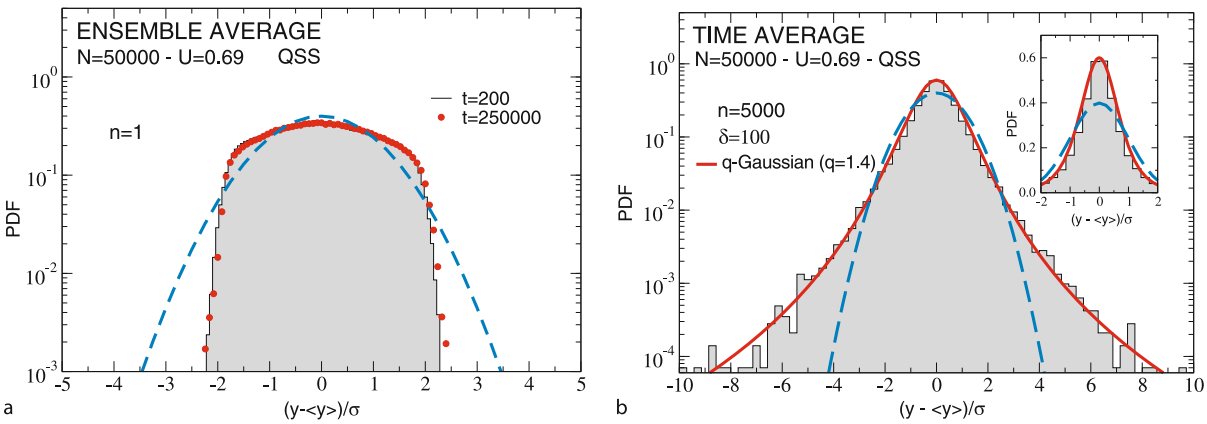

Entropy, Figure 7

Distribution of velocities for the HMF model at the quasi-stationary state (whose duration appears to diverge when $N \to \infty$). The blue curves indicate a Gaussian, for comparison. See details in [137]

ary state [14,15,31,46,62,67,68,77,103,111,122,123,154,170,174,176,177,179]. Also, recent numerical indications suggest the validity of a dynamical version of the q-generalized central limit theorem [175]. The possible connections between all these various properties is still in its infancy.

Long-range-interacting many-body Hamiltonians

A wide class of long-range-interacting N-body classical Hamiltonians exhibits collective states whose Lyapunov spectrum has a maximal value that vanishes in the $N \to \infty$ limit. As such, they constitute natural candidates for studying whether the concepts derived from the nonadditive entropy S_q are applicable. A variety of properties have been calculated, through molecular dynamics, for various systems, such as Lennard–Jones-like fluids, XY and Heisenberg ferromagnets, gravitational-like models, and others. One or more long-standing quasi-stationary states (infinitely long-standing in the limit $N \to \infty$) are typically observed before the terminal entrance into thermal equilibrium. Properties such as distribution of velocities and angles, correlation functions, Lyapunov spectrum, metastability, glassy states, aging, time-dependence of the temperature in isolated systems, energy whenever thermal contact with a large thermostat at a given temperature is allowed, diffusion, order parameter, and others, are typically focused on. An ongoing debate exists, also involving Vlasov-like equations, Lynden–Bell statistics, among others. The breakdown of ergodicity that emerges in various situations makes the whole discussion rich and complex. The activity of the research nowadays in this area is illustrated in papers such as [21,26,45,53,56,57,63,104,119,121,126,127,132,133,134,135,136,142,169,200]. A quite remarkable molecular-dynamical result has been obtained for a paradigmatic long-range Hamiltonian: the distribution of *time averaged* velocities sensibly differs from that found for the *ensemble-averaged* velocities, and has been shown to be numerically consistent with a q-Gaussian [137], as shown in Fig. 7. This result provides strong support to a conjecture made long ago: see Fig. 4 at p. 8 of [157].

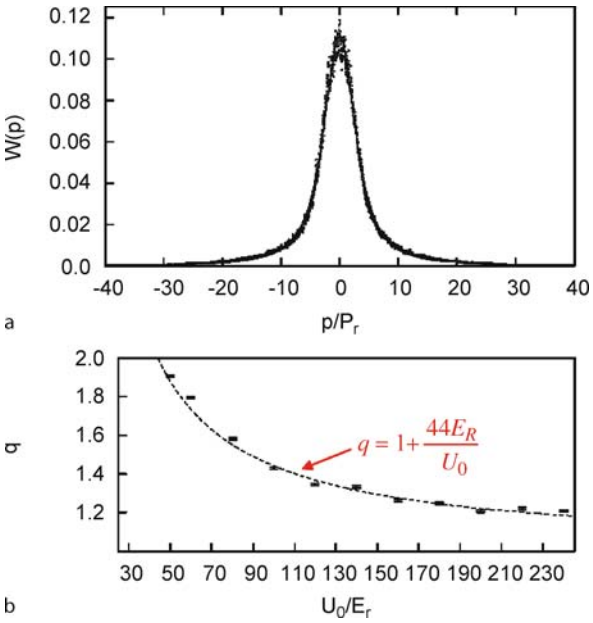

Entropy, Figure 8

Quantum Monte Carlo simulations in [81]: a Velocity distribution (superimposed with a q-Gaussian); b Index q (superimposed with Lutz prediction [110], by courtesy of the authors)

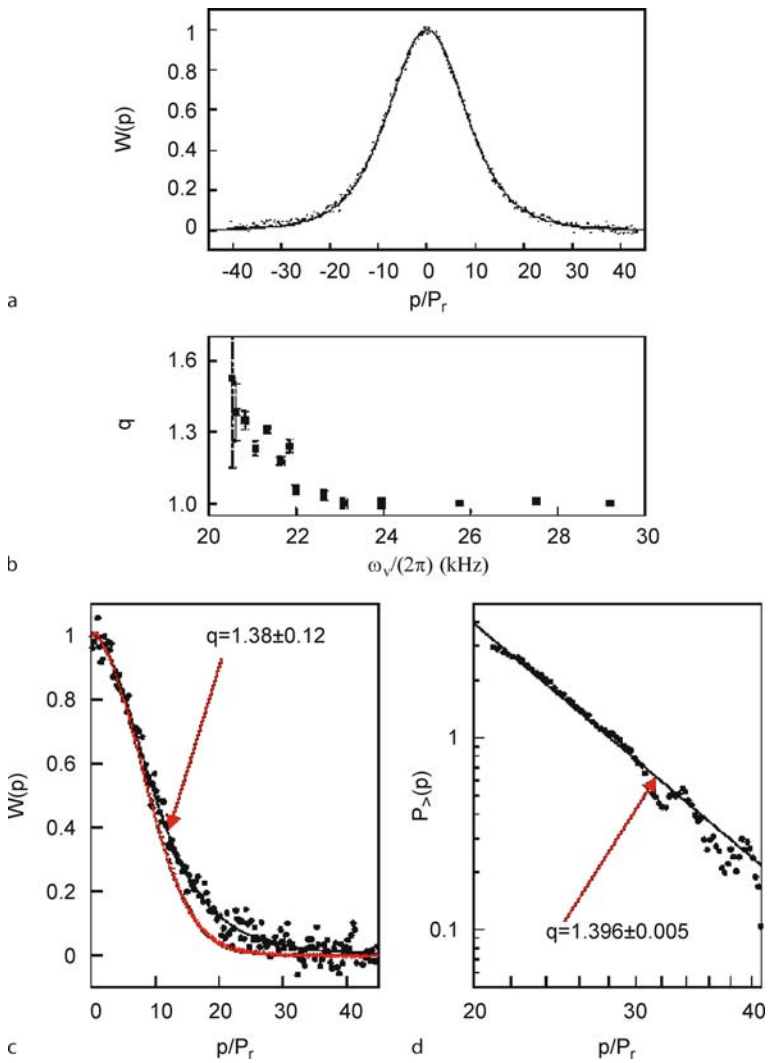

Entropy, Figure 9

Experiments in [81]: a Velocity distribution (superimposed with a *q*-Gaussian); b Index *q* as a function of the frequency; c Velocity distribution (superimposed with a *q*-Gaussian; the *red curve* is a Gaussian); d Tail of the velocity distribution (superimposed with the asymptotic power-law of a *q*-Gaussian). [By courtesy of the authors]

Stochastic differential equations Quite generic Fokker–Planck equations are currently being studied. Aspects such as fractional derivatives, nonlinearities, space-dependent diffusion coefficients are being focused on, as well as their connections to entropic forms, and associated generalized Langevin equations [20,23,24,70,128,168,214]. Quite recently, computational (see Fig. 8) and experimental (see Fig. 9) verifications of Lutz' 2003 prediction [110] have been exhibited [81], namely about the q-Gaussian form of the velocity distribution of cold atoms in dissipative optical lattices, with $q = 1 + 44E_R/U_0$(E_R and U_0 being energy parameters of the optical lattice). These experimental verifications are in variance with some of those exhibited previously [96], namely double-Gaussians. Although it is naturally possible that the experimental conditions have not been exactly equivalent, this interesting question remains open at the present time. A hint might be hidden in the recent results [62] obtained for a quite different problem, namely the size distributions of avalanches; indeed, at a critical state, a q-Gaussian shape was obtained, whereas, at a noncritical state, a double-Gaussian was observed.

Quantum entanglement and quantum chaos The nonlocal nature of quantum physics implies phenomena that are somewhat analogous to those originated by classical long-range interactions. Consequently, a variety of studies are being developed in connection with the entropy S_q [3,36,58,59,60,61,155,156,195]. The same happens with some aspects of quantum chaos [11,180,210,211].

Astrophysics, geophysics, economics, linguistics, cognitive psychology, and other interdisciplinary applications Applications are available and presently searched in many areas of physics (plasmas, turbulence, nuclear collisions, elementary particles, manganites), but also in interdisciplinary sciences such astrophysics [38,47,48,49,78,84,87,101,109,129,196], geophysics [4,5,6,7,8,9,10,62,208], economics [25,50,51,52,80,139,141,197,215], linguistics [118], cognitive psychology [181], and others.

Global optimization, image and signal processing Optimizing algorithms and related techniques for signal and image processing are currently being developed using the entropic concepts presented in this article [17,35,72,75,95,105,114,120,164,166,191].

Superstatistics and other generalizations The methods discussed here have been generalized along a variety of lines. These include Beck–Cohen superstatistics [32,34,65,190], crossover statistics [194,196], spectral statistics [201]. Also, a huge variety of entropies have been introduced which generalize in different manners the BG entropy, or even focus on other possibilities. Their number being nowadays over forty, we mention here just a few of them: see [18,44,69,98,99,115].

Acknowledgments

Among the very many colleagues towards which I am deeply grateful for profound and long-lasting comments along many years, it is a must to explicitly thank S. Abe, E.P. Borges, E.G.D. Cohen, E.M.F. Curado, M. Gell-Mann, R.S. Mendes, A. Plastino, A.R. Plastino, A.K. Rajagopal, A. Rapisarda and A. Robledo.

Bibliography

1. Abe S (2000) Axioms and uniqueness theorem for Tsallis entropy. Phys Lett A 271:74–79
2. Abe S (2002) Stability of Tsallis entropy and instabilities of Renyi and normalized Tsallis entropies: A basis for q-exponential distributions. Phys Rev E 66:046134
3. Abe S, Rajagopal AK (2001) Nonadditive conditional entropy and its significance for local realism. Physica A 289:157–164
4. Abe S, Suzuki N (2003) Law for the distance between successive earthquakes. J Geophys Res (Solid Earth) 108(B2):2113
5. Abe S, Suzuki N (2004) Scale-free network of earthquakes. Europhys Lett 65:581–586
6. Abe S, Suzuki N (2005) Scale-free statistics of time interval between successive earthquakes. Physica A 350:588–596
7. Abe S, Suzuki N (2006) Complex network of seismicity. Prog Theor Phys Suppl 162:138–146
8. Abe S, Suzuki N (2006) Complex-network description of seismicity. Nonlinear Process Geophys 13:145–150
9. Abe S, Sarlis NV, Skordas ES, Tanaka H, Varotsos PA (2005) Optimality of natural time representation of complex time series. Phys Rev Lett 94:170601
10. Abe S, Tirnakli U, Varotsos PA (2005) Complexity of seismicity and nonextensive statistics. Europhys News 36:206–208
11. Abul AY-M (2005) Nonextensive random matrix theory approach to mixed regular-chaotic dynamics. Phys Rev E 71:066207
12. Albert R, Barabasi AL (2000) Phys Rev Lett 85:5234–5237
13. Alemany PA, Zanette DH (1994) Fractal random walks from a variational formalism for Tsallis entropies. Phys Rev E 49:R956–R958
14. Ananos GFJ, Tsallis C (2004) Ensemble averages and nonextensivity at the edge of chaos of one-dimensional maps. Phys Rev Lett 93:020601
15. Ananos GFJ, Baldovin F, Tsallis C (2005) Anomalous sensitivity to initial conditions and entropy production in standard maps: Nonextensive approach. Euro Phys J B 46:409–417
16. Andrade RFS, Pinho STR (2005) Tsallis scaling and the long-range Ising chain: A transfer matrix approach. Phys Rev E 71:026126
17. Andricioaei I, Straub JE (1996) Generalized simulated annealing algorithms using Tsallis statistics: Application to conformational optimization of a tetrapeptide. Phys Rev E 53:R3055–R3058
18. Anteneodo C, Plastino AR (1999) Maximum entropy approach to stretched exponential probability distributions. J Phys A 32:1089–1097
19. Anteneodo C, Tsallis C (1998) Breakdown of exponential sensitivity to initial conditions: Role of the range of interactions. Phys Rev Lett 80:5313–5316
20. Anteneodo C, Tsallis C (2003) Multiplicative noise: A mechanism leading to nonextensive statistical mechanics. J Math Phys 44:5194–5203
21. Antoniazzi A, Fanelli D, Barre J, Chavanis P-H, Dauxois T, Ruffo S (2007) Maximum entropy principle explains quasi-stationary states in systems with long-range interactions: The example of the Hamiltonian mean-field model. Phys Rev E 75:011112
22. Arevalo R, Garcimartin A, Maza D (2007) A non-standard statistical approach to the silo discharge. Eur Phys J Special Topics 143:191–197
23. Assis PC Jr, da Silva LR, Lenzi EK, Malacarne LC, Mendes RS (2005) Nonlinear diffusion equation, Tsallis formalism and exact solutions. J Math Phys 46:123303
24. Assis PC Jr, da Silva PC, da Silva LR, Lenzi EK, Lenzi MK (2006) Nonlinear diffusion equation and nonlinear external force: Exact solution. J Math Phys 47:103302
25. Ausloos M, Ivanova K (2003) Dynamical model and nonextensive statistical mechanics of a market index on large time windows. Phys Rev E 68:046122

26. Baldovin F, Orlandini E (2006) Incomplete equilibrium in long-range interacting systems. Phys Rev Lett 97:100601
27. Baldovin F, Robledo A (2002) Sensitivity to initial conditions at bifurcations in one-dimensional nonlinear maps: Rigorous nonextensive solutions. Europhys Lett 60:518–524
28. Baldovin F, Robledo A (2002) Universal renormalization-group dynamics at the onset of chaos in logistic maps and nonextensive statistical mechanics. Phys Rev E 66:R045104
29. Baldovin F, Robledo A (2004) Nonextensive Pesin identity. Exact renormalization group analytical results for the dynamics at the edge of chaos of the logistic map. Phys Rev E 69:R045202
30. Baldovin F, Robledo A (2005) Parallels between the dynamics at the noise-perturbed onset of chaos in logistic maps and the dynamics of glass formation. Phys Rev E 72:066213
31. Baldovin F, Moyano LG, Majtey AP, Robledo A, Tsallis C (2004) Ubiquity of metastable-to-stable crossover in weakly chaotic dynamical systems. Physica A 340:205–218
32. Beck C, Cohen EGD (2003) Superstatistics. Physica A 322:267–275
33. Beck C, Schlogl F (1993) Thermodynamics of Chaotic Systems. Cambridge University Press, Cambridge
34. Beck C, Cohen EGD, Rizzo S (2005) Atmospheric turbulence and superstatistics. Europhys News 36:189–191
35. Ben A Hamza (2006) Nonextensive information-theoretic measure for image edge detection. J Electron Imaging 15: 013011
36. Batle J, Plastino AR, Casas M, Plastino A (2004) Inclusion relations among separability criteria. J Phys A 37:895–907
37. Batle J, Casas M, Plastino AR, Plastino A (2005) Quantum entropies and entanglement. Intern J Quantum Inf 3:99–104
38. Bernui A, Tsallis C, Villela T (2007) Deviation from Gaussianity in the cosmic microwave background temperature fluctuations. Europhys Lett 78:19001
39. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D-U (2006) Phys Rep 424:175–308
40. Bologna M, Tsallis C, Grigolini P (2000) Anomalous diffusion associated with nonlinear fractional derivative Fokker-Planck-like equation: Exact time-dependent solutions. Phys Rev E 62:2213–2218
41. Boltzmann L (1896) Vorlesungen über Gastheorie. Part II, ch I, paragraph 1. Leipzig, p 217; (1964) Lectures on Gas Theory (trans: Brush S). Univ. California Press, Berkeley
42. Boon JP, Tsallis C (eds) (2005) Nonextensive Statistical Mechanics: New Trends, New Perspectives. Europhysics News 36(6):185–231
43. Borges EP (2004) A possible deformed algebra and calculus inspired in nonextensive thermostatistics. Physica A 340:95–101
44. Borges EP, Roditi I (1998) A family of non-extensive entropies. Phys Lett A 246:399–402
45. Borges EP, Tsallis C (2002) Negative specific heat in a Lennard–Jones-like gas with long-range interactions. Physica A 305:148–151
46. Borges EP, Tsallis C, Ananos GFJ, Oliveira PMC (2002) Nonequilibrium probabilistic dynamics at the logistic map edge of chaos. Phys Rev Lett 89:254103
47. Burlaga LF, Vinas AF (2004) Multiscale structure of the magnetic field and speed at 1 AU during the declining phase of solar cycle 23 described by a generalized Tsallis PDF. J Geophys Res Space – Phys 109:A12107
48. Burlaga LF, Vinas AF (2005) Triangle for the entropic index q of non-extensive statistical mechanics observed by Voyager 1 in the distant heliosphere. Physica A 356:375–384
49. Burlaga LF, Ness NF, Acuna MH (2006) Multiscale structure of magnetic fields in the heliosheath. J Geophys Res Space – Phys 111:A09112
50. Borland L (2002) A theory of non-gaussian option pricing. Quant Finance 2:415–431
51. Borland L (2002) Closed form option pricing formulas based on a non-Gaussian stock price model with statistical feedback. Phys Rev Lett 89:098701
52. Borland L, Bouchaud J-P (2004) A non-Gaussian option pricing model with skew. Quant Finance 4:499–514
53. Cabral BJC, Tsallis C (2002) Metastability and weak mixing in classical long-range many-rotator system. Phys Rev E 66:065101(R)
54. Calabrese P, Cardy J (2004) JSTAT – J Stat Mech Theory Exp P06002
55. Callen HB (1985) Thermodynamics and An Introduction to Thermostatistics, 2nd edn. Wiley, New York
56. Chavanis PH (2006) Lynden-Bell and Tsallis distributions for the HMF model. Euro Phys J B 53:487–501
57. Chavanis PH (2006) Quasi-stationary states and incomplete violent relaxation in systems with long-range interactions. Physica A 365:102–107
58. Canosa N, Rossignoli R (2005) General non-additive entropic forms and the inference of quantum density operstors. Physica A 348:121–130
59. Cannas SA, Tamarit FA (1996) Long-range interactions and nonextensivity in ferromagnetic spin models. Phys Rev B 54:R12661–R12664
60. Caruso F, Tsallis C (2007) Extensive nonadditive entropy in quantum spin chains. In: Abe S, Herrmann HJ, Quarati P, Rapisarda A, Tsallis C (eds) Complexity, Metastability and Nonextensivity. American Institute of Physics Conference Proceedings, vol 965. New York, pp 51–59
61. Caruso F, Tsallis C (2008) Nonadditive entropy reconciles the area law in quantum systems with classical thermodynamics. Phys Rev E 78:021101
62. Caruso F, Pluchino A, Latora V, Vinciguerra S, Rapisarda A (2007) Analysis of self-organized criticality in the Olami–Feder–Christensen model and in real earthquakes. Phys Rev E 75:055101(R)
63. Campa A, Giansanti A, Moroni D (2002) Metastable states in a class of long-range Hamiltonian systems. Physica A 305:137–143
64. Csiszar I (1978) Information measures: A critical survey. In: Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, and the European Meeting of Statisticians, 1974. Reidel, Dordrecht
65. Cohen EGD (2005) Boltzmann and Einstein: Statistics and dynamics – An unsolved problem. Boltzmann Award Lecture at Statphys-Bangalore-2004. Pramana 64:635–643
66. Condat CA, Rangel J, Lamberti PW (2002) Anomalous diffusion in the nonasymptotic regime. Phys Rev E 65:026138
67. Coraddu M, Meloni F, Mezzorani G, Tonelli R (2004) Weak insensitivity to initial conditions at the edge of chaos in the logistic map. Physica A 340:234–239
68. Costa UMS, Lyra ML, Plastino AR, Tsallis C (1997) Power-law sensitivity to initial conditions within a logistic-like family

of maps: Fractality and nonextensivity. Phys Rev E 56:245–250
69. Curado EMF (1999) General aspects of the thermodynamical formalism. Braz J Phys 29:36–45
70. Curado EMF, Nobre FD (2003) Derivation of nonlinear Fokker-Planck equations by means of approximations to the master equation. Phys Rev E 67:021107
71. Curado EMF, Tsallis C (1991) Generalized statistical mechanics: connection with thermodynamics. Phys J A 24:L69-L72; [Corrigenda: 24:3187 (1991); 25:1019 (1992)]
72. Cvejic N, Canagarajah CN, Bull DR (2006) Image fusion metric based on mutual information and Tsallis entropy. Electron Lett 42:11
73. Daniels KE, Beck C, Bodenschatz E (2004) Defect turbulence and generalized statistical mechanics. Physica D 193:208–217
74. Daroczy Z (1970) Information and Control 16:36
75. de Albuquerque MP, Esquef IA, Mello ARG, de Albuquerque MP (2004) Image thresholding using Tsallis entropy. Pattern Recognition Lett 25:1059–1065
76. de Meneses MDS, da Cunha SD, Soares DJB, da Silva LR (2006) In: Sakagami M, Suzuki N, Abe S (eds) Complexity and Nonextensivity: New Trends in Statistical Mechanics. Prog Theor Phys Suppl 162:131–137
77. de Moura FABF, Tirnakli U, Lyra ML (2000) Convergence to the critical attractor of dissipative maps: Log-periodic oscillations, fractality and nonextensivity. Phys Rev E 62:6361–6365
78. de Oliveira HP, Soares ID, Tonini EV (2004) Role of the nonextensive statistics in a three-degrees of freedom gravitational system. Phys Rev D 70:084012
79. de Souza AMC, Tsallis C (1997) Student's t- and r-distributions: Unified derivation from an entropic variational principle. Physica A 236:52–57
80. de Souza J, Moyano LG, Queiros SMD (2006) On statistical properties of traded volume in financial markets. Euro Phys J B 50:165–168
81. Douglas P, Bergamini S, Renzoni F (2006) Tunable Tsallis distributions in dissipative optical lattices. Phys Rev Lett 96:110601
82. Fermi E (1936) Thermodynamics. Dover, New York, p 53
83. Ferri GL, Martinez S, Plastino A (2005) Equivalence of the four versions of Tsallis' statistics. J Stat Mech P04009
84. Ferro F, Lavagno A, Quarati P (2004) Non-extensive resonant reaction rates in astrophysical plasmas. Euro Phys J A 21:529–534
85. Fulco UL, da Silva LR, Nobre FD, Rego HHA, Lucena LS (2003) Effects of site dilution on the one-dimensional long-range bond-percolation problem. Phys Lett A 312:331–335
86. Frank TD (2005) Nonlinear Fokker–Planck Equations – Fundamentals and Applications. Springer, Berlin
87. Gervino G, Lavagno A, Quarati P (2005) *CNO* reaction rates and chemical abundance variations in dense stellar plasma. J Phys G 31:S1865–S1868
88. Gibbs JW (1902) Elementary Principles in Statistical Mechanics – Developed with Especial Reference to the Rational Foundation of Thermodynamics. C Scribner, New York; Yale University Press, New Haven, 1948; OX Bow Press, Woodbridge, Connecticut, 1981
89. Ginsparg P, Moore G (1993) Lectures on 2D Gravity and 2D String Theory. Cambridge University Press, Cambridge; hep-th/9304011, p 65
90. Grigera JR (1996) Extensive and non-extensive thermodynamics. A molecular dyanamic test. Phys Lett A 217:47–51
91. Hasegawa H (2006) Nonextensive aspects of small-world networks. Physica A 365:383–401
92. Havrda J, Charvat F (1967) Kybernetika 3:30
93. Hernandez H-S, Robledo A (2006) Fluctuating dynamics at the quasiperiodic onset of chaos, Tsallis q-statistics and Mori's q-phase thermodynamics. Phys A 370:286–300
94. Hilhorst HJ, Schehr G (2007) A note on *q*-Gaussians and non-Gaussians in statistical mechanics. J Stat Mech P06003
95. Jang S, Shin S, Pak Y (2003) Replica-exchange method using the generalized effective potential. Phys Rev Lett 91:058305
96. Jersblad J, Ellmann H, Stochkel K, Kastberg A, Sanchez L-P, Kaiser R (2004) Non-Gaussian velocity distributions in optical lattices. Phys Rev A 69:013410
97. Jund P, Kim SG, Tsallis C (1995) Crossover from extensive to nonextensive behavior driven by long-range interactions. Phys Rev B 52:50–53
98. Kaniadakis G (2001) Non linear kinetics underlying generalized statistics. Physica A 296:405–425
99. Kaniadakis G, Lissia M, Scarfone AM (2004) Deformed logarithms and entropies. Physica A 340:41–49
100. Khinchin AI (1953) Uspekhi Matem. Nauk 8:3 (Silverman RA, Friedman MD, trans. Math Found Inf Theory. Dover, New York)
101. Kronberger T, Leubner MP, van Kampen E (2006) Dark matter density profiles: A comparison of nonextensive statistics with *N*-body simulations. Astron Astrophys 453:21–25
102. Latora V, Baranger M (1999) Kolmogorov-Sinai entropy rate versus physical entropy. Phys Rev Lett 82:520–523
103. Latora V, Baranger M, Rapisarda A, Tsallis C (2000) The rate of entropy increase at the edge of chaos. Phys Lett A 273:97–103
104. Latora V, Rapisarda A, Tsallis C (2001) Non-Gaussian equilibrium in a long-range Hamiltonian system. Phys Rev E 64:056134
105. Lemes MR, Zacharias CR, Dal Pino A Jr (1997) Generalized simulated annealing: Application to silicon clusters. Phys Rev B 56:9279–9281
106. Lenzi EK, Anteneodo C, Borland L (2001) Escape time in anomalous diffusive media. Phys Rev E 63:051109
107. Lesche B (1982) Instabilities of Rényi entropies. J Stat Phys 27:419–422
108. Lindhard J, Nielsen V (1971) Studies in statistical mechanics. Det Kongelige Danske Videnskabernes Selskab Matematisk-fysiske Meddelelser (Denmark) 38(9):1–42
109. Lissia M, Quarati P (2005) Nuclear astrophysical plasmas: Ion distributions and fusion rates. Europhys News 36:211–214
110. Lutz E (2003) Anomalous diffusion and Tsallis statistics in an optical lattice. Phys Rev A 67:051402(R)
111. Lyra ML, Tsallis C (1998) Nonextensivity and multifractality in low-dimensional dissipative systems. Phys Rev Lett 80:53–56
112. Mann GM, Tsallis C (eds) (2004) Nonextensive Entropy – Interdisciplinary Applications. Oxford University Press, New York
113. Marsh JA, Fuentes MA, Moyano LG, Tsallis C (2006) Influence of global correlations on central limit theorems and entropic extensivity. Physica A 372:183–202
114. Martin S, Morison G, Nailon W, Durrani T (2004) Fast and accurate image registration using Tsallis entropy and simultaneous perturbation stochastic approximation. Electron Lett 40(10):20040375

115. Masi M (2005) A step beyond Tsallis and Renyi entropies. Phys Lett A 338:217–224
116. Mayoral E, Robledo A (2004) Multifractality and nonextensivity at the edge of chaos of unimodal maps. Physica A 340:219–226
117. Mayoral E, Robledo A (2005) Tsallis' *q* index and Mori's *q* phase transitions at edge of chaos. Phys Rev E 72:026209
118. Montemurro MA (2001) Beyond the Zipf–Mandelbrot law in quantitative linguistics. Physica A 300:567–578
119. Montemurro MA, Tamarit F, Anteneodo C (2003) Aging in an infinite-range Hamiltonian system of coupled rotators. Phys Rev E 67:031106
120. Moret MA, Pascutti PG, Bisch PM, Mundim MSP, Mundim KC (2006) Classical and quantum conformational analysis using Generalized Genetic Algorithm. Phys A 363:260–268
121. Moyano LG, Anteneodo C (2006) Diffusive anomalies in a long-range Hamiltonian system. Phys Rev E 74:021118
122. Moyano LG, Majtey AP, Tsallis C (2005) Weak chaos in large conservative system – Infinite-range coupled standard maps. In: Beck C, Benedek G, Rapisarda A, Tsallis C (eds) Complexity, Metastability and Nonextensivity. World Scientific, Singapore, pp 123–127
123. Moyano LG, Majtey AP, Tsallis C (2006) Weak chaos and metastability in a symplectic system of many long-range-coupled standard maps. Euro Phys J B 52:493–500
124. Moyano LG, Tsallis C, Gell-Mann M (2006) Numerical indications of a *q*-generalised central limit theorem. Europhys Lett 73:813–819
125. Nivanen L, Le Mehaute A, Wang QA (2003) Generalized algebra within a nonextensive statistics. Rep Math Phys 52:437–444
126. Nobre FD, Tsallis C (2003) Classical infinite-range-interaction Heisenberg ferromagnetic model: Metastability and sensitivity to initial conditions. Phys Rev E 68:036115
127. Nobre FD, Tsallis C (2004) Metastable states of the classical inertial infinite-range-interaction Heisenberg ferromagnet: Role of initial conditions. Physica A 344:587–594
128. Nobre FD, Curado EMF, Rowlands G (2004) A procedure for obtaining general nonlinear Fokker-Planck equations. Physica A 334:109–118
129. Oliveira HP, Soares ID (2005) Dynamics of black hole formation: Evidence for nonextensivity. Phys Rev D 71:124034
130. Penrose O (1970) Foundations of Statistical Mechanics: A Deductive Treatment. Pergamon Press, Oxford, p 167
131. Plastino AR, Plastino A (1995) Non-extensive statistical mechanics and generalized Fokker-Planck equation. Physica A 222:347–354
132. Pluchino A, Rapisarda A (2006) Metastability in the Hamiltonian Mean Field model and Kuramoto model. Physica A 365:184–189
133. Pluchino A, Rapisarda A (2006) Glassy dynamics and nonextensive effects in the *HMF* model: the importance of initial conditions. In: Sakagami M, Suzuki N, Abe S (eds) Complexity and Nonextensivity: New Trends in Statistical Mechanics. Prog Theor Phys Suppl 162:18–28
134. Pluchino A, Latora V, Rapisarda A (2004) Glassy dynamics in the HMF model. Physica A 340:187–195
135. Pluchino A, Latora V, Rapisarda A (2004) Dynamical anomalies and the role of initial conditions in the *HMF* model. Physica A 338:60–67
136. Pluchino A, Rapisarda A, Latora V (2005) Metastability and anomalous behavior in the HMF model: Connections to nonextensive thermodynamics and glassy dynamics. In: Beck C, Benedek G, Rapisarda A, Tsallis C (eds) Complexity, Metastability and Nonextensivity. World Scientific, Singapore, pp 102–112
137. Pluchino A, Rapisarda A, Tsallis C (2007) Nonergodicity and central limit behavior in long-range Hamiltonians. Europhys Lett 80:26002
138. Prato D, Tsallis C (1999) Nonextensive foundation of Levy distributions. Phys Rev E 60:2398–2401
139. Queiros SMD (2005) On non-Gaussianity and dependence in financial in time series: A nonextensive approach. Quant Finance 5:475–487
140. Queiros SMD, Tsallis C (2007) Nonextensive statistical mechanics and central limit theorems II – Convolution of *q*-independent random variables. In: Abe S, Herrmann HJ, Quarati P, Rapisarda A, Tsallis C (eds) Complexity, Metastability and Nonextensivity. American Institute of Physics Conference Proceedings, vol 965. New York, pp 21–33
141. Queiros SMD, Moyano LG, de Souza J, Tsallis C (2007) A nonextensive approach to the dynamics of financial observables. Euro Phys J B 55:161–168
142. Rapisarda A, Pluchino A (2005) Nonextensive thermodynamics and glassy behavior. Europhys News 36:202–206; Erratum: 37:25 (2006)
143. Rapisarda A, Pluchino A (2005) Nonextensive thermodynamics and glassy behaviour in Hamiltonian systems. Europhys News 36:202–206; Erratum: 37:25 (2006)
144. Rego HHA, Lucena LS, da Silva LR, Tsallis C (1999) Crossover from extensive to nonextensive behavior driven by long-range $d = 1$ bond percolation. Phys A 266:42–48
145. Renyi A (1961) In: Proceedings of the Fourth Berkeley Symposium, 1:547 University California Press, Berkeley; Renyi A (1970) Probability theory. North-Holland, Amsterdam
146. Robledo A (2004) Aging at the edge of chaos: Glassy dynamics and nonextensive statistics. Physica A 342:104–111
147. Robledo A (2004) Universal glassy dynamics at noise-perturbed onset of chaos: A route to ergodicity breakdown. Phys Lett A 328:467–472
148. Robledo A (2004) Criticality in nonlinear one-dimensional maps: RG universal map and nonextensive entropy. Physica D 193:153–160
149. Robledo A (2005) Intermittency at critical transitions and aging dynamics at edge of chaos. Pramana-J Phys 64:947–956
150. Robledo A (2005) Critical attractors and *q*-statistics. Europhys News 36:214–218
151. Robledo A (2006) Crossover from critical to chaotic attractor dynamics in logistic and circle maps. In: Sakagami M, Suzuki N, Abe S (eds) Complexity and Nonextensivity: New Trends in Statistical Mechanics. Prog Theor Phys Suppl 162:10–17
152. Robledo A, Baldovin F, Mayoral E (2005) Two stories outside Boltzmann-Gibbs statistics: Mori's *q*-phase transitions and glassy dynamics at the onset of chaos. In: Beck C, Benedek G, Rapisarda A, Tsallis C (eds) Complexity, Metastability and Nonextensivity. World Scientific, Singapore, p 43
153. Rodriguez A, Schwammle V, Tsallis C (2008) Strictly and asymptotically scale-invariant probabilistic models of *N* correlated binary random variables havin *q*-Gaussians as *N* -> infinity Limiting distributions. J Stat Mech P09006

154. Rohlf T, Tsallis C (2007) Long-range memory elementary 1D cellular automata: Dynamics and nonextensivity. Physica A 379:465–470
155. Rossignoli R, Canosa N (2003) Violation of majorization relations in entangled states and its detection by means of generalized entropic forms. Phys Rev A 67:042302
156. Rossignoli R, Canosa N (2004) Generalized disorder measure and the detection of quantum entanglement. Physica A 344:637–643
157. Salinas SRA, Tsallis C (eds) (1999) Nonextensive Statistical Mechanics and Thermodynamics. Braz J Phys 29(1)
158. Sampaio LC, de Albuquerque MP, de Menezes FS (1997) Nonextensivity and Tsallis statistic in magnetic systems. Phys Rev B 55:5611–5614
159. Santos RJV (1997) Generalization of Shannon' s theorem for Tsallis entropy. J Math Phys 38:4104–4107
160. Sato Y, Tsallis C (2006) In: Bountis T, Casati G, Procaccia I (eds) Complexity: An unifying direction in science. Int J Bif Chaos 16:1727–1738
161. Schutzenberger PM (1954) Contributions aux applications statistiques de la theorie de l' information. Publ Inst Statist Univ Paris 3:3
162. Shannon CE (1948) A Mathematical Theory of Communication. Bell Syst Tech J 27:379–423; 27:623–656; (1949) The Mathematical Theory of Communication. University of Illinois Press, Urbana
163. Sharma BD, Mittal DP (1975) J Math Sci 10:28
164. Serra P, Stanton AF, Kais S, Bleil RE (1997) Comparison study of pivot methods for global optimization. J Chem Phys 106:7170–7177
165. Soares DJB, Tsallis C, Mariz AM, da Silva LR (2005) Preferential Attachment growth model and nonextensive statistical mechanics. Europhys Lett 70:70–76
166. Son WJ, Jang S, Pak Y, Shin S (2007) Folding simulations with novel conformational search method. J Chem Phys 126:104906
167. Stigler SM (1999) Statistics on the table – The history of statistical concepts and methods. Harvard University Press, Cambridge
168. Silva AT, Lenzi EK, Evangelista LR, Lenzi MK, da Silva LR (2007) Fractional nonlinear diffusion equation, solutions and anomalous diffusion. Phys A 375:65–71
169. Tamarit FA, Anteneodo C (2005) Relaxation and aging in long-range interacting systems. Europhys News 36:194–197
170. Tamarit FA, Cannas SA, Tsallis C (1998) Sensitivity to initial conditions and nonextensivity in biological evolution. Euro Phys J B 1:545–548
171. Thistleton W, Marsh JA, Nelson K, Tsallis C (2006) unpublished
172. Thurner S (2005) Europhys News 36:218–220
173. Thurner S, Tsallis C (2005) Nonextensiveaspects of self-organized scale-free gas-like networks. Europhys Lett 72:197–204
174. Tirnakli U, Ananos GFJ, Tsallis C (2001) Generalization of the Kolmogorov–Sinai entropy: Logistic – like and generalized cosine maps at the chaos threshold. Phys Lett A 289:51–58
175. Tirnakli U, Beck C, Tsallis C (2007) Central limit behavior of deterministic dynamical systems. Phys Rev E 75:040106(R)
176. Tirnakli U, Tsallis C, Lyra ML (1999) Circular-like maps: Sensitivity to the initial conditions, multifractality and nonextensivity. Euro Phys J B 11:309–315
177. Tirnakli U, Tsallis C (2006) Chaos thresholds of the *z*-logistic map: Connection between the relaxation and average sensitivity entropic indices. Phys Rev E 73:037201
178. Tisza L (1961) Generalized Thermodynamics. MIT Press, Cambridge, p 123
179. Tonelli R, Mezzorani G, Meloni F, Lissia M, Coraddu M (2006) Entropy production and Pesin-like identity at the onset of chaos. Prog Theor Phys 115:23–29
180. Toscano F, Vallejos RO, Tsallis C (2004) Random matrix ensembles from nonextensive entropy. Phys Rev E 69:066131
181. Tsallis AC, Tsallis C, Magalhaes ACN, Tamarit FA (2003) Human and computer learning: An experimental study. Complexus 1:181–189
182. Tsallis C Regularly updated bibliography at http://tsallis.cat.cbpf.br/biblio.htm
183. Tsallis C (1988) Possible generalization of Boltzmann–Gibbs statistics. J Stat Phys 52:479–487
184. Tsallis C (2004) What should a statistical mechanics satisfy to reflect nature? Physica D 193:3–34
185. Tsallis C (2004) Dynamical scenario for nonextensive statistical mechanics. Physica A 340:1–10
186. Tsallis C (2005) Is the entropy S_q extensive or nonextensive? In: Beck C, Benedek G, Rapisarda A, Tsallis C (eds) Complexity, Metastability and Nonextensivity. World Scientific, Singapore
187. Tsallis C (2005) Nonextensive statistical mechanics, anomalous diffusion and central limit theorems. Milan J Math 73:145–176
188. Tsallis C, Bukman DJ (1996) Anomalous diffusion in the presence of external forces: exact time-dependent solutions and their thermostatistical basis. Phys Rev E 54:R2197–R2200
189. Tsallis C, Queiros SMD (2007) Nonextensive statistical mechanics and central limit theorems I – Convolution of independent random variables and *q*-product. In: Abe S, Herrmann HJ, Quarati P, Rapisarda A, Tsallis C (eds) Complexity, Metastability and Nonextensivity. American Institute of Physics Conference Proceedings, vol 965. New York, pp 8–20
190. Tsallis C, Souza AMC (2003) Constructing a statistical mechanics for Beck-Cohen superstatistics. Phys Rev E 67:026106
191. Tsallis C, Stariolo DA (1996) Generalized simulated annealing. Phys A 233:395–406; A preliminary version appeared (in English) as Notas de Fisica/CBPF 026 (June 1994)
192. Tsallis C, Levy SVF, de Souza AMC, Maynard R (1995) Statistical-mechanical foundation of the ubiquity of Levy distributions in nature. Phys Rev Lett 75:3589–3593; Erratum: (1996) Phys Rev Lett 77:5442
193. Tsallis C, Mendes RS, Plastino AR (1998) The role of constraints within generalized nonextensive statistics. Physica A 261:534–554
194. Tsallis C, Bemski G, Mendes RS (1999) Is re-association in folded proteins a case of nonextensivity? Phys Lett A 257:93–98
195. Tsallis C, Lloyd S, Baranger M (2001) Peres criterion for separability through nonextensive entropy. Phys Rev A 63:042104
196. Tsallis C, Anjos JC, Borges EP (2003) Fluxes of cosmic rays: A delicately balanced stationary state. Phys Lett A 310:372–376
197. Tsallis C, Anteneodo C, Borland L, Osorio R (2003) Nonextensive statistical mechanics and economics. Physica A 324:89–100
198. Tsallis C, Mann GM, Sato Y (2005) Asymptotically scale-invariant occupancy of phase space makes the entropy S_q extensive. Proc Natl Acad Sci USA 102:15377–15382

199. Tsallis C, Mann GM, Sato Y (2005) Extensivity and entropy production. In: Boon JP, Tsallis C (eds) Nonextensive Statistical Mechanics: New Trends, New perspectives. Europhys News 36:186–189
200. Tsallis C, Rapisarda A, Pluchino A, Borges EP (2007) On the non-Boltzmannian nature of quasi-stationary states in long-range interacting systems. Physica A 381:143–147
201. Tsekouras GA, Tsallis C (2005) Generalized entropy arising from a distribution of q-indices. Phys Rev E 71:046144
202. Umarov S, Tsallis C (2007) Multivariate generalizations of the q–central limit theorem. cond-mat/0703533
203. Umarov S, Tsallis C, Steinberg S (2008) On a q-central limit theorem consistent with nonextensive statistical mechanics. Milan J Math 76. doi:10.1007/s00032-008-0087-y
204. Umarov S, Tsallis C, Gell-Mann M, Steinberg S (2008) Symmetric (q, α)-stable distributions. Part I: First representation. cond-mat/0606038v2
205. Umarov S, Tsallis C, Gell-Mann M, Steinberg S (2008) Symmetric (q, α)-stable distributions. Part II: Second representation. cond-mat/0606040v2
206. Upadhyaya A, Rieu J-P, Glazier JA, Sawada Y (2001) Anomalous diffusion and non-Gaussian velocity distribution of Hydra cells in cellular aggregates. Physica A 293:549–558
207. Vajda I (1968) Kybernetika 4:105 (in Czech)
208. Varotsos PA, Sarlis NV, Tanaka HK, Skordas ES (2005) Some properties of the entropy in the natural time. Phys Rev E 71:032102
209. Wehrl A (1978) Rev Modern Phys 50:221
210. Weinstein YS, Lloyd S, Tsallis C (2002) Border between between regular and chaotic quantum dynamics. Phys Rev Lett 89:214101
211. Weinstein YS, Tsallis C, Lloyd S (2004) On the emergence of nonextensivity at the edge of quantum chaos. In: Elze H-T (ed) Decoherence and Entropy in Complex Systems. Lecture notes in physics, vol 633. Springer, Berlin, pp 385–397
212. White DR, Kejzar N, Tsallis C, Farmer JD, White S (2005) A generative model for feedback networks. Phys Rev E 73:016119
213. Wilk G, Wlodarczyk Z (2004) Acta Phys Pol B 35:871–879
214. Wu JL, Chen HJ (2007) Fluctuation in nonextensive reaction-diffusion systems. Phys Scripta 75:722–725
215. Yamano T (2004) Distribution of the Japanese posted land price and the generalized entropy. Euro Phys J B 38:665–669
216. Zanette DH, Alemany PA (1995) Thermodynamics of anomalous diffusion. Phys Rev Lett 75:366–369

Entropy in Ergodic Theory

Jonathan L. F. King
University of Florida, Gainesville, USA

Article Outline

Glossary
Definition of the Subject
Entropy example: *How many questions?*
Distribution Entropy
A Gander at Shannon's Noisy Channel Theorem
The Information Function
Entropy of a Process
Entropy of a Transformation
Determinism and Zero-Entropy
The Pinsker–Field and K-Automorphisms
Ornstein Theory
Topological Entropy
Three Recent Results
Exodos
Bibliography

Glossary

Some of the following definitions refer to the "Notation" paragraph immediately below. Use ***mpt*** for '*measure-preserving transformation*'.

Measure space A ***measure space*** $(X, \mathcal{X}, \mu)$ is a set X, a ***field*** (that is, a σ-algebra) $\mathcal{X}$ of subsets of X, and a countably-additive measure $\mu : \mathcal{X} \to [0, \infty]$. (We often just write (X, μ), with the field implicit.) For a collection $\mathcal{C} \subset \mathcal{X}$, use $\mathrm{Fld}(\mathcal{C}) \supset \mathcal{C}$ for the smallest field including $\mathcal{C}$. The number $\mu(B)$ is the "***μ-mass of*** B".

Measure-preserving map A ***measure-preserving map*** $\psi : (X, \mathcal{X}, \mu) \to (Y, \mathcal{Y}, \nu)$ is a map $\psi : X \to Y$ such that the inverse image of each $B \in \mathcal{Y}$ is in $\mathcal{X}$, and $\mu(\psi^{-1}(B)) = \nu(B)$. A ***(measure-preserving)*** *transformation* is a measure-preserving map $T : (X, \mathcal{X}, \mu) \to (X, \mathcal{X}, \mu)$. Condense this notation to $(T : X, \mathcal{X}, \mu)$ or $(T : X, \mu)$.

Probability space A ***probability space*** is a measure space (X, μ) with $\mu(X) = 1$; this μ is a ***probability measure***. All our maps/transformations in this article are on probability spaces.

Factor map A ***factor map***

$$\psi : (T : X, \mathcal{X}, \mu) \to (S : Y, \mathcal{Y}, \nu)$$

is a measure-preserving map $\psi : X \to Y$ which intertwines the transformations, $\psi \circ T = S \circ \psi$. And ψ is an ***isomorphism*** if – after deleting a nullset (a mass-zero set) in each space – this ψ is a bijection and ψ^{-1} is also a factor map.

Almost everywhere (a.e.) A measure-theoretic statement holds ***almost everywhere***, abbreviated ***a.e.***, if it holds off of a nullset. (Eugene Gutkin once remarked to me that the problem with Measure Theory is … that you have to say "almost everywhere", almost everywhere.) For example, $B \overset{a.e.}{\supset} A$ means that $\mu(B \setminus A)$ is zero. The *a.e.* will usually be implicit.

Probability vector A ***probability vector*** $\vec{v} = (v_1, v_2, \dots)$ is a list of non-negative reals whose sum is 1. We generally assume that probability vectors and partitions (see

below) have *finitely* many components. Write "*countable probability vector/partition*", when finitely or denumerably many components are considered.

Partition A ***partition*** $\mathsf{P} = (A_1, A_2, \dots)$ splits X into pairwise disjoint subsets $A_i \in \mathcal{X}$ so that the disjoint union $\bigsqcup_i A_i$ is all of X. Each A_i is an ***atom*** of P. Use $|\mathsf{P}|$ or ${}^{\#}\mathsf{P}$ for the number of atoms. When P partitions a probability space, then it yields a probability vector $\vec{v}$, where $v_j := \mu(A_j)$. Lastly, use $\mathsf{P}\langle x\rangle$ to denote the P-atom that owns x.

Fonts We use the font $\mathcal{H}, \mathcal{E}, \mathcal{I}$ for *distribution-entropy*, *entropy* and the *information function*. In contrast, the script font $\mathcal{ABC}\dots$ will be used for collections of sets; usually subfields of $\mathcal{X}$. Use $\mathbb{E}(\cdot)$ for the (conditional) expectation operator.

Notation $\mathbb{Z}$ = integers, $\mathbb{Z}_+$ = positive integers, and $\mathbb{N}$ = natural numbers $= 0, 1, 2, \dots$. (Some well-meaning folk use $\mathbb{N}$ for $\mathbb{Z}_+$, saying '*Nothing could be more natural than the positive integers*'. And this is why $0 \in \mathbb{N}$. Use $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$ for the ***ceiling*** and ***floor*** functions; $\lfloor\cdot\rfloor$ is also called the **"greatest-integer function"**. For an interval $J := [a, b) \subset [-\infty, +\infty]$, let $[a \dots b)$ denote the ***interval of integers*** $J \cap \mathbb{Z}$ (with a similar convention for closed and open intervals). E. g., $(\mathrm{e} \dots \pi] = (\mathrm{e} \dots \pi) = \{3\}$.
For subsets A and B of the same space, Ω, use $A \subset B$ for inclusion and $A \subsetneqq B$ for *proper* inclusion. The difference set $B \smallsetminus A$ is $\{\omega \in B \mid \omega \notin A\}$. Employ A^c for the complement $\Omega \smallsetminus A$. Since we work in a probability space, if we let $x := \mu(A)$, then a convenient convention is to have

$$x^c \quad \text{denote} \quad 1 - x\,,$$

since then $\mu(A^c)$ equals x^c.
Use $A \triangle B$ for the ***symmetric difference*** $[A \smallsetminus B] \cup [B \smallsetminus A]$. For a collection $\mathcal{C} = \{E_j\}_j$ of sets in Ω, let the ***disjoint union*** $\bigsqcup_j E_j$ or $\bigsqcup(\mathcal{C})$ represent the union $\bigcup_j E_j$ *and also assert* that the sets are pairwise disjoint.
Use "$\forall_{\text{large}} n$" to mean: "$\exists n_0$ such that $\forall n > n_0$". To refer to left hand side of an Eq. (20), use LhS(20); do analogously for RhS(20), the right hand side.

Definition of the Subject

The word '*entropy*' (originally German, *Entropie*) was coined by Rudolf Julius Emanuel Clausius circa 1865 [2,3], taken from the Greek $\epsilon\nu\tau\rho o\pi\iota\alpha$, '*a turning towards*'. This article thus begins (*Prolegomenon*, "introduction") and ends (*Exodos*[1], "the path out") in Greek.

[1] This is the Greek spelling.

Clausius, in his coinage, was referring to the thermodynamic notion in physics. Our focus in this article, however, will be the concept in measurable and topological dynamics. (Entropy in differentiable dynamics[2] would require an article by itself.) Shannon's 1948 paper [6] on Information Theory, then Kolmogorov's [4] and Sinai's [7] generalization to dynamical systems, will be our starting point. Our discussion will be of the one-dimensional case, where the acting-group is $\mathbb{Z}$.

> "My greatest concern was what to call it. I thought of calling it '*information*', but the word was overly used, so I decided to call it '*uncertainty*'. John von Neumann had a better idea, he told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function goes by that name in statistical mechanics. In the second place, and more important, **nobody knows what entropy really is**, so in a debate you will always have the advantage.'" (Shannon as quoted in [59])

Entropy example: *How many questions?*

Imagine a dartboard, Fig. 1, split in five regions $A, \dots, E$ with known probabilities. Blindfolded, you throw a dart at the board. What is the expected number V of Yes/No questions needed to ascertain the region in which the dart landed?

Solve this by always dividing the remaining probability in half. '*Is it A?*' – if Yes, then $V = 1$. Else: '*Is it B or C?*' – if Yes, then '*Is it B?*' – if No, then the dart landed in C, and $V = 3$ was the number of questions. Evidently $V = 3$ also for regions B, D, E. Using "log" to denote base-2 logarithm[3], the expected number of questions[4] is thus

$$\begin{aligned}\mathbb{E}(V) &= \frac{1}{2}\cdot 1 + \frac{1}{8}\cdot 3 + \frac{1}{8}\cdot 3 + \frac{1}{8}\cdot 3 + \frac{1}{8}\cdot 3 \\ &= \sum_{j=0}^{4} p_j \log\left(\frac{1}{p_j}\right) \overset{\text{note}}{=} 2\,. \qquad (1)\end{aligned}$$

[2] For instance, see [15,18,24,25].

[3] In this paper, unmarked logs will be to base-2. In entropy theory, it does not matter much what base is used, but base-2 is convenient for computing entropy for messages described in bits. When using the natural logarithm, some people refer to the unit of information as a ***nat***. In this paper, I have picked bits rather than nats.

[4] This is holds when each probability p is a reciprocal power of two. For general probabilities, the "expected number of questions" interpretation holds in a weaker sense: Throw N darts independently at N copies of the dartboard. Efficiently ask Yes/No questions to determine where *all* N darts landed. Dividing by N, then sending $N \to \infty$, will be the $p \log\left(\frac{1}{p}\right)$ sum of Eq. 1.

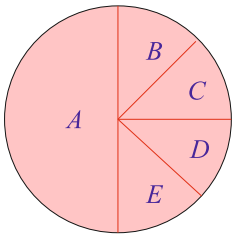

Entropy in Ergodic Theory, Figure 1
This dartboard is a probability space with a 5-set partition. The atoms have probabilities $\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}$. This probability distribution will be used later in Meshalkin's example

Letting $\vec{v} := \left(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right)$ be the probability vector, we can write this expectation as

$$\mathbb{E}(V) = \sum_{x \in \vec{v}} \eta(x) .$$

Here, $\eta : [0,1] \to [0,\infty)$ is the important function[5]

$$\eta(x) := x \log(1/x) , \quad \text{so extending by continuity gives } \eta(0) = 0 . \tag{2}$$

An interpretation of "$\eta(x)$" is the number of questions needed to winnow down to an event of probability x.

Distribution Entropy

Given a probability vector $\vec{v}$, define its ***distribution entropy*** as

$$\mathcal{H}(\vec{v}) := \sum_{x \in \vec{v}} \eta(x) . \tag{3}$$

This article will use the term ***distropy*** for '*distribution entropy*', reserving the word ***entropy*** for the corresponding dynamical concept, when there is a notion of *time* involved. Getting ahead of ourselves, the *entropy* of a stationary process is the asymptotic average value that its distropy decays to, as we look at larger and larger finite portions of the process.

An equi-probable vector $\vec{v} := \left(\frac{1}{K}, \overset{K}{\ldots}, \frac{1}{K}\right)$ evidently has $\mathcal{H}(\vec{v}) = \log(K)$. On a probability space, the **"distropy of partition** P**"**, written $\mathcal{H}(\mathsf{P})$ or $\mathcal{H}(A_1, A_2, \ldots)$ shall mean the distropy of probability vector $j \mapsto \mu(A_j)$.

A (finite) partition necessarily has finite distropy. A *countable* partition can have finite distropy, e. g. $\mathcal{H}\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots\right) = 2$. One could also have infinite distropy: Consider a piece $B \subset X$ of mass $1/2^N$. Splitting B into 2^k many equal-mass atoms gives an η-sum of $2^k(k+N)/(2^k 2^N)$. Setting $k = k_N := 2^N - N$ makes this η-sum equal 1; so splitting the pieces of $X = \bigsqcup_{N=1}^{\infty} B_N$, with $\mu(B_N) = \frac{1}{2^N}$, yields an ∞-distropy partition.

Function η

The $\eta(x) = x \log(1/x)$ function[6] has vertical tangent at $x = 0$, maximum at $1/e$ and, when graphed in nats slope -1 at $x = 1$.

Consider partitions P and Q on the same space (X, μ). Their ***join***, written $\mathsf{P} \vee \mathsf{Q}$, has atoms $A \cap B$, for each pair $A \in \mathsf{P}$ and $B \in \mathsf{Q}$. They are ***independent***, written $\mathsf{P} \perp \mathsf{Q}$ if $\mu(A \cap B) = \mu(A)\mu(B)$ for each A, B pair. We write $\mathsf{P} \succcurlyeq \mathsf{Q}$, and say that "$\mathsf{P}$ ***refines*** Q", if each P-atom is a subset of some Q-atom. Consequently, each Q-atom is a union of P-atoms.

Recall, for δ a real number, our convention that δ^c means $1 - \delta$, in analogy with $\mu(B^c)$ equaling $1 - \mu(B)$ on a probability space.

Distropy Fact

For partitions $\mathsf{P}, \mathsf{Q}, \mathsf{R}$ on probability space (X, μ):

(a) $\mathcal{H}(\mathsf{P}) \le \log(^{\#}\mathsf{P})$, with equality IFF P is an equi-mass partition.
(b) $\mathcal{H}(\mathsf{Q} \vee \mathsf{R}) \le \mathcal{H}(\mathsf{Q}) + \mathcal{H}(\mathsf{R})$, with equality IFF $\mathsf{Q} \perp \mathsf{R}$.
(c) For $\delta \in \left[0, \frac{1}{2}\right]$, the function $\delta \mapsto \mathcal{H}(\delta, \delta^c)$ is strictly increasing.
(d) $\mathsf{R} \preccurlyeq \mathsf{P}$ implies $\mathcal{H}(\mathsf{R}) \le \mathcal{H}(\mathsf{P})$, with equality IFF $\mathsf{R} \overset{a.e.}{=} \mathsf{P}$.

Proof Use the strict concavity of $\eta(\cdot)$, together with Jensen's inequality. □

Remark 1 Although we will not discuss it in this paper, most distropy statements remain true with '*partition*' replaced by '*countable partition of finite distropy*'.

[5]There does not seem to be a standard name for this function. I use η, since an uppercase η looks like an H, which is the letter that Shannon used to denote what I am calling distribution-entropy.

[6]Curiosity: Just in this paragraph we compute distropy in ***nats***, that is, using natural logarithm. Given a small probability $p \in [0,1]$ and setting $x := 1/p$, note that $\eta(p) = \frac{\log(x)}{x} \approx 1/\pi(x)$, where $\pi(x)$ denotes the number of prime numbers less-equal x. (This approximation is a weak form of the Prime Number Theorem.) Is there any actual connection between the '*approximate distropy*' function $\mathcal{H}_\pi(\vec{\mathbf{p}}) := \sum_{p \in \vec{\mathbf{p}}} 1/\pi(1/p)$ and Number Theory, other than a coincidence of growth rate?

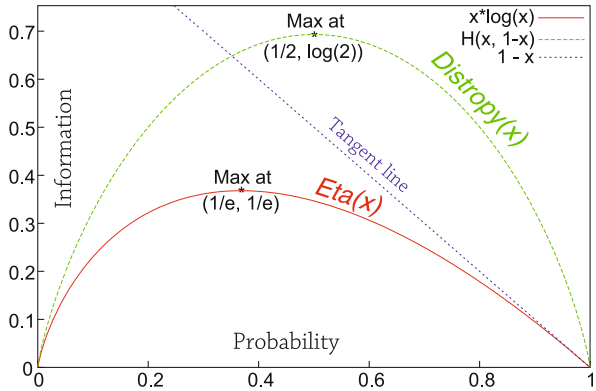

Entropy in Ergodic Theory, Figure 2
Using natural log, here are the graphs of: $\eta(x)$ in *solid red*, $\mathcal{H}(x, x^c)$ in *dashed green*, $1-x$ in *dotted blue*. Both $\eta(x)$ and $\mathcal{H}(x, x^c)$ are strictly convex-down. The $1-x$ line is tangent to $\eta(x)$ at $x = 1$

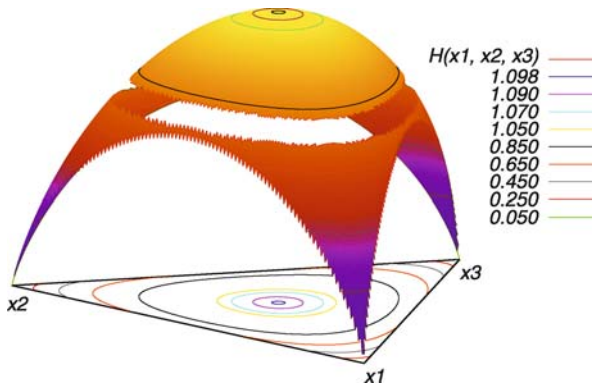

Entropy in Ergodic Theory, Figure 3
Using natural log: The graph of $\mathcal{H}(x_1, x_2, x_3)$ in barycentric coordinates; a slice has been removed, between $z = 0.745$ and $z = 0.821$. The three arches are copies of the distropy curve from Fig. 2.

Binomial Coefficients

The dartboard gave an example where distropy arises in a natural way. Here is a second example.

For a small $\delta > 0$, one might guess that the binomial coefficient $\binom{n}{\delta n}$ grows asymptotically (as $n \to \infty$) like $2^{\varepsilon n}$, for some small ε. But what is the correct relation between ε and δ?

Well, Stirling's formula $n! \approx [n/\mathrm{e}]^n$ gives

$$\frac{n!}{[\delta n]![\delta^c n]!} \approx \frac{n^n}{[\delta n]^{[\delta n]}[\delta^c n]^{[\delta^c n]}} = \frac{1}{[\delta^{\delta n}[\delta^c]^{\delta^c n}]} \quad (\text{recall } \delta^c = 1-\delta)\,.$$

Thus $\frac{1}{n}\log\binom{n}{\delta n} \approx \mathcal{H}(\delta, \delta^c)$. But by means of the above distropy inequalities, we get an inequality true for *all* n, not just asymptotically.

Lemma 2 (Binomial Lemma) *Fix a $\delta \in \left[0, \frac{1}{2}\right]$ and let $\mathbf{H} := \mathcal{H}(\delta, \delta^c)$. Then for each $n \in \mathbb{Z}_+$:*

$$\sum_{j \in [0 \ldots \delta n]} \binom{n}{j} \le 2^{\mathbf{H}n}\,. \tag{4}$$

Proof Let $X \subset \{0,1\}^n$ be the set of $\vec{x}$ with

$$^{\#}\{i \in [1 \ldots n] \mid x_i = 1\} \le \delta n\,.$$

On X, let $\mathsf{P}_1, \mathsf{P}_2, \ldots$ be the coordinate partitions; e. g. $\mathsf{P}_7 = (A_7, A_7^c)$, where $A_7 := \{\vec{x} \mid x_7 = 1\}$. Weighting each point by $\frac{1}{|X|}$, the uniform distribution on X, gives that $\mu(A_7) \le \delta$. So $\mathcal{H}(\mathsf{P}_7) \le \mathbf{H}$, by (c) in Sect. "Distropy Fact".

Finally, the join $\mathsf{P}_1 \vee \cdots \vee \mathsf{P}_n$ separates the points of X. So

$$\begin{aligned}\log(^{\#}X) &= \mathcal{H}(\mathsf{P}_1 \vee \cdots \vee \mathsf{P}_n)\\ &\le \mathcal{H}(\mathsf{P}_1) + \ldots + \mathcal{H}(\mathsf{P}_n)\\ &\le \mathbf{H}n\,,\end{aligned}$$

making use of (a),(b) in "Distropy Fact". And $^{\#}X$ equals LhS (4). □

A Gander at Shannon's Noisy Channel Theorem

We can restate the Binomial lemma using the ***Hamming metric*** on $\{0,1\}^n$,

$$\mathrm{Dist}(\vec{x}, \vec{y}) := {}^{\#}\{i \in [1 \ldots n] \mid x_i \ne y_i\}\,.$$

Use $\mathrm{Bal}(\vec{x}, r)$ for the open radius-r ball centered at $\vec{x}$, and

$$\overline{\mathrm{Bal}}(\vec{x}, r) := \{\vec{y} \mid \mathrm{Dist}(\vec{x}, \vec{y}) \le r\}$$

for the closed ball. The above lemma can be interpreted as saying that

$$|\overline{\mathrm{Bal}}(\vec{x}, \delta n)| \le 2^{\mathcal{H}(\delta, \delta^c)n}\,, \quad \text{for each} \quad \vec{x} \in \{0,1\}^n\,. \tag{5}$$

Corollary 3 *Fix $n \in \mathbb{Z}_+$ and $\delta \in \left[0, \frac{1}{2}\right]$, and let*

$$\mathbf{H} := \mathcal{H}(\delta, \delta^c)\,.$$

*Then there is a set $C \subset \{0,1\}^n$, with $^{\#}C \ge 2^{[1-\mathbf{H}]n}$, that is **strongly** δn-**separated**. I. e., $\mathrm{Dist}(\vec{x}, \vec{y}) > \delta n$ for each distinct pair $\vec{x}, \vec{y} \in C$.*

Noisy Channel

Shannon's theorem says that a noisy channel has a ***channel capacity***. Transmitting *above* this speed, there is a minimum error-rate (depending how much "*above*") that no error-correcting code can fix. Conversely, one can transmit *below* – but arbitrarily close to – the channel capacity, and encode the data so as to make the error-rate less than any given ε. We use Corollary 3 to show the existence of such codes, in the simplest case where the noise[7] is a binary independent-process (a "Bernoulli" process, in the language later in this article).

We have a channel which can pass one bit per second. Alas, there is a fixed noise-probability $\nu \in [0, \frac{1}{2})$ so that a bit in the channel is perturbed into the other value. Each perturbation is independent of all others. Let $\mathbf{H} := \mathcal{H}(\nu, \nu^c)$. The value $[1 - \mathbf{H}]$ bits-per-second is the ***channel capacity*** of this noise-afflicted channel.

Encoding/Decoding Encode using an **"k, n–block-code"**; an injective map $F\colon \{0,1\}^k \to \{0,1\}^n$. The source text is split into consecutive k-bit blocks. A block $\vec{x} \in \{0,1\}^k$ is encoded to $F(\vec{x}) \in \{0,1\}^n$ and then sent through the channel, where it comes out perturbed to $\vec{\alpha} \in \{0,1\}^n$. The *transmission rate* is thus k/n bits per second.

For this example, we fix a radius $r > 0$ to determine the decoding map,

$$D_r\colon \{0,1\}^n \to \{Oops\} \sqcup \{0,1\}^k .$$

We set $D_r(\vec{\alpha})$ to $\vec{z}$ if there is a *unique* $\vec{z}$ with $F(\vec{z}) \in \overline{\mathrm{Bal}}(\vec{\alpha}, r)$; else, set $D_r(\vec{\alpha}) := Oops$.

One can think of the noise as a $\{0,1\}$-independent-process, with $\mathrm{Prob}(1) = \nu$, which is added mod-2 to the signal-process. Suppose we can arrange that the set $\{F(\vec{x})) \mid \vec{x} \in \{0,1\}^k\}$ of codewords, is a strongly r-separated-set. Then

> The probability that a block is mis-decoded is the probability, flipping a ν-coin n times that we get more than r many Heads. (6)

Theorem 4 (Shannon) *Fix a noise-probability $\nu \in [0, \frac{1}{2})$ and let $\mathbf{H} := \mathcal{H}(\nu, \nu^c)$. Consider a rate $R < [1 - \mathbf{H}]$ and an $\varepsilon > 0$. Then $\forall_{\text{large}} n$ there exists a k and a code $F\colon \{0,1\}^k \to \{0,1\}^n$ so that: The F-code transmits bits at faster than R bits-per-second, and with error-rate $< \varepsilon$.*

[7] The noise-process is assumed to be *independent* of the signal-process. In contrast, when the perturbation is highly dependent on the signal, then it is sometimes called ***distortion***.

Proof Let $\mathbf{H}' := \mathcal{H}(\delta, \delta^c)$, where $\delta > \nu$ was chosen so close to ν that

$$\delta < \frac{1}{2} \quad \text{and} \quad 1 - \mathbf{H}' > R . \tag{7}$$

Pick a large n for which

$$\frac{k}{n} > R , \quad \text{where} \quad k := \lfloor [1 - \mathbf{H}']n \rfloor . \tag{8}$$

By Corollary 3, there is a strongly δn-separated-set $C \subset \{0,1\}^n$ with ${}^{\#}C \geq 2^{[1-\mathbf{H}']n}$. So C is big enough to permit an injection $F : \{0,1\}^k \to C$. Courtesy Eq. (6), the probability of a decoding error is that of getting more than δn many Heads in flipping a ν-coin n times. Since $\delta > \nu$, the Weak Law of Large Numbers guarantees – once n is large enough – that this probability is less than the given ε. □

The Information Function

Agree to use $\mathsf{P} = (A_1, \ldots), \mathsf{Q} = (B_1, \ldots), \mathsf{R} = (C_1, \ldots)$ for partitions, and $\mathcal{F}, \mathcal{G}$ for fields.

With $\mathfrak{C}$ a (finite or infinite) family of subfields of $\mathcal{X}$, their ***join*** $\bigvee_{\mathcal{G} \in \mathfrak{C}} \mathcal{G}$ is the smallest field $\mathcal{F}$ such that $\mathcal{G} \subset \mathcal{F}$, for each $\mathcal{G} \in \mathfrak{C}$. A partition Q can be interpreted also as a field; namely, the field of unions of its atoms. A join of denumerably many partitions will be interpreted as a field, but a join of *finitely* many, $\mathsf{P}_1 \vee \cdots \vee \mathsf{P}_N$, will be viewed as a partition *or* as a field, depending on context.

Conditioning a partition P on a positive-mass set B, let $\mathsf{P}|B$ be the probability vector $A \mapsto \frac{\mu(A \cap B)}{\mu(B)}$. Its distropy is

$$\mathcal{H}(\mathsf{P}|B) = \sum_{A \in \mathsf{P}} \log\left(\frac{1}{\mu(A \cap B)/\mu(B)}\right) \frac{\mu(A \cap B)}{\mu(B)} .$$

So conditioning P on a *partition* Q gives conditional distropy

$$\begin{aligned} \mathcal{H}(\mathsf{P}|\mathsf{Q}) &= \sum_{B \in \mathsf{Q}} \mathcal{H}(\mathsf{P}|B)\mu(B) \\ &= \sum_{A \in \mathsf{P}, B \in \mathsf{Q}} \log\left(\frac{1}{\mu(A \cap B)/\mu(B)}\right)\mu(A \cap B) . \end{aligned} \tag{9}$$

A "dartboard" interpretation of $\mathcal{H}(\mathsf{P}|\mathsf{Q})$ is

> The expected number of questions to ascertain the P-atom that a random dart $x \in X$ fell in, given that we are told its Q-atom.

For a set $A \subset X$, use $\mathbf{1}_A : X \to \{0,1\}$ for its *indicator function*; $\mathbf{1}_A(x) = 1$ IFF $x \in A$. The ***information function*** of partition P, a map $\mathcal{I}_{\mathsf{P}} : X \to [0,\infty)$, is

$$\mathcal{I}_{\mathsf{P}} := \sum_{A\in\mathsf{P}} \log\left(\frac{1}{\mu(A)}\right)\mathbf{1}_A(\cdot)\,. \tag{10}$$

The information function has been defined so that its expectation is the distropy of P,

$$\mathbb{E}(\mathcal{I}_{\mathsf{P}}) = \int_X \mathcal{I}_{\mathsf{P}}(\cdot)\mathrm{d}\mu = \mathcal{H}(\mathsf{P})\,.$$

Conditioning on a Field

For a subfield $\mathcal{F} \subset \mathcal{X}$, recall that each function $g \in \mathbb{L}^1(\mu)$ has a *conditional expectation* $\mathbb{E}(g|\mathcal{F}) \in \mathbb{L}^1(\mu)$. It is the *unique* $\mathcal{F}$-measurable function with

$$\forall B \in \mathcal{F}: \quad \int_B \mathbb{E}(g|\mathcal{F})\mathrm{d}\mu = \int_B g\mathrm{d}\mu\,.$$

Returning to distropy ideas, use $\mu(A|\mathcal{F})$ for the ***conditional probability*** function; it is the conditional expectation $\mathbb{E}(\mathbf{1}_A|\mathcal{F})$. So the ***conditional information function*** is

$$\mathcal{I}_{\mathsf{P}|\mathcal{F}}(x) := \sum_{A\in\mathsf{P}} \log\left(\frac{1}{\mu(A|\mathcal{F}(x))}\right)\mathbf{1}_A(x)\,. \tag{11}$$

Its integral

$$\mathcal{H}(\mathsf{P}|\mathcal{F}) := \int \mathcal{I}_{\mathsf{P}|\mathcal{F}}\mathrm{d}\mu\,,$$

is the ***conditional distropy*** of P on $\mathcal{F}$.

When $\mathcal{F}$ is the field of unions of atoms from some partition Q, then the number $\mathcal{H}(\mathsf{P}|\mathcal{F})$ equals the $\mathcal{H}(\mathsf{P}|\mathsf{Q})$ from Eq. (9).

Write $\mathcal{G}_j \nearrow \mathcal{F}$ to indicate that fields $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \ldots$ are nested, and that $\mathrm{Fld}\left(\bigcup_1^\infty \mathcal{G}_j\right) = \mathcal{F}$, a.e. The Martingale Convergence Theorem (p. 103 in [20]) gives (c) below.

Conditional-Distropy Fact Consider partitions $\mathsf{P}, \mathsf{Q}, \mathsf{R}$ and fields $\mathcal{F}$ and $\mathcal{G}_j$. Then

(a) $0 \le \mathcal{H}(\mathsf{P}|\mathcal{F}) \le \mathcal{H}(\mathsf{P})$, with equality IFF $\mathsf{P} \overset{a.e.}{\subset} \mathcal{F}$, respectively, $\mathsf{P}\perp\mathcal{F}$.
(b) $\mathcal{H}(\mathsf{Q}\vee\mathsf{R}|\mathcal{F}) \le \mathcal{H}(\mathsf{Q}|\mathcal{F}) + \mathcal{H}(\mathsf{R}|\mathcal{F})$.
(c) Suppose $\mathcal{G}_j \nearrow \mathcal{F}$. Then $\mathcal{H}(\mathsf{P}|\mathcal{G}_j) \searrow \mathcal{H}(\mathsf{P}|\mathcal{F})$.
(d) $\mathcal{H}(\mathsf{Q}\vee\mathsf{R}) = \mathcal{H}(\mathsf{Q}|\mathsf{R}) + \mathcal{H}(\mathsf{R})$.
(d') $\mathcal{H}(\mathsf{Q}\vee\mathsf{R}_1|\mathsf{R}_0) = \mathcal{H}(\mathsf{Q}|\mathsf{R}_1\vee\mathsf{R}_0) + \mathcal{H}(\mathsf{R}_1|\mathsf{R}_0)$.

Imagining our dartboard (Fig. 1) divided by superimposed partitions Q and R, equality (d) can interpreted as saying: '*You can efficiently discover where the dart landed in both partitions, by first asking efficient questions about* R, *then – based on where it landed in* R *– asking intelligent questions about* Q.'

Entropy of a Process

Consider a transformation $(T\colon X,\mu)$ and partition $\mathsf{P} = (A_1, A_2, \ldots)$. Each "time" n determines a partition $\mathsf{P}_n := T^n\mathsf{P}$, whose jth-atom is $T^{-n}(A_j)$. The ***process*** T,P refers to how T acts on the subfield $\bigvee_0^\infty \mathsf{P}_n \subset \mathcal{X}$. (An alternative view of a process is as a stationary sequence $V_0, V_1, \ldots$ of random variables $V_n\colon X \to \mathbb{Z}_+$, where $V_n(x) := j$ because x is in the jth-atom of P_n.)

Write $\mathcal{E}(T,\mathsf{P})$ or $\mathcal{E}^T(\mathsf{P})$ for the "***entropy***" of the T,P process". It is the limit of the ***conditional-distropy-numbers***

$$c_n := \mathcal{H}(\mathsf{P}_0|\mathsf{P}_1\vee\mathsf{P}_2\vee\cdots\vee\mathsf{P}_{n-1})\,.$$

This limit exists since $\mathcal{H}(\mathsf{P}) = c_1 \ge c_2 \ge \cdots \ge 0$.

Define the ***average-distropy-number*** $\frac{1}{n}h_n$, where

$$h_n := \mathcal{H}(\mathsf{P}_0\vee\mathsf{P}_1\vee\cdots\vee\mathsf{P}_{n-1})\,.$$

Certainly $h_n = c_n + \mathcal{H}(\mathsf{P}_1\vee\cdots\vee\mathsf{P}_{n-1}) = c_n + h_{n-1}$, since T is measure preserving. Induction gives $h_n = \sum_{j=1}^n c_j$. So the Cesàro averages $\frac{1}{n}h_n$ converge to the entropy.

Theorem 5 *The entropy of process* $(T,P\colon X,\mathcal{X},\mu)$ *equals*

$$\lim_{n\to\infty}\frac{1}{n}\mathcal{H}(P_0\vee\cdots\vee P_{n-1}) = \lim_{n\to\infty}\mathcal{H}\left(P_0\middle|\bigvee_1^n P_j\right) = \mathcal{H}\left(P_0\middle|\bigvee_1^\infty P_j\right).$$

Both limits are non-increasing. The entropy $\mathcal{E}^T(P) \ge 0$, *with equality IFF* $P \subset \bigvee_1^\infty P_j$. *And* $\mathcal{E}^T(P) \le \mathcal{H}(P)$, *with equality IFF* T,P *is an independent process.*

Generators

We henceforth only discuss *invertible* mpts, that is, when T^{-1} is itself an mpt. Viewing the atoms of P as "letters", then, each $x \in X$ has a T,P-***name***

$$\ldots \mathrm{x}_{-2}\mathrm{x}_{-1}\mathrm{x}_0\mathrm{x}_1\mathrm{x}_2\mathrm{x}_3\ldots,$$

where x_n is $\mathsf{P}\langle T^n(x)\rangle$, the P-letter owning $T^n(x)$.

A partition P ***generates*** (the whole field) under $(T\colon X,\mu)$, if $\bigvee_{-\infty}^\infty T^n\mathsf{P} =_\mu \mathcal{X}$. It turns out[8] that

[8] I am now at liberty to reveal that our X has always been a ***Lebesgue space***, that is, measure-isomorphic to an interval of $\mathbb{R}$ to-

P generates IFF P ***separates points***. That is, after deleting a (T-invariant) nullset, distinct points of X have distinct T, P-names.

A finite set $[1 \dots L]$ of integers, interpreted as an ***alphabet,*** yields the ***shift space*** $X := [1 \dots L]^{\mathbb{Z}}$ of doubly-infinite sequences $x = (\dots \mathrm{x}_{-1}\mathrm{x}_0\mathrm{x}_1 \dots)$. The ***shift*** $T\colon X \to X$ acts on X by

$$T(x) := [n \mapsto \mathrm{x}_{n+1}] \,.$$

The time-zero partition P separates points, under the action of the shift. This L-atom ***time-zero partition*** has $\mathsf{P}\langle x\rangle = \mathsf{P}\langle y\rangle$ IFF $\mathrm{x}_0 = \mathrm{y}_0$. So no matter what shift-invariant measure is put on X, the time-zero partition will generate under the action of T.

Time Reversibility

A transformation need not be isomorphic to its inverse. Nonetheless, the average-distropy-numbers show that $\mathcal{E}(T^{-1}, \mathsf{P}) = \mathcal{E}(T, \mathsf{P})$; although this is not obvious from the conditioning-definition of entropy. Alternatively,

$$\begin{aligned}
&\mathcal{H}\left(\mathsf{P}_0 \middle| \bigvee_1^n \mathsf{P}_j\right) \\
&\quad = \mathcal{H}(\mathsf{P}_0 \vee \dots \vee \mathsf{P}_n) - \mathcal{H}(\mathsf{P}_1 \vee \dots \vee \mathsf{P}_n) \\
&\quad = \mathcal{H}(\mathsf{P}_{-n} \vee \dots \vee \mathsf{P}_0) - \mathcal{H}(\mathsf{P}_{-n} \vee \dots \vee \mathsf{P}_{-1}) \\
&\quad = \mathcal{H}\left(\mathsf{P}_0 \middle| \bigvee_{-1}^{-n} \mathsf{P}_j\right) . \qquad (12)
\end{aligned}$$

Bernoulli Processes

A probability vector $\vec{v} := (v_1, \dots, v_L)$ can be viewed as a measure on alphabet $[1 \dots L]$. Let $\mu_{\vec{v}}$ be the resulting product measure on $X := [1 \dots L]^{\mathbb{Z}}$, with T the shift on X and P the time-zero partition. The independent process $(T, \mathsf{P}\colon X, \mu_{\vec{v}})$ is called, by ergodic theorists, a ***Bernoulli process***. Not necessarily consistently, we tend to refer to the underlying *transformation* as a ***Bernoulli shift***.

The $\left(\frac{1}{2}, \frac{1}{2}\right)$-Bernoulli and the $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$-Bernoulli have different process-entropies, but perhaps their underlying transformations are isomorphic? Prior to the Kolmogorov–Sinai definition of entropy[9] of a *transformation*, this question remained unanswered.

Entropy of a Transformation

The Kolmogorov–Sinai definition of the entropy of an mpt is

$$\mathcal{E}(T) := \sup\{\mathcal{E}^T(\mathsf{Q}) \mid \mathsf{Q} \text{ a partition on } X\} \,.$$

Certainly entropy is an isomorphism invariant – but is it useful? After all, the supremum of *distropies* of partitions is always infinite (on non-atomic spaces) and one might fear that the same holds for entropies. The key observation (restated in Lemma 8c and proved below) was this, from [4] and [7].

Theorem 6 (Kolmogorov–Sinai Theorem) *If P generates under T, then $\mathcal{E}(T) = \mathcal{E}(T, \mathsf{P})$.*

Thereupon the $\left(\frac{1}{2}, \frac{1}{2}\right)$ and $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ Bernoulli-shifts are *not* isomorphic, since their respective entropies are $\log(2) \neq \log(3)$.

Wolfgang Krieger later proved a converse to the Kolmogorov–Sinai theorem.

Theorem 7 (Krieger Generator Theorem, 1970) *Suppose T ergodic. If $\mathcal{E}(T) < \infty$, then T has a generating partition. Indeed, letting K be the smallest integer $K > \mathcal{E}(T)$, there is a K-atom generator.*[10]

Proof See Rudolph [21], or § 5.1 in ► Joinings in Ergodic Theory, where Krieger's theorem is stated in terms of joinings. □

Entropy Is Continuous

Given *ordered* partitions $\mathsf{Q} = (\mathbf{B_1}, \dots)$ and $\mathsf{Q}' = (\mathbf{B'_1}, \dots)$, extend the shorter by null-atoms until $|\mathsf{Q}| = |\mathsf{Q}'|$. Let $Fat := \bigsqcup_j [B_j \cap B'_j]$; this set should have mass close to 1 if Q and Q' are almost the same partition. Define a new partition

$$\mathsf{Q} \triangle \mathsf{Q}' := \{Fat\} \sqcup \{B_i \cap B'_j \mid \text{with } i \neq j\} \,.$$

(In other words, take $\mathsf{Q} \vee \mathsf{Q}'$ and coalesce, into a single atom, all the $B_k \cap B'_k$ sets.) Topologize the space of parti-

gether with countably many point-atoms (points with positive mass). The equivalence of *generating* and *separating* is a technical theorem, due to Rokhlin.

Assuming μ to be Lebesgue is not much of a limitation. For instance, if μ is a finite measure on any Polish space, then μ extends to a Lebesgue measure on the μ-completion of the Borel sets. To not mince words: All spaces are Lebesgue spaces unless you are actively *looking* for trouble.

[9] This is sometimes called ***measure(-theoretic)*** *entropy* or (perhaps unfortunately) ***metric entropy***, to distinguish it from topological entropy. Tools known prior to entropy, such as *spectral* properties, did *not* distinguish the two Bernoulli-shifts; see ► Spectral Theory of Dynamical Systems for the definitions.

[10] It is an easier result, undoubtedly known much earlier, that every ergodic T has a *countable* generating partition – possibly of ∞-distropy.

tions by saying[11] that $\mathsf{Q}^{(L)} \to \mathsf{Q}$ when $\mathcal{H}\left(\mathsf{Q}\triangle\mathsf{Q}^{(L)}\right) \to 0$. Then Lemma 8b says that process-entropy varies continuously with varying the partition.

Lemma 8 *Fix a mpt $(T\colon X, \mu)$. For partitions P, Q, Q', define $R := Q\triangle Q'$ and let $\delta := \mathcal{H}(R)$. Then*

(a) $|\mathcal{H}(Q) - \mathcal{H}(Q')| \le \delta$. *(Distropy varies continuously with the partition.)*
(b) $|\mathcal{E}^T(Q) - \mathcal{E}^T(Q')| \le \delta$. *(Process-entropy varies continuously with the partition.)*
(c) For all partitions $Q \subset \mathrm{Fld}(T, P)\colon \mathcal{E}^T(Q) \le \mathcal{E}^T(P)$.

Proof (of (a)) Evidently $\mathsf{Q}' \vee \mathsf{R} = \mathsf{Q}' \vee \mathsf{Q} = \mathsf{Q} \vee \mathsf{R}$. So $\mathcal{H}(\mathsf{Q}') \le \mathcal{H}(\mathsf{Q} \vee \mathsf{R}) \le \mathcal{H}(\mathsf{Q}) + \delta$. □

Proof (of (b)) As above,

$$\mathcal{H}\left(\bigvee_1^N \mathsf{Q}'_j\right) \le \mathcal{H}\left(\bigvee_1^N \mathsf{Q}_j\right) + \mathcal{H}\left(\bigvee_1^N \mathsf{R}_j\right).$$

Sending $N \to \infty$ gives $\mathcal{E}^T(\mathsf{Q}') \le \mathcal{E}^T(\mathsf{Q}) + \mathcal{E}^T(\mathsf{R})$. Finally, $\mathcal{E}^T(\mathsf{R}) \le \mathcal{H}(\mathsf{R})$ and so $\mathcal{E}^T(\mathsf{Q}') \le \mathcal{E}^T(\mathsf{Q}) + \delta$. □

Proof (of (c)) Let $K := |\mathsf{Q}|$. Then there is a sequence of K-set partitions $\mathsf{Q}^{(L)} \to \mathsf{Q}$ with $\mathsf{Q}^{(L)} \preccurlyeq \bigvee_{-L}^{L} \mathsf{P}_\ell$. By above, $\mathcal{E}^T(\mathsf{Q}^{(L)}) \to \mathcal{E}^T(\mathsf{Q})$, so showing that

$$\mathcal{E}^T\left(\bigvee_{-L}^{L} \mathsf{P}_\ell\right) \overset{?}{\le} \mathcal{E}^T(\mathsf{P})$$

will suffice. Note that

$$h_N := \mathcal{H}\left(\bigvee_{n=0}^{N-1} T^n\left(\bigvee_{-L}^{L} \mathsf{P}_\ell\right)\right) = \mathcal{H}\left(\bigvee_{j=-L}^{N-1+L} \mathsf{P}_j\right).$$

So $\frac{1}{N}h_N \le \frac{1}{N}\mathcal{H}\left(\bigvee_0^{N-1} \mathsf{P}_j\right) + \frac{1}{N}2L\mathcal{H}(\mathsf{P})$. Now send $N \to \infty$. □

Entropy Is Not Continuous

The most common topology placed on the space Ω of mpts is the *coarse topology*[12] that Halmos discusses in his "little red book" [14].

[11] On the set of ordered K-set partitions (with K fixed) this convergence is the same as: $\mathsf{Q}^{(L)} \to \mathsf{Q}$ when $\mu\left(Fat\left(\mathsf{Q}^{(L)}, \mathsf{Q}\right)\right) \to 1$.

An alternative approach is the ***Rokhlin metric***, $\mathrm{Dist}(\mathsf{P}, \mathsf{Q}) := \mathcal{H}(\mathsf{P}|\mathsf{Q}) + \mathcal{H}(\mathsf{Q}|\mathsf{P})$, which has the advantage of working for *unordered* partitions.

[12] i. e, $S_n \to T$ IFF $\forall A \in \mathcal{X} : \mu({S_n}^{-1}(A)\triangle T^{-1}(A)) \to 0$; this is a metric-topology, since our probability space is countably generated. This can be restated in terms of the unitary operator U_T on $\mathbb{L}^2(\mu)$, where $U_T(f) := f \circ T$. Namely, $S_n \to T$ in the coarse topology IFF $U_{S_n} \to U_T$ in the strong operator topology.

The Rokhlin lemma (see p. 33 in [21]) implies that the isomorphism-class of *each* ergodic mpt is *dense* in Ω, (e. g., see p. 77 in [14]) disclosing that the $S \mapsto \mathcal{E}(S)$ map is exorbitantly discontinuous.

Indeed, the failure happens already for process-entropy with respect to a fixed partition. A Bernoulli process T, P has positive entropy. Take mpts $S_n \to T$, each isomorphic to an irrational rotation. Then each $\mathcal{E}(S_n, \mathsf{P})$ is zero, as shown in the later section "Determinism and Zero-Entropy".

Further Results When $\mathcal{F}$ is a T-invariant subfield, agree to use $T\restriction_{\mathcal{F}}$ for "T restricted to $\mathcal{F}$", which is a factor (see Glossary) of T. Transformations T and S are ***weakly isomorphic*** if each is isomorphic to a factor of the other.

The foregoing entropy tools make short shrift of the following.

Lemma 9 (Entropy Lemma) *Consider T-invariant subfields $\mathcal{G}_j$ and $\mathcal{F}$.*

(a) Suppose $\mathcal{G}_j \nearrow \mathcal{F}$. Then $\mathcal{E}T\restriction_{\mathcal{G}_j} \nearrow \mathcal{E}(T\restriction_{\mathcal{F}})$. In particular, $\mathcal{G} \subset \mathcal{F}$ implies that $\mathcal{E}(T\restriction_{\mathcal{G}}) \le \mathcal{E}(T\restriction_{\mathcal{F}})$, so entropy is an invariant of weak-isomorphism.
(b) $\mathcal{E}(T\restriction_{\mathcal{G}_1\vee\mathcal{G}_2\vee\ldots}) \le \sum_j \mathcal{E}(T\restriction_{\mathcal{G}_j})$.
And $\mathcal{E}(T, Q_1 \vee Q_2 \vee \ldots) \le \sum_j \mathcal{E}(T, Q_j)$.
(c) For mpts $(S_j\colon Y_j, \nu_j)$: $\mathcal{E}(S_1 \times S_2 \times \cdots) = \sum_j \mathcal{E}(S_j)$.
(d) $\mathcal{E}(T^{-1}) = \mathcal{E}(T)$. *More generally,* $\mathcal{E}(T^n) = |n| \cdot \mathcal{E}(T)$.

Meshalkin's Map

In the wake of Kolmogorov's 1958 entropy paper, for two Bernoulli-shifts to be isomorphic one now knew that they had to have equal entropies. Meshalkin provided the first non-trivial example in 1959 [45].

Let $S : Y \to Y$ be the Bernoulli-shift over the "letter" alphabet $\{\mathtt{E}, \mathtt{D}, \mathtt{P}, \mathtt{N}\}$, with probability distribution $\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$. The letters $\mathtt{E}, \mathtt{D}, \mathtt{P}, \mathtt{N}$ stand for *Even, oDd, Positive, Negative*, and will be used to describe the code (isomorphism) between the processes.

Use $T : X \to X$ for the Bernoulli-shift over "digit" alphabet $\{0, +1, -1, +2, -2\}$, with probability distribution $\left(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right)$. Both distributions $\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$ and $\left(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right)$ have distropy log(4).

After deleting invariant nullsets from X and Y, the construction will produce a measure-preserving isomorphism $\psi : X \to Y$ so that $T \circ \psi = \psi \circ S$.

The Code In X, consider this point x:

$$\ldots\ \mathbf{0}\ \mathbf{0}\ \mathbf{0}\ -1\ \mathbf{0}\ \mathbf{0}\ +1\ +2\ -1\ +1\ \mathbf{0}\ \ldots$$

Regard each **0** as a left-parenthesis, and each non-zero as a right-parenthesis. Link them according to the legal way of matching parentheses, as shown in the top row, below:

$$\begin{array}{ccccccccccc} \mathbf{0} & \mathbf{0} & \mathbf{0} & -1 & \mathbf{0} & \mathbf{0} & +1 & +2 & -1 & +1 & \mathbf{0} \\ \mathtt{P} & \mathtt{N} & \mathtt{N} & \mathtt{D} & \mathtt{P} & \mathtt{P} & \mathtt{D} & \mathtt{E} & \mathtt{D} & \mathtt{D} & ? \end{array}$$

The leftmost **0** is linked to the rightmost +1, as indicated by the longest-overbar. The left/right-parentheses form a $\left(\frac{1}{2}, \frac{1}{2}\right)$-random-walk. Since this random walk is recurrent, every position in x will be linked (except for a nullset of points x).

Below each **0**, write "P" or "N" as the **0** is linked to a *positive* or *negative* digit. And below the other digits, write "E" or "D" as the digit is *even* or *odd*. So the upper name in X is mapped to the lower name, a point $y \in Y$.

This map $\psi : X \to Y$ carries the upstairs $\left(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right)$ distribution to $\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$, downstairs. It takes some arguing to show that independence is preserved.

The inverse map ψ^{-1} views D and E as right-parentheses, and P and N as left. Above D, write the odd digit $+\mathbf{1}$ or $-\mathbf{1}$, as this D is linked to Positive or Negative.

Markov Shifts

A Bernoulli process T, P has independence $\mathsf{P}_{(-\infty\ldots 0]} \perp \mathsf{P}_1$ whereas a ***Markov process*** is a bit less aloof:

> The infinite Past $\mathsf{P}_{(-\infty\ldots 0]}$ doesn't provide any more information about Tomorrow than Today did.

That is, the conditional distribution $\mathsf{P}_1 | \mathsf{P}_{(-\infty\ldots 0]}$ equals $\mathsf{P}_1 | \mathsf{P}_0$. Equivalently,

$$\mathcal{H}(\mathsf{P}_1 | \mathsf{P}_0) = \mathcal{H}\left(\mathsf{P}_1 | \mathsf{P}_{(-\infty\ldots 0]}\right) \overset{\text{note}}{=} \mathcal{E}(T, \mathsf{P}) . \tag{13}$$

The simplest non-trivial Markov process $(T, \mathsf{P} : X, \mu)$ is over a two-letter alphabet $\{\mathtt{a}, \mathtt{b}\}$, and has transition graph Fig. 4, for some choice of transition probabilities s and c. The graph's Markov matrix is

$$\mathbf{M} = [m_{i,j}]_{i,j} = \begin{bmatrix} s & c \\ 1 & 0 \end{bmatrix} ,$$

where $c = 1 - s$, and $m_{i,j}$ denotes the probability of going from state i to state j.

If Today's distribution on the two states is the probability-vector $\vec{v} := [p_{\mathtt{a}} \quad p_{\mathtt{b}}]$, then Tomorrow's is the product $\vec{v}\mathbf{M}$. So a *stationary* process needs $\vec{v}\mathbf{M} = \vec{v}$. This equation has the unique solution $p_{\mathtt{a}} = \frac{1}{1+c}$ and $p_{\mathtt{b}} = \frac{c}{1+c}$. An example of computing the probability of a word or

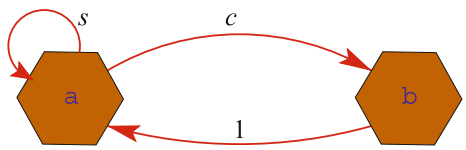

Entropy in Ergodic Theory, Figure 4
Call the transition probabilities $s :=$ Prob(a→a) for *stay*, and $c :=$ Prob a→b for *change*. These are non-negative reals, and $s + c = 1$

cylinder set; (see Sect. "The Carathéodory Construction" in ▶ Measure Preserving Systems) in the process, is

$$\begin{aligned} \mu_s(\mathtt{baaaba}) &= p_{\mathtt{b}} m_{\mathtt{ba}} m_{\mathtt{aa}} m_{\mathtt{aa}} m_{\mathtt{ab}} m_{\mathtt{ba}} \\ &= \frac{c}{1+c} \cdot 1 \cdot s \cdot s \cdot c \cdot 1 . \end{aligned}$$

The subscript on μ_s indicates the dependence on the transition probabilities; let's also mark the mpt and call it T_s. Using Eq. (13), the entropy of our Markov map is

$$\begin{aligned} \mathcal{E}(T_s) &= p_{\mathtt{a}} \cdot \mathcal{H}(s, c) + p_{\mathtt{b}} \cdot \underbrace{\mathcal{H}(1, 0)}_{=0} \\ &= \frac{-1}{1+c} \cdot [s \log(s) + c \log(c)] . \end{aligned} \tag{14}$$

Determinism and Zero-Entropy

Irrational rotations have zero-entropy; let's reveal this in two different ways.

Equip $X := [0, 1)$ with "length" (Lebesgue) measure and wrap it into a circle. With "$\oplus$" denoting addition mod-1, have $T : X \to X$ be the rotation $T(x) := x \oplus \boldsymbol{\alpha}$, where the rotation number $\boldsymbol{\alpha}$ is irrational. Pick distinct points $y_0, z_0 \in X$, and let P be the partition whose two atoms are the intervals $[y_0, z_0)$ and $[z_0, y_0)$, wrapping around the circle.

The T-orbit of each point x is dense[13] in X. In particular, y_0 has dense orbit, so P separates points – hence generates – under T. Our goal, thus, is $\mathcal{E}(T, \mathsf{P}) \overset{?}{=} 0$.

Rotations are Deterministic

The forward T-orbit of each point is dense. This is true for y_0, and so the *backward* T, P-name of each x actually tells us which point x is. I. e., $\mathsf{P} \subset \bigvee_{-\infty}^{-1} T^n \mathsf{P}$, which is our definition of "process T, P is ***deterministic***". Our P being finite, this determinism implies that $\mathcal{E}(T, \mathsf{P})$ is zero, by Theorem 5.

[13] Fix an $\varepsilon > 0$ and an $N > 1/\varepsilon$. Points $x, T(x), \ldots, T^N(x)$ have some two at distance less than $\frac{1}{N}$; say, $\mathrm{Dist}(T^i(x), T^j(x)) < \varepsilon$, for some $0 \le i < j \le N$. Since T is an isometry, $\varepsilon > \mathrm{Dist}(x, T^k(x)) > 0$, where $k := j - i$. So the T^k-orbit of x is ε-dense.

Counting Names in a Rotation The $\mathsf{P}_0 \vee \cdots \vee \mathsf{P}_{n-1}$ partition places n translates of points y_0 and of z_0, cutting the circle into at most $2n$ intervals. Thus $\mathcal{H}(\mathsf{P}_0 \vee \ldots \vee \mathsf{P}_{n-1}) \leq \log(2n)$. And $\frac{1}{n}\log(2n) \to 0$.

Alternatively, the below SMB-theorem implies, for an ergodic process T, P, that the number of length-n names is approximately $2^{\mathcal{E}(T,\mathsf{P})\cdot n}$; this, after discarding small mass from the space. But the growth of $n \mapsto 2n$ is sub-exponential and so, for our rotation, $\mathcal{E}(T, \mathsf{P})$ must be zero.

Theorem 10 (Shannon–McMillan–Breiman Theorem (SMB-Theorem)) *Set* $\mathbf{E} := \mathcal{E}(T, P)$, *where the tuple* $(T, P\colon X, \mu)$ *is an ergodic process. Then the average information function*

$$\frac{1}{n}\mathcal{I}_{P_{[0\ldots n)}}(x) \overset{n\to\infty}{\longrightarrow} \mathbf{E}\,, \quad \textit{for a.e.} \quad x \in X\,. \tag{15}$$

The functions $f_n := \mathcal{I}_{P_{[0\ldots n)}}$ *converge to the constant function* $\mathbf{E}$ *both in the* $\mathbb{L}^1$*-norm and in probability.* [14]

Proof See the texts of Karl Petersen (p. 261 in [20]), or Dan Rudolph (p. 77 in [21]).

□

Consequences Recall that $\mathsf{P}_{[0\ldots n)}$ means $\mathsf{P}_0 \vee \mathsf{P}_1 \vee \cdots \vee \mathsf{P}_{n-1}$, where $\mathsf{P}_j := T^j\mathsf{P}$. As usual, $\mathsf{P}_{[0\ldots n)}\langle x\rangle$ denotes the $\mathsf{P}_{[0\ldots n)}$-atom owning x.

Having deleted a nullset, we can restate Eq. (15) to now say that $\forall\varepsilon, \forall x, \forall_{\text{large}} n$:

$$1/2^{[\mathbf{E}+\varepsilon]n} \leq \mu\big(\mathsf{P}_{[0\ldots n)}\langle x\rangle\big) \leq 1/2^{[\mathbf{E}-\varepsilon]n}\,. \tag{16}$$

This has the following consequence. Fixing a number $\delta > 0$, we consider any set with $\mu(B) \geq \delta$ and count the number of n-names of points in B. The SMB-Thm implies

$$\forall\varepsilon, \forall_{\text{large}} n, \forall B \overset{\mu}{\geq} \delta: \quad \big|\{n\text{-names in } B\}\big| \geq 2^{[\mathbf{E}-\varepsilon]n}. \tag{17}$$

Rank-1 Has Zero-Entropy

There are several equivalent definitions for "rank-1 transformation", several of which are discussed in the introduction of [28]. (See Chap. 6 in [13] as well as [51] and [27] for examples of stacking constructions.)

A **rank-1 transformation** $(T\colon X, \mu)$ admits a generating partition P and a sequence of Rokhlin stacks $S_n \subset X$, with heights going to ∞, and with $\mu(S_n) \to 1$. Moreover, each of these Rokhlin stacks is P-monochromatic, that is, each level of the stack lies entirely in some atom of P.

Taking a stack of some height $2n$, let $B = B_n$ be the union of the bottom n levels of the stack. There are at most n many length-n names starting in B_n, by monochromaticity. Finally, $\mu(B_n)$ is almost $\frac{1}{2}$, so is certainly larger than $\delta := \frac{1}{3}$. Thus Eq. (17) shows that our rank-1 T has zero entropy.

Cautions on Determinism's Relation to Zero-Entropy

A finite-valued process T, P has zero-entropy iff $\mathsf{P} \subset \bigvee_{-\infty}^{-1} \mathsf{P}_j$. Iterating gives

$$\bigvee_{0}^{\infty} \mathsf{P}_j \subset \bigvee_{-\infty}^{-1} \mathsf{P}_j\,,$$

i. e., the future is measurable with respect to the past.

This was the case with the rotation, where a point's past uniquely identified the point, thus telling us its future.

While determinism and zero-entropy mean the same thing for finite-valued processes, this fails catastrophically for real-valued (i. e., continuum-valued) processes, as shown by an example of the author's. A stationary real-valued process $\mathbf{V} = \ldots V_{-1}\ V_0\ V_1\ V_2\ \ldots$ is constructed in [40] which is simultaneously

> ***strongly deterministic***: The two values V_0, V_1 determine all of $\mathbf{V}$, future and past.

and **non-consecutively independent**. This latter means that for each bi-infinite increasing integer sequence $\{n_j\}_{j=-\infty}^{\infty}$ with no consecutive pair (always $1 + n_j < n_{j+1}$), then the list of random variables $\ldots V_{n_{-1}}\ V_{n_0}\ V_{n_1}\ V_{n_2}\ \ldots$ is an independent process.

Restricting the random variables to be *countably*-valued, how much of the example survives? Joint work with Kalikow [39] produced a countably-valued stationary $\mathbf{V}$ which is non-consecutively independent as well as deterministic. (Strong determinism is ruled out, due to cardinality considerations.) A side-effect of the construction is that $\mathbf{V}$'s time-reversal $n \mapsto V_{-n}$ is **not** deterministic.

The Pinsker–Field and K-Automorphisms

Consider the collection of ***zero-entropy sets***,

$$\mathcal{Z} = \mathcal{Z}_T := \{A \in \mathcal{X} \mid \mathcal{E}(T, (A, A^c)) = 0\}\,. \tag{18}$$

Courtesy of Lemma 9b, $\mathcal{Z}$ is a T-invariant *field*, and

$$\forall \mathsf{Q} \subset \mathcal{Z}\colon \mathcal{E}(T, \mathsf{Q}) = 0\,. \tag{19}$$

The ***Pinsker field***[15] of T is this $\mathcal{Z}$. It is maximal with respect to Eq. (19). Unsurprisingly, the ***Pinsker factor*** $T\restriction_{\mathcal{Z}}$ has

[14] In engineering circles, this is called the Almost-everywhere equipartition theorem.

[15] Traditionally, this called the *Pinsker algebra* where, in this context, "algebra" is understood to mean "σ-algebra".

zero entropy, that is, $\mathcal{E}(T \restriction_{\mathcal{Z}}) = 0$. A transformation T is said to have ***completely-positive entropy*** if it has no (non-trivial) zero-entropy factors. That is, its Pinsker field $\mathcal{Z}_T$ is the trivial field, $\emptyset := \{\varnothing, X\}$.

K-Processes

Kolmogorov introduced the notion of a ***K-process*** or ***Kolmogorov process***, in which the present becomes asymptotically independent of the distant past. The asymptotic past of the T, P process is called its ***tail field***, where

$$\mathrm{Tail}(T, \mathsf{P}) := \bigcap_{M=1}^{\infty} \bigvee_{j=-\infty}^{-M} \mathsf{P}_j \,.$$

This T, P is a K-process if $\mathrm{Tail}(T, \mathsf{P}) = \emptyset$. This turns out to be equivalent to what we might a call a strong form of *"sensitive dependence on initial conditions"*: For each fixed length L, the distant future

$$\bigvee_{j \in (G \ldots G+L]} \mathsf{P}_j$$

becomes more and more independent of

$$\bigvee_{j \in (-\infty \ldots 0]} \mathsf{P}_j \,,$$

as the gap $G \to \infty$.

A transformation T is a ***Kolmogorov automorphism*** if it possesses a generating partition P for which T, P is a K-process.

Here is a theorem that relates the "asymptotic forgetfulness" of $\mathrm{Tail}(T, \mathsf{P}) = \emptyset$, to the lack of determinism implied by having no zero-entropy factors (See Walters, [23], p. 113. Related results appear in Berg [44]).

Theorem 11 (Pinsker-Algebra Theorem) *Suppose P is a generating partition for an ergodic T. Then* $\mathrm{Tail}(T, P)$ *equals $\mathcal{Z}_T$.*

Since $\mathcal{Z}_T$ does not depend on P, this means that all generating partitions have the same tail field, and therefore K-ness of T can be detected from any generator.

Another non-evident fact follows from the above. The ***future field*** of T, P is defined to be $\mathrm{Tail}\,(T^{-1}\mathsf{P})$. It is not obvious that if the present is independent of the distant past, then it is automatically independent of the distant future. (Indeed, the precise definitions are important; witness the *Cautions on determinism* section.) But since the entropy of a process, $\mathcal{E}(T, (A, A^c))$, equals the entropy of the time-reversed process $\mathcal{E}(T^{-1}, (A, A^c))$, it follows that $\mathcal{Z}_T$ equals $\mathcal{Z}_{T^{-1}}$.

Ornstein Theory

In 1970, Don Ornstein [46] solved the long-standing problem of showing that entropy was a complete isomorphism-invariant of Bernoulli transformations; that is, that two independent processes with same entropy necessarily have the same underlying transformation. (Earlier, Sinai [52]) had shown that two such Bernoulli maps were ***weakly isomorphic***, that is, each isomorphic to a factor of the other.)

Ornstein introduced the notion of a process being ***finitely determined***, see [46] for a definition, proved that a transformation T was Bernoulli IFF it had a finitely-determined generator IFF every partition was finitely-determined with respect to T, and showed that entropy completely classified the finitely-determined processes up to isomorphism.

This seminal result led to a vast machinery for proving transformations to be Bernoulli, as well as classification and structure theorems [47,51,53]. Showing that the class of K-automorphisms far exceeds the Bernoulli maps, Ornstein and Shields produced in [48] an uncountable family of non-isomorphic K-automorphisms all with the same entropy.

Topological Entropy

Adler, Konheim and McAndrew, in 1965, published the first definition of *topological entropy* in the eponymous article [33]. Here, $T\colon X \to X$ is a continuous self-map of a compact topological space. The role of atoms is played by open sets. Instead of a finite partition, one uses a finite[16] ***open-cover*** $\mathcal{V} = \{U_j\}_{j=1}^{L}$, i. e. each ***patch*** U_j is open, and their union $\bigcup(\mathcal{V}) = X$. (Henceforth, 'cover' means "open cover".)

Let $\mathrm{Card}(\mathcal{V})$ be the *minimum* cardinality over all subcovers.

$$\mathrm{Card}(\mathcal{V}) := \mathrm{Min}\left\{{}^{\sharp}\mathcal{V}' \mid \mathcal{V}' \subset \mathcal{V} \text{ and } \bigcup(\mathcal{V}') = X\right\} ,$$

and let

$$\mathcal{H}(\mathcal{V}) = \mathcal{H}_{\mathrm{top}}(\mathcal{V}) := \log(\mathrm{Card}(\mathcal{V})) \,.$$

Analogous to the definitions for partitions, define

$$\mathcal{V} \vee \mathcal{W} := \{V \cap W \mid V \in \mathcal{V} \text{ and } W \in \mathcal{W}\} \,;$$
$$T\mathcal{V} := \{T^{-1}(U) \mid U \in \mathcal{V}\}$$
$$\text{and } \mathcal{V}_{[0\ldots n)} := \mathcal{V}_0 \vee \mathcal{V}_1 \vee \cdots \vee \mathcal{V}_{n-1} \,;$$
$$\mathcal{W} \succcurlyeq \mathcal{V} \,, \text{ if each } \mathcal{W}\text{-patch is a subset of some } \mathcal{V}\text{-patch.}$$

[16] Because we only work on a compact space, we can omit "finite". Some generalizations of topological entropy to non-compact spaces require that only *finite* open-covers be used [37].

The $T, \mathcal{V}$-***entropy*** is

$$\mathcal{E}^T(\mathcal{V}) = \mathcal{E}(T, \mathcal{V}) = \mathcal{E}_{\text{top}}(T, \mathcal{V}) := \limsup_{n\to\infty} \frac{1}{n} \mathcal{H}_{\text{top}}\big(\mathcal{V}_{[0\ldots n)}\big) . \tag{20}$$

And the ***topological entropy*** of T is

$$\mathcal{E}_{\text{top}}(T) := \sup_{\mathcal{V}} \mathcal{E}_{\text{top}}(T, \mathcal{V}) , \tag{21}$$

taken over all open covers $\mathcal{V}$.

Thus $\mathcal{E}_{\text{top}}$ counts, in some sense, the growth rate in the number of T-orbits of length n.

Evidently, topological entropy is an isomorphism invariant. Two continuous maps $T\colon X \to X$ and $S\colon Y \to Y$ are ***topologically conjugate*** (as *isomorphism* is called in this category) if there exists a homeomorphism $\psi\colon X \to Y$ with $\psi T = S\psi$.

Lemma 12 (Subadditive Lemma) *Consider a sequence* $\mathbf{s} = (s_l)_1^\infty \subset [-\infty, \infty]$ *satisfying* $s_{k+l} \le s_k + s_l$, *for all* $k, l \in \mathbb{Z}$. *Then the following limit exists in* $[-\infty, \infty]$, *and* $\lim_{n\to\infty} \frac{s_n}{n} = \inf_n \frac{s_n}{n}$.

Topological entropy, or **"top ent"** for short, satisfies many of the relations of measure-entropy.

Lemma 13

(a) $\mathcal{V} \preccurlyeq \mathcal{W}$ *implies*

$$\mathcal{H}(\mathcal{V}) \le \mathcal{H}(\mathcal{W}) \quad \textit{and} \quad \mathcal{E}(T, \mathcal{V}) \le \mathcal{E}(T, \mathcal{W}) .$$

(b) $\mathcal{H}(\mathcal{V} \vee \mathcal{W}) \le \mathcal{H}(\mathcal{V}) + \mathcal{H}(\mathcal{W})$.

(c) $\mathcal{H}(T(\mathcal{V})) \le \mathcal{H}(\mathcal{V})$, *with equality if* T *is surjective. Also,* $\mathcal{E}(T, \mathcal{V}) \le \mathcal{H}(\mathcal{V})$.

(d) In Eq. (20), the $\lim_{n\to\infty} \frac{1}{n} \mathcal{H}\big(\mathcal{V}_{[0\ldots n)}\big)$ *exists.*

(e) Suppose T *is a homeomorphism. Then*

$$\mathcal{E}(T^{-1}, \mathcal{V}) = \mathcal{E}(T, \mathcal{V}) ,$$

for each cover $\mathcal{V}$. *Consequently,* $\mathcal{E}_{\text{top}}(T^{-1}) = \mathcal{E}_{\text{top}}(T)$.

(f) Suppose $\mathfrak{C}$ *is a collection of covers such that: For each cover* $\mathcal{W}$, *there exists a* $\mathcal{V} \in \mathfrak{C}$ *with* $\mathcal{V} \succcurlyeq \mathcal{W}$. *Then* $\mathcal{E}_{\text{top}}(T)$ *equals the supremum of* $\mathcal{E}_{\text{top}}(T, \mathcal{V})$, *just taken over those* $\mathcal{V} \in \mathfrak{C}$.

(g) For all $\ell \in \mathbb{N}$:

$$\mathcal{E}_{\text{top}}\Big(T^\ell\Big) = \ell \mathcal{E}_{\text{top}}(T) .$$

Proof (of (c)) Let $\mathcal{C} \preccurlyeq \mathcal{V}$ be a min-cardinality subcover. Then $T\mathcal{C}$ is a subcover of $T\mathcal{V}$. So Card $T\mathcal{V} \le |T\mathcal{C}| = |\mathcal{C}|$.

As for entropy, inequality (b) and the foregoing give $\mathcal{H}(\mathcal{V}_{[0\ldots n)}) \le \mathcal{H}(\mathcal{V})n$. □

Proof (of (d)) Set $s_n := \mathcal{H}(\mathcal{V}_{[0\ldots n)})$. Then

$$s_{k+l} \le s_k + \mathcal{H}\Big(T^k\big(\mathcal{V}_{[0\ldots l)}\big)\Big) \le s_k + s_l ,$$

by (b) and (c), and so the Subadditive Lemma 12, applies. □

Proof (of (g)) WLOG, $\ell = 3$. Given $\mathcal{V}$ a cover, triple it to $\widehat{\mathcal{V}} := \mathcal{V} \cap T\mathcal{V} \cap T^2\mathcal{V}$; so

$$\bigvee_{j\in[0\ldots N)} [T^3]^j\big(\widehat{\mathcal{V}}\big) = \bigvee_{i\in[0\ldots 3N)} T^i(\mathcal{V}) .$$

Thus $\mathcal{H}\Big(T^3, \widehat{\mathcal{V}}, N\Big) = \mathcal{H}(T, \mathcal{V}, 3N)$, extending notation. Part (d) and sending $N \to \infty$, gives $\mathcal{E}(T^3, \widehat{\mathcal{V}}) = 3\mathcal{H}(T, \mathcal{V})$.

Lastly, take covers such that

$$\mathcal{E}\Big(T^3, \mathcal{C}^{(k)}\Big) \to \mathcal{E}_{\text{top}}\big(T^3\big) \quad \text{and}$$
$$\mathcal{E}\Big(T, \mathcal{D}^{(k)}\Big) \to \mathcal{E}_{\text{top}}(T) ,$$

as $k\to\infty$. Define $\mathcal{V}^{(k)} := \mathcal{C}^{(k)} \vee \mathcal{D}^{(k)}$. Apply the above to $\mathcal{V}^{(k)}$, then send $k\to\infty$. □

Using a Metric

From now on, *our space is a compact metric space* (X, d).

Dinaburg [36] and Bowen [34,35], gave alternative, equivalent, definitions of topological entropy, in the compact metric-space case, that are often easier to work with than covers. Bowen gave a definition also when X is not compact[17] (see [35] and Chap. 7 in [23]).

Metric Preliminaries An ***ε-ball-cover*** comprises finitely many balls, all of radius ε. Since our space is compact, every cover $\mathcal{V}$ has a ***Lebesgue number*** $\varepsilon > 0$. *I. e., for each* $z \in X$, the $\mathrm{Bal}(z, \varepsilon)$ lies entirely inside at least one $\mathcal{V}$-patch. (In particular, there is an ε-ball-cover which *refines* $\mathcal{V}$.) Let $\mathrm{LEB}(\mathcal{V})$ be the supremum of the Lebesgue numbers. Courtesy of Lemma 13f we can

> Fix a "universal" list $\mathcal{V}^{(1)} \preccurlyeq \mathcal{V}^{(2)} \preccurlyeq \ldots$, with $\mathcal{V}^{(k)}$ a $\frac{1}{k}$-ball-cover. For every $T\colon X \to X$, then, the $\lim_k \mathcal{E}(T, \mathcal{V}^{(k)})$ computes $\mathcal{E}_{\text{top}}(T)$.

[17] When X is not compact, the definitions need not coincide; e. g. [37]. And topologically-equivalent metrics, but which are not uniformly equivalent, may give the same T different entropies (see p. 171 in [23]).

An ε-Microscope Three notions are useful in examining a metric space (X, m) at scale ε. Subset $A \subset X$ is an ***ε-separated-set***, if $\mathrm{m}(z, z') \geq \varepsilon$ for all distinct $z, z' \in A$. Subset $F \subset X$ is ***ε-spanning*** if $\forall x \in X, \exists z \in F$ with $\mathrm{m}(x, z) < \varepsilon$.

Lastly, a cover $\mathcal{V}$ is ***ε-small*** if $\mathrm{Diam}(U) < \varepsilon$, for each $U \in \mathcal{V}$.

You Take the High Road and I'll Take the Low Road There are several routes to computing top-ent, some via maximization, others, minimization. Our foregoing discussion computed $\mathcal{E}_{\mathrm{top}}(T)$ by a family of ***sizes*** $f_k(n) = f_k^T(n)$, depending on a parameter k which specifies the fineness of scale. (In Sect. "Metric Preliminaries", this k is an integer; in the original definition, an open cover.) Define two numbers:

$$\widehat{\mathsf{L}}^f(k) := \limsup_{n\to\infty} \frac{1}{n} \log f_k(n) \quad \text{and}$$
$$\underline{\mathsf{L}}^f(k) := \liminf_{n\to\infty} \frac{1}{n} \log f_k(n) \,. \tag{22}$$

Finally, let $\mathcal{E}^f(T) := \sup_k \widehat{\mathsf{L}}^f(k)$. If the limit exists in Eq. (22) then agree to write $\mathsf{L}^f(k)$ for the common value.

The A-K-M definition used the size $f_{\mathcal{V}}(n) := \mathrm{Card}(\mathcal{V}_{[0\ldots n)})$, where

$$\mathrm{Card}(\mathcal{W}) := \text{Minimum cardinality of a subcover from } \mathcal{W}.$$

Here are three metric-space sizes $f_\varepsilon(n)$:

$$\mathrm{Sep}(n, \varepsilon) := \textit{Maximum} \text{ cardinality of a } \mathrm{d}_n\text{–}\varepsilon\text{-separated set.}$$
$$\mathrm{Spn}(n, \varepsilon) := \text{Minimum cardinality of a } \mathrm{d}_n\text{–}\varepsilon\text{-spanning set.}$$
$$\mathrm{Cov}(n, \varepsilon) := \text{Minimum cardinality of a } \mathrm{d}_n\text{–}\varepsilon\text{-small cover.}$$

These use a list $(\mathrm{d}_n)_{n=1}^\infty$ of progressively finer metrics on X, where

$$\mathrm{d}_N(x, y) := \mathrm{Max}_{j\in[0\ldots N)}\, \mathrm{d}\Big(T^j(x), T^j(y)\Big) \,.$$

Theorem 14 (All-Roads-Lead-to-Rome Theorem) *Fix ε and let $\mathcal{W}$ be any d-ε-small cover. Then*

(i) $\forall n$:
$\mathrm{Cov}(n, 2\varepsilon) \leq \mathrm{Spn}(n, \varepsilon) \leq \mathrm{Sep}(n, \varepsilon) \leq \mathrm{Card}(\mathcal{W}_{[0\ldots n)})$.

(ii) Take a cover $\mathcal{V}$ and a $\delta < \mathrm{LEB}(\mathcal{V})$. Then $\forall n$: $\mathrm{Card}(\mathcal{V}_{[0\ldots n)}) \leq \mathrm{Cov}(n, \delta)$.

(iii) The limit $L^{\mathrm{Cov}}(\varepsilon) = \lim_n \frac{1}{n}\log(\mathrm{Cov}(n, \varepsilon))$ exists in $[0 \ldots \infty)$.

(iv)
$$\mathcal{E}^{\mathrm{Sep}}(T) = \mathcal{E}^{\mathrm{Spn}}(T) = \mathcal{E}^{\mathrm{Cov}}(T) = \mathcal{E}^{\mathrm{Card}}(T) \overset{\text{by defn}}{=} \mathcal{E}_{\mathrm{top}}(T) \,.$$

Proof (of (i)) Take $F \subset X$, a min-cardinality d_n-ε-spanning set. So $\bigcup_{z\in F} D_z = X$, where

$$D_z := \mathrm{d}_n\text{-Bal}(z, \varepsilon) \overset{\text{note}}{=} \bigcap_{j=0}^{n-1} T^{-j}\Big(\mathrm{Bal}\Big(T^j z, \varepsilon\Big)\Big) \,.$$

This $\mathcal{D} := \{D_z\}_z$ is a cover, and it is d_n–2ε–small. Thus $\mathrm{Cov}(n, 2\varepsilon) \leq |\mathcal{D}| = |F|$.

For any metric, a *maximal* ε-separated-set is automatically ε-spanning; adjoin a putative unspanned point to get a larger separated set.

Let A be a max-cardinality d_n-ε-separated set. Take $\mathcal{C}$, a min-cardinality subcover of $\mathcal{W}_{[0\ldots n)}$. For each $z \in A$, pick a $\mathcal{C}$-patch $C_z \ni z$. Could some pair $x, y \in A$ pick the same C? Well, write $C = \bigcap_{j=0}^{n-1} T^{-j}(W_j)$, with each $W_j \in \mathcal{W}$. For every $j \in [0 \ldots n)$, then,

$$\mathrm{d}(T^j(x), T^j(y)) \leq \mathrm{Diam}(W_j) < \varepsilon \,.$$

Hence $\mathrm{d}_n(x, y) < \varepsilon$; so $x = y$. Accordingly, the $z \mapsto C_z$ map is injective, whence $|A| \leq |\mathcal{C}|$. □

Proof (of (ii)) Choose a min-cardinality d_n-δ-small cover $\mathcal{C}$. For each $C \in \mathcal{C}$ and $j \in [0 \ldots n)$, the d-$\mathrm{Diam}(T^j C) < \delta$. So there is a $\mathcal{V}$-patch $V_{C,j} \supset T^j(C)$. Hence

$$\mathcal{V}_{[0\ldots n)} \overset{\text{note}}{\ni} \bigcap_{j=0}^{n-1} T^{-j}\left(V_{C,j}\right) \supset C \,.$$

Thus $\mathcal{V}_{[0\ldots n)} \preccurlyeq \mathcal{C}$. So

$$\mathrm{Card}\big(\mathcal{V}_{[0\ldots n)}\big) \leq \mathrm{Card}(\mathcal{C}) \leq |\mathcal{C}| = \mathrm{Cov}(n, \delta) \,.$$ □

Proof (of (iii)) To upper-bound $\mathrm{Cov}(k + l, \varepsilon)$ let $\mathcal{V}$ and $\mathcal{W}$ be min-cardinality ε-small covers, respectively, for metrics d_k and d_l. Then $\mathcal{V} \cap T^l(\mathcal{W})$ is a ε-small for d_{k+l}. Consequently $\mathrm{Cov}(k+l, \varepsilon) \leq \mathrm{Cov}(k, \varepsilon) \cdot \mathrm{Cov}(l, \varepsilon)$. Thus $n \mapsto \log(\mathrm{Cov}(n, \varepsilon))$ is subadditive. □

Proof (of (iv)) Pick a $\mathcal{V}$ from the list in Sect. "Metric Preliminaries", choose some $2\varepsilon < \mathrm{LEB}(\mathcal{V})$ followed by an ε-small $\mathcal{W}$ from Sect. "Metric Preliminaries". Pushing $n \to \infty$ gives

$$\mathsf{L}^{\mathrm{Card}}(\mathcal{V}) \leq \mathsf{L}^{\mathrm{Cov}}(2\varepsilon) \leq \begin{matrix} \widehat{\mathsf{L}}^{\mathrm{Spn}}(\varepsilon) \leq \widehat{\mathsf{L}}^{\mathrm{Sep}}(\varepsilon) \\ \underline{\mathsf{L}}^{\mathrm{Spn}}(\varepsilon) \leq \underline{\mathsf{L}}^{\mathrm{Sep}}(\varepsilon) \end{matrix} \leq \mathsf{L}^{\mathrm{Card}}(\mathcal{W}) \,. \tag{23}$$

Now send $\mathcal{V}$ and $\mathcal{W}$ along the list in Sect. "Metric Preliminaries". □

Pretension Topological entropy takes its values in $[0, \infty]$. A useful corollary of Eq. (23) can be stated in terms of any Distance$(\cdot, \cdot)$ which topologizes $[0, \infty]$ as a compact interval.

> For each continuous $T: X \to X$ on a compact metric-space, the Distance$\left(\widehat{\mathsf{L}}^{\mathrm{Sep}}(\varepsilon), \underline{\mathsf{L}}^{\mathrm{Sep}}(\varepsilon)\right)$ goes to zero as $\varepsilon \searrow 0$. Consequently, we can pretend that the
>
> $$\mathsf{L}^{\mathrm{Sep}}(\varepsilon) = \lim_{n\to\infty} \frac{1}{n} \log(\mathrm{Sep}(n, \varepsilon)) \tag{24}$$
>
> limit exists, in arguments that subsequently send $\varepsilon \searrow 0$. Ditto for $\mathsf{L}^{\mathrm{Spn}}(\varepsilon)$.

This will be used during the proof of the **Variational Principle**. But first, here are two entropy computations which illustrate the efficacy in having several characterizations of topological entropy.

$\mathcal{E}_{\mathrm{top}}$(Isometry) = 0 Suppose $(T: X, \mathrm{d})$ is a distance-preserving map of a compact metric-space. Fixing ε, a set is d_n-ε-separated IFF it is d-ε-separated. Thus $\mathrm{Sep}(n, \varepsilon)$ does *not* grow with n. So each $\widehat{\mathsf{L}}^{\mathrm{Sep}}(\varepsilon)$ is zero.

Topological Markov Shifts Imagine ourselves back in the days when computer data is stored on large reels of fast-moving magnetic tape. One strategy to maximize the density of binary data stored is to *not* put timing-marks (which take up space) on the tape. This has the defect that when the tape-writer writes, say, 577 consecutive 1-bits, then the tape-reader may erroneously count 578 copies of 1. We sidestep this flaw by first encoding our data so as to avoid the $11\overset{577}{\cdots}1$ word, then writing to tape.

Generalize this to a finite alphabet Q and a finite list $\mathcal{F}$ of disallowed Q-words. Extend each word to a common length $K+1$; now $\mathcal{F} \subset \mathsf{Q}^{K+1}$. The resulting "$K$-step **TMS** (***topological Markov shift***)" is the shift on the set of doubly-∞ Q-names having no substring in $\mathcal{F}$. In the above magnetic-tape example, $K = 576$. Making it more realistic, suppose that some string of zeros, say $00\overset{574}{\cdots}0$, is also forbidden[18]. Extending to length 577, we get 2^3=8 new disallowed words of form $00\overset{574}{\cdots}0b_1b_2b_3$.

We ***recode*** to a 1-step **TMS** (just called a **TMS** or a ***subshift of finite type***) over the alphabet $\mathsf{P} := \mathsf{Q}^K$. Each outlawed Q-word $w_0w_1\cdots w_K$ engenders a length-2 forbidden P-word $(w_0, \ldots, w_{K-1})(w_1, \ldots, w_K)$. The resulting **TMS** is topologically conjugate to the original K-step. The *allowed* length-2 words can be viewed as the edges in a directed-graph and the set of points $x \in X$ is the set of doubly-∞ paths through the graph. Once trivialities removed, this X is a Cantor set and the shift $T: X \to X$ is a homeomorphism.

[18] Perhaps the Ø-bad-length, 574, is shorter than the 1-bad-length because, say, Øs take less tape-space than 1s and so – being written more densely – cause ambiguity sooner.

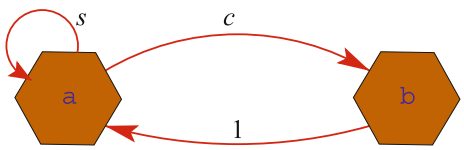

Entropy in Ergodic Theory, Figure 5
Ignoring the labels on the edges, for the moment, the *Golden shift*, T, acts on the space of doubly-infinite paths through this graph. The space can be represented as a subset $X_{\mathrm{Gold}} \subset \{\mathtt{a}, \mathtt{b}\}^{\mathbb{Z}}$, namely, the set of sequences with no two consecutive $\mathtt{b}$ letters.

The Golden Shift As the simplest example, suppose our magnetic-tape is constrained by the Markov graph, Fig. 5 that we studied measure-theoretically in Fig. 4.

We want to store the text of *The Declaration of Independence* on our magnetic tape. Imagining that English is a stationary process, we'd like to encode English into this Golden **TMS** as efficiently as possible. We seek a shift-invariant measure μ on X_{Gold} of *maximum entropy*, should such exist.

View $\mathsf{P}=\{\mathtt{a}, \mathtt{b}\}$ as the time-zero partition on X_{Gold}; that is, name $x = \ldots \mathrm{x}_{-1}\mathrm{x}_0\mathrm{x}_1\mathrm{x}_2 \ldots$, is in atom $\mathtt{b}$ IFF letter x_0 is "b". Any measure μ gives conditional probabilities

$$\mu(\mathtt{a}|\mathtt{a}) =: s\,, \qquad \mu(\mathtt{b}|\mathtt{a}) =: c\,,$$
$$\mu(\mathtt{a}|\mathtt{b}) \overset{\text{note}}{=} 1\,, \qquad \mu(\mathtt{b}|\mathtt{b}) \overset{\text{note}}{=} 0\,.$$

But recall, $\mathcal{E}(T) = \mathcal{H}(\mathsf{P}_1|\mathsf{P}_{[-\infty\ldots 0)}) \le \mathcal{H}(\mathsf{P}_1|\mathsf{P}_0)$. So among all measures that make the conditional distribution $\mathsf{P}|\mathtt{a}$ equal (s, c), the *unique* one maximizing entropy is the (s, c)-Markov-process. Its entropy, derived in Eq. (14), is

$$\begin{aligned} f(s) &:= \frac{1}{2-s}\mathcal{H}(s, 1-s) \\ &= \frac{-1}{2-s}\left[s\log(s) + (1-s)\log(1-s)\right]\,. \end{aligned} \tag{25}$$

Certainly $f(0) = f(1) = 0$, so f's maximum occurs at the (it turns out) *unique* point $\widehat{s}$ where the derivative $f'(\widehat{s})$ equals zero. This $\widehat{s} = (-1+\sqrt{5})/2$. Plugging in, the maximum entropy supportable by the Golden Shift is

$$\begin{aligned} \mathrm{MaxEnt} = \frac{2}{5-\sqrt{5}}\Bigg[&\frac{-1+\sqrt{5}}{2}\log\left(\frac{2}{-1+\sqrt{5}}\right) \\ &+ \frac{3-\sqrt{5}}{2}\log\left(\frac{2}{3-\sqrt{5}}\right)\Bigg]\,. \end{aligned} \tag{26}$$

Exponentiating, the number of μ-typical n-names grows like G^n, where

$$\mathrm{G} := \left[\frac{2}{-1+\sqrt{5}}^{\frac{-1+\sqrt{5}}{5-\sqrt{5}}}\right]\left[\frac{2}{3-\sqrt{5}}^{\frac{3-\sqrt{5}}{5-\sqrt{5}}}\right] . \tag{27}$$

This expression[19] looks unpleasant to simplify – it isn't even obviously an algebraic number – and yet topological entropy will reveal its familiar nature. This, because the Variational Principle (proved in the next section) says that the top-ent of a system is the supremum of measure-entropies supportable by the system.

Top-ent of the Golden Shift For a moment, let's work more generally on an arbitrary subshift (a closed, shift-invariant subset) $X \subset \mathsf{Q}^{\mathbb{Z}}$, where Q is a finite alphabet. Here, the transformation is always the shift – but the *space* is varying – so agree to refer to the top-ent as $\mathcal{E}_{\mathrm{top}}(X)$. Let $\mathrm{Names}_X(n)$ be the number of distinct words in the set $\{x \restriction_{[0\ldots n)} \mid x \in X\}$. Note that a metric inducing the product-topology on $\mathsf{Q}^{\mathbb{Z}}$ is

$$\mathrm{d}(x, x') := \frac{1}{1+|m|} , \tag{28}$$

for the smallest $|m|$ with $x_m \neq x'_m$.

Lemma 15 *Consider a subshift X. Then the*

$$\lim_{n\to\infty} \frac{1}{n} \log(\mathrm{Names}_X(n))$$

exists in $[0, \infty]$, *and equals* $\mathcal{E}_{top}(X)$.

Proof With $\varepsilon \in (0, 1)$ fixed, two n-names are d_n-ε-separated IFF they are not the same name. Hence $\mathrm{Sep}(n, \varepsilon) = \mathrm{Names}_X(n)$. □

To compute $\mathcal{E}_{\mathrm{top}}(X_{\mathrm{Gold}})$, declare that a word is "golden" if it appears in some $x \in X_{\mathrm{Gold}}$. Each $[n+1]$-golden word ending in $\mathtt{a}$ has form $w\mathtt{a}$, where w is n-golden. An $[n+1]$-golden word ending in $\mathtt{b}$, must end in $\mathtt{ab}$ and so has form $w\mathtt{ab}$, where w is $[n-1]$-golden. Summing up,

$$\begin{aligned}&\mathrm{Names}_{X_{\mathrm{Gold}}}(n+1)\\ &\qquad = \mathrm{Names}_{X_{\mathrm{Gold}}}(n) + \mathrm{Names}_{X_{\mathrm{Gold}}}(n-1) .\end{aligned}$$

This is the Fibonacci recurrence, and indeed, these are *the* Fibonacci numbers, since $\mathrm{Names}_{X_{\mathrm{Gold}}}(0) = 1$ and $\mathrm{Names}_{X_{\mathrm{Gold}}}(1) = 2$. Consequently, we have that

$$\mathrm{Names}_{X_{\mathrm{Gold}}}(n) \sim \mathrm{Const} \cdot \boldsymbol{\lambda}^n ,$$

where $\boldsymbol{\lambda} = \frac{1+\sqrt{5}}{2}$ is the Golden Ratio. So the sesquipedalian number G from Eq. (27) is simply $\boldsymbol{\lambda}$, and $\mathcal{E}_{\mathrm{top}}(X_{\mathrm{Gold}}) = \log(\boldsymbol{\lambda})$.

Since $\log(\boldsymbol{\lambda}) \approx 0.694$, each thousand bits written on tape (subject to the "no $\mathtt{bb}$ substrings" constraint) can carry at most 694 bits of information.

Top-ent of a General ***TMS*** A (finite) digraph G engenders a **TMS** $T\colon X_{\mathrm{G}} \to X_{\mathrm{G}}$, as well as a $\{0, 1\}$-valued adjacency matrix $\mathbf{A}{=}\mathbf{A}_{\mathrm{G}}$, where $a_{i,j}$ is the number of directed-edges from state i to j. (Here, each $a_{i,j}$ is 0 or 1.) The (i, j)-entry in power $\mathbf{A}^n$ is automatically the number of length-n paths from i to j. Employing the matrix-norm $\|\mathbf{M}\| := \sum_{i,j} |m_{i,j}|$, then,

$$\|\mathbf{A}\|^n = \mathrm{Names}_X(n) .$$

Happily Gelfand's formula (see 10.13 in [58] or Spectral_radius in [60]) applies: *For an arbitrary (square) complex matrix*,

$$\lim_{n\to\infty} \|\mathbf{A}^n\|^{\frac{1}{n}} = \mathrm{SpecRad}(\mathbf{A}) . \tag{29}$$

This right hand side, the ***spectral radius*** of $\mathbf{A}$, means the maximum of the absolute values of $\mathbf{A}$'s eigenvalues. So the top-ent of a **TMS** is thus the

$$\begin{aligned}\mathcal{E}_{\mathrm{top}}(X_{\mathrm{G}}) &= \mathrm{SpecRad}(\mathbf{A}_{\mathrm{G}})\\ &:= \mathrm{Max}\left\{|e| \;\middle|\; e \text{ is an eigenvalue of } \mathbf{A}_{\mathrm{G}}\right\} .\end{aligned} \tag{30}$$

The $(\mathtt{a}, \mathtt{b})$-adjacency matrix of Fig. 5 is

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix},$$

whose eigenvalues are $\boldsymbol{\lambda}$ and $\frac{-1}{\boldsymbol{\lambda}}$.

Labeling Edges Interpret $(s, c, 1)$ simply as edge-labels in Fig. 5. The set of doubly-∞ paths can also be viewed as a subset $Y_{\mathrm{Gold}} \subset \{s, c, 1\}^{\mathbb{Z}}$, and it too is a **TMS**. The shift on Y_{Gold} is conjugate (topologically isomorphic) to the shift on X_{Gold}, so they *a fortiori* have the same top-ent, $\log(\boldsymbol{\lambda})$. The $(s, c, 1)$-adjacency matrix is

$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} .$$

Its $|\cdot|$-largest eigenvalue is still $\boldsymbol{\lambda}$, as it must.

Now we make a new graph. We modify Fig. 5 by manufacturing a total of two s-edges, seven c-edges, and three edges $1_1, 1_2, 1_3$. Give these $2 + 7 + 3$ edges twelve *distinct* labels. We *could* compute the resulting **TMS**-entropy

[19] A popular computer-algebra-system was not, at least under my inexpert tutelage, able to simplify this. However, once top-ent gave the correct answer, the software *was* able to detect the *equality*.

from the corresponding 12×12 adjacency matrix. Alternatively, look at the $(\mathtt{a}, \mathtt{b})$-adjacency matrix

$$\mathbf{A} := \begin{bmatrix} 2 & 7 \\ 3 & 0 \end{bmatrix} .$$

The roots of its characteristic polynomial are $1 \pm \sqrt{22}$. Hence $\mathcal{E}_{\text{top}}$ of this 12-symbol **TMS** is $\log(1 + \sqrt{22})$.

The Variational Principle

Let $\mathcal{M} := \mathcal{M}(X, \mathrm{d})$ be the set of Borel probability measures, and $\mathcal{M}(T) := \mathcal{M}(T\colon X, \mathrm{d})$ the set of T-invariant $\mu \in \mathcal{M}$. Assign

$$\mathrm{EntSup}(T) := \sup \{\mathcal{E}_\mu(T) \mid \mu \in \mathcal{M}(T)\} .$$

Theorem 16 (Variational principle (Goodson)) $\mathrm{EntSup}(T) = \mathcal{E}_{\text{top}}(T)$.

This says that top-ent is the top entropy – *if* there is a measure μ which realizes the supremum. There doesn't have to be. Choose a sequence of metric-systems $(S_k\colon Y_k, \mathrm{m}_k)$ whose entropies *strictly* increase $\mathcal{E}_{\text{top}}(S_k) \nearrow L$ to some limit in $(0, \infty]$. Let $(S_\infty\colon Y_\infty, \mathrm{m}_\infty)$ be the identity-map on a 1-point space. Define a new system $(T\colon X, \mathrm{d})$, where $X := \bigsqcup_{k \in [1 \ldots \infty]} Y_k$. Have $T(x) := S_k(x)$, for the unique k with $Y_k \ni x$. As for the metric, on Y_k let d be a scaled version of m_k, so that the d-Diam(Y_k) is less than $1/2^k$. Finally, for points in *distinct* components, $x \in Y_k$ and $z \in Y_\ell$, decree that $\mathrm{d}(x, z) := \left|2^{-k} - 2^{-\ell}\right|$. Our T is continuous, and is a homeomorphism if each of the S_k is. Certainly $\mathcal{E}_{\text{top}}(T) = L > \mathcal{E}_{\text{top}}(S_k)$, for every $k \in [1 \ldots \infty]$.

If L is *finite* then there is *no* measure μ of maximal entropy; for μ must give mass to *some* Y_k; this pulls the entropy below L, since there are no compensatory components with entropy exceeding L.

In contrast, when $L = \infty$ then there *is* a maximal-entropy measure (put mass $1/2^j$ on some component Y_{k_j}, where $k_j \nearrow \infty$ swiftly); indeed, there are continuum-many maximal-entropy measures. *But* there is no [20] *ergodic* measure of maximal entropy.

For a concrete $L = \infty$ example, let S_k be the shift on $[1 \ldots k]^{\mathbb{Z}}$.

[20] The ergodic measures are the extreme points of $\mathcal{M}(T)$; call them $\mathcal{M}_{\text{Erg}}(T)$. This $\mathcal{M}(T)$ is the set of barycenters obtained from Borel probability measures on $\mathcal{M}_{\text{Erg}}(T)$ (see Krein-Milman_theorem, Choquet_theory in [60]). In this instance, what explains the failure to have an *ergodic* maximal-entropy measure? Let μ_k be an invariant ergodic measure on Y_k. These measures *do* converge to the one-point (ergodic) probability measure μ_∞ on Y_∞. But the map $\mu \mapsto \mathcal{E}_\mu(T)$ is not continuous at μ_∞.

Topology on $\mathcal{M}$ Let's arrange our tools for establishing the Variational Principle. The argument will follow Misiurewicz's proof, adapted from the presentations in [23] and [11].

Equip $\mathcal{M}$ with the ***weak-*** $*$ topology. [21] An $A \subset X$ is ***μ-nice*** if its topological boundary $\partial(A)$ is μ-null. And a *partition* is ***μ-nice*** if each atom is.

Proposition 17 *If $\alpha_L \to \mu$ and $A \subset X$ is μ-nice, then $\alpha_L(A) \to \mu(A)$.*

Proof Define operator $\mathcal{U}(D) := \limsup_L \alpha_L(D)$. It suffices to show that $\mathcal{U}(A) \le \mu(A)$. For since A^{c} is μ-nice too, then $\mathcal{U}(A^{\mathrm{c}}) \le \mu(A^{\mathrm{c}})$. Thus $\lim_L \alpha_L(A)$ exists, and equals $\mu(A)$.

Because $C := \overline{A}$ is closed, the continuous functions $f_N \searrow \mathbf{1}_C$ pointwise, where

$$f_N(x) := 1 - \mathrm{Min}(N \cdot \mathrm{d}(x, C), 1) .$$

By the Monotone Convergence theorem, then,

$$\int f_N \mathrm{d}\mu \xrightarrow{N} \mu(C) .$$

And $\mu(C) = \mu(A)$, since A is nice. Fixing N, then, it suffices to establish $\mathcal{U}(A) \le \int f_N \mathrm{d}\mu$. But f_N is continuous, so

$$\begin{aligned} \int f_N \mathrm{d}\mu &= \limsup_{L \to \infty} \int f_N \mathrm{d}\alpha_L \\ &\ge \limsup_{L \to \infty} \int \mathbf{1}_A \mathrm{d}\alpha_L = \mathcal{U}(A) . \end{aligned}$$ □

Corollary 18 *Suppose $\alpha_L \to \mu$, and partition P is μ-nice. Then $\mathcal{H}_{\alpha_L}(\mathsf{P}) \to \mathcal{H}_\mu(\mathsf{P})$.*

The ***diameter*** of partition P is $\mathrm{Max}_{A \in \mathsf{P}} \mathrm{Diam}(A)$.

Proposition 19 *Take $\mu \in \mathcal{M}$ and $\varepsilon > 0$. Then there exists a μ-nice partition with Diam$(\mathsf{P}) < \varepsilon$.*

Proof Centered at an x, the uncountably many balls $\{\mathrm{Bal}(x, r) \mid r \in (0, \varepsilon)\}$ have disjoint boundaries. So all but countably many are μ-nice; pick one and call it B_x. Compactness gives a finite nice cover, say, $\{B_1, \ldots, B_7\}$, at different centers. Then the partition $\mathsf{P} := (A_1, \ldots, A_7)$ is nice, [22] where $A_k := B_k \setminus \bigcup_{j=1}^{k-1} B_j$. □

Here is a consequence of Jensen's inequality.

[21] Measures $\alpha_L \to \mu$ IFF $\int f \mathrm{d}\alpha_L \to \int f \mathrm{d}\mu$, for each continuous $f\colon X \to \mathbb{R}$. This metrizable topology makes $\mathcal{M}$ compact. Always, $\mathcal{M}(T)$ is a non-void compact subset (see ▶ Measure Preserving Systems.)

[22] For any two sets $B, B' \subset X$, the union $\partial B \cup \partial B'$ is a superset of the three boundaries $\partial(B \cup B')$, $\partial(B \cap B')$, $\partial(B \setminus B')$.

Lemma 20 (Distropy-Averaging Lemma) *For $\mu, \nu \in \mathcal{M}$, a partition R, and a number $t \in [0,1]$,*

$$t\mathcal{H}_\mu(R) + t^c\mathcal{H}_\nu(R) \le \mathcal{H}_{t\mu+t^c\nu}(R)\,.$$

Strategy for $\mathrm{EntSup}(T) \ge \mathcal{E}_{\mathrm{top}}(T)$. Choose an $\varepsilon > 0$. For $L = 1, 2, 3, \ldots$, take a maximal (L, ε)–separated–set $F_L \subset X$, then define

$$\mathbf{F} = \mathbf{F}_\varepsilon := \limsup_{L\to\infty} \frac{1}{L}\log(|F_L|)\,.$$

Let $\varphi_L()$ be the equi-probable measure on F_L; each point has weight $1/|F_L|$. The desired invariant measure μ will come from the Cesàro averages

$$\alpha_L := \frac{1}{L}\sum_{\ell\in[0\ldots L)} T^\ell\varphi_L\,,$$

which get more and more invariant.

Lemma 21 *Let μ be any weak-$*$ accumulation point of the above $\{\alpha_L\}_1^\infty$. (Automatically, μ is T-invariant.) Then $\mathcal{E}_\mu(T) \ge \mathbf{F}$. Indeed, if Q is any μ-nice partition with $\mathrm{Diam}(\mathsf{Q}) < \varepsilon$, then $\mathcal{E}_\mu(T, \mathsf{Q}) \ge \mathbf{F}$.*

Tactics As usual, $\mathsf{Q}_{[0\ldots N)}$ means $\mathsf{Q}_0 \vee \mathsf{Q}_1 \vee \ldots \vee \mathsf{Q}_{N-1}$. Our goal is

$$\forall N: \quad \mathbf{F} \overset{?}{\le} \frac{1}{N}\mathcal{H}_\mu\big(\mathsf{Q}_{[0\ldots N)}\big)\,. \tag{31}$$

Fix N and $\mathsf{P} := \mathsf{Q}_{[0\ldots N)}$, and a $\delta > 0$. It suffices to verify: $\forall_{\mathrm{large}} L \gg N$,

$$\frac{1}{L}\log(|F_L|) \overset{?}{\le} \delta + \frac{1}{N}\mathcal{H}_{\alpha_L}(\mathsf{P})\,, \tag{32}$$

since this and Corollary 17 will prove Eq. (31): Pushing $L \to \infty$ along the sequence that produced μ essentially sends LhS(32) to $\mathbf{F}$, courtesy Eq. (24). And RhS(32) goes to $\delta + \frac{1}{N}\mathcal{H}_\mu(\mathsf{P})$, by Corollary 17, since P is μ-nice. Descending $\delta \searrow 0$, hands us the needed Eq. (31).

Remark 22 The idea in the following proof is to mostly fill interval $[0..L)$ with N-blocks, starting with a offset $K \in [0..N)$. Averaging over the offset will create a Cesàro average over each N-block. Averaging over the N-blocks will allow us to compute distropy with respect to the averaged measure, α_L.

Proof (of Eq. (32)) Since L is frozen, agree to use φ for the φ_L probability measure.

Our d_L-ε-separated set F_L has at most *one* point in any given atom of $\mathsf{Q}_{[0\ldots L)}$, thereupon

$$\log(|F_L|) = \mathcal{H}_\varphi\big(\mathsf{Q}_{[0\ldots L)}\big)\,.$$

Regardless of the "offset" $K \in [0\ldots N)$, we can always fit $C := \lfloor \frac{L-N}{N} \rfloor$ many N-blocks into $[0\ldots L)$. Denote by $\mathcal{G}(K) := [K\ldots K+CN)$, this union of N-blocks, the ***good*** set of indices. Unsurprisingly, $\mathcal{B}(K) := [0\ldots L) \smallsetminus \mathcal{G}(K)$ is the ***bad*** index-set. Therefore,

$$\mathcal{H}_\varphi\big(\mathsf{Q}_{[0\ldots L)}\big) \le \overbrace{\mathcal{H}_\varphi\Big(\bigvee_{j\in\mathcal{B}(K)}\mathsf{Q}_j\Big)}^{\mathrm{Bad}(K)} + \overbrace{\mathcal{H}_\varphi\Big(\bigvee_{j\in\mathcal{G}(K)}\mathsf{Q}_j\Big)}^{\mathrm{Good}(K)}\,. \tag{33}$$

Certainly $\mathrm{Bad}(K) \le 3N\log(|\mathsf{Q}|)$. So

$$\frac{1}{NL}\sum_{K\in[0\ldots N)}\mathrm{Bad}(K) \le \frac{3N}{L}\log(|\mathsf{Q}|)\,.$$

This is less than δ, since L is large. Applying $\frac{1}{NL}\sum_{K\in[0\ldots N)}$ to Eq. (28) now produces

$$\frac{1}{L}\log(|F_L|) \le \delta + \frac{1}{NL}\sum_K \mathrm{Good}(K)\,. \tag{34}$$

Note

$$\bigvee_{j\in\mathcal{G}(K)} T^j(\mathsf{Q}) = \bigvee_{c\in[0\ldots C)} T^{K+cN}(\mathsf{P})\,.$$

So

$$\mathrm{Good}(K) \le \sum_c \mathcal{H}_\varphi(T^{K+cN}\mathsf{P})\,.$$

This latter, by definition, equals $\sum_c \mathcal{H}_{T^{K+cN}(\varphi)}(\mathsf{P})$. We conclude that

$$\begin{aligned}\frac{1}{NL}\sum_K \mathrm{Good}(K) &\le \frac{1}{NL}\sum_K\sum_c \mathcal{H}_{T^{K+cN}\varphi}(\mathsf{P})\\ &\le \frac{1}{NL}\sum_{\ell\in[0\ldots L)} \mathcal{H}_{T^\ell\varphi}(\mathsf{P})\,,\\ &\qquad \text{by adjoining a few translates of } \mathsf{P},\\ &\le \frac{1}{N}\mathcal{H}_{\alpha_L}(\mathsf{P})\,,\\ &\qquad \text{by the Distropy-averaging lemma 18,}\end{aligned}$$

since α_L is the average $\frac{1}{L}\sum_\ell T^\ell\varphi$. Thus Eq. (34) implies Eq. (32), our goal. □

Proof (of $\mathrm{EntSup}(T) \le \mathcal{E}_{\mathrm{top}}(T)$*)* Fix a T-invariant μ. For partition $\mathsf{Q} = (B_1, \ldots, B_K)$, choose a compact set $A_k \subset B_k$ with $\mu(B_k \setminus A_k)$ small. (This can be done, since μ is automatically a regular measure [58].) Letting $D := \left[\bigsqcup_i A_i\right]^{\mathrm{c}}$ and $\mathsf{P} := (D, A_1, \ldots, A_K)$, we can have made $\mathcal{H}(\mathsf{P}|\mathsf{Q})$ as small as desired. Courtesy of Lemma 8b, then, we only need consider partitions of the form that P has.

Open-cover $\mathcal{V} = (U_1, \ldots, U_K)$ has patches $U_k := D \cup A_k$. What atoms of, say, $\mathsf{P}_{[0\ldots3)}$, can the intersection $U_9 \cap T^{-1}(U_2) \cap T^{-2}(U_5)$ touch? Only the eight atoms

$$(D \text{ or } A_9) \cap T^{-1}(D \text{ or } A_2) \cap T^{-2}(D \text{ or } A_5)\,.$$

Thus $^\sharp\mathsf{P}_{[0\ldots n)} \le 2^n \cdot {}^\sharp\mathcal{V}_{[0\ldots n)}$. (Here, $^\sharp()$ counts the number of non-void atoms/patches.) So

$$\begin{aligned}\frac{1}{n}\mathcal{H}_\mu\left(\mathsf{P}_{[0\ldots n)}\right) &\le 1 + \frac{1}{n}\log\left({}^\sharp\mathcal{V}_{[0\ldots n)}\right)\\ &\le 1 + 1 + \mathcal{E}_{\mathrm{top}}(T)\,;\end{aligned}$$

this last inequality, when n is large. The upshot: $\mathcal{E}_\mu(T) \le 2 + \mathcal{E}_{\mathrm{top}}(T)$.

Applied to a power T^ℓ, this asserts that $\mathcal{E}_\mu(T^\ell) \le 2 + \mathcal{E}_{\mathrm{top}}(T^\ell)$. Thus

$$\mathcal{E}_\mu(T) \le \frac{2}{\ell} + \mathcal{E}_{\mathrm{top}}(T)\,,$$

using Lemma 9d and using Lemma 13g. Now coax $\ell \to \infty$. □

Three Recent Results

Having given an survey of older results in measure-theoretic entropy and in topological entropy, let us end this survey with a brief discussion of a few recent results, chosen from many.

Ornstein–Weiss: Finitely-Observable Invariant

In a landmark paper [10], Ornstein and Weiss show that all "finitely observable" properties of ergodic processes are secretly entropy; indeed, they are continuous functions of entropy. This was generalized by Gutman and Hochman [9]; some of the notation below is from their paper.

Here is the setting. Consider an ergodic process, on a non-atomic space, taking on only finitely many values in $\mathbb{N}$; let $\mathcal{C}$ be some family of such processes. An ***observation scheme*** is a metric space (Ω, d) and a sequence of functions $\mathbf{S} = (S_n)_1^\infty$, where S_n maps $\mathbb{N} \times \overset{n}{\ldots} \times \mathbb{N}$ into Ω. On a point $\vec{x} \in \mathbb{N}^\infty$, the scheme ***converges*** if

$$n \mapsto S_n(x_1, x_2, \ldots x_n) \tag{35}$$

converges in Ω. And on a particular process $\mathcal{X}$, say that $\mathbf{S}$ ***converges***, if $\mathbf{S}$ converges on a.e. $\vec{x}$ in $\mathcal{X}$.

A function $J\colon \mathcal{C} \to \Omega$ is isomorphism invariant if, whenever the underlying transformations of two processes $\mathcal{X}, \mathcal{X}' \in \mathcal{C}$ are isomorphic, then $J(\mathcal{X}) = J(\mathcal{X}')$. Lastly, say that $\mathbf{S}$ "converges to J", if for each $\mathcal{X} \in \mathcal{C}$, scheme $\mathbf{S}$ converges to the value $J(\mathcal{X})$.

The work of David Bailey [38], a student of Ornstein, produced an observation scheme for entropy. The Lempel-Ziv algorithm [43] was another entropy observer, with practical application.

Ornstein and Weiss provided entropy schemes in [41] and [42]. Their recent paper *"Entropy is the only finitely-observable invariant"* [10], gives a converse, a uniqueness result.

Theorem 23 (Ornstein, Weiss) *Suppose J is a finitely observable function, defined on all ergodic finite-valued processes. If J is an isomorphism invariant, then J is a continuous function of the entropy.*

Gutman–Hochman: Finitely-Observable Extension

Extending the Ornstein–Weiss result, Yonatan Gutman and Michael Hochman, in [9] proved that it holds even when the isomorphism invariant, J, is well-defined only on certain subclasses of the set of all ergodic processes. In particular they obtain the following result on three classes of zero-entropy transformations.

Theorem 24 (Gutman, Hochman) *Suppose $J()$ is a finitely observable invariant on one of the following classes:*

(i) The Kronecker systems; the class of systems with pure point spectrum.
(ii) The zero-entropy mild mixing processes.
(iii) The zero-entropy strong mixing processes.

Then $J()$ is constant.

Entropy of Actions of Free Groups

Consider $(G, \mathcal{G})$, a topological group and its Borel field (sigma-algebra). Let $\mathcal{G} \times \mathcal{X}$ be the field on $G \times X$ generated by the two coordinate-subfields. A map

$$\phi\colon GxX \to X \tag{36}$$

is ***measurable*** if

$$\psi^{-1}(\mathcal{X}) \subset \mathcal{G} \times \mathcal{X}\,.$$

Use $\psi^g(x)$ for $\psi(g, x)$.

This map in Eq. (36) is a (measure-preserving) ***group action*** if $\forall g, h \in G$: $\psi^g \circ \psi^h = \psi^{gh}$, and each $\psi^g : X \to X$ is measure preserving.

This encyclopedia article has only discussed entropy for $\mathbb{Z}$-actions, i. e., when $G = \mathbb{Z}$. The ergodic theorem, our definition of entropy, and large parts of ergodic theory, involve taking averages (of some quantity of interest) over larger and larger "pieces of time". In $\mathbb{Z}$, we typically use the intervals $I_n := [0 \ldots n)$. When G is $\mathbb{Z} \times \mathbb{Z}$, we might average over squares $I_n \times I_n$.

The *amenable groups* are those which possess, in a certain sense, larger and larger averaging sets. Parts of ergodic theory have been carried over to actions of amenable groups, e. g. [49] and [55]. Indeed, much of the Bernoulli theory was extended to certain amenable groups by Ornstein and Weiss, [50].

The stereotypical example of a *non*-amenable group, is a free group (on more than one generator). But recently, Lewis Bowen [8] succeeded in extending the definition of entropy to actions of finite-rank free groups.

Theorem 25 (Lewis Bowen) *Let G be a finite-rank free group. Then two Bernoulli G-actions are isomorphic IFF they have the same entropy.*

The paper introduces a new isomorphism invariant, the "f invariant", and shows that, for Bernoulli actions, the f invariant agrees with entropy, that is, with the distropy of the independent generating partition.

Exodos

Ever since the pioneering work of Shannon, and of Kolmogorov and Sinai, entropy has been front and center as a major tool in Ergodic Theory. Simply *mentioning* all the substantial results in entropy theory would dwarf the length of this encyclopedia article many times over. And, as the above three results (cherry-picked out of many) show, Entropy shows no sign of fading away…

Bibliography

Historical

1. Adler R, Weiss B (1967) Entropy, a complete metric invariant for automorphisms of the torus. Proc Natl Acad Sci USA 57:1573–1576
2. Clausius R (1864) Abhandlungen ueber die mechanische Wärmetheorie, vol 1. Vieweg, Braunschweig
3. Clausius R (1867) Abhandlungen ueber die mechanische Wärmetheorie, vol 2. Vieweg, Braunschweig
4. Kolmogorov AN (1958) A new metric invariant of transitive automorphisms of Lebesgue spaces. Dokl Akad Nauk SSSR 119(5):861–864
5. McMillan B (1953) The basic theorems of information theory. Ann Math Stat 24:196–219
6. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423,623–656
7. Sinai Y (1959) On the Concept of Entropy of a Dynamical System, Dokl Akad Nauk SSSR 124:768–771

Recent Results

8. Bowen L (2008) A new measure-conjugacy invariant for actions of free groups. http://www.math.hawaii.edu/%7Elpbowen/notes11.pdf
9. Gutman Y, Hochman M (2006) On processes which cannot be distinguished by finitary observation. http://arxiv.org/pdf/math/0608310
10. Ornstein DS, Weiss B (2007) Entropy is the only finitely-observable invariant. J Mod Dyn 1:93–105; http://www.math.psu.edu/jmd

Ergodic Theory Books

11. Brin M, Stuck G (2002) Introduction to dynamical systems. Cambridge University Press, Cambridge
12. Cornfeld I, Fomin S, Sinai Y (1982) Ergodic theory. Grundlehren der Mathematischen Wissenschaften, vol 245. Springer, New York
13. Friedman NA (1970) Introduction to ergodic theory. Van Nostrand Reinhold, New York
14. Halmos PR (1956) Lectures on ergodic theory. The Mathematical Society of Japan, Tokyo
15. Katok A, Hasselblatt B (1995) Introduction to the modern theory of dynamical systems. (With a supplementary chapter by Katok and Leonardo Mendoza). Encyclopedia of Mathematics and its Applications, vol 54. Cambridge University Press, Cambridge
16. Keller G, Greven A, Warnecke G (eds) (2003) Entropy. Princeton Series in Applied Mathematics. Princeton University Press, Princeton
17. Lind D, Marcus B (1995) An introduction to symbolic dynamics and coding. Cambridge University Press, Cambridge,
18. Mañé R (1987) Ergodic theory and differentiable dynamics. Ergebnisse der Mathematik und ihrer Grenzgebiete, ser 3, vol 8. Springer, Berlin
19. Parry W (1969) Entropy and generators in ergodic theory. Benjamin, New York
20. Petersen K (1983) Ergodic theory. Cambridge University Press, Cambridge
21. Rudolph DJ (1990) Fundamentals of measurable dynamics. Clarendon Press, Oxford
22. Sinai Y (1994) Topics in ergodic theory. Princeton Mathematical Series, vol 44. Princeton University Press, Princeton
23. Walters P (1982) An introduction to ergodic theory. Graduate Texts in Mathematics, vol 79. Springer, New York

Differentiable Entropy

24. Ledrappier F, Young L-S (1985) The metric entropy of diffeomorphisms. Ann Math 122:509–574

25. Pesin YB (1977) Characteristic Lyapunov exponents and smooth ergodic theory. Russ Math Surv 32:55–114
26. Young L-S (1982) Dimension, entropy and Lyapunov exponents. Ergod Theory Dyn Syst 2(1):109–124

Finite Rank

27. Ferenczi S (1997) Systems of finite rank. Colloq Math 73(1):35–65
28. King JLF (1988) Joining-rank and the structure of finite rank mixing transformations. J Anal Math 51:182–227

Maximal-Entropy Measures

29. Buzzi J, Ruette S (2006) Large entropy implies existence of a maximal entropy measure for interval maps. Discret Contin Dyn Syst 14(4):673–688
30. Denker M (1976) Measures with maximal entropy. In: Conze J-P, Keane MS (eds) Théorie ergodique, Actes Journées Ergodiques, Rennes, 1973/1974. Lecture Notes in Mathematics, vol 532. Springer, Berlin, pp 70–112
31. Misiurewicz M (1973) Diffeomorphism without any measure with maximal entropy. Bull Acad Polon Sci Sér Sci Math Astron Phys 21:903–910

Topological Entropy

32. Adler R, Marcus B (1979) Topological entropy and equivalence of dynamical systems. Mem Amer Math Soc 20(219)
33. Adler RL, Konheim AG, McAndrew MH (1965) Topological entropy. Trans Am Math Soc 114(2):309–319
34. Bowen R (1971) Entropy for group endomorphisms and homogeneous spaces. Trans Am Math Soc 153:401–414. Errata 181:509–510 (1973)
35. Bowen R (1973) Topological entropy for noncompact sets. Trans Am Math Soc 184:125–136
36. Dinaburg EI (1970) The relation between topological entropy and metric entropy. Sov Math Dokl 11:13–16
37. Hasselblatt B, Nitecki Z, Propp J (2005) Topological entropy for non-uniformly continuous maps. http://www.citebase.org/abstract?id=oai:arXiv.org:math/0511495

Determinism and Zero-Entropy, and Entropy Observation

38. Bailey D (1976) Sequential schemes for classifying and predicting ergodic processes. Ph D Dissertation, Stanford University
39. Kalikow S, King JLF (1994) A countably-valued sleeping stockbroker process. J Theor Probab 7(4):703–708
40. King JLF (1992) Dilemma of the sleeping stockbroker. Am Math Monthly 99(4):335–338
41. Ornstein DS, Weiss B (1990) How sampling reveals a process. Ann Probab 18(3):905–930
42. Ornstein DS, Weiss B (1993) Entropy and data compression schemes. IEEE Trans Inf Theory 39(1):78–83
43. Ziv J, Lempel A (1977) A universal algorithm for sequential data compression. IEEE Trans Inf Theory 23(3):337–343

Bernoulli Transformations, K-Automorphisms, Amenable Groups

44. Berg KR (1975) Independence and additive entropy. Proc Am Math Soc 51(2):366–370; http://www.jstor.org/stable/2040323
45. Meshalkin LD (1959) A case of isomorphism of Bernoulli schemes. Dokl Akad Nauk SSSR 128:41–44
46. Ornstein DS (1970) Bernoulli shifts with the same entropy are isomorphic. Adv Math 5:337–352
47. Ornstein DS (1974) Ergodic theory randomness and dynamical systems, Yale Math Monographs, vol 5. Yale University Press, New Haven
48. Ornstein DS, Shields P (1973) An uncountable family of K-automorphisms. Adv Math 10:63–88
49. Ornstein DS, Weiss B (1983) The Shannon–McMillan–Briman theorem for a class of amenable groups. Isr J Math 44(3):53–60
50. Ornstein DS, Weiss B (1987) Entropy and isomorphism theorems for actions of amenable groups. J Anal Math 48:1–141
51. Shields P (1973) The theory of Bernoulli shifts. University of Chicago Press, Chicago
52. Sinai YG (1962) A weak isomorphism of transformations having an invariant measure. Dokl Akad Nauk SSSR 147:797–800
53. Thouvenot J-P (1975) Quelques propriétés des systèmes dynamiques qui se décomposent en un produit de deux systèmes dont l'un est un schéma de Bernoulli. Isr J Math 21:177–207
54. Thouvenot J-P (1977) On the stability of the weak Pinsker property. Isr J Math 27:150–162

Abramov Formula

55. Ward T, Zhang Q (1992) The Abramov–Rohlin entropy addition formula for amenable group actions. Monatshefte Math 114:317–329

Miscellaneous

56. Newhouse SE (1989) Continuity properties of entropy. Ann Math 129:215–235
57. http://www.isib.cnr.it/control/entropy/
58. Rudin W (1973) Functional analysis. McGraw-Hill, New York
59. Tribus M, McIrvine EC (1971) Energy and information. Sci Am 224:178–184
60. Wikipedia, http://en.wikipedia.org/wiki/. pages: http://en.wikipedia.org/wiki/Spectral_radius, http://en.wikipedia.org/wiki/Information_entropy

Books and Reviews

Boyle M, Downarowicz T (2004) The entropy theory of symbolic extensions. Invent Math 156(1):119–161
Downarowicz T, Serafin J (2003) Possible entropy functions. Isr J Math 135:221–250
Hassner M (1980) A non-probabilistic source and channel coding theory. Ph D Dissertation, UCLA
Katok A, Sinai YG, Stepin AM (1977) Theory of dynamical systems and general transformation groups with an invariant measure. J Sov Math 7(6):974–1065

Entropy Maximization and Species Abundance

BILL SHIPLEY
Département de Biologie, Université de Sherbrooke, Sherbrooke, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
Information Theory Basis of Entropy Maximization
Fixed Species Pools: Predicting Community Composition
Species Abundance Distributions
Future Directions
Bibliography

Glossary

Species richness the number of species per unit area; a synonym is species density.

Species abundance the number of individuals or amount of living biomass per species per unit area.

Per capita instantaneous growth rate the net number of new individuals produced by an average individual in a population at time t.

Deterministic chaos a pattern of change over time that is perfectly determined by initial conditions, but for which tiny changes in initial conditions result is divergences in the dynamics such that the pattern appears random.

Attractor a set of values (a point, a curve or higher – including a fractal – dimensional object) towards which a dynamic trajectory will move.

Basin of attraction the set of values from which any dynamic trajectory will move towards the same attractor.

Maximum entropy formalism a formal method for producing probability estimates that agree with certain constraints specifying available information about a system but that are otherwise maximally uninformative.

Maximally uninformative (or ignorance) prior a probability distribution that encodes the basic structure of a logical problem as given by the definition of the variable, but that contains no other information.

Definition of the Subject

An understanding of the determinants of biodiversity is an important goal of academic ecology and the protection of biodiversity is an important goal in conservation. Biodiversity has two components: the number (or density) of species and the relative abundance of species. Both components vary over different spatial and temporal scales. The allocation of limiting resources to different species in an ecological community is reflected in the abundance of each species. The notion of a "community" in ecology (as in more general usages of this word) is poorly defined but will here mean the total number of organisms found within a fixed area of space (a site). Abundance is measured in different ways but the most exact measure is the total mass of living tissues (i. e. biomass) found in a given species at the site. Relative abundance is the proportional abundance of each species relative to the total abundance of all species at the site. The species "pool" is the set of species that can potentially disperse into the site, including those that are not found within the community because they have been excluded due either to biotic interactions or to an inability to survive the abiotic conditions of the site. The vector of abundances, $\boldsymbol{a} = \{a_i\}$, or relative abundances, $\boldsymbol{ra} = \{ra_i\}$, of each species in the species pool that is observed at the site is a description of the community structure. The change in this vector over time is a description of the community dynamics.

As such, community structure and dynamics are fundamental properties of ecological communities and relate to basic questions posed by ecologists: Which species will be found at a site? Which of these will be rare? Which species will dominate? How many species will be found? How is the flow of matter and energy through the organisms affected by community structure? How will the community structure change if the environmental conditions of the site change? Every one of these questions involves complex dynamics of large numbers of interactions of organisms. Because of this, answers to these questions will have to deal with this complexity. Most of the results that follow are based on plant communities, since these are the only ones that have been studied to date using entropy maximization, but animal communities will be discussed when looking at future directions. The answers to such questions have important applied implications for forestry, agriculture and conservation.

Introduction

Four aspects of ecological communities are obvious to even a casual observer of nature. First, the organisms found at any given site are distributed into different species. Second, the distribution of individuals and biomass, and therefore of resources, is not equal between species. It is not simply that some species are common and

some are rare, but that a small number of species in every community make up the majority of the biomass and most species in every community make up a minority of the biomass. Third, which species are common, which are rare, and which are absent, differ across sites having different physical (abiotic) environmental conditions. Fourth, which species are common, which are rare, and which are absent, change systematically at the same site over time following a major disturbance (i. e. ecological succession). Explaining and predicting these four properties of ecological communities is a major goal of ecological science.

Although much progress has been made at the empirical level, a general theoretical explanation of these properties with good predictive ability under field conditions is still largely missing despite almost a century of effort by influential theoretical ecologists. All of these theoretical efforts were based on dynamic models derived from population-level demography and based on an analogy with Newtonian mechanics. The purpose of this essay is to contrast this majority tradition in ecological modeling of communities with a more recent statistical mechanistic approach based on maximization of entropy or relative entropy.

Population dynamic models begin with two undeniable properties of any biological population. First, because each mature member of the population can potentially reproduce, population growth is a multiplicative process defining a geometric series. This is the first fundamental insight of Malthus that inspired Darwin [1] and Wallace [2] to propose the theory of evolution by natural selection: every biological population has the *potential* for exponential growth. If $n_i(t)$ is the number of individuals of species i alive at time t, then the per capita instantaneous growth rate at time t, $r_i(t)$, is defined as:

$$\frac{1}{n_i(t)}\frac{\mathrm{d}n_i(t)}{\mathrm{d}t} = r_i(t) \ . \tag{1}$$

If $r_i(t)$ is constant over time then integration of Eq. (1) leads to exponential growth: $n_i(t) = n_i(0)\mathrm{e}^{r_i t}$. Per capita growth rates can also be expressed as difference equations by an obvious modification of Eq. (1). The second undeniable property of any biological population is that exponential growth cannot persist indefinitely without resources becoming exhausted. For instance a single bacterium with a mass of 1 picogram that divides by binary fission once every 20 minutes would, if this rate of growth continued without interruption, produce offspring weighting the mass of the Earth in a little under two days. Therefore there must be a negative feedback such that $r_i(t)$ decreases in some fashion as population size, $n_i(t)$, increases. The simplest type of feedback leads to the well-known logistic (or Lotka–Volterra) equation, shown in Eq. (2). Lotka [3] published his model in 1926 although the original idea is due to Pierre–François Verhulst [4]. Volterra [5] published the genesis of this model in 1926. Hutchinson [4] gives a good historical description of the origins of this equation.

$$\frac{1}{n_i(t)}\frac{\mathrm{d}n_i(t)}{\mathrm{d}t} = r_i\left(\frac{K_i - \alpha_{ii}n_i(t)}{K_i}\right) \ . \tag{2}$$

Here r_i is the maximum per capita growth rate, attained when $n_i \to 0$ and therefore when all competition disappears. K_i is the carrying capacity, attained when negative feedback reduces birth rates and increases death rates such that per capita birth and death rates are equal and leading to a constant population size with $r_i(t) = 0$. The coefficient α_{ii} (the per capita competition coefficient) is the amount by which each additional individual of species i decreases the per capita growth rate below the maximum (r_i). The extension to a community of S species leads to a system of simultaneous differential equation that are shown in Eq. (3), where the $S \times S$ matrix of competition coefficients is called the community competition matrix.

$$\frac{1}{n_i(t)}\frac{\mathrm{d}n_i(t)}{\mathrm{d}t} = r_i\left(\frac{K_i - \sum_{j=1}^{S}\alpha_{ij}n_i(t)}{K_i}\right) \ . \tag{3}$$

Equation (3) has been modified in many ways, including the addition of time lags, of nonlinear terms, of replacing constants by stochastic terms [6], and replacing the generalized logistic equation by other feedback functions. A more recent modification has been to replace the phenomenological competition coefficients by a dynamic coupling of organisms having a population feedback following a Monod-type equation with an explicit description of dynamic changes in resource levels [7]. This essentially extends the demographic approach to resources as well as organisms. Despite this body of theoretical work, the ability of such demographically-based models to actually predict the structure and dynamics of real ecological communities under field conditions is very weak. In fact, even simplified multispecies communities of phytoplankton [8] or fruit flies reared under controlled laboratory conditions [9] are not well described by these equations despite their intuitive logic. Of course, given the large number of parameters that must be estimated in such models, they are not applicable to real ecological communities in practice.

Recent theoretical and empirical work, involving very simple one-species systems in controlled laboratory conditions, has suggested one reason why such demographic models have low predictive ability even in simplified sys-

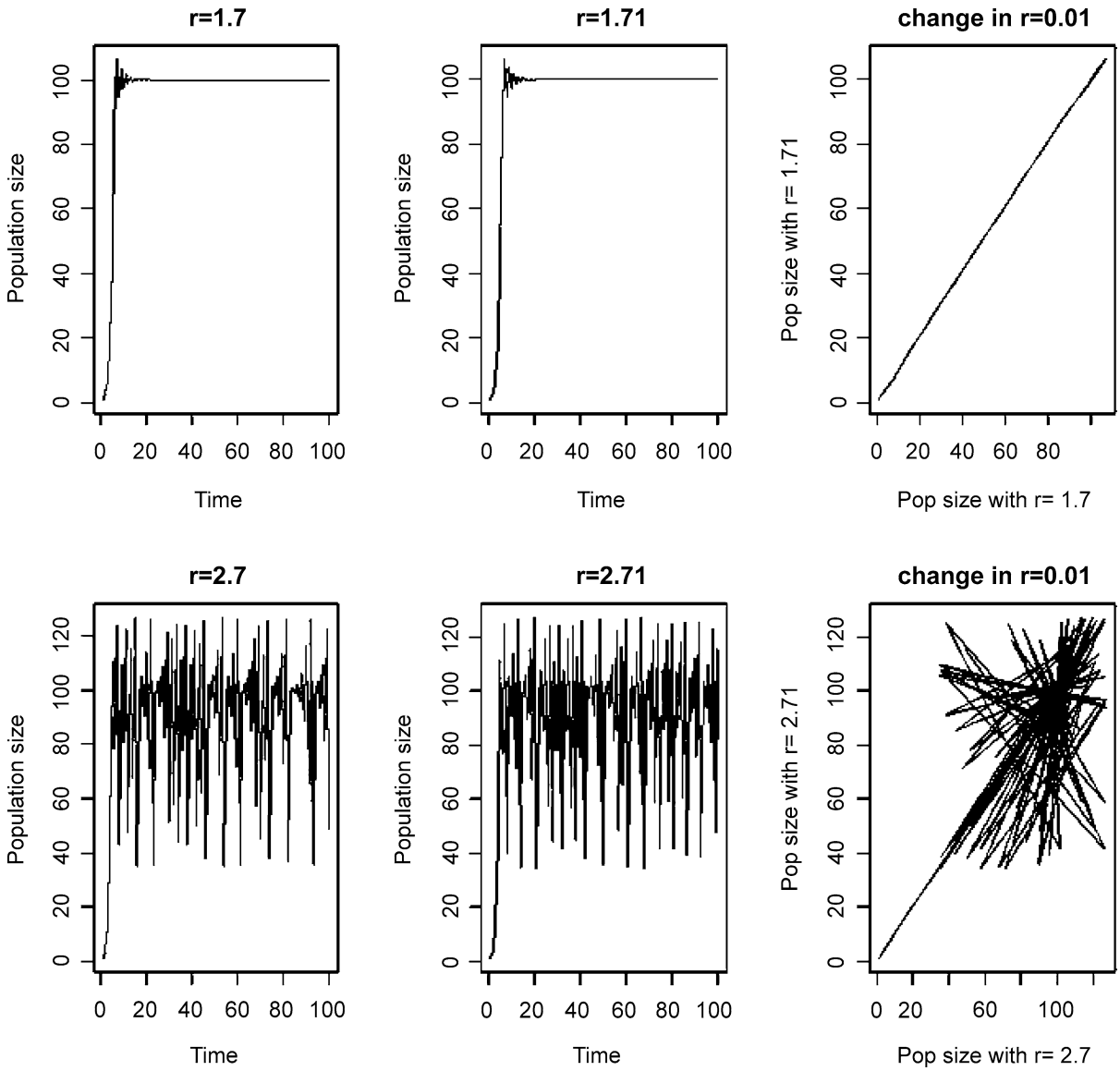

Entropy Maximization and Species Abundance, Figure 1

Time series of the deterministic discrete logistic equation showing extreme sensitivity to initial conditions. *Top row*: Time series with the maximum per capita growth rate (*r*) set at either 1.7 or 1.71 and the relationship of population sizes at each time *t*. *Bottom row*: Time series with the maximum per capita growth rate (*r*) set at either 2.7 or 2.71 and the relationship of population sizes at each time *t*. Note that the same change in the maximum per capita growth rate ($\Delta = 0.01$) will produce changes in population sizes that are very small (*top right*) or very large and essentially random (*bottom right*)

tems. It has been known since May [10] that the simple logistic equation (Eq. (1)) can display deterministic chaos over some ranges of the three parameters (r, α, K). In fact, the basic structure of all such dynamic demographic models in ecology involves nonlinearity and feedback and these two properties are necessary for deterministic chaos. Thus, such models can be sensitive to initial conditions because of the presence of chaotic attractors in phase space.

To illustrate the problem, consider the logistic equation for a single population written as a difference equation: $N(t+1) = N(t) + N(t)\, r\, ((K - N(t))/(K))$. There are no random variables in this equation and, given the values of r, K and $N(0)$, the subsequent population sizes, $N(t)$, over time are completely deterministic and predictable. The first row of Fig. 1 (top row) shows, from left to right, the changes in population size over time when

$r = 1.7$, when $r = 1.71$ and the relationship between the population sizes at each time when comparing these two dynamics. We see the classic discrete version of the logistic equation and the tiny change (0.01) in the value of r is mirrored by the tiny change in the population sizes over time. The second row of Fig. 1 shows the same thing except that the values of r are 2.7 versus 2.71. The behavior of the population size over time looks completely random. More importantly, the same tiny change in r (0.01) results in trajectories of population size that rapidly diverge and that quickly become essentially independent. In other words, making an error of 0.01 in the estimation of the per capita growth rate would result in completely incorrect predictions of population size. Yet, if r were measured in a real field population, a difference of 0.01 would be far less than the measurement error even for a huge sample size. In such a case the population dynamics is, for all practical purposes, unpredictable. Depending on the equation and the part of parameter space that is being explored, even changes in parameter values of 10^{-6} can produce equally spectacular differences; this is an ecological example of Edward Lorenz' famous "butterfly effect" [11].

A good real-life example of this is the series of modeling and experimental studies described in Constantino et al. [12]. The experimental system was extremely simple: Flour beatles were grown over many generations in standardized containers in the laboratory. Each container received a fixed amount of food at regular intervals. Temperature, humidity and other environmental conditions are maintained at fixed values. A simple nonlinear dynamic model was developed to capture the essential elements of the population dynamics of this species under these carefully controlled conditions. An analysis of the model showed which areas of parameter space resulted in chaotic dynamics (as well as other dynamic properties) and the experimental conditions were then manipulated according to the model predictions in order to bring the measured parameters into these regions of parameter space. The experiments reproduced the various predictions of chaotic dynamics and showed that even very small differences in parameter values could quickly lead to dynamic trajectories that rapidly diverge and soon become indistinguishable from independent random variables.

These experimental results are very bad news for models attempting to model community assembly from this reductionist perspective of population dynamics. If even such a simplified system – a single species growing in controlled conditions in the laboratory – can exhibit such complicated dynamics, then it seems very likely that much more complicated systems, involving hundreds of competing species as well as their predators, prey, parasites and pathogens existing in constantly fluctuating environmental conditions over many different temporal and spatial scales, will also show at least such complicated dynamics. If sensitivity to initial conditions, a signature of deterministic stochasticity, exists then it seems certain that accurate predictions are beyond the reach of such models in field conditions where parameter measurements are always estimated with a substantial degree of error.

These results, and the inherent difficulty in measuring the large number of parameter values required by dynamic demographic models, suggest that a new approach is needed to link properties of species to community structure.

Information Theory Basis of Entropy Maximization

Conceptually, the entire enterprize of modeling community structure is one of describing the relationship between macroscopic properties of communities (i. e. abundances of each species or the distribution of abundances) and microscopic properties of resource allocation when there are a large number of interacting entities. Stated in this way it is tempting to draw an analogy between, on the one hand, the patterns in ecological communities and the individuals that make up such communities and, on the other hand, the relationship between the macroscopic properties of a gas (pressure, volume, temperature) and the microscopic properties of individual molecules. Historically, the relationship between macroscopic properties of gases (for instance, temperature) and microscopic properties of molecules, for instance, the amount of kinetic and potential energy possessed by each one, was not obtained by solving a system of differential equations for the community of molecules, but rather by a statistical mechanistic approach based on macroscopic constraints [13,14]. Statistical mechanics tries to understand the macroscopic behavior of complex systems, involving large numbers of interacting microscopic components, from a probabilistic viewpoint. Although Boltzmann sought a dynamic explanation, this explanation was refuted by Jaynes [15]. More recent theoretical [16] and experimental studies [17] have confirmed Jaynes claim. Maxwell [14] saw that the second law of thermodynamics was a statistical, not a dynamic and deterministic, law; the seemingly deterministic behavior of macroscopic systems was due to the fact that such behavior was the most likely behavior by a very large margin.

Historically, Bolzmann derived his distribution from microstate counting arguments, but Gibbs [18] showed how this, and other distributions from statistical mechanics, could be derived by maximizing entropy subject to

certain physical constraints. As a way of illustration, consider N entities that can exist in one of S different states. The number of different ways (W) that these N entities can be distributed into these S states such that the macrostate distribution (i. e. the number n_i of entities expressing each state i) is the same is given by the multinomial formula (Eq. (4a)). Taking logarithms, applying Sterling's approximation ($\ln(x!) \approx x\ln(x) - x$ for large values of x), and re-expressing the numbers (n_i) as proportions ($p_i = n_i/N$), we get Shannon's information entropy (Eq. (4b)). We then see that maximizing the number of ways (W) that different microstates will produce the same macroscopic structure, consistent with certain physical constraints, is equivalent to maximizing the information entropy of $\boldsymbol{p} = \{p_i\}$ subject to these constraints. This link between counting microstates and entropy breaks down when the states vary continuously rather than being discrete with a fixed number, and a more general link is needed; this is explained below. In what follows we will attempt to develop a macroscopic description of ecological communities following this information theoretic approach to statistical mechanics. Although this approach began with Gibbs it was developed into its modern and generalized form, in the context of the Maximum Entropy Formalism, by Jaynes [19,20,21].

$$W = \frac{N!}{n_1!n_2!\ldots n_S!} \tag{4a}$$

$$\frac{\ln(W)}{N} \approx H = -\sum_{i=1}^{S} p_i \ln(p_i) \tag{4b}$$

One begins by defining the average information content of a probability distribution using a Bayesian interpretation of probabilities, recognizing that the set of relative abundances $\boldsymbol{ra} = \{ra_i\}$ has the formal properties of a probability distribution. To emphasize this equivalence, we will express relative abundances as Bayesian probabilities $\boldsymbol{ra} = \{p_i\}$ where p_i is the probability that a unit of resource will be allocated to species i given certain information about macroscopic properties of the community. The uncertainty of the information contained in p_i is defined as $I_i = -\ln(p_i)$ and the average uncertainty of information content of a probability distribution, $\sum p_i I_i$, is the information (or Shannon) entropy (Eq. (4b)). For instance, if our available information is sufficient to perfectly predict which of S different states will be expressed by an entity before it happens, then we would have zero uncertainty, we would ascribe $p_i = 1$ to the correct state and ascribe $p_j = 0\,(i \neq j)$ to all others, and the Information entropy would be zero. If, on the other hand, our available information was simply that there are S different and mutually exclusive states *and nothing else*, then we would ascribe $p_i = 1/S$ for all these states and the Information entropy would be maximal: $-\ln(S)$.

Equation (4b) is valid when the different possible states are discrete and fixed in number. A more general expression, valid for continuous states or when the number of states is not fixed, is the relative entropy expressed with reference to a maximally uninformative (or ignorance) prior probability distribution $\mathbf{q}$ (Eq. (5). The ignorance prior $\mathbf{q}$ is chosen using transformation groups such that $\mathbf{q}$ is invariant under changes in scale or location; this is equivalent to a scale-invariant measure on the probability space [21]. For a fixed number of unordered discrete states, $\mathbf{q}$ is the discrete uniform distribution and, in this special case, the relative entropy (Eq. (5) equals the entropy (Eq. (4b)). The relative entropy is the negative of the Kullback–Leibler divergence [22] in Bayesian inference, which measures the divergence between two probability distributions ($\mathbf{p}$, $\mathbf{q}$).

$$\mathrm{RH} = -\sum p_i \ln\left(\frac{p_i}{q_i}\right) \tag{5a}$$

$$\mathrm{RH} = -\int p_i \ln\left(\frac{p_i}{q_i}\right). \tag{5b}$$

The macroscopic constraints on the behavior of the system are expressed as linear functions of products of $\mathbf{p}$ and traits of each of the i states (x_{ij}) in the form of j macroscopic constraints on means (Eq. (6)); examples will be given shortly.

$$\bar{X}_j = \sum x_{ij} p_i\,. \tag{6}$$

From basic axioms of inductive reasoning [21] we want to ascribe probability values (p_i) that agree with the information encoded in the constraint equations, but that do not imply any additional information beyond this available macroscopic information. In other words, they agree with what we know (the macroscopic constraints), but that is otherwise maximally uncertain (i. e. that doesn't assume further information for which we are lacking). Since Eqs. (4) and (5) each measure the average uncertainty of information content of $\mathbf{p}$, maximizing uncertainty subject to the constraints is equivalent to maximizing the (relative) entropy subject to the constraints. This is done using the method of Lagrange multipliers. We first specify our objective function (Q, Eq. (7)) which consists of the relative entropy (RH) plus each of the j constraint equations multiplied by its Lagrange multiplier (λ_j) and solve for the values of $\mathbf{p}$ that maximize Q (Eq. (8)). The solution is a generalized exponential distribution (Eq. (9)). Analytical solutions to the values of λ are possible when there

are only a few constraint equations and numerical methods [23,24] are used for more complicated situations.

$$Q = \mathrm{RH} + \lambda_0 \left(1 - \sum p_i\right) + \lambda_1 \left(\overline{X}_1 - \sum p_i x_{i1}\right) + \cdots + \lambda_j \left(\overline{X}_j - \sum p_i x_{ij}\right) \quad (7)$$

$$\frac{\partial Q}{\partial p_i} = -\left(\ln p_i + 1\right) + \ln q_i - \lambda_0 - \sum_{j=1}^{K} x_{ij}\lambda_j = 0 \quad (8)$$

$$p_i = \frac{q_i \mathrm{e}^{-\sum_{j=1}^{K} x_{ij}\lambda_j}}{\sum_i q_i \mathrm{e}^{-\sum_{j=1}^{K} x_{ij}\lambda_j}} . \quad (9)$$

The reason why entropy maximization works (when the correct physical constraints are identified) is that if the same macroscopic behavior arises every time we apply the same physical constraints, then these constraints must be sufficient in the explanation and the myriad complexities of the microscopic dynamics must be irrelevant: the overwhelming majority of microstates that are compatible with the macroscopic constraints look the same at the macroscopic level. At the macroscopic level it does not matter which of the macroscopically redundant microstates the system finds itself.

Fixed Species Pools: Predicting Community Composition

This section deals with the case in which the total number of species that can potentially arrive at a given site is known, as are the values of certain morphological, physiological or phenological traits of each species. This list of potential species is called the species pool. In this case, the goal is to predict the relative abundance of each of these species in the pool. A more general, but less precise problem arises when one does not know the number (and therefore the specific identity) of the species making up such a pool. This more general problem will be developed later.

Imagine a collection of genotypes possessing different values of some heritable morphological, physiological or phenological trait x. If different values of this trait cause differences in per capita growth rates of these genotypes then natural selection will result in increases in relative abundance of these genotypes according to the general replicator equation [25] of natural selection (Eq. (10)).

$$\frac{\mathrm{d}p_i(t)}{\mathrm{d}t} = p_i(t)\left(r_i(t) - \bar{r}(t)\right) . \quad (10)$$

If we assume for simplicity that the relationship between per capita growth rates and the trait is linear plus a random error term (ε) with zero mean, i. e. $r_i(t) = a(t) + b(t)\, x_i + \varepsilon_i$, then the general replicator equation is:

$$\frac{\mathrm{d}p_i(t)}{\mathrm{d}t} = p_i(t)\, b(t)\left(x_i - \bar{X}(t)\right) . \quad (11)$$

Equation (11), or more complicated versions if the relationship between per capita growth rates and traits is nonlinear, defines a dynamic for community composition. The dynamics of such equations can be very complex, including the presence of chaotic attractors. However complex the dynamic might be, it is a function of the difference between the trait of each genotype, x_i, and the average trait value at a given time, $\bar{X}(t) = \sum p_i(t)\, X_i$. The trajectory of $\bar{X}$ over time is a consequence of this dynamic and is constrained by it.

The breeder's equation of quantitative genetics (Eqs. (12a), (12b)), or its multivariate equivalent [26], gives the amount by which $\bar{X}$ changes in one time unit (formally, before and after a single selection event). In this equation $h^2(t)$ is the heritability of the trait, varying between 1 (trait values are perfectly transmitted to offspring) and 0 (trait values are independent between parents and offspring) and $S(t)$ is the selection gradient at time t, i. e. the amount by which birth/death probabilities bias genotypes possessing different values of the trait. This equation therefore describes how the average trait value will behave in a given environment and is subject to natural selection. Because these average trait values change systematically over time and across different environmental conditions, they provide information about the trait distribution and are constraints (Eq. (6)). Therefore, if we know the values of these average trait values, and we know the trait values of each species, then we have the same formal mathematical problem as given in the Maximum Entropy Formalism.

$$\bar{X}(t+1) = h^2(t)\, S(t) + \bar{X}(t) \quad (12a)$$

$$\bar{X}(t) = \sum_{j=0}^{t} h^2(j)\, S(j) + \bar{X}(0) \quad (12b)$$

Natural selection, which describes the degree to which genotypes possessing particular traits in a given environmental setting have biased probabilities of reproduction and death, therefore represents a physical constraint on the dynamics of the system. We don't have to know the details of this dynamic so long as we can predict, or measure, these average trait values at the level of the entire community. Natural selection among potentially interbreeding genotypes (i. e. genotypes belonging

to the same species) leads to evolution as genotypes possessing the favored trait replace those of genotypes selected against. However, natural selection among genotypes that are reproductively isolated (i. e. genotypes belonging to different species) cannot lead to evolution since genes causing the particular phenotypic trait cannot be transmitted across the species barrier. In this case, natural selection leads to changes in the relative abundance of different species. In other words, natural selection among genotypes of different species leads to community assembly.

Before describing a field test of this statistical mechanistic approach, it is useful to look at a simple simulation. Consider a species pool consisting of six species, each species having one relevant trait whose value for species i is $(x_i = 1, 2, \ldots, 6)$. A neutral process of community assembly [27] would consist of each individual of each species having exactly the same probabilities of survival and reproduction irrespective of the trait value, so that any actual differences in population growth rates (i. e. $r_i(t)$, the per capita growth rates) are due purely to sampling fluctuations. We begin with 60 individuals per species and, in each generation, each species adds one new individual (reproduces) with a probability of 1/6 and loses one existing individual (death) with a probability of 1/6. Figure 2a shows the results of a simulation over 5000 generations and Fig. 2b shows the change in the average trait value (a community aggregated trait) over time in the community. Because probabilities of birth and death, and thus the effect of natural selection, are independent of the phenotypic trait, the average trait value fluctuates randomly. Notice that, since the trait value has no causal relationship to the probabilities of survival and reproduction and therefore does not constrain the population dynamics, the average trait value is very close to the value predicted by the maximally uninformative prior for this problem which is a discrete uniform distribution ($p_i = 1/6$). Using the average trait value at generation 5000 as a constraint and maximizing the entropy conditional on this average trait value gives, as expected, a maximum entropy distribution which is almost a discrete uniform distribution (Fig. 2c). In other words, since the phenotypic traits did not constrain the dynamics, the average trait value provided no new information that was not already present in the maximally uninformative prior.

Now, we repeat the simulation except that the probabilities of birth and death are functions of the trait x_i: larger values of x_i increase the probability of an individual of species i dying, while larger values of x_i decrease the probability of an individual of species i reproducing. Figure 3 shows the result of a typical simulation run.

Now, since there is a causal connection between the trait value and the probabilities of birth and death, these trait values constrain the population dynamics of each species and therefore the average trait value reflects the direction of natural selection as specified in the breeder's equation. Since natural selection favors individuals with low values of trait x, the average trait value (the community aggregated trait value) deceases over time. Since the trait constrains the dynamics, the average trait value provides information. Using the average trait value at the end of this simulation (generation 1000) and maximizing the entropy conditional on this average trait value correctly predicts the relative abundance of each species at this time (Fig. 3c).

It is possible to make these simulations much more complicated and realistic but the basic result is the same. If trait values are causally connected to probabilities of survival, reproduction, immigration and emigration then the population dynamics of the species will be constrained by them as specified in the general replicator equation (Eq. (11)) and the changes in these average trait values over time will be described by the breeder's equation or its multivariate version if (as there always is) there is more than one trait involved in determining fitness. In fact, since the general replicator equation is a mathematical expression of the notion of evolutionary fitness that links heritable phenotypic traits to probabilities of survival and reproduction relative to competing genotypes, such community average trait values are tracking the constraint of fitness over time and this is why they provide information on relative abundances.

An empirical example can be found in the context of secondary succession in an herbaceous plant community. Secondary succession is the well-known process of re-colonization and subsequent vegetation change following a major disturbance that kills off an existing plant community. A classic example is the pattern of vegetation change over time following abandonment of an agricultural field. If one lists the species and their relative abundances each year following abandonment in different regions and compares these lists then the patterns are unintelligible. Different species will be found in each list. On the other hand, if one lists the values of key morphological, physiological and phenological traits related to the ability of species to survive and reproduce that are found over time then very consistent patterns emerge. If one concentrates on the traits possessed by the more abundant species in each list over time, the patterns are even more consistent. Immediately following abandonment the common species will possess traits that allow them to rapidly colonize the newly opened space [28]. For instance, species

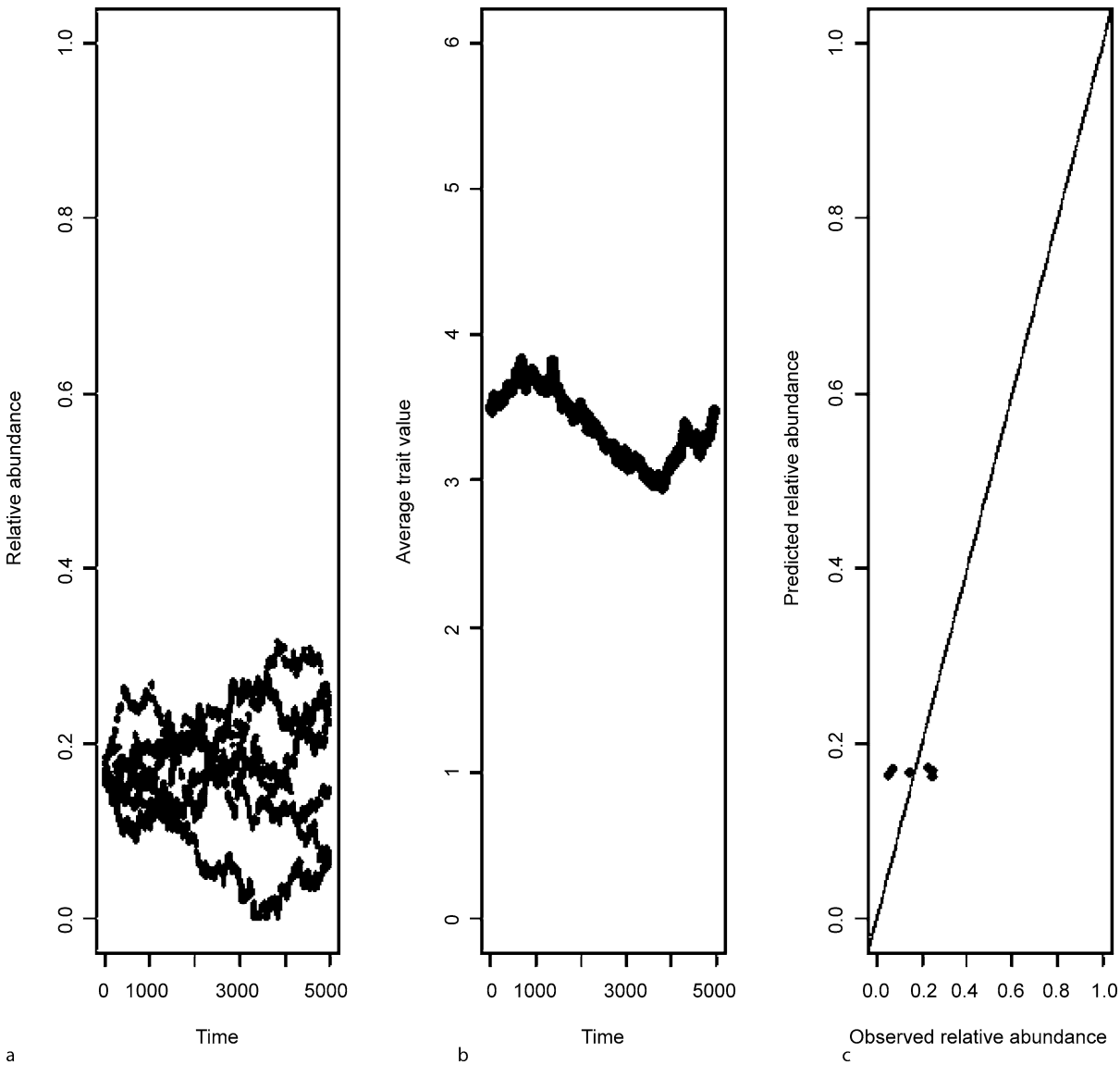

Entropy Maximization and Species Abundance, Figure 2

a An example of a neutral dynamic involving six species. Probabilities of birth and death are equal and all fluctuations in population numbers are due to sampling fluctuations. **b** Changes in the community averaged (aggregated) trait value over time that is produced by such a neutral dynamic. **c** Predicted relative abundances, given the community aggregated trait value at the end of the series, using the Maximum Entropy Formalism

having rapid growth rate and having small seeds, and especially seeds with appendages that allow for long distance dispersal, are very common. Species whose individuals produce large numbers of seeds are common since larger numbers of seeds increase the chances of some arriving at the site. Species adapted to sites having frequent large mortality events reproduce very rapidly following germination and this limits their size. Small plants that are required to produce many seeds must further decrease the size of each one, and so on. Such ruderal species have physiologies that are adapted to quickly acquire available resources and allocate then to production rather than maintenance and this leads to a correlated series of traits (specific leaf area, maximum photosynthetic rate, tissue nutrient supplies, tissue turnover rates) that maximize production of new tissues over maintenance of existing ones [29,30]. As a vegetation

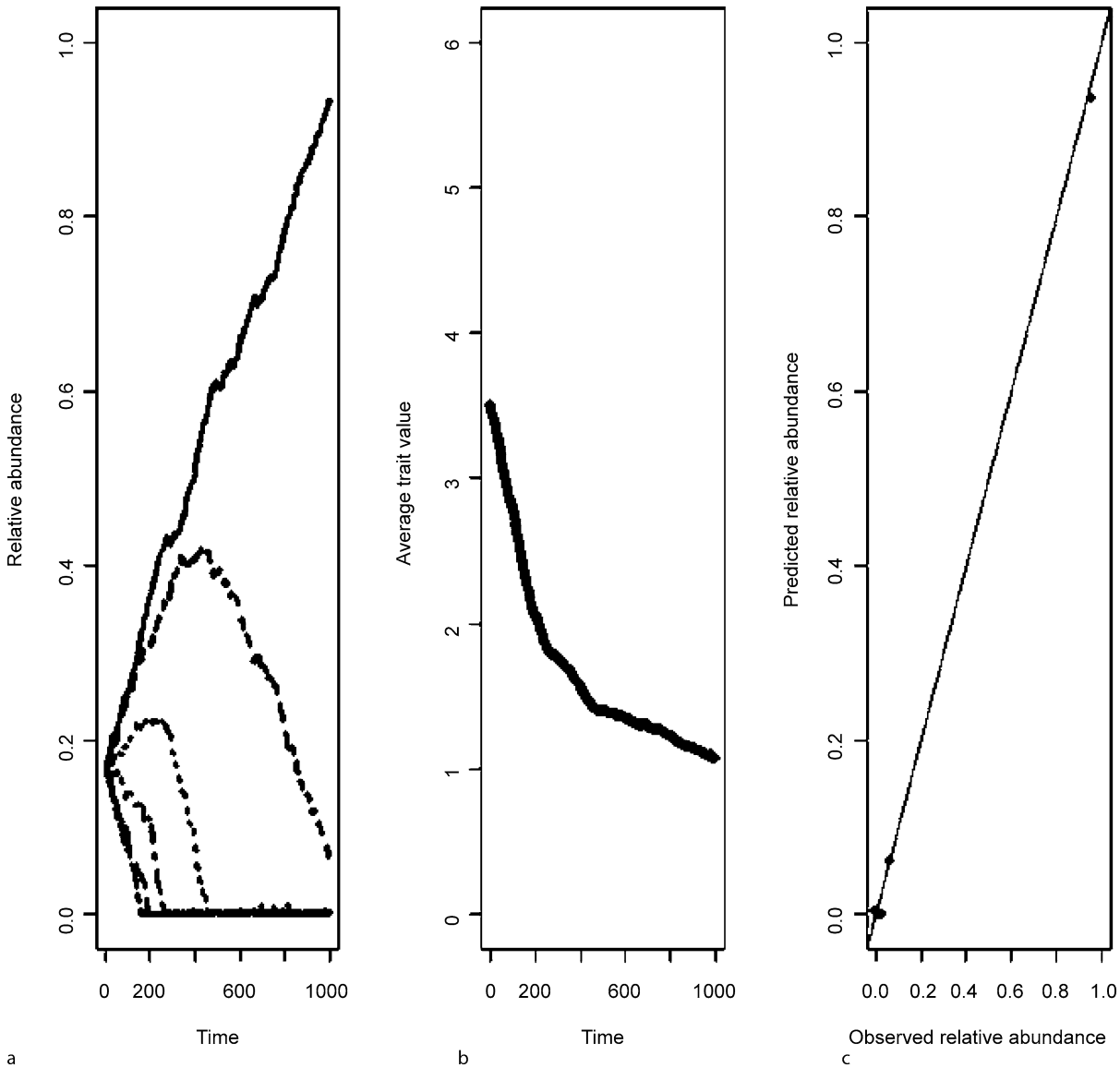

Entropy Maximization and Species Abundance, Figure 3

a An example of a dynamic involving six species in which one trait determines the differential probabilities of birth and death plus fluctuations in population numbers due to sampling fluctuations. **b** Changes in the community averaged (aggregated) trait value over time that is produced by such a dynamic. **c** Predicted relative abundances, given the community aggregated trait value at the end of the series, using the Maximum Entropy Formalism

cover develops, these ruderal species are replaced by a series of other species, larger and better able to competitively suppress those coming before, and this leads to changes in the suite of traits possessed by such species.

A recent structural equation model linking such plant traits to time following abandonment of vineyards in southern France was published by Vile et al. [31]. To test the ability of a maximum entropy approach to predicting secondary succession, Shipley et al. [32] used the same series of 10 sites. The sites varied from two to 42 years post-abandonment and estimates of the above-ground biomass of each species at each site were obtained from two $0.5 \times 0.5\,m^2$ quadrants in each. All were within a $4\,km^2$ distance and all had the same history of land use before abandonment. Using the 30 species in these that were more than transients in these 12 sites as the species

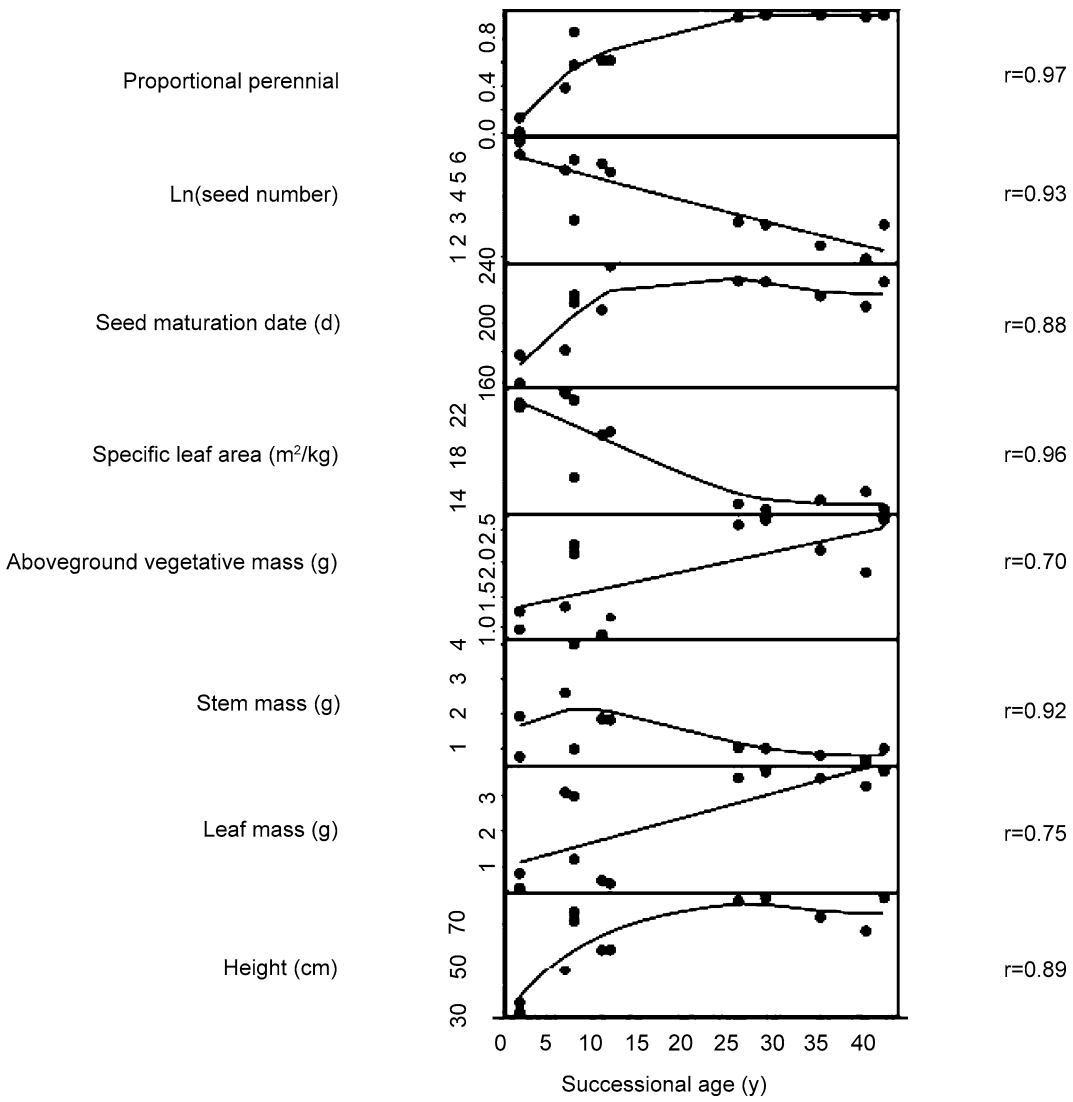

Entropy Maximization and Species Abundance, Figure 4

Each point **represents the community aggregated value of the trait measured in one field, whose successional age (years since abandonment) is given on the ordinate.** ***Lines*** **are regression (cubic spline) smoothers. Also shown is the Pearson correlation between observed and predicted values. Figure taken from [32]**

pool, and using eight traits measured on each of these 30 species, we calculated the community average (aggregated) values for each trait in each site. Figure 4 shows the results. As can be seen, there are clear systematic trends over time in these community averaged traits and these are consistent with trends reported by others.

If these traits are causally linked to fitness of these 30 species over this successional sequence then they will constrain the actual population dynamics as individual plants possessing different values of these traits immigrate into the sites, compete and reproduce. Using these eight community averaged trait values for each site, we then calculated the relative abundances assuming that these traits are sufficient to determine the population dynamics except for random sampling fluctuation. We did this by finding the distributions of relative abundance having maximum entropy after constraining then to agree with the community averaged traits. This was done using an algorithm published by Della Pietra et al. [24]. Figure 5 shows the result.

As can be seen, the predicted and observed values are close and highly significant. Clearly, there are prediction errors and this is probably be due to a combination of

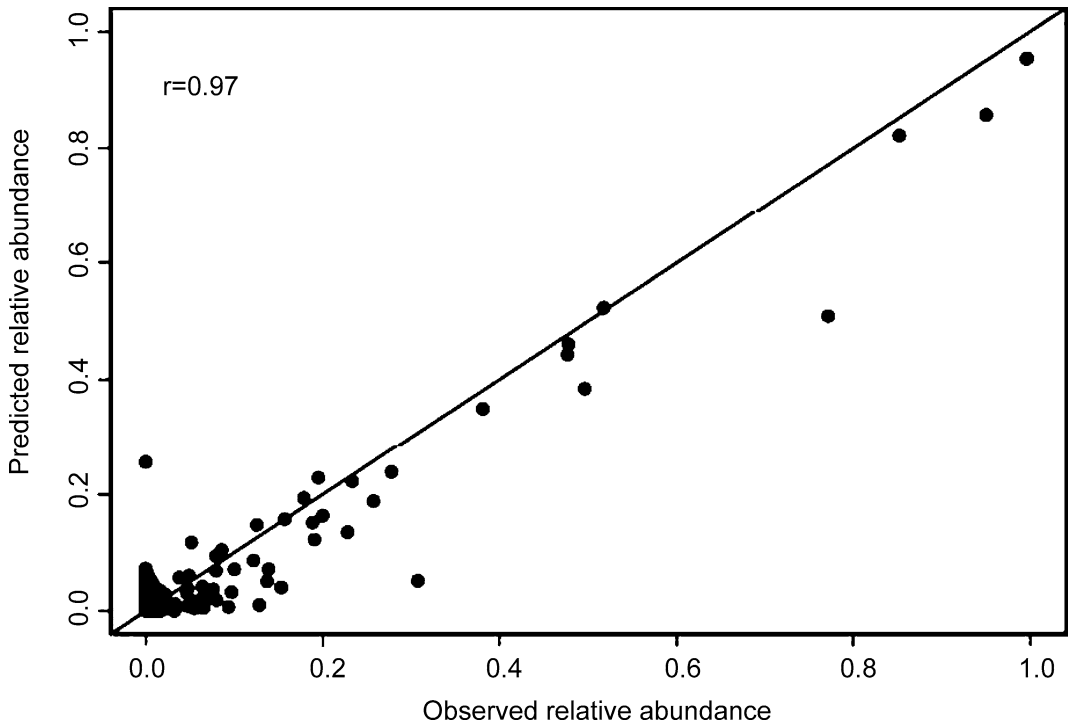

Entropy Maximization and Species Abundance, Figure 5
Observed and predicted relative abundances of each of 30 species in each of 10 different fields. Predicted relative abundances were obtained by maximizing the entropy of the distribution of relative abundances for each field conditional on the eight community aggregated traits shown in Fig. 4

missing constraints (i. e. unmeasured traits that contribute to fitness), measurement error of the trait values and random demographic fluctuations due to the fact that the numbers of individuals per site are sufficiently small that probabilities of survival and reproduction and actual frequencies of survival and reproduction differ due to sampling effects.

Species Abundance Distributions

The previous section considered the case in which we know which species occur in the species pool and in which interest resides in predicting the relative abundance of each species given certain traits. In such a situation the number of species (i. e. states in which biomass can be allocated) is fixed. However, one can also imagine situations in which the number and identity of species in the species pool is unknown. In this case we cannot predict the abundance of any particular species but we do know the distribution of abundances of the species that are present. This leads to the species abundance distribution (SAD) which describes the probability that a species will have an abundance of n individuals.

The empirical study of species abundance distributions has a long history in ecology. The first known treatment was published in Japanese by Motomira [33], as cited in Hutchinson [4]. The proposed distribution was the geometric series (Eq. (13)). Corbet [34] published a species abundance distribution based on species of butterflies and this distribution, like all others since, shows that a few species are very common and most species are very rare. C.B. Williams had collected similar data for moths and so these two authors, in collaboration with the statistician and evolutionary theorist R.A. Fisher, put forward the log-series distribution (Eq. (14)) as a statistical description of the species abundance relationship [35]. Pueyo [36], applying a Taylor series expansion to the log-series distribution, found that small deviations from a log-series gave a bounded power law (Eq. (15)). F.W. Preston, working with birds, plotted the number of species having different numbers of individuals (i. e. abundance) by binning values on a logarithmic scale and showed that, in this form, the species abundance distribution was a log-normal distribution [37,38], although any variable that must be greater than zero and is strongly right-skewed will tend to show a curve similar to a log-normal distribution after binning.

More recently, so-called neutral models of population dynamics [27,39] have been shown to recreate these patterns. Neutral models assume that all species have equal fitnesses (thus probabilities of survival, reproduction and migration) and thus all changes in population size are due to sampling fluctuations. Hubbell [27], in particular, has proposed a slightly different SAD, the zero-sum multinomial. However, differentiating between these different distributions using empirical data has proved problematic because they are very similar over most of the range of observed values [40]. Even more problematic is that authors searching for specifically biological explanations for these patterns tend to ignore that such statistical patterns are commonly found in non-biological systems. Significantly for our purposes, Preston [41] noted the similarity of species abundance distributions and gas laws as well as the distribution of wealth. Limpert et al. [42] and Nekola and Brown [43] document other cases. If the patterns transcend ecology then the explanation for the patterns must also transcend ecology.

$$P(n) = \frac{\Psi}{n} \tag{13}$$

$$p(n) = \frac{\phi x^n}{n} = \frac{\phi\left(\mathrm{e}^{-\beta}\right)^n}{n} = \frac{\phi \mathrm{e}^{-\beta n}}{n} \tag{14}$$

$$p(n) = \frac{\phi \mathrm{e}^{-\beta_1 n}}{n^{\beta_2}} \tag{15}$$

Very recently, two different groups have proposed that the common patterns of species abundance distributions can be more parsimoniously explained by the same statistical mechanistic approach as described above. The first hurtle in applying a maximum entropy approach to this area is to obtain the maximally uninformative prior. In the case of a fixed species pool the maximally uninformative prior is well-known: a discrete uniform distribution. However, when the number of species in the species pool is unknown and abundance is measured in numbers of individuals (a discrete variable) the prior is not clear. If abundance was measured as biomass, a continuous variable, then the maximally uninformative prior is known based on a consideration of invariance under transformations of scale and location [21] and is a Jeffrey's prior. Pueyo et al. [44], again based on a consideration of invariance under transformations of scale and location have shown that when abundance is measured as counts of individuals the same maximally uninformative prior is obtained: $q(n) = 1/n$.

The first constraint that one might propose is a limit on the total number of individuals (N) that can exist at a site, at least when averaged over time. This follows from the fact that the total amount of resources within a fixed area is itself fixed but also follows from basic postulates of per capita growth, namely that as populations increase the increased competition for resources leads to a decrease in per capita reproductive rates and an increase in per capita death rates, until the per capita growth rate is zero. The second constraint is that the total number of species (S) that can exist at a site is also fixed, at least when averaged over time. This follows from the theory of island biogeography [45] where, as the available resources at a site decrease, the rate of introduction of new species decreases and the rate of local extinction of species already present increases, each due to increased competition. If we accept these two macroscopic constraints, then the average number of individuals per species must also be constrained (Eq. (16)).

$$\bar{n} = \frac{N}{S} \approx k_1 \,. \tag{16}$$

To find the predicted proportion of species having abundance n in the community, and assuming that the only constraint acting on the system is to limit the average number of individuals per species (Eq. (16)), we must maximize the relative entropy conditional on this one constraint. In other words, find the values of $\mathbf{p}$ that maximize Eq. (5a), subject to the constraint given in Eq. (16) plus the constraint on normalization ($\sum p_i = 1$). The solution, given in Eq. (17), is Fisher's log-series distribution where Z is a constant.

$$p_i = \frac{q_i \mathrm{e}^{-\lambda_1 n_i}}{\sum_i q_i \mathrm{e}^{-\lambda_1 n_i}} = \frac{\mathrm{e}^{-\lambda_1 n_i}}{n_i Z} \,. \tag{17}$$

We know that, in any real ecological community, the process of community assembly involves a myriad of complicated interactions between the individuals of each species with each other and with those of other species plus the interactions of each with the physical properties of the environment. Equation (17) says that the statistical pattern of species relative abundance that was fit to empirical data by Fisher, Corbet and Williams is simply a consequence of two constraints: the total number of individuals at the site is (approximately) constant and the total number of species at the site is (approximately) constant. Equation (17) doesn't deny the existence of the complicated biological interactions that are involved in community assembly but does claim that these interactions are irrelevant with respect to the macroscopic pattern, *except* in how they constrain the total number of individuals and species. Pueyo et al. [44] call this an idiosyncratic model of community assembly: the exact process (or model) determining the abundance of species i is independent of the

process determining the abundance of every other species, therefore knowing the abundance of species i provides no information about the abundance of species j. Thus, from a macroscopic perspective, the abundances of each species at any point in time can be viewed as independent and identically distributed variables even though independent does not imply no interactions and identically distributed does not imply ecologically equivalent.

Can we imagine other macroscopic constraints that might exist in real communities? The per capita growth rate of species i over one small unit of time (Δt) is given as $r_i\,(t + \mathrm{A}t) = \ln\,(n_i\,(t+1)) - \ln\,(n_i\,(t))$. If each species in the community is at a stable equilibrium by time t then $r_i(t) \approx 0$ for all species and so the following macroscopic constraint must hold:

$$\frac{\sum \ln\,(n\,(t+1))}{S} = \frac{\sum \ln\,(n\,(t))}{S} = k_2\,. \tag{18}$$

Given the per capita growth rate describes exponential growth at time t, and since we are imagining a community that is filled with individuals, it seems quite reasonable that Eq. (18) will hold; if it did not hold over any extended time then there would be huge swings in total biomass, and this is not observed. Maximizing the relative entropy conditional on the constraints in Eqs. (16) and (18), we recover the bounded power law of Pueyo [36], as given in Eq. (15).

Finally, we could imagine a case in which the community dynamics are such that the population trajectories are in a basin of attraction but in which random external perturbations result in total resource levels fluctuations around some average values so that the per capita growth rates, $r_i(t)$, are zero over a longer time scale but can randomly deviate from zero at any instant. The fact that the trajectories are in a basin of attraction means that there will be a constraint on how far the $r_i(t)$ values can deviate from zero and this will place a constraint on the variance of the $r_i(t)$ values and this would produce a bounded log-normal distribution when maximizing the relative entropy. Since this last constraint is quite speculative at this point, we won't explore this point further.

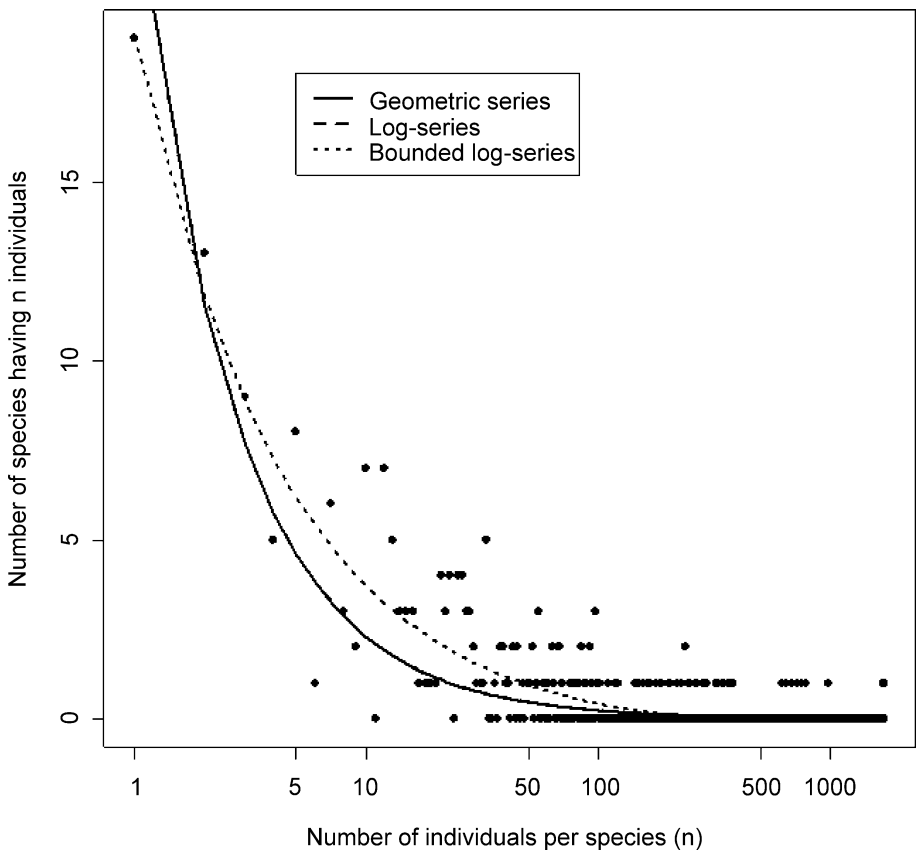

Entropy Maximization and Species Abundance, Figure 6
Distribution of abundances of tree species in a 50 hectare plot in Barro Colorado Island (Panama). Three theoretical curves, derived the maximization of relative entropy, are shown. The Geometric and Log-series curves are indistinguishable at this scale

To place these results in an empirical context, I have fit the geometric series of Motomura, the log-series of Fisher, and the bounded log-series of Pueyo to the extensive data set of tree species on Barro Colorado Island (Panama); these data come from the appendix of Volkov [46]. This data set consists of a complete census of each tree whose stem diameter at breast height is greater than 1 cm within a 50 hectare area [47]. The resulting curves are given in Eq. (19) and the result is shown in Fig. 6. The fitted curves for the geometric series and the log-series are not distinguishable and are not significantly different ($F_{1,1715} = 2.78,\ p = 0.10$) but the fitted curves for the log-series and the bounded log-series are highly significant ($F_{1,1715} = 534.43,\ p \ll 0.001$).

$$p_i = \frac{1}{933.32n} \tag{19a}$$

$$p_i = \frac{e^{-(4.263e-4)n}}{935.2n} \tag{19b}$$

$$p_i = \frac{e^{-(6.78e-3)n}}{1121n^{0.68}} \tag{19c}$$

Future Directions

Before discussing future directions, it is important to appreciate how little published work has been done in ecology using entropy maximization. Plank studied statistical mechanistic properties of the Lotka–Volterra equations [48,49,50]. A few studies have considered some theoretical aspects of entropy maximization in food webs [51,52,53,54]. Levich and co-workers [55,56,57] applied entropy maximization to algal communities but without including maximally uninformative priors. Besides [44], discussed above, I am not aware of any other work in this area. Clearly, future directions are many and varied.

At a theoretical level it will be important to more precisely link the dynamics of natural selection and the allocation of energy and resources to individuals to microstate properties of communities. In this context Dewar's fluctuation theorem [58,59,60] might prove fruitful since it links entropy maximization with those dynamic trajectories of microstates that maximize the production of entropy. If possible this could open up a thermodynamic interpretation of community assembly and possibly even of evolution by natural selection.

Those few empirical results that have been obtained so far come from plant communities. In part, this is because it is easier to census plant communities, and therefore to get reasonably accurate determinations of abundance for all species present, but the main reason is probably idiosyncratic: the application of these ideas to ecology is very recent and those very few ecologists who have explored them have been trained in plant ecology. The extension of these methods to animal communities is an obvious future direction. The use of constraints based on food webs seems particularly promising since such constraints can be linked to basic thermodynamic principles of energy transfer across trophic levels. As an example, homeotherms and poikilotherms have different basal metabolic rates and so the proportion of these two types of animals should constrain the length of food chains. Metabolic rate scales allometrically with body size [61] and so the average body size in a community should be constrained by the total amount of available energy and this, in turn, will be related to potential evapotranspiration (a function of average temperatures and precipitation). The ratio of predators to prey also scales with body size and so this might also serve as a constraint on the assembly of animal communities.

At an empirical level, it is clear that the success or failure of this approach resides in our ability to properly identify those physical constraints that actually control community dynamics and how such constraints might change as a function of environmental variables. This requires further work in order to make statistical inferences in the context of maximum entropy models and, especially, in detecting lack of fit. When the entities being allocated to different states are discrete and mutually independent and the states themselves are also discrete and fixed in number then the problem is a classical one. Given such assumptions then the statistic $-2N\sum p_i \ln p_i$ follows a chi-squared distribution with $N - k - 1$ degrees of freedom ($k =$ the number of constraints). However, such conditions rarely apply to ecological problems either because the sequential allocations are not mutually independent or because the total number (N) is unknown.

Roderick Dewar and Annabel Porté (INRA Centre de Bordeaux–Aquitaine, France) have been exploring other directions (pers. comm.). For instance, if we consider that the total amount of available resources (R) in an area is limited, that the organisms are using all of these resources, and that each species has a characteristic per capita rate of resource use, then the following constraint must hold: $\sum p_i r_i = R$. If this is combined with some well-known scaling laws in ecology relating body size with per capita resource use then, even without knowing the number of identity of species in a species pool, one can derive the species abundance distribution. Using the exponent of the Shannon entropy as an expression of the expected number of species, they explore how species richness ($\hat{S}$, the number of species per unit area) should vary as a function of R at different scales. At local spatial scales it is well known that $\hat{S}$ has a so-called humped-back distribution [62] with

respect to productivity, and therefore R. At such local spatial scales, it is reasonable to consider R fixed and S (the number of species in the species pool) to be finite. At spatial scales large enough to include substantial variation in mean annual temperature and precipitation $\hat{S}$ scales allometrically with potential evapotranspiration [63] and it is perhaps reasonable to consider R free. Can these contrasting patterns of biodiversity at different spatial scales be accounted for by a statistical mechanistic approach in which constraints on R and S change?

Bibliography

1. Darwin C (1859) On the origin of species. John Murray, London
2. Wallace AR (1858) On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. III. On the tendency of varieties to depart indefinitely from the original type. J Proc Linnean Soc Lond 3:53–62
3. Lotka AJ (1925) Elements of physical biology. Williams and Williams, Baltimore
4. Hutchinson GE (1978) An introduction to population ecology. Yale University Press, New Haven
5. Volterra V (1926) Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. Mem R Acad dei Lincei (Ser. 6) 2:31–113
6. Roughgarden J (1979) Theory of population genetics and evolutionary ecology: An introduction. Macmillan Publishing, New York
7. Tilman D (1982) Resource competition and community structure. Princeton University Press, Princeton
8. Vandermeer JH (1969) The competitive structure of communities: An experimental approach with protozoa. Ecology 50: 362–371
9. Gilpin ME, Carpenter MP, Pomerantz MJ (1986) The assembly of a laboratory community: Multispecies competition in drosophila. In: Diamond JM, Case T (eds) Community ecology. Harper and Row, Cambridge
10. May RA (1974) Biological populations with nonoverlapping generations: Stable points, stable cycles, and chaos. Science 186:645–647
11. Hilborn RC (2004) Seagulls, butterflies, and grasshoppers: A brief history of the butterfly effect in nonlinear dynamics. Am J Phys 72:425–427
12. Costantino RF et al (2005) Nonlinear stochastic population dynamics: The flour beetle tribolium as an effective tool of discovery. Advances in ecological research: Population dynamics and laboratory ecology, vol 37. Academic Press, London, pp 101–141
13. Boltzmann L (1898) Lectures on gas theory. Dover Publications, New York
14. Maxwell JC (1871) Theory of heat. Dover Publications, New York
15. Jaynes ET (1971) Violations of boltzmann's h theorem in real gases. Phys Rev A4:747–751
16. Evans DJ, Searles DJ (2002) The fluctuation theorem. Adv Phys 51:1529–1585
17. Wang GM, Sevick EM, Mittag E, Searles DJ, Evans DJ (2002) Experimental demonstration of violations of the second law of thermodynamics for small systems and short time scales. Phys Rev Lett 89:50601
18. Gibbs JW (1902) Elementary principles in statistical mechanics. Yale University Press, New Haven
19. Jaynes ET (1957) Information theory and statistical mechanics i. Phys Rev 106:620–630
20. Jaynes ET (1957) Information theory and statistical mechanics ii. Phys Rev 108:171–190
21. Jaynes ET (2003) Probability theory. The logic of science. Cambridge University Press, Cambridge
22. Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Statist 22:79–86
23. Agmon N, Alhassid Y, Levine RD (1979) An algorithm for finding the distribution of maximal entropy. J Comput Phys 30:250–258
24. Della Pietra S, Della Pietra V, Lafferty J (1997) Inducing features of random fields. IEEE Trans Pattern Anal Mach Intell 19: 1–13
25. Schuster P, Sigmund K (1983) Replicator dynamics. J Theor Biol 100:533–538
26. Roff DA (1997) Evolutionary quantitative genetics. Chapman and Hall, NY
27. Hubbell SP (2001) The unified neutral theory of biodiversity and biogeography. Princeton University Press, Princeton
28. Grime JP (2007) Comparative plant ecology. Castlepoint Press, Colvend
29. Shipley B, Lechoweicz MJ, Wright I, Reich PB (2006) Fundamental trade-offs generating the worldwide leaf economics spectrum. Ecology 87:535–541
30. Wright IJ et al (2004) The worldwide leaf economics spectrum. Nature 428:821–827
31. Vile D, Shipley B, Garnier E (2006) A structural equation model to integrate changes in functional strategies during old-field succession. Ecology 87:504–517
32. Shipley B, Vile D, Garnier E (2006) From plant traits to plant communities: A statistical mechanistic approach to biodiversity. Science 314:812–814
33. Motomura I (1932) A statistical treatment of associations [in japanese]. Jpn J Zool 44:379–383
34. Corbet AS (1941) The distribution of butterflies in the malay peninsula (lepid.). Proc Roy Entomol Soc Ser A 16:101–116
35. Fisher RA, Corbet AS, Williams CB (1943) The relation between the number of species and the number of individuals in a random sample from an animal population. J Animal Ecol 12: 42–58
36. Pueyo S (2006) Diversity: Between neutrality and structure. Oikos 112:392–405
37. Preston FW (1948) The commonness, and rarity, of species. Ecology 29:254–283
38. Preston FW (1962) The canonical distribution of commonness and rarity. Ecology 43:185–215
39. Bell G (2000) The distribution of abundance in neutral communities. Am Nat 155:606–617
40. McGill BJ, Maurer BA, Weiser MD (2006) Empirical evaluation of neutral theory. Ecology 87:1411–1423
41. Preston FW (1950) Gas laws and wealth laws. Sci Mon 71:309–311
42. Limpert E, Stahel WA, Abbt M (2001) Log-normal distributions across the sciences: Keys and clues. Bioscience 51:341–352

43. Nekola JC, Brown JH (2007) The wealth of species: Ecological communities, complex systems and the legacy of frank preston. Ecol Lett 10:188–196
44. Pueyo S, He F, Zillio T (2007) The maximum entropy formalism and the idiosyncratic theory of biodiversity. Ecol Lett 10:1017–1028
45. MacArthur RH (1972) Geographical ecology: Patterns in the distribution of species. Harper and Row, NY
46. Volkov I, Banavar JR, He FL, Hubbell SP, Maritan A (2005) Density dependence explains tree species abundance and diversity in tropical forests. Nature 438:658–661
47. Hubbell SP, Foster RB (1990) Structure, dynamics and equilibrium status of old-growth forest on barro colorado island. In: Gentry A (ed) Four neotropical forests. Yale University Press, New Haven
48. Plank M (1995) Hamiltonian structures for the *n*-dimensional Lotka–Volterra equations. J Math Phys 36:3520–3534
49. Plank M (1996) Bi-hamiltonian systems and Lotka–Volterra equations: A three-dimensional classification. Nonlinearity 9:887–896
50. Plank M (1999) On the dynamics of Lotka–Volterra equations having an invariant hyperplane. Siam J Appl Math 59:1540–1551
51. Kerner EH (1964) Gibbs ensemble: Biological ensemble. Gordon and Brench, New York
52. Kerner EH (1979) The gibbs grand ensemble and the eco-genetic gap. In: Levine RD, Tribus M (eds) The maximum entropy formalism. Massachusetts Institute of Technology, Massachusetts, pp 468–476
53. Wagensberg J, Valls J (1987) The [extended] maximum entropy formalism and the statistical structure of ecosystems. Bull Math Biol 48:531–538
54. Lurie D, Wasenburg J (1985) An extremal principle for biomass diversity in ecology. In: Lamprecht I, Zotin AI (eds) Thermodynamics and regulation of biological processes. De Gruyter, Berlin, pp 577–271
55. Levich AP (1988) What are the possible theoretical principles in the ecology of communities? In: Kull K, Tiivel T (eds) Lectures in theoretical biology. Valgus, Tallinn, pp 121–127
56. Alexeyev VL, Levich AP (1997) A search for maximum species abundances in ecological communities under conditional diversity optimization. Bull Math Biol 59:649–677
57. Levich AP (2000) Variational modelling theorems and algocoenoses functioning principles. Ecol Model 131:207–227
58. Dewar R (2003) Information theory explanation of the fluctuation theorem, maximum entropy production and self-organized criticality in non-equilibrium stationary states. J Phys Math Gen 36:631–641
59. Dewar R (2004) Maximum entropy production and non-equilibrium statistical mechanics. In: Kliedon A, Lorenz RD (eds) Non-equilibrium thermodynamics and the production of entropy: Life, earth and beyond. Springer, Berlin, pp 41–56
60. Dewar R (2005) Maximum entropy production and the fluctuation theorem. J Phys Math Gen 38:L371–L381
61. Peters RH (1983) The ecological implications of body size. Cambridge University Press, Cambridge
62. Al-Mufti MM, Sydes CL, Furness SB, Grime JP, Band SR (1977) A quantitative analysis of shoot phenology and dominance in herbaceous vegetation. J Ecol 65:759–791
63. Currie DJ, Paquin V (1987) Large-scale biogeographic patterns of species richness of trees. Nature 329:326–327

Ergodicity and Mixing Properties

Anthony Quas
Department of Mathematics and Statistics,
University of Victoria, Victoria, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
Basics and Examples
Ergodicity
Ergodic Decomposition
Mixing
Hyperbolicity and Decay of Correlations
Future Directions
Bibliography

Glossary

Bernoulli shift Mathematical abstraction of the scenario in statistics or probability in which one performs repeated independent identical experiments.

Markov chain A probability model describing a sequence of observations made at regularly spaced time intervals such that at each time, the probability distribution of the subsequent observation depends only on the current observation and not on prior observations.

Measure-preserving transformation A map from a measure space to itself such that for each measurable subset of the space, it has the same measure as its inverse image under the map.

Measure-theoretic entropy A non-negative (possibly infinite) real number describing the complexity of a measure-preserving transformation.

Product transformation Given a pair of measure-preserving transformations: T of X and S of Y, the product transformation is the map of $X \times Y$ given by $(T \times S)(x, y) = (T(x), S(y))$.

Definition of the Subject

Many physical phenomena in equilibrium can be modeled as measure-preserving transformations. Ergodic theory is the abstract study of these transformations, dealing in particular with their long term average behavior.

One of the basic steps in analyzing a measure-preserving transformation is to break it down into its simplest possible components. These simplest components are its ergodic components, and on each of these components, the system enjoys the ergodic property: the long-term time

average of any measurement as the system evolves is equal to the average over the component. Ergodic decomposition gives a precise description of the manner in which a system can be split into ergodic components.

A related (stronger) property of a measure-preserving transformation is mixing. Here one is investigating the correlation between the state of the system at different times. The system is mixing if the states are asymptotically independent: as the times between the measurements increase to infinity, the observed values of the measurements at those times become independent.

Introduction

The term ergodic was introduced by Boltzmann [8,9] in his work on statistical mechanics, where he was studying Hamiltonian systems with large numbers of particles. The system is described at any time by a point of *phase space*, a subset of $\mathbb{R}^{6N}$ where N is the number of particles. The configuration describes the 3-dimensional position and velocity of each of the N particles. It has long been known that the Hamiltonian (i. e. the overall energy of the system) is invariant over time in these systems. Thus, given a starting configuration, all future configurations as the system evolves lie on the same *energy surface* as the initial one.

Boltzmann's *ergodic hypothesis* was that the trajectory of the configuration in phase space would fill out the entire energy surface. The term ergodic is thus an amalgamation of the Greek words for work and path. This hypothesis then allowed Boltzmann to conclude that the long-term average of a quantity as the system evolves would be equal to its average value over the phase space.

Subsequently, it was realized that this hypothesis is rarely satisfied. The ergodic hypothesis was replaced in 1911 by the *quasi-ergodic hypothesis* of the Ehrenfests [17] which stated instead that each trajectory is dense in the energy surface, rather than filling out the entire energy surface. The modern notion of ergodicity (to be defined below) is due to Birkhoff and Smith [7]. Koopman [44] suggested studying a measure-preserving transformation by means of the associated isometry on Hilbert space, $U_T\colon L^2(X) \to L^2(X)$ defined by $U_T(f) = f \circ T$. This point of view was used by von Neumann [91] in his proof of the mean ergodic theorem. This was followed closely by Birkhoff [6] proving the pointwise ergodic theorem. An ergodic measure-preserving transformation enjoys the property that Boltzmann first intended to deduce from his hypothesis: that long-term averages of an observable quantity coincide with the integral of that quantity over the phase space.

These theorems allow one to deduce a form of independence on the average: given two sets of configurations A and B, one can consider the volume of the phase space consisting of points that are in A at time 0 and in B at time t. In an ergodic measure-preserving transformation, if one computes the average of the volumes of these regions over time, the ergodic theorems mentioned above allow one to deduce that the limit is simply the product of the volume of A and the volume of B. This is the weakest mixing-type property. In this article, we will outline a rather full range of mixing properties with ergodicity at the weakest end and the Bernoulli property at the strongest end.

We will set out in some detail the various mixing properties, basing our study on a number of concrete examples sitting at various points of this hierarchy. Many of the mixing properties may be characterized in terms of the Koopman operators operators mentioned above (i. e. they are *spectral properties*), but we will see that the strongest mixing properties are not spectral in nature.

We shall also see that there are connections between the range of mixing properties that we discuss and measure-theoretic entropy. In measure-preserving transformations that arise in practice, there is a correlation between strong mixing properties and positive entropy, although many of these properties are logically independent.

One important issue for which many questions remain open is that of higher-order mixing. Here, the issue is if instead of asking that the observations at two times separated by a large time T be approximately independent, one asks whether if one makes observations at more times, each pair suitably separated, the results can be expected to be approximately independent. This issue has an analogue in probability theory, where it is well-known that it is possible to have a collection of random variables that are pairwise independent, but not mutually independent.

Basics and Examples

In this article, except where otherwise stated, the measure-preserving transformations that we consider will be defined on probability spaces.

More specifically, given a measurable space $(X, \mathcal{B})$ and a probability measure μ defined on $\mathcal{B}$, a *measure-preserving transformation* of $(X, \mathcal{B}, \mu)$ is a $\mathcal{B}$-measurable map $T\colon X \to X$ such that $\mu(T^{-1}B) = \mu(B)$ for all $B \in \mathcal{B}$.

While this definition makes sense for arbitrary measures, not simply probability measures, most of the results and definitions below only make sense in the probability measure case. Sometimes it will be helpful to make the assumption that the underlying probability space is

a Lebesgue space (that is, the space together with its completed σ-algebra agrees up to a measure-preserving bijection with the unit interval with Lebesgue measure and the usual σ-algebra of Lebesgue measurable sets). Although this sounds like a strong restriction, in practice it is barely a restriction at all, as almost all of the spaces that appear in the theory (and all of those that appear in this article) turn out to be Lebesgue spaces. For a detailed treatment of the theory of Lebesgue spaces, the reader is referred to Rudolph's book [76]. The reader is referred also to the chapter on ► Measure Preserving Systems.

While many of the definitions that we shall present are valid for both invertible and non-invertible measure-preserving transformations, the strongest mixing conditions are most useful in the case of invertible transformations.

It will be helpful to present a selection of simple examples, relative to which we will be able to explore ergodicity and the various notions of mixing. These examples and the lemmas necessary to show that they are measure-preserving transformations as claimed may be found in the books of Petersen [64], Rudolph [76] and Walters [92]. More details on these examples can also be found in the chapter on ► Ergodic Theory: Basic Examples and Constructions.

Example 1 (Rotation on the circle) Let $\alpha \in \mathbb{R}$. Let $R_\alpha\colon [0,1) \to [0,1)$ be defined by $R_\alpha(x) = x + \alpha \bmod 1$. It is straightforward to verify that R_α preserves the restriction of Lebesgue measure λ to $[0,1)$ (it is sufficient to check that $\lambda(R_\alpha^{-1}(J)) = \lambda(J)$ for an interval J).

Example 2 (Doubling Map) Let $M_2\colon [0,1) \to [0,1)$ be defined by $M_2(x) = 2x \bmod 1$. Again, Lebesgue measure is invariant under M_2 (to see this, one observes that for an interval J, $M_2^{-1}(J)$ consists of two intervals, each of half the length of J). This may be generalized in the obvious way to a map M_k for any integer $k \geq 2$.

Example 3 (Interval Exchange Transformation) The class of interval exchange transformations was introduced by Sinai [85]. An interval exchange transformation is the map obtained by cutting the interval into a finite number of pieces and permuting them in such a way that the resulting map is invertible, and restricted to each interval is an order-preserving isometry.

More formally, one takes a sequence of positive lengths $\ell_1, \ell_2, \ldots, \ell_k$ summing to 1 and a permutation π of $\{1, \ldots, k\}$ and defines $a_i = \sum_{j<i} \ell_j$ and $b_i = \sum_{\pi(j)<\pi(i)} \ell_j$ (again with $b_0 = 0$). The interval exchange transformation defined by $(\ell_1, \ldots, \ell_k)$ and π is the map $T\colon [0,1) \to [0,1)$ defined by $T|_{[a_i, a_{i+1})}(x) = x + (b_i - a_i)$. It is straightforward to check that any such interval exchange transformation preserves Lebesgue measure on the unit interval.

Example 4 (Bernoulli Shift) Let A be a finite set and fix a vector $(p_i)_{i\in A}$ of positive numbers that sum to 1. Let $A^{\mathbb{N}}$ denote the set of sequences of the form $x_0x_1x_2\ldots$, where $x_n \in A$ for each $n \in \mathbb{N}$ and let $A^{\mathbb{Z}}$ denote the set of bi-infinite sequences of the form $\ldots x_{-2}x_{-1} \cdot x_0x_1x_2\ldots$ (the $\cdot$ is a placeholder that allows us to distinguish (for example) between the sequences $\ldots 01010 \cdot 10101 \ldots$ and $\ldots 10101 \cdot 01010 \ldots$).

We define a map (the *shift map*) S on $A^{\mathbb{N}}$ by $(S(x))_n = x_{n+1}$ and define S on $A^{\mathbb{Z}}$ by the same formula. Note that S is invertible as a transformation on $A^{\mathbb{Z}}$ but non-invertible as a transformation on $A^{\mathbb{N}}$.

We need to equip $A^{\mathbb{N}}$ and $A^{\mathbb{Z}}$ with measures. This is done by defining the measure of a preferred class of sets, checking certain consistency conditions and appealing to the Kolmogorov extension theorem. Here the preferred sets are the *cylinder sets*. Given $m \leq n$ in the invertible case and a sequence $a_m \ldots a_n$, we let $[a_m \ldots a_n]_m^n$ denote $\{x \in A^{\mathbb{Z}}\colon x_m = a_m, \ldots, x_n = a_n\}$ and define $\mu([a_m \ldots a_n]_m^n) = p_{a_m} p_{a_{m+1}} \ldots p_{a_n}$. This is then shown to uniquely define a measure μ on the σ-algebra of $A^{\mathbb{Z}}$ generated by the cylinder sets. It is immediate to see that for any cylinder set C, $\mu(S^{-1}C) = \mu(C)$, and it follows that S is a measure-preserving transformation of $(A^{\mathbb{Z}}, \mathcal{B}, \mu)$. The construction is exactly analogous in the non-invertible case. See the chapter on ► Measure Preserving Systems or the books of Walters [92] or Rudolph [76] for more details of defining measures in these systems.

The class of Bernoulli shifts will play a distinguished role in what follows.

Example 5 (Markov Shift) The spaces $A^{\mathbb{N}}$ and $A^{\mathbb{Z}}$ are exactly as above, as is the shift map. All that changes is the measure.

To define a Markov shift, we need a stochastic matrix P (i. e. a matrix with non-negative entries whose rows sum to 1) with rows and columns indexed by A and a left eigenvector π for P with eigenvalue 1 with the property that the entries of π are non-negative and sum to 1. The existence of such an eigenvector is a consequence of the Perron–Frobenius theory of positive matrices. Provided that the matrix P is irreducible (for each a and a' in A, there is an $n > 0$ such that $P^n_{a,a'} > 0$), the eigenvector π is unique.

Given the pair (P, π), one defines the measure of a cylinder set by $\mu([a_m \ldots a_n]_m^n) = \pi_{a_m} P_{a_m a_{m+1}} \ldots P_{a_{n-1} a_n}$ and extends μ as before to a probability measure on $A^{\mathbb{N}}$ or $A^{\mathbb{Z}}$.

Example 6 (Hard Sphere Gases and Billiards) We wish to model the behavior of a gas in a bounded region. We make the assumption that the gas consists of a large number N

of identical balls which move at constant velocity until two balls collide, whereupon they elastically swap momentum along the direction of contact. The phase space for this system is a region of $\mathbb{R}^{6N}$ (with N 3-dimensional position vectors and N 3-dimensional velocity vectors). More abstractly, the system is equivalent to the motion of a single point particle in a region of $\mathbb{R}^M \times \mathbb{R}^M$ (with the first M-vector representing position and the second representing velocity). The system is constrained in that its position is required to lie in a bounded region S of $\mathbb{R}^M$ with a piecewise smooth boundary. The system evolves by moving the position at a constant rate in the direction of the velocity vector until the point reaches ∂S, at which time the component of the velocity parallel to the normal to ∂S is reversed. This then defines a flow (i. e. a family of maps $(T_t)_{t\in\mathbb{R}}$ satisfying $T_{t+s} = T_t \circ T_s$) on the phase space. Since the magnitude of the velocity is conserved, it is convenient to restrict to flows with speed 1. This system is clearly the closest of the examples that we consider to the situation envisaged by Boltzmann. Perhaps not surprisingly, proofs of even the most basic properties for this system are much harder than those for the other examples that we consider.

We will need to make use of the concept of measure-theoretic isomorphism. Two measure-preserving transformations T of $(X, \mathcal{B}, \mu)$ and S of $(Y, \mathcal{F}, \nu)$ are *measure-theoretically isomorphic* (or just *isomorphic*) if there exist measurable maps $g\colon X \to Y$ and $h\colon Y \to X$ such that

1. $g \circ h$ and $h \circ g$ agree with the respective identity maps almost everywhere;
2. $\mu(g^{-1}F) = \nu(F)$ and $\nu(h^{-1}B) = \mu(B)$ for all $F \in \mathcal{F}$ and $B \in \mathcal{B}$; and
3. $S \circ g(x) = g \circ T(x)$ for μ-almost every x (or equivalently $T \circ h(y) = h \circ S(y)$ for ν-almost every y).

Measure-theoretic isomorphism is the basic notion of 'sameness' in ergodic theory. It is in some sense quite weak, so that systems may be isomorphic that feel very different (for example, as we discuss later, the time one map of a geodesic flow is isomorphic to a Bernoulli shift). For comparison, the notion of sameness in topological dynamical systems (topological conjugacy) is far stronger.

As an example of measure-theoretic isomorphism, it may be seen that the doubling map is isomorphic to the one-sided Bernoulli shift on $\{0, 1\}$ with $p_0 = p_1 = 1/2$ (the map g takes an $x \in [0, 1)$ to the sequence of 0's and 1's in its binary expansion (choosing the sequence ending with 0's, for example, if x is of the form $p/2^n$) and the inverse map h takes a sequence of 0's and 1's to the point in $[0, 1)$ with that binary expansion.)

Given a measure-preserving transformation T of a probability space $(X, \mathcal{B}, \mu)$, T is associated to an isometry of $L^2(X, \mathcal{B}, \mu)$ by $U_T(f) = f \circ T$. This operator is known as the *Koopman Operator*. In the case where T is invertible, the operator U_T is unitary. Two measure-preserving transformations T and S of $(X, \mathcal{B}, \mu)$ and $(Y, \mathcal{F}, \nu)$ are *spectrally isomorphic* if there is a Hilbert space isomorphism Θ from $L^2(X, \mathcal{B}, \mu)$ to $L^2(Y, \mathcal{F}, \nu)$ such that $\Theta \circ U_T = U_S \circ \Theta$. As we shall see below, spectral isomorphism is a strictly weaker property than measure-theoretic isomorphism.

Since in ergodic theory, measure-theoretic isomorphism is the basic notion of sameness, all properties that are used to describe measure-preserving systems are required to be invariant under measure-theoretic isomorphism (i. e. if two measure-preserving transformations are measure-theoretically isomorphic, the first has a given property if and only if the second does). On the other hand, we shall see that some mixing-type properties are invariant under spectral isomorphism, while others are not. If a property is invariant under spectral isomorphism, we say that it is a *spectral property*.

There are a number of mixing type properties that occur in the probability literature (α-mixing, β-mixing, ϕ-mixing, ψ-mixing etc.) (see Bradley's survey [12] for a description of these conditions). Many of these are stronger than the Bernoulli property, and are therefore not preserved under measure-theoretic isomorphism. For this reason, these properties are not widely used in ergodic theory, although β-mixing turns out to be equivalent to the so-called *weak Bernoulli* property (which turns out to be stronger than the Bernoulli property that we discuss in this article – see Smorodinsky's paper [87]) and α-mixing is equivalent to strong-mixing.

A basic construction (see the article on ► Ergodic Theory: Basic Examples and Constructions) that we shall require in what follows is the product of a pair of measure-preserving transformations: given transformations T of $(X, \mathcal{B}, \mu)$ and S of $(Y, \mathcal{F}, \nu)$, we define the *product transformation* $T \times S\colon (X \times Y, \mathcal{B} \otimes \mathcal{F}, \mu \times \nu)$ by $(T \times S)(x, y) = (Tx, Sy)$.

One issue that we face on occasion is that it is sometimes convenient to deal with invertible measure-preserving transformations. It turns out that given a non-invertible measure-preserving transformation, there is a natural way to uniquely associate an invertible measure-preserving transformation transformation sharing almost all of the ergodic properties of the original transformation. Specifically, given a non-invertible measure-preserving transformation T of $(X, \mathcal{B}, \mu)$, one lets $\overline{X} = \{(x_0, x_1, \dots)\colon x_n \in X \text{ and } T(x_n) = x_{n-1} \text{ for all } n\}$, $\overline{\mathcal{B}}$

be the σ-algebra generated by sets of the form $\bar{A}_n = \{\bar{x} \in \overline{X}: x_n \in A\}$, $\bar{\mu}(\bar{A}_n) = \mu(A)$ and $\overline{T}(x_0, x_1, \dots) = (T(x_0), x_0, x_1, \dots)$. The transformation $\overline{T}$ of $(\overline{X}, \overline{\mathcal{B}}, \overline{\mu})$ is called the *natural extension* of the transformation T of $(X, \mathcal{B}, \mu)$ (see the chapter on ► Ergodic Theory: Basic Examples and Constructions for more details). In situations where one wants to use invertibility, it is often possible to pass to the natural extension, work there and then derive conclusions about the original non-invertible transformation.

Ergodicity

Given a measure-preserving transformation $T: X \to X$, if $T^{-1}A = A$, then $T^{-1}A^c = A^c$ also. This allows us to decompose the transformation X into two pieces A and A^c and study the transformation T separately on each. In fact the same situation holds if $T^{-1}A$ and A agree up to a set of measure 0. For this reason, we call a set A *invariant* if $\mu(T^{-1}A \Delta A) = 0$.

Returning to Boltzmann's ergodic hypothesis, existence of an invariant set of measure between 0 and 1 would be a bad situation as his essential idea was that the orbit of a single point would 'see' all of X, whereas if X were decomposed in this way, the most that a point in A could see would be all of A, and similarly the most that a point in A^c could see would be all of A^c.

A measure-preserving transformation will be called *ergodic* if it has no non-trivial decomposition of this form. More formally, let T be a measure-preserving transformation of a probability space $(X, \mathcal{B}, \mu)$. The transformation T is said to be ergodic if for all invariant sets, either the set or its complement has measure 0.

Unlike the remaining concepts that we discuss in this article, this definition of ergodicity applies also to infinite measure-preserving transformations and even to certain non-measure-preserving transformations. See Aaronson's book [1] for more information.

The following lemma is often useful:

Lemma 1 *Let $(X, \mathcal{B}, \mu)$ be a probability space and let $T: X \to X$ be a measure-preserving transformation. Then T is ergodic if and only if the only measurable functions f satisfying $f \circ T = f$ (up to sets of measure 0) are constant almost everywhere.*

For the straightforward proof, we notice that if the condition in the lemma holds and A is an invariant set, then $\mathbf{1}_A \circ T = \mathbf{1}_A$ almost everywhere, so that $\mathbf{1}_A$ is an a. e. constant function and so A or A^c is of measure 0. Conversely, if f is an invariant function, we see that for each α, $\{x: f(x) < \alpha\}$ is an invariant set and hence of measure 0 or 1. It follows that f is constant almost everywhere. We remark for future use that it is sufficient to check that the bounded measurable invariant functions are constant.

The following corollary of the lemma shows that ergodicity is a spectral property.

Corollary 2 *Let T be a measure-preserving transformation of the probability space $(X, \mathcal{B}, \mu)$. Then T is ergodic if and only if 1 is a simple eigenvalue of U_T.*

The ergodic theorems mentioned earlier due to von Neumann and Birkhoff are the following (see also the chapter on ► Ergodic Theorems).

Theorem 3 (von Neumann Mean Ergodic Theorem [91]) *Let T be a measure-preserving transformation of the probability space $(X, \mathcal{B}, \mu)$. For $f \in L^2(X, \mathcal{B}, \mu)$, let $A_N f = 1/N(f + f \circ T + \cdots + f \circ T^{N-1})$. Then for all $f \in L^2(X, \mathcal{B}, \mu)$, $A_N f$ converges in L^2 to an invariant function f^*.*

Theorem 4 (Birkhoff Pointwise Ergodic Theorem [6]) *Let T be a measure-preserving transformation of the probability space $(X, \mathcal{B}, \mu)$. Let $f \in L^1(X, \mathcal{B}, \mu)$. Let $A_N f$ be as above. Then for μ-almost every $x \in X$, $(A_N f(x))$ is a convergent sequence.*

Of these two theorems, the pointwise ergodic theorem is the deeper result, and it is straightforward to deduce the mean ergodic theorem from the pointwise ergodic theorem. The mean ergodic theorem was reproved very concisely by Riesz [71] and it is this proof that is widely known now. Riesz's proof is reproduced in Parry's book [63]. There have been many different proofs given of the pointwise ergodic theorem. Notable amongst these are the argument due to Garsia [23] and a proof due to Katznelson and Weiss [40] based on work of Kamae [35], which appears in a simplified form in work of Keane and Petersen [42].

If the measure-preserving transformation T is ergodic, then by virtue of Lemma 1, the limit functions appearing in the ergodic theorems are constant. One sees that the constant is simply the integral of f with respect to μ, so that in this situation $A_N f(x)$ converges to $\int f \, d\mu$ in norm and pointwise almost everywhere, thereby providing a justification of Boltzmann's original claim: for ergodic measure-preserving transformations, *time averages agree with spatial averages*. In the case where T is not ergodic, it is also possible to identify the limit in the ergodic theorems: we have $f^* = \mathbb{E}(f|\mathcal{I})$, where $\mathcal{I}$ is the σ-algebra of T-invariant sets.

Note that the set on which the almost everywhere convergence in Birkhoff's theorem takes place depends on the L^1 function f that one is considering. Straightforward considerations show that there is no single full

measure set that works simultaneously for all L^1 functions. In the case where X is a compact metric space, it is well known that $C(X)$, the space of continuous functions on X with the uniform norm has a countable dense set, $(f_n)_{n\geq 1}$ say. If the invariant measure μ is ergodic, then for each n, there is a set B_n of measure 1 such that for all $x \in B_n$, $A_N f_n(x) \to \int f_n \, d\mu$. Letting $B = \bigcap_n B_n$, one obtains a full measure set such that for all n and all $x \in B$, $A_N f_n(x) \to \int f_n \, d\mu$. A simple approximation argument then shows that for all $x \in B$ and all $f \in C(X)$, $A_N f(x) \to \int f \, d\mu$. A point x with this property is said to be *generic* for μ. The observations above show that for an ergodic invariant measure μ, we have $\mu\{x\colon x \text{ is generic for } \mu\} = 1$.

If T is ergodic, but T^n is not ergodic for some n, then one can show that the space X splits up as $A_1, \ldots, A_d$ for some $d|n$ in such a way that $T(A_i) = A_{i+1}$ for $i < d$ and $T(A_d) = A_1$ with T^n acting ergodically on each A_i. The transformation T is *totally ergodic* if T^n is ergodic for all $n \in \mathbb{N}$. One can check that a non-invertible transformation T is ergodic if and only if its natural extension is ergodic.

The following lemma gives an alternative characterization of ergodicity, which in particular relates it to mixing.

Lemma 5 (Ergodicity as a Mixing Property) *Let T be a measure-preserving transformation of the probability space $(X, \mathcal{B}, \mu)$. Then T is ergodic if and only if for all f and g in L^2,*

$$\frac{1}{N}\sum_{n=0}^{N} \langle f, g \circ T^n \rangle \to \langle f, 1\rangle\langle 1, g\rangle.$$

In particular, if T is ergodic, then $(1/N)\sum_{n=0}^{N-1} \mu(A \cap T^{-n}B) \to \mu(A)\mu(B)$ for all measurable sets A and B.

Proof Suppose that T is ergodic. Then the left-hand side of the equality is equal to $\langle f, (1/N)\sum_{n=0}^{N-1} g \circ T^n\rangle$. The mean ergodic theorem shows that the second term converges in L^2 to the constant function with value $\int g \, d\mu = \langle g, 1\rangle$, and the equality follows.

Conversely, if the equation holds for all f and g in L^2, suppose that A is an invariant set. Let $f = g = \mathbf{1}_A$. Then since $g \circ T^n = \mathbf{1}_A$ for all n, the left-hand side is $\langle \mathbf{1}_A, \mathbf{1}_A\rangle = \mu(A)$. On the other hand, the right-hand side is $\mu(A)^2$, so that the equation yields $\mu(A) = \mu(A)^2$, and $\mu(A)$ is either 0 or 1 as required.

Taking $f = \mathbf{1}_A$ and $g = \mathbf{1}_B$ for measurable sets A and B gives the final statement. □

We now examine the ergodicity of the examples presented above. Firstly, for the rotation of the circle, we claim that the transformation is ergodic if and only if the 'angle' α is irrational. To see this, we argue as follows. If $\alpha = p/q$, then we see that $f(x) = e^{2\pi i q x}$ is a non-constant R_α-invariant function, and hence R_α is not ergodic. On the other hand, if α is irrational, suppose f is a bounded measurable invariant function. Since f is bounded, it is an L^2 function, and so f may be expressed in L^2 as a Fourier series: $f = \sum_{n\in\mathbb{Z}} c_n e_n$ where $e_n(x) = e^{2\pi i n x}$. We then see that $f \circ R_\alpha = \sum_{n\in\mathbb{Z}} e^{2\pi i n\alpha} c_n e_n$. In order for f to be equal in L^2 to $f \circ R_\alpha$, they must have the same Fourier coefficients, so that $c_n = e^{2\pi i n\alpha} c_n$ for each n. Since α is irrational, this forces $c_n = 0$ for all $n \neq 0$, so that f is constant as required.

The doubling map and the Bernoulli shift are both ergodic, although we defer proof of this for the time being, since they in fact have the strong-mixing property. A Markov chain with matrix P and vector π is ergodic if and only if for all i and j in A with $\pi_i > 0$ and $\pi_j > 0$, there exists an $n \geq 0$ with $P^n_{ij} > 0$. This follows from the ergodic theorem for Markov chains (which is derived from the Strong Law of Large Numbers) (see [18] for details). In particular, if the underlying Markov chain is irreducible, then the measure is ergodic.

In the case of interval exchange transformations, there is a simple necessary condition on the permutation for irreducibility, namely for $1 \leq j < k$, we do not have $\pi\{1, \ldots, j\} = \{1, \ldots, j\}$. Under this condition, Masur [49] and Veech [88] independently showed that for almost all values of the sequence of lengths $(\ell_i)_{1\leq i\leq k}$, the interval exchange transformation is ergodic. (In fact they showed the stronger condition of *unique ergodicity*: that the transformation has no other invariant measure than Lebesgue measure. This implies that Lebesgue measure is ergodic, because if there were a non-trivial invariant set, then the restriction of Lebesgue measure to that set would be another invariant measure).

For the hard sphere systems, there are no results on ergodicity in full generality. Important special cases have been studied by Sinai [84], Sinai and Chernov [86], Krámli, Simányi and Szász [45], Simányi and Szász [81], Simnyi [79,80] and Young [95].

Ergodic Decomposition

We already observed that if a transformation is not ergodic, then it may be decomposed into parts. Clearly if these parts are not ergodic, they may be further decomposed. It is natural to ask whether the transformation can be decomposed into ergodic parts, and if so what form does the decomposition take? In fact such a decomposition does exist, but rather than decompose the transforma-

tion, it is necessary to decompose the measure into ergodic pieces. This is known as ergodic decomposition.

The set of invariant measures for a measurable map T of a measurable space $(X, \mathcal{B})$ to itself forms a simplex. General functional analytic considerations (due to Choquet [14,15] – see also Phelps' account [66] of this theory) mean that it is possible to write any member of the simplex as an integral-convex combination of the extreme points. Further, the extreme points of the simplex may be identified as precisely the ergodic invariant measures for T. It follows that any invariant probability measure μ for T may be uniquely expressed in the form

$$\mu(A) = \int_{M_{\mathrm{erg}}(X,T)} \nu(A)\,\mathrm{d}m(\nu)\,,$$

where $M_{\mathrm{erg}}(X, T)$ denotes the set of ergodic T-invariant measures on X and m is a measure on $M_{\mathrm{erg}}(X, T)$.

We will give a proof of this theorem in the special case of a continuous transformation of a compact space. Our proof is based on the Birkhoff ergodic theorem and the Riesz Representation Theorem identifying the dual space of the space of continuous functions on a compact space as the set of bounded signed measures on the space (see Rudin's book [75] for details). We include it here because this special case covers many cases that arise in practice, and because few of the standard ergodic theory references include a proof of ergodic decomposition. An exception to this is Rudolph's book [76] which gives a full proof in the case that X is a Lebesgue space. This is based on a detailed development of the theory of these spaces and builds measures using conditional exceptions. Kalikow's notes [32] give a brief outline of a proof similar to that which follows. Oxtoby [62] also wrote a survey article containing much of the following (and much more besides).

Theorem 6 *Let X be a compact metric space, $\mathcal{B}$ be the Borel σ-algebra, μ be an invariant Borel probability measure and T be a continuous measure-preserving transformation of $(X, \mathcal{B}, \mu)$. Then for each $x \in X$, there exists an invariant Borel measure μ_x such that:*

1. *For $f \in L^1(X, \mathcal{B}, \mu)$, $\int f\,\mathrm{d}\mu = \int \left(\int f\,\mathrm{d}\mu_x\right)\,\mathrm{d}\mu(x)$;*
2. *Given $f \in L^1(X, \mathcal{B}, \mu)$, for μ-almost every $x \in X$, one has $A_N f(x) \to \int f\,\mathrm{d}\mu_x$;*
3. *The measure μ_x is ergodic for μ-almost every $x \in X$.*

Notice that conclusion (2) shows that μ_x can be understood as the distribution on the phase space "seen" if one starts the system in an initial condition of x. This interpretation of the measures μ_x corresponds closely with the ideas of Boltzmann and the Ehrenfests in the formulation of the ergodic and quasi-ergodic hypotheses, which can be seen as demanding that μ_x is equal to μ for (almost) all x.

Proof The proof will be divided into 3 main steps: defining the measures μ_x, proving measurability with respect to x and proving ergodicity of the measures.

Step 1: Definition of μ_x

Given a function $f \in L^1(X, \mathcal{B}, \mu)$, Birkhoff's theorem states that for μ-almost every $x \in X$, $(A_N f(x))$ is convergent. It will be convenient to denote the limit by $\tilde{f}(x)$. Let $f_1, f_2, \ldots$ be a sequence of continuous functions that is dense in $C(X)$. For each k, there is a set B_k of x's measure 1 for which $(A_n f_k(x))_{n=1}^{\infty}$ is a convergent sequence. Intersecting these gives a set B of full measure such that for $x \in B$, for each $k \geq 1$, $A_n f_k(x)$ is convergent. A simple approximation argument shows that for $x \in B$ and f an arbitrary continuous function, $A_n f(x)$ is convergent. Given $x \in B$, define a map $L_x\colon C(X) \to \mathbb{R}$ by $L_x(f) = \tilde{f}(x)$. This is a continuous linear functional on $C(X)$, and hence by the Riesz Representation Theorem there exists a Borel measure μ_x such that $\tilde{f}(x) = \int f\,\mathrm{d}\mu_x$ for each $f \in C(X)$ and $x \in B$. Since $L_x(f) \geq 0$ when f is a non-negative function and $L_x(1) = 1$, the measure μ_x is a probability measure. Since $L_x(f \circ T) = L_x(f)$ for $f \in C(X)$, one can check that μ_x must be an invariant probability measure. For $x \notin B$, simply define $\mu_x = \mu$. Since B^c is a set of measure 0, this will not affect any of the statements that we are trying to prove.

Now for f continuous, we have $A_N f$ is a bounded sequence of functions with $A_N f(x)$ converging to $\int f\,\mathrm{d}\mu_x$ almost everywhere and $\int A_N f\,\mathrm{d}\mu = \int f\,\mathrm{d}\mu$ since T is measure-preserving. It follows from the bounded convergence theorem that for $f \in C(X)$,

$$\int f\,\mathrm{d}\mu = \int \left(\int f\,\mathrm{d}\mu_x\right)\,\mathrm{d}\mu(x)\,. \tag{1}$$

Step 2: Measurability of $x \mapsto \mu_x(A)$

Lemma 7 *Let $C \in \mathcal{B}$ satisfy $\mu(C) = 0$. Then $\mu_x(C) = 0$ for μ-almost every $x \in X$.*

Proof Using regularity of Borel probability measures (see Rudin's book [75] for details), there exist open sets $U_1 \supset U_2 \supset \cdots \supset C$ with $\mu(U_k) < 1/k$. There exist continuous functions $g_{k,m}$ with $(g_{k,m}(x))_{m=1}^{\infty}$ increasing to $\mathbf{1}_{U_k}$ everywhere (e. g. $g_{k,m}(x) = \min(1, m \cdot d(x, U_k^c))$). By (1), we have $\int(\int g_{k,m}\,\mathrm{d}\mu_x)\,\mathrm{d}\mu(x) < 1/k$ for all k, m. Note that $\int g_{k,m}\,\mathrm{d}\mu_x = \lim_{n\to\infty} A_n g_{k,m}(x)$ is a measurable function of x, so that using the monotone convergence theorem (taking the limit in m), $x \mapsto \int \mathbf{1}_{U_k}\,\mathrm{d}\mu_x = \mu_x(U_k)$ is measurable and $\int(\int \mathbf{1}_{U_k}\,\mathrm{d}\mu_x)\,\mathrm{d}\mu(x) \leq 1/k$. We now see that $x \mapsto \lim_{k\to\infty} \mu_x(U_k) = \mu_x(\bigcap U_k)$

is also measurable, and by monotone convergence we see $\int \mu_x(\bigcap U_k)\,\mathrm{d}\mu(x) = 0$. It follows that $\mu_x(\bigcap U_k) = 0$ for μ-almost every x. Since $\bigcap U_k \supset C$, the lemma follows. □

Given a set $A \in \mathcal{B}$, let f_k be a sequence of continuous functions (uniformly bounded by 1) satisfying $\|f_k - \mathbf{1}_A\|_{L^1(\mu)} < 2^{-n}$, so that in particular $f_k(x) \to \mathbf{1}_A(x)$ for μ-almost every x. For each k, $x \mapsto \int f_k\,\mathrm{d}\mu_x = \lim_{n\to\infty} A_n f_k(x)$ is a measurable function. By Lemma 7, for μ-almost every x, $f_k \to 1_A$ μ_x-almost everywhere, so that by the bounded convergence theorem $\lim_{k\to\infty} \int f_k\,\mathrm{d}\mu_x = \mu_x(A)$ for μ-almost every x. Since the limit of measurable functions is measurable, it follows that $x \mapsto \mu_x(A)$ is measurable for any measurable set $A \in \mathcal{B}$.

This allows us to define a measure ν by $\nu(A) = \int \mu_x(A)\,\mathrm{d}\mu(x)$. For a bounded measurable function f, we have $\int f\,\mathrm{d}\nu = \int(\int f\,\mathrm{d}\mu_x)\,\mathrm{d}\mu(x)$. Since this agrees with $\int f\,\mathrm{d}\mu$ for continuous functions by (1), it follows that $\mu = \nu$. Conclusion (1) of the theorem now follows easily.

Given $f \in L^1(X)$, we let (f_k) be a sequence of continuous functions such that $\|f_k - f\|_{L^1(\mu)}$ is summable. This implies that $\|f_k - f\|_{L^1(\mu_x)}$ is summable for μ-almost every x and in particular, $\int f_k\,\mathrm{d}\mu_x \to \int f\,\mathrm{d}\mu_x$ for almost every x. On the other hand, by the remark following the statement of Birkhoff's theorem, we have $\tilde{f}_k = \mathbb{E}(f_k|\mathcal{I})$ so that $\|\tilde{f} - \tilde{f}_k\|_{L^1(\mu)}$ is summable and $\tilde{f}_k(x) \to \tilde{f}(x)$ for μ-almost every x. Combining these two statements, we see that for μ-almost every x, we have

$$\tilde{f}(x) = \lim_{k\to\infty} \tilde{f}_k(x) = \lim_{k\to\infty} \int f_k\,\mathrm{d}\mu_x = \int f\,\mathrm{d}\mu_x\,.$$

This establishes conclusion (2) of the theorem.

Step 3: Ergodicity of μ_x

We have shown how to disintegrate the invariant measure μ as an integral combination of μ_x's, and we have interpreted the μ_x's as describing the average behavior starting from x. It remains to show that the μ_x's are ergodic measures.

Fix for now a continuous function f and a number $0 < \epsilon < 1$. Since $A_n f(x) \to \tilde{f}(x)$ μ-almost everywhere, there exists an N such that $\mu\{x\colon |A_N f(x) - \tilde{f}(x)| > \epsilon/2\} < \epsilon^3/8$.

We now claim the following:

$$\mu\big\{x\colon \mu_x\{y\colon |\tilde{f}(y) - \textstyle\int f\,\mathrm{d}\mu_x| > \epsilon\} > \epsilon\big\} < \epsilon\,. \qquad (2)$$

To see this, note that $\{y\colon |\tilde{f}(y) - \int f\,\mathrm{d}\mu_x| > \epsilon\} \subset \{y\colon |\tilde{f}(y) - A_N f(y)| > \epsilon/2\} \cup \{y\colon |A_N f(y) - \tilde{f}(x)| > \epsilon/2\}$, so that if $\mu_x\{y\colon |\tilde{f}(y) - \int f\,\mathrm{d}\mu_x| > \epsilon\} > \epsilon$, then either $\mu_x\{y\colon |\tilde{f}(y) - A_N f(y)| > \epsilon/2\} > \epsilon/2$ or $\mu_x\{y\colon |A_N f(y) - \tilde{f}(x)| > \epsilon/2\} > \epsilon/2$. We show that the set of x's satisfying each condition is small.

Firstly, we have $\epsilon^3/8 > \mu\{y\colon |\tilde{f}(y) - A_N f(y)| > \epsilon/2\} = \int \mu_x\{y\colon |\tilde{f}(y) - A_N f(y)| > \epsilon/2\}\,\mathrm{d}\mu(x)$, so that $\mu\{x\colon \mu_x\{y\colon |\tilde{f}(y) - A_N f(y)| > \epsilon/2\} > \epsilon/2\} < \epsilon^2/4 < \epsilon/2$.

For the second term, given $c \in \mathbb{R}$, let $F_c(x) = |A_N f(x) - c|$ and $G(x) = F_{\tilde{f}(x)}(x)$. Note that

$$\int F_{\tilde{f}(x)}(y)\,\mathrm{d}\mu_x(y) = \lim_{n\to\infty} A_n F_{\tilde{f}(x)}(x) = \lim_{n\to\infty} A_n G(x)$$

(using the facts that $y \mapsto F_{\tilde{f}(x)}(y)$ is a continuous function and that since $\tilde{f}(x)$ is an invariant function, $F_{\tilde{f}(x)}(T^k x) = G(T^k x)$). Since $\int G(x)\,\mathrm{d}\mu(x) < \epsilon^3/8$, it follows that $\int F_{\tilde{f}(x)}(y)\,\mathrm{d}\mu_x(y) \le \epsilon^2/4$ except on a set of x's of measure less than $\epsilon/2$. Outside this bad set, we have $\mu_x\{y\colon |A_N f(y) - \tilde{f}(x)| > \epsilon/2\} < \epsilon/2$ so that $\mu\{x\colon \mu_x\{y\colon |A_N f(y) - \tilde{f}(x)| > \epsilon/2\} > \epsilon/2\} < \epsilon/2$ as required.

This establishes our claim (2) above. Since $\epsilon > 0$ is arbitrary, it follows that for each $f \in C(X)$, for μ-almost every x, μ_x-almost every y satisfies $\tilde{f}(y) = \int f\,\mathrm{d}\mu_x$. As usual, taking a countable dense sequence (f_k) in $C(X)$, it is the case that for all k and μ-almost every x, $\tilde{f}_k(y) = \int f_k\,\mathrm{d}\mu_x$ μ_x-almost everywhere. Let the set of x's with this property be D. We claim that for $x \in D$, μ_x is ergodic. Suppose not. Then let $x \in D$ and let J be an invariant set of μ_x measure between δ and $1-\delta$ for some $\delta > 0$. Then by density of $C(X)$ in $L^1(\mu_x)$, there exists an f_k with $\|f_k - \mathbf{1}_J\|_{L^1(\mu_x)} < \delta$. Since $\mathbf{1}_J$ is an invariant function, we have $\tilde{\mathbf{1}}_J = \mathbf{1}_J$. On the other hand, $\tilde{f}_k$ is a constant function. It follows that $\|\tilde{f}_k - \tilde{\mathbf{1}}_J\|_{L^1(\mu_x)} \ge \delta > \|f_k - \mathbf{1}_J\|_{L_1(\mu_x)}$. This contradicts the identification of the limit as a conditional expectation and concludes the proof of the theorem. □

Mixing

As mentioned above, ergodicity may be seen as an independence on average property. More specifically, one wants to know whether in some sense $\mu(A \cap T^{-n}B)$ converges to $\mu(A)\mu(B)$ as $n \to \infty$. Ergodicity is the property that there is convergence in the Césaro sense. Weak-mixing is the property that there is convergence in the strong Césaro sense. That is, a measure-preserving transformation T is *weak-mixing* if

$$\frac{1}{N}\sum_{n=0}^{N-1} |\mu(A \cap T^{-n}B) - \mu(A)\mu(B)| \to 0 \quad \text{as} \quad N \to \infty\,.$$

In order for T to be *strong-mixing*, we require simply $\mu(A \cap T^{-N}B) \to \mu(A)\mu(B)$ as $N \to \infty$. It is clear that strong-mixing implies weak-mixing and weak-mixing implies ergodicity.

If T^d is not ergodic (so that $T^{-d}A = A$ for some A of measure strictly between 0 and 1), then $|\mu(T^{-nd}A \cap A) - \mu(A)^2| = \mu(A)(1-\mu(A))$, so that T is not weak-mixing.

An alternative characterization of weak-mixing is as follows:

Lemma 8 *The measure-preserving transformation T is weak-mixing if and only if for every pair of measurable sets A and B, there exists a subset J of $\mathbb{N}$ of density 1 (i. e. $\#(J \cap \{1,\dots,N\})/N \to 1$) such that*

$$\lim_{n\to\infty\, n\notin J} \mu(A \cap T^{-n}B) = \mu(A)\mu(B)\,. \tag{3}$$

By taking a countable family of measurable sets that are dense (with respect to the metric $d(A,B) = \mu(A\Delta B)$) and taking a suitable intersection of the corresponding J sets, one shows that for a given weak-mixing measure-preserving transformation, there is a single set $J \subset \mathbb{N}$ such that (3) holds for *all* measurable sets A and B (see Petersen [64] or Walters [92] for a proof).

We show that an irrational rotation of the circle is not weak-mixing as follows: let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ and let A be the interval $[\frac{1}{4}, \frac{3}{4})$. There is a positive proportion of n's in the natural numbers (in fact proportion 1/3) with the property that $|T^n(\frac{1}{2}) - \frac{1}{2}| < \frac{1}{6}$. For these n's $\mu(A \cap T^{-n}A) > \frac{1}{3}$, so that in particular $|\mu(A \cap T^{-n}A) - \mu(A)\mu(A)| > \frac{1}{12}$. Clearly this precludes the required convergence to 0 in the definition of weak-mixing, so that an irrational rotation is ergodic but not weak-mixing. Since $R_\alpha^n = R_{n\alpha}$, the earlier argument shows that R_α^n is ergodic, so that R_α is totally ergodic.

On the other hand, we show that any Bernoulli shift is strong-mixing. To see this, let A and B be arbitrary measurable sets. By standard measure-theoretic arguments, A and B may each be approximated arbitrarily closely by a finite union of cylinder sets. Since if A' and B' are finite unions of cylinder sets, we have that $\mu(A' \cap T^{-n}B')$ is *equal* to $\mu(A')\mu(B')$ for large n, it is easy to deduce that $\mu(A \cap T^{-n}B) \to \mu(A)\mu(B)$ as required. Since the doubling map is measure-theoretically isomorphic to a one-sided Bernoulli shift, it follows that the doubling map is also strong-mixing.

Similarly, if a Markov Chain is irreducible (i. e. for any states i and j, there exists an $n \geq 0$ such that $P_{ij}^n > 0$) and aperiodic (there is a state i such that $\gcd\{n\colon P_{ii}^n > 0\} = 1$), then given any pair of cylinder sets A' and B' we have by standard theorems of Markov chains $\mu(A' \cap T^{-n}B') \to \mu(A')\mu(B')$. The same argument as above then shows that an aperiodic irreducible Markov Chain is strong-mixing. On the other hand, if a Markov chain is periodic ($d = \gcd\{n\colon P_{ii}^n > 0\} > 0$), then letting $A = B = \{x\colon x_0 = i\}$, we have that $\mu(A \cap T^{-n}B) = 0$ whenever $d \nmid n$. It follows T^d is not ergodic, so that T is not weak-mixing.

Both weak- and strong-mixing have formulations in terms of functions:

Lemma 9 *Let T be a measure-preserving transformation of the probability space $(X, \mathcal{B}, \mu)$.*

1. *T is weak-mixing if and only if for every $f, g \in L^2$ one has*

$$\frac{1}{N}\sum_{n=0}^{N-1} |\langle f, g \circ T^n\rangle - \langle f,1\rangle\langle 1,g\rangle| \to 0 \text{ as } N \to \infty\,.$$

2. *T is strong-mixing if and only if for every $f, g \in L^2$, one has*

$$\langle f, g \circ T^N\rangle \to \langle f,1\rangle\langle 1,g\rangle \text{ as } N \to \infty\,.$$

Using this, one can see that both mixing conditions are spectral properties.

Lemma 10 *Weak- and strong-mixing are spectral properties.*

Proof Suppose S is a weak-mixing transformation of $(Y, \mathcal{F}, \nu)$ and the transformation T of $(X, \mathcal{B}, \mu)$ is spectrally isomorphic to S by the Hilbert space isomorphism Θ. Then for $f, g \in L^2(X, \mathcal{B}, \mu)$, $\langle f, g \circ T^n\rangle_X - \langle f,1\rangle_X\langle 1,g\rangle_X = \langle \Theta(f), \Theta(g) \circ S^n\rangle_Y - \langle \Theta(f), \Theta(1)\rangle_Y\langle \Theta(1), \Theta(g)\rangle_Y$. Since 1 is an eigenfunction of U_T with eigenvalue 1, $\Theta(1)$ is an eigenfunction of U_S with an eigenvalue 1, so since S is ergodic, $\Theta(1)$ must be a constant function. Since Θ preserves norms, $\Theta(1)$ must have a constant value of absolute value 1 and hence $\langle f, g \circ T^n\rangle_X - \langle f,1\rangle_X\langle 1,g\rangle_X = \langle \Theta(f), \Theta(g) \circ S^n\rangle_Y - \langle \Theta(f), 1\rangle_Y\langle 1, \Theta(g)\rangle_Y$. It follows from Lemma 9 that T is weak-mixing.

A similar proof shows that strong-mixing is a spectral property. □

Both weak- and strong-mixing properties are preserved by taking natural extensions.

Recent work of Avila and Forni [4] shows that for interval exchange transformations of $k \geq 3$ intervals with the underlying permutation satisfying the non-degeneracy condition above, almost all divisions of the interval (with respect to Lebesgue measure on the $k-1$-dimensional simplex) lead to weak-mixing transformations. On the

other hand, work of Katok [36] shows that no interval exchange transformation is strong-mixing.

It is of interest to understand the behavior of the 'typical' measure-preserving transformation. There are a number of Baire category results addressing this. In order to state them, one needs a set of measure-preserving transformations and a topology on them. As mentioned earlier, it is effectively no restriction to assume that a transformation is a Lebesgue-measurable map on the unit interval preserving Lebesgue measure. The classical category results are then on the collection of invertible Lebesgue-measure preserving transformations of the unit interval. One topology on these is the 'weak' topology, where a sub-base is given by sets of the form $N(T, A, \epsilon) = \{S\colon \lambda(S(A)\Delta T(A)) < \epsilon\}$. With respect to this topology, Halmos [26] showed that a residual set (i. e. a dense G_δ set) of invertible measure-preserving transformations is weak-mixing (see also work of Alpern [3]), while Rokhlin [72] showed that the set of strong-mixing transformations is meagre (i. e. a nowhere dense F_σ set), allowing one to conclude that with respect to this topology, the typical transformation is weak- but not strong-mixing.

As often happens in these cases, even when a certain kind of behavior is typical, it may not be simple to exhibit concrete examples. In this case, a well-known example of a transformation that is weak-mixing but not strong-mixing was given by Chacon [13].

While on the face of it the formulation of weak-mixing is considerably less natural than that of strong-mixing, the notion of weak-mixing turns out to be extremely natural from a spectral point of view. Given a measure-preserving transformation T, let U_T be the Koopman operator described above. Since this operator is an isometry, any eigenvalue must lie on the unit circle. The constant function 1 is always an eigenfunction with eigenvalue 1. If T is ergodic and g and h are eigenfunctions of U_T with eigenvalue λ, then $g\bar{h}$ is an eigenfunction with eigenvalue 1, hence invariant, so that $g = Kh$ for some constant K. We see that for ergodic transformations, up to rescaling, there is at most one eigenfunction with any given eigenvalue.

If U_T has a non-constant eigenfunction f, then one has $|\langle U_T^n f, f\rangle| = \|f\|^2$ for each n, whereas by Cauchy–Schwartz, $|\langle f, 1\rangle|^2 < \|f\|^2$. It follows that $|\langle U_T^n f, f\rangle - \langle f, 1\rangle\langle 1, f\rangle| \geq c$ for some positive constant c, so that using Lemma 9, T is not weak-mixing.

Using the spectral theorem, the converse is shown to hold.

Theorem 11 *The measure-preserving transformation T is weak-mixing if and only U_T has no non-constant eigenfunctions.*

Of course this also shows that weak-mixing is a spectral property. Equivalently, this says that the transformation T is weak-mixing if and only if the apart from the constant eigenfunction, the operator U_T has only continuous spectrum (that is, the operator has no other eigenfunctions). For a very nice and concise development of the part of spectral theory relevant to ergodic theory, the reader is referred to the Appendix in Parry's book [63].

Using this theory, one can establish the following:

Theorem 12

1. *T is weak-mixing if and only if $T \times T$ is ergodic;*
2. *If T and S are ergodic, then $T \times S$ is ergodic if and only if U_S and U_T have no common eigenvalues other than 1.*

Proof The main factor in the proof is that the eigenvalues of $U_{T\times S}$ are precisely the set of $\alpha\beta$, where α is an eigenvalue of U_T and β is an eigenvalue of U_S. Further, the eigenfunctions of $U_{T\times S}$ with eigenvalue γ are spanned by eigenfunctions of the form $f \otimes g$, where f is an eigenfunction of U_T, g is an eigenfunction of U_S, and the product of the eigenvalues is γ.

Suppose that T is weak-mixing. Then the only eigenfunction is the constant function, so that the only eigenfunction of $U_{T\times T}$ is the constant function, proving that $T \times T$ is ergodic. Conversely, if U_T has an eigenvalue (so that $f \circ T = \alpha T$ for some non-constant f) then $f \otimes \bar{f}$ is a non-constant invariant function of $T \times T$ so that $T \times T$ is not ergodic.

For the second part, if U_S and U_T have a common eigenvalue other than 1 (say $f \circ T = \alpha f$ and $g \circ T = \alpha g$), then $f \otimes \bar{g}$ is a non-constant invariant function. Conversely, if $T \times S$ has a non-constant invariant function h, then h can be decomposed into functions of the form $f \otimes g$, where f and g are eigenfunctions of U_T and U_S respectively with eigenvalues α and β satisfying $\alpha\beta = 1$. Since the eigenvalues of S are closed under complex conjugation, we see that U_T and U_S have a common eigenvalue other than 1 as required. □

For a measure-preserving transformation T, we let K be the subspace of L^2 spanned by the eigenfunctions of U_T. It is a remarkable fact that K may be identified as $L^2(X, \mathcal{B}', \mu)$ where $\mathcal{B}'$ is a sub-σ-algebra of $\mathcal{B}$. The space K is called the *Kronecker factor* of T. The terminology comes from the fact that any sub-σ-algebra $\mathcal{F}$ of $\mathcal{B}$ gives rise to a factor mapping $\pi\colon (X, \mathcal{B}, \mu) \to (X, \mathcal{F}, \mu)$ with $\pi(x) = x$. By construction $L^2(X, \mathcal{B}', \mu)$ is the closed linear span of the eigenfunctions of T considered as a measure-preserving transformation of $(X, \mathcal{B}', \mu)$. By the Discrete Spectrum Theorem of Halmos and von Neumann [27],

T acting on $(X, \mathcal{B}', \mu)$ is measure-theoretically isomorphic to a rotation on a compact group. This allows one to split $L^2(X, \mathcal{B}, \mu)$ as $L^2(X, \mathcal{B}', \mu) \oplus L^2_c(X, \mathcal{B}, \mu)$, where, as mentioned above the first part is the discrete spectrum part, spanned by eigenfunctions, and the second part is the continuous spectrum part, consisting of functions whose spectral measure is continuous. Since we have split L^2 into a discrete part and a continuous part, it is natural to ask whether the underlying transformation T can be split up in some way into a weak-mixing part and a discrete spectrum (compact group rotation) part, somewhat analogously to the ergodic decomposition. Unfortunately, there is no-such decomposition available. However for some applications, for example to multiple recurrence (starting with the work of Furstenberg [20,21]), the decomposition of L^2 (possibly into more complicated parts) plays a crucial role (see the chapters on ► Ergodic Theory: Recurrence and ► Ergodic Theory: Interactions with Combinatorics and Number Theory).

For non-invertible measure-preserving transformations, the transformation is weak- or strong-mixing if and only if its natural extension has that property.

The understanding of weak-mixing in terms of the discrete part of the spectrum of the operator also extends to total ergodicity. T^n is ergodic if and only if T has no eigenvalues of the form $e^{2\pi i p/n}$ other than 1. From this it follows that an ergodic measure-preserving transformation T is totally ergodic if and only if it has no *rational spectrum* (i. e. no eigenvalues of the form $e^{2\pi i p/q}$ other than the simple eigenvalue 1).

An intermediate mixing condition between strong- and weak- mixing is that a measure-preserving transformation is *mild-mixing* if whenever $f \circ T^{n_i} \to f$ for an L^2 function f and a sequence $n_i \to \infty$, then f is a.e. constant. Clearly mild-mixing is a spectral property. If a transformation has an eigenfunction f, then it is straightforward to find a sequence n_i such that $f \circ T^{n_i} \to f$, so we see that mild-mixing implies weak-mixing. To see that strong-mixing implies mild-mixing, suppose that T is strong-mixing and that $f \circ T^{n_i} \to f$. Then we have $\int f \circ T^{n_i} \bar{f} \to \|f\|^2$. On the other hand, the strong mixing property implies that $\int f \circ T^{n_i} \bar{f} \to |\langle f, 1\rangle|^2$. The equality of these implies that f is a. e. constant. Mild-mixing has a useful reformulation in terms of ergodicity of general (not necessarily probability) measure-preserving transformations: A transformation T is mild-mixing if and only if for every conservative ergodic measure-preserving transformation S, $T \times S$ is ergodic. See Furstenberg and Weiss' article [22] for further information on mild-mixing.

The strongest spectral property that we consider is that of having countable Lebesgue spectrum. While we will avoid a detailed discussion of spectral theory in this article, this is a special case that can be described simply. Specifically, let T be an invertible measure-preserving transformation. Then T has countable Lebesgue spectrum if there is a sequence of functions $f_1, f_2, \ldots$ such that $\{1\} \cup \{U_T^n f_j \colon n \in \mathbb{Z}, j \in \mathbb{N}\}$ forms an orthonormal basis for $L^2(X)$.

To see that this property is stronger than strong-mixing, we simply observe that it implies that $\langle U_T^t U_T^n f_j, U_T^m f_k\rangle \to 0$ as $t \to \infty$. Then by approximating f and g by their expansions with respect to a finite part of the basis, we deduce that $\langle U_T^n f, g\rangle \to \langle f, 1\rangle\langle 1, g\rangle$ as required. Since already strong-mixing is atypical from the topological point of view, it follows that countable Lebesgue spectrum has to be atypical. In fact, Yuzvinskii [96] showed that the typical invertible measure-preserving transformation has simple singular spectrum.

The property of countable Lebesgue spectrum is by definition a spectral property. Since it completely describes the transformation up to spectral isomorphism, there can be no stronger spectral properties. The remaining properties that we shall examine are invariant under measure-theoretic isomorphisms only.

An invertible measure-preserving transformation T of $(X, \mathcal{B}, \mu)$ is said to be K (for Kolmogorov) if there is a sub-σ-algebra $\mathcal{F}$ of $\mathcal{B}$ such that

1. $\bigcap_{n=1}^{\infty} T^{-n}\mathcal{F}$ is the trivial σ-algebra up to sets of measure 0 (i. e. the intersection consists only of null sets and sets of full measure).
2. $\bigvee_{n=1}^{\infty} T^n \mathcal{F} = \mathcal{B}$ (i. e. the smallest σ-algebra containing $T^n\mathcal{F}$ for all $n > 0$ is $\mathcal{B}$).

The K property has a useful reformulation in terms of entropy as follows: T is K if and only if for every non-trivial partition $\mathcal{P}$ of X, the entropy of T with respect to the partition $\mathcal{P}$ is positive: T has *completely positive entropy*. See the chapter on ► Entropy in Ergodic Theory for the relevant definitions. The equivalence of the K property and completely positive entropy was shown by Rokhlin and Sinai [74]. For a general transformation T, one can consider the collection of all subsets B of X such that with respect to the partition $\mathcal{P}_B = \{B, B^c\}$, $h(\mathcal{P}_B) = 0$. One can show that this is a σ-algebra. This σ-algebra is known as the *Pinsker* σ-algebra. The above reformulation allows us to say that a transformation is K if and only if it has a trivial Pinsker σ-*algebra*.

The K property implies countable Lebesgue spectrum (see Parry's book [63] for a proof). To see that K is not implied by countable Lebesgue spectrum, we point out that certain measure-preserving transformations derived from Gaussian systems (see for example the paper of Parry and

Newton [52]) have countable Lebesgue spectrum but zero entropy.

The fact that (two-sided) Bernoulli shifts have the K property follows from Kolmogorov's 0–1 law by taking $\mathcal{F} = \bigvee_{n=0}^{\infty} T^{-n}\mathcal{P}$, where $\mathcal{P}$ is the partition into cylinder sets (see Williams's book [93] for details of the 0–1 law).

Although the K property is explicitly an invertible property, it has a non-invertible counterpart, namely exactness. A transformation T of $(X, \mathcal{B}, \mu)$ is *exact* if $\bigcap_{n=0}^{\infty} T^{-n}\mathcal{B}$ consists entirely of null sets and sets of measure 1. It is not hard to see that a non-invertible transformation is exact if and only if its natural extension is K.

The final and strongest property in our list is that of being measure-theoretically isomorphic to a Bernoulli shift. If T is measure-theoretically isomorphic to a Bernoulli shift, we say that T has the *Bernoulli property*. While in principle this could apply to both invertible and non-invertible transformations, in practice the definition applies to a large class of invertible transformations, but occurs comparatively seldom for non-invertible transformations. For this reason, we will restrict ourselves to a discussion of the Bernoulli property for invertible transformations (see however work of Hoffman and Rudolph [29] and Heicklen and Hoffman [28] for work on the one-sided Bernoulli property).

In the case of invertible Bernoulli shifts, Ornstein [53,58] developed in the early 1970s a powerful isomorphism theory, showing that two Bernoulli shifts are measure-theoretically isomorphic if and only if they have the same entropy. Entropy had already been identified as an invariant by Kolmogorov and Sinai [43,82], so this established that it was a complete invariant for Bernoulli shifts. Keane and Smorodinsky [41] gave a proof which showed that two Bernoulli shifts of the same entropy are isomorphic using a conjugating map that is continuous almost everywhere. With other authors, this theory was extended to show that the property of being isomorphic to a Bernoulli shift applied to a surprisingly large class of measure-preserving transformations (e. g. geodesic flows on manifolds of constant negative curvature (Ornstein and Weiss [60]), aperiodic irreducible Markov chains (Friedman and Ornstein [19]), toral automorphisms (Katznelson [39]) and more generally many Gibbs measures for hyperbolic dynamical systems (see the book of Bowen [11])).

Initially, it was conjectured that the properties of being K and Bernoulli were the same, but since then a number of measure-preserving transformations that are K but not Bernoulli have been identified. The earliest was due to Ornstein [55]. Ornstein and Shields [59] then provided an uncountable family of non-isomorphic K automorphisms. Katok [37] gave an example of a smooth diffeomorphism that is K but not Bernoulli; and Kalikow [33] gave a very natural probabilistic example of a transformation that has this property (the T, T^{-1} process).

While in systems that one regularly encounters there is a correlation between positive entropy and the stronger mixing properties that we have discussed, these properties are logically independent (for example taking the product of a Bernoulli shift and the identity transformation gives a positive entropy transformation that fails to be ergodic; also, the zero entropy Gaussian systems with countable Lebesgue spectrum mentioned above have relatively strong mixing properties but zero entropy).

In many of the mixing criteria discussed above we have considered a pair of sets A and B and asked for asymptotic independence of A and B (so that for large n, A and $T^{-n}B$ become independent). It is natural to ask, given a finite collection of sets $A_0, A_1, \ldots, A_k$, under what conditions $\mu(A_0 \cap T^{-n_1}A_1 \cap \cdots \cap T^{-n_k}A_k)$ converges to $\prod_{j=0}^{k} \mu(A_j)$.

A measure-preserving transformation is said to be *mixing of order* $k+1$ if for all measurable sets $A_0, \ldots, A_k$,

$$\lim_{n_1\to\infty, n_{j+1}-n_j\to\infty} \mu\big(A_0 \cap T^{-n_1}A_1 \cap \cdots \cap T^{-n_k}A_k\big) = \prod_{j=0}^{k} \mu(A_j)\,.$$

An outstanding open question asked by Rokhlin [73] appearing already in Halmos' 1956 book [27] is to determine whether mixing (i. e. mixing of order 2) implies mixing of all orders. Kalikow [34] showed that mixing implies mixing of all orders for rank 1 transformations (existence of rank one mixing transformations having been previously established by Ornstein in [54]). Later Ryzhikov [77] used joining methods to establish the result for transformations with finite rank, and Host [30] also used joining methods to establish the result for measure-preserving transformations with singular spectrum, but the general question remains open.

It is not hard to show using martingale arguments that K automorphisms and hence all Bernoulli measure-preserving transformations are mixing of all orders.

For weak-mixing transformations, Furstenberg [21] has established the following *weak-mixing of all orders* statement: if a measure-preserving transformation T is weak-mixing, then given sets $A_0, \ldots, A_k$, there is a subsequence J of the integers of density 0 such that

$$\lim_{n\to\infty\, n\notin J} \mu\big(A_0 \cap T^{-n}A_1 \cap \cdots \cap T^{-kn}A_k\big) = \prod_{i=0}^{k} \mu(A_i)\,.$$

Bergelson [5] generalized this by showing that

$$\lim_{n\to\infty, n\notin J} \mu\big(A_0 \cap T^{-p_1(n)}A_1 \cap \cdots \cap T^{-p_k(n)}A_k\big) = \prod_{i=0}^{k} \mu(A_i)$$

whenever $p_1(n), \ldots, p_k(n)$ are non-constant integer-valued polynomials such that $p_i(n) - p_j(n)$ is unbounded for $i \neq j$. The method of proof of both of these results was a Hilbert space version of the van der Corput inequality of analytic number theory. Furstenberg's proof played a key role in his ergodic proof [20] of Szemerédi's theorem on the existence of arbitrarily long arithmetic progressions in a subset of the integers of positive density (see the chapter on ▶ Ergodic Theory: Interactions with Combinatorics and Number Theory for more information about this direction of study).

The conclusions that one draws here are much weaker than the requirement for mixing of all orders. For mixing of all orders, it was required that provided the gaps between $0, n_1, \ldots, n_k$ diverge to infinity, one achieves asymptotic independence, whereas for these weak-mixing results, the gaps are increasing along prescribed sequences with regular growth properties.

It is interesting to note that the analogous question of whether mixing implies mixing of all orders is known to fail in higher-dimensional actions. Here, rather than a $\mathbb{Z}$ action, in which there is a single measure-preserving transformation (so that the integer n acts on a point $x \in X$ by mapping it to $T^n x$), one takes a $\mathbb{Z}^d$ action. For such an action, one has d commuting transformations $T_1, \ldots, T_d$ and a vector $(n_1, \ldots, n_d)$ acts on a point x by sending it to $T_1^{n_1} \cdots T_d^{n_d} x$. Ledrappier [46] studied the following two-dimensional action. Let $X = \{x \in \{0,1\}^{\mathbb{Z}^2} : x_{\mathbf{v}} + x_{\mathbf{v}+\mathbf{e}_1} + x_{\mathbf{v}+\mathbf{e}_2} = 0 \pmod 2\}$ and let $T_i(x)_{\mathbf{v}} = x_{\mathbf{v}+\mathbf{e}_i}$. Since X is a compact Abelian group, it has a natural measure μ invariant under the group operations (the Haar measure). It is not hard to show that this system is mixing (i. e. given any measurable sets A and B, $\mu(A \cap T_1^{-n_1} T_2^{-n_2} B) \to \mu(A)\mu(B)$ as $\|(n_1, n_2)\| \to \infty$). Ledrappier showed that the system fails to be 3-mixing. Subsequently Masser [48] established necessary and sufficient conditions for similar higher-dimensional algebraic actions to be mixing of order k but not order $k+1$ for any given k.

Hyperbolicity and Decay of Correlations

One class of systems in which the stronger mixing properties are often found is the class of smooth systems possessing uniform hyperbolicity (i. e. the tangent space to the manifold at each point splits into stable and unstable subspaces $E^s(x)$ and $E^u(x)$ such that the $\|DT|_{E_s(x)}\| \leq a < 1$ for all x and $\|DT^{-1}|_{E_u(x)}\| \leq a$ and $DT(E_s(x)) = E_s(T(x))$ and $DT(E_u(x)) = E_u(T(x))$). In some cases similar conclusions are found in systems possessing non-uniform hyperbolicity. See Katok and Hasselblatt's book [38] for an overview of hyperbolic dynamical systems, as well as the chapter in this volume on ▶ Smooth Ergodic Theory.

In the simple case of expanding piecewise continuous maps of the interval (that is, maps for which the absolute value of the derivative is uniformly bounded below by a constant greater than 1), it is known that if they are totally ergodic and topologically transitive (i. e. the forward images of any interval cover the entire interval), then provided that the map has sufficient smoothness (e. g. the map is C^1 and the derivative satisfies a certain additional summability condition), the map has a unique absolutely continuous invariant measure which is exact and whose natural extension is Bernoulli (see the paper of Góra [25] for results of this type proved under some of the mildest hypotheses). These results were originally established for maps that were twice continuously differentiable, and the hypotheses were progressively weakened, approaching, but never meeting, C^1. Subsequent work of Quas [68,69] provided examples of C^1 expanding maps of the interval for which Lebesgue measure was invariant, but respectively not ergodic and not weak-mixing. Some of the key tools in controlling mixing in one-dimensional expanding maps that are absent in the C^1 case are bounded distortion estimates. Here, there is a constant $1 \leq C < \infty$ such that given any interval I on which some power T^n of T acts injectively and any sub-interval J of I, one has $1/C \leq (|T^n J|/|T^n I|)/(|J|/|I|) \leq C$. An early place in which bounded distortion estimates appear is the work of Rényi [70].

One important class of results for expanding maps establishes an exponential decay of correlations. Here, one starts with a pair of smooth functions f and g and one estimates $\int f \cdot g \circ T^n \, d\mu - \int f \, d\mu \int g \, d\mu$, where μ is an absolutely continuous invariant measure. If μ is mixing, we expect this to converge to 0. In fact though, in good cases this converges to 0 at an exponential rate for each pair of functions f and g belonging to a sufficiently smooth class. In this case, the measure-preserving transformation T is said to have exponential decay of correlations. See Liverani's article [47] for an introduction to a method of establishing this based on cones. Exponential decay of correlations implies in particular that the natural extension is Bernoulli.

Hu [31] has studied the situation of maps of the interval for which the derivative is bigger than 1 everywhere

except at a fixed point, where the local behavior is of the form $x \mapsto x + x^{1+\alpha}$ for $0 < \alpha < 1$. In this case, rather than exhibiting exponential decay of correlations, the map has polynomial decay of correlations with a rate depending on α.

In Young's survey [94], a variety of techniques are outlined for understanding the strong ergodic properties of non-uniformly hyperbolic diffeomorphisms. In her article [95], methods are introduced for studying many classes of non-uniformly hyperbolic systems by looking at suitably high powers of the map, for which the power has strong hyperbolic behavior. The article shows how to understand the ergodic behavior of these systems. These methods are applied (for example) to billiards, one-dimensional quadratic maps and Hénon maps.

Future Directions

Problem 1 (Mixing of all orders) Does mixing imply mixing of all orders? Can the results of Kalikow, Ryzhikov and Host be extended to larger classes of measure-preserving transformations? Thouvenot observed that it is sufficient to establish the result for measure-preserving transformations of entropy 0. This observation (whose proof is based on the Pinsker σ-algebra) was stated in Kalikow's paper [34] and is reproduced as Proposition 3.2 in recent work of de la Rue [16] on the mixing of all orders problem.

Problem 2 (Multiple weak-mixing) As mentioned above, Bergelson [5] showed that if T is a weak-mixing transformation, then there is a subset J of the integers of density 0 such that

$$\lim_{n\to\infty, n\notin J} \mu\big(A_0 \cap T^{-p_1(n)}A_1 \cap \cdots \cap T^{-p_k(n)}A_k\big) = \prod_{i=0}^{k} \mu(A_i)$$

whenever $p_1(n), \ldots, p_k(n)$ are non-constant integer-valued polynomials such that $p_i(n) - p_j(n)$ is unbounded for $i \neq j$. It is natural to ask what is the most general class of times that can replace the sequences $(p_1(n)), \ldots, (p_k(n))$. In unpublished notes, Bergelson and Håland considered as times the values taken by a family of integer-valued generalized polynomials (those functions of an integer variable that can be obtained by the operations of addition, multiplication, addition of or multiplication by a real constant and taking integer parts (e. g. $g(n) = \lfloor \sqrt{2}\lfloor \pi n \rfloor + \lfloor \sqrt{3} n \rfloor^2 \rfloor$)). They conjectured necessary and sufficient conditions for the analogue of Bergelson's weak-mixing polynomial ergodic theorem to hold, and proved the conjecture in certain cases.

In a recent paper of McCutcheon and Quas [50], the analogous question was addressed in the case where T is a mild-mixing transformation.

Problem 3 (Pascal adic transformation) Vershik [89,90] introduced a family of transformations known as the adic transformations. The underlying spaces for these transformations are certain spaces of paths on infinite graphs, and the transformations act by taking a path to its lexicographic neighbor. Amongst the adic transformations, the so-called Pascal adic transformation (so-called because the underlying graph resembles Pascal's triangle) has been singled out for attention in work of Petersen and others [2,10,51,65]. In particular, it is unresolved whether this transformation is weak-mixing with respect to any of its ergodic measures. Weak-mixing has been shown by Petersen and Schmidt to follow from a number-theoretic condition on the binomial coefficients [2,10].

Problem 4 (Weak Pinsker Conjecture) Pinsker [67] conjectured that in a measure-preserving transformation with positive entropy, one could express the transformation as a product of a Bernoulli shift with a system with zero entropy. This conjecture (now known as the *Strong Pinsker Conjecture*) was shown to be false by Ornstein [56,57]. Shields and Thouvenot [78] showed that the collection of transformations that can be written as a product of a zero entropy transformation with a Bernoulli shift is closed in the so-called $\bar{d}$-metric that lies at the heart of Ornstein's theory.

It is, however, the case that if $T\colon X \to X$ has entropy $h > 0$, then for all $h' \leq h$, T has a factor S with entropy h' (this was originally proved by Sinai [83] and reproved using the Ornstein machinery by Ornstein and Weiss in [61]). The *Weak Pinsker Conjecture* states that if a measure-preserving transformation T has entropy $h > 0$, then for all $\epsilon > 0$, T may be expressed as a product of a Bernoulli shift and a measure-preserving transformation with entropy less than ϵ.

Bibliography

1. Aaronson J (1997) An Introduction to Infinite Ergodic Theory. American Mathematical Society, Providence
2. Adams TM, Petersen K (1998) Binomial coefficient multiples of irrationals. Monatsh Math 125:269–278
3. Alpern S (1976) New proofs that weak mixing is generic. Invent Math 32:263–278
4. Avila A, Forni G (2007) Weak-mixing for interval exchange transformations and translation flows. Ann Math 165:637–664
5. Bergelson V (1987) Weakly mixing pet. Ergodic Theory Dynam Systems 7:337–349
6. Birkhoff GD (1931) Proof of the ergodic theorem. Proc Nat Acad Sci 17:656–660

7. Birkhoff GD, Smith PA (1924) Structural analysis of surface transformations. J Math 7:345–379
8. Boltzmann L (1871) Einige allgemeine Sätze über Wärmegleichgewicht. Wiener Berichte 63:679–711
9. Boltzmann L (1909) Wissenschaftliche Abhandlungen. Akademie der Wissenschaften, Berlin
10. Boshernitzan M, Berend D, Kolesnik G (2001) Irrational dilations of Pascal's triangle. Mathematika 48:159–168
11. Bowen R (1975) Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms. Springer, Berlin
12. Bradley RC (2005) Basic properties of strong mixing conditions. A survey and some open questions. Probab Surv 2:107–144
13. Chacon RV (1969) Weakly mixing transformations which are not strongly mixing. Proc Amer Math Soc 22:559–562
14. Choquet G (1956) Existence des représentations intégrales au moyen des points extrémaux dans les cônes convexes. C R Acad Sci Paris 243:699–702
15. Choquet G (1956) Unicité des représentations intégrales au moyen de points extrémaux dans les cônes convexes réticulés. C R Acad Sci Paris 243:555–557
16. de la Rue T (2006) 2-fold and 3-fold mixing: why 3-dot-type counterexamples are impossible in one dimension. Bull Braz Math Soc (NS) 37(4):503–521
17. Ehrenfest P, Ehrenfest T (1911) Begriffliche Grundlage der statistischen Auffassung in der Mechanik. No. 4. In: Encyclopädie der mathematischen Wissenschaften. Teubner, Leipzig
18. Feller W (1950) An Introduction to Probability and its Applications. Wiley, New York
19. Friedman NA, Ornstein DS (1970) On isomorphism of weak Bernoulli transformations. Adv Math 5:365–394
20. Furstenberg H (1977) Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. J Analyse Math 31:204–256
21. Furstenberg H (1981) Recurrence in Ergodic Theory and Combinatorial Number Theory. Princeton University Press, Princeton
22. Furstenberg H, Weiss B (1978) The finite multipliers of infinite ergodic transformations. In: The Structure of Attractors in Dynamical Systems. Proc Conf, North Dakota State Univ, Fargo, N.D., 1977. Springer, Berlin
23. Garsia AM (1965) A simple proof of E. Hopf's maximal ergodic theory. J Math Mech 14:381–382
24. Girsanov IV (1958) Spectra of dynamical systems generated by stationary Gaussian processes. Dokl Akad Nauk SSSR 119:851–853
25. Góra P (1994) Properties of invariant measures for piecewise expanding one-dimensional transformations with summable oscillations of derivative. Ergodic Theory Dynam Syst 14:475–492
26. Halmos PA (1944) In general a measure-preserving transformation is mixing. Ann Math 45:786–792
27. Halmos P (1956) Lectures on Ergodic Theory. Chelsea, New York
28. Hoffman C, Heicklen D (2002) Rational maps are *d*-adic bernoulli. Ann Math 156:103–114
29. Hoffman C, Rudolph DJ (2002) Uniform endomorphisms which are isomorphic to a Bernoulli shift. Ann Math 76:79–101
30. Host B (1991) Mixing of all orders and pairwise independent joinings of systems with singular spectrum. Israel J Math 76:289–298
31. Hu H (2004) Decay of correlations for piecewise smooth maps with indifferent fixed points. Ergodic Theory Dynam Syst 24:495–524
32. Kalikow S () Outline of ergodic theory. Notes freely available for download. See http://www.math.uvic.ca/faculty/aquas/kalikow/kalikow.html
33. Kalikow S (1982) T, T^{-1} transformation is not loosely Bernoulli. Ann Math 115:393–409
34. Kalikow S (1984) Twofold mixing implies threefold mixing for rank one transformations. Ergodic Theory Dynam Syst 2:237–259
35. Kamae T (1982) A simple proof of the ergodic theorem using non-standard analysis. Israel J Math 42:284–290
36. Katok A (1980) Interval exchange transformations and some special flows are not mixing. Israel J Math 35:301–310
37. Katok A (1980) Smooth non-Bernoulli K-automorphisms. Invent Math 61:291–299
38. Katok A, Hasselblatt B (1995) Introduction to the Modern Theory of Dynamical Systems. Cambridge, Cambridge
39. Katznelson Y (1971) Ergodic automorphisms of $\mathbb{T}^n$ are Bernoulli shifts. Israel J Math 10:186–195
40. Katznelson Y, Weiss B (1982) A simple proof of some ergodic theorems. Israel J Math 42:291–296
41. Keane M, Smorodinsky M (1979) Bernoulli schemes of the same entropy are finitarily isomorphic. Ann Math 109:397–406
42. Keane MS, Petersen KE (2006) Nearly simultaneous proofs of the ergodic theorem and maximal ergodic theorem. In: Dynamics and Stochastics: Festschrift in Honor of M.S. Keane. Institute of Mathematical Statistics, pp 248–251, Bethesda MD
43. Kolmogorov AN (1958) New metric invariant of transitive dynamical systems and endomorphisms of Lebesgue spaces. Dokl Russ Acad Sci 119:861–864
44. Koopman BO (1931) Hamiltonian systems and Hilbert space. Proc Nat Acad Sci 17:315–218
45. Krámli A, Simányi N, Szász D (1991) The *K*-property of three billiard balls. Ann Math 133:37–72
46. Ledrappier F (1978) Un champ Markovien peut être d'entropie nulle et mélangeant. C R Acad Sci Paris, Sér A-B 287:561–563
47. Liverani C (2004) Decay of correlations. Ann Math 159:1275–1312
48. Masser DW (2004) Mixing and linear equations over groups in positive characteristic. Israel J Math 142:189–204
49. Masur H (1982) Interval exchange transformations and measured foliations. Ann Math 115:169–200
50. McCutcheon R, Quas A (2007) Generalized polynomials and mild mixing systems. Canad J Math (to appear)
51. Méla X, Petersen K (2005) Dynamical properties of the Pascal adic transformation. Ergodic Theory Dynam Syst 25:227–256
52. Newton D, Parry W (1966) On a factor automorphism of a normal dynamical system. Ann Math Statist 37:1528–1533
53. Ornstein DS (1970) Bernoulli shifts with the same entropy are isomorphic. Adv Math 4:337–352
54. Ornstein DS (1972) On the root problem in ergodic theory. In: Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), vol II: Probability theory. Univ. California Press, pp 347–356
55. Ornstein DS (1973) An example of a Kolmogorov automorphism that is not a Bernoulli shift. Adv Math 10:49–62
56. Ornstein DS (1973) A K-automorphism with no square root and Pinsker's conjecture. Adv Math 10:89–102

57. Ornstein DS (1973) A mixing transformation for which Pinsker's conjecture fails. Adv Math 10:103–123
58. Ornstein DS (1974) Ergodic Theory, Randomness, and Dynamical Systems. Yale University Press, Newhaven
59. Ornstein DS, Shields PC (1973) An uncountable family of K-automorphisms. Adv Math 10:89–102
60. Ornstein DS, Weiss B (1973) Geodesic flows are Bernoullian. Israel J Math 14:184–198
61. Ornstein DS, Weiss B (1975) Unilateral codings of Bernoulli systems. Israel J Math 21:159–166
62. Oxtoby JC (1952) Ergodic sets. Bull Amer Math Soc 58:116–136
63. Parry W (1981) Topics in Ergodic Theory. Cambridge, Cambridge
64. Petersen K (1983) Ergodic Theory. Cambridge, Cambridge
65. Petersen K, Schmidt K (1997) Symmetric Gibbs measures. Trans Amer Math Soc 349:2775–2811
66. Phelps R (1966) Lectures on Choquet's Theorem. Van Nostrand, New York
67. Pinsker MS (1960) Dynamical systems with completely positive or zero entropy. Soviet Math Dokl 1:937–938
68. Quas A (1996) A C^1 expanding map of the circle which is not weak-mixing. Israel J Math 93:359–372
69. Quas A (1996) Non-ergodicity for C^1 expanding maps and g-measures. Ergodic Theory Dynam Systems 16:531–543
70. Rényi A (1957) Representations for real numbers and their ergodic properties. Acta Math Acad Sci Hungar 8:477–493
71. Riesz F (1938) Some mean ergodic theorems. J Lond Math Soc 13:274–278
72. Rokhlin VA (1948) A 'general' measure-preserving transformation is not mixing. Dokl Akad Nauk SSSR Ser Mat 60:349–351
73. Rokhlin VA (1949) On endomorphisms of compact commutative groups. Izvestiya Akad Nauk SSSR Ser Mat 13:329–340
74. Rokhlin VA, Sinai Y (1961) Construction and properties of invariant measurable partitions. Dokl Akad Nauk SSSR 141:1038–1041
75. Rudin W (1966) Real and Complex Analysis. McGraw Hill, New York
76. Rudolph DJ (1990) Fundamentals of Measurable Dynamics. Oxford, Oxford University Press
77. Ryzhikov VV (1993) Joinings and multiple mixing of the actions of finite rank. Funct Anal Appl 27:128–140
78. Shields P, Thouvenot J-P (1975) Entropy zero × Bernoulli processes are closed in the $\bar{d}$-metric. Ann Probab 3:732–736
79. Simányi N (2003) Proof of the Boltzmann–Sinai ergodic hypothesis for typical hard disk systems. Invent Math 154:123–178
80. Simányi N (2004) Proof of the ergodic hypothesis for typical hard ball systems. Ann Henri Poincaré 5:203–233
81. Simányi N, Szász D (1999) Hard ball systems are completely hyperbolic. Ann Math 149:35–96
82. Sinai YG (1959) On the notion of entropy of a dynamical system. Dokl Russ Acad Sci 124:768–771
83. Sinai YG (1964) On a weak isomorphism of transformations with invariant measure. Mat Sb (NS) 63:23–42
84. Sinai YG (1970) Dynamical systems with elastic reflections. Ergodic properties of dispersing billiards. Uspehi Mat Nauk 25:141–192
85. Sinai Y (1976) Introduction to Ergodic Theory. Princeton, Princeton, translation of the 1973 Russian original
86. Sinai YG, Chernov NI (1987) Ergodic properties of some systems of two-dimensional disks and three-dimensional balls. Uspekhi Mat Nauk 42:153–174, 256
87. Smorodinsky M (1971) A partition on a Bernoulli shift which is not weakly Bernoulli. Math Systems Th 5:201–203
88. Veech W (1982) Gauss measures for transformations on the space on interval exchange maps. Ann Math 115:201–242
89. Vershik A (1974) A description of invariant measures for actions of certain infinite-dimensional groups. Soviet Math Dokl 15:1396–1400
90. Vershik A (1981) Uniform algebraic approximation of shift and multiplication operators. Soviet Math Dokl 24:97–100
91. von Neumann J (1932) Proof of the quasi-ergodic hypothesis. Proc Nat Acad Sci USA 18:70–82
92. Walters P (1982) An Introduction to Ergodic Theory. Springer, Berlin
93. Williams D (1991) Probability with Martingales. Cambridge, Cambridge
94. Young LS (1995) Ergodic theory of differentiable dynamical systems. In: Real and Complex Dynamical Systems (Hillerød, 1993). Kluwer, Dordrecht, pp 293–336
95. Young LS (1998) Statistical properties of dynamical systems with some hyperbolicity. Ann Math 147:585–650
96. Yuzvinskii SA (1967) Metric automorphisms with a simple spectrum. Soviet Math Dokl 8:243–245

Ergodic Theorems

Andrés del Junco
Department of Mathematics, University of Toronto, Toronto, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
Ergodic Theorems for Measure-Preserving Maps
Generalizations to Continuous Time and Higher–Dimensional Time
Pointwise Ergodic Theorems for Operators
Subadditive and Multiplicative Ergodic Theorems
Entropy and the Shannon–McMillan–Breiman Theorem
Amenable Groups
Subsequence and Weighted Theorems
Ergodic Theorems and Multiple Recurrence
Rates of Convergence
Ergodic Theorems for Non-amenable Groups
Future Directions
Bibliography

Glossary

Dynamical system in its broadest sense, any set X, with a map $T\colon X \to X$. The classical example is: X is a set

whose points are the states of some physical system and the state x is succeeded by the state Tx after one unit of time.

Iteration repeated applications of the map T above to arrive at the state of the system after n units of time.

Orbit of x the forward images $x, Tx, T^2X \ldots$ of $x \in X$ under iteration of T. When T is invertible one may consider the forward, backward or two-sided orbit of x.

Automorphism a dynamical system $T\colon X \to X$, where X is a measure space and T is an invertible map preserving measure.

Ergodic average if f is a function on X let $A_n f(x) = n^{-1} \sum_{i=0}^{n-1} f(T^i x)$; the average of the values of f over the first n points in the orbit of x.

Ergodic theorem an assertion that ergodic averages converge in some sense.

Mean ergodic theorem an assertion that ergodic averages converge with respect to some norm on a space of functions.

Pointwise ergodic theorem an assertion that ergodic averages $A_n f(x)$ converge for some or all $x \in X$, usually for a.e. x.

Stationary process a sequence $(X_1, X_2, \ldots)$ of random variables (real or complex-valued measurable functions) on a probability space whose joint distributions are invariant under shifting $(X_1, X_2, \ldots)$ to $(X_2, X_3, \ldots)$.

Uniform distribution a sequence $\{x_n\}$ in $[0, 1]$ is uniformly distributed if for each interval $I \subset [0, 1]$, the time it spends in I is asymptotically proportional to the length of I.

Maximal inequality an inequality which allows one to bound the pointwise oscillation of a sequence of functions. An essential tool for proving pointwise ergodic theorems.

Operator any linear operator U on a vector space of functions on X, for example one arising from a dynamical system T by setting $Uf(x) = f(Tx)$. More generally any linear transformation on a real or complex vector space.

Positive contraction an operator T on a space of functions endowed with a norm $\|\cdot\|$ such that T maps positive functions to positive functions and $\|Tf\| \le |f|$.

Definition of the Subject

Ergodic theorems are assertions about the long-term statistical behavior of a dynamical system. The subject arose out of Boltzmann's ergodic hypothesis which sought to equate the spatial average of a function over the set of states in a physical system having a fixed energy with the time average of the function observed by starting with a particular state and following its evolution over a long time period.

Introduction

Suppose that $(X, \mathcal{B}, \mu)$ is a measure space and $T\colon (X, \mathcal{B}, \mu) \to (X, \mathcal{B}, \mu)$ is a measurable and measure-preserving transformation, that is $\mu(T^{-1}E) = \mu(E)$ for all $E \in \mathcal{B}$. One important motivation for studying such maps is that a Hamiltonian physical system (see the article by Petersen in this collection) gives rise to a one-parameter group $\{T_t : t \in \mathbb{R}\}$ of maps in the phase space of the system which preserve Lebesgue measure. The ergodic theorem of Birkhoff asserts that for $f \in L_1(\mu)$

$$\frac{1}{n} \sum_{i=0}^{n-1} f(T^i x) \tag{1}$$

converges a.e. and that if T is *ergodic* (to be defined shortly) then the limit is $\int f \mathrm{d}\mu$. This may be viewed as a justification for Boltzmann's ergodic hypothesis that "space averages equal time averages". See Zund [214] for some history of the ergodic hypothesis. For physicists, then, the problem is reduced to showing that a given physical system is ergodic, which can be very difficult. However ergodic systems arise in many natural ways in mathematics. One example is rotation $z \mapsto \lambda z$ of the unit circle if λ is a complex number of modulus one which is not a root of unity. Another is the shift transformation on a sequence of i.i.d. random variables, for example a coin-tossing sequence. Another is an automorphism of a compact Abelian group. Often a transformation possesses an invariant measure which is not obvious at first sight. Knowledge of such a measure can be a very useful tool. See Petersen's article for more examples.

If $(X, \mathcal{B}, \mu)$ is a probability space and T is ergodic then Birkhoff's ergodic theorem implies that if A is a measurable subset of X then for almost every x, the frequency with which x visits A is asymptotically equal to $\mu(A)$, a very satisfying justification of intuition. For example, applying this to the coin-tossing sequence one obtains the *strong law of large numbers* which asserts that almost every infinite sequence of coin tosses has tails occurring with asymptotic frequency $\frac{1}{2}$. One also obtains Borel's theorem on normal numbers which asserts that for almost all $x \in [0, 1]$ each digit $0, 1, 2, \ldots, 9$ occurs with limiting frequency $\frac{1}{10}$. The so-called *continued fraction transformation* $x \mapsto x^{-1} \mod 1$ on $(0, 1)$ has a finite invariant measure $\frac{dx}{1+x}$. (Throughout this article $x \mod 1$ denotes the

fractional part of x.) Applying Birkhoff's theorem then gives precise information about the frequency of occurrence of any $n \in \mathbb{N}$ in the continued fraction expansion of x, for a.e. x. See for example Billingsley [34].

These are the classical roots of the subject of ergodic theorems. The subject has evolved from these simple origins into a vast field in its own right, quite independent of physics or probability theory. Nonetheless it still has close ties to both these areas and has also forged new links with many other areas of mathematics.

Our purpose here is to give a broad overview of the subject in a historical perspective. There are several excellent references, notably the books of Krengel [135] and Tempelman [194] which give a good picture of the state of the subject at the time they appeared. There has been tremendous progress since then. The time is ripe for a much more comprehensive survey of the field than is possible here.

Many topics are necessarily absent and many are only glimpsed. For example this article will not touch on random ergodic theorems. See the articles [75,143] for some references on this topic.

I thank Mustafa Akcoglu, Ulrich Krengel, Michael Lin, Dan Rudolph and particularly Joe Rosenblatt and Vitaly Bergelson for many helpful comments and suggestions. I would like to dedicate this article to Mustafa Akcoglu who has been such an important contributor to the development of ergodic theorems over the past 40 years. He has also played a vital role in my mathematical development as well as of many other mathematicians. He remains a source of inspiration to me, and a valued friend.

Ergodic Theorems for Measure-Preserving Maps

Suppose $(X, \mathcal{B}, \mu)$ is a measure space. A (measure-preserving) *endomorphism* of $(X, \mathcal{B}, \mu)$ is a measurable mapping $T\colon X \to X$ such that $\mu(T^{-1}E) = \mu(E)$ for any measurable subset $E \subset X$. If T has a measurable inverse then one says that it is an *automorphism*. In this article the unqualified terms "endomorphism" or "automorphism" will always mean measure-preserving. T is called *ergodic* if for all measurable E one has

$$T^{-1}E = E \Rightarrow \mu(E) = 0 \quad \text{or} \quad \mu(E^c) = 0\,.$$

A very general class of examples comes from the notion of a *stationary stochastic process* in probability theory. A stochastic process is a sequence of measurable functions $f_1, f_2, \ldots$ on a probability space $(X, \mathcal{B}, \mu)$ taking values in a measurable space (Y, C). The *distribution* of the process, a measure ν on $Y^{\mathbb{N}}$, is defined as the image of μ under the map $(f_1, f_2, \ldots) : X \to Y^{\mathbb{N}}$. ν captures all the essential information about the process $\{f_i\}$. In effect one may view any process $\{f_i\}$ as a probability measure on $Y^{\mathbb{N}}$. The process is said to be *stationary* if ν is invariant under the left shift transformation S on $Y^{\mathbb{N}}$, $S(y)(i) = y(i+1)$, that is S is an endomorphism of the probability space $(Y^{\mathbb{N}}, \nu)$.

From a probabilistic point of view the most natural examples of stationary stochastic processes are independent identically distributed processes, namely the case when ν is a product measure $\lambda^{\mathbb{N}}$ for some measure λ on (Y, C). More generally one can consider a stationary Markov process defined by transition probabilities on the state space Y and an invariant probability on Y. See for example Chap. 7 in [50], also Sect. "Pointwise Ergodic Theorems for Operators" below.

The first and most fundamental result about endomorphisms is the celebrated recurrence theorem of Poincaré [173]).

Theorem 1 *Suppose μ is finite, $A \in \mathcal{B}$ and $\mu(A) > 0$. Then for a.e. $x \in A$ there is an $n > 0$ such that $T^n x \in A$, in fact there are infinitely many such n.*

It may be viewed as an ergodic theorem, in that it is a qualitative statement about how x behaves under iteration of T. For a proof, observe that if $E \subset A$ is the measurable set of points which never return to A then for each $n > 0$ the set $T^{-n}E$ is disjoint from E. Applying T^{-m} one gets also $T^{-(n+m)}E \cap T^{-m}E = \emptyset$. Thus $E, T^{-1}E, T^{-2}E, \ldots$ is an infinite sequence of disjoint sets all having measure $\mu(E)$ so $\mu(E) = 0$ since $\mu(X) < \infty$. Much later Kac [120] formulated the following quantitative version of Poincaré's theorem.

Theorem 2 *Suppose that $\mu(X) = 1$, T is ergodic and let $r_A x$ denote the time of first return to A, that is $r_A(x)$ is the least $n > 0$ such that $T^n x \in A$. Then*

$$\frac{1}{\mu(A)} \int_A r_A \, \mathrm{d}\mu = \frac{1}{\mu(A)}\,, \tag{2}$$

that is, the expected value of the return time to A is $\mu(A)^{-1}$.

Koopman [131] made the observation that associated to an automorphism T there is a unitary operator $U = U_T$ defined on the Hilbert space $L_2(\mu)$ by the formula $Uf = f \circ T$. This led von Neumann [151] to prove his mean ergodic theorem.

Theorem 3 *Suppose $\mathcal{H}$ is a Hilbert space, U is a unitary operator on $\mathcal{H}$ and let P denote the orthogonal projection on the subspace of U-invariant vectors. Then for any $x \in \mathcal{H}$ one has*

$$\left\| \frac{1}{n} \sum_{i=0}^{n-1} U^i x - Px \right\| \to 0 \quad \text{as } n \to \infty\,. \tag{3}$$

Von Neumann's theorem is usually quoted as above but to be historically accurate he dealt with unitary operators indexed by a continuous time parameter.

Inspired by von Neumann's theorem Birkhoff very soon proved his pointwise ergodic theorem [35]. In spite of the later publication by von Neumann his result did come first. See [29,214] for an interesting discussion of the history of the two theorems and of the interaction between Birkhoff and von Neumann.

Theorem 4 *Suppose $(X, \mathcal{B}, \mu)$ is a measure space and T is an endomorphism of $(X, \mathcal{B}, \mu)$. Then for any $f \in L_1 = L_1(\mu)$ there is a T-invariant function $g \in L_1$ such that*

$$A_n f(x) = \frac{1}{n} \sum_{i=0}^{n-1} f(T^i x) \to g(x) \quad a.e. \tag{4}$$

Moreover if μ is finite then the convergence also holds with respect to the L_1 norm and one has $\int_E g \, \mathrm{d}\mu = \int_E f d\mu$ for all T-invariant subsets E.

Again this formulation of Birkhoff's theorem is not historically accurate as he dealt with a smooth flow on a manifold. It was soon observed that the theorem, and its proof, remain valid for an abstract automorphism of a measure space although the realization that T need not be invertible seems to have taken a little longer.

The notation $A_n f = A_n(T) f$ as above will occur often in the sequel. Whenever T is an endomorphism one uses the notation $Tf = f \circ T$ and with this notation $A_n(T) = \frac{1}{n} \sum_{i=0}^{n-1} T^i$. When the scalars ($\mathbb{R}$ or $\mathbb{C}$) for an L_1 space are not specified the notation should be understood as referring to either possibility. In most of the theorems in this article the complex case follows easily from the real and any indications about proofs will refer to the real case.

Although von Neumann originally used spectral theory to prove his result, there is a quick proof, attributed to Riesz by Hopf in his 1937 book [100], which uses only elementary properties of Hilbert space. Let I denote the (closed) subspace of U-invariant vectors and I' the (usually not closed) subspace of vectors of the form $f - Uf$. It is easy to check that any vector orthogonal to I' must be in I, whence the subspace $I + I'$ is dense in $\mathcal{H}$. For any vector of the form $x = y + y'$, $y \in I$, $y' \in I'$ it is clear that $A_n x = \frac{1}{n} \sum_{i=0}^{n-1} U^i x$ converges to $y = Px$, since $A_n y = y$ and if $y' = z - Uz$ then the telescoping sum $A_n y' = n^{-1}(z - U^n z)$ converges to 0. This establishes the desired convergence for $x \in I + I'$ and it is easy to extend it to the closure of $I + I'$ since the operators A_n are *contractions* of $\mathcal{H}$ ($\|A_n\| \le 1$). Lorch [147] used a soft argument in a similar spirit to extend non Neumann's theorem from the case of a unitary operator on a Hilbert space to that of an arbitrary linear contraction on any reflexive Banach space. Sine [188] gave a necessary and sufficient condition for the strong convergence of the ergodic averages of a contraction on an arbitrary Banach space.

Birkhoff's theorem has the distinction of being one of the most reproved theorems of twentieth century mathematics. One approach to the pointwise convergence, parallel to the argument to argument just seen, is to find a dense subspace E of L_1 so that the convergence holds for all $f \in E$ and then try to extend the convergence to all $f \in L_1$ by an approximation argument. The first step is not too hard. For simplicity assume that μ is a finite measure. As in the proof of von Neumann's theorem the subspace E spanned by the T-invariant L_1 functions together with functions of the form $g - Tg$, $g \in L_1$, is dense in $L_1(\mu)$. This can be seen by using the Hahn–Banach theorem and the duality of L_1 and L_∞. (Here one needs finiteness of μ to know that $L_\infty \subset L_1$.) The pointwise convergence of $A_n f$ for invariant f is trivial and for $f = g - Tg$ it follows from telescoping of the sum and the fact that $n^{-1} T^n g \to 0$ a.e. This last can be shown by using the Borel–Cantelli lemma. The second step, extending pointwise convergence, as opposed to norm convergence, for f in a dense subspace to all f in L_1 is a delicate matter, requiring a *maximal inequality*.

Roughly speaking a maximal inequality is an inequality which bounds the pointwise oscillation of $A_n f$ in terms of the norm of f. The now standard maximal inequality in the context of Birkhoff's theorem is the following, due to Kakutani and Yosida [209]. Birkhoff's proof of his theorem includes a weaker version of this result. Let $S_n f = n A_n f$, the nth partial sum of the iterates of f.

Theorem 5 *Given any real $f \in L_1$ let $A = \bigcup_{n \ge 1} \{S_n f \ge 0\}$. Then*

$$\int_A f \, \mathrm{d}\mu \ge 0 \,. \tag{5}$$

Moreover if one sets $Mf = \sup_{n \ge 1} A_n f$ then for any $\alpha > 0$

$$\mu\{Mf > \alpha\} \le \frac{1}{\alpha} \|f\|_1 \tag{6}$$

A distributional inequality such as (6) will be referred to as a *weak* L_1 inequality. Note that (6) follows easily from (5) by applying (5) to $f - \alpha$, at least in the case when μ is finite. With the maximal inequality in hand it is straightforward to complete the proof of Birkhoff's theorem. For a real-valued function f let

$$\operatorname{Osc} f = \limsup A_n f - \liminf A_n f \,. \tag{7}$$

$\operatorname{Osc} f = 0$ a.e. if and only if $A_n f$ converges a.e.(to a possibly infinite limit). One has $\limsup A_n f \le Mf \le M|f|$ and by symmetry $\liminf A_n f \ge M|f|$, so $\operatorname{Osc} f \le 2M|f|$. To establish the convergence of $A_n f$ for a real-valued $f \in L_1$ let $\epsilon > 0$ and write $f = g + h$ with $g \in E$ (the subspace where convergence has already been established), $h \in L^1$ and $\|h\| < \epsilon$. Then since $\operatorname{Osc} g = 0$ one has $\operatorname{Osc} f = \operatorname{Osc} h$. Thus for any fixed $\alpha > 0$, using (6)

$$\mu\{\operatorname{Osc} f > \alpha\} = \mu\{\operatorname{Osc} h > \alpha\} \le \mu\left\{Mh > \frac{\alpha}{2}\right\} \le \frac{2\|h\|_1}{\alpha} < \frac{2\epsilon}{\alpha} . \quad (8)$$

Since $\epsilon > 0$ was arbitrary one concludes that $\mu\{\operatorname{Osc} f > \alpha\} = 0$ and since $\alpha > 0$ is arbitrary it follows that $\mu\{\operatorname{Osc} f > 0\} = 0$, establishing the a.e. convergence. Moreover a simple application of Fatou's lemma shows that the limiting function is in L_1, hence finite a.e.

There are many proofs of (5). Two of particular interest are Garsia's [89], perhaps the shortest and most mysterious, and the proof via the filling scheme of Chacón and Ornstein [61], perhaps the most intuitive, which goes like this. Given a function $g \in L_1$ write $g^+ = \max(g, 0)$, $g^- = g^+ - g$ and let $Ug = Tg^+ - g^-$. Interpretation: the region between the graph of g^+ and the X-axis is a sheaf of vertical spaghetti sticks, the intervals $[0, g^+(x)]$, x in X, and $-g^-$ is a hole. Now move the spaghetti (horizontally) by T and then let it drop (vertically) into the hole leaving a new hole and a new sheaf which are the negative and positive parts of Ug. Now let $E' = \bigcup_{n\ge1}\{U^n f \ge 0\}$, the set of points x at which the hole is eventually filled after finitely many iterations of U.

The key point is that $E = E'$. Indeed if $S_n f(x) \ge 0$ for some n, then the total linear height of sticks over $x, Tx, \ldots T^{n-1}x$ is greater than the total linear depth of holes at these points. The only way that spaghetti can escape from these points is by first filling the hole at x, which shows $x \in E'$. Similar thinking shows that if $x \in E'$ and the hole at x is filled *for the first time* at time n then $S_n f(x) \ge 0$, so $x \in E$, and that all the spaghetti that goes into the hole at x comes from points $T^i x$ which belong to E'. This shows that $E = E'$ and that the part of the hole lying beneath E is eventually filled by spaghetti coming from $E' = E$. Thus the amount of spaghetti over E is no less than the size of the hole under E, that is $\int_E f \, d\mu \ge 0$.

Most proofs of Birkhoff's theorem use a maximal inequality in some form but a few avoid it altogether, for example [126,186]. It is also straightforward to deduce Birkhoff's theorem directly from a maximal inequality, as Birkhoff does, without first establishing convergence on a dense subspace. However the technique of proving a pointwise convergence theorem by finding an appropriate dense subspace and a suitable maximal inequality has proved extremely useful, not only in ergodic theory.

Indeed, in some sense maximal inequalities are unavoidable: this is the content of the following principle proved already in 1926 by Banach. The principle has many formulations; the following one is a slight simplification of the one to be found in [135]. Suppose B is a Banach space, $(X, \mathcal{B}, \mu)$ is a finite measure space and let E denote the space of μ-equivalence classes of measurable real-valued functions on X. A linear map $T\colon B \to E$ is said to be *continuous in measure* if for each $\epsilon > 0$

$$\|x_n - x\| \to 0 \Rightarrow \mu\{|Tx_n - Tx| > \epsilon\} \to 0 .$$

Suppose that T_n is a sequence of linear maps from B to E which are continuous in measure and let $Mx = \sup_n |T_n x|$. Of course if $T_n x$ converges a.e. to a finite limit then $Mx < \infty$ a.e.

Theorem 6 (Banach principle) *Suppose $Mx < \infty$ a.e. for each $x \in X$. Then there is a function $C(\lambda)$ such that $C(\lambda) \to 0$ as $\lambda \to \infty$ such that for all $x \in B$ and $\lambda > 0$ one has*

$$\mu\{Mx \ge \lambda\|x\|\} \le C(\lambda) . \quad (9)$$

Moreover the set of x for which $T_n x$ converges a.e. is closed in B.

The first chapter of Garsia's book [89] contains a nice introduction to the Banach principle.

It should be noted that, for general integrable f, Mf need not be in L^1. However if $f \in L^p$ for $p > 1$ then Mf does belong to L^p and one has the estimate

$$\|Mf\|_p \le \frac{p}{p-1}\|f\|_p . \quad (10)$$

This can be derived from (6) using also the (obvious) fact that $\|Mf\|_\infty \le \|f\|_\infty$. Any such estimate on the norm of a maximal function will be called a *strong L_p* inequality. It also follows from (6) that if $\mu(X)$ is finite and $f \in L \log L$, that is $\int |f| \log^+ |f| d\mu < \infty$, then $Mf \in L^1$. In fact Ornstein [160] has shown that the converse of this last statement holds provided T is ergodic.

There is a special setting where one has uniform convergence in the ergodic theorem. Suppose T is a homeomorphism of a compact metric space X. By a theorem of Krylov and Bogoliouboff [138] there is at least one probability measure on the Borel σ-algebra of X which is invariant under T. T is said to be *uniquely ergodic* if there is only one Borel probability measure, say μ, invariant under T. It is easy to see that when this is the

case then T is an ergodic automorphism of $(X, \mathcal{B}, \mu)$. As an example, if α is an irrational number then the rotation $z \mapsto e^{2\pi i\alpha} z$ is a uniquely ergodic transformation of the circle $\{|z| = 1\}$. Equivalently $x \mapsto x + \alpha \mod 1$ is a uniquely ergodic map on $[0, 1]$. A quick way to see this is to show that the Fourier co-efficients $\hat{\mu}(n)$ of any invariant probability μ are zero for $n \neq 0$. The Jewett–Krieger theorem (see Jewett [109] and Krieger [137]) guarantees that unique ergodicity is ubiquitous in the sense that any automorphism of a probability space is measure-theoretically isomorphic to a uniquely ergodic homeomorphism. The following important result is due to Oxtoby [165]).

Theorem 7 *If T is uniquely ergodic, μ is its unique invariant probability measure and $f \in C(X)$ then the ergodic averages $A_n(f)$ converge uniformly to $\int f d\mu$.*

This result can be proved along the same lines as the proof given above of von Neumann's theorem. In a nutshell, one uses the fact that the dual of $C_{\mathbb{R}}(X)$ is the space of finite signed measures on X and the unique ergodicity to show that functions of the form $f - f \circ T$ together with the invariant functions (which are just constant functions) span a dense subspace of $C(X)$.

A sequence $\{x_n\}$ in the interval $[0, 1]$ is said to be *uniformly distributed* if $\frac{1}{n}\sum_{i=1}^{n} f(x_n) \to \int_0^1 f(x)\mathrm{d}x$ for any $f \in C[0, 1]$ (equivalently for any Riemann integrable f or for any $f = 1_I$ where I is any subinterval of $[0, 1]$). As a simple application of Theorem 7 one obtains the linear case of the following result of Weyl [202].

Theorem 8 *If α is irrational and $p(x)$ is any non-constant polynomial with integer coefficients then the sequence $\{p(n)\alpha \mathrm{mod} 1\}$ is uniformly distributed.*

Furstenberg [84], see also [86], has shown that Weyl's result, in full generality, can be deduced from the unique ergodicity of certain affine transformations of higher dimensional tori. For polynomials of degree $k > 1$ Weyl's result is usually proved by inductively reducing to the case $k = 1$, using the following important lemma of van der Corput (see for example [139]).

Theorem 9 *Suppose that for each fixed $h > 0$ the sequence*

$$\{x_{n+h} - x_n \mod 1\}_n$$

is uniformly distributed. Then $\{x_n\}$ is uniformly distributed.

When μ is infinite and T is ergodic the limiting function in Birkhoff's theorem is 0 a.e. In 1937 Hopf [100] proved a generalization of Birkhoff's theorem which is more meaningful in the case of an infinite invariant measure. It is a special case of a later theorem of Hurewicz [105], which we will discuss first.

Suppose that $(X, \mathcal{B}, \mu)$ is a σ-finite measure space. If ν is another σ-finite measure on $\mathcal{B}$ write $\nu \ll \mu$ (ν is *absolutely continuous* relative to μ) if $\mu(E) = 0$ implies $\nu(E) = 0$ and one writes $\nu \sim \mu$ if $\nu \ll \mu$ and $\mu \ll \nu$. Consider a *non-singular automorphism* $\tau : X \to X$, meaning that τ is measurable with a measurable inverse and that $\mu(E) = 0$ if and only if $\mu(\tau E) = 0$. In other words $\nu = \mu \circ \tau \sim \mu$. By the Radon–Nikodym theorem there is a function $\rho \in L_1(\mu)$ such that $\rho > 0$ a.e. and $\nu(E) = \int_E \rho \mathrm{d}\mu$ for all measurable E. In order to obtain an associated operator T on L_1 which is an (invertible) isometry one defines

$$Tf(x) = \rho(x) f(\tau x) . \tag{11}$$

The dual operator on L_∞ is then given by $T^* g = g \circ \tau^{-1}$.

If ν is a σ-finite measure equivalent to μ which is invariant under τ then the ergodic theory of τ can be reduced to the measure-preserving case using σ. The interesting case is when there is no such ν. It was an open problem for some time whether there is always an equivalent invariant measure. In 1960 Ornstein [163] gave an example of a τ which does not have an equivalent invariant measure. It is curious that, with hindsight, such examples were already known in the fifties to people studying von Neumann algebras.

For $f \in L_1$ let $S_n f = \sum_{i=0}^{n} T^i f$. τ is said to be *conservative* if there is no set E with $\mu(E) > 0$ such that $\tau^{-i} E$, $i = 0, 1, \ldots$ are pairwise disjoint, that is if the Poincaré recurrence theorem remains valid. For example, the shift on $\mathbb{Z}$ is not conservative. Hurewicz [105] proved the following ratio ergodic theorem.

Theorem 10 *Suppose τ is conservative, $f, g \in L_1$ and $g(x) > 0$ a.e. Then $S_n f / S_n g$ converges a.e. to a τ-invariant limit h. If τ is ergodic then $h = \frac{\int f d\mu}{\int g d\mu}$.*

In the case when μ is τ-invariant one has $Tf = f \circ \tau$. If μ is invariant and finite, taking $g = 1$ one recovers Birkhoff's theorem. If μ is invariant and σ-finite then Hurewicz's theorem becomes the theorem of Hopf alluded to earlier.

Wiener and Wintner [204] proved the following variant of Birkhoff's theorem.

Theorem 11 (Wiener–Wintner) *Suppose T is an automorphism of a probability space $(X, \mathcal{B}, \mu)$. Then for any $f \in L_1$ there is a subset $X' \subset X$ of measure one such that for each $x \in X'$ and $\lambda \in \mathbb{C}$ of modulus 1 the sequence $\frac{1}{n}\sum_{i=0}^{n-1} \lambda^i T^i f(x)$ converges.*

It is an easy consequence of the ergodic theorem that one has a.e. convergence for a given f and λ but the point here

is that the set on which the convergence occurs is independent of λ.

Generalizations to Continuous Time and Higher-Dimensional Time

A (measure-preserving) flow is a one-parameter group $\{T_t, t \in \mathbb{R}\}$ of automorphisms of $(X, \mathcal{B}, \mu)$, that is $T_{t+s} = T_t T_s$, such that $T_t x$ is measurable as a function of (t, x). It will always be implicitly assumed that the map $(t, x) \mapsto T_t x$ from $\mathbb{R} \times X$ to X is measurable. Theorem 4 generalizes to flows by replacing sums with integrals and this generalization follows without difficulty from Theorem 4. (As already observed this observation reverses the historical record.) Theorem 4 may be viewed as a theorem about the "discrete flow" $\{T_n = T^n : n \in \mathbb{Z}\}$. Wiener was the first to generalize Birkhoff's theorem to families of automorphisms $\{T_g\}$ indexed by groups more general than $\mathbb{R}$ or $\mathbb{Z}$.

A measure-preserving flow is an *action* of $\mathbb{R}$ while a single automorphism corresponds to an action of $\mathbb{Z}$. A (measure-preserving) action of a group G is a homomorphism $T: g \mapsto T_g$ from G into the group of automorphisms of a measure space $(X, \mathcal{B}, \mu)$ (satisfying the appropriate joint measurability condition in case G is not discrete). Suppose now that $G = \mathbb{R}^k$ or $\mathbb{Z}^k$ and T is an action of G on $(X, \mathcal{B}, \mu)$. In the case of $\mathbb{Z}^k$ an action amounts to an arbitrary choice of commuting maps $T_1, \ldots T_k$, $T_i = T(e_i)$ where e_i is the standard basis of $\mathbb{Z}^k$. In the case of $\mathbb{R}^k$ one must specify k commuting flows.

Let m denote counting measure on G in case $G = \mathbb{Z}^k$ and Lebesgue measure in case $G = \mathbb{R}^k$. For any subset E of G with $m(E) < \infty$ let

$$A_E f(x) = \frac{1}{m(E)} \int_E f(T_g x) \mathrm{d}m(g) \,. \tag{12}$$

One may then ask whether $A_E f$ converges to a limit, either in the mean or pointwise, as E varies through some sequence of sets which "grow large" or, in case $G = \mathbb{R}^k$, "shrink to 0". The second case is referred to as a *local ergodic theorem*. In the case of ergodic theorems at infinity the continuous and discrete theories are rather similar and often the continuous analogue of a discrete result can be deduced from the discrete result.

In Wiener [203] proved the following result for actions of $G = \mathbb{R}^k$ and ergodic averages over Euclidean balls $B_r = \{x \in \mathbb{R}^k : \|x\|_2 \le r\}$.

Theorem 12 *Suppose T is an action of $\mathbb{R}^k$ on $(X, \mathcal{B}, \mu)$ and $f \in L_1(\mu)$.*

(a) *For $f \in L_1$ $\lim_{r\to\infty} A_{B_r} f = g$ exists a.e. If μ is finite the convergence also holds with respect to the L_1-norm, g is T-invariant and $\int_I g d\mu = \int_I f \,\mathrm{d}\mu$ for every T-invariant set I.*

(b) *$\lim_{r\to 0} A_{B_r} f = f$ a.e.*

The local aspect of Wiener's theorem is closely related to the Lebesgue differentiation theorem, see for example Proposition 3.5.4 of [132], which, in its simplest form, states that for $f \in L^1(\mathbb{R}^k, m)$ one has a.e. convergence of $\frac{1}{m(B_r)} \int_{B_r} f(x+t)\mathrm{d}t$ to $f(x)$ as $r \to 0$. The local ergodic theorem implies Lebesgue's theorem, simply by considering the action of $\mathbb{R}^k$ on itself by translation. In fact the local ergodic theorem can also be deduced from Lebesgue's theorem by a simple application of Fubini's theorem (see for example [135], Chap. 1, Theorem 2.4 in the case $k = 1$).

The key point in Wiener's proof is the following weak L_1 maximal inequality, similar to (6).

Theorem 13 *Let $Mf = \sup_{r>0} |A_{B_r} f|$. Then one has*

$$\mu\{Mf > \alpha\} \le \frac{C}{\alpha} \|f\|_1 \,, \tag{13}$$

where C is a constant depending only on the dimension d. (In fact one may take $C = 3^d$.)

In the case when T is the action of R^k on itself by translation (13) is the well-known maximal inequality for the Hardy–Littlewood maximal function ([132], Lemma 3.5.3). Wiener proves (13) by way of the following *covering lemma*. If B is a ball in $\mathbb{R}^k$ let B' denote the concentric ball with three times the radius.

Theorem 14 *Suppose a compact subset K of $\mathbb{R}^k$ is covered by a (finite) collection $\mathcal{U}$ of open balls. Then there exist pairwise disjoint $B_i \in \mathcal{U}$, $i = 1, \ldots k$ such that $\bigcup_i B_i'$ covers K.*

To find the B_i it suffices to let B_1 be the largest ball in $\mathcal{U}$, then B_2 the largest ball which does not intersect B_1 and in general B_n the largest ball which does not intersect $\bigcup_{i=1}^{n-1} B_i$. Then it is not hard to argue that the B_i' cover K.

In general, a covering lemma is, roughly speaking, an assertion that, given a cover $\mathcal{U}$ of a set K in some space (often a group), one may find a subcollection $\mathcal{U}' \subset \mathcal{U}$ which covers a substantial part of K and is in some sense efficient in that $\sum_{U \in \mathcal{U}'} 1_U \le C$, C some absolute constant. Covering lemmas play an important role in the proofs of many maximal inequalities. See Sect. 3.5 of [132] for a discussion of several of the best-known classical covering lemmas.

As Wiener was likely aware, the same kind of covering argument easily leads to maximal inequalities and ergodic theorems for averages over "sufficiently regular" sets. For example, in the case of $G = \mathbb{Z}^k$ use the standard total order $\prec$ on G and for $n \in \mathbb{N}^k$ let $S_n = \{m \in Z^k : 0 \prec m \prec$

$n\}$. Let $e(n)$ denote the maximum value of n_i/n_j. Then one has the following result.

Theorem 15 *For any $e > 0$ and any integrable f*

$$\lim_{n\to\infty, e(n)<e} A_{S_n} f$$

exists a.e. and the limit is characterized by the same properties as in Theorem 12. Moreover there is a maximal inequality

$$\mu\left\{\sup_{e(n)<e} |A_{S_n} f| > \alpha\right\} \le \frac{C}{\alpha}\|f\|_1 , \tag{14}$$

where C is a constant depending only on e and the dimension k.

If one does not impose a bound on $e(n)$ the a.e. convergence of $A_{S_n} f$ may fail for $f \in L_1$, (although it does hold for any increasing sequence of n's by Theorem 30 below). Nonetheless Dunford [72] and independently Zygmund [215] showed that one does have unrestricted convergence if $f \in L_p$ for some $p > 1$, provided μ is finite. Let $T_1, \ldots, T_k$ be any k (possibly non-commuting!) automorphisms of a finite measure space $(X, \mathcal{B}, \mu)$. For $n = (n_1, n_2, \ldots, n_k)$ let $T_n = T_1^{n_1} \ldots T_k^{n_k}$. (This is not an action of $\mathbb{Z}^k$ unless the T_i commute with each other.) As before when $F \subset \mathbb{Z}^k$ is a finite subset write $A_F f = \frac{1}{|F|}\sum_{n\in F} T_n f$. Finally let $P_j f = \lim_n A_n(T_j) f$.

Theorem 16 *For $f \in L_p$*

$$\lim_{n\to\infty} A_{S_n} f = P_1 \ldots P_k f \tag{15}$$

both a.e. and in L_p.

The proof of Theorem 16 uses repeated applications of Birkhoff's theorem and hinges on (10).

Somewhat surprisingly Hurewicz's theorem was not generalized to higher dimensions until the very recent work of Feldman [79]. In fact the theorem fails if one considers averages over the cube $[0, n-1]^d$ in $\mathbb{Z}^d$. However Feldman was able to prove a suitably formulated generalization of Hurewicz's theorem for averages over symmetric cubes.

For $f \in L^1(\mathbb{R})$ the classical *Hilbert transform* is defined for a.e. t by

$$Hf(t) = \frac{1}{\pi}\lim_{\epsilon\to 0^+}\int_{|s|>\epsilon} \frac{f(t-s)}{s} ds \tag{16}$$

Let $H^* f(t) = \frac{1}{\pi}\sup_{\epsilon>0} |\int_{|s|>\epsilon} \frac{f(t-s)}{s} ds|$, the corresponding maximal function. The proof that the limit (16) exists a.e. is based on the maximal inequality

$$m\{Hf > \lambda\} \le \frac{C}{\lambda}\|f\|_1 , \tag{17}$$

where C is an absolute constant. See Loomis [146] (1946) for a proof of (17) using real-variable methods. Cotlar [65] proved the existence of the following ergodic analogue of the Hilbert transform (actually for the n-dimensional Hilbert transform).

Theorem 17 *Suppose $\{T_t\}$ is a measure-preserving flow on a probability space $(X, \mathcal{B}, \mu)$ and $f \in L^1$. Then*

$$\lim_{\epsilon\to 0}\int_{\epsilon<t<\epsilon^{-1}} \frac{f(T_t x)}{t} dt \tag{18}$$

exists for a.a. x.

Motivated by Cotlar's result Calderón [54]proved a general *transfer principle* which allows one to transfer maximal inequalities and convergence theorems for functions of a real or integer variable to the ergodic setting. Although stated for functions on $\mathbb{R}$ it applies equally well to $\mathbb{R}^k$ or $\mathbb{Z}^k$. This principle subsumes Birkhoff's theorem, Wiener's theorem and Cotlar's result. For simplicity only a special case of Calderón's result, in the discrete case, will be stated here.

Let T be an automorphism of a probability space $(X, \mathcal{B}, \mu)$, let m denote counting measure on $\mathbb{Z}$ and suppose σ is a probability measure on $\mathbb{Z}$. For $g \in L_1(m)$ define $\sigma(g)(n) = \sum_{i\in\mathbb{Z}} \sigma(i) g(n+i)$. For $f \in L_1(\mu)$ let $\sigma(T) f = \sum_{i\in\mathbb{Z}} \sigma(i) T^i f$, which is easily seen to converge a.e. and in L_1-norm. Given a fixed sequence σ_n of probabilities define $Mg = \sup_n \sigma_n g$ and $M(T) f = \sup_n \sigma_n(T) f$.

Theorem 18 (Calderón) *Suppose there is a constant C such that*

$$m\{Mg > \lambda\} < \frac{C}{\lambda}\|g\|_1 \quad \textit{for all } g \in L_1(m) . \tag{19}$$

Then one also has

$$\mu\{M(T) f > \lambda\} < \frac{C}{\lambda}\|f\|_1 \quad \textit{for all } T \quad \textit{and } g \in L_1(\mu) . \tag{20}$$

In other words, in order to prove a maximal inequality for general T it suffices to prove it in case T is the shift map on the integers. The idea of transference could already be seen in Wiener's proof of the $\mathbb{R}^n$ ergodic theorem. Transfer principles in various forms of have become an important tool in the study of ergodic theorems. See Bellow [18] for a very readable overview.

Pointwise Ergodic Theorems for Operators

Early in the history of ergodic theorems there were attempts to generalize the ergodic theorem to more general

linear operators on L_p spaces, that is, operators which do not arise by composition with a mapping of X . In the case $p = 1$ the main motivation for this comes from the theory of Markov processes.

If $(X, \mathcal{B})$ is a measurable space a *sub-stochastic kernel* on X is a non-negative function P on $X \times \mathcal{B}$ such that

(a) for each $x \in X P(x, \cdot) = P_x$ is a measure on $\mathcal{B}$ such that $P_x(X) \leq 1$ and
(b) $P(\cdot, A)$ is a measurable function for each $A \in \mathcal{B}$.

It is most intuitive to think about the *stochastic* case, namely when each P_x is a probability measure. One then views $P(x, A)$ as the probability that the point x moves into the set A in one unit of time, so one has stochastic dynamics as opposed to deterministic dynamics, namely the case when $P_x = \delta_{Tx}$ for a map T. In this case the measures P_x are called *transition probabilities*.

If μ is a σ-finite measure on X one may define the measure $P\mu = \int P_x \mathrm{d}\mu(x)$. $P\nu$ is also meaningful if ν is a finite signed measure. The case when $P\mu = \mu$ is the stochastic analogue of measure-preserving dynamics and the case when $P\mu \ll \mu$ is the analogue of non-singular dynamics. It is easy to see that given any σ-finite measure λ there is always a μ such that $\lambda \ll \mu$ and $P\mu \ll \mu$. Let $\tilde{L}_1$ denote the space of finite signed measures ν such that $\nu \ll \mu$, which is identified with $L_1 = L_1(\mu, \mathbb{R})$ via the Radon–Nikodym theorem. If $P\mu \ll \mu$ then P maps $\tilde{L}_1(\mu)$ into itself so the restriction of P is an operator T on $L_1(\mu)$. T is a positive contraction, that is $\|T\| \leq 1$ and T maps non-negative functions to non-negative functions. As proved in, for example, [153], every positive contraction arises in this way from a substochastic kernel under the assumption that $(X, \mathcal{B})$ is standard. This simply means that there is some complete metric on X for which $\mathcal{B}$ is the σ-algebra of Borel sets. Virtually all measurable spaces encountered in analysis are standard so this should be viewed as a technicality only. See [153], [80] and [135] for more about the relation between kernels and positive contractions.

The case when X is finite, P is stochastic and μ is a a probability measure is classical in probability theory. P and μ determine a probability measure ν on $X^{\mathbb{N}}$ characterized by its values on cylinder sets, namely for all $x_1, \ldots, x_n \in X$

$$\nu\{x \in X^{\mathbb{N}} : x(i) = x_i, 1 \leq i \leq n-1\} = \mu(x_1) \prod_{i=1}^{n-1} P_{x_i}(x_{i+1}) . \tag{21}$$

The co-ordinate functions $X_i(x) = x(i)$ on the space $X^{\mathbb{N}}$ endowed with the probability ν form a *Markov process*, which will be stationary if and only if μ is P-invariant. For a general X the analogous construction is possible provided X is standard.

Hopf [101] initiated the systematic study of positive L_1 contractions and proved the following ergodic theorem.

Theorem 19 *Suppose $(X, \mathcal{B}, \mu)$ is a probability space and T is a positive contraction on $L^1(\mu)$ satisfying $T1 = 1$ and $T^*1 = 1$. Then $\lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} T^k f$ exists a.e.*

The importance of Hopf's article lies less in this convergence result than in the methods he developed. He proved that the maximal inequality (5) generalizes to all positive L_1 contractions T and used this to obtain the decomposition of X into its conservative and dissipative parts C and D, characterized by the fact that for any $p \in L_1$ such that $p > 0$ a.e. $\sum_{k=0}^{\infty} T^k p$ is infinite a.e. on C and finite a.e. on D. These results are the cornerstone of much of the subsequent work on L_1 contractions.

Theorem 19 contains Birkhoff's theorem for an automorphism τ of $(X, \mathcal{B}, \mu)$, simply by defining $Tf = f \circ \tau$. In fact one can also deduce Theorem 19 from Birkhoff's theorem (if one assumes only that X is standard). Indeed the hypotheses of Hopf's theorem imply that the kernel P associated to T is stochastic ($T^*1 = 1$) and and that μ is P-invariant ($T1 = 1$). Hopf's theorem then follows by applying Birkhoff's theorem to the shift on the stationary Markov process associated to P and μ. In fact Kakutani [123] (see also Doob [71]) had already made essentially the same observation, except that his result assumes the stationary Markov process to be already given.

In 1955 Dunford and Schwartz [73] made essential use of Hopf's work to prove the following result.

Theorem 20 *Suppose $\mu(X) = 1$ and T is a (not necessarily positive) contraction with respect to both the L_1 and L_∞ norms. Then the conclusion of Hopf's theorem remains valid.*

Note that the assumption that T contracts the L_∞-norm is meaningful, as $L_\infty \subset L_1$. The proof of the result is reduced to the positive case by defining a positive contraction $|T|$ analogous to the total variation of a complex measure.

Then in 1960 Chacón and Ornstein [61] proved a definitive ratio ergodic theorem for all positive contractions of L_1 which generalizes both Hopf's theorem and the Hurewicz theorem.

Theorem 21 *Suppose T is a positive contraction of $L_1(\mu)$, where μ is σ-finite, $f, g \in L_1$ and $g \geq 0$. Then*

$$\frac{\sum_{i=0}^{n} T^i f}{\sum_{i=0}^{n} T^i g} \tag{22}$$

converges a.e. on the set $\{\sum_{i=0}^{\infty} T^i g > 0\}$.

In 1963 Chacón [58] proved a very general theorem for non-positive operators which includes the Chacón–Ornstein theorem as well as the Dunford–Schwartz theorem.

Theorem 22 *Suppose T is a contraction of L_1 and $p_n \geq 0$ is a sequence of measurable functions with the property that*

$$g \in L_1, |g| \leq p_n \Rightarrow |Tg| \leq p_{n+1} . \tag{23}$$

Then

$$\frac{\sum_{i=0}^{n} T^i f}{\sum_{i=0}^{n} p_i} \tag{24}$$

converges a.e. to a finite limit on the set $\{\sum_{i=0}^{\infty} p_i > 0\}$.

If T is an L_1, L_∞-contraction and $p_n = 1$ for all n then the hypotheses of this theorem are satisfied, so Theorem 22 reduces to the result of Dunford and Schwartz. See [59] for a concise overview of all of the above theorems in this section and the relations between them.

The identification of the limit in the Chacón–Ornstein theorem on the conservative part C of X is a difficult problem. It was solved by Neveu [152] in case $C = X$, and in general by Chacón [57]. Chacón [60] has shown that there is a non-singular automorphism τ of $(X, \mathcal{B}, \mu)$ such that for the associated invertible isometry T of $L_1(\mu)$ given by (11) there is an $f \in L_1(\mu)$ such that $\limsup A_n f = \infty$ and $\liminf A_n f = -\infty$.

In 1975 Akcoglu [3] solved a major open problem when he proved the following celebrated theorem.

Theorem 23 *Suppose $T\colon L_p \mapsto L_p$ is a positive contraction. Then $A_n f = \frac{1}{n}\sum_{i=0}^{n-1} T^i f$ converges a.e. Moreover one has the strong L_p inequality*

$$\left\| \sup_n |A_n f| \right\|_p \leq \frac{p}{p-1} \|f\|_p . \tag{25}$$

As usual, the maximal inequality (25) is the key to the convergence. Note that it is identical in form to (10) the classical strong L_p inequality for automorphisms. (25) was proved by A. Ionesco–Tulcea (now Bellow) [106] in the case of positive invertible isometries of L_p. It is a result of Banach [16], see also [140], that in this case T arises from a non-singular automorphism τ of $(X, \mathcal{B}, \mu)$ in the form $Tf = \rho^{1/p} f \circ \tau$. By a series of reductions Bellow was able to show that in this case (25) can be deduced from (10).

Akcoglu's brilliant idea was to consider a *dilation* S of T which is a positive invertible isometry on a larger L_p space $\tilde{L}_p = L_p(Y, C, \nu)$. What this means is that there is a positive isometric injection $D\colon L_p \to \tilde{L}_p$ and a positive projection P on $\tilde{L}_p$ whose range is $D(L_p)$ such that $DT^n = PS^n D$ for all $n \geq 0$. Given the existence of such an S it is not hard to deduce (25) for T from (25) for S. In fact Akcoglu constructs a dilation only in the case when L_p is finite dimensional and shows how to reduce the proof of (25) to this case. In the finite dimensional case the construction is very concrete and P is a conditional expectation operator. Later Akcoglu and Kopp [7] gave a construction in the general case. It is noteworthy that the proof of Akcoglu's theorem consists ultimately of a long string of reductions to the classical strong L_p inequality (10), which in turn is a consequence of (5).

Subadditive and Multiplicative Ergodic Theorems

Consider a family $\{X_{n,m}\}$ of real-valued random variables on a probability space indexed by the set of pairs $(n, m) \in \mathbb{Z}^2$ such that $0 \leq n < m$. $\{X_{(n,m)}\}$ is called a (stationary) *subadditive process* if

(a) the joint distribution of $\{X_{n,m}\}$ is the same as that of $\{X_{n+1,m+1}\}$
(b) $X_{n,m} \leq X_{n,l} + X_{l,m}$ whenever $n < l < m$.

Denoting the index set by $\{n < m\} \subset \mathbb{Z}^2$ the distribution of the process is a measure μ on $\mathbb{R}^{\{n<m\}}$ which is invariant under the shift $Tx(n, m) = x(n + 1, m + 1)$. Thus there is no loss of generality in assuming that there is an underlying endomorphism T such that $X_{n,m} \circ T = X_{n+1,m+1}$. In 1968 Kingman [129] proved the following generalization of Birkhoff's theorem. $\gamma = \inf \frac{1}{n} \int X_{0,n} \mathrm{d}\mu$ is called the *time constant* of the process.

Theorem 24 *If the $X_{n,m}$ are integrable and $\gamma > -\infty$ then $\frac{1}{n} X_{0,n}$ converges a.e. and in L_1-norm to a $\mathbb{T}$-invariant limit $\bar{X} \in L_1(\mu)$ satisfying $\int \bar{X} d\mu = \gamma$.*

It is easy to deduce from the above that if one assumes only that $X_{0,1}^+$ is integrable then $\frac{1}{n} X_{0,n}$ still converges a.e. to a T-invariant limit $\bar{X}$ taking values in $[-\infty, \infty)$.

Subadditive processes first arose in the work of Hammersley and Welsh [99] on percolation theory. Here is an example. Let G be the graph with vertex set $\mathbb{Z}^2$ and with edges joining every pair of nearest neighbors. Let E denote the edge set and let $\{T_e : e \in E\}$ be non-negative integrable i.i.d. random variables. To each finite path P in G associate the "travel time" $T(P) = \sum_{e \in E} T_e$. For integers $m > n \geq 0$ let $X_{n,m}$ be the infimum of $T(P)$ over all paths P joining $(0, n)$ to $(0, m)$. This is a subadditive process with $0 \leq \gamma < \int T_e \, \mathrm{d}\mu$ and it is not hard to see that the underlying endomorphism is ergodic. Thus Kingman's theorem yields the result that $\frac{1}{n} X_{0,n} \to \gamma$ a.e.

Suppose now that T is an ergodic automorphism of a probability space $(X, \mathcal{B}, \mu)$ and P is a function on X

taking values in the space of $d \times d$ real matrices. Define $P_{n,m} = P(T^{m-1}x)P(T^{m-2}x)\ldots P(T^n x)$ and let $P_{n,m}(i,j)$ denote the i,j entry of $P_{n,m}$. Then $X_{n,m} = \log(\|P_{n,m}\|)$ (use any matrix norm) is a subadditive process so one obtains the first part of the following result of Furstenberg and Kesten [87] (1960) originally proved by more elaborate methods. The second part can also be deduced from the subadditive theorem with a little more work. See Kingman [130] for details and for some other applications of subadditive processes.

Theorem 25

(a) *Suppose* $\int \log^+(|P|)d\mu < \infty$. *Then* $\|P_{0,n}\|^{1/n}$ *converges a.e. to a finite limit.*
(b) *Suppose that for each* i,j $P(i,j)$ *is a strictly positive function such that* $\log P(i,j)$ *is integrable. Then the limit* $p = \lim(P_{0,n}(i,j))^{\frac{1}{n}}$ *exists a.e. and is independent of* i *and* j.

Partial results generalizing Kingman's theorem to the multiparameter case were obtained by Smythe [189] and Nguyen [157]. In 1981 Akcoglu and Krengel [8] obtained a definitive multi-parameter subadditive theorem. They consider an action $\{T_m\}$ of the semigroup $G = \mathbb{Z}^d_{\geq 0}$ by endomorphisms of a measure space $(X, \mathcal{B}, \mu)$. Using the standard total ordering $\prec$ of G an *interval* in G is any set of the form $\{k \in G : m \prec k \prec n\}$ for any $m \prec n \in G$. Let $\mathcal{I}$ denote the set of non-empty intervals. Reversing the direction of the inequality, they define a *superadditive* process as a collection of integrable functions $F_I, I \in \mathcal{I}$, such that

(a) $F_I \circ T_m = F_{I+m}$,
(b) $F_I \geq F_{I_1} + \ldots + F_{I_k}$ whenever I is the disjoint union of $I_1, \ldots, I_k$ and
(c) $\gamma = \sup_{I \in \mathcal{I}} |I|^{-1} \int F_I \, d\mu < \infty$.

A sequence $\{I_n\}$ of sets in $\mathcal{I}$ is called *regular* if there is an increasing sequence I'_n such that $I_n \subset I'_n$ and $|I'_n| \leq C|I_n|$ for some constant C.

Theorem 26 (Akcoglu–Krengel) *Suppose* F_I *is a superadditive process and* $\{I_n\}$ *is regular. Then* $\frac{1}{|I_n|}F_{I_n}$ *converges a.e.*

$\{F_I\}$ is additive if the inequality in (b) is replaced by equality. In this case $F_I = \sum_{n \in I} f \circ T_n$ where f is an integrable function. Thus in the additive case the Akcoglu–Krengel result is a theorem about ordinary multi-dimensional ergodic averages, which is in fact a special case of an earlier result of Tempelman [196] (see Sect. "Amenable Groups" below).

Kingman's proof of Theorem 24 hinged on the existence of a certain (typically non-unique) decomposition for subadditive processes. Akcoglu and Krengel's proof of the multi-parameter result does not depend on a Kingman-type decomposition, in fact they show that there is no such decomposition in general. They prove a weak maximal inequality

$$\mu\{\sup |I_n|^{-1} F_{I_n} > \lambda\} < \frac{C}{\lambda}\gamma \,, \tag{26}$$

where C is a constant depending only on the dimension, and show that this is sufficient to prove their result. In the case $d = 1$ the Akcoglu–Krengel argument provides a new and more natural proof of Kingman's theorem, similar in spirit to Wiener's arguments.

Akcoglu and Sucheston [9] have proved a ratio ergodic theorem for subadditive processes with respect to a positive L_1 contraction, generalizing both the Chacón–Ornstein theorem and Kingman's theorem.

In 1968 Oseledec [164] proved his celebrated multiplicative ergodic theorem, which gives very precise information about the random matrix products studied by Furstenberg and Kesten. His theorem is an important tool for the study of Lyapunov exponents in differentiable dynamics, see notably Pesin [169]. If A is a $d \times d$ matrix let $\|A\| = \sup\{\|Ax\| : \|x\| = 1\}$ where $\|x\|$ is the Euclidean norm on $\mathbb{R}^n$.

Theorem 27 *Suppose* T *is an endomorphism of the probability space* $(X, \mathcal{B}, \mu)$. *Suppose* P *is a measurable function on* X *whose values are* $d \times d$ *real matrices such that* $\int \log^+ \|P\| d\mu < \infty$ *and let* $P_n(x) = P(T^{n-1}x)P(T^{n-2}x)\ldots P(x)$. *Then there is a* T*-invariant subset* X' *of* X *with measure 1 such that for* $x \in X'$ *the following hold.*

(a) $\lim_{n\to\infty}(P_n^*(x)P_n(x))^{\frac{1}{2n}} = A(x)$ *exists.*
(b) *Let* $0 \leq \exp\lambda_1(x) \leq \exp\lambda_2(x) \leq \ldots \leq \exp\lambda_r(x)$ *be the distinct eigenvalues of* $A(x)$ ($r = r(x)$ *may depend on* x *and* λ_1 *may be* $-\infty$) *with multiplicities* $m_1(x), \ldots m_r(x)$. *Let* $E_i(x)$ *be the eigenspace corresponding to* $\exp(\lambda_i(x))$ *and set*

$$F_i(x) = E_1(x) + \ldots + E_i(x) \,.$$

Then for each $u \in F_i(x) \backslash F_{i-1}(x)$

$$\lim \frac{1}{n} \log \|P_n(x)u\| = \lambda_i(x) \,.$$

(c) *The functions* m_i *and* λ_i *are* T*-invariant.*
(d) *If* T *is ergodic,* $\det P(x) = 1$ *a.e. and*

$$\limsup_n \frac{1}{n} \int \log \|P_n\| d\mu > 0$$

then the λ_i *are constants,* $\lambda_1 < 0$ *and* $\lambda_r > 0$.

Raghunathan [174] gave a much shorter proof of Oseledec's theorem, valid for matrices with entries in a locally compact normed field. He showed that it could be reduced to the Furstenberg–Kesten theorem by considering the exterior powers of P. Ruelle [180] extended Oseledec's theorem to the case where P takes values in the set of bounded operators on a Hilbert space. Walters [199] has given a proof (under slightly stronger hypotheses) which avoids the matrix calculations and tools from multilinear algebra used in other proofs.

Entropy and the Shannon–McMillan–Breiman Theorem

The notion of entropy was introduced by Shannon in his landmark work [185] which laid the foundations for a mathematical theory of information. Suppose $(X, \mathcal{B}, \mu)$ is a probability space, P is a finite measurable partition of X and T is an automorphism of $(X, \mathcal{B}, \mu)$. $P(x)$ denotes the atom of P containing x. The *entropy* of P is

$$\begin{aligned} h(P) &= -\sum_{p\in P} \mu(p)\log(\mu(p)) \\ &= -\int \log(\mu(P(x))\,\mathrm{d}\mu(x) \geq 0\,. \end{aligned} \tag{27}$$

$-\log(\mu(A))$ may be viewed as a quantitative measure of the amount of information contained in the statement that a randomly chosen $x \in X$ happens to belong to A. So $h(P)$ is the expected information if one is about to observe which atom of P a randomly chosen point falls in. See Billingsley [34] for more motivation of this concept. See also the article in this collection on entropy by J. King or any introductory book on ergodic theory, e. g. Petersen [171].

If P and Q are partitions $P \bigvee Q$ denotes the common refinement which consists of all sets $p \cap q$, $p \in P$, $q \in Q$. It is intuitive and not hard to show that $h(P \bigvee Q) \leq h(P) + h(Q)$. Now let $P_0^n = \bigvee_{i=0}^{n-1} T^{-i}P$ and $h_n = h(P_0^n)$. The subadditivity of entropy implies that $h_{n+m} \leq h_n + h_m$, so by a well-known elementary lemma the limit

$$h(P, T) = \lim_n \frac{h_n}{n} = \inf_n \frac{h_n}{n} \geq 0 \tag{28}$$

exists.

If one thinks of $P(x)$ as a measurement performed on the space X and Tx as the state succeeding x after one second has elapsed then $h(P, T)$ is the expected information per second obtained by repeating the experiment every second for a very long time. See [177] for an alternative and very useful approach to $h(P, T)$ via *name-counting*.

The following result, which is known as the Shannon–McMillan–Breiman theorem, has proved to be of fundamental importance in ergodic theory, notably, for example, in the proof of Ornstein's celebrated isomorphism theorem for Bernoulli shifts [159].

Theorem 28 *If T is ergodic then*

$$\lim_{n\to\infty} -\frac{1}{n}\log(\mu(P_0^{n-1}(x)) = h(P, T) \tag{29}$$

a.e. and in L_1-norm.

In other words, the actual information obtained per second by observing x over time converges to the constant $h(P, T)$, namely the limiting expected information per second. Shannon [185] formulated Theorem 28 and proved convergence in probability. McMillan [149] proved L_1 convergence and Breiman [49] obtained the a.e. convergence.

The original proofs of a.e. convergence used the martingale convergence theorem, were not very intuitive and did not generalize to $\mathbb{Z}^n$-actions, where the martingale theorem is not available. Ornstein and Weiss [161] (1983) found a beautiful and more natural argument which bypasses the martingale theorem and allows generalization to a class of groups which includes $\mathbb{Z}^n$.

Amenable Groups

Let G be any countable group and $T = \{T_g\}$ an action of G by automorphisms of a probability space. Suppose σ is a complex measure on G, that is, $\{\sigma(g)\}_{g\in G}$ is an absolutely summable sequence. Let T_g act on functions via $T_g f = f \circ T_g$, so T_g is an isometry of L_p for every $1 \leq p \leq \infty$. Let $\sigma(T) = \sum_{g\in G} \sigma(g)T_g$. A very general framework for formulating ergodic theorems is to consider a sequence $\{\sigma_n\}$ and ask whether $\sigma_n(T)f$ converges, a.e. or in mean, for f in some L_p space. When $\sigma_n(T)f$ converges for all actions T and all f in L_p in p-norm or a.e. then one says that σ_n is *mean* or *pointwise good* in L_p. When the σ_n are probability measures it is natural to call such results weighted ergodic theorems and this terminology is retained for complex σ as well.

Birkhoff's theorem says that if $G = \mathbb{Z}$ and σ_n is the normalized counting measure on $\{0, 1, \ldots n-1\}$ then $\{\sigma_n\}$ is pointwise good in L_1. This section will be concerned only with sequences $\{\sigma_n\}$ such that σ_n is normalized counting measure on a finite subset $F_n \subset G$ so one speaks of mean or pointwise good sequences $\{F_n\}$. A natural condition to require of $\{F_n\}$, which will ensure that the limiting function is invariant, is that it be asymptoti-

cally (left) invariant in the sense that

$$\frac{|gF_n \triangle F_n|}{|F_n|} \to 0 \quad \forall g \in G\,. \tag{30}$$

Such a sequence is called a *Følner sequence* and a group G is *amenable* if it has a Følner sequence. As in most of this article G is restricted to be a discrete countable group for simplicity but most of the results to be seen actually hold for a general locally compact group.

Amenability of G is equivalent to the existence of a finitely additive left invariant probability measure on G. It is not hard to see that any Abelian, and more generally any solvable, group is amenable. On the other hand the free group F_2 on two generators is not amenable. See Paterson [167] for more information on amenable groups. The Følner property by itself is enough to give a mean ergodic theorem.

Theorem 29 *Any Følner sequence is mean good in L_p for $1 \le p < \infty$.*

The proof of this result is rather similar to the proof of Theorem 3. In fact Theorem 29 is only a special case of quite general results concerning amenable semi-groups acting on abstract Banach spaces. See the book of Paterson [167] for more on this.

Turning to pointwise theorems, the Følner condition alone does not yield a pointwise theorem, even when $G = \mathbb{Z}$ and the F_n are intervals. For example Akcoglu and del Junco [6] have shown that when $G = \mathbb{Z}$ and $F_n = [n, n + \sqrt{n}] \cap \mathbb{Z}$ the pointwise ergodic theorem fails for any aperiodic T and for some characteristic function f. See also del Junco and Rosenblatt [119].

The following pointwise result of Tempelman [196] is often quoted. A Følner sequence $\{F_n\}$ is called *regular* if there is a constant C such that $|F_n^{-1}F_n| \le C|F_n|$ and there is an increasing sequence F_n' such that $F_n \subset F_n'$ and $|F_n'| \le C|F_n|$.

Theorem 30 *Any regular Følner sequence is pointwise good in L_1.*

In case the F_n are intervals in $\mathbb{Z}^n$ this result can be proved by a variant of Wiener's covering argument and in the general case by an abstraction thereof. The condition $|F_n^{-1}F_n| \le C|F_n|$ captures the property of rectangles which is needed for the covering argument. Emerson [78] independently proved a very similar result.

The work on ergodic theorems for abstract locally compact groups was pioneered by Calderón [53] who built on Wiener's methods. The main result in this paper is somewhat technical but it already contains the germ of Tempelman's theorem. Other ergodic theorems for amenable groups, whose main interest lies in the case of continuous groups, include Tempelman [195], Renaud [175], Greenleaf [93] and Greenleaf and Emerson[94]. The discrete versions of these results are all rather close to Tempelman's theorem.

Among pointwise theorems for discrete groups Tempelman's result was essentially the best available for a long time. It was not known whether every amenable group had a Følner sequence which is pointwise good for some L_p. In 1988 Shulman [187] introduced the notion of a *tempered* Følner sequence $\{F_n\}$, namely one for which

$$\left|\bigcup_{i<n} F_i^{-1}F_n\right| < C|F_n|\,, \tag{31}$$

for some constant C. The advantage of the tempered condition is that any Følner sequence has a tempered subsequence, and in particular any amenable group has a tempered Følner sequence.

Shulman proved a maximal inequality in L_2 for such F_n which implies that $\{F_n\}$ is pointwise good in L_2. An account of this work may be found in Section 5.6 of Tempelman's book [194].

Lindenstrauss [145] was able to extend the result to L_1.

Theorem 31 *Any tempered Følner sequence is pointwise good in L_1.*

The key new idea in his proof is to use a probabilistic argument to establish a covering lemma. In the discrete case Ornstein and Weiss [201] have given a non-probabilistic proof of Lindenstrauss's covering lemma. Lindenstrauss also generalizes the a.e. convergence in the Shannon–McMillan–Breiman theorem to this setting. L_1 convergence was already established by Kieffer [128] in 1975.

Subsequence and Weighted Theorems

In this section G and T are as in the previous section and σ_n is a sequence of complex measures on G.

This section will be concerned with conditions on σ_n that ensure that it is mean or pointwise good. For the most part G will be $\mathbb{Z}$.

Hopf's ergodic theorem gives a class of examples for free. Choose any probability measure λ on G, define the operator $T_\lambda = \lambda(T)$ and observe that $(T_\lambda)^n = T_{\lambda^{*n}}$, where λ^{*n} denotes the convolution power. Since T_λ satisfies the hypotheses of Hopf's theorem it follows that the sequence $\sigma_n = \frac{1}{n}\sum_{i=0}^{n-1}\lambda^{*i}$ is pointwise good in L_1.

Another sort of example is given by choosing a sequence g_n in G and letting $\sigma_n = \frac{1}{n}\sum_{i=0}^{n-1}\delta_{g_n}$. If one has convergence for such a sequence one speaks of a *subsequence ergodic theorem*. This section will mainly focus on

the case $G = \mathbb{Z}$ and sequences which are the increasing enumeration of a subset $S \subset N$. Given $S \subset \mathbb{N}$ write $\sigma_{S,n}$ for the corresponding probabilities σ_n and say S is good if $\sigma_{S,n}$ is good.

Perhaps the first subsequence ergodic theorem is due to Blum and Hanson [38] who proved that an automorphism T of a probability space is *strongly mixing* if and only if $\frac{1}{n}\sum_{i=0}^{n-1} T^{m_i} f$ converges in L_2 norm for every $f \in L_2$ and every increasing $\{m_i\}$. Strong mixing means that $\mu(T^{-n}A \cap B) \to \mu(A)\mu(B)$. In 1969 Brunel and Keane [51] proved the first pointwise subsequence ergodic theorem

Theorem 32 *Suppose that T is a translation on a compact Abelian group G with Haar measure λ, $g \in G$, and $E \subset G$ is any Borel set with $\lambda(E) > 0$ and $\lambda(\partial E) = 0$. Let $S = \{i > 0 : T^i g \in E\}$. Then S is pointwise good in L_1.*

Krengel [133] constructed the first example of a sequence $S \subset \mathbb{N}$ which is pointwise universally bad, in the strong sense that for any aperiodic T the a.e. convergence of $\sigma_{S,n}(T)f$ fails for some characteristic function f. Bellow [17] proved that any lacunary sequence (meaning $a_{n+1} > ca_n$ for some $c > 1$) is pointwise universally bad in L_1.

Later Akcoglu at al. [1] were able to show that for lacunary sequences $\{\sigma_{S,n}\}$ is even *strongly sweeping out*. A sequence $\{\sigma_n\}$ of probability measures on $\mathbb{Z}$ is said to be strongly sweeping out if for any ergodic T and for all $\delta > 0$ there is a characteristic function f with $\int f \mathrm{d}\mu < \delta$ such that $\limsup \sigma_n(T)f = 1$ a.e. It is not difficult to show that if $\{\sigma_n\}$ is strongly sweeping out then there are characteristic functions f such that $\liminf \sigma_n(T)f = 0$ and $\limsup \sigma_n(T)f = 1$. Thus for lacunary sequences the ergodic theorem fails in the worst possible way.

Bellow and Losert [21] gave the first example of a sequence $S \subset \mathbb{Z}$ of density 0 which is universally good for pointwise convergence, answering a question posed by Furstenberg. They construct an S which is pointwise good in L_1. This paper also contains a good overview of the progress on weighted and subsequence ergodic theorems at that time.

Weyl's theorem on uniform distribution (Theorem 9) suggests the possibility of an ergodic theorem for the sequence $\{n^2\}$. It is not hard to see that $\{n^2\}$ is mean good in L_2. In fact the spectral theorem and the dominated convergence theorem show that it is enough to prove that the L_∞-bounded sequence of functions $\frac{1}{n}\sum_{i=0}^{n-1} z^{n^2}$ on the unit circle converges at each point z of the unit circle. When z is not a root of unity the sequence converges to 0 by Weyl's result and when z is a root of unity the convergence is trivial because $\{z^{n^2}\}$ is periodic. In 1987 Bourgain [39,43] proved his celebrated pointwise ergodic theorem for polynomial subsequences.

Theorem 33 *If p is any polynomial with rational coefficients taking integer values on the integers then $S = \{p(n)\}$ is pointwise good in L_2.*

The first step in Bourgain's argument is to reduce the problem of proving a maximal inequality to the case of the shift map on the integers, via Calderón's transfer principle. Then the problem is transferred to the circle by using Fourier transforms. At this point the problem becomes a very delicate question about exponential sums and a whole arsenal of tools is brought to bear. See Rosenblatt and Wierdl [176] and Quas and Wierdl [28](Appendix B) for nice expositions of Bourgain's methods.

Bourgain subsequently improved this to all L_p, $p > 1$ and also extended it to sequences $\{[q(n)]\}$ where now q is an arbitrary real polynomial and $[\cdot]$ denotes the greatest integer function. He also announced that his methods can be used to show that the sequence of primes is pointwise good in L_p for any $p > \frac{1+\sqrt{3}}{2}$. Wierdl [205] (1988) soon extended the result for primes to all $p > 1$.

Theorem 34 *The primes are pointwise good in L_p for $p > 1$.*

It has remained a major open question for quite some time whether any of these results hold for $p = 1$. In 2005 there appeared a preprint of Mauldin and Buczolich [148], which remains unpublished, showing that polynomial sequences are L_1-universally bad.

Another major result of Bourgain's is the so-called return times theorem [44]. A simplification of Bourgain's original proof was published jointly with Furstenberg, Katznelson and Ornstein as an appendix to an article [47] of Bourgain. To state it let us agree to say that a a sequence of complex numbers $\{a(n)\}_{n\geq 0}$ has property P if the sequence of complex measures $\sigma_n = \frac{1}{n}\sum_{i=0}^{n-1} a(i)\delta_i$ has property P, where δ_i denotes the point mass at i.

Theorem 35 (Bourgain) *Suppose T is an automorphism of a probability space $(X, \mathcal{B}, \mu)$, $1 \leq p, q \leq \infty$ are conjugate exponents and $f \in L_p(\mu)$. Then for almost all x the sequence $\{f(T^n x)\}$ is pointwise good in L_q.*

Applying this to characteristic functions $f = 1_E$ one sees that the return time sequence $\{i > 0 : S^i x \in E\}$ is good for pointwise convergence in L_1. Theorem 32 is a very special case. It is also easy to see that Theorem 35 contains the Wiener–Wintner theorem.

In 1998 Rudolph [179] proved a far-reaching generalization of the return times theorem using the technique of *joinings*. For an introduction to joinings see the article

by de la Rue in this collection and also Thouvenot [198], Glasner [90] (2003) and Rudolph's book [177]. Rudolph's result concerns the convergence of multiple averages

$$\frac{1}{N}\sum_{n=0}^{N-1}\prod_{j=1}^{k} f_j(T_j^n x) \tag{32}$$

where each T_j is an automorphism of a probability space $(X_j, \mathcal{B}_j, \mu_j)$ and the f_j are L_∞ functions. The point is that the convergence occurs whenever each $x_j \in X_j'$, sets of measure one which may be chosen sequentially for $j = 1, \ldots, k$ without knowing what T_i or f_i are for any $i > j$. He actually proves something stronger, namely he identifies an intrinsic property of a sequence $\{a_i\}$, which he calls *fully generic*, such that the following hold.

(a) The constant sequence $\{1\}$ is fully generic.
(b) If $\{a_i\}$ is fully generic then for any T and $f \in L_\infty$ the sequence $a_i f(T^i x)$ is fully generic for almost all x.
(c) Fully generic implies pointwise good in L_∞.

The definition of fully generic will not be quoted here as it is somewhat technical.

For a proof of the basic return times theorem using joinings see Rudolph [178]. Assani, Lesigne and Rudolph [13] took a first step towards the multiple theorem, a Wiener–Wintner version of the return times theorem. Also Assani [11] independently gave a proof of Rudolph's result in the case when all the T_j are weakly mixing.

Ornstein and Weiss [162] have proved the following version of the return times theorem for abstract discrete groups. As with $\mathbb{Z}$, let us say that a sequence $\{a_g\}_{g\in G}$ of complex numbers has property P for $\{F_n\}$ if the sequence $\sigma_n = \frac{1}{|F_n|}\sum_{g\in F_n} a(g)\delta_g$ of complex measures has property P.

Theorem 36 *Suppose that the increasing Følner sequence $\{F_n\}$ satisfies the Tempelman condition $\sup_n |F_n^{-1}F_n|/|F_n| < \infty$ and $\bigcup F_n = G$. If $b \in L_\infty$ then for a.a. x the sequence $\{b(T_g x)\}$ is pointwise good in L_1 for $\{F_n\}$.*

Recently Demeter, Lacey, Tao and Thiele [67] have proved that the return times theorem remains valid for any $1 < p \le \infty$ and $q \ge 2$. On the other hand Assani, Buczolich and Mauldin [14] (2005) showed that it fails for $p = q = 1$.

Bellow, Jones and Rosenblatt have a series of papers [22,23,24,25] studying general weighted averages associated to a sequence σ_n of probability measures on $\mathbb{Z}$, and, in some cases, more general groups. The following are a few of their results. [23] is concerned with $\mathbb{Z}$-actions and *moving block* averages given by $\sigma_n = m_{I_n}$, where the I_n are finite intervals and m_I denotes normalized counting measure on I. They resolve the problem completely, obtaining a checkable necessary and sufficient condition for such a sequence to be pointwise good in L_1.

[24] gives sufficient conditions on a sequence σ_n for it to be pointwise good in L_p, $p > 1$, via properties of the Fourier transforms $\hat{\sigma}_n$. A particular consequence is that if $\lim_{n\to\infty}\sum_{k\in\mathbb{Z}} |\sigma_n(k) - \sigma_n(k-1)| = 0$ then $\{\sigma_n\}$ has a subsequence which is pointwise good in L_p, $p > 1$. In [25] they obtain convergence results for sequences $\sigma_n = \sigma^n$, the convolution powers of a probability measure σ. A consequence of one of their main results is that if the expectation $\sum_{k\in\mathbb{Z}} k\sigma(k)$ is zero, the second moment $\sum_{k\in\mathbb{Z}} k^2\sigma(k)$ is finite and σ is *aperiodic* (its support is not contained in any proper coset in $\mathbb{Z}$) then σ^n is pointwise good in L_p for $p > 1$.

Bellow and Calderón [19] later showed that this last result is valid also for $p = 1$. This is a consequence of the following sufficient condition for a sequence T to satisfy a weak L_1 inequality. Given an automorphism of a probability space $(X, \mathcal{B}, \mu)$ let $Mf = \sup|\sigma_n(T)f(x)|$ be the maximal operator associated to $\{\sigma_n\}$.

Theorem 37 (Bellow and Calderón) *Suppose there is an $\alpha \in (0, 1]$ and $C > 0$ such that for each $n > 1$ one has*

$$|\sigma_n(x+y) - \sigma_n(x)| \le C\frac{|y|^\alpha}{|x|^{1+\alpha}} \quad \text{for all } x, y \in \mathbb{Z} \quad \text{such that } 0 < 2|y| \le |x|$$

Then there is a constant D such that

$$\mu\{Mf > \lambda\} \le \frac{D}{\lambda}\|f\|_1 \quad \text{for all } T, \quad f \in L_1(\mu) \quad \text{and } \lambda > 0\,.$$

Ergodic Theorems and Multiple Recurrence

Suppose $S \subset \mathbb{N}$. The *upper density* of S is

$$\bar{d}(S) = \limsup_n \frac{|S \cap [1, n]|}{n}\,. \tag{33}$$

and the *density* $d(S)$ is the limit of the same quantity, if it exists. In 1975 Szemerédi [190] proved the following celebrated theorem, answering an old question of Erdős and Turán.

Theorem 38 *Any subset of $\mathbb{N}$ with positive upper density contains an arithmetic progression of length k for each $k \ge 1$.*

This result has a distinctly ergodic-theoretic flavor. Letting T denote the shift map on $\mathbb{Z}$, it says that for each k there

is an n such that $S' = \bigcap_{i=1}^{k} T^{-in} S$ is non-empty. In fact the result gives more: there is an n for which $\bar{d}(S') > 0$. In this light Szemerédi's theorem becomes a multiple recurrence theorem for the shift map on $\mathbb{N}$, equipped with the invariant "measure-like" quantity $\bar{d}$. Of course $\bar{d}$ is not even finitely additive so it is *not* a measure. d, however, is at least finitely additive, when defined, and $d(\mathbb{N}) = 1$.

This point of view suggests the following multiple recurrence theorem.

Theorem 39 *Suppose T is an automorphism of a probability space $(X, \mathcal{B}, \mu)$, $\mu(B) > 0$ and $k \geq 1$. Then there is an $n > 0$ such that $\mu(\bigcap_{i=1}^{k} T^{in} B) > 0$.*

In 1977, Furstenberg [85] proved the following ergodic theorem which implies the multiple recurrence theorem. He also established a general *correspondence principle* which puts the shaky analogy between the multiple recurrence theorem and Szemerédi's theorem on a firm footing and allows each to be deduced from the other. Thus he obtained an ergodic theoretic proof of Szemerédi's combinatorial result.

Theorem 40 *Suppose T is an automorphism of a probability space $(X, \mathcal{B}, \mu)$, $f \in L_\infty$, $f \geq 0$, $\int f d\mu > 0$ and $k \geq 1$. Then*

$$\liminf_N \frac{1}{N} \sum_{n=0}^{N-1} \int \prod_{i=1}^{k} T^{in} f d\mu > 0 \,. \tag{34}$$

Furstenberg's result opened the door to the study of so-called ergodic Ramsey theory which has yielded a vast array of deep results in combinatorics, many of which have no non-ergodic proof as yet. The focus of this article is not on this direction but the reader is referred to Furstenberg's book [86] for an excellent introduction and to Bergelson [27,28] for surveys of later developments. There is also the article by Frantzikinakis and McCutcheon in this collection.

Furstenberg's proof relies on a deep structure theorem for a general automorphism which was also developed independently by Zimmer [213], [212] in a more general context. A *factor* of T is any sub-σ-algebra $\mathcal{F} \subset \mathcal{B}$ such that $T(\mathcal{F}) = \mathcal{F}$. (It is more accurate to think of the factor as the action of T on the measure space $(X, \mathcal{F}, \mu|\mathcal{F})$.) The structure theorem asserts that there is a transfinite increasing sequence of factors $\{\mathcal{F}_\alpha\}$ of T such that the following conditions hold.

(a) $\mathcal{F}_{\alpha+1}$ is *compact* relative to $\mathcal{F}_\alpha$.
(b) $\mathcal{F}_\alpha = \bigvee_{\beta<\alpha} \mathcal{F}_\beta$ whenever α is a limit ordinal.
(c) $\mathcal{B}$ is *weakly mixing* relative to $\bigvee \mathcal{F}_\alpha$.

($\bigvee$ denotes the σ-algebra generated by a collection of σ-algebras.) $\bigvee_\alpha \mathcal{F}_\alpha$ is called the *maximal distal factor* of T and if $\bigvee \mathcal{F}_\alpha = \mathcal{B}$ then T is called *distal*. The definitions of the relative properties in (a) and (c) above are somewhat technical so only their absolute versions (i. e. relative to the trivial σ-algebra $\{\emptyset, X\}$) will be described here.

T is said to be compact if for each $f \in L_2$ the orbit $\{T^i f : i \in \mathbb{Z}\}$ is pre-compact in the norm topology of L_2. This turns out to be equivalent to the statement that T is a translation on a compact Abelian group endowed with its Haar measure. The property of weak mixing is a fundamental notion in ergodic theory which has many equivalent definitions. The most appropriate for our purposes is that T is weakly mixing if it has no compact factors. This turns out to be equivalent to the ergodicity of $T \times T$ acting on the product measure space $(X, \mathcal{B}, \mu) \times (X, \mathcal{B}, \mu)$.

The verification of 37 in the case of compact T is rather easy. In this case it is not hard to prove that for any $f \in L_\infty$ and $\epsilon > 0$ the set $\{n \in \mathbb{Z} : \|T^n f - f\|_2 < \epsilon\}$ has bounded gaps and (34) follows easily. In the case when T is weakly mixing (34) is a consequence of the following theorem which Furstenberg proves in [85] (as a warm-up for it's much harder relative version).

Theorem 41 *If T is weakly mixing and $f_1, f_2, \ldots, f_k$ are L_∞ functions then*

$$\lim_N \frac{1}{N} \sum_{n=0}^{N-1} \int \prod_{i=1}^{k} T^{in} f_i d\mu = \prod_{i=1}^{k} \int f_i \tag{35}$$

Later Bergelson [26] showed that the result can be obtained easily by an induction argument using the following Hilbert space generalization of van der Corput's lemma.

Theorem 42 (Bergelson) *Suppose $\{x_n\}$ is a bounded sequence of vectors in Hilbert space such that for each $h > 0$ one has $\frac{1}{N} \sum_{n=0}^{N-1} \langle x_{n+h}, x_n \rangle \to 0$ as $N \to \infty$. Then $\| \frac{1}{N} \sum_{n=0}^{N-1} x_n \| \to 0$.*

Ryzhikov has also given a beautiful short proof of Theorem 41 using joinings ([182]). Bergelson's van der Corput lemma and variants of it have been a key tool in subsequent developments in ergodic Ramsey theory and in the convergence results to be be discussed in this section. Bergelson [26] used it to prove the following mean ergodic theorem for weakly mixing automorphisms.

Theorem 43 *Suppose T is weakly mixing, $f_1, \ldots, f_k$ are L_∞ functions and $p_1, \ldots, p_k$ are polynomials with rational coefficients taking integer values on the integers such that no*

$p_i - p_j$ is constant for $i \neq j$. Then

$$\lim_{N\to\infty} \left\| \frac{1}{N} \sum_{n=0}^{N-1} \prod_{i=1}^{k} T^{p_i(n)} f_i - \prod_{i=1}^{k} \int f_i d\mu \right\| = 0 . \quad (36)$$

Theorems 40 and 41 immediately raise the question of convergence of the multiple averages $\frac{1}{N}\sum_{n=0}^{N-1}\prod_{i=1}^{k} T^{\text{in}} f_i$ for a general T. Several authors obtained partial results on the question of mean convergence. It was finally resolved only recently by Host and Kra [104], who proved the following landmark theorem.

Theorem 44 *Suppose $f_1, f_2, \ldots, f_k \in L_\infty$. Then there is a $g \in L_\infty$ such that*

$$\lim \left\| \frac{1}{N} \sum_{n=0}^{N-1} \prod_{i=1}^{k} T^{\text{in}} f_i - g \right\|_2 = 0 . \quad (37)$$

Independently and somewhat later Ziegler [211]obtained the same result by somewhat different methods. Furstenberg had already established Theorem 44 for $k = 2$ in [85]. It was proved for $k = 3$ in the case of a totally ergodic T by Conze and Lesigne [63]and in general by Host and Kra [102]. It can also be obtained using the methods developed by Furstenberg and Weiss [88]. In this paper Furstenberg and Weiss proved a result for polynomial powers of T in the case $k = 2$. They also formalized the key notion of a *characteristic factor*. A factor C of T is said to be characteristic for the averages (37) if, roughly speaking, the L_2 limiting behavior of the averages is unchanged when any one of the f_i's is replaced by its conditional expectation on C. This means that the question of convergence of these averages may be reduced to the case when f_i are all C-measurable. So the problem is to find the right (smallest) characteristic factor and prove convergence for that factor.

The importance of characteristic factors was already apparent in Furstenberg's original paper [85], where he showed that the maximal distal factor is characteristic for the averages (37). In fact he showed that for a given k a k-step distal factor is characteristic. (An automorphism is k-step distal if it is the top rung in a k-step ladder of factors as in the Furstenberg–Zimmer structure-theorem.) It turns out, though, that the right characteristic factor for (37) is considerably smaller. In their seminal paper [63] Conze and Lesigne identified the characteristic factor for $k = 3$, now called the Conze–Lesigne factor. As shown in [104], and [211], the characteristic factor for a general k is (isomorphic to) an inverse limit of k-step nilflows. A k-step nilflow is a compact homogeneous space N/Γ of a k-step nilpotent Lie group N, endowed with its unique left-invariant probability measure, on which T acts via left translation by an element of N. Ergodic properties of nilflows have been studied for some time in ergodic theory, for example in Parry [166]. In this way the problem of L_2-convergence of (37) is reduced to the case when T is a nilflow. In this case one has more: the averages converge pointwise by a result of Leibman [141] (See also Ziegler [210]).

There have already been a good many generalizations of (37). Host and Kra [103], Frantzikinakis and Kra [82, 83], and Leibman [141] have proved results which replace linear powers of T by polynomial powers. In increasing degrees of generality Conze and Lesigne [63], Frankzikinakis and Kra [81] and Tao [192] have obtained results which replace the maps $T, T^2, \ldots, T^k$ in (37) by commuting maps $T_1, \ldots, T_k$. Bergelson and Leibman [30,31] have obtained results, both positive and negative, in the case of two noncommuting maps.

In the direction of pointwise convergence the only general result is the following theorem of Bourgain [48] which asserts pointwise convergence in the case $k = 2$.

Theorem 45 *Suppose S and T are powers of a single automorphism R and $f, g \in L_\infty$. Then $\frac{1}{N}\sum_{n=1}^{N} f(T^n x) g(S^n x)$ converges a.e.*

When T is a K-automorphism Derrien and Lesigne [68] have proved that the averages (35) converge pointwise to the product of the integrals, even with polynomial powers of T replacing the linear powers.

Gowers [91] has given a new proof of Szemerédi's theorem by purely finite methods using only harmonic analysis on $\mathbb{Z}_n$. His results give better quantitative estimates in the statement of finite versions of Szemerédi's theorem. Although his proof contains no ergodic theory it is to some extent guided by Furstenberg's approach.

This section would be incomplete without mentioning the spectacular recent result of Green and Tao [92] on primes in arithmetic progression and the subsequent extensions of the Green-Tao theorem due to Tao [191] and Tao and Ziegler [193].

Rates of Convergence

There are many results which say in various ways that, in general, there are no estimates for the rate of convergence of the averages $A_n f$ in Birkhoff's theorem. For example there is the following result of Krengel [134].

Theorem 46 *Suppose $\lim_{n\to\infty} c_n = \infty$ and T is any ergodic automorphism of a probability space. Then there is a bounded measurable f with $\int f \, d\mu = 0$ such that $\limsup c_n A_n f = \infty$ a.e.*

See Part 1 of Derriennic [69] for a selection of other results in this direction. In spite of these negative results one can obtain quantitative estimates by reformulating the ergodic theorem in various ways. Bishop [37] proved the following result which is purely finite and constructive in nature and evidently implies the a.e. convergence in Birkhoff's theorem. If $y = (y_1, \ldots, y_n)$ is a finite sequence of real numbers and $a < b$, an *upcrossing* of y over $[a, b]$ is a minimal integer interval $[k, l] \subset [1, n]$ satisfying $y_k < a$ and $y_l > b$.

Theorem 47 *Suppose T is an automorphism of the probability space $(X, \mathcal{B}, \mu)$. Let $U(n, a, b, f, x)$ be the number of upcrossings of the sequence $A_0 f(x), \ldots, A_n f(x)$ over $[a, b]$. Then for every n*

$$\mu\{x : U(n, a, b, f, x) > k\} \leq \frac{\|f\|_1}{(b-a)k} . \tag{38}$$

Ivanov [107] has obtained the following stronger upcrossing inequality for an arbitrary *positive* measurable f, which also implies Birkhoff's theorem.

Theorem 48 *For any positive measurable f and $0 < a < b$*

$$\mu\{x : U(n, a, b, f, x) > k\} \leq \left(\frac{a}{b}\right)^k . \tag{39}$$

Note the exponential decay and the remarkable fact that the estimate does not depend on f. Ivanov has also obtained the following result (Theorem 23 in [121]) about *fluctuations* of $A_n f$. An *ϵ-fluctuation* of a real sequence $y = (y_1 \ldots y_n)$ is a minimal integer interval $[k, l]$ satisfying $|y_k - y_l| \geq \epsilon$. If $f \in L_1(\mathbb{R})$ let $F(\epsilon, f, x)$ be the number of ϵ-fluctuations of the sequence $\{A_n f(x)\}_{n=0}^{\infty}$.

Theorem 49

$$\mu\{x : F(\epsilon, f, x) \geq k\} \leq C \frac{(\log k)^{\frac{1}{2}}}{k} \tag{40}$$

where C is a constant depending only on $\|f\|_1/\epsilon$. If $f \in L_\infty$ then

$$\mu\{x : F(\epsilon, f, x) \geq k\} \leq A e^{-Bk} , \tag{41}$$

where A and B are constants depending only on $\|f\|_\infty/\epsilon$.

See Kachurovskii's survey [121] of results on rates of convergence in the ergodic theorem and also the article of Jones et al. [115] for more results on oscillation type inequalities.

Under special assumptions on f and T it is possible to give more precise results on speed of convergence. If the sequence $T^i f$ is independent then there is a vast literature in probability theory giving very precise results, for example the central limit theorem and the law of the iterated logarithm (see, for example, [50]). See the surveys of Derrienic [69] (2006) and of Merlevède, Peligrad and Utev [150] for results on the central limit theorem for dynamical systems.

Ergodic Theorems for Non-amenable Groups

Guivarch [95] was the first to prove an ergodic theorem for a general pair of non-commuting unitary operators. Much work has been done in the last 15 years on mean and pointwise theorems for non-amenable groups. See the extensive survey by Nevo [155]. Here is just one result of Nevo [154] as a sample. Suppose G is a discrete group and $\{\sigma_n\}$ is a sequence of probability measures on G. Say that $\{\sigma_n\}$ is *mean ergodic* if for any unitary representation π of G on a Hilbert space $\mathcal{H}$ and any $x \in \mathcal{H}$ one has $\sum_{g \in G} \sigma_n(g)\pi(g)x$ converges in norm to the projection of x on the subspace of π-invariant vectors. Say that $\{\sigma_n\}$ is *pointwise ergodic* if for any measure preserving action T of G on a probability space and $f \in L_2$ one has $\sum_{g \in G} \sigma_n(g) T_g f$ converges a.e. to the projection of f on the subspace of T-invariant functions.

Theorem 50 *Let G be the free group on k generators, $k \geq 1$. Let τ_n be the normalized counting measure on the set of elements whose word length (in terms of the generators and their inverses) is n. Let $\sigma_n = (\tau_n + \tau_{n+1})/2$ and $\sigma'_n = \frac{1}{n}\sum_{i=1}^{n} \sigma_i$. Then σ_n and σ'_n are mean and pointwise ergodic but τ_n is not mean ergodic.*

Future Directions

This section will be devoted to a few open problems and some questions. Many of these have been suggested by my colleagues acknowledged in the introduction.

The topic of convergence of Furstenberg's multiple averages has seen some remarkable achievements in the last few years and will likely continue to be vital for some time to come. The question of pointwise convergence of multiple averages is completely open, beyond Bourgain's result (Theorem 45). Even extending Bourgain's result to three different powers of R or to any two commuting automorphisms S and T would be a very significant achievement. Another natural question is whether one has pointwise convergence of the averages of the sequence $f(T^n(x))g(T^{n^2}(x))$, which are mean convergent by the result of Furstenberg and Weiss [88] (which is now subsumed in much more general results).

A long-standing open problem relating to Akcoglu's ergodic theorem (Theorem 23) for positive contractions of L_p is whether it can be extended to power-bounded operators T (this means that $\|T\|^n$ is bounded). It is also an open question whether it extends to non-positive contractions, excepting the case $p = 2$ where Burkholder [52] has shown that it fails.

There are many natural questions in the area of subsequence and weighted ergodic theorems. For example, which of Bourgain's several pointwise theorems can be extended to L_1? Are there other natural subsequences of an arithmetic character which have density 0 and for which an ergodic theorem is valid, either mean or pointwise, and in what L_p spaces might such theorems be valid? Are there higher dimensional analogues?

Since lacunary sequences are bad, to have any sort of pointwise theorem $l_n = \log \frac{a_n}{a_{n+1}}$ must get close to 0 and for simplicity let us assume that $\lim l_n = 0$. How fast must the convergence be in order to get an ergodic theorem? Jones and Wierdl [113] have shown that if $l_n > (\log n)^{-\frac{1}{2}+\epsilon}$ then the pointwise ergodic theorem fails in L_2 while Jones, Lacey and Wierdl [116] have shown that an only slightly faster rate permits a sequence which is pointwise good in L_2. How well or badly does the ergodic theorem succeed or fail depending on the rate of convergence of l_n? In particular is there a (slow) rate which still guarantees strong sweeping out? [116] contains some interesting conjectures in this direction.

There are also interesting questions concerning the mean and pointwise ergodic theorems for subsequences which are chosen randomly in some sense. See Bourgain [42] and [116] for some results in this direction. Again [116] contains some interesting conjectures along these lines.

In a recent paper [32] Bergelson and Leibman prove some very interesting and surprising results about the distribution of *generalized polynomials*. A generalized polynomial is any function which can be built starting with polynomials in $\mathbb{R}[x]$ using the operations of addition, multiplication and taking the greatest integer. As a consequence they derive a generalization of von Neumann's mean ergodic theorem to averages along generalized polynomial sequences. The following is a special case.

Theorem 51 *Suppose p is a generalized polynomial taking integer values on the integers and U is a unitary operator on $\mathcal{H}$. Then $\frac{1}{n}\sum_{i=0}^{n-1} U^{p(i)}x$ is norm convergent for all $x \in \mathcal{H}$.*

This begs the question: does one have pointwise convergence? If so, this would be a far-reaching generalization of Bourgain's polynomial ergodic theorem.

There are also lots of questions concerning the nature of Følner sequences $\{F_n\}$ in an amenable group which give a pointwise theorem. For example Lindenstrauss [145] has shown that in the *lamplighter* group, a semi-direct product of $\mathbb{Z}$ with $\bigoplus_{i\in\mathbb{Z}} \mathbb{Z}/2\mathbb{Z}$ on which Z acts by the shift, there is no sequence satisfying Tempelman's condition and that any $\{F_n\}$ satisfying the Shulman condition must grow super-exponentially. So, it is natural to ask for slower rates of growth. In particular, in any amenable group is there always a sequence $\{F_n\}$ which is pointwise good and grows at most exponentially? Can one do better either in general or in particular groups?

Lindenstrauss's theorem at least guarantees the existence of Følner sequences which are pointwise good in L_1 but in particular groups there are often natural sequences which one hopes might be good. For example in $\bigoplus_{i=1}^{\infty} \mathbb{Z}$ one may take F_n to be a cube based at 0 of sidelength l_n and dimension d_n (that is, all but the first d_n co-ordinates are zero), where both sequences increase to ∞. What conditions on l_n and d_n will give a good sequence? Note that no such sequence is regular in Tempelman's sense. If $d_n = n$ then $\{l_n\}$ must be superexponential to ensure Shulman's condition. Can one do better? What about $l_n = d_n = n$?

Bibliography

1. Akcoglu M, Bellow A, Jones RL, Losert V, Reinhold–Larsson K, Wierdl M (1996) The strong sweeping out property for lacunary sequences, Riemann sums, convolution powers, and related matters. Ergodic Theory Dynam Systems 16(2):207–253
2. Akcoglu M, Jones RL, Rosenblatt JM (2000) The worst sums in ergodic theory. Michigan Math J 47(2):265–285
3. Akcoglu MA (1975) A pointwise ergodic theorem in L_p-spaces. Canad J Math 27(5):1075–1082
4. Akcoglu MA, Chacon RV (1965) A convexity theorem for positive operators. Z Wahrsch Verw Gebiete 3:328–332 (1965)
5. Akcoglu MA, Chacon RV (1970) A local ratio theorem. Canad J Math 22:545–552
6. Akcoglu MA, del Junco A (1975) Convergence of averages of point transformations. Proc Amer Math Soc 49:265–266
7. Akcoglu MA, Kopp PE (1977) Construction of dilations of positive L_p-contractions. Math Z 155(2):119–127
8. Akcoglu MA, Krengel U (1981) Ergodic theorems for superadditive processes. J Reine Angew Math 323:53–67
9. Akcoglu MA, Sucheston L (1978) A ratio ergodic theorem for superadditive processes. Z Wahrsch Verw Gebiete 44(4):269–278
10. Alaoglu L, Birkhoff G (1939) General ergodic theorems. Proc Nat Acad Sci USA 25:628–630
11. Assani I (2000) Multiple return times theorems for weakly mixing systems. Ann Inst H Poincaré Probab Statist 36(2):153–165
12. Assani I (2003) Wiener–Wintner ergodic theorems. World Scientific Publishing Co. Inc., River Edge, NJ
13. Assani I, Lesigne E, Rudolph D (1995) Wiener–Wintner return-times ergodic theorem. Israel J Math 92(1–3):375–395

14. Assani I, Buczolich Z, Mauldin RD (2005) An L^1 counting problem in ergodic theory. J Anal Math 95:221–241
15. Auslander L, Green L, Hahn F (1963) Flows on homogeneous spaces. With the assistance of Markus L, Massey W and an appendix by Greenberg L Annals of Mathematics Studies, No 53. Princeton University Press, Princeton, NJ
16. Banach S (1993) Théorie des opérations linéaires. Éditions Jacques Gabay, Sceaux, reprint of the 1932 original
17. Bellow A (1983) On "bad universal" sequences in ergodic theory II. In: Belley JM, Dubois J, Morales P (eds) Measure theory and its applications. Lecture Notes in Math, vol 1033. Springer, Berlin, pp 74–78
18. Bellow A (1999) Transference principles in ergodic theory. In: Christ M, Kenig CE, Sadowsky C (eds) Harmonic analysis and partial differential equations, Chicago Lectures in Math. Univ Chicago Press, Chicago, pp 27–39
19. Bellow A, Calderón A (1999) A weak-type inequality for convolution products. In: Christ M, Kenig CE, Sadowsky C (eds) Harmonic analysis and partial differential equations, Chicago Lectures in Math. Univ Chicago Press, Chicago, pp 41–48
20. Bellow A, Jones R (eds) (1991) Almost everywhere convergence, II. Academic Press, Boston
21. Bellow A, Losert V (1985) The weighted pointwise ergodic theorem and the individual ergodic theorem along subsequences. Trans Amer Math Soc 288(1):307–345
22. Bellow A, Jones R, Rosenblatt J (1989) Almost everywhere convergence of powers. In: Edgar GA, Sucheston L (eds) Almost everywhere convergence. Academic, Boston, pp 99–120
23. Bellow A, Jones R, Rosenblatt J (1990) Convergence for moving averages. Ergodic Theory Dynam Systems 10(1):43–62
24. Bellow A, Jones RL, Rosenblatt J (1992) Almost everywhere convergence of weighted averages. Math Ann 293(3):399–426
25. Bellow A, Jones R, Rosenblatt J (1994) Almost everywhere convergence of convolution powers. Ergodic Theory Dynam Systems 14(3):415–432
26. Bergelson V (1987) Weakly mixing PET. Ergodic Theory Dynam Systems 7(3):337–349
27. Bergelson V (1996) Ergodic Ramsey theory—an update. In: Pollicott M, Schmidt K (eds) Ergodic theory of Z^d actions. London Math Soc Lecture Note Ser, vol 228. Cambridge Univ Press, Cambridge, pp 1–61
28. Bergelson V (2006) Combinatorial and Diophantine applications of ergodic theory. In: Hasselblatt B, Katok A (eds) Handbook of dynamical systems, vol 1B, Appendix A by Leibman A, Appendix B by Quas A, Wierdl M. Elsevier, Amsterdam, pp 745–869
29. Bergelson V (2007) Some historical remarks and modern questions around the ergodic theorem. Internat Math Nachrichten 205:1–10
30. Bergelson V, Leibman A (2002) A nilpotent Roth theorem. Invent Math 147(2):429–470
31. Bergelson V, Leibman A (2004) Failure of the Roth theorem for solvable groups of exponential growth. Ergodic Theory Dynam Systems 24(1):45–53
32. Bergelson V, Leibman A (2007) Distribution of values of bounded generalized polynomials. Acta Math 198(2):155–230
33. Berkson E, Bourgain J, Gillespie TA (1991) On the almost everywhere convergence of ergodic averages for power-bounded operators on L^p-subspaces. Integral Equ Operator Theory 14(5):678–715
34. Billingsley P (1965) Ergodic theory and information. Wiley, New York
35. Birkhoff GD (1931) Proof of the ergodic theorem. Proc Nat Acad Sci USA 17:656–660
36. Bishop E (1966) An upcrossing inequality with applications. Michigan Math J 13:1–13
37. Bishop E (1967/1968) A constructive ergodic theorem. J Math Mech 17:631–639
38. Blum JR, Hanson DL (1960) On the mean ergodic theorem for subsequences. Bull Amer Math Soc 66:308–311
39. Bourgain J (1987) On pointwise ergodic theorems for arithmetic sets. C R Acad Sci Paris Sér I Math 305(10):397–402
40. Bourgain J (1988) Almost sure convergence and bounded entropy. Israel J Math 63(1):79–97
41. Bourgain J (1988) An approach to pointwise ergodic theorems. In: Lindenstrauss J, Milman VD (eds) Geometric aspects of functional analysis (1986/87). Lecture Notes in Math, vol 1317. Springer, Berlin, pp 204–223
42. Bourgain J (1988) On the maximal ergodic theorem for certain subsets of the integers. Israel J Math 61(1):39–72
43. Bourgain J (1988) On the pointwise ergodic theorem on L^p for arithmetic sets. Israel J Math 61(1):73–84
44. Bourgain J (1988) Temps de retour pour les systèmes dynamiques. C R Acad Sci Paris Sér I Math 306(12):483–485
45. Bourgain J (1988) Temps de retour pour les systèmes dynamiques. C R Acad Sci Paris Sér I Math 306(12):483–485
46. Bourgain J (1989) Almost sure convergence in ergodic theory. In: Edgar GA, Sucheston L (eds) Almost everywhere convergence. Academic, Boston, pp 145–151
47. Bourgain J (1989) Pointwise ergodic theorems for arithmetic sets. Inst Hautes Études Sci Publ Math (69):5–45, with an appendix by the author, Furstenberg H, Katznelson Y, Ornstein DS
48. Bourgain J (1990) Double recurrence and almost sure convergence. J Reine Angew Math 404:140–161
49. Breiman L (1957) The individual ergodic theorem of information theory. Ann Math Statist 28:809–811
50. Breiman L (1968) Probability. Addison–Wesley, Reading
51. Brunel A, Keane M (1969) Ergodic theorems for operator sequences. Z Wahrsch Verw Gebiete 12:231–240
52. Burkholder DL (1962) Semi–Gaussian subspaces. Trans Amer Math Soc 104:123–131
53. Calderon AP (1953) A general ergodic theorem. Ann Math (2) 58:182–191
54. Calderón AP (1968) Ergodic theory and translation-invariant operators. Proc Natl Acad Sci USA 59:349–353
55. Calderón AP (1968) Ergodic theory and translation-invariant operators. Proc Natl Acad Sci USA 59:349–353
56. Calderon AP, Zygmund A (1952) On the existence of certain singular integrals. Acta Math 88:85–139
57. Chacon RV (1962) Identification of the limit of operator averages. J Math Mech 11:961–968
58. Chacon RV (1963) Convergence of operator averages. In: Wright FB (ed) Ergodic theory. Academic, New York, pp 89–120
59. Chacon RV (1963) Linear operators in L_1. In: Wright FB (ed) Ergodic theory. Academic, New York, pp 75–87
60. Chacon RV (1964) A class of linear transformations. Proc Amer Math Soc 15:560–564

61. Chacon RV, Ornstein DS (1960) A general ergodic theorem. Illinois J Math 4:153–160
62. Conze JP, Lesigne E (1984) Théorèmes ergodiques pour des mesures diagonales. Bull Soc Math France 112(2):143–175
63. Conze JP, Lesigne E (1988) Sur un théorème ergodique pour des mesures diagonales. In: Probabilités, Publ Inst Rech Math Rennes, vol 1987. Univ Rennes I, Rennes, pp 1–31
64. Cotlar M (1955) On ergodic theorems. Math Notae 14:85–119 (1956)
65. Cotlar M (1955) A unified theory of Hilbert transforms and ergodic theorems. Rev Mat Cuyana 1:105–167 (1956)
66. Day M (1942) Ergodic theorems for Abelian semigroups. Trans Amer Math Soc 51:399–412
67. Demeter C, Lacey M, Tao T, Thiele C (2008) Breaking the duality in the return times theorem. Duke Math J 143(2):281–355
68. Derrien JM, Lesigne E (1996) Un théorème ergodique polynomial ponctuel pour les endomorphismes exacts et les K-systèmes. Ann Inst H Poincaré Probab Statist 32(6):765–778
69. Derriennic Y (2006) Some aspects of recent works on limit theorems in ergodic theory with special emphasis on the "central limit theorem". Discrete Contin Dyn Syst 15(1):143–158
70. Derriennic Y (2006) Some aspects of recent works on limit theorems in ergodic theory with special emphasis on the "central limit theorem". Discrete Contin Dyn Syst 15(1):143–158
71. Doob JL (1938) Stochastic processes with an integral-valued parameter. Trans Amer Math Soc 44(1):87–150
72. Dunford N (1951) An individual ergodic theorem for non-commutative transformations. Acta Sci Math Szeged 14:1–4
73. Dunford N, Schwartz J (1955) Convergence almost everywhere of operator averages. Proc Natl Acad Sci USA 41:229–231
74. Dunford N, Schwartz JT (1988) Linear operators, Part I. Wiley Classics Library, Wiley, New York. General theory, with the assistance of Bade WG, Bartle RG, Reprint of the 1958 original, A Wiley–Interscience Publication
75. Durand S, Schneider D (2003) Random ergodic theorems and regularizing random weights. Ergodic Theory Dynam Systems 23(4):1059–1092
76. Eberlein WF (1949) Abstract ergodic theorems and weak almost periodic functions. Trans Amer Math Soc 67:217–240
77. Edgar G, Sucheston L (eds) (1989) Almost everywhere convergence. Academic Press, Boston
78. Emerson WR (1974) The pointwise ergodic theorem for amenable groups. Amer J Math 96:472–487
79. Feldman J (2007) A ratio ergodic theorem for commuting, conservative, invertible transformations with quasi-invariant measure summed over symmetric hypercubes. Ergodic Theory Dynam Systems 27(4):1135–1142
80. Foguel SR (1969) The ergodic theory of Markov processes. In: Van Nostrand Mathematical Studies, No 21. Van Nostrand Reinhold, New York
81. Frantzikinakis N, Kra B (2005) Convergence of multiple ergodic averages for some commuting transformations. Ergodic Theory Dynam Systems 25(3):799–809
82. Frantzikinakis N, Kra B (2005) Polynomial averages converge to the product of integrals. Israel J Math 148:267–276
83. Frantzikinakis N, Kra B (2006) Ergodic averages for independent polynomials and applications. J London Math Soc (2) 74(1):131–142
84. Furstenberg H (1967) Disjointness in ergodic theory, minimal sets, and a problem in Diophantine approximation. Math Systems Theory 1:1–49
85. Furstenberg H (1977) Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. J Analyse Math 31:204–256
86. Furstenberg H (1981) Recurrence in ergodic theory and combinatorial number theory. Princeton University Press, Princeton, NJ, m B Porter Lectures
87. Furstenberg H, Kesten H (1960) Products of random matrices. Ann Math Statist 31:457–469
88. Furstenberg H, Weiss B (1996) A mean ergodic theorem for $(1/N)\sum_{n=1}^{N} f(T^n x)g(T^{n^2}x)$. In: Bergelson V, March P, Rosenblatt J (eds) Convergence in ergodic theory and probability. Ohio State Univ Math Res Inst Publ, vol 5. de Gruyter, Berlin, pp 193–227
89. Garsia AM (1970) Topics in almost everywhere convergence. In: Lectures in Advanced Mathematics, vol 4. Markham, Chicago
90. Glasner E (2003) Ergodic theory via joinings, Mathematical Surveys and Monographs, vol 101. American Mathematical Society, Providence
91. Gowers WT (2001) A new proof of Szemerédi's theorem. Geom Funct Anal 11(3):465–588
92. Green B, Tao T (2004) The primes contain arbitrarily large arithmetic progressions. http://arxivorg/abs/mathNT/0404188
93. Greenleaf FP (1973) Ergodic theorems and the construction of summing sequences in amenable locally compact groups. Comm Pure Appl Math 26:29–46
94. Greenleaf FP, Emerson WR (1974) Group structure and the pointwise ergodic theorem for connected amenable groups. Adv Math 14:153–172
95. Guivarc'h Y (1969) Généralisation d'un théorème de von Neumann. C R Acad Sci Paris Sér A-B 268:A1020–A1023
96. Halmos PR (1946) An ergodic theorem. Proc Natl Acad Sci USA 32:156–161
97. Halmos PR (1949) A nonhomogeneous ergodic theorem. Trans Amer Math Soc 66:284–288
98. Halmos PR (1960) Lectures on ergodic theory. Chelsea, New York
99. Hammersley JM, Welsh DJA (1965) First-passage percolation, subadditive processes, stochastic networks, and generalized renewal theory. In: Proc Internat Res Semin. Statist Lab, University of California, Berkeley. Springer, New York, pp 61–110
100. Hopf E (1937) Ergodentheorie. In: Ergebnisse der Mathematik und ihrer Grenzgebiete, vol 5. Springer, Berlin
101. Hopf E (1954) The general temporally discrete Markoff process. J Rational Mech Anal 3:13–45
102. Host B, Kra B (2001) Convergence of Conze–Lesigne averages. Ergodic Theory Dynam Systems 21(2):493–509
103. Host B, Kra B (2005) Convergence of polynomial ergodic averages. Israel J Math 149:1–19
104. Host B, Kra B (2005) Nonconventional ergodic averages and nilmanifolds. Ann Math (2) 161(1):397–488
105. Hurewicz W (1944) Ergodic theorem without invariant measure. Ann Math (2) 45:192–206
106. Tulcea AI (1964) Ergodic preperties of positive isometries. Bull AMS 70:366–371

107. Ivanov VV (1996) Geometric properties of monotone functions and the probabilities of random oscillations. Sibirsk Mat Zh 37(1):117–15
108. Ivanov VV (1996) Oscillations of averages in the ergodic theorem. Dokl Akad Nauk 347(6):736–738
109. Jewett RI (1969/1970) The prevalence of uniquely ergodic systems. J Math Mech 19:717–729
110. Jones R, Rosenblatt J, Tempelman A (1994) Ergodic theorems for convolutions of a measure on a group. Illinois J Math 38(4):521–553
111. Jones RL (1987) Necessary and sufficient conditions for a maximal ergodic theorem along subsequences. Ergodic Theory Dynam Systems 7(2):203–210
112. Jones RL (1993) Ergodic averages on spheres. J Anal Math 61:29–45
113. Jones RL, Wierdl M (1994) Convergence and divergence of ergodic averages. Ergodic Theory Dynam Systems 14(3):515–535
114. Jones RL, Olsen J, Wierdl M (1992) Subsequence ergodic theorems for L^p contractions. Trans Amer Math Soc 331(2):837–850
115. Jones RL, Kaufman R, Rosenblatt JM, Wierdl M (1998) Oscillation in ergodic theory. Ergodic Theory Dynam Systems 18(4):889–935
116. Jones RL, Lacey M, Wierdl M (1999) Integer sequences with big gaps and the pointwise ergodic theorem. Ergodic Theory Dynam Systems 19(5):1295–1308
117. Jones RL, Rosenblatt JM, Wierdl M (2001) Oscillation inequalities for rectangles. Proc Amer Math Soc 129(5):1349–1358 (electronic)
118. Jones RL, Rosenblatt JM, Wierdl M (2003) Oscillation in ergodic theory: higher dimensional results. Israel J Math 135:1–27
119. del Junco A, Rosenblatt J (1979) Counterexamples in ergodic theory and number theory. Math Ann 245(3):185–197
120. Kac M (1947) On the notion of recurrence in discrete stochastic processes. Bull Amer Math Soc 53:1002–1010
121. Kachurovskii AG (1996) Rates of convergence in ergodic theorems. Uspekhi Mat Nauk 51(4(310)):73–124
122. Kachurovskii AG (1996) Spectral measures and convergence rates in the ergodic theorem. Dokl Akad Nauk 347(5):593–596
123. Kakutani S (1940) Ergodic theorems and the Markoff process with a stable distribution. Proc Imp Acad Tokyo 16:49–54
124. Kalikow S, Weiss B (1999) Fluctuations of ergodic averages. In: Proceedings of the Conference on Probability, Ergodic Theory, and Analysis, vol 43. pp 480–488
125. Kamae T (1982) A simple proof of the ergodic theorem using nonstandard analysis. Israel J Math 42(4):284–290
126. Katok A, Hasselblatt B (1995) Introduction to the modern theory of dynamical systems, Encyclopedia of Mathematics and its Applications, vol 54. Cambridge University Press, Cambridge, with a supplementary chapter by Katok A and Mendoza L
127. Katznelson Y, Weiss B (1982) A simple proof of some ergodic theorems. Israel J Math 42(4):291–296
128. Kieffer JC (1975) A generalized Shannon–McMillan theorem for the action of an amenable group on a probability space. Ann Probability 3(6):1031–1037
129. Kingman JFC (1968) The ergodic theory of subadditive stochastic processes. J Roy Statist Soc Ser B 30:499–510
130. Kingman JFC (1976) Subadditive processes. In: École d'Été de Probabilités de Saint–Flour, V–1975. Lecture Notes in Math, vol 539. Springer, Berlin, pp 167–223
131. Koopman B (1931) Hamiltonian systems and transformations in hilbert spaces. Proc Natl Acad Sci USA 17:315–318
132. Krantz SG, Parks HR (1999) The geometry of domains in space. Birkhäuser Advanced Texts: Basler Lehrbücher. Birkhäuser Boston Inc, Boston, MA
133. Krengel U (1971) On the individual ergodic theorem for subsequences. Ann Math Statist 42:1091–1095
134. Krengel U (1978/79) On the speed of convergence in the ergodic theorem. Monatsh Math 86(1):3–6
135. Krengel U (1985) Ergodic theorems. In: de Gruyter studies in mathematics, vol 6. de Gruyter, Berlin, with a supplement by Antoine Brunel
136. Krengel U, Lin M, Wittmann R (1990) A limit theorem for order preserving nonexpansive operators in L_1. Israel J Math 71(2):181–191
137. Krieger W (1972) On unique ergodicity. In: Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, vol II. Probability Theory, University of California Press, Berkeley, pp 327–346
138. Kryloff N, Bogoliouboff N (1937) La théorie générale de la mesure dans son application à l'étude des systèmes dynamiques de la mécanique non linéaire. Ann Math (2) 38(1):65–113
139. Kuipers L, Niederreiter H (1974) Uniform distribution of sequences. Wiley, New York, Pure and Applied Mathematics
140. Lamperti J (1958) On the isometries of certain function-spaces. Pacific J Math 8:459–466
141. Leibman A (2005) Convergence of multiple ergodic averages along polynomials of several variables. Israel J Math 146:303–315
142. Leibman A (2005) Pointwise convergence of ergodic averages for polynomial actions of $\mathbb{Z}^d$ by translations on a nilmanifold. Ergodic Theory Dynam Systems 25(1):215–225
143. Lemańczyk M, Lesigne E, Parreau F, Volný D, Wierdl M
144. Lesigne E (1989) Théorèmes ergodiques pour une translation sur un nilvariété. Ergodic Theory Dynam Systems 9(1):115–126
145. Lindenstrauss E (1999) Pointwise theorems for amenable groups. Electron Res Announc Amer Math Soc 5:82–90
146. Loomis LH (1946) A note on the Hilbert transform. Bull Amer Math Soc 52:1082–1086
147. Lorch ER (1939) Means of iterated transformations in reflexive vector spaces. Bull Amer Math Soc 45:945–947
148. Mauldin D, Buczolich Z (2005) Concepts behind divergent ergodic averages along the squares. In: Assani I (ed) Ergodic theory and related fields. Contemp Math, vol 430. Amer Math Soc, Providence, pp 41–56
149. McMillan B (1953) The basic theorems of information theory. Ann Math Statistics 24:196–219
150. Merlevède F, Peligrad M, Utev S (2006) Recent advances in invariance principles for stationary sequences. Probab Surv 3:1–36
151. von Neumann J (1932) Proof of the quasi-ergodic hypothesis. Proc Natl Acad Sci USA 18:70–82
152. Neveu J (1961) Sur le théorème ergodique ponctuel. C R Acad Sci Paris 252:1554–1556
153. Neveu J (1965) Mathematical foundations of the calculus of probability. Translated by Amiel Feinstein, Holden-Day, San Francisco

154. Nevo A (1994) Harmonic analysis and pointwise ergodic theorems for noncommuting transformations. J Amer Math Soc 7(4):875–902
155. Nevo A (2006) Pointwise ergodic theorems for actions of groups. In: Hasselblatt B, Katok A (eds) Handbook of dynamical systems vol 1B. Elsevier, Amsterdam, pp 871–982
156. Nevo A, Stein EM (1994) A generalization of Birkhoff's pointwise ergodic theorem. Acta Math 173(1):135–154
157. Nguyen XX (1979) Ergodic theorems for subadditive spatial processes. Z Wahrsch Verw Gebiete 48(2):159–176
158. Orey S (1971) Lecture notes on limit theorems for Markov chain transition probabilities. In: Van Nostrand Reinhold Mathematical Studies, no 34. Van Nostrand Reinhold Co, London
159. Ornstein D (1970) Bernoulli shifts with the same entropy are isomorphic. Adv Math 4:337–352 (1970)
160. Ornstein D (1971) A remark on the Birkhoff ergodic theorem. Illinois J Math 15:77–79
161. Ornstein D, Weiss B (1983) The Shannon–McMillan–Breiman theorem for a class of amenable groups. Israel J Math 44(1): 53–60
162. Ornstein D, Weiss B (1992) Subsequence ergodic theorems for amenable groups. Israel J Math 79(1):113–127
163. Ornstein DS (1960) On invariant measures. Bull Amer Math Soc 66:297–300
164. Oseledec VI (1968) A multiplicative ergodic theorem. Characteristic Ljapunov, exponents of dynamical systems. Trudy Moskov Mat Obšč 19:179–210
165. Oxtoby JC (1952) Ergodic sets. Bull Amer Math Soc 58:116–136
166. Parry W (1969) Ergodic properties of affine transformations and flows on nilmanifolds. Amer J Math 91:757–771
167. Paterson A (1988) Amenability, Mathematical Surveys and Monographs, vol 29. American Mathematical Society, Providence, RI
168. Peck JEL (1951) An ergodic theorem for a noncommutative semigroup of linear operators. Proc Amer Math Soc 2:414–421
169. Pesin JB (1977) Characteristic Ljapunov exponents, and smooth ergodic theory. Uspehi Mat Nauk 32(4(196)):55–112,287
170. Petersen K (1983) Another proof of the existence of the ergodic Hilbert transform. Proc Amer Math Soc 88(1):39–43
171. Petersen K (1989) Ergodic theory, Cambridge Studies in Advanced Mathematics, vol 2. Cambridge University Press, Cambridge
172. Pitt HR (1942) Some generalizations of the ergodic theorem. Proc Cambridge Philos Soc 38:325–343
173. Poincaré H (1987) Les méthodes nouvelles de la mécanique céleste. Tome I, II, III. Les Grands Classiques Gauthier–Villars. Librairie Scientifique et Technique Albert Blanchard, Paris
174. Raghunathan MS (1979) A proof of Oseledec's multiplicative ergodic theorem. Israel J Math 32(4):356–362
175. Renaud PF (1971) General ergodic theorems for locally compact groups. Amer J Math 93:52–64
176. Rosenblatt JM, Wierdl M (1995) Pointwise ergodic theorems via harmonic analysis. In: Peterson KE, Salama IA (eds) Ergodic theory and its connections with harmonic analysis, London Math Soc Lecture Note Ser, vol 205. Cambridge University Press, Cambridge, pp 3–151
177. Rudolph DJ (1990) Fundamentals of measurable dynamics. In: Fundamentals of measurable dynamics: Ergodic theory on Lebesgue spaces. Oxford Science Publications, The Clarendon Press, Oxford University Press, New York
178. Rudolph DJ (1994) A joinings proof of Bourgain's return time theorem. Ergodic Theory Dynam Systems 14(1):197–203
179. Rudolph DJ (1998) Fully generic sequences and a multiple-term return-times theorem. Invent Math 131(1):199–228
180. Ruelle D (1982) Characteristic exponents and invariant manifolds in Hilbert space. Ann Math (2) 115(2):243–290
181. Ryll–Nardzewski C (1951) On the ergodic theorems, II. Ergodic theory of continued fractions. Studia Math 12:74–79
182. Ryzhikov V (1994) Joinings, intertwining operators, factors and mixing properties of dynamical systems. Russian Acad Sci Izv Math 42:91–114
183. Shah NA (1998) Invariant measures and orbit closures on homogeneous spaces for actions of subgroups generated by unipotent elements. In: Dani SG (ed) Lie groups and ergodic theory. Tata Inst Fund Res Stud Math, vol 14. Tata Inst Fund Res, Bombay, pp 229–271
184. Shalom Y (1998) Random ergodic theorems, invariant means and unitary representation. In: Dani SG (ed) Lie groups and ergodic theory. Tata Inst Fund Res Stud Math, vol 14. Tata Inst Fund Res, Bombay, pp 273–314
185. Shannon CE (1948) A mathematical theory of communication. Bell System Tech J 27:379–423, 623–656
186. Shields PC (1987) The ergodic and entropy theorems revisited. IEEE Trans Inform Theory 33(2):263–266
187. Shulman A (1988) Maximal ergodic theorems on groups. Dep Lit NIINTI No. 2184
188. Sine R (1970) A mean ergodic theorem. Proc Amer Math Soc 24:438–439
189. Smythe RT (1976) Multiparameter subadditive processes. Ann Probability 4(5):772–782
190. Szemerédi E (1975) On sets of integers containing no k elements in arithmetic progression. Acta Arith 27:199–245
191. Tao T (2005) The gaussian primes contain arbitrarily shaped constellations. http://arxivorg/abs/math/0501314
192. Tao T (2008) Norm convergence of multiple ergodic averages for commuting transformations. Ergod Theory Dyn Syst 28(2):657–688
193. Tao T, Ziegler T (2006) The primes contain arbitrarily long polynomial progressions. http://frontmathucdavisedu/06105050
194. Tempelman A (1992) Ergodic theorems for group actions, Mathematics and its Applications, vol 78. Kluwer, Dordrecht, informational and thermodynamical aspects, Translated and revised from the 1986 Russian original
195. Tempel'man AA (1967) Ergodic theorems for general dynamical systems. Dokl Akad Nauk SSSR 176:790–793
196. Tempel'man AA (1972) Ergodic theorems for general dynamical systems. Trudy Moskov Mat Obšč 26:95–132
197. Tempel'man AA (1972) A generalization of a certain ergodic theorem of Hopf. Teor Verojatnost i Primenen 17:380–383
198. Thouvenot JP (1995) Some properties and applications of joinings in ergodic theory. In: Peterson KE, Salama IA (eds) Ergodic theory and its connections with harmonic analysis, London Math Soc Lecture Note Ser, vol 205. Cambridge University Press, Cambridge, pp 207–235

199. Walters P (1993) A dynamical proof of the multiplicative ergodic theorem. Trans Amer Math Soc 335(1):245–257
200. Weber M (1998) Entropie métrique et convergence presque partout, Travaux en Cours, vol 58. Hermann, Paris
201. Weiss B (2003) Actions of amenable groups. In: Bezuglyi S, Kolyada S (eds) Topics in dynamics and ergodic theory, London Math Soc Lecture Note Ser, vol 310. Cambridge University Press, Cambridge, pp 226–262
202. Weyl H (1916) Über die Gleichverteilung von Zahlen mod Eins. Math Ann 77(3):313–352
203. Wiener N (1939) The ergodic theorem. Duke Math J 5(1):1–18
204. Wiener N, Wintner A (1941) Harmonic analysis and ergodic theory. Amer J Math 63:415–426
205. Wierdl M (1988) Pointwise ergodic theorem along the prime numbers. Israel J Math 64(3):315–336 (1989)
206. Wittmann R (1995) Almost everywhere convergence of ergodic averages of nonlinear operators. J Funct Anal 127(2): 326–362
207. Yosida K (1940) An abstract treatment of the individual ergodic theorem. Proc Imp Acad Tokyo 16:280–284
208. Yosida K (1940) Ergodic theorems of Birkhoff–Khintchine's type. Jap J Math 17:31–36
209. Yosida K, Kakutani S (1939) Birkhoff's ergodic theorem and the maximal ergodic theorem. Proc Imp Acad, Tokyo 15:165–168
210. Ziegler T (2005) A non-conventional ergodic theorem for a nilsystem. Ergodic Theory Dynam Systems 25(4):1357–1370
211. Ziegler T (2007) Universal characteristic factors and Furstenberg averages. J Amer Math Soc 20(1):53–97 (electronic)
212. Zimmer RJ (1976) Ergodic actions with generalized discrete spectrum. Illinois J Math 20(4):555–588
213. Zimmer RJ (1976) Extensions of ergodic group actions. Illinois J Math 20(3):373–409
214. Zund JD (2002) George David Birkhoff and John von Neumann: a question of priority and the ergodic theorems, 1931–1932. Historia Math 29(2):138–156
215. Zygmund A (1951) An individual ergodic theorem for noncommutative transformations. Acta Sci Math Szeged 14:103–110

Ergodic Theory: Basic Examples and Constructions

Matthew Nicol[1], Karl Petersen[2]
[1] Department of Mathematics, University of Houston, Houston, USA
[2] Department of Mathematics, University of North Carolina, Chapel Hill, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Examples
Constructions
Future Directions
Bibliography

Glossary

- A transformation T of a measure space $(X, \mathcal{B}, \mu)$ is *measure-preserving* if $\mu(T^{-1}A) = \mu(A)$ for all measurable $A \in \mathcal{B}$.
- A measure-preserving transformation $(X, \mathcal{B}, \mu, T)$ is *ergodic* if $T^{-1}(A) = A \pmod{\mu}$ implies $\mu(A) = 0$ or $\mu(A^c) = 0$ for each measurable set $A \in \mathcal{B}$.
- A measure-preserving transformation $(X, \mathcal{B}, \mu, T)$ of a probability space is *weak-mixing* if $\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} |\mu(T^{-i}A \cap B) - \mu(A)\mu(B)| = 0$ for all measurable sets $A, B \in \mathcal{B}$.

  A measure-preserving transformation $(X, \mathcal{B}, \mu, T)$ of a probability space is *strong-mixing* if $\lim_{n\to\infty} \mu(T^{-n}A \cap B) = \mu(A)\mu(B)$ for all measurable sets $A, B \in \mathcal{B}$.
- A continuous transformation T of a compact metric space X is *uniquely ergodic* if there is only one T-invariant Borel probability measure on X. A continous transformation of a topological space X is *topologically mixing* for any two open sets $U, V \subset X$ there exists $N > 0$ such that $T^{-n}(U) \cap V \neq \emptyset$, for each $n \geq N$.
- Suppose $(X, \mathcal{B}, \mu)$ is a probability space. A *finite partition* $\mathcal{P}$ of X is a finite collection of disjoint (mod μ, i. e., up to sets of measure 0) measurable sets $\{P_1, \ldots, P_n\}$ such that $X = \cup P_i \pmod{\mu}$. The *entropy* of $\mathcal{P}$ with respect to μ is $H(\mathcal{P}) = -\sum_i \mu(P_i) \ln \mu(P_i)$ (other bases are sometimes used for the logarithm).
- The *metric (or measure-theoretic) entropy of T with respect to $\mathcal{P}$* is $h_\mu(T, \mathcal{P}) = \lim_{n\to\infty} \frac{1}{n} H(\mathcal{P} \vee \cdots \vee T^{-n+1}(\mathcal{P}))$, where $\mathcal{P} \vee \cdots \vee T^{-n+1}(\mathcal{P})$ is the partition of X into sets of points with the same coding with respect to $\mathcal{P}$ under T^i, $i = 0, \ldots, n-1$. That is x, y are in the same set of the partition $\mathcal{P} \vee \cdots \vee T^{-n+1}(\mathcal{P})$ if and only if $T^i(x)$ and $T^i(y)$ lie in the same set of the partition $\mathcal{P}$ for $i = 0, \ldots, n-1$.
- The *metric entropy* $h_\mu(T)$ of $(X, \mathcal{B}, \mu, T)$ is the supremum of $h_\mu(T, \mathcal{P})$ over all finite measurable partitions $\mathcal{P}$.
- If T is a continuous transformation of a compact metric space X, then the *topological entropy* of T is the supremum of the metric entropies $h_\mu(T)$, where the supremum is taken over all T-invariant Borel probability measures.
- A system $(X, \mathcal{B}, \mu, T)$ is *loosely Bernoulli* if it is isomorphic to the first-return system to a subset of positive measure of an irrational rotation or a (positive or infinite entropy) Bernoulli system.

- Two systems are *spectrally isomorphic* if the unitary operators that they induce on their L^2 spaces are unitarily equivalent.
- A *smooth dynamical system* consists of a differentiable manifold M and a differentiable map $f: M \to M$. The degree of differentiability may be specified.
- Two submanifolds S_1, S_2 of a manifold M *intersect transversely* at $p \in M$ if $T_p(S_1) + T_p(S_2) = T_p(M)$.
- An (ϵ-) small C^r *perturbation* of a C^r map f of a manifold M is a map g such that $d_{C^r}(f, g) < \epsilon$ i. e. the distance between f and g is less than ϵ in the C^r topology.
- A map T of an interval $I = [a, b]$ is *piecewise smooth* (C^k for $k \geq 1$) if there is a finite set of points $a = x_1 < x_2 < \cdots < x_n = b$ such that $T|(x_i, x_{i+1})$ is C^k for each i. The degree of differentiability may be specified.
- A measure μ on a measure space $(X, \mathcal{B})$ is *absolutely continuous* with respect to a measure ν on $(X, \mathcal{B})$ if $\nu(A) = 0$ implies $\mu(A) = 0$ for all measurable $A \in \mathcal{B}$.
- A Borel measure μ on a Riemannian manifold M is *absolutely continuous* if it is absolutely continuous with respect to the Riemannian volume on M.
- A measure μ on a measure space $(X, \mathcal{B})$ is *equivalent* to a measure ν on $(X, \mathcal{B})$ if μ is absolutely continuous with respect to ν and ν is absolutely continuous with respect to μ.

Definition of the Subject

Measure-preserving systems are a common model of processes which evolve in time and for which the rules governing the time evolution don't change. For example, in Newtonian mechanics the planets in a solar system undergo motion according to Newton's laws of motion: the planets move but the underlying rule governing the planets' motion remains constant. The model adopted here is to consider the time-evolution as a transformation (either a map in discrete time or a flow in continuous time) on a probability space or more generally a measure space. This is the setting of the subject called ergodic theory. Applications of this point of view include the areas of statistical physics, classical mechanics, number theory, population dynamics, statistics, information theory and economics. The purpose of this chapter is to present a flavor of the diverse range of examples of measure-preserving transformations which have played a role in the development and application of ergodic theory and smooth dynamical systems theory. We also present common constructions involving measure-preserving systems. Such constructions may be considered a way of putting 'building-block' dynamical systems together to construct examples or decomposing a complicated system into simple 'building-blocks' to understand it better.

Introduction

In this chapter we collect a brief list of some important examples of measure-preserving dynamical systems, which we denote typically by $(X, \mathcal{B}, \mu, T)$ or $(T, X, \mathcal{B}, \mu)$ or slight variations. These examples have played a formative role in the development of dynamical systems theory, either because they occur often in applications in one guise or another or because they have been useful simple models to understand certain features of dynamical systems. There is a fundamental difference in the dynamical properties of those systems which display hyperbolicity: roughly speaking there is some exponential divergence of nearby orbits under iteration of the transformation. In differentiable systems this is associated with the derivative of the transformation possessing eigenvalues of modulus greater than one on a 'dynamically significant' subset of phase space. Hyperbolicity leads to complex dynamical behavior such as positive topological entropy, exponential divergence of nearby orbits ("sensitivity to initial conditions") often coexisting with a dense set of periodic orbits. If ϕ, ψ are sufficiently regular functions on the phase space X of a hyperbolic measure-preserving transformation (T, X, μ), then typically we have fast decay of correlations in the sense that

$$\left| \int_X \phi(T^n x)\psi(x)\,\mathrm{d}\mu - \int \phi\,\mathrm{d}\mu \int \psi\,\mathrm{d}\mu \right| \leq Ca(n)$$

where $a(n) \to 0$. If $a(n) \to 0$ at an exponential rate we say that the system has exponential decay of correlations. A theme in dynamical systems is that the time series formed by sufficiently regular observations on systems with some degree of hyperbolicity often behave statistically like independent identically distributed random variables.

At this point it is appropriate to point out two pervasive differences between the usual probabilistic setting of a stationary stochastic process $\{X_n\}$ and the (smooth) dynamical systems setting of a time series of observations on a measure-preserving system $\{\phi \circ T^n\}$. The most crucial is that for deterministic dynamical systems the time series is usually not an independent process, which is a common assumption in the strictly probabilistic setting. Even if some weak-mixing is assumed in the probabilistic setting it is usually a mixing condition on the σ-algebras $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ generated by successive random variables, a condition which is not natural (and usually very difficult to check) for dynamical systems. Mix-

ing conditions on dynamical systems are given more naturally in terms of the mixing of the sets of the σ-algebra $\mathcal{B}$ of the probability space $(X, \mathcal{B}, \mu)$ under the action of T and not by mixing properties of the σ-algebras generated by the random variables $\{\phi \circ T^n\}$. The other difference is that in the probabilistic setting, although $\{X_n\}$ satisfy moment conditions, usually no regularity properties, such as the Hölder property or smoothness, are assumed. In contrast in dynamical systems theory the transformation T is often a smooth or piecewise smooth transformation of a Riemannian manifold X and the observation $\phi\colon X \to \mathbb{R}$ is often assumed continuous or Hölder. The regularity of the observation ϕ turns out to play a crucial role in proving properties such as rates of decay of correlation, central limit theorems and so on.

An example of a hyperbolic transformation is an expanding map of the unit interval $T(x) = (2x)$ (where (x) is x modulo the integers). Here the derivative has modulus 2 at all points in phase space. This map preserves Lebesgue measure, has positive topological entropy, Lebesgue almost every point x has a dense orbit and periodic points for the map are dense in $[0, 1)$.

Non-hyperbolic systems are of course also an important class of examples, and in contrast to hyperbolic systems they tend to model systems of 'low complexity', for example systems displaying quasiperiodic behavior. The simplest non-trivial example is perhaps an irrational rotation of the unit interval $[0, 1)$ given by a map $T(x) = (x + \alpha)$, $\alpha \in \mathbb{R} \setminus \mathbb{Q}$. T preserves Lebesgue measure, every point has a dense orbit (there are no periodic orbits), yet the topological entropy is zero and nearby points stay the same distance from each other under iteration under T.

There is a natural notion of equivalence for measure-preserving systems. We say that measure-preserving systems $(T, X, \mathcal{B}, \mu)$ and (S, Y, C, ν) are *isomorphic* if (possibly after deleting sets of measure 0 from X and Y) there is a one-to-one onto measurable map $\phi\colon X \to Y$ with measurable inverse ϕ^{-1} such that $\phi \circ T = S \circ \phi$ μ a.e. and $\mu(\phi^{-1}(A)) = \nu(A)$ for all $A \in C$. If X, Y are compact topological spaces we say that T is *topologically conjugate* to S if there exists a homeomorphism $\phi\colon X \to Y$ such that $\phi \circ T = S \circ \phi$. In this case we call ϕ a *conjugacy*. If ϕ is C^r for some $r \geq 1$ we will call ϕ a C^r*-conjugacy* and similarly for other degrees of regularity.

We will consider $X = [0, 1)(\mathrm{mod}\, 1)$ as a representation of the unit circle $S^1 = \{z \in \mathbb{C} : |z| = 1\}$ (under the map $x \to e^{2\pi i x}$) and similarly represent the k-dimensional torus $T^k = S^1 \times \cdots \times S^1$ (k-times). If the σ-algebra is clear from the context we will write (T, X, μ) instead of $(T, X, \mathcal{B}, \mu)$ when denoting a measure-preserving system.

Examples

Rigid Rotation of a Compact Group

If G is a compact group equipped with Haar measure and $a \in G$, then the transformation $T(x) = ax$ preserves Haar measure and is called a *rigid rotation* of G. If G is abelian and the transformation is ergodic (in this setting transitivity implies ergodicity), then the transformation is uniquely ergodic. Such systems always have zero topological entropy.

The simplest example of such a system is a circle rotation. Take $X = [0, 1)(\mathrm{mod}\, 1)$, with

$$T(x) = (x + \alpha) \quad \text{where} \quad \alpha \in \mathbb{R}\,.$$

Then T preserves Lebesgue (Haar) measure and is ergodic (in fact uniquely ergodic) if and only if α is irrational. Similarly, the map

$$T(x_1, \ldots, x_k) = (x_1 + \alpha_1, \ldots, x_k + \alpha_k), \qquad \text{where} \quad \alpha_1, \ldots, \alpha_k \in \mathbb{R}\,,$$

preserves k-dimensional Lebesgue (Haar) measure and is ergodic (uniquely ergodic) if and only if there are no integers $m_1, \ldots, m_k$, not all 0, which satisfy $m_1\alpha_1 + \cdots + m_k\alpha_k \in \mathbb{Z}$.

Adding Machines

Let $\{k_i\}_{i\in\mathbb{N}}$ be a sequence of integers with $k_i \geq 2$. Equip each cyclic group $\mathbb{Z}_{k_i}$ with the discrete topology and form the product space $\Sigma = \prod_{i=1}^{\infty} \mathbb{Z}_{k_i}$ equipped with the product topology. An *adding machine* corresponding to the sequence $\{k_i\}_{i\in\mathbb{N}}$ is the topological space $\Sigma = \prod_{i=1}^{\infty} \mathbb{Z}_{k_i}$ together with the map

$$\sigma\colon \Sigma \to \Sigma$$

defined by

$$\sigma(k_1 - 1, k_2 - 1, \ldots) = (0, 0, \ldots) \qquad \text{if each entry in the } \mathbb{Z}_{k_i} \text{ component is } k_i - 1\,,$$

while

$$\sigma(k_1 - 1, k_2 - 1, \ldots, k_n - 1, x_1, x_2, \ldots) = (0, \overbrace{0, \ldots, 0}^{n \text{ times}}, x_1 + 1, x_2, x_3, \ldots)$$

when $x_1 \neq k_{n+1} - 1$.

The map σ may be thought of as "add one and carry" and also as mapping each point to its successor in a certain order. See Sect. "Adic Transformations" for generalizations. If each $k_i = 2$ then the system is called the *dyadic*

(or von Neumann-Kakutani) adding machine or *2-odometer*. Adding machines give examples of naturally occurring minimal systems of low orbit complexity in the sense that the topological entropy of an adding machine is zero. In fact if f is a continuous map of an interval with zero topological entropy and S is a closed, topologically transitive invariant set without periodic orbits, then the restriction of f to S is topologically conjugate to the dyadic adding machine (Theorem 11.3.13 in [46]).

We say a non-empty set Λ is an *attractor* for a map T if there is an open set U containing Λ such that $\Lambda = \cap_{n\geq 0} T^n(U)$ (other definitions are found in the literature). The dyadic adding machine is topologically conjugate to the Feigenbaum attractor at the limit point of period doubling bifurcations (see Sect. "Unimodal Maps"). Furthermore, attractors for continuous unimodal maps of the interval are either periodic orbits, transitive cycles of intervals, or Cantor sets on which the dynamics is topologically conjugate to an adding machine [32].

Interval Exchange Maps

A map $T\colon [0,1] \to [0,1]$ is an *interval exchange transformation* if it is defined in the following way. Suppose that π is a permutation of $\{1,\ldots,n\}$ and $l_i > 0$, $i = 1,\ldots,n$, is a sequence of subintervals of I (open or closed) with $\sum_i l_i = 1$. Define t_i by $l_i = t_i - t_{i-1}$ with $t_0 = 0$. Suppose also that σ is an n-vector with entries ± 1. T is defined by sending the interval $t_{i-1} \leq x < t_{i-1}$ of length l_i to the interval

$$\sum_{\pi(j)<\pi(i)} l_{\pi(j)} \leq x < \sum_{\pi(j)>\pi(i)} l_{\pi(j)}$$

with orientation preserved if the ith entry of σ is $+1$ and orientation reversed if the ith entry of σ is -1. Thus on each interval l_i, T has the form $T(x) = \sigma_i x + a_i$, where σ_i is ± 1. If $\sigma_i = 1$ for each i, the transformation is called *orientation preserving*.

The transformation T has finitely many discontinuities (at the endpoints of each l_i), and modulo this set of discontinuities is smooth. T is also invertible (neglecting the finite set of discontinuities) and preserves Lebesgue measure. These maps have zero topological entropy and arise naturally in studies of polygonal billiards and more generally area-preserving flows. There are examples of minimal but non-ergodic interval exchange maps [56,69].

Full Shifts and Shifts of Finite Type

Given a finite set (or alphabet) $A = \{0,\ldots,d-1\}$, take $X = \Omega^+(A) = A^{\mathbb{N}}$ (or $X = A^{\mathbb{Z}}$) the sets of one-sided (two-sided) sequences, respectively, with entries from A. For example sequences in $A^{\mathbb{N}}$ have the form $x = \cdot x_0 x_1 \ldots x_n \ldots$. A *cylinder set* $C(y_{n_1},\ldots,y_{n_k})$, $y_{n_i} \in A$, of length k is a subset of X defined by fixing k entries; for example,

$$C(y_{n_1},\ldots,y_{n_k}) = \{x\colon x_{n_1} = y_{n_1},\ldots,x_{n_k} = y_{n_k}\}.$$

We define the set A^k to consist of all cylinders $C(y_1,\ldots,y_k)$ determined by fixing the first k entries, i. e. an element of A^k is specified by fixing the first k entries of a sequence $\cdot x_0 \cdots x_k$ by requiring $x_i = y_i$, $i = 0,\ldots,k$.

Let $p = (p_0,\ldots,p_{d-1})$ be a probability vector: all $p_i \geq 0$ and $\sum_{i=0}^{d-1} p_i = 1$. For any cylinder $B = C(b_1,\ldots,b_k) \in A^k$, define

$$g_k(B) = p_{b_1} \ldots p_{b_k}\,. \tag{1}$$

It can be shown that these functions on A^k extend to a shift-invariant measure μ_p on $A^{\mathbb{N}}$ (or $A^{\mathbb{Z}}$) called product measure. (See the article on ► Measure Preserving Systems.) The space $A^{\mathbb{N}}$ or $A^{\mathbb{Z}}$ may be given a metric by defining

$$d(x,y) = \begin{cases} 1 & \text{if } x_0 \neq y_0\,; \\ \frac{1}{2^{|n|}} & \text{if } x_n \neq y_n \text{ and } x_i = y_i \text{ for } |i| < n\,. \end{cases}$$

The *shift* $\sigma(.x_0x_1\cdots x_n\cdots) = .x_1x_2\cdots x_n\cdots$ is ergodic with respect to μ_p. The measure-preserving system $(\Omega,\mathcal{B},\mu,\sigma)$ (with $\mathcal{B}$ the σ-algebra of Borel subsets of $\Omega(A)$, or its completion), is denoted by $\mathcal{B}(p)$ and is called the *Bernoulli shift* determined by p. This system models an infinite number of independent repetitions of an experiment with finitely many outcomes, the ith of which has probability p_i on each trial.

These systems are mixing of all orders (i. e. σ^n is mixing for all $n \geq 1$) and have countable Lebesgue spectrum (hence are all spectrally isomorphic). Kolmogorov and Sinai showed that two of them cannot be isomorphic unless they have the same entropy; Ornstein [81] showed the converse. $\mathcal{B}(1/2,1/2)$ is isomorphic to the Lebesgue-measure-preserving transformation $x \to 2x \bmod 1$ on $[0,1]$; similarly, $\mathcal{B}(1/3,1/3,1/3)$ is isomorphic to $x \to 3x \bmod 1$. Furstenberg asked whether the only nonatomic measure invariant for both $x \to 2x \bmod 1$ and $x \to 3x \bmod 1$ on $[0,1]$ is Lebesgue measure. Lyons [68] showed that if one of the actions is K, then the measure must be Lebesgue, and Rudolph [100] showed the same thing under the weaker hypothesis that one of the actions has positive entropy. For further work on this question, see [50,87].

This construction can be generalized to model one-step finite-state Markov stochastic processes as dynamical systems. Again let $A = \{0,\ldots,d-1\}$, and let $p = (p_0,\ldots,$

p_{d-1}) be a probability vector. Let P be a $d \times d$ *stochastic matrix* with rows and columns indexed by A. This means that all entries of P are nonnegative, and the sum of the entries in each row is 1. We regard P as giving the transition probabilities between pairs of elements of A. Now we define for any cylinder $B = C(b_1, \ldots, b_k) \in A^k$

$$\mu_{p,P}(B) = p_{b_1} P_{b_1 b_2} P_{b_2 b_3} \ldots P_{b_{k-1} b_k} \,. \tag{2}$$

It can be shown that $\mu_{p,P}$ extends to a measure on the Borel σ-algebra of $\Omega^+(A)$, and its completion. (See the article on ► Measure Preserving Systems.) The resulting stochastic process is a (one-step, finite-state) *Markov process*. If p and P also satisfy

$$pP = p \,, \tag{3}$$

then the Markov process is stationary. In this case we call the (one or two-sided) measure-preserving system the *Markov shift* determined by p and P.

Aperiodic and irreducible Markov chains (those for which a power of the transition matrix P has all entries positive) are strongly mixing, in fact are isomorphic to Bernoulli shifts (usually by means of a complicated measure-preserving recoding).

More generally we say a set $\Lambda \subset A^{\mathbb{Z}}$ is a *subshift* if it is compact and invariant under σ. A subshift Λ is said to be of *finite type* (SFT) if there exists an $d \times d$ matrix $M = (a_{ij})$ such that all entries are 0 or 1 and $x \in \Lambda$ if and only if $a_{x_i x_{i+1}} = 1$ for all $i \in \mathbb{Z}$. Shifts of finite type are also called *topological Markov chains*. There are many invariant measures for a non-trivial shift of finite type. For example the orbit of each periodic point is the support of an invariant measure. An important role in the theory, derived from motivations of statistical mechanics, is played by *equilibrium measures* (or *equilibrium states*) for continuous functions $\phi\colon \Lambda \to \mathbb{R}$, i. e. those measures μ which maximize $\{h_\sigma(\mu) + \int_\Lambda \phi \, \mathrm{d}\mu\}$ over all shift-invariant probability measures, where $h_\sigma(\mu)$ is the measure-theoretic entropy of σ with respect to μ. The study of full shifts or shifts of finite type has played a prominent role in the development of the hyperbolic theory of dynamical systems as physical systems with 'chaotic' dynamics 'typically' possess an invariant set with induced dynamics topologically conjugate to a shift of finite type (see the discussion by Smale in p. 147 in [108]). Dynamical systems in which there are transverse homoclinic connections are a common example (Theorem 5.3.5 in [45]). Furthermore in certain settings positive metric entropy implies the existence of shifts of finite type. One result along these lines is a theorem of Katok [53]. Let $h_{top}(f)$ denote the topological entropy of a map f and $h_\mu(f)$ denote metric entropy with respect to an invariant measure μ.

Theorem 1 (Katok) *Suppose $T\colon M \to M$ is a $C^{1+\epsilon}$ diffeomorphism of a closed manifold and μ is an invariant measure with positive metric entropy (i. e. $h_\mu(T) > 0$). Then for any $0 < \epsilon < h_\mu(T)$ there exists an invariant set Λ topologically conjugate to a transitive shift of finite type with $h_{top}(T|\Lambda) > h_\mu(T) - \epsilon$.*

More Examples of Subshifts

We consider some further examples of systems that are given by the shift transformation on a subset of the set of (usually doubly-infinite) sequences on a finite alphabet, usually $\{0, 1\}$. Associated with each subshift is its *language*, the set of all finite blocks seen in all sequences in the subshift. These languages are *extractive* (or *factorial*) (every subword of a word in the language is also in the language) and *insertive* (or *extendable*) (every word in the language extends on both sides to longer words in the language). In fact these two properties characterize the languages (subsets of the set of finite-length words on an alphabet) associated with subshifts.

Prouhet–Thue–Morse An interesting (and often rediscovered) element of $\{0, 1\}^{\mathbb{Z}_+}$ is produced as follows. Start with 0 and at each stage write down the opposite ($0' = 1, 1' = 0$) or mirror image of what is available so far. Or, repeatedly apply the *substitution* $0 \to 01, 1 \to 10$.

0

0 1

0 1 10

0 1 10 0110

⋮

The nth entry is the sum, mod 2, of the digits in the dyadic expansion of n. Using Keane's *block multiplication* [55] according to which if B is a block, $B \times 0 = B, B \times 1 = B'$, and $B \times (\omega_1 \cdots \omega_n) = (B \times \omega_1) \cdots (B \times \omega_n)$, we may also obtain this sequence as

$$0 \times 01 \times 01 \times 01 \times \cdots \,.$$

The orbit closure of this sequence is uniquely ergodic (there is a unique shift-invariant Borel probability measure, which is then necessarily ergodic). It is isomorphic to a skew product (see Sect. "Skew Products") over the von Neumann-Kakutani adding machine, or odometer (see Sect. "Adding Machines"). *Generalized Morse systems*, that is, orbit closures of sequences like $0 \times 001 \times 001 \times 001 \times \cdots$, are also isomorphic to skew products over compact group rotations.

Chacon System This is the orbit closure of the sequence generated by the substitution $0 \to 0010, 1 \to 1$. It is uniquely ergodic and is one of the first systems shown to be weakly mixing but not strongly mixing. It is *prime* (has no nontrivial factors) [30], and in fact has *minimal self joinings* [31]. It also has a nice description by means of cutting up the unit interval and stacking the pieces, using spacers (see Sect. "Cutting and Stacking"). This system has singular spectrum. It is not known whether or not its Cartesian square is loosely Bernoulli.

Sturmian Systems Take the orbit closure of the sequence $\omega_n = \chi_{[1-\alpha,1)}(n\alpha)$, where α is irrational. This is a uniquely ergodic system that is isomorphic to rotation by α on the unit interval. These systems have *minimal complexity* in the sense that the number of n-blocks grows as slowly as possible $(n+1)$ [29].

Toeplitz Systems A bi-infinite sequence (x_i) is a Toeplitz sequence if the set of integers can be decomposed into arithmetic progressions such that each x_i is constant on each arithmetic progression. A shift space X is a Toeplitz shift if it is the closure of the orbit of a Toeplitz sequence. It is possible to construct Toeplitz shifts which are uniquely ergodic and isomorphic to a rotation on a compact abelian group [34].

Sofic Systems These are images of SFT's under continuous factor maps (finite codes, or block maps). They correspond to *regular languages*—languages whose words are recognizable by finite automata. These are the same as the languages defined by *regular expressions*—finite expressions built up from $\emptyset$ (empty set), ϵ (empty word), + (union of two languages), $\cdot$ (all concatenations of words from two languages), and * (all finite concatenations of elements). They also have the characteristic property that the family of all *follower sets* of all blocks seen in the system is a finite family; similarly for *predecessor sets*. These are also generated by *phase-structure grammars* which are *linear*, in the sense that every production is either of the form $A \to Bw$ or $A \to w$, where A and B are variables and w is a string of terminals (symbols in the alphabet of the language).

(A *phase-structure grammar* consists of alphabets V of *variables* and A of *terminals*, a set of *productions*, which is finite set of pairs of words (α, w), usually written $\alpha \to w$, of words on $V \cup A$, and a *start symbol* S. The associated language consists of all words on the alphabet A of terminals which can be made by starting with S and applying a finite sequence of productions.)

Sofic systems typically support many invariant measures (for example they have many periodic points) but topologically transitive ones (those with a dense orbit) have a unique measures of maximal entropy (see [65]).

Context-free Systems These are generated by phase-structure grammars in which all productions are of the form $A \to w$, where A is a variable and w is a string of variables and terminals.

Coded Systems These are systems all of whose blocks are concatenations of some (finite or infinite) list of blocks. These are the same as the closures of increasing sequences of SFT's [60]. Alternatively, they are the closures of the images under finite edge-labelings of irreducible countable-state topological Markov chains. They need not be context-free. Squarefree languages are not coded, in fact do not contain any coded systems of positive entropy. See [13,14,15].

Smooth Expanding Interval Maps

Take $X = [0,1)(\text{mod } 1)$, $m \in \mathbb{N}$, $m > 1$ and define

$$T(x) = (mx).$$

Then T preserves Lebesgue measure μ (recall that T preserves μ if $\mu(T^{-1}A) = \mu(A)$ for all $A \in \mathcal{B}$). Furthermore it can be shown that T is ergodic.

This simple map exemplifies many of the characteristics of systems with some degree of hyperbolicity. It is isomorphic to a Bernoulli shift. The map has positive topological entropy and exponential divergence of nearby orbits, and Hölder functions have exponential decay of correlations and satisfy the central limit theorem and other strong statistical properties [20].

If $m = 2$ the system is isomorphic to a model of tossing a fair coin, which is a common example of randomness. To see this let $\mathcal{P} = \{P_0 = [0, 1/2), P_1 = [1/2, 1]\}$ be a partition of $[0,1]$ into two subintervals. We code the orbit under T of any point $x \in [0,1)$ by 0's and 1's by letting $x_k = i$ if $T^k x \in P_i$, $k = 0, 1, 2, \ldots$. The map $\phi\colon X \to \{0,1\}^{\mathbb{N}}$ which associates a point x to its *itinerary* in this way is a measure-preserving map from (X, μ) to $\{0,1\}^{\mathbb{N}}$ equipped with the Bernoulli measure from $p_0 = p_1 = \frac{1}{2}$. The map ϕ satisfies $\phi \circ T = \sigma \circ \phi$, μ a.e. and is invertible a.e., hence is an isomorphism. Furthermore, reading the binary expansion of x is equivalent to following the orbit of x under T and noting which element of the partition $\mathcal{P}$ is entered at each time. Borel's theorem on normal numbers (base m) may be seen as a special case of the Birkhoff Ergodic Theorem in this setting.

Piecewise C^2 Expanding Maps The main statistical features of the examples in Sect. "Smooth Expanding Interval Maps" generalize to a broader class of expanding maps of the interval. For example:

Let $X = [0, 1]$ and let $\mathcal{P} = \{I_1, \ldots, I_n\}$ $(n \geq 2)$ be a partition of X into intervals (closed, half-open or open) such that $I_i \cap I_j = \emptyset$ if $i \neq j$. Let I_i^o denote the interior of I_i. Suppose $T\colon X \to X$ satisfies:

(a) For each $i = 1, \ldots, n$, $T|I_i$ has a C^2 extension to the closure $\bar{I}_i$ of I_i and $|T'(x)| \geq \alpha > 1$ for all $x \in I_i^o$.
(b) $T(I_j) = \cup_{i \in P_j} I_i$ Lebesgue a.e. for some non-empty subset $P_j \subset \{1, \ldots, n\}$.
(c) For each I_j there exists n_j such that $T^{n_j}(I_j) = [0, 1]$ Lebesgue a.e.

Then T has an invariant measure μ which is absolutely continuous with respect to Lebesgue measure m, and there exists $C > 0$ such that $\frac{1}{C} \leq \frac{d\mu}{dm} \leq C$. Furthermore T is ergodic with respect to μ and displays the same statistical properties listed above for the C^2 expanding maps [20]. (See the "Folklore Theorem" in the article on ► Measure Preserving Systems.)

More Interval Maps

Continued Fraction Map This is the map $T\colon [0, 1] \to [0, 1]$ given by $Tx = 1/x \bmod 1$, and it corresponds to the shift $[0; a_1, a_2, \ldots] \to [0; a_2, a_3, \ldots]$ on the continued fraction expansions of points in the unit interval (a map on $\mathbb{N}^{\mathbb{N}}$). It preserves a unique finite measure equivalent to Lebesgue measure, the *Gauss measure* $dx/(\log 2)(1 + x)$. It is Bernoulli with entropy $\pi^2/6 \log 2$ (in fact the natural partition into intervals is a weak Bernoulli generator, for the definition and details see [91]). By using the Ergodic Theorem, Khintchine and Lévy showed that

$$(a_1 \cdots a_n)^{1/n} \to \prod_{k=1}^{\infty} \left[1 + \frac{1}{k^2 + 2k}\right]^{\frac{\log k}{\log 2}} \text{ a.e. as } n \to \infty\,;$$

$$\text{if } [0; a_1, \ldots, a_n] = \frac{p_n}{q_n}, \text{ then } \frac{1}{n} \log q_n \to \frac{\pi^2}{12 \log 2} \text{ a.e.}\,;$$

$$\frac{1}{n} \log \left| x - \frac{p_n(x)}{q_n(x)} \right| \to \frac{\pi^2}{6 \log 2} \text{ a.e.}\,;$$

and if m is Lebesgue measure (or any equivalent measure) and μ is Gauss measure, then for each interval I, $m(T^{-n} I) \to \mu(I)$, in fact exponentially fast, with a best constant $0.30366\ldots$ see [10,75].

The Farey Map This is the map $U\colon [0, 1] \to [0, 1]$ given by $Ux = x/(1 - x)$ if $0 \leq x \leq 1/2$, $Ux = (1 - x)/x$ if $1/2 \leq x \leq 1$. It is ergodic for the σ-finite infinite measure dx/x (Rényi and Parry). It is also ergodic for the *Minkowski measure* d, which is a measure of maximal entropy. This map corresponds to the shift on the *Farey tree* of rational numbers which provide the *intermediate convergence* (best one-sided) as well as the continued fraction (best two-sided) rational approximations to irrational numbers. See [61,62].

f-Expansions Generalizing the continued fraction map, let $f\colon [0, 1] \to [0, 1]$ and let $\{I_n\}$ be a finite or infinite partition of $[0, 1]$ into subintervals. We study the map f by coding itineraries with respect to the partition $\{I_n\}$. For many examples, absolutely continuous (with respect to Lebesgue measure) invariant measures can be found and their dynamical properties determined. See [104].

β-Shifts This is the special case of f-expansions when $f(x) = \beta x \bmod 1$ for some fixed $\beta > 1$. This map of the interval is called the *β-transformation*. With a proper choice of partition, it is represented by the shift on a certain subshift of the set of all sequences on the alphabet $D = \{0, 1, \ldots, \lfloor \beta \rfloor\}$, called the *$\beta$-shift*. A point x is expanded as an infinite series in negative powers of β with coefficients from this set; $d_\beta(x)_n = \lfloor \beta f^n(x) \rfloor$. (By convention terminating expansions are replaced by eventually periodic ones.) A one-sided sequence on the alphabet D is in the β-shift if and only if all of its shifts are lexicographically less than or equal to the expansion $d_\beta(1)$ of 1 base β. A one-sided sequence on the alphabet D is the valid expansion of 1 for some β if and only if it lexicographically dominates all its shifts. These were first studied by Bissinger [11], Everett [35], Rényi [93] and Parry [84,85]; there are good summaries by Bertrand-Mathis [9] and Blanchard [12].

For $\beta = \frac{1+\sqrt{5}}{2}$, $d_\beta(1) = 10101010\ldots$.
For $\beta = \frac{3}{2}$, $d_\beta(1) = 101000001\ldots$ (not eventually periodic).
Every β-shift is coded.
The topological entropy of a β-shift is $\log \beta$. There is a unique measure of maximal entropy $\log \beta$.
A β-shift is a shift of finite type if and only if the β-expansion of 1 is finite. It is sofic if and only if the expansion of 1 is eventually periodic. If β is a Pisot–Vijayaragavhan number (algebraic integer all of whose conjugates have modulus less than 1), then the β-shift is sofic. If the β-shift is sofic, then β is a Perron number (algebraic integer of maximum modulus among its conjugates).

Theorem 2 (Parry [86]) *Every strongly transitive (for every nonempty open set U, $\cup_{n>0} T^n U = X$) piecewise mono-*

tonic map on $[0,1]$ *is topologically conjugate to a* β*-transformation.*

Gaussian Systems

Consider a real-valued stationary process $\{f_k\colon -\infty < k < \infty\}$ on a probability space $(\Omega, \mathcal{F}, P)$. The process (and the associated measure-preserving system consisting of the shift and a shift-invariant measure on $\mathbb{R}^{\mathbb{Z}}$) is called *Gaussian* if for each $d \geq 1$, any d of the f_k form an $\mathbb{R}^d$-valued Gaussian random variable on Ω: this means that with $E(f_k) = m$ for all k and

$$A_{ij} = \int_\Omega (f_{k_i} - m)(f_{k_j} - m)\mathrm{d}P = C(k_i - k_j) \qquad \text{for } i, j = 1, \ldots, d\,,$$

where $C(\cdot)$ is a function, for each Borel set $B \subset \mathbb{R}$,

$$P\{\omega\colon (f_{k_1}(\omega), \ldots, f_{k_d}(\omega)) \in B\} = \frac{1}{2\pi^{d/2}\sqrt{\det A}} \int_B \exp\left[-\frac{1}{2}(x - (m, \ldots, m))^{tr} \cdot A^{-1}(x - (m, \ldots, m))\right] \mathrm{d}x_1 \cdots \mathrm{d}x_d\,.$$

where A is a matrix with entries (A_{ij}). The function $C(k)$ is positive semidefinite and hence has an associated measure σ on $[0, 2\pi)$ such that

$$C(k) = \int_0^{2\pi} \mathrm{e}^{ikt}\,\mathrm{d}\sigma(t)\,.$$

Theorem 3 (de la Rue [33]) *The Gaussian system is ergodic if and only if the "spectral measure" σ is continuous (i. e., nonatomic), in which case it is also weakly mixing. It is mixing if and only if $C(k) \to 0$ as $|k| \to \infty$. If σ is singular with respect to Lebesgue measure, then the entropy is 0; otherwise the entropy is infinite.*

For more details see [28].

Hamiltonian Systems

(This paragraph is from the article on ► Measure Preserving Systems.) Many systems that model physical situations can be studied by means of Hamilton's equations. The state of the entire system at any time is specified by a vector $(q, p) \in \mathbb{R}^{2n}$, the *phase space*, with q listing the coordinates of the positions of all of the particles, and p listing the coordinates of their momenta. We assume there is a time-independent *Hamiltonian function* $H(q, p)$ such that the time development of the system satisfies *Hamilton's equations*:

$$\frac{\mathrm{d}q_i}{\mathrm{d}t} = \frac{\partial H}{\partial p_i}, \quad \frac{\mathrm{d}p_i}{\mathrm{d}t} = -\frac{\partial H}{\partial q_i}, \quad i = 1, \ldots, n\,. \tag{4}$$

Often in applications the Hamiltonian function is the sum of kinetic and potential energy:

$$H(q, p) = K(p) + U(q)\,. \tag{5}$$

Solving these equations with initial state (q, p) for the system produces a flow $(q, p) \to T_t(q, p)$ in phase space which moves (q, p) to its position $T_t(q, p)t$ units of time later. According to *Liouville's formula* Theorem 3.2 in [69], this flow preserves Lebesgue measure on $\mathbb{R}^{2n}$. Calculating $\mathrm{d}H/\mathrm{d}t$ by means of the Chain Rule

$$\frac{\mathrm{d}H}{\mathrm{d}t} = \sum_i \left(\frac{\partial H}{\partial p_i}\frac{\mathrm{d}p_i}{\mathrm{d}t} + \frac{\partial H}{\partial q_i}\frac{\mathrm{d}q_i}{\mathrm{d}t}\right)$$

and using Hamilton's equations shows that H is constant on orbits of the flow, and thus each set of constant energy $X(H_0) = \{(q, p)\colon H(q, p) = H_0\}$ is an invariant set. There is a natural invariant measure on a constant energy set $X(H_0)$ for the restricted flow, namely the measure given by rescaling the volume element $\mathrm{d}S$ on $X(H_0)$ by the factor $1/||\nabla H||$.

Billiard Systems These form an important class of examples in ergodic theory and dynamical systems, motivated by natural questions in physics, particularly the behavior of gas models. Consider the motion of a particle inside a bounded region D in $\mathbb{R}^d$ with piecewise smooth (C^1 at least) boundaries. In the case of planar billiards we have $d = 2$. The particle moves in a straight line with constant speed until it hits the boundary, at which point it undergoes a perfectly elastic collision with the angle of incidence equal to the angle of reflection and continues in a straight line until it next hits the boundary. It is usual to normalize and consider unit speed, as we do in this discussion for convenience. We take coordinates (x, v) given by the Euclidean coordinates in $x \in D$ together with a direction vector $v \in S^{d-1}$. A flow ϕ_t is defined with respect to Lebesgue almost every (x, v) by translating x a distance t defined by the direction vector v, taking account of reflections at boundaries. ϕ_t preserves a measure absolutely continuous with respect to Riemannian volume on (x, v) coordinates. The flow we have described is called a *billiard flow*. The corresponding *billiard map* is formed by taking the Poincaré map corresponding to the cross-section given by the boundary ∂D. We will describe the planar billiard map; the higher dimensional generalization is clear. The billiard map is a map $T\colon \partial D \to \partial D$, where ∂D is coordinatized by (s, θ), $s \in [0, L)$, where L is the length of ∂D and $\theta \in (0, \pi)$ measures the angle that inward pointing vec-

tors make with the tangent line to ∂D at s. Given a point (s, θ), the angle θ defines an oriented line $l(s, \theta)$ which intersects ∂D in two points s and s'. Reflecting l in the tangent line to ∂D at the point s' gives another oriented line passing through s' with angle θ' (measured with respect to the angular coordinate system based at s'). The billiard map is the map $T(s, \theta) = (s', \theta')$. T preserves a measure $\mu = \sin\theta \mathrm{d}s \times \mathrm{d}\theta$. The billiard flow may be modeled as a suspension flow over the billiard map (see Sect. "Suspension Flows").

If the region D is a polygon in the plane (or polyhedron in $\mathbb{R}^d$), then ∂D consists of the faces of the polyhedron. The dynamical behavior of the billiard map or flow in regions with only flat (non-curved) boundaries is quite different to that of billiard flows or maps in regions D with strictly convex or strictly concave boundaries. The topological entropy of a flat polygonal billiard is zero. Research interest focuses on the existence and density of periodic or transitive orbits. It is known that if all the angles between sides are rational multiples of π then there are periodic orbits [17,74,112] and they are dense in the phase space [16]. It is also known that a residual set of polygonal billiards are topologically transitive and ergodic [58,117].

On the other hand, billiard maps in which ∂D has strictly convex components are physical examples of non-uniformly hyperbolic systems (with singularities). The meaning of concave or convex varies in the literature. We will consider a billiard flow inside a circle to be a system with a strictly concave boundary, while a billiard flow on the torus from which a circle has been excised to be a billiard flow with strictly convex boundary.

The class of billiards with some strictly convex boundary components, sometimes called *dispersing billiards* or *Sinai billiards*, was introduced by Sinai [106] who proved many of their fundamental properties. Lazutkin [63] proved that planar billiards with generic strictly concave boundary are not ergodic. Nevertheless Bunimovich [22,23] produced a large of billiard systems, *Bunimovich billiards*, with strictly concave boundary segments (perhaps with some flat boundaries as well) which were ergodic and non-uniformly hyperbolic. For more details see [26,54,66,109]. We will discuss possibly the simplest example of a dispersing billiard, namely a toral billiard with a single convex obstacle. Take the torus T^2 and consider a single strictly convex subdomain S with C^∞ boundary. The domain of the billiard map is $[0, L) \times (0, \pi)$, where L is the length of ∂S. The measure $\sin(\theta)\mathrm{d}s \times \mathrm{d}\theta$ is preserved. If the curvature of ∂S is everywhere non-zero, then the billiard map T has positive topological entropy, periodic points are dense, and in fact the system is isomorphic to a Bernoulli shift [41].

KAM-Systems and Stably Non-Ergodic Behavior A celebrated theorem of Kolmogorov, Arnold and Moser (the KAM theorem) implies that the set of ergodic area-preserving diffeomorphisms of a compact surface without boundary is not dense in the C^r topology for $r \geq 4$. This has important implications, in that there are natural systems in which ergodicity is not generic. The constraint of perturbing in the class of area-preserving diffeomorphisms is an appropriate imposition in many physical models. We will take the version of the KAM theorem as given in Theorem 5.1 in [69] (original references include [3,59,79]). An elliptic fixed point for an area-preserving diffeomorphism T of a surface M is called a *non-degenerate elliptic fixed point* if there is a local C^r, $r \geq 4$, change of coordinates h so that in polar coordinates

$$hTh^{-1}(r, \theta) = (r, \theta + \alpha_0 + \alpha_1 r) + F(r, \theta) ,$$

where all derivatives of F up to order 3 vanish, $\alpha_1 \neq 0$ and $\alpha_0 \neq 0, \frac{\pm\pi}{2}, \pi, \frac{\pm 2\pi}{3}$. A map of the form

$$\tau(r, \theta) = (r, \theta + \alpha_0 + \alpha_1 r) ,$$

where $\alpha_1 \neq 0$, is called a *twist map*. Note that a twist map leaves invariant the circle $r = k$, for any constant k, and rotates each invariant curve by a rigid rotation $\alpha_1 r$, the magnitude of the rotation depending upon r. With respect to two-dimensional Lebesgue measure a twist map is certainly not ergodic.

Theorem 4 *Suppose T is a volume-preserving diffeomorphism of class C^r, $r \geq 4$, of a surface M. If x is a non-degenerate elliptic fixed point, then for every $\epsilon > 0$ there exists a neighborhood U_ϵ of x and a set $U_{0,\epsilon} \subset U_\epsilon$ with the properties:*

(a) $U_{0,\epsilon}$ is a union of T-invariant simple closed curves of class C^{r-1} containing x in their interior.
(b) The restriction of T to each such invariant curve is topologically conjugate to an irrational rotation.
(c) $m(U_\epsilon - U_{0,\epsilon}) \leq \epsilon m(U_\epsilon)$, where m is Lebesgue measure on M.

It is possible to prove the existence of a C^r volume preserving diffeomorphism ($r \geq 4$) with a non-degenerate elliptic fixed point and also show that if T possesses a non-degenerate elliptic fixed point then there is a neighborhood V of T in the C^r topology on volume-preserving diffeomorphisms such that each $T' \in V$ possesses a non-degenerate elliptic fixed point Chapter II, Sect. 6 in [69]. As a corollary we have

Corollary 1 *Let M be a compact surface without boundary and* $\mathrm{Diff}^r(M)$ *the space of C^r area-preserving diffeomorphisms with the C^r topology. Then the set of $T \in \mathrm{Diff}^r(M)$ which are ergodic with respect to the probability measure determined by normalized area is not dense in* $\mathrm{Diff}^r(M)$ *for* $r \geq 4$.

Smooth Uniformly Hyperbolic Diffeomorphisms and Flows

Time series of measurements on deterministic dynamical systems sometimes display limit laws exhibited by independent identically distributed random variables, such as the central limit theorem, and also various mixing properties. The models of hyperbolicity we discuss in this section have played a key role in showing how this phenomenon of 'chaotic behavior' arises in deterministic dynamical systems. Hyperbolic sets and their associated dynamics have also been pivotal in studies of structural stability. A smooth system is C^r *structurally stable* if a small perturbation in the C^r topology gives rise to a system which is topologically conjugate to the original. When modeling a physical system it is desirable that slight changes in the modeling parameters do not greatly affect the qualitative or quantitative behavior of the ensemble of orbits considered as a whole. The orbit of a point may change drastically under perturbation (especially if the system has sensitive dependence on initial conditions) but the collection of all orbits should ideally be 'similar' to the original unperturbed system. In the latter case one would hope that statistical properties also vary only slightly under perturbation. Structural stability is one, quite strong, notion of stability. The conclusion of a body of work on structural stability is that a system is C^1 structurally stable if and only if it is uniformly hyperbolic and satisfies a technical assumption called strong transversality (see below for details).

Suppose M is a C^1 compact Riemannian manifold equipped with metric d and tangent space TM with norm $\|\,\|$. Suppose also that $U \subset M$ is a non-empty open subset and $T\colon U \to T(U)$ is a C^1 diffeomorphism. A compact T invariant set $\Lambda \subset U$ is called a *hyperbolic set* if there is a splitting of the tangent space T_pM at each point $p \in \Lambda$ into two invariant subspaces, $T_pM = E^{\mathrm{u}}(p) \oplus E^{\mathrm{s}}(p)$, and a number $0 < \lambda < 1$ such that for $n \geq 0$

$$\|D_pT^n v\| \leq C\lambda^n \|v\| \text{ for } v \in E^{\mathrm{s}}(p)\,,$$

$$\|D_pT^{-n} v\| < C\lambda^n \|v\| \text{ for } v \in E^{\mathrm{u}}(p)\,.$$

The subspace E^{u} is called the *unstable* or *expanding subspace* and the subspace E^{s} the *stable* or *contracting subspace*. The stable and unstable subspaces may be integrated to produce *stable* and *unstable manifolds*

$$W^{\mathrm{s}}(p) = \{y\colon d(T^n p, T^n y) \to 0\} \quad \text{as } n \to \infty\,,$$

$$W^{\mathrm{u}}(p) = \{y\colon d(T^{-n} p, T^{-n} y) \to 0\} \quad \text{as } n \to \infty\,.$$

The stable and unstable manifolds are immersions of Euclidean spaces of the same dimension as $E^{\mathrm{s}}(p)$ and $E^{\mathrm{u}}(p)$, respectively, and are of the same differentiability as T. Moreover, $T_p(W^{\mathrm{s}}(p)) = E^{\mathrm{s}}(p)$ and $T_p(W^{\mathrm{u}}(p)) = E^{\mathrm{u}}(p)$. It is also useful to define *local stable manifolds* and *local unstable manifolds* by

$$W^{\mathrm{s}}_\epsilon(p) = \{y \in W^{\mathrm{s}}(p)\colon d(T^n p, T^n y) < \epsilon\} \text{ for all } n \geq 0\,,$$

$$W^{\mathrm{u}}_\epsilon(p) = \{y \in W^{\mathrm{u}}(p)\colon d(T^{-n} p, T^{-n} y) < \epsilon\} \quad \text{for all } n \geq 0\,.$$

Finally we discuss the notion of strong transversality. We say a point x is *non-wandering* if for each open neighborhood U of x there exists an $n > 0$ such that $T^n(U) \cap U \neq \emptyset$. The NW set of non-wandering points is called the *non-wandering set*. We say a dynamical system has the *strong transversal property* if $W^{\mathrm{s}}(x)$ intersects $W^{\mathrm{u}}(y)$ transversely for each pair of points $x, y \in NW$. In the C^r, $r \geq 1$ topology Robbin [94], de Melo [78] and Robinson [95,96] proved that dynamical systems with the strong transversal property are structurally stable, and Robinson [97] in addition showed that strong transversality was also necessary. Mañé [70] showed that a C^1 structurally stable diffeomorphism must be uniformly hyperbolic and Hayashi [47] extended this to flows. Thus a C^1 diffeomorphism or flow on a compact manifold is structurally stable if and only it is uniformly hyperbolic and satisfies the strong transversality condition.

Geodesic Flow on Manifold of Negative Curvature The study of the geodesic flow on manifolds of negative sectional curvature by Hedlund and Hopf was pivotal to the development of the ergodic theory of hyperbolic systems. Suppose that M is a geodesically complete Riemannian manifold. Let $\gamma_{p,v}(t)$ be the geodesic with $\gamma_{p,v}(0) = p$ and $\dot\gamma_{p,v}(0) = v$, where $\dot\gamma_{p,v}$ denotes the derivative with respect to time t. The geodesic flow is a flow ϕ_t on the tangent bundle TM of M, $\phi_t\colon \mathbb{R} \times TM \to TM$, defined by

$$\phi_t(p, v) = (\gamma_{p,v}(t), \dot\gamma_{p,v}(t))\,.$$

where $(p, v) \in TM$. Since geodesics have constant speed, if $\|v\| = 1$ then $\|\gamma_{p,v}(t)\| = 1$ for all t, and thus the unit tangent bundle $T^1M = \{(p, v) \in TM\colon \|v\| = 1\}$ is preserved

under the geodesic flow. The geodesic flow and its restriction to the unit tangent bundle both preserve a volume form, Liouville measure. In 1934 Hedlund [48] proved that the geodesic flow on the unit tangent bundle of a surface of strictly negative constant sectional curvature is ergodic, and in 1939 Hopf [49] extended this result to manifolds of arbitrary dimension and strictly negative (not necessarily constant) curvature. Hopf's technique of proof of ergodicity (Hopf argument) was extremely influential and used the foliation of the tangent space into stable and unstable manifolds. For a clear exposition of this technique, and the property of absolute continuity of the foliations into stable and unstable manifolds, see [66]. The geodesic flow on manifolds of constant negative sectional curvature is an Anosov flow (see Sect. "Anosov Systems"). We remark that for surfaces sectional curvature is the same as Gaussian curvature. Recently the time-one map of the geodesic flow on the unit tangent bundle of a surface with constant negative curvature, which is a partially hyperbolic system (see Sect. "Partially Hyperbolic Dynamical Systems"), was shown to be stably ergodic [44], so the geodesic flow is still playing a major role in the development of ergodic theory.

Horocycle Flow All surfaces endowed with a Riemannian metric of constant negative curvature are quotients of the upper half-plane $\mathcal{H}^+ := \{x + iy \in \mathbb{C} : y > 0\}$ with the metric $\mathrm{d}s^2 = \frac{\mathrm{d}x^2+\mathrm{d}y^2}{y^2}$, whose sectional curvature is -1. The orientation-preserving isometries of this metric are exactly the linear fractional (also known as Möbius) transformations:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2,\mathbb{R}), \quad [z \in \mathcal{H}^+ \mapsto \frac{az+b}{cz+d} \in \mathcal{H}^+] .$$

Since each matrix $\pm I$ corresponds to the identity transformation, we consider matrices in $PSL(2,\mathbb{R}) := SL(2,\mathbb{R})/\{\pm I\}$.

The unit tangent bundle, $S\mathcal{H}^+$, of the upper half-plane can be identified with $PSL(2,\mathbb{R})$. Then the geodesic flow corresponds to the transformations

$$t \in \mathbb{R} \mapsto \begin{pmatrix} \mathrm{e}^t & 0 \\ 0 & \mathrm{e}^{-t} \end{pmatrix}$$

seen as acting on $PSL(2,\mathbb{R})$. The unstable foliation of an element $A \in PSL(2,\mathbb{R}) \cong S\mathcal{H}^+$ is given by

$$\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} A, \quad t \in \mathbb{R} ,$$

and the flow along this foliation, given by

$$t \in \mathbb{R} \mapsto \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} ,$$

is called the *horocycle flow*. Similarly for the flow induced on the unit tangent bundle of each quotient of the upper half-plane by a discrete group of linear fractional transformations.

The geodesic and horocycle flows acting on a (finite-volume) surface of constant negative curvature form the fundamental example of a transverse pair of actions. The geodesic flow often has many periodic orbits and many invariant measures, has positive entropy, and is in fact Bernoulli with respect to the natural measure [82], while the horocycle flow is often uniquely ergodic [39,72] and of entropy zero, although mixing of all orders [73]. See [46] for more details.

Markov Partitions and Coding If $(X, T, \mathcal{B}, \mu)$ is a dynamical system then a finite partition of X induces a coding of the orbits and a semi-conjugacy with a subshift on a symbol space (it may not of course be a full conjugacy). For hyperbolic systems a special class of partitions, *Markov partitions*, induce a conjugacy for the invariant dynamics to a subshift of finite type. A Markov partition $\mathcal{P}$ for an invariant subset Λ of a diffeomorphism T of a compact manifold M is a finite collection of sets R_i, $1 \le i \le n$ called *rectangles*. The rectangles have the property, for some $\epsilon > 0$, if $x, y \in R_i$ then $W^{\mathrm{s}}_\epsilon(x) \cap W^{\mathrm{u}}_\epsilon(y) \in R_i$. This is sometimes described as being closed under *local product structure*. We let $W^{\mathrm{u}}(x, R_i)$ denote $W^{\mathrm{u}}_\epsilon(x) \cap R_i$ and $W^{\mathrm{s}}(x, R_i)$ denote $W^{\mathrm{s}}_\epsilon(x) \cap R_i$. Furthermore we require for all i, j:

(1) Each R_i is the closure of its interior.
(2) $\Lambda \subset \cup_i R_i$
(3) $R_i \cap R_j = \partial R_i \cap \partial R_j$ if $i \neq j$
(4) if $x \in R_i^o$ and $T(x) \in R_j^o$ then $W^{\mathrm{u}}(T(x), R_j) \subset T(W^{\mathrm{u}}(x, R_i))$ and $W^{\mathrm{s}}(x, R_i) \subset T^{-1}(W^{\mathrm{u}}(T(x), R_j))$

Anosov Systems An *Anosov diffeomorphism* [2] is a uniformly hyperbolic system in which the entire manifold is a hyperbolic set. Thus an Anosov diffeomorphism is a C^1 diffeomorphism T of M with a DT-invariant splitting (which is a continuous splitting) of the tangent space $TM(x)$ at each point p into a disjoint sum

$$T_pM = E^{\mathrm{u}}(p) \oplus E^{\mathrm{s}}(p)$$

and there exist constants $0 < \lambda < 1$, constant $C > 0$ such that $\|DT^n v\| < C\lambda^n \|v\|$ for all $v \in E^{\mathrm{s}}(p)$ and $\|DT^{-n} w\| \le C\lambda^n \|w\|$ for all $w \in E^{\mathrm{u}}(p)$.

A similar definition holds for Anosov flows $\phi : \mathbb{R} \times M \to M$. A flow is *Anosov* if there is a splitting of the tangent bundle into flow-invariant subspaces $E^{\mathrm{u}}, E^{\mathrm{s}}, E^{\mathrm{c}}$ so

$D_p\phi_t E_p^s = E_{\phi_t(p)}^s$, $D_p\phi_t E_p^u = E_{\phi_t(p)}^u$ and $D_p\phi_t E_p^c = E_{\phi_t(p)}^c$, and at each point $p \in M$

$$T_pM = E_p^s \oplus E_p^u \oplus E_p^c$$
$$\|(D_p\phi_t)v\| < C\lambda^t\|v\| \text{ for } v \in E^s(p)$$
$$\|(D_p\phi_{-t})v\| < C\lambda^t\|v\| \text{ for } v \in E^u(p)$$

for some $0 < \lambda < 1$. The tangent to the flow direction $E^c(p)$ is a neutral direction:

$$\|(D_p\phi_t)v\| = \|v\| \text{ for } v \in E^c(p)\,.$$

Anosov proved that Anosov flows and diffeomorphisms which preserve a volume form are ergodic [2] and are also structurally stable. Sinai [105] constructed Markov partitions for Anosov diffeomorphisms and hence coded trajectories via a subshift of finite type. Using ideas from statistical physics in [107] Sinai constructed Gibbs measures for Anosov systems. An SRB measure (see Sect. "Physically Relevant Measures and Strange Attractors") is a type of Gibbs measure corresponding to the potential $-\log|\det(DT|_{E^u})|$ and is characterized by the property of absolutely continuous conditional measures on unstable manifolds.

The simplest examples of Anosov diffeomorphisms are perhaps the two-dimensional hyperbolic toral automorphisms (the $n > 2$ generalization is clear). Suppose A is a 2×2 matrix with integer entries

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

such that $\det(A) = 1$ and A has no eigenvalues of modulus 1. Then A defines a transformation of the two-dimensional torus $T^2 = S^1 \times S^1$ such that if $v \in T^2$,

$$v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}\,,$$

then

$$Av = \begin{pmatrix} (av_1 + bv_2) \\ (cv_1 + dv_2) \end{pmatrix}\,.$$

A preserves Lebesgue (or Haar) measure and is ergodic. A prominent example of such a matrix is

$$\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\,,$$

which is sometimes called the Arnold Cat Map. Each point with rational coordinates $(p_1/q_1, p_2/q_2)$ is periodic. There are two eigenvalues $1/\lambda < 1 < \lambda = (3+\sqrt{5})/2$ with orthogonal eigenvectors, and the projections of the eigenspaces from $\mathbb{R}^2$ to T^2 are the stable and unstable subspaces.

Axiom A Systems In the case of Anosov diffeomorphisms the splitting into contracting and expanding bundles holds on the entire phase space M. A system $T\colon M \to M$ is an *Axiom A system* if the non-wandering set NW is a hyperbolic set and periodic points are dense in the non-wandering set. $NW \subset M$ may have Lebesgue measure zero. A set $\Lambda \subset M$ is *locally maximal* if there exists an open set U such that $\Lambda = \cap_{n\in\mathbb{Z}} T^n(U)$. The solenoid and horseshoe discussed below are examples of locally maximal sets. Bowen [18] constructed Markov partitions for Axiom A diffeomorphisms . Ruelle imported ideas from statistical physics, in particular the idea of an equilibrium state and the variational principle, to the study of Axiom A systems (see [101,102]) This work extended the notion of Gibbs measure and other ideas from statistical mechanics, introduced by Sinai for Anosov systems [107], into Axiom A systems.

One achievement of the Axiom A program was the Smale Decomposition Theorem, which breaks the dynamics of Axiom A systems into locally maximal sets and describes the dynamics on each [18,19,108].

Theorem 5 (Spectral Decomposition Theorem) *If T is Axiom A then there is a unique decomposition of the non-wandering set NW of T*

$$NW = \Lambda_1 \cup \cdots \cup \Lambda_k$$

as a disjoint union of closed, invariant, locally maximal hyperbolic sets Λ_i such that T is transitive on each Λ_i. Furthermore each Λ_i may be further decomposed into a disjoint union of closed sets Λ_i^j, $j = 1, \ldots, n_i$ such that T^{n_i} is topologically mixing on each Λ_i^j and T cyclically permutes the Λ_i^j.

Horseshoe Maps This type of map was introduced by Steven Smale in the 1960's and has played a pivotal role in the development of dynamical systems theory. It is perhaps the canonical example of an Axiom A system [108] and is conjugate to a full shift on 2 symbols. Let S be a unit square in $\mathbb{R}^2$ and let T be a diffeomorphism of S onto its image such that $S \cap T(S)$ consists of two disjoint horizontal strips S_0 and S_1. Think of stretching S uniformly in the horizontal direction and contracting uniformly in the vertical direction to form a long thin rectangle, and then bending the rectangle into the shape of a horseshoe and laying the straight legs of the horseshoe back on the unit square S. This transformation may be realized by a diffeomorphism and we may also require that T restricted to $T^{-1}S_i$, $i = 0, 1$, acts as a linear map. The restriction of T to the maximal invariant set $H = \cap_{i=-\infty}^{\infty} T^i S$ is a Smale horseshoe map. H is a Cantor set, the product of a Cantor

set in the horizontal direction and a Cantor set in the vertical direction. The conjugacy with the shift on two symbols is realized by mapping $x \in H$ to its itinerary with respect to the sets S_0 and S_1 under powers of T (positive and negative powers).

Solenoids The solenoid is defined on the solid torus X in $\mathbb{R}^3$ which we coordinatize as a circle of two-dimensional solid disks, so that

$$X = \{(\theta, z) \colon \theta \in [0,1) \text{ and } |z| \le 1, z \in \mathbb{C}\}$$

The transformation $T \colon X \to X$ is given by

$$T(\theta, z) = \left(2\theta \pmod 1\,, \frac{1}{4}z + \frac{1}{2}e^{2\pi i\theta}\right)$$

Geometrically the transformation stretches the torus to twice its length, shrinks its diameter by a factor of 4, then twists it and doubles it over, placing the resultant object without self-intersection back inside the original solid torus. $T(X)$ intersects each disk $D_c = \{(\theta, z) \colon \theta = c\}$ in two smaller disks of $\frac{1}{4}$ the diameter. The transformation T contracts volume by a factor of 2 upon each application, yet there is expansion in the θ direction ($\theta \to 2\theta$). The solenoid $A = \cap_{n \ge 0} T^n(X)$ has zero Lebesgue measure, is T-invariant and is (locally) topologically a line segment cross a two-dimensional Cantor set (A intersects each disk D_c in a Cantor set). The set A is an *attractor*, in that all points inside X limit under iteration by T upon A. $T \colon A \to A$ is an Axiom A system.

Partially Hyperbolic Dynamical Systems

Partially hyperbolic dynamical systems are a generalization of uniformly hyperbolic systems in that an invariant central direction is allowed but the contraction in the central direction is strictly weaker than the contraction in the contracting direction and the expansion in the central direction is weaker than the expansion in the expanding direction. More precisely, suppose M is a C^1 compact (adapted) Riemannian manifold equipped with metric d and tangent space TM with norm $\|\,\|$. A C^1 diffeomorphism T of M is a partially hyperbolic diffeomorphism if there is a nontrivial continuous DT invariant splitting of the tangent space T_pM at each point p into a disjoint sum

$$T_pM = E^{\mathrm{u}}(p) \oplus E^{\mathrm{c}}(p) \oplus E^{\mathrm{s}}(p)$$

and continuous positive functions $m, M, \tilde{\gamma}, \gamma$ such that

- E^{s} is contracted:
  if $v^{\mathrm{s}} \in E^{\mathrm{s}}(x) \setminus \{0\}$ then $\frac{\|D_pT^n v^{\mathrm{s}}\|}{\|v^{\mathrm{s}}\|} \le m(p) < 1$;
- E^{u} is expanded:
  if $v^{\mathrm{u}} \in E^{\mathrm{u}}(x) \setminus \{0\}$ then $\frac{\|D_pTv^{\mathrm{u}}\|}{\|v^{\mathrm{u}}\|} \ge M(p) > 1$;
- E^{c} is uniformly dominated by E^{u} and E^{s}:
  if $v^{\mathrm{c}} \in E^{\mathrm{c}}(x) \setminus \{0\}$ then there are numbers $\tilde{\gamma}(p), \gamma(p)$ such that $m(p) < \tilde{\gamma}(p) \le \frac{\|D_pTv^{\mathrm{c}}\|}{\|v^{\mathrm{c}}\|} \le \gamma(p) < M(p)$.

The notion of partial hyperbolicity was introduced by Brin and Pesin [21] who proved existence and properties, including absolute continuity, of invariant foliations in this setting. There has been intense recent interest in partially hyperbolic systems primarily because significant progress has been made in establishing that certain volume-preserving partially hyperbolic systems are 'stably ergodic'—that is, they are ergodic and under small (C^r topology) volume-preserving perturbations remain ergodic. This phenomenon had hitherto been restricted to uniformly hyperbolic systems. For recent developments, and precise statements, on stable ergodicity of partially hyperbolic systems see [24,92].

Compact Group Extensions of Uniformly Hyperbolic Systems A natural example of a partially hyperbolic system is given by a compact group extension of an Anosov diffeomorphism. If the following terms are not familiar see Sect. "Constructions" on standard constructions. Suppose that (T, M, μ) is an Anosov diffeomorphism, G is a compact connected Lie group and $h \colon M \to G$ is a differentiable map. The skew product $F \colon M \times G \to M \times G$ given by

$$F(x, g) = (Tx, h(x)g)$$

has a central direction in its tangent space corresponding to the Lie algebra LG of G (as a group element h acts isometrically on G so there is no expansion or contraction) and uniformly expanding and contracting bundles corresponding to those of the tangent space of $T \colon M \to M$. Thus $T(M \times G) = E^{\mathrm{u}} \oplus LG \oplus E^{\mathrm{s}}$.

Time-One Maps of Anosov Flows Another natural context in which partial hyperbolicity arises is in time-one maps of uniformly hyperbolic flows. Suppose $\phi_t \colon \mathbb{R} \times M \to M$ is an Anosov flow. The diffeomorphism $\phi_1 \colon M \to M$ is a partially hyperbolic diffeomorphism with central direction given by the flow direction. There is no expansion or contraction in the central direction.

Non-Uniformly Hyperbolic Systems

The assumption of uniform hyperbolicity is quite restrictive and few 'chaotic systems' found in applications are likely to exhibit uniform hyperbolicity. A natural weakening of this assumption, and one that is non-trivial and

greatly extends the applicability of the theory, is to require the hyperbolic splitting (no longer uniform) to hold only at almost every point of phase space. A systematic theory was built by Pesin [89,90] on the assumption that the system has non-zero Lyapunov exponents μ almost everywhere where μ is a Lebesgue equivalent invariant probability measure . Recall that a number λ is a *Lyapunov exponent* for $p \in M$ if $\|D_p T^n v\| \sim e^{\lambda n}$ for some unit vector $v \in T_p M$. Oseledet's theorem [83] (see also p. 232 in [113]), which is also called the Multiplicative Ergodic Theorem, implies that if T is a C^1 diffeomorphism of M then for any T-invariant ergodic measure μ almost every point has well-defined Lyapunov exponents. One of the highlights of Pesin theory is the following structure theorem: If $T\colon M \to M$ is a $C^{1+\epsilon}$ diffeomorphism with a T-invariant Lebesgue equivalent Borel measure μ such that T has non-zero Lyapunov exponents with respect to μ then T has at most a countable number of ergodic components $\{C_i\}$ on each of which the restriction of T is either Bernoulli or Bernoulli times a rotation (by which we mean the support of $\mu_i = \mu_{|C_i}$ consists of a finite number n_i of sets $\{S_1^i, \dots S_{n_i}^i\}$ cyclically permuted and T^{n_i} is Bernoulli when restricted to each S_j^i) [90,114]. This structure theorem has been generalized to SRB measures with non-zero Lyapunov exponents [64,90].

Physically Relevant Measures and Strange Attractors

(This paragraph is from the article on ► Measure Preserving Systems.) For Hamiltonian systems and other volume-preserving systems it is natural to consider ergodicity (and other statistical properties) of the system with respect to Lebesgue measure. In dissipative systems a measure equivalent to Lebesgue may not be invariant (for example the solenoid). Nevertheless Lebesgue measure has a distinguished role since sampling by experimenters is done with respect to Lebesgue measure. The idea of a physically relevant measure μ is that it determines the statistical behavior of a positive Lebesgue measure set of orbits, even though the support of μ may have zero Lebesgue measure. An example of such a situation in the uniformly hyperbolic setting is the solenoid Λ, where the attracting set Λ has Lebesgue measure zero and is (locally) topologically the product of a two-dimensional Cantor set and a line segment. Nevertheless Λ determines the behavior of all points in a solid torus in $\mathbb{R}^3$. More generally, suppose that $T\colon M \to M$ is a diffeomorphism on a compact Riemannian manifold and that m is a version of Lebesgue measure on M, given by a smooth volume form. Although Lebesgue measure m is a distinguished physically relevant measure, m may not be invariant under T, and the system may even be volume contracting in the sense that $m(T^n A) \to 0$ for all measurable sets A. Nevertheless an experimenter might observe long-term "chaotic" behavior whenever the state of the system gets close to some compact invariant set X which attracts a positive m-measure of orbits in the sense that these orbits limit on X. Possibly $m(X) = 0$, so that X is effectively invisible to the observer except through its effects on orbits not contained in X. The dynamics of T restricted to X can in fact be quite complicated—maybe a full shift, or a shift of finite type, or some other complicated topological dynamical system. Suppose there is a T-invariant measure μ supported on X such that for all continuous functions $\phi\colon M \to \mathbb{R}$

$$\frac{1}{n}\sum_{k=0}^{n-1} \phi \circ T^k(x) \to \int_X \phi \, \mathrm{d}\mu \,, \tag{6}$$

for a positive m-measure of points $x \in M$. Then the long-term equilibrium dynamics of an observable set of points $x \in M$ (i. e. a set of points of positive m measure) is described by (X, T, μ). In this situation μ is described as a *physical measure*. There has been a great deal of research on the properties of systems with attractors supporting physical measures.

In the dissipative non-uniformly hyperbolic setting the theory of 'physically relevant' measures is best developed in the theory of SRB (for Sinai, Ruelle and Bowen) measures. These dynamically invariant measures may be supported on a set of Lebesgue measure zero yet determine the asymptotic behavior of points in a set of positive Lebesgue measure.

If T is a diffeomorphism of M and μ is a T-invariant Borel probability measure with positive Lyapunov exponents which may be integrated to unstable manifolds, then we call μ an *SRB measure* if the conditional measure μ induces on the unstable manifolds is absolutely continuous with respect to the Riemannian volume element on these manifolds. The reason for this definition is technical but is motivated by the following observations. Suppose that the diffeomorphism has no zero Lyapunov exponents with respect to μ. Since T is a diffeomorphism, this implies T has negative Lyapunov exponents as well as positive Lyapunov exponents and corresponding local stable manifolds as well as local unstable manifolds. Suppose that a T-invariant set A consists of a union of unstable manifolds and is the support of an ergodic SRB measure μ and that $\phi\colon M \to \mathbb{R}$ is a continuous function. Since μ has absolutely continuous conditional measures on unstable manifolds with respect to conditional Lebesgue measure on the unstable manifolds, almost every point x in the

union of unstable manifolds U satisfies

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \phi \circ T^j(x) = \int \phi \, \mathrm{d}\mu \,. \tag{7}$$

If $y \in W^s_\epsilon(x)$ for such an $x \in U$ then $d(T^n x, T^n y) \to \infty$ and hence (7) implies

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \phi \circ T^j(y) = \int \phi \, \mathrm{d}\mu \,.$$

Furthermore, if the holonomy between unstable manifolds defined by sliding along stable manifolds is absolutely continuous (takes sets of zero Lebesgue measure on W^u to sets of zero Lebesgue measure on W^u), there is a positive Lebesgue measure of points (namely an unstable manifold and the union of stable manifolds through it) satisfying (7). Thus an SRB measure with absolutely continuous holonomy maps along stable manifolds is a physically relevant measure. If the stable foliation possesses this property it is called *absolutely continuous*. An Axiom A attractor for a C^2 diffeomorphism is an example of an SRB attractor [19,101,102,107]. The examples we have given of SRB measures and attractors and measures have been uniformly hyperbolic.

Recently much progress has been made in understanding the statistical properties of non-uniformly hyperbolic systems by using a tower (see Sect. "Induced Transformations") to construct SRB measures. We refer to Young's original papers [115,116], the book by Baladi [5] and to [114] for a recent survey on SRB measures in the non-uniformly hyperbolic setting.

Unimodal Maps Unimodal Maps of an interval are simple examples of non-uniformly hyperbolic dynamical systems that have played an important role in the development of dynamical systems theory. Suppose $I \subset \mathbb{R}$ is an interval; for simplicity we take $I = [0, 1]$. A *unimodal map* is a map $T\colon [0, 1] \to [0, 1]$ such that there exists a point $0 < c < 1$ and

- T is C^2;
- $T'(x) > 0$ for $x < c$, $T'(x) < 0$ for $x > c$;
- $T'(c) = 0$.

Such a map is clearly not uniformly expanding, as $|T'(x)| < 1$ for points in a neighborhood of c. The family of maps $T_\mu(x) = \mu x(1 - x)$, $0 < \mu \le 4$, is a family of unimodal maps with $c = 1/2$ and $T_2(1/2) = 1/2$, $T_4(1/2) = 1$.

We could have taken the interval I to be $[-1, 1]$ or indeed any interval with an obvious modification of the definition above. A well-studied family of unimodal maps in this setting is the *logistic family* $f_a\colon [-1, 1] \to [-1, 1]$, $f_a(x) = 1 - ax^2$, $a \in (0, 2]$. The families f_a and T_μ are equivalent under a smooth coordinate change, so statements about one family may be translated into statements about the other.

Unimodal maps are studied because of the insights they offer into transitions from regular or periodic to chaotic behavior as a parameter (e. g. μ or a) is varied, the existence of absolutely continuous measures, and rates of decay of correlations of regular observations for non-uniformly hyperbolic systems.

Results of Jakobson [52] and Benedicks and Carleson [6] implies that in the case of the logistic family there is a positive Lebesgue measure set of a such that f_a has an absolutely continuous ergodic invariant measure μ_a. It has been shown by Keller and Nowicki [57] (see alsoYoung [116]) that if f_a is mixing with respect to μ_a then the decay of correlations for Lipshitz observations on I is exponential. It is also known that the set of a such that f_a is mixing with respect to μ_a has positive Lebesgue measure. There is a well-developed theory concerning bifurcations the maps T_μ undergo as μ varies [27]. We briefly describe the period-doubling route to chaos in the family $T_\lambda(x) = \lambda x(1 - x)$. For a nice account see [46]. We let c_λ denote the fixed point $\frac{\lambda-1}{\lambda}$. For $3 < \lambda \le 1 + \sqrt{6}$, all points in $[0, 1]$ except for $0, c_\lambda$ and their preimages are attracted to a unique periodic orbit $O(p_\lambda)$ of period 2. There is a monotone sequence of parameter values λ_n ($\lambda_1 = 3$) such that for $\lambda_n < \lambda \le \lambda_{n+1}$, T_λ has a unique attracting periodic orbit $O(\lambda_n)$ of period 2^n and for each $k = 1, 2, \ldots, n - 1$ a unique repelling orbit of period 2^k. All points in the interval $[0, 1]$ except for the repelling periodic orbits and their preimages are attracted to the attracting periodic orbit of period 2^n. At $\lambda = \lambda_n$ the periodic orbit $O(\lambda_n)$ undergoes a period-doubling bifurcation. Feigenbaum [36] found that the limit $\delta = \frac{\lambda_n - \lambda_{n-1}}{\lambda_{n+1} - \lambda_n} \sim 4.699\ldots$ exists and that in a wide class of unimodal maps this period-doubling cascade occurs and the differences between successive bifurcation parameters give the same limiting ratio, an example of *universality*. At the end of the period-doubling cascade at a parameter $\lambda_\infty \sim 3.569\ldots$, T_{λ_∞} has an invariant Cantor set C (the Feigenbaum attractor) which is topologically conjugate to the dyadic adding machine coexisting with isolated repelling orbits of period 2^n, $n = 0, 1, 2, \ldots$. There is a unique repelling orbit of period 2^n for $n \ge 1$ along with two fixed points. The Cantor set C is the ω-limit set for all points that are not periodic or preimages of periodic orbits. C is the set of accumulation points of periodic orbits. Despite this picture of incredible complexity the topological entropy is zero for $\lambda \le \lambda_\infty$. For $\lambda > \lambda_\infty$ the map T_λ

has positive topological entropy and infinitely many periodic orbits whose periods are not powers of 2. For each $\lambda \geq \lambda_\infty$, T_λ possesses an invariant Cantor set which is repelling for $\lambda > \lambda_\infty$. We say that T_λ is *hyperbolic* if there is only one attracting periodic orbit and the only recurrent sets are the attracting periodic orbit, repelling periodic orbits and possibly a repelling invariant Cantor set. It is known that the set of $\lambda \in [0, 4]$ for which T_λ is hyperbolic is open and dense [43]. Remarkably, by Jakobson's result [52] there is also a positive Lebesgue measure set of parameters λ for which T_λ has an absolutely continuous invariant measure μ_λ with a positive Lyapunov exponent.

Intermittent Maps Maps of the unit interval $T\colon [0, 1] \to [0, 1]$ which are expanding except at the point $x = 0$, where they are locally of form $T(x) \sim x + x^{1+\alpha}$, $\alpha > 0$, have been extensively studied both for the insights they give into rates of decay of correlations for non-uniformly hyperbolic systems (hyperbolicity is lost at the point $x = 0$, where the derivative is 1) and for their use as models of intermittent behavior in turbulence [71]. A fixed point where the derivative is 1 is sometimes called an *indifferent fixed point*. It is a model of *intermittency* in the sense that orbits close to 1 will stay close for many iterates (since the expansion is very weak there) and hence a time series of observations will be quite uniform for long periods of time before displaying chaotic type behavior after moving away from the indifferent fixed into that part of the domain where the map is uniformly expanding.

A particularly simple model [67] is provided by

$$T(x) = \begin{cases} x(1 + 2^\alpha x^\alpha) & \text{if } x \in [0, 1/2)\,; \\ 2x - 1 & \text{if } x \in [1/2, 1]\,. \end{cases}$$

For $\alpha = 0$ the map is uniformly expanding and Lebesgue measure is invariant. In this case the rate of decay of correlations for Hölder observations is exponential. For $0 < \alpha < 1$ the map has an SRB measure μ_α with support the unit interval. For $\alpha \geq 1$ there are no absolutely continuous invariant probability measures though there are σ-finite absolutely continuous measures. Upper and lower polynomial bounds on the rate of decay of observations on such maps have been given as a function of $0 < \alpha < 1$ and the regularity of the observable. For details see [51,67,103].

Hénon Diffeomorphisms The Henón family of diffeomorphisms was introduced and studied as Poincaré maps for the Lorenz system of equations. It is a two-parameter two-dimensional family which shares many characteristics with the logistic family and for small $b > 0$ may be considered a two-dimensional 'perturbation' of the logistic family. The parametrized mapping is defined as

$$T_{a,b}(x, y) = (1 - ax^2 + y, bx)\,,$$

so $T_{a,b}\colon \mathbb{R}^2 \to \mathbb{R}^2$ with $0 < a < 2$ and $b > 0$. Benedicks and Carleson [7] showed that for a positive-measure set of parameters (a, b), $T_{a,b}$ has a topologically transitive attractor $\Lambda_{a,b}$. Benedicks and Young [8] later proved that for a positive-measure set of parameters (a, b), $T_{a,b}$ has a topologically transitive SRB attractor $\Lambda_{a,b}$ with SRB measure $\mu_{a,b}$ and that $(T_{a,b}, \Lambda_{a,b}, \mu_{a,b})$ is isomorphic to a Bernoulli shift.

Complex Dynamics

Complex dynamics is concerned with the behavior of rational maps

$$\frac{\alpha_1 z^d + \alpha_2 z^{d-1} + \cdots \alpha_{d+1}}{\beta_1 z^d + \beta_2 z^{d-1} + \cdots \beta_{d+1}}$$

of the extended complex plane $\bar{\mathbb{C}}$ to itself, in which the domain is $\mathbb{C}$ completed with the point at infinity (called the Riemann sphere). Recall that a family F of meromorphic functions is called *normal* on a domain D if every sequence possesses a subsequence that converges uniformly (in the spherical metric $\bar{\mathbb{C}} \sim S^2$) on compact subsets of D. A family is *normal at a point* $z \in \bar{\mathbb{C}}$ if it is normal on a neighborhood of z. The *Fatou set* $F(R) \subset \bar{\mathbb{C}}$ of a rational map $R\colon \bar{\mathbb{C}} \to \bar{\mathbb{C}}$ is the set of points $z \in \bar{\mathbb{C}}$ such that the family of forward iterates $\{R^n\}_{n\geq 0}$ is normal at z. The *Julia set* $J(R)$ is the complement of the Fatou set $F(R)$. The Fatou set is open and hence the Julia set is a closed set. Another characterization in the case $d > 1$ is that $J(R)$ is the closure of the set of all repelling periodic orbits of $R\colon \bar{\mathbb{C}} \to \bar{\mathbb{C}}$. Both $F(R)$ and $J(R)$ are invariant under R. The dynamics of greatest interest is the restriction $R\colon J(R) \to J(R)$. The Julia set often has a complicated fractal structure. In the case that $R_a(z) = z^2 - a$, $a \in \mathbb{C}$, the *Mandelbrot set* is defined as the set of a for which the orbit of the origin 0 is bounded. The topology of the Mandelbrot set has been the subject of intense research. The study of complex dynamics is important because of the fascinating and complicated dynamics displayed and also because techniques and results in complex dynamics have direct implications for the behavior of one-dimensional maps. For more details see [25].

Infinite Ergodic Theory

We may also consider a measure-preserving transformation (T, X, μ) of a measure space such that $\mu(X) = \infty$. For example X could be the real line equipped with Lebesgue measure. This setting also arises with compact X

in applications. For example, suppose $T\colon [0,1] \to [0,1]$ is the simple model of intermittency given in Sect. "Intermittent Maps" and $\gamma \in (1,2)$. Then T possesses an absolutely continuous invariant measure μ with support $[0,1]$, but $\mu([0,1]) = \infty$. The Radon–Nikodym derivative of μ with respect to Lebesgue measure m exists but is not in $L^1(m)$.

In this setting we say a measurable set A is a *wandering set* for T if $\{T^{-n}A\}_{n=0}^{\infty}$ are disjoint. Let $D(T)$ be the measurable union of the collection of wandering sets for T. The transformation T is *conservative* with respect to μ if $(X \setminus D(T)) = X \pmod{\mu}$ (see the article on ► Measure Preserving Systems). It is usually necessary to assume T conservative with respect to μ to say anything interesting about its behavior. For example if $T(x) = x + \alpha$, $\alpha > 0$, is a translation of the real line then $D(T) = X$. The definition of ergodicity in this setting remains the same: T is *ergodic* if $A \in \mathcal{B}$ and $T^{-1}A = A \bmod \mu$ implies that $\mu(A) = 0$ or $\mu(A^c) = 0$. However the equivalence of ergodicity of T with respect to μ and the equality of time and space averages for $L^1(\mu)$ functions no longer holds. Thus in general μ ergodic does not imply that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \phi \circ T^j(x) = \int_X \phi \, \mathrm{d}\mu \ \mu \text{ a.e. } x \in X$$

for all $\phi \in L^1(\mu)$. In the example of the intermittent map with $\gamma \in (1,2)$ the orbit of Lebesgue almost every $x \in X$ is dense in X, yet the fraction of time spent near the indifferent fixed point $x = 0$ tends to one for Lebesgue almost every $x \in X$. In fact it may be shown Sect. 2.4 in [1] that when $\mu(x) = \infty$ there are no constants $a_n > 0$ such that

$$\lim_{n\to\infty} \frac{1}{a_n} \sum_{i=0}^{n-1} \phi \circ T^j(x) = \int_X \phi \, \mathrm{d}\mu \ \mu \text{ a.e. } x \in X$$

Nevertheless it is sometimes possible to obtain distributional limits, rather than almost sure limits, of Birkhoff sums under suitable normalization. We refer the reader to Aaronson's book [1] for more details.

Constructions

We give examples of some of the standard constructions in dynamical systems. Often these constructions appear in modeling situations (for example skew products are often used to model systems which react to inputs from other systems, continuous time systems are often modeled as suspension flows over discrete-time dynamics) or to reduce systems to simpler components (often a factor system or induced system is simpler to study). Unless stated otherwise, in the sequel we will be discussing measure-preserving transformations on Lebesgue spaces (see the article on ► Measure Preserving Systems).

Products

Given measure-preserving systems $(X, \mathcal{B}, \mu, T)$ and $(Y, \mathcal{C}, \nu, S)$, their *product* consists of their completed product measure space with the transformation $T \times S\colon X \times Y \to X \times Y$ defined by $(T \times S)(x, y) = (Tx, Sy)$ for all $(x, y) \in X \times Y$. Neither ergodicity nor transitivity is in general preserved by taking products; for example the product of an irrational rotation on the unit circle with itself is not ergodic. For a list of which mixing properties are preserved in various settings by forming a product see [113]. Given any countable family of measure-preserving transformations on probability spaces, their direct product is defined similarly.

Factors

We say that a measure-preserving system $(Y, \mathcal{C}, \nu, S)$ is a *factor* of a measure-preserving system $(X, \mathcal{B}, \mu, T)$ if (possibly after deleting a set of measure 0 from X) there is a measurable onto map $\phi\colon X \to Y$ such that

$$\begin{aligned} &\phi^{-1}\mathcal{C} \subset \mathcal{B}, \\ &\phi T = S\phi, \quad \text{and} \\ &\mu T^{-1} = \nu. \end{aligned} \tag{8}$$

For Lebesgue spaces, there is a correspondence of factors of $(X, \mathcal{B}, \mu, T)$ and T-invariant complete sub-σ-algebras of $\mathcal{B}$. According to Rokhlin's theory of Lebesgue spaces [98] factors also correspond to certain partitions of X (see the article on ► Measure Preserving Systems). A factor map $\phi\colon X \to Y$ between Lebesgue spaces is an *isomorphism* if and only if it has a measurable inverse, or equivalently $\phi^{-1}\mathcal{C} = \mathcal{B}$ up to sets of measure 0.

Skew Products

If $(X, \mathcal{B}, \mu, T)$ is a measure-preserving system, $(Y, \mathcal{C}, \nu)$ is a measure-space, and $\{S_x : x \in X\}$ is a family of measure-preserving maps $Y \to Y$ such that the map that takes (x, y) to $S_x y$ is jointly measurable in the two variables x and y, then we may define a *skew product system* consisting of the product measure space of X and Y equipped with product measure $\mu \times \nu$ together with the measure-preserving map $T \ltimes S\colon X \times Y \to X \times Y$ defined by

$$(T \ltimes S)(x, y) = (Tx, S_x y)\,. \tag{9}$$

The space Y is called the *fiber* of the skew product and the space X the *base*. Sometimes in the literature the word

skew product has a more general meaning and refers to the structure $(T \ltimes S)(x, y) = (Tx, S_x y)$ (without any assumption of measure-preservation), where the action of the map on the fiber Y is determined or 'driven' by the map $T: X \to X$.

Some common examples of skew products include:

Random Dynamical Systems Suppose that X indexes a collection of mappings $S_x: Y \to Y$. We may have a transformation $T: X \to X$ which is a full shift. Then the sequence of mappings $\{S_{T^n x}\}$ may be considered a (random) choice of a mapping $Y \to Y$ from the set $\{S_x : x \in X\}$. The projection onto Y of the orbits of $(Tx, S_x y)$ give the orbits of a point $y \in Y$ under a random composition of maps $S_{T^n x} \circ \cdots \circ S_{Tx} \circ S_x$. More generally we could consider the choice of maps S_x that are composed to come from any ergodic dynamical system, (T, X, μ) to model the effect of perturbations by a stationary ergodic 'noise' process.

Group Extensions of Dynamical Systems Suppose Y is a group, ν is a measure on Y invariant under a left group action, and $S_x y := g(x)y$ is given by a group-valued function $g: X \to Y$. In this setting g is often called a *cocycle*, since upon defining $g^{(n)}(x)$ by $(T \ltimes S)^{(n)}(x, y) = (T^n x, g^{(n)}(x)y)$ we have a cocycle relation, namely $g^{(m+n)}(x) = g^{(m)}(T^n x)g^{(n)}(x)$. Group extensions arise often in modeling systems with symmetry [37]. Common examples are provided by a random composition of matrices from a group of matrices (or more generally from a set of matrices which may form a group or not).

Induced Transformations

Since by the Poincaré Recurrence Theorem (see [113]) a measure-preserving transformation $(T, X, \mu, \mathcal{B})$ on a probability space is recurrent, given any set B of positive measure, the return-time function

$$n_B(x) = \inf\{n \geq 1 : T^n x \in B\} \tag{10}$$

is finite μ a.e. We may define the *first-return map* by

$$T_B x = T^{n_B(x)} x \,. \tag{11}$$

Then (after perhaps discarding as usual a set of measure 0) $T_B: B \to B$ is a measurable transformation which preserves the probability measure $\mu_B = \mu/\mu(B)$. The system $(B, \mathcal{B} \cap B, \mu_B, T_B)$ is called an *induced*, *first-return* or *derived transformation*. If $(T, X, \mu, \mathcal{B})$ is ergodic then $(B, \mathcal{B} \cap B, \mu_B, T_B)$ is ergodic, but the converse is not in general true.

The construction of the transformation T_B allows us to represent the forward orbit of points in B via a *tower* or *skyscraper* over B. For each $n = 1, 2, \ldots$, let

$$B_n = \{x \in B : n_B(x) = n\} \,. \tag{12}$$

Then $\{B_1, B_2, \ldots\}$ form a partition of B, which we think of as the bottom floor or base of the tower. The next floor is made up of $TB_2, TB_3, \ldots$, which form a partition of $TB \setminus B$, and so on. All these sets are disjoint. A *column* is a part of the tower of the form $B_n \cup TB_n \cup \cdots \cup T^{n-1}B_n$ for some $n = 1, 2, \ldots$. The action of T on the entire tower is pictured as mapping each x not at the top of its column straight up to the point Tx above it on the next level, and mapping each point on the top level to $T^{n_B} x \in B$. An equivalent way to describe the transformation on the tower is to write for each n and $j < n$, $T^j B_n$ as $\{(x, j) : x \in B_n\}$, and then the transformation F on the tower is

$$F(x, l) = \begin{cases} (x, l+1) & \text{if } l < n_B(x) - 1 \,; \\ (T^{n_B(x)} x, 0) & \text{if } l = n_B(x) - 1 \,. \end{cases}$$

If T preserves a measure μ, then F preserves $\mu \times \mathrm{d}l$, where l is counting measure.

Sometimes the process of inducing yields an induced map which is easier to analyze (perhaps it has stronger hyperbolicity properties) than the original system. Sometimes it is possible to 'lift' ergodic or statistical properties from an induced system to the original system, so the process of inducing plays an important role in the study of statistical properties of dynamical systems [77].

It is possible to generalize the tower construction and relax the condition that $n_B(x)$ is the first-return time function. We may take a measurable set $B \subset X$ of positive μ measure and define for almost every point $x \in B$ a *height* or *ceiling* function $R: B \to \mathbb{N}$ and take a countable partition $\{X_n\}$ of B into the sets on which R is constant. We define the *tower* as the set $\Delta := \{(x, l) : x \in B, 0 \leq l < R(x)\}$ and the *tower map* $F: \Delta \to \Delta$ by

$$F(x, l) = \begin{cases} (x, l+1) & \text{if } l < R(x) - 1 \,; \\ (T^{R(x)} x, 0) & \text{if } l = R(x) - 1 \,. \end{cases}$$

In this setting, if $\int_B R(x)\,\mathrm{d}\mu < \infty$, we may define an F-invariant probability measure on Δ as $\frac{\mu}{C(R,B)} \times \mathrm{d}l$, where $\mathrm{d}l$ is counting measure and $C(R, B)$ is the normalizing constant $C(R, B) = \mu(B) \int_B R(x)\mathrm{d}\mu$. This viewpoint is connected with the construction of systems by cutting and stacking—see Sect. "Cutting and Stacking".

Suspension Flows

The tower construction has an analogue in which the height function R takes values in $\mathbb{R}$ rather than $\mathbb{N}$. Such towers are commonly used to model dynamical systems with continuous time parameter. Let (T, X, μ) be a measure-preserving system and $R\colon X \to (0, \infty)$ a measurable "ceiling" function on X. The set

$$X^R = \{(x, t)\colon 0 \le R(x) < t\}\,, \tag{13}$$

with measure ν given locally by the product of μ on X with Lebesgue measure m on $\mathbb{R}$, is a measure space in a natural way. If μ is a finite measure and R is integrable with respect to μ then ν is a finite measure. We define an action of $\mathbb{R}$ on X^R by letting each point x flow at unit speed up the vertical lines $\{(x, t)\colon 0 \le t < R(x)\}$ under the graph of R until it hits the ceiling, then jump to Tx, and so on. More precisely, defining $R_n(x) = R(x) + \cdots + R(T^n x)$,

$$T_s(x, t) = \begin{cases} (x, s + t) \\ \quad \text{if } 0 \le s + t < R(x), \\ (Tx, s + t - R(x)) \\ \quad \text{if } R(x) \le s + t < R(x) + R(Tx) \\ \cdots \\ (T^n x, s + t - [R(x) + \cdots + R(T^{n-1}x)]) \\ \quad \text{if } R_{n-1}(x) \le s + t < R_n(x)\,. \end{cases} \tag{14}$$

Ergodicity of (T, X, μ) implies the ergodicity of (T_s, X^R, ν).

Cutting and Stacking

Several of the most interesting examples in ergodic theory have been constructed by this method; in fact, because of Rokhlin's Lemma (see Sect. "Rokhlin's Lemma") every ergodic measure-preserving transformation on a Lebesgue space is isomorphic to one constructed by cutting and stacking. For example, the von Neumann-Kakutani adding machine (or 2-odometer) (Sect. "Adding Machines"), the Chacon weakly mixing but not strongly mixing system (Sect. "Chacon System"), Ornstein's mixing rank one examples (see e. g. p. 160 ff. in [80]), and many more.

We construct a Lebesgue measure-preserving transformation T on an interval X (bounded or maybe unbounded) by defining it as a translation on each of a pairwise disjoint countable collection of subintervals. The construction proceeds by stages, at each stage defining T on an additional part of X, until eventually T is defined a.e.

At each stage X is represented as a *tower*, which is defined to be a disjoint union of *columns*. A *column* is defined to be a finite disjoint union of intervals of equal length, which are numbered from 0, for the "floor", to the last one, for the "roof", and which we picture as lying each above the preceding-numbered interval. T is defined on each *level* of a column (i. e. each interval in the column) except the roof by mapping it by translation to the next higher interval in the column.

At stage 0, we have just one column, consisting of all of X as the floor, and T is not defined anywhere. To pass from one stage to the next, the columns are *cut* and *stacked*. This means that each column is divided, by vertical cuts, into a disjoint union of subcolumns of equal height (but maybe not equal width), and then some of these subcolumns are stacked above others (of the same width) so as to form a new tower. This allows the definition of T to be extended to some parts of X that were previously tops of towers, since they now may have levels above them. (Sometimes columns of height 1 are thought of as forming a reservoir for "spacers" to be inserted between subcolumns that are being stacked.) If the measure of the union of the tops of the columns tends to 0, eventually T becomes defined a.e. This description in words can be made precise with cumbersome notation, but the process can also be given a neater graphical description, which we sketch in the next section.

Adic Transformations

A.M. Vershik has introduced a family of models, called *adic* or *Bratteli–Vershik* transformations, into ergodic theory and dynamical systems. One begins with a graph which is arranged in levels, finitely many vertices on each level, with connections only from each level to the adjacent ones. The space X consists of the set of all infinite paths in this graph; it is a compact metric space in a natural way. We are given an order on the set of edges into each vertex, and then X is partially ordered as follows: x and y are comparable if they agree from some point on, in which case we say that $x < y$ if at the last level n where they traverse different edges, the edge x_n of x is smaller than the edge y_n of y. A map T is defined by letting Tx be the smallest y that is larger than x, if there is one. In nice situations, T is a homeomorphism after defining it and its inverse on perhaps countably many maximal and minimal elements. Invariant measures can sometimes be defined by assigning weights to edges, which are then multiplied to define the measure of each cylinder set. This is a nice combinatorial way to present the cutting and stacking method of constructing m.p.t.'s, allows for more convenient analysis of questions such as orbit equivalence, and leads to the construction of many interesting examples, such as those

based on the Pascal or Euler graphs [4,38,76]. Odometers and generalizations are natural examples of adic systems. Vershik showed that in fact every ergodic measure-preserving transformation on a Lebesgue space is isomorphic to a uniquely ergodic adic transformation. See [111].

Rokhlin's Lemma

The following result is the fundamental starting point for many constructions in ergodic theory, from representing arbitrary systems in terms of cutting and stacking or adic systems, to constructing useful partitions and symbolic codings of abstract systems, to connecting convergence theorems in abstract ergodic theory with those in harmonic analysis. It allows us to picture arbitrarily long stretches of the action of a measure-preserving transformation as a translation within the set of integers. In the ergodic nonatomic case the statement follows readily from the construction of induced transformations.

Lemma 1 (Rokhlin's Lemma) *Let $T: X \to X$ be a measure-preserving transformation on a probability space $(X, \mathcal{B}, \mu)$. Suppose that $(X, \mathcal{B}, \mu)$ is nonatomic and $T: X \to X$ is ergodic, or, more generally, $(T, X, \mathcal{B}, \mu)$ is aperiodic: that is to say, the set $\{x \in X: \text{ there is } n \in \mathbb{N} \text{ such that } T^n x = x\}$ of periodic points has measure 0. Then given $n \in \mathbb{N}$ and $\epsilon > 0$, there is a measurable set $B \subset X$ such that the sets $B, TB, \dots, T^{n-1}B$ are pairwise disjoint and $\mu(\cup_{k=0}^{n-1} T^k B) > 1 - \epsilon$.*

Inverse Limits

Suppose that for each $i = 1, 2, \dots$ we have a Lebesgue probability space $(X_i, \mathcal{B}_i, \mu_i)$ and a measure-preserving transformation $T_i: X_i \to X_i$. Suppose also that for each $i \le j$ there is a factor map $\phi_{ji}: (T_j, X_j, \mathcal{B}_j, \mu_j) \to (T_i, X_i, \mathcal{B}_i, \mu_i,)$, such that each ϕ_{jj} is the identity on X_j and $\phi_{ji}\phi_{kj} = \phi_{ki}$ whenever $k \ge j \ge i$. Let

$$X = \{x \in \Pi_{i=1}^{\infty} X_i : \phi_{ji} x_j = x_i \text{ for all } j \ge i\} . \tag{15}$$

For each j, let $\pi_j: X \to X_j$ be the projection defined by $\pi_j x = x_j$.

Let $\mathcal{B}$ be the smallest σ-algebra of subsets of X which contains all the $\pi_j^{-1}\mathcal{B}_j$. Define μ on each $\pi_j^{-1}\mathcal{B}_j$ by

$$\mu(\pi_j^{-1} B) = \mu_j(B) \quad \text{for all } B \in \mathcal{B}_j . \tag{16}$$

Because $\phi_{ji}\pi_j = \pi_i$ for all $j \ge i$, the $\pi_j^{-1}\mathcal{B}_j$ are increasing, and so their union is an algebra. The set function μ can, with some difficulty, be shown to be countably additive on this algebra: since we are dealing with Lebesgue spaces, by means of measure-theoretic isomorphisms it is possible to replace the entire situation by compact metric spaces and continuous maps, then use regularity of the measures involved—see p. 137 ff. in [88]. Thus by Carathéodory's Theorem (see the article on ► Measure Preserving Systems) μ extends to all of $\mathcal{B}$.

Define $T: X \to X$ by $T(x_j) = (T_j x_j)$. Then $(T, X, \mathcal{B}, \mu)$ is a measure-preserving system such that any system which has all the $(T_j, X_j, \mathcal{B}_j, \mu_j)$ as factors, also has $(T, X, \mathcal{B}, \mu)$ a factor.

Natural Extension

The natural extension is a way to produce an invertible system from a non-invertible system. The original system is a factor of its natural extension and its orbit structure and ergodic properties are captured by the natural extension, as will be seen from its construction. Let $(T, X, \mathcal{B}, \mu)$ be a measure-preserving transformation of a Lebesgue probability space. Define

$$\Omega := \{(x_0, x_1, x_2, \dots): x_n = T(x_{n+1}), x_n \in X, n = 0, 1, 2, \dots\}$$

with $\sigma: \Omega \to \Omega$ defined by $\sigma((x_0, x_1, x_2, \dots)) = (T(x_0), x_0, x_1, \dots)$. The map σ is invertible on Ω. Given the invariant measure μ we define the invariant measure for the natural extension $\tilde{\mu}$ on Ω by defining it first on cylinder sets $C(A_0, A_1, \dots, A_k)$ by

$$\tilde{\mu}(C(A_0, A_1, \dots, A_k)) = \mu(T^{-k}(A_0) \cap T^{-k+1}(A_1) \cdots \cap T^{-k+i}(A_i) \cap \cdots \cap A_k)$$

and then extending it to Ω using Kolmogorov's extension theorem. We think of $(x_0, x_1, x_2, \dots)$ as being an inverse branch of $x_0 \in X$ under the mapping $T: X \to X$. The maps $\sigma, \sigma^{-1}: \Omega \to \Omega$ are ergodic with respect to $\tilde{\mu}$ if $(T, X, \mathcal{B}, \mu)$ is ergodic [113]. If $\pi: \Omega \to X$ is projection onto the first component i. e. $\pi(x_0, \dots, x_n, \dots) = x_0$ then $\pi \circ \sigma^n(x_0, \dots, x_n, \dots) = T^n(x_0)$ for all x_0 and thus the natural extension yields information about the orbits of X under T.

The natural extension is an inverse limit. Let $(X, \mathcal{B}, \mu)$ be a Lebesgue probability space and $T: X \to X$ a map such that $T^{-1}\mathcal{B} \subset \mathcal{B}$ and $\mu T^{-1} = \mu$. For each $i = 1, 2, \dots$ let $(T_i, X_i, \mathcal{B}_i, \mu_i) = (T, X, \mathcal{B}, \mu)$, and $\phi_{ji} = T^{j-i}$ for each $j > i$. Then the inverse limit $(\hat{T}, \hat{X}, \hat{\mathcal{B}}, \hat{\mu})$ of this system is an *invertible* measure-preserving system which is the natural extension of $(T, X, \mathcal{B}, \mu)$. We have

$$\hat{T}^{-1}(x_1, x_2, \dots) = (x_2, x_3, \dots) . \tag{17}$$

The original system $(T, X, \mathcal{B}, \mu)$ is a factor of $(\hat{T}, \hat{X}, \hat{\mathcal{B}}, \hat{\mu})$ (using any π_i as the factor map), and any factor mapping from an invertible system onto $(T, X, \mathcal{B}, \mu)$ consists

of a factor mapping onto $(\hat{T}, \hat{X}, \hat{\mathcal{B}}, \hat{\mu})$ followed by projection onto the first coordinate.

Joinings

Given measure-preserving systems $(T, X, \mathcal{B}, \mu)$ and (S, Y, C, ν), a *joining* of the two systems is a $T \times S$-invariant measure P on their product measurable space that projects to μ and ν, respectively, under the projections of $X \times Y$ to X and Y, respectively. That is, if $\pi_1 \colon X \times Y \to X$ is the projection onto the first component i. e. $\pi_1(x, y) = x$ then $P(\pi_1^{-1}(A)) = \mu(A)$ for all $A \in \mathcal{B}$ and similarly for $\pi_2 \colon X \times Y \to Y$.

This concept is the ergodic-theoretic version of the notion in probability theory of a *coupling*. The product measure $\mu \times \nu$ is always a joining of the two systems. If product measure is the only joining of the two systems, then we say that they are disjoint and write $X \perp Y$ [40]. If $\mathcal{D}$ is any family of systems, we write $\mathcal{D}^{\perp}$ for the family of all measure-preserving systems which are disjoint from every system in $\mathcal{D}$. Extensive recent accounts of the use of joinings in ergodic theory are in [42,99,110].

Future Directions

The basic examples and constructions presented here are idealized, and many of the underlying assumptions (such as uniform hyperbolicity) are seldom satisfied in applications, yet they have given important insights into the behavior of real-world physical systems. Recent developments have improved our understanding of the ergodic properties of non-uniformly and partially hyperbolic systems. The ergodic properties of deterministic systems will continue to be an active research area for the foreseeable future. The directions will include, among others: establishing statistical and ergodic properties under weakened dependence assumptions; the study of systems which display 'anomalous statistics'; the study of the stability and typicality of ergodic behavior and mixing in dynamical systems; the ergodic theory of infinite-dimensional systems; advances in number theory (see the sections on Szemerédi and Ramsey theory); research into models with non-singular rather than invariant measures; and infinite-measure systems. Other chapters in this Encyclopedia discuss in more detail these and other topics.

Bibliography

Primary Literature

1. Aaronson J (1997) An introduction to infinite ergodic theory, Mathematical Surveys and Monographs, vol 50. American Mathematical Society, Providence. MR 1450400 (99d:28025)
2. Anosov DV (1967) Geodesic flows on closed riemannian manifolds of negative curvature. Trudy Mat Inst Steklov 90:209. MR 0224110 (36 #7157)
3. Arnol′d VI (1963) Small denominators and problems of stability of motion in classical and celestial mechanics. Uspehi Mat Nauk 18(6 (114)):91–192. MR 0170705 (30 #943)
4. Bailey S, Keane M, Petersen K, Salama IA (2006) Ergodicity of the adic transformation on the euler graph. Math Proc Cambridge Philos Soc 141(2):231–238. MR 2265871 (2007m:37010)
5. Baladi V (2000) Positive transfer operators and decay of correlations. Advanced Series in Nonlinear Dynamics, vol 16. World Scientific Publishing Co Inc, River Edge. MR 1793194 (2001k:37035)
6. Benedicks M, Carleson L (1985) On iterations of $1 - ax(X, \mathcal{B}, \mu)2$ on $(-1, 1)$. Ann Math (2) 122(1):1–25. MR 799250 (87c:58058)
7. Benedicks M, Carleson L (1991) The dynamics of the hénon map. Ann Math (2) 133(1):73–169. MR 1087346 (92d:58116)
8. Benedicks M, Young LS (1993) Sinaï-bowen-ruelle measures for certain hénon maps. Invent Math 112(3):541–576. MR 1218323 (94e:58074)
9. Bertrand-Mathis A (1986) Développement en base θ; répartition modulo un de la suite $(x\theta(X, \mathcal{B}, \mu)n)_{n\geq 0}$; langages codés et θ-shift. Bull Soc Math France 114(3):271–323. MR 878240 (88e:11067)
10. Billingsley P (1978) Ergodic Theory and Information. Robert E. Krieger Publishing Co, Huntington, N.Y., reprint of the 1965 original. MR 524567 (80b:28017)
11. Bissinger BH (1944) A generalization of continued fractions. Bull Amer Math Soc 50:868–876. MR 0011338 (6,150h)
12. Blanchard F (1989) β-expansions and symbolic dynamics. Theoret Comput Sci 65(2):131–141. MR 1020481 (90j:54039)
13. Blanchard F, Hansel G (1986) Systèmes codés. Theoret Comput Sci 44(1):17–49. MR 858689 (88m:68029)
14. Blanchard F, Hansel G (1986) Systèmes codés et limites de systèmes sofiques. C R Acad Sci Paris Sér I Math 303(10):475–477. MR 865864 (87m:94009)
15. Blanchard F, Hansel G (1991) Sofic constant-to-one extensions of subshifts of finite type. Proc Amer Math Soc 112(1):259–265. MR 1050016 (91m:54050)
16. Boshernitzan M, Galperin G, Krüger T, Troubetzkoy S (1998) Periodic billiard orbits are dense in rational polygons. Trans Amer Math Soc 350(9):3523–3535. MR 1458298 (98k:58179)
17. Boshernitzan MD (1992) Billiards and rational periodic directions in polygons. Amer Math Monthly 99(6):522–529. MR 1166001 (93d:51043)
18. Bowen R (1970) Markov partitions for axiom A diffeomorphisms. Amer J Math 92:725–747. MR 0277003 (43 #2740)
19. Bowen R (1975) Equilibrium states and the ergodic theory of Anosov diffeomorphisms. In: Lecture notes in mathematics, vol 470. Springer, Berlin. MR 0442989 (56 #1364)
20. Boyarsky A, Góra P (1997) Laws of chaos: Invariant measures and dynamical systems in one dimension. Probability and its Applications. Birkhäuser, Boston. MR 1461536 (99a:58102)
21. Brin MI, Pesin JB (1974) Partially hyperbolic dynamical systems. Izv Akad Nauk SSSR Ser Mat 38:170–212. MR 0343316 (49 #8058)
22. Bunimovich LA (1974) The ergodic properties of certain billiards. Funkcional Anal i Priložen 8(3):73–74. MR 0357736 (50 #10204)

23. Bunimovich LA (1979) On the ergodic properties of nowhere dispersing billiards. Comm Math Phys 65(3):295–312. MR 530154 (80h:58037)
24. Burns K, Pugh C, Shub M, Wilkinson A (2001) Recent results about stable ergodicity. In: Smooth ergodic theory and its applications, Seattle, WA, 1999. Proc Sympos Pure Math, vol 69. Amer Math Soc, Providence, RI, pp 327–366. MR 1858538 (2002m:37042)
25. Carleson L, Gamelin TW (1993) Complex dynamics. Universitext: Tracts in Mathematics. Springer, New York. MR 1230383 (94h:30033)
26. Chernov N, Markarian R (2006) Chaotic billiards, Mathematical Surveys and Monographs, vol 127. American Mathematical Society, Providence, RI. MR 2229799 (2007f:37050)
27. Collet P, Eckmann JP (1980) Iterated maps on the interval as dynamical systems, Progress in Physics, vol 1. Birkhäuser, Boston. MR 613981 (82j:58078)
28. Cornfeld IP, Fomin SV, Sinaĭ YG (1982) Ergodic Theory, Grundlehren der mathematischen Wissenschaften (Fundamental Principles of Mathematical Sciences), vol 245. Springer, New York, translated from the Russian by A. B. Sosinskiĭ. MR 832433 (87f:28019)
29. Coven EM, Hedlund GA (1973) Sequences with minimal block growth. Math Systems Theory 7:138–153. MR 0322838 (48 #1199)
30. del Junco A (1978) A simple measure-preserving transformation with trivial centralizer. Pacific J Math 79(2):357–362. MR 531323 (80i:28034)
31. del Junco A, Rahe M, Swanson L (1980) Chacon's automorphism has minimal self-joinings. J Analyse Math 37:276–284. MR 583640 (81j:28027)
32. de Melo W, van Strien S (1993) One-dimensional dynamics. Ergebnisse der Mathematik und ihrer Grenzgebiete (3) (Results in Mathematics and Related Areas (3)), vol 25. Springer, Berlin. MR 1239171 (95a:58035)
33. de la Rue T (1993) Entropie d'un système dynamique gaussien: cas d'une action de $\mathbf{z}(X, \mathcal{B}, \mu)d$. C R Acad Sci Paris Sér I Math 317(2):191–194. MR 1231420 (94c:28022)
34. Downarowicz T (2005) Survey of odometers and Toeplitz flows. In: Algebraic and topological dynamics. Contemp Math, vol 385. Amer Math Soc, Providence, RI, pp 7–37. MR 2180227 (2006f:37009)
35. Everett CJ (1946) Representations for real numbers. Bull Amer Math Soc 52:861–869. MR 0018221 (8,259c)
36. Feigenbaum MJ (1978) Quantitative universality for a class of nonlinear transformations. J Statist Phys 19(1):25–52. MR 0501179 (58 #18601)
37. Field M, Nicol M (2004) Ergodic theory of equivariant diffeomorphisms: Markov partitions and stable ergodicity. Mem Amer Math Soc 169(803):viii+100. MR 2045641 (2005g:37041)
38. Frick SB, Petersen K () Random permutations and unique fully supported ergodicity for the Euler adic transformation. Ann Inst H Poincaré Prob Stat. To appear
39. Furstenberg H (1973) The unique ergodicity of the horocycle flow. In: Recent advances in topological dynamics (Proc Conf, Yale Univ, New Haven, Conn, 1972; in honor of Gustav Arnold Hedlund). Lecture Notes in Math, vol 318. Springer, Berlin, pp 95–115. MR 0393339 (52 #14149)
40. Furstenberg H (1967) Disjointness in ergodic theory, minimal sets, and a problem in diophantine approximation. Math Systems Theory 1:1–49. MR 0213508 (35 #4369)
41. Gallavotti G, Ornstein DS (1974) Billiards and bernoulli schemes. Comm Math Phys 38:83–101. MR 0355003 (50 #7480)
42. Glasner E (2003) Ergodic Theory via Joinings. Mathematical Surveys and Monographs, vol 101. American Mathematical Society, Providence, RI. MR 1958753 (2004c:37011)
43. Graczyk J, Światek G (1997) Generic hyperbolicity in the logistic family. Ann Math (2) 146(1):1–52. MR 1469316 (99b:58079)
44. Grayson M, Pugh C, Shub M (1994) Stably ergodic diffeomorphisms. Ann Math (2) 140(2):295–329. MR 1298715 (95g:58128)
45. Guckenheimer J, Holmes P (1990) Nonlinear oscillations, dynamical systems, and bifurcations of vector fields. Applied Mathematical Sciences, vol 42. Springer, New York, revised and corrected reprint of the 1983 original. MR 1139515 (93e:58046)
46. Hasselblatt B, Katok A (2003) A First Course in Dynamics. Cambridge University Press, with a panorama of recent developments. MR 1995704 (2004f:37001)
47. Hayashi S (1997) Connecting invariant manifolds and the solution of the $c(X, \mathcal{B}, \mu)1$ stability and ω-stability conjectures for flows. Ann Math (2) 145(1):81–137. MR 1432037 (98b:58096)
48. Hedlund GA (1934) On the metrical transitivity of the geodesics on closed surfaces of constant negative curvature. Ann Math (2) 35(4):787–808. MR 1503197
49. Hopf E (1939) Statistik der geodätischen Linien in Mannigfaltigkeiten negativer Krümmung. Ber Verh Sächs Akad Wiss Leipzig 91:261–304. MR 0001464 (1,243a)
50. Host B (1995) Nombres normaux, entropie, translations. Israel J Math 91(1-3):419–428. MR 1348326 (96g:11092)
51. Hu H (2004) Decay of correlations for piecewise smooth maps with indifferent fixed points. Ergodic Theory Dynam Systems 24(2):495–524. MR 2054191 (2005a:37064)
52. Jakobson MV (1981) Absolutely continuous invariant measures for one-parameter families of one-dimensional maps. Comm Math Phys 81(1):39–88. MR 630331 (83j:58070)
53. Katok A (1980) Lyapunov exponents, entropy and periodic orbits for diffeomorphisms. Inst Hautes Études Sci Publ Math 51:137–173. MR 573822 (81i:28022)
54. Katok A, Strelcyn JM, Ledrappier F, Przytycki F (1986) Invariant manifolds, entropy and billiards; smooth maps with singularities. Lecture Notes in Mathematics, vol 1222. Springer, Berlin. MR 872698 (88k:58075)
55. Keane M (1968) Generalized morse sequences. Z Wahrscheinlichkeitstheorie Verw Gebiete 10:335–353. MR 0239047 (39 #406)
56. Keane M (1977) Non-ergodic interval exchange transformations. Israel J Math 26(2):188–196. MR 0435353 (55 #8313)
57. Keller G, Nowicki T (1992) Spectral theory, zeta functions and the distribution of periodic points for Collet-Eckmann maps. Comm Math Phys 149(1):31–69. MR 1182410 (93i:58123)
58. Kerckhoff S, Masur H, Smillie J (1986) Ergodicity of billiard flows and quadratic differentials. Ann Math (2) 124(2):293–311. MR 855297 (88f:58122)
59. Kolmogorov AN (1954) On conservation of conditionally periodic motions for a small change in Hamilton's function. Dokl Akad Nauk SSSR (NS) 98:527–530. MR 0068687 (16,924c)
60. Krieger W (2000) On subshifts and topological Markov chains. In: Numbers, information and complexity (Bielefeld, 1998). Kluwer, Boston, pp 453–472. MR 1755380 (2001g:37010)

61. Lagarias JC (1991) The Farey shift. Manuscript
62. Lagarias JC (1992) Number theory and dynamical systems. In: The unreasonable effectiveness of number theory (Orono, ME, 1991). Proc Sympos Appl Math, vol 46. Amer Math Soc, Providence, RI, pp 35–72. MR 1195841 (93m:11143)
63. Lazutkin VF (1973) Existence of caustics for the billiard problem in a convex domain. Izv Akad Nauk SSSR Ser Mat 37:186–216. MR 0328219 (48 #6561)
64. Ledrappier F (1984) Propriétés ergodiques des mesures de sinaï. Inst Hautes Études Sci Publ Math 59:163–188. MR 743818 (86f:58092)
65. Lind D, Marcus B (1995) An Introduction to Symbolic Dynamics and Coding. Cambridge University Press, Cambridge. MR 1369092 (97a:58050)
66. Liverani C, Wojtkowski MP (1995) Ergodicity in hamiltonian systems. In: Dynamics reported. Dynam Report Expositions Dynam Systems (N.S.), vol 4. Springer, Berlin, pp 130–202. MR 1346498 (96g:58144)
67. Liverani C, Saussol B, Vaienti S (1999) A probabilistic approach to intermittency. Ergodic Theory Dynam Systems 19(3):671–685. MR 1695915 (2000d:37029)
68. Lyons R (1988) On measures simultaneously 2- and 3-invariant. Israel J Math 61(2):219–224. MR 941238 (89e:28031)
69. Mañé R (1987) Ergodic theory and differentiable dynamics, Ergebnisse der Mathematik und ihrer Grenzgebiete (3) (Results in Mathematics and Related Areas (3)), vol 8. Springer, Berlin, translated from the Portuguese by Silvio Levy. MR 889254 (88c:58040)
70. Mañé R (1988) A proof of the $c(X, \mathcal{B}, \mu)1$ stability conjecture. Inst Hautes Études Sci Publ Math 66:161–210. MR 932138 (89e:58090)
71. Manneville P, Pomeau Y (1980) Different ways to turbulence in dissipative dynamical systems. Phys D 1(2):219–226. MR 581352 (81h:58041)
72. Marcus B (1975) Unique ergodicity of the horocycle flow: variable negative curvature case. Israel J Math 21(2-3):133–144, Conference on Ergodic Theory and Topological Dynamics (Kibbutz Lavi, 1974). MR 0407902 (53 #11672)
73. Marcus B (1978) The horocycle flow is mixing of all degrees. Invent Math 46(3):201–209. MR 0488168 (58 #7731)
74. Masur H (1986) Closed trajectories for quadratic differentials with an application to billiards. Duke Math J 53(2):307–314. MR 850537 (87j:30107)
75. Mayer DH (1991) Continued fractions and related transformations. In: Ergodic theory, symbolic dynamics, and hyperbolic spaces (Trieste, 1989). Oxford Sci Publ. Oxford Univ Press, Oxford, pp 175–222. MR 1130177
76. Méla X, Petersen K (2005) Dynamical properties of the pascal adic transformation. Ergodic Theory Dynam Systems 25(1):227–256. MR 2122921 (2005k:37012)
77. Melbourne I, Török A (2004) Statistical limit theorems for suspension flows. Israel J Math 144:191–209. MR 2121540 (2006c:37005)
78. de Melo W (1973) Structural stability of diffeomorphisms on two-manifolds. Invent Math 21:233–246. MR 0339277 (49 #4037)
79. Moser J (1962) On invariant curves of area-preserving mappings of an annulus. Nachr Akad Wiss Göttingen Math-Phys Kl II 1962:1–20. MR 0147741 (26 #5255)
80. Nadkarni MG (1998) Spectral Theory of Dynamical Systems. Birkhäuser Advanced Texts: Basler Lehrbücher. (Birkhäuser Advanced Texts: Basel Textbooks), Birkhäuser, Basel. MR 1719722 (2001d:37001)
81. Ornstein D (1970) Bernoulli shifts with the same entropy are isomorphic. Advances in Math 4:337–352. MR 0257322 (41 #1973)
82. Ornstein DS, Weiss B (1973) Geodesic flows are bernoullian. Israel J Math 14:184–198. MR 0325926 (48 #4272)
83. Oseledec VI (1968) A multiplicative ergodic theorem. Characteristic Ljapunov exponents of dynamical systems. Trudy Moskov Mat Obšč 19:179–210. MR 0240280 (39 #1629)
84. Parry W (1960) On the β-expansions of real numbers. Acta Math Acad Sci Hungar 11:401–416. MR 0142719 (26 #288)
85. Parry W (1964) Representations for real numbers. Acta Math Acad Sci Hungar 15:95–105. MR 0166332 (29 #3609)
86. Parry W (1966) Symbolic dynamics and transformations of the unit interval. Trans Amer Math Soc 122:368–378. MR 0197683 (33 #5846)
87. Parry W (1996) Squaring and cubing the circle—rudolph's theorem. In: Ergodic theory of $\mathbf{Z}(X, \mathcal{B}, \mu)d$ actions (Warwick, 1993-1994). London Math Soc Lecture Note Ser, vol 228. Cambridge Univ Press, Cambridge, pp 177–183. MR 1411219 (97h:28009)
88. Parthasarathy KR (2005) Probability Measures on Metric Spaces. AMS Chelsea Publishing, Providence, RI, reprint of the 1967 original. MR 2169627 (2006d:60004)
89. Pesin JB (1976) Families of invariant manifolds that correspond to nonzero characteristic exponents. Izv Akad Nauk SSSR Ser Mat 40(6):1332–1379, 1440. MR 0458490 (56 #16690)
90. Pesin JB (1977) Characteristic Ljapunov exponents, and smooth ergodic theory. Uspehi Mat Nauk 32(4 (196)):55–112, 287. MR 0466791 (57 #6667)
91. Phillips E, Varadhan S (eds) (1975) Ergodic Theory. Courant Institute of Mathematical Sciences New York University, a seminar held at the Courant Institute of Mathematical Sciences, New York University, New York, 1973–1974; With contributions by S. Varadhan, E. Phillips, S. Alpern, N. Bitzenhofer and R. Adler. MR 0486431 (58 #6177)
92. Pugh C, Shub M (2004) Stable ergodicity. Bull Amer Math Soc (NS) 41(1):1–41 (electronic), with an appendix by Alexander Starkov. MR 2015448 (2005f:37011)
93. Rényi A (1957) Representations for real numbers and their ergodic properties. Acta Math Acad Sci Hungar 8:477–493. MR 0097374 (20 #3843)
94. Robbin JW (1971) A structural stability theorem. Ann Math (2) 94:447–493. MR 0287580 (44 #4783)
95. Robinson C (1975) Errata to: "structural stability of vector fields" (ann. of math. (2) 99:154–175 (1974)). Ann Math (2) 101:368. MR 0365630 (51 #1882)
96. Robinson C (1976) Structural stability of $c(X, \mathcal{B}, \mu)1$ diffeomorphisms. J Differential Equations 22(1):28–73. MR 0474411 (57 #14051)
97. Robinson RC (1973) $c(X, \mathcal{B}, \mu)r$ structural stability implies Kupka-Smale. In: Dynamical systems, (Proc Sympos, Univ Bahia, Salvador, 1971). Academic Press, New York, pp 443–449. MR 0334282 (48 #12601)
98. Rohlin VA (1952) On the fundamental ideas of measure theory. Amer Math Soc Translation 1952(71):55. MR 0047744 (13,924e)
99. Rudolph DJ (1990) Fundamentals of Measurable Dynamics: Ergodic theory on Lebesgue spaces. Oxford Science Publica-

tions. The Clarendon Press Oxford University Press, New York. MR 1086631 (92e:28006)
100. Rudolph DJ (1990) ×2 and ×3 invariant measures and entropy. Ergodic Theory Dynam Systems 10(2):395–406. MR 1062766 (91g:28026)
101. Ruelle D (1976) A measure associated with axiom-a attractors. Amer J Math 98(3):619–654. MR 0415683 (54 #3763)
102. Ruelle D (1978) Thermodynamic formalism: the mathematical structures of classical equilibrium statistical mechanics. Encyclopedia of Mathematics and its Applications, vol 5. Addison-Wesley Publishing Co, Reading, MA, with a foreword by Giovanni Gallavotti and Gian-Carlo Rota. MR 511655 (80g:82017)
103. Sarig O (2002) Subexponential decay of correlations. Invent Math 150(3):629–653. MR 1946554 (2004e:37010)
104. Schweiger F (1995) Ergodic Theory of Fibred Systems and Metric Number Theory. Oxford Science Publications. The Clarendon Press Oxford University Press, New York. MR 1419320 (97h:11083)
105. Sinaĭ JG (1968) Markov partitions and u-diffeomorphisms. Funkcional Anal i Priložen 2(1):64–89. MR 0233038 (38 #1361)
106. Sinaĭ JG (1970) Dynamical systems with elastic reflections. Ergodic properties of dispersing billiards. Uspehi Mat Nauk 25(2 (152)):141–192. MR 0274721 (43 #481)
107. Sinaĭ JG (1972) Gibbs measures in ergodic theory. Uspehi Mat Nauk 27(4(166)):21–64. MR 0399421 (53 #3265)
108. Smale S (1980) The mathematics of time. Essays on dynamical systems, economic processes, and related topics. Springer, New York. MR 607330 (83a:01068)
109. Tabachnikov S (2005) Geometry and billiards, Student Mathematical Library, vol 30. American Mathematical Society, Providence, RI. MR 2168892 (2006h:51001)
110. Thouvenot JP (1995) Some properties and applications of joinings in ergodic theory. In: Ergodic theory and its connections with harmonic analysis (Alexandria, 1993). London Math Soc Lecture Note Ser, vol 205. Cambridge Univ Press, Cambridge, pp 207–235. MR 1325699 (96d:28017)
111. Vershik AM, Livshits AN (1992) Adic models of ergodic transformations, spectral theory, substitutions, and related topics. In: Representation theory and dynamical systems. Adv Soviet Math, vol 9. Amer Math Soc, Providence, RI, pp 185–204. MR 1166202 (93i:46131)
112. Vorobets YB, Gal′perin GA, Stëpin AM (1992) Periodic billiard trajectories in polygons: generation mechanisms. Uspekhi Mat Nauk 47(3(285)):9–74, 207, (Russian with Russian summary), English translation: (1992) Russian Math Surveys 47(3):5–80. MR 1185299 (93h:58088)
113. Walters P (1982) An Introduction to Ergodic Theory. Graduate Texts in Mathematics, vol 79. Springer, New York. MR 648108 (84e:28017)
114. Young LS (1993) Ergodic theory of chaotic dynamical systems. In: From Topology to Computation: Proceedings of the Smalefest (Berkeley, CA, 1990). Springer, New York, pp 201–226. MR 1246120 (94i:58112)
115. Young LS (1998) Statistical properties of dynamical systems with some hyperbolicity. Ann Math (2) 147(3):585–650. MR 1637655 (99h:58140)
116. Young LS (1999) Recurrence times and rates of mixing. Israel J Math 110:153–188. MR 1750438 (2001j:37062)
117. Zemljakov AN, Katok AB (1975) Topological transitivity of billiards in polygons. Mat Zametki 18(2):291–300. MR 0399423 (53 #3267)

Books and Reviews

Baladi V (2000) Positive transfer operators and decay of correlations. Advanced Series in Nonlinear Dynamics, vol 16. World Scientific Publishing Co Inc, River Edge. MR 1793194 (2001k:37035)

Billingsley P (1978) Ergodic Theory and Information. Robert E. Krieger Publishing Co, Huntington, N.Y., pp xiii+194, reprint of the 1965 original. MR 524567 (80b:28017)

Billingsley P (1995) Probability and Measure. Wiley Series in Probability and Mathematical Statistics, 3rd edn. Wiley, New York, pp xiv+593, A Wiley-Interscience Publication. MR 1324786 (95k:60001)

Bonatti C, Díaz LJ, Viana M (2005) Dynamics beyond uniform hyperbolicity: A global geometric and probabilistic perspective; Mathematical Physics, III. In: Encyclopaedia of Mathematical Sciences,vol. 102. Springer, Berlin, pp xviii+384. MR 2105774 (2005g:37001)

Boyarsky A, Góra P (1997) Laws of chaos: Invariant measures and dynamical systems in one dimension. Probability and its Applications. Birkhäuser, Boston. MR 1461536 (99a:58102)

Brin M, Stuck G (2002) Introduction to dynamical systems. Cambridge University Press, Cambridge. MR 1963683 (2003m:37001)

Carleson L, Gamelin TW (1993) Complex dynamics. Universitext: Tracts in Mathematics. Springer, New York. MR 1230383 (94h:30033)

Chernov N, Markarian R (2006) Chaotic billiards, Mathematical Surveys and Monographs, vol 127. American Mathematical Society, Providence, RI. MR 2229799 (2007f:37050)

Collet P, Eckmann JP (1980) Iterated maps on the interval as dynamical systems, Progress in Physics, vol 1. Birkhäuser, Boston. MR 613981 (82j:58078)

Cornfeld IP, Fomin SV, Sinaĭ YG (1982) Ergodic Theory, Grundlehren der mathematischen Wissenschaften (Fundamental Principles of Mathematical Sciences), vol 245. Springer, New York, translated from the Russian by A. B. Sosinskiĭ. MR 832433 (87f:28019)

Denker M, Grillenberger C, Sigmund K (1976) Ergodic Theory on Compact Spaces. Lecture Notes in Mathematics, vol 527. Springer, Berlin. MR 0457675 (56 #15879)

Friedman NA (1970) Introduction to Ergodic Theory. Van Nostrand Reinhold Mathematical Studies, No 29. Van Nostrand Reinhold Co, New York. MR 0435350 (55 #8310)

Glasner E (2003) Ergodic Theory via Joinings, Mathematical Surveys and Monographs, vol 101. American Mathematical Society, Providence, RI. MR 1958753 (2004c:37011)

Halmos PR (1960) Lectures on Ergodic Theory. Chelsea Publishing Co, New York. MR 0111817 (22 #2677)

Hasselblatt B, Katok A (2003) A First Course in Dynamics: With a panorama of recent developments. Cambridge University Press, Cambridge. MR 1995704 (2004f:37001)

Hopf E (1937) Ergodentheorie, 1st edn. Ergebnisse der Mathematik und ihrer Grenzgebiete; 5. Bd, 2, J. Springer, Berlin

Jacobs K (1965) Einige neuere Ergebnisse der Ergodentheorie. Jber Deutsch Math-Verein 67(Abt 1):143–182. MR 0186789 (32 #4244)

Keller G (1998) Equilibrium States in Ergodic Theory. In: London Mathematical Society Student Texts, vol 42. Cambridge University Press, Cambridge, pp x+178. MR 1618769 (99e:28022)

Mañé R (1987) Ergodic theory and differentiable dynamics. Ergebnisse der Mathematik und ihrer Grenzgebiete (3) (Results in Mathematics and Related Areas (3)), vol 8. Springer, Berlin, translated from the Portuguese by Silvio Levy. MR 889254 (88c:58040)

Nadkarni MG (1998) Spectral Theory of Dynamical Systems. Birkhäuser Advanced Texts: Basler Lehrbücher. (Birkhäuser Advanced Texts: Basel Textbooks), Birkhäuser, Basel. MR 1719722 (2001d:37001)

Katok A, Hasselblatt B (1995) Introduction to the Modern Theory of Dynamical Systems, Encyclopedia of Mathematics and its Applications, vol 54. Cambridge University Press, Cambridge, With a supplementary chapter by Katok and Leonardo Mendoza. MR 1326374 (96c:58055)

Parry W, Pollicott M (1990) Zeta functions and the periodic orbit structure of hyperbolic dynamics. Astérisque (187-188):268. MR 1085356 (92f:58141)

Ornstein DS, Rudolph DJ, Weiss B (1982) Equivalence of measure preserving transformations. Mem Amer Math Soc 37(262):xii+116. MR 653094 (85e:28026)

Petersen K (1989) Ergodic Theory, Cambridge Studies in Advanced Mathematics, vol 2. Cambridge University Press, Cambridge. corrected reprint of the 1983 original. MR 1073173 (92c:28010)

Royden HL (1988) Real Analysis, 3rd edn. Macmillan Publishing Company, New York. MR 1013117 (90g:00004)

Rudolph DJ (1990) Fundamentals of Measurable Dynamics: Ergodic theory on Lebesgue spaces. Oxford Science Publications. The Clarendon Press Oxford University Press, New York. MR 1086631 (92e:28006)

Schweiger F (1995) Ergodic Theory of Fibred Systems and Metric Number Theory. Oxford Science Publications. The Clarendon Press Oxford University Press, New York. MR 1419320 (97h:11083)

Thouvenot JP (1995) Some properties and applications of joinings in ergodic theory. In: Ergodic theory and its connections with harmonic analysis (Alexandria, 1993). London Math Soc Lecture Note Ser, vol 205. Cambridge Univ Press, Cambridge, pp 207–235. MR 1325699 (96d:28017)

Walters P (1982) An Introduction to Ergodic Theory, Graduate Texts in Mathematics, vol 79. Springer, New York. MR 648108 (84e:28017)

Ergodic Theory of Cellular Automata

Marcus Pivato
Department of Mathematics, Trent University, Peterborough, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
Invariant Measures for CA
Limit Measures and Other Asymptotics
Measurable Dynamics
Entropy
Future Directions and Open Problems
Acknowledgments
Bibliography

Glossary

Configuration space and the shift Let $\mathbb{M}$ be a finitely generated group or monoid (usually abelian). Typically, $\mathbb{M} = \mathbb{N} := \{0, 1, 2, \ldots\}$ or $\mathbb{M} = \mathbb{Z} := \{\ldots, -1, 0, 1, 2, \ldots\}$, or $\mathbb{M} = \mathbb{N}^E$, $\mathbb{Z}^D$, or $\mathbb{Z}^D \times \mathbb{N}^E$ for some $D, E \in \mathbb{N}$. In some applications, $\mathbb{M}$ could be nonabelian (although usually amenable), but to avoid notational complexity we will generally assume $\mathbb{M}$ is abelian and additive, with operation '+'.

Let $\mathcal{A}$ be a finite set of symbols (called an *alphabet*). Let $\mathcal{A}^{\mathbb{M}}$ denote the set of all functions $\mathbf{a}\colon \mathbb{M} \longrightarrow \mathcal{A}$, which we regard as $\mathbb{M}$-indexed *configurations* of elements in $\mathcal{A}$. We write such a configuration as $\mathbf{a} = [a_{\mathsf{m}}]_{\mathsf{m}\in\mathbb{M}}$, where $a_{\mathsf{m}} \in \mathcal{A}$ for all $\mathsf{m} \in \mathbb{M}$, and refer to $\mathcal{A}^{\mathbb{M}}$ as *configuration space*.

Treat $\mathcal{A}$ as a discrete topological space; then $\mathcal{A}$ is compact (because it is finite), so $\mathcal{A}^{\mathbb{M}}$ is compact in the Tychonoff product topology. In fact, $\mathcal{A}^{\mathbb{M}}$ is a *Cantor space*: it is compact, perfect, totally disconnected, and metrizable. For example, if $\mathbb{M} = \mathbb{Z}^D$, then the standard metric on $\mathcal{A}^{\mathbb{Z}^D}$ is defined $d(\mathbf{a}, \mathbf{b}) = 2^{-\Delta(\mathbf{a},\mathbf{b})}$, where $\Delta(\mathbf{a}, \mathbf{b}) := \min\{|\mathsf{z}|; a_{\mathsf{z}} \neq b_{\mathsf{z}}\}$.

Any $\mathsf{v} \in \mathbb{M}$, determines a continuous *shift map* $\sigma^{\mathsf{v}}\colon \mathcal{A}^{\mathbb{M}} \longrightarrow \mathcal{A}^{\mathbb{M}}$ defined by $\sigma^{\mathsf{v}}(\mathbf{a})_{\mathsf{m}} = a_{\mathsf{m}+\mathsf{v}}$ for all $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$ and $\mathsf{m} \in \mathbb{M}$. The set $\{\sigma^{\mathsf{v}}\}_{\mathsf{v}\in\mathbb{M}}$ is then a continuous $\mathbb{M}$-action on $\mathcal{A}^{\mathbb{M}}$, which we denote simply by "σ".

If $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$ and $\mathbb{U} \subset \mathbb{M}$, then we define $\mathbf{a}_{\mathbb{U}} \in \mathcal{A}^{\mathbb{U}}$ by $\mathbf{a}_{\mathbb{U}} := [a_{\mathsf{u}}]_{\mathsf{u}\in\mathbb{U}}$. If $\mathsf{m} \in \mathbb{M}$, then strictly speaking, $\mathbf{a}_{\mathsf{m}+\mathbb{U}} \in \mathcal{A}^{\mathsf{m}+\mathbb{U}}$; however, it will often be convenient to 'abuse notation' and treat $\mathbf{a}_{\mathsf{m}+\mathbb{U}}$ as an element of $\mathcal{A}^{\mathbb{U}}$ in the obvious way.

Cellular automata Let $\mathbb{H} \subset \mathbb{M}$ be some finite subset, and let $\phi\colon \mathcal{A}^{\mathbb{H}} \longrightarrow \mathcal{A}$ be a function (called a *local rule*). The *cellular automaton* (CA) determined by ϕ is the function $\Phi\colon \mathcal{A}^{\mathbb{M}} \longrightarrow \mathcal{A}^{\mathbb{M}}$ defined by $\Phi(\mathbf{a})_{\mathsf{m}} = \phi(\mathbf{a}_{\mathsf{m}+\mathbb{H}})$ for all $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$ and $\mathsf{m} \in \mathbb{M}$. Curtis, Hedlund and Lyndon showed that cellular automata are exactly the continuous transformations of $\mathcal{A}^{\mathbb{M}}$ which commute with all shifts (see Theorem 3.4 in [58]). We refer to $\mathbb{H}$ as the *neighborhood* of Φ. For example, if $\mathbb{M} = \mathbb{Z}$, then typically $\mathbb{H} := [-\ell \ldots r] := \{-\ell, 1-\ell, \ldots, r-1, r\}$ for some *left radius* $\ell \geq 0$ and *right radius* $r \geq 0$. If $-\ell \geq 0$, then ϕ can either define CA on $\mathcal{A}^{\mathbb{N}}$ or define a *one-sided* CA on $\mathcal{A}^{\mathbb{Z}}$. If $\mathbb{M} = \mathbb{Z}^D$, then typically

$\mathbb{H} \subseteq [-R \ldots R]^D$, for some *radius* $R \geq 0$. Normally we assume that ℓ, r, and R are chosen to be minimal. Several specific classes of CA will be important to us:

Linear CA Let $(\mathcal{A}, +)$ be a finite abelian group (e. g. $\mathcal{A} = \mathbb{Z}_{/p}$, where $p \in \mathbb{N}$; usually p is prime). Then Φ is a *linear* CA (LCA) if the local rule ϕ has the form

$$\phi(\mathbf{a}_{\mathbb{H}}) := \sum_{\mathsf{h} \in \mathbb{H}} \varphi_{\mathsf{h}}(a_{\mathsf{h}}), \quad \forall \mathbf{a}_{\mathbb{H}} \in \mathcal{A}^{\mathbb{H}}, \tag{1}$$

where $\varphi_{\mathsf{h}} \colon \mathcal{A} \longrightarrow \mathcal{A}$ is an endomorphism of $(\mathcal{A}, +)$, for each $\mathsf{h} \in \mathbb{H}$. We say that Φ has *scalar coefficients* if, for each $\mathsf{h} \in \mathbb{H}$, there is some scalar $c_{\mathsf{h}} \in \mathbb{Z}$, so that $\varphi_{\mathsf{h}}(a_{\mathsf{h}}) := c_{\mathsf{h}} \cdot a_{\mathsf{h}}$; then $\phi(\mathbf{a}_{\mathbb{H}}) := \sum_{\mathsf{h} \in \mathbb{H}} c_{\mathsf{h}} a_{\mathsf{h}}$. For example, if $\mathcal{A} = (\mathbb{Z}_{/p}, +)$, then *all* endomorphisms are scalar multiplications, so *all* LCA have scalar coefficients.

If $c_{\mathsf{h}} = 1$ for all $\mathsf{h} \in \mathbb{H}$, then Φ has local rule $\phi(\mathbf{a}_{\mathbb{H}}) := \sum_{\mathsf{h} \in \mathbb{H}} a_{\mathsf{h}}$; in this case, Φ is called an *additive cellular automaton*; see ► Additive Cellular Automata.

Affine CA If $(\mathcal{A}, +)$ is a finite abelian group, then an *affine CA* is one with a local rule $\phi(\mathbf{a}_{\mathbb{H}}) := c + \sum_{\mathsf{h} \in \mathbb{H}} \varphi_{\mathsf{h}}(a_{\mathsf{h}})$, where c is some constant and where $\varphi_{\mathsf{h}} \colon \mathcal{A} \longrightarrow \mathcal{A}$ are endomorphisms of $(\mathcal{A}, +)$. Thus, Φ is an LCA if $c = 0$.

Permutative CA Suppose $\Phi \colon \mathcal{A}^{\mathbb{Z}} \longrightarrow \mathcal{A}^{\mathbb{Z}}$ has local rule $\phi \colon \mathcal{A}^{[-\ell \ldots r]} \longrightarrow \mathcal{A}$. Fix $\mathbf{b} = [b_{1-\ell}, \ldots, b_{r-1}, b_r] \in \mathcal{A}^{(-\ell \ldots r]}$. For any $a \in \mathcal{A}$, define $[a\ \mathbf{b}] := [a, b_{1-\ell}, \ldots, b_{r-1}, b_r] \in \mathcal{A}^{[-\ell \ldots r]}$. We then define the function $\phi_{\mathbf{b}} \colon \mathcal{A} \longrightarrow \mathcal{A}$ by $\phi_{\mathbf{b}}(a) := \phi([a\ \mathbf{b}])$. We say that Φ is *left-permutative* if $\phi_{\mathbf{b}} \colon \mathcal{A} \longrightarrow \mathcal{A}$ is a permutation (i. e. a bijection) for all $\mathbf{b} \in \mathcal{A}^{(-\ell \ldots r]}$. Likewise, given $\mathbf{b} = [b_{-\ell}, \ldots, b_{r-1}] \in \mathcal{A}^{[-\ell \ldots r)}$ and $c \in \mathcal{A}$, define $[\mathbf{b}\ c] := [b_{-\ell}, b_{1-\ell} \ldots, b_{r-1}, c] \in \mathcal{A}^{[-\ell \ldots r]}$, and define ${}_{\mathbf{b}}\phi \colon \mathcal{A} \longrightarrow \mathcal{A}$ by ${}_{\mathbf{b}}\phi(c) := \phi([\mathbf{b}c])$; then Φ is *right-permutative* if ${}_{\mathbf{b}}\phi \colon \mathcal{A} \longrightarrow \mathcal{A}$ is a permutation for all $\mathbf{b} \in \mathcal{A}^{[-\ell \ldots r)}$. We say Φ is *bipermutative* if it is both left- and right-permutative. More generally, if $\mathbb{M}$ is any monoid, $\mathbb{H} \subset \mathbb{M}$ is any neighborhood, and $\mathsf{h} \in \mathbb{H}$ is any fixed coordinate, then we define h-permutativity for a CA on $\mathcal{A}^{\mathbb{M}}$ in the obvious fashion.

For example, suppose $(\mathcal{A}, +)$ is an abelian group and Φ is an affine CA on $\mathcal{A}^{\mathbb{Z}}$ with local rule $\phi(\mathbf{a}_{\mathbb{H}}) = c + \sum_{h=-\ell}^{r} \varphi_h(a_h)$. Then Φ is left-permutative iff $\varphi_{-\ell}$ is an automorphism, and right-permutative iff φ_r is an automorphism. If $\mathcal{A} = \mathbb{Z}_{/p}$, and p is prime, then *every* nontrivial endomorphism is an automorphism(because it is multiplication by a nonzero element of $\mathbb{Z}_{/p}$, which is a field), so in this case, *every* affine CA is permutative in every coordinate of its neighborhood (and in particular, bipermutative). If $\mathcal{A} \neq \mathbb{Z}_{/p}$, however, then not all affine CA are permutative.

Permutative CA were introduced by Hedlund [58], §6, and are sometimes called *permutive* CA. Right permutative CA on $\mathcal{A}^{\mathbb{N}}$ are also called *toggle automata*. For more information, see Sect. 7 of ► Topological Dynamics of Cellular Automata.

Subshifts A *subshift* is a closed, σ-invariant subset $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$. For any $\mathbb{U} \subset \mathbb{M}$, let $\mathbf{X}_{\mathbb{U}} := \{\mathbf{x}_{\mathbb{U}}; \mathbf{x} \in \mathbf{X}\} \subset \mathcal{A}^{\mathbb{U}}$. We say $\mathbf{X}$ is a *subshift of finite type* (SFT) if there is some finite $\mathbb{U} \subset \mathbb{M}$ such that $\mathbf{X}$ is entirely described by $\mathbf{X}_{\mathbb{U}}$, in the sense that $\mathbf{X} = \{\mathbf{x} \in \mathcal{A}^{\mathbb{M}}; \mathbf{x}_{\mathbb{U}+\mathsf{m}} \in \mathbf{X}_{\mathbb{U}}, \forall \mathsf{m} \in \mathbb{M}\}$.

In particular, if $\mathbb{M} = \mathbb{Z}$, then a (two-sided) *Markov subshift* is an SFT $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}}$ determined by a set $\mathbf{X}_{\{0,1\}} \subset \mathcal{A}^{\{0,1\}}$ of *admissible transitions*; equivalently, $\mathbf{X}$ is the set of all bi-infinite directed paths in a digraph whose vertices are the elements of $\mathcal{A}$, with an edge $a \rightsquigarrow b$ iff $(a, b) \in \mathbf{X}_{\{0,1\}}$. If $\mathbb{M} = \mathbb{N}$, then a *one-sided Markov subshift* is a subshift of $\mathcal{A}^{\mathbb{N}}$ defined in the same way.

If $D \geq 2$, then an SFT in $\mathcal{A}^{\mathbb{Z}^D}$ can be thought of as the set of admissible 'tilings' of $\mathbb{R}^D$ by Wang tiles corresponding to the elements of $\mathbf{X}_{\mathbb{U}}$. (Wang tiles are unit squares (or (hyper)cubes) with various 'notches' cut into their edges (or (hyper)faces) so that they can only be juxtaposed in certain ways.)

A subshift $\mathbf{X} \subseteq \mathcal{A}^{\mathbb{Z}^D}$ is *strongly irreducible* (or *topologically mixing*) if there is some $R \in \mathbb{N}$ such that, for any disjoint finite subsets $\mathbb{V}, \mathbb{U} \subset \mathbb{Z}^D$ separated by a distance of at least R, and for any $\mathbf{u} \in \mathbf{X}_{\mathbb{U}}$ and $\mathbf{v} \in \mathbf{X}_{\mathbb{V}}$, there is some $\mathbf{x} \in \mathbf{X}$ such that $\mathbf{x}_{\mathbb{U}} = \mathbf{u}$ and $\mathbf{x}_{\mathbb{V}} = \mathbf{v}$. Please see ► Symbolic Dynamics for more about subshifts.

Measures For any finite subset $\mathbb{U} \subset \mathbb{M}$, and any $\mathbf{b} \in \mathcal{A}^{\mathbb{U}}$, let $\langle \mathbf{b} \rangle := \{\mathbf{a} \in \mathcal{A}^{\mathbb{M}}; \mathbf{a}_{\mathbb{U}} := \mathbf{b}\}$ be the *cylinder set* determined by $\mathbf{b}$. Let $\mathfrak{B}$ be the sigma-algebra on $\mathcal{A}^{\mathbb{M}}$ generated by all cylinder sets. A (probability) measure μ on $\mathcal{A}^{\mathbb{M}}$ is a countably additive function $\mu \colon \mathfrak{B} \longrightarrow [0, 1]$ such that $\mu[\mathcal{A}^{\mathbb{M}}] = 1$. A measure on $\mathcal{A}^{\mathbb{M}}$ is entirely determined by its values on cylinder sets. We will be mainly concerned with the following classes of measures:

Bernoulli measure Let β_0 be a probability measure on $\mathcal{A}$. The *Bernoulli measure* induced by β_0 is the measure β on $\mathcal{A}^{\mathbb{M}}$ such that, for any finite subset $\mathbb{U} \subset \mathbb{M}$, and any $\mathbf{a} \in \mathcal{A}^{\mathbb{U}}$, if $U := |\mathbb{U}|$, then $\beta[\langle \mathbf{a} \rangle] = \prod_{\mathsf{h} \in \mathbb{H}} \beta_0(a_{\mathsf{h}})$.

Invariant measure Let μ be a measure on $\mathcal{A}^{\mathbb{M}}$, and let $\Phi : \mathcal{A}^{\mathbb{M}} \longrightarrow \mathcal{A}^{\mathbb{M}}$ be a cellular automaton. The measure $\Phi\mu$ is defined by $\Phi\mu(\mathbf{B}) = \mu(\Phi^{-1}(\mathbf{B}))$, for any $\mathbf{B} \in \mathfrak{B}$. We say that μ is *Φ-invariant* (or that Φ is *μ-preserving*) if $\Phi\mu = \mu$. For more information, see ► Ergodic Theory: Basic Examples and Constructions.

Uniform measure Let $A := |\mathcal{A}|$. The uniform measure η on $\mathcal{A}^{\mathbb{M}}$ is the Bernoulli measure such that, for any finite subset $\mathbb{U} \subset \mathbb{M}$, and any $\mathbf{b} \in \mathcal{A}^{\mathbb{U}}$, if $U := |\mathbb{U}|$, then $\mu[\langle \mathbf{b} \rangle] = 1/A^U$.

The *support* of a measure μ is the smallest closed subset $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$ such that $\mu[\mathbf{X}] = 1$; we denote this by $\mathsf{supp}(\mu)$. We say μ has *full support* if $\mathsf{supp}(\mu) = \mathcal{A}^{\mathbb{M}}$ – equivalently, $\mu[\mathbf{C}] > 0$ for *every* cylinder subset $\mathbf{C} \subset \mathcal{A}^{\mathbb{M}}$.

Notation Let $\mathsf{CA}(\mathcal{A}^{\mathbb{M}})$ denote the set of all cellular automata on $\mathcal{A}^{\mathbb{M}}$. If $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$, then let $\mathsf{CA}(\mathbf{X})$ be the subset of all $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})$ such that $\Phi(\mathbf{X}) \subseteq \mathbf{X}$. Let $\mathfrak{M}_{\mathsf{ens}}(\mathcal{A}^{\mathbb{M}})$ be the set of all probability measures on $\mathcal{A}^{\mathbb{M}}$, and let $\mathfrak{M}_{\mathsf{ens}}(\mathcal{A}^{\mathbb{M}}; \Phi)$ be the subset of Φ-invariant measures. If $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$, then let $\mathfrak{M}_{\mathsf{ens}}(\mathbf{X})$ be the set of probability measures μ with $\mathsf{supp}(\mu) \subseteq \mathbf{X}$, and define $\mathfrak{M}_{\mathsf{ens}}(\mathbf{X}; \Phi)$ in the obvious way.

Font conventions Upper case calligraphic letters $(\mathcal{A}, \mathcal{B}, \mathcal{C}, \ldots)$ denote finite alphabets or groups. Upper-case bold letters $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots)$ denote subsets of $\mathcal{A}^{\mathbb{M}}$ (e. g. subshifts), lowercase bold-faced letters $(\mathbf{a}, \mathbf{b}, \mathbf{c}, \ldots)$ denote elements of $\mathcal{A}^{\mathbb{M}}$, and Roman letters $(a, b, c, \ldots)$ are elements of $\mathcal{A}$ or ordinary numbers. Lower-case sans-serif $(\ldots, \mathsf{m}, \mathsf{n}, \mathsf{p})$ are elements of $\mathbb{M}$, upper-case hollow font $(\mathbb{U}, \mathbb{V}, \mathbb{W}, \ldots)$ are subsets of $\mathbb{M}$. Upper-case Greek letters $(\Phi, \Psi, \ldots)$ are functions on $\mathcal{A}^{\mathbb{M}}$ (e. g. CA, block maps), and lowercase Greek letters $(\phi, \psi, \ldots)$ are other functions (e. g. local rules, measures.)

Acronyms in square brackets (e. g. ► Topological Dynamics of Cellular Automata) indicate cross-references to related entries in the Encyclopedia; these are listed at the end of this article.

Definition of the Subject

Loosely speaking, a *cellular automaton* (CA) is the 'discrete' analogue of a partial differential evolution equation: it is a spatially distributed, discrete-time, symbolic dynamical system governed by a local interaction rule which is invariant in space and time. In a CA, 'space' is discrete (usually the D-dimensional lattice, $\mathbb{Z}^D$) and the local statespace at each point in space is also discrete (a finite 'alphabet', usually denoted by $\mathcal{A}$).

A *measure-preserving dynamical system* (MPDS) is a dynamical system equipped with an invariant probability measure. Any MPDS can be represented as a stationary stochastic process (SSP) and vice versa; 'chaos' in the MPDS can be quantified via the information-theoretic 'entropy' of the corresponding SSP. An MPDS Φ on a statespace $\mathbf{X}$ also defines a unitary linear operator Φ_* on the Hilbert space $\mathbf{L}^2(\mathbf{X})$; the spectral properties of Φ_* encode information about the global periodic structure and long-term informational asymptotics of Φ. *Ergodic theory* is the study of MPDSs and SSPs, and lies at the interface between dynamics, probability theory, information theory, and unitary operator theory.

Please refer to the Glossary for precise definitions of 'CA', 'MPDS', etc. Also, see ► Ergodic Theory: Basic Examples and Constructions for an introduction to ergodic theory.

Introduction

The study of CA as symbolic dynamical systems began with Hedlund [58], and the study of CA as MPDSs began with Coven and Paul [24] and Willson [144]. (Further historical details will unfold below, where appropriate.) The ergodic theory of CA is important for several reasons:

- CA are topological dynamical systems (► Topological Dynamics of Cellular Automata, ► Chaotic Behavior of Cellular Automata). We can gain insight into the topological dynamics of a CA by identifying its invariant measures, and then studying the corresponding measurable dynamics.
- CA are often proposed as stylized models of spatially distributed systems in statistical physics – for example, as microscale models of hydrodynamics, or of atomic lattices (► Cellular Automata Modeling of Physical Systems). In this context, the distinct invariant measures of a CA correspond to distinct 'phases' of the physical system (► Phase Transitions in Cellular Automata).
- CA can also act as information-processing systems (► Cellular Automata, Universality of, ► Cellular Automata as Models of Parallel Computation). Ergodic theory studies the 'informational' aspect of dynamical systems, so it is particularly suited to explicitly 'informational' dynamical systems like CA.

Article Roadmap

In Sect. "Invariant Measures for CA", we characterize the invariant measures for various classes of CA. Then, in

Sect. "Limit Measures and Other Asymptotics", we investigate which measures are 'generic' in the sense that they arise as the attractors for some large class of initial conditions. In Sect. "Measurable Dynamics" we study the mixing and spectral properties of CA as measure-preserving dynamical systems. Finally, in Sect. "Entropy", we look at entropy. These sections are logically independent, and can be read in any order.

Invariant Measures for CA

The Uniform Measure vs. Surjective Cellular Automata

The uniform measure η plays a central role in the ergodic theory of cellular automata, because of the following result.

Theorem 1 *Let $\mathbb{M} = \mathbb{Z}^D \times \mathbb{N}^E$, let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})$ and let η be the uniform measure on $\mathcal{A}^{\mathbb{M}}$. Then (Φ preserves η) $\Longleftrightarrow$ (Φ is surjective).*

Proof sketch "$\Longrightarrow$" If Φ preserves η, then Φ must map $\mathsf{supp}(\eta)$ onto itself. But $\mathsf{supp}(\eta) = \mathcal{A}^{\mathbb{M}}$; hence Φ is surjective.

"$\Longleftarrow$" The case $D = 1$ follows from a result of W.A. Blankenship and Oscar S. Rothaus, which first appeared in Theorem 5.4 in [58]. The Blankenship–Rothaus Theorem states that, if $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ is surjective and has neighborhood $[-\ell \ldots r]$, then for any $k \in \mathbb{N}$ and any $\mathbf{a} \in \mathcal{A}^k$, the Φ-preimage of the cylinder set $\langle \mathbf{a} \rangle$ is a disjoint union of exactly $A^{r+\ell}$ cylinder sets of length $k + r + \ell$; it follows that $\mu[\Phi^{-1}\langle \mathbf{a} \rangle] = A^{r+\ell}/A^{k+r+\ell} = A^{-k} = \mu\langle \mathbf{a} \rangle$. This result was later reproved by Kleveland (see Theorem 5.1 in [74]). The special case $\mathcal{A} = \{0, 1\}$ also appeared in Theorem 2.4 in [131].

The case $D \geq 2$ follows from the multidimensional version of the Blankenship–Rothaus Theorem, which was proved by Maruoka and Kimura (see Theorem 2 in [93]) (their proof assumes that $D = 2$ and that Φ has a 'quiescent' state, but neither hypothesis is essential). Alternately, "$\Longleftarrow$" follows from recent, more general results of Meester, Burton, and Steif; see Example 9 below. □

Example 2 Let $\mathbb{M} = \mathbb{Z}$ or $\mathbb{N}$ and consider CA on $\mathcal{A}^{\mathbb{M}}$.

(a) Say that Φ is *bounded-to-one* if there is some $B \in \mathbb{N}$ such that every $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$ has at most B preimages. Then (Φ is bounded-to-one) $\Longleftrightarrow$ (Φ is surjective).

(b) Any posexpansive CA on $\mathcal{A}^{\mathbb{M}}$ is surjective (see Subsect. "Posexpansive and Permutative CA" below).

(c) Any left- or right-permutative CA on $\mathcal{A}^{\mathbb{Z}}$ (or right-permutative CA on $\mathcal{A}^{\mathbb{N}}$) is surjective. This includes, for example, most linear CA.

Hence, in any of these cases, Φ preserves the uniform measure.

Proof For **(a)**, see Theorem 5.9 in [58], or Corollary 8.1.20, p. 271 in [81]. For **(b)**, see Proposition 2.2 in [9], in the case $\mathcal{A}^{\mathbb{N}}$; their argument also works for $\mathcal{A}^{\mathbb{Z}}$.

Part **(c)** follows from (b) because any permutative CA is posexpansive (Proposition 11 below). There is also a simple direct proof for a right-permutative CA on $\mathcal{A}^{\mathbb{N}}$: using right-permutativity, you can systematically construct a preimage of any desired image sequence, one entry at a time. See Theorem 6.6 in [58] for the proof in $\mathcal{A}^{\mathbb{Z}}$. □

The surjectivity of a one-dimensional CA can be determined in finite time using certain combinatorial tests (▶ Topological Dynamics of Cellular Automata). However, for $D \geq 2$, it is formally undecidable whether an arbitrary CA on $\mathcal{A}^{\mathbb{Z}^D}$ is surjective (▶ Tiling Problem and Undecidability in Cellular Automata). This problem is sometimes referred to as the *Garden of Eden problem*, because an element of $\mathcal{A}^{\mathbb{Z}^D}$ with no Φ-preimage is called a *Garden of Eden* (GOE) configuration for Φ (because it could only ever occur at the 'beginning of time'). However, it is known that a CA is surjective if it is 'almost injective' in a certain sense, which we now specify.

Let $(\mathbb{M}, +)$ be any monoid, and let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})$ have neighborhood $\mathbb{H} \subset \mathbb{M}$. If $\mathbb{B} \subset \mathbb{M}$ is any subset, then we define

$$\overline{\mathbb{B}} := \mathbb{B} + \mathbb{H} = \{\mathsf{b} + \mathsf{h}\,;\, \mathsf{b} \in \mathbb{B}, \mathsf{h} \in \mathbb{H}\}\,; \quad \text{and} \quad \partial\mathbb{B} := \overline{\mathbb{B}} \cap \overline{\mathbb{B}^{\mathsf{C}}}\,.$$

If $\mathbb{B}$ is finite, then so is $\overline{\mathbb{B}}$ (because $\mathbb{H}$ is finite). If Φ has local rule $\phi\colon \mathcal{A}^{\mathbb{H}} \longrightarrow \mathcal{A}$, then ϕ induces a function $\Phi_{\mathbb{B}}\colon \mathcal{A}^{\overline{\mathbb{B}}} \longrightarrow \mathcal{A}^{\mathbb{B}}$ in the obvious fashion. A $\mathbb{B}$-*bubble* (or $\mathbb{B}$-*diamond*) is a pair $\mathbf{b}, \mathbf{b}' \in \mathcal{A}^{\overline{\mathbb{B}}}$ such that:

$$\mathbf{b} \neq \mathbf{b}'\,; \quad \mathbf{b}_{\partial\mathbb{B}} = \mathbf{b}'_{\partial\mathbb{B}}\,; \quad \text{and} \quad \Phi_{\mathbb{B}}(\mathbf{b}) = \Phi_{\mathbb{B}}(\mathbf{b}')\,.$$

Suppose $\mathbf{a}, \mathbf{a}' \in \mathcal{A}^{\mathbb{M}}$ are two configurations such that

$$\mathbf{a}_{\overline{\mathbb{B}}} = \mathbf{b}\,, \quad \mathbf{a}'_{\overline{\mathbb{B}}} = \mathbf{b}'\,, \quad \text{and} \quad \mathbf{a}_{\mathbb{B}^{\mathsf{C}}} = \mathbf{a}'_{\mathbb{B}^{\mathsf{C}}}\,.$$

Then it is easy to verify that $\Phi(\mathbf{a}) = \Phi(\mathbf{a}')$. We say that $\mathbf{a}$ and $\mathbf{a}'$ form a *mutually erasable pair* (because Φ 'erases' the difference between $\mathbf{a}$ and $\mathbf{a}'$). Figure 1 is a schematic representation of this structure in the case $D = 1$ (hence

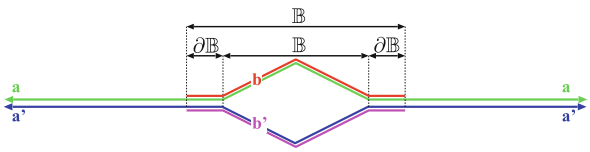

Ergodic Theory of Cellular Automata, Figure 1
A 'diamond' in $\mathcal{A}^{\mathbb{Z}}$

the term 'diamond'). If $D = 2$, then $\mathbf{a}$ and $\mathbf{a}'$ are like two membranes which are glued together everywhere except for a $\mathbb{B}$-shaped 'bubble'. We say that Φ is *pre-injective* if any (and thus, all) of the following three conditions hold:

- Φ admits no bubbles.
- Φ admits no mutually erasable pairs.
- For any $\mathbf{c} \in \mathcal{A}^{\mathbb{M}}$, if $\mathbf{a}, \mathbf{a}' \in \Phi^{-1}\{\mathbf{c}\}$ are distinct, then $\mathbf{a}$ and $\mathbf{a}'$ must differ in infinitely many locations.

For example, any injective CA is preinjective (because a mutually erasable pair for Φ gives two distinct Φ-preimages for some point). More to the point, however, if $\mathbb{B}$ is finite, and Φ admits a $\mathbb{B}$-bubble $(\mathbf{b}, \mathbf{b}')$, then we can embed N disjoint copies of $\mathbb{B}$ into $\mathbb{M}$, and thus, by making various choices between $\mathbf{b}$ and $\mathbf{b}'$ on different translates, we obtain a configuration with 2^N distinct Φ-preimages (where N is arbitrarily large). But if some configurations in $\mathcal{A}^{\mathbb{M}}$ have such a large number of preimages, then other configurations in $\mathcal{A}^{\mathbb{M}}$ must have very few preimages, or even none. This leads to the following result:

Theorem 3 (Garden of Eden) *Let $\mathbb{M}$ be a finitely generated amenable group* (e. g. $\mathbb{M} = \mathbb{Z}^D$). *Let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{M}})$.*
(a) *Φ is surjective if and only if Φ is pre-injective.*
(b) *Let $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$ be a strongly irreducible SFT such that $\Phi(\mathbf{X}) \subseteq \mathbf{X}$. Then $\Phi(\mathbf{X}) = \mathbf{X}$ if and only if $\Phi|_{\mathbf{X}}$ is pre-injective.*

Proof **(a)** The case $\mathbb{M} = \mathbb{Z}^2$ was originally proved by Moore [103] and Myhill [104]; see ► Cellular Automata and Groups. The case $\mathbb{M} = \mathbb{Z}$ was implicit Hedlund (Lemma 5.11, and Theorems 5.9 and 5.12 in [58]). The case when $\mathbb{M}$ is a finite-dimensional group was proved by Machi and Mignosi [92]. Finally, the general case was proved by Ceccherini-Silberstein, Machi, and Scarabotti (see Theorem 3 in [20]), see ► Cellular Automata and Groups.

(b) The case $\mathbb{M} = \mathbb{Z}$ is Corollary 8.1.20 in [81] (actually this holds for any sofic subshift); see also Fiorenzi [38]. The general case is Corollary 4.8 in [39]. □

Corollary 4 (Incompressibility) *Suppose $\mathbb{M}$ is a finitely generated amenable group and $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{M}})$. If Φ is injective, then Φ is surjective.*

Remark 5 (a) A *cellular network* is a CA-like system defined on an infinite, locally finite digraph, with different local rules at different nodes. By assuming a kind of 'amenability' for this digraph, and then imposing some weak global statistical symmetry conditions on the local rules, Gromov (see Theorem 8.F' in [51]) has generalized the GOE Theorem 3 to a large class of such cellular networks (which he calls 'endomorphisms of symbolic algebraic varieties'). See also [19].

(b) In the terminology suggested by Gottschalk [46], Incompressibility Corollary 4 says that the group $\mathbb{M}$ is *surjunctive*; Gottschalk claims that 'surjunctivity' was first proved for all residually finite groups by Lawton (unpublished); see ► Cellular Automata and Groups. For a recent direct proof (not using the GOE theorem), see Weiss (see Theorem 1.6 in [143]). Weiss also defines *sofic* groups (a class containing both residually finite groups and amenable groups) and shows that Corollary 4 holds whenever $\mathbb{M}$ is a sofic group (see Theorem 3.2 in [143]); see also ► Cellular Automata and Groups.

(c) If $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$ is an SFT such that $\Phi(\mathbf{X}) \subseteq \mathbf{X}$, then Corollary 4 holds as long as $\mathbf{X}$ is 'semi-strongly irreducible'; see Fiorenzi (see Corollary 4.10 in [40]).

Invariance of Maxentropy Measures

If $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^D}$ is any subshift with topological entropy $h_{\text{top}}(\mathbf{X}, \sigma)$, and $\mu \in \mathfrak{M}_{\text{ens}}(\mathbf{X}, \sigma)$ has measurable entropy $h(\mu, \sigma)$, then in general, $h(\mu, \sigma) \leq h_{\text{top}}(\mathbf{X}, \sigma)$; we say μ is a *measure of maximal entropy* (or *maxentropy measure*) if $h(\mu, \sigma) = h_{\text{top}}(\mathbf{X}, \sigma)$. (See Example 75(a) for definitions.)

Every subshift admits one or more maxentropy measures. If $D = 1$ and $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}}$ is an irreducible subshift of finite type (SFT), then Parry (see Theorem 10 in [110]) showed that $\mathbf{X}$ admits a *unique* maxentropy measure $\eta_{\mathbf{X}}$ (now called the *Parry measure*); see Theorem 8.10, p. 194 in [142] or Sect. 13.3, pp. 443–444 in [81]. Theorem 1 is then a special case of the following result:

Theorem 6 (Coven, Paul, Meester and Steif) *Let $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^D}$ be an SFT having a unique maxentropy measure $\eta_{\mathbf{X}}$, and let $\Phi \in \mathcal{CA}(\mathbf{X})$. Then Φ preserves $\eta_{\mathbf{X}}$ if and only if $\Phi(\mathbf{X}) = \mathbf{X}$.*

Proof The case $D = 1$ is Corollary 2.3 in [24]. The case $D \geq 2$ follows from Theorem 2.5(iii) in [95], which states: if $\mathbf{X}$ and $\mathbf{Y}$ are SFTs, and $\Phi : \mathbf{X} \longrightarrow \mathbf{Y}$ is a factor mapping, and μ is a maxentropy measure on $\mathbf{X}$, then $\Phi(\mu)$ is a maxentropy measure on $\mathbf{Y}$. □

For example, if $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}}$ is an irreducible SFT and $\eta_{\mathbf{X}}$ is its Parry measure, and $\Phi(\mathbf{X}) = \mathbf{X}$, then Theorem 6 says $\Phi(\eta_{\mathbf{X}}) = \eta_{\mathbf{X}}$, as observed by Coven and Paul (see Theorem 5.1 in [24]). Unfortunately, higher-dimensional SFTs do *not*, in general, have unique maxentropy measures. Burton and Steif [14] provided a plethora of examples of such nonuniqueness, but they also gave a sufficient condition for uniqueness of the maxentropy measure, which we now explain.

Let $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^D}$ be an SFT and let $\mathbb{U} \subset \mathbb{Z}^D$. For any $\mathbf{x} \in \mathbf{X}$, let $\mathbf{x}_{\mathbb{U}} := [x_{\mathsf{u}}]_{\mathsf{u}\in\mathbb{U}}$ be its 'projection' to $\mathcal{A}^{\mathbb{U}}$, and let $\mathbf{X}_{\mathbb{U}} := \{\mathbf{x}_{\mathbb{U}}; \mathbf{x} \in \mathbf{X}\} \subseteq \mathcal{A}^{\mathbb{U}}$. Let $\mathbb{V} := \mathbb{U}^{\mathsf{C}} \subset \mathbb{Z}^D$. For any $\mathbf{u} \in \mathcal{A}^{\mathbb{U}}$ and $\mathbf{v} \in \mathcal{A}^{\mathbb{V}}$, let $[\mathbf{uv}]$ denote the element of $\mathcal{A}^{\mathbb{Z}^D}$ such that $[\mathbf{uv}]_{\mathbb{U}} = \mathbf{u}$ and $[\mathbf{uv}]_{\mathbb{V}} = \mathbf{v}$. Let

$$\mathbf{X}^{(\mathbf{u})} := \{\mathbf{v} \in \mathcal{A}^{\mathbb{V}}\,;\ \ [\mathbf{uv}] \in \mathbf{X}\}$$

be the set of all "$\mathbf{X}$-admissible completions" of $\mathbf{u}$ (thus, $\mathbf{X}^{(\mathbf{u})} \neq \emptyset \Leftrightarrow \mathbf{u} \in \mathbf{X}_{\mathbb{U}}$). If $\mu \in \mathfrak{M}_{\mathsf{ers}}(\mathcal{A}^{\mathbb{Z}^D})$, and $\mathbf{u} \in \mathcal{A}^{\mathbb{U}}$, then let $\mu^{(\mathbf{u})}$ denote the *conditional measure* on $\mathcal{A}^{\mathbb{V}}$ induced by $\mathbf{u}$. If $\mathbb{U}$ is finite, then $\mu^{(\mathbf{u})}$ is just the restriction of μ to the cylinder set $\langle\mathbf{u}\rangle$. If $\mathbb{U}$ is infinite, then the precise definition of $\mu^{(\mathbf{u})}$ involves a 'disintegration' of μ into 'fibre measures' (we will suppress the details).

Let $\mu_{\mathbb{U}}$ be the projection of μ onto $\mathcal{A}^{\mathbb{U}}$. If $\mathsf{supp}(\mu) \subseteq \mathbf{X}$, then $\mathsf{supp}(\mu_{\mathbb{U}}) \subseteq \mathbf{X}_{\mathbb{U}}$, and for any $\mathbf{u} \in \mathcal{A}^{\mathbb{U}}$, $\mathsf{supp}(\mu^{(\mathbf{u})}) \subseteq \mathbf{X}^{(\mathbf{u})}$. We say that μ is a *Burton–Steif measure* on $\mathbf{X}$ if:

(1) $\mathsf{supp}(\mu) = \mathbf{X}$; and
(2) For any $\mathbb{U} \subset \mathbb{Z}^D$ whose complement $\mathbb{U}^{\mathsf{C}}$ is finite, and for $\mu_{\mathbb{U}}$-almost any $\mathbf{u} \in \mathbf{X}_{\mathbb{U}}$, the measure $\mu^{(\mathbf{u})}$ is uniformly distributed on the (finite) set $\mathbf{X}^{(\mathbf{u})}$.

For example, if $\mathbf{X} = \mathcal{A}^{\mathbb{Z}^D}$, then the only Burton–Steif measure is the uniform Bernoulli measure. If $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}}$ is an irreducible SFT, then the only Burton–Steif measure is the Parry measure. If $r > 0$ and $\mathbb{B} := [-r \ldots r]^D \subset \mathbb{Z}^D$, and $\mathbf{X}$ is an SFT determined by a set of admissible words $\mathbf{X}_{\mathbb{B}} \subset \mathcal{A}^{\mathbb{B}}$, then it is easy to check that any Burton–Steif measure μ on $\mathbf{X}$ must be a Markov random field with interaction range r.

Theorem 7 (Burton and Steif) *Let $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^D}$ be a subshift of finite type.*

(a) *Any maxentropy measure on $\mathbf{X}$ is a Burton–Steif measure.*
(b) *If $\mathbf{X}$ is strongly irreducible, then any Burton–Steif measure on $\mathbf{X}$ is a maxentropy measure for $\mathbf{X}$.*

Proof **(a)** and **(b)** are Propositions 1.20 and 1.21 of [15], respectively. For a proof in the case when $\mathbf{X}$ is a symmetric nearest-neighbor subshift of finite type, see Propositions 1.19 and 4.1 of [14], respectively. □

Any subshift admits at least one maxentropy measure, so any SFT admits at least one Burton–Steif measure. Theorems 6 and 7 together imply:

Corollary 8 *If $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^D}$ is an SFT which admits a unique Burton–Steif measure $\eta_{\mathbf{X}}$, then $\eta_{\mathbf{X}}$ is the unique maxentropy measure for $\mathbf{X}$. Thus, if $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})$ and $\Phi(\mathbf{X}) = \mathbf{X}$, then $\Phi(\eta_{\mathbf{X}}) = \eta_{\mathbf{X}}$.*

Example 9 If $\mathbf{X} = \mathcal{A}^{\mathbb{Z}^D}$, then we get Theorem 1, because the unique Burton–Steif measure on $\mathcal{A}^{\mathbb{Z}^D}$ is the uniform Bernoulli measure.

Remark If $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$ is a subshift admitting a unique maxentropy measure μ, and $\mathsf{supp}(\mu) = \mathbf{X}$, then Wiess (see Theorem 4.2 in [143]) has observed that $\mathbf{X}$ automatically satisfies Incompressibility Corollary 4. In particular, this applies to any SFT having a unique Burton–Steif measure.

Periodic Invariant Measures

If $P \in \mathbb{N}$, then a sequence $\mathbf{a} \in \mathcal{A}^{\mathbb{Z}}$ is *P-periodic* if $\sigma^P(\mathbf{a}) = \mathbf{a}$. If $A := |\mathcal{A}|$, then there are exactly A^P such sequences, and a measure μ on $\mathcal{A}^{\mathbb{Z}}$ is called *P-periodic* if μ is supported entirely on these P-periodic sequences. More generally, if $\mathbb{M}$ is any monoid and $\mathbb{P} \subset \mathbb{M}$ is any submonoid, then a configuration $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$ is *$\mathbb{P}$-periodic* if $\sigma^{\mathsf{p}}(\mathbf{a}) = \mathbf{a}$ for all $\mathsf{p} \in \mathbb{P}$. (For example, if $\mathbb{M} = \mathbb{Z}$ and $\mathbb{P} := P\mathbb{Z}$, then the $\mathbb{P}$-periodic configurations are the P-periodic sequences). Let $\mathcal{A}^{\mathbb{M}/\mathbb{P}}$ denote the set of $\mathbb{P}$-periodic configurations. If $P := |\mathbb{M}/\mathbb{P}|$, then $|\mathcal{A}^{\mathbb{M}/\mathbb{P}}| = A^P$. A measure μ is called *$\mathbb{P}$-periodic* if $\mathsf{supp}(\mu) \subseteq \mathcal{A}^{\mathbb{M}/\mathbb{P}}$.

Proposition 10 *Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})$. If $\mathbb{P} \subset \mathbb{M}$ is any submonoid and $|\mathbb{M}/\mathbb{P}|$ is finite, then there exists a $\mathbb{P}$-periodic, Φ-invariant measure.*

Proof sketch If $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})$, then $\Phi(\mathcal{A}^{\mathbb{M}/\mathbb{P}}) \subseteq \mathcal{A}^{\mathbb{M}/\mathbb{P}}$. Thus, if μ is $\mathbb{P}$-periodic, then $\Phi^t(\mu)$ is $\mathbb{P}$-periodic for all $t \in \mathbb{N}$. Thus, the Cesàro limit of the sequence $\{\Phi^t(\mu)\}_{t=1}^{\infty}$ is $\mathbb{P}$-periodic and Φ-invariant. This Cesàro limit exists because $\mathcal{A}^{\mathbb{M}/\mathbb{P}}$ is finite. □

These periodic measures have finite (hence discrete) support, but by convex-combining them, it is easy to obtain (nonergodic) Φ-invariant measures with countable, dense support. When studying the invariant measures of CA, we usually regard these periodic measures (and their convex combinations) as somewhat trivial, and concentrate instead on invariant measures supported on aperiodic configurations.

Posexpansive and Permutative CA

Let $\mathbb{B} \subset \mathbb{M}$ be a finite subset, and let $\mathcal{B} := \mathcal{A}^{\mathbb{B}}$. If $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})$, then we define a continuous function $\Phi_{\mathbb{B}}^{\mathbb{N}} : \mathcal{A}^{\mathbb{M}} \longrightarrow \mathcal{B}^{\mathbb{N}}$ by

$$\Phi_{\mathbb{B}}^{\mathbb{N}}(\mathbf{a}) := [\mathbf{a}_{\mathbb{B}}; \Phi(\mathbf{a})_{\mathbb{B}}; \Phi^2(\mathbf{a})_{\mathbb{B}}; \Phi^3(\mathbf{a})_{\mathbb{B}}; \ldots] \in \mathcal{B}^{\mathbb{N}}\,. \quad (2)$$

Clearly, $\Phi_{\mathbb{B}}^{\mathbb{N}} \circ \Phi = \sigma \circ \Phi_{\mathbb{B}}^{\mathbb{N}}$. We say that Φ is *$\mathbb{B}$-posexpansive* if $\Phi_{\mathbb{B}}^{\mathbb{N}}$ is injective. Equivalently, for any $\mathbf{a}, \mathbf{a}' \in$

$\mathcal{A}^{\mathbb{M}}$, if $\mathbf{a} \neq \mathbf{a}'$, then there is some $t \in \mathbb{N}$ such that $\Phi^t(\mathbf{a})_{\mathbb{B}} \neq \Phi^t(\mathbf{a}')_{\mathbb{B}}$. We say Φ is *positively expansive* (or *posexpansive*) if Φ is $\mathbb{B}$-posexpansive for some finite $\mathbb{B}$ (it is easy to see that this is equivalent to the usual definition of positive expansiveness a topological dynamical system). For more information see Sect. 8 ► Topological Dynamics of Cellular Automata.

Thus, if $\mathbf{X} := \Phi_{\mathbb{B}}^{\mathbb{N}}(\mathcal{A}^{\mathbb{M}}) \subset \mathcal{B}^{\mathbb{N}}$, then $\mathbf{X}$ is a compact, shift-invariant subset of $\mathcal{B}^{\mathbb{N}}$, and $\Phi_{\mathbb{B}}^{\mathbb{N}} : \mathcal{A}^{\mathbb{M}} \longrightarrow \mathbf{X}$ is an isomorphism from the system $(\mathcal{A}^{\mathbb{M}}, \Phi)$ to the one-sided subshift $(\mathbf{X}, \sigma)$, which is sometimes called the *canonical factor* or *column shift* of Φ. The easiest examples of posexpansive CA are one-dimensional, permutative automata.

Proposition 11 ***(a)*** *Suppose $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{N}})$ has neighborhood $[r \ldots R]$, where $0 \leq r < R$. Let $\mathbb{B} := [0 \ldots R)$ and let $\mathcal{B} := \mathcal{A}^{\mathbb{B}}$. Then*

$$(\Phi \text{ is right permutative}) \iff (\Phi \text{ is } \mathbb{B}\text{-posexpansive, and } \Phi_{\mathbb{B}}^{\mathbb{N}}(\mathcal{A}^{\mathbb{N}}) = \mathcal{B}^{\mathbb{N}}).$$

(b) *Suppose $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ has neighborhood $[-L \ldots R]$, where $-L < 0 < R$. Let $\mathbb{B} := [-L \ldots R)$, and let $\mathcal{B} := \mathcal{A}^{\mathbb{B}}$. Then*

$$(\Phi \text{ is bipermutative}) \iff (\Phi \text{ is } \mathbb{B}\text{-posexpansive, and } \Phi^{\mathbb{N}_{\mathbb{B}}}(\mathcal{A}^{\mathbb{Z}}) = \mathcal{B}^{\mathbb{N}}).$$

Thus, one-sided, right-permutative CA and two-sided, bipermutative CA are both topologically conjugate the one-sided full shift $(\mathcal{B}^{\mathbb{N}}, \sigma)$, where $\mathcal{B}$ is an alphabet with $|\mathcal{A}|^{R+L}$ symbols (setting $L = 0$ in the one-sided case).

Proof Suppose $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$ (where $\mathbb{M} = \mathbb{N}$ or $\mathbb{Z}$). Draw a picture of the spacetime diagram for Φ. For any $t \in \mathbb{N}$, and any $\mathbf{b} \in \mathcal{B}^{[0 \ldots t)}$, observe how (bi)permutativity allows you to reconstruct a unique $\mathbf{a}_{[-tL \ldots tR)} \in \mathcal{A}^{[-tL \ldots tR)}$ such that $\mathbf{b} = (\mathbf{a}_{\mathbb{B}}, \Phi(\mathbf{a})_{\mathbb{B}}, \Phi^2(\mathbf{a})_{\mathbb{B}}, \ldots, \Phi^{t-1}(\mathbf{a})_{\mathbb{B}})$. By letting $t \to \infty$, we see that the function $\Phi_{\mathbb{B}}^{\mathbb{N}}$ is a bijection between $\mathcal{A}^{\mathbb{M}}$ and $\mathcal{B}^{\mathbb{N}}$. □

Remark 12 (a) The idea of Proposition 11 is implicit in Theorem 6.7 in [58], but it was apparently first stated explicitly by Shereshevsky and Afraĭmovich (see Theorem 1 in [130]). It was later rediscovered by Kleveland (see Corollary 7.3 in [74]) and Fagnani and Margara (see Theorem 3.2 in [35]).

(b) Proposition 11(b) has been generalized to higher dimensions by Allouche and Skordev (see Proposition 1 in [3]), who showed that, if $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}^D})$ is permutative in the 'corner' entries of its neighborhood, then Φ is conjugate to a full shift $(\mathcal{K}^{\mathbb{N}}, \sigma)$, where $\mathcal{K}$ is an uncountable, compact space.

Proposition 11 is quite indicative of the general case. Posexpansiveness occurs only in one-dimensional CA, in which it takes a very specific form. To explain this, suppose $(\mathbb{M}, \cdot)$ is a group with finite generating set $\mathbb{G} \subset \mathbb{M}$. For any $r > 0$, let $\mathbb{B}(r) := \{\mathsf{g}_1 \cdot \mathsf{g}_2 \cdots \mathsf{g}_r ; \mathsf{g}_1, \ldots, \mathsf{g}_r \in \mathbb{G}\}$. The *dimension* (or *growth degree*) of $(\mathbb{M}, \cdot)$ is defined $\dim(\mathbb{M}, \cdot) := \limsup_{r \to \infty} \log |\mathbb{B}(r)| / \log(r)$; see [50] or [49]. It can be shown that this number is independent of the choice of generating set $\mathbb{G}$, and is always an integer. For example, $\dim(\mathbb{Z}^D, +) = D$. If $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$ is a subshift, then we define its topological entropy $h_{\text{top}}(\mathbf{X})$ with respect to $\dim(\mathbb{M})$ in the obvious fashion (see Example 75(a)).

Theorem 13 *Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})$.*

(a) *If $\mathbb{M} = \mathbb{Z}^D \times \mathbb{N}^E$ with $D + E \geq 2$, then Φ cannot be posexpansive.*

(b) *If $\mathbb{M}$ is any group with $\dim(\mathbb{M}) \geq 2$, and $\mathbf{X} \subseteq \mathcal{A}^{\mathbb{M}}$ is any subshift with $h_{\text{top}}(\mathbf{X}) > 0$, and $\Phi(\mathbf{X}) \subseteq \mathbf{X}$, then the system $(\mathbf{X}, \Phi)$ cannot be posexpansive.*

(c) *Suppose $\mathbb{M} = \mathbb{Z}$ or $\mathbb{N}$, and Φ has neighborhood $[-L \ldots .R] \subset \mathbb{M}$. Let $\overline{L} := \max\{0, L\}$, $\overline{R} := \max\{0, R\}$ and $\mathbb{B} := [-\overline{L} \ldots \overline{R})$. If Φ is posexpansive, then Φ is $\mathbb{B}$-posexpansive.*

Proof **(a)** is Corollary 2 in [127]; see also Theorem 4.4 in [37]. Part **(b)** follows by applying Theorem 1.1 in [128] to the natural extension of $(\mathbf{X}, \Phi)$.

(c) The case $\mathbb{M} = \mathbb{Z}$ is Proposition 7 in [75]. The case $\mathbb{M} = \mathbb{N}$ is Proposition 2.3 in [9]. □

Proposition 11 says bipermutative CA on $\mathcal{A}^{\mathbb{Z}}$ are conjugate to full shifts. Using his formidable theory of *textile systems*, Nasu extended this to *all* posexpansive CA on $\mathcal{A}^{\mathbb{Z}}$.

Theorem 14 (Nasu's) *Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ and let $\mathbb{B} \subset \mathbb{Z}$. If Φ is $\mathbb{B}$-posexpansive, then $\Phi_{\mathbb{B}}^{\mathbb{N}}(\mathcal{A}^{\mathbb{Z}}) \subseteq \mathcal{B}^{\mathbb{N}}$ is a one-sided SFT which is conjugate to a one-sided full shift $C^{\mathbb{N}}$ for some alphabet C with $|C| \geq 3$.*

Proof sketch The fact that $\mathbf{X} := \Phi^{\mathbb{N}_{\mathbb{B}}}(\mathcal{A}^{\mathbb{Z}})$ is an SFT follows from Theorem 10 in [75] or Theorem 10.1 in [76]. Next, Theorem 3.12(1) on p. 49 of [105] asserts that, if Φ is *any* surjective endomorphism of an irreducible, aperiodic, SFT $\mathbf{Y} \subseteq \mathcal{A}^{\mathbb{Z}}$, and $(\mathbf{Y}, \Phi)$ is itself conjugate to an SFT, then $(\mathbf{Y}, \Phi)$ is actually conjugate to a full shift $(C^{\mathbb{N}}, \sigma)$ for some alphabet C with $|C| \geq 3$. Let $\mathbf{Y} := \mathcal{A}^{\mathbb{Z}}$ and invoke Kůrka's result.

For a direct proof not involving textile systems, see Theorem 4.9 in [86]. □

Remark 15 (a) See Theorem 60(d) for an 'ergodic' version of Theorem 14.

(b) In contrast to Proposition 11, Nasu's Theorem 14 does *not* say that $\Phi_{\mathbb{B}}^{\mathbb{N}}(\mathcal{A}^{\mathbb{Z}})$ itself is a full shift – only that it is conjugate to one.

If $(\mathbf{X}, \mu; \Psi)$ is a measure-preserving dynamical system (MPDS) with sigma-algebra $\mathfrak{B}$, then a *one-sided generator* is a finite partition $\mathfrak{P} \subset \mathfrak{B}$ such that $\bigvee_{t=0}^{\infty} \Psi^{-t}\mathfrak{P} \underset{\mu}{=} \mathfrak{B}$. If $\mathfrak{P}$ has C elements, and C is a finite set with $|C| = C$, then $\mathfrak{P}$ induces an essentially injective function $\mathfrak{p}\colon \mathbf{X} \longrightarrow C^{\mathbb{N}}$ such that $\mathfrak{p} \circ \Psi = \sigma \circ \mathfrak{p}$. Thus, if $\lambda := \mathfrak{p}(\mu)$, then $(\mathbf{X}, \mu; \Psi)$ is measurably isomorphic to the (one-sided) stationary stochastic process $(C^{\mathbb{N}}, \lambda; \sigma)$. If Ψ is invertible, then a (two-sided) *generator* is a finite partition $\mathfrak{P} \subset \mathfrak{B}$ such that $\bigvee_{t=-\infty}^{\infty} \Psi^{t}\mathfrak{P} \underset{\mu}{=} \mathfrak{B}$. The Krieger Generator Theorem says every finite-entropy, invertible MPDS has a generator; indeed, if $h(\Psi, \mu) \leq \log_2(C)$, then $(\mathbf{X}, \mu; \Psi)$ has a generator with C or less elements (▶ Ergodic Theory: Basic Examples and Constructions). If $|C| = C$, then once again, $\mathfrak{P}$ induces a measurable isomorphism from $(\mathbf{X}, \mu; \Psi)$ to a two-sided stationary stochastic process $(C^{\mathbb{Z}}, \lambda; \sigma)$, for some stationary measure λ on $\mathcal{A}^{\mathbb{Z}}$.

Corollary 16 (Universal Representation) *Let $\mathbb{M} = \mathbb{N}$ or $\mathbb{Z}$, and let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})$ have neighborhood $\mathbb{H} \subset \mathbb{M}$. Suppose that*

either *$\mathbb{M} = \mathbb{N}$, Φ is right-permutative, and $\mathbb{H} = [r \ldots R]$ for some $0 \leq r < R$, and then let $C := R \log_2 |\mathcal{A}|$;*

or *$\mathbb{M} = \mathbb{Z}$, Φ is bipermutative, and $\mathbb{H} = [-L \ldots R]$, and then let $C := (\overline{L} + \overline{R}) \log_2 |\mathcal{A}|$ where $\overline{L} := \max\{0, L\}$ and $\overline{R} := \max\{0, R\}$;*

or *$\mathbb{M} = \mathbb{Z}$ and Φ is positively expansive, and $h_{\text{top}}(\mathcal{A}^{\mathbb{M}}, \Phi) = \log_2(C)$ for some $C \in \mathbb{N}$.*

(a) *Let $(\mathbf{X}, \mu; \Psi)$ be any MPDS with a one-sided generator having at most C elements. Then there exists $\nu \in \mathfrak{M}_{\mathsf{cms}}(\mathcal{A}^{\mathbb{M}}; \Phi)$ such that the system $(\mathcal{A}^{\mathbb{M}}, \nu; \Phi)$ is measurably isomorphic to $(\mathbf{X}, \mu; \Psi)$.*

(b) *Let $(\mathbf{X}, \mu; \Psi)$ be an invertible MPDS, with measurable entropy $h(\mu, \phi) \leq \log_2(C)$. Then there exists $\nu \in \mathfrak{M}_{\mathsf{cms}}(\mathcal{A}^{\mathbb{M}}; \Phi)$ such that the natural extension of the system $(\mathcal{A}^{\mathbb{M}}, \nu; \Phi)$ is measurably isomorphic to $(\mathbf{X}, \mu, \Psi)$.*

Proof Under each of the three hypotheses, Proposition 11 or Theorem 14 yields a topological conjugacy $\Gamma\colon (C^{\mathbb{N}}, \sigma) \longrightarrow (\mathcal{A}^{\mathbb{M}}, \Phi)$, where C is a set of cardinality C.
(a) As discussed above, there is a measure λ on $C^{\mathbb{N}}$ such that $(C^{\mathbb{N}}, \lambda; \sigma)$ is measurably isomorphic to $(\mathbf{X}, \mu, \Psi)$. Thus, $\nu := \Gamma[\lambda]$ is a Φ-invariant measure on $\mathcal{A}^{\mathbb{M}}$, and $(\mathcal{A}^{\mathbb{M}}, \nu; \Phi)$ is isomorphic to $(C^{\mathbb{N}}, \mu, \Psi)$ via Γ.
(b) As discussed above, there is a measure λ on $C^{\mathbb{Z}}$ such that $(C^{\mathbb{Z}}, \lambda; \sigma)$ is measurably isomorphic to $(\mathbf{X}, \mu, \Psi)$. Let $\lambda_{\mathbb{N}}$ be the projection of λ to $C^{\mathbb{N}}$; then $(C^{\mathbb{N}}, \lambda_{\mathbb{N}}; \sigma)$ is a one-sided stationary process. Thus, $\nu := \Gamma[\lambda_{\mathbb{N}}]$ is a Φ-invariant measure on $\mathcal{A}^{\mathbb{M}}$, and $(\mathcal{A}^{\mathbb{M}}, \nu; \Phi)$ is isomorphic to $(C^{\mathbb{N}}, \lambda_{\mathbb{N}}; \sigma)$ via Γ. Thus, the natural extension of $(\mathcal{A}^{\mathbb{M}}, \nu; \Phi)$ is isomorphic to the natural extension of $(C^{\mathbb{N}}, \lambda_{\mathbb{N}}; \sigma)$, which is $(C^{\mathbb{Z}}, \lambda; \sigma)$, which is in turn isomorphic to $(\mathbf{X}, \mu; \Psi)$. □

Remark 17 The Universal Representation Corollary implies that studying the measurable dynamics of the CA Φ with respect to some arbitrary Φ-invariant measure ν will generally tell us nothing whatsoever about Φ. For these measurable dynamics to be meaningful, we must pick a measure on $\mathcal{A}^{\mathbb{M}}$ which is somehow 'natural' for Φ. First, this measure should be shift-invariant (because one of the defining properties of CA is that they commute with the shift). Second, we should seek a measure which has maximal Φ-entropy or is distinguished in some other way. (In general, the measures ν given by the Universal Representation Corollary will neither be σ-invariant, nor have maximal entropy for Φ.)

If $\Phi_{\mathbb{N}} \in \mathsf{CA}(\mathcal{A}^{\mathbb{N}})$, and $\Phi_{\mathbb{Z}} \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ is the CA obtained by applying the same local rule to all coordinates in $\mathbb{Z}$, then $\Phi_{\mathbb{Z}}$ can *never* be posexpansive: if $\mathbb{B} = [-B \ldots B]$, and $\mathbf{a}, \mathbf{a}' \in \mathcal{A}^{\mathbb{Z}}$ are any two sequences such that $\mathbf{a}_{(-\infty \ldots -B)} \neq \mathbf{a}'_{(-\infty \ldots -B)}$, then $\Phi^t(\mathbf{a})_{\mathbb{B}} = \Phi^t(\mathbf{a}')_{\mathbb{B}}$ for all $t \in \mathbb{N}$, because the local rule of Φ only propagates information to the left. Thus, in particular, the posexpansive CA on $\mathcal{A}^{\mathbb{Z}}$ are completely unrelated to the posexpansive CA on $\mathcal{A}^{\mathbb{N}}$. Nevertheless, posexpansive CA on $\mathcal{A}^{\mathbb{N}}$ behave quite similarly to those on $\mathcal{A}^{\mathbb{Z}}$.

Theorem 18 *Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{N}})$ have neighborhood $[r \ldots R]$, where $0 \leq r < R$, and let $\mathbb{B} := [0 \ldots R)$. Suppose Φ is posexpansive. Then:*

(a) *$\mathbf{X} := \Phi_{\mathbb{B}}^{\mathbb{N}}(\mathcal{A}^{\mathbb{N}}) \subseteq \mathcal{B}^{\mathbb{N}}$ is a topologically mixing SFT.*

(b) *The topological entropy of Φ is $\log_2(k)$ for some $k \in \mathbb{N}$.*

(c) *If η is the uniform measure on $\mathcal{A}^{\mathbb{N}}$, then $\Phi_{\mathbb{B}}^{\mathbb{N}}(\eta)$ is the Parry measure on $\mathbf{X}$. Thus, η is the maxentropy measure for Φ.*

Proof See Corollary 3.7 and Theorems 3.8 and 3.9 in [9] or Theorem 4.8(1,2,4) in [86]. □

Remark (a) See Theorem 58 for an 'ergodic' version of Theorem 18.

(b) The analog of Nasu's Theorem 14 (i. e. conjugacy to a full shift) is *not* true for posexpansive CA on $\mathcal{A}^{\mathbb{N}}$. See [13] for a counterexample.

(c) If $\Phi\colon \mathcal{A}^{\mathbb{N}} \longrightarrow \mathcal{A}^{\mathbb{N}}$ is *invertible*, then we define the function $\Phi_{\mathbb{B}}^{\mathbb{Z}}\colon \mathcal{A}^{\mathbb{N}} \longrightarrow \mathcal{B}^{\mathbb{Z}}$ by extending the defini-

tion of $\Phi_{\mathbb{B}}^{\mathbb{N}}$ to negative times. We say that Φ is *expansive* if $\Phi_{\mathbb{B}}^{\mathbb{Z}}$ is bijective for some finite $\mathbb{B} \subset \mathbb{N}$. Expansiveness is a much weaker condition than *positive* expansiveness. Nevertheless, the analog of Theorem 18(a) is true: if $\Phi\colon \mathcal{A}^{\mathbb{N}} \longrightarrow \mathcal{A}^{\mathbb{N}}$ is invertible and expansive, then $\mathcal{B}^{\mathbb{Z}}$ is conjugate to a (two-sided) subshift of finite type; see Theorem 1.3 in [106].

Measure Rigidity in Algebraic CA

Theorem 1 makes the uniform measure η a 'natural' invariant measure for a surjective CA Φ. However, Proposition 10 and Corollary 16 indicate that there are many other (unnatural) Φ-invariant measures as well. Thus, it is natural to seek conditions under which the uniform measure η is the unique (or almost unique) measure which is Φ-invariant, shift-invariant, and perhaps 'nondegenerate' in some other sense – a phenomenon which is sometimes called *measure rigidity*. Measure rigidity has been best understood when Φ is compatible with an underlying algebraic structure on $\mathcal{A}^{\mathbb{M}}$.

Let $\star\colon \mathcal{A}^{\mathbb{M}} \times \mathcal{A}^{\mathbb{M}} \longrightarrow \mathcal{A}^{\mathbb{M}}$ be a binary operation ('multiplication') and let $\bullet^{-1}\colon \mathcal{A}^{\mathbb{M}} \longrightarrow \mathcal{A}^{\mathbb{M}}$ be an unary operation ('inversion') such that $(\mathcal{A}^{\mathbb{M}}, \star)$ is a group, and suppose both operations are continuous and commute with all $\mathbb{M}$-shifts; then $(\mathcal{A}^{\mathbb{M}}, \star)$ is called a *group shift*. For example, if $(\mathcal{A}, \cdot)$ is itself a finite group, and $\mathcal{A}^{\mathbb{M}}$ is treated as a Cartesian product and endowed with componentwise multiplication, then $(\mathcal{A}^{\mathbb{M}}, \cdot)$ is a group shift. However, not all group shifts arise in this manner; see [70,71,72,73,124]. If $(\mathcal{A}^{\mathbb{M}}, \star)$ is a group shift, then a *subgroup shift* is a closed, shift-invariant subgroup $\mathbf{G} \subset \mathcal{A}^{\mathbb{M}}$ (i. e. $\mathbf{G}$ is both a subshift and a subgroup).

If $(\mathbf{G}, \star)$ is a subgroup shift, then the *Haar measure* on $\mathbf{G}$ is the unique probability measure $\eta_{\mathbf{G}}$ on $\mathbf{G}$ which is invariant under translation by all elements of $\mathbf{G}$. That is, if $\mathbf{g} \in \mathbf{G}$, and $\mathbf{U} \subset \mathbf{G}$ is any measurable subset, and $\mathbf{U} \star \mathbf{g} := \{\mathbf{u} \star \mathbf{g}; \mathbf{u} \in \mathbf{U}\}$, then $\eta_{\mathbf{G}}[\mathbf{U} \star \mathbf{g}] = \eta_{\mathbf{G}}[\mathbf{U}]$. In particular, if $\mathbf{G} = \mathcal{A}^{\mathbb{M}}$, then $\eta_{\mathbf{G}}$ is just the uniform Bernoulli measure on $\mathcal{A}^{\mathbb{M}}$. The Haar measure is a maxentropy measure on $\mathbf{G}$ (see Subsect. "Invariance of Maxentropy Measures").

If $(\mathcal{A}^{\mathbb{M}}, \star)$ is a group shift, and $\mathbf{G} \subseteq \mathcal{A}^{\mathbb{M}}$ is a subgroup shift, and $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})$, then Φ is called an *endomorphic* (or *algebraic*) CA on $\mathbf{G}$ if $\Phi(\mathbf{G}) \subseteq \mathbf{G}$ and $\Phi\colon \mathbf{G} \longrightarrow \mathbf{G}$ is an endomorphism of $(\mathbf{G}, \star)$ as a topological group. Let $\mathsf{ECA}(\mathbf{G}, \star)$ denote the set of endomorphic CA on $\mathbf{G}$. For example, suppose $(\mathcal{A}, +)$ is abelian, and let $(\mathbf{G}, \star) := (\mathcal{A}^{\mathbb{M}}, +)$ with the product group structure; then the endomorphic CA on $\mathcal{A}^{\mathbb{M}}$ are exactly the linear CA. However, if $(\mathcal{A}, \cdot)$ is a *non*abelian group, then endomorphic CA on $(\mathcal{A}^{\mathbb{M}}, \cdot)$ are *not* the same as multiplicative CA.

Even in this context, CA admit many nontrivial invariant measures. For example, it is easy to check the following:

Proposition 19 *Let $\mathcal{A}^{\mathbb{M}}$ be a group shift and let $\Phi \in \mathsf{ECA}(\mathcal{A}^{\mathbb{M}}, \star)$. Let $\mathbf{G} \subseteq \mathcal{A}^{\mathbb{M}}$ be any Φ-invariant subgroup shift; then the Haar measure on $\mathbf{G}$ is Φ-invariant.*

For example, if $(\mathcal{A}, +)$ is any nonsimple abelian group, and $(\mathcal{A}^{\mathbb{M}}, +)$ has the product group structure, then $\mathcal{A}^{\mathbb{M}}$ admits many nontrivial subgroup shifts; see [73]. If Φ is any linear CA on $\mathcal{A}^{\mathbb{M}}$ with scalar coefficients, then *every* subgroup shift of $\mathcal{A}^{\mathbb{M}}$ is Φ-invariant, so Proposition 19 yields many nontrivial Φ-invariant measures. To isolate η as a unique measure, we must impose further restrictions. The first nontrivial results in this direction were by Host, Maass, and Martínez [61]. Let $h(\Phi, \mu)$ be the entropy of Φ relative to the measure μ (see Sect. "Entropy" for definition).

Proposition 20 *Let $\mathcal{A} := \mathbb{Z}_{/p}$, where p is prime. Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ be a linear CA with neighborhood $\{0, 1\}$, and let $\mu \in \mathfrak{M}_{\mathrm{ens}}(\mathcal{A}^{\mathbb{Z}}; \Phi, \sigma)$. If μ is σ-ergodic, and $h(\Phi, \mu) > 0$, then μ is the Haar measure η on $\mathcal{A}^{\mathbb{Z}}$.*

Proof See Theorem 12 in [61]. □

A similar idea is behind the next result, only with the roles of Φ and σ reversed. If μ is a measure on $\mathcal{A}^{\mathbb{N}}$, and $\mathbf{b} \in \mathcal{A}^{[1\ldots\infty)}$, then we define the conditional measure $\mu^{(\mathbf{b})}$ on $\mathcal{A}$ by $\mu^{(\mathbf{b})}(a) := \mu[x_0 = a | \mathbf{x}_{[1\ldots\infty)} = \mathbf{b}]$, where $\mathbf{x}$ is a μ-random sequence. For example, if μ is a Bernoulli measure, then $\mu^{(\mathbf{b})}(a) = \mu[x_0 = a]$, independent of $\mathbf{b}$; if μ is a Markov measure, then $\mu^{(\mathbf{b})}(a) = \mu[x_0 = a | x_1 = b_1]$.

Proposition 21 *Let $(\mathcal{A}, \cdot)$ be any finite (possibly nonabelian) group, and let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{N}})$ have multiplicative local rule $\phi\colon \mathcal{A}^{\{0,1\}} \longrightarrow \mathcal{A}$ defined by $\phi(a_0, a_1) := a_0 \cdot a_1$. Let $\mu \in \mathfrak{M}_{\mathrm{ens}}(\mathcal{A}^{\mathbb{Z}}; \Phi, \sigma)$. If μ is Φ-ergodic, then there is some subgroup $C \subset \mathcal{A}$ such that, for every $\mathbf{b} \in \mathcal{A}^{[1\ldots\infty]}$, $\mathsf{supp}(\mu^{(\mathbf{b})})$ is a right coset of C, and $\mu^{(\mathbf{b})}$ is uniformly distributed on this coset.*

Proof See Theorem 3.1 in [113]. □

Example 22 Let Φ and μ be as in Proposition 21. Let η be the Haar measure on $\mathcal{A}^{\mathbb{N}}$

(a) μ has *complete connections* if $\mathsf{supp}(\mu^{(\mathbf{b})}) = \mathcal{A}$ for μ-almost all $\mathbf{b} \in \mathcal{A}^{[1\ldots\infty)}$. Thus, if μ has complete connections in Proposition 21, then $\mu = \eta$.

(b1) Suppose $h(\mu, \sigma) > h_0 := \max\{\log_2 |C|\,;\, C$ a proper subgroup of $\mathcal{A}\}$. Then $\mu = \eta$.

(b2) In particular, suppose $\mathcal{A} = (\mathbb{Z}_{/p}, +)$, where p is prime; then $h_0 = 0$. Thus, if Φ has local rule $\phi(a_0, a_1) := a_0 + a_1$, and μ is any σ-invariant, Φ-ergodic measure with $h(\mu, \sigma) > 0$, then $\mu = \eta$. This is closely analogous to Proposition 20, but 'dual' to it, because the roles of Φ and σ are reversed in the ergodicity and entropy hypotheses.

(c) If $C \subset \mathcal{A}$ is a subgroup, and μ is the Haar measure on the subgroup shift $C^{\mathbb{N}} \subset \mathcal{A}^{\mathbb{N}}$, then μ satisfies the conditions of Proposition 21. Other, less trivial possibilities also exist (see Examples 3.2(b,c) in [113]).

If μ is a measure on $\mathcal{A}^{\mathbb{Z}}$, and $\mathbf{X}, \mathbf{Y} \subset \mathcal{A}^{\mathbb{Z}}$, then we say $\mathbf{X}$ *essentially equals* $\mathbf{Y}$ and write $\mathbf{X} \underset{\mu}{=} \mathbf{Y}$ if $\mu[\mathbf{X} \triangle \mathbf{Y}] = 0$. If $n \in \mathbb{N}$, then let

$$\mathfrak{I}_n(\mu) := \left\{ \mathbf{X} \subset \mathcal{A}^{\mathbb{Z}} \,;\; \sigma^n(\mathbf{X}) \underset{\mu}{=} \mathbf{X} \right\}$$

be the sigma-algebra of subsets of $\mathcal{A}^{\mathbb{Z}}$ which are 'essentially' σ^n-invariant. Thus, μ is σ-ergodic if and only if $\mathfrak{I}_1(\mu)$ is trivial (i. e. contains only sets of measure zero or one). We say μ is *totally σ-ergodic* if $\mathfrak{I}_n(\mu)$ is trivial for all $n \in \mathbb{N}$ (▶ Ergodicity and Mixing Properties).

Let $(\mathcal{A}^{\mathbb{Z}}, *)$ be any group shift. The identity element $\mathbf{e}$ of $(\mathcal{A}^{\mathbb{Z}}, *)$ is a constant sequence. Thus, if $\Phi \in \mathsf{ECA}(\mathcal{A}^{\mathbb{Z}}, *)$ is surjective, then $\ker(\Phi) := \{\mathbf{a} \in \mathcal{A}^{\mathbb{Z}};\ \Phi(\mathbf{a}) = \mathbf{e}\}$ is a finite, shift-invariant subgroup of $\mathcal{A}^{\mathbb{Z}}$ (i. e. a finite collection of σ-periodic sequences).

Proposition 23 *Let $(\mathcal{A}^{\mathbb{Z}}, *)$ be a (possibly nonabelian) group shift, and let $\Phi \in \mathsf{ECA}(\mathcal{A}^{\mathbb{Z}}, *)$ be bipermutative, with neighborhood $\{0, 1\}$. Let $\mu \in \mathfrak{M}_{\mathrm{ers}}(\mathcal{A}^{\mathbb{Z}}; \Phi, \sigma)$. Suppose that:*

(IE) *μ is totally ergodic for σ;* ***(H)*** *$h(\Phi, \mu) > 0$; and*

(K) *$\ker(\Phi)$ contains no nontrivial σ-invariant subgroups.*

Then μ is the Haar measure on $\mathcal{A}^{\mathbb{Z}}$.

Proof See Theorem 5.2 in [113]. □

Example 24 If $\mathcal{A} = \mathbb{Z}_{/p}$ and $(\mathcal{A}^{\mathbb{Z}}, +)$ is the product group, then Φ is a linear CA and condition (c) is automatically satisfied, so Proposition 23 becomes a special case of Proposition 20.

If $\Phi \in \mathsf{ECA}(\mathcal{A}^{\mathbb{Z}}, *)$, then we have an increasing sequence of finite, shift-invariant subgroups $\ker(\Phi) \subseteq \ker(\Phi^2) \subseteq \ker(\Phi^3) \subseteq \cdots$. If $\mathbf{K}(\Phi) := \bigcup_{n=1}^{\infty} \ker(\Phi^n)$, then $\mathbf{K}(\Phi)$ is a countable, shift-invariant subgroup of $(\mathcal{A}^{\mathbb{Z}}, *)$.

Theorem 25 *Let $(\mathcal{A}^{\mathbb{Z}}, +)$ be an abelian group shift, and let $\mathbf{G} \subseteq \mathcal{A}^{\mathbb{Z}}$ be a subgroup shift. Let $\Phi \in \mathsf{ECA}(\mathbf{G}, +)$ be bipermutative, and let $\mu \in \mathfrak{M}_{\mathrm{ers}}(\mathbf{G}; \Phi, \sigma)$. Suppose:*

(I) *$\mathfrak{I}_{kP}(\mu) = \mathfrak{I}_1(\mu)$, where P is the lowest common multiple of the σ-periods of all elements in $\ker(\Phi)$, and $k \in \mathbb{N}$ is any common multiple of all prime factors of $|\mathcal{A}|$.*

(H) *$h(\Phi, \mu) > 0$.*

Furthermore, suppose that either:

(E1) *μ is ergodic for the $\mathbb{N} \times \mathbb{Z}$ action (Φ, σ);*

(K1) *Every infinite, σ-invariant subgroup of $\mathbf{K}(\Phi) \cap \mathbf{G}$ is dense in $\mathbf{G}$;*

or:

(E2) *μ is σ-ergodic;*

(K2) *Every infinite, (Φ, σ)-invariant subgroup of $\mathbf{K}(\Phi) \cap \mathbf{G}$ is dense in $\mathbf{G}$.*

Then μ is the Haar measure on $\mathbf{G}$.

Proof See Theorems 3.3 and 3.4 of [122], or Théorèmes V.4 and V.5 on p. 115 of [120]. In the special case when $\mathbf{G}$ is irreducible and has topological entropy $\log_2(p)$ (where p is prime), Sobottka has given a different and simpler proof, by using his theory of 'quasigroup shifts' to establish an isomorphism between Φ and a linear CA on $\mathbb{Z}_{/p}$, and then invoking Theorem 7. See Theorems 7.1 and 7.2 of [136], or Teoremas IV.3.1 and IV.3.2 on pp. 100–101 of [134]. □

Example 26 (a) Let $\mathcal{A} := \mathbb{Z}_{/p}$, where p is prime. Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ be linear, and suppose that $\mu \in \mathfrak{M}_{\mathrm{ers}}(\mathcal{A}^{\mathbb{Z}}; \Phi, \sigma)$ is (Φ, σ)-ergodic, $h(\Phi, \mu) > 0$, and $\mathfrak{I}_{p(p-1)}(\mu) = \mathfrak{I}_1(\mu)$. Setting $k = p$ and $P = p - 1$ in Theorem 25, we conclude that μ is the Haar measure on $\mathcal{A}^{\mathbb{Z}}$. For Φ with neighborhood $\{0, 1\}$, this result first appeared as Theorem 13 in [61].

(b) If $(\mathcal{A}^{\mathbb{Z}}, *)$ is abelian, then Proposition 23 is a special case of Theorem 25 [hypothesis **(IE)** of the former implies hypotheses **(I)** and **(E2)** of the latter, while **(K)** implies **(K2)**]. Note, however, that Proposition 23 also applies to nonabelian groups.

An *algebraic $\mathbb{Z}^D$-action* is an action of $\mathbb{Z}^D$ by automorphisms on a compact abelian group $\mathbf{G}$. For example, if $\mathbf{G} \subseteq \mathcal{A}^{\mathbb{Z}^D}$ is an abelian subgroup shift, then σ is an algebraic $\mathbb{Z}^D$-action. The invariant measures of algebraic $\mathbb{Z}^D$-actions have been studied in Schmidt (see Sect. 29 in [124]), Silberger (see Sect. 7 in [132]), and Einsiedler [28,29].

If $\Phi \in \mathsf{CA}(\mathbf{G})$, then a *complete history* for Φ is a sequence $(\mathbf{g}_t)_{t \in \mathbb{Z}} \in \mathbf{G}^{\mathbb{Z}}$ such that $\Phi(\mathbf{g}_t) = \mathbf{g}_{t+1}$ for all $t \in \mathbb{Z}$. Let $\Phi^{\mathbb{Z}}(\mathbf{G}) \subset \mathbf{G}^{\mathbb{Z}} \subseteq (\mathcal{A}^{\mathbb{Z}^D})^{\mathbb{Z}} \cong \mathcal{A}^{\mathbb{Z}^{D+1}}$ be the set of all complete histories for Φ; then $\Phi^{\mathbb{Z}}(\mathbf{G})$ is a subshift of $\mathcal{A}^{\mathbb{Z}^{D+1}}$. If $\Phi \in \mathsf{ECA}[\mathbf{G}]$, then $\Phi^{\mathbb{Z}}(\mathbf{G})$ is itself an abelian subgroup shift, and the shift action of $\mathbb{Z}^{D+1}$ on $\Phi^{\mathbb{Z}}(\mathbf{G})$ is thus an algebraic $\mathbb{Z}^{D+1}$-action. Any (Φ, σ)-invariant measure on $\mathbf{G}$ extends in the obvious way to a σ-invariant measure on $\Phi^{\mathbb{Z}}(\mathbf{G})$. Thus, any result about the invariant measures (or rigidity) of algebraic $\mathbb{Z}^{D+1}$-actions can

be translated immediately into a result about the invariant measures (or rigidity) of endomorphic cellular automata.

Proposition 27 *Let $\mathbf{G} \subseteq \mathcal{A}^{\mathbb{Z}^D}$ be an abelian subgroup shift and let $\Phi \in ECA(\mathbf{G})$. Suppose $\mu \in \mathfrak{M}_{\text{eas}}(\mathbf{G}; \Phi, \sigma)$ is (Φ, σ)-totally ergodic, and has entropy dimension $d \in [1 \ldots D]$ (see Subsect. "Entropy Geometry and Expansive Subdynamics"). If the system $(\mathbf{G}, \mu; \Phi, \sigma)$ admits no factors whose d-dimensional measurable entropy is zero, then there is a Φ-invariant subgroup shift $\mathbf{G}' \subseteq \mathbf{G}$ and some element $\mathbf{x} \in \mathbf{G}$ such that μ is the translated Haar measure on the 'affine' subset $\mathbf{G}' + \mathbf{x}$.*

Proof This follows from Corollary 2.3 in [29]. □

If we remove the requirement of 'no zero-entropy factors', and instead require $\mathbf{G}$ and Φ to satisfy certain technical algebraic conditions, then μ must be the Haar measure on $\mathbf{G}$ (see Theorem 1.2 in [29]). These strong hypotheses are probably necessary, because in general, the system $(\mathbf{G}, \sigma, \Phi)$ admits uncountably many distinct nontrivial invariant measures, even if $(\mathbf{G}, \sigma, \Phi)$ is *irreducible*, meaning that $\mathbf{G}$ contains no proper, infinite, Φ-invariant subgroup shifts:

Proposition 28 *Let $\mathbf{G} \subseteq \mathcal{A}^{\mathbb{Z}^D}$ be an abelian subgroup shift, let $\Phi \in ECA(\mathbf{G})$, and suppose $(\mathbf{G}, \sigma, \Phi)$ is irreducible. For any $s \in [0, 1)$, there exists a (Φ, σ)-ergodic measure $\mu \in \mathfrak{M}_{\text{eas}}(\mathbf{G}; \Phi, \sigma)$ such that $h(\mu, \Phi^n \circ \sigma^z) = s \cdot h_{\text{top}}(\mathbf{G}, \Phi^n \circ \sigma^z)$ for every $n \in \mathbb{N}$ and $z \in \mathbb{Z}^D$.*

Proof This follows from Corollary 1.4 in [28]. □

Let $\mu \in \mathfrak{M}_{\text{eas}}(\mathcal{A}^{\mathbb{M}}; \sigma)$ and let $\mathbb{H} \subset \mathbb{M}$ be a finite subset. We say that μ is $\mathbb{H}$-*mixing* if, for any $\mathbb{H}$-indexed collection $\{\mathbf{U}_{\mathbf{h}}\}_{\mathbf{h} \in \mathbb{H}}$ of measurable subsets of $\mathcal{A}^{\mathbb{M}}$,

$$\lim_{n \to \infty} \mu \left[\bigcap_{\mathbf{h} \in \mathbb{H}} \sigma^{n\mathbf{h}}(\mathbf{U}_{\mathbf{h}}) \right] = \prod_{\mathbf{h} \in \mathbb{H}} \mu\left[\mathbf{U}_{\mathbf{h}}\right] .$$

For example, if $|\mathbb{H}| = H$, then any H-multiply σ-mixing measure (see Subsect. "Mixing and Ergodicity") is $\mathbb{H}$-mixing.

Proposition 29 *Let $\mathbf{G} \subseteq \mathcal{A}^{\mathbb{Z}^D}$ be an abelian subgroup shift and let $\Phi \in ECA(\mathbf{G})$ have neighborhood $\mathbb{H}$ (with $|\mathbb{H}| \geq 2$). Suppose $(\mathbf{G}, \sigma, \Phi)$ is irreducible, and let $\mu \in \mathfrak{M}_{\text{eas}}(\mathcal{A}^{\mathbb{Z}^D}; \Phi, \sigma)$. Then μ is $\mathbb{H}$-mixing if and only if μ is the Haar measure of $\mathbf{G}$.*

Proof This follows from [124], Corollary 29.5, p. 289 (note that Schmidt uses 'almost minimal' to mean 'irreducible'). A significant generalization of Proposition 29 appears in [115] □

The Furstenberg Conjecture

Let $\mathbb{T}^1 = \mathbb{R}/\mathbb{Z}$ be the circle group, which we identify with the interval $[0, 1)$. Define the functions $\times_2, \times_3 : \mathbb{T}^1 \longrightarrow \mathbb{T}^1$ by $\times_2(t) = 2t \pmod 1$ and $\times_3(t) = 3t \pmod 1$. Clearly, these maps commute, and preserve the Lebesgue measure on $\mathbb{T}^1$. Furstenberg [44] speculated that the *only* nonatomic $\times_2$- and $\times_3$-invariant measure on $\mathbb{T}^1$ was the Lebesgue measure. Rudolph [119] showed that, if ρ is $(\times_2, \times_3)$-invariant measure and *not* Lebesgue, then the systems $(\mathbb{T}^1, \rho, \times_2)$ and $(\mathbb{T}^1, \rho, \times_3)$ have zero entropy; this was later generalized in [60,69]. It is not known whether any nonatomic measures exist on $\mathbb{T}^1$ which satisfy Rudolph's conditions; this is considered an outstanding problem in abstract ergodic theory.

To see the connection between Furstenberg's Conjecture and cellular automata, let $\mathcal{A} = \{0, 1, 2, 3, 4, 5\}$, and define the surjection $\Psi : \mathcal{A}^{\mathbb{N}} \longrightarrow \mathbb{T}^1$ by mapping each $\mathbf{a} \in \mathcal{A}^{\mathbb{N}}$ to the element of $[0, 1)$ having $\mathbf{a}$ as its base-6 expansion. That is:

$$\Psi(a_0, a_1, a_2, \ldots) := \sum_{n=0}^{\infty} \frac{a^n}{6^{n+1}} .$$

The map Ψ is injective everywhere except on the countable set of sequences ending in $[000\ldots]$ or $[555\ldots]$ (on this set, Ψ is 2-to-1). Furthermore, Ψ defines a semiconjugacy from $\times_2$ and $\times_3$ into two CA on $\mathcal{A}^{\mathbb{N}}$. Let $\mathbb{H} := \{0, 1\}$, and define local maps $\xi_2, \xi_3 : \mathcal{A}^{\mathbb{H}} \longrightarrow \mathcal{A}$ as follows:

$$\xi_2(a_0, a_1) = \left[2a_0\right]_6 + \left\lceil \frac{a_1}{3} \right\rceil \text{ and}$$
$$\xi_3(a_0, a_1) = \left[3a_0\right]_6 + \left\lceil \frac{a_1}{2} \right\rceil ,$$

where, $[a]_6$ is the least residue of a, mod 6. If $\Xi_p \in \mathsf{CA}(\mathcal{A}^{\mathbb{N}})$ has local map ξ_p (for $p = 2, 3$), then it is easy to check that Ξ_p corresponds to multiplication by p in base-6 notation. In other words, $\Psi \circ \times_p = \Xi_p \circ \Psi$ for $p = 2, 3$.

If λ is the Lebesgue measure on $\mathbb{T}^1$, then $\Psi(\lambda) = \eta$, where η is the uniform Bernoulli measure on $\mathcal{A}^{\mathbb{N}}$. Thus, η is Ξ_2- and Ξ_3-invariant, and Furstenberg's Conjecture asserts that η is the *only* nonatomic measure on $\mathcal{A}^{\mathbb{N}}$ which is both Ξ_2- and Ξ_3-invariant. The shift map $\sigma : \mathcal{A}^{\mathbb{N}} \longrightarrow \mathcal{A}^{\mathbb{N}}$ corresponds to multiplication by 6 in base-6 notation. Hence, $\Xi_2 \circ \Xi_3 = \sigma$. From this it follows that a measure μ is (Ξ_2, Ξ_3)-invariant if and only if μ is (Ξ_2, σ)-invariant if and only if μ is (σ, Ξ_3)-invariant. Thus, Furstenberg's Conjecture equivalently asserts that η is the only stationary, Ξ_3-invariant nonatomic measure on $\mathcal{A}^{\mathbb{N}}$, and Rudolph's result asserts that η is the only such nonatomic measure with nonzero entropy; this is analogous to the

'measure rigidity' results of Subsect. "Measure Rigidity in Algebraic CA". The existence of zero-entropy, (σ, Ξ_3)-invariant, nonatomic measures remains an open question.

Remark 30 (a) There is nothing special about 2 and 3; the same results hold for any pair of prime numbers.

(b) Lyons [85] and Rudolph and Johnson [68] have also established that a wide variety of $\times_2$-invariant probability measures on $\mathbb{T}^1$ will weak* converge, under the iteration of $\times_3$, to the Lebesgue measure (and vice versa). In the terminology of Subsect. "Asymptotic Randomization by Linear Cellular Automata", these results immediately translate into equivalent statements about the 'asymptotic randomization' of initial probability measures on $\mathcal{A}^{\mathbb{N}}$ under the iteration of Ξ_2 or Ξ_3.

Domains, Defects, and Particles

Suppose $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$, and there is a collection of Φ-invariant subshifts $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_N \subset \mathcal{A}^{\mathbb{Z}}$ (called *phases*). Any sequence $\mathbf{a}$ can be expressed a finite or infinite concatenation

$$\mathbf{a} = [\ldots \mathbf{a}_{-2}\, \mathbf{d}_{-2}\, \mathbf{a}_{-1}\, \mathbf{d}_{-1}\, \mathbf{a}_0\, \mathbf{d}_0\, \mathbf{a}_1\, \mathbf{d}_1\, \mathbf{a}_2 \cdots] ,$$

where each *domain* $\mathbf{a}_k$ is a finite word (or half-infinite sequence) which is admissible to phase $\mathbf{P}_n$ for some $n \in [1 \ldots N]$, and where each *defect* $\mathbf{d}_k$ is a (possibly empty) finite word (note that this decomposition may not be unique). Thus, $\Phi(\mathbf{a}) = \mathbf{a}'$, where

$$\mathbf{a}' = [\ldots \mathbf{a}'_{-2}\, \mathbf{d}'_{-2}\, \mathbf{a}'_{-1}\, \mathbf{d}'_{-1}\, \mathbf{a}'_0\, \mathbf{d}'_0\, \mathbf{a}'_1\, \mathbf{d}'_1\, \mathbf{a}'_2 \cdots] ,$$

and, for every $k \in \mathbb{Z}$, $\mathbf{a}'_k$ belongs to the same phase as $\mathbf{a}_k$. We say that Φ has *stable phases* if, for any such $\mathbf{a}$ and $\mathbf{a}'$ in $\mathcal{A}^{\mathbb{Z}}$, it is the case that, for all $k \in \mathbb{Z}$, $|\mathbf{d}'_k| \leq |\mathbf{d}_k|$. In other words, the defects do not grow over time. However, they may propagate sideways; for example, $\mathbf{d}'_k$ may be slightly to the right of $\mathbf{d}_k$, if the domain $\mathbf{a}'_k$ is larger than $\mathbf{a}_k$, while the domain $\mathbf{a}'_{k+1}$ is slightly smaller than $\mathbf{a}_{k+1}$. If $\mathbf{a}_k$ and $\mathbf{a}_{k+1}$ belong to different phases, then the defect $\mathbf{d}_k$ is sometimes called a *domain boundary* (or 'wall', or 'edge particle'). If $\mathbf{a}_k$ and $\mathbf{a}_{k+1}$ belong to the same phase, then the defect $\mathbf{d}_k$ is sometimes called a *dislocation* (or 'kink').

Often $\mathbf{P}_n = \{\mathbf{p}\}$ where $\mathbf{p} = [\ldots ppp \ldots]$ is a constant sequence, or each $\mathbf{P}_n$ consists of the σ-orbit of a single periodic sequence. More generally, the phases $\mathbf{P}_1, \ldots, \mathbf{P}_N$ may be subshifts of finite type. In this case, most sequences in $\mathcal{A}^{\mathbb{Z}}$ can be fairly easily and unambiguously decomposed into domains separated by defects. However, if the phases are more complex (e. g. sofic shifts), then the exact definition of a 'defect' is actually fairly complicated – see [114] for a rigorous discussion.

Example 31 Let $\mathcal{A} = \{0, 1\}$ and let $\mathbb{H} = \{-1, 0, 1\}$. *Elementary cellular automaton* (ECA) #184 is the CA $\Phi : \mathcal{A}^{\mathbb{Z}} \longrightarrow \mathcal{A}^{\mathbb{Z}}$ with local rule $\phi : \mathcal{A}^{\mathbb{H}} \longrightarrow \mathcal{A}$ given as follows: $\phi(a_{-1}, a_0, a_1) = 1$ if $a_0 = a_1 = 1$, or if $a_{-1} = 1$ and $a_0 = 0$. On the other hand, $\phi(a_{-1}, a_0, a_1) = 0$ if $a_{-1} = a_0 = 0$, or if $a_1 = 0$ and $a_0 = 1$. Heuristically, each '1' represents a 'car' moving cautiously to the right on a single-lane road. During each iteration, each car will advance to the site in front of it, unless that site is already occupied, in which case the car will remain stationary. ECA#184 exhibits one stable phase $\mathbf{P}$, given by the 2-periodic sequence $[\ldots 0101.0101 \ldots]$ and its translate $[\ldots 1010.1010 \ldots]$ (here the decimal point indicates the zeroth coordinate), and Φ acts on $\mathbf{P}$ like the shift. The phase $\mathbf{P}$ admits two dislocations of width 2. The dislocation $\mathbf{d}_0 = [00]$ moves uniformly to the right, while the dislocation $\mathbf{d}_1 = [11]$ moves uniformly to the left. In the traffic interpretation, $\mathbf{P}$ represents freely flowing traffic, $\mathbf{d}_0$ represents a stretch of empty road, and $\mathbf{d}_1$ represents a traffic jam.

Example 32 Let $\mathcal{A} := \mathbb{Z}_{/N}$, and let $\mathbb{H} := [-1 \ldots 1]$. The one-dimensional, N-color *cyclic cellular automaton* (CCA_N) $\Phi : \mathcal{A}^{\mathbb{Z}} \longrightarrow \mathcal{A}^{\mathbb{Z}}$ has local rule $\phi : \mathcal{A}^{\mathbb{H}} \longrightarrow \mathcal{A}$ defined:

$$\phi(\mathbf{a}) := \begin{cases} a_0 + 1 & \text{if there is some } \mathsf{h} \in \mathbb{H} \\ & \qquad \text{with } a_{\mathsf{h}} = a_0 + 1 ; \\ a_0 & \text{otherwise} . \end{cases}$$

(here, addition is mod N). The CCA has phases $\mathbf{P}_0, \mathbf{P}_1, \ldots, \mathbf{P}_{N-1}$, where $\mathbf{P}_a = \{[\ldots aaa \ldots]\}$ for each $a \in \mathcal{A}$. A domain boundary between $\mathbf{P}_a$ and $\mathbf{P}_{a-1}$ moves with constant velocity towards the $\mathbf{P}_{a-1}$ side. All other domain boundaries are stationary.

In a *particle cellular automaton* (PCA), $\mathcal{A} = \{\emptyset\} \sqcup \mathcal{P}$, where $\mathcal{P}$ is a set of 'particle types' and $\emptyset$ represents a vacant site. Each particle $p \in \mathcal{P}$ is assigned some (constant) velocity vector $\mathsf{v}(p) \in (-\mathbb{H})$ (where $\mathbb{H}$ is the neighborhood of the automaton). Particles propagate with constant velocity through $\mathbb{M}$ until two particles try to simultaneously enter the same site in the lattice, at which point the outcome is determined by a *collision rule*: a stylized 'chemical reaction equation'. For example, an equation "$p_1 + p_2 \rightsquigarrow p_3$" means that, if particle types p_1 and p_2 collide, they coalesce to produce a particle of type p_3. On the other hand, "$p_1 + p_2 \rightsquigarrow \emptyset$" means that the two particles annihilate on contact. Formally, given a set of velocities and collision

rules, the local rule $\phi\colon \mathcal{A}^{\mathbb{H}} \longrightarrow \mathcal{A}$ is defined

$$\phi(\mathbf{a}) := \begin{cases} p & \text{if there is a unique } \mathsf{h} \in \mathbb{H} \text{ and } p \in \mathcal{P} \text{ with} \\ & a_{\mathsf{h}} = p \text{ and } \mathsf{v}(p) = -\mathsf{h} \ : \\ q & \text{if } \{p \in \mathcal{P}\,;\ a_{-\mathsf{v}(p)} = p\} = \{p_1, p_2, \ldots, p_n\}, \\ & \text{and } p_1 + \cdots + p_n \rightsquigarrow q\,. \end{cases}$$

Example 33 The one-dimensional *ballistic annihilation model* (BAM) contains two particle types: $\mathcal{P} = \{\pm 1\}$, with the following rules:

$$\mathsf{v}(1) = 1\,, \quad \mathsf{v}(-1) = -1\,, \quad \text{and} \quad -1 + 1 \rightsquigarrow \emptyset\,.$$

(This CA is sometimes also called *Just Gliders*.) Thus, $a_{\mathsf{z}} = 1$ if the cell z contains a particle moving to the right with velocity 1, whereas $a_{\mathsf{z}} = -1$ if the cell z contains a particle moving left with velocity -1, and $a_{\mathsf{z}} = \emptyset$ if cell z is vacant. Particles move with constant velocity until they collide with oncoming particles, at which point both particles are annihilated. If $\mathcal{B} := \{\pm 1, \emptyset\}$ and $\mathbb{H} = [-1 \ldots 1] \subset \mathbb{Z}$, then we can represent the BAM using $\Psi \in \mathsf{CA}(\mathcal{B}^{\mathbb{Z}})$ with local rule $\psi\colon \mathcal{B}^{\mathbb{H}} \longrightarrow \mathcal{B}$ defined:

$$\psi(b_{-1}, b_0, b_1) := \begin{cases} -1 & \text{if } b_1 = -1 \text{ and} \\ & \qquad b_{-1}, b_0 \in \{-1, \emptyset\}\,; \\ 1 & \text{if } b_{-1} = 1 \text{ and } b_0, b_1 \in \{1, \emptyset\}\,; \\ \emptyset & \text{otherwise}\,. \end{cases}$$

Particle CA can be seen as 'toy models' of particle physics or microscale chemistry. More interestingly, however, one-dimensional PCA often arise as factors of coalescent-domain CA, with the 'particles' tracking the motion of the defects.

Example 34 (a) Let $\mathcal{A} := \{0, 1\}$ and let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ be ECA#184. Let $\mathcal{B} := \{\pm 1, 0\}$, and let $\Psi \in \mathsf{CA}(\mathcal{B}^{\mathbb{Z}})$ be the BAM. Let $\mathbb{G} := \{0, 1\}$, and let $\Gamma\colon \mathcal{A}^{\mathbb{Z}} \longrightarrow \mathcal{B}^{\mathbb{Z}}$ be the block map with local rule $\gamma\colon \mathcal{A}^{\mathbb{G}} \longrightarrow \mathcal{B}$ defined

$$\gamma(a_0, a_1) := 1 - a_0 - a_1 = \begin{cases} 1 & \text{if } [a_0, a_1] = [0, 0] = \mathbf{d}_0\,; \\ -1 & \text{if } [a_0, a_1] = [1, 1] = \mathbf{d}_1\,; \\ 0 & \text{otherwise}\,. \end{cases}$$

Then $\Gamma \circ \Phi = \Psi \circ \Gamma$; in other words, the BAM is a factor of ECA#184, and tracks the motion of the dislocations.
(b) Again, let $\Psi \in \mathsf{CA}(\mathcal{B}^{\mathbb{Z}})$ be the BAM. Let $\mathcal{A} = \mathbb{Z}_{/3}$, and let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ be the 3-color CCA. Let $\mathbb{G} := \{0, 1\}$, and let $\Gamma\colon \mathcal{A}^{\mathbb{Z}} \longrightarrow \mathcal{B}^{\mathbb{Z}}$ be the block map with local rule $\gamma\colon \mathcal{A}^{\mathbb{G}} \longrightarrow \mathcal{B}$ defined

$$\gamma(a_0, a_1) := (a_0 - a_1) \bmod 3\,.$$

Then $\Gamma \circ \Phi = \Psi \circ \Gamma$; in other words, the BAM is a factor of CCA_3, and tracks the motion of the domain boundaries.

Thus, it is often possible to translate questions about coalescent domain CA into questions about particle CA, which are generally easier to study. For example, the invariant measures of the BAM have been completely characterized.

Proposition 35 *Let $\mathcal{B} = \{\pm 1, 0\}$, and let $\Psi\colon \mathcal{B}^{\mathbb{Z}} \longrightarrow \mathcal{B}^{\mathbb{Z}}$ be the BAM.*

(a) *The sets $\mathbf{R} := \{0, 1\}^{\mathbb{Z}}$ and $\mathbf{L} := \{0, -1\}^{\mathbb{Z}}$ are Ψ-invariant, and Ψ acts as a right-shift on $\mathbf{R}$ and as a left-shift on $\mathbf{L}$.*

(b) *Let $\mathbf{L}^+ := \{0, -1\}^{\mathbb{N}}$ and $\mathbf{R}^- := \{0, 1\}^{-\mathbb{N}}$, and let*

$$\mathbf{X} := \big\{\mathbf{a} \in \mathcal{A}^{\mathbb{Z}}\,;\ \exists\, z \in \mathbb{Z} \text{ such that } \mathbf{a}_{(-\infty \ldots z]} \in \mathbf{R}^- \text{ and } \mathbf{a}_{[z \ldots \infty)} \in \mathbf{L}^+\big\}\,.$$

Then $\mathbf{X}$ is Ψ-invariant. For any $\mathbf{x} \in \mathbf{X}$, Ψ acts as a right shift on $\mathbf{a}_{(-\infty \ldots z)}$, and as a left-shift on $\mathbf{x}_{(z \ldots \infty)}$. (The boundary point z executes some kind of random walk.)

(c) *Any Ψ-invariant measure on $\mathcal{A}^{\mathbb{Z}}$ can be written in a unique way as a convex combination of four measures δ_0, ρ, λ, and μ, where: δ_0 is the point mass on the 'vacuum' configuration $[\ldots 0\,0\,0 \ldots]$, ρ is any shift-invariant measure on $\mathbf{R}$, λ is any shift-invariant measure on $\mathbf{L}$, and μ is a measure on $\mathbf{X}$.*
Furthermore, there exist shift-invariant measures μ_- and μ_+ on $\mathbf{R}^-$ and $\mathbf{L}^+$, respectively, such that, for μ-almost all $\mathbf{x} \in \mathbf{X}$, $\mathbf{x}_{(-\infty \ldots z]}$ is μ_--distributed and $\mathbf{x}_{[z \ldots \infty)}$ is μ_+-distributed.

Proof **(a)** and **(b)** are obvious; **(c)** is Theorem 1 in [8]. □

Remark 36 (a) Proposition 35(c) can be immediately translated into a complete characterization of the invariant measures of ECA#184, via the factor map Γ in Example 34(a); see [8], Theorem 3. Likewise, using the factor map in Example 34(b) we get a complete characterization of the invariant measures for CCA_3.

(b) Proposition 48 and Corollaries 49 and 50 describe the limit measures of the BAM, CCA_3, and ECA#184. Also, Blank [10] has characterized invariant measures for a broad class of multilane, multi-speed traffic models (including ECA#184); see Remark 51(b).

(c) Kůrka [78] has defined, for any $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$, a construction similar to the set $\mathbf{X}$ in Proposition 35(b). For any $n \in \mathbb{N}$ and $z \in \mathbb{Z}$, let $\mathbf{S}_{z,n}$ be the set of fixed points of $\Phi^n \circ \sigma^z$; then $\mathbf{S}_{z,n}$ is a subshift of finite type, which Kůrka calls a *signal subshift* with *velocity* $v = z/n$. (For example, if Φ is the BAM, then $\mathbf{R} = \mathbf{S}_{1,1}$ and $\mathbf{L} = \mathbf{S}_{-1,1}$.)

Now, suppose that $z_1/n_1 > z_2/n_2 > \cdots > z_J/n_J$. The *join* of the signal subshifts $\mathbf{S}_{z_1,n_1}, \mathbf{S}_{z_2,n_2}, \ldots, \mathbf{S}_{z_J,n_J}$ is the set $\mathbf{S}$ of all infinite sequences $[\mathbf{a}_1\ \mathbf{a}_2\ \ldots\ \mathbf{a}_J]$, where for all $j \in [1..J]$, $\mathbf{a}_j$ is a (possibly empty) finite word or (half-)infinite sequence admissible to the subshift $\mathbf{S}_{z_j,n_j}$. (For example, if $\mathbf{S}$ is the join of $\mathbf{S}_{1,1} = \mathbf{R}$ and $\mathbf{S}_{-1,1} = \mathbf{L}$ from Proposition 35(a), then $\mathbf{S} = \mathbf{L} \cup \mathbf{X} \cup \mathbf{R}$.) It follows that $\mathbf{S} \subseteq \Phi(\mathbf{S}) \subseteq \Phi^2(\mathbf{S}) \subseteq \cdots$. If we define $\Phi^\infty(\mathbf{S}) := \bigcup_{t=0}^\infty \Phi^t(\mathbf{S})$, then $\Phi^\infty(\mathbf{S}) \subseteq \Phi^\infty(\mathcal{A}^\mathbb{Z})$, where $\Phi^\infty(\mathcal{A}^\mathbb{Z}) := \bigcap_{t=0}^\infty \Phi^t(\mathcal{A}^\mathbb{Z})$ is the omega limit set of Φ (see Proposition 5 in [78]). The support of any Φ-invariant measure must be contained in $\Phi^\infty(\mathcal{A}^\mathbb{Z})$, so invariant measures may be closely related to the joins of signal subshifts. See ▶ Topological Dynamics of Cellular Automata for more information.

In the case of the BAM, it is not hard to check that $\Phi^\infty(\mathbf{S}) = \mathbf{S} = \Phi^\infty(\mathcal{A}^\mathbb{Z})$; this suggests an alternate proof of Proposition 35(c). It would be interesting to know whether a conclusion analogous to Proposition 35(c) holds for other $\Phi \in \mathsf{CA}(\mathcal{A}^\mathbb{Z})$ such that $\Phi^\infty(\mathcal{A}^\mathbb{Z})$ is a join of signal subshifts.

Limit Measures and Other Asymptotics

Asymptotic Randomization by Linear Cellular Automata

The results of Subsect. "Measure Rigidity in Algebraic CA" suggest that the uniform Bernoulli measure η is the 'natural' measure for algebraic CA, because η is the unique invariant measure satisfying any one of several collections of reasonable criteria. In this section, we will see that η is 'natural' in quite another way: it is the unique limit measure for linear CA from a large set of initial conditions.

If $\{\mu_n\}_{n=1}^\infty$ is a sequence of measures on $\mathcal{A}^\mathbb{M}$, then this sequence *weak* converges* to the measure μ_∞ ("$wk*\lim_{n\to\infty} \mu_n = \mu_\infty$") if, for all cylinder sets $\mathbf{B} \subset \mathcal{A}^\mathbb{M}$, $\lim_{n\to\infty} \mu_n[\mathbf{B}] = \mu_\infty[\mathbf{B}]$. Equivalently, for all continuous functions $f\colon \mathcal{A}^\mathbb{M} \longrightarrow \mathbb{C}$, we have

$$\lim_{n\to\infty} \int_{\mathcal{A}^\mathbb{M}} f \, d\mu_n = \int_{\mathcal{A}^\mathbb{M}} f \, d\mu_\infty .$$

The *Cesàro average* (or *Cesàro limit*) of $\{\mu_n\}_{n=1}^\infty$ is $wk*\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^N \mu_n$, if this limit exists.

Let $\mu \in \mathfrak{M}_{\mathrm{ces}}(\mathcal{A}^\mathbb{M})$ and let $\Phi \in \mathsf{CA}(\mathcal{A}^\mathbb{M})$. For any $t \in \mathbb{N}$, the measure $\Phi^t\mu$ is defined by $\Phi^t\mu(\mathbf{B}) = \mu(\phi^{-t}(\mathbf{B}))$, for any measurable subset $\mathbf{B} \subset \mathcal{A}^\mathbb{M}$. We say that Φ *asymptotically randomizes* μ if the Cesàro average of the sequence $\{\phi^n\mu\}_{n=1}^\infty$ is η. Equivalently, there is a subset $\mathbb{J} \subset \mathbb{N}$ of density 1, such that

$$wk*\lim_{\substack{j\to\infty \\ j\in\mathbb{J}}} \Phi^j\mu = \eta .$$

The uniform measure η is the measure of maximal entropy on $\mathcal{A}^\mathbb{M}$. Thus, asymptotic randomization is kind of 'Second Law of Thermodynamics' for CA.

Let $(\mathcal{A}, +)$ be a finite abelian group, and let Φ be a linear cellular automaton (LCA) on $\mathcal{A}^\mathbb{M}$. Recall that Φ has *scalar coefficients* if there is some finite $\mathbb{H} \subset \mathbb{M}$, and integer coefficients $\{c_\mathsf{h}\}_{\mathsf{h}\in\mathbb{H}}$ so that Φ has a local rule of the form

$$\phi(\mathbf{a}_\mathbb{H}) := \sum_{\mathsf{h}\in\mathbb{H}} c_\mathsf{h} a_\mathsf{h} , \tag{3}$$

An LCA Φ is *proper* if Φ has scalar coefficients as in Eq. (3), and if, furthermore, for any prime divisor p of $|\mathcal{A}|$, there are at least two $\mathsf{h}, \mathsf{h}' \in \mathbb{H}$ such that $c_\mathsf{h} \not\equiv 0 \not\equiv c_{\mathsf{h}'} \bmod p$. For example, if $\mathcal{A} = \mathbb{Z}_{/n}$ for some $n \in \mathbb{N}$, then every LCA on $\mathcal{A}^\mathbb{M}$ has scalar coefficients; in this case, Φ is proper if, for every prime p dividing n, at least two of these coefficients are coprime to p. In particular, if $\mathcal{A} = \mathbb{Z}_{/p}$ for some prime p, then Φ is proper as long as $|\mathbb{H}| \geq 2$.

Let $\mathsf{PLCA}(\mathcal{A}^\mathbb{M})$ be the set of proper linear CA on $\mathcal{A}^\mathbb{M}$. If $\mu \in \mathfrak{M}_{\mathrm{ces}}(\mathcal{A}^\mathbb{M})$, recall that μ has *full support* if $\mu[\mathbf{B}] > 0$ for every cylinder set $\mathbf{B} \subset \mathcal{A}^\mathbb{M}$.

Theorem 37 *Let $(\mathcal{A}, +)$ be a finite abelian group, let $\mathbb{M} := \mathbb{Z}^D \times \mathbb{N}^E$ for some $D, E \geq 0$, and let $\Phi \in \mathsf{PLCA}(\mathcal{A}^\mathbb{M})$. Let μ be any Bernoulli measure or Markov random field on $\mathcal{A}^\mathbb{M}$ having full support. Then Φ asymptotically randomizes μ.*

History Theorem 37 was first proved for simple one-dimensional LCA randomizing Bernoulli measures on $\mathcal{A}^\mathbb{Z}$, where $\mathcal{A}$ was a cyclic group. In the case $\mathcal{A} = \mathbb{Z}_{/2}$, Theorem 37 was independently proved for the *nearest-neighbor XOR* CA (having local rule $\phi(a_{-1}, a_0, a_1) = a_{-1} + a_1 \bmod 2$) by Miyamoto [100] and Lind [82]. This result was then generalized to $\mathcal{A} = \mathbb{Z}_{/p}$ for any prime p by Cai and Luo [16]. Next, Maass and Martínez [87] extended the Miyamoto/Lind result to the *binary Ledrappier* CA (local rule $\phi(a_0, a_1) = a_0 + a_1 \bmod 2$). Soon after, Ferrari et al. [36] considered the case when $\mathcal{A}$ was an abelian group of order p^k (p prime), and proved Theorem 37 for any *Ledrappier* CA (local rule $\phi(a_0, a_1) = c_0 a_0 + c_1 a_1$, where $c_0, c_1 \not\equiv 0 \bmod p$) acting on any measure on $\mathcal{A}^\mathbb{Z}$ having full support and 'rapidly decaying correlations' (see Part II(a) below). For example, this includes any Markov measure on $\mathcal{A}^\mathbb{Z}$ with full support. Next, Pivato and Yassawi [116] generalized Theorem 37 to *any* PLCA acting on any fully supported N-step Markov chain on $\mathcal{A}^\mathbb{Z}$

or any nontrivial Bernoulli measure on $\mathcal{A}^{\mathbb{Z}^D \times \mathbb{N}^E}$, where $\mathcal{A} = \mathbb{Z}_{/p^k}$ (p prime). Finally, Pivato and Yassawi [117] proved Theorem 37 in full generality, as stated above.

The proofs of Theorem 37 and its variations all involve two parts:

Part I. A careful analysis of the local rule of Φ^t (for all $t \in \mathbb{N}$), showing that the neighborhood of Φ^t grows large as $t \to \infty$ (and in some cases, contains large 'gaps').

Part II. A demonstration that the measure μ exhibits 'rapidly decaying correlations' between widely separated elements of $\mathbb{M}$; hence, when these elements are combined using Φ^t, it is as if we are summing independent random variables.

Part I Any linear CA with scalar coefficients can be written as a 'Laurent polynomial of shifts'. That is, if Φ has local rule (3), then for any $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$,

$$\Phi(\mathbf{a}) := \sum_{\mathsf{h} \in \mathbb{H}} c_{\mathsf{h}} \sigma^{\mathsf{h}}(\mathbf{a}) \quad \text{(where we add configurations componentwise)}.$$

We indicate this by writing "$\Phi = F(\sigma)$", where $F \in \mathbb{Z}[x_1^{\pm 1}, x_2^{\pm 1}, \ldots, x_D^{\pm 1}]$ is the D-variable Laurent polynomial defined:

$$F(x_1, \ldots, x_D) := \sum_{(h_1, \ldots, h_D) \in \mathbb{H}} c_{\mathsf{h}} x_1^{h_1} x_2^{h_2} \ldots x_D^{h_D}.$$

For example, if Φ is the *nearest-neighbor XOR* CA, then $\Phi = \sigma^{-1} + \sigma^1 = F(\sigma)$, where $F(x) = x^{-1} + x$. If Φ is a *Ledrappier* CA, then $\Phi = c_0 \mathbf{Id} + c_1 \sigma^1 = F(\sigma)$, where $F(x) = c_0 + c_1 x$.

It is easy to verify that, if F and G are two such polynomials, and $\Phi = F(\sigma)$ while $\Gamma = G(\sigma)$, then $\Phi \circ \Gamma = (F \cdot G)(\sigma)$, where $F \cdot G$ is the product of F and G in the polynomial ring $\mathbb{Z}[x_1^{\pm 1}, x_2^{\pm 1}, \ldots, x_D^{\pm 1}]$. In particular, this means that $\Phi^t = F^t(\sigma)$ for all $t \in \mathbb{N}$. Thus, iterating an LCA is equivalent to computing the powers of a polynomial.

If $\mathcal{A} = \mathbb{Z}_{/p}$, then we can compute the coefficients of F^t modulo p. If p is prime, then this can be done using a result of Lucas [84], which provides a formula for the binomial coefficient $\binom{a}{b}$ in terms of the base-p expansions of a and b. For example, if $p = 2$, then Lucas' theorem says that Pascal's triangle, modulo 2, looks like a 'discrete Sierpinski triangle', made out of 0's and 1's. (This is why fragments of the Sierpinski triangle appear frequently in the spacetime diagrams of linear CA on $\mathcal{A} = \mathbb{Z}_{/2}$, a phenomenon which has inspired much literature on 'fractals and automatic sequences in cellular automata'; see [2,4,5,6,7,52,53,54,55,56,57,94,138,139,140,145,146,147,148,149].) Thus, Lucas' Theorem, along with some combinatorial lemmas about the structure of base-p expansions, provides the machinery for Part I.

Part II There are two approaches to analyzing probability measures on $\mathcal{A}^{\mathbb{M}}$; one using renewal theory, and the other using harmonic analysis.

II(a) Renewal theory This approach was developed by Maass, Martínez and their collaborators. Loosely speaking, if $\mu \in \mathfrak{M}_{\mathrm{ens}}(\mathcal{A}^{\mathbb{Z}}, \sigma)$ has sufficiently large support and sufficiently rapid decay of correlations (e. g. a Markov chain), and $\mathbf{a} \in \mathcal{A}^{\mathbb{Z}}$ is a μ-random sequence, then we can treat $\mathbf{a}$ as if there is a sparse, randomly distributed set of 'renewal times' when the normal stochastic evolution of $\mathbf{a}$ is interrupted by independent, random 'errors'. By judicious use of Part I described above, one can use this 'renewal process' to make it seem as though Φ^t is summing independent random variables.

For example, if $(\mathcal{A}, +)$ be an abelian group of order p^k where p is prime, and $\mu \in \mathfrak{M}_{\mathrm{ens}}(\mathcal{A}^{\mathbb{Z}}; \sigma)$ has *complete connections* (see Example 22(a)) and *summable decay* (which means that a certain sequence of coefficients (measuring long-range correlation) decays fast enough that its sum is finite), and $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ is a Ledrappier CA, then Ferrari et al. (see Theorem 1.3 in [36]) showed that Φ asymptotically randomizes μ. (For example, this applies to any N-step Markov chain with full support on on $\mathcal{A}^{\mathbb{Z}}$.) Furthermore, if $\mathcal{A} = \mathbb{Z}_{/p} \times \mathbb{Z}_{/p}$, and $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ has linear local rule $\phi\left(\left[\begin{smallmatrix} x_0 \\ y_0 \end{smallmatrix}\right], \left[\begin{smallmatrix} x_1 \\ y_1 \end{smallmatrix}\right]\right) = (y_0, x_0 + y_1)$, then Maass and Martínez [88] showed that Φ randomizes any Markov measure with full support on $\mathcal{A}^{\mathbb{Z}}$. Maass and Martínez again handled Part II using renewal theory. However, in this case, Part I involves some delicate analysis of the (noncommutative) algebra of the matrix-valued coefficients; unfortunately, their argument does not generalize to other LCA with noncommuting, matrix-valued coefficients. (However, Proposition 8 of [117] suggests a general strategy for dealing with such LCA).

II(b) Harmonic analysis This approach to Part II was implicit in the early work of Lind [82] and Cai and Luo [16], but was developed in full generality by Pivato and Yassawi [116,117,118]. We regard $\mathcal{A}^{\mathbb{M}}$ as a direct product of copies of the group $(\mathcal{A}, +)$, and endow it with the product group structure; then $(\mathcal{A}^{\mathbb{M}}, +)$ a compact abelian topological group. A *character* on $(\mathcal{A}^{\mathbb{M}}, +)$ is a continuous group homomorphism $\chi : \mathcal{A}^{\mathbb{M}} \longrightarrow \mathbb{T}$, where $\mathbb{T} := \{c \in \mathbb{C}; |c| = 1\}$ is the unit circle group. If μ is

a measure on $\mathcal{A}^{\mathbb{M}}$, then the *Fourier coefficients* of μ are defined: $\hat{\mu}[\boldsymbol{\chi}] = \int_{\mathcal{A}^{\mathbb{M}}} \boldsymbol{\chi}\, d\mu$, for every character $\boldsymbol{\chi}$.

If $\boldsymbol{\chi} : \mathcal{A}^{\mathbb{M}} \longrightarrow \mathbb{T}$ is any character, then there is a unique finite subset $\mathbb{K} \subset \mathbb{M}$ (called the *support* of $\boldsymbol{\chi}$) and a unique collection of nontrivial characters $\chi_{\mathsf{k}} : \mathcal{A} \longrightarrow \mathbb{T}$ for all $\mathsf{k} \in \mathbb{K}$, such that,

$$\boldsymbol{\chi}(\mathbf{a}) = \prod_{\mathsf{k}\in\mathbb{K}} \chi_{\mathsf{k}}(a_{\mathsf{k}}) , \quad \forall\, \mathbf{a} \in \mathcal{A}^{\mathbb{M}} . \tag{4}$$

We define $\mathsf{rank}[\boldsymbol{\chi}] := |\mathbb{K}|$. The measure μ is called *harmonically mixing* if, for all $\epsilon > 0$, there is some R such that for all characters $\boldsymbol{\chi}$, $(\mathsf{rank}[\boldsymbol{\chi}] \geq R) \implies (|\hat{\mu}[\boldsymbol{\chi}]| < \epsilon)$.

The set $\mathfrak{Sm}(\mathcal{A}^{\mathbb{M}})$ of harmonically mixing measures on $\mathcal{A}^{\mathbb{M}}$ is quite inclusive. For example, if μ is any (N-step) Markov chain with full support on $\mathcal{A}^{\mathbb{Z}}$, then $\mu \in \mathfrak{Sm}(\mathcal{A}^{\mathbb{Z}})$ (see Propositions 8 and 10 in [116]), and if $\nu \in \mathfrak{M}_{\mathrm{ens}}(\mathcal{A}^{\mathbb{Z}})$ is absolutely continuous with respect to this μ, then $\nu \in \mathfrak{Sm}(\mathcal{A}^{\mathbb{Z}})$ also (see Corollary 9 in [116]). If $\mathcal{A} = \mathbb{Z}_{/p}$ (p prime) then any nontrivial Bernoulli measure on $\mathcal{A}^{\mathbb{M}}$ is harmonically mixing (see Proposition 6 in [116]). Furthermore, if $\mu \in \mathfrak{M}_{\mathrm{ens}}(\mathcal{A}^{\mathbb{Z}}; \sigma)$ has complete connections and summable decay, then $\mu \in \mathfrak{Sm}(\mathcal{A}^{\mathbb{Z}})$ (see Theorem 23 in [61]). If $\mathfrak{M} := \mathfrak{M}_{\mathrm{ens}}(\mathcal{A}^{\mathbb{M}}; \mathbb{C})$ is the set of all complex-valued measures on $\mathcal{A}^{\mathbb{M}}$, then $\mathfrak{M}$ is *Banach algebra* (i. e. it is a vector space under the obvious definition of addition and scalar multiplication for measures, and a Banach space under the total variation norm, and finally, since $\mathcal{A}^{\mathbb{M}}$ is a topological group, $\mathfrak{M}$ is a ring under convolution). Then $\mathfrak{Sm}(\mathcal{A}^{\mathbb{M}})$ is an ideal in $\mathfrak{M}$, is closed under the total variation norm, and is dense in the weak* topology on $\mathfrak{M}$ (see Propositions 4 and 7 in [116]).

Finally, if μ is any Markov random field on $\mathcal{A}^{\mathbb{M}}$ which is *locally free* (which roughly means that the boundary of any finite region does not totally determine the interior of that region), then $\mu \in \mathfrak{Sm}(\mathcal{A}^{\mathbb{M}})$ (see Theorem 1.3 in [118]). In particular, this implies:

Proposition 38 *If $(\mathcal{A}, +)$ is any finite group, and $\mu \in \mathfrak{M}_{\mathrm{ens}}(\mathcal{A}^{\mathbb{M}})$ is any Markov random field with full support, then μ is harmonically mixing.*

Proof This follows from Theorem 1.3 in [118]. It is also a special case of Theorem 15 in [117]. □

If $\boldsymbol{\chi}$ is a character, and Φ is a LCA, then $\boldsymbol{\chi} \circ \Phi^t$ is also a character, for any $t \in \mathbb{N}$ (because it is a composition of two continuous group homomorphisms). We say Φ is *diffusive* if there is a subset $\mathbb{J} \subset \mathbb{N}$ of density 1, such that, for every character $\boldsymbol{\chi}$ of $\mathcal{A}^{\mathbb{M}}$,

$$\lim_{\mathbb{J}\ni j\to\infty} \mathsf{rank}[\boldsymbol{\chi} \circ \Phi^j] = \infty .$$

Proposition 39 *Let $(\mathcal{A}, +)$ be any finite abelian group and let $\mathbb{M}$ be any monoid. If μ is harmonically mixing and Φ is diffusive, then Φ asymptotically randomizes μ.*

Proof See Theorem 12 in [117]. □

Proposition 40 *Let $(\mathcal{A}, +)$ be any abelian group and let $\mathbb{M} := \mathbb{Z}^D \times \mathbb{N}^E$ for some $D, E \geq 0$. If $\Phi \in \mathit{PLCA}(\mathcal{A}^{\mathbb{M}})$, then Φ is diffusive.*

Proof The proof uses Lucas' theorem, as described in Part I above. See Theorem 15 in [116] for the case $\mathcal{A} = \mathbb{Z}_{/p}$ when p prime. See Theorem 6 in [117] for the case when $\mathcal{A}$ is any cyclic group. That proof easily extends to any finite abelian group $\mathcal{A}$: write $\mathcal{A}$ as a product of cyclic groups and decompose Φ into separate automata over these cyclic factors. □

Proof of Theorem 37 Combine Propositions 38, 39, and 40.

□

Remark (a) Proposition 40 can be generalized: we do not need the coefficients of Φ to be integers, but merely to be a collection of automorphisms of $\mathcal{A}$ which commute with one another (so that Lucas' theorem from Part I is still applicable). See Theorem 9 in [117].

(b) For simplicity, we stated Theorem 37 for measures with full support; however, Proposition 39 actually applies to many Markov random fields *without* full support, because harmonic mixing only requires 'local freedom' (see Theorem 1.3 in [118]). For example, the support of a Markov chain on $\mathcal{A}^{\mathbb{Z}}$ is Markov subshift. If $\mathcal{A} = \mathbb{Z}_{/p}$ (p prime), then Proposition 39 yields asymptotic randomization of the Markov chain as long as the transition digraph of the underlying Markov subshift admits at least *two* distinct paths of length 2 between any pair of vertices in $\mathcal{A}$. More generally, if $\mathbb{M} = \mathbb{Z}^D$, then the support of any Markov random field on $\mathcal{A}^{\mathbb{Z}^D}$ is an SFT, which we can regard as the set of all tilings of $\mathbb{R}^D$ by a certain collection of Wang tiles. If $\mathcal{A} = \mathbb{Z}_{/p}$ (p prime), then Proposition 39 yields asymptotic randomization of the Markov random field as long as the underlying Wang tiling is flexible enough that any hole can always be filled in at least two ways; see Sect. 1 in [118].

Remark 41 (Generalizations and Extensions) (a) Pivato and Yassawi (see Thm 3.1 in [118]) proved a variation of Theorem 60 where diffusion (of Φ) is replaced with a slightly stronger condition called *dispersion*, so that harmonic mixing (of μ) can be replaced with a slightly weaker condition called *dispursion mixing* (DM). It is unknown whether all proper linear CA are dispersive, but a very large class are (including, for example, $\Phi = \mathbf{Id} + \sigma$).

Any uniformly mixing measure with positive entropy is DM (see Theorem 5.2 in [118]); this includes, for example, any mixing *quasimarkov measure* (i. e. the image of a Markov measure under a block map; these are the natural measures supported on sofic shifts). Quasimarkov measures are not, in general, harmonically mixing (see Sect. 2 in [118]), but this result shows they are still asymptotically randomized by most linear CA.

(b) Suppose $\mathbf{G} \subset \mathcal{A}^{\mathbb{Z}^D}$ is a σ-transitive *subgroup shift* (see Subsect. "Measure Rigidity in Algebraic CA" for definition), and let $\Phi \in \mathsf{PLCA}(\mathbf{G})$. If $\mathbf{G}$ satisfies an algebraic condition called the *follower lifting property* (FLP) and μ is any Markov random field with $\mathsf{supp}(\mu) = \mathbf{G}$, then Maass, Martínez, Pivato, and Yassawi [89] have shown that Φ asymptotically randomizes μ to a maxentropy measure on $\mathbf{G}$. Furthermore, if $D = 1$, then this maxentropy measure is the Haar measure on $\mathbf{G}$. In particular, if $\mathcal{A}$ is an abelian group of prime-power order, then *any* transitive Markov subgroup $\mathbf{G} \subset \mathcal{A}^{\mathbb{Z}}$ satisfies the FLP, so this result holds for any multistep Markov measure on $\mathbf{G}$. See also [90] for the special case when $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ has local rule $\phi(x_0, x_1) = x_0 + x_1$. In the special case when Φ has local rule $\phi(x_0, x_1) = c_0 x_0 + c_1 x_1 + a$, the result has been extended to measures with complete connections and summable decay; see Teorema III.2.1, p. 71 in [134] or see Theorem 1 in [91].

(c) All the aforementioned results concern asymptotic randomization of initial measures with nonzero entropy. Is nonzero entropy either necessary or sufficient for asymptotic randomization? First let $\mathbf{X}_N \subset \mathcal{A}^{\mathbb{Z}}$ be the set of N-periodic points (see Subsect. "Periodic Invariant Measures") and suppose $\mathsf{supp}(\mu) \subseteq \mathbf{X}_N$. Then the Cesàro limit of $\{\Phi^t(\mu)\}_{t\in\mathbb{N}}$ will also be a measure supported on $\mathbf{X}_N$, so μ_∞ cannot be the uniform measure on $\mathcal{A}^{\mathbb{Z}}$. Nor, in general, will μ_∞ be the uniform measure on $\mathbf{X}_N$; this follows from Jen's (1988) exact characterization of the limit cycles of linear CA acting on $\mathbf{X}_N$.

What if μ is a *quasiperiodic* measure, such as the unique σ-invariant measure on a Sturmian shift? There exist quasiperiodic measures on $(\mathbb{Z}_{/2})^{\mathbb{Z}}$ which are *not* asymptotically randomized by the Ledrappier CA (see Sect. 15 in [112]). But it is unknown whether this extends to all quasiperiodic measures or all linear CA.

There is also a measure μ on $\mathcal{A}^{\mathbb{Z}}$ which has zero σ-entropy, yet is still asymptotically randomized by Φ (see Sect. 8 in [118]). Loosely speaking, μ is a Toeplitz measure with a very low density of 'bit errors'. Thus, μ is 'almost' deterministic (so it has zero entropy), but by sufficiently increasing the density of 'bit errors', we can introduce just enough randomness to allow asymptotic randomization to occur.

(d) Suppose $(\mathcal{G}, \cdot)$ is a *nonabelian* group and $\Phi\colon \mathcal{G}^{\mathbb{Z}} \longrightarrow \mathcal{G}^{\mathbb{Z}}$ has *multiplicative* local rule $\phi(\mathbf{g}) := g_{\mathsf{h}_1}^{n_1} g_{\mathsf{h}_2}^{n_2} \cdots g_{\mathsf{h}_J}^{n_J}$, for some $\{\mathsf{h}_1, \ldots, \mathsf{h}_J\} \subset \mathbb{Z}$ (possibly not distinct) and $n_1, \ldots, n_J \in \mathbb{N}$. If $\mathcal{G}$ is nilpotent, then $\mathcal{G}$ can be decomposed into a tower of abelian group extensions; this induces a structural decomposition of Φ into a tower of skew products of 'relative' linear CA. This strategy was first suggested by Moore [102], and was developed by Pivato (see Theorem 21 in [111]), who proved a version of Theorem 37 in this setting.

(e) Suppose $(\mathcal{Q}, \star)$ is a *quasigroup* – that is, $\bigstar$ is a binary operation such that for any $q, r, s \in \mathcal{Q}$, $(q \star r = q \star s) \iff (r = s) \iff (r \star q = s \star q)$. Any finite *associative* quasigroup has an identity, and any associative quasigroup with an identity is a group. However there are also many nonassociative finite quasigroups. If we define a 'multiplicative' CA $\Phi\colon \mathcal{Q}^{\mathbb{Z}} \longrightarrow \mathcal{Q}^{\mathbb{Z}}$ with local rule $\phi\colon \mathcal{Q}^{\{0,1\}} \longrightarrow \mathcal{Q}$ given by $\phi(q_0, q_1) = q_0 \star q_1$, then it is easy to see that Φ is bipermutative if and only if $(\mathcal{Q}, \star)$ is a quasigroup. Thus, quasigroups seem to provide the natural algebraic framework for studying bipermutative CA; this was first proposed by Moore [101], and later explored by Host, Maass, and Martínez (see Sect. 3 in [61]), Pivato (see Sect. 2 in [113]), and Sobottka [134,135,136].

Note that $\mathcal{Q}^{\mathbb{Z}}$ is a quasigroup under componentwise $\bigstar$-multiplication. A *quasigroup shift* is a subshift $\mathbf{X} \subset \mathcal{Q}^{\mathbb{Z}}$ which is also a subquasigroup; it follows that $\Phi(\mathbf{X}) \subseteq \mathbf{X}$. If $\mathbf{X}$ and Φ satisfy certain strong algebraic conditions, and $\mu \in \mathfrak{M}_{\mathsf{cns}}(\mathbf{X}; \sigma)$ has complete connections and summable decay, then the sequence $\{\Phi^t \mu\}_{t=1}^{\infty}$ Cesàro-converges to a maxentropy measure μ_∞ on $\mathbf{X}$ (thus, if $\mathbf{X}$ is irreducible, then μ_∞ is the Parry measure; see Subsect. "Invariance of Maxentropy Measures"). See Theorem 6.3(i) in [136], or Teorema IV.5.3, p. 107 in [134].

Hybrid Modes of Self-Organization

Most cellular automata do not asymptotically randomize; instead they seem to weak* converge to limit measures concentrated on small (i. e. low-entropy) subsets of the statespace $\mathcal{A}^{\mathbb{M}}$ – a phenomenon which can be interpreted as a form of 'self-organization'. Exact limit measures have been computed for a few CA. For example, let $\mathcal{A} = \{0, 1, 2\}$ and let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}^D})$ be the *Greenberg–Hastings* model (a simple model of an excitable medium). Durrett and Steif [26] showed that, if $D \geq 2$ and μ is any Bernoulli measure on $\mathcal{A}^{\mathbb{Z}^D}$, then $\mu_\infty := wk * \lim_{t\to\infty} \Phi^t \mu$ exists; μ_∞-almost all points are 3-periodic for Φ, and although μ_∞ is not a Bernoulli measure, the system $(\mathcal{A}^{\mathbb{Z}^D}, \mu_\infty, \sigma)$ is measurably isomorphic to a Bernoulli system.

In other cases, the limit measure cannot be exactly computed, but can still be estimated. For example, let $\mathcal{A} = \{\pm 1\}$, $\theta \in (0,1)$, and $R > 0$, and let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ be the (R,θ)-*threshold voter CA* (where each cell computes the fraction of its radius-R neighbors which disagree with its current sign, and negates its sign if this fraction is at least θ). Durret and Steif [27] and Fisch and Gravner [43] have described the long-term behavior of Φ in the limit as $R \to \infty$. If $\theta < 1/2$, then every initial condition falls into a two-periodic orbit (and if $\theta < 1/4$, then every cell simply alternates its sign). Let η be the uniform Bernoulli measure on $\mathcal{A}^{\mathbb{Z}}$; if $1/2 < \theta$, then for any finite subset $\mathbb{B} \subset \mathbb{Z}$, if R is large enough, then 'most' initial conditions (relative to η) converge to orbits that are fixed inside $\mathbb{B}$. Indeed, there is a critical value $\theta_c \approx 0.6469076$ such that, if $\theta_c < \theta$, and R is large enough, then 'most' initial conditions (for η) are already fixed inside $\mathbb{B}$; see also [137] for an analysis of behavior at the critical value.

In still other cases, the limit measure is is known to exist, but is still mysterious; this is true for the Cesàro limit measures of *Coven* CA, for example see [87], Theorem 1. However, for most CA, it is difficult to even show that limit measures exist. Except for the linear CA of Subsect. "Asymptotic Randomization by Linear Cellular Automata", there is no large class of CA whose limit measures have been exactly characterized. Often, it is much easier to study the dynamical asymptotics of CA at a purely topological level.

If $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})$, then $\mathcal{A}^{\mathbb{M}} \supseteq \Phi(\mathcal{A}^{\mathbb{M}}) \supseteq \Phi^2(\mathcal{A}^{\mathbb{M}}) \supseteq \cdots$. The *limit set* of Φ is the nonempty subshift $\Phi^{\infty}(\mathcal{A}^{\mathbb{M}}) := \bigcap_{t=1}^{\infty} \Phi^t(\mathcal{A}^{\mathbb{M}})$. For any $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$, the *omega-limit set* of $\mathbf{a}$ is the set $\omega(\mathbf{a}, \Phi)$ of all cluster points of the Φ-orbit $\{\Phi^t(\mathbf{a})\}_{t=1}^{\infty}$. A closed subset $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$ is a (Conley) *attractor* if there exists a clopen subset $\mathbf{U} \supseteq \mathbf{X}$ such that $\Phi(\mathbf{U}) \subseteq \mathbf{U}$ and $\mathbf{X} = \bigcap_{t=1}^{\infty} \Phi^t(\mathbf{U})$. It follows that $\omega(\Phi, \mathbf{u}) \subseteq \mathbf{X}$ for all $\mathbf{u} \in \mathbf{U}$. For example, $\Phi^{\infty}(\mathcal{A}^{\mathbb{M}})$ is an attractor (let $\mathbf{U} := \mathcal{A}^{\mathbb{M}}$). The topological attractors of CA were analyzed by Hurley [63,65,66], who discovered severe constraints on the possible attractor structures a CA could exhibit; see Sect. 9 of ► Topological Dynamics of Cellular Automata and ► Cellular Automata, Classification of.

Within pure topological dynamics, attractors and (omega) limit sets are the natural formalizations of the heuristic notion of 'self-organization'. The corresponding formalization in pure ergodic theory is the weak* limit measure. However, both weak* limit measures and topological attractors fail to adequately describe the sort of self-organization exhibited by many CA. Thus, several 'hybrid' notions self-organization have been developed, which combine topological and measurable criteria. These hybrid notions are more flexible and inclusive than purely topological notions. However, they do not require the explicit computation (or even the existence) of weak* limit measures, so in practice they are much easier to verify than purely ergodic notions.

Milnor–Hurley μ-attractors If $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$ is a closed subset, then for any $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$, we define $d(\mathbf{a}, \mathbf{X}) := \inf_{\mathbf{x} \in \mathbf{X}} d(\mathbf{a}, \mathbf{x})$. If $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})$, then the *basin* (or *realm*) of $\mathbf{X}$ is the set

$$\begin{aligned}\mathsf{Basin}(\mathbf{X}) &:= \left\{\mathbf{a} \in \mathcal{A}^{\mathbb{M}}\,;\ \lim_{t\to\infty} d\left(\Phi^t(\mathbf{a}), \mathbf{X}\right) = 0\right\} \\ &= \left\{\mathbf{a} \in \mathcal{A}^{\mathbb{M}}\,;\ \omega(\mathbf{a}, \Phi) \subseteq \mathbf{X}\right\}.\end{aligned}$$

Suppose $\Phi(\mathbf{X}) \subseteq \mathbf{X}$. If $\mu \in \mathfrak{M}_{\mathsf{ens}}(\mathcal{A}^{\mathbb{M}})$, then $\mathbf{X}$ is a *μ-attractor* if $\mu[\mathsf{Basin}(\mathbf{X})] > 0$; we call $\mathbf{X}$ a *lean μ-attractor* if in addition, $\mu[\mathsf{Basin}(\mathbf{X})] > \mu[\mathsf{Basin}(\mathbf{Y})]$ for any proper closed subset $\mathbf{Y} \subsetneq \mathbf{X}$. Finally, a μ-attractor $\mathbf{X}$ is *minimal* if $\mu[\mathsf{Basin}(\mathbf{Y})] = 0$ for any proper closed subset $\mathbf{Y} \subsetneq \mathbf{X}$. For example, if $\mathbf{X}$ is a μ-attractor, and $(\mathbf{X}, \Phi)$ is minimal as a dynamical system, then $\mathbf{X}$ is a minimal μ-attractor. This concept was introduced by Milnor [96,97] in the context of smooth dynamical systems; its ramifications for CA were first explored by Hurley [64,65].

If $\mu \in \mathfrak{M}_{\mathsf{ens}}(\mathcal{A}^{\mathbb{Z}^D}, \sigma)$, then μ is *weakly σ-mixing* if, for any measurable sets $\mathbf{U}, \mathbf{V} \subset \mathcal{A}^{\mathbb{Z}^D}$, there is a subset $\mathbb{J} \subset \mathbb{Z}^D$ of density 1 such that $\lim_{\mathbb{J} \ni \mathsf{j} \to \infty} \mu[\sigma^{\mathsf{j}}(\mathbf{U}) \cap \mathbf{V}] = \mu[\mathbf{U}] \cdot \mu[\mathbf{V}]$ (see Subsect. "Mixing and Ergodicity"). For example, any Bernoulli measure is weakly mixing. A subshift $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^D}$ is *σ-minimal* if $\mathbf{X}$ contains no proper nonempty subshifts. For example, if $\mathbf{X}$ is just the σ-orbit of some σ-periodic point, then $\mathbf{X}$ is σ-minimal.

Proposition 42 *Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})$, let $\mu \in \mathfrak{M}_{\mathsf{ens}}(\mathcal{A}^{\mathbb{M}}, \sigma)$, and let $\mathbf{X}$ be a μ-attractor.*

(a) *If μ is σ-ergodic, and $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$ is a subshift, then $\mu[\mathsf{Basin}(\mathbf{X})] = 1$.*

(b) *If $\mathbb{M}$ is countable, and $\mathbf{X}$ is σ-minimal subshift with $\mu[\mathsf{Basin}(\mathbf{X})] = 1$, then $\mathbf{X}$ is lean.*

(c) *Suppose $\mathbb{M} = \mathbb{Z}^D$ and μ is weakly σ-mixing.*

(i) If $\mathbf{X}$ is a minimal μ-attractor, then $\mathbf{X}$ is a subshift, so $\mu[\mathsf{Basin}(\mathbf{X})] = 1$, and thus $\mathbf{X}$ is the only lean μ-attractor of Φ.

(ii) If $\mathbf{X}$ is a Φ-periodic orbit which is also a lean μ-attractor, then $\mathbf{X}$ is minimal, $\mu[\mathsf{Basin}(\mathbf{X})] = 1$, and $\mathbf{X}$ contains only constant configurations.

Proof **(a)** If $\mathbf{X}$ is σ-invariant, then $\mathsf{Basin}(\mathbf{X})$ is also σ-invariant; hence $\mu[\mathsf{Basin}(\mathbf{X})] = 1$ because μ is σ-ergodic.

(b) Suppose $\mathbf{Y} \subsetneq \mathbf{X}$ was a proper closed subset with $\mu[\mathsf{Basin}(\mathbf{Y})] = 1$. For any $\mathsf{m} \in \mathbb{M}$, it is easy

to check that $\mathsf{Basin}(\sigma^{\mathsf{m}}[\mathbf{Y}]) = \sigma^{\mathsf{m}}[\mathsf{Basin}(\mathbf{Y})]$. Thus, if $\widetilde{\mathbf{Y}} := \bigcap_{\mathsf{m}\in\mathbb{M}} \sigma^{\mathsf{m}}(\mathbf{Y})$, then $\mathsf{Basin}(\widetilde{\mathbf{Y}}) = \bigcap_{\mathsf{m}\in\mathbb{M}} \sigma^{\mathsf{m}}[\mathsf{Basin}(\mathbf{Y})]$, so $\mu[\mathsf{Basin}(\widetilde{\mathbf{Y}})] = 1$ (because $\mathbb{M}$ is countable). Thus, $\widetilde{\mathbf{Y}}$ is nonempty, and is a subshift of $\mathbf{X}$. But $\mathbf{X}$ is σ-minimal, so $\widetilde{\mathbf{Y}} = \mathbf{X}$, which means $\mathbf{Y} = \mathbf{X}$. Thus, $\mathbf{X}$ is a lean μ-attractor.

(c) In the case when μ is a Bernoulli measure, **(c)[i]** is Theorem B in [64] or Proposition 2.7 in [65], while **(c)[ii]** is Theorem A in [65]. Hurley's proofs easily extend to the case when μ is weakly σ-mixing. The only property we require of μ is this: for any nontrivial measurable sets $\mathbf{U}, \mathbf{V} \subset \mathcal{A}^{\mathbb{Z}^D}$, and any $\mathsf{z} \in \mathbb{Z}^D$, there is some $\mathsf{x}, \mathsf{y} \in \mathbb{Z}^D$ with $\mathsf{z} = \mathsf{x} - \mathsf{y}$, such that $\mu[\sigma^{\mathsf{y}}(\mathbf{U}) \cap \mathbf{V}] > 0$ and $\mu[\sigma^{\mathsf{x}}(\mathbf{U}) \cap \mathbf{V}] > 0$. This is clearly true if μ is weakly mixing (because if $\mathbb{J} \subset \mathbb{Z}^D$ has density 1, then $\mathbb{J} \cap (\mathsf{z} + \mathbb{J}) \neq \emptyset$ for any $\mathsf{z} \in \mathbb{Z}^D$).

Proof sketch for **(c)[i]** If $\mathbf{X}$ is a (minimal) μ-attractor, then so is $\sigma^{\mathsf{y}}(\mathbf{X})$, and $\mathsf{Basin}[\sigma\mathsf{y}(\mathbf{X})] = \sigma\mathsf{y}(\mathsf{Basin}[\mathbf{X}])$. Thus, weak mixing yields $\mathsf{x}, \mathsf{y} \in \mathbb{Z}^D$ such that $\mathsf{Basin}[\sigma\mathsf{x}(\mathbf{X})] \cap \mathsf{Basin}[\mathbf{X}]$ and $\mathsf{Basin}[\sigma\mathsf{y}(\mathbf{X})] \cap \mathsf{Basin}[\mathbf{X}]$ are both nontrivial. But the basins of distinct minimal μ-attractors must be disjoint; thus $\sigma^{\mathsf{x}}(\mathbf{X}) = \mathbf{X} = \sigma^{\mathsf{y}}(\mathbf{X})$. But $\mathsf{x} - \mathsf{y} = \mathsf{z}$, so this means $\sigma^{\mathsf{z}}(\mathbf{X}) = \mathbf{X}$. This holds for all $\mathsf{z} \in \mathbb{Z}^D$, so $\mathbf{X}$ is a subshift, so (a) implies $\mu[\mathsf{Basin}(\mathbf{X})] = 1$. □

Section 4 of [64] contains several examples showing that the minimal topological attractor of Φ can be different from its minimal μ-attractor. For example, a CA can have different minimal μ-attractors for different choices of μ. On the other hand, there is a CA possessing a minimal topological attractor but with no minimal μ-attractors for any Bernoulli measure μ.

Hilmy–Hurley Centers

Let $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$. For any closed subset $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$, we define

$$\mu_{\mathbf{a}}[\mathbf{X}] := \liminf_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}_{\mathbf{X}}(\Phi^t(\mathbf{a})) .$$

(Thus, if μ is a Φ-ergodic measure on $\mathcal{A}^{\mathbb{M}}$, then Birkhoff's Ergodic Theorem asserts that $\mu_{\mathbf{a}}[\mathbf{X}] = \mu[\mathbf{X}]$ for μ-almost all $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$). The *center* of $\mathbf{a}$ is the set:

$$\mathsf{Cent} < (\mathbf{a}, \Phi) := \bigcap \left\{ \text{closed subsets } \mathbf{X} \subseteq \mathcal{A}^{\mathbb{M}} \; ; \; \mu_{\mathbf{a}}[\mathbf{X}] = 1 \right\}.$$

Thus, $\mathsf{Cent}(\mathbf{a}, \Phi)$ is the smallest closed subset such that $\mu_{\mathbf{a}}[\mathsf{Cent}(\mathbf{a}, \Phi)] = 1$. If $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$ is closed, then the *well* of $\mathbf{X}$ is the set

$$\mathsf{Well}(\mathbf{X}) := \left\{ \mathbf{a} \in \mathcal{A}^{\mathbb{M}} \; ; \; \mathsf{Cent}(\mathbf{a}, \Phi) \subseteq \mathbf{X} \right\} .$$

If $\mu \in \mathfrak{M}_{\mathsf{cons}}(\mathcal{A}^{\mathbb{M}})$, then $\mathbf{X}$ is a *μ-center* if $\mu[\mathsf{Well}(\mathbf{X})] > 0$; we call $\mathbf{X}$ a *lean μ-center* if in addition, $\mu[\mathsf{Well}(\mathbf{X})] > \mu[\mathsf{Well}(\mathbf{Y})]$ for any proper closed subset $\mathbf{Y} \subsetneq \mathbf{X}$. Finally, a μ-center $\mathbf{X}$ is *minimal* if $\mu[\mathsf{Well}(\mathbf{Y})] = 0$ for any proper closed subset $\mathbf{Y} \subsetneq \mathbf{X}$. This concept was introduced by Hilmy [59] in the context of smooth dynamical systems; its ramifications for CA were first explored by Hurley [65].

Proposition 43 *Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})$, let $\mu \in \mathfrak{M}_{\mathsf{cons}}(\mathcal{A}^{\mathbb{M}}, \sigma)$, and let $\mathbf{X}$ be a μ-center.*

(a) *If μ is σ-ergodic, and $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$ is a subshift, then $\mu[\mathsf{Well}(\mathbf{X})] = 1$.*

(b) *If $\mathbb{M}$ is countable, and $\mathbf{X}$ is σ-minimal subshift with $\mu[\mathsf{Well}(\mathbf{X})] = 1$, then $\mathbf{X}$ is lean.*

(c) *Suppose $\mathbb{M} = \mathbb{Z}^D$ and μ is weakly σ-mixing. If $\mathbf{X}$ is a minimal μ-center, then $\mathbf{X}$ is a subshift, $\mathbf{X}$ is the only lean μ-center, and $\mu[\mathsf{Well}(\mathbf{X})] = 1$.*

Proof **(a)** and **(b)** are very similar to the proofs of Proposition 42(a,b).

(c) is proved for Bernoulli measures as Theorem B in [65]. The proof is quite similar to Proposition 42(c)[i], and again, we only need μ to be weakly mixing. □

Section 4 of [65] contains several examples of minimal μ-centers which are not μ-attractors. In particular, the analogue of Proposition 42(c)[ii] is false for μ-centers.

Kůrka–Maass μ-limit Sets If $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})$ and $\mu \in \mathfrak{M}_{\mathsf{cons}}(\mathcal{A}^{\mathbb{M}}, \sigma)$, then Kůrka and Maass define the *μ-limit set* of Φ:

$$\Lambda(\Phi, \mu) := \bigcap \left\{ \text{closed subsets } \mathbf{X} \subset \mathcal{A}^{\mathbb{M}} \; ; \; \lim_{t\to\infty} \Phi^t \mu(\mathbf{X}) = 1 \right\} .$$

It suffices to take this intersection only over all *cylinder* sets $\mathbf{X}$. By doing this, we see that $\Lambda(\Phi, \mu)$ is a subshift of $\mathcal{A}^{\mathbb{M}}$, and is defined by the following property: for any finite $\mathbb{B} \subset \mathbb{M}$ and any word $\mathbf{b} \in \mathcal{A}^{\mathbb{B}}$, $\mathbf{b}$ is admissible to $\Lambda(\Phi, \mu)$ if and only if $\liminf_{t\to\infty} \Phi^t \mu[\mathbf{b}] > 0$.

Proposition 44 *Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})$ and $\mu \in \mathfrak{M}_{\mathsf{cons}}(\mathcal{A}^{\mathbb{M}}, \sigma)$.*

(a) *If $wk * \lim_{t\to\infty} \Phi^t \mu = \nu$, then $\Lambda(\Phi, \mu) = \mathsf{supp}(\nu)$.*

Suppose $\mathbb{M} = \mathbb{Z}$.

(b) *If Φ is surjective and has an equicontinuous point, and μ has full support on $\mathcal{A}^{\mathbb{Z}}$, then $\Lambda(\Phi, \mu) = \mathcal{A}^{\mathbb{Z}}$.*

(c) *If Φ is left- or right-permutative and μ is connected* (see below), *then $\Lambda(\Phi, \mu) = \mathcal{A}^{\mathbb{Z}}$.*

Proof For **(a)**, see Proposition 2 in [79]. For **(b,c)**, see Theorems 2 and 3 in [77]; for earlier special cases of these results, see also Propositions 4 and 5 in [79]. □

Remark 45 (a) In Proposition 44(c), the measure μ is *connected* if there is some constant $C > 0$ such that, for any finite word $\mathbf{b} \in \mathcal{A}^*$, and any $a \in \mathcal{A}$, we have $\mu[\mathbf{b}\ a] \geq C \cdot \mu[\mathbf{b}]$ and $\mu[a\ \mathbf{b}] \geq C \cdot \mu[\mathbf{b}]$.

For example, any Bernoulli, Markov, or N-step Markov measure with full support is connected. Also, any measure with 'complete connections' (see Example 22(a)) is connected.

(b) Proposition 44(a) shows that μ-limit sets are closely related to the weak* limits of measures. Recall from Subsect. "Asymptotic Randomization by Linear Cellular Automata" that the uniform Bernoulli measure η is the weak* limit of a large class of initial measures under the action of linear CA. Presumably the same result should hold for a much larger class of *permutative* CA, but so far this is unproven, except in some special cases [see Remarks 41(d,e)]. Proposition .44(a,c) implies that the limit measure of a permutative CA (if it exists) must have full support – hence it can't be 'too far' from η.

Kůrka's Measure Attractors Let $\mathfrak{M}^{\sigma}_{\mathrm{inv}} := \mathfrak{M}_{\mathrm{ers}}(\mathcal{A}^{\mathbb{M}}, \sigma)$ have the weak* topology, and define $\Phi_* : \mathfrak{M}^{\sigma}_{\mathrm{inv}} \longrightarrow \mathfrak{M}^{\sigma}_{\mathrm{inv}}$ by $\Phi_*(\mu) = \mu \circ \Phi^{-1}$. Then Φ_* is continuous, so we can treat $(\mathfrak{M}^{\sigma}_{\mathrm{inv}}, \Phi_*)$ itself as a compact topological dynamical system. The "weak* limit measures" of Φ are simply the attracting fixed points of $(\mathfrak{M}^{\sigma}_{\mathrm{inv}}, \Phi_*)$. However, even if the Φ_*-orbit of a measure μ does not weak* converge to a fixed point, we can still consider the omega-limit set of μ. In particular, the limit set $\Phi_*^{\infty}(\mathfrak{M}^{\sigma}_{\mathrm{inv}})$ is the union of the omega-limit sets of all σ-invariant initial measures under Φ_*. Kůrka defines the *measure attractor* of Φ:

$$\mathsf{MeasAttr}(\Phi) := \overline{\bigcup \{\mathsf{supp}(\mu)\,;\ \ \mu \in \Phi_*^{\infty}(\mathfrak{M}^{\sigma}_{\mathrm{inv}})\}} \subseteq \mathcal{A}^{\mathbb{M}}.$$

(The bar denotes topological closure.) A configuration $\mathbf{a} \in \mathcal{A}^{\mathbb{Z}^D}$ is *densely recurrent* if any word which occurs in $\mathbf{a}$ does so with nonzero frequency. Formally, for any finite $\mathbb{B} \subset \mathbb{Z}^D$

$$\limsup_{N \to \infty} \frac{\#\left\{\mathsf{z} \in [-N \ldots N]^D\,;\ \ \mathbf{a}_{\mathbb{B}+\mathsf{z}} = \mathbf{a}_{\mathbb{B}}\right\}}{(2N+1)^D} > 0\,.$$

If $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^D}$ is a subshift, then the *densely recurrent subshift* of $\mathbf{X}$ is the closure $\mathbf{D}$ of the set of all densely recurrent points in $\mathbf{X}$. If $\mu \in \mathfrak{M}^{\sigma}_{\mathrm{inv}}(\mathbf{X})$, then the Birkhoff Ergodic Theorem implies that $\mathsf{supp}(\mu) \subseteq \mathbf{D}$; see Proposition 8.8 in [1], p. 164. From this it follows that $\mathfrak{M}^{\sigma}_{\mathrm{inv}}(\mathbf{X}) = \mathfrak{M}^{\sigma}_{\mathrm{inv}}(\mathbf{D})$. On the other hand, $\mathbf{D} = \bigcup \{\mathsf{supp}(\mu)\,;\ \ \mu \in \mathfrak{M}^{\sigma}_{\mathrm{inv}}(\mathbf{D})\}$. In other words, densely recurrent subshifts are the only subshifts which are 'covered' by their own set of shift-invariant measures.

Proposition 46 *Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})[\mathcal{A}^{\mathbb{Z}^D}]$. Let $\mathbf{D}$ be the densely recurrent subshift of $\Phi^{\infty}(\mathcal{A}^{\mathbb{Z}^D})$. Then $\mathbf{D} = \mathsf{MeasAttr}(\Phi)$, and $\Phi^{\infty}(\mathfrak{M}^{\sigma}_{\mathrm{inv}}) = \mathfrak{M}_{\mathrm{ers}}(\mathbf{D}, \sigma)$.*

Proof Case $D = 1$ is Proposition 13 in [78]. The same proof works for $D \geq 2$. □

Synthesis

The various hybrid modes of self-organization are related as follows:

Proposition 47 *Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})$.*

(a) *Let $\mu \in \mathfrak{M}_{\mathrm{ers}}(\mathcal{A}^{\mathbb{M}}, \sigma)$ and let $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$ be any closed set.*

[i] If $\mathbf{X}$ is a topological attractor and μ has full support, then $\mathbf{X}$ is a μ-attractor.
[ii] If $\mathbf{X}$ is a μ-attractor, then $\mathbf{X}$ is a μ-center.
[iii] Suppose $\mathbb{M} = \mathbb{Z}^D$, and that μ is weakly σ-mixing. Let $\mathbf{Y}$ be the intersection of all topological attractors of Φ. If Φ has a minimal μ-attractor $\mathbf{X}$, then $\mathbf{X} \subseteq \mathbf{Y}$.
[iv] If μ is σ-ergodic, then $\Lambda(\Phi, \mu) \subseteq \bigcap \{\mathbf{X} \subseteq \mathcal{A}^{\mathbb{M}}\,;\ \mathbf{X}$ a subshift and μ-attractor$\} \subseteq \Phi^{\infty}(\mathcal{A}^{\mathbb{Z}^D})$.
[v] Thus, if μ is σ-ergodic and has full support, then

$$\Lambda(\Phi, \mu) \subseteq \bigcap \left\{\mathbf{X} \subseteq \mathcal{A}^{\mathbb{M}}\,;\ \ \mathbf{X} \text{ a subshift and a topological attractor}\right\}.$$

[vi] If $\mathbf{X}$ is a subshift, then $(\Lambda(\Phi, \mu) \subseteq \mathbf{X}) \iff (\omega(\Phi_*, \mu) \subseteq \mathfrak{M}^{\sigma}_{\mathrm{inv}}(\mathbf{X}))$.

(b) *Let $\mathbb{M} = \mathbb{Z}^D$. Let $\mathfrak{B}$ be the set of all Bernoulli measures on $\mathcal{A}^{\mathbb{Z}^D}$, and for any $\beta \in \mathfrak{B}$, let $\mathbf{X}_\beta$ be the minimal β-attractor for Φ (if it exists).*
There is a comeager subset $\mathbf{A} \subset \mathcal{A}^{\mathbb{Z}^D}$ such that $\overline{\bigcup_{\beta \in \mathfrak{B}} \mathbf{X}_\beta} \subseteq \bigcap_{\mathbf{a} \in \mathbf{A}} \omega(\mathbf{a}, \Phi)$.

(c) *$\mathsf{MeasAttr}(\Phi) = \overline{\bigcup \left\{\Lambda(\Phi, \mu)\,;\ \ \mu \in \mathfrak{M}^{\sigma}_{\mathrm{inv}}(\mathcal{A}^{\mathbb{M}})\right\}}$.*

(d) *If $\mathbb{M} = \mathbb{Z}^D$, then $\mathsf{MeasAttr}(\Phi) \subseteq \Phi^{\infty}(\mathcal{A}^{\mathbb{Z}^D})$.*

Proof **(a)[i]:** If $\mathbf{U}$ is a clopen subset and $\Phi^{\infty}(\mathbf{U}) = \mathbf{X}$, then $\mathbf{U} \subseteq \mathsf{Basin}(\mathbf{X})$; thus, $0 < \mu[\mathbf{U}] \leq \mu[\mathsf{Basin}(\mathbf{X})]$, where the "<" is because μ has full support.

(a)[ii]: For any $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$, it is easy to see that $\mathsf{Cent}(\mathbf{a}, \Phi) \subseteq \omega(\mathbf{a}, \Phi)$. Thus, $\mathsf{Well}(\mathbf{X}) \supseteq \mathsf{Basin}(\mathbf{X})$. Thus, $\mu[\mathsf{Well}(\mathbf{X})] \geq \mu[\mathsf{Basin}(\mathbf{X})] > 0$.

(a)[iii] is Proposition 3.3 in [64]. (Again, Hurley states and proves this in the case when μ is a Bernoulli measure, but his proof only requires weak mixing).

(a)[iv]: Let $\mathbf{X}$ be a subshift and a μ-attractor; we claim that $\Lambda(\Phi, \mu) \subseteq \mathbf{X}$. Proposition 42(a) says $\mu[\mathsf{Basin}(\mathbf{X})] = 1$.

Let $\mathbb{B} \subset \mathbb{M}$ be any finite set. If $\mathbf{b} \in \mathcal{A}^{\mathbb{B}} \setminus \mathbf{X}_{\mathbb{B}}$, then

$$\left\{\mathbf{a} \in \mathcal{A}^{\mathbb{M}}\,;\ \exists\, T \in \mathbb{N} \text{ such that } \forall t \geq T\,,\, \Phi^t(\mathbf{a})_{\mathbb{B}} \neq \mathbf{b}\right\} \supseteq \mathsf{Basin}(\mathbf{X})\,.$$

It follows that the left-hand set has μ-measure 1, which implies that $\lim_{t\to\infty} \Phi^t \mu\langle \mathbf{b}\rangle = 0$ – hence $\mathbf{b}$ is a forbidden word in $\Lambda(\Phi, \mu)$.

Thus, all the words forbidden in $\mathbf{X}$ are also forbidden in $\Lambda(\Phi, \mu)$. Thus $\Lambda(\Phi, \mu) \subseteq \mathbf{X}$. (The case $\mathbb{M} = \mathbb{Z}$ of (a)[iv] appears as Prop. 1 in [79] and as Prop. II.27, p. 67 in [120]; see also Cor. II.30 in [120] for a slightly stronger result.)

(a)[v] follows from (a)[iv] and (a)[i].

(a)[vi] is Proposition 1 in [77] or Proposition 10 in [78]; the argument is fairly similar to (a)[iv]. (Kůrka assumes $\mathbb{M} = \mathbb{Z}$, but this is not necessary.)

(b) is Proposition 5.2 in [64].

(c) Let $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$ be a subshift and let $\mathfrak{M}^{\sigma}_{\mathrm{inv}} = \mathfrak{M}^{\sigma}_{\mathrm{inv}}(\mathcal{A}^{\mathbb{M}})$. Then

$$\begin{aligned}
&(\mathsf{Meas}\mathsf{Attr}(\Phi) \subseteq \mathbf{X})\\
&\iff (\mathsf{supp}(\nu) \subseteq \mathbf{X}\,,\ \ \forall\, \nu \in \Phi_*^{\infty}(\mathfrak{M}^{\sigma}_{\mathrm{inv}}))\\
&\iff (\nu \in \mathfrak{M}^{\sigma}_{\mathrm{inv}}(\mathbf{X})\,,\ \ \forall\, \nu \in \Phi_*^{\infty}(\mathfrak{M}^{\sigma}_{\mathrm{inv}}))\\
&\iff (\Phi_*^{\infty}(\mathfrak{M}^{\sigma}_{\mathrm{inv}}) \subseteq \mathfrak{M}^{\sigma}_{\mathrm{inv}}(\mathbf{X}))\\
&\iff (\omega(\Phi_*, \mu) \subseteq \mathfrak{M}^{\sigma}_{\mathrm{inv}}(\mathbf{X})\,,\ \ \forall\, \mu \in \mathfrak{M}^{\sigma}_{\mathrm{inv}})\\
&\underset{(*)}{\iff} (\Lambda(\Phi, \mu) \subseteq \mathbf{X}\,,\ \ \forall\, \mu \in \mathfrak{M}^{\sigma}_{\mathrm{inv}})\\
&\iff (\bigcup\{\Lambda(\Phi, \mu)\ \ \mu \in \mathfrak{M}^{\sigma}_{\mathrm{inv}}\} \subseteq \mathbf{X})\,.
\end{aligned}$$

where $(*)$ is by (a)[vi]. It follows that $\mathsf{Meas}\mathsf{Attr}(\Phi) = \overline{\bigcup\{\Lambda(\Phi, \mu)\,;\ \ \mu \in \mathfrak{M}^{\sigma}_{\mathrm{inv}}(\mathcal{A}^{\mathbb{M}})\}}$.

(d) follows immediately from Proposition 46. □

Examples and Applications The most natural examples of these hybrid modes of self-organization arise in the *particle cellular automata* (PCA) introduced in Subsect. "Domains, Defects, and Particles". The long-term dynamics of a PCA involves a steady reduction in particle density, as particles coalesce or annihilate one another in collisions. Thus, presumably, for almost any initial configuration $\mathbf{a} \in \mathcal{A}^{\mathbb{Z}}$, the sequence $\{\Phi^t(\mathbf{a})\}_{t=1}^{\infty}$ should converge to the subshift $\mathbf{Z}$ of configurations containing no particles (or at least, no particles of certain types), as $t \to \infty$. Unfortunately, this presumption is generally *false* if we interpret 'convergence' in the strict topological dynamical sense: the occasional particles will continue to wander near the origin at arbitrarily large times in the future orbit of $\mathbf{a}$ (albeit with diminishing frequency), so $\omega(\mathbf{a}, \Phi)$ will *not* be contained in $\mathbf{Z}$. However, the presumption becomes true if we instead employ one of the more flexible hybrid notions introduced above. For example, most initial probability measures μ should converge, under iteration of Φ to a measure concentrated on configurations with few or no particles; hence we expect that $\Lambda(\Phi, \mu) \subseteq \mathbf{Z}$. As discussed in Subsect. "Domains, Defects, and Particles", a result about self-organization in a PCA can sometimes be translated into an analogous result about self-organization in associated coalescent-domain CA.

Proposition 48 *Let $\mathcal{A} = \{0, \pm 1\}$ and let $\Psi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ be the Ballistic Annihilation Model* (BAM) *from* Example 33. *Let $\mathbf{R} := \{0, 1\}^{\mathbb{Z}}$ and $\mathbf{L} := \{0, -1\}^{\mathbb{Z}}$.*

(a) *If $\mu \in \mathfrak{M}_{\mathrm{eas}}(\mathcal{A}^{\mathbb{Z}}, \sigma)$, then $\nu = wk * \lim_{t\to\infty} \Psi^t(\mu)$ exists, and has one of three forms: either $\nu \in \mathfrak{M}_{\mathrm{eas}}(\mathbf{R}, \sigma)$, or $\nu \in \mathfrak{M}_{\mathrm{eas}}(\mathbf{L}, \sigma)$, or $\nu = \delta_0$, the point mass on the sequence $\mathbf{0} = [\ldots 000 \ldots]$.*

(b) *Thus, the measure attractor of Φ is $\mathbf{R} \cup \mathbf{L}$* (note that $\mathbf{R} \cap \mathbf{L} = \{\mathbf{0}\}$).

(c) *In particular, if μ is a Bernoulli measure on $\mathcal{A}^{\mathbb{Z}}$ with $\mu[+1] = \mu[-1]$, then $\nu = \delta_0$.*

(d) *Let μ be a Bernoulli measure on $\mathcal{A}^{\mathbb{Z}}$.*

[i] If $\mu[+1] > \mu[-1]$, then $\mathbf{R}$ is a μ-attractor – i. e. $\mu[\mathsf{Basin}(\mathbf{R})] > 0$.

[ii] If $\mu[+1] < \mu[-1]$, then $\mathbf{L}$ is a μ-attractor.

[iii] If $\mu[+1] = \mu[-1]$, then $\{\mathbf{0}\}$ is not a μ-attractor, because $\mu[\mathsf{Basin}\{\mathbf{0}\}] = 0$. However, $\Lambda(\Phi, \mu) = \{\mathbf{0}\}$.

Proof (a) is Theorem 6 of [8], and (b) follows from (a). (c) follows from Theorem 2 of [42]. (d)[i,ii] were first observed by Gilman (see Sect. 3, pp. 111–112 in [45], and later by Kůrka and Maass (see Example 4 in [80]). (d)[iii] follows immediately from (c): the statement $\Lambda(\Phi, \mu) = \{\mathbf{0}\}$ is equivalent to asserting that $\lim_{t\to\infty} \Phi^t \mu[\pm 1] = 0$, which a consequence of (c). Another proof of (d)[iii] is Proposition 11 in [80]; see also Example 3 in [79] or Prop. II.32, p. 70 in [120]. □

Corollary 49 *Let $\mathcal{A} = \mathbb{Z}_{/3}$, let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ be the CCA_3* (see Example 32), *and let η be the uniform Bernoulli measure on $\mathcal{A}^{\mathbb{Z}}$. Then $wk * \lim\limits_{t\to\infty} \Phi^t(\eta) = \frac{1}{3}(\delta_0 + \delta_1 + \delta_2)$, where δ_a is the point mass on the sequence $[\ldots aaa \ldots]$ for each $a \in \mathcal{A}$.*

Proof Combine Proposition 48(c) with the factor map Γ in Example 34(b). See Theorem 1 of [42] for details. □

Corollary 50 *Let $\mathcal{A} = \{0, 1\}$, let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ be ECA#184* (see Example 31).

(a) *$\mathsf{MeasAttr}(\Phi) = \mathbf{R} \cup \mathbf{L}$, where $\mathbf{R} \subset \mathcal{A}^{\mathbb{Z}}$ is the set of sequences not containing* [11], *and $\mathbf{L} \subset \mathcal{A}^{\mathbb{Z}}$ is the set of sequences not containing* [00].

(b) *If η is the uniform Bernoulli measure on $\mathcal{A}^{\mathbb{Z}}$, then $wk * \lim_{t\to\infty} \Phi^t(\eta) = \frac{1}{2}(\delta_0 + \delta_1)$, where δ_0 and δ_1 are the point masses on* [...010.101...] *and* [...101.010...].

(c) *Let μ be a Bernoulli measure on $\mathcal{A}^{\mathbb{Z}}$.*

[i] If $\mu[0] > \mu[1]$, then $\mathbf{R}$ is a μ-attractor – i.e. $\mu[\mathsf{Basin}(\mathbf{R})] > 0$.

[ii] If $\mu[0] < \mu[1]$, then $\mathbf{L}$ is a μ-attractor.

Proof sketch Let Γ be the factor map from Example 34(a). To prove (a), apply Γ to Proposition 48(b); see Example 26, Sect. 9 in [78] for details. To prove (b), apply Γ to Proposition 48(c); see Proposition 12 in [80] for details. To prove (c), apply Γ to Proposition 48(d)[i,ii]; □

Remark 51 (a) The other parts of Proposition 48 can likewise be translated into equivalent statements about the measure attractors and μ-attractors of CCA_3 and ECA#184.

(b) Recall that ECA#184 is a model of single-lane, traffic, where each car is either stopped or moving rightwards at unit speed. Blank [10] has extended Corollary 50(c) to a much broader class of CA models of multi-lane, multi-speed traffic. For any such model, let $\mathbf{R} \subset \mathcal{A}^{\mathbb{Z}}$ be the set of 'free flowing' configurations where each car has enough space to move rightwards at its maximum possible speed. Let $\mathbf{L} \subset \mathcal{A}^{\mathbb{Z}}$ be the set of 'jammed' configurations where the cars are so tightly packed that the jammed clusters can propagate (leftwards) through the cars at maximum speed. If μ is any Bernoulli measure, then $\mu[\mathsf{Basin}(\mathbf{R})] = 1$ if the μ-average density of cars is greater than 1/2, whereas $\mu[\mathsf{Basin}(\mathbf{L})] = 1$ if the density is less than 1/2 Theorems 1.2 and 1.3 in [10]. Thus, $\mathbf{L} \sqcup \mathbf{R}$ is a (non-lean) μ-attractor, although not a topological attractor Lemma 2.13 in [10].

Example 52 A *cyclic addition and ballistic annihilation model* (CABAM) contains the same 'moving' particles ± 1 as the BAM (Example 33), but also has one or more 'stationary' particle types. Let $3 \le N \in \mathbb{N}$, and let $\mathcal{P} = \{1, 2, \ldots, N-1\} \subset \mathbb{Z}_{/N}$, where we identify $N-1$ with -1, modulo N. It will be convenient to represent the 'vacant' state $\emptyset$ as 0; thus, $\mathcal{A} = \mathbb{Z}_{/N}$. The particles 1 and –1 have velocities and collisions as in the BAM, namely:

$$\mathsf{v}(1) = 1\,, \quad \mathsf{v}(-1) = -1\,, \quad \text{and} \quad -1 + 1 \rightsquigarrow \emptyset\,.$$

We set $\mathsf{v}(p) = 0$, for all $p \in [2 \ldots N-2]$, and employ the following collision rule:

$$\text{If } p_{-1} + p_0 + p_1 \equiv q, \quad (\text{mod } N), \quad \text{then} \quad p_{-1} + p_0 + p_1 \rightsquigarrow q. \tag{5}$$

(here, any one of p_{-1}, p_0, p_1, or q could be 0, signifying vacancy). For example, if $N = 5$ and a (rightward moving) type $+1$ particle strikes a (stationary) type 3 particle, then the $+1$ particle is annihilated and the 3 particle turns into a (stationary) 4 particle. If another $+1$ particle hits the 4 particle, then both are annihilated, leaving a vacancy (0).

Let $\mathcal{B} = \mathbb{Z}_{/N}$, and let $\Psi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ be the CABAM. Then the set of fixed points of Ψ is $\mathbf{F} = \{\mathbf{f} \in \mathcal{B}^{\mathbb{Z}}; f_{\mathsf{z}} \neq \pm 1, \forall \mathsf{z} \in \mathbb{Z}\}$. Note that, if $\mathbf{b} \in \mathsf{Basin}[\mathbf{F}]$ – that is, if $\omega(\mathbf{b}, \Psi) \subseteq \mathbf{F}$ – then in fact $\lim_{t\to\infty} \Psi^t(\mathbf{b})$ exists and is a Ψ-fixed point.

Proposition 53 *Let $\mathcal{B} = \mathbb{Z}_{/N}$, let $\Psi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ be the CABAM, and let η be the uniform Bernoulli measure on $\mathcal{B}^{\mathbb{Z}}$. If $N \ge 5$, then $\mathbf{F}$ is a 'global' η-attractor – that is, $\eta[\mathsf{Basin}(\mathbf{F})] = 1$. However, if $N \le 4$, then $\eta[\mathsf{Basin}(\mathbf{F})] = 0$.*

Proof See Theorem 1 of [41]. □

Let $\mathcal{A} = \mathbb{Z}_{/N}$ and let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ be the N-color CCA from Example 32. Then the set of fixed points of Φ is $\mathbf{F} = \{\mathbf{f} \in \mathcal{A}^{\mathbb{Z}}; f_{\mathsf{z}} - f_{\mathsf{z}+1} \neq \pm 1, \forall \mathsf{z} \in \mathbb{Z}\}$. Note that, if $\mathbf{a} \in \mathsf{Basin}[\mathbf{F}]$, then in fact $\lim_{t\to\infty} \Phi^t(\mathbf{a})$ exists and is a Φ-fixed point.

Corollary 54 *Let $\mathcal{A} = \mathbb{Z}_{/N}$, let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ be the N-color CCA, and let η be the uniform Bernoulli measure on $\mathcal{A}^{\mathbb{Z}}$ If $N \ge 5$, then $\mathbf{F}$ is a 'global' η-attractor – that is, $\eta[\mathsf{Basin}(\mathbf{F})] = 1$. However, if $N \le 4$, then $\eta[\mathsf{Basin}(\mathbf{F})] = 0$.*

Proof sketch Let $\mathcal{B} = \mathbb{Z}_{/N}$ and let $\Psi \in \mathsf{CA}(\mathcal{B}^{\mathbb{Z}})$ be the N-particle CABAM. Construct a factor map $\Gamma\colon \mathcal{A}^{\mathbb{Z}} \longrightarrow \mathcal{B}^{\mathbb{Z}}$ with local rule $\gamma(a_0, a_1) := (a_0 - a_1)$ mod N, similar to Example 34(b). Then $\Gamma \circ \Phi = \Psi \circ \Gamma$, and the Ψ-particles track the Φ-domain boundaries. Now apply Γ to Proposition 53. □

Example 55 Let $\mathcal{A} = \{0, 1\}$ and let $\mathbb{H} = \{-1, 0, 1\}$. *Elementary Cellular Automaton #18* is the one-dimensional CA with local rule $\phi\colon \mathcal{A}^{\mathbb{H}} \longrightarrow \mathcal{A}$ given: $\phi[100] = 1 = \phi[001]$, and $\phi(\mathbf{a}) = 0$ for all other $\mathbf{a} \in \mathcal{A}^{\mathbb{H}}$.

Empirically, ECA#18 has one stable phase: the *odd sofic shift* $\mathbf{S}$, defined by the $\mathcal{A}$-labeled digraph ① ⇆ ⓪ ⇆ ⓪. In other words, a sequence is admissible to $\mathbf{S}$ as long as an pair of consecutive ones are separated by an *odd* number of zeroes. Thus, a *defect* is any word of the form $10^{2m}1$ (where 0^{2m} represents $2m$ zeroes) for any $m \in \mathbb{N}$. Thus, defects can be arbitrarily large, they can grow and move arbitrarily quickly, and they can coalesce across arbitrarily large distances. Thus, it is impossible to construct a particle CA which tracks the motion of these defects. Nevertheless, in computer simulations, one can visually follow the moving defects through time, and they appear to per-

form random walks. Over time, the density of defects decreases as they randomly collide and annihilate. This was empirically observed by Grassberger [47,48] and Boccara et al. [11]. Lind (see Sect. 5 in [82]) conjectured that this gradual elimination of defects caused almost all initial conditions to converge, in some sense, to $\mathbf{S}$ under application of Φ.

Eloranta and Numelin [34] proved that the defects of Φ individually perform random walks. However, the motions of neighboring defects are highly correlated. They are not *independent* random walks, so one cannot use standard results about stochastic interacting particle systems to conclude that the defect density converges to zero. To solve problems like this, Kůrka [77] developed a theory of 'particle weight functions' for CA.

Let $\mathcal{A}^*$ be the set of all finite words in the alphabet $\mathcal{A}$. A *particle weight function* is a bounded function $p\colon \mathcal{A}^* \longrightarrow \mathbb{N}$, so that, for any $\mathbf{a} \in \mathcal{A}^{\mathbb{Z}}$, we interpret

$$\#_p(\mathbf{a}) := \sum_{r=0}^{\infty} \sum_{z\in\mathbb{Z}} p(\mathbf{a}_{[z\ldots z+r]}) \quad \text{and}$$

$$\delta_p(\mathbf{a}) := \sum_{r=0}^{\infty} \lim_{N\to\infty} \frac{1}{2N} \sum_{z=-N}^{N} p(\mathbf{a}_{[z\ldots z+r]})$$

to be, respectively the 'number of particles' and 'density of particles' in configuration $\mathbf{a}$ (clearly $\#_p(\mathbf{a})$ is finite if and only if $\delta_p(\mathbf{a}) = 0$). The function p can count the single-letter 'particles' of a PCA, or the short-length 'domain boundaries' found in ECA#184 and the CCA of Examples 31 and 32. However, p can also track the arbitrarily large defects of ECA#18. For example, define $p_{18}(10^{2m}1) = 1$ (for any $m \in \mathbb{N}$), and define $p_{18}(\mathbf{a}) = 0$ for all other $\mathbf{a} \in \mathcal{A}^*$.

Let $\mathbf{Z}_p := \{\mathbf{a} \in \mathcal{A}^{\mathbb{Z}}; \#_p(\mathbf{a}) = 0\}$ be the set of *vacuum configurations*. (For example, if $p = p_{18}$ as above, then $\mathbf{Z}_p$ is just the odd sofic shift $\mathbf{S}$.) If the iteration of a CA Φ decreases the number (or density) of particles, then one expects $\mathbf{Z}_p$ to be a limit set for Φ in some sense. Indeed, if $\mu \in \mathfrak{M}^{\sigma}_{\mathrm{Mrv}} := \mathfrak{M}_{\mathrm{eqs}}(\mathcal{A}^{\mathbb{Z}}, \sigma)$, then we define $\Delta_p(\mu) := \int_{\mathcal{A}^{\mathbb{Z}}} \delta_p \, \mathrm{d}\mu$. If Φ is 'p-decreasing' in a certain sense, then Δ_p acts as a Lyapunov function for the dynamical system $(\mathfrak{M}^{\sigma}_{\mathrm{Mrv}}, \Phi_*)$. Thus, with certain technical assumptions, we can show that, if $\mu \in \mathfrak{M}^{\sigma}_{\mathrm{Mrv}}$ is connected, then $\Lambda(\Phi, \mu) \subseteq \mathbf{Z}_p$ (see Theorem 8 in [77]). Furthermore, under certain conditions, $\mathsf{Meas}\mathsf{Attr}(\Phi) \subseteq \mathbf{Z}_p$ (see Theorem 7 in [77]). Using this machinery, Kůrka proved:

Proposition 56 *Let $\Phi\colon \mathcal{A}^{\mathbb{Z}} \longrightarrow \mathcal{A}^{\mathbb{Z}}$ be ECA#18, and let $\mathbf{S} \subset \mathcal{A}^{\mathbb{Z}}$ be the odd sofic shift. If $\mu \in \mathfrak{M}_{\mathrm{eqs}}(\mathcal{A}^{\mathbb{Z}}, \sigma)$ is connected, then $\Lambda(\Phi, \mu) \subseteq \mathbf{S}$.*

Proof See Example 6.3 of [77]. □

Measurable Dynamics

If $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})$ and $\mu \in \mathfrak{M}_{\mathrm{eqs}}(\mathcal{A}^{\mathbb{M}}, \Phi)$, then the triple $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is a *measure-preserving dynamical system* (MPDS), and thus, amenable to the methods of classical ergodic theory.

Mixing and Ergodicity

If $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})$, then the topological dynamical system $(\mathcal{A}^{\mathbb{M}}, \Phi)$ is *topologically transitive* (or *topologically ergodic*) if, if, for any open subsets $\mathbf{U}, \mathbf{V} \subseteq \mathcal{A}^{\mathbb{M}}$, there exists $t \in \mathbb{N}$ such that $\mathbf{U} \cap \Phi^{-t}(\mathbf{V}) \neq \emptyset$. Equivalently, there exists some $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$ whose orbit $\mathcal{O}(\mathbf{a}) := \{\Phi^t(\mathbf{a})\}_{t=0}^{\infty}$ is dense in $\mathcal{A}^{\mathbb{M}}$. If $\mu \in \mathfrak{M}_{\mathrm{eqs}}(\mathcal{A}^{\mathbb{M}}, \Phi)$, then the system $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is *ergodic* if, for any nontrivial measurable $\mathbf{U}, \mathbf{V} \subseteq \mathcal{A}^{\mathbb{M}}$, there exists some $t \in \mathbb{N}$ such that $\mu[\mathbf{U} \cap \Phi^{-t}(\mathbf{V})] > 0$. The system $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is *totally ergodic* if $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi^n)$ is ergodic for every $n \in \mathbb{N}$. The system $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is *(strongly) mixing* if, for any nontrivial measurable $\mathbf{U}, \mathbf{V} \subseteq \mathcal{A}^{\mathbb{M}}$.

$$\lim_{t\to\infty} \mu\left[\mathbf{U} \cap \Phi^{-t}(\mathbf{V})\right] = \mu[\mathbf{U}] \cdot \mu[\mathbf{V}] \,. \tag{6}$$

The system $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is *weakly mixing* if the limit (6) holds as $n\to\infty$ along an increasing subsequence $\{t_n\}_{n=1}^{\infty}$ of *density one* – i. e. such that $\lim_{n\to\infty} t_n/n = 1$. For any $M \in \mathbb{N}$, we say $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is *M-mixing* if, for any measurable $\mathbf{U}_0, \mathbf{U}_1, \ldots, \mathbf{U}_M \subseteq \mathcal{A}^{\mathbb{M}}$.

$$\lim_{\substack{|t_n - t_m| \to \infty \\ \forall n \neq m \in [0\ldots M]}} \mu\left[\bigcap_{m=0}^{M} \Phi^{-t_m}(\mathbf{U}_m)\right] = \prod_{m=0}^{M} \mu[\mathbf{U}_m] \tag{7}$$

(thus, 'strong' mixing is 1-mixing). We say $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is *multimixing* (or *mixing of all orders*) if $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is M-mixing for all $M \in \mathbb{N}$.

We say $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is a *Bernoulli endomorphism* if its natural extension (▶ Ergodic Theory: Basic Examples and Constructions) is measurably isomorphic to a system $(\mathcal{B}^{\mathbb{Z}}, \beta; \sigma)$, where $\beta \in \mathfrak{M}_{\mathrm{eqs}}(\mathcal{B}^{\mathbb{Z}}; \sigma)$ is a Bernoulli measure. We say $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is a *Kolmogorov endomorphism* if its natural extension is a Kolmogorov (or "K") automorphism; see ▶ Ergodicity and Mixing Properties. The following chain of implications is well-known; see ▶ Ergodicity and Mixing Properties.

Theorem 57 *Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})$, let $\mu \in \mathfrak{M}_{\mathrm{eqs}}(\mathcal{A}^{\mathbb{M}}; \Phi)$, and let $\mathbf{X} = \mathrm{supp}(\mu)$. Then $\mathbf{X}$ is a compact, Φ-invariant set. Furthermore:*
(μ, Φ) is Bernoulli $\Longrightarrow$ (μ, Φ) is Kolmogorov $\Longrightarrow$ (μ, Φ) is multimixing $\Longrightarrow$ (μ, Φ) is mixing

$\Longrightarrow$ *(μ, Φ) is weakly mixing* $\Longrightarrow$ *(μ, Φ) is totally ergodic* $\Longrightarrow$ *(μ, Φ) is ergodic* $\Longrightarrow$ *The system $(\mathbf{X}, \Phi)$ is topologically transitive* $\Longrightarrow$ *$\Phi : \mathbf{X} \longrightarrow \mathbf{X}$ is surjective.*

Theorem 58 *Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{N}})$ be posexpansive* (see *Subsect. "Posexpansive and Permutative CA"*). *Then $(\mathcal{A}^{\mathbb{N}}, \Phi)$ has topological entropy $\log_2(k)$ for some $k \in \mathbb{N}$, Φ preserves the uniform measure η, and $(\mathcal{A}^{\mathbb{N}}, \eta; \Phi)$ is a uniformly distributed Bernoulli endomorphism on an alphabet of cardinality k.*

Proof Extend the argument of Theorem 18. See Corollary 3.10 in [9] or Theorem 4.8(5) in [86]. □

Example 59 Suppose $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{N}})$ is right-permutative, with neighborhood $[r \ldots R]$, where $0 \leq r < R$. Then $h_{\text{top}}(\Phi) = \log_2(|\mathcal{A}|^R)$, so Theorem 58 says that $(\mathcal{A}^{\mathbb{N}}, \eta; \Phi)$ is a uniformly distributed Bernoulli endomorphism on the alphabet $\mathcal{B} := \mathcal{A}^R$.

In this case, it is easy to see this directly. If $\Phi_{\mathbb{B}}^{\mathbb{N}} : \mathcal{A}^{\mathbb{N}} \longrightarrow \mathcal{B}^{\mathbb{N}}$ is as in Eq. (2), then $\beta := \Phi_{\mathbb{B}}^{\mathbb{N}}(\eta)$ is the uniform Bernoulli measure on $\mathcal{B}^{\mathbb{N}}$, and $\Phi_{\mathbb{B}}^{\mathbb{N}}$ is an isomorphism from $(\mathcal{A}^{\mathbb{N}}, \mu; \Phi)$ to $(\mathcal{B}^{\mathbb{N}}, \beta; \sigma)$.

Theorem 60 *Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ have neighborhood $[L \ldots R]$. Suppose that*
either (a) *$0 \leq L < R$ and Φ is right-permutative;*
or (b) *$L < R \leq 0$ and Φ is left-permutative;*
or (c) *$L < R$ and Φ is bipermutative;*
or (d) *Φ is posexpansive.*
Then Φ preserves the uniform measure η, and $(\mathcal{A}^{\mathbb{Z}}, \eta; \Phi)$ is a Bernoulli endomorphism.

Proof For cases **(a)** and **(b)**, see Theorem 2.2 in [125]. For case **(c)**, see Theorem 2.7 in [125] or Corollary 7.3 in [74]. For **(d)**, extend the argument of Theorem 14; see Theorem 4.9 in [86]. □

Remark Theorem 60(c) can be extended to some higher-dimensional permutative CA using Proposition 1 in [3]; see Remark 12(b).

Theorem 61 *Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ have neighborhood $[L \ldots R]$. Suppose that*
either (a) *Φ is surjective and $0 < L \leq R$;*
or (b) *Φ is surjective and $L \leq R < 0$;*
or (c) *Φ is right-permutative and $R \neq 0$;*
or (d) *Φ is left-permutative and $L \neq 0$.*
Then Φ preserves η, and $(\mathcal{A}^{\mathbb{Z}}, \eta; \Phi)$ is a Kolmogorov endomorphism.

Proof Cases **(a)** and **(b)** are Theorem 2.4 in [125]. Cases **(c)** and **(d)** are from [129]. □

Corollary 62 *Any CA satisfying the hypotheses of Theorem 61 is multimixing.*

Proof This follows from Theorems 57 and 61. See also Theorem 3.2 in [131] for a direct proof that any CA satisfying hypotheses (a) or (b) is 1-mixing. See Theorem 6.6 in [74] for a proof that any CA satisfying hypotheses (c) or (d) is multimixing. □

Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}^D})$ have neighborhood $\mathbb{H}$. An element $\mathsf{x} \in \mathbb{H}$ is *extremal* if $\langle \mathsf{x}, \mathsf{x} \rangle > \langle \mathsf{x}, \mathsf{h} \rangle$ for all $\mathsf{h} \in \mathbb{H} \setminus \{\mathsf{x}\}$. We say Φ is *extremally permutative* if Φ is permutative in some extremal coordinate.

Theorem 63 *Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}^D})$ and let η be the uniform measure. If Φ is extremally permutative, then $(\mathcal{A}^{\mathbb{Z}^D}, \eta; \Phi)$ is mixing.*

Proof See Theorem A in [144] for the case $D = 2$ and $\mathcal{A} = \mathbb{Z}_{/2}$. Willson described Φ as 'linear' in an extremal coordinate (which is equivalent to permutative when $\mathcal{A} = \mathbb{Z}_{/2}$), and then concluded that Φ was 'ergodic' – however, he did this by explicitly showing that Φ was mixing. His proof technique easily generalizes to any extremally permutative CA on any alphabet, and any $D \geq 1$. □

Theorem 64 *Let $\mathcal{A} = \mathbb{Z}_{/m}$. Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}^D})$ have linear local rule $\phi : \mathcal{A}^{\mathbb{H}} \longrightarrow \mathcal{A}$ given by $\phi(\mathbf{a}_{\mathbb{H}}) = \sum_{\mathsf{h} \in \mathbb{H}} c_{\mathsf{h}} \cdot a_{\mathsf{h}}$, where $c_{\mathsf{h}} \in \mathbb{Z}$ for all $\mathsf{h} \in \mathbb{H}$. Let η be the uniform measure on $\mathcal{A}^{\mathbb{Z}^D}$. The following are equivalent:*
(a) *Φ preserves η and $(\mathcal{A}^{\mathbb{Z}^D}, \eta, \Phi)$ is ergodic.*
(b) *$(\mathcal{A}^{\mathbb{Z}^D}, \Phi)$ is topologically transitive.*
(c) *$\gcd\{c_{\mathsf{h}}\}_{0 \neq \mathsf{h} \in \mathbb{H}}$ is coprime to m.*
(d) *For all prime divisors p of m, there is some nonzero $\mathsf{h} \in \mathbb{H}$ such that c_{h} is not divisible by p.*

Proof Theorem 3.2 in [18]; see also [17]. For a different proof in the case $D = 2$, see Theorem 6 in [123]. □

Spectral Properties

If $\mu \in \mathfrak{M}_{\mathsf{cns}}(\mathcal{A}^{\mathbb{M}})$, then let $\mathbf{L}^2_\mu = \mathbf{L}^2(\mathcal{A}^{\mathbb{M}}, \mu)$ be the set of measurable functions $f : \mathcal{A}^{\mathbb{M}} \longrightarrow \mathbb{C}$ such that $\|f\|_2 := (\int_{\mathcal{A}^{\mathbb{M}}} |f|^2 \, \mathrm{d}\mu)^{1/2}$ is finite. If $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})$ and $\mu \in \mathfrak{M}_{\mathsf{cns}}(\mathcal{A}^{\mathbb{M}}, \Phi)$, then Φ defines a unitary linear operator $\Phi_* : \mathbf{L}^2_\mu \longrightarrow \mathbf{L}^2_\mu$ by $\Phi_*(f) = f \circ \Phi$ for all $f \in \mathbf{L}^2_\mu$. If $f \in \mathbf{L}^2_\mu$, then f is an *eigenfunction* of Φ, with *eigenvalue* $c \in \mathbb{C}$, if $\Phi_*(f) = c \cdot f$. By definition of Φ_*, any eigenvalue must be an element of the unit circle $\mathbb{T} := \{c \in \mathbb{C} \; ; \; |c| = 1\}$. Let $\mathbb{S}_\Phi \subset \mathbb{T}$ be the set of all eigenvalues of Φ, and for any $s \in \mathbb{S}_\Phi$, let $\mathbf{E}_s(\Phi) := \{f \in \mathbf{L}^2_\mu \; ; \; \Phi_* f = s f\}$ be the corresponding eigenspace. For example, if f is constant μ-almost everywhere, then $f \in \mathbf{E}_1(\Phi)$. Let $\mathbf{E}(\Phi) := \bigsqcup_{s \in \mathbb{S}_\Phi} \mathbf{E}_s(\Phi)$. Note that $\mathbb{S}_\Phi$ is a group. Indeed, if $s_1, s_2 \in \mathbb{S}_\Phi$, and $f_1 \in \mathbf{E}_{s_1}$ and $f_2 \in \mathbf{E}_{s_2}$,

then $(f_1 f_2) \in \mathbf{E}_{s_1 s_2}$ and $(1/f_1) \in \mathbf{E}_{1/s_1}$. Thus, $\mathbb{S}_\Phi$ is called the *spectral group* of Φ.

If $s \in \mathbb{S}_\Phi$, then heuristically, an s-eigenfunction is an 'observable' of the dynamical system $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ which exhibits quasiperiodically recurrent behavior. Thus, the spectral properties of Φ characterize the 'recurrent aspect' of its dynamics (or the lack thereof). For example:

- $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is ergodic $\iff$ $\mathbf{E}_1(\Phi)$ contains only constant functions $\iff$ $\dim[\mathbf{E}_s(\Phi)] = 1$ for all $s \in \mathbb{S}_\Phi$.
- $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is weakly mixing (see Subsect. "Mixing and Ergodicity") $\iff$ $\mathbf{E}(\Phi)$ contains only constant functions $\iff$ $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is ergodic and $\mathbb{S}_\Phi = \{1\}$.

We say $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ has *discrete spectrum* if $\mathbf{L}^2_\mu$ is spanned by $\mathbf{E}(\Phi)$. In this case, $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is measurably isomorphic to an MPDS defined by translation on a compact abelian group (e. g. an irrational rotation of a torus, an odometer, etc.).

If $\mu \in \mathfrak{M}_{\mathrm{erg}}(\mathcal{A}^{\mathbb{M}}, \sigma)$, then there is a natural unitary $\mathbb{M}$-action on $\mathbf{L}^2_\mu$, where $\sigma^{\mathsf{m}}_*(f) = f \circ \sigma^{\mathsf{m}}_*$. A *character* of $\mathbb{M}$ is a monoid homomorphism $\chi : \mathbb{M} \longrightarrow \mathbb{T}$. The set $\widehat{\mathbb{M}}$ of all characters is a group under pointwise multiplication, called the *dual group* of $\mathbb{M}$. If $f \in \mathbf{L}^2_\mu$ and $\chi \in \widehat{\mathbb{M}}$, then f is a *χ-eigenfunction* of $(\mathcal{A}^{\mathbb{M}}, \mu; \sigma)$ if $\sigma^{\mathsf{m}}_*(f) = \chi(\mathsf{m}) \cdot f$ for all $\mathsf{m} \in \mathbb{M}$; then χ is called a *eigencharacter*. The *spectral group* of $(\mathcal{A}^{\mathbb{M}}, \mu; \sigma)$ is then the subgroup $\mathbb{S}_\sigma \subset \widehat{\mathbb{M}}$ of all eigencharacters. For any $\chi \in \mathbb{S}_\sigma$, let $\mathbf{E}_\chi(\sigma)$ be the corresponding eigenspace, and let $\mathbf{E}(\sigma) := \bigsqcup_{\chi \in \mathbb{S}_\Phi} \mathbf{E}_\chi(\sigma)$.

- $(\mathcal{A}^{\mathbb{M}}, \mu; \sigma)$ is ergodic $\iff$ $\mathbf{E}_1(\sigma)$ contains only constant functions $\iff$ $\dim[\mathbf{E}_\chi(\sigma)] = 1$ for all $\chi \in \mathbb{S}_\sigma$.
- $(\mathcal{A}^{\mathbb{M}}, \mu; \sigma)$ is weakly mixing $\iff$ $\mathbf{E}(\sigma)$ contains only constant functions $\iff$ $(\mathcal{A}^{\mathbb{M}}, \mu; \sigma)$ is ergodic and $\mathbb{S}_\sigma = \{\mathbf{1}\}$.

$(\mathcal{A}^{\mathbb{M}}, \mu; \sigma)$ has *discrete spectrum* if $\mathbf{L}^2_\mu$ is spanned by $\mathbf{E}(\sigma)$. In this case, the system $(\mathcal{A}^{\mathbb{M}}, \mu; \sigma)$ is measurably isomorphic to an action of $\mathbb{M}$ by translations on a compact abelian group.

Example 65 Let $\mathbb{M} = \mathbb{Z}$; then any character $\chi : \mathbb{Z} \longrightarrow \mathbb{T}$ has the form $\chi(n) = c^n$ for some $c \in \mathbb{T}$, so a χ-eigenfunction is just a eigenfunction with eigenvalue c. In this case, the aforementioned spectral properties for the $\mathbb{Z}$-action by shifts are equivalent to the corresponding spectral properties of the CA $\Phi = \sigma^1$. Bernoulli measures and irreducible Markov chains are weakly mixing. On other hand, several important classes of symbolic dynamical systems have discrete spectrum, including Sturmian shifts, constant-length substitution shifts, and regular Toeplitz shifts; see ▶ Symbolic Dynamics and ▶ Dynamics of Cellular Automata in Non-compact Spaces.

Proposition 66 *Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})$, and let $\mu \in \mathfrak{M}_{\mathrm{erg}}(\mathcal{A}^{\mathbb{M}}; \Phi, \sigma)$ be σ-ergodic.*

(a) $\mathbf{E}(\sigma) \subseteq \mathbf{E}(\Phi)$.

(b) *If $(\mathcal{A}^{\mathbb{M}}, \mu; \sigma)$ has discrete spectrum, then so does $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$.*

(c) *Suppose μ is Φ-ergodic. If $(\mathcal{A}^{\mathbb{M}}, \mu; \sigma)$ is weakly mixing, then so is $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$.*

Proof **(a)** Suppose $\chi \in \widehat{\mathbb{M}}$ and $f \in \mathbf{E}_\chi$. Then $f \circ \Phi \in \mathbf{E}_\chi$ also, because for all $\mathsf{m} \in \mathbb{M}$, $f \circ \Phi \circ \sigma^{\mathsf{m}} = f \circ \sigma^{\mathsf{m}} \circ \Phi = \chi(\mathsf{m}) \cdot f \circ \Phi$. But if $(\mathcal{A}^{\mathbb{M}}, \mu; \sigma)$ is ergodic, then $\dim[\mathbf{E}_\chi(\sigma)] = 1$; hence $f \circ \Phi$ must be a scalar multiple of f. Thus, f is also an eigenfunction for Φ. **(b)** follows from (a).

(c) By reversing the roles of Φ and σ in (a), we see that $\mathbf{E}(\Phi) \subseteq \mathbf{E}(\sigma)$. But if $(\mathcal{A}^{\mathbb{M}}, \mu; \sigma)$ is weakly mixing, then $\mathbf{E}(\sigma) = \{\text{constant functions}\}$. Thus, $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is also weakly mixing. □

Example 67 (a) Let μ be any Bernoulli measure on $\mathcal{A}^{\mathbb{M}}$. If μ is Φ-invariant and Φ-ergodic, then $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is weakly mixing (because $(\mathcal{A}^{\mathbb{M}}, \mu; \sigma)$ is weakly mixing).

(b) Let $P \in \mathbb{N}$ and suppose μ is a Φ-invariant measure supported on the set $\mathbf{X}_P$ of P-periodic sequences (see Proposition 10). Then $(\mathcal{A}^{\mathbb{Z}}, \mu; \sigma)$ has discrete spectrum (with rational eigenvalues). But $\mathbf{X}_P$ is finite, so the system $(\mathbf{X}_P, \Phi)$ is also periodic; hence $(\mathcal{A}^{\mathbb{Z}}, \mu; \Phi)$ also has discrete spectrum (with rational eigenvalues).

(c) Downarowicz [25] has constructed an example of a regular Toeplitz shift $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}}$ and $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ (not the shift) such that $\Phi(\mathbf{X}) \subseteq \mathbf{X}$. Any regular Toeplitz shift is uniquely ergodic, and the unique shift-invariant measure μ has discrete spectrum; thus, $(\mathcal{A}^{\mathbb{Z}}, \mu; \Phi)$ also has discrete spectrum.

Aside from Examples 67(b,c), the literature contains no examples of discrete-spectrum, invariant measures for CA; this is an interesting area for future research.

Entropy

Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{M}})$. For any finite $\mathbb{B} \subset \mathbb{M}$, let $\mathcal{B} := \mathcal{A}^{\mathbb{B}}$, let $\Phi^{\mathbb{N}}_{\mathbb{B}} : \mathcal{A}^{\mathbb{N}} \longrightarrow \mathcal{B}^{\mathbb{N}}$ be as in Eq. (2), and let $\mathbf{X} := \Phi^{\mathbb{N}}_{\mathbb{B}}(\mathcal{A}^{\mathbb{M}}) \subseteq \mathcal{A}^{\mathbb{N}}$; then define

$$H_{\mathrm{top}}(\mathbb{B}; \Phi) := h_{\mathrm{top}}(\mathbf{X}) = \lim_{T \to \infty} \frac{1}{T} \log_2(\#\mathbf{X}_{[0 \ldots T)}) .$$

If $\mu \in \mathfrak{M}_{\text{erg}}(\mathcal{A}^{\mathbb{M}}, \Phi)$, let $\nu := \Phi_{\mathbb{B}}^{\mathbb{N}}(\mu)$; then ν is a σ-invariant measure on $\mathcal{B}^{\mathbb{N}}$. Define

$$H_\mu(\mathbb{B}; \Phi) := h_\nu(\sigma) = -\lim_{T\to\infty} \frac{1}{T} \sum_{\mathbf{b}\in\mathcal{B}^{[0\ldots T)}} \nu[\mathbf{b}] \log_2(\nu[\mathbf{b}]).$$

The *topological entropy* of $(\mathcal{A}^{\mathbb{M}}, \Phi)$ and the *measurable entropy* of $(\mathcal{A}^{\mathbb{M}}, \Phi, \mu)$ are then defined

$$h_{\text{top}}(\Phi) := \sup_{\substack{\mathbb{B}\subset\mathbb{M}\\ \text{finite}}} H_{\text{top}}(\mathbb{B}; \Phi) \quad \text{and}$$

$$h_\mu(\Phi) := \sup_{\substack{\mathbb{B}\subset\mathbb{M}\\ \text{finite}}} H_\mu(\mathbb{B}; \Phi)\,.$$

The famous *Variational Principle* states that $h_{\text{top}}(\Phi) = \sup\{h_\mu(\Phi); \mu \in \mathfrak{M}_{\text{erg}}(\mathcal{A}^{\mathbb{M}}; \Phi)\}$; see ► Entropy in Ergodic Theory or Sect. 10 ► Topological Dynamics of Cellular Automata.

If $\mathbb{M}$ has more than one dimension (e. g. $\mathbb{M} = \mathbb{Z}^D$ or $\mathbb{N}^D$ for $D \geq 2$) then most CA on $\mathcal{A}^{\mathbb{M}}$ have infinite entropy. Thus, entropy is mainly of interest in the case $\mathbb{M} = \mathbb{Z}$ or $\mathbb{N}$. Coven [23] was the first to compute the topological entropy of a CA; he showed that $h_{\text{top}}(\Phi) = 1$ for a large class of left-permutative, one-sided CA on $\{0, 1\}^{\mathbb{N}}$ (which have since been called *Coven CA*). Later, Lind [83] showed how to construct CA whose topological entropy was any element of a countable dense subset of $\mathbb{R}_+$, consisting of logarithms of certain algebraic numbers. Theorems 14 and 18(b) above characterize the topological entropy of posexpansive CA. However, Hurd et al. [62] showed that there is no algorithm which can compute the topological entropy of an arbitrary CA; see ► Tiling Problem and Undecidability in Cellular Automata.

Measurable entropy has also been computed for a few special classes of CA. For example, if $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ is bipermutative with neighborhood $\{0, 1\}$ and $\mu \in \mathfrak{M}_{\text{erg}}(\mathcal{A}^{\mathbb{Z}}, \Phi; \sigma)$ is σ-ergodic, then $h_\mu(\Phi) = \log_2(K)$ for some integer $K \leq |\mathcal{A}|$ (see Thm 4.1 in [113]). If η is the uniform measure, and Φ is posexpansive, then Theorems 58 and 60 above characterize $h_\eta(\Phi)$. Also, if Φ satisfies the conditions of Theorem 61, then $h_\eta(\Phi) > 0$, and furthermore, all factors of the MPDS $(\mathcal{A}^{\mathbb{Z}}, \mu; \Phi)$ also have positive entropy.

However, unlike abstract dynamical systems, CA come with an explicit spatial 'geometry'. The most fruitful investigations of CA entropy are those which have interpreted entropy in terms of how information propagates through this geometry.

Lyapunov Exponents

Wolfram [150] suggested that the propagation speed of 'perturbations' in a one-dimensional CA Φ could transform 'spatial' entropy [i. e. $h(\sigma)$] into 'temporal' entropy [i. e. $h(\Phi)$]. He compared this propagation speed to the 'Lyapunov exponent' of a smooth dynamical system: it determines the exponential rate of divergence between two initially close Φ-orbits (see pp. 172, 261 and 514 in [151]). Shereshevsky [126] formalized Wolfram's intuition and proved the conjectured entropy relationship; his results were later improved by Tisseur [141]. Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$, let $\mathbf{a} \in \mathcal{A}^{\mathbb{Z}}$, and let $\mathsf{z} \in \mathbb{Z}$. Define

$$\mathbf{W}_{\mathsf{z}}^+(\mathbf{a}) := \left\{\mathbf{w} \in \mathcal{A}^{\mathbb{Z}}\,;\ \mathbf{w}_{[z\ldots\infty)} = \mathbf{a}_{[z\ldots\infty)}\right\}, \quad \text{and}$$

$$\mathbf{W}_{\mathsf{z}}^-(\mathbf{a}) := \left\{\mathbf{w} \in \mathcal{A}^{\mathbb{Z}}\,;\ \mathbf{w}_{(-\infty\ldots z]} = \mathbf{a}_{(-\infty\ldots z]}\right\}.$$

Thus, we obtain each $\mathbf{w} \in \mathbf{W}_{\mathsf{z}}^+(\mathbf{a})$ (respectively $\mathbf{W}_{\mathsf{z}}^-(\mathbf{a})$) by 'perturbing' $\mathbf{a}$ somewhere to the left (resp. right) of coordinate z. Next, for any $t \in \mathbb{N}$, define

$$\widetilde{\Lambda}_t^+(\mathbf{a}) := \min\left\{\mathsf{z} \in \mathbb{N}\,;\ \Phi^t\left[\mathbf{W}_0^+(\mathbf{a})\right] \subseteq \mathbf{W}_{\mathsf{z}}^+\left(\Phi^t[\mathbf{a}]\right)\right\},$$

and

$$\widetilde{\Lambda}_t^-(\mathbf{a}) := \min\left\{\mathsf{z} \in \mathbb{N}\,;\ \Phi^t\left[\mathbf{W}_0^-(\mathbf{a})\right] \subseteq \mathbf{W}_{-\mathsf{z}}^-\left(\Phi^t[\mathbf{a}]\right)\right\}.$$

Thus, $\widetilde{\Lambda}_t^\pm$ measures the farthest distance which any perturbation of $\mathbf{a}$ at coordinate 0 could have propagated by time t. Next, define $\Lambda_t^\pm(\mathbf{a}) := \max_{\mathsf{z}\in\mathbb{Z}} \widetilde{\Lambda}_t^\pm[\sigma^z(\mathbf{a})]$. Then Shereshevsky [126] defined the *(maximum) Lyapunov exponents*

$$\lambda^+(\Phi, \mathbf{a}) := \lim_{t\to\infty} \frac{1}{t}\Lambda_t^+(\mathbf{a})\,, \quad \text{and}$$

$$\lambda^-(\Phi, \mathbf{a}) := \lim_{t\to\infty} \frac{1}{t}\Lambda_t^-(\mathbf{a})\,,$$

whenever these limits exist. Let $\mathbf{G}(\Phi) := \{\mathbf{g} \in \mathcal{A}^{\mathbb{Z}}; \lambda^\pm(\Phi, \mathbf{g})$ both exist$\}$. The subset $\mathbf{G}(\Phi)$ is 'generic' within $\mathcal{A}^{\mathbb{Z}}$ in a very strong sense, and the Lyapunov exponents detect 'chaotic' topological dynamics.

Proposition 68 *Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$.*

(a) *Let $\mu \in \mathfrak{M}_{\text{erg}}(\mathcal{A}^{\mathbb{Z}}; \sigma)$. Suppose that either:* [i] *μ is also Φ-invariant; or:* [ii] *μ is σ-ergodic and $\mathsf{supp}(\mu)$ is a Φ-invariant subset. Then $\mu(\mathbf{G}) = 1$.*

(b) *The set $\mathbf{G}$ and the functions $\lambda^\pm(\Phi, \bullet)$ are (Φ, σ)-invariant. Thus, if μ is either Φ-ergodic or σ-ergodic, then there exist constants $\lambda_\mu^\pm(\Phi) \geq 0$ such that $\lambda^\pm(\Phi, \mathbf{g}) = \lambda_\mu^\pm(\Phi)$ for μ- all $\mathbf{g} \in \mathbf{G}$.*

(c) *If Φ is posexpansive, then there is a constant $c > 0$ such that $\lambda^\pm(\Phi, \mathbf{g}) \geq c$ for all $\mathbf{g} \in \mathbf{G}$.*

(d) *Let η be the uniform Bernoulli measure. If Φ is surjective, then $h_{\text{top}}(\Phi) \leq \left(\lambda_\eta^+(\Phi) + \lambda_\eta^-(\Phi)\right) \cdot \log|\mathcal{A}|$.*

Proof **(a)** follows from the fact that, for any $\mathbf{a} \in \mathcal{A}^{\mathbb{Z}}$, the sequence $[\Lambda_t^\pm(\mathbf{a})]_{t\in\mathbb{N}}$ is subadditive in t. Condition [i]

is Theorem 1 in [126], and follows from Kingman's subadditive ergodic theorem. Condition [ii] is Proposition 3.1 in [141].

(b) is clear by definition of $\lambda^{\pm}$. **(c)** is Theorem 5.2 in [37]. **(d)** is Proposition 5.3 in [141]. □

For any Φ-ergodic $\mu \in \mathfrak{M}_{\text{ergs}}(\mathcal{A}^{\mathbb{Z}}; \Phi, \sigma)$, Shereshevsky (see Theorem 2 in [126]) showed that $h_\mu(\Phi) \leq \left(\lambda^+_\mu(\Phi) + \lambda^-_\mu(\Phi)\right) \cdot h_\mu(\sigma)$. Tisseur later improved this estimate. For any $T \in \mathbb{N}$, let

$$\begin{aligned}\widetilde{I}^+_T(\mathbf{a}) := \min\big\{\mathsf{z} \in \mathbb{N}\,;\ \forall\, t \in [1\ldots T]\,,\\ \Phi^t\left[\mathbf{W}^+_{-\mathsf{z}}(\mathbf{a})\right] \subseteq \mathbf{W}^+_0\left(\Phi^t[\mathbf{a}]\right)\big\} \quad \text{and}\\ \widetilde{I}^-_T(\mathbf{a}) := \min\big\{\mathsf{z} \in \mathbb{N}\,;\ \forall\, t \in [1\ldots T]\,,\\ \Phi^t\left[\mathbf{W}^-_{\mathsf{z}}(\mathbf{a})\right] \subseteq \mathbf{W}^-_0\left(\Phi^t[\mathbf{a}]\right)\big\}\,.\end{aligned}$$

Next, for any $\mu \in \mathfrak{M}_{\text{ergs}}(\mathcal{A}^{\mathbb{Z}}, \sigma)$, define $\widehat{I}^{\pm}_T(\mu) := \int_{\mathcal{A}^{\mathbb{Z}}} \widetilde{I}^{\pm}_T(\mathbf{a})\,\mathrm{d}\mu[\mathbf{a}]$.
Tisseur then defined the *average Lyapunov exponents*: $I^{\pm}_\mu(\Phi) := \liminf_{T\to\infty} \widehat{I}^{\pm}_T(\mu)/T$.

Theorem 69 *Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ and let $\mu \in \mathfrak{M}_{\text{ergs}}(\mathcal{A}^{\mathbb{Z}}; \sigma)$.*

(a) *If supp(μ) is Φ-invariant, then $I^+_\mu(\Phi) \leq \lambda^+_\mu(\Phi)$ and $I^-_\mu(\Phi) \leq \lambda^-_\mu(\Phi)$, and one or both inequalities are sometimes strict.*

(b) *If μ is σ-ergodic and Φ-invariant, then $h_\mu(\Phi) \leq \left(I^+_\mu(\Phi) + I^-_\mu(\Phi)\right) \cdot h_\mu(\sigma)$, and this inequality is sometimes strict.*

(c) *If supp(μ) contains Φ-equicontinuous points, then $I^+_\mu(\Phi) = I^-_\mu(\Phi) = h_\mu(\Phi) = 0$.*

Proof See [141]: **(a)** is Proposition 3.2 and Example 6.1; **(b)** is Theorem 5.1 and Example 6.2; and **(c)** is Proposition 5.2. □

Directional Entropy

Milnor [98,99] introduced directional entropy to capture the intuition that information in a CA propagates in particular directions with particular 'velocities', and that different CA 'mix' information in different ways. Classical entropy is unable to detect this informational anisotropy. For example, if $\mathcal{A} = \{0, 1\}$ and $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ has local rule $\phi(a_0, a_1) = a_0 + a_1 \pmod 2$, then $h_{\text{top}}(\Phi) = 1 = h_{\text{top}}(\sigma)$, despite the fact that Φ vigorously 'mixes' information together and propagates any 'perturbation' outwards in an expanding cone, whereas σ merely shifts information to the left in a rigid and essentially trivial fashion.

If $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}^D})$, then a *complete history* for Φ is a sequence $(\mathbf{a}_t)_{t\in\mathbb{Z}} \in (\mathcal{A}^{\mathbb{Z}^D})^{\mathbb{Z}} \cong \mathcal{A}^{\mathbb{Z}^{D+1}}$ such that $\Phi(\mathbf{a}_t) = \mathbf{a}_{t+1}$ for all $t \in \mathbb{Z}$. Let $\mathbf{X}^{\text{Hist}} := \mathbf{X}^{\text{Hist}}(\Phi) \subset \mathcal{A}^{\mathbb{Z}^{D+1}}$ be the subshift of all complete histories for Φ, and let σ be the $\mathbb{Z}^{D+1}$ shift action on $\mathbf{X}^{\text{Hist}}$; then $(\mathbf{X}^{\text{Hist}}; \sigma)$ is conjugate to the natural extension of the system $(\mathbf{Y}; \Phi, \sigma)$, where $\mathbf{Y} := \Phi^\infty(\mathcal{A}^{\mathbb{M}}) := \bigcap_{t=1}^\infty \Phi^t(\mathcal{A}^{\mathbb{Z}^D})$ is the omega-limit set of Φ. If $\mu \in \mathfrak{M}_{\text{ergs}}(\mathcal{A}^{\mathbb{Z}^D}; \Phi, \sigma)$, then $\mathsf{supp}(\mu) \subseteq \mathbf{Y}$, and μ extends to a σ-invariant measure $\tilde{\mu}$ on $\mathbf{X}^{\text{Hist}}$ in the obvious way.

Let $\vec{\mathrm{v}} = (v_0; v_1, \ldots, v_D) \in \mathbb{R} \times \mathbb{R}^D \cong \mathbb{R}^{D+1}$. For any bounded open subset $\mathrm{B} \subset \mathbb{R}^{D+1}$ and $T > 0$, let $\mathrm{B}(T\vec{\mathrm{v}}) := \{\mathrm{b} + t\vec{\mathrm{v}}; \mathrm{b} \in \mathrm{B} \text{ and } t \in [0, T]\}$ be the 'sheared cylinder' in $\mathbb{R}^{D+1}$ with cross-section B and length $T|\vec{\mathrm{v}}|$ in the direction $\vec{\mathrm{v}}$, and let $\mathbb{B}(T\vec{\mathrm{v}}) := \mathrm{B}(T\vec{\mathrm{v}}) \cap \mathbb{Z}^{D+1}$. Let $\mathbf{X}^{\text{Hist}}{}_{\mathbb{B}(T\vec{\mathrm{v}})} := \left\{\mathbf{x}_{\mathbb{B}(T\vec{\mathrm{v}})}\,;\ \mathbf{x} \in \mathbf{X}^{\text{Hist}}(\Phi)\right\}$. We define

$$H_{\text{top}}(\Phi; \mathrm{B}, \vec{\mathrm{v}}) := \limsup_{T\to\infty} \frac{1}{T} \log_2[\#\mathbf{X}^{\text{Hist}}{}_{\mathbb{B}(T\vec{\mathrm{v}})}]\,; \quad \text{and}$$

$$H_\mu(\Phi; \mathrm{B}, \vec{\mathrm{v}}) := -\limsup_{T\to\infty} \frac{1}{T} \sum_{\mathbf{x}\in\mathbf{X}^{\text{Hist}}{}_{\mathbb{B}(T\vec{\mathrm{v}})}} \tilde{\mu}[\mathbf{x}] \log_2(\tilde{\mu}[\mathbf{x}])\,.$$

We then define the *$\vec{\mathrm{v}}$-directional topological entropy* and *$\vec{\mathrm{v}}$-directional μ-entropy* of Φ by

$$h_{\text{top}}(\Phi; \vec{\mathrm{v}}) := \sup_{\substack{\mathrm{B}\subset\mathbb{R}^{D+1}\\ \text{open \& bounded}}} h_{\text{top}}(\Phi; \mathrm{B}, \vec{\mathrm{v}})\,; \quad \text{and} \tag{8}$$

$$h_\mu(\Phi; \vec{\mathrm{v}}) := \sup_{\substack{\mathrm{B}\subset\mathbb{R}^{D+1}\\ \text{open \& bounded}}} h_\mu(\Phi; \mathrm{B}, \vec{\mathrm{v}})\,. \tag{9}$$

Proposition 70 *Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}^D})$ and let $\mu \in \mathfrak{M}_{\text{ergs}}(\mathcal{A}^{\mathbb{Z}^D}; \Phi, \sigma)$.*

(a) *Directional entropy is homogeneous. That is, for any $\vec{\mathrm{v}} \in \mathbb{R}^{D+1}$ and $r > 0$, $h_{\text{top}}(\Phi, r\vec{\mathrm{v}}) = r \cdot h_{\text{top}}(\Phi, \vec{\mathrm{v}})$ and $h_\mu(\Phi, r\vec{\mathrm{v}}) = r \cdot h_\mu(\Phi, \vec{\mathrm{v}})$.*

(b) *If $\vec{\mathrm{v}} = (t; \mathbf{z}) \in \mathbb{Z} \times \mathbb{Z}^D$, then $h_{\text{top}}(\Phi, \vec{\mathrm{v}}) = h_{\text{top}}(\Phi^t \circ \sigma^{\mathbf{z}})$ and $h_\mu(\Phi, \vec{\mathrm{v}}) = h_\mu(\Phi^t \circ \sigma\mathbf{z})$.*

(c) *There is an extension of the $\mathbb{Z} \times \mathbb{Z}^D$-system $(\mathbf{X}^{\text{Hist}}, \Phi; \sigma)$ to an $\mathbb{R} \times \mathbb{R}^D$-system $(\widetilde{\mathbf{X}}, \widetilde{\Phi}, \widetilde{\sigma})$ such that, for any $\vec{\mathrm{v}} = (t; \vec{u}) \in \mathbb{R} \times \mathbb{R}^D$ we have $h_{\text{top}}(\Phi, \vec{\mathrm{v}}) = h_{\text{top}}(\widetilde{\Phi}^t \circ \widetilde{\sigma}^{\vec{u}})$.*

For any $\mu \in \mathfrak{M}_{\text{ergs}}(\mathcal{A}^{\mathbb{Z}^D}; \Phi, \sigma)$, there is an extension $\tilde{\mu} \in \mathfrak{M}_{\text{ergs}}(\widetilde{\mathbf{X}}; \widetilde{\Phi}, \widetilde{\sigma})$ such that for any $\vec{\mathrm{v}} = (t; \vec{u}) \in \mathbb{R} \times \mathbb{R}^D$ we have $h_\mu(\Phi, \vec{\mathrm{v}}) = h_{\tilde{\mu}}(\widetilde{\Phi}^t \circ \widetilde{\sigma}^{\vec{u}})$.

Proof **(a,b)** follow from the definition. **(c)** is Proposition 2.1 in [109]. □

Remark 71 Directional entropy can actually be defined for any continuous $\mathbb{Z}^{D+1}$-action on a compact metric space, and in particular, for any subshift of $\mathcal{A}^{\mathbb{Z}^{D+1}}$. The directional entropy of a CA Φ is then just the directional

entropy of the subshift $\mathbf{X}^{\text{Hist}}(\Phi)$. Proposition 70 holds for any subshift.

Directional entropy is usually infinite for multidimensional CA (for the same reason that classical entropy is usually infinite). Thus, most of the analysis has been for one-dimensional CA. For example, Kitchens and Schmidt (see Sect. 1 in [72]) studied the directional topological entropy of one-dimensional linear CA, while Smillie (see Proposition 1.1 in [133]) computed the directional topological entropy for ECA#184. If Φ is linear, then the function $\vec{v} \mapsto h_{\text{top}}(\Phi, \vec{v})$ is piecewise linear and convex, but if Φ is ECA#184, it is neither.

If $\vec{v}$ has rational entries, then Proposition 70(a,b) shows that $h(\Phi, \vec{v})$ is a rational multiple of the classical entropy of some composite CA, which can be computed through classical methods. However, if $\vec{v}$ is irrational, then $h(\Phi, \vec{v})$ is quite difficult to compute using the formulae (8) and (9), and Proposition 70(c), while theoretically interesting, is not very computationally useful. Can we compute $h(\Phi, \vec{v})$ as the limit of $h(\Phi, \vec{v}_k)$ where $\{\vec{v}_k\}_{k=1}^{\infty}$ is a sequence of *rational* vectors tending to $\vec{v}$? In other words, is directional entropy *continuous* as a function of $\vec{v}$? What other properties has $h(\Phi, \vec{v})$ as a function of $\vec{v}$?

Theorem 72 *Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ and let $\mu \in \mathfrak{M}_{\text{cons}}(\mathcal{A}^{\mathbb{Z}}; \Phi)$.*

(a) *The function $\mathbb{R}^2 \ni \vec{v} \mapsto h_\mu(\Phi, \vec{v}) \in \mathbb{R}$ is continuous.*

(b) *Suppose there is some $(t, z) \in \mathbb{N} \times \mathbb{Z}$ with $t \geq 1$, such that $\Phi^t \circ \sigma^z$ is posexpansive. Then the function $\mathbb{R}^2 \ni \vec{v} \mapsto h_{\text{top}}(\Phi, \vec{v}) \in \mathbb{R}$ is convex, and thus, Lipschitz-continuous.*

(c) *However, there exist other $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$ for which the function $\mathbb{R}^2 \ni \vec{v} \mapsto h_{\text{top}}(\Phi, \vec{v}) \in \mathbb{R}$ is not continuous.*

(d) *Suppose Φ has neighborhood $[-\ell \ldots r] \subset \mathbb{Z}$. If $\vec{v} = (t; x) \in \mathbb{R}^2$, then let $z_\ell := x - \ell t$ and $z_r := x + rt$. Let $L := \log|\mathcal{A}|$.*

[i] Suppose $z_\ell \cdot z_r \geq 0$. Then $h_\mu(\Phi; \vec{v}) \leq \max\{|z_\ell|, |z_r|\} \cdot L$. Furthermore:

- *If Φ is right-permutative, and $|z_\ell| \leq |z_r|$, then $h_\mu(\Phi; \vec{v}) = |z_r| \cdot L$.*
- *If Φ is left-permutative, and $|z_r| \leq |z_\ell|$, then $h_\mu(\Phi; \vec{v}) = |z_r| \cdot L$.*

[ii] Suppose $z_\ell \cdot z_r \leq 0$. Then $h_\mu(\Phi; \vec{v}) \leq |z_r - z_\ell| \cdot L$. Furthermore, if Φ is bipermutative in this case, then $h_\mu(\Phi; \vec{v}) = |z_r - z_\ell| \cdot L$.

Proof **(a)** is Corollary 3.3 in [109], while **(b)** is Théorème III.11 and Corollaire III.12, pp. 79–80 in [120]. **(c)** is Proposition 1.2 in [133].

(d) summarizes the main results of [21]. See also Example 6.2 in [99] for an earlier analysis of permutative CA in the case $r = \ell = 1$; see also Example 6.4 in [12] and Sect. 1 in Sect. 1 in [72] for the special case when Φ is linear. □

Remark 73 (a) In fact, the conclusion of Theorem 72(b) holds as long as Φ has any *posexpansive directions* (even irrational ones). A posexpansive direction is analogous to an *expansive subspace* (see Subsect. "Entropy Geometry and Expansive Subdynamics"), and is part of Sablik's theory of 'directional dynamics' for one-dimensional CA; see Remark 84(b) below. Using this theory, Sablik has also shown that $h_\mu(\Phi; \vec{v}) = 0 = h_{\text{top}}(\Phi, \vec{v})$ whenever $\vec{v}$ is an *equicontinuous direction* for Φ, whereas $h_\mu(\Phi; \vec{v}) \neq 0 \neq h_{\text{top}}(\Phi, \vec{v})$ whenever $\vec{v}$ is a *right-* or *left posexpansive direction* for Φ. See Sect. §III.4.5–Sect. §III.4.6, pp. 86–88 in [120].

(b) Courbage and Kamiński have defined a 'directional' version of the Lyapunov exponents introduced in Subsect. "Lyapunov Exponents". If $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$, $\mathbf{a} \in \mathcal{A}^{\mathbb{Z}}$ and $\vec{v} = (t; z) \in \mathbb{N} \times \mathbb{Z}$, then $\lambda^{\pm}_{\vec{v}}(\Phi, \mathbf{a}) := \lambda^{\pm}(\Phi^t \circ \sigma^z, \mathbf{a})$, where $\lambda^{\pm}$ are defined as in Subsect. "Lyapunov Exponents". If $\vec{v} \in \mathbb{R}^2$ is irrational, then the definition of $\lambda^{\pm}_{\vec{v}}(\Phi, \mathbf{a})$ is somewhat more subtle. For any Φ and $\mathbf{a}$, the function $\mathbb{R}^2 \ni \vec{v} \mapsto \lambda^{\pm}_{\vec{v}}(\Phi, \mathbf{a}) \in \mathbb{R}$ is homogeneous and continuous (see Lemma 2 and Proposition 3 in [22]). If $\mu \in \mathfrak{M}_{\text{cons}}(\mathcal{A}^{\mathbb{Z}}; \Phi, \sigma)$ is σ-ergodic, then $\lambda^{\pm}_{\vec{v}}(\Phi, \bullet)$ is constant μ-almost everywhere, and is related to $h_\mu(\Phi; \vec{v})$ through an inequality exactly analogous to Theorem 69(b); see Theorem 1 in [22].

Cone Entropy For any $\vec{v} \in \mathbb{R}^{D+1}$, any angle $\theta > 0$, and any $N > 0$, we define

$$\mathbb{K}(N\vec{v}, \theta) := \Big\{ \mathsf{z} \in \mathbb{Z}^{D+1} \,;\ |\mathsf{z}| \leq N|\vec{v}| \quad \text{and} \quad \mathsf{z} \bullet \vec{v}/|\mathsf{z}||\vec{v}| \geq \cos(\theta) \Big\} .$$

Geometrically, this is the set of all $\mathbb{Z}^{D+1}$-lattice points in a cone of length $N|\vec{v}|$ which subtends an angle of 2θ around an axis parallel to $\vec{v}$, and which has its apex at the origin. If $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}^D})$, then let $\mathbf{X}^{\text{Hist}}(N\vec{v}, \theta) := \{\mathbf{x}_{\mathbb{K}(N\vec{v},\theta)} \,;\ \mathbf{x} \in \mathbf{X}^{\text{Hist}}(\Phi)\}$. If $\mu \in \mathfrak{M}_{\text{cons}}(\mathcal{A}^{\mathbb{Z}^D}; \Phi)$, and $\tilde{\mu}$ is the extension of μ to $\mathbf{X}^{\text{Hist}}$, then the *cone entropy* of (Φ, μ) in direction $\vec{v}$ is defined

$$h^{\text{cone}}_\mu(\Phi, \vec{v}) := -\lim_{\theta \searrow 0} \lim_{N \to \infty} \frac{1}{N} \sum_{\mathbf{x} \in \mathbf{X}^{\text{Hist}}(N\vec{v},\theta)} \tilde{\mu}[\mathbf{x}] \log_2(\tilde{\mu}[\mathbf{x}]) .$$

Park [107,108] attributes this concept to Doug Lind. Like directional entropy, cone entropy can be defined for any

continuous $\mathbb{Z}^{D+1}$-action, and is generally infinite for multidimensional CA. However, for one-dimensional CA, Park has proved:

Theorem 74 *If $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$, $\mu \in \mathfrak{M}_{\mathrm{eqs}}(\mathcal{A}^{\mathbb{Z}}; \Phi)$ and $\vec{v} \in \mathbb{R}^2$, then $h_\mu^{\mathrm{cone}}(\Phi, \vec{v}) = h_\mu(\Phi, \vec{v})$.*

Proof See Theorem 1 in [108]. □

Entropy Geometry and Expansive Subdynamics

Directional entropy is the one-dimensional version of a multidimensional 'entropy density' function, which was introduced by Milnor [99] to address the fact that classical and directional entropy are generally infinite for multidimensional CA. Milnor's ideas were then extended by Boyle and Lind [12], using their theory of *expansive subdynamics*.

Let $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^{D+1}}$ be a subshift, and let $\mu \in \mathfrak{M}_{\mathrm{eqs}}(\mathbf{X}; \sigma)$. For any bounded $\mathsf{B} \subset \mathbb{R}^{D+1}$, let $\mathbb{B} := \mathsf{B} \cap \mathbb{Z}^{D+1}$, let $\mathbf{X}_{\mathsf{B}} := \mathbf{X}_{\mathbb{B}}$, and then define

$$H_{\mathbf{X}}(\mathsf{B}) := \log_2 |\mathbf{X}_{\mathsf{B}}| \quad \text{and}$$
$$H_\mu(\mathsf{B}) := -\sum_{\mathbf{x} \in \mathbf{X}_{\mathsf{B}}} \mu[\mathbf{x}] \log_2(\mu[\mathbf{x}]) \,.$$

The *topological entropy dimension* $\dim(\mathbf{X})$ is the smallest $d \in [0 \ldots D+1]$ having some constant $c > 0$ such that, for any finite $\mathsf{B} \subset \mathbb{R}^{D+1}$, $H_{\mathbf{X}}(\mathsf{B}) \leq c \cdot \mathrm{diam}\,[\mathsf{B}]^d$. The *measurable entropy dimension* $\dim(\mu)$ is defined similarly, only with H_μ in place of $H_{\mathbf{X}}$. Note that $\dim(\mu) \leq \dim(\mathbf{X})$, because $H_\mu(\mathsf{B}) \leq H_{\mathbf{X}}(\mathsf{B})$ for all $\mathsf{B} \subset \mathbb{R}^{D+1}$.

For any bounded $\mathsf{B} \subset \mathbb{R}^{D+1}$ and 'scale factor' $s > 0$, let $s\mathsf{B} := \{s\mathsf{b}; \mathsf{b} \in \mathsf{B}\}$. For any radius $r > 0$, let $(s\mathsf{B})^r := \left\{\mathsf{x} \in \mathbb{R}^{D+1}\,;\; d(\mathsf{x}, s\mathsf{B}) \leq r\right\}$. Define the *d-dimensional topological entropy density* of B by

$$h_{\mathbf{X}}^d(\mathsf{B}) := \sup_{r>0} \limsup_{s \to \infty} H_{\mathbf{X}}[(s\mathsf{B})^r]/s^d \,. \tag{10}$$

Define *d-dimensional measurable entropy density* $h_\mu^d(\mathsf{B})$ similarly, only using H_μ instead of $H_{\mathbf{X}}$. Note that, for any $d < \dim(\mathbf{X})$ [respectively, $d < \dim(\mu)$], $h_{\mathbf{X}}^d(\mathsf{B})$ [resp. $h_\mu^d(\mathsf{B})$] will be infinite, whereas for for any $d > \dim(\mathbf{X})$ [resp. $d > \dim(\mu)$], $h_{\mathbf{X}}^d(\mathsf{B})$ [resp. $h_\mu^d(\mathsf{B})$] will be zero; hence $\dim(\mathbf{X})$ [resp. $\dim(\mu)$] is the unique value of d for which the function $h_{\mathbf{X}}^d$ [resp. h_μ^d] defined in Eq. (10) could be nontrivial.

Example 75 (a) If $d = D+1$, and B is a unit cube centered at the origin, then $h_{\mathbf{X}}^{D+1}(\mathsf{B})$ (resp. $h_\mu^{D+1}(\mathsf{B})$) is just the classical $(D+1)$-dimensional topological (resp. measurable) entropy of $\mathbf{X}$ (resp. μ) as a $(D+1)$-dimensional subshift (resp. random field); see ► Entropy in Ergodic Theory.

(b) However, the most important case for Milnor [99] (and us) is when $\mathbf{X} = \mathbf{X}^{\mathrm{Hist}}(\Phi)$ for some $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}^D})$. In this case, $\dim(\mu) \leq \dim(\mathbf{X}) \leq D < D+1$. In particular, if $d = 1$, then for any $\vec{v} \in \mathbb{R}^{D+1}$, if $\mathsf{B} := \left\{r\vec{v}; r \in [0,1]\right\}$, then $h_{\mathbf{X}}^1(\mathsf{B}) = h_{\mathrm{top}}(\Phi; \vec{v})$ and $h_\mu^1(\mathsf{B}) = h_\mu(\Phi; \vec{v})$ are *directional entropies* of Subsect. "Directional Entropy".

For any $d \in [0 \ldots D+1]$, let λ^d be the d-dimensional Hausdorff measure on $\mathbb{R}^{D+1}$ such that, if $\mathsf{P} \subset \mathbb{R}^{D+1}$ is any *d-plane* (i. e. a d-dimensional linear subspace of $\mathbb{R}^{D+1}$), then λ^d restricts to the d-dimensional Lebesgue measure on P.

Theorem 76 *Let $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^{D+1}}$ be a subshift, and let $\mu \in \mathfrak{M}_{\mathrm{eqs}}(\mathbf{X}; \sigma)$. Let $d = \dim(\mathbf{X})$ (or $\dim(\mu)$) and let h^d be $h_{\mathbf{X}}^d$ (or h_μ^d). Let $B, C \subset \mathbb{R}^{D+1}$ be compact sets. Then*

(a) *$h^d(B)$ is well-defined and finite.*
(b) *If $B \subseteq C$ then $h^d(B) \leq h^d(C)$.*
(c) *$h^d(B \cup C) \leq h^d(B) + h^d(C)$.*
(d) *$h^d(B + \vec{v}) = h^d(B)$ for any $\vec{v} \in \mathbb{R}^{D+1}$.*
(e) *$h^d(sB) = s^d \cdot h^d(B)$ for any $s > 0$.*
(f) *There is some constant c such that $h_d(B) \leq c\lambda^d(B)$ for all compact $B \subset \mathbb{R}^{D+1}$.*
(g) *If $d \in \mathbb{N}$, then for any d-plane $P \subset \mathbb{R}^{D+1}$, there is some $\mathfrak{H}^d(P) \geq 0$ such that $h^d(B) = \mathfrak{H}^d(P) \cdot \lambda^d(B)$ for any compact subset $B \subset P$ with $\lambda^d(\partial B) = 0$.*
(h) *There is a constant $\overline{H}_{\mathbf{X}}^d < \infty$ such that $\mathfrak{H}_{\mathbf{X}}^d(P) \leq \overline{H}_{\mathbf{X}}^d$ for all d-planes P.*

Proof See Theorems 1 and 2 and Corollary 1 in [99], or see Theorems 6.2, 6.3, and 6.13 in [12]. □

Example 77 Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}^D})$ and let $\mathbf{X} := \mathbf{X}^{\mathrm{Hist}}(\Phi)$. If $\mathsf{P} := \{0\} \times \mathbb{R}^D$, then $\mathfrak{H}^D(\mathsf{P})$ is the classical D-dimensional entropy of the omega limit set $\mathbf{Y} := \Phi^\infty(\mathcal{A}^{\mathbb{Z}^D})$; heuristically, this measures the asymptotic level of 'spatial disorder' in $\mathbf{Y}$. If $\mathsf{P} \subset \mathbb{R}^{D+1}$ is some other D-plane, then $\mathfrak{H}^D(\mathsf{P})$ measures some combination of the 'spatial disorder' of $\mathbf{Y}$ with the dynamical entropy of Φ.

Let $d \in [1 \ldots D+1]$, and let $\mathsf{P} \subset \mathbb{R}^{D+1}$ be a d-plane. For any $r > 0$, let $\mathsf{P}(r) := \{\mathsf{z} \in \mathbb{Z}^{D+1}; d(\mathsf{z}, \mathsf{P}) < r\}$. We say P is *expansive* for $\mathbf{X}$ if there is some $r > 0$ such that, for any $\mathbf{x}, \mathbf{y} \in \mathbf{X}$, $(\mathbf{x}_{\mathsf{P}(r)} = \mathbf{y}_{\mathsf{P}(r)}) \iff (\mathbf{x} = \mathbf{y})$. If P is spanned by d rational vectors, then $\mathsf{P} \cap \mathbb{Z}^{D+1}$ is a rank-d sublattice $\mathbb{L} \subset \mathbb{Z}^{D+1}$, and P is expansive if and only if the induced $\mathbb{L}$-action on $\mathbf{X}$ is expansive. However, if P is 'irrational', then expansiveness is a more subtle concept; see Sect. 2 in [12] for more information.

If $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}^D})$ and $\mathbf{X} = \mathbf{X}^{\mathrm{Hist}}(\Phi)$, then Φ is *quasi-invertible* if $\mathbf{X}$ admits an expansive D-plane P (this is a natural extension of Milnor's (1988; §7) definition in terms of

'causal cones'). Heuristically, if we regard $\mathbb{Z}^{D+1}$ as 'space-time' (in the spirit of special relativity), then $\mathbb{P}$ can be seen as 'space', and any direction transversal to $\mathbb{P}$ can be interpreted as the flow of 'time'.

Example 78 (a) If Φ is invertible, then it is quasi-invertible, because $\{0\} \times \mathbb{R}^D$ is an expansive D-plane (recall that the zeroth coordinate is time).
(b) Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}})$, so that $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^2}$. Let Φ have neighborhood $[-\ell \ldots r]$, with $-\ell \leq 0 \leq r$, and let $\mathbb{L} \subset \mathbb{R}^2$ be a line with slope S through the origin (Fig. 2).

[i] If Φ is right-permutative, and $0 < S \leq \frac{1}{\ell+1}$, then $\mathbb{L}$ is expansive for $\mathbf{X}$.
[ii] If Φ is left-permutative, and $\frac{-1}{r+1} \leq S < 0$, then $\mathbb{L}$ is expansive for $\mathbf{X}$.
[iii] If Φ is bipermutative, and $\frac{-1}{r+1} \leq S < 0$ or $0 < S \leq \frac{1}{\ell+1}$, then $\mathbb{L}$ is expansive for $\mathbf{X}$.
[iv] If Φ is posexpansive (see Subsect. "Posexpansive and Permutative CA") then the 'time' axis $\mathbb{L} = \mathbb{R} \times \{0\}$ is expansive for $\mathbf{X}$.

Hence, in any of these cases, Φ is quasi-invertible. (Presumably, something similar is true for multidimensional permutative CA.)

Proposition 79 *Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}^D})$, let $\mathbf{X} = \mathbf{X}^{\mathrm{Hist}}(\Phi)$, let $\mu \in \mathfrak{M}_{\mathrm{erg}}(\mathbf{X};\sigma)$, and let $\mathfrak{H}^d$ and $\overline{H}^d_{\mathbf{X}}$ be as in Theorem 76(g,h).*
(a) *If $\mathfrak{H}^D_{\mathbf{X}}(\{0\} \times \mathbb{R}^D) = 0$, then $\overline{H}^D_{\mathbf{X}} = 0$.*
(b) *Let $d \in [1 \ldots D]$, and suppose that $\mathbf{X}$ admits an expansive d-plane. Then:*
[i] $\dim(\mathbf{X}) \leq d$;
[ii] There is a constant $\overline{H}^d_\mu < \infty$ such that $\mathfrak{H}^d_\mu(P) \leq \overline{H}^d_\mu$ for all d-planes P;
[iii] If $\mathfrak{H}^d(P) = 0$ for some expansive d-plane P, then $\overline{H}^d = 0$.

Proof **(a)** is Corollary 3 in [99], **(b)[i]** is Corollary 1.4 in [128], and **(b)[ii]** is Theorem 6.19(2) in [12].

(b)[iii]: See Theorem 6.3(4) in [12] for "$\overline{H}^d_{\mathbf{X}} = 0$". See Theorem 6.19(1) in [12] for "$\overline{H}^d_\mu = 0$". □

If $d \in [1 \ldots D+1]$, then a *d-frame* in $\mathbb{R}^{D+1}$ is a d-tuple $\mathsf{F} := (\vec{v}_1, \ldots, \vec{v}_d)$, where $\vec{v}_1, \ldots, \vec{v}_d \in \mathbb{R}^{D+1}$ are linearly independent. Let $\mathfrak{Frame}(D+1, d)$ be the set of all d-frames in $\mathbb{R}^{D+1}$; then $\mathfrak{Frame}(D+1, d)$ is an open subset of $\mathbb{R}^{D+1} \times \cdots \times \mathbb{R}^{D+1} := \mathbb{R}^{(D+1)} \times d]$. Let

$$\mathfrak{E}_{\mathfrak{xpans}}(\mathbf{X}, d) := \{\mathsf{F} \in \mathfrak{Frame}(D+1, d)\,; \quad \mathrm{span}(\mathsf{F}) \text{ is expansive for } \mathbf{X}\}\,.$$

Then $\mathfrak{E}_{\mathfrak{xpans}}(\mathbf{X}, d)$ is an open subset of $\mathfrak{Frame}(D+1, d)$, by Lemma 3.4 in [12]. A connected component of $\mathfrak{E}_{\mathfrak{xpans}}(\mathbf{X}, d)$ is called an *expansive component* for $\mathbf{X}$. For any $\mathsf{F} \in \mathfrak{Frame}(D+1, d)$, let $[\mathsf{F}]$ be the d-dimensional parallelepiped spanned by F, and let $\mathfrak{h}^d_{\mathbf{X}}(\mathsf{F}) := h^d_{\mathbf{X}}([\mathsf{F}]) = \mathfrak{H}^d_{\mathbf{X}}(\mathrm{span}(\mathsf{F})) \cdot \lambda^d([\mathsf{F}])$, where the last equality is by Theorem 76(g). The next result is a partial extension of Theorem 72(b).

Proposition 80 *Let $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^{D+1}}$ be a subshift, suppose $d := \dim(\mathbf{X}) \in \mathbb{N}$, and let $\mathfrak{C} \subset \mathfrak{E}_{\mathfrak{xpans}}(\mathbf{X}, d)$ be an expan-*

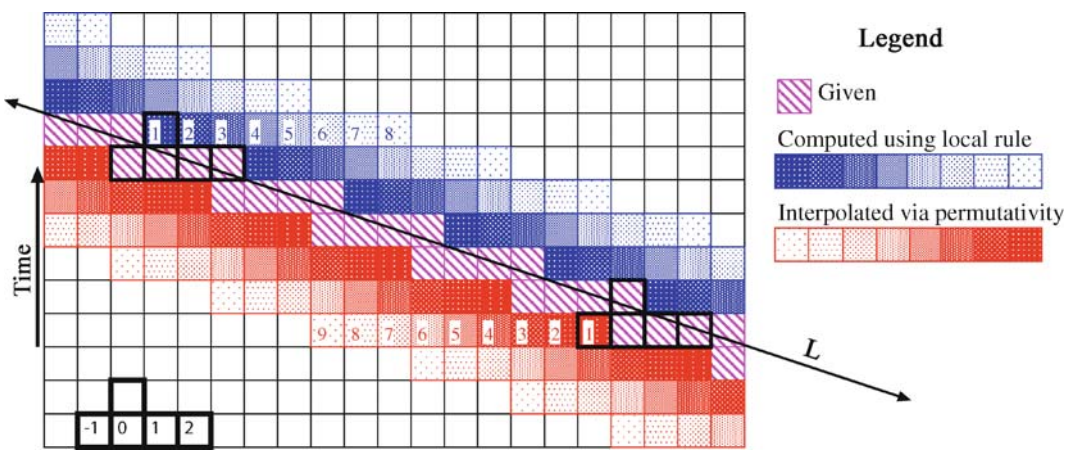

Ergodic Theory of Cellular Automata, Figure 2
Example 78(b)[ii]: A left permutative CA Φ is quasi-invertible. In this picture, $[-\ell \ldots r] = [-1 \ldots 2]$, and $\mathbb{L}$ is a line of slope −1/3. If x ∈ X and we know the entries of x in a neighborhood of $\mathbb{L}$, then we can reconstruct the rest of x as shown. Entries above $\mathbb{L}$ are directly computed using the local rule of Φ. Entries below $\mathbb{L}$ are interpolated via left-permutativity. In both cases, the reconstruction occurs in consecutive diagonal lines, whose order is indicated by shading from darkest to lightest in the figure

sive component. Then the function $\mathfrak{h}^d_{\mathbf{X}} : \mathfrak{C} \longrightarrow \mathbb{R}$ *is convex in each of its d distinct* $\mathbb{R}^{D+1}$*-valued arguments. Thus,* $\mathfrak{h}^d_{\mathbf{X}}$ *is Lipschitz-continuous on* $\mathfrak{C}$.

Proof See Theorem 6.9(1,4) in [12]. □

For measurable entropy, we can say much more. Recall that a *d-linear form* is a function $\omega : \mathbb{R}^{(D+1)\times d} \longrightarrow \mathbb{R}$ which is linear in each of its d distinct $\mathbb{R}^{D+1}$-valued arguments and antisymmetric.

Theorem 81 *Let* $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^{D+1}}$ *be a subshift and let* $\mu \in \mathfrak{M}_{\mathrm{ers}}(\mathbf{X};\sigma)$. *Suppose* $d := \dim(\mu) \in \mathbb{N}$, *and let* $\mathfrak{C} \subset \mathfrak{E}_{\mathrm{xpns}}(\mathbf{X}, d)$ *be an expansive component for* $\mathbf{X}$. *Then there is a d-linear form* $\omega : \mathbb{R}^{(D+1)\times d} \longrightarrow \mathbb{R}$ *such that* $\mathfrak{h}^d_\mu$ *agrees with* ω *on* $\mathfrak{C}$.

Proof Theorem 6.16 in [12]. □

If $\overline{H}^d_\mu \neq 0$, then Theorem 81 means that there is an orthogonal $(D+1-d)$-frame $\mathbb{W} := (\vec{\mathrm{w}}_{d+1}, \ldots, \vec{\mathrm{w}}_{D+1})$ (transversal to all frames in $\mathfrak{C}$) such that, for any d-frame $\mathbb{V} := (\vec{\mathrm{v}}_1, \ldots, \vec{\mathrm{v}}_d) \in \mathfrak{C}$,

$$\mathfrak{h}^d_\mu(\mathbb{V}) = \det(\vec{\mathrm{v}}_1, \ldots, \vec{\mathrm{v}}_d; \vec{\mathrm{w}}_{d+1}, \ldots, \vec{\mathrm{w}}_{D+1}) \,. \tag{11}$$

Thus, the d-plane orthogonal to $\{\vec{\mathrm{w}}_{d+1}, \ldots, \vec{\mathrm{w}}_{D+1}\}$ is the d-plane which maximizes $\mathfrak{S}^d_\mu$ – this is the d-plane manifesting the most rapid decay of correlation with distance. On the other hand, $\mathrm{span}(\mathbb{W})$ is the $(D+1-d)$-plane along which correlations decay the most slowly. Also, if $\mathbb{V} \in \mathfrak{C}$, then Eq. (11) implies that $\mathfrak{C}$ cannot contain any frame spanning $\mathrm{span}(\mathbb{V})$ with reversed orientation (e. g. an odd permutation of $\mathbb{V}$), because entropy is nonnegative.

Example 82 Let $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}^D})$ be quasi-invertible, and let $\mathbb{P}$ be an expansive D-plane for $\mathbf{X} := \mathbf{X}^{\mathrm{Hist}}(\Phi)$ (see Example 78). The D-frames spanning $\mathbb{P}$ fall into two expansive components (related by orientation-reversal); let $\mathfrak{C}$ be union of these two components. Let $\mu \in \mathfrak{M}_{\mathrm{ers}}(\mathcal{A}^{\mathbb{Z}^D}; \Phi)$, and extend μ to a σ-invariant measure on $\mathbf{X}$. In this case, Theorem 81 is equivalent to Theorem 4 in [99], which says there a vector $\vec{\mathrm{w}} \in \mathbb{R}^{D+1}$ such that, for any D-frame $(\vec{\mathrm{v}}_1, \ldots, \vec{\mathrm{v}}_D) \in \mathfrak{C}$, $\mathfrak{h}^d_\mu(\mathsf{F}) = \left|\det(\vec{\mathrm{v}}_1, \ldots, \vec{\mathrm{v}}_D; \vec{\mathrm{w}})\right|$. Thus, $\mathfrak{S}^d_\mu(\mathbb{P})$ is maximized when $\mathbb{P}$ is the hyperplane orthogonal to $\vec{\mathrm{w}}$. Heuristically, $\vec{\mathrm{w}}$ points in the direction of minimum correlation decay (or maximum 'causality') – the direction which could most properly be called 'time' for the MPDS (Φ, μ).

Theorem 81 yields the following generalization the Variational Principle:

Theorem 83 *Let* $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^{D+1}}$ *be a subshift and suppose* $d := \dim(\mathbf{X}) \in \mathbb{N}$.

(a) *If* $\mathsf{F} \in \mathfrak{E}_{\mathrm{xpns}}(\mathbf{X}, d)$, *then there exists* $\mu \in \mathfrak{M}_{\mathrm{ers}}(\mathbf{X};\sigma)$ *such that* $\mathfrak{h}^d_{\mathbf{X}}(\mathsf{F}) = \mathfrak{h}^d_\mu(\mathsf{F})$.

(b) *Let* $\mathfrak{C} \subset \mathfrak{E}_{\mathrm{xpns}}(\mathbf{X}, d)$ *be an expansive component for* $\mathbf{X}$. *There exists some* $\mu \in \mathfrak{M}_{\mathrm{ers}}(\mathbf{X};\sigma)$ *such that* $\mathfrak{h}^d_{\mathbf{X}} = \mathfrak{h}^d_\mu$ *on* $\mathfrak{C}$ *if and only if* $\mathfrak{h}^d_{\mathbf{X}}$ *is a d-linear form on* $\mathfrak{C}$.

Proof Proposition 6.24 and Theorem 6.25 in [12]. □

Remark 84 (a) If $\mathbf{G} \subset \mathcal{A}^{\mathbb{Z}^D}$ is an abelian subgroup shift and $\Phi \in \mathsf{ECA}(\mathbf{G})$, then $\mathbf{X}^{\mathrm{Hist}}(\Phi)$ is a subgroup shift of $\mathcal{A}^{\mathbb{Z}^{D+1}}$, which can be viewed as an algebraic $\mathbb{Z}^{D+1}$-action (see discussion prior to Proposition 27). In this context, the expansive subspaces of $\mathbf{X}^{\mathrm{Hist}}(\Phi)$ have been completely characterized by Einsiedler et al. (see Theorem 8.4 in [33]). Furthermore, certain dynamical properties (such as positive entropy, completely positive entropy, or Bernoullicity) are common amongst all elements of each expansive component of $\mathbf{X}^{\mathrm{Hist}}(\Phi)$ (see Theorem 9.8 in [33]) (this sort of 'commonality' within expansive components was earlier emphasized by Boyle and Lind (see [12])). If $\mathbf{X}^{\mathrm{Hist}}(\Phi)$ has entropy dimension 1 (e. g. Φ is a one-dimensional linear CA), the structure of $\mathbf{X}^{\mathrm{Hist}}(\Phi)$ has been thoroughly analyzed by Einsiedler and Lind [30]. Finally, if $\mathbf{G}_1$ and $\mathbf{G}_2$ are subgroup shifts, and $\Phi_k \in \mathsf{ECA}(\mathbf{G}_k)$ and $\mu_k \in \mathfrak{M}_{\mathrm{ers}}(\mathbf{G}_k; \Phi, \sigma)$ for $k = 1, 2$, with $\dim(\mu_1) = \dim(\mu_2) = 1$, then Einsiedler and Ward [32] have given conditions for the measure-preserving systems $(\mathbf{G}_1, \mu_1; \Phi_1, \sigma)$ and $(\mathbf{G}_2, \mu_2; \Phi_2, \sigma)$ to be disjoint.

(b) Boyle and Lind's 'expansive subdynamics' concerns expansiveness along certain directions in the space-time diagram of a CA. Recently, M. Sablik has developed a theory of *directional dynamics*, which explores other topological dynamical properties (such as equicontinuity and sensitivity to initial conditions) along spatiotemporal directions in a CA; see [120], Chapitre II or [121].

Future Directions and Open Problems

1. We now have a fairly good understanding of the ergodic theory of linear and/or 'abelian' CA. The next step is to extend these results to CA with nonlinear and/or nonabelian algebraic structures. In particular:
   (a) Almost all the measure rigidity results of Subsect. "Measure Rigidity in Algebraic CA" are for endomorphic CA on abelian group shifts, except for Propositions 21 and 23. Can we extend these results to CA on nonabelian group shifts or other permutative CA?
   (b) Likewise, the asymptotic randomization results of Subsect. "Asymptotic Randomization by Linear Cellular Automata" are almost exclusively

for linear CA with scalar coefficients, and for $\mathbb{M} = \mathbb{Z}^D \times \mathbb{N}^E$. Can we extend these results to LCA with noncommuting, matrix-valued coefficients? (The problem is: if the coefficients do not commute, then the 'polynomial representation' and Lucas' theorem become inapplicable.) Also, can we obtain similar results for multiplicative CA on *nonabelian* groups? (See Remark 41(d).) What about other permutative CA? (See Remark 41(e).) Finally, what if $\mathbb{M}$ is a nonabelian group? (For example, Lind and Schmidt (unpublished) [31] have recently investigated algebraic actions of the discrete Heisenberg group.)

2. Cellular automata are often seen as models of spatially distributed computation. Meaningful 'computation' could possibly occur when a CA interacts with a highly structured initial configuration (e. g. a substitution sequence), whereas such computation is probably impossible in the roiling cauldron of noise arising from a mixing, positive entropy measure (e. g. a Bernoulli measure or Markov random field). Yet almost all the results in this article concern the interaction of CA with such mixing, positive-entropy measures. We are starting to understand the topological dynamics of CA acting on non-mixing and/or zero-entropy symbolic dynamical systems, (e. g. substitution shifts, automatic shifts, regular Toeplitz shifts, and quasisturmian shifts); see ► Dynamics of Cellular Automata in Non-compact Spaces. However, almost nothing is known about the interaction of CA with the natural invariant measures on these systems. In particular:
   (a) The invariant measures discussed in Sect. "Invariant Measures for CA" all have nonzero entropy (see, however, Example 67(c)). Are there any nontrivial zero-entropy measures for interesting CA?
   (b) The results of Subsect. "Asymptotic Randomization by Linear Cellular Automata" all concern the asymptotic randomization of initial measures with nonzero entropy, except for Remark 41(c). Are there similar results for zero-entropy measures?
   (c) Zero-entropy systems often have an appealing combinatorial description via cutting-and-stacking constructions, Bratteli diagrams, or finite state machines. Likewise, CA admit a combinatorial description (via local rules). How do these combinatorial descriptions interact?
3. As we saw in Subsect. "Domains, Defects, and Particles", and also in Propositions 48–56, emergent defect dynamics can be a powerful tool for analyzing the measurable dynamics of CA. Defects in one-dimensional CA generally act like 'particles', and their 'kinematics' is fairly well-understood. However, in higher dimensions, defects can be much more topologically complicated (e. g. they can look like curves or surfaces), and their evolution in time is totally mysterious. Can we develop a theory of multidimensional defect dynamics?
4. Almost all the results about mixing and ergodicity in Subsect. "Mixing and Ergodicity" are for one-dimensional (mostly permutative) CA and for the uniform measure on $\mathcal{A}^{\mathbb{Z}}$. Can similar results be obtained for other CA and/or measures on $\mathcal{A}^{\mathbb{Z}}$? What about CA in $\mathcal{A}^{\mathbb{Z}^D}$ for $D \geq 2$?
5. Let μ be a (Φ, σ)-invariant measure on $\mathcal{A}^{\mathbb{M}}$. Proposition 66 suggests an intriguing correspondence between certain spectral properties (namely, weak mixing and discrete spectrum) for the system $(\mathcal{A}^{\mathbb{M}}, \mu; \sigma)$ and those for the system $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$. Does a similar correspondence hold for other spectral properties, such as continuous spectrum, Lebesgue spectral type, spectral multiplicity, rigidity, or mild mixing?
6. Let $\mathbf{X} \in \mathcal{A}^{\mathbb{Z}^{D+1}}$ be a subshift admitting an expansive D-plane $\mathrm{P} \subset \mathbb{R}^{D+1}$. As discussed in Subsect. "Entropy Geometry and Expansive Subdynamics", if we regard $\mathbb{Z}^{D+1}$ as 'spacetime', then we can treat P as a 'space', and a transversal direction as 'time'. Indeed, if P is spanned by rational vectors, then the Curtis–Hedlund–Lyndon theorem implies that $\mathbf{X}$ is isomorphic to the history shift of some invertible $\Phi \in \mathsf{CA}(\mathcal{A}^{\mathbb{Z}^D})$ acting on some Φ-invariant subshift $\mathbf{Y} \subseteq \mathcal{A}^{\mathbb{Z}^D}$ (where we embed $\mathbb{Z}^D$ in P). If P is irrational, then this is not the case; however, $\mathbf{X}$ still seems very much like the history shift of a spatially distributed symbolic dynamical system, closely analogous to a CA, except with a continually fluctuating 'spatial distribution' of state information, and perhaps with occasional nonlocal interactions. For example, Proposition 79(b)[i] implies that $\dim(\mathbf{X}) \leq D$, just as for a CA. How much of the theory of invertible CA can be generalized to such systems?

I will finish with the hardest problem of all. Cellular automata are tractable mainly because of their *homogeneity*: CA are embedded in a highly regular spatial geometry (i. e. a lattice or other Cayley digraph) with the same local rule everywhere. However, many of the most interesting spatially distributed symbolic dynamical systems are not nearly this homogeneous. For example:

- CA are often proposed as models of spatially distributed physical systems. Yet in many such systems (e. g. living tissues, quantum 'foams'), the underlying geometry is not a flat Euclidean space, but a curved manifold. A good discrete model of such a manifold

can be obtained through a Voronoi tessellation of sufficient density; a realistic symbolic dynamical model would be a CA-like system defined on the dual graph of this Voronoi tessellation.

- As mentioned in question #3, defects in multidimensional CA may have the geometry of curves, surfaces, or other embedded submanifolds (possibly with varying nonzero thickness). To model the evolution of such a defect, we could treat it as a CA-like object whose underlying geometry is an (evolving) manifold, and whose local rules (although partly determined by the local rule of the original CA) are spatially heterogenous (because they are also influenced by incoming information from the ambient 'nondefective' space).
- The CA-like system arising in question #6 has a D-dimensional planar geometry, but the distribution of 'cells' within this plane (and, presumably, the local rules between them) are constantly fluctuating.

More generally, any topological dynamical system on a Cantor space can be represented as a *cellular network*: a CA-like system defined on an infinite digraph, with different local rules at different nodes. Gromov [51] has generalized the Garden of Eden Theorem 3 to this setting (see Remark 5(a)). However, other than Gromov's work, basically nothing is known about such systems. Can we generalize any of the theory of cellular automata to cellular networks? Is it possible to develop a nontrivial ergodic theory for such systems?

Acknowledgments

I would like to thank François Blanchard, Mike Boyle, Maurice Courbage, Doug Lind, Petr Kůrka, Servet Martínez, Kyewon Koh Park, Mathieu Sablik, Jeffrey Steif, and Marcelo Sobottka, who read draft versions of this article and made many invaluable suggestions, corrections, and comments. (Any errors which remain are mine.) To Reem.

Bibliography

1. Akin E (1993) The general topology of dynamical systems, Graduate Studies in Mathematics, vol 1. American Mathematical Society, Providence
2. Allouche JP (1999) Cellular automata, finite automata, and number theory. In: Cellular automata (Saissac, 1996), Math. Appl., vol 460. Kluwer, Dordrecht, pp 321–330
3. Allouche JP, Skordev G (2003) Remarks on permutive cellular automata. J Comput Syst Sci 67(1):174–182
4. Allouche JP, von Haeseler F, Peitgen HO, Skordev G (1996) Linear cellular automata, finite automata and Pascal's triangle. Discret Appl Math 66(1):1–22
5. Allouche JP, von Haeseler F, Peitgen HO, Petersen A, Skordev G (1997) Automaticity of double sequences generated by one-dimensional linear cellular automata. Theoret Comput Sci 188(1-2):195–209
6. Barbé A, von Haeseler F, Peitgen HO, Skordev G (1995) Coarse-graining invariant patterns of one-dimensional two-state linear cellular automata. Internat J Bifur Chaos Appl Sci Eng 5(6):1611–1631
7. Barbé A, von Haeseler F, Peitgen HO, Skordev G (2003) Rescaled evolution sets of linear cellular automata on a cylinder. Internat J Bifur Chaos Appl Sci Eng 13(4):815–842
8. Belitsky V, Ferrari PA (2005) Invariant measures and convergence properties for cellular automaton 184 and related processes. J Stat Phys 118(3-4):589–623
9. Blanchard F, Maass A (1997) Dynamical properties of expansive one-sided cellular automata. Israel J Math 99:149–174
10. Blank M (2003) Ergodic properties of a simple deterministic traffic flow model. J Stat Phys 111(3-4):903–930
11. Boccara N, Naser J, Roger M (1991) Particle-like structures and their interactions in spatiotemporal patterns generated by one-dimensional deterministic cellular automata. Phys Rev A 44(2):866–875
12. Boyle M, Lind D (1997) Expansive subdynamics. Trans Amer Math Soc 349(1):55–102
13. Boyle M, Fiebig D, Fiebig UR (1997) A dimension group for local homeomorphisms and endomorphisms of onesided shifts of finite type. J Reine Angew Math 487:27–59
14. Burton R, Steif JE (1994) Non-uniqueness of measures of maximal entropy for subshifts of finite type. Ergodic Theory Dynam Syst 14(2):213–235
15. Burton R, Steif JE (1995) New results on measures of maximal entropy. Israel J Math 89(1-3):275–300
16. Cai H, Luo X (1993) Laws of large numbers for a cellular automaton. Ann Probab 21(3):1413–1426
17. Cattaneo G, Formenti E, Manzini G, Margara L (1997) On ergodic linear cellular automata over Z_m. In: STACS 97 (Lübeck). Lecture Notes in Computer Science, vol 1200. Springer, Berlin, pp 427–438
18. Cattaneo G, Formenti E, Manzini G, Margara L (2000) Ergodicity, transitivity, and regularity for linear cellular automata over Z_m. Theoret Comput Sci 233(1-2):147–164
19. Ceccherini-Silberstein T, Fiorenzi F, Scarabotti F (2004) The Garden of Eden theorem for cellular automata and for symbolic dynamical systems. In: Random walks and geometry, de Gruyter, Berlin, pp 73–108
20. Ceccherini-Silberstein TG, Machì A, Scarabotti F (1999) Amenable groups and cellular automata. Ann Inst Fourier (Grenoble) 49(2):673–685
21. Courbage M, Kamiński B (2002) On the directional entropy of $\mathbb{Z}^2$-actions generated by cellular automata. Studia Math 153(3):285–295
22. Courbage M, Kamiński B (2006) Space-time directional Lyapunov exponents for cellular automata. J Stat Phys 124(6):1499–1509
23. Coven EM (1980) Topological entropy of block maps. Proc Amer Math Soc 78(4):590–594
24. Coven EM, Paul ME (1974) Endomorphisms of irreducible subshifts of finite type. Math Syst Theory 8(2):167–175
25. Downarowicz T (1997) The royal couple conceals their mutual relationship: a noncoalescent Toeplitz flow. Israel J Math 97:239–251

26. Durrett R, Steif JE (1991) Some rigorous results for the Greenberg–Hastings model. J Theoret Probab 4(4):669–690
27. Durrett R, Steif JE (1993) Fixation results for threshold voter systems. Ann Probab 21(1):232–247
28. Einsiedler M (2004) Invariant subsets and invariant measures for irreducible actions on zero-dimensional groups. Bull London Math Soc 36(3):321–331
29. Einsiedler M (2005) Isomorphism and measure rigidity for algebraic actions on zero-dimensional groups. Monatsh Math 144(1):39–69
30. Einsiedler M, Lind D (2004) Algebraic $\mathbb{Z}^d$-actions on entropy rank one. Trans Amer Math Soc 356(5):1799–1831 (electronic)
31. Einsiedler M, Rindler H (2001) Algebraic actions of the discrete Heisenberg group and other non-abelian groups. Aequationes Math 62(1–2):117–135
32. Einsiedler M, Ward T (2005) Entropy geometry and disjointness for zero-dimensional algebraic actions. J Reine Angew Math 584:195–214
33. Einsiedler M, Lind D, Miles R, Ward T (2001) Expansive subdynamics for algebraic $\mathbb{Z}^d$-actions. Ergodic Theory Dynam Systems 21(6):1695–1729
34. Eloranta K, Nummelin E (1992) The kink of cellular automaton Rule 18 performs a random walk. J Stat Phys 69(5-6):1131–1136
35. Fagnani F, Margara L (1998) Expansivity, permutivity, and chaos for cellular automata. Theory Comput Syst 31(6):663–677
36. Ferrari PA, Maass A, Martínez S, Ney P (2000) Cesàro mean distribution of group automata starting from measures with summable decay. Ergodic Theory Dynam Syst 20(6):1657–1670
37. Finelli M, Manzini G, Margara L (1998) Lyapunov exponents versus expansivity and sensitivity in cellular automata. J Complexity 14(2):210–233
38. Fiorenzi F (2000) The Garden of Eden theorem for sofic shifts. Pure Math Appl 11(3):471–484
39. Fiorenzi F (2003) Cellular automata and strongly irreducible shifts of finite type. Theoret Comput Sci 299(1-3):477–493
40. Fiorenzi F (2004) Semi-strongly irreducible shifts. Adv Appl Math 32(3):421–438
41. Fisch R (1990) The one-dimensional cyclic cellular automaton: a system with deterministic dynamics that emulates an interacting particle system with stochastic dynamics. J Theoret Probab 3(2):311–338
42. Fisch R (1992) Clustering in the one-dimensional three-color cyclic cellular automaton. Ann Probab 20(3):1528–1548
43. Fisch R, Gravner J (1995) One-dimensional deterministic Greenberg–Hastings models. Complex Systems 9(5):329–348
44. Furstenberg H (1967) Disjointness in ergodic theory, minimal sets, and a problem in Diophantine approximation. Math Systems Theory 1:1–49
45. Gilman RH (1987) Classes of linear automata. Ergodic Theory Dynam Syst 7(1):105–118
46. Gottschalk W (1973) Some general dynamical notions. In: Recent advances in topological dynamics (Proc Conf Topological Dynamics, Yale Univ, New Haven, 1972; in honor of Gustav Arnold Hedlund), Lecture Notes in Math, vol 318. Springer, Berlin, pp 120–125
47. Grassberger P (1984) Chaos and diffusion in deterministic cellular automata. Phys D 10(1-2):52–58, cellular automata (Los Alamos, 1983)
48. Grassberger P (1984) New mechanism for deterministic diffusion. Phys Rev A 28(6):3666–3667
49. Grigorchuk RI (1984) Degrees of growth of finitely generated groups and the theory of invariant means. Izv Akad Nauk SSSR Ser Mat 48(5):939–985
50. Gromov M (1981) Groups of polynomial growth and expanding maps. Inst Hautes Études Sci Publ Math 53:53–73
51. Gromov M (1999) Endomorphisms of symbolic algebraic varieties. J Eur Math Soc (JEMS) 1(2):109–197
52. von Haeseler F, Peitgen HO, Skordev G (1992) Pascal's triangle, dynamical systems and attractors. Ergodic Theory Dynam Systems 12(3):479–486
53. von Haeseler F, Peitgen HO, Skordev G (1993) Cellular automata, matrix substitutions and fractals. Ann Math Artificial Intelligence 8(3-4):345–362, theorem proving and logic programming (1992)
54. von Haeseler F, Peitgen HO, Skordev G (1995) Global analysis of self-similarity features of cellular automata: selected examples. Phys D 86(1-2):64–80, chaos, order and patterns: aspects of nonlinearity– the "gran finale" (Como, 1993)
55. von Haeseler F, Peitgen HO, Skordev G (1995) Multifractal decompositions of rescaled evolution sets of equivariant cellular automata. Random Comput Dynam 3(1-2):93–119
56. von Haeseler F, Peitgen HO, Skordev G (2001) Self-similar structure of rescaled evolution sets of cellular automata. I. Internat J Bifur Chaos Appl Sci Eng 11(4):913–926
57. von Haeseler F, Peitgen HO, Skordev G (2001) Self-similar structure of rescaled evolution sets of cellular automata. II. Internat J Bifur Chaos Appl Sci Eng 11(4):927–941
58. Hedlund GA (1969) Endormorphisms and automorphisms of the shift dynamical system. Math Syst Theory 3:320–375
59. Hilmy H (1936) Sur les centres d'attraction minimaux des systeémes dynamiques. Compositio Mathematica 3:227–238
60. Host B (1995) Nombres normaux, entropie, translations. Israel J Math 91(1-3):419–428
61. Host B, Maass A, Martìnez S (2003) Uniform Bernoulli measure in dynamics of permutative cellular automata with algebraic local rules. Discret Contin Dyn Syst 9(6):1423–1446
62. Hurd LP, Kari J, Culik K (1992) The topological entropy of cellular automata is uncomputable. Ergodic Theory Dynam Syst 12(2):255–265
63. Hurley M (1990) Attractors in cellular automata. Ergodic Theory Dynam Syst 10(1):131–140
64. Hurley M (1990) Ergodic aspects of cellular automata. Ergodic Theory Dynam Syst 10(4):671–685
65. Hurley M (1991) Varieties of periodic attractor in cellular automata. Trans Amer Math Soc 326(2):701–726
66. Hurley M (1992) Attractors in restricted cellular automata. Proc Amer Math Soc 115(2):563–571
67. Jen E (1988) Linear cellular automata and recurring sequences in finite fields. Comm Math Phys 119(1):13–28
68. Johnson A, Rudolph DJ (1995) Convergence under $\times_q$ of $\times_p$ invariant measures on the circle. Adv Math 115(1):117–140
69. Johnson ASA (1992) Measures on the circle invariant under multiplication by a nonlacunary subsemigroup of the integers. Israel J Math 77(1-2):211–240
70. Kitchens B (2000) Dynamics of *mathbfZ*^{d} actions on Markov subgroups. In: Topics in symbolic dynamics and applications (Temuco, 1997), London Math Soc Lecture Note Ser, vol 279. Cambridge Univ Press, Cambridge, pp 89–122

71. Kitchens B, Schmidt K (1989) Automorphisms of compact groups. Ergodic Theory Dynam Syst 9(4):691–735
72. Kitchens B, Schmidt K (1992) Markov subgroups of $(\mathbf{Z}/2\mathbf{Z})^{\mathbf{Z}^2}$. In: Symbolic dynamics and its applications (New Haven, 1991), Contemp Math, vol 135. Amer Math Soc, Providence, pp 265–283
73. Kitchens BP (1987) Expansive dynamics on zero-dimensional groups. Ergodic Theory Dynam Syst 7(2):249–261
74. Kleveland R (1997) Mixing properties of one-dimensional cellular automata. Proc Amer Math Soc 125(6):1755–1766
75. Kůrka P (1997) Languages, equicontinuity and attractors in cellular automata. Ergodic Theory Dynam Systems 17(2):417–433
76. Kůrka P (2001) Topological dynamics of cellular automata. In: Codes, systems, and graphical models (Minneapolis, 1999), IMA vol Math Appl, vol 123. Springer, New York, pp 447–485
77. Kůrka P (2003) Cellular automata with vanishing particles. Fund Inform 58(3-4):203–221
78. Kůrka P (2005) On the measure attractor of a cellular automaton. Discret Contin Dyn Syst (suppl.):524–535
79. Kůrka P, Maass A (2000) Limit sets of cellular automata associated to probability measures. J Stat Phys 100(5-6):1031–1047
80. Kůrka P, Maass A (2002) Stability of subshifts in cellular automata. Fund Inform 52(1-3):143–155, special issue on cellular automata
81. Lind D, Marcus B (1995) An introduction to symbolic dynamics and coding. Cambridge University Press, Cambridge
82. Lind DA (1984) Applications of ergodic theory and sofic systems to cellular automata. Phys D 10(1-2):36–44, cellular automata (Los Alamos, 1983)
83. Lind DA (1987) Entropies of automorphisms of a topological Markov shift. Proc Amer Math Soc 99(3):589–595
84. Lucas E (1878) Sur les congruences des nombres eulériens et les coefficients différentiels des functions trigonométriques suivant un module premier. Bull Soc Math France 6:49–54
85. Lyons R (1988) On measures simultaneously 2- and 3-invariant. Israel J Math 61(2):219–224
86. Maass A (1996) Some dynamical properties of one-dimensional cellular automata. In: Dynamics of complex interacting systems (Santiago, 1994), Nonlinear Phenom. Complex Systems, vol 2. Kluwer, Dordrecht, pp 35–80
87. Maass A, Martínez S (1998) On Cesàro limit distribution of a class of permutative cellular automata. J Stat Phys 90(1–2): 435–452
88. Maass A, Martínez S (1999) Time averages for some classes of expansive one-dimensional cellular automata. In: Cellular automata and complex systems (Santiago, 1996), Nonlinear Phenom. Complex Systems, vol 3. Kluwer, Dordrecht, pp 37–54
89. Maass A, Martínez S, Pivato M, Yassawi R (2006) Asymptotic randomization of subgroup shifts by linear cellular automata. Ergodic Theory Dynam Syst 26(4):1203–1224
90. Maass A, Martínez S, Pivato M, Yassawi R (2006) Attractiveness of the Haar measure for the action of linear cellular automata in abelian topological Markov chains. In: Dynamics and Stochastics: Festschrift in honour of Michael Keane, vol 48 of, Lecture Notes Monograph Series of the IMS, vol 48. Institute for Mathematical Statistics, Beachwood, pp 100–108
91. Maass A, Martínez S, Sobottka M (2006) Limit measures for affine cellular automata on topological Markov subgroups. Nonlinearity 19(9):2137–2147, http://stacks.iop.org/0951-7715/19/2137
92. Machì A, Mignosi F (1993) Garden of Eden configurations for cellular automata on Cayley graphs of groups. SIAM J Discret Math 6(1):44–56
93. Maruoka A, Kimura M (1976) Condition for injectivity of global maps for tessellation automata. Inform Control 32(2):158–162
94. Mauldin RD, Skordev G (2000) Random linear cellular automata: fractals associated with random multiplication of polynomials. Japan J Math (NS) 26(2):381–406
95. Meester R, Steif JE (2001) Higher-dimensional subshifts of finite type, factor maps and measures of maximal entropy. Pacific J Math 200(2):497–510
96. Milnor J (1985) Correction and remarks: "On the concept of attractor". Comm Math Phys 102(3):517–519
97. Milnor J (1985) On the concept of attractor. Comm Math Phys 99(2):177–195
98. Milnor J (1986) Directional entropies of cellular automaton-maps. In: Disordered systems and biological organization (Les Houches, 1985), NATO Adv Sci Inst Ser F Comput Syst Sci, vol 20. Springer, Berlin, pp 113–115
99. Milnor J (1988) On the entropy geometry of cellular automata. Complex Syst 2(3):357–385
100. Miyamoto M (1979) An equilibrium state for a one-dimensional life game. J Math Kyoto Univ 19(3):525–540
101. Moore C (1997) Quasilinear cellular automata. Phys D 103(1–4):100–132, lattice dynamics (Paris, 1995)
102. Moore C (1998) Predicting nonlinear cellular automata quickly by decomposing them into linear ones. Phys D 111(1–4):27–41
103. Moore EF (1963) Machine models of self reproduction. Proc Symp Appl Math 14:17–34
104. Myhill J (1963) The converse of Moore's Garden-of-Eden theorem. Proc Amer Math Soc 14:685–686
105. Nasu M (1995) Textile systems for endomorphisms and automorphisms of the shift. Mem Amer Math Soc 114(546): viii+215
106. Nasu M (2002) The dynamics of expansive invertible onesided cellular automata. Trans Amer Math Soc 354(10):4067–4084 (electronic)
107. Park KK (1995) Continuity of directional entropy for a class of $\mathbf{Z}^2$-actions. J Korean Math Soc 32(3):573–582
108. Park KK (1996) Entropy of a skew product with a $\mathbf{Z}^2$-action. Pacific J Math 172(1):227–241
109. Park KK (1999) On directional entropy functions. Israel J Math 113:243–267
110. Parry W (1964) Intrinsic Markov chains. Trans Amer Math Soc 112:55–66
111. Pivato M (2003) Multiplicative cellular automata on nilpotent groups: structure, entropy, and asymptotics. J Stat Phys 110(1-2):247–267
112. Pivato M (2005) Cellular automata versus quasisturmian shifts. Ergodic Theory Dynam Syst 25(5):1583–1632
113. Pivato M (2005) Invariant measures for bipermutative cellular automata. Discret Contin Dyn Syst 12(4):723–736
114. Pivato M (2007) Spectral domain boundaries cellular automata. Fundamenta Informaticae 77(special issue), available at: http://arxiv.org/abs/math.DS/0507091
115. Pivato M (2008) Module shifts and measure rigidity in linear cellular automata. Ergodic Theory Dynam Syst (to appear)

116. Pivato M, Yassawi R (2002) Limit measures for affine cellular automata. Ergodic Theory Dynam Syst 22(4):1269–1287
117. Pivato M, Yassawi R (2004) Limit measures for affine cellular automata. II. Ergodic Theory Dynam Syst 24(6):1961–1980
118. Pivato M, Yassawi R (2006) Asymptotic randomization of sofic shifts by linear cellular automata. Ergodic Theory Dynam Syst 26(4):1177–1201
119. Rudolph DJ (1990) $\times 2$ and $\times 3$ invariant measures and entropy. Ergodic Theory Dynam Syst 10(2):395–406
120. Sablik M (2006) Étude de l'action conjointe d'un automate cellulaire et du décalage: Une approche topologique et ergodique. Ph D thesis, Université de la Méditerranée, Faculté des science de Luminy, Marseille
121. Sablik M (2008) Directional dynamics for cellular automata: A sensitivity to initial conditions approach. submitted to Theor Comput Sci 400(1–3):1–18
122. Sablik M (2008) Measure rigidity for algebraic bipermutative cellular automata. Ergodic Theory Dynam Syst 27(6):1965–1990
123. Sato T (1997) Ergodicity of linear cellular automata over $\mathbf{Z}_m$. Inform Process Lett 61(3):169–172
124. Schmidt K (1995) Dynamical systems of algebraic origin, Progress in Mathematics, vol 128. Birkhäuser, Basel
125. Shereshevsky MA (1992) Ergodic properties of certain surjective cellular automata. Monatsh Math 114(3-4):305–316
126. Shereshevsky MA (1992) Lyapunov exponents for one-dimensional cellular automata. J Nonlinear Sci 2(1):1–8
127. Shereshevsky MA (1993) Expansiveness, entropy and polynomial growth for groups acting on subshifts by automorphisms. Indag Math (NS) 4(2):203–210
128. Shereshevsky MA (1996) On continuous actions commuting with actions of positive entropy. Colloq Math 70(2):265–269
129. Shereshevsky MA (1997) K-property of permutative cellular automata. Indag Math (NS) 8(3):411–416
130. Shereshevsky MA, Afraĭmovich VS (1992/93) Bipermutative cellular automata are topologically conjugate to the one-sided Bernoulli shift. Random Comput Dynam 1(1):91–98
131. Shirvani M, Rogers TD (1991) On ergodic one-dimensional cellular automata. Comm Math Phys 136(3):599–605
132. Silberger S (2005) Subshifts of the three dot system. Ergodic Theory Dynam Syst 25(5):1673–1687
133. Smillie J (1988) Properties of the directional entropy function for cellular automata. In: Dynamical systems (College Park, 1986–87), Lecture Notes in Math, vol 1342. Springer, Berlin, pp 689–705
134. Sobottka M (2005) Representación y aleatorización en sistemas dinámicos de tipo algebraico. Ph D thesis, Universidad de Chile, Facultad de ciencias físicas y matemáticas, Santiago
135. Sobottka M (2007) Topological quasi-group shifts. Discret Continuous Dyn Syst 17(1):77–93
136. Sobottka M (to appear 2007) Right-permutative cellular automata on topological Markov chains. Discret Continuous Dyn Syst. Available at http://arxiv.org/abs/math/0603326
137. Steif JE (1994) The threshold voter automaton at a critical point. Ann Probab 22(3):1121–1139
138. Takahashi S (1990) Cellular automata and multifractals: dimension spectra of linear cellular automata. Phys D 45(1–3):36–48, cellular automata: theory and experiment (Los Alamos, NM, 1989)
139. Takahashi S (1992) Self-similarity of linear cellular automata. J Comput Syst Sci 44(1):114–140
140. Takahashi S (1993) Cellular automata, fractals and multifractals: space-time patterns and dimension spectra of linear cellular automata. In: Chaos in Australia (Sydney, 1990), World Sci Publishing, River Edge, pp 173–195
141. Tisseur P (2000) Cellular automata and Lyapunov exponents. Nonlinearity 13(5):1547–1560
142. Walters P (1982) An introduction to ergodic theory, Graduate Texts in Mathematics, vol 79. Springer, New York
143. Weiss B (2000) Sofic groups and dynamical systems. Sankhyā Ser A 62(3):350–359
144. Willson SJ (1975) On the ergodic theory of cellular automata. Math Syst Theory 9(2):132–141
145. Willson SJ (1984) Cellular automata can generate fractals. Discret Appl Math 8(1):91–99
146. Willson SJ (1984) Growth rates and fractional dimensions in cellular automata. Phys D 10(1-2):69–74, cellular automata (Los Alamos, 1983)
147. Willson SJ (1986) A use of cellular automata to obtain families of fractals. In: Chaotic dynamics and fractals (Atlanta, 1985), Notes Rep Math Sci Eng, vol 2. Academic Press, Orlando, pp 123–140
148. Willson SJ (1987) Computing fractal dimensions for additive cellular automata. Phys D 24(1-3):190–206
149. Willson SJ (1987) The equality of fractional dimensions for certain cellular automata. Phys D 24(1-3):179–189
150. Wolfram S (1985) Twenty problems in the theory of cellular automata. Physica Scripta 9:1–35
151. Wolfram S (1986) Theory and Applications of Cellular Automata. World Scientific, Singapore

Ergodic Theory: Fractal Geometry

Jörg Schmeling
Center for Mathematical Sciences, Lund University, Lund, Sweden

Article Outline

Glossary
Definition of the Subject
Introduction
Preliminaries
Brief Tour Through Some Examples
Dimension Theory of Low-Dimensional Dynamical Systems – Young's Dimension Formula
Dimension Theory of Higher-Dimensional Dynamical Systems
Hyperbolic Measures
General Theory
Endomorphisms
Multifractal Analysis
Future Directions
Bibliography

Glossary

Dynamical system A (discrete time) dynamical system describes the time evolution of a point in phase space. More precisely a space X is given and the time evolution is given by a map $T: X \to X$. The main interest is to describe the asymptotic behavior of the trajectories (orbits) $T^n(x)$, i. e. the evolution of an initial point $x \in X$ under the iterates of the map T. More generally one is interested in obtaining information on the geometrically complicated invariant sets or measures which describe the asymptotic behavior.

Fractal geometry Many objects of interest (invariant sets, invariant measures etc.) exhibit a complicated structure that is far from being smooth or regular. The aim of fractal geometry is to study those objects. One of the main tools is the fractal dimension theory that helps to extract important properties of geometrically "irregular" sets.

Definition of the Subject

The connection between fractal geometry and dynamical system theory is very diverse. There is no unified approach and many of the ideas arose from significant examples. Also the dynamical system theory has been shown to have a strong impact on classical fractal geometry. In this article there are first presented some examples showing nontrivial results coming from the application of dimension theory. Some of these examples require a deeper knowledge of the theory of smooth dynamical systems then can be provided here. Nevertheless, the flavor of these examples can be understood. Then there is a brief overview of some of the most developed parts of the application of fractal geometry to dynamical system theory. Of course a rigorous and complete treatment of the theory cannot be given. The cautious reader may wish to check the original papers. Finally, there is an outlook over the most recent developments. This article is by no means meant to be complete. It is intended to give some of the ideas and results from this field.

Introduction

In this section some of the aspects of fractal geometry in dynamical systems are pointed out. Some notions that are used will be defined later on and I intend only to give a flavor of the applications. The nonfamiliar reader will find the definitions in the corresponding sections and can return to this section later. The *geometry* of many invariant sets or invariant measures of dynamical systems (including attractors, measures defining the statistical properties) look very complicated at *all scales*, and their geometry is impossible to describe using standard geometric tools. For some important classes of dynamical systems, these complicated structures are intensively studied using notions of dimension. In many cases it becomes possible to relate these notions of dimension to other fundamental dynamical characteristics, such as Lyapunov exponents, entropies, pressure, etc.

On the other hand tools from dynamical systems, especially from ergodic theory and thermodynamic formalism, are extremely useful to explore the fractal properties of the objects in question. This includes dimensions of limit sets of geometric constructions (the standard Cantor set being the most famous example), which a priori, are not related to dynamical systems [46,100]. Many dimension formulas for asymptotic sets of dynamical systems are obtained by means of Bowen-type formulas, i. e. as roots of some functionals arising from thermodynamic formalism.

The dimension of a set is a subtle characteristic which measures the *geometric complexity* of the set at arbitrarily fine scales. There are many notions of dimension, and most definitions involve a measurement of geometric complexity at scale ε (which ignores the irregularities of the set at size less than ε) and then considers the limiting measurement as $\varepsilon \to 0$. A priori (and in general) these different notions can be different. An important result is the affirmative solution of the Eckmann–Ruelle conjecture by Barreira, Pesin and Schmeling [17], which says that for smooth nonuniformly hyperbolic systems, the pointwise dimension is almost everywhere constant with respect to a hyperbolic measure. This result implies that many dimension characteristics for the measure coincide.

The deep connection between dynamical systems and dimension theory seems to have been first discovered by Billingsley [21] through several problems in number theory.

Another link between dynamical systems and dimension theory is through Pesin's theory of dimension-like characteristics. This general theory is a unification of many notions of dimension along with many fundamental quantities in dynamical system such as entropies, pressure, etc.

However, there are numerous examples of dynamical systems exhibiting pathological behavior with respect to fractal geometrical characteristics. In particular higher-dimensional systems seem to be as complicated as general objects considered in geometric measure theory. Therefore, a clean and unified theory is still not available.

The study of characteristic notions like entropy, exponents or dimensions is an essential issue in the theory of dynamical systems. In many cases it helps to classify or to understand the dynamics. Most of these charac-

teristics were introduced for different questions and concepts. For example, entropy was introduced to distinguish nonisomorphic systems and appeared to be a complete invariant for Bernoulli systems (Ornstein). Later, the thermodynamic formalism (see [107]) introduced new quantities like the pressure. Bowen [27] and also Ruelle discovered a remarkable connection between the thermodynamic formalism and the dimension theory for invariant sets. Since then many efforts were taken to find the relations between all these different quantities. It occurred that the dimension of invariant sets or measures carries lots of information about the system, combining its combinatorial complexity with its geometric complexity. Unfortunately it is extremely difficult to compute the dimension in general. The general flavor is that local divergence of orbits and global recurrence cause complicated global behavior (chaos). It is impossible to study the exact (infinite) trajectory of all orbits. One way out is to study the statistical properties of "typical" orbits by means of an invariant measure. Although the underlying system might be smooth the invariant measures may often be singular.

Preliminaries

Throughout the article the following situation is considered. Let M be a compact Riemannian manifold without boundary. On M is acting a dynamical system generated by a $C^{1+\alpha}$ diffeomorphism $T\colon M \to M$. The presence of a dynamical system provides several important additional tools and methods for the theory of fractal dimensions. Also the theory of fractal dimensions allows one to draw deep conclusions about the dynamical system. The importance and relevance of the study of fractal dimension will be explained in later sections.

In the next sections some of the most important tools in the fractal theory of dynamical systems are considered. The definitions given here are not necessarily the original definitions but rather the ones which are closer to contemporary use. More details can be found in [95].

Some Ergodic Theory

Ergodic theory is a powerful method to analyze statistical properties of dynamical systems. All the following facts can be found in standard books on ergodic theory like [103,124].

The main idea in ergodic theory is to relate global quantities to observations along single orbits. Let us consider an **invariant measure**: $\mu(f^{-1}A) = \mu(A)$ for all measurable sets A. Such a measure "selects" typical trajectories. It is important to note that the properties vary with the invariant measures. Any such invariant measure can be decomposed into elementary parts (ergodic components).

An invariant measure is called **ergodic** if for any invariant set $A = T^{-1}A$ one has $\mu(A)\mu(M \setminus A) = 0$ (with the agreement $0 \cdot \infty = 0$), i. e. from the measure-theoretic point of view there are no nontrivial invariant subsets.

The importance of ergodic **probability** measures (i. e. $\mu(M) = 1$) lies in the following theorem of Birkhoff

Theorem 1 (Birkhoff) *Let μ be an ergodic probability measure and $\varphi \in L^1(\mu)$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \varphi(T^k x) = \int_M \varphi \, \mathrm{d}\mu \quad \mu - a.e\,.$$

Hausdorff Dimension

With $Z \subset \mathbb{R}^N$ and $s \geq 0$ one defines

$$m_{\mathrm{H}}(s, Z) = \lim_{\delta\to 0} \inf_{\{B_i\}} \Big\{ \sum_i \operatorname{diam}(B_i)^s : \sup_i \operatorname{diam}(B_i) < \varepsilon \text{ and } \bigcup_i B_i \supset Z \Big\}\,.$$

Note that this limit exists. $m_{\mathrm{H}}(s, Z)$ is called the **s-dimensional outer Hausdorff measure of Z**. It is immediate that there exists a unique value s^*, called the **Hausdorff dimension of Z**, at which $m_{\mathrm{H}}(s, Z)$ jumps from ∞ to 0.

In general it is very hard to find optimal coverings and hence it is often impossible to compute the Hausdorff dimension of a set. Therefore a simpler notion – the lower and upper **Box dimension** – was introduced. The difference to the Hausdorff dimension is that the covering balls are assumed to have the same radius ε. Since then the limit as $\varepsilon \to 0$ does not have to exist one arrives at the notion of the upper and lower dimension.

Dimension of a Measure

Definition 1 Let $Z \subset \mathbb{R}^N$ and let μ be a probability measure supported on Z. Define the **Hausdorff dimension of the measure μ** by

$$\dim_{\mathrm{H}}(\mu) \equiv \inf_{K\subset Z\colon \mu(K)=1} \dim_{\mathrm{H}}(K).$$

Pointwise Dimension

Most invariant sets or measures are not strongly self-similar, i. e. the local geometry at arbitrarily fine scales might look different from point to point. Therefore, the notion

of pointwise dimension with respect to a Borel probability measure is defined.

Let μ be a Borel probability measure. By $B(x,\varepsilon)$ the ball with center x and radius ε is denoted. The **pointwise dimension** of the measure μ at the point x is defined as

$$d_\mu(x) := \lim_{\varepsilon\to 0} \frac{\log \mu(B(x,\varepsilon))}{\log \varepsilon}$$

if the above limit exists. If $d_\mu(x) = d$, then for small ε the measure of small balls scales as $\mu(B(x,\varepsilon)) \asymp \varepsilon^d$.

Proposition 1 *Suppose μ is a probability measure supported on $Z \subset \mathbb{R}^N$. Then $d_\mu(x) = d$ for μ – almost all $x \in Z$ implies* $\dim_H(\mu) = d$.

One should not take the existence of a local dimension (even for *good* measures) for granted. Later on it will be seen that the spectrum of the pointwise dimension (dimension spectrum) is a main object of study in classical multifractal analysis.

The dimension of a measure or a set of measures is its geometric complexity. However, under the presence of a dynamical system one also wants to measure the dynamical (combinatorial) complexity of the system. This leads to the notion of entropy.

Dimension-Like Characteristics and Topological Entropy

Pesin's theory of dimension-like characteristics provides a unified treatment of dimensions and important dynamical quantities like entropies and pressure. The topological entropy of a continuous map f with respect to a subset Z in a metric space (X,ρ) (in particular $X = M$ – a Riemannian manifold) can be defined as a dimension-like characteristic. For each $n \in \mathbb{N}$ and $\varepsilon > 0$, define the **Bowen ball** $B_n(x,\varepsilon) = \{y \in X\colon \rho(T^i(x), T^i(y)) \le \varepsilon \text{ for } 0 \le i \le n\}$. Then let

$$m_h(Z,\alpha,\mathrm{e},n) := \lim_{n\to\infty} \inf\left\{\sum_i \mathrm{e}^{\alpha n_i} : n_i > n, \bigcup_i B_{n_i}(x,\mathrm{e}) \supset Z\right\} .$$

This gives rise to an outer measure that jumps from ∞ to 0 at some value α^*. This threshold value α^* is called the **topological entropy of Z** (at scale e). However, in many situations this value does not depend on e. The topological entropy is denoted by $h_{\mathrm{top}}(T|Z)$.

If Z is f – invariant and compact, this definition of topological entropy coincides with the usual definition of topological entropy [124].

The **entropy h_μ of a measure** μ is defined as $h_\mu = \inf\{h_{\mathrm{top}}(T|Z)\colon \mu(Z) = 1\}$. For ergodic measures this definition coincides with the Kolmogorov–Sinai entropy (see [95]).

One has to note that in the definition of entropy metric ("round") balls are substituted by Bowen balls and the metric diameter by the "depth" of the Bowen ball. Therefore, the relation between entropy and dimension is determined by the relation of "round" balls to "oval" dynamical Bowen balls. If one understands how metric balls can be *efficiently* used to cover dynamical balls one can use the dynamical and relatively easy relation to compute notion of entropy to determine the dimension. However, in higher dimensions this relation is by far nontrivial. A heuristic argument for comparing "round" balls with dynamical balls is given in Subsect. "The Kaplan–Yorke Conjecture".

The Pressure Functional

A useful tool in the dimension analysis of dynamical systems is the pressure functional. It was originally defined by means of statistical physics (thermodynamic formalism) as the free energy (or pressure) of a potential φ (see for example [107]). However, in this article a dimension-like definition (see [95]) is more suitable. Again an outer measure using Bowen balls will be used. Let $\varphi\colon M \to \mathbb{R}$ be a continuous function and

$$m_P(Z,\alpha,\varepsilon,n,\varphi) = \lim_{n\to\infty} \inf\left\{\sum_i \exp\left(-\alpha n_i + \sup_{x\in B_{n_i}(x,\varepsilon)} \sum_{k=0}^{n} \varphi(T^k x)\right)\right\} .$$

This defines an outer measure that jumps from ∞ to 0 as α increases.The threshold value α^* is called the **topological pressure** of the potential φ denoted by $P(\varphi)$. In many situations it does not depend on ε.

There is also a third way of defining the pressure in terms of a variational principle (see [124]):

$$P(\varphi) = \sup_{\mu-\text{invariant}} \left(h_\mu + \int_M \varphi \,\mathrm{d}\mu\right)$$

Brief Tour Through Some Examples

Before describing the fractal theory of dynamical systems in more detail some ideas about its role are presented. The application of dimension theory has many different aspects. At this point some (but by far not all) important examples are considered that should give the reader some

feeling about the importance and wide use of dimension theory in dynamical system theory.

Dimension of Conformal Repellers: Ruelle's Pressure Formula

Computing or estimating a dimension via a *pressure formula* is a fundamental technique. Explicit properties of pressure help to analyze subtle characteristics. For example, the smooth dependence of the Hausdorff dimension of basic sets for Axiom- A surface diffeomorphisms on the derivative of the map follows from smoothness of pressure.

Ruelle proved the following pressure formula for the Hausdorff dimension of a conformal repeller. A conformal repeller J is an invariant set $T(J) = J = \{x \in M\colon f^n x \in V \forall n \in \mathbb{N}$ and some neighborhood V of $J\}$ such that for any $x \in J$ the differential $D_x T = a(x)\,\mathrm{Iso}_x$ where $a(x)$ is a scalar with $|a(x)| > 1$ and Iso_x an isometry of the tangent space $T_x M$.

Theorem 2 ([108]) *Let $T\colon J \to J$ be a conformal repeller, and consider the function $t \to P(-t \log |D_x T|)$, where P denotes pressure. Then*

$$P(-s \log |D_x T|) = 0 \quad \Longleftrightarrow \quad s = \dim_H(J)$$

Iterated Function Systems

In fractal geometry one of the most-studied objects are iterated function systems, where there are given n contractions $F_1, \cdots, F_n$ of $\mathbb{R}^d$. The unique compact set J fulfilling $J = \bigcup_i F_i(J)$ is called the attractor of the iterated function system (see [42]). One question is to evaluate its Hausdorff dimension. Often (for example under the open set condition) the attractor J can be represented as the repeller of a piecewise expanding map $T\colon \mathbb{R}^d \to \mathbb{R}^d$ where the F_i are the inverse branches of the map T. In general it is by far not trivial to determine the dimension of J or even the dimension of a measure sitting on J. The following example explains some of those difficulties.

Let $1/2 < \lambda < 1$ and consider the maps $F_i\colon [0,1] \to [0,1]$ given by $F_1(x) = \lambda x$ and $F_2(x) = \lambda x + (1 - \lambda)$. Then the images of F_1, F_2 have an essential overlap and $J = [0,1]$. If one randomizes this construction in the way that one applies both maps each with probability 1/2 a probability measure is induced on J. This measure might be absolute continuous with respect to Lebesgue measure or not. Already Erdös realized that for some special values of λ (for example for the inverse of the golden mean) the induced measure is singular. In a breakthrough paper B. Solomyak ([119]) proved that for a.e. λ the induced measure is absolutely continuous. A main ingredient in the proof is a transversality condition in the parameter space: the images of arbitrary two random samples of the (infinite) applications of the maps F_i have to cross with nonzero speed when the parameter λ changes. This is a general mechanism which allows one to handle more general situations.

Homoclinic Bifurcations for Dissipative Surface Diffeomorphisms

Homoclinic tangencies and their bifurcations play a fundamental role in the theory of dynamical systems [87,88,89]. Systems with homoclinic tangencies have a complicated and subtle quasi-local behavior. Newhouse showed that homoclinic tangencies can persist under small perturbations, and that horseshoes may co-exist with infinitely many sinks in a neighborhood of the homoclinic orbit and hence the system is not hyperbolic (**Newhouse phenomenon**).

Let $T_\mu\colon M^2 \to M^2$ be a smooth parameter family of surface diffeomorphisms that exhibits for $\mu = 0$ an invariant hyperbolic set Λ_0 (horseshoe) and undergoes a homoclinic bifurcation. The Hausdorff dimension of the hyperbolic set Λ_0 for T_0 determines whether hyperbolicity is the *typical* dynamical phenomenon near T_0 or not. If $\dim_{\mathrm{H}} \Lambda_0 < 1$, then hyperbolicity is the prevalent dynamical phenomenon near f_0. This is not the case if $\dim_{\mathrm{H}} \Lambda_0 > 1$.

More precisely, let $\mathcal{NW}_\mu$ denote the set of nonwandering points of T_μ in an open neighborhood of Λ_0 after the homoclinic bifurcation. Let ℓ denote Lebesgue measure.

Theorem 3 ([87,88]) *If $\dim_H \Lambda_0 < 1$, then*

$$\lim_{\mu_0 \to 0} \frac{\ell\{\mu \in [0, \mu_0]\colon \mathcal{NW}_\mu \text{ is hyperbolic}\}}{\mu_0} = 1\,.$$

The hyperbolicity co-exists with the Newhouse phenomena for a residual set of parameter values.

Theorem 4 (Palis–Yoccoz) *If $\dim_H \Lambda_0 > 1$, then*

$$\lim_{\mu_0 \to 0} \frac{\ell\{\mu \in [0, \mu_0]\colon \mathcal{NW}_\mu \text{ is hyperbolic}\}}{\mu_0} < 1\,.$$

Some Applications to Number Theory

Sometimes dimension problems in number theory can be transferred to dynamical systems and attacked using tools from dynamics.

Example Diadic expansion of numbers:

Consider a real number expanded in base 2, i. e. $x = \sum_{n=1}^{\infty} x_n/2^n$. Let

$$X_p = \left\{ x\colon \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} x_k = p \right\} .$$

Borel showed that $\ell(X_{1/2}) = 1$, where ℓ denotes Lebesgue measure. This result is an easy consequence of the Birkhoff ergodic theorem applied to the characteristic function of the digit 1 (which in this simple case is the Strong Law of Large Numbers for i.i.d. processes).

One can ask how large is the set X_p in general. Eggleston [39] discovered the following wonderful dimension formula, which Billingsley [21] interpreted in terms of dynamics and reproved using tools from ergodic theory.

Theorem 5 ([39]) *The Hausdorff dimension of X_p is given by*

$$\dim_H(X_p) = (1/\log 2)\big[-p\log p - (1-p)\log(1-p)\big] .$$

The underlying dynamical system is $E_2(x) = 2x \mod 1$ and the dimension is the dimension of the Bernoulli p measure. Other cases included that by Rényi who proposed generalization of the base d expansion from integer base d to noninteger base β. In this case the underlying dynamical system is the (in general non-Markovian) beta shift $\beta(x) = \beta x \mod 1$ studied in [110].

There are many investigations concerning the approximation of real numbers by rationals by using dynamical methods. The underlying dynamical system in this case is the Gauss map $T(x) = (1/x) \mod 1$. This map is uniformly expanding, but has infinitely many branches.

Example Continued fraction approximation of numbers

Consider the continued fraction expansion of $x \in [0,1]$, i. e.,

$$x = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{a_4 + \ddots}}}} = [a_1, a_2, a_3, a_4, \cdots] ,$$

and the approximants $p_n(x)/q_n(x) = [a_1, a_2, \ldots, a_n]$ (i. e. the approximation given by the finite continued fraction after step n).

The set of numbers which admit a *faster approximation* by rational numbers is defined as follows. For $\tau \geq 2$, let $\mathcal{F}_\tau$

$$\mathcal{F}_\tau = \left\{ x \in [0,1]\colon \left| x - \frac{p}{q} \right| \leq \frac{1}{q^\tau} \quad \text{for infinitely many} \quad p/q \right\} .$$

It is well known that this set has zero measure for each $\tau \geq 2$. Jarnik [58] computed the Hausdorff dimension of $\mathcal{F}_\tau$ and showed that $\dim_{\mathrm{H}}(\mathcal{F}_\tau) = 2/\tau$. However, nowadays there are methods from dynamical systems which not only allow unified proofs but also can handle more subtle subsets of the reals defined by some properties of their continued fraction expansion (see for example [3,105]).

Infinite Iterated Function Systems and Parabolic Systems

In the previous section a system with infinitely many branches appeared. This can be regarded as an iterated function system with infinitely many maps F_i. This situation is quite general. If one considers a (one-dimensional) system with a parabolic (indifferent) fixed point, i. e. there is a fixed point where the derivative has absolute value equal to 1, one often uses an induced system. For this one chooses nearby the parabolic point a higher iterate of the map in order to achieve uniform expansion away from the parabolic point. This leads to infinitely many branches since the number of iterates has to be increased the closer the parabolic point is. The main difference to the finite iterated function system is that the setting is no longer compact and many properties of the pressure functional are lost.

Mauldin and Urbański and others (see for example [3,75,76,78]) developed a thermodynamic formalism adapted to the pressure functional for infinite iterated function systems. Besides noncompactness one of the main problems is the loss of analyticity (phase transitions) and convexity of the pressure functional for infinite iterated function systems.

Complex Dynamics

Let $T\colon \mathbb{C} \to \mathbb{C}$ be a polynomial and J its Julia set (repeller of this system). If this set is hyperbolic, i. e. the derivative at each point has absolute value larger than 1 the study of the dimension can be related to the study of a finite iterated function system. However, in the presence of a parabolic fixed point this leads to an infinite iterated function system.

If one considers the coefficients of the polynomial as parameters one often sees a qualitative change in the

asymptotic behavior. For example, the classical Mandelbrodt set for polynomials $z^2 + c$ is the locus of values $c \in C$ for which the orbit of the origin stays bounded. This set is well known to be fractal. However, its complete description is still not available.

Embedology and Computational Aspects of Dimension

Tools from dynamical systems are becoming increasingly important to study the time evolution of deterministic systems in engineering and the physical and biological sciences. One of the main ideas is to model a "real world" system by a smooth dynamical system which possesses a strange attractor with a natural ergodic invariant measure. When studying a complicated real world system, one can measure only a very small number of variables. The challenge is to reconstruct the underlying attractor from the time measurement of a scaler quantity. An idealized measurement is considered as a function $h\colon M^n \to \mathbb{R}$.

The main tool researchers currently use to *reconstruct* the model system is called attractor reconstruction (see papers and references in [86]). This method is based on embedding with time delays (see the influential paper [33], where the authors attribute the idea of delay coordinates to Ruelle), where one attempts to reconstruct the attractor for the model using a single long trajectory. Then one considers the points in R^{p+1} defined by $(x_k, x_{k+\tau}, x_{k+2\tau}, \ldots, x_{k+p\tau})$.

Takens [122] showed that for a smooth $T\colon M^n \to M^n$ and for typical smooth h, the mapping $\varphi\colon M^n \to \mathbb{R}^{2n+1}$ defined by $x \to (h(x), h(f^\tau(x)), \cdots, h(f^{2n\tau}(x))$ is an embedding. Since the box dimension of the attractor Λ may be much less than the dimension of the ambient manifold n, an interesting mathematical question is whether there exists $p < 2n+1$ such that the mapping on the attractor $\varphi\colon \Lambda \to \mathbb{R}^p$ defined by $x \to (h(x), h(f(x)), \ldots, h(f^p(x))$ is $1-1$? It is known that for a typical smooth h the mapping φ is $1-1$ for $p > 2\dim_B(\Lambda)$ [29].

Denjoy Systems

This section will give some ideas indicating the principle difficulties that arise in systems with low complexity. Contrary to hyperbolic systems (each vector in the tangent space is either contracted or expanded) finer mechanisms determine the local behavior of the scaling of balls. While in hyperbolic systems the dynamical scaling of small balls is exponential in a low-complexity system this scaling is subexponential and hence the linearization error is of the same magnitude. Up to now there is no general dimension theory for low complexity systems.

A specific example presented here is considered in [67]. Poincaré showed that to each orientation preserving homeomorphism of the circle $S^1 = \mathbb{R}/\mathbb{Z}$ is associated a unique real parameter $\alpha \in [0,1)$, called the rotation number, so that the ordered orbit structure of T is the same as that of the rigid rotation R_α, where $R_\alpha(t) = (t+\alpha) \bmod 1$, provided that α is irrational. Half a century later, Denjoy [35] constructed examples of C^1 diffeomorphisms that are not conjugate (via a homeomorphism) to rotations. This was improved later on by Herman [55]. In these examples, the minimal set of T is necessarily a Cantor set Ω.

The arithmetic properties of the rotation number have a strong effect on the properties of T. One area that has been well understood is the relation between the differentiability of T, the differentiability of the conjugation and the arithmetic properties of the rotation number. (See, for example, Herman [55]) Without stating any precise theorem, the results differ sharply for Diophantine and for Liouville rotation numbers (definition follows). Roughly speaking the conjugating map is always regular for Diophantine rotation numbers while it might be not smooth at all for Liuoville rotation numbers.

Definition 2 An irrational number α is of Diophantine class $\nu = \nu(\alpha) \in \mathbb{R}^+$ if

$$\|q\alpha\| < \frac{1}{q^\mu}$$

has infinitely many solutions in integers q for $\mu < \nu$ and at most finitely many for $\mu > \nu$ where $\|\cdot\|$ denotes the distance to the nearest integer.

In [67] the effect of the rotation number on the dimension of Ω is studied. There the main result is

Theorem 6 *Assume that $0 < \delta < 1$ and that $\alpha \in (0,1)$ is of Diophantine class $\nu \in (0,\infty)$. Then an orientation preserving $C^{1+\delta}$ diffeomorphism of the circle with rotation number α and minimal set Ω_α^δ satisfies*

$$\dim_H \Omega_\alpha^\delta \geq \frac{\delta}{\nu}\,.$$

Furthermore, these results are sharp, i. e. the standard Denjoy examples attain the minimum.

Return Times and Dimension

Recently an interesting connection between the pointwise dimensions, multifractal analysis, and recurrence behavior of trajectories was discovered [1,11,23]. Roughly speaking,

given an ergodic probability measure μ the return time asymptotics (as the neighborhood of the point shrinks) of μ-a.e point is determined by the pointwise dimension of μ at this point. The deeper understanding of this relation would help to get a unified approach to dimensions, exponents, entropies, recurrence times and correlation decay.

Dimension Theory of Low-Dimensional Dynamical Systems – Young's Dimension Formula

In this section a remarkable extension of Ruelle's dimension formula by Young [128] for the dimension of a measure is discussed.

Theorem 7 *Let $T\colon M^2 \to M^2$ be a C^2 surface diffeomorphism and let μ be an ergodic measure. Then*

$$\dim_H(\mu) = h_\mu(f)\left(\frac{1}{\lambda^1} - \frac{1}{\lambda^2}\right),$$

where $\lambda^1 \le \lambda^2$ are the two Lyapunov exponents for μ.

In [79], Manning and McCluskey prove the following dimension formula for a basic set (horseshoe) of an Axiom-A surface diffeomorphism which is a set-version of Young's formula.

Theorem 8 *Let Λ be a basic set for a C^2 Axiom-A surface diffeomorphism $T\colon M^2 \to M^2$. Then $\dim_H(\Lambda) = s_1 + s_2$, where s_1 and s_2 satisfy*

$$P(-s\log\|DT_x|E^u_x\|) = 0$$
$$P(s\log\|DT_x|E^s_x\|) = 0\,.$$

where E^s and E^u are the stable and unstable directions, respectively.

Some Remarks on Dimension Theory for Low-Dimensional versus High-Dimensional Dynamical Systems

Unlike lower dimensions (one, two, or conformal repellers), for higher-dimensional dynamical systems there are no general dimension formulas (besides the Ledrappier–Young formula), and in general dimension theory is much more difficult. This is due to several problems:

1) The geometry of the Bowen balls differs in a substantial way from round balls.
2) Number theoretic properties of some scaling rates [104, 106] enter into dimension calculations in ways they do not in low dimensions (see Subsect. "Iterated Function Systems").
3) The dimension theory of sets is often reduced to the theory of invariant measures. However, there is no invariant measure of full dimension in general and measure-theoretic considerations do not apply [79].
4) The stable and unstable foliations for higher dimensional systems are typically not C^1 [51,109]. Hence, to split the system into an expanding and a contracting part is far more subtle.

Dimension Theory of Higher-Dimensional Dynamical Systems

Here an example of a hyperbolic attractor in dimension 3 is considered to highlight some of the difficulties. Let $\triangle$ denote the unit disc in $\mathbb{R}^2$. Let $f\colon S^1 \times \triangle \to S^1 \times \triangle$ be of the form

$$T(t,x,y) = (\varphi(t), \psi^1(t,x), \psi^2(t,y))\,,$$

with

$$0 < \max_{S^1\times\triangle} \frac{\partial}{\partial x}\psi^1(t,x) < \min_{S^1\times\triangle} \frac{\partial}{\partial y}\psi^2(t,y) < \lambda < 1\,.$$

The limit set

$$\Lambda := \bigcap_{n\in\mathbb{N}} T^n(S^1 \times \triangle)$$

is called the attractor or the *solenoid*. It is an example of a structurally stable basic set and is one of the fundamental examples of a uniformly hyperbolic attractor.

The following result can be proved.

Theorem 9 ([24,52]) *For all t, the thermodynamic pressure*

$$P\left(\dim_H \Lambda^s_t \log\left|\frac{\partial}{\partial y}\psi^2(t,y)\right|\right) = 0\,.$$

In particular, the stable dimension is independent of the stable section.

In this particular case the invariant axes for strong and weak contraction split the system smoothly and the difficulty is to show that the strong contraction is dominated by the weaker. In particular one has to ensure that effects as described in Subsect. "Iterated Function Systems" do not appear. In the general situation this is not the case and one lacks a similar theorem in the general situation. In particular, the unstable foliation is not better than Hölder and does not provide a "nice" coordinate.

Hyperbolic Measures

Given $x \in M$ and a vector $v \in T_x M$ the **Lyapunov exponent** is defined as

$$\chi(x, v) = \lim_{n\to\infty} \frac{1}{n} \log \|D_x T^n v\| \,.$$

provided this limit exists. For fixed x, the numbers $\chi(x, \cdot)$ attain only finitely many values. By ergodicity they are μ-a.e. constant, and the corresponding values are denoted by

$$\chi_1 \le \cdots \le \chi_p \,,$$

where $p = \dim M$. Denote by s (s stands for stable) the largest index such that $\chi_s < 0$.

Definition 3 The invariant ergodic measure μ is said to be **hyperbolic** if $\chi_i \neq 0$ for every $i = 1, \ldots, p$.

The Kaplan–Yorke Conjecture

In this section a heuristic argument for the dimension of an invariant ergodic measure will be given. This argument uses a specific cover of a dynamical ball by "round" balls. These ideas are essentially developed by Kaplan and Yorke [61] for the estimation of the dimension of an invariant set. Their estimates always provide an upper bound for the dimension. Kaplan and Yorke conjectured that in typical situations these estimates also provide a lower bound for the dimension of an attractor. Ledrappier and Young [72,73] showed that in case this holds for an invariant measure, then this measure has to be very special: an SBR-measure (after Sinai, Ruelle, and Bowen) (an SRB-measure is a measure that describes the asymptotic behavior for a set of initial points of positive Lebesgue measure and has absolutely continuous conditional measures on unstable manifolds).

Consider a small ball B in the phase space. The image $T^n B$ is almost an ellipsoid with axes of length

$$e^{\chi_1 n}, \cdots, e^{\chi_p n} \,.$$

For $1 \le i \le s$, cover $T^n B$ by balls of radius $e^{\chi_i n}$. Then approximately

$$\frac{\exp[\chi_{i+1} n]}{\exp[\chi_i n]} \cdots \frac{\exp[\chi_p n]}{\exp[\chi_i n]} \,.$$

balls are needed for covering. The dimension can be estimated from above by

$$\dim_B \Lambda \le \frac{\sum_{j>i} \chi_j}{|\chi_i|} + (p - i) := \dim_L^i \Lambda \,. \tag{1}$$

This is the Kaplan–Yorke formula.

General Theory

In this section the dimension theory of higher-dimensional dynamical systems is investigated. Most developed is this theory for invariant measures. There is an important connection between Lyapunov exponents and the measure theoretic entropy and dimensions of measures that will be presented here.

Let μ be an ergodic invariant measure. The Oseledec and Pesin Theory guarantee that local stable manifolds exist at μ-a.e. point. As for the Kaplan–Yorke formula the idea is to consider the contributions to the entropy and to the dimension in the directions of χ_i .

Historically the first connections between exponents and entropy were discovered by Margulis and Ruelle. They proved that

$$h_\mu(f) \le -\sum_{i=1}^{s} \chi_i$$

for a C^1 diffeomorphism T. Pesin [91] showed that this inequality is actually an equality if the measure μ is essentially the Riemannian volume on unstable manifolds. Ledrappier and Young [72] showed that this is indeed a necessary condition. They also provided an exact formula:

Theorem 10 (Ledrappier–Young [72,73]) *With $d^0 = 0$ for a C^2 diffeomorphism holds*

$$h_\mu(f) = -\sum_{i=1}^{s} \chi_i \left(d^i - d^{i-1}\right)$$

where d^i are the dimensions of the (conditional) measure on the ith unstable leaves.

The proof of this theorem is difficult and uses the theory of *nonuniform hyperbolic systems (Pesin theory)*.

In dimension 1 and 2 the above theorem resembles Ruelle's and Young's theorems.

The reader should note that the above theorem includes also the existence of the pointwise dimension along the stable and unstable direction. Here the question arises whether this implies the existence of the pointwise dimension itself.

The Existence of the Pointwise Dimension for Hyperbolic Measure – the Eckmann–Ruelle Conjecture

In [17], Barreira, Pesin, and Schmeling prove that every hyperbolic measure has an *almost* local product structure, i. e., the measure of a small ball can be approximated by the product of the stable conditional measure of the sta-

ble component and the unstable conditional measure of the unstable component, up to small exponential error. This was used to prove the existence of the pointwise dimension of every hyperbolic measure almost everywhere. Moreover, the pointwise dimension is the sum of the contributions from its stable and unstable part. This implies that most dimension-type characteristics of the measure (including the Hausdorff dimension, box dimension, and information dimension) coincide, and provides a rigorous mathematical justification of the concept of fractal dimension for hyperbolic measures. The existence of the pointwise dimension for hyperbolic measures had been conjectured long before by Eckmann and Ruelle.

The hypotheses of this theorem are sharp. Ledrappier and Misiurewicz [70] constructed an example of a non-hyperbolic measure for which the pointwise dimension is not constant a.e. In [101], Pesin and Weiss present an example of a Hölder homeomorphism with Hölder constant arbitrarily close to one, where the pointwise dimension for the unique measure of maximal entropy does not exist a.e. There is also a one-dimensional example by Cutler [34].

Endomorphisms

The previous section indicates that the dimensional properties of hyperbolic measures under invariant conditions for a diffeomorphism are understood. However, partial differential equations often generate only semi-flows and the corresponding dynamical system is noninvertible. Also, Poincaré sections sometimes introduce singularities. For such dynamical systems the theory of diffeomorphisms does not apply. However, the next theorem allows under some conditions application of this theory. It essentially rules out similar situations as considered in Subsect. "Iterated Function Systems".

Definition 4 A system (possibly with singularities) is *almost surely invertible* if it is invertible on a set of full measure. This implies that a full measure set of points has unique forward and backward trajectories.

Theorem 11 (Schmeling–Troubetzkoy [114]) *A two-dimensional system with singularities is almost surely invertible (w.r.t. an SRB–measure) if and only if Young's formula holds.*

Multifractal Analysis

A group of physicists [50] suggested the idea of a multifractal analysis.

The Dynamical Characteristic View of Multifractal Analysis

The aim of multifractal analysis is an attempt to understand the fine structure of the level sets of the fundamental asymptotic quantities in ergodic theory (e. g., Birkhoff averages, local entropy, Lyapunov exponents). For ergodic measures these quantities are a.e. constant, however may depend on the underlying ergodic measure. Important elements of multifractal analysis entail determining the range of values these characteristics attain, an analysis of the dimension of the level sets, and an understanding of the sets where the limits do not exist. A general concept of multifractal analysis was proposed by Barreira, Schmeling and Pesin [16]. An important field of applications of multifractal analysis is to describe sets of real numbers that have constraints on their digits or continued fraction expansion.

General Multifractal Formalism

In this section the abstract theory of multifractal analysis is described. Let X, Y be two measurable spaces and $g\colon X \setminus \mathcal{B} \to Y$ be any measurable function where $\mathcal{B}$ is a measurable (possibly empty) subset of X (in the standard applications $Y = \mathbb{R}$ or $Y = \mathbb{C}$). The associated **multifractal decomposition** of X is defined as

$$X = \mathcal{B} \cup \bigcup_{\alpha \in Y} K_\alpha^g$$

where

$$K_\alpha^g := \{x \in X\colon g(x) = \alpha\}$$

For a given set function $G\colon 2^X \to \mathbb{R}$ the **multifractal spectrum** is defined by

$$\mathcal{F}(\alpha) := G(K_\alpha^g)\,.$$

At this point some classical and concrete examples of this general framework are considered.

The Entropy Spectrum

Let μ be an ergodic invariant measure for $T\colon X \to X$. If one sets

$$g^E(x) := h_\mu(x) \in Y = \mathbb{R}$$

and

$$G^E(Z) = h_{\text{top}}(T|Z)$$

the associated multifractal spectrum $f_\mu^E(\alpha)$ is called the **entropy spectrum**.

The Dimension Spectrum

This is the **classical multifractal spectrum**.

Let μ be an invariant ergodic measure on a complete separable metric space X.

Set

$$g^D(x) := d_\mu(x) \in Y = \mathbb{R}$$

and

$$G^D(Z) = \dim_{\mathrm{H}} Z\,.$$

The associated multifractal spectrum $f_\mu^D(\alpha)$ is called the **dimension spectrum**.

The Lyapunov Spectrum

Let

$$g^L(x) := \chi(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \psi(\mathbb{T}^k x) \in Y = \mathbb{R}$$

where $\psi(x) = \log|D_x T|$ and

$$G^D(Z) = \dim_{\mathrm{H}} Z \quad \text{or} \quad G^E(Z) = h_{\mathrm{top}}(\sigma|Z)\,.$$

The associated multifractal spectra $f_L^D(\alpha)$ and $f_L^E(\alpha)$ are called **Lyapunov spectra**.

It was observed by H. Weiss [126] and Barreira, Pesin and Schmeling [16] that for conformal repellers

$$f_L^D(\alpha) = f_{\mu_h}^D\left(\frac{\log 2}{\alpha}\right) \tag{2}$$

where the dimension spectrum on the right-hand side is with respect to the measure of maximal entropy, and

$$f_L^E(\alpha) = f_{\mu_D}^E(\dim_{\mathrm{H}} \Lambda \cdot \alpha) \tag{3}$$

where the entropy spectrum on the right-hand side is with respect to the measure of maximal dimension.

The following list summarizes the state of the art for the dynamical characteristic multifractal analysis of dynamical systems. The precise statements can be found in the original papers.

- [102,126] For conformal repellers and Axiom-A surface diffeomorphisms, a complete multifractal analysis exists for the Lyapunov exponent.
- [16,102,126] For mixing a subshift of finite type, a complete multifractal analysis exists for the Birkhoff average for a Hölder continuous potential and for the local entropy for a Gibbs measure with Hölder continuous potential.
- [12,97] There is a complete multifractal analysis for hyperbolic flows.
- [123] There is a generalization of the multifractal analysis on subshifts with specification and continuous potentials.
- [13,19] There is an analysis of "mixed" spectra like the dimension spectrum of local entropies and also an analysis of joint level sets determined by more than one (measurable) function.
- [105] For the Gauss map (and a class of nonuniformly hyperbolic maps) a complete multifractal analysis exists for the Lyapunov exponent.
- [57] A general approach to multifractal analysis for repellers with countably many branches is developed. It shows in contrary to finitely many branches features of nonanalytic behavior.

In the first three statements the multifractal spectra are analytic concave functions that can be computed by means of the Legendre transform of the pressure functional with respect to a suitable chosen family of potentials. In the remaining items this is no longer the case. Analyticity and convexity properties of the pressure functional are lost. However, the authors succeeded to provide a satisfactory theory in these cases.

Multifractal Analysis and Large Deviation Theory

There are deep connections between large deviation theory and multifractal analysis. The variational formula for pressure is an important tool in the analysis, and can be viewed (and proven) as a large deviation result [41]. Some authors use large deviation theory as a tool to effect multifractal analysis.

Future Directions

The dimension theory is fast developing and of great importance in the theory of dynamical systems. In the most ideal situations (low dimensions and hyperbolicity) a generally far reaching and powerful theory has been developed. It uses ideas from statistical physics, fractal geometry, probability theory and other fields.

Unfortunately, the richness of this theory does not carry over to higher-dimensional systems. However, recent developments have shown that it is possible to obtain a general theory for the dimension of measures. Part of this theory is the development of the analytic tools of nonuniformly hyperbolic systems.

Therefore, the dimension theory of dynamical systems is far from complete. In particular, it is usually difficult to apply the general theory to concrete examples, for instance

if one really wants to *compute* the dimension. The general theory does not provide a way to compute the dimension but gives rather connections to other characteristics. Moreover, in the presence of neutral directions (zero Lyapunov exponents) one encounters all the difficulties arising in low-complexity systems.

Another important open problem is to understand the dimension theory of invariant sets in higher-dimensional spaces. One way would be to relate the dimension of sets to the dimension of measures. Such a connection is not clear. The reason is that most systems do not exhibit a measure whose dimension coincides with the dimension of its support (invariant set). But there are some reasons to conjecture that any compact invariant set of an expanding map in any dimension carries a measure of maximal dimension (see [48,63]). If this conjecture is true one obtains an invariant measure whose unstable dimension coincides with the unstable dimension of the invariant set. There is also a measure of maximal stable dimension. Combining these two measures one could establish an analogous theory for invariant sets as for invariant measures.

Last but not least one has to mention the impact of the dimension theory of dynamical systems on other fields. This new point of view makes in many cases the posed problems more tractable. This is illustrated in examples from number theory, geometric limit constructions and others. The applications of the dimension theory of dynamical systems to other questions seem to be unlimited.

Bibliography

Primary Literature

1. Afraimovich VS, Schmeling J, Ugalde J, Urias J (2000) Spectra of dimension for Poincaré recurrences. Discret Cont Dyn Syst 6(4):901–914
2. Afraimovich VS, Chernov NI, Sataev EA (1995) Statistical Properties of 2D Generalized Hyperbolic Attractors. Chaos 5:238–252
3. Aihua F, Yunping J, Jun W (2005) Asymptotic Hausdorff dimensions of Cantor sets associated with an asymptotically non-hyperbolic family. Ergod Theory Dynam Syst 25(6):1799–1808
4. Alexander J, Yorke J (1984) Fat Baker's transformations. Erg Th Dyn Syst 4:1–23
5. Ambroladze A, Schmeling J (2004) Lyapunov exponents are not stable with respect to arithmetic subsequences. In: Fractal geometry and stochastics III. Progr Probab 57. Birkhäuser, Basel, pp 109–116
6. Artin E (1965) Ein mechanisches System mit quasi-ergodischen Bahnen, Collected papers. Addison Wesley, pp 499–501
7. Barreira L () Variational properties of multifractal spectra. IST preprint
8. Barreira L (1995) Cantor sets with complicated geometry and modeled by general symbolic dynamics. Random Comp Dyn 3:213–239
9. Barreira L (1996) A non-additive thermodynamic formalism and applications to dimension theory of hyperbolic dynamical systems. Erg Th Dyn Syst 16:871–927
10. Barreira L (1996) A non-additive thermodynamic formalism and dimension theory of hyperbolic dynamical systems. Math Res Lett 3:499–509
11. Barreira L, Saussol B (2001) Hausdorff dimension of measure via Poincaré recurrence. Comm Math Phys 219(2):443–463
12. Barreira L, Saussol B (2001) Multifractal analysis of hyperbolic flows. Comm Math Phys 219(2):443–463
13. Barreira L, Saussol B (2001) Variational principles and mixed multifractal spectra. Trans Amer Math Soc 353(10):3919–3944 (electronic)
14. Barreira L, Schmeling J (2000) Sets of "Non-typical" Points Have Full Topological Entropy and Full Hausdorff Dimension. Isr J Math 116:29–70
15. Barreira L, Pesin Y, Schmeling J (1997) Multifractal spectra and multifractal rigidity for horseshoes. J Dyn Contr Syst 3:33–49
16. Barreira L, Pesin Y, Schmeling J (1997) On a General Concept of Multifractal Rigidity: Multifractal Spectra For Dimensions, Entropies, and Lyapunov Exponents. Multifractal Rigidity. Chaos 7:27–38
17. Barreira L, Pesin Y, Schmeling J (1999) Dimension and Product Structure of Hyperbolic Measures. Annals Math 149:755–783
18. Barreira L, Saussol B, Schmeling J (2002) Distribution of frequencies of digits via multifractal analysis. J Number Theory 97(2):410–438
19. Barreira L, Saussol B, Schmeling J (2002) Higher–dimensional multifractal analysis. J Math Pures Appl (9) 81(1):67–91
20. Belykh VP (1982) Models of discrete systems of phase synchronization. In: Shakhildyan VV, Belynshina LN (eds) Systems of Phase Synchronization. Radio i Svyaz, Moscow, pp 61–176
21. Billingsley P (1978) Ergodic Theory and Information. Krieger
22. Blinchevskaya M, Ilyashenko Y (1999) Estimate for the Entropy Dimension Of The Maximal Attractor For k—Constracting Systems In An Infinite-Dimensional Space. Russ J Math Phys 6(1):20–26
23. Boshernitzan M (1993) Quantitative recurrence results. Invent Math 113:617–631
24. Bothe H-G (1995) The Hausdorff dimension of certain solenoids. Erg Th Dyn Syst 15:449–474
25. Bousch T (2000) Le poisson n'a pas d'arêtes. Ann IH Poincaré (Prob-Stat) 36(4):489–508
26. Bowen R (1973) Topological entropy for noncompact sets. Trans Amer Math Soc 184:125–136
27. Bowen R (1979) Hausdorff Dimension Of Quasi-circles. Publ Math IHES 50:11–25
28. Bylov D, Vinograd R, Grobman D, Nemyckii V (1966) Theory of Lyapunov exponents and its application to problems of stability. Izdat "Nauka", Moscow (in Russian)
29. Casdagli M, Sauer T, Yorke J (1991) Embedology. J Stat Phys 65:589–616
30. Ciliberto S, Eckmann JP, Kamphorst S, Ruelle D (1971) Liapunov Exponents from Times. Phys Rev A 34
31. Collet P, Lebowitz JL, Porzio A (1987) The Dimension Spectrum of Some Dynamical Systems. J Stat Phys 47:609–644
32. Constantin P, Foias C (1988) Navier-Stokes Equations. Chicago U Press

33. Cruchfield J, Farmer D, Packard N, Shaw R (1980) Geometry from a Time Series. Phys Rev Lett 45:712–724
34. Cutler C (1990) Connecting Ergodicity and Dimension in Dynamical Systems. Ergod Th Dynam Syst 10:451–462
35. Denjoy A (1932) Sur les courbes défines par les équations différentielles á la surface du tore. J Math Pures Appl 2:333–375
36. Ding M, Grebogi C, Ott E, Yorke J (1993) Estimating correlation dimension from a chaotic times series: when does the plateau onset occur? Phys D 69:404–424
37. Dodson M, Rynne B, Vickers J (1990) Diophantine approximation and a lower bound for Hausdorff dimension. Mathematika 37:59–73
38. Douady A, Oesterle J (1980) Dimension de Hausdorff Des Attracteurs. CRAS 290:1135–1138
39. Eggleston HG (1952) Sets of Fractional Dimension Which Occur in Some Problems of Number Theory. Proc Lond Math Soc 54:42–93
40. Ellis R (1984) Large Deviations for a General Class of Random Vectors. Ann Prob 12:1–12
41. Ellis R (1985) Entropy, Large Deviations, and Statistical Mechanics. Springer
42. Falconer K (1990) Fractal Geometry, Mathematical Foundations and Applications. Cambridge U Press, Cambridge
43. Fan AH, Feng DJ, Wu J (2001) Recurrence, dimension and entropy. J Lond Math Soc 64(2):229–244
44. Frederickson P, Kaplan J, Yorke E, Yorke J (1983) The Liapunov Dimension Of Strange Attractors. J Differ Equ 49:185–207
45. Frostman O (1935) Potential d'équilibre Et Capacité des Ensembles Avec Quelques Applications à la Théorie des Fonctions. Meddel Lunds Univ Math Sem 3:1–118
46. Furstenberg H (1967) Disjointness in Ergodic Theory, Minimal Sets, and a Problem in Diophantine Approximation. Math Syst Theory 1:1–49
47. Furstenberg H (1970) Intersections of Cantor Sets and Transversality of Semigroups I. In: Problems in Analysis. Sympos Salomon Bochner, Princeton Univ. Princeton Univ Press, pp 41–59
48. Gatzouras D, Peres Y (1996) The variational principle for Hausdorff dimension: A survey, in Ergodic theory of $\mathbb{Z}^d$ actions. In: Pollicott M et al (ed) Proc of the Warwick symposium. Cambridge University Press. Lond Math Soc Lect Note Ser 228:113–125
49. Grassberger P, Procaccia I, Hentschel H (1983) On the Characterization of Chaotic Motions, Lect Notes. Physics 179:212–221
50. Halsey T, Jensen M, Kadanoff L, Procaccia I, Shraiman B (1986) Fractal Measures and Their Singularities: The Characterization of Strange Sets. Phys Rev A 33(N2):1141–1151
51. Hasselblatt B (1994) Regularity of the Anosov splitting and of horospheric foliations. Ergod Theory Dynam Syst 14(4):645–666
52. Hasselblatt B, Schmeling J (2004) Dimension product structure of hyperbolic sets. In: Modern dynamical systems and applications. Cambridge Univ Press, Cambridge, pp 331–345
53. Henley D (1992) Continued Fraction Cantor Sets, Hausdorff Dimension, and Functional Analysis. J Number Theory 40:336–358
54. Hentschel HGE, Procaccia I (1983) The Infinite Number of Generalized Dimensions of Fractals and Strange Attractors. Physica 8D:435–444
55. Herman MR (1979) Sur la conjugaison différentiable des difféomorphismes du cercle á des rotations. Publications de l'Institute de Mathématiques des Hautes Études Scientifiques 49:5–234
56. Hunt B (1996) Maximal Local Lyapunov Dimension Bounds The Box Dimension Of Chaotic Attractors. Nonlinearity 9:845–852
57. Iommi G (2005) Multifractal analysis for countable Markov shifts. Ergod Theory Dynam Syst 25(6):1881–1907
58. Jarnik V (1931) Über die simultanen diophantischen Approximationen. Math Zeitschr 33:505–543
59. Jenkinson O (2001) Rotation, entropy, and equilibrium states. Trans Amer Math Soc 353:3713–3739
60. Kalinin B, Sadovskaya V (2002) On pointwise dimension of non-hyperbolic measures. Ergod Theory Dynam Syst 22(6):1783–1801
61. Kaplan JL, Yorke JA (1979) Functional differential equations and approximation of fixed points. Lecture Notes. In: Mathematics vol 730. Springer, Berlin, pp 204–227
62. Katznelson Y, Weiss B (1982) A simple proof of some ergodic theorems. Isr J of Math 42:291–296
63. Kenyon R, Peres Y (1996) Measure of full dimension on affine invariant sets. Erg Th Dyn Syst 16:307–323
64. Kesseböhmer M (1999) Multifrakale und Asymptotiken grosser Deviationen. Thesis U Göttingen, Göttingen
65. Kingman JFC (1968) The ergodic theory of subadditive stochastic processes. J Royal Stat Soc B30:499–510
66. Kleinbock D, Margulis G (1998) Flows on Homogeneous Spaces and Diophantine Approximation on Manifold. Ann Math 148:339–360
67. Kra B, Schmeling J (2002) Diophantine classes, dimension and Denjoy maps. Acta Arith 105(4):323–340
68. Ledrappier F (1981) Some Relations Between Dimension And Lyapounov Exponents. Comm Math Phys 81:229–238
69. Ledrappier F (1986) Dimension of invariant measures. Proceedings of the conference on ergodic theory and related topics II (Georgenthal, 1986). Teubner–texte Math 94:137–173
70. Ledrappier F, Misiurewicz M (1985) Dimension of Invariant Measures for Maps with Exponent Zero. Ergod Th Dynam Syst 5:595–610
71. Ledrappier F, Strelcyn JM (1982) A proof of the estimate from below in Pesin's entropy formula. Ergod Theory Dynam Syst 2:203–219
72. Ledrappier F, Young LS (1985) The Metric Entropy Of Diffeomorphisms. I Characterization Of Measures Satisfying Pesin's Entropy Formula. Ann Math 122:509–539
73. Ledrappier F, Young LS (1985) The Metric Entropy Of Diffeomorphisms, II. Relations Between Entropy, Exponents and Dimension. Ann Math 122:540–574
74. Lopes A (1989) The Dimension Spectrum of the Maximal Measure. SIAM J Math Analysis 20:1243–1254
75. Mauldin RD, Urbański M (1996) Dimensions and measures in infinite iterated function systems. Proc Lond Math Soc 73(1):105–154
76. Mauldin RD, Urbański M (2002) Fractal measures for parabolic IFS. Adv Math 168(2):225–253
77. Mauldin DR, Urbański M (2003) Graph directed Markov systems. Geometry and dynamics of limit sets Cambridge Tracts in Mathematics, 148. Cambridge University Press, Cambridge

78. Mauldin RD; Urbański M (2000) Parabolic iterated function systems. Ergod Theory Dynam Syst 20(5):1423–1447
79. McCluskey H, Manning A (1983) Hausdorff Dimension For Horseshoes. Erg Th Dyn Syst 3:251–260
80. Moran P (1946) Additive Functions Of Intervals and Hausdorff Dimension. Proceedings Of Cambridge Philosophical Society 42:15–23
81. Moreira C, Yoccoz J (2001) Stable Intersections of Cantor Sets with Large Hausdorff Dimension. Ann of Math (2) 154(1):45–96
82. Mãné R (1981) On the Dimension of Compact Invariant Sets for Certain Nonlinear Maps. Lecture Notes in Mathematics, vol 898. Springer
83. Mãné R (1990) The Hausdorff Dimension of Horseshoes of Surfaces. Bol Soc Bras Math 20:1–24
84. Neunhäuserer J (1999) An analysis of dimensional theoretical properties of some affine dynamical systems. Thesis. Free University Berlin, Berlin
85. Oseledets V (1968) A multiplicative ergodic theorem. Liapunov characteristic numbers for dynamical systems. Trans Moscow Math Soc 19:197–221
86. Ott E, Sauer T, Yorke J (1994) Part I Background. In: Coping with chaos. Wiley Ser Nonlinear Sci, Wiley, New York, pp 1–62
87. Palis J, Takens F (1987) Hyperbolicity And The Creation Of Homoclinic Orbits. Ann Math 125:337–374
88. Palis J, Takens F (1993) Hyperbolicity And Sensitive Chaotic Dynamics At Homoclinic Bifurcations. Cambridge U Press, Cambridge
89. Palis J, Takens F (1994) Homoclinic Tangencies For Hyperbolic Sets Of Large Hausdorff Dimension. Acta Math 172:91–136
90. Palis J, Viana M (1988) On the continuity of Hausdorff dimension and limit capacity for horseshoes. Lecture Notes in Math, vol 1331. Springer
91. Pesin Y (1977) Characteristic exponents and smooth ergodic theory. Russian Math Surveys 32(4):55–114
92. Pesin Y (1992) Dynamical systems with generalized hyperbolic attractors: hyperbolic, ergodic and topological properties. Erg Th Dyn Syst 12:123–151
93. Pesin Y (1993) On Rigorous Mathematical Definition of Correlation Dimension and Generalized Spectrum for Dimensions. J Statist Phys 71(3–4):529–547
94. Pesin Y (1997) Dimension Theory In Dynamical Systems: Rigorous Results And Applications. Cambridge U Press, Cambridge
95. Pesin Y (1997) Dimension theory in dynamical systems: contemporary views and applications. In: Chicago Lectures in Mathematics. Chicago University Press, Chicago
96. Pesin Y, Pitskel' B (1984) Topological pressure and the variational principle for noncompact sets. Funct Anal Appl 18:307–318
97. Pesin Y, Sadovskaya V (2001) Multifractal analysis of conformal axiom A flows. Comm Math Phys 216(2):277–312
98. Pesin Y, Tempelman A (1995) Correlation Dimension of Measures Invariant Under Group Actions. Random Comput Dyn 3(3):137–156
99. Pesin Y, Weiss H (1994) On the Dimension of Deterministic and Random Cantor-like Sets. Math Res Lett 1:519–529
100. Pesin Y, Weiss H (1996) On The Dimension Of Deterministic and Random Cantor-like Sets, Symbolic Dynamics, And The Eckmann-Ruelle Conjecture. Comm Math Phys 182:105–153
101. Pesin Y, Weiss H (1997) A Multifractal Analysis of Equilibrium Measures For Conformal Expanding Maps and Moran-like Geometric Constructions. J Stat Phys 86:233–275
102. Pesin Y, Weiss H (1997) The Multifractal Analysis of Gibbs Measures: Motivation. Mathematical Foundation and Examples. Chaos 7:89–106
103. Petersen K (1983) Ergodic theory. Cambridge studies in advanced mathematics 2. Cambridge Univ Press, Cambridge
104. Pollicott M, Weiss H (1994) The Dimensions Of Some Self Affine Limit Sets In The Plane And Hyperbolic Sets. J Stat Phys 77:841–866
105. Pollicott M, Weiss H (1999) Multifractal Analysis for the Continued Fraction and Manneville-Pomeau Transformations and Applications to Diophantine Approximation. Comm Math Phys 207(1):145–171
106. Przytycki F, Urbański M (1989) On Hausdorff Dimension of Some Fractal Sets. Studia Math 93:155–167
107. Ruelle D (1978) Thermodynamic Formalism. Addison-Wesley
108. Ruelle D (1982) Repellers For Real Analytic Maps. Erg Th Dyn Syst 2:99–107
109. Schmeling J (1994) Hölder Continuity of the Holonomy Maps for Hyperbolic basic Sets II. Math Nachr 170:211–225
110. Schmeling J (1997) Symbolic Dynamics for Beta-shifts and Self-Normal Numbers. Erg Th Dyn Syst 17:675–694
111. Schmeling J (1998) A dimension formula for endomorphisms – the Belykh family. Erg Th Dyn Syst 18:1283–1309
112. Schmeling J (1999) On the Completeness of Multifractal Spectra. Erg Th Dyn Syst 19:1–22
113. Schmeling J (2001) Entropy Preservation under Markov Codings. J Stat Phys 104(3–4):799–815
114. Schmeling J, Troubetzkoy S (1998) Dimension and invertibility of hyperbolic endomorphisms with singularities. Erg Th Dyn Syst 18:1257–1282
115. Schmeling J, Weiss H (2001) Dimension theory and dynamics. AMS Proceedings of Symposia in Pure Mathematics series 69:429–488
116. Series C (1985) The Modular Surface and Continued Fractions. J Lond Math Soc 31:69–80
117. Simon K (1997) The Hausdorff dimension of the general Smale–Williams solenoidwith different contraction coefficients. Proc Am Math Soc 125:1221–1228
118. Simpelaere D (1994) Dimension Spectrum of Axiom-A Diffeomorphisms, II. Gibbs Measures. J Stat Phys 76:1359–1375
119. Solomyak B (1995) On the random series $\sum \pm\lambda^n$ (an Erdös problem). Ann of Math (2) 142(3):611–625
120. Solomyak B (2004) Notes on Bernoulli convolutions. In: Fractal geometry and applications: a jubilee of Benoît Mandelbrot. Proc Sympos Pure Math, 72, Part 1. Amer Math Soc, Providence, pp 207–230
121. Stratmann B (1995) Fractal Dimensions for Jarnik Limit Sets of Geometrically Finite Kleinian Groups; The Semi-Classical Approach. Ark Mat 33:385–403
122. Takens F (1981) Detecting Strange Attractors in Turbulence. Lecture Notes in Mathematics, vol 898. Springer
123. Takens F, Verbitzki E (1999) Multifractal analysis of local entropies for expansive homeomorphisms with specification. Comm Math Phys 203:593–612
124. Walters P (1982) Introduction to Ergodic Theory. Springer
125. Weiss H (1992) Some Variational Formulas for Hausdorff Dimension, Topological Entropy, and SRB Entropy for Hyperbolic Dynamical System. J Stat Phys 69:879–886

126. Weiss H (1999) The Lyapunov Spectrum Of Equilibrium Measures for Conformal Expanding Maps and Axiom-A Surface Diffeomorphisms. J Statist Phys 95(3–4):615–632
127. Young LS (1981) Capacity of Attractors. Erg Th Dyn Syst 1:381–388
128. Young LS (1982) Dimension, Entropy, and Lyapunov Exponents. Erg Th Dyn Syt 2:109–124

Books and Reviews

Bowen R (1975) Equilibrium states and the ergodic theory of Anosov diffeomorphisms. Lecture Notes in Mathematics, vol 470. Springer
Eckmann JP, Ruelle D (1985) Ergodic Theory Of Chaos And Strange Attractors. Rev Mod Phys 57:617–656
Federer H (1969) Geometric measure theory. Springer
Hasselblatt B, Katok A (2002) Handbook of Dynamical Systems, vol 1, Survey 1. Principal Structures. Elsevier
Katok A (1980) Lyapunov exponents, entropy and periodic orbits for diffeomorphisms. Inst Hautes Études Sci Publ Math 51:137–173
Katok A, Hasselblatt B (1995) Introduction to the Modern Theory of Dynamical Systems. Cambridge Univ Press, Cambridge
Keller G (1998) Equilibrium states in ergodic theory. In: London Mathematical Society Student Texts 42. Cambridge University Press, Cambridge
Mañé R (1987) Ergodic theory and differentiable dynamics. In: Ergebnisse der Mathematik und ihrer Grenzgebiete 3, Band 8. Springer
Mario R, Urbański M (2005) Regularity properties of Hausdorff dimension in infinite conformal iterated function systems. Ergod Theory Dynam Syst 25(6):1961–1983
Mattila P (1995) Geometry of sets and measures in Euclidean spaces. In: Fractals and rectifiability. Cambridge University Press, Cambridge
Pugh C, Shub M (1989) Ergodic attractors. Trans Amer Math Soc 312(1):1–54
Takens F (1988) Limit capacity and Hausdorff dimension of dynamically defined Cantor sets. Lecture Notes in Math, vol 1331. Springer

Ergodic Theory on Homogeneous Spaces and Metric Number Theory

Dmitry Kleinbock
Department of Mathematics, Brandeis University, Waltham, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Basic Facts
Connection with Dynamics on the Space of Lattices
Diophantine Approximation with Dependent Quantities: The Set-Up
Further Results
Future Directions
Acknowledgment
Bibliography

Glossary

Diophantine approximation Diophantine approximation refers to approximation of real numbers by rational numbers, or more generally, finding integer points at which some (possibly vector-valued) functions attain values close to integers.

Metric number theory Metric number theory (or, specifically, metric Diophantine approximation) refers to the study of sets of real numbers or vectors with prescribed Diophantine approximation properties.

Homogeneous spaces A homogeneous space G/Γ of a group G by its subgroup Γ is the space of cosets $\{g\Gamma\}$. When G is a Lie group and Γ is a discrete subgroup, the space G/Γ is a smooth manifold and locally looks like G itself.

Lattice; unimodular lattice A lattice in a Lie group is a discrete subgroup of finite covolume; unimodular stands for covolume equal to 1.

Ergodic theory The study of statistical properties of orbits in abstract models of dynamical systems.

Hausdorff dimension A nonnegative number attached to a metric space and extending the notion of topological dimension of "sufficiently regular" sets, such as smooth submanifolds of real Euclidean spaces.

Definition of the Subject

The theory of Diophantine approximation, named after Diophantus of Alexandria, in its simplest set-up deals with the approximation of real numbers by rational numbers. Various higher-dimensional generalizations involve studying values of linear or polynomial maps at integer points. Often a certain "approximation property" is fixed, and one wants to characterize the set of numbers (vectors, matrices) which share this property, by means of certain measures (Lebesgue, or Hausdorff, or some other interesting measures). This is usually referred to as *metric* Diophantine approximation.

The starting point for the theory is an elementary fact that $\mathbb{Q}$, the set of rational numbers, is dense in $\mathbb{R}$, the reals. In other words, every real number can be approximated by rationals: for any $y \in \mathbb{R}$ and any $\varepsilon > 0$ there ex-

ists $p/q \in \mathbb{Q}$ with

$$|y - p/q| < \varepsilon \,. \tag{1}$$

To answer questions like "how well can various real numbers be approximated by rational numbers? i. e., how small can ε in (1) be chosen for varying $p/q \in \mathbb{Q}$?", a natural approach has been to compare the accuracy of the approximation of y by p/q to the "complexity" of the latter, which can be measured by the size of its denominator q in its reduced form. This seemingly simple set-up has led to introducing many important Diophantine approximation properties of numbers/vectors/matrices, which show up in various fields of mathematics and physics, such as differential equations, KAM theory, transcendental number theory.

Introduction

As the first example of refining the statement about the density of $\mathbb{Q}$ in $\mathbb{R}$, consider a theorem by Kronecker stating that for any $y \in \mathbb{R}$ and any $c > 0$, there exist infinitely many $q \in \mathbb{Z}$ such that

$$|y - p/q| < c/|q| \quad \text{i. e. } |qy - p| < c \tag{2}$$

for some $p \in \mathbb{Z}$. A comparison of (1) and (2) shows that it makes sense to multiply both sides of (1) by q, since in the right hand side of (2) one would still be able to get very small numbers. In other words, approximation of y by p/q translates into approximating integers by integer multiples of y.

Also, if y is irrational, (p, q) can be chosen to be relatively prime, i. e. one gets infinitely many different rational numbers p/q satisfying (2). However if $y \in \mathbb{Q}$ the latter is no longer true for small enough c. Thus it seems to be more convenient to talk about pairs (p, q) rather than $p/q \in \mathbb{Q}$, avoiding a necessity to consider the two cases separately.

At this point it is convenient to introduce the following central definition: if ψ is a function $\mathbb{N} \to \mathbb{R}_+$ and $y \in \mathbb{R}$, say that y is *ψ-approximable* (notation: $y \in \mathcal{W}(\psi)$) if there exist infinitely many $q \in \mathbb{N}$ such that

$$|qy - p| < \psi(q) \tag{3}$$

for some $p \in \mathbb{Z}$. Because of Kronecker's Theorem, it is natural to assume that $\psi(x) \to 0$ as $x \to \infty$. Often ψ will be assumed non-increasing, although many results do not require monotonicity of ψ.

One can similarly consider a higher-dimensional version of the above set-up. Note that $y \in \mathbb{R}$ in the above formulas plays the role of a linear map from $\mathbb{R}$ to another copy of $\mathbb{R}$, and one asks how close values of this map at integers are from integers. It is natural to generalize it by taking a linear operator Y from $\mathbb{R}^n$ to $\mathbb{R}^m$ for fixed $m, n \in \mathbb{N}$, that is, an $m \times n$-matrix (interpreted as a system of m linear forms Y_i on $\mathbb{R}^n$). We will denote by $M_{m,n}$ the space of $m \times n$ matrices with real coefficients. For ψ as above, one says that $Y \in M_{m,n}$ is *ψ-approximable* (notation: $Y \in \mathcal{W}_{m,n}(\psi)$) if there are infinitely many $\mathbf{q} \in \mathbb{Z}^n$ such that

$$\|Y\mathbf{q} + \mathbf{p}\| \le \psi(\|\mathbf{q}\|) \tag{4}$$

for some $\mathbf{p} \in \mathbb{Z}^m$. Here $\|\cdot\|$ is the supremum norm on $\mathbb{R}^k$ given by $\|\mathbf{y}\| = \max_{1 \le i \le k} |y_i|$. (This definition is slightly different from the one used in [58], where powers of norms were considered).

Traditionally, one of the main goals of metric Diophantine approximation has been to understand how big the sets $\mathcal{W}_{m,n}(\psi)$ are for fixed m, n and various functions ψ. Of course, (4) is not the only interesting condition that can be studied; various modifications of the approximation properties can also be considered. For example the Oppenheim Conjecture, now a theorem of Margulis [69] and a basis for many important recent developments [22,34,35], states that indefinite irrational quadratic forms can take arbitrary small values at integer points; Littlewood's conjecture, see (18) below, deals with a similar statement about products of linear forms. See the article ► Ergodic Theory: Rigidity by Nitica and surveys [32,70] for details.

We remark that the standard tool for studying Diophantine approximation properties of real numbers ($m = n = 1$) is the continued fraction expansion, or, equivalently, the Gauss map $x \mapsto 1/x \mod 1$ of the unit interval, see [49]. However the emphasis of this survey lies in higher-dimensional theory, and the dynamical system described below can be thought of as a replacement for the continued fraction technique applicable in the one-dimensional case. Additional details about interactions between ergodic theory and number theory can be found in the article by Nitica mentioned above, in ► Ergodic Theory: Recurrence by Frantzikinakis and McCutcheon and ► Ergodic Theory: Interactions with Combinatorics and Number Theory by Ward, as well as in the survey papers [32,33,50,58,66,70,71].

Here is a brief outline of the rest of the article. In the next section we survey basic results, some classical, some obtained relatively recently, in metric Diophantine approximation. Sect. "Connection with Dynamics on the Space of Lattices" is devoted to a description of the connection between Diophantine approximation and dynamics, specifically flows on the space of

lattices. In Sect. "Diophantine Approximation with Dependent Quantities: The Set-Up" and Sect. "Further Results", we specialize to the set-up of Diophantine approximation on manifolds, or, more generally, approximation properties of vectors with respect to measures satisfying some natural conditions, and show how applications of homogeneous dynamics contributed to important recent developments in the field. Sect. "Future Directions" mentions several open questions and directions for further investigation.

Basic Facts

General references for this section: [17,80].

The simplest choice for functions ψ happens to be the following: let us denote $\psi_{c,\nu}(x) = cx^{-\nu}$. It was shown by Dirichlet in 1842 that with the choice $c = 1$ and $\nu = n/m$, all $Y \in M_{m,n}$ are ψ-approximable. Moreover, Dirichlet's Theorem states that for any $Y \in M_{m,n}$ and for any $t > 0$ there exist $\mathbf{q} = (q_1, \dots, q_n) \in \mathbb{Z}^n \smallsetminus \{0\}$ and $\mathbf{p} = (p_1, \dots, p_m) \in \mathbb{Z}^m$ satisfying the following system of inequalities:

$$\|Y\mathbf{q} - \mathbf{p}\| < \mathrm{e}^{-t/m} \quad \text{and} \quad \|\mathbf{q}\| \le \mathrm{e}^{t/n} . \tag{5}$$

From this it easily follows that $\mathcal{W}_{m,n}(\psi_{1,n/m}) = M_{m,n}$. In fact, it is this paper of Dirichlet which gave rise to his box principle. Later another proof of the same result was given by Minkowski. The constant $c = 1$ is not optimal: the smallest value of c for which $\mathcal{W}_{1,1}(\psi_{c,1}) = \mathbb{R}$ is $1/\sqrt{5}$, and the optimal constants are not known in higher dimensions, although some estimates can be given [80].

Systems of linear forms which do not belong to $\mathcal{W}_{m,n}(\psi_{c,n/m})$ for some positive c are called *badly approximable*; that is, we set

$$\mathrm{BA}_{m,n} \stackrel{\text{def}}{=} M_{m,n} \smallsetminus \cup_{c>0} \mathcal{W}_{m,n}(\psi_{c,n/m}) .$$

Their existence in arbitrary dimensions was shown by Perron. Note that a real number y ($m = n = 1$) is badly approximable if and only if its continued fraction coefficients are uniformly bounded. It was proved by Jarnik [46] in the case $m = n = 1$ and by Schmidt in the general case [78] that badly approximable matrices form a set of full Hausdorff dimension: that is, $\dim(\mathrm{BA}_{m,n}) = mn$.

On the other hand, it can be shown that each of the sets $\mathcal{W}_{m,n}(\psi_{c,n/m})$ for any $c > 0$ has full Lebesgue measure, and hence the complement $\mathrm{BA}_{m,n}$ to their intersection has measure zero. This is a special case of a theorem due to Khintchine [48] in the case $n = 1$ and to Groshev [42] in full generality, which gives the precise condition on the function ψ under which the set of ψ-approximable matrices has full measure. Namely, if ψ is non-increasing (this assumption can be removed in higher dimensions but not for $n = 1$, see [29]), then λ-almost no (resp. λ-almost every) $Y \in M_{m,n}$ is ψ-approximable, provided the sum

$$\sum_{k=1}^{\infty} k^{n-1} \psi(k)^m \tag{6}$$

converges (resp. diverges). (Here and hereafter λ stands for Lebesgue measure). This statement is usually referred to as the Khintchine–Groshev Theorem. The convergence case of this theorem follows in a straightforward manner from the Borel–Cantelli Lemma, but the divergence case is harder. It was reproved and sharpened in 1960 by Schmidt [76], who showed that if the sum (6) diverges, then for almost all Y the number of solutions to (4) with $\|\mathbf{q}\| \le N$ is asymptotic to the partial sum of the series (6) (up to a constant), and also gave an estimate for the error term.

A special case of the convergence part of the theorem shows that $\mathcal{W}_{m,n}(\psi_{1,\nu})$ has measure zero whenever $\nu > n/m$. Y is said to be *very well approximable* if it belongs to $\mathcal{W}_{m,n}(\psi_{1,\nu})$ for some $\nu > n/m$. That is,

$$\mathrm{VWA}_{m,n} \stackrel{\text{def}}{=} \cup_{\nu > n/m} \mathcal{W}_{m,n}(\psi_{1,\nu}) .$$

More specifically, let us define the *Diophantine exponent* $\omega(Y)$ of Y (sometimes called "the exact order" of Y) to be the supremum of $\nu > 0$ for which $Y \in \mathcal{W}_{m,n}(\psi_{1,\nu})$. Then $\omega(Y)$ is always not less than n/m, and is equal to n/m for Lebesgue-a.e. Y; in fact, $\mathrm{VWA}_{m,n} = \{Y \in M_{m,n} : \omega(Y) > n/m\}$.

The Hausdorff dimension of the null sets $\mathcal{W}_{m,n}(\psi_{1,\nu})$ was computed independently by Besicovitch [14] and Jarnik [45] in the one-dimensional case and by Dodson [26] in general: when $\nu > n/m$, one has

$$\dim\left(\mathcal{W}_{m,n}(\psi_{1,\nu})\right) = (n-1)m + \frac{m+n}{\nu+1} . \tag{7}$$

See [27] for a nice exposition of ideas involved in the proof of both the aforementioned formula and the Khintchine–Groshev Theorem.

Note that it follows from (7) that the null set $\mathrm{VWA}_{m,n}$ has full Hausdorff dimension. Matrices contained in the intersection

$$\bigcap_{\nu} \mathcal{W}_{m,n}(\psi_{1,\nu}) = \{Y \in M_{m,n} : \omega(Y) = \infty\}$$

are called *Liouville* and form a set of Hausdorff dimension $(n-1)m$, that is, to the dimension of Y for which $Y\mathbf{q} \in \mathbb{Z}$ for some $\mathbf{q} \in \mathbb{Z}^n \smallsetminus \{0\}$ (the latter belong to $\mathcal{W}_{m,n}(\psi)$ for any positive ψ).

Note also that the aforementioned properties behave nicely with respect to transposition; this is described by the so-called Khintchine's Transference Principle (Chap. V in [17]). For example, $Y \in \mathrm{BA}_{m,n}$ if and only if $Y^T \in \mathrm{BA}_{n,m}$, and $Y \in \mathrm{VWA}_{m,n}$ if and only if $Y^T \in \mathrm{VWA}_{n,m}$. In particular, many problems related to approximation properties of vectors ($n = 1$) and linear forms ($m = 1$) reduce to one another.

We refer the readers to [43] and [8] for very detailed and comprehensive recent accounts of various further aspects of the theory.

Connection with Dynamics on the Space of Lattices

General references for this section: [4,86].

Interactions between Diophantine approximation and the theory of dynamical systems has a long history. Already in Kronecker's Theorem one can see a connection. Indeed, the statement of the theorem can be rephrased as follows: the points on the orbit of 0 under the rotation of the circle $\mathbb{R}/\mathbb{Z}$ by y approach the initial point 0 arbitrarily closely. This is a special case of the Poincare Recurrence Theorem in measurable dynamics. And, likewise, all the aforementioned properties of $Y \in M_{m,n}$ can be restated in terms of recurrence properties of the $\mathbb{Z}^n$-action on the m-dimensional torus $\mathbb{R}^m/\mathbb{Z}^m$ given by $\mathbf{x} \mapsto Y\mathbf{x} \mod \mathbb{Z}^m$. In other words, fixing Y gives rise to a dynamical system in which approximation properties of Y show up.

However the theme of this section is a different dynamical system, whose phase space is (essentially) the space of parameters Y, and which can be used to read the properties of Y from the behavior of the associated trajectory.

It has been known for a long time (see [81] for a historical account) that Diophantine properties of real numbers can be coded by the behavior of geodesics on the quotient of the hyperbolic plane by $\mathrm{SL}_2(\mathbb{Z})$. In fact, the latter flow can be viewed as the suspension flow of the Gauss map mentioned at the end of Sect. "Introduction". There have been many attempts to construct a higher-dimensional analogue of the Gauss map so that it captures all the features of simultaneous approximation, see [47,63,65] and references therein. On the other hand, it seems to be more natural and efficient to generalize the suspension flow itself, and this is where one needs higher rank homogeneous dynamics.

As was mentioned above, in the basic set-up of simultaneous Diophantine approximation one takes a system of m linear forms $Y_1, \dots, Y_m$ on $\mathbb{R}^n$ and looks at the values of $|Y_i(\mathbf{q}) + p_i|$, $p_i \in \mathbb{Z}$, when $\mathbf{q} = (q_1, \dots, q_n) \in \mathbb{Z}^n$ is far from 0. The trick is to put together

$$Y_1(\mathbf{q}) + p_1, \dots, Y_m(\mathbf{q}) + p_m \quad \text{and} \quad q_1, \dots, q_n,$$

and consider the collection of vectors

$$\left\{ \begin{pmatrix} Y\mathbf{q} + \mathbf{p} \\ \mathbf{q} \end{pmatrix} \middle| \, \mathbf{p} \in \mathbb{Z}^m, \, \mathbf{q} \in \mathbb{Z}^n \right\} = L_Y \mathbb{Z}^k$$

where $k = m + n$ and

$$L_Y \stackrel{\text{def}}{=} \begin{pmatrix} I_m & Y \\ 0 & I_n \end{pmatrix}, \quad Y \in M_{m,n} \,. \tag{8}$$

This collection is a unimodular lattice in $\mathbb{R}^k$, that is, a discrete subgroup of $\mathbb{R}^k$ with covolume 1. Our goal is to keep track of vectors in such a lattice having small projections onto the first m components of $\mathbb{R}^k$ and big projections onto the last n components. This is where dynamics comes into the picture. Denote by g_t the one-parameter subgroup of $\mathrm{SL}_k(\mathbb{R})$ given by

$$g_t = \mathrm{diag}(\underbrace{\mathrm{e}^{t/m}, \dots, \mathrm{e}^{t/m}}_{m\text{times}}, \underbrace{\mathrm{e}^{-t/n}, \dots, \mathrm{e}^{-t/n}}_{n\text{times}}) \,. \tag{9}$$

The vectors in the lattice $L_Y \mathbb{Z}^k$ are moved by the action of g_t, $t > 0$, and a special role is played by the moment t when the "small" and "big" projections equalize.

That is, one is led to consider a new dynamical system. Its phase space is the space of unimodular lattices in $\mathbb{R}^k$, which can be naturally identified with the homogeneous space

$$\Omega_k \stackrel{\text{def}}{=} G/\Gamma, \quad \text{where } G = \mathrm{SL}_k(\mathbb{R}) \text{ and } \Gamma = \mathrm{SL}_k(\mathbb{Z}) \,, \tag{10}$$

and the action is given by left multiplication by elements of the subgroup (9) of G, or perhaps other subgroups $H \subset G$. Study of such systems has a rich history; for example, they are known to be ergodic and mixing whenever H is unbounded [74]. What is important in this particular case is that the space Ω_k happens to be noncompact, and its structure at infinity is described via Mahler's Compactness Criterion, see Chap. V in [4]: a sequence of lattices $g_i \mathbb{Z}^k$ goes to infinity in $\Omega_k \iff$ there exists a sequence $\{\mathbf{v}_i \in \mathbb{Z}^k \smallsetminus \{0\}\}$ such that $g_i(\mathbf{v}_i) \to 0$ as $i \to \infty$. Equivalently, for $\varepsilon > 0$ consider a subset K_ε of Ω_k consisting of lattices with no nonzero vectors of norm less than ε; then all the sets K_ε are compact, and every compact subset of Ω_k is contained in one of them. Moreover, one can choose a metric on Ω_k such that $\mathrm{dist}(\Lambda, \mathbb{Z}^k)$ is, up to a uniform multiplicative constant,

equal to $-\log \min_{\mathbf{v}\in\Lambda\smallsetminus\{0\}} \|\mathbf{v}\|$ (see [25]); then the length of the smallest nonzero vector in a lattice Λ will determine how far away is this lattice in the "cusp" of Ω_k.

Using Mahler's Criterion, it is not hard to show that $Y \in \mathrm{BA}_{m,n}$ if and only if the trajectory

$$\{g_t L_Y \mathbb{Z}^k : t \in \mathbb{R}_+\} \tag{11}$$

is bounded in Ω_k. This was proved by Dani [20] in 1985, and later generalized in [57] to produce a criterion for Y to be ψ-approximable for any non-increasing function ψ. An important special case is a criterion for a system of linear forms to be very well approximable: $Y \in \mathrm{VWA}_{m,n}$ if and only if the trajectory (11) has linear growth, that is, there exists a positive γ such that $\mathrm{dist}(g_t L_Y \mathbb{Z}^k, \mathbb{Z}^k) > \gamma t$ for an unbounded set of $t > 0$.

This correspondence allows one to link various Diophantine and dynamical phenomena. For example, from the results of [55] on abundance of bounded orbits on homogeneous spaces one can deduce the aforementioned theorem of Schmidt [78]: the set $\mathrm{BA}_{m,n}$ has full Hausdorff dimension. And a dynamical Borel–Cantelli Lemma established in [57] can be used for an alternative proof of the Khintchine–Groshev Theorem; see also [87] for an earlier geometric approach. Note that both proofs are based on the following two properties of the g_t-action: mixing, which forces points to return to compact subsets and makes preimages of cusp neighborhoods quasi-independent, and hyperbolicity, which implies that the behavior of points on unstable leaves is generic. The latter is important since the orbits of the group $\{L_Y \mathbb{Z}^k : Y \in M_{m,n}\}$ are precisely the unstable leaves with respect to the g_t-action.

We note that other types of Diophantine problems, such as conjectures of Oppenheim and Littlewood mentioned in the previous section, can be reduced to statements involving Ω_k by means of the same principle: Mahler's Criterion is used to relate small values of some function at integer points to excursions to infinity in Ω_k of orbit of the stabilizer of this function.

Other important and useful recent applications of homogeneous dynamics to metric Diophantine approximation are related to the circle of ideas roughly called "Diophantine approximation with dependent quantities" (terminology borrowed from [84]), to be surveyed in the next two sections.

Diophantine Approximation with Dependent Quantities: The Set-Up

General references for this section: [12,84].

Here we restrict ourselves to Diophantine properties of vectors in $\mathbb{R}^n$. In particular, we will look more closely at the set of very well approximable vectors, which we will simply denote by VWA, dropping the subscripts. In many cases it does not matter whether one works with row or column vectors, in view of the duality remark made at the end of Sect. "Basic Facts".

We begin with a non-example of an application of dynamics to Diophantine approximation: a celebrated and difficult theorem which currently, to the best of the author's knowledge, has no dynamical proof. Suppose that $\mathbf{y} = (y_1, \dots, y_n) \in \mathbb{R}^n$ is such that each y_i is algebraic and $1, y_1, \dots, y_n$ are linearly independent over $\mathbb{Q}$. It was established by Roth for $n = 1$ [75] and then generalized to arbitrary n by Schmidt [79], that $\mathbf{y}$ as above necessarily belongs to the complement of VWA. In other words, vectors with very special algebraic properties happen to follow the behavior of a generic vector in $\mathbb{R}^n$.

We would like to view the above example as a special case of a general class of problems. Namely, suppose we are given a Radon measure μ on $\mathbb{R}^n$. Let us say that μ is *extremal* [85] if μ-a.e. $\mathbf{y} \in \mathbb{R}^n$ is not very well approximable. Further, define the *Diophantine exponent* $\omega(\mu)$ of μ to be the μ-essential supremum of the function $\omega(\cdot)$; in other words,

$$\omega(\mu) \stackrel{\text{def}}{=} \sup\big\{\nu \,|\, \mu\big(\mathcal{W}(\psi_{1,\nu})\big) > 0\big\}\,.$$

Clearly it only depends on the measure class of μ. If μ is naturally associated with a subset $\mathcal{M}$ of $\mathbb{R}^n$ supporting μ (for example, if $\mathcal{M}$ is a smooth submanifold of $\mathbb{R}^n$ and μ is the measure class of the Riemannian volume on $\mathcal{M}$, or, equivalently, the pushforward $\mathbf{f}_*\lambda$ of λ by a smooth map $\mathbf{f}$ parametrizing $\mathcal{M}$), one defines the Diophantine exponent $\omega(\mathcal{M})$ of $\mathcal{M}$ to be equal to that of μ, and says that $\mathcal{M}$ is extremal if $\mathbf{f}(\mathbf{x})$ is not very well approximable for λ-a.e. $\mathbf{x}$.

Then $\omega(\mu) \geq n$ for any μ, and $\omega(\lambda) = \omega(\mathbb{R}^n)$ is equal to n. The latter justifies the use of the word "extremal": μ is *extremal* if $\omega(\mu)$ is equal to n, i. e. attains the smallest possible value. The aforementioned results of Roth and Schmidt then can be interpreted as the extremality of atomic measures supported on algebraic vectors without rational dependence relations.

Historically, the first measure (other than λ) to be considered in the set-up described above was the pushforward of λ by the map

$$\mathbf{f}(x) = (x, x^2, \dots, x^n)\,. \tag{12}$$

The extremality of $\mathbf{f}_*\lambda$ for $\mathbf{f}$ as above was conjectured in 1932 by K. Mahler [67] and proved in 1964 by Sprindžuk [82,83]. It was important for Mahler's study of transcendental numbers: this result, roughly speaking, says

that almost all transcendental numbers are "not very algebraic". At about the same time Schmidt [77] proved the extremality of $\mathbf{f}_*\lambda$ when $\mathbf{f}\colon I \to \mathbb{R}^2$, $I \subset \mathbb{R}$, is C^3 and satisfies

$$\begin{vmatrix} f_1'(x) & f_2'(x) \\ f_1''(x) & f_2''(x) \end{vmatrix} \neq 0 \quad \text{for } \lambda\text{-a.e. } x \in I\,;$$

in other words, the curve parametrized by $\mathbf{f}$ has nonzero curvature at almost all points. Since then, a lot of attention has been devoted to showing that measures $\mathbf{f}_*\lambda$ are extremal for other smooth maps $\mathbf{f}$.

To describe a broader class of examples, recall the following definition. Let $\mathbf{x} \in \mathbb{R}^d$ and let $\mathbf{f} = (f_1, \dots, f_n)$ be a C^k map from a neighborhood of $\mathbf{x}$ to $\mathbb{R}^n$. Say that $\mathbf{f}$ is *nondegenerate at* $\mathbf{x}$ if $\mathbb{R}^n$ is spanned by partial derivatives of $\mathbf{f}$ at $\mathbf{x}$ up to some order. Say that $\mathbf{f}$ is *nondegenerate* if it is nondegenerate at λ-a.e. $\mathbf{x}$. It was conjectured by Sprindžuk [84] in 1980 that $\mathbf{f}_*\lambda$ for real analytic nondegenerate $\mathbf{f}$ are extremal. Many special cases were established since then (see [12] for a detailed exposition of the theory and many related results), but the general case stood open until the mid-1990s [56], when Sprindžuk's conjecture was proved using the dynamical approach (later Beresnevich [6] succeeded in establishing and extending this result without use of dynamics). The proof in [56] uses the correspondence outlined in the previous section plus a measure estimate for flows on the space of lattices which is described below.

In the subsequent work the method of [56] was adapted to a much broader class of measures. To define it we need to introduce some more notation and definitions. If $\mathbf{x} \in \mathbb{R}^d$ and $r > 0$, denote by $B(\mathbf{x}, r)$ the open ball of radius r centered at x. If $B = B(\mathbf{x}, r)$ and $c > 0$, cB will denote the ball $B(\mathbf{x}, cr)$. For $B \subset \mathbb{R}^d$ and a real-valued function f on B, let

$$\|f\|_B \stackrel{\text{def}}{=} \sup_{\mathbf{x}\in B} |f(\mathbf{x})|\,.$$

If ν is a measure on $\mathbb{R}^d$ such that $\nu(B) > 0$, define $\|f\|_{\nu,B} \stackrel{\text{def}}{=} \|f\|_{B\cap \operatorname{supp}\nu}$; this is the same as the $L^\infty(\nu)$-norm of $f|_B$ if f is continuous and B is open. If $D > 0$ and $U \subset \mathbb{R}^d$ is an open subset, let us say that ν is *D-Federer on U* if for any ball $B \subset U$ centered at $\operatorname{supp}\nu$ one has $\frac{\nu(3B)}{\nu(B)} < D$ whenever $3B \subset U$. This condition is often called "doubling" in the literature. See [54,72] for examples and references. ν is called *Federer* if for ν-a.e. $\mathbf{x} \in \mathbb{R}^d$ there exist a neighborhood U of x and $D > 0$ such that ν is D-Federer on U.

Given $C, \alpha > 0$, open $U \subset \mathbb{R}^d$ and a measure ν on U, a function $f\colon U \to \mathbb{R}$ is called *(C,α)-good on U with respect to* ν if for any ball $B \subset U$ centered in $\operatorname{supp}\nu$ and any $\varepsilon > 0$ one has

$$\nu\big(\{\mathbf{x} \in B\colon |f(\mathbf{x})| < \varepsilon\}\big) \le C\left(\frac{\varepsilon}{\|f\|_{\nu,B}}\right)^\alpha \nu(B)\,. \tag{13}$$

This condition was formally introduced in [56] for ν being Lebesgue measure, and in [54] for arbitrary ν. A basic example is given by polynomials, and the upshot of the above definition is the formalization of a property needed for the proof of several basic facts [19,21,68] about polynomial maps into the space of lattices.

In [54] a strengthening of this property was considered: f was called *absolutely (C,α)-good on U with respect to* ν if for B and ε as above one has

$$\nu\big(\{\mathbf{x} \in B : |f(\mathbf{x})| < \varepsilon\}\big) \le C\left(\frac{\varepsilon}{\|f\|_{B}}\right)^\alpha \nu(B)\,. \tag{14}$$

There is no difference between (13) and (14) when ν has full support, but it turns out to be useful for describing measures supported on proper (e. g. fractal) subsets of $\mathbb{R}^d$.

Now suppose that we are given a measure ν on $\mathbb{R}^d$, an open $U \subset \mathbb{R}^d$ with $\nu(U) > 0$ and a map $\mathbf{f} = (f_1, \dots, f_n)\colon \mathbb{R}^d \to \mathbb{R}^n$. Following [62], say that a pair $(\mathbf{f}, \nu)$ is *(absolutely) good on U* if any linear combination of $1, f_1, \dots, f_n$ is (absolutely) (C,α)-good on U with respect to ν. If for ν-a.e. $\mathbf{X}$ there exists a neighborhood U of $\mathbf{X}$ and $C, \alpha > 0$ such that ν is (absolutely) (C,α)-good on U, we will say that the pair $(\mathbf{f}, \nu)$ is *(absolutely) good*.

Another relevant notion is the nonplanarity of $(\mathbf{f}, \nu)$. Namely, $(\mathbf{f}, \nu)$ is said to be *nonplanar* if whenever B is a ball with $\nu(B) > 0$, the restrictions of $1, f_1, \dots, f_n$ to $B \cap \operatorname{supp}\nu$ are linearly independent over $\mathbb{R}$; in other words, $\mathbf{f}(B \cap \operatorname{supp}\nu)$ is not contained in any proper affine subspace of $\mathbb{R}^n$. Note that absolutely good implies both good and nonplanar, but the converse is in general not true.

Many examples of (absolutely) good and nonplanar pairs $(\mathbf{f}, \nu)$ can be found in the literature. Already the case $n = d$ and $\mathbf{f} = \mathrm{Id}$ is very interesting. A measure μ on $\mathbb{R}^n$ is said to be *friendly* (resp., *absolutely friendly*) if and only if it is Federer and the pair (Id, μ) is good and nonplanar (resp., absolutely good). See [54,88,89] for many examples. An important class of measures is given by limit measures of irreducible system of self-similar or self-conformal contractions satisfying the Open Set Condition [44]; those are shown to be absolutely friendly in [54]. The prime example is the middle-third Cantor set on the real line. The term "friendly" was cooked up as a loose abbreviation for "Federer, nonplanar and decaying", and later proved to be particularly friendly in dealing with problems arising in metric number theory, see e. g. [36].

Also let us say that a pair $(\mathbf{f}, \nu)$ is *nondegenerate* if $\mathbf{f}$ is nondegenerate at ν-a.e. $\mathbf{X}$. When ν is Lebesgue measure on $\mathbb{R}^d$, it is proved in Proposition 3.4 in [56], that a nondegenerate $(\mathbf{f}, \nu)$ is good and nonplanar. The same conclusion is derived in Proposition 7.3 in [54], assuming that ν is absolutely friendly. Thus volume measures on smooth nondegenerate manifolds are friendly, but not absolutely friendly.

It turns out that all the aforementioned examples of measures can be proved to be extremal by a generalization of the argument from [56]. Specifically, let ν be a Federer measure on $\mathbb{R}^d$, U an open subset of $\mathbb{R}^d$, and $\mathbf{f}\colon U \to \mathbb{R}^n$ a continuous map such that the pair $(\mathbf{f}, \nu)$ is good and nonplanar; then $\mathbf{f}_*\nu$ is extremal. This can be derived from the Borel–Cantelli Lemma, the correspondence described in the previous section, and the following measure estimate: if ν, U and $\mathbf{f}$ are as above, then for ν-a.e. $\mathbf{x}_0 \in U$ there exists a ball $B \subset U$ centered at $\mathbf{x}_0$ and $\tilde{C}, \alpha > 0$ such that for any $t \in \mathbb{R}_+$ and any $\varepsilon > 0$,

$$\nu\big(\{\mathbf{x} \in B : g_t L_{\mathbf{f}(\mathbf{x})}\mathbb{Z}^{n+1} \notin K_\varepsilon\}\big) < \tilde{C}\varepsilon^\alpha \,. \tag{15}$$

Here g_t is as in (9) with $m = 1$ (assuming that the row vector viewpoint is adopted). This is a quantitative way of saying that for fixed t, the "flow" $\mathbf{x} \mapsto g_t L_{\mathbf{f}(\mathbf{x})}\mathbb{Z}^{n+1}$, $B \to \Omega_{n+1}$, cannot diverge, and in fact must spend a big (uniformly in t) proportion of time inside compact sets K_ε.

The inequality (15) is derived from a general "quantitative non-divergence" estimate, which can be thought of a substantial generalization of theorems of Margulis and Dani [19,21,68] on non-divergence of unipotent flows on homogeneous spaces. One of its most general versions [54] deals with a measure ν on $\mathbb{R}^d$, a continuous map $h\colon \widetilde{B} \to G$, where $\widetilde{B}$ is a ball in $\mathbb{R}^d$ centered at $\operatorname{supp}\nu$ and G is as in (10). To describe the assumptions on h, one needs to employ the combinatorial structure of lattices in $\mathbb{R}^k$, and it will be convenient to use the following notation: if V is a nonzero rational subspace of $\mathbb{R}^k$ and $g \in G$, define $\ell_V(g)$ to be the covolume of $g(V \cap \mathbb{Z}^k)$ in gV. Then, given positive constants C, D, α, there exists $C_1 = C_1(d, k, C, \alpha, D) > 0$ with the following property. Suppose ν is D-Federer on $\widetilde{B}$, $0 < \rho \le 1$, and h is such that for each rational $V \subset \mathbb{R}^k$

(i) $\ell_V \circ h$ is (C, α)-good on $\tilde{B}$ with respect to ν, and
(ii) $\|\ell_V \circ h\|_{\nu,B} \ge \rho$, where $B = 3^{-(k-1)}\tilde{B}$. Then
(iii) for any positive $\varepsilon \le \rho$, one has

$$\nu\big(\{\mathbf{x} \in B : h(\mathbf{x})\mathbb{Z}^k \notin K_\varepsilon\}\big) \le C_1(\varepsilon\rho)^\alpha \nu(B) \,. \tag{16}$$

Taking $h(\mathbf{x}) = g_t L_{\mathbf{f}(\mathbf{x})}$ and unwinding the definitions of good and nonplanar pairs, one can show that (i) and (ii) can be verified for some balls B centered at ν-almost every point, and derive (15) from (16).

Further Results

The approach to metric Diophantine approximation using quantitative non-divergence, that is, the implication (i) + (ii) $\Rightarrow$ (iii), is not omnipotent. In particular, it is difficult to use when more precise results are needed, such as for example computing/estimating the Hausdorff dimension of the set of $\psi_{1,\nu}$-approximable vectors on a manifold. See [9,10] for such results. On the other hand, the dynamical approach can often treat much more general objects that its classical counterpart, and also can be perturbed in a lot of directions, producing many generalizations and modifications of the main theorems from the preceding section.

One of the most important of them is the so-called *multiplicative* version of the set-up of Sect. "Diophantine Approximation with Dependent Quantities: The Set-Up". Namely, define functions $\Pi(\mathbf{x}) \stackrel{\text{def}}{=} \prod_i |x_i|$ and $\Pi_+(\mathbf{x}) \stackrel{\text{def}}{=} \prod_i \max(|x_i|, 1)$ Then, given a function $\psi\colon \mathbb{N} \to \mathbb{R}_+$, one says that $Y \in M_{m,n}$ is *multiplicatively ψ-approximable* (notation: $Y \in \mathcal{W}^\times_{m,n}(\psi)$) if there are infinitely many $\mathbf{q} \in \mathbb{Z}^n$ such that

$$\Pi(Y\mathbf{q} + \mathbf{p})^{1/m} \le \psi\big(\Pi_+(\mathbf{q})^{1/n}\big) \tag{17}$$

for some $\mathbf{p} \in \mathbb{Z}^m$. Since $\Pi(\mathbf{x}) \le \Pi_+(\mathbf{x}) \le \|\mathbf{x}\|^k$ for $\mathbf{x} \in \mathbb{R}^k$, any ψ-approximable Y is multiplicatively ψ-approximable; but the converse is in general not true, see e. g. [37]. However if one, as before, considers the family $\{\psi_{1,\nu}\}$, the critical parameter for which the drop from full measure to measure zero occurs is again n/m. That is, if one defines the *multiplicative Diophantine exponent* $\omega^\times(Y)$ of Y by $\omega^\times(Y) \stackrel{\text{def}}{=} \sup\{\nu\colon Y \in \mathcal{W}^\times_{m,n}(\psi_{1,\nu})\}$, then clearly $\omega^\times(Y) \ge \omega(Y)$ for all Y, and yet $\omega^\times(Y) = n/m$ for λ-a.e. $Y \in M_{m,n}$.

Now specialize to $\mathbb{R}^n$ (by the same duality principle as before, it does not matter whether to think in terms of row or column vectors, but we will adopt the row vector set-up), and define the *multiplicative exponent* $\omega^\times(\mu)$ of a measure μ on $\mathbb{R}^n$ by $\omega^\times(\mu) \stackrel{\text{def}}{=} \sup\big\{\nu | \mu\big(\mathcal{W}^\times(\psi_{1,\nu})\big) > 0\big\}$; then $\omega^\times(\lambda) = n$. Following Sprindžuk [85], say that μ is *strongly extremal* if $\omega^\times(\mu) = n$. It turns out that all the results mentioned in the previous section have their multiplicative analogues; that is, the measures described there happen to be strongly extremal. This was conjectured by A. Baker [1] for the curve (12), and then by Sprindžuk in 1980 [85] for analytic nondegenerate manifolds. (We remark that only very few results in this set-up can be obtained by the standard methods, see e. g. [10]). The proof

of this stronger statement is based on using the multi-parameter action of

$$g_{\mathbf{t}} = \operatorname{diag}(e^{t_1+\cdots+t_n}, e^{-t_1}, \ldots, e^{-t_n}), \quad \text{where } \mathbf{t} = (t_1, \ldots, t_n)$$

instead of g_t considered in the previous section. One can show that the choice $h(\mathbf{x}) = g_{\mathbf{t}} L_{\mathbf{f}(\mathbf{x})}$ allows one to verify (i) and (ii) uniformly in $\mathbf{t} \in \mathbb{R}^n_+$, and the proof is finished by applying a multi-parameter version of the correspondence described in Sect. "Connection with Dynamics on the Space of Lattices". Namely, one can show that $\mathbf{y} \in \mathrm{VWA}^{\times}_{1,n}$ if and only if the trajectory $\{g_{\mathbf{t}} L_{\mathbf{y}} \mathbb{Z}^k : \mathbf{t} \in \mathbb{R}^n_+\}$ grows linearly, that is, for some $\gamma > 0$ one has $\operatorname{dist}(g_{\mathbf{t}} L_Y \mathbb{Z}^{n+1}, \mathbb{Z}^{n+1}) > \gamma \|\mathbf{t}\|$ for an unbounded set of $\mathbf{t} \in \mathbb{R}^n_+$. A similar correspondence was recently used in [30] to prove that the set of exceptions to Littlewood's Conjecture, which, using the terminology introduced above, can be called *badly multiplicatively approximable* vectors:

$$\mathrm{BA}^{\times}_{n,1} \stackrel{\text{def}}{=} \mathbb{R}^n \smallsetminus \bigcup_{c>0} \mathcal{W}^{\times}_{n,1}(\psi_{c,1/n}) = \left\{ \mathbf{y} : \inf_{q \in \mathbb{Z} \smallsetminus \{0\}, \mathbf{p} \in \mathbb{Z}^n} |q| \cdot \Pi(q\mathbf{y} - \mathbf{p}) > 0 \right\}, \quad (18)$$

has Hausdorff dimension zero. This was done using a measure rigidity result for the action of the group of diagonal matrices on the space of lattices. See [18] for an implicit description of this correspondence and [32,66,70] for more detail.

The dynamical approach also turned out to be fruitful in studying Diophantine properties of pairs $(\mathbf{f}, \nu)$ for which the nonplanarity condition fails. Note that obvious examples of non-extremal measures are provided by proper affine subspaces of $\mathbb{R}^n$ whose coefficients are rational or are well enough approximable by rational numbers. On the other hand, it is clear from a Fubini argument that almost all translates of any given subspace are extremal. In [51] the method of [56] was pushed further to produce criteria for the extremality, as well as the strong extremality, of arbitrary affine subspaces $\mathcal{L}$ of $\mathbb{R}^n$. Further, it was shown that if $\mathcal{L}$ is extremal (resp. strongly extremal), then so is any smooth submanifold of $\mathcal{L}$ which is nondegenerate in $\mathcal{L}$ at a.e. point. (The latter property is a straightforward generalization of the definition of nondegeneracy in $\mathbb{R}^n$: a map $\mathbf{f}$ is *nondegenerate in $\mathcal{L}$ at* $\mathbf{x}$ if the linear part of $\mathcal{L}$ is spanned by partial derivatives of $\mathbf{f}$ at $\mathbf{x}$). In other words, extremality and strong extremality pass from affine subspaces to their nondegenerate submanifolds.

A more precise analysis makes it possible to study Diophantine exponents of measures with supports contained in arbitrary proper affine subspaces of $\mathbb{R}^n$. Namely, in [53] it is shown how to compute $\omega(\mathcal{L})$ for any $\mathcal{L}$, and furthermore proved that if ν is a Federer measure on $\mathbb{R}^d$, U an open subset of $\mathbb{R}^d$, and $\mathbf{f}\colon U \to \mathbb{R}^n$ a continuous map such that the pair $(\mathbf{f}, \nu)$ is good and nonplanar in $\mathcal{L}$, then $\omega(\mathbf{f}_* \nu) = \omega(\mathcal{L})$. Here we say, generalizing the definition from Sect."Diophantine Approximation with Dependent Quantities: The Set-Up", that $(\mathbf{f}, \nu)$ is *nonplanar in $\mathcal{L}$* if for any ball B with $\nu(B) > 0$, the $\mathbf{f}$-image of $B \cap \operatorname{supp} \nu$ is not contained in any proper affine subspace of $\mathcal{L}$. (It is easy to see that for a smooth map $\mathbf{f}\colon U \to \mathcal{L}$, $(\mathbf{f}, \lambda)$ is good and nonplanar in $\mathcal{L}$ whenever $\mathbf{f}$ is nondegenerate in $\mathcal{L}$ at a.e. point). It is worthwhile to point out that these new applications require a strengthening of the measure estimate described at the end of Sect. "Diophantine Approximation with Dependent Quantities: The Set-Up": it was shown in [53] that (i) and (ii) would still imply (iii) if ρ in (ii) is replaced by $\rho^{\dim V}$.

Another application concerns badly approximable vectors. Using the dynamical description of the set $\mathrm{BA} \subset \mathbb{R}^n$ due to Dani [20], it turns out to be possible to find badly approximable vectors inside supports of certain measures on $\mathbb{R}^n$. Namely, if a subset K of $\mathbb{R}^n$ supports an absolutely friendly measure, then $\mathrm{BA} \cap K$ has Hausdorff dimension not less than the Hausdorff dimension of this measure. In particular, it proves that limit measures of irreducible system of self-similar/self-conformal contractions satisfying the Open Set Condition, such as e. g. the middle-third Cantor set on the real line, contain subsets of full Hausdorff dimension consisting of badly approximable vectors. This was established in [60] and later independently in [64] using a different approach. See also [36] for a stronger result.

The proof in [60] uses quantitative nondivergence estimates and an iterative procedure, which requires the measure in question to be absolutely friendly and not just friendly. A similar question for even the simplest not-absolutely friendly measures is completely open. For example, it is not known whether there exist uncountably many badly approximable pairs of the form (x, x^2). An analogous problem for atomic measures supported on algebraic numbers, that is, a "badly approximable" version of Roth's Theorem, is currently beyond reach as well — there are no known badly approximable (or, for that matter, well approximable) algebraic numbers of degree bigger than two.

It has been recently understood that the quantitative nondivergence method can be applied to the question of improvement to Dirichlet's Theorem (see the beginning of Sect. "Basic Facts"). Given a positive $\varepsilon < 1$, let us say that Dirichlet's Theorem *can be ε-improved* for $Y \in M_{m,n}$, writing $Y \in \mathrm{DI}_{m,n}(\varepsilon)$, if for every sufficiently large t the

system

$$\|Y\mathbf{q}-\mathbf{p}\| < \varepsilon e^{-t/m} \quad \text{and} \quad \|\mathbf{q}\| < \varepsilon e^{t/n} \tag{19}$$

(that is, (5) with the right hand side terms multiplied by ε) has a nontrivial integer solution $(\mathbf{p},\mathbf{q})$. It is a theorem of Davenport and Schmidt [24] that $\lambda\big(\mathrm{DI}_{m,n}(\varepsilon)\big)=0$ for any $\varepsilon<1$; in other words, Dirichlet's Theorem cannot be improved for Lebesgue-generic systems of linear forms. By a modification of the correspondence between dynamics and approximation, (19) is easily seen to be equivalent to $g_t L_Y \mathbb{Z}^k \in K_\varepsilon$, and since the complement to K_ε has nonempty interior for any $\varepsilon<1$, the result of Davenport and Schmidt follows from the ergodicity of the g_t-action on Ω_k.

Similar questions with λ replaced by $\mathbf{f}_*\lambda$ for some specific smooth maps $\mathbf{f}$ were considered in [2,3,15,23]. For example, [15], Theorem 7, provides an explicitly computable constant $\varepsilon_0=\varepsilon_0(n)$ such that for $\mathbf{f}$ as in (12),

$$\mathbf{f}_*\lambda\big(\mathrm{DI}_{1,n}(\varepsilon)\big)=0 \quad \text{for } \varepsilon<\varepsilon_0\,.$$

This had been previously done in [23] for $n=2$ and in [2] for $n=3$. In [62] this is extended to a much broader class of measures using estimates described in Sect. "Diophantine Approximation with Dependent Quantities: The Set-Up". In particular, almost every point of any nondegenerate smooth manifold is proved not to lie in $\mathrm{DI}(\varepsilon)$ for small enough ε depending only on the manifold. Earlier this was done in [61] for the set of *singular* vectors, defined as the intersection of $\mathrm{DI}(\varepsilon)$ over all positive ε; those correspond to divergent g_t-trajectories. As before, the advantage of the method is allowing a multiplicative generalization of the Dirichlet-improvement set-up; see [62] for more detail.

It is also worthwhile to mention that a generalization of the measure estimate discussed in Sect. "Diophantine Approximation with Dependent Quantities: The Set-Up" was used in [13] to estimate the measure of the set of points $\mathbf{x}$ in a ball $B\subset\mathbb{R}^d$ for which the system

$$\begin{cases} |\mathbf{f}(\mathbf{x})\cdot\mathbf{q}+p|<\varepsilon \\ |\mathbf{f}'(\mathbf{x})\cdot\mathbf{q}|<\delta \\ |q_i|<Q_i, \quad i=1,\dots,n, \end{cases}$$

where $\mathbf{f}$ is a smooth nondegenerate map $B\to\mathbb{R}^n$, has a nonzero integer solution. For that, $L_{\mathbf{f}(\mathbf{x})}$ as in (15) has to be replaced by the matrix

$$\begin{pmatrix} 1 & 0 & \mathbf{f}(\mathbf{x}) \\ 0 & 1 & \mathbf{f}'(\mathbf{x}) \\ 0 & 0 & I_n \end{pmatrix},$$

and therefore (i) and (ii) turn into more complicated conditions, which nevertheless can be checked when $\mathbf{f}$ is smooth and nondegenerate and ν is Lebesgue measure. This has resulted in proving the convergence case of Khintchine–Groshev Theorem for nondegenerate manifolds [13], in both standard and multiplicative versions. The aforementioned estimate was also used in [7] for the proof of the divergence case, and in [38,40] for establishing the convergence case of Khintchine–Groshev theorem for affine hyperplanes and their nondegenerate submanifolds. This generalized results obtained by standard methods for the curve (12) by Bernik and Beresnevich [6,11].

Finally, let us note that in many of the problems mentioned above, the ground field $\mathbb{R}$ can be replaced by $\mathbb{Q}_p$, and in fact several fields can be taken simultaneously, thus giving rise to the S-arithmetic setting where $S=\{p_1,\dots,p_s\}$ is a finite set of normalized valuations of $\mathbb{Q}$, which may or may not include the infinite valuation (cf. [83,90]). The space of lattices in $\mathbb{R}^{n+1}$ is replaced there by the space of lattices in $\mathbb{Q}_S^{n+1}$, where $\mathbb{Q}_S$ is the product of the fields $\mathbb{R}$ and $\mathbb{Q}_{p_1},\dots,\mathbb{Q}_{p_s}$. This is the subject of the paper [59], where S-arithmetic analogues of many results reviewed in Sect. "Diophantine Approximation with Dependent Quantities: The Set-Up" have been established. Similarly one can consider versions of the above theorems over local fields of positive characteristic [39]. See also [52] where Sprindžuk's solution [83] of the complex case of Mahler's Conjecture has been generalized (the latter involves studying small values of linear forms with coefficients in $\mathbb{C}$ at real integer points), and [31] which establishes a p-adic analogue of the result of [30] on the set of exceptions to Littlewood's Conjecture.

Future Directions

Interactions between ergodic theory and number theory have been rapidly expanding during the last two decades, and the author has no doubts that new applications of dynamics to Diophantine approximation will emerge in the near future. Specializing to the topics discussed in the present paper, it is fair to say that the list of "further results" contained in the previous section is by no means complete, and many even further results are currently in preparation. This includes: extending proofs of extremality and strong extremality of certain measures to the set-up of systems of linear forms (namely, with $\min(m,n)>1$; this was mentioned as work in progress in [56]); proving Khintchine-type theorems (both convergence and divergence parts) for p-adic and S-arithmetic nondegenerate manifolds, see [73] for results in this direction; extending [38,40] to establish Khintchine-type theo-

rems for submanifolds of arbitrary affine subspaces. The work of Druţu [28], who used ergodic theory on homogeneous spaces to compute the Hausdorff dimension of the intersection of $\mathcal{W}_{n,1}(\psi_{1,\nu})$, $\nu > 1$, with some rational quadratic hypersurfaces in $\mathbb{R}^n$ deserves a special mention; it is plausible that using this method one can treat more general situations. Several other interesting open directions are listed in [41], Section 9, in the final sections of papers [7,54], in the book [15], and in surveys by Frantzikinakis–McCutcheon, Nitica and Ward in this volume.

Acknowledgment

The work on this paper was supported in part by NSF Grant DMS-0239463.

Bibliography

1. Baker A (1975) Transcendental number theory. Cambridge University Press, London
2. Baker RC (1976) Metric diophantine approximation on manifolds. J Lond Math Soc 14:43–48
3. Baker RC (1978) Dirichlet's theorem on diophantine approximation. Math Proc Cambridge Phil Soc 83:37–59
4. Bekka M, Mayer M (2000) Ergodic theory and topological dynamics of group actions on homogeneous spaces. Cambridge University Press, Cambridge
5. Beresnevich V (1999) On approximation of real numbers by real algebraic numbers. Acta Arith 90:97–112
6. Beresnevich V (2002) A Groshev type theorem for convergence on manifolds. Acta Mathematica Hungarica 94:99–130
7. Beresnevich V, Bernik VI, Kleinbock D, Margulis GA (2002) Metric Diophantine approximation: the Khintchine–Groshev theorem for nondegenerate manifolds. Moscow Math J 2:203–225
8. Beresnevich V, Dickinson H, Velani S (2006) Measure theoretic laws for lim sup sets. Mem Amer Math Soc:179:1–91
9. Beresnevich V, Dickinson H, Velani S (2007) Diophantine approximation on planar curves and the distribution of rational points. Ann Math 166:367–426
10. Beresnevich V, Velani S (2007) A note on simultaneous Diophantine approximation on planar curves. Ann Math 337: 769–796
11. Bernik VI (1984) A proof of Baker's conjecture in the metric theory of transcendental numbers. Dokl Akad Nauk SSSR 277:1036–1039
12. Bernik VI, Dodson MM (1999) Metric Diophantine approximation on manifolds. Cambridge University Press, Cambridge
13. Bernik VI, Kleinbock D, Margulis GA (2001) Khintchine-type theorems on manifolds: convergence case for standard and multiplicative versions. Int Math Res Notices 2001:453–486
14. Besicovitch AS (1929) On linear sets of points of fractional dimensions. Ann Math 101:161–193
15. Bugeaud Y (2002) Approximation by algebraic integers and Hausdorff dimension. J London Math Soc 65:547–559
16. Bugeaud Y (2004) Approximation by algebraic numbers. Cambridge University Press, Cambridge
17. Cassels JWS (1957) An introduction to Diophantine approximation. Cambridge University Press, New York
18. Cassels JWS, Swinnerton-Dyer H (1955) On the product of three homogeneous linear forms and the indefinite ternary quadratic forms. Philos Trans Roy Soc London Ser A 248:73–96
19. Dani SG (1979) On invariant measures, minimal sets and a lemma of Margulis. Invent Math 51:239–260
20. Dani SG (1985) Divergent trajectories of flows on homogeneous spaces and diophantine approximation. J Reine Angew Math 359:55–89
21. Dani SG (1986) On orbits of unipotent flows on homogeneous spaces. II. Ergodic Theory Dynam Systems 6:167–182
22. Dani SG, Margulis GA (1993) Limit distributions of orbits of unipotent flows and values of quadratic forms. In: IM Gelfand Seminar. American Mathematical Society, Providence, pp 91–137
23. Davenport H, Schmidt WM (1970) Dirichlet's theorem on diophantine approximation. In: Symposia Mathematica. INDAM, Rome, pp 113–132
24. Davenport H, Schmidt WM (1969/1970) Dirichlet's theorem on diophantine approximation. II. Acta Arith 16:413–424
25. Ding J (1994) A proof of a conjecture of C. L. Siegel. J Number Theory 46:1–11
26. Dodson MM (1992) Hausdorff dimension, lower order and Khintchine's theorem in metric Diophantine approximation. J Reine Angew Math 432:69–76
27. Dodson MM (1993) Geometric and probabilistic ideas in the metric theory of Diophantine approximations. Uspekhi Mat Nauk 48:77–106
28. Druţu C (2005) Diophantine approximation on rational quadrics. Ann Math 333:405–469
29. Duffin RJ, Schaeffer AC (1941) Khintchine's problem in metric Diophantine approximation. Duke Math J 8:243–255
30. Einsiedler M, Katok A, Lindenstrauss E (2006) Invariant measures and the set of exceptions to Littlewood's conjecture. Ann Math 164:513–560
31. Einsiedler M, Kleinbock D (2007) Measure rigidity and *p*-adic Littlewood-type problems. Compositio Math 143:689–702
32. Einsiedler M, Lindenstrauss E (2006) Diagonalizable flows on locally homogeneous spaces and number theory. In: Proceedings of the International Congress of Mathematicians. Eur Math Soc, Zürich, pp 1731–1759
33. Eskin A (1998) Counting problems and semisimple groups. In: Proceedings of the International Congress of Mathematicians. Doc Math, Berlin, pp 539–552
34. Eskin A, Margulis GA, Mozes S (1998) Upper bounds and asymptotics in a quantitative version of the Oppenheim conjecture. Ann Math 147:93–141
35. Eskin A, Margulis GA, Mozes S (2005) Quadratic forms of signature (2,2) and eigenvalue spacings on rectangular 2-tori. Ann Math 161:679–725
36. Fishman L (2006) Schmidt's games on certain fractals. Israel S Math (to appear)
37. Gallagher P (1962) Metric simultaneous diophantine approximation. J London Math Soc 37:387–390
38. Ghosh A (2005) A Khintchine type theorem for hyperplanes. J London Math Soc 72:293–304
39. Ghosh A (2007) Metric Diophantine approximation over a local field of positive characteristic. J Number Theor 124:454–469
40. Ghosh A (2006) Dynamics on homogeneous spaces and Diophantine approximation on manifolds. Ph D Thesis, Brandeis University, Walthum

41. Gorodnik A (2007) Open problems in dynamics and related fields. J Mod Dyn 1:1–35
42. Groshev AV (1938) Une théorème sur les systèmes des formes linéaires. Dokl Akad Nauk SSSR 9:151–152
43. Harman G (1998) Metric number theory. Clarendon Press, Oxford University Press, New York
44. Hutchinson JE (1981) Fractals and self-similarity. Indiana Univ Math J 30:713–747
45. Jarnik V (1928-9) Zur metrischen Theorie der diophantischen Approximationen. Prace Mat-Fiz 36:91–106
46. Jarnik V (1929) Diophantischen Approximationen und Hausdorffsches Mass. Mat Sb 36:371–382
47. Khanin K, Lopes-Dias L, Marklof J (2007) Multidimensional continued fractions, dynamical renormalization and KAM theory. Comm Math Phys 270:197–231
48. Khintchine A (1924) Einige Sätze über Kettenbrüche, mit Anwendungen auf die Theorie der Diophantischen Approximationen. Math Ann 92:115–125
49. Khintchine A (1963) Continued fractions. P Noordhoff Ltd, Groningen
50. Kleinbock D (2001) Some applications of homogeneous dynamics to number theory. In: Smooth ergodic theory and its applications. American Mathematical Society, Providence, pp 639–660
51. Kleinbock D (2003) Extremal subspaces and their submanifolds. Geom Funct Anal 13:437–466
52. Kleinbock D (2004) Baker–Sprindžuk conjectures for complex analytic manifolds. In: Algebraic groups and Arithmetic. Tata Inst Fund Res, Mumbai, pp 539–553
53. Kleinbock D (2008) An extension of quantitative nondivergence and applications to Diophantine exponents. Trans AMS, to appear
54. Kleinbock D, Lindenstrauss E, Weiss B (2004) On fractal measures and diophantine approximation. Selecta Math 10: 479–523
55. Kleinbock D, Margulis GA (1996) Bounded orbits of nonquasi-unipotent flows on homogeneous spaces. In: Sinaĭ's Moscow Seminar on Dynamical Systems. American Mathematical Society, Providence, pp 141–172
56. Kleinbock D, Margulis GA (1998) Flows on homogeneous spaces and Diophantine approximation on manifolds. Ann Math 148:339–360
57. Kleinbock D, Margulis GA (1999) Logarithm laws for flows on homogeneous spaces. Invent Math 138:451–494
58. Kleinbock D, Shah N, Starkov A (2002) Dynamics of subgroup actions on homogeneous spaces of Lie groups and applications to number theory. In: Handbook on Dynamical Systems, vol 1A. Elsevier Science, North Holland, pp 813–930
59. Kleinbock D, Tomanov G (2007) Flows on *S*-arithmetic homogeneous spaces and applications to metric Diophantine approximation. Comm Math Helv 82:519–581
60. Kleinbock D, Weiss B (2005) Badly approximable vectors on fractals. Israel J Math 149:137–170
61. Kleinbock D, Weiss B (2005) Friendly measures, homogeneous flows and singular vectors. In: Algebraic and Topological Dynamics. American Mathematical Society, Providence, pp 281–292
62. Kleinbock D, Weiss B (2008) Dirichlet's theorem on diophantine approximation and homogeneous flows. J Mod Dyn 2:43–62
63. Kontsevich M, Suhov Y (1999) Statistics of Klein polyhedra and multidimensional continued fractions. In: Pseudoperiodic topology. American Mathematical Society, Providence, pp 9–27
64. Kristensen S, Thorn R, Velani S (2006) Diophantine approximation and badly approximable sets. Adv Math 203:132–169
65. Lagarias JC (1994) Geodesic multidimensional continued fractions. Proc London Math Soc 69:464–488
66. Lindenstrauss E (2007) Some examples how to use measure classification in number theory. In: Equidistribution in number theory, an introduction. Springer, Dordrecht, pp 261–303
67. Mahler K (1932) Über das Mass der Menge aller S-Zahlen. Math Ann 106:131–139
68. Margulis GA (1975) On the action of unipotent groups in the space of lattices. In: Lie groups and their representations (Budapest, 1971). Halsted, New York, pp 365–370
69. Margulis GA (1989) Discrete subgroups and ergodic theory. In: Number theory, trace formulas and discrete groups (Oslo, 1987). Academic Press, Boston, pp 377–398
70. Margulis GA (1997) Oppenheim conjecture. In: Fields Medalists' lectures. World Sci Publishing, River Edge, pp 272–327
71. Margulis GA (2002) Diophantine approximation, lattices and flows on homogeneous spaces. In: A panorama of number theory or the view from Baker's garden. Cambridge University Press, Cambridge, pp 280–310
72. Mauldin D, Urbański M (1996) Dimensions and measures in infinite iterated function systems. Proc London Math Soc 73:105–154
73. Mohammadi A, Salehi Golsefidy A (2008) *S*-Arithmetic Khintchine-Type Theorem. Preprint
74. Moore CC (1966) Ergodicity of flows on homogeneous spaces. Amer J Math 88:154–178
75. Roth KF (1955) Rational Approximations to Algebraic Numbers. Mathematika 2:1–20
76. Schmidt WM (1960) A metrical theorem in diophantine approximation. Canad J Math 12:619–631
77. Schmidt WM (1964) Metrische Sätze über simultane Approximation abhängiger Größen. Monatsh Math 63:154–166
78. Schmidt WM (1969) Badly approximable systems of linear forms. J Number Theory 1:139–154
79. Schmidt WM (1972) Norm form equations. Ann Math 96: 526–551
80. Schmidt WM (1980) Diophantine approximation. Springer, Berlin
81. Sheingorn M (1993) Continued fractions and congruence subgroup geodesics. In: Number theory with an emphasis on the Markoff spectrum (Provo, UT, 1991). Dekker, New York, pp 239–254
82. Sprindžuk VG (1964) More on Mahler's conjecture. Dokl Akad Nauk SSSR 155:54–56
83. Sprindžuk VG (1969) Mahler's problem in metric number theory. American Mathematical Society, Providence
84. Sprindžuk VG (1979) Metric theory of Diophantine approximations. VH Winston & Sons, Washington DC
85. Sprindžuk VG (1980) Achievements and problems of the theory of Diophantine approximations. Uspekhi Mat Nauk 35: 3–68
86. Starkov A (2000) Dynamical systems on homogeneous spaces. American Mathematical Society, Providence
87. Sullivan D (1982) Disjoint spheres, approximation by imaginary quadratic numbers, and the logarithm law for geodesics. Acta Math 149:215–237

88. Stratmann B, Urbanski M (2006) Diophantine extremality of the Patterson measure. Math Proc Cambridge Phil Soc 140:297–304
89. Urbanski M (2005) Diophantine approximation of self-conformal measures. J Number Theory 110:219–235
90. Želudevič F (1986) Simultane diophantische Approximationen abhängiger Grössen in mehreren Metriken. Acta Arith 46: 285–296

Ergodic Theory: Interactions with Combinatorics and Number Theory

Tom Ward
School of Mathematics, University of East Anglia, Norwich, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Ergodic Theory
Frequency of Returns
Ergodic Ramsey Theory and Recurrence
Orbit-Counting as an Analogous Development
Diophantine Analysis as a Toolbox
Future Directions
Bibliography

Glossary

Almost everywhere (abbreviated a. e.) A property that makes sense for each point x in a measure space $(X, \mathcal{B}, \mu)$ is said to hold almost everywhere (or a. e.) if the set $N \subset X$ on which it does not hold satisfies $N \in \mathcal{B}$ and $\mu(N) = 0$.

Čech–Stone compactification of $\mathbb{N}$, $\beta\mathbb{N}$ A compact Hausdorff space that contains $\mathbb{N}$ as a dense subset with the property that any map from $\mathbb{N}$ to a compact Hausdorff space K extends uniquely to a continuous map $\beta\mathbb{N} \to K$. This property and the fact that $\beta\mathbb{N}$ is a compact Hausdorff space containing $\mathbb{N}$ characterizes $\beta\mathbb{N}$ up to homeomorphism.

Curvature An intrinsic measure of the curvature of a Riemannian manifold depending only on the Riemannian metric; in the case of a surface it determines whether the surface is locally convex (positive curvature), locally saddle-shaped (negative) or locally flat (zero).

Diophantine approximation Theory of the approximation of real numbers by rational numbers: how small can the distance from a given irrational real number to a rational number be made in terms of the denominator of the rational?

Equidistributed A sequence is equidistributed if the asymptotic proportion of time it spends in an interval is proportional to the length of the interval.

Ergodic A measure-preserving transformation is ergodic if the only invariant functions are equal to a constant a. e.; equivalently if the transformation exhibits the convergence in the quasi-ergodic hypothesis.

Ergodic theory The study of statistical properties of orbits in abstract models of dynamical systems; more generally properties of measure-preserving (semi-) group actions on measure spaces.

Geodesic (flow) The shortest path between two points on a Riemannian manifold; such a geodesic path is uniquely determined by a starting point and the initial tangent vector to the path (that is, a point in the unit tangent bundle). The transformation on the unit tangent bundle defined by flowing along the geodesic defines the geodesic flow.

Haar measure (on a compact group) If G is a compact topological group, the unique measure μ defined on the Borel sets of G with the property that $\mu(A + g) = \mu(A)$ for all $g \in G$ and $\mu(G) = 1$.

Measure-theoretic entropy A numerical invariant of measure-preserving systems that reflects the asymptotic growth in complexity of measurable partitions refined under iteration of the map.

Mixing A measure-preserving system is mixing if measurable sets (events) become asymptotically independent as they are moved apart in time (under iteration).

(Quasi) Ergodic hypothesis The assumption that, in a dynamical system evolving in time and preserving a natural measure, there are some reasonable conditions under which the 'time average' along orbits of an observable (that is, the average value of a function defined on the phase space) will converge to the 'space average' (that is, the integral of the function with respect to the preserved measure).

Recurrence Return of an orbit in a dynamical system close to its starting point infinitely often.

S-Unit theorems A circle of results stating that linear equations in fields of zero characteristic have only finitely many solutions taken from finitely-generated multiplicative subgroups of the multiplicative group of the field (apart from infinite families of solutions arising from vanishing sub-sums).

Topological entropy A numerical invariant of topological dynamical systems that measures the asymptotic

growth in the complexity of orbits under iteration. The *variational principle* states that the topological entropy of a topological dynamical system is the supremum over all invariant measures of the measure-theoretic entropies of the dynamical systems viewed as measurable dynamical systems.

Definition of the Subject

Number theory is a branch of pure mathematics concerned with the properties of numbers in general, and integers in particular. The areas of most relevance to this article are *Diophantine analysis* (the study of how real numbers may be approximated by rational numbers, and the consequences for solutions of equations in integers); *analytic number theory*, and in particular asymptotic estimates for the number of primes smaller than X as a function of X; *equidistribution*, and questions about how the digits of real numbers are distributed. Combinatorics is concerned with identifying structures in discrete objects; of most interest here is that part of combinatorics connected with Ramsey theory, asserting that large subsets of highly structured objects must automatically contain large replicas of that structure. Ergodic theory is the study of asymptotic behavior of group actions preserving a probability measure; it has proved to be a powerful part of dynamical systems with wide applications.

Introduction

Ergodic theory, part of the mathematical study of dynamical systems, has pervasive connections with number theory and combinatorics. This article briefly surveys how these arise through a small sample of results. Unsurprisingly, many details are suppressed, and of course the selection of topics reflects the author's interests far more than it does the full extent of the flow of ideas between ergodic theory and number theory. In addition the selection of topics has been chosen in part to be complementary to those in related articles in the Encyclopedia. A particularly enormous lacuna is the theory of arithmetic dynamical systems itself – the recent monograph by Silverman [94] gives a comprehensive overview.

More sophisticated aspects of this connection – in particular the connections between ergodic theory on homogeneous spaces and Diophantine analysis – are covered in the articles ► Ergodic Theory on Homogeneous Spaces and Metric Number Theory and ► Ergodic Theory: Rigidity; more sophisticated overviews of the connections with combinatorics may be found in the article ► Ergodic Theory: Recurrence.

Ergodic Theory

While the early origins of ergodic theory lie in the quasi-ergodic hypothesis of classical Hamiltonian dynamics, the mathematical study of ergodic theory concerns various properties of group actions on measure spaces, including but not limited to several special branches:

1. The classical study of single measure-preserving transformations.
2. Measure-preserving actions of $\mathbb{Z}^d$; more generally of countable amenable groups.
3. Measure-preserving actions of $\mathbb{R}^d$ and more general amenable groups, called flows.
4. Measure-preserving actions of lattices in Lie groups.
5. Measure-preserving actions of Lie groups.

The ideas and conditions surrounding the quasi-ergodic hypothesis were eventually placed on a firm mathematical footing by developments starting in 1890. For a single measure-preserving transformation $T\colon X \to X$ of a probability space $(X, \mathcal{B}, \mu)$, Poincaré [74] showed a *recurrence theorem*: if $E \in \mathcal{B}$ is any measurable set, then for a. e. $x \in E$ there is an infinite set of return times, $0 < n_1 < n_2 < \cdots$ with $T^{n_j}(x) \in E$ (of course Poincaré noted this in a specific setting, concerned with a natural invariant measure for the "three-body" problem in planetary motion).

Poincaré's qualitative result was made quantitative in the 1930s, when von Neumann [105] used the approach of Koopman [52] to show the *mean ergodic theorem*: if $f \in L^2(\mu)$ then there is some $\overline{f} \in L^2(\mu)$ for which

$$\left\| \frac{1}{N} \sum_{n=0}^{N-1} f \circ T^n - \overline{f} \right\|_2 \longrightarrow 0 \quad \text{as} \quad N \longrightarrow \infty;$$

clearly $\overline{f}$ then has the property that $\|\overline{f} - \overline{f} \circ T\|_2 = 0$ and $\int \overline{f}\, \mathrm{d}\mu = \int f\, \mathrm{d}\mu$. Around the same time, Birkhoff [13] showed the more delicate *pointwise ergodic theorem*: for any $g \in L^1(\mu)$ there is some $\overline{g} \in L^1(\mu)$ for which

$$\frac{1}{N} \sum_{n=0}^{N-1} g(T^n x) \to \overline{g}(x) \text{ a. e.};$$

again it is then clear that $\overline{g}(Tx) = \overline{g}(x)$ a. e. and $\int \overline{g}\, \mathrm{d}\mu = \int g\, \mathrm{d}\mu$.

The map T is called *ergodic* if the invariance condition forces the function (f or g) to be equal to a constant a. e. Thus an ergodic map has the property that the *time* or *ergodic average* $(1/N)\sum_{n=0}^{N-1} f \circ T^j$ converges to the *space average* $\int \mathrm{d}\mu$. An overview of ergodic theorems and their

many extensions may be found in the article ► Ergodic Theorems.

Thus ergodic theory at its most basic level makes strong statements about the asymptotic behavior of orbits of a dynamical system as seen by *observables* (measurable functions on the space X). Applying the ergodic theorem to the indicator function of a measurable set A shows that ergodicity guarantees that a. e. orbit spends an asymptotic proportion of time in A equal to the volume $\mu(A)$ of that set (as measured by the invariant measure). This points to the start of the pervasive connections between ergodic theory and number theory – but as this and other articles relate, the connections extend far beyond this.

Frequency of Returns

In this section we illustrate the way in which a dynamical point of view may unify, explain and extend quite disparate results from number theory.

Normal Numbers

Borel [15] showed (as a consequence of what became the Borel–Cantelli Lemma in probability) that a. e. real number (with respect to Lebesgue measure) is *normal* to every base: that is, has the property that any block of k digits in the base-r expansion appears with asymptotic frequency r^{-k}.

Continued Fraction Digits

Analogs of normality results for the continued fraction expansion of real numbers were found by Khinchin, Kuz'min, Lévy and others. Any irrational $x \in [0,1]$ has a unique expansion as a continued fraction

$$x = \cfrac{1}{a_1(x) + \cfrac{1}{a_2(x) + \cfrac{1}{a_3(x) + \cdots}}}$$

and, just as in the case of the familiar base-r expansion, it turns out that the digits $(a_n(x))$ obey precise statistical rules for a. e. x. Gauss conjectured that the appearance of individual digits would obey the law

$$\frac{1}{N}|\{k\colon 1 \le k \le N, a_k(x) = j\}| \longrightarrow \frac{2\log(1+j) - \log j - \log(2+j)}{\log 2}\,. \quad (1)$$

This was eventually proved by Kuz'min [54] and Lévy [59], and the probability distribution of the digits is the *Gauss–Kuz'min law*. Khinchin [50] developed this further, showing for example that

$$\begin{aligned}\lim_{n\to\infty} (a_1(x)a_2(x)\ldots a_n(x))^{1/n} &= \prod_{n=1}^{\infty}\left(\frac{(n+1)^2}{n(n+2)}\right)^{\log n/\log 2}\\ &= 2.68545\ldots \quad \text{for a. e. } x\,.\end{aligned}$$

Lévy [60] showed that the denominator $q_n(x)$ of the nth convergent $(p_n(x))/(q_n(x))$ (the rational obtained by truncating the continued fraction expansion of x at the nth term) grows at a specific exponential rate,

$$\lim_{n\to\infty} \frac{1}{n}\log q_n(x) = \frac{\pi^2}{12\log 2} \quad \text{for a. e. } x\,.$$

First Digits

The astronomer Newcomb [67] noted that the first digits of large collections of numerical data that are not dimensionless have a specific and non-uniform distribution:

> "The law of probability of the occurrence of numbers is such that all mantissæ of their logarithms are equally probable."

This is now known as Benford's Law, following his popularization and possible rediscovery of the phenomenon [6]. In both cases, this was an empirical observation eventually made rigorous by Hill [42]. Arnold (App. 12 in [3]), pointed out the dynamics behind this phenomena in certain cases, best illustrated by the statistical behavior of the sequence $1, 2, 4, 8, 1, 3, 6, 1, \ldots$ of first digits of powers of 2. Empirically, the digit 1 appears about 30% of the time, while the digit 9 appears about 5% of the time.

Equidistribution

Weyl [109] (and, separately, Bohl [14] and Sierpiński [91]) found an important instance of *equidistribution*. Writing $\{\cdot\}$ for the fractional part, a sequence (a_n) of real numbers is said to be equidistributed modulo 1 if, for any interval $[a,b] \subset [0,1)$,

$$\frac{1}{N}|\{k\colon 1 \le k \le N, \{a_k\} \in [a,b]\}| \to (b-a) \quad \text{as} \quad N \to \infty\,;$$

equivalently if

$$\frac{1}{N}\sum_{k=1}^{N} f(a_k) \to \int_0^1 f(t)\mathrm{d}t \quad \text{as} \quad N \to \infty$$

for all continuous functions f. Weyl showed that the sequence $\{n\alpha\}$ is equidistributed if and only if α is irrational. This result was refined and extended in many directions; for example, Hlawka [44] and others found rates for the convergence in terms of the discrepancy of the sequence, Weyl [110] proved equidistribution for $\{n^2\alpha\}$, and Vinogradov for $\{p_n\alpha\}$ where p_n is the nth prime.

The Ergodic Context

All the results of this section are manifestations of various kinds of convergence of ergodic averages. Borel's theorem on normal numbers is an immediate consequence of the fact that Lebesgue measure on $[0,1)$ is invariant and ergodic for the map $x \mapsto bx$ modulo 1 with $b \geqslant 2$. The asymptotic properties of continued fraction digits are all a consequence of the fact that the *Gauss measure* defined by

$$\mu(A) = \frac{1}{\log 2}\int_A \frac{\mathrm{d}x}{1+x} \quad \text{for } A \subset [0,1]$$

is invariant and ergodic for the Gauss map $x \mapsto \{\frac{1}{x}\}$, and the orbit of an irrational number under the Gauss map determine the digits appearing in the continued fraction expansion much as the orbit under the map $x \mapsto bx \pmod 1$ determines the digits in the base b expansion.

The results on equidistribution and the frequency of first digits are related to ergodic averaging of a different sort. For example, writing $R_\alpha(t) = t + \alpha$ modulo 1 for the circle rotation by α, the first digit of 2^n is the digit j if and only if

$$\log_{10} j \leq R_{\log_{10}(2)}(0) < \log_{10}(j+1)\,.$$

Thus the asymptotic frequency of appearance concerns the orbit of a *specific point*. In order to see what this means, consider a continuous map $T\colon X \to X$ of a compact metric space (X, d). The space $\mathcal{M}(T)$ of Borel probability measures on the Borel σ-algebra of (X, d) is a non-empty compact convex set in the weak*-topology, each extreme point is an ergodic measure for T, and these ergodic measures are mutually singular. If $\mathcal{M}(T)$ is not a singleton and $\mu_1, \mu_2 \in \mathcal{M}(T)$ are distinct ergodic measures, then for a continuous function f with $\int_X f\,\mathrm{d}\mu_1 \neq \int_X f\,\mathrm{d}\mu_2$ it is clear that the ergodic averages $\frac{1}{N}\sum_{n=0}^{N-1} f(T^n x)$ must converge to $\int_X f\,\mathrm{d}\mu_1$ a. e. with respect to μ_1 and to $\int_X f\,\mathrm{d}\mu_2$ a. e. with respect to μ_2. Thus the presence of many invariant measures for a continuous map means that ergodic averages along the orbits of specific points need not converge to the space average with respect to a chosen invariant measure.

In the extreme situation of *unique ergodicity* (a single invariant measure, which is necessarily an extreme point of $\mathcal{M}(T)$ and hence ergodic) the convergence of ergodic averages is much more uniform. Indeed, if T is uniquely ergodic with $\mathcal{M}(T) = \{\mu\}$ then, for any continuous function $f\colon X \to \mathbb{R}$,

$$\frac{1}{N}\sum_{n=0}^{N-1} f(T^n x) \longrightarrow \int_X f\,\mathrm{d}\mu \text{ uniformly in } x$$

(see Oxtoby [69]). The circle rotation R_α is uniquely ergodic for irrational α, leading to the equidistribution results.

The ergodic viewpoint on equidistribution also places equidistribution results in a wider context. Weyl's result that $\{n^2\alpha\}$ (indeed, the fractional part of any polynomial with at least one irrational coefficient) is equidistributed for irrational α was given another proof by Furstenberg [29] using the notion of unique ergodicity. These methods were then used in the study of nilsystems (translations on quotients of nilpotent Lie groups) by Auslander, Green and Hahn [4] and Parry [70], and these nilsystems play an essential role for polynomial (and other non-conventional) ergodic averaging (see Host and Kra [45] and Leibman [58] in the polynomial case; Host and Kra [46] in the multiple linear case). Remarkably, nilsystems are starting to play a role within combinatorics – an example is the work on the asymptotic number of 4-step arithmetic progressions in the primes by Green and Tao [40]. Pointwise ergodic theorems have also been found along sequences other than $\mathbb{N}$; notably for integer-valued polynomials and along the primes for L^2 functions by Bourgain [16,17]. For more details, see the survey paper of del Junco on ergodic theorems.

Ergodic Ramsey Theory and Recurrence

In 1927 van der Waerden proved a conjecture attributed to Baudet: if the natural numbers are written as a disjoint union of finitely many sets,

$$\mathbb{N} = C_1 \sqcup C_2 \sqcup \cdots \sqcup C_r\,, \tag{2}$$

then there must be one set C_j that contains arbitrarily long arithmetic progressions. That is, there is some $j \in \{1,\ldots,r\}$ such that for any $k \geqslant 1$ there are $a \geqslant 1$ and $n \geqslant 1$ with

$$a, a+n, a+2n, \ldots, a+(k-1)n \in C_j\,.$$

The original proof appears in van der Waerden's paper [103], and there is a discussion of how he found the proof in [104].

Work of Furstenberg and Weiss [34] and others placed the theorem of van der Waerden in the context of topological dynamics, giving alternative proofs. Specifically, van der Waerden's theorem is a consequence of *topological multiple recurrence*: the return of points under iteration in a topological dynamical system close to their starting point along finite sequences of times. The same approach readily gives dynamical proofs of Rado's extension [76] of van der Waerden's theorem, and of Hindman's theorem [43]. The theorems of Rado and Hindman introduce a new theme: given a set $A = \{n_1, n_2, \dots\}$ of natural numbers, write $FS(A)$ for the set of numbers obtained as finite sums $n_{i_1} + \cdots + n_{i_j}$ with $i_1 < i_2 < \cdots < i_j$. Rado showed that for any large n there is some C_s containing some $FS(A)$ for a set A of cardinality n. Hindman showed that there is some C_s containing some $FS(A)$ for an *infinite* set A.

In the theorem of van der Waerden, it is clear that for any reasonable notion of "proportion" or "density" one of the sets C_j must occupy a positive proportion of $\mathbb{N}$. A set $A \subset \mathbb{N}$ is said to have *positive upper density* if there are sequences (M_i) and (N_i) with $N_i - M_i \to \infty$ as $i \to \infty$ such that

$$\lim_{i\to\infty} \frac{1}{N_i - M_i} |\{a \in A\colon M_i < a < N_i\}| > 0 .$$

Erdős and Turán [26] conjectured the stronger statement that any subset of $\mathbb{N}$ with positive upper density must contain arbitrary long arithmetic progressions. This statement was shown for arithmetic progressions of length 3 by Roth [79] in 1952, then for length 4 by Szemerédi [96] in 1969. The general result was eventually proved by Szemerédi [97] in 1975 in a lengthy and extremely difficult argument.

Furstenberg saw that Szemerédi's Theorem would follow from a deep extension of the Poincaré recurrence phenomena described in Sect. "Ergodic Theory" and proved that extension [30] (see also the survey article by Furstenberg, Katznelson and Ornstein [33]). The *multiple recurrence* result of Furstenberg says that for any measure-preserving system $(X, \mathcal{B}, \mu, T)$ and set $A \in \mathcal{B}$ with $\mu(A) > 0$, and for any $k \in \mathbb{N}$,

$$\liminf_{N-M\to\infty} \frac{1}{N-M+1} \cdot \sum_{n=M}^{N} \mu\left(A \cap T^{-n}A \cap T^{-2n}A \cap \cdots \cap T^{-kn}A\right) > 0 .$$

An immediate consequence is that in the same setting there must be some $n \geqslant 1$ for which

$$\mu\left(A \cap T^{-n}A \cap T^{-2n}A \cap \cdots \cap T^{-kn}A\right) > 0 . \qquad (3)$$

A general *correspondence principle*, due to Furstenberg, shows that statements in combinatorics like Szemerédi's Theorem are equivalent to statements in ergodic theory like (3).

This opened up a significant new field of *ergodic Ramsey theory*, in which methods from dynamical systems and ergodic theory are used to produce new results in infinite combinatorics. For an overview, see the articles ▶ Ergodic Theory on Homogeneous Spaces and Metric Number Theory, ▶ Ergodic Theory: Rigidity, ▶ Ergodic Theory: Recurrence and the survey articles of Bergelson [7,8,10]. The field is too large to give an overview here, but a few examples will give a flavor of some of the themes.

Call a set $R \subset \mathbb{Z}$ a *set of recurrence* if, for any finite measure-preserving invertible transformation T of a finite measure space $(X, \mathcal{B}, \mu)$ and any set $A \in \mathcal{B}$ with $\mu(A) > 0$, there are infinitely many $n \in R$ for which $\mu(A \cap T^{-n}A) > 0$. Thus Poincaré recurrence is the statement that $\mathbb{N}$ is a set of recurrence. Furstenberg and Katznelson [31] showed that if $T_1, \dots, T_k$ form a family of commuting measure-preserving transformations and A is a set of positive measure, then

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mu(T_1^{-n}A \cap \cdots \cap T_k^{-n}A) > 0 .$$

This remarkable multiple recurrence implies a multi-dimensional form of Szemerédi's theorem. Recently, Gowers has found a non-ergodic proof of this [38].

Furstenberg also gave an ergodic proof of Sárközy's theorem [80]: if $p \in \mathbb{Q}[t]$ is a polynomial with $p(\mathbb{Z}) \subset \mathbb{Z}$ and $p(0) = 0$, then $\{p(n)\}_{n>0}$ is a set of recurrence. This was extended to multiple polynomial recurrence by Bergelson and Leibman [11].

Topology and Coloring Theorems

The existence of idempotent ultrafilters in the Čech–Stone compactification $\beta\mathbb{N}$ gives rise to an algebraic approach to many questions in topological dynamics (this notion has its origins in the work of Ellis [24]). Using these methods, results like Hindman's finite sums theorem find elegant proofs, and many new results in combinatorics have been found. For example, in the partition (2) there must be one set C_j containing a triple x, y, z solving $x - y = z^2$.

A deeper application is to improve a strengthening of Kronecker's theorem. To explain this, recall that a set S is called *IP* if there is a sequence (n_i) of natural numbers (which do not need to be distinct) with the property that S contains all the terms of the sequence and all finite sums of terms of the sequence with distinct indices.

A set S is called $I\!P^*$ if it has non-empty intersection with *every* $I\!P$ set, and a set S is called $I\!P^*_+$ if there is some $t \in \mathbb{Z}$ for which $S - t$ is $I\!P^*$. Thus being $I\!P^*$ (or $I\!P^*_+$) is an extreme form of 'fatness' for a set. Now let $1, \alpha_1, \ldots, \alpha_k$ be numbers that are linearly independent over the rationals, and for any $d \in \mathbb{N}$ and kd non-empty intervals $I_{ij} \subset [0,1]$ $(1 \le i \le d, 1 \le j \le k)$, let

$$D = \{n \in \mathbb{N} : \{n^i \alpha_j\} \in I_{ij} \text{ for all } i, j\} .$$

Kronecker showed that if $d = 1$ then D is non-empty; Hardy and Littlewood showed that D is infinite, and Weyl showed that D has positive density. Bergelson [9] uses these algebraic methods to improve the result by showing that D is an $I\!P^*_+$ set.

Polynomialization and *IP*-sets

As mentioned above, Bergelson and Leibman [11] extended multiple recurrence to a polynomial setting. For example, let

$$\{p_{i,j} : 1 \le i \le k, 1 \le j \le t\}$$

be a collection of polynomials with rational coefficients and $p_{i,j}(\mathbb{Z}) \subset \mathbb{Z}$, $p_{i,j}(0) = 0$. Then if $\mu(A) > 0$, we have

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mu\left(\bigcap_{i=1}^{k} \left(\prod_{j=1}^{t} T_j^{p_{i,j}(n)}\right)^{-1} A\right) > 0 .$$

Using the Furstenberg correspondence principle, this gives a multi-dimensional polynomial Szemerédi theorem: If $P : \mathbb{Z}^r \to \mathbb{Z}^\ell$ is a polynomial mapping with the property that $P(0) = 0$, and $F \subset \mathbb{Z}^r$ is a finite configuration, then any set $S \subset \mathbb{Z}^\ell$ of positive upper Banach density contains a set of the form $u + P(nF)$ for some $u \in \mathbb{Z}^\ell$ and $n \in \mathbb{N}$.

In a different direction, motivated in part by Hindman's theorem, the multiple recurrence results generalize to $I\!P$-sets. Furstenberg and Katznelson [32] proved a linear $I\!P$-multiple recurrence theorem in which the recurrence is guaranteed to occur along an $I\!P$-set. A combinatorial proof of this result has been found by Nagle, Rödl and Schacht [66]. Bergelson and McCutcheon [12] extended these results by proving a polynomial $I\!P$-multiple recurrence theorem. To formulate this, make the following definitions. Write $\mathcal{F}$ for the family of non-empty finite subsets of $\mathbb{N}$, so that a sequence indexed by $\mathcal{F}$ is an $I\!P$-set. More generally, an $\mathcal{F}$-sequence $(n_\alpha)_{\alpha\in\mathcal{F}}$ taking values in an abelian group is called an $I\!P$-sequence if $n_{\alpha\cup\beta} = n_\alpha + n_\beta$ whenever $\alpha \cap \beta = \emptyset$. An $I\!P$-ring is a set of the form $\mathcal{F}^{(1)} = \{\bigcup_{i\in\beta} \alpha_i : \beta \in \mathcal{F}\}$ where $\alpha_1 < \alpha_2 < \cdots$ is a sequence in $\mathcal{F}$, and $\alpha < \beta$ means $a < b$ for all $a \in \alpha, b \in \beta$; write $\mathcal{F}^m_<$ for the set of m-tuples $(\alpha_1, \ldots, \alpha_m)$ from $\mathcal{F}^m$ with $\alpha_i < \alpha_j$ for $i < j$. Write $PE(m,d)$ for the collection of all expressions of the form $T(\alpha_1, \ldots, \alpha_m) = \prod_{i=1}^r T_i^{p_i((n^{(b)}_{\alpha_j})_{1\le b\le k,\, 1\le j\le m})}$, $(\alpha_1, \ldots, \alpha_m) \in (F \cup \emptyset)^m_<$, where each p_i is a polynomial in a $k \times m$ matrix of variables with integer coefficients and zero constant term with degree $\le d$. Then for every $m, t \in \mathbb{N}$, there is an $I\!P$-ring $\mathcal{F}^{(1)}$, and an $a = a(A, m, t, d) > 0$, such that, for every set of polynomial expressions $\{S_0, \ldots, S_t\} \subset PE(m,d)$,

$$I\!P - \lim_{(\alpha_1,\ldots,\alpha_m)\in(\mathcal{F}^{(1)})^m_<} \mu\left(\bigcap_{i=0}^{t} S_i(\alpha_1, \ldots, \alpha_m)^{-1} A\right) > 0 .$$

There are a large number of deep combinatorial consequences of this result, not all of which seem accessible by other means.

Sets of Primes

In a remarkable development, Szemerédi's theorem and some of the ideas behind ergodic Ramsey theory joined results of Goldston and Yıldırım [37] in playing a part in Green and Tao's proof [39] that the set of primes contains arbitrarily long arithmetic progressions. This profound result is surveyed from an ergodic point of view in the article of Kra [53]. As with Szemerédi's theorem itself, this result has been extended to a polynomial setting by Tao and Ziegler [101]. Given integer-valued polynomials $f_1, \ldots, f_k \in \mathbb{Z}[t]$ with

$$f_1(0) = \cdots = f_k(0) = 0$$

and any $\varepsilon > 0$, Tao and Ziegler proved that there are infinitely many integers x, m with $1 \le m \le x^\varepsilon$ for which $x + f_1(m), \ldots, x + f_k(m)$ are primes.

Orbit-Counting as an Analogous Development

Some of the connections between number theory and ergodic theory arise through developments that are analogous but not directly related. A remarkable instance of this concerns the long history of attempts to count prime numbers laid alongside the problem of counting closed orbits in dynamical systems.

Counting Orbits and Geodesics

Consider first the fundamental arithmetic function $\pi(X) = |\{p \le X : p \text{ is prime}\}|$. Tables of primes prepared by Felkel and Vega in the 18th century led Legendre to suggest that $\pi(X)$ is approximately $x/(\log(X) - 1.08)$. Gauss,

using both computational evidence and a heuristic argument, suggested that $\pi(X)$ is approximated by

$$\mathrm{li}(X) = \int_2^X \frac{\mathrm{d}t}{\log t} .$$

Both of these suggestions imply the well-known asymptotic formula

$$\pi(X) \sim \frac{X}{\log X} . \tag{4}$$

Riemann brought the analytic ideas of Dirichlet and Chebyshev (who used the zeta function to find a weaker version of (4) with upper and lower bounds for the quantity $\frac{\pi(X)\log(X)}{X}$) to bear by proposing that the zeta function

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_p \left(1 - p^{-s}\right)^{-1} , \tag{5}$$

already studied by Euler, would connect properties of the primes to analytic methods. An essential step in these developments, due to Riemann, is the meromorphic extension of ζ from the region $\Re(s) > 1$ in (5) to the whole complex plane and a functional equation relating the value of the extension at s to the value at $1 - s$. Moreover, Riemann showed that the extension has readily understood real zeros, and that all the other zeros he could find were symmetric about $\Re(s) = \frac{1}{2}$. The *Riemann hypothesis* asserts that zeros in the region $0 < \Re(s) < 1$ all lie on the line $\Re(s) = \frac{1}{2}$, and this remains open.

Analytic properties of the Riemann zeta function were used by Hadamard and de la Vallée Poussin to prove (4), the *Prime Number Theorem*, in 1896. Tauberian methods developed by Wiener and Ikehara [111] later gave different approaches to the Prime Number Theorem.

These ideas initiated the widespread use of zeta functions in several parts of mathematics, but it was not until the middle of the 20th century that Selberg [88] introduced a zeta function dealing directly with quantities arising in dynamical systems: the lengths of closed geodesics on surfaces of constant curvature -1. The geodesic flow acts on the unit tangent bundle to the surface by moving a point and unit tangent vector at that point along the unique geodesic they define at unit speed. Closed geodesics are then in one-to-one correspondence with periodic orbits of the associated geodesic flow on the unit tangent bundle, and it is in this sense that the quantities are dynamical. The function defined by Selberg takes the form

$$Z(s) = \prod_{\tau} \prod_{k=0}^{\infty} \left(1 - \mathrm{e}^{-(s+k)|\tau|}\right) ,$$

in which τ runs over all the closed geodesics, and $|\tau|$ denotes the length of the geodesic. In a direct echo of the Riemann zeta function, Selberg found an analytic continuation to the complex plane, and showed that the zeros of Z lie on the real axis or on the line $\Re(s) = \frac{1}{2}$ (the analogue of the Riemann hypothesis for Z; see also the paper of Hejhal [41]). The zeros of Z are closely connected to the eigenvalues for the Laplace–Beltrami operator, and thus give information about the lengths of closed geodesics via Selberg's trace formula in the same paper. Huber [47] and others used this approach to give an analogue of the prime number theorem for closed geodesics – a *prime orbit theorem*.

Sinai [92] considered closed geodesics on a manifold M with negative curvature bounded between $-R^2$ and $-r^2$, and found the bounds

$$(\dim(M) - 1)r \le \liminf_{T\to\infty} \frac{\log \pi(T)}{T} \le \limsup_{T\to\infty} \frac{\log \pi(T)}{T} \le (\dim(M) - 1)R$$

for the number $\pi(T)$ of closed geodesics of multiplicity one with length less than T, analogous to Chebyshev's result.

The essential dynamical feature behind the geodesic flow on a manifold of negative curvature is that it is an example of an *Anosov flow* [2]. These are smooth dynamical $\mathbb{R}$-actions (equivalently, first-order differential equations on Riemannian manifolds) with the property that the tangent bundle has a continuously varying splitting into a direct sum $E^u \oplus E^s \oplus E^o$ and the action of the differential of the flow acts on E^u as an exponential expansion, on E^s as an exponential contraction, E^o is the one-dimensional bundle of vectors that are tangent to orbits, and the expansion and contraction factors are bounded. In the setting of Anosov flows, the natural orbit counting function is $\pi(X) = |\{\tau : \tau \text{ a closed orbit of length } |\tau| \le X\}|$. Margulis [62,63] generalized the picture to weak-mixing Anosov flows by showing a prime orbit theorem of the form

$$\pi(X) \sim \frac{\mathrm{e}^{h_{\mathrm{top}} X}}{h_{\mathrm{top}} X} \tag{6}$$

for the counting function $\pi(X) = |\{\tau : \tau \text{ a closed orbit of length } |\tau| \le X\}|$ where as before h_{top} denotes the topological entropy of the flow. Integral to Margulis' work is a result on the spatial distribution of the closed geodesics reflected in a flow-invariant probability measure, now called the Margulis measure.

Anosov also studied discrete dynamical systems with similar properties: diffeomorphisms of compact manifolds

with a similar splitting of the tangent space (though in this setting E^o disappears). The archetypal Anosov diffeomorphism is a hyperbolic toral automorphism of the sort considered in Subsect. "Orbit Growth and Convergence"; for such automorphisms of the 2-torus Adler and Weiss [1] constructed Markov partitions, allowing the dynamics of the toral automorphism to be modeled by a topological Markov shift, and used this to determine when two such automorphisms are measurably isomorphic. Sinai [93], Ratner [77], Bowen [18,20] and others developed the construction of Markov partitions in general for Anosov diffeomorphisms and flows.

Around the same time, Smale [95] introduced a more permissive hyperbolicity axiom for diffeomorphisms, Axiom A. Maps satisfying Axiom A are diffeomorphisms satisfying the same hypothesis as that of Anosov diffeomorphisms, but only on the set of points that return arbitrarily close under the action of the flow (or iteration of the map).

Thus Markov partitions, and with them associated transfer operators became a substitute for the geometrical Laplace–Beltrami operators of the setting considered by Selberg. Bowen [19] extended the uniform distribution result of Margulis to this setting and found an analogue of Chebychev's theorem for closed orbits. Parry [71] (in a restricted case) and Parry and Pollicott [73] went on to prove the prime orbit theorem in this more general setting. The methods are an adaptation of the Ikehara–Wiener–Tauberian approach to the prime number theorem.

Thus many facets of the prime number theorem story find their echoes in the study of closed orbits for hyperbolic flows: the role played by meromorphic extensions of suitable zeta functions, Tauberian methods, and so on. Moreover, related results from number theory have analogues in dynamics, for example Mertens' theorem [65] in the work of Sharp [89] and Noorani [68] and Dirichlet's theorem in work of Parry [72].

The "elementary" proof (not using analytic methods) of the prime number theorem by Erdős [25] and Selberg [87] (see the survey by Goldfeld [36] for the background to the results and the unfortunate priority dispute) has an echo in some approaches to orbit-counting problems from an elementary (non-Tauberian) perspective, including work of Lalley [56] on special flows and Everest, Miles, Stevens and the author [27] in the algebraic setting.

In a different direction Lalley [55] found orbit asymptotics for closed orbits satisfying constraints in the Axiom A setting without using Tauberian theorems. His more direct approach is still analytic, using complex transfer operators (the same objects used to by Parry and Pollicott to study the dynamical zeta function at complex values) and indeed somewhat parallels a Tauberian argument.

Further resonances with number theory arise here. For example, there are results on the distribution of closed orbits for group extensions (analogous to Chebotarev's theorem) and for orbits with homological constraints (see Sharp [90], Katsuda and Sunada [49]).

Of course the great diversity of dynamical systems subsumed in the phrase "prime orbit theorem" creates new problems and challenges, and in particular if there is not much geometry to work with then the reliance on Markov partitions and transfer operators makes it difficult to find higher-order asymptotics.

Dolgopyat [22] has nonetheless managed to push the Markov methods to obtain uniform bounds on iterates of the associated transfer operators to the region $\Re(s) > \sigma_0$ with $\sigma_0 < 1$. This result has wide implications; an example most relevant to the analogy with number theory is the work of Pollicott and Sharp [75] in which Dolgopyat's result is used to show that for certain geodesic flows there is a two-term prime orbit theorem of the form

$$\pi(X) = \mathrm{li}\left(\mathrm{e}^{h_{\mathrm{top}}X}\right) + \mathrm{O}\left(\mathrm{e}^{cX}\right)$$

for some $c < h_{\mathrm{top}}$.

For non-positive curvature manifolds less is known: Knieper [51] finds upper and lower bounds for the function counting closed geodesics on rank-1 manifolds of non-positive curvature of the form

$$A\frac{\mathrm{e}^{hX}}{X} \le \pi(X) \le B\mathrm{e}^{hX}$$

for constants $A, B > 0$.

Counting Orbits for Group Endomorphisms

A prism through which to view some of the deeper issues that arise in Subsect. "Counting Orbits and Geodesics" is provided by group endomorphisms. The price paid for having simple closed formulas for all the quantities involved is of course a severe loss of generality, but the diversity of examples illustrates many of the phenomena that may be expected in more general settings when hyperbolicity is lost.

Consider an endomorphism $T\colon X \to X$ of a compact group with the property that $\mathrm{F}_n(T) < \infty$ for all $n \geqslant 1$. The number of closed orbits of length n under T is then

$$\mathrm{O}_n(T) = \frac{1}{n}\sum_{d|n}\mu(n/d)\mathrm{F}_d(T)\,. \tag{7}$$

In simple situations (hyperbolic toral automorphisms for example) it is straightforward to show that

$$\pi_T(X) = |\{\tau : \tau \text{ a closed orbit under } T \text{ of length} \leq X\}| \sim \frac{e^{(X+1)h_{\text{top}}(T)}}{X} . \tag{8}$$

Waddington [106] considered quasihyperbolic toral automorphisms, showing that the asymptotic (8) in this case is multiplied by an explicit almost-periodic function bounded away from zero and infinity.

This result has been extended further into non-hyperbolic territory, which is most easily seen via the so-called connected S-integer dynamical systems introduced by Chothi, Everest and the author [21]. Fix an algebraic number field $\mathbb{K}$ with set of places $P(\mathbb{K})$ and set of infinite places $P_\infty(\mathbb{K})$, an element of infinite multiplicative order $\xi \in \mathbb{K}^*$, and a finite set $S \subset P(\mathbb{K}) \setminus P_\infty(\mathbb{K})$ with the property that $|\xi|_w \leq 1$ for all $w \notin S \cup P_\infty(\mathbb{K})$. The associated ring of S-integers is

$$R_S = \{x \in \mathbb{K} : |x|_w \leq 1 \text{ for all } w \notin S \cup P_\infty(\mathbb{K})\} .$$

Let X be the compact character group of R_S, and define the endomorphism $T : X \to X$ to be the dual of the map $x \mapsto \xi x$ on R_S. Following Weil [108], write $\mathbb{K}_w$ for the completion at w, and for w finite, write r_w for the maximal compact subring of $\mathbb{K}_w$. Notice that if $S = P$ then $R_S = \mathbb{K}$ and $\mathrm{F}_n(T) = 1$ for all $n \geqslant 1$ by the product formula for $\mathbb{A}$-fields. As the set S shrinks, more and more periodic orbits come into being, and if S is as small as possible (given ξ) then the resulting system is (more or less) hyperbolic or quasi-hyperbolic.

For S finite, it turns out that there are still sufficiently many periodic orbits to have the growth rate result (10), but the asymptotic (8) is modified in much the same way as Waddington observed for quasi-hyperbolic toral automorphisms:

$$\liminf_{X\to\infty} \frac{X\pi_T(X)}{e^{(X+1)h_{\text{top}}(T)}} > 0 \tag{9}$$

and there is an associated pair (X^*, a_T), where X^* is a compact group and $a_T \in X^*$, with the property that if $a_T^{N_j}$ converges in X^* as $j \to \infty$, then there is convergence in (9).

A simple special case will illustrate this. Taking $\mathbb{K} = \mathbb{Q}$, $\xi = 2$ and $S = \{3\}$ gives a compact group endomorphism T with

$$\mathrm{F}_n(T) = (2^n - 1)|2^n - 1|_3 .$$

For this example the results of [21] are sharper: The expression in (9) converges along (X_j) if and only if 2^{X_j} converges in the ring of 3-adic integers $\mathbb{Z}_3$, the expression has uncountably many limit points, and the upper and lower limits are transcendental.

Similarly, the dynamical analogue of Mertens' theorem found by Sharp may be found for S-integer systems with S finite. Writing

$$\mathcal{M}_T(N) = \sum_{|\tau|\leq N} \frac{1}{e^{h(T)|\tau|}} ,$$

it is shown in [21] that for an ergodic S-integer map T with $\mathbb{K} = \mathbb{Q}$ and S finite, there are constants $k_T \in \mathbb{Q}$ and C_T such that

$$\mathcal{M}_T(N) = k_T \log N + C_T + \mathrm{O}(1/N) .$$

Without the restriction that $\mathbb{K} = \mathbb{Q}$, it is shown that there are constants $k_T \in \mathbb{Q}$, C_T and $\delta > 0$ with

$$\mathcal{M}_T(N) = k_T \log N + C_T + \mathrm{O}(N^{-\delta}) .$$

Diophantine Analysis as a Toolbox

Many problems in ergodic theory and dynamical system exploit ideas and results from number theory in a direct way; we illustrate this by describing a selection of dynamical problems that call on particular parts of number theory in an essential way. The example of mixing in Subsect. "Mixing and Additive Relations in Fields" is particularly striking for two reasons: the results needed from number theory are relatively recent, and the ergodic application directly motivated a further development in number theory.

Orbit Growth and Convergence

The analysis of periodic orbits – how their number grows as the length grows and how they spread out through space – is of central importance in dynamics (see Katok [48] for example). An instance of this is that for many simple kinds of dynamical systems $T : X \to X$ (where T is a continuous map of a compact metric space (X, d)) the logarithmic growth rate of the number of periodic points exists and coincides with the topological entropy $h(T)$ (an invariant giving a quantitative measure of the average rate of growth in orbit complexity under T). That is, writing

$$\mathrm{F}_n(T) = |\{x \in X : T^n x = x\}| ,$$

we find

$$\frac{1}{n} \log \mathrm{F}_n(T) \longrightarrow h_{\text{top}}(T) \tag{10}$$

for many of the simplest dynamical systems. For example, if $X = \mathbb{T}^r$ is the r-torus and $T = T_A$ is the automorphism of the torus corresponding to a matrix A in $GL_r(\mathbb{Z})$, then T_A is ergodic with respect to Lebesgue measure if and only if no eigenvalue of A is a root of unity. Under this assumption, we have

$$\mathrm{F}_n(T_A) = \prod_{i=1}^{r} |\lambda_i^n - 1|$$

and

$$h_{\mathrm{top}}(T_A) = \sum_{i=1}^{r} \log \max\{1, |\lambda_i|\} \tag{11}$$

where $\lambda_1, \ldots, \lambda_r$ are the eigenvalues of A. It follows that the convergence in (10) is clear under the assumption that T_A is *hyperbolic* (that is, no eigenvalue has modulus one). Without this assumption the convergence is less clear: for $r \geqslant 4$ the automorphism T_A may be ergodic without being hyperbolic. That is, while no eigenvalues are unit roots some may have unit modulus. As pointed out by Lind [61] in his study of these *quasihyperbolic* automorphisms, the convergence (10) does still hold for these systems, but this requires a significant Diophantine result (the theorem of Gel'fond [35] suffices; one may also use Baker's theorem [5]). Even further from hyperbolicity lie the family of S-integer systems [21,107]; their orbit-growth properties are intimately tied up with Artin's conjecture on primitive roots and prime divisors of linear recurrence sequences.

Mixing and Additive Relations in Fields

The problem of higher-order mixing for commuting group automorphisms provides a striking example of the dialogue between ergodic theory and number theory, in which deep results from number theory have been used to solve problems in ergodic theory, and questions arising in ergodic theory have motivated further developments in number theory.

An action T of a countable group Γ on a probability space $(X, \mathcal{B}, \mu)$ is called *k-fold mixing* or *mixing on* $(k+1)$ *sets* if

$$\mu\left(A_0 \cap T^{-g_1} A_1 \cap \cdots \cap T^{-g_k} A_k\right) \longrightarrow \mu(A_0) \cdots \mu(A_k) \tag{12}$$

as

$$g_i g_j^{-1} \longrightarrow \infty \quad \text{for } i \neq j$$

with the convention that $g_0 = 1_\Gamma$, for any sets $A_0, \ldots, A_k \in \mathcal{B}$; $g_n \to \infty$ in Γ means that for any finite set $F \subset \Gamma$ there is an N with $n > N \implies g_n \notin F$. For $k = 1$ the property is called simply *mixing*. This notion for single transformations goes back to the foundational work of Rohlin [78], where he showed that ergodic group endomorphisms are mixing of all orders (and so the notion is not useful for distinguishing between group endomorphisms as measurable dynamical systems). He raised the (still open) question of whether any measure-preserving transformation can be mixing without being mixing of all orders.

A class of group actions that are particularly easy to understand are the *algebraic dynamical systems* studied systematically by Schmidt [83]: here X is a compact abelian group, each T^g is a continuous automorphism of X, and μ is the Haar measure on X. Schmidt [82] related mixing properties of algebraic dynamical systems with $\Gamma = \mathbb{Z}^d$ to statements in arithmetic, and showed that a mixing action on a connected group could only fail to mix in a certain way. Later Schmidt and the author [85] showed that for X connected, mixing implies mixing of all orders. The proof proceeds by showing that the result is exactly equivalent to the following statement: if $\mathbb{K}$ is a field of characteristic zero, and G is a finitely generated subgroup of the multiplicative group $\mathbb{K}^\times$, then the equation

$$a_1 x_1 + \cdots + a_n x_n = 1 \tag{13}$$

for fixed $a_1, \ldots, a_n \in \mathbb{K}^\times$ has a finite number of solutions $x_1, \ldots, x_n \in G$ for which no subsum $\sum_{i \in I} a_i x_i$ with $I \subsetneq \{1, \ldots, n\}$ vanishes. The bound on the number of solutions to (13) follows from the profound extensions to W. Schmidt's subspace theorem in Diophantine geometry [86] by Evertse and Schlickewei (see [28,81,102] for the details).

The argument in [85] may be cast as follows: failure of k-fold mixing in a connected algebraic dynamical system implies (via duality) an infinite set of solutions to an equation of the shape (13) in some field of characteristic zero. The S-unit theorem means that this can only happen if there is some proper subsum that vanishes infinitely often. This infinite family of solutions to a homogeneous form of (13) with fewer terms can then be translated back via duality to show that the system fails to mix for some strictly lower order, proving that mixing implies mixing of all orders by induction.

Mixing properties for algebraic dynamical systems without the assumption of connectedness are quite different, and in particular it is possible to have mixing actions that are not mixing of all orders. This is a simple consequence of the fact that the constituents of a disconnected algebraic dynamical system are associated with fields of positive characteristic, where the presence of the

Frobenius automorphism can prevent higher-order mixing. Ledrappier [57] pointed this out via examples of the following shape. Let

$$X = \left\{x \in \mathbb{F}_2^{\mathbb{Z}^2} : x_{(a+1,b)} + x_{(a,b)} + x_{(a,b+1)} = 0 \pmod 2\right\}$$

and define the $\mathbb{Z}^2$-action T to be the natural shift action,

$$(T^{(n,m)}x)_{(a,b)} = x_{(a+n,b+m)} .$$

It is readily seen that this action is mixing with respect to the Haar measure. The condition $x_{(a+1,b)} + x_{(a,b)} + x_{(a,b+1)} = 0 \pmod 2$ implies that, for any $k \geqslant 1$,

$$x_{(0,2^k)} = \sum_{j=0}^{2^k} \binom{2^k}{j} x_{(j,0)} = x_{(0,0)} + x_{(2^k,0)} \pmod 2 \quad (14)$$

since every entry in the 2^kth row of Pascal's triangle is even apart from the first and the last. Now let $A = \{x \in X : x_{(0,0)} = 0\}$ and let $x_* \in X$ be any element with $x_{(0,0)} = 1$. Then X is the disjoint union of A and $A + x_*$, so

$$\mu(A) = \mu(A + x_*) = \tfrac{1}{2} .$$

However, (14) shows that

$$x \in A \cap T_{-(2^k,0)}A \implies x \in T_{-(0,2^k)}A ,$$

so

$$A \cap T_{-(2^k,0)}A \cap T_{-(0,2^k)}(A + x_*) = \emptyset$$

for all $k \geqslant 1$, which shows that T cannot be mixing on three sets.

The full picture of higher-order mixing properties on disconnected groups is rather involved; see Schmidt's monograph [83]. A simple illustration is the construction by Einsiedler and the author [23] of systems with any prescribed order of mixing. When such systems fail to be mixing of all orders, they fail in a very specific way – along dilates of a specific *shape* (a finite subset of $\mathbb{Z}^d$). In the example above, the shape that fails to mix is $\{(0,0),(1,0),(0,1)\}$. This gives an order of mixing as detected by shapes; computing this is in principle an algebraic problem. On the other hand, there is a more natural definition of the order of mixing, namely the largest k for which (12) holds; computing this is in principle a Diophantine problem. A conjecture emerged (formulated explicitly by Schmidt [84]) that for any algebraic dynamical system, if every set of cardinality $r \geqslant 2$ is a mixing shape, then the system is mixing on r sets.

This question motivated Masser [64] to prove an appropriate analogue of the S-unit theorem on the number of solutions to (13) in positive characteristic as follows. Let H be a multiplicative group and fix $n \in \mathbb{N}$. An infinite subset $A \subset H^n$ is called *broad* if it has both of the following properties:

- if $h \in H$ and $1 \leq j \leq n$, then there are at most finitely many $(a_1, \ldots, a_n)$ in A with $a_j = h$;
- if $n \geqslant 2$, $h \in H$ and $1 \leq i < j \leq n$ then there are at most finitely many $(a_1, \ldots, a_n) \in H$ with $a_i a_j^{-1} = h$.

Then Masser's theorem says the following. Let $\mathbb{K}$ be a field of characteristic $p > 0$, let G be a finitely-generated subgroup of $\mathbb{K}^\times$ and suppose that the equation

$$a_1 x_1 + \cdots + a_n x_n = 1$$

has a broad set of solutions $(x_1, \ldots, x_n) \in G^n$ for some constants $a_1, \ldots, a_n \in \mathbb{K}^\times$. Then there is an $m \leq n$, constants $b_1, \ldots, b_m \in \mathbb{K}^\times$ and some $(g_1, \ldots, g_m) \in G^m$ with the following properties:

- $g_j \neq 1$ for $1 \leq j \leq m$;
- $g_i g_j^{-1} \neq 1$ for $1 \leq i < j \leq m$;
- there are infinitely many k for which

$$b_1 g_1^k + b_2 g_2^k + \cdots + b_m g_m^k = 1 .$$

The proof that shapes detect the order of mixing in algebraic dynamics then proceeds much as in the connected case.

Future Directions

The interaction between ergodic theory, number theory and combinatorics continues to expand rapidly, and many future directions of research are discussed in the articles ► Ergodic Theory on Homogeneous Spaces and Metric Number Theory, ► Ergodic Theory: Rigidity and ► Ergodic Theory: Recurrence. Some of the directions most relevant to the examples discussed in this article include the following.

The recent developments mentioned in Subsect. "Sets of Primes" clearly open many exciting prospects involving finding new structures in arithmetically significant sets (like the primes). The original conjecture of Erdős and Turán [26] asked if $\sum_{a \in A \subset \mathbb{N}} \frac{1}{a} = \infty$ is sufficient to force the set A to contain arbitrary long arithmetic progressions, and remains open. This would of course imply both Szemerédi's theorem [97] and the result of Green and Tao [39] on arithmetic progressions in the primes. More generally, it is clear that there is still much to come from the dialogue subsuming the four parallel proofs of

Szemerédi's: one by purely combinatorial methods, one by ergodic theory, one by hypergraph theory, and one by Fourier analysis and additive combinatorics. For an overview, see the survey papers of Tao [98,99,100].

In the context of the orbit-counting results in Sect. "Orbit-Counting as an Analogous Development", a natural problem is to on the one hand obtain finer asymptotics with better control of the error terms, and on the other to extend the situations that can be handled. In particular, relaxing the hypotheses related to hyperbolicity (or negative curvature) is a constant challenge.

Bibliography

Primary Literature

1. Adler RL, Weiss B (1970) Similarity of automorphisms of the torus. Memoirs of the American Mathematical Society, No. 98. American Mathematical Society, Providence
2. Anosov DV (1967) Geodesic flows on closed Riemannian manifolds of negative curvature. Trudy Mat Inst Steklov 90:209
3. Arnol'd VI, Avez A (1968) Ergodic problems of classical mechanics. Translated from the French by A Avez. Benjamin Inc, New York
4. Auslander L, Green L, Hahn F (1963) Flows on homogeneous spaces. Annals of Mathematics Studies, No 53. Princeton University Press, Princeton
5. Baker A (1990) Transcendental number theory. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2nd edn
6. Benford F (1938) The law of anomalous numbers. Proc Amer Philos Soc 78:551–572
7. Bergelson V (1996) Ergodic Ramsey theory – an update. In: Ergodic theory of $\mathbf{Z}^d$ actions (Warwick, 1993–1994), vol 228. London Math Soc Lecture Note Ser. Cambridge University Press, Cambridge, pp 1–61
8. Bergelson V (2000) Ergodic theory and Diophantine problems. In: Topics in symbolic dynamics and applications (Temuco, 1997), vol 279. London Math Soc Lecture Note Ser. Cambridge University Press, Cambridge, pp 167–205
9. Bergelson V (2003) Minimal idempotents and ergodic Ramsey theory. In: Topics in dynamics and ergodic theory, vol 310. London Math Soc, Lecture Note Series, Cambridge University Press, Cambridge, pp 8–39
10. Bergelson V (2006) Combinatorial and Diophantine applications of ergodic theory. In: Handbook of dynamical systems, vol 1B. Elsevier, Amsterdam, pp 745–869
11. Bergelson V, Leibman A (1996) Polynomial extensions of van der Waerden's and Szemerédi's theorems. J Amer Math Soc 9(3):725–753
12. Bergelson V, McCutcheon R (2000) An ergodic IP polynomial Szemerédi theorem. Mem Amer Math Soc 146(695):viii–106
13. Birkhoff GD (1931) Proof of the ergodic theorem. Proc Natl Acad Sci USA 17:656–660
14. Bohl P (1909) Über ein in der Theorie der säkularen Störungen vorkommendes Problem. J Math 135:189–283
15. Borel E (1909) Les probabilités denombrables et leurs applications arithmetiques. Rend Circ Math Palermo 27:247–271
16. Bourgain J (1988) An approach to pointwise ergodic theorems. In: Geometric aspects of functional analysis (1986/87), vol 1317. Lecture Notes in Math, pp 204–223. Springer, Berlin
17. Bourgain J (1988) On the maximal ergodic theorem for certain subsets of the integers. Israel J Math 61(1):39–72
18. Bowen R (1970) Markov partitions for Axiom A diffeomorphisms. Amer J Math 92:725–747
19. Bowen R (1972) The equidistribution of closed geodesics. Amer J Math 94:413–423
20. Bowen R (1973) Symbolic dynamics for hyperbolic flows. Amer J Math 95:429–460
21. Chothi V, Everest G, Ward T (1997) *S*-integer dynamical systems: periodic points. J Reine Angew Math 489:99–132
22. Dolgopyat D (1998) On decay of correlations in Anosov flows. Ann Math 147(2):357–390
23. Einsiedler M, Ward T (2003) Asymptotic geometry of nonmixing sequences. Ergodic Theory Dyn Syst 23(1):75–85
24. Ellis R (1969) Lectures on topological dynamics. Benjamin Inc, New York
25. Erdős P (1949) On a new method in elementary number theory which leads to an elementary proof of the prime number theorem. Proc Natl Acad Sci USA 35:374–384
26. Erdős P, Turán P (1936) On some sequences of integers. J London Math Soc 11:261–264
27. Everest G, Miles R, Stevens S, Ward T (2007) Orbit-counting in non-hyperbolic dynamical systems. J Reine Angew Math 608:155–182
28. Evertse JH, Schlickewei HP (1999) The absolute subspace theorem and linear equations with unknowns from a multiplicative group. In: Number theory in progress, vol 1 (Zakopane-Kościelisko, 1997). de Gruyter, Berlin, pp 121–142
29. Furstenberg H (1961) Strict ergodicity and transformation of the torus. Amer J Math 83:573–601
30. Furstenberg H (1977) Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. J Analyse Math 31:204–256
31. Furstenberg H, Katznelson Y (1979) An ergodic Szemerédi theorem for commuting transformations. J Analyse Math 34:275–291
32. Furstenberg H, Katznelson Y (1985) An ergodic Szemerédi theorem for IP-systems and combinatorial theory. J Analyse Math 45:117–168
33. Furstenberg H, Katznelson Y, Ornstein D (1982) The ergodic theoretical proof of Szemerédi's theorem. Bull Amer Math Soc 7(3):527–552
34. Furstenberg H, Weiss B (1979) Topological dynamics and combinatorial number theory. J Analyse Math 34:61–85
35. Gel'fond AO (1960) Transcendental and algebraic numbers. Translated from the first Russian edition by Leo F Boron. Dover Publications, New York
36. Goldfeld D (2004) The elementary proof of the prime number theorem: an historical perspective. In: Number theory (2003). Springer, New York, pp 179–192
37. Goldston DA, Yıldırım CY (2005) Small gaps between primes I. arXiv:math.NT/0504336
38. Gowers WT (2007) Hypergraph regularity and the multidimensional Szemerédi Theorem. Ann of Math 166:897–946
39. Green B, Tao T (2004) The primes contain arbitrarily long arithmetic progressions. arXiv:math.NT/0404188
40. Green B, Tao T (2006) Linear equations in primes. arXiv:math.NT/0606088

41. Hejhal DA (1976) The Selberg trace formula and the Riemann zeta function. Duke Math J 43(3):441–482
42. Hill TP (1995) Base-invariance implies Benford's law. Proc Amer Math Soc 123(3):887–895
43. Hindman N (1974) Finite sums from sequences within cells of a partition of $\mathbb{N}$. J Comb Theory Ser A 17:1–11
44. Hlawka E (1964) Discrepancy and uniform distribution of sequences. Compositio Math 16:83–91
45. Host B, Kra B (2005) Convergence of polynomial ergodic averages. Israel J Math 149:1–19, Probability in mathematics
46. Host B, Kra B (2005) Nonconventional ergodic averages and nilmanifolds. Ann Math 161(1):397–488
47. Huber H (1959) Zur analytischen Theorie hyperbolischen Raumformen und Bewegungsgruppen. Math Ann 138:1–26
48. Katok A (1980) Lyapunov exponents, entropy and periodic orbits for diffeomorphisms. Inst Hautes Études Sci Publ Math (51):137–173
49. Katsuda A, Sunada T (1990) Closed orbits in homology classes. Inst Hautes Études Sci Publ Math (71):5–32
50. Khinchin AI (1964) Continued fractions. The University of Chicago Press, Chicago
51. Knieper G (1997) On the asymptotic geometry of nonpositively curved manifolds. Geom Funct Anal 7(4):755–782
52. Koopman B (1931) Hamiltonian systems and transformations in Hilbert spaces. Proc Natl Acad Sci USA 17:315–318
53. Kra B (2006) The Green-Tao theorem on arithmetic progressions in the primes: an ergodic point of view. Bull Amer Math Soc 43(1):3–23
54. Kuz'min RO (1928) A problem of Gauss. Dokl Akad Nauk, pp 375–380
55. Lalley SP (1987) Distribution of periodic orbits of symbolic and Axiom A flows. Adv Appl Math 8(2):154–193
56. Lalley SP (1988) The "prime number theorem" for the periodic orbits of a Bernoulli flow. Amer Math Monthly 95(5):385–398
57. Ledrappier F (1978) Un champ markovien peut être d'entropie nulle et mélangeant. C R Acad Sci Paris Sér A-B 287(7):A561–A563
58. Leibman A (2005) Convergence of multiple ergodic averages along polynomials of several variables. Israel J Math 146:303–315
59. Levy P (1929) Sur les lois de probabilité dont dependent les quotients complets et incomplets d'une fraction continue. Bull Soc Math France 57:178–194
60. Levy P (1936) Sur quelques points de la théorie des probabilités dénombrables. Ann Inst H Poincaré 6(2):153–184
61. Lind DA (1982) Dynamical properties of quasihyperbolic toral automorphisms. Ergodic Theory Dyn Syst 2(1):49–68
62. Margulis GA (1969) Certain applications of ergodic theory to the investigation of manifolds of negative curvature. Funkcional Anal i Priložen 3(4):89–90
63. Margulis GA (2004) On some aspects of the theory of Anosov systems. Springer Monographs in Mathematics. Springer, Berlin.
64. Masser DW (2004) Mixing and linear equations over groups in positive characteristic. Israel J Math 142:189–204
65. Mertens F (1874) Ein Beitrag zur analytischen Zahlentheorie. J Reine Angew Math 78:46–62
66. Nagle B, Rödl V, Schacht M (2006) The counting lemma for regular *k*-uniform hypergraphs. Random Structures Algorithms 28(2):113–179
67. Newcomb S (1881) Note on the frequency of the use of digits in natural numbers. Amer J Math 4(1):39–40
68. Noorani MS (1999) Mertens' theorem and closed orbits of ergodic toral automorphisms. Bull Malaysian Math Soc 22(2):127–133
69. Oxtoby JC (1952) Ergodic sets. Bull Amer Math Soc 58:116–136
70. Parry W (1969) Ergodic properties of affine transformations and flows on nilmanifolds. Amer J Math 91:757–771
71. Parry W (1983) An analogue of the prime number theorem for closed orbits of shifts of finite type and their suspensions. Israel J Math 45(1):41–52
72. Parry W (1984) Bowen's equidistribution theory and the Dirichlet density theorem. Ergodic Theory Dyn Syst 4(1):117–134
73. Parry W, Pollicott M (1983) An analogue of the prime number theorem for closed orbits of Axiom A flows. Ann Math 118(3):573–591
74. Poincaré H (1890) Sur le probléme des trois corps et les équations de la Dynamique. Acta Math 13:1–270
75. Pollicott M, Sharp R (1998) Exponential error terms for growth functions on negatively curved surfaces. Amer J Math 120(5):1019–1042
76. Rado R (1933) Studien zur Kombinatorik. Math Z 36(1):424–470
77. Ratner M (1973) Markov partitions for Anosov flows on *n*-dimensional manifolds. Israel J Math 15:92–114
78. Rohlin VA (1949) On endomorphisms of compact commutative groups. Izvestiya Akad Nauk SSSR Ser Mat 13:329–340
79. Roth K (1952) Sur quelques ensembles d'entiers. C R Acad Sci Paris 234:388–390
80. Sárközy A (1978) On difference sets of sequences of integers. III. Acta Math Acad Sci Hungar 31(3-4):355–386
81. Schlickewei HP (1990) *S*-unit equations over number fields. Invent Math 102(1):95–107
82. Schmidt K (1989) Mixing automorphisms of compact groups and a theorem by Kurt Mahler. Pacific J Math 137(2):371–385
83. Schmidt K (1995) Dynamical systems of algebraic origin, vol 128. Progress in Mathematics. Birkhäuser, Basel
84. Schmidt K (2001) The dynamics of algebraic $\mathbb{Z}^d$-actions. In: European Congress of Mathematics, vol I (Barcelona, 2000), vol 201. Progress in Math. Birkhäuser, Basel, pp 543–553
85. Schmidt K, Ward T (1993) Mixing automorphisms of compact groups and a theorem of Schlickewei. Invent Math 111(1):69–76
86. Schmidt WM (1972) Norm form equations. Ann Math 96(2):526–551
87. Selberg A (1949) An elementary proof of the prime-number theorem. Ann Math 50(2):305–313
88. Selberg A (1956) Harmonic analysis and discontinuous groups in weakly symmetric Riemannian spaces with applications to Dirichlet series. J Indian Math Soc 20:47–87
89. Sharp R (1991) An analogue of Mertens' theorem for closed orbits of Axiom A flows. Bol Soc Brasil Mat 21(2):205–229
90. Sharp R (1993) Closed orbits in homology classes for Anosov flows. Ergodic Theory Dyn Syst 13(2):387–408
91. Sierpiński W (1910) Sur la valeur asymptotique d'une certaine somme. Bull Intl Acad Polonmaise des Sci et des Lettres (Cracovie), pp 9–11
92. Sinaĭ JG (1966) Asymptotic behavior of closed geodesics on compact manifolds with negative curvature. Izv Akad Nauk SSSR Ser Mat 30:1275–1296

93. Sinaĭ JG (1968) Construction of Markov partitionings. Funkcional Anal i Priložen 2(3):70–80
94. Silverman JH (2007) The Arithmetic of Dynamical Systems, vol 241. Graduate Texts in Mathematics. Springer, New York
95. Smale S (1967) Differentiable dynamical systems. Bull Amer Math Soc 73:747–817
96. Szemerédi E (1969) On sets of integers containing no four elements in arithmetic progression. Acta Math Acad Sci Hungar 20:89–104
97. Szemerédi E (1975) On sets of integers containing no *k* elements in arithmetic progression. Acta Arith 27:199–245
98. Tao T (2005) The dichotomy between structure and randomness, arithmetic progressions, and the primes. arXiv: math/0512114v2
99. Tao T (2006) Arithmetic progressions and the primes. Collect Math, vol extra, Barcelona, pp 37–88
100. Tao T (2007) What is good mathematics? arXiv:math/ 0702396v1
101. Tao T, Ziegler T (2006) The primes contain arbitrarily long polynomial progressions. arXiv:math.NT/0610050
102. van der Poorten AJ, Schlickewei HP (1991) Additive relations in fields. J Austral Math Soc Ser A 51(1):154–170
103. van der Waerden BL (1927) Beweis einer Baudet'schen Vermutung. Nieuw Arch Wisk 15:212–216
104. van der Waerden BL (1971) How the proof of Baudet's conjecture was found. In: Studies in Pure Mathematics (Presented to Richard Rado). Academic Press, London, pp 251–260
105. von Neumann J (1932) Proof of the quasi-ergodic hypothesis. Proc Natl Acad Sci USA 18:70–82
106. Waddington S (1991) The prime orbit theorem for quasihyperbolic toral automorphisms. Monatsh Math 112(3):235–248
107. Ward T (1998) Almost all *S*-integer dynamical systems have many periodic points. Ergodic Theory Dyn Syst 18(2):471–486
108. Weil A (1967) Basic number theory. Die Grundlehren der mathematischen Wissenschaften, Band 144. Springer, New York
109. Weyl H (1910) Über die Gibbssche Erscheinung und verwandte Konvergenzphänomene. Rendiconti del Circolo Matematico di Palermo 30:377–407
110. Weyl H (1916) Uber die Gleichverteilung von Zahlen mod Eins. Math Ann 77:313–352
111. Wiener N (1932) Tauberian theorems. Ann Math 33(1):1–100

Books and Reviews

Cornfeld IP, Fomin SV, Sinaĭ YG (1982) Ergodic theory. Springer, New York
Dajani K, Kraaikamp C (2002) Ergodic theory of numbers, vol 29. In: Carus Mathematical Monographs. Mathematical Association of America, Washington DC
Denker M, Grillenberger C, Sigmund K (1976) Ergodic theory on compact spaces. Lecture Notes in Mathematics, vol 527. Springer, Berlin
Furstenberg H (1981) Recurrence in ergodic theory and combinatorial number theory. Princeton University Press, Princeton
Glasner E (2003) Ergodic theory via joinings, vol 101. Mathematical Surveys and Monographs. American Mathematical Society, Providence
Iosifescu M, Kraaikamp C (2002) Metrical theory of continued fractions, vol 547. Mathematics and its Applications. Kluwer, Dordrecht
Katok A, Hasselblatt B (1995) Introduction to the modern theory of dynamical systems, vol 54. In: Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge
Krengel U (1985) Ergodic theorems, vol 6. de Gruyter Studies in Mathematics. de Gruyter, Berlin
McCutcheon R (1999) Elemental methods in ergodic Ramsey theory, vol 1722. Lecture Notes in Mathematics. Springer, Berlin
Petersen K (1989) Ergodic theory, vol 2. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge
Schweiger F (1995) Ergodic theory of fibred systems and metric number theory. Oxford Science Publications. The Clarendon Press Oxford University Press, New York
Totoki H (1969) Ergodic theory. Lecture Notes Series, No 14. Matematisk Institut, Aarhus Universitet, Aarhus
Walters P (1982) An introduction to ergodic theory, vol 79. Graduate Texts in Mathematics. Springer, New York

Ergodic Theory, Introduction to

Bryna Kra
Northwestern University, Evanston, USA

Ergodic theory lies at the intersection of many areas of mathematics, including smooth dynamics, statistical mechanics, probability, harmonic analysis, and group actions. Problems, techniques, and results are related to many other areas of mathematics, and ergodic theory has had applications both within mathematics and to numerous other branches of science. Ergodic theory has particularly strong overlap with other branches of dynamical systems; to clarify what distinguishes it from other areas of dynamics, we start with a quick overview of dynamical systems.

Dynamical systems is the study of systems that evolve with time. The evolution of a dynamical system is given by some fixed rule that determines the states of the system a short time into the future, given only the present states. Reflecting the origins of the subject in celestial mechanics, the set of states through which the system evolves with time is called an orbit. Many important concepts in dynamical systems are related to understanding the orbits in the system: Do the orbits fill out the entire space? Do orbits collapse? Do orbits return to themselves? What are statistical properties of the orbits? Are orbits stable under perturbation? For simple dynamical systems, knowing the individual orbits is often sufficient to answer such questions. However, in most dynamical systems it is impossible to write down explicit formulae for orbits, and even when one can, many systems are too complicated to be understood just in terms of individual orbits. The orbits may only be known approximately, some orbits may appear to be random while others exhibit regular behavior, and vary-

ing the parameters defining the system may give rise to qualitatively different behaviors. The various branches of dynamical systems have been developed for understanding long term properties of the orbits.

To make the notion of a dynamical system more precise, let X denote the collection of all states of the system. The evolution of these states is given by some fixed rule $T: X \to X$, dictating where each state $x \in X$ is mapped. An application of the transformation $T: X \to X$ corresponds to the passage of a unit of time and for a positive integer n, the map $T^n = T \circ T \circ \ldots \circ T$ denotes the composition of T with itself taken n times. Given a state $x \in X$, the orbit of the point x under the transformation T is the collection of iterates $x, Tx, T^2x, \ldots$ of the state x. Thus the single transformation T generates a semigroup of transformations acting on X, by considering the powers T^n. More generally, one can consider a family of transformations $\{T_t : t \in \mathbb{R}\}$ with each $T_t: X \to X$. Assuming that $T_0(x) = x$ and that $T_{t+s}(x) = T_t(T_s(x))$ for all states $x \in X$ and all real t and s, this models the evolution of continuous time in a system. Autonomous differential equations are examples of such continuous time systems.

In almost all cases of interest, the space X has some underlying structure which is preserved by the transformation T. Different underlying structures X and different properties of the transformation T give rise to different branches of dynamical systems. When X is a smooth manifold and $T: X \to X$ is a differentiable mapping, one is in the framework of differentiable dynamics. When X is a topological space and $T: X \to X$ is a continuous map, one is in the framework of topological dynamics. When X is a measure space and $T: X \to X$ is a measure preserving map, one is in the framework of ergodic theory. These categories are not mutually exclusive, and the relations among them are deep and interesting. Some of these relations are explored in the articles ▶ Topological Dynamics, ▶ Symbolic Dynamics, and ▶ Smooth Ergodic Theory.

To further explain the role of ergodic theory, a few definitions are needed. The state space X is assumed to be a measure space, endowed with a σ-algebra $\mathcal{B}$ of measurable sets and a measure μ. The measure μ assigns each set $B \in \mathcal{B}$ a non-negative number (its measure), and usually one assumes that $\mu(X) = 1$ (thus μ is a probability measure). The transformation $T: X \to X$ is assumed to be a measurable and measure preserving map: For all $B \in \mathcal{B}$, $\mu(T^{-1}B) = \mu(B)$. The quadruple $(X, \mathcal{B}, \mu, T)$ is called a measure preserving system. For the precise definitions and background on measure preserving transformations, see the article ▶ Measure Preserving Systems. An extensive discussion of examples of measure preserving systems and basic constructions used in ergodic theory is given in ▶ Ergodic Theory: Basic Examples and Constructions.

The origins of ergodic theory are in the nineteenth century work of Boltzmann on the foundations of statistical mechanics. Boltzmann hypothesized that for large systems of interacting particles in equilibrium, the "time average" is equal to the "space average". This question can be reformulated in the context of modern terminology. Assume that $(X, \mathcal{B}, \mu, T)$ is a measure preserving system and that $f: X \to \mathbb{R}$ is some measurement taken on the system. Thus if $x \in X$ is a state, evaluating the sequence

$$f(x)\,, \quad f(Tx)\,, \quad f(T^2x), \ \ldots$$

can be viewed as successive values of this measurement. Boltzmann's question can be phrased as: Under what conditions is the time mean equal to the space mean? In short, when does

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T^n x) = \int_X f \, \mathrm{d}\mu \ ?$$

Boltzmann hypothesized that if orbits went "everywhere" in the space, then such a conclusion would hold.

The study of the equality of space and time averages has been a major direction of research in ergodic theory. The long term behavior of the average

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T^n x)\,,$$

and especially the existence of this limit, is a basic question. Roughly speaking, the ergodic theorem states that starting at almost any initial point, the distribution of its iterates obeys some asymptotic law. This, and more general convergence questions, are addressed in the article ▶ Ergodic Theorems.

Perhaps the earliest result in ergodic theory is the Poincaré Recurrence Theorem: In a finite measure space, some iterate of any set with positive measure intersects the set itself in a set of positive measure. More generally, the qualitative behavior of orbits is used to understand conditions under which the time average is equal to the space average of the system (see the article ▶ Ergodic Theory: Recurrence). If the time average is equal almost everywhere to the space average, then the system is said to be ergodic. Ergodicity is a key notion, giving a simple expression for the time average of an arbitrary function. Moreover, using the Ergodic Decomposition Theorem, the study of arbitrary measure preserving systems can be reduced to ergodic ones. Ergodicity and related properties of a system

are discussed in the article ► Ergodicity and Mixing Properties.

Another central problem in ergodic theory is the classification of measure preserving systems. There are various notions of equivalence, and a classical approach to checking if systems are equivalent is finding invariants that are preserved under the equivalence. This subject, including an introduction to Ornstein Theory, is covered in the article ► Isomorphism Theory in Ergodic Theory. A map $T\colon X \to X$ determines an associated unitary operator $U = U_T$ defined on $L^2(X)$ by

$$U_T f(x) = f(Tx)\,.$$

There are also numerical invariants that can be assigned to a system, for example entropy, and this is discussed in the article ► Entropy in Ergodic Theory.

When two systems are not equivalent, one would like to understand what properties they do have in common. An essential tool in such a classification is the notion of joinings (see the article ► Joinings in Ergodic Theory). Roughly speaking, a joining is a way of embedding two systems in the same space. When this can be done in a nontrivial manner, one obtains information on properties shared by the systems.

Some systems have predictable behavior and can be classified according to the behavior of individual points and their iterates. Others have behavior that is too complex or unpredictable to be understood on the level of orbits. Ergodic theory provides a statistical understanding of such systems and this is discussed in the article ► Chaos and Ergodic Theory. A prominent role in chaotic dynamical systems is played by one dimensional Gibbs measure and by equilibrium states and (see the article ► Pressure and Equilibrium States in Ergodic Theory). Rigidity theory addresses the opposite case, studying what kinds of properties in a system are obstructions to general chaotic behavior. The role of ergodic theory in this area is discussed in the article ► Ergodic Theory: Rigidity.

Another important class of systems arises when one relaxes the condition that the transformation T preserves the measure of sets on X, only requiring that the transformation preserve the negligible sets. Such systems are discussed in ► Joinings in Ergodic Theory.

Ergodic theory has seen a burst of recent activity, and most of this activity comes from interaction with other fields. Historically, ergodic theory has interacted with numerous fields, including other areas of dynamics, probability, statistical mechanics, and harmonic analysis. More recently, ergodic theory and its techniques have been imported into number theory and combinatorics, proving new results that have yet to be proved by other methods, and in turn, combinatorial problems have given rise to new areas of research within ergodic theory itself. Problems related to Diophantine approximation are discussed in ► Ergodic Theory on Homogeneous Spaces and Metric Number Theory and ► Ergodic Theory: Rigidity and ones related to combinatorial problems are addressed in ► Ergodic Theory: Interactions with Combinatorics and Number Theory and in ► Ergodic Theory: Recurrence. Interaction with problems that are geometric in nature, in particular dimension theory, is discussed in ► Ergodic Theory: Fractal Geometry.

Ergodic Theory: Non-singular Transformations

Alexandre I. Danilenko[1], Cesar E. Silva[2]
[1] Institute for Low Temperature Physics & Engineering, Ukrainian National Academy of Sciences, Kharkov, Ukraine
[2] Department of Mathematics, Williams College, Williamstown, USA

Article Outline

Glossary
Definition of the Subject
Basic Results
Panorama of Examples
Mixing Notions and multiple recurrence
Topological Group Aut(X, μ)
Orbit Theory
Smooth Nonsingular Transformations
Spectral Theory for Nonsingular Systems
Entropy and Other Invariants
Nonsingular Joinings and Factors
Applications. Connections with Other Fields
Concluding Remarks
Bibliography

Glossary

Nonsingular dynamical system Let $(X, \mathcal{B}, \mu)$ be a standard Borel space equipped with a σ-finite measure. A Borel map $T\colon X \to X$ is a *nonsingular transformation* of X if for any $N \in \mathcal{B}$, $\mu(T^{-1}N) = 0$ if and only if $\mu(N) = 0$. In this case the measure μ is called *quasi-invariant* for T; and the quadruple $(X, \mathcal{B}, \mu, T)$ is called a *nonsingular dynamical system*. If $\mu(A) = \mu(T^{-1}A)$ for all $A \in \mathcal{B}$ then μ is said to be

invariant under T or, equivalently, T is *measure-preserving*.

Conservativeness T is *conservative* if for all sets A of positive measure there exists an integer $n > 0$ such that $\mu(A \cap T^{-n}A) > 0$.

Ergodicity T is *ergodic* if every measurable subset A of X that is invariant under T (i. e., $T^{-1}A = A$) is either μ-null or μ-conull. Equivalently, every Borel function $f\colon X \to \mathbb{R}$ such that $f \circ T = f$ is constant a. e.

Types II, II_1, II_∞ and III Suppose that μ is non-atomic and T ergodic (and hence conservative). If there exists a σ-finite measure ν on $\mathcal{B}$ which is equivalent to μ and invariant under T then T is said *to be of type II*. It is easy to see that ν is unique up to scaling. If ν is finite then T is *of type II_1*. If ν is infinite then T is of type II_∞. If T is not of type II then T is said *to be of type III*.

Definition of the Subject

An abstract measurable dynamical system consists of a set X (phase space) with a transformation $T\colon X \to X$ (evolution law or time) and a finite or σ-finite measure μ on X that specifies a class of negligible subsets. Nonsingular ergodic theory studies systems where T respects μ in a weak sense: the transformation preserves only the class of negligible subsets but it may not preserve μ. This survey is about dynamics and invariants of nonsingular systems. Such systems model 'non-equilibrium' situations in which events that are impossible at some time remain impossible at any other time. Of course, the first question that arises is whether it is possible to find an equivalent invariant measure, i. e. pass to a hidden equilibrium without changing the negligible subsets? It turns out that there exist systems which do not admit an equivalent invariant finite or even σ-finite measure. They are of our primary interest here. In a way (Baire category) most of systems are like that.

Nonsingular dynamical systems arise naturally in various fields of mathematics: topological and smooth dynamics, probability theory, random walks, theory of numbers, von Neumann algebras, unitary representations of groups, mathematical physics and so on. They also can appear in the study of probability preserving systems: some criteria of mild mixing and distality, a problem of Furstenberg on disjointness, etc. We briefly discuss this in Sect. "Applications. Connections with Other Fields". Nonsingular ergodic theory studies all of them from a general point of view:

- What is the qualitative nature of the dynamics?
- What are the orbits?
- Which properties are typical withing a class of systems?
- How do we find computable invariants to compare or distinguish various systems?

Typically there are two kinds of results: some are extensions to nonsingular systems of theorems for finite measure-preserving transformations (for instance, the entire Sect. "Basic Results") and the other are about new properly 'nonsingular' phenomena (see Sect. "Mixing Notions and Multiple Recurrence" or Sect. "Orbit Theory"). Philosophically speaking, the dynamics of nonsingular systems is more diverse comparatively with their finite measure-preserving counterparts. That is why it is usually easier to construct counterexamples than to develop a general theory. Because of shortage of space we concentrate only on invertible transformations, and we have not included as many references as we had wished. Nonsingular endomorphisms and general group or semigroup actions are practically not considered here (with some exceptions in Sect. "Applications. Connections with Other Fields" devoted to applications). A number of open problems are scattered through the entire text.

We thank J. Aaronson, J.R. Choksi, V.Ya. Golodets, M. Lemańczyk, F. Parreau, E. Roy for useful remarks.

Basic Results

This section includes the basic results involving conservativeness and ergodicity as well as some direct nonsingular counterparts of the basic machinery from classic ergodic theory: mean and pointwise ergodic theorems, Rokhlin lemma, ergodic decomposition, generators, Glimm–Effros theorem and special representation of nonsingular flows. The historically first example of a transformation of type III (due to Ornstein) is also given here with full proof.

Nonsingular Transformations

In this paper we will consider only *invertible* nonsingular transformations, i. e. those which are bijections when restricted to an invariant Borel subset of full measure. Thus when we refer to a nonsingular dynamical system $(X, \mathcal{B}, \mu, T)$ we shall assume that T is an invertible nonsingular transformation. Of course, each measure ν on $\mathcal{B}$ which is *equivalent* to μ, i. e. μ and ν have the same null sets, is also quasi-invariant under T. In particular, since μ is σ-finite, T admits an equivalent quasi-invariant probability measure. For each $i \in \mathbb{Z}$, we denote by ω_i^μ or ω_i the Radon–Nikodym derivative $\mathrm{d}(\mu \circ T^i)/\mathrm{d}\mu \in L^1(X, \mu)$. The derivatives satisfy the cocycle equation $\omega_{i+j}(x) = \omega_i(x)\,\omega_j(T^i x)$ for a. e. x and all $i, j \in \mathbb{Z}$.

Basic Properties of Conservativeness and Ergodicity

A measurable set W is said to be *wandering* if for all $i, j \geq 0$ with $i \neq j$, $T^{-i}W \cap T^{-j}W = \emptyset$. Clearly, if T has a wandering set of positive measure then it cannot be conservative. A nonsingular transformation T is *incompressible* if whenever $T^{-1}C \subset C$, then $\mu(C \setminus T^{-1}C) = 0$. A set W of positive measure is said to be *weakly wandering* if there is a sequence $n_i \to \infty$ such that $T^{n_i}W \cap T^{n_j}W = \emptyset$ for all $i \neq j$. Clearly, a finite measure-preserving transformation cannot have a weakly wandering set. Hajian and Kakutani [83] showed that a nonsingular transformation T admits an equivalent finite invariant measure if and only if T does not have a weakly wandering set.

Proposition 1 (see e. g. [123]) *Let $(X, \mathcal{B}, \mu, T)$ be a nonsingular dynamical system. The following are equivalent:*

(i) T is conservative.
(ii) For every measurable set A, $\mu(A \setminus \bigcup_{n=1}^{\infty} T^{-n}A) = 0$.
(iii) T is incompressible.
(iv) Every wandering set for T is null.

Since any finite measure-preserving transformation is incompressible, we deduce that it is conservative. This is the statement of the classical Poincaré recurrence lemma. If T is a conservative nonsingular transformation of $(X, \mathcal{B}, \mu)$ and $A \in \mathcal{B}$ a subset of positive measure, we can define an *induced transformation* T_A of the space $(A, \mathcal{B} \cap A, \mu \restriction A)$ by setting $T_A x := T^n x$ if $n = n(x)$ is the smallest natural number such that $T^n x \in A$. T_A is also conservative. As shown in [179], if $\mu(X) = 1$ and T is conservative and ergodic, $\int_A \sum_{i=0}^{n(x)-1} \omega(x)\,\mathrm{d}\mu(x) = 1$, which is a nonsingular version of the well-known Kaçs formula.

Theorem 2 (Hopf Decomposition, see e. g. [3]) *Let T be a nonsingular transformation. Then there exist disjoint invariant sets $C, D \in \mathcal{B}$ such that $X = C \sqcup D$, T restricted to C is conservative, and $D = \bigsqcup_{n=-\infty}^{\infty} T^n W$, where W is a wandering set. If $f \in L^1(X, \mu)$, $f > 0$, then $C = \{x\colon \sum_{i=0}^{n-1} f(T^i x)\,\omega_i(x) = \infty \text{ a. e.}\}$ and $D = \{x\colon \sum_{i=0}^{n-1} f(T^i x)\,\omega_i(x) < \infty \text{ a. e.}\}$.*

The set C is called the *conservative part* of T and D is called the *dissipative part* of T.

If T is ergodic and μ is non-atomic then T is automatically conservative. The translation by 1 on the group $\mathbb{Z}$ furnished with the counting measure is an example of an ergodic non-conservative (infinite measure-preserving) transformation.

Proposition 4 *Let $(X, \mathcal{B}, \mu, T)$ be a nonsingular dynamical system. The following are equivalent:*

(i) T is conservative and ergodic.
(ii) For every set A of positive measure, $\mu(X \setminus \bigcup_{n=1}^{\infty} T^{-n}A) = 0$. (In this case we will say A sweeps out.)
(iii) For every measurable set A of positive measure and for a. e. $x \in X$ there exists an integer $n > 0$ such that $T^n x \in A$.
(iv) For all sets A and B of positive measure there exists an integer $n > 0$ such that $\mu(T^{-n}A \cap B) > 0$.
(v) If A is such that $T^{-1}A \subset A$, then $\mu(A) = 0$ or $\mu(A^c) = 0$.

This survey is mainly about systems of type III. For some time it was not quite obvious whether such systems exist at all. The historically first example was constructed by Ornstein in 1960.

Example 5 (Ornstein [149]) Let $A_n = \{0, 1, \ldots, n\}$, $\nu_n(0) = 0.5$ and $\nu_n(i) = 1/(2n)$ for $0 < i \leq n$ and all $n \in \mathbb{N}$. Denote by (X, μ) the infinite product probability space $\bigotimes_{n=1}^{\infty}(A_n, \nu_n)$. Of course, μ is non-atomic. A point of X is an infinite sequence $x = (x_n)_{n=1}^{\infty}$ with $x_n \in A_n$ for all n. Given $a_1 \in A_1, \ldots, a_n \in A_n$, we denote the cylinder $\{x = (x_i)_{i=1}^{\infty} \in X\colon x_1 = a_1, \ldots, x_n = a_n\}$ by $[a_1, \ldots, a_n]$. Define a Borel map $T\colon X \to X$ by setting

$$(Tx)_i = \begin{cases} 0, & \text{if } i < l(x) \\ x_i + 1, & \text{if } i = l(x) \\ x_i, & \text{if } i > l(x)\,, \end{cases} \tag{1}$$

where $l(x)$ is the smallest number l such that $x_l \neq l$. It is easy to verify that T is a nonsingular transformation of (X, μ) and

$$\frac{\mathrm{d}\mu \circ T}{\mathrm{d}\mu}(x) = \prod_{n=1}^{\infty} \frac{\nu_n((Tx)_n)}{\nu_n(x_n)} = \begin{cases} \dfrac{(l(x)-1)!}{l(x)}, & \text{if } x_{l(x)} = 0 \\ (l(x)-1)!\,, & \text{if } x_{l(x)} \neq 0\,. \end{cases}$$

We prove that T is of type III by contradiction. Suppose that there exists a T-invariant σ-finite measure ν equivalent to μ. Let $\varphi := \mathrm{d}\mu/\mathrm{d}\nu$. Then

$$\omega_i^{\mu}(x) = \varphi(x)\,\varphi(T^i x)^{-1} \text{ for a. a. } x \in X \text{ and all } i \in \mathbb{Z}. \tag{2}$$

Fix a real $C > 1$ such that the set $E_C := \varphi^{-1}([C^{-1}, C]) \subset X$ is of positive measure. By a standard approximation argument, for each sufficiently large n, there is a cylinder $[a_1, \ldots, a_n]$ such that $\mu(E_C \cap [a_1, \ldots, a_n]) > 0.9\mu([a_1, \ldots, a_n])$. Since $\nu_{n+1}(0) = 0.5$, it follows that $\mu(E_C \cap [a_1, \ldots, a_n, 0]) > 0.8\mu([a_1, \ldots, a_n, 0])$. Moreover, by the pigeon hole principle there is $0 < i \leq n+1$ with $\mu(E_C \cap$

$[a_1, \dots, a_n, i]) > 0.8\mu([a_1, \dots, a_n, i])$. Find $N_n > 0$ such that $T^{N_n}[a_1, \dots, a_n, 0] = [a_1, \dots, a_n, i]$. Since $\omega^{\mu}_{N_n}$ is constant on $[a_1, \dots, a_n, 0]$, there is a subset $E_0 \subset E_C \cap [a_1, \dots, a_n, 0]$ of positive measure such that $T^{N_n}E_0 \subset E_C \cap [a_1, \dots, a_n, i]$. Moreover, $\omega^{\mu}_{N_n}(x) = \nu_{n+1}(i)/\nu_{n+1}(0) = (n+1)^{-1}$ for a. a. $x \in [a_1, \dots, a_n, 0]$. On the other hand, we deduce from (2) that $\omega^{\mu}_{N_n}(x) \geq C^{-2}$ for all $x \in E_0$, a contradiction.

Mean and Pointwise Ergodic Theorems. Rokhlin Lemma

Let $(X, \mathcal{B}, \mu, T)$ be a nonsingular dynamical system. Define a unitary operator U_T of $L^2(X, \mu)$ by setting

$$U_T f := \sqrt{\frac{\mathrm{d}(\mu \circ T)}{\mathrm{d}\mu}} \cdot f \circ T \,. \tag{3}$$

We note that U_T preserves the cone of positive functions $L^2_+(X, \mu)$. Conversely, every positive unitary operator in $L^2(X, \mu)$ that preserves $L^2_+(X, \mu)$ equals U_T for a μ-nonsingular transformation T.

Theorem 6 (von Neumann mean Ergodic Theorem, see e. g. [3]) *If T has no μ-absolutely continuous T-invariant probability, then $n^{-1}\sum_{i=0}^{n-1} U_T^i \to 0$ in the strong operator topology.*

Denote by $\mathcal{I}$ the sub-σ-algebra of T-invariant sets. Let $\mathbb{E}_\mu[.|\mathcal{I}]$ stand for the conditional expectation with respect to $\mathcal{I}$. Note that if T is ergodic, then $\mathbb{E}_\mu[f|\mathcal{I}] = \int f \,\mathrm{d}\mu$. Now we state a nonsingular analogue of Birkhoff's pointwise ergodic theorem, due to Hurewicz [105] and in the form stated by Halmos [84].

Theorem 7 (Hurewicz pointwise Ergodic Theorem) *If T is conservative, $\mu(X) = 1$, $f, g \in L^1(X, \mu)$ and $g > 0$, then*

$$\frac{\sum_{i=0}^{n-1} f(T^i x)\,\omega_i(x)}{\sum_{i=0}^{n-1} g(T^i x)\,\omega_i(x)} \to \frac{\mathbb{E}_\mu[f|\mathcal{I}]}{\mathbb{E}_\mu[g|\mathcal{I}]} \quad \text{as } n \to \infty \quad \text{for a. e. } x \,.$$

A transformation T is *aperiodic* if the T-orbit of a. e. point from X is infinite. The following classical statement can be deduced easily from Proposition 1.

Lemma 8 (Rokhlin's lemma [161]) *Let T be an aperiodic nonsingular transformation. For each $\varepsilon > 0$ and integer $N > 1$ there exists a measurable set A such that the sets $A, TA, \dots, T^{N-1}A$ are disjoint and $\mu(A \cup TA \cup \dots \cup T^{N-1}A) > 1 - \varepsilon$.*

This lemma was refined later (for ergodic transformations) by Lehrer and Weiss as follows.

Theorem 9 (ε-free Rokhlin lemma [132]) *Let T be ergodic and μ non-atomic. Then for a subset $B \subset X$ and any N for which $\bigcup_{k=0}^{\infty} T^{-kN}(X \setminus B) = X$, there is a set A such that the sets $A, TA, \dots, T^{N-1}A$ are disjoint and $A \cup TA \cup \dots \cup T^{N-1}A \supset B$.*

The condition $\bigcup_{k=0}^{\infty} T^{-kN}(X \setminus B) = X$ holds of course for each $B \neq X$ if T is *totally ergodic*, i. e. T^p is ergodic for any p, or if N is prime.

Ergodic Decomposition

A proof of the following theorem may be found in [3].

Theorem 10 (Ergodic Decomposition Theorem) *Let T be a conservative nonsingular transformation on a standard probability space $(X, \mathcal{B}, \mu)$. There there exists a standard probability space $(Y, \nu, \mathcal{A})$ and a family of probability measures μ_y on $(X, \mathcal{B})$, for $y \in Y$, such that*

(i) For each $A \in \mathcal{B}$ the map $y \mapsto \mu_y(A)$ is Borel and for each $A \in \mathcal{B}$

$$\mu(A) = \int \mu_y(A)\,\mathrm{d}\nu(y) \,.$$

(ii) For $y, y' \in Y$ the measures μ_y and $\mu_{y'}$ are mutually singular.

(iii) For each $y \in Y$ the transformation T is nonsingular and conservative, ergodic on $(X, \mathcal{B}, \mu_y)$.

(iv) For each $y \in Y$

$$\frac{\mathrm{d}\mu \circ T}{\mathrm{d}\mu} = \frac{\mathrm{d}\mu_y \circ T}{\mathrm{d}\mu_y} \quad \mu_y\text{-a. e} \,.$$

(v) (Uniqueness) If there exists another probability space $(Y', \nu', \mathcal{A}')$ and a family of probability measures $\mu'_{y'}$ on $(X, \mathcal{B})$, for $y' \in Y'$, satisfying (i)–(iv), then there exists a measure-preserving isomorphism $\theta\colon Y \to Y'$ such that $\mu_y = \mu'_{\theta y}$ for ν-a. e. y.

It follows that if T preserves an equivalent σ-finite measure then the system $(X, \mathcal{B}, \mu_y, T)$ is of type II for a. a. y. The space $(Y, \nu, \mathcal{A})$ is called *the space of T-ergodic components*.

Generators

It was shown in [157,162] that a nonsingular transformation T on a standard probability space $(X, \mathcal{B}, \mu)$ has a *countable generator*, i. e. a countable partition $\mathcal{P}$ so that $\bigvee_{n=-\infty}^{\infty} T^n \mathcal{P}$ generates the measurable sets. It was refined by Krengel [126]: if T is of

type II_∞ or III then there exists a generator P consisting of two sets only. Moreover, given a sub-σ-algebra $\mathcal{F} \subset \mathcal{B}$ such that $\mathcal{F} \subset T\mathcal{F}$ and $\bigcup_{k>0} T^k\mathcal{F} = \mathcal{B}$, the set $\{A \in \mathcal{F} \mid (A, X \setminus A)$ is a generator of $T\}$ is dense in $\mathcal{F}$. It follows, in particular, that T is isomorphic to the shift on $\{0,1\}^{\mathbb{Z}}$ equipped with a quasi-invariant probability measure.

The Glimm–Effros Theorem

The classical Bogoliouboff–Krylov theorem states that each homeomorphism of a compact space admits an ergodic invariant probability measure [33]. The following statement by Glimm [76] and Effros [61] is a "nonsingular" analogue of that theorem. (We consider here only a particular case of $\mathbb{Z}$-actions.)

Theorem 11 *Let X be a Polish space and $T\colon X \to X$ an aperiodic homeomorphism. Then the following are equivalent:*

(i) T has a recurrent point x, i. e. $x = \lim_{n\to\infty} T^{n_i}x$ for a sequence $n_1 < n_2 < \cdots$.
(ii) There is an orbit of T which is not locally closed.
(iii) There is no a Borel set which intersects each orbit of T exactly once.
(iv) There is a continuous probability Borel measure μ on X such that (X, μ, T) is an ergodic nonsingular system.

A natural question arises: under the conditions of the theorem how many such μ can exists? It turns out that there is a wealth of such measures. To state a corresponding result we first write an important definition.

Definition 12 Two nonsingular systems $(X, \mathcal{B}, \mu, T)$ and $(X, \mathcal{B}', \mu', T')$ are called *orbit equivalent* if there is a one-to-one bi-measurable map $\varphi\colon X \to X$ with $\mu' \circ \varphi \sim \mu$ and such that φ maps the T-orbit of x onto the T'-orbit of $\varphi(x)$ for a. a. $x \in X$.

The following theorem was proved in [116,128] and [174].

Theorem 13 *Let (X, T) be as in Theorem 11. Then for each ergodic dynamical system (Y, C, ν, S) of type II_∞ or III, there exist uncountably many mutually disjoint Borel measures μ on X such that $(X, T, \mathcal{B}, \mu)$ is orbit equivalent to (Y, C, ν, S).*

On the other hand, T may not have any finite invariant measure. Indeed, let T be an irrational rotation on the circle $\mathbb{T}$ and X a non-empty T-invariant G_δ subset of $\mathbb{T}$ of full Lebesgue measure. Let (X, T) contain a recurrent point. Then the unique ergodicity of $(\mathbb{T}, T)$ implies that (X, T) has no finite invariant measures.

Let T be an aperiodic Borel transformation of a standard Borel space X. Denote by $\mathcal{M}(T)$ the set of all ergodic T-nonsingular continuous measures on X. Given $\mu \in \mathcal{M}(T)$, let $N(\mu)$ denote the family of all Borel μ-null subsets. Shelah and Weiss showed [178] that $\bigcap_{\mu\in\mathcal{M}(T)} N(\mu)$ coincides with the collection of all Borel T-wandering sets.

Special Representations of Ergodic Flows

Nonsingular flows ($= \mathbb{R}$-actions) appear naturally in the study of orbit equivalence for systems of type III (see Sect. "Orbit Theory"). Here we record some basic notions related to nonsingular flows. Let $(X, \mathcal{B}, \mu)$ be a standard Borel space with a σ-finite measure μ on $\mathcal{B}$. A nonsingular *flow* on (X, μ) is a Borel map $S\colon X \times \mathbb{R} \ni (x, t) \mapsto S_t x \in X$ such that $S_t S_s = S_{t+s}$ for all $s, t \in \mathbb{R}$ and each S_t is a nonsingular transformation of (X, μ). Conservativeness and ergodicity for flows are defined in a similar way as for transformations.

A very useful example of a flow is a flow built under a function. Let $(X, \mathcal{B}, \mu, T)$ be a nonsingular dynamical system and f a positive Borel function on X such that $\sum_{i=0}^{\infty} f(T^i x) = \sum_{i=0}^{\infty} f(T^{-i}x) = \infty$ for all $x \in X$. Set $X^f := \{(x, s)\colon x \in X, 0 \le s < f(x)\}$. Define μ^f to be the restriction of the product measure $\mu \times \mathrm{Leb}$ on $X \times \mathbb{R}$ to X^f and define, for $t \ge 0$,

$$S_t^f(x, s) := \left(T^n x, s + t - \sum_{i=0}^{n-1} f(T^i x)\right),$$

where n is the unique integer that satisfies

$$\sum_{i=0}^{n-1} f(T^i x) < s + t \le \sum_{i=0}^{n} f(T^i x).$$

A similar definition applies when $t < 0$. In particular, when $0 < s + t < \varphi(x)$, $S_t^f(x, s) = (x, s + t)$, so that the flow moves the point (x, s) up t units, and when it reaches $(x, \varphi(x))$ it is sent to $(Tx, 0)$. It can be shown that $S^f = (S_t^f)_{t\in\mathbb{R}}$ is a free μ^f-nonsingular flow and that it preserves μ^f if and only if T preserves μ [148]. It is called the *flow built under the function φ with the base transformation T*. Of course, S^f is conservative or ergodic if and only if so is T.

Two flows $S = (S_t)_{t\in\mathbb{R}}$ on $(X, \mathcal{B}, \mu)$ and $V = (V_t)_{t\in\mathbb{R}}$ on (Y, C, ν) are said to be *isomorphic* if there exist invariant co-null sets $X' \subset X$ and $Y' \subset Y$ and an invertible nonsingular map $\rho\colon X' \to Y'$ that interwines the actions of the flows: $\rho \circ S_t = V_t \circ \rho$ on X' for all t. The following nonsingular version of Ambrose–Kakutani representation theorem was proved by Krengel [120] and Kubo [130].

Theorem 14 *Let S be a free nonsingular flow. Then it is isomorphic to a flow built under a function.*

Rudolph showed that in the Ambrose–Kakutani theorem one can choose the function φ to take two values. Krengel [122] showed that this can also be assumed in the nonsingular case.

Panorama of Examples

This section is devoted entirely to examples of nonsingular systems. We describe here the most popular (and simple) constructions of nonsingular systems: odometers, nonsingular Markov odometers, tower transformations, rank-one and finite rank systems and nonsingular Bernoulli shifts.

Nonsingular Odometers

Given a sequence m_n of natural numbers, we let $A_n := \{0, 1, \dots, m_n - 1\}$. Let ν_n be a probability on A_n and $\nu_n(a) > 0$ for all $a \in A_n$. Consider now the infinite product probability space $(X, \mu) := \bigotimes_{n=1}^{\infty}(A_n, \nu_n)$. Assume that $\prod_{n=1}^{\infty} \max\{\nu_n(a) \mid a \in A_n\} = 0$. Then μ is non-atomic. Given $a_1 \in A_1, \dots, a_n \in A_n$, we denote by $[a_1, \dots, a_n]$ the cylinder $x = (x_i)_{i>0} \mid x_1 = a_1, \dots, x_n = a_n$. If $x \neq (0, 0, \dots)$, we let $l(x)$ be the smallest number l such that the lth coordinate of x is not $m_l - 1$. We define a Borel map $T\colon X \to X$ by (1) if $x \neq (m_1, m_2, \dots)$ and put $Tx := (0, 0, \dots)$ if $x = (m_1, m_2, \dots)$. Of course, T is isomorphic to a rotation on a compact monothetic totally disconnected Abelian group. It is easy to check that T is μ-nonsingular and

$$\frac{\mathrm{d}\mu \circ T}{\mathrm{d}\mu}(x) = \prod_{n=1}^{\infty} \frac{\nu_n((Tx)_n)}{\nu_n(x_n)} = \frac{\nu_{l(x)}(x_{l(x)} + 1)}{\nu_{l(x)}(x_{l(x)})} \prod_{n=1}^{l(x)-1} \frac{\nu_n(0)}{\nu_n(m_n - 1)}$$

for a. a. $x = (x_n)_{n>0} \in X$. It is also easy to verify that T is ergodic. It is called the *nonsingular odometer* associated to $(m_n, \nu_n)_{n=1}^{\infty}$. We note that Ornstein's transformation (Example 5) is a nonsingular odometer.

Markov Odometers

We define Markov odometers as in [54]. An ordered Bratteli diagram B [102] consists of

(i) a vertex set V which is a disjoint union of finite sets $V^{(n)}$, $n \geq 0$, V_0 is a singleton;
(ii) an edge set E which is a disjoint union of finite sets $E^{(n)}$, $n > 0$;
(iii) source mappings $s_n\colon E^{(n)} \to V^{(n-1)}$ and range mappings $r_n\colon E^{(n)} \to V^{(n)}$ such that $s_n^{-1}(v) \neq \emptyset$ for all $v \in V^{(n-1)}$ and $r_n^{-1}(v) \neq \emptyset$ for all $v \in V^{(n)}$, $n > 0$;
(iv) a partial order on E so that $e, e' \in E$ are comparable if and only if $e, e' \in E^{(n)}$ for some n and $r_n(e) = r_n(e')$.

A *Bratteli compactum* X_B of the diagram B is the space of infinite paths

$$\{x = (x_n)_{n>0} \mid x_n \in E^{(n)} \text{ and } r(x_n) = s(x_{n+1})\}$$

on B. X_B is equipped with the natural topology induced by the product topology on $\prod_{n>0} E^{(n)}$. We will assume always that the diagram is *essentially simple*, i. e. there is only one infinite path $x_{\max} = (x_n)_{n>0}$ with x_n maximal for all n and only one $x_{\min} = (x_n)_{n>0}$ with x_n minimal for all n. The *Bratteli–Vershik* map $T_B\colon X_B \to X_B$ is defined as follows: $Tx_{\max} = x_{\min}$. If $x = (x_n)_{n>0} \neq x_{\max}$ then let k be the smallest number such that x_k is not maximal. Let y_k be a successor of x_k. Let $(y_1, \dots, y_k)$ be the unique path such that $y_1, \dots, y_{k-1}$ are all minimal. Then we let $T_B x := (y_1, \dots, y_k, x_{k+1}, x_{k+2}, \dots)$. It is easy to see that T_B is a homeomorphism of X_B. Suppose that we are given a sequence $P^{(n)} = (P^{(n)}_{(v,e) \in V^{n-1} \times E^{(n)}})$ of stochastic matrices, i. e.

(i) $P^{(n)}_{v,e} > 0$ if and only if $v = s_n(e)$ and
(ii) $\sum_{\{e \in E^{(n)} \mid s_n(e) = v\}} P^{(n)}_{v,e} = 1$ for each $v \in V^{(n-1)}$.

For $e_1 \in E^{(1)}, \dots, e_n \in E^{(n)}$, let $[e_1, \dots, e_n]$ denote the cylinder $\{x = (x_j)_{j>0} \mid x_1 = e_1, \dots, x_n = e_n\}$. Then we define a *Markov measure* on X_B by setting

$$\mu_P([e_1, \dots, e_n]) = P^1_{s_1(e_1), e_1} P^2_{s_2(e_2), e_2} \cdots P^n_{s_n(e_n), e_n}$$

for each cylinder $[e_1, \dots, e_n]$. The dynamical system (X_B, μ_P, T_B) is called a *Markov odometer*. It is easy to see that every nonsingular odometer is a Markov odometer where the corresponding $V^{(n)}$ are all singletons.

Tower Transformations

This construction is a discrete analogue of flow under a function. Given a nonsingular dynamical system (X, μ, T) and a measurable map $f\colon X \to \mathbb{N}$, we define a new dynamical system (X^f, μ^f, T^f) by setting

$$X^f := \{(x, i) \in X \times \mathbb{Z}_+ \mid 0 \leq i < f(x)\},$$
$$\mathrm{d}\mu^f(x, i) := \mathrm{d}\mu(x) \quad \text{and}$$
$$T^f(x, i) := \begin{cases} (x, i+1), & \text{if } i + 1 < f(x) \\ (Tx, 0), & \text{otherwise}. \end{cases}$$

Then T^f is μ^f-nonsingular and $(\mathrm{d}\mu^f \circ T^f/\mathrm{d}\mu^f)(x,i) = (\mathrm{d}\mu \circ T/\mathrm{d}\mu)(x)$ for a. a. $(x,i) \in X^f$. This transformation is called the (Kakutani) *tower over T with height function f*. It is easy to check that T^f is conservative if and only if T is conservative; T^f is ergodic if and only if T is ergodic; T^f is of type III if and only if T is of type III. Moreover, the induced transformation $(T^f)_{X\times\{0\}}$ is isomorphic to T. Given a subset $A \subset X$ of positive measure, T is the tower over the induced transformation T_A with the first return time to A as the height function.

Rank-One Transformations. Chacón Maps. Finite Rank

The definition uses the process of "cutting and stacking". We construct by induction a sequence of columns C_n. A *column* C_n consists of a finite sequence of bounded intervals (left-closed, right-open) $C_n = \{I_{n,0}, \dots, I_{n,h_n-1}\}$ of *height* h_n. A column C_n determines a *column map* T_{C_n} that sends each interval $I_{n,i}$ to the interval above it $I_{n,i+1}$ by the unique orientation-preserving affine map between the intervals. T_{C_n} remains undefined on the top interval I_{n,h_n-1}. Set $C_0 = \{[0,1)\}$ and let $\{r_n > 2\}$ be a sequence of positive integers, let $\{s_n\}$ be a sequence of functions $s_n\colon \{0,\dots,r_n-1\} \to \mathbb{N}_0$, and let $\{w_n\}$ be a sequence of probability vectors on $\{0,\dots,r_n-1\}$. If C_n has been defined, column C_{n+1} is defined as follows. First "cut" (i. e., subdivide) each interval $I_{n,i}$ in C_n into r_n subintervals $I_{n,i}[j]$, $j = 0,\dots,r_n-1$, whose lengths are in the proportions $w_n(0)\colon w_n(1)\colon \cdots \colon w_n(r_n-1)$. Next place, for each $j = 0,\dots,r_n-1$, $s_n(j)$ new subintervals above $I_{n,h_n-1}[j]$, all of the same length as $I_{n,h_n-1}[j]$. Denote these intervals, called *spacers*, by $S_{n,0}[j],\dots,S_{n,s_n(j)-1}[j]$. This yields, for each $j \in \{0,\dots,r_n-1\}$, r_n subcolumns each consisting of the subintervals

$$I_{n,0}[j],\dots,I_{n,h_n-1}[j]$$
$$\text{followed by the spacers } S_{n,0}[j],\dots,S_{n,s_n(j)-1}[j]\,.$$

Finally each subcolumn is stacked from left to right so that the top subinterval in subcolumn j is sent to the bottom subinterval in subcolumn $j+1$, for $j = 0,\dots,r_n-2$ (by the unique orientation-preserving affine map between the intervals). For example, $S_{n,s_n(0)-1}[0]$ is sent to $I_{n,0}[1]$. This defines a new column C_{n+1} and new column map $T_{C_{n+1}}$, which remains undefined on its top subinterval. Let X be the union of all intervals in all columns and let μ be Lebesgue measure restricted to X. We assume that as $n \to \infty$ the maximal length of the intervals in C_n converges to 0, so we may define a transformation T of (X,μ) by $Tx := \lim_{n\to\infty} T_{C_n}x$. One can verify that T is well-defined a. e. and that it is nonsingular and ergodic. T is said to be the *rank-one* transformation associated with $(r_n, w_n, s_n)_{n=1}^{\infty}$. If all the probability vectors w_n are uniform the resulting transformation is measure-preserving. The measure is infinite (σ-finite) if and only if the total mass of the spacers is infinite. In the case $r_n = 3$ and $s_n(0) = s_n(2) = 0$, $s_n(1) = 1$ for all $n \geq 0$, the associated rank-one transformation is called a *nonsingular Chacón map*.

It is easy to see that every nonsingular odometer is of rank-one (the corresponding maps s_n are all trivial). Each rank-one map T is a tower over a nonsingular odometer (to obtain such an odometer reduce T to a column C_n).

A rank N transformation is defined in a similar way. A nonsingular transformation T is said to be of *rank N or less* if at each stage of its construction there exits N disjoint columns, the levels of the columns generate the σ-algebra and the Radon–Nikodym derivative of T is constant on each non-top level of every column. T is said to be of *rank N* if it is of rank N or less and not of rank $N-1$ or less. A rank N transformation, $N \geq 2$, need not be ergodic.

Nonsingular Bernoulli Transformations – Hamachi's Example

A *nonsingular Bernoulli* transformation is a transformation T such that there exists a countable generator $\mathcal{P}$ (see Subsect. "Generators") such that the partitions $T^n\mathcal{P}$, $n \in \mathbb{Z}$, are mutually independent and such that the Radon–Nikodym derivative ω_1 is measurable with respect to the sub-σ-algebra $\bigvee_{n=-\infty}^{0} T^n\mathcal{P}$.

In [87], Hamachi constructed examples of conservative nonsingular Bernoulli transformations, hence ergodic (see Subsect. "Weak Mixing, Mixing, K-Property"), with a 2-set generating partition that are of type III. Krengel [121] asked if there are of type II_∞ examples of nonsingular Bernoulli automorphisms and the question remains open. Hamachi's construction is the left-shift on the space $X = \prod_{n=-\infty}^{\infty}\{0,1\}$. The measure is a product $\mu = \prod_{n=-\infty}^{\infty}\mu_n$ where $\mu_n = (1/2, 1/2)$ for $n \geq 0$ and for $n < 0$ μ_n is chosen carefully alternating on large blocks between the uniform measure and different non-uniform measures. Kakutani's criterion for equivalence of infinite product measures is used to verify that μ is nonsingular.

Mixing Notions and multiple recurrence

The study of mixing and multiple recurrence are central topics in classical ergodic theory [33,70]. Unfortunately, these notions are considerably less 'smooth' in the world of nonsingular systems. The very concepts of any kind of mixing and multiple recurrence are not well understood in view of their ambiguity. Below we discuss nonsingular

systems possessing a surprising diversity of such properties that seem equivalent but are different indeed.

Weak Mixing, Mixing, K-Property

Let T be an ergodic conservative nonsingular transformation. A number $\lambda \in \mathbb{C}$ is an L^∞*-eigenvalue* for T if there exists a nonzero $f \in L^\infty$ so that $f \circ T = \lambda f$ a. e. It follows that $|\lambda| = 1$ and f has constant modulus, which we assume to be 1. Denote by $e(T)$ the set of all L^∞-eigenvalues of T. T is said to be *weakly mixing* if $e(T) = \{1\}$. We refer to Theorem 2.7.1 in [3] for proof of the following Keane's ergodic multiplier theorem: given an ergodic probability preserving transformation S, the product transformation $T \times S$ is ergodic if and only if $\sigma_S(e(T)) = 0$, where σ_S denotes the measure of (reduced) maximal spectral type of the unitary U_S (see (3)). It follows that T is weakly mixing if and only $T \times S$ is ergodic for every ergodic probability preserving S. While in the finite measure-preserving case this implies that $T \times T$ is ergodic, it was shown in [5] that there exits a weakly mixing nonsingular T with $T \times T$ not conservative, hence not ergodic. In [11], a weakly mixing T was constructed with $T \times T$ conservative but not ergodic. A nonsingular transformation T is said to be *doubly ergodic* if for all sets of positive measure A and B there exists an integer $n > 0$ such that $\mu(A \cap T^{-n}A) > 0$ and $\mu(A \cap T^{-n}B) > 0$. Furstenberg [70] showed that for finite measure-preserving transformations double ergodicity is equivalent to weak mixing. In [20] it is shown that for nonsingular transformations weak mixing does not imply double ergodicity and double ergodicity does not imply that $T \times T$ is ergodic.

T is said to have *ergodic index* k if the Cartesian product of k copies of T is ergodic but the product of $k+1$ copies of T is not ergodic. If all finite Cartesian products of T are ergodic then T is said to have *infinite ergodic index*. Parry and Kakutani [113] constructed for each $k \in \mathbb{N} \cup \{\infty\}$, an infinite Markov shift of ergodic index k. A stronger property is *power weak mixing*, which requires that for all nonzero integers $k_1, \ldots, k_r$ the product $T^{k_1} \times \cdots \times T^{k_r}$ is ergodic [47]. The following examples were constructed in [12,36,38]:

(i) power weakly mixing rank-one transformations,
(ii) non-power weakly mixing rank-one transformations with infinite ergodic index,
(iii) non-power weakly mixing rank-one transformations with infinite ergodic index and such that $T^{k_1} \times \cdots \times T^{k_r}$ are all conservative, $k_1, \ldots, k_r \in \mathbb{Z}$,

of types II_∞ and III (and various subtypes of III, see Sect. "Orbit Theory"). Thus we have the following scale of properties (equivalent to weak mixing in the probability preserving case), where every next property is strictly stronger than the previous ones:

$$\begin{aligned} T \text{ is weakly mixing } &\Leftarrow T \text{ is doubly ergodic} \\ &\Leftarrow T \times T \text{ is ergodic} \\ &\Leftarrow T \times T \times T \text{ is ergodic} \\ &\Leftarrow \cdots \\ &\Leftarrow T \text{ has infinite ergodic index} \\ &\Leftarrow T \text{ is power weakly mixing .} \end{aligned}$$

We also mention a recent example of a power weakly mixing transformation of type II_∞ which embeds into a flow [46].

We now consider several attempts to generalize the notion of (strong) mixing. Given a sequence of measurable sets $\{A_n\}$ let $\sigma_k(\{A_n\})$ denote the σ-algebra generated by $A_k, A_{k+1}, \ldots$. A sequence $\{A_n\}$ is said to be *remotely trivial* if $\bigcap_{k=0}^\infty \sigma_k(\{A_n\}) = \{\emptyset, X\} \bmod \mu$, and it is *semi-remotely trivial* if every subsequence contains a subsequence that is remotely trivial. Krengel and Sucheston [124] define a nonsingular transformation T of a σ-finite measure space to be *mixing* if for every set A of finite measure the sequence $\{T^{-n}A\}$ is semi-remotely trivial, and *completely mixing* if $\{T^{-n}A\}$ is semi-remotely trivial for all measurable sets A. They show that T is completely mixing if and only if it is type II_1 and mixing for the equivalent finite invariant measure. Thus there are no type III and II_∞ completely mixing nonsingular transformations on probability spaces. We note that this definition of mixing in infinite measure spaces depends on the choice of measure inside the equivalence class (but it is independent if we replace the measure by an equivalent measure with the same collection of sets of finite measure).

Hajian and Kakutani showed [83] that an ergodic infinite measure-preserving transformation T is either of *zero type*: $\lim_{n\to\infty}\mu(T^{-n}A \cap A) = 0$ for all sets A of finite measure, or of *positive type*: $\limsup_{n\to\infty}\mu(T^{-n}A \cap A) > 0$ for all sets A of finite positive measure. T is mixing if and only if it is of zero type [124]. For $0 \le \alpha \le 1$ Kakutani suggested a related definition of α-type: an infinite measure preserving transformation is *of α-type* if $\limsup_{n\to\infty} \mu(A \cap T^n A) = \alpha\mu(A)$ for every subset A of finite measure. In [153] examples of ergodic transformations of any α-type and a transformation of not any type were constructed.

It may seem that mixing is stronger than any kind of nonsingular weak mixing considered above. However, it is not the case: if T is a weakly mixing infinite measure preserving transformation of zero type and S is an ergodic probability preserving transformation then $T \times S$ is er-

godic and of zero type. On the other hand, the L^∞-spectrum $e(T \times S)$ is nontrivial, i. e. $T \times S$ is not weakly mixing, whenever S is not weakly mixing. We also note that there exist rank-one infinite measure-preserving transformations T of zero type such that $T \times T$ is not conservative (hence not ergodic) [11]. In contrast to that, if T is of positive type all of its finite Cartesian products are conservative [7]. Another result that suggests that there is no good definition of mixing in the nonsingular case was proved recently in [110]. It is shown there that while the mixing finite measure-preserving transformations are measurably sensitive, there exists no infinite measure-preserving system that is measurably sensitive. (Measurable sensitivity is a measurable version of the strong sensitive dependence on initial conditions – a concept from topological theory of chaos.)

A nonsingular transformation T of $(X, \mathcal{B}, \mu)$ is called *K-automorphism* [180] if there exists a sub-σ-algebra $\mathcal{F} \subset \mathcal{B}$ such that $T^{-1}\mathcal{F} \subset \mathcal{F}$, $\bigcap_{k\geq 0} T^{-k}\mathcal{F} = \{\emptyset, X\}$, $\bigvee_{k=0}^{+\infty} T^k\mathcal{F} = \mathcal{B}$ and the Radon–Nikodym derivative $\mathrm{d}\mu \circ T/\mathrm{d}\mu$ is $\mathcal{F}$-measurable (see also [156] for the case when T is of type II_∞; the authors in [180] required T to be conservative). Evidently, a nonsingular Bernoulli transformation (see Subsect. "Nonsingular Bernoulli Transformations – Hamachi's Example") is a K-automorphism. Parry [156] showed that a type II_∞ K-automorphism is either dissipative or ergodic. Krengel [121] proved the same for a class of Bernoulli nonsingular transformations, and finally Silva and Thieullen extended this result to nonsingular K-automorphisms [180]. It is also shown in [180] that if T is a nonsingular K-automorphism, for any ergodic nonsingular transformation S, if $S \times T$ is conservative, then it is ergodic. It follows that a conservative nonsingular K-automorphism is weakly mixing. However, it does not necessarily have infinite ergodic index [113]. Krengel and Sucheston [124] showed that an infinite measure-preserving conservative K-automorphism is mixing.

Multiple and Polynomial Recurrence

Let p be a positive integer. A nonsingular transformation T is called *p-recurrent* if for every subset B of positive measure there exists a positive integer k such that

$$\mu(B \cap T^{-k}B \cap \cdots \cap T^{-kp}B) > 0\,.$$

If T is p-recurrent for any $p > 0$, then it is called *multiply recurrent*. It is easy to see that T is 1-recurrent if and only if it is conservative. T is called *rigid* if $T^{n_k} \to \mathrm{Id}$ for a sequence $n_k \to \infty$. Clearly, if T is rigid then it is multiply recurrent. Furstenberg showed [70] that every finite measure-preserving transformation is multiply recurrent. In contrast to that Eigen, Hajian and Halverson [64] constructed for any $p \in \mathbb{N} \cup \{\infty\}$, a nonsingular odometer of type II_∞ which is p-recurrent but not $(p+1)$-recurrent. Aaronson and Nakada showed in [7] that an infinite measure preserving Markov shift T is p-recurrent if and only if the product $T \times \cdots \times T$ (p times) is conservative. It follows from this and [5] that in the class of ergodic Markov shifts infinite ergodic index implies multiple recurrence. However, in general this is not true. It was shown in [12,45] and [82] that for each $p \in \mathbb{N} \cup \{\infty\}$ there exist

(i) power weakly mixing rank-one transformations and
(ii) non-power weakly mixing rank-one transformations with infinite ergodic index

which are p-recurrent but not $(p+1)$-recurrent (the latter holds when $p \neq \infty$, of course).

A subset A is called *p-wandering* if $\mu(A \cap T^kA \cap \cdots \cap T^{pk}A) = 0$ for each k. Aaronson and Nakada established in [7] a p-analogue of Hopf decomposition (see Theorem 2).

Proposition 15 *If $(X, \mathcal{B}, \mu, T)$ is conservative aperiodic nonsingular dynamical system and $p \in \mathbb{N}$ then $X = C_p \cup D_p$, where C_p and D_p are T-invariant disjoint subsets, D_p is a countable union of p-wandering sets, $T \restriction C_p$ is p-recurrent and $\sum_{k=1}^{\infty} \mu(B \cap T^{-k}B \cap \cdots \cap T^{-dk}B) = \infty$ for every $B \subset C_p$.*

Let T be an infinite measure-preserving transformation and let $\mathcal{F}$ be a σ-finite factor (i. e., invariant subalgebra) of T. Inoue [106] showed that for each $p > 0$, if $T \restriction \mathcal{F}$ is p-recurrent then so is T provided that the extension $T \to T \restriction \mathcal{F}$ is isometric. It is unknown yet whether the latter assumption can be dropped. However, partial progress was recently achieved in [140]: if $T \restriction \mathcal{F}$ is multiply recurrent then so is T.

Let $\mathcal{P} := \{q \in \mathbb{Q}[t] \mid q(\mathbb{Z}) \subset \mathbb{Z} \text{ and } q(0) = 0\}$. An ergodic conservative nonsingular transformation T is called *p-polynomially recurrent* if for every $q_1, \ldots, q_p \in \mathcal{P}$ and every subset B of positive measure there exists $k \in \mathbb{N}$ with

$$\mu(B \cap T^{q_1(k)}B \cap \cdots \cap T^{q_p(k)}B) > 0\,.$$

If T is p-polynomially recurrent for every $p \in \mathbb{N}$ then it is called *polynomially recurrent*. Furstenberg's theorem on multiple recurrence was significantly strengthened in [17], where it was shown that every finite measure-preserving transformation is polynomially recurrent. However, Danilenko and Silva [45] constructed

(i) nonsingular transformations T which are p-polynomially recurrent but not $(p+1)$-polynomially recurrent (for each fixed $p \in \mathbb{N}$),

(ii) polynomially recurrent transformations T of type II_∞,
(iii) rigid (and hence multiply recurrent) transformations T which are not polynomially recurrent.

Moreover, such T can be chosen inside the class of rank-one transformations with infinite ergodic index.

Topological Group Aut(X, μ)

Let $(X, \mathcal{B}, \mu)$ be a standard probability space and let $\mathrm{Aut}(X, \mu)$ denote the group of all nonsingular transformations of X. Let ν be a finite or σ-finite measure equivalent to μ; the subgroup of the ν-preserving transformations is denoted by $\mathrm{Aut}_0(X, \nu)$. Then $\mathrm{Aut}(X, \mu)$ is a simple group [62] and it has no outer automorphisms [63]. Ryzhikov showed [169] that every element of this group is a product of three involutions (i. e. transformations of order 2). Moreover, a nonsingular transformation is a product of two involutions if and only if it is conjugate to its inverse by an involution.

Inspired by [85], Ionescu Tulcea [107] and Chacon and Friedman [21] introduced the *weak* and the *uniform* topologies respectively on $\mathrm{Aut}(X, \mu)$. The weak one – we denote it by d_w – is induced from the weak operator topology on the group of unitary operators in $L^2(X, \mu)$ by the embedding $T \mapsto U_T$ (see Subsect. "Mean and Pointwise Ergodic Theorems. Rokhlin Lemma"). Then $(\mathrm{Aut}(X, \mu), d_w)$ is a Polish topological group and $\mathrm{Aut}_0(X, \nu)$ is a closed subgroup of $\mathrm{Aut}(X, \mu)$. This topology will not be affected if we replace μ with any equivalent measure. We note that T_n weakly converges to T if and only if $\mu(T_n^{-1}A \triangle T^{-1}A) \to 0$ for each $A \in \mathcal{B}$ and $\mathrm{d}(\mu \circ T_n)/\mathrm{d}\mu \to \mathrm{d}(\mu \circ T)/\mathrm{d}\mu$ in $L^1(X, \mu)$. Danilenko showed in [34] that $(\mathrm{Aut}(X, \mu), d_w)$ is contractible. It follows easily from the Rokhlin lemma that periodic transformations are dense in $\mathrm{Aut}(X, \mu)$.

For each $p \geq 1$, one can also embed $\mathrm{Aut}(X, \mu)$ into the isometry group of $L^p(X, \mu)$ via a formula similar to (3) but with another power of the Radon–Nikodym derivative in it. The strong operator topology on the isometry group induces the very same weak topology on $\mathrm{Aut}(X, \mu)$ for all $p \geq 1$ [24].

It is natural to ask which properties of nonsingular transformations are typical in the sense of Baire category. The following technical lemma (see see [24,68]) is an indispensable tool when considering such problems.

Lemma 16 *The conjugacy class of each aperiodic transformation T is dense in* $\mathrm{Aut}(X, \mu)$ *endowed with the weak topology.*

Using this lemma and the Hurewicz ergodic theorem Choksi and Kakutani [24] proved that the ergodic transformations form a dense G_δ in $\mathrm{Aut}(X, \mu)$. The same holds for the subgroup $\mathrm{Aut}_0(X, \nu)$ [24,170]. Combined with [107] the above implies that the ergodic transformations of type III is a dense G_δ in $\mathrm{Aut}(X, \mu)$. For further refinement of this statement we refer to Sect. "Orbit Theory".

Since the map $T \mapsto T \times \cdots \times T$ (p times) from $\mathrm{Aut}(X, \mu)$ to $\mathrm{Aut}(X^p, \mu^{\otimes p})$ is continuous for each $p > 0$, we deduce that the set $\mathcal{E}_\infty$ of transformations with infinite ergodic index is a G_δ in $\mathrm{Aut}(X, \mu)$. It is non-empty by [113]. Since this $\mathcal{E}_\infty$ is invariant under conjugacy, it is dense in $\mathrm{Aut}(X, \mu)$ by Lemma 16. Thus we obtain that $\mathcal{E}_\infty$ is a dense G_δ. In a similar way one can show that $\mathcal{E}_\infty \cap \mathrm{Aut}_0(X, \nu)$ is a dense G_δ in $\mathrm{Aut}_0(X, \nu)$ (see also [24,26,170] for original proofs of these claims).

The rigid transformations form a dense G_δ in $\mathrm{Aut}(X, \mu)$. It follows that the set of multiply recurrent nonsingular transformations is residual [13]. A finer result was established in [45]: the set of polynomially recurrent transformations is also residual.

Given $T \in \mathrm{Aut}(X, \mu)$, we denote the *centralizer* $\{S \in \mathrm{Aut}(X, \mu) \mid ST = TS\}$ of T by $C(T)$. Of course, $C(T)$ is a closed subgroup of $\mathrm{Aut}(X, \mu)$ and $C(T) \supset \{T^n \mid n \in \mathbb{Z}\}$. The following problems solved recently (by the efforts of many authors) for probability preserving systems are still open for the nonsingular case. Are the properties:

(i) T has square root;
(ii) T embeds into a flow;
(iii) T has non-trivial invariant sub-σ-algebra;
(iv) $C(T)$ contains a torus of arbitrary dimension

typical (residual) in $\mathrm{Aut}(X, \mu)$?

The *uniform* topology on $\mathrm{Aut}(X, \mu)$, finer than d_w, is defined by the metric

$$d_u(T, S) = \mu(\{x\colon Tx \neq Sx\}) + \mu(\{x\colon T^{-1}x \neq S^{-1}x\}).$$

This topology is also complete metric. It depends only on the measure class of μ. However the uniform topology is not separable and that is why it is of less importance in ergodic theory. We refer to [21,24,27] and [68] for the properties of d_u.

Orbit Theory

Orbit theory is, in a sense, the most complete part of nonsingular ergodic theory. We present here the seminal Krieger's theorem on orbit classification of ergodic nonsingular transformations in terms of ratio sets and associated flows. Examples of transformations of various types

III_λ, $0 \le \lambda \le 1$ are also given here. Next, we consider the outer conjugacy problem for automorphisms of the orbit equivalence relations. This problem is solved in terms of a simple complete system of invariants. We discuss also a general theory of cocycles (of nonsingular systems) taking values in locally compact Polish groups and present an important orbit classification theorem for cocycles. This theorem is an analogue of the aforementioned result of Krieger. We complete the section by considering ITPFI-systems and their relation to AT-flows.

Full Groups. Ratio Set and Types III_λ, $0 \le \lambda \le 1$

Let T be a nonsingular transformation of a standard probability space $(X, \mathcal{B}, \mu)$. Denote by $\mathrm{Orb}_T(x)$ the T-orbit of x, i. e. $\mathrm{Orb}_T(x) = \{T^n x \mid n \in \mathbb{Z}\}$. The *full group* $[T]$ of T consists of all transformations $S \in \mathrm{Aut}(X, \mu)$ such that $Sx \in \mathrm{Orb}_T(x)$ for a. a. x. If T is ergodic then $[T]$ is topologically simple (or even algebraically simple if T is not of type II_∞) [62]. It is easy to see that $[T]$ endowed with the uniform topology d_u is a Polish group. If T is ergodic then $([T], d_u)$ is contractible [34].

The *ratio set* $r(T)$ of T was defined by Krieger [126] and as we shall see below it is the key concept in the orbit classification (see Definition 1). The ratio set is a subset of $[0, +\infty)$ defined as follows: $t \in r(T)$ if and only if for every $A \in \mathcal{B}$ of positive measure and each $\epsilon > 0$ there is a subset $B \subset A$ of positive measure and an integer $k \neq 0$ such that $T^k B \subset A$ and $|\omega_k^\mu(x) - t| < \epsilon$ for all $x \in B$. It is easy to verify that $r(T)$ depends only on the equivalence class of μ and not on μ itself. A basic fact is that $1 \in r(T)$ if and only if T is conservative. Assume now T to be conservative and ergodic. Then $r(T) \cap (0, +\infty)$ is a closed subgroup of the multiplicative group $(0, +\infty)$. Hence $r(T)$ is one of the following sets:

(i) $\{1\}$;
(ii) $\{0, 1\}$; in this case we say that T is of *type* III_0,
(iii) $\{\lambda^n \mid n \in \mathbb{Z}\} \cup \{0\}$ for $0 < \lambda < 1$; then we say that T is of *type* III_λ,
(iv) $[0, +\infty)$; then we say that T is of *type* III_1.

Krieger showed that $r(T) = \{1\}$ if and only if T is of type II. Hence we obtain a further subdivision of type III into subtypes III_0, III_λ, or III_1.

Example 17 (i) Fix $\lambda \in (0, 1)$. Let $\nu_n(0) := 1/(1 + \lambda)$ and $\nu_n(1) := \lambda/(1 + \lambda)$ for all $n = 1, 2, \ldots$. Let T be the nonsingular odometer associated with the sequence $(2, \nu_n)_{n=1}^\infty$ (see Subsect. "Nonsingular Odometers"). We claim that T is of type III_λ. Indeed, the group Σ of finite permutations of $\mathbb{N}$ acts on X by $(\sigma x)_n = x_{\sigma^{-1}(n)}$, for all $n \in \mathbb{N}$, $\sigma \in \Sigma$ and $x = (x_n)_{n=1}^\infty \in X$. This action preserves μ. Moreover, it is ergodic by the Hewitt–Savage 0–1 law. It remains to notice that $(\mathrm{d}\mu \circ T/\mathrm{d}\mu)(x) = \lambda$ on the cylinder $[0]$ which is of positive measure.
(ii) Fix positive reals ρ_1 and ρ_2 such that $\log \rho_1$ and $\log \rho_2$ are rationally independent. Let $\nu_n(0) := 1/(1 + \rho_1 + \rho_2)$, $\nu_n(1) := \rho_1/(1 + \rho_1 + \rho_2)$ and $\nu_n(2) := \rho_2/(1 + \rho_1 + \rho_2)$ for all $n = 1, 2, \ldots$. Then the nonsingular odometer associated with the sequence $(3, \nu_n)_{n=1}^\infty$ is of type III_1. This can be shown in a similar way as (i).

Non-singular odometer of type III_0 will be constructed in Example 19 below.

Maharam Extension, Associated Flow and Orbit Classification of Type III Systems

On $X \times \mathbb{R}$ with the σ-finite measure $\mu \times \kappa$, where $\mathrm{d}\kappa(y) = \exp(y)\mathrm{d}y$, consider the transformation

$$\widetilde{T}(x, y) := \left(Tx, y - \log \frac{\mathrm{d}\mu \circ T}{\mathrm{d}\mu}(x)\right) .$$

We call it the *Maharam extension* of T (see [136], where these transformations were introduced). It is measure-preserving and it commutes with the flow $S_t(x, y) := (x, y + t)$, $t \in \mathbb{R}$. It is conservative if and only if T is conservative [136]. However $\widetilde{T}$ is not necessarily ergodic. Let (Z, ν) denote the space of $\widetilde{T}$-ergodic components. Then $(S_t)_{t\in\mathbb{R}}$ acts nonsingularly on this space. The restriction of $(S_t)_{t\in\mathbb{R}}$ to (Z, ν) is called the *associated flow* of T. The associated flow is ergodic whenever T is ergodic. It is easy to verify that the isomorphism class of the associated flow is an invariant of the orbit equivalence of the underlying system.

Proposition 18 ([90])

(i) T is of type II if and only if its associated flow is the translation on $\mathbb{R}$, i. e. $x \mapsto x + t$, $x, t \in \mathbb{R}$,
(ii) T is of type III_λ, $0 \le \lambda < 1$ if and only if its associated flow is the periodic flow on the interval $[0, -\log \lambda)$, i. e. $x \mapsto x + t \mod (-\log \lambda)$,
(iii) T is of type III_1 if and only if its associated flow is the trivial flow on a singleton or, equivalently, $\widetilde{T}$ is ergodic,
(iv) T is of type III_0 if and only if its associated flow is non-transitive.

Example 19 Let $A_n = \{0, 1, \ldots, 2^{2^n}\}$ and $\nu_n(0) = 0.5$ and $\nu_n(i) = 0.5 \cdot 2^{-2^n}$ for all $0 < i \le 2^n$. Let T be the nonsingular odometer associated with $(2^{2^n} + 1, \nu_n)_{n=0}^\infty$. It is straightforward that the associated flow of T is the flow built under the constant function 1 with the probability preserving 2-adic odometer (associated with $(2, \kappa_n)_{n=1}^\infty$,

$\kappa_n(0) = \kappa_n(1) = 0.5$) as the base transformation. In particular, T is of type III_0.

A natural problem arises: to compute Krieger's type (or the ratio set) for the nonsingular odometers – the simplest class of nonsingular systems. Some partial progress was achieved in [56,141,152], etc. However in the general setting this problem remains open.

The map $\Psi\colon \mathrm{Aut}(X,\mu) \ni T \mapsto \widetilde{T} \in \mathrm{Aut}(X\times\mathbb{R},\mu\times\kappa)$ is a continuous group homomorphism. Since the set $\mathcal{E}$ of ergodic transformations on $X\times\mathbb{R}$ is a G_δ in $\mathrm{Aut}(X\times\mathbb{R},\mu\times\kappa)$ (see Sect. "Topological Group Aut(X, μ)"), the subset $\Psi^{-1}(\mathcal{E})$ of type III_1 ergodic transformations on X is also G_δ. The latter subset is nonempty in view of Example 17(ii). Since it is invariant under conjugacy, we deduce from Lemma 16 that the set of ergodic transformations of type III_1 is a dense G_δ in $(\mathrm{Aut}(X,\mu), d_w)$ [23,159].

Now we state the main result of this section – Krieger's theorem on orbit classification for ergodic transformations of type III. It is a far reaching generalization of the basic result by H. Dye: any two ergodic probability preserving transformations are orbit equivalent [60].

Theorem 20 (Orbit equivalence for type III systems [125]–[129]) *Two ergodic transformations of type III are orbit equivalent if and only if their associated flows are isomorphic. In particular, for a fixed* $0<\lambda\le 1$, *any two ergodic transformations of type* III_λ *are orbit equivalent.*

The original proof of this theorem is rather complicated. Simpler treatment of it can be found in [90] and [117].

We also note that every nontransitive ergodic flow can be realized as the associated flow of a type III_0 transformation. However it is somewhat easier to construct a $\mathbb{Z}^2$-action of type III_0 whose associated flow is the given one. For this, we take an ergodic nonsingular transformation Q on a probability space $(Z,\mathcal{B},\lambda)$ and a measure-preserving transformation R of an infinite σ-finite measure space $(Y,\mathcal{F},\nu)$ such that there is a continuous homomorphism $\pi:\mathbb{R}\to C(R)$ with $(\mathrm{d}\nu\circ\pi(t)/\mathrm{d}\nu)(y)=\exp(t)$ for a. a. y (for instance, take a type III_1 transformation T and put $R:=\widetilde{T}$ and $\pi(t):=S_t$). Let $\varphi\colon Z\to\mathbb{R}$ be a Borel map with $\inf_Z\varphi>0$. Define two transformations R_0 and Q_0 of $(Z\times Y,\lambda\times\nu)$ by setting:

$$R_0(x,y):=(x,Ry)\,,\quad Q_0(x,y)=(Qx,U_xy)\,,$$

where $U_x=\pi(\varphi(x)-\log(\mathrm{d}\mu\circ Q/\mathrm{d}\mu)(x))$. Notice that R_0 and Q_0 commute. The corresponding $\mathbb{Z}^2$-action generated by these transformations is ergodic. Take any transformation $V\in\mathrm{Aut}(Z\times Y,\lambda\times\nu)$ whose orbits coincide with the orbits of the $\mathbb{Z}^2$-action. (According to [29], any ergodic nonsingular action of any countable amenable group is orbit equivalent to a single transformation.) Then V is of type III_0. It is now easy to verify that the associated flow of V is the special flow built under $\varphi\circ Q^{-1}$ with the base transformation Q^{-1}. Since Q and φ are arbitrary, we deduce the following from Theorem 14.

Theorem 21 *Every nontransitive ergodic flow is an associated flow of an ergodic transformation of type* III_0.

In [129] Krieger introduced a map Φ as follows. Let T be an ergodic transformation of type III_0. Then the associated flow of T is a flow built under function with a base transformation $\Phi(T)$. We note that the orbit equivalence class of $\Phi(T)$ is well defined by the orbit equivalent class of T. If $\Phi^n(T)$ fails to be of type III_0 for some $1\le n<\infty$ then T is said to *belong to Krieger's hierarchy*. For instance, the transformation constructed in Example 19 belongs to Krieger's hierarchy. Connes gave in [28] an example of T such that $\Phi(T)$ is orbit equivalent to T (see also [73] and [90]). Hence T is not in Krieger's hierarchy.

Normalizer of the Full Group. Outer Conjugacy Problem

Let

$$N[T]=\{R\in\mathrm{Aut}(X,\mu)\mid R[T]R^{-1}=[T]\}\,,$$

i. e. $N[T]$ is the *normalizer* of the full group $[T]$ in $\mathrm{Aut}(X,\mu)$. We note that a transformation R belongs to $N[T]$ if and only if $R(\mathrm{Orb}_T(x))=\mathrm{Orb}_T(Rx)$ for a. a. x. To define a topology on $N[T]$ consider the T-orbit equivalence relation $\mathcal{R}_T\subset X\times X$ and a σ-finite measure $\mu_{\mathcal{R}}$ on $\mathcal{R}_T$ given by $\mu_{\mathcal{R}_T}=\int_X\sum_{y\in\mathrm{Orb}_T(x)}\delta_{(x,y)}\mathrm{d}\mu(x)$. For $R\in N[T]$, we define a transformation $i(R)\in\mathrm{Aut}(\mathcal{R}_T,\mu_{\mathcal{R}_T})$ by setting $i(R)(x,y):=(Rx,Ry)$. Then the map $R\mapsto i(R)$ is an embedding of $N[T]$ into $\mathrm{Aut}(\mathcal{R}_T,\mu_{\mathcal{R}_T})$. Denote by τ the topology on $N[T]$ induced by the weak topology on $\mathrm{Aut}(\mathcal{R}_T,\mu_{\mathcal{R}_T})$ via i [34]. Then $(N[T],\tau)$ is a Polish group. A sequence R_n converges to R in $N[T]$ if $R_n\to R$ weakly (in $\mathrm{Aut}(X,\mu)$) and $R_nTR_n^{-1}\to RTR^{-1}$ uniformly (in $[T]$).

Given $R\in N[T]$, denote by $\widetilde{R}$ the Maharam extension of R. Then $\widetilde{R}\in N[\widetilde{T}]$ and it commutes with $(S_t)_{t\in\mathbb{R}}$. Hence it defines a nonsingular transformation mod R on the space (Z,ν) of the associated flow $W=(W_t)_{t\in\mathbb{R}}$ of T. Moreover, mod R belongs to the centralizer $C(W)$ of W in $\mathrm{Aut}(Z,\nu)$. Note that $C(W)$ is a closed subgroup of $(\mathrm{Aut}(Z,\nu),d_w)$.

Let T be of type II_∞ and let μ' be the invariant σ-finite measure equivalent to μ. If $R\in N[T]$ then it is easy

to see that the Radon–Nikodym derivative $\mathrm{d}\mu' \circ R/\mathrm{d}\mu'$ is invariant under T. Hence it is constant, say c. Then $\operatorname{mod} R = \log c$.

Theorem 22 ([86,90]) *If T is of type III then the map* $\operatorname{mod}\colon N[T] \to C(W)$ *is a continuous onto homomorphism. The kernel of this homomorphism is the τ-closure of $[T]$. Hence the quotient group $N[T]/\overline{[T]}^{\tau}$ is (topologically) isomorphic to $C(W)$. In particular, $\overline{[T]}^{\tau}$ is co-compact in $N[T]$ if and only if W is a finite measure-preserving flow with a pure point spectrum.*

The following theorem describes the homotopical structure of normalizers.

Theorem 23 ([34]) *Let T be of type II or III_λ, $0 \le \lambda < 1$. The group $\overline{[T]}^{\tau}$ is contractible. $N[T]$ is homotopically equivalent to $C(W)$. In particular, $N[T]$ is contractible if T is of type II. If T is of type III_λ with $0 < \lambda < 1$ then $\pi_1(N[T]) = \mathbb{Z}$.*

The *outer period* $p(R)$ of $R \in N[T]$ is the smallest positive integer n such that $R^n \in [T]$. We write $p(R) = 0$ if no such n exists.

Two transformations R and R' in $N[T]$ are called *outer conjugate* if there are transformations $V \in N[T]$ and $S \in [T]$ such that $VRV^{-1} = R'S$. The following theorem provides convenient (for verification) necessary and sufficient conditions for the outer conjugacy.

Theorem 24 ([30] for type II and [18] for type III) *Transformations $R, R' \in N[T]$ are outer conjugate if and only if $p(R) = p(R')$ and* $\operatorname{mod} R$ *is conjugate to* $\operatorname{mod} R'$ *in the centralizer of the associated flow for T.*

We note that in the case T is of type II, the second condition in the theorem is just $\operatorname{mod} R = \operatorname{mod} R'$. It is always satisfied when T is of type II_1.

Cocycles of Dynamical Systems. Weak Equivalence of Cocycles

Let G be a locally compact Polish group and λ_G a left Haar measure on G. A Borel map $\varphi\colon X \to G$ is called a *cocycle* of T. Two cocycles φ and φ' are *cohomologous* if there is a Borel map $b\colon X \to G$ such that

$$\varphi'(x) = b(Tx)^{-1}\varphi(x)b(x)$$

for a. a. $x \in X$. A cocycle cohomologous to the trivial one is called a *coboundary*. Given a dense subgroup $G' \subset G$, then every cocycle is cohomologous to a cocycle with values in G' [81]. Each cocycle φ extends to a (unique) map $\alpha_\varphi\colon \mathcal{R}_T \to G$ such that $\alpha_\varphi(Tx, x) = \varphi(x)$ for a. a. x and $\alpha_\varphi(x, y)\alpha_\varphi(y, z) = \alpha_\varphi(x, z)$ for a. a. $(x, y), (y, z) \in \mathcal{R}_T$. α_φ is called the *cocycle of $\mathcal{R}_T$ generated by φ*. Moreover, φ and φ' are cohomologous via b as above if and only if α_φ and $\alpha_{\varphi'}$ are *cohomologous* via b, i. e. $\alpha_\varphi(x, y) = b(x)^{-1}\alpha_{\varphi'}(x, y)b(y)$ for $\mu_{\mathcal{R}_T}$-a. a. $(x, y) \in \mathcal{R}_T$. The following notion was introduced by Golodets and Sinelshchikov [78,81]: two cocycles φ and φ' are *weakly equivalent* if there is a transformation $R \in N[T]$ such that the cocycles α_φ and $\alpha'_\varphi \circ (R \times R)$ of $\mathcal{R}_T$ are cohomologous. Let $\mathcal{M}(X, G)$ denote the set of Borel maps from X to G. It is a Polish group when endowed with the topology of convergence in measure. Since T is ergodic, it is easy to deduce from Rokhlin's lemma that the cohomology class of any cocycle is dense in $\mathcal{M}(X, G)$. Given $\varphi \in \mathcal{M}(X, G)$, we define the *$\varphi$-skew product extension* T_φ of T acting on $(X \times G, \mu \times \lambda_G)$ by setting $T_\varphi(x, g) := (Tx, \varphi(x)g)$. Thus Maharam extension is (isomorphic to) the Radon–Nikodym cocycle-skew product extension. We now specify some basic classes of cocycles [19,35,81,173]:

(i) φ is called *transient* if T_φ is of type I.
(ii) φ is called *recurrent* if T_φ is conservative (equivalently, T_φ is not transient).
(iii) *φ has dense range in G* if T_φ is ergodic.
(iv) φ is called *regular* if φ cobounds with dense range into a closed subgroup H of G (then H is defined up to conjugacy).

These properties are invariant under the cohomology and the weak equivalence. The Radon–Nikodym cocycle ω_1 is a coboundary if and only if T is of type II. It is regular if and only if T is of type II or III_λ, $0 < \lambda \le 1$. It has dense range (in the multiplicative group $\mathbb{R}^*_+$) if and only if T is of type III_1. Notice that ω_1 is never transient (since T is conservative).

Schmidt introduced in [176] an invariant $R(\varphi) := \{g \in G \mid \varphi - g \text{ is recurrent}\}$. He showed in particular that

(i) $R(\varphi)$ is a cohomology invariant,
(ii) $R(\varphi)$ is a Borel set in G,
(iii) $R(\log \omega_1) = \{0\}$ for each aperiodic conservative T,
(iv) there are cocycles φ such that $R(\varphi)$ and $G \setminus R(\varphi)$ are dense in G,
(v) if $\mu(X) = 1$, $\mu \circ T = \mu$ and $\varphi\colon X \to \mathbb{R}$ is integrable then $R(\varphi) = \{\int \varphi \,\mathrm{d}\mu\}$.

We note that (v) follows from Atkinson theorem [15]. A nonsingular version of this theorem was established in [183]: if T is ergodic and μ-nonsingular and $f \in L^1(\mu)$ then

$$\liminf_{n\to\infty} \left| \sum_{j=0}^{n-1} f(T^j x)\omega_j(x) \right| = 0 \quad \text{for a. a. } x$$

if and only if $\int f \,\mathrm{d}\mu = 0$.

Since T_φ commutes with the action of G on $X \times G$ by inverted right translations along the second coordinate, this action induces an ergodic G-action $W_\varphi = (W_\varphi(g))_{g\in G}$ on the space (Z, ν) of T_φ-ergodic components. It is called the *Mackey range (or Poincaré flow)* of φ [66,135,173,188]. We note that φ is regular (and cobounds with dense range into $H \subset G$) if and only if W_φ is transitive (and H is the stabilizer of a point $z \in Z$, i. e. $H = \{g \in G \mid W_\varphi(g)z = z\}$). Hence every cocycle taking values in a compact group is regular.

It is often useful to consider the *double cocycle* $\varphi_0 := \varphi \times \omega_1$ instead of φ. It takes values in the group $G \times \mathbb{R}_+^*$. Since T_{φ_0} is exactly the Maharam extension of T_φ, it follows from [136] that φ_0 is transient or recurrent if and only if φ is transient or recurrent respectively.

Theorem 25 (Orbit classification of cocycles [81]) *Let $\varphi, \varphi' : X \to G$ be two recurrent cocycles of an ergodic transformation T. They are weakly equivalent if and only if their Mackey ranges W_{φ_0} and $W_{\varphi_0'}$ are isomorphic.*

Another proof of this theorem was presented in [65].

Theorem 26 *Let T be an ergodic nonsingular transformation. Then there is a cocycle of T with dense range in G if and only if G is amenable.*

It follows that if G is amenable then the subset of cocycles of T with dense range in G is a dense G_δ in $\mathcal{M}(X, G)$ (just adapt the argument following Example 19). The 'only if' part of Theorem 26 was established in [187]. The 'if' part was considered by many authors in particular cases: G is compact [186], G is solvable or amenable almost connected [79], G is amenable unimodular [108], etc. The general case was proved in [78] and [100] (see also a recent treatment in [9]).

Theorem 21 is a particular case of the following result.

Theorem 27 ([10,65,80]) *Let G be amenable. Let V be an ergodic nonsingular action of $G \times \mathbb{R}_+^*$. Then there is an ergodic nonsingular transformation T and a recurrent cocycle φ of T with values in G such that V is isomorphic to the Mackey range of the double cocycle φ_0.*

Given a cocycle $\varphi \in \mathcal{M}(X, G)$ of T, we say that a transformation $R \in N[T]$ is *compatible with* φ if the cocycles α_φ and $\alpha_\varphi \circ (R \times R)$ of $\mathcal{R}_T$ are cohomologous. Denote by $D(T, \varphi)$ the group of all such R. It has a natural Polish topology which is stronger than τ [41]. Since $[T]$ is a normal subgroup in $D(T, \varphi)$, one can consider the outer conjugacy equivalence relation inside $D(T, \varphi)$. It is called *φ-outer conjugacy*. Suppose that G is Abelian. Then an analogue of Theorem 24 for the φ-outer conjugacy is established in [41]. Also, the cocycles φ with $D(T, \varphi) = N[T]$ are described there.

ITPFI Transformations and AT-Flows

A nonsingular transformation T is called *ITPFI*[1] if it is orbit equivalent to a nonsingular odometer (associated to a sequence $(m_n, \nu_n)_{n=1}^\infty$, see Subsect. "Nonsingular Odometers"). If the sequence m_n can be chosen bounded then T is called ITPFI of bounded type. If $m_n = 2$ for all n then T is called ITPFI$_2$. By [74], every ITPFI-transformation of bounded type is ITPFI$_2$. A remarkable characterization of ITPFI transformations in terms of their associated flows was obtained by Connes and Woods [31]. We first single out a class of ergodic flows. A nonsingular flow $V = (V_t)_{t\in\mathbb{R}}$ on a space (Ω, ν) is called *approximate transitive (AT)* if given $\epsilon > 0$ and $f_1, \dots, f_n \in L_+^1(X, \mu)$, there exists $f \in L_+^1(X, \mu)$ and $\lambda_1, \dots, \lambda_n \in L_+^1(\mathbb{R}, \mathrm{d}t)$ such that

$$\left\| f_j - \int_{\mathbb{R}} f \circ V_t \frac{\mathrm{d}\nu \circ V_t}{\mathrm{d}\nu} \lambda_j(t)\mathrm{d}t \right\|_1 < \epsilon$$

for all $1 \le j \le n$. A flow built under a constant ceiling function with a funny rank-one [67] probability preserving base transformation is AT [31]. In particular, each ergodic finite measure-preserving flow with a pure point spectrum is AT.

Theorem 28 ([31]) *An ergodic nonsingular transformation is ITPFI if and only if its associated flow is AT.*

The original proof of this theorem was given in the framework of von Neumann algebras theory. A simpler, purely measure theoretical proof was given later in [96] (the 'only if' part) and [88] (the 'if' part). It follows from Theorem 28 that every ergodic flow with pure point spectrum is the associated flow of an ITPFI transformation. If the point spectrum of V is $\theta\Gamma$, where Γ is a subgroup of $\mathbb{Q}$ and $\theta \in \mathbb{R}$, then V is the associated flow of an ITPFT$_2$ transformation [91].

Theorem 29 ([54]) *Each ergodic nonsingular transformation is orbit equivalent to a Markov odometer (see Subsect. "Markov Odometers").*

The existence of non-ITPFI transformations and ITPFI transformations of unbounded type was shown in [127]. In [55], an explicit example of a non-ITPFI Markov odometer was constructed.

Smooth Nonsingular Transformations

Diffeomorphisms of smooth manifolds equipped with smooth measures are commonly considered as physically natural examples of dynamical systems. Therefore

[1] This abbreviates 'infinite tensor product of factors of type I' (came from the theory of von Neumann algebras).

the construction of smooth models for various dynamical properties is a well established problem of the modern (probability preserving) ergodic theory. Unfortunately, the corresponding 'nonsingular' counterpart of this problem is almost unexplored. We survey here several interesting facts related to the topic.

For $r \in \mathbb{N} \cup \{\infty\}$, denote by $\mathrm{Diff}_+^r(\mathbb{T})$ the group of orientation preserving C^r-diffeomorphisms of the circle $\mathbb{T}$. Endow this set with the natural Polish topology. Fix $T \in \mathrm{Diff}_+^r(\mathbb{T})$. Since $\mathbb{T} = \mathbb{R}/\mathbb{Z}$, there exists a C^1-function $f\colon \mathbb{R} \to \mathbb{R}$ such that $T(x+\mathbb{Z}) = f(x) + \mathbb{Z}$ for all $x \in \mathbb{R}$. The *rotation number* $\rho(T)$ of T is the limit

$$\lim_{n\to\infty} \underbrace{(f \circ \cdots \circ f)}_{n \text{ times}}(x) (\mathrm{mod}\, 1)\ .$$

The limit exists and does not depend on the choice of x and f. It is obvious that T is nonsingular with respect to Lebesgue measure $\lambda_{\mathbb{T}}$. Moreover, if $T \in \mathrm{Diff}_+^r(\mathbb{T})$ and $\rho(T)$ is irrational then the dynamical system $(\mathbb{T}, \lambda_{\mathbb{T}}, T)$ is ergodic [33]. It is interesting to ask: which Krieger's type can such systems have?

Katznelson showed in [114] that the subset of type III C^∞-diffeomorphisms and the subset of type $\mathrm{II}_\infty C^\infty$-diffeomorphisms are dense in $\mathrm{Diff}_+^\infty(\mathbb{T})$. Hawkins and Schmidt refined the idea of Katznelson from [114] to construct, for every irrational number $\alpha \in [0,1)$ which is not of constant type (i. e. in whose continued fraction expansion the denominators are not bounded) a transformation $T \in \mathrm{Diff}_+^2(\mathbb{T})$ which is of type III_1 and $\rho(T) = \alpha$ [97]. It should be mentioned that class C^2 in the construction is essential, since it follows from a remarkable result of Herman that if $T \in \mathrm{Diff}_+^3(\mathbb{T})$ then under some condition on α (which determines a set of full Lebesgue measure), T is measure theoretically (and topologically) conjugate to a rotation by $\rho(T)$ [101]. Hence T is of type II_1.

In [94], Hawkins shows that every smooth paracompact manifold of dimension ≥ 3 admits a type III_λ diffeomorphism for every $\lambda \in [0,1]$. This extends a result of Herman [100] on the existence of type III_1 diffeomorphisms in the same circumstances.

It is also of interest to ask: which free ergodic flows are associated with smooth dynamical systems of type III_0? Hawkins proved that any free ergodic C^∞-flow on a smooth, connected, paracompact manifold is the associated flow for a C^∞-diffeomorphism on another manifold (of higher dimension) [95].

A nice result was obtained in [115]: if $T \in \mathrm{Diff}_+^2(\mathbb{T})$ and the rotation number of T has unbounded continued fraction coefficients then $(\mathbb{T}, \lambda_{\mathbb{T}}, T)$ is ITPFI. Moreover, a converse also holds: given a nonsingular odometer R, the set of orientation-preserving C^∞-diffeomorphisms of the circle which are orbit equivalent to R is C^∞-dense in the Polish set of all C^∞-orientation-preserving diffeomorphisms with irrational rotation numbers. In contrast to that, Hawkins constructs in [93] a type III_0 C^∞-diffeomorphism of the 4-dimensional torus which is not ITPFI.

Spectral Theory for Nonsingular Systems

While the spectral theory for probability preserving systems is developed in depth, the spectral theory of nonsingular systems is still in its infancy. We discuss below some problems related to L^∞-spectrum which may be regarded as an analogue of the discrete spectrum. We also include results on computation of the maximal spectral type of the 'nonsingular' Koopman operator for rank-one nonsingular transformations.

L^∞-Spectrum and Groups of Quasi-Invariance

Let T be an ergodic nonsingular transformation of $(X, \mathcal{B}, \mu)$. A number $\lambda \in \mathbb{T}$ belongs to the L^∞-spectrum $e(T)$ of T if there is a function $f \in L^\infty(X,\mu)$ with $f \circ T = \lambda f$. f is called an *L^∞-eigenfunction* of T corresponding to λ. Denote by $\mathcal{E}(T)$ the group of all L^∞-eigenfunctions of absolute value 1. It is a Polish group when endowed with the topology of converges in measure. If T is of type II_1 then the L^∞-eigenfunctions are $L^2(\mu')$-eigenfunctions of T, where μ' is an equivalent invariant probability measure. Hence $e(T)$ is countable. Osikawa constructed in [151] the first examples of ergodic nonsingular transformations with uncountable $e(T)$.

We state now a nonsingular version of the von Neumann–Halmos discrete spectrum theorem. Let $Q \subset \mathbb{T}$ be a countable infinite subgroup. Let K be a compact dual of Q_d, where Q_d denotes Q with the discrete topology. Let $k_0 \in K$ be the element defined by $k_0(q) = q$ for all $q \in Q$. Let $R\colon K \to K$ be defined by $Rk = k + k_0$. The system (K, R) is called a *compact group rotation*. The following theorem was proved in [6].

Theorem 30 *Assume that the L^∞-eigenfunctions of T generate the entire σ-algebra $\mathcal{B}$. Then T is isomorphic to a compact group rotation equipped with an ergodic quasi-invariant measure.*

A natural question arises: which subgroups of $\mathbb{T}$ can appear as $e(T)$ for an ergodic T?

Theorem 31 ([1,143]) *$e(T)$ is a Borel subset of $\mathbb{T}$ and carries a unique Polish topology which is stronger than the usual topology on $\mathbb{T}$. The Borel structure of $e(T)$ under this topology agrees with the Borel structure inherited from $\mathbb{T}$. There is a Borel map $\psi\colon e(T) \ni \lambda \mapsto \psi_\lambda \in \mathcal{E}(T)$ such that $\psi_\lambda \circ T = \lambda\psi_\lambda$ for each λ. Moreover, $e(T)$ is of Lebesgue*

measure 0 and it can have an arbitrary Hausdorff dimension.

A proper Borel subgroup E of $\mathbb{T}$ is called

(i) *weak Dirichlet* if $\limsup_{n\to\infty} \widehat{\lambda}(n) = 1$ for each finite complex measure λ supported on E;
(ii) *saturated* if $\limsup_{n\to\infty} |\widehat{\lambda}(n)| \geq |\lambda(E)|$ for each finite complex measure λ on $\mathbb{T}$, where $\widehat{\lambda}(n)$ denote the nth Fourier coefficient of λ.

Every countable subgroup of $\mathbb{T}$ is saturated.

Theorem 32 *$e(T)$ is σ-compact in the usual topology on $\mathbb{T}$ [104] and saturated [104,139].*

It follows that $e(T)$ is weak Dirichlet (this fact was established earlier in [175]).

It is not known if every Polish group continuously embedded in $\mathbb{T}$ as a σ-compact saturated group is the eigenvalue group of some ergodic nonsingular transformation. This is the case for the so-called H_2-groups and the groups of quasi-invariance of measures on $\mathbb{T}$ (see below). Given a sequence n_j of positive integers and a sequence $a_j \geq 0$, the set of all $z \in \mathbb{T}$ such that $\sum_{j=1}^{\infty} a_j |1 - z^{n_j}|^2 < \infty$ is a group. It is called an *H_2-group*. Every H_2-group is Polish in an intrinsic topology stronger than the usual circle topology.

Theorem 33 ([104])

(i) Every H_2-group is a saturated (and hence weak Dirichlet) σ-compact subset of $\mathbb{T}$.
(ii) If $\sum_{j=0}^{\infty} a_j = +\infty$ then the corresponding H_2-group is a proper subgroup of $\mathbb{T}$.
(iii) If $\sum_{j=0}^{\infty} a_j (n_j/n_{j+1})^2 < \infty$ then the corresponding H_2-group is uncountable.
(iv) Any H_2-group is $e(T)$ for an ergodic nonsingular compact group rotation T.

It is an open problem whether every eigenvalue group $e(T)$ is an H_2-group. It is known however that $e(T)$ is close 'to be an H_2-group': if a compact subset $L \subset \mathbb{T}$ is disjoint from $e(T)$ then there is an H_2-group containing $e(T)$ and disjoint from L.

Example 34 ([6], see also [151]) Let (X, μ, T) be the nonsingular odometer associated to a sequence $(2, \nu_j)_{j=1}^{\infty}$. Let n_j be a sequence of positive integers such that $n_j > \sum_{i<j} n_i$ for all j. For $x \in X$, we put $h(x) := n_{l(x)} - \sum_{j<l(x)} n_j$. Then h is a Borel map from X to the positive integers. Let S be the tower over T with height function h (see Subsect. "Tower Transformations"). Then $e(S)$ is the H_2-group of all $z \in \mathbb{T}$ with $\sum_{j=1}^{\infty} \nu_j(0)\nu_j(1)|1 - z^{n_j}|^2 < \infty$.

It was later shown in [104] that if $\sum_{j=1}^{\infty} \nu_j(0)\nu_j(1)(n_j/n_{j+1})^2 < \infty$ then the L^∞-eigenfunctions of S generate the entire σ-algebra, i. e. S is isomorphic (measure theoretically) to a nonsingular compact group rotation.

Let μ be a finite measure on $\mathbb{T}$. Let $H(\mu) := \{z \in \mathbb{Z} \mid \delta_z * \mu \sim \mu\}$, where $*$ means the convolution of measures. Then H_μ is a group called the *group of quasi-invariance of* μ. It has a Polish topology whose Borel sets agree with the Borel sets which $H(\mu)$ inherits from $\mathbb{T}$ and the injection map of $H(\mu)$ into $\mathbb{T}$ is continuous. This topology is induced by the weak operator topology on the unitary group in the Hilbert space $L^2(\mathbb{T}, \mu)$ via the map $H(\mu) \ni z \mapsto U_z$, $(U_z f)(x) = \sqrt{(\mathrm{d}(\delta_z * \mu)/\mathrm{d}\mu)(x)} f(xz)$ for $f \in L^2(\mathbb{T}, \mu)$. Moreover, $H(\mu)$ is saturated [104]. If $\mu(H(\mu)) > 0$ then either $H(\mu)$ is countable or μ is equivalent to $\lambda_{\mathbb{T}}$ [137].

Theorem 35 ([6]) *Let μ be an ergodic with respect to the $H(\mu)$-action by translations on $\mathbb{T}$. Then there is a compact group rotation (K, R) and a finite measure on K quasi-invariant and ergodic under R such that $e(R) = H(\mu)$. Moreover, there is a continuous one-to-one homomorphism $\psi : e(R) \to E(R)$ such that $\psi_\lambda \circ R = \lambda \psi_\lambda$ for all $\lambda \in e(R)$.*

It was shown by Aaronson and Nadkarni [6] that if $n_1 = 1$ and $n_j = a_j a_{j-1} \cdots a_1$ for positive integers $a_j \geq 2$ with $\sum_{j=1}^{\infty} a_j^{-1} < \infty$ then the transformation S from Example 34 does not admit a continuous homomorphism $\psi : e(S) \to E(S)$ with $\psi_\lambda \circ T = \lambda \psi_\lambda$ for all $\lambda \in e(S)$. Hence $e(S) \neq H(\mu)$ for any measure μ satisfying the conditions of Theorem 35.

Assume that T is an ergodic nonsingular compact group rotation. Let $\mathcal{B}_0$ be the σ-algebra generated by a subcollection of eigenfunctions. Then $\mathcal{B}_0$ is invariant under T and hence a factor (see Sect. "Nonsingular Joinings and Factors") of T. It is not known if every factor of T is of this form. It is not even known whether every factor of T must have non-trivial eigenvalues.

Unitary Operator Associated with a Nonsingular System

Let $(X, \mathcal{B}, \mu, T)$ be a nonsingular dynamical system. In this subsection we consider spectral properties of the unitary operator U_T defined by (3). First, we note that the spectrum of T is the entire circle $\mathbb{T}$ [147]. Next, if U_T has an eigenvector then T is of type II_1. Indeed, if there are $\lambda \in \mathbb{T}$ and $0 \neq f \in L^2(X, \mu)$ with $U_T f = \lambda f$ then the measure ν, $\mathrm{d}\nu(x) := |f(x)|^2 \mathrm{d}\mu(x)$, is finite, T-invariant and equivalent to μ. Hence if T is of type III or II_∞ then the maximal spectral type σ_T of U_T is continuous. Another 'restriction' on σ_T was recently found in [166]: no

Foïaş-Strătilă measure is absolutely continuous with respect to σ_T if T is of type II_∞. We recall that a symmetric measure on $\mathbb{T}$ possesses *Foïaş-Strătilă property* if for each ergodic probability preserving system (Y, ν, S) and $f \in L^2(Y, \nu)$, if σ is the spectral measure of f then f is a Gaussian random variable [134]. For instance, measures supported on Kronecker sets possess this property.

Mixing is an L^2-spectral property for type II_∞ transformations: T is mixing if and only if σ_T is a Rajchman measure, i. e. $\widehat{\sigma}_T(n) := \int z^n \, \mathrm{d}\sigma_T(z) \to 0$ as $|n| \to \infty$. Also, T is mixing if and only if $n^{-1} \sum_{i=0}^{n-1} U_T^{k_i} \to 0$ in the strong operator topology for each strictly increasing sequence $k_1 < k_2 < \cdots$ [124]. This generalizes a well known theorem of Blum and Hanson for probability preserving maps. For comparison, we note that ergodicity is not an L^2-spectral property of infinite measure preserving systems.

Now let T be a rank-one nonsingular transformation associated with a sequence $(r_n, w_n, s_n)_{n=1}^\infty$ as in Subsect. "Rank-One Transformations. Chacón Maps. Finite Rank".

Theorem 36 ([25,104]) *The spectral multiplicity of U_T is 1 and the maximal spectral type σ_T of U_T (up to a discrete measure in the case T is of type II_1) is the weak limit of the measures ρ_k defined as follows:*

$$\mathrm{d}\rho_k(z) = \prod_{j=1}^{k} w_j(0)|P_j(z)|^2 \, \mathrm{d}z \,,$$

where $P_j(z) := 1 + \sqrt{w_j(1)/w_j(0)} z^{-R_{1,j}} + \cdots + \sqrt{w_j(m_j - 1)/w_j(0)} z^{-R_{r_j-1,j}}$, $z \in \mathbb{T}$, $R_{i,j} := ih_{j-1} + s_j(0) + \cdots + s_j(i)$, $1 \le i \le r_k - 1$ and h_j is the height of the jth column.

Thus the maximal spectral type of U_T is given by a so-called *generalized Riesz product*. We refer the reader to [25,103,104,148] for a detailed study of Riesz products: their convergence, mutual singularity, singularity to $\lambda_\mathbb{T}$, etc.

It was shown in [6] that $H(\sigma_T) \supset e(T)$ for any ergodic nonsingular transformation T. Moreover, σ_T is ergodic under the action of $e(T)$ by translations if T is isomorphic to an ergodic nonsingular compact group rotation. However it is not known:

(i) Whether $H(\sigma_T) = e(T)$ for all ergodic T.
(ii) Whether ergodicity of σ_T under $e(T)$ implies that T is an ergodic compact group rotation.

The first claim of Theorem 36 extends to the rank N nonsingular systems as follows: if T is an ergodic nonsingular transformation of rank N then the spectral multiplicity of U_T is bounded by N (as in the finite measure-preserving case). It is not known whether this claim is true for a more general class of transformations which are defined as rank N but without the assumption that the Radon–Nikodym cocycle is constant on the tower levels.

Entropy and Other Invariants

Let T be an ergodic conservative nonsingular transformation of a standard probability space $(X, \mathcal{B}, \mu)$. If $\mathcal{P}$ is a finite partition of X, we define the entropy $H(\mathcal{P})$ of $\mathcal{P}$ as $H(\mathcal{P}) = -\sum_{P \in \mathcal{P}} \mu(P) \log \mu(P)$. In the study of measure-preserving systems the classical (Kolmogorov–Sinai) entropy proved to be a very useful invariant for isomorphism [33]. The key fact of the theory is that if $\mu \circ T = \mu$ then the limit $\lim_{n\to\infty} n^{-1} H(\bigvee_{i=1}^n T^{-i}\mathcal{P})$ exists for every $\mathcal{P}$. However if T does not preserve μ, the limit may no longer exist. Some efforts have been made to extend the use of entropy and similar invariants to the nonsingular domain. These include Krengel's entropy of conservative measure-preserving maps and its extension to nonsingular maps, Parry's entropy and Parry's nonsingular version of Shannon–McMillan–Breiman theorem, critical dimension by Mortiss and Dooley, etc. Unfortunately, these invariants are less informative than their classical counterparts and they are more difficult to compute.

Krengel's and Parry's Entropies

Let S be a conservative measure-preserving transformation of a σ-finite measure space $(Y, \mathcal{E}, \nu)$. The *Krengel entropy* [119] of S is defined by

$$h_{\mathrm{Kr}}(S) = \sup\{\nu(E)h(S_E) \mid 0 < \nu(E) < +\infty\} \,,$$

where $h(S_E)$ is the finite measure-preserving entropy of S_E. It follows from Abramov's formula for the entropy of induced transformation that $h_{\mathrm{Kr}}(S) = \mu(E)h(S_E)$ whenever E *sweeps out*, i. e. $\bigcup_{i\ge 0} S^{-i}E = X$. A generic transformation from $\mathrm{Aut}_0(X, \mu)$ has entropy 0. Krengel raised a question in [119]: does there exist a zero entropy infinite measure-preserving S and a zero entropy finite measure-preserving R such that $h_{\mathrm{Kr}}(S \times R) > 0$? This problem was recently solved in [44] (a special case was announced by Silva and Thieullen in an October 1995 AMS conference (unpublished)):

(i) if $h_{\mathrm{Kr}}(S) = 0$ and R is distal then $h_{\mathrm{Kr}}(S \times R) = 0$;
(ii) if R is not distal then there is a rank-one transformation S with $h_{\mathrm{Kr}}(S \times R) = \infty$.

We also note that if a conservative $S \in \mathrm{Aut}_0(X, \mu)$ commutes with another transformation R such that

$\nu \circ R = c\nu$ for a constant $c \neq 1$ then $h_{\mathrm{Kr}}(S)$ is either 0 or ∞ [180].

Now let T be a type III ergodic transformation of $(X, \mathcal{B}, \mu)$. Silva and Thieullen define an entropy $h^*(T)$ of T by setting $h^*(T) := h_{\mathrm{Kr}}(\widetilde{T})$, where $\widetilde{T}$ is the Maharam extension of T (see Subsect. "Maharam Extension, Associated Flow and Orbit Classification of Type III Systems"). Since $\widetilde{T}$ commutes with transformations which 'multiply' $\widetilde{T}$-invariant measure, it follows that $h^*(T)$ is either 0 or ∞.

Let T be the standard III_λ-odometer from Example 17(i). Then $h^*(T) = 0$. The same is true for a so-called ternary odometer associated with the sequence $(3, \nu_n)_{n=1}^{\infty}$, where $\nu_n(0) = \nu_n(2) = \lambda/(1+2\lambda)$ and $\nu_n(1) = \lambda/(1+\lambda)$ [180]. It is not known however whether every ergodic nonsingular odometer has zero entropy. On the other hand, it was shown in [180] that $h^*(T) = \infty$ for every K-automorphism.

The Parry entropy [158] of S is defined by

$$h_{\mathrm{Pa}}(S) := \left\{ H(S^{-1}\mathfrak{F}|\mathfrak{F}) \mid \mathfrak{F} \text{ is a } \sigma\text{-finite subalgebra of } \mathfrak{B} \text{ such that } \mathfrak{F} \subset S^{-1}\mathfrak{F} \right\} .$$

Parry showed [158] that $h_{\mathrm{Pa}}(S) \leq h_{\mathrm{Kr}}(S)$. It is still an open question whether the two entropies coincide. This is the case when S is of rank one (since $h_{\mathrm{Kr}}(S) = 0$) and when S is quasi-finite [158]. The transformation S is called *quasi-finite* if there exists a subset of finite measure $A \subset Y$ such that the first return time partition $(A_n)_{n>0}$ of A has finite entropy. We recall that $x \in A_n \iff n$ is the smallest positive integer such that $T^n x \in A$. An example of non-quasi-finite ergodic infinite measure preserving transformation was constructed recently in [8].

Parry's Generalization of Shannon–MacMillan–Breiman Theorem

Let T be an ergodic transformation of a standard non-atomic probability space $(X, \mathcal{B}, \mu)$. Suppose that $f \circ T \in L^1(X, \mu)$ if and only if $f \in L^1(X, \mu)$. This means that there is $K > 0$ such that $K^{-1} < (\mathrm{d}\mu \circ T)/(\mathrm{d}\mu)(x) < K$ for a. a. x. Let $\mathcal{P}$ be a finite partition of X. Denote by $C_n(x)$ the atom of $\bigvee_{i=0}^{n} T^{-i}\mathcal{P}$ which contains x. We put $\omega_{-1} = 0$. Parry shows in [155] that

$$\frac{\sum_{j=0}^{n} \log \mu\big(C_{n-j}(T^j x)\big)(\omega_j(x) - \omega_{j-1}(x))}{\sum_{i=0}^{n} \omega_j(x)} \to H\left(P \,\middle|\, \bigvee_{i=1}^{\infty} T^{-1}P\right) - \int_X \log E\left(\frac{\mathrm{d}\mu \circ T}{\mathrm{d}\mu} \,\middle|\, \bigvee_{i=0}^{\infty} T^{-i}\mathcal{P}\right) \mathrm{d}\mu$$

for a. a. x. Parry also shows that under the aforementioned conditions on T,

$$\frac{1}{n}\left(\sum_{j=0}^{n} H\left(\bigvee_{i=0}^{j} T^{-j}\mathcal{P}\right) - \sum_{j=0}^{n-1} H\left(\bigvee_{i=1}^{j+1} T^{-j}\mathcal{P}\right)\right) \to H\left(\mathcal{P} \,\middle|\, \bigvee_{i=1}^{\infty} T^{-i}\mathcal{P}\right) .$$

Critical Dimension

The critical dimension introduced by Mortiss [146] measures the order of growth for sums of Radon–Nikodym derivatives. Let $(X, \mathcal{B}, \mu, T)$ be an ergodic nonsingular dynamical system. Given $\delta > 0$, let

$$X_\delta := \left\{ x \in X \,\middle|\, \liminf_{n\to\infty} \frac{\sum_{i=0}^{n-1} \omega_i(x)}{n^\delta} > 0 \right\} \text{ and} \tag{4}$$

$$X^\delta := \left\{ x \in X \,\middle|\, \liminf_{n\to\infty} \frac{\sum_{i=0}^{n-1} \omega_i(x)}{n^\delta} = 0 \right\} . \tag{5}$$

Then X_δ and X^δ are T-invariant subsets.

Definition 37 ([57,146]) The *lower critical dimension* $\alpha(T)$ of T is $\sup\{\delta \mid \mu(X_\delta) = 1\}$. The *upper critical dimension* $\beta(T)$ of T is $\inf\{\delta \mid \mu(X^\delta) = 1\}$.

It was shown in [57] that the lower and upper critical dimensions are invariants for isomorphism of nonsingular systems. Notice also that

$$\alpha(T) = \liminf_{n\to\infty} \frac{\log\left(\sum_{i=1}^{n} \omega_i(x)\right)}{\log n} \text{ and}$$

$$\beta(T) = \limsup_{n\to\infty} \frac{\log\left(\sum_{i=1}^{n} \omega_i(x)\right)}{\log n} .$$

Moreover, $0 \leq \alpha(T) \leq \beta(T) \leq 1$. If T is of type II_1 then $\alpha(T) = \beta(T) = 1$. If T is the standard III_λ-odometer from Example 17 then $\alpha(T) = \beta(T) = \log(1+\lambda) - \lambda/(1+\lambda)\log\lambda$.

Theorem 38

(i) *For every $\lambda \in [0,1]$ and every $c \in [0,1]$ there exists a nonsingular odometer of type III_λ with critical dimension equal to c [145].*

(ii) For every $c \in [0,1]$ there exists a nonsingular odometer of type II_∞ with critical dimension equal to c [57].

Let T be the nonsingular odometer associated with a sequence $(m_n, \nu_n)_{n=1}^\infty$. Let $s(n) = m_1 \cdots m_n$ and let $H(\mathcal{P}_n)$ denote the entropy of the partition of the first n coordinates with respect to μ. We now state a nonsingular version of Shannon–MacMillan–Breiman theorem for T from [57].

Theorem 39 *Let m_i be bounded from above. Then*

(i)

$$\alpha(T) = \liminf_{n\to\infty} \inf \frac{-\sum_{i=1}^n \log m_i(x_i)}{\log s(n)} = \liminf_{n\to\infty} \frac{H(\mathcal{P}_n)}{\log s(n)}$$

and

(ii)

$$\beta(T) = \limsup_{n\to\infty} \inf \frac{-\sum_{i=1}^n \log m_i(x_i)}{\log s(n)} = \limsup_{n\to\infty} \frac{H(\mathcal{P}_n)}{\log s(n)}$$

for a. a. $x = (x_i)_{i\geq 1} \in X$.

It follows that in the case when $\alpha(T) = \beta(T)$, the critical dimension coincides with $\lim_{n\to\infty} H(\mathcal{P}_n)/(\log s(n))$. In [145] this expression (when it exists) was called *AC-entropy* (average coordinate). It also follows from Theorem 39 that if T is an odometer of bounded type then $\alpha(T^{-1}) = \alpha(T)$ and $\beta(T^{-1}) = \beta(T)$. In [58], Theorem 39 was extended to a subclass of Markov odometers. The critical dimensions for Hamachi shifts (see Subsect. "Nonsingular Bernoulli Transformations – Hamachi's Example") were investigated in [59]:

Theorem 40 *For any $\epsilon > 0$, there exists a Hamachi shift S with $\alpha(S) < \epsilon$ and $\beta(S) > 1 - \epsilon$.*

Nonsingular Restricted Orbit Equivalence

In [144] Mortiss initiated study of a nonsingular version of Rudolph's restricted orbit equivalence [167]. This work is still in its early stages and does not yet deal with any form of entropy. However she introduced nonsingular orderings of orbits, defined sizes and showed that much of the basic machinery still works in the nonsingular setting.

Nonsingular Joinings and Factors

The theory of joinings is a powerful tool to study probability preserving systems and to construct striking counterexamples. It is interesting to study what part of this machinery can be extended to the nonsingular case. However, the definition of nonsingular joining is far from being obvious. Some progress was achieved in understanding 2-fold joinings and constructing prime systems of any Krieger type. As far as we know the higher-fold nonsingular joinings have not been considered so far. It turned out however that an alternative coding technique, predating joinings in studying the centralizer and factors of the classical measure-preserving Chacón maps, can be used as well to classify factors of Cartesian products of some nonsingular Chacón maps.

Joinings, Nonsingular MSJ and Simplicity

In this section all measures are probability measures. A *nonsingular joining* of two nonsingular systems $(X_1, \mathcal{B}_1, \mu_1, T_1)$ and $(X_2, \mathcal{B}_2, \mu_2, T_2)$ is a measure $\hat{\mu}$ on the product $\mathcal{B}_1 \times \mathcal{B}_2$ that is nonsingular for $T_1 \times T_2$ and satisfies: $\hat{\mu}(A \times X_2) = \mu_1(A)$ and $\hat{\mu}(X_1 \times B) = \mu_2(B)$ for all $A \in \mathcal{B}_1$ and $B \in \mathcal{B}_2$. Clearly, the product $\mu_1 \times \mu_2$ is a nonsingular joining. Given a transformation $S \in C(T)$, the measure μ_S given by $\mu_S(A \times B) := \mu(A \cap S^{-1}B)$ is a nonsingular joining of (X, μ, T) and $(X, \mu \circ S^{-1}, T)$. It is called a *graph*-joining since it is supported on the graph of S. Another important kind of joinings that we are going to define now is related to factors of dynamical systems. Recall that given a nonsingular system $(X, \mathcal{B}, \mu, T)$, a sub-σ-algebra $\mathcal{A}$ of $\mathcal{B}$ such that $T^{-1}(\mathcal{A}) = \mathcal{A} \bmod \mu$ is called a *factor* of T. There is another, equivalent, definition. A nonsingular dynamical system $(Y, \mathcal{C}, \nu, S)$ is called a factor of T if there exists a measure-preserving map $\varphi\colon X \to Y$, called a *factor map*, with $\varphi T = S\varphi$ a. e. (If φ is only nonsingular, ν may be replaced with the equivalent measure $\mu \circ \varphi^{-1}$, for which φ is measure-preserving.) Indeed, the sub-σ-algebra $\varphi^{-1}(\mathcal{C}) \subset \mathcal{B}$ is T-invariant and, conversely, any T-invariant sub-σ-algebra of $\mathcal{B}$ defines a factor map by immanent properties of standard probability spaces, see e. g. [3]. If φ is a factor map as above, then μ has a disintegration with respect to φ, i. e., $\mu = \int \mu_y \mathrm{d}\nu(y)$ for a measurable map $y \mapsto \mu_y$ from Y to the probability measures on X so that $\mu_y(\varphi^{-1}(y)) = 1$, the measure $\mu_{S\varphi(x)} \circ T$ is equivalent to $\mu_{\varphi(x)}$ and

$$\frac{\mathrm{d}\mu \circ T}{\mathrm{d}\mu}(x) = \frac{\mathrm{d}\nu \circ S}{\mathrm{d}\nu}(\varphi(x)) \frac{\mathrm{d}\mu_{S\varphi(x)} \circ T}{\mathrm{d}\mu_{\varphi(x)}}(x) \tag{6}$$

for a. e. $x \in X$. Define now the *relative product* $\hat{\mu} = \mu \times_\varphi \mu$ on $X \times X$ by setting $\hat{\mu} = \int \mu_y \times \mu_y \, \mathrm{d}\nu(y)$. Then

it is easy to deduce from (6) that $\hat{\mu}$ is a nonsingular self-joining of T.

We note however that the above definition of joining is not satisfactory since it does not reduce to the classical definition when we consider probability preserving systems. Indeed, the following result was proved in [168].

Theorem 41 *Let $(X_1, \mathcal{B}_1, \mu_1, T_1)$ and $(X_2, \mathcal{B}_2, \mu_2, T_2)$ be two finite measure-preserving systems such that $T_1 \times T_2$ is ergodic. Then for every $\lambda, 0 < \lambda < 1$, there exists a nonsingular joining $\hat{\mu}$ of μ_1 and μ_2 such that $(T_1 \times T_2, \hat{\mu})$ is ergodic and of type III_λ.*

It is not known however if the nonsingular joining $\hat{\mu}$ can be chosen in every orbit equivalence class. In view of the above, Rudolph and Silva [168] isolate an important subclass of joining. It is used in the definition of a nonsingular version of minimal self-joinings.

Definition 42

(i) A nonsingular joining $\hat{\mu}$ of (X_1, μ_1, T_1) and (X_2, μ_2, T_2) is *rational* if there exit measurable functions $c^1 \colon X_1 \to \mathbb{R}_+$ and $c^2 \colon X_2 \to \mathbb{R}_+$ such that

$$\begin{aligned}\hat{\omega}_1^{\hat{\mu}}(x_1, x_2) &= \omega_1^{\mu_1}(x_1)\omega_1^{\mu_2}(x_2)c^1(x_1)\\ &= \omega_1^{\mu_1}(x_1)\omega_1^{\mu_2}(x_2)c^2(x_2) \quad \hat{\mu}\text{ a. e.}\end{aligned}$$

(ii) A nonsingular dynamical system $(X, \mathcal{B}, \mu, T)$ has *minimal self-joinings (MSJ)* over a class $\mathcal{M}$ of probability measures equivalent to μ, if for every $\mu_1, \mu_2 \in \mathcal{M}$, for every rational joining $\hat{\mu}$ of μ_1, μ_2, a. e. ergodic component of $\hat{\mu}$ is either the product of its marginals or is the graph-joining supported on T^j for some $j \in \mathbb{Z}$.

Clearly, product measure, graph-joinings and the relative products are all rational joinings. Moreover, a rational joining of finite measure-preserving systems is measure-preserving and a rational joining of type II_1's is of type II_1 [168]. Thus we obtain the finite measure-preserving theory as a special case. As for the definition of MSJ, it depends on a class $\mathcal{M}$ of equivalent measures. In the finite measure-preserving case $\mathcal{M} = \{\mu\}$. However, in the nonsingular case no particular measure is distinguished. We note also that Definition 42(ii) involves some restrictions on all rational joinings and not only ergodic ones as in the finite measure-preserving case. The reason is that an ergodic component of a nonsingular joining needs not be a joining of measures equivalent to the original ones [2]. For finite measure-preserving transformations, MSJ over $\{\mu\}$ is the same as the usual 2-fold MSJ [49].

A nonsingular transformation T on $(X, \mathcal{B}, \mu)$ is called *prime* if its only factors are $\mathcal{B}$ and $\{X, \emptyset\}$ mod μ. A (non-empty) class $\mathcal{M}$ of probability measures equivalent to μ is said to be *centralizer stable* if for each $S \in C(T)$ and $\mu_1 \in \mathcal{M}$, the measure $\mu_1 \circ S$ is in $\mathcal{M}$.

Theorem 43 ([168]) *Let $(X, \mathcal{B}, \mu, T)$ be a ergodic non-atomic dynamical system such that T has MSJ over a class $\mathcal{M}$ that is centralizer stable. Then T is prime and the centralizer of T consists of the powers of T.*

A question that arises is whether if such nonsingular dynamical system (not of type II_1) exist. Expanding on Ornstein's original construction from [150], Rudolph and Silva construct in [168], for each $0 \le \lambda \le 1$, a nonsingular rank-one transformation T_λ that is of type III_λ and that has MSJ over a class $\mathcal{M}$ that is centralizer stable. Type II_∞ examples with analogues properties were also constructed there. In this connection it is worth to mention the example by Aaronson and Nadkarni [6] of II_∞ ergodic transformations that have no factor algebras on which the invariant measure is σ-finite (except for the trivial and the entire ones); however these transformations are not prime.

A more general notion than MSJ called *graph self-joinings (GSJ)*, was introduced [181]: just replace the the words "on T^j for some $j \in \mathbb{Z}$" in Definition 3(ii) with "on S for some element $S \in C(T)$". For finite measure-preserving transformations, GSJ over $\{\mu\}$ is the same as the usual 2-fold simplicity [49]. The famous Veech theorem on factors of 2-fold simple maps (see [49]) was extended to nonsingular systems in [181] as follows: if a system $(X, \mathcal{B}, \mu, T)$ has GSJ then for every non-trivial factor $\mathcal{A}$ of T there exists a locally compact subgroup H in $C(T)$ (equipped with the weak topology) which acts smoothly (i. e. the partition into H-orbits is measurable) and such that $\mathcal{A} = \{B \in \mathcal{B} \mid \mu(hB \triangle B) = 0 \text{ for all } h \in H\}$. It follows that there is a cocycle φ from $(X, \mathcal{A}, \mu \restriction \mathcal{A})$ to H such that T is isomorphic to the φ-skew product extension $(T \restriction A)_\varphi$ (see Subsect. "Cocycles of Dynamical Systems. Weak Equivalence of Cocycles"). Of course, the ergodic nonsingular odometers and, more generally, ergodic nonsingular compact group rotation (see Subsect. "L^∞-Spectrum and Groups of Quasi-Invariance") have GSJ. However, except for this trivial case (the Cartesian square is non-ergodic) plus the systems with MSJ from [168], no examples of type III systems with GSJ are known. In particular, no smooth examples have been constructed so far. This is in sharp contrast with the finite measure preserving case where abundance of simple (or close to simple) systems are known (see [39,40,49,182]).

Nonsingular Coding and Factors of Cartesian Products of Nonsingular Maps

As we have already noticed above, the nonsingular MSJ theory was developed in [168] only for 2-fold self-joinings. The reasons for this were technical problems with extending the notion of rational joinings form 2-fold to n-fold self-joinings. However while the 2-fold nonsingular MSJ or GSJ properties of T are sufficient to control the centralizer and the factors of T, it is not clear whether it implies anything about the factors or centralizer of $T \times T$. Indeed, to control them one needs to know the 4-fold joinings of T. However even in the finite measure-preserving case it is a long standing open question whether 2-fold MSJ implies n-fold MSJ. That is why del Junco and Silva [51] apply an alternative – nonsingular coding – techniques to classify the factors of Cartesian products of nonsingular Chacón maps. The techniques were originally used in [48] to show that the classical Chacón map is prime and has trivial centralizer. They were extended to nonsingular systems in [50].

For each $0 < \lambda < 1$ we denote by T_λ the Chacón map (see Subsect. "Rank-One Transformations. Chacón Maps. Finite Rank") corresponding the sequence of probability vectors $w_n = (\lambda/(1+2\lambda), 1/(1+2\lambda), \lambda/(1+2\lambda))$ for all $n > 0$. One can verify that the maps T_λ are of type III_λ. (The classical Chacón map corresponds to $\lambda = 1$.) All of these transformations are defined on the same standard Borel space $(X, \mathcal{B})$. These transformations were shown to be power weakly mixing in [12]. The centralizer of any finite Cartesian product of nonsingular Chacón maps is computed in the following theorem.

Theorem 44 ([51]) *Let $0 < \lambda_1 < \ldots < \lambda_k \leq 1$ and $n_1, \ldots, n_k$ be positive integers. Then the centralizer of the Cartesian product $T_{\lambda_1}^{\otimes n_1} \times \ldots \times T_{\lambda_k}^{\otimes n_k}$ is generated by maps of the form $U_1 \times \ldots \times U_k$, where each U_i, acting on the n_i-dimensional product space X^{n_i}, is a Cartesian product of powers of T_{λ_i} or a co-ordinate permutation on X^{n_i}.*

Let π denote the permutation on $X \times X$ defined by $\pi(x, y) = (y, x)$ and let $\mathcal{B}^{2\odot}$ denote the symmetric factor, i. e. $\mathcal{B}^{2\odot} = \{A \in \mathcal{B} \otimes \mathcal{B} \mid \pi(A) = A\}$. The following theorem classifies the factors of the Cartesian product of any two nonsingular type III_λ, $0 < \lambda < 1$, or type II_1 Chacón maps.

Theorem 45 ([51]) *Let T_{λ_1} and T_{λ_2} be two nonsingular Chacón systems. Let $\mathcal{F}$ be a factor algebra of $T_{\lambda_1} \times T_{\lambda_2}$.*

(i) If $\lambda_1 \neq \lambda_2$ then $\mathcal{F}$ is equal mod 0 *to one of the four algebras $\mathcal{B} \otimes \mathcal{B}$, $\mathcal{B} \otimes \mathcal{N}$, $\mathcal{N} \otimes \mathcal{B}$, or $\mathcal{N} \otimes \mathcal{N}$, where $\mathcal{N} = \{\emptyset, X\}$.*

(ii) If $\lambda_1 = \lambda_2$ then $\mathcal{F}$ is equal mod 0 *to one of the following algebras $\mathcal{B} \otimes C$, $\mathcal{B} \otimes \mathcal{N}$, $\mathcal{N} \otimes C$, $\mathcal{N} \otimes \mathcal{N}$, or $(T^m \times \mathrm{Id})\mathcal{B}^{2\odot}$ for some integer m.*

It is not hard to obtain type III_1 examples of Chacón maps for which the previous two theorems hold. However the construction of type II_∞ and type III_0 nonsingular Chacón transformations is more subtle as it needs the choice of ω_n to vary with n. In [92], Hamachi and Silva construct type III_0 and type II_∞ examples, however the only property proved for these maps is ergodicity of their Cartesian square. More recently, Danilenko [38] has shown that all of them (in fact, a wider class of nonsingular Chacón maps of all types) are power weakly mixing.

In [22], Choksi, Eigen and Prasad asked whether there exists a zero entropy, finite measure-preserving mixing automorphism S, and a nonsingular type III automorphism T, such that $T \times S$ has no Bernoulli factors. Theorem 45 provides a partial answer (with a mildly mixing only instead of mixing) to this question: if S is the finite measure-preserving Chacón map and T is a nonsingular Chacón map as above, the factors of $T \times S$ are only the trivial ones, so $T \times S$ has no Bernoulli factors.

Applications. Connections with Other Fields

In this – final – section we shed light on numerous mathematical sources of nonsingular systems. They come from the theory of stochastic processes, random walks, locally compact Cantor systems, horocycle flows on hyperbolic surfaces, von Neumann algebras, statistical mechanics, representation theory for groups and anticommutation relations, etc. We also note that such systems sometimes appear in the context of probability preserving dynamics (see also a criterium of distality in Subsect. "Krengel's and Parry's Entropies").

Mild Mixing

An ergodic finite measure-preserving dynamical system $(X, \mathcal{B}, \mu, T)$ is called *mildly mixing* if for each non-trivial factor algebra $\mathcal{A} \subset \mathcal{B}$, the restriction $T \restriction \mathcal{A}$ is not rigid. For equivalent definitions and extensions to actions of locally compact groups we refer to [3] and [177]. There is an interesting criterium of the mild mixing that involves nonsingular systems: T is mildly mixing if and only if for each ergodic nonsingular transformation S, the product $T \times S$ is ergodic [71]. Furthermore, $T \times S$ is then orbit equivalent to S [98]. Moreover, if R is a nonsingular transformation such that $R \times S$ is ergodic for any ergodic nonsingular S then R is of type II_1 (and mildly mixing) [177].

Disjointness and Furstenberg's Class $\mathcal{W}^{\perp}$

Two probability preserving systems (X, μ, T) and (Y, ν, S) are called *disjoint* if $\mu \times \nu$ is the only $T \times S$-invariant probability measure on $X \times Y$ whose coordinate projections are μ and ν respectively. Furstenberg in [69] initiated studying the class $\mathcal{W}^{\perp}$ of transformations disjoint from all weakly mixing ones. Let $\mathcal{D}$ denote the class of distal transformations and $\mathcal{M}(\mathcal{W}^{\perp})$ the class of multipliers of $\mathcal{W}^{\perp}$ (for the definitions see [75]). Then $\mathcal{D} \subset \mathcal{M}(\mathcal{W}^{\perp}) \subset \mathcal{W}^{\perp}$. In [43] and [133] it was shown by constructing explicit examples that these inclusions are strict. We record this fact here because nonsingular ergodic theory was the key ingredient of the arguments in the two papers pertaining to the theory of probability preserving systems. The examples are of the form $T_{\varphi,S}(x, y) = (Tx, S_{\varphi(x)}y)$, where T is an ergodic rotation on (X, μ), $(S_g)_{g\in G}$ a mildly mixing action of a locally compact group G on Y and $\varphi\colon X \to G$ a measurable map. Let W_φ denote the Mackey action of G associated with φ and let (Z, κ) be the space of this action. The key observation is that there exists an affine isomorphism of the simplex of $T_{\varphi,S}$-invariant probability measures whose pullback on X is μ and the simplex of $W_\varphi \times S$ quasi-invariant probability measures whose pullback on Z is κ and whose Radon–Nikodym cocycle is measurable with respect to Z. This is a far reaching generalization of Furstenberg theorem on relative unique ergodicity of ergodic compact group extensions.

Symmetric Stable and Infinitely Divisible Stationary Processes

Rosinsky in [163] established a remarkable connection between structural studies of stationary stochastic processes and ergodic theory of nonsingular transformations (and flows). For simplicity we consider only real processes in discrete time. Let $X = (X_n)_{n\in\mathbb{Z}}$ be a measurable stationary symmetric α-stable (SαS) process, $0 < \alpha < 2$. This means that any linear combination $\sum_{k=1}^{n} a_k X_{j_k}$, $j_k \in \mathbb{Z}$, $a_k \in \mathbb{R}$ has an SαS-distribution. (The case $\alpha = 2$ corresponds to Gaussian processes.) Then the process admits a spectral representation

$$X_n = \int_Y f_n(y)\, M(\mathrm{d}y)\,, \quad n \in \mathbb{Z}\,, \tag{7}$$

where $f_n \in L^\alpha(Y, \mu)$ for a standard σ-finite measure space $(Y,\mathcal{B},\mu)$ and M is an independently scattered random measure on $\mathcal{B}$ such that $E\exp(iuM(A)) = \exp(-|u|^\alpha \mu(A))$ for every $A \in \mathcal{B}$ of finite measure. By [163], one can choose the kernel $(f_n)_{n\in\mathbb{Z}}$ in a special way: there are a μ-nonsingular transformation T and measurable maps $\varphi\colon X \to \{-1, 1\}$ and $f \in L^\alpha(Y, \mu)$ such that $f_n = U^n f$, $n \in \mathbb{Z}$, where U is the isometry of $L^\alpha(X, \mu)$ given by $Ug = \varphi \cdot (\mathrm{d}\mu \circ T/\mathrm{d}\mu)^{1/\alpha} \cdot g \circ T$. If, in addition, the smallest T-invariant σ-algebra containing $f^{-1}(\mathcal{B}_{\mathbb{R}})$ coincides with $\mathcal{B}$ and $\mathrm{Supp}\,\{f \circ T^n\colon n \in \mathbb{Z}\} = Y$ then the pair (T, φ) is called minimal. It turns out that minimal pairs always exist. Moreover, two minimal pairs (T, φ) and (T', φ') representing the same SαS process are equivalent in some natural sense [163]. Then one can relate ergodic-theoretical properties of (T, φ) to probabilistic properties of $(X_n)_{n\in\mathbb{Z}}$. For instance, let $Y = C \sqcup D$ be the Hopf decomposition of Y (see Theorem 2). We let $X_n^D := \int_D f_n(y)\, M(\mathrm{d}y)$ and $X_n^C := \int_C f_n(y)\, M(\mathrm{d}y)$. Then we obtain a unique (in distribution) decomposition of X into the sum $X^D + X^C$ of two independent stationary SαS-processes.

Another kind of decomposition was considered in [171]. Let P be the largest invariant subset of Y such that $T \restriction P$ has a finite invariant measure. Partitioning Y into P and $N := Y \setminus N$ and restricting the integration in (7) to P and N we obtain a unique (in distribution) decomposition of X into the sum $X^P + X^N$ of two independent stationary SαS-processes. Notice that the process X is ergodic if and only if $\mu(P) = 0$.

Recently, Roy considered a more general class of *infinitely divisible (ID)* stationary processes [165]. Using Maruyama's representation of the characteristic function of an ID process X without Gaussian part he singled out the Lévy measure Q of X. Then Q is a shift invariant σ-finite measure on $\mathbb{R}^{\mathbb{Z}}$. Decomposing the dynamical system $(\mathbb{R}^{\mathbb{Z}}, \tau, Q)$ in various natural ways (Hopf decomposition, 0-type and positive type, so-called 'rigidity free' part and its complement) he obtains corresponding decompositions for the process X. Here τ stands for the shift on $\mathbb{R}^{\mathbb{Z}}$.

Poisson Suspensions

Poisson suspensions are widely used in statistical mechanics to model ideal gas, Lorentz gas, etc (see [33]). Let $(X, \mathcal{B}, \mu)$ be a standard σ-finite non-atomic measure space and $\mu(X) = \infty$. Denote by $\widetilde{X}$ the space of unordered countable subsets of X. It is called the space of *configurations*. Fix $t > 0$. Let $A \in \mathcal{B}$ have positive finite measure and let $j \in \mathbb{Z}_+$. Denote by $[A, j]$ the subset of all configurations $\widetilde{x} \in \widetilde{X}$ such that $\#(\widetilde{x} \cap A) = j$. Let $\widetilde{\mathcal{B}}$ be the σ-algebra generated by all $[A, j]$. We define a probability measure $\widetilde{\mu}_t$ on $\widetilde{\mathcal{B}}$ by two conditions:

(i) $\widetilde{\mu}_t([A, j]) = \frac{(t\mu(A))^j}{j!} \exp(-t\mu(A))$;
(ii) if $A_1, \ldots, A_p$ are pairwise disjoint then $\widetilde{\mu}_t(\bigcap_{k=1}^{p} [A_k, j_k]) = \prod_{k=1}^{p} \widetilde{\mu}_t([A_k, j_k])$.

If T is a μ-preserving transformation of X and $\widetilde{x} = (x_1, x_2, \dots)$ is a configuration then we set $\widetilde{T}\omega := (Tx_1, Tx_2, \dots)$. It is easy to verify that $\widetilde{T}$ is a $\widetilde{\mu}$-preserving transformation of $\widetilde{X}$. The dynamical system $(\widetilde{X}, \widetilde{\mathcal{B}}, \widetilde{\mu}, \widetilde{T})$ is called the *Poisson suspension* above $(X, \mathcal{B}, \mu, T)$. It is ergodic if and only if T has no invariant sets of finite positive measure. There is a canonical representation of $L^2(\widetilde{X}, \widetilde{\mu})$ as the Fock space over $L^2(X, \mu)$ such that the unitary operator $U_{\widetilde{T}}$ is the 'exponent' of U_T. Thus, the maximal spectral type of $U_{\widetilde{T}}$ is $\sum_{n\geq 0}(n!)^{-1}\sigma^{*n}$, where σ is a measure of the maximal spectral type of U_T. It is easy to see that a σ-finite factor of T corresponds to a factor (called Poissonian) of $\widetilde{T}$. Moreover, a σ-finite measure-preserving joining (with σ-finite projections) of two infinite measure-preserving transformations T_1 and T_2 generates a joining (called Poissonian) of $\widetilde{T_1}$ and $\widetilde{T_2}$ [52,164]. Thus we see a similarity with the well studied theory of Gaussian dynamical systems [134]. However, the Poissonian case is less understood. There was a recent progress in this field. Parreau and Roy constructed Poisson suspensions whose ergodic self-joinings are all Poissonian [154]. In [111] partial solutions of the following (still open) problems are found:

(i) whether the Pinsker factor of $\widetilde{T}$ is Poissonian,
(ii) what is the relationship between Krengel's entropy of T, Parry's entropy of T and Kolmogorov–Sinai entropy of $\widetilde{T}$.

Recurrence of Random Walks with Non-stationary Increments

Using nonsingular ergodic theory one can introduce the notion of recurrence for random walks obtained from certain non-stationary processes. Let T be an ergodic nonsingular transformation of a standard probability space $(X, \mathcal{B}, \mu)$ and let $f\colon X \to \mathbb{R}^n$ a measurable function. Define for $m \geq 1$, $Y_m\colon X \to \mathbb{R}^n$ by $Y_m := \sum_{n=0}^{m-1} f \circ T^n$. In other words, $(Y_m)_{m\geq 1}$ is the random walk associated with the (non-stationary) process $(f \circ T^n)_{n\geq 0}$. Let us call this random walk *recurrent* if the cocycle f of T is recurrent (see Subsect. "Cocycles of Dynamical Systems. Weak Equivalence of Cocycles"). It was shown in [176] that in the case $\mu \circ T = \mu$, i. e. the process is stationary, this definition is equivalent to the standard one.

Boundaries of Random Walks

Boundaries of random walks on groups retain valuable information on the underlying groups (amenability, entropy, etc.) and enable one to obtain integral representation for harmonic functions of the random walk [112,186,187]. Let G be a locally compact group and ν a probability measure on G. Let T denote the (one-sided) shift on the probability space $(X, \mathcal{B}_X, \mu) := (G, \mathcal{B}_G, \nu)^{\mathbb{Z}_+}$ and $\varphi\colon X \to G$ a measurable map defined by $(y_0, y_1, \dots) \mapsto y_0$. Let T_φ be the φ-skew product extension of T acting on the space $(X \times G, \mu \times \lambda_G)$ (for non-invertible transformations the skew product extension is defined in the very same way as for invertible ones, see Subsect. "Cocycles of Dynamical Systems. Weak Equivalence of Cocycles"). Then T_φ is isomorphic to the *homogeneous random walk* on G with jump probability ν. Let $\mathcal{I}(T_\varphi)$ denote the sub-σ-algebra of T_φ-invariant sets and let $\mathcal{F}(T_\varphi) := \bigcap_{n>0} T_\varphi^{-n}(\mathcal{B}_X \otimes \mathcal{B}_G)$. The former is called the *Poisson boundary* of T_φ and the latter one is called the *tail boundary* of T_φ. Notice that a nonsingular action of G by inverted right translations along the second coordinate is well defined on each of the two boundaries. The two boundaries (or, more precisely, the G-actions on them) are ergodic. The Poisson boundary is the Mackey range of φ (as a cocycle of T). Hence the Poisson boundary is amenable [187]. If the support of ν generates a dense subgroup of G then the corresponding Poisson boundary is weakly mixing [4]. As for the tail boundary, we first note that it can be defined for a wider family of *non-homogeneous* random walks. This means that the jump probability ν is no longer fixed and a sequence $(\nu_n)_{n>0}$ of probability measures on G is considered instead. Now let $(X, \mathcal{B}_X, \mu) := \prod_{n>0}(G, \mathcal{B}_G, \nu)$. The one-sided shift on X may not be nonsingular now. Instead of it, we consider the tail equivalence relation $\mathcal{R}$ on X and a cocycle $\alpha\colon \mathcal{R} \to G$ given by $\alpha(x, y) = x_1 \cdots x_n y_n^{-1} \cdots y_1$, where $x = (x_i)_{i>0}$ and $y = (y_i)_{i>0}$ are $\mathcal{R}$-equivalent and n in the smallest integer such that $x_i = y_i$ for all $i > n$. The tail boundary of the random walk on G with time dependent jump probabilities $(\nu_n)_{n>0}$ is the Mackey G-action associated with α. In the case of homogeneous random walks this definition is equivalent to the initial one. Connes and Woods showed [32] that the tail boundary is always amenable and AT. It is unknown whether the converse holds for general G. However it is true for $G = \mathbb{R}$ and $G = \mathbb{Z}$: the class of AT-flows coincides with the class of tail boundaries of the random walks on $\mathbb{R}$ and a similar statement holds for $\mathbb{Z}$ [32]. Jaworsky showed [109] that if G is countable and a random walk is homogeneous then the tail boundary of the random walk possesses a so-called SAT-property (which is stronger than AT).

Classifying σ-Finite Ergodic Invariant Measures

The description of ergodic finite invariant measures for topological (or, more generally, standard Borel) systems

is a well established problem in the classical ergodic theory [33]. On the other hand, it seems impossible to obtain any useful information about the system by analyzing the set of all ergodic quasi-invariant (or just σ-finite invariant) measures because this set is wildly huge (see Subsect. "The Glimm–Effros Theorem"). The situation changes if we impose some restrictions on the measures. For instance, if the system under question is a homeomorphism (or a topological flow) defined on a locally compact Polish space then it is natural to consider the class of (σ-finite) invariant Radon measures, i. e. measures taking finite values on the compact subsets. We give two examples.

First, the seminal results of Giordano, Putnam and Skau on the topological orbit equivalence of compact Cantor minimal systems were extended to locally compact Cantor minimal (l.c.c.m.) systems in [37] and [138]. Given a l.c.c.m. system X, we denote by $\mathcal{M}(X)$ and $\mathcal{M}_1(X)$ the set of invariant Radon measures and the set of invariant probability measures on X. Notice that $\mathcal{M}_1(X)$ may be empty [37]. It was shown in [138] that two systems X and X' are topologically orbit equivalent if and only if there is a homeomorphism of X onto X' which maps bijectively $\mathcal{M}(X)$ onto $\mathcal{M}(X')$ and $\mathcal{M}_1(X)$ onto $\mathcal{M}_1(X')$. Thus $\mathcal{M}(X)$ retains an important information on the system – it is 'responsible' for the topological orbit equivalence of the underlying systems. Uniquely ergodic l.c.c.m. systems (with unique up to scaling infinite invariant Radon measure) were constructed in [37].

The second example is related to study of the smooth horocycle flows on tangent bundles of hyperbolic surfaces. Let $\mathbb{D}$ be the open disk equipped with the hyperbolic metric $|\mathrm{d}z|/(1-|z|^2)$ and let Möb($\mathbb{D}$) denote the group of Möbius transformations of $\mathbb{D}$. A hyperbolic surface can be written in the form $M := \Gamma\backslash\text{Möb}(\mathbb{D})$, where Γ is a torsion free discrete subgroup of Möb($\mathbb{D}$). Suppose that Γ is a nontrivial normal subgroup of a lattice Γ_0 in Möb($\mathbb{D}$). Then M is a regular cover of the finite volume surface $M_0 := \Gamma_0\backslash\text{Möb}(\mathbb{D})$. The group of deck transformations $G = \Gamma_0/\Gamma$ is finitely generated. The horocycle flow $(h_t)_{t\in\mathbb{R}}$ and the geodesic flow $(g_t)_{t\in\mathbb{R}}$ defined on the unit tangent bundle $T^1(\mathbb{D})$ descend naturally to flows, say h and g, on $T^1(M)$. We consider the problem of classification of the h-invariant Radon measures on M. According to Ratner, h has no finite invariant measures on M if G is infinite (except for measures supported on closed orbits). However there are infinite invariant Radon measures, for instance the volume measure. In the case when G is free Abelian and Γ_0 is co-compact, every homomorphism $\varphi\colon G \to \mathbb{R}$ determines a unique up to scaling ergodic invariant Radon measure (e.i.r.m.) m on $T^1(M)$ such that $m \circ \mathrm{d}D = \exp(\varphi(D))m$ for all $D \in G$ [16] and every e.i.r.m. arises this way [172]. Moreover all these measures are quasi-invariant under g. In the general case, an interesting bijection is established in [131] between the e.i.r.m. which are quasi-invariant under g and the 'nontrivial minimal' positive eigenfunctions of the hyperbolic Laplacian on M.

Von Neumann Algebras

There is a fascinating and productive interplay between nonsingular ergodic theory and von Neumann algebras. The two theories alternately influenced development of each other. Let $(X, \mathcal{B}, \mu, T)$ be a nonsingular dynamical system. Given $\varphi \in L^\infty(X, \mu)$ and $j \in \mathbb{Z}$, we define operators A_φ and U_j on the Hilbert space $L^2(Z \times \mathbb{Z}, \mu \times \nu)$ by setting

$$(A_\varphi f)(x, i) := \varphi(T^i x) f(x, i)\,,$$
$$(U_j f)(x, i) := f(x, i - j)\,.$$

Then $U_j A_\varphi U_j^* = A_{\varphi\circ T^j}$. Denote by $\mathcal{M}$ the von Neumann algebra (i. e. the weak closure of the $*$-algebra) generated by A_φ, $\varphi \in L^\infty(X, \mu)$ and U_j, $j \in \mathbb{Z}$. If T is ergodic and aperiodic then $\mathcal{M}$ is a factor, i. e. $\mathcal{M} \cap \mathcal{M}' = \mathbb{C}1$, where $\mathcal{M}'$ denotes the algebra of bounded operators commuting with $\mathcal{M}$. It is called a *Krieger's* factor. Murray–von Neumann–Connes' type of $\mathcal{M}$ is exactly the Krieger's type of T. The flow of weights of $\mathcal{M}$ is isomorphic to the associated flow of T. Two Krieger's factors are isomorphic if and only if the underlying dynamical systems are orbit equivalent [129]. Moreover, a number of important problems in the theory of von Neumann algebras such as classification of subfactors, computation of the flow of weights and Connes' invariants, outer conjugacy for automorphisms, etc. are intimately related to the corresponding problems in nonsingular orbit theory. We refer to [42,66,73,74,89,142] for details.

Representations of CAR

Representations of canonical anticommutation relations (CAR) is one of the most elegant and useful chapters of mathematical physics, providing a natural language for many body quantum physics and quantum field theory. By a representation of CAR we mean a sequence of bounded linear operators $a_1, a_2, \ldots$ in a separable Hilbert space $\mathcal{K}$ such that $a_j a_k + a_k a_j = 0$ and $a_j a_k^* + a_k^* a_j = \delta_{j,k}$.

Consider $\{0, 1\}$ as a group with addition mod 2. Then $X = \{0, 1\}^{\mathbb{N}}$ is a compact Abelian group. Let $\Gamma := \{x = (x_1, x_2, \ldots)\colon \lim_{n\to\infty} x_n = 0\}$. Then Γ is a dense countable subgroup of X. It is generated by elements γ_k whose k-coordinate is 1 and the other ones

are 0. Γ acts on X by translations. Let μ be an ergodic Γ-quasi-invariant measure on X. Let $(C_k)_{k\geq 1}$ be Borel maps from X to the group of unitary operators in a Hilbert space $\mathcal{H}$ satisfying $C_k^*(x) = C_k(x+\delta_k)$, $C_k(x)C_l(x+\delta_l) = C_l(x)C_k(x+\delta_k)$, $k \neq l$ for a. a. x. In other words, $(C_k)_{k\geq 1}$ defines a cocycle of the Γ-action. We now put $\widetilde{\mathcal{H}} := L^2(X,\mu)\otimes\mathcal{H}$ and define operators a_k in $\widetilde{\mathcal{H}}$ by setting

$$(a_k f)(x) = (-1)^{x_1+\cdots+x_{k-1}}(1-x_k) \times C_k(x)\sqrt{\frac{\mathrm{d}\mu\circ\delta_k}{\mathrm{d}\mu}(x)}f(x+\delta_k)\,,$$

where $f\colon X\to\mathcal{H}$ is an element of $\widetilde{\mathcal{H}}$ and $x=(x_1,x_2,\ldots)\in X$. It is easy to verify that $a_1,a_2,\ldots$ is a representation of CAR. The converse was established in [72] and [77]: every factor-representation (this means that the von Neumann algebra generated by all a_k is a factor) of CAR can be represented as above for some ergodic measure μ, Hilbert space $\mathcal{H}$ and a Γ-cocycle $(C_k)_{k\geq 1}$. Moreover, using nonsingular ergodic theory Golodets [77] constructed for each $k=2,3,\ldots,\infty$, an irreducible representation of CAR such that $\dim\mathcal{H}=k$. This answered a question of Gårding and Wightman [72] who considered only the case $k=1$.

Unitary Representations of Locally Compact Groups

Nonsingular actions appear in a systematic way in the theory of unitary representations of groups. Let G be a locally compact second countable group and H a closed normal subgroup of G. Suppose that H is commutative (or, more generally, of type I, see [53]). Then the natural action of G by conjugation on H induces a Borel G-action, say α, on the dual space $\widehat{H}$ – the set of unitarily equivalent classes of irreducible unitary representations of H. If now $U=(U_g)_{g\in G}$ is a unitary representation of G in a separable Hilbert space then by applying Stone decomposition theorem to $U\restriction H$ one can deduce that α is nonsingular with respect to a measure μ of the 'maximal spectral type' for $U\restriction H$ on $\widehat{H}$. Moreover, if U is irreducible then α is ergodic. Whenever μ is fixed, we obtain a one-to-one correspondence between the set of cohomology classes of irreducible cocycles for α with values in the unitary group on a Hilbert space $\mathcal{H}$ and the subset of $\widehat{G}$ consisting of classes of those unitary representations V for which the measure associated to $V\restriction H$ is equivalent to μ. This correspondence is used in both directions. From information about cocycles we can deduce facts about representations and vise versa [53,118].

Concluding Remarks

While some of the results that we have cited for nonsingular $\mathbb{Z}$-actions extend to actions of locally compact Polish groups (or subclasses of Abelian or amenable ones), many natural questions remain open in the general setting. For instance: what is Rokhlin lemma, or the pointwise ergodic theorem, or the definition of entropy for nonsingular actions of general countable amenable groups? The theory of abstract nonsingular equivalence relations [66] or, more generally, nonsingular groupoids [160] and polymorphisms [184] is also a beautiful part of nonsingular ergodic theory that has nice applications: description of semifinite traces of AF-algebras, classification of factor representations of the infinite symmetric group [185], path groups [14], etc. Nonsingular ergodic theory is getting even more sophisticated when we pass from $\mathbb{Z}$-actions to noninvertible endomorphisms or, more generally, semigroup actions (see [3] and references therein). However, due to restrictions of space we do not consider these issues in our survey.

Bibliography

1. Aaronson J (1983) The eigenvalues of nonsingular transformations. Isr J Math 45:297–312
2. Aaronson J (1987) The intrinsic normalizing constants of transformations preserving infinite measures. J Analyse Math 49:239–270
3. Aaronson J (1997) An Introduction to Infinite Ergodic Theory. Amer Math Soc, Providence
4. Aaronson J, Lemańczyk M (2005) Exactness of Rokhlin endomorphisms and weak mixing of Poisson boundaries, Algebraic and Topological Dynamics. Contemporary Mathematics, vol 385. Amer Math Soc, Providence 77–88
5. Aaronson J, Lin M, Weiss B (1979) Mixing properties of Markov operators and ergodic transformations, and ergodicity of cartesian products. Isr J Math 33:198–224
6. Aaronson J, Nadkarni M (1987) L_∞ eigenvalues and L_2 spectra of nonsingular transformations. Proc Lond Math Soc 55(3):538–570
7. Aaronson J, Nakada H (2000) Multiple recurrence of Markov shifts and other infinite measure preserving transformations. Isr J Math 117:285–310
8. Aaronson J, Park KK (2008) Predictability, entropy and information of infinite transformations, preprint. ArXiv:0705:2148v3
9. Aaronson J, Weiss B (2004) On Herman's theorem for ergodic, amenable group extensions of endomorphisms. Ergod Theory Dynam Syst 5:1283–1293
10. Adams S, Elliott GA, Giordano T (1994) Amenable actions of groups. Trans Amer Math Soc 344:803–822
11. Adams T, Friedman N, Silva CE (1997) Rank-One Weak Mixing for Nonsingular Transformations. Isr J Math 102:269–281
12. Adams T, Friedman N, Silva CE (2001) Rank one power weak mixing for nonsingular transformations. Ergod Theory Dynam Systems 21:1321–1332

13. Ageev ON, Silva CE (2002) Genericity of rigid and multiply recurrent infinite measure-preserving and nonsingular transformations. In: Proceedings of the 16th Summer Conference on General Topology and its Applications. Topology Proc 26(2):357–365
14. Albeverio S, Hoegh-Krohn R, Testard D, Vershik AM (1983) Factorial representations of Path groups. J Funct Anal 51: 115–231
15. Atkinson G (1976) Recurrence of co-cycles and random walks. J Lond Math Soc 13:486–488
16. Babillot M, Ledrappier F (1998) Geodesic paths and horocycle flow on abelian covers. In: Lie groups and ergodic theory. Tata Inst Fund Res Stud Math 14, Tata Inst Fund Res, Bombay, pp 1–32
17. Bergelson V, Leibman A (1996) Polynomial extensions of van der Waerden's and Semerédi's theorems. J Amer Math Soc 9:725–753
18. Bezuglyi SI, Golodets VY (1985) Groups of measure space transformations and invariants of outer conjugation for automorphisms from normalizers of type III full groups. J Funct Anal 60(3):341–369
19. Bezuglyi SI, Golodets VY (1991) Weak equivalence and the structures of cocycles of an ergodic automorphism. Publ Res Inst Math Sci 27(4):577–625
20. Bowles A, Fidkowski L, Marinello A, Silva CE (2001) Double ergodicity of nonsingular transformations and infinite measure-preserving staircase transformations. Ill J Math 45(3):999–1019
21. Chacon RV, Friedman NA (1965) Approximation and invariant measures. Z Wahrscheinlichkeitstheorie Verw Gebiete 3: 286–295
22. Choksi JR, Eigen S, Prasad V (1989) Ergodic theory on homogeneous measure algebras revisited. In: Mauldin RD, Shortt RM, Silva CE (eds) Measure and measurable dynamics. Contemp Math 94, Amer Math Soc, Providence, pp 73–85
23. Choksi JR, Hawkins JM, Prasad VS (1987) Abelian cocylces for nonsingular ergodic transformations and the genericity of type III_1 transformations. Monat fur Math 103:187–205
24. Choksi JR, Kakutani S (1979) Residuality of ergodic measurable transformations and of ergodic transformations which preserve an infinite measure. Ind Univ Math J 28:453–469
25. Choksi JR, Nadkarni MG (1994) The maximal spectral type of a rank one transformation. Canad Math Bull 37(1):29–36
26. Choksi JR, Nadkarni MG (2000) Genericity of nonsingular transformations with infinite ergodic index. Colloq Math 84/85:195–201
27. Choksi JR, Prasad VS (1983) Approximation and Baire category theorems in ergodic theory. In: Belley JM, Dubois J, Morales P (eds) Measure theory and its applications. Lect Notes Math 1033. Springer, Berlin, pp 94–113
28. Connes A (1975) On the hierarchy of W Krieger. Ill J Math 19:428–432
29. Connes A, Feldman J, Weiss B (1981) An amenable equivalence relation is generated by a single transformation. Ergod Theory Dynam Systems 1:431–450
30. Connes A, Krieger W (1977) Measure space automorphisms, the normalizers of their full groups, and approximate finiteness. J Funct Anal 24(4):336–352
31. Connes A, Woods EJ (1985) Approximately transitive flows and ITPFI factors. Ergod Theory Dynam Syst 5(2):203–236
32. Connes A, Woods EJ (1989) Hyperfinite von Neumann algebras and Poisson boundaries of time dependent random walks. Pac J Math 37:225–243
33. Cornfeld IP, Fomin VS, Sinaï YG (1982) Ergodic theory. Grundlehren der Mathematischen Wissenschaften, vol 245. Springer, New York
34. Danilenko AI (1995) The topological structure of Polish groups and groupoids of measure space transformations. Publ Res Inst Math Sci 31(5):913–940
35. Danilenko AI (1998) Quasinormal subrelations of ergodic equivalence relations. Proc Amer Math Soc 126(11): 3361–3370
36. Danilenko AI (2001) Funny rank one weak mixing for nonsingular Abelian actions. Isr J Math 121:29–54
37. Danilenko AI (2001) Strong orbit equivalence of locally compact Cantor minimal systems. Int J Math 12:113–123
38. Danilenko AI (2004) Infinite rank one actions and nonsingular Chacon transformations. Ill J Math 48(3):769–786
39. Danilenko AI (2007) On simplicity concepts for ergodic actions. J d'Anal Math 102:77–118
40. Danilenko AI (2007) (C, F)-actions in ergodic theory. In: Kapranov M, Kolyada S, Manin YI, Moree P, Potyagailo L (eds) Geometry and Dynamics of Groups and Spaces. Progr Math 265:325–351
41. Danilenko AI, Golodets YV (1996) On extension of cocycles to normalizer elements, outer conjugacy, and related problems. Trans Amer Math Soc 348(12):4857–4882
42. Danilenko AI, Hamachi T (2000) On measure theoretical analogues of the Takesaki structure theorem for type III factors. Colloq Math 84/85:485–493
43. Danilenko AI, Lemańczyk M (2005) A class of multipliers for $\mathcal{W}^{\perp}$. Isr J Math 148:137–168
44. Danilenko AI, Rudolph DJ: Conditional entropy theory in infinite measure and a question of Krengel. Isr J Math, to appear
45. Danilenko AI, Silva CE (2004) Multiple and polynomial recurrence for Abelian actions in infinite measure. J Lond Math Soc 2 69(1):183–200
46. Danilenko AI, Solomko AV: Infinite measure preserving flows with infinite ergodic index. Colloq Math, to appear
47. Day S, Grivna B, McCartney E, Silva CE (1999) Power Weakly Mixing Infinite Transformations. N Y J Math 5:17–24
48. del Junco A (1978) A simple measure-preserving transformation with trivial centralizer. Pac J Math 79:357–362
49. del Junco A, Rudolph DJ (1987) On ergodic actions whose self-joinings are graphs. Ergod Theory Dynam Syst 7:531–557
50. del Junco A, Silva CE (1995) Prime type III_λ automorphisms: An instance of coding techniques applied to nonsingular maps. In: Takahashi Y (ed) Fractals and Dynamics. Plenum, New York, pp 101–115
51. del Junco A, Silva CE (2003) On factors of nonsingular Cartesian products. Ergod Theory Dynam Syst 23(5):1445–1465
52. Derriennic Y, Frączek K, Lemańczyk M, Parreau F (2008) Ergodic automorphisms whose weak closure of off-diagonal measures consists of ergodic self-joinings. Colloq Math 110:81–115
53. Dixmier J (1969) Les C^*-algèbres et leurs représentations. Gauthier-Villars Editeur, Paris
54. Dooley AH, Hamachi T (2003) Nonsingular dynamical systems, Bratteli diagrams and Markov odometers. Isr J Math 138:93–123

55. Dooley AH, Hamachi T (2003) Markov odometer actions not of product type. Ergod Theory Dynam Syst 23(3):813–829
56. Dooley AH, Klemes I, Quas AN (1998) Product and Markov measures of type III. J Aust Math Soc Ser A 65(1):84–110
57. Dooley AH, Mortiss G: On the critical dimension of product odometers, preprint
58. Dooley AH, Mortiss G (2006) On the critical dimension and AC entropy for Markov odometers. Monatsh Math 149:193–213
59. Dooley AH, Mortiss G (2007) The critical dimensions of Hamachi shifts. Tohoku Math J 59(2):57–66
60. Dye H (1963) On groups of measure-preserving transformations I. Amer J Math 81:119–159, and II, Amer J Math 85: 551–576
61. Effros EG (1965) Transformation groups and C^*-algebras. Ann Math 81(2):38–55
62. Eigen SJ (1981) On the simplicity of the full group of ergodic transformations. Isr J Math 40(3–4):345–349
63. Eigen SJ (1982) The group of measure preserving transformations of [0,1] has no outer automorphisms. Math Ann 259:259–270
64. Eigen S, Hajian A, Halverson K (1998) Multiple recurrence and infinite measure preserving odometers. Isr J Math 108:37–44
65. Fedorov A (1985) Krieger's theorem for cocycles, preprint
66. Feldman J, Moore CC (1977) Ergodic equivalence relations, cohomology, and von Neumann algebras. I. Trans Amer Math Soc 234:289–324
67. Ferenczi S (1985) Systèmes de rang un gauche. Ann Inst H Poincaré Probab Statist 21(2):177–186
68. Friedman NA (1970) Introduction to Ergodic Theory. Van Nostrand Reinhold Mathematical Studies, No 29. Van Nostrand Reinhold Co., New York
69. Furstenberg H (1967) Disjointness in ergodic theory, minimal sets and diophantine approximation. Math Syst Theory 1: 1–49
70. Furstenberg H (1981) Recurrence in Ergodic Theory and Combinatorial Number Theory. Princeton University Press, Princeton
71. Furstenberg H, Weiss B (1978) The finite multipliers of infinite ergodic transformations, The structure of attractors in dynamical systems. In: Markley NG, Martin JC, Perrizo W (eds) Lecture Notes in Math 668. Springer, Berlin, pp 127–132
72. Gårding L, Wightman AS (1954) Representation of anticommutation relations. Proc Nat Acad Sci USA 40:617–621
73. Giordano T, Skandalis G (1985) Krieger factors isomorphic to their tensor square and pure point spectrum flows. J Funct Anal 64(2):209–226
74. Giordano T, Skandalis G (1985) On infinite tensor products of factors of type I_2. Ergod Theory Dynam Syst 5:565–586
75. Glasner E (1994) On the multipliers of $\mathcal{W}^{\perp}$. Ergod Theory Dynam Syst 14:129–140
76. Glimm J (1961) Locally compact transformation groups. Trans Amer Math Soc 101:124–138
77. Golodets YV (1969) A description of the representations of anticommutation relations. Uspehi Matemat Nauk 24(4):43–64
78. Golodets YV, Sinel'shchikov SD (1983) Existence and uniqueness of cocycles of ergodic automorphism with dense range in amenable groups. Preprint FTINT AN USSR, pp 19–83
79. Golodets YV, Sinel'shchikov SD (1985) Locally compact groups appearing as ranges of cocycles of ergodic Z-actions. Ergod Theory Dynam Syst 5:47–57
80. Golodets YV, Sinel'shchikov SD (1990) Amenable ergodic actions of groups and images of cocycles. Dokl Akad Nauk SSSR 312(6):1296–1299, in Russian
81. Golodets YV, Sinel'shchikov SD (1994) Classification and structure of cocycles of amenable ergodic equivalence relations. J Funct Anal 121(2):455–485
82. Gruher K, Hines F, Patel D, Silva CE, Waelder R (2003) Power weak mixing does not imply multiple recurrence in infinite measure and other counterexamples. N Y J Math 9:1–22
83. Hajian AB, Kakutani S (1964) Weakly wandering sets and invariant measures. Trans Amer Math Soc 110:136–151
84. Halmos PR (1946) An ergodic theorem. Proc Nat Acad Sci USA 32:156–161
85. Halmos PR (1956) Lectures on ergodic theory. Publ Math Soc Jpn 3
86. Hamachi T (1981) The normalizer group of an ergodic automorphism of type III and the commutant of an ergodic flow. J Funct Anal 40:387–403
87. Hamachi T (1981) On a Bernoulli shift with nonidentical factor measures. Ergod Theory Dynam Syst 1:273–283
88. Hamachi T (1992) A measure theoretical proof of the Connes–Woods theorem on AT-flows. Pac J Math 154:67–85
89. Hamachi T, Kosaki H (1993) Orbital factor map. Ergod Theory Dynam Syst 13:515–532
90. Hamachi T, Osikawa M (1981) Ergodic groups of automorphisms and Krieger's theorems. Semin Math Sci 3, Keio Univ
91. Hamachi T, Osikawa M (1986) Computation of the associated flows of ITPFI_2 factors of type III_0. In: Geometric methods in operator algebras. Pitman Res Notes Math Ser 123, Longman Sci Tech, Harlow, pp 196–210
92. Hamachi T, Silva CE (2000) On nonsingular Chacon transformations. Ill J Math 44:868–883
93. Hawkins JM (1982) Non-ITPFI diffeomorphisms. Isr J Math 42:117–131
94. Hawkins JM (1983) Smooth type III diffeomorphisms of manifolds. Trans Amer Math Soc 276:625–643
95. Hawkins JM (1990) Diffeomorphisms of manifolds with nonsingular Poincaré flows. J Math Anal Appl 145(2):419–430
96. Hawkins JM (1990) Properties of ergodic flows associated to product odometers. Pac J Math 141:287–294
97. Hawkins J, Schmidt K (1982) On C^2-diffeomorphisms of the circle which are of type III_1. Invent Math 66(3):511–518
98. Hawkins J, Silva CE (1997) Characterizing mildly mixing actions by orbit equivalence of products. In: Proceedings of the New York Journal of Mathematics Conference, June 9–13 1997. N Y J Math 3A:99–115
99. Hawkins J, Woods EJ (1984) Approximately transitive diffeomorphisms of the circle. Proc Amer Math Soc 90(2):258–262
100. Herman M (1979) Construction de difféomorphismes ergodiques, preprint
101. Herman M-R (1979) Sur la conjugaison differentiable des diffeomorphismes du cercle a des rotations. Inst Hautes Etudes Sci Publ Math 49:5–233, in French
102. Herman RH, Putnam IF, Skau CF (1992) Ordered Bratteli diagrams, dimension groups and topological dynamics. Int J Math 3(6):827–864
103. Host B, Méla J-F, Parreau F (1986) Analyse harmonique des mesures. Astérisque 135–136:1–261
104. Host B, Méla J-F, Parreau F (1991) Nonsingular transformations and spectral analysis of measures. Bull Soc Math France 119:33–90

105. Hurewicz W (1944) Ergodic theorem without invariant measure. Ann Math 45:192–206
106. Inoue K (2004) Isometric extensions and multiple recurrence of infinite measure preserving systems. Isr J Math 140: 245–252
107. Ionescu Tulcea A (1965) On the category of certain classes of transformations in ergodic theory. Trans Amer Math Soc 114:261–279
108. Ismagilov RS (1987) Application of a group algebra to problems on the tail σ-algebra of a random walk on a group and to problems on the ergodicity of a skew action. Izv Akad Nauk SSSR Ser Mat 51(4):893–907
109. Jaworsky W (1994) Strongly approximatevely transitive actions, the Choquet–Deny theorem, and polynomial growth. Pac J Math 165:115–129
110. James J, Koberda T, Lindsey K, Silva CE, Speh P (2008) Measurable sensitivity. Proc Amer Math Soc 136(10):3549–3559
111. Janvresse E, Meyerovitch T, de la Rue T, Roy E: Poisson suspensions and entropy of infinite transformations, preprint
112. Kaimanovich VA, Vershik AM (1983) Random walks on groups: boundary and entropy. Ann Probab 11:457–490
113. Kakutani S, Parry W (1963) Infinite measure preserving transformations with "mixing". Bull Amer Math Soc 69:752–756
114. Katznelson Y (1977) Sigma-finite invariant measures for smooth mappings of the circle. J Anal Math 31:1–18
115. Katznelson Y (1979) The action of diffeomorphism of the circle on the Lebesgue measure. J Anal Math 36:156–166
116. Katznelson Y, Weiss B (1972) The construction of quasi-invariant measures. Isr J Math 12:1–4
117. Katznelson Y, Weiss B (1991) The classification of nonsingular actions, revisited. Ergod Theory Dynam Syst 11:333–348
118. Kirillov AA (1978) Elements of the theory of representations. Nauka, Moscow
119. Krengel U (1967) Entropy of conservative transformations. Z Wahrscheinlichkeitstheorie Verw Gebiete 7:161–181
120. Krengel U (1969) Darstellungssätze für Strömungen und Halbströmungen, vol II. Math Ann 182:1–39
121. Krengel U (1970) Transformations without finite invariant measure have finite strong generators. In: Contributions to Ergodic Theory and Probability. Proc Conf Ohio State Univ, Columbus, Ohio. Springer, Berlin, pp 133–157
122. Krengel U (1976) On Rudolph's representation of aperiodic flows. Ann Inst H Poincaré Sect B (NS) 12(4):319–338
123. Krengel U (1985) Ergodic Theorems. De Gruyter Studies in Mathematics, Berlin
124. Krengel U, Sucheston L (1969) On mixing in infinite measure spaces. Z Wahrscheinlichkeitstheorie Verw Gebiete 13: 150–164
125. Krieger W (1969) On nonsingular transformations of a measure space, vol I, II. Z Wahrscheinlichkeitstheorie Verw Gebiete 11:83–119
126. Krieger W (1970) On the Araki–Woods asymptotic ratio set and nonsingular transformations of a measure space. In: Contributions to Ergodic Theory and Probability. Proc Conf Ohio State Univ, Columbus, Ohio. In: Lecture Notes in Math, vol 160. Springer, Berlin, pp 158–177
127. Krieger W (1972) On the infinite product construction of nonsingular transformations of a measure space. Invent Math 15:144–163; Erratum in 26:323–328
128. Krieger W (1976) On Borel automorphisms and their quasi-invariant measures. Math Z 151:19–24
129. Krieger W (1976) On ergodic flows and isomorphism of factors. Math Ann 223:19–70
130. Kubo I (1969) Quasi-flows. Nagoya Math J 35:1–30
131. Ledrappier F, Sarig O (2007) Invariant measures for the horocycle flow on periodic hyperbolic surfaces. Isr J Math 160:281–315
132. Lehrer E, Weiss B (1982) An ε-free Rokhlin lemma. Ergod Theory Dynam Syst 2:45–48
133. Lemańczyk M, Parreau F (2003) Rokhlin extensions and lifting disjointness. Ergod Theory Dynam Syst 23:1525–1550
134. Lemańczyk M, Parreau F, Thouvenot J-P (2000) Gaussian automorphisms whose ergodic self-joinings are Gaussian. Fund Math 164:253–293
135. Mackey GW (1966) Ergodic theory and virtual group. Math Ann 166:187–207
136. Maharam D (1964) Incompressible transformations. Fund Math LVI:35–50
137. Mandrekar V, Nadkarni M (1969) On ergodic quasi-invariant measures on the circle group. J Funct Anal 3:157–163
138. Matui H (2002) Topological orbit equivalence of locally compact Cantor minimal systems. Ergod Theory Dynam Syst 22:1871–1903
139. Méla J-F (1983) Groupes de valeurs propres des systèmes dynamiques et sous-groupes saturés du cercle. CR Acad Sci Paris Sér I Math 296(10):419–422
140. Meyerovitch T (2007) Extensions and Multiple Recurrence of infinite measure preserving systems, preprint. ArXiv: http://arxiv.org/abs/math/0703914
141. Moore CC (1967) Invariant measures on product spaces. Proc Fifth Berkeley Symp. University of California Press, Berkeley, pp 447–459
142. Moore CC (1982) Ergodic theory and von Neumann algebras. Proc Symp Pure Math 38:179–226
143. Moore CC, Schmidt K (1980) Coboundaries and homomorphisms for nonsingular actions and a problem of H Helson. Proc Lond Math Soc 3 40:443–475
144. Mortiss G (2000) A non-singular inverse Vitali lemma with applications. Ergod Theory Dynam Syst 20:1215–1229
145. Mortiss G (2002) Average co-ordinate entropy. J Aust Math Soc 73:171–186
146. Mortiss G (2003) An invariant for nonsingular isomorphism. Ergod Theory Dynam Syst 23:885–893
147. Nadkarni MG (1979) On spectra of nonsingular transformations and flows. Sankhya Ser A 41(1–2):59–66
148. Nadkarni MG (1998) Spectral theory of dynamical systems. In: Birkhäuser Advanced Texts: Basler Lehrbücher. Birkhäuser, Basel
149. Ornstein D (1960) On invariant measures. Bull Amer Math Soc 66:297–300
150. Ornstein D (1972) On the Root Problem in Ergodic Theory. In: Proc Sixth Berkeley Symp Math Stat Probab. University of California Press, Berkley, pp 347–356
151. Osikawa M (1977/78) Point spectra of nonsingular flows. Publ Res Inst Math Sci 13:167–172
152. Osikawa M (1988) Ergodic properties of product type odometers. Springer Lect Notes Math 1299:404–414
153. Osikawa M, Hamachi T (1971) On zero type and positive type transformations with infinite invariant measures. Mem Fac Sci Kyushu Univ 25:280–295
154. Parreau F, Roy E: Poisson joinings of Poisson suspensions, preprint

155. Parry W (1963) An ergodic theorem of information theory without invariant measure. Proc Lond Math Soc 3 13:605–612
156. Parry W (1965) Ergodic and spectral analysis of certain infinite measure preserving transformations. Proc Amer Math Soc 16:960–966
157. Parry W (1966) Generators and strong generators in ergodic theory. Bull Amer Math Soc 72:294–296
158. Parry W (1969) Entropy and generators in ergodic theory. WA Benjamin, New York, Amsterdam
159. Parthasarathy KR, Schmidt K (1977) On the cohomology of a hyperfinite action. Monatsh Math 84(1):37–48
160. Ramsay A (1971) Virtual groups and group actions. Adv Math 6:243–322
161. Rokhlin VA (1949) Selected topics from the metric theory of dynamical systems. Uspekhi Mat Nauk 4:57–125
162. Rokhlin VA (1965) Generators in ergodic theory, vol II. Vestnik Leningrad Univ 20(13):68–72, in Russian, English summary
163. Rosinsky J (1995) On the structure of stationary stable processes. Ann Probab 23:1163–1187
164. Roy E (2005) Mesures de Poisson, infinie divisibilité et propriétés ergodiques. Thèse de doctorat de l'Université Paris 6
165. Roy E (2007) Ergodic properties of Poissonian ID processes. Ann Probab 35:551–576
166. Roy E: Poisson suspensions and infinite ergodic theory, preprint
167. Rudolph DJ (1985) Restricted orbit equivalence. Mem Amer Math Soc 54(323)
168. Rudolph D, Silva CE (1989) Minimal self-joinings for nonsingular transformations. Ergod Theory Dynam Syst 9:759–800
169. Ryzhikov VV (1994) Factorization of an automorphism of a full Boolean algebra into the product of three involutions. Mat Zametki 54(2):79–84,159; in Russian. Translation in: Math Notes 54(1–2):821–824
170. Sachdeva U (1971) On category of mixing in infinite measure spaces. Math Syst Theory 5:319–330
171. Samorodnitsky G (2005) Null flows, positive flows and the structure of stationary symmetric stable processes. Ann Probab 33:1782–1803
172. Sarig O (2004) Invariant measures for the horocycle flows on Abelian covers. Invent Math 157:519–551
173. Schmidt K (1977) Cocycles on ergodic transformation groups. Macmillan Lectures in Mathematics, vol 1. Macmillan Company of India, Delhi
174. Schmidt K (1977) Infinite invariant measures in the circle. Symp Math 21:37–43
175. Schmidt K (1982) Spectra of ergodic group actions. Isr J Math 41(1–2):151–153
176. Schmidt K (1984) On recurrence. Z Wahrscheinlichkeitstheorie Verw Gebiete 68:75–95
177. Schmidt K, Walters P (1982) Mildly mixing actions of locally compact groups. Proc Lond Math Soc 45:506–518
178. Shelah S, Weiss B (1982) Measurable recurrence and quasi-invariant measures. Isr Math J 43:154–160
179. Silva CE, Thieullen P (1991) The subadditive ergodic theorem and recurrence properties of Markovian transformations. J Math Anal Appl 154(1):83–99
180. Silva CE, Thieullen P (1995) A skew product entropy for nonsingular transformations. J Lond Math Soc 2 52:497–516
181. Silva CE, Witte D (1992) On quotients of nonsingular actions whose self-joinings are graphs. Int J Math 5:219–237
182. Thouvenot J-P (1995) Some properties and applications of joinings in ergodic theory. In: Ergodic theory and its connections with harmonic analysis (Alexandia, 1993), pp 207–235. Lond Math Soc Lect Notes Ser 205. Cambridge Univ Press, Cambridge
183. Ullman D (1987) A generalization of a theorem of Atkinson to non-invariant measures. Pac J Math 130:187–193
184. Vershik AM (1983) Manyvalued mappings with invariant measure (polymorphisms) and Markov processes. J Sov Math 23:2243–2266
185. Vershik AM, Kerov SV (1985) Locally semisimple algebras. In: Combinatorial theory and K_0-functor. Mod Probl Math 26:3–56
186. Zimmer RJ (1977) Random walks on compact groups and the existence of cocycles. Isr J Math 26:84–90
187. Zimmer RJ (1978) Amenable ergodic group actions and an application to Poisson boundaries of random walks. J Funct Anal 27:350–372
188. Zimmer RJ (1984) Ergodic theory and semisimple Lie groups. Birkhäuser, Basel, Boston

Ergodic Theory: Recurrence

Nikos Frantzikinakis, Randall McCutcheon
Department of Mathematics, University of Memphis, Memphis, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Quantitative Poincaré Recurrence
Subsequence Recurrence
Multiple Recurrence
Connections with Combinatorics and Number Theory
Future Directions
Bibliography

Glossary

Almost every, essentially Given a Lebesgue measure space $(X, \mathcal{B}, \mu)$, a property $P(x)$ predicated of elements of X is said to hold for almost every $x \in X$, if the set $X \setminus \{x\colon P(x) \text{ holds}\}$ has zero measure. Two sets $A, B \in \mathcal{B}$ are essentially disjoint if $\mu(A \cap B) = 0$.

Conservative system Is an infinite measure preserving system such that for no set $A \in \mathcal{B}$ with positive measure are $A, T^{-1}A, T^{-2}A, \ldots$ pairwise essentially disjoint.

(c_n)-Conservative system If $(c_n)_{n\in\mathbb{N}}$ is a decreasing sequence of positive real numbers, a conservative

ergodic measure preserving transformation T is (c_n)-conservative if for some non-negative function $f \in L^1(\mu)$, $\sum_{n=1}^{\infty} c_n f(T^n x) = \infty$ a. e.

Doubling map If $\mathbb{T}$ is the interval $[0,1]$ with its endpoints identified and addition performed modulo 1, the (non-invertible) transformation $T\colon \mathbb{T} \to \mathbb{T}$, defined by $Tx = 2x \bmod 1$, preserves Lebesgue measure, hence induces a measure preserving system on $\mathbb{T}$.

Ergodic system Is a measure preserving system $(X, \mathcal{B}, \mu, T)$ (finite or infinite) such that every $A \in \mathcal{B}$ that is T-invariant (i. e. $T^{-1}A = A$) satisfies either $\mu(A) = 0$ or $\mu(X \setminus A) = 0$. (One can check that the rotation R_α is ergodic if and only if α is irrational, and that the doubling map is ergodic).

Ergodic decomposition Every measure preserving system $(X, \mathcal{X}, \mu, T)$ can be expressed as an integral of ergodic systems; for example, one can write $\mu = \int \mu_t \mathrm{d}\lambda(t)$, where λ is a probability measure on $[0,1]$ and μ_t are T-invariant probability measures on $(X, \mathcal{X})$ such that the systems $(X, \mathcal{X}, \mu_t, T)$ are ergodic for $t \in [0,1]$.

Ergodic theorem States that if $(X, \mathcal{B}, \mu, T)$ is a measure preserving system and $f \in L^2(\mu)$, then $\lim_{N\to\infty} \left\| \frac{1}{N} \sum_{n=1}^{N} T^n f - P_f \right\|_{L^2(\mu)} = 0$, where P_f denotes the orthogonal projection of the function f onto the subspace $\{f \in L^2(\mu)\colon Tf = f\}$.

Hausdorff a-measure Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system endowed with a μ-compatible metric d. The Hausdorff a-measure $\mathcal{H}_a(X)$ of X is an outer measure defined for all subsets of X as follows: First, for $A \subset X$ and $\varepsilon > 0$ let $\mathcal{H}_{a,\varepsilon}(A) = \inf\{\sum_{i=1}^{\infty} r_i^a\}$, where the infimum is taken over all countable coverings of A by sets $U_i \subset X$ with diameter $r_i < \varepsilon$. Then define $\mathcal{H}_a(A) = \limsup_{\varepsilon\to 0} \mathcal{H}_{a,\varepsilon}(A)$.

Infinite measure preserving system Same as measure preserving system, but $\mu(X) = \infty$.

Invertible system Is a measure preserving system $(X, \mathcal{B}, \mu, T)$ (finite or infinite), with the property that there exists $X_0 \in X$, with $\mu(X \setminus X_0) = 0$, and such that the transformation $T\colon X_0 \to X_0$ is bijective, with T^{-1} measurable.

Measure preserving system Is a quadruple $(X, \mathcal{B}, \mu, T)$, where X is a set, $\mathcal{B}$ is a σ-algebra of subsets of X (i. e. $\mathcal{B}$ is closed under countable unions and complementation), μ is a probability measure (i. e. a countably additive function from $\mathcal{B}$ to $[0,1]$ with $\mu(X) = 1$), and $T\colon X \to X$ is measurable (i. e. $T^{-1}A = \{x \in X : Tx \in A\} \in \mathcal{B}$ for $A \in \mathcal{B}$), and μ-preserving (i. e. $\mu(T^{-1}A) = \mu(A)$). Moreover, throughout the discussion we assume that the measure space $(X, \mathcal{B}, \mu)$ is Lebesgue (see Sect. 1.0 in [2]).

μ-Compatible metric Is a separable metric on X, where $(X, \mathcal{B}, \mu)$ is a probability space, having the property that open sets are measurable.

Positive definite sequence Is a complex-valued sequence $(a_n)_{n\in\mathbb{Z}}$ such that for any $n_1, \ldots, n_k \in \mathbb{Z}$ and $z_1, \ldots, z_k \in \mathbb{C}$, $\sum_{i,j=1}^{k} a_{n_i - n_j} z_i \overline{z_j} \geq 0$.

Rotations on $\mathbb{T}$ If $\mathbb{T}$ is the interval $[0,1]$ with its endpoints identified and addition performed modulo 1, then for every $\alpha \in \mathbb{R}$ the transformation $R_\alpha\colon \mathbb{T} \to \mathbb{T}$, defined by $R_\alpha x = x + \alpha \bmod 1$, preserves Lebesgue measure on $\mathbb{T}$ and hence induces a measure preserving system on $\mathbb{T}$.

Syndetic set Is a subset $E \subset \mathbb{Z}$ having bounded gaps. If G is a general discrete group, a set $E \subset G$ is syndetic if $G = FE$ for some finite set $F \subset G$.

Upper density Is the number $\overline{d}(\Lambda) = \limsup_{N\to\infty}(|\Lambda \cap \{-N, \ldots, N\}|)/(2N+1)$, where $\Lambda \subset \mathbb{Z}$ (assuming the limit to exist). Alternatively for measurable $E \subset \mathbb{R}^m$, $\overline{D}(E) = \limsup_{l(S)\to\infty}(m(S \cap E))/(m(S))$, where S ranges over all cubes in $\mathbb{R}^m$, and $l(S)$ denotes the length of the shortest edge of S.

Notation The following notation will be used throughout the article: $Tf = f \circ T$, $\{x\} = x - [x]$, D-$\lim_{n\to\infty}(a_n) = a \leftrightarrow \overline{d}\big(\{n : |a_n - a| > \varepsilon\}\big) = 0$ for every $\varepsilon > 0$.

Definition of the Subject

The basic principle that lies behind several recurrence phenomena is that the typical trajectory of a system with finite volume comes back infinitely often to any neighborhood of its initial point. This principle was first exploited by Poincaré in his 1890 King Oscar prize-winning memoir that studied planetary motion. Using the prototype of an ergodic-theoretic argument, he showed that in any system of point masses having fixed total energy that restricts its dynamics to bounded subsets of its phase space, the typical state of motion (characterized by configurations and velocities) must recur to an arbitrary degree of approximation.

Among the recurrence principle's more spectacularly counterintuitive ramifications is that isolated ideal gas systems that do not lose energy will return arbitrarily closely to their initial states, even when such a return entails a decrease in entropy from equilibrium, in apparent contradiction to the second law of thermodynamics. Such concerns, previously canvassed by Poincaré himself, were more infamously expounded by Zermelo [74] in 1896. Subsequent clarifications by Boltzmann, Maxwell and others led to an improved understanding of the second law's primarily statistical nature. (For an interesting historical/philosophical

discussion, see [68]; also [10]. For a probabilistic analysis of the likelihood of observing second law violations in small systems over short time intervals, see [28]).

These discoveries had a profound impact in dynamics, and the theory of measure preserving transformations (ergodic theory) evolved from these developments. Since then, the Poincaré recurrence principle has been applied to a variety of different fields in mathematics, physics, and information theory. In this article we survey the impact it has had in ergodic theory, especially as pertains to the field of *ergodic Ramsey theory*. (The heavy emphasis herein on the latter reflects authorial interest, and is not intended to transmit a proportionate image of the broader landscape of research relating to recurrence in ergodic theory.) Background information we assume in this article can be found in the books [35,63,71] (▶ Measure Preserving Systems).

Introduction

In this section we shall give several formulations of the Poincaré recurrence principle using the language of ergodic theory. Roughly speaking, the principle states that in a finite (or conservative) measure preserving system, every set of positive measure (or almost every point) comes back to itself infinitely many times under iteration. Despite the profound importance of these results, their proofs are extremely simple.

Theorem 1 (Poincaré Recurrence for Sets) *Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system and $A \in \mathcal{B}$ with $\mu(A) > 0$. Then $\mu(A \cap T^{-n}A) > 0$ for infinitely many $n \in \mathbb{N}$.*

Proof Since T is measure preserving, the sets $A, T^{-1}A, T^{-2}A, \ldots$ have the same measure. These sets cannot be pairwise essentially disjoint, since then the union of finitely many of them would have measure greater than $\mu(X) = 1$. Therefore, there exist $m, n \in \mathbb{N}$, with $n > m$, such that $\mu(T^{-m}A \cap T^{-n}A) > 0$. Again since T is measure preserving, we conclude that $\mu(A \cap T^{-k}A) > 0$, where $k = n - m > 0$. Repeating this argument for the iterates $A, T^{-m}A, T^{-2m}A, \ldots$, for all $m \in \mathbb{N}$, we easily deduce that $\mu(A \cap T^{-n}A) > 0$ for infinitely many $n \in \mathbb{N}$. □

We remark that the above argument actually shows that $\mu(A \cap T^{-n}A) > 0$ for some $n \leq [\frac{1}{\mu(A)}] + 1$.

Theorem 2 (Poincaré Recurrence for Points) *Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system and $A \in \mathcal{B}$. Then for almost every $x \in A$ we have that $T^n x \in A$ for infinitely many $n \in \mathbb{N}$.*

Proof Let B be the set of $x \in A$ such that $T^n x \notin A$ for all $n \in \mathbb{N}$. Notice that $B = A \setminus \bigcup_{n\in\mathbb{N}} T^{-n}A$; in particular, B is measurable. Since the iterates $B, T^{-1}B, T^{-2}B, \ldots$ are pairwise essentially disjoint, we conclude (as in the proof of Theorem 1) that $\mu(B) = 0$. This shows that for almost every $x \in A$ we have that $T^n x \in A$ for some $n \in \mathbb{N}$. Repeating this argument for the transformation T^m in place of T for all $m \in \mathbb{N}$, we easily deduce the advertised statement. □

Next we give a variation of Poincaré recurrence for measure preserving systems endowed with a compatible metric:

Theorem 3 (Poincaré Recurrence for Metric Systems) *Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system, and suppose that X is endowed with a μ-compatible metric. Then for almost every $x \in X$ we have*

$$\liminf_{n\to\infty} d(x, T^n x) = 0 .$$

The proof of this result is similar to the proof of Theorem 2 (see p. 61 in [35]). Applying this result to the doubling map $Tx = 2x$ on $\mathbb{T}$, we get that for almost every $x \in X$, every string of zeros and ones in the dyadic expansion of x occurs infinitely often.

We remark that all three formulations of the Poincaré Recurrence Theorem that we have given hold for conservative systems as well. See, e. g., [2] for details.

This article is structured as follows. In Sect. "Quantitative Poincaré Recurrence" we give a few quantitative versions of the previously mentioned qualitative results. In Sects. "Subsequence Recurrence" and "Multiple Recurrence" we give several refinements of the Poincaré recurrence theorem, by restricting the scope of the return time n, and by considering multiple intersections (for simplicity we focus on $\mathbb{Z}$-actions). In Sect. "Connections with Combinatorics and Number Theory" we give various implications of the recurrence results in combinatorics and number theory (▶ Ergodic Theory: Interactions with Combinatorics and Number Theory). Lastly, in Sect. "Future Directions" we give several open problems related to the material presented in Sects. "Subsequence Recurrence" to "Connections with Combinatorics and Number Theory".

Quantitative Poincaré Recurrence

Early Results

For applications it is desirable to have quantitative versions of the results mentioned in the previous section. For example one would like to know how large $\mu(A \cap T^{-n}A)$ can be made and for how many n.

Theorem 4 (Khintchine [55]) *Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system and $A \in \mathcal{B}$. Then for every $\varepsilon > 0$ we*

have $\mu(A \cap T^{-n}A) > \mu(A)^2 - \varepsilon$ for a set of $n \in \mathbb{N}$ that has bounded gaps.

By considering the doubling map $Tx = 2x$ on $\mathbb{T}$ and letting $A = \mathbf{1}_{[0,1/2)}$, it is easy to check that the lower bound of the previous result cannot be improved. We also remark that it is not possible to estimate the size of the gap by a function of $\mu(A)$ alone. One can see this by considering the rotations $R_k x = x + 1/k$ for $k \in \mathbb{N}$, defined on $\mathbb{T}$, and letting $A = \mathbf{1}_{[0,1/3]}$.

Concerning the second version of the Poincaré recurrence theorem, it is natural to ask whether for almost every $x \in X$ the set of return times $S_x = \{n \in \mathbb{N} : T^n x \in A\}$ has bounded gaps. This is not the case, as one can see by considering the doubling map $Tx = 2x$ on $\mathbb{T}$ with the Lebesgue measure, and letting $A = \mathbf{1}_{[0,1/2)}$. Since Lebesgue almost every $x \in \mathbb{T}$ contains arbitrarily large blocks of ones in its dyadic expansion, the set S_x has unbounded gaps. Nevertheless, as an easy consequence of the Birkhoff ergodic theorem [19], one has the following:

Theorem 5 *Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system and $A \in \mathcal{B}$ with $\mu(A) > 0$. Then for almost every $x \in X$ the set $S_x = \{n \in \mathbb{N} : T^n x \in A\}$ has well defined density and $\int \mathrm{d}(S_x)\, \mathrm{d}\mu(x) = \mu(A)$. Furthermore, for ergodic measure preserving systems we have $\mathrm{d}(S_x) = \mu(A)$ a. e.*

Another question that arises naturally is, given a set A with positive measure and an $x \in A$, how long should one wait until some iterate $T^n x$ of x hits A? By considering an irrational rotation R_α on $\mathbb{T}$, where α is very near to, but not less than, $\frac{1}{100}$, and letting $A = 1_{[0,1/2]}$, one can see that the first return time is a member of the set $\{1, 50, 51\}$. So it may come as a surprise that the average first return time does not depend on the system (as long as it is ergodic), but only on the measure of the set A.

Theorem 6 (Kac [51]) *Let $(X, \mathcal{B}, \mu, T)$ be an ergodic measure preserving system and $A \in \mathcal{B}$ with $\mu(A) > 0$. For $x \in X$ define $R_A(x) = \min\{n \in \mathbb{N} : T^n x \in A\}$. Then for $x \in A$ the expected value of $R_A(x)$ is $1/\mu(A)$, i. e. $\int_A R_A(x)\, \mathrm{d}\mu = 1$.*

More Recent Results

As we mentioned in the previous section, if the space X is endowed with a μ-compatible metric d, then for almost every $x \in X$ we have that $\liminf_{n\to\infty} d(x, T^n x) = 0$. A natural question is, how much iteration is needed to come back within a small distance of a given typical point? Under some additional hypothesis on the metric d we have the following answer:

Theorem 7 (Boshernitzan [20]) *Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system endowed with a μ-compatible metric d. Assume that the Hausdorff a-measure $\mathcal{H}_a(X)$ of X is σ-finite (i. e., X is a countable union of sets X_i with $\mathcal{H}_a(X_i) < \infty$). Then for almost every $x \in X$,*

$$\liminf_{n\to\infty} \left\{ n^{\frac{1}{a}} \cdot d(x, T^n x) \right\} < \infty \,.$$

Furthermore, if $\mathcal{H}_a(X) = 0$, then for almost every $x \in X$,

$$\liminf_{n\to\infty} \left\{ n^{\frac{1}{a}} \cdot d(x, T^n x) \right\} = 0 \,.$$

One can see from rotations by "badly approximable" vectors $\alpha \in \mathbb{T}^k$ that the exponent $1/a$ in the previous theorem cannot be improved. Several applications of Theorem 7 to billiard flows, dyadic transformations, symbolic flows and interval exchange transformations are given in [20]. For a related result dealing with mean values of the limits in Theorem 7 see [67].

An interesting connection between rates of recurrence and entropy of an ergodic measure preserving system was established by Ornstein and Weiss [62], following earlier work of Wyner and Ziv [73]:

Theorem 8 (Ornstein and Weiss [62]) *Let $(X, \mathcal{B}, \mu, T)$ be an ergodic measure preserving system and $\mathcal{P}$ be a finite partition of X. Let $P_n(x)$ be the element of the partition $\bigvee_{i=0}^{n-1} T^{-i}\mathcal{P} = \{\bigcap_{i=0}^{n-1} T^{-i} P^{(i)} : P^{(i)} \in \mathcal{P}, 0 \le i < n\}$ that contains x. Then for almost every $x \in X$, the first return time $R_n(x)$ of x to $P_n(x)$ is asymptotically equivalent to $e^{h(T,\mathcal{P})n}$, where $h(T, \mathcal{P})$ denotes the entropy of the system with respect to the partition $\mathcal{P}$. More precisely,*

$$\lim_{n\to\infty} \frac{\log R_n(x)}{n} = h(T, \mathcal{P}) \,.$$

An extension of the above result to some classes of infinite measure preserving systems was given in [42].

Another connection of recurrence rates, this time with the local dimension of an invariant measure, is given by the next result:

Theorem 9 (Barreira [4]) *Let $(X, \mathcal{B}, \mu, T)$ be an ergodic measure preserving system. Define the upper and lower recurrence rates*

$$\underline{R}(x) = \liminf_{r\to 0} \frac{\log \tau_r(x)}{-\log r} \quad \textit{and}$$
$$\overline{R}(x) = \limsup_{r\to 0} \frac{\log \tau_r(x)}{-\log r} \,,$$

where $\tau_r(x)$ is the first return time of $T^k x$ to $B(x,r)$, and the upper and lower pointwise dimensions

$$\underline{d}_\mu(x) = \liminf_{r\to 0} \frac{\log \mu(B(x,r))}{\log r} \quad \textit{and}$$
$$\overline{d}_\mu(x) = \limsup_{r\to 0} \frac{\log \mu(B(x,r))}{\log r} .$$

Then for almost every $x \in X$, we have

$$\underline{R}(x) \le \underline{d}_\mu(x) \quad \textit{and} \quad \overline{R}(x) \le \overline{d}_\mu(x) .$$

Roughly speaking, this theorem asserts that for typical $x \in X$ and for small r, the first return time of x to $B(x,r)$ is at most $r^{-d_\mu(x)}$. Since $\underline{d}_\mu(x) \le \mathcal{H}_a(X)$ for almost every $x \in X$, we can conclude the first part of Theorem 7 from Theorem 9. For related results the interested reader should consult the survey [5] and the bibliography therein.

We also remark that the previous results and related concepts have been applied to estimate the dimension of certain strange attractors (see [49] and the references therein) and the entropy of some Gibbsian systems [25].

We end this section with a result that connects "wandering rates" of sets in infinite measure preserving systems with their "recurrence rates". The next theorem follows easily from a result about lower bounds on ergodic averages for measure preserving systems due to Leibman [57]; a weaker form for conservative, ergodic systems can be found in Aaronson [1].

Theorem 10 *Let $(X, \mathcal{B}, \mu, T)$ be an infinite measure preserving system, and $A \in \mathcal{B}$ with $\mu(A) < \infty$. Then for all $N \in \mathbb{N}$,*

$$\left(\frac{\mu\left(\bigcup_{n=0}^{N-1} T^{-n} A\right)}{N} \cdot \sum_{n=0}^{N-1} \mu(A \cap T^{-n} A) \right) \ge \frac{1}{2} \cdot (\mu(A))^2 .$$

Subsequence Recurrence

In this section we discuss what restrictions one can impose on the set of return times in the various versions of the Poincaré recurrence theorem. We start with:

Definition 11 Let $R \subset \mathbb{Z}$. Then R is a set of:

(a) *Recurrence* if for any invertible measure preserving system $(X, \mathcal{B}, \mu, T)$, and $A \in \mathcal{B}$ with $\mu(A) > 0$, there is some nonzero $n \in R$ such that $\mu(A \cap T^{-n} A) > 0$.
(b) *Topological recurrence* if for every compact metric space (X, d), continuous transformation $T: X \to X$ and every $\varepsilon > 0$, there are $x \in X$ and nonzero $n \in R$ such that $d(x, T^n x) < \varepsilon$.

It is easy to check that the existence of a single $n \in R$ satisfying the previous recurrence conditions actually guarantees the existence of infinitely many $n \in R$ satisfying the same conditions. Moreover, if R is a set of recurrence then one can see from existence of some T-invariant measure μ that R is also a set of topological recurrence. A (complicated) example showing that the converse is not true was given by Kriz [56].

Before giving a list of examples of sets of (topological) recurrence, we discuss some necessary conditions: A set of topological recurrence must contain infinitely many multiples of every positive integer, as one can see by considering rotations on $\mathbb{Z}_d$, $d \in \mathbb{N}$. Hence, the sets $\{2n+1, n \in \mathbb{N}\}$, $\{n^2+1, n \in \mathbb{N}\}$, $\{p+2, p \text{ prime}\}$ are not good for (topological) recurrence. If $(s_n)_{n\in\mathbb{N}}$ is a lacunary sequence (meaning $\liminf_{n\to\infty}(s_{n+1}/s_n) = \rho > 1$), then one can construct an irrational number α such that $\{s_n\alpha\} \in [\delta, 1-\delta]$ for all large $n \in \mathbb{N}$, where $\delta > 0$ depends on ρ (see [54], for example). As a consequence, the sequence $(s_n)_{n\in\mathbb{N}}$ is not good for (topological) recurrence.

Lastly, we mention that by considering product systems, one can immediately show that any set of (topological) recurrence R is partition regular, meaning that if R is partitioned into finitely many pieces then at least one of these pieces must still be a set of (topological) recurrence. Using this observation, one concludes for example that any union of finitely many lacunary sequences is not a set of recurrence.

We present now some examples of sets of recurrence:

Theorem 12 *The following are sets of recurrence:*

(i) *Any set of the form $\bigcup_{n\in\mathbb{N}} \{a_n, 2a_n, \ldots, na_n\}$ where $a_n \in \mathbb{N}$.*
(ii) *Any IP-set, meaning a set that consists of all finite sums of the members of some infinite set.*
(iii) *Any difference set $S - S$, meaning a set that consists of all possible differences of the members of some some infinite set S.*
(iv) *The set $\{p(n), n \in \mathbb{N}\}$ where p is any nonconstant integer polynomial with $p(0) = 0$ [35,66] (In fact we only have to assume that the range of the polynomial contains multiples of an arbitrary positive integer [53]).*
(v) *The set $\{p(n), n \in S\}$, where p is an integer polynomial with $p(0) = 0$ and S is any IP-set [12].*
(vi) *The set of values of an admissible generalized polynomial (this class contains in particular the*

smallest function algebra G containing all integer polynomials having zero constant term and such that if $g_1, \ldots, g_k \in G$ and $c_1, \ldots, c_k \in \mathbb{R}$ then $[\![\sum_{i=1}^k c_i g_i]\!] \in G$, where $[\![x]\!] = [x + \frac{1}{2}]$ denotes the integer nearest to x) [13].

(vii) *The set of shifted primes $\{p - 1, p \text{ prime}\}$, and the set $\{p + 1, p \text{ prime}\}$* [66].

(viii) *The set of values of a random non-lacunary sequence. (Pick $n \in \mathbb{N}$ independently with probability b_n where $0 \le b_n \le 1$ and $\lim_{n\to\infty} nb_n = \infty$. The resulting set is almost surely a set of recurrence. If $\limsup_{n\to\infty} nb_n < \infty$ then the resulting set is almost surely a finite union of sets, each of which is the range of some lacunary sequence, hence is not a set of recurrence). Follows from* [22].

Showing that the first three sets are good for recurrence is a straightforward modification of the argument used to prove Theorem 1. Examples $(iv) - (viii)$ require more work.

A criterion of Kamae and Mendés-France [53] provides a powerful tool that may be used in many instances to establish that a set R is a set of recurrence. We mention a variation of their result:

Theorem 13 (Kamae and Mendés-France [53]) *Suppose that $R = \{a_1 < a_2 < \ldots\}$ is a subset of $\mathbb{N}$ such that:*

(i) *The sequence $\{a_n\alpha\}_{n\in\mathbb{N}}$ is uniformly distributed in $\mathbb{T}$ for every irrational α.*

(ii) *The set $R_m = \{n \in \mathbb{N} : m|a_n\}$ has positive density for every $m \in \mathbb{N}$.*

Then R is a set of recurrence.

We sketch a proof for this result. First, recall Herglotz's theorem: if $(a_n)_{n\in\mathbb{Z}}$ is a positive definite sequence, then there is a unique measure σ on the torus $\mathbb{T}$ such that $a_n = \int e^{2\pi i n t}\, d\sigma(t)$. The case of interest to us is $a_n = \int_{\mathbb{T}} f(x) \cdot f(T^n x)\, d\mu$, where T is measure preserving and $f \in L^\infty(\mu)$; (a_n) is positive definite, and we call $\sigma = \sigma_f$ the *spectral measure* of f.

Let now $(X, \mathcal{B}, \mu, T)$ be a measure preserving system and $A \in \mathcal{B}$ with $\mu(A) > 0$. Putting $f = \mathbf{1}_A$, one has

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^N \int f(x) \cdot f(T^{a_n} x)\, d\mu = \int_{\mathbb{T}} \lim_{N\to\infty} \left(\frac{1}{N} \sum_{n=1}^N e^{2\pi i a_n t} \right) d\sigma_f(t) . \quad (1)$$

For t irrational the limit inside the integral is zero (by condition (i)), so the last integral can be taken over the rational points in $\mathbb{T}$. Since the spectral measure of a function orthogonal to the subspace

$$\mathcal{H} = \overline{\{f \in L^2(\mu) : \text{ there exists } k \in \mathbb{N} \text{ with } T^k f = f\}} \quad (2)$$

has no rational point masses, we can easily deduce that when computing the first limit in (1), we can replace the function f by its orthogonal projection g onto the subspace $\mathcal{H}$ (g is again nonnegative and $g \neq 0$). To complete the argument, we approximate g by a function g' such that $T^m g' = g'$ for some appropriately chosen m, and use condition (ii) to deduce that the limit of the average (1) is positive.

In order to apply Theorem 13, one uses the standard machinery of uniform distribution. Recall *Weyl's criterion*: a real-valued sequence $(x_n)_{n\in\mathbb{N}}$ is uniformly distributed mod 1 if for every non-zero $k \in \mathbb{Z}$,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^N e^{2\pi i k x_n} = 0 .$$

This criterion becomes especially useful when paired with van der Corput's so-called third principal property: if, for every $h \in \mathbb{N}$, $(x_{n+h} - x_n)_{n\in\mathbb{N}}$ is uniformly distributed mod 1, then $(x_n)_{n\in\mathbb{N}}$ is uniformly distributed mod 1. Using the foregoing criteria and some standard (albeit nontrivial) exponential sum estimates, one can verify for example that the sets (iv) and (vii) in Theorem 12 are good for recurrence.

In light of the connection elucidated above between uniform distribution mod 1 and recurrence, it is not surprising that van der Corput's method has been adapted by modern ergodic theorists for use in establishing recurrence properties *directly*.

Theorem 14 (Bergelson [7]) *Let $(x_n)_{n\in\mathbb{N}}$ be a bounded sequence in a Hilbert space. If*

$$D\text{-}\lim_{m\to\infty} \left(\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^N \langle x_{n+m}, x_n \rangle \right) = 0 ,$$

then

$$\lim_{N\to\infty} \left\| \frac{1}{N} \sum_{n=1}^N x_n \right\| = 0 .$$

Let us illustrate how one uses this "van der Corput trick" by showing that $S = \{n^2 : n \in \mathbb{N}\}$ is a set of recurrence. We will actually establish the following stronger

fact: If $(X, \mathcal{B}, \mu, T)$ is a measure preserving system and $f \in L^\infty(\mu)$ is nonnegative and $f \neq 0$ then

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \int f(x) \cdot f(T^{n^2} x)\, \mathrm{d}\mu > 0\,. \tag{3}$$

Then our result follows by setting $f = \mathbf{1}_A$ for some $A \in \mathcal{B}$ with $\mu(A) > 0$.

The main idea is one that occurs frequently in ergodic theory; split the function f into two components, one of which contributes zero to the limit appearing in (3), and the other one being much easier to handle than f. To do this consider the T-invariant subspace of $L^2(X)$ defined by

$$\mathcal{H} = \overline{\{f \in L^2(\mu)\colon \text{ there exists } k \in \mathbb{N} \text{ with } T^k f = f\}}\,. \tag{4}$$

Write $f = g + h$ where $g \in \mathcal{H}$ and $h \perp \mathcal{H}$, and expand the average in (3) into a sum of four averages involving the functions g and h. Two of these averages vanish because iterates of g are orthogonal to iterates of h. So in order to show that the only contribution comes from the average that involves the function g alone, it suffices to establish that

$$\lim_{N\to\infty} \left\| \frac{1}{N} \sum_{n=1}^{N} T^{n^2} h \right\|_{L^2(\mu)} = 0\,. \tag{5}$$

To show this we will apply the Hilbert space van der Corput lemma. For given $h \in \mathbb{N}$, we let $x_n = T^{n^2} h$ and compute

$$\begin{aligned}
&\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \langle x_{n+m}, x_n \rangle \\
&\quad = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \int T^{n^2+2nm+m^2} h \cdot T^{n^2} h\, \mathrm{d}\mu \\
&\quad = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \int T^{2nm}(T^{m^2} h) \cdot h\, \mathrm{d}\mu\,.
\end{aligned}$$

Applying the ergodic theorem to the transformation T^{2m} and using the fact that $h \perp \mathcal{H}$, we get that the last limit is 0. This implies (5).

Thus far we have shown that in order to compute the limit in (3) we can assume that $f = g \in \mathcal{H}$ (g is also nonnegative and $g \neq 0$). By the definition of $\mathcal{H}$, given any $\varepsilon > 0$, there exists a function $f' \in \mathcal{H}$ such that $T^k f' = f'$ for some $k \in \mathbb{N}$ and $\|f - f'\|_{L^2(\mu)} \leq \varepsilon$. Then the limit in (3) is at least $1/k$ times the limit

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \int f(x) \cdot f(T^{(kn)^2} x)\, \mathrm{d}\mu.$$

Applying the triangle inequality twice we get that this is greater or equal than

$$\begin{aligned}
&\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \int f'(x) \cdot f'(T^{(kn)^2} x)\, \mathrm{d}\mu - c \cdot \varepsilon \\
&\quad = \int (f'(x))^2\, \mathrm{d}\mu - 2\varepsilon \\
&\quad \geq \left(\int f'(x)\, \mathrm{d}\mu \right)^2 - c \cdot \varepsilon,
\end{aligned}$$

for some constant c that does not depend on ε (we used that $T^k f' = f'$ and the Cauchy–Schwartz inequality). Choosing ε small enough we conclude that the last quantity is positive, completing the proof.

Multiple Recurrence

Simultaneous multiple returns of positive measure sets to themselves were first considered by H. Furstenberg [34], who gave a new proof of Szemerédi's theorem [69] on arithmetic progressions by deriving it from the following theorem:

Theorem 15 (Furstenberg [34]) *Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system and $A \in \mathcal{B}$ with $\mu(A) > 0$. Then for every $k \in \mathbb{N}$, there is some $n \in \mathbb{N}$ such that*

$$\mu(A \cap T^{-n} A \cap \cdots \cap T^{-kn} A) > 0\,. \tag{6}$$

Furstenberg's proof came by means of a new structure theorem allowing one to decompose an arbitrary measure preserving system into component elements exhibiting one of two extreme types of behavior: *compactness*, characterized by regular, "almost periodic" trajectories, and *weak mixing*, characterized by irregular, "quasi-random" trajectories. On $\mathbb{T}$, these types of behavior are exemplified by rotations and by the doubling map, respectively. To see the point, imagine trying to predict the initial digit of the dyadic expansion of $T^n x$ given knowledge of the initial digits of $T^i x$, $1 \leq i < n$. We use the case $k = 2$ to illustrate the basic idea.

It suffices to show that if $f \in L^\infty(\mu)$ is nonnegative and $f \neq 0$, one has

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \int f(x) \cdot f(T^n x) \cdot f(T^{2n} x)\, \mathrm{d}\mu > 0\,. \tag{7}$$

An ergodic decomposition argument enables us to assume that our system is ergodic. As in the earlier case of the squares, we split f into "almost periodic" and "quasi-random" components. Let $\mathcal{K}$ be the closure in L^2 of the subspace spanned by the eigenfunctions of T, i. e. the functions $f \in L^2(\mu)$ that satisfy $f(Tx) = e^{2\pi i \alpha} f(x)$ for some

$\alpha \in \mathbb{R}$. We write $f = g + h$, where $g \in \mathcal{K}$ and $h \perp \mathcal{K}$. It can be shown that $g, h \in L^\infty(\mu)$ and g is again nonnegative with $g \neq 0$. We expand the average in (7) into a sum of eight averages involving the functions g and h. In order to show that the only non-zero contribution to the limit comes from the term involving g alone, it suffices to establish that

$$\lim_{N\to\infty} \left\| \frac{1}{N} \sum_{n=1}^{N} T^n g \cdot T^{2n} h \right\|_{L^2(\mu)} = 0 \,, \tag{8}$$

(and similarly with h and g interchanged, and with $g = h$, which is similar). To establish (8), we use the Hilbert space van der Corput lemma on $x_n = T^n g \cdot T^{2n} h$. Some routine computations and a use of the ergodic theorem reduce the task to showing that

$$\text{D-}\lim_{m\to\infty} \left(\int h(x) \cdot h(T^{2m} x) \, \mathrm{d}\mu \right) = 0 \,.$$

But this is well known for $h \perp \mathcal{K}$ (in virtue of the fact that for $h \perp \mathcal{K}$ the spectral measure σ_h is continuous, for example).

We are left with the average (7) when $f = g \in \mathcal{K}$. In this case f can be approximated arbitrarily well by a linear combination of eigenfunctions, which easily implies that given $\varepsilon > 0$ one has $\|T^n f - f\|_{L^2(\mu)} \leq \varepsilon$ for a set of $n \in \mathbb{N}$ with bounded gaps. Using this fact and the triangle inequality, one finds that for a set of $n \in \mathbb{N}$ with bounded gaps,

$$\int f(x) \cdot f(T^n x) \cdot f(T^{2n} x) \, \mathrm{d}\mu \geq \left(\int f \, \mathrm{d}\mu \right)^3 - c \cdot \varepsilon$$

for a constant c that is independent of ε. Choosing ε small enough, we get (7).

The new techniques developed for the proof of Theorem 15 have led to a number of extensions, many of which have to date only ergodic proofs. To expedite discussion of some of these developments, we introduce a definition:

Definition 16 Let $R \subset \mathbb{Z}$ and $k \in \mathbb{N}$. Then R is a set of *k-recurrence* if for every invertible measure preserving system $(X, \mathcal{B}, \mu, T)$ and $A \in \mathcal{B}$ with $\mu(A) > 0$, there is some nonzero $n \in R$ such that

$$\mu(A \cap T^{-n} A \cap \cdots \cap T^{-kn} A) > 0 \,.$$

The notions of k-recurrence are distinct for different values of k. An example of a difference set that is a set of 1-recurrence but not a set of 2-recurrence was given in [34]; sets of k-recurrence that are not sets of $(k+1)$-recurrence for general k were given in [31] ($R_k = \{n \in \mathbb{N} : \{n^{k+1}\sqrt{2}\} \in [1/4, 3/4]\}$ is such).

Aside from difference sets, the sets of (1-)recurrence given in Theorem 12 may well be sets of k-recurrence for every $k \in \mathbb{N}$, though this has not been verified in all cases. Let us summarize the current state of knowledge. The following are sets of k-recurrence for every k: Sets of the form $\bigcup_{n\in\mathbb{N}} \{a_n, 2a_n, \ldots, na_n\}$ where $a_n \in \mathbb{N}$ (this follows from a uniform version of Theorem 15 that can be found in [15]). Every IP-set [37]. The set $\{p(n), n \in \mathbb{N}\}$ where p is any nonconstant integer polynomial with $p(0) = 0$ [16], and more generally, when the range of the polynomial contains multiples of an arbitrary integer [33]. The set $\{p(n), n \in S\}$ where p is an integer polynomial with $p(0) = 0$ and S is any IP-set [17]. The set of values of an admissible generalized polynomial [60]. Moreover, the set of shifted primes $\{p - 1, p \text{ prime}\}$, and the set $\{p + 1, p \text{ prime}\}$ are sets of 2-recurrence [32].

More generally, one would like to know for which sequences of integers $a_1(n), \ldots, a_k(n)$ it is the case that for every invertible measure preserving system $(X, \mathcal{B}, \mu, T)$ and $A \in \mathcal{B}$ with $\mu(A) > 0$, there is some nonzero $n \in \mathbb{N}$ such that

$$\mu(A \cap T^{-a_1(n)} A \cap \cdots \cap T^{-a_k(n)} A) > 0 \,. \tag{9}$$

Unfortunately, a criterion analogous to the one given in Theorem 13 for 1-recurrence is not yet available for k-recurrence when $k > 1$. Nevertheless, there have been some notable positive results, such as the following:

Theorem 17 (Bergelson and Leibman [16]) *Let $(X, \mathcal{B}, \mu, T)$ be an invertible measure preserving system and $p_1(n), \ldots, p_k(n)$ be integer polynomials with zero constant term. Then for every $A \in \mathcal{B}$ with $\mu(A) > 0$, there is some $n \in \mathbb{N}$ such that*

$$\mu(A \cap T^{-p_1(n)} A \cap \cdots \cap T^{-p_k(n)} A) > 0 \,. \tag{10}$$

Furthermore, it has been shown that the n in (10) can be chosen from any IP set [17], and the polynomials $p_1, \ldots, p_k$ can be chosen to belong to the more general class of admissible generalized polynomials [60].

Very recently, a new boost in the area of multiple recurrence was given by a breakthrough of Host and Kra [50]. Building on work of Conze and Lesigne [26,27] and Furstenberg and Weiss [41] (see also the excellent survey [52], exploring close parallels with [45] and the seminal paper of Gowers [43]), they isolated the structured component (or factor) of a measure preserving system that one needs to analyze in order to prove various multiple recurrence and convergence results. This allowed them, in particular, to prove existence of L^2 limits for the so-called "Furstenberg ergodic averages" $\frac{1}{N}\sum_{n=1}^{N} \prod_{i=0}^{k} f(T^{in} x)$,

which had been a major open problem since the original ergodic proof of Szemerédi's theorem. Subsequently Ziegler in [75] gave a new proof of the aforementioned limit theorem and established minimality of the factor in question. It turns out that this minimal component admits of a purely algebraic characterization; it is a *nilsystem*, i. e. a rotation on a homogeneous space of a nilpotent Lie group. This fact, coupled with some recent results about nilsystems (see [58,59] for example), makes the analysis of some otherwise intractable multiple recurrence problems much more manageable. For example, these developments have made it possible to estimate the size of the multiple intersection in (6) for $k = 2, 3$ (the case $k = 1$ is Theorem 4):

Theorem 18 (Bergelson, Host and Kra [14]) *Let $(X, \mathcal{B}, \mu, T)$ be an ergodic measure preserving system and $A \in \mathcal{B}$. Then for $k = 2, 3$ and for every $\varepsilon > 0$,*

$$\mu(A \cap T^{-n}A \cap \cdots \cap T^{-kn}A) > \mu^{k+1}(A) - \varepsilon \qquad (11)$$

for a set of $n \in \mathbb{N}$ with bounded gaps.

Based on work of Ruzsa that appears as an appendix to the paper, it is also shown in [14] that a similar estimate fails for ergodic systems (with any power of $\mu(A)$ on the right hand side) when $k \geq 4$. Moreover, when the system is nonergodic it also fails for $k = 2, 3$, as can be seen with the help of an example in [6]. Again considering the doubling map $Tx = 2x$ and the set $A = [0, 1/2]$, one sees that the positive results for $k \leq 3$ are sharp. When the polynomials $n, 2n, \ldots, kn$ are replaced by linearly independent polynomials $p_1, p_2, \ldots, p_k$ with zero constant term, similar lower bounds hold for every $k \in \mathbb{N}$ without assuming ergodicity [30]. The case where the polynomials $n, 2n, 3n$ are replaced with general polynomials p_1, p_2, p_3 with zero constant term is treated in [33].

Connections with Combinatorics and Number Theory

The combinatorial ramifications of ergodic-theoretic recurrence were first observed by Furstenberg, who perceived a correspondence between recurrence properties of measure preserving systems and the existence of structures in sets of integers having positive upper density. This gave rise to the field of ergodic Ramsey theory, in which problems in combinatorial number theory are treated using techniques from ergodic theory. The following formulation is from [8].

Theorem 19 *Let Λ be a subset of the integers. There exists an invertible measure preserving system $(X, \mathcal{B}, \mu, T)$ and a set $A \in \mathcal{B}$ with $\mu(A) = \overline{d}(\Lambda)$ such that*

$$\overline{d}(\Lambda \cap (\Lambda - n_1) \cap \ldots \cap (\Lambda - n_k)) \geq \mu(A \cap T^{-n_1}A \cap \cdots \cap T^{-n_k}A)\,, \qquad (12)$$

for all $k \in \mathbb{N}$ and $n_1, \ldots, n_k \in \mathbb{Z}$.

Proof The space X will be taken to be the sequence space $\{0, 1\}^{\mathbb{Z}}$, $\mathcal{B}$ is the Borel σ-algebra, while T is the shift map defined by $(Tx)(n) = x(n + 1)$ for $x \in \{0, 1\}^{\mathbb{Z}}$, and A is the set of sequences x with $x(0) = 1$. So the only thing that depends on Λ is the measure μ which we now define. For $m \in \mathbb{N}$ set $\Lambda^0 = \mathbb{Z} \setminus \Lambda$ and $\Lambda^1 = \Lambda$. Using a diagonal argument we can find an increasing sequence of integers $(N_m)_{m\in\mathbb{N}}$ such that $\lim_{m\to\infty} |\Lambda \cap [1, N_m]|/N_m = \overline{d}(\Lambda)$ and such that

$$\lim_{m\to\infty} \frac{\left| \begin{array}{c} (\Lambda^{i_1} - n_1) \cap (\Lambda^{i_2} - n_2) \cap \cdots \\ \cap (\Lambda^{i_r} - n_r) \cap [1, N_m] \end{array} \right|}{N_m} \qquad (13)$$

exists for every $n_1, \ldots, n_r \in \mathbb{Z}$, and $i_1, \ldots, i_r \in \{0, 1\}$. For $n_1, n_2, \ldots, n_r \in \mathbb{Z}$, and $i_1, i_2, \ldots, i_r \in \{0, 1\}$, we define the measure μ of the *cylinder set* $\{x(n_1) = i_1, x(n_2) = i_2, \ldots, x(n_r) = i_r\}$ to be the limit (13). Thus defined, μ extends to a premeasure on the algebra of sets generated by cylinder sets and hence by Carathéodory's extension theorem [24] to a probability measure on $\mathcal{B}$. It is easy to check that $\mu(A) = \overline{d}(\Lambda)$, the shift transformation T preserves the measure μ and (12) holds. □

Using this principle for $k = 1$, one may check that any set of recurrence is *intersective*, that is intersects $E - E$ for every set E of positive density. Using it for $n_1 = n, n_2 = 2n, \ldots, n_k = kn$, together with Theorem 15, one gets an ergodic proof of Szemerédi's theorem [69], stating that every subset of the integers with positive upper density contains arbitrarily long arithmetic progressions (conversely, one can easily deduce Theorem 15 from Szemerédi's theorem, and that intersective sets are sets of recurrence). Making the choice $n_1 = n^2$ and using part (iv) of Theorem 13, we get an ergodic proof of the surprising result of Sárközy [66] stating that every subset of the integers with positive upper density contains two elements whose difference is a perfect square. More generally, using Theorem 19, one can translate all of the recurrence results of the previous two sections to results in combinatorics. (This is not straightforward for Theorem 18 because of the ergodicity assumption made there. We refer the reader to [14] for the combinatorial consequence of this result). We mention explicitly only the combinatorial consequence of Theorem 17:

Theorem 20 (Bergelson and Leibman [16]) *Let $\Lambda \subset \mathbb{Z}$ with $\overline{d}(\Lambda) > 0$, and $p_1, \ldots, p_k$ be integer polynomials with zero constant term. Then Λ contains infinitely many configurations of the form $\{x, x + p_1(n), \ldots, x + p_k(n)\}$.*

The ergodic proof is the only one known for this result, even for patterns of the form $\{x, x + n^2, x + 2n^2\}$ or $\{x, x + n, x + n^2\}$.

Ergodic-theoretic contributions to the field of geometric Ramsey theory were made by Furstenberg, Katznelson, and Weiss [40], who showed that if E is a positive upper density subset of $\mathbb{R}^2$ then: *(i)* E contains points with any large enough distance (see also [21] and [29]), *(ii)* Every δ-neighborhood of E contains three points forming a triangle congruent to any given large enough dilation of a given triangle (in [21] it is shown that if the three points lie on a straight line one cannot always find three points with this property in E itself). Recently, a generalization of property *(ii)* to arbitrary finite configurations of $\mathbb{R}^m$ was obtained by Ziegler [76].

It is also worth mentioning some recent exciting connections of multiple recurrence with some structural properties of the set of prime numbers. The first one is in the work of Green and Tao [45], where the existence of arbitrarily long arithmetic progressions of primes was demonstrated, the authors, in addition to using Szemerédi's theorem outright, use several ideas from its ergodic-theoretic proofs, as appearing in [34] and [39]. The second one is in the recent work of Tao and Ziegler [70], where a quantitative version of Theorem 17 was used to prove that the primes contain arbitrarily long polynomial progressions. Furthermore, several recent results in ergodic theory, related to the structure of the minimal characteristic factors of certain multiple ergodic averages, play an important role in the ongoing attempts of Green and Tao to get asymptotic formulas for the number of k-term arithmetic progressions of primes up to x (see for example [46] and [47]). This project has been completed for $k = 3$, thus verifying an interesting special case of the Hardy–Littlewood k-tuple conjecture predicting the asymptotic growth rate of $N_{a_1,\ldots,a_k}(x) =$ the number of configurations of primes having the form $\{p, p + a_1, \ldots, p + a_k\}$ with $p \leq x$.

Finally, we remark that in this article we have restricted attention to multiple recurrence and Furstenberg correspondence for $\mathbb{Z}$ actions, while in fact there is a wealth of literature on extensions of these results to general commutative, amenable and even non-amenable groups. For an excellent exposition of these and other recent developments the reader is referred to the surveys [9] and [11]. Here, we give just one notable combinatorial corollary to some work of this kind, a density version of the classical Hales–Jewett coloring theorem [48].

Theorem 21 (Furstenberg and Katznelson [38]) *Let $W_n(A)$ denote the set of words of length n with letters in the alphabet $A = \{a_1, \ldots, a_k\}$. For every $\varepsilon > 0$ there exists $N_0 = N_0(\varepsilon, k)$ such that if $n \geq N_0$ then any subset S of $W_n(A)$ with $|S| \geq \varepsilon k^n$ contains a combinatorial line, i. e., a set consisting of k n-letter words, having fixed letters in l positions, for some $0 \leq l < n$, the remaining $n - l$ positions being occupied by a variable letter x, for $x = a_1, \ldots, a_k$. (For example, in $W_4(A)$ the sets $\{(a_1, x, a_2, x) : x \in A\}$ and $\{(x, x, x, x), : x \in A\}$ are combinatorial lines).*

At first glance, the uninitiated reader may not appreciate the importance of this "master" density result, so it is instructive to derive at least one of its immediate consequences. Let $A = \{0, 1, \ldots, k - 1\}$ and interpret $W_n(A)$ as integers in base k having at most n digits. Then a combinatorial line in $W_n(A)$ is an arithmetic progression of length k – for example, the line $\{(a_1, x, a_2, x) : x \in A\}$ corresponds to the progression $\{m, m + n, m + 2n, m + 3n\}$, where $m = a_1 + a_2 d^2$ and $n = d + d^3$. This allows one to deduce Szemerédi's theorem. Similarly, one can deduce from Theorem 21 multidimensional and IP extensions of Szemerédi's theorem [36,37], and some related results about vector spaces over finite fields [37]. Again, the only known proof for the density version of the Hales–Jewett theorem relies heavily on ergodic theory.

Future Directions

In this section we formulate a few open problems relating to the material in the previous three sections. It should be noted that this selection reflects the authors' interests, and does not strive for completeness.

We start with an intriguing question of Katznelson [54] about sets of topological recurrence. A set $S \subset \mathbb{N}$ is a set of Bohr recurrence if for every $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$ and $\varepsilon > 0$ there exists $s \in S$ such that $\{s\alpha_i\} \in [0, \varepsilon] \cup [1 - \varepsilon, 1)$ for $i = 1, \ldots, k$.

Problem 1 *Is every set of Bohr recurrence a set of topological recurrence?*

Background for this problem and evidence for a positive answer can be found in [54,72]. As we mentioned in Sect. "Subsequence Recurrence", there exists a set of topological recurrence (and hence Bohr recurrence) that is not a set of recurrence.

Problem 2 *Is the set $S = \{l!2^m 3^n : l, m, n \in \mathbb{N}\}$ a set of recurrence? Is it a set of k-recurrence for every $k \in \mathbb{N}$?*

It can be shown that S is a set of Bohr recurrence. Theorem 13 cannot be applied since the uniform distribution condition fails for some irrational numbers α. A relevant question was asked by Bergelson in [9]: "Is the set $S = \{2^m 3^n : m, n \in \mathbb{N}\}$ good for single recurrence for weakly mixing systems?"

As we mentioned in Sect. "Multiple Recurrence", the set of primes shifted by 1 (or -1) is a set of 2-recurrence [32].

Problem 3 *Show that the sets $\mathbb{P} - 1$ and $\mathbb{P} + 1$, where $\mathbb{P}$ is the set of primes, are sets of k-recurrence for every $k \in \mathbb{N}$.*

As remarked in [32], a positive answer to this question will follow if some uniformity conjectures of Green and Tao [47] are verified.

We mentioned in Sect. "Subsequence Recurrence" that random non-lacunary sequences (see definition there) are almost surely sets of recurrence.

Problem 4 *Show that random non-lacunary sequences are almost surely sets of k-recurrence for every $k \in \mathbb{N}$.*

The answer is not known even for $k = 2$, though, in unpublished work, Wierdl and Lesigne have shown that the answer is positive for random sequences with at most quadratic growth. We refer the reader to the survey [65] for a nice exposition of the argument used by Bourgain [22] to handle the case $k = 1$.

It was shown in [31] that if S is a set of 2-recurrence then the set of its squares is a set of recurrence for circle rotations. The same method shows that it is actually a set of Bohr recurrence.

Problem 5 *If $S \subset \mathbb{Z}$ is a set of 2-recurrence, is it true that $S^2 = \{s^2 : s \in S\}$ is a set of recurrence?*

A similar question was asked in [23]: "If S is a set of k-recurrence for every k, is the same true of S^2?".

One would like to find a criterion that would allow one to deduce that a sequence is good for double (or higher order) recurrence from some uniform distribution properties of this sequence.

Problem 6 *Find necessary conditions for double recurrence similar to the one given in Theorem 13.*

It is now well understood that such a criterion should involve uniform distribution properties of some generalized polynomials or 2-step nilsequences.

We mentioned in Sect. "Connections with Combinatorics and Number Theory" that every positive density subset of $\mathbb{R}^2$ contains points with any large enough distance. Bourgain [21] constructed a positive density subset E of $\mathbb{R}^2$, a triangle T, and numbers $t_n \to \infty$, such that E does not contain congruent copies of all t_n-dilations of T. But the triangle T used in this construction is degenerate, which leaves the following question open:

Problem 7 *Is it true that every positive density subset of $\mathbb{R}^2$ contains a triangle congruent to any large enough dilation of a given non-degenerate triangle?*

For further discussion on this question the reader can consult the survey [44].

The following question of Aaronson and Nakada [1] is related to a classical question of Erdős concerning whether every $K \subset \mathbb{N}$ such that $\sum_{n \in K} 1/n = \infty$ contains arbitrarily long arithmetic progressions:

Problem 8 *Suppose that $(X, \mathcal{B}, \mu, T)$ is a $\{1/n\}$-conservative ergodic measure preserving system. Is it true that for every $A \in \mathcal{B}$ with $\mu(A) > 0$ and $k \in \mathbb{N}$ we have $\mu(A \cap T^{-n}A \cap \cdots \cap T^{-kn}A) > 0$ for some $n \in \mathbb{N}$?*

The answer is positive for the class of Markov shifts, and it is remarked in [1] that if the Erdős conjecture is true then the answer will be positive in general. The converse is not known to be true. For a related result showing that multiple recurrence is preserved by extensions of infinite measure preserving systems see [61].

Our next problem is motivated by the question whether Theorem 21 has a polynomial version (for a precise formulation of the general conjecture see [9]). Not even this most special consequence of it is known to hold.

Problem 9 *Let $\varepsilon > 0$. Does there exist $N = N(\varepsilon)$ having the property that every family P of subsets of $\{1, \ldots, N\}^2$ satisfying $|P| \geq \varepsilon 2^{N^2}$ contains a configuration $\{A, A \cup (\gamma \times \gamma)\}$, where $A \subset \{1, \ldots, N\}^2$ and $\gamma \subset \{1, \ldots, N\}$ with $A \cap (\gamma \times \gamma) = \emptyset$?*

A measure preserving action of a general countably infinite group G is a function $g \to T_g$ from G into the space of measure preserving transformations of a probability space X such that $T_{gh} = T_g T_h$. It is easy to show that a version of Khintchine's recurrence theorem holds for such actions: if $\mu(A) > 0$ and $\varepsilon > 0$ then $\{g : \mu(A \cap T_g A) > (\mu(A))^2 - \varepsilon\}$ is syndetic. However it is unknown whether the following ergodic version of Roth's theorem holds.

Problem 10 *Let (T_g) and (S_g) be measure preserving G-actions of a probability space X that commute in the sense that $T_g S_h = S_h T_g$ for all $g, h \in G$. Is it true that for all positive measure sets A, the set of g such that $\mu(A \cap T_g A \cap S_g A) > 0$ is syndetic?*

We remark that for general (possibly amenable) groups G not containing arbitrarily large finite subgroups nor elements of infinite order, it is not known whether one can

find a *single* such $g \neq e$. On the other hand, the answer is known to be positive for general G in case $(T_g^{-1} S_g)$ is a G-action [18]; even under such strictures, however, it is unknown whether a triple recurrence theorem holds.

Bibliography

1. Aaronson J (1981) The asymptotic distribution behavior of transformations preserving infinite measures. J Analyse Math 39:203–234
2. Aaronson J (1997) An introduction to infinite ergodic theory. Mathematical Surveys and Monographs 50. American Mathematical Society, Providence
3. Aaronson J, Nakada H (2000) Multiple recurrence of Markov shifts and other infinite measure preserving transformations. Israel J Math 117:285–310
4. Barreira L (2001) Hausdorff dimension of measures via Poincaré recurrence. Comm Math Phys 219:443–463
5. Barreira L (2005) Poincaré recurrence: old and new. XIVth International Congress on Mathematical Physics, World Sci Publ Hackensack, NJ, pp 415–422
6. Behrend F (1946) On sets of integers which contain no three in arithmetic progression. Proc Nat Acad Sci 23:331–332
7. Bergelson V (1987) Weakly mixing PET. Ergod Theory Dynam Syst 7:337–349
8. Bergelson V (1987) Ergodic Ramsey Theory. In: Simpson S (ed) Logic and Combinatorics. Contemporary Mathematics 65. American Math Soc, Providence, pp 63–87
9. Bergelson V (1996) Ergodic Ramsey Theory – an update. In: Pollicot M, Schmidt K (eds) Ergodic theory of $\mathbb{Z}^d$-actions. Lecture Note Series 228. London Math Soc, London, pp 1–61
10. Bergelson V (2000) The multifarious Poincaré recurrence theorem. In: Foreman M, Kechris A, Louveau A, Weiss B (eds) Descriptive set theory and dynamical systems. Lecture Note Series 277. London Math Soc, London, pp 31–57
11. Bergelson V (2005) Combinatorial and diophantine applications of ergodic theory. In: Hasselblatt B, Katok A (eds) Handbook of dynamical systems, vol 1B. Elsevier, pp 745–841
12. Bergelson V, Furstenberg H, McCutcheon R (1996) IP-sets and polynomial recurrence. Ergod Theory Dynam Syst 16:963–974
13. Bergelson V, Håland-Knutson I, McCutcheon R (2006) IP Systems, generalized polynomials and recurrence. Ergod Theory Dynam Syst 26:999–1019
14. Bergelson V, Host B, Kra B (2005) Multiple recurrence and nilsequences. Inventiones Math 160(2):261–303
15. Bergelson V, Host B, McCutcheon R, Parreau F (2000) Aspects of uniformity in recurrence. Colloq Math 84/85(2):549–576
16. Bergelson V, Leibman A (1996) Polynomial extensions of van der Waerden's and Szemerédi's theorems. J Amer Math Soc 9:725–753
17. Bergelson V, McCutcheon R (2000) An ergodic IP polynomial Szemerédi theorem. Mem Amer Math Soc 146:viii–106
18. Bergelson V, McCutcheon R (2007) Central sets and a noncommutative Roth theorem. Amer J Math 129:1251–1275
19. Birkhoff G (1931) A proof of the ergodic theorem. Proc Nat Acad Sci 17:656–660
20. Boshernitzan M (1993) Quantitative recurrence results. Invent Math 113:617–631
21. Bourgain J (1986) A Szemerédi type theorem for sets of positive density in $\mathbb{R}^k$. Israel J Math 54(3):307–316
22. Bourgain J (1988) On the maximal ergodic theorem for certain subsets of the positive integers. Israel J Math 61:39–72
23. Brown T, Graham R, Landman B (1999) On the set of common differences in van der Waerden's theorem on arithmetic progressions. Canad Math Bull 42:25–36
24. Carathéodory C (1968) Vorlesungen über reelle Funktionen, 3rd edn. Chelsea Publishing Co, New York
25. Chazottes J, Ugalde E (2005) Entropy estimation and fluctuations of hitting and recurrence times for Gibbsian sources. Discrete Contin Dyn Syst Ser B 5(3):565–586
26. Conze J, Lesigne E (1984) Théorèmes ergodiques pour des mesures diagonales. Bull Soc Math France 112(2):143–175
27. Conze J, Lesigne E (1988) Sur un théorème ergodique pour des mesures diagonales. Probabilités, Publ Inst Rech Math Rennes 1987-1, Univ Rennes I, Rennes, pp 1–31
28. Evans D, Searles D (2002) The fluctuation theorem. Adv Phys 51:1529–1585
29. Falconer K, Marstrand J (1986) Plane sets with positive density at infinity contain all large distances. Bull Lond Math Soc 18:471–474
30. Frantzikinakis N, Kra B (2006) Ergodic averages for independent polynomials and applications. J Lond Math Soc 74(1): 131–142
31. Frantzikinakis N, Lesigne E, Wierdl M (2006) Sets of k-recurrence but not $(k+1)$-recurrence. Annales de l'Institut Fourier 56(4):839–849
32. Frantzikinakis N, Host B, Kra B (2007) Multiple recurrence and convergence for sets related to the primes. J Reine Angew Math 611:131–144. Available at http://arxiv.org/abs/math/0607637
33. Frantzikinakis N (2008) Multiple ergodic averages for three polynomials and applications. Trans Am Math Soc 360(10):5435–5475. Available at http://arxiv.org/abs/math/0606567
34. Furstenberg H (1977) Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. J Anal Math 71:204–256
35. Furstenberg H (1981) Recurrence in ergodic theory and combinatorial number theory. Princeton University Press, Princeton
36. Furstenberg H, Katznelson Y (1979) An ergodic Szemerédi theorem for commuting transformations. J Analyse Math 34:275–291
37. Furstenberg H, Katznelson Y (1985) An ergodic Szemerédi theorem for IP-systems and combinatorial theory. J Analyse Math 45:117–168
38. Furstenberg H, Katznelson Y (1991) A density version of the Hales–Jewett theorem. J Analyse Math 57:64–119
39. Furstenberg H, Katznelson Y, Ornstein D (1982) The ergodic theoretical proof of Szemerédi's theorem. Bull Amer Math Soc (NS) 7(3):527–552
40. Furstenberg H, Katznelson Y, Weiss B (1990) Ergodic theory and configurations in sets of positive density. Mathematics of Ramsey theory. Algorithms Combin 5. Springer, Berlin, pp 184–198
41. Furstenberg H, Weiss B (1996) A mean ergodic theorem for $(1/N)\sum_{n=1}^{N} f(T^n x) g(T^{n^2} x)$. Convergence in ergodic theory and probability (Columbus, OH, 1993), Ohio State Univ Math Res Inst Publ 5, de Gruyter, Berlin, pp 193–227
42. Galatolo S, Kim DH, Park KK (2006) The recurrence time for ergodic systems with infinite invariant measures. Nonlinearity 19:2567–2580

43. Gowers W (2001) A new proof of Szemerédi's theorem. Geom Funct Anal 11:465–588
44. Graham RL (1994) Recent trends in Euclidean Ramsey theory. Trends in discrete mathematics. Discret Math 136(1–3):119–127
45. Green B, Tao T (2008) The primes contain arbitrarily long arithmetic progressions. Ann Math 167:481–547. Available at http://arxiv.org/abs/math/0404188
46. Green B, Tao T (to appear) Quadratic uniformity of the Möbius function. Annales de l'Institut Fourier. Available at http://arxiv.org/abs/math.NT/0606087
47. Green B, Tao T () Linear equations in primes. Ann Math (to appear). Available at http://arxiv.org/abs/math.NT/0606088
48. Hales A, Jewett R (1963) Regularity and positional games. Trans Amer Math Soc 106:222–229
49. Hasley T, Jensen M (2004) Hurricanes and butterflies. Nature 428:127–128
50. Host B, Kra B (2005) Nonconventional ergodic averages and nilmanifolds. Ann Math 161:397–488
51. Kac M (1947) On the notion of recurrence in discrete stochastic processes. Bull Amer Math Soc 53:1002–10010
52. Kra B (2006) The Green-Tao theorem on arithmetic progressions in the primes: an ergodic point of view. Bull Amer Math Soc (NS) 43:3–23
53. Kamae T, Mendés-France M (1978) Van der Corput's difference theorem. Israel J Math 31:335–342
54. Katznelson Y (2001) Chromatic numbers of Cayley graphs on $\mathbb{Z}$ and recurrence. Paul Erdős and his mathematics (Budapest, 1999). Combinatorica 21(2):211–219
55. Khintchine A (1934) Eine Verschärfung des Poincaréschen "Wiederkehrsatzes". Comp Math 1:177–179
56. Kriz I (1987) Large independent sets in shift invariant graphs. Solution of Bergelson's problem. Graphs Combinatorics 3:145–158
57. Leibman A (2002) Lower bounds for ergodic averages. Ergod Theory Dynam. Syst 22:863–872
58. Leibman A (2005) Pointwise convergence of ergodic averages for polynomial sequences of rotations of a nilmanifold. Ergod Theory Dynam Syst 25:201–213
59. Leibman A (2005) Pointwise convergence of ergodic averages for polynomial actions of $\mathbb{Z}^d$ by translations on a nilmanifold. Ergod Theory Dynam Syst 25:215–225
60. McCutcheon R (2005) FVIP systems and multiple recurrence. Israel J Math 146:157–188
61. Meyerovitch T (2007) Extensions and multiple recurrence of infinite measure preserving systems. Preprint. Available at http://arxiv.org/abs/math/0703914
62. Ornstein D, Weiss B (1993) Entropy and data compression schemes. IEEE Trans Inform Theory 39:78–83
63. Petersen K (1989) Ergodic theory. Cambridge Studies in Advanced Mathematics 2. Cambridge University Press, Cambridge
64. Poincaré H (1890) Sur le problème des trois corps et les équations de la dynamique. Acta Math 13:1–270
65. Rosenblatt J, Wierdl M (1995) Pointwise ergodic theorems via harmonic analysis. Ergodic theory and its connections with harmonic analysis (Alexandria, 1993). London Math Soc, Lecture Note Series 205, Cambridge University Press, Cambridge, pp 3–151
66. Sárközy A (1978) On difference sets of integers III. Acta Math Acad Sci Hungar 31:125–149
67. Shkredov I (2002) Recurrence in the mean. Mat Zametki 72(4):625–632; translation in Math Notes 72(3–4):576–582
68. Sklar L (2004) Philosophy of Statistical Mechanics. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy (Summer 2004 edn), http://plato.stanford.edu/archives/sum2004/entries/statphys-statmech/
69. Szemerédi E (1975) On sets of integers containing no k elements in arithmetic progression. Acta Arith 27:299–345
70. Tao T, Ziegler T () The primes contain arbitrarily long polynomial progressions. Acta Math (to appear). Available at http://www.arxiv.org/abs/math.DS/0610050
71. Walters P (1982) An introduction to ergodic theory. Graduate Texts in Mathematics, vol 79. Springer, Berlin
72. Weiss B (2000) Single orbit dynamics. CBMS Regional Conference Series in Mathematics 95. American Mathematical Society, Providence
73. Wyner A, Ziv J (1989) Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. IEEE Trans Inform Theory 35:1250–1258
74. Zermelo E (1896) Über einen Satz der Dynamik und die mechanische Wärmetheorie. Annalen der Physik 57:485–94; English translation, On a theorem of dynamics and the mechanical theory of heat. In: Brush SG (ed) Kinetic Theory. Oxford, 1966, II, pp 208–17
75. Ziegler T (2007) Universal characteristic factors and Furstenberg averages. J Amer Math Soc 20:53–97
76. Ziegler T (2006) Nilfactors of $\mathbb{R}^m$-actions and configurations in sets of positive upper density in $\mathbb{R}^m$. J Anal Math 99:249–266

Ergodic Theory: Rigidity

Viorel Niţică[1,2]

[1] West Chester University, West Chester, USA

[2] Institute of Mathematics, Bucharest, Romania

Article Outline

Glossary
Definition of the Subject
Introduction
Basic Definitions and Examples
Differentiable Rigidity
Local Rigidity
Global Rigidity
Measure Rigidity
Future Directions
Acknowledgment
Bibliography

Glossary

Differentiable rigidity Differentiable rigidity refers to finding invariants to the differentiable conjugacy of dynamical systems, and, more general, group actions.

Local rigidity Local rigidity refers to the study of perturbations of homomorphisms from discrete or continuous groups into diffeomorphism groups.

Global rigidity Global rigidity refers to the classification of all group actions on manifolds satisfying certain conditions.

Measure rigidity Measure rigidity refers to the study of invariant measures for actions of abelian groups and semigroups.

Lattice A lattice in a Lie group is a discrete subgroup of finite covolume.

Conjugacy Two elements g_1, g_2 in a group G are said to be conjugated if there exists an element $h \in G$ such that $g_1 = h^{-1} g_2 h$. The element h is called conjugacy.

C^k Conjugacy Two diffeomorphisms ϕ_1, ϕ_2 acting on the same manifold M are said to be C^k-conjugated if there exists a C^k diffeomorphism h of M such that $\phi_1 = h^{-1} \circ \phi_2 \circ h$. The diffeomorphism h is called C^k conjugacy.

Definition of the Subject

As one can see from this volume, chaotic behavior of complex dynamical systems is prevalent in nature and in large classes of transformations. Rigidity theory can be viewed as the counterpart to the generic theory of dynamical systems which often investigates chaotic dynamics for a typical transformation belonging to a large class. In rigidity one is interested in finding obstructions to chaotic, or generic, behavior. This often leads to rather unexpected classification results. As such, rigidity in dynamics and ergodic theory is difficult to define precisely and the best approach to this subject is to study various results and themes that developed so far. A classification is offered below in local, global, differentiable and measurable rigidity. One should note that all branches are strongly intertwined and, at this stage of the development of the subject, it is difficult to separate them.

Rigidity is a well developed and prominent topic in modern mathematics. Historically, rigidity has two main origins, one coming from the study of lattices in semi-simple Lie groups, and one coming from the theory of hyperbolic dynamical systems. From the start, ergodic theory was an important tool used to prove rigidity results, and a strong interdependence developed between these fields. Many times a result in rigidity is obtained by combining techniques from the theory of lattices in Lie groups with techniques from hyperbolic dynamical systems and ergodic theory. Among other mathematical disciplines using results and contributing to this field one can mention representation theory, smooth, continuous and measurable dynamics, harmonic and spectral analysis, partial differential equations, differential geometry, and number theory. Additional details about the appearance of rigidity in ergodic theory as well as definitions for some terminology used in the sequel can be found in the articles ► Ergodic Theory on Homogeneous Spaces and Metric Number Theory by Kleinbock, ► Ergodic Theory: Recurrence by Frantzikinakis, McCutcheon, and ► Ergodic Theory: Interactions with Combinatorics and Number Theory by Ward. The theory of hyperbolic dynamics is presented in the article ► Hyperbolic Dynamical Systems by Viana and in the article ► Smooth Ergodic Theory by Wilkinson.

Introduction

The first results about classification of lattices in semi-simple Lie groups were local, aimed at trying to understand the space of small perturbations of a given linear representation. A major contributor was Weil [110,111,112], who proved local rigidity of linear representations for large classes of groups, in particular lattices. Another breakthrough was the contribution of Kazhdan [66], who introduced property (T), allowing to show that large classes of lattices are finitely generated. Rigidity theory matured due to the remarkable global rigidity results obtained by Mostow [90] and Margulis [82], leading to a complete classification of lattices in large classes of semi-simple Lie groups.

Briefly, a hyperbolic (Anosov) dynamical system is one that exhibits strong expansion and contraction along complementary directions. An early contribution introducing this class of objects is the paper of Smale [106], in which basic examples and techniques are introduced. A breakthrough came with the results of Anosov [1], who proved structural stability of the Anosov systems and ergodicity of the geodesic flow on a manifold of negative curvature. Motivated by questions arising in mathematical physics, chaos theory and other areas, hyperbolic dynamics emerged as one of the major fields of contemporary mathematics. From the beginning, a major unsolved problem in the field was the classification of Anosov diffeomorphisms and flows.

In the 80s a change in philosophy occurred, partially motivated by a program introduced by Zimmer [114]. The goal of the program was to classify the smooth actions of higher rank semi-simple Lie groups and of their (irreducible) lattices on compact manifolds. It was expected that any such lattice action that preserves a smooth volume form and is ergodic can be reduced to one of the following standard models: isometric actions, linear actions on infranilmanifolds, and left translations on compact ho-

mogeneous spaces. This original conjecture was disproved by Katok, Lewis (see [56]): by blowing up a linear nilmanifold-action at some fixed points they exhibit real-analytic, volume preserving, ergodic lattice actions on manifolds with complicated topology.

Nevertheless, imposing extra assumptions on a higher rank action, for example the existence of some hyperbolicity, allows local and global classification results. The concept of *Anosov action*, that is, an action that contains at least one Anosov diffeomorphism, was introduced for general groups by Pugh, Shub [99]. The significant differences between the classical $\mathbb{Z}$ and $\mathbb{R}$ cases and those of higher rank lattices, or at a more basic level, of higher rank abelian groups, went unnoticed for a while. The surge of activity in the 80s allowed these differences to surface: for lattices in the work of Hurder, Katok, Lewis, Zimmer (see [43,55,56,63]); and for higher rank abelian groups in the work of Katok, Lewis [55] and Katok, Spatzier [59]. As observed in these papers, local and global rigidity are typical for such Anosov actions. This generated additional research which is summarized in Sects. "Local Rigidity" and "Global Rigidity".

Differentiable rigidity is covered in Sect. "Differentiable Rigidity". An interesting problem is to find moduli for the C^k conjugacy, $k \geq 1$, of Anosov diffeomorphisms and flows. This was tackled so far only for low dimensional cases ($n = 2$ for diffeomorphisms and $n = 3$ for flows). Another direction that can be included here refers to finding obstructions for higher transverse regularity of the stable/unstable foliation of a hyperbolic system. A spin-off of the research done so far, which is of high interest by itself, and has applications to local and global rigidity, consists of results lifting the regularity of solutions of cohomological equations over hyperbolic systems. In turn, these results motivated a more careful study of analytic questions about lifting the regularity of real valued continuous functions that enjoy higher regularity along webs of foliations. We also include in this section rigidity results for cocycles over higher rank abelian actions. These are crucial to the proof of local rigidity of higher rank abelian group actions. A more detailed presentation of the material relevant to differentiable rigidity can be found in the forthcoming monograph [58].

Measure rigidity refers to the study of invariant measures under actions of abelian groups and semigroups. If the actions are hyperbolic, higher-rank, and satisfy natural algebraic and irreducibility assumptions, one expects the invariant measures to be rare. This direction was started by a question of Furstenberg, asking if any nonatomic probability measure on the circle, invariant and ergodic under multiplications by 2 and 3, is the Lebesgue measure. An early contribution is that of Rudolph [103], who answered positively if the action has an element of strictly positive entropy. Katok, Spatzier [61] extended the question to more general higher rank abelian actions, such as actions by linear automorphisms of tori and Weyl chamber flows. A related direction is the study of the invariant sets and measures under the action of horocycle flows, where important progress was made by Ratner [100,101] and earlier by Margulis [16,81,83]. An application of these results present in the last papers is the proof of the long standing Oppenheim's conjecture, about the density of the values of the quadratic forms at integer points. Recent developments due to and Einsiedler, Katok, Lindenstrauss [20] give a partial answer to another outstanding conjecture in number theory, Littlewood's conjecture, and emphasize measure rigidity as one of a more promising directions in rigidity. More details are shown in Sect. "Measure Rigidity".

Four other recent surveys of rigidity theory, each one with a fair amount of overlap but also complementary in part to the present one, that discuss various aspects of the field and its significance are written by Fisher [23], Lindenstrauss [68], Nițică, Török [95], and Spatzier [107]. Among these, [23] concentrates mostly on local and global rigidity, [95] on differentiable rigidity, [68] on measure rigidity, and [107] gives a general overview.

Here is a word of caution for the reader. Many times, instead of the most general results, we present an example that contains the essence of what is available. Also, several important facts that should have been included, are left out. This is because stating complete results would require more space than allocated to this material. The limited knowledge of the author also plays a role here. He apologizes for any obvious omissions and hopes that the bibliography will help fill the gaps.

Basic Definitions and Examples

A detailed introduction to the theory of Anosov systems and hyperbolic dynamics is given in the monograph [51]. The proofs of the basic results for diffeomorphisms stated below can be found there. The proofs for flows are similar. Surveys about hyperbolic dynamics in this volume are the article ► Hyperbolic Dynamical Systems by Viana and the article ► Smooth Ergodic Theory by Wilkinson.

Consider a compact differentiable manifold M and $f: M \to M$ a C^1 diffeomorphism. Let TM be the tangent bundle of M, and $Df: TM \to TM$ be the derivative of f. The map f is said to be an *Anosov diffeomorphism* if there is a smooth Riemannian metric $\|\cdot\|$ on M, which induces a metric d_M called *adapted*, a number $\lambda \in (0,1)$, and

a continuous Df invariant splitting $TM = E^s \oplus E^u$ such that

$$\|Df v\| \le \lambda \|v\|\,, v \in E^s\,, \quad \|Df^{-1} v\| \le \lambda \|v\|\,, v \in E^u\,.$$

For each $x \in M$ there is a pair of embedded C^1 discs $W^s_{\rm loc}(x)$, $W^u_{\rm loc}(x)$, called the local stable manifold and the local unstable manifold at x, respectively, such that:

1. $T_x W^s_{\rm loc}(x) = E^s(x)\,,\ T_x W^u_{\rm loc}(x) = E^u(x)$;
2. $f(W^s_{\rm loc}(x)) \subset W^s_{\rm loc}(fx)\,,\ f^{-1}(W^u_{\rm loc}(x)) \subset W^u_{\rm loc}(f^{-1}x)$;
3. For any $\mu \in (\lambda, 1)$, there exists a constant $C > 0$ such that for all $n \in \mathbb{N}$,

$$d_M(f^n x, f^n y) \le C\mu^n d_M(x, y)\,, \quad \text{for} \quad y \in W^s_{\rm loc}(x)\,,$$
$$d_M(f^{-n} x, f^{-n} y) \le C\mu^n d_M(x, y)\,, \quad \text{for} \quad y \in W^u_{\rm loc}(x)\,.$$

The local stable (unstable) manifolds can be extended to global stable (unstable) manifolds $W^s(x)$ and $W^u(x)$ which are well defined and smoothly injectively immersed. These global manifolds are the leaves of global foliations W^s and W^u of M. In general, these foliations are only continuous, but their leaves are differentiable.

Let $\phi\colon \mathbb{R} \times M \to M$ be a C^1 flow. The flow ϕ is said to be an *Anosov flow* if there is a Riemannian metric $\|\cdot\|$ on M, a constant $0 < \lambda < 1$, and a continuous Df invariant splitting $TM = E^s \oplus E^0 \oplus E^u$ such that for all $x \in M$ and $t > 0$:

1. $\frac{\rm d}{{\rm d}t}|_{t=0}\phi^t \in E^c_x \setminus \{0\}\,,\ \dim E^c_x = 1$,
2. $\|D\phi^t v\| \le \lambda^t \|v\|\,,\ v \in E^s$,
3. $\|D\phi^{-t} v\| \le \lambda^t \|v\|\,,\ v \in E^u$.

For each $x \in M$ there is a pair of embedded C^1 discs $W^s_{\rm loc}(x)\,, W^u_{\rm loc}(x)$, called the local (strong) stable manifold and the local (strong) unstable manifold at x, respectively, such that:

1. $T_x W^s_{\rm loc}(x) = E^s(x)\,,\ T_x W^u_{\rm loc}(x) = E^u(x)$;
2. $\phi^t(W^s_{\rm loc}(x)) \subset W^s_{\rm loc}(\phi^t x)\,,$
   $\phi^{-t}(W^u_{\rm loc}(x)) \subset W^u_{\rm loc}(\phi^{-t} x)$ for $t > 0$;
3. For any $\mu \in (\lambda, 1)$, there exists a constant $C > 0$ such that for all $n \in \mathbb{N}$,

$$d_M(\phi^t x\,, \phi^t y) \le C\mu^t d_M(x, y)\,, \quad \text{for} \quad y \in W^s_{\rm loc}(x)\,, t > 0\,,$$
$$d_M(\phi^{-t} x\,, \phi^{-t} y) \le C\mu^t d_M(x, y)\,, \quad \text{for} \quad y \in W^u_{\rm loc}(x)\,, t > 0\,.$$

The local stable (unstable) manifolds can be extended to global stable (unstable) manifolds $W^s(x)$ and $W^u(x)$. These global manifolds are the leaves of global foliations W^s and W^u of M. One can also define weak stable and weak unstable foliations with leaves given by $W^{cs}(x) = \cup_{t\in\mathbb{R}}(W^s(x))$ and $W^{cu}(x) = \cup_{t\in\mathbb{R}}(W^u(x))$, which have as tangent distributions $E^{cs} = E^c \oplus E^s$ and $E^{cu} = E^c \oplus E^s$. In general, all these foliations are only continuous, but their leaves are differentiable.

Any Anosov diffeomorphism is *structurally stable*, that is, any C^1 diffeomorphism that is C^1 close to an Anosov diffeomorphism is topologically conjugate to the unperturbed one via a Hölder homeomorphism. An Anosov flow is structurally stable in the orbit equivalence sense: any C^1 small perturbation of an Anosov flow has the orbit foliation topologically conjugate via a Hölder homeomorphism to the orbit foliation of the unperturbed flow.

Let $SL(n, \mathbb{R})$ be the group of all n-dimensional square matrices with real valued entries of determinant 1. Let $SL(n, \mathbb{Z}) \subset SL(n, \mathbb{R})$ be the subgroup with integer entries. Basic examples of Anosov diffeomorphisms are automorphisms of the n-torus $\mathbb{T}^n = \mathbb{R}^n/\mathbb{Z}^n$ induced by hyperbolic matrices in $SL(n, \mathbb{Z})$. A hyperbolic matrix is one that has only nonzero eigenvalues, all away in absolute value from 1. A specific example of such matrix in $SL(2, \mathbb{Z})$ is

$$\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\,.$$

Basic examples of Anosov flows are given by the geodesic flows of surfaces of constant negative curvature. The unitary bundle of such a surface can be realized as $M = \Gamma \setminus PSL(2, \mathbb{R})$, where $PSL(2, \mathbb{R}) = SL(2, \mathbb{R})/\{\pm 1\}$ and Γ is a cocompact lattice in $PSL(2, \mathbb{R})$. The action of the geodesic flow on M is induced by right multiplication with elements in the diagonal one-parameter subgroup

$$\left\{ \begin{pmatrix} {\rm e}^{t/2} & 0 \\ 0 & {\rm e}^{-t/2} \end{pmatrix}, t \in \mathbb{R} \right\}\,.$$

A related transformation, which is not hyperbolic, but will be of interest in this presentation, is the *horocycle flow* induced by right multiplication on M by elements in the one parameter subgroup

$$\left\{ \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}, t \in \mathbb{R} \right\}\,.$$

Of interest in this survey are also actions of more general groups than $\mathbb{Z}$ and $\mathbb{R}$. Typical examples of higher rank $\mathbb{Z}^k$ Anosov actions are constructed on tori using groups of units in number fields. See [65] for more details about this construction. A particular example of Anosov $\mathbb{Z}^2$-action on $\mathbb{T}^3$ is induced by the hyperbolic matrices:

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 8 & 2 \end{pmatrix}\,, \quad B = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 1 & 8 & 4 \end{pmatrix}\,.$$

One can check, by looking at the eigenvalues, that A and B are not multiples of the same matrix. Moreover, A and B commute.

Typical examples of higher rank Anosov $\mathbb{R}^k$-actions are given by Weyl chamber flows, which we now describe using some notions from the theory of Lie groups. A good reference for the background in Lie group theory necessary here is the book of Helgason [40]. Note that for a hyperbolic element of such an action the center distribution is k-dimensional and coincides with the tangent distribution to the orbit foliation of $\mathbb{R}^k$.

Let G be a semi-simple connected real Lie group of the noncompact type, with Lie algebra $\mathfrak{g}$. Let $K \subset G$ be a maximal compact subgroup that gives a Cartan decomposition $\mathfrak{g} = \mathfrak{k} + \mathfrak{p}$, where $\mathfrak{k}$ is the Lie algebra of K and $\mathfrak{p}$ is the orthogonal complement of $\mathfrak{k}$ with respect to the Killing form of $\mathfrak{g}$. Let $\mathfrak{a} \subset \mathfrak{p}$ be a maximal abelian subalgebra and $A = \exp \mathfrak{a}$ be the corresponding subgroup. The simultaneous diagonalization of $ad_{\mathfrak{g}}(\mathfrak{a})$ gives the decomposition

$$\mathfrak{g} = \mathfrak{g} + \sum_{\lambda \in \Lambda} \mathfrak{g}_\lambda , \quad \mathfrak{g}_0 = \mathfrak{a} + \mathfrak{m} ,$$

where Λ is the set of restricted roots. The spaces $\mathfrak{g}_\lambda$ are called *root spaces*. A point $H \in \mathfrak{a}$ is called *regular* if $\lambda(H) \neq 0$ for all $\lambda \in \Lambda$. Otherwise it is called *singular*. The set of regular elements consists of the complement of a union of finitely many hyperplanes. Its components are cones in $\mathfrak{a}$ called *Weyl chambers*. The faces of the Weyl chambers are called *Weyl chamber walls*.

Let M be the centralizer of A in K. Suppose Γ is an irreducible torsion-free cocompact lattice in G. Since A commutes with M, the action of A by right translations on $\Gamma \setminus G$ descends to an A-action on $N := \Gamma \setminus G/M$. This action is called a *Weyl chamber flow*. Any Weyl chamber flow is an Anosov action, that is, has an element that acts hyperbolically transversally to the orbit foliation of A. Note that all maximal connected $\mathbb{R}$ diagonalizable subgroups of G are conjugate and their common dimension is called the $\mathbb{R}$-rank of G. If the $\mathbb{R}$-rank k of G is higher than 2, then the Weyl chamber flow is a higher rank hyperbolic $\mathbb{R}^k$-action.

An example of semi-simple Lie group is $SL(n, \mathbb{R})$. Let A be the diagonal subgroup of matrices with positive entries in $SL(n, \mathbb{R})$. An example of Weyl chamber flow that will be discussed in the sequel is the action of A by right translations on $\Gamma \setminus SL(n, \mathbb{R})$, where Γ is a cocompact lattice. In this case the centralizer M is trivial. The rank of this action is $n - 1$. The picture of the Weyl chambers for $n = 3$ is shown in Fig. 1. The signs that appear in each chamber are the signs of half of the Lyapunov exponents of a regular element from the chamber with respect to a certain fixed basis. For this action, the Lyapunov exponents appear in pairs of opposite signs.

ERGODIC THEORY: RIGIDITY

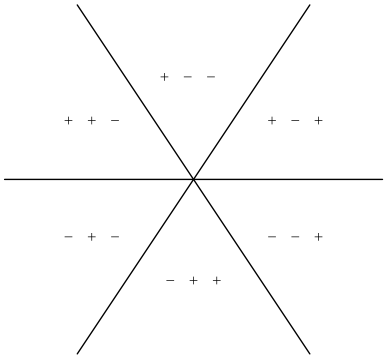

Ergodic Theory: Rigidity, Figure 1
Weyl chambers for $SL(3, \mathbb{R})$

An example of higher rank lattice Anosov action that will be discussed in the sequel is the standard action of $SL(n, \mathbb{Z})$ on the torus $\mathbb{T}^n$, $(A, x) \mapsto Ax, A \in SL(n, \mathbb{Z}), x \in \mathbb{T}^n$. $SL(n, \mathbb{Z})$ is a (noncocompact!) lattice in $SL(n, \mathbb{R})$. As shown in [55], this action is generated by Anosov diffeomorphisms.

We describe now a class of dynamical systems more general than the hyperbolic one. A C^1 diffeomorphism f of a compact differentiable manifold M is called *partially hyperbolic* if there exists a continuous invariant splitting of the tangent bundle $TM = E^s \oplus E^0 \oplus E^u$ such the the derivative of f expands E^u much more than E^0, and contracts E^s much more than E^0. See [9,41] and [10] for the theory of partially hyperbolic diffeomorphisms. E^s and E^u are called stable, respectively unstable distributions, and are integrable. E^0 is called center distribution and, in general, it is not integrable. A structural stability result proved by Hirsch, Pugh, Shub [41], that is a frequently used tool in rigidity, shows that, if E^0 is integrable to a smooth foliation, then any perturbation $\tilde{f}$ of f is partially hyperbolic and has an integrable center foliation. Moreover, the center foliations of $\tilde{f}$ of f are mapped one into the other by a homeomorphism that conjugates the maps induced on the factor spaces of the center foliations by $\tilde{f}$ of f respectively.

We review now basic facts about cocycles. These basic definitions refer to several regularity classes: measurable, continuous, or differentiable. Let G be a group acting on a set M, and denote the action $G \times M \to M$ by $(g, x) \mapsto gx$; thus $(g_1 g_2)x = g_1(g_2 x)$. Let Γ be a group with unit 1_Γ. M is usually endowed with a measurable, continuous, or differentiable structure. A *cocycle* β over

the action is a function $\beta: G \times M \to \Gamma$ such that

$$\beta(g_1 g_2, x) = \beta(g_1, g_2 x)\beta(g_2, x)\,, \tag{1}$$

for all $g_1, g_2 \in G, x \in M$. Note that any group representation $\pi: G \to \Gamma$ defines a cocycle called *constant cocycle*. The trivial representation defines the *trivial cocycle*.

A natural equivalence relation for the set of cocycles is given by cohomology. Two cocycles $\beta_1, \beta_2: G \times M \to \Gamma$ are cohomologous if there exists a map $P: M \to \Gamma$, called *transfer map*, such that

$$\beta_1(g, x) = P(gx)\beta_2(g, x)P(x)^{-1}\,, \tag{2}$$

for all $g \in G, x \in M$.

Differentiable Rigidity

We start by reviewing cohomological results. Several basic notions are already defined in Sect. "Basic Definitions and Examples". In this section we assume that the cocycles are at least continuous. A cocycle $\beta: G \times M \to \Gamma$ over an action $(g, x) \mapsto gx, g \in G, x \in M$, is said to satisfy *closing conditions* if for any $g \in G$ and $x \in M$ such that $gx = x$, one has $\beta(g, x) = 1_\Gamma$. Note that closing conditions are necessary in order for a cocycle to be cohomologous to the trivial one. Since a $\mathbb{Z}$-cocycle is determined by a function $\beta: M \to \Gamma, \beta(x) := \beta(1, x)$, the closing conditions become

$$f^n x = x \quad \text{implies} \quad \beta(f^{n-1}x) \ldots \beta(x) = 1_\Gamma\,, \tag{3}$$

where $f: M \to M$ is the function that implements the $\mathbb{Z}$-action.

The first cohomological results for hyperbolic systems were obtained by Livshits [70,71]. Let M be a compact Riemannian manifold with a $\mathbb{Z}$-action implemented by a topologically transitive Anosov C^1 diffeomorphism f. Then an α-Hölder function $\beta: M \to \mathbb{R}$ determines a cocycle cohomologous to a trivial cocycle if and only if β satisfies the closing conditions (3). The transfer map is α-Hölder. Moreover, for each Hölder class α and each finite dimensional Lie group Γ, there is a neighborhood U of the identity in Γ such that an α-Hölder function $\beta: M \to \Gamma$ determines a cocycle cohomologous to the trivial cocycle if and only if β satisfies the closing conditions (3). The transfer map is again α-Hölder. Similar results are true for Anosov flows.

Using Fourier analysis, Veech [108] extended Livshits's result to real valued cocycles over $\mathbb{Z}$-actions induced by ergodic endomorphisms of an n-dimensional torus, not necessarily hyperbolic. For cocycles with values in abelian groups, the question of two arbitrary cocycles being cohomologous reduces to the question of an arbitrary cocycle being cohomologous to a trivial one. This is not the case for cocycles with values in non-abelian groups. Parry [98] extended Livshits's criteria to one for cohomology of two arbitrary cocycles with values in compact Lie groups. Parry's result was generalized by Schmidt [104] to cocycles with values in Lie groups that, in addition, satisfy a center bunching condition. Niţică, Török [92] extended Livshits's result to cocycles with values in the group $\mathrm{Diff}^k(M)$ of C^k diffeomorphism of a compact manifold M with stably trivial bundle, $k \geq 3$. Examples of such manifolds are the tori and the spheres. In this case, the transfer map takes values in $\mathrm{Diff}^{k-3}(M)$, and it is Hölder with respect to a natural metric on $\mathrm{Diff}^{k-3}(M)$. In [92] one can also find a generalization of Livshits's result to generic Anosov actions, that is, actions generated by families of Anosov diffeomorphisms that do not interchange the stable and unstable directions of elements in the family. An example of such an action is the standard action of $SL(n, \mathbb{Z})$ on the n-dimensional torus.

A question of interest is the following: if two C^k cocycles, $1 \leq k \leq \omega$, over a hyperbolic action, are cohomologous through a continuous/measurable transfer map P, what can be said about the higher regularity of P? For real valued cocycles the question can be reduced to one about cohomologically trivial cocycles. Livshits showed that for a real valued C^1 cocycle cohomologous to a constant the transfer map is C^1. He also obtained C^∞ regularity results if the action is given by hyperbolic automorphisms of a torus. After preliminary results by Guillemin and Kazhdan for geodesic flows on surfaces of negative curvature, for general hyperbolic systems the question was answered positively by de la Llave, Marco, Moriyon [76] in the C^∞ case and by de la Llave [74] in the real analytic case. Niţică, Török [93] considered the lift of regularity for a transfer map between two cohomologous cocycles with values in a Lie group or a diffeomorphism group. In contrast to the case of cocycles cohomologous to trivial ones, here one needs to require for the transfer map a certain amount of Hölder regularity that depends on the ratio between the expansion/contraction that appears in the base and the expansion/contraction introduced by the cocycle in the fiber. This assumption is essential, as follows from a counterexample found by de la Llave's [73].

Useful tools in this development have been results from analysis that lift the regularity of a continuous real valued function which is assumed to have higher regularity along pairs of transverse Hölder foliations. Many times the foliations are the stable and unstable ones associated to a hyperbolic system. Journé [46] proved the $C^{n,\alpha}$ regularity of a continuous function that is $C^{n,\alpha}$ along two transverse continuous foliations with $C^{n,\alpha}$ leaves. If one

is interested only in C^∞ regularity, a convenient alternative is a result of Hurder, Katok [44]. This has a simpler proof and can be applied to the more general situation in which the function is regular along a web of transverse foliations. A real analytic regularity result along these lines belongs to de la Llave [74]. In certain problems, for example when working with Weyl chamber flows, it is difficult to control the regularity in enough directions to span the whole tangent space. Nevertheless, the tangent space can be generated if one consider higher brackets of good directions. A C^∞ regularity result for this case belongs to Katok, Spatzier [60]. In order to apply this result, the foliations need to be C^∞ not only along the leaves, but also transversally.

An application of the above regularity results is to questions about transverse regularity of the stable and unstable foliations of the geodesic flow on certain C^∞ surfaces of nonpositive curvature. For compact negatively curved C^∞ surfaces, E. Hopf showed that these foliations are C^1, and it follows from the work of Anosov that individual leaves are C^∞. Hurder, Katok [44] showed that once the weak-stable and weak-unstable foliations of a volume-preserving Anosov flow on a compact 3-manifold are C^2, they are C^∞.

Another application of the regularity results is to the study of invariants for C^k conjugacy of hyperbolic systems. By structural stability, a small C^1 perturbation of a hyperbolic system is C^0 conjugate to the unperturbed one. The conjugacy, in general, is only Hölder. If the conjugacy is C^1 then it preserves the eigenvalues of the derivative at the periodic orbit. The following two results describe the invariants of smooth and real analytic conjugacy of low dimensional hyperbolic systems. They are proved in a series of papers written in various combinations by de la Llave, Marco, Moryion [72,74,75,79,80].

Let X, Y be two $C^\infty(C^\omega)$ transitive Anosov vector fields on a compact three-dimensional manifold. If they are C^0 conjugate and the eigenvalues of the derivative at the corresponding periodic orbits are the same, then the conjugating homeomorphism is C^∞ (C^ω). In particular, any C^1 conjugacy is C^∞ (C^ω).

Assume now that f, g are two C^∞ (C^ω) Anosov diffeomorphisms on a compact two dimensional manifold. If they are C^0 conjugate and the eigenvalues of the derivative at the corresponding periodic orbits are the same, then the conjugating diffeomorphism is C^∞ (C^ω). In particular, any C^1 conjugacy is C^∞ (C^ω).

An important direction was initiated by Katok, Spatzier [59] who studied cohomological results over hyperbolic $\mathbb{Z}^k$ or $\mathbb{R}^k$-actions, $k \geq 2$. They show that real valued smooth/Hölder cocycles over typical classes of hyperbolic $\mathbb{Z}^k$ or $\mathbb{R}^k$, $k \geq 2$, actions are smoothly/Hölder cohomologous to constants. These results cover, in particular, actions by hyperbolic automorphisms of a torus, and Weyl chamber flows. The proofs rely on harmonic analysis techniques, such as Fourier transform and group representations for semi-simple Lie groups.

A geometric method for cocycle rigidity was developed in [64]. One constructs a differentiable form using invariant structures along stable/unstable foliations, and the commutativity of the action. The form is exact if and only if the cocycle is cohomologous to a constant one. The method covers actions on nilmanifolds satisfying a condition called TNS (Totally Non-Symplectic). This condition means that the action is higher rank abelian hyperbolic, and that the tangent space is a direct sum of invariant distributions, with each pair of these included in the stable distribution of a hyperbolic element of the action. The method was also applied to small (i. e. close to identity on a set of generators) Lie group valued cocycles. A related paper is [96] which contains rigidity results for cocycles over (TNS) actions with values in compact Lie groups. In this situation the number of cohomology classes is finite. An example of (TNS) action is given by the action of a maximal diagonalizable subgroup of $SL(n, \mathbb{Z})$ on $\mathbb{T}^n$.

Recently Damjanović, Katok [14] developed a new method that was applied to the action of the matrix diagonal group on $\Gamma \setminus SL(n, \mathbb{R})$. They use techniques from [54], where one finds cohomology invariants for cocycles over partially hyperbolic actions that satisfy accessibility property. Accessibility means that one can connect any two points from the manifold supporting the partially hyperbolic dynamical system by transverse piecewise smooth paths included in stable/unstable leaves. This notion was introduced by Brin, Pesin [9] and it is playing a crucial role in the recent surge of activity in the field of partially hyperbolic diffeomorphisms. See [10] for a recent survey of the subject. The cohomology invariants described in [54] are heights of the cocycle over cycles constructed in the base out of pieces inside stable/unstable leaves. They provide a complete set of obstructions for solving the cohomology equation. A new tool introduced in [14] is algebraic K-theory [88]. The method can be extended to cocycles with non-abelian range. In [57] one finds related results for small cocycles with values in a Lie group or the diffeomorphism group of a compact manifold.

The equivalent of the Livshits theorem in the higher-rank setting appears to be a description of the highest cohomology rather than the first cohomology. Indeed, for higher rank partially hyperbolic actions of the torus, the intermediate cohomologies are trivial, while for the high-

est one the closing conditions characterize the cohomology classes. This behavior provides a generalization of Veech cohomological result and of Katok, Spatzier cohomological result for toral automorphisms, and was discovered by A. Katok, S. Katok [52,53].

Flaminio, Forni [27] studied the cohomological equation over the horocycle flow. It is shown that there are infinitely many obstructions to the existence of a smooth solution. Moreover, if these obstructions vanish, then one can solve the cohomological equation. In [28] it is shown a similar result for cocycles over area preserving flows on compact higher-genus surfaces under certain assumptions that hold generically. Mieczkowski [87] extended these techniques and studied the cohomology of parabolic higher rank abelian actions. All these results rely on non-commutative Fourier analysis, more specifically representation theory of $SL(2,\mathbb{R})$ and $SL(2,\mathbb{C})$.

Local Rigidity

Let Γ be a finitely (or compactly) generated group, G a topological group, and $\pi:\Gamma\to G$ a homomorphism. The target of local rigidity theory is to understand the space of perturbations of various homomorphisms π. Trivial perturbations of a homomorphism arise from conjugation by an arbitrary element of G. In order to rule them out, one says that π is *locally rigid* if any nearby homomorphism π', (that is, π' close to π on a finite or compact set of generators of Γ), is conjugate to π by an element $g\in G$, that is, $\pi(\gamma)=g\pi'(\gamma)g^{-1}$ for all $\gamma\in\Gamma$. If G is path-wise connected, one can also consider *deformation rigidity*, meaning that any nearby continuous path of homomorphisms is conjugated to the initial one via a continuous path of elements in G that has an end in the identity.

Initial results on local rigidity are about embeddings of lattices into semi-simple Lie groups. The main results belong to Weil [110,111,112]. He showed that if G is a semi-simple Lie group that is not locally isomorphic to $SL(2,\mathbb{R})$ and if $\Gamma\subset G$ is an irreducible cocompact lattice, then the natural embedding of Γ into G is locally rigid. Earlier results were obtained by Selberg [105], Calabi, Vesentini [12], and Calabi [11]. Selberg proved the local rigidity of the natural embedding of cocompact lattices into $SL(n,\mathbb{R})$. His proof used the dynamics of iterates of matrices, in particular the existence of singular directions, or walls of Weyl chambers, in the maximal diagonalizable subgroups of $SL(n,\mathbb{R})$. Selberg's approach inspired Mostow [90] to use the boundaries at infinity in his proof of strong rigidity of lattices, which in turn was crucial to the development of superrigidity due to Margulis [82]. See Sect. "Global Rigidity" for more details.

Recall that hyperbolic dynamical systems are structurally stable. Thus they are, in a certain sense, locally rigid. We introduce now a precise definition of local rigidity in the infinite-dimensional setup. The fact that for general group actions one needs to consider different regularities for the actions, perturbations and conjugacies is apparent from the description of structural stability for Anosov systems.

A C^k action α of a finitely generated discrete group Γ on a manifold M, that is, a homomorphism $\alpha:\Gamma\to\mathrm{Diff}^k(M)$, is said to be $C^{k,l,p,r}$ *locally rigid* if any C^l perturbation $\tilde\alpha$ which is C^p close to α on a family of generators, is C^r conjugate to α, i.e there exists a C^r diffeomorphism $H:M\to M$ which conjugates $\tilde\alpha$ to α, that is, $H\circ\alpha(g)=\tilde\alpha(g)\circ H$ for all $g\in\Gamma$. Note that for Anosov $\mathbb{Z}$-actions, $C^{1,1,1,0}$ rigidity is known as *structural stability*. One can also introduce the notion of *deformation rigidity* if the initial action and the perturbation are conjugate by a continuous path of diffeomorphisms that has an end coinciding to the identity.

A weaker notion of local rigidity can be defined in the presence of invariant foliations for the initial group action and for the perturbation. The map H is now required to preserve the leaves of the foliations and to conjugate only after factorization by the invariant foliations. The importance of this notion is apparent from the leafwise conjugacy structural stability theorem of Hirsch, Pugh, Shub [41]. See Sect. "Basic Definitions and Examples". Moreover, for Anosov flows this is the natural notion of structural stability, and appears by taking the invariant foliation to be the one-dimensional orbit foliation. For more general actions, of lattices or higher rank abelian groups, this property is often used in combination to cocycle rigidity in order to show local rigidity. We discuss more about this when we review local rigidity results for partially hyperbolic actions.

We summarize now several developments in local rigidity that emerged in the 80s. Initial results [67,115] were about *infinitesimal rigidity*, that is, a weaker version of local rigidity suitable for discrete groups representations in infinite-dimensional spaces of smooth vector fields. Then Hurder [43] proved $C^{\infty,\infty,1,\infty}$ deformation rigidity and Katok, Lewis, Zimmer [55,56,63] proved $C^{\infty,\infty,1,\infty}$ local rigidity of the standard action of $SL(n,\mathbb{Z})$, $n\geq 3$, on the n-dimensional torus. In these results crucial use was made of the presence of an Anosov element in the action. Due to the uniqueness of the conjugacy coming from structural stability, one has a continuous candidate for the conjugacy between the actions. Margulis, Qian [85] used the existence of a spanning family of directions that are hyperbolic for certain elements of the action to show lo-

cal rigidity of partially hyperbolic actions that are not hyperbolic. Another important tool present in many proofs is Margulis and Zimmer superrigidity. These results allow one to produce a measurable conjugacy for the perturbation. Then one shows that the conjugacy has higher regularity using the presence of hyperbolicity. Having enough directions to span the whole tangent space is essential to lift the regularity. A cocycle to which superrigidity can be applied is the derivative cocycle.

The study of local rigidity of partially hyperbolic actions that contain a compact factor was initiated by Nițică, Török [92,94]. Let $n \geq 3$ and $d \geq 1$. Let ρ be the action of $SL(n, \mathbb{Z})$ on $\mathbb{T}^{n+d} = \mathbb{T}^n \times \mathbb{T}^d$ given by $\rho(A)(x, y) = (Ax, y)$, $x \in \mathbb{T}^n$, $y \in \mathbb{T}^d$, $A \in SL(n, \mathbb{Z})$. Then, for $K \geq 1$, [92] shows that ρ is $C^{\infty,\infty,5,K-1}$ deformation rigid. The proof is based on three results in hyperbolic dynamics: the generalization of Livshits's cohomological results to cocycles with values in diffeomorphism groups, the extension of Livshits's result to general Anosov actions, and a version of the Hirsch, Pugh, Shub structural stability theorem improving the regularity of the conjugacy.

Assume now $n \geq 3$ and $K \geq 1$. If ρ is the action of $SL(n, \mathbb{Z})$ on $\mathbb{T}^{n+1} = \mathbb{T}^n \times \mathbb{T}$ given by $\rho(A)(x, y) = (Ax, y)$, $x \in \mathbb{T}^n$, $y \in \mathbb{T}$, $A \in SL(n, \mathbb{Z})$, [94] shows that the action ρ is $C^{\infty,\infty,2,K-1}$ locally rigid. Ingredients in the proof are two rigidity results, one about TNS actions, and one about actions of property (T) groups. A locally compact group has property (T) if the trivial representation is isolated in the Fell topology. This means that if G acts on a Hilbert space unitarily and it has almost invariant vectors, then it has invariant vectors. Hirsch–Pugh–Shub theorem implies that perturbations of abelian partially hyperbolic actions of product type are conjugated to skew-products of abelian Anosov actions via cocycles with values in diffeomorphism groups. In addition, the TNS property implies that the sum of the stable and unstable distributions of any regular element of the perturbation is integrable. The leaves of the integral foliation are closed, covering the base simply. Thus one obtains a conjugacy between the perturbation and a product action. Property (T) is used to show that the conjugacy reduces the perturbed action to a family of perturbations of hyperbolic actions. But the last ones are already known to be conjugate to the hyperbolic action in the base.

Recent important progress in the question of local rigidity of lattice actions was made by Fisher, Margulis [24,25,26]. Their proofs are modeled along the proof of Weil's local rigidity result [112] and use an analog of Hamilton's [39] hard implicit function theorem. Let G be a connected semi-simple Lie group with all simple factors of rank at least two, and $\Gamma \subset G$ a lattice. The main result shows that a volume preserving affine action ρ of G or Γ on a compact smooth manifold X is $C^{\infty,\infty,\infty,\infty}$ locally rigid. Lower regularity results are also available. A component of the proof shows that if Γ is a group with property (T), X a compact smooth manifold, and ρ a smooth action of Γ on X by Riemannian isometries, then ρ is $C^{\infty,\infty,\infty,\infty}$ locally rigid. An earlier local rigidity result for this type of actions by cocompact lattices was obtained by Benveniste [5].

Many lattices act naturally on "boundaries" of type G/P, where G is a semi-simple algebraic Lie group and P is a parabolic subgroup. An example is given by $G = SL(2, \mathbb{R})$ and P the subgroup in G consisting of upper triangular matrices. Local rigidity results for this type of actions were found by Ghys [34], Kanai [50] and Katok, Spatzier [62].

Starting with the work of Katok and Lewis, a related direction was the study of local rigidity for higher rank abelian actions. They prove in [55] the $C^{\infty,\infty,1,\infty}$ local rigidity of the action of a $\mathbb{Z}^n$ maximal diagonalizable (over $\mathbb{R}$) subgroup of $SL(n+1, \mathbb{Z})$, $n \geq 2$, acting on the torus $\mathbb{T}^{n+1}$. These type of results were later pushed forward by Katok, Spatzier [62]. Using the theory of nonstationary normal forms developed in [38] and [37] by Katok, Guysinsky, they proved several local rigidity results. The first one assumes that G is a semisimple Lie group with all simple factors of rank at least two, Γ a lattice in G, N a nilpotent Lie group and Λ a lattice in N. Then any Anosov affine action of Γ on N/Λ is $C^{\infty,\infty,1,\infty}$ locally rigid. Second, let $\mathbb{Z}^d$ be a group of affine transformations of N/Λ for which the derivatives are simultaneously diagonalizable over $\mathbb{R}$ with no eigenvalues on the unit circle. Then the $\mathbb{Z}^d$-action on N/Λ is $C^{\infty,\infty,1,\infty}$ locally rigid. A related result for continuous groups is the $C^{\infty,\infty,1,\infty}$ local rigidity (after factorization by the orbit foliation) of the action of a maximal abelian $\mathbb{R}$-split subgroup in an $\mathbb{R}$-split semi-simple Lie group of real rank at least two on G/Λ, where Λ is a cocompact lattice in G.

One can also study rigidity of higher rank abelian partially hyperbolic actions that are not hyperbolic. Natural examples appear as automorphisms of tori and as variants of Weyl chamber flows. For the case of ergodic actions by automorphisms of a torus, this was investigated using a version of the KAM (Kolmogorov, Arnold, Moser) method by Damianović, Katok [15]. As usual in the KAM method, one starts with a linearization of the conjugacy equation. At each step of the iterative KAM scheme, some twisted cohomological equations are solved. The existence of the solutions is forced by the ergodicity of the action and the higher rank assumptions. Diophantine conditions present in this case allow to control the fixed loss of regu-

larity which is necessary for the convergence of these solutions to a conjugacy.

Global Rigidity

The first remarkable result in global rigidity belongs to Mostow [90]. For G a connected non-compact semi-simple Lie group not locally isomorphic to $SL(2,\mathbb{R})$, and two irreducible cocompact lattices $\Gamma_1, \Gamma_2 \subset G$, Mostow showed that any isomorphism θ from Γ_1 into Γ_2 extends to an isomorphism of G into itself. G has an involution σ whose fixed set is a maximal compact subgroup K. One constructs the symmetric Riemannian space $X = G/K$. To each chamber of X corresponds a parabolic group and these parabolic groups are endowed with a Tits geometry similar to the projective geometry of lines, planes etc. formed in the classical case when $G = PGL(n,\mathbb{R})$. The proof of Mostow's result starts by building a θ-equivariant pseudo-isometric map $\phi\colon G/K_1 \to G/K_2$. The map ϕ induces an incidence preserving θ-equivariant isomorphism ϕ_0 of the Tits geometries. By Tits' generalized fundamental theorem of projective geometry, ϕ_0 is induced by an isomorphism of G. Finally, $\theta(\gamma) = \phi_0 \cdot \gamma \cdot \phi_0^{-1}$ gives the desired conclusion.

The next remarkable result in global rigidity is Margulis' superrigidity theorem. An account of this development can be found in the monograph [82]. For large classes of irreducible lattices in semi-simple Lie groups, this result classifies all finite dimensional representations. Let G be a semi-simple simply connected Lie group of rank higher than two and $\Gamma < G$ an irreducible lattice. Then any linear representation π of Γ is almost the restriction of a linear representation of G. That is, there exists a linear representation π_1 of G and a bounded image representation π_2 of Γ such that $\pi = \pi_1\pi_2$. The possible representations π_2 are also classified by Margulis up to some facts concerning finite image representations. As in the case of Mostow's result, the proof involved the analysis of a map defined on the boundary at infinity. In this case the map is studied using deep results from dynamics like the multiplicativity ergodic theorem of Oseledec [97] or the theory of random walks on groups developed by Furstenberg [31]. An important consequence of Margulis superrigidity result is the arithmeticity of irreducible lattices in connected semi-simple Lie groups of rank higher than two. A basic example of arithmetic lattice can be obtained by taking the integer points in a semi-simple Lie group that is a matrix group, like taking $SL(n,\mathbb{Z})$ inside $SL(n,\mathbb{R})$. Special cases of superrigidity theorems were proved by Corlette [13] and Gromov, Schoen [36] for the rank one groups $Sp(1,n)$ and respectively F_4 using the theory of harmonic maps. A consequence is the arithmeticity of lattices in these groups. Some of these results are put into differential geometric setting in [89].

Margulis supperrigidity result was extended to cocycles by Zimmer. A detailed exposition, including a self contained presentation of several rigidity results of Margulis, can be found in the monograph [113]. We mention here a version of this result that can be found in [24]. Let M be a compact manifold, H a matrix group, $P = M \times H$, and Γ a lattice in a simply connected, semi-simple Lie group with all factors of rank higher that two. Assume that Γ acts on M and H in a way that makes the projection from P to M equivariant. Moreover, the action of Γ on P is measure preserving and ergodic. Then there exists a measurable map $s\colon M \to H$, a representation $\pi\colon G \to H$, a compact subgroup $K < H$ which commute with $\pi(G)$ and a measurable map $k\colon \Gamma \times M \to K$ such that $\gamma \cdot s(m) = k(\gamma, m) \cdot \pi(\gamma) \cdot s(\gamma \cdot m)$. One can easily check from the last equation that k is a cocycle. So, up to a measurable change of coordinates given by the map s, the action of Γ on P is a compact extension via a cocycle of a linear representation of G.

Developing further the method of Mostow for studying the Tits building associated to a symmetric space of non-positive curvature led Ballman, Brin, Eberlein, Spatzier [2,3] to a number of characterizations of symmetric spaces. In particular, they showed that if M is a complete Riemannian manifold of non-positive curvature, finite volume, with simply connected cover, irreducible and of rank at least two, then M is isometric to a symmetric space with the connected component of $\mathrm{Isom}(M)$ having no compact factors.

A topological rigidity theorem has been proved by Farrell, Jones [21]. They showed that if N is a complete connected Riemannian manifold whose sectional curvature lies in a closed interval included in $(-\infty, 0]$, and M is a topological manifold of dimension greater than 5, then any proper homotopy equivalence $f\colon M \to N$ is properly homotopic to a homeomorphism. In particular, if M and N are both compact connected negatively curved Riemannian manifolds with isomorphic fundamental groups, then M and N are homeomorphic.

Likewise to the case of local rigidity, a source of inspiration for results in global rigidity was the theory of hyperbolic systems, in particular their classification. The only known examples of Anosov diffeomorphisms are hyperbolic automorphisms of infranilmanifolds. Moreover, any Anosov diffeomorphism on an infranilmanifold is topologically conjugate to a hyperbolic automorphism [29,78]. It is conjectured that any Anosov diffeomorphism is topo-

logically conjugate to a hyperbolic automorphism of an infranilmanifold. Partial results are obtained in [91], where the conjecture is proved for Anosov diffeomorphisms with codimension one stable/unstable foliation. The proof of the general conjecture eluded efforts done so far. It is not even known if any Anosov diffeomorphism is topologically transitive, that is, if it has a dense orbit. A few positive results are available. Let M be a C^∞ compact manifold endowed with a C^∞ affine connection. Let f be a topologically transitive Anosov diffeomorphism preserving the connection and such that the stable and unstable distributions are C^∞. Then Benoist, Labourie [4] proved that f is C^∞ conjugate to a hyperbolic automorphism of an infranilmanifold.

The situation for Anosov flows is somehow different. As shown in [30], there exist Anosov flows that are not topologically transitive, so a general analog of the conjecture is false. Nevertheless, for the case of codimension one stable or unstable foliation, it is conjectured in [109] that any Anosov flow on a manifold of dimension greater than three admits a global cross-section. This would imply that the flow is topologically conjugate to the suspension of a linear automorphism of a torus.

For actions of groups larger than $\mathbb{Z}$, or $\mathbb{R}$, global classification results are more abundant. A useful strategy in these type of results, which are quite technical, is to start by obtaining a measurable description of the action, most of the time using Margulis–Zimmer superrigidity results, and then use extra assumptions on the action, such as the presence of a hyperbolic element, or the presence of an invariant geometric structure, or both, in order to show that the measurable model is actually continuous or even differentiable. For actions of higher rank Lie groups and their lattices some representative papers are by Katok, Lewis, Zimmer [63] and Goetze, Spatzier [35]. For actions of higher rank abelian groups see Kalinin, Spatzier [49].

Measure Rigidity

Measure rigidity is the study of invariants measures for actions of one parameter and multiparameter abelian groups and semigroups acting on manifolds. Typical situations when interesting rigidity phenomena appear are for one parameter unipotent actions and higher rank hyperbolic actions, discrete or continuous.

A unipotent matrix is one all of whose eigenvalues are one. An important case where the action of a unipotent flow appears is that of the horocycle flow. The invariant measures for it were studied by Furstenberg [33], who showed that the horocycle flow on a compact surface is uniquely ergodic, that is, the ergodic measure is unique. Dani and Smillie [17] extended this result to the case of non-compact surfaces, with the only other ergodic measures appearing being those supported on compact horocycles. An important breakthrough is the work of Margulis [81], who solved a long standing question in number theory, Oppenheim's conjecture. The conjecture is about the density properties of the values of an indefinite quadratic form in three or more variables, provided the form is not proportional to a rational form. The proof of the conjecture is based on the study of the orbits for unipotent flows acting by translations on the homogenous space $SL(n,\mathbb{Z}) \setminus SL(n,\mathbb{R})$. All these results were special cases of the Raghunathan conjecture about the structure of the orbits of the actions of unipotent flows on homogenous spaces. Raghunathan's conjecture was proved in full generality by Ratner [100,101]. Borel, Prasad [8] raised the question of an analog of Raghunathan's conjecture for S-algebraic groups. S-algebraic groups are products of real and p-adic algebraic groups. This was answered independently by Margulis, Tomanov [86] and Ratner [102].

A basic example of higher rank abelian hyperbolic action is given by the action of $S_{m,n}$, the multiplicative semigroup of endomorphisms generated by the multiplication by m and n, two nontrivial integers, on the one dimensional torus $\mathbb{T}^1$. In a pioneering paper [32] Furstenberg showed that for m, n that are not powers of the same integer the action of $S_{m,n}$ has a unique closed, infinite invariant set, namely $\mathbb{T}^1$ itself. Since there are many closed, infinite invariant sets for multiplication by m, and by n, this result shows a remarkable rigidity property of the joint action. Furstenberg's result was generalized later by Berend for other group actions on higher dimensional tori and on other compact abelian groups in [6] and [7].

Furstenberg further opened the field by raising the following question:

Conjecture 1 *Let μ be a $S_{m,n}$-invariant and ergodic probability measure on $\mathbb{T}^1$. Then μ is either an atomic measure supported on a finite union of (rational) periodic orbits or μ is the Lebesgue measure.*

While the statement appears to be simple, proving it has been elusive. The first partial result was given by Lyons [77] under the strong additional assumption that the measure makes one of the endomorphisms generating the action exact. Later Rudolph [103] and Johnson [45] weaken the exactness assumption and proved that μ must be the Lebesgue measure provided that multiplication by m (or multiplication by n) has positive entropy with respect to μ. Their results have been proved again using slightly different methods by Feldman [22]. A further extension was given by Host [42].

Katok proposed another example for which measure rigidity can be fruitfully tested, the $\mathbb{Z}^2$-action induced by two commuting hyperbolic automorphisms on the torus $\mathbb{T}^3$. An example of such action is shown in Sect. "Basic Definitions and Examples". One can also consider the action induced by a $\mathbb{Z}^{n-1}$ maximal abelian group of hyperbolic automorphisms acting on the torus $\mathbb{T}^n$. Katok, Spatzier [61] developed a more geometric technique allowing to prove measure rigidity for these actions if they have an element acting with positive entropy. Their technique can be applied in higher generality if the action is irreducible in a strong sense and, in addition, it has individual ergodic elements or it is TNS. See also [47]. This method is based on the study of conditional measures induced by the invariant measure on certain invariant foliations that appear naturally in the presence of a hyperbolic action. Besides stable and unstable foliations, one can also consider various intersections of them. Einsiedler, Lindenstrauss [19] were able to eliminate the ergodicity and TNS assumptions.

Yet another interesting example of higher rank abelian action is given by Weyl chamber flows. These do not satisfy the TNS condition. Einsiedler, Katok [18] proved that if G is $SL(n,\mathbb{R})$, $\Gamma \subset G$ is a lattice, H is the subgroup of positive diagonal matrices in G, and μ a H-invariant and ergodic measure on G/Γ such that the entropy of μ with respect to all one parameter subgroups of H is positive, then μ is the G invariant measure on G/Γ.

These results are useful in the investigation of several deep questions in number theory. Define $X = SL(3,\mathbb{Z}) \setminus SL(3,\mathbb{R})$ the diagonal subgroup of matrices with positive entries in $SL(3,\mathbb{R})$. This space is not compact but is endowed with a unique translation invariant probability measure. The diagonal subgroup

$$H = \left\{ \begin{pmatrix} e^s & 0 & 0 \\ 0 & e^t & 0 \\ 0 & 0 & e^{-s-t} \end{pmatrix} : s, t \in \mathbb{R} \right\}$$

acts naturally on X by right translations. It was conjectured by Margulis [83] that any compact H-invariant subset of X is a union of compact H-orbits. A positive solution to this conjecture implies a long standing conjecture of Littlewood:

Conjecture 2 *Let $\|x\|$ denote the distance between the real number x and the closest integer. Then*

$$\liminf_{n\to\infty} n\|n\alpha\|\|n\beta\| = 0 \tag{4}$$

for any real numbers α and β.

A partial result was obtained by Einsiedler, Katok, Lindenstrauss [20] who proved that the set of pairs $(\alpha,\beta) \in \mathbb{R}^2$ for which (4) is not satisfied has Hausdorff dimension zero. Applications of these techniques to questions in quantum ergodicity were found by Lindenstrauss [69].

A current direction in measure rigidity is attempting to classify the invariant measures under rather weak assumptions about the higher rank abelian action, like the homotopical data for the action. Kalinin and Katok [48] proved that any $\mathbb{Z}^k$, $k \geq 2$, smooth action α on a $k+1$-dimensional torus whose elements are homotopic to the corresponding elements of an action by hyperbolic automorphisms preserves an absolutely continuous measure.

Future Directions

An important open problem in differential rigidity is to find invariants for the C^k conjugacy of the perturbations of higher dimensional hyperbolic systems. For Anosov diffeomorphisms, de la Llave counterexample [73] shows that this extension is not possible for a four dimensional example that appears as a direct product of two dimensional Anosov diffeomorphisms. Indeed, there are C^∞ perturbations of the product that are only C^k conjugate to the unperturbed system for any $k \geq 1$. In the positive direction, Katok conjectured that generalizations are possible for the diffeomorphism induced by an irreducible hyperbolic automorphism of a torus. One can also investigate this question for Anosov flows.

Examples of higher rank Lie groups can be obtained by taking products of rank one Lie groups. Many actions of irreducible lattices in these type of groups are believed to be locally rigid, but the techniques available so far cannot be applied. A related problem is to study local rigidity in low regularity classes, for example the local rigidity of homomorphisms from higher rank lattices into homeomorphism groups. More problems related to local rigidity can be found in [23].

An important problem in global rigidity, emphasized by Katok and Spatzier, is to show that, up to differentiable conjugacy, any higher rank Anosov or partially hyperbolic action is algebraic under the assumption that it is sufficiently irreducible. The irreducibility assumption is needed in order to exclude actions obtained by successive application of products, extensions, restrictions and time changes from basic ingredients which include some rank one actions.

Another problem of current research in measure rigidity is to develop a counterpart of Ratner's theory for the case of actions by hyperbolic higher rank abelian groups on homogenous spaces. It was conjectured by Katok, Spatzier [61] that the invariant measures for such actions given by toral automorphisms or Weyl chamber flows are

essentially algebraic, that is, supported on closed orbits of connected subgroups. Margulis in [84] extended this conjecture to a rather general setup addressing both the topological and measurable aspects of the problem. More details about actions on homogenous spaces, as well as connections to diophantine approximation, can be found in the article ► Ergodic Theory on Homogeneous Spaces and Metric Number Theory by Kleinbock.

Acknowledgment

This research was supported in part by NSF Grant DMS-0500832.

Bibliography

Primary Literature

1. Anosov DV (1967) Geodesic flows on closed Riemannian manifolds with negative curvature. Proc Stek Inst 90:1–235
2. Ballman W, Brin M, Eberlein P (1985) Structure of manifolds of non-negative curvature. I. Ann Math 122:171–203
3. Ballman W, Brin M, Spatzier R (1985) Structure of manifolds of non-negative curvature. II. Ann Math 122:205–235
4. Benoist Y, Labourie F (1993) Flots d'Anosov á distribuitions stable et instable différentiables. Invent Math 111:285–308
5. Benveniste EJ (2000) Rigidity of isometric lattice actions on compact Riemannian manifolds. Geom Func Anal 10:516–542
6. Berend D (1983) Multi-invariant sets on tori. Trans AMS 280:509–532
7. Berend D (1984) Multi-invariant sets on compact abelian groups. Trans AMS 286:505–535
8. Borel A, Prasad G (1992) Values of isotropic quadratic forms at S-integral points. Compositio Math 83:347–372
9. Brin MI, Pesin YA (1974) Partially hyperbolic dynamical systems. Izvestia 38:170–212
10. Burns K, Pugh C, Shub M, Wilkinson A (2001) Recent results about stable ergodicity. In: Katok A, Pesin Y, de la Llave R, Weiss H (eds) Smooth ergodic theory and its applications, Seattle, 1999. Proc Symp Pure Math, vol 69. AMS, Providence, pp 327–366
11. Calabi E (1961) On compact Riemannian manifolds with constant curvature. I. Proc Symp Pure Math, vol 3. AMS, Providence, pp 155–180
12. Calabi E, Vesentini E (1960) On compact, locally symmetric Kähler manifolds. Ann Math 17:472–507
13. Corlette K (1992) Archimedian superrigidity and hyperbolic geometry. Ann Math 135:165–182
14. Damianović D, Katok A (2005) Periodic cycle functionals and cocycle rigidity for certain partially hyperbolic actions. Disc Cont Dynam Syst 13:985–1005
15. Damianović D, Katok A (2007) Local rigidity of partially hyperbolic actions KAM, I.method and $\mathbb{Z}^k$-actions on the torus. Preprint available at http://www.math.psu.edu/katok_a
16. Dani SG, Margulis GA (1990) Values of quadratic forms at integer points: an elementary approach. Enseign Math 36:143–174
17. Dani SG, Smillie J (1984) Uniform distributions of horocycle orbits for Fuchsian groups. Duke Math J 51:185–194
18. Einsiedler M, Katok A (2003) Invariant measures on G/Γ for split simple Lie groups. Comm Pure Appl Math 56:1184–1221
19. Einsiedler M, Lindenstrauss E (2003) Rigidity properties of $\mathbb{Z}^d$-actions on tori and solenoids. Elec Res Ann AMS 9:99–110
20. Einsiedler M, Katok A, Lindenstrauss E (2006) Invariant measures and the set of exceptions to Littlewood conjecture. Ann Math 164:513–560
21. Farrell FT, Jones LE (1989) A topological analog of Mostow's rigidity theorem. J AMS 2:237–370
22. Feldman J (1993) A generalization of a result of R Lyons about measures on [0,1). Isr J Math 81:281–287
23. Fisher D (2006) Local rigidity of group actions: past, present, future. In: Dynamics, Ergodic Theory and Geometry (2007). Cambridge University Press
24. Fisher D, Margulis GA (2003) Local rigidity for cocycles. In: Surv Diff Geom VIII. International Press, Cambridge, pp 191–234
25. Fisher D, Margulis GA (2004) Local rigidity of affine actions of higher rank Lie groups and their lattices. 2003
26. Fisher D, Margulis GA (2005) Almost isometric actions, property T, and local rigidity. Invent Math 162:19–80
27. Flaminio L, Forni G (2003) Invariant distributions and time averages for horocycle flows. Duke Math J 119:465–526
28. Forni G (1997) Solutions of the cohomological equation for area-preserving flows on compact surfaces of higher genus. Ann Math 146:295–344
29. Franks J (1970) Anosov diffeomorphisms. In: Chern SS, Smale S (eds) Global Analysis (Proc Symp Pure Math, XIV, Berkeley 1968). AMS, Providence, pp 61–93
30. Franks J, Williams R (1980) Anomalous Anosov flows. In: Global theory of dynamical systems, Proc Inter Conf Evanston, 1979. Lecture Notes in Mathematics, vol 819. Springer, Berlin, pp 158–174
31. Furstenberg H (1963) A Poisson formula for semi-simple Lie groups. Ann Math 77:335–386
32. Furstenberg H (1967) Disjointness in ergodic theory, minimal sets, and a problem in Diophantine approximation. Math Syst Theor 1:1–49
33. Furstenberg H (1973) The unique ergodicity of the horocycle flow. In: Recent advances in topological dynamics, Proc Conf Yale Univ, New Haven 1972. Lecture Notes in Mathematics, vol 318. Springer, Berlin, pp 95–115
34. Ghys E (1985) Actions localement libres du groupe affine. Invent Math 82:479–526
35. Goetze E, Spatzier R (1999) Smooth classification of Cartan actions of higher rank semi-simple Lie groups and their lattices. Ann Math 150:743–773
36. Gromov M, Schoen R (1992) Harmonic maps into singular spaces and p-adic superrigidity for lattices in groups of rank one. Publ Math IHES 76:165–246
37. Guysinsky M (2002) The theory of nonstationary normal forms. Erg Th Dyn Syst 22:845–862
38. Guysinsky M, Katok A (1998) Normal forms and invariant geometric structures for dynamical systems with invariant contracting foliations. Math Res Lett 5:149–163
39. Hamilton R (1982) The inverse function theorem of Nash and Moser. Bull AMS 7:65–222
40. Helgason S (1978) Differential Geometry, Lie Groups and Symmetric Spaces. Academic Press, New York
41. Hirsch M, Pugh C, Shub M (1977) Invariant Manifolds. Lecture Notes in Mathematics, vol 583. Springer, Berlin

42. Host B (1995) Nombres normaux, entropie, translations. Isr J Math 91:419–428
43. Hurder S (1992) Rigidity of Anosov actions of higher rank lattices. Ann Math 135:361–410
44. Hurder S, Katok A (1990) Differentiability, rigidity and Godbillon-Vey classes for Anosov flows. Publ Math IHES 72:5–61
45. Johnson AS (1992) Measures on the circle invariant under multiplication by a nonlacunary subsemigroup of integers. Isr J Math 77:211–240
46. Journé JL (1988) A regularity lemma for functions of several variables. Rev Mat Iberoam 4:187–193
47. Kalinin B, Katok A (2002) Measurable rigidity and disjointness for $\mathbb{Z}^k$-actions by toral automorphisms. Erg Theor Dyn Syst 22:507–523
48. Kalinin B, Katok A (2007) Measure rigidity beyond uniform hyperbolicity: invariant measures for Cartan actions on tori. J Modern Dyn 1:123–146
49. Kalinin B, Spatzier R (2007) On the classification of Cartan actions. Geom Func Anal 17:468–490
50. Kanai M (1996) A new approach to the rigidity of discrete group actions. Geom Func Anal 6:943–1056
51. Katok A, Hasselblatt B (1995) Introduction to the modern theory of dynamical systems. Encyclopedia of Mathematics and its Applications 54. Cambridge University Press, Cambridge
52. Katok A, Katok S (1995) Higher cohomology for abelian groups of toral automorphisms. Erg Theor Dyn Syst 15:569–592
53. Katok A, Katok S (2005) Higher cohomology for abelian groups of toral automorphisms. II. The partially hyperbolic case, and corrigendum. Erg Theor Dyn Syst 25:1909–1917
54. Katok A, Kononenko A (1996) Cocycles' stability for partially hyperbolic systems. Math Res Lett 3:191–210
55. Katok A, Lewis J (1991) Local rigidity for certain groups of toral automorphisms. Isr J Math 75:203–241
56. Katok A, Lewis J (1996) Global rigidity results for lattice actions on tori and new examples of vol preserving actions. Isr J Math 93:253–280
57. Katok A, Niţică V (2007) Rigidity of higher rank abelian cocycles with values in diffeomorphism groups. Geometriae Dedicata 124:109–131
58. Katok A, Niţică V () Differentiable rigidity of abelian group actions. Cambridge University Press (to appear)
59. Katok A, Spatzier R (1994) First cohomology of Anosov actions of higher rank abelian groups and applications to rigidity. Publ Math IHES 79:131–156
60. Katok A, Spatzier R (1994) Subelliptic estimates of polynomial differential operators and applications to rigidity of abelian actions. Math Res Lett 1:193–202
61. Katok A, Spatzier R (1996) Invariant measures for higher-rank abelian actions. Erg Theor Dyn Syst 16:751–778; Katok A, Spatzier R (1998) Corrections to: Invariant measures for higher-rank abelian actions.; (1996) Erg Theor Dyn Syst 16:751–778; Erg Theor Dyn Syst 18:503–507
62. Katok A, Spatzier R (1997) Differential rigidity of Anosov actions of higher rank abelian groups and algebraic lattice actions. Trudy Mat Inst Stek 216:292–319
63. Katok A, Lewis J, Zimmer R (1996) Cocycle superrigidity and rigidity for lattice actions on tori. Topology 35:27–38
64. Katok A, Niţică V, Török A (2000) Non-abelian cohomology of abelian Anosov actions. Erg Theor Dyn Syst 2:259–288
65. Katok A, Katok S, Schmidt K (2002) Rigidity of measurable structure for $\mathbb{Z}^d$-actions by automorphisms of a torus. Comm Math Helv 77:718–745
66. Kazhdan DA (1967) On the connection of a dual space of a group with the structure of its closed subgroups. Funkc Anal Prilozen 1:71–74
67. Lewis J (1991) Infinitezimal rigidity for the action of $SL_n(\mathbb{Z})$ on $\mathbb{T}^n$. Trans AMS 324:421–445
68. Lindenstrauss E (2005) Rigidity of multiparameter actions. Isr Math J 149:199–226
69. Lindenstrauss E (2006) Invariant measures and arithmetic quantum unique ergodicity. Ann Math 163:165–219
70. Livshits A (1971) Homology properties of Y-systems. Math Zametki 10:758–763
71. Livshits A (1972) Cohomology of dynamical systems. Izvestia 6:1278–1301 he Livšic cohomology equation. Ann Math 123:537–611
72. de la Llave R (1987) Invariants for smooth conjugacy of hyperbolic dynamical systems. I. Comm Math Phys 109:369–378
73. de la Llave R (1992) Smooth conjugacy and S-R-B measures for uniformly and non-uniformly hyperbolic dynamical systems. Comm Math Phys 150:289–320
74. de la Llave R (1997) Analytic regularity of solutions of Livshits's cohomology equation and some applications to analytic conjugacy of hyperbolic dynamical systems. Erg Theor Dyn Syst 17:649–662
75. de la Llave R, Moriyon R (1988) Invariants for smooth conjugacy of hyperbolic dynamical systems. IV. Comm Math Phys 116:185–192
76. de la Llave R, Marco JM, Moriyon R (1986) Canonical perturbation theory of Anosov systems and regularity results for the Livšic cohomology equation. Ann Math 123:537–611
77. Lyons R (1988) On measures simultaneously 2- and 3-invariant. Isr J Math 61:219–224
78. Manning A (1974) There are no new Anosov diffeomorphisms on tori. Amer J Math 96:422–429
79. Marco JM, Moriyon R (1987) Invariants for smooth conjugacy of hyperbolic dynamical systems. I. Comm Math Phys 109:681–689
80. Marco JM, Moriyon R (1987) Invariants for smooth conjugacy of hyperbolic dynamical systems. III. Comm Math Phys 112:317–333
81. Margulis GA (1989) Discrete subgroups and ergodic theory. In: Number theory, trace formulas and discrete groups, Oslo, 1987. Academic Press, Boston, pp 277–298
82. Margulis GA (1991) Discrete subgroups of semi-simple Lie groups. Springer, Berlin
83. Margulis GA (1997) Oppenheim conjecture. In: Fields Medalists Lectures, vol 5. World Sci Ser 20th Century Math. World Sci Publ, River Edge, pp 272–327
84. Margulis GA (2000) Problems and conjectures in rigidity theory. In: Mathematics: frontiers and perspectives. AMS, Providence, pp 161–174
85. Margulis GA, Qian N (2001) Local rigidity of weakly hyperbolic actions of higher rank real Lie groups and their lattices. Erg Theor Dyn Syst 21:121–164
86. Margulis GA, Tomanov G (1994) Invariant measures for actions of unipotent groups over local fields of homogenous spaces. Invent Math 116:347–392

87. Mieczkowski D (2007) The first cohomology of parabolic actions for some higher-rank abelian groups and representation theory. J Modern Dyn 1:61–92
88. Milnor J (1971) Introduction to algebraic *K*-theory. Princeton University Press, Princeton
89. Mok N, Siu YT, Yeung SK (1993) Geometric superrigidity. Invent Math 113:57–83
90. Mostow GD (1973) Strong rigidity of locally symmetric spaces. Ann Math Studies 78. Princeton University Press, Princeton
91. Newhouse SE (1970) On codimension one Anosov diffeomorphisms. Amer J Math 92:761–770
92. Niţică V, Török A (1995) Cohomology of dynamical systems and rigidity of partially hyperbolic actions of higher rank lattices. Duke Math J 79:751–810
93. Niţică V, Török A (1998) Regularity of the transfer map for cohomologous cocycles. Erg Theor Dyn Syst 18:1187–1209
94. Niţică V, Török A (2001) Local rigidity of certain partially hyperbolic actions of product type. Erg Theor Dyn Syst 21:1213–1237
95. Niţică V, Török A (2002) On the cohomology of Anosov actions. In: Rigidity in dynamics and geometry, Cambridge, 2000. Springer, Berlin, pp 345–361
96. Niţică V, Török A (2003) Cocycles over abelian TNS actions. Geom Ded 102:65–90
97. Oseledec VI (1968) A multiplicative ergodic theorem. Characteristic Lyapunov, exponents of dynamical systems. Trudy Mosk Mat Obsc 19:179–210
98. Parry W (1999) The Livšic periodic point theorem for nonabelian cocycles. Erg Theor Dyn Syst 19:687–701
99. Pugh C, Shub M (1972) Ergodicity of Anosov actions. Invent Math 15:1–23
100. Ratner M (1991) On Ragunathan's measure conjecture. Ann Math 134:545–607
101. Ratner M (1991) Ragunathan's topological conjecture and distributions of unipotent flows. Duke Math J 63:235–280
102. Ratner M (1995) Raghunathan's conjecture for Cartesians products of real and p-adic Lie groups. Duke Math J 77:275–382
103. Rudolph D (1990) $\times 2$ and $\times 3$ invariant measures and entropy. Erg Theor Dyn Syst 10:395–406
104. Schmidt K (1999) Remarks on Livšic' theory for nonabelian cocycles. Erg Theor Dyn Syst 19:703–721
105. Selberg A (1960) On discontinuous groups in higher-dimensional symmetric spaces. In: Contributions to function theory. Inter Colloq Function Theory, Bombay. Tata Institute of Fundamental Research, pp 147–164
106. Smale S (1967) Differentiable dynamical systems. Bull AMS 73:747–817
107. Spatzier R (1995) Harmonic analysis in rigidity theory. In: Ergodic theory and its connections with harmonic analysis. Alexandria, 1993. London Math Soc Lect Notes Ser, vol 205. Cambridge University Press, Cambridge, pp 153–205
108. Veech WA (1986) Periodic points and invariant pseudomeasures for toral endomorphisms. Erg Theor Dyn Syst 6:449–473
109. Verjovsky A (1974) Codimension one Anosov flows. Bul Soc Math Mex 19:49–77
110. Weil A (1960) On discrete subgroups of Lie groups. I. Ann Math 72:369–384
111. Weil A (1962) On discrete subgroups of Lie groups. II. Ann Math 75:578–602
112. Weil A (1964) Remarks on the cohomology of groups. Ann Math 80:149–157
113. Zimmer R (1984) Ergodic theory and semi-simple groups. Birhhäuser, Boston
114. Zimmer R (1987) Actions of semi-simple groups and discrete subgroups. Proc Inter Congress of Math (1986). AMS, Providence, pp 1247–1258
115. Zimmer R (1990) Infinitesimal rigidity of smooth actions of discrete subgroups of Lie groups. J Diff Geom 31:301–322

Books and Reviews

de la Harpe P, Valette A (1989) La propriété (T) de Kazhdan pour les groupes localement compacts. Astérisque 175
Feres R (1998) Dynamical systems and semi-simple groups: An introduction. Cambridge Tracts in Mathematics, vol 126. Cambridge University Press, Cambridge
Feres R, Katok A (2002) Ergodic theory and dynamics of *G*-spaces. In: Handbook in Dynamical Systems, 1A. Elsevier, Amsterdam, pp 665–763
Gromov M (1988) Rigid transformation groups. In: Bernard D, Choquet-Bruhat Y (eds) Géométrie Différentielle (Paris, 1986). Hermann, Paris, pp 65–139; Travaux en Cours. 33
Kleinbock D, Shah N, Starkov A (2002) Dynamics of subgroup actions on homogeneous spaces of Lie groups and applications to number theory. In: Handbook in Dynamical Systems, 1A. Elsevier, Amsterdam, pp 813–930
Knapp A (2002) Lie groups beyond an introduction, 2nd edn. Progress in Mathematics, 140. Birkhäuser, Boston
Raghunathan MS (1972) Discrete subgroups of Lie groups. Springer, Berlin
Witte MD (2005) Ratner's theorems on unipotent flows. Chicago Lectures in Mathematics. University of Chicago Press, Chicago

Evacuation as a Communication and Social Phenomenon

Douglas Goudie
Australian Centre for Disaster Studies, School of Earth and Environmental Science, James Cook University, Townsville, Australia

Article Outline

Glossary
Definition of the Subject
Introduction
Concepts, Language and Mathematically Modeling the Propensity to Evacuate
Mathematical Modeling
Effective Risk Communication
Integrating Theory and Implementation
Institutional Barriers to Greater Community Self-Help
Experiences and Lessons – Some Case Studies
Discussion
Acknowledgment
Bibliography

Glossary

Community A group of neighbors or people with a commonality of association and generally defined by location, shared experience, or function [59].

Community empowerment Internally and externally nurtures a community to accept that residents live in a hazard zone, and they choose to do things as a group to maximize their safety.

Community safety group Existing community groups (such as Neighborhood Watch) and individuals, working with formal response organizations form a coherent affiliation in and near a hazard zone, to help maximize safety and care for all community members.

Disaster The interface between an extreme physical event and a vulnerable human population [81].

Disaster lead time The time taken from first detection of a natural disaster threat to the likely time of impact on humans or human structures.

Disaster threat A natural extreme event which may impact on a community.

Effective risk communication . That which motivates people to maximize their own safety.

Emergency An actual or imminent event which endangers or threatens to endanger life, property or the environment, and which requires a significant and coordinated response [55].

Evacuation People relocating to safely escape hazardous disaster impacts. To move from a high danger zone to relative safety.

Hazard A source of potential harm or a situation with a potential to cause loss. A situation or condition with potential for loss or harm to the community or environment [55]. Hazard is synonymous with 'source of risk' [25].

Hazard zone Defined geographic areas which may be subject to a natural disaster impact of flood, bushfire, storm surge, destructive winds, earthquake, landslide or damaging hail. Hazard zones include major accident sites, including industrial, transport or mining precincts; or biological or terrorist threat or impact, or from-source predicted area(s) of pandemic spread.

Mitigation Any efforts taken which may reduce the impact of a threat.

Prevention Measures to eliminate or reduce the incidence or severity of emergencies [55].

Ramp-up preparations The final set of preparations and precautionary evacuations taken ahead of a forecast disaster impact. This includes earlier final actions than precipitated by formal organizations.

Risk treatment options Measures that modify the characteristics of hazards, communities and environments to reduce risk, e. g. prevention, preparedness, response and recovery [55].

Vulnerability comprises 'resilience' and 'susceptibility'. 'Resilience' is related to 'existing controls' and the capacity to reduce or sustain harm. 'Susceptibility' is related to 'exposure' [25].

Definition of the Subject

This article intends to show how system and complexity science can contribute to an understanding and improvement of evacuation processes, especially considering the roles of engaged communities at risk, the concepts of community self-help, and clear communication about local threats and remedies.

This article shows researchers in Complexity and Systems Science (CSS) a social sciences approach to maximize effective and precautionary evacuation, maximize safety, minimize loss and speed full recovery. The computational and analytical modeling tools of CSS may be considered to apply to a complex interaction of community awareness, inclination to accept the reality of a natural disaster threat, along with achieving background and final preparations to maximize safety and recovery from a natural disaster impact. This article may stimulate CSS researchers to develop detailed models of the complex systems and complexity of melding information from Weather Bureaus and Disaster Managers, via contacts and intervening media to communities at risk, with the shared social goal of maximizing safety. This social sciences task requires cross-disciplinary approaches of respect and response.

The old disaster management model lacked the predictive and rapid communication systems now available and developing in disaster predictive models (such as a flood maps). An approach to modeling the great complexity of human behavior responding to threat is provided. Such a model must include people's prior knowledge of a threat type, and consider such fine detail as the overarching language used in a country with threat zones, and the dominant languages of all under threat. It is hoped this article stimulates CSS models to further engage in this social good of helping people get safe and stay safe through natural disasters by providing predictive tool to Authorities to better inform and encourage those at risk to action, including the possible need for precautionary evacuations ahead of a predicted impact.

Disaster management in Australia, and increasingly, globally, is focused on mitigation as part of a 'threat continuum', from acceptance that some locations are vulner-

able to a hazard impact, through to recovery [13]. Emergency warnings and a possible need to evacuate are embedded as 'spikes' on that continuum. Thus, this article stresses the importance of developing ways; incentives, to mobilize aware at-risk community members to precautionary self-evacuation. For this to happen, people need to know and internalize the reality that they are in a hazard zone.

Thus, in the cost-effective philosophy of engendering self-help, the process of understanding the complexity of achieving the shared social goal in maximizing safety and minimizing loss is to engender creation of empowered communities with a high motivation for safety-oriented and precautionary action. This is likely to lead to minimized loss and disruption, and maximized recovery. This article details many elements of that process, and invites detailed development of the Sustainability Implementation Research to achieve that goal through CSS.

To model the path to collective safety, the complexity of the dynamics at play need to be clarified: impact preparedness, including possible evacuation, is a communication and social issue.

This article demonstrates that acknowledgment of hazard zones, developing community acceptance of threat and needed action needs to be at the individual, household and community levels. Evacuation modeling is needed only for those whose homes may be at real threat of a disaster impact. For those living in a hazard zone, a fully informed community, who have internalized the reality of the threat and have worked for maximum background preparation, and have mechanisms to receive alerts and warnings of a looming threat; a community predisposed to precautionary evacuations will result.

Capturing this complexity is the challenge for modelers. Evacuation is about hazard zone residents actively monitoring a looming threat via refined communication channels detailed in this article, within a developed social predisposition to act. Some examples are given. For consideration by scientists and students internationally, this article introduces the Communication Safety Triangle and the Seven Steps to Community Safety on the Preparedness Continuum, within the new research frame of Sustainability Implementation Research (SIR).

Introduction

The purpose of this article is to share with modelers and complexity and system scientists the social and communication issues of modeling effective safety strategies to a natural disaster threat. It is hoped, through the approaches and processes described in this article, that modelers will more clearly link physical threats with warnings and community engagement.

This article first looks at the definitions and language used in risk communication and effective warnings, leading to informal and formal evacuations, then considers some Australian policies relating to emergency management. Theories related to risk communication are presented, with examples of evacuation issues provided from Indigenous communities, and from non English speaking households. The needed conceptual shift to self-help is placed within the larger theoretical frame of paradigms and paradigms shifts.

An example of including residents to internalize threats is given, followed by a more general example of transport evacuation.

A discussion of international evacuation issues precedes a broader view of some of the institutional barriers which may restrict the uptake of the seven step approach to an aware, informed community, relying on accurate information and choosing to self evacuate as a precaution. These issues are discussed, considering effective ways of allowing people to know that they are at risk so they are inclined to evacuate themselves, as a practice. These approaches can be used or tested by other scholars. Recommendations and a summary of the key issues to maximize voluntary and safe evacuations finish this article.

Some Key Evacuation Issues

In an era of increasing social self-help [7,55] and community empowerment [13,25,26], a part of global efforts to embrace Ecologically Sustainable Development [8,62] is to increasingly see evacuation as a social phenomenon.

This article provides a conceptual frame, the Communication Safety Triangle (CST) which includes responsible media telling people at risk what they need to know and *seven steps to community safety* (7SCS). The 7SCS help guide emergency managers and modelers to treat the possible need for evacuation as a decision-making process where the community should be aware of the potential threats and receive clear, detailed and reliable information on the possible need to evacuate, so most residents in a high impact zone self-evacuate in a precautionary way, as a practice. Examples provided in this article illustrate this new, *sustainability implementation research* (SIR) approach to disaster management.

Warnings precede a perceived need to evacuate. In the USA, the need for an integrative approach to warnings is identified: "There is a major need for better coordination among the warning providers, more effective delivery mechanisms, better education of those at risk, and new

ways for building partnerships among the many public and private groups involved" [63].

Sustainability implementation is the way to achieve a sustainable future. Disaster management, effective risk communication and community self-help provide a stark and comprehensible example of what sustainability implementation means, how it will benefit societies, and will help channel us into a safer, more sustainable future. Within this approach, scientists and students become major agents for sustainability implementation, as models can illustrate the needed paths to achievement.

Evacuation Overview

There are three types of disasters which may require evacuation: human induced and natural disasters with and without lead times. This article is focused on precautionary evacuations ahead of natural disasters with lead times (Table 1). The research and conceptual frames draw on and are applicable to all communities in hazard zones. Defining, modeling and effectively communicating to at-risk residents and travelers about specific geographic hazards and safety-oriented behaviors are core elements of successful precautionary evacuations.

Within Table 1, evacuation may be a response to an emerging hazard of indicated intensity, direction and speed. This article advocates development of informed and directed communities which will actively respond to a communicated threat, with the vulnerable moving themselves early to places of safety, or being helped by other community members to move.

Concepts, Language and Mathematically Modeling the Propensity to Evacuate

Precautionary self evacuation pivots on knowing who is at risk. With the help of researchers, modelers; planning and community involvement can prepare people and their valuables to minimize loss. Very public hazard maps will help people internalize that they are in a hazard zone, and what they should do. The concept is not new or alarming: every public building has emergency exit routes marked, with all that implies. Air flight comes with the mandatory emergency preparation presentation. Minute or recurrent risks to where we live or travel to should be no less public, nor more alarming than a fire drill in a public building.

Hazard Types – Little or Considerable Warning Times

Hazards may or may not have lead times (Table 1). Sudden onset threats include: tsunami, earthquake, major eruption, major toxic spill or discharge, mine disaster, terrorist threat or attack. Signaled threats include: – cyclone, flood, fire and destructive winds. This article focuses on disasters with sufficient warning periods to be able to evacuate the vulnerable away from the predicted worst impact areas. Table 1 that with sudden onset impacts, sheltering to sur-

Evacuation as a Communication and Social Phenomenon, Table 1
Evacuation decision matrix – short to long warning times

Evacuation decision matrix, Evacuation around *sudden onset impact*			
Hazard:	Landslide	Earthquake	Tsunami
Possible safety-oriented response	Stay in strong structure. Move across slope as soon as possible.	Get into the open or shelter under strong structure.	Immediately flee to higher ground.

Precautionary Evacuation (PE) decisions with lead time – signaled threats			
Considerations for evacuation decision	Hazard		
	Destructive wind/cyclone	Fire	Flood
1. Vulnerability of present environment	If likely to be in a storm surge, must PE; If shelter weak, must PE	House material, surrounds, water available. If poor, PE	May be inundated= PE; may be cut off, consider PE
2. If 1 OK: vulnerability of individuals (e. g. weak)	PE first	Asthmatics and less able: PE	Judgments of flood height. If in any doubt, PE
3a. Distance to safe shelter	The further, PE earlier	The further, PE earlier	The further, PE earlier
3b. Safety along exit route	Know in preparation	Know in preparation	Know in preparation
3c. Means of travel	Reliability and suitability	Reliability and suitability	Reliability and suitability
4. Community cohesion	Help available	Help available	Help available

vive the first few minutes is critical, then moving to open; stable ground is important to avoid further aftershocks or landslide. Always take direction from local authorities. For natural disasters with lead times, Table 1 shows that, if you will clearly be safe where you are, stay put, and prepare as best you can. If, in the worst case, you may not be safe, move early (Table 1). The vulnerability of individuals needs to be considered. In a bush fire, for instance, the general advice is: if the property is well prepared [7], stay and defend. If, however, you may not be able to cope with the psychological terror of staying through the fire front, fully prepare your property, and then leave early (Table 1). If the exit route may be blocked by flood waters, or by dense smoke or fire, evacuation needs to precede that obstruction. In the 'background' phase of community disaster preparation, all such possible obstacles to a clear escape route (including gridlock congestion) need to be factored in to the timing of precautionary evacuation (Table 1).

Finally, as seen in the Woodgate Beach example of Sect. "Experiences and Lessons – Some Case Studies", the level of community support in ensuring all residents are aware of and prepared for the 'ramp-up' phase of a possible disaster impact, and receive the warning as soon as possible, the able-bodied will help the less able to get out of harm's way early in the threat period, as a safety-oriented practice, minimizing the demands of the formal response groups as the likelihood of impact increases.

All effective communication involves sending signals and having them received and processed, then incorporated into the receiver's world view [27,40,41,76,81,82]. Effective risk communication [4,13], Sect. "Effective Risk Communication", motivates people to act to maximize their own safety. This may not happen if people have not internalized that the threat is real. They are in denial. Alternatively, people may be ignorant of the threat, or why it should be taken seriously. Salter [77] categorizes ignorance from pure ignorance to acts of ignoring.

Disaster Definitions

A disaster may be seen as a negative impact of a hazard on a community as measure of vulnerability. The language of disaster mitigation evolved and is increasingly practiced since the late 1990s i. e. [13,96]. Risk is seen as a function of probability and consequence, related to exposure and the level of force embedded in the threatening hazard.

Boughton [6] points out that natural disasters are usually extremely rare for the individuals concerned but they can cause massive impacts. Because Australia is so vast, overall there are reasonably frequent natural disasters. However, in most locations they are rare indeed. Boughton [6] argues that a "natural disaster" is a natural event in which the community life is seriously and traumatically disrupted. Embracing the way forward with 'structural mitigation', "… a key step in preventing natural disasters is to prevent building damage." [6]. Like "disasters", "community" is difficult to define [84]. As developed in Sect. "Integrating Theory and Implementation", "community" is the collection of people in a close geographic area, particularly focused on near neighbors and supportive friends of those in or near a known hazard zone.

What to Communicate

Having the right words or approaches in place *as policy* does not automatically guarantee community safety-oriented responses to disruptive warnings. Since 1989 the approach to cope with disasters has been *prevention, preparedness, response and recovery* training courses. This helps focus all concerned on the temporal sequence. The language could perhaps be refined to talk about acceptance that a threat exists, background, then 'ramp-up' (final) preparations to safe shelter; impact, then orderly return and recovery. Classically, 'response' was seen as the final, near impact flurry of 'lock down' activity [57].

Deeper than language is our conceptual frame (Sect. "Integrating Theory and Implementation"). If researchers encouraged disaster managers to move 'response' behavior back a day; a few hours to 'precautionary, early response', many of Lewis's [57] legitimate concerns would be addressed. Some disaster managers may see themselves as dramatic figures in the early impact phase of a disaster, rather than calm, precautionary minimizers of risk. This culture has changed greatly over the last decade (Sect. "Integrating Theory and Implementation"). For many reasons, there is a clear divide between the US Federal response to Cyclone Katrina (2005) and to the California fires (October, 2007). This shift from passive to active can be encouraged by modelers. As part of effective risk communication to encourage precautionary impact preparedness, research with remote Indigenous communities and recently arrived, non-English speaking refugees (Sect. "Integrating Theory and Implementation") showed the importance of accurate, plain English, and the use of images.

Yates [93] argues mitigation efforts need to be refined to make sure they are focused on issues from the relevant local communities. Much of the problem of non response seems that 'the message' to take care does not effectively get through to the target [68]. It is to do with communica-

tion, with signals sent, signals received, and their interpretation.

The types of received and interpreted messages are explored specifically in Sect. "Integrating Theory and Implementation", which develops the intellectual foundation to more fully understand the semiotics of risk communication. The following outlines the concepts and relevance of semiotics.

Semiotics is *the study of signs* including words, sounds and such things as 'body language'. By the 'message conveying' principles of semiotics we can understand how various authors on disasters and evacuations approach the topic, and which cultural signs and symbols they manipulate. For example, a sign showing a person running upslope ahead of an exaggerated tsunami wave contains all the preferred imagery to tell people what to do in a tsunami warning. Images, as seen for locating oneself in airports or finding evacuation stairs, do not use words.

The next section lays the foundations for a mathematical modeling of propensity for householder self-evacuation, centered on knowledge and refined communications acting on residents in hazard zones who are predisposed to act in their own safety.

Mathematical Modeling

Foundations of the Household Safety Preparedness and Action Index: the HSPA Index

This section does not consider refined algorithms to define how successfully people may respond to *Authority's Perceived Need to Evacuate (APNE)*. This section considers the form and ingredients of how this could be done; the elements and their possible interactions to predict the success of APNE. Unfortunately, in the previously cited Canberra 2003 bush fires, the APNE did not appear universally, nor was it properly conveyed to the residents in dire risk.

Rohrmann [75] has clearly mapped a process from the warning signal to the hoped-for response. The information and processing flow to internalized inferences is called intuitive heuristics [74]. Renn & Rohrmann [74] basically argue that people process probabilistic information on the likelihood of them needing to move, aligned and convergent with researchers like Thompson [87]. Hence the need for the seven step process (Fig. 8) to trigger a paradigm shift for people in danger zones.

As seen in formula 1, there are finite evacuation triggers (ET), ranging from no lead time LT_0, such as an earthquake, to a LT of days (LT_{xd}), such as a cyclone. So a key component of a successful response to an APNE is the time in which to disseminate the warning (LT_{0-xd};Table 1) Also, Impact Severity (IS) is central to safety and loss.

As detailed in Sect. "Integrating Theory and Implementation" and Fig. 8; *The seven steps to community safety*, there is a convincing body of research which indicated that People's Propensity to Respond (PPR) to an emergency warning (whether that response be to finalize preparation to stay and defend the property or evacuate as a precaution) is strongly linked to their Acceptance they are In a Hazard Zone (AIHZ).

Community Resilience (linkage and inter-support and communication – CR) is also important; along with a potential evacuee's Knowledge Base of the hazard and safety-maximizing behavior (KB). Thinking Through to Recovery (TR) is also important. As per points 6 and 7 in the 7SCS; Fig. 8, the medium of warning delivery, and the quality of the warning information (Medium & Message: MM) is a factor in the response of those under threat. Overall, Sorenson and Mileti [81] believe that increased credibility of the warning source means the specific warning will be more effective. They also provide evidence that the electronic mass media produced the most believable public warnings. This underlines why development of formal links between all types and levels of media information about hazards and preparation from the weather bureau and emergency managers is so important. Having formal links with the media, coupled with web-delivered 'real-terrain' simulations of the hazard will help produce a powerful and effective 'active warning' regime.

Issues of Exit Route(s) (ER), Possible Travel Mode (PTM) may be crucial. A normally free-flowing exit route may be blocked by the disaster itself, or others trying to flee. This may involve accidents. People's psychological and physical states of Well-Being (WB) will also affect potential effective evacuation decisions and accomplishment. Authors like Rohrmann [75] have attempted to stylistically model some of this complexity.

Many of the earlier recognized impediments to full preparedness, like unrealistic optimism (Heller et al., [46]) are incorporated into one overarching factor of formula (1): People's Propensity to Respond (PPR), including the centrally important feature of individual households fully accepting they are in a hazard zone, and that they need to engage in background preparation, 'listen up' around a hazard threat, and undertake their own, precautionary final preparations.

Cutter et al. [17] developed 11 factors indicating vulnerability caused by a major disaster, using principal components analysis within factor analysis. The resultant Social Vulnerability Index (SoVI), necessarily weights the 11 components according to the percentage of variation

of analysis of counties on the USA as to resident's vulnerability. Personal wealth, age and housing density head up the contributions of the 11 variables used to determine the SoVI of USA counties. Cutter's SoVI contributes to the likelihood that proper preparations; including for a timely evacuation, would occur, either at a broad regional or more individual level. This SoIV is included in the HPSAI below. Its weighting remains to be tested. A final issue overarching many others is Institutional Barriers to change (IB), detailed in Sect. "Institutional Barriers to Greater Community Self-Help". The support or otherwise of the media can play a critical role in effective crisis communication – most messages are likely to be through the media, so the media is part of the systems complexity blend: Media Support (MS).

From the weight of SIR and the conceptual framework described in this article, the following generic formula expressing the above factors likely to impact on householder propensity for precautionary evacuations may entice other, more theoretical researchers, to develop the needed algorithms. What follows is purely a design base for others to develop predictive modeling of ***People's Propensity and Capacity to Successfully Evacuate (PPCSE – or safely and actively stay)*** PPCSE is better named the ***Household Preparedness and Safety Action Index: the HPSA Index*** . The following is a synthesis of Bayesian Logic (Hoeting et al. [45]) and Eigenvalues.

People's Propensity and Capacity to Successfully Evacuate (PPCSE) or Household Preparedness and Safety Actions Index (HPSAI):

$$\begin{aligned} &\text{HPSAI } \alpha \\ &\text{f(PPR)(AIHZ)(LT}_{0-xd}) \\ &\text{(APNE)(ET)(IS)(AIHZ)(CR)(TR)(MM)(ER)} \\ &\qquad\qquad\text{(PTM)(WB)(SoIV)(IB)(MS)} \end{aligned} \tag{1}$$

Formula (1) anticipates constants to 'weigh' each factor according to its contributing importance on the resultant HPSAI (see [45] p 384) for detail of the Bayesian modeling approach.

Resultant safety (*safety* being a combination of successfully evacuating to a safe place, or sheltering within the impact zone in a safe place) of those under threat will be greatest (Fig. 1) where the factors of formula (1) tend to intersect in the positive quadrant of an Eigen plane (Fig. 1) [94]. Equally, outcomes of loss will occur with, for example, short lead times, insufficient warnings of evacuation triggers or lack of belief that residents or travelers are in a hazard zone. A journal search showed few links between disasters and Eigenvalues. An exception is Fowler et al. (2007, [32]). With global warming and increased populations increasingly encroaching into obvious hazard zones, modeling to influence planning and to maximize HPSAI must be a major growth industry.

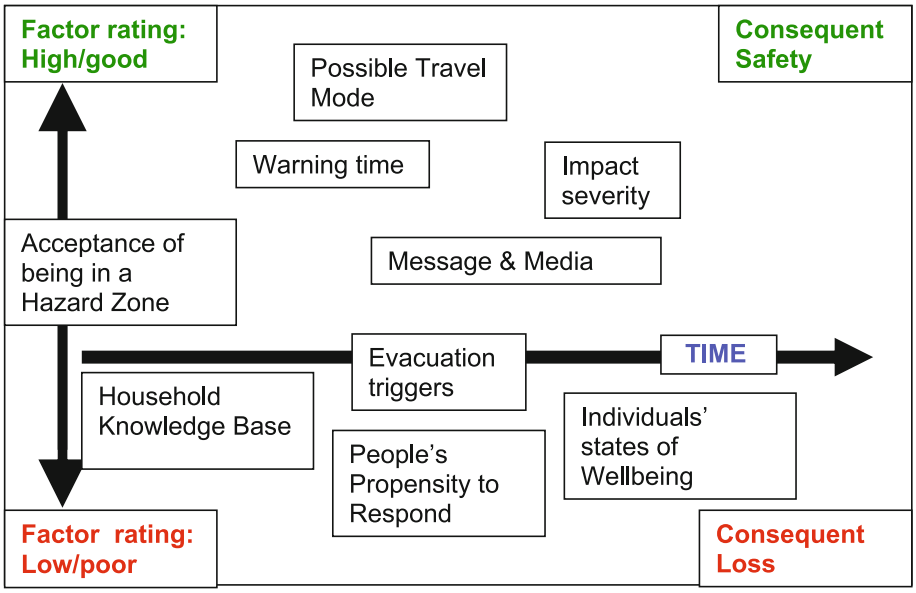

Evacuation as a Communication and Social Phenomenon, Figure 1
Factors in the Eigen plane determining *Household Preparedness and Safety Action Index*

Calibrating the Model: the Greek Fires, August '07

[98] or California in late October '07 [99] show that if various of the above variables or likely variable clusters or Eigenvalues from factor analysis have low values, people will not be placed in a strong, precautionary position of safety. This was true of fire-threatened residents in Canberra in 2003 (http://www.abc.net.au/canberra/bushfires/), all ultimately leading back to the mathematical weight which needs to be given to each point on in the 7SCS: especially accepting that the threat is real (and possibly imminent).

Once the above-described preparedness paradigm is internalized, it is difficult to experience any disasters news coverage without feeling some despair: cars submerged or floating in flood waters; people fleeing for their lives ahead of imminent immolation; any reports of lost lives through natural disasters with lead times of hours or longer. There are few excuses for this to remain so. In most parts of the developed world, there are prevailing technologies to detect and rapidly convey threat. Researchers are well placed to conceive, trial and link the threat information directly to prepared populations; to close the gap between knowing and acting to maximize safety, as a social exercise of community engagement and communication refinement.

Cyclone Larry [54] showed that, with general community acceptance of the reality of the threat, good warnings and a lead time of about 20 h, combined with well developed disaster management and media cooperation in getting the needed messages to people well primed to the 7SCS, there was no loss of life, although about 5000 people were directly subjected to destructive winds of a category 4/5 cyclone. Residents of the impact zone generally followed the 7SCS, supporting the general approach of formula (1): the community was prepared and acted in a precautionary way to maximize their own safety.

The above model design offers elements/factors to maximize *People's Propensity and Capacity to Successfully Evacuate (PPCSE – or safely and actively stay); the Household Safety Preparedness and Action Index (HSPAI), so* strengthening a paradigm among Agencies and hazard-zone residents of 'self-help' and 'community engagement' approaches; less passive than earlier approaches used in considering people's vulnerability. This generic approach is supported by model testing by Schadschneider et al. (► Evacuation Dynamics: Empirical Results, Modeling and Applications).

The Australian examples detailed in this article, including the national shift to mitigation (COAG 2004 [13]), best illustrated in the Woodgate Beach example (Sect. "Experiences and Lessons – Some Case Studies"), show that the conceptual frame and thus variables chosen to mathematically model evacuation propensity and capacity are more important than the results of any model using 'passive' variables or variable clusters. The advantages and means of achieving this shift from community passivity to partnerships; place- and people-based community empowerment and self-help, form the basis of this article. The above mathematical approach may help validate and guide this necessary conceptual shift.

The next section considers policy, laws, concepts and possible safety-oriented actions that words, images and stories convey to vulnerable residents in hazard zones. In Australia, policy and legislative requirements are indicated, which aim to responsibly maximize community safety [13,24,25,26,27,28,44,70]. Some new concepts, epitomized by new language phrases (with their embedded meanings) are discussed in Sect. "Integrating Theory and Implementation", especially *Community safety groups*. New phrases include 'Social Burnoffs' and 'Practice evacuations'. Like "The Communication Safety Triangle" and "Seven steps to community safety", it is hoped these expressions enter local, state and national lexicons and modeling in disaster management.

Effective Risk Communication

Policy and Laws on Hazard Preparedness and Evacuation

All risk communication operates within government policies and legislation. Forced evacuation may be normal in some jurisdictions for all threats; outlawed in others. Policy may encourage passivity, or be energetically pro-active in achieving the 7SCS, both in urban planning and building materials, and community engagement. In Australia, for instance, there are many federal and regional guides to urban growth, e. g. [2]. A new urban paradigm is, perhaps, best expressed as: "Think globally, act locally, respond personally" ([2] pnp, [3] p 1). This paradigm shift can easily embrace and be informed by computer modeling of individual's behavior, demonstrated in recent Springer's Journal of Systems Science and Complexity articles, such as Kikuchi T & Nakamori Y [97]. Agent model analysis to explore effects of interaction and environment on individual performance [97]

From all decisions of urban development being in the hands of 'experts', government policy increasingly requires a dialog with local residents, through public participation. The first barrier to ecologically sustainable urban development is "… belief systems – doubting a problem exists, or supporting the status quo. [Solutions include] … consciousness raising campaigns, public partici-

pation in decision-making, demonstration projects and incentives and disincentives" ([76] p 46). This is the KB of formula (1).

This article stresses that response models, Authorities and all residents are constrained (and empowered) by national and regional laws and policies. Research can only work within a prevailing frame, and will be welcomed by or influence prevailing policy. It is thus useful to have clear goals to achieve social or environmental 'good' [13,44,70]. This requires greater community involvement in mitigation, and responsible media helping to mitigate disaster impacts (MS of formula 1). In Australia, the national radio broadcaster has a binding agreement with the weather bureau and disaster managers to read issued warnings verbatim; report and engage community members in safety oriented behavior, by providing relevant information. Local community media are often willing to develop the same responsible functions.

Risk Communication's Core Goals

Risk communicators advise "individuals and communities to respond appropriately to a threat in order to reduce the risk of death, injury, property loss and damage."[42]. Risk communicators and disaster managers need to formally work closely with the media to maximize social benefit.

Although it may be intended that information flows and is received accurately; that desired behavior will result, and that only communication techniques are important [49], humans tend to be irrational and optimistic, and only hear what they want to hear. It is not what our message *is*, but what, if anything, the listener *does* with our message. To have any chance of 'success', information needs to have meaning which is shared between those who construct and send the warning, and those for whom the warning is meant to inform and motivate to action.

'Action statements' (what the at-risk person, family or community should actually do to minimize damage) are seen as central to the whole purpose of risk communication [78]. Kasperson & Stallen [49], along with Salter [77,78] and others detail risk communication messages in terms of content, clarity, understandability, consistency, relevance, accuracy, certainty, frequency, channel, credibility, public participation, ethnicity, age, gender, roles, responsibility, elements, sequencing, synopsis, prognosis, location, action, warning timing, and action statements. It is not a case of saying: "a category three cyclone will pass over Smithfield". It is more a case of making sure the members of Smithfield hear that message, and feel moved to and competent to take well-understood personal risk-minimizing actions.

Knowledge, Self, World Views and Messages

For precautionary evacuations to be successful, the seven steps to safety (Sect. "Integrating Theory and Implementation") have occurred; or people have been coerced by authorities, or they have seen others depart, and felt insecure, so they leave as a follower of the 'innovation' of the 'norm': precautionary evacuation. This occurred in the desktop exercise in an isolated settlement in March 2007, detailed in Sect. "Experiences and Lessons – Some Case Studies".

It is now difficult to understand why people needed prompting to take evasive action against the forecast Brisbane floods in 1974, but many ignored the prompts. Authorities believed houses and roads would be flooded, so people should finalize travel or evacuation early [10]. For those in the flood zone, 88% of a later survey sample reported evacuating their home. Some took this step very early, but 67% of respondents only made preparations immediately before leaving home. Twelve percent only left on foot or in boats *after waters entered the main living areas of their homes*. Almost 22% of respondents said they made no preparations, mainly because the threat was not recognized in time [10]. This well documented 'poor' crisis communication can help calibrate formula 1.

A key safety message surrounding floods is not to enter flood waters. Figure 2 is one example of using images to help convey the reality of a threat, and the pitfalls of belated action.

Effectively Conveying Meaning

Because language is pivotal to acceptance of risk and conveyed warnings of need for evacuations, it is important that all participants reasonably agree on word meanings. This section starts in the comfort zone of simple definitions, then considers semiotics, considers core issues of world views (paradigms), and finishes with the uneasy realities of our knowledge base, our epistemological orientation. Sect. "Integrating Theory and Implementation" shows that some cultures do not carry the background cultural experience to absorb the meaning of cyclone or bushfire – there is no shared experience or 'stories' of the power of these extreme forces of nature. Some of the underlying knowledge foundations of context, intent and behavioral motivation need to be considered – how humans construct, transfer, acquire and use knowledge.

Imparting meteorological knowledge, then warnings, to target audiences to engender safety-oriented responses is a complex exercise in social empowerment, explored in Sect. "Experiences and Lessons – Some Case Studies". As information promulgators, information and warning

Evacuation as a Communication and Social Phenomenon, Figure 2
What flood-threatened residents need to consider (Photos courtesy Townsville City Council)

sources should understand a little of *Perception* (the raw data from the outside world entering an organism via any of the five senses), *Cognition* (internal processing, analyzing, information storage and processing), *Attitudes* (how we think and feel about particular issues, implying a predisposition to specific action), *Language use* and links to *Behavior*.

An intellectual framework to risk communication is provided by Rohrmann [75]. It considers 'the message features': the recipient features, social influences and context which influences individual risk assessment and management, including preventative action. Rohrmann and Handmers' publications [41,42,43,75] inform the Communication Safety Triangle and *the seven steps to community safety* provided in this article.

Philosophy for Policy Review: Crying Wolf or Worse – Applying the Precautionary Principle

From the 1990s a strong issue of debate in risk communication has been "the right to know" [4]. Some disaster managers wish to avoid false alarms, which may cause 'concern fatigue'. This can be seen as an institutional barrier to change (Sect. "Institutional Barriers to Greater Community Self-Help"). 'Avoiding undue alarm' is in conflict with the right to know, and the precautionary principle of ESD.

The 'precautionary' approach is supported by the Economic Commission for Asia and the Pacific (ECAP), the World Meteorological Organization and the Red Cross Societies. The alerting of the community and its responsible authorities must begin, at least provisionally, as soon as the existence of a tropical cyclone over the seas bordering the country is known ([24] p 16). According to ECAP et al. [24], the warning challenge is less clear for predicted localized downpours and flash flooding – how much effort should be taken to warn – what is the message, how do you keep it to the affected area, and what do you want people to do? These questions resonated in Australia after billion dollar hail damage in Sydney in April 1999, or damaging flash floods in Melbourne in December 2003.

Precautionary Evacuations

Handmer [43] reports an evacuation of 250,000 Dutch ahead of a flood threat in 1995. Eighty-eight percent of people surveyed in broad post-emergency surveys "believe that evacuation was appropriate." [43] p 24. In part this may be because of floods experienced two years prior. Good skills in dealing with the mass media appear to have helped in the effective precautionary evacuation. The Dutch experience showed a willingness to evacuate again in future, even though the threatening flood waters did not inundate to the level feared. This compares favorably with the 2005 boat owners' responses to Cyclone Ingrid in Port Douglas, North Queensland (Sect. "Institutional Barriers to Greater Community Self-Help"). Both Handmer and Goudie's research [33,36,38,41,42,43,52,53,54] show clearly that people would rather practice (make a precautionary evacuation) than incur loss. It was treated as a learning experience. The 'boy who cried wolf' argument is not acceptable. This is an important message researchers can test and convey to partnering Disaster Managers.

Have Clear, Consistent Messages

Part of the *seven steps to community safety* is clear, reliable, explicit languages and images. Salter et al. [78] point out that the use of meteorological category systems such as 'minor', 'moderate' or 'major' carry unambiguous information about the level of disruption likely from a particular flood. Language used should not be for the convenience of the warning agencies. Its function is to convey clear unambiguous messages to the threatened public.

Use Past Events to 'Make the Threat Real'

The purpose of risk communication is to make people perceive the threat as real, and to successfully motivate safety-oriented action. Boughton (1992, p 6) argues that awareness of hazards and disasters can be fostered by "drawing attention to media coverage of hazards in other places". Images of large scale floods and evacuation in Holland in 1995 [42] may help prompt a future flood evacuation.

The Changing Politics of Risk Communication

Risk communication is often laden with values and political implications (Sect. "Institutional Barriers to Greater Community Self-Help"). For instance, in the 1990s in Cairns, north Queensland, it was argued that the reason for not having detailed local cyclone surge inundation maps made available at the corner store level was that such information may have a negative impact on local land prices. This appeared more a political decision than an attempt at effective risk communication. In 2007, such maps are freely available on the local Council web site [9].

Sirens or Not

No or short lead time disasters are a powerful argument for warning sirens to alert people, perhaps at 2 am, to "listen to local media now". Lives will be saved in future tsunamis and bush fires with the reintroduction of sirens, as insistent triggers to "find out more now". However, "large numbers of sirens are needed to cover populated areas and to be loud enough to be heard indoors by most people. Sirens are expensive to install and maintain and can only provide limited information" [63] p 33. Fixed public address systems or those on vehicles may be used. People who hear such sirens will be encouraged to tune to local media, and to phone others in the threat zone, to warn them of the alert. Sorenson and Mileti [81] believe sirens are most effective if used on populations without other ways of receiving the warning. Wider use would appear prudent, especially with short lead-time threats.

Media Roles

As community media moves from 'sensationalism' [13, 71], 2007 research by Goudie confirms that community media organizations now want to become responsible conveyors of safety-oriented information to people at risk. After the Canberra fires of 2003, where 300 homes and four lives were lost in the nation's 'bush capital', all local media signed binding agreements with emergency authorities to faithfully convey provided safety information. The Communication Safety Triangle envisages local media and householders drawing detailed local threat information from the internet, with media conveying that directly to readers, watchers and listeners. This forms the basis of future 'world best practice' risk communication. In the preparation for threat impact, the reliable information will help people make informed decisions, rather that remaining trapped and inactive in uncertainty.

Risk communication is complex, involving many values, predispositions and distorting lenses. Rohrmann's [75] fine risk communication explanations show that we may tell the people at risk, but they may not interpret as intended. Clear, consistent warnings in plain English, with clear images of the threat, showing safety-oriented behavior are needed from reliable sources. Warnings, seen on a continuum of risk and preparation actions, should be able to be discussed and reinforced with information from other sources. This is most likely to produce safety-oriented behavior, with the constrained and clear assistance of the media, as responsible agents for community safety.

Given the previously fraught nature of risk communication, Sect. "Institutional Barriers to Greater Community Self-Help" provides a conceptual framework, with Australian examples of current community safety theory and implementation, preceding some detailed examples.

Integrating Theory and Implementation

This section introduces a triangular model to best ensure community safety. With disaster managers central, the three elements are the community, local media and the internet. Also, there is a continuous spectrum of seven steps, from accepting there is a risk, through early preparation, final (ramp up) preparation, which may include evacuation, surviving the hazard impact and achieving smooth recovery. An exploration of motivation leads to a discussion of 'world view', including some insights into remote Indigenous communities [36], and recently arrived, non-English speaking immigrants. The conceptual frame and steps are intended to help guide risk communicators and potential modelers through the issues of acceptance, preparations, evacuation and functional recovery. Within CSS, all these factors have some bearing on consequent behavior.

Changing Values and Roles

In an attempt to understand why we support or ignore certain messages relevant to our safety, and thus facilitate modeling, this section considers the Dominant Social Paradigm and the New Environmental Paradigm.

A paradigm is a coherent world view, "a mental image of social reality that guides expectations in a society" ([23] p 10, [29]). Shared paradigms change. The underlying philosophy shift toward disaster mitigation generates a distinct policy move toward safe, self-helping communities. Thus, the goal of disaster managers, effective risk communicators and computer modelers becomes one of championing or demonstrating the efficacy of self-help techniques and information to and within communities.

World Views Which Accept Responsibilities of Living in a Hazard Zone

'Why do we do what we do?' has long been a central exploration of psychology. Precautionary safety behavior, including self-evacuations, may depend on a person's world view. Stern et al. [83] developed a simple and elegant model to help explain human behavior, starting with a person's position in a social structure, with constraints and incentives which generate values. Values determine general beliefs, leading to a consistent world view, specific beliefs and attitudes, which predisposes intent and helps explain behavior ([83] Fig. 3). Loosely translated, behavior results from a linked sequence starting with: the where, when and to whom of an individuals birth, the surrounding circumstances and 'cultural sheath' of their early childhood experiences, leading to acceptance or rebellion against prevailing social norms, determining an evolving world view. Once we have our coherent world view, beliefs and values give us our 'predisposition to act', our 'intent' which precedes action/behavior.

Many things now mold and modify world views, including the media, and normative world views are mutable, changeable. Aligned starkly with our shared and collective biological survival urge, and much concerted efforts by scientists and conservationists for decades; publicizing by people like Al Gore, and the authority of England's uptake of the 2006 Stern Report, there is now a global paradigm shift to the urgent need for behavioral and technological change to minimize the gross impacts of global warming, pertinent to increased disasters and needed evacuation readiness. Modeling specific hazard zones' strengths and weaknesses and broadcasting prediction simulations to the internet and TV will help mobilize those at risk.

Behavior is explained by:

1. a person's *position in a social structure*,
2. with *constraints and incentives* as generators of *values*, which lead to
3. general *beliefs*,
4. *world view*,
5. *specific beliefs and attitudes*, generating
6. *intent*, which helps explain
7. *Behavior*.

Evacuation as a Communication and Social Phenomenon, Figure 3
Stern's 1995 behavioral explanation model

Why We Do What We Do

A behavioral model of causality ([91] Fig. 4) shows relationships between reported attitudes and actual behavior. However, a complex model proposed by Kitchin [51], with the strength of including social and environmental interactions shows why we do what we do: Kitchin's model includes a person's 'working and long term' memory. Internal information is processed within 'real world' context, such as cost [82], which may be processed within the 'it can't happen to me' cognitive frame [61] of subjective reality. These complex but quantifiable attributes will contribute to modeling maximum likelihood to accept, prepare and act to maximize safety.

Learning from 18 Remote Indigenous Communities

Much of the rest of this section underlines why modelers will need to deal in great detail with 'cultural' aspects of massed responses to safety threats.

The Stern model (Fig. 3) helps explain why there is such a strong sense of self-help in remote Indigenous communities [27,29,80,94] helps explain why there is such a strong sense of self-help in remote Indigenous communities [27,29,80,94]. Elders decide responses to threats, there are historic and immediate constraints, generating a value system where community members need to look out for each other.

In many traditional Australian Aboriginal stories, most people drown in flood disasters, often as punishment when people do not take care of each other [36,39,64,87,88]. If general beliefs embrace self help, including the felt need to ensure safety, and a resultant intent to achieve community safety, this should lead to safety-centered behavior.

Each community ([36] Fig.5) was not greatly concerned by weather extremes (values), but each relied on and respected their traditional reading of threats, and information from the Bureau of Meteorology (the Bureau).

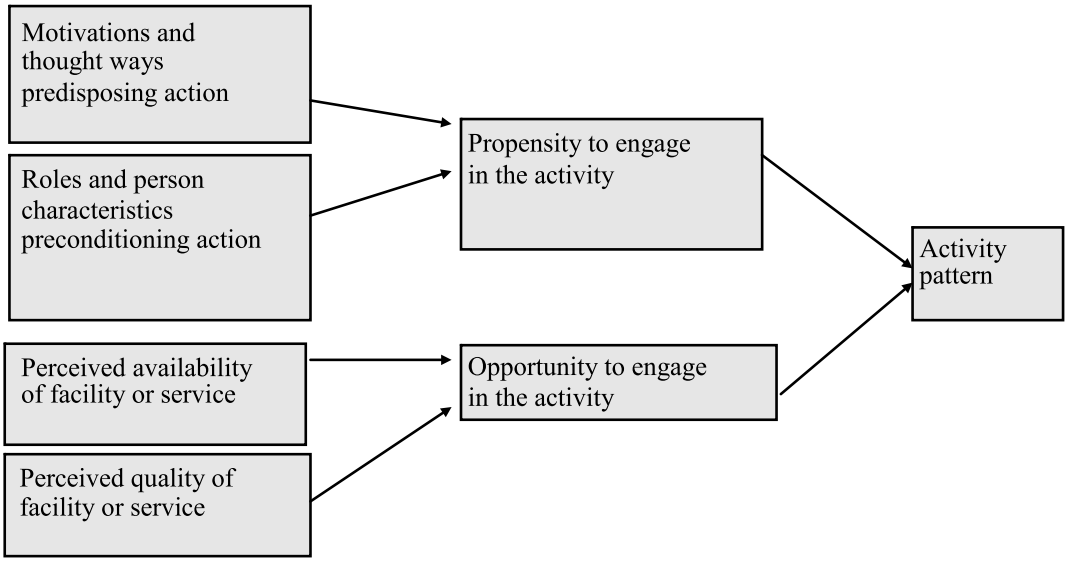

Evacuation as a Communication and Social Phenomenon, Figure 4
Possible determinants of activity patterns (from [91])

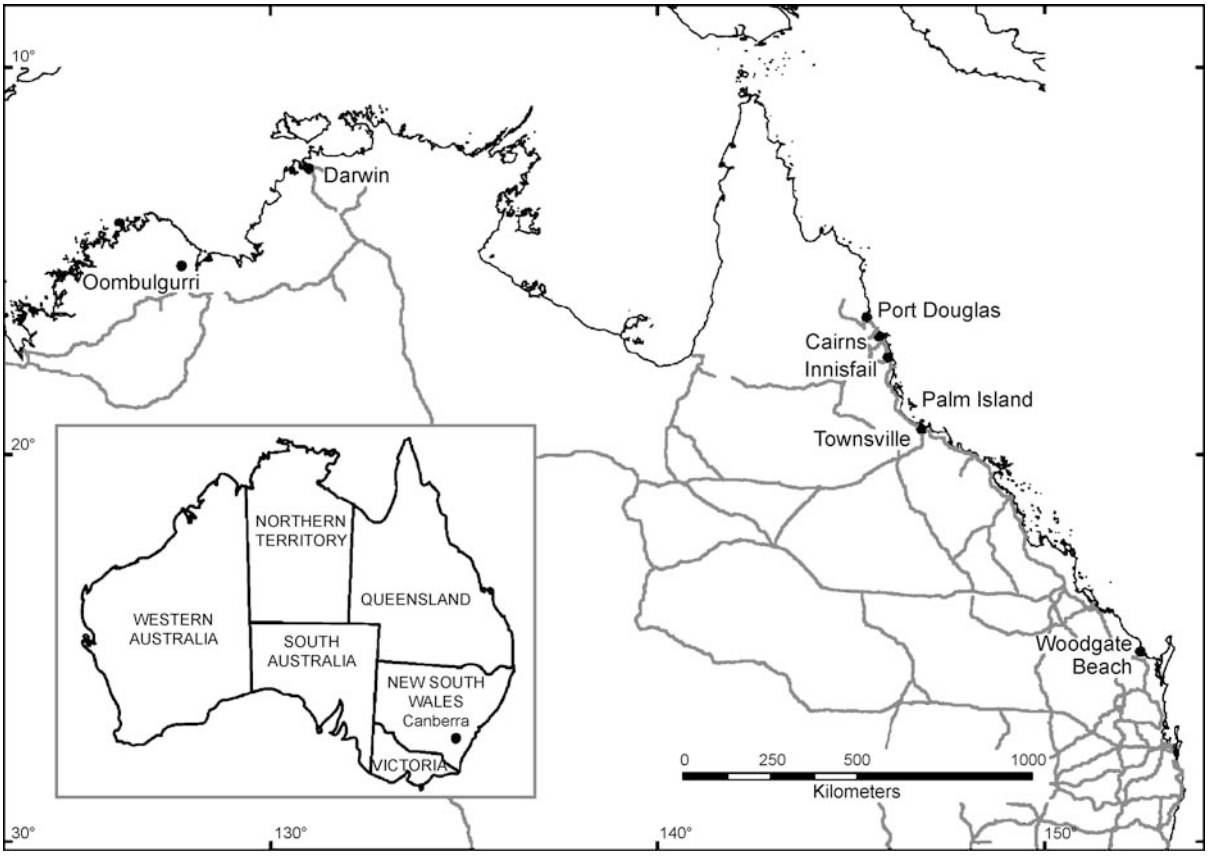

Evacuation as a Communication and Social Phenomenon, Figure 5
Australian study sites

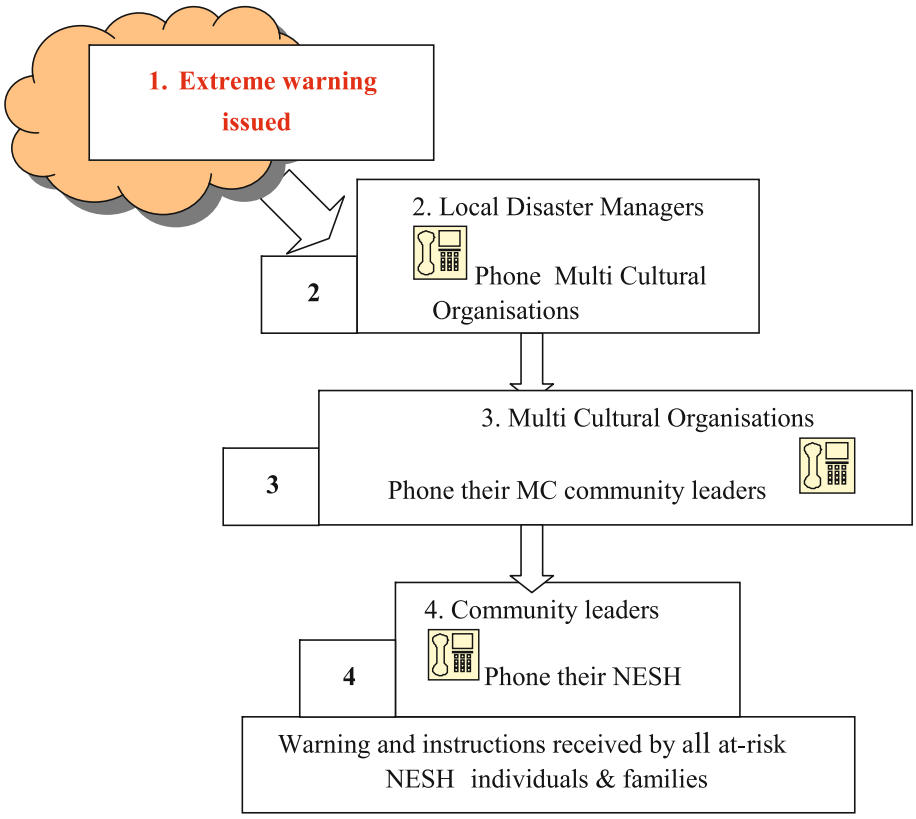

Evacuation as a Communication and Social Phenomenon, Figure 6
Multicultural phone tree disaster warnings model

The world view is that flooding or worse may happen, and that the Bureau will provide adequate warning. Responses may be inhibited by the Social Vulnerability Index (SoVI) of formula 1.

Community Links, Phone Trees and the Web in Risk Communication for Non-English Speaking Households (NESH)

Many multicultural organizations and focus group meetings in 2005 helped develop the warnings phone tree model for NESH in Fig. 6. NESH rarely listen to the English language media. With about 30,000 NES people arriving in Australia each year [18], there is Federal government recognition of special emergency management needs [28]. This way of getting evacuation warnings through was developed once interviews revealed that practically all NESH have a mobile phone, and are closely connected with their nearest government funded Multicultural Organization and their 'community leader'; hence the phone tree. Modeling is of and for the real world, so research like this reported NESH study is needed to see what 'complex systems' may be possible. Further recommendations to develop a multilingual warning web site with a simple guide to the seven steps, in relevant languages, to be accessed and used by MCO and NESH training sessions is being considered by authorities. This provides another example of the internet's latent role in effective risk communication.

The Australian Northern Territory has cyclone preparedness kits with action guides in 8 languages; the NSW Rural Fire Service has information in 27 languages. This shows the importance of modeling all possible communication avenues, encapsulated as Medium & Message: MM in the general formula for Household Preparedness and Safety Actions Index (HPSAI).

An Holistic Approach to Community Risk Management

The seven steps for effective warnings involves community networking, 'responsible' media, and, increased web

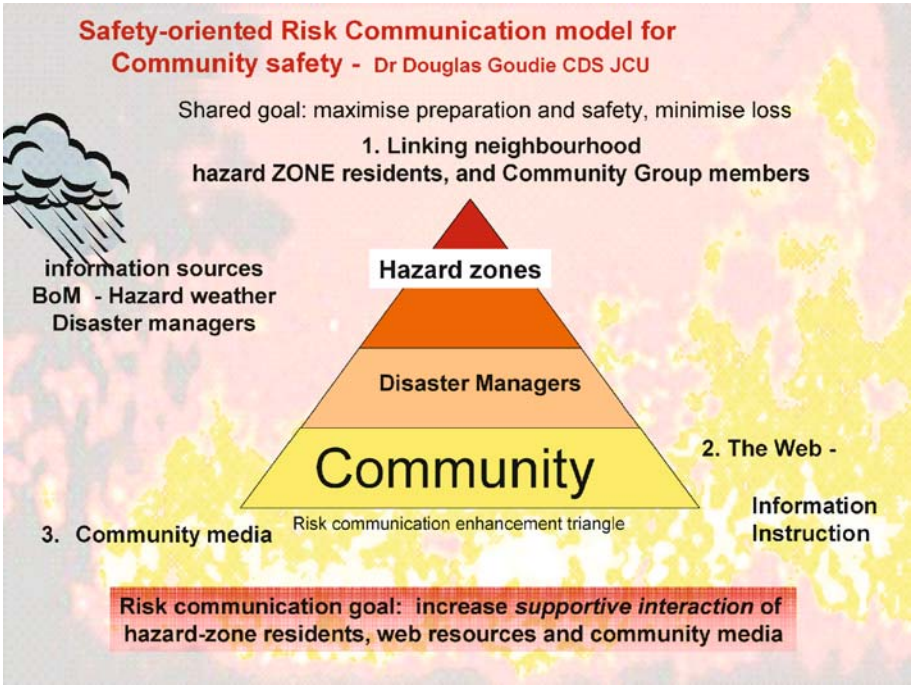

Evacuation as a Communication and Social Phenomenon, Figure 7
The Communication Safety Triangle

use. A 'continuum' approach to hazards which may lead to a need to evacuate is important to modelers because it is important to all people in hazard zones: they will be able to access more accurate, detailed and timely information of any looming threat, assisting Emergency Managers because a self-help public will decrease demands on formal evacuation, rescue and property protection [93]. Recovery will be less arduous [67] if hazard impact is minimized. Reducing disaster impacts will reduce costs, to national benefit. This gives strength to supporting the very demanding goal of developing the model of formula 1.

Risk Managers

Researchers can work with risk managers to achieve "… better, timely warnings and advice on safe action during fire events" [3]. With risk managers central, Fig. 7 suggests that enhancing community links [7]; web information for residents and media outlets, and cooperation of community media with fire managers [89] will more empower householders to embrace self-help in fire safety. Figure 7, the Community Safety Triangle, with the "Seven steps to community safety" (Fig. 8) provides the conceptual frame to enable maximum safety modeling.

Core of Risk Communication for Community Safety Through Natural Disasters

Building Community Links and Refining Media Delivery will Change the Household Preparedness and Safety Actions Index

The disaster safety benefits of enhanced community links fit positively with broader social policy [47]. Stronger community links will help ensure that threat information is easily accessed and shared; and the need to actively self-protect is internalized at the neighborhood level. Residents will benefit by enhanced community links (with such innovations as social burn offs and Community Safety Groups) in improving their general quality of local social interaction. Changing Community Resilience (CR of the formula) will then change overall preparedness, so pilot tests of change to CR will show quantifiable changes to the Household Preparedness and Safety Actions Index. Disaster web site managers will be providing a product which is rationalized to reduce national duplication of core information. This will positively change Medium & Message, MM, of the formula. Residents with few English language skills will benefit by having disaster information delivered,

Goal: Maximise safety and recovery, and minimise loss in hazard zones.

1. Encourage those in hazard zones to accept that the risks are real.
2. Help create an aware, informed community, predisposed to safety-oriented action, as a precaution; as a practice.
3. Encourage information-sharing and support among friends, neighbours, family.
4. Provide 'what to do' (action) information, via reliable sources, including web and community media delivered for background and preparation.
5. Encourage people to think right through to impact and recovery.
6. When a threat is closing in, warning messages will clearly convey: *this is real, this is coming at me. I need to make safe where I am, or move early to somewhere much safer. I will not travel during the impact period.*
7. Provide timely, effective threat warnings and fine location and forecast weather detail, and recommended local responses.

Evacuation as a Communication and Social Phenomenon, Figure 8
Seven steps to community safety

via the web, in language and concepts that can be internalized and acted upon.

Handmer [43] suggests that a flood, for instance is actually 'owned' by the communities at risk. Individuals and organizations within these communities actively seek out information and mobilize their personal networks for action. In this way of looking at the warning process, the warning specialists act as mediators between the threat and the threatened. Local knowledge is used and the whole response process remains focused on safety and loss-minimization. For this plausible vision to become fact, local capacity building will need to proceed apace with these more formal efforts to inform and maximize community safety.

The importance of 'informal' information sources and community links are shown in Table 2, from about 200 people interviewed immediately after Cyclone Larry in 2006 [53], showing that about 90% are deeply dependent on social and family support, if only for reassurance.

Finally for community enhancement within the CST model, there is current and lucid national disaster mitigation policy [13] http://www.dotars.gov.au/localgovt/ndr/nat_disaster_report/naturaldis.pdf : in the statement of the paradigm shift [13], p 13, "Principle 7 – Reform Commitment 7: develop jointly improved national practices in community awareness, education, and warnings which can be tailored to suit State, Territory and local circumstances."

Evacuation as a Communication and Social Phenomenon, Table 2
Cyclone zone people have much personal contact

Contact with other relatives	% of total response (rounded)
Yes	25
Lots	20
Mobile contact	15
Landline Phone	30
No	10
Frequency of neighbor contact	
Often or lots	50
A bit	20
Helped/contacted during eye	15
None	15

Community Media

"The Media: Media organizations, particularly public and private radio and television organizations, have responsibilities in ensuring that timely and appropriate warnings and advice on disasters is broadcast to communities at the request of relevant authorities. They also have a role to play in educating the community about natural disaster issues" ([13] p. 18).

The Web

Rationalized web information delivery: One national generic provider and up keeper of core information, known to all, is recommended; to which state and local government will add their unique detail. A rationalized disaster information web delivery can properly incorporate broad contents, fine (and zoom-in findable) and multi lingual information.

Simulations of predicted near-term fire or other threat movement based on the above, akin to micro scale modeling of cyclone impact forecasts will help revolutionize the way people react to fire and flood threat. With about one in three households having the internet, prior to likely power loss residents can draw on enhanced fine detail of the current fire situation, to assist in their actual decision making. Before Cyclone Larry struck, residents with web access to the Bureau copied cyclone forecast maps and distributed

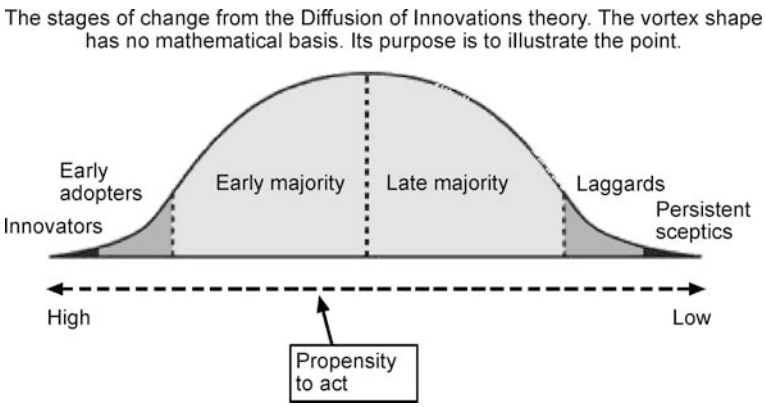

Evacuation as a Communication and Social Phenomenon, Figure 9
O'Neil's 2004 innovation uptake model

them to neighbors [53]. All the preceding is encapsulated in Figs. 8 and 9.

Images Tell a Story

Images such as cyclone forecast track maps in risk communication convey more information that words alone [36,53]. With this feedback from real residents, it becomes clear that linking fire zone residents, for instance, to educative web fire images becomes an important goal (e. g. [100]).

Community Story Telling Bonds Neighbors

The multi-pronged approach [33,52,54] fosters community 'story-telling' links [50,59,90] relating to community self-help and safety. Strengthening community bonds, working with enhanced web resources and community media [14,72] will increase community internalization of risk [31], enhancing the likelihood of safety-oriented responses, leading to and possibly including evacuation. People need to accept the reality of the threat, indeed, feel some anxiety about the threat to help drive the intent to seek more information, or the intent to prepare [68].

With easier access to relevant web information; with greater detail of current fire behavior and *nowcasts*; fire zone residents and community radio announcers can describe the looming threat, helping timely preparations and the monumental 'stay or go' decision. Community radio, like the National public radio, the ABC, will deliver authoritative and timely risk communication directly from the refined web information.

Theory and emerging practice converge on using refined web-delivered material to households and their neighbors, and to local media, to inform and motivate residents through the whole continuum of the seven steps to community safety. The next section considers institutional barriers to change, followed by extensive research findings and lessons from Australia in disaster management to develop effective warnings and self evacuations.

Institutional Barriers to Greater Community Self-Help

Institutional barriers to change take many forms, from 'unconsciously' avoiding consideration of the extreme event as 'too hard', to a misuse of the power relationships within bureaucracies because of fear of change, or malice. Clear examples of the 'too hard syndrome' mingled with entrenched vested interests has been denial of links between smoking and cancer, or delayed uptake of global warming mitigation.

There were Institutional Barriers for disaster managers to consider land-based flooding from torrential rains preceding a cyclone (hurricane/typhoon) [34] in many cyclone-prone areas. The Queensland State Planning policy [70] now explicitly refers to Probable Maximum Flood. Including maximum flooding is now part of emergency manager's internal planning base. This means the Exit Routes (ER, formula 1) are now considered in planned evacuations.

Since the early 90s researchers like Boughton [6] have suggested having drills for schools and other institutions in readiness for possible earthquakes, cyclones or other hazards. This author supports precautionary evacuation practices. There are, of course, liability barriers to easily undertaking practice evacuations. There is also uncertainty of threat, which may restrain some risk managers [86].

Institutions have cultures which may passively express antipathy to a paradigm shift toward sustainability, whilst being required to usher in sustainability [19,37]. Emergency Management Australia, the peak national body, has the slogan: *Safe, sustainable communities*. However, there is a multitude of conservative forces representing the Dominant Social Paradigm restricting innovation, despite sustainability policy. People engaging in sustainability implementation may come into conflict with institutional representatives of the old paradigm.

With most development taking place in urban settings, concepts of urban sustainability attempt to merge two different fields of human endeavor – how we modify our landscape, our built environment, and how we behave in that environment. Disaster mitigation is obliged to fuse these two seemingly disparate fields. There may be resistance to that. That resistance needs to be included in any mass disaster movement modeling. The following subsections indicate some reasons for IBs which modelers may need to consider in the IB factor.

Political Insecurity, Real Estate Values, or Undue Alarm

The tension between people's right to know, government duty of care and politicians' perceptions of probability of risk on 'their shift' form a complex interplay. If maximizing safety is the shared goal, all risk communication theory suggests people will behave appropriately with the realistic threat information, motivation and 'how to' instruction on safety-oriented behavior leading into, through and recovering from a natural disaster impact.

Paradigms of Politics: the Real Estate Industry and Vested Interests

Broughton [6] argues that if people know of the threats, they are likely to support politicians who make sound decisions for community survival. Some plans are made but they are hidden away for various financial and political reasons. This is a case in which attitudinal changes on the part of those communities may change the priorities of the decision makers and promote the interests of the community.

Those most empowered to assist in implementing projects that require independence, initiative, local links and knowledge may be the ones who prove most obstructive to innovation [1,48,56,69].

Action Research [16] sets out to research, develop and document socially and environmentally desired outcomes within sustainability principles. In this inclusive and 'engagement-focused' research, gaining support from the top is crucial to the success of any organizations' efforts at 'social change' [66]. If innovation uptake is viewed as a normative uptake process [12,65] and Fig. 9, then the gradual, long term paradigm shift to sustainability [30] is plausible and has a conceptual frame.

Uptake of the Communication Safety Triangle and the seven steps to community safety (Figs. 7 and 8) will make them the disaster managers' 'norm' over time, simply because that is the direction of social evolution in disaster management. It fits social policy and advanced disaster management approaches, and the technology and willingness of regional media and residents is there. All that is needed is the testing and roll-out by more innovative disaster managers. The same applies for Community Safety Groups and variants of *over-managed, locally controlled near-house "social burn-offs"* in the case of fire risk management.

Bureaucratic processes need to support the stated goals of their own work units, but individuals may be ambiguous, contradictory, belated, bullying, ill informed and, due perhaps to time pressure or arrogance, quite destructive. Such individuals may be 'corporate psychopaths' (http://www.abc.net.au/catalyst/stories/s1360571.htm). Instead, innovators may be seen as threatening; troublesome. Any innovative pilot project must develop strategies that will maximize change within their unique situation. Ideally, that happens within a supportive parent organization. The next section discusses the emergent issues of effective risk communication and precautionary self-evacuations.

Experiences and Lessons – Some Case Studies

This section considers disaster impacts and impacts from north Queensland and from fire zones in SE Australia to illustrate the generally positive points of approaches already outlined: living flesh for modelers to understand how subtle, complex but do-able successful community preparedness and willingness to act for maximum safety can be. Researching these events, gaining community and disaster managers feedback on risk communication; working closely with the Australian Bureau of Meteorology over 15 years, through the Federal disaster 'paradigm shift' to mitigation ([13] p 13) all inform the emergent CST and seven steps to community safety.

Cairns and Storm Surge Considering Exit Routes (ER, Formula 1)

Cairns City, North Queensland, is centered on land less than 2 m above high tide, and subject to cyclone (storm) surge of up to 5 m. A storm surge tracks just behind the eye of a cyclone, a low mound of sea water, perhaps 50 km

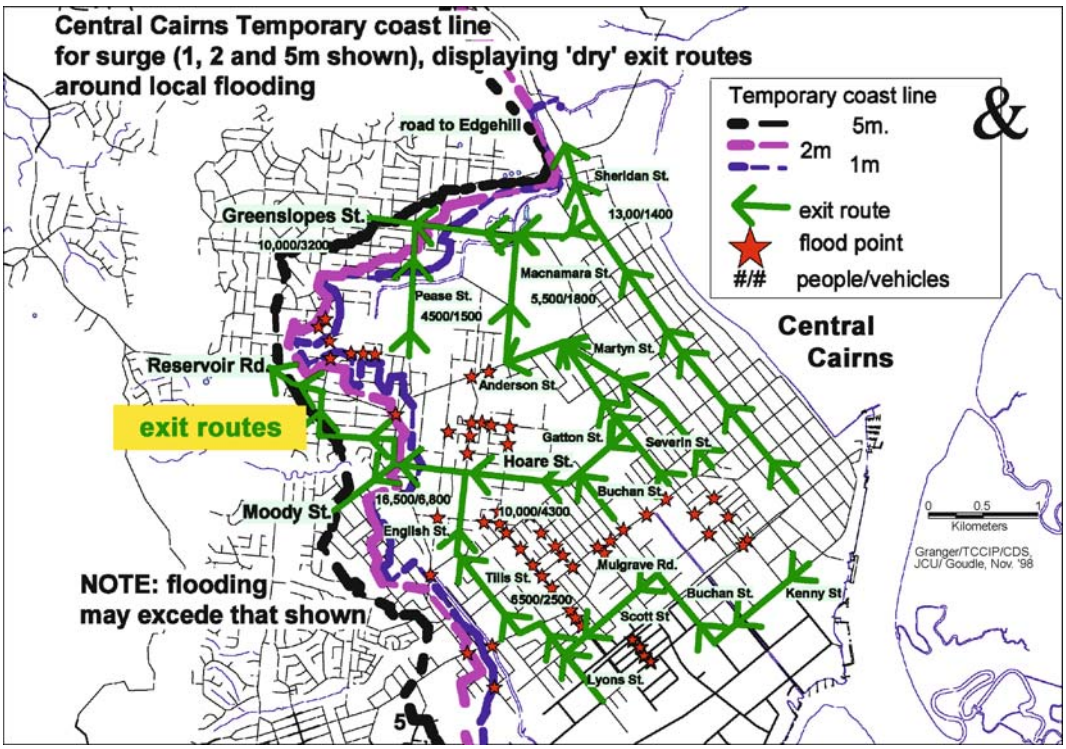

Evacuation as a Communication and Social Phenomenon, Figure 10
Cairns flood prone evacuation routes and schemata of wave reach and temporary coast line from 1998 information

wide and up to 3 m above normal sea level. It may flow overland for perhaps 4 h, as destructive winds to 280 km/h tear at structures, and churning seas; both laden with pounding debris, behave as battering rams and missiles. Because cyclones have such long lead times from modern electronic detection to landfall, results of this deadly combination [60] will be widespread destruction and loss of infrastructure, but not life, with precautionary evacuations [38,72]. At the macro-observation scale, people and vehicles in motion within confining environments tend to behave like fluids, capable of modeling. Constrictions of flow paths cause congestion, as described by Helbing et al., (► Pedestrian, Crowd and Evacuation Dynamics), so early and precautionary evacuation will help minimize the likelihood of grid-lock.

Land-based flooding may be a core issue to effective evacuation [34,38,95], with up to 40,000 people in the vulnerable central city area and northern beach suburbs needing early, precautionary evacuation [79], Fig. 10. Cairns City Council now has a storm surge map on its public web site ([101], posted in 2006), like the public and informative flood map for the Redlands Shire, SE of Brisbane, Queensland ([102]).

Within the philosophical and moral frame of people's "right to know", these local governments are using the internet to inform people they are in a hazard zone, the first step in making the threat real to those residents. They have made the paradigm shift to providing fine-detailed background information which says to residents: "you are in a hazard zone, you may need to do things. Listen for warnings and be prepared to act in a precautionary way."

Figure 11 shows that after a devastating flood in 1997, the flood-affected residents of Cloncurry, NW Queensland town would do things differently, with better warnings, when faced with rising flood waters. Remote automatic flood monitoring devices were requested, but the local downpour over a fully flooded, vast and flat landscape appeared to be the cause of the flood rising 2 m higher than any flood in the prior hundred years. Precautionary evacuation of the low-lying homes would have prevented much heartache and loss of valued possessions [52].

In 2005, north Queensland was threatened by cyclone Ingrid. Newspaper-reading residents were left in no doubt about the threat (Fig. 12); a good example of the clear warning role played by the media, and non-language image used to convey meaning (Sect. "Discussion").

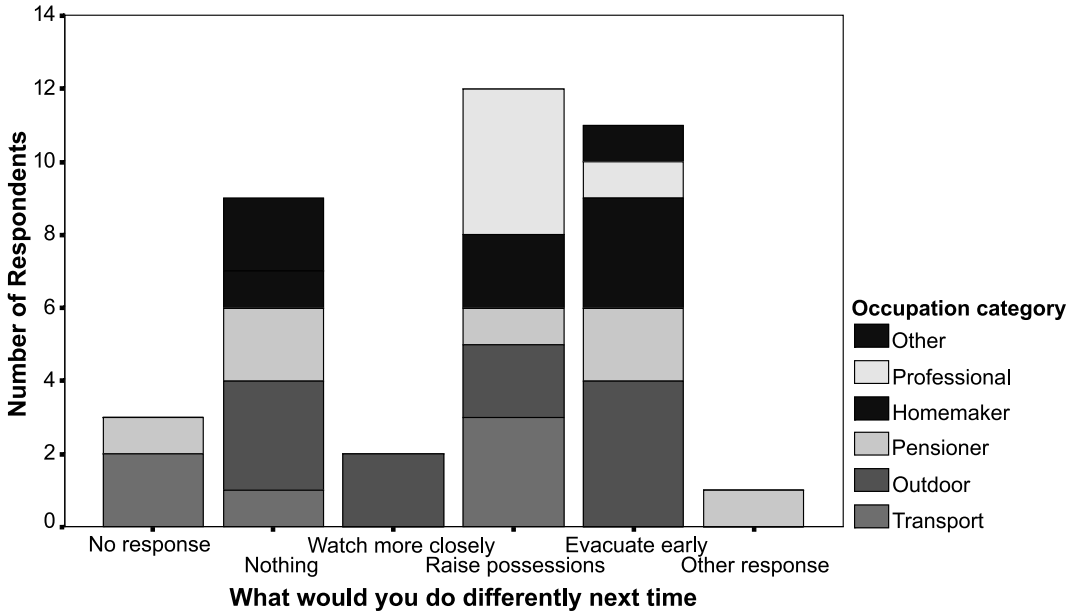

Evacuation as a Communication and Social Phenomenon, Figure 11
Cloncurry reflective of likely responses to future floods

A portion of a Weather Bureau warning media bulletin for Cyclone Ingrid follows, showing, verbatim, what the nation radio broadcaster, the Indigenous Radio Network and most responsible media outlet relayed on. This is clear information, including the possible impacts (M& M of formula 1).

> Media: For immediate broadcast. Transmitters in the area Cape Grenville to Cooktown are requested to use the Standard Emergency Warning Signal.
>
> TOP PRIORITY
>
> TROPICAL CYCLONE ADVICE NUMBER 14
>
> Issued by the Bureau of Meteorology, Brisbane
>
> Issued at 10:56 am on Wednesday the 9th of March 2005
>
> A Cyclone WARNING is current for communities between Cape Grenville and
>
> Cooktown. The warning extends inland across central Cape York Peninsula.
>
> A Cyclone WATCH is current for coastal and island communities on the eastern Gulf of Carpentaria between Weipa and Kowanyama.
>
> The watch south to the Gilbert River Mouth has been canceled.
>
> At 10:00 am EST SEVERE TROPICAL CYCLONE Ingrid, Category 4, with central pressure 935 hPa, was located near latitude 13.5 south longitude 145.5 east, which is about 140 km northeast of Cape Melville and 260 km east of Coen. The cyclone was moving westward at 11 km/h.
>
> Severe Tropical Cyclone Ingrid poses a serious threat to the far north.
>
> Queensland coast with very destructive wind gusts to 280 km/hr near the center. Gales are expected to develop between Cape Grenville and Cooktown during the afternoon. Destructive winds are expected between Coen and Cape Flattery overnight. The very destructive core of the cyclone is expected near the coast between Coen and Cape Melville on Thursday morning.
>
> Coastal residents between Coen and Cape Flattery are specifically warned of the dangerous storm tide as the cyclone crosses the coast early Thursday. The sea is likely to rise steadily to a level significantly above the highest tides of the year with damaging waves, strong currents and flooding of low-lying areas extending some way inland. People living in areas likely to be affected by this flooding should be prepared to evacuate if advised to do so.

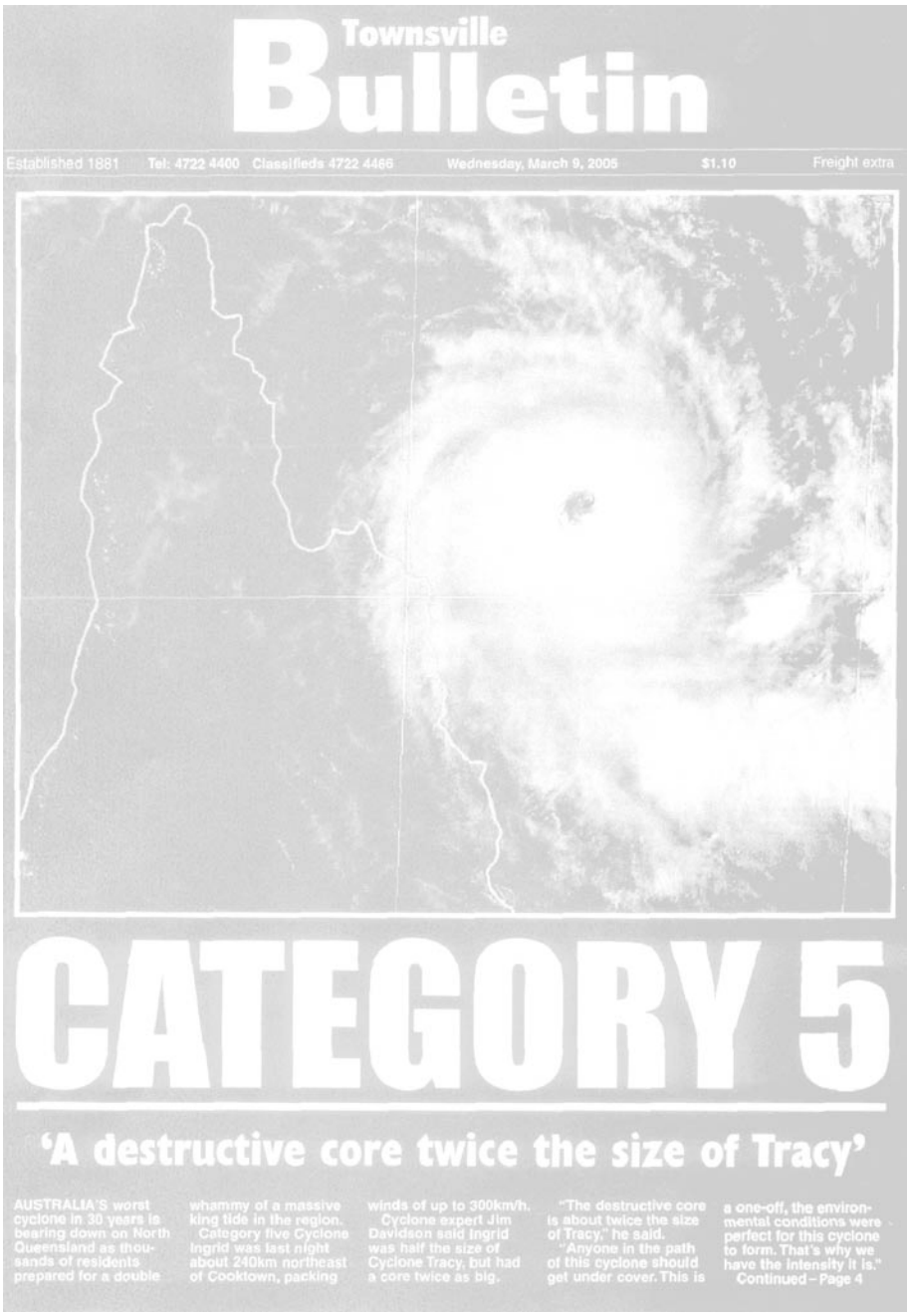

Townsville Bulletin

Established 1881 Tel: 4722 4400 Classifieds 4722 4466 Wednesday, March 9, 2005 $1.10 Freight extra

CATEGORY 5

'A destructive core twice the size of Tracy'

AUSTRALIA'S worst cyclone in 30 years is bearing down on North Queensland as thousands of residents prepared for a double whammy of a massive king tide in the region.

Category five Cyclone Ingrid was last night about 240km northeast of Cooktown, packing winds of up to 300km/h.

Cyclone expert Jim Davidson said Ingrid was half the size of Cyclone Tracy, but had a core twice as big.

"The destructive core is about twice the size of Tracy," he said.

"Anyone in the path of this cyclone should get under cover. This is a one-off, the environmental conditions were perfect for this cyclone to form. That's why we have the intensity it is."

Continued – Page 4

Evacuation as a Communication and Social Phenomenon, Figure 12
Media coverage of threatening Cyclone Ingrid, 2005

> Very heavy rain can be expected to develop on the coast and ranges north of Cooktown.

Research in Port Douglas, NQ was conducted immediately following Cyclone Ingrid, to help clarify the 'Boy who cried wolf' hypothesis about 'concern fatigue' over precautionary evacuations (Sect. "Institutional Barriers to Greater Community Self-Help"). Feedback from marine tourist operators and tourist businesses that removed about 60 large vessels to safe, up-creek moorings, or fully

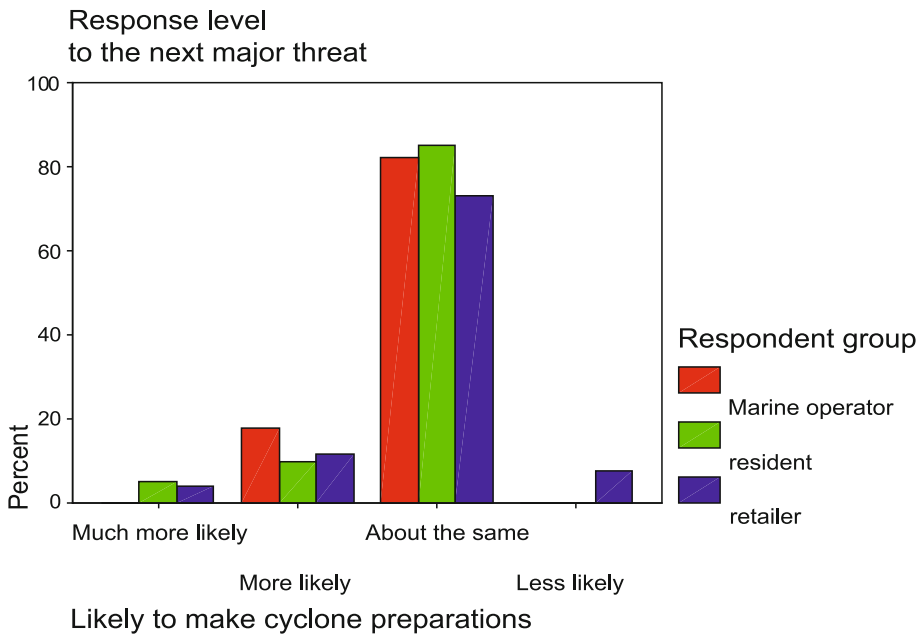

Evacuation as a Communication and Social Phenomenon, Figure 13
People of Port Douglas appreciated a precautionary evacuation

shuttered their premises and raised all stock. These preparations were arduous, and costly. The cyclone veered away from Port Douglas, with virtually no impact. Despite that, people were glad of the precautionary evacuation, as a practice (Fig. 13), reinforcing the findings of the Netherlands flood study [43]. People appreciate practice, as part of the new paradigm of a precautionary approach to hazards.

From Those Impacted by Larry, 2006

After Cyclone Larry, a powerful Category 4/5 severe cyclone which damaged the Innisfail region, Goudie lead a social science research team into the impact area, interviewing households for 4 days (King et al. [53,54]), from the 150 householders survey, focused on risk communication, we learned: as with fire zone residents and those who experienced Cyclone Larry, that most people, once they have experienced a major disaster, maintain a healthy respect and inclination to act ahead of any future such threat. Hence the advocated merit of encouraging people with experience to 'tell their story'; and community (Media and Message and Community Resilience of formula 1) to help make the threat real to others. The next section is, perhaps, at the leading edge of where disaster management will go: living community self-help. Such communities can provide pilot locations to 'calibrate' formula 1.

Woodgate Beach and Community Safety Groups (CSGs)

This section reports a collaborative process between the author and many others, mainly residents of Woodgate Beach, Queensland (Fig. 5), to develop a *Preparation and Evacuation Plan* for residents, formal response groups, the Local Government Disaster Management Group (LDMG) and Shire Council.

Through the 18 month consultation and development process, the plan needed to be realistic and achievable, relying mainly on locals working together under the LDMG and local SES group, aiming to optimize each element of formula 1.

Community networking and self-responsibility for community safety is profoundly developed in the isolated, 700 strong, Central-coast Queensland settlement. The volunteer community safety groups were made up of dedicated people. Goudie led development of a community-based evacuation plan, aiming to: ensure maxim preparation and minimum impact on property and residents; optimizing formula 1.

There were 5 meetings, of up to 80 residents and representatives of all formal response groups, weather and earthquake experts. Meetings included researcher-led two-day 'table-top' evacuation exercise. Woodgate Beach did not develop an evacuation plan, but a preparations and evacuation plan:

Preparations and Evacuation Self-Help Approach

To nurture aware, informed residents preparing to be safe through, and recover from natural disaster impacts.

Overarching Evacuation Approach

- Identify vulnerable areas or houses
- Evacuate caravan park, visitors, people at risk unable to easily move themselves

The Planning Process:

1. DEFINE THE THREATS – Fire, flood, wind, cyclone surge, earth movement, tsunami.
2. MOVE FROM WHERE TO WHERE? Fires can be fought; floods, storm surge and cyclonic winds, tsunami or earthquake cannot be. In all threats, make sure your property is as secure from impact as possible.
3. IDENTIFY THE VULNERABLE – Which buildings, infrastructure and people may be in the threat zone.

Threats and Treatments An all-of-community approach has developed an annual round of public education projects, press releases and pamphlets to inform residents and tourists of the annual cycle of dangers to be wary of, from cyclones to fire care and management. The Council newsletter will be a consistent source of information on threats and how to minimize risks (M & M), from the flying debris of a wind storm to being patient at a flood-swollen creek crossing.

A sequenced approach, where the aged and vulnerable are moved first from the highest risk areas will be taken, as a practice, as a precaution. To highlight the stages of disaster and possible evacuation planning, the background preparations phase is included in this article:

Background Preparations

1. Woodgate Beach residents recognize and act on the need for background preparations to minimize the impacts of all hazard impacts. This includes property maintenance and upkeep. This maximizes PPR of formula 1.
2. Provide newcomers with an information pack, including a copy of extracts of this Community preparations and evacuation plan (CPEP). All community members, including tourists, the elderly, infirm, and needy are incorporated into this CPEP. This enhances the KB of formula 1.
3. Provide dot points on evacuation for local residents in the *Disaster Preparedness Information Kit*, delivered to each household.
4. Expose tourists to the essence of local threats and what will be expected of them: leave early, unless they or their vehicles can actively help, under direction.
5. Define safe shelters – preferably with friends or compatible households. Organizing possible billets for any major impact on portions or all of Woodgate Beach can form a key function of the Community Safety Group. This is the CR of formula 1.
6. Go through whole plan, and address matters like the caravan park needing auxiliary power for fuel pumping before the cyclone season.

The Community Safety Group Approach to Disaster Management This approach is detailed to provide a guide for researchers to use as a framework for other communities:

The Community Safety Group

Purpose encourage:

1. Early Warning Alert The CSG is an affiliation of existing community groups and neighborhood-level residents who make first contact with 'walking-distance' neighbors as soon as anyone hears of a warning that a natural disaster may be approaching their area.

2. Final Preparations (Ramp-Up) Activation The neighbor-level CSG will provide early local motivation for final safety preparations.

Recovery – Thinking Through to Recovery (TR) of formula 1.

Recovery is now seen as part of the preparations package, rather that just looking to minimizing impact. The developing approach is to see the whole threat event as one continuous process: from awareness and structural preparedness, through initial communication of threat, to final precautionary preparations and impact and rapid recovery to a fully functional community.

International Snaps

The Center for Disaster Development within The Northumbria University, Newcastle on Tyne, coastal north east England specializes in recovery and response with an emphasis on development long term recovery and resilience-building. Embedded in this approach is to undertake mitigation. Interviewed by Goudie in May 2005, the Director reported "We use the approach that local knowledge is always drawn on, and that people involved should be in control", KB and CR of formula (1) and agrees "all disaster management is under the umbrella of sustainable development". Other interviewed staff support precautionary

evacuation as a practice, and, independent of the media, an alert signal for people to tune in to the media.

The Shetland Islands

Like Australia's emerging approach to self-help communities, The Shetland Council's Emergency Management Planner (2005 interview) gave strong practical support to the approach of an informed, aware, self-activating community ready to act on reliable, clear, how-to emergency warnings, ranging from making safe if intrinsically out of the main impact zones, to early 'self-evacuation'. This convergence from differently evolved and slightly different disaster management systems is grounds for optimism that the evacuation approach and formula expounded in this article has widely applicable merit.

Remote communities like the Shetlands have lessons for the mainstream in taking responsibility and perforce being self reliant and oriented to robust self-help. Such isolated communities represent matured examples of the informed, aware communities, predisposed to precautionary action that mainstream populations now aspire to. Community-building is an international aspiration, achievable in urban settings.

Hurricane Katrina

Hurricane Katrina, USA late August, 2005 (Fig. 14) has deliberately not been included in this evacuation analysis. With days of clear warning, general formal approaches to precautionary action which underpin the CST and the seven steps to safety and recovery were underplayed.

Figure 14 is included to remind all readers that hazards are real threats to real people in real hazard zones. If all the parameters to maximizing formula 1, such as Institutional Barriers are not optimal to safety, great distress and loss can ensue.

This section has provided a cross-section of disaster threats, evolving reasons to empower communities, and some factors to include in modeling greater community safety. After discussion in the following section, some recommendations and future directions are provided.

Modeling not only the natural hazard but the various social and communication parameters will provide a good basis for not only simulating impact; say, of the extent of flood waters; but will also highlight which impact areas are the most and least likely to properly act in their own best safety.

Discussion

Risk Communication Theory and Residents in Hazard Zones

People need to know that an impact is possible before they will willingly evacuate. The concept of risk characterization [42,46,68,75,77,83] makes clear that people need a practical understanding (*the possibility of impact is real,*

Evacuation as a Communication and Social Phenomenon, Figure 14
What no-one should have to stay through – Katrina '05 http://news.bbc.co.uk/1/hi/in_pictures/4194032.stm

I fully accept that) to then illuminate practical choices. 'Internalizing' that *a threat is real and what you need to do to maximize your safety* may be the core goal of risk communication. This internalization and safety-oriented action may lead to a choice of precautionary evacuations. People need to make informed decisions, thus leading to decision-driven activity.

Delivering Real-Time Warnings

The idea of subjective uncertainty [74] is well displayed by fire-zone residents who, despite all authority efforts to have them commit to an early decision to stay and defend or to leave their properties early, recurrently reported that they would decide on the day whether they would stay or go. There are tragic and recent international examples of resident's decisions made on too little understanding or information.

The bushfire 'stay or go' decision point explored in 2007 research by the author with Australian fire-zone residents clarifies that decision impact on those under threat – packing up their valuables, children and pets and fleeing their house, almost certainly putting it at risk of burning down; as opposed to staying in their house and experiencing the terror of the developing bushfire, was a decision many preferred to delay. Further, many acknowledged that to leave early with all of the disruption, only to learn that the fire was not an actual threat to their property helped induces people to delay that decision point.

Subjective uncertainty is a psychological problem. The outcome of current research is to encourage the threat information gatherers and providers; the weather bureau and disaster managers, to refine information detail and use all currently available modes of dissemination to provide the threatened with the maximum amount of fine details to make that evacuation decision in an informed and timely manner. This is a clear example of where theory and the practical views of hazard zone residents converge.

The 2007 southeast Australian fire zone research shows that people who are more obviously at risk from bushfire are far more prepared than people who are at risk from an occasional but as potentially destructive bushfire. The literature [13,24,31] shows that the less frequent the event, the less prepared people are likely to be. Aligned to the goals of this publication, the formula explained in this article on the realities and complex elements of combining physical with human geography to the 'social good' of maximizing safety around disaster impacts is a great potential application of complexity and systems science. This complexity includes the Media (Media Support MS of formula 1).

The Media

The media plays an ongoing role in socialization and the development of normative values. Dominick [20] argues that the media plays a key role in the cognitive development of individuals. Cognition is the act of coming to know something. Mass media can play a defining role in people's awareness and responses to disaster threats. The CST embraces local media as an active agent in providing needed local information. Local media outlets can draw on and enhance Internet information to inform readers and listeners of 'how to' actions to maximize their safety. In this way the mass media can help clarify facts to help make contentious decisions (which may result in life or major property loss) by providing the fine details to reduce the uncertainty surrounding decisions that householders under threat need to make. The seven steps to community safety (Sect. "Integrating Theory and Implementation") underline this core need for individuals in hazard zones to accept they are at risk and for the Internet and local media to act as providers of relevant, local, timely information for informed decision-making.

The media can play a powerful role in mobilizing communities [58], but they need to have accurate and timely information from disaster managers. The information which fire zone residents in southeastern Australia in 2007 asked for from the Internet is in full keeping with theory, and a recognition that greater access to current facts will reduce uncertainty in decision-making.

Triggering Action: Modeling and Simulating This Geographic, Communications and Psychological Complexity

The reason step one in the seven step process is so important is to counteract a psychological defense against accepting threat. We need to have a sense of personal invulnerability to get us through each day [75]. If we were scared of all possible problems we would cease to operate. The risk literature from Handmer, Goudie, Rohrmann, Salter [34,42,75,77] and many others underlines the importance of the clarity in the description of the threat and safety oriented action. Thus Acceptance they are In a Hazard Zone (AIHZ) of formula 1 is central to possible consequent safety-oriented behavior.

One reason given for risk communication failing is that it is not clear what should be done. The seven steps include a requirement that 'what to do' information be an integral part of any warning as a prelude to evacuation. The clear message is that the process of people excepting that a risk is real needs to well precede any actual impact. This is underlined by Kasperson & Stallen [49], who also

stress the importance of time perception and time horizons. Hence the importance of Knowledge Base of the hazard and safety-maximizing behavior (KB) of formula 1.

Timely Preparation, Final Actions and Evacuations

Timeliness of warnings is paramount to the effectiveness of warnings [85]. In the early 1990s there was much psychology research on perceptions of risk probabilities, individual psychometric studies on perceptions of future time, time orientation, planning horizons, and prediction of future events. In 2007 Australian fire zone residents say they want fine, detailed information as a threat approaches, so the uncertainty of what they may face is minimized. Their planning horizons are generally well oriented to maximum safety, but the crucial 'stay or go' decision will be based on information as close to impact as allows them adequate time to act safely.

Svenson [85] and others make clear that it is not only what you may chose to do, but also when you do it. In the case of remote area warnings of flooding, for instance, when flood waters may take a week to block down-catchment roads, traveling earlier than planned may be the best way to avoid the fate of the truck shown in Fig. 15. This is an example of an 'Active warning image', along with the image of the person standing on the car roof in the flooded road crossing (Fig. 16). Images like these are of use to remind people to travel before or after expected flooding – not during the flood.

Evacuation as a Communication and Social Phenomenon, Figure 15
Nearly crossing a flooded road – Tully NQ 2004, From: Townsville Bulletin, 28/4/4.

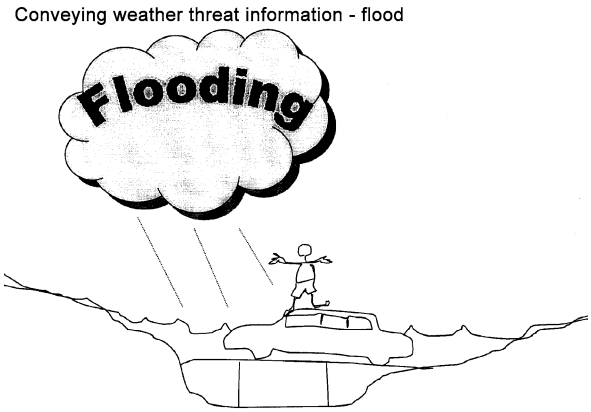

Evacuation as a Communication and Social Phenomenon, Figure 16
An image dissuading people from driving in to flood waters

In 2007, in a two day threat and evacuation exercise in Woodgate Beach operated on the underlined imperative, from the outset, that evacuation ahead of a storm surge would need to be completed not as cyclonic winds struck, but six hours before landfall. Once the winds gusted above 100 km per hour, all peoples, including disaster managers, must be in safe shelter above cyclone surge height.

Disasters and evacuations are issues of people, information, time and space. People in threat zones are entitled to accurate and timely localized information, detailing the threat and likely time-linked movement. People also want to know, via the internet or battery operated radios, what authorities are doing. That information may influence their own actions. What they ask for is the enhancement of their 'risk decision landscape' [83].

In 1974 communication management of and responses to disasters in Darwin and Brisbane shocked Australians. Darwin's Cyclone Tracy killed 65 people, and the Brisbane floods killed 16 [10,11]. These catastrophes taught Australian emergency planners much about the importance of effective warnings, sharing honest, complete and open information in a timely way to emergency managers, emergency workers, and those at risk to ensure sound preparation and responses. The Australian natural disasters of 1974, with major flooding elsewhere in Australia that year, also reinforced the importance of community and family ties to get people through the often profound emotional trauma allied with major natural disasters. The detailed documentation of Cyclone Tracey can help guide development of predictive models of disaster preparedness, communications and responses. With so few well document cases, a Bayesian Logic approach could be used, some well documented disasters, communications and response cases to develop the model, others to test and refine it.

The more recent and current disaster impacts all point to the importance of communities being properly informed. Indeed, the literature on knowledge, risk and inclusion of residents in the bushfire planning process is well described by Goldammer [32]. Issues of increased encroachment on to bush edges along with climate change are causing increased international concern over fire threats, often to places without much collective wisdom over the reality of those fire threats, or the nature of the strategies needed to maximize community safety.

South Eastern Australia is seen as one of the most bushfire-prone environments in the world [92]. Boura [7], an Australian fire manager, describes the development of community fireguard groups in Victoria. These groups are managed and instructed by the Country Fire Authority, but they are residents within very localized neighborhoods. They have their counterparts in the ACT and some other Australian States. They draw on detailed, often technically advanced information from the weather bureau and fire agencies to increase local awareness and action. This is a real playing out of the theories of social amplification developed by psychologists and risk communication theorists [73]; advanced planning techniques [16] and aligned with the way many feel about the direction of emergency management in Australia [15].

Linking the People with the Message

The foci on culture, community and social frameworks, either formal or informal neighborly links, is considered by Douglas [21]. Douglas argues that there is a need for people to understand that risk and danger exist where they live, despite a low probability that an impact may occur in any given hazard season. There is a large body of psychology which considers why people ignore clear messages which may maximize their safety [12]. Douglas speaks of artificially distorted world views. Douglas posits such bias is rooted in over-simple views of heroic and bourgeois fiction. Changing such a normative world view has been discussed in this article in terms of a paradigm shift, well displayed by the Australian Government [13]. A model constructed from formula 1 needs to include all the subtle complex issues of social profiling.

Woodgate Beach (Sect. "Experiences and Lessons – Some Case Studies"), perhaps because of physical isolation, is fully embracing self-help, also displaying a deep culture of volunteerism that can be nurtured and emulated in 'urban villages'. Sustainable urban planning concepts of nodes or activity centers are now well evolved as the hub of 'urban villages' [35]. What is happening in Woodgate Beach can be used as a model for any formula of developed social collective will to self-help, ultimately whether urban and surrounded by other neighborhoods, or more physically isolated.

Like the Ferny Creek (Victoria) community who agitated for a fixed bushfire siren after three of their neighbors were burned to death in the flash fire of 1997, community action needs a few individuals of vision and drive, residents of 'neighborhoods' can initiate and embrace safety-oriented behaviors and structures. The best thing the insti-

Evacuation as a Communication and Social Phenomenon, Figure 17
Initiatives to empower residents to be prepared for fire

tutional systems can do is draw out, nurture and encourage these individuals to mobilize a focus on gaining threat knowledge and acting to increase community responses. Part of the community engagement in the fact they are in a threat zone, as step 1 to action, is shown in Fig. 17; a great road-side banners initiative by one section of the Victorian Country Fire Authority to make the fire threat and implications real for fire-prone residents. Figure 17 shows, in few words, that if you intend to evacuate, do it early. Figure 17 says that the support and creativity to maximize communication effectiveness (the opposite of Institutional Barriers) helps contribute to a positive Medium and Message (MM) of formula 1.

Traditional Weather Warnings

Traditional weather warning signs include ant movement for looming flood, and general bird movement for strong winds, including cyclones [64,87,88]. On Palm Island, when the birds and animals go quiet, it signals that a major storm may be on the way [36,39]. The buildup to the summer monsoon rains is the universal experience of still, humid and hot conditions, continuing in intensity and cloud until the rains come. All cultures in all hazard zones have their traditional knowledge. Cultural Knowledge Bases (KB of formula 1) needs to be modeled into any mathematical prediction of likely propensity to respond to a natural disaster threat.

Words and Images

Word and image use are critical to effective communication. A newspaper graphic ahead of the terrorist-threatened 2004 Olympic Games (Fig. 18) conveys much about the world we now occupy, and uses no words.

The need for clear messages, most likely to provoke a precautionary response, is supported by the risk communication literature of Sect. "Effective Risk Communication" and the communication and cognitive theory of Sect. "Integrating Theory and Implementation". If this knowledge is melded to requests of remote Indigenous peoples and NESH, the safety goal of clear, plain words and images will become the risk communication norm. There are many examples of images conveying meaning, such as the internationally used figure running upstairs, with an arrow; signifying an evacuation route. The symbol for tsunami; a stylized figure racing up a steep slope with a very large wave following, threatening to engulf them is also without words, but conveys all we need to know about evacuating ahead of a tsunami: get up slope immediately.

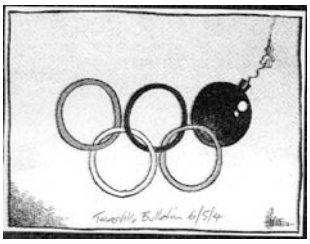

Evacuation as a Communication and Social Phenomenon, Figure 18
An image of terror in 2004 – Source: Townsville Bulletin

Images and Warning Maps

Risk communication is usually about attempts to prompt considered action by a person or community. Effective communication should make the future threat real in present thinking. Alternative responses should be outlined, along with likely consequences. To further prompt a considered and active response to the 'action warning', the consequences of inaction or a range of defensive actions should be made lucid. 'Preferred' behavior should seem reasonably attractive to 'target individuals' [85]. The theoretical overview of risk communication in Sect. "Effective Risk Communication" explains why the message must be clear to the target audience. It needs to have some cognitive content to get people thinking about how it may impact on them, and what the alternative outcomes for them may be if the predicted impact strikes where they are. Stimulus and local risk simulations, as embedded in the Redlands shire flood map [102], should be saturation broadcast into all hazard zones. The CST should be attempted with at-risk groups; people in threat zones, encouraging them to properly think about the real threat and to indicate to people what inaction may bring. A range of safety-oriented actions should also be presented.

Modeling Risk Communication into and Through Communities

Efforts have been made since the 1960s to see how well people understood the hazard. These studies continue [5]. The model of hearing, understanding, believing and feeling that the information is personalized so that those at risk will act has been well understood since the early 90s [81]. The general issues of credibility [18,69,70] still apply. Studies reported by Sorenson and Mileti [81] show increased knowledge as a result of risk communication efforts. People do become more aware of hazards and their personal place within the hazard threats. Unfortunately the link between knowledge and behavior remains tenuous [35].

Knowing that the normal first warning of major disruptive weather comes via the evening TV news in remote settlements, simulations will carry a high embedded likelihood of safety-oriented community response.

Sorenson and Mileti [81] showed that the believability of warnings increases as people get more warnings from officials with high credibility, and that women tend to believe emergency warnings more than men. They also showed that people higher up the socioeconomic ladder tend to believe warnings more than their counterparts. Minority groups have lower than average belief in warnings while people with a high knowledge of the hazard tend to find warnings more credible.

New Cultures to Hazard Zones

The issues discussed in this article from remote Indigenous communities and from non-English-speaking households make clear the importance of using plain English. Work with representatives of both groups made clear that plain English and images were necessary to convey the concept of danger to people in hazard zones, concurring with the expositions of Douglas [21]. The simpler the language the better. Douglas argues that even the word 'risk' could be dropped and 'danger' used instead. Goudie's work with recent Somali arrivals in Australia showed that, like Japanese tourists and many other people from non-English-speaking backgrounds, they understood the word and concept of 'danger', but did not, for instance, understand the word 'severe'. The importance of language clarity *to the intended audiences* in the MM expression of formula 1 needs emphasis.

For risk communication and evacuation advice to penetrate to all people in hazard zones, the language and images must be clear, simple, and compelling. If risk communicators construct campaigns of engagement which work for such 'marginalized' groups, the 'mainstream' can be easily included in that campaign.

Playing the Odds

Renn and Rohrmann [74] make clear that the assumption that risk judgments and evaluations are universal processes independent of social status or national heritage are unreasonable. Goudie's Australian research with multicultural organizations shows most clearly that there are many populations in hazard zones and that each of those groups must be addressed in ways that make the hazard real to them. Findings from the Cyclone Larry research [53] also show that there are many groups other than cultural, such as 'social isolates' – the socially disenfranchised, who need special attention around disaster threats to ensure their safety.

Douglas [21] argues that risk perception and thus response is also an issue of moral and political issues. Westerners tend to consider the probability of an impact in terms of gambling whereas other cultures may need other triggers to internalize the threat and motivate them to action. The concept of a low probability event needing to be taken seriously is discussed by Douglas and others. Residents of Florida are well used to a near-annual evacuation ahead of frequent cyclones. Residents of the coastal communities of Queensland may be less likely to take the warning seriously, because they believe that there is no real chance that they will be personally affected by a cyclone. The whole issue of low probability and high impact events needs to be stressed to relevant impact zone residents.

The first and hardest step in an effective community evacuation remains convincing people in the projected impact zone that the looming threat is real. Renn and Rohrmann [74] suggest that to help make that threat real, project the expected number of fatalities or the catastrophic potential or where the threat may come from.

Spreading the Warning

Handmer [43] recommends that the professional warning agencies should attempt to harness the "personal informational networks of individuals within formal communication systems, and by assuring that formal warning advice is consistent with local norms and behavior" ([43] p 27). Shifting the normative values of recalcitrant disaster managers, and residents in ignorance or denial in urban risk zones becomes the key task of the *seven steps to community safety*, along with the amalgamating approach of the *community safety triangle*.

Part of the strength of the CST is that it taps into and informs an existing social predisposition for people to talk to each other, particularly in times of common threat (King and Goudie 2006). Using emergent technologies to provide real-time information to community media and households helps realize aware informed communities, predisposed to action.

With literature and research results discussed throughout this article, this discussion underlines the importance of modeling local norms, and interactive, safety-focused behavior. Emerging concepts which may be tested in further research are Community Safety groups (Sect. "Experiences and Lessons – Some Case Studies") and Social Burnoffs or flood evacuation practices. The main thrust of this work is to encourage interested modelers to accurately simulate how communities understand and act on

the need for safety-oriented action, where people are inclusive at the neighborhood level, acting as a self-preserving group, no matter what the ethnicity or state of surrounding populations. Neighborhood cohesion and empowerment is current social policy consistent with ESD philosophies and principles, enshrined in current planning laws such as IPA 1997 [46].

To illustrate the power of the internet, consider that any online reader globally can click on [103] and see the social aspects of Queensland's sustainability planning law. Indeed, click on and draw from any of the web sites given. If we were not suffering from information overload and time paucity, the paradigm shift to sustainability implementation would be rapid, the modeling acted on by disaster managers and financial cost-benefit analysts. Until that time, models will help convince and guide people in testable and demonstrable ways to increase community safety. Our very busyness is perhaps the main barrier to reducing global warming and natural disaster impacts. An alternative way into a more sustainable future is centered on local needs-meeting, including nurturing more local community cohesion and sharing of reliable, safety-oriented information and safety-oriented actions.

Future Directions in Modeling Disaster Preparedness and Evacuations as a Social Issue

As global warming and climate change intensifies, the need to model for preparedness and evacuations will increase with more frequent and extreme weather events. The total and fine detail of all needed background preparations and ramp-up preparations are too numerous and arduous for formal disaster management organizations to implement alone, hence the increased need to promote and strongly support the role of community engagement, of community empowerment and nurturing self-help in maximizing effective natural disaster preparation, including evacuations.

Australia has matured disaster policy, law and evolving practice, all embedded in concepts of ecologically sustainable development. Researchers can access, model and test local applicability of some of the Australian experiences and culture of community self-help. Sustainability Implementation Research will become universal as the era of last-minute organizational flurries whilst goading an ignorant and passive population at threat is superseded by prepared and bonded communities who are primed to receive well-delivered warnings, often sourced from the web, and who move themselves to safety in plenty of time. Modeling this will help sell the approach to conservative disaster managers.

The Sustainability Implementation Research, including data-based simulations introduced in this article, is the logical next step to 'action research' of the social sciences and will become the norm in approaching all issues of sustainability where the policies are mature, but agencies are unsure how to implement the paradigm shift; the behavioral and technological shift, to agreed long-term goals.

Meaningful consultation is a defining requirement of sustainability planning and good modeling, so all future social research of merit will take the Human Geographer's approach and "ask the local residents, involve the local residents." Community empowerment, and its modeling means local residents are entitled to help mold their own future. With hazard management, the shared SIR goal is to empower hazard-zone communities to be aware of the threat, be basically prepared for any warning, and act, as a community, in a precautionary way to maximize safety, minimize loss, and speed recovery.

As we move forward, plain language(s), clear, widely displayed hazard maps and images of like hazard impacts will help energize hazard-zone residents to the threats and their own needed safety-oriented behavior.

Researchers and Complexity and Systems Scientists have a moral if onerous obligation to challenge any layer of government clearly exercising or imposing barriers to helping people internalize any threats, then be supported in getting safe and staying safe, recognizing that change may threaten some individuals or sections within bureaucracies. As disaster management moves from a militaristic model of command and control to community empowerment and self help, the developing potential of the web as a key information source into hazard zones – ultimately web-to-air – will bear the fruit of greater community safety and minimized loss. Web-to-TV in disaster warning will be a fine conduit for showing residents-at-risk where the threat may be in relation to their home in a few hours. That immediacy of moving images is a compelling motivator for safety-oriented action.

In future, researchers, modelers and others involved in maximizing community safety will embrace some variations of the *communications safety triangle* and *the seven steps to community safety*, simply because they make sense; are highly cost-effective and easily web-refined from international to local conditions, populations and threats.

Acknowledgment

I most thank my research guardian and mentor over 15 years, Prof. David King, Director of both the Australian Center for Disaster Studies and of the Center for Tropi-

cal Urban and Regional Planning. David allowed me freedom to develop as an 'evidence-based' scientist, to conceive core approaches to sustainability implementation research; productive in helping render positive change in both disaster management and sustainability planning.

Thanks to Australian Bureau of Meteorology staff for their interactive support to improve risk communications, listening to what real people in real hazard zones experience, how they hear warnings, and how the medium and the messages can be and are optimized. The Bureau embraces the core goal to motivate safety-oriented action by people in hazard zones. The Bureau listens and improves the message and the delivery.

The bushfire research of '06 & 07 was funded by the Australian Bushfire Cooperative Research Center, supported by the University of Tasmania.

The 14 years of research reported in this article was not possible without the contributions of Authorities and more than 1000 Australians, old and new, who opened their organizations or doors to myself or research team members and shared their hazard experiences and specific warning needs. Thanks all.

Bibliography

Primary Literature

1. ACF (2004) Institutions for Sustainability. Australian Conservation Foundation ACF 1–37. http://www.acfonline.org.au/uploads/res/res_tp007.pdf. Accessed 2008
2. AMCORD (1995) AMCORD95, Australian Model Code of Residential Development. Department of Housing and Regional Development. Aust Govt Printing Service, Canberra
3. Australian Bushfire CRC (2005) Bushfire CRC/Research/Community Self Sufficiency for Fire Safety/Effective Risk Communication. http://www.bushfirecrc.com/research/c41/c41.html. Accessed 2008
4. Baram M (1991) Rights and duties concerning the availability of environmental risk information to the public. In: Kasperson RE, Stallen PJM (eds) Communicating Risks to the Public – International Perspectives. Kluwer, Dordrecht/Boston/London, p 481
5. Berry L, King D (1998) Tropical cyclone awareness and education issues for far north Queensland school students – Storm Watchers. Aust J Emerg Manag 13:6
6. Boughton GN (1992) Education on Natural Hazards, The Macedon Digest. Aust J Disaster Manag 7(2):4–7
7. Boura J (1998) Community Fireguard: Creating partnerships with the community to minimise the impact of bushfire. Aust J Emerg Manag 13:59–64
8. Bruntland GH (1988) Our Common Future. The World Commission for the Environment and Development. Alianza Publications. http://tilastokeskus.fi/abo2004/foredrag/hoglund_pp.pdf, http://www.erf.es/eng/empresa/brundtland.html, http://web.uvic.ca/~stucraw/Lethbridge/MyArticles/Brundtland.htm. Accessed 2008
9. Cairns City Council (2007) Storm surge maps. http://www.cairns.qld.gov.au/cairns/files/StormTideMaps/index.pdf. Accessed 2008
10. Chamberlain ER, Hartshorn AE, Muggleston H, Short P, Svensson H, Western JS (1981a) Queensland flood report Australia Day (1974). Australian Government Publishing Service, Canberra, p 38
11. Chamberlain ER, Doube L, Milne G, Rolls M, Western JS (1981b) The Experience of cyclone Tracy. Australian Government Publishing Service, Canberra, p 150
12. Cialdini R, Reno R, Kallgren C (1990) A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. J Pers Soc Psychol 58(6):1015–1026
13. COAG (2004) Natural Disasters in Australia, reforming mitigation, relief and recovery arrangements. http://www.ema.gov.au/agd/EMA/rwpattach.nsf/VAP/(756EDFD270AD704EF00C15CF396D6111)~COAG+Report+on+Natural+Disasters+in+Australia+-+August+2002.pdf/$file/COAG+Report+on+Natural+Disasters+in+Australia+-+August+2002.pdf. Accessed 2008, Council of Australian Governments Commonwealth of Australia, 201
14. Cohen E, White P, Hughes P (2006) Bushfire and the Media Reports 1–3. Latrobe University and BCRC, La Trobe University
15. Cronan K (1998) Foundations of emergency management. The Aust J Emerg Manag 1(13):20–23
16. Cuthill M (2004) Community well-being – the ultimate goal of democratic governance. Qld Plan 44(2):8–11
17. Cutter SL, Boruff BJ, Shirley WL (2003) Social Vulnerability to Environmental Hazards. Soc Sci Q 84(2):242–261
18. DIMIA (2004) Department of Immigration, Multicultural and Indigenous Affairs. http://www.humanrights.gov.au/racial_discrimination/face_facts/mig_htm.htm#q. 2005
19. Dolphin RR, Richard R, Ying F (2000) Is Corporate Communications a Strategic Function. Manag Decis 38(2):99–106
20. Dominick JR (1993) The dynamics of mass communication. Mc Graw-Hill Inc, Columbus, p 616
21. Douglas M (1992) Risk and Blame. Essays in cultural theory. Routledge, London/New York
22. Drabek TE (1994) Disaster evacuation and the tourist industry. Program on environment and behaviour monograph 57. University of Colorado, Colorado
23. Dunlap RE, Van Liere KD (1978) The 'new environmental paradigm': a proposed measuring instrument and preliminary results. J Envtl Ed 9(4):10–19
24. ECAP (1997) Guidelines for disaster prevention and preparedness in tropical cyclone areas. Economic Commission for Asia and the Pacific, the World Meteorological Organisation and the Red Cross Societies, Geneva/Bangkok
25. EMA (2000) Emergency Risk Management Applications Guide. The Australian Emergency Manuals Series. Emergency Management Australia. Dickson ACT:Emergency Management Australia, p 4
26. EMA (2002) Research agenda for Emergency Management. March. EMA Research and development Strategy for "Safer Sustainable Communities". Dickson ACT:Emergency Management Australia
27. EMA (2002) Indigenous Communities and Emergency Management. Emergency Management Australia, Canberra, p 22
28. EMA (2002) Guidelines for Emergency Managers working with Culturally and Linguistically Diverse Communities. http://www.ema.gov.au/ema/rwpattach.

nsf/viewasattachmentpersonal/AFD7467016783EA8CA256CB30036EF42/$file/caldsept2002.pdf. Accessed 2008
29. EMAI (1998) Emergency Management Australia Information Service. Report of the strategic planning conference on the development of enhanced awareness education programs and materials for remote Aboriginal and Torres Strait Islander communities. Darwin, May 1997. Conference Proceedings, EMA, Mt Macedon Victoria
30. Fien J (1993) Education for the environment – critical curriculum theorising and environmental education. Deakin University, Geelong, Victoria Australia
31. Finnis K, Johnston D, Paton D (2004) Volcanic Hazard Risk Perceptions in New Zealand. Tephra, Earth and atmospheric sciences University of AlbertaEdmonton, Alberta, Canada, pp 60–64, http://www.civildefence.govt.nz/memwebsite.nsf/Files/Tephra%20v21%20chapters/$file/volcanichazardrisk.pdf, 2008
32. Fowler KL, Kling ND, Larson MD (2007) Eigenvalues. Bus Soc 46:1. March. 88–103. Sage Publications
33. Goldammer J (2005) Wildland fire – rising threats and vulnerabilities at the interface with society. In: Know risk. United Nations Tudor Rose, T. Jeggle, United Nations, ed; UN International Strategy for Disaster Reduction (UN-ISDR), Geneva, 376 p, pp 322–3
34. Goudie D, King D (1999) Cyclone surge and community preparedness. Austn J Emerg Manag 13:1:454–60. http://www.ema.gov.au/agd/EMA/rwpattach.nsf/viewasattachmentpersonal/(85FE07930A2BB4482E194CD03685A8EB)~Cyclone_surge_and_community_preparedness.pdf/$file/Cyclone_surge_and_community_preparedness.pdf, 2008
35. Goudie D (2001) Toward Sustainable Urban Travel. Ph D thesis. James Cook University. http://eprints.jcu.edu.au/967/, 2008
36. Goudie D (2004) Disruptive weather warnings and weather knowledge in remote Australian Indigenous communities. Web-based report http://www.tesag.jcu.edu.au/CDS/Pages/reports/Gou_iwwrpt/index.shtml?Id=23, 2008
37. Goudie D (2005) Sustainability planning: pushing against institutional barriers. Ecosystems and Sustainable Development V. WIT Press. WIT Transactions on Ecology and the Environment 81(5):215–224, www.witpress.com, 2008
38. Goudie D (2007) Transport and Evacuation Planning. In: King D, Cottrell A (eds) Communities Living With Hazards. Centre for Disaster Studies, James Cook University with Queensland Department of Emergency Services, 293, James Cook University, North Queensland Australia, pp 48–62
39. Goudie D (2007) Oral histories about weather hazards in northern Australia. In: King D, Cottrell A (eds) Communities Living With Hazards. Centre for Disaster Studies, James Cook University with Queensland Department of Emergency Services, 293, James Cook University, North Queensland Australia, pp 102–125
40. Gurmankin AD, Baron J, Armstrong K (2004) Intended message versus received message in hypothetical physician risk communication: exploring the gap. Risk Analysis 24(5):1337–1347
41. Handmer J (1992) Can we have too much warning time? A study of Rockhampton, Australia. The Macedon Digest. Aust J Disaster Manag 7:2 p 8–10
42. Handmer J (2000) Are Flood Warnings Futile? Risk Commun emergencies 2(e):1–14. http://www.massey.ac.nz/~trauma/issues/2000-2/handmer.htm, 2008
43. Handmer J (2001) Improving flood warnings in Europe: a research and policy agenda. Environ Hazards 3(2001):19–28
44. Heller K, Alexander DB, Gatz M, Knight BG, Rose T (2005) Social And Personal Factors As Predictors Of Earthquake Preparation: The Role Of Support Provision, Network Discussion, Negative Affect, Age, and Education. J Appl Soc Psychol 35(2):399–422
45. Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian Model Averaging: A Tutorial. Stat Sci 14(4):382–417
46. IPA (1997) The Integrated Planning Act. Queensland Government, http://www.legislation.qld.gov.au/LEGISLTN/CURRENT/I/integplana97.pdf, 2008
47. ISR (2007) Australian Policyonline. Institute for Social Research, Swinburne University of Technology, http://www.apo.org.au/linkboard/results.chtml?filename_num=117732, 2008
48. Jarach M (1989) Overview of the literature on barriers to the diffusion of renewable energy sources in agriculture. Appl En 32(2):117–131
49. Kasperson RE, Stallen PJM (1991) Communicating Risks to the Public International Perspectives. Kluwer, Dordrecht/Boston/London
50. Kim YC, Ball-Rokeach SJ (2006) Civic engagement from a communication infrastructure perspective. Commun Theory 16:173–197
51. Kitchin RM (1996) Increasing the integrity of cognitive mapping research: appraising conceptual schemata of environment-behaviour interaction. Progress Hum Geogr 20(1):56–84
52. King D, Goudie D (1998) Breaking through the disbelief – the March 1997 floods at Cloncurry. Even the duck swam away. Aust J Emerg Manag 4:12 29–33
53. King D, Goudie D (2006) Cyclone Larry, March 2006 Post Disaster Residents Survey. Centre for Disaster Studies, James Cook University, with the Australian Bureau of Meteorology P77 http://www.tesag.jcu.edu.au/CDS/Pages/reports/Larry_mainreport.pdf, 2008
54. King D, Goudie D, Dominey-Howes D (2006) Cyclone knowledge and household preparation – some insights from Cyclone Larry Report on how well Innisfail prepared for Cyclone Larry. The Aust J Emerg Manag 21:3 52–59 http://www.ema.gov.au/agd/EMA/rwpattach.nsf/VAP/(A80860EC13A61F5BA8C1121176F6CC3C)~AJEM_EMA_Larry_Aug2006.pdf/$file/AJEM_EMA_Larry_Aug2006.pdf. Accessed August 2006
55. Kobb P (2000) Emergency Risk Management Applications Guide. Emerg Manag Aust, Dickson, A.C.T.: Emergency Management Australia. Australian Emergency Manuals Series; 05 http://www.ema.gov.au/agd/EMA/rwpattach.nsf/VAP/(383B7EDC29CDE21FBA276BBBCE12CDC0)~Manual+05a.pdf/$file/Manual+05a.pdf, 2008
56. Leibovitz J (2003) Institutional barriers to associative city-region governance: the politics of institution-building and economic governance in 'Canada's Technology Triangle.' Urban Studies: 40(13):2613–2642
57. Lewis C (2006) Risk Management and Prevention Strategies. Aust J Emerg Manag 21(3):47–51

58. Lichtenberg J, Maclean D (1991) The role of the media in risk communication. In: Kasperson RE, Stallen PJM (ed) Communicating Risks to the Public – International Perspectives. Kluwer, Dordrecht/Boston/London, p 481
59. Lidstone J (2006) Blazer to the Rescue! The role of puppetry in enhancing fire prevention and preparedness for young children. Aust J Emerg Manag 21(2):17–28
60. Loudness RS (1977) Tropical Cyclones in the Australian Region July 1909 to June 1975. AGPS, Canberra
61. McKenna F (1993) It won't happen to me: Unrealistic optimism or illusion of control. Brit J Psych 84:39–50
62. Munro DA (1995) Ecologically sustainable development – is it possible? How will we recognise it? In: Sivakumar M, Messer J (eds) Protecting the future – ESD in action. Futureworld, Wollongong
63. NDSI Working Group (2000) Effective Disaster Warnings. Working Group on Natural Disaster Information Systems. Subcommittee on Natural Disaster Reduction National Science and Technology Council Committee on Environment and Natural Resources. Executive Office of the President of the United States of America, 56, http://www.incident.com/cap/docs/NDIS_rev_Oct27.pdf, 2008
64. Napurrurlarlu NO, Jakamarrarlu NP (1988) Ngawarra-Kurlu. Yuendumu B.R.D.U., Darwin, p 19
65. O'Neill P (2004) Why don't they listen – Developing a risk communication model to promote community safety behaviour. The International Emergency Management Society, 11th Annual Conference Proceedings, Melbourne, Victoria, Australia, May 18–21 2004, pp 160–169
66. Young J, O'Neill P (1999) A social marketing framework for the development of effective public awareness programs. http://www.ses.nsw.gov.au/multiattachments/2740/documentname/A_Social_Marketing_Framework_for_the_Development_of_Effective_Public_Awareness_Programs.pdf, 2008. NSW Australia
67. Paton D (2003) Stress in Disaster Response: A risk management approach. Disaster Prev Manag 12(3):203–209
68. Paton D, Smith L, Johnston D (2005) When good intentions turn bad: promoting natural Hazard preparedness. Aust J Emerg Manag 20:25–30
69. Phillips R (1994) Long Range Planning. Lond 27(4):143–145
70. QG & QES (2003) State Planning Policy. Mitigating the adverse impacts of flood, bushfire and landslide. State planning policy 1/03. Dept Local Government and planning, & Dept of Emergency Services, http://www.emergency.qld.gov.au/publications/spp/pdf/spp.pdf, 2008
71. Quarantelli EL (2002) The role of the Mass Communication system in natural and technological disasters and possible extrapolation to terrorism situations. Risk Manag Int J 4(4):7–21
72. Raggatt P, Butterworth E, Morrissey S (1993) Issues in Natural Disaster Management: Community Response to the Threat of Tropical Cyclones in Australia. Disaster Prev Manag 2(3): 12–21
73. Renn O, Levine D (1991) Credibility and trust in risk communication. In: Kasperson RE, Stallen PJM (ed) Communicating Risks to the Public – International Perspectives. Kluwer, Dordrecht/Boston/London, p 481
74. Renn O, Rohrmann B (2000) Cross-cultural risk perception, A survey of empirical studies. Kluwer, Dordrecht, p 240
75. Rohrmann B (2000) A socio-psychological model for analysing risk communication processes. Australas J Disaster Trauma Stud 2000(2) http://www.massey.ac.nz/~trauma/issues/2000-2/rohrmann.htm, 2008
76. Rounsefell V (1992) Unified human settlement ecology. In: Birkeland J (ed) Design for sustainability: A sourcebook of integrated eco-logical solutions. Earthscan Publications, London, S4.2:78–83
77. Salter J (1992) The Nature of the disaster – more than just the meanings of words: Some reflections on definitions, doctrine and concepts. The Macedon Digest. Aust J Disaster Manag 7(2):1–3
78. Salter J, Bally J, Elliott J, Packham D (1993) Natural disasters: protecting vulnerable communities. In: Merriman PA, Browitt CW (eds) Conference Proceedings. London, 13–15 October 1993, Royal Society (Great Britain)
79. Sheppard E (1986) Modelling and predicting aggregate flows. In: Hanson S (ed) The geography of urban transportation. Guilford Press, New York/London, pp 91–110
80. Skertchly A, Skertchly K (2000) Message Sticks – Hazard Mitigation Visual Language. EMA, ACT, Dickson, Australian Capital Territory
81. Sorenson J, Mileti D (1991) Risk communication in emergencies. In: Kasperson RE, Stallen PJM (ed) Communicating Risks to the Public – International Perspectives. Kluwer, Dordrecht/Boston/London, p 481
82. Stern P (1992) Psychological dimensions of global environmental change. Ann Rev Psychol 43:269–302
83. Stern PC, Fineberg HV (1996) Understanding Risk, Informing decisions in a democratic society. National Academy Press, Washington DC, pp 249
84. Sullivan M (2003) Communities and their experience of emergencies. Aust J Emerg Manag 18(1):19–26
85. Svenson O (1991) The time dimension in perception and communication of risk. In: Kasperson RE, Stallen PJM (ed) Communicating Risks to the Public – International Perspectives. Kluwer, Dordrecht/Boston/London, p 481
86. Thompson KM (2002) Variability and uncertainty meet risk management and risk communication. Risk Analysis 22(3):647–654
87. Utemorrah D, Clendon M (2000) Dumbi the owl. In: Kimberley Language Resource Centre (ed) Worrorra Lalai, Worrorra Dreamtime Stories. KLRC, Halls Creek WA, pp 113
88. Utemorrah D (1980) How The People Were All Drown(ed). In: Mowanjum (ed) Visions of Mowanjum: Aboriginal writings from the Kimberley. Rigby, Adelaide
89. Wakefield SE, Elliot SJ (2003) Constructing the news: the role of local newspapers in environmental risk communication. Prof Geogr 55(2):216–266
90. Wall M (2006) The case study method and management learning: making the most of a strong story telling tradition in emergency services management education. Aust J Emerg Manag 21(2):11–16
91. Walmsley DJ (1988) Urban living, the individual in the city. Longman scientific and technical, Longman, London, p 104
92. Woods F, Gabriel P (2005) Individual responsibility and statewide strategies: bushfire in Victoria, Australia. In: Know risk. United Nations Tudor Rose, UK, Jeggle T (ed) Leicester, LE1 5RA, UK 376. pp 326–8
93. Yates J (1992) Assisting the community to plan: A pilot program in Western Australia. The Macedon Digest. Aust J Disaster Manag 7(2):12

94. Yates J (1997) Federalism and disaster mitigation in remote Aboriginal communities in Western Australia. Spring, AJEM, 25–32, Emergency Management Australia, Mt Macedon Victoria Australia. Publisher: Grey Worldwide Canberra Australia
95. Yeo S (2002) Natural Hazards. Flooding in Australia: A review of events in 1998. 25:177–191. Department of Physical Geography, Macquarie University, NSW, http://www.springerlink.com/content/n160g0121800n742/ Natural Hazards 25(2):177–191, 2002. Kluwer Academic Publishers. Printed in the Netherlands. http://www.springerlink.com/content/n160g0121800n742/fulltext.pdf
96. Zamecka A, Buchanan G (2000) Disaster risk management. Queensland Department of Emergency Services, Brisbane, p 115
97. Kikuchi T, Nakamori Y (2007) Agent model analysis to explore effects of interaction and environment on individual performance. J Syst Sci Complexity 2007 20:1–17. Springer Science + Business Media, Inc
98. http://news.bbc.co.uk/1/hi/world/europe/6963373.stm BBC (2005) Access to 2008, In pictures: Hurricane onslaught
99. http://news.bbc.co.uk/2/hi/americas/7055721.stm BBC (2007) Californians flee as fires rage
100. Mornington Peninsula Shire Council (2006) Fire wise fire management. http://www.mornpen.vic.gov.au/Files/FireWiseFireManagementBooklet.pdf
101. Cairns City Council (2006) Storm tide maps. http://www.hazardsaustralia.info/Mapping.html
102. http://maps.redland.qld.gov.au/website/redemapexternal%5Fv2%5F03/Default.aspx, Redlands Shire Accessed 2008 Flood Maps
103. Queensland Government (1997) Integrated planning act. http://www.legislation.qld.gov.au/LEGISLTN/CURRENT/I/integplana97.pdf

Books and Reviews

Bushnell S, Cottrell A, Spillman M, Lowe D (2006) Thuringowa bushfire case study. Understanding Communities. Project BCRC Program C Community self-sufficiency for fire safety. Bushfire CRC, JCU CDS, Townsville, Australia

Granger KJ, Smith DI (1995) Storm tide impact and consequence modelling: some preliminary observations. Math Comput Model 21:9:15–21

Roth W (1897) Ethnological Studies among the North West Central Queensland Aborigines. Queensland government, Brisbane/London

FEMA (2008) Prepare for a Wildfire. Federal Emergency Management Agency. U.S. Department of Homeland Security Washington, DC http://www.fema.gov/hazard/wildfire/wf_prepare.shtm

Australian Bureau of Meteorology (2008) Protecting yourself and your home. Bushfire weather. BoM. Melbourne Australia http://www.bom.gov.au/inside/services_policy/fire_ag/bushfire/protect.htm

Emergency Management Australia (2003). Community safety Bushfire action guide. EMA. Dickson, Australian Capital Territory http://www.ema.gov.au/agd/ema/emainternet.nsf/Page/RWP07C6046B98D07DB8CA256C5A00230553

ABC (2006) Bushfire summer. Australian Broadcasting Commission. Melbourne, www.abc.net.au/bushfire

Geoscience Australia (2005) Sentinel. Commonwealth of Australia. Canberra, ACT, Australia, http://sentinel.ga.gov.au/acres/sentinel/index.shtml

Rural Fire Service (2008) Fire Safety Information. New South Wales Rural Fire Service. NSW Government, Sydney, http://www.rfs.nsw.gov.au/dsp_content.cfm?CAT_ID=515 NSW Rural Fire Service, with information in 27 languages

CFA (2008) Country Fire Authority. Melbourne, Victoria, Australia, http://www.cfa.vic.gov.au

ESA (2008) Community Education. Emergency Services Agency, Australian Capital Territory, http://www.esa.act.gov.au/esawebsite/content_esa/community_education/community_education.html

Evacuation Dynamics: Empirical Results, Modeling and Applications

Andreas Schadschneider[1,2], Wolfram Klingsch[3], Hubert Klüpfel[4], Tobias Kretz[5], Christian Rogsch[3], Armin Seyfried[6]

[1] Institut für Theoretische Physik, Universität zu Köln, Köln, Germany
[2] Interdisziplinäres Zentrum für Komplexe Systeme, Bonn, Germany
[3] Institute for Building Material Technology and Fire Safety Science, University of Wuppertal, Wuppertal, Germany
[4] TraffGo HT GmbH, Duisburg, Germany
[5] PTV Planung Transport Verkehr AG, Karlsruhe, Germany
[6] Jülich Supercomputing Centre, Research Centre Jülich, Jülich, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Empirical Results
Modeling
Applications
Future Directions
Acknowledgments
Bibliography

Glossary

Pedestrian A person traveling on foot. In this article, other characterizations are used depending on the context, e. g., agent or particle.

Crowd A large group of pedestrians moving in the same area, but not necessarily in the same direction.

Evacuation The movement of persons from a dangerous place due to the threat or occurrence of a disastrous event. In normal situations this is called "egress" instead.

Flow The flow or current J is defined as the number of persons passing a specified cross-section per unit time. The common unit of flow is "persons per second". Specific flow is the flow per unit cross-section. The maximal flow supported by a facility (or a part of it) is called "capacity".

Fundamental diagram In traffic engineering (and physics): density-dependence of the flow: $J(\rho)$. Due to the hydrodynamic relation $J = \rho v b$ equivalent representations used frequently are $v = v(\rho)$ or $v = v(J)$. The fundamental diagram is probably the most important quantitative characterization of traffic systems.

Lane formation In bidirectional flows, lanes are often dynamically formed in which all pedestrians move in the same direction.

Bottleneck A limited resource for pedestrian flows, for example a door, a narrowing in a corridor, or stairs, i. e., a location of reduced capacity. At bottlenecks jamming occurs if the inflow is larger than the capacity. Other phenomena that can be observed are the formation of lanes and the zipper-effect.

Microscopic models Models which represent each pedestrian separately with individual properties like walking velocity or route choice behavior and the interactions between them. Typical models that belong to this class are cellular automata and the social-force model.

Macroscopic models Models which do not distinguish individuals. The description is based on aggregate quantities, e. g., appropriate densities. Typical models belonging to this class are fluid-dynamic approaches. Hand calculation methods which are based on related ideas and are often used in the field of (fire-safety) engineering belong to this class as well.

Crowd disaster An accident in which the specific behavior of the crowd is a relevant factor, e. g., through competitive and non-adaptive behavior. In the media, it is often called "panic" which is a controversial concept in crowd dynamics and should thus be avoided.

Definition of the Subject

Today, there are many occasions on which a large number of people gathers in a rather small area. Office buildings and apartment houses grow larger and more complex. Very large events related to sports, entertainment or cultural and religious events are held all over the world on a regular basis. This brings about serious safety issues for the participants and for the organizers who must be prepared for any case of emergency or critical situation. Usually in such cases the participants must be guided away from the dangerous area as quickly as possible. Therefore the understanding of the dynamics of large groups of people is very important.

In general, evacuation is egress from an area, a building or a vessel due to a potential or actual threat. In the cases described above, the dynamics of the evacuation processes are quite complex due to the large number of people and their interaction, external factors such as fire, complex building geometries, etc. Evacuation dynamics must be described and understood on different levels: physical, physiological, psychological, and social. Accordingly, the scientific investigation of evacuation dynamics involves many research areas and disciplines. The system "evacuation process" (i. e., the population and the environment) can be modeled on many different levels of detail, ranging from hydro-dynamic models to artificial intelligence and multi-agent systems. There are at least three aspects of evacuation dynamics that motivate its scientific investigation:

1) As in most many-particle systems several interesting collective phenomena can be observed that need to be explained;
2) Models need to be developed that are able to reproduce pedestrian dynamics in a realistic way, and
3) Pedestrian dynamics must be applied to facility design and to emergency preparation and management.

The investigation of evacuation dynamics is a difficult problem that requires close collaboration between different fields. The origin of the apparent complexity lies in the fact that one is concerned with a many-'particle' system with complex interactions that are not fully understood. Typically the systems are far from equilibrium and so are, e. g., sensitive to boundary conditions. Motion and behavior are influenced by several external factors and often crowds can be rather inhomogeneous.

In this article we want to deal with these problems from different perspectives and will not only review the theoretical background, but will also discuss some concrete applications.

Introduction

The awareness that emergency exits are one of the most important factors to ensure the safety of persons in buildings can be traced back more than 100 years. Disasters due to the fires in the Ring theater in Vienna and the urban theater in Nizza in 1881 resulted in several hundred fatal-

ities and led to a rethinking of the safety in buildings [24]. First, attempts were made to improve safety by using non-flammable building materials. However, the disaster at the Troquois Theater in Chicago with more than 500 fatalities, where only the decorations burned, demonstrated the need for more effective measures. This was a starting point for studying the influences of emergency exits and thus the dynamics of pedestrian streams [24,32].

In recent years there have been two major evacuation incidents which gained immense global attention. First, there was the capsizing of the Baltic Sea ferry MV Estonia (September 28, 1994, 852 casualties) [100] and, of course, the terrorist attacks of 9/11 (2,749 casualties). Other prominent examples of the possible tragic outcomes of the dynamics of pedestrian crowds are the Hillsborough stadium disaster in Sheffield (April 15, 1989, 96 casualties) [182], the accident at Bergisel (December 4, 1999, 5 casualties) [189], the stampede in Baghdad (August 30, 2005, 1,011 casualties), the tragedy at the concert of "The Who" (December 3, 1979, 11 casualties) [73] and – very early – the events at the crowning ceremony of Tsar Nicholas II in St. Petersburg in May 1896 with 1,300 to 3,000 fatalities (sources vary considerably) [168]. In the past, tragic accidents have happened frequently in Mecca during the Hajj (1990: 1,426, 1994: 270, 1997: 343, 1998: 107, 2001: 35, 2003: 14, 2004: 244, and 2006: 364 casualties). What stands out is that the initiating events are very diverse and range from external human aggression (terrorism) to external physical dangers (fire) and rumors to various shades of greedy behavior in absence of any external danger.

Many authors have pointed out that the results of experts' investigations and the way the media typically reports about an accident very often differ strongly [17,77,109,155,156,178]. Public discussion has a much greater tendency to identify "panic" as the cause of a disaster, while expert commissions often conclude that there either was no panic at all, or panic was merely a result of some other preceding phenomenon.

This article first discusses the empirical basis of pedestrian dynamics in Sect. "Empirical Results". Here we introduce the basic observables and describe the main qualitative and quantitative results, focusing on collective phenomena and the fundamental diagram. It is emphasized that even for the most basic quantities, no consensus about basic behavior has been reached.

In Sect. "Modeling" various model approaches that have been applied to the description of pedestrian dynamics are reviewed.

Section "Applications" discusses more practical issues and gives a few examples for applications to safety analysis. In this regard, prediction of evacuation times is an important problem as legal regulations must often be fulfilled. Here, commercial software tools are available. A comparison shows that the results must be interpreted with care.

Empirical Results

Overview

Pedestrians are three-dimensional objects and a complete description of their highly developed and complicated motion sequence is rather difficult. Therefore, in pedestrian and evacuation dynamics, pedestrian motion is usually treated as two-dimensional by considering the vertical projection of the body.

In the following sections we review the present knowledge of empirical results. These are relevant not only as a basis for the development of models, but also for applications such as safety studies and legal regulations.

We start with the phenomenological description of collective effects. Some of these are known from everyday experience and will serve as benchmark tests for any kind of modeling approach. Any model that does not reproduce these effects is missing some essential part of the dynamics. Next, the foundations of a quantitative description are laid by introducing the fundamental observables of pedestrian dynamics. Difficulties arise from different conventions and definitions. Then pedestrian dynamics in several simple scenarios (corridor, stairs etc.) are discussed. Surprisingly, even for these simple cases no consensus about the basic quantitative properties exists. Finally, more complex scenarios are discussed which are combinations of the simpler elements. Investigations of scenarios such as evacuations of large buildings or ships suffer even more from lack of reliable quantitative and sometimes even qualitative results.

Collective Effects

One of the reasons why the investigation of pedestrian dynamics is attractive for physicists is the large variety of interesting collective effects and self-organization phenomena that can be observed. These macroscopic effects reflect the individuals' microscopic interactions and thus give important information for any modeling approach.

Jamming Jamming and clogging typically occur for high densities at locations where the inflow exceeds capacity. Locations with reduced capacity are called *bottlenecks*. Typical examples are exits (Fig. 1) or narrowings. This kind of jamming phenomenon does not depend strongly on the microscopic dynamics of the

Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 1
Clogging near a bottleneck. The shape of the clog is discussed in more detail in Subsect. "Theoretical Results"

particles. Rather it is a consequence of an exclusion principle: space occupied by one particle is not available for others.

This clogging effect is typical for a bottleneck situation. It is important for practical applications, especially evacuation simulations.

Other types of jamming occur in the case of counterflow where two groups of pedestrians mutually block each other. This happens typically at high densities and when it is not possible to turn around and move back, e. g., when the flow of people is large.

Density waves Density waves in pedestrian crowds can be generally characterized as quasi-periodic density variations in space and time. A typical example is the movement in a densely crowded corridor (e. g., in subway-stations close to the density that causes a complete halt of motion) where phenomena similar to stop-and-go vehicular traffic can be observed, e. g., density fluctuations in a longitudinal direction that move backwards (opposite to the movement direction of the crowd) through the corridor. More specifically, for the situation on the Jamarat Bridge in Makkah (during the Hajj pilgrimage 2006), stop-and-go waves have been reported. At densities of 7 persons per m^2 upstream, moving stop-and-go waves of period 45 s have been observed that lasted for 20 minutes [59]. Fruin reports, that "at occupancies of about 7 persons per square meter the crowd becomes almost a fluid mass. Shock waves can be propagated through the mass sufficient to lift people off their feet and propel them distances of 3 m (10 ft) or more." [36].

Lane formation In counterflow, i. e., two groups of people moving in opposite directions, (dynamically varying) lanes are formed where people move in just one direction [135,139,197]. In this way, strong interactions with oncoming pedestrians are reduced which is more comfortable and allows higher walking speeds.

The occurrence of lane formation does not require a preference of moving on one side. It also occurs in situations without left- or right-preference. However, cultural differences for the preferred side have been observed. Although this preference is not essential for the phenomenon itself, it has an influence on the kind of lanes formed and their order.

Several quantities for the quantitative characterization of lane formation have been proposed. Yamori [197] has introduced a band index which is basically the ratio of pedestrians in lanes to their total number. In [13] a characterization of lane formation through the (transversal) velocity profiles at fixed positions has been proposed. Lane formation has also been predicted to occur in colloidal mixtures driven by an external field [15,28,158]. Here, an order parameter $\phi = \frac{1}{N}\langle\sum_{j=1}^{N}\phi_j\rangle$ has been introduced where $\phi_j = 1$ if the lateral distance to all other particles of the other type is larger than a typical density-dependent length scale, and $\phi_j = 0$ otherwise.

The number of lanes can vary considerably with the total width of the flow. Figure 2 shows a street in the city center of Cologne during World Youth Day in Cologne (August 2005) where two comparatively large lanes have been formed.

The number of lanes usually is not constant and might change in time, even if there are relatively small changes in density. The number of lanes in opposite directions is not always identical. This can be interpreted as a sort of spontaneous symmetry breaking.

Quantitative empirical studies of lane formation are rare. Experimental results have been reported in [94] where two groups with varying relative sizes had to pass each other in a corridor with a width of 2 m. On one hand, similar to [197] a variety of different lane patterns were observed, ranging from 2 to 4 lanes. On the other hand, in spite of this complexity, surprisingly

Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 2
The "Hohe Straße" in Cologne during World Youth Day 2005. The yellow line is the border of the two walking directions

large flows could be measured: the sum of (specific) flow and counterflow was between 1.8 and 2.8 persons per meter per second and exceeded the specific flow for one-directional motion ($\approx$ 1.4 P/ms).

Oscillations In counterflow at bottlenecks, e. g., doors, one can sometimes observe oscillatory changes of the direction of motion. Once a pedestrian is able to pass the bottleneck it becomes easier for others to follow in the same direction until somebody is able to pass the bottleneck (e. g., through a fluctuation) in the opposite direction.

Patterns at intersections At intersections, various collective patterns of motion can be formed. A typical example is short-lived roundabouts which make motion more efficient. Even if these are connected with small detours, the formation of these patterns can be favorable since they allow for "smoother" motion.

Emergency situations, "panic" In emergency situations various collective phenomena have been reported that have sometimes misleadingly been attributed to *panic behavior*. However, there is strong evidence that this is not the case. Although a precise accepted definition of *panic* is missing, usually certain aspects are associated with this concept [77]. Typically "panic" is assumed to occur in situations where people compete for scarce or dwindling resources (e. g., safe space or access to an exit) which leads to selfish, asocial or even completely irrational behavior and contagion that affects large groups. A closer investigation of many crowd disasters has revealed that most of the above characteristics have played almost no role and most of the time have not been observed at all (see e. g. [73]). Often the reason for these accidents is much simpler, e. g., in several cases the capacity of the facilities was too small for the actual pedestrian traffic, e. g., Luschniki Stadium Moskau (October 20, 1982), Bergisel (December 4, 1999), pedestrian bridge Kobe (Akashi) (July 21, 2001) [186]. Therefore the term "panic" should be avoided, *crowd disaster* being a more appropriate characterization. Also it should be kept in mind that in dangerous situations it is *not* irrational to fight for resources (or for your own life), if everybody else does this [18,113]. Only from the outside is this behavior perceived as irrational since it might lead to a catastrophe [178]. The latter aspect is therefore better described as *non-adaptive behavior*. We will discuss these issues in more detail in Subsect. "Evacuations: Empirical Results".

Observables

Before we review experimental studies in this section, the commonly used observables are introduced.

The flow J of a pedestrian stream gives the number of pedestrians crossing a fixed location of a facility per unit of time. Usually it is taken as a scalar quantity since only the flow normal to some cross-section is considered. There are various methods to measure flow. The most natural approach is to determine the times t_i at which pedestrians pass a fixed measurement location. The time gaps $\Delta t_i = t_{i+1} - t_i$ between two consecutive pedestrians i and $i+1$ are directly related to the flow

$$J = \frac{1}{\langle \Delta t_i \rangle} \quad \text{with } \langle \Delta t_i \rangle = \frac{1}{N} \sum_{i=1}^{N} (t_{i+1} - t_i) = \frac{t_{N+1} - t_1}{N} . \tag{1}$$

Another possibility for measuring the flow of a pedestrian stream is borrowed from fluid dynamics. The flow through a facility of width b is determined by the average density ρ and the average speed v of a pedestrian stream as

$$J = \rho v b = J_s b . \tag{2}$$

where the *specific flow* [1]

$$J_s = \rho v \tag{3}$$

gives the flow per unit-width. This relation is also known as *hydrodynamic relation*.

There are several problems concerning the way in which velocities, densities or time gaps are measured and

[1] In strictly one-dimensional motion often a line density (dimension: 1/length) is used. Then the *flow* is given by $J = \rho v$.

the conformance of the two definitions of flow. The flow according to Eq. (1) is usually measured as a mean value over time at a certain location, while the measurement of the density in Eq. (2) is connected with an instantaneous mean value over space. This can lead to a bias caused by underestimation of fast moving pedestrians at the average over space compared to the mean value of the flow over time at a single measurement line (see the discussion for vehicular traffic e. g., in [51,81,102]). Furthermore, most experimental studies measuring the flow according to Eq. (2) combine for technical reasons an *average* velocity of a single pedestrian over time with an *instantaneous* density. To ensure a correspondence of the mean values the average velocity of all pedestrians contributing to the density at a certain instant has to be considered. However this procedure is very time consuming and not realized in practice up to now. Moreover, the fact that the dimension of the test section has usually the same order of magnitude as the extent of the pedestrians can influence the averages over space. These all are possible factors why different measurements can differ in a large way, see discussion in Subsect. "Fundamental Diagram".

Another way to quantify the pedestrian load of facilities has been proposed by Fruin [35]. The "pedestrian area module" is given by the reciprocal of the density. Thompson and Marchant [184] introduced the so-called "inter-person distance" d, which is measured between center coordinates of the assessing and obstructing persons. According to the "pedestrian area module" Thompson and Marchant call $\sqrt{1/\rho}$ the "average inter-person distance" for a pedestrian stream of evenly spaced persons [184]. An alternative definition is introduced in [58] where the local density is obtained by averaging over a circular region of radius R,

$$\rho(\mathbf{r}, t) = \sum_j f(\mathbf{r}_j(t) - \mathbf{r}) \,, \tag{4}$$

where $\mathbf{r}_j(t)$ are the positions of the pedestrians j encompassed by $\mathbf{r}$ and $f(\dots)$ is a Gaussian, distance-dependent weight function.

In contrast to the density definitions above, Predtechenskii and Milinskii [151] consider the ratio of the sum of the projection area f_j of the bodies and the total area of the pedestrian stream A, defining the (dimensionless) density $\tilde{\rho}$ as

$$\tilde{\rho} = \frac{\sum_j f_j}{A} \,, \tag{5}$$

a quantity known as *occupancy* in the context of vehicular traffic. Since the projection area f_j depends strongly on the type of person (e. g., it is much smaller for a child than for an adult), the densities for different pedestrian streams consisting of the same number of persons and the same stream area can be quite different.

Beside technical problems due to camera distortions and camera perspective there are several conceptual problems, such as the association of averaged with instantaneous quantities, the need to choose an observation area in the same order of magnitude as the extent of a pedestrian together with the definition of the density of objects with nonzero extent and much more. A detailed analysis of the ways in which measurement influences the relations is necessary but still lacking.

Fundamental Diagram

The fundamental diagram describes the empirical relation between density ρ and flow J. The name indicates its importance and naturally it has been the subject of many investigations. Due to the hydrodynamic relation (3) there are three equivalent forms: $J_s(\rho)$, $v(\rho)$ and $v(J_s)$. In applications the relation is a basic input for engineering methods developed for the design and dimensioning of pedestrian facilities [35,136,150]. Furthermore, it is a quantitative benchmark for models of pedestrian dynamics [21,86,112,175].

In this section we will concentrate on planar facilities such as sidewalks, corridors and halls. For various facilities such as floors, stairs or ramps, the shape of the diagrams differ, but in general it is assumed that the fundamental diagrams for the same type of facilities but having different widths merge into one diagram for the specific flow J_s. In first order this is confirmed by measurements on different widths [49,135,139,142]. However, Navin and Wheeler observed in narrow sidewalks more orderly movement leading to slightly higher specific flows than for wider sidewalks [135]. A natural lower bound for the independence of the specific flow from the width is given by the body size and the asymmetry in movement possibilities of the human body. Surprisingly, Kretz et al. found an increase of the specific flow for bottlenecks with $b \leq 0.7\,\mathrm{m}$ [93]. This will be discussed in more detail later. For the following discussion we assume facility widths larger than $b = 0.6\,\mathrm{m}$ and use the most common representations $J_s(\rho)$ and $v(\rho)$.

Figure 3 shows various fundamental diagrams used in planning guidelines and measurements of two selected empirical studies representing the overall range of the data. The comparison reveals that specifications and measurements disagree considerably. In particular, the maximum of the function giving the capacity $J_{s,\max}$ ranges from $1.2\,(\mathrm{ms})^{-1}$ to $1.8\,(\mathrm{ms})^{-1}$, the density value where the max-

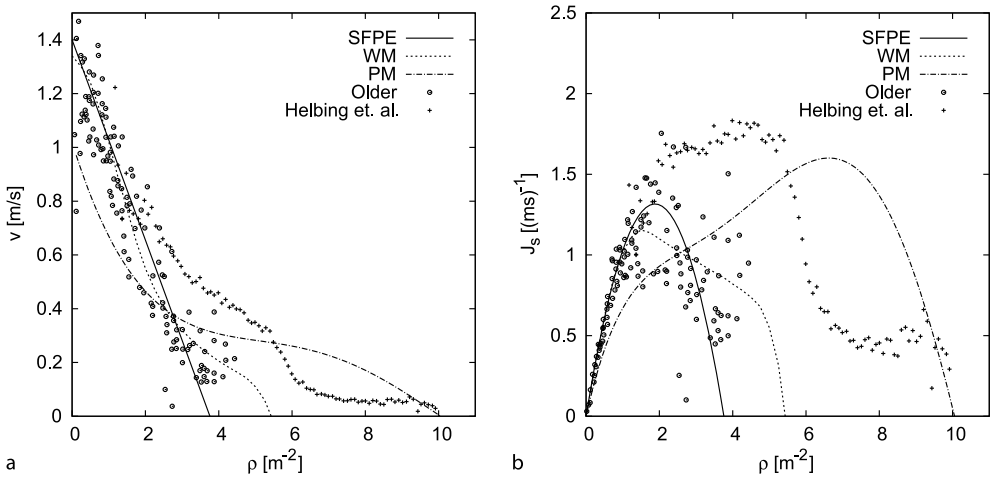

Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 3
Fundamental diagrams for pedestrian movement in planar facilities. The lines refer to specifications according to planning guidelines (SFPE Handbook [136]), Predtechenskii and Milinskii (PM) [150], Weidmann (WM) [192]). Data points give the range of experimental measurements (Older [142] and Helbing [58])

imum flow is reached (ρ_c) ranges from 1.75 m^{-2} to 7 m^{-2} and, most notably, the density ρ_0, where the velocity approaches zero due to overcrowding, ranges from 3.8 m^{-2} to 10 m^{-2}.

Several explanations for these deviations have been suggested, including cultural and population differences [58,116], differences between uni- and multidirectional flow [99,135,154], short-ranged fluctuations [154], influence of psychological factors given by the incentive of the movement [150] and, partially related to the latter, the type of traffic (commuters, shoppers) [139].

It seems that the most elaborate fundamental diagram is given by Weidmann who collected 25 data sets. An examination of the data which were included in Weidmann's analysis shows that most measurements with densities larger then $\rho = 1.8\,m^{-2}$ are performed on multidirectional streams [135,139,140,142,148]. But data gained by measurements on strictly unidirectional streams has also been considered [35,49,188]. Thus Weidmann neglected differences between uni- and multidirectional flow in accordance with Fruin, who states in his often cited book [35] that the fundamental diagrams of multidirectional and unidirectional flow differ only slightly. This disagrees with results of Navin and Wheeler [135] and Lam et al. [99] who found a reduction of the flow in dependence of directional imbalances. Here lane formation in bidirectional flow has to be considered. Bidirectional pedestrian flow includes unordered streams as well as lane-separated and thus quasi-unidirectional streams in opposite directions. A more detailed discussion and data can be found in [99,135,154]. A surprising finding is that the sum of flow and counterflow in corridors is larger than the unidirectional flow and for equally distributed loads it can be twice the unidirectional flow [94].

Another explanation is given by Helbing et al. [58] who argue that cultural and population differences are responsible for the deviations between Weidmann and their data. In contrast to this interpretation the data of Hanking and Wright [49] gained by measurements in the London subway (UK) are in good agreement with the data of Mori and Tsukaguchi [115] measured in the central business district of Osaka (Japan), both on strictly uni-directional streams. This brief discussion clearly shows that up to now there is no consensus about the origin of the discrepancies between different fundamental diagrams and how one can explain the shape of the function.

However, all diagrams agree in one characteristic: velocity decreases with increasing density. As the discussion above indicates there are many possible reasons and causes for the velocity reduction. For the movement of pedestrians along a line, a linear relation between speed and the inverse of the density was measured in [174]. The speed for walking pedestrians depends also linearly on the step size [192] and the inverse of the density can be regarded as the required length for one pedestrian to move. Thus it seems that smaller step sizes caused by a reduction of the available space with increasing density is, at least for a certain density region, one cause for the decrease of speed. However, this is only a starting point for a more elaborated modeling of the fundamental diagram.

Bottleneck Flow

The flow of pedestrians through bottlenecks shows a rich variety of phenomena, e. g., the formation of lanes at the entrance to the bottleneck [64,66,93,176], clogging and blockages at narrow bottlenecks [24,57,93,121,122,150] or some special features of bidirectional bottleneck flow [57]. Moreover, the estimation of bottleneck capacities by the maxima of fundamental diagrams is an important tool for the design and dimensioning of pedestrian facilities.

Capacity and Bottleneck Width One of the most important practical questions is how the capacity of a bottleneck rises with increasing width. Studies of this dependence can be traced back to the beginning of the last century [24,32] and, up to now, have been discussed controversially. As already mentioned in the context of the fundamental diagram there are multiple possible influences on pedestrian flow and thus on the capacity. In the following, the major findings are outlined, demonstrating the complexity of the system and documenting a controversial discussion over one hundred years.

At first sight, a stepwise increase of capacity with the width appears to be natural if lanes are formed. For independent lanes, where pedestrians in one lane are not influenced by those in others, the capacity increases only if an additional lane can be formed. This is reflected in the stepwise enlargement of exit width, which has been a requirement of several building codes and design recommendations. See e. g., the discussion in [146] for the USA and GB and [130] for Germany. e. g.; the German building code requires an exit width (e. g., for a door) to be at least 90 cm plus 60 cm for every 200 persons. Independently from this simple lane model, Hoogendoorn and Daamen [64,66] measured by a laboratory experiment the trajectories of pedestrians passing a bottleneck. The trajectories show that inside a bottleneck the formation of lanes occurs, resulting from the zipper effect occurring on entry to the bottleneck. Due to the zipper effect, a self-organization phenomenon leading to an optimization of the available space and velocity; the lanes are not independent and thus do not allow passing (Fig. 4). The empirical results of [64,66] indicate a distance between lanes of $d \approx 0.45$ m, independent of the bottleneck width b, implying a stepwise increase of capacity. However, the investigation was restricted to two values ($b = 1.0$ m and $b = 2.0$ m) of the width.

In contrast, the study [176] considered more values of the width and found that the lane distance increases continuously as illustrated in Fig. 4. Moreover it was shown that a continuous increase of the lane distance leads to a very weak dependence on its width of the density and velocity inside the bottleneck. Thus in reference to Eq. (2) the flow does not necessarily depend on the number of lanes. This is consistent with common guidelines and handbooks [2] which assume that the capacity is a linear function of the width [35,136,150,192]. It is given by the maximum of the fundamental diagram and in reference to the specific flow concept introduced in Subsect. "Observables", Eqs. (2), (3), the maximum grows linearly with the facility width. To find a conclusive judgment on the question if the capacity grows continuously with the width the results of different laboratory experiments [93,121,122,132,176] are compared in [176].

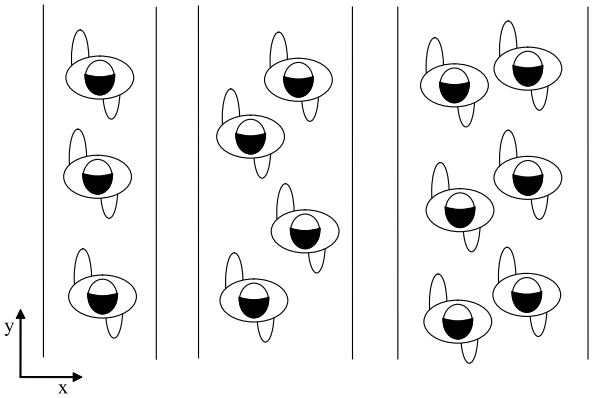

Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 4
A sketch of the zipper effect with continuously increasing lane distances in *x*: The distance in the walking direction decreases with increasing lateral distance. Density and velocities are the same in all cases, but the flow increases continuously with the width of the section

In the following we discuss the data of flow measurement collected in Fig. 5. The corresponding setups are sketched in Fig. 6. First, note that all presented data are taken under laboratory conditions where the test persons are advised to move normally. The data by Muir et al. [121], who studied the evacuation of airplanes (see Fig. 6b), seem to support the stepwise increase of flow with the width. They show constant flow values for $b > 0.6$ m. But the independence of flow over the large range from $b = 0.6$ m to $b = 1.8$ m indicates that in this special setup the flow is not restricted by the bottleneck width. Moreover, it was shown in [176] by determination of the trajectories that the distance between lanes changes continuously, invalidating the basic assumption leading to a stepwise increasing flow. Thus all collected data for flow measurements in Fig. 5 are compatible with a continuous

[2] One exception is the German MVStättV [130], see above.

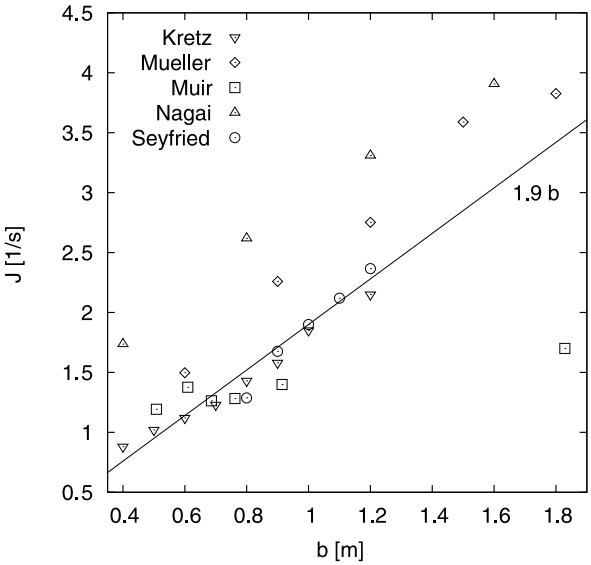

Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 5
Influence of the width of a bottleneck on the flow. Experimental data [121,122,132,176] of different types of bottlenecks and initial conditions. All data are taken under laboratory conditions where the test persons are advised to move normally

and almost linear increase with the bottleneck width for $b > 0.6\,\mathrm{m}$.

The data in Fig. 5 differ considerably in values of bottleneck capacity. In particular, the flow values of Nagai [132] and Müller [122] are much higher than the maxima of empirical fundamental diagrams (see Subsect. "Fundamental Diagram"). The influence of "panic" or pushing can be excluded since in all experiments the participants were instructed to move normally. The comparison of the different experimental setups (Fig. 6) shows that the exact geometry of the bottleneck is of only minor influence on the flow, while a high initial density in front of the bottleneck can increase the resulting flow values. This is confirmed by the study of Nagai et al., see Figure 6 in [132]. There it is shown that for $b = 1.2\,\mathrm{m}$ the flow grows from $J = 1.04\,\mathrm{s}^{-1}$ to $3.31\,\mathrm{s}^{-1}$ when the initial density is increased from $0.4\,\mathrm{m}^{-2}$ to $5\,\mathrm{m}^{-2}$.

The linear dependence of the flow on the width has a natural limitation due to the nonzero body-size and the asymmetry given by the sequence of movement in steps. Movement of pedestrians through bottlenecks smaller than shoulder width requires a rotation of the body. Kretz et al. found in their experiment [93] that the specific flow J_s increases if the width decreases from $b = 0.7\,\mathrm{m}$ to $b = 0.4\,\mathrm{m}$.

Connection Between Bottleneck Flow and Fundamental Diagrams An interesting question is how the bottleneck flow is connected to the fundamental diagram. General results for driven diffusive systems [149] show that boundary conditions only *select* between the states of the undisturbed system instead of creating completely different ones. Therefore it is surprising that the measured maximal flow at bottlenecks can exceed the maximum of the empirical fundamental diagram. These questions are related to the common jamming criterion. Generally, it is assumed that a jam occurs if the incoming flow exceeds the capacity of the bottleneck. In this case one expects the

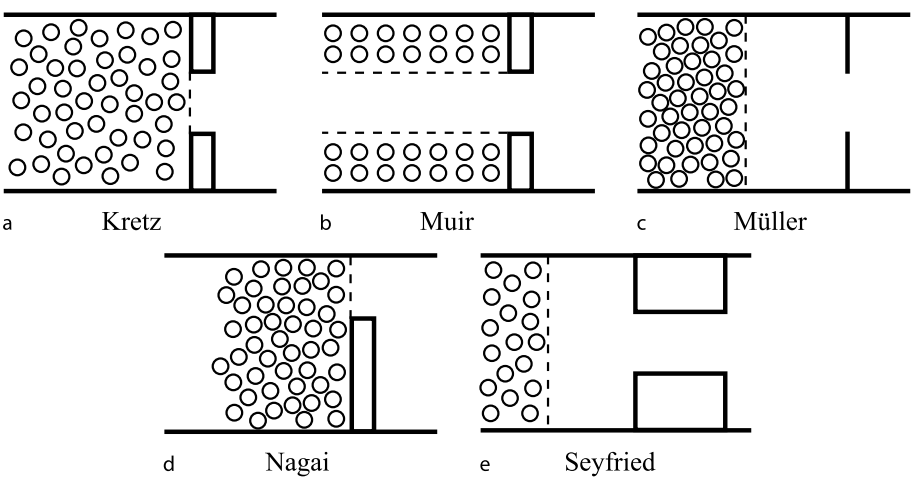

Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 6
Outlines of the experimental arrangements under which the data shown in Fig. 5 were taken

flow through the bottleneck to continue with the capacity (or lower values).

The data presented in [176] show a more complicated picture. While the density in front of the bottleneck amounts to $\rho \approx 5.0(\pm 1)\,\mathrm{m}^{-2}$, the density inside the bottleneck tunes around $\rho \approx 1.8\,\mathrm{m}^{-2}$. The observation that the density inside the bottleneck is lower than in front of the bottleneck is consistent with measurements of Daamen and Hoogendoorn [20] and the description given by Predtechenskii and Milinskii in [150]. The latter assumes that in the case of a jam the flow through the bottleneck is determined by the flow in front of the bottleneck. The density inside the jam will be higher than the density associated with the capacity. Thus the reduced flow in front of the bottleneck causes a flow through the bottleneck smaller than the bottleneck capacity. Correspondingly the associated density is also smaller than that at capacity. But the discussion above cannot explain why the capacities measured at bottlenecks are significantly higher than the maxima of empirical fundamental diagrams and cast doubts on the common jamming criterion. Possible unconsidered influences are stochastic flow fluctuations, non-stationarity of the flow, flow interferences due to the necessity of local organization or changes of the incentive during the access into the bottleneck.

Blockages in Competitive Situations As stated above all data collected in Fig. 5 are gained by runs where the test persons were instructed to move normally. By definition a bottleneck is a limited resource and it is possible that under competitive situations pedestrian flow through bottlenecks is different from the flow in normal situations. One qualitative difference to normal situations is the occurrence of blockages. Regarding the term 'panic' one has to bear in mind that for the occurrence of blockages some kind of reward is essential, while the emotional state of the test persons is not. This was a result of a very interesting and often cited study by Mintz [113]. First experiments with real pedestrians have been performed by Dieckmann [24] in 1911 as a reaction to many fatalities in theater fires at the end of the 19th century. In these small scale experiments test persons were instructed to go through great trouble to pass the door as fast as possible. Even in the first run he observed a stable "wedging". In [150] it is described how these obstruction occurs due to the formation of arches in front of the door under high pressure. This is very similar to the well-known phenomenon of *arching* occurring in the flow of granular materials through narrow openings [194].

Systematic studies including the influence of the shape and width of the bottleneck and comparisons with flow values under normal situations have been performed by Müller and Muir et al. [121,122]. Müller found that funnel-like geometries support the formation of arches and thus blockages. For further discussion, one must distinguish between temporary blockages and stable blockages leading to a zero flow. For the setup sketched in Fig. 6c Müller found that temporary blockages occur only for $b < 1.8\,\mathrm{m}$. For $b \leq 1.2\,\mathrm{m}$ the flow shows strong pulsing due to unstable blockages. Temporal disruptions of the flow appear for $b \leq 1.0\,\mathrm{m}$. In comparison to normal situations the flow is higher, and in general the occurrence of blockages decreases with width. However a surprising result is that for narrow bottlenecks, increasing the width can be counterproductive since it also increases the probability of blockages. Muir et al. for example note that in their setup (Fig. 6b) the enlargement of the width from $b = 0.5\,\mathrm{m}$ to $b = 0.6\,\mathrm{m}$ leads to an increase of temporary blockages. The authors explain this by differences in the perception of the situation by the test persons. While the smaller width is clearly passable only for one person, the wider width may lead to the perception that the bottleneck is sufficiently wide to allow two persons to pass through. How many people have direct access to the bottleneck is clearly influenced by the width of the corridor in front of the bottleneck. Also, Müller found hints that flow under competitive situations did not increase in general with the bottleneck width. He notes an optimal ratio of 0.75:1 between the bottleneck width and the width of the corridor in front of the bottleneck.

To reduce the occurrence of blockages, and thus evacuation times, Helbing et al. [54,55,83] suggested putting a column (asymmetrically) in front of a bottleneck. It should be emphasized that this theoretical prediction was made under the assumption that the system parameters, i. e., the basic behavior of the pedestrians, does not change in the presence of the column. This is highly questionable in real situations where a column can be perceived as an additional obstacle or can make it difficult to find the exit. In experiments [57] an increase of the flow of about 30% has been observed for a door with $b = 0.82\,\mathrm{m}$. But this experiment was performed only for one width and the discussion above indicates the strong influence of the specific setup used. Independent of this uncertainty this concept is limited, as the occurrence of stable arches, to narrow bottlenecks. In practice narrow bottlenecks are not suitable for a large number of people and an opening in a room has other important functionalities, which would be restricted by a column.

Another finding is the observation that the total flow at bottlenecks with bidirectional movement is higher than it is for unidirectional flows [57].

Stairs

In most evacuation scenarios stairs are important elements that are a major determinant for the evacuation time. Due to their physical dimension, which is often smaller than other parts of a building, or due to a reduced walking speed, stairs generally must be considered as bottlenecks for the flow of evacuees. For the movement on stairs, just as for the movement on flat terrain, the fundamental diagram is of central interest. Compared to the latter there are more degrees of freedom, which influence the fundamental diagram:

- One has to distinguish between upward and downward movement.
- The influence of riser height and tread width (which determine the incline) has to be taken into account.
- For upward motion exhaustion effects lead to a strong time dependence of the free speed.

It is probably a consequence of the existence of a continuum of fundamental diagrams in dependence of the incline that there are no generally accepted fundamental diagrams for movement on stairs. However, there are studies on various details—mostly the free speed—of motion on stairs in dependence of the incline [35,38,39,46], conditions (comfortable, normal, dangerous) [151], age and sex [35], tread width [33], and the length of a stair [95]; and in consideration of various disabilities [11].

In addition there are some compilations or "meta studies": Graat [46] compiled a list of capacity measurements and Weidmann [192] built an average of 58 single studies and found an average for the horizontal upstairs speed—the speed when the motion is projected to the horizontal level—of 0.610 m/s.

Depending on various parameters, the aforesaid studies report horizontal upward walking speeds varying over a wide range from 0.391 to 1.16 m/s. Interestingly, on one and the same short stairs it could be observed [95] that people on average walked faster up- than downwards.

There is also a model where the upstairs speed is calculated from the stair geometry (riser and tread) [183] and an empirical investigation of the collision avoidance behavior on stairs [37].

On stairs (up- as well as downward) people like to put their hand on the handrail, i. e., they tend to walk close to walls, even if there is no counterflow. This is in contrast to movement on flat terrain, where at least in situations of low density there is a tendency to keep some distance from walls.

The movement on stairs is typically associated with a reduction of the walking speed. For upward motion this follows from the increased physical effort required. This has two aspects: first, there is the physical potential energy that a pedestrian has to supply if he wants to rise in height; second, the motion process itself is more exertive – the leg has to be lifted higher – than during motion on a level, even if this motion process is executed only on the spot. Concerning the potential energy there is no comparable effect for people going downstairs. But still one can observe jams forming at the upper end of downstairs streams. These are due to the slight hesitation that occurs when pedestrians synchronize their steps with the geometry of the (down-)stairs ahead. Therefore the bottleneck character of downstairs is less a consequence of the speed on the stairs itself and more of the transition from planar to downward movement, at least as long as the steps are not overly steep.

Evacuations: Empirical Results

Up to now this section has focused on empirical results for pedestrian motion in rather simple scenarios. As we have seen there are many open questions where no consensus has been reached, sometimes even about the qualitative aspects. This becomes even more relevant for full-scale descriptions of evacuations from large buildings or cruise ships. These are typically a combination of many of the simpler elements, so a lack of reliable information is not surprising. In the following we will discuss several complex scenarios in more detail.

Evacuation Experiments In the case of an emergency, the movement of a crowd usually is more straightforward than in the general case. Commuters in a railway station, for example, or visitors of a building might have complex itineraries which are usually represented by origin-destination matrices. In the case of an evacuation, however, the aims and routes are known and usually the same, i. e., the exits and the egress routes. This is the reason why an evacuation process is rather strictly limited in space and time, i. e., its beginning and end are well-defined: the sound of the alarm, initial position of all persons, safe areas (final position of all persons), and the time at which the last person reaches the safe area. When all people have left a building or vessel and reached a safe area (or the lifeboats or liferafts), then the evacuation is finished. Therefore, it is also possible to perform evacuation trials and measure overall evacuation times. Before we go into details, we will clarify three different aspects of data on evacuation processes:

(1) The definition and parts of evacuation time,
(2) The different sources of data, and
(3) The application of these data.

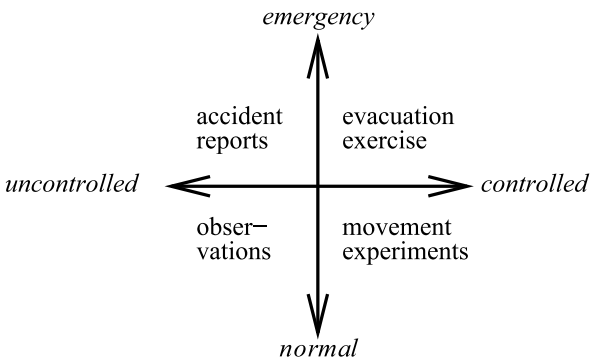

Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 7
Empirical data can be roughly classified according to controlled/uncontrolled and emergency/normal situations

Concerning the evacuation time, five different phases can be distinguished [48,118,153]:

(1) Detection time,
(2) Awareness time,
(3) Decision time,
(4) Reaction time, and
(5) Movement time.

In IMO's regulations [118,119], the first four are grouped together into *response time*. Usually, this time is called *pre-movement time*

One possible scheme for the classification of data on evacuation processes is shown in the following Fig. 7.

Please note that not only data obtained from uncontrolled or emergency situations can be used in the context of evacuation assessment. Knowledge about bottleneck capacities (i. e., flows through doors and on stairs) is especially important when assessing the layout of a building with respect to evacuation. The purpose of empirical data in the context of evacuation processes (and modeling in general) is threefold [43,71]:

(1) Identify parameters (factors that influence the evacuation process, e. g., bottleneck widths and capacities),
(2) Quantify (calibrate) those parameters, e. g., flow through a bottleneck in persons per meter per second, and
(3) Validate simulation results, e. g., compare the overall evacuation time measured in an evacuation with simulation or calculation results.

The validation is usually based on data from the evacuation of complete buildings, aircraft, trains or ships. These are available from two different sources:

(1) Full scale evacuation trials and
(2) Real evacuations.

Evacuation trials are usually observed and videotaped. Reports of real evacuation processes are obtained from eyewitness records and a posteriori incident investigations. Since the setting of a complete evacuation is not experimental, it is hardly possible to measure microscopic features of the crowd motion. Therefore, calibration of parameters is usually not the main purpose in evacuation trials; rather, they are carried out to gain knowledge about the overall evacuation process, the behavior of the persons, to identify the governing influences/parameters and to validate simulation results.

One major concern in evacuation exercises is the well-being of the participants. Due to practical, financial, and ethical constraints, an evacuation trial cannot be, by nature, realistic. Therefore, an evacuation exercise does not convey the increased stress of a real evacuation. To draw conclusions on the evacuation process, the walking speed observed in an exercise should not be assumed to be higher in a real evacuation [145]. Along the same lines of argument, a simplified evacuation analysis based on, e. g., a hydro-dynamic model can predict an evacuation exercise, and the same constraints apply for its results concerning the prediction of evacuation times and the evacuation process. If population parameters (such as gender, age, walking speed, etc.) are explicitly stated in the model, increased stress can be simulated by adapting these parameters.

In summary, evacuation exercises are just too expensive, time consuming, and dangerous to be a standard measure for evacuation analysis. An evacuation exercise organized by the UK Marine Coastguard Agency on the Ro-Ro ferry "Stena Invicta" held in Dover Harbor in 1996 cost more than 10,000 GBP [117]. This is one major argument for the use of evacuation simulations based on hydro-dynamic models and calculations.

Panic, Herding, and Similar Conjectured Collective Phenomena As already mentioned earlier in Subsect. "Collective Effects", the concept of "panic" and its relevance for crowd disasters is rather controversial. It is usually used to describe irrational and unsocial behavior. In the context of evacuations, empirical evidence shows that this type of behavior is rare [3,17,77,178]. On the other hand there are indications that fear might be "contagious" [22]. Related concepts like "herding" and "stampede" imply a certain similarity between the behavior of human crowds and animal behavior. This terminology is quite often used in the public media. *Herding* has been described in animal experiments [166] and is difficult to measure in human crowds. However, it seems to be natural that herding exists in certain situations, e. g., limited

visibility due to failing lights or strong smoke when exits are hard to find.

Panic As stated earlier, "panic" behavior is usually characterized by selfish and anti-social behavior which through contagion affects large groups and even leads to completely irrational actions. Often it is assumed, especially in the media, to occur in situations where people compete for scarce or dwindling resources, which in the case of emergencies are safe space or access to an exit. However, this point of view does not stand close scrutiny and it has turned out that this behavior has played no role at all in many tragic events [73,77]. For these incidents *crowd disaster* is a much more appropriate characterization.

Furthermore, lack of social behavior seems to be more frequent during so called "acquisitive panics" or "crazes" [179] than during "flight panics". That is, social behavior seems to be less stable if there is something to gain than if there is some external danger which threatens all members of a group. Examples of crazes (acquisitive panics) include the Victoria Hall Disaster (1883) [150], the crowning ceremony of Tsar Nicholas II (1896) [168], a governmental Christmas celebration in Aracaju (2001), the distribution of free Saris in Uttar Pradesh (2004), and the opening of an IKEA store in Jeddah (2004). Crowd accidents which occur at rock concerts and religious events as well bear more similarities with crazes than with panics.

However, it is not the case that altruism and cooperation increase with danger. The events during the capsizing of the MV Estonia (see Sect. 16.6 of [100]) show some behavioral threshold: faced with immediate life-threatening danger, most people struggle for their own survival or that of close relatives.

Herding Herding in a broad context means "go with the flow" or "follow the crowd". Like "panic", the term "herding" is often used in the context of stock market crashes, i. e., causing an avalanche effect. Like "panic" the term is usually not well defined and is used in an allegoric way. Therefore, it is advisable to avoid the term in a scientific context (apart from zoology, of course). Furthermore, "herding", "stampede", and "panic" have a strong connotation of "deindividuation". The conjecture of an automatic deindividuation caused by large crowds [101] has been replaced by a social attachment theory ("the typical response to a variety of threats and disasters is not to flee but to seek the proximity of familiar persons and places") [109].

Stampede Stampede is – like herding – a term from zoology where herds of large mammals, such as buffalo, collectively run in one direction and might overrun any obstacles. This is dangerous for human observers if they cannot get out of the way. The term "stampede" is sometimes used for crowd accidents [73], too. It is furthermore assumed to be highly correlated with panic. When arguing along those lines, a stampede might be the result of "crowd panic" or vice versa.

Shock or Density Waves Shock waves are reported for rock concerts [180] and religious events [2,58]. They might result in people standing close to each other falling down. Pressures in dense crowds of up to 4, 450 N/m^2 have been reported.

Although empirical data on crowd disasters exist, e. g., in the form of reports from survivors or even video footage, it is almost impossible to derive quantitative results from them. Models that aim at describing such scenarios make predictions for certain counter-intuitive phenomena that should occur. In the faster-is-slower effect [54] a higher desired velocity leads to a slower movement of a large crowd. In the freezing-by-heating effect [53] increasing the fluctuations can lead to a more ordered state. For a thorough discussion we refer to [54,55] and references therein. However, from a statistical point of view there is insufficient data to decide the relevance of these effects in real emergency situations, not least because it is almost impossible to perform "realistic" experiments.

Sources of Empirical Data on Evacuation Processes
The evacuation of a building can either be an isolated process (due to fire restricted to this building, a bomb threat, etc.) or it can be part of the evacuation of a complete area. We will focus on the single building evacuation, here. For the evacuation of complete areas, e. g., because of flooding or hurricanes, cf. [157] *and references therein*.

For passenger ships, a distinction between High Speed Craft (HSC), Ro-Ro passenger ferries, and other passenger vessels (cruise ships) is made. High Speed Craft do not have cabins and the seating arrangement is similar to aircraft. Therefore, there is a separate guideline for HSC [119]. A performance-based evacuation analysis at an early stage of design is required for HSC and Ro-Pax. There is currently no such requirement for cruise ships. For an overview of IMO's requirements and the historical development up to 2001 cf. [27]. In addition to the five components for the overall evacuation time listed above, there are three more specific to ships:

(6) Preparation time (for the life-saving appliances, i. e., lifeboats, life-rafts, davits, chutes),
(7) Embarkation time, and
(8) Launching time.

Therefore, the evacuation procedure on ships is more complex than for buildings. Additionally, SAR (Search And Rescue) is an integral part of ship evacuation.

For High Speed Craft, the time limit is 17 minutes for evacuation [70], for Ro-Ro passenger ships it is 60 minutes [118], and for all other passenger ships (e. g., cruise ships) it is 60 minutes if the number of main vertical zones is less or equal to five and 80 minutes otherwise [118]. For HSC, no distinction is made between assembly and embarkation phases.

For aircraft, the approach can be compared to that of HSC. First, an evacuation test is mandatory and there is a time limit of 90 seconds that has to be complied to in the test [31].

In many countries there is no strict criterion for the maximum evacuation time of buildings. The requirements are based on minimum exit widths and maximum escape path lengths.

A number of real evacuations has been investigated and reports are publicly available. Among the most recent ones are: Beverly Hills Club [12], MGM Grand Hotel, [12], retail store [4], department store [1], World Trade Center [47] and www.wtc.nist.gov, high-rise buildings [144, 173], theater [191] for buildings, High Speed Craft "Sleipner" [138] for HSC, an overview up to 1998 [143], exit width variation [121], double deck aircraft [74], another overview for aircraft [120], and for trains [43,169].

Modeling

A comprehensive theory of pedestrian dynamics has to take into account three different levels of behavior (Fig. 8). At the *strategic level*, pedestrians decide which activities they like to perform and the order of these activities. With the choices made at the strategic level, the *tactical level* concerns the short-term decisions made by the pedestrians, e. g., choosing the precise route taking into account obstacles, density of pedestrians etc. Finally, the *operational level* describes the actual walking behavior of pedestrians, e. g., their immediate decisions necessary to avoid collisions etc.

Processes at the strategic and tactical level are usually considered to be exogenous to pedestrian simulation. Here information from other disciplines (sociology, psychology etc.) is required. In the following we will mostly be concerned with the operational level, although some of the models that we are going to describe allow us to take into account certain elements of behavior at the tactical level as well.

Modeling on the operational level is usually based on variations of models from physics. Indeed the motion of pedestrian crowds shares certain similarities with fluids and the flow of granular materials. The goal is to find models which are as simple as possible, but at the same time can reproduce "realistic" behavior in the sense that the empirical observations are reproduced. Therefore, based on the experience from physics, pedestrians are often modeled as simple "particles" that interact with each other.

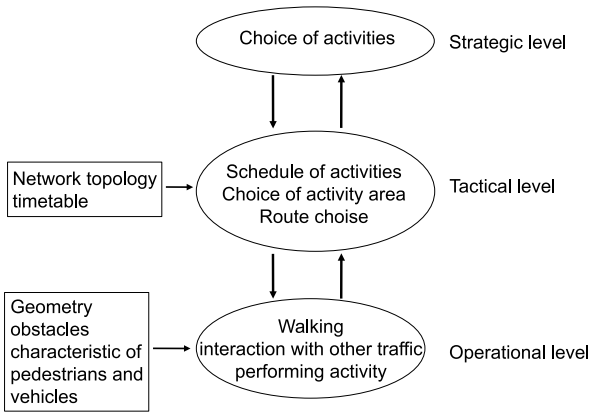

Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 8
The different levels of modeling pedestrian behavior (after [19, 65])

There are several characteristics which can be used to classify the modeling approaches:

Microscopic vs. macroscopic In microscopic models each individual is represented separately. Such an approach allows us to introduce different types of pedestrians with individual properties as well as issues such as route choice. In contrast, in macroscopic models, individuals cannot be distinguished. Instead the state of the system is described by densities, usually a mass density derived from the positions of the persons and a corresponding locally averaged velocity.

Discrete vs. continuous Each of the three basic variables for a description of a system of pedestrians, namely space, time and state variable (e. g., velocities), can be either discrete (i. e., an integer number) or continuous (i. e., a real number). Here all combinations are possible. In a cellular automaton approach all variables are by definition discrete, whereas in hydrodynamic models all are continuous. These are the most common choices, but other combinations are used as well. Sometimes for a cellular automata approach a continuous time variable is also allowed. In computer simulation this is realized through a *random-sequential update* where at each step the particle or site to be up-

dated (moved) is chosen randomly (from *all* particles or sites, respectively). A discrete time is usually realized through a *parallel* or *synchronous update* where all particles or sites are moved at the same time. This introduces a timescale. In so-called coupled map lattices time is discrete, whereas space and state variables are continuous.

Deterministic vs. stochastic The dynamics of pedestrians can either be deterministic or stochastic. In the first case the behavior at a certain time is completely determined by the present state. In stochastic models, behavior is controlled by certain probabilities such that the agents can react differently in the same situation. This is one of the lessons learnt from the theory of complex systems where it has been shown for many examples that through introduction of stochasticity into rather simple systems very complex behavior can be generated. On the other hand, the stochasticity in the models reflects our lack of knowledge of the underlying physical processes that, e. g., determine the decision-making of the pedestrians. Through stochastic behavioral rules it often becomes possible to generate a rather realistic representation of complex systems such as pedestrian crowds.

This "intrinsic" stochasticity should be distinguished from "noise". Sometimes external noise terms are added to the *macroscopic* observables, such as position or velocity. Often the main effect of these terms is to avoid certain special configurations which are considered to be unrealistic, like completely blocked states. Otherwise the behavior is very similar to the deterministic case. For true stochasticity, on the other hand, the deterministic limit usually has very different properties from the generic case.

Rule-based vs. force-based Interactions between the agents can be implemented in at least two different ways: In a rule-based approach agents make "decisions" based on their current situation, the nature of their neighborhood as well as their goals, etc. It focuses on the *intrinsic properties* of the agents and thus the rules are often justified from psychology. In force-based models, agents "feel" a force exerted by others and the infrastructure. They therefore emphasize *extrinsic properties* and their relevance for the motion of the agents. This is a physical approach based on the observation that the presence of others leads to deviations from a straight motion. In analogy to Newtonian mechanics a force is made responsible for these accelerations.

Cellular automata are typically rule-based models, whereas, e. g., the social-force model belongs to the force-based approaches. However, sometimes a clear distinction cannot be made; many models combine aspects of both approaches.

High vs. low fidelity *Fidelity* here refers to the apparent realism of the modeling approach. High fidelity models try to capture the complexity of decision making, actions, etc. that constitute pedestrian motion in a realistic way. In contrast, in the simplest models pedestrians are represented by particles without any intelligence. Usually the behavior of these particles is determined by "forces". This approach can be extended, e. g., by allowing different "internal" states of the particles so that they react differently to the same force depending on the internal state. This can be interpreted as some kind of "intelligence" and leads to more complex approaches, like multi-agent models. Roughly speaking, the number of parameters in a model is a good measure for fidelity in the sense introduced here, but note that higher fidelity does not necessarily mean that empirical observations are reproduced better!

It should be mentioned that a clear classification according to the characteristics outlined here is not always possible. In the following we will describe some model classes in more detail.

Fluid-dynamic and Gas kinetic Models

Pedestrian dynamics has some obvious similarities with fluids. For example, the motion around obstacles appears to follow "streamlines". Motion at intermediate densities is restricted (short-ranged correlations). Therefore it is not surprising that, very much like for vehicular dynamics, the earliest models of pedestrian dynamics took inspiration from hydrodynamics or gas-kinetic theory [50,61,68,69]. Typically these macroscopic models are deterministic, force-based and of low fidelity.

Henderson [60,61] has tried to establish an analogy of large crowds with a classical gas. From measurements of motion in different crowds in a low density ("gaseous") phase he found good agreement of the velocity distribution functions with Maxwell–Boltzmann distribution [60].

Motivated by this observation, he later developed a fluid-dynamic theory of pedestrian flow [61]. Describing the interactions between the pedestrians as a collision process where the particles exchange momenta and energy, a homogeneous crowd can be described by the well-known kinetic theory of gases. However, the interpretation of the quantities is not entirely clear, e. g., what the analogues of pressure and temperature are in the context of pedestrian

motion. Temperature could be identified with the velocity variance, which is related to the distribution of desired velocities, whereas the pressure expresses the desire to move against a force in a certain direction.

The applicability of classical hydrodynamic models is based on several conservation laws. The conservation of mass, corresponding to conservation of the total number of pedestrians, is expressed through a continuity equation of the form

$$\frac{\partial \rho(\mathbf{r}, t)}{\partial t} + \nabla \cdot \mathbf{J}(\mathbf{r}, t) = 0 \,, \tag{6}$$

which connects the local density $\rho(\mathbf{r}, t)$ with the current $\mathbf{J}(\mathbf{r}, t)$. This equation can be generalized to include source and sink terms. However, the assumption of conservation of energy and momentum is not true for interactions between pedestrians which in general do not even satisfy Newton's Third Law ("actio = reaction"). In [50] several other differences to normal fluids were pointed out, e. g., the anisotropy of interactions or the fact that pedestrians usually have an individual preferred direction of motion.

In [50] a better founded fluid-dynamical description was derived on the basis of a gas kinetic model which describes the system in terms of a density function $f(\mathbf{r}, \mathbf{v}, t)$. The dynamics of this function are determined by Boltzmann's transport equation that describes its change for a given state as difference of inflow and outflow due to binary collisions.

An important new aspect in pedestrian dynamics is the existence of desired directions of motion which allows us to distinguish different groups μ of particles. The corresponding densities f_μ change in time due to four different effects:

1. A relaxation term with characteristic time τ describes tendency of pedestrians to approach their intended velocities.
2. The interaction between pedestrians is modeled by a Stosszahlansatz as in the Boltzmann equation. Here, pair interactions between types μ and ν occur with a total rate that is proportional to the densities f_μ and f_ν.
3. Pedestrians are allowed to change from type μ to ν which, e. g., accounts for turning left or right at a crossing.
4. Additional gain and loss terms allow us to model entrances and exits where pedestrian can enter or leave the system.

The resulting fluid-dynamic equations derived from this gas kinetic approach are similar to that of ordinary fluids. However, due to the different types of pedestrians, corresponding to individuals who have approximately the same desired velocity, one actually obtains a set of coupled equations describing several interacting fluids. These equations contain additional characteristic terms describing the approach to the intended velocity and the change of fluid-type due to interactions in avoidance maneuvers.

Equilibrium is approached through the tendency to walk with the intended velocity, not through interactions as in ordinary fluids. Momentum and energy are not conserved in pedestrian motion, but the relaxation towards the intended velocity describes a tendency to restore these quantities.

Unsurprisingly for a macroscopic approach, the gas-kinetic models have problems at low densities. For a discussion, see e. g. [50].

Hand Calculation method For practical applications effective engineering tools have been developed from the hydrodynamical description. In engineering these are often called *hand calculation methods*. One could also classify some of them as queuing models since the central idea is to describe pedestrian dynamics as flow on a network with links of limited capacities. These methods allow us to calculate evacuation times in a relatively simple way that does not require any simulations. Parameters entering in the calculations can be adapted to the situation that is studied. Often they are based on empirical results, e. g., evacuation trials. Details about this kind of model can be found in Subsect. "Calculation of Evacuation Times".

Social-Force Models

The social-force model [52] is a deterministic continuum model in which interactions between pedestrians are implemented by using the concept of a *social force* or *social field* [103]. It is based on the idea that changes in behavior can be understood in terms of fields or forces. Applied to pedestrian dynamics, the social force $\mathbf{F}_j^{(\mathrm{soc})}$ represents the influence of the environment (other pedestrians, infrastructure) and changes the velocity $\mathbf{v}_j$ of pedestrian j. Thus it is responsible for acceleration which justifies the interpretation as a force. The basic equation of motion for a pedestrian of mass m_j is then of the general form

$$\frac{\mathrm{d}\mathbf{v}_j}{\mathrm{d}t} = \mathbf{f}_j^{(\mathrm{pers})} + \mathbf{f}_j^{(\mathrm{soc})} + \mathbf{f}_j^{(\mathrm{phys})} \tag{7}$$

where $\mathbf{f}_j^{(\mathrm{soc})} = \frac{1}{m_j}\mathbf{F}_j^{(\mathrm{soc})} = \sum_{l \neq j} \mathbf{f}_{jl}^{(\mathrm{soc})}$ is the total (specific) force due to other pedestrians. $\mathbf{f}_j^{(\mathrm{pers})}$ denotes a "personal" force which makes a pedestrian attempt to move with his or her own preferred velocity $\mathbf{v}_j^{(0)}$ and thus acts as a driving

term. It is given by

$$\mathbf{f}_j^{(\mathrm{pers})} = \frac{\mathbf{v}_j^{(0)} - \mathbf{v}_j}{\tau_j} \tag{8}$$

where τ_j is reaction or acceleration time. In high density situations, physical forces $\mathbf{f}_{jl}^{(\mathrm{phys})}$ also become important, e. g., friction and compression when pedestrians make contact.

The most important contribution to the social force $\mathbf{f}_j^{(\mathrm{soc})}$ comes from the territorial effect, i. e., the private sphere. Pedestrians feel uncomfortable if they get too close to others, which effectively leads to a repulsive force between them. Similar effects are observed for the environment, e. g., people prefer not to walk too close to walls.

Since social forces are difficult to determine empirically, some assumptions must be made. Usually an exponential form is assumed. Describing the pedestrians as disks of radius R_j and position (of the center of mass) $\mathbf{r}_j$, the typical structure of the force between the pedestrians is described by [54]

$$\mathbf{f}_{jl}^{(\mathrm{soc})} = A_j \exp\left[\frac{R_{jl} - \Delta r_{jl}}{\xi_j}\right] \mathbf{n}_{jl} \tag{9}$$

with $R_{jl} = R_j + R_l$, the sum of the disk radii, $\Delta r_{jl} = |\mathbf{r}_j - \mathbf{r}_l|$, the distance between the centers of mass, $\mathbf{n}_{jl} = \mathbf{r}_j - \mathbf{r}_l / \Delta r_{jl}$, the normalized vector pointing from pedestrian l to j. A_j can be interpreted as strength, ξ_j as the range of the interactions.

The appeal of the social-force model is given mainly by analogy to Newtonian dynamics. For the solution of the equations of motion of Newtonian many-particle systems, well-founded molecular dynamics techniques exists. However, in most studies so far the distinctions between pedestrian and Newtonian dynamics are not discussed in detail. A straightforward implementation of the equations of motion neglecting these distinctions can lead to unrealistic movement of single pedestrians. For example, negative velocities in the main moving direction cannot be excluded in general even if asymmetric interactions (violating Newton's Third Law) between the pedestrians are chosen. Another effect is the occurrence of velocities higher then the preferred velocity $v_j{}^{(0)}$ due to the forces on pedestrians in the moving direction. To prevent this effect, additional restrictions for the degrees of freedom must be introduced, see for example [52], or the superposition of forces has to be discarded [175]. A general discussion of the limited analogy between Newtonian dynamics and the social-force model as well as the consequences for model implementations is still missing.

Apart from the ad hoc introduction of interactions, the structure of the social-force model can also be derived from an extremal principle [62,63]. It follows under the assumption that pedestrian behavior is determined by the desire to minimize a certain cost function which takes into account not only kinematic aspects and walking comfort, but also deviations from a planned route.

Cellular Automata

Cellular automata (CA) are rule-based dynamical models that are discrete in space, time and state variable which in the case of traffic usually corresponds to velocity. Discreteness in time means that the positions of the agents are updated in well defined steps. In computer simulations this is realized through a *parallel* or *synchronous* update where all pedestrians move at the same time. The time step corresponds to a natural timescale Δt which could be identified, e. g., with some reaction time. This can be used for the calibration of the model which is essential for making quantitative predictions. A natural space discretization can be derived from the maximal densities observed in dense crowds which gives the minimal space requirement of one person. Usually each cell in the CA can be occupied by only one particle (exclusion principle) so that this space requirement can be identified with the cell size. In this way, a maximal density of 6.25 P/m^2 [192] leads to a cell size of $40 \times 40\,\mathrm{cm}^2$. Sometimes finer discretizations are more appropriate (see Subsect. "Theoretical Results). In this case pedestrians correspond to extended particles that occupy more than one cell (e. g., four cells). The exclusion principle and the modeling of humans as non-compressible particles mimics short-range repulsive interactions, i. e., the "private-sphere".

The dynamics are usually defined by rules which specify transition probabilities for motion to one of the neighboring cells (Fig. 9). The models differ in the specification of these probabilities as well in that of the "neighborhood". For deterministic models, all but one are of probability zero.

The first cellular automata (CA) models [7,41,89,129] for pedestrian dynamics can be considered two-dimensional variants of the asymmetric simple exclusion process (ASEP) (for reviews, see [9,23,172]) or models for city or highway traffic [6,16,133] based on it. Most of these models represent pedestrians by particles without any internal degrees of freedom. They can move to one of the neighboring cells based on certain transition probabilities which are determined by three factors:

(1) The desired direction of motion, e. g., to find the shortest connection,

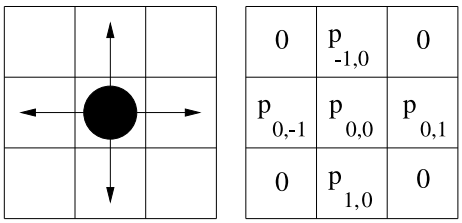

Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 9
A particle, its possible directions of motion and the corresponding transition probabilities p_{ij} for the case of a von Neumann neighborhood

(2) Interactions with other pedestrians, and
(3) Interactions with the infrastructure (walls, doors, etc.).

Fukui–Ishibashi Model One of the first CA models for pedestrian dynamics was proposed by Fukui and Ishibashi [40,41] and is based on a two-dimensional variant of the ASEP. They studied bidirectional motion in a long corridor where particles moving in opposite directions were updated alternatingly. Particles move deterministically in their desired direction; only if the desired cell is occupied by an oppositely moving particle do they make a random sidestep.

Various extensions and variations of the model have been proposed, e. g., an asymmetric variant [129] where walkers prefer lane changes to the right, different update types [193], simultaneous (exchange) motion of pedestrians standing "face-to-face" [72], or the possibility of backstepping [107]. The influence of the shape of the particles has been investigated in [131]. Also other geometries [128, 181] and extensions to full 2-dimensional motion have been studied in various modifications [106,107,127]

Blue–Adler Model The model of Blue and Adler [7,8] is based on a variant of the Nagel–Schreckenberg model [133] of highway traffic. Pedestrian motion is considered in analogy to a multi-lane highway. The structure of the rules is similar to the basic two-lane rules suggested in [159]. The update is performed in four steps which are applied to all pedestrians in parallel. In the first step each pedestrian chooses a preferred lane. In the second step the lane changes are performed. In the third step the velocities are determined based on the available gap in the new lanes. Finally, in the fourth step the pedestrians move forward according to the velocities determined in the previous step.

In counterflow situations head-on-conflicts occur. These are resolved stochastically and with some probability opposing pedestrians are allowed to exchange positions within one time step. Note that the motion of a single pedestrian (not interacting with others) is deterministic otherwise.

Unlike the Fukui–Ishibashi model, motion is not restricted to nearest-neighbor sites. Instead, pedestrians can have different velocities $v_{\max}$ which correspond to the maximal number of cells they are allowed to move forward. In contrast to vehicular traffic, acceleration to $v_{\max}$ can be assumed to be instantaneous in pedestrian motion.

In order to study the effects of inhomogeneities, the pedestrians are assigned different maximal velocities $v_{\max}$. Fast walkers have $v_{\max} = 4$, standard walkers $v_{\max} = 3$ and slow walkers $v_{\max} = 2$. The cell size is assumed to be $50\,\text{cm} \times 50\,\text{cm}$. The best agreement with empirical observations has been achieved with 5% slow and 5% fast walkers [8]. Furthermore the fundamental diagram in more complex situations, such as bi- or four-directional flows, has been investigated.

Gipps–Marksjös Model A more sophisticated discrete model was suggested by Gipps and Marksjös [45] in 1985. One motivation for developing a discrete model was the limited computer power at that time. Therefore a discrete model, which reproduces the properties of pedestrian motion realistically, was in many respects a real improvement over the existing continuum approaches.

Interactions between pedestrians are assumed to be repulsive, anticipating the idea of social forces (see Subsect. "Social-Force Models"). The pedestrians move on a grid of rectangular cells of size $0.5 \times 0.5\,\text{m}$. To each cell a score is assigned based on its proximity to other pedestrians. This score represents the repulsive interactions and actual motion is then determined by the competition between these repulsions and the gain of approaching the destination. Applying this procedure to all pedestrians, a potential value is assigned to each cell which is the sum of the individual contributions. The pedestrian then selects the cell of its nine neighbors (Moore neighborhood) which leads to the maximum benefit. This benefit is defined as the difference between the gain of moving closer to the destination and the cost of moving closer to other pedestrians as represented by the potential. This requires a suitable chosen gain function P.

The updating is done sequentially to avoid conflicts of several pedestrians trying to move to the same position. In order to model different velocities, faster pedestrians are updated more frequently. Note that the model dynamics are deterministic.

Floor Field CA Floor field CA [13,14,83,167] can also be considered as an extension of the ASEP. However, the transition probabilities to neighboring cells are no longer

fixed but vary dynamically. This is motivated by the process of chemotaxis (see [5] for a review) used by some insects (e. g., ants) for communication. They create a chemical trace to guide other individuals to food sources. In this way a complex trail system is formed that has many similarities with human transport networks.

In the approach of [13] the pedestrians also create a trace. In contrast to chemotaxis, however, this trace is only virtual, although one could assume that it corresponds to some abstract representation of the path in the mind of the pedestrians. Although this is mainly a technical trick which reduces interactions to local ones that allow efficient simulations in arbitrary geometries, one could also think of the trail as representation of the paths in the mind of a pedestrian. The locality becomes important in complex geometries as no algorithm is required to check whether the interaction between particles is screened by walls, etc. The number of interaction terms always grows linearly with the number of particles.

The translation into local interactions is achieved by the introduction of so-called *floor fields*. The transition probabilities for all pedestrians depend on the strength of the floor fields in their neighborhood in such a way that transitions in the direction of larger fields are preferred. The *dynamic floor field* D_{ij} corresponds to a virtual trace which is created by the motion of the pedestrians and in turn influences the motion of other individuals. Furthermore it has its own dynamics, namely through diffusion and decay, which leads to a dilution and finally the vanishing of the trace after some time. The *static floor field* S_{ij} does not change with time since it only takes into account the effects of the surroundings. Therefore it exists even without any pedestrians present. It allows us to model, e. g., preferred areas, walls and other obstacles. Figure 10 shows the static floor field used for the simulation of evacuations from a room with a single door. Its strength decreases with increasing distance from the door. Since the pedestrians prefer motion into the direction of larger fields, this is already sufficient to find the door.

Coupling constants control the relative influence of both fields. For a strong coupling to the static field pedestrians will choose the shortest path to the exit. This corresponds to a 'normal' situation. A strong coupling to the dynamic field implies a strong herding behavior where pedestrians try to follow the lead of others. This often happens in emergency situations.

The model uses a fully parallel update. Therefore conflicts can occur where different particles choose the same destination cell. This is relevant for high density situations and happens in all models with parallel update if motion in different directions is allowed. Conflicts have been considered a technical problem for a long time and usually the dynamics have been modified in order to avoid them. The simplest method is to update pedestrians sequentially instead of using fully parallel dynamics. However, this leads to other problems, e. g., the identification of the relevant timescale. Therefore it has been suggested in [84,85] to take these conflicts seriously as an important part of the dynamics.

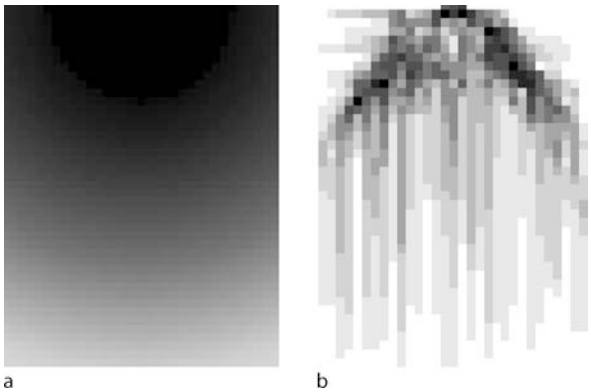

Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 10
***Left*: Static floor field for the simulation of an evacuation from a large room with a single door. The door is located in the middle of the upper boundary and the field strength increases with increasing intensity. *Right*: Snapshot of the dynamical floor field created by people leaving the room**

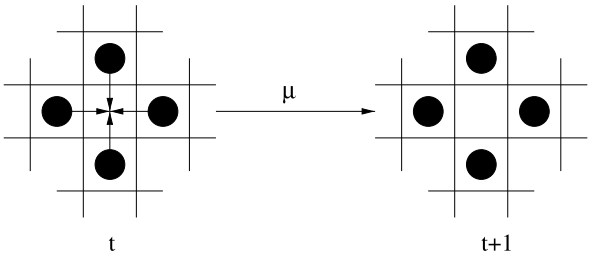

Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 11
Refused movement due to the friction parameter μ (for $m = 4$)

For the floor field model it has been shown in [85] that the behavior becomes more realistic if not all conflicts are resolved by allowing one pedestrian to move while the others stay at their positions. Instead with probability $\mu \in [0, 1]$, which is called the friction parameter, the movement of *all* involved pedestrians is denied [85] (see Fig. 11).

This allows one to describe clogging effects between the pedestrians in a much more detailed way [85]. μ works as some kind of local pressure between the pedestrians.

If μ is high, the pedestrians handicap each other trying to reach their desired target sites. This local effect can have enormous influence on macroscopic quantities like flow and evacuation time [85]. Note that the kind of friction introduced here only influences interacting particles, not the average velocity of a freely moving pedestrian.

Surprisingly, the qualitative behaviors of the floor field model and the social-force models are very similar despite the fact that the interactions are very different. In the floor field model interactions are attractive, whereas in the social-force model they are repulsive. However, in the latter interactions are between particle densities. In contrast, in the floor field model the particle density interacts with the velocity density.

Other Approaches

Lattice-gas models In 1986, Frisch, Hasslacher, and Pomeau [34] showed that one does not have to take into account detailed molecular motion within fluids in order to obtain a realistic picture of (2d) fluid dynamics. They proposed a lattice gas model [164,165] on a triangular lattice with hexagonal symmetry, which is similar in spirit to CA models, but the exclusion principle is relaxed: particles with different velocities are allowed to occupy the same site. Note that the allowed velocities differ only in the direction, not absolute value. The dynamics are based on a succession of collision and propagation that can be chosen in such a way that the coarse-grained averages of this microscopic dynamic is asymptotically equivalent to the Navier–Stokes equations of incompressible fluids.

In [108] a kind of mesoscopic approach inspired by these lattice gas models has been suggested as a model for pedestrian dynamics. In analogy with the description of transport phenomena in fluids (e. g., the Boltzmann equation) the dynamics are based on a succession of collision and propagation.

Pedestrians are modeled as particles, moving on a triangular lattice, which have a preferred direction of motion $\mathbf{c}_F$. However, the particles do not strictly follow this direction but also have a tendency to move with the flow. Furthermore, at high densities the crowd motion is influenced by a kind of friction which slows down the pedestrians. This is achieved by reducing the number of individuals allowed to move to neighboring sites.

As in a lattice gas model [165], the dynamics now consists of two steps. In the *propagation step* each pedestrian moves to the neighbor site in the direction of its velocity vector. In the *collision step* the particles interact and new velocities (directions) are determined. In contrast to physical systems, momentum, etc., does not need to be conserved during the collision step. These considerations lead to a collision step that takes into account the favorite direction $\mathbf{c}_F$, the local density (the number of pedestrians at the collision site), and a quantity called mobility at all neighbor sites which is a normalized measure of the local flow after the collision.

Optimal-Velocity Model The optimal velocity (OV) model originally introduced for the description of highway traffic can be generalized to higher dimensions [134] which allows its application to pedestrian dynamics.

In the two-dimensional extension of the OV model the equation of motion for particle i is given by

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{x}_i(t) = a\left\{\mathbf{V}_0 + \sum_j \mathbf{V}(\mathbf{x}_j(t) - \mathbf{x}_i(t)) - \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}_i(t)\right\}, \quad (10)$$

where $\mathbf{x}_i = (x_i, y_i)$ is the position of particle i. It can be considered as a special case of the general social-force model (7) without physical forces. The optimal-velocity function

$$\mathbf{V}(\mathbf{x}_j - \mathbf{x}_i) = f(r_{ij})(1 + \cos\varphi)\mathbf{n}_{ij}\,, \quad (11)$$

$$f(r_{ij}) = \alpha\{\tanh\beta(r_{ij} - b) + c\}, \quad (12)$$

where $r_{ij} = |\mathbf{x}_j - \mathbf{x}_i|$, $\cos\varphi = (x_j - x_i)/r_{ij}$ and $\mathbf{n}_{ij} = (\mathbf{x}_j - \mathbf{x}_i)/r_{ij}$ is determined by interactions with other pedestrians. $\mathbf{V}_0$ is a constant vector that represents a 'desired velocity' at which an isolated pedestrian would move. The strength of the interaction depends on the distance r_{ij} between the ith and jth particles, and on the angle φ between the directions of $\mathbf{x}_j - \mathbf{x}_i$ and the current velocity $\mathrm{d}\mathbf{x}_i/\mathrm{d}t$. Due to the term $(1 + \cos\varphi)$, a particle reacts more sensitively to particles in front than to those behind.

Now two cases can be distinguished: repulsive and attractive interactions. The former is relevant for pedestrian dynamics whereas the latter is more suitable for biological motion. Therefore, for pedestrian motion one chooses $c = 1$ which implies $f < 0$.

A detailed analysis [134] shows that the model exhibits a rich phase diagram including the formation of various patterns.

Other Models We briefly mention a few other model approaches that have been suggested. In [10] a discretized version of the social-force model has been introduced and shown to reproduce qualitatively the observed collective phenomena.

In [141] a magnetic force model has been proposed where pedestrians and their goals are treated as magnetic poles of opposite sign.

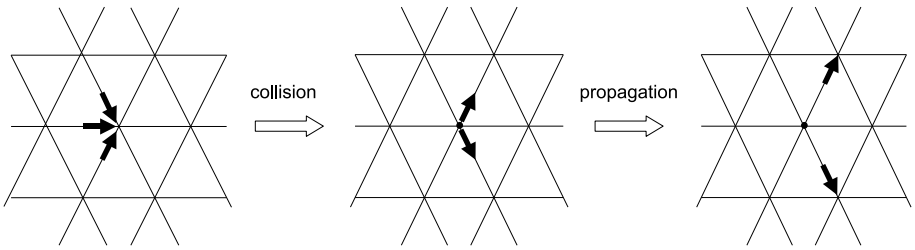

Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 12
The dynamics of lattice gas models proceed in two steps. Pedestrians coming from neighboring sites interact in the collision step where velocities are redistributed. In the propagation step the pedestrians move to neighbor sites in directions determined by the collision step

Another class of models is based on ideas from queuing theory. In principle, some hand calculation methods can be considered as macroscopic queuing models. Typically, rooms are represented as nodes in the queuing network and links correspond to doors. In microscopic approaches, in the movement process each agent chooses a new node, e. g., according to some probability [105].

Theoretical Results

As emphasized in Subsect. "Collective Effects", the collective effects observed in the motion of pedestrian crowds are a direct consequence of microscopic dynamics. These effects are reproduced quite well by some models, e. g., the social-force and floor-field model, at least on a qualitative level. As mentioned before, the qualitative behavior of the two models is rather similar despite the very different implementation of the interactions. This indicates a certain robustness of the collective phenomena observed.

As an example we discuss the formation of lanes in counterflow formation. Empirically one observes a strong tendency to follow immediately in the "wake" of another person heading in the same direction. Such lane formation was reproduced in the social-force model [52,53] as well as in the floor-field model [13,76] (see Fig. 13). While the formation of lanes in general is essential to avoid deadlocks and thus keep the chance of reproducing realistic fluxes, the number of direction changes per meter cross section is a parameter which in reality crucially depends on the situation [76]: the longer a counterflow situation is assumed to persist, the fewer lanes per meter cross section can be found. The correct reproduction of counterflow is an issue for an accommodating animation, but more or less unimportant for macroscopic observables. This is probably the

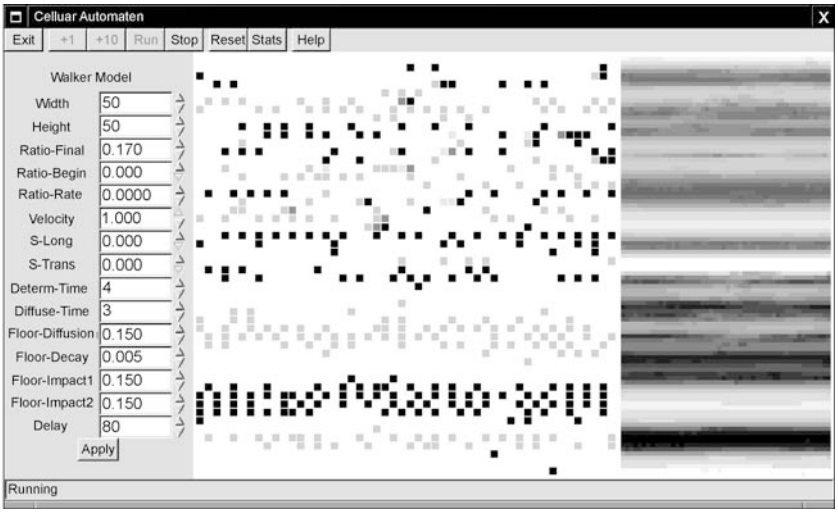

Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 13
Lane formation in the floor-field model. The central window is the corridor and the light and dark squares are right- and left-moving pedestrians, respectively. In the bottom part well-separated lanes can be observed whereas in the top part the motion is still disordered. The right part of the figure shows the floor fields for the right-movers (*upper half*) and left-movers (*lower half*)

main reason why there seems to have been little effort put into the attempt to reproduce different "kinds" of lane formation in a controlled, situation-dependent manner.

On the quantitative side, the fundamental diagram is the first and most serious test for any model. Since most quantitative results rely on the fundamental diagram, it can be considered the most important characteristic of pedestrian dynamics. It is not only relevant for movement in a corridor or through a bottleneck, but also as an important determinant of evacuation times. However, as emphasized earlier, there is currently no consensus on the empirical form of the fundamental diagram. Therefore, a calibration of the model parameters is currently difficult.

Most cellular automata models are based on the asymmetric simple exclusion process. This strictly one-dimensional stochastic process has a fundamental diagram which is symmetric around density $\rho = 1/2$. Lane changes in two-dimensional extensions lead to only a small shift towards smaller densities. Despite the discrepancies in the empirical results, an almost symmetric fundamental diagram can be excluded.

Based on the experience with modeling of highway traffic [16,133], models with higher speeds have been introduced which naturally lead to an asymmetric fundamental diagram. Typically this is implemented by allowing the agents to move more than one cell per update step [82,86,87,92,195,196]. These model variants have been shown to be flexible enough to reproduce, e. g., Weidmann's fundamental diagram for the flow in a corridor [192] with high precision. Usually in the simulations a homogeneous population is assumed. However, in reality, different pedestrians have different properties such as walking speed, motivation, etc. This is easily taken into account in every microscopic model. There are many parameters that could potentially have an influence on the fundamental diagram. However, the current empirical situation does not allow to decide this question.

Another problem occurring in CA models has its origin in the discreteness of space. Through the choice of the lattice discretization, space is no longer isotropic. Motion in directions not parallel to the main axis of the lattice are difficult to realize and can only be approximated by a sequence of steps parallel to the main directions.

Higher velocities also require the extension of the neighborhood of a particle which is no longer identical to the cells adjacent to the current position. A natural definition of "neighborhood" corresponds to those cells that could be reached within one time step. In this way the introduction of higher velocities also reduces the problem of space isotropy as the neighborhoods become more isotropic for larger velocities.

Other solutions to this problem have been proposed. One way is to count the number of diagonal steps and let the agent suspend from moving following certain rules which depend on the number of diagonal steps [171]. A similar idea is to sum up the real distance that an agent has moved during one round: a diagonal step counts $\sqrt{2}$ and a horizontal or vertical step counts 1. An agent has to finish its round as soon as this sum is bigger than its speed [87]. A third possibility – which works for arbitrary speeds – is to assign selection probabilities to each of the four lattice positions adjacent to the exact final position [195,196]. Naturally these probabilities are inversely proportional to the square area between the exact final position and the lattice point, as in this case the probabilities are normalized by construction if one has a square lattice with points on all integer number combinations. However, one also could think of other methods to calculate the probability.

For the social-force model, the specification of the repulsive interaction (with and without hard core, exponential or reciprocal with distance) as well as the parameter sets for the forces changes in different publications [52,53,54,114]. In [55] the authors state that "most observed self-organization phenomena are quite insensitive to the specification of interaction forces". However, at least for the fundamental diagram, a relation connected with all phenomena in pedestrian dynamics, this statement is questionable. As remarked in [56] the reproduction of the fundamental diagram "requires a less simple specification of the repulsive interaction forces". Indeed in [175] it was shown that the choice of hard-core forces or repulsive soft interactions as well as the particular parameter set can strongly influence the resulting fundamental diagram regarding qualitative as well as quantitative effects.

Also a more realistic behavior at higher densities requires a modification of the basic model. Here the use of density-dependent desired velocities leads to a reduction of the otherwise unrealistically large number of collisions [10].

The particular specification of forces and the previously mentioned problem with Newton's Third law can lead in principle to some unwanted effects, such as momentary velocities larger than the preferred velocity [52] or the penetration of pedestrians into each other or into walls [98]. It is possible that these effects can be suppressed for certain parameter sets by contact or friction forces, but the general appearance is not excluded. Only in the first publication [52] are restrictions for the velocity explicitly formulated to prevent velocities larger than the intended speed; other authors tried to improve the model by introducing more parameters [98]. But additional parame-

ter and artificial restrictions of variables diminish the simplicity and thus the attractiveness of the model. A general discussion of how to deal with these problems of the social-force model and a verification that the observed phenomena are not limited to a certain specification of the interaction and a special parameter set is up to now still missing.

While realistic reproduction within the empirical range of these macroscopic observables, especially the fundamental diagram, is absolutely essential to guarantee safety standards in evacuation simulations, and while a user should always be distrustful of models where no fundamental diagram has ever been published, it is by no means sufficient to exclusively check for the realism of macroscopic observables. On the microscopic level there are a large number of phenomena which need to be reproduced realistically, be it just to make a simulation animation look realistic or because microscopic effects can often easily influence macroscopic observables.

If one compares simulations of bottleneck flows with real events, one observes that in simulations the form of the queue in front of bottlenecks is often a half-circle, while in reality it is drop- or wedge-shaped. In most cases this discrepancy probably does not have an influence on the simulated evacuation time, but it is interesting to note where it originates from. Most simulation models implicitly or explicitly use some kind of utility maximization to steer the pedestrians – with the utility being foremost inversely proportional to the distance from the nearest exit. This obviously leads to half-circle-shaped queues in front of bottlenecks. So wherever one observes queues different than half-circles, people have exchanged their normal "utility function based on the distance" with something else. One such alternative utility function could be that people are just curious about what is inside or behind the bottleneck, so they seek a position where they can look into it. A more probable explanation would be that in any case it is the time distance not the spatial distance which is is sought to be minimized. As anyone knows what the inescapable loss in time a bottleneck means for the whole waiting group, the precise waiting spot is not that important. However, in societies with a strong feeling for egality, people would strongly wish to equally distribute the waiting time and keep a first-in-first-out principle, which can best be accomplished and controlled when the queue is more or less one-dimensional, respectively just as wide as the bottleneck itself.

Finally it should be mentioned that theoretical investigations based on simulations of models for pedestrian dynamics have led to the prediction of some surprising and counter-intuitive collective phenomena, such as the reduction of evacuation times through additional columns near exits (see Subsect. "Bottleneck Flow") or the faster-is-slower [54] and freezing-by-heating effect [53]. However, so far the empirical evidence for the relevance or even the occurrence of these effects in real situations is rather scarce.

Applications

In the following section we discuss more practical aspects of based on the modeling concepts presented in Sect. "Modeling". Tools of different sophistication have been developed that are nowadays routinely used in safety analysis. The latter becomes more and more relevant since many public facilities must fulfill certain legal standards. As an example we mention aircrafts which must be evacuated within 90 seconds. The simulations etc. are already used in the planning stages because changes of the design at a later stage are difficult and expensive.

For this kind of safety analysis tools of different sophistication have been developed. Some of them mainly are able to predict just evacuation times whereas others are based on microscopic simulations which allow also to study various external influences (fire, smoke, . . .) in much detail.

Calculation of Evacuation Times

The basic idea of hand calculation methods has already briefly been described at the end of Subsect. "Fluid-Dynamic and Gas Kinetic Models". Here we want to discuss its practical aspects in more detail.

The approach has been developed since the middle of the 1950s [185]. The basic idea of these methods is the assumption that people can be modeled to behave like fluids. Knowledge of the flow (see Eq. 1) and the technical data of the facility are then sufficient to evaluate evacuation times, etc.

Hand calculation method can be divided into two major approaches: methods with "dynamic" flow [35,42,78,79,80,136,151,152,163,192] and methods with "fixed" flow [110,123,124,125,126,137,145,173,185]. As methods with "dynamic" flow we cite methods where the pedestrian flow is dependent on the density of the pedestrian stream (see Subsect. "Observables") in the selected facility, thus the flow can be obtained from fundamental diagrams (see Subsect. "Fundamental Diagram") or it is explicitly prescribed in the chosen method. This flow can change during movement through the building, e. g., by using stairs, thus the pedestrian stream has a "dynamic" flow. Methods with "fixed" flow do not use this concept of relationship between density and flow. In these methods selected

facilities (e. g., stairs or doors) have a fixed flow which is independent from the density, which is usually not used in these methods. The "fixed" flow is usually based upon empirical and measured data of flow, which are specified for a special type of building, such as high-rise buildings or railway stations, for example. Because of much simplification, in these "fixed" flow methods a calculation can always be done very quickly.

Methods with "dynamic" flow allow one to describe the condition of the pedestrian flow in every part of a selected building or environment, because they are mostly based upon the continuity equation, thus it is possible to calculate different kinds of buildings. This allows the user to calculate transitions from wide to narrow, floor to door, floor to stair, etc. The disadvantage is that some these methods are very elaborate and time-intensive. But in general, a method with "dynamic" flow is not complicated to calculate, thus we want to divide hand calculation methods in simple [35,42,110,123,124,125,126,136,137,145,152,163,173,185,192] and complex [78,79,80,151] for evacuation calculation. All of these hand calculation methods are able to predict total evacuation times for a selected building, but differences between different methods still exist. Thus the user has to ensure that he is familiar with assumptions made by each method to ensure that a result is interpreted in a correct way [161].

Simulation of Evacuation Processes

Before we go into the details of evacuation simulation, let us briefly clarify its scope and limitations and contrast it to other methods used in evacuation analysis. When analyzing evacuation processes, three different approaches can be identified:

(1) Risk assessment,
(2) Optimization, and
(3) Simulation.

The aim and result of risk-assessment is a list of events and their consequences (e. g., damage, financial loss, loss of life), i. e., usually an event tree with probabilities and expectation values for financial loss. Optimization aims at, roughly speaking, minimizing the evacuation time and reducing the area and duration of congestion. And finally, simulation describes a system with respect to its function and behavior by investigating a model of the system. This model is usually non-analytic, so does not provide explicit equations for the calculation of, e. g., evacuation time. Of course, simulations are used for "optimization" in a more general sense, too, i. e., they can be part of an optimization. This holds for risk assessment, too, if simulations are used to determine the outcomes of the different scenarios in the event tree.

In evacuation analysis the system is, generally speaking, a group of persons in an environment. More specifically, four components (sub-systems/sub-models) of the system *evacuation process* can be identified:

(1) Geometry,
(2) Environment,
(3) Population, and
(4) Hazards [43].

Any evacuation simulation must at least take into account (1) and (3). The behavior of the persons (which can be described on the strategic, tactical, and operational level—see Sect. "Modeling") is part of the population sub-model. An alternative way of describing behavior is according to its algorithmic representation: no behavior modeling – functional analogy – implicit representation (equation) – rule based – artificial intelligence [43].

In the context of evacuation, hazards are first of all fire and smoke, which then require a toxicity sub-model, e. g., the fractional effective dose model (FED), to assess the physiological effect of toxic gases and temperature [25]. Further hazards to take into account might be earthquakes, flooding, or in the case of ships, list, heel, or roll motion. The sub-model environment comprises all other influences that affect the evacuation process, e. g., exit signs, surface texture, public address system, etc.

In summary, aims of an evacuation analysis and simulation are to provide feedback and hints for improvement at an early stage of design, information for safer and more rigorous regulations, improvement of emergency preparedness, training of staff, and accident investigation [43]. They usually do not provide direct results on the probability of a scenario or a systematic search for optimal geometries.

Calculation of Overall Evacuation Time, Identification of Congestion, and Corrective Actions The scope of this section is to show general results that can be obtained by evacuation simulations. They are general in the sense that they can basically be obtained by any stochastic and microscopic model, i. e., apart from these two requirements, the results are not model specific. In detail, five different results of evacuation simulations can be distinguished:

(1) Distribution of evacuation times,
(2) Evacuation curve (number of persons evacuated vs. time),

(3) Sequence of the evacuation (e. g., snapshots/screenshots at specific times, e. g., every minute), and
(4) Identification of congestion, usually based on density and time.

The last point (4), in particular, needs some more explanation: congestion is defined based on density. Notwithstanding the difficulties of measuring density, we suggest density as the most suitable criterion for the identification of congestion. In addition to the mere occurrence of densities exceeding a certain threshold (say 3.5 persons per square meter), the time this threshold is exceeded is another necessary condition for a sensible definition of congestion. In the case presented here, 10% of the overall evacuation time is used. Both criteria are in accordance with the IMO regulations [118].

Based on these results, evacuation time and areas of congestion, corrective actions can be taken. The most straightforward measure would be a change of geometry, i. e., shorter or wider escape paths (floors, stairs, doors). This can be directly put into the geometry sub-model, the simulation can be re-run, and the result checked. Secondly, the signage, and therefore the orientation capability, can be improved. This is not as straightforward as geometrical changes. It does depend more heavily on the model characteristics as to how these changes influence the evacuation sequence.

We will not go into these details in the following two sections but rather show two typical examples for evacuation simulations and the results obtained. We will also not discuss the results in detail, since they are of an illustrative nature in the context of this article. The following examples are based on investigations that have been performed using a cellular automaton model which is described along with the simulation program in [90,111].

Simulation Example 1 – Hotel The first example we show is a hotel with 8069 persons. In Fig. 15 only the ground floor is shown. There are nine floors altogether. The upper floors influence the ground floor only via the stair landings and the exits adjacent to them. Most of the 8069 persons are initially located in the ground floor, since the theater and conference area is located there. The upper floors are mainly covering bedrooms and some small conference areas.

The first step in our example (which might well be a useful recipe for evacuation analyses in general and is again in accordance with [118]) is to perform a statistical analysis. To this end, 500 samples are simulated. The evacuation time of a single run is the time it takes for all persons to get out. In this context, no fire or smoke are taken into account. Since there are stochastic influences in the model used, the significant overall evacuation time is taken to be the 95-percentile (cf. Fig. 14). Finally, the maximum, minimum, mean, and significant values for the evacuation curve (number of persons evacuated vs. time) are also shown in Fig. 14.

The next figure (Fig. 16) shows the cumulated density. The thresholds (red areas) are 3.5 persons per square meter and 10% of the overall evacuation time (in this case 49 seconds). The overall evacuation time is 8:13 minutes (493 seconds). This value is obtained by taking the 95-percentile of the frequency distribution for the overall evacuation times (cf. Fig. 14).

Of course, a distribution of overall evacuation times (for one scenario, i. e., the same initial parameters) can only be obtained by a stochastic model. In a deterministic model only one single value is calculated for the overall evacuation time. The variance of the overall evacuation times is due to two effects in the model used here: the initial position of the persons is determined anew at the beginning of each simulation run since only the statistical properties of the overall population are set and the motion of the persons is governed by partially stochastic rules (e. g., probabilistic parameters).

Simulation Example 2 – Passenger Ship The second example we will show is a ship. The major difference from the previous example is the addition of (1) the assembly phase and (2) embarkation and launching.

$$\begin{aligned} T &= A + \frac{2}{3}(E + L) \\ &= f_{\text{safety}} \cdot (t_{\text{react}} + t_{\text{walk}}) + \frac{2}{3}(E + L) \\ &\leq 60 \text{ minutes} . \end{aligned}$$

Embarkation and launching time $(E + L)$ are required to be less than 30 minutes. For the sake of the evacuation analysis at an early design stage, the sum of embarkation and launching time can be assumed to be 30 minutes. Therefore, the requirement for A is 40 minutes. Alternatively, the embarkation and launching time can be determined by an evacuation trial.

Figure 17 shows the layout, initial population distribution (night case), density plot for the day case, and density plot for the night case. The reaction times are different for the day and the night case: 3 to 7 minutes (equally distributed) in the one and 7 to 13 minutes in the other. The longer reaction time in the night case results in less congestion (cf. Fig. 17). Both cases must be done in the analysis according to [118]. Additionally, a secondary night and day case are required (making up four cases altogether). In these secondary cases the main vertical zone (MVZ) lead-

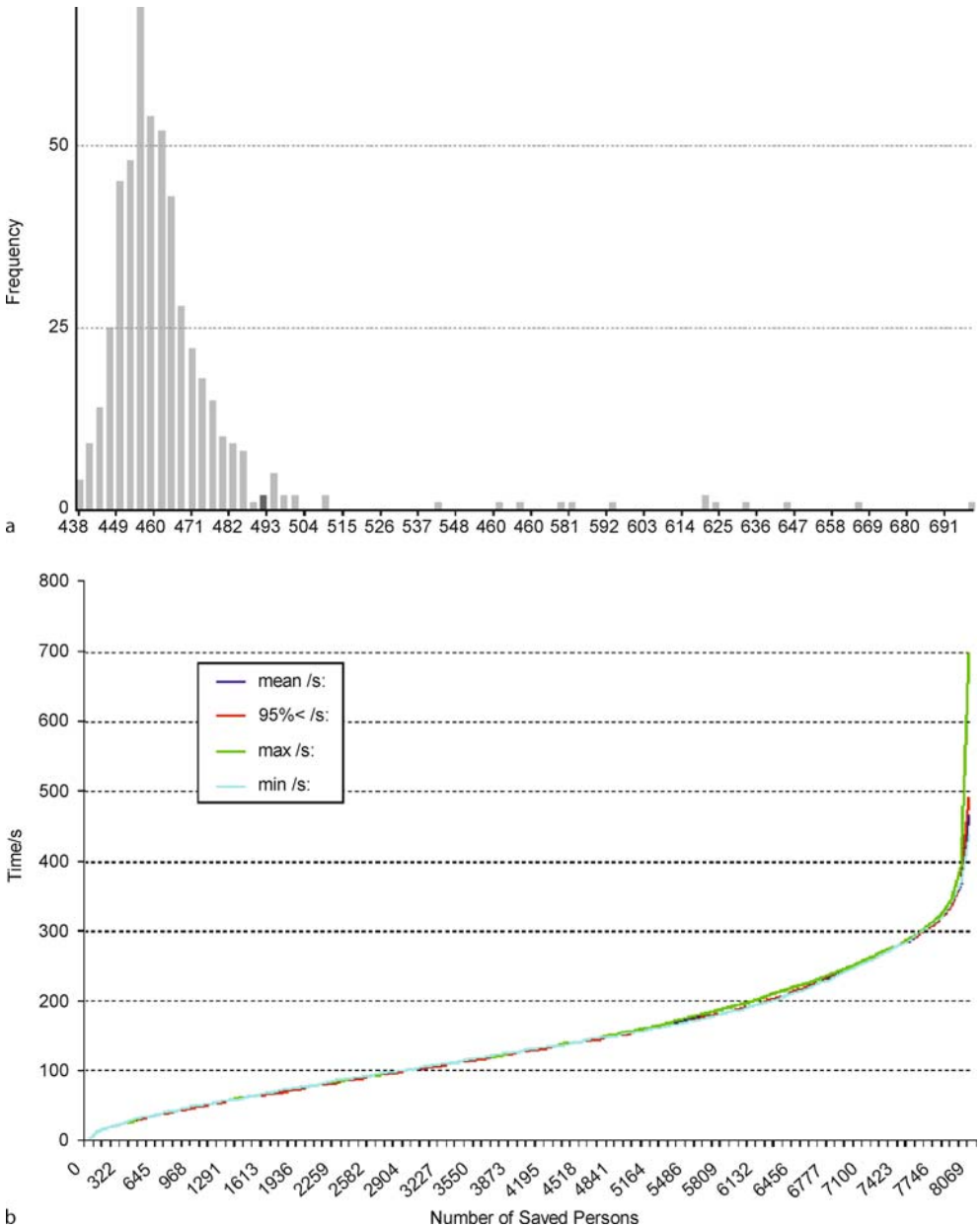

Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 14
Frequency distribution for the overall evacuation time (a) and evacuation curve (b)

ing to the longest overall individual assembly time is identified, and then either half of the stairway capacity in this zone is assumed to be not available, or 50% of the persons initially located in this zone must be led via one neighboring zone to the assembly station.

In the same way as shown for the two examples, simulations can be performed for other types of buildings and vessels. This technique has been applied to various passengers ships [112] to football stadiums [88] and the World Youth Day 2005 [88], the Jamarat Bridge in Makkah [88], a movie theater and schools (mainly for calibration and validation) [90] and airports [171]. Of course, many examples of applications based on various models can be found in the literature. For an overview, the proceedings of the PED conference series are an excellent starting point [44, 170,190].

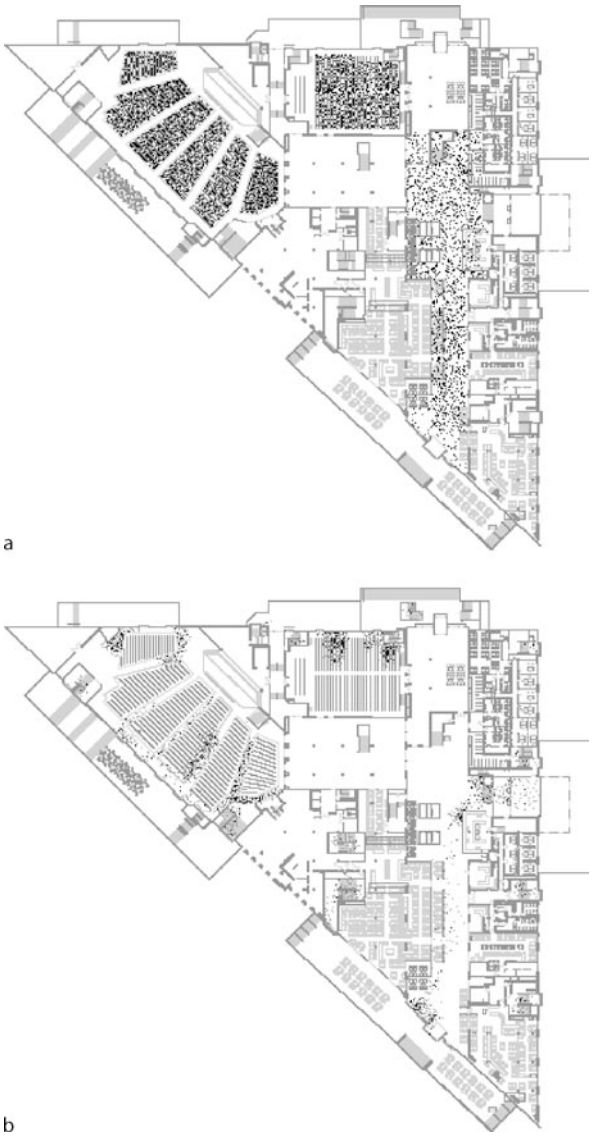

Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 15
Initial population distribution and situation after two minutes

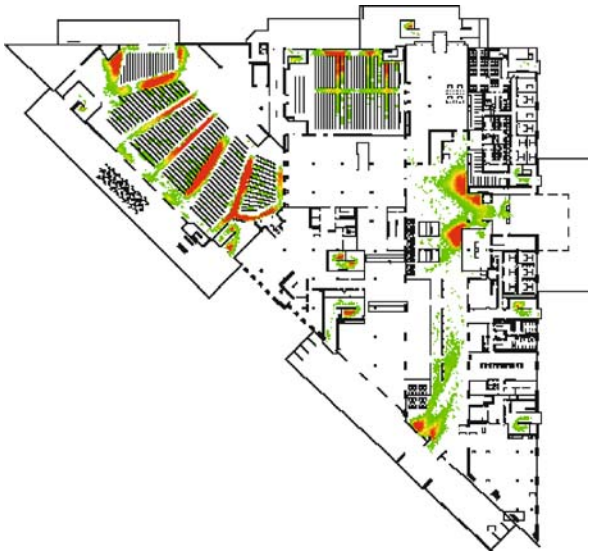

Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 16
Density plot, i. e., cumulated person density exceeding 3.5 persons per square meter and 10% of overall evacuation time

Comparison of Commercial Software Tools

From a practical point of view, application of models for pedestrian dynamics and evacuation processes becomes more and more relevant in safety analysis. This has led to the development of a number of software tools that, with different sophistication, help us study many aspects without risking the health of test persons in evacuation trials.

There are commercial, as well as non-commercial software tools. All tools might be based on different models [97,187]. They have become very popular since the middle of the 1990s. A first comparison of different commercial software tools can be found in [191], where they were said to produce "reasonable results". Further comparisons of real evacuation data with software tools or hand calculation methods can be found in [29,67,91,96,104,160,161,177]. But results predicted by different commercial software tools can differ by up to 40% for the same building [96]. Results may differ, too, when calculating with different assumptions, e. g., different reaction times, use of more or less detailed stair models, or when calculating with a real occupant load in contrast to an uncertainty analysis [96,104]. Contrary to these results, another study [161] shows that calculations with different software tools are able to predict total evacuation times for high-rise buildings and there are no large differences as shown in [96]. In [161] the results of an evacuation trial and simulations with different commercial software tools differed for selected floors of a highrise building. The densities were very low in this instance. In this case human behavior has a very large influence on the evacuation time. By contrast, evacuations at medium or high densities, human behavior has a smaller influence on the evacuation time of selected areas because congestion appears and continues larger than in low density situations – thus people reach the exit while congestion is still a factor [162]. In low density situations congestions are very rare, thus people move narrowly with free walking velocity through the building [162].

But the results presented in [161] also show that commercial software tools sometimes have problems with the

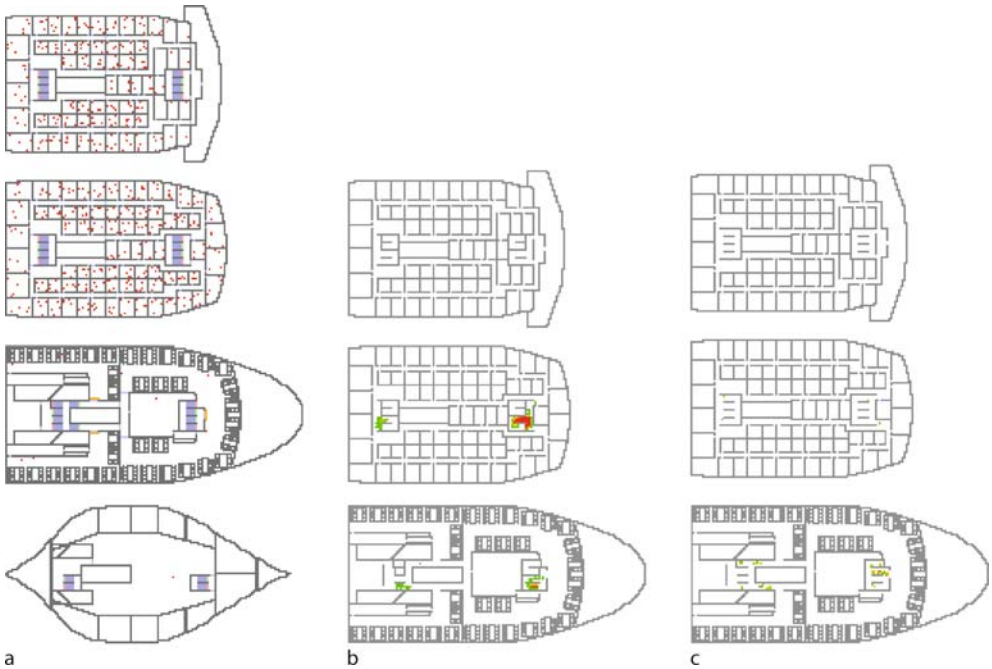

Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 17
Initial distribution for the night case, density plot for the day case, and density plot for the night case for the "AENEAS steamliner"

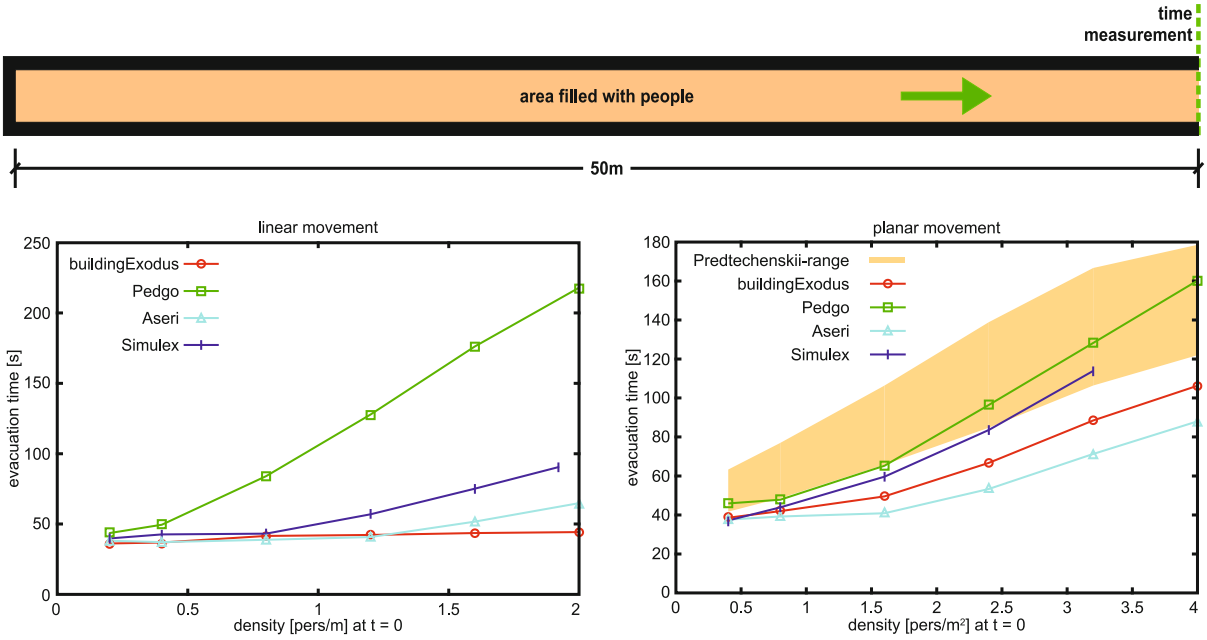

Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 18
Comparison of different software tools by simulating linear (*left*, narrow floor) and planar (*right*, 2 m wide floor) movement [162]

empirical relationship of density and walking speed (see Fig. 18). Furthermore, it is very important how boundary conditions are implemented in these tools (see Fig. 19), and the investigation of a simple scenario of a single room using different software tools shows results differing by about a factor of two (see Fig. 19) [161]. In this case all software tools predict a congestion at the exit. Furthermore it is possible that the implemented algorithm fails [161].

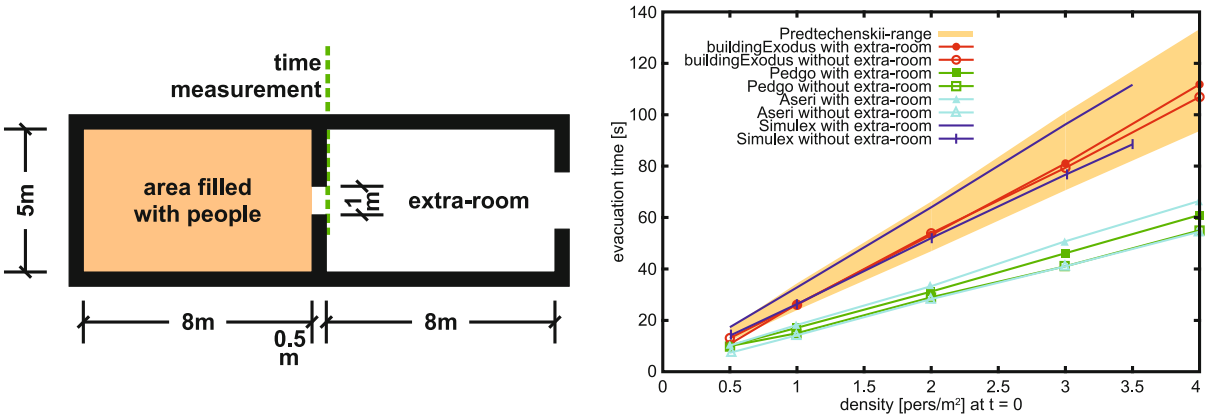

Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 19
Comparison of different software tools by simulating a simple room geometry [162]

Thus for the user it is hard to know which algorithms are implemented in closed-source tools so that such a tool must be considered as "black box" [147]. It is also quite difficult to compare results about density and appearing congestions calculated by different software tools [162] and so it is questionable how these results should be interpreted. But, as pointed out earlier, reliable empirical data are often missing so that a validation of software tools or models is quite difficult [162].

Future Directions

The discussion has shown that the problem of crowd dynamics and evacuation processes is far from being well understood. One big problem is experimental basis. As in many human systems, it is difficult to perform controlled experiments on a sufficiently large scale. This would be necessary since data from actual emergency situations is usually not available, at least in sufficient quality. Progress should be possible by using modern video and computer technology which should allow us, in principle, to extract precise data even for the trajectories of individuals.

The full understanding of the complex dynamics of evacuation processes requires collaboration between engineering, physics, computer science, psychology, etc. Engineering in cooperation with computer science will lead to an improved empirical basis. Methods from physics allow us to develop simple but realistic models that capture the main aspects of the dynamics. Psychology is then needed to understand the interactions between individuals in sufficient detail to get a reliable set of 'interaction' parameters for the physical models.

In the end, we hope these joint efforts will lead to realistic models for evacuation processes that not only allow us to study these in the planning stages of facilities, but even allow for dynamical real-time evacuation control in case an emergency occurs.

Acknowledgments

The authors would like to acknowledge the contribution of Tim Meyer-König (the developer of PedGo) and Michael Schreckenberg, Ansgar Kirchner, Bernhard Steffen for many fruitful discussions and valuable hints.

Bibliography

Primary Literature

1. Abe K (1986) The Science of Human Panic. Brain, Tokyo (in Japanese)
2. AlGadhi SAH, Mahmassani HS, Herman R (2002) A speed-concentration relation for bi-directional crowd movements with strong interaction. In: Schreckenberg M, Sharma SD (eds) Pedestrian and Evacuation Dynamics. Springer, Berlin, pp 3–20
3. American Sociological Association (2002) In disasters, panic is rare, altruism dominates. Technical report, American Sociological Association
4. Ashe B, Shields TJ (1999) Analysis and modelling of the unannounced evacuation of a large retail store. Fire Mater 23: 333–336
5. Ben-Jacob E (1997) From snowflake formation to growth of bacterial colonies, Part II. Cooperative formation of complex colonial patterns. Contemp Phys 38:205
6. Biham O, Middleton AA, Levine D (1992) Self-organization and a dynamical transition in traffic-flow models. Phys Rev A 46:R6124
7. Blue VJ, Adler JL (2000) Cellular automata microsimulation of bi-directional pedestrian flows. J Trans Res Board 1678: 135–141
8. Blue VJ, Adler JL (2002) Flow capacities from cellular automata modeling of proportional spilts of pedestrians by direction. In: Schreckenberg M, Sharma SD (eds) Pedestrian and Evacuation Dynamics. Springer, Berlin, pp 115–121

9. Blythe RA et al (2007) Nonequilibrium steady states of matrix product form: a solver's guide. Math Theor 40:R333–R441, doi:10.1088/1751-8113/40/46/R01
10. Bolay K (1998) Nichtlineare Phänomene in einem fluiddynamischen Verkehrsmodell. Diploma Thesis, Stuttgart University
11. Boyce KE, Shields TJ, Silcock GWH (1999) Toward the Characterization of Building Occupancies for Fire Safety Engineering: Capabilities of Disabled People Moving Horizontally and on an Incline. Fire Technol 35:51–67
12. Bryan JL (1995) Behavioral response to fire and smoke. In: DiNenno PJ, Beyler CL, Custer RLP, Walton WD, Watts JM, Drysdale D, Hall JR (eds) SFPE Handbook of Fire Protection Engineering, 2nd edn. National Fire Protection Association, Quincy, p 263
13. Burstedde C, Klauck K, Schadschneider A, Zittartz J (2001) Simulation of pedestrian dynamics using a two-dimensional cellular automaton. Physica A 295:507–525
14. Burstedde C, Kirchner A, Klauck K, Schadschneider A, Zittartz J (2002) Cellular automaton approach to pedestrian dynamics – applications. In: Schreckenberg M, Sharma SD (eds) Pedestrian and Evacuation Dynamics. Springer, Berlin, pp 87–98
15. Chakrabarti J, Dzubiella J, Löwen H (2004) Reentrance effect in the lane formation of driven colloids. Phys Rev E 70:012401
16. Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. Phys Rep 329(4–6):199–329
17. Clarke L (2002) Panic: Myth or reality? Contexts 1(3):21–26
18. Coleman JS (1990) Foundation of Social Theory. Belknap, Cambridge, Chap 9
19. Daamen W (2004) Modelling Passenger Flows in Public Transport Facilities. Ph.D. thesis, Technical University of Delft
20. Daamen W, Hoogendoorn SP (2006) Flow-density relations for pedestrian traffic. In: Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) Traffic and Granular Flow 05. Springer, Berlin, pp 315–322
21. Daamen W, Bovy PHL, Hoogendoorn SP (2002) Modelling pedestrians in transfer stations. In: Schreckenberg M, Sharma SD (eds) Pedestrian and Evacuation Dynamics. Springer, Berlin, pp 59–73
22. de Gelder B, Snyder J, Greve D, Gerard G, Hadjikhani N (2004) Fear fosters flight: A mechanism for fear contagion when perceiving emotion expressed by a whole body. Proc Natl Acad Sci 101(47):16701–16706
23. Derrida B (1998) An exactly soluble non-equilibrium system: The asymmetric simple exclusion process. Phys Rep 301:65
24. Dieckmann D (1911) Die Feuersicherheit in Theatern. Jung, München (in German)
25. DiNenno PJ (ed) (2002) SFPE Handbook of Fire Protection Engineering, 3rd edn. National Fire Protection Association, Bethesda
26. DiNenno PJ, Beyler CL, Custer RLP, Walton WD, Watts JM, Drysdale D, Hall JR (eds) (1995) SFPE Handbook of Fire Protection Engineering, 2nd edn. National Fire Protection Association, Quincy
27. Dogliani M (2002) An overview of present and underdevelopment IMO's requirements concerning evacuation from ships. In: Schreckenberg M, Sharma SD (eds) Pedestrian and Evacuation Dynamics. Springer, Berlin, pp 339–354
28. Dzubiella J, Hoffmann GP, Löwen H (2002) Lane formation in colloidal mixtures driven by an external field. Phys Rev E 65:021402
29. Ehm M, Linxweiler J (2004) Berechnungen von Evakuierungszeiten bei Sonderbauten mit dem Programm buildingExodus. Technical report, TU Braunschweig
30. El Yacoubi S, Chopard B, Bandini S (eds) (2006) Cellular Automata – 7th International Conference on Cellular Automata for Research and Industry, ACRI 2006, Perpignan. Springer, Berlin
31. Federal Aviation Administration FAA (1990) Emergency evacuation – cfr sec. 25.803. Regulation CFR Sec. 25.803
32. Fischer H (1933) Über die Leistungsfähigkeit von Türen, Gängen und Treppen bei ruhigem, dichtem Verkehr. Dissertation, Technische Hochschule Dresden (in German)
33. Frantzich H (1996) Study of movement on stairs during evacuation using video analysing techniques. Technical report, Department of Fire Safety Engineering, Lund Institute of Technology
34. Frisch U, Hasslacher B, Pomeau Y (1986) Lattice-gas automata for the Navier-Stokes equation. Phys Rev Lett 56:1505
35. Fruin JJ (1971) Pedestrian Planning and Design. Metropolitan Association of Urban Designers and Environmental Planners, New York
36. Fruin JJ (1993) The causes and prevention of crowd disasters. In: Smith RA, Dickie JF (eds) Engineering for Crowd Safety. Amsterdam, Elsevier
37. Fujiyama T (2006) Collision avoidance of pedestrians on stairs. Technical report, Centre for Transport Studies. University College London, London
38. Fujiyama T, Tyler N (2004) An explicit study on walking speeds of pedestrians on stairs. In: 10th International Conference on Mobility and Transport for Elderly and Disabled People, Hamamatsu, Japan, May 2004
39. Fujiyama T, Tyler N (2004) Pedestrian Speeds on Stairs: An Initial Step for a Simulation Model. In: Proceedings of 36th Universities Transport Studies Group Conference, Life Science Centre, Newcastle upon Tyne, Jan 2004
40. Fukui M, Ishibashi Y (1999) Jamming transition in cellular automaton models for pedestrians on passageway. J Phys Soc Jpn 68:3738
41. Fukui M, Ishibashi Y (1999) Self-organized phase transitions in cellular automaton models for pedestrians. J Phys Soc Jpn 68:2861
42. Galbreath M (1969) Time of evacuation by stairs in high buildings. Fire Res Note 8, NRCC
43. Galea ER (2002) Simulating evacuation and circulation in planes, trains, buildings and ships using the EXODUS software. In: Schreckenberg M, Sharma SD (eds) Pedestrian and Evacuation Dynamics. Springer, Berlin pp 203–226
44. Galea ER (ed) (2003) Pedestrian and Evacuation Dynamics 2003. CMS Press, London
45. Gipps PG, Marksjö B (1985) A micro-simulation model for pedestrian flows. Math Comput Simul 27:95–105
46. Graat E, Midden C, Bockholts P (1999) Complex evacuation; effects of motivation level and slope of stairs on emergency egress time in a sports stadium. Saf Sci 31:127–141
47. Grosshandler W, Sunder S, Snell J (2003) Building and fire safety investigation of the world trade center disaster. In: Galea ER (ed) Pedestrian and Evacuation Dynamics 2003. CMS Press, London, pp 279–281

48. Hamacher HW, Tjandra SA (2002) Mathematical modelling of evacuation problems – a state of the art. In: Schreckenberg M, Sharma SD (eds) Pedestrian and Evacuation Dynamics. Springer, Berlin, pp 227–266
49. Hankin BD, Wright RA (1958) Passenger flow in subways. Oper Res Q 9:81–88
50. Helbing D (1992) A fluid-dynamic model for the movement of pedestrians. Complex Syst 6:391–415
51. Helbing D (2001) Traffic and related self-driven many-particle systems. Rev Mod Phys 73:1067–1141
52. Helbing D, Molnár P (1995) Social force model for pedestrian dynamics. Phys Rev E 51:4282–4286
53. Helbing D, Farkas I, Vicsek T (2000) Freezing by heating in a driven mesoscopic system. Phys Rev Let 84:1240–1243
54. Helbing D, Farkas I, Vicsek T (2000) Simulating dynamical features of escape panic. Nature 407:487–490
55. Helbing D, Farkas I, Molnár P, Vicsek T (2002) Simulation of pedestrian crowds in normal and evacuation situations. In: Schreckenberg M, Sharma SD (eds) Pedestrian and Evacuation Dynamics. Springer, Berlin, pp 21–58
56. Helbing D, Buzna L, Werner T (2003) Self-organized pedestrian crowd dynamics and design solutions. Traffic Forum, pp 2003-12
57. Helbing D, Buzna L, Johansson A, Werner T (2005) Self-organized pedestrian crowd dynamics: Experiments, simulations, and design solutions. Transp Sci 39:1–24
58. Helbing D, Johansson A, Al-Abideen HZ (2007) The dynamics of crowd disasters: An empirical study. Phys Rev E 75:046109
59. Helbing D, Johannson A, Al-Abideen HZ (2007) Crowd turbulence: the physics of crowd disasters. In: The Fifth International Conference on Nonlinear Mechanics, ICNM-V, Shanghai, pp 967–969
60. Henderson LF (1971) The statistics of crowd fluids. Nature 229:381–383
61. Henderson LF (1974) On the fluid mechanics of human crowd motion. Transp Res 8:509–515
62. Hoogendoorn SP (2003) Walker behaviour modelling by differential games. In: Emmerich H, Nestler B, Schreckenberg M (eds) Interface and transport dynamics. Lecture notes in Computational Science and Engineering, vol 32. Springer, Berlin, pp 275–294
63. Hoogendoorn SP, Bovy PHL (2003) Simulation of pedestrian flows by optimal control and differential games. Optim Control Appl Meth 24:153
64. Hoogendoorn SP, Daamen W (2005) Pedestrian behavior at bottlenecks. Transp Sci 39 2:0147–0159
65. Hoogendoorn SP, Bovy PHL, Daamen W (2002) Microscopic pedestrian wayfinding and dynamics modelling. In: Schreckenberg M, Sharma SD (eds) Pedestrian and Evacuation Dynamics. Springer, Berlin, pp 123–154
66. Hoogendoorn SP, Daamen W, Bovy PHL (2003) Microscopic pedestrian traffic data collection and analysis by walking experiments: Behaviour at bottlenecks. In: Galea ER (ed) Pedestrian and Evacuation Dynamics 2003. CMS Press, London, pp 89–100
67. Hoskin K (2004) Fire protection and evacuation procedures of stadia venues in new zealand. Master's thesis, University of Canterburry
68. Hughes RL (2000) The flow of large crowds of pedestrians. Math Comput Simul 53:367–370
69. Hughes RL (2002) A continuum theory for the flow of pedestrians. Transp Res Part B 36:507–535
70. International Maritime Organization (IMO) (2000) International Code of Safety for High-Speed Craft, 2000 (2000 HSC Code). Technical report, Resolution MSC 97(73)
71. International Organization for Standardization (2000) ISO-TR-13387-8-1999 Fire safety engineering – part 8: Life safety – occupant behaviour, location and condition. Technical report
72. Jian L, Lizhong Y, Daoling Z (2005) Simulation of bi-direction pedestrian movement in corridor. Physica A 354:619
73. Johnson NR (1987) Panic at "The Who Concert Stampede": An Empirical Assessment. Soc Probl 34(4):362–373
74. Jungermann H, Göhlert C (2000) Emergency evacuation from double-deck aircraft. In: Cottam MP, Harvey DW, Pape RP, Tait J (eds) Foresight and Precaution. Proceedings of ESREL 2000, SARS and SRA. Europe Annual Conference, Rotterdam, pp 989–992
75. Kashiwagi T (ed) (1994) Fire Safety Science – 4th international Symposium Proceedings. Interscience, West Yard House, Guildford. The International Association for Fire Safety Science. Grove, London
76. Kaufman M (2007) Lane Formation in Counterflow Situations of Pedestrian Traffic. Master's thesis, Universität Duisburg-Essen
77. Keating JP (1982) The myth of panic. Fire J May:57–62
78. Kendik E (1983) Determination of the evacuation time pertinent to the projected area factor in the event of total evacuation of high-rise office buildings via staircases. Fire Saf J 5:223–232
79. Kendik E (1984) Die Berechnung der Personenströme als Grundlage für die Bemessung von Gehwegen in Gebäuden und um Gebäude. Ph.D. thesis, TU Wien
80. Kendik E (1986) Designing escape routes in buildings. Fire Technol 22:272–294
81. Kerner BS (2004) The Physics of Traffic. Springer, Heidelberg
82. Kirchner A (2003) Modellierung und statistische Physik biologischer und sozialer Systeme. Dissertation, Universität zu Köln
83. Kirchner A, Schadschneider A (2002) Simulation of evacuation processes using a bionics-inspired cellular automaton model for pedestrian dynamics. Physica A 312:260–276
84. Kirchner A, Namazi A, Nishinari K, Schadschneider A (2003) Role of conflicts in the floor field cellular automaton model for pedestrian dynamics. In: Galea ER (ed) Pedestrian and Evacuation Dynamics 2003. CMS Press, London, pp 51
85. Kirchner A, Nishinari K, Schadschneider A (2003) Friction effects and clogging in a cellular automaton model for pedestrian dynamics. Phys Rev E 67:056122
86. Kirchner A, Klüpfel H, Nishinari K, Schadschneider A, Schreckenberg M (2004) Discretization effects and the influence of walking speed in cellular automata models for pedestrian dynamics. J Stat Mech 10:P10011
87. Klüpfel H (2003) A Cellular Automaton Model for Crowd Movement and Egress Simulation. Dissertation, University Duisburg–Essen
88. Klüpfel H (2006) The simulation of crowds at very large events. In: Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) Traffic and Granular Flow 05. Springer, Berlin, pp 341–346
89. Klüpfel H, Meyer-König T, Wahle J, Schreckenberg M (2000) Microscopic simulation of evacuation processes on passen-

ger ships. In: Bandini S, Worsch T (eds) Theory and Practical Issues on Cellular Automata. Springer, Berlin
90. Klüpfel H, Meyer-König T, Schreckenberg M (2001) Empirical data on an evacuation exercise in a movie theater. Technical report, University Duisburg-Essen
91. Ko SY (2003) Comparison of evacuation times using Simulex and EvacuatioNZ based on trial evacuations. Fire Engineering Research Report 03/9, University of Canterbury
92. Kretz T (2007) Pedestrian Traffic – Simulation and Experiments. Dissertation, Universität Duisburg-Essen
93. Kretz T, Grünebohm A, Schreckenberg M (2006) Experimental study of pedestrian flow through a bottleneck. J Stat Mech P10014
94. Kretz T, Grünebohm A, Kaufmann M, Mazur F, Schreckenberg M (2006) Experimental study of pedestrian counterflow in a corridor. J Stat Mech P10001
95. Kretz T, Grünebohm A, Keßel A, Klüpfel H, Meyer-König T, Schreckenberg M (2008) Upstairs walking speed distributions on a long stair. Saf Sci 46:72–78
96. Kuligowski ED, Milke JA (2005) A perfomance-based egress analysis of a hotel building using two models. J Fire Prot Eng 15:287–305
97. Kuligowski ED, Peacock RD (2005) A review of building evacuation models. Technical report 1471. National Institute of Standards and Technology, Gaithersburg
98. Lakoba TI, Kaup DJ, Finkelstein NM (2005) Modifications of the Helbing-Molnár-Farkas-Vicsek social force model for pedestrian evolution. Simulation 81(5):339–352
99. Lam WHK, Lee JYS, Chan KS, Goh PK (2003) A generalised function for modeling bi-directional flow effects on indoor walkways in Hong Kong. Transp Res A: Policy Pract 37:789–810
100. Laur U, Jaakula H, Metsaveer J, Lehtola K, Livonen H, Karppinen T, Eksborg AL, Rosengren H, Noord O (1997) Final Report on the Capsizing on 28 September 1994 in the Baltic Sea of the Ro-Ro Passenger Vessel MV Estonia. Technical report. The Joint Accident Investigation Commission of Estonia, Finland and Sweden
101. LeBon G (1895) Lois Psychlogiques De L'evolution Des Peuples. Alcan, Paris
102. Leutzbach W (1988) Introduction to the Theory of Traffic Flow. Springer, Berlin
103. Lewin K (1951) Field Theory in Social Science. Harper, New York
104. Lord J, Meacham B, Moore A, Fahy R, Proulx G (2005) Guide for evaluating the predictive capabilities of computer egress models. Technical report NIST GCR 06–886, NIST, Gaithersburg
105. Lovas GG (1994) Modeling and simulation of pedestrian traffic flow. Transp Res B 28V:429
106. Maniccam S (2003) Traffic jamming on hexagonal lattice. Physica A 321:653
107. Maniccam S (2005) Effects of back step and update rule on congestion of mobile objects. Physica A 346:631
108. Marconi S, Chopard B (2002) A multiparticle lattice gas automata model for a crowd. In: Cellular Automata. Lecture Notes Computer Science, vol 2493. Springer, Berlin, pp 231
109. Mawson AR (2005) Understanding mass panic and other collective responses to threat and disaster. Psychiatry 68:95–113
110. Melinek SJ, Booth S (1975) An analysis of evacuation times and the movement of crowds in buildings. Technical report CP 96/75, BRE
111. Meyer-König T, Klüpfel H, Schreckenberg M (2001) A microscopic model for simulating mustering and evacuation processes onboard passenger ships. In: KH Drager (ed) Proceedings of the International Emergency Management Society Conference. The International Emergency Management Society, Oslo
112. Meyer-König T, Klüpfel H, Schreckenberg M (2002) Assessment and analysis of evacuation processes on passenger ships by microscopic simulation. In: Schreckenberg M, Sharma SD (eds) Pedestrian and Evacuation Dynamics. Springer, Berlin, pp 297–302
113. Mintz A (1951) Non-adaptive group behaviour. J Abnorm Soc Psychol 46:150–159
114. Molnár P (1995) Modellierung und Simulation der Dynamik von Fußgängerströmen. Dissertation, Universität Stuttgart
115. Mori M, Tsukaguchi H (1987) A new method for evaluation of level of service in pedestrian facilities. Transp Res 21A(3): 223–234
116. Morrall JF, Ratnayake LL, Seneviratne PN (1991) Comparison of CBD pedestrian characteristics in Canada and Sri Lanka. In: Transportation Research Record 1294. TRB, National Research Council, Washington DC, pp 57–61
117. MSA (1997) Report on Exercise Invicta. Technical report. Marine Safety Agency, Southhampton
118. MSC-Circ.1033. Interim guidelines for evacuation analyses for new and existing passenger ships. Technical report, International Maritime Organization, Marine Safety Committee, London, June, 6th 2002. MSC/Circ. 1033
119. MSC-Circ.1166. Guidelines for a simplified evacuation analysis for high-speed passenger craft. Technical report, International Maritime Organisation, 2005
120. Muir HC (1997) Airplane of the 21st century: Challenges in safety and survivability. International Conference on Aviation Safety and Security in the 21st Century, White House Commission on Aviation Safety and Security, Washington
121. Muir HC, Bottomley DM, Marrison C (1996) Effects of motivation and cabin configuration on emergency aircraft evacuation behavior and rates of egress. Int J Aviat Psychol 6(1):57–77
122. Müller K (1981) Zur Gestaltung und Bemessung von Fluchtwegen für die Evakuierung von Personen aus Bauwerken auf der Grundlage von Modellversuchen. Dissertation, Technische Hochschule Magdeburg
123. Müller W (1966) Die Beurteilung von Treppen als Rückzugsweg in mehrgeschossigen Gebäuden. Unser Brandschutz – Wissenschaftlich-Technische Beil 3:65–70; to be continued in 4/1966
124. Müller W (1966) Die Beurteilung von Treppen als Rückzugsweg in mehrgeschossigen Gebäuden. Unser Brandschutz – Wissenschaftlich-Technische Beil 4:93–96; continuation from 3/1966
125. Müller W (1968) Die Überschneidung der Verkehrsströme bei dem Berechnen der Räumungszeit von Gebäuden. Unser Brandschutz – Wissenschaftlich-Technische Beil 4:87–92
126. Müller W (1970) Untersuchung über zulässige Räumungszeiten und die Bemessung von Rückzugswegen in Gebäuden. Habilitation, TU Dresden, Dresden

127. Muramatsu M, Nagatani T (2000) Jamming transition in two-dimensional pedestrian traffic. Physica A 275:281–291
128. Muramatsu M, Nagatani T (2000) Jamming transition of pedestrian traffic at crossing with open boundary conditions. Physica A 286:377–390
129. Muramatsu M, Irie T, Nagatani T (1999) Jamming transition in pedestrian counter flow. Physica A 267:487–498
130. Argebau (2005) MVStättV – Erläuterungen: Musterverordnung über den Bau und Betrieb von Versammlungsstätten. Erläuterungen, Juni 2005
131. Nagai R, Nagatani T (2006) Jamming transition in counter flow of slender particles on square lattice. Physica A 366:503
132. Nagai R, Fukamachi M, Nagatani T (2006) Evacuation of crawlers and walkers from corridor through an exit. Physica A 367:449–460
133. Nagel K, Schreckenberg M (1992) A cellular automaton model for freeway traffic. J Phys I 2:2221
134. Nakayama A, Hasebe K, Sugiyama Y (2005) Instability of pedestrian flow and phase structure in a two-dimensional optimal velocity model. Phys Rev E 71:036121
135. Navin PD, Wheeler RJ (1969) Pedestrian flow characteristics. Traffic Eng 39:31–36
136. Nelson HE, Mowrer FW (2002) Emergency movement. In: DiNenno PJ (ed) SFPE Handbook of Fire Protection Engineering, 3rd edn. National Fire Protection Association, Bethesda, p 367
137. National Fire Protection Association (2007) NFPA 130: Standard for Fixed Guideway Transit and Passenger Rail Systems.
138. Norwegian Ministry of Justice and Police (2000) The High-Speed Craft MS Sleipner Disaster, 26 November 1999. Official Norwegian Reports 2000:31, Oslo
139. Oeding D (1963) Verkehrsbelastung und Dimensionierung von Gehwegen und anderen Anlagen des Fußgängerverkehrs. Forschungsbericht, vol 22. Technische Hochschule Braunschweig
140. O'Flaherty CA, Parkinson MH (1972) Movement in a city centre footway. Traffic Eng Control, p 434
141. Okazaki S, Matsushita S (1993) A study of simulation model for pedestrian movement with evacuation and queuing. In: Smith RA, Dickie JF (eds) Proceedings International Conference Engineering Crowd Safety. Elsevier, Amsterdam, pp 271
142. Older SJ (1968) Movement of pedestrians on footways in shopping streets. Traffic Eng Control 10:160–163
143. Owen M, Galea ER, Lawrence PJ, Filippidis L (1998) AASK – aircraft accident statistics and knowledge: a database of human experience in evacuation, derived from aviation accident reports. Aero J 102:353–363
144. Pauls JL (1971) Evacuation drill held in the b. c. hydro building, 26 june 1969. Building Research Note 80, National Republican Congressional Committee
145. Pauls JL (1995) Movement of people. In: DiNenno PJ, Beyler CL, Custer RLP, Walton WD, Watts JM, Drysdale D, Hall JR (eds) SFPE Handbook of Fire Protection Engineering, 2nd edn. National Fire Protection Association, Quincy, p 263
146. Pauls JL, Fruin JJ, Zupan JM (2007) Minimum stair width for evacuytion, overtaking movement and counterflow – technical bases and suggestions for the past, present and future. In: Waldau N, Gattermann P, Knoflacher H, Schreckenberg M (eds) Pedestrian and Evacuation Dynamics 2005. Springer, Berlin, pp 57–69
147. Paulsen T, Soma H, Schneider V, Wiklund J, Lovas G (1995) Evaluation of simulation models of evacuating from complex spaces. SINTEF Report STF75 A95020. SINTEF, Trondheim
148. Polus A, Joseph JL, Ushpiz A (1983) Pedestrian flow and level of service. J Transp Eng 109(1):46–56
149. Popkov V, Schütz G (1999) Steady-state selection in driven diffusive systems with open boundaries. Europhys Lett 48:257
150. Predtechenskii VM, Milinskii AI (1969) Planing for foot traffic flow in buildings. Amerind Publishing, New Dehli, 1978. Translation of: Proekttirovanie Zhdaniis Uchetom Organizatsii Dvizheniya Lyuddskikh Potokov, Stroiizdat Publishers, Moscow
151. Predtetschenski WM, Milinski AI (1971) Personenströme in Gebäuden – Berechnungsmethoden für die Modellierung. Müller, Köln-Braunsfeld
152. Predtetschenski WM, Cholstschewnikow WW, Völkel H (1972) Vereinfachte Berechnung der Umformung von Personenströmen auf Wegabschnitten mit begrenzter Länge. Unser Brandschutz Wissenschaftlich-Technische Beil 6:90–94
153. Purser DA, Bensilium M (2001) Quantification of behaviour for engineering design standards and escape time calculations. Saf Sci 38(2):158–182
154. Pushkarev B, Zupan JM (1975) Capacity of walkways. Transp Res Rec 538:1–15
155. Quarantelli EL (1960) Images of withdrawal behavior in disasters: Some basic misconceptions. Soc Probl 8:63–79
156. Quarantelli EL (2001) The sociology of panic. In: Smelser NJ, Baltes PB (eds) International Encyclopedia of the Social and Behavioral Sciences. Pergamon, New York, pp 11020–11030
157. Revi A, Singh AK (2007) Cyclone and storm surge, pedestrian evacuation and emergency response in India. In: Waldau N, Gattermann P, Knoflacher H, Schreckenberg M (eds) Pedestrian and Evacuation Dynamics 2005. Springer, Berlin, pp 119–130
158. Rex M, Löwen H (2007) Lane formation in oppositely charged colloids driven by an electric field: Chaining and two-dimensional crystallization. Phys Rev E 75:051402
159. Rickert M, Nagel K, Schreckenberg M, Latour A (1996) Two lane traffic simulations using cellular automata. Physica A 231:534
160. Rogsch C (2005) Vergleichende Untersuchungen zur dynamischen Simulation von Personenströmen. Technical report JUEL-4185. Forschungszentrum Jülich
161. Rogsch C, Klingsch W, Seyfried A, Weigel H (2007) How reliable are commercial software-tools for evacuation calculation? In: Interflam 2007 – Conference Proceedings. Interscience Communication Ltd, Greenwich, London, pp 235–245
162. Rogsch C, Klingsch W, Seyfried A, Weigel H (2007) Prediction accuracy of evacuation times for high-rise buildings and simple geometries by using different software-tools. In Traffic and Granular Flow 2007. Preprint
163. Roitman MJ (1966) Die Evakuierung von Menschen aus Bauwerken. Staatsverlag der Deutschen Demokratischen Republik
164. Rothman DH, Zaleski S (1994) Lattice-gas models of phase separation: Interfaces, phase transitions, and multiphase flow. Rev Mod Phys 66:1417
165. Rothman DH, Zaleski S (1997) Lattice-Gas Cellular Automata. Cambridge University Press, Cambridge
166. Saloma C, Perez GJ (2007) Herding in real escape panic. In: Waldau N, Gattermann P, Knoflacher H, Schreckenberg M

(eds) Pedestrian and Evacuation Dynamics 2005. Springer, Berlin, pp 471–479
167. Schadschneider A (2002) Cellular automaton approach to pedestrian dynamics – theory. In: Schreckenberg M, Sharma SD (eds) Pedestrian and Evacuation Dynamics. Springer, Berlin, pp 75–86
168. Schelajew J, Schelajewa E, Semjonow N (2000) Nikolaus II. Der letzte russische Zar. Bechtermünz, Augsburg
169. Schneider U, Kath K, Oswald M, Kirchberger H (2006) Evakuierung und Verhalten von Personen im Brandfall unter spezieller Berücksichtigung von schienengebundenen Fahrzeugen. Technical report 12, TU Wien
170. Schreckenberg M, Sharma SD (eds) (2007) Pedestrian and Evacuation Dynamics. Springer, Berlin
171. Schultz M, Lehmann S, Fricke H (2007) A discrete microscopic model for pedestrian dynamics to manage emergency situations in airport terminals. In: Waldau N, Gattermann P, Knoflacher H, Schreckenberg M (eds) Pedestrian and Evacuation Dynamics 2005. Springer, Berlin, pp 389–395
172. Schütz GM (2001) Exactly solvable models for many-body systems. In: Domb C, Lebowitz JL (eds) Phase Transitions and Critical Phenomena, vol 19. Academic Press, Amsterdam
173. Seeger PG, John R (1978) Untersuchung der Räumungsabläufe in Gebäuden als Grundlage für die Ausbildung von Rettungswegen, Teil III: Reale Räumungsversuche. Technical report T395. Forschungsstelle für Brandschutztechik an der Universität Karlsruhe (TH)
174. Seyfried A, Steffen B, Klingsch W, Boltes M (2005) The fundamental diagram of pedestrian movement revisited. J Stat Mech P10002
175. Seyfried A, Steffen B, Lippert T (2006) Basics of modelling the pedestrian flow. Physica A 368:232–238
176. Seyfried A, Rupprecht T, Passon O, Steffen B, Klingsch W, Boltes M (2007) Capacity estimation for emergency exits and bootlenecks. In: Interflam 2007 – Conference Proceedings. Interscience Communication Ltd, Greenwich, London
177. Shestopal VO, Grubits SJ (1994) Evacuation model for merging traffic flows in multi-room and multi-storey buildings. In: Kashiwagi T (ed) Fire Safety Science – 4th international Symposium Proceedings. Interscience, West Yard House, Guildford. The International Association for Fire Safety Science. Grove, London, pp 625–632
178. Sime JD (1990) The Concept of Panic. In: Canter D (ed) Fires and Human Behaviour, vol 1. Wiley, London, pp 63–81
179. Smelser NJ (1962) Theory of Collective Behavior. Free Press, New York
180. Still KG (2001) Crowd Dynamics, Ph.D. thesis, University of Warwick
181. Tajima Y, Nagatani T (2002) Clogging transition of pedestrian flow in t-shaped channel. Physica A 303:239–250
182. Taylor PM (1990) The Hillsborough Stadium Disaster: Inquiry Final Report. Technical report, Great Britain Home Office
183. Templer J (1992) The Staircase. MIT Press, Cambridge
184. Thompson PA, Marchant EW (1994) Simulex; developing new computer modelling techniques for evaluation. In: Kashiwagi T (ed) Fire Safety Science – 4th international Symposium Proceedings. Interscience, West Yard House, Guildford. The International Association for Fire Safety Science. Grove, London, pp 613–624
185. Togawa K (1955) Study on fire escapes basing on the observation of multitude currents. Report of the building research institute. Ministry of Construction, Japan (in Japanese)
186. Tsuji Y (2003) Numerical simulation of pedestrian flow at high densities. In: Galea ER (ed) Pedestrian and Evacuation Dynamics 2003. CMS Press, London, p 27
187. Tubbs JS, Meacham B (2007) Egress Design Solutions – A Guide to Evacuation and Crowd Management Planning. Wiley, New Jersey
188. Virkler MR, Elayadath S (1994) Pedestrian density characteristics and shockwaves. In: Akcelik R (ed) Proceedings of the Second International Symposium on Highway Capacity, vol 2. Australian Road Research Board, Sydney, pp 671–684
189. Waldau N (2002) Massenpanik in Gebäuden. Diploma thesis, Technische Universität Wien
190. Waldau N, Gattermann P, Knoflacher H, Schreckenberg M (eds) (2006) Pedestrian and Evacuation Dynamics 2005. Springer, Berlin
191. Weckman LS, Mannikkö S (1999) Evacuation of a theatre: Exercise vs calculations. Fire Mater 23:357–361
192. Weidmann U (1993) Transporttechnik der Fußgänger – Transporttechnische Eigenschaften des Fußgängerverkehrs (Literaturauswertung). Schriftenreihe des IVT 90, ETH Zürich, 3 1993. Zweite, ergänzte Auflage (in German)
193. Weifeng F, Lizhong Y, Weicheng F (2003) Simulation of bidirectional pedestrian movement using a cellular automata model. Physica A 321:633–640
194. Wolf DE, Grassberger P (eds) (1996) Friction, Arching, Contact Dynamics. World Scientific, Singapore
195. Yamamoto K, Kokubo S, Nishinari K (2006) New approach for pedestrian dynamics by real-coded cellular automata (rca). In: El Yacoubi S, Chopard B, Bandini S (eds) Cellular Automata – 7th International Conference on Cellular Automata for Research and Industry, ACRI 2006, Perpignan. Springer, Berlin, pp 728–731
196. Yamamoto K, Kokubo S, Nishinari K (2007) Simulation for pedestrian dynamics by real-coded cellular automata (rca). Physica A 379:654
197. Yamori K (1998) Going with the flow: Micro-macro dynamics in the macrobehavioral patterns of pedestrian crowds. Psychol Rev 105(3):530–557

Books and Reviews

Chopard B, Droz M (1998) Cellular automaton modeling of physical systems. Cambridge University Press, Cambridge

Chowdhury D, Nishinari K, Santen L, Schadschneider A (2008) Stochastic transport in complex systems: From molecules to vehicles. Elsevier, Amsterdam

DiNenno PJ (ed) (2002) SFPE Handbook of Fire Protection Engineering. National Fire Protection Association, Quincy

Galea ER (ed) (2003) Pedestrian and Evacuation Dynamics '03. CMS Press, London

Ped-Net collaboration. Webpage www.ped-net.org (including discussion forum)

Predtechenskii VM, Milinskii AI (1978) Planing for foot traffic flow in buildings. Amerint Publishing, New Delhi

Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) (2007) Traffic and Granular Flow '05. Springer, Berlin (see also previous issues of this conference series)

Tubbs JS, Meacham BJ (2007) Egress Design Solution – A Guide to Evacuation and Crowd Management Planning. Wiley, New Jersey

Waldau N, Gattermann P, Knoflacher H, Schreckenberg M (eds) (2007) Pedestrian and Evacuation Dynamics '05. Springer, Berlin

Evolutionary Game Theory

William H. Sandholm
Department of Economics, University of Wisconsin, Madison, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Normal Form Games
Static Notions of Evolutionary Stability
Population Games
Revision Protocols
Deterministic Dynamics
Stochastic Dynamics
Local Interaction
Applications
Future Directions
Acknowledgments
Bibliography

Glossary

Deterministic evolutionary dynamic A deterministic evolutionary dynamic is a rule for assigning population games to ordinary differential equations describing the evolution of behavior in the game. Deterministic evolutionary dynamics can be derived from *revision protocols*, which describe choices (in economic settings) or births and deaths (in biological settings) on an agent-by-agent basis.

Evolutionarily stable strategy (ESS) In a symmetric normal form game, an evolutionarily stable strategy is a (possibly mixed) strategy with the following property: a population in which all members play this strategy is resistant to invasion by a small group of mutants who play an alternative mixed strategy.

Normal form game A normal form game is a strategic interaction in which each of n players chooses a strategy and then receives a payoff that depends on all agents' choices choices of strategy. In a *symmetric two-player normal form game*, the two players choose from the same set of strategies, and payoffs only depend on own and opponent's choices, not on a player's identity.

Population game A population game is a strategic interaction among one or more large populations of agents. Each agent's payoff depends on his own choice of strategy and the distribution of others' choices of strategies. One can generate a population game from a normal form game by introducing random matching; however, many population games of interest, including congestion games, do not take this form.

Replicator dynamic The replicator dynamic is a fundamental deterministic evolutionary dynamic for games. Under this dynamic, the percentage growth rate of the mass of agents using each strategy is proportional to the excess of the strategy's payoff over the population's average payoff. The replicator dynamic can be interpreted biologically as a model of natural selection, and economically as a model of imitation.

Revision protocol A revision protocol describes both the timing and the results of agents' decisions about how to behave in a repeated strategic interaction. Revision protocols are used to derive both deterministic and stochastic evolutionary dynamics for games.

Stochastically stable state In Game-theoretic models of stochastic evolution in games are often described by irreducible Markov processes. In these models, a population state is stochastically stable if it retains positive weight in the process's stationary distribution as the level of noise in agents' choices approaches zero, or as the population size approaches infinity.

Definition of the Subject

Evolutionary game theory studies the behavior of large populations of agents who repeatedly engage in strategic interactions. Changes in behavior in these populations are driven either by natural selection via differences in birth and death rates, or by the application of myopic decision rules by individual agents.

The birth of evolutionary game theory is marked by the publication of a series of papers by mathematical biologist John Maynard Smith [137,138,140]. Maynard Smith adapted the methods of traditional game theory [151,215], which were created to model the behavior of rational economic agents, to the context of biological natural selection. He proposed his notion of an *evolutionarily stable strategy* (*ESS*) as a way of explaining the existence of ritualized animal conflict.

Maynard Smith's equilibrium concept was provided with an explicit dynamic foundation through a differential

equation model introduced by Taylor and Jonker [205]. Schuster and Sigmund [189], following Dawkins [58], dubbed this model the *replicator dynamic*, and recognized the close links between this game-theoretic dynamic and dynamics studied much earlier in population ecology [132,214] and population genetics [73]. By the 1980s, evolutionary game theory was a well-developed and firmly established modeling framework in biology [106].

Towards the end of this period, economists realized the value of the evolutionary approach to game theory in social science contexts, both as a method of providing foundations for the equilibrium concepts of traditional game theory, and as a tool for selecting among equilibria in games that admit more than one. Especially in its early stages, work by economists in evolutionary game theory hewed closely to the interpretation set out by biologists, with the notion of ESS and the replicator dynamic understood as modeling natural selection in populations of agents genetically programmed to behave in specific ways. But it soon became clear that models of essentially the same form could be used to study the behavior of populations of active decision makers [50,76,133,149,167,191]. Indeed, the two approaches sometimes lead to identical models: the replicator dynamic itself can be understood not only as a model of natural selection, but also as one of imitation of successful opponents [35,188,216].

While the majority of work in evolutionary game theory has been undertaken by biologists and economists, closely related models have been applied to questions in a variety of fields, including transportation science [143,150,173,175,177,197], computer science [72,173,177], and sociology [34,62,126,225,226]. Some paradigms from evolutionary game theory are close relatives of certain models from physics, and so have attracted the attention of workers in this field [141,201,202,203]. All told, evolutionary game theory provides a common ground for workers from a wide range of disciplines.

Introduction

This article offers a broad survey of the theory of evolution in games. Section "Normal Form Games" introduces normal form games, a simple and commonly studied model of strategic interaction. Section "Static Notions of Evolutionary Stability" presents the notion of an evolutionarily stable strategy, a static definition of stability proposed for this normal form context.

Section "Population Games" defines population games, a general model of strategic interaction in large populations. Section "Revision Protocols" offers the notion of a revision protocol, an individual-level description of behavior used to define the population-level processes of central concern.

Most of the article concentrates on these population-level processes: Section "Deterministic Dynamics" considers deterministic differential equation models of game dynamics; Section "Stochastic Dynamics" studies stochastic models of evolution based on Markov processes; and Sect. "Local Interaction" presents deterministic and stochastic models of local interaction. Section "Applications" records a range of applications of evolutionary game theory, and Sect. "Future Directions" suggests directions for future research. Finally, Sect. "Bibliography" offers an extensive list of primary references.

Normal Form Games

In this section, we introduce a very simple model of strategic interaction: the symmetric two-player normal form game. We then define some of the standard solution concepts used to analyze this model, and provide some examples of games and their equilibria. With this background in place, we turn in subsequent sections to evolutionary analysis of behavior in games.

In a *symmetric two-player normal form game*, each of the two players chooses a (pure) strategy from the finite set S, which we write generically as $S = \{1, \ldots, n\}$. The game's *payoffs* are described by the matrix $A \in \mathbf{R}^{n \times n}$. Entry A_{ij} is the payoff a player obtains when he chooses strategy i and his opponent chooses strategy j; this payoff does not depend on whether the player in question is called player 1 or player 2.

The fundamental solution concept of noncooperative game theory is Nash equilibrium [151]. We say that the pure strategy $i \in S$ is a *symmetric Nash equilibrium* of A if

$$A_{ii} \geq A_{ji} \quad \text{for all } j \in S. \tag{1}$$

Thus, if his opponent chooses a symmetric Nash equilibrium strategy i, a player can do no better than to choose i himself.

A stronger requirement on strategy i demands that it be superior to all other strategies regardless of the opponent's choice:

$$A_{ik} > A_{jk} \quad \text{for all } j, k \in S. \tag{2}$$

When condition (2) holds, we say that strategy i is *strictly dominant* in A.

Example 1 The game below, with strategies C ("cooperate") and D ("defect"), is an instance of a *Prisoner's*

Dilemma:

$$\begin{array}{c|cc} & C & D \\ \hline C & 2 & 0 \\ D & 3 & 1 \end{array}.$$

(To interpret this game, note that $A_{CD} = 0$ is the payoff to cooperating when one's opponent defects.) Since $1 > 0$, defecting is a symmetric Nash equilibrium of this game. In fact, since $3 > 2$ and $1 > 0$, defecting is even a strictly dominant strategy. But since $2 > 1$, both players are better off when both cooperate than when both defect.

In many instances, it is natural to allow players to choose *mixed* (or *randomized*) *strategies*. When a player chooses mixed strategy from the simplex $X = \{x \in \mathbf{R}^n_+ : \sum_{i \in S} x_i = 1\}$, his behavior is stochastic: he commits to playing pure strategy $i \in S$ with probability x_i.

When either player makes a randomized choice, we evaluate payoffs by taking expectations: a player choosing mixed strategy x against an opponent choosing mixed strategy y garners an expected payoff of

$$x'Ay = \sum_{i \in S} \sum_{j \in S} x_i A_{ij} y_j. \tag{3}$$

In biological contexts, payoffs are *fitnesses*, and represent levels of reproductive success relative to some baseline level; Eq. (3) reflects the idea that in a large population, expected reproductive success is what matters. In economic contexts, payoffs are *utilities*: a numerical representation of players' preferences under which Eq. (3) captures players' choices between uncertain outcomes [215].

The notion of Nash equilibrium extends easily to allow for mixed strategies. Mixed strategy x is a *symmetric Nash equilibrium* of A if

$$x'Ax \geq y'Ax \quad \text{for all } y \in X. \tag{4}$$

In words, x is a symmetric Nash equilibrium if its expected payoff against itself is at least as high as the expected payoff obtainable by any other strategy y against x. Note that we can represent the pure strategy $i \in S$ using the mixed strategy $e_i \in X$, the ith standard basis vector in $\mathbf{R}^n$. If we do so, then definition (4) restricted to such strategies is equivalent to definition (1).

We illustrate these ideas with a few examples.

Example 2 Consider the *Stag Hunt* game:

$$\begin{array}{c|cc} & H & S \\ \hline H & h & h \\ S & 0 & s \end{array}.$$

Each player in the Stag Hunt game chooses between hunting hare (H) and hunting stag (S). A player who hunts hare always catches one, obtaining a payoff of $h > 0$. But hunting stag is only successful if both players do so, in which case each obtains a payoff of $s > h$. Hunting stag is potentially more profitable than hunting hare, but requires a coordinated effort.

In the Stag Hunt game, H and S (or, equivalently, e_H and e_S) are symmetric pure Nash equilibria. This game also has a symmetric mixed Nash equilibrium, namely $x^* = (x^*_H, x^*_S) = (\frac{s-h}{s}, \frac{h}{s})$. If a player's opponent chooses this mixed strategy, the player's expected payoff is h whether he chooses H, S, or any mixture between the two; in particular, x^* is a best response against itself.

To distinguish between the two pure equilibria, we might focus on the one that is *payoff dominant*, in that it achieves the higher joint payoff. Alternatively, we can concentrate on the *risk dominant* equilibrium [89], which utilizes the strategy preferred by a player who thinks his opponent is equally likely to choose either option (that is, against an opponent playing mixed strategy $(x_H, x_S) = (\frac{1}{2}, \frac{1}{2})$). In the present case, since $s > h$, equilibrium S is payoff dominant. Which strategy is risk dominant depends on further information about payoffs. If $s > 2h$, then S is risk dominant. But if $s < 2h$, H is risk dominant: evidently, payoff dominance and risk dominance need not agree.

Example 3 In the *Hawk–Dove game* [139], the two players are animals contesting a resource of value $v > 0$. The players choose between two strategies: display (D) or escalate (E). If both display, the resource is split; if one escalates and the other displays, the escalator claims the entire resource; if both escalate, then each player is equally likely to claim the entire resource or to be injured, suffering a cost of $c > v$ in the latter case.

The payoff matrix for the Hawk–Dove game is therefore

$$\begin{array}{c|cc} & D & E \\ \hline D & \frac{1}{2}v & 0 \\ E & v & \frac{1}{2}(v-c) \end{array}.$$

This game has no symmetric Nash equilibrium in pure strategies. It does, however, admit the symmetric mixed equilibrium $x^* = (x^*_D, x^*_E) = (\frac{c-v}{c}, \frac{v}{c})$. (In fact, it can be shown that every symmetric normal form game admits at least one symmetric mixed Nash equilibrium [151].)

In this example, our focus on symmetric behavior may seem odd: rather than randomizing symmetrically, it seems more natural for players to follow an asymmetric Nash equilibrium in which one player escalates and the

other displays. But the symmetric equilibrium is the most relevant one for understanding natural selection in populations whose members are randomly matched in pairwise contests –;see Sect. "Static Notions of Evolutionary Stability".

Example 4 Consider the class of *Rock–Paper–Scissors* games:

$$\begin{array}{c|ccc} & R & P & S \\ \hline R & 0 & -l & w \\ P & w & 0 & -l \\ S & -l & w & 0 \end{array} .$$

Here $w > 0$ is the benefit of winning the match and $l > 0$ the cost of losing; ties are worth 0 to both players. We call this game *good RPS* if $w > l$, so that the benefit of winning the match exceeds the cost of losing, *standard RPS* if $w = l$, and *bad RPS* if $w < l$. Regardless of the values of w and l, the unique symmetric Nash equilibrium of this game, $x^* = (x_R^*, x_P^*, x_S^*) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, requires uniform randomization over the three strategies.

Static Notions of Evolutionary Stability

In introducing game-theoretic ideas to the study of animal behavior, Maynard Smith advanced this fundamental principle: that the evolutionary success of (the genes underlying) a given behavioral trait can depend on the prevalences of all traits. It follows that natural selection among the traits can be modeled as random matching of animals to play normal form games [137,138,139,140]. Working in this vein, Maynard Smith offered a stability concept for populations of animals sharing a common behavioral trait – that of playing a particular mixed strategy in the game at hand. Maynard Smith's concept of evolutionary stability, influenced by the work of Hamilton [87] on the evolution of sex ratios, defines such a population as stable if it is resistant to invasion by a small group of mutants carrying a different trait.

Suppose that a large population of animals is randomly matched to play the symmetric normal form game A. We call mixed strategy $x \in X$ an *evolutionarily stable strategy* (*ESS*) if

$$x'A((1-\varepsilon)x + \varepsilon y) > y'A((1-\varepsilon)x + \varepsilon y) \quad \text{for all } \varepsilon \le \bar{\varepsilon}(y) \text{ and } y \neq x . \tag{5}$$

To interpret condition (5), imagine that a population of animals programmed to play mixed strategy x is invaded by a group of mutants programmed to play the alternative mixed strategy y. Equation (5) requires that regardless of the choice of y, an incumbent's expected payoff from a random match in the post-entry population exceeds that of a mutant so long as the size of the invading group is sufficiently small.

The definition of ESS above can also be expressed as a combination of two conditions:

$$x'Ax \ge y'Ax \quad \text{for all } y \in X; \tag{4}$$

$$\text{For all } y \neq x,\ [x'Ax = y'Ax] \text{ implies that } [x'Ay > y'Ay] . \tag{6}$$

Condition (4) is familiar: it requires that the incumbent strategy x be a best response to itself, and so is none other than our definition of symmetric Nash equilibrium. Condition (6) requires that if a mutant strategy y is an alternative best response against the incumbent strategy x, then the incumbent earns a higher payoff against the mutant than the mutant earns against itself.

A less demanding notion of stability can be obtained by allowing the incumbent and the mutant in condition (6) to perform equally well against the mutant:

$$\text{For all } y \in X,\ [x'Ax = y'Ax] \text{ implies that } [x'Ay \ge y'Ay] . \tag{7}$$

If x satisfies conditions (4) and (7), it is called a *neutrally stable strategy* (*NSS*) [139].

Let us apply these stability notions to the games introduced in the previous section. Since every ESS and NSS must be a Nash equilibrium, we need only consider whether the Nash equilibria of these games satisfy the additional stability conditions, (6) and (7).

Example 5 In the Prisoner's Dilemma game (Example 1), the dominant strategy D is an ESS.

Example 6 In the Stag Hunt game (Example 2), each pure Nash equilibrium is an ESS. But the mixed equilibrium $(x_H^*, x_S^*) = (\frac{s-h}{s}, \frac{h}{s})$ is not an ESS: if mutants playing either pure strategy enter the population, they earn a higher payoff than the incumbents in the post-entry population.

Example 7 In the Hawk–Dove game (Example 3), the mixed equilibrium $(x_D^*, x_E^*) = (\frac{c-v}{c}, \frac{v}{c})$ is an ESS. Maynard Smith used this and other examples to explain the existence of ritualized fighting in animals. While an animal who escalates always obtains the resource when matched with an animal who merely displays, a population of escalators is unstable: it can be invaded by a group of mutants who display, or who merely escalate less often.

Example 8 In Rock–Paper–Scissors games (Example 4), whether the mixed equilibrium $x^* = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ is evolutionarily stable depends on the relative payoffs to winning

and losing a match. In good RPS ($w > l$), x^* is an ESS; in standard RPS ($w = l$), x^* is a NSS but not an ESS, while in bad RPS ($w < l$), x^* is neither an ESS nor an NSS. The last case shows that neither evolutionary nor neutrally stable strategies need exist in a given game.

The definition of an evolutionarily stable strategy has been extended to cover a wide range of strategic settings, and has been generalized in a variety of directions. Prominent among these developments are set-valued versions of ESS: in rough terms, these concepts consider a set of mixed strategies $Y \subset X$ to be stable if the no population playing a strategy in the set can be invaded successfully by a population of mutants playing a strategy outside the set. [95] provides a thorough survey of the first 15 years of research on ESS and related notions of stability; key references on set-valued evolutionary solution concepts include [15,199,206].

Maynard Smith's notion of ESS attempts to capture the dynamic process of natural selection using a static definition. The advantage of this approach is that his definition is often easy to check in applications. Still, more convincing models of natural selection should be explicitly dynamic models, building on techniques from the theories of dynamical systems and stochastic processes. Indeed, this thoroughgoing approach can help us understand whether and when the ESS concept captures the notion of robustness to invasion in a satisfactory way.

The remainder of this article concerns explicitly dynamic models of behavior. In addition to being dynamic rather than static, these models will differ from the one considered in this section in two other important ways as well. First, rather than looking at populations whose members all play a particular mixed strategy, the dynamic models consider populations in which different members play different pure strategies. Second, instead of maintaining a purely biological point of view, our dynamic models will be equally well-suited to studying behavior in animal and human populations.

Population Games

Population games provide a simple and general framework for studying strategic interactions in large populations whose members play pure strategies. The simplest population games are generated by random matching in normal form games, but the population game framework allows for interactions of a more intricate nature.

We focus here on games played by a single population (i. e., games in which all agents play equivalent roles). We suppose that there is a unit mass of agents, each of whom chooses a pure strategy from the set $S = \{1, \ldots, n\}$. The aggregate behavior of these agents is described by a *population state* $x \in X$, with x_j representing the proportion of agents choosing pure strategy j. We identify a *population game* with a continuous vector-valued payoff function $F\colon X \to \mathbf{R}^n$. The scalar $F_i(x)$ represents the payoff to strategy i when the population state is x.

Population state x^* is a *Nash equilibrium* of F if no agent can improve his payoff by unilaterally switching strategies. More explicitly, x^* is a Nash equilibrium if

$$x_i^* > 0 \quad \text{implies that } F_i(x) \geq F_j(x) \quad \text{for all } j \in S. \tag{8}$$

Example 9 Suppose that the unit mass of agents are randomly matched to play the symmetric normal form game A. At population state x, the (expected) payoff to strategy i is the linear function $F_i(x) = \sum_{j \in S} A_{ij} x_j$; the payoffs to all strategies can be expressed concisely as $F(x) = Ax$. It is easy to verify that x^* is a Nash equilibrium of the population game F if and only if x^* is a symmetric Nash equilibrium of the symmetric normal form game A.

While population games generated by random matching are especially simple, many games that arise in applications are not of this form. In the biology literature, games outside the random matching paradigm are known as *playing the field* models [139].

Example 10 Consider the following model of highway congestion [17,143,166,173]. A pair of towns, Home and Work, are connected by a network of *links*. To commute from Home to Work, an agent must choose a *path* $i \in S$ connecting the two towns. The payoff the agent obtains is the negation of the delay on the path he takes. The delay on the path is the sum of the delays on its constituent links, while the delay on a link is a function of the number of agents who use that link.

Population games embodying this description are known as a *congestion games*. To define a congestion game, let Φ be the collection of links in the highway network. Each strategy $i \in S$ is a route from Home to Work, and so is identified with a set of links $\Phi_i \subseteq \Phi$. Each link ϕ is assigned a *cost function* $c_\phi\colon \mathbf{R}_+ \to \mathbf{R}$, whose argument is link ϕ's *utilization level* u_ϕ:

$$u_\phi(x) = \sum_{i \in \rho(\phi)} x_i\,, \quad \text{where } \rho(\phi) = \{i \in S\colon \phi \in \Phi_i\}$$

The payoff of choosing route i is the negation of the total delays on the links in this route:

$$F_i(x) = -\sum_{\phi \in \Phi_i} c_\phi(u_\phi(x))\,.$$

Since driving on a link increases the delays experienced by other drivers on that link (i. e., since highway congestion involves *negative externalities*), cost functions in models of highway congestion are increasing; they are typically convex as well. Congestion games can also be used to model positive externalities, like the choice between different technological standards; in this case, the cost functions are decreasing in the utilization levels.

Revision Protocols

We now introduce foundations for our models of evolutionary dynamics. These foundations are built on the notion of a revision protocol, which describes both the timing and results of agents' myopic decisions about how to continue playing the game at hand [24,35,96,175,217]. Revision protocols will be used to derive both the deterministic dynamics studied in Sect. "Deterministic Dynamics" and the stochastic dynamics studied in Sect. "Stochastic Dynamics"; similar ideas underlie the local interaction models introduced in Sect. "Local Interaction".

Definition

Formally, a *revision protocol* is a map $\rho\colon \mathbf{R}^n \times X \to \mathbf{R}_+^{n\times n}$ that takes the payoff vectors π and population states x as arguments, and returns nonnegative matrices as outputs. For reasons to be made clear below, scalar $\rho_{ij}(\pi, x)$ is called the *conditional switch rate* from strategy i to strategy j.

To move from this notion to an explicit model of evolution, let us consider a population consisting of $N < \infty$ members. (A number of the analyzes to follow will consider the limit of the present model as the population size N approaches infinity – see Sects. "Mean Dynamics", "Deterministic Approximation", and "Stochastic Stability via Large Population Limits".) In this case, the set of feasible social states is the finite set $\mathcal{X}^N = X \cap \frac{1}{N}\mathbf{Z}^n = \{x \in X\colon Nx \in \mathbf{Z}^n\}$, a grid embedded in the simplex X.

A revision protocol ρ, a population game F, and a population size N define a continuous-time evolutionary process – a Markov process $\{X_t^N\}$ – on the finite state space $\mathcal{X}^N$. A one-size-fits-all description of this process is as follows. Each agent in the society is equipped with a "stochastic alarm clock". The times between rings of of an agent's clock are independent, each with a rate R exponential distribution. The ringing of a clock signals the arrival of a revision opportunity for the clock's owner. If an agent playing strategy $i \in S$ receives a revision opportunity, he switches to strategy $j \neq i$ with probability ρ_{ij}/R. If a switch occurs, the population state changes accordingly, from the old state x to a new state y that accounts for the agent's change in strategy.

While this interpretation of the evolutionary process can be applied to any revision protocol, simpler interpretations are sometimes available for protocols with additional structure. The examples to follow illustrate this point.

Examples

Imitation Protocols and Natural Selection Protocols In economic contexts, revision protocols of the form

$$\rho_{ij}(\pi, x) = x_j \hat{\rho}_{ij}(\pi, x) \tag{9}$$

are called *imitation protocols* [35,96,216]. These protocols can be given a very simple interpretation: when an agent receives a revision opportunity, he chooses an opponent at random and observes her strategy. If our agent is playing strategy i and the opponent strategy j, the agent switches from i to j with probability proportional to $\hat{\rho}_{ij}$. Notice that the value of the population share x_j is not something the agent need know; this term in (9) accounts for the agent's observing a randomly chosen opponent.

Example 11 Suppose that after selecting an opponent, the agent imitates the opponent only if the opponent's payoff is higher than his own, doing so in this case with probability proportional to the payoff difference:

$$\rho_{ij}(\pi, x) = x_j[\pi_j - \pi_i]_+.$$

This protocol is known as *pairwise proportional imitation* [188].

Protocols of form (9) also appear in biological contexts, [144], [153,158], where in these cases we refer to them as *natural selection protocols*. The biological interpretation of (9) supposes that each agent is programmed to play a single pure strategy. An agent who receives a revision opportunity dies, and is replaced through asexual reproduction. The reproducing agent is a strategy j player with probability $\rho_{ij}(\pi, x) = x_j \hat{\rho}_{ij}(\pi, x)$, which is proportional both to the number of strategy j players and to some function of the prevalences and fitnesses of all strategies. Note that this interpretation requires the restriction

$$\sum_{j\in S} \rho_{ij}(\pi, x) \equiv 1.$$

Example 12 Suppose that payoffs are always positive, and let

$$\rho_{ij}(\pi, x) = \frac{x_j\, \pi_j}{\sum_{k\in S} x_k\, \pi_k}. \tag{10}$$

Understood as a natural selection protocol, (10) says that the probability that the reproducing agent is a strategy j player is proportional to $x_j \pi_j$, the aggregate fitness of strategy j players.

In economic contexts, we can interpret (10) as an imitative protocol based on repeated sampling. When an agent's clock rings he chooses an opponent at random. If the opponent is playing strategy j, the agent imitates him with probability proportional to π_j. If the agent does not imitate this opponent, he draws a new opponent at random and repeats the procedure.

Direct Evaluation Protocols In the previous examples, only strategies currently in use have any chance of being chosen by a revising agent (or of being the programmed strategy of the newborn agent). Under other protocols, agents' choices are not mediated through the population's current behavior, except indirectly via the effect of behavior on payoffs. These *direct evaluation protocols* require agents to directly evaluate the payoffs of the strategies they consider, rather than to indirectly evaluate them as under an imitative procedure.

Example 13 Suppose that choices are made according to the *logit choice rule*:

$$\rho_{ij}(\pi, x) = \frac{\exp(\eta^{-1}\pi_j)}{\sum_{k\in S}\exp(\eta^{-1}\pi_k)}. \tag{11}$$

The interpretation of this protocol is simple. Revision opportunities arrive at unit rate. When an opportunity is received by an i player, he switches to strategy j with probability $\rho_{ij}(\pi, x)$, which is proportional to an exponential function of strategy j's payoffs. The parameter $\eta > 0$ is called the *noise level*. If η is large, choice probabilities under the logit rule are nearly uniform. But if η is near zero, choices are optimal with probability close to one, at least when the difference between the best and second best payoff is not too small.

Additional examples of revision protocols can be found in the next section, and one can construct new revision protocols by taking linear combinations of old ones; see [183] for further discussion.

Deterministic Dynamics

Although antecedents of this approach date back to the early work of Brown and von Neumann [45], the use of differential equations to model evolution in games took root with the introduction of the replicator dynamic by Taylor and Jonker [205], and remains an vibrant area of research; Hofbauer and Sigmund [108] and Sandholm [183] offer recent surveys. In this section, we derive a deterministic model of evolution: the *mean dynamic* generated by a revision protocol and a population game. We study this deterministic model from various angles, focusing in particular on local stability of rest points, global convergence to equilibrium, and nonconvergent limit behavior.

While the bulk of the literature on deterministic evolutionary dynamics is consistent with the approach we take here, we should mention that other specifications exist, including discrete time dynamics [5,59,131,218], and dynamics for games with continuous strategy sets [41,42,77,100,159,160] and for Bayesian population games [62,70,179]. Also, deterministic dynamics for extensive form games introduce new conceptual issues; see [28,30,51,53,55] and the monograph of Cressman [54].

Mean Dynamics

As described earlier in Sect. "Definition", a revision protocol ρ, a population game F, and a population size N define a Markov process $\{X_t^N\}$ on the finite state space $\mathcal{X}^N$. We now derive a deterministic process – the *mean dynamic* – that describes the expected motion of $\{X_t^N\}$. In Sect. "Deterministic Approximation", we will describe formally the sense in which this deterministic process provides a very good approximation of the behavior of the stochastic process $\{X_t^N\}$, at least over finite time horizons and for large population sizes. But having noted this result, we will focus in this section on the deterministic process itself.

To compute the expected increment of $\{X_t^N\}$ over the next $\mathrm{d}t$ time units, recall first that each of the N agents receives revision opportunities via a rate R exponential distribution, and so expects to receive $R\mathrm{d}t$ opportunities during the next $\mathrm{d}t$ time units. If the current state is x, the expected number of revision opportunities received by agents currently playing strategy i is approximately $Nx_i\, R\mathrm{d}t$. Since an i player who receives a revision opportunity switches to strategy j with probability ρ_{ij}/R, the expected number of such switches during the next $\mathrm{d}t$ time units is approximately $Nx_i\, \rho_{ij}\mathrm{d}t$. Therefore, the expected change in the number of agents choosing strategy i during the next $\mathrm{d}t$ time units is approximately

$$N\left(\sum_{j\in S} x_j\rho_{ji}(F(x), x) - x_i \sum_{j\in S}\rho_{ij}(F(x), x)\right)\mathrm{d}t. \tag{12}$$

Dividing expression (12) by N and eliminating the time differential $\mathrm{d}t$ yields a differential equation for the rate of

change in the *proportion* of agents choosing strategy i:

$$\dot{x}_i = \sum_{j \in S} x_j \rho_{ji}(F(x), x) - x_i \sum_{j \in S} \rho_{ij}(F(x), x). \tag{M}$$

Equation (M) is the *mean dynamic* (or *mean field*) generated by revision protocol ρ in population game F. The first term in (M) captures the inflow of agents to strategy i from other strategies, while the second captures the outflow of agents to other strategies from strategy i.

Examples

We now describe some examples of mean dynamics, starting with ones generated by the revision protocols from Sect. "Examples". To do so, we let

$$\overline{F}(x) = \sum_{i \in S} x_i F_i(x)$$

denote the *average payoff* obtained by the members of the population, and define the *excess payoff* to strategy i,

$$\hat{F}_i(x) = F_i(x) - \overline{F}(x)\,,$$

to be the difference between strategy i's payoff and the population's average payoff.

Example 14 In Example 11, we introduced the pairwise proportional imitation protocol $\rho_{ij}(\pi, x) = x_j[\pi_j - \pi_i]_+$. This protocol generates the mean dynamic

$$\dot{x}_i = x_i \hat{F}_i(x)\,. \tag{13}$$

Equation (13) is the *replicator dynamic* [205], the best-known dynamic in evolutionary game theory. Under this dynamic, the percentage growth rate $\dot{x}_i/x_i$ of each strategy currently in use is equal to that strategy's current excess payoff; unused strategies always remain so. There are a variety of revision protocols other than pairwise proportional imitation that generate the replicator dynamic as their mean dynamics; see [35,96,108,217].

Example 15 In Example 12, we assumed that payoffs are always positive, and introduced the protocol $\rho_{ij} \propto x_j \pi_j$, which we interpreted both as a model of biological natural selection and as a model of imitation with repeated sampling. The resulting mean dynamic,

$$\dot{x}_i = \frac{x_i F_i(x)}{\sum_{k \in S} x_k F_k(x)} - x_i = \frac{x_i \hat{F}_i(x)}{\overline{F}(x)}, \tag{14}$$

is the *Maynard Smith replicator dynamic* [139]. This dynamic only differs from the standard replicator dynamic (13) by a change of speed, with motion under (14) being relatively fast when average payoffs are relatively low. (In multipopulation models, the two dynamics are less similar, and convergence under one does not imply convergence under the other – see [183,216].)

Example 16 In Example 13 we introduced the logit choice rule $\rho_{ij}(\pi, x) \propto \exp(\eta^{-1}\pi_j)$. The corresponding mean dynamic,

$$\dot{x}_i = \frac{\exp(\eta^{-1} F_i(x))}{\sum_{k \in S} \exp(\eta^{-1} F_k(x))} - x_i, \tag{15}$$

is called the *logit dynamic* [82].

If we take the noise level η to zero, then the probability with which a revising agent chooses the best response approaches one whenever the best response is unique. At such points, the logit dynamic approaches the *best response dynamic* [84]:

$$\dot{x} \in B^F(x) - x, \tag{16}$$

where

$$B^F(x) = \operatorname{argmax}_{y \in X}\, y' F(x)$$

defines the (*mixed*) *best response correspondence* for game F. Note that unlike the other dynamics we consider here, (16) is defined not by an ordinary differential equation, but by a differential inclusion, a formulation proposed in [97].

Example 17 Consider the protocol

$$\rho_{ij}(\pi, x) = \left[\pi_j - \sum\nolimits_{k \in S} x_k \pi_k\right]_+ .$$

When an agent's clock rings, he chooses a strategy at random; if that strategy's payoff is above average, the agent switches to it with probability proportional to its excess payoff. The resulting mean dynamic,

$$\dot{x}_i BM = [\hat{F}_i(x)]_+ - x_i \sum_{k \in S} [\hat{F}_k(x)]_+,$$

is called the *Brown–von Neumann–Nash* (*BNN*) *dynamic* [45]; see also [98,176,194,200,217].

Example 18

Consider the revision protocol

$$\rho_{ij}(\pi, x) = [\pi_j - \pi_i]_+ .$$

When an agent's clock rings, he selects a strategy at random. If the new strategy's payoff is higher than his cur-

Evolutionary Game Theory, Table 1
Five basic deterministic dynamics

Revision Protocol	Mean Dynamic	Name and source
$\rho_{ij} = x_j[\pi_j - \pi_i]_+$	$\dot{x}_i = x_i \hat{F}_i(x)$	Replicator [205]
$\rho_{ij} = \dfrac{\exp(\eta^{-1}\pi_j)}{\sum_{k\in S}\exp(\eta^{-1}\pi_k)}$	$\dot{x}_i = \dfrac{\exp(\eta^{-1}F_i(x))}{\sum_{k\in S}\exp(\eta^{-1}F_k(x))} - x_i$	Logit [82]
$\rho_{ij} = 1_{\{j=\operatorname{argmax}_{k\in S}\pi_k\}}$	$\dot{x} \in B^F(x) - x$	Best response [84]
$\rho_{ij} = \left[\pi_j - \sum_{k\in S} x_k\pi_k\right]_+$	$\dot{x}_i = [\hat{F}_i(x)]_+ - x_i \sum_{j\in S}[\hat{F}_j(x)]_+$	BNN [45]
$\rho_{ij} = [\pi_j - \pi_i]_+$	$\dot{x}_i = \sum_{j\in S} x_j[F_i(x) - F_j(x)]_+ - x_i \sum_{j\in S}[F_j(x) - F_i(x)]_+$	Smith [197]

rent strategy's payoff, he switches strategies with probability proportional to the difference between the two payoffs. The resulting mean dynamic,

$$\dot{x}_i = \sum_{j\in S} x_j[F_i(x) - F_j(x)]_+ - x_i \sum_{j\in S}[F_j(x) - F_i(x)]_+, \tag{17}$$

is called the *Smith dynamic* [197]; see also [178].

We summarize these examples of revision protocols and mean dynamics in Table 1.

Figure 1 presents phase diagrams for the five basic dynamics when the population is randomly matched to play standard Rock–Paper–Scissors (Example 4). In the phase diagrams, colors represent speed of motion: within each diagram, motion is fastest in the red regions and slowest in the blue ones.

The phase diagram of the replicator dynamic reveals closed orbits around the unique Nash equilibrium $x^* = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Since this dynamic is based on imitation (or on reproduction), each face and each vertex of the simplex X is an invariant set: a strategy initially absent from the population will never subsequently appear.

The other four dynamics pictured are based on direct ecaluation, allowing agents to select strategies that are currently unused. In these cases, the Nash equilibrium is the sole rest point, and attracts solutions from all initial conditions. (In the case of the logit dynamic, the rest point happens to coincide with the Nash equilibrium only because of the symmetry of the game; see [101,104].) Under the logit and best response dynamics, solution trajectories quickly change direction and then accelerate when the best response to the population state changes; under the BNN and especially the Smith dynamic, solutions approach the Nash equlibrium in a less angular fashion.

Evolutionary Justification of Nash Equilibrium

One of the goals of evolutionary game theory is to justify the prediction of Nash equilibrium play. For this justification to be convincing, it must be based on a model that makes only mild assumptions about agents' knowledge about one another's behavior. This sentiment can be captured by introducing two desiderata for revision protocols:

(C) *Continuity:* ρ is Lipschitz continuous.

(SD) *Scarcity of data:* ρ_{ij} only depends on π_i, π_j, and x_j.

Continuity (C) asks that revision protocols depend continuously on their inputs, so that small changes in aggregate behavior do not lead to large changes in players' responses. *Scarcity of data* (SD) demands that the conditional switch rate from strategy i to strategy j only depend on the payoffs of these two strategies, so that agents need only know those facts that are most germane to the decision at hand [183]. (The dependence of ρ_{ij} on x_j is included to allow for dynamics based on imitation.) Protocols that respect these two properties do not make unrealistic demands on the amount of information that agents in an evolutionary model possess.

Our two remaining desiderata impose restrictions on mean dynamics $\dot{x} = V^F(x)$, linking the evolution of aggregate behavior to incentives in the underlying game.

(NS) *Nash stationarity:*

$$V^F(x) = \mathbf{0} \quad \text{if and only if } x \in NE(F).$$

(PC) *Positive correlation:*

$$V^F(x) \neq \mathbf{0} \quad \text{implies that } V^F(x)'F(x) > 0.$$

Nash stationarity (NS) is a restriction on stationary states: it asks that the rest points of the mean dynamic be pre-

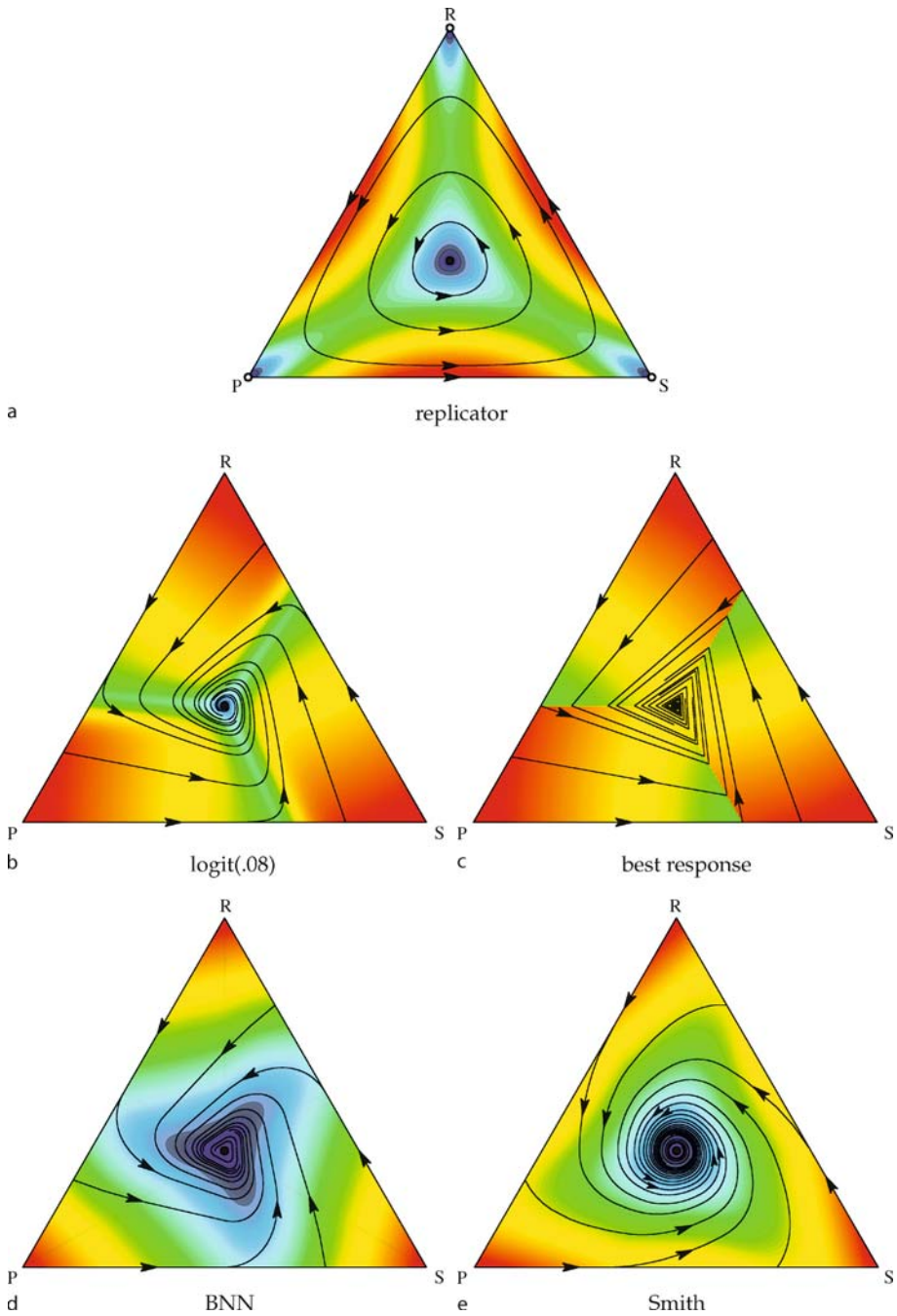

Evolutionary Game Theory, Figure 1
Five basic deterministic dynamics in standard Rock–Paper–Scissors. Colors represent speeds: *red* is fastest, *blue* is slowest

cisely the Nash equilibria of the game being played. *Positive correlation* (PC) is a restriction on disequilibrium adjustment: it requires that away from rest points, strategies' growth rates be positively correlated with their payoffs. Condition (PC) is among the weakest of the many conditions linking growth rates of evolutionary dynamics and payoffs in the underlying game; for alternatives, see [76,110,149,162,170,173,200].

Evolutionary Game Theory, Table 2
Families of deterministic evolutionary dynamics and their properties; yes* indicates that a weaker or alternate form of the property is satisfied

Dynamic	Family	(C)	(SD)	(NS)	(PC)
Replicator	Imitation	yes	yes	no	yes
Best response		no	yes*	yes*	yes*
Logit	Perturbed best response	yes	yes*	no	no
BNN	Excess payoff	yes	no	yes	yes
Smith	Pairwise comparison	yes	yes	yes	yes

In Table 2, we report how the the five basic dynamics fare under the four criteria above. For the purposes of justifying the Nash prediction, the most important row in the table is the last one, which reveals that the Smith dynamic satisfies all four desiderata at once: while the revision protocol for the Smith dynamic (see Example 18) requires only limited information on the part of the agents who employ it, this information is enough to ensure that rest points of the dynamic and Nash equilibria coincide.

In fact, the dynamics introduced above can be viewed as members of families of dynamics that are based on similar revision protocols and that have similar qualitative properties. For instance, the Smith dynamic is a member of the family of *pairwise comparison* dynamics [178], under which agents only switch to strategies that outperform their current choice. For this reason, the exact functional forms of the previous examples are not essential to establishing the properties noted above.

In interpreting these results, it is important to remember that Nash stationarity only concerns the rest points of a dynamic; it says nothing about whether a dynamic will converge to Nash equilibrium from an arbitrary initial state. The question of convergence is addressed in Sects. "Global Convergence" and "Nonconvergence". There we will see that in some classes of games, general guarantees of convergence can be obtained, but that there are some games in which no reasonable dynamic converges to equilibrium.

Local Stability

Before turning to the global behavior of evolutionary dynamics, we address the question of local stability. As we noted at the onset, an original motivation for introducing game dynamics was to provide an explicitly dynamic foundation for Maynard Smith's notion of ESS [205]. Some of the earliest papers on evolutionary game dynamics [105,224] established that being an ESS is a sufficient condition for asymptotically stablity under the replicator dynamic, but that it is not a necessary condition. It is curious that this connection obtains despite the fact that ESS is a stability condition for a population whose members all play the same mixed strategy, while (the usual version of) the replicator dynamic looks at populations of agents choosing among different pure strategies.

In fact, the implications of ESS for local stability are not limited to the replicator dynamic. Suppose that the symmetric normal form game A admits a symmetric Nash equilibrium that places positive probability on each strategy in S. One can show that this equilibrium is an ESS if and only if the payoff matrix A is negative definite with respect to the tangent space of the simplex:

$$z'Az < 0 \quad \text{for all } z \in TX = \left\{\hat{z} \in \mathbf{R}^n : \sum_{i \in S} \hat{z}_i = 0\right\}. \tag{18}$$

Condition (18) and its generalizations imply local stability of equilibrium not only under the replicator dynamic, but also under a wide range of other evolutionary dynamics: see [52,98,99,102,111,179] for further details.

The papers cited above use linearization and Lyapunov function arguments to establish local stability. An alternative approach to local stability analysis, via index theory, allows one to establish restrictions on the stability properties of all rest points at once – see [60].

Global Convergence

While analyses of local stability reveal whether a population will return to equilibrium after a small disturbance, they do not tell us whether the population will approach equilibrium from an arbitrary disequilibrium state. To establish such global convergence results, we must restrict attention to classes of games defined by certain interesting payoff structures. These structures appear in applications, lending strong support for the Nash prediction in the settings where they arise.

Potential Games A *potential game* [17,106,143,166,173,181] is a game that admits a *potential function*: a scalar val-

ued function whose gradient describes the game's payoffs. In a full potential game $F\colon \mathbf{R}^n_+ \to \mathbf{R}^n$ (see [181]), all information about incentives is captured by the potential function $f\colon \mathbf{R}^n_+ \to \mathbf{R}$, in the sense that

$$\nabla f(x) = F(x) \quad \text{for all } x \in \mathbf{R}^n_+. \tag{19}$$

If F is smooth, then it is a full potential game if and only if it satisfies *full externality symmetry*:

$$\frac{\partial F_i}{\partial x_j}(x) = \frac{\partial F_j}{\partial x_i}(x) \quad \text{for all } i, j \in S \text{ and } x \in \mathbf{R}^n_+. \tag{20}$$

That is, the effect on the payoff to strategy i of adding new strategy j players always equals the effect on the payoff to strategy j of adding new strategy i players.

Example 19 Suppose a single population is randomly matched to play the symmetric normal form game $A \in \mathbf{R}^{n\times n}$, generating the population game $F(x) = Ax$. We say that A exhibits *common interests* if the two players in a match always receive the same payoff. This means that $A_{ij} = A_{ji}$ for all i and j, or, equivalently, that the matrix A is symmetric. Since $DF(x) = A$, this is precisely what we need for F to be a full potential game. The full potential function for F is $f(x) = \frac{1}{2}x'Ax$, which is one-half of the average payoff function $\bar{F}(x) = \sum_{i\in S} x_i F_i(x) = x'Ax$. The common interest assumption defines a fundamental model from population genetics, this assumption reflects the shared fate of two genes that inhabit the same organism [73,106,107].

Example 20 In Example 10, we introduced congestion games, a basic model of network congestion. To see that these games are potential games, observe that an agent taking path $j \in S$ affects the payoffs of agents choosing path $i \in S$ through the marginal increases in congestion on the links $\phi \in \Phi_i \cap \Phi_j$ that the two paths have in common. But since the marginal effect of an agent taking path i on the payoffs of agents choosing path j is identical, full externality symmetry (20) holds:

$$\frac{\partial F_i}{\partial x_j}(x) = -\sum_{\phi\in\Phi_i\cap\Phi_j} c'_\phi(u_\phi(x)) = \frac{\partial F_j}{\partial x_i}(x).$$

In congestion games, the potential function takes the form

$$f(x) = -\sum_{\phi\in\Phi} \int_0^{u_\phi(x)} c_\phi(z)\,\mathrm{d}z,$$

and so is typically unrelated to aggregate payoffs,

$$\overline{F}(x) = \sum_{i\in S} x_i F_i(x) = -\sum_{\phi\in\Phi} u_\phi(x)\, c_\phi(u_\phi(x)).$$

However, potential is proportional to aggregate payoffs if the cost functions c_ϕ are all monomials of the same degree [56,173].

Population state x is a Nash equilibrium of the potential game F if and only if it satisfies the Kuhn–Tucker first order conditions for maximizing the potential function f on the simplex X [17,173]. Furthermore, it is simple to verify that any dynamic $\dot{x} = V^F(x)$ satisfying positive correlation (PC) ascends the potential function:

$$\tfrac{\mathrm{d}}{\mathrm{d}t} f(x_t) = \nabla f(x_t)'\dot{x}_t = F(x_t)'V^F(x_t) \geq 0.$$

It then follows from classical results on Lyapunov functions that any dynamic satisfying positive correlation (PC) converges to a connected set of rest points. If the dynamic also satisfies Nash stationarity (NS), these sets consist entirely of Nash equilibria. Thus, in potential games, very mild conditions on agents' adjustment rules are sufficient to justify the prediction of Nash equilibrium play.

In the case of the replicator dynamic, one can say more. On the interior of the simplex X, the replicator dynamic for the potential game F is a *gradient system* for the potential function f (i. e., it always ascends f in the direction of maximum increase). However, this is only true after one introduces an appropriate Riemannian metric on X [123,192]. An equivalent statement of this result, due to [2], is that the replicator dynamic is the gradient system for f under the usual Euclidean metric if we stretch the state space X onto the radius 2 sphere. This stretching is accomplished using the *Akin transformation* $H_i(x) = 2\sqrt{x_i}$, which emphasizes changes in the use of rare strategies relative to changes in the use of common ones [2,4,185]. (There is also a dynamic that generates the gradient system for f on X under the usual metric: the so-called *projection dynamic* [130,150,185].)

Example 21 Consider evolution in *123 Coordination*:

$$\begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & 1 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 3 & 0 & 0 & 3 \end{array}.$$

Figure 2a presents a phase diagram of the replicator dynamic on its natural state space X, drawn atop of a contour plot of the potential function $f(x) = \frac{1}{2}((x_1)^2 + 2(x_2)^2 + 3(x_3)^2)$. Evidently, all solution trajectories ascend this function and converge to one of the seven symmetric Nash equilibria, with trajectories from all but a measure

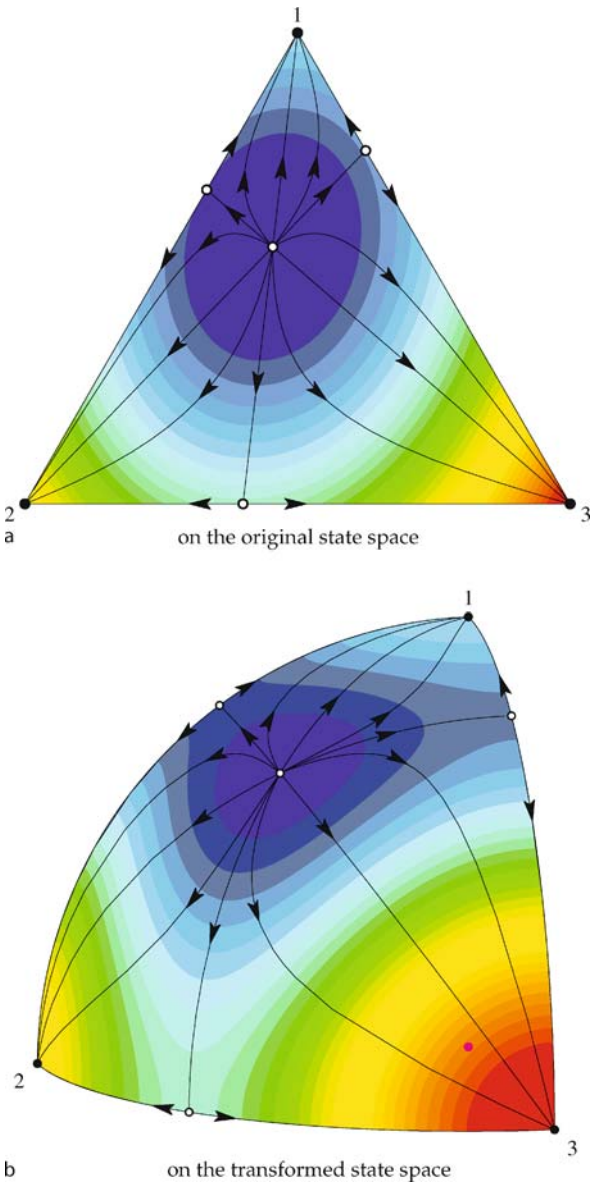

Evolutionary Game Theory, Figure 2
The replicator dynamic in 123 Coordination. Colors represent the value of the game's potential function

zero set of initial conditions converging to one of the three pure equilibria.

Figure 2b presents another phase diagram for the replicator dynamic, this time after the solution trajectories and the potential function have been transported to the surface of the radius 2 sphere using the Akin transformation. In this case, solutions cross the level sets of the potential function orthogonally, moving in the direction that increases potential most quickly.

Stable Games A population game F is a *stable game* [102] if

$$(y - x)'(F(y) - F(x)) \leq 0 \quad \text{for all } x, y \in X. \tag{21}$$

If the inequality in (21) always holds strictly, then F is a *strictly stable game*.

If F is smooth, then F is a stable game if and only if it satisfies *self-defeating externalities*:

$$z' DF(x) z \leq 0 \quad \text{for all } z \in TX \text{ and } x \in X, \tag{22}$$

where $DF(x)$ is the derivative of $F\colon X \to \mathbf{R}^n$ at x. This condition requires that the improvements in the payoffs of strategies to which revising agents are switching are always exceeded by the improvements in the payoffs of strategies which revising agents are abandoning.

Example 22 The symmetric normal form game A is *symmetric zero-sum* if A is skew-symmetric (i. e., if $A = -A'$), so that the payoffs of the matched players always sum to zero. (An example is provided by the standard Rock–Paper–Scissors game (Example 4).) Under this assumption, $z'Az = 0$ for all $z \in \mathbf{R}^n$; thus, the population game generated by random matching in A, $F(x) = Ax$, is a stable game that is not strictly stable.

Example 23 Suppose that A satisfies the interior ESS condition (18). Then (22) holds strictly, so $F(x) = Ax$ is a strictly stable game. Examples satisfying this condition include the Hawk–Dove game (Example 3) and any good Rock–Paper–Scissors game (Example 4).

Example 24 A *war of attrition* [33] is a symmetric normal form game in which strategies represent amounts of time committed to waiting for a scarce resource. If the two players choose times i and $j > i$, then the j player obtains the resource, worth v, while both players pay a cost of c_i: once the first player leaves, the other seizes the resource immediately. If both players choose time i, the resource is split, so payoffs are $\frac{v}{2} - c_i$ each. It can be shown that for any resource value $v \in \mathbf{R}$ and any increasing cost vector $c \in \mathbf{R}^n$, random matching in a war of attrition generates a stable game [102].

The flavor of the self-defeating externalities condition (22) suggests that obedience of incentives will push the population toward some "central" equilibrium state. In fact, the set of Nash equilibria of a stable game is always convex, and in the case of strictly stable games, equilibrium is unique. Moreover, it can be shown that the replicator dynamic converges to Nash equilibrium from all interior initial conditions in any strictly stable

Evolutionary Game Theory, Table 3
Lyapunov functions for five basic deterministic dynamics in stable games

Dynamic	Lyapunov function for stable games
Replicator	$H_{x^*}(x) = \sum_{i \in S(x^*)} x_i^* \log \frac{x_i^*}{x_i}$
Logit	$\tilde{G}(x) = \max_{y \in \text{int}(X)} \left(y' \hat{F}(x) - \eta \sum_{i \in S} y_i \log y_i\right) + \eta \sum_{i \in S} x_i \log x_i$
Best response	$G(x) = \max_{i \in S} \hat{F}_i(x)$
BNN	$\Gamma(x) = \frac{1}{2} \sum_{i \in S} [\hat{F}_i(x)]_+^2$
Smith	$\Psi(x) = \frac{1}{2} \sum_{i \in S} \sum_{j \in S} x_i [F_j(x) - F_i(x)]_+^2$

game [4,105,224], and that the direct evaluation dynamics introduced above converge to Nash equilibrium from all initial conditions in all stable games, strictly stable or not [98,102,104,197]. In each case, the proof of convergence is based on the construction of a Lyapunov function that solutions of the relevant dynamic descend. The Lyapunov functions for the five basic dynamics are presented in Table 3.

Interestingly, the convergence results for direct evaluation dynamics are not restricted to the dynamics listed in Table 3, but extend to other dynamics in the same families (cf Table 2). But compared to the conditions for convergence in potential games, the conditions for convergence in stable games demand additional structure on the adjustment process [102].

Perturbed Best Response Dynamics in Supermodular Games Supermodular games are defined by the property that higher choices by one's opponents (with respect to the natural ordering on $S = \{1, \ldots, n\}$) make one's own higher strategies look relatively more desirable. Let the matrix $\Sigma \in \mathbf{R}^{(n-1)\times n}$ satisfy $\Sigma_{ij} = 1$ if $j > i$ and $\Sigma_{ij} = 0$ otherwise, so that $\Sigma x \in \mathbf{R}^{n-1}$ is the "decumulative distribution function" corresponding to the "density function" x. The population game F is a *supermodular game* if it exhibits *strategic complementarities*:

$$\text{If } \Sigma y \geq \Sigma x, \text{ then } F_{i+1}(y) - F_i(y) \geq F_{i+1}(x) - F_i(x) \quad \text{for all } i < n \text{ and } x \in X. \tag{23}$$

If F is smooth, condition (23) is equivalent to

$$\frac{\partial(F_{i+1} - F_i)}{\partial(e_{j+1} - e_j)}(x) \geq 0 \quad \text{for all } i, j < n \text{ and } x \in X. \tag{24}$$

Example 25 Consider this model of *search with positive externalities*. A population of agents choose levels of search effort in $S = \{1, \ldots, n\}$. The payoff to choosing effort i is

$$F_i(x) = m(i)\, b(a(x)) - c(i),$$

where $a(x) = \sum_{k \leq n} k x_k$ is the aggregate search effort, b is some increasing benefit function, m is an increasing multiplier function, and c is an arbitrary cost function. Notice that the benefits from searching are increasing in both own search effort and in the aggregate search effort. It is easy to check that F is a supermodular game.

Complementarity condition (23) implies that the agents' best response correspondence is monotone in the stochastic dominance order, which in turn ensures the existence of minimal and maximal Nash equilibria [207]. One can take advantage of the monotoncity of best responses in studying evolutionary dynamics by appealing to the theory of monotone dynamical systems [196]. To do so, one needs to focus on dynamics that respect the monotonicity of best responses and that also are smooth, so that the the theory of monotone dynamics can be applied. It turns out that the logit dynamic satisfies these criteria; so does any perturbed best response dynamic defined in terms of stochastic payoff perturbations. In supermodular games, these dynamics define cooperative differential equations; consequently, solutions of these dynamics from almost every initial condition converge to an approximate Nash equilibrium [104].

Imitation Dynamics in Dominance Solvable Games Suppose that in the population game F, strategy i is a strictly dominated by strategy j: $F_i(x) < F_j(x)$ for all $x \in X$. Consider the evolution of behavior under the repli-

cator dynamic (13). Since for this dynamic we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\frac{x_i}{x_j} &= \frac{\dot{x}_i x_j - \dot{x}_j x_i}{(x_j)^2}\\ &= \frac{x_i \hat{F}_i(x) x_j - x_j \hat{F}_j(x) x_i}{(x_j)^2}\\ &= \frac{x_i}{x_j}\left(\hat{F}_i(x) - \hat{F}_j(x)\right),\end{aligned}$$

solutions from every interior initial condition converge to the face of the simplex where the dominated strategy is unplayed [3]. It follows that the replicator dynamic converges in games with a strictly dominant strategy, and by iterating this argument, one can show that this dynamic converges to equilibrium in any game that can be solved by iterative deletion of strictly dominated strategies. In fact, this argument is not specific to the replicator dynamic, but can be shown to apply to a range of dynamics based on imitation [110,170]. Even in games which are not dominance solvable, arguments of a similar flavor can be used to restrict the long run behavior of imitative dynamics to better-reply closed sets [162]; see Sect. "Convergence to Equilibria and to Better–Reply Closed Sets" for a related discussion.

While the analysis here has focused on imitative dynamics, it is natural to expect that elimination of dominated strategies will extend to any reasonable evolutionary dynamic. But we will see in Sect. "Survival of Dominated Strategies" that this is not the case: the elimination of dominated strategies that obtains under imitative dynamics is the exception, not the rule.

Nonconvergence

The previous section revealed that when certain global structural conditions on payoffs are satisfied, one can establish global convergence to equilibrium under various classes of evolutionary dynamics. Of course, if these conditions are not met, convergence cannot be guaranteed. In this section, we offer examples to illustrate some of the possibilities for nonconvergent limit behavior.

Conservative Properties of the Replicator Dynamic in Zero-Sum Games In Sect. "Stable Games", we noted that in strictly stable games, the replicator dynamic converges to Nash equilibrium from all interior initial conditions. To prove this, one shows that interior solutions descend the function

$$H_{x^*}(x) = \sum_{i \in S(x^*)} x_i^* \log \frac{x_i^*}{x_i},$$

until converging to its minimizer, the unique Nash equilibrium x^*.

Now, random matching in a symmetric zero-sum game generates a population game that is stable, but not strictly stable (Example 22). In this case, for each interior Nash equilibrium x^*, the function H_{x^*} is a constant of motion for the replicator dynamic: its value is fixed along every interior solution trajectory.

Example 26 Suppose that agents are randomly matched to play the symmetric zero-sum game A, given by

$$\begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & -1 & 0 & 1 \\ 2 & 1 & 0 & -1 & 0 \\ 3 & 0 & 1 & 0 & -1 \\ 4 & -1 & 0 & 1 & 0 \end{array}.$$

The Nash equilibria of $F(x) = Ax$ are the points on the line segment NE connecting states $(\frac{1}{2}, 0, \frac{1}{2}, 0)$ and $(0, \frac{1}{2}, 0, \frac{1}{2})$, a segment that passes through the barycenter $x^* = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. Figure 3 shows solutions to the replicator dynamic that lie on the level set $H_{x^*}(x) = .58$. Evidently, each of these solutions forms a closed orbit.

Although solution trajectories of the replicator dynamic do not converge in zero-sum games, it can be proved that the the time average of each solution trajectory converges to Nash equilibrium [190].

The existence of a constant of motion is not the only conservative property enjoyed by replicator dynamics for symmetric zero-sum games: these dynamics are also vol-

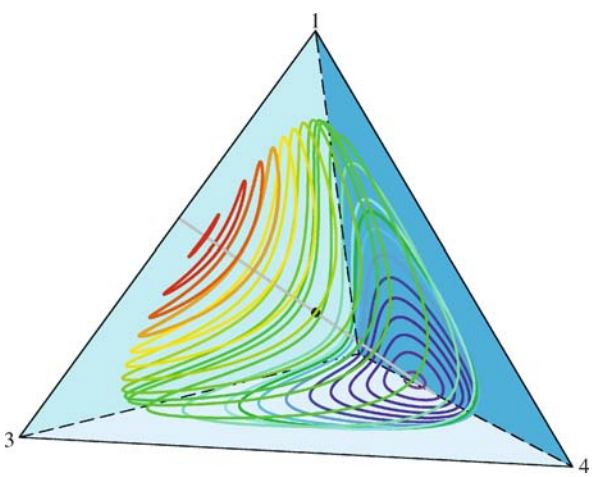

Evolutionary Game Theory, Figure 3
Solutions of the replicator dynamic in a zero-sum game. The solutions pictured lie on the level set $H_{x^*}(x) = .58$

ume preserving after an appropriate change of speed or change of measure [5,96].

Games with Nonconvergent Dynamics The conservative properties described in the previous section have been established only for the replicator dynamic (and its distant relative, the projection dynamic [185]). Inspired by Shapley [193], many researchers have sought to construct games in which large classes of evolutionary dynamics fail to converge to equilibrium.

Example 27 Suppose that players are randomly matched to play the following symmetric normal form game [107,109]:

	1	2	3	4
1	0	0	-1	ε
2	ε	0	0	-1
3	-1	ε	0	0
4	0	-1	ε	0

.

When $\varepsilon = 0$, the payoff matrix $A^\varepsilon = A^0$ is symmetric, so F^0 is a potential game with potential function $f(x) = \frac{1}{2}x'A^0x = -x_1x_3 - x_2x_4$. The function f attains its minimum of $-\frac{1}{4}$ at states $v = (\frac{1}{2}, 0, \frac{1}{2}, 0)$ and $w = (0, \frac{1}{2}, 0, \frac{1}{2})$, has a saddle point with value $-\frac{1}{8}$ at the Nash equilibrium $x^* = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, and attains its maximum of 0 along the closed path of Nash equilibria γ consisting of edges $\overline{e_1e_2}$, $\overline{e_2e_3}$, $\overline{e_3e_4}$, and $\overline{e_4e_1}$.

Let $\dot{x} = V^F(x)$ be an evolutionary dynamic that satisfies Nash stationarity (NS) and positive correlation (PC), and that is based on a revision protocol that is continuous (C). If we apply this dynamic to game F^0, then the foregoing discussion implies that all solutions to $\dot{x} = V^{F^0}(x)$ whose initial conditions ξ satisfy $f(\xi) > -\frac{1}{8}$ converge to γ. The Smith dynamic for F^0 is illustrated in Fig. 4a.

Now consider the same dynamic for the game F^ε, where $\varepsilon > 0$. By continuity (C), the attractor γ of V^{F^0} continues to an attractor γ^ε of V^{F^ε} whose basin of attraction approximates that of γ under $\dot{x} = V^{F^0}(x)$ (Fig. 4b). But since the unique Nash equilibrium of F^ε is the barycenter x^*, it follows that solutions from most initial conditions converge to an attractor far from any Nash equilibrium.

Other examples of games in which many dynamics fail to converge include monocyclic games [22,83,97,106], Mismatching Pennies [91,116], and the hypnodisk game [103]. These examples demonstrate that there is no evolutionary dynamic that converges to Nash equilibrium regardless of the game at hand. This suggests that in general, analyses of long run behavior should not restrict attention to equilibria alone.

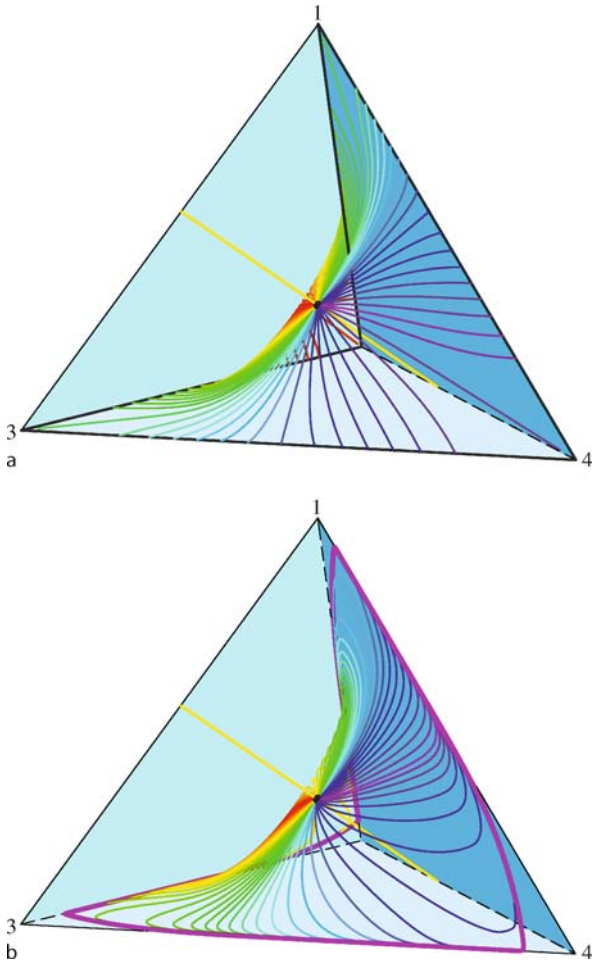

Evolutionary Game Theory, Figure 4
Solutions of the Smith dynamic in a the potential game F^0; b the perturbed potential game F^ε, $\varepsilon = \frac{1}{10}$

Chaotic Dynamics We have seen that deterministic evolutionary game dynamics can follow closed orbits and approach limit cycles. We now show that they also can behave chaotically.

Example 28 Consider evolution under the replicator dynamic when agents are randomly matched to play the symmetric normal form game below [13,195], whose lone interior Nash equilibrium is the barycenter $x^* = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$:

	1	2	3	4
1	0	-12	0	22
2	20	0	0	-10
3	-21	-4	0	35
4	10	-2	2	0

.

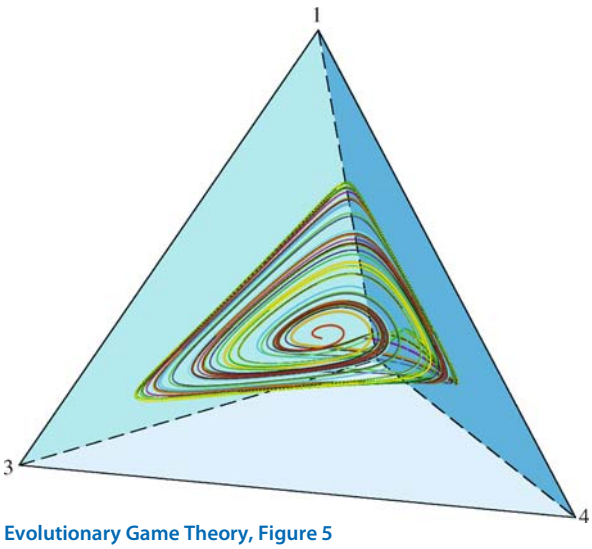

Evolutionary Game Theory, Figure 5
Chaotic behavior under the replicator dynamic

Figure 5 presents a solution to the replicator dynamic for this game from initial condition $x_0 = (.24, .26, .25, .25)$. This solution spirals clockwise about x^*. Near the rightmost point of each circuit, where the value of x_3 gets close to zero, solutions sometimes proceed along an "outside" path on which the value of x_3 surpasses .6. But they sometimes follow an "inside" path on which x_3 remains below .4, and at other times do something in between. Which of these alternatives occurs is difficult to predict from approximate information about the previous behavior of the system.

While the game in Example 28 has a complicated payoff structure, in multipopulation contexts one can find chaotic evolutionary dynamics in very simple games [187].

Survival of Dominated Strategies In Sect. "Imitation Dynamics in Dominance Solvable Games", we saw that dynamics based on imitation eliminate strictly dominated strategies along solutions from interior initial conditions. While this result seems unsurprising, it is actually extremely fragile: [25,103] prove that dynamics that satisfy continuity (C), Nash stationarity (NS), and positive correlation (PC) and that are not based exclusively on imitation must fail to eliminate strictly dominated strategies in some games. Thus, evolutionary support for a basic rationality criterion is more tenuous than the results for imitative dynamics suggest.

Example 29 Figure 6a presents the Smith dynamic for "bad RPS with a twin":

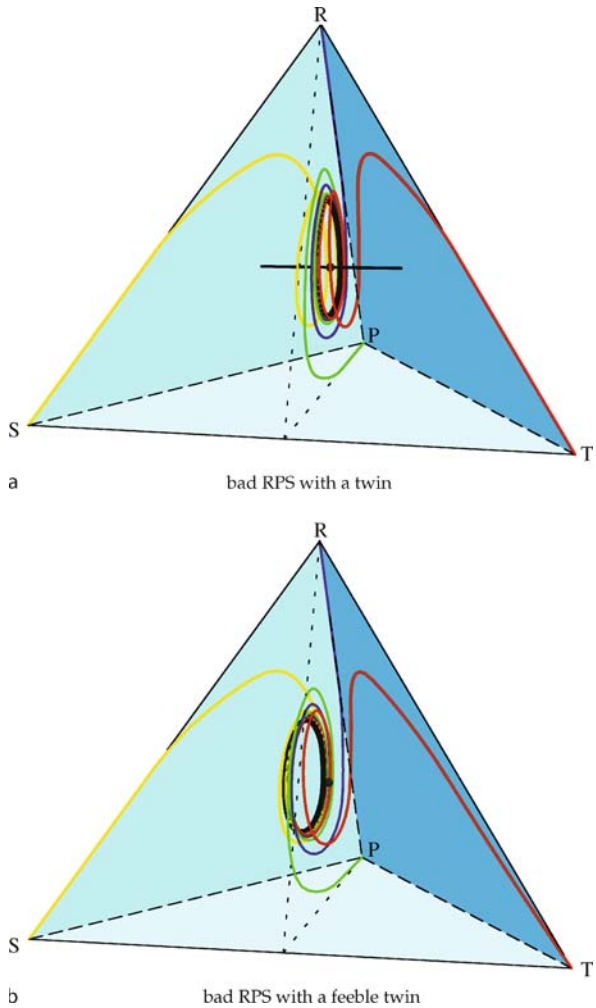

Evolutionary Game Theory, Figure 6
The Smith dynamic in two games

	R	P	S	T
R	0	−2	1	1
P	1	0	−2	−2
S	−2	1	0	0
T	−2	1	0	0

.

The Nash equilibria of this game are the states on line segment $NE = \{x^* \in X : x^* = (\frac{1}{3}, \frac{1}{3}, c, \frac{1}{3} - c)\}$, which is a repellor under the Smith dynamic. Under this dynamic, strategies gain players at rates that depend on their payoffs, but lose players at rates proportional to their current usage levels. It follows that when the dynamics are not at rest, the proportions of players choosing strategies 3 and 4 become equal, so that the dynamic approaches the plane $P = \{x \in X : x_3 = x_4\}$ on which the twins receive equal

weight. Since the usual three-strategy version of bad RPS, exhibits cycling solutions here on the plane P approach a closed orbit away from any Nash equilibrium.

Figure 6b presents the Smith dynamic in "bad RPS with a feeble twin",

	R	P	S	T
R	0	-2	1	1
P	1	0	-2	-2
S	-2	1	0	0
T	$-2-\epsilon$	$1-\epsilon$	$-\epsilon$	$-\epsilon$

with $\varepsilon = \frac{1}{10}$. Evidently, the attractor from Fig. 6a moves slightly to the left, reflecting the fact that the payoff to Twin has gone down. But since the new attractor is in the interior of X, the strictly dominated strategy Twin is always played with probabilities bounded far away from zero.

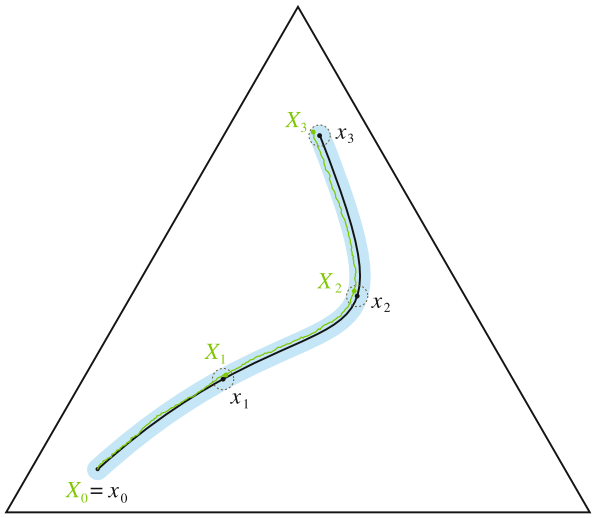

Evolutionary Game Theory, Figure 7
Deterministic approximation of the Markov process $\{X_t^N\}$

Stochastic Dynamics

In Sect. "Revision Protocols" we defined the stochastic evolutionary process $\{X_t^N\}$ in terms of a simple model of myopic individual choice. We then turned to the study of deterministic dynamics, which we claimed could be used to approximate the stochastic process $\{X_t^N\}$ over finite time spans and for large population sizes. In this section, we turn our attention to the stochastic process $\{X_t^N\}$ itself. After offering a formal version of the deterministic approximation result, we investigate the long run behavior of $\{X_t^N\}$, focusing on the questions of convergence to equilibrium and selection among multiple stable equilibria.

Deterministic Approximation

In Sect. "Revision Protocols", we defined the Markovian evolutionary process $\{X_t^N\}$ from a revision protocol ρ, a population game F, and a finite population size N. In Sect. "Mean Dynamics", we argued that the expected motion of this process is captured by the mean dynamic

$$\dot{x}_i = V_i^F(x) = \sum_{j\in S} x_j \rho_{ji}(F(x), x) - x_i \sum_{j\in S} \rho_{ij}(F(x), x). \tag{M}$$

The basic link between the Markov process $\{X_t^N\}$ and its mean dynamic (M) is provided by *Kurtz's Theorem* [127], variations and extensions of which have been offered in a number of game-theoretic contexts [24,29,43,44,175,204]. Consider the sequence of Markov processes $\{\{X_t^N\}_{t\geq 0}\}_{N=N_0}^{\infty}$, supposing that the initial conditions X_0^N converge to $x_0 \in X$. Let $\{x_t\}_{t\geq 0}$ be the solution to the mean dynamic (M) starting from x_0. Kurtz's Theorem tells us that for each finite time horizon $T < \infty$ and error bound $\varepsilon > 0$, we have that

$$\lim_{N\to\infty} P\left(\sup_{t\in[0,T]} \left|X_t^N - x_t\right| < \varepsilon\right) = 1. \tag{25}$$

Thus, when the population size N is large, nearly all sample paths of the Markov process $\{X_t^N\}$ stay within ε of a solution of the mean dynamic (M) through time T. By choosing N large enough, we can ensure that with probability close to one, X_t^N and x_t differ by no more than ε for all times t between 0 and T (Fig. 7).

The intuition for this result comes from the law of large numbers. At each revision opportunity, the increment in the process $\{X_t^N\}$ is stochastic. Still, at most population states the expected number of revision opportunities that arrive during the brief time interval $I = [t, t + \mathrm{d}t]$ is large – in particular, of order $N\mathrm{d}t$. Since each opportunity leads to an increment of the state of size $\frac{1}{N}$, the size of the overall change in the state during time interval I is of order $\mathrm{d}t$. Thus, during this interval there are a large number of revision opportunities, each following nearly the same transition probabilities, and hence having nearly the same expected increments. The law of large numbers therefore suggests that the change in $\{X_t^N\}$ during this interval should be almost completely determined by the expected motion of $\{X_t^N\}$, as described by the mean dynamic (M).

Convergence to Equilibria and to Better–Reply Closed Sets

Stochastic models of evolution can also be used to address directly the question of convergence to equilibrium [61,78,117,118,125,143,172,219]. Suppose that a society of agents is randomly matched to play an (asymmetric) normal form game that is *weakly acyclic in better replies*: from each strategy profile, there exists a sequence of profitable unilateral deviations leading to a Nash equilibrium. If agents switch to strategies that do at least as well as their current one against the choices of random samples of opponents, then the society will eventually escape any better-response cycle, ultimately settling upon a Nash equilibrium.

Importantly, many classes of normal form games are weakly acyclic in better replies: these include potential games, dominance solvable games, certain supermodular games, and certain *aggregative games*, in which each agent's payoffs only depend on opponents' behavior through a scalar aggregate statistic. Thus, in all of these cases, simple stochastic better-reply procedures are certain to lead to Nash equilibrium play.

Outside these classes of games, one can narrow down the possibilities for long run behavior by looking at *better-reply closed sets*: that is, subsets of the set of strategy profiles that cannot be escaped without a player switching to an inferior strategy (cf. [16,162]). Stochastic better-reply procedures must lead to a cluster of population states corresponding to a better-reply closed set; once the society enters such a cluster, it never departs.

Stochastic Stability and Equilibirum Selection

To this point, we used stochastic evolutionary dynamics to provide foundations for deterministic dynamics and to address the question of convergence to equilibrium. But stochastic evolutionary dynamics introduce an entirely new possibility: that of obtaining unique long-run predictions of play, even in games with multiple locally stable equilibria. This form of analysis, which we consider next, was pioneered by Foster and Young [74], Kandori, Mailath, and Rob [119], and Young [219], building on mathematical techniques due to Freidlin and Wentzell [75].

Stochastic Stability To minimize notation, let us describe the evolution of behavior using a discrete-time Markov chain $\{X_k^{N,\varepsilon}\}_{k=0}^{\infty}$ on $\mathcal{X}^N$, where the parameter $\varepsilon > 0$ represents the level of "noise" in agents' decision procedures. The noise ensures that the Markov chain is irreducible and aperiodic: any state in $\mathcal{X}^N$ can be reached from any other, and there is positive probability that a period passes without a change in the state.

Under these conditions, the Markov chain $\{X_k^{N,\varepsilon}\}$ admits a unique *stationary distribution*, $\mu^{N,\varepsilon}$, a measure on the state space $\mathcal{X}^N$ that is invariant under the Markov chain:

$$\sum_{x\in\mathcal{X}^N} \mu^{N,\varepsilon}(x)\, P\big(X_{k+1}^{N,\varepsilon} = y \,\big|\, X_k^{N,\varepsilon} = x\big) = \mu^{N,\varepsilon}(y) \qquad \text{for all } y \in \mathcal{X}^N.$$

The stationary distribution describes the long run behavior of the process $\{X_t^{N,\varepsilon}\}$ in two distinct ways. First, $\mu^{N,\varepsilon}$ is the *limiting distribution* of $\{X_t^{N,\varepsilon}\}$:

$$\lim_{k\to\infty} P\big(X_k^{N,\varepsilon} = y \,\big|\, X_0^{N,\varepsilon} = x\big) = \mu^{N,\varepsilon}(y) \qquad \text{for all } x, y \in \mathcal{X}^N.$$

Second, $\mu^{N,\varepsilon}$ almost surely describes the *limiting empirical distribution* of $\{X_t^{N,\varepsilon}\}$:

$$P\Big(\lim_{K\to\infty} \frac{1}{K}\sum_{k=0}^{K-1} 1_{\{X_k^{N,\varepsilon}\in A\}} = \mu^{N,\varepsilon}(A)\Big) = 1 \qquad \text{for any } A \subseteq \mathcal{X}^N.$$

Thus, if most of the mass in the stationary distribution $\mu^{N,\varepsilon}$ were placed on a single state, then this state would provide a unique prediction of long run behavior.

With this motivation, consider a sequence of Markov chains $\{\{X_k^{N,\varepsilon}\}_{k=0}^{\infty}\}_{\varepsilon\in(0,\bar{\varepsilon})}$ parametrized by noise levels ε that approach zero. Population state $x \in \mathcal{X}^N$ is said to be *stochastically stable* if it retains positive weight in the stationary distributions of these Markov chains as ε becomes arbitrarily small:

$$\lim_{\varepsilon\to 0} \mu^{N,\varepsilon}(x) > 0.$$

When the stochastically stable state is unique, it offers a unique prediction of play that is relevant over sufficiently long time spans.

Bernoulli Arrivals and Mutations Following the approach of many early contributors to the literature, let us consider a model of stochastic evolution based on *Bernoulli arrivals of revision opportunities* and *best responses with mutations*. The former assumption means that during each discrete time period, each agent has probability $\theta \in (0, 1]$ of receiving an opportunity to update his strategy. This assumption differs than the one we proposed in Sect. "Revision Protocols"; the key new implication is

that all agents may receive revision opportunities simultaneously. (Models that assume this directly generate similar results.) The latter assumption posits that when an agent receives a revision opportunity, he plays a best response to the current strategy distribution with probability $1-\varepsilon$, and chooses a strategy at random with probability ε.

Example 30 Suppose that a population of N agents is randomly matched to play the Stag Hunt game (Example 2):

$$\begin{array}{cc} & \begin{array}{cc} H & S \end{array} \\ \begin{array}{c} H \\ S \end{array} & \boxed{\begin{array}{cc} h & h \\ 0 & s \end{array}} \end{array}.$$

Since $s > h > 0$, hunting hare and hunting stag are both symmetric pure equilibria; the game also admits the symmetric mixed equilibrium $x^* = (x_H^*, x_S^*) = (\frac{s-h}{s}, \frac{h}{s})$.

If more than fraction x_H^* of the agents hunt hare, then hare is the unique best response, while if more than fraction x_S^* of the agents hunt stag, then stag is the unique best response. Thus, under any deterministic dynamic that respects payoffs, the mixed equilibrium x^* divides the state space into two basins of attraction, one for each of the two pure equilibria.

Now consider our stochastic evolutionary process. If the noise level ε is small, this process typically behaves like a deterministic process, moving quickly toward one of the two pure states, $e_H = (1,0)$ or $e_S = (0,1)$, and remaining there for some time. But since the process is ergodic, it will eventually leave the pure state it reaches first, and in fact will switch from one pure state to the other infinitely often.

To determine the stochastically stable state, we must compute and compare the "improbabilities" of these transitions. If the current state is e_H, a transition to e_S requires mutations to cause roughly Nx_S^* agents to switch to the suboptimal strategy S, sending the population into the basin of attraction of e_S; the probability of this event is of order $\varepsilon^{Nx_S^*}$. Similarly, to transit from e_S to e_H, mutations must cause roughly $Nx_H^* = N(1-x_S^*)$ to switch from S to H; this probability of this event is of order $\varepsilon^{N(1-x_S^*)}$.

Which of these rare events is more likely ones depends on whether x_S^* is greater than or less than $\frac{1}{2}$. If $s > 2h$, so that $x_S^* < \frac{1}{2}$, then $\varepsilon^{Nx_S^*}$ is much smaller than $\varepsilon^{N(1-x_S^*)}$ when ε is small; thus, state e_S is stochastically stable (Fig. 8a). If instead $s < 2h$, so that $x_S^* > \frac{1}{2}$, then $\varepsilon^{N(1-x_S^*)} < \varepsilon^{Nx_S^*}$, so e_H is stochastically stable (Fig. 8b).

These calculations show that *risk dominance* – being the optimal response against a uniformly randomizing opponent – drives stochastic stability 2×2 games. In particular, when $s < 2h$, so that risk dominance and payoff dominance disagree, stochastic stability favors the former over the latter.

This example illustrates how under Bernoulli arrivals and mutations, stochastic stability analysis is based on *mutation counting*: that is, on determining how many simultaneous mutations are required to move from each equilibrium into the basin of attraction of each other equilibrium. In games with more than two strategies, completing the argument becomes more complicated than in the example above: the analysis, typically based on the tree-analysis techniques of [75,219], requires one to account for the relative difficulties of transitions between all pairs of equilibria. [68] develops a streamlined method of computing the stochastically stable state based on radius-coradius calculations; while this approach is not always sufficiently fine

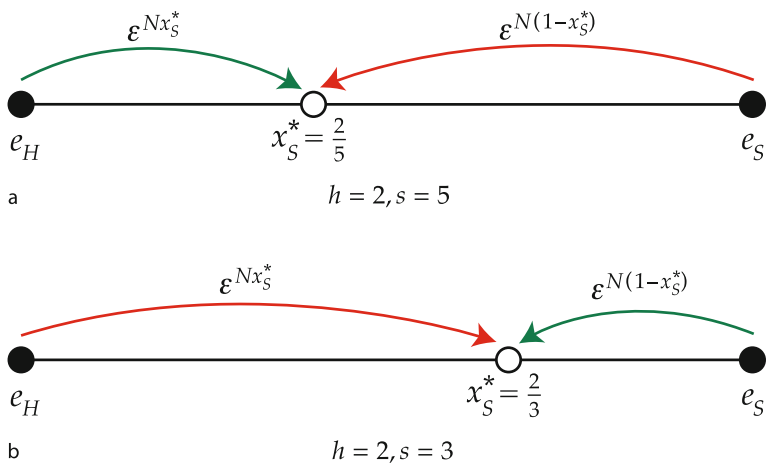

Evolutionary Game Theory, Figure 8
Equilibrium selection via mutation counting in Stag Hunt games

to yield a complete analysis, in the cases where it works it can be considerably simpler to apply than the tree-analysis method.

These techniques have been employed successfully to variety of classes of games, including pure coordination games, supermodular games, games satisfying "bandwagon" properties, and games with equilibria that satisfy generalizations of risk dominance [68,120,121,134]. A closely related literature uses stochastic stability as a basis for evaluating traditional solution concepts for extensive form games [90,115,122,128,152,168,169].

A number of authors have shown that variations on the Bernoulli arrivals and mutations model can lead to different equilibrium selection results. For instance, [165,211] show that if choices are determined from the payoffs from a single round of matching (rather than from expected payoffs), the payoff dominant equilibrium rather than the risk dominant equilibrium is selected. If choices depend on strategies' relative performances rather than their absolute performances, then long run behavior need not resemble a Nash equilibrium at all [26,161,171,198]. Finally, if the probability of mutation depends on the current population state, then any recurrent set of the unperturbed process (e. g., any pure equilibrium of a coordination game) can be selected in the long run if the mutation rates are specified in an appropriate way [27]. This last result suggests that mistake probabilities should be provided with an explicit foundation, a topic we take up in Sect. "Poisson Arrivals and Payoff Noise".

Another important criticism of the stochastic stability literature concerns the length of time needed for its predictions to become relevant [31,67]. If the population size N is large and the mutation rate ε is small, then the probability ε^{cN} that a transition between equilibria occurs during given period is miniscule; the waiting time between transitions is thus enormous. Indeed, if the mutation rate falls over time, or if the population size grows over time, then ergodicity may fail, abrogating equilibrium selection entirely [163,186]. These analyses suggest that except in applications with very long time horizons, the unique predictions generated by analyses of stochastic stability may be inappropriate, and that modelers would do better to focus on history-dependent predictions of the sort provided by deterministic models. At the same time, there are frameworks in which stochastic stability becomes relevant much more quickly. The most important of these are local interaction models, which we discuss in Sect. "Local Interaction".

Poisson Arrivals and Payoff Noise Combining the assumption of Bernoulli arrivals of revision opportunities with that of best responses with mutations creates a model in which the probabilities of transitions between equilibria are easy to compute: one can focus on events in which large numbers of agents switch to a suboptimal strategy at once, each doing so with the same probability. But the simplicity of this argument also highlights the potency of the assumptions behind it.

An appealing alternative approach is to model stochastic evolution using *Poisson arrivals of revision opportunities* and *payoff noise* [29,31,38,39,63,135,145,209,210,222]. (One can achieve similar effects by looking at models defined in terms of stochastic differential equations; see [18,48,74,79,113].) By allowing revision opportunities to arrive in continuous time, as we did in Sect. "Revision Protocols", we ensure that agents do not receive opportunities simultaneously, ruling out the simultaneous mass revisions that drive the Bernoulli arrival model. (One can accomplish the same end using a discrete time model by assuming that one agent updates during each period; the resulting process is a random time change away from the Poisson arrivals model.)

Under Poisson arrivals, transitions between equilibria occur gradually, as the population works its way out of basins of attraction one agent at a time. In this context, the mutation assumption becomes particularly potent, ensuring that the probabilities of suboptimal choices do not vary with their payoff consequences. Under the alternative assumption of payoff noise, one supposes that agents play best responses to payoffs that are subject to random perturbations drawn from a fixed multivariate distribution. In this case, suboptimal choices are much more likely near basin boundaries, where the payoffs of second-best strategies are not much less than those of optimal ones, than they are at stable equilibria, where payoff differences are larger.

Evidently, assuming Poisson arrivals and payoff noise means that stochastic stability cannot be assessed by way of mutation counting. To determine the unlikelihood of escaping from an equilibrium's basin of attraction, one must not only account for the "width" of the basin of attraction (i. e., the number of suboptimal choices needed to escape it), but also for its "depth" (the unlikelihood of each of these choices). In two-strategy games this is not difficult to accomplish: in this case the evolutionary process is a birth-and-death chain, and its stationary distribution can be expressed using an explicit formula. Beyond this case, one can employ the Freidlin and Wentzell [75] machinery, although doing so tends to be computationally demanding.

This computational burden is less in models that retain Poisson arrivals, but replace perturbed optimization with decision rules based on imitation and mutation [80].

Because agents imitate successful opponents, the population spends the vast majority of periods on the edges of the simplex, implying that the probabilities of transitions between vertices can be determined using birth-and-death chain methods [158]. As a consequence, one can reduce the problem of finding the stochastically stable state in an n strategy coordination game to that of computing the limiting stationary distribution of an n state Markov chain.

Stochastic Stability via Large Population Limits The approach to stochastic stability followed thus far relies on small noise limits: that is, on evaluating the limit of the stationary distributions $\mu^{N,\varepsilon}$ as the noise level ε approaches zero. Binmore and Samuelson [29] argue that in the contexts where evolutionary models are appropriate, the amount of noise in agents decisions is not negligible, so that taking the low noise limit may not be desirable. At the same time, evolutionary models are intended to describe behavior in large populations, suggesting an alternative approach: that of evaluating the limit of the stationary distributions $\mu^{N,\varepsilon}$ as the population size N grows large.

In one respect, this approach complicates the analysis. When N is fixed and ε varies, each stationary distribution $\mu^{N,\varepsilon}$ is a measure on the fixed state space $\mathcal{X}^N = \{x \in X : Nx \in \mathbf{Z}^n\}$. But when ε is fixed and N varies, the state space $\mathcal{X}^N$ varies as well, and one must introduce notions of weak convergence of probability measures in order to define stochastic stability.

But in other respects taking large population limits can make analysis simpler. We saw in Sect. "Deterministic Approximation" that by taking the large population limit, we can approximate the finite-horizon sample paths of the stochastic evolutionary process $\{X_t^{N,\varepsilon}\}$ by solutions to the mean dynamic (M). Now we are concerned with infinite horizon behavior, but it is still reasonable to hope that the large population limit will again reduce some of our computations to a calculus problems.

As one might expect, this approach is easiest to follow in the two-strategy case, where for each fixed population size N, the evolutionary process $\{X_t^{N,\varepsilon}\}$ is a birth-and-death chain. When one takes the large population limit, the formulas for waiting times and for the stationary distribution can be evaluated using integral approximations [24,29,39,222]. Indeed, the approximations so obtained take an appealing simple form [182].

The analysis becomes more complicated beyond the two-strategy case, but certain models have proved amenable to analysis. For instance [80], characterizes large population stochastic stability in models based on imitation and mutation. Imitation ensures that the population spends nearly all periods on the edges of the simplex X, and the large population limit makes evaluating the probabilities of transitions along these edges relatively simple.

If one supposes that agents play best responses to noisy payoffs, then one must account directly for the behavior of the process $\{X_t^{N,\varepsilon}\}$ in the interior of the simplex. One possibility is to combine the deterministic approximation results from Sect. "Deterministic Approximation" with techniques from the theory of stochastic approximation [20,21] to show that the large N limiting stationary distribution is concentrated on attractors of the mean dynamic. By combining this idea with convergence results for deterministic dynamics from Sect. "Global Convergence", Ref. [104] shows that the limiting stationary distribution must be concentrated around equilibrium states in potential games, stable games, and supermodular games.

The results in [104] do not address the question of equilibrium selection. However, for the specific case of logit evolution in potential games, a complete characterization of the large population limit of the process $\{X_t^{N,\varepsilon}\}$ has been obtained [23]. By combining deterministic approximation results, which describe the usual behavior of the process within basins of attraction, with a large deviations analysis, which characterizes the rare escapes from basins of attraction, one can obtain a precise asymptotic formula for the large N limiting stationary distribution. This formula accounts both for the typical procession of the process along solutions of the mean dynamic, and for the rare sojourns of the process against this deterministic flow.

Local Interaction

All of the game dynamics considered so far have been based implicitly on the assumption of *global interaction*: each agent's payoffs depend directly on all agents' actions. In many contexts, one expects to the contrary that interactions will be local in nature: for instance, agents may live in fixed locations and interact only with neighbors. In addition to providing a natural fit for these applications, *local interaction* models respond to some of the criticisms of the stochastic stability literature. At the same time, once one moves beyond relatively simple cases, local interaction models become exceptionally complicated, and so lend themselves to methods of analysis very different from those considered thus far.

Stochastic Stability and Equilibrium Selection Revisited

In Sect. "Stochastic Stability and Equilibirum Selection", we saw the prediction of risk dominant equilibrium play provided by stochastic stability models is subverted by the waiting-time critique: namely, that the length of time re-

quired before this equilibrium is reached may be extremely long. Ellison [67,68] shows that if interactions are local, then selection of the risk dominant equilibrium persists, and waiting times are no longer an issue.

Example 31 In the simplest local interaction model, a population of N agents are located at N distinct positions around a circle. During each period of play, each agent plays the Stag Hunt game (Examples 2 and 30) with his two nearest neighbors, following the same action against both of his opponents. If we suppose that $s \in (h, 2h)$, so that hunting hare is the risk dominant strategy, then by definition, an agent whose neighbors play different strategies finds it optimal to choose H himself.

Now suppose that there are Bernoulli arrivals of revision opportunities, and that decisions are based on best responses and rare mutations. To move from the all S state to the all H state, it is enough that a single agent mutates S to H. This one mutation begins a chain reaction: the mutating agent's neighbors respond optimally by switching to H themselves; they are followed in this by their own neighbors; and the contagion continues until all agents choose H. Since a single mutation is always enough to spur the transition from all S to all H, the expected wait before this transition is small, even when the population is large.

In contrast, the transition back from all H to all S is extremely unlikely. Even if all but one of the agents simultaneously mutate to S, the contagion process described above will return the population to the all-H state. Thus, while the transition from all-S to all-H occurs quickly, the reverse transition takes even longer than in the global interaction setting.

The local interaction approach to equilibrium selection has been advanced in a variety of directions: by allowing agents to choose their locations [69], or to pay a cost to choose different strategies against different opponents [86], and by basing agents' decisions on the attainment of aspiration levels [11], or on imitation of successful opponents [9,10]. A portion of this literature initiated by Blume develops connections between local interaction models in evolutionary game theory with models from statistical mechanics [36,37,38,124,141]. These models provide a point of departure for research on complex spatial dynamics in games, which we consider next.

Complex Spatial Dynamics

The local interaction models described above address the questions of convergence to equilibrium and selection among multiple equilibria. In the cases where convergence and selection results obtain, behavior in these models is relatively simple, as most periods are spent with most agents coordinating on a single strategy. A distinct branch of the literature on evolution and local interaction focuses on cases with complex dynamics, where instead of settling quickly into a homogeneous, static configuration, behavior remains in flux, with multiple strategies coexisting for long periods of time.

Example 32 Cooperating is a dominated strategy in the Prisoner's Dilemma, and is not played in equilibrium in finitely repeated versions of this game. Nevertheless, a pair of Prisoner's Dilemma tournaments conducted by Axelrod [14] were won by the strategy Tit-for-Tat, which cooperates against cooperative opponents and defects against defectors. Axelrod's work spawned a vast literature aiming to understand the persistence of individually irrational but socially beneficial behavior.

To address this question, Nowak and May [153,154,155,156,157] consider a population of agents who are repeatedly matched to play the Prisoner's Dilemma

$$\begin{array}{cc} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \boxed{\begin{array}{cc} 1 & -\varepsilon \\ g & 0 \end{array}} \end{array},$$

where the greedy payoff g exceeds 1 and $\varepsilon > 0$ is small. The agents are positioned on a two-dimensional grid. During each period, each agent plays the Prisoner's Dilemma with the eight agents in his (Moore) neighborhood. In the simplest version of the model, all agents simultaneously update their strategies at the end of each period. If an agent's total payoff that period is as high as that of any of neighbor, he continues to play the same strategy; otherwise, he switches to the strategy of the neighbor who obtained the highest payoff.

Since defecting is a dominant strategy in the Prisoner's Dilemma, one might expect the local interaction process to converge to a state at which all agents defect, as would be the case in nearly any model of global interaction. But while an agent is always better off defecting himself, he also is better off the more of his neighbors cooperate; and since evolution is based on imitation, cooperators tend to have more cooperators as neighbors than do defectors.

In Figs. 9–11, we present snapshots of the local interaction process for choices of the greedy payoff g from each of three distinct parameter regions. If $g > \frac{5}{3}$ (Fig. 9), the process quickly converges to a configuration containing a few rectangular islands of cooperators in a sea of defectors; the exact configuration depending on the initial conditions. If instead $g < \frac{8}{5}$ (Fig. 10), the process moves towards a configuration in which agents other than those in a "web" of defectors cooperate. But for $g \in (\frac{8}{5}, \frac{5}{3})$ (Fig. 11), the sys-

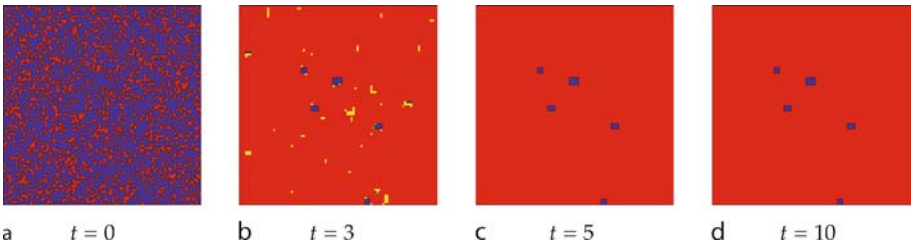

Evolutionary Game Theory, Figure 9
Local interaction in a Prisoner's Dilemma; greedy payoff $g = 1.7$. In Figs. 9–11, agents are arrayed on a 100×100 grid with periodic boundaries (i. e., a torus). Initial conditions are random with 75% cooperators and 25% defectors. Agents update simultaneously, imitating the neighbor who earned the highest payoff. *Blue* cells represent cooperators who also cooperated last period, *green* cells represent new cooperators; *red* cells represent defectors who also defected last period, *yellow* cells represent new defectors. (Figs. 9–11 created using VirtualLabs [92])

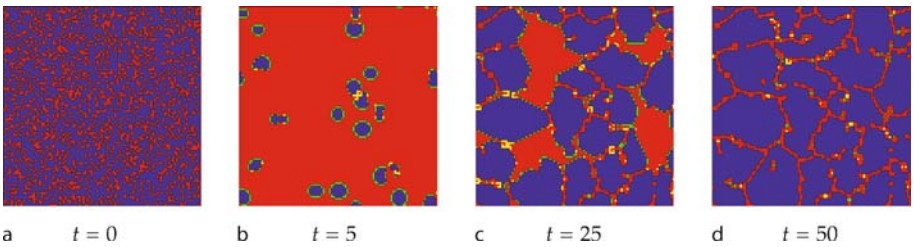

Evolutionary Game Theory, Figure 10
Local interaction in a Prisoner's Dilemma; greedy payoff $g = 1.55$

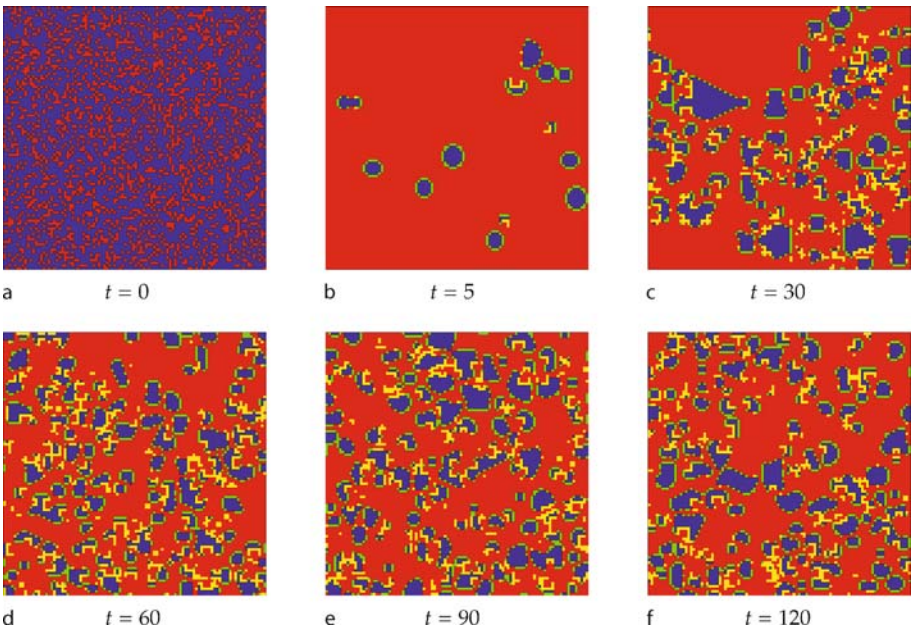

Evolutionary Game Theory, Figure 11
Local interaction in a Prisoner's Dilemma; greedy payoff $g = 1.65$

tem evolves in a complicated fashion, with clusters of cooperators and of defectors forming, expanding, disappearing, and reforming. But while the configuration of behavior never stabilizes, the proportion of cooperators appears to settle down to about .30.

The specification of the dynamics considered above, based on simultaneous updating and certain imitation of the most successful neighbor, presents a relatively favorable environment for cooperative behavior. Nevertheless, under Poisson arrivals of revision opportunities, or probabilistic decision rules, or both, cooperation can persist for very long periods of time for values of g significantly larger than 1 [154,155].

The literature on complex spatial dynamics in evolutionary game models is large and rapidly growing, with the evolution of behavior in the spatial Prisoners' Dilemma being the single most-studied environment. While analyses are typically based on simulations, analytical results have been obtained in some relatively simple settings [71,94].

Recent work on complex spatial dynamics has considered games with three or more strategies, including Rock–Paper–Scissors games, as well as public good contribution games and Prisoner's Dilemmas with voluntary participation. Introducing more than two strategies can lead to qualitatively novel dynamic phenomena, including large-scale spatial cycles and traveling waves [93,202,203]. In addition to simulations, the analysis of complex spatial dynamics is often based on approximation techniques from non-equilibrium statistical physics, and much of the research on these dynamics has appeared in the physics literature. [201] offers a comprehensive survey of work on this topic.

Applications

Evolutionary game theory was created with biological applications squarely in mind. In the prehistory of the field, Fisher [73] and Hamilton [87] used game-theoretic ideas to understand the evolution of sex ratios. Maynard Smith [137,138,139,140] introduced his definition of ESS as a way of understanding ritualized animal conflicts. Since these early contributions, evolutionary game theory has been used to study a diverse array of biological questions, including mate choice, parental investment, parent-offspring conflict, social foraging, and predator-prey systems. For overviews of research on these and other topics in biology, see [65,88].

The early development of evolutionary game theory in economics was motivated primarily by theoretical concerns: the justification of traditional game-theoretic solution concepts, and the development of methods for equilibrium selection in games with multiple stable equilibria. More recently, evolutionary game theory has been applied to concrete economic environments, in some instances as a means of contending with equilibrium selection problems, and in others to obtain an explicitly dynamic model of the phenomena of interest. Of course, these applications are most successful when the behavioral assumptions that underlie the evolutionary approach are appropriate, and when the time horizon needed for the results to become relevant corresponds to the one germane to the application at hand.

Topics in economics theoretical studied using the methods of evolutionary game theory range from behavior in markets [1,6,7,8,12,19,64,112,129,212], to bargaining and hold-up problems [32,46,57,66,164,208,220,221,222], to externality and implementation problems [47,49,136,174,177,180], to questions of public good provision and collective action [146,147,148]. The techniques described here are being applied with increasing frequency to problems of broader social science interest, including residential segregation [40,62,142,222,223,225,226] and cultural evolution [34,126], and to the study of behavior in transportation and computer networks [72,143,150,173,175,177,197]. A proliferating branch of research extends the approaches described in this article to address the evolution of structure and behavior in social networks; a number of recent books [85,114,213] offer detailed treatments of work in this domain.

Future Directions

Evolutionary game theory is a maturing field; many basic theoretical issues are well understood, but many difficult questions remain. It is tempting to say that stochastic and local interaction models offer the more open terrain for further explorations. But while it is true that we know less about these models than about deterministic evolutionary dynamics, even our knowledge of the latter is limited: while dynamics on one and two dimensional state spaces, and for games satisfying a few interesting structural assumptions, are well-understood, the dynamics of behavior in the vast majority of many-strategy games are not.

The prospects for further applications of the tools of evolutionary game theory are brighter still. In economics, and in other social sciences, the analysis of mathematical models has too often been synonymous with the computation and evaluation of equilibrium behavior. The questions of whether and how equilibrium will come to be are often ignored, and the possibility of long-term disequilibrium behavior left unmentioned. For settings in which its

assumptions are tenable, evolutionary game theory offers a host of techniques for modeling the dynamics of economic behavior. The exploitation of the possibilities for a deeper understanding of human social interactions has hardly begun.

Acknowledgments

The figures in Sects. "Deterministic Dynamics" and "Local Interaction" were created using Dynamo [184] and VirtualLabs [92], respectively. I am grateful to Caltech for its hospitality as I completed this article, and I gratefully acknowledge financial support under NSF Grant SES-0617753.

Bibliography

1. Agastya M (2004) Stochastic stability in a double auction. Games Econ Behav 48:203–222
2. Akin E (1979) The geometry of population genetics. Springer, Berlin
3. Akin E (1980) Domination or equilibrium. Math Biosci 50: 239–250
4. Akin E (1990) The differential geometry of population genetics and evolutionary games. In: Lessard S (ed) Mathematical and statistical developments of evolutionary theory. Kluwer, Dordrecht, pp 1–93
5. Akin E, Losert V (1984) Evolutionary dynamics of zero-sum games. J Math Biol 20:231–258
6. Alós-Ferrer C (2005) The evolutionary stability of perfectly competitive behavior. Econ Theory 26:497–516
7. Alós-Ferrer C, Ania AB, Schenk-Hoppé KR (2000) An evolutionary model of Bertrand oligopoly. Games Econ Behav 33:1–19
8. Alós-Ferrer C, Kirchsteiger G, Walzl M (2006) On the evolution of market institutions: The platform design paradox. Unpublished manuscript, University of Konstanz
9. Alós-Ferrer C, Weidenholzer S (2006) Contagion and efficiency. J Econ Theory forthcoming, University of Konstanz and University of Vienna
10. Alós-Ferrer C, Weidenholzer S (2006) Imitation, local interactions, and efficiency. Econ Lett 93:163–168
11. Anderlini L, Ianni A (1996) Path dependence and learning from neighbors. Games Econ Behav 13:141–177
12. Ania AB, Tröger T, Wambach A (2002) An evolutionary analysis of insurance markets with adverse selection. Games Econ Behav 40:153–184
13. Arneodo A, Coullet P, Tresser C (1980) Occurrence of strange attractors in three-dimensional Volterra equations. Phys Lett 79A:259–263
14. Axelrod R (1984) The evolution of cooperation. Basic Books, New York
15. Balkenborg D, Schlag KH (2001) Evolutionarily stable sets. Int J Game Theory 29:571–595
16. Basu K, Weibull JW (1991) Strategy sets closed under rational behavior. Econ Lett 36:141–146
17. Beckmann M, McGuire CB, Winsten CB (1956) Studies in the economics of transportation. Yale University Press, New Haven
18. Beggs AW (2002) Stochastic evolution with slow learning. Econ Theory 19:379–405
19. Ben-Shoham A, Serrano R, Volij O (2004) The evolution of exchange. J Econ Theory 114:310–328
20. Benaïm M (1998) Recursive algorithms, urn processes, and the chaining number of chain recurrent sets. Ergod Theory Dyn Syst 18:53–87
21. Benaïm M, Hirsch MW (1999) On stochastic approximation algorithms with constant step size whose average is cooperative. Ann Appl Probab 30:850–869
22. Benaïm M, Hofbauer J, Hopkins E (2006) Learning in games with unstable equilibria. Unpublished manuscript, Université de Neuchâtel, University of Vienna and University of Edinburgh
23. Benaïm M, Sandholm WH (2007) Logit evolution in potential games: Reversibility, rates of convergence, large deviations, and equilibrium selection. Unpublished manuscript, Université de Neuchâtel and University of Wisconsin
24. Benaïm M, Weibull JW (2003) Deterministic approximation of stochastic evolution in games. Econometrica 71:873–903
25. Berger U, Hofbauer J (2006) Irrational behavior in the Brown-von Neumann-Nash dynamics. Games Econ Behav 56:1–6
26. Bergin J, Bernhardt D (2004) Comparative learning dynamics. Int Econ Rev 45:431–465
27. Bergin J, Lipman BL (1996) Evolution with state-dependent mutations. Econometrica 64:943–956
28. Binmore K, Gale J, Samuelson L (1995) Learning to be imperfect: The ultimatum game. Games Econ Behav 8:56–90
29. Binmore K, Samuelson L (1997) Muddling through: Noisy equilibrium selection. J Econ Theory 74:235–265
30. Binmore K, Samuelson L (1999) Evolutionary drift and equilibrium selection. Rev Econ Stud 66:363–393
31. Binmore K, Samuelson L, Vaughan R (1995) Musical chairs: Modeling noisy evolution. Games Econ Behav 11:1–35
32. Binmore K, Samuelson L, Peyton Young H (2003) Equilibrium selection in bargaining models. Games Econ Behav 45:296–328
33. Bishop DT, Cannings C (1978) A generalised war of attrition. J Theor Biol 70:85–124
34. Bisin A, Verdier T (2001) The economics of cultural transmission and the dynamics of preferences. J Econ Theory 97:298–319
35. Björnerstedt J, Weibull JW (1996) Nash equilibrium and evolution by imitation. In: Arrow KJ et al. (eds) The Rational Foundations of Economic Behavior. St. Martin's Press, New York, pp 155–181
36. Blume LE (1993) The statistical mechanics of strategic interaction. Games Econ Behav 5:387–424
37. Blume LE (1995) The statistical mechanics of best reponse strategy revision. Games Econ Behav 11:111–145
38. Blume LE (1997) Population games. In: Arthur WB, Durlauf SN, Lane DA (eds) The economy as an evolving complex system II. Addison–Wesley, Reading pp 425–460
39. Blume LE (2003) How noise matters. Games Econ Behav 44:251–271
40. Bøg M (2006) Is segregation robust? Unpublished manuscript, Stockholm School of Economics
41. Bomze IM (1990) Dynamical aspects of evolutionary stability. Monatshefte Mathematik 110:189–206
42. Bomze IM (1991) Cross entropy minimization in uninvadable states of complex populations. J Math Biol 30:73–87

43. Börgers T, Sarin R (1997) Learning through reinforcement and the replicator dynamics. J Econ Theory 77:1–14
44. Boylan RT (1995) Continuous approximation of dynamical systems with randomly matched individuals. J Econ Theory 66:615–625
45. Brown GW, von Neumann J (1950) Solutions of games by differential equations. In: Kuhn HW, Tucker AW (eds) Contributions to the theory of games I, volume 24 of Annals of Mathematics Studies. Princeton University Press, Princeton, pp 73–79
46. Burke MA, Peyton Young H (2001) Competition and custom in economic contracts: A case study of Illinois agriculture. Am Econ Rev 91:559–573
47. Cabrales A (1999) Adaptive dynamics and the implementation problem with complete information. J Econ Theory 86:159–184
48. Cabrales A (2000) Stochastic replicator dynamics. Int Econ Rev 41:451–481
49. Cabrales A, Ponti G (2000) Implementation, elimination of weakly dominated strategies and evolutionary dynamics. Rev Econ Dyn 3:247–282
50. Crawford VP (1991) An "evolutionary" interpretation of Van Huyck, Battalio, and Beil's experimental results on coordination. Games Econ Behav 3:25–59
51. Cressman R (1996) Evolutionary stability in the finitely repeated prisoner's dilemma game. J Econ Theory 68:234–248
52. Cressman R (1997) Local stability of smooth selection dynamics for normal form games. Math Soc Sci 34:1–19
53. Cressman R (2000) Subgame monotonicity in extensive form evolutionary games. Games Econ Behav 32:183–205
54. Cressman R (2003) Evolutionary dynamics and extensive form games. MIT Press, Cambridge
55. Cressman R, Schlag KH (1998) On the dynamic (in)stability of backwards induction. J Econ Theory 83:260–285
56. Dafermos S, Sparrow FT (1969) The traffic assignment problem for a general network. J Res Nat Bureau Stand B 73:91–118
57. Dawid H, Bentley MacLeod W (2008) Hold-up and the evolution of investment and bargaining norms. Games Econ Behav forthcoming62:26–52
58. Dawkins R (1976) The selfish gene. Oxford University Press, Oxford
59. Dekel E, Scotchmer S (1992) On the evolution of optimizing behavior. J Econ Theory 57:392–407
60. Demichelis S, Ritzberger K (2003) From evolutionary to strategic stability. J Econ Theory 113:51–75
61. Dindoš M, Mezzetti C (2006) Better-reply dynamics and global convergence to Nash equilibrium in aggregative games. Games Econ Behav 54:261–292
62. Dokumacı E, Sandholm WH (2007) Schelling redux: An evolutionary model of residential segregation. Unpublished manuscript, University of Wisconsin
63. Dokumacı E, Sandholm WH (2007) Stochastic evolution with perturbed payoffs and rapid play. Unpublished manuscript, University of Wisconsin
64. Droste E, Hommes, Tuinstra J (2002) Endogenous fluctuations under evolutionary pressure in Cournot competition. Games Econ Behav 40:232–269
65. Dugatkin LA, Reeve HK (eds)(1998) Game theory and animal behavior. Oxford University Press, Oxford
66. Ellingsen T, Robles J (2002) Does evolution solve the hold-up problem? Games Econ Behav 39:28–53
67. Ellison G (1993) Learning, local interaction, and coordination. Econometrica 61:1047–1071
68. Ellison G (2000) Basins of attraction, long run equilibria, and the speed of step-bystep evolution. Rev Econ Stud 67:17–45
69. Ely JC (2002) Local conventions. Adv Econ Theory 2:1(30)
70. Ely JC, Sandholm WH (2005) Evolution in Bayesian games I: Theory. Games Econ Behav 53:83–109
71. Eshel I, Samuelson L, Shaked A (1998) Altruists, egoists, and hooligans in a local interaction model. Am Econ Rev 88:157–179
72. Fischer S, Vöcking B (2006) On the evolution of selfish routing. Unpublished manuscript, RWTH Aachen
73. Fisher RA (1930) The genetical theory of natural selection. Clarendon Press, Oxford
74. Foster DP, Peyton Young H (1990) Stochastic evolutionary game dynamics. Theor Popul Biol 38:219–232 also in Corrigendum 51:77–78 (1997)
75. Freidlin MI, Wentzell AD (1998) Random perturbations of dynamical systems, 2nd edn. Springer, New York
76. Friedman D (1991) Evolutionary games in economics. Econometrica 59:637–666
77. Friedman D, Yellin J (1997) Evolving landscapes for population games. Unpublished manuscript, UC Santa Cruz
78. Friedman JW, Mezzetti C (2001) Learning in games by random sampling. J Econ Theory 98:55–84
79. Fudenberg D, Harris C (1992) Evolutionary dynamics with aggregate shocks. J Econ Theory 57:420–441
80. Fudenberg D, Imhof LA (2006) Imitation processes with small mutations. J Econ Theory 131:251–262
81. Fudenberg D, Imhof LA (2008) Monotone imitation dynamics in large populations. J Econ Theory 140:229–245
82. Fudenberg D, Levine DK (1998) Theory of learning in games. MIT Press, Cambridge
83. Gaunersdorfer A, Hofbauer J (1995) Fictitious play, shapley polygons, and the replicator equation. Games Econ Behav 11:279–303
84. Gilboa I, Matsui A (1991) Social stability and equilibrium. Econometrica 59:859–867
85. Goyal S (2007) Connections: An introduction to the economics of networks. Princeton University Press, Princeton
86. Goyal S, Janssen MCW (1997) Non-exclusive conventions and social coordination. J Econ Theory 77:34–57
87. Hamilton WD (1967) Extraordinary sex ratios. Science 156:477–488
88. Hammerstein P, Selten R (1994) Game theory and evolutionary biology. In: Aumann RJ, Hart S (eds) Handbook of Game Theory. vol 2, chap 28, Elsevier, Amsterdam, pp 929–993
89. Harsanyi JC, Selten R (1988) A General Theory of equilibrium selection in games. MIT Press, Cambridge
90. Hart S (2002) Evolutionary dynamics and backward induction. Games Econ Behav 41:227–264
91. Hart S, Mas-Colell A (2003) Uncoupled dynamics do not lead to Nash equilibrium. Am Econ Rev 93:1830–1836
92. Hauert C (2007) Virtual Labs in evolutionary game theory. Software http://www.univie.ac.at/virtuallabs. Accessed 31 Dec 2007
93. Hauert C, De Monte S, Hofbauer J, Sigmund K (2002) Volunteering as Red Queen mechanism for cooperation in public goods games. Science 296:1129–1132

94. Herz AVM (1994) Collective phenomena in spatially extended evolutionary games. J Theor Biol 169:65–87
95. Hines WGS (1987) Evolutionary stable strategies: A review of basic theory. Theor Popul Biol 31:195–272
96. Hofbauer J (1995) Imitation dynamics for games. Unpublished manuscript, University of Vienna
97. Hofbauer J (1995) Stability for the best response dynamics. Unpublished manuscript, University of Vienna
98. Hofbauer J (2000) From Nash and Brown to Maynard Smith: Equilibria, dynamics and ESS. Selection 1:81–88
99. Hofbauer J, Hopkins E (2005) Learning in perturbed asymmetric games. Games Econ Behav 52:133–152
100. Hofbauer J, Oechssler J, Riedel F (2005) Brown-von Neumann-Nash dynamics: The continuous strategy case. Unpublished manuscript, University of Vienna
101. Hofbauer J, Sandholm WH (2002) On the global convergence of stochastic fictitious play. Econometrica 70:2265–2294
102. Hofbauer J, Sandholm WH (2006) Stable games. Unpublished manuscript, University of Vienna and University of Wisconsin
103. Hofbauer J, Sandholm WH (2006) Survival of dominated strategies under evolutionary dynamics. Unpublished manuscript, University of Vienna and University of Wisconsin
104. Hofbauer J, Sandholm WH (2007) Evolution in games with randomly disturbed payoffs. J Econ Theory 132:47–69
105. Hofbauer J, Schuster P, Sigmund K (1979) A note on evolutionarily stable strategies and game dynamics. J Theor Biol 81:609–612
106. Hofbauer J, Sigmund K (1988) Theory of evolution and dynamical systems. Cambridge University Press, Cambridge
107. Hofbauer J, Sigmund K (1998) Evolutionary games and population dynamics. Cambridge University Press, Cambridge
108. Hofbauer J, Sigmund K (2003) Evolutionary game dynamics. Bull Am Math Soc (New Series) 40:479–519
109. Hofbauer J, Swinkels JM (1996) A universal Shapley example. Unpublished manuscript, University of Vienna and Northwestern University
110. Hofbauer J, Weibull JW (1996) Evolutionary selection against dominated strategies. J Econ Theory 71:558–573
111. Hopkins E (1999) A note on best response dynamics. Games Econ Behav 29:138–150
112. Hopkins E, Seymour RM (2002) The stability of price dispersion under seller and consumer learning. Int Econ Rev 43:1157–1190
113. Imhof LA (2005) The long-run behavior of the stochastic replicator dynamics. Ann Appl Probab 15:1019–1045
114. Jackson MO Social and economic networks. Princeton University Press, Princeton, forthcoming
115. Jacobsen HJ, Jensen M, Sloth B (2001) Evolutionary learning in signalling games. Games Econ Behav 34:34–63
116. Jordan JS (1993) Three problems in learning mixed-strategy Nash equilibria. Games Econ Behav 5:368–386
117. Josephson J (2008) Stochastic better reply dynamics in finite games. Econ Theory, 35:381–389
118. Josephson J, Matros A (2004) Stochastic imitation in finite games. Games Econ Behav 49:244–259
119. Kandori M, Mailath GJ, Rob R (1993) Learning, mutation, and long run equilibria in games. Econometrica 61:29–56
120. Kandori M, Rob R (1995) Evolution of equilibria in the long run: A general theory and applications. J Econ Theory 65:383–414
121. Kandori M, Rob R (1998) Bandwagon effects and long run technology choice. Games Econ Behav 22:84–120
122. Kim Y-G, Sobel J (1995) An evolutionary approach to pre-play communication. Econometrica 63:1181–1193
123. Kimura M (1958) On the change of population fitness by natural selection. Heredity 12:145–167
124. Kosfeld M (2002) Stochastic strategy adjustment in coordination games. Econ Theory 20:321–339
125. Kukushkin NS (2004) Best response dynamics in finite games with additive aggregation. Games Econ Behav 48:94–110
126. Kuran T, Sandholm WH (2008) Cultural integration and its discontents. Rev Economic Stud 75:201–228
127. Kurtz TG (1970) Solutions of ordinary differential equations as limits of pure jump Markov processes. J Appl Probab 7:49–58
128. Kuzmics C (2004) Stochastic evolutionary stability in extensive form games of perfect information. Games Econ Behav 48:321–336
129. Lahkar R (2007) The dynamic instability of dispersed price equilibria. Unpublished manuscript, University College London
130. Lahkar R, Sandholm WH The projection dynamic and the geometry of population games. Games Econ Behav, forthcoming
131. Losert V, Akin E (1983) Dynamics of games and genes: Discrete versus continuous time. J Math Biol 17:241–251
132. Lotka AJ (1920) Undamped oscillation derived from the law of mass action. J Am Chem Soc 42:1595–1598
133. Mailath GJ (1992) Introduction: Symposium on evolutionary game theory. J Econ Theory 57:259–277
134. Maruta T (1997) On the relationship between risk-dominance and stochastic stability. Games Econ Behav 19:221–234
135. Maruta T (2002) Binary games with state dependent stochastic choice. J Econ Theory 103:351–376
136. Mathevet L (2007) Supermodular Bayesian implementation: Learning and incentive design. Unpublished manuscript, Caltech
137. Maynard Smith J (1972) Game theory and the evolution of fighting. In: Maynard Smith J On Evolution. Edinburgh University Press, Edinburgh, pp 8–28
138. Maynard Smith J (1974) The theory of games and the evolution of animal conflicts. J Theor Biol 47:209–221
139. Maynard Smith J (1982) Evolution and the theory of games. Cambridge University Press, Cambridge
140. Maynard Smith J, Price GR (1973) The logic of animal conflict. Nature 246:15–18
141. Miękisz J (2004) Statistical mechanics of spatial evolutionary games. J Phys A 37:9891–9906
142. Möbius MM (2000) The formation of ghettos as a local interaction phenomenon. Unpublished manuscript, MIT
143. Monderer D, Shapley LS (1996) Potential games. Games Econ Behav 14:124–143
144. Moran PAP (1962) The statistical processes of evolutionary theory. Clarendon Press, Oxford
145. Myatt DP, Wallace CC (2003) A multinomial probit model of stochastic evolution. J Econ Theory 113:286–301
146. Myatt DP, Wallace CC (2007) An evolutionary justification for thresholds in collective-action problems. Unpublished manuscript, Oxford University
147. Myatt DP, Wallace CC (2008) An evolutionary analysis of the volunteer's dilemma. Games Econ Behav 62:67–76

148. Myatt DP, Wallace CC (2008) When does one bad apple spoil the barrel? An evolutionary analysis of collective action. Rev Econ Stud 75:499–527
149. Nachbar JH (1990) "Evolutionary" selection dynamics in games: Convergence and limit properties. Int J Game Theory 19:59–89
150. Nagurney A, Zhang D (1997) Projected dynamical systems in the formulation, stability analysis and computation of fixed demand traffic network equilibria. Transp Sci 31:147–158
151. Nash JF (1951) Non-cooperative games. Ann Math 54:287–295
152. Nöldeke G, Samuelson L (1993) An evolutionary analysis of backward and forward induction. Games Econ Behav 5:425–454
153. Nowak MA (2006) Evolutionary dynamics: Exploring the equations of life. Belknap/Harvard, Cambridge
154. Nowak MA, Bonhoeffer S, May RM (1994) More spatial games. Int J Bifurc Chaos 4:33–56
155. Nowak MA, Bonhoeffer S, May RM (1994) Spatial games and the maintenance of cooperation. Proc Nat Acad Sci 91:4877–4881
156. Nowak MA, May RM (1992) Evolutionary games and spatial chaos. Nature 359:826–829
157. Nowak MA, May RM (1993) The spatial dilemmas of evolution. Int J Bifurc Chaos 3:35–78
158. Nowak MA, Sasaki A, Taylor C, Fudenberg D (2004) Emergence of cooperation and evolutionary stability in finite populations. Nature 428:646–650
159. Oechssler J, Riedel F (2001) Evolutionary dynamics on infinite strategy spaces. Econ Theory 17:141–162
160. Oechssler J, Riedel F (2002) On the dynamic foundation of evolutionary stability in continuous models. J Econ Theory 107:141–162
161. Rhode P, Stegeman M (1996) A comment on "learning, mutation, and long run equilibria in games". Econometrica 64:443–449
162. Ritzberger K, Weibull JW (1995) Evolutionary selection in normal form games. Econometrica 63:1371–1399
163. Robles J (1998) Evolution with changing mutation rates. J Econ Theory 79:207–223
164. Robles J (2008) Evolution, bargaining and time preferences. Econ Theory 35:19–36
165. Robson A, Vega-Redondo F (1996) Efficient equilibrium selection in evolutionary games with random matching. J Econ Theory 70:65–92
166. Rosenthal RW (1973) A class of games possessing pure strategy Nash equilibria. Int J Game Theory 2:65–67
167. Samuelson L (1988) Evolutionary foundations of solution concepts for finite, two-player, normal-form games. In: Vardi MY (ed) Proc. of the Second Conference on Theoretical Aspects of Reasoning About Knowledge (Pacific Grove, CA, 1988), Morgan Kaufmann Publishers, Los Altos, pp 211–225
168. Samuelson L (1994) Stochastic stability in games with alternative best replies. J Econ Theory 64:35–65
169. Samuelson L (1997) Evolutionary games and equilibrium selection. MIT Press, Cambridge
170. Samuelson L, Zhang J (1992) Evolutionary stability in asymmetric games. J Econ Theory 57:363–391
171. Sandholm WH (1998) Simple and clever decision rules in a model of evolution. Econ Lett 61:165–170
172. Sandholm WH (2001) Almost global convergence to *p*-dominant equilibrium. Int J Game Theory 30:107–116
173. Sandholm WH (2001) Potential games with continuous player sets. J Econ Theory 97:81–108
174. Sandholm WH (2002) Evolutionary implementation and congestion pricing. Rev Econ Stud 69:81–108
175. Sandholm WH (2003) Evolution and equilibrium under inexact information. Games Econ Behav 44:343–378
176. Sandholm WH (2005) Excess payoff dynamics and other well-behaved evolutionary dynamics. J Econ Theory 124:149–170
177. Sandholm WH (2005) Negative externalities and evolutionary implementation. Rev Econ Stud 72:885–915
178. Sandholm WH (2006) Pairwise comparison dynamics. Unpublished manuscript, University of Wisconsin
179. Sandholm WH (2007) Evolution in Bayesian games II: Stability of purified equilibria. J Econ Theory 136:641–667
180. Sandholm WH (2007) Pigouvian pricing and stochastic evolutionary implementation. J Econ Theory 132:367–382
181. Sandholm WH (2007) Large population potential games. Unpublished manuscript, University of Wisconsin
182. Sandholm WH (2007) Simple formulas for stationary distributions and stochastically stable states. Games Econ Behav 59:154–162
183. Sandholm WH Population games and evolutionary dynamics. MIT Press, Cambridge, forthcoming
184. Sandholm WH, Dokumacı E (2007) Dynamo: Phase diagrams for evolutionary dynamics. Software http://www.ssc.wisc.edu/~whs/dynamo
185. Sandholm WH, Dokumacı E, Lahkar R The projection dynamic and the replicator dynamic. Games Econ Behav, forthcoming
186. Sandholm WH, Pauzner A (1998) Evolution, population growth, and history dependence. Games Econ Behav 22:84–120
187. Sato Y, Akiyama E, Doyne Farmer J (2002) Chaos in learning a simple two-person game. Proc Nat Acad Sci 99:4748–4751
188. Schlag KH (1998) Why imitate, and if so, how? A boundedly rational approach to multi-armed bandits. J Econ Theory 78:130–156
189. Schuster P, Sigmund K (1983) Replicator dynamics. J Theor Biol 100:533–538
190. Schuster P, Sigmund K, Hofbauer J, Wolff R (1981) Selfregulation of behaviour in animal societies I: Symmetric contests. Biol Cybern 40:1–8
191. Selten R (1991) Evolution, learning, and economic behavior. Games Econ Behav 3:3–24
192. Shahshahani S (1979) A new mathematical framework for the study of linkage and selection. Mem Am Math Soc 211
193. Shapley LS (1964) Some topics in two person games. In: Dresher M, Shapley LS, Tucker AW (eds) Advances in game theory. vol 52 of Annals of Mathematics Studies. Princeton University Press, Princeton, pp 1–28
194. Skyrms B (1990) The Dynamics of Rational Deliberation. Harvard University Press, Cambridge
195. Skyrms B (1992) Chaos in game dynamics. J Log Lang Inf 1:111–130
196. Smith HL (1995) Monotone Dynamical Systems: An introduction to the theory of competitive and cooperative systems. American Mathematical Society, Providence, RI

197. Smith MJ (1984) The stability of a dynamic model of traffic assignment –an application of a method of Lyapunov. Transp Sci 18:245–252
198. Stegeman M, Rhode P (2004) Stochastic Darwinian equilibria in small and large populations. Games Econ Behav 49:171–214
199. Swinkels JM (1992) Evolutionary stability with equilibrium entrants. J Econ Theory 57:306–332
200. Swinkels JM (1993) Adjustment dynamics and rational play in games. Games Econ Behav 5:455–484
201. Szabó G, Fáth G (2007) Evolutionary games on graphs. Phys Rep 446:97–216
202. Szabó G, Hauert C (2002) Phase transitions and volunteering in spatial public goods games. Phys Rev Lett 89:11801(4)
203. Tainaka K-I (2001) Physics and ecology of rock-paper-scissors game. In: Marsland TA, Frank I (eds) Computers and games, Second International Conference (Hamamatsu 2000), vol 2063 in Lecture Notes in Computer Science. Springer, Berlin, pp 384–395
204. Tanabe Y (2006) The propagation of chaos for interacting individuals in a large population. Math Soc Sci 51:125–152
205. Taylor PD, Jonker L (1978) Evolutionarily stable strategies and game dynamics. Math Biosci 40:145–156
206. Thomas B (1985) On evolutionarily stable sets. J Math Biol 22:105–115
207. Topkis D (1979) Equilibrium points in nonzero-sum *n*-person submodular games. SIAM J Control Optim 17:773–787
208. Tröger T (2002) Why sunk costs matter for bargaining outcomes: An evolutionary approach. J Econ Theory 102:28–53
209. Ui T (1998) Robustness of stochastic stability. Unpublished manuscript, Bank of Japan
210. van Damme E, Weibull JW (2002) Evolution in games with endogenous mistake probabilities. J Econ Theory 106:296–315
211. Vega-Redondo F (1996) Evolution, games, and economic behaviour. Oxford University Press, Oxford
212. Vega-Redondo F (1997) The evolution of Walrasian behavior. Econometrica 65:375–384
213. Vega-Redondo F (2007) Complex social networks. Cambridge University Press, Cambridge
214. Volterra V (1931) Lecons sur la Theorie Mathematique de la Lutte pour la Vie. Gauthier–Villars, Paris
215. von Neumann J, Morgenstern O (1944) Theory of games and economic behavior. Prentice–Hall, Princeton
216. Weibull JW (1995) Evolutionary game theory. MIT Press, Cambridge
217. Weibull JW (1996) The mass action interpretation. Excerpt from "The work of John Nash in game theory: Nobel Seminar, December 8, 1994". J Econ Theory 69:165–171
218. Weissing FJ (1991) Evolutionary stability and dynamic stability in a class of evolutionary normal form games. In: Selten R (ed) Game Equilibrium Models I. Springer, Berlin, pp 29–97
219. Peyton Young H (1993) The evolution of conventions. Econometrica 61:57–84
220. Peyton Young H (1993) An evolutionary model of bargaining. J Econ Theory 59:145–168
221. Peyton Young H (1998) Conventional contracts. Review Econ Stud 65:773–792
222. Peyton Young H (1998) Individual strategy and social structure. Princeton University Press, Princeton
223. Peyton Young H (2001) The dynamics of conformity. In: Durlauf SN, Peyton Young H (eds) Social dynamics. Brookings Institution Press/MIT Press, Washington/Cambridge, pp 133–153
224. Zeeman EC (1980) Population dynamics from game theory. In: Nitecki Z, Robinson C (eds) Global theory of dynamical systems (Evanston, 1979). number 819 in Lecture Notes in Mathematics. Springer, Berlin, pp 472–497
225. Zhang J (2004) A dynamic model of residential segregation. J Math Sociol 28:147–170
226. Zhang J (2004) Residential segregation in an all-integrationist world. J Econ Behav Organ 24:533–550

Evolution of Culture, Memetics

Francis Heylighen, Klaas Chielens
Vrije Universiteit Brussel, Brussels, Belgium

Article Outline

Glossary
Definition of the Subject
Introduction
Defining the Meme
Dynamics of Meme Replication and Spread
Social Structures
Computer Simulations of Cultural Evolution
Selection Criteria for Memes
Parasitic Memes
Empirical Tests
Future Directions
Bibliography

Glossary

Culture The attitudes, beliefs, and behaviors that, for a certain group, define their general way of life and that they have taken over from others.

Cultural evolution The development of culture over time, as conceptualized through the mechanisms of variation and natural selection of cultural elements.

Replicator An information pattern that is able to make copies of itself, typically with the help of another system. Examples are genes, memes, and (computer) viruses.

Meme A cultural replicator; a unit of imitation or communication.

Memeplex (or meme complex) A collection of mutually supporting memes, which tend to replicate together.

Memetics The theoretical and empirical science that studies the replication, spread and evolution of memes.

Fitness The overall success rate of a replicator, as determined by its degree of adaptation to its environment, and the three requirements of longevity, fecundity and copying-fidelity.

Longevity The duration that an individual replicator survives.

Fecundity The speed of reproduction of a replicator, as measured by the number of copies made per time unit.

Copying-fidelity The degree to which a replicator is accurately reproduced.

Vertical transmission Transmission of traits (memes or genes) from parents to offspring.

Horizontal transmission Transmission of traits between individuals of the same generation.

Memotype A meme in the form of information held in an individual's memory.

Mediotype A meme as expressed in an external medium, such as a text, an artefact, a song, or a behavior.

Sociotype The group or community of individuals who hold a particular meme in their memory.

Definition of the Subject

Cultural traits are transmitted from person to person, similarly to genes or viruses. Cultural evolution therefore can be understood through the same basic mechanisms of reproduction, spread, variation, and natural selection that underlie biological evolution. This implies a shift from genes as units of biological information to a new type of units of cultural information: *memes*. The concept of meme can be defined as an information pattern, held in an individual's memory, which is capable of being copied to another individual's memory. Memetics can then be defined as the theoretical and empirical science that studies the replication, spread and evolution of memes. Memes differ in their degree of fitness, i. e. adaptedness to the socio-cultural environment in which they propagate. Fitter memes will be more successful in being communicated, "infecting" more individuals, thus spreading over a larger population. This biological analogy allows us to apply Darwinian concepts and theories to model cultural evolution.

Introduction

The transmission of cultural traits is a process that in many ways resembles the spread of an infectious disease: the carrier of a certain idea, behavior or attitude directly or indirectly communicates this idea to another person, who now also becomes a carrier, ready to "infect" further people. For example, after you heard your neighbor whistling a catchy tune a couple of times, you may well start whistling it yourself, thus being ready to infect some more people with the tune. Similarly, after you hear your friends recommend a new electronic tool they have bought, you may well buy one yourself, and, if you like it, start recommending it to those acquaintances who do not know it yet. Thus, cultural traits can be seen as analogous to *mind viruses* [18,28], *idea viruses* [41] or *thought contagions* [59], which are reproduced from mind to mind via imitation or communication. A truly successful trait is one that spreads like an epidemic, infecting the whole of the population, in order to end up as a stable, endemic component of that population's culture. For example, the tune may become part of the repertoire of "evergreens" that everyone knows, and the tool may become as widespread as the mobile phone or color television.

This virus metaphor is attractive in that it suggests a new perspective and new methods, such as epidemiology [4], for studying the dynamics of culture. However, in order to turn it into a well-founded scientific theory, we need a deeper understanding of the underlying assumptions and implications of this analogy. For this, we can turn to the science that studies viruses and other self-reproducing systems: biology.

It is an old idea to see a correspondence between cultural and biological evolution, with cultural entities undergoing similar processes of variation, reproduction and natural selection as organisms or genes. Around the end of the 18th century Western linguists discovered the similarities between different languages. Sir William Jones gave birth to the field of language evolution studies, more specifically in the search for the origin of languages, and their "common descent" [72]. The German linguist August Schleicher attempted to recreate this common ancestor of languages, publishing tree-diagrams of languages as early as 1853, six years before Darwin published his Origin of Species. In an 1870 article one can already read: "How does a new style of architecture prevail? How, again, does fashion change? (...) or take language itself (...) it is the idea of 'natural selection' that was wanted" [65]. The American philosopher and psychologist William James [55] pointed out in a presentation to the Harvard Natural History Society that: "A remarkable parallel, ..., obtains between the facts of social evolution on the one hand, and of zoological evolution as expounded by Mr. Darwin on the other."

By the end of the 20th century, the parallel study of cultural and biological evolution got a new impetus with the introduction, by Richard Dawkins [27] (first edition 1976), of the concept of *meme* (for a review see [7]). A meme, named in analogy with a gene, is defined as a cultural replicator, i. e. an element of culture such as a tradition, belief, idea, melody, or fashion, that can be held in memory and

transmitted or copied to the memory of another person. The core idea of memetics is that memes differ in their degree of fitness, i. e. adaptedness to the socio-cultural environment in which they propagate [29,50]. Mutations and recombinations of existing ideas will produce a variety of memes that compete with each other for the attention of people. Fitter memes will be more successful in being communicated, "infecting" more individuals, thus spreading over a larger population. The resulting evolutionary dynamics is one of variation creating new meme variants, followed by natural selection retaining only the ones that are most fit. Thus, the Darwinian principle of the survival of the fittest can be seen to underlie cultural evolution as well as biological evolution [3,5,6,33,58].

The memetic perspective on culture is complementary to the traditional social science perspective, which focuses on the characteristics of the individuals and groups communicating rather than on the characteristics of the information being communicated. This does not imply a "memetic reductionism", which would deny individual control over what you communicate. It just notes that in many cases the dynamics of information propagation and the ensuing evolution of culture can be modeled more simply from the "meme's point of view" than by analyzing the conscious or unconscious intentions of the communicating agents.

Over the past thirty years, several models of cultural evolution have been proposed that study the propagation of memes or similarly defined cultural traits e. g. "culturgens" [58] or "mnemons" [21]. Most of those models are purely theoretical, proposing various conceptualizations, implications and speculations based on the memetic perspective e. g. [15,31,38,54,57]. Some studies are mathematical in nature, applying techniques from mathematical genetics or epidemiology to quantitatively estimate the spread of particular types of memes within a population e. g. [17,23,58,60]. Others are computational, simulating the transmission of knowledge or behaviors between software agents e. g. [12,19,39]. A few are observational case studies, where the spread of a particular cultural phenomenon, such as a chain letter, an urban legend, or a social stereotype, is investigated qualitatively or quantitatively e. g. [10,24,42,68].

However, in spite of these advances, the memetic perspective on culture is not very well developed yet, and remains controversial [2,3,34]. There are several reasons why memeticists have not yet been able to convince the bulk of social and cultural scientists of the soundness of their approach.

First, the analogy with the gene, and its embodiment as DNA, seems to indicate that a meme should have a clear, well-delineated, stable structure. (Although one should note that natural selection was proposed by Darwin well before genes were postulated by Mendel, and a century before their structure was elucidated by Watson and Crick). Cultural entities, such as beliefs, ideas, fashions, and norms, on the other hand are typically ambiguous, difficult to delimit and constantly changing. Memetic models that are based on "hard", explicitly defined units, therefore, only seem applicable to a very small subset of cultural phenomena, such as chain letters. However, the biological analogy does not imply such rigidity: unlike higher organisms, the genes of bacteria and viruses too are in a flux, constantly mutating and exchanging bits of DNA with other organisms, but that does not imply that they do not obey evolutionary principles.

A second criticism of the memetic approach is that people are not passive "vehicles" or "carriers" of ideas and beliefs, the way they may carry viruses. Individuals actively interpret the information they receive in the light of their existing knowledge and values, and on the basis of that may decide to reject, accept, or modify the information that is communicated to them. In order words, individuals and groups actively intervene in the formulation and propagation of culture. In that sense, cultural evolution is Lamarckian rather than purely Darwinian.

A final criticism is that memetic models have not yet been sufficiently subjected to empirical tests [25,34]. Part of the reason is that most memetic theories do not make sufficiently concrete predictions to be falsifiable by observation. Most of these theories remain very speculative—often hardly better than a form of "armchair philosophy". Moreover, until now there simply have been very few empirical studies of how memes propagate, whether in the laboratory e. g. [62] or in real life e. g. [10], and even fewer links have been established between these observations and theoretical or mathematical models.

We will try to address these criticisms in the remainder of this article. First, we will discuss the issue of how to define a meme in an as accurate way as possible. Then we will review the process of transmission of memes between individuals, emphasizing the active role played by an individual's cognitive structure. This will give us a basis to review the dynamics of memetic propagation across a population, and the mathematical and simulation models that have been used to study it. To introduce empirical tests, we will first discuss the criteria that determine the fitness of a meme, specifying which memes are most likely to spread. We will then summarize a few experiments and case studies in which the predictive value of such selection criteria was tested. Finally, we will discuss some potential future applications of memetic research.

Defining the Meme

Replicators

The original definition of a meme by Dawkins [27] was based on the concept of *replicator*. A replicator is a system that is able to make copies of itself, typically with the help of some other system. Examples include real and computer viruses, which need respectively a cell and a computer processor to make copies of themselves. The fundamental example discussed by Dawkins is the gene, the string of DNA that carries the information on how to make a protein, and that is copied with the help of the cellular machinery whenever a cell divides. A meme too is a replicator, as it is copied whenever information is transmitted from one individual to another via communication or imitation.

Because replicators can be reproduced in different quantities, they are subject to natural selection: the one that tends to produce the largest number of replicas over an extended time span will win the competition with less productive replicators. To succeed in this, according to Dawkins [27], a good replicator should exhibit the following characteristics:

Longevity: The longer any instance of the replicating pattern survives, the more copies can be made of it. A drawing made by etching lines in the sand is likely to be erased before anybody could have reproduced it.

Fecundity: The faster the rate of copying, the more the replicator will spread. An industrial printing press can churn out many more copies of a pamphlet than an office-copying machine.

Copying-fidelity: The more accurate or faithful the copy, the more will remain of the initial pattern after several rounds of copying. If a painting is reproduced by making photocopies from photocopies, the picture will quickly become unrecognizable.

Dawkins called memes the "new" replicators, in the sense that they appeared very recently compared to genes. The reason for this evolution is clear: the typically human ability of imitation, i. e. learning new ideas, knowledge or behavior by copying what another individual already learnt, provides a tremendous shortcut for the multiple experiences of trial-and-error that are otherwise necessary to discover a useful new behavior pattern [21]. While some other animals are capable of limited imitation—e. g. songbirds learn songs from each other, and apes can imitate simple behaviors [16]—this capability is best developed in humans [15]. This accounts for our ability to develop a culture that is passed on from generation to generation, thus accumulating ever more useful knowledge in the course of its evolution. In that sense, memes can be seen to be responsible for the extremely fast development of human society and its subsequent dominance of the ecosystem.

Memes vs. Genes

When we compare the two most important replicators, genes and memes, we immediately notice a number of fundamental differences. Genes can only be transmitted from parent to offspring. Memes can in principle be transmitted between any two individuals. For genes to be transmitted, you need one generation. Memes can be transmitted in the span of minutes. Meme propagation is also much faster than gene spreading, because gene replication is restricted by the relatively small number of offspring a single parent can have, whereas the number of individuals that can take over a meme from a single individual is almost unlimited. Moreover, it seems much easier for memes to undergo variation, since the information in the nervous system is more plastic than that in the DNA, and since individuals can come into contact with many more different sources of novel memes. On the other hand, selection processes too are more efficient because of "vicarious" selection [21]: the carrier of a meme does not need to be killed in order to eliminate an inadequate meme; it suffices that he witnesses or hears about the troubles of another individual due to that same meme.

The conclusion is that cultural evolution will be several orders of magnitude faster and more efficient than genetic evolution. It should not surprise us then that during the last ten thousand years, humans have hardly changed on the genetic level, whereas their culture has undergone the most radical developments. In practice the superior *evolvability* of memes also means that in cases where genetic and memetic replicators are in competition, we would expect the memes to win, even though the genes would start with the advantage of a well-established, stable structure [15], as we will discuss further when reviewing computer simulations of such dual evolution.

While memes have a much higher fecundity than genes, their plasticity implies a much lower copying-fidelity: a message as received and understood by an individual will rarely be identical to the one that was expressed, as illustrated by the many misunderstandings and reinterpretations during communication. Yet, we should not conclude from this that effective communication is impossible: if you believed that, you would not be reading this article, hoping to assimilate the main ideas presented by its authors. The reason for such mixture of accurate transmission with creative reinterpretation is that, most fundamentally, humans are cognitive agents. This means that they

process incoming information depending on the knowledge they already have and the computing machinery they are endowed with, selectively retain some of that information in their memory, and selectively express some of that information to other agents. Generally, the transmission of information by an agent will change both the agent, who has learned something new, and the information, which will be affected by the knowledge the agent already had.

Therefore, a meme reaching an agent, if it is reproduced at all, will typically be transmitted in a changed form, possibly recombined with other information learned earlier. This explains why it is often so difficult to define or pinpoint an individual meme. In that sense, cultural evolution is Lamarckian: characteristics acquired during the lifetime of the meme's carrier can be transmitted to later carriers. Lamarckian evolution, while not being Darwinian in the strict sense, is still subject to the principle of natural selection: acquired characteristics too will be passed on selectively, depending on their fitness. Natural selection by definition will pick out the memes who survive this transmission process and are favorable among other memes thanks to possible variations they underwent in this process (although it is likely that most changes will not be beneficial to the favorability). Therefore, the fittest memes, such as certain songs, religious beliefs, scientific laws, or brand names, will have a stable, recognizable identity, even though they may differ in appearance, as exemplified by the many renditions of a song or joke. All such memes together define the culture shared by a community.

This identity will be reinforced by positive feedback that characterizes the non-linear interaction between meme and carrier: the more people encounter a particular version of a meme, the more they will tend to adapt their own version to this common prototype, the more commonly they will express this version, and thus the more people will encounter it. In this way, a variety of versions that are constantly being exchanged within the same group will tend to *converge* to a single, canonical version [8]. A newcomer to this group with a variant version will be extensively subjected to the accepted version, and is likely to eventually give in to this *conformist pressure* by adopting the majority version [17].

This *winner-takes-all* dynamics, where the initially most frequent variant comes to dominate all others, is elegantly illustrated by computer simulations of the evolution of language, in which many communicating individuals who use different words for the same concept quickly converge on a single word [69]. Similarly, most systems of ethics or religious belief tend to actively suppress any variant from their canonical version. This explains why in spite of the great variability of memes, we generally have no problems determining whether an individual belongs to a certain religious or linguistic group [51]. Note that such non-linear reinforcement does not exist for genes, since genes are transmitted only once, from parent to offspring. Moreover, once a gene is given, it can no longer be affected by the presence of other versions in the population.

Another fundamental difference between memes and genes is that for memes there is no equivalent for the traditional distinction between *genotype* (the information carried by the genes and passed on to the next generation) and *phenotype* (the specific appearance of an organism as determined by genes and environmental influences). In biological evolution, the genotype is the site of evolutionary variation (since variations in the phenotype are not passed on during reproduction) and the phenotype the site of selection (since it is the organism as a whole that survives and reproduces, or is eliminated). In memetics, we can distinguish three levels:

1. The *memotype* denotes the information as held in an individual's memory;
2. The *mediotype* denotes that information as expressed in an external medium, such as a text, an artefact, a song, or a behavior;
3. The *sociotype* denotes the group or community of individuals who hold that information in their memory [15].

Variation and selection take place on all three levels. A memotype can vary or be eliminated (forgotten) while residing in an individual's brain. A mediotype can similarly mutate (e. g. via a printing error) or be lost, and a sociotype can change when new individuals are added to the group, who may introduce different memes, or be eliminated (as when an unsuccessful tribe dies out). In conclusion, the processes of variation and selection, while analogous at the deepest level, are much more complex for memes than for genes.

Delimiting the Memetic Unit

What are the elements that make up a meme? In order to analyze meme structure, we can use some concepts from cognitive science, the discipline that studies mental content. Perhaps the most popular unit used to represent knowledge in artificial intelligence is the *production rule*. It has the form "if condition, then action". The action leads in general to the activation of another condition or category. A production rule can thus be analyzed as a combination of even more primitive elements: two *concepts* or *categories* and a *connection* (the "then" part, which

makes the first category entail the second one). For example, a meme like "God is omnipotent" can be interpreted as "if a phenomenon is God (it belongs to the category of God-like entities), then that phenomenon is omnipotent (it belongs to the category of omnipotent entities)".

Production rules are connected when the output condition (action) of the one matches the input condition of the other. This makes it possible to construct complex cognitive systems on the basis of simple rules. In memetics, such systems are called *meme complexes* or *memeplexes*. For example, a scientific theory or a religious system of belief may be represented as a collection of mutually connected propositions or production rules, such as "God is omnipotent", "God is good", "God punishes bad people", "if you steal, you are bad", etc. This collection of rules together determines a knowledge system that allows making inferences, such as "if you steal, God will punish you". Even more concrete perceptual or behavioral memes, such as a tune, might be modeled in this way, as concatenations of production rules of the type "if C (musical note distinguished), then E (note produced and subsequently distinguished)", "if E, then A", and so on. (In fact, genetic information too can be modeled using networks of "if… then" productions: a DNA string is activated by the presence of certain proteins (condition) to which it responds by producing specific other proteins (action), see [56]).

Production rules—or at least a simplified, binary representation of them, called "classifiers"—can be used to build computer simulations of cognitive evolution, using genetic algorithms, i. e. algorithmically applied operators that perform the equivalents of mutation, recombination, and selection on the basis of "fitness" on such strings [53]. Although classifier models generally do not take into account distinct carriers, this looks like a promising road to study the evolution of memeplexes formally and computationally. As we will see later, though, simulations of cultural evolution are usually limited to the mutation and spread of simple memes, ignoring the cognitive structures and processes that support inferences and that create new meme(plexe)s out of combinations of existing ones.

Even if we would model memes as connected sets of production rules, we still have the problem of how many production rules define a single meme(plex). If we call a religion or a scientific theory a meme, it is clear that this will encompass a very large number of interconnected rules. In practice it will be impossible to enumerate all rules, or to define sharp boundaries between the rules that belong to the meme and those that do not. For example, should you believe in the existence of Hell, the creation of the world in seven days, and the virginity of Mary to be called a Catholic?

A pragmatic criterion that can be used in this regard is to define a meme or memeplex as the smallest collection of propositions or memory items that tends to replicate as a whole cf. [73]. For example, a proposition like "God is omnipotent" on its own, without specification of God's other characteristics, is much too abstract to be clearly understood or applied, and as such is unlikely to replicate well. However, in combination with a number of other propositions, like "God is good", "God is the creator of the world", etc., that flesh out, apply, and support this abstract idea, the package will make much more sense, and be more likely to be passed on to other individuals. Similarly, the first three notes of a melody are unlikely to be remembered as a unit, but the first eight, as in the beginning of Beethoven's fifth symphony, may well be.

It remains that often we can add or subtract a few production rules (such as the virginity of Mary) from a memeplex without significantly changing its chances of replication. Therefore, in practice it will rarely be possible to determine the precise boundaries of a meme(plex). However, this should not detract us from considering memes while analyzing cultural evolution. Indeed, the same problem besets genetic models of biological evolution: as yet, it is in practice impossible to specify the exact combination of DNA codons that determine the gene for, say, fair skin, big ears or altruism. The biochemical definition of a gene as a string of DNA that codes for one protein is not very useful when studying evolution, since most practical functions require a combination of proteins, most proteins exhibit a combination of functions, and much of the DNA is non-coding, but therefore not necessarily useless, as it may contain control information that determines the activation of other DNA strings.

As Dawkins [27] notes, we do not need to know the constitutive elements or boundaries of a gene in order to explain the evolution of particular characteristics, such as altruism or fair skin, for which such a gene would be responsible. It is sufficient that we can distinguish the effects of that gene from the effects of its rival genes (alleles). If we can determine the fitness differences resulting from these effects, then we can make predictions about which type of genes will win the competition in a particular situation, and thus which characteristics the species is most likely to evolve. For example, knowing that people with lighter skin need less sunlight to produce sufficient vitamin D, we can predict that in Northern regions natural selection will favor genes for light skin over genes for dark skin—whatever DNA codons make up these respective genes.

The same applies to memes. If, for example, we observe that one meme (say Catholicism) induces its carriers to have more children than its competitors (say Anglican-

ism), and that the children tend to take over their memes from their parents, then, *all other things being equal*, we can predict that after sufficient time this meme will dominate in the population. This prediction does not require any explicit definition of the meme of Catholicism, but only the ability to distinguish it from its competitors. Of course, in practice it is never the case that all other things are equal, but that is the predicament of all scientific modeling: we must always simplify, and ignore potentially important influences. The question is to do that as wisely as possible, and to maximally include relevant variables without making the model too complex.

Dynamics of Meme Replication and Spread

To be replicated, a meme must pass successfully through four subsequent stages:

1. *Assimilation* by an individual, who thereby becomes a carrier or *host* of the meme;
2. *Retention* in that individual's memory;
3. *Expression* by the individual in language, behavior, or another form that can be perceived by others;
4. *Transmission* of the thus created message or mediotype to one or more other individuals.

This last stage is followed again by stage 1, thus closing the replication loop. At each stage there is selection, meaning that some memes will be eliminated. Let us look in more detail at the mechanisms governing these four stages.

Assimilation

A successful meme must be able to "infect" a new host, that is, enter into its memory, and thus acquire its *memotype* form. Let us assume that a meme is presented to a potential new host. "Presented" means either that the individual encounters an existing mediotype form of a meme, or that he or she independently discovers the meme, by observation of outside phenomena or by thought, i. e. recombination of existing cognitive elements. To be assimilated, the presented meme must be respectively *noticed*, *understood* and *accepted* by the host. Noticing requires that the mediotype be sufficiently salient to attract the host's attention. Understanding means that the host recognizes the meme as something that fits in with his or her cognitive system. The mind is not a blank slate on which any idea can be impressed. To be understood, a new idea or phenomenon must connect to cognitive structures that are already available to the individual. Finally, a host that has understood a new idea must also be willing to believe it or to take it seriously. For example, although you are likely to understand the proposition that your car was built by little green men from Mars, you are unlikely to accept that proposition without very strong evidence. Therefore, you will in general not memorize it, and the meme will not manage to infect you.

Retention

The second stage of memetic replication is the retention of the meme in memory. The longer the meme stays, the more opportunities it will have to spread further by infecting other hosts. This is Dawkins's [27] *longevity* characteristic for replicators.

Just like assimilation, retention is characterized by strong selection, which few memes will survive. Indeed, most of the things we hear, see or understand during the day are not stored in memory for longer than a few hours. Although you may have clearly assimilated the news that the national party won the Swaziland elections with 54% of the votes, you are unlikely to remember this a week later—unless you live in Swaziland, perhaps. Retention will depend on how important the idea is to you, and how often it is repeated, either by recurrent encounter or by internal rehearsal.

Expression

To be communicated to other individuals, a meme must emerge from its storage as memory pattern or memotype and enter its mediotype phase, i. e. assume a physical shape that can be perceived by others. This process may be called "expression". The most obvious medium for expression is speech. Other common means are text, pictures, behavior, and the creation of artifacts such as tools, buildings or works of art. Expression does not require the conscious decision of the host to communicate the meme. A meme can be expressed simply by the way somebody walks or manipulates an object, or by what he or she wears.

Some retained memes will never be expressed, for example because the host does not consider the meme interesting enough for others to know, uses it unconsciously without it showing up in his or her behavior, does not know how to express it, or wants to keep it secret. On the other hand, the host may be convinced that the meme is so important that it must be expressed again and again to everybody he or she meets.

Transmission

To reach another individual, an expression needs a physical carrier or medium that is sufficiently stable to transmit the expression without too much loss or deformation.

Speech, for example, uses sound to transmit an expression, while text will be transmitted through ink on paper or electrical impulses in a wire. The expression will take the form of a physical signal, modulating the carrier into a specific shape—the mediotype—from which the original meme can be re-derived. For example, mediotypes can be books, photographs, artifacts or CD-ROMs.

Selection at the transmission stage happens through either elimination of certain memes, when the mediotype is destroyed or gets corrupted before it is perceived by another individual, or through differential multiplication, when the mediotype is reproduced into many copies. For example, a manuscript may be put into the shredder or turned into a book that is printed in millions of copies. Especially since the emergence of mass media and mass manufacturing, the transmission stage is the one where the contrast between successful and unsuccessful memes is largest, and where selection can have the largest impact.

Meme Fitness

The overall survival rate of a meme m can be expressed as the meme *fitness* $F(m)$, which measures the expected number $N(m)$ of memes at the next time step or generation $t+1$ divided by the average number of memes at the present time t. This fitness can be expressed in a simplified model as the product of the survival/multiplication rates for each of the four stages, respectively assimilation A, retention R, expression E and transmission T:

$$F(m) \equiv \frac{N(m,t+1)}{N(m,t)} = A(m).P(m).E(m).T(m)$$

A denotes the proportion of mediotypes encountered by the host that are assimilated. R represents the proportion of these assimilated memes that are retained in memory. Therefore, $A \leqslant 1$, $R \leqslant 1$. E is the number of times a retained meme is expressed by the host. T is the number of potential new hosts reached by a copy of the expression. Unlike A and R, E and T do not have an upper bound, although E is likely to be more restricted than T. Note that F is zero as soon as one of its components (A, R, E, T) is zero. This expresses the fact that a meme must successfully pass through *all* four stages in order to replicate. Also note that for a meme to spread ($F > 1$), you must have at least $E > 1$ or $T > 1$.

Dynamics of Spread

From the standard definition of fitness F, we can derive the rate of growth for the number $N(t)$ of meme copies at time t. This determines the speed with which the meme spreads through the population of carriers:

$$\frac{\mathrm{d}N}{\mathrm{d}t} \cong \frac{N(t+1)-N(t)}{1} = (F-1).N$$

This results in a traditional exponential growth if $F > 1$, exponential decay (and eventual extinction) if $F < 1$, and stability if $F = 1$. This model is too simple if the population is finite. In that case, we need to take into account the total size of the population of potential carriers K, which functions as the "carrying capacity" of the socio-cultural environment in which a meme proliferates. The increase in the number $N(t)$ of memes can be represented by the following Verhulst type of equation:

$$\frac{\mathrm{d}N}{\mathrm{d}t} = (F-1).N\left(1-\frac{N}{K}\right) .$$

This equation expresses the fact that the growth in meme number ($\mathrm{d}N$) is in first instance proportional to the number (N) that is already there—since more memes produce more copies of themselves—, but eventually limited by the number K of potential hosts in the population, so that growth becomes zero when the population reaches this limit ($N = K$). The function $N(t)$ that is the solution to this differential equation is the logistic function with its characteristic *sigmoid* (S-like) shape.

Interactions Between Memes

The dynamics of a single growing meme population $N(m)$ could be extended to several interacting memes $N_i = N(m_i)$. Here we should add an interaction term A_{ij} which describes the strength of the influence of meme i on meme j. This influence can be positive ($A_{ij} > 0$), which means that an increase in i produces an increase in j, i. e. i helps j to grow. A negative influence ($A_{ij} < 0$) means that the growth of i suppresses the growth of j. A neutral relation ($A_{ij} = 0$) means that the spread of the one does not influence the spread of the other. This applies to memes from independent domains, such as "God exists" and "apples are healthy". If we now consider the reciprocal influence (A_{ji}), we can distinguish the following specific types of interaction:

- $A_{ij} > 0$, $A_{ji} > 0$: the memetic species i and j can be seen as *mutualists*, that help each other to spread, e. g. by reinforcing each others' message. An example could be "God is good" and "God is great".
- $A_{ij} < 0$, $A_{ji} < 0$: i and j are rivals or *competitors* [12]: an increase in the one produces a decrease in the other. Examples are "God is good" and "God does not exist".

- $A_{ij} > 0$, $A_{ji} < 0$: i and j stand in a *predator-prey* type of relationship, i. e. i grows at the expense of j. This may happen when i (e. g. relativity theory) is a more advanced version of j (e. g. Newtonian mechanics), so that carriers of j would quickly convert to i, but non-carriers of j would be more difficult to convince of i's value.

The overall dynamics can be represented by the following system of non-linear differential $\sum_j$ equations:

$$\frac{\mathrm{d}N_i}{\mathrm{d}t} = N_i \left(\sum_j A_{ij} N_j + B_i \right)$$

A_{ii}, the influence of meme i on itself will here normally always be negative and equal to $(1 - F_i)/K$, while $B_i = F_i - 1$, as in the previous equation for a single meme. Such dynamical models quickly become very complex to solve, but are not fundamentally different from traditional growth and competition models used in population biology, epidemiology, or studies of the diffusion of innovations [12]. However, they do not take into account the dependence of a meme on its carrier, nor the specific communication channels between carriers.

Social Structures

One way to make the model more realistic without adding too much complication is to consider the structure of the social space in which the potential carriers of a meme reside. Here we make the additional assumption of continuity, namely that a meme cannot jump from one carrier to another without there being some form of proximity or relationship between the carriers.

Horizontal Transmission and the Evolution of Cooperation

The simplest form of relationship is the one between parents and their offspring. Parent-to-child transmission (or more generally transmission between generations) is called *vertical transmission* [23]. Memes belonging to domains such as religion, language, ethics, and general culture are commonly transmitted in this way. This form of propagation is analogous to the transmission of genes. Therefore "vertical" models of cultural evolution find results similar to those of biological evolution. This means that vertically transmitted memes, such as established religions, will typically reinforce or elaborate genetically transmitted behavioral patterns and thus directly contribute to biological fitness [26].

The same does not apply to *horizontally transmitted* culture, i. e. memes exchanged between members of the same generation [23]. Here what is good for a meme (e. g. slavish imitation of fads and fashions) is not necessarily good for the biological individual or gene pool, since genes and memes are subjected to different kinds of natural selection. This may promote the evolution of parasitic memes that are deleterious to their carriers, as we will discuss further.

However, in addition to the fact that it spreads new information more quickly, horizontal transmission also offers another benefit that vertical transmission lacks. A classic problem in biological evolution is the *evolution of cooperation* [27,51]: given that genes are selected to promote their own good, with a disregard or even hostility toward any rivals that compete for their scarce resources, how can we explain cooperative or altruistic behavior where an individual invests more in helping another than in his or her own good? In the animal world, cases of altruism, such as among social insects, are usually explained via *kin selection*: individuals will help others as long as these are related to them, i. e. share their genes. In human society, however, people often help strangers that are totally unrelated. The initially proposed explanation of *group selection*, namely that groups of individuals that help each other survive better than groups of selfish individuals, has the shortcoming that, within altruist groups, it are the selfish profiteers that do best, and thus spread their genes most [27].

Horizontal transmission of cooperation norms solves this problem, since the members of a cultural group are all *memetically related* to each other, sharing their memes rather than their genes. Therefore, cultural kin selection will extend to all members of the group [36]. This entails a selective pressure for memes to support the fitness of the whole group of their carriers, e. g. by promoting cooperation. Moreover, selfish profiteers will not be able to undermine the cooperation produced by such altruism-promoting memes because of conformist pressure [17,51], or what we have called "winner-takes-all": when one meme establishes a majority position it will eventually get imposed on *all* members of the group, thus suppressing the appearance of selfish dissidents—or at least not allowing them to make any converts and thus spread their memes. This cultural solution to the cooperation paradox in biological evolution appears to have been developed more or less independently by different meme theorists [15,17,36,47].

Topologies of Communication

Horizontal transmission will generally follow existing social or geographical topologies. This can be modeled in

two different ways:

1. Individuals are situated in a *space* (typically a two-dimensional plane, or its discrete equivalent, a two-dimensional lattice of cells);
2. Individuals are considered as nodes in a (social) *network*, which are connected by ties of acquaintance or trust.

The basic assumption in these models is that memes diffuse *continuously* across the space or network. This means that, in first instance, communications are considered to be *local*, i. e. agents exchange memes only with their direct neighbors in the space or social network. The neighbor can pass on the meme to its neighbors, and so on, so that a meme eventually may spread across the whole population.

When a population consists of different clusters or local communities, that have little communication with each other, this will typically lead to different cultures establishing themselves in different communities [8,17]. The reason is that intense communication within each community will produce a "winner-takes-all" dynamics where by chance or local adaptation one of several variant memes becomes dominant. Memes from other communities, however, will only rarely be encountered, so that they will generally not receive enough reinforcement to displace the established memes.

Recent research in complex networks, including social networks, has shown that such networks commonly have a *scale-free structure* [1]. This means that a few agents, the so-called "hubs" of the network, have a great many social ties, while most agents only have a few links. The implication for cultural diffusion is that memes hosted by "hub" agents will have a disproportionately large effect, and are much more likely to spread widely. A similar effect has been observed in the spread of sexually transmitted diseases, such as AIDS, where the infection of a few hubs in the network (in this case individuals with a large number of sexual partners) may make the difference between a large-scale epidemic and a few isolated infections. This observation has provided inspiration to researchers in "viral" marketing, who look for methods to make publicity for a brand or product by creating a "buzz", i. e. a positive message about their product that is propagated via word-of-mouth [63,64]. Their strategies focus on identifying and targeting the "opinion-leaders" within a community, i. e. those central individuals that many know and tend to imitate.

Although it is in principle possible to make analytical models of the propagation of memes across space or across networks, calculating the precise spread in a realistic environment is far too difficult. Therefore, these processes are typically explored via multi-agent computer simulations.

Computer Simulations of Cultural Evolution

Cultural transmission of rules, norms or information is a common ingredient in many social simulations (e. g. [8,20,32,37,45], that are based on an "artificial society" of interacting software agents [35]. However, such memetic propagation is often added merely as one of the many assumptions within a complicated model of a specific type of socio-cultural evolution, such as the evolution of a shared vocabulary [69] or of cooperation norms [45]. There have been relatively few simulations that have explored cultural evolution in the broadest sense. We will now discuss some typical examples that illustrate the wider issue.

Probably the first explicitly memetic simulation, *Meme and Variations*, was made by Gabora [39] (first written 1992). The assumptions underlying this, and related simulations of cultural diffusion e. g. [9,30] are the following: agents search the best solution for a particular problem. They can either find a solution on their own through trial-and-error, or they can take over a solution from another agent, by observing the solutions each of their neighbors has found and imitating the best one. The result of the simulation is that the agents collectively find the best solutions if they partially imitate others, partially explore individually. If they only imitate, there is no creativity and the best solution cannot be improved. If they only explore individually, lots of search is needed to merely rediscover what was already known elsewhere. In the ideal situation, which is achieved by trying out different parameter values for the simulation until one has found the optimal mix of innovation and imitation, good solutions will spread very quickly throughout the population, but this without preventing the discovery of even better solutions by certain agents.

This simulation investigated the relative effectiveness of, and interaction between, *individual learning* and *cultural diffusion*k. An older classic simulation [52] investigated the relative effectiveness of, and interaction between, *individual learning* and *genetic evolution*. Inspired by this work, Best [14] studied the three-way interactions between individual learning, genetic evolution, and cultural evolution. In Best's simulation, agents can acquire knowledge that allows them to maximize their fitness in three ways: 1) by inheriting it, possibly with variations, from their parents (vertical, genetic transmission); 2) by copying it from another, fitter agent (horizontal, cultural transmission); 3) by individually discovering it via trial-and-error. The sim-

ulation showed that cultural transmission, just like individual learning, can enhance genetic evolution, accelerating its convergence to the optimal solution. Moreover, cultural transmission appeared superior to individual learning in that it produced convergence more quickly.

Best [14] also examined the situation in which cultural and genetic evolution pursue opposite goals, and found that in this case genetic evolution normally wins the competition. However, Bull, Holland and Blackmore [19] further investigated this situation by allowing cultural evolution to be much more rapid than genetic evolution, as is normally the case. They found that under these conditions memetic effects are stronger than genetic effects, and the only way genes can still keep some control over the process is by evolving mechanisms to filter out particularly harmful memes.

These simulations of cultural evolution are still rather simplistic, in the sense that agents literally copy any knowledge exhibited by a fitter agent. In practice, individuals do not a priori know which individual is fitter, and when they receive a message, this information will interact with the knowledge they already had. Van Overwalle and Heylighen [71] have proposed a more realistic simulation model in which agents do not just copy a message, but actively "reinterpret" it, based on their previous experience. Agents are modeled as simple neural networks that learn from experience. A message corresponds to a pattern of activation over the nodes in such a network, and communication to the spread of that activation from agent to agent via variable inter-agent connections. The strength of the connection between two agents represents the degree of trust of the one in the information received from the other. This trust is learned on the basis of the degree to which information received previously from that agent is confirmed by own knowledge. Unlike most multi-agent simulations, the Van Overwalle and Heylighen [71] model is supported by solid empirical evidence, in that it manages to accurately reproduce the results of several classic communication experiments, including the Lyons and Kashima [61,62] study of meme transmission that we will discuss in a later section.

Selection Criteria for Memes

Since mathematical models and computer simulations of meme spread necessarily have to make plenty of simplifying assumptions, and cannot incorporate all the specific social, psychological, linguistic and cultural factors that influence the propagation of a meme, they are not very useful in predicting which concrete memes will be successful and which will not. Yet, such predictions are necessary if we want to arrive at an empirically testable theory, which can be applied to practical problems. One way to arrive at a more practical, predictive model is to formulate general selection criteria that distinguish fitter memes from less fit ones. All other things being equal, *a meme that scores better on one of these criteria is predicted to become more numerous* than a meme that scores worse. This is a falsifiable hypothesis that can be tested through experiments or observations. It suffices to operationalize the criteria so that satisfaction of a criterion can be objectively measured.

Many authors have proposed criteria for memetic success, and a few (e. g. [22,44,48,49,50]) have compiled lists of such criteria. Since these proposals, while related, are all different, we need to examine more clearly what is needed for a good list of criteria. First, such criteria should be formulated to be as much as possible independent or non-overlapping, so that a piece of information can vary along one dimension of evaluation without varying along the others. Second, without becoming too restrictive, they should be defined as precisely, concretely and unambiguously as possible, so that different observers using these criteria can come to the same conclusions.

To illustrate the importance of these methodological considerations, let us review some proposed criteria, and point to their shortcomings. For example, one might naively propose that fit memes should be *attractive* to their receivers. While this is true in a general sense, it helps us very little in operationalizing meme fitness, as we cannot say what makes a meme attractive without becoming much more explicit about its properties. A somewhat more sophisticated hypothesis may propose that good memes should be *communicable* [68]. Again, this is obviously correct, but communicability has so many different aspects, depending on the meme itself, its audience, the used medium, etc., that we might as well say that it simply should be a good meme. A more specific criterion, used e. g. by [46], is *plausibility*. The problem here is that people may use very different procedures to estimate plausibility, e. g. by looking at the source of the information, the available evidence for it, or their own previous experience.

We will here summarize the criteria proposed by Heylighen [48,49,50], which are based on an extensive review of relevant cognitive, social and communicative mechanisms. At the most abstract level, there are three classes of entities that information depends on: the object that it refers to, the subject who assimilates and remembers it, and the communication process that is used to transmit it between subjects. These determine three categories of selection criteria, *objective*, *subjective* and *intersubjective*:

Objective Criteria

Distinctiveness Information that refers to something precise, distinct or detailed, can be confirmed more easily by observation.

Invariance Information that remains valid over a wide range of contexts or situations, is more stable and broadly applicable.

Evidence Information that is supported by independent observations, is more reliable.

Subjective Criteria

Utility Information that is valuable or useful to its carrier is more likely to be remembered and passed on.

Affectivity Information that provokes strong emotions is more likely to be remembered and passed on: this typically stimulates instinctive reactions, such as fear, desire or disgust [46].

Coherence The better information fits in with the knowledge that individuals already have, the more easily they will understand and accept it [70]

Simplicity Short, simple messages are easier to assimilate, remember and transmit.

Novelty Information that is unexpected will attract more attention.

Repetition Repeated exposure to the same message helps it to be assimilated and retained.

Intersubjective Criteria

Publicity The more effort an individual puts into spreading a message, the more people will receive it.

Formality Messages formulated explicitly and unambiguously are less likely to be misinterpreted.

Expressivity Information must be easy to express in a given language or medium.

Authority An authoritative, trustworthy source of the information makes it more likely to be accepted.

Conformity Information confirmed by many people is more easily accepted [17].

Collective utility Information, if adopted by a group, may help the group to function better, and therefore to grow or function as a model for others. Examples are standards, linguistic conventions, and traffic rules.

Parasitic Memes

Memes being communicated undergo natural selection. Some memes are transmitted easily, thus reaching a large number of people, while others are rejected, misunderstood, forgotten or otherwise eliminated from circulation. This means that the memes best adapted to the underlying cognitive and communicative processes will spread farthest. We may assume that our brain, general culture, and social structures have evolved so as to maximize the fitness of society and its members. This means that they should be good at assimilating useful memes, and at rejecting bad ones, e. g. [15]. For example, we know that we should not accept things without evidence, and that some sources are more reliable than other. Insofar that these socio-cognitive guidelines, as exemplified by the above list of selection criteria, efficiently filter out poor-quality information, successful memes will also increase the fitness of their carriers.

However, since no system is foolproof, these mechanisms will not always be reliable. This leaves a niche for memes to evolve that propagate well, apparently satisfying the criteria that people intuitively use, but without delivering any benefit to their carriers. We may call such memes *selfish* [47] or *parasitic* cf. [26], as they free ride on the effort invested by individuals to gather and communicate useful information. Such information parasites succeed by faking the criteria that we use to recognize high-quality information. This is similar to the way many biological organisms mimic other phenomena, such as viruses that mimic the cell's own DNA, so that they are reproduced for free by the cellular machinery. Memes have therefore been described as "mind viruses" [18,28], since they similarly exploit our cognitive machinery to get themselves replicated.

There are plenty of examples of such parasitic memes. Perhaps the most studied from a memetic point of view are *chain letters*, whose only purpose is to have themselves replicated and sent to as many people as possible [11,42]. A more modern variant are *virus hoaxes* [24,25], i. e. email messages that warn the receivers for a non-existent type of computer virus, and urge them to pass on this warning to as many people as possible. Probably the most dangerous information parasites are certain *religious cults* [26], which indoctrinate their followers to make as many converts as possible, while isolating them from alternative sources of information, so that they tend to develop a view of reality that is so distorted that it may end fatally, as in the mass suicides of the Heaven's Gate cult. *Pseudosciences* too can be dangerous, parading as solid scientific theories, but asserting statements that at best are not supported by the facts, like in astrology, at worst fatally wrong, like in certain quack cures for cancer. Somewhat more benign are *urban legends* and various *rumors* and *fads*, which tend to spread in waves, being passed on from person to person but without any authoritative source or real evidence. Bangerter and Heath [10] have tracked the evolution of

one such legend, the *Mozart effect* (i. e. the unfounded belief that babies listening to classical music become more intelligent), starting from its source: a scientific experiment that merely found that after listening to music adults temporarily scored better on certain tests—perhaps simply because they were more relaxed.

Several memeticists, e. g. [18,28,60], have investigated the many tricks that memes use to appear more acceptable than they deserve to be, including the following. *Self-justification* means that the components of a memeplex mutually justify each other, but without independent support. An example is: "God exists because the Bible says so", and "You should believe the Bible because it is the word of God". *Self-reinforcement* means that a meme stimulates its host to rehearse itself, e. g. by repeated study, meditation, prayer, etc. *Intolerance* means that a meme indoctrinates its host to a priori reject any potentially competing memes. *Proselytism* occurs when a meme urges its host to maximally spread the meme to other hosts, like in the case of chain letters. *Baiting* occurs when a meme promises its carriers a reward if they only accept and spread the meme. All these tactics are common in religious cults, which promise their adherents that they will go to Heaven if they believe the teachings, pray, and spread the word, while they will burn in Hell if they dare to doubt [28].

Parasitic memes have been the subject and inspiration of most empirical approaches to memetics, since their spread is relatively easy to track, and since they are prime illustrations of the way in which cultural evolution is fundamentally different from genetic evolution.

Empirical Tests

Memetics has often been challenged and has known some very virulent critics. One of the main criticisms is that there are no empirical data to back up the theories that were put forth, and in that sense memetics is merely a way of thinking rather than a scientific discipline [34,40,43]. This criticism is to some degree justifiable. The lack of a universally accepted meme definition and the vagueness of meme boundaries cf. [2] indeed make empirical studies less evident. Yet, there have already been a few empirical studies of meme propagation in different condition, both in the laboratory, e. g. [61,62], and in real life, e. g. [11,24,42]. More generally, we must note that memetics is an approach that illuminates important aspects of culture, society and communication that more traditional approaches, such as sociology, psychology or history, tend to neglect. Empirical tests cannot confirm or falsify this perspective as a whole, but merely specific implementations of it. This is analogous to the observation that experiments in psychology or sociology can test particular theories within their field, but not the field as such.

Within the memetics field, one simple way to test specific theories is by considering the memetic selection criteria they imply [25]. We can then measure the apparent success rate of different memes, and examine its correlation with the degree to which the memes fulfill the proposed selection criteria. Heath, Bell and Sternberg [46] applied this approach to investigate the criteria of *affectivity* (which they call *emotional selection*) and *plausibility*. From the affects, they focused on disgust because this is a relatively simple emotion whose strength is easy to measure. When comparing different urban legends that contained an element of disgust (e. g. the story of a man who discovers a dead rat in the cola bottle he has just been drinking from), they found that the more disgusting variations typically had spread more widely than the less disgusting ones. The same applied to more plausible variations. Chielens [24] used a similar method to examine the spread of virus hoaxes. He used both expert and non-expert surveys to score different hoaxes on different criteria, including novelty, simplicity, utility, authority and proselytism. In this case, the spread of the hoax correlated most strongly with its *novelty*. Schaller, Conway and Tanchuk [68] examined the correlation between the *communicability* of traits, i. e. the probability with which subjects would speak about a particular trait, and the frequency of these traits. They found that for traits used to describe different ethnic groups, the most communicable traits are also the ones that are most widely spread and persistent in society. Pocklington and Best [66] and [12,13] used automatic text analysis to measure how often a certain discussion subject was mentioned in a particular discussion group on the net. They found evidence for memetic competition between mutually *incoherent* subjects, meaning that an increase in the one correlated with a decrease in the other.

While these studies merely observed existing patterns of spread, Lyons and Kashima [61,62] performed a laboratory experiment in which they deliberately produced a memetic transmission chain. They created a fictional story and asked the participants in the experiment to subsequently tell the story from person 1 to person 2, from 2 to 3, and so one, like in the game of "telephone" or "Chinese Whispers." The story involved a fictitious tribe, the Jamayans, about which all participants had received some background information. After several experiments, a statistical analysis of the story elements that remained at the end of the transmission chain (i. e. as reported by the last person to hear the story) found a number of systematic selection effects. These seem to confirm four of the previously mentioned criteria:

1. *Coherence*: Elements inconsistent with the background information were more likely to be left out;
2. *Novelty*: Elements that the speakers assumed were already known by the listeners were more likely to be left out;
3. *Simplicity*: Details or embellishments that did not affect the story line tended to be left out;
4. *Conformity*: When the participants were told that the majority of them believed that the Jamayans were, e. g., peaceful, they were more likely to leave out elements inconsistent with this fact than if they thought that this was only a minority opinion.

Experiments on memes "in the wild" are more difficult to control than in such a laboratory situation. Because of ethical concerns it is hardly possible to release a well-doctored meme upon the world to observe how it evolves. Another problem is that it is hard to track memes once they are released. Even when feedback is asked or if mechanical trackers are included in the meme (for example, images included in an email message that are automatically downloaded from a controlled server each time the message is opened), it is hard to see whether the meme has already been changed by the time these devices are triggered.

In conclusion, empirical memetics research remains in its infancy. However, experiments have already shown that one can quantitatively confirm or falsify a number of nontrivial memetic predictions. It is to be hoped that many more studies along these lines will be carried out. As long as memetics has not been thoroughly investigated empirically, it is likely to remain a theoretical niche framework used only for description rather than prediction.

Future Directions

The theory of memetics and cultural evolution holds out great promises for a better understanding, anticipation and control of fundamental social problems that depend on the propagation of ideas and behaviors. This will require more extensive empirical observations and tests, more detailed computer models of the interaction between memes and their hosts, and a better conceptualization of what a meme precisely is. A first basic result that should come out of such research is a concrete and reliable list of criteria that characterize successful memes, i. e. ideas or cultural traits that propagate widely and easily across large populations. This would allow us not only to recognize such memes, but to some degree to design or improve them.

The ability to create successful memes is the Holy Grail of marketing research [41], which is constantly on the look-out for techniques to create a "buzz" and have their publicity message or brand name [63,64] become as widely known as possible. Another application of these principles lies in public education. For example, if the government makes a campaign to convince people to stop smoking, it would be very useful to have the campaign designed according to sound memetic principles. This should take into account both the characteristics of the message itself (e. g. being sufficiently simple and unambiguous), of the intended audience (e. g. being consistent with what the audience already believes, while being sufficiently novel to attract their attention), and of the way it is transmitted (e. g. having the meme expressed in a common medium by people considered trustworthy).

Memetic selection criteria can be applied not only positively, to help spread a beneficial idea, but negatively, to prevent or suppress harmful memes. Examples are the idea that it is cool to smoke, false rumors and scares that may promote panic or accentuate social prejudice, fundamentalist ideologies that incite hatred or terrorism, and dangerous superstitions, such as the belief that you can cure AIDS by having sex with a virgin. A better understanding of memetic dynamics may help us to understand how such mind viruses arise and spread. It may moreover help us to "immunize" the population by educating them about basic memetics, so as not to be misled by apparently plausible—but fundamentally misleading—cults, fads and superstitions [18,28].

Another basic result of future memetic research should be a complex dynamical model of the interactions between individuals, groups, and the memes they carry. This should allow making longer-term predictions about the interactions between different groups and subcultures within our globalizing society. A crucial issue in this regard is whether minority cultures will eventually be assimilated into the majority, or on the contrary become polarized, asserting their divergent habits and beliefs ever more forcefully [8,71]. Two concrete examples are minority languages, such as the Welsh still spoken in Britain, where there is a tendency for the subculture to be slowly erased by the majority culture, and the culture of Islamic immigrants in Europe, where there is a tendency towards polarization in the sense of increased radicalism. Neither complete assimilation nor polarization are desirable outcomes, but at first sight they seem like the most likely results of the "winner-takes-all" dynamics created by the pressure to conform to the group one has most contact with. A more detailed theory of cultural evolution may help us to find a middle way that preserves cultural diversity without exacerbating conflicts, and to pinpoint the crucial factors that can steer the dynamics in one direction rather than another.

Bibliography

Primary Literature

1. Albert R, Barabasi AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
2. Atran S (2001) The Trouble with memes: Inference versus imitation in cultural creation. Hum Nat 12:351–381
3. Aunger R (2001) Darwinizing Culture: The Status of Memetics As a Science. University Press, Oxford
4. Aunger R (2002) Exposure versus susceptibility in the epidemiology of 'everyday' beliefs. J Cogn Cult 2(2):113–154
5. Aunger R (2003) Cultural transmission and diffusion. In: Nadel L (ed) Encyclopedia of Cognitive Science. MacMillan, London
6. Aunger R (2003) The Electric Meme: A New Model of How We Think. Simon and Schuster, New York
7. Aunger R (2004) Memes, In: Kuper A, Kuper J (eds) The Social Science Encyclopedia, 3rd edn. Routledge, London
8. Axelrod R (1997) The Dissemination of Culture: A Model with Local Convergence and Global Polarization. J Confl Resolut 41:203–26
9. Baldassarre G, Parisi D (1999) Trial-and-Error Learning, Noise and Selection in Cultural Evolution: A Study Through Artificial Life Simulations. In: Proceedings of the AISB'99 Symposium on Imitation in Animals and Artifacts
10. Bangerter A, Heath C (2004) The Mozart effect: Tracking the evolution of a scientific legend. Br J Soc Psychol 43:605–623
11. Bennet C, Li M, Ma B (2003) Chain Letters & Evolutionary Histories. Sci Am 288(6):76–81
12. Best ML (1997) Models for interacting populations of memes: Competition and niche behavior. J Memetics 1(2) http://jom-emit.cfpm.org/1997/vol1/best_ml.html
13. Best ML (1998) An ecology of text: Using text retrieval to study alife on the net. J Artif Life 3(4):261–287
14. Best ML (1999) How culture can guide evolution: An inquiry into gene/meme enhancement. J Adapt Behav 7(3/4):289–306
15. Blackmore S (2000) The Meme Machine. Oxford University Press, Oxford
16. Bonner JT (1980) The Evolution of Culture in Animals. Princeton University Press, Princeton
17. Boyd R, Richerson PJ (1985) Culture and the evolutionary process. Chicago University Press, Chicago
18. Brodie R (1996) Virus of the mind: The new science of the meme. Integral Press, Seattle
19. Bull L, Holland W, Blackmore S (2001) On meme-gene coevolution. Artif Life 6(3):227–235
20. Bura S (1994) MINIMEME: Of life and death in the Noosphere. In: Cliff D, Husbands P, Meyer JA Wilson SW (eds) From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior. MIT Press, Cambridge, pp 479–486
21. Campbell DT (1974) Evolutionary Epistemology. In: Schilpp PA (ed) The Philosophy of Karl Popper. Open Court Publish, La Salle, pp 413–463
22. Castelfranchi C (2001) Towards a Cognitive Memetics: Socio-Cognitive Mechanisms for Memes Selection and Spreading. J Memetics Evol Models Inf Transm 5
23. Cavalli-Sforza LL, Feldman MW (1981) Cultural transmission and evolution: A quantitative approach. Princeton University Press, Princeton
24. Chielens K (2003) The Viral Aspects of Language: A Quantitative Research of Memetic Selection Criteria. Master's thesis, Vrije Universiteit Brussel. http://memetics.chielens.net/master/thesis.pdf
25. Chielens K, Heylighen F (2005) Operationalization of Meme Selection Criteria: procedures to empirically test memetic hypotheses. In: Proceedings AISB 2005
26. Cullen B (1999) Parasite Ecology and the Evolution of Religion, In: Heylighen F, Bollen J, Riegler A (eds) The Evolution of Complexity. Kluwer, Dordrecht
27. Dawkins R (1989) The Selfish Gene, 2nd edn. Oxford University Press, Oxford
28. Dawkins R (1993) Viruses of the mind, In: Dahlbom B (ed) Dennett and his Critics: Demystifying the Mind. Blackwell, USA, pp 12–27
29. de Jong M (1999) Survival of the institutionally fittest concepts. J Memetics Evol Models Inf Transm 3. http://jom-emit.cfpm.org/1999/vol3/de_jong_m.html
30. Denaro D, Parisi D (1996) Cultural evolution in a population of neural networks. In: Marinaro M, Tagliaferri R (eds) Neural nets: Wirn96. Springer, New York
31. Dennett D (1995) Darwin's dangerous idea. Hammondsworth, Penguin
32. Doran J (1998) Simulating Collective Misbelief. J Artif Soc Soc Simul 1:1. http://www.soc.surrey.ac.uk/JASSS/1/1/3.html
33. Durham WH (1991) Coevolution: Genes, culture and human diversity. Stanford University Press, Stanford
34. Edmonds B (2002) Three Challenges for the Survival of Memetics. J Memetics Evol Models Inf Transm 6. http://jom-emit.cfpm.org/2002/vol6/edmonds_b_letter.html
35. Epstein J, Axtell R (1996) Growing Artificial Societies: Social Science from the Bottom Up. MIT Press, London
36. Evers JR (1998) A justification of societal altruism according to the memetic application of Hamilton's rule. Proc 16th Int Congress on Cybernetics. Namur: Association Internat de Cybernétique, pp 437–442
37. Flentge F, Polani D, Uthmann T (2001) Modelling the Emergence of Possession Norms using Memes. J Artif Soc Soc Simul 4(4). http://www.soc.surrey.ac.uk/JASSS/4/4/3.html
38. Flinn MV, Alexander RD (1982) Culture theory: The developing synthesis from biology. Hum Ecol 10:383–400
39. Gabora L (1995) Meme and variations: A Computer Model of Cultural Evolution. In: Nadel L, Stein D (eds) 1993 Lectures in Complex Systems. Addison-Wesley, Boston, pp 471–485
40. Gil-White FJ (2005) Common Misunderstandings of Memes (and Genes). The Promise and the Limits of the Genetic Analogy to Cultural Transmission Processes. In: Hurley S, Chater N (eds) Perspectives on Imitation. From Neuroscience to Social Science, vol 2. MIT Press, Cambridge, pp 317–338
41. Godin S (2002) Unleashing the Ideavirus. Simon and Schuster, London
42. Goodenough OR, Dawkins R (2002) The 'St Jude' mind virus. Nature 371:23–24
43. Greenberg M (2005) Goals versus Memes: Explanation in the Theory of Cultural Evolution. In: Hurley S, Chater N (ed) Perspectives on Imitation. From Neuroscience to Social Science, vol 2. MIT Press, Cambridge, pp 339–353
44. Hale-Evans R (1995) Memetics: A Systems Metabiology working report. http://ron.ludism.org/memetics.html

45. Hales D (1998) Selfish memes & selfless agents – Altruism in the swap shop. In: Proceedings of the 3rd International Conference on Multi-Agent Systems. IEEE Press, Los Gatos
46. Heath C, Bell C, Sternberg E (2001) Emotional Selection in Memes: The Case of Urban Legends. J Pers Soc Psychol 81(6):1028–1041
47. Heylighen F (1992) Selfish Memes and the Evolution of Cooperation. J Ideas 2(4):77–84
48. Heylighen F (1993) Selection C riteria for the Evolution of Knowledge. In: Proc 13th Int Congress on Cybernetics Association Internat de Cybernétique, Namur, pp 524–528
49. Heylighen F (1997) Objective, subjective and intersubjective selectors of knowledge. Evol Cogn 3(1):63–67
50. Heylighen F (1998) What makes a meme successful? Selection Criteria for Cultural Evolution. Proc 16th Int Congress on Cybernetics. Association Internat de Cybernétique, Namur
51. Heylighen F, Campbell DT (1995) Selection of Organization at the Social Level: obstacles and facilitators of metasystem transitions. World Futures: J General Evol 45:181–212
52. Hinton GE, Nowlan SJ (1987) How Learning Can Guide Evolution. Complex Syst 1:495–502
53. Holland JH, Holyoak KJ, Nisbett RE, Thagard PR (1986) Induction: processes of inference, learning and discovery. MIT Press, Cambridge
54. Hull DL (1982) The naked meme. In: Plotkin HC (ed) Development and culture: Essays in evolutionary epistemology. Chichester, Wiley, pp 272–327
55. James W (1880) Great Men and their Environment. http://www.des.emory.edu/mfp/jgreatmen.html. Accessed 28 Sept 2005
56. Kauffman SA (1993) The Origins of Order: Self-Organization and Selection in Evolution. Oxford University Press, New York
57. Lake MW (1998) Digging for memes: The role of material objects in cultural evolution. In: Renfrew C, Scarre C (eds) Cognition and material culture: the archeology of symbolic storage. McDonald Institute Monographs, Cambridge pp 77–88
58. Lumsden C, Wilson E (1981) Genes, Mind, and Culture: the Coevolutionary Process. Harvard University Press, Cambridge
59. Lynch A (1996) Thought contagion: How beliefs spread through society: The new science of memes. Basic Books, New York
60. Lynch A (1998) Units, events and dynamics in memetic evolution. J Memetics Evol Models Inf Transm 2. http://www.cpm.mmu.ac.uk/jomemit/1998/vol2/lynch_a.html
61. Lyons A, Kashima Y (2001) The reproduction of culture: Communication processes tend to maintain cultural stereotypes. Soc Cogn 19:372–394
62. Lyons A, Kashima Y (2003) How Are Stereotypes Maintained Through Communication? J Pers Soc Psychol 85:989–1005
63. Marsden P (2002) Brand positioning: meme's the word. Mark Intell Plan 20(5):307–12
64. Marsden P, Kirby J (2005) Connected Marketing: The Buzz, Viral and Word of Mouth Revolution. Elsevier, London
65. Müller M (1870) The Science of Language. Nature 2:256–259
66. Pocklington R, Best ML (1997) Cultural evolution and units of selection in replicating text. J Theor Biol 188:79–87
67. Rogers EM (2003) Diffusion of Innovations, Fifth Edition. Free Press, New York
68. Schaller M, Conway LG III, Tanchuk TL (2002) Selective pressures on the once and future contents of ethnic stereotypes: Effects of the 'communicability' of traits. J Pers Soc Psychol 82:861–877
69. Steels L (1998) Synthesising the origins of language and meaning using co-evolution, selforganisation and level formation. In: Hurford J (ed) Evolution of Human Language. Edinburgh Univ Press, Edinburg
70. Thagard P (1989) Explanatory Coherence. Behav Brain Sci 12:435–467
71. Van Overwalle F, Heylighen F (2006) Talking Nets: A Multi-Agent Connectionist Approach to Communication and Trust between Individuals. Psychol Rev 113:606–627
72. Van Wyhe J (2005) The Descent of words: evolutionary thinking 1780–1880. Endeavour 29(3):94–100
73. Wilkins JS (1998) What's in a meme? Reflections from the perspective of the history and philosophy of evolutionary biology. J Memetics Evol Models Inf Transm 2. http://jom-emit.cfpm.org/1998/vol2/wilkins_js.html

Books and Reviews

Benzon W (1996) Culture as an evolutionary arena. J Soc Evol Syst 19:321–362 http://www.newsavanna.com/wlb/CE/Arena/Arena00.shtml

Blackmore S (2000) The Meme Machine. University Press, Oxford

Blackmore S (2001) Evolution And Memes: The Human Brain As A Selective Imitation Device. Cybern Syst 32(1–2):225–255; http://www.ingentaconnect.com/content/tandf/ucbs;jsessionid=1b1oec7s7li7s.henrietta

Boyd R, Richerson PJ (1985) Culture and the evolutionary process. Chicago University Press, Chicago

Edmonds B (1998) On Modelling in Memetics. J Memetics Evol Models Inf Transm 2. http://jom-emit.cfpm.org/1998/vol2/edmonds_b.html

Fog A (1999) Cultural Selection. Kluwer, Dordrecht

Heyes CM (1994) Imitation and culture: Longevity, fecundity and fidelity in social transmission. In: Galef BG, Mainardi M, Valsecchi P (eds) Behavioural Aspects of Feeding. Harwood, Chur, pp 271–287

Jesiek BK (2003) Betwixt the Popular and Academic: The Histories and Origins of Memetics. Master Thesis, Virginia Tech

Marsden P (1998) Memetics and Social Contagion: Two Sides of the Same Coin. J Memetics Evol Models Inf Transm 2. http://jom-emit.cfpm.org/1998/vol2/marsden_p.html

Moritz E (1990) Memetic Science: I – General Introduction. J Ideas 1:1–23

Evolution in Materio

Simon Harding[1], Julian F. Miller[2]
[1] Department of Computer Science, Memorial University, St. John's, Canada
[2] Department of Electronics, University of York, Heslington, UK

Article Outline

Glossary
Definition of the Subject
Introduction

Evolutionary Algorithms
Evolution in Materio: Historical Background
Evolution in Materio: Defining Suitable Materials
Evolution in Materio Is Verified with Liquid Crystal
Evolution in Materio Using Liquid Crystal: Implementational Details
The Computational Power of Materials
Future Directions
Bibliography

Glossary

Evolutionary algorithm A computer algorithm loosely inspired by Darwinian evolution.

Generate-and-test The process of generating a potential solution to a computational problem and testing it to see how good a solution it is. The idea behind it is that no human ingenuity is employed to make good solutions more likely.

Genotype A string of information that encodes a potential solution instance of a problem and allows its suitability to be assessed.

Evolution in materio The method of applying computer controlled evolution to manipulate or configure a physical system.

Liquid crystal Substances that have properties between those of a liquid and a crystal.

Definition of the Subject

Evolution in materio refers to the use of computers running search algorithms, called evolutionary algorithms, to find the values of variables that should be applied to material systems so that they carry out useful computation. Examples of such variables might be the location and magnitude of voltages that need to be applied to a particular physical system. Evolution in materio is a methodology for programming materials that utilizes physical effects that the human programmer need not be aware of. It is a general methodology for obtaining analogue computation that is specific to the desired problem domain. Although a form of this methodology was hinted at in the work of Gordon Pask in the 1950s it was not convincingly demonstrated until 1996 by Adrian Thompson, who showed that physical properties of a digital chip could be exploited by computer controlled evolution. This article describes the first demonstration that such a method can be used to obtain specific analogue computation in a non-silicon based physical material (liquid crystal). The work is important for a number of reasons. Firstly, it proposes a general method for building analogue computational devices. Secondly it explains how previously unknown physical effects may be utilized to carry out computations. Thirdly, it presents a method that can be used to *discover* useful physical effects that can form the basis of future computational devices.

Introduction

Physical Computation

Classical computation is founded on a mathematical model of computation based on an abstract (but physically inspired) machine called a Turing Machine [1]. A Turing machine is a machine that can write or erase symbols on a possibly infinite one dimensional tape. Its actions are determined by a table of instructions that determine what the machine will write on the tape (by moving one square left or right) given its state (stored in a state register) and the symbol on the tape. Turing showed that the calculations that could be performed on such a machine accord with the notion of computation in mathematics. The Turing machine is an abstraction (partly because it uses a possibly infinite tape) and to this day it is still not understood what limitations or extensions to the computational power of Turing's model might be possible using real physical processes. Von Neumann and others at the Institute for Advanced Study at Princeton devised a design for a computer based on the ideas of Turing that has formed the foundation of modern computers. Modern computers are digital in operation. Although they are made of physical devices (i. e. transistors), computations are made on the basis of whether a voltage is above or below some threshold. Prior to the invention of digital computers there have been a variety of analogue computing machines. Some of these were purely mechanical (e. g. an abacus, a slide-rule, Charles Babbage's difference engine, Vannevar Bush's Differential Analyzer) but later computing machines were built using operational amplifiers [2].

There are many aspects of computation that were deliberately ignored by Turing in his model of computation. For instance, speed, programmability, parallelism, openess, adaptivity are not considered. The speed at which an operation can be performed is clearly an important issue since it would be of little use to have a machine that can calculate any computable function but takes an arbitrarily large amount of time to do so. Programmability is another issue that is of great importance. Writing programs directly in the form of instruction tables that could be used with a device based on a Turing is extremely tedious. This is why many high-level computer languages have been devised. The general issue of how to subdivide a computer program into a number of parallel executing processes so that the intended computation is carried out as quickly as

possible is still unsolved. Openness refers to systems that can interact with an external environment during their operation. Openness is exhibited strongly in biological systems where new resources can be added or removed either by an external agency or by the actions taken by the system itself. Adaptivity refers to the ability of systems to change their characteristics in response to an environment.

In addition to these aspects, the extent to which the underlying physics affects both the abstract notion of computation and its tractability has been brought to prominence through the discovery of quantum computation, where Deutsch pointed out that Turing machines implicitly use assumptions based on physics [3]. He also showed that through 'quantum parallelism' certain computations could be performed much more quickly than on classical computers. Other forms of physical computation that have recently been explored are: reaction-diffusion systems [4], DNA computing [5,6] and synthetic biology [7].

In the UK a number of Grand Challenges in computing research have been proposed [8], in particular 'Journeys in Non-Classical Computation' [9,10] seeks to explore, unify and generalize many diverse non-classical computational paradigms to produce a mature science of computation.

Toffoli argued that 'Nothing Makes Sense in Computing Except in the Light of Evolution' [11]. He argues firstly that a necessary but not sufficient condition for a computation to have taken place, is when a novel function is produced from a fixed and finite repertoire of components (i. e. logic gates, protein molecules). He suggests that a sufficient condition requires *intention*. That is to say, we cannot argue that computation has taken place unless a system has arisen for a higher purpose (this is why he insists on intention as being a prerequisite for computation). Otherwise, almost everything is carrying out some form of computation (which is not a helpful point of view). Thus a Turing machine does not carry out computations unless it has been programmed to do so, and since natural evolution constructs organisms that have an increased chance of survival (the higher 'purpose') we can regard them as carrying out computations. It is in this sense that Toffoli points to the fundamental role of evolution in the definition of a computation as it has provided animals with the ability to have intention.

This brings us to one of the fundamental questions in computation. How can we program a physical system to perform a particular computation? The dominant method used to answer this question has been to construct logic gates and from these build a von Neumann machine (i. e. a digital computer). The mechanism that has been used to devise a computer program to carry out a particular computation is the familiar top-down design process, where ultimately the computation is represented using Boolean operations. According to Conrad this process leads us to pay "The Price of Programmability" [12], whereby in conventional programming and design we proceed by excluding many of the processes that may lead to us solving the problem at hand. Natural evolution does not do this. It is noteworthy that natural evolution has constructed systems of extraordinary sophistication, complexity and computational power. We argue that it is not possible to construct computational systems of such power using a conventional methodology and that complex software systems that directly utilize physical effects will require some form of search process akin to natural evolution together with a way of manipulating the properties of materials. We suggest that some form of evolution ought to be an appropriate methodology for arriving at *physical* systems that compute. In this chapter we discuss work that has adopted this methodology. We call it evolution in materio.

Evolutionary Algorithms

Firstly we propose that to overcome the limitations of a top-down design process, we should use a more unconstrained design technique that is more akin to a process of generate-and-test. However, a guided search method is also required that spends more time in areas of the search space that confer favorable traits for computation. One such approach is the use of evolutionary algorithms. These algorithms are inspired by the Darwinian concepts of survival of the fittest and the genetic inheritance of information. Using a computer, a population of randomly generated solutions is systematically tested, selected and modified until a solution has been found [13,14,15].

As in nature, a genetic algorithm optimizes a population of individuals by selecting the ones that are best suited to solving a problem and allowing their genetic make-up to propagate into future generations. It is typically guided only by the evolutionary process and often contains very limited domain specific knowledge. Although these algorithms are bio-inspired, it is important that any analogies drawn with nature are considered only as analogies.

Their lack of specialization for a problem makes genetic algorithms ideal search techniques where little is known about a problem. As long as a suitable representation is chosen along with a fitness function that allows for ease of movement around a search space, a GA can search vast problem spaces rapidly. Another feature of their behavior is that provided that the genetic representation chosen is sufficiently expressive the algorithm can explore potential solutions that are unconventional. A human de-

signer normally has a set of predefined rules and strategies that they adopt to solve a problem. These preconceptions may prevent trying a new method, and may prevent the designer using a better solution. A genetic algorithm does not necessarily require such domain knowledge. Evolutionary algorithms have been shown to be competitive or surpass human designed solutions in a number of different areas. The largest conference on evolutionary computation called GECCO has an annual session on evolutionary approaches that have produced human competitive scientific and technological results. Moreover the increase in computational power of computers makes such results increasingly more likely.

Many different versions of genetic algorithms exist. Variations in representations and genetic operators change the performance characteristics of the algorithm, and depending on the problem, people employ a variety of modifications of the basic algorithm. However, all the algorithms follow a similar basic set of steps.

Firstly the numbers or physical variables that are required to define a potential solution have to be identified and encoded into a data representation that can be manipulated inside a computer program. This is referred to as the encoding step. The representation chosen is of crucial importance as it is possible to inadvertedly choose overly constrained representations which limits the portions of the space of potential solutions that will be considered by the evolutionary algorithm. Generally the encoded information is referred to as a genotype and genotypes are sometimes divided into a number of separate strings called chromosomes. Each entry in the chromosome string is an allele, and one or more of these make up a gene.

The second step is to create inside the computer a number of independently generated genotypes whose alleles have been chosen with uniform probability from the allowed set of values. This collection of genotypes is called a population.

In its most basic form, an individual genotype is a single chromosome made of 1 s and 0 s. However, it is also common to use integer and floating-point numbers if they more appropriate for the task at hand. Combinations of different representations can also be used within the same chromosome, and that is the approach used in the work described in this article. Whatever representation is used, it should be able to adequately describe the individual and provide a mechanism where its characteristics can be transferred to future generations without loss of information.

Each of these individuals is then decoded into its phenotype, the outward, physical manifestation of the individual and tested to see how well the candidate solution solves the problem at hand. This is usually returned as a number that is referred to as the *fitness* of the genotype. Typically it is this phase in a genetic algorithm that is the most time consuming.

The next stage is to select what genetic information will proceed to the next generation. In nature the fitness function and selection are essentially the same – individuals that are better suited to the environment survive to reproduce and pass on their genes. In the genetic algorithm a procedure is applied to determine what information gets to proceed.

Genetic algorithms are often generational – where all the old population is removed before moving to the next generation, in nature this process is much less algorithmic. However, to increase the continuity of information between generations, some versions of the algorithm use elitism, where the fittest individuals are always selected for promotion to the next generation. This ensures that good solutions are not lost from the population, but it may have the side effect of causing the genetic information in the population to converge too quickly so that the search stagnates on a sub-optimal solution.

To generate the next population, a procedure analogous to sexual reproduction occurs. For example, two individuals will be selected and they will then have their genetic information combined together to produce the genotype for the offspring. This process is called recombination or crossover. The genotype is split into sections at randomly selected points called crossover points. A "simple" GA has only one of these points, however it is possible to perform this operation at multiple points.

Sections of the two chromosomes are then put together to form a new individual. This individual shares some of the characteristics of both parents. There are many different ways to choose which members of the population to breed with each other, the aim in general is to try and ensure that fit individuals get to reproduce with other fit individuals. Individuals can be selected with a probability proportional to their relative fitness or selected through some form of tournament, which may choose two or more chromosomes at random from the population and select the fittest.

In natural recombination, errors occur when the DNA is split and combined together. Also, errors in the DNA of a cell can occur at any time under the influence of a mutagen, such as radiation, a virus or toxic chemical. The genetic algorithm also has mutations. A number of alleles are selected at random and modified in some way. For a binary GA, the bit may be flipped, in a real-numbered GA a random value may be added to or subtracted from the previous allele.

Although GAs often have both mutation and crossover, it is possible to just use mutation. A mutation only approach has in some cases been demonstrated to work, and often crossover is seen as a macro mutation operator – effectively changing large sections of a chromosome.

After the previous operations have been carried out, the new individuals in the population are then retested and their new fitness scores calculated. Eventually this process leads to an increase in the average fitness of the population, and so the population moves closer toward a solution. This cycle of test, select and reproduce is continued until a solution is found (or some other termination condition is reached), at which point the algorithm stops. The performance of a genetic algorithm is normally measured in terms of the number of evaluations required to find a solution of a given quality.

Evolution in Materio: Historical Background

It is arguable that 'evolution in materio' began in 1958 in the work of Gordon Pask who worked on experiments to grow neural structures using electrochemical assemblages [16,17,18,19]. Gordon Pask's goal was to create a device sensitive to either sound or magnetic fields that could perform some form of signal processing – a kind of ear. He realized he needed a system that was rich in structural possibilities, and chose to use a metal solution. Using electric currents, wires can be made to self-assemble in an acidic aqueous metal-salt solution (e. g. ferrous sulphate). Changing the electric currents can alter the structure of these wires and their positions – the behavior of the system can be modified through external influence. Pask used an array of electrodes suspended in a dish containing the metal-salt solution, and by applying current (either transiently or a slowly changing source) was able to build iron wires that responded differently to two different frequencies of sound – 50 Hz and 100 Hz.

Pask had developed a system whereby he could manually train the wire formation in such a way that no complete specification had to be given – a complete paradigm shift from previous engineering techniques which would have dictated the position and behavior of every component in the system. His training technique relied on making changes to a set of resistors, and updating the values with given probabilities – in effect a test-randomly modify-test cycle. We would today recognize this algorithm as some form of evolutionary, hill climbing strategy – with the test stage as the fitness evaluation.

In 1996 Adrian Thompson started what we might call the modern era of evolution in materio. He was investigating whether it was possible to build working electronic circuits using unconstrained evolution (effectively, generate-and-test) using a re-configurable electronic silicon chip called an Field Programmable Gate Array (FPGA). Carrying out evolution by defining configurations of actual hardware components is known as *intrinsic* evolution. This is quite possible using FPGAs which are devices that have a two-dimensional array of logic functions that a configuration bit string defines and connects together. Thompson had set himself the task of evolving a digital circuit that could discriminate between an applied 1 kHz or 10 kHz applied signal [20,21]. He found that computer controlled evolution of the configuring bit strings could relatively easily solve this problem. However, when he analyzed the successful circuits he found to his surprise that they worked by utilizing subtle electrical properties of the silicon. Despite painstaking analysis and simulation work he was unable to explain how, or what property was being utilized. This lack of knowledge of how the system works, of course, prevents humans from designing systems that are intended to exploit these subtle and complex physical characteristics. However, it does not prevent exploitation through artificial evolution. Since then a number of researchers have demonstrated the viability of intrinsic evolution in silicon devices [21,22,23,24,25,26,27].

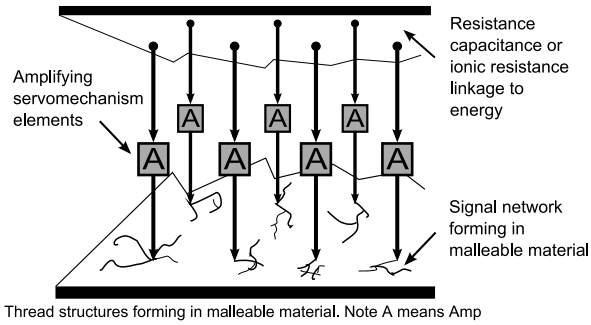

Evolution in Materio, Figure 1
Pask's experimental set up for growing dendritic wires in ferrous sulphate solution [17]

The term *evolution in materio* was first coined by Miller and Downing [28]. They argued that the lesson that should be drawn from the work of [21] is that evolution may be used to exploit the properties of a wider range of materials than silicon.

In summary, evolution in materio can be described as:

> Exploitation, using an unconstrained evolutionary algorithm, of the non-linear properties of a malleable or programmable material to perform a desired function by altering its physical or electrical configuration.

Evolution in materio is a subset of a research field known as evolvable hardware. It aims to exploit properties of physical systems with much fewer preconditions and constraints than is usual, and it deliberately tries to avoid paying Conrad's 'Price of Programmability'. However, to get access to physically rich systems, we may have to discard devices designed with human programming in mind. Such devices are often based on abstract idealizations of processes occurring in the physical world. For example, FPGAs are considered as digital, but they are fundamentally analogue devices that have been constrained to behave in certain, human understandable ways. This means that intrinsically complex physical processes are carefully manipulated to represent extremely simple effects (e. g. a rapid switch from one voltage level to another). Unconstrained evolution, as demonstrated by Thompson, allows for the analogue properties of such devices to be effectively utilized.

We would expect physically rich systems to exhibit non-linear properties – they will be complex systems. This is because physical systems generally have huge numbers of parts interacting in complex ways. Arguably, humans have difficulty working with complex systems, and the use of evolution enables us to potentially overcome these limitations when dealing with such systems.

When systems are abstracted, the relationship to the physical world becomes more distant. This is highly convenient for human designers who do not wish to understand, or work with, hidden or subtle properties of materials. Exploitation through evolution reduces the need for abstraction, as it appears evolution is capable of discovering and utilizing any physical effects it can find. The aim of this new methodology in computation is to evolve special purpose computational processors. By directly exploiting physical systems and processes, one should be able to build extremely fast and efficient computational devices. It is our view that computer controlled evolution is a universal methodology for doing this. Of course, von Neumann machines (i. e. digital computers) are *individually* universal and this is precisely what confers their great utility in modern technology, however this universality comes at a price. They ignore the rich computational possibilities of materials and try to create operations that are close to a mathematical abstraction. Evolution in materio is a universal methodology for producing specific, highly tuned computational devices.

It is important not to underestimate the real practical difficulties associated with using an unconstrained design process. Firstly the evolved behavior of the material may be extremely sensitive to the specific properties of the material sample, so each piece would require individual training. Thompson originally experienced this difficulty, however in later work he showed that it was possible to evolve the configuration of FPGAs so that they produced reliable behavior in a variety of environmental conditions [29].

Secondly, the evolutionary algorithm may utilize physical aspects of any part of the training set-up. Both of these difficulties have already been experienced [21,23]. A third problem can be thought of as "the wiring problem". The means to supply huge amounts of configuration data to a tiny sample. This problem is a very fundamental one. It suggests that if we wish to exploit the full physical richness of materials we might have to allow the material to grow its own wires and be self-wiring. This has profound implications for intrinsic evolution as artificial hardware evolution requires complete reconfigurability, this implies that one would have to be able to "wipe-clean" the evolved wiring and start again with a new artificial genotype. This might be possible by using nanoparticles that assemble into nanowires. These considerations bring us to an important issue in evolution in materio. Namely, the problem of choosing a suitable materials that can be exploited by computer controlled evolution.

Evolution in Materio: Defining Suitable Materials

The obvious characteristic required by a candidate material is the ability to reconfigure it in some way. Liquid crystal, clay, salt solutions etc can be readily configured either electrically or mechanically; their physical state can be adjusted, and readjusted, by applying a signal or force. In contrast (excluding its electrical properties) the physical properties of an FPGA would remain unchanged during configuration. It is also desirable to bulk configure the system. It would be infeasible to configure every molecule in the material, so the material should support the ability to be reconfigured over large areas using a small amount of configuration.

The material needs to perform some form of transformation (or computation) on incident signals that we apply. To do this, the material will have to interfere with the incident signal and perform a modification to it. We will need to be able to observe this modification, in order to extract the result of the computation. To perform a non-trivial computation, the material should be capable of performing complex operations upon the signal. Such capabilities would be maximized if the system exhibited non-linear behavior when interacting with input signals.

In summary, we can say that for a material to be useful to evolution in materio it should have the following properties:

- Modify incident signals in observable ways.
- The components of a system (i. e. the molecules within a material) interact with each other locally such that non-linear effects occur at either the local or global levels.
- It is possible to configure the state of the material locally.
- It is possible to observe the state of the material – either as a whole or in one or more locations.
- For practical reasons we can state that the material should be reconfigurable, and that changes in state should be temporary or reversible.

Miller and Downing [28] identified a number of physical systems that have some, if not all, of these desirable properties. They identified liquid crystal as the most promising in this regard as it digitally writable, reconfigurable and works at a molecular level. Most interestingly, it is an example of mesoscopic organization. Some people have argued that it is within such systems that emergent, organized behavior can occur [30]. Liquid crystals also exhibit the phenomenon of self-assembly. They form a class of substances that are being designed and developed in a field of chemistry called Supramolecular Chemistry [31]. This is a new and exciting branch of chemistry that can be characterized as 'the designed chemistry of the intermolecular bond'. Supramolecular chemicals are in a permanent process of being assembled and disassembled. It is interesting to consider that conceptually liquid crystals appear to sit on the 'edge of chaos' [32] in that they are fluids (chaotic) that can be ordered, under certain circumstances.

Liquid Crystal

Liquid crystal (LC) is commonly defined as a substance that can exist in a mesomorphic state [33,34]. Mesomorphic states have a degree of molecular order that lies between that of a solid crystal (long-range positional and orientational) and a liquid, gas or amorphous solid (no long-range order). In LC there is long-range orientational order but no long-range positional order.

LC tends to be transparent in the visible and near infrared and quite absorptive in UV. There are three distinct types of LC: lyotropic, polymeric and thermotropic. Lyotropic LC is obtained when an appropriate amount of material is dissolved in a solvent. Most commonly this is formed by water and amphiphilic molecules: molecules with a hydrophobic part (water insoluble) and a hydrophillic part (strongly interacting with water). Polymeric LC is basically a polymer version of the aromatic LC discussed. They are characterized by high viscosity and include vinyls and Kevlar. Thermotropic LC (TLC) is the most common form and is widely used. TLC exhibit various liquid crystalline phases as a function of temperature. They can be depicted as rod-like molecules and interact with each other in distinctive ordered structures. TLC exists in three main forms: nematic, cholesteric and smectic. In nematic LC the molecules are positionally arranged randomly but they all share a common alignment axis. Cholesteric LC (or chiral nematic) is like nematic however they have a chiral orientation. In smectic LC there is typically a layered positionally disordered structure. The three types A, B and C are defined as follows. In type A the molecules are oriented in alignment with the natural physical axes (i. e normal to the glass container), however in type C the common molecular axes of orientation is at an angle to the container. LC molecules typically are dipolar. Thus the organization of the molecular dipoles give another order of symmetry to the LC. Normally the dipoles would be randomly oriented. However in some forms the natural molecular dipoles are aligned with one another. This gives rise to ferroelectric and ferrielectric forms.

There is a vast range of different types of liquid crystal. LC of different types can be mixed. LC can be doped (as in Dye-Doped LC) to alter their light absorption characteristics. Dye-Doped LC film has been made that is optically addressable and can undergo very large changes in refractive index [35]. There are Polymer-Dispersed Liquid Crystals, which can have tailored, electrically controlled light refractive properties. Another interesting form of LC being actively investigated is Discotic LC. These have the form of disordered stacks (1-dimensional fluids) of disc-shaped molecules on a two-dimensional lattice. Although discotic LC is an electrical insulator, it can be made to conduct by doping with oxidants [36]. The oxidants are incorporated into the fluid hydrocarbon chain matrix (between disks). LC is widely known as useful in electronic displays, however, there are in fact, many non-display applications too. There are many applications of LC (especially ferroelectric LC) to electrically controlled light modulation: phase modulation, optical correlation, optical interconnects and switches, wavelength filters, optical neural networks. In the latter case a ferroelectric LC is used to encode the weights in a neural network [37].

Conducting and Electroactive Polymers

Conducting polymer composites have been made that rapidly change their microwave reflection coefficient when an electric field is applied. When the field is removed, the composite reverts to its original state. Experiments have shown that the composite can change from one state to the other in the order of 100 ms [38]. Also, some poly-

mers exhibit electrochromism. These substances change their reflectance when a voltage is applied. This can be reversed by a change in voltage polarity [39]. Electroactive polymers [40] are polymers that change their volume with the application of an electric field. They are particularly interesting as voltage controlled artificial muscle. Organic semiconductors also look promising especially when some damage is introduced. Further details of electronic properties of polymers and organic crystals can be found in [41].

Voltage Controlled Colloids

Colloids are suspensions of particles of sub-micron sizes in a liquid. The phase behavior of colloids is not fully understood. Simple colloids can self assemble into crystals, while multi-component suspensions can exhibit a rich variety of crystalline structures. There are also electrorheological fluids. These are suspensions of extremely fine nonconducting particles in an electrically insulating fluid. The viscosity of these fluids can be changed in a reversible way by large factors in response to an applied electric field in times of the order of milliseconds [42]. Also colloids can also be made in which the particles are charged making them easily manipulatable by suitable applied electric fields. Even if the particles are not charged they may be moved through the action of applied fields using a phenomenon known as dielectrophoresis which is the motion of polarized but electrically uncharged particles in nonuniform electric fields [43]. In work that echoes the methods of Pask nearly four decades ago, dielectrophoresis has been used to grow tiny gold wires through a process of self-assembly [44].

Langmuir–Blodgett Films

Langmuir–Blodgett films are molecular monolayers of organic material that can be transferred to a solid substrate [45]. They usually consist of hydrophillic heads and hydrophobic tails attached to the substrate. Multiple monolayers can be built and films can be built with very accurate and regular thicknesses. By arranging an electrode layer above the film it seems feasible that the local electronic properties of the layers could be altered. These systems look like feasible systems whose properties might be exploitable through computer controlled evolution of the voltages.

Kirchoff–Lukasiewicz Machines

Work by Mills [46,47] also demonstrates the use of materials in computation. He has designed an 'Extended Analog Computer' (EAC) that is a physical implementation of

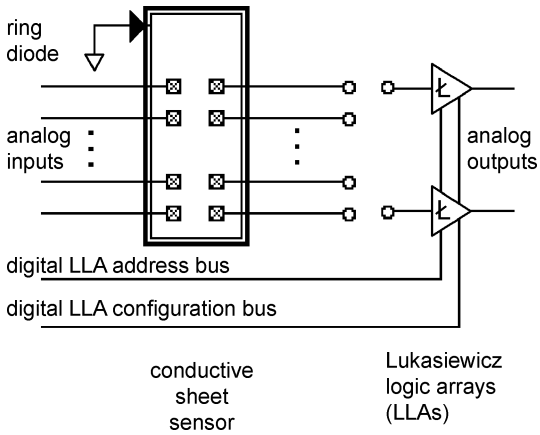

Evolution in Materio, Figure 2
Kirchhoff–Lukasiewicz Machine

a Kirchhoff–Lukasiewicz Machine (KLM) [46]. The machines are composed of logical function units connected to a conductive media, typically a conductive polymer sheet. The logical units implement Lukasiewicz Logic – a type of multi-valued logic [47]. Figure 2 shows how the Lukasiewicz Logic Arrays (LLA) are connected to the conductive polymer. The LLA bridge areas of the sheet together. The logic units measure the current at one point, perform a transformation and then apply a current source to the other end of the bridge.

Computation is performed by applying current sinks and sources to the conductive polymer and reading the output from the LLAs. Different computations can be performed that are determined by the location of applied signals in the conducting sheet and the configuration of the LLAs. Hence, computation is performed by an interaction of the physics described by Kirchoff's laws and the Lukasiewicz Logic units. Together they form a physical device that can solve certain kinds of partial differential equations. Using this form of analogue computation, a large number of these equations can be solved in nanoseconds – much faster than on a conventional computer. The speed of computation is dependent on materials used and how they are interfaced to digital computers, but it is expected that silicon implementations will be capable of finding tens of millions of solutions to the equations per second.

Examples of computation so far implemented in this system include robot control, control of a cyclotron beam [48], models of biological systems (including neural networks) [49] and radiosity based image rendering.

One of the most interesting feature of these devices is the programming method. It is very difficult to understand the actual processes used by the system to perform

computation, and until recently most of the reconfiguration has been done manually. This is difficult as the system is not amenable to traditional software development approaches. However, evolutionary algorithms can be used to automatically define the parameters of the LLAs and the placement of current sinks and sources. By defining a suitable fitness function, the configuration of the EAC can be evolved – which removes the need for human interaction and for knowledge of the underlying system.

Although it is clear that such KLMs are clearly using the physical properties of a material to perform computation, the physical state of the material is not reconfigured (i. e., programmed), only the currents in the sheet are changed.

Evolution in Materio Is Verified with Liquid Crystal

Harding [50] has verified Miller's intuition about the suitability of liquid crystal as an evolvable material by demonstrating that it is relatively easy to configure liquid crystal to perform various forms of computation.

In 2004, Harding constructed an analogue processor that utilizes the physical properties of liquid crystal for computation. He evolved the configuration of the liquid crystal to discriminate between two square waves of many different frequencies. This demonstrated, for the first time, that the principle of using computer-controlled evolution was a viable and powerful technique for using non-silicon materials for computation. The analogue processor consists of a passive liquid crystal display mounted on a reconfigurable circuit, known as an evolvable motherboard. The motherboard allows signals and configuration voltages to be routed to physical locations in the liquid crystal.

Harding has shown that many different devices can be evolved in liquid crystal including:

- Tone discriminator. A device was evolved in liquid crystal that could differentiate many different frequencies of square wave. The results were competitive, if not superior to those evolved in the FPGA.
- Logic gates. A variety of two input logic gates were evolved, showing that liquid crystal could behave in a digital fashion. This indicates that liquid crystal is capable of universal computation.
- Robot controller. An obstacle avoidance system for a simple exploratory robot was evolved. The results were highly competitive, with solutions taking fewer evaluations to find compared to other work on evolved robot controllers.

One of the surprising findings in this work has been that it turns out to be relatively easy to evolve the configuration of liquid crystal to solve tasks; i. e., only 40 generations of a modest population of configurations are required to evolve a very good frequency discriminator, compared to the thousands of generations required to evolve a similar circuit on an FPGA. This work has shown that evolving such devices in liquid crystal is easier than when using conventional components, such as FPGAs. The work is a clear demonstration that evolutionary design can produce solutions that are beyond the scope of human design.

Evolution in Materio Using Liquid Crystal: Implementational Details

An evolvable motherboard (EM) [23] is a circuit that can be used to investigate intrinsic evolution. The EM is a reconfigurable circuit that rewires a circuit under computer control. Previous EMs have been used to evolve circuits containing electronic components [23,51] – however they can also be used to evolve in materio by replacing the standard components with a candidate material.

An EM is connected to an Evolvatron. This is essentially a PC that is used to control the evolutionary processes. The Evolvatron also has digital and analog I/O, and can be used to provide test signals and record the response of the material under evolution.

The Liquid Crystal Evolvable Motherboard (LCEM) is a circuit that uses four cross-switch matrix devices to dynamically configure the circuits connecting to the liquid crystal. The switches are used to wire the 64 connections on the LCD to one of 8 external connections. The external connections are: input voltages, grounding, signals and connections to measurement devices. Each of the external connectors can be wired to any of the connections to the LCD.

The external connections of the LCEM are connected to the Evolvatron's analogue inputs and outputs. One connection was assigned for the incident signal, one for measurement and the other for fixed voltages. The value of the fixed voltages is determined by the evolutionary algorithm, but is constant throughout each evaluation.

In these experiments the liquid crystal glass sandwich was removed from the display controller it was originally mounted on, and placed on the LCEM. The display has a large number of connections (in excess of 200), however because of PCB manufacturing constraints we are limited in the size of connection we can make, and hence the number of connections. The LCD is therefore roughly positioned over the pads on the PCB, with many of the PCB pads touching more than one of the connectors on the LCD. This means that we are applying configuration voltages to several areas of LC at the same time.

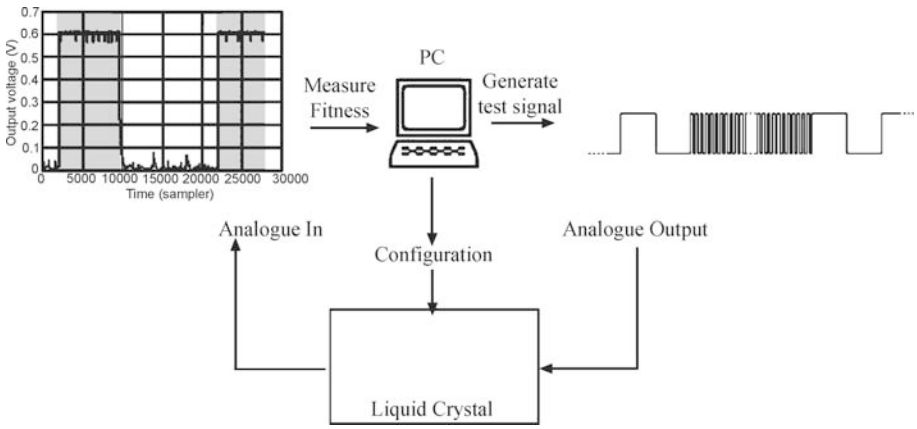

Evolution in Materio, Figure 3
Equipment configuration

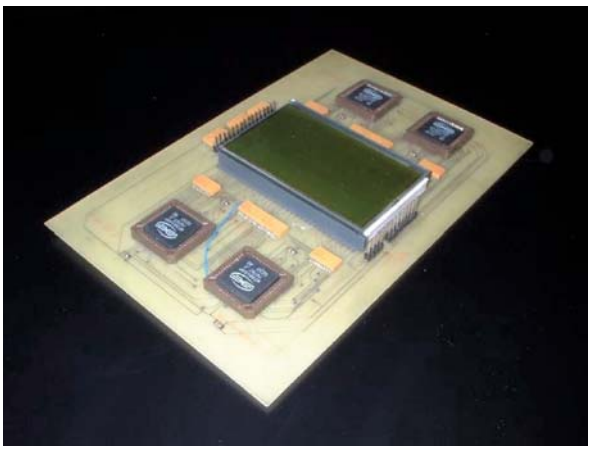

Evolution in Materio, Figure 4
The LCEM

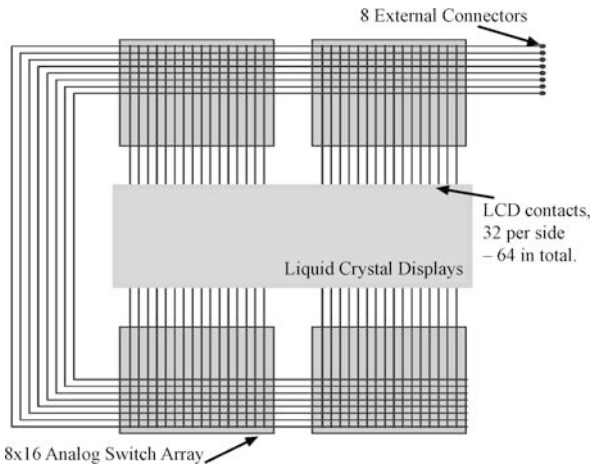

Evolution in Materio, Figure 5
Schematic of LCEM

Unfortunately neither the internal structure nor the electrical characteristics of the LCD are known. This raises the possibility that a configuration may be applied that would damage the device. The wires inside the LCD are made of an extremely thin material that could easily be burnt out if too much current flows through them. To guard against this, each connection to the LCD is made through a 4.7 Kohm resistor in order to provide protection against short circuits and to help limit the current in the LCD. The current supplied to the LCD is limited to 100 mA. The software controlling the evolution is also responsible for avoiding configurations that may endanger the device (such as short circuits).

It is important to note that other than the control circuitry for the switch arrays there are no other active components on the motherboard – only analog switches, smoothing capacitors, resistors and the LCD are present.

Stability and Repeatability Issues

When the liquid crystal display is observed while solving a problem it is seen that some regions of the liquid display go dark indicating that the local molecular direction has been changed. This means that the configuration of the liquid crystal is changing while signals are being applied. To draw an analogy with circuit design, the incident signals would be changing component values or changing the circuit topology, which would have an affect on the behavior of the system. This is likely to be detrimental to the measured performance of the circuit. When a solution is evolved, the fitness function automatically measures it stability over the period of the evaluation. Changes made by the incident signals can be considered part of the genotype-phenotype mapping. Solutions that cannot cope with

their initial configurations being altered will achieve a low score. However, the fitness function cannot measure the behavior beyond the end of the evaluation time. Therein lies the difficulty, in evolution in materio long term stability cannot be guaranteed.

Another issue concerns repeatability. When a configuration is applied to the liquid crystal the molecules are unlikely go back to exactly where they were when this configuration was tried previously. Assuming, that there is a strong correlation between genotype and phenotype, then it is likely that evolution will cope with this extra noise. However, if evolved devices are to be useful one needs to be sure that previously evolved devices will function in the same way as they did when originally evolved.

In [27] it is noted that the behavior of circuits evolved intrinsically can be influenced by previous configurations – therefore their behavior (and hence fitness) is dependent not only on the currently evaluated individual's configuration but on those that came before. It is worth noting that this is precisely what happens in natural evolution. For example, in a circuit, capacitors may still hold charge from a previously tested circuit. This charge would then affect the circuits operation, however if the circuit was tested again with no stored charge a different behavior would be expected and a different fitness score would be obtained. Not only does this affect the ability to evolve circuits, but would mean that some circuits are not valid. Without the influence of the previously evaluated circuits the current solution may not function as expected. It is expected that such problems will have analogies in evolution in materio. The configurations are likely to be highly sensitive to initial conditions (i. e. conditions introduced by previous configurations).

Dealing with Environmental Issues

A major problem when working with intrinsic evolution is separating out the computation allegedly being carried out by the target device, and that actually done by the material being used. For example, whilst trying to evolve an oscillator Bird and Layzell discovered that evolution was using part of the circuit for a radio antenna, and picking up emissions from the environment [22]. Layzell also found that evolved circuits were sensitive to whether or not a soldering iron was plugged in (not even switched on) in another part of the room [23]!

An evolved device is not useful if it is highly sensitive to its environment in unpredictable ways, and it will not always be clear what environmental effects the system is using. It would be unfortunate to evolve a device for use in a space craft, only to find out it fails to work once out of range of a local radio tower!

To minimize these risks, we will need to check the operation of evolved systems under different conditions. We will need to test the behavior of a device using a different set up in a different location. It will be important to know if a particular configuration only works with one particular sample of a given material.

The Computational Power of Materials

In [52], Lloyd argued that the theoretical computing power of a kilogram of material is far more than is possible with a kilogram of traditional computer. He notes that computers are subject to the laws of physics, and that these laws place limits on the maximum speed they can operate and the amount of information it can process. Lloyd shows that if we were fully able to exploit a material, we would get an enormous increase in computing power. For example, with 1 kg of matter we should be able to perform roughly 5×10^{50} operations per second, and store 10^{31} bits. Amazingly, contemporary quantum computers do operate near these theoretical limits [52].

A small amount of material also contains a large number of components (regardless of whether we consider the molecular or atomic scale). This leads to some interesting thoughts. If we can exploit materials at this level, we would be able to do a vast amount of computation in a small volume. A small size also hints at low power consumption, as less energy has to be spent to perform an operation. Many components also provide a mechanism for reliability through redundancy. A particularly interesting observation, especially when considered in terms of non Von-Neumann computation, is the massive parallelism we may be able to achieve. The reason that systems such as quantum, DNA and chemical computation can operate so quickly is that many operations are performed at the same time. A programmable material might be capable of performing vast numbers of tasks simultaneously, and therefore provide a computational advantage.

In commercial terms, small is often synonymous with low cost. It may be possible to construct devices using cheaply available materials. Reliability may not be an issue, as the systems could be evolved to be massively fault tolerant using their intrinsic redundancy. Evolution is capable of producing novel designs. Koza has already rediscovered circuits that infringe on recent patents, and his genetic programming method has 'invented' brand new circuit designs [53]. Evolving in materio could produce many novel designs, and indeed given the infancy of pro-

grammable materials all designs may be unique and hence patentable.

Future Directions

The work described here concerning liquid crystal computational devices is at an early stage. We have merely demonstrated that it is possible to evolve configurations of voltages that allow a material to perform desired computations. Any application that ensues from this work is unlikely to be a replacement for a simple electronic circuit. We can design and build those very successfully. What we have difficulty with is building complex, fault tolerant systems for performing complex computation. It appears that nature managed to do this. It used a simple process of a repetitive test and modify, and it did this in a universe of unimaginable physical complexity. If nature can exploit the physical properties of a material and its surroundings through evolution, then so should we.

There are many important issues that remain to be addressed. Although we have made some suggestions about materials worthy of investigation, it is at present unclear which materials are most suitable. An experimental platform needs to be constructed that allows many materials to be tried and investigated. The use of microelectrode arrays in a small volume container would allow this. This would also have the virtue of allowing internal signals in the materials to be inspected and potentially understood.

We need materials that are rapidly configurable. They must not be fragile and sensitive to minute changes in physical setup. They must be capable of maintaining themselves in a stable configuration. The materials should be complex and allow us to carry out difficult computations more easily than conventional means. One would like materials that can be packaged into small volumes. The materials should be relatively easily interfaced with. So far, material systems have been configured by applying a constant configuration pattern, however this may not be appropriate for all systems. It may be necessary to put the physical system under some form of responsive control, in order to program and then keep the behavior stable.

We may or may not know if a particular material can be used to perform some form of computation. However, we can treat our material as a "black box", and using evolution as a search technique, automatically discover what, if any, computations our black box can perform. The first step is to build an interface that will allow us to communicate with a material. Then we will use evolution to find a configuration we can apply using this platform, and then attempt to find a mapping from a given problem to an input suitable for that material, and a mapping from the materials response to an output. If this is done correctly, we might be automatically able to tell if a material can perform computation, and then classify the computation.

When we evolve in materio, using mappings evolved in software, how can we tell when the material is giving us any real benefit? The lesson of evolution in materio has been that the evolved systems can be very difficult to analyze, and the principal obstacle to the analysis is the problem of separating out the computational role that each component plays in the evolved system. These issues are by no means just a problem for evolution in materio. They may be an inherent part of complex evolved systems. Certainly the understanding of biological systems are providing immense challenges to scientists.

The single most important aspect that suggests that evolution in materio has a future is that natural evolution has produced immensely sophisticated material computational systems. It would seem foolish to ignore this and merely try to construct computational devices that operate according to one paradigm of computation (i. e. Turing). Oddly enough, it is precisely the sophistication of the latter that allows us to attempt the former.

Bibliography

Primary Literature

1. Turing AM (1936) On computable numbers, with an application to the entscheidungsproblem. Proc Lond Math Soc 42(2):230–265
2. Bissell C (2004) A great disappearing act: the electronic analogue computer. In: IEEE Conference on the History of Electronics, 28–30 June
3. Deutsch D (1985) Quantum theory, the church–turing principle and the universal quantum computer. Proc Royal Soc Lond A 400:97–117
4. Adamatzky A, Costello BDL, Asai T (2005) Reaction-Diffusion Computers. Elsevier, Amsterdam
5. Adleman LM (1994) Molecular computation of solutions to combinatorial problems. Science 266(11):1021–1024
6. Amos M (2005) Theoretical and Experimental DNA Computation. Springer, Berlin
7. Weiss R, Basu S, Hooshangi S, Kalmbach A, Karig D, Mehreja R, Netravali I (2003) Genetic circuit building blocks for cellular computation, communications, and signal processing. Nat Comput 2(1):47–84
8. UK Computing Research Committee (2005) Grand challenges in computer research. http://www.ukcrc.org.uk/grand_challenges/
9. Stepney S, Braunstein SL, Clark JA, Tyrrell A, Adamatzky A, Smith RE, Addis T, Johnson C, Timmis J, Welch P, Milner R, Partridge D (2005) Journeys in non-classical computation I: A grand challenge for computing research. Int J Parallel Emerg Distrib Syst 20(1):5–19
10. Stepney S, Braunstein S, Clark J, Tyrrell A, Adamatzky A, Smith R, Addis T, Johnson C, Timmis J, Welch P, Milner R, Partridge D

(2006) Journeys in non-classical computation II: Initial journeys and waypoints. Int J Parallel, Emerg Distrib Syst 21(2):97–125
11. Toffoli T (2005) Nothing makes sense in computing except in the light of evolution. Int J Unconv Comput 1(1):3–29
12. Conrad M (1988) The price of programmability. The Universal Turing Machine. pp 285–307
13. Goldberg D (1989) Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading, Massachusetts
14. Holland J (1992) Adaptation in Natural and Artificial Systems. 2nd edn. MIT Press, Cambridge
15. Mitchell M (1996) An introduction to genetic algorithms. MIT Press, Cambridge, MA, USA
16. Pask G (1958) Physical analogues to the growth of a concept. In: Mechanization of Thought Processes. Symposium 10, National Physical Laboratory, pp 765–794
17. Pask G (1959) The natural history of networks. In: Proceedings of International Tracts. In: Computer Science and Technology and their Application. vol 2, pp 232–263
18. Cariani P (1993) To evolve an ear: epistemological implications of gordon pask's electrochemical devices. Syst Res 3:19–33
19. Pickering A (2002) Cybernetics and the mangle: Ashby, beer and pask. Soc Stud Sci 32:413–437
20. Thompson A, Harvey I, Husbands P (1996) Unconstrained evolution and hard consequences. In: Sanchez E, Tomassini M (eds) Towards Evolvable Hardware: The evolutionary engineering approach. LNCS, vol 1062. Springer, Berlin, pp 136–165
21. Thompson A (1996) An evolved circuit, intrinsic in silicon, entwined with physics. ICES 390–405
22. Bird J, Layzell P (2002) The evolved radio and its implications for modelling the evolution of novel sensors. In: Proceedings of Congress on Evolutionary Computation, pp 1836–1841
23. Layzell P (1998) A new research tool for intrinsic hardware evolution. Proceedings of The Second International Conference on Evolvable Systems: From Biology to Hardware. LNCS, vol 1478. Springer, Berlin, pp 47–56
24. Linden DS, Altshuler EE (2001) A system for evolving antennas in-situ. In: 3rd NASA / DoD Workshop on Evolvable Hardware. IEEE Computer Society, pp 249–255
25. Linden DS, Altshuler EE (1999) Evolving wire antennas using genetic algorithms: A review. In: 1st NASA / DoD Workshop on Evolvable Hardware. IEEE Computer Society, pp 225–232
26. Stoica A, Zebulum RS, Guo X, Keymeulen D, Ferguson MI, Duong V (2003) Silicon validation of evolution-designed circuits. In: Proceedings. NASA/DoD Conference on Evolvable Hardware, pp 21–25
27. Stoica A, Zebulum RS, Keymeulen D (2000) Mixtrinsic evolution. In: Proceedings of the Third International Conference on Evolvable Systems: From Biology to Hardware (ICES2000). Lecture Notes in Computer Science, vol 1801. Springer, Berlin, pp 208–217
28. Miller JF, Downing K (2002) Evolution in materio: Looking beyond the silicon box. In: Proceedings of NASA/DoD Evolvable Hardware Workshop, pp 167–176
29. Thompson A (1998) On the automatic design of robust electronics through artificial evolution. In: Sipper M, Mange D, Pérez-Uribe A (eds) Evolvable Systems: From Biology to Hardware, vol 1478. Springer, New York, pp 13–24
30. Laughlin RB, Pines D, Schmalian J, Stojkovic BP, Wolynes P (2000) The middle way. Proc Nat Acd Sci 97(1):32–37
31. Lindoy LF, Atkinson IM (2000) Self-assembly in Supramolecular Systems. Royal Society of Chemistry
32. Langton C (1991) Computation at the edge of chaos: Phase transitions and emergent computation. In: Emergent Computation, pp 12–37. MIT Press
33. Demus D, Goodby JW, Gray GW, Spiess HW, Vill V (eds) (1998) Handbook of Liquid Crystals, vol 4. Wiley-VCH, ISBN 3-527-29502-X, pp 2180
34. Khoo IC (1995) Liquid Crystals: physical properties and nonlinear optical phenomena. Wiley
35. Khoo IC, Slussarenko S, Guenther BD, Shih MY, Chen P, Wood WV (1998) Optically induced space-charge fields, dc voltage, and extraordinarily large nonlinearity in dye-doped nematic liquid crystals. Opt Lett 23(4):253–255
36. Chandrasekhar S (1998) Columnar, discotic nematic and lamellar liquid crystals: Their structure and physical properties. In: Handbook of Liquid Crystals, vol 2B. Wiley-VCH pp 749–780
37. Crossland WA, Wilkinson TD (1998) Nondisplay applications of liquid crystals. In: Handbook of Liquid Crystals, vol 1. Wiley-VCH, pp 763–822
38. Wright PV, Chambers B, Barnes A, Lees K, Despotakis A (2000) Progress in smart microwave materials and structures. Smart Mater Struct 9:272–279
39. Mortimer RJ (1997) Electrochromic materials. Chem Soc Rev 26:147–156
40. Bar-Cohen Y (2001) Electroactive Polymer (EAP) Actuators as Artificial Muscles – Reality, Potential and Challenges. SPIE Press
41. Pope M, Swenberg CE (1999) Electronic Processes of Organic Crystals and Polymers. Oxford University Press, Oxford
42. Hao T (2005) Electrorheological Fluids: The Non-aqueous Suspensions. Elsevier Science
43. Khusid B, Activos A (1996) Effects of interparticle electric interactions on dielectrophoresis in colloidal suspensions. Phys Rev E 54(5):5428–5435
44. Khusid B, Activos A (2001) Hermanson KD, Lumsdon SO, Williams JP, Kaler EW, Velev OD. Science 294:1082–1086
45. Petty MC (1996) Langmuir–Blodgett Films: An Introduction. Cambridge University Press, Cambridge
46. Mills JW (1995) Polymer processors. Technical Report TR580, Department of Computer Science, University of Indiana
47. Mills JW, Beavers MG, Daffinger CA (1989) Lukasiewicz logic arrays. Technical Report TR296, Department of Computer Science, University of Indiana
48. Mills JW (1995) Programmable vlsi extended analog computer for cyclotron beam control. Technical Report TR441, Department of Computer Science, University of Indiana
49. Mills JW (1995) The continuous retina: Image processing with a single sensor artificial neural field network. Technical Report TR443, Department of Computer Science, University of Indiana
50. Harding S, Miller JF (2004) Evolution in materio: A tone discriminator in liquid crystal. In: In Proceedings of the Congress on Evolutionary Computation 2004 (CEC'2004), vol 2, pp 1800–1807
51. Crooks J (2002) Evolvable analogue hardware. Meng project report, The University Of York
52. Lloyd S (2000) Ultimate physical limits to computation. Nature 406:1047–1054
53. Koza JR (1999) Human-competitive machine intelligence by means of genetic algorithms. In: Booker L, Forrest S, Mitchell M, Riolo R (eds) Festschrift in honor of John H Holland. Center for the Study of Complex Systems, Ann Arbor, pp 15–22

Books and Reviews

Analog Computer Museum and History Center: Analog Computer Reading List. http://dcoward.best.vwh.net/analog/readlist.htm

Bringsjord S (2001) In Computation, Parallel is Nothing, Physical Everything. Minds and Machines 11(1)

Feynman RP (2000) Feyman Lectures on Computation. Perseus Books Group

Fifer S (1961) Analogue computation: theory, techniques, and applications. McGraw-Hill, New York

Greenwood GW, Tyrrell AM (2006) Introduction to Evolvable Hardware: A Practical Guide for Designing Self-Adaptive Systems. Wiley-IEEE Press

Hey AJG (ed) (2002) Feynman and Computation. Westview Press

Penrose R (1989) The Emperor's New Mind, Concerning Computers, Minds, and the Laws of Physics. Oxford University, Oxford

Piccinini G: The Physical Church–Turing Thesis: Modest or Bold? http://www.umsl.edu/~piccininig/CTModestorBold5.htm

Raichman N, Ben-Jacob N, Segev R (2003) Evolvable Hardware: Genetic Search in a Physical Realm. Phys A 326:265–285

Sekanina L (2004) Evolvable Components: From Theory to Hardware Implementations, 1st edn. Springer, Heidelberg

Siegelmann HT (1999) Neural Networks and Analog Computation, Beyond the Turing Limits. Birkhauser, Boston

Sienko T, Adamatzky A, Rambidi N, Conrad M (2003) Molecular Computing. MIT Press

Thompson A (1999) Hardware Evolution: Automatic Design of Electronic Circuits in Reconfigurable Hardware by Artificial Evolution, 1st edn. Springer, Heidelberg

Evolving Cellular Automata

Martin Cenek, Melanie Mitchell
Computer Science Department,
Portland State University, Portland, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Cellular Automata
Computation in CAs
Evolving Cellular Automata with Genetic Algorithms
Previous Work on Evolving CAs
Coevolution
Other Applications
Future Directions
Acknowledgments
Bibliography

Glossary

Cellular automaton (CA) Discrete-space and discrete-time spatially extended lattice of cells connected in a regular pattern. Each cell stores its state and a state-transition function. At each time step, each cell applies the transition function to update its state based on its local neighborhood of cell states. The update of the system is performed in synchronous steps – i. e., all cells update simultaneously.

Cellular programming A variation of genetic algorithms designed to simultaneously evolve state transition rules and local neighborhood connection topologies for non-homogeneous cellular automata.

Coevolution An extension to the genetic algorithm in which candidate solutions and their "environment" (typically test cases) are evolved simultaneously.

Density classification A computational task for binary CAs: the desired behavior for the CA is to iterate to an all-1s configuration if the initial configuration has a majority of cells in state 1, and to an all-0s configuration otherwise.

Genetic algorithm (GA) A stochastic search method inspired by the Darwinian model of evolution. A population of candidate solutions is evolved by reproduction with variation, followed by selection, for a number of generations.

Genetic programming A variation of genetic algorithms that evolves genetic trees.

Genetic tree Tree-like representation of a transition function, used by genetic programming algorithm.

Lookup table (LUT) Fixed-length table representation of a transition function.

Neighborhood Pattern of connectivity specifying to which other cells each cell is connected.

Non-homogeneous cellular automaton A CA in which each cell can have its own distinct transition function and local neighborhood connection pattern.

Ordering A computational task for one-dimensional binary CAs with fixed boundaries: The desired behavior is for the CA to iterate to a final configuration in which all initial 0 states migrate to the left-hand side of the lattice and all initial 1 states migrate to the right-hand side of the lattice.

Particle Periodic, temporally coherent boundary between two regular domains in a set of successive CA configurations. Particles can be interpreted as carrying information about the neighboring domains. Collisions between particles can be interpreted as the processing of information, with the resulting information carried by new particles formed by the collision.

Regular domain Region defined by a set of successive CA configurations that can be described by a simple regular language.

Synchronization A computational task for binary CAs: the desired behavior for the CA is to iterate to a tem-

poral oscillation between two configurations: all cells have state 1 and all cells have state 0s.

Transition function Maps a local neighborhood of cell states to an update state for the center cell of that neighborhood.

Definition of the Subject

Evolving cellular automata refers to the application of evolutionary computation methods to evolve cellular automata transition rules. This has been used as one approach to automatically "programming" cellular automata to perform desired computations, and as an approach to model the evolution of collective behavior in complex systems.

Introduction

In recent years, the theory and application of cellular automata (CAs) has experienced a renaissance, due to advances in the related fields of reconfigurable hardware, sensor networks, and molecular-scale computing systems. In particular, architectures similar to CAs can be used to construct physical devices such as field configurable gate arrays for electronics, networks of robots for environmental sensing and nano-devices embedded in interconnect fabric used for fault tolerant nanoscale computing. Such devices consist of networks of simple components that communicate locally without centralized control. Two major areas of research on such networks are (1) *programming* – how to construct and configure the locally connected components such that they will collectively perform a desired task; and (2) *computation theory* – what types of tasks are such networks able to perform efficiently, and how does the configuration of components affect the computational capability of these networks?

This article describes research into one particular automatic programming method: the use of genetic algorithms (GAs) to evolve cellular automata to perform desired tasks. We survey some of the leading approaches to evolving CAs with GAs, and discuss some of the open problems in this area.

Cellular Automata

A *cellular automaton* (CA) is a spatially extended lattice of locally connected simple processors (cells). CAs can be used both to model physical systems and to perform parallel distributed computations.

In a CA, each cell maintains a discrete state and a transition function that maps the cell's current state to its next state. This function is often represented as a lookup table (LUT). The LUT stores all possible configurations of a cell's local neighborhood, which consists of its own current state and the state of its neighboring cells. Change of state is performed in discrete time steps: the entire lattice is updated synchronously. There are many possible definitions of a *neighborhood*, but here we will define a neighborhood as the cell to be updated and the cells adjacent to it at a distance of radius r. The number of entries in the LUT will be s^N, where s is the number of possible states and N is the total number of cells in the neighborhood: $(2r+1)^d$ for a square shaped neighborhood in a d-dimensional lattice, also known as a *Moore neighborhood*. CAs typically are given *periodic boundary conditions*, which treat the lattice as a torus.

To transform a cell's state, the values of the cell's state and those of its neighbors are encoded as a lookup index to the LUT that stores a value representing the cell's new state (Fig. 1: left) [8,16,59]. For the scope of this article, we will focus on homogeneous binary CAs, which means that all cells in the CAs have the same LUT and each cell has one of two possible states, $s \in \{0, 1\}$. Figure 1 shows the mechanism of updates in a homogeneous one-dimensional two-state CA with a neighborhood radius $r = 1$.

CAs were invented in the 1940s by Stanislaw Ulam and John von Neumann. Ulam used CAs as a mathematical abstraction to study the growth of crystals, and von Neumann used them as an abstraction of a physical system with the concepts of a cell, state and transition function in order to study the logic of self-reproducing sys-

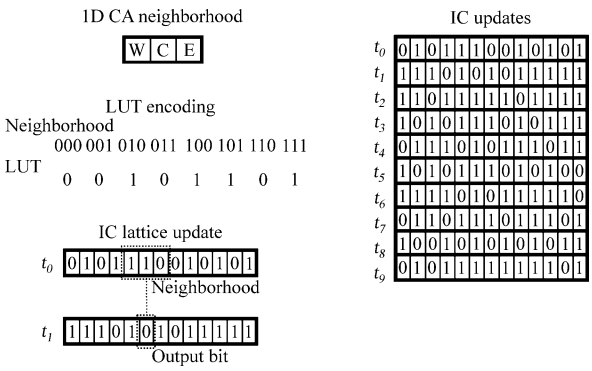

Evolving Cellular Automata, Figure 1

***Left top*: A one-dimensional neighborhood of three cells (radius 1): Center cell, West neighbor, and East neighbor. *Left middle*: A sample look-up table in which all possible neighborhood configurations are listed, along with the update state for the center cell in each neighborhood. *Left bottom*: Mechanism of update in a one dimensional binary CA of length 13: t_0 is the initial configuration at time 0 and t_1 is the initial configuration at next time step. *Right*: The sequence of synchronous updates starting at the initial state t_0 and ending at state t_9**

tems [8,11,55]. Von Neumann's seminal work on CAs had great significance. Science after the industrial revolution was primarily concerned with energy, force and motion, but the concept of CAs shifted the focus to information processing, organization, programming, and most importantly, control [8]. The universal computational ability of CAs was realized early on, but harnessing this power continues to intrigue scientists [8,11,32,55].

Computation in CAs

In the early 1970s John Conway published a description of his deceptively simple Game of Life CA [18]. Conway proved that the Game of Life, like von Neumann's self-reproducing automaton, has the power of a universal Turing machine: any program that can be run on a Turing machine can be simulated by the Game of Life with the appropriate initial configuration of states. This initial configuration (IC) encodes both the input and the program to be run on that input. It is interesting that so simple a CA as the Game of Life (as well as even simpler CAs – see chapter 11 in [60]) has the power of a universal computer. However, the actual application of CAs as universal computers is, in general, impractical due to the difficulty of encoding a given program and input as an IC, as well as very long simulation times.

An alternative use of CAs as computers is to design a CA to perform a particular computational task. In this case, the initial configuration is the input to the program, the transition function corresponds to the program performing the specific task, and some set of final configurations is interpreted as the output of the computation. The intermediate configurations comprise the actual computation being done.

Examples of tasks for which CAs have been designed include location management in mobile computing networks [50], classification of initial configuration densities [38], pseudo-random number generation [51], multi-agent synchronization [47], image processing [26], simulation of growth patterns of material microstructures [5], chemical reactions [35], and pedestrian dynamics [45].

The problem of designing a CA to perform a task requires defining a cell's local neighborhood and boundary conditions, and constructing a transition function for cells that produces the desired input-output mapping. Given a CA's states, neighborhood radius, boundary conditions, and initial configuration, it is the LUT values that must be set by the "programmer" so that the computation will be performed correctly over all inputs.

In order to study the application of genetic algorithms to designing CAs, substantial experimentation has been done using the *density classification* (or *majority classification*) task. Here, "density" refers to the fraction of 1s in the initial configuration. In this task, a binary-state CA must iterate to an all-1s configuration if the initial configuration has a majority of cells in state 1, and iterate to an all-0s configuration otherwise. The maximum time allowed for completing this computation is a function of the lattice size.

One "naïve" solution for designing the LUT for this task would be *local majority voting*: set the output bit to 1 for all neighborhood configurations with a majority of 1s, and 0 otherwise. Figure 2 gives two space-time diagrams illustrating the behavior of this LUT in a one-dimensional binary CA with $N = 149$, and $r = 3$, where N denotes the number of cells in the lattice, and r is the neighborhood radius.

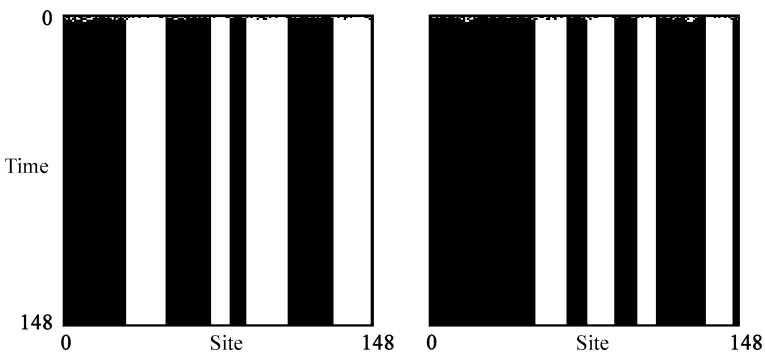

Evolving Cellular Automata, Figure 2
Two space-time diagrams illustrating the behavior of the "naïve" local majority voting rule, with lattice size $N = 149$, neighborhood radius $r = 3$, and number of time steps $M = 149$. *Left*: initial configuration has a majority of 0s. *Right*: initial configuration has a majority of 1s. Individual cells are colored black for state 1 and white for state 0. (Reprinted from [37] with permission of the author.)

Each diagram shows an initial configuration of 149 cells (horizontal) iterating over 149 time steps (vertical, down the page). The left-hand diagram has an initial configuration with a majority of 0 (white) cells, and the right-hand diagram has an initial configuration with a majority of 1 (black) cells. In neither case does the CA produce the "correct" global behavior: an all-0s configuration for the left diagram and an all-1s configuration for the right diagram. This illustrates the general fact that human intuition often fails when trying to capture emergent collective behavior by manipulating individual bits in the lookup table that reflect the settings of the local neighborhood.

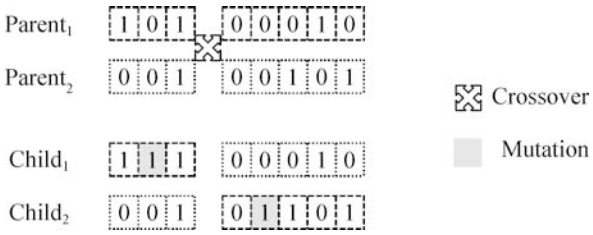

Evolving Cellular Automata, Figure 4
Reproduction applied to *Parent*$_1$ and *Parent*$_2$ producing *Child*$_1$ and *Child*$_2$. The one-point crossover is performed at a randomly selected crossover point (bit 3) and a mutation is performed on bits 2 and 5 in *Child*$_1$ and *Child*$_2$ respectively

Evolving Cellular Automata with Genetic Algorithms

Genetic Algorithms (GAs) are a group of stochastic search algorithms, inspired by the Darwinian model of evolution, that have proved successful for solving various difficult problems [3,4,36].

A GA works as follows: (1) A population of individuals ("chromosomes") representing candidate solutions to a given problem is initially generated at random. (2) The fitness of each individual is calculated as a function of its quality as a solution. (3) The fittest individuals are then selected to be the parents of a new generation of candidate solutions. Offspring are created from parents via copying, random mutation, and crossover. Once a new generation of individuals is created, the process returns to step two. This entire process is iterated for some number of generations, and the result is (hopefully) one or more highly fit individuals that are good solutions to the given problem.

GAs have been used by a number of groups to evolve LUTs for binary CAs [2,10,14,15,34,43,47,51]. The individuals in the GA population are LUTs, typically encoded as binary strings. Figure 3 shows a mechanism of encoding LUTs from a particular neighborhood configuration. For example: the decimal value for the neighborhood 11010 is 26. The updated value for the neighborhood's center cell 11**0**10 is retrieved from the 26th position in the LUT, updating cell's value to 1.

The fitness of a LUT is a measure of how well the corresponding CA performs a given task after a fixed number of time steps, starting from a number of *test* initial configurations. For example, given the density classification task, the fitness of a LUT is calculated by running the corresponding CA on some number k of random initial configurations, and returning the fraction of those k on which the CA produces the correct final configuration (all 1s for initial configurations with majority 1s, all 0s otherwise). The set of random test ICs is typically regenerated at each generation.

For LUTs represented as bit strings, crossover is applied to two parents by randomly selecting a crossover point, so that each child inherits one segment of bits from each parent. Next, each child is subject to a mutation, where the genome's individual bits are subject to a bit complement with a very low probability. An example of the reproduction process is illustrated in Fig. 4 for a lookup

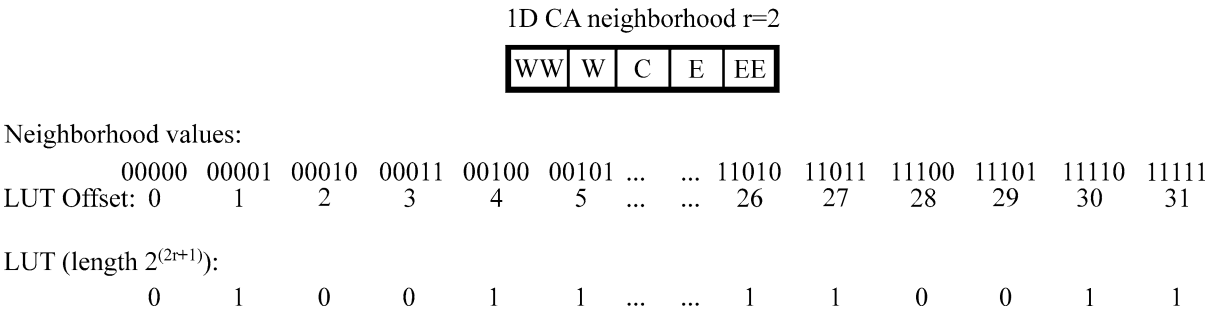

Evolving Cellular Automata, Figure 3
Lookup table encoding for 1D CA with neighborhood $r = 2$. All permutations of neighborhood values are encoded as an offset to the LUT. The LUT bit represents a new value for the center cell of the neighborhood. The binary string (LUT) encodes an individual's chromosome used by evolution

table representation of $r = 1$. Here, one of two children is chosen for survival at random and placed in an offspring population. This process is repeated until the offspring population is filled. Before a new evolutionary cycle begins, the newly created population of offspring replaces the previous population of parents.

Previous Work on Evolving CAs

Von Neumann's self-reproducing automaton was the first construction that showed that CAs can perform universal computation [55], meaning that the CAs are capable, in principle, of performing any desired computation. However, in general it was unknown how to effectively "program" CAs to perform computations or what information processing dynamics CAs could best use to accomplish a task. In the 1980s and 1990s, a number of researchers attempted to determine how the generic dynamical behavior of a CA might be related to its ability to perform computations [18,19,21,33,58]. In particular, Langton defined a parameter on CA LUTs, λ, that he claimed correlated with computational ability. In Langton's work, λ is a function of the state-update values in the LUT; for binary CAs, λ is defined as the fraction of 1s in the state-update values.

Computation at the Edge of Chaos

Packard [40] was the first to use a genetic algorithm to evolve CA LUTs in order to test the hypothesis that LUTs with a critical value of λ will have maximal computational capability. Langton had shown that generic CA behavior seemed to undergo a sequence of phase transitions – from simple to "complex" to chaotic – as λ was varied. Both Langton and Packard believed that the "complex" region was necessary for non-trivial computation in CAs, thus the phrase "computation at the edge of chaos" was coined [33,40]. Packard's experiments indicated that CAs evolved by GAs to perform the density classification task indeed tended to exhibit critical λ values. However, this conclusion was not replicated in later work [38]. Correlations between λ (or other statistics of LUTs) and computational capability in CAs have been hinted at in further work, but have not been definitively established. A major problem is the difficulty of quantifying "computational capability" in CAs beyond the general (and not very practical) capability of universal computation.

Computation via CA "Particles"

While Mitchell, Hraber, and Crutchfield were not able to replicate Packard's results on λ, they were able to show that genetic algorithms can indeed evolve CAs to perform computations [38]. Using earlier work by Hanson and Crutchfield on characterizing computation in CAs [20,21], Das, Mitchell and Crutchfield gave an information-processing interpretation of the dynamics exhibited by the evolved CAs in terms of *regular domains* and *particles* [21]. This work was extended by Das, Crutchfield, Mitchell, and Hanson [14] and Hordijk, Crutchfield and Mitchell [24].

In particular these groups showed that when regular domains – patterns described by simple regular languages – are filtered out of CA space-time behavior, the boundaries between these domains become forefront and can be interpreted as information-carrying "particles". These particles can characterize non-trivial computation carried out by CAs [15,21].

The information-carrying role of particles becomes clear when applied to CAs evolved by the GA for the density classification task. Figure 5, left, shows typical behavior of the best CAs evolved by the GA. The CA contains three regular domains: all white (0*), all black (1*), and checkerboard ((01)*). Figure 5, right, shows the particles remaining after the regular domains are filtered out. Each particle has an origin and velocity, and carries information about the neighboring regions [37]. Hordijk et al. [24] showed that a small set of particles and their interactions can explain the computational behavior (i. e., the fitness) of the evolved cellular automata. Crutchfield et al. [13] describe how the analysis of evolved CAs in terms of particles can also explain how the GA evolved CAs with high fitness.

Land and Belew [31] proved that no two-state homogeneous CA can perform the density classification task perfectly. However, the maximum possible performance for CAs on this task is not known.

The density classification task remains a popular benchmark for studying the evolution of CAs with GAs, since the task requires collective behavior: the decision about the global density of the IC is based on information only from each local neighborhood. Das et al. [14] also used GAs to evolve CAs to perform a global synchronization task, which requires that, starting from any initial configuration, all cells of the CA will synchronize their states (to all 1s or 0s) and in the next time step all cells must change state to the opposite value. Again, this behavior requires global coordination based on local communication. Das et al. showed that an analysis in terms of particles and their interactions was also possible for this task.

Genetic Programming

Andre et al. [2] applied genetic programming (GP), a variation of GAs, to the density classification task. GP method-

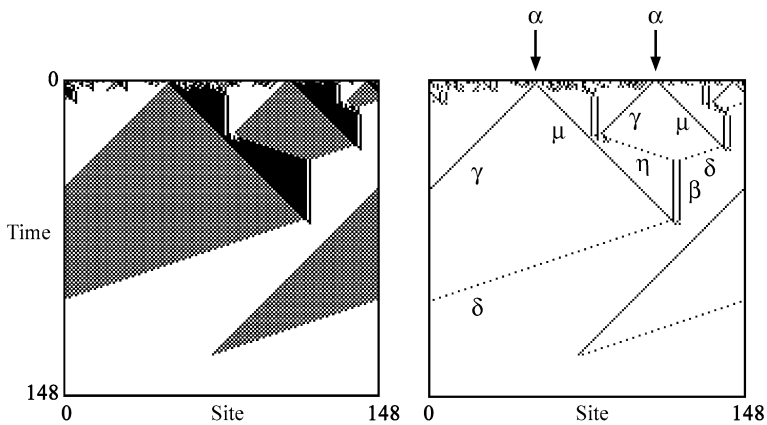

Evolving Cellular Automata, Figure 5

Analysis of a GA evolved CA for density classification task. *Left*: The original spacetime diagram containing particle strategies in a CA evolved by GA. The regions of regular domains are all white, all black, or have a checkerboard pattern. *Right*: Spacetime diagram after regular domains are filtered out. (Reprinted from [37] with permission of the author.)

ology also uses a population of evolving candidate solutions, and the principles of reproduction and survival are the same for both GP and GAs. The main difference between these two methods is the encoding of individuals in the population. Unlike the binary strings used in GAs, individuals in a GP population have tree structures, made up of *function* and *terminal* nodes. The *function* nodes (internal nodes) are operators from a pre-defined function set, and the *terminal* nodes (leaves) represent operands from a terminal set. The fitness value is obtained by evaluating the tree on a set of test initial configurations. The crossover operator is applied to two parents by swapping randomly selected sub-trees, and the mutation operation is performed on a single node by creating a new node or by changing its value (Fig. 6) [29,30].

The GP algorithm evolved CAs whose performance is slightly higher than the performance of the best CAs evolved by a traditional GA.

Unlike traditional GAs that use crossover and mutation to evolve fixed length genome solutions, GP trees evolve to different sizes or shapes, and the subtrees can be substituted out and added to the function set as automatically defined functions. According to Andre et al., this allows GP to better explore the "regularities, symmetries, homogeneities, and modularities of the problem domain" [2]. The best-evolved CAs by GP revealed more complex particles and particle interactions that than the CAs found by the EvCA group [13,24]. It is unclear whether the improved results were due to the GP representation or to the increased population sizes and computation time used by Andre et al.

Parallel Cellular Machines

The field of evolving CAs has grown in several directions. One important area is evolving non-homogeneous cellular automata [22,47,48,54]. Each cell of a non-homogeneous CA contains two independently evolving chromosomes. One represents the LUT for the cell (different cells can have different LUTs), and the second represents the neighborhood connections for the cell. Both the LUTs and the cell's connectivity can be evolved at the same time. Since a task is performed by a collection of cells with different LUTs, there is no single best performing individual; the fitness is a measure of the collective behavior of the cells' LUTs and their neighborhood assignments [46,48].

One of many tasks studied by Sipper was the global ordering task [47]. Here, the CA has fixed rather than periodic boundaries, so the "left" and "right" parts of the CA lattice are defined. The *ordering* in any given IC pattern will place all 0s on the left, followed by all 1s on the right. The initial density of the IC has to be preserved in the final configuration. Sipper designed a *cellular programming algorithmj* to co-evolve multiple LUTs and their neighborhood topologies. Cellular programming carries out the same steps as the conventional GA (initialization, evaluation, reproduction, replacement), but each cell reproduces only with its local neighbors. The LUTs and connectivity chromosomes from the locally connected sites are the only potential parents for the reproduction and replacement of cell's LUTs and the connectivity tables respectively. The cell's limited connectivity results in genetically diverse population. If a current population has a cell with

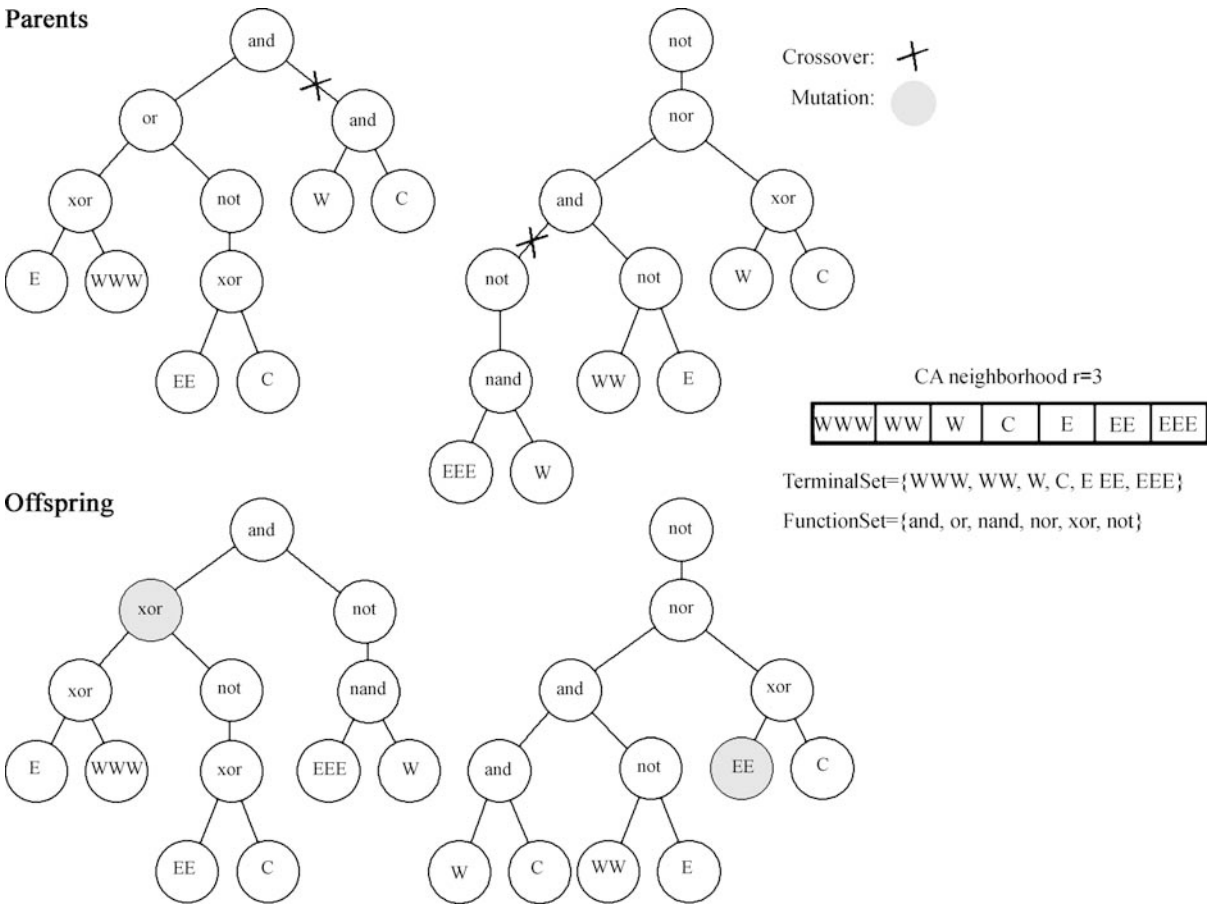

Evolving Cellular Automata, Figure 6

An example of the encoding of individuals in a GP population, similar to one used in [2]. The function set here consists of the logical operators {and, or, not, nand, nor, and xor}. The terminal set represents the states of cells in a CA neighborhood, here {Center, East, West, EastOfEast WestOfWest, EastOfEastOfEast, WestOfWestOfWest}. The figure shows the reproduction of *Parent*$_1$ and *Parent*$_2$ by crossover with subsequent mutation to produce *Child*$_1$ and *Child*$_2$

a high-fitness LUT, its LUT will not be directly inherited by a given cell unless they are connected. The connectivity chromosome causes spatial isolation that allows evolution to explore multiple CA rules as a part of a collective solution [47,48].

Sipper exhaustively tested all homogeneous CAs with $r = 1$ on the ordering task, and found that the best performing rule (rule 232) correctly ordered 71% of 1000 randomly generated ICs. The cellular programming algorithm evolved a non-homogeneous CA that outperformed the best homogeneous CA. The evolutionary search identified multiple rules that the non-homogeneous CA used as the components in the final solution. The rules composing the collective CA solution were classified as state preserving or repairing the incorrect ordering of the neighborhood bits. The untested hypothesis is that the cellular programming algorithm can discover multiple important rules (partial traits) that compose more complex collective behavior.

Coevolution

Coevolution is an extension of the GA, introduced by Hillis [23], inspired by host-parasite coevolution in nature. The main idea is that randomly generated test cases will not continually challenge evolving candidate solutions. Coevolution solves this problem by evolving two populations – candidate solutions and test cases – also referred to as hosts and parasites. The hosts obtain high fitness by performing well on many of the parasites, whereas the parasites obtain high fitness by being difficult for the hosts. Simultaneously coevolving both populations engages hosts

and parasites in a mutual competition to achieve increasingly better results [7,17,56].

Successful applications of coevolutionary learning include discovery of minimal sorting networks, training artificial neural networks for robotics, function induction from data, and evolving game strategies [9,23,41,44,56,57]. Coevolution also improved upon GA results on evolving CA rules for density classification [28].

In the context of evolving CAs, the LUT candidate solutions are hosts, and the ICs are parasites. The fitness of a host is a fraction of correctly evaluated ICs from the parasite population. The fitness of a parasite is a function of the number of hosts that failed to correctly classify it.

Pagie et al. and Mitchell et al. among others, have found that embedding the host and parasite populations in a spatial grid, where hosts and parasites compete and evolve locally, significantly improves the performance of coevolution on evolving CAs [39,41,42,57].

Other Applications

The examples described in previous sections illustrate the power and versatility of genetic algorithms used to evolve desired collective behavior in CAs. The following are some additional examples of applications of CAs evolved by GAs.

CAs are most commonly used for modeling physical systems. CAs evolved by GAs modeled multi-phase fluid flow in porous material [61]. A 3D CA represented a pore model, and the GA evolved the permeability characteristics of the model to match the fluid flow pattern collected from the sample data. Another example is the modeling of physical properties of material microstructures [5]. An alternative definition of CAs (*effector automata*) represented a 2D cross-section of a material. The rule table specified the next location of the neighborhood's center cell. The results show that the GA evolved rules that reconstructed microstructures in the sample superalloy.

Network theory and topology studies for distributed sensor networks rely on connectivity and communication among its components. Evolved CAs for location management in mobile computing networks is an application in this field [50]. The cells in the mobile networks are mapped to CA cells where each cell is either a reporting or non-reporting cell. Subrata and Zomaya's study used three network datasets that assigned unique communication costs to each cell. A GA evolved the rules that designate each cell as reporting or not while minimizing the communication costs in the network. The results show that the GA found optimal or near optimal rules to determine which cells in a network are reporting. Sipper also hinted at applying his cellular programming algorithm to non-homogeneous CAs with non-standard topology to evolve network topology assignments [47].

Chopra and Bender applied GAs to evolve CAs to predict protein secondary structure [10]. The 1D CA with $r = 5$ represents interactions among local fragments of a protein chain. A GA evolved the weights for each of the neighboring fragments that determine the shape of the secondary protein structure. The algorithm achieved superior results in comparison with some other protein-secondary-structure prediction algorithms.

Built-In Self-Test (BIST) is a test method widely used in the design and production of hardware components. A combination of a *selfish gene algorithm* (a GA variant) and CAs were used to program the BIST architecture [12]. The individual CA cells correspond to the circuitry's input terminals, and the transition function serves as a test pattern generator. The GA identified CA rules that produce input test sequences that detect circuitry faults. The results achieved are comparable with previously proposed GA-based methods but with lower overhead.

Computer vision is a fast growing research area where CAs have been used for low-level image processing. The cellular programming algorithm has evolved non-homogeneous CAs to perform image thinning, finding and enhancing an object's rectangle boundaries, image shrinking, and edge detection [47].

Future Directions

Initial work on evolving two-dimensional CAs with GAs was done by Sipper [47] and Jimènez-Morales, Crutchfield, and Mitchell [27]. An extension of domain-particle analysis for 2D CAs is needed in order to analyze the information processing of CAs and to identify the epochs of innovations in evolutionary learning.

Spatially extended coevolution was successfully used to evolve high performance CAs for density classification. Parallel cellular machines also used spatial embedding of their components and found better performing CAs than the homogeneous CAs evolved by a traditional GA. The hypothesis is that spatially extended search techniques are successful more often than non-spatial techniques because spatial embedding enforces greater genetic diversity and, in the case of coevolution, more effective competition between hosts and parasites. This hypothesis deserves more detailed investigation.

Additional important research topics include the study of the error resiliency and the effect of noise on both the information processing in CAs and evolution of CAs. How successful is evolutionary learning in noisy environment?

What is the impact of failing CA components on information processing and evolutionary adaptation? Similarly, to make CAs more realistic as models of physical systems, evolving CAs with asynchronous cell updates is an important topic for future research. A number of groups have shown that CAs and similar decentralized spatially extended systems using asynchronous updates can have very different behavior from those using synchronous updates (e. g., [1,6,25,49,53]). An additional topic for future research is the effect of connectivity network structure on the behavior and computational capability of CAs. Some work along these lines has been done by Teuscher [52].

Acknowledgments

This work has been funded by the Center on Functional Engineered Nano Architectonics (FENA), through the Focus Center Research Program of the Semiconductor Industry Association.

Bibliography

1. Alba E, Giacobini M, Tomassini M, Romero S (2002) Comparing synchronous and asynchronous cellular genetic algorithms. In: Guervos MJJ et al (eds) Parallel problem solving from nature. PPSN VII, Seventh International Conference. Springer, Berlin, pp 601–610
2. Andre D, Bennett FH III, Koza JR (1996) Evolution of intricate long-distance communication signals in cellular automata using genetic programming. In: Artificial life V: Proceedings of the fifth international workshop on the synthesis and simulation of living systems. MIT Press, Cambridge
3. Ashlock D (2006) Evolutionary computation for modeling and optimization. Springer, New York
4. Back T (1996) Evolutionary algorithms in theory and practice. Oxford University Press, New York
5. Basanta D, Bentley PJ, Miodownik MA, Holm EA (2004) Evolving cellular automata to grow microstructures. In: Genetic programming: 6th European Conference. EuroGP 2003, Essex, UK, April 14–16, 2003. Proceedings. Springer, Berlin, pp 77–130
6. Bersini H, Detours V (2002) Asynchrony induces stability in cellular automata based models. In: Proceedings of the IVth conference on artificial life. MIT Press, Cambridge, pp 382–387
7. Bucci A, Pollack JB (2002) Order-theoretic analysis of coevolution problems: Coevolutionary statics. In: GECCO 2002 Workshop on Understanding Coevolution: Theory and Analysis of Coevolutionary Algorithms, vol 1. Morgan Kaufmann, San Francisco, pp 229–235
8. Burks A (1970) Essays on cellular automata. University of Illinois Press, Urban
9. Cartlidge J, Bullock S (2004) Combating coevolutionary disengagement by reducing parasite virulence. Evol Comput 12(2):193–222
10. Chopra P, Bender A (2006) Evolved cellular automata for protein secondary structure prediction imitate the determinants for folding observed in nature. Silico Biol 7(0007):87–93
11. Codd EF (1968) Cellular automata. ACM Monograph series, New York
12. Corno F, Reorda MS, Squillero G (2000) Exploiting the selfish gene algorithm for evolving cellular automata. IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN'00) 06:6577
13. Crutchfield JP, Mitchell M, Das R (2003) The evolutionary design of collective computation in cellular automata. In: Crutchfield JP, Schuster PK (eds) Evolutionary Dynamics – Exploring the Interplay of Selection, Neutrality, Accident, and Function. Oxford University Press, New York, pp 361–411
14. Das R, Crutchfield JP, Mitchell M, Hanson JE (1995) Evolving globally synchronized cellular automata. In: Eshelman L (ed) Proceedings of the sixth international conference on genetic algorithms. Morgan Kaufmann, San Francisco, pp 336–343
15. Das R, Mitchell M, Crutchfield JP (1994) A genetic algorithm discovers particle-based computation in cellular automata. In: Davidor Y, Schwefel HP, Männer R (eds) Parallel Problem Solving from Nature-III. Springer, Berlin, pp 344–353
16. Farmer JD, Toffoli T, Wolfram S (1984) Cellular automata: Proceedings of an interdisciplinary workshop. Elsevier Science, Los Alamos
17. Funes P, Sklar E, Juille H, Pollack J (1998) Animal-animat coevolution: Using the animal population as fitness function. In: Pfeiffer R, Blumberg B, Wilson JA, Meyer S (eds) From animals to animats 5: Proceedings of the fifth international conference on simulation of adaptive behavior. MIT Press, Cambridge, pp 525–533
18. Gardner M (1970) Mathematical games: The fantastic combinations of John Conway's new solitaire game "Life". Sci Am 223:120–123
19. Grassberger P (1983) Chaos and diffusion in deterministic cellular automata. Physica D 10(1–2):52–58
20. Hanson JE (1993) Computational mechanics of cellular automata. Ph D Thesis, Univeristy of California at Berkeley
21. Hanson JE, Crutchfield JP (1992) The attractor-basin portrait of a cellular automaton. J Stat Phys 66:1415–1462
22. Hartman H, Vichniac GY (1986) Inhomogeneous cellular automata (inca). In: Bienenstock E, Fogelman F, Weisbuch G (eds) Disordered Systems and Biological Organization, vol F20. Springer, Berlin, pp 53–57
23. Hillis WD (1990) Co-evolving parasites improve simulated evolution as an optimization procedure. Physica D 42:228–234
24. Hordijk W, Crutchfield JP, Mitchell M (1996) Embedded-particle computation in evolved cellular automata. In: Toffoli T, Biafore M, Leão J (eds) Physics and Computation 1996. New England Complex Systems Institute, Cambridge, pp 153–158
25. Huberman BA, Glance NS (1993) Evolutionary games and computer simulations. Proc Natl Acad Sci 90:7716–7718
26. Ikebe M, Amemiya Y (2001) VMoS cellular-automaton circuit for picture processing. In: Miki T (ed) Brainware: Bio-inspired architectures and its hardware implementation, vol 6 of FLSI Soft Computing, chapter 6. World Scientific, Singapore, pp 135–162
27. Jiménez-Morales F, Crutchfield JP, Mitchell M (2001) Evolving two-dimensional cellular automata to perform density classification: A report on work in progress. Parallel Comput 27(5):571–585
28. Juillé H, Pollack JB (1998) Coevolutionary learning: A case study. In: Proceedings of the fifteenth international conference on machine learning (ICML-98). Morgan Kaufmann, San Francisco, pp 24–26

29. Koza JR (1992) Genetic programming: On the programming of computers by means of natural selection. MIT Press, Cambridge
30. Koza JR (1994) Genetic programming II: Automatic discovery of reusable programs. MIT Press, Cambridge
31. Land M, Belew RK (1995) No perfect two-state cellular automata for density classification exists. Phys Rev Lett 74(25): 5148–5150
32. Langton C (1986) Studying artificial life with cellular automata. Physica D 10D:120
33. Langton C (1990) Computation at the edge of chaos: Phase transitions and emergent computation. Physica D 42:12–37
34. Lohn JD, Reggia JA (1997) Automatic discovery of self-replicating structures in cellular automata. IEEE Trans Evol Comput 1(3):165–178
35. Madore BF, Freedman WL (1983) Computer simulations of the Belousov-Zhabotinsky reaction. Science 222:615–616
36. Mitchell M (1996) An introduction to genetic algorithms. MIT Press, Cambridge
37. Mitchell M (1998) Computation in cellular automata: A selected review. In: Gramss T, Bornholdt S, Gross M, Mitchell M, Pellizzari T (eds) Nonstandard Computation. VCH, Weinheim, pp 95–140
38. Mitchell M, Hraber PT, Crutchfield JP (1993) Revisiting the edge of chaos: Evolving cellular automata to perform computations. Complex Syst 7:89–130
39. Mitchell M, Thomure MD, Williams NL (2006) The role of space in the success of coevolutionary learning. In: Rocha LM, Yaeger LS, Bedau MA, Floreano D, Goldstone RL, Vespignani A (eds) Artificial life X: Proceedings of the tenth international conference on the simulation and synthesis of living systems. MIT Press, Cambridge, pp 118–124
40. Packard NH (1988) Adaptation toward the edge of chaos. In: Kelso JAS, Mandell AJ, Shlesinger M (eds) Dynamic patterns in complex systems. World Scientific, Singapore, pp 293–301
41. Pagie L, Hogeweg P (1997) Evolutionary consequences of coevolving targets. Evol Comput 5(4):401–418
42. Pagie L, Mitchell M (2002) A comparison of evolutionary and coevolutionary search. Int J Comput Intell Appl 2(1):53–69
43. Reynaga R, Amthauer E (2003) Two-dimensional cellular automata of radius one for density classification task $\rho = \frac{1}{2}$. Pattern Recogn Lett 24(15):2849–2856
44. Rosin C, Belew R (1997) New methods for competitive coevolution. Evol Comput 5(1):1–29
45. Schadschneider A (2001) Cellular automaton approach to pedestrian dynamics – theory. In: Pedestrian and evacuation dynamics. Springer, Berlin, pp 75–86
46. Sipper M (1994) Non-uniform cellular automata: Evolution in rule space and formation of complex structures. In: Brooks RA, Maes P (eds) Artificial life IV. MIT Press, Cambridge, pp 394–399
47. Sipper M (1997) Evolution of parallel cellular machines: The cellular programming approach. Springer, Heidelberg
48. Sipper M, Ruppin E (1997) Co-evolving architectures for cellular machines. Physica D 99:428–441
49. Sipper M, Tomassini M, Capcarrere M (1997) Evolving asynchronous and scalable non-uniform cellular automata. In: Proceedings of the international conference on artificial neural networks and genetic algorithms (ICANNGA97). Springer, Vienna, pp 382–387
50. Subrata R, Zomaya AY (2003) Evolving cellular automata for location management in mobile computing networks. IEEE Trans Parallel Distrib Syst 14(1):13–26
51. Tan SK, Guan SU (2007) Evolving cellular automata to generate nonlinear sequences with desirable properties. Appl Soft Comput 7(3):1131–1134
52. Teuscher C (2006) On irregular interconnect fabrics for self-assembled nanoscale electronics. In: Tyrrell AM, Haddow PC, Torresen J (eds) 2nd ieee international workshop on defect and fault tolerant nanoscale architectures, NANOARCH'06. Lecture Notes in Computer Science, vol 2602. ACM Press, New York, pp 60–67
53. Teuscher C, Capcarrere MS (2003) On fireflies, cellular systems, and evolware. In: Tyrrell AM, Haddow PC, Torresen J (eds) Evolvable systems: From biology to hardware. Proceedings of the 5th international conference, ICES2003. Lecture Notes in Computer Science, vol 2602. Springer, Berlin, pp 1–12
54. Vichniac GY, Tamayo P, Hartman H (1986) Annealed and quenched inhomogeneous cellular automata. J Stat Phys 45:875–883
55. von Neumann J (1966) Theory of Self-Reproducing Automata. University of Illinois Press, Champaign
56. Wiegand PR, Sarma J (2004) Spatial embedding and loss of gradient in cooperative coevolutionary algorithms. Parallel Probl Solving Nat 1:912–921
57. Williams N, Mitchell M (2005) Investigating the success of spatial coevolution. In: Proceedings of the 2005 conference on genetic and evolutionary computation. Washington DC, pp 523–530
58. Wolfram S (1984) Universality and complexity in cellular automata. Physica D 10D:1
59. Wolfram S (1986) Theory and application of cellular automata. World Scientific Publishing, Singapore
60. Wolfram S (2002) A new kind of science. Wolfram Media, Champaign
61. Yu T, Lee S (2002) Evolving cellular automata to model fluid flow in porous media. In: 2002 Nasa/DoD conference on evolvable hardware (EH '02). IEEE Computer Society, Los Alamitos, pp 210

Evolving Fuzzy Systems

Plamen Angelov
Intelligent Systems Research Laboratory, Digital Signal Processing Research Group, Communication Systems Department, Lancaster University, Lancaster, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Evolving Clustering
Evolving TS Fuzzy Systems
Evolving Fuzzy Classifiers
Evolving Fuzzy Controllers

Application Case Studies
Future Directions
Acknowledgments
Bibliography

Glossary

Evolving system In the context of this article the term 'evolving' is used in the sense of the self-development of a system (in terms of both its structure and parameters) based on the stream of data coming to the system on-line and in real-time from the environment and the system itself. The system is assumed to be mathematically described by a set of fuzzy rules of the form:

$$\begin{aligned} &Rule^i: \\ &IF\ \ (Input_1\ is\ close\ to\ prototype_1^i) \\ &\quad AND\ \ldots\ AND\ \ (Input_n\ is\ close\ to\ prototype_n^i) \\ &THEN\ \ (Output^i = \overline{Inputs}^T\ \ ConseqPara\ ms) \end{aligned} \tag{1}$$

In this sense, this definition strictly follows the meaning of the English word "evolving" as described in [34], p. 294, namely "unfolding; developing; being developed, naturally and gradually". Contrast this to the definition of "evolutionary" in the same source, which is "development of more complicated forms of life (plants, animals) from earlier and simpler forms". The terms evolutionary or genetic are also associated with such phenomena (respectively operators that mimic these) as chromosome crossover, mutation, selection and reproduction, parents and off-springs [32]. Evolving (fuzzy and neuro-fuzzy) systems do not deal with such phenomena. They rather consider a *gradual development* of the underlying (fuzzy or neuro-fuzzy) system structure.

Fuzzy system structure Structure of a fuzzy (or neuro-fuzzy) system is constituted of a set of fuzzy rules (1). Each fuzzy rule is composed of antecedent (IF) and consequents (THEN) parts. They are linguistically expressed. The antecedent part consists of a number of fuzzy sets that are linked with fuzzy logic aggregators such as conjunction, disjunction, more rarely, negation [43]. In the above example, a conjunction (logical AND) is used. It can be mathematically described by so-called t-norms or t-conorms between membership functions. The most popular membership functions are Gaussian, triangular, trapezoidal [73]. The consequent part of the fuzzy rules in the so-called Takagi–Sugeno (TS) form is represented by mathematical functions (usually linear). The structure of the TS fuzzy system can also be represented as a neural network with a specific (five layer) composition (Fig. 1). Therefore, these systems are also called neuro-fuzzy (NF).

The number of fuzzy rules and inputs (which in case of classification problems are also called features or attributes) is also a part of the structure.

The first layer consists of neurons corresponding to the membership functions of a wparticular fuzzy set. This layer takes the inputs, x and gives as output the degree, μ to which these fuzzy descriptors are satisfied. The second layer represents the antecedent parts

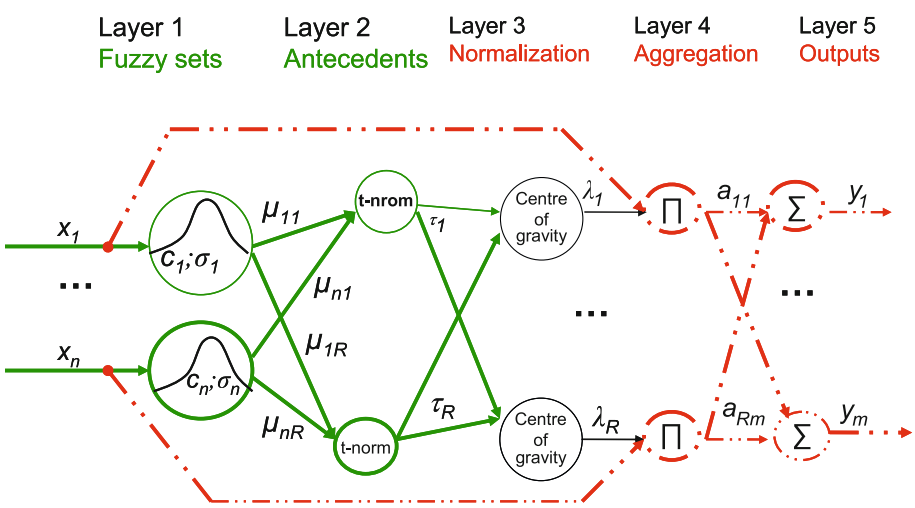

Evolving Fuzzy Systems, Figure 1
Structure of the (neuro-fuzzy) system of TS type

of the fuzzy rules. It takes as inputs the membership function values and gives as output the firing level of the ith rule, τ_i. The third layer of the network takes as inputs the firing levels of the respective rule, τ_i and gives as output the normalized firing level, λ_i as "center of gravity" [43] of τ_i. As an alternative one can use the "winner takes all" operator. This operator is used usually in classification, while the "center of gravity" is preferred for time-series prediction and general system modeling and control. The fourth layer aggregates the antecedent and the consequent part that represents the local sub-systems (singletons or hyper planes). Finally, the last 5th layer forms the total output of the NF system. It performs a weighed summation of local subsystems.

Fuzzy system parameters Parameters of the NF system of TS type include the center, c and spread, σ of the Gaussians or parameters of the triangular (or trapezoidal) membership functions. An example of a Gaussian type membership function can be given as:

$$\mu = \mathrm{e}^{-\frac{1}{2}\left(\frac{d}{r}\right)^2} \tag{2}$$

where d denotes distance between a data sample (point in the data space) and a prototype/cluster center (focal point of a fuzzy set); r is the radius of the cluster (spread of the membership function).

Note that the distance can be represented by Euclidean (the most typical example), Mahalonobis [33], cosine etc. forms.

These parameters are associated with the antecedent part of the system. Consequent part parameters are coefficients of the (usually) linear functions, singleton coefficients or coefficients of more complex functions (e. g. exponential) if such ones are used.

$$y_i = a_{i0} + a_{i1}x_1 + \cdots + a_{in}x_n \tag{3}$$

where a denotes parameters of the consequent part; x denote the inputs (features); i is the index of the ith fuzzy rule; n is the number (dimensionality) of the inputs (features).

Potential Potential is a mathematical measure of the data density. It is calculated at a data point, z and represents numerically the accumulated proximity (*density*) of the data surrounding this data point. It resembles the probability distribution used in so-called Parzen windows [33] and is described in [26,72] by a Gaussian-like function:

$$P(z) = \mathrm{e}^{-\frac{1}{2r}\overline{\sigma}^2} \tag{4}$$

where $z = [x, y]$ denotes the joint (input/output) vector;

$\overline{\sigma}_k^2 = \frac{1}{k-1}\sum_{i=1}^{k-1} d^2(z_k, z_i)$ is the variance of the data in terms of the cluster center.

In [3,9] the Cauchy function is used which has the same properties as the Gaussian but is suitable for recursive calculations.

$$P(z) = \frac{1}{1+\overline{\sigma}^2}\,. \tag{5}$$

Age of a cluster or fuzzy rule The age of the (evolving) cluster is defined as the accumulated time of appearance of the samples that form the cluster which support that fuzzy rule.

$$A^i = k - \frac{\sum_{l=1}^{S_k^i} k_l}{S_k^i} \tag{6}$$

where k denotes the current time instant; S_k^i denotes the support of the cluster that is the number of data samples (points) that are in the zone of influence of the cluster (formed by its radius). It is derived by simple counting of data samples (points) at the moment of their arrival (when they are first read) and assigned to the nearest cluster [10].

The values of A vary from 0 to k and the derivative of A in respect to time is always less or equal to 1 [17]. An "old" cluster (fuzzy rule) has not been updated recently. A "young" cluster (fuzzy rule) is one that has predominantly new samples or recent ones. The (first and second) derivatives of the *age* are very informative and useful for detection of data "*shift*" and "*drift*" [17].

Definition of the Subject

Evolving Fuzzy Systems (EFS) are a class of Fuzzy Rule-based (FRB) and Neuro-Fuzzy (NF) systems that have both their parameters and underlying structure self-adapting, self-developing, self-learning from the data in on-line mode and, possibly, in real-time. The concept was conceived at the beginning of this century [2,5]. Parallel investigations have led to similar developments in neural networks (NN) [41,42]. EFS have the significant advantage compared to the evolving NN of being linguistically tractable and transparent. EFS have been instrumental in the emergence of new branches of *evolving* clustering algorithms [3], *evolving* classifiers [16,51], *evolving* time-series predictors [9,47], *evolving* fuzzy controllers [4], *evolving* fault detectors [30] etc. Over the last years EFS has demonstrated a wide range of applications spanning robotics [76]

and defense [24] to biomedical [70] and industrial process [29] data processing in real-time, new generations of self-calibrating, self-adapting sensors [52,53], speech [37] and image processing [56] etc. EFS have the potential to revolutionize such areas as autonomous systems [66], intelligent sensors [45], early cancer detection and diagnosis; they are instrumental in raising the so-called machine intelligence quotient [74] by developing systems that self-adapt in real-time to the dynamically changing environment and to internal changes that occur in the system itself (e. g. wearing, contamination, performance degradation, faults etc.). Although the terms *intelligent* and *artificial intelligence* have been used often during the last several decades the technical systems that claim to have such features are in reality far from true intelligence. One of the main reasons is that true intelligence is evolving, it is not fixed. EFS are the first mathematical constructs that combine the approximate reasoning typical for humans represented by the fuzzy inference with the dynamically evolving structure and respective formal mathematically sound learning mechanisms to implement it.

Introduction

Fuzzy Sets and Fuzzy Logic were introduced by Lotfi Zadeh in 1965 in his seminal paper [71]. During the last decade of the previous century there was an increase of the various applications of fuzzy logic-based systems mainly due to the introduction of fuzzy logic controllers (FLC) by Ebrahim Mamdani in 1975 [54], the introduction of the fuzzily blended linear systems construct called Takagi–Sugeno (TS) fuzzy systems in 1985 [65], and the theoretical proof that FRB systems are universal approximators (that is any arbitrary non-linear function in the [0; 1] range can be asymptotically approximated by a FRB system [68]). Historically, the FRB systems where first being designed based entirely or predominantly on human expert knowledge [54,71]. This offers advantages and was a novel technique at that time for incorporating uncertain, subjective information, preferences, experience, intuition, which are difficult or impossible to be described otherwise. However, it poses enormous difficulties for the process of designing and routine use of these systems, especially in real industrial environments and in on-line and real-time modes. TS fuzzy systems made possible the development of efficient algorithms for their design not only in off-line, but also in on-line mode [14]. This is facilitated by their dual nature – they combine a fuzzy linguistic premise (antecedent) part with a functional (usually linear) consequent part [65]. With the invention of the concept of EFS [2,5] the problem of the design was completely automated and data-driven. This means, EFS systems self-develop their model, respectively system structure as well as adapt their parameters "from scratch" on the fly using experimental data and efficient recursive learning mechanisms. Human expert knowledge is not compulsory, not limiting, not essential (especially if it is difficult to obtain in real-time). This does not necessarily mean that such knowledge is prohibited or not possible to be used. On the contrary, the concept of EFS makes possible the use of such knowledge in initialization stages, even during the learning process itself, but this is not essential, it is optional. Examples of EFS are intelligent sensors for oil refineries [52,53], autonomous self-localization algorithms used by mobile robots [75,76], smart agents for machine health monitoring and prognosis in the car industry [30], smart systems for automatic classification of images in CD production process [51] etc. This is a new promising area of research and new applications in different branches of industry are emerging.

Evolving Clustering

Data Clustering and, Fuzzy Clustering in particular, are methods for grouping the data based on their similarity, density in the data space and proximity. Partitioning of the data into clusters can be done off-line (using a batch set of data, performing iterative computations over this set, minimizing certain criteria/cost function) or on-line, incrementally. Examples of incremental clustering approaches are self-organizing maps (SOM) conceived by Teuvo Kohonen in the early 1980s [44], adaptive resonance theory (ART) by Stephen Grossberg conceived in the same period [25] etc. Clustering is a type of unsupervised learning technique where the correct examples are not provided. Usually, the number of clusters is pre-specified, e. g. in SOM the number of nodes of the map is pre-defined; the number of neighbors, k in the k-nearest neighbor clustering method [33] is also supposed to be provided, the number C in the fuzzy c-means (FCM) fuzzy clustering algorithm by Jim Bezdek should also be provided [22]. Usually these approaches rely on a threshold and are very sensitive to the specific values of this threshold. Most of the existing approaches are also mean-based (i. e. they use the mean of all data or mean of groups of data). The problem is that the mean is a virtual (non-existing and possibly infeasible) point in the data space.

In contrast, the evolving clustering method eClustering conceived in the last decade [3] does not need the number of clusters, the threshold or any other parameter to be pre-specified. It is parameter-free and starts "from scratch" to cluster the data based on their density distri-

bution alone. It is based on the recursive calculation of the potential (5). eClustering is prototype-based (some of the data points are used as prototypes of cluster centers). The procedure of the evolving clustering approach starts from scratch assuming that the first available data point is a center of a cluster. This assumption is temporary and if a priori knowledge exists the procedure can start with an initial set of cluster centers that will be further refined. The coordinates of the first cluster center are formed from the coordinates of the first data point ($z_1^* \leftarrow z_1$). Its potential is set to the ideal value, $P_1(z_1) \rightarrow 1$. Starting from the **next** data point which is read in real-time, the following steps are performed for each new data point:

- calculate its potential, $P_k(z_k)$;
- update the potential of the existing cluster centers (because their potential has been affected by adding a new data point);
- compare the potential of the new data point with the potential of the previously existing centers. On the basis of this comparison and the membership of the existing clusters one of the following actions is taken:
  (**add** a new cluster center based on the new data point) **OR** (**remove** the cluster that describes well the new point which brings an increment to the potential) **AND** (**replace** it with a cluster formed around the new point) **OR** (**ignore** (do not change the cluster structure)).

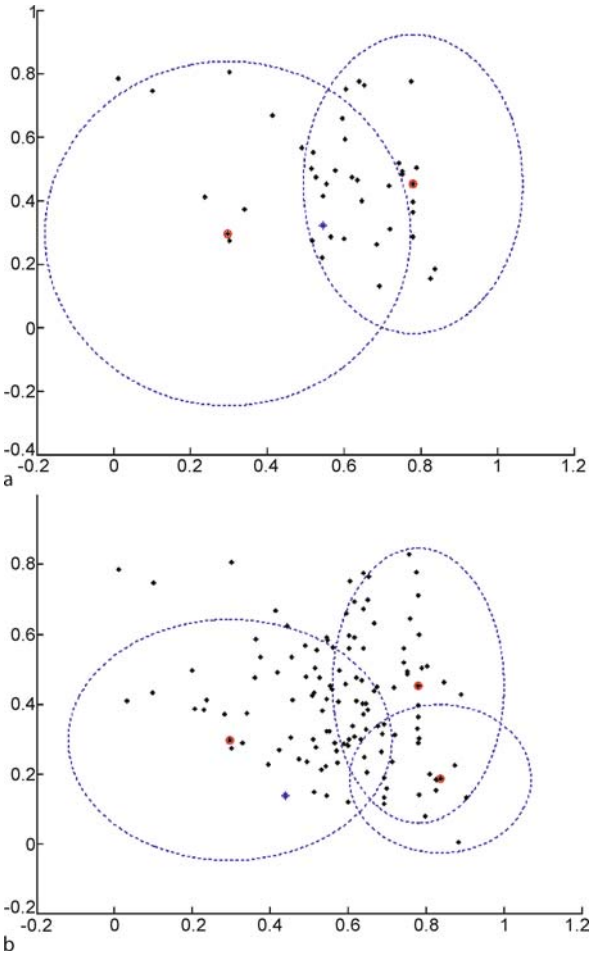

Evolving Fuzzy Systems, Figure 2
The Evolving Clustering method applied to data concerning NOx emissions; a top plot-after 43 samples are read (after 43 s because the sampling rate is 1 sample/second or 1 Hz); b bottom plot – after 124 samples are read (after 124 s)

The process is illustrated in Fig. 2 for the data of NOx emissions from a car exhaust [13].

One can see that the clustering evolves (number of clusters increases from two to three, their position and their radius has changed. Note that in this experiment only two normalized to the range [0; 1] inputs (features), namely the engine output torque in N/m, x_1 and pressure in the second cylinder in Pa, x_2 are used. For more details on eClustering, please, consult the papers from the bibliography, and especially [2,3,9,16].

Evolving TS Fuzzy Systems

TS fuzzy systems [as illustrated in Fig. 1 and described in a very general form in Eq. (1)] were first introduced in 1985 [65] in the form:

$$\begin{aligned} \Re_i : \textbf{\textit{IF}}\ (x_1\ \textit{is}\ A_{i1})\ \textbf{\textit{AND}}\ \ldots\ \textbf{\textit{AND}}\ (x_n\ \textit{is}\ A_{in}) \\ \textbf{\textit{THEN}}\ (y_i = a_{i0} + a_{i1}x_1 + \cdots + a_{in}x_n) \end{aligned} \tag{7}$$

where A_i denotes the ith fuzzy rule ($i = [1, R]$); R is the number of fuzzy rules; $\boldsymbol{x}$ is the input vector; $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T$; A_{ij} denotes the antecedent fuzzy sets, $j \in \{1, n\}$; y_i is the output of the ith linear subsystem; a_{il} are its parameters, $l \in \{0, n\}$.

The structure of the TS system (number of fuzzy rules), antecedent part of the rules, number of inputs etc., are supposed to be known and are fixed. The data may be provided to the TS system in off-line or in on-line manner. Different data-driven techniques were developed to identify the best in terms of certain (local or global) error minimization criteria such as using a (recursive) least squares technique [65], using genetic algorithms [7,62] etc. The overall output is found to be a weighted sum of local outputs produced by each fuzzy rule:

$$y = \sum_{i=1}^{R} \lambda^i y^i \,. \tag{8}$$

Where the weights, λ represent the normalized firing level of the respective fuzzy rule and can be determined by:

$$\lambda^i = \frac{\overset{n}{\underset{j=1}{T}} \mu_j^i(x_j)}{\sum_{l=1}^{R} \overset{n}{\underset{j=1}{T}} \mu_j^l(x_j)} . \tag{9}$$

In a vector form the above equations can be represented as:

$$y = \boldsymbol{\psi}^{\mathrm{T}} \boldsymbol{\theta} \tag{10}$$

where $\boldsymbol{\psi} = [\lambda_1 x_e^{\mathrm{T}}, \lambda_2 x_e^{\mathrm{T}}, \ldots, \lambda_R x_e^{\mathrm{T}}]^{\mathrm{T}}$ is the vector of weighted extended inputs; $x_e^{\mathrm{T}} = [1, x^{\mathrm{T}}]$;
$\boldsymbol{\theta} = \left[\pi_1^{\mathrm{T}}, \pi_2^{\mathrm{T}}, \ldots, \pi_R^{\mathrm{T}}\right]^{\mathrm{T}}$ is the vector of parameters;

$$\pi^i = \begin{bmatrix} \alpha_{01}^i & \alpha_{02}^i & \ldots & \alpha_{0m}^i \\ \alpha_{11}^i & \alpha_{12}^i & \ldots & \alpha_{1m}^i \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_{n1}^i & \alpha_{n2}^i & \ldots & \alpha_{nm}^i \end{bmatrix}$$

are the parameters of the m local linear sub-systems.

The assumption that the TS fuzzy system structure has to be known a priori was for the first time questioned in [2,5] and ultimately in [9] with the proposal of evolving TS (eTS) systems. In [15] a further extension of the eTS system was proposed, namely that they can have many outputs. In this way, the multi-input-multi-output eTS systems were introduced (MIMO-eTS).

eTS is a very flexible and powerful tool for time-series prediction, prognosis, modeling non-stationary phenomena, intelligent sensors etc. The algorithm for its learning from streaming data in real-time has two basic phases, which can both be performed very quickly (in one time step between the arrival of two data samples – the current one and the next one). The learning mechanism proposed in [9] is computationally very efficient because it is fully recursive. The two phases include:

(a) Data space partitioning and based on this forming and update of the fuzzy rule-base structure;
(b) Learning parameters of the consequent part of the fuzzy rules.

Note that the partitioning of the data space serves in eTS identification a different purpose compared to the purpose of data space partitioning in eClustering. In eTS there are outputs and the aim is to find such (perhaps overlapping) clustering of the input-output joint data space that fragments the input-output relationship into locally valid simpler (possibly linear) dependences. In eClustering the aim is to cluster the input data space into distinctive regions. Other than that, the first phase of the eTS model identification is the same as the procedure in the eClustering method described above.

The second phase of the learning is parameter identification. It can be performed using a fuzzily weighted version [9] of the well-known recursive least squares (RLS) method [50]. One can perform either local (11) or global (12) identification by minimizing different cost functions [9]:

$$J_L = \sum_{i=1}^{R} \left(Y - X^{\mathrm{T}}\pi_i\right)^{\mathrm{T}} \Lambda_i \left(Y - X^{\mathrm{T}}\pi_i\right) \tag{11}$$

$$J_G = \left(Y - \Psi^{\mathrm{T}}\boldsymbol{\theta}\right)^{\mathrm{T}} \left(Y - \Psi^{\mathrm{T}}\boldsymbol{\theta}\right) . \tag{12}$$

In one of the cases (when a local cost function is used) the result will be a better approximation locally of the overall non-linear function by the local linear sub-models. The pay-off is, however, a poorer overall approximation. This is, however, compensated by a simpler and computationally more efficient procedure (if we use a locally valid cost function the covariance matrices of much smaller size can be used and they require much less memory space and time to perform computations) [9].

Evolving Fuzzy Classifiers

Classification is a problem that has been well studied and a large number of conventional approaches exist to address this problem [33]. Most of them, however, are designed to operate in batch mode and do not change their structure on-line (do not capture new patterns that may be present in the streaming data once the classifier is built). Off-line pre-trained classifiers may be good for certain scenarios, but they need to be redesigned or retrained if the circumstances change. There are also so-called incremental (or on-line) classifiers which work on a "sample-by-sample" basis and only require the features of that sample plus a small amount of aggregated information (a rule-base, a small number of variables needed for recursive calculations). They do not require all the history of the data stream (all previously seen data samples). Sometimes they are also called one-pass (each sample is processed only once at a time and is then discarded from the memory).

FRB systems have been successfully applied to a range of classification tasks including, but not limited to, decision making, fault detection, pattern recognition, image processing [46]. FRB systems have become one of the alternative frameworks for classifier design together with the more established Bayesian classifiers, decision trees [33], neural network-based classifiers [57], and support-vector machines (SVM) [67]. The task of the classifier is to map

the set of features of the sample data onto the set of class labels. A particular advantage of the FRB classifiers is that they are linguistic in form while being also proven universal approximators [68].

In the framework of the concept of *evolving fuzzy systems* a family of evolving fuzzy classifiers, eClass was proposed in [16,17,51]. The first type of evolving fuzzy classifiers, eClass0 has the typical structure of a fuzzy classifier [46] that differs from structure (1) by the consequent part only:

$$\begin{aligned} R^i:\ & \textbf{\textit{IF}}\ \ (\textit{Feature}_1\ \textit{is close to prototype}_1^i) \\ & \textbf{\textit{AND}}\ \ \ldots\ \ \textbf{\textit{AND}}\ \ (\textit{Feature}_n\ \textit{is close to prototype}_n^i) \\ & \textbf{\textit{THEN}}\ \ (\textit{ClassLabel}^i) \end{aligned} \tag{13}$$

The output of eClass0, in the same way as typical fuzzy classifiers [46] provides the label of the class (0, 1 etc.) directly. In this sense, it is not a TS fuzzy system, but is closer to the Mamdani-type fuzzy systems [54]. The main difference of the eClass0 from the typical classifiers [46] is its ability to evolve, to expand the set of fuzzy rules that enables it to capture new data patterns, to adapt to possibly changing characteristics of the streaming data [16,17]. The inference in eClass0 is produced using the so-called "winner takes all" rule [33,46]:

$$Label = Label^{i*};\ i^* = \arg\max_{i=1}^{R}\left(\mathop{T}_{j=1}^{n}\mu_j^i\left(x_j\right)\right). \tag{14}$$

It is much easier and faster to build and train eClass0 in real-time, but the results of the classification can be further significantly improved if the classifier structure is assumed to be of TS-type. eClass is designed for on-line applications with an evolving (self-developing) FRB structure. The antecedents of the FRB are formed from the data stream around highly descriptive focal points (prototypes) per class in the input-output space. The features vector, $\boldsymbol{x}$ is augmented with the class label, L to form a joint input-output vector $\boldsymbol{z} = \left[\boldsymbol{x}^{\mathrm{T}}, L\right]^{\mathrm{T}}$. The eClustering algorithm is applied ***per class*** (not over all the data). In this way, information granules (primitive forms of knowledge) [36] are formed **in real-time** around descriptive data samples, represented linguistically by fuzzy sets. This on-line algorithm works similarly to adaptive control [19] and estimation [39] – in the period between two samples two phases are performed: (1) class prediction (classification); (2) classifier update or evolution. During the first phase the class label is not known and is being predicted; during the second phase, however, it is known and is used as supervisory information to update the classifier (including its ***structure evolution*** as well as its parameters update).

An alternative structure of the fuzzy classifier, eClass1 is based on the TS-type fuzzy system which has a consequent part of functional type as described in (7). The architecture of eClass1 differs significantly from the architecture of eClass0 and the typical FRB [46]. It performs a regression over the features. Having in mind that the classification surface in a data stream changes dynamically the goal of the evolving fuzzy classifier eClass1 is to evolve a rule-base which takes these changes into account by adapting parameters of the FRB (spreads, consequent parameters) as well as the focal points and the size of the rule-base. The output of each rule is a real (not integer as in the typical fuzzy classifiers) value, which if normalized represents the possibility of a data sample to be of certain class [16,17]:

$$\overline{y}^i = \frac{y^i}{\sum_{i=1}^{R} y^i}. \tag{15}$$

The overall output of the classifier is then taken as a weighted average (not as winner takes all as in typical fuzzy classifiers) of the normalized outputs of each fuzzy rule:

$$y = \sum_{i=1}^{R}\frac{\mathop{T}_{j=1}^{n}\mu_j^i}{\sum_{l=1}^{R}\mathop{T}_{j=1}^{n}\mu_j^l}\overline{y^i}. \tag{16}$$

This output is then used to discriminate between the classes. If the problem has two classes (A and B) then the target values are, obviously, 0 for one of the two classes (e. g. Class A) and 1 for the other one (Class B) or vice versa. To discriminate in this case one can simply use a threshold of 0.5. All the outputs that are above 0.5 are being classified as Class B while all the outputs below 0.5 are classified as Class A or vice versa:

$$\begin{aligned} & \textbf{\textit{IF}}\ \ (y > 0.5) \\ & \textbf{\textit{THEN}}\ \ (\textit{Class}\ \ A) \\ & \textbf{\textit{ELSE}}\ \ (\textit{Class}\ \ B) \end{aligned} \tag{17}$$

When the problem has more than two classes, one can apply MIMO eTS where each of the K outputs corresponds to the possibility that a data sample belongs to a certain class (as discussed above). It is interesting to note that it is possible to use MIMO eTS for a two-class problem. In order to do this, one needs to have vectors that represent the target outputs, for example $\boldsymbol{y} = [1\,0]$ for Class A and

$\boldsymbol{y} = [0\,1]$ for Class B or vice versa. In eClass1-MIMO the label is determined by the highest value of the discriminator, $\overline{\boldsymbol{y}}_l$:

$$Label = Label^{i*}; i^* = \arg\max_{l=1}^{K} \overline{y}_l \tag{18}$$

where K denotes the number of classes.

Evolving Fuzzy Controllers

Fuzzy logic controllers have been applied in a range of industrial processes [48,59] around the word including in home appliances [21]. The structure of the controller, however, is often decided in an ad hoc manner [54,59] and parameters are tuned off-line using various numerical techniques such as genetic algorithms for example [27,32,63]. In reality, however, even if a validation test has been made beforehand, there is no guarantee that a controller designed in this way will perform satisfactory if the object of control or its environment change [4]. The reasons could be aging, wearing, change in the mode of operation or development of a fault, seasonal changes in the environment etc. An effective mechanism for tackling such problems known from the classical control theory is adaptation [19]. It is well developed for the linear models and controllers [38], but not for a general (very often highly non-linear, complex and uncertain) case [2]. Adaptive control theory assumes a linear model with a fixed structure and applies to parameters only [19,38].

The concept of evolving fuzzy systems has been applied to the control problem in [2,4] in terms of self-developing the structure of a controller from experimental data in a data-driven manner based on the indirect adaptive learning scheme proposed initially by Psaltis [60] and developed further using NN by Anderson [1]. The indirect learning (IL) control scheme is based on the approximation of the inverse dynamics of the plant. The IL-based control scheme is a model-free concept. It feeds back the integrated (or delayed one-step back) output signal instead of feeding back the error between the plant output and the reference signal as represented in Fig. 3.

Figure 3 represents only the basic concept of the approach. It has two phases and the switching between them can be represented by an imaginary switch knob. When the imaginary knob, K is in position "1" the controller is used and we are in phase "Control". When the imaginary knob is in position "2" the controller learns, self-develops and we are in phase "Learning". During the supervisory learning phase, the true output signal ($\boldsymbol{y}_{k+1}$) at the time-instant $(k+1)$ is fed back and the knob is in position "2". The controller also receives a signal that is a delayed true output, $\boldsymbol{y}_k$. The controller has as an output the value of the control signal, $\boldsymbol{u}_k$. During the control phase (when the knob is in position "1") the input is determined entirely based on the reference signal (ref) as an alternative to the predicted next step output, $\boldsymbol{y}_{k+1}$. In this way, the controller already trained in the previous learning phases produces such a control signal (u_k), which brings the output of the plant at the next time step ($\boldsymbol{y}_{k+1}$) close to the reference signal (ref).

The IL scheme was taken further in [2,4] by implementing the controller as an evolving FRB system of TS-type. The original works realized the controller as a NN that was trained *off-line* based on a batch set of training data for the control action and output triplets of the form $[\boldsymbol{y}_k, \boldsymbol{y}_{k+1}, \boldsymbol{u}_k]$ for $k = 1, 2, \ldots, N$; where N denotes the

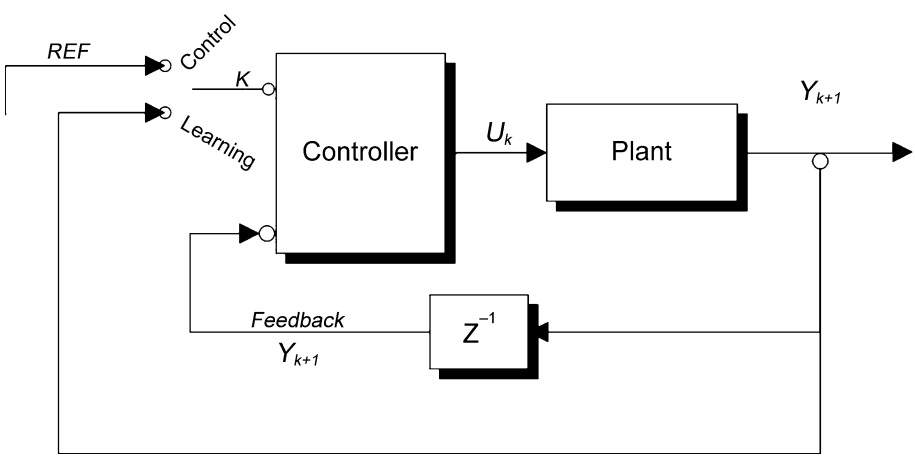

Evolving Fuzzy Systems, Figure 3
Indirect learning-based control scheme

number of training data samples. However, learning techniques for NN are iterative and, therefore training of the NN-controller as described in [1,60] is performed *off-line*. Additionally, NN suffers from the important disadvantage in comparison to the FRB systems that they are not transparent. In [2,4] the basic scheme of IL control is taken further by adding a disturbance and by using eTS to realize the controller. This scheme was implemented on temperature-control problems.

Application Case Studies

Self-Calibrating Intelligent Sensors for Process Industries

So-called intelligent or inferential sensors have been adopted by the process industries (chemical, petro-chemical, manufacturing) for several decades [31]. The main reason is that they provide accurate real time estimates of difficult to measure otherwise parameters or can replace (substitute) expensive measurements such as gas emissions, biomass, melt index, etc. They use as inputs the available ("hard") sensors for easy to measure physical variables, such as temperatures, pressures, and flows which are also cheaper. The main disadvantage of the currently existing inferential or "soft" sensors is that significant efforts must be made based on batch sets of data to develop and maintain the mathematical models that support them (neural networks, statistical models etc.). Any process changes outside the conditions used for off-line model development can lead to significant performance deterioration which, in turn, requires maintenance and recalibration.

Evolving Fuzzy Systems offer an effective opportunity to develop "soft" sensors that are more flexible, self-calibrating, and thus, more "intelligent" [45]. Several applications of EFS-based soft sensors, in particular for oil refineries [52,53] and propylene production [12] were reported. An important advantage of the evolving sensors is that they extract human-interpretable knowledge in the form of linguistic fuzzy rules. For example, Fig. 4 illustrates membership functions of the fuzzy sets (in values in the range [0;1] on the vertical axis) that describe the density of the crude, d in gram per liter (g/l) on the horizontal axis. The evolving fuzzy sensor (eSensor) implemented in the oil refinery at Santa Cruz, Tenerife, Spain predicts in real-time the temperature of the heavy naphtha (hn) T^{hn}, °C in degrees Celsius when it evaporates 95% liquid volume, according to the ASTM D86-04b standard based on real-time measurements of:

- The pressure of the tower, p, measured in kg/cm^{2}g
- The amount of the product taking off, P, represented in %
- The density of the crude, d in g/l (illustrated in Fig. 4)
- Temperature of the column overhead, T^{co} in °C
- Temperature of the naphtha extraction, T^{ne} in °C

An expert in the area of oil refining processes can easily visually distinguish between the heavy crude and light crude represented by respective membership functions derived automatically from the data in real-time.

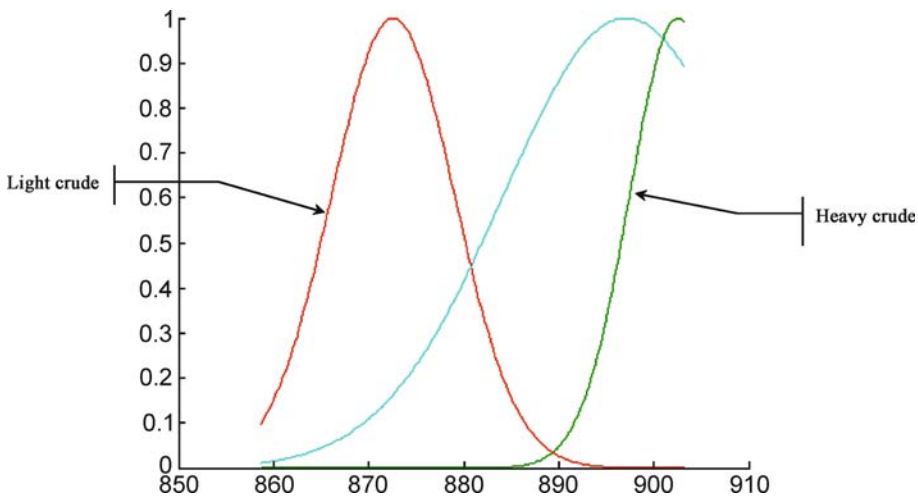

Evolving Fuzzy Systems, Figure 4
Fuzzy sets for different fractions of the crude that contribute to the different quality of the end product (naphtha in this case) can be extracted automatically in real-time from the data stream using eSensor

Adaptive Real-Time Classifiers (Image, Land-Mark, Robotic)

Presently the data that are to be processed in industry, in defense and other real-life applications are not only huge in volume, but very often they are in the form of data *streams* [28]. This requires not only precise classifiers, but also dynamically evolvable classifiers. For example, in mobile robotics, an autonomous vehicle produces a video stream while operating in a completely unknown environment that needs to be processed [75,76]. The Evolving Clustering method was used to automatically generate a fuzzy rule-base that describes the landmarks discovered without any prior learning, "on the fly" by a mobile robot Pioneer 3DX [58] exploring a completely unknown environment [76]. The mobile robot is using its on-board pan-tilt zoom camera to produce a video stream. The frames were grasped and processed by the eClustering algorithm based on a 3-dimensional color-vector (R, G, B). Fuzzy rules of the following form were extracted from the data automatically:

Note that the landmarks that were identified automatically represented real objects on the route of the mobile robot such as the underpass of Lancaster University from the example. They were identified to be distinctive from the surrounding background by eClustering. The fuzzy rule base (Fig. 5) has evolved "from scratch" based on the video information and the data distribution only.

Predictive Models (Air-Conditioning, Financial Time-Series, Benchmark Data Sets)

There are different techniques that can be used for predictive models, such as ARMAX models [50], neural networks [34] non-evolving (fixed structure) fuzzy rule-based models [65,73]. Evolving fuzzy systems, however, offer additionally the capability to have a predictor that evolves following the dynamic changes of the data by gradually adapting not only its parameters, but also its structure. In [6] the problem of predicting the characteristic temperature difference across a coil in a heat exchanger of an air-conditioning unit installed in a real building in Iowa, USA is considered. The evolving fuzzy rule-based model develops its structure and parameters based on the data of the flow rate entering the coil, moisture content of the air entering the coil, temperature of the chilled water, and control signal to the valve as illustrated in Fig. 6. The model proved to work satisfactorily in all season conditions due to its ability to adapt to the changes in the environment (different seasons) as demonstrated in Fig. 6c. When pre-trained (based on 400 samples) and fixed as both structure and parameters the performance deteriorated unacceptably in changing seasonal conditions as seen in Fig. 6a. A partial re-training improved significantly the results (Fig. 6b), but still this was less valid when the season changed (at around sample 912) and the model structure evolution has stopped (at sample 1000). When, the model structure evolution continued uninterrupted the result was a satisfactory performance in all seasons as seen in Fig. 6c.

Fault Detection and Prognostics

Evolving clustering and eTS fuzzy systems were applied in Ford Motor Company to machine health monitoring and prognosis in [30]. The ability of the evolving clustering method to form new clusters that represent different operating modes of a machine was exploited and different types of faults (incipient or drastic) were automatically identified based on the difference in the cluster formation. A prediction of the direction of movement of the cluster centers was used for prediction of possible faults and of the end-of-life of the machine.

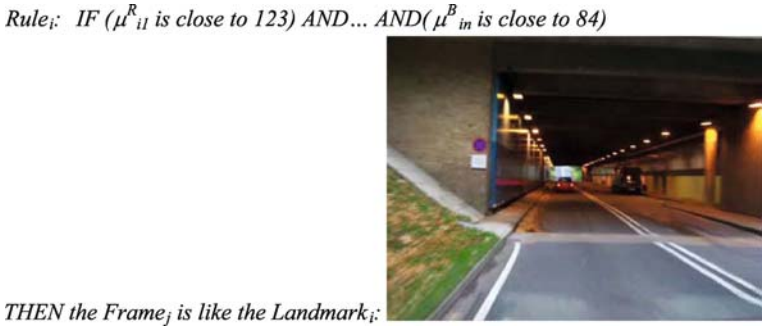

Evolving Fuzzy Systems, Figure 5
Fuzzy rule describing a Landmark (underpass at Lancaster University campus) that was discovered automatically by eClustering using video streaming data

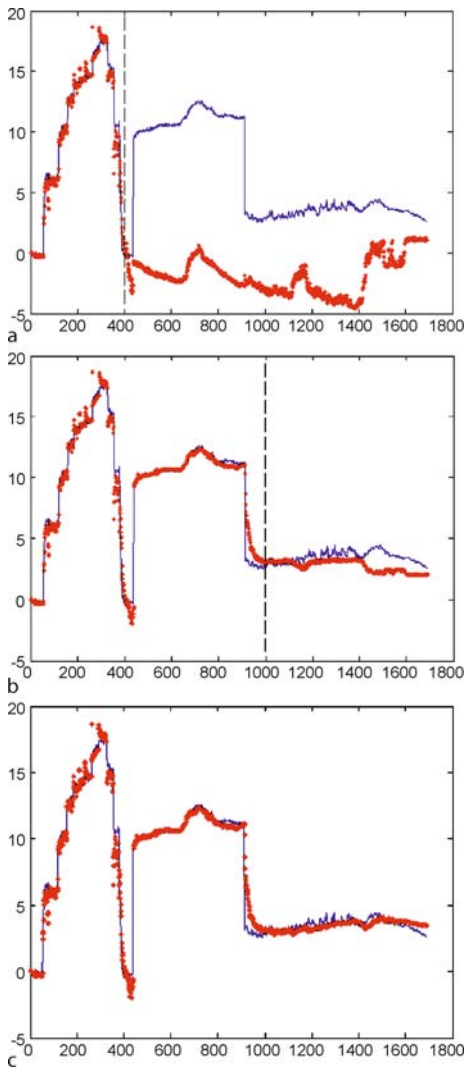

Evolving Fuzzy Systems, Figure 6
A predictive model of the characteristic temperature difference across a coil in a heat exchanger of an air-conditioning unit installed in a real building in Iowa, USA. a An off-line pre-trained model used in two different seasons (summer and spring); b Evolving FRB model trained and used during the summer (up to sample 1000) and then having its structure fixed during the spring season; c Evolving FRB model left to evolve (self-develop) during the whole period of usage (both spring and summer seasons)

Speech Signal Reconstruction

The ETS fuzzy system is used in [37] for error concealment in the next-generation Voice over Internet Protocol (VoIP) communication receivers. It is used in combination with parametric speech coders of analysis-by-synthesis-type. eTS MIMO [15] is used to predict the missing values of the linear spectral pairs (LSP) that will allow one to reconstruct the lost in transmission packets. The eTS fuzzy model used ten inputs (current LSP parameter values) and ten outputs (predicted one step/20 milliseconds ahead LSP values). This research was a joint work between Lancaster University and Nokia-UK and aims the development of the next generation intelligent decoders at the receiver that will be able to conceal lost packets with a size of 80 to 160 ms without significant deterioration to the quality of service (QoS) in VoIP transmission.

Future Directions

The area of evolving fuzzy systems is in its infancy and has already demonstrated a remarkable success in addressing some of the most vibrating issues of the development, application and implementation of truly intelligent systems in a wide variety of branches of industry and real-life problems [11]. It opens the door for future developments that are related to the areas of autonomous systems, early cancer diagnosis and prognosis of the progression, even to the identification of structural changes in biological cells that correspond to the evolution of the disease. In the area of intelligent self-maintaining sensors the process industry can benefit from more flexible and smarter solutions. The problems that are yet to be addressed and can mark the future development of this vibrant area are; (1) collaboration aspects between two or more evolving fuzzy system-based intelligent systems (autonomous robots, intelligent sensors etc.); (2) further flexibility of the systems in terms of real-time self-analysis, optimal features and input selection, rule aggregation mechanism adaptation etc.; (3) even more flexible system structure architectures such as hierarchical, decentralized; (4) more robust learning algorithms that take care of missing data, different sampling intervals etc.

From a broader prospective, the future developments of this discipline will influence and are closely related to similar developments in the area of communication networks (self-adaptive networks [64]), self-validating soft sensors [61], autonomous aerial, ground-based, and underwater vehicles [20,40,49] etc. The area is closely related to the developments in the area of neural networks [34,48], so-called autonomous mental development [23] and cognitive psychology [55], mining data streams [28]. One can also expect more hardware implementations (the first hardware implementation of eClustering was reported in 2005 [8]). From the point of view of mathematical fundamentals and learning it is also closely related to adaptive filters theory [69] and the recent developments in particle filters [17] will certainly influence the future, more ef-

ficient techniques that will be developed in this emerging and highly potential branch of research.

Acknowledgments

The author would like to thank Mr. Xiaowei Zhou for his assistance in producing the illustrative material, Dr. Jose Macias Hernandez for kindly providing real data from the oil refinery CEPSA, Santa Cruz, Tenerife, Spain, Dr. Richard Buswell, Loughborough University and ASHRAE (RP-1020) for the real air conditioning data, and Dr. Edwin Lughofer from Johannes Kepler University of Linz, Austria for providing real data from car engines.

Bibliography

Primary Literature

1. Andersen HC, Teng FC, Tsoi AC (1994) Single Net Indirect Learning Architecture. IEEE Trans Neural Netw 5:1003–1005
2. Angelov P (2002) Evolving Rule-based Models: A Tool for Design of Flexible Adaptive Systems. Springer, Heidelberg
3. Angelov P (2004) An Approach for Fuzzy Rule-base Adaptation using On-line Clustering. Int J Approx Reason 35(3): 275–289
4. Angelov PP (2004) A Fuzzy Controller with Evolving Structure. Inf Sci 161:21–35
5. Angelov P, Buswell R (2001) Evolving Rule-based Models: A Tool for Intelligent Adaptation. In: Proc of the Joint 9th IFSA World Congress and 20th NAFIPS Intern Conf, Vancouver, 25–28 July 2001. IEEE Press, USA, pp 1062–1066
6. Angelov P, Buswell R (2002) Identification of Evolving Rule-based Models. IEEE Trans Fuzzy Syst 10(5):667–677
7. Angelov P, Buswell R (2003) Automatic Generation of Fuzzy Rule-based Models from Data by Genetic Algorithms. Inf Sci 150(1/2):17–31
8. Angelov P, Everett M (2005) EvoMap: On-Chip Implementation of Intelligent Information Modelling using EVOlving MAPping. Lancaster University, Lancaster, pp 1–15
9. Angelov P, Filev D (2004) An approach to on-line identification of evolving Takagi–Sugeno models. IEEE Trans Syst Man Cybern part B Cybern 34(1):484–498
10. Angelov P, Filev D (2005) Simpl_eTS: A Simplified Method for Learning Evolving Takagi–Sugeno Fuzzy Models. In: Proc of The 2005 IEEE Intern. Conf. on Fuzzy Systems FUZZ-IEEE – 2005, Reno 2005, pp 1068–1073
11. Angelov P, Filev D, Kasabov N, Cordon O (eds) (2006) Evolving Fuzzy Systems. Proc of the 2nd Int Symposium on Evolving Fuzzy Systems, Ambleside, 7–9 Sept 2006. pp 1–350
12. Angelov P, Kordon A, Zhou X (2008) Adaptive Inferential Sensors based on Evolving Fuzzy Models: An Industrial Case Study. IEEE Trans Fuzzy Syst (under review)
13. Angelov P, Lughofer E, Klement PE (2005) Two Approaches for Data – Driven Design of Evolving Fuzzy Systems: eTS and FLEXFIS. In: Proc of The 2005 North American Fuzzy Information Processing Society, NAFIPS Annual Conference, Ann Arbor, June 2005, pp 31–35
14. Angelov P, Victor J, Dourado A, Filev D (2004) On-line evolution of Takagi–Sugeno Fuzzy Models. In: Proc of the 2nd IFAC Workshop on Advanced Fuzzy and Neural Control, Oulu, 16–17 Sept 2004, pp 67–72
15. Angelov P, Xydeas C, Filev D (2004) On-line Identification of MIMO Evolving Takagi–Sugeno Fuzzy Models. In: Proc of the Intern. Joint Conf. on Neural Networks and Intern. Conf. on Fuzzy Systems, IJCNN-FUZZ-IEEE, Budapest, 25–29 July 2004, pp 55–60
16. Angelov P, Zhou X, Klawonn F (2007) Evolving Fuzzy Rule-based Classifiers. In: Proc of the First 2007 IEEE International Conference on Computational Intelligence Applications for Signal and Image Processing – a part of the IEEE Symposium Series on Computational Intelligence, SSCI-2007, Honolulu, 1–5 April 2007, pp 220–225
17. Angelov P, Zhou X, Lughofer E, Filev D (2007) Architectures of Evolving Fuzzy Rule-based Classifiers. In: Proc of the 2007 IEEE International Conference on Systems, Man, and Cybernetics, Montreal, 7–10 Oct 2007, pp 2050–2055
18. Arulampalam MS, Maskell S, Gordon N (2002) A Tutorial on Particle Filters for On-line Nonlinear Non-Gaussian Bayesian Tracking. IEEE Trans Signal Process 50(2):174–188
19. Astroem KJ, Wittenmark B (1994) Adaptive Control. Prentice Hall, Upper Saddle River
20. Azimi-Sadjadi MR, Yao D, Jamshidi AA, Dobeck GJ (2002) Underwater Target Classification in Changing Environments Using an Adaptive Feature Mapping. IEEE Trans Neural Netw 13(5):1099–1111
21. Badami VV, Chbat NW (1998) Home appliances get smart. IEEE Spectrum 35(8):36–43
22. Bezdek J (1974) Cluster Validity with Fuzzy Sets. J Cybern 3(3):58–71
23. Bonarini A, Lazaric A, Restelli M, Vitali P (2006) Self-Development Framework for Reinforcement Learning Agents. In: Proc of the 5th Intern. Conf. on Development and Learning, ICDL-06 New Delhi
24. Carline D, Angelov PP, Clifford R (2005) Agile Collaborative Autonomous Agents for Robust Underwater Classification Scenarios. In: the Proceedings of the Underwater Defense Technology Conference, Amsterdam, June 2005
25. Carpenter GA, Grossberg S (2003) Adaptive Resonance Theory. In: Arbib MA (ed) The Handbook of Brain Theory and Neural Networks, 2nd edn. MIT Press, Cambridge, pp 87–90
26. Chiu SL (1994) Fuzzy model identification based on cluster estimation. J Intell Fuzzy Syst 2:267–278
27. Cordon O, Gomide F, Herrera F, Hoffmann F, Magdalena L (2004) Ten years of genetic fuzzy systems: Current framework and new trends. Fuzzy Sets Syst 141(1):5–31
28. Domingos P, Hulten G (2001) Catching up with the data: Research issues in mining data streams. In: Proc of the Workshop on Research Issues in Data Mining and Knowledge Discovery, Santa Barbara
29. Filev D, Larson T, Ma L (2000) Intelligent Control for Automotive Manufacturing – Rule-based Guided Adaptation. In: Proc of the IEEE Conference on Industrial Electronics, IECON-2000, Nagoya Oct 2000, pp 283–288
30. Filev D, Tseng F (2006) Novelty detection-based Machine Health Prognostics. In: Proc of the 2006 Int Symposium on Evolving Fuzzy Systems. IEEE Press, USA, pp 193–199
31. Fortuna L, Graziani S, Rizzo A, Xibilia MG (2007) Soft sensors for Monitoring and Control of In Industrial Processes. Springer, London

32. Goldberg DE (1989) Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading
33. Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, Heidelberg
34. Hornby AS (1974) Oxford Advance Learner's Dictionary. Oxford University Press, Oxford
35. Huang G-B, Saratchandran P, Sundarajan N (2005) A generalized growing and pruning RBF (GGAP – RBF) neural network for function approximation. IEEE Trans Neural Netw 16(1):57–67
36. Ishibuchi H, Nakashima T, Nii M (2004) Classification and Modeling with Linguistic Granules: Advanced Information Processing. Springer, Berlin
37. Jones E, Angelov P, Xydeas C (2006) Recovery of LSP Coefficients in VoIP Systems using Evolving Takagi–Sugeno Fuzzy MIMO Models. In: Proc of the 2006 Intern. Symposium on Evolving Fuzzy Systems, Ambelside, 7–9 Sept 2006, pp 208–214
38. Kailath T, Sayed AH, Hassibi B (2000) Linear Estimation. Prentice Hall, Upper Saddle River
39. Kalman RE (1960) A New Approach to linear filtering and prediction problem. Transactions of the American Society of Mechanical Engineering, ASME, Ser. D. J Basic Eng 8:34–45
40. Kanakakis V, Valavanis KP, Tsourveloudis NC (2004) Fuzzy-Logic Based Navigation of Underwater Vehicles. J Intell Robotic Syst 40:45–88
41. Kasabov N (2001) Evolving fuzzy neural networks for on-line supervised/unsupervised, knowledge-based learning. IEEE Trans. on Systems, Man and Cybernetics – part B. Cybernetics 31:902–918
42. Kasabov N, Song Q (2002) DENFIS: Dynamic Evolving Neural-Fuzzy Inference System and Its Application for Time-Series Prediction. IEEE Trans Fuzzy Syst 10(2):144–154
43. Klir G, Folger T (1988) Fuzzy Sets, Uncertainty and Information. Prentice Hall, Englewood Cliffs
44. Kohonen T (1995) Self-Organizing Maps. Series in Inf Sci, vol 30. Springer, Heidelberg
45. Kordon A (2006) Inferential Sensors as Potential Application Area of Intelligent Evolving Systems. 2006 International Symposium on Evolving Fuzzy Systems, Ambleside, 7–9 September 2006, key note presentation
46. Kuncheva L (2000) Fuzzy Classifiers. Physica, Heidelberg
47. Leng G, McGuinty TM, Prasad G (2005) An approach for online extraction of fuzzy rules using a self-organizing fuzzy neural network. Fuzzy Sets Syst 150(2):211–243
48. Lin F-J, Lin C-H, Shen P-H (2001) Self-constructing fuzzy neural network speed controller for permanent-magnet synchronous motor drives. IEEE Trans Fuzzy Syst 9(5):751–759
49. Liu PX, Meng MQ-X (2004) On-line Data-Driven Fuzzy Clustering with Applications to Real-time Robotic Tracking. IEEE Trans Fuzzy Syst 12(4):516–523
50. Ljung L (1987) System Identification: Theory for the User. Prentice-Hall, New Jersey
51. Lughofer E, Angelov P, Zhou X (2007) Evolving Single- and Multi-Model Fuzzy Classifiers with FLEXFIS-Class. In: Proc of the 2007 IEEE International Conference on Fuzzy Systems, London, 23–26 July 2007, pp 363–368
52. Macias J, Angelov P, Zhou X (2006) Predicting quality of the crude oil distillation using evolving Takagi–Sugeno fuzzy models. In: Proc of the 2006 International Symposium on Evolving Fuzzy Systems, Ambleside, 7–9 Sept 2006, pp 201–207
53. Macias-Hernandez JJ, Angelov P, Zhou X (2007) Soft Sensor for Predicting Crude Oil Distillation Side Streams using Takagi Sugeno Evolving Fuzzy Models. In: Proc of the 2007 IEEE Int Conf on Syst, Man, and Cybernetics, Montreal, 7–10 Oct. 2007, pp 3305–3310
54. Mamdani EH, Assilian S (1975) An Experiment in Linguistic Synthesis with a Fuzzy Logic Controller. Int J Man-Mach Stud 7:1–13
55. Massaro DW (1991) Integration versus Interactive Activation: The Joint Influence of Stimulus and Context in Perception. Cogn Psychol 23:558–614
56. Memon MA, Angelov P, Ahmed H (2006) An Approach to Real-Time Color-based Object Tracking. In: Proc 2006 International Symposium on Evolving Fuzzy Systems, Ambleside, 7–9 Sept 2006, pp 81–87
57. Nauck D, Kruse R (1997) A Neuro-fuzzy method to learn fuzzy classification rules from data. Fuzzy Sets Syst 89:277–288
58. Pioneer-3DX (2004) User Guide. ActiveMedia Robotics, Amherst
59. Procyk TJ, Mamdani EH (1979) A linguistic self-organizing process controller. Automatica 15:15–30
60. Psaltis D, Sideris A, Yamamura AA (1988) A Multilayered Neural Network Controller. IEEE Trans Control Syst Manag 8:17–21
61. Qin SJ, Yue H, Dunia R (1997) Self-validating inferential sensors with application to air emission monitoring. Ind Eng Chem Res 36:1675–1685
62. Setnes M, Roubos H (2000) Ga-fuzzy modeling and classification: complexity and performance. IEEE Trans Fuzzy Syst 8(5):509–522
63. Shimojima K, Fukuda T, Hashegawa Y (1995) Self-Tuning Modeling with Adaptive Membership Function, Rules, and Hierarchical Structure based on Genetic Algorithm. Fuzzy Sets Syst 71:295–309
64. Sifalakis M, Hutchison D (2004) From Active Networks to Cognitive Networks. In: Proc of the ICRC Dagstuhl Seminar 04411 on Service Management and Self-Organization in IP-based Networks. Waden, October 2004
65. Takagi T, Sugeno M (1985) Fuzzy identification of systems and its application to modeling and control. IEEE Trans Syst Man Cybern B – Cybern 15:116–132
66. Valavanis K (2006) Unmanned Vehicle Navigation and Control: A Fuzzy Logic Perspective. In: Proc of the 2006 International Symposium on Evolving Fuzzy Systems. Ambleside, 7–9 Sept. 2006, pp 200–207
67. Vapnik VN (1998) The Statistical Learning Theory. Springer, Berlin
68. Wang L-X (1992) Fuzzy Systems are Universal Approximators. In: Proc of the First IEEE International Conference on Fuzzy Systems, FUZZ-IEEE – 1992, San Diego, pp 1163–1170
69. Widrow B, Stearns S (1985) Adaptive Signal Processing. Prentice Hall, Englewood Cliffs
70. Xydeas C, Angelov P, Chiao S, Reoullas M (2006) Advances in EEG Signals Classification via Dependant HMM models and Evolving Fuzzy Classifiers. Int J Comput Biol Medicine, special issue on Intell Technol Bio-Inform Medicine 36(10):1064–1083
71. Yager R (2006) Learning Methods for Intelligent Evolving Systems. In: Proc 2006 International Symposium on Evolving Fuzzy Systems. Ambelside, 7–9 Sept. 2006, pp 3–7

72. Yager RR, Filev DP (1993) Learning of Fuzzy Rules by Mountain Clustering. In: Proc of the SPIE Conf. on Application of Fuzzy Logic Technology, Boston, pp 246–254
73. Yager RR, Filev DP (1994) Essentials of Fuzzy Modeling and Control. Wiley, New York
74. Zadeh LA (1993) Soft Computing. Introductory Lecture for the 1st European Congress on Fuzzy and Intelligent Technologies EUFIT'93, Aachen, pp vi–vii
75. Zhou X-W, Angelov P (2006) Real-Time joint Landmark Recognition and Classifier Generation by an Evolving Fuzzy System. In: Proc of the 2006 IEEE World Congress on Computational Intelligence, WCCI-2006, Vancouver, 16–21 July 2006, pp 6314–6321
76. Zhou X, Angelov P (2007) An approach to autonomous self-localization of a mobile robot in completely unknown environment using evolving fuzzy rule-based classifier. In: Proc of the First 2007 IEEE Int Symposium on Computational Intelligence Applications for Defense and Security – a part of the IEEE Symposium Series on Computational Intelligence, SSCI-2007, Honolulu, 1–5 April 2007, pp 131–138

Books and Reviews

Angelov P, Xydeas C (2006) Fuzzy Systems Design: Direct and Indirect Approaches. Int J Soft Comput, special issue on New Trends in Fuzzy Modeling part I: Novel Approaches 10(9): 836–849

Angelov P, Zhou X (2006) Evolving fuzzy systems from data streams in real-time. Proc 2006 International Symposium on Evolving Fuzzy Systems, Ambleside, 7–9 Sept. 2006, pp 29–35

Bentley PJ (2000) Evolving Fuzzy Detectives: An Investigation into the Evolution of Fuzzy Rules. In: Suzuki, Roy, Ovasks, Furuhashi, Dote (eds) Soft Computing in Industrial Applications. Springer, London

Chan Z, Kasabov N (2004) Evolutionary computation for on-line and off-line parameter tuning of evolving fuzzy neural networks. Int J Comput Intell Appl 4(3):309–319

Fayyad UM, Piatetsky-Shapiro G, Smyth P (1996) From Data Mining to Knowledge Discovery: An Overview, Advances in Knowledge Discovery and Data Mining. MIT Press, Boston

Fritzke B (1994) Growing cell structures – a self-organizing network for unsupervised and supervised learning. Neural Netw 7(9):1441–1460

Futschik M, Reeve A, Kasabov N (2003) Evolving connectionist systems for knowledge discovery from gene expression data of cancer tissue. Artif Intell Med 28:165–189

Hopner F, Klawonn F (2000) Obtaining interpretable fuzzy models from fuzzy clustering and fuzzy regression. In: Proc of the 4th Intern Conf on Knowledge-based Intelligent Engineering Systems (KES), Brighton, pp 162–165

Huang G-B, Saratchandran P, Sundarajan N (2005) A generalized growing and pruning RBF (GGAP-RBF) neural network for function approximation. IEEE Trans Neural Netw 16(1):57–67

Huang L, Song Q, Kasabov N (2005) Evolving Connectionist Systems Based Role Allocation of Robots for Soccer Playing. In: Proc of the Joint 2005 International Symposium on Intelligent Control and 13th Mediterranean Conference on Control and Automation, ISIC – MED – 2005, Limassol, 27–29 June 2005

Jang JSR (1993) ANFIS: Adaptive Network-based Fuzzy Inference Systems. IEEE Trans. on Syst Man Cybernt B – Cybern 23(3):665–685

Juang C-F, Lin X-T (1999) A recurrent self-organizing neural fuzzy inference network. IEEE Trans Neural Netw 10:828–845

Kasabov N (2006) Adaptation and Interaction in Dynamical Systems: Modelling and Rule Discovery Through Evolving Connectionist Systems. Appl Soft Comput 6(3):307–322

Kasabov N (2006) Evolving connectionist systems: Brain-, gene-, and, quantum inspired computational intelligence. Springer, London

Kasabov N, Chan Z, Song Q, Greer D (2005) Evolving neuro-fuzzy systems with evolutionary parameter self-optimisation. In: Do Adaptive Smart Systems exist? Series Study in Fuzziness, vol 173. Physica, Heidelberg

Kim K, Baek J, Kim E, Park M (2005) TSK Fuzzy model based on-line identification. In: Proc of the 11th International Fuzzy Systems Association, IFSA World Congress, Beijing, pp 1435–1439

Klinkenberg R, Joachims T (2000) Detection concept drift with support vector machines. Proc of the 7th International Conference on Machine Learning (ICML). Morgan Kaufman, Stanford University, pp 487–494

Marin-Blazquez JG, Shen Q (2002) From approximative to descriptive fuzzy classifiers. IEEE Trans Fuzzy Syst 10(4):484–497

Marshall MR, Song Q, Ma TM, MacDonell S, Kasabov N (2005) Evolving Connectionist System versus Algebraic Formulae for Prediction of Renal Function from Serum Creatinine. Kidney Int 6:1944–1954

Ozawa S, Pang S, Kasabov N (2004) A Modified Incremental Principal Component Analysis for On-Line Learning of Feature Space and Classifier. Lecture Notes in Artificial Intelligence LNAI, vol 3157. Springer, Berlin, pp 231–240

Pang S, Ozawa S, Kasabov N (2005) Incremental Linear Discriminant Analysis for Classification of Data Streams. IEEE Trans Syst Man Cybern B – Cybern 35(5):905–914

Plat J (1991) A resource allocation network for function interpolation. Neural Comput 3(2):213–225

Widmer G, Kubat M (1996) Learning in the presence of concept drift and hidden contexts. Mach Learn 23(1):69–101

Existence and Uniqueness of Solutions of Initial Value Problems

Gianne Derks
Department of Mathematics,
University of Surrey, Guildford, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Existence
Uniqueness
Continuous Dependence on Initial Conditions
Extended Concept of Differential Equation
Further Directions
Bibliography

Glossary

Ordinary differential equation An ordinary differential equation is a relation between a (vector) function $u: I \to \mathbb{R}^m$, (I an interval in $\mathbb{R}$, $m \in \mathbb{N}$) and its derivatives. A function u, which satisfies this relation, is called a solution.

Initial value problem An initial value problem is an ordinary differential equation with a prescribed value for the solution at one instance of its variable, often called the initial time. The initial value is the pair consisting of the initial time and the prescribed value of the solution.

Vector field For an ordinary differential equation of the form $\frac{\mathrm{d}u}{\mathrm{d}t} = f(t, u)$, the function $f: \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m$ is called the vector field. Thus, at a solution, the vector field gives the tangent to the solution.

Flow map The flow map describes the solution of an ordinary differential equation for varying initial values. Hence, for an ordinary differential equation $\frac{\mathrm{d}u}{\mathrm{d}t} = f(t, u)$, the flow map is a function $\Phi: \mathbb{R} \times \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m$ such that $\Phi(t, t_0, u_0)$ is a solution of the ordinary differential equation, starting at $t = t_0$ in $u = u_0$.

Functions: bounded, continuous, uniformly Lipschitz continuous Let D be a connected set in $\mathbb{R}^k$ with $k \in \mathbb{N}$ and let $f: D \to \mathbb{R}^m$ be a function on D:

- The function f is *bounded* if there is some $M > 0$ such that $|f(x)| \leq M$ for all $x \in D$.
- The function f is *continuous* on D if $\lim_{x \to x_0} f(x) = f(x_0)$, for every $x_0 \in D$. A continuous function on a bounded and closed set D is bounded. The function f is *equicontinuous* or *uniform continuous* if the convergence of the limits in the definition of continuity is uniform for all $x_0 \in D$, i. e., for every $\varepsilon > 0$, there is a $\delta > 0$, such that for all $x_0 \in D$ and all $x \in D$ with $|x - x_0| < \delta$ it holds that $|f(x) - f(x_0)| < \varepsilon$. A continuous function on a compact interval is equicontinuous.
- Let $D \subset \mathbb{R} \times \mathbb{R}^m$. The function $f: D \to \mathbb{R}^m$ is *uniformly Lipschitz continuous on D with respect to its second variable*, if f is continuous on D and there exists some constant $L > 0$ such that

$$|f(t, u) - f(t, u)| \leq L|u - v| \,, \qquad \text{for all } (t, u),\ (t, v) \in D \,.$$

  The constant L is called the *Lipschitz constant*.

Pointwise and uniform convergence A sequence of functions $\{u_n\}$ with $u_n: I \to \mathbb{R}^m$ is Pointwise convergent if $\lim_{n\to\infty} u_n(t)$ exists for every $t \in I$. The sequence of functions $\{u_n\}$ is uniform convergent with limit function $u: I \to \mathbb{R}^m$ if $\lim_{n\to\infty} \sup\{|u - u_n| \mid t \in I\} = 0$. A sequence of pointwise convergent, equicontinuous functions is uniform convergent and the limit function is equicontinuous.

Notation

$\dot{u}$	Derivative of u, i. e., $\frac{\mathrm{d}u}{\mathrm{d}t}$
$u^{(k)}$	The kth derivative of u, i. e., $\frac{\mathrm{d}^k u}{\mathrm{d}t^k}$
$I_a(t_0)$	The closed interval $[t_0, t_0 + a]$
$B_b(u_0)$	The closed ball with radius b about u_0 in $\mathbb{R}^m$, i. e., $B_b(u_0) := \{u \in \mathbb{R}^m \mid \lvert u - u_0\rvert \leq b\}$
$\lVert u\rVert_\infty$	The supremum norm for a bounded function $u: I \to \mathbb{R}^m$, i. e., $\lVert u\rVert_\infty = \sup\{\lvert u(t)\rvert \mid t \in I\}$

Definition of the Subject

Many problems in physics, engineering, biology, economics, etc., can be modeled as relations between observables or states and their derivatives, hence as differential equations. When only derivatives with respect to one variable play a role, the differential equation is called an ordinary differential equation. The field of differential equations has a long history, starting with Newton and Leibniz in the seventeenth century. In the beginning of the study of differential equations, the focus is on finding explicit solutions as the emphasis is on solving the underlying physical problems. But soon one starts to wonder: If a starting point for a solution of a differential equation is given, does the solution always exist? And if such a solution exists, how long does it exist and is there only one such solution? These are the questions of existence and uniqueness of solutions of initial value problems. The first existence result is given in the middle of the nineteenth century by Cauchy. At the end of the nineteenth century and the beginning of the twentieth century, substantial progress is made on the existence and uniqueness of solutions of initial value problems and currently the heart of the topic is quite well understood. But there are many open questions as soon as one considers delay equations, functional differential equations, partial differential equations or stochastic differential equations. Another area of intensive current research, which uses the existence and uniqueness of differential equations, is the area of finite- and infinite-dimensional dynamical systems.

Introduction

As indicated before, an initial value problem is the problem of finding a solution of an ordinary differential equation with a given initial condition. To be precise, let

D be an open, connected set in $\mathbb{R} \times \mathbb{R}^m$. Given are a function $f\colon D \to \mathbb{R}^m$ and a point $(t_0, u_0) \in D$. The initial value problem is the problem of finding an interval $I \subset \mathbb{R}$ and a function $u\colon I \to \mathbb{R}^m$ such that $t_0 \in I$, $\{(t, u(t)) \mid t \in I\} \subset D$ and

$$\frac{\mathrm{d}u}{\mathrm{d}t} = f(t, u(t)) \,, \quad t \in I \,, \quad \text{with } u(t_0) = u_0 \,. \tag{1}$$

The function f is called the vector field and the point (t_0, u_0) the initial value. One often writes the derivative as $\frac{\mathrm{d}u}{\mathrm{d}t} = \dot{u}$. If f is continuous, then the initial value problem (1) is equivalent to finding a function $u\colon I \to \mathbb{R}^m$ such that

$$u(t) = u_0 + \int_{t_0}^{t} f(\tau, u(\tau))\mathrm{d}\tau \,, \quad \text{for } t \in I \,. \tag{2}$$

If f is continuous and a solution exists, then clearly this solution is continuously differentiable on I.

It might seem restrictive to only consider first-order systems; however, most higher-order equations can be written as a system of first-order equations. Consider for example the nth order differential equation for $v\colon I \to \mathbb{R}$

$$v^{(n)}(t) = F\left(t, v, \dot{v}, \ldots, v^{(n-1)}\right) \,,$$

with initial conditions

$$v^{(k)}(t_0) = v_k \,, \quad k = 0, \ldots, n-1 \,.$$

This can be written as a first-order system by using the vector function $u\colon I \to \mathbb{R}^m$ defined as $u = (v, \dot{v}, \ldots, v^{(n-1)})$. The equivalent initial value problem for u is

$$\dot{u} = f(t, u) = \begin{pmatrix} u_2 \\ \vdots \\ u_n \\ F(t, u_1, u_2, \ldots, u_n) \end{pmatrix} \,,$$

$$\text{with } u(t_0) = \begin{pmatrix} v_0 \\ v_1 \\ \vdots \\ v_{n-1} \end{pmatrix} \,.$$

Obviously, this system is not a unique representation, there are many other ways of obtaining first-order systems from the nth-order problem.

In the beginning of the study of differential equations, the focus is on finding explicit solutions. The first existence theorem is by Cauchy [3] in the middle of the nineteenth century and an initial value problem is also called a *Cauchy problem*. At the end of the nineteenth century substantial progress is made on the existence of solutions of an initial value problem when Peano [20] shows that if f is continuous, then a solution exists near the initial value (t_0, u_0), i. e., there is local existence of solutions. Global existence of solutions of initial value problems needs more than smoothness, as is illustrated by the following example.

Example 1 Consider the initial value problem on $D = \mathbb{R} \times \mathbb{R}$ given as

$$\dot{u} = u^2 \quad \text{and} \quad u(0) = u_0$$

for some $u_0 > 0$. As can be verified easily, the solution is $u(t) = u_0/(1 - tu_0)$, for $t \in (-\infty, 1/u_0)$. This solution cannot be extended for $t \geq 1/u_0$, even though the vector field $f(t, u) = u^2$ is infinitely differentiable on the full domain. As can be seen from later theorems, the lack of global existence is related to the fact that the vector field f is unbounded.

Once existence of solutions of initial value problems is known, the next question is if such a solution is unique. As can be seen from the following example, continuity of f is not sufficient for uniqueness of solutions.

Example 2 Consider the following initial value problem on $D = \mathbb{R} \times \mathbb{R}$:

$$\dot{u} = |u|^\alpha \quad \text{and} \quad u(0) = 0 \,.$$

If $0 < \alpha < 1$, then there is an infinite number of solutions. Two obvious solutions for $t \in [0, \infty)$ are $\overline{u}(t) = 0$ and $\hat{u}(t) = ((1-\alpha)t)^{\frac{1}{1-\alpha}}$. But these solutions are members of a large family of solutions. For any $c \geq 0$, the following functions are solutions for $I = \mathbb{R}$:

$$u_c(t) = \begin{cases} 0 \,, & t < c \\ ((1-\alpha)(t-c))^{\frac{1}{1-\alpha}} \,, & t \geq c \,. \end{cases}$$

A sufficient condition for uniqueness of solutions of the initial value problem is uniform Lipschitz continuity of the vector field f in its second variable u. Although the ideas behind this theorem go back to Cauchy and Lipschitz, Picard [21] and Lindelöf [18] are usually credited with this result. They used the so-called method of successive approximations to prove the uniqueness result; see Sect. "Uniqueness" for details.

Having determined existence and uniqueness, the next issue is the relation between a solution and its initial value. This is the topic of Sect. "Continuous Dependence on Initial Conditions". The concept of a flow map is first introduced. Roughly speaking, the flow map is the solution,

with the initial value as an extra variable. If the vector field is uniformly Lipschitz continuous in its second variable, then the flow map is continuous. Thus, a small change in the initial value, will only give a small change in the solution (locally).

The concept of an initial value problem can be extended to noncontinuous vector fields f and existence theorems can be shown if the vector fields satisfy the so-called Carathéodory conditions. More details can be found in Sect. "Extended Concept of Differential Equation". This chapter ends with a discussion of initial value problems in more general differential equations and the use of existence and uniqueness of solutions of initial value problems in dynamical systems.

Existence

Cauchy [3] seems to have been the first one to publish an existence result for differential equations, using an approximation of solutions by joined line segments. This work was extended considerably by Peano [20] in 1890.

Theorem 1 (Cauchy–Peano Theorem) *Let $a, b > 0$, define $I_a(t_0) := [t_0, t_0 + a]$ and $B_b(u_0) := \{u \in \mathbb{R}^m \mid |u - u_0| \le b\}$ and assume that the cylinder $S := I_a(t_0) \times B_b(u_0) \subset D$. Let the vector field f be a continuous function on the cylinder S. This implies that f is bounded on S, say, $M = \max\{|f(t,u)| \mid (t,u) \in S\}$.*

Define the parameter $\rho = \min\left(a, \frac{b}{M}\right)$. Then there exists a solution $u(t)$, with $t_0 \le t \le t_0 + \rho$, which solves the initial value problem (1).

The theorem can also be phrased for an interval $[t_0 - \rho, t_0]$ or $[t_0 - \rho, t_0 + \rho]$. The parameter ρ gives a time interval such that that the solution $(t, u(t))$ is guaranteed to be within S. From the integral formulation of the initial value problem (2), it follows that

$$|u(t) - u_0| \le \int_{t_0}^{t} |f(\tau, u(\tau))| \mathrm{d}\tau \le M(t - t_0) .$$

To guarantee that $u(t)$ is within S, the condition $t - t_0 \le \rho$ is sufficient as it gives $M(t - t_0) \le b$ and $t - t_0 \le a$. In Fig. 1, this is illustrated in the case $D \subset \mathbb{R} \times \mathbb{R}$.

To prove the Cauchy–Peano existence theorem, a construction which goes back to Cauchy is used. This so-called Cauchy–Euler construction of approximate solutions uses joined line segments which are such that the tangent of each line segment is given by the vector field evaluated at the start of the line segment. To be specific, for any $N \in \mathbb{N}$, the interval $I_\rho(t_0) = [t_0, t_0 + \rho]$ is divided in N equal parts with intermediate points $t_k := t_0 + \frac{k}{N}\rho$, $k = 1, \ldots, N$. The approximate solution is the function $u_N : I_\rho(t_0) \to \mathbb{R}$, with

$$\begin{aligned} u_N(t_0) &= u_0 , \\ u_N(t) &= u_N(t_{k-1}) + f(t_{k-1}, u_N(t_{k-1}))\,(t - t_{k-1}) , \\ t_{k-1} &< t \le t_k , \quad k = 1, \ldots, N , \end{aligned} \tag{3}$$

see also Fig. 2. One sees immediately that function u_N is continuous and piecewise differentiable and, using the bound on the vector field f,

$$|u_N(t) - u_N(s)| \le M|t - s| , \quad \text{for all } t, s \in I_\rho(t_0) . \tag{4}$$

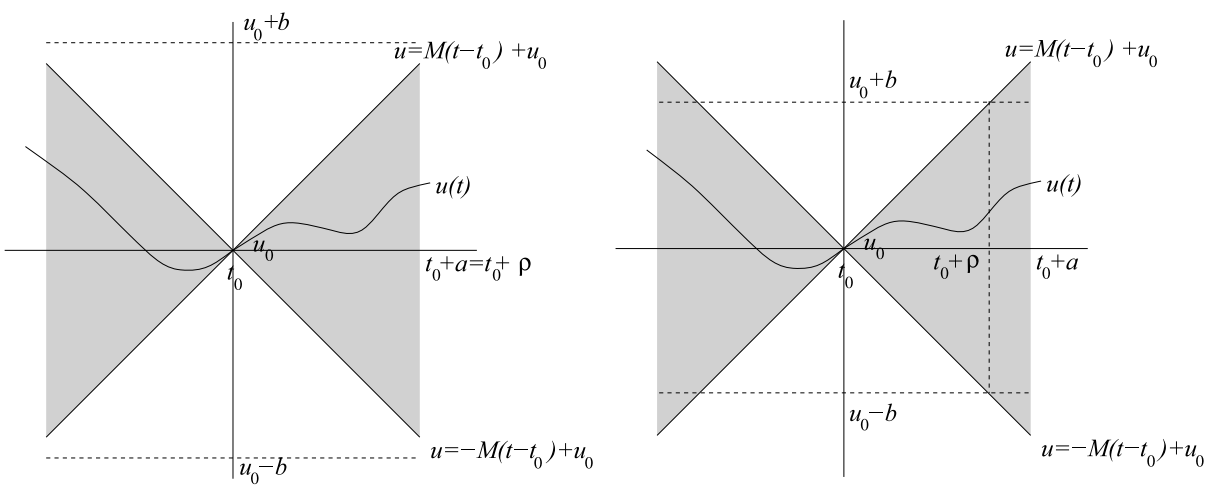

Existence and Uniqueness of Solutions of Initial Value Problems, Figure 1
The parameter $\rho = \min(\frac{b}{M}, a)$ guarantees that the solution is within the *shaded area S, on the left* in the case $a < \frac{b}{M}$ and *on the right* if $a > \frac{b}{M}$

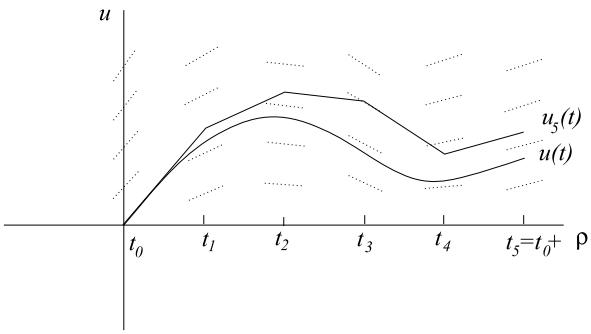

Existence and Uniqueness of Solutions of Initial Value Problems, Figure 2
The approximate solution u_N, with $N = 5$, and the solution $u(t)$. The *dotted lines* show the tangent given by the vector field $f(t, u)$ at the points (t_i, u) for various values of u and $i = 0, \ldots, 5$

This estimate implies that the sequence of functions $\{u_N\}$ is uniformly bounded and that the functions u_N are equicontinuous. Indeed, the estimate above gives that for any $t \in I_\rho(t_0)$ and any $N \in \mathbb{N}$,

$$|u_N(t)| \le |u_N(t_0)| + M|t - t_0| \le |u_0| + M\rho\,.$$

And equicontinuity follows immediately from (4) as M does not depend on N.

Arzela–Ascoli's lemma states that a sequence of functions which is uniformly bounded and equicontinuous on a compact set has a subsequence which is uniformly convergent on this compact set [5]. As the sequence $\{u_N\}$ is uniformly bounded and equicontinuous on the compact set $I_\rho(t_0)$, it follows immediately that it has a convergent subsequence. This convergent subsequence is denoted by u_{N-k} and its limit by u. To prove the Cauchy–Peano theorem, we will show that this limit function u satisfies the integral equation (2) and hence is a solution of the initial value problem (1).

Proof First it will be shown that if N is sufficiently large, then the functions u_N are close to solutions of the differential equation in the following sense: for all $\varepsilon > 0$, there is some N_0 such that for all $N > N_0$

$$\begin{aligned} &|\dot{u}_N(t) - f(t, u_N(t))| < \varepsilon\,, \\ &t \in I_\rho(t_0)\,, \quad t \ne t_k\,, \quad k = 0, \ldots, N\,. \end{aligned} \tag{5}$$

Let $\varepsilon > 0$. As f is a continuous function on the compact set S, it follows that f is uniform continuous. Thus, there is some δ_ε such that for all (t, u), $(s, v) \in S$ with $|t - s| + |u - v| < \delta_\varepsilon$

$$|f(t, u) - f(s, v)| < \varepsilon\,. \tag{6}$$

Now define $N_0 = \left\lceil \frac{(M+1)\rho}{\delta_n} \right\rceil$ and let $N > N_0$. Then for any $t \in I_\rho(t_0)$, $t \ne t_k$, $k = 0, \ldots, N$, we have $\dot{u}_N(t) = f(t_l, u_N(t_l))$, where l is such that $t_l < t < t_{l+1}$. Hence, (6) gives

$$|\dot{u}_N(t) - f(t, u_N(t))| = |f(t_l, u_N(t_l)) - f(t, u_N(t))| < \varepsilon$$

as (4) shows that $|t_l - t| + |u_N(t_l) - u_N(t)| < (1 + M)|t_l - t| < \frac{(1+M)\rho}{N} \le \frac{(1+M)\rho}{N_0} \le \delta_\varepsilon$.

Next it will be shown that (5) implies that the functions u_N almost satisfy the integral equation (2) for N sufficiently large. From (5), it follows that $f(t, u_N(t)) - \varepsilon < \dot{u}_N(t) < f(t, u_N(t)) + \varepsilon$ for all $N > N_0$ and all $t \in I_\rho(t_0)$, except at the special points t_k, $k = 0, \ldots, N$. Hence, for any $k = 1, \ldots, N$ and all $t_{k-1} < t \le t_k$, this gives

$$\begin{aligned} u_N(t) &< u_N(t_{k-1}) + \int_{t_{k-1}}^{t} \big(f(\tau, u_N(\tau)) + \varepsilon\big)\,\mathrm{d}\tau \\ &\le u_N(t_{k-1}) + \int_{t_{k-1}}^{t} f(\tau, u_N(\tau))\mathrm{d}\tau + \frac{\varepsilon\rho}{N}\,. \end{aligned}$$

Thus, also for any $k = 1, \ldots, N$, $u_N(t_k) - u_N(t_{k-1}) < \int_{t_{k-1}}^{t_k} f(\tau, u_N(\tau))\mathrm{d}\tau + \frac{\varepsilon\rho}{N}$ and hence

$$\begin{aligned} u_N(t_k) - u_N(t_0) &= \sum_{j=1}^{k} \big[u_N(t_j) - u_N(t_{j-1})\big] \\ &< \int_{t_0}^{t_k} f(\tau, u_N(\tau))\mathrm{d}\tau + \varepsilon\rho\,. \end{aligned}$$

Combination of the last two results gives that for any $t \in I_\rho(t_0)$

$$u_N(t) < u_N(t_0) + \int_{t_0}^{t} \big(f(\tau, u_N(\tau)) + \varepsilon\big)\,\mathrm{d}\tau + \varepsilon\rho\,.$$

In a similar way it can be shown that $u_N(t) > u_N(t_0) + \int_{t_0}^{t} f(\tau, u_N(\tau))\mathrm{d}\tau - \rho\varepsilon$ and hence for all $N > N_0$, the following holds:

$$\left| u_N(t) - u_N(t_0) - \int_{t_0}^{t} f(\tau, u_N(\tau)\mathrm{d}\tau \right| < \varepsilon\rho\,, \quad \text{for all } t \in I_\rho(t_0)\,. \tag{7}$$

Thus, the function u_N satisfies the integral equation (2) up to an order ε error if N is sufficiently large.

As the subsequence u_{N_k} converges uniformly to u and f is continuous, this implies that $\int_{t_0}^{t} f(\tau, u_{N_k}(\tau)) \to \int_{t_0}^{t} f(\tau, u(\tau))\mathrm{d}\tau$. Thus, from (7), it can be concluded that u satisfies the integral equation (2) exactly. □

Note that the full sequence $\{u_N\}$ does not necessarily converge. A counterexample can be found in exercise 12 in Chap. 1 of [5]. It is based on the initial value problem $\dot{u} = |u|^{\frac{1}{4}} \operatorname{sgn}(u) + t \sin(\pi/t)$ with $u(0) = 0$, and shows that on a small interval near $t_0 = 0$, the even approximations are bounded below by a strictly positive constant, while the odd approximations are bounded above by a strictly negative constant.

However, if it is known that the solution of the initial value problem is unique, then the full sequence of approximate solutions $\{u_N\}$ must converge to the solution. This can be seen by a contradiction argument. If there is a unique solution, but the sequence $\{u_N\}$ is not convergent, then there is a convergent subsequence and a remaining set of functions which does not converge to the unique solution. But Arzela–Ascoli's lemma can be applied to the remaining set as well; thus, there exists a convergent subsequence in the remaining set, which must converge to a different solution. This is not possible as the solution is unique.

In case of the initial value problem as presented in Example 2, the sequence of Cauchy–Euler approximate solutions $\{u_N\}$ converges to the zero function; hence, no subsequence is required to obtain a solution. As there are many solutions of this initial value problem, this implies that not all solutions of the initial value problem in Example 2 solutions can be approximated by the Cauchy–Euler construction of approximate solutions as defined in (3). As indicated in exercise 2.2 in Chap. 2 of [13], it is possible to get any solution by using appropriate Lipschitz continuous approximations of the vector field f, instead of the original vector field itself, in the Cauchy–Euler construction. In the example below this is illustrated for the vector field $f(t,u) = \sqrt{|u|}$.

Example 3 Let $f\colon \mathbb{R} \times \mathbb{R}$ be defined as $f(t,u) = \sqrt{|u|}$. First we define the approximate functions $\hat{f}_n\colon [0,2] \times [0,\infty) \to [0,\infty)$ as

$$\hat{f}_n(t,u) = \begin{cases} \sqrt{u}\,, & u > \frac{1}{n} \\ \frac{1}{\sqrt{n}} - \frac{1}{2}\sqrt{n}\left(\frac{1}{n} - u\right)\,, & u \le \frac{1}{n} \end{cases}$$

and the definition of the Euler–Cauchy approximate functions is modified to

$$u_N(t) = u_N(t_{k-1}) + \hat{f}_N(t_{k-1}, u_N(t_{k-1}))\,(t - t_{k-1}),$$
$$t_{k-1} < t \le t_k\,, \quad k = 1, \ldots, N\,.$$

This sequence converges to the solution $u(t) = 4t^2$ for $t \in [0,2]$.

Next define the approximate functions $\tilde{f}_n\colon [0,2] \times [0,\infty) \to [0,\infty)$ as

$$\tilde{f}_n(t,u) = \begin{cases} \sqrt{u}\,, & t \le 1\,,\ u > \frac{1}{n}\,, \\ \frac{1}{2n} - \left(\frac{1}{2n} - u\right) \\ \quad + a_2\left(\frac{1}{2n} - u\right)^2 \\ \quad + a_3\left(\frac{1}{n} - u\right)^3\,, & t \le 1\,,\ \frac{1}{2n} < u \le \frac{1}{n}\,, \\ u\,, & t \le 1\,,\ u \le \frac{1}{2n}\,, \end{cases}$$

with $a_2 = 4n(\sqrt{n}-1)$ and $a_3 = 12n^3(3\sqrt{n}-2)$, $\tilde{f}_n(t,u) = \hat{f}_n(t,u)$, for $t \ge 1 + \frac{1}{n}$ and and a smooth connection between those two components of the approximation for $a < t < 1 + \frac{1}{n}$. Then the related approximate solutions will converge to the solution $u(t) = 4(t-1)^2$, $1 \le t \le 2$ and $u(t) = 0$, $0 \le t < 1$. In a similar way all other solutions presented in Example 2 can be obtained.

Example 2 gives a connected and closed family of solutions to the initial value problem. The following theorem shows this is typical.

Theorem 2 ([15]) *Assume that the assumptions of Theorem 1 are satisfied. Let $t_0 < c \le t_0 + \rho$ and define A_c to be the set of points that can be reached at time $t = c$ by some solution of the initial value problem. Then A_c is closed, bounded and connected.*

If the initial value problem is one-dimensional, i. e., $u \in \mathbb{R}$, then this implies that the set of points reached by possible solutions at time $t = c$ is the empty set, a point or a closed interval.

A proof of this theorem, based on [19], can be found in Theorem 4.1 in Chap. 2 of [13].

The existence theorems presented so far give existence of solutions near the initial value (t_0, u_0). But this local result can be extended to a global one.

Theorem 3 (Extension Theorem) *Let f be continuous on D. If $u(t)$ is a solution of the initial value problem (1) on some interval, then the solution can be extended to a maximal interval of existence (t_-, t_+) such that $(t, u(t))$ converges to the boundary of D as $t \uparrow t_+$ or $t \downarrow t_-$ (where $t_\pm = \pm\infty$ is possible).*

Proof Let D_n be open subsets of D such that $\cup_{n\in\mathbb{N}} D_n = D$, the closures $\overline{D_n}$ are compact and $\overline{D_n} \subset D_{n+1}$ (e. g., $D_n = \{(t,u) \in \mathbb{R} \mid |(t,u)| < n, \operatorname{dist}((t,u), \partial D) > 1/n\}$, see [13]). First it will be shown that for all $n \in \mathbb{N}$, there is some ρ_n such that for each $(t_0, u_0) \in D_n$, the initial value problem has a solution on the interval

$[t_0 - \rho_n, t_0 + \rho_n]$. Indeed, for each $n \in \mathbb{N}$, the function f is continuous on the compact set $\overline{D_n}$. Hence, there is some $M_n > 0$ such that $|f(t,u)| \le M_n$ for all $(t,u) \in \overline{D_n}$. Furthermore, for each $n \in \mathbb{N}$, the distance $d_n := \mathrm{dist}(\overline{D_n}, D_{n+1}) > 0$. Thus, the Cauchy–Peano theorem implies that for each $(t_0, u_0) \in \overline{D_n}$, the solution of the initial value problem exists on the interval $[t_0 - \rho_n, t_0 + \rho_n]$, with $\rho_n = 1/2 \min(d_n/M_{n+1}, d_n)$.

If the solution $u(t)$ is defined on an interval I, which is not a right maximal interval, then the argument above shows that the right endpoint of I can be included in the interval of existence. So it can be assumed that the interval I is of the form $[a_1, a_2]$. The continuity of the solution u gives that the set $\{(t, u(t)) \mid t \in [a_1, a_2]\}$ is a compact set in the open set D; hence, there is some $n \in \mathbb{N}$ such that $\{(t, u(t)) \mid t \in [a_1, a_2]\} \subset D_n \subset \overline{D_n}$. But this implies that the solution can be extended to the interval $[a_1, 1_2 + \rho_n]$. And if $\{(t, u(t)) \mid t \in [a_1, a_2 + \rho_n]\} \subset D_n$, then this can be repeated. Since $\overline{D_n}$ is compact, there is some $k \in \mathbb{N}$ such that $\{(t, u(t)) \mid t \in [a_1, a_2 + k\rho_n]\} \not\subset D_n$; hence, there exists some $(t_n, u(t_n))$ such that $(t_n, u(t_n)) \in D_{n+1} \backslash D_n$. This argument can be repeated for each $n \in \mathbb{N}$ to give sequence $(t_n, u(t_n)) \in D_{n+1} \backslash D_n$, with $t_n < t_{n+1}$. Thus, the sequence t_n is monotone increasing. Either it is unbounded, which implies that the solution exists on an interval $[a_1, \infty)$ and hence $t_+ = \infty$ or it is bounded and hence is convergent to some limit t_+. Similarly, the sequence $(t_n, u(t_n))$ is either an unbounded sequence in D or has a cluster point on the boundary of D. In either case it is clear that the solution cannot be extended outside the interval $[a_1, t_+)$. A similar argument can be used for the left endpoint and it can be concluded that there exists a maximal interval of existence of the form (t_-, t_+).

Finally it will be shown that the solution converges to the boundary of D by a contradiction argument. Assume that the solution does not converge to the boundary if $t \uparrow t_+$. Then there is some $\varepsilon_0 > 0$ and a sequence $\{\tau_n\}$ with $t_+ - \tau_n < \frac{1}{n}$ and $d(\tau_n, u(\tau_n)), D) > \varepsilon_0$. This implies that there is some $N \in \mathbb{N}$ such that the sequence $\{(\tau_n, u(\tau_n))\} \subset D_N$. With the arguments above, this implies that the solution can be extend to an interval including t_+, which contradicts t_+ being the maximal right point. □

Example 1 gives an example of a solution for which $u(t) \to \infty$ if $t \to \frac{1}{u_0}$, the boundary of its interval of existence. Example 2 gives an example of a solution with an unbounded interval of existence. Example 4 shows that the endpoint can also be related to the failure of continuity of f.

Example 4 Consider the initial value problem with $D = \mathbb{R} \times (0, \infty)$ and

$$\dot{u} = -\frac{1}{u}, \quad \text{and} \quad u(0) = 2\,.$$

Then the solution is $u(t) = \sqrt{4 - 2t}$ for $t \in (-\infty, 2)$. If $t \to 2$, then $(t, u(t)) \to (2, 0)$, hence the boundary of D and the point where the continuity of f fails.

If the initial value problem has a nonunique solution, then the maximal interval of extension in the extension theorem will in general depend on the initial solution $u(t)$.

Uniqueness

Uniqueness follows if the vector field f is Lipschitz continuous with respect to its second variable u as shown first by Picard [21] and Lindelöf [18]. Their proof is based again on approximate solutions, but the approximate solutions are smoother than the Cauchy–Euler ones. The approximate solutions are obtained by successive iterations and are defined as

$$u_0(t) := u_0 \quad \text{and}$$
$$u_{n+1}(t) := u_0 + \int_{t_0}^{t} f(\tau, u_n(\tau))\mathrm{d}\tau\,, \quad n \in \mathbb{N}\,, \quad t \in I\,. \tag{8}$$

Sometimes this iteration is called Picard iteration. Clearly, this iteration is based on the integral formula (2). The iteration process fails if $u_n(\tau) \notin D$ for some $\tau \in I$. If f is continuous, then there is some interval I for which the sequence is well-defined, as follows from the proof below.

Theorem 4 (Picard–Lindelöf Theorem) *Let $a, b > 0$ be such that the cylinder $S := I_a(t_0) \times B_b(u_0) \subset D$. Let the vector field f be a uniformly Lipschitz continuous function on the cylinder S with respect to its second variable u and let M be the upper bound on f, i.e., $M = \max\{f(t,u) \mid (t,u) \in S\}$.*

Define the parameter $\rho = \min\left(a, \frac{b}{M}\right)$. For $t \in [t_0, t_0 + \rho]$, there exists a unique solution $u(t)$ of the initial value problem (1).

Successive iterations play an important role in proving existence for more general differential equations. Thus, although the existence of solutions already follows from the Cauchy–Peano theorem, we will prove again existence by using the successive iterations instead of the Cauchy–Euler approximations as the ideas in the proof can be extended to more general differential equations.

Proof By using the bound on the vector field f, we will show that if $(t, u_n(t)) \in S$ for all $t \in [t_0, t_0 + \rho_0]$, then $(t, u_{n+1}(t)) \in S$ for all $t \in [t_0, t_0 + \rho_0]$. Indeed, for any $t \in [t_0, t_0 + \rho_0]$,

$$\begin{aligned} |u_{n+1}(t) - u_0| &\le \int_{t_0}^{t} |f(\tau, u_n(\tau))| \mathrm{d}\tau \\ &\le \int_{t_0}^{t} M \mathrm{d}\tau \le M(t - t_0) \le b\,; \end{aligned} \tag{9}$$

thus, $|u_{n+1}(t) - u_0| \le M\rho \le b$ and therefore $u_{n+1}(t) \in B_b(u_0)$.

The boundedness of f also ensures that the functions $\{u_n\}$ are equicontinuous as for any $n \in \mathbb{N}$ and any $t_0 \le t_1 < t_2 \le \rho + t_0$:

$$|u_n(t_1) - u_n(t_2)| \le \int_{t_1}^{t_2} |f(\tau, u_n(\tau))| \mathrm{d}\tau \le M|t_2 - t_1|\,.$$

Up to this point, we have only used the boundedness of f (which follows from the continuity of f). But for the next part, in which it will be shown that for any $t \in [t_0, t_0 + \rho_0]$ the sequence $\{u_n(t)\}$ is a Cauchy sequence in $\mathbb{R}^m$, we will need the Lipschitz continuity. Let L be the Lipschitz constant of f on S, then for any $n \in \mathbb{N}$ and $t \in [t_0, t_0 + \rho_0]$, we have

$$\begin{aligned} |u_{n+1}(t) - u_n(t)| &\le \int_{t_0}^{t} |f(\tau, u_n(\tau)) - f(\tau, u_{n-1}(\tau))| \mathrm{d}\tau \\ &\le \int_{t_0}^{t} L\,|u_n(\tau) - u_{n-1}(\tau)| \mathrm{d}\tau \\ &\le L\,\|u_n - u_{n-1}\|_\infty \int_{t_0}^{t} \mathrm{d}\tau \\ &= L\,\|u_n - u_{n-1}\|_\infty\,(t - t_0)\,, \end{aligned}$$

where $\|u_n - u_{n-1}\|_\infty = \sup\{|u_n(\tau) - u_{n-1}(\tau)| \mid t_0 \le \tau \le t_0 + \rho\}$. This implies that

$$\begin{aligned} |u_{n+2}(t) - u_{n+1}(t)| &\le \int_{t_0}^{t} L\,|u_{n+1}(\tau) - u_n(\tau)| \mathrm{d}\tau \\ &\le L^2\,\|u_n - u_{n-1}\|_\infty \int_{t_0}^{t} (\tau - t_0) \mathrm{d}\tau \\ &= \frac{L^2}{2}\,\|u_n - u_{n-1}\|_\infty\,(t - t_0)^2\,. \end{aligned}$$

Repeating this process, it follows for any $k \in \mathbb{N}$ that

$$\begin{aligned} |u_{n+k+1}(t) - u_{n+k}(t)| &\le \frac{L^k}{k!}\,\|u_n - u_{n-1}\|_\infty\,(t - t_0)^k \\ &\le \frac{L^k \rho^k}{k!}\,\|u_n - u_{n-1}\|_\infty\,. \end{aligned}$$

By using the triangle inequality repeatedly this gives for any $n,\ k, \in \mathbb{N}$ and $t \in [t_0, t_0 + \rho_0]$,

$$\begin{aligned} |u_{n+k+1}(t) - u_n(t)| &\le \sum_{i=0}^{k} |u_{n+i+1}(t) - u_{n+i}(t)| \\ &\le \sum_{i=0}^{k} \frac{L^{n+i}\rho^{n+i}}{(n+i)!}\,\|u_1 - u_0\|_\infty \\ &\le \frac{(L\rho)^n}{n!}\,\|u_1 - u_0\|_\infty \sum_{i=0}^{k} \frac{L^i \rho^i}{i!} \\ &\le \frac{(L\rho)^n}{n!}\,b\,\mathrm{e}^{L\rho}\,. \end{aligned}$$

Thus, for any $t \in [t_0, t_0 + \rho_0]$, the sequence $\{u_n(t)\}$ is a Cauchy sequence in $\mathbb{R}^m$; hence, the sequence has a limit, which will be denoted by $u(t)$. In other words, the sequence of functions $\{u_n\}$ is pointwise convergent in $I_\rho(t_0)$. We have already seen that the functions u_n are equicontinuous; hence, the convergence is uniform and the limit function $u(t)$ is equicontinuous [23].

To see that u satisfies the integral equation (2), observe that for any $t \in [t_0, t_0 + \rho_0]$

$$\begin{aligned} &\left| u(t) - u_0 - \int_{t_0}^{t} f(\tau, u(\tau)) \mathrm{d}\tau \right| \\ &\quad \le |u(t) - u_{n+1}(t)| \\ &\qquad + \left| \int_{t_0}^{t} |f(\tau, u_n(\tau)) - f(\tau, u(\tau))| \mathrm{d}\tau \right| \\ &\quad \le |u(t) - u_{n+1}(t)| + L\|u - u_n\|_\infty (t - t_0) \\ &\quad \le (1 + L\rho)\,(\|u - u_{n+1}\|_\infty + \|u - u_n\|_\infty) \end{aligned}$$

for any $n \in \mathbb{N}$. As the sequence $\{u_n\}$ is uniformly convergent, this implies that $u(t)$ satisfies the integral equation (2).

Finally the uniqueness will be proved using techniques similar to those in "Existence." Assume that there are two solutions u and v of the integral equation (2). Hence, for any $t_0 \le t \le t_0 + \rho$,

$$\begin{aligned} |u(t) - v(t)| &\le \int_{t_0}^{t} |f(\tau, u(\tau)) - f(\tau, v(\tau))| \mathrm{d}\tau \\ &\le L\|u - v\|_\infty (t - t_0)\,. \end{aligned}$$

Thus, for $k = 1$, it holds that

$$|u(t) - v(t)| \le \frac{L^k \|u - v\|_\infty (t - t_0)^k}{k!}\,, \quad \text{for any } t_0 \le t \le t_0 + \rho\,. \tag{10}$$

For the induction step, assume that (10) holds for some $k \in \mathbb{N}$. Then for any $t_0 \le t \le t_0 + \rho$,

$$|u(t) - v(t)| \le L \int_{t_0}^{t} \frac{L^k \|u - v\|_\infty (\tau - t_0)^k}{k!} \mathrm{d}\tau = \frac{L^{k+1} \|u - v\|_\infty (t - t_0)^{k+1}}{(k+1)!} .$$

By using the principle of induction, it follows that (10) holds for any $k \in \mathbb{N}$. This implies that for any $t_0 \le t \le t_0 + \rho$, the solutions satisfy $|u(t) - v(t)| \le \frac{(L\rho)^{k+1} \|u-v\|_\infty}{(k+1)!}$ for any $k \in \mathbb{N}$. Since the expression on the right converges to 0 for $k \to \infty$, this implies that $u(t) = v(t)$ for all $t \in [t_0, t_0 + \rho]$ and hence the solution is unique. □

Implicitly, the proof shows that the integral equation gives rise to a contraction on the space of continuous functions with the supremum norm and this could also have been used to show existence and uniqueness; see [11,25], where Schauder's fixed-point theorem is used to prove the Picard–Lindelöf theorem. With Schauder's fixed-point theorem, existence and uniqueness can be shown for much more general differential equations than ordinary differential equations.

The iteration process in (8) does not necessarily have to start with the initial condition. It can start with any continuously differentiable function $u_0(t)$ and the interation process will still converge. In specific problems, there are often obvious choices for the initial function u_0 which give a faster convergence or are more efficient, for instance, when proving unboundedness.

Techniques similar to those in the proof can be used to derive a bound on the difference between the successive approximations and the solution, showing that

$$|u_n(t) - u(t)| \le \frac{ML^n (t - t_0)^n}{(n+1)!} , \quad t \in [t_0, t_0 + \rho] ;$$

see [13]. This illustrates that the convergence is fast for t near t_0, but might be slow further away: see Fig. 3 for an example.

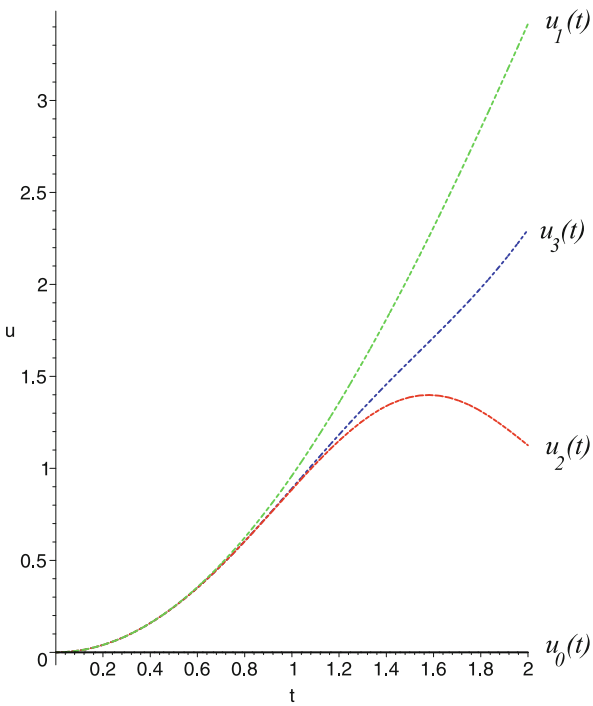

Existence and Uniqueness of Solutions of Initial Value Problems, Figure 3
The first four iterations for initial value problem with the vector field $f(t, u) = t \cos u + \sin t$ and the initial value $u(0) = 0$. Note that the convergence in the interval [0, 1] is very fast (two steps seem sufficient for graphical purposes), but near $t = 2$, the convergence has not yet been reached in three steps

Continuous Dependence on Initial Conditions

In this section, the vector field f is uniformly Lipschitz continuous on D with respect to its second variable u and its Lipschitz constant is denoted by L. By combining the Picard–Lindelöf theorem (Theorem 4) and the extension theorem (Theorem 3), it follows that for every $(t_0, u_0) \in D$, there exists a unique solution of (1) passing through (t_0, u_0) with a maximal interval of existence $(t_-(t_0, u_0), t_+(t_0, u_0))$. The *trajectory* through (t_0, u_0) is the set of points $(t, u(t))$, where $u(t)$ solves (1) and $t_-(t_0, u_0) < t < t_+(t_0, u_0)$. Now define the set $E \subset \mathbb{R}^{n+2}$ as

$$E := \{ (t, t_0, u_0) \mid t_-(t_0, u_0) < t < t_+(t_0, u_0), \quad (t_0, u_0) \in D \} .$$

The *flow map* $\Phi : E \to \mathbb{R}^m$ is a mapping which describes the solution of the initial value problem with the initial condition varying through D, i. e.,

$$\frac{\mathrm{d}}{\mathrm{d}t} \Phi(t, t_0, u_0) = f(t, \Phi(t, t_0, u_0)) \quad \text{and}$$
$$\Phi(t_0, t_0, u_0) = u_0 , \quad \text{for } (t, t_0, u_0) \in E .$$

From the integral equation (2), it follows that the flow map satisfies

$$\Phi(t, t_0, u_0) = u_0 + \int_{t_0}^{t} f(\tau, \Phi(\tau, t_0, u_0)) \mathrm{d}\tau , \quad \text{for } (t, t_0, u_0) \in E .$$

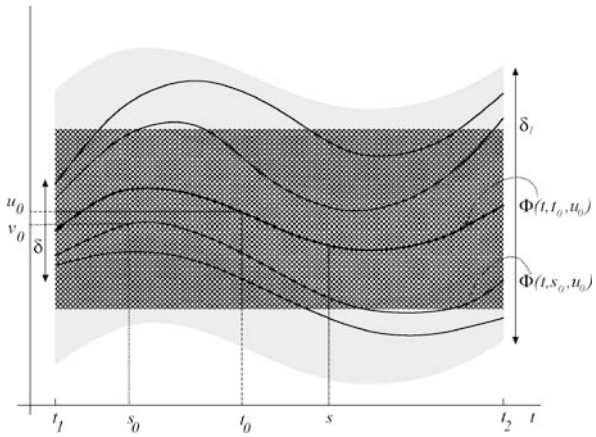

Existence and Uniqueness of Solutions of Initial Value Problems, Figure 4
The flow map $\Phi(t, t_0, u_0)$. Note the group property $\Phi(t, s, \Phi(s, t_0, u_0)) = \Phi(t, t_0, u_0)$. The *lightly shaded area* is the tube U_{δ_1} and the *darker shaded area* is the tube U_δ as used in the proof of Theorem 6. If a solution starts from within U_δ, then it will stay within the tube U_{δ_1} for $t_1 \le t \le t_2$, as the solution $\Phi(t, s_0, v_0)$ demonstrates

Furthermore, the uniqueness implies for the flow map that starting at the point (t_0, u_0), flowing to $(t, \Phi(t, t_0, u_0))$ is the same as starting at the point (t_0, u_0) and flowing to the intermediate point $(s, \Phi(s, t_0, u_0))$ and continuing to the point $(t, \Phi(t, t_0, u_0))$. This implies that the flow map has a group structure: for any $(t_0, u_0) \in D$ and any $s, t \in (t_-(t_0, u_0), t_+(t_0, u_0))$, the flow map satisfies $\Phi(t, s, \Phi(s, t_0, u_0)) = \Phi(t, t_0, u_0)$; the identity map is given by $\Phi(t_0, t_0, u_0) = u_0$; and by combining the last two results, it follows that $\Phi(t_0, t, \Phi(t, t_0, u_0)) = \Phi(t_0, t_0, u_0) = u_0$, and hence there is an inverse. See also Fig. 4 for a sketch of the flow map for $D \subset \mathbb{R} \times \mathbb{R}$.

To show that the flow map is continuous in all its variables, Gronwall's lemma [9] will be very useful. There are many versions of Gronwall's lemma; the one presented here follows [1].

Lemma 5 (Gronwall's Lemma) *Let $\psi\colon [a, b] \to \mathbb{R}$ and $\phi\colon [a, b] \to \mathbb{R}$ be nonnegative continuous functions on an interval $[a, b]$. Let $K \ge 0$ be some nonnegative constant such that*

$$\phi(t) \le K + \int_a^t \psi(\tau)\phi(\tau)d\tau\,, \quad \textit{for all } a \le t \le b\,.$$

Then

$$\phi(t) \le K \exp\left[\int_a^t \psi(\tau)d\tau\right], \quad \textit{for all } a \le t \le b\,.$$

Proof Define $F(t) = K + \int_a^t \psi(\tau)\phi(\tau)\mathrm{d}\tau$, for $a \le t \le b$. Then the assumption in the lemma gives $\phi(t) \le F(t)$ for $a \le t \le b$. Furthermore, F is differentiable, with $\dot{F} = \phi\psi$; hence, for $a \le t \le b$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(F(t)\exp\left[-\int_a^t \psi(\tau)\mathrm{d}\tau\right]\right) &= \psi(t)\phi(t)\exp\left[-\int_a^t \psi(\tau)\mathrm{d}\tau\right] \\ &\quad - F(t)\psi(t)\exp\left[-\int_a^t \psi(\tau)\mathrm{d}\tau\right] \le 0\end{aligned}$$

as $\phi(t) \le F(t)$. Integrating the left-hand side gives

$$F(t)\exp\left[-\int_a^t \psi(\tau)\mathrm{d}\tau\right] \le F(a) = K\,,$$

which implies for ϕ that $\phi(t) \le F(t) \le K\exp\left[\int_a^t \psi(\tau)\mathrm{d}\tau\right]$. □

Gronwall's lemma will be used to show that small variations in the initial conditions give locally small variations in the solution, or

Theorem 6 *The flow map $\Phi\colon E \to \mathbb{R}^{n+2}$ is continuous in E.*

Proof Let $(t, t_0, u_0) \in E$. As $t, t_0 \in (t_-(t_0, u_0), t_+(t_0, u_0))$, there is some $t_1 < t_2$ such that $t, t_0 \in (t_1, t_2)$, $[t_1, t_2] \subset (t_-(t_0, u_0), t_+(t_0, u_0))$ and the solution $\Phi(s, t_0, u_0)$ exists for any $s \in [t_1, t_2]$. First we will show that there is some tube around the solution curve $\{(s, \Phi(s, t_0, u_0)) \mid t_1 \le s \le t_2\}$ such that for any (s_0, v_0) in this tube, the solution $\Phi(s, s_0, v_0)$ exists for $t_1 \le s \le t_2$.

As D is open, there is some $\delta_1 > 0$ such that the closed tube $U_{\delta_1} := \{(s, v) \mid |\Phi(s, t_0, u_0) - v| \le \delta_1, t_1 \le s \le t_2\} \subset D$. As f is Lipschitz continuous, hence continuous, there is some $M > 0$ such that $|f(s, v)| \le M$ for all $(s, v) \in U_{\delta_1}$. Recall that the Lipschitz constant of f on D is denoted by L. Now define $\delta := \delta_1\,\mathrm{e}^{-L(t_2 - t_1)} < \delta_1$ and the open tube

$$U_\delta = \{(s, v) \mid |\Phi(s, t_0, u_0) - v| < \delta,\ t_1 < s < t_2\}\,,$$

thus, $U_\delta \subset U_{\delta_1} \subset D$. Thus, for every $(s_0, v_0) \in U_\delta$, the solution $\Phi(s, s_0, v_0)$ exists for s in some closed interval in $[t_1, t_2]$ around s_0 and the solution in this interval satisfies

$$\Phi(s, s_0, v_0) = v_0 + \int_{s_0}^s f(\tau, \Phi(\tau, s_0, v_0))\mathrm{d}\tau\,,$$

see also Fig. 4. Furthermore, for any $s \in [t_1, t_2]$, we have

$$\begin{aligned}
&\Phi(s, t_0, u_0) \\
&\quad = \Phi(s, s_0, \Phi(s_0, t_0, u_0)) \\
&\quad = \Phi(s_0, t_0, u_0) + \int_{s_0}^{s} f(\tau, \Phi(\tau, s_0, \Phi(s_0, t_0, u_0)))\mathrm{d}\tau \\
&\quad = \Phi(s_0, t_0, u_0) + \int_{s_0}^{s} f(\tau, \Phi(\tau, t_0, u_0))\mathrm{d}\tau \,.
\end{aligned}$$

Subtracting these expressions gives

$$\begin{aligned}
&|\Phi(s, s_0, v_0) - \Phi(s, t_0, u_0)| \\
&\quad \le |v_0 - \Phi(s_0, t_0, u_0)| \\
&\qquad + \int_{s_0}^{s} |f(\tau, \Phi(\tau, s_0, v_0)) - f(\tau, \Phi(\tau, t_0, u_0))|\,\mathrm{d}\tau \\
&\quad \le \delta + \int_{s_0}^{s} L\,|\Phi(\tau, s_0, v_0) - \Phi(\tau, t_0, u_0)|\mathrm{d}\tau \,,
\end{aligned}$$

so Gronwall's lemma implies that $|\Phi(s, s_0, v_0) - \Phi(s, t_0, u_0)| \le \delta \mathrm{e}^{L(s-s_0)} \le \delta_1$. Hence, for any s, $\Phi(s, s_0, v_0) \in U_{\delta_1} \subset D$, and hence this solution can be extended to its maximal interval of existence, which contains the interval $[t_1, t_2]$. So it can be concluded that for any $(s_0, v_0) \in U_\delta$, the solution $\Phi(s, s_0, v_0)$ exists for any $s \in [t_1, t_2]$.

Next we will show continuity of the flow map in its last two variables, i. e., in the initial value. For any $(s_0, v_0) \in U_\delta$ and $t_1 \le t \le t_2$, we have

$$\begin{aligned}
&|\Phi(t, s_0, v_0) - \Phi(t, t_0, u_0)| \\
&\quad \le |v_0 - u_0| \\
&\qquad + \left| \int_{s_0}^{t} f(\tau, \Phi(\tau, s_0, v_0))\mathrm{d}\tau - \int_{t_0}^{t} f(\tau, \Phi(\tau, t_0, u_0))\mathrm{d}\tau \right| \\
&\quad \le |v_0 - u_0| + \left| \int_{s_0}^{t_0} |f(\tau, \Phi(\tau, s_0, v_0))|\mathrm{d}\tau \right| \\
&\qquad + \left| \int_{t_0}^{t} |f(\tau, \Phi(\tau, t_0, u_0)) - f(\tau, \Phi(\tau, s_0, v_0))|\mathrm{d}\tau \right| \\
&\quad \le |v_0 - u_0| + M|t_0 - s_0| \\
&\qquad + \left| \int_{t_0}^{t} L|\Phi(\tau, t_0, u_0) - \Phi(\tau, s_0, v_0)|\mathrm{d}\tau \right| \,.
\end{aligned}$$

Thus, Gronwall's lemma implies that

$$\begin{aligned}
&|\Phi(t, s_0, v_0) - \Phi(t, t_0, u_0)| \\
&\quad \le (|v_0 - u_0| + M|t_0 - s_0|)\,\mathrm{e}^{L|t-t_0|} \\
&\quad \le (|v_0 - u_0| + M|t_0 - s_0|)\,\mathrm{e}^{L|t_2-t_1|} \,,
\end{aligned}$$

and it follows that Φ is continuous in its last two arguments. The continuity of Φ in its first argument follows immediately from the fact that the solution of the initial value problem is continuous. □

If the vector field f is smooth, then the flow map Φ is smooth as well.

Theorem 7 *Let $f \in C^1(D, R^m)$. Then the flow map $\Phi \in C^1(E, R^m)$ and*

$$\begin{aligned}
&\det(D_{u_0}\Phi(t, t_0, u_0)) \\
&\quad = \exp\left(\int_{t_0}^{t} \mathrm{tr}\left(D_u f(\tau, \Phi(\tau, t_0, u_0))\right) d\tau \right)
\end{aligned}$$

for any $(t, t_0, u_0) \in E$.

In this theorem, $\mathrm{tr}(D_u f(\tau, \Phi(\tau, t_0, u_0)))$ stands for the trace of the matrix $D_u f(\tau, \Phi(\tau, t_0, u_0))$. For second-order linear systems, this identity is known as Abel's identity and $\det(D_{u_0}\Phi(t, t_0, u_0))$ is the Wronskian.

The proof of Theorem 7 uses the fact that $D_{u_0}\Phi(t, t_0, u_0)$ satisfies the linear differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t} D_{u_0}\Phi(t, t_0, u_0) = D_u f(t, \Phi(t, t_0, u_0))\, D_{u_0}\Phi(t, t_0, u_0),$$

with the initial condition $D_{u_0}\Phi(t_0, t_0, u_0)) = I$. This last fact follows immediately from $\Phi(t_0, t_0, u_0) = u_0$. And the linear differential equation follows by differentiating the differential equation for $\Phi(t, t_0, u_0)$ with respect to u_0. The full details of the proof can be found, for example, in [5,11].

Extended Concept of Differential Equation

Until now, we have looked for continuously differentiable functions $u(t)$, which satisfy the initial value problem (1). But the initial value problem can be phrased for less smooth functions as well. For example, one can define a solution as an absolute continuous function $u(t)$ which satisfies (1). A function $u\colon I \to \mathbb{R}^m$ is *absolutely continuous* on I if for every positive number ε, no matter how small, there is a positive number δ small enough so that whenever a sequence of pairwise disjoint subintervals $[s_k, t_k]$ of I, $k = 1, 2, \ldots, N$ satisfies $\sum_{k=1}^{N} |t_k - s_k| < \delta$ then

$$\sum_{k=1}^{n} d\,(u(t_k), u(s_k)) < \varepsilon \,.$$

An absolute continuous function has a derivative almost everywhere and is uniformly continuous, thus continuous [24]. Thus, the initial value problem for an absolute continuous function can be stated almost everywhere.

For the existence of absolute continuous solutions, the continuity condition for the vector field in the Cauchy–

Peano theorem (Theorem 1) is replaced by the so-called Carathéodory conditions on the vector field. A function $f\colon D \to \mathbb{R}^m$ satisfies the Carathéodory conditions if [2]

- $f(t,u)$ is Lebesgue measurable in t for each fixed u;
- $f(t,u)$ is continuous in u for each fixed t;
- for each compact set $A \subset D$, there is a measurable function $m_A(t)$ such that $|f(t,u)| \le m_A(t)$ for any $(t,u) \in A$.

Similar theorems to the ones in the previous sections about existence and uniqueness can be stated in this case.

Theorem 8 (Carathéodory) *If D is an open set in $\mathbb{R}^{n+1}$ and $f\colon D \to \mathbb{R}^m$ satisfies the Carathéodory conditions on D, then for every $(t_0,u_0) \in D$ there is a solution of the initial value problem (1) through (t_0,u_0).*

Theorem 9 *If D is an open set in $\mathbb{R}^{n+1}$, $f\colon D \to \mathbb{R}^m$ satisfies the Carathéodory conditions on D and $u(t)$ satisfies the initial value problem (1) on some interval, then there exists a maximal open interval of existence. Furthermore, if (t_-,t_+) denotes the maximal interval of existence, then the solution $u(t)$ tends to the boundary of D if $t \downarrow t_-$ or $t \uparrow t_+$.*

Theorem 10 *If D is an open set in $\mathbb{R}^{n+1}$, $f\colon D \to \mathbb{R}^m$ satisfies the Carathéodory conditions on D and for every compact set $U \subset D$ there is an integrable function $L_U(t)$ such that*

$$|f(t,u)-f(t,v)| \le L_U\,|u-v|, \quad \text{for all } (t,u),\ (t,v) \in U,$$

then for every $(t_0,u_0) \in D$, there is a unique solution $\Phi(t,t_0,u_0)$ of the initial value problem (1). The domain E of the flow map Φ is open and Φ is continuous in E.

Proofs of those theorems can be found, for example, in Sect. I.5 in [11].

Many other generalizations are possible, for example, to vector fields which are discontinuous in u. Such vector fields play an important role in control theory. More details can be found in [6,7,17].

Further Directions

As follows from the previous sections, the theory of initial value problems for ordinary differential equations is quite well understood, at least for continuous vector fields. For more general differential equations, the situation is quite different. Consider for example a retarded differential equation, hence a differential equation with delay effects [12]. For retarded differential equations, there are local existence and uniqueness theorems which are quite similar to the ones for ordinary differential equations as presented in the earlier sections and can be proved by using Schauder's fixed-point theorem. But not all solutions can be extended as in the extension theorem (Theorem 3); see Chap. 2 in [12].

If one considers differential equations with more variables, i. e., partial differential equations, then there are no straightforward general existence and uniqueness theorems. For partial differential equations, the existence and uniqueness depends very much on the details of the partial differential equation. One possible theorem is the Cauchy–Kowaleskaya theorem, which applies to partial differential equations of the form $\partial_t^m u = F(t,u,\partial_x u,\partial_t u,\partial_{tx}^2 u,\dots)$, where the total number of derivatives of u in F should be less than or equal to m [22]. But there are still many open questions as well. A famous open question is the existence and uniqueness of solutions of the three-dimensional Navier–Stokes equations, a system of equations which describes fluid motion. Existence and uniqueness is known for a fluid in a two-dimensional domain, but not in a three-dimensional domain. The question is one of the seven "Millennium Problems," stated as prize problems at the beginning of the third millennium by the Clay Mathematics Institute [26].

Existence and uniqueness results are also used in dynamical systems. An area of dynamical systems is bifurcation theory and for bifurcation theory the smooth dependence on parameters is crucial. The following theorem gives sufficient conditions for the smooth dependence parameters; see Theorem 3.3 in Chap. 1 of [11].

Theorem 11 *If the vector field depends on parameters $\mu \in \mathbb{R}^k$, i. e., $f\colon D \times \mathbb{R}^k \to \mathbb{R}^m$ and $f \in C^1(D \times \mathbb{R}^k, \mathbb{R}^m)$, then the flow map is continuously differentiable with respect to its parameters μ. Furthermore, $D_\mu \Phi$ satisfies an inhomogeneous linear equation*

$$\begin{aligned}\frac{d}{dt} D_\mu \Phi(t,t_0,u_0,\mu) &= D_u f(t,\Phi(t,t_0,u_0,\mu),\mu)\, D_\mu \Phi(t,t_0,u_0,\mu) \\ &\quad + D_\mu f(t,\Phi(t,t_0,u_0,\mu),\mu)\end{aligned}$$

with initial condition $D_\mu \Phi(t_0,t_0,u_0,\mu) = 0$.

Results and overviews of bifurcation theory can be found, for example, in [4,8,10,14,16].

Another area of dynamical systems is stability theory. Roughly speaking, a solution is called stable if other solutions which start near this solution stay near it for all time. Note that the local continuity result in Theorem 6 is not a stability result as it only states that nearby solutions will stay nearby for a short time. For stability results, long-time existence of nearby solutions is a crucial property [10,14].

Bibliography

Primary Literature

1. Bellman R (1943) The stability of solutions of linear differential equations. Duke Math J 10:643–647
2. Carathéodory C (1918) Vorlesungen über Reelle Funktionen. Teubner, Leipzig (reprinted: (1948) Chelsea Publishing Company, New York)
3. Cauchy AL (1888) Oevres complètes (1) 6. Gauthiers-Villars, Paris
4. Chow S-N, Hale JK (1982) Methods of bifurcation theory. Springer, New York
5. Coddington EA, Levinson N (1955) Theory of Ordinary Differential Equations. McGraw-Hill, New York
6. Filippov AF (1988) Differential equations with discontinuous righthand sides. Kluwer, Dordrecht
7. Flugge-Lotz I (1953) Discontinuous automatic control. Princeton University Press, Princeton
8. Golubitsky M, Stewart I, Schaeffer DG (1985–1988) Singularities and groups in bifurcation theory, vol 1 and 2. Springer, New York
9. Gronwall TH (1919) Note on the derivative with respect to a parameter of the solutions of a system of differential equations. Ann Math 20:292–296
10. Guckenheimer J, Holmes P (1983) Nonlinear oscillations, dynamical systems, and bifurcations of vector fields. Springer, New York
11. Hale JK (1969) Ordinary Differential Equations. Wiley, New York
12. Hale JK, Verduyn Lunel SM (1993) Introduction to Functional Differential Equations. Springer, New York
13. Hartman P (1973) Ordinary Differential Equations. Wiley, Baltimore
14. Iooss G, Joseph DD (1980) Elementary stability and bifurcation theory. Springer, New York
15. Kneser H (1923) Ueber die Lösungen eines Systems gewöhnlicher Differentialgleichungen das der Lipschitzschen Bedingung nicht genügt. S-B Preuss Akad Wiss Phys-Math Kl 171–174
16. Kuznetsov YA (1995) Elements of applied bifurcation analysis. Springer, New York
17. Lee B, Markus L (1967) Optimal Control Theory. Wiley, New York
18. Lindelöf ME (1894) Sur l'application de la méthode des approximations successives aux équations différentielles ordinaires du premier ordre. Comptes rendus hebdomadaires des séances de l'Académie des sciences 114:454–457
19. Müller M (1928) Beweis eines Satzes des Herrn H. Kneser über die Gesamtheit der Lösungen, die ein System gewöhnlicher Differentialgleichungen durch einen Punkt schickt. Math Zeit 28:349–355
20. Peano G (1890) Démonstration de l'integrabilité des équations differentielles ordinaires. Math Ann 37:182–228
21. Picard É (1890) Mémoire sur la théorie de équations aux dérivées partielles et la méthode des approximations successives. J Math, ser 4, 6:145–210
22. Rauch J (1991) Partial Differential Equations. Springer, New York
23. Rudin W (1976) Principles of Mathematical Analysis, 3rd edn. McGraw-Hill, New York
24. Rudin W (1987) Real and Complex Analysis, 3rd edn. McGraw-Hill, New York
25. Zeidler E (1995) Applied functional analysis. Vol 1 Applications to mathematical physics. Springer, New York
26. Clay Mathematics Institute (2000) The Millenium Problems: Navier Stokes equation. http://www.claymath.org/millennium/Navier-Stokes_Equations/

Books and Reviews

Arnold VI (1992) Ordinary Differential Equations. Springer, Berlin

Arrowsmith DK, Place CM (1990) An Introduction to Dynamical Systems. Cambridge University Press, Cambridge

Braun M (1993) Differential Equations and their Applications. Springer, Berlin

Brock WA, Malliaris AG (1989) Differential Equations, stability and chaos in Dynamic Economics. Elsevier, Amsterdam

Grimshaw R (1993) Nonlinear Ordinary Differential Equations. CRC Press, Boca Raton

Ince EL (1927) Ordinary differential equations. Longman, Green, New York

Jordan DW, Smith P (1987) Nonlinear Differential Equations. Oxford University Press, Oxford

Werner H, Arndt H (1986) Gewöhnliche Differentialgleichungen: eine Einführung in Theorie und Praxis. Springer, Berlin

Exobiology and Complexity

Eric J. Chaisson
Wright Center, Tufts University, Massachusetts, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Arrow of Time
Non-equilibrium Thermodynamics
Big-Bang Cosmology
Measuring Complexity
Complexity and Evolution, Broadly Considered
Conclusions and Future Directions
Bibliography

Glossary

Complexity A state of intricacy, complication, variety, or involvement, as in the interconnected parts of a system – a quality of having many interacting, different components.

Cosmic evolution A grand synthesis of the many varied changes in the assembly and composition of radiation, matter, and life throughout the history of the Universe.

Cosmology The study of the structure, evolution, and destiny of the Universe.

Energy The ability to do work or to cause change.

Energy rate density The amount of energy flowing through a system per unit time per unit mass.

Evolution Any process of growth and change with time, including an accumulation of historical information; in its broadest sense, both developmental and generational change.

Exobiology The study of the origin, evolution, and distribution of past and present life in the Universe; also known as astrobiology or bioastronomy.

Thermodynamics The study of the macroscopic changes in the energy of a system, for which temperature is a central property.

Definition of the Subject

Recent research, guided by theoretical searches for unification as much as by compilation of huge new databases, suggests that complex systems throughout Nature are localized, temporary islands of ordered structures within vastly larger, disordered environments beyond those systems. All such complex systems – including, for example, stars, life, and society – can be shown to obey quantitatively the principles of non-equilibrium thermodynamics, and all can be modeled in a common, integral manner by analyzing the energy passing through those systems. The concept of energy flow does seem to be as universal a process as anything yet found in Nature for the origin, maintenance, and evolution of ordered, complex systems. The optimization of such energy flows acts as an agent of evolution broadly considered, thereby affecting, and to some extent unifying, all of physical, biological, and cultural evolution.

More specifically, non-equilibrium thermodynamics, especially the energy flows resulting from contrasting temporal behaviors of matter and radiation energy densities, can generally explain the cosmic environments needed for the emergence of increasingly ordered structures over time. Furthermore, a necessary (though perhaps not sufficient) condition for the natural flow of energy, and hence for the growth of complexity, is the expansion of the Universe itself. Among all of Nature's diverse systems, energy – acquired, stored, and expressed – is a principal driver of the rising complexity among galaxies, stars, planets and life-forms throughout the cosmos. Neither new science nor appeals to non-science are required to appreciate the outstanding hierarchy of evolutionary change, from atoms to galaxies, from cells to society.

One way to approach the topic of exobiology and complexity – also known as astrobiology or bioastronomy – is to place it within the grand context of cosmic evolution. This interdisciplinary subject seeks to combine all the natural sciences into a unified whole, thereby effectively creating a new scientific worldview for the 21st century. Evolution, broadly considered, has indeed become a powerful unifying concept in all of science. Life itself, including complex life, seems to be a natural, but not necessarily inevitable, result of the way things complexify in an expanding Universe.

Introduction

Cosmic evolution is the study of the sum total of the many varied developmental and generational changes in the assembly and composition of radiation, matter, and life throughout all space and across all time. These are the physical, biological, and cultural changes that have produced, in turn and among many other complex systems, our Galaxy, our Sun, our Earth, and ourselves. The result is an inclusive evolutionary synthesis bridging a wide variety of scientific specialties – physics, astronomy, geology, chemistry, biology, and anthropology – a genuine scientific narrative of epic proportions extending from the beginning of time to the present, from big bang to humankind.

The general idea of evolution – change writ large – extends well beyond the subject of biology, granting it a powerful unifying potential in all of science. Unquestionably, change is widespread throughout all of Nature, much as the Greek philosopher Heraclitus asserted 25 centuries ago: "All flows … nothing stays". Yet questions remain: How realistic is the quest for interdisciplinary unification? Can we reconcile the observed constructiveness of cosmic evolution with the inherent destructiveness of thermodynamics? Specifically, how have the amazing examples of order all around us arisen from chaos – and how does all this fit into complexity science?

We especially seek to understand the origins of the many diverse structures spanning the Universe today, notably those characterized by the term "complexity" – a state of intricacy, complication, variety, or involvement, as in the interconnected parts of a system. (In this article, no definitional distinctions are made among the words "order", "form", and "complexity"; we address only a general understanding of an entire spectrum of structures often described by the intuitive usage of the term complexity.) Particularly intriguing is the increase of complexity over the course of time, indeed dramatically so (with some exceptions) within the past half-billion years since the Cambrian period on Earth. Perhaps some underlying principle, a general law, or an ongoing process does create, organize, and maintain all complex structures in the

Universe, enabling us to study Nature's many changes on uniform, common ground – in the same quantitative way and with the same mental tools, in other words "on the same page".

Both theory and experiment, as well as computer simulations, suggest affirmative answers to some of the above questions: Islands of ordered complexity – namely, open systems that are galaxies, stars, planets, and life-forms – are numerically more than balanced by great seas of increasing disorder elsewhere in the environments beyond those systems. All quantitatively agrees with the valued principles of thermodynamics, especially non-equilibrium thermodynamics. Furthermore, energy flows produced largely by the expanding cosmos do seem to be as universal a factor in the origin of structured systems as anything yet found in Nature. The optimization of such energy flows might well act as the motor of evolution broadly conceived, thereby affecting all of physical, biological, and cultural evolution, the combination of which constitutes cosmic evolution.

Therefore, a general outline for this article is:
Cosmic Evolution

- Subsets in time: physical evolution → biological evolution → cultural evolution.

Other researchers have addressed life and complexity in a cosmic setting, originally Chambers [16], who anonymously wrote a pre-Darwinian study of wide interdisciplinary insight, and notably Shapley [62], who pioneered a "cosmography" that classified all known structures according to increasing dimensions. Among others, Spencer [64] championed the idea of growing complexity in biological and cultural evolution, Henderson [31] regarded the whole evolutionary process, both physical and organic, as one and the same, Whitehead [72] sought to broaden scientific thinking with his "organic philosophy", von Bertalanffy [75] championed a systems theoretic approach to physical, biological, and social studies, and Sagan [57], Reeves [56], Jantsch [33] and Chaisson [7] widely advanced the concept of complex (intelligent) life within a cosmological framework.

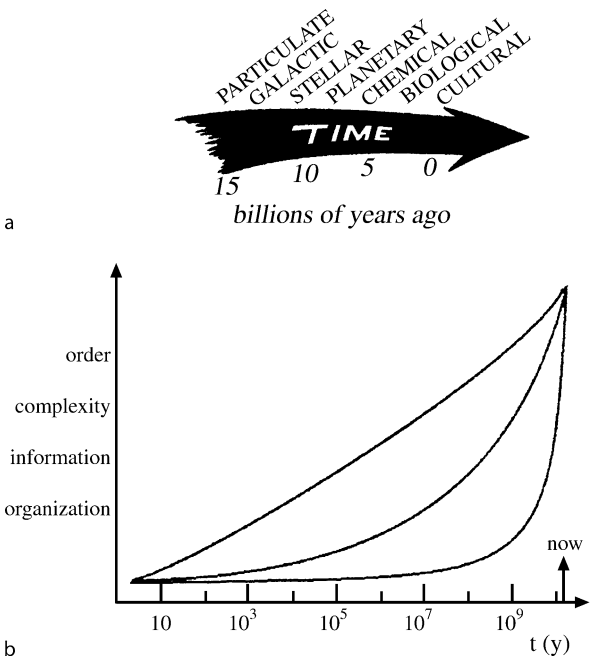

Exobiology and Complexity, Figure 1
a An arrow of time symbolically chronicles the principal epochs of cosmic history, from the beginning of the Universe ~14 billion years ago (at left) to the present (at right). Labeled diagonally across the top are 7 major evolutionary phases (corresponding to the main historical epochs) that have produced, in turn, increasing amounts of order and complexity among all material systems: particulate, galactic, stellar, planetary, chemical, biological, and cultural. Cosmic evolution encompasses all of these phases, each of which represents a coarse temporal duration when the emergence of key systems flourished in Nature. Time is assumed to flow linearly and irreversibly, much as other basic tenets are presumed, including the fixed character of physical law and the idea that 2 + 2 = 4 everywhere. b Sketched here qualitatively is the rise of order, form, and complexity typifying the evolution of *localized* material structures throughout the history of the Universe. This family of curves depicts a widespread, innate feeling, and not a rigid proof, that the complexity of ordered structures has *generally* increased over the course of time. It is unknown if this rise of complexity has been linear, exponential, or even faster (as drawn here for the 3 curves); current research aims to specify this curve and to describe it quantitatively

Arrow of Time

Figure 1a sketches Nature's main historical epochs diagonally atop the so-called arrow of time [3]; these 7 epochs correspond to the major evolutionary phases comprising the whole of cosmic evolution [14]. Regardless of its shape or orientation, such an arrow symbolizes a *sequence* of events that have changed systems from simplicity to complexity, from inorganic to organic, from chaos in the early Universe to order more recently. That sequence, as determined by a large body of post-Renaissance data, accords well with the idea that a chain of knowledge – a loose continuity along an impressive hierarchy of complexity – links, in turn:

- The evolution of primal energy into elementary particles and then atoms
- The evolution of those atoms into galaxies and stars
- The evolution of stars into heavy elements

- The evolution of those elements into the molecular building blocks of life
- The evolution of those molecules into life itself
- The evolution of advanced life forms into intelligence
- The evolution of intelligent life into a cultured and technological civilization.

Despite the extreme specialization of modern science, evolution marks no disciplinary boundaries; complexity science is a truly interdisciplinary topic. A more specific outline for this article is then:
Cosmic Evolution

- Subsets: physical evolution → biological evolution → cultural evolution
- Phases: particulate → galactic → stellar → planetary → chemical → biological → cultural.

Accordingly, the most familiar kind of evolution – biological evolution, or neo-Darwinism – is just one, albeit important, subset of a broader evolutionary scenario including much more than life on Earth. In short, what Darwinian change does for plants and animals, cosmic evolution aspires to do for all things. And if Darwinism created a revolution in understanding by helping to free us from the anthropocentric belief that humans differ from other life-forms on our planet, then cosmic evolution extends that intellectual revolution by treating matter on Earth and in our bodies no differently from that in the stars and galaxies far beyond.

Note that time's arrow does not imply that primitive, "lower" life-forms have biologically changed directly into advanced, "higher" organisms, any more than galaxies have physically changed into stars, or stars into planets. Rather, with time – much time – the environmental conditions suitable for spawning simple life eventually changed into those favoring the biological origin and evolution of more complex species. Likewise, in the earlier Universe, the physical evolution of environments ripe for galactic formation eventually gave way more recently to conditions conducive to stellar and planetary formation. Now, at least on Earth, cultural evolution dominates, since our local planetary environment has once more changed to foster greater, societal complexity. Change in the surrounding environments usually precedes change in organized systems, and the resulting changes for those systems selected to endure have *generally* been toward greater amounts of diverse order and complexity.

Anthropocentrism is neither intended nor implied by the arrow of time; the arrow is not pointing at humankind. Anthropic principles notwithstanding, no logic supports the idea that the Universe was conceived in order to produce specifically us. Humans are not the pinnacle or culmination of the cosmic-evolutionary scenario, nor are we likely to be the only technologically competent beings that have emerged in the organically rich Universe. The arrow merely provides a convenient symbol, artistically suggesting the building of increasingly complex structures, from spiral galaxies to rocky planets to thinking beings.

Figure 1b graphs the widespread impression that material systems have become more organized and complex, especially in relatively recent times. This family of curves graphs "islands" of complexity that comprise ordered systems per se – whether massive stars, colorful flowers, or busy urban centers – not their vastly, indeed increasingly disorganized surroundings. Modern science aims to explain this rise of complexity and to do so with accepted scientific principles and observational or experimental data.

Non-equilibrium Thermodynamics

Cosmic evolution, as understood today, is governed largely by the laws of physics, particularly those of thermodynamics. However, this does not mean classical reductionism, for here we seek to model change guided by a combination of randomness and determinism, of chance and necessity. Nor does the cosmic-evolutionary narrative employ mere equilibrium thermodynamics – the kind most often used to explain closed systems isolated from their surroundings and having maximum entropy states. All structures observed in Nature, among them most notably galaxies, stars, planets, and life-forms, are demonstrably open, non-equilibrium systems, with flows of energy in and out being an important feature. And it is this energy, often called available, or "free" energy that helps to build structures [27,28,38,54,60,74].

By utilizing energy, order can be achieved temporarily, or at least the environmental conditions made conducive for the potential rise of order within open systems ripe for growth. Energy flow plays a vital role in the creation, maintenance, and fate of complex systems – all quantitatively in accord with the second law of thermodynamics, which demands an overall rise in disorder. None of Nature's ordered structures, not even life itself, is a violation (or even a circumvention) of the second law. Considering both any ordered system as well as its disordered surroundings, non-equilibrium thermodynamics shows that the net entropy of the system and its environment always increases. (Quantitative details for many such systems can be found in [10].)

Energy is now recognized as a key ingredient, not only for biological systems such as plants and animals, but also

for physical systems such as stars and galaxies, indeed as well for social systems such as a city's inward flow of food and resources amidst its outward flow of products and wastes (Weber [69]; Dyke [21]). The analysis is much the same for all open systems, provided they are modeled in broad, interdisciplinary ways; energy flow seems indispensable for any system's origin and evolution.

Figure 2, adapted from the work of Prigogine et al. [55] and Salk [58], graphically diagrams the emergence of structure in the presence of energy flow. Physicists relate to the type of curves drawn in part (a); biologists are more familiar with those in part (b). By crossing certain energy thresholds that depend on a system's status, bifurcations can occur, fostering the origin of whole new structures that display surprising amounts of coherent behavior. Such "dissipative" structures can export some of their entropy (or expel some of their energy) into their external environments. Accordingly, order is created and sustained by routine consumption of substances rich in energy, followed by discharge of substances low in energy. This process, often misnamed, is not really *self*-ordering; it is ordering in the presence of energy. "Self-assembling" systems demonstrate an essential tension between energy inflow and dissipative outflow; such systems do no function magically by themselves.

The emergence of order from a condition where originally there was none (or less of it) is relatively straightforward [54,59]. Fluctuations – random deviations from some average, equilibrium value of, for example, density, temperature, or pressure – inevitably yet stochastically appear in any natural system having many degrees of freedom. Normally, as in equilibrium thermodynamics, such instabilities regress in time and disappear; they come and go by chance since the statistical fluctuations diffuse as quickly as they arise. Even in an isolated system, such internal fluctuations can generate local, microscopic reductions in entropy, but the second law ensures that they will always balance out. Microscopic temperature fluctuations, for instance, are said to be thermally relaxed, and entropy remains maximized in such systems. Nor can an open system *near* equilibrium change spontaneously to new and interesting structures. But should those fluctuations become too great for an open system to damp, that system can then depart far from equilibrium and have a chance to reorganize. Such restructuring generates a "dynamic steady state", provided the amplified fluctuations are continuously driven and stabilized by a flow of energy from the surroundings – namely, provided the energy flow rate exceeds the thermal relaxation rate. Systematic, coherent cycling is often the result, since under these conditions the spontaneous creation of macroscopic structures dissipates

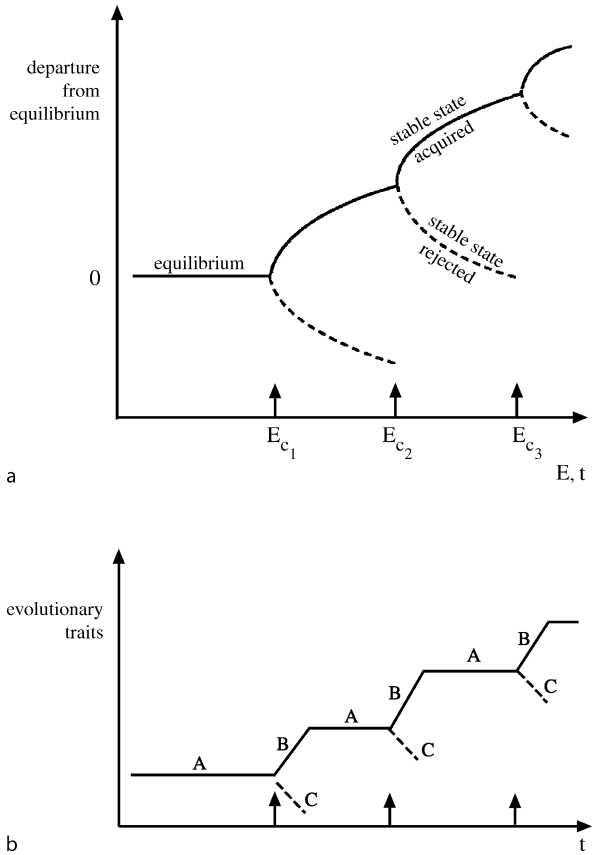

Exobiology and Complexity, Figure 2

a The departure of an open system from equilibrium is drawn here as a function of both time, *t*, and energy, E. The time axis makes clear that this is an historical, evolutionary process, whereas the parallel energy axis tracks free energy flowing through the open system as a vital part of that process. At certain critical energies, labeled here E_c, a system can spontaneously change, or bifurcate, into new, non-equilibrium, dynamic steady states. Statistical fluctuations – that's chance – affect which fork the system selects – that's necessity – upon bifurcation (*vertical arrows*), namely which spatial structure is achieved. Not all new systems survive (*solid curve*); some are rejected (*dashed curve*). The process, as always, is a mixture of randomness and determinism, therefore the end result is inherently unpredictable, as with all of evolution. b Events in evolutionary biology mimic those of the diagram in a, although the results here are richer in structural detail, system function, and energy flow. In phases marked A, a species survives and thus persists until the environment changes (*vertical arrows*), after which further evolution occurs – along phase B toward renewed survival or phase C toward extinction. Both upwardly rising graphs (drawn by *solid lines* for both parts of this figure) imply neither progress nor inevitability, but they do suggest a *general trend* toward increasing complexity with time – a trend that cannot be denied among organized systems observed throughout Nature

energy more rapidly than the ensuing, and damaging, heat can damp the gradients and destroy those structures. Furthermore, since each successive reordering often causes more complexity than the preceding one, such systems become even more susceptible to fluctuations. Complexity itself consequently creates the conditions for greater instability, which in turn provides an opportunity for greater reordering. Nothing is guaranteed; thermodynamics specifies what can happen, not what actually does happen. The resulting phenomenon – termed "order through fluctuations" – is a distinctly evolutionary one, complete with feedback loops that help drive some systems further from equilibrium. And as the energy consumption and resulting complexity accelerate, so does the evolutionary process. This is the realm of true thermo*dynamics*, the older, established subject of that name more properly labeled "thermostatics".

Numerous examples abound throughout Nature, and not just among physical systems, but for biological and social systems as well. Naturally occurring phenomena such as convection cells, river eddies, atmospheric storms, and even artificially made devices such as kitchen refrigerators and coherent lasers among an array of many physical systems that experience coherent order when amply fed with sufficient energy, all display enhanced order when energy flows exceed some threshold. Biological systems also obey the rules of non-equilibrium thermodynamics, for we and our living relatives are demonstrable examples of dynamic steady states that emerge and function via energetically enhanced neo-Darwinism (though biologists often worry that such statements aim to reduce biology to physics – whereas in reality physics is broadened to include biology.) As are Lamarckian-type cultural systems of more recent times also dynamic steady states, for among the bricks and chips that civilization has built, energy has been a principal driver (although, again, sociologists and anthropologists often loathe their subjects being treated thermodynamically). The result is that life and its cultural inventions differ not in kind, but merely in degree – specifically, degree of complexity – among numerous ordered systems evident in Nature.

Big-Bang Cosmology

The origin of Nature's many complex structures depends on the flow of free energy. And this, like the arrow of time itself, is a direct consequence of the expanding Universe – a much tested "standard cosmological model" based largely on three-fold observations of distant receding galaxies, microwave background radiation, and light-element abundances. Time marches on and free energies flow because the cosmos dynamically evolves – building, maintaining, and often destroying systems. Indeed, it is cosmic expansion, and probably nothing more, that has caused the entire Universe to depart from its initial state of thermodynamic equilibrium. (Thus, the free energies are inevitable, not the resulting systems per se – which is why it's called "available", or potential, energy freely *capable* of doing work.) The stark contrast between localized hot stars and the vast, cold interstellar space surrounding them now guarantees a state of non-equilibrium, a cosmic condition that has pertained for billions of years [26,40].

Matter

Although modern cosmology stipulates that matter only later emerged from the primordial radiation of the early Universe, it is pedagogically useful to quantify first the role of matter and thereafter the primacy of radiation. In this way, perhaps the greatest change in the history of the Universe – the transformation from radiation to matter – can be mathematically justified.

Imagine an arbitrary shell of mass, m, and radius, r, expanding isotropically with the Universe at a velocity, v, from some central point. The sphere within the shell is not necessarily meant to represent the entire Universe, only an extremely large gas cloud within it – in fact, larger than the extent of a typical galaxy supercluster (∼50 Mpc across), which comprises the highest rank in the known hierarchy of matter assemblages in the Universe. Invoking the principle of energy conservation, we find the Friedmann–Lemaitre equation that describes a family of models for the Universe in bulk,

$$H^2 - \frac{8}{3}\pi G\rho_{\mathrm{m}} = -kR^{-2} .$$

Here, H is Hubble's constant (a measure of galaxy recession in the expanding Universe), G is the universal gravitational constant, ρ_{m} is the matter density, and k is a time-dependent curvature constant. R is a scale factor which relates the radius, r, at any time, t, in cosmic history to the current radius, r_0, at the present time – namely, $r = Rr_0$. Solutions to the above equation specify three general models for the Universe:

- The Universe is "open" (i. e., k negative) and thus recedes forevermore
- The Universe is "closed" (i. e., k positive), meaning it eventually stops and thereafter contracts to a point much like that from which it began
- The Universe is precisely balanced between the open and closed models, in which case it eternally expands

and never contracts (because it can never reach infinity).

Even if the Universe is, as now suspected, accelerating in its outward expansion, the effect of "dark energy" (that supposedly causes the acceleration) on stars and galaxies is minimal, and that on smaller structures like planets and their organized living systems is inconsequential; cosmic acceleration likely affects only the dynamics of the Universe on the largest scales and not those of organized systems that are controlled by local energies within it.

The simplest case ($k = 0$, also known as the Einstein–deSitter solution) leads to the critical density for closure,

$$\rho_{\mathrm{m,c}} = 3H^2/8\pi G\ ,$$

which, when evaluated for G and H (~70 km/s/Mpc), equals $\sim 10^{-29}$ g/cm^3. This is ~6 atoms in each cubic meter of space, or about a million times thinner than in the region between Earth and the Moon. Whether the actual current density, on average everywhere, is smaller or larger than this value, making the Universe open or closed, respectively, is currently unknown; there is too much uncertainty concerning "dark (non-baryonic) matter" that is implied, but not found yet needed, to gravitationally bind galaxies and their clusters. However, the above-noted acceleration of the Universe does imply that it is expanding ever faster, thereby giving it an open geometry that will recede toward infinity forevermore.

To follow the evolution of matter throughout cosmic history (up to the present), we appeal to the conservation of material particles in the huge sphere postulated above, $\rho_{\mathrm{m}} = \rho_{\mathrm{m},0}R^{-3}$, substitute into the special ($k = 0$) case of the Friedmann–Lemaitre equation, and manipulate,

$$\int \mathrm{d}t = \left(\frac{8}{3}\pi G\rho_{\mathrm{m,c}}\right)^{-0.5} \int R^{0.5}\mathrm{d}R\ .$$

The result suggests that the Universe is $\sim 14 \times 10^9$ y old (±~10%). This equation additionally stipulates how the average matter density thins with time,

$$\rho_{\mathrm{m}} \approx 10^6 t^{-2}\ ,$$

where ρ_{m} is expressed in g/cm^3 and t in seconds.

Figure 3 plots this evolution of matter throughout all of universal history. This run of density, ρ_{m}, in standard, big-bang cosmology demonstrates the essence of change on the largest scales – the broadest view of the biggest picture. Here displayed in this one plot is the thermodynamic history of the whole Universe, so the curve for ρ_{m} in this figure (as well as the curve for T discussed in the next section) pertain to nothing in particular, just everything in general.

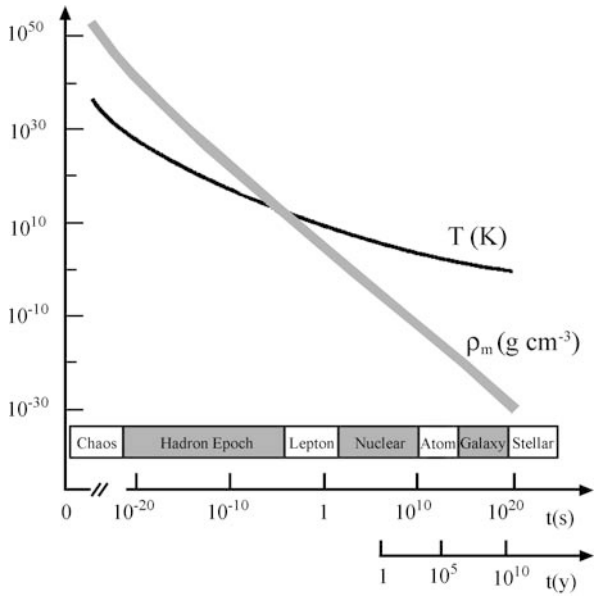

Exobiology and Complexity, Figure 3
Log-log plot of the density (ρ_{m}) of matter on average and the temperature (T) of radiation on average, over the course of all time, to date. The thick width of the density curve displays the range of uncertainty in total mass density, whose true value depends on the amount of "dark matter" in the Universe. By contrast, the cold cosmic background temperature is very accurately measured today (2.7 K), and its thin curve here is equally accurately extrapolated back into the hot, early Universe. Recent findings that cosmic expansion is accelerating should not much affect these curves

Radiation

The same analysis regarding matter can be applied to radiation in order to follow the change of temperature with time. Again, for the simplest $k = 0$ case,

$$H^2 = \frac{8\pi G\rho_{r,c}}{3R^4}\ ,$$

where ρ_r is the equivalent mass density of radiation. Here the R^4 term derives from the fact that radiation scales not only as the volume ($\alpha\ R^3$) but also by one additional factor of R because radiation (unlike matter) is also affected linearly by the Doppler shift. And noting that $\rho_r c^2 = \mathrm{a}T^4$, where a is the universal radiation constant for any blackbody emitter and T is the temperature of radiation, we find the temporal dependence of average temperature throughout all time (in seconds),

$$T \approx 10^{10} t^{-0.5}\ .$$

The universal radiation, having begun in a fiery expansion (popularly called the "big bang"), has now cooled to 2.7 K, the average value of the cosmic microwave back-

ground measured today by radio telescopes on the ground and satellites in orbit [2].

Figure 3 also plots this run of T versus t. Again, for emphasis, the two curves in this figure show the prime twin trends of big-bang cosmology: the cooling and thinning of radiation and matter, based largely on observations of the microwave background radiation and of the distant receding galaxies.

For the first few hundred millennia of the Universe, radiation reigned supreme over matter. Life was nonexistent and matter only a submicroscopic precipitate suspended in a glowing, chaotic fireball. All space was flooded with high-frequency photons, especially light, x-rays, and γ-rays, ensuring a non-structured, undifferentiated, (virtually) informationless, and highly uniform plasma. Matter and radiation were intimately coupled to each other – thermalized and equilibrated. Structure of any sort had yet to emerge; the energy density of radiation was too great. If single protons captured single electrons to form simple hydrogen atoms, the radiation was then so fierce as to destroy those atoms immediately. However, as the Universe expanded with time, the energy of radiation decreased faster than the energy equivalently contained in matter.

To confirm this statement, compare the energy densities of radiation and matter, and especially how these two quantities have change over time. First convert the matter density derived earlier to an equivalent energy density by invoking the Einsteinian mass (m)-energy (E) relation, $E = mc^2$ – that is, by multiplying the above equation for ρ_m by c^2. Now, $\sim 14 \times 10^9$ y after the big bang, $\rho_{m,0}c^2 \approx 10^{-9}$ erg/cm^3, whereas $aT_0^4 \approx 4 \times 10^{-13}$ erg/cm^3; thus currently, $\rho_{m,0}c^2 > aT_0^4$ by several orders of magnitude, proving that matter is now in firm control (gravitationally) of cosmic changes, despite the Universe still being flooded today with long-wavelength radiation. However, given that $\rho_m c^2$ scales as R^{-3} and aT^4 scales as R^{-4}, there must have been a time in the past when $\rho_m c^2 = aT^4$, and an even earlier time when $\rho_m c^2 < aT^4$. Manipulation of the above equations shows that these two energy densities crossed at $t \approx 10^4$ y, well less than a million years after the big bang. Figure 4 is a graphical representation of this paragraph.

This crossover represents a preeminent change in all of cosmic history. The event, $\rho_m c^2 = aT^4$, separates the *Radiation Era* from the *Matter Era*, and designates the time ($\sim 10^4$ y) when the Universe gradually began to become transparent. Thermal equilibrium was destroyed and symmetry broken, causing the radiative fireball and disorganized matter to decouple; it was as though a fog had lifted. Photons, previously scattered aimlessly and destructively

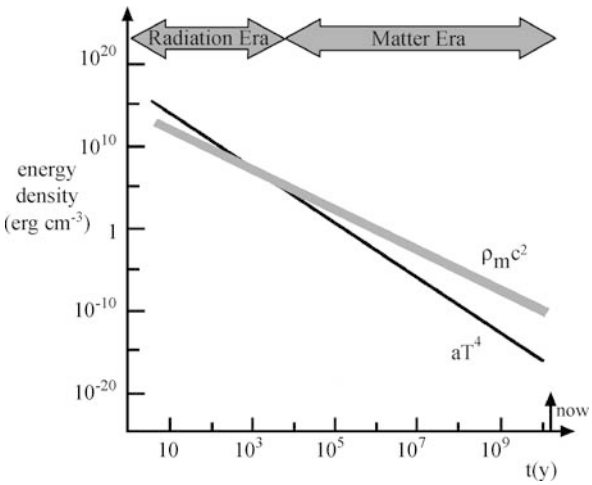

Exobiology and Complexity, Figure 4
The temporal behavior of both matter energy density ($\rho_m c^2$) and radiation energy density (aT^4) illustrates perhaps the greatest change in all of natural history. Where the two curves intersect, neutral atoms began to form; by $t \approx 10^5$ y after the big bang the Radiation Era had changed into the Matter Era. A uniform, featureless state describing the early Universe was thus naturally transformed into one in which order and complexity were thereafter possible

by subatomic material particles (especially free electrons) in the expanding, hot, opaque plasma of the Radiation Era, were no longer so affected once the electrons were bound into atoms of the Matter Era. This crucial and dramatic change was over by $\sim 4 \times 10^5$ y, when the last remnants of the early ionized plasma state had finally transformed into neutral matter. The 2.7-K microwave radiation reaching Earth today is a relic of this critical phase transition, having streamed unimpeded (except for being greatly redshifted, $z \sim 10^3$) across space and time for most of the age of the Universe, granting a "view" of this grandest of all evolutionary events that occurred long ago.

With the onset of the Matter Era, matter literally began dominating radiation. Natural history became more interesting, for then structures could begin to form. The results of inevitable change, induced gradients, energy flows, and evolved systems, over billions of years and minus the details, are galaxies, stars, planets, and life-forms, one byproduct of which is intelligence – at least on Earth. And this, in turn, has anthropogenically changed nearly everything on our planet.

Life

Now ~ 14 billion years after the beginning of space and time, the *Life Era* has begun, at least locally on Earth

(and possibly at many other places in the Universe). Here, the emergence of technologically intelligent life heralds a whole new era wherein life has gradually begun to dominate matter. This second of two great transformations was not caused by the origin of life per se several billion years ago; rather, it is technologically advanced life that differs significantly from primitive life and from other types of clustered inanimate matter scattered throughout the Universe. This is not an anthropocentric statement; we differ because we are the only species capable of knowing our past and worrying about our future, the only one able to control matter (albeit locally), much as matter evolved to control radiation long ago. Intelligent life on Earth is literally taking matter into its own hands – manipulating matter and energy, altering genes and terrestrial environments, indeed potentially changing evolution itself.

Some central questions before us are these: What caused the changes amid a wide spectrum of ordered structures throughout cosmic history and how has complexity increased with time? Have humans actually become the agents of change on Earth, able to tinker with both matter and energy, including now modifying genes and environments more than they affect us? How did the neural network within our human brains acquire the sophistication needed to fashion societies, weapons, cathedrals, philosophies, etc? In short, what caused us to become sentient enough to contemplate our complex selves?

Measuring Complexity

To appreciate the crux of the historical appearance of structured matter and life, we return to the greater cosmic environment and to some of the thermodynamic issues raised earlier. In brief, when the Universe broke its symmetry a few thousand centuries after the big bang, equilibrium was also destroyed. Temperature gradients were thereafter established naturally owing to the expansion of the cosmos. And that meant free energy began flowing, in fact increasingly so as the temperatures of matter and radiation diverged with time. These are the environmental conditions that are favorable for the potential growth of order, form, and complexity.

Cosmic Environment for the Growth of Complexity

When matter and radiation were still equilibrated in the Radiation Era, only a single temperature is needed to describe the thermal history of the Universe; the absence of any thermal gradients imply (virtually) zero information content, or zero macroscopic order, in the early Universe. However, once the Matter Era began, the gas-energy equilibrium was destroyed and a single temperature is insufficient to specify the bulk evolution of the cosmos. Since the random motions of the H and He atoms failed to keep pace with the rate of general expansion of the atoms away from one another [40], the matter cooled faster, $T_m \approx 6 \times 10^{16} t^{-1}$, than the radiation, $T_r \approx 10^{10} t^{-0.5}$. Figure 5a displays this thermal gradient, which has grown wider since $t \approx 10^5$ y.

Such a thermal gradient is the clear signature of a heat engine, and it is this ever-widening gradient that has enabled matter, in the main, to "build things" increasingly complex. Theoretically at least, the environmental conditions after 10^5 y naturally allowed a rise in "negentropy" [60] or "information content" [61] – both factors

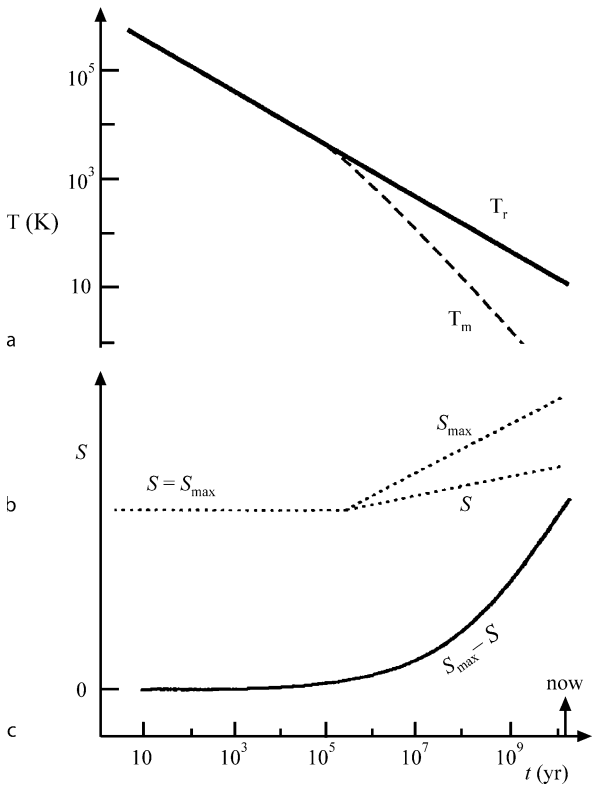

Exobiology and Complexity, Figure 5
a In the expanding Universe, the temperatures of matter and radiation separated once these quantities became fully decoupled at ~10^5 y. Since that time, the Universe has been in a non-equilibrium state. b *S* increases less rapidly than S_{max}, once the symmetry of equilibrium broke when matter and radiation decoupled at ~10^5 y. By contrast, in the early, equilibrated Universe, $S = S_{max}$ for the prevailing conditions. c The potential for the growth of order, $S_{max} - S$, has increased ever since the start of the Matter Era. This potential rise of order compares well with the subjectively drawn curves of Fig. 1b, thus providing a theoretical basis for the growth of system complexity

qualitatively synonymous with the term "complexity" [42]. But, as noted below, in practice both such terms are overly vague and subject to interpretation [5,73], so we resort here to a more conventional use of entropy, as agreed upon by most thermodynamicists. The important point – without getting lost in dubious semantics or contentious definitions – is that such non-equilibrium states are suitable, indeed apparently necessary, for the emergence of order, thus it can be reasoned that *cosmic expansion itself is the prime mover for the gradual construction of a hierarchy of structures throughout the Universe.*

Figure 5b plots the run of entropy, S, for a thermal gradient typical of a heat engine, but here graphed for the whole Universe. This is a not a mechanical device running with idealized Newtonian precision, but a global engine capable of potentially doing work as locally emerging systems interact with their environments – especially those systems able to take advantage of increasing flows of free energy resulting from cosmic expansion and its naturally growing gradients. Although thermal and chemical (but not gravitational) entropy must have been maximized in the early Universe, hence complexity in the form of any structures then nonexistent, after decoupling the environmental conditions became favorable for the potential growth of order, taken here to mean a "lack of disorder." At issue was timing: As ρ decreased, the equilibrium reaction rates ($\alpha\rho$) fell below the cosmic expansion rate ($\alpha\rho^{1/2}$) and non-equilibrium states froze in. Thus we have a paradoxical yet significant result that, in an expanding Universe, both the disorder (i. e. net entropy) and the order (maximum possible entropy minus actual entropy at any given time) can increase simultaneously – the former globally and the latter locally. All the more interesting when comparing the shape of this curve of potentially rising order, $S_{max} - S$ in Fig. 5c, with our earlier intuited sketch of rising complexity in Fig. 1b [12,25,41].

Free Energy Rate Density

Theory aside, have the many diverse real structures known to exist in the Universe displayed this sort of progressive increase in order during the course of time? The answer is generally yes. At issue again is how to best characterize complexity numerically, given the varied connotations that this term presents for many researchers [46,47]. In biology alone, much as their inability to reach consensus on a definition of life, biologists cannot agree on a complexity metric. Some count non-junk genome size [66], others employ structural morphology or behavioral flexibility [4], while still others chart numbers of cell types in organisms [36] or appeal to cellular specialization [49]. All these attributes of life have qualitative usefulness, yet all are hard to quantify in practical terms; nor do they apply to non-living things. If progress is to be made assessing a wide spectrum of complex systems in Nature, our analysis must extend beyond mere words, indeed beyond biology.

Putting aside as unhelpful (in the sense that it is too ambiguous and controversial) the above-noted concept of information content [32,34] as well as the concept of negative entropy (or negentropy, which Schroedinger [60] first adopted but then quickly abandoned), we return to the quantity with greatest appeal to physical intuition – energy. Given that energy – the ability to do work or to cause change – is the most universal currency known in the natural sciences, it might reasonably be expected to have a central role in any attempted unification of physical, biological, and cultural evolution.

Energy does act as an underlying, universal driver like no other in all of modern science. Whether living or non-living, dynamical systems need flows of energy to endure. If stars don't convert gravitational potential into heat and light, they would collapse; if plants don't photosynthesize sunlight, they would shrivel and decay; if humans don't eat, we too would die. Likewise, society's fuel is energy: Resources come in and wastes go out, all the while civilization goes about its daily business.

Not that energy has been ignored in previous studies of systems' origin and assembly. Physicists (e. g., Morrison [52] and Dyson [22]), biologists (Lotka [44]; Morowitz [51]; Fox [24]), and ecologists (Odum [53], Ulanowitz [68], and Smil [63]), to cite only a few researchers, have noted energy's organizational abilities. But the quantity of choice cannot be energy alone, for a star clearly has more energy than an amoeba, a galaxy much more than a single cell. Yet any living system is surely more complex than any inanimate object. Thus, absolute energies are not as suitable as normalized values, which depend on a system's size, composition, and efficiency. Nor are maximum energy principles or minimum entropy states [43] likely relevant; rather, organizational complexity is more likely governed by the *optimum* use of energy – not too little as to starve a system, yet not too much as to destroy it.

To characterize complexity objectively – that is, to normalize all such ordered systems on the same, level page – a kind of energy density is useful, much like the competing energy densities of radiation and matter that dictated changing events in the earlier Universe (Fig. 4). In fact, for a proper treatment of the thermodynamics of non-equilibrium open systems, it is the *rate* at which free energy flows through such systems of given mass that is most practical. Hence, *free energy rate density*, symbolized by Φ_m, is an

operational term whose meaning, measurement, and units are clearly understood. In this way, neither new science nor appeals to nonscience are needed to justify the impressive hierarchy of the cosmic-evolutionary story, from stars to plants to society.

Note that this complexity metric is not an original term; free energy rate density is a mere revision of an old one. Moreover, for years the same term has been labeled differently by specialized researchers: Φ_m is familiar to astronomers as the luminosity-to-mass ratio, to physicists as the power density, to geologists as the specific radiant flux, to biologists as the specific metabolic rate, and to engineers as the power-to-mass ratio. Free energy rate density is central to many varied subjects; all the more reason to use it to build a true interdisciplinary subject and to use it in search of unity across the spectrum of all the natural sciences [8,9,10,11].

Exobiology and Complexity, Table 1
Free energy rate densities for several representative systems

System	Duration (10^6 y)	Φ_m (erg/s/g)
Galaxy (Milky Way)	12,000	0.5
Star (main-sequence Sun)	10,000	2
Planet (Earth's climasphere)	5000	75
Plant (Earth's biosphere)	3000	900
Animal (hominid body)	10	20,000
Brain (human cranium)	1	150,000
Society (modern culture)	0	500,000

Table 1 lists values of Φ_m, in units of erg/sec/g, for seven representative systems (and their specific, computed cases in parentheses). Also listed is the duration, in millions of years, for each type of structure, dating back to their origins in the observational record. Clearly, Φ_m

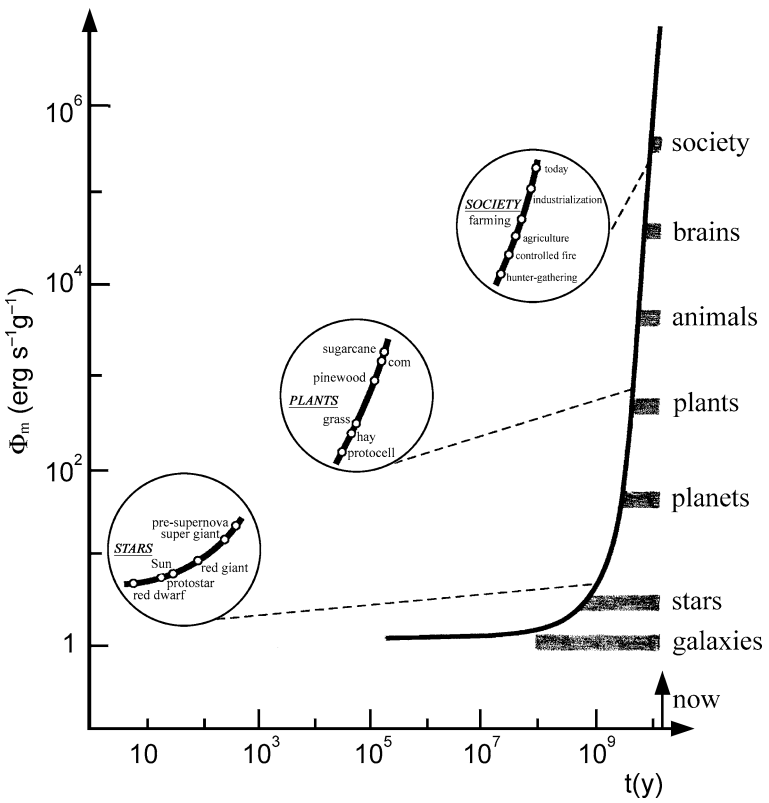

Exobiology and Complexity, Figure 6
Increase in free energy rate density, Φ_m, plotted as *horizontal histograms* when various open structures prospered in Nature, has been especially rapid in the last few billion years, much as expected from subjective intuition (Fig. 1b) and objective thermodynamics (Fig. 5c). The *drawn curve* approximates the increase in normalized energy flows characterizing order, form, and structure for a range of systems throughout the history of the Universe. The *circled insets* show greater detail of further measurements or calculations of free energy rate density for three representative systems – stars, plants, and society – typifying physical, biological, and cultural evolution, respectively. The *data in those circled insets* are discussed in Sect. "Complexity and Evolution, Broadly Considered". (Adapted from [9,10])

has increased as more intricately ordered systems have emerged throughout cosmic history, and dramatically so in relatively recent times.

The modeled flow of energy through a wide variety of open systems, alive or not, resembles the intuitive rise in complexity implied by Fig. 1b; it also mimics the potential rise of order in the above thermodynamic analysis of Fig. 5c. Complexity (at least as treated here energetically for localized structures) has indeed quantitatively increased over the course of natural history, and at a rate faster than exponential in more recent times ([9,11,13]; for details of energy computations and modeling, consult [10]).

Figure 6 plots the results listed in Table 1, where the Φ_m values are graphed as horizontal histograms for various systems' evolutionary durations to date. As expected:

- Stars and planets have small energy rate densities, $\Phi_m = 1\text{–}10^2$ erg/s/g
- Plants and animals have larger energy rate densities, $10^3\text{–}10^5$ erg/s/g
- Human societies have the largest known energy rate densities, $\sim 10^6$ erg/s/g.

Note that, although the total energy flowing through a star or planet is much greater than that through our individual bodies or brains, the *specific* rate (per unit mass) is larger for the latter – in fact, roughly a million times greater Φ_m for the human body than for the Sun.

This is not to say, by any means, that galaxies evolved into stars, or stars into planets, or planets into life. Rather, this analysis contends that galaxies gave rise to environments suited to the birth of stars, that some stars spawned environments conducive to the formation of planets, and that countless planets likely fostered environments ripe for the origin of life. Cosmic evolution, to repeat, incorporates both developmental and generational change.

Nor do these evolutionary phases, or historical durations, have well-determined start and stop times – or stop times necessarily at all. The horizontal histograms of Fig. 6 serve to stress that each of these phases once begun did not end; stars and galaxies, for example, first emerged in the earlier Universe, as also implied by the diagonal phases atop the arrow of time in Fig. 1a, but both such system types continue on presently originating, developing, and evolving, as do plants and animals that emerged much later. As depicted by those histograms yet unlike customary geological periods that do have set time intervals, currently all evolutionary phases noted in Figs. 1 and 6 operate simultaneously and indefinitely.

We thus arrive at a comprehensible reconciliation of the evident destructiveness of thermodynamics with the observed constructiveness of cosmic evolution. The sources and sinks of such energy flows passing through complex yet disparate entities such as stars, plants, and civilization all relate back to the time of thermal decoupling in the early Universe, when the conditions naturally emerged for the origin and evolution of order and organization.

Complexity and Evolution, Broadly Considered

Evolution should not be the sole province of biology, nor should its utility be of value only to life scientists. Darwin [19] never used the word "evolution" as a noun, in fact only once as a verb in the very last sentence of his classic book, *On the Origin of Species*. Nor need the principle of natural selection be the only mechanism of evolutionary change, past or present.

Actually, the term "selection" is itself a misnomer, for no known agent in Nature deliberately selects. Selection is not an active force or promoter of change as much as a passive mechanism that weeds out the unfit. As such, selected systems are simply those that remain after all the poorly adapted or less fortunate ones have been removed from a population of such systems. A better term might be "non-random elimination" [48]. What we really aim to explain are the adverse circumstances responsible for the deletion of some members of a group. Accordingly, selection can be generally taken to mean favorable interaction of any system with its environment – a more liberal interpretation that also helps widen the concept of evolution.

Selection works alongside the flow of resources into and out of all open systems, not just life-forms. Ordered systems are selected partly for their ability to utilize energy; and this energy is the "force", if there is any at all, in evolution. Broadly considered, selection occurs in the inanimate world as well as among animate objects, often providing a formative step in the production of order. It is energy flow and natural selection that together, working in tandem, underlie the "self" -assembly sketched in Fig. 2 – the former driving initial systems beyond equilibrium, the latter aiding the emergence of higher order in those systems that survive.

A handful of cases will suffice, among many others so documented [14], to illustrate the action of this energy-selection duo among a spectrum of increasingly ordered systems in successive phases of cosmic evolution:

- Red-giant stars are more complex than main-sequence stars
- Eukaryotes are more complex than prokaryotes
- Plants are more complex than protists
- Animals are more complex than plants
- Mammals are more complex than reptiles

- Brains are more complex than bodies
- Industrial society is more complex than hunter-gatherers.

Whether physical evolution of galaxies, stars and planets, or biological evolution of plants and animals on Earth, or cultural evolution of our technological civilization, a rather remarkable ranking order is apparent among all known organized structures. Stars, life, and society, all share a significant, common conclusion: Basic differences, both within and among these categories, are in degree, not in kind – namely, in degree of complexity arising from ongoing cosmic evolution. To justify this, consider below in greater detail each of the three major subsets of cosmic evolution noted in Sect. "Introduction".

Physical Evolution

Stars are good examples of physical evolution. Growing complexity can serve as an indicator of stellar aging – a developmental process – allowing stars to be judged as their interiors undergo cycles of nuclear fusion that result in greater thermal and chemical gradients. More data are needed to describe the increasingly differentiated, onion-like layers of fused heavy elements within highly evolved stars; more energy also flows per unit mass. Stellar size, color, brightness, and composition all change while passing on up the hierarchy of complexity for all stars, each stage using more free energy rate density:

- From protostars at "birth" ($\Phi_m \approx 0.5$ erg/s/g)
- To main-sequence stars at "mid-life" (~ 2)
- To red-giant stars in "old age" ($\sim 10^2$),
- To pre-supernovae near "death" ($\sim 5 \times 10^2$).

Those parenthetical values are the stars' increased energy rate densities, plotted among other values in the lower circled inset of Fig. 6. At least as regards energy flow, material resources, and structural integrity while experiencing change, stars have much in common with life. This is not to say that stars are alive, which is an occasional misinterpretation of such a broad statement. Nor do stars evolve in the strict and limited biological sense; most researchers would agree that stars *develop*. Yet close parallels are apparent among stars, including selection, adaptation, and perhaps even a kind of stellar reproduction – a generational process – reminiscent of the following Malthusian-inspired scenario:

Galactic clouds spawn clusters of stars, only a few of which (the more massive ones unlike the Sun) enable other, subsequent populations of stars to emerge in turn, with each generation's offspring showing slight variations, especially among the heavy elements contained within. Waves of "sequential star formation" [23] propagate through many such clouds like slow-motion chain reactions over eons of time – shock waves from the death of old stars triggering the birth of new ones – neither any one kind of star displaying a dramatic increase in number nor the process of regeneration ever being perfect. Those massive stars selected by Nature to endure the fusion needed to produce heavy elements are in fact the very same stars that often produce shocks to create new populations of stars, thereby both episodically and gradually enriching the interstellar medium with greater elemental complexity on timescales measured in millions of millennia. As always, the necessary though perhaps not sufficient conditions for the growth of complexity depend on the environmental circumstances and on the availability of energy flows in such (here, galactic) environments. All of which is reminiscent of stellar "evolution", minus any genes, inheritance, or overt function, for these are the value-added qualities of biological evolution that go well beyond physical evolution.

Continuing on and throughout the physical evolutionary subset of cosmic evolution, a *general* trend prevails, at least as pertains to Earth's environment that set the stage for life:

- Young rocky planets have greater Φ_m (~ 10 erg/s/g) than normal stars and galaxies (~ 1)
- Hydrothermal vents on at least one of those planets have more (~ 50)
- Planetary climaspheres, such as Earth's ocean-air interface, have even more (~ 100).

Note that some physical systems seem to be exceptions to the above findings, but upon closer inspection they are not exceptional at all. For example, that supernovae have very high values of Φ_m ($\gg 10^6$ erg/s/g) does not violate our complexity metric. The reason is that supernovae are not organized systems, in fact just the opposite; as excellent examples of totally disorganized explosions of massive stars, they have too much energy flow that is well outside the optimal range for stars, and thus we should not expect to properly plot chaotic supernovae among other clearly ordered systems in Fig. 6. Pre-supernovae are noted there, representing an advanced stage of stellar evolution and growing complexity prior to explosion, but supernovae themselves are destructive events more typical of retreat from complexity toward simplicity. Likewise, bombs, flames, and many other damaging events do have large energy throughput yet do not belong on this curve, thus do not partake of a general trend toward rising complexity in Nature.

Biological Evolution

Plants are good examples of biological evolution. Here, we trace increasing complexity among plant life on Earth where neo-Darwinism is clearly at work, making use of free energy rate densities well higher than those for galaxies, stars, and planets. As shown in the middle circled inset of Fig. 6, energy-flow diagnostics display a definite increase in complexity among various plants that locally and temporarily decrease entropy. The most dominant process in Earth's biosphere – photosynthesis – well illustrates that complexity hierarchy [29]:

- From simple hay or grass ($\Phi_m \approx 5 \times 10^2$ erg/s/g)
- To inefficient pinewood ($\sim 3 \times 10^3$)
- To more efficient corn ($\sim 6 \times 10^3$)
- To well cultivated sugarcane ($\sim 10^4$).

System functionality and genetic inheritance are two factors, above and beyond mere system structure, which help to enhance order among animate systems that are clearly living compared to inanimate systems that are clearly not. Unsurprisingly, more complex life-forms require the acquisition of more energy per unit mass per unit time for their well being.

Energy flows through plants as captured solar energy during the act of photosynthesis converts H_2O and CO_2 into nourishing carbohydrates; the previous low-grade disordering sunlight becomes, in a relative sense, a higher-grade ordering form of energy compared to the even lower-grade (infrared) energy re-emitted by Earth. Likewise, as regards previously discussed physical evolution, energy flows through stars as gravitational potential energy during the act of star formation converts into radiation released by mature stars; high-grade energy produced by gravitational and nuclear events yield greater (thermal and elemental) organization, yet only at the expense of their environments into which stars emit low-grade light abundant in entropy. Either way, energy is a fuel for evolution, fostering some systems to utilize increased power densities while driving others to destruction and extinction.

Onward across the bush of life (or the arrow of time) – cells, tissues, organs, organisms – much the same metric holds for animals (all in units of erg/s/g):

- Cold-blooded reptiles have greater Φ_m ($\sim 10^4$) than globally averaged plants ($\sim 10^3$)
- Warm-blooded mammals typically have more ($\sim 5 \times 10^4$)
- Some birds, during complex flight, can achieve even more ($\sim 7 \times 10^4$).

Human life itself can also be examined on finer scale to show how energy usage continues upward (per unit mass) for more complex tasks [30,63]:

- Laboring humans have greater Φ_m ($\sim 6 \times 10^4$) than sedentary humans ($\sim 2 \times 10^4$)
- Vigorously bicycling and intricately sewing humans have more ($\sim 10^5$)
- Thinking human brains themselves have even more ($\sim 2 \times 10^5$).

Starting with life's precursor molecules (the realm of chemical evolution) and all the way to human brains exemplifying the most complex clump of animate matter known (neurological evolution), the same *general* trend characterizes the complexity of plants and animals as for stars and planets: The greater the perceived complexity of the system, the greater the flow of free energy density through that system – either to build it, or to maintain it, or both.

No strong distinctions are made here among Φ_m values for members of the animal kingdom, except to note that they are nearly all within a factor of ten of one another, confined between those for photosynthesizing plants on the one hand and central nervous systems on the other. The results are broadly consistent with measured specific metabolic rates scaling inversely with body mass, $M^{-1/4}$, among a wide variety of animal species [37,70]. Suffice it to say that animals in the main and in accord with Fig. 6 fit well within the complexity trends for the major evolutionary stages of life and for the intermediate phases of cosmic evolution.

Note, however, as for some non-living systems above, a minority of living systems seem exceptional, their values of Φ_m somewhat out of bounds among other equally advanced biological systems. Occasional life-forms also display retreat from complexity, such as some bats that move deeper into caves over generations and thus gradually lose their eyesight, or snakes and whales that eventually lost legs over time. Exceptions, real and apparent, to any rule will likely occur in a biosphere so rich in numbers and diversity as ours on Earth. For example, respiring bacteria are problematic at face value, having Φ_m values as much as 10^6 erg/s/g [45], thus comparable to higher forms of life. But microbes are so highly metabolic only when environmental resources warrant; none of them respire continuously. Measured rates are often quoted for peak periods of high reproductivity. By contrast, more than three-quarters of all soil bacteria are virtually dormant and thus have Φ_m values orders of magnitude less while eking out a living in nutrient-poor environments. When all microbial rates are time-weighted, microbes' average values range in the

thousands of erg/s/g, as expected for systems of intermediate complexity. Likewise and to note just one other seemingly exceptional animal, the Komodo Dragon can consume 80% of its body weight at one meal, yet not need another meal for a month – however, its time-averaged metabolic rate is much less than its maximum rate while eating.

Birds are another case in point, as they are well known to have high specific metabolic rates ($\sim 3 \times 10^4$ erg/s/g) during periods of peak activity, such as when earnestly foraging for food for their nestlings. But, once again, upon closer inspection, they are recognized not to be exceptions at all. That the smallest animals have the highest such rates is often taken [70] as an explanation of their frequent eating habits (hummingbirds ingest up to half of their body mass daily), extreme levels of activity (bumblebees flap their wings up to 160 times per second), and relatively short lifespans (few years typically, given the heavy toll on their metabolic functions); those are operational tasks, namely, function, not structure. Given that birds and bees normally function in a three-dimensional aerial environment while solving advanced tasks in spatial geometry, materials science, aeronautical engineering, molecular biochemistry, and social stratification, then perhaps they ought to have large values of Φ_m. That birds, while airborne, have higher values than for resting humans should not surprise us since we ourselves have not solved the art of flying, an admittedly complex task. By contrast, when bicycling vigorously or sewing intricately, our specific metabolic rates do exceed even those of birds in flight as noted above. Moreover, when humans do fly, aided by built aircraft, machine values of Φ_m are indeed higher ($\sim 10^7$ erg/s/g) than for even the most impressively ingesting hummingbirds, as discussed in the next section on cultural evolution.

Cultural Evolution

Society is a good example of cultural evolution. Here, the cosmic-evolutionary chronicle continues, yet with greater normalized energy flows to power our obviously complex civilization. As plotted in the upper circled inset of Fig. 6, social progress can be tracked, again in terms of energy consumption, for a variety of human-related cultural advances among our hominid ancestors. Quantitatively, that same energy rate density increases:

- From hunter-gatherers of a few million years ago ($\Phi_m \approx 10^4$ erg/s/g)
- To agriculturists of several thousand years ago ($\sim 10^5$)
- To industrialists of two hundred years ago ($\sim 5 \times 10^5$)
- To western society today, on average ($\sim 10^6$).

That a cluster of brainy organisms working collectively in a social group is more energy intensive per capita (and thus more complex) than each of its individual human members – at least as regards the present criterion for order of free energy rate usage per capita – is a good example of a "whole greater than the sum of its parts", in this case for the open, non-equilibrated society that constitutes modern civilization [15].

The road to today's technological society was unquestionably built with increased energy use, as has been earlier recognized by many cultural historians (e. g., White [71]; Cook [18]; Brown [6]; Jervis [35]; McNeill and McNeill [50]), who noted the importance of rising energy expenditure per capita, a factor also more recently emphasized by practitioners of "big history" (Christian [17]; Spier [65]; Aunger [1]), a newly emerging subject that treats conventional history more deeply, indeed parallels the scenario of cosmic evolution.

Machines, too, and not just computer chips, but also ordinary motors and engines that typified the fast-paced economy of the 20th century, can be cast in evolutionary terms – though here the mechanism is less Darwinian than Lamarckian [39], with the latter's emphasis on accumulation of acquired traits. Either way, energy remains a driver, and with rapidly accelerating pace. Aircraft engines, for example, display clear evolutionary trends as engineering improvement and customer selection over generations of products have made engines more intricate, complex, and efficient, all the while utilizing enriched flows of energy density [63]:

- Gas-guzzling SUVs have greater Φ_m ($\sim 10^6$erg/s/g) than model-T automobiles ($\sim 10^5$)
- Boeing-747 jumbo jets of the last few decades have more ($\sim 10^7$)
- Military F-117 stealth aircraft of the present have even more ($\sim 10^8$).

Finer-scale evolutionary analysis of many technological advancements display evident progress toward greater complexity, such as for the typical American passenger car over the past two decades that can be cast in terms of growing horsepower-to-weight ratios provided by the US Highway Traffic Safety Administration: $\Phi_m = 5.9 \times 10^5$ erg/s/g in 1978, 6.8×10^5 in 1988, and 8.3×10^5 in 1998. Not surprisingly, silicon chips – a cultural icon of today's vibrant, digitized 21st-century economy – have immense flows of energy density, currently reaching values of $\sim 10^{10}$ erg/s/g mostly caused by chip miniaturization despite reduced power consumption.

Rare exceptions in cultural evolution's apparent drive toward greater complexity sometimes cause regression to-

ward simpler systems, much as for minor aspects of the physical and biological subsets of cosmic evolution. Collapse of civilizations, either internally (because of societal conflict) or externally (owing to environmental change), that then resort to social chaos is are examples of infrequent retreat from society's overall drive toward greater complexity [20,67].

Occasional exceptions aside, increasingly sophisticated technological gadgets, under the Lamarckian pressure of dealer competition and customer selection, do in fact show increases in Φ_m values with product improvement over the years. Not only can the cultural evolution of machines be traced and their Φ_m values computed as noted above for engines, but similar advances can also be tracked for a whole array of silicon-based devices now inundating our global economy. In keeping with the upper part of the curve in Fig. 6, many of these cultural devices do have complexity measures comparable to, and often greater than, biological systems, including brains. Technology clearly allows individual humans to accomplish things that cannot be done by us alone, and usually faster too, which partly explains why most of society continues to embrace technology, despite its pitfalls, to aid our senses and improve our increasingly complex lives.

Conclusions and Future Directions

This article has taken the liberty of extrapolating the word "evolution" in an intentionally broad way to analyze change on all spatial and temporal scales. Within the grand context of cosmic evolution, common threads have been identified linking a wide spectrum of ordered structures during an extremely long period of natural history, from big bang to humankind. More than any other single factor, energy flow seems to be a principal means whereby Nature's diverse systems naturally became increasingly complex in an expanding Universe, including not only galaxies, stars and planets, but also lives, brains and civilization.

The scenario of cosmic evolution accords well with observations demonstrating an entire hierarchy of structures to have emerged, in turn, throughout the history of the Universe: particles, galaxies, stars, planets, life, intelligence, and culture. As a general trend, an overall increase in complexity is apparent with the relentless march of time, without any progress, purpose or design implied. With cosmic evolution as our guide, we can begin to understand the environmental conditions needed for matter to have become increasingly ordered, organized, and complex. This rise in order, form, and structure violates no laws of physics, and certainly not those of modern thermodynamics. Nor is the idea of ubiquitous change novel to modern worldviews. What is new and exciting is the way that frontier, non-equilibrium science now helps us unify a holistic cosmology wherein complex life plays an integral role – namely, to address the origin and evolution of all things by means of logic, rationality, and the methods of natural science.

When studying complexity in Nature, some researchers prefer the concept of information rather than energy, often becoming displeased when the former is put aside as done in this article. But information content has had a muddled history full of assorted interpretations – meaningful information, the value of information, Shannon information, algorithmic information, raw information. Furthermore, no one has yet shown quantitatively and unambiguously, that information content rises throughout the ages for physical, biological, and cultural systems. A useful future research direction would clarify the role of the information sciences in complexity studies and cosmic evolution, including the possibility that information is merely other forms of energy – energy acquired, energy stored, and energy expressed.

By contrast, it is encouraging that a single quantity such as free energy rate density, defined here clearly and with units well understood, affects all ordered systems, given that some systems are regulated by gravity and others practically not. Thermodynamics does pertain to all such systems universally, whether massive enough like stars subject to gravity or less so as for life-forms governed mostly by electromagnetism. Energy flow is a common feature of every open, non-equilibrium system, and it is insightful not only that one such quantity is uniformly applicable but also that it seems to map reasonably well the rise of complexity among many known systems. Gravitational force in physics, natural selection in biology, and technological innovation in culture are all examples of diversified actions that can give rise to accelerated rates of change at locales much smaller than the Universe per se – such as the islands of order that are stars, life, and civilization itself. Indeed, our use of energy wisely and optimally will likely guide our fate along the future arrow of time, for we humans are also part of the cosmic-evolutionary scenario.

Humankind is now moving toward a time, possibly as soon as within a few generations, when we shall no longer be able to expect Nature to easily provide for our own survival. Rather, civilization on Earth will either have to adapt to the natural environment with ever-accelerating speed, or to generate artificially controlled environments (either on or beyond Earth) needed for our ecological existence. From two of Nature's most advanced yet locally ordered systems – society and machines – will likely emerge

a symbiotic technoculture, the epitome (thus far as best we know) of complexity in the Universe – a new technology-based system that will likely require even greater values of energy rate density, as the curve in Fig. 6 continues racing upward. Can humanity endure despite its own increasing complexity? Or will our species transform into some other intricate entity as complexity continues to rise?

I thank *la Fondation Wright de Geneve* for support and encouragement of this research, which comprises the intellectual theme of the Wright Center for Science Education at Tufts University. The author is also associated with the Harvard College Observatory, where he teaches a course on the topic of cosmic evolution and complexity science, for which a multi-media web site has been built http://www.tufts.edu/as/wright_center/cosmic_evolution.

Bibliography

1. Aunger R (2007) Major transitions in 'big' history. Tech Forecast Soc Chang 68:27
2. Bennett CL et al (2003) Wilkinson Microwave Anisotropy Probe (WMAP) basic results. Astrophys. J Suppl Ser 148:1
3. Blum HF (1968) Time's arrow and evolution. Princeton University Press, Princeton
4. Bonner JT (1988) Evolution of complexity. Princeton University Press, Princeton
5. Brooks DR, Wiley EO (1988) Evolution as entropy. University of Chicago Press, Chicago
6. Brown H (1976) Energy in our future. Ann Rev Energy 1:1
7. Chaisson EJ (1981) Cosmic dawn: origins of matter and life. Atlantic Monthly Press, Boston
8. Chaisson EJ (1987) The life era (appendix). Atlantic Monthly Press, New York
9. Chaisson EJ (1998) The cosmic environment for the growth of complexity. BioSystems 46:13
10. Chaisson EJ (2001) Cosmic evolution: the rise of complexity in nature. Harvard University Press, Cambridge
11. Chaisson EJ (2003) A unifying concept for astrobiology. Int J Astrobio 2:91
12. Chaisson EJ (2004) Complexity; an energetics agenda. Complex J Santa Fe Inst 9:14
13. Chaisson EJ (2005) Non-equilibrium thermodynamics in an energy-rich universe. In: Kleidon A, Lorenz RD (eds) Non-equilibrium Thermodynamics and the Production of Entropy. Springer, Berlin
14. Chaisson EJ (2006) Epic of evolution: seven ages of the cosmos. Columbia University Press, New York
15. Chaisson EJ (2009) Cosmic evolution: state of the science. In: Dick S (ed) Cosmos and Culture. NASA, Washington
16. Chambers R (1844) Vestiges of the natural history of creation. Churchill, London
17. Christian D (2004) Maps of time: introduction to big history. University California Press, Berkeley
18. Cook E (1971) The flow of energy in an industrial society. Sci Am 224:135
19. Darwin C (1859) On the origin of species. J Murray, London
20. Diamond J (2005) Collapse: how societies choose to fail or succeed. Viking, New York
21. Dyke C (1988) Cities as dissipative structures. In: Weber BH et al (eds) Entropy, Information, and Evolution. MIT Press, Cambridge
22. Dyson F (1979) Time without end: physics and biology in an open universe. Rev Mod Phys 51:447
23. Elmegreen B, Lada C (1977) Sequential star formation of subgroups in OB associations. Astrophys J 214:725
24. Fox RF (1988) Energy and the evolution of life. Freeman, San Francisco
25. Frautschi S (1982) Entropy in an expanding universe. Science 217:593
26. Gold T (1962) The arrow of time. Am J Phys 30:403
27. Haken H (1978) Synergetics. Springer, Berlin
28. Haken H (1975) Cooperative phenomena in systems far from thermal equiblibrium and in nonphysical systems. Rev Mod Phys 47:67
29. Halacy DS (1977) Earth, water, wind and sun. Harper & Row, New York
30. Hammond KA, Diamond J (1997) Maximal sustained energy budgets in humans and animals. Nature 386:457
31. Henderson L (1913) Fitness of the environment. Macmillan, New York
32. Hofkirchner W (ed) (1999) The quest for a unified theory of information. Gordon & Breach, Amsterdam
33. Jantsch E (1980) Self-organizing universe. Pergamon, Oxford
34. Jaynes ET (1957) Information theory and statistical mechanics. Phys Rev 108:171
35. Jervis R (1997) System effects: complexity in political and social life. Princeton University Press, Princeton
36. Kauffman S (1993) The origins of order. Oxford University, Press
37. Kleiber M (1961) The fire of life. Wiley, New York
38. Kleidon A, Lorenz RD (eds) (2005) Non-equilibrium thermodynamics and the production of entropy. Springer, Berlin
39. Lamarck J-B (1809) Philosophie Zoologique. Editions du Seuil, Paris
40. Layzer D (1976) The arrow of time. Astrophys J 206:559
41. Layzer D (1988) Growth of order in the universe. In: Weber BH et al (eds) Entropy, Information, and Evolution. MIT Press, Cambridge
42. Lewin R (1992) Complexity. Macmillan, New York
43. Lineweaver C (2005) Cosmological and biological reproducibility: limits on maximum entropy production principle. In: Kleidon A, Lorenz RD (eds) Non-equilibrium Thermodynamics and the Production of Entropy. Springer, Berlin
44. Lotka A (1922) Contribution to the energetics of evolution. Proc Nat Acad Sci USA 8:147
45. Margulis L, Sagan D (1986) Microcosmos. Simon & Schuster, New York
46. Marijuan PC et al (eds) (1996) First conference on foundations of information sciences. Biosystems 38:87
47. Matsuno K (1989) Protobiology: Physical Basis of Biology. CRC Press, Florida
48. Mayr E (1997) This is biology. Harvard University Press, Cambridge
49. McMahon T, Bonner JT (1983) On size and life. Freeman, San Francisco
50. McNeill JR, McNeill WH (2003) The human web. Norton, New York
51. Morowitz HJ (1968) Energy flow in biology. Academic Press, New York

52. Morrison P (1964) A thermodynamic characterization of self-reproduction. Rev Mod Phys 36:517
53. Odum HT (1971) Environment, power, and society. Wiley, New York
54. Prigogine I (1961) Introduction to thermodynamics of irreversible processes. Wiley, New York
55. Prigogine I, Nicolis G, Babloyantz A (1972) Thermodynamics of evolution. Physics Today 11:23
56. Reeves H (1981) Patience dans l'azur: l'evolution cosmique. Editions du Seuil, Paris
57. Sagan C (1980) Cosmos. Random House, New York
58. Salk J (1982) An evolutionary approach to world problems. UNESCO, Paris
59. Schneider ED, Kay JJ (1995) Order from disorder: thermodynamics of complexity in biology. In: Murphy M, O'Neill L (eds) What is life. Cambridge University Press, Cambridge
60. Schroedinger E (1944) What is life? Cambridge University Press, Cambridge
61. Shannon CE, Weaver W (1949) Mathematical theory of communication. University of Illinois Press, Champaign-Urbana
62. Shapley H (1930) Flights from chaos. McGraw Hill, New York
63. Smil V (1999) Energies. MIT Press, Cambridge
64. Spencer H (1896) A system of synthetic philosophy. Williams and Norgate, London
65. Spier F (2005) How big history works. Soc Evol Hist 4:25
66. Szathmary E, Maynard Smith J (1995) The major evolutionary transitions. Nature 374:227
67. Tainter JA (1988) The collapse of complex societies. Cambridge University Press, Cambridge
68. Ulanowicz RE (1986) Growth and development. Springer, Berlin
69. Weber BH, Depew DJ, Smith JD (eds) (1988) Entropy, information, and evolution. MIT Press, Cambridge
70. West GB, Brown JH, Enquist BJ (1999) Fourth dimension of life. Science 284:1677
71. White LA (1959) The evolution of culture. McGraw-Hill, New York
72. Whitehead AN (1925) Science and the modern world. Macmillan, New York
73. Wicken JS (1987) Evolution, thermodynamics, and information. Oxford University Press, Oxford
74. von Bertalanffy L (1932) Theoretische biologie. Borntraeger, Berlin
75. von Bertalanffy L (1968) General system theory. Braziller, New York

Exobiology (theoretical), Complexity in

Axel Brandenburg
Nordita, AlbaNova University Center, Stockholm, Sweden

Article Outline

Glossary
Definition of the Subject
Introduction
Homochirality
Establishing Hereditary Information
Alteration of the Environment by Early Life
Conclusions
Bibliography

Glossary

Chiral, achiral and racemic A molecule is chiral if its three-dimensional structure is different from its mirror image. Such molecules tend to be optically active and turn the polarization plane of linearly polarized light in the right- or left-handed sense. Correspondingly, they are referred to as D- and L-forms, which stand for dextrorotatory and levorotatory molecules. An achiral molecule is mirror-symmetric and does not have this property. A substance is racemic if it consists of equally many left- and right-handed molecules. A polymer is said to be isotactic if all its elements have the same chirality.

Enantiomers and enantiomeric excess Enantiomers are a pair of chiral molecules that have opposite handedness, but are otherwise identical. Enantiomeric excess, usually abbreviated as e.e., is a normalized measure of the degree by which one handedness dominates over the other one. It is defined as the ratio of the difference to the sum of the two concentrations, so e.e. always falls between -1 and $+1$.

Epimerization and racemization Epimerization is a spontaneous change of handedness of one sub-unit in a polymer. Racemization indicates the loss of a preferred handedness in a substance.

Catalysis and auto-catalysis Catalysts are agents that lower the reaction barrier. A molecule reacts with the catalyst, but at the end of the reaction, the catalyst emerges unchanged. This is called catalysis. In auto-catalysis the catalyst is a target molecule itself, so this process leads to exponential amplification of the concentration of this molecule by using some substrate. Biological catalysts are referred to as enzymes.

Nucleotides and nucleic acids Nucleotides are monomers of nucleic acids, e. g., of RNA (ribonucleic acid) or DNA (deoxyribonucleic acid). They contain one of four nucleobases (often just called bases) that can pair in a specific way. Nucleotides can form polymers, and their sequence carries genetic information. One speaks about a polycondensation reaction instead of polymerization because one water molecule is removed in this step. Other nucleotides of interest include peptide nucleic acid or PNA. Here the backbone is made of peptides instead of sugar phosphate.

Peptides and amino acids Amino acids are molecules of the general form NH_3-CHR-COOH, where R stands for the rest, which makes the difference between different amino acids. For glycine, the simplest amino acid, we have R=H, so two of the bonds on the central C atom are the same and the molecule is, therefore, chiral. For alanine, R=CH_3, so all four bonds on the central C atom are different, therefore this molecule is chiral. A peptide is a polymer generated through a polycondensation reaction of amino acids. Peptides are also referred to as proteins.

Solar constant and albedo The solar constant is the total flux of energy from the Sun above the Earth's atmosphere. Its current value is $S = 1.37\,\mathrm{kW\,m^{-2}}$, but it was about 30% lower when the solar system was young (10^8 yr ago, say), so S is not a constant. The albedo A is the fraction of the Sun's energy that is reflected from the surface of the Earth, e. g., by clouds and snow and, to a lesser extent, by land masses and oceans.

Photosynthesis and carbon fixation Photosynthesis uses light to reduce CO_2 and to produce oxygen either as free molecular oxygen or in some other chemical form. This process removes CO_2 from the atmosphere and produces biomass, which is written in simplistic form as $(CH_2O)_n$. This process is referred to as carbon fixation.

Life A preliminary definition of life involves replication and death, coupled to a metabolism that utilizes any sort of available energy. Life is characterized further by natural selection to adapt to environmental changes and to utilize available niches. A proper definition of life is difficult given that all life on Earth can be traced back to a single common ancestor. Any definition of life may need to be adjusted if extraterrestrial or artificial life is discovered.

Definition of the Subject

Astrobiology is concerned with questions regarding possible origins of life on Earth and elsewhere in the Universe. Although to date there has been no detection of extraterrestrial life, it is generally assumed that life could be widespread, provided certain conditions of habitability are met. A common implicit hypothesis in astrobiology is that life can emerge spontaneously once certain environmental conditions are met. This implies that there may well have been multiple geneses, separated only by global extinction events, such as major impacts by other celestial bodies [11].

Four important discoveries can be named that have provided impetus to the field of astrobiology.

1. More than 300 extrasolar planets have been discovered since 1995, providing explicit targets for detecting life outside the solar system.
2. Recent Mars missions have provided evidence for liquid water on the surface of Mars in the past and possibly even in the present time. This has fostered the search for techniques to detect microbial life on Mars.
3. On earth the carbon in very old sedimentary rocks dating back 3.8 Gyr ago shows a consistently lower ^{13}C to ^{12}C abundance ratio, which is normally indicative of life. This lends support to the notion that life may have been present as soon as the Earth's surface became hospitable.
4. The discovery of extremophiles on Earth has considerably extended the definition of habitability to include extreme temperatures, pressures and pH values, high salinity as well as high radiation levels. This has raised hopes of finding life elsewhere in our solar system.

Astrobiology thus comprises several scientific disciplines: astronomy, geology, chemistry, and biology. Therefore, much of the original literature tends to appear in journals in these various fields. We should also mention that there are technological attempts being made to produce artificial life [35]. While this approach is not aimed at reproducing the origin of life on Earth, it may still be useful for prompting our imaginations in understanding the transition from nonliving to living matter.

Introduction

Since the early days of nonlinear dynamics and non-equilibrium thermodynamics, it has been clear that one of the ultimate applications of this theory might be to facilitate an understanding of the transition from non-living to living matter. The main reason is obviously that living systems are very far from equilibrium – as indicated by the high degree of order, and hence the low entropy, of living systems relative to their environment.

As early as 1952, Turing [44] proposed the idea that chemical reaction-diffusion systems might provide a tool for studying biochemical pattern formation, which has increased our understanding of the laws of nature far from equilibrium, where life occurs. This idea was followed up in the late 1960s by Prigogine [33,34] who suggested that dissipative structures have great importance in establishing a physical description of living matter. A general theory of autocatalytic molecular evolution was developed in 1971 by Eigen [15], who argued that in a single micro-environment, only a single handedness can result from a single event. In particular, the famous chicken and egg problem that occurs in biology at different levels was identi-

fied as a Hopf bifurcation. A Hopf bifurcation describes the spontaneous emergence of an oscillating solution once some stability threshold has been crossed. The mathematics of this is familiar to any physicist, but it requires that the equations describing the relevant physics are known. In biology, it is not even clear that the various phenomena can be described by equations. A first detailed attempt in this direction was, indeed, that of Eigen. However, the equations governing the emergence of life are only phenomenological ones. Nevertheless, these approaches are invaluable in that they help give the origin of life question a mathematical basis.

One of the earliest anticipated forms of life that is still similar to present life is the RNA world [21], whereby simple RNA molecules with functional behavior self-reproduce using genetic information encoded either in the same or in other participating RNA molecules. Obviously, there are tremendous difficulties, given that RNA is too complicated a molecule to be synthesized abiologically. A significantly simpler molecule is peptide nucleic acid or PNA [30], in which the backbone consists of peptide instead of sugar phosphate. Nevertheless, the difficulty of producing RNA remains.

There is no firm idea where on Earth such molecular replication may have originally taken place. Frequently discussed scenarios include hydrothermal vent systems [37], and also beach scenarios that are subject to tides leading to cyclic changes in concentration [24] as well as to repeated wetting and drying [7].

An early experiment that contributed significantly to the research into the origins of life was the Urey–Miller experiment [26], which demonstrated the spontaneous production of amino acids in a reducing atmosphere consisting of H_2O vapor, CH_4, NH_3, and H_2 with an energy supply in the form of sparks. More recent experiments also allow for the presence of CO_2, which now seems unavoidable on the early Earth, given that it is continuously replenished through outgassing by volcanoes.

In the following, we review some areas in which there has been considerable cross-fertilization between astrobiology and nonlinear dynamics. We begin by discussing a phenomenon that is believed to have taken place around the time of the origin of life, namely the establishment of a definitive handedness of biomolecules that is inherent in DNA and RNA (D-form) and in amino acids (L-form). Next, we discuss constraints on the evolution of hereditary information and, finally, we review some models that characterize the alteration of the terrestrial environment by early life. In addition to any of these physical effects, random fluctuations lead inevitably to local imbalances between the concentrations of molecules of D- and L-form. In the following, we discuss mechanisms that can lead to an exponential amplification of enantiomeric excess. For a recent review of these ideas see [32].

Homochirality

Theories of a chemical origin of life involve polymerization of nucleotides that carry and utilize genetic information. Ribonucleotides possess chirality, i. e., they are different from their mirror images. All known life forms use ribonucleotides of the so-called D-form (right-handed), as opposed to the L-form (left-handed). These two molecules are referred to as opposite *enantiomers*. In most cases these different enantiomers are optically active, i. e., they turn the polarization plane of linearly polarized light in a right-handed or left-handed sense.

Any non-enzymatic synthesis of ribonucleotides would have produced a mixture of equally many right- and left-handed building blocks. Technically this is referred to as a *racemic* mixture of these molecules. However, it is known experimentally that, in a racemic mixture of mononucleotides, polymerization is quickly terminated after the first or second polymerization step [23]. This is generally referred to as *enantiomeric cross-inhibition*, which was long thought to be a serious obstacle to a chemical origin of life. It was therefore thought to be necessary that life evolved only in a homochiral environment. Moreover, it would then be necessary that the degree of enantiomeric purity must have been very high. This is important because it rules out a number of physical mechanisms based on the enantioselective effects of circularly polarized radiation, magnetic fields, and the parity-breaking property of the electroweak force.

The Frank-Mechanism

A general mechanism for producing complete homochirality was proposed in 1953 by Frank [19] based on the assumed effects of auto-catalysis and what he called mutual antagonism. In fact, the enantiomeric cross-inhibition mentioned above can be thought of as a possible example of mutual antagonism. Frank's model is characterized by the following set of three reactions:

$$D + S \xrightarrow{k_C} D + D\,, \tag{1}$$

$$L + S \xrightarrow{k_C} L + L\,, \tag{2}$$

$$D + L \xrightarrow{k_I} DL\,, \tag{3}$$

where D and L denote monomers of the two enantiomers, S is a substrate from which monomers could be formed

via auto-catalysis, and DL are inactive dimers that are lost from the system. (At this level of simplification no distinction is made between DL and LD. This simplification will later be relaxed.) The parameters k_C and k_I characterize the reaction speeds. These reactions translate to the following set of equations for the concentrations of D, L, S, and DL,

$$\frac{d}{dt}[D] = +k_C[S][D] - k_I[D][L]\,, \tag{4}$$

$$\frac{d}{dt}[L] = +k_C[S][L] - k_I[D][L]\,, \tag{5}$$

$$\frac{d}{dt}[S] = -k_C[S]\Big([D] + [L]\Big)\,, \tag{6}$$

$$\frac{d}{dt}[DL] = +2k_I[D][L]\,. \tag{7}$$

These equations imply that the total mass of all building blocks (including the substrate), is constant, i. e. $[D] + [L] + [S] + [DL] = \text{const} \equiv M$.

This system of equations describes the continued autocatalytic production of DL, D and L until the substrate S is exhausted, i. e., $[S] = 0$. However, as long as $[S]$ is still finite, the asymmetry, $\mathcal{A} = [D] - [L]$, grows quasi-exponentially, proportional to $\exp(\int [S]\,dt)$. A numerical example of this is shown in Fig. 1.

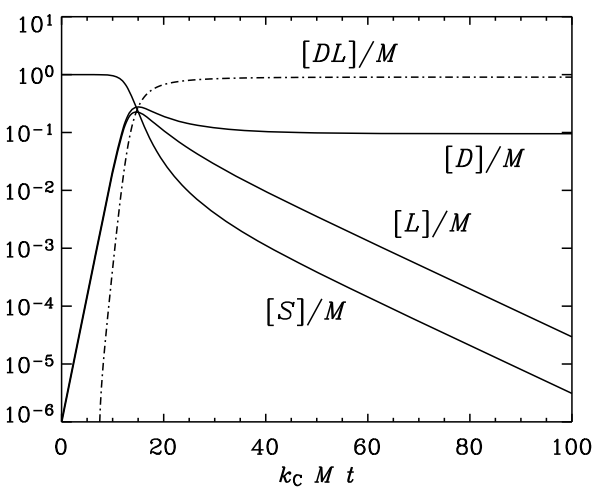

Exobiology (theoretical), Complexity in, Figure 1
Solution of Eqs. (4)–(6) for $k_I = k_C$. Both D and L grow exponentially until $[D] + [L]$ becomes comparable to the constantly declining substrate concentration $[S]$. At the same time the production of DL removes an equal amount of D and L, but this effect primarily affects those enantiomers that are already in the minority. In this calculation an initial asymmetry (here 10%) of $[D] - [L]$ grows until saturation. At the end, $[D]$ has reached 100% enantiomeric excess, but this happened at the expense of producing a large number of inactive heterochiral dimers DL

In the numerical example above, we started with very small initial concentrations. Another possibility is to start with a perturbed racemic solution. The racemic solution is given by $[D] = [L] = \lambda/k_I$, where $\lambda = k_C[S]$ is the instantaneous growth rate due to auto-catalysis. Under the assumption that λ can be treated as a constant (i. e., when the system is still nearly racemic), a linear stability analysis shows that the enantiomeric excess,

$$\text{e.e.} = \frac{[D] - [L]}{[D] + [L]} \tag{8}$$

grows exponentially. This means that the racemic solution is unstable and that the mechanism for achieving homochirality is based on a linear instability.

Continued Polymerization

There is a priori no good reason to permit the production of heterochiral dimers DL, but not homochiral dimers DD and LL, i. e.,

$$D + D \xrightarrow{k_S} DD\,, \tag{9}$$

$$L + L \xrightarrow{k_S} LL\,. \tag{10}$$

The importance of such reactions was stressed in a review by Blackmond [1], who also introduced an additional modification that consists in the assumption that, rather than monomers, the homochiral dimers DD and LL catalyze the production of monomers, i. e., reactions (1) and (2) are replaced by

$$DD + S \xrightarrow{k_C} DD + D\,, \tag{11}$$

$$LL + S \xrightarrow{k_C} LL + L\,. \tag{12}$$

This model is similar to the original Frank model provided there is a way of getting rid of those homochiral dimers that are in the minority. This requires enantiomeric cross-inhibition for dimers to form heterochiral trimers, i. e., we need the additional reactions

$$DD + L \xrightarrow{k_I} DDL\,, \tag{13}$$

$$LL + D \xrightarrow{k_I} LLD\,. \tag{14}$$

A solution to the corresponding reaction equations is given in Fig. 2. Reaction calorimetry supports the assumption that dimers and not monomers are the relevant catalysts [1]. This seems to apply, in particular, to the first autocatalytic reaction ever found that enhances enantiomeric excess [42]. In this reaction (sometimes referred to as the Soai reaction), the substrate is pyridine-3-carbaldehyde and the chiral molecule of either D- or L-form is 3-pyridyl

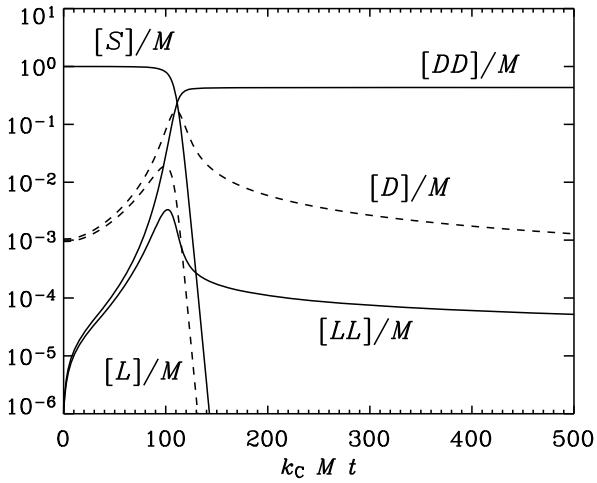

Exobiology (theoretical), Complexity in, Figure 2
Solution of Eqs. (4), (6) supplemented by kinetic equations corresponding to the reactions Eqs. (9), (14), for $k_S = k_I = k_C$. Again, an initial 10% asymmetry of $[D] - [L]$ grows until $[D] + [L]$ becomes comparable to the constantly declining substrate concentration $[S]$. The monomers polymerize into dimers *DD* and *LL*. Toward the end, $[DD]$ reaches 100% enantiomeric excess

alkanol, which thus acts as an asymmetric autocatalyst to produce more of itself. In this reaction, however, dialkylzinc acts as an additional achiral catalyst. While the Soai reaction is important as a first explicit example of an autocatalytic reaction that enhances the enantiomeric excess, it is not normally regarded as directly important for astrobiology.

The polymerization model was developed by Sandars [39], who included arbitrarily many polymerization steps of the form

$$D_n + D \xrightarrow{k_S} D_{n+1}\,, \tag{15}$$

$$L_n + L \xrightarrow{k_S} L_{n+1}\,, \tag{16}$$

$$D_n + L \xrightarrow{k_S} D_n L\,, \tag{17}$$

$$L_n + D \xrightarrow{k_S} L_n D\,. \tag{18}$$

The basic outcome of this and similar models is always the same as in the original Frank model, except that the polymerization model is also capable of displaying interesting wave-like dynamics in time-dependent histograms of different polymers [5].

Spatially Extended Models

In reality, there are limits as to the degree to which a system can be considered fully mixed. In general, $[D]$ and $[L]$ should be functions of time *and* space, i. e. $[D] = [D](t, \boldsymbol{x})$ and $[L] = [L](t, \boldsymbol{x})$. Assuming that there is only molecular diffusion, the relevant reaction equations are to be supplemented by additional diffusion terms,

$$\frac{\mathrm{d}}{\mathrm{d}t}[D_n] = R_n^{(D)} + \kappa\nabla^2[D_n]\,, \tag{19}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}[L_n] = R_n^{(L)} + \kappa\nabla^2[L_n]\,, \tag{20}$$

where $R_n^{(D)}$ and $R_n^{(L)}L$ are the right hand sides of the reaction equations.

If there were only one type of handedness, the resulting equation would be reminiscent of the Fisher equation [29],

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \lambda(1-f)f + \kappa\nabla^2 f\,, \tag{21}$$

which admits propagating front solutions with front speed $v_{\text{front}} = 2\sqrt{\kappa\lambda}$. Here, f could represent the local concentration of some disease in models of the spread of epidemics, for example.

In the present case, there are two fields, one of each handedness. It is instructive to refer to these fields as populations, which is suggestive of their ability to replicate, migrate, become extinct, and to compete against a population of opposite handedness. Each population is able to expand into unpopulated space at a speed given approximately by v_{front}, but once two opposing handednesses come into contact, there is an impasse and the propagation comes to a halt. A snapshot of a one-dimensional model illustrating polymer length as a function of position is shown in Fig. 3 for populations of opposite handedness that have come into contact.

The overall dynamics of symmetry breaking are well characterized by a low order truncation, where the model is truncated at $n = 2$ and the evolution of the $n = 1$ modes is assumed to be enslaved by the evolution of the $n = 2$ modes [5]. An example of such a solution is shown in Fig. 4, which shows the evolution in a space-time diagram, where two populations of opposite handedness expand into unpopulated space until two opposite populations come into contact.

In two and three dimensions, a front between two opposing enantiomers is (in general) curved, in which case it can propagate diffusively in the direction of curvature. This is caused by the fact that the inner front between two populations is slightly shorter than the outer one. (Only the immediate proximity of a front matters; what lies behind it is irrelevant if it is of the same handedness.) Indeed, on a two-dimensional surface the inner front is shorter than the outer one by $2\pi d$ – independent of radius. Here, d is the front thickness, which is of the order of $d \approx (\kappa/\lambda)^{1/2}$.

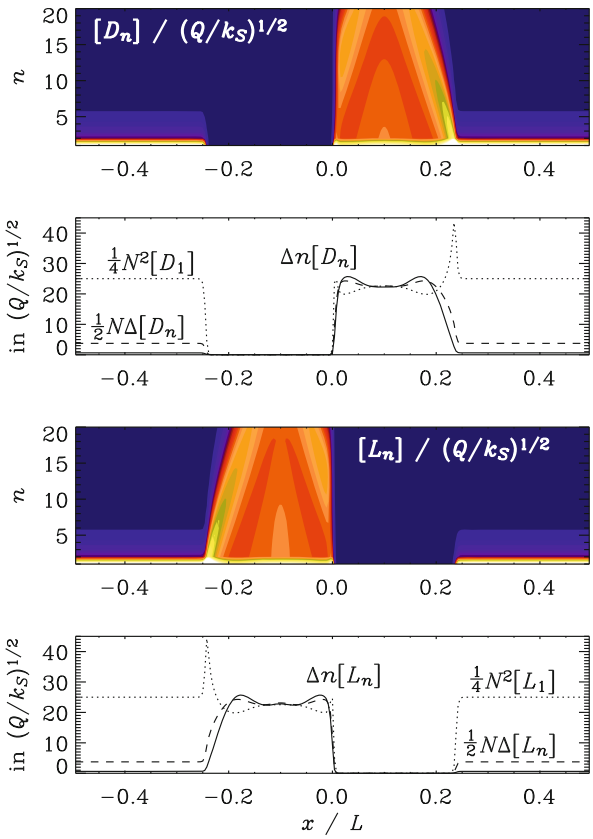

Exobiology (theoretical), Complexity in, Figure 3
Color/gray scale plots of $[D_n]$ and $[L_n]$ for $t/\tau_{\text{diff}} = 0.8$ as a function of x and n, and the corresponding dependencies of $\sum_{n=1}^{N} n[D_n]$ and $\sum_{n=1}^{N} n[L_n]$ (*solid line*), compared with $\frac{1}{2}N\sum_{n=1}^{N}[D_n]$ and $\frac{1}{2}N\sum_{n=1}^{N}[L_n]$ (*dashed line*), and $\frac{1}{4}N^2[D_1]$ and $\frac{1}{4}N^2[L_1]$ (*dotted line*), all in units of $(Q/k_S)^{1/2}$. The normalized diffusivity is $\kappa/(L^2\lambda_0) = 10^{-2}$ and $N = 20$. Adapted from [4]

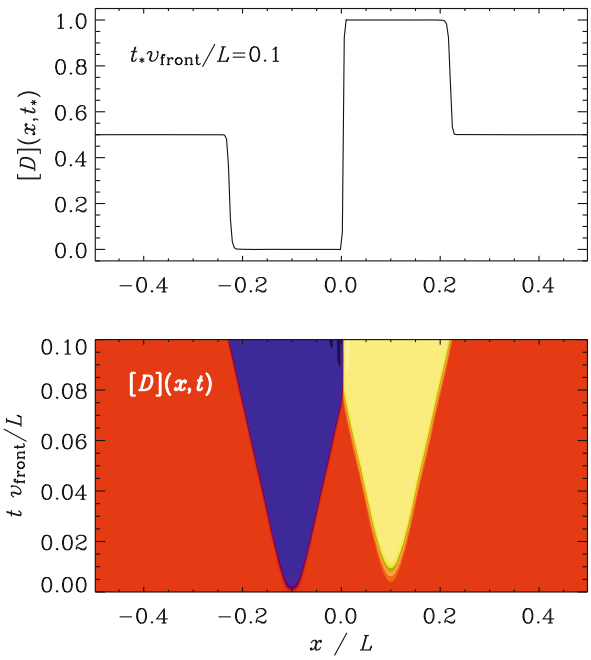

Exobiology (theoretical), Complexity in, Figure 4
Profile of $[D](x, t_*)$ and space-time diagram of $[D](x, t)$ for the one-dimensional problem without advection and an initial perturbation corresponding to a weak (amplitude 0.01) right-handed excess at $x/L = 0.1$ (marked in *white* or *yellow*) and a somewhat stronger (amplitude 0.3) left-handed excess at $x/L = -0.1$ (marked in *dark* or *blue*). Note the propagation of fronts with constant speed if the exterior is racemic (i. e. $[D] = [L] = 1/2$, shown in medium shades or red) and a non-propagating front when the chirality is opposite on the two sides of the front. The normalized diffusivity is $\kappa/(L^2\lambda_0) = 10^{-2}$, i. e. the same as in Fig. 3. Adapted from [4]

It turns out that in two dimensions the rate of change of the integrated asymmetry, $\mathcal{A} = \int([D] - [L])\,\mathrm{d}^2x$, depends only on the *number* of topologically distinct rings or islands. Once an island is wiped out, the rate of change of $\mathcal{A}$ changes abruptly and then stays constant until the next island gets wiped out. So the enantiomeric excess,

$$\text{e.e.} = \frac{\int([D] - [L])\,\mathrm{d}^2x}{\int([D] + [L])\,\mathrm{d}^2x}\,, \tag{22}$$

increases with time in a piecewise linear fashion.

Even if at each point homochirality could be reached rapidly (time scale λ^{-1}), global homochirality requires that one population wipes out the other one completely. Diffusion is usually too slow to lead to any significant mixing and hence to global homochirality. However, there could be circumstances where such mixing is sped up by something like "turbulent" transport. In the case of the Earth, the slowest relevant transport is in the Earth's mantle, part of which is now associated with what is called the deep biosphere. Assuming that multiple geneses of life is possible, this would raise the question of whether a simultaneous co-existence of different handednesses on different parts of the early Earth would have been possible. It is, however, unclear whether this possibility could have left any traces that would still be detectable today.

Another approach to solving the problem of spatially extended chemistry is by means of cellular automata. In this approach, points on a mesh can take different states corresponding to molecules of right or left handedness, achiral substrate molecules, or even empty states. An example of such a calculation by Shibata et al. [41] is shown in Fig. 5. Again, there are patches of populations of opposite handedness that grow and wipe each other out, such that in the end only one handedness survives.

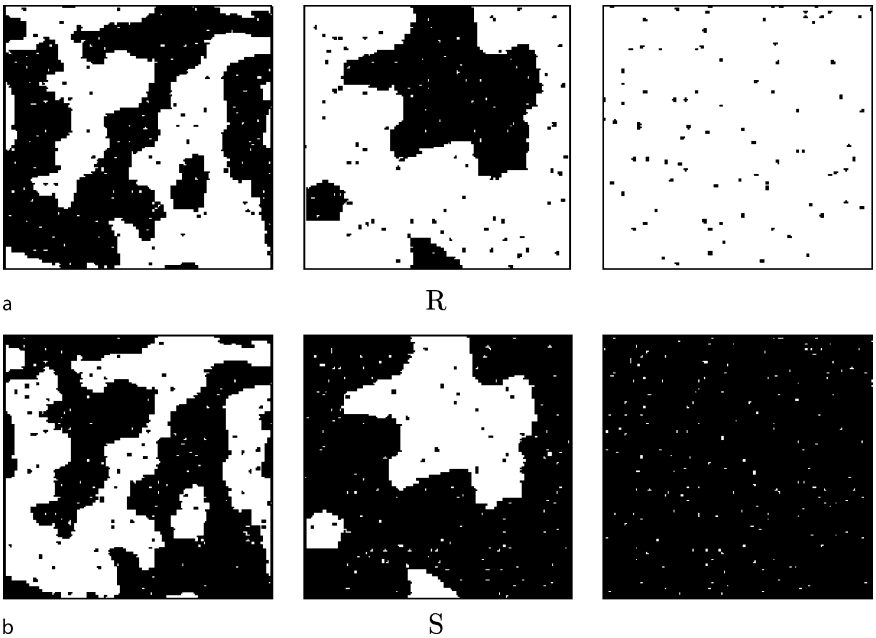

Exobiology (theoretical), Complexity in, Figure 5
The evolution of molecules of D- and L-forms is shown in the upper and lower panels, respectively. Note the tendency toward complete homochiralization by gradually filling up isolated islands with the chirality of the surrounding molecules. The dimensionless times are 50, 250, and 1750 from left to right. Adapted from [41]

Epimerization

An interesting alternative to the Frank-type mechanism is a set of reactions based primarily on a phenomenon called epimerization, i. e., the spontaneous change of handedness in one part of the polymer. This mechanism is important in the chemistry of amino acids. Plasson et al. [31] identified four reactions: activation, polymerization, epimerization, and depolymerization as necessary ingredients that can, under certain conditions, lead to an instability of the racemic state with a bifurcation toward full homochirality. They called this the APED model, whose reactions can be summarized as follows:

A: activation:

$$L \xrightarrow{a} L^* , \quad D \xrightarrow{a} D^* , \tag{23}$$

P: polymerization:

$$L^* + L \xrightarrow{p} LL , \quad D^* + D \xrightarrow{p} DD , \tag{24}$$

$$L^* + D \xrightarrow{\alpha p} LD , \quad D^* + L \xrightarrow{\alpha p} DL , \tag{25}$$

E: epimerization:

$$LD \xrightarrow{e} DD , \quad DL \xrightarrow{e} LL , \tag{26}$$

D: depolymerization:

$$LL \xrightarrow{h} L + L, \quad DD \xrightarrow{h} D + D. \tag{27}$$

This minimal subset of reactions is shown in Fig. 6.

Compared with the Frank model, a major advantage of the APED model is that no hypothetical auto-catalysis is required. Indeed, all these reactions exist in principle, although it is as yet unclear what kind of manipulation of the environment is required to make all these reactions happen. Another advantage is that the system is closed, so no inflow or outflow of matter is required. The system is maintained away from equilibrium by energy input though the activation of amino acids.

Given that there is neither auto-catalysis nor enantiomeric cross-inhibition, one wonders whether the APED model still shares some similarities with Frank's original model. Some degree of similarity is immediately seen by writing the APED reactions in sequential form in one line, i. e.,

$$D^* + L \xrightarrow{\alpha p} DL \xrightarrow{e} LL \xrightarrow{h} L + L , \tag{28}$$

$$L^* + D \xrightarrow{\alpha p} LD \xrightarrow{e} DD \xrightarrow{h} D + D . \tag{29}$$

This shows that, as long as the reaction rates for epimerization and depolymerization are not limiting factors, we

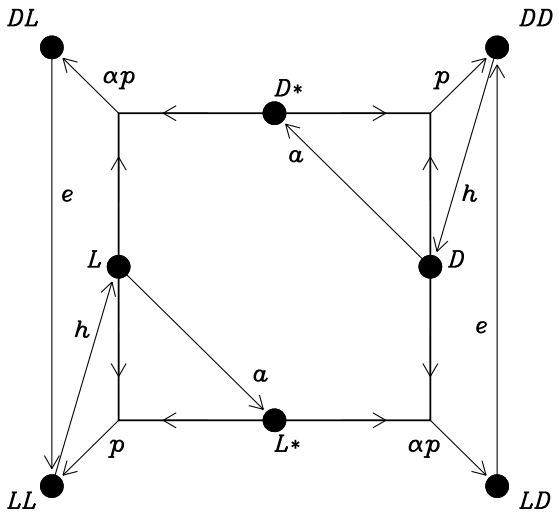

Exobiology (theoretical), Complexity in, Figure 6
Representation of the minimal set of reactions necessary for allowing a spontaneous transition to homochirality. Adapted from [6]

have essentially the reactions

$$D^* \xrightarrow{\alpha p L} L \quad \text{and} \quad L^* \xrightarrow{\alpha p D} D\,. \tag{30}$$

This way of writing these reactions emphasizes the roles of L and D in catalyzing the conversion of D^* into L and L^* into D, respectively. Just like the mechanism of mutual antagonism, these reactions disfavor a racemic state, but instead of producing unreactive waste, these reactions produce directly one of two possible homochiral states.

In addition, there are reactions of the form

$$L^* + L \xrightarrow{p} LL \xrightarrow{h} L + L\,, \tag{31}$$

$$D^* + D \xrightarrow{p} DD \xrightarrow{h} D + D\,. \tag{32}$$

These reactions simulate the autocatalytic conversion of L^* into L by L and of D^* into D by D. Again, linear analysis establishes that the racemic state is unstable provided α is in the range $0 < \alpha < 1$; see [6,31].

In conclusion we can say that the homochirality of life-bearing molecules might well have originated from the chemical reactions that led to their formation. Thus, the hypothetical RNA world may have been born into an environment surrounded by homochiral peptides (as described in Sect. "Epimerization"), or, alternatively, homochirality may have emerged as a consequence of enantiomeric cross-inhibition during the first stages of the RNA world (as discussed in Sect. "Continued Polymerization").

In the next section we discuss some issues regarding possible strategies for establishing a primitive information-carrying system. This is also based on catalysis, but catalysis in the production of molecules other than itself.

Establishing Hereditary Information

So far, we have ignored the fact that polymers can consist of different amino acid or nucleotide units, even though they would all have the same handedness. Therefore, such molecules could, in principle, carry information. Once such polymers can replicate, the question arises as to how to prevent them from becoming extinct due to errors in the copying process, and, instead, how to allow them to compete against parasites. It is generally believed that early self-replicating systems had a substantial error rate associated with each replication event. A certain error rate is obviously necessary for facilitating Darwinian evolution by natural selection, but it must be small enough to prevent extinction.

Assuming that with each generation a species produces σ offspring where the length of the genome is N bits, and that the probability for a copying error at any position in the genome is p, the necessary condition for long-term survival is given by [15,16]

$$pN < \ln \sigma\,. \tag{33}$$

The significance of this formula is illustrated in Fig. 7 with the help of a numerical example where the selective advan-

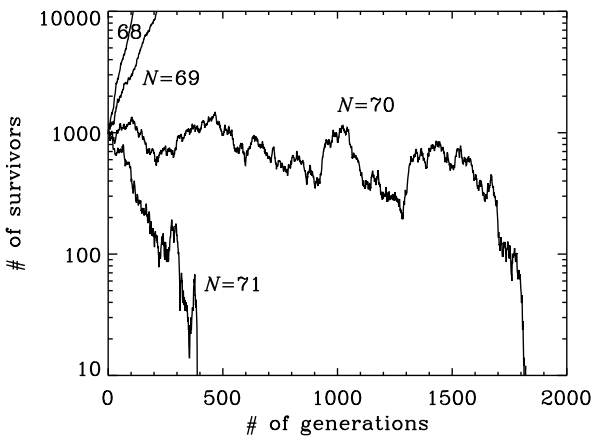

Exobiology (theoretical), Complexity in, Figure 7
Number of survivors as a function of the number of generations in a numerical example with an error rate of $p = 0.01$. This initial number of survivors is 1000. Note that for a genome length of $N = 70$ and 71, the population dies out after 1800 and 400 generations, respectively. For $N = 69$ and less the number of survivors increases exponentially. This is compatible with the survival criterion $N \lesssim (\ln 2)/0.01 = 69.3$, derived from Eq. (33)

tage (i. e., the multiplication factor) is chosen to be $\sigma = 2$, the error rate is $p = 0.01$, and four different values of N between 68 and 71 are used. In this numerical experiment, σ new offspring are produced, but with a probability p that an error is introduced at each of the N positions. Following an error, only the intact copies can produce further offspring.

According to Eq. (33) the maximum genome length is, with the parameters of our example, $(\ln 2)/0.01 = 69.3$. This is compatible with Fig. 7, which shows that the dividing line between extinction and long-term survival is between $N = 69$ and 70. For contemporary genomes N is of the order of 10^8, and p is of the order of 10^{-8} [14], or below, depending on the efficiency of error-correcting mechanisms available in contemporary organisms.

The first replicating systems are likely to have rather high error rates and no correction mechanism, making it virtually impossible to carry sufficient information for building more complex replicators. This difficulty can be removed by invoking the concept of hypercycles [17], whereby the full genetic information is carried collectively by several smaller systems (smaller N), each one small enough to obey Eq. (33). Mathematically, such a system can be described by the following set of reactions [2]:

$$I_i \xrightarrow{r_i} 2I_i \,, \tag{34}$$

$$I_i + I_{i-1} \xrightarrow{k_i} 2I_i + I_{i-1} \,. \tag{35}$$

Assuming, furthermore, that resources are limited, the total number of molecules, $M = \sum_i [I_i]$, is taken to be constant, i. e., I_i is assumed to be siphoned off from the system at a rate ϕ that is independent of i. Mathematically, such a system can be described by the following set of ordinary differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}[I_i] = r_i[I_i] + k_i[I_i][I_{i-1}] - \phi[I_i] \,, \tag{36}$$

where

$$\phi = \sum_i \Big(r_i[I_i] + k_i[I_i][I_{i-1}] \Big) \Big/ \sum_i [I_i] \tag{37}$$

is the factor that keeps the total number of molecules constant. The kinetic coefficient r_i models the residual effects of birth and death, while k_i is the kinetic coefficient for the catalytic production of I_i, where I_{i-1} acts as a catalyst. The evolution of number densities in a model of five hypercycles is shown in Fig. 8 for a case in which all $k_i = k$ and $r_i = r$ are chosen to be the same for all values of i.

An interesting situation arises when the effects of parasites are included. Boerlijst & Hogeweg [2] considered an example in which a parasite was coupled to I_2; see Fig. 9. The effect on the above model is shown in Fig. 10 where $k_{\text{para}} = 2k$ and $r_{\text{para}} = r$ have been chosen. One sees that not much happens for a long time when the parasite is turned on. This is because the parasite has to grow to a level at which it can affect the entire system. When this point is reached, all components of the system decay exponentially – including the parasite itself. Unfortunately, the system can never recover from this disaster, so the hypercycle theory seems to have a problem.

Again, spatial extent can significantly alter the situation. Using a cellular automata approach, Boerlijst & Hogeweg [2] and Boerlijst [3] showed that the nonlinear spatial dynamics of spiral waves gives the system stabil-

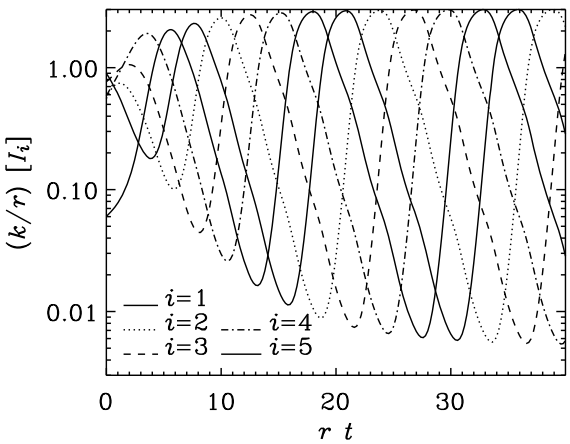

Exobiology (theoretical), Complexity in, Figure 8
Evolution of the number densities of five hypercycles with equal parameters. Note that peaks of I_1 (*solid line*) are followed by peaks of I_2 (*dotted line*) and I_3 (*dashed line*), and so forth. Time is measured in units of r^{-1} and concentrations are measured in units of r/k

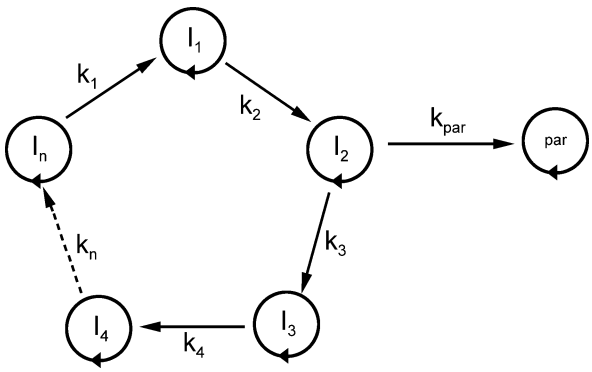

Exobiology (theoretical), Complexity in, Figure 9
Sketch showing the coupling of several hypercycles together with a parasite coupled to species I_2. Adapted from [2]

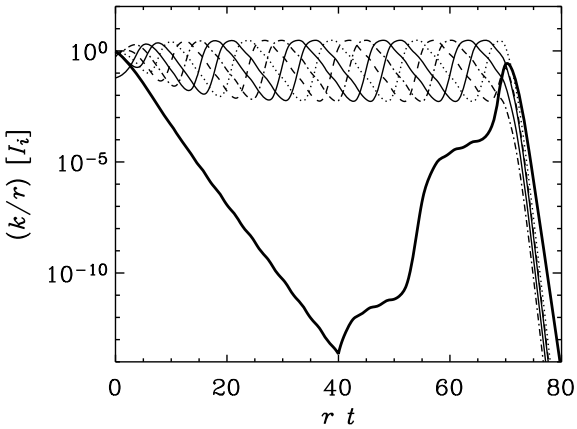

Exobiology (theoretical), Complexity in, Figure 10
Evolution of the number densities of five hypercycles with equal parameters and a parasite where $k_{para} = 2k_i$ and $r_{para} = r$. Time is measured in units of r^{-1} and concentrations are measured in units of r/k

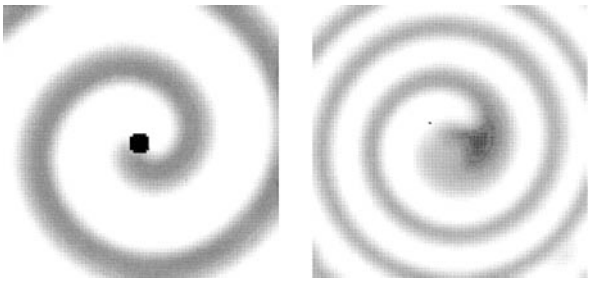

Exobiology (theoretical), Complexity in, Figure 11
Spatial patterns in the cellular automata approach of Boerlijst and Hogeweg [2]. Courtesy of Boerlijst MC

ity against otherwise deadly parasitic species. Interestingly enough, this approach tends to produce spiraling interfaces between different species; see Fig. 11

These equations are, by nature, similar to other chemical reaction-diffusion equations where several different substances catalyze each other's reactions. A particularly exciting example is the famous Belousov–Zhabotinsky reaction, where malonic acid, $CH_2(COOH)_2$ is oxidized in the presence of bromate ions, BrO_3^-. To initiate the reaction, cerium is used as a catalyst to donate ions, although other metal ions may also be used. The color depends on the state of the cerium as it changes from Ce^{3+} to Ce^{4+} or, if iron is used, from Fe^{2+} to Fe^{3+}. The resulting reactions are of the form [28]

$$A + Y \xrightarrow{k_1} X\,, \tag{38}$$

$$X + Y \xrightarrow{k_2} P\,, \tag{39}$$

$$B + X \xrightarrow{k_3} 2X + Z\,, \tag{40}$$

$$2X \xrightarrow{k_4} Q\,, \tag{41}$$

$$Z \xrightarrow{k_5} Y\,, \tag{42}$$

where $X = HBrO_2$, $Y = Br^-$, $Z = Ce^{4+}$, $A = B = BrO_3^-$, P and Q are reaction products that do not contribute further to the reactions, and $k_1, \ldots, k_5$ are known rate constants. The reactions above lead to kinetic equations of the form

$$\frac{\partial[X]}{\partial t} = k_1[A][Y] - k_2[X][Y] + k_3[A][X] - k_4[X]^2\,, \tag{43}$$

$$\frac{\partial[Y]}{\partial t} = -k_1[A][Y] - k_2[X][Y] + k_5[Z]\,, \tag{44}$$

$$\frac{\partial[Z]}{\partial t} = 2k_3[A][X] - k_5[Z]\,. \tag{45}$$

This model of reaction equations is called the Oregonator, which refers to the affiliation of the authors at the time of publication [18].

If spatial extent is included via diffusion terms, this reaction exhibits, in certain cases, spiral patterns similar to those in the model of Boerlijst & Hogeweg [2]. In Fig. 12 we reproduce the pattern obtained by Zhang et al. [48] for a slightly modified model consisting of only two partial differential equations. Depending on the value of a certain control parameter in their model, spiral patterns of different size are produced. An extensive review of the physics of pattern formation in different settings is given by Cross and Hohenberg [10].

The connection between pattern formation and the origin of life may seem rather remote. However, the equations governing chemical pattern formation illustrate some of the critical steps that are thought to have played a role in the origins of life. In particular, the fact that different chemical compounds catalyze each other in a productive manner is an essential property behind the model proposed by Eigen. The Belousov–Zhabotinsky reaction also illustrates the phenomenon of auto-catalysis where, in the presence of A, the molecule X catalyzes the production of more X by using A as a substrate and producing Z as an additional side product.

The possibility of self-replication was demonstrated for simple RNA molecules by Spiegelman [27] back in the late 1960s. Now, there are examples of simple peptide chains that can catalyze the production of each other [38]. However, a serious shortcoming of any of the above examples is the fact that there is no possibility of natural selec-

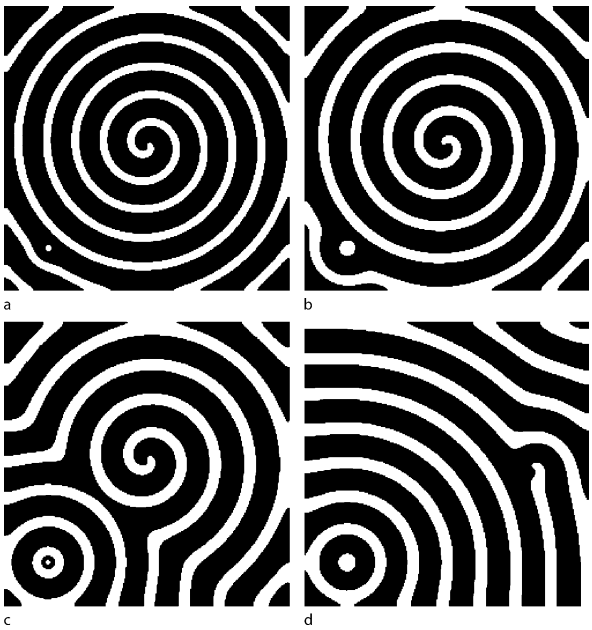

Exobiology (theoretical), Complexity in, Figure 12
Spiral and ring-like patterns for the modified reaction equations by Zhang et al. [48]. On the boundaries a no-flux condition has been adopted, i. e., the normal components of all gradients vanish. Adapted from [48]

tion and hence Darwinian evolution. So, as far as the question of the origin of life is concerned, this pathway must be considered a dead end.

In summary, one can say that there are similarities in the mathematics of producing homochirality and in establishing hereditary information in the composition of the first replicating polymers. However, in the latter case, even less is known about the detailed nature of such polymers and their catalytic properties. A particularly important aspect is the possibility of spatial extent, which can substantially modify the behavior of any chemical system. In the present case, as shown in Ref. [2], the possibility of spatial extent is critical for stabilizing the system against destruction by parasites. The model also exhibits spiral pattern formation that has been at the heart of early work by Prigogine and others in connection with early ideas on biogenesis.

Alteration of the Environment by Early Life

In this last section, we discuss some physics problems within astrobiology that illustrate how life, once it has formed, might affect the environment of the early Earth and how it led to a planet so markedly different from a planet that does not harbour life.

Global Energy Balance of the Earth

The young Sun was about 30% fainter than today, and yet the young Earth was covered with liquid water and had temperatures higher than nowadays. This was caused by the presence of greenhouse gases such as water vapor, carbon dioxide, and probably methane. Life is responsible for reducing CO_2 to compounds of the form $(CH_2O)_n$ and similar, and for oxidizing various minerals to produce O_2. The resulting decrease of CO_2 weakens the greenhouse effect, so in this sense the emergence of life has an essentially cooling effect on the Earth's overall climate.

Without atmosphere, the planet would cool like a black body at a rate proportional to the local flux $\sigma_{SB} T^4$, where σ_{SB} is the Stefan–Boltzmann constant. Integrated over the entire surface of the planet, this corresponds to a loss of $4\pi R_E^2 \sigma_{SB} T^4$, which would need to be balanced against the rate of energy received by solar radiation. The solar "constant" is $S = 1.37\,\mathrm{kW\,m^{-2}}$ and the total energy projected onto the disk of the Earth is $(1-A)\pi R_E^2 S$, where A is the albedo, i. e., the fraction of energy reflected from the Earth. The resulting blackbody temperature would be

$$T_0 = \left[(1-A)\frac{S}{4\sigma_{SB}}\right]^{1/4} . \tag{46}$$

Using $A = 0.3$ and $\sigma_{SB} = 5.67 \times 10^{-8}\,\mathrm{W\,m^{-2}\,K^{-1}}$), the temperature of the Earth would be 255 K or about −18°C.

In the presence of an atmosphere, the rate of cooling is modified to $\sigma_{SB} T_{eff}^4$, where T_{eff} is the effective temperature equivalent to that of a black body. A positive greenhouse effect corresponds to $T_{eff} < T$, so the cooling is reduced and the atmosphere heats up according to the vertically integrated energy equation

$$C\frac{dT_0}{dt} = (1-A)\frac{S}{4} - \sigma_{SB} T_{eff}^4 , \tag{47}$$

where C is the vertically integrated specific heat.

The value of the effective temperature can be obtained from a radiative transfer calculation. A simplified model calculation[1] yields

$$T_{eff}^4 = \frac{\ell}{\ell_{crit}} T_0^4 , \tag{48}$$

[1]Under the assumption of local isotropy (Eddington approximation) radiative equilibrium implies that the flux is proportional to the negative gradient of the radiative energy density aT^4, where a is the radiation-density constant, so

$$\boldsymbol{F} = -\frac{1}{3} c\ell \nabla (aT^4) .$$

Here, c is the speed of light and ℓ is the mean free path of photons. The latter can be expressed in terms of the opacity κ and the density ρ

where T_0 is the surface temperature, ℓ is an averaged mean free path of photons and $\ell_{\rm crit}$ is the critical value above which there is a positive greenhouse effect. Again, a simplified calculation suggests $\ell_{\rm crit} = 3H/16 \approx 0.19H$, where $H = \mathcal{R}T/(\mu g) \approx 8\,\mathrm{km}$ is the pressure scale height of the atmosphere. So, an increase in opacity leads to a decrease of the cooling and hence to an increase in the surface temperature.

Another interpretation is that the greenhouse gases shift the radiating surface by a certain amount, $\ell_{\rm g}$, upward. The value of $\ell_{\rm g}$ is related to ℓ. Ditlevsen [13] uses $\ell_{\rm g} = 3\,\mathrm{km}$. He also noted that a more accurate lapse rate of the temperature is $\mathrm{d}T/\mathrm{d}z = 10\,\mathrm{K\,km^{-1}}$ instead of $T_0/H = 40\,\mathrm{K\,km^{-1}}$, so that the temperature gain caused by greenhouse gases is $\ell_{\rm g} \times \mathrm{d}T/\mathrm{d}z = 30\,\mathrm{K}$. The reason for a shallower temperature gradient is the presence of convection that causes the specific entropy s to be nearly constant with height. In that case the temperature gradient is just the adiabatic one, $(\mathrm{d}T/\mathrm{d}z)_{\rm ad} = g/c_{\rm p}$, where c_p is the specific heat at fixed pressure, which in turn is related to the universal gas constant and the specific weight via $\mathcal{R}/\mu = c_{\rm p} - c_{\rm v}$, where c_v is the specific heat at fixed volume and $c_{\rm p}/c_{\rm v} = \gamma$ is the ratio of specific heats. With these formulae one does indeed get [2]

$$\left(\frac{\mathrm{d}T}{\mathrm{d}z}\right)_{\rm ad} = \left(1 - \frac{1}{\gamma}\right)\frac{\mu g}{\mathcal{R}} \approx 10\,\mathrm{K\,km^{-1}}\,, \qquad (49)$$

where we have used $\gamma = 7/5$ for air molecules with 5 degrees of freedom (3 for translation and 2 for rotation).

At certain times over the history of the Earth, other greenhouse gases such as methane may have played an important role in keeping the Earth above freezing temperatures. Indeed, the burial of oxides in the crust allowed methane to build up in the atmosphere, which may have led to concentrations of a few thousand times greater than modern levels. UV radiation in the upper atmosphere breaks up methane into its components, letting H_2 escape into space, leading to a net gain of oxygen, that comes ultimately from H_2O.

According to a model of Catling et al. [8] methane (CH_4) may have been important 2.7 . . . 2.3 Gyr ago, just before the famous Snowball Earth deep freeze of the planet [22]. As discussed above, the associated loss of hydrogen may have led to a gradual accumulation of oxygen in the atmosphere, which then terminated the methane era and led to the Snowball Earth event. This event lasted until the continuous CO_2 production from volcanoes accumulated to large amounts so that the resulting greenhouse effect became sufficient to initiate partial melting of the ice cover.

Response to Changes in Greenhouse Gases

As we knew first from global climate models [20] and later from simplified models [9] using Eq. (47), with a relatively simple piecewise linear temperature dependence of $A(T)$, there can be three different equilibrium temperatures. This is illustrated in Fig. 13, where we compare the graph of $\sigma_{\rm SB}T_{\rm eff}^4$ versus surface temperature T_0 with the net radiation $(1-A)S/4$. Here, $A(T)$ has been arranged such that $A = A_{\rm hot}$ for $T \geq T_{\rm max}$ (corresponding to no ice coverage) $A = A_{\rm cold}$ for $T \leq T_{\rm min}$ (corresponding to full ice coverage).

Ditlevsen [13] used Eq. (47) to study the response of the system to variable greenhouse gas concentrations. As the amount of CO_2 increases, the equilibrium temperature increases. Obviously, when the system is on the lower fixed point initially, there must be a critical CO_2 concentration above which the solution will jump discontinuously to the upper branch; see Fig. 14.

It is generally accepted that the rate of weathering increases with increasing temperature. This provides a stabilizing effect on the climate. As T increases, the rate of weathering increases, removing more CO_2 from the atmosphere, reducing the greenhouse effect, and thus leading to cooling. Ditlevsen [13] introduced the assumption that there is a continuous source of CO_2 through outgassing from volcanoes and a temperature-dependent sink of CO_2 from weathering when T exceeds a critical temperature $T_{\rm w}$, but no weathering for $T < T_{\rm w}$ [46]. This leads to a self-regulating effect for $T < T_{\rm w}$, which Ditlevsen calls a greenhouse thermostat. Whenever $T < T_{\rm w}$, since there is then no weathering and hence no sink of CO_2, greenhouse gases will build up until the Earth's temperature has reached the value $T_{\rm w}$; see Fig. 15. This is the mechanism that is believed to have caused the early Earth to be above freezing through most of its history – with the

as $\ell = (\kappa\rho)^{-1}$. Hydrostatic equilibrium can be written in the form

$$\boldsymbol{g} = -\frac{\mathcal{R}}{\mu}\nabla T\,,$$

where $\mathbf{g}$ is the gravitational acceleration, $\mathcal{R}$ is the universal gas constant, and μ is the mean molecular weight. These equations can be solved by a polytrope, i. e., $T = T_0(1 - z/H)$ and $\rho = \rho_0(1 - z/H)^3$, where z is the distance from the surface and H is the vertical pressure scale height. This leads to a condition of the form Eq. (48) where $\ell_{\rm crit} = 3H/16 \approx 0.19H$ is the critical mean free path of photons

[2] Hydrostatic equilibrium can be written as $-\rho^{-1}\nabla p - \nabla\phi = 0$, where p is the pressure and $\phi = gz + \text{const}$ is the gravitational potential. Using the thermodynamic relation $-\rho^{-1}\nabla p = -\nabla h + T\nabla s = 0$, where $h = c_{\rm p}T$ is the specific enthalpy and $\nabla s = 0$ for adiabatic stratification, we have $\mathrm{d}(c_{\rm p}T)/\mathrm{d}z = g$

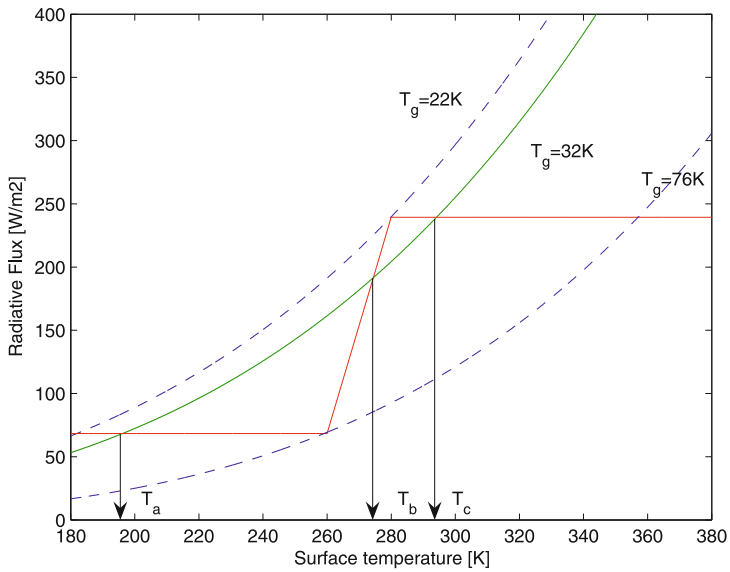

Exobiology (theoretical), Complexity in, Figure 13
Plot of $\sigma_{SB} T_{eff}^4$ versus surface temperature T_0 for three different greenhouse temperature shifts, T_g, compared with the net radiation $\frac{1}{4}(1 - A)$ for a simple piecewise linear function $A(T)$. Courtesy of P. Ditlevsen [13]

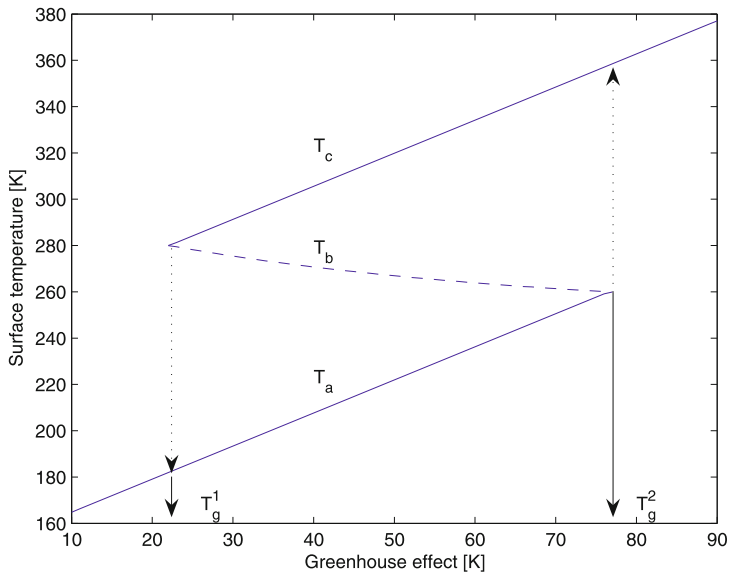

Exobiology (theoretical), Complexity in, Figure 14
Equilibrium temperature as a function of CO_2 concentration. Courtesy of P. Ditlevsen [13]

exception of intermediate Snowball Earth-like events that are caused by the emergence of other sinks of greenhouse gases, such as the onset of aerobic photosynthesis or the enhanced formation of mountain topography that leads to an increase in the erosion rate and, hence, weathering.

The Daisyworld Model

Life can also affect the planet's albedo, as has been demonstrated by Lovelock [47] in his Daisyworld model. For a tutorial on the Daisyworld model see [45]. This model also makes use of Eq. (47), but now the planet's albedo A is af-

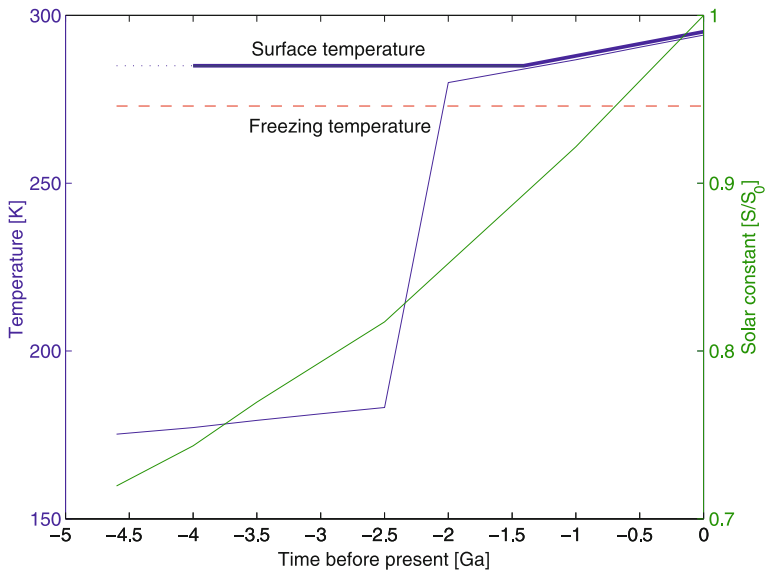

Exobiology (theoretical), Complexity in, Figure 15
Dependence of surface temperature on time under the assumption of a continuous source of outgassing of CO_2 and the onset of a CO_2 sink for $T > T_w$. Courtesy of P. Ditlevsen [13]

fected by the plant population which is simplistically represented by black and white plants or flowers (daisies) with local albedos A_1 and A_2, respectively. So the total albedo is a weighted average of the form

$$A = \sum_{i=1}^{3} \alpha_i A_i \,, \tag{50}$$

where A_3 is the albedo of the unpopulated surface. The weights α_i depend on the surface coverage of the respective regions and obey evolution equations that are, in turn, governed by a temperature-dependent growth term, $\beta(T_i)$, and a fixed death rate, γ. The resulting equations for the rate of change of the albedo are

$$\frac{\mathrm{d}\alpha_i}{\mathrm{d}t} = [\beta(T_i) - \gamma]\alpha_i \,, \tag{51}$$

where $i = 1$ for black and $i = 2$ for white plant populations, $\beta(T)$ is assumed to be different from zero in the range $T_{\min} < T < T_{\max}$ with a maximum at $T_{\mathrm{aver}} = \frac{1}{2}(T_{\min} + T_{\max})$. The weight for the unpopulated surface follows from the normalization $\sum \alpha_i = 1$, so $\alpha_3 = 1 - \alpha_1 - \alpha_2$.

The temperatures are higher in the regions of black plants and lower in regions of white plants according to the formula

$$T_i^4 = (A - A_i)q + T_0^4 \,, \tag{52}$$

where q is a parameter that must be smaller than a critical value,

$$q < q_{\mathrm{crit}} = S/(4\sigma_{\mathrm{SB}}) \,, \tag{53}$$

in order that heat flows against the temperature gradient [40]. For $q = q_{\mathrm{crit}}$ the temperature is uniform for different values of A_i, while for $q < q_{\mathrm{crit}}$ the regions of high albedo are cooler and those of low albedo warmer. Note also that Eq. (52) preserves heat balance, i. e., $\sum \alpha_i T_i^4 = T_0^4$.

The important point in the Daisyworld model is the fact that, for a certain range of S, the surface temperature of the planet, T_0, is stabilized in a certain temperature range around the optimal value close to T_{aver}; see Fig. 16.

Saunders makes another remarkable point. He showed that by changing the model to allow for Darwinian evolution such that each plant species works with an optimized temperature dependence, so $\beta(T_i) \to \beta_i(T_i)$ is modified to become dependent on i, the overall result changes only very little. More importantly, the range over which the model can stabilize the planet's temperature shrinks, making the planet as a whole more vulnerable. Although the amount of shrinkage is small, it emphasizes the dangers associated with adopting changes that lead only to short-term benefits. Saunders emphasizes in his work that the ability of life to regulate the surface temperature of a planet is not associated with natural selection as in the concept of Darwinian evolution. More generally he warns, therefore,

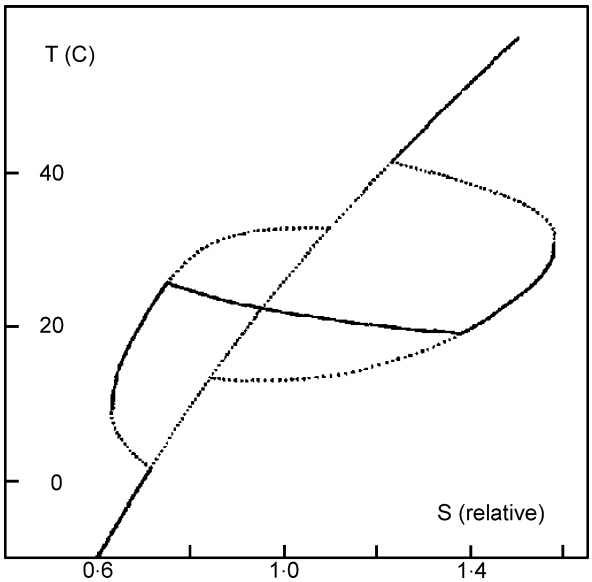

Exobiology (theoretical), Complexity in, Figure 16
Temperature (in Celsius) versus relative irradiation (normalized to the average temperature). Note that the temperature is stabilized around the value T_{aver}, provided the energy input S is within a certain range. Adapted from [40]

that not everything that is to an advantage needs to be the result of natural selection [40].

Oxidation of the Earth's Crust

It has recently been proposed that, in addition to the effects discussed above, life may also have profound effects on the Earth's crust. A possible scenario has recently been discussed by Rosing et al. [36]. The idea is that photosynthetic life may tap large amounts of solar energy that were used to reduce carbon from CO_2 to compounds of the form $(CH_2O)_n$ and similar, via reactions of the form

$$CO_2 + H_2O + h\nu \rightarrow CH_2O + O_2\,, \tag{54}$$

where $h\nu$ denotes the energy taken from solar radiation. Furthermore, and even more surprisingly, the oxygen produced by photosynthesis may have been critical in oxidizing iron in the continental crust. Although other factors also played a substantial role, it is clear that biological processes can speed up the oxidation process substantially. Comparing the oceanic crust with the continental crust, a major difference is the enhanced fraction of SiO_2 (57% in the continental crust compared to 50% in the oceanic crust).

Granite, being one of the lightest rock types, was eventually able to escape subduction and to produce stable continents about 3.8 Gyr ago. This is also the time of the oldest rock findings on Earth. Given that the rise of continents on the early Earth is associated with granite formation, the presence of granite on silicon-bearing rocky planets might thus be a possible biomarker for photosynthesis [36].

Although this idea is speculative, it may be supported quantitatively as follows. Firstly, the present day production rate of organically fixated carbon is estimated to be 9×10^{15} mol C yr^{-1} [12,25]. The amount of energy required for this can be calculated by using the fact that it costs 477 kJ to transfer one mole of carbon to hexose. The energy required for this is then 300 mW m^{-2}. Rosing et al. [36] argue that this amount could be supplied by only 0.1% of the effective solar energy flux, $S/4$. Assuming that the amount of carbon burial, relevant to estimating the usable fraction of oxygen for iron oxidation, is also about 0.1%, this corresponds to about 10^{13} mol C yr^{-1}. This would yield a comparable iron oxidation rate. Rosing et al. [36] argue further that annual basalt production contributes about 10^{14}mol Fe yr^{-1}, so a fraction of the magmatic iron flux could be used for building up the mantle reservoir of ferric iron.

In conclusion, the presence of life can lead to significant alterations of the planet in a number of different ways, as is quite clearly demonstrated by some of the differences between Earth and its neighboring planets Venus and Mars. Only the Earth has extensive reservoirs of oxygen and of granite. Within limits, the presence of life on a planet can also have a stabilizing effect on its climate. The relevant mathematical modeling of some of these processes resembles, in many ways, those encountered earlier in studies of homochirality and of the spread of hereditary information on the early Earth.

Conclusions

Astrobiology has developed into a rapidly growing research field involving expertise from a number of neighboring disciplines. Nonlinear dynamics and nonequilibrium thermodynamics find applications in all these subfields. Here, we have elaborated on a few such aspects. Closest to the onset of life is, perhaps, the emergence of homochirality of biomolecules. Given that RNA has been proven to form longer polymers only in a homochiral environment, one would expect that homochirality must be a prerequisite to the emergence of life at the level of a replicating RNA world. On the other hand, the very mechanism causing the polymerization to terminate, namely enantiomeric cross-inhibition, can also be the mechanism responsible for causing 100% homochirality by destroying RNA molecules whose chirality is already in the minority. This would, however, require the possibility of auto-

catalysis, which can be avoided in another scenario where a closed peptide system is kept away from equilibrium by continuous activation of amino acids.

Chemically speaking, the stabilization of a definite chirality is in some models similar to the subsequent establishment of hereditary information, in that catalysis plays a crucial role. Furthermore, in both cases, the possibility of chemistry in an extended system is crucial. On the one hand, spatial extent gives rise to the possibility of coexistence of life forms of opposite handedness on the early Earth. On the other hand, spatial extent can be critical in allowing the system to find unpopulated locations fast enough to avoid being overwhelmed by the effects of parasites that tap the same resources that are required for the maintenance and development of hereditary information.

Finally, life is invariably coupled to some kind of metabolism that is ultimately powered by solar energy. This clearly affects the environment by reducing carbon and oxidizing the crust of the Earth and, over the last two billion years, the atmosphere. How much these alterations of the environment are due to biological processes is less obvious. However, it is clear that biological factors greatly speed up weathering on the Earth. The extent of biologically induced alterations of the continental crust, for example, may therefore best be tested using quantitatively accurate model calculations. The outcome may ultimately hinge on energetic considerations and on the efficiency of photosynthesis as a solar energy collector.

With the scope of being able to explore, in the near future, not only the planets and other celestial bodies in the solar system in much more detail, but also planets of other planetary systems, research in astrobiology quickly develops into a field that will be driven more and more by new data, making this field less susceptible to speculation. Therefore, it is important to be prepared for upcoming discoveries in this field. Finally, it should be emphasized that astrobiology is efficient in communicating science to the general public, which may provide an additional boost to the field.

Bibliography

Primary Literature

1. Blackmond DG (2004) Asymmetric auto-catalysis and its implications for the origin of homochirality. Proc Natl Acad Sci 101:5732–5736
2. Boerlijst MC, Hogeweg P (1991) Spiral wave structure in prebiotic evolution – hypercycles stable against parasites. Phys D 48:17–28
3. Boerlijst MC, (2000) Spiral ans spots: Novel evolution phenomena through spatial self-structing. In: Dieckmann U, Law R, Metz H (eds) The geometry of ecological interactions: Simplifying spatial complexity. Cambridge University Press, Campridge, pp 171–182
4. Brandenburg A, Multamäki T (2004) How long can left and right handed life forms coexist? Int J Astrobiol 3:209–219
5. Brandenburg A, Andersen AC, Höfner S, Nilsson M (2005) Homochiral growth through enantiomeric cross-inhibition. Orig Life Biosph 35:225–241
6. Brandenburg A, Lehto HJ, Lehto KM (2007) Homochirality in an early peptide world. Astrobiol 7:725–732
7. Bywater RP, Conde-Frieboesk K (2005) Did Life Begin on the Beach? Astrobiol 5:568–574
8. Catling DD, Zahnle KJ, McKay CP (2001) Biogenic methane, hydrogen escape, and the irreversible oxidation of early Earth. Science 293:839–843
9. Crafoord C, Källén E (1978) A note on the condition for existence of more than one steady-state solution in Budyko-Sellers type models. J Atmos Sci 35:1123–1125
10. Cross MC, Hohenberg PC (1993) Pattern formation outside of equilibrium. Rev Mod Phys 65:851–1112
11. Davies PCW, Lineweaver CH (2005) Finding a second sample of life on Earth. Astrobiol 5:154–163
12. Des Marais DJ (2000) When did photosynthesis emerge on Earth. Science 289:1703–1705
13. Ditlevsen PD (2005) A climatic thermostat making Earth habitable. Int J Astrobiol 4:3–7
14. Dyson FJ (1999) Origins of life. Cambridge University Press, Cambridge
15. Eigen M (1971) Selforganization of matter and evolution of biological macromolecules. Naturwissenschaften 58:465–523
16. Eigen M (2002) Error catastrophe and antiviral strategy. Proc Natl Acad Sci 99:13374–13376
17. Eigen M, Schuster P (1977) The hypercycle. Naturwissenschaften 64:541–565
18. Field RJ, Körö E, Noyes RM (1972) Oscillations in chemical systems, part 2. Thorough analysis of temporal oscillations in the bromate-cerium-malonic acid system. J Am Chem Soc 94:8649–8664
19. Frank FC (1953) On spontaneous asymmetric synthesis. Biochim Biophys Acta 11:459–464
20. Ghil M (1976) Climate stability for a Sellers-type model. J Theor Biol 33:3–20
21. Gilbert W (1986) Origin of life – the RNA world. Nature 319:618–618
22. Hoffman PF, Kaufman AJ, Halverson GP, Schrag DP (1998) A Neoproterozoic Snowball Earth. Science 281:1342–1346
23. Joyce GF, Visser GM, van Boeckel CAA, van Boom JH, Orgel LE, Westrenen J (1984) Chiral selection in poly(C)-directed synthesis of oligo(G). Nature 310:602–603
24. Lathe R (2004) Fast tidal cycling and the origin of life. Icarus 168:18–22
25. Martin JH, Knauer GA, Karl DM, Broenkow WW (1987) Vertex - carbon cycling in the Northeast Pacific. Deep-Sea Res Part A 34:267–285
26. Miller SL (1953) A production of amino acids under possible primitive Earth conditions. Science 117:528–529
27. Mills DR, Peterson RL, Spiegelman S (1967) An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. Proc Natl Acad Sci 58:217–224
28. Murray JD (1974) On a model for the temporal oscillations in the Belousov–Zhabotinsky reaction. J Chem Phys 61:3610–3613

29. Murray JD (2002) Mathematical Biology, An introduction. Springer, New York
30. Nielsen PE (1993) Peptide nucleic acid (PNA): A model structure for the primordial genetic material. Orig Life Evol Biosph 23:323–327
31. Plasson R, Bersini H, Commeyras A (2004) Recycling Frank: spontaneous emergence of homochirality in noncatalytic systems. Phys Rev 101:16733–16738
32. Plasson R, Kondepudi DK, Bersini H, Commeyras A, Asakura K (2007) Emergence of homochirality in far-from-equilibrium systems: Mechanisms and role in prebiotic chemistry. Chirality 19:589–600
33. Prigogine I, Lefever R (1968) Symmetry breaking instabilities in dissipative systems II. J Chem Phys 48:1695–1700
34. Prigogine I, Nicolis G (1967) On symmetry-breaking instabilities in dissipative systems. Chem J Phys 46:3542–3550
35. Rasmussen S, Chen L, Nilsson M, Abe S (2003) Bridging nonliving and living matter. Artif Life 9:269–316
36. Rosing MT, Bird DK, Sleep NH, Glassley W, Albarede F (2006) The rise of continents–An essay on the geologic consequences of photosynthesis. Palaeogeogr Palaeoclimatol Palaeoecol 232:99–113
37. Russell M (2006) First life. Am Sci 94:32–39
38. Saghatelian A, Yokobayashi Y, Soltani K, Ghadiri MR (2001) A chiroselective peptide replicator. Nature 409:797–801
39. Sandars PGH (2003) A toy model for the generation of homochirality during polymerization. Orig Life Biosph 33:575–587
40. Saunders PT (1994) Evolution without natural selection: further implications of the Daisyworld parable. J Theor Biol 166:365–373
41. Shibata R, Saito Y, Hyuga H (2006) Diffusion accelerates and enhances chirality selection. Phys Rev 74:026117
42. Soai K, Shibata T, Morioka H, Choji K (1995) Asymmetric autocatalysis and amplification of enantiomeric excess of a chiral molecule. Nature 378:767–768
43. Taylor FW (1991) The greenhouse effect and climate change. Rep Prog Phys 54:881–918
44. Turing AM (1952) The chemical basis of morphogenesis. Phil Trans Roy Soc B237:37–72
45. von Bloh W, Block A, Parade M, Schellnhuber HJ (1999) Tutorial Modelling of geosphere-biosphere interactions: the effect of percolation-type habitat fragmentation. Rep Prog Phys A 266:186–106
46. Walker JCG, Hays PB, Kasting JF (1981) A negative feedback mechanism for the long-term stabilization of the Earth's surface temperature. J Geophys Res 86:9776–9782
47. Watson AJ, Lovelock JE (1983) Biological homeostasis of the global environment: the parable of Daisyworld. Tellus B 35:284–289
48. Zhang XZ, Liao HM, Zhou LQ, Quyang Q (2004) Pattern selection in the Belousov–Zhabotinsky reaction with the addition of an activating reactant. J Phys Chem B 108:16990–16994

Books and Reviews

Barbieri M (2003) The organic codes: an introduction to semantic biology. CUP, Cambridge
Brack A (1998) The molecular origins of life. Cambridge University Press, Cambridge
Darwin C (1859) The origin of species by means of natural section. Reprinted by: Penguin books, London, 1985
Haken H (1983) Synergetics – An Introduction. Springer, Berlin
Lovelock JE (1995) Gaia. A new look at life on Earth. Oxford University Press, Oxford
Lunine J (2003) Astrobiology: a multi-disciplinary approach. Pearson Higher Education, Addison-Wesley, San Francisco
Prigogine I (1980) From being to becoming. Time and complexity in the physical sciences. Freeman, New York
Rauchfuß H (2005) Chemische Evolution und der Ursprung des Lebens. Springer, Berlin
Ward PD, Brownlee D (2000) Rare Earth: why complex life is uncommon in the Universe? Copernicus, New York

Extreme Events in Socio-economic and Political Complex Systems, Predictability of

Vladimir Keilis-Borok[1,2], Alexandre Soloviev[2,3], Allan Lichtman[4]
[1] Institute of Geophysics and Planetary Physics and Department of Earth and Space Sciences, University of California, Los Angeles, USA
[2] International Institute of Earthquake Prediction Theory and Mathematical Geophysics, Russian Academy of Science, Moscow, Russia
[3] Abdus Salam International Centre for Theoretical Physics, Trieste, Italy
[4] American University, Washington D.C., USA

Article Outline

Glossary
Definition of the Subject
Introduction
Common Elements of Data Analyzes
Elections
US Economic Recessions
Unemployment
Homicide Surges
Summary: Findings and Emerging Possibilities
Bibliography

Glossary

Complexity A definitive feature of nonlinear systems of interacting elements. It comprises high instability with respect to initial and boundary conditions, and complex but non-random behavior patterns ("order in chaos").

Extreme events Rare events having a large impact. Such events are also known as critical phenomena, disas-

ters, catastrophes, and crises. They persistently reoccur in hierarchical complex systems created, separately or jointly, by nature and society.

Fast acceleration of unemployment (FAU) The start of a strong and lasting increase of the unemployment rate.

Pattern recognition of rare events The methodology of artificial intelligence' kind aimed at studying distinctive features of complex phenomena, in particular – at formulating and testing hypotheses on these features.

Premonitory patterns Patterns of a complex system's behavior that emerge most frequently as an extreme event approaches.

Recession The American National Bureau of Economic Research defines recession as "a significant decline in economic activity spread across the economy, lasting more than a few months". A recession may involve simultaneous decline in coincident measures of overall economic activity such as industrial production, employment, investment, and corporate profits.

Start of the homicide surge (SHS) The start of a strong and lasting increase in the smoothed homicide rate.

Definition of the Subject

At stake in the development of accurate and reliable methods of prediction for social systems is the capacity of scientific reason to improve the human condition. Today's civilization is highly vulnerable to crises arising from extreme events generated by complex and poorly understood systems. Examples include external and civil wars, terrorist attacks, crime waves, economic downturns, and famines, to name just a few. Yet more subtle effects threaten modern society, such as the inability of democratic systems to produce policies responsive to challenges like climate change, global poverty, and resource depletion.

Our capacity to predict the course of events in complex social systems is inherently limited. However, there is a new and promising approach to predicting and understanding complex systems that has emerged through the integration of studies in the social sciences and the mathematics of prediction. This entry describes and analyzes that approach and its real-world applications. These include algorithmic prediction of electoral fortunes of incumbent parties, economic recessions, surges of unemployment, and outbursts of crimes. This leads to important inferences for averting and responding to impending crises and for improving the functioning of modern democratic societies.

That approach was successfully applied also to natural disasters such as earthquakes. Ultimately, improved prediction methods enhance our capacity for understanding the world and for protecting and sustaining our civilization.

Extreme events. Hierarchical complex systems persistently generate extreme events – the rare fast changes that have a strong impact on the system. Depending on connotation they are also known as critical phenomena, disasters, catastrophes, and crises. This article examines the development and application of the algorithmic prediction of extreme socio-economic and political events.

The prediction problem is formulated as follows:

given are time series that describe dynamics of the system up to the current moment of time t and contain potential precursors of an extreme event;

to predict whether an extreme event will or will not occur during the subsequent time period $(t, t + \tau)$; if the answer is "yes", this will be the "*period of alarm*".

As the time goes by, predictions form a discrete sequence of alarms. The possible outcomes of such a prediction are shown in Fig. 1. The actual outcome is determined unambiguously, since the extreme events are identified independently of the prediction either by the actual happening (e. g. by an election result) or by a separate algorithm (e. g. homicide surge) after they occur.

Such "yes or no" prediction is aimed not at analyzing the whole dynamics of the system, but only at identifying the occurrence of rare extreme events. In a broad field of prediction studies this prediction is different from and complementary to the classical Kolmogoroff–Wiener prediction of continuous functions, and to traditional cause-and-effect analysis.

The problem includes estimating the predictions' accuracy: the rates of false alarms and failures to predict, and the total duration of alarms in relation to the total time considered. These characteristics represent the inevitable

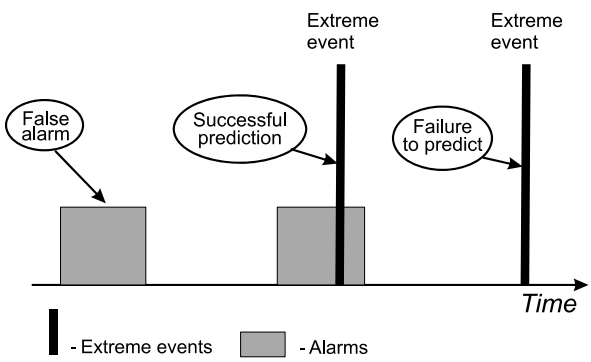

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 1
Possible outcomes of prediction

probabilistic component of prediction; they provide for statistical validation of a prediction algorithm and for optimizing preparedness to predicted events (e. g. recessions or crime surges).

Twofold importance. The prediction problem is pivotal in two areas:

- *Fundamental understanding of complex systems*. Prediction algorithms quantitatively define phenomena that anticipate extreme events. Such quantitative definition is pivotal for fundamental understanding of a complex system where these events occur, including the intertwined mechanisms of system's development and its basic features, e. g. multiple scaling, correlation range, clustering, fragmentation etc. (see Sects. "Common Elements of Data Analyzes", "Elections", "US Economic Recessions", "Unemployment"). The understanding of complex systems remains a major unsolved problem of modern science, tantamount to transforming our understanding of the natural and human world.
- *Disaster preparedness*. On the practical side prediction is pivotal for coping with a variety of disasters, commonly recognized as major threats to the survival and sustainability of our civilization (e. g. [22]; see also materials of G8-UNESCO World Forum on "Education, Innovation and Research: New Partnership for Sustainable Development", http://g8forum.ictp.it). The reliable advance prediction of extreme events can save lives, contribute to social and economic stability, and to improving the governing of modern societies.

Introduction

Predictability vs. Complexity: The Need for Holistic Approach [7,12,13,15,17,27,32]

Natural science had for many centuries regarded the Universe as a completely predictable machine. As Pierre Simon de Laplace wrote in 1776, "… if we knew exactly the laws of nature and the situation of the universe at the initial moment, we could predict exactly the situation of the same universe at a succeeding moment." However, at the turn of the 20th century (1905) Jules Henry Poincare discovered, that "… this is not always so. It may happen that small differences in the initial conditions will produce very great ones in the final phenomena. Prediction becomes impossible".

This instability to initial conditions is indeed a definitive attribute of complex systems. Nonetheless, through the robust integral description of such systems, it is possible to discover regular behavior patterns that transcend the inherent complexity. For that reason studying complexity requires the holistic approach that proceeds from the whole to details, as opposed to the reductionism approach that proceeds from details to the whole. It is in principle not possible "to understand a complex system by breaking it apart" [13].

Among the regular behavior patterns of complex systems are "premonitory" ones that emerge more frequently as an extreme event approaches. These premonitory patterns make complex systems predictable. The accuracy of predictions, however, is inevitably limited due to the systems' complexity and observational errors.

Premonitory patterns and extreme events are consecutive manifestations of a system's dynamics. These patterns may not trigger extreme events but merely signal the growth of instability, making the system ripe for the emergence of extreme events.

Methodology

The prediction algorithms described here are based on discovering premonitory patterns. The development of the algorithms requires the integration of complementary methods:

- Theoretical and numerical modeling of complex systems; this includes "universal"models considered in statistical physics and non-linear dynamics (e. g. [1,3,5,8,12,15,20,42]), and system-specific models, if available.
- Exploratory data analysis.
- Statistical analysis of limited samples, which is relevant since the prediction targets are by definition rare.
- Practical expertise, even if it is intuitive.
- Risk analysis and theory of optimal control for optimizing prediction strategy along with disaster preparedness.

Pattern Recognition of Rare Events This methodology provides an efficient framework for integrating diverse information into prediction algorithms [4,11,19]. This methodology has been developed by the artificial intelligence school of I. Gelfand for the study of rare phenomena of a highly complex origin. In terminology of pattern recognition, the "object of recognition" is the time moment t. The problem is to recognize whether it belongs to the period of alarm, i. e. to a time interval Δ preceding an extreme event. An alarm starts when certain combinations of premonitory patterns emerges.

Several features of that methodology are important for predicting extreme events in the absence of a complete

closed theory that would unambiguously define a prediction algorithm. First, this kind of pattern recognition relies on simple, robust parameters that overcome the bane of complexity analysis – incomplete knowledge of the system's causal mechanisms and chronic imperfections in the available data. In its efficient robustness, pattern recognition of rare events is akin to exploratory data analysis as developed by J. Tukey [50]. Second, unlike other statistical methods, e. g. regression analysis, that methodology can be used for small samples such as presidential elections or economic recessions. Also, it integrates quantitative and judgmental parameters and thereby more fully captures the full dimensions of the prediction problem than procedures that rely strictly on quantitative variables.

Summing up, the methodology described here can help in prediction when there are (1) many causal variables, (2) qualitative knowledge about which variables are important, and (3) limited amounts of data [2].

Besides societal predictions, pattern recognition of rare events has been successfully applied in seismology and earthquake prediction (e. g. [11,19,20,44,46]), geological prospecting (e. g. [45]) and in many other fields. Review can be found in [21,47]. Tutorial materials are available at the web site of the Abdus Salam International Centre for Theoretical Physics (http://cdsagenda5.ictp.it/full_display.php?da=a06219).

Validation of Prediction Algorithms The algorithms include many adjustable elements, from selecting the data and defining the prediction targets, to specifying numerical parameters involved. In lieu of theory that would unambiguously determine these elements they have to be developed retrospectively, by "predicting" past extreme events. The application of the methodology to known events creates the danger of self-deceptive data-fitting: As J. von Neumann put it "*with four exponents I can fit an elephant*". The proper validation of the prediction algorithms requires three consecutive tests.

- *Sensitivity analysis*: testing whether predictions are sensitive to variations of adjustable elements.
- *Out of sample analysis*: application of an algorithm to past data that has not been used in the algorithm's development. The test is considered successful if algorithm retains its accuracy.
- *Predicting future events* – the only decisive test of a prediction algorithm (see for example Sect. "Elections" below).

A highly efficient tool for such tests is the error Diagram, showing major characteristics of prediction accuracy [33,34,35,36,37,38,39]. Its example is given in Fig. 10. Exhaustive sets of these tests are described in [10,11,24,52].

Common Elements of Data Analyzes

The methodology discussed above was used for predicting various kinds of extreme events, as illustrated in the next four Sections. Naturally, from case to case this methodology was used in different ways, according to specifics of phenomena considered. However in all cases data analysis has essential common elements described below.

Sequence of analysis comprises four stages: (i) Defining prediction targets. (ii) Choosing the data (time series), where premonitory patterns will be looked for and summing up a priori constrains on these patterns. (iii) Formulating hypothetical definition of these patterns and developing prediction algorithm; determining the error diagram. (iv) Validating and optimizing that algorithm.

Preliminary transformation of raw data. In predicting recessions (Sect. "US Economic Recessions"), fast acceleration of unemployment (Sect. "Unemployment") and crime surges (Sect. "Homicide Surges") raw data were time series of relevant monthly indicators, hypothetically containing premonitory patterns. Let $f(m)$ be such an indicator, with integer m showing time in months. Premonitory behavior of some indicators is better captured by their linear trends.

Let $W^f(l/q,p)$ be the local linear least-squares regression of a function $f(m)$ within the sliding time window (q,p):

$$W^f(l/q,p) = K^f(q,p)l + B^f(q,p), \quad q \le l \le p\,, \quad (1)$$

where integers l, q, and p stand for time in months.

Premonitory behavior of most indicators was captured by the following two functions:

- The trend of $f(m)$ in the s months long window, $(m-s,m)$. For brevity we denote

$$K^f(m/s) = K^f(m-s,m) \quad (2)$$

- The deviation of $f(m)$ from extrapolation of its long-term regression (i. e. regression on a long time window $(q,m-1)$:

$$R^f(m/q) = f(m) - W^f(m/q,m-1)\,. \quad (3)$$

Both functions can be used for prediction since their values do not depend on the information about the future (after the month m) which would be anathema in prediction.

Discretization. The prediction algorithms use one or several premonitory patterns. Each pattern is defined at

the lowest – binary – level of resolution, as 0 or 1, distinguishing only the presence of absence of a pattern at each moment of time. Then the objects of recognition are described by binary vectors of the same length. This ensures the robustness of the prediction algorithms.

Simple algorithm called Hamming distance is used for classification of binary vectors in applications considered here, [14,20,28]. Each vector is either premonitory or not. Analyzing the samples of vectors of each class ("the learning material"), the algorithm determines a reference binary vector ("kernel") with components typical for the premonitory vector. Let D be the Hamming distance of a vector from the kernel (the number of non-coinciding binary components). The given vector is recognized as premonitory class, if D is below a certain threshold D^*. This criterion takes advantage of the clustering of precursors in time.

Summing up, these elements of the pattern recognition approach are common for its numerous applications, their diversity notwithstanding. Experience in the specific applications is described in Sects. "Elections", "US Economic Recessions", "Unemployment", "Homicide Surges". The conceptual summary of the accumulated experience is given in the final Sect. "Summary: Findings and Emerging Possibilities".

Elections

This Section describes algorithms for predicting the outcome of the US Presidential and mid-term Senatorial elections [28,29,30,31]. Elections' time is set by the law as follows.

- National elections are held every even-numbered year, on the first Tuesday after the first Monday in November (i. e., between November 2 and November 8, inclusively).
- Presidential elections are held once every 4 years, i. e. on every other election day. People in each of the 50 states and District of Columbia are voting separately for "electors" pledged to one or another of the Presidential candidates. These electors make up the "Electoral College" which directly elects the President. Since 1860, when the present two-party system was basically established, the Electoral College reversed the decision of the popular vote only three times, in 1888, 1912, and 2000. Algorithmic prediction of such reversals is not developed so far.
- A third of Senators are elected for a 6-year term every election day; "mid-term" elections held in the middle of a Presidential term are considered here.

Methodology

The prediction target is an electoral defeat of an "incumbent" party, i. e. the party holding the contested seat. Accordingly, the prediction problem is formulated as whether the incumbent party will retain this seat or lose it to the challenging party (*and not whether Republican or Democrat will win*). As is shown below, that formulation is crucial for predicting the outcomes of elections considered.

Data. The pre-election situation is described by robust common sense parameters defined at the lowest (binary) level of resolution, as the *yes* or *no* answers to the questionnaires given below (Tables 1, 2). The questions are formulated in such a way that the answer *no* favors the victory of the challenging party. According to the Hamming distance analysis (Sect. "Common Elements of Data Analyzes") the victory of the challenging party is predicted when the number of answers *no* exceeds a threshold D^*.

Mid-term Senatorial Elections

The prediction algorithm was developed by a retrospective analysis of the data on three elections, 1974, 1978, and 1982. The questionnaire is shown in Table 1. Victory of the challenger is predicted if the number of answers *no* is 5 or more [28,29,30].

The meaning of these questions may be broader than their literal interpretation. For example, financial contributions (key 5 in Table 2) not only provide the resources required for an effective campaign, but may also constitute a poll in which the preferences are weighed by the money attached.

Predicting future elections. This algorithm (without any changes from year to year and from state to state) was applied in advance to the five subsequent elections, 1986–2002. Predictions are shown in Fig. 2. Altogether, 150 seats were put up for election. For each seat a separate prediction was made, 128 predictions were correct, and 22 – wrong.

Statistical significance of this score is 99.9%. In other words the probability to get such a score by chance is below 0.1% [28,29,30]. For some elections these predictions might be considered as trivial, since they coincide with prevailing expectation of experts. Such elections are identified by *Congressional Review*. Eliminating them from the score still results in 99% significance.

Presidential Elections

The prediction algorithm was developed by a retrospective analysis of the data on the past 31 elections, 1860–1980;

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Table 1
Questionnaire for mid-term Senatorial Elections [28]

1.	(Incumbency): The incumbent -party candidate is the sitting senator.
2.	(Stature): The incumbent -party candidate is a major national figure.
3.	(Contest): There was no serious contest for the incumbent -party nomination.
4.	(Party mandate): The incumbent party won the seat with 60% or more of the vote in the previous election.
5.	(Support): The incumbent -party candidate outspends the challenger by 10% or more.
6.	(Obscurity): The challenging -party candidate is not a major national figure or a past or present governor or member of Congress.
7.	(Opposition): The incumbent party is not the party of the President.
8.	(Contest): There is no serious contest for the challenging -party nomination (the nominee gains a majority of the votes cast in the first primary and beats the second-place finisher at least two to one).

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Table 2
Questionnaire for Presidential elections [29,30]

KEY 1	(Party Mandate): After the midterm elections, the incumbent party holds more seats in the US House of Representatives than it did after the previous midterm elections.
KEY 2	(Contest): There is no serious contest for the incumbent -party nomination.
KEY 3	(Incumbency): The incumbent -party candidate is the sitting president.
KEY 4	(Third party): There is significant third-party or independent campaign.
KEY 5	(Short-term economy): The economy is not in recession during the election campaign.
KEY 6	(Long-term economy): Real per -capita economic growth during the term equals or exceeds mean growth during the previous two terms.
KEY 7	(Policy change): The incumbent administration effects major changes in national policy.
KEY 8	(Social unrest): There is no sustained social unrest during the term.
KEY 9	(Scandal): The incumbent administration is unattained by a major scandal.
KEY 10	(Foreign/military failure): The incumbent administration suffers no major failure in foreign or military affairs.
KEY 11	(Foreign/military success): The incumbent administration achieves a major success in foreign or military affairs.
KEY 12	(Incumbent charisma): The incumbent -party candidate is charismatic or a national hero.
KEY 13	(Challenger charisma): The challenging -party candidate is not charismatic or a national hero.

that covers the period between victories of A. Lincoln and R. Reagan inclusively. The questionnaire is shown in Table 2. Victory for the challenger is predicted if the number of answers *no* is 6 or more [29,30].

Predicting of future elections. This algorithm (without any changes from year to year state) was applied in advance to the six subsequent elections, 1984–2004. Predictions are shown in Fig. 3. All of them happened to be correct. In 2000 the decision of popular majority was reversed by the Electoral College; such reversals are not targeted by this algorithm [29,30].

Understanding Elections

Collective behavior. The finding that aggregate-level parameters can reliably anticipate the outcome of both presidential and senatorial elections points to an electoral behavior highly integrated not only for the nation as a whole but also within the diverse American states.

- A presidential election is determined by the collective, integrated estimation of performance of incumbent administration during the previous four years.
- In case of senatorial elections the electorate has more diffused expectations of performance but puts more importance on political experience and status than in the case of presidential elections. Senate incumbents, unlike presidential ones, do not suffer from a bad economy or benefit from a good one. (This suggests that rather than punishing the party holding a Senate seat for hard times, the voters may instead regard the incumbent party as a safe port in a storm).

Similarity. For each election year in all states the outcomes of elections follow the same pattern that transcends the diversities of the situations in each of the individual elections.

The same pattern of the choice of the US President prevails since 1860, i. e. since election of A Lincoln, despite

0	1	2	3	4	5	6	7
			OK98				
			CO98				
			FL98				
			GA98				
			HA98	TN02			
			ID98	SC02			
			MA98	NC02			
			ND98	NE02			
			PN98	KY02			
			SD98	IA02			
			UT98	CO02			
			FL94	AL02			
			HA94	AK98			
			IN94	CA98			
			MT94	CT98			
			NB94	NE98			
			NJ94	OR98			
			TX94	SC98			
			WA94	VT98			
		AS98	WV94	WA98			
		KA98	WI94	CT94			
		LA98	AK90	MD94			
		MI98	IN90	NV94			
		NH98	KN90	WY94			
		MS94	ME90	CO90			
	AL98	NM94	MA90	HA90			
	AZ98	ND94	MT90	KY90			
	IO98	RI94	NB90	MI90			
	DL94	VT94	NC90	AZ86			
	MA94	AS90	TX90	CO86			
	NY94	IO90	WY90	ID86			
	AL90	MS90	AR86	LA86			
	DE90	NM90	CA86	NY86			
	IL90	OR90	IL86	OK86	WI98	MN94	
	LA90	RI90	IN86	WI86	CA94	MO94	
	OK90	SD90	IA86	**NC86**	ID90	VA94	
	SC90	VA90	NH86	**WA86**	PA86	NH90	
	TN90	WV90	OR86	**MN90**	**IL98**	**IN98**	
	HI86	AK86	VT86	**OK94**	**ME94**	**OH98**	
	OH86	CT86	**TN94**	**PA94**	**AL86**	**MI94**	
UT94	SC86	KS86	**TX02**	**TN294**	**FL86**	**MD86**	**KY98**
GA90	UT86	KY86	**OK02**	**NC98**	**GA86**	**NV86**	**AZ94**
NJ90	**NH02**	**ND86**	**NJ02**	**NY98**	**MO86**	**SD86**	**OH94**
0	1	2	3	4	5	6	7

OK98 – incumbent won, **KY98** – challenger won, errors are highlighted.

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 2
Made-in-advance predictions of the mid-term senatorial elections (1986–2002). Each election is represented by the two-letter state abbreviation with the election year shown by two last digits. Each column shows elections with certain number *D* of answers "*no*" to the questionnaire given in Table 1 (such answers are favorable to challenging party). Value of *D*, indicated at the top, is the Hamming distance from the kernel

D (number of answers NO)	0	1	2	3	4	5	6	7	8	9
Predictions published months in advance			*1984*	*1988*	*2004*	*2000** *1996*	**1992**			
Learning	*1904*	*1956* *1936*	*1944* *1940* *1868*	*1964* *1928* *1916* *1908* *1872* *1864*	*1972* *1924* *1880*	*1948* *1888**	**1912*** **1892**	**1884** **1860**	**1980** **1976** **1968** **1952** **1932** **1920** **1896**	**1960** **1876***

1904 years when incumbent won popular vote
1892 years when challenger won popular vote
* years when popular vote was reversed by electoral vote

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 3
Division of presidential elections (1860–2004) by the number *D* of answers "*no*" to the questionnaire given in Table 2 (such answers are favorable to challenging party). *D* is the Hamming distance from the kernel

all the overwhelming changes in the electorate, the economy, the social order and the technology of politics during these 130 years. (For example, the electorate of 1860 did not include the groups, which constitute 3/4 of present electorate, such as women, African Americans, or most of the citizens of the Latin American, South European, Eastern European, and Jewish descent [30].

An alternative (and more traditional) concept of American elections focuses on the division of voters into interest and attitudinal groups. By this concept the goal of the contestants is to attract the maximum number of voting blocks with minimal antagonism from other blocks. Electoral choice depends strongly on the factors irrelevant to the essence of the electoral dilemma (e. g. on the campaign tactics). The drawbacks of this concept are discussed in [18,30]. In sum, the work on presidential and senatorial elections described above suggests the following new ways of understanding American politics and perhaps the politics of other societies as well.

1. Fundamental shifts in the composition of the electorate, the technology of campaigning, the prevailing economic and social conditions, and the key issues of campaigns do not necessarily change the pragmatic basis on which voters choose their leaders.
2. It is governing not campaigning that counts in the outcomes of presidential elections.
3. Different factors may decide the outcome of executive as compared to legislative elections.
4. Conventional campaigning will not improve the prospects for candidates faced with an unfavorable combination of fundamental historical factors. Disadvantaged candidates have an incentive to adopt innovative campaigns that break the pattern of conventional politics.
5. All candidates would benefit from using campaigns to build a foundation for governing in the future.

US Economic Recessions

US National Bureau of Economic Research (NBER) has identified the seven recessions that occurred in the US since 1960 (Table 3). The starting points of a recession and of the recovery from it follow the months marked by a peak and a trough of economic activity, respectively.

A peak indicates the last month before a recession, and a trough – the last month of a recession.

Prediction targets considered are the first month after the peak and after the trough ("the turns to the worst and to the best", respectively). The start of the first recession, in 1960, is not among the targets, since the data do not cover a sufficient period of time preceding the recession.

The data used for prediction comprise the following six monthly leading economic indicators obtained from the CITIBASE data base, Jan. 1960–June 2000 (abbreviations are the same, as in [49]).

G10FF = FYGT10 – FEDFUN Difference between the annual interest rate on 10 year US Treasury bonds, and federal fund annual interest rate.

IP Industrial Production, total: index of real (constant dollars, dimensionless) output in the entire economy.

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Table 3
American Economic Recessions since 1960

#	Peaks	Troughs
1	1960:04	1961:02
2	1969:12	1970:11
3	1973:11	1975:03
4	1980:01	1980:07
5	1981:07	1982:11
6	1990:07	1991:03
7	2001:03	2001:11

This represents mainly the manufacturing industry, because of the difficulties in measuring the quantity of the output in services (such as travel agents, banking, etc.).

LHELL Index of "help wanted" advertising. This is put together by a private publishing company that measures the amount of job advertising (column-inches) in a number of major newspapers.

LUINC Average weekly number of people claiming unemployment insurance.

INVMTQ Total inventories in manufacturing and trade, in real dollars. Includes intermediate inventories (for example held by manufacturers, ready to be sent to retailers) and final goods inventories (goods on the shelves in stores).

FYGM3 Interest rate on 90 day US treasury bills at an annual rate (in percent).

These indicators were already known [48,49], as those that correlate with a recession's approach.

Prediction of a Recession Start

Single indicators exhibit the following premonitory patterns:

G10FF: small value
IP and INVMTQ: small deviation from the long-term trend R^f (3)
FYGM3: large deviation from the long-term trend R^f
LHELL: small trend K^f (2)
LUINC: large trend K^f

The prediction algorithm triggers an alarm after a month when most of the patterns emerge simultaneously. It lasts Δ months and can be extended by the same rule, if premonitory patterns keep emerging. Formal quantitative definition of the algorithm can be found in [23] along with its validation by sensitivity and out-of-sample analyzes.

Alarms and recessions are juxtaposed in Fig. 4. We see that five recessions occurring between 1961 and 2000 were predicted by an alarm. The sixth recession started in April 2001, one month before the corresponding alarm. (Recession of 1960 was not considered for prediction, since data analyzed start just before it.)

Only the first six recessions listed in Table 1 were considered in the developing of the algorithm [23]. Duration of each alarm was between 1 and 14 months. Total duration of all alarms was 38 months, or 13.6% of the time interval considered. There were no false alarms. No alarms were yielded so far by subsequent prediction in advance and no recession was identified during that time.

Prediction of a Recession End

Prediction targets are the starting points of recovery from recessions; these points are indicated in the last column of Table 3.

The data comprise the same six indicators that indicate the approach of a recession (see Subsect. "Prediction of a Recession Start"); they are analyzed only within the recessions' periods.

Data analysis shows intriguing regularity illustrated in Fig. 5:

- Financial indicators change in opposite directions before the recession and before the recovery.
- Economic indicators change in the same direction before the recession and the recovery; but the change is stronger before the recovery, i. e., the economic situation worsens.

Prediction algorithm is formulated in the same terms as in the previous case but an alarm is triggered after *three* consecutive months when most of the patterns emerge simultaneously. The alarms predict when the recovery will start. Alarms and prediction targets are juxtaposed in Fig. 6. Du-

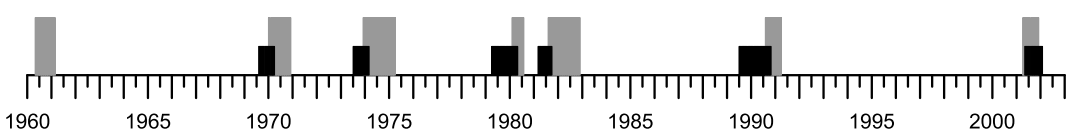

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 4
Alarms (*black bars*) and recessions (*gray bars*)

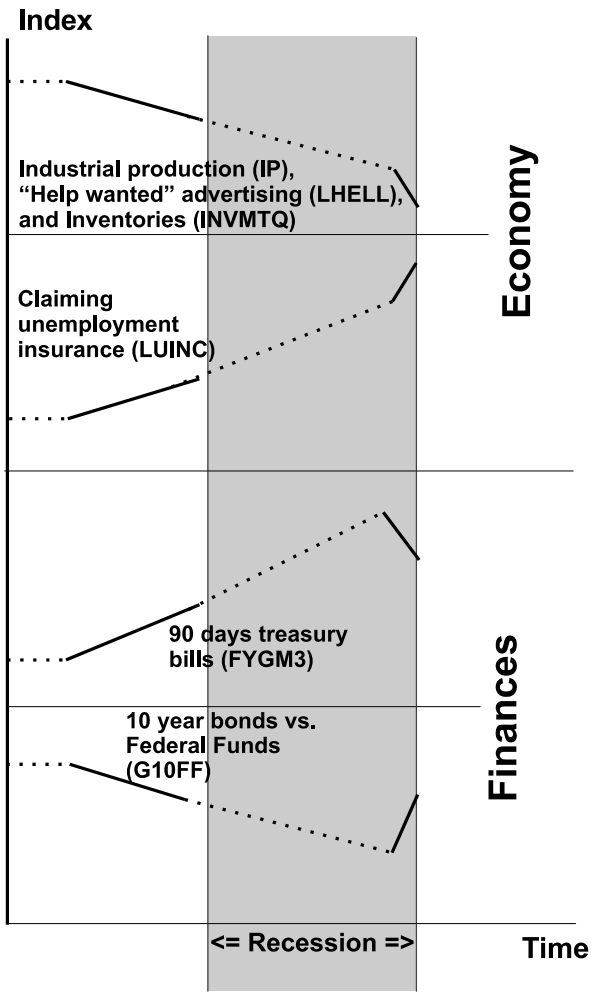

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 5
Premonitory changes of indicators before the start of a recession and before its end. See explanations in the text

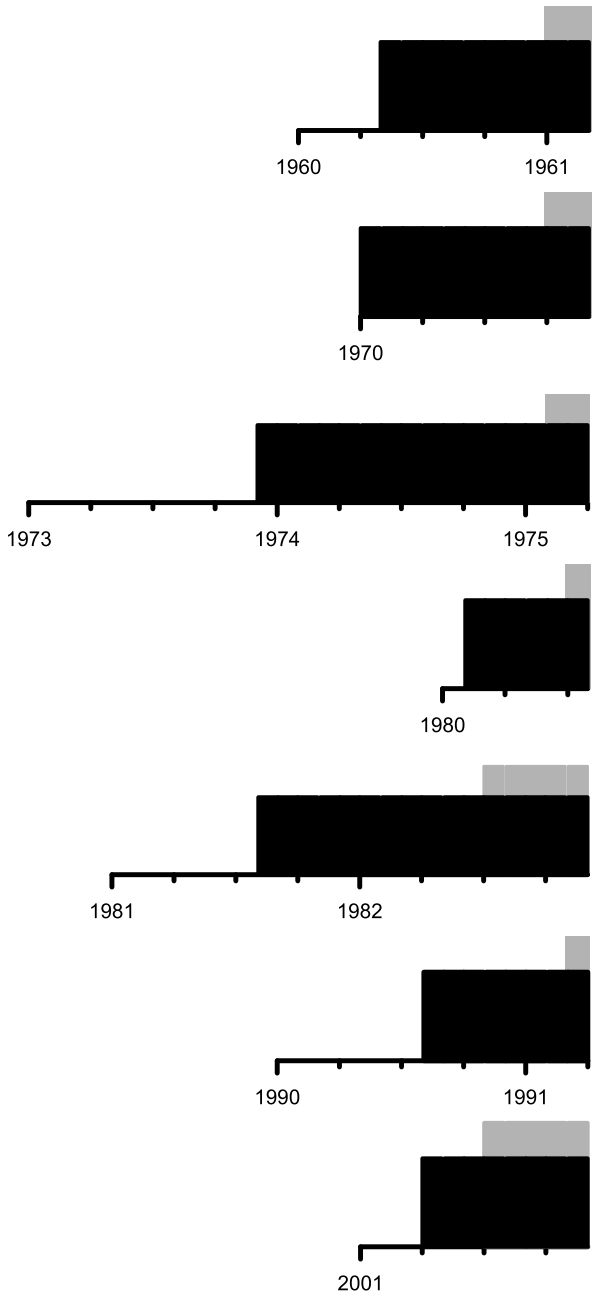

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 6
Prediction of recovery from a recession. *Black bars* – periods of recessions. *Gray bars* – alarms preceding the end of a recession

ration of a single alarm is one to five months. Total duration of alarms is 16 months, which is 22% of time covered by all recessions. There are neither false alarms nor failures to predict.

Unemployment

Here we describe uniform prediction of the sharp and lasting unemployment surge in France, Germany, Italy, and the USA [25].

Prediction Target

A prediction target is schematically illustrated in Fig. 7. Thin curve shows monthly unemployment with seasonal variations. On the thick curve seasonal variations are smoothed away. The arrow indicates a sharp upward bend of the smoothed curve. The moment of that bend is the prediction target. It is called by the acronym *FAU*, for "Fast Acceleration of Unemployment".

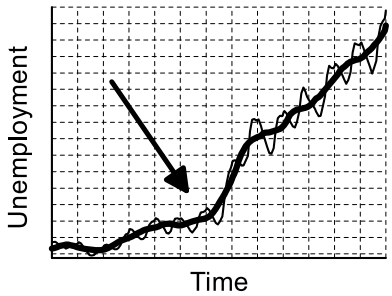

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 7
Fast acceleration of unemployment (*FAU*): schematic definition. *Thin line* – monthly unemployment; with seasonal variations. *Thick line* – monthly unemployment, with seasonal variations smoothed away. The *arrow* indicates a *FAU* – the sharp bend of the smoothed curve. The moment of a *FAU* is the target of prediction

Smoothing was done as follows: Let $u(m)$ be number of unemployed in a month $m = 1, 2, \ldots$. After smoothing out the seasol variation we obtain time series $U(m) = W^u(m/m - 6, m + 6)$; this is the linear regression over the year-long time interval $(m - 6, m + 6)$. A natural robust measure of unemployment acceleration at the time m is the bend of the linear trend of U; in notations used in (1) this is the function $F(m/s) = K^U(m + s, m) - K^U(m, m - s)$. The *FAU*s are identified by the local maxima of $F(m)$ exceeding a certain threshold F. The time $m^\star$ and the height $F^\star$ of such a maximum are, respectively, the time and the magnitude of a *FAU*. Subsequent local minimum of $F(m)$ identifies the month m_e when acceleration ends. Figure 8 shows thus defined *FAU*s for France.

The Data

The analysis has been initially made for France and three groups of data have been analyzed.

- *Composite macroeconomic indicators of national economy*
  1. **IP**: Industrial production indicator, composed of weighted production levels in numerous sectors of the economy, in % relative to the index for 1990.
  2. **L**: Long-term interest rate on 10-year government bonds, in %.
  3. **S**: Short-term interest rate on 3-month bills, in %.
- *Characteristics of more narrow areas of French economy*
  4. **NC**: The number of new passenger car registrations, in thousands of units.

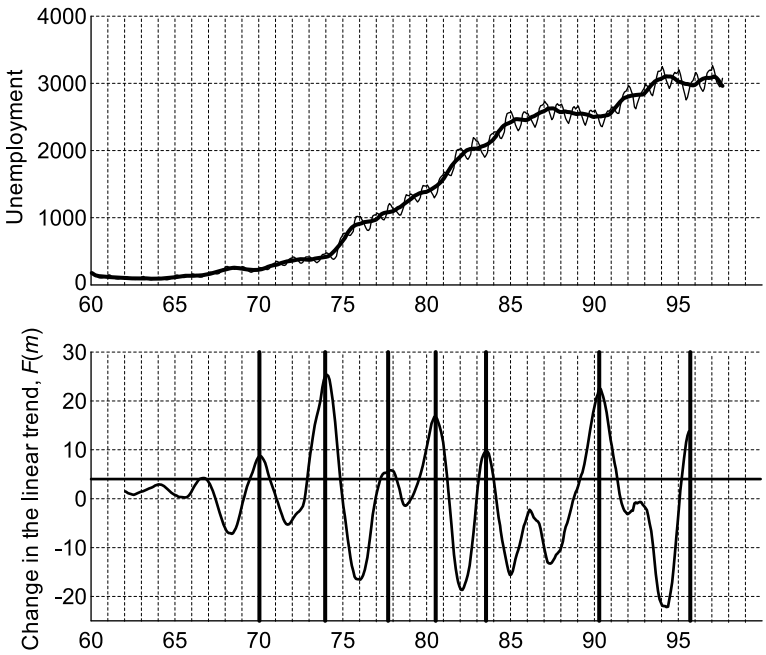

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 8
Unemployment in France. *Top:* Monthly unemployment, thousands of people. *Thin line*: $u(m)$, data from the OECD database; note the seasonal variations. *Thick line*: $U(m)$, data smoothed over one year. *Bottom:* Determination of *FAU*s. $F(m)$ shows the change in the linear trend of unemployment $U(m)$. *FAU*s are attributed to the local maxima of $F(m)$ exceeding threshold $F = 4.0$ shown by *horizontal line*. The *thick vertical lines* show moments of the *FAU*s

5. **EI**: Expected prospects for the national industrial sector.
6. **EP**: Expected prospects for manufacturers.
7. **EO**: Estimated volume of current orders.

Indicators 5–7 distinguish only "good" and "bad" expectations determined polling 2,500 manufacturers, whose answers are by the size of their businesses.

- *Indicators related to US economy.*
  8. **FF/$**: Value of US dollar in French francs.
  9. **AR**: The state of the American economy: is it close to a recession or not? This indicator shows the presence or absence of a current pre-recession alarm (see Subsect. "Prediction of a Recession Start").

The data bases with above indicators for Europe are issued by the Organization for Economic Cooperation and Development [43] and the International Monetary Fund [16].

American analogues of indicators **IP**, **L**, and **S** are provided by CITIBASE; they are described in Sect. "US Economic Recessions>" under abbreviations **IP**, **FYGM3** and **FIGT10** respectively.

Prediction

Single indicators exhibit the following premonitory behavior.

- Steep upward trends of composite indicators (#1–#3). This behavior reflects "overheating" of the economy and may sound counterintuitive for industrial production (#1), since the rise of production is supposed to create more jobs. However, a particularly steep rise may create oversupply.
- Steep downward trends of economic expectations by general public (#4) and business community (#5–#8).
- Proximity of an American recession (#9). Before analysis was made such and opposite precursors might be expected for equally plausible reasons, so that this finding, if further confirmed, does provide a constraint on understanding unemployment's dynamics.

Among different combinations of indicators the macroeconomic ones (#1–#3) jointly give relatively better predictions, with smallest rates of errors and highest stability in sensitivity tests.

Retrospective prediction. Macroeconomic indicators were used jointly in the Hamming distance prediction algorithm (Sect. "Common Elements of Data Analyzes"). Being robust and self-adjusting to regional conditions, this algorithm was applied without any changes to the four countries considered here.

Alarms and *FAU*s are juxtaposed in Fig. 9. Error diagram in Fig. 10 shows quality of prediction for different countries. For US the quality is lower than for European countries, though still higher than in random predictions.

Prediction of the future FAUs was launched for USA. The results are shown in Fig. 11. It shows that by January 2008 two correct predictions have been made, without ether false alarms or failures to predict. In November

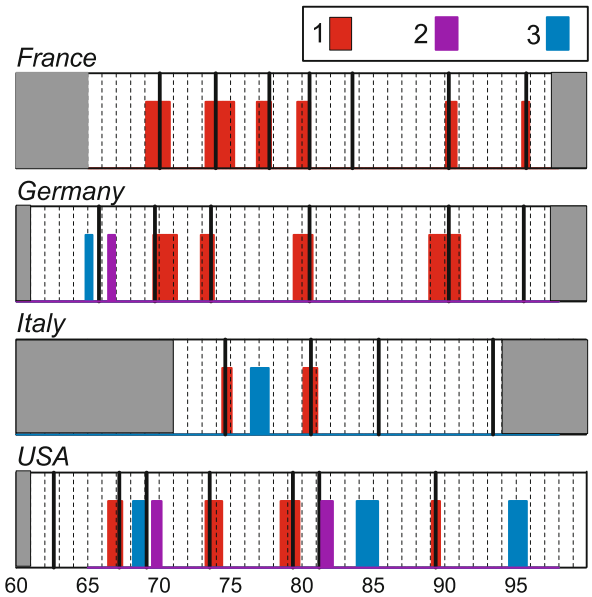

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 9
Retrospective predictions for four countries: *FAU*s and alarms obtained by the prediction algorithm. The *thick vertical lines* show the moments of *FAU*s in a country. *Bars* – the alarms with different outcome: 1 – alarms that predict *FAU*s, 2 – alarms starting shortly after *FAU*s within the periods of unemployment surge, 3 – false alarms. *Shaded areas* on both sides indicate the times, for which data on economic indicators were unavailable

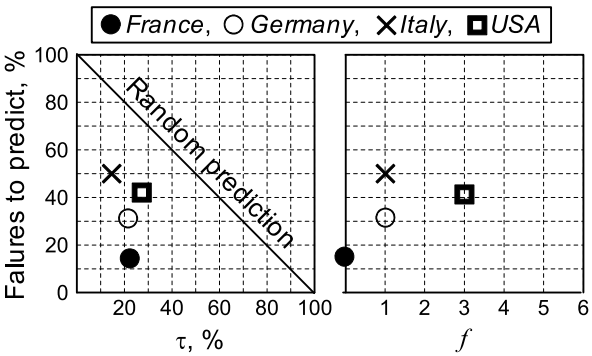

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 10
Error diagram for prediction of *FAU*s in different countries; τ is total duration of alarms in % to the time interval considered, f – total number of false alarms

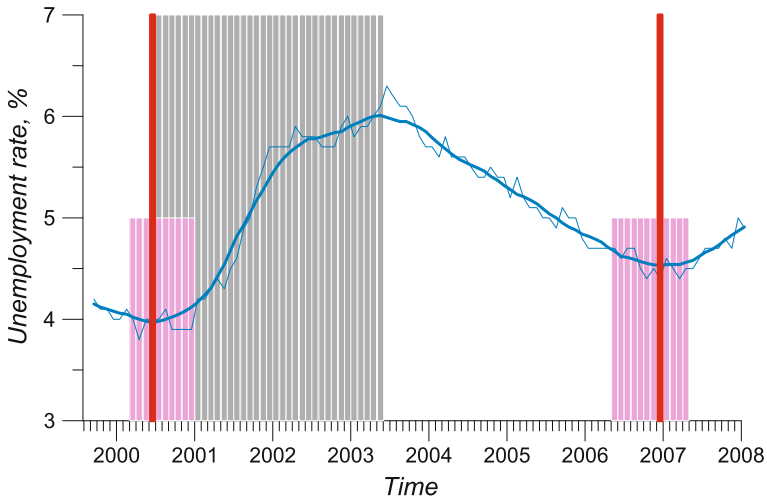

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 11
Experiment in predicting future FAUs, September (1999)–January (2008). *Thin blue curve* shows monthly unemployment rate in USA, according to data of Bureau of Labor Statistics, US Department of Labor (http://www.data.bls.gov). *Thick curve* shows this rate with seasonal variation smoothed away. *Vertical red lines* show prediction targets – the moments of *FAU*, *gray bar* – the period of unemployment's growth; *pink bars* – periods of alarms

2006 the second prediction was filed on the web site of the Anderson School of Management, University of California, Los Angeles (http://www.uclaforecast.com/). This started the documented experiment in testing the algorithm by predicting future *FAU*s on that website.

Homicide Surges

This section analyzes the prediction of homicide rates in an American megacity – Los Angeles, CA [24].

Prediction Target

A prediction target is the start of a sharp and lasting acceleration of the homicide rate; it is called by the acronym SHS, for "Start of the Homicide Surge." It is formally determined by the analysis of monthly homicides rates, with seasonal variations smoothed out, as described in Subsect. "Prediction Target". Prediction targets thus identified are shown by vertical lines in Figs. 12 and 14 below.

The Data

The analyzed data include monthly rates of the homicides and 11 types of lesser crimes, listed in Table 2. Definitions of these crimes are given in [6].

The data are taken from two sources:

- The National Archive of Criminal Justice Data, placed on the web site (NACJD), 1975–1993.

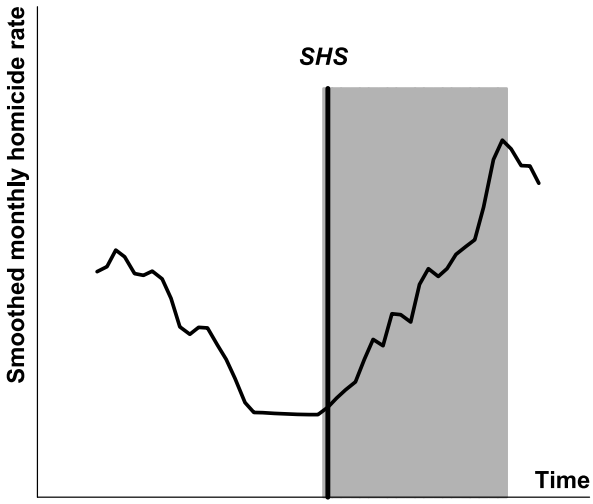

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 12
Target of prediction – the Start of the Homicide Surge ("*SHS*"); schematic definition. *Gray bar* marks the period of homicide surge

- Data bank of the Los Angeles Police Department (LAPD) Information Technology Division), 1990–2003.

The algorithm does not use socio-economic determinants of crime, or other data that might be also useful. The objective was to develop a simple, efficient prediction model; development of comprehensive causal model would be a complementary objective.

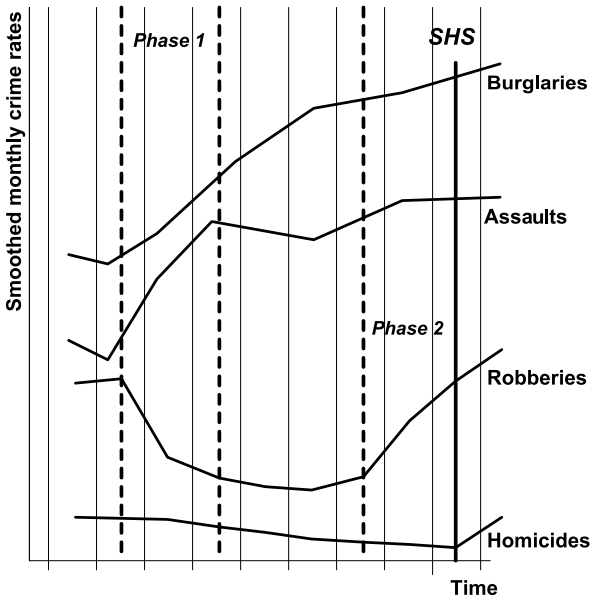

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 13
Scheme of premonitory changes in crime statistics

Prediction

Premonitory behavior of indicators is illustrated in Fig. 13. The first phase is characterized by an escalation of burglaries and assaults, but not of robberies. Later on, closer to a homicide surge, robberies also increase.

The Prediction algorithm based on Hamming distance (see Sect. "Common Elements of Data Analyzes") uses seven indicators listed in Table 4. Other five indicators marked by * are used in sensitivity tests; and the homicide rate is used for identification of targets *SHS*.

Alarms and homicide surges are juxtaposed in Fig. 14. The *SHS* episode in November 1994 has occurred simultaneously with the corresponding alarm. It is captured by an alarm, which starts in the month of *SHS* without a lead time. Prediction missed the October 1999 episode: it occurred two months *before* the start of the corresponding alarm. Such delays should be taken into account for validating the algorithm. Note, however, that the last prediction did remain informative.

Altogether alarms occupy 15% of the time considered. During phase 2 (as defined in Fig. 13) this rate might be reduced [24].

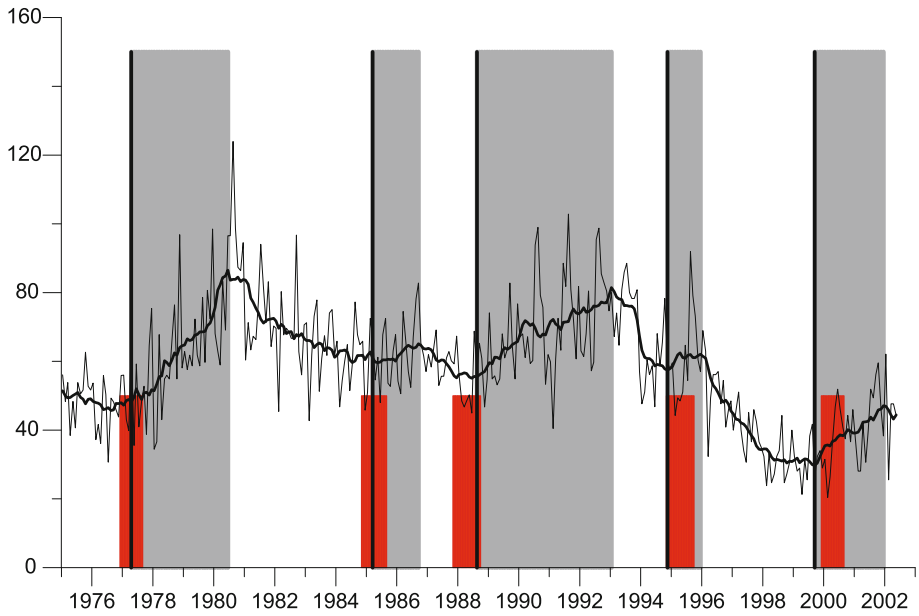

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 14
Performance of prediction algorithm through 1975–2002. *Thin curve* – original time series, total monthly number of homicides in Los Angeles city, per 3,000,000 inhabitants. Data from NACJD [6] have been used for 1975–1993 and from the Data Bank of the Los Angeles Police Department (LAPD Information Technology Division) for subsequent 9 years. *Thick curve* – smoothed series, with seasonal variations eliminated. *Vertical lines* show the targets of prediction – episodes of *SHS* (Subsect. "Prediction Target"). *Gray bars* show the periods of homicide surge. *Red bars* show the alarms declared by the prediction algorithm [24]

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Table 4
Types of crimes considered (after [6])

Homicide	Robberies	Assaults	Burglaries
• All	• All	• All*	• Unlawful not forcible entry
	• With firearms	• With firearms	• Attempted forcible entry*
	• With knife or cutting instrument	• With knife or cutting instrument	
	• With other dangerous weapon	• With other dangerous weapon*	
	• Strong -arm robberies*	• Aggravated injury assaults*	

*Analyzed in sensitivity tests only

Summary: Findings and Emerging Possibilities

The findings described above enhance predictive understanding of extreme events and indicate yet untapped possibilities for further R&D in that field.

Pattern Recognition Approach

Information extracted from the already available data is indeed increased by this approach. To each problem considered here one may apply the following conclusion of J. Stock, a leading expert in the field: "Prediction/of recessions/requires fitting non-linear, high-dimensional models to a handful of observations generated by a possibly non-stationary economic environment.... The evidence presented here suggests that these simple binary transformations of economic indicators have significant predictive content for recessions. It is striking that these models, in which the information in the data is reduced to binary indicators, have predictive contents comparable to or, in many cases, better than that of more conventional models." Importantly, this is achieved by using not more detailed data and models, but more robust aggregation (Subsect. "Predictability vs. Complexity: The Need for Holistic Approach").

Partial "universality" of premonitory patterns is established by broad research in modeling and data analysis. This includes the common definition of the patterns, their self-adjustment, scaling, and similarity [9,10,20,26,42]; see also references in Sects. "Elections", "US Economic Recessions", "Unemployment", "Homicide Surges").

Relation to "cause and effect" analysis (*perpetrators or witnesses?*). Premonitory patterns might be either "perpetrators" contributing to causing extreme events, or the "witnesses" – parallel manifestations of the system's development. The cause that triggered a specific extreme event is usually identified, at least in retrospect. It may be, for example, a certain governmental decision, a change in the international situation, a natural disaster, the depletion of natural resources etc. However an actual extreme event might materialize only if the system is destabilized and "ripe" for it. Patterns of each kind signal such a ripe situation.

What premonitory patters to use for prediction? Existing theories and experience reduce the number of such patterns, but too many of them remain hypothetically promising and have to be chosen by a trial and error procedure. Inevitably a prediction algorithm begins with a limited number of promising patterns. They should be sufficient for prediction, but other patterns may be equally or more useful and should be considered in further development of the algorithm. Most relevant "perpetrators" might not be included in the most useful patterns (e. g. due to their sensitivity to too many factors).

Relation to policy-making: *prediction and disaster preparedness*. Reliable predictions of future extreme events in complex societal systems would allow policy-makers to take remedial action before rather than after the onset of such afflictions as economic disasters, crime surges, etc. As in case of military intelligence predictions would be useful if their accuracy is known, albeit not necessarily high. Analysis of error diagrams allows to regulate the tradeoff between the rates of failures to predict and false alarms according to the needs of a decision-maker.

Relation to governing and campaigning. The findings presented here for the USA elections show that top elected officials would have better chances for reelection, if they focus on effective governing, and not on rhetoric, packaging and image-making. Candidates will benefit themselfes and their parties if they run substantive campaigns that build a foundation for governing during the next term.

Further Possibilities

A wealth of *yet untapped data and models* is readily available for the continuation of the kinds of studies described and analyzed in this article. Following are some immediate possibilities; specific examples can be found in the given references.

- *Continuing experiments in advance prediction*, for which the above findings set up a base (Sect. "Elections"). Successes and errors are equally important [37,38].
- *Incorporating other available data into the analysis* (Sects. "US Economic Recessions","Unemployment")
- *Predicting the same kind of extreme events in different contexts* (Sect. "Unemployment")
- *Predicting the end of a crisis* (Sect. "US Economic Recessions").
- *Multistage prediction with several lead times* (Sect. "Homicide Surges")

Less imminent, but within reach are:

- *"Universal" scenarios of extreme development and low-parametric definition of an ensemble of premonitory patterns* [9,51,52].
- *Validation of an algorithm and joint optimization of prediction and preparedness strategy* [38].
- *Developing prediction algorithms for other types of extreme events*.

The authors would be glad to provide specific information upon request.

Generalizations

The problems considered here have the following common features:

- *The absence of a closed theory* that would unambiguously determine prediction methodology. This leads to the need for intense intertwining of mathematics, statistical physics and non-linear dynamics, a range of societal sciences, and practical experience (Subsect. "Methodology"). In reality this requires long-term collaboration of respective experts. As can be seen from the references to Sects. "Elections", "US Economic Recessions", "Unemployment", "Homicide Surges" previous applications inevitably involved the teams of such experts.
- *Predictions in advance* is the only final validation of the results obtained.
- *The need for holistic analysis* driven to extreme robustness.
- *Considerable, albeit limited, universality* of the premonitory phenomena.

Two classical quotations shed the light on these features:

A. N. Kolmogoroff. "It became clear for me that it is unrealistic to have a hope for the creation of a pure theory [of the turbulent flows of fluids and gases] closed in itself. Due to the absence of such a theory we have to rely upon the hypotheses obtained by processing of the experimental data."

M. Gell-Mann: "… if the parts of a complex system or the various aspects of a complex situation, all defined in advance, are studied carefully by experts on those parts or aspects, and the results of their work are pooled, an adequate description of the whole system or situation does not usually emerge. … The reason, of course, is that these parts or aspects are typically entangled with one another. … We have to supplement the partial studies with a transdisciplinary crude look at the whole."

In the general scheme of things the problem considered belongs to a much wider field – the quest for a universal theory of complex systems extended to predicting extreme events – the Holy Grail of complexity studies. This quest encompasses the natural and human-made complex systems that comprise what some analysts have called "the global village". It requires entirely new applications of modern science, such as algebraic geometry, combinatorics, and thermodynamics. As a means for anticipating, preventing and responding to natural and manmade disasters and for improving the outcomes of economic and political systems, the methods described here may hold one key for the survival and sustainability of our civilization.

Bibliography

Primary Literature

1. Allègre CJ, Le Mouël J-L, Ha Duyen C, Narteau C (1995) Scaling organization of fracture tectonics (SOFT) and earthquake mechanism. Phys Earth Planet Inter 92:215–233
2. Armstrong JS, Cuzan AG (2005) Index methods for forecasting: An application to american presidential elections. Foresight Int J Appl Forecast 3:10–13
3. Blanter EM, Shnirman MG, Le Mouël JL, Allègre CJ (1997) Scaling laws in blocks dynamics and dynamic self-organized criticality. Phys Earth Planet Inter 99:295–307
4. Bongard MM, Vaintsveig MI, Guberman SA, Izvekova ML, Smirnov MS (1966) The use of self-learning prog in the detection of oil containing layers. Geol Geofiz 6:96–105
5. Burridge R, Knopoff L (1967) Model and theoretical seismicity. Bull Seismol Soc Am 57:341–360
6. Carlson SM (1998) Uniform crime reports: Monthly weapon-specific crime and arrest time series 1975–1993 (National, State, 12-City Data), ICPSR 6792 Inter-university Consortium for Political and Social Research. Ann Arbor
7. Farmer JD, Sidorowich J (1987) Predicting chaotic time series. Phys Rev Lett 59:845
8. Gabrielov A, Keilis-Borok V, Zaliapin I, Newman WI (2000) Critical transitions in colliding cascades. Phys Rev E 62:237–249
9. Gabrielov A, Keilis-Borok V, Zaliapin I (2007) Predictability of extreme events in a branching diffusion model. arXiv:0708.1542

10. Gabrielov AM, Zaliapin IV, Newman WI, Keilis-Borok VI (2000) Colliding cascade model for earthquake prediction. Geophys J Int 143(2):427–437
11. Gelfand IM, Guberman SA, Keilis-Borok VI, Knopoff L, Press F, Ranzman IY, Rotwain IM, Sadovsky AM (1976) Pattern recognition applied to earthquake epicenters in California. Phys Earth Planet Inter 11:227–283
12. Gell-Mann M (1994) The quark and the jaguar: Adventures in the simple and the complex. Freeman, New York
13. Crutchfield JP, Farmer JD, Packard NH, Shaw RS (1986) Chaos Sci Am 255:46–57
14. Gvishiani AD, Kosobokov VG (1981) On found of the pattern recognition results applied to earthquake-prone areas. Izvestiya Acad Sci USSR. Phys Earth 2:21–36
15. Holland JH (1995) Hidden order: How adaptation builds complexity. Addison, Reading
16. IMF (1997) International monetary fund, international financial statistics. CD-ROM
17. Kadanoff LP (1976) Scaling, universality and operator algebras. In: Domb C, Green MS (eds) Phase transitions and critical phenomena, vol 5a. Academic, London, pp 1–34
18. Keilis-Borok VI, Lichtman AJ (1993) The self-organization of American society in presidential and senatorial elections. In: Kravtsov YA (ed) Limits of predictability. Springer, Berlin, pp 223–237
19. Keilis-Borok VI, Press F (1980) On seismological applications of pattern recognition. In: Allègre CJ (ed) Source mechanism and earthquake prediction applications. Editions du centre national du la recherché scientifique, Paris, pp 51–60
20. Keilis-Borok VI, Soloviev AA (eds) (2003) Nonlinear dynamics of the lithosphere and earthquake prediction. Springer, Berlin
21. Keilis-Borok V, Soloviev A (2007) Pattern recognition methods and algorithms. Ninth workshop on non-linear dynamics and earthquake prediction, Trieste ICTP 1864-11
22. Keilis-Borok VI, Sorondo MS (2000) (eds) Science for survival and sustainable development. The proceedings of the study-week of the Pontifical Academy of Sciences, 12–16 March 1999. Pontificiae Academiae Scientiarvm Scripta Varia, Vatican City
23. Keilis-Borok V, Stock JH, Soloviev A, Mikhalev P (2000) Pre-recession pattern of six economic indicators in the USA. J Forecast 19:65–80
24. Keilis-Borok VI, Gascon DJ, Soloviev AA, Intriligator MD, Pichardo R, Winberg FE (2003) On predictability of homicide surges in megacities. In: Beer T, Ismail-Zadeh A (eds) Risk science and sustainability. Kluwer, Dordrecht (NATO Sci Ser II Math, Phys Chem 112), pp 91–110
25. Keilis-Borok VI, Soloviev AA, Allègre CB, Sobolevskii AN, Intriligator MD (2005) Patterns of macroeconomic indicators preceding the unemployment rise in Western Europe and the USA. Pattern Recogn 38(3):423–435
26. Keilis-Borok V, Soloviev A, Gabrielov A, Zaliapin I (2007) Change of scaling before extreme events in complex systems. In: Proceedings of the plenary session on "predictability in science: Accuracy and limitations", Pontificiae Academiae Scientiarvm Scripta Varia, Vatican City
27. Kravtsov YA (ed) (1993) Limits of predictability. Springer, Berlin
28. Lichtman AJ, Keilis-Borok VI (1989) Aggregate-level analysis and prediction of midterm senatorial elections in the United States, 1974–1986. Proc Natl Acad Sci USA 86(24):10176–10180
29. Lichtman AJ (1996) The keys to the White House. Madison Books, Lanham
31. Lichtman AJ (2005) The keys to the White House: Forecast for 2008. Foresight Int J Appl Forecast 3:5–9
30. Lichtman AJ (2008) The keys to the White House, 2008 edn. Rowman/Littlefield, Lanham
32. Ma Z, Fu Z, Zhang Y, Wang C, Zhang G, Liu D (1990) Earthquake prediction: Nine major earthquakes in china. Springer, New York
33. Mason IB (2003) Binary events. In: Jolliffe IT, Stephenson DB (eds) Forecast verification. A practitioner's guide in atmospheric science. Wiley, Chichester, pp 37–76
34. Molchan GM (1990) Strategies in strong earthquake prediction. Phys Earth Planet Inter 61:84–98
35. Molchan GM (1991) Structure of optimal strategies of earthquake prediction. Tectonophysics 193:267–276
36. Molchan GM (1994) Models for optimization of earthquake prediction. In: Chowdhury DK (ed) Computational seismology and geodynamics, vol 1. Am Geophys Un, Washington, pp 1–10
37. Molchan GM (1997) Earthquake prediction as a decision-making problem. Pure Appl Geophys 149:233–237
38. Molchan GM (2003) Earthquake prediction strategies: A theoretical analysis. In: Keilis-Borok VI, Soloviev AA (eds) Nonlinear dynamics of the lithosphere and earthquake prediction. Springer, Berlin, pp 209–237
39. Molchan G, Keilis-Borok V (2008) Earthquake prediction: Probabilistic aspect. Geophys J Int 173(3):1012–1017
40. NACJD: http://www.icpsr.umich.edu/NACJD/index.html
41. NBER: http://www.nber.org/cycles/cyclesmain.html
42. Newman W, Gabrielov A, Turcotte DL (eds) (1994) Nonlinear dynamics and predictability of geophysical phenomena. Am Geophys Un, Int Un Geodesy Geophys, Washington
43. OECD (1997) Main economic indicators: Historical statistics 1960–1996. Paris, CD-ROM
44. Press F, Briggs P (1975) Chandler wobble, earthquakes, rotation and geomagnetic changes. Nature 256:270–273, London
45. Press F, Briggs P (1977) Pattern recognition applied to uranium prospecting. Nature 268:125–127
46. Press F, Allen C (1995) Patterns of seismic release in the southern California region. J Geophys Res 100(B4):6421–6430
47. Soloviev A (2007) Application of the pattern recognition techniques to earthquake-prone areas determination. Ninth workshop on non-linear dynamics and earthquake prediction, Trieste ICTP 1864-9
48. Stock JH, Watson MW (1989) New indexes of leading and coincident economic indicators. NBER Macroecon Ann 4:351–394
49. Stock JH, Watson MW (1993) A procedure for predicting recessions with leading indicators. In: Stock JH, Watson MW (eds) Business cycles, indicators, and forecasting (NBER Studies in Business Cycles, vol 28), pp 95–156
50. Tukey JW (1977) Exploratory data analysis. Addison-wesley series in behavioral science: Quantitative methods. Addison, Reading
51. Turcotte DL, Newman WI, Gabrielov A (2000) A statistical physics approach to earthquakes. In: Geocomplexity and the physics of earthquakes. Am Geophys Un, Washington
52. Zaliapin I, Keilis-Borok V, Ghil M (2003) A Boolean delay model of colliding cascades, II: Prediction of critical transitions. J Stat Phys 111(3–4):839–861

Books and Reviews

Bongard MM (1967) The problem of recognition. Nauka, Moscow
Brito DL, Intriligator MD, Worth ER (1998) In: Eliassson G, Green C (eds) Microfoundations of economic growth: A Schumpeterian perspective. University of Michigan Press, Ann Arbor
Bui Trong L (2003) Risk of collective youth violence in french suburbs. A clinical scale of evaluation, an alert system. In: Beer T, Ismail-Zadeh A (eds) Risk science and sustainability. Kluwer, Dordrecht (NATO Sci Ser II Math Phys Chem 112)
Engle RF, McFadden DL (1994) (eds) Handbook of econometrics, vol 4. North-Holland, Amsterdam
Klein PA, Niemira MP (1994) Forecasting financial and economic cycles. Wiley, New York
Messner SF (1983) Regional differences in the economic correlates of the urban homicide rate. Criminology 21:477–488
Mitchell WC (1951) What happens during business cycles: A progress report. NBER, New York
Mitchell WC, Burns AF (1946) Measuring business cycles. NBER, New York
Moore GH (ed) (1961) Business cycle indicators. NBER, New York
Mostaghimi M, Rezayat F (1996) Probability forecast of a downturn in US economy using classical statistical theory. Empir Econ 21:255–279
Watson MW (1994) In: Engle RF, McFadden DL (eds) Handbook of econometrics, vol IV. North-Holland, Amsterdam

Extreme Value Statistics

Mario Nicodemi
Department of Physics, University of Warwick, Coventry, UK

Article Outline

Glossary
Definition of the Subject
Introduction
The Extreme Value Distributions
The Generalized Extreme Value Distribution
Domains of Attraction and Examples
Future Directions
Bibliography

Glossary

Random variable When a coin is tossed two random outcomes are permitted: head or tail. These outcomes can be mapped to numbers in a process which defines a 'random variable': for instance, 'head' and 'tail' could be mapped respectively to +1 and −1. More generally, any function mapping the outcomes of a random process to real numbers is defined a random variable [15]. More technically, a random variable is any function from a probability space to some measurable space, i. e., the space of admitted values of the variable, e. g., real numbers with the Borel σ-algebra. The amount of rainfall in a day or the daily price variation of a stock are two more examples. It's worth to stress that, formally, the outcome of a given random experiment is not a random variable: the random variable is the function describing all the possible outcomes as numbers. Finally, two random variables are said independent when the outcome of either of them has no influence on the other.

Probability distribution The probability of either outcomes, 'head' and 'tail', in tossing a coin is 50%. Similarly, a discrete random variable, X, with values $\{x_1, x_2, \dots\}$ has an associate discrete probability distribution of occurrence $\{p_1, p_2, \dots\}$. More generally, for a random variable on real numbers, X, the corresponding probability distribution [15] is the function returning the probability to find a value of X within a given interval $[x_1, x_2]$ (where x_1 and x_2 are real numbers): $\Pr[x_1 \le X \le x_2]$. In particular, the random variable, X, is fully characterized by its cumulative distribution function, $F(x)$, which is: $F(x) = \Pr[X < x]$ for any x in $\mathcal{R}$. The probability distribution density, $f(x)$, can be often defined as the derivative of $F(x)$: $f(x) = \mathrm{d}F(x)/\mathrm{d}x$.
The probability distribution of two independent random variables, X and Y, is the product of the distributions, F_X and F_Y, of X and Y: $F(x, y) \equiv \Pr[X < x; Y < y] = \Pr[X < x] \cdot \Pr[Y < y] = F_X(x) \cdot F_Y(y)$.

Expected value The expected value [15] of a random variable is its average outcome over many independent experiments. Consider, for instance, a discrete random variable, X, with values in the set $\{x_1, x_2, \dots\}$ and the corresponding probability for each of these values $\{p_1, p_2, \dots\}$. In probability theory, the expected, or average, value of X (denoted $E(X)$) is just the sum: $E(X) = \sum x_i p_i$. For instance, if you have an asset which can give two returns $\{x_1, x_2\}$ with probability $\{p_1, p_2\}$, its expected return is $x_1 p_1 + x_2 p_2$.
In case we have a random variable defined on real numbers and $F(x)$ is its probability distribution function, the expected value of X is: $E(X) = \int X \mathrm{d}F$. As for some $F(x)$ the above integral may not exist, the 'expected value' of a random variable is not always defined.

Variance and moments The variance [15] of a probability distribution is a measure of the average deviations from the mean of the related random variable. In probability theory, the variance is usually defined as the mean squared deviation, $E((X - E(X))^2)$, i. e., the expected value of $(X - E(X))^2$. The square root of the

variance is named the standard deviation and is a more sensible measure of fluctuations of X around $E(X)$. Alike $E(X)$, for some distributions the variance may not exist. In general, the expected value of the kth power of X, $E(X^k)$, is called the kth moment of the distribution.

The Central limit theorem The Central Limit Theorem [15] is a very important result in probability theory stating that the sum of N independent identically-distributed random variables, with finite average and variance, has a Gaussian probability distribution in the limit $N \to \infty$, irrespective of the underlying distributions of the random variables. The domain of attraction of the Gaussian as a limit distribution is, thus, very large and can explain why the Gaussian is so frequently encountered. The theorem is, in practice, very useful since many real random processes have a finite average and variance and are approximately independent.

Definition of the Subject

Extreme value theory is concerned with the statistical properties of the extreme events related to a random variable (see Fig. 1), and the understanding and applications of their probability distributions. The methods and the practical use of such a theory have been developed in the last 60 years, though, many complex real-life problems have only recently been tackled. Many disciplines use the tools of extreme value theory including meteorology, hydrology, ocean wave modeling, and finance to name just a few.

For example, in economics, extreme value theory is currently used by actuaries to evaluate and price insurance against the probability of rare but financially catastrophic events. An other application is for the estimation of Value at Risk. In hydrology, the theory is applied by environmental risk agencies to calculate, for example, the height of sea-walls to prevent flooding. Similarly, extreme value theory is also used to set strength boundaries in engineering materials, as well as for material fatigue and reliability in buildings (e. g., bridges, oil rigs), and estimating pollution levels.

This paper aims to give a simple, self contained, introduction to the motivations and basic ideas behind the development of extreme value theory, and briefly covers a few more technical topics such as extreme r order statistics and the generalization of extreme value distribution theory. We refer to textbooks on probability theory, such as [15] (or, for simplicity, to the Glossary), for the definition of the basic notions of probability used here.

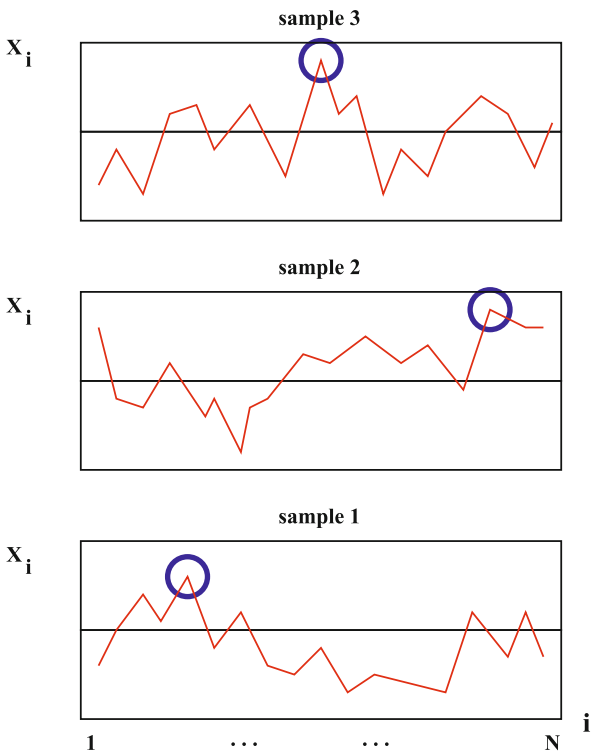

Extreme Value Statistics, Figure 1
We show three samples, 1 to 3, each with N realizations of a random variable $X_i (i \in \{1, \ldots, N\})$. Extreme value theory is concerned with the statistical properties of occurrence of the extreme values in those samples, such as the maxima (*circled points*)

The extensive research on extreme value statistics is reviewed in excellent books published over the past years, e. g., [6,11,12,13,14,17,18,21,24]; we provide here only an overview of the basic concepts and tools. To make this paper as self contained as possible, the Glossary gives a beginner introduction to all the elementary notions of probability theory encountered in the following sections.

Introduction

Extreme events, exceeding the typical expected value of a random variable, can have substantial relevance to problems raising in disciplines as diverse as sciences, engineering and economics.

Extreme value theory is a sub-field of applied statistics, early developed [13,14,17,18] by mathematician such as Fisher, Tippett, Gnedenko, and, in particular, Emil Julius Gumbel, dealing precisely with the problems related to extreme events (see Fig. 1). One of its key point is the so called 'three types theorem', relating the properties of

the distribution of probability of the underlying stochastic variable to its extreme value distributions, i. e., the limiting distributions for the extreme (minimum or maximum) value of a large collection of random observations. Interestingly, for a comparatively large class of random variables, the theory points out that only a few species of limit extreme value distributions are found.

In some respect, the 'three types theorem' can be considered the analogous of the well known central limit theorem applying to ordinary sums, or averages, of random variables. From a practical point of view it is as important, since it opens a way to estimate the asymptotic distribution of extreme values without any a priori knowledge, or guess assumption, on the parent distribution. In this way, we have a solid ground to estimate the parameters of the limit distributions along with their confidence intervals, an issue crucial, for instance, to proper risk assessment.

In Finance, for example, market regulators and financial institutions face the important and complex task to estimate and manage risk. Assessing the probability of rare and extreme events is, of course, a crucial issue and reliable measures of risk are needed to minimize undesirable effects on portfolios from large fluctuations in market conditions, e. g., exchange rates or prices of assets. Similar issues about risk and reliability are faced in insurance and banking, which are deeply concerned with unusually large fluctuations. Extreme value theory provides the solid theoretical foundation needed for the statistical modeling of such events and proper computation of risk and related confidence intervals.

The study of natural and environmental hazards is also strongly interested to extreme events; for instance, reported hydrology and meteorology applications of extreme event theory concern flood frequency analysis, estimation of precipitation probabilities, or extreme tide levels. Predictions of events such as strong heat waves, rainfall, occurrence of huge sea waves are deeply grounded on such a theory as well. Analogous problems are found in telecommunications and transport systems, such as traffic data analysis, Internet traffic, queuing modeling; problems from material science, pollutants hazards on health add examples from a very long list of related phenomena.

It is impossible to summarize here the huge, often technical, literature on all these topics and we refer to the general books cited in the bibliography. Actually, to give an idea of the variety of applications of the theory, we only mention a few more example from still an other class of disciplines, Physical Sciences. In Physics, for instance, the equilibrium low-temperature properties of disordered systems are characterized by the statistics of extremely low-energy states. Several problems in this class, including the Random Energy Model and models for decaying Burgers turbulence, have been connected to extreme value distributions [3]. In GaAs films, extreme values in Gaussian $1/f$ correlations of voltage fluctuations were shown to follow one of the limit distributions of extreme value theory, the Gumbel asymptote [1]. Hierarchically correlated random variables representing the energies of directed polymers [9] and maximal heights of growing self-affine surfaces [20] exhibits extreme value statistics as well. The Fisher–Tippett–Gumbel asymptote is involved in distribution of extreme height fluctuations for Edwards–Wilkinson relaxation of fluctuating interfaces on small-world-coupled interacting systems [16]. A connection was also established between the energy level density of a gas of non-interacting bosons and the distribution laws of extreme value statistics [7].

In case of systems with correlated variables, the application of the extreme value theory is far from trivial. A theorem states that the statistics of maxima of stationary Gaussian sequences, with suitable correlations, asymptotically converges to a Gumbel distribution [2]. Evidences supporting similar results were derived from numerical simulations and analysis of long-term correlated exponentially distributed signals [10]. In general, however, the scenario is non trivial. For instance, in Physics a variant of Gumbel distribution was observed in turbulence [4] and derived in the two-dimensional XY model [5], which are systems where correlations play an important role. Similarly, correlated extreme value statistics were discovered in the Sneppen depinning model [8]. In models for fluctuating correlated interfaces, such as Edwards–Wilkinson and Kardar–Parisi–Zhang equations, an exact solution for the distribution of maximal heights was recently derived and it turns out to be an Airy function [19].

After the above brief picture of the field, we illustrate next the general properties of extreme value distributions of independent random variables.

The Extreme Value Distributions

Extreme value distributions are the limit distributions of extremes (either maxima or minima) of a set of random variables (see Fig. 1). For definiteness, we will deal here with maxima, as minima can be seen as 'maxima' of a set of variables with opposite signs.

Consider a set $\{X_1, X_2, \ldots, X_N\}$ of N independent identically distributed random variables, X_i, with a cumulative distribution function, $F(x) \equiv \Pr\{X_i \leq x\}$. The maximum in the set, $Y_N = \mathrm{Max}\{X_1, X_2, \ldots, X_N\}$, has a distribution function, $H_N(x)$, which is simply related to

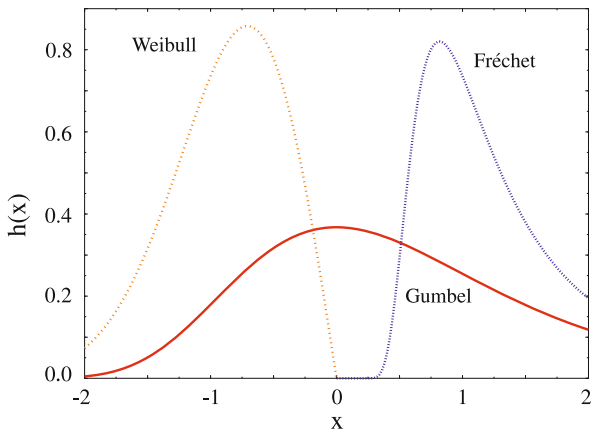

Extreme Value Statistics, Figure 2
As an example, we plot in this figure the Gumbel density distribution, $h(x) = dH(x)/dx$, from Eq. (3), and the Fréchet and Weibull density distributions from Eqs. (4) and (5), for $\alpha = 2$

F, since by definition of Y_N, we have:

$$\begin{aligned} H_N(x) &\equiv \Pr\{Y_N \le x\} \\ &= \Pr\{X_1 \le x, X_2 \le x, \ldots X_N \le x\} \\ &= \Pr\{X_1 \le x\} \cdot \Pr\{X_2 \le x\} \cdot \ldots \cdot \Pr\{X_N \le x\} \\ &= F^N(x)\,. \end{aligned} \tag{1}$$

In the limit of large samples, $N \to \infty$, it is possible to show that, under some general hypotheses on F described below, we can find a suitable sequence of scaling constants a_N and b_N, such that the scaled variable $y_N = (Y_N - b_N)/a_N$ has a non degenerate probability distribution function $H(y)$. Specifically, as $N \to \infty$, the distribution $\Pr\{y_N \le y\}$ has a non trivial well defined limit $H(y)$:

$$\begin{aligned} \Pr\{y_N \le y\} &= \Pr\{(Y_N - b_N)/a_N \le y\} \\ &= \Pr\{Y_N \le a_N y + b_N\} \\ &= F^N(a_N y + b_N) \longrightarrow H(y) \;\text{ for }\; N \to \infty\,. \end{aligned} \tag{2}$$

For a given underlying distribution, F, the individuation of the precise sequence of scaling constants a_N and b_N required in Eq. (2) is a non trivial technical problem in the mathematics of extreme values [18], which we briefly discuss in the next sections. Such an issue is overshadowed by the simplicity of the result of the 'three types theorem', which states that there are only three types (apart from a scaling transformation of the variable) of limiting distribution $H(y)$ (see Fig. 2):

I) Gumbel type:

$$H(y) = \exp[-\exp(-y)] \tag{3}$$

with $-\infty < y < \infty$;

II) Fréchet type:

$$H(y) = \exp[-y^{-\alpha}] \tag{4}$$

where α is a fixed exponent and $0 < y < \infty$ (with $H(y) = 0$ for $y < 0$);

III) Weibull type:

$$H(y) = \exp[-(-y)^{\alpha}] \tag{5}$$

where α is a fixed exponent and $-\infty < y < 0$ (with $H(y) = 1$ for $y > 0$).

The Generalized Extreme Value Distribution

In the extreme value statistics literature, the three types of limiting distributions, Gumbel, Fréchet and Weibull, are often represented as a single family including all of them, the so called generalized extreme value distribution:

$$H(y;\mu,\sigma,\xi) = \exp\left[-\left(1+\xi\frac{y-\mu}{\sigma}\right)^{-1/\xi}\right] \tag{6}$$

with support in the interval where $1 + \xi(y-\mu)/\sigma > 0$, as otherwise H is either zero or one. Out of the three parameters μ, σ, ξ of Eq. (6), ξ is called the 'shape' parameter and it is especially important as it selects the specific type of asymptote:

I) The case $\xi = 0$ corresponds to the Gumbel asymptote, since it is easy to show that

$$\lim_{\xi\to 0} H(y;\mu,\sigma,\xi) = \exp\left[-\exp\left(\frac{-y-\mu}{\sigma}\right)\right]; \tag{7}$$

II) Similarly, the case $\xi > 0$ corresponds to the Fréchet asymptote of Eq. (4), where the exponent is $\alpha = 1/\xi$;

III) And, finally, the case $\xi < 0$ corresponds to the Weibull asymptote of Eq. (5), where the exponent is $\alpha = -1/\xi$.

The parameters μ and σ of Eq. (6) are called the 'location' and the 'scale' parameters, since they are related to the moments of the generalized extreme value distribution of Eq. (6). Figure 3 plots the effects of changes in μ and σ on the form of $H(y)$ from Eq. (6) in the Gumbel case, $\xi = 0$. It is possible to show [18] that the kth moment is finite only if $\xi < 1/k$. The mean, which exists only if $\xi < 1$, can be expressed in the following general form

$$E(y) = \mu + \frac{\sigma}{\xi}[\Gamma(1-\xi) - 1]\,, \tag{8}$$

where $\Gamma(x)$ is the Gamma function. In the Gumbel limit, $\xi \to 0$, the above result is simplified to: $E(y) = \mu + \sigma\gamma$, where $\gamma = 0.577\ldots$ is the Euler γ constant.

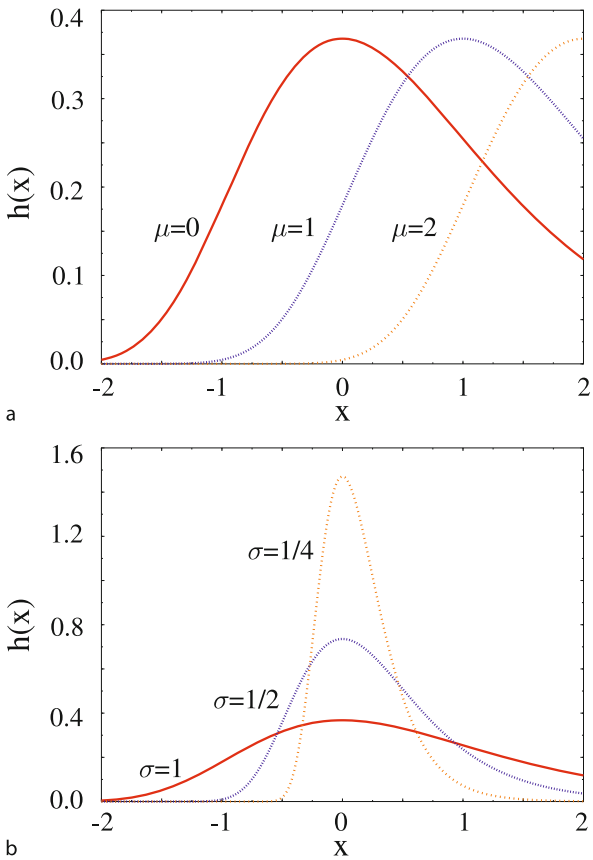

Extreme Value Statistics, Figure 3
In this figure we show the effects of the μ and σ parameters on the appearance of the generalized extreme value density distribution, $h(x) = dH(x)/dx$, from Eq. (6), in the case $\xi = 0$, i. e., the Gumbel type. In the *upper panel*, we plot $h(x)$ for $\sigma = 1$ and $\mu = 0, 1, 2$. In the *lower panel*, we plot $h(x)$ for $\mu = 0$ and $\sigma = 1/4, 1/2, 1$

Analogously, the variance, existing for $\xi < 1/2$, can be written as:

$$E\left([y - E(y)]^2\right) = \left(\frac{\sigma}{\xi}\right)^2 \left[\Gamma(1-2\xi) - \Gamma^2(1-\xi)\right], \quad (9)$$

which in the $\xi \to 0$ limit becomes $E\left([y - E(y)]^2\right) = \sigma^2\pi^2/6$.

The *r* Largest Order Statistics

The results on distributions of the maximum (or minimum) discussed above can be extended to the set of the rth largest value of an ensemble. Consider a set $\{X_1, X_2, \ldots, X_N\}$ of N identically distributed random variables which, for simplicity of notation, are arranged in order of magnitude: $X_1 < X_2 < \cdots < X_N$. As before, $F(x) \equiv \Pr\{X_i \le x\}$ is their common cumulative distribution function. The statistics of X_N and X_1 are the distribution of, respectively, the maximum and the minimum seen before. Similarly, X_r (with $r \in \{1, N\}$) is called the r (largest) order statistic.

The r order statistic has a distribution function, $H_r(x)$, simply related to F

$$\begin{aligned} H_r(x) &\equiv \Pr\{X_r \le x\} \\ &= \Pr\{X_1 \le x, X_2 \le x, \ldots, X_r \le x\} \\ &\quad \cdot \Pr\{X_{r+1} > x, \ldots, X_N > x\} \\ &= \sum_{i=r}^{N} \frac{N!}{(N-i)!\,i!} F^i(x)[1 - F(x)]^{N-i}\,. \end{aligned} \quad (10)$$

The theory of the generalized extreme value distribution can be extended to the r order statistic. Actually, in the limit $N \to \infty$, if a suitable sequence of scaling constants, a_N and b_N, can be found such that the scaled maximum variable $y_N = (X_N - b_N)/a_N$ has a limit distribution function $H(y)$ given in Eq. (6), then the r order statistic has a limit distribution which can be easily expressed in terms of $H(y)$ [18].

In order to describe a broader panorama of the available results in extreme value theory, we give here some details on the more general case of the limit probability distribution density of the vector of the first r largest values $(y_1, y_2, \ldots, y_r) = ((X_N - b_N)/a_N, (X_{N-1} - b_N)/a_N, \ldots, (X_{N-r+1} - b_N)/a_N)$. Such a limit distribution density can be shown to be [18]

$$\begin{aligned} h(y_1, \ldots, y_r) = \frac{1}{\sigma^r} \exp\Bigg[&-\left(1 + \xi\frac{y_r - \mu}{\sigma}\right)^{-\frac{1}{\xi}} \\ &- \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{r} \ln\left(1 + \xi\frac{y_i - \mu}{\sigma}\right)\Bigg]\,. \end{aligned} \quad (11)$$

Most of the other results of the previous sections can be generalized to the r order statistics, as shown for instance in [18].

Domains of Attraction and Examples

The problem of finding the domains of attraction of the classes of limiting distribution is a complex, partially still open, topic in extreme value theory [18]. Even in the case of independent identically distributed random variables, understanding which asymptote a given distribution, F, converges to and which is the sequence of scaling constants, a_N and b_N, can be a non trivial task. As the extreme events of a random variable are characterized by the tail of

F, a simple approximate approach to guess the domain of attraction F falls into is to consider its behavior for large x. We summarize below a few well known examples having a broad validity which can guide practical applications of the theory to the case where the random variables are independent and identically distributed.

E I) Many common distributions, $F(x)$, have exponential tails in x, very important examples being the Gaussian and the exponential distributions. In this case, their extreme value statistic is the Gumbel asymptote.
A more formal condition for F to belong to the domain of attraction of the Gumbel limiting distribution was established by von Mises. Take a function F and denote $x_{\max}$ the largest value in its support, i. e., where $F(x_{\max}) = 1$ (the point $x_{\max}$ can be also infinite). Consider the derivative $f(x) = \mathrm{d}F(x)/\mathrm{d}x$ and the rate of $F(x)$ in approaching 1 as $x \to x_{\max}$: when $x \to x_{\max}$, if

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{1-F(x)}{f(x)}\right] \to 0 \tag{12}$$

then $\Pr\{y_N \le y\}$ tends to the Gumbel asymptote given in Eq. (3).
The above criterion can be rephrased in a more colloquial way: the Gumbel type is the limiting distribution when $1 - F(x)$ decays faster than a polynomial for $x \to x_{\max}$. Beyond the Gaussian and exponential distributions, the lognormal, the Gamma, the Weibull, the Benktander-type I and II, and many more common distributions, with $x_{\max}$ either finite or infinite, belong to this class.

E II) Distributions such as the Pareto, Cauchy, Student, Burr have the Fréchet asymptote. More generally, when $x_{\max}$ is infinite and F has a power law tail for $x \to \infty$

$$1 - F(x) \simeq x^{-\alpha} \tag{13}$$

with an exponent $\alpha > 0$, then the domain of attraction of the extreme value statistics is the Fréchet type given in Eq. (4), with precisely the same exponent α of Eq. (13).

E III) Finally, when $x_{\max}$ is finite and F has a power law behavior for $x \to x_{\max}$

$$1 - F(x) \simeq (x_{\max} - x)^{-\alpha} \tag{14}$$

with an exponent $\alpha > 0$, then the domain of attraction of the extreme value statistics is the Weibull type of Eq. (5), with the same exponent α of Eq. (14). The Uniform and Beta distributions have, for instance, the Weibull asymptote.

Extremes of Correlated Random Variables

When the underlying random variables are not independent, as in many cases of practical relevance ranging from Meteorology to Finance, the problem to individuate the form, or even the existence, of the limiting distribution is, in general, open. The existing broad technical literature on the topic [18] shows that the three types, Gumbel, Fréchet and Weibull, summarized in the generalized extreme value distribution of Eq. (6), often arise as well. For instance, a recent theorem has shown that in the case of stationary Gaussian sequences with suitable correlations the distribution of maxima asymptotically follows the Gumbel type [2]. Analysis of numerical simulations of long-term correlated exponentially distributed signals has given evidences supporting similar conclusions [10]. Sometimes, when considering 'time' series of N correlated variables, the approximate rule of thumb that N must be much bigger than the 'correlation length' of the sequence is used as a guide to decide whether Eq. (6) is likely to be the right asymptote.

Some of the example mentioned in the Introduction can help, however, in delineating the strong limits of the above approximate criteria and the lack of a general picture. For instance, in the XY model for magnetic systems used in Statistical Physics, in the Kosterliz–Thouless low temperature phase the magnetization has a distribution which is a generalized Gumbel [5], but not the one in Eq. (3), a result expected to hold in a broad class of systems. In models for fluctuating interfaces, developing correlations, described by Edwards–Wilkinson and Kardar–Parisi–Zhang like equations, it has been derived that the exact distribution of maximal heights is an Airy function [19]. These examples show the variety of situations which can arise in practical cases and indicate that the theorems derived for independent identically distributed variables must be applied with caution.

Future Directions

In the sections above, we reviewed at an introductory level the mathematics of extreme value theory, with a special focus on the 'three types theorem' on the limiting distributions. We also discussed their domains of attraction and many examples on random extreme events. We have not covered, instead, other important, though, more technical and still evolving topics such as the theoretical approach to the problem of 'exceedances over thresholds', and the the methodology for estimating from real sample data the

parameters of extreme distributions, such as maximum likelihood and Bayesian methods. These are covered, for instance, in the general references listed in the bibliography. Actually, there is a number of excellent textbooks on these topics, ranging from the original book by E.J. Gumbel [17], to more recent volumes illustrating in details extreme value theory in the formal framework of the theory of probability [11,13,14,18]. Volumes more focused on applications to Finance and Insurances are, e. g., [6,11,12,21], as applications to climate, hydrology and meteorology research are found in [6,12,21,24]. Finally, there is a number of more technical review papers on the topic, including [10,22,23].

Bibliography

1. Antal T, Droz M, Gyrgyi G, Racz Z (2001) Phys Rev Lett 87:240601
2. Berman SM (1964) Ann Math Stat 35:502
3. Bouchaud J-P, Mézard M (1997) J Phys A 30:7997
4. Bramwell ST, Holdsworth PCW, Pinton J-F (1998) Nature (London) 396:552
5. Bramwell ST, Christensen K, Fortin J-Y, Holdsworth PCW, Jensen HJ, Lise S, Lopez JM, Nicodemi M, Pinton J-F, Sellitto M (2000) Phys Rev Lett 84:3744
6. Bunde A, Kropp J, Schellnhuber H-J (eds) (2002) The science of disasters-climate disruptions, heart attacks, and market crashes. Springer, Berlin
7. Comtet A, Leboeuf P, Majumdar SN (2007) Phys Rev Lett 98:070404
8. Dahlstedt K, Jensen HJ (2001) J Phys A 34:11193; [Inspec] [ISI]
9. Dean DS, Majumdar SN (2001) Phys Rev E 64:046121
10. Eichner JF, Kantelhardt JW, Bunde A, Havlin S (2006) Phys Rev E 73:016130
11. Embrechts P, Klüppelberg C, Mikosch T, Karatzas I, Yor M (eds) (1997) Modelling extremal events. Springer, Berlin
12. Finkenstadt B, Rootzen H (2004) Extreme values in finance, telecommunications, and the environment. Chapman and Hall/CRC Press, London
13. Galambos J (1978) The asymptotic theory of extreme order statistics. Wiley, New York
14. Galambos J, Lechner J, Simin E (eds) (1994) Extreme value theory and applications. Kluwer, Dordrecht
15. Gnedenko BV (1998) Theory of probability. CRC, Boca Raton, FL
16. Guclu H, Korniss G (2004) Phys Rev E 69:065104(R)
17. Gumbel EJ (1958) Statistics of extremes. Columbia University Press, New York
18. Leadbetter MR, Lindgren G, Rootzen H (1983) Extremes and related properties of random sequences and processes. Springer, New York
19. Majumdar SN, Comtet A (2004) Phys Rev Lett 92:225501; J Stat Phys 119, 777 (2005)
20. Raychaudhuri S, Cranston M, Przybyla C, Shapir Y (2001) Phys Rev Lett 87:136101
21. Reiss RD, Thomas M (2001) Statistical analysis of extreme values: with applications to insurance, finance, hydrology, and other fields. Birkhäuser, Basel
22. Smith RL (2003) Statistics of extremes, with applications in environment, insurance and finance, chap 1. In: Statistical analysis of extreme values: with applications to insurance, finance, hydrology, and other fields. Birkhäuser, Basel
23. Smith RL, Tawn JA, Yuen HK (1990) Statistics of multivariate extremes. Int Stat Rev 58:47
24. v. Storch H, Zwiers FW (2001) Statistical analysis in climate research. Cambridge University Press, Cambridge

F

Fair Division*

Steven J. Brams
Department of Politics, New York University,
New York, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Single Heterogeneous Good
Several Divisible Goods
Indivisible Goods
Conclusions
Future Directions
Bibliography

Glossary

Efficiency An allocation is efficient if there is no other allocation that is better for one player and at least as good for all the other players.

Envy-freeness An allocation is envy-free if each player thinks it receives at least a tied-for-largest portion and so does not envy the portion of any other player.

Equitability An allocation is equitable if each player values the portion that it receives the same as every other player values its portion.

Definition of the Subject

Cutting a cake, dividing up the property in an estate, determining the borders in an international dispute – such allocation problems are ubiquitous. *Fair division* treats all these problems and many more through a rigorous analysis of procedures for allocating goods, or deciding who wins on what issues, in a dispute.

Introduction

The literature on fair division has burgeoned in recent years, with five academic books [1,13,23,28,32] and one popular book [15] providing overviews. In this review, I will give a brief survey of three different literatures: (i) the division of a single heterogeneous good (e. g., a cake with different flavors or toppings); (ii) the division, in whole or part, of several divisible goods; and (iii) the allocation of several indivisible goods. In each case, I assume the different people, called *players*, may have different preferences for the items being divided.

For (i) and (ii), I will describe and illustrate procedures for dividing divisible goods fairly, based on different criteria of fairness. For (iii), I will discuss problems that arise in allocating indivisible goods, illustrating trade-offs that must be made when different criteria of fairness cannot all be satisfied simultaneously.

Single Heterogeneous Good

The metaphor I use for a single heterogeneous good is a cake, with different flavors or toppings, that cannot be cut into pieces that have exactly the same composition. Unlike a sponge or layer cake, different players may like different pieces – even if they have the same physical size – because they are not homogeneous.

Some of the cake-cutting procedures that have been proposed are discrete, whereby players make cuts with a knife – usually in a sequence of steps – but the knife is not allowed to move continuously over the cake. Moving-knife procedures, on the other hand, permit such continuous movement and allow players to call "stop" at any point at which they want to make a cut or mark.

There are now about a dozen procedures for dividing a cake among three players, and two procedures for dividing a cake among four players, such that each player is assured of getting a most valued or tied-for-most-valued piece, and there is an upper bound on the number of cuts that must be made [16]. When a cake is so divided, no

* Adapted from Barry R. Weingast and Donald Wittman (eds) *Oxford Handbook of Political Economy* (Oxford University Press, 2006) by permission of Oxford University Press.

player will envy another player, resulting in an *envy-free division*.

In the literature on cake-cutting, two assumptions are commonly made:

1. The goal of each player is to maximize the minimum-size piece (*maximin piece*) that he or she can guarantee for himself or herself, regardless of what the other players do. To be sure, a player might do better by not following such a *maximin strategy*; this will depend on the strategy choices of the other players. However, all players are assumed to be *risk-averse*: They never choose strategies that might yield them more valued pieces if they entail the possibility of giving them less than their maximin pieces.
2. The preferences of the players over the cake are continuous. Consider a procedure in which a knife moves across a cake from left to right and, at any moment, the piece of the cake to the left of the knife is A and the piece to the right is B. The continuity assumption enables one to use the intermediate-value theorem to say the following: If, for some position of the knife, a player views piece A as being more valued than piece B, and for some other position he or she views piece B as being more valued than piece A, then there must be some intermediate position such that the player values the two pieces exactly the same.

Only two 3-person procedures [2,30], and no 4-person procedure, make an envy-free division with the minimal number of cuts ($n-1$ cuts if there are n players). A cake so cut ensures that each player gets a single connected piece, which is especially desirable in certain applications (e. g., land division).

For two players, the well-known procedure of "I cut the cake, you choose a piece," or "cut-and-choose," leads to an envy-free division if the players choose maximin strategies. The cutter divides the cake 50-50 in terms of his or her preferences. (Physically, the two pieces may be of different size, but the cutter values them the same.) The chooser takes the piece he or she values more and leaves the other piece for the cutter (or chooses randomly if the two pieces are tied in his or her view). Clearly, these strategies ensure that each player gets at least half the cake, as he or she values it, proving that the division is envy-free.

But this procedure does not satisfy certain other desirable properties [7,22]. For example, if the cake is, say, half vanilla, which the cutter values at 75 percent, and half chocolate, which the chooser values at 75 percent, a "pure" vanilla-chocolate division would be better for the cutter than the divide-and-choose division, which gives him or her exactly 50% percent of the value of the cake.

The moving-knife equivalent of "I cut, you choose" is for a knife to move continuously across the cake, say from left to right. Assume that the cake is cut when one player calls "stop." If each of the players calls "stop" when he or she perceives the knife to be at a 50-50 point, then the first player to call "stop" will produce an envy-free division if he or she gets the left piece and the other player gets the right piece. (If both players call "stop" at the same time, the pieces can be randomly assigned to the two players.)

To be sure, if the player who would truthfully call "stop" first knows the other player's preference and delays calling "stop" until just before the knife would reach the other player's 50-50 point, the first player can obtain a greater-than-50-percent share on the left. However, the possession of such information by the cutter is not generally assumed in justifying cut-and-choose, though it does not undermine an envy-free division.

Surprisingly, to go from two players making one cut to three players making two cuts cannot be done by a discrete procedure if the division is to be envy-free.[1] The 3-person discrete procedure that uses the fewest cuts is one discovered independently by John L. Selfridge and John H. Conway about 1960; it is described in, among other places, Brams and Taylor (1996) and Robertson and Webb (1998) and requires up to five cuts.

Although there is no discrete 4-person envy-free procedure that uses a bounded number of cuts, Brams, Taylor, and Zwicker (1997) and Barbanel and Brams (2004) give moving-knife procedures that require up to 11 and 5 cuts, respectively. The Brams-Taylor-Zwicker (1997) procedure is arguably simpler because it requires fewer simultaneously moving knives. Peterson and Su (2002) give a 4-person envy-free moving-knife procedure for chore division, whereby each player thinks he or she receives the least undesirable chores, that requires up to 16 cuts.

To illustrate ideas, I describe next the Barbanel–Brams [2] 3-person, 2-cut envy-free procedure, which is based on the idea of squeezing a piece by moving two knives simultaneously. The Barbanel–Brams [2] 4-person, 5-cut envy-free procedure also uses this idea, but it is considerably more complicated and will not be described here.

The latter procedure, however, is not as complex as Brams and Taylor's [12] general n-person discrete procedure. Their procedure illustrates the price one must pay for an envy-free procedure that works for all n, because it places no upper bound on the number of cuts that are required to produce an envy-free division; this is also true of other n-person envy-free procedures [25,27]. While the

[1] [28], pp. 28–29; additional information on the minimum numbers of cuts required to give envy-freeness is given in [19], and [29].

number of cuts needed depends on the players' preferences over the cake, it is worth noting that Su's [31] approximate envy-free procedure uses the minimal number of cuts at a cost of only small departures from envy-freeness.[2]

I next describe the Barbanel–Brams 3-person, 2-cut envy-free procedure, called the *squeezing procedure* [2]. I refer to players by number – player 1, player 2, and so on – calling even-numbered players "he" and odd-numbered players "she." Although cuts are made by two knives in the end, initially one player makes "marks," or virtual cuts, on the line segment defining the cake; these marks may subsequently be changed by another player before the real cuts are made.

Squeezing procedure. A referee moves a knife from left to right across a cake. The players are instructed to call "stop" when the knife reaches the 1/3 point for each. Let the first player to call "stop" be player 1. (If two or three players call "stop" at the same time, randomly choose one.) Have player 1 place a mark at the point where she calls "stop" (the right boundary of piece A in the diagram below), and a second mark to the right that bisects the remainder of the cake (the right boundary of piece B below). Thereby player 1 indicates the two points that, for her, trisect the cake into pieces A, B, and C, which will be assigned after possible modifications.

```
   A       B       C
/-----|-----|-----/
        1       1
```

Because neither player 2 nor player 3 called "stop" before player 1 did, each of players 2 and 3 thinks that piece A is at most 1/3. They are then asked whether they prefer piece B or piece C. There are three cases to consider:

1. If players 2 and 3 each prefer a different piece – one player prefers piece B and the other piece C – we are done: Players 1, 2, and 3 can each be assigned a piece that they consider to be at least tied for largest.
2. Assume players 2 and 3 both prefer piece B. A referee places a knife at the right boundary of B and moves it to the left. At the same time, player 1 places a knife at the left boundary of B and moves it to the right in such a way that the value of the cake traversed on the left (by B's knife) and on the right (by the referee's knife) are equal for player 1. Thereby pieces A and C increase equally in player 1's eyes. At some point, piece B will be diminished sufficiently to a new piece, labeled B′ – in either player 2's or player 3's eyes – to tie with either piece A′ or C′, the enlarged A and C pieces. Assume player 2 is the first, or tied for the first, to call "stop" when this happens; then give player 3 piece B′, which she still thinks is the most valued or the tied-for-most-valued piece. Give player 2 the piece he thinks ties for the most value with piece B′ (say, piece A′), and give player 1 the remaining piece (piece C′), which she thinks ties for the most value with the other enlarged piece (A′). Clearly, each player will think that he or she received at least a tied-for-most-valued piece.
3. Assume players 2 and 3 both prefer piece C. A referee places a knife at the right boundary of B and moves it to the right. Meanwhile, player 1 places a knife at the left boundary of B and moves it to the right in such a way as to maintain the equality, in her view, of pieces A and B as they increase. At some point, piece C will be diminished sufficiently to C′ – in either player 2's or player 3's eyes – to tie with either piece A′ or B′, the enlarged A and B pieces. Assume player 2 is the first, or the tied for the first, to call "stop" when this happens; then give player 3 piece C′, which she still thinks is the most valued or the tied-for-most-valued piece. Give player 2 the piece he thinks ties for the most value with piece C′ (say, piece A′), and give player 1 the remaining piece (piece B′), which she thinks ties for the most value with the other enlarged piece (A′). Clearly, each player will think that he or she received at least a tied-for-most-valued piece.

Note that who moves a knife or knives varies, depending on what stage is reached in the procedure. In the beginning, I assume a referee moves a single knife, and the first player to call "stop" (player 1) then trisects the cake. But, at the next stage of the procedure, in cases (2) and (3), it is a referee and player 1 that move two knives simultaneously, "squeezing" what players 2 and 3 consider to be the most-valued piece until it eventually ties, for one of them, with one of the two other pieces.

Several Divisible Goods

Most disputes – divorce, labor-management, merger-acquisition, and international – involve only two parties, but they frequently involve several homogeneous goods that must be divided, or several issues that must be resolved.[3]

[2] See [10,20,26] for other approaches, based on bidding, to the housemates problem discussed in [31]. On approximate solutions to envy-freeness, see [33]. For recent results on pie-cutting, in which radial cuts are made from the center of a pie to divide it into wedge-shaped pieces, see [3,8].

[3] Dividing several homogeneous goods is very different from cake-cutting. Cake-cutting is most applicable to a problem like land division, in which hills, dales, ponds, and trees form an incongruous mix, making it impossible to give all or one thing (e. g., trees) to one play-

As an example of the latter, consider an executive negotiating an employment contract with a company. The issues before them are (1) bonus on signing, (2) salary, (3) stock options, (4) title and responsibilities, (5) performance incentives, and (6) severance pay [14].

The procedure I describe next, called *adjusted winner* (AW), is a 2-player procedure that has been applied to disputes ranging from interpersonal to international ([15]).[4] It works as follows. Two parties in a dispute, after perhaps long and arduous bargaining, reach agreement on (i) what issues need to be settled and (ii) what winning and losing means for each side on each issue. For example, if the executive wins on the bonus, it will presumably be some amount that the company considers too high but, nonetheless, is willing to pay. On the other hand, if the executive loses on the bonus, the reverse will hold.

Thus, instead of trying to negotiate a specific compromise on the bonus, the company and the executive negotiate upper and lower bounds, the lower one favoring the company and the upper one favoring the executive. The same holds true on other issues being decided, including non-monetary ones like title and responsibilities.

Under AW, each side will always win on some issues. Moreover, the procedure guarantees that both the company and the executive will get at least 50% of what they desire, and often considerably more.

To implement AW, each side secretly distributes 100 points across the issues in the dispute according to the importance it attaches to winning on each. For example, suppose that the company and the executive distribute their points as follows, illustrating that the company cares more about the bonus than the executive (it would be a bad precedent for it to go too high), whereas the reverse is true for severance pay (the executive wants to have a cushion in the event of being fired):

	Issues	Company	Executive
1.	Bonus	*10*	5
2.	Salary	35	*40*
3.	Stock Options	15	*20*
4.	Title and Responsibilities	*15*	10
5.	Performance Incentives	*15*	5
6.	Severance Pay	10	*20*
	Total	100	100

The italicized figures show the side that wins initially on each issue by placing more points on it. Notice that whereas the company wins a total of $10 + 15 + 15 = 40$ of its points, the executive wins a whopping $40+20+20 = 80$ of its points.

This outcome is obviously unfair to the company. Hence, a so-called *equitability adjustment* is necessary to equalize the points of the two sides. This adjustment transfers points from the initial winner (the executive) to the loser (the company).

The key to the success of AW – in terms of a mathematical guarantee that no win-win potential is lost – is to make the transfer in a certain order (for a proof, see [13], pp. 85–94). That is, of the issues initially won by the executive, look for the one on which the two sides are in closest agreement, as measured by the quotient of the winner's points to the loser's points. Because the winner-to-loser quotient on the issue of salary is $40/35 = 1.14$, and this is smaller than on any other issue on which the executive wins (the next-smallest quotient is $20/15 = 1.33$ on stock options), some of this issue must be transferred to the company.

But how much? The point totals of the company and the executive will be equal when the company's winning points on issues 1, 4, and 5, plus x percent of its points on salary (left side of equation below), equal the executive's winning points on issues 2, 3, and 6, minus x percent of its points on salary (right side of equation):

$$\begin{aligned} 40 + 35x &= 80 - 40x \\ 75x &= 40\,. \end{aligned}$$

Solving for x gives $x = 8/15 \approx 0.533$. This means that the executive will win about 53% on salary, and the company will lose about 53% (i. e., win about 47%), which is almost a 50-50 compromise between the low and high figures they negotiated earlier, only slightly favoring the executive.

This compromise ensures that both the company and the executive will end up with exactly the same total number of points after the equitability adjustment:

$$40 + 35(.533) = 80 - 40(.533) \approx 58.7\,.$$

On all other issues, either the company or the executive gets its way completely (and its winning points), as it should since it valued these issues more than the other side.

Thus, AW is essentially a winner-take-all procedure, except on the one issue on which the two sides are closest and which, therefore, is the one subject to the equitability adjustment. On this issue a split will be necessary, which

er. By contrast, in property division it is possible to give all of one good to one player. Under certain conditions, 2-player cake division, and the procedure to be discussed next (adjusted winner), are equivalent [22].

[4] A website for AW can be found at http://www.nyu.edu/projects/adjustedwinner. Procedures applicable to more than two players are discussed in [13,15,23,32].

will be easier if the issue is a quantitative one, like salary, than a more qualitative one like title and responsibilities.[5]

Still, it should be possible to reach a compromise on an issue like title and responsibilities that reflects the percentages the relative winner and relative loser receive (53% and 47% on salary in the example). This is certainly easier than trying to reach a compromise on each and every issue, which is also less efficient than resolving them all at once according to AW.[6]

In the example, each side ends up with, in toto, almost 59% of what it desires, which will surely foster greater satisfaction than would a 50-50 split down the middle on each issue. In fact, assuming the two sides are truthful, there is no better split for both, which makes the AW settlement *efficient*.

In addition, it is *equitable*, because each side gets exactly the same amount above 50%, with this figure increasing the greater the differences in the two sides' valuations of the issues. In effect, AW makes optimal trade-offs by awarding issues to the side that most values them, except as modified by the equitability adjustment that ensures that both sides do equally well (in their own subjective terms, which may not be monetary). On the other hand, if the two sides have unequal claims or entitlements – as specified, for example, in a contract – AW can be modified to give each side shares of the total proportional to its specified claims.

Can AW be manipulated to benefit one side? It turns out that exploitation of the procedure by one side is practically impossible unless that side knows exactly how the other side will allocate its points. In the absence of such information, attempts at manipulation can backfire miserably, with the manipulator ending up with less than the minimum 50 points its honesty guarantees it [13,15].

While AW offers a compelling resolution to a multi-issue dispute, it requires careful thought to delineate what the issues being divided are, and tough bargaining to determine what winning and losing means on each. More specifically, because the procedure is an additive point scheme, the issues need to be made as independent as possible, so that winning or losing on one does not substantially affect how much one wins or loses on others. To the degree that this is not the case, it becomes less meaningful to use the point totals to indicate how well each side does.

[5] AW may require the transfer of more than one issue, but at most one issue must be divided in the end.

[6] A procedure called *proportional allocation* (PA) awards issues to the players in proportion to the points they allocate to them. While inefficient, PA is less vulnerable to strategic manipulation than AW, with which it can be combined ([13], pp. 75–80).

The half dozen issues identified in the executive-compensation example overlap to an extent and hence may not be viewed as independent (after all, might not the bonus be considered part of salary?). On the other hand, they might be reasonably thought of as different parts of a compensation *package*, over which the disputants have different preferences that they express with points. In such a situation, losing on the issues you care less about than the other side will be tolerable if it is balanced by winning on the issues you care more about.

Indivisible Goods

The challenge of dividing up indivisible goods, such as a car, a boat, or a house in a divorce, is daunting, though sometimes such goods can be shared (usually at different times). The main criteria I invoke are *efficiency* (there is no other division better for everybody, or better for some players and not worse for the others) and *envy-freeness* (each player likes its allocation at least as much as those that the other players receive, so it does not envy anybody else). But because efficiency, by itself, is not a criterion of fairness (an efficient allocation could be one in which one player gets everything and the others nothing), I also consider other criteria of fairness besides envy-freeness, including Rawlsian and utilitarian measures of welfare (to be defined).

I present two paradoxes, from a longer list of eight in [4],[7] that highlight difficulties in creating "fair shares" for everybody. But they by no means render the task impossible. Rather, they show how dependent fair division is on the fairness criteria one deems important and the trade-offs one considers acceptable. Put another way, achieving fairness requires some consensus on the ground rules (i. e., criteria), and some delicacy in applying them (to facilitate trade-offs when the criteria conflict).

I make five assumptions. First, players rank indivisible items but do not attach cardinal utilities to them. Second, players cannot compensate each other with side payments (e. g., money) – the division is only of the indivisible items. Third, players cannot randomize among different allocations, which is a way that has been proposed for "smoothing out" inequalities caused by the indivisibility of items. Fourth, all players have positive values for every item. Fifth, a player prefers one set S of items to a different set T if (i) S has as many items as T and (ii) for every item t in T and not in S, there is a distinct item s in S and not T that the player prefers to t. For example, if a player

[7] For a more systematic treatment of conflicts in fairness criteria and trade-offs that are possible, see [5,6,9,11,18,21].

ranks four items in order of decreasing preference, 1 2 3 4, I assume that it prefers

- the set {1,2} to {2,3}, because {1} is preferred to {3}; and
- the set {1,3} to {2,4}, because {1} is preferred to {2} and {3} is preferred to {4},

whereas the comparison between sets {1,4} and {2,3} could go either way.

Paradox 1. A unique envy-free division may be inefficient.

Suppose there is a set of three players, {A, B, C}, who must divide a set of six indivisible items, {1, 2, 3, 4, 5, 6}. Assume the players rank the items from best to worst as follows:

A : 1 2 3 4 5 6
B : 4 3 2 1 5 6
C : 5 1 2 6 3 4

The unique envy-free allocation to (A, B, C) is ({1,3}, {2,4}, {5,6}), or for simplicity (13, 24, 56), whereby A and B get their best and 3rd-best items, and C gets its best and 4th-best items. Clearly, A prefers its allocation to that of B (which are A's 2nd-best and 4th-best items) and that of C (which are A's two worst items). Likewise, B and C prefer their allocations to those of the other two players. Consequently, the division (13, 24, 56) is envy-free: All players prefer their allocations to those of the other two players, so no player is envious of any other.

Compare this division with (12, 34, 56), whereby A and B receive their two best items, and C receives, as before, its best and 4th-best items. This division *Pareto-dominates* (13, 24, 56), because two of the three players (A and B) prefer the former allocation, whereas both allocations give player C the same two items (56).

It is easy to see that (12, 34, 56) is Pareto-optimal or efficient: No player can do better with some other division without some other player or players doing worse, or at least not better. This is apparent from the fact that the only way A or B, which get their two best items, can do better is to receive an additional item from one of the two other players, but this will necessarily hurt the player who then receives fewer than its present two items. Whereas C can do better without receiving a third item if it receives item 1 or item 2 in place of item 6, this substitution would necessarily hurt A, which will do worse if it receives item 6 for item 1 or 2.

The problem with efficient allocation (12, 34, 56) is that it is not *assuredly* envy-free. In particular, C will envy A's allocation of 12 (2nd-best and 3rd-best items for C) if it prefers these two items to its present allocation of 56 (best and 4th-best items for C). In the absence of information about C's preferences for subsets of items, therefore, we cannot say that efficient allocation (12, 34, 56) is envy-free.[8]

But the real bite of this paradox stems from the fact that not only is inefficient division (13, 24, 56) envy-free, but it is uniquely so – there is no other division, including an efficient one, that guarantees envy-freeness. To show this in the example, note first that an envy-free division must give each player its best item; if not, then a player might prefer a division, like envy-free division (13, 24, 56) or efficient division (12, 34, 56), that does give each player its best item, rendering the division that does not do so envy-possible or envy-ensuring. Second, even if each player receives its best item, this allocation cannot be the only item it receives, because then the player might envy any player that receives two or more items, *whatever* these items are.

By this reasoning, then, the only possible envy-free divisions in the example are those in which each player receives two items, including its top choice. It is easy to check that no efficient division is envy-free. Similarly, one can check that no inefficient division, except (13, 24, 56), is envy-free, making this division uniquely envy-free.

Paradox 2. Neither the Rawlsian maximin criterion nor the utilitarian Borda-score criterion may choose a unique efficient and envy-free division.

Unlike the example illustrating paradox 1, efficiency and envy-freeness are compatible in the following example:

A: 1 2 3 4 5 6
B: 5 6 2 1 4 3
C: 3 6 5 4 1 2

There are three efficient divisions in which (A, B, C) each get two items: (i) (12, 56, 34); (12, 45, 36); (iii) (14, 25, 36). Only (iii) is envy-free: Whereas C might prefer B's 56 allocation in (i), and B might prefer A's 12 allocation in (ii), no player prefers another player's allocation in (iii).

Now consider the following *Rawlsian maximin criterion* to distinguish among the efficient divisions: Choose

[8]Recall that an *envy-free* division of indivisible items is one in which, no matter how the players value subsets of items consistent with their rankings, no player prefers any other player's allocation to its own. If a division is not envy-free, it is *envy-possible* if a player's allocation *may* make it envious of another player, depending on how it values subsets of items, as illustrated for player C by division (12, 34, 56). It is *envy-ensuring* if it causes envy, independent of how the players value subsets of items. In effect, a division that is envy-possible has the potential to cause envy. By comparison, an envy-ensuring division always causes envy, and an envy-free division never causes envy.

a division that maximizes the minimum rank of items that players receive, making a worst-off player as well off as possible.[9] Because (ii) gives a 5th-best item to B, whereas (i) and (iii) give players, at worst, a 4th-best item, the latter two divisions satisfy the Rawlsian maximin criterion.

Between these two, (i), which is envy-possible, is arguably better than (iii), which is envy-free: (i) gives the two players that do not get a 4th-best item their two best items, whereas (iii) does not give B its two best items.[10]

Now consider what a modified Borda count would also give the players under each of the three efficient divisions. Awarding 6 points for obtaining a best item, 5 points for obtaining a 2nd-best item, …, 1 point for obtaining a worst item in the example, (ii) and (iii) give the players a total of 30 points, whereas (i) gives the players a total of 31 points.[11] This criterion, which I call the *utilitarian Borda-score criterion*, gives the nod to division (i); the Borda scores provide a measure of the overall utility or welfare of the players. Thus, neither the Rawlsian maximin criterion nor the utilitarian Borda-score criterion guarantees the selection of the unique efficient and envy-free division of (iii).

Conclusions

The squeezing procedure I illustrated for dividing up a cake among three players ensures efficiency and envy-freeness, but it does not satisfy equitability. Whereas adjusted winner satisfies efficiency, envy-freeness, and equitability for two players dividing up several divisible goods, all these properties cannot be guaranteed if there are more than two players. Finally, the two paradoxes relating to the fair division of indivisible good, which are independent of the procedure used, illustrate new difficulties – that no division may satisfy either maximin or utilitarian notions of welfare and, at the same time, be efficient and envy-free.

[9] This is somewhat different from Rawls's (1971) proposal to maximize the utility of the player with minimum utility, so it might be considered a modified Rawlsian criterion. I introduce a rough measure of utility next with a modified Borda count.

[10] This might be considered a second-order application of the maximin criterion: If, for two divisions, players rank the worst item any player receives the same, consider the player that receives a next-worst item in each, and choose the division in which this item is ranked higher. This is an example of a *lexicographic decision rule*, whereby alternatives are ordered on the basis of a most important criterion; if that is not determinative, a next-most important criterion is invoked, and so on, to narrow down the set of feasible alternatives.

[11] The standard scoring rules for the Borda count in this 6-item example would give 5 points to a best item, 4 points to a 2nd-best item, …, 0 points to a worst item. I depart slightly from this standard scoring rule to ensure that each player obtains some positive value for all items, including its worst choice, as assumed earlier.

Future Directions

Patently, fair division is a hard problem, whatever the things being divided are. While some conflicts are ineradicable, as the paradoxes demonstrate, the trade-offs that best resolve these conflicts are by no means evident. Understanding these may help to ameliorate, if not solve, practical problems of fair division, ranging from the splitting of the marital property in a divorce to determining who gets what in an international dispute.

Bibliography

1. Barbanel JB (2005) The Geometry of Efficient Fair Division. Cambridge University Press, New York
2. Barbanel JB, Brams SJ (2004) Cake Division with Minimal Cuts: Envy-Free Procedures for 3 Persons, 4 Persons, and Beyond. Math Soc Sci 48(3):251–269
3. Barbanel JB, Brams SJ (2007) Cutting a Pie Is Not a Piece of Cake. Am Math Month (forthcoming)
4. Brams SJ, Edelman PH, Fishburn PC (2001) Paradoxes of Fair Division. J Philos 98(6):300–314
5. Brams SJ, Edelman PH, Fishburn PC (2004) Fair Division of Indivisible Items. Theory Decis 55(2):147–180
6. Brams SJ, Fishburn PC (2000) Fair Division of Indivisible Items Between Two People with Identical Preferences: Envy-Freeness, Pareto-Optimality, and Equity. Soc Choice Welf 17(2):247–267
7. Brams SJ, Jones MA, Klamler C (2006) Better Ways to Cut a Cake. Not AMS 35(11):1314–1321
8. Brams SJ, Jones MA, Klamler C (2007) Proportional Pie Cutting. Int J Game Theory 36(3–4):353–367
9. Brams SJ, Kaplan TR (2004) Dividing the Indivisible: Procedures for Allocating Cabinet Ministries in a Parliamentary System. J Theor Politics 16(2):143–173
10. Brams SJ, Kilgour MD (2001) Competitive Fair Division. J Political Econ 109(2):418–443
11. Brams SJ, King DR (2004) Efficient Fair Division: Help the Worst Off or Avoid Envy? Ration Soc 17(4):387–421
12. Brams SJ, Taylor AD (1995) An Envy-Free Cake Division Protocol. Am Math Month 102(1):9–18
13. Brams SJ, Taylor AD (1996) Fair Division: From Cake-Cutting to Dispute Resolution. Cambridge University Press, New York
14. Brams SJ, Taylor AD (1999a) Calculating Consensus. Corp Couns 9(16):47–50
15. Brams SJ, Taylor AD (1999b) The Win-Win Solution: Guaranteeing Fair Shares to Everybody. W.W. Norton, New York
16. Brams SJ, Taylor AD, Zwicker SW (1995) Old and New Moving-Knife Schemes. Math Intell 17(4):30–35
17. Brams SJ, Taylor AD, Zwicker WS (1997) A Moving-Knife Solution to the Four-Person Envy-Free Cake Division Problem. Proc Am Math Soc 125(2):547–554
18. Edelman PH, Fishburn PC (2001) Fair Division of Indivisible Items Among People with Similar Preferences. Math Soc Sci 41(3):327–347
19. Even S, Paz A (1984) A Note on Cake Cutting. Discret Appl Math 7(3):285–296
20. Haake CJ, Raith MG, Su FE (2002) Bidding for Envy-Freeness: A Procedural Approach to *n*-Player Fair Division Problems. Soc Choice Welf 19(4):723–749

21. Herreiner D, Puppe C (2002) A Simple Procedure for Finding Equitable Allocations of Indivisible Goods. Soc Choice Welf 19(2):415–430
22. Jones MA (2002) Equitable, Envy-Free, and Efficient Cake Cutting for Two People and Its Application to Divisible Goods. Math Mag 75(4):275–283
23. Moulin HJ (2003) Fair Division and Collective Welfare. MIT Press, Cambridge
24. Peterson E, Su FE (2000) Four-Person Envy-Free Chore Division. Math Mag 75(2):117–122
25. Pikhurko O (2000)On Envy-Free Cake Division. Am Math Month 107(8):736–738
26. Potthoff RF (2002) Use of Linear Programming to Find an Envy-Free Solution Closest to the Brams-Kilgour Gap Solution for the Housemates Problem. Group Decis Negot 11(5):405–414
27. Robertson JM, Webb WA (1997) Near Exact and Envy-Free Cake Division. Ars Comb 45:97–108
28. Robertson J, Webb W (1998) Cake-Cutting Algorithms: Be Fair If You Can. AK Peters, Natick
29. Shishido H, Zeng DZ (1999) Mark-Choose-Cut Algorithms for Fair and Strongly Fair Division. Group Decis Negot 8(2): 125–137
30. Stromquist W (1980) How to Cut a Cake Fairly. Am Math Month 87(8):640–644
31. Su FE (1999) Rental Harmony: Sperner's Lemma in Fair Division. Am Math Month 106:922–934
32. Young HP (1994) Equity in Theory and Practice. Princeton University Press, Princeton
33. Zeng DZ (2000) Approximate Envy-Free Procedures. Game Practice: Contributions from Applied Game Theory. Kluwer Academic Publishers, Dordrecht, pp. 259–271

Field Computation in Natural and Artificial Intelligence

Bruce J. MacLennan
Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Basic Principles
Field Computation in the Brain
Examples of Field Computation
Field Computers
Future Directions
Bibliography

Glossary

Axon Nerve fiber adapted to the efficient, reliable, active transmission of neural impulses between locations in the brain or body.

Dendrite Nerve fibers adapted to the (primarily passive) sensing and integration of signals from other neurons, which are transmitted to the neuron cell body.

Dirac delta function A *distribution* or *generalized function* that is defined to be infinite at the origin, zero everywhere else, and to have unit area (or volume). More generally, such a function but with its infinite point located elsewhere than the origin. Dirac delta functions are idealized impulses and exist as limit objects in Hilbert spaces.

Eigenfield An eigenfield of a linear operator has the property of passing through the operator with its shape unchanged and only its amplitude possibly modified. Equivalent to an eigenfunction of the operator, but stresses the function's role as a field.

Field A continuous distribution of continuous quantity. Mathematically, an element of an appropriate space, such as a Hilbert space, of continuous-valued functions over a continuum. See also *phenomenological field* and *structural field*.

Field computation Computation in which data are represented by fields, or by other representations that can be mathematically modeled by fields.

Field space A suitably constrained set of fields. Generally field spaces are taken to be subspaces of Hilbert spaces.

Field transformation Functions between field spaces; more generally, functions whose input and/or outputs are fields. Synonymous with *operator* in this article.

Functional A scalar-valued function of functions, and in particular a scalar-valued field transformation.

Idempotency An operation is idempotent when repeating it several times has the same effect as doing it once.

Impulse response The response of a system to an input that is an idealized impulse (a *Dirac delta function*, q.v.).

Microfeature Features of a stimulus or representation that are much smaller and at a lower level than ordinary (macro-)features, which are the sort of properties for which natural languages have words. Typically microfeatures have meaning (are interpretable) only in the context of many other microfeatures. Pixels are examples of microfeatures of images.

Nullcline In a two-dimensional system of differential equations ($\dot{u}_k = f_k(u, v)$, $k = 1, 2$), the lines along which each of the derivatives is zero ($f_k(u, v) = 0$).

Operator A function of functions (i. e., a functions whose inputs and/or outputs are functions), and in particular a function whose inputs and/or outputs are fields. Operators may be linear or nonlinear. Synonymous, in this article, with *field transformation*.

Orthonormal (ON) A set of vectors, fields, or functions is orthonormal if they are: (1) mutually orthogonal (i. e., inner products of distinct elements are 0), and (2) individually normalized (i. e., inner products of elements with themselves are 1).

Phenomenological field A physical distribution of quantity that for practical purposes may be treated mathematically as a continuous distribution of continuous quantity, although it is not so in reality (cf. *structural field*).

Physical realizability A field is physically realizable if it can be represented in some physical medium.

Population coding Neural representation in which a population of neurons jointly represent a stimulus or other information. Each individual neuron is broadly tuned to a range of stimuli, but collectively they can represent a stimulus accurately.

Post–Moore's law computing Refers to computing paradigms that will be important after the expiration of Moore's Law [48], which predicts a doubling of digital logic density every two years.

Projection A systematic pattern of axonal connections from one region of a brain to another.

Radial basis function (RBF) One of a set of real-valued functions, each of whose value decreases with distance from a central point (different for each function). The set as a whole satisfies some appropriate criteria of completeness (ability to approximate a class of functions).

Receptive field The receptive field of a sensory neuron is the set of stimuli to which it responds. By extension, the receptive field of a non-sensory neuron is the set of inputs (from other neurons) to which it responds. Each neuron has a *receptive field profile* which describes the extent to which particular patterns of input stimulate or inhibit activity in the neuron (and so, in effect, its receptive field is fuzzy-boundaried).

Structural field A physical field that is in reality a continuous distribution of continuous quantity (cf. *phenomenological field*).

Synapse A connection between neurons, often from the axon of one to the dendrite of another. Electrical impulses in the pre-synaptic neuron cause neurotransmitter molecules to be secreted into the synapses between the neurons. These chemicals bind to receptors in the post-synaptic neuron membrane, and cause fluctuations in the membrane potential.

Transfer function A function expressing the effect of a linear system on its input, expressed in terms of its effect on the amplitude and phase of each component frequency.

Unit doublet A generalized function that is the derivative of the Dirac delta function (q.v.). It is zero except infinitesimally to the left of the origin, where it is $+\infty$, and infinitesimally to the right of the origin, where it is $-\infty$.

Definition of the Subject

A *field* may be defined as a spatially continuous distribution of continuous quantity. The term is intended to include physical fields, such as electromagnetic fields and potential fields, but also patterns of electrical activity over macroscopic regions of neural cortex. Fields include two-dimensional representations of information, such as optical images and their continuous Fourier transforms, and one-dimensional images, such as audio signals and their spectra, but, as will be explained below, fields are not limited to two or three dimensions. A *field transformation* is a mathematical operation or function that operates on one or more fields in parallel yielding one or more fields as results. Since, from a mathematical standpoint, fields are defined over a continuous domain, field transformations operate with *continuous parallelism*. Some examples of field transformations are point-wise summation and multiplication of fields, Fourier and wavelet transforms, and convolutions and correlations.

Field computation is a model of computation in which information is represented primarily in fields and in which information processing is primarily by means of field transformations. Thus is may be described as *continuously analog computing* (see ► Analog Computation). Field computation may be *feed-forward*, in which one or more fields progress through a series of field transformations from input to output, or it may be *recurrent*, in which there is feedback from later stages of the field computation back to earlier stages. Furthermore, field computations can proceed in discrete sequential steps (similar to digital program execution, but with each step applying a field transformation), or in continuous time according to partial differential equations.

A distinction is often made in science between *structural fields* and *phenomenological fields*. Structural fields are physically continuous distributions of continuous quantity, such as gravitational fields and electromagnetic fields. Phenomenological fields are distributions of quantity that can be treated mathematically as though they are continuous, even if they are not physically continuous. For example, the velocity field of a macroscopic volume of fluid is a phenomenological field, because it is not physically continuous (each discrete molecule has its own velocity), but can be treated as though it is. Similarly, a macro-

scopic charge distribution is a phenomenological field because charge is quantized but can be treated as a continuous quantity for many purposes. Although structural fields are sometimes used, often field computation is concerned with phenomenological fields, that is, with information that can be treated as a continuous distribution of continuous quantity, regardless of whether it is physically continuous. Practically, this means that quantization in both the distribution and the quantity must be sufficiently fine that they can be modeled mathematically by continua.

One of the goals of field computation is to provide a mathematical language for describing information processing in the brain and in future large artificial neural networks intended to exhibit brain-scale intelligence. Neural computation is qualitatively different from ordinary digital computation. Computation on an ordinary computer can be characterized as *deep but narrow*; that is, the processor operates on relatively few data values at a time, but the operations are very rapid, and so many millions of operations can be executed each second. Even on a modern parallel computer, the degree of parallelism is modest, on the order of thousands, whereas even a square millimeter of cortex has at least 146 000 neurons operating in parallel (see p. 51 in [11]). On the other hand, since neurons are quite slow (responding on the order of milliseconds), the "100-Step Rule" says that there can be at most about 100 sequential processing stages between sensory input and response [19]. Therefore, neural computation is *shallow but wide*; that is, it uses relatively few sequential stages, but each operates with a very high degree of parallelism (on the order of many millions). In addition to its speed, modern electronic digital arithmetic is relatively precise compared to the analog computation of neurons (at most about one digit of precision) (see p. 378 in [43]). Therefore we can conclude that neuronal information processing operates according to quite different principles to ordinary digital computing.

It is not unreasonable to suppose that achieving an artificial intelligence that is comparable to the natural intelligence of mammals will require a similar information processing architecture; in any case that seems to be a promising research direction. Therefore we should be aiming toward components with computational capabilities comparable to neurons and densities of at least 15 million per square centimeter. Fortunately, the brain demonstrates that these components do not have to be high-speed, high-precision devices, nor do they have to be precisely connected, for the detailed connections can be established through self-organization and learning. The theory of field computation can contribute in two ways: first, by providing a mathematical framework for understanding information in massively parallel analog computation systems, such as the brain, and second, by suggesting how to exploit relatively homogeneous masses of computational materials (e. g., thin films, new nanostructured materials). For the same reasons, field computers may provide an attractive alternative for "post-Moore's law computing".

Introduction

The term "field computation" dates from 1987 [32], but examples of field computation are much older. For example, G Kirchhoff (1824–87) and others developed the *field analogy method* in which partial differential equation (PDE) problems are solved by setting up an analogous physical system and measuring it [26]. Thus a two-dimensional boundary value problem, for example determining a steady-state temperature or magnetic field distribution, could solved by setting up an analogous system with a conductive sheet or a shallow tank containing an electrolytic solution (see p. 34 in [63]). When the boundary conditions were applied, the system computed the steady-state solution field in parallel and at electronic speed. The resulting potential field could not be displayed directly at that time, and so it was necessary to probe the field at discrete points and plot the equipotential lines; later devices allowed the equipotentials to be traced more or less automatically (see pp. 2-6 in [70]). In either case, setting up the problem and reading out the results were much slower than the field computation, which was comparatively instantaneous. Three-dimensional PDEs were similarly solved with tanks containing electrolytic solutions (see pp. 2-5–6 in [70]). (For more on conductive sheet and electrolytic tanks methods, see Chap. 9 in [65].)

Non-electronic field computation methods were also developed in the nineteenth century, but continued to be used through the first half of the twentieth century to solve the complex PDEs that arise in practical engineering (see pp. 2-8–9 in [70]). For example, so called "rubber-sheet computers" were used to compute the complex electric fields in vacuum tubes. A thin elastic membrane represented the field, and rods or plates pushing the sheet down from above or up from below represented constant negative or positive potential sources. The sheet assumed the shape of the electrical potential field, which could be viewed directly. By altering the rods and plates and observing the effects on the sheet, the engineer could develop an intuitive understanding of the field's dependence on the potential sources. These simple mechanical devices used effectively instantaneous field computation to display the steady-state field's dependence on the boundary conditions.

Electrolytic tanks and conductive and elastic sheets are all examples of the use of structural fields in computation, but other mechanical field computers used discrete approximations of spatially continuous fields, and therefore made use of phenomenological fields. For example, "pin-and-rod systems", which were developed in the nineteenth century, exploited the fact that equipotential lines and flux (or stream) lines always cross at right angles (see pp. 2-9–11 in [70]). A (two-dimensional) field was represented by two arrays of flexible but stiff wires, representing the flux and equipotential lines. At each crossing point was a pin with two perpendicular holes drilled through it, through which the crossing wires passed. The pins were loose enough that they could move on the wires, while maintaining, of course, their relative position and the perpendicular crossings of the wires. To solve a PDE problem (for example, determining the pressure potentials and streamlines of a non-tubulent flow through a nozzle), the edges of the pin-and-rod system were bent to conform to the boundary conditions; the rest of the system adjusted itself to the steady-state solution field. Like the rubber-sheet computers, pin-and-rod systems allowed the solution field to be viewed directly and permitted exploration of the effects on the solution of changes in the boundary conditions.

Through the first half of the twentieth century, *network analyzers* were popular electronic analog computers, which were often used for field computation (see pp. 35–40 in [63]) This was similar to the field analogy method, but a discrete network of resistors or resistive elements replaced such continuous conducting media as the electrolytic tank and conductive sheet. Nevertheless, a sufficiently fine mesh of resistive elements may be treated as a phenomenological field, and network analyzers were used to solve PDE problems (see pp. 2-6–8 in [70]). Boundary conditions were defined by applying voltages to the appropriate locations in the network, and the resulting steady-state field values were determined by measuring the corresponding nodes in the network. As usual, it was possible to monitor the effects of boundary condition changes on particular locations in the field, and to plot them automatically or display them on an oscilloscope.

The field computers discussed so far were suited to determining the steady-state solution of a system of PDEs given specified boundary conditions; as a consequence they were sometimes called *field plotters* or *potential analyzers* (see pp. 2-3 in [70]). These are essentially static problems, although, as we have seen, it was possible to simulate and monitor changes in the (relatively) steady-state solution as a consequence of (relatively slowly) changing boundary conditions. On the other hand, truly dynamic problems, which simulated the evolution of a field in time, could be addressed by *reactive networks*, that is, networks incorporating capacitive and inductive elements as well as resistors (see pp. 2-11–13 in [70]). For example an *RC network analyzer*, which had capacitance at each of the nodes of the resistor network, could solve the diffusion equation, for the charge on the capacitors corresponded to the concentration of the diffusing substance at corresponding locations in the medium. An *RLC network analyzer* had inductance, as well as resistance and capacitance, at each node, and so it was able to address a wider class of PDEs, including wave equations.

Although these twentieth-century field computers were constructed from discrete resistors, capacitors, and inductors, which limited the size of feasible networks, analog VLSI and emerging fabrication technologies will permit the implementation of much denser devices incorporating these and similar field computation techniques (see Sect. "Field Computers").

The following section will present the mathematical foundations and notation for field computation; Hilbert spaces provide the basic mathematical framework. Next we discuss examples of field computation in the brain, especially in its computational maps. Fields appear in a number of contexts, including activity at the axon hillocks, in patterns of axonal connection between areas, and in patterns of synaptic connection to dendrites. The following section presents examples of field computation in the brain and in other natural and artificial systems, including fields for sensorimotor processing, excitable media, and diffusion processes. Next we consider special topics in field computation in cognition, including the separation of information (semantics) from pragmatics, and the analysis of discrete symbols as field excitations. We also consider the relevance of universal mutivariate approximation theorems to general-purpose field computation. Then we discuss hardware specifically oriented toward field computation, including electronic, optical, and chemical technologies. Finally, we consider future directions for field computation research.

Basic Principles

Mathematical Definitions

Mathematically, a field is (generally continuous) function $\phi\colon \Omega \to K$ defined over some bounded domain Ω (often a compact subset of a Euclidean space) and taking values in an algebraic field K. Typically K is the real numbers, but in some applications it is the complex numbers or a vector space (see Sects. "Neuronal Fields", "Diffusion Processes", "Motion in Direction Fields", and "Gabor Wavelets and Coherent States" below).

As usual, the value of a field ϕ at a point $u \in \Omega$ of its domain can be denoted by $\phi(u)$, but we more often use the notation ϕ_u with the same meaning. The latter is especially convenient for time-varying fields. For example, the value of a field ϕ at point u and time t can be denoted by $\phi_u(t)$ rather than $\phi(u, t)$. The entire field at a particular time t is then written $\phi(t)$. As is commonly done in mathematics, we may consider ϕ to be a variable implicitly defined over all $u \in \Omega$. (In this article lower-case Greek letters are usually used for fields. We occasionally use bold-face numbers, such as $\mathbf{0}$ and $\mathbf{1}$, for constant-valued fields; thus $\mathbf{0}_u = 0$ for all $u \in \Omega$. When it is necessary to make the field's domain explicit, we write $\mathbf{0}_\Omega$, $\mathbf{1}_\Omega$, etc.)

For practical field computation (e. g., in natural and artificial intelligence) we are interested in fields that can be realized in some physical medium, which places restrictions on the space of allowable fields. These restrictions vary somewhat for different physical media (e. g., neural cortex or optical fields), but we can specify a few general conditions for *physical realizability*. Generally, fields are defined over a bounded domain, although sometimes we are interested in fields that are extended in time with no prespecified bound (e. g., an auditory signal). Furthermore, since most media cannot represent unbounded field amplitudes, it is reasonable to assume that a field's range of variation is also bounded (e. g., $|\phi_u| \leq B$ for all $u \in \Omega$). In addition, most media will not support unbounded gradients, so the field's derivatives are bounded. Indeed, physically realizable fields are band-limited in both the spatial and temporal domains. Although different assumptions apply in different applications, from a mathematical perspective it is generally convenient to assume that fields are *uniformly continuous square-integrable functions* (defined below), and therefore that they belong to a Hilbert function space. In any case we use the notation $\Phi_K(\Omega)$ for a physically realizable space of K'' valued fields over a domain Ω, and write $\Phi(\Omega)$ when the fields' values are clear from context.

The foregoing considerations suggest that the inner product of fields is an important concept, and indeed it is fundamental to Hilbert spaces Therefore, if ϕ and ψ are two real-valued fields with the same domain, $\phi, \psi \in \Phi(\Omega)$, we define their inner product in the usual way:

$$\langle \phi \mid \psi \rangle = \int_\Omega \phi_u \psi_u \mathrm{d}u \,.$$

If the fields are complex-valued, then we take the complex conjugate of one of the fields:

$$\langle \phi \mid \psi \rangle = \int_\Omega \phi_u^* \psi_u \mathrm{d}u \,.$$

For vector-valued fields $\boldsymbol{\phi}, \boldsymbol{\psi} \in \Phi_{\mathbb{R}^n}(\Omega)$ we may define

$$\langle \boldsymbol{\phi} \mid \boldsymbol{\psi} \rangle = \int_\Omega \boldsymbol{\phi}_u \cdot \boldsymbol{\psi}_u \mathrm{d}u \,,$$

where $\boldsymbol{\phi}_u \cdot \boldsymbol{\psi}_u$ is the ordinary scalar product on the vector space $\mathbb{R}^n$. Finally, the inner-product norm $\|\phi\|$ is defined in the usual way:

$$\|\phi\|^2 = \langle \phi \mid \phi \rangle \,.$$

As previously remarked, the elements of a Hilbert space are required to be square-integrable ("finite energy"): $\|\phi\| < \infty$.

Field Transformations

A *field transformation* or *operator* is any continuous function that maps one or more input fields into one or more output fields. In the simplest case a field transformation $F\colon \Phi(\Omega) \to \Phi(\Omega')$ maps a field in the input space $\Phi(\Omega)$ into a field in the output space $\Phi(\Omega')$.

We do not want to exclude degenerate field transformations, which operate on a field to produce a single real number, for example, or operate on a scalar value to produce a field. (An example of the former is the norm operation, $\|\cdot\|$, and an example of the latter is the operator that produces a constant-valued field over a domain.) In these cases we can consider the inputs or outputs to belong to a space $\Phi(\Omega)$ in which Ω is a singleton set. For example, the real numbers can be treated as fields in

$$\Phi^0 = \Phi_{\mathbb{R}}(\{0\}) \,.$$

Since $\mathbb{R}$ and Φ^0 are isomorphic, we will ignore the difference between them when no confusion can result.

Another class of simple field transformations are the *local transformations*, in which each point of the output field is a function of the corresponding point in the input field. In the simplest case, the same function is applied at each point. Suppose that for input field $\phi \in \Phi_J(\Omega)$, the output field $\psi \in \Phi_K(\Omega)$ is defined $\psi_u = f(\phi_u)$, where $f\colon J \to K$. Then we write $\overline{f}\colon \Phi_J(\Omega) \to \Phi_K(\Omega)$ for the local transformation $\psi = \overline{f}(\phi)$. For example, $\overline{\log}(\phi)$ applies the log function to every element of ϕ and returns field of the results. More generally, we may apply a different function (from a parameterized family) at each point of the input field. Suppose $F\colon \Omega \times J \to K$, then we define $\overline{F}\colon \Phi_J(\Omega) \to \Phi_K(\Omega)$ so that if $\psi = \overline{F}(\phi)$, then $\psi_u = F(u, \phi_u)$.

Field transformations may be linear or nonlinear. The most common linear transformations are *integral operators of Hilbert–Schmidt type*, which are the field analogs of

matrix-vector products. Let $\phi \in \Phi(\Omega)$ and $L \in \Phi(\Omega' \times \Omega)$ be square-integrable fields; then the product $L\phi = \psi \in \Phi(\Omega')$ is defined:

$$\psi_u = \int_\Omega L_{uv}\phi_v \mathrm{d}v \,.$$

L is called the *kernel* of the operator. It is easy to show that physically realizable linear operators have a Hilbert–Schmidt kernel, because physically realizable fields and the operators on them are band-limited [33]. Therefore they can be computed by field products of the form $L\phi$.

According to the Riesz Representation Theorem (e. g., Sect. 12.4 in [9]), a continuous linear functional (real-valued operator) $L\colon \Phi(\Omega) \to \mathbb{R}$ has a *representer*, which is a field $\rho \in \Phi(\Omega)$ such that $L\phi = \langle \rho \mid \phi \rangle$. However, since linear operators are continuous if and only if they are bounded, and since practical field transformations are bounded, all practical linear functionals have representers.

We define multilinear products in the same way. Suppose $\phi_k \in \Phi(\Omega_k)$, for $k = 1, \ldots, n$, and $M \in \Phi(\Omega' \times \Omega_n \times \cdots \Omega_2 \times \Omega_1)$. Then $M\phi_1\phi_2\cdots\phi_n = \psi \in \Phi(\Omega')$ is defined

$$\psi_u = \int_{\Omega_n} \cdots \int_{\Omega_2} \int_{\Omega_1} M_{uv_n\cdots v_2 v_1}\phi_1(v_1)\phi_2(v_2) \cdots \phi_n(v_n)\mathrm{d}v_1\mathrm{d}v_2\cdots\mathrm{d}v_n \,.$$

Again, physically realizable multilinear operators are band limited, and so they can be computed by this kind of multilinear product [33].

Like the field analogs of inner products and matrix-vector products, it is also convenient to define an analog of the outer product. For $\phi \in \Phi(\Omega)$ and $\psi \in \Phi(\Omega')$ we define the outer product $\phi \wedge \psi \in \Phi(\Omega \times \Omega')$ so that $(\phi \wedge \psi)_{(u,v)} = \phi_u\psi_v$, for $u \in \Omega$, $v \in \Omega'$. Inner, outer, and field products are related as follows for $\phi, \chi \in \Phi(\Omega)$ and $\psi \in \Phi(\Omega')$:

$$\phi(\chi \wedge \psi) = \langle \phi \mid \chi \rangle \psi = (\psi \wedge \chi)\phi \,.$$

In the simplest kind of field computation (corresponding to a feed-forward neural network), an input field ϕ is processed through one or more field transformations $F_1, \ldots, F_n$ to yield an output field ψ:

$$\psi = F_n(\cdots F_1(\phi)\cdots) \,.$$

This includes cases in which the output field is the continuously-varying image of a time-varying input field,

$$\psi(t) = F_n(\cdots F_1(\phi(t))\cdots) \,.$$

More complex feed-forward computations may involve additional input, output, and intermediate fields, which might be variable or not.

In an ordinary artificial neural network, the activity y_i of neuron i in one layer is defined by the activities $x_1, \ldots, x_n$ of the neurons in the preceding layer by an equation such as

$$y_i = s\left(\sum_{j=1}^{N} W_{ij}x_j + b_i\right) \,, \tag{1}$$

where W_{ij} is the weight or strength of the connection from neuron j to neuron i, b_i is a *bias term*, and $s\colon \mathbb{R} \to \mathbb{R}$ is a *sigmoid function*, that is, a non-decreasing, bounded continuous function. (The hyperbolic tangent is a typical example.) The field computation analog is obtained by taking the number of neurons in each layer to the continuum limit. That is, the activities ψ_u in one neural field ($u \in \Omega'$) are defined by the values ϕ_v in the input field ($v \in \Omega$) by this equation:

$$\psi_u = \int_\Omega L_{uv}\phi_v \mathrm{d}v + \beta_u \,,$$

where $L \in \Phi(\Omega' \times \Omega)$ is an *interconnection field* and $\beta \in \Phi(\Omega')$ is a *bias field*. More compactly,

$$\psi = \bar{s}(L\phi + \beta) \,. \tag{2}$$

Typically, the input is processed through a series of layers, each with its own weights and biases. Analogously, in field computation we may have an N-layer neural field computation, $\phi_k = \bar{s}(L_k\phi_{k-1} + \beta_k)$, $\quad k = 1, \ldots, N$, where $\phi_0 \in \Phi(\Omega_0)$ is the input, $\phi_N \in \Phi(\Omega_N)$ is the output, $L_k \in \Phi(\Omega_k \times \Omega_{k-1})$ are the interconnection fields, and $\beta_k \in \Phi(\Omega_k)$ are the bias fields. Other examples of neural-network style field computing are discussed later (Sect. "Examples of Field Computation").

Many important field computation algorithms are iterative, that is, they sequentially modify one or more fields at discrete moments of time. They are analogous to ordinary computer programs, except that the variables contain fields rather than scalar quantities (integers, floating-point numbers, characters, etc., and arrays of these). Since the current value of a field variable may depend on its previous values, iterative field computations involve *feedback*. Examples of iterative algorithms include field computation analogs of neural network algorithms that adapt in discrete steps (e. g., ordinary back-propagation), and re-

current neural networks, which have feedback from later layers to earlier layers.

Analog field computers, like ordinary analog computers, can operate in continuous time, defining the continuous evolution of field variable by differential equations. For instance, $\dot{\phi} = F(\phi)$ is a simple first-order field-valued differential equation, which can be written $\mathrm{d}\phi_u(t)/\mathrm{d}t = F_u[\phi(t)]$. An example is the familiar diffusion equation $\dot{\phi} = k^2\nabla^2\phi$.

Continuously varying fields arise in a number of contexts in natural and artificial intelligence. For example, sensorimotor control (in both animals and robots) depends on the processing of continuously varying input fields (e. g., visual images or auditory signals) and their transformation into continuously varying output signals (e. g., to control muscles or mechanical effectors). One of the advantages of field computing for these applications is that the fields are processed in parallel, as they are in the brain. Often we find continuous field computation in optimization problems, in adaptation and learning, and in the solution of other continuous problems. For example, a field representing the interpretation of perceptual data (such as stereo disparity) may be continuously converging to the optimal interpretation or representation of the data.

Optimization problems are sometimes solved by continuous gradient ascent or descent on a potential surface defined by a functional F over a field space ($F\colon \Phi(\Omega) \to \mathbb{R}$), where $F(\phi)$ defines the "goodness" of solution ϕ. Gradient ascent is implemented by $\dot{\phi} = r\nabla F(\phi)$, where r is the rate of ascent. This and other examples are discussed in Sect. "Gradient Processes", but the use of the gradient raises the issue of the derivatives of field transformations, such as F, which we now address.

Derivatives of Field Transformations

Since fields are functions, field spaces are function spaces (generally, Hilbert spaces), and therefore the derivatives of field transformations are the derivatives of operators over function spaces (see § 251G in [42]). There are two common definitions of the differentiation of operators on Hilbert spaces (more generally, on Banach spaces), the Fréchet and the Gâteaux derivatives, which turn out to be the same for field transformations [33]. Therefore suppose that $F\colon \Phi(\Omega) \to \Phi(\Omega')$ is a field transformation and that U is an open subset of $\Phi(\Omega)$. Then $D \in \mathcal{L}(\Phi(\Omega), \Phi(\Omega'))$, the space of bounded linear operators from $\Phi(\Omega)$ to $\Phi(\Omega')$, is called the *Fréchet differential* of F at $\phi \in U$ if for all $\alpha \in \Phi(\Omega)$ such that $\phi + \alpha \in U$ there is an $E\colon \Phi(\Omega) \to \Phi(\Omega')$ such that,

$$F(\phi + \alpha) = F(\phi) + D(\alpha) + E(\alpha)$$

and

$$\lim_{\|\alpha \to 0\|} \frac{\|E(\alpha)\|}{\|\alpha\|} = 0 \,.$$

The *Fréchet derivative* $F'\colon \Phi(\Omega) \to \mathcal{L}(\Phi(\Omega), \Phi(\Omega'))$ is defined by $F'(\phi) = D$, which is the locally linear approximation to F at ϕ.

Similarly $\mathrm{d}F\colon \Phi(\Omega) \times \Phi(\Omega) \to \Phi(\Omega')$ is a *Gâteaux derivative* of F if for all $\alpha \in U$ the following limit exists:

$$\mathrm{d}F(\phi, \alpha) = \lim_{s \to 0} \frac{F(\phi + s\alpha) - F(\phi)}{s} = \left.\frac{\mathrm{d}F(\phi + s\alpha)}{\mathrm{d}s}\right|_{s=0} .$$

If the Fréchet derivative exists, then the two derivatives are identical, $\mathrm{d}F(\phi, \alpha) = F'(\phi)(\alpha)$ for all $\alpha \in \Phi(\Omega)$.

Based on the analogy with finite-dimensional spaces, we define $\nabla F(\phi)$, the gradient of F at ϕ, to be the field $K \in \Phi(\Omega' \times \Omega)$ satisfying $F'(\phi)(\alpha) = K\alpha$ for all α in $\Phi(\Omega)$. That is, $F'(\phi)$ is an integral operator with kernel $K = \nabla F(\phi)$; note that $F'(\phi)$ is an operator but $\nabla F(\phi)$ is a field. The field analog of a directional derivative is then defined:

$$\nabla_\alpha F(\phi) = [\nabla F(\phi)]\alpha = F'(\phi)(\alpha) \,.$$

Because of their importance, it is worth highlighting the gradients of functionals (real-valued operators on fields). According to the preceding definition, the gradient of a functional $F\colon \Phi(\Omega) \to \Phi^0$ will be a two-dimensional field $\nabla F(\phi) \in \Phi(\{0\} \times \Omega)$. (Recall $\Phi^0 = \Phi(\{0\})$.) However, when confusion is unlikely, it is more convenient to define $\nabla F(\phi) = \gamma \in \Phi(\Omega)$, where γ is the representer of $F'(\phi)$. Then $F'(\phi)(\alpha) = \langle \gamma \mid \alpha \rangle = \langle \nabla F(\phi) \mid \alpha \rangle$.

Higher order derivatives of field operators are defined in the obvious way, but it is important to note that each derivative is of "higher type" that the preceding. That is, we have seen that if $F\colon \Phi(\Omega) \to \Phi(\Omega')$, then $\mathrm{d}F\colon \Phi(\Omega)^2 \to \Phi(\Omega')$, where $\Phi(\Omega)^2 = \Phi(\Omega) \times \Phi(\Omega)$. Similarly, $\mathrm{d}^n F\colon \Phi(\Omega)^{n+1} \to \Phi(\Omega')$. Also, as $F'\colon \Phi(\Omega) \to \mathcal{L}(\Phi(\Omega), \Phi(\Omega'))$, so $F''\colon \Phi(\Omega) \to \mathcal{L}(\Phi(\Omega), \mathcal{L}(\Phi(\Omega), \Phi(\Omega')))$ and, in general,

$$F^{(n)}\colon \Phi(\Omega) \to \overbrace{\mathcal{L}(\Phi(\Omega), \mathcal{L}(\Phi(\Omega), \cdots, \mathcal{L}(\Phi(\Omega),}^{n} \Phi(\Omega'))\cdots)) \,.$$

Corresponding to higher-order derivatives are higher-order gradients:

$$\begin{aligned} \mathrm{d}F^n(\phi, \alpha_1, \ldots, \alpha_n) &= \nabla^n F(\phi)\alpha_1 \cdots \alpha_n \\ &= \nabla^n F(\phi)(\alpha_n \wedge \cdots \wedge \alpha_1) \\ &= \nabla_{\alpha_n} \cdots \nabla_{\alpha_1} F(\phi) \,. \end{aligned}$$

For reference, we state the chain rules for Fréchet and Gâteaux derivatives:

$$(F \circ G)'(\phi)(\alpha) = F'[G(\phi)][G'(\phi)(\alpha)] \,, \tag{3}$$

$$\mathrm{d}(F \circ G)(\phi, \alpha) = \mathrm{d}F[G(\phi), \mathrm{d}G(\phi, \alpha)] \,. \tag{4}$$

Just as a real function can be expanded in a Taylor series around a point to obtain a polynomial approximation, there is a corresponding theorem in functional analysis that allows the expansion of an operator around a fixed field. This suggests an approach to general-purpose computation based on *field polynomials* [32], but there are also other approaches suggested by neural networks (see Sect. "Universal Approximation" below). We begin with a formal statement of the theorem.

Theorem 1 (Taylor) *Suppose that U is any open subset of $\Phi(\Omega)$ and that $F\colon \Phi(\Omega) \to \Phi(\Omega')$ is a field transformation that is C^n in U (that is, its first n derivatives exist). Let $\phi \in U$ and $\alpha \in \Phi(\Omega)$ such that $\phi + s\alpha \in U$ for all $s \in [0, 1]$. Then:*

$$F(\phi + \alpha) = \sum_{k=0}^{n-1} \frac{d^k F(\phi, \overbrace{\alpha, \ldots, \alpha}^{k})}{k!} + R_n(\phi, \alpha) \,,$$

where

$$R_n(\phi, \alpha) = \int_0^1 \frac{(1-s)^{n-1} d^n F(\phi + s\alpha, \overbrace{\alpha, \ldots, \alpha}^{n})}{(n-1)!} \mathrm{d}s \,.$$

Here the "zeroth derivative" is defined in the obvious way: $\mathrm{d}^0 F(\phi) = F(\phi)$.

If the first n gradients exist (as they will for physically realizable fields and transformations), then the Taylor approximation can be written:

$$F(\phi + \alpha) = F(\phi) + \sum_{k=1}^{n} \frac{\nabla_\alpha^k F(\phi)}{k!} + R_n(\phi, \alpha) \,.$$

However, $\nabla_\alpha^k F(\phi) = \nabla^k F(\phi) \alpha^{(k)}$, where $\alpha^{(k)}$ is the k-fold outer product:

$$\alpha^{(k)} = \overbrace{\alpha \wedge \alpha \wedge \cdots \wedge \alpha}^{k} \,.$$

If we define the fields $\Gamma_k = \nabla^k F(\phi)$, then we can see this approximation as a "field polynomial":

$$F(\phi + \alpha) \approx F(\phi) + \sum_{k=1}^{n} \frac{\Gamma_k \alpha^{(k)}}{k!} \,.$$

Such an approximation may be computed by a field analog of "Horner's rule", which is especially appropriate for computation in a series of layers similar to a neural network. Thus $F(\phi + \alpha) \approx G_0(\alpha)$, where

$$G_k(\alpha) = \Gamma_k + \frac{G_{k+1}(\alpha)}{k+1} \alpha \,,$$

for $k = 0, \ldots, n$, $\Gamma_0 = F(\phi)$, and $G_{n+1}(\alpha) = 0$.

Field Computation in the Brain

There are a number of contexts in mammalian brains in which information representations are usefully treated as fields, and information processing as field computation. These include neuronal cell bodies, patterns of axonal projection, and synapses. Of course, all of these are discrete structures, but in many cases the numbers are sufficiently large (e. g., 146×10^3 neurons/mm^2: see p. 51 in [11]) that the representations are usefully treated as fields; that is, they are *phenomenological fields*). (We omit discussing the intriguing possibility that the brain's electromagnetic field may affect conscious experience [44,50].)

Neuronal Fields

Computational maps, in which significant information is mapped to cortical location, are found throughout the brain [27]. For example, *tonotopic maps* in auditory cortex have systematic arrangements of neurons that respond to particular pitches, and *retinotopic maps* in visual cortex respond systematically to patches of color, edges, and other visual features projected onto the retina. Other *topographic maps* in somatosensory cortex and motor cortex systematically reflect sensations at particular locations in the body, or control motor activity at those locations, respectively. The number of identified maps is very large and there are probably many that have not been identified. And while some are quite large and can be investigated by fMRI and other noninvasive imaging techniques, other are less than a square millimeter in size [27]. However, even a 0.1 mm^2 map may have tens of thousands of neurons, and thus be analyzed reasonably as a field.

In mathematical terms, let $\mathcal{X}$ be a space of features represented by a cortical map. These might be microfeatures of a sensory stimulus (e. g., oriented edges at particular retinal locations) or motor neurons (e. g., controlling muscle fibers in particular locations). These examples are peripheral features, but $\mathcal{X}$ might represent patterns of activity in nonperipheral groups of neurons (e. g., in other cortical maps). Let Ω be a two-dimensional manifold corresponding to a cortical map representing $\mathcal{X}$. There will a piecewise continuous function $\mu\colon \mathcal{X} \to \Omega$ so that $\mu(x)$

is the cortical location corresponding to feature $x \in \mathcal{X}$. The mapping μ may be only piecewise continuous since $\mathcal{X}$ may be of higher dimension that Ω. (This is the reason, for example, that we find stripes in striate cortex; it is a consequence of mapping a higher dimensional space into a lower one.)

Typically, the activity $\phi_{\mu(x)}$ at a cortical location $\mu(x)$ will reflect the degree of presence of the feature x in the map's input. Furthermore, the responses of neurons are often broadly tuned, therefore the response at a location $\mu(x')$ will generally be a decreasing function $r[d(x, x')]$ of the distance $d(x, x')$, where d is some appropriate metric on $\mathcal{X}$. Therefore an input feature x will generate a response field $\phi = \xi(x)$ given by

$$\phi_{\mu(x')} = r[d(x, x')] \,.$$

If a number of features $x_1, \ldots, x_n$ are simultaneously present in the input, then the activity in the map may be a superposition of the activities due to the individual features:

$$\xi(x_1) + \cdots + \xi(x_n).$$

Furthermore, a sensory or other input, represented as a subset $\mathcal{X}' \subset \mathcal{X}$ of the feature space, generates a corresponding field,

$$\xi(\mathcal{X}') = \int_{\mathcal{X}'} \xi(x) \mathrm{d}x$$

(with an appropriate definition of integration for $\mathcal{X}$, which usually can be taken to be a measure space). (See Sect. "Nonlinear Computation via Topographic Maps" for more on computation on superpositions of inputs via topographic maps.)

The preceding discussion of cortical maps refers somewhat vaguely to the "activity" of neurons, which requires clarification. In cortical maps the represented microfeatures are correlated most closely with the location of the neuronal cell body, which often interacts with nearby neurons. Therefore, when a cortical map is treated mathematically as a field, there are several physical quantities that can be interpreted as the field's value ϕ_u at a particular cortical location u. Although the choice depends somewhat on the purpose of the analysis, the most common interpretation of $\phi_u(t)$ will be the instantaneous spiking frequency at time t of the neuron at location u. We will refer to $\phi(t) \in \Phi(\Omega)$ as the *neuronal field* (at time t) associated with the neurons u in the map Ω.

The relative phase of neural impulses is sometimes relevant to neural information processing [25]. For example, the relative phase with which action potentials arrive a neuron's dendrites can determine whether or not the induced post-synaptic potentials add. In these cases it may be convenient to treat neural activity as a complex-valued field, $\psi(t) \in \Phi_{\mathbb{C}}(\Omega)$, which can be written in polar form:

$$\psi(t) = \rho(t) \mathrm{e}^{i\phi(t)} \,.$$

Then the magnitude (or modulus) field $\rho(t)$ may represent the impulse rate and the phase (or argument) field $\phi(t)$ may represent the relative phase of the impulses. That is, $\rho_u(t)$ is the rate of neuron u (at time t) and $\phi_u(t)$ is its phase. For example, in a *bursting neuron* (which generates impulses in clusters), $\rho(t)$ can represent the impulse rate within the clusters and $\phi(t)$ the relative phase of the clusters. More generally, in a complex-valued neuronal field, the phase part may represent microfeatures of stimulus, while the magnitude part represent pragmatic characteristics of the microfeatures, such as their importance, confidence, or urgency. (Such dual representations, comprising semantics and pragmatics, are discussed in Sect. "Information Fields".)

Synaptic and Dendritic Fields

The surface of each neuron's dendritic tree and soma (cell body) is a complicated two-dimensional manifold Ω_{m}, and so the electrical field across the neuron's membrane is naturally treated as a two-dimensional potential field $\phi \in \Phi(\Omega_{\mathrm{m}})$. Synaptic inputs create electrical disturbances in this field, which, to a first approximation, propagate passively according to the cable equations (see pp. 25–31 in [3]). However, there are also nonlinear effects due to voltage-gated ion channels etc. (see p. 381 in [58]). Therefore the membrane field obeys a nonlinear PDE (partial differential equation) dependent on a synaptic input field ϵ:

$$M(\epsilon, \phi, \dot{\phi}, \ddot{\phi}, \ldots) = 0 \,.$$

The electrical field ϕ on the membrane includes the field ϕ_a at the axon hillock $a \in \Omega_{\mathrm{m}}$. This voltage determines the rate at which the neuron generates action potentials (APs, nerve impulses), which constitute the neuron's contribution to a neuronal field. The dependence of the impulse rate r on the membrane field, $r(t) = A_r[\phi(t)]$, which is approximately linear (that is, the rate is proportional to the depolarization, relative to the resting potential, at the axon hillock). To a first approximation, the dendritic tree implements an approximately linear (adaptive) analog filter on its input field [35,36]. Some purposes require a more detailed analysis, which looks at the time-varying action potential $V(t)$, rather than at the instantaneous impulse rate, as a function of the membrane field, $V(t) = A_V[\phi(t)]$.

Many neurons have tens of thousands of synaptic inputs (see p. 304 in [3]), and so these quantitative properties can be treated as a field over a domain Ω, which is a subset of the dendritic membrane. The post-synaptic potential ϵ_s at synapse s is a result of the synaptic efficacy σ_s and the pre-synaptic axonal impulse rate ζ_s. The synaptic efficacy is the composite effect of the number of receptors for the neurotransmitter released by the incoming axon, as well as of other factors, such as the dependence of neurotransmitter flux on the impulse rate. Some learning processes (e. g., long-term potentiation) alter the synaptic efficacy field σ.

However, because synaptic transmission involves the diffusion of neurotransmitter molecules across the synaptic cleft, the subsequent binding to and unbinding from receptors, and the opening and closing of ion channels, the post-synaptic potential is not a simple product, $\epsilon_s = \sigma_s \zeta_s$. Rather, the synaptic system filters the input field. To a first approximation we may analyze it as a linear system S:

$$\begin{pmatrix} \epsilon \\ \dot{\psi} \end{pmatrix} = \mathrm{S}(\sigma) \begin{pmatrix} \zeta \\ \psi \end{pmatrix} ,$$

where ψ represents the internal state of the synaptic system (concentrations of neurotransmitter in the clefts, receptor and ion channel states, etc.). The parameter σ shows the system's dependence on the synaptic efficacies. The preceding equation is an abbreviation for the following system (in which we suppress the σ parameter):

$$\epsilon = S_{11}\zeta + S_{12}\psi ,$$
$$\dot{\psi} = S_{21}\zeta + S_{22}\psi ,$$

in which the products are Hilbert–Schmidt integral operators (that is, the S_{ij} are fields operating on the input and state fields).

Axonal Projection Fields

Bundles of axons form *projections* from one brain region to another; through the pattern of their connections they may effect certain field transformations (explained below). The input is typically a neuronal field $\phi \in \Phi(\Omega)$ defined over the source region Ω. At their distal ends the axons branch and form synapses with the dendritic trees of the neurons in the destination region. Since each axon may form synapses with many destination neurons, and each neuron may receive synapses from many axons, it is convenient to treat all the synapses of the destination neurons as forming one large synaptic system S, where the synaptic efficacies σ_u range over all the synapses u in the destination region, $u \in \Omega'$. Correspondingly we can consider the field $\zeta \in \Phi(\Omega')$ of pre-synaptic inputs ζ_u to all of these synapses. The axons and their synapses define an *axonal projection system* P, which is, to a first approximation, a linear system:

$$\begin{pmatrix} \zeta \\ \dot{\alpha} \end{pmatrix} = \mathrm{P} \begin{pmatrix} \phi \\ \alpha \end{pmatrix} ,$$

where α represents the internal state of the axonal projection system.

The function of axons is to transmit nerve impulses over relatively long distances with no change of amplitude or waveform. However, there is a transmission delay, and different axons in a projection may introduce different delays. Thus an axonal projection may change the phase relationships of the input field, in addition to introducing an overall delay. On the basis of our analysis the axonal projection as a linear system, we can express the Laplace transform Z of the pre-synaptic field $\zeta(t)$ in terms of the transfer function H^{S} of the projection and the Laplace transform Φ of the input field $\phi(t)$:

$$Z(s) = H^{\mathrm{S}}(s)\Phi(s)$$

(where s is the conjugate variable of time). Note that all the variables refer to fields, and so this equation means

$$Z_u(s) = \int_\Omega H^{\mathrm{S}}_{uv}(s)\Phi_v(s)\mathrm{d}v ,$$

where $H^{\mathrm{S}}_{uv}(s) \in \Phi_{\mathbb{C}}(\Omega' \times \Omega)$ is the (complex-valued) transfer function to synapse u from input neuron v. Since the effects of the axons are pure delays, the transfer function is imaginary:

$$H^{\mathrm{S}}_{uv}(s) = \exp(-i\Delta_{uv}s) ,$$

where Δ_{uv} is the delay imposed by the axon from neuron v to synapse u. Thus the delay field $\Delta \in \Phi(\Omega' \times \Omega)$ defines the effect of the axonal projection on the input field.

The system S comprising all the synapses of the destination neurons is also characterized by a transfer function $H^{\mathrm{S}}(s)$; that is, $E(s) = H^{\mathrm{S}}(s)Z(s)$, where $E(s)$ is the Laplace transform of the post-synaptic field $\epsilon(t)$. Therefore the combined effect of the axonal projection and the synapses is $E(s) = H^{\mathrm{SP}}(s)\Phi(s)$, where the composite transfer function is $H^{\mathrm{SP}}(s) = H^{\mathrm{S}}(s)H^{\mathrm{P}}(s)$. Note that this a field equation, which abbreviates

$$H^{\mathrm{SP}}_{uv}(s) = \int_\Omega pH^{\mathrm{S}}_{uw}(s)H^{\mathrm{P}}_{wv}\mathrm{d}w .$$

The transfer function $H^{\mathrm{SP}}_{uv}(s)$ has a corresponding *impulse response* $\eta^{\mathrm{SP}}_{uv}(t)$, which represents the post-synaptic response at u to a mathematical impulse (Dirac delta function) injected at v. (For Dirac delta functions, see Glossary

and Sect. "Approximation of Spatial Integral and Differential Operators".) The impulse response characterizes the effect of signal transmission to u from v as follows:

$$\epsilon_u(t) = \int_{\Omega'} \eta^{\mathrm{SP}}_{uv}(t) \otimes \phi_v(t) \mathrm{d}v \,,$$

where "$\otimes$" represents convolution in the time domain. This may be abbreviated as a field equation, $\epsilon(t) = \eta^{\mathrm{SP}}(t) \otimes \phi(t)$.

Since axonal projections largely determine the *receptive fields* of the destination neurons, it will be worthwhile to consider the relation of the projection field to the neuronal field at the destination region. Therefore, let ψ_w represent the output signal of a destination neuron w in response to an input field ϕ. We may write

$$\psi_w(t) = F_w[\eta^{\mathrm{SP}}(t) \otimes \phi(t)],$$

where F_w represents the (possibly nonlinear) function computed by neuron w on the subset of the post-synaptic signal $\eta^{\mathrm{SP}}(t) \otimes \phi(t)$ in its dendritic tree. Therefore, the destination neuronal field is given by the field equation $\psi = F[\eta^{\mathrm{SP}} \otimes \phi]$. Many neurons behave as "leaky integrators" (see pp. 52–54 in [3]), which are approximately linear, and in these cases the combined effect of the axonal projection, synaptic field, and destination neurons is a linear operator applied to the input signal, $\psi(t) = L\phi(t)$.

Examples of Field Computation

Neural-Network-like Computation

Many neural network approaches to artificial intelligence can be adapted easily to field computation, effectively by taking the number of neurons in a layer to the continuum limit. For example, as discussed in Sect. "Field Transformations", $\psi = \bar{s}(L\phi + \beta)$ (Eq. 2) is the field analog of one layer of a neural net, that is, a continuum neural net, with interconnection field L and bias field β.

Discrete Basis Function Networks *Radial basis function (RBF) networks* are a familiar and useful class of artificial neural networks, which have similarities to neural networks in the brain [29,52]. Indeed, RBF networks are inspired by the observation that many sensory neurons are tuned to a point in sensory space and that their response falls off continuously with distance from that central point (recall Sect. "Neuronal Fields"). RBFs are usually defined over finite-dimensional spaces, but the extension to fields is straight-forward. Therefore we will consider a set of functionals $r_1, r_2, \ldots,$ where $r_j \colon \Phi(\Omega) \to [0,1]$. Typically we restrict our attention to finite sets of basis functionals, but we include the infinite case for generality. The intent is that each r_j is tuned to a different field input η_j, its "focal field", and that $r_j(\phi)$ represents the closeness of ϕ to the focal field η_j.

If all the RBFs have the same *receptive field profile*, that is, the same fall-off of response with increasing distance from the focal field, then we can write $r_j(\phi) = r(\|\phi - \eta_j\|)$, where the receptive field profile is defined by a $r \colon [0,\infty) \to [0,1]$ that is monotonically decreasing with $r(0) = 1$ and $r(x) \longrightarrow 0$ as $x \longrightarrow \infty$.

As is well known, the inner product is frequently used as a measure of similarity. Expanding the difference in terms of the inner product yields:

$$\|\phi - \eta_j\|^2 = \|\phi\|^2 - 2\langle \phi \mid \eta_j \rangle + \|\eta_j\|^2 \,.$$

The inverse relation between the inner product and distance is especially obvious if, as is often the case (see Sect. "Information Fields"), the input and focal fields are normalized ($\|\phi\| = 1 = \|\eta_j\|$); then:

$$\|\phi - \eta_j\|^2 = 2 - 2\langle \phi \mid \eta_j \rangle \,.$$

Therefore, RBFs with an identical fall-off of response can be defined in terms of a fixed function $c \colon [-1,1] \to [0,1]$ applied to the inner product, $r_j(\phi) = c(\langle \phi \mid \eta_j \rangle)$, where the monotonically increasing function c equals 1 when $\phi = \eta_j$ and equals 0 when the fields are maximally different ($\phi = -\eta_j$). That is, for normalized fields $\langle \phi \mid \eta_j \rangle \in [-1,1]$, and so $c(-1) = 0$, $c(1) = 1$.

Such RBFs are closely related to familiar artificial neurons (Eq. 1). Indeed, we may define $r_j(\phi) = c(\langle \eta_j \mid \phi \rangle + b_j)$, where $c \colon \mathbb{R} \to [0,1]$ is a sigmoidal activation function and b_j is the bias term. Here the input ϕ to the neuron is a field, as is its receptive field profile η_j, which is the focal field defined by the neuron's interconnection field.

Generally, neurons are quite broadly tuned, and so individual RBFs do not characterize the input very precisely, but with an appropriate distribution of focal fields the collection of RBFs can characterize the input accurately, a process known as *coarse coding* (e. g., pp. 91–96 in [58]; [59]). Therefore the discrete ensemble of RBFs compute a representation $\mathbf{p}(\phi)$ of the input given by $p_j(\phi) = r_j(\phi)$.

When information is represented in some way we must consider the adequacy of the representation for our information processing goals. In general, it is not necessary that a representation function $\mathbf{p}$ preserve all characteristics and distinctions of the input space; indeed often the function of representation is to extract the relevant features of the input for subsequent processing. Nevertheless it will be worthwhile to consider briefly RBF-like rep-

resentations that do not lose any information. A Hilbert function space is isomorphic (indeed, isometric) to the space ℓ_2 of square-summable sequences; that is, there is a one-to-one correspondence between fields and the infinite sequences of their generalized Fourier coefficients. Therefore let $\beta_1, \beta_2, \ldots$ be any orthonormal (ON) basis of $\Phi(\Omega)$ and define $p_j(\phi) = \langle \beta_j \mid \phi \rangle$. Define $\mathbf{p}\colon \Phi(\Omega) \to \ell_2$ so that $\mathbf{p}(\phi)$ is the infinite sequence of generalized Fourier coefficients, $(p_1(\phi), p_2(\phi), \ldots)$. Mathematically, we can always find an m such that the first m coefficients approximate the fields as closely as we like; practically, physically realizable fields are band-limited, and so they have only a finite number of nonzero Fourier coefficients. Therefore, we may use $\mathbf{p}^m\colon \Phi(\Omega) \to \mathbb{R}^m$ to compute the m-dimensional representations (relative to an understood ON basis):

$$\mathbf{p}^m(\phi) = (p_1(\phi), p_2(\phi), \ldots, p_m(\phi))^{\mathrm{T}} .$$

Continua of Basis Functions In the preceding section we looked at the field computation of a discrete, typically finite, set of basis functionals. This is appropriate when the basis elements are relatively few in number and there is no significant topological relation among them. In the brain, however, large masses of neurons typically have a significant topological relation (e. g., they may form a topographic map (Sect. "Neuronal Fields"), and so we are interested in cases in which each point in an output field ψ is a result of applying a different basis function to the input field. Suppose $\phi \in \Phi(\Omega)$ and $\psi \in \Phi(\Omega')$. For all $u \in \Omega'$ we want $\psi_u = R(u, \phi)$, where $R\colon \Omega' \times \Phi(\Omega) \to \Phi(\Omega')$. That is, R defines a family of functionals in which, for each u, $R(u, —)$ has a different focal field, which varies continuously with u.

For example, suppose we want ψ_u to be an inner-product comparison of ϕ with the focal field η^u: $\psi_u = c(\langle \eta^u \mid \phi \rangle)$. Since $\langle \eta^u \mid \phi \rangle = \int_\Omega \eta^u_v \phi_v \mathrm{d}v$, define the field $\mathrm{H} \in \Phi(\Omega' \times \Omega)$ by $\mathrm{H}_{uv} = \eta^u_v$. Then a point in the output field is given by $\psi_u = c[(\mathrm{H}\phi)_u]$, and the entire field is computed by:

$$\psi = \bar{c}(\mathrm{H}\phi) . \tag{5}$$

This is, of course, the field analog of one layer of a neural net (Eq. 2), but with no bias field. In a similar way we can define a continuum of RBFs: $\psi_u = r(\|\phi - \eta^u\|)$.

Spatial Correlation and Convolution A special case of Eq. (5) rises when all the focal fields η^u are the same shape but centered on different points $u \in \Omega$. That is, $\eta^u_v = \varrho(v - u)$, where $\varrho \in \Phi(\Omega)$ is the common shape of the focal fields (their *receptive field profile*). In this case,

$$\langle \eta^u \mid \phi \rangle = \int_\Omega \varrho(v - u)\phi(v)\mathrm{d}v .$$

This is simply the *cross-correlation* of ϱ and ϕ, which we may write $\varrho \star \phi$. In general,

$$(\psi \star \phi)_u = \int_\Omega \psi(v - u)\phi(v)\mathrm{d}v , \tag{6}$$

which gives the correlation of ψ and ϕ at a relative displacement u. Therefore in this case the RBF field is given by $\psi = \bar{c}(\varrho \star \phi)$. If the receptive field ϱ is symmetric, $\varrho(-x) = \varrho(x)$, then

$$\langle \eta^u \mid \phi \rangle = \int_\Omega \varrho(u - v)\phi(v)\mathrm{d}v ,$$

which is $\varrho \otimes \phi$, the *convolution* of ϱ and ϕ. In general,

$$(\psi \otimes \phi)_u = \int_\Omega \psi(u - v)\phi(v)\mathrm{d}v . \tag{7}$$

Hence $\psi = \bar{s}(\varrho \otimes \phi)$ when ϱ is symmetric. Computation of these fields by means of convolution or correlation rather than by the integral operator (Eq. 5) may be more convenient on field computers that implement convolution or correlation directly.

Approximation of Spatial Integral and Differential Operators Correlation and convolution (Eqs. 6, 7) can be used to implement many useful linear operators, in particular spatial integral and differential operators. Of course these linear operations can be implemented by a field product with the appropriate Hilbert–Schmidt kernel, but convolution and correlation make use of lower dimensional fields than the kernel.

For example, suppose we want to compute the indefinite spatial integral of a field $\phi \in \Phi(\mathbb{R})$. That is, we want to compute $\psi = \int \phi$ defined by $\psi_x = \int_{-\infty}^{x} \phi_y \mathrm{d}y$. This can be computed by the convolution $\psi = \upsilon \otimes \phi$ where υ is the *Heaviside* or *unit step field* on $\mathbb{R}$:

$$\upsilon_x = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} .$$

The Heaviside field is discontinuous, and therefore it may not be physically realizable, but obviously it may be approximated arbitrarily closely by a continuous field.

Spatial differentiation is important in image processing in nervous systems and artificial intelligence systems. In the one-dimesional case, for $\phi \in \Phi(\mathbb{R})$ we want $\phi' \in$

$\Phi(\mathbb{R})$, where $\phi'_u = \mathrm{d}\phi_u/\mathrm{d}u$. To express this as a convolution we may begin by considering the *Dirac delta function* or *unit impulse function* δ, which is the derivative of the unit step function, $\delta(x) = \upsilon'(x)$. This is a *generalized function* or *distribution* with the following properties:

$$\delta(0) = +\infty,$$
$$\delta(x) = 0, x \neq 0,$$
$$\int_{-\infty}^{+\infty} \delta(x)\mathrm{d}x = 1 .$$

Obviously such a function is not physically realizable (more on that shortly), but such functions exist as limit objects in Hilbert spaces. The Dirac delta satisfies the following "sifting property":

$$\phi_x = \int_{-\infty}^{+\infty} \delta(x-y)\phi(y)\mathrm{d}y ;$$

that is, the Dirac delta is an identity for convolution, $\phi = \delta \otimes \phi$. Now observe:

$$\phi'_x = \mathrm{D}_x \int_{-\infty}^{+\infty} \delta(x-y)\phi(y)\mathrm{d}y$$
$$= \int_{-\infty}^{+\infty} \delta'(x-y)\phi(y)\mathrm{d}y ,$$

where δ' is the derivative of the Dirac delta. It is called the *unit doublet* and has the property of being zero everywhere except infinitesimally to the left of the origin, where it is $+\infty$, and infinitesimally to the right of the origin, where it is $-\infty$. Thus the spatial derivative of a field can be computed by convolution with the unit doublet: $\phi' = \delta' \otimes \phi$.

Obviously, neither the unit impulse (Dirac delta) nor the unit doublet is physically realizable, but both may be approximated arbitrarily closely by physically realizable fields. For example, the delta function can be approximated by a sufficiently sharp Gaussian field γ (i. e., $\gamma_x = \sqrt{r/\pi}\exp(-rx^2)$ for sufficiently large r). Corresponding to the sifting property $\phi = \delta \otimes \phi$ we have Gaussian smoothing $\phi \approx \gamma \otimes \phi$, which is a typical effect of the limited bandwidth of physically realizable fields in cortex and other physical media. Similarly, the unit doublet can be approximated by a *derivative of Gaussian (DoG)* field γ', where $\gamma'_x = \mathrm{d}\gamma_x/\mathrm{d}x$. Thus, the spatial derivative can be approximated the convolution $\phi' \approx \gamma' \otimes \phi$. Indeed, in the nervous system we find neurons with approximately DoG receptive field profiles. (These derivative formulas are perhaps more intuitively expressed in terms of correlation, $\phi' = (-\delta') \star \phi \approx (-\gamma') \star \phi$, since this is more easily related to the difference, $\phi_{x+\epsilon} - \phi_{x-\epsilon}$.)

If ψ is a two-dimensional field, $\psi \in \Phi(\mathbb{R}^2)$, it is easy to show that the partial derivative along the first dimension can be computed by convolution with $\delta' \wedge \delta$, and along the second by convolution with $\delta \wedge \delta'$. The partial derivatives may be approximated by convolutions with $\gamma' \wedge \gamma$ and $\gamma \wedge \gamma'$. The divergence of a field can be computed by a two-dimensional convolution with the sum of these fields:

$$\nabla \cdot \psi = (\delta' \wedge \delta + \delta \wedge \delta') \otimes \psi \approx (\gamma' \wedge \gamma + \gamma \wedge \gamma') \otimes \psi .$$

Similarly the gradient is

$$\nabla\psi = [(\delta' \wedge \delta) \otimes \psi]\mathbf{i} + [(\delta \wedge \delta') \otimes \psi]\mathbf{j} ,$$

where $\mathbf{i} \in \Phi_{\mathbb{R}^2}(\mathbb{R}^2)$ is a constant vector field of unit vectors in the x direction, $\mathbf{i}_{(x,y)} = (1,0)$, and $\mathbf{j}$ is a similar field in the y direction. It is approximated by

$$\nabla\psi \approx [(\gamma' \wedge \gamma) \otimes \psi]\mathbf{i} + [(\gamma \wedge \gamma') \otimes \psi]\mathbf{j} . \quad (8)$$

To compute the Laplacian we need the second partial derivatives, but note that for a one-dimensional field $\phi'' = \delta' \otimes (\delta' \otimes \phi) = (\delta' \otimes \delta') \otimes \phi = \delta'' \otimes \phi$, where δ'' is the second derivative of the Dirac function (a "unit triplet"). Hence, for two-dimensional ψ

$$\nabla^2\psi = (\delta'' \wedge \delta + \delta \wedge \delta'') \otimes \psi \approx (\gamma'' \wedge \gamma + \gamma \wedge \gamma'') \otimes \psi, \quad (9)$$

where γ'' is the second derivative of the Gaussian, a typical (inverted) "Mexican hat function" with the center-surround receptive-field profile often found in the nervous system. These formulas extend in the obvious way to higher-dimensional fields.

Change of Field Domain

We have seen that physically realizable linear operators are integral operators, and therefore can be computed by field products of the form $K\phi$. However, the kernel K might not be physically realizable if its dimension is too high. For example, suppose $L\colon \Phi(\Omega) \to \Phi(\Omega)$ is a linear operator on two-dimensional visual images; that is, Ω is a bounded subset of two-dimensional Euclidean space. Its kernel K, satisfying $K\phi = L(\phi)$, will be a four-dimensional field $K \in \Phi(\Omega \times \Omega)$, and therefore physically unrealizable. Therefore we need means for realizing or approximating high-dimensional fields in three or fewer spatial dimensions.

The simplest way to accomplish this is to represent fields of higher-dimensional spaces by corresponding fields over lower dimensional spaces. For example, to represent $\phi \in \Phi(\Omega)$ by $\psi \in \Phi(\Omega')$, suppose $\beta_1, \beta_2, \ldots$ is an ON basis for $\Phi(\Omega)$, as $\eta_1, \eta_2, \ldots$ is for $\Phi(\Omega')$. Then,

let the generalized Fourier coefficients of ϕ be used as the coefficients to compute a corresponding ψ. Observe:

$$\psi = \sum_k \eta_k \langle \beta_k \mid \phi \rangle = \sum_k (\eta_k \wedge \beta_k)\phi \,.$$

(Of course, a finite sum is sufficient for physically realizable fields.) Therefore the change of basis can be implemented by the kernel $K = \sum_k \eta_k \wedge \beta_k$. By this means, any Hilbert–Schmidt operator on two-dimensional fields can be implemented by a physically realizable field product: represent the input by a one-dimensional field, generate the one-dimensional representation of the output by a product with a two-dimensional kernel, and convert this representation to the output field. Specifically, suppose $\phi \in \Phi(\Omega)$, $\psi \in \Phi(\Omega')$, and $L\colon \Phi(\Omega) \to \Phi(\Omega')$ is a Hilbert–Schmidt linear operator. The three-dimensional kernel $\mathrm{H} = \sum_k \eta_k \wedge \beta_k \in \Phi([0,1] \times \Omega)$ will be used to generate a one-dimensional representation of the two-dimensional input, $\mathrm{H}\phi \in \Phi([0,1])$. Similarly, the two-dimensional output will be generated by $\Theta = \sum_j \zeta_j \wedge \eta_j \in \Phi(\Omega' \times [0,1])$, where $\zeta_1, \zeta_2, \ldots$ is an ON basis for $\Phi(\Omega')$. It is easy to show that the required two-dimensional kernel $K \in \Phi([0,1]^2)$ such that $L = \Theta K \mathrm{H}$ is just

$$K = \sum_{jk} \langle \zeta_j \mid L\beta_k \rangle (\eta_j \wedge \eta_k) \,.$$

We have seen (see Sect. "Neural-Network-Like Computation") that field computation can often be implemented by neural-network-style computation on finite-dimensional spaces. For example, a linear field transformation (of Hilbert–Schmidt type) can be factored through the *eigenfield basis*. Let $\epsilon_1, \epsilon_2, \ldots$ be the eigenfields of L with corresponding eigenvalues $e_1, e_2, \ldots$: $L\epsilon_k = e_k \epsilon_k$. The eigenfields can be chosen to be orthonormal (ON), and, since $\Phi(\Omega)$ is a Hilbert space, only a finite number of the eigenvalues are greater than any fixed bound, so ϕ can be approximated arbitrarily closely by a finite sum $\phi \approx \sum_{k=1}^m c_k \epsilon_k$, where $c_k = \langle \epsilon_k \mid \phi \rangle$; that is, ϕ is represented by the finite-dimensional vector $\mathbf{c}$. The discrete set of coefficients $c_1, \ldots, c_m$ is not a field because there is no significant topological relationship among them; also, typically, m is relatively small.

The output ψ is computed by a finite sum, $\psi \approx \sum_{k=1}^m \epsilon_k e_k c_k$. In terms of neural computation, we have a finite set of neurons $k = 1, \ldots, m$ whose receptive field profiles are the eigenfields, so that they compute $e_k c_k = e_k \langle \epsilon_k \mid \phi \rangle$. The outputs of these neurons amplitude-modulate the generation of the individual eigenfields ϵ_k, whose superposition yields the output ψ.

It is not necessary to factor the operator through the eigenfield basis. To see this, suppose $L\colon \Phi(\Omega) \to \Phi(\Omega')$ and that the fields β_k are an ON basis for $\Phi(\Omega)$ and that the fields ζ_j are an ON basis for $\Phi(\Omega')$. Represent the input by a finite-dimensional vector $\mathbf{c}$, where $c_k = \langle \beta_k \mid \phi \rangle$. Then the output ψ can be represented by the finite-dimensional vector $\mathbf{d}$, where $d_j = \langle \zeta_j \mid \psi \rangle$. (Since the input and output spaces are both Hilbert spaces, only a finite number of these coefficients are greater than any fixed bound.) It is easy to show $\mathbf{d} = \mathrm{M}\mathbf{c}$, where $M_{jk} = \langle \zeta_j \mid L\beta_k \rangle$ (the Hilbert–Schmidt theorem). In neural terms, a first layer of neurons with receptive field profiles β_k compute the discrete representation $c_k = \langle \beta_k \mid \phi \rangle$. Next, a layer of linear neurons computes the linear combinations $d_j = \sum_{k=1}^m M_{jk} c_k$ in order to control the amplitudes of the output basis fields in the output superposition $\psi \approx \sum_{j=1}^n d_j \zeta_j$. In this way, an arbitrary linear field transformation may be computed through a neural representation of relatively low dimension.

If a kernel has too high dimension to be physically realizable, it is not necessary to completely factor the product through a discrete space; rather, one or more dimensions can be replaced by a discrete set of basis functions and the others performed by field computation. To see the procedure, suppose we have a linear operator $L\colon \Phi(\Omega) \to \Phi(\Omega')$ with kernel $K \in \Phi(\Omega' \times \Omega)$, where $\Omega = \Omega_1 \times \Omega_2$ is of too great dimension. Let $\psi = K\phi$ and observe

$$\psi_u = \int_\Omega K_{uv} \phi_v \mathrm{d}v = \int_{\Omega_1} \int_{\Omega_2} K_{uxy} \phi_x(y) \mathrm{d}y \mathrm{d}x \,,$$

where we consider $\phi_v = \phi_{xy}$ as a function of y, $\phi_x(y)$. Expand ϕ_x in terms of an ON basis of $\Phi(\Omega_2)$, $\beta_1, \beta_2, \ldots$:

$$\phi_x = \sum_k \langle \phi_x \mid \beta_k \rangle \beta_k \,.$$

Note that

$$\langle \phi_x \mid \beta_k \rangle = \int_{\Omega_2} \phi_{xy} \beta_k(y) \mathrm{d}y = (\phi\beta_k)_x \,,$$

where $\phi\beta_k \in \Phi(\Omega_1)$. Rearranging the order of summation and integration,

$$\begin{aligned} \psi_u &= \sum_k \int_{\Omega_1} \int_{\Omega_2} K_{uxy} \beta_k(y) (\phi\beta_k)_x \mathrm{d}y \mathrm{d}x \\ &= \sum_k [K\beta_k(\phi\beta_k)]_u \,. \end{aligned}$$

Hence, $\psi = \sum_k K\beta_k(\phi\beta_k)$. Let $J_k = K\beta_k$ to obtain a lower-dimensional field computation:

$$L(\phi) = \sum_k J_k(\phi\beta_k) \,.$$

Note that $J_k \in \Phi(\Omega' \times \Omega_1)$ and all the other fields are of lower dimension than $K \in \Phi(\Omega' \times \Omega)$. As usual, for physically realizable fields, a finite summation is sufficient.

We can discretize $\Phi(\Omega_1)$ by a similar process, which also can be extended straightforwardly to cases where several dimensions must be discretized. Normally we will discretize the dimension that will have the fewest generalized Fourier coefficients, given the bandwidth of the input fields.

The foregoing example discretized one dimension of the input space, but it is also possible to discretize dimensions of the output space. Therefore suppose $L\colon \Phi(\Omega) \to \Phi(\Omega')$ with kernel $K \in \Phi(\Omega' \times \Omega)$, where $\Omega' = \Omega_1 \times \Omega_2$ is of too great dimension. Suppose $\zeta_1, \zeta_2, \ldots$ are an ON basis for $\Phi(\Omega_1)$. Consider $\psi_u = \psi_{xy}$ as a function of x, expand, and rearrange:

$$\begin{aligned}\psi_{xy} &= \sum_k \zeta_k(x) \int_{\Omega_1} \zeta_k(x')\psi_{x'y}\mathrm{d}x' \\ &= \sum_k \zeta_k(x) \int_{\Omega} \int_{\Omega_1} \zeta_k(x')K_{x'yv}\mathrm{d}x'\phi_v\mathrm{d}v \\ &= \sum_k \zeta_k(x)[(\zeta_k K)\phi]_y \,.\end{aligned}$$

Hence $\psi = \sum_k \zeta_k \wedge [(\zeta_k K)\phi]$. Let $J_k = \zeta_k K \in \Phi(\Omega_2 \times \Omega)$ and we can express the computation with lower dimensional fields:

$$L(\phi) = \sum_k \zeta_k \wedge J_k \phi \,.$$

Other approaches to reducing the dimension of fields are described elsewhere [33].

The converse procedure, using field computation to implement a matrix vector product, is also useful, since a field computer may have better facilities for field computation than for computing with vectors. Therefore suppose M is an $m \times n$ matrix, $\mathbf{c} \in \mathbb{R}^n$, and that we want to compute $\mathbf{d} = \mathrm{M}\mathbf{c}$ by a field product $\psi = K\phi$. The input vector will be represented by $\phi \in \Phi(\Omega)$, where we choose a field space $\Phi(\Omega)$ for which the first n ON basis elements $\beta_1, \ldots, \beta_n$ are physically realizable. The field representation is given by $\phi = \sum_{k=1}^n c_k \beta_k$. Analogously, the output is represented by a field $\psi \in \Phi(\Omega')$ given by $\psi = \sum_{j=1}^m d_k \zeta_k$, for ON basis fields $\zeta_1, \ldots, \zeta_m$. The required kernel $K \in \Phi(\Omega' \times \Omega)$ is given by

$$K = \sum_{j=1}^m \sum_{k=1}^n M_{ij}(\zeta_j \wedge \beta_k) \,.$$

To see this, observe:

$$\begin{aligned}K\phi &= \sum_{jk} M_{jk}(\zeta_j \wedge \beta_k)\phi \\ &= \sum_{jk} M_{jk}\zeta_j \langle \beta_k \mid \phi \rangle \\ &= \sum_j \zeta_j \sum_k M_{jk} c_k \\ &= \sum_j \zeta_j d_j \,.\end{aligned}$$

Diffusion Processes

Diffusion processes are useful in both natural and artificial intelligence. For example, it has been applied to path planning through a maze [66] and to optimization and constraint-satisfaction problems, such as occur in image processing and motion estimation [45,67]. Natural systems, such as developing embryos and colonies of organisms, use diffusion as a means of massively parallel search and communication.

A simple diffusion equation has the form $\dot{\phi} = d\nabla^2\phi$ with $d > 0$. On a continuous-time field computer that provides the Laplacian operator (∇^2) diffusion can be implemented directly by this equation. With sequential computation, the field will be iteratively updated in discrete steps:

$$\phi := \phi + d\nabla^2\phi \,.$$

If the Laplacian is not provided as a primitive operation, then its effect can be approximated by a spatial convolution with a suitable field ϱ (see Sect. "Spatial Correlation and Convolution"). In sequential computation we may use $\phi := (1-d)\phi + d\varrho \otimes \phi$, where ϱ is an appropriate two-dimensional Gaussian or similarly shaped field. In continuous time, we may use $\dot{\phi} = d\varrho \otimes \phi$, where $\varrho = \gamma'' \wedge \gamma + \gamma \wedge \gamma''$ (Eq. (9), Sect. "Approximation of Spatial Integral and Differential Operators"), where γ is an appropriate one-dimensional Gaussian and γ'' is its second derivative (or similarly shaped fields).

Reaction-diffusion systems combine diffusion in two or more fields with local nonlinear reactions among the fields ► Reaction-Diffusion Computing. A typical reaction-diffusion system over fields $\phi^1, \ldots, \phi^n \in \Phi(\Omega)$ has the form:

$$\begin{aligned}\dot{\phi^1} &= \overline{F_1}(\phi^1, \ldots, \phi^n) + d_1\nabla^2\phi^1, \\ \dot{\phi^2} &= \overline{F_2}(\phi^1, \ldots, \phi^n) + d_2\nabla^2\phi^2, \\ &\vdots \\ \dot{\phi^n} &= \overline{F_n}(\phi^1, \ldots, \phi^n) + d_n\nabla^2\phi^n \,,\end{aligned}$$

where the $d_k > 0$, and the local reactions $\overline{F_k}$ apply at each point $u \in \Omega$ of the fields: $F_k(\phi_u^1, \ldots, \phi_u^n)$. With obvious extension of the notation, this can be written as a differential equation on a vector field:

$$\dot{\boldsymbol{\phi}} = \overline{\mathbf{F}}(\boldsymbol{\phi}) + \mathrm{D}\nabla^2\boldsymbol{\phi} \,,$$

where $\mathrm{D} = \mathrm{diag}(d_1, \ldots, d_n)$ is a diagonal matrix of diffusion rates.

Embryological development and many other biological processes of self-organization are controlled by local reaction to multiple diffusing chemicals (e. g., Chap. 7 in Bar-Yam [5], Chap. 3 in Solé and Goodwin [64]); these are examples of natural field computation, a subject pioneered by AM Turing [71]. For example, simple *activator-inhibitor systems* can generate *Turing patterns*, which are reminiscent of animal skin and hair-coat pigmentation patterns (e. g., Chap. 7 in Bar-Yam [5]). In the simplest case, these involve an activator (α) and an inhibitor (β), which diffuse at different rates, and a nonlinear interaction which increases both when $\alpha > \beta$, and decreases them otherwise (for example p. 668 in [5]):

$$\dot{\alpha} = \frac{k_1\alpha^2}{\beta(1 + k_5\alpha^2)} - k_2\alpha + d_\alpha\nabla^2\alpha,$$
$$\dot{\beta} = k_3\alpha^2 - k_4\beta + d_\beta\nabla^2\beta \,.$$

Reaction-diffusion systems have been applied experimentally in several image-processing applications, where they have been used to restore broken contours, detect edges, and improve contrast (pp. 26–31 in [1]). In general, diffusion accomplishes (high-frequency) noise filtering and the reaction is used for contrast enhancement.

A Adamatzky and his colleagues have used chemical implementation of reaction-diffusion systems to construct Voronoi diagrams around points and other two-dimensional objects (see Chap. 2 in [2]). Voronoi diagrams have been applied to collision-free path planning, nearest-neighbor pattern classification, and many other problems (see pp. 32–33 in [2]). They also demonstrated a chemical field computer on a mobile robot to implement a reaction-diffusion path planning (see Chap. 4 in [2]).

Excitable media are an important class of reaction-diffusion system, which are found, for example, in the brain, cardiac tissue, slime mold aggregation, and many other natural systems. In the simplest cases these comprise an *excitation field* ϵ and a *recovery field* ρ coupled by local nonlinear reactions:

$$\dot{\epsilon} = \overline{F}(\epsilon, \rho) + d_\epsilon\nabla^2\epsilon \,,$$
$$\dot{\rho} = \overline{G}(\epsilon, \rho) + d_\rho\nabla^2\rho \,.$$

Typically $G(e, r)$ is positive for large e and negative for large r, while along the *nullcline* $F(e, r) = 0$, r has a roughly cubic dependence on e, with $F(e, r) < 0$ for large values of r and > 0 for small ones. The intersection of the nullclines defines the system's stable state, and small perturbations return to the stable state. However excitation above a threshold will cause the excitation to increase to a maximum, after which the system becomes first refractory (unexcitable), then partially excitable with an elevated threshold, and finally back to its excitable, resting state. Excitation spreads to adjacent regions, but the refractory property assures that propagation takes the form of a unidirectional wave of constant amplitude. Characteristic circular and spiral waves appear in two-dimensional media. Excitable media are useful for rapid, efficient communication. For example, masses of slime mold amoebas (*Dictyostelium discoideum*) act as an excitable medium in which the propagating waves accelerate aggregation of the amoebas into a mound (see pp. 12–24 in [64]).

Many self-organizing systems and structures in biological systems involve reaction-diffusion processes, chemical gradients, excitable media, and other instances of field computation.

For example, J-L Deneubourg [15] has described the construction of equally-spaced pillars in termite nests in terms of three interrelated two-dimensional fields: ϕ, the concentration of cement pheromone in the air, σ, the amount of deposited cement with active pheromone, and τ the density of termites carrying cement (see also pp. 188–193 in [8], pp. 399–400 in [10], and pp. 151–157 in [64]). The amount of deposited cement with pheromone increases as it is deposited by the termites and decreases as the pheromone evaporates into the air: $\dot{\sigma} = k_1\tau - k_2\sigma$. The pheromone in the air is increased by this evaporation, but also decays and diffuses at specified rates: $\dot{\phi} = k_2\sigma - k_4\phi + d_\phi\nabla^2\phi$. Laden termites enter the system at a uniform rate r, deposit their cement (k_1), wander a certain amount (modeled by diffusion at rate d_τ), but also exhibit *chemotaxis*, that is, motion up the gradient of pheromone concentration:

$$\dot{\tau} = r - k_1\tau + d_\tau\nabla^2\tau - k_5\nabla\cdot(\tau\overline{\times}\nabla\phi) \,,$$

where $\overline{\times}$ represents the point-wise (local) product, $(\phi\overline{\times}\psi)_u = \phi_u\psi_u$. See Fig. 1 for this model expressed as a field computation.

In addition to reaction-diffusion systems, chemical gradients, chemotaxis, and other field processes are essential to self-organization in morphogenesis, which can be understood in terms of field computation [14].

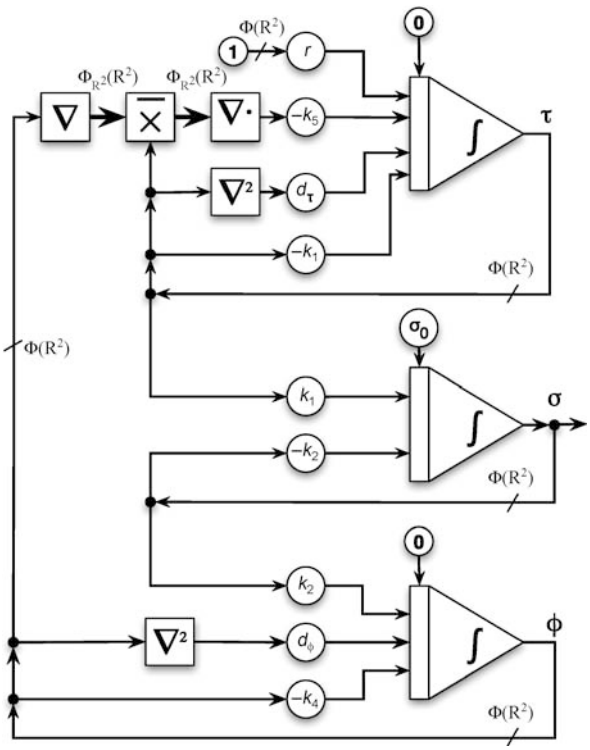

Field Computation in Natural and Artificial Intelligence, Figure 1
Field computation of Deneubourg's model of pillar construction by termites

Motion in Direction Fields

For an example of field computation in motor control, we may consider Georgopoulos' [21] explanation of the population coding of direction. In this case the feature space $\mathcal{D}$ represents directions in three-dimensional space, which we may identify with normalized three-dimensional vectors $\mathbf{d} \in \mathcal{D}$. Each neuron $u \in \Omega$ has a preferred direction $\boldsymbol{\eta}_u \in \mathcal{D}$ to which it responds most strongly, and it is natural to define u as the location in the map corresponding to this direction, $u = \mu(\boldsymbol{\eta}_u)$. However, Georgopoulos has shown that the direction is represented (more accurately and robustly) by a population code, in which the direction is represented by a neuronal field. Specifically, the activity ϕ_u of a neuron (above a base level) is proportional to the cosine of the angle between its preferred direction $\boldsymbol{\eta}_u$ and the direction $\mathbf{d}$ to be encoded. In particular, since the cosine of the angle between normalized vectors is equal to their scalar product, $\phi_u \propto \mathbf{d} \cdot \boldsymbol{\eta}_u$. A neurally plausible way of generating such a field is with a layer of radial basis functions (see Sect. "Continua of Basis Functions"), $\phi_u = r(\|\mathbf{d} - \boldsymbol{\eta}_u\|)$, where $r(x) = 1 - x^2/2$; then $\phi_u = \mathbf{d} \cdot \boldsymbol{\eta}_u$ [39].

Field computation is also used to update direction fields in the brain. For example, a remembered two-dimensional location, relative to the retina, must be updated when the eye moves [17,18]. In particular, if the direction field ϕ has a peak representing the remembered direction, and the eye moves in the direction $\mathbf{v}$, then this peak has to move in the direction $-\mathbf{v}$ in compensation. More specifically, if $\mathbf{v}$ is a two-dimensional vector defining the direction of eye motion, then the change in the direction field is given by the differential field equation, $\dot{\phi} = \mathbf{v} \cdot \nabla\phi$, where the gradient is a two-dimensional vector field (retinal coordinates). (That is, $\partial\phi(\mathbf{d}, t)/\partial t = \mathbf{v} \cdot \nabla_{\mathbf{d}}\phi(\mathbf{d}, t)$.) To see this, note that *behind* the moving peak $\nabla\phi$ and $-\mathbf{v}$ point in the same direction, and therefore $(-\mathbf{v}) \cdot \nabla\phi$ is positive; hence $\dot{\phi}$ is negative. Conversely, $\dot{\phi}$ is positive in front of the peak. Each component of the gradient may be approximated by convolution with a derivative-of-Gaussian (DoG) field, in accord with Eq. 8, which can be computed by neurons with DoG receptive field profiles. (Additional detail can be found elsewhere [39].)

RA Anderson [4] describes how transformations between retinal coordinates and head- or body-centered coordinates can be understood as transformations between field representations in area 7a of the posterior parietal cortex. For example, a minimum in a field may represent the destination of a motion (such as a saccade) in head-centered space, and then the gradient represents paths from other locations to that destination [39]. Further, the effects of motor neurons often correspond to vector fields [6,22].

Nonlinear Computation via Topographic Maps

As discussed in Sects. "Neuronal Fields" and "Motion in Direction Fields", the brain often represents scalar or vector quantities by topographic or computational maps, in which fields are defined over the range of possible values and a particular value is represented by a field with a peak of activity at the corresponding location. That is, a value $x \in \Omega$ is represented by a field $\phi_x \in \Phi(\Omega)$ that is distinctly peaked at x. For mathematical convenience we can idealize ϕ_x as a Dirac delta function (unit impulse) centered at x: δ_x, where $\delta_x(u) = \delta(u - x)$. That is, δ_x is an idealized topographic representation of x.

For every function $f\colon \Omega \to \Omega'$, with $y = f(x)$, there is a corresponding linear transformation of a topographic representation of its input, $\delta_x \in \Phi(\Omega)$, into a topographic representation of its output, $\delta_y \in \Phi(\Omega')$. It is easy to show that the kernel $K \in \Phi(\Omega' \times \Omega)$ of this operation is

$$K = \int_{\Omega} \delta_{f(x)} \wedge \delta_x \mathrm{d}x \,,$$

which is essentially a *graph* of the function f. That is, we can compute an arbitrary, possibly *nonlinear* function $y = f(x)$ by a *linear* operation on the corresponding computational maps, $\delta_y = K\delta_x$.

To avoid the use of Dirac delta functions, we can expand them into generalized Fourier series; for example, $\delta_x = \sum_k \beta_k \langle \beta_k \mid \delta_x \rangle = \sum_k \beta_k \beta_k(x)$. This expansion yields

$$\begin{aligned} K &= \int_\Omega \left(\sum_j \zeta_j \zeta_j[f(x)] \right) \wedge \left(\sum_k \beta_k \beta_k(x) \right) \mathrm{d}x \\ &= \sum_{j,k} \zeta_j \wedge \beta_k \int_\Omega \zeta_j[f(x)] \beta_k(x) \mathrm{d}x \\ &= \sum_{j,k} \zeta_j \wedge \beta_k \langle \zeta_j \circ f \mid \beta_k \rangle , \end{aligned}$$

where $\zeta_j \circ f$ is the composition of ζ_j and f: $(\zeta_j \circ f)(x) = \zeta_j[f(x)]$. A physically realizable approximation to K is obtained by limiting the summations to finite sets of physically realizable basis functions. (This has the effect of blurring the graph of f.)

Computation on topographic maps has a number attractive advantages. These are simple mathematical consequences of the linearity of topographic computation, but it will be informative to look at their applications in neural information processing. For example, transformation of input superpositions compute superpositions of the corresponding outputs in parallel: $K(\delta_x + \delta_{x'}) = \delta_{f(x)} + \delta_{f(x')}$ (recall Sect. "Neuronal Fields").

Since an input value is encoded by the position of the peak of a field rather than by its amplitude, the amplitude can be used for pragmatic characteristics of the input, such as its importance or certainty (see Sect. "Information Fields"). These pragmatic characteristics are preserved by topographic computation, $K(p\delta_x) = p\delta_{f(x)}$. Therefore if we have two (or more) inputs $x, x' \in \Omega$ with corresponding pragmatic scale factors $p, p' \in \mathbb{R}$, then the corresponding outputs carry the same factors, $K(p\delta_x + p'\delta_{x'}) = p\delta_{f(x)} + p'\delta_{f(x')}$. For example, if the inputs are weighted by confidence or importance, then the corresponding outputs will be similarly weighted. Further, if several inputs generate the same output, then their pragmatic scale factors will sum; for example if $f(x) = f(x')$, then $K(p\delta_x + p'\delta_{x'}) = (p + p')\delta_{f(x)}$. Thus, a number of inputs that are individually relatively unimportant (or uncertain) could contribute to a single output that is relatively important (or certain).

Finite superpositions of inputs are easily extended to the continuum case. For example, suppose that ϕ_x is the pragmatic scale factor associated with x, for all $x \in \Omega$ (for example, ϕ_x might be the probability of input x). We can think of the field ϕ as a continuum of weighted delta functions, $\phi = \int_\Omega \phi_x \delta_x \mathrm{d}x$. Applying the kernel to this field yields a corresponding continuum of weighted outputs, $K\phi = \int_\Omega \phi_x \delta_{f(x)} \mathrm{d}x \in \Phi(\Omega')$, where each point of the output field gives the total of the pragmatic scale factors (e. g., probabilities) of the inputs leading to the corresponding output value:

$$(K\phi)_y = \int_{\{x \mid y = f(x)\}} \phi_x \mathrm{d}x \, .$$

Therefore, by topographic computation, a transformation of an input probability distribution yields the corresponding output probability distribution.

We have remarked that the brain often uses *coarse coding*, in which a population of broadly-tuned neurons collectively represent a value with high precision (see Sect. "Discrete Basis Function Networks"). If ϕ is the coarse coding of input x, then its maximum will be at x and its amplitude will decrease with distance from x, $\phi_u = r(\|u - x\|)$. Similarly, $K\phi$ will be a coarse coding of the output $f(x)$ induced by the coarse coding of the input. As discussed in Sect. "Spatial Correlation and Convolution", if all the neurons have the same receptive field profile ϱ, then the effect of coarse coding is a convolution or correlation of ϱ with the input map.

Gabor Wavelets and Coherent States

In 1946 D Gabor presented a theory of information based on application to arbitrary signals of the Heisenberg–Weyl derivation of the quantum mechanical Uncertainty Principle [20]. Although he derived it for functions of time, it is easily generalizable to fields (square-integrable functions) over any finite-dimensional Euclidean space (reviewed elsewhere [34]). Therefore, for $\Omega \subset \mathbb{R}^n$, let $\psi \in \Phi(\Omega)$ be an arbitrary (possible complex-valued) field (assumed, as usual, to have a finite norm, that is, to be square-integrable; see Sect. "Mathematical Definitions"). To characterize this field's locality in space, we can measure its spread (or uncertainty) along each of the n spatial dimensions x_k by the root mean square deviation of x_k (assumed to have 0 mean):

$$\Delta x_k = \|x_k \psi(\mathbf{x})\| = \sqrt{\int_\Omega \psi_{\mathbf{x}}^* x_k^2 \psi_{\mathbf{x}} \mathrm{d}\mathbf{x}} \, ,$$

where $\mathbf{x} = (x_1, \ldots, x_n)^{\mathrm{T}} \in \mathbb{R}^n$. Consider also the Fourier transform $\Psi(\mathbf{u})$ of $\psi(\mathbf{x})$, the spread or uncertainty of which, in the frequency domain, can be quantified in

a similar way:

$$\Delta u_k = \|(u_k - \bar{u})\Psi(\mathbf{u})\| = \sqrt{\int_\Omega \Psi_\mathbf{u}^* u_k^2 \Psi_\mathbf{u} \mathrm{d}\mathbf{u}} \,.$$

It is straight-forward to show that the joint localization in any two conjugate variables (i. e., x_k in the space domain and u_k in the spatial-frequency domain) is limited by the *Gabor Uncertainty Principle*: $\Delta x_k \Delta u_k \geq 1/4\pi$.

This principle limits the information carrying capacity of any physically-realizable signal, so it is natural to ask if any function achieves the theoretical minimum, $\Delta x_k \Delta u_k = 1/4\pi$. Gabor showed that this minimum is achieved by what we may call the *Gabor elementary fields*, which have the form:

$$\Gamma_\mathbf{pu}(\mathbf{x}) = \exp\left[-\pi \|\mathrm{A}(\mathbf{x}-\mathbf{p})\|^2\right] \exp[2\pi i \mathbf{u}\cdot(\mathbf{x}-\mathbf{p})] \,.$$

The second, imaginary exponential defines a plane wave originating at $\mathbf{p}$ with a frequency and direction determined by the *wave vector* $\mathbf{u}$. The first exponential defines a Gaussian envelope centered at $\mathbf{p}$ with a shape determined by the diagonal *aspect matrix* $\mathrm{A} = \mathrm{diag}(\alpha_1, \ldots, \alpha_n)$, which determines the spread of the function along each of the space and frequency axes:

$$\Delta x_k = \frac{\alpha_k}{2\sqrt{\pi}} \,, \quad \Delta u_k = \frac{\alpha_k^{-1}}{2\sqrt{\pi}} \,.$$

Gaussian-modulated complex exponentials of this form correspond to the *coherent states* of quantum mechanics.

Each Gabor elementary field defines a cell in $2n$-dimensional "Gabor space" with volume $(4\pi)^{-n}$. He explained that these correspond to elementary units of information, which he called *logons*, since a field of finite spatial extent and bandwidth occupies a finite region in Gabor space, which determines its *logon content*. It may be computed by

$$N = \prod_{k=1}^{n} \frac{X_k}{\Delta x_k} \frac{U_k}{\Delta u_k} = (4\pi)^n \prod_{k=1}^{n} X_k U_k \,,$$

where X_k is the width of the field along the kth axis, and U_k its bandwidth on that axis, that is, a field's logon content is $(4\pi)^n$ times its Gabor-space volume.

The set of Gabor elementary functions are complete, and so any finite-energy function can be expanded into a series (see pp. 656–657 in [24]): $\psi = \sum_{k=1}^{N} c_k \Gamma_k$, where $\Gamma_1, \ldots, \Gamma_N$ are the Gabor fields corresponding to the cells occupied by ψ, and the c_k are complex coefficients. These N complex coefficients are the information conveyed by ψ, each corresponding to a logon or degree of freedom in the signal.

The Gabor elementary functions are not orthogonal, and so the coefficients cannot be computed by the inner product, $\langle \Gamma_k \mid \psi \rangle$. (They do form a *tight frame*, a very useful but weaker condition, under some conditions (see p. 1275 in [12]); see MacLennan [34] for additional discussion of the non-orthogonality issue.) On the other hand, it is easy to find the coefficients by minimization of the approximation error [13]. Let $\hat{\psi}(\mathbf{c}) = \sum_{k=1}^{N} c_k \Gamma_k$ and define the error $\mathcal{E} = \|\hat{\psi}(\mathbf{c}) - \psi\|^2$. This is a standard least-squares problem (cf. Sect. "Universal Approximation"), which can be solved by matrix calculation or by gradient descent on the error surface. It is easy to show that $\partial\mathcal{E}/\partial c_k = 2\langle \Gamma_k \mid \hat{\psi}(\mathbf{c}) - \psi \rangle$, and therefore gradient descent is given by $\dot{c}_k = r\langle \Gamma_k \mid \psi - \hat{\psi}(\mathbf{c}) \rangle$ for some rate $r > 0$.

There is considerable evidence (reviewed elsewhere [34]) that approximate Gabor representations are used in primary visual cortex, and there is also evidence that Gabor representations are used for generating motor signals (see pp. 139–144 in Pribram [53], Pribram et al. [54]).

Information Fields

J.J. Hopfield [25] observed that in some cases a neural impulse train can be understood as transmitting two signals: (1) the information content, encoded in the *phase* of the impulses relative to some global or local "clock", and (2) some other pragmatic characteristic of the information (such as importance, urgency, or confidence), encoded in the *rate* of the impulses. Such a combination of phase-encoded semantics and rate-encoded pragmatics may be common in the nervous system. Already in his *Laws of Thought* (1854), George Boole recognized idempotency as characteristic of information: repeating a message does not change its meaning, but it may affect its pragmatic import. The distinction is implicit in our typographic conventions; consider:

YES NO
YES NO

The horizontal distinction is semantic, but the vertical is pragmatic. More generally, following a distinction that has been made in quantum mechanics (see pp. 35–36 in [7]), we may say that the *form* of the signal *guides* the resulting action, but its *magnitude* determines the *amount* of action.

Similarly in field computation it may be useful to represent information by a field's shape and pragmatics by its magnitude; that is, pragmatics depends on the total amount of "stuff", semantics on it disposition (also a holistic property). The magnitude of such an *information field*

is given by its norm $\|\psi\|$, where we normally mean the inner-product norm of the Hilbert space, $\|\psi\|^2 = \langle\psi \mid \psi\rangle$ (which we can think of as "energy"), but other norms may be appropriate, depending on the relevant sense of the "amount" of action. The semantics of such fields is determined by their form, which we may identify with the normalization of the field, $N(\psi) = \psi/\|\psi\|$ (for nonzero fields). Idempotency is expressed by the identity $N(z\psi) = N(\psi)$ for all $z \neq 0$.

Therefore, it is reasonable that the entropy of a field depends on its shape, but not its magnitude:

$$S(\psi) = \int_\Omega \frac{\psi_u}{\|\psi\|} \log \frac{\psi_u}{\|\psi\|} \mathrm{d}u$$
$$= \int_\Omega N(\psi)_u \log N(\psi)_u \mathrm{d}u = \langle N(\psi) \mid \overline{\log} N(\psi)\rangle \,.$$

It is perhaps unsurprising that similar issues arise in quantum mechanics and field computation, for they are both formulated in the language of Hilbert spaces. For example, a quantum mechanical state ψ is taken to be undetermined with respect to magnitude, so that $z\psi$ is the same state as ψ for any nonzero complex number z (see p. 17 in [16]). Therefore, the state is conventionally taken to the normalized, $\|\psi\| = 1$, so that its square is a probability density function, $\rho_x = |\psi_x|^2$.

Independence of magnitude is also characteristic of the quantum potential, which led Bohm and Hiley [7] to characterize this field as *active information*. For example, if we write the wave function in polar form, $\psi_x = R_x \mathrm{e}^{iS_x/h}$, then the motion of a single particle is given by (see pp. 28–29 in [7]):

$$\frac{\partial S_x}{\partial t} + \frac{(\nabla S_x)^2}{2m} + V_x + Q_x = 0 \,,$$

where the quantum potential is defined:

$$Q_x = -\frac{\hbar^2}{2m} \frac{\nabla^2 R_x}{R_x} \,.$$

Since the Laplacian $\nabla^2 R_x$ is scaled by R_x, the quantum potential depends only on the *local form* of the wavefunction ψ, not on its magnitude. From this perspective, the particle moves under its own energy, but the quantum potential controls the energy.

Field Representations of Discrete Symbols

Quantum field theory treats discrete particles as quantized excitations of a field. This observation suggests analogous means by which field computation can represent and manipulate discrete symbols and structures, such as those employed in symbolic AI. It also provides potential models for neural representation of words and categories, especially in computational maps, which may illuminate how discrete symbol processing interacts with continuous image processing. From this perspective, discrete symbol manipulation is an emergent property of continuous field computation, which may help to explain the flexibility of human symbolic processes, such as language use and reasoning [36,37,38].

Mathematically, discrete symbols have the *discrete topology*, which is defined by the *discrete metric*, for which the distance between any two distinct objects is 1: $d(x,x) = 0$ and $d(x,y) = 1$ for $x \neq y$. Therefore we will consider various field representations of symbols that have this property. For example, discrete symbols could be represented by localized, non-overlapping patterns of activity in a computational map. In particular, symbols could be represented by Dirac delta functions, for which $\langle\delta_x \mid \delta_x\rangle = 1$ and $\langle\delta_x \mid \delta_y\rangle = 0$ for $x \neq y$. Here we may let $d(x,y) = 1 - \langle\delta_x \mid \delta_y\rangle$. More realistically, symbols could be represented by physically realizable normalized fields ϕ_x with little or no overlap between the representations of different symbols: $\langle\phi_x \mid \phi_y\rangle \approx 0$ for $x \neq y$. Indeed, any sufficiently large set of orthonormal fields may be used to represent discrete symbols. Fields may seem like an inefficient way to represent discrete symbols, and so it is worth observing that with at least 146 000 neurons per square millimeter, a one hundred thousand-word vocabulary could be represented in a few square millimeters of cortex.

Since the meaning of these fields is conveyed by the location of activity peak in the map, that is, by the shape of the field rather than its amplitude, the field's amplitude can be used for pragmatic scale factors, as previously discussed (see Sect. "Nonlinear Computation via Topographic Maps"). This could be used, for example, to convey the confidence or probability of a word or verbal category, or another pragmatic factor, such as loudness (cf. Sect. "Information Fields").

Wave packets (coherent states, Gabor elementary functions) are localized patterns of oscillation resulting from the superposition of a number of nonlocal oscillators with a Gaussian distribution of frequencies [34]. The relative phase of these oscillators determines the position of the wave packet within its field of activity. Therefore different phase relationships may determine field representations for different discrete symbols. The amplitude of the wave packet could represent pragmatic information, and frequency could be used for other purposes, for example for *symbol binding*, with bound symbols having the same frequency. Continuous phase control could be used to control the motion of wave packets in other representa-

tions, such as direction fields (Sect. "Motion in Direction Fields").

Gradient Processes

Many optimization algorithms and adaptive processes are implemented by gradient ascent or gradient descent. Because of its physical analogies, it is more convenient to think of optimization as decreasing a *cost function* rather than increasing some *figure of merit*. For example, the function might represent the difficulty of a motor plan or the incoherence in an interpretation of sensory data (such as stereo disparity).

Therefore suppose that $U\colon \Phi(\Omega) \to \mathbb{R}$ is a functional that defines the undesirability of a field; the goal is to vary ϕ so that $U(\phi)$ decreases down a path of "steepest descent". (By analogy with physical systems, we may call U a *potential function* and think of gradient descent as a *relaxation process* that decreases the potential.) The change in the potential U is given by the chain rule for field transformations (Eq. 3):

$$\begin{aligned}\dot{U}(t) &= (U \circ \phi)'(t,1)\\ &= U'[\phi(t)][\phi'(t)(1)]\\ &= \langle \nabla U[\phi(t)] \mid \dot{\phi}(t)\rangle\,.\end{aligned}$$

More briefly, suppressing the dependence on time, $\dot{U} = \langle \nabla U(\phi) \mid \dot{\phi}\rangle$. To guarantee $\dot{U} \le 0$ we let $\dot{\phi} = -r\nabla U(\phi)$ with a rate $r > 0$ for gradient descent. Then,

$$\begin{aligned}\dot{U} &= \langle \nabla U(\phi) \mid \dot{\phi}\rangle\\ &= \langle \nabla U(\phi) \mid -r\nabla U(\phi)\rangle\\ &= -r\|\nabla U(\phi)\|^2 \le 0\,.\end{aligned}$$

Therefore, gradient descent decreases U so long as the gradient is nonzero. (More generally, of course, so long as the trajectory satisfies $\langle \nabla U(\phi) \mid \dot{\phi}\rangle < 0$ the potential will decrease.)

Often the potential takes the form of a *quadratic functional*:

$$U(\phi) = \phi K \phi + L\phi + c\,,$$

where $K \in \Phi(\Omega \times \Omega)$, $\phi K \phi = \int_\Omega \int_\Omega \phi_u K_{uv} \phi_v \mathrm{d}u\mathrm{d}v$, L is a linear functional, and $c \in \mathbb{R}$. We require the *coupling field* K to be symmetric: $K_{uv} = K_{vu}$ for all $u, v \in \Omega$; typically it reflects the importance of correlated activity between any two locations u and v in ϕ. By the Riesz Representation Theorem (Sect. "Field Transformations") this quadratic functional may be written

$$U(\phi) = \phi K \phi + \langle \rho \mid \phi\rangle + c\,,$$

where $\rho \in \Phi(\Omega)$. The field gradient of such a functional is especially simple:

$$\nabla U(\phi) = 2K\phi + \rho\,.$$

In many cases $\rho = \mathbf{0}$ and then gradient descent is a linear process: $\dot{\phi} = -rK\phi$.

This process can be understood as follows. Notice that $-K_{uv}$ decreases with the *coupling* between locations u and v in a field and reflects the inverse variation of the potential with the coherence of the activity at those sites (i. e., the potential measures lack of coherence). That is, if $K_{uv} > 0$ then the potential will be lower to the extent that activity at u *covaries* with activity at v (since then $-\phi_u K_{uv} \phi_v \le 0$), and if $K_{uv} < 0$, the potential will be lower to the extent they *contravary*. Therefore, the gradient descent process $\dot{\phi} = -rK\phi$ changes ϕ_u to maximally decrease the potential in accord with the covariances and contravariances with other areas as defined by K: $\dot{\phi}_u = -r\int_\Omega K_{uv}\phi_v \mathrm{d}v$. The gradient descent will stop when it produces a field ϕ^* for which $-rK\phi^* = \mathbf{0}$, that is, a field in the *null space* of K (the set of all $\phi \in \Phi(\Omega)$ such that $K\phi = \mathbf{0}$).

Universal Approximation

A system of *universal computation* provides a limited range of facilities that can be programmed or otherwise set up to implement any computation in a large and interesting class. The most familiar example is the Universal Turing Machine (UTM), which can be programmed to emulate any Turing machine, and therefore can implement any (Church–Turing) computable function. While this model of universal computation has been important in the theory of digital computation, other models may be more relevant in for other computing paradigms [40,41] (see also ► Analog Computation).

Models of universal computation are important for both theory and practice. First, they allow the theoretical power of a computing paradigm to be established. For example, what cannot be computed by a UTM cannot be computed by a Turing machine or by any computer equivalent to a Turing machine. Conversely, if a function in Church–Turing computable, then it can be computed on a UTM or any equivalent machine (such as a programmable, general-purpose digital computer). Second, a model of universal computation for a computing paradigm provides a starting point for designing a general-purpose computer for that paradigm. Of course, there are many engineering problems that must be solved to design a practical general-purpose computer, but a model of universal computation establishes a theoretical foundation.

In the context of field computing there are several approaches to universal computation. One approach to universal field computation is based on a kind of field polynomial approximation based on the Taylor series for field transformations (see Sect. "Derivatives of Field Transformations") [32,33]. Another approach relies on a variety of "universal approximation theorems" for real functions, which are themselves generalizations of Fourier-series approximation (see pp. 208–209, 249–250, 264-265, 274–278, 290–294 in [23]). To explain this approach we will begin with the problem of interpolating a field transformation $F\colon \Phi(\Omega) \to \Phi(\Omega')$ specified by the samples $F(\phi_k) = \psi^k$, $k = 1, \ldots, P$. Further, we require the interpolating function to have the form

$$\hat{\psi} = \sum_{j=1}^{H} r_j(\phi)\alpha_j \,,$$

for some H, where the $r_j\colon \Phi(\Omega) \to \mathbb{R}$ are fixed nonlinear functionals (real-valued field transformation), and the $\alpha_j \in \Phi(\Omega')$ are determined by the samples so as to minimize the sum-of-squares error defined by $\mathcal{E} = \sum_{k=1}^{P} \|\hat{\psi}^k - \psi^k\|^2$, where $\hat{\psi}^k = \sum_{j=1}^{H} r_j(\phi_k)\alpha_j$. (A regularization term can be added if desired (see Chap. 5 in [23]).)

A field, as an element of a Hilbert space, has the same norm as the (infinite) sequence of its generalized Fourier coefficiants (with respect to some ON basis). Let $\zeta_1, \zeta_2, \ldots$ be a basis for $\Phi(\Omega')$, and we can compute the Fourier coefficients of $\hat{\psi}^k - \psi^k$ as follows:

$$\begin{aligned}\langle \zeta_i \mid \hat{\psi}^k - \psi^k \rangle &= \left\langle \zeta_i \,\middle|\, \sum_{j=1}^{H} r_j(\phi_k)\alpha_j - \psi^k \right\rangle \\ &= \left[\sum_{j=1}^{H} r_j(\phi_k)\langle \zeta_i \mid \alpha_j \rangle \right] - \langle \zeta_i \mid \psi^k \rangle \,.\end{aligned}$$

Let $R_{kj} = r_j(\phi_k)$, $A_{ji} = \langle \zeta_i \mid \alpha_j \rangle$, and $Y_{ki} = \langle \zeta_i \mid \psi^k \rangle$. Then, $\langle \zeta_i \mid \hat{\psi}^k - \psi^k \rangle = \sum_{j=1}^{H} R_{kj}A_{ji} - Y_{ki}$. The fields may approximated arbitrarily closely by the first N Fourier coefficients, in which case R, A, and Y are ordinary matrices. Then $\|\hat{\psi}^k - \psi^k\|^2 \approx \sum_{i=1}^{N} E_{ki}^2$, where $\mathrm{E} = \mathrm{RA} - \mathrm{Y}$. Therefore the approximate total error is $\hat{\mathcal{E}} = \sum_{k=1}^{P}\sum_{i=1}^{N} E_{ki}^2$, or $\hat{\mathcal{E}} = \|\mathrm{E}\|_F^2$ (the squared Frobenius norm).

This is a standard least-squares minimization problem, and, as is well known (see pp. 371–373 in [28]), the error is minimized by $\mathrm{A} = \mathrm{R}^{+}\mathrm{Y}$, where R^{+} is the *Moore–Penrose pseudoinverse* of the interpolation matrix R: $\mathrm{R}^{+} = (\mathrm{R}^{\mathrm{T}}\mathrm{R})^{-1}\mathrm{R}^{\mathrm{T}}$. From A we can compute the required fields to approximate F: $\alpha_j = \sum_{i=1}^{N} A_{ji}\zeta_i$.

For universality, we require that the approximation error can be made arbitrarily small, which depends on the choice of the basis functionals r_j, as can be learned from multivariable interpolation theory. Therefore, we represent the input fields by their first M generalized Fourier coefficients, an approximation that can be made as accurate as we like. Let $\beta_1, \beta_2, \ldots$ be an ON basis for $\Phi(\Omega)$ and let $\mathbf{p}^M\colon \Phi(\Omega) \to \mathbb{R}^M$ compute this finite-dimensional representation: $p_j^M(\phi) = \langle \beta_j \mid \phi \rangle$. We will approximate $r_j(\phi) \approx s_j[\mathbf{p}^M(\phi)]$, for appropriate functions $s_j\colon \mathbb{R}^M \to \mathbb{R}$, $j = 1, \ldots, H$. That is, we are approximating the field transformation F by

$$F(\phi) \approx \sum_{j=1}^{H} s_j[\mathbf{p}^M(\phi)]\alpha_j \,.$$

Now let $S_{kj} = s_j[\mathbf{p}^M(\phi_k)]$, and we have corresponding finite-dimensional interpolation conditions $\mathrm{Y} = \mathrm{SA}$ with the best least-square solution $\mathrm{A} = \mathrm{S}^{+}\mathrm{Y}$.

Various universal approximation theorems tell us that, given an appropriate choice of basis functions $s_1, \ldots, s_H$, any continuous function $\mathbf{f}\colon \mathbb{R}^M \to \mathbb{R}^N$ can be approximated arbitrarily closely by a linear combination of these functions (see pp. 208–209 in [23]). That is, the error $\hat{\mathcal{E}} = \|\mathrm{SA} - \mathrm{Y}\|_F^2$ can be made as small as we like. Therefore, appropriate choices for the s_j imply corresponding choices for the basis functionals r_j.

For example, one universal class of basis functions has the form $s_j(\mathbf{x}) = c(\mathbf{w}_j \cdot \mathbf{x} + b_j)$, for any nonconstant, bounded, monotone-increasing continuous function c (see p. 208 in [23]). This form is common in artificial neural networks, where $\mathbf{w}_j$ is a vector of neuron j's input weights (connection strengths) and b_j is its bias. To find the corresponding basis functional, $r_j(\phi) = s_j[\mathbf{p}^M(\phi)]$, observe

$$\begin{aligned}\mathbf{w}_j \cdot \mathbf{p}^M(\phi) + b_j &= \sum_{k=1}^{M} w_{jk} p_k^M(\phi) + b_j \\ &= \sum_{k=1}^{M} w_{jk}\langle \beta_k \mid \phi \rangle + b_j \\ &= \left\langle \sum_{k=1}^{M} w_{jk}\beta_k \,\middle|\, \phi \right\rangle + b_j \,.\end{aligned}$$

Therefore, let $\varpi_j = \sum_{k=1}^{M} w_{jk}\beta_k$, and we see that a universal class of functionals has the form:

$$r_j(\phi) = c(\langle \varpi_j \mid \phi \rangle + b_j) \,. \tag{10}$$

Thus, in this field analog of an artificial neuron, the input field ϕ is matched to the neuron's interconnection field ϖ_j.

Another universal class is the *radial basis functions*, $s_j(\mathbf{x}) = r(\|\mathbf{x} - \mathbf{c}^j\|)$, where the radial function r is monotonically decreasing, and the centers $\mathbf{c}^j$ are either fixed or dependent on the function to be approximated. A corresponding universal class of field functions has the form:

$$r_j(\phi) = r(\|\phi - \eta_j\|) \,, \tag{11}$$

where each field $\eta_j = \sum_i c_i^j \zeta_i$ causes the maximal response of the corresponding basis function r_j. Furthermore, if we set $H = P$ and $\eta_j = \phi_j$, then the matrix R is invertible for a wide variety of radial functions r (see pp. 264–265 in [23]).

Thus familiar methods of universal approximation can be transferred to field computation, which reveals simple classes of field transformations that are universal. This implies that universal field computers can be designed around a small number of simple functions (e. g., field summation, inner product, monotonic real functions).

Field Computers

Structure

As previously explained (see Sect. "Definition"), fields do not have to be physically continuous in either variation or spatial extension (that is, in range or domain), so long as the discretization is sufficiently fine that a continuum is a practical approximation. Therefore, field computation can be implemented with ordinary serial or parallel digital computing systems (as it has been in the past). However, field computation has a distinctively different approach to information representation and processing; computation tends to be shallow (in terms of operations applied), but very wide, "massively parallel" in the literal sense of computing with an effectively continuous *mass* of processors. Therefore field computation provides opportunities for the exploitation of novel computing media that may not be suitable for digital computation. For example, as the brain illustrates how relatively slow, low precision analog computing devices can be used to implement intelligent information processing via field computation, so electronic field computers may exploit massive assemblages of low-precision analog devices, which may be imprecisely fabricated, located, and interconnected. Other possibilities are optical computing in which fields are represented by optical wavefronts, molecular computation based on films of bacteriorhodopsin or similar materials, chemical computers based on reaction-diffusion systems, and "free space computing" based on the interactions of charge carriers and electrical fields in homogeneous semiconductors (see Sect. "Field Computing Hardware").

Field computation is a kind of analog computation, and so there are two principal time domains in which field computation can take place, sequential time and continuous time (see ► Analog Computation). In *sequential* computation, operations take place in discrete steps in an order prescribed by a program. Therefore, sequential field computation is similar to ordinary digital computation, except that the individual program steps may perform massively parallel analog field operations. For example, a field assignment statement, such as:

$$\psi := \phi + \psi \,;$$

updates that field variable ψ to contain the sum of ϕ and the previous value of ψ.

In *continuous-time computation* the fields vary continuously in time, generally according to differential equations in which time is the independent variable; this has been the mode of operation of most analog computers in the past. In this case, a simple dependence, such as $\psi = \phi + \chi$, is assumed to have an implicit time parameter, $\psi(t) = \phi(t) + \chi(t)$, which represents the real time of computation. Since continuous-time programs are often expressed by differential equations, these computers usually provide hardware for definite integration of functions with respect to time:

$$\psi(t) = \psi_0 + \int_0^t F[\phi(\tau)]\mathrm{d}\tau \,. \tag{12}$$

Continuous-time programs are expressed by circuit diagrams (variable-dependency diagrams) rather than by textual programs such as used in digital computer programming (see Fig. 1 for an example). Although special-purpose analog and digital computers are appropriate for many purposes, already in the first half of the twentieth century the value of general-purpose (programmable) digital and analog computers had been recognized (see ► Analog Computation). Therefore it will be worthwhile to consider briefly the sort of facilities we may expect to find in a general-purpose field computer (whether operating in sequential or continuous time).

We have seen that the following facilities are sufficient for universal computation (Sect. "Universal Approximation"): multiplication of fields by scalars, local (point-wise) addition of fields ($\psi_u = \phi_u + \chi_u$), and some means of computing appropriate basis functionals. Neural-net style functionals (Eq. 10) require inner product and any nonconstant, bounded, monotone-increasing scalar function (i. e., a sigmoid function). Radial basis functionals (Eq. 11) require the norm (which can be computed with the in-

ner product) and any non-constant, bounded, monotone-decreasing scalar function. (Point-wise subtraction can be implemented, of course, by scalar multiplication and point-wise addition.) These are modest requirements, and we can expect practical field computers to have additional facilities.

In addition, continuous-time field computers will implement definite integration with respect to time (Eq. 12), which is used to implement field processes defined by differential equations. The equations are implemented in terms of the operations required for universal computation or in terms of others, discussed next.

Additional useful operations for general-purpose field computing include matrix-vector style field products (Hilbert–Schmidt integral operators), outer product, convolution, cross-correlation, normalization, local (point-wise) product and quotient ($\psi_u = \phi_u \chi_u$, $\psi_u = \phi_u / \chi_u$), and various other local operations ($\overline{\log}$, $\overline{\exp}$, etc.). Operations on vector fields can be implemented by scalar field operations on the vector components (Cartesian or polar); in this manner, vector fields of any finite dimension can be processed. If vector fields and operations on them are provided by the hardware, then it is useful if these operations include conversions between scalar and vector fields (e. g., between vector fields and their Cartesian or polar coordinate fields). Other useful vector field operations include point-wise scalar products between vector fields ($\boldsymbol{\psi}_u = \boldsymbol{\phi}_u \cdot \boldsymbol{\chi}_u$), gradient ($\nabla$), Laplacian ($\nabla^2$), divergence ($\nabla\cdot$), and point-wise scalar-vector multiplication ($\boldsymbol{\psi}_u = \phi_u \boldsymbol{\chi}_u$). Scalar analog computation is a degenerate case of field computation (since scalars correspond to fields in $\Phi(\{0\})$), and so practical general-purpose field computers will include the facilities typical of analog computers (see ► Analog Computation).

The Extended Analog Computer

One interesting proposal for a general-purpose field computer is the *Extended Analog Computer* (EAC) of LA Rubel, which was a consequence of his conviction that the brain is an analog computer [55]. However, Rubel and others had shown that the existing model of a general-purpose analog computer (GPAC), the abstract *differential analyzer* defined by CE Shannon, had relatively severe theoretical limitations, and so it did not seem adequate as a model of the brain (see ► Analog Computation) [30,51,56,60,61]. Like Shannon's differential analyzer, the EAC is an abstract machine intended for theoretical investigation of the power of analog computation, not a proposal for a practical computer [57]; nevertheless, some actual computing devices have been based on it.

The EAC is structured in a series of levels, each building on those below it, taking outputs from the lower layers and applying analog operations to them to produce its own outputs. The inputs to the lowest layer are a finite number of "settings", which can be thought of real-numbers (e. g., set by a continuously adjustable knob). This layer is able to combine the inputs with real constants to compute polynomials over which it can integrate to generate differentially algebraic functions; this layer is effectively equivalent to Shannon's GPAC. Each layer provides a number of analog devices, including "boundary-value-problem boxes", which can solve systems of PDEs subject to boundary conditions and other constraints. That is, these conceptual devices solve field computation problems. Although for his purposes Rubel was not interested in implementation, he did remark that PDE solvers might be implemented by physical processes that obeyed the same class of PDEs as the problem (e. g., using physical diffusion to solve diffusion problems). This of course is precisely the old field analogy method, which was also used in network analyzers (recall Sect. "Introduction"). Rubel was able to show that the EAC is able to solve an extremely large class of problems, but the extent of its power has not been determined (see ► Analog Computation).

Field Computing Hardware

Research in field computing hardware is ongoing and a comprehensive survey is beyond the scope of this article; a few examples must suffice.

Although the EAC was intended as a conceptual machine (for investigating the limits of analog computing), JW Mills has demonstrated several hardware devices inspired by it [46,47]. In these the diffusion of electrons in bulk silicon or conductive gels is used to solve diffusion equations subject to given boundary conditions, an technique he describes as "computing with empty space". This approach, in which a physical system satisfying certain PDEs is used to solve problems involving similar PDEs, is a contemporary version of the "field analogy method" developed by Kirchhoff and others (see Sect. "Introduction").

Adamatzky and his colleagues have investigated chemical field computers for implementing reaction-diffusion equations [1,2]; see Sect. "Diffusion Processes" and ► Reaction-Diffusion Computing. These use variants of the Belousov–Zhabotinsky Reaction and similar chemical reactions. Although the chemical reactions proceed relatively slowly, they are massively parallel: at the molecular level ("molar parallelism"). Also, Chaps. 6–8 in Adamatzky et al. [2] have designed both analog and digital electronic reaction-diffusion computers. M Perŭs and his col-

leagues have investigated the use of quantum holography to implement field analogues of neural-network algorithms [31,49].

Several investigators have explored optical implementations of field computers. For example, Skinner et al. [62] used self-lensing media, which respond nonlinearly to applied irradience, to implement feed-forward neural networks trained by back-propagation. Tõkés et al. [68,69] have been developing an optical field computer using bacteriorhodopsin as a medium.

Future Directions

In the future field computation can be expected to provide an increasingly important analytical and intuitive framework for understanding massively parallel analog computation in natural and artificial intelligence.

First, field computation will provide a theoretical framework for understanding information processing in the brain in terms of cortical maps and, more generally, at a level between anatomical structures and individual neurons or small neural circuits. This will require improved understanding of information processing in terms of field computation, which will benefit from cognitive neuroscience research, but also contribute new computational concepts to it. Increased understanding of neural field computation will improve our ability to design very large artificial neural networks, which will be more attractive as massively parallel neurocomputing hardware is developed.

Traditionally, artificial intelligence has approached knowledge representation from the perspective of discrete, language-like structures, which are difficult to reconcile with the massively parallel analog representations found in the cortex. Therefore field computation will provide an alternative framework for understanding knowledge representation and inference, which will be more compatible with neuroscience but also provide a basis for understanding cognitive phenomena such as context sensitivity, perception, sensorimotor coordination, image-based cognition, analogical and metaphorical thinking, and nonverbal intelligences (kinesthetic, emotional, aesthetic, etc.).

As we have seen, concepts from field computation may be applied to understanding the collective intelligence of large groups of organisms. This approach permits separating the abstract computational principles from the specifics of their realization by particular organisms, and therefore permits their application to other organisms or artificial systems. For example, principles of field computation governing the self-organization of groups of organisms are applicable to distributed robotics; in particular, they will provide a foundation for controlling very large population of microrobots or nanobots.

Embryological morphogenesis is naturally expressed in terms of field computation, since the differentiation and self-organization of an (initially homogeneous) cell mass is governed by continuous distributions of continuous quantity. Therefore, field computation provides a vehicle for rising above the specifics of particular signaling molecules, mechanisms of cell migration, etc. in order to understand development in abstract or formal terms. Understanding morphogenesis in terms of field computation will facilitate applying its principles to other systems in which matter self-organizes into complex structures. In particular, field computation will suggest means for programming the reorganization of matter for nanotechnological applications and for describing the behavior of adaptive "smart" materials.

As we approach the end of Moore's Law [48], future improvements in computing performance will depend on developing new computing paradigms not based in sequential digital computation (see also ► Analog Computation). Improvements in both speed and density can be achieved by matching data representations and computational operations to the physical processes that realize them, which are primarily continuous and parallel in operation. Indeed, many of these processes are described in terms of fields or involve physical fields (i. e., phenomenological or structural fields). Therefore field computation points toward many structural processes that might be used for computation and provides a framework for understanding how best to use them. Thus we anticipate that field computation will play an important role in post-Moore's Law computing.

Bibliography

Primary Literature

1. Adamatzky A (2001) Computing in nonlinear media and automata collectives. Institute of Physics Publishing, Bristol
2. Adamatzky A, De Lacy Costello B, Asai T (2005) Reaction-diffusion computers. Elsevier, Amsterdam
3. Anderson JA (1995) An introduction to neural networks. MIT Press, Cambridge
4. Anderson RA (1995) Coordinate transformations and motor planning in posterior parietal cortex. In: Gazzaniga MS (ed) The cognitive neurosciences. MIT Press, Cambridge, pp 519–32
5. Bar-Yam Y (1997) Dynamics of complex systems. Perseus Books, Reading
6. Bizzi E, Mussa-Ivaldi FA (1995) Toward a neurobiology of coordinate transformation. In: Gazzaniga MS (ed) The cognitive neurosciences. MIT Press, Cambridge, pp 495–506
7. Bohm D, Hiley BJ (1993) The undivided universe: an ontological interpretation of quantum theory. Routledge, New York

8. Bonabeau E, Dorigo M, Theraulaz G (1999) Swarm intelligence: from natural to artificial systems. Santa Fe Institute Studies in the Sciences of Complexity. Oxford University Press, New York
9. Brachman G, Narici L (1966) Functional analysis. Academic Press, New York
10. Camazine S, Deneubourg J-L, Franks NR, Sneyd G, Theraulaz J, Bonabeau E (2001) Self-organization in biological systems. Princeton University Press, Princeton
11. Changeux J-P (1985) Neuronal man: the biology of mind. Oxford University Press, Oxford, tr. by L. Garey
12. Daubechies I, Grossman A, Meyer Y (1986) Painless non-orthogonal expansions. J Math Phys 27:1271–1283
13. Daugman JG (1993) An information-theoretic view of analog representation in striate cortex. In: Schwartz EL (ed) Computational neuroscience. MIT Press, Cambridge, pp 403–423
14. Davies JA (2005) Mechanisms of morphogensis. Elsevier, Amsterdam
15. Deneubourg JL (1977) Application de l'ordre par fluctuation à la description de certaines étapes de la construction du nid chez les termites. Insectes Sociaux 24:117–130
16. Dirac PAM (1958) The principles of quantum mechanics, 4th edn. Oxford University Press, Oxford
17. Droulez J, Berthoz A (1991) The concept of dynamic memory in sensorimotor control. In: Humphrey DR, Freund H-J (eds) Motor control: concepts and issues. Wiley, New York, pp 137–161
18. Droulez J, Berthoz A (1991) A neural network model of sensoritopic maps with predictive short-term memory properties. Proc Natl Acad Sci USA 88:9653–9657
19. Feldman JA, Ballard DH (1982) Connectionist models and their properties. Cogn Sci 6(3):205–254
20. Gabor D (1946) Theory of communication. J Inst Electr Eng 93(III):429–457
21. Georgopoulos AP (1995) Motor cortex and cognitive processing. In: The Cognitive Neurosciences. MIT Press, Cambridge, pp 507–517
22. Goodman SJ, Anderson RA (1989) Microstimulation of a neural-network model for visually guided saccades. J Cogn Neurosci 1:317–326
23. Haykin S (1999) Neural networks: a comprehensive foundation, 2nd edn. Prentice-Hall, Upper Saddle River
24. Heil CE, Walnut DF (1989) Continuous and discrete wavelet transforms. SIAM Rev 31(4):628–666
25. Hopfield JJ (1995) Pattern recognition computation using action potential timing for stimulus response. Nature 376:33–36
26. Kirchhoff G (1845) Ueber den Durchgang eines elektrischen Stromes durch eine Ebene, insbesondere durch eine kreisförmige. Ann Phys Chemie 140/64(4):497–514
27. Knudsen EJ, du Lac S, Esterly SD (1987) Computational maps in the brain. Ann Rev Neurosc 10:41–65
28. Leon SJ (1986) Linear algebra with applications, 2nd edn. Macmillan, New York
29. Light WA (1992) Ridge functions, sigmoidal functions and neural networks. In: Cheney EW, Chui CK, Schumaker LL (eds) Approximation theory VII. Academic Press, Boston, pp 163–206
30. Lipshitz L, Rubel LA (1987) A differentially algebraic replacment theorem. Proc Am Math Soc 99(2):367–72
31. Loo CK, Peruš M, Bischof H (2004) Associative memory based image and object recognition by quantum holography. Open Syst Inf Dyn 11(3):277–289
32. MacLennan BJ (1987) Technology-independent design of neurocomputers: the universal field computer. In: Caudill M, Butler C (eds) Proceedings of the IEEE First International Conference on Neural Networks, vol 3. IEEE Press, Piscataway, pp 39–49
33. MacLennan BJ (1990) Field computation: a theoretical framework for massively parallel analog computation, parts I–IV. Technical Report CS-90-100. Department of Computer Science, University of Tennessee, Knoxville, Available from www.cs.utk.edu/~mclennan
34. MacLennan BJ (1991) Gabor representations of spatiotemporal visual images. Technical Report CS-91-144. Department of Computer Science, University of Tennessee, Knoxville, Available from www.cs.utk.edu/~mclennan
35. MacLennan BJ (1993) Information processing in the dendritic net. In: Karl HP (ed) Rethinking neural networks: quantum fields and biological data. Lawrence Erlbaum, Hillsdale, pp 161–197
36. MacLennan BJ (1994) Continuous computation and the emergence of the discrete. In: Karl HP (ed) Origins: brain and self-organization. Lawrence Erlbaum, Hillsdale, pp 121–151
37. MacLennan BJ (1994) Image and symbol: continuous computation and the emergence of the discrete. In: Honavar V, Uhr L (eds) Artificial intelligence and neural networks: steps toward principled integration. Academic Press, New York, pp 207–224
38. MacLennan BJ (1995) Continuous formal systems: a unifying model in language and cognition. In: Proc. of the IEEE Workshop on Architectures for Semiotic Modeling and Situation Analysis in Large Complex Systems. IEEE Press, Piscataway, pp 161–172
39. MacLennan BJ (1997) Field computation in motor control. In: Morasso PG, Sanguineti V (eds) Self-organization, computational maps and motor control. Elsevier, Amsterdam, pp 37–73
40. MacLennan BJ (2003) Transcending Turing computability. Minds Mach 13:3–22
41. MacLennan BJ (2004) Natural computation and non-Turing models of computation. Theor Comput Sci 317:115–145
42. Mathematical Society of Japan (1980) Encyclopedic dictionary of mathematics. MIT Press, Cambridge
43. McClelland JL, Rumelhart DE, PDP Research Group (1986) Parallel distributed processing: explorations in the microstructure of cognition, vol 2. Psychological and biological models. MIT Press, Cambridge
44. McFadden J (2002) Synchronous firing and its influence on the brain's electromagnetic field: evidence for an electromagnetic field theory of consciousness. J Conscious Stud 9(4):23–50
45. Miller MI, Roysam B, Smith KR, O'Sullivan JA (1991) Representing and computing regular languages on massively parallel networks. IEEE Trans Neural Netw 2:56–72
46. Mills JW (1996) The continuous retina: Image processing with a single-sensor artificial neural field network. In: Proc. IEEE Conference on Neural Networks. IEEE Press, Piscataway
47. Mills JW, Himebaugh B, Kopecky B, Parker M, Shue C, Weilemann C (2006) "Empty space" computes: the evolution of an unconventional supercomputer. In: Proc. of the 3rd Conference on Computing Frontiers. ACM Press, New York, pp 115–126
48. Moore GE (1965) Cramming more components onto integrated circuits. Electronics 38(8):114–117
49. Peruš M (1998) A quantum information-processing "algorithm" based on neural nets. In: Wang P, Georgiou G, Janikow C, Yao Y (eds) Joint conference on information sci-

ences, vol II. Association for Intelligent Machinery, New York, pp 197–200
50. Pockett S (2000) The nature of consciousness: a hypothesis. Writers Club Press, San Jose
51. Pour-El MB (1974) Abstract computability and its relation to the general purpose analog computer (some connections between logic, differential equations and analog computers). Trans Am Math Soc 199:1–29
52. Powell MJD (1987) Radial basis functions for multivariable interpolation: a review. In: Algorithms for approximation. Clarendon, New York, pp 143–167
53. Pribram KH (1991) Brain and perception: holonomy and structural in figural processing. Lawrence Erlbaum, Hillsdale
54. Pribram KH, Sharafat A, Beekman GJ (1984) Frequency encoding in motor systems. In: Whiting HTA (ed) Human motor actions: Bernstein reassessed. Elsevier, New York, pp 121–156
55. Rubel LA (1985) The brain as an analog computer. J Theor Neurobiol 4:73–81
56. Rubel LA (1988) Some mathematical limitations of the general-purpose analog computer. Adv Appl Math 9:22–34
57. Rubel LA (1993) The extended analog computer. Adv Appl Math 14:39–50
58. Rumelhart DE, McClelland JL, PDP Research Group (1986) Parallel distributed processing: explorations in the microstructure of cognition, vol 1. Foundations. MIT Press, Cambridge
59. Sanger TD (1996) Probability density estimation for the interpretation of neural population codes. J Neurophysiol 76:2790–2793
60. Shannon CE (1941) Mathematical theory of the differential analyzer. J Math Phys Mass Inst Tech 20:337–354
61. Shannon CE (1993) Mathematical theory of the differential analyzer. In: Sloane NJA, Wyner AD (eds) Claude Elwood Shannon: collected papers. IEEE Press, New York, pp 496–513
62. Skinner SR, Behrman EC, Cruz-Cabrera AA, Steck JE (1995) Neural network implementation using self-lensing media. Appl Opt 34:4129–4135
63. Small JS (2001) The analogue alternative: the electronic analogue computer in Britain and the USA, 1930–1975. Routledge, London, New York
64. Solé R, Goodwin B (2000) Signs of life: how complexity pervades biology. Basic Books, New York
65. Soroka WW (1954) Analog methods in computation and simulation. McGraw-Hill, New York
66. Steinbeck O, Tóth A, Showalter K (1995) Navigating complex labyrinths: optimal paths from chemical waves. Science 267:868–871
67. Ting P-Y, Iltis RA (1994) Diffusion network architectures for implementation of Gibbs samplers with applications to assignment problems. IEEE Trans Neural Netw 5:622–638
68. Tőkés S, Orzó L, Ayoub A (2003) Two-wavelength POAC (programmable opto-electronic analogic computer) using bacteriorhodopsin as dynamic holographic material. In: Proc. of ECCTD '03 Conference, vol 3. Krakow, pp 97–100
69. Tőkés S, Orzó L, Váró G, Dér A, Ormos P, Roska T (2001) Programmable analogic cellular optical computer using bacteriorhodopsin as analog rewritable image memory. In: Dér A, Keszthelyi L (eds) Bioelectronic applications of photochromic pigments. IOS Press, Amsterdam, pp 54–73
70. Truitt TD, Rogers AE (1960) Basics of analog computers. John F. Rider, New York
71. Turing AM (1952) The chemical basis of morphogenesis. Philos Trans Royal Soc B 237:37–72

Books and Reviews

Bachman G, Narici L (1966) Functional analysis. Academic Press, New York

Berberian SK (1961) Introduction to Hilbert space. Oxford, New York

MacLennan BJ (1991) Field computation: a theoretical framework for massively parallel analog computation, parts I–IV. Technical Report CS-90-100, Dept. of Computer Science, University of Tennessee, Knoxville. Available from http:www.cs.utk.edu/~mclennan

MacLennan BJ (2008) Foundations of Field Computation. In preparation. Available from http:www.cs.utk.edu/~mclennan

Field Theoretic Methods

Uwe Claus Täuber
Department of Physics, Center for Stochastic Processes in Science and Engineering, Virginia Polytechnic Institute and State University, Blacksburg, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Correlation Functions and Field Theory
Discrete Stochastic Interacting Particle Systems
Stochastic Differential Equations
Future Directions
Acknowledgments
Bibliography

Glossary

Absorbing state State from which, once reached, an interacting many-particle system cannot depart, not even through the aid of stochastic fluctuations.

Correlation function Quantitative measure of the correlation of random variables; usually set to vanish for statistically independent variables.

Critical dimension Borderline dimension d_c above which mean-field theory yields reliable results, while for $d \leq d_c$ fluctuations crucially affect the system's large scale behavior.

External noise Stochastic forcing of a macroscopic system induced by random external perturbations, such as thermal noise from a coupling to a heat bath.

Field theory A representation of physical processes through continuous variables, typically governed by an exponential probability distribution.

Generating function Laplace transform of the probability distribution; all moments and correlation functions follow through appropriate partial derivatives.

Internal noise Random fluctuations in a stochastic macroscopic system originating from its internal kinetics.

Langevin equation Stochastic differential equation describing time evolution that is subject to fast random forcing.

Master equation Evolution equation for a configurational probability obtained by balancing gain and loss terms through transitions into and away from each state.

Mean-field approximation Approximative analytical approach to an interacting system with many degrees of freedom wherein spatial and temporal fluctuations as well as correlations between the constituents are neglected.

Order parameter A macroscopic density corresponding to an extensive variable that captures the symmetry and thereby characterizes the ordered state of a thermodynamic phase in thermal equilibrium. Nonequilibrium generalizations typically address appropriate stationary values in the long-time limit.

Perturbation expansion Systematic approximation scheme for an interacting and/or nonlinear system that involves a formal expansion about an exactly solvable simplication by means of a power series with respect to a small coupling.

Definition of the Subject

Traditionally, complex macroscopic systems are often described in terms of ordinary differential equations for the temporal evolution of the relevant (usually collective) variables. Some natural examples are particle or population densities, chemical reactant concentrations, and magnetization or polarization densities; others involve more abstract concepts such as an apt measure of activity, etc. Complex behavior often entails (diffusive) spreading, front propagation, and spontaneous or induced pattern formation. In order to capture these intriguing phenomena, a more detailed level of description is required, namely the inclusion of spatial degrees of freedom, whereupon the above quantities all become local density fields. Stochasticity, i. e., randomly occuring propagation, interactions, or reactions, frequently represents another important feature of complex systems. Such stochastic processes generate *internal noise* that may crucially affect even long-time and large-scale properties. In addition, other system variables, provided they fluctuate on time scales that are fast compared to the characteristic evolution times for the relevant quantities of interest, can be (approximately) accounted for within a Langevin description in the form of *external* additive or multiplicative noise.

A quantitative mathematical analysis of complex spatio-temporal structures and more generally cooperative behavior in stochastic interacting systems with many degrees of freedom typically relies on the study of appropriate *correlation functions*. *Field-theoretic*, i. e., spatially continuous, representations both for random processes defined through a master equation and Langevin-type stochastic differential equations have been developed since the 1970s. They provide a general framework for the computation of correlation functions, utilizing powerful tools that were originally developed in quantum many-body as well as quantum and statistical field theory. These methods allow us to construct systematic approximation schemes, e. g., *perturbative expansions* with respect to some parameter (presumed small) that measures the strength of fluctuations. They also form the basis of more sophisticated renormalization group methods which represent an especially potent device to investigate scale-invariant phenomena.

Introduction

Stochastic Complex Systems

Complex systems consist of many interacting components. As a consequence of either these interactions and/or the kinetics governing the system's temporal evolution, correlations between the constituents emerge that may induce cooperative phenomena such as (quasi-)periodic oscillations, the formation of spatio-temporal patterns, and phase transitions between different macroscopic states. These are characterized in terms of some appropriate collective variables, often termed *order parameters*, which describe the large-scale and long-time system properties. The time evolution of complex systems typically entails random components: either, the kinetics itself follows stochastic rules (certain processes occur with given probabilities per unit time); or, we project our ignorance of various fast microscopic degrees of freedom (or our lack of interest in their detailed dynamics) into their treatment as stochastic noise.

An exact mathematical analysis of nonlinear stochastic systems with many interacting degrees of freedom is usually not feasible. One therefore has to resort to either computer simulations of corresponding stochastic cellu-

lar automata, or approximative treatments. A first step, which is widely used and often provides useful qualitative insights, consists of ignoring spatial and temporal fluctuations, and just studying equations of motion for ensemble-averaged order parameters. In order to arrive at closed equations, additional simplifications tend to be necessary, namely the factorization of correlations into powers of the mean order parameter densities. Such approximations are called *mean-field* theories; familiar examples are rate equations for chemical reaction kinetics or Landau–Ginzburg theory for phase transitions in thermal equilibrium. Yet in some situations mean-field approximations are insufficient to obtain a satisfactory quantitative description (see, e. g., the recent work collected in [1,2]). Let us consider an illuminating example.

Example: Lotka–Volterra Model

In the 1920s, Lotka and Volterra independently formulated a mathematical model to describe emerging periodic oscillations respectively in coupled autocatalytic chemical reactions, and in the Adriatic fish population (see, e. g., [3]). We shall formulate the model in the language of population dynamics, and treat it as a stochastic system with two species A (the 'predators') and B (the 'prey'), subject to the following reactions: predator death $A \to \emptyset$, with rate μ; prey proliferation $B \to B + B$, with rate σ; predation interaction $A + B \to A + A$, with rate λ. Obviously, for $\lambda = 0$ the two populations decouple; while the predators face extinction, the prey population will explode. The average predator and prey population densities $a(t)$ and $b(t)$ are governed by the linear differential equations $\dot{a}(t) = -\mu a(t)$ and $\dot{b}(t) = \sigma b(t)$, whose solutions are exponentials. Interesting competition arises as a consequence of the nonlinear process governed by the rate λ. In an exact representation of the system's temporal evolution, we would now need to know the probability of finding an $A - B$ pair at time t. Moreover, in a spatial Lotka–Volterra model, defined on a d-dimensional lattice, say, on which the individual particles can move via nearest-neighbor hopping, the predation reaction should occur only if both predators and prey occupy the same or adjacent sites. The evolution equations for the mean densities $a(t)$ and $b(t)$ would then have to be respectively amended by the terms $\pm\lambda\langle a(x,t)b(x,t)\rangle$. Here $a(x,t)$ and $b(x,t)$ represent local concentrations, the brackets denote the ensemble average, and $\langle a(x,t)b(x,t)\rangle$ represents $A - B$ cross correlations.

In the rate equation approximation, it is assumed that the local densities are uncorrelated, whereupon $\langle a(x,t)b(x,t)\rangle$ factorizes to $\langle a(x,t)\rangle\langle b(x,t)\rangle = a(t)b(t)$. This yields the famous deterministic Lotka–Volterra equations

$$\dot{a}(t) = \lambda a(t)b(t) - \mu a(t), \quad \dot{b}(t) = \sigma b(t) - \lambda a(t)b(t). \quad (1)$$

Within this mean-field approximation, the quantity $K(t) = \lambda[a(t) + b(t)] - \sigma \ln a(t) - \mu \ln b(t)$ (essentially the system's Lyapunov function) is a constant of motion, $\dot{K}(t) = 0$. This results in regular nonlinear population oscillations, whose frequency and amplitude are fully determined by the initial conditions, a rather unrealistic feature. Moreover Eqs. (1) are known to be unstable with respect to various model modifications (as discussed in [3]).

In contrast with the rate equation predictions, the original stochastic spatial Lotka–Volterra system displays much richer behavior (a recent overview is presented in [4]): The predator–prey coexistence phase is governed, for sufficiently large values of the predation rate, by an incessant sequence of 'pursuit and evasion' wave fronts

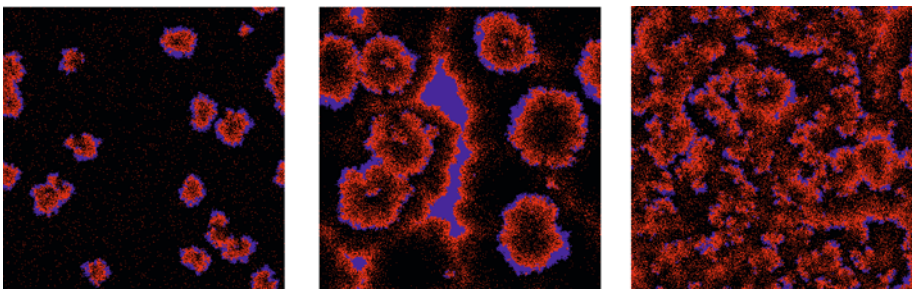

Field Theoretic Methods, Figure 1
Snapshots of the time evolution (*left* to *right*) of activity fronts emerging in a stochastic Lotka–Volterra model simulated on a 512 × 512 lattice, with periodic boundary conditions and site occupation numbers restricted to 0 or 1. For the chosen reaction rates, the system is in the species coexistence phase (with rates $\sigma = 4.0$, $\mu = 0.1$, and $\lambda = 2.2$), and the corresponding mean-field fixed point a focus. The *red*, *blue*, and *black dots* respectively represent predators *A*, prey *B*, and empty sites Ø. Reproduced with permission from [4]

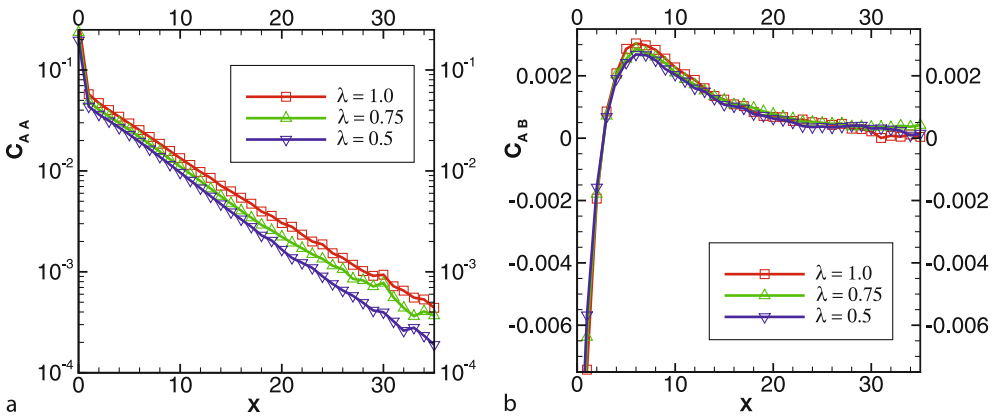

Field Theoretic Methods, Figure 2
Static correlation functions a $C_{AA}(x)$ (note the logarithmic scale), and b $C_{AB}(x)$, measured in simulations on a 1024 × 1024 lattice without any restrictions on the site occupations. The reaction rates were $\sigma = 0.1$, $\mu = 0.1$, and λ was varied from 0.5 (*blue triangles, upside down*), 0.75 (*green triangles*), to 1.0 (*red squares*). Reproduced with permission from [5]

that form quite complex dynamical patterns, as depicted in Fig. 1, which shows snapshots taken in a two-dimensional lattice Monte Carlo simulation where each site could at most be occupied by a single particle. In finite systems, these correlated structures induce erratic population oscillations whose features are independent of the initial configuration. Moreover, if locally the prey 'carrying capacity' is limited (corresponding to restricting the maximum site occupation number per lattice site), there appears an extinction threshold for the predator population that separates the active coexistence regime through a continuous phase transition from a state wherein at long times $t \to \infty$ only prey survive. With respect to the predator population, this represents an *absorbing state*: Once all A particles have vanished, they cannot be produced by the stochastic kinetics.

A quantitative characterization of the emerging spatial structures utilizes equal-time correlation functions such as $C_{AA}(x - x', t) = \langle a(x,t)a(x',t)\rangle - a(t)^2$ and $C_{AB}(x - x', t) = \langle a(x,t)b(x',t)\rangle - a(t)b(t)$, computed at some large time t in the (quasi-)stationary state. These are shown in Fig. 2 as measured in computer simulations for a stochastic Lotka–Volterra model (but here no restrictions on the site occupation numbers of the A or B particles were implemented). The $A - A$ (and $B - B$) correlations obviously decay essentially exponentially with distance x, $C_{AA}(x) \propto C_{BB}(x) \propto e^{-|x|/\xi}$, with roughly equal correlation lengths ξ for the predators and prey. The cross-correlation function $C_{AB}(x)$ displays a maximum at six lattice spacings; these positive correlations indicate the spatial extent of the emerging activity fronts (prey followed by the predators). At closer distance, the A and B particles become *anti*-correlated ($C_{AB}(x) < 0$ for $|x| < 3$): prey would not survive close encounters with the predators. In a similar manner, one can address temporal correlations. These appear prominently in the space-time plot of Fig. 3 obtained for a Monte Carlo run on a one-dimensional lattice (no site occupation restrictions), indicating localized population explosion and extinction events.

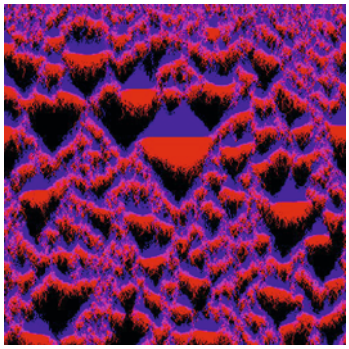

Field Theoretic Methods, Figure 3
Space-time plot (space horizontal, with periodic boundary conditions; time vertical, proceeding downward) showing the temporal evolution of a one-dimensional stochastic Lotka–Volterra model on 512 lattice sites, but without any restrictions on the site occupation numbers (*red*: predators, *blue*: prey, *magenta*: sites occupied by both species; rates: $\sigma = 0.1$, $\mu = 0.1$, $\lambda = 0.1$). Reproduced with permission from [5]

Correlation Functions and Field Theory

The above example demonstrates that stochastic fluctuations and correlations induced by the dynamical inter-

actions may lead to important features that are not adequately described by mean-field approaches. We thus require tools that allow us to systematically account for fluctuations in the mathematical description of stochastic complex systems and evaluate characteristic correlations. Such a toolbox is provided through *field theory* representations that are conducive to the identification of underlying symmetries and have proven useful starting points for the construction of various approximation schemes. These methods were originally devised and elaborated in the theory of (quantum and classical) many-particle systems and quantum fields ([6,7,8,9,10,11,12,13] represent a sample of recent textbooks).

Generating Functions

The basic structure of these field theories rests in a (normalized) exponential probability distribution $\mathcal{P}[S_i]$ for the N relevant variables S_i, $i = 1, \ldots, N$: $\int \prod_{i=1}^{N} \mathrm{d}S_i \, \mathcal{P}[S_i] = 1$, where the integration extends over the allowed range of values for the S_i; i. e.,

$$\begin{aligned} \mathcal{P}[S_i] &= \frac{1}{\mathcal{Z}} \exp\left(-\mathcal{A}[S_i]\right) , \\ \mathcal{Z} &= \int \prod_{i=1}^{N} \mathrm{d}S_i \exp\left(-\mathcal{A}[S_i]\right) . \end{aligned} \tag{2}$$

In canonical equilibrium statistical mechanics, $\mathcal{A}[S_i] = \mathcal{H}[S_i]/k_\mathrm{B}T$ is essentially the Hamiltonian, and the normalization is the partition function $\mathcal{Z}$. In Euclidean quantum field theory, the action $\mathcal{A}[S_i]$ is given by the Langrangian.

All observables $\mathcal{O}$ should be functions of the basic degrees of freedom S_i; their ensemble average thus becomes

$$\begin{aligned} \langle \mathcal{O}[S_i] \rangle &= \int \prod_{i=1}^{N} \mathrm{d}S_i \mathcal{O}[S_i] \mathcal{P}[S_i] \\ &= \frac{1}{\mathcal{Z}} \int \prod_{i=1}^{N} \mathrm{d}S_i \mathcal{O}[S_i] \exp\left(-\mathcal{A}[S_i]\right) . \end{aligned} \tag{3}$$

If we are interested in n-point correlations, i. e., expectation values of the products of the variables S_i, it is useful to define a *generating function*

$$\mathcal{W}[j_i] = \left\langle \exp \sum_{i=1}^{N} j_i S_i \right\rangle , \tag{4}$$

with $\mathcal{W}[j_i = 0] = 1$. Notice that $\mathcal{W}[j_i]$ formally is just the Laplace transform of the probability distribution $\mathcal{P}[S_i]$. The correlation functions can now be obtained via partial derivatives of $\mathcal{W}[j_i]$ with respect to the sources j_i:

$$\langle S_{i_1} \ldots S_{i_n} \rangle = \frac{\partial}{\partial j_{i_1}} \cdots \frac{\partial}{\partial j_{i_n}} \mathcal{W}[j_i] \Big|_{j_i=0} . \tag{5}$$

Connected correlation functions or cumulants can be found by similar partial derivatives of the logarithm of the generating function:

$$\langle S_{i_1} \ldots S_{i_n} \rangle_c = \frac{\partial}{\partial j_{i_1}} \cdots \frac{\partial}{\partial j_{i_n}} \ln \mathcal{W}[j_i] \Big|_{j_i=0} , \tag{6}$$

e. g., $\langle S_i \rangle_c = \langle S_i \rangle$, and $\langle S_i S_j \rangle_c = \langle S_i S_j \rangle - \langle S_i \rangle \langle S_j \rangle = \langle (S_i - \langle S_i \rangle)(S_j - \langle S_j \rangle) \rangle$.

Perturbation Expansion

For a Gaussian action, i. e., a quadratic form $\mathcal{A}_0[S_i] = \frac{1}{2} \sum_{ij} S_i A_{ij} S_j$ (for simplicity we assume real variables S_i), one may readily compute the corresponding generating function $\mathcal{W}_0[j_i]$. After diagonalizing the symmetric $N \times N$ matrix A_{ij}, completing the squares, and evaluating the ensuing Gaussian integrals, one obtains

$$\begin{aligned} \mathcal{Z}_0 &= \frac{(2\pi)^{N/2}}{\sqrt{\det A}} , \\ \mathcal{W}_0[j_i] &= \exp\left(\frac{1}{2} \sum_{i,j=1}^{N} j_i A^{-1}_{ij} j_j \right) , \quad \left\langle S_i S_j \right\rangle_0 = A^{-1}_{ij} . \end{aligned} \tag{7}$$

Thus, the two-point correlation functions in the Gaussian ensemble are given by the elements of the inverse harmonic coupling matrix. An important special property of the Gaussian ensemble is that all n-point functions with odd n vanish, whereas those with even n factorize into sums of all possible permutations of products of two-point functions A^{-1}_{ij} that can be constructed by pairing up the variables S_i (Wick's theorem). For example, the four-point function reads $\langle S_i S_j S_k S_l \rangle_0 = A^{-1}_{ij} A^{-1}_{kl} + A^{-1}_{ik} A^{-1}_{jl} + A^{-1}_{il} A^{-1}_{jk}$.

Let us now consider a general action, isolate the Gaussian contribution, and label the remainder as the nonlinear, anharmonic, or interacting part, $\mathcal{A}[S_i] = \mathcal{A}_0[S_i] + \mathcal{A}_\int[S_i]$. We then observe that

$$\begin{aligned} \mathcal{Z} &= \mathcal{Z}_0 \left\langle \exp\left(-\mathcal{A}_\int[S_i]\right) \right\rangle_0 , \\ \langle \mathcal{O}[S_i] \rangle &= \frac{\left\langle \mathcal{O}[S_i] \exp\left(-\mathcal{A}_\int[S_i]\right) \right\rangle_0}{\left\langle \exp\left(-\mathcal{A}_\int[S_i]\right) \right\rangle_0} , \end{aligned} \tag{8}$$

where the index 0 indicates that the expectation values are computed in the Gaussian ensemble. The nonlinear terms

in Eq. (8) may now be treated perturbatively by expanding the exponentials in the numerator and denominator with respect to the interacting part $\mathcal{A}_{\int}[S_i]$:

$$\langle \mathcal{O}[S_i] \rangle = \frac{\left\langle \mathcal{O}[S_i] \sum_{\ell=0}^{\infty} \frac{1}{\ell!} \left(-\mathcal{A}_{\int}[S_i] \right)^{\ell} \right\rangle_0}{\left\langle \sum_{\ell=0}^{\infty} \frac{1}{\ell!} \left(-\mathcal{A}_{\int}[S_i] \right)^{\ell} \right\rangle_0} . \tag{9}$$

If the interaction terms are polynomial in the variables S_i, Wick's theorem reduces the calculation of n-point functions to a summation of products of Gaussian two-point functions. Since the number of contributing terms grows factorially with the order ℓ of the perturbation expansion, graphical representations in terms of Feynman diagrams become very useful for the classification and evaluation of the different contributions to the perturbation series. Basically, they consist of lines representing the Gaussian two-point functions ('propagators') that are connected to vertices that stem from the (polynomial) interaction terms; for details, see, e. g., [6,7,8,9,10,11,12,13].

Continuum Limit and Functional Integrals

Discrete spatial degrees of freedom are already contained in the above formal description: for example, on a d-dimensional lattice with N_d sites the index i for the fields S_i merely needs to entail the site labels, and the total number of degrees of freedom is just $N = N_d$ times the number of independent relevant quantities. Upon discretizing time, these prescriptions can be extended in effectively an additional dimension to systems with temporal evolution. We may at last take the *continuum limit* by letting $N \to \infty$, while the lattice constant and elementary time step tend to zero in such a manner that macroscopic dynamical features are preserved. Formally, this replaces sums over lattice sites and time steps with spatial and temporal integrations; the action $\mathcal{A}[S_i]$ becomes a functional of the fields $S_i(x,t)$; partial derivatives turn into functional derivatives; and functional integrations $\int \prod_{i=1}^{N} \mathrm{d}S_i \to \int \mathcal{D}[S_i]$ are to be inserted in the previous expressions. For example, Eqs. (3), (4) and (6) become

$$\langle \mathcal{O}[S_i] \rangle = \frac{1}{\mathcal{Z}} \int \mathcal{D}[S_i] \mathcal{O}[S_i] \exp\left(-\mathcal{A}[S_i]\right) , \tag{10}$$

$$\mathcal{W}[j_i] = \left\langle \exp \int \mathrm{d}^d x \int \mathrm{d}t \sum_i j_i(x,t) S_i(x,t) \right\rangle , \tag{11}$$

$$\left\langle \prod_{j=1}^{n} S_{i_j}(x_j,t_j) \right\rangle_c = \prod_{j=1}^{n} \frac{\delta}{\delta j_{i_j}(x_j,t_j)} \ln \mathcal{W}[j_i] \Big|_{j_i=0} . \tag{12}$$

Thus we have arrived at a continuum field theory. Nevertheless, we may follow the procedures outlined above; specifically, the perturbation expansion expressions (8) and (9) still hold, yet with arguments $S_i(x,t)$ that are now fields depending on continuous space-time parameters.

More than thirty years ago, Janssen and De Dominicis independently derived a mapping of the stochastic kinetics defined through nonlinear Langevin equations onto a field theory action ([14,15]; reviewed in [16]). Almost simultaneously, Doi constructed a Fock space representation and therefrom a stochastic field theory for classical interacting particle systems from the master equation describing the corresponding stochastic processes [17,18]. His approach was further developed by several authors into a powerful method for the study of internal noise and correlation effects in reaction-diffusion systems ([19,20,21,22,23]; for recent reviews, see [24,25]). We shall see below that the field-theoretic representations of both classical master and Langevin equations require *two* independent fields for each stochastic variable. Otherwise, the computation of correlation functions and the construction of perturbative expansions fundamentally works precisely as sketched above. But the underlying causal temporal structure induces important specific features such as the absence of 'vacuum diagrams' (closed response loops): the denominator in Eq. (2) is simply $\mathcal{Z} = 1$. (For unified and more detailed descriptions of both versions of dynamic stochastic field theories, see [26,27].)

Discrete Stochastic Interacting Particle Systems

We first outline the mapping of stochastic interacting particle dynamics as defined through a master equation onto a field theory action [17,18,19,20,21,22,23]. Let us denote the configurational probability for a stochastically evolving system to be in state α at time t with $P(\alpha;t)$. Given the transition rates $W_{\alpha\to\beta}(t)$ from states α to β, a *master equation* essentially balances the transitions into and out of each state:

$$\frac{\partial P(\alpha;t)}{\partial t} = \sum_{\beta\neq\alpha} \left[W_{\beta\to\alpha}(t) P(\beta;t) - W_{\alpha\to\beta}(t) P(\alpha;t) \right] . \tag{13}$$

The dynamics of many complex systems can be cast into the language of 'chemical' reactions, wherein certain particle species (upon encounter, say) transform into different species with fixed (time-independent) reaction rates. The 'particles' considered here could be atoms or molecules in chemistry, but also individuals in population dynamics (as in our example in Sect. "Example: Lotka–Volterra Model"), or appropriate effective degrees of freedom governing the system's kinetics, such as domain walls in mag-

nets, etc. To be specific, we envision our particles to propagate via unbiased random walks (diffusion) on a d-dimensional hypercubic lattice, with the reactions occuring according to prescribed rules when particles meet on a lattice site. This stochastic interacting particle system is then at any time fully characterized by the number of particles $n_A, n_B, \ldots$ of each species $A, B, \ldots$ located on any lattice site. The following describes the construction of an associated field theory action. As important examples, we briefly discuss annihilation reactions and absorbing state phase transitions.

Master Equation and Fock Space Representation

The formal procedures are best explained by means of a simple example; thus consider the irreversible binary annihilation process $A + A \to A$, happening with rate λ. In terms of the occupation numbers n_i of the lattice sites i, we can construct the master equation associated with these on-site reactions as follows. The annihilation process locally changes the occupation numbers by one; the transition rate from a state with n_i particles at site i to $n_i - 1$ particles is $W_{n_i \to n_i - 1} = \lambda n_i(n_i - 1)$, whence

$$\frac{\partial P(n_i;t)}{\partial t} = \lambda(n_i+1)n_i P(n_i+1;t) - \lambda n_i(n_i-1)P(n_i;t) \tag{14}$$

represents the master equation for this reaction at site i. As an initial condition, we can for example choose a Poisson distribution $P(n_i) = \bar{n}_0^{n_i} \mathrm{e}^{-\bar{n}_0}/n_i!$ with mean initial particle density $\bar{n}_0$. In order to capture the complete stochastic dynamics, we just need to add similar contributions describing other processes, and finally sum over all lattice sites i.

Since the reactions all change the site occupation numbers by integer values, a Fock space representation (borrowed from quantum mechanics) turns out particularly useful. To this end, we introduce the harmonic oscillator or bosonic ladder operator algebra $[a_i, a_j] = 0 = [a_i^\dagger, a_j^\dagger]$, $[a_i, a_j^\dagger] = \delta_{ij}$, from which we construct the particle number eigenstates $|n_i\rangle$, namely $a_i|n_i\rangle = n_i|n_i - 1\rangle$, $a_i^\dagger|n_i\rangle = |n_i + 1\rangle$, $a_i^\dagger a_i|n_i\rangle = n_i|n_i\rangle$. (Notice that a different normalization than in ordinary quantum mechanics has been employed here.) A general state with n_i particles on sites i is obtained from the 'vacuum' configuration $|0\rangle$, defined via $a_i|0\rangle = 0$, through the product $|\{n_i\}\rangle = \prod_i a_i^{\dagger n_i}|0\rangle$.

To implement the stochastic kinetics, we introduce a formal state vector as a linear combination of all possible states weighted by the time-dependent configurational probability:

$$|\Phi(t)\rangle = \sum_{\{n_i\}} P(\{n_i\};t)|\{n_i\}\rangle \,. \tag{15}$$

Simple manipulations then transform the linear time evolution according to the master equation into an 'imaginary-time' Schrödinger equation

$$\frac{\partial|\Phi(t)\rangle}{\partial t} = -H|\Phi(t)\rangle \,, \quad |\Phi(t)\rangle = \mathrm{e}^{-Ht}|\Phi(0)\rangle \tag{16}$$

governed by a stochastic quasi-Hamiltonian (rather, the Liouville time evolution operator). For on-site reaction processes, $H_{\mathrm{reac}} = \sum_i H_i(a_i^\dagger, a_i)$ is a sum of local contributions; e. g., for the binary annihilation reaction, $H_i(a_i^\dagger, a_i) = -\lambda(1 - a_i^\dagger)a_i^\dagger a_i^2$. It is a straightforward exercise to construct the corresponding expressions within this formalism for the generalization $kA \to \ell A$,

$$H_i(a_i^\dagger, a_i) = -\lambda\left(a_i^{\dagger\ell} - a_i^{\dagger k}\right)a_i^k, \tag{17}$$

and for nearest-neighbor hopping with rate D between adjacent sites $\langle ij\rangle$,

$$H_{\mathrm{diff}} = D\sum_{<ij>}\left(a_i^\dagger - a_j^\dagger\right)\left(a_i - a_j\right) \,. \tag{18}$$

The two contributions for each process may be interpreted as follows: The first term in Eq. (17) corresponds to the actual process, and describes how many particles are annihilated and (re-)created in each reaction. The second term encodes the 'order' of each reaction, i. e., the number operator $a_i^\dagger a_i$ appears to the kth power, but in the normal-ordered form $a_i^{\dagger k} a_i^k$, for a kth-order process. These procedures are readily adjusted for reactions involving multiple particle species. We merely need to specify the occupation numbers on each site and correspondingly introduce additional ladder operators $b_i, c_i, \ldots$ for each new species, with $[a_i, b_i^\dagger] = 0 = [a_i, c_i^\dagger]$ etc. For example, consider the reversible reaction $kA + \ell B \rightleftharpoons mC$ with forward rate λ and backward rate σ; the associated reaction Hamiltonian reads

$$H_{\mathrm{reac}} = -\sum_i\left(c_i^{\dagger m} - a_i^{\dagger k} b_i^{\dagger\ell}\right)\left(\lambda a_i^k b_i^\ell - \sigma c_i^m\right) \,. \tag{19}$$

Similarly, for the Lotka–Volterra model of Sect. "Example: Lotka–Volterra Model", one finds

$$H_{\mathrm{reac}} = -\sum_i\Big[\mu\left(1 - a_i^\dagger\right)a_i + \sigma\left(b_i^\dagger - 1\right)b_i^\dagger b_i + \lambda\left(a_i^\dagger - b_i^\dagger\right)a_i^\dagger a_i b_i\Big] \,. \tag{20}$$

Note that all the above quasi-Hamiltonians are non-Hermitean operators, which naturally reflects the creation and destruction of particles.

Our goal is to compute averages and correlation functions with respect to the configurational probability $P(\{n_i\};t)$. Returning to a single-species system (again, the generalization to many particle species is obvious), this is accomplished with the aid of the projection state $\langle\mathcal{P}| = \langle 0|\prod_i \mathrm{e}^{a_i}$, for which $\langle\mathcal{P}|0\rangle = 1$ and $\langle\mathcal{P}|a_i^\dagger = \langle\mathcal{P}|$, since $[\mathrm{e}^{a_i}, a_j^\dagger] = \mathrm{e}^{a_i}\delta_{ij}$. For the desired statistical averages of observables (which must all be expressible as functions of the occupation numbers $\{n_i\}$), one obtains

$$\langle\mathcal{O}(t)\rangle = \sum_{\{n_i\}} \mathcal{O}(\{n_i\})P(\{n_i\};t) = \langle\mathcal{P}|\mathcal{O}(\{a_i^\dagger a_i\})|\Phi(t)\rangle\,. \tag{21}$$

For example, as a consequence of probability conservation, $1 = \langle\mathcal{P}|\Phi(t)\rangle = \langle\mathcal{P}|\mathrm{e}^{-Ht}|\Phi(0)\rangle$. Thus necessarily $\langle\mathcal{P}|H = 0$; upon commuting $\mathrm{e}^{\sum_i a_i}$ with H, the creation operators are shifted $a_i^\dagger \to 1 + a_i^\dagger$, whence this condition is fulfilled provided $H_i(a_i^\dagger \to 1, a_i) = 0$, which is indeed satisfied by our above explicit expressions (17) and (18). Through this prescription, we may replace $a_i^\dagger a_i \to a_i$ in all averages; e. g., the particle density becomes $a(t) = \langle a_i(t)\rangle$.

In the bosonic operator representation above, we have assumed that no restrictions apply to the particle occupation numbers n_i on each site. If $n_i \le 2s+1$, one may instead employ a representation in terms of spin s operators. For example, particle exclusion systems with $n_i = 0$ or 1 can thus be mapped onto non-Hermitean spin 1/2 'quantum' systems (for recent overviews, see [28,29]). Specifically in one dimension, such representations in terms of integrable spin chains have been very fruitful. An alternative approach uses the bosonic theory, but incorporates the site occupation restrictions through exponentials in the number operators $\mathrm{e}^{-a_i^\dagger a_i}$ [30].

Continuum Limit and Field Theory

As a next step, we follow an established route in quantum many-particle theory [8] and proceed towards a field theory representation through constructing the path integral equivalent to the 'Schrödinger' dynamics (16) based on coherent states, which are right eigenstates of the annihilation operator, $a_i|\phi_i\rangle = \phi_i|\phi_i\rangle$, with complex eigenvalues ϕ_i. Explicitly, $|\phi_i\rangle = \exp\left(-\frac{1}{2}|\phi_i|^2 + \phi_i a_i^\dagger\right)|0\rangle$, and these coherent states satisfy the overlap formula $\langle\phi_j|\phi_i\rangle = \exp\left(-\frac{1}{2}|\phi_i|^2 - \frac{1}{2}|\phi_j|^2 + \phi_j^*\phi_i\right)$, and the (over-)completeness relation $\int\prod_i \mathrm{d}^2\phi_i|\{\phi_i\}\rangle\langle\{\phi_i\}| = \pi$. Upon splitting the temporal evolution (16) into infinitesimal increments, standard procedures (elaborated in detail in [25]) eventually yield an expression for the configurational average

$$\langle\mathcal{O}(t)\rangle \propto \int\prod_i \mathrm{d}\phi_i\,\mathrm{d}\phi_i^*\,\mathcal{O}(\{\phi_i\})\mathrm{e}^{-\mathcal{A}[\phi_i^*,\phi_i;t]}\,, \tag{22}$$

which is of the form (3), with the action

$$\mathcal{A}[\phi_i^*,\phi_i;t_f] = \sum_i\Bigg(-\phi_i(t_f) + \int_0^{t_f}\mathrm{d}t\left[\phi_i^*\frac{\partial\phi_i}{\partial t} + H_i(\phi_i^*,\phi_i)\right] - \bar{n}_0\phi_i^*(0)\Bigg)\,, \tag{23}$$

where the first term originates from the projection state, and the last one stems from the initial Poisson distribution. Through this procedure, in the original quasi-Hamiltonian the creation and annihilation operators $a_i^\dagger$ and a_i are simply replaced with the complex numbers ϕ_i^* and ϕ_i.

Finally, we proceed to the continuum limit, $\phi_i(t) \to \psi(\mathbf{x},t)$, $\phi_i^*(t) \to \hat{\psi}(\mathbf{x},t)$. The 'bulk' part of the action then becomes

$$\mathcal{A}[\hat{\psi},\psi] = \int\mathrm{d}^d x \cdot \int\mathrm{d}t\left[\hat{\psi}\left(\frac{\partial}{\partial t} - D\nabla^2\right)\psi + \mathcal{H}_{\mathrm{reac}}(\hat{\psi},\psi)\right]\,, \tag{24}$$

where the discrete hopping contribution (18) has naturally turned into a continuum diffusion term. We have thus arrived at a *microscopic* field theory for stochastic reaction–diffusion processes, without invoking any assumptions on the form or correlations of the internal reaction noise. Note that we require two independent fields $\hat{\psi}$ and ψ to capture the stochastic dynamics. Actions of the type (24) may serve as a basis for further systematic coarse-graining, constructing a perturbation expansion as outlined in Sect. "Perturbation Expansion", and perhaps a subsequent renormalization group analysis [25,26,27]. We remark that it is often useful to perform a shift in the field $\hat{\psi}$ about the mean-field solution, $\hat{\psi}(x,t) = 1 + \widetilde{\psi}(x,t)$. For occasionally, the resulting field theory action allows the derivation of an equivalent Langevin dynamics, see Sect. "Stochastic Differential Equations" below.

Annihilation Processes

Let us consider our simple single-species example $kA \to \ell A$. The reaction part of the corresponding field

theory action reads

$$\mathcal{H}_{\text{reac}}(\hat{\psi}, \psi) = -\lambda \left(\hat{\psi}^{\ell} - \hat{\psi}^{k}\right) \psi^{k} , \tag{25}$$

see Eq. (17). It is instructive to study the *classical field equations*, namely $\delta\mathcal{A}/\delta\psi = 0$, which is always solved by $\hat{\psi} = 1$, reflecting probability conservation, and $\delta\mathcal{A}/\delta\hat{\psi} = 0$, which, upon inserting $\hat{\psi} = 1$ yields

$$\frac{\partial\psi(x,t)}{\partial t} = D\nabla^2\psi(x,t) - (k-\ell)\lambda\psi(x,t)^k , \tag{26}$$

i. e., the mean-field equation for the local particle density $\psi(x,t)$, supplemented with a diffusion term. For $k = 1$, the particle density grows ($k < \ell$) or decays ($k > \ell$) exponentially. The solution of the rate equation for $k > 1$, $a(t) = \langle\psi(x,t)\rangle = \left[a(0)^{1-k} + (k-l)(k-1)\lambda t\right]^{-1/(k-1)}$ implies a divergence within a finite time for $k < \ell$, and an algebraic decay $\sim (\lambda t)^{-1/(k-1)}$ for $k > \ell$.

The full field theory action, which was derived from the master equation defining the very stochastic process, provides a means of systematically including fluctuations in the mathematical treatment. Through a dimensional analysis, we can determine the (upper) *critical dimension* below which fluctuations become sufficiently strong to alter these power laws. Introducing an inverse length scale κ, $[x] \sim \kappa^{-1}$, and applying diffusive temporal scaling, $[Dt] \sim \kappa^{-2}$, and $[\hat{\psi}(x,t)] \sim \kappa^0$, $[\psi(x,t)] \sim \kappa^d$ in d spatial dimensions, the reaction rate in terms of the diffusivity scales according to $[\lambda/D] \sim \kappa^{2-(k-1)d}$. In large dimensions, the kinetics is *reaction-limited*, and at least qualitatively correctly described by the mean-field rate equation. In low dimensions, the dynamics becomes *diffusion-limited*, and the annihilation reactions generate depletion zones and spatial particle anti-correlations that slow down the density decay. The nonlinear coupling λ/D becomes dimensionless at the boundary critical dimension $d_c(k) = 2/(k-1)$ that separates these two distinct regimes. Thus in physical dimensions, intrinsic stochastic fluctuations are relevant only for pair and triplet annihilation reactions. By means of a renormalization group analysis (for details, see [25]) one finds for $k = 2$ and $d < d_c(2) = 2$: $a(t) \sim (Dt)^{-d/2}$ [21,22], as confirmed by exact solutions in one dimension. Precisely at the critical dimension, the mean-field decay laws acquire logarithmic corrections, namely $a(t) \sim (Dt)^{-1}\ln(Dt)$ for $k = 2$ at $d_c(2) = 2$, and $a(t) \sim \left[(Dt)^{-1}\ln(Dt)\right]^{1/2}$ for $k = 3$ at $d_c(3) = 1$. Annihilation reaction between different species (e. g., $A + B \to \emptyset$) may introduce additional correlation effects, such as particle segregation and the confinement of active dynamics to narrow reaction zones [23]; a recent overview can be found in [25].

Active to Absorbing State Phase Transitions

Competition between particle production and decay processes leads to even richer scenarios, and can induce genuine nonequilibrium transitions that separate 'active' phases (wherein the particle densities remain nonzero in the long-time limit) from 'inactive' stationary states (where the concentrations ultimately vanish). A special but abundant case are *absorbing states*, where, owing to the absence of any agents, stochastic fluctuations cease entirely, and no particles can be regenerated [31,32]. These occur in a variety of systems in nature ([33,34] contain extensive discussions of various model systems); examples are chemical reactions involving an inert state $\emptyset$, wherefrom no reactants A are released anymore, or stochastic population dynamics models, combining diffusive migration of a species A with asexual reproduction $A \to 2A$ (with rate σ), spontaneous death $A \to \emptyset$ (at rate μ), and lethal competition $2A \to A$ (with rate λ). In the inactive state, where no population members A are left, clearly all processes terminate. Similar effective dynamics may be used to model certain nonequilibrium physical systems, such as the domain wall kinetics in Ising chains with competing Glauber and Kawasaki dynamics. Here, spin flips $\uparrow\uparrow\downarrow\downarrow\to\uparrow\uparrow\uparrow\downarrow$ and $\uparrow\uparrow\downarrow\uparrow\to\uparrow\uparrow\uparrow\uparrow$ may be viewed as domain wall (A) hopping and pair annihilation $2A \to \emptyset$, whereas spin exchange $\uparrow\uparrow\downarrow\downarrow\to\uparrow\downarrow\uparrow\downarrow$ represents a branching process $A \to 3A$. Notice that the para- and ferromagnetic phases respectively map onto the active and inactive 'particle' states. The ferromagnetic state becomes absorbing if the spin flip rates are taken at zero temperature.

The reaction quasi-Hamiltonian corresponding to the stochastic dynamics of the aforementioned population dynamics model reads

$$\mathcal{H}_{\text{reac}}(\hat{\psi}, \psi) = \left(1 - \hat{\psi}\right)\left(\sigma\hat{\psi}\psi - \mu\psi - \lambda\hat{\psi}\psi^2\right) . \tag{27}$$

The associated rate equation is the Fisher–Kolmogorov equation (see Murray 2002 [3])

$$\dot{a}(t) = (\sigma - \mu)a(t) - \lambda a(t)^2 , \tag{28}$$

which yields both inactive and active phases: For $\sigma < \mu$ we have $a(t \to \infty) \to 0$, whereas for $\sigma > \mu$ the density eventually saturates at $a_s = (\sigma - \mu)/\lambda$. The explicit time-dependent solution $a(t) = a(0)a_s/[a(0) + [a_s - a(0)]e^{(\mu-\sigma)t}]$ shows that both stationary states are approached exponentially in time. They are separated by a continuous nonequilibrium phase transition at $\sigma = \mu$, where the temporal decay becomes algebraic, $a(t) = a(0)/[1+a(0)\lambda t]) \to 1/(\lambda t)$ as $t \to \infty$, independent of the initial density $a(0)$. As in second-order equilibrium phase transitions, however, critical fluctuations are expected to

invalidate the mean-field power laws in low dimensions $d < d_c$.

If we now shift the field $\hat{\psi}$ about its stationary value 1 and rescale according to $\hat{\psi}(\mathbf{x}, t) = 1 + \sqrt{\sigma/\lambda}\,\widetilde{S}(\mathbf{x}, t)$ and $\psi(\mathbf{x}, t) = \sqrt{\lambda/\sigma}\,S(\mathbf{x}, t)$, the (bulk) action becomes

$$\mathcal{A}\left[\widetilde{S}, S\right] = \int \mathrm{d}^d x \int \mathrm{d}t \left[\widetilde{S}\left(\frac{\partial}{\partial t} + D\left(r - \nabla^2\right)\right) S - u\left(\widetilde{S} - S\right)\widetilde{S}S + \lambda \widetilde{S}^2 S^2\right]. \quad (29)$$

Thus, the three-point vertices have been scaled to identical coupling strengths $u = \sqrt{\sigma\lambda}$, which in fact represents the effective coupling of the perturbation expansion. Its scaling dimension is $[u] = \mu^{2-d/2}$, whence we infer the upper critical dimension $d_c = 4$. The four-point vertex $\propto \lambda$, with $[\lambda] = \mu^{2-d}$, is then found to be irrelevant in the renormalization group sense, and can be dropped for the computation of universal, asymptotic scaling properties. The action (29) with $\lambda = 0$ is known as Reggeon field theory [35]; it satisfies a characteristic symmetry, namely invariance under so-called rapidity inversion $S(\mathbf{x}, t) \leftrightarrow -\widetilde{S}(\mathbf{x}, -t)$. Remarkably, it has moreover been established that the field theory action (29) describes the scaling properties of critical directed percolation clusters [36,37,38]. The fluctuation-corrected universal power laws governing the vicinity of the phase transition can be extracted by renormalization group methods (reviewed for directed percolation in [39]). Table 1 compares the analytic results obtained in an ε expansion about the critical dimension ($\epsilon = 4 - d$) with the critical exponent values measured in Monte Carlo computer simulations [33,34].

According to a conjecture originally formulated by Janssen and Grassberger, any continuous nonequilibrium phase transition from an active to an absorbing state in a system governed by Markovian stochastic dynamics that is decoupled from any other slow variable, and in the absence of special additional symmetries or quenched randomness, should in fact fall in the directed percolation universality class [38,40]. This statement has indeed been confirmed in a large variety of model sytems (many examples are listed in [33,34]). It even pertains to multi-species generalizations [41], and applies for instance to the predator extinction threshold in the stochastic Lotka–Volterra model with restricted site occupation numbers mentioned in Sect. "Example: Lotka–Volterra Model" [4].

Stochastic Differential Equations

This section explains how dynamics governed by *Langevin-type stochastic differential equations* can be represented through a field-theoretic formalism [14,15,16]. Such a description is especially useful to capture the effects of external noise on the temporal evolution of the relevant quantities under consideration, which encompasses the case of thermal noise induced by the coupling to a heat bath in thermal equilibrium at temperature T. The underlying assumption in this approach is that there exists a natural *separation of time scales* between the slow variables S_i, and all other degrees of freedom ζ_i which in comparison fluctuate rapidly, and are therefore summarily gathered in zero-mean noise terms, assumed to be uncorrelated in space and time,

$$\begin{aligned} \langle \zeta_i(x, t)\rangle &= 0\,, \\ \left\langle \zeta_i(x, t)\zeta_j(x', t')\right\rangle &= 2L_{ij}[S_i]\delta(x - x')\delta(t - t')\,. \end{aligned} \quad (30)$$

Here, the noise correlator $2L_{ij}[S_i]$ may be a function of the slow system variables S_i, and also contain operators such as spatial derivatives. A general set of coupled Langevin-type stochastic differential equations then takes the form

$$\frac{\partial S_i(t)}{\partial t} = F_i[S_i] + \zeta_i\,, \quad (31)$$

where we may decompose the 'systematic forces' into reversible terms of microscopic origin and relaxational contributions that are induced by the noise and drive the system towards its stationary state (see below), i. e.: $F_i[S_i] = F_i^{\mathrm{rev}}[S_i] + F_i^{\mathrm{rel}}[S_i]$. Both ingredients may contain nonlinear terms as well as mode couplings between

Field Theoretic Methods, Table 1

Comparison of the values for the critical exponents of the directed percolation universality class measured in Monte Carlo simulations with the analytic renormalization group results within the $\epsilon = 4 - d$ expansion: ξ denotes the correlation length, t_c the characteristic relaxation time, a_s the saturation density in the active state, and $a_c(t)$ the critical density decay law

Scaling exponent	d = 1	d = 2	d = 4 − ϵ
$\xi \sim \lvert\tau\rvert^{-\nu}$	$\nu \approx 1.100$	$\nu \approx 0.735$	$\nu = 1/2 + \epsilon/16 + O(\epsilon^2)$
$t_c \sim \xi^z \sim \lvert\tau\rvert^{-z\nu}$	$z \approx 1.576$	$z \approx 1.73$	$z = 2 - \epsilon/12 + O(\epsilon^2)$
$a_s \sim \lvert\tau\rvert^{\beta}$	$\beta \approx 0.2765$	$\beta \approx 0.584$	$\beta = 1 - \epsilon/6 + O(\epsilon^2)$
$a_c(t) \sim t^{-\alpha}$	$\alpha \approx 0.160$	$\alpha \approx 0.46$	$\alpha = 1 - \epsilon/4 + O(\epsilon^2)$

different variables. Again, we first introduce the abstract formalism, and then proceed to discuss relaxation to thermal equilibrium as well as some examples for nonequilibrium Langevin dynamics.

Field Theory Representation of Langevin Equations

The shortest and most general route towards a field theory representation of the Langevin dynamics (31) with noise correlations (30) starts with one of the most elaborate ways to expand unity, namely through a product of functional delta functions (for the sake of compact notations, we immediately employ a functional integration language, but in the end all the path integrals are defined through appropriate discretizations in space and time):

$$\begin{aligned} 1 &= \int \prod_i \mathcal{D}[S_i] \\ &\quad \cdot \prod_{(x,t)} \delta\left(\frac{\partial S_i(x,t)}{\partial t} - F_i[S_i](x,t) - \zeta_i(x,t)\right) \\ &= \int \prod_i \mathcal{D}[i\widetilde{S}_i]\mathcal{D}[S_i] \exp\Big[-\int \mathrm{d}^d x \\ &\quad \cdot \int \mathrm{d}t \sum_i \widetilde{S}_i \left(\frac{\partial S_i}{\partial t} - F_i[S_i] - \zeta_i\right)\Big] . \end{aligned} \tag{32}$$

In the second line we have used the Fourier representation of the (functional) delta distribution by means of the purely imaginary auxiliary variables $\widetilde{S}_i$ (also called Martin–Siggia–Rose response fields [42]). Next we require the explicit form of the noise probability distribution that generates the correlations (30); for simplicity, we may employ the Gaussian

$$\begin{aligned} \mathcal{W}[\zeta_i] \propto \exp\Big[-\frac{1}{4} \int \mathrm{d}^d x \\ \cdot \int_0^{t_f} \mathrm{d}t \sum_{ij} \zeta_i(x,t) \left[L_{ij}^{-1} \zeta_j(x,t)\right]\Big] . \end{aligned} \tag{33}$$

Inserting the identity (32) and the probability distribution (33) into the desired stochastic noise average of any observable $\mathcal{O}[S_i]$, we arrive at

$$\begin{aligned} \langle \mathcal{O}[S_i]\rangle_\zeta &\propto \int \prod_i \mathcal{D}[i\widetilde{S}_i]\mathcal{D}[S_i] \\ &\exp\left[-\int \mathrm{d}^d x \int \mathrm{d}t \sum_i \widetilde{S}_i \left(\frac{\partial S_i}{\partial t} - F_i[S_i]\right)\right] \mathcal{O}[S_i] \\ &\cdot \int \prod_i \mathcal{D}[\zeta_i] \exp\Big(-\int \mathrm{d}^d x \int \mathrm{d}t \\ &\qquad \cdot \sum_i \Big[\frac{1}{4}\zeta_i \sum_j L_{ij}^{-1}\zeta_j - \widetilde{S}_i \zeta_i\Big]\Big) . \end{aligned} \tag{34}$$

Subsequently evaluating the Gaussian integrals over the noise ζ_i yields at last

$$\begin{aligned} \langle \mathcal{O}[S_i]\rangle_\zeta &= \int \prod_i \mathcal{D}[S_i]\mathcal{O}[S_i]\mathcal{P}[S_i] , \\ \mathcal{P}[S_i] &\propto \int \prod_i \mathcal{D}[i\widetilde{S}_i] \mathrm{e}^{-\mathcal{A}[\widetilde{S}_i, S_i]} , \end{aligned} \tag{35}$$

with the statistical weight governed by the Janssen–De Dominicis 'response' functional [14,15]

$$\begin{aligned} \mathcal{A}[\widetilde{S}_i, S_i] &= \int \mathrm{d}^d x \int_0^{t_f} \mathrm{d}t \\ &\cdot \sum_i \left[\widetilde{S}_i \left(\frac{\partial S_i}{\partial t} - F_i[S]\right) - \widetilde{S}_i \sum_j L_{ij} \widetilde{S}_j\right] . \end{aligned} \tag{36}$$

It should be noted that in the above manipulations, we have omitted the functional determinant from the variable change $\{\zeta_i\} \to \{S_i\}$. This step can be justified through applying a forward (Itô) discretization (for technical details, see [16,27,43]). Normalization implies $\int \prod_i \mathcal{D}[i\widetilde{S}_i]\mathcal{D}[S_i]\mathrm{e}^{-\mathcal{A}[\widetilde{S}_i, S]_i} = 1$. The first term in the action (36) encodes the temporal evolution according to the systematic terms in the Langevin Equations (31), whereas the second term specifies the noise correlations (30). Since the auxiliary fields appear only quadratically, they could be eliminated via completing the squares and Gaussian integrations. This results in the equivalent Onsager–Machlup functional which however contains squares of the nonlinear terms and the inverse of the noise correlator operators; the form (36) is therefore usually more convenient for practical purposes. The Janssen–De Dominicis functional (36) takes the form of a $(d+1)$-dimensional statistical field theory with again *two* independent sets of fields S_i and $\widetilde{S}_i$. It may serve as a starting point for systematic approximation schemes including perturbative expansions, and subsequent renormalization group treatments. Causality is properly incorporated in this formalism which has important technical implications [16,27,43].

Thermal Equilibrium and Relaxational Critical Dynamics

Consider the dynamics of a system that following some external perturbation relaxes towards thermal equilibrium governed by the canonical Boltzmann distribution at fixed temperature T,

$$\mathcal{P}_{\mathrm{eq}}[S_i] = \frac{1}{\mathcal{Z}(T)} \exp\left(-\mathcal{H}[S_i]/k_{\mathrm{B}}T\right) . \tag{37}$$

The relaxational term in the Langevin Equation (31) can then be specified as

$$F_i^{\mathrm{rel}}[S_i] = -\lambda_i \frac{\delta \mathcal{H}[S_i]}{\delta S_i} , \tag{38}$$

with Onsager coefficients λ_i; for nonconserved fields, λ_i is a positive relaxation rate. On the other hand, if the variable S_i is a conserved quantity (such as the energy density), there is an associated continuity equation $\partial S_i/\partial t + \nabla \cdot J_i = 0$, with a conserved current that is typically given by a gradient of the field S_i: $J_i = -D_i \nabla S_i + \ldots$; as a consequence, the fluctuations of the fields S_i will relax diffusively with diffusivity D_i, and $\lambda_i = -D_i \nabla^2$ becomes a spatial Laplacian.

In order for $\mathcal{P}(t) \to \mathcal{P}_{\mathrm{eq}}$ as $t \to \infty$, the stochastic Langevin dynamics needs to satisfy *two* conditions, which can be inferred from the associated Fokker–Planck equation [27,44]. First, the reversible probability current is required to be divergence-free in the space spanned by the fields S_i:

$$\int \mathrm{d}^d x \sum_i \frac{\delta}{\delta S_i(x)} \left(F_i^{\mathrm{rev}}[S_i] \mathrm{e}^{-\mathcal{H}[S_i]/k_{\mathrm{B}}T} \right) = 0 . \tag{39}$$

This condition severely constrains the reversible force terms. For example, for a system whose microscopic time evolution is determined through the Poisson brackets $Q_{ij}(x,x') = \{S_i(x), S_j(x')\} = -Q_{ji}(x',x)$ (to be replaced by commutators in quantum mechanics), one finds for the reversible mode-coupling terms [44]

$$F_i^{\mathrm{rev}}[S_i](x) = -\int \mathrm{d}^d x' \cdot \sum_j \left[Q_{ij}(x,x') \frac{\delta \mathcal{H}[S_i]}{\delta S_j(x')} - k_{\mathrm{B}} T \frac{\delta Q_{ij}(x,x')}{\delta S_j(x')} \right] . \tag{40}$$

Second, the noise correlator in Eq. (30) must be related to the Onsager relaxation coefficients through the Einstein relation

$$L_{ij} = k_{\mathrm{B}} T \lambda_i \delta_{ij} . \tag{41}$$

To provide a specific example, we focus on the case of purely relaxational dynamics (i. e., reversible force terms are absent entirely), with the (mesoscopic) Hamiltonian given by the Ginzburg–Landau–Wilson free energy that describes second-order phase transitions in thermal equilibrium for an n-component order parameter S_i, $i = 1, \ldots, N$ [6,7,8,9,10,11,12,13]:

$$\mathcal{H}[S_i] = \int \mathrm{d}^d x \sum_{i=1}^{N} \left[\frac{r}{2} [S_i(x)]^2 + \frac{1}{2} [\nabla S_i(x)]^2 + \frac{u}{4!} [S_i(x)]^2 \sum_{j=1}^{N} [S_j(x)]^2 \right] , \tag{42}$$

where the control parameter $r \propto T - T_{\mathrm{c}}$ changes sign at the critical temperature T_{c}, and the positive constant u governs the strength of the nonlinearity. If we assume that the order parameter itself is not conserved under the dynamics, the associated response functional reads

$$\mathcal{A}\left[\widetilde{S}_i, S_i\right] = \int \mathrm{d}^d x \int \mathrm{d}t \cdot \sum_i \widetilde{S}_i \left(\frac{\partial}{\partial t} + \lambda_i \frac{\delta \mathcal{H}[S_i]}{\delta S_i} - k_{\mathrm{B}} T \lambda_i \widetilde{S}_i \right) . \tag{43}$$

This case is frequently referred to as model A critical dynamics [45]. For a diffusively relaxing conserved field, termed model B in the classification of [45], one has instead

$$\mathcal{A}\left[\widetilde{S}_i, S_i\right] = \int \mathrm{d}^d x \int \mathrm{d}t \cdot \sum_i \widetilde{S}_i \left(\frac{\partial}{\partial t} - D_i \nabla^2 \frac{\delta \mathcal{H}[S_i]}{\delta S_i} + k_{\mathrm{B}} T D_i \nabla^2 \widetilde{S}_i \right) . \tag{44}$$

Consider now the external fields h_i that are thermodynamically conjugate to the mesoscopic variables S_i, i. e., $\mathcal{H}(h_i) = \mathcal{H}(h_i = 0) - \int \mathrm{d}^d x \sum_i h_i(x) S_i(x)$. For the simple relaxational models (43) and (44), we may thus immediately relate the dynamic susceptibility to two-point correlation functions that involve the auxiliary fields $\widetilde{S}_i$ [43], namely

$$\chi_{ij}(x-x', t-t') = \left. \frac{\delta \langle S_i(x,t) \rangle}{\delta h_j(x',t')} \right|_{h_i=0} = k_{\mathrm{B}} T \lambda_i \left\langle S_i(x,t) \widetilde{S}_j(x',t') \right\rangle \tag{45}$$

for nonconserved fields, while for model B dynamics

$$\chi_{ij}(x-x', t-t') = -k_{\mathrm{B}} T D_i \left\langle S_i(x,t) \nabla^2 \widetilde{S}_j(x',t') \right\rangle . \tag{46}$$

Finally, in thermal equilibrium the dynamic response and correlation functions are related through the fluctuation-dissipation theorem [43]

$$\chi_{ij}(x-x', t-t') = \Theta(t-t') \frac{\partial}{\partial t'} \left\langle S_i(x,t) S_j(x',t') \right\rangle . \tag{47}$$

Driven Diffusive Systems and Interface Growth

We close this section by listing a few intriguing examples for Langevin sytems that describe genuine out-of-equilibrium dynamics. First, consider a driven diffusive lattice gas (an overview is provided in [46]), namely a particle system with conserved total density with biased diffusion in a specified ('‖') direction. The coarse-grained Langevin equation for the scalar density fluctuations thus becomes spatially anisotropic [47,48],

$$\frac{\partial S(x,t)}{\partial t} = D\left(\nabla_\perp^2 + c\nabla_\parallel^2\right) S(x,t) + \frac{Dg}{2}\nabla_\parallel S(x,t)^2 + \zeta(x,t), \tag{48}$$

and similarly for the conserved noise with $\langle\zeta\rangle = 0$,

$$\langle\zeta(x,t)\zeta(x',t')\rangle = -2D\left(\nabla_\perp^2 + \tilde{c}\nabla_\parallel^2\right)\delta(x-x')\delta(t-t'). \tag{49}$$

Notice that the drive term $\propto g$ breaks both the system's spatial reflection symmetry as well as the Ising symmetry $S \to -S$. In one dimension, Eq. (48) coincides with the noisy Burgers equation [49], and since in this case (only) the condition (39) is satisfied, effectively represents a system with equilibrium dynamics. The corresponding Janssen–De Dominicis response functional reads

$$\mathcal{A}\left[\widetilde{S}, S\right] = \int \mathrm{d}^d x \int \mathrm{d}t \widetilde{S}\left[\frac{\partial S}{\partial t} - D\left(\nabla_\perp^2 + c\nabla_\parallel^2\right) S + D\left(\nabla_\perp^2 + \tilde{c}\nabla_\parallel^2\right)\widetilde{S} - \frac{Dg}{2}\nabla_\parallel S^2\right]. \tag{50}$$

It describes a 'massless' theory, hence we expect the system to generically display scale-invariant features, without the need to tune to a special point in parameter space. The large-scale scaling properties can be analyzed by means of the dynamic renormalization group [47,48].

Another famous example for generic scale invariance emerging in a nonequilibrium system is curvature-driven interface growth, as captured by the Kardar–Parisi–Zhang equation [50]

$$\frac{\partial S(x,t)}{\partial t} = D\nabla^2 S(x,t) + \frac{Dg}{2}[\nabla S(x,t)]^2 + \zeta(x,t), \tag{51}$$

with again $\langle\zeta\rangle = 0$ and the noise correlations

$$\langle\zeta(x,t)\zeta(x',t')\rangle = 2D\delta(x-x')\delta(t-t'). \tag{52}$$

(For more details and intriguing variants, see e. g. [51, 52,53].) The associated field theory action

$$\mathcal{A}\left[\widetilde{S}, S\right] = \int \mathrm{d}^d x \cdot \int \mathrm{d}t\left[\widetilde{S}\left(\frac{\partial S}{\partial t} - D\nabla^2 S - \frac{Dg}{2}[\nabla S]^2\right) - D\widetilde{S}^2\right] \tag{53}$$

encodes surprisingly rich behavior including a kinetic roughening transition separating two distinct scaling regimes in dimensions $d > 2$ [51,52,53].

Future Directions

The rich phenomenology in many complex systems is only inadequately captured within widely used mean-field approximations, wherein both statistical fluctuations and correlations induced by the subunits' interactions or the system's kinetics are neglected. Modern computational techniques, empowered by recent vast improvements in data storage and tact frequencies, as well as the development of clever algorithms, are clearly invaluable in the theoretical study of model systems displaying the hallmark features of complexity. Yet in order to gain a deeper understanding and to maintain control over the typically rather large parameter space, numerical investigations need to be supplemented by analytical approaches. The field-theoretic methods described in this article represent a powerful set of tools to systematically include fluctuations and correlations in the mathematical description of complex stochastic dynamical systems composed of many interacting degrees of freedom. They have already been very fruitful in studying the intriguing physics of highly correlated and strongly fluctuating many-particle systems. Aside from many important quantitative results, they have provided the basis for our fundamental understanding of the emergence of universal macroscopic features.

At the time of writing, the transfer of field-theoretic methods to problems in chemistry, biology, and other fields such as sociology has certainly been initiated, but is still limited to rather few and isolated case studies. This is understandable, since becoming acquainted with the intricate technicalities of the field theory formalism requires considerable effort. Also, whereas it is straightforward to write down the actions corresponding the stochastic processes defined via microscopic classical discrete master or mesoscopic Langevin equations, it is usually not that easy to properly extract the desired information about large-scale structures and long-time asymptotics. Yet if successful, one tends to gain insights that are not accessible by any other means. I therefore anticipate that the now well-

developed methods of quantum and statistical field theory, with their extensions to stochastic dynamics, will find ample successful applications in many different areas of complexity science. Naturally, further approximation schemes and other methods tailored to the questions at hand will have to be developed, and novel concepts be devised. I look forward to learning about and hopefully also participating in these exciting future developments.

Acknowledgments

The author would like to acknowledge financial support through the US National Science Foundation grant NSF DMR-0308548. This article is dedicated to the victims of the terrible events at Virginia Tech on April 16, 2007.

Bibliography

1. Lindenberg K, Oshanin G, Tachiya M (eds) (2007) J Phys: Condens Matter 19(6): Special issue containing articles on Chemical kinetics beyond the textbook: fluctuations, many-particle effects and anomalous dynamics; see: http://www.iop.org/EJ/toc/0953-8984/19/6
2. Alber M, Frey E, Goldstein R (eds) (2007) J Stat Phys 128(1/2): Special issue on Statistical physics in biology; see: http://springerlink.com/content/j4q1ln243968/
3. Murray JD (2002) Mathematical biology, vols. I, II, 3rd edn. Springer, New York
4. Mobilia M, Georgiev IT, Täuber UC (2007) Phase transitions and spatio-temporal fluctuations in stochastic lattice Lotka–Volterra models. J Stat Phys 128:447–483. several movies with Monte Carlo simulation animations can be accessed at http://www.phys.vt.edu/~tauber/PredatorPrey/movies/
5. Washenberger MJ, Mobilia M, Täuber UC (2007) Influence of local carrying capacity restrictions on stochastic predator-prey models. J Phys: Condens Matter 19:065139, 1–14
6. Ramond P (1981) Field theory – a modern primer. Benjamin/Cummings, Reading
7. Amit DJ (1984) Field theory, the renormalization group, and critical phenomena. World Scientific, Singapore
8. Negele JW, Orland H (1988) Quantum many-particle systems. Addison-Wesley, Redwood City
9. Parisi G (1988) Statistical field theory. Addison-Wesley, Redwood City
10. Itzykson C, Drouffe JM (1989) Statistical field theory. Cambridge University Press, Cambridge
11. Le Bellac M (1991) Quantum and statistical field theory. Oxford University Press, Oxford
12. Zinn-Justin J (1993) Quantum field theory and critical phenomena. Clarendon Press, Oxford
13. Cardy J (1996) Scaling and renormalization in statistical physics. Cambridge University Press, Cambridge
14. Janssen HK (1976) On a Lagrangean for classical field dynamics and renormalization group calculations of dynamical critical properties. Z Phys B 23:377–380
15. De Dominicis C (1976) Techniques de renormalisation de la théorie des champs et dynamique des phénomènes critiques. J Physique (France) Colloq 37:C247–C253
16. Janssen HK (1979) Field-theoretic methods applied to critical dynamics. In: Enz CP (ed) Dynamical critical phenomena and related topics. Lecture Notes in Physics, vol 104. Springer, Heidelberg, pp 26–47
17. Doi M (1976) Second quantization representation for classical many-particle systems. J Phys A: Math Gen 9:1465–1477
18. Doi M (1976) Stochastic theory of diffusion-controlled reactions. J Phys A: Math Gen 9:1479–1495
19. Grassberger P, Scheunert M (1980) Fock-space methods for identical classical objects. Fortschr Phys 28:547–578
20. Peliti L (1985) Path integral approach to birth-death processes on a lattice. J Phys (Paris) 46:1469–1482
21. Peliti L (1986) Renormalisation of fluctuation effects in the $A + A \to A$ reaction. J Phys A: Math Gen 19:L365–L367
22. Lee BP (1994) Renormalization group calculation for the reaction $kA \to \emptyset$. J Phys A: Math Gen 27:2633–2652
23. Lee BP, Cardy J (1995) Renormalization group study of the $A + B \to \emptyset$ diffusion-limited reaction. J Stat Phys 80:971–1007
24. Mattis DC, Glasser ML (1998) The uses of quantum field theory in diffusion-limited reactions. Rev Mod Phys 70:979–1002
25. Täuber UC, Howard MJ, Vollmayr-Lee BP (2005) Applications of field-theoretic renormalization group methods to reaction-diffusion problems. J Phys A: Math Gen 38:R79–R131
26. Täuber UC (2007) Field theory approaches to nonequilibrium dynamics. In: Henkel M, Pleimling M, Sanctuary R (eds) Ageing and the glass transition. Lecture Notes in Physics, vol 716. Springer, Berlin, pp 295–348
27. Täuber UC, Critical dynamics: a field theory approach to equilibrium and nonequilibrium scaling behavior. To be published at Cambridge University Press, Cambridge. for completed chapters, see: http://www.phys.vt.edu/~tauber/utaeuber.html
28. Schütz GM (2000) Exactly solvable models for many-body systems far from equilibrium. In: Domb C, Lebowitz JL (eds) Phase transitions and critical phenomena, vol 19. Academic Press, London
29. Stinchcombe R (2001) Stochastic nonequilibrium systems. Adv Phys 50:431–496
30. Van Wijland F (2001) Field theory for reaction-diffusion processes with hard-core particles. Phys Rev E 63:022101, 1–4
31. Chopard B, Droz M (1998) Cellular automaton modeling of physical systems. Cambridge University Press, Cambridge
32. Marro L, Dickman R (1999) Nonequilibrium phase transitions in lattice models. Cambridge University Press, Cambridge
33. Hinrichsen H (2000) Nonequilibrium critical phenomena and phase transitions into absorbing states. Adv Phys 49:815–958
34. Ódor G (2004) Phase transition universality classes of classical, nonequilibrium systems. Rev Mod Phys 76:663–724
35. Moshe M (1978) Recent developments in Reggeon field theory. Phys Rep 37:255–345
36. Obukhov SP (1980) The problem of directed percolation. Physica A 101:145–155
37. Cardy JL, Sugar RL (1980) Directed percolation and Reggeon field theory. J Phys A: Math Gen 13:L423–L427
38. Janssen HK (1981) On the nonequilibrium phase transition in reaction-diffusion systems with an absorbing stationary state. Z Phys B 42:151–154
39. Janssen HK, Täuber UC (2005) The field theory approach to percolation processes. Ann Phys (NY) 315:147–192
40. Grassberger P (1982) On phase transitions in Schlögl's second model. Z Phys B 47:365–374

41. Janssen HK (2001) Directed percolation with colors and flavors. J Stat Phys 103:801–839
42. Martin PC, Siggia ED, Rose HA (1973) Statistical dynamics of classical systems. Phys Rev A 8:423–437
43. Bausch R, Janssen HK, Wagner H (1976) Renormalized field theory of critical dynamics. Z Phys B 24:113–127
44. Chaikin PM, Lubensky TC (1995) Principles of condensed matter physics. Cambridge University Press, Cambridge
45. Hohenberg PC, Halperin BI (1977) Theory of dynamic critical phenomena. Rev Mod Phys 49:435–479
46. Schmittmann B, Zia RKP (1995) Statistical mechanics of driven diffusive systems. In: Domb C, Lebowitz JL (eds) Phase transitions and critical phenomena, vol 17. Academic Press, London
47. Janssen HK, Schmittmann B (1986) Field theory of long time behaviour in driven diffusive systems. Z Phys B 63:517–520
48. Leung KT, Cardy JL (1986) Field theory of critical behavior in a driven diffusive system. J Stat Phys 44:567–588
49. Forster D, Nelson DR, Stephen MJ (1977) Large-distance and long-time properties of a randomly stirred fluid. Phys Rev A 16:732–749
50. Kardar M, Parisi G, Zhang YC (1986) Dynamic scaling of growing interfaces. Phys Rev Lett 56:889–892
51. Barabási AL, Stanley HE (1995) Fractal concepts in surface growth. Cambridge University Press, Cambridge
52. Halpin-Healy T, Zhang YC (1995) Kinetic roughening phenomena, stochastic growth, directed polymers and all that. Phys Rep 254:215–414
53. Krug J (1997) Origins of scale invariance in growth processes. Adv Phys 46:139–282

Finance, Agent Based Modeling in

Sebastiano Manzan
Department of Economics and Finance, Baruch College CUNY, New York, USA

Article Outline

Glossary
Definition of the Subject
Introduction
The Standard RE Model
Analytical Agent-Based Models
Computational Agent-Based Models
Other Applications in Finance
Future Directions
Bibliography

Glossary

Rational expectations (RE) An assumption often introduced in economic models. It assumes that agents subjective distribution is equal to the true probability distribution of a random variable. The implication is that expectation errors are purely random.

Bounded rationality The assumption that agents have limited ability to acquire and process information and to solve complex economic problems. These limitations imply that expectations can diverge from RE.

Efficient markets hypothesis (EMH) An application of rational expectations to asset prices. The EMH assumes that asset prices reflect all available information. It implies that asset prices behave like a random walk process and their changes are purely random.

Artificial financial markets A market populated by agents that have bounded rational expectations and learning from available information. Trading in these markets occurs based on traditional price setting mechanisms or more realistic mechanisms inspired by electronic markets.

Definition of the Subject

Finance can be broadly defined as studying the allocation of resources over time in an uncertain environment. Consumers are interested in saving part of their current income and transfer it for consumption in the future (e. g., saving for retirement). On the other hand, firms are looking to raise capital to finance productive investments that will payoff in the future. In both decisions, the future is uncertain and individuals and firms are required to evaluate the risks involved in buying an asset (e. g., stocks and bonds) or investing in a project.

The traditional modeling approach in finance is to introduce strong assumptions on the behavior of agents. They are assumed to have perfect knowledge of the structure of the economy and to correctly process the available information. Based on these two assumptions, agents are able to form Rational Expectations (RE) such that their beliefs are not systematically wrong (in other words, the forecasting errors are random). Common sense suggests that these assumptions impose unreasonable requirements on the cognitive and computational abilities of agents. In practice, investors and firms are trying to learn to behave "*rationally*" in an economic system that is continuously evolving and where information is imperfect. In addition, there is an increasing amount of empirical evidence that is not consistent with RE theories.

These limitations have motivated an interest in finance to relax the strong assumptions on agents' behavior. Agent-based modeling contributes to this literature by assuming that consumers and firms have limited computational abilities (also known as bounded rationality) and learning (rather than knowing) the mechanisms governing the economy. These models have two main targets. First, to determine the conditions that lead a population of

bounded-rational interacting agents to produce an aggregate behavior that resembles the one of a RE representative agent model. Second, they aim at explaining the empirical facts and anomalies that the standard approach fails to explain.

This entry is structured as follows. In Sect. "Introduction" we discuss in more detail the application of agent-based modeling in finance. In particular, most of the early literature has focused on one specific aspect of financial economics, asset pricing. Sects. "The Standard RE Model" to "Computational Agent-Based Models" introduce the standard asset pricing model and describe the agent-based approaches that have been proposed in the literature. Sect. "Other Applications in Finance" presents some (more recent) applications of agent-based models to corporate finance and market microstructure and, finally, Sect. "Future Directions" discusses some possible future directions on the application of agent-based models in finance.

Introduction

The goal of asset pricing models is to provide an explanation for the "fair" valuation of a financial asset paying an uncertain cash flow. A key role in asset pricing models is played by agents expectations regarding the future cash flow of the asset. Standard economic models assume that agents have *Rational Expectations* (RE). The RE hypothesis is the outcome of some more basic assumptions on agents behavior: they know and use all the information available, they have unlimited computational ability, and rationality is common knowledge in the population. Common sense and introspection suggest that these are quite strong assumptions if the target is to build a realistic model of agents behavior. A justification for assuming RE in asset pricing models is provided by [37]:

> … this hypothesis (like utility maximization) is not "behavioral": it does not describe the way agents think about their environment, how they learn, process information, and so forth. It is rather a property likely to be (approximately) possessed by the outcome of this unspecified process of learning and adapting.

Agent-based models try to address the issues left unspecified by the RE proponents: how do agents learn and process the information available? In other words, how do they form expectations? In fact, in the intent of the RE proponents, rationality is simply a property of the outcome (e. g., asset price) rather than an assumption about the subjective expectation formation process.

The innovative aspect of the agent-based approach is that it explicitly models "*this unspecified process of learning and adaptation*" (in Lucas's words). The common elements of the wide range of agent-based asset pricing models are:

Expectations agents hold subjective expectations that are *bounded rational*, that is, they are based on processing the available (and possibly imperfect and costly) information and that evolve over time. Agent-based models explicitly specify the way individuals form their expectation, instead of leaving it totally unspecified as in the RE approach.

Heterogeneity agents have different subjective expectations about the future due to heterogeneity in the way they process or interpret information. The RE setup suppresses agents heterogeneity: given the same information set, there is only one way to be rational and agents are thus homogeneous.

Evolution agents evolve in the sense that they abandon a belief if it performs poorly. Instead, rational models typically rely on the latent assumption that non-rational agents will not survive a (unspecified) process of evolutionary market competition.

Based on these basic ingredients, the agent based literature has now grown in different directions and we can distinguish two clearly defined approaches to agent-based modeling. The main difference between them is how they combine the different characteristics discussed above:

Analytical models these models assume that there are many expectation types and agents switch between them according to a deterministic or stochastic process. In the deterministic case, evolution is based on the past performance of the beliefs: agents discard belief types that perform badly compared to the other available. Instead, models with stochastic switching assume that a process governs the imitation and mutation of types, with possible additional features of herding. These models are simple and analytically tractable.

Computational models agents beliefs can change (or mutate) over time, due to the evolutionary selection of the best performing beliefs. Contrary to the analytical approach, the computational models renounce to analytical tractability in order to investigate more realistic expectation formation processes. Most of these models adopt fitness criteria (e. g., a Genetic Algorithm) to model the evolution of expectations.

The first aim of the agent-based literature is to understand whether introducing less restrictive assumptions on

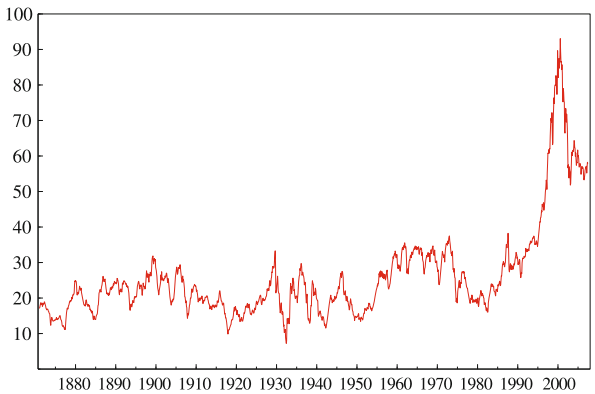

Finance, Agent Based Modeling in, Figure 1
Monthly Price-to-Dividend Ratio for the S&P500 Index from 1871 to 2006

agents behavior (bounded rationality, heterogeneity, and evolutionary selection of expectations) is consistent with the economy converging to the RE equilibrium. If this is the case, it can be argued that relaxing the homogeneity and rationality of agents represents a feasible way to describe the way individuals learn and adapt to achieve an outcome consistent with RE. The second aim of this literature is to provide an explanation for the empirical behavior of asset prices. To illustrate the main stylized facts of financial returns, we consider the Standard & Poors 500 (S&P500) Composite Index (a U.S. equity index). Figure 1 shows the Price-to-Dividend (PD) ratio from 1871 until 2006 at the monthly frequency. It is clear that the PD ratio fluctuates significantly with some periods of extreme valuations, as in the late 1990s. The debate on the reasons for these fluctuations has not reached (yet) a widely accepted conclusion. On the one hand, there are RE models that explain the variation in the PD ratio by changes in the risk premium, i. e., the ex-ante rate of return required by agents to invest in the risky asset. Instead, other models attribute these swings to irrational expectations of investors, that are prone to optimism (and overvaluation) when asset prices are increasing. The two explanations are not mutually excluding since both factors might contribute to explain the observed fluctuations of the PD ratio.

Figure 2 considers the S&P500 Index from 1977 until 2007 at the daily frequency. The figure shows the returns (defined as the percentage change of the price of a financial asset) and the absolute value of the returns. Figure 3 describes the statistical properties of returns, such as the histogram and the autocorrelation function of the returns and absolute returns. The main stylized facts of daily returns are:

Volatility clustering returns alternate periods of high and low volatility (or variability). In calm periods, returns oscillate within a small range, while in turbulent periods they display a much wider range of variation. This is a feature common to different asset classes (e. g., equities, exchange rates, and bonds). The time series of returns and absolute returns on the S&P500 in Fig. 2 clearly show this pattern. In certain periods returns vary in a narrow interval between $\pm 1\%$, while in other periods their variability is higher (e. g., between ± 3 and 5%).

Leptokurtic distribution the distribution of asset returns has a sharp peak around the mean and fat tails (compared to the normal distribution). Large events (positive and negative) are more likely to occur compared to what is expected under the assumption of normality. This property emerges clearly from the top plot of Fig. 3 that shows the histogram of the S&P500 returns and the normal distribution (based on the estimated mean and variance).

No serial correlation returns do not display significant linear serial correlation. The autocorrelation function of the returns (mid-plot of Fig. 3) is close to 0 at all lags considered.

Persistence in volatility on the other hand, volatility (measured by absolute or square returns) has significant linear dependence. The autocorrelation of the absolute returns in Fig. 3 is about 0.1 (and still significant) at lag 100.

Another relevant fact that is short of explanations is the large trading volume that occurs in financial markets. A model in which everybody is rational and knows that everybody else is rational cannot account for the existence of such relevant volume of trade. Agent-based models aim at explaining this phenomenon based on the assumption that agents hold heterogeneous expectations. Volume can arise, for example, if an optimistic agent is willing to buy an asset from a pessimistic agent (that is willing to sell). An interesting feature of trading volume is its asymmetry during markets cycles: it is typically high when financial markets are booming, and low when the prices are decreasing. There is also empirical evidence that trading volume and volatility are correlated, suggesting that the same economic mechanism might be able to explain both phenomena.

Summarizing, the aim of the agent-based approach to asset pricing is to introduce more realistic assumptions on the way agents form expectations, learn from new information, and adapt to a changing environment. The re-

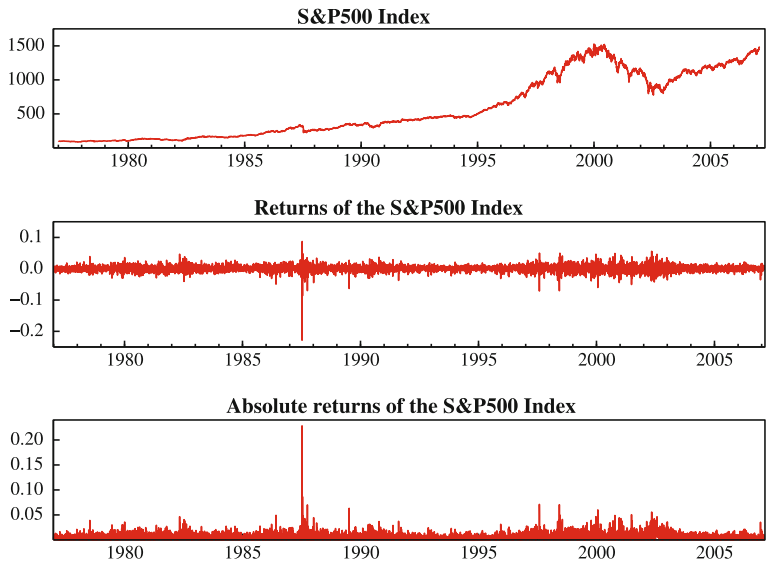

Finance, Agent Based Modeling in, Figure 2
Daily observations of the S&P500 Index from 1977 to 2007. (*top*) Time series of the Index, (*middle*) the returns, (*bottom*) the absolute value of returns

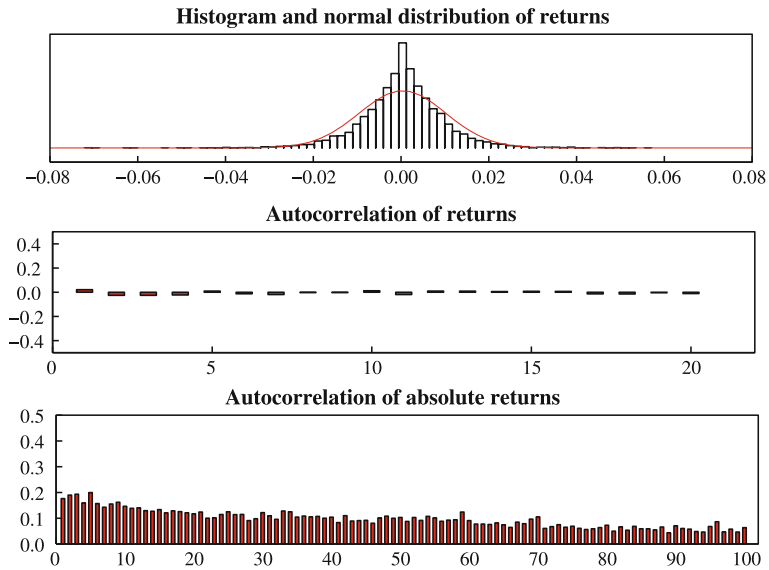

Finance, Agent Based Modeling in, Figure 3
Statistical properties of the S&P500 returns: (*top*) histogram and normal distribution, (*middle*) autocorrelation function (max lag 20) for the returns, (*bottom*) autocorrelation function (max lag 100) for the absolute returns

search questions the agent-based approach is trying to answer are:

1. Under what conditions are these models able to reproduce the RE equilibrium (although starting from a more general setup where agents are not – a priori – assumed to have RE)?
2. Another issue is the empirical validity of these models: are they able to explain the empirical features of financial returns that standard RE models fail to account for?

In the following Sections, we describe some of the most well-known examples of agent-based models in finance, both in the *analytical* and *computational* group. However,

we first introduce a basic RE model that is the starting point for most of the literature. In Sect. "Other Applications in Finance" we discuss other interesting applications of agent-based models in finance.

The Standard RE Model

We briefly consider the standard asset pricing model that is used as a benchmark in the agent-based literature. A more detailed discussion can be found in [25]. The model assumes that agent i faces the choice of investing her wealth among two assets: a riskless asset that pays a constant return r, and a risky asset that pays a stochastic dividend in period t denoted by D_t. A typical assumption is that agents have Constant Absolute Risk Aversion (CARA) preferences defined as $U(W_i) = -e^{-\lambda W_i}$, where $U(\cdot)$ indicates the utility function, W_i denotes the wealth of agent i and λ is the coefficient of absolute risk aversion. These preferences imply the following demand of shares of the risky asset, $X_{i,t}$:

$$X_{i,t} = \frac{E_{i,t}(P_{t+1} + D_{t+1}) - (1 + r)P_t}{\lambda \sigma^2_{i,t}(P_{t+1} + D_{t+1})} , \tag{1}$$

where P_t is the price of the risky asset in period t, $E_{i,t}(\cdot)$ is the conditional expectation of agent i about next-period payoff of the risky investment, and $\sigma^2_{i,t}(\cdot)$ is the conditional variance of the payoff for agent i. Agents buy shares of the risky asset ($X_{i,t} > 0$) if they expect the return of a share to be higher compared to investing the same amount (P_t) in the riskless asset.

The equilibrium price of the risky asset is such that the aggregate demand and supply are equal. Assuming that there are S number of shares of the risky asset available, the equilibrium condition is

$$S = \sum_i X_{i,t} . \tag{2}$$

The aggregation across different individuals is simplified by assuming a representative agent with expectation $E_t(P_{t+1} + D_{t+1})$ (and similarly for the conditional variance) for all i's in the economy. This is equivalent to assume that agents are homogeneous in their expectation about the future payoff of the risky asset. In addition, assuming that the representative agent holds RE, it can be shown that the equilibrium price of the risky asset is a linear function of D_t given by

$$P_t = a + bD_t ,$$

where a and b are constant (and related to the structural parameters of the model).

There is an extensive literature that aims at relaxing the strong restrictions imposed by the RE hypothesis. Models of *rational learning* assume that agents (typically in a representative agent setup) have to learn (rather than know) the structure of the economy, e. g., the parameters governing the cash flow process. In this case, agents are rational in the sense that they process optimally the information available. However, they do not hold rational expectations since they have imperfect knowledge of the structure of the economy. An alternative route followed in the literature is to consider the effect of *behavioral* biases in the expectation formation process. A comparison of the vast literature on rational learning and behavioral models is provided by [6].

Agent-based models build on these extensions of the basic asset pricing model by considering both rational learning and irrational expectations in a richer economic structure where agents hold heterogeneous expectations. We will now discuss some of the most well-known models in the analytical and computational agent-based literature and deal with their main differences.

Analytical Agent-Based Models

The analytical models assume that the population of agents can choose among a small number of beliefs (or predictors) about next period payoff of the risky asset. Heterogeneity is introduced by allowing agents with different predictors to co-exist, and learning might occur if they are allowed to switch between different beliefs in an evolutionary way.

These models can be described as follows. Assume there are a set of H belief types publicly available to agents. Denote the belief of type h (for $h = 1, \ldots, H$) about next period payoff by $E_{h,t}(P_{t+1} + D_{t+1})$ and the conditional variance by $\sigma^2_{h,t}(P_{t+1} + D_{t+1})$. Since these models depart from the assumption of RE, they typically introduce a behavioral assumption that the beliefs are either of the *fundamentalist* or the *trend-following* type. [20] and [21] conducted survey studies of exchange rate traders and found that their expectations could be represented as trend-following in the short-run, but fundamentalist in the long run. Fundamentalist expectations are characterized by the belief that the market price is anchored to the asset fundamental valuation and deviations (of the price from the fundamental) are expected to disappear over time. In this case, the belief uses both information about the asset price and the dividend process (that drives the fundamental value) to form an expectation about the future. On the other hand, trend-following expectations exploit only information contained in the price series to extrapolate the future

dynamics. These types of beliefs are obviously not consistent with the RE equilibrium although they are supported by empirical evidence of their widespread use in financial markets.

Another key assumption of agent-based models concerns the evolution of beliefs: agents switch between expectations based on their past performance or because of interaction with other agents in the population. It is possible to distinguish two families of models with different evolutionary dynamics:

Deterministic evolution agents switch between the different beliefs based on a deterministic function. Typically, the switching is determined by past forecast accuracy of the predictors or their realized profits.

Stochastic evolution a stochastic process governs the switching of agents between beliefs.

Deterministic Evolution

An example of an agent-based model with deterministic evolution is proposed by [7]. A simple version of their model assumes that there are only two types of beliefs: fundamentalists and trend-followers. Some simplifying assumptions are used in deriving the equilibrium price: the dividend process D_t is assumed to be $i.i.d$ (with mean $\bar{D}$) and agents have homogeneous expectations about the dividend process. In this case, the expectation about next period payoff $E_{h,t}(P_{t+1} + D_{t+1})$ in Eq. (1) becomes $E_{h,t}(P_{t+1}) + \bar{D}$.

Lets denote by $P^*(= \bar{D}/r)$ the constant RE fundamental price. The fundamentalist type has the following belief:

$$E_{\mathrm{F},t}(P_{t+1}) = P^* + g_{\mathrm{F}}(P_{t-1} - P^*) \,. \tag{3}$$

When $0 < g_{\mathrm{F}} < 1$, fundamentalists believe the asset price will revert toward its fundamental value, and g_{F} can be interpreted as the speed at which this adjustment is expected to occur. This model assumes that when agents form their belief at time t they actually do not observe the realized asset price for period t. This explains the fact that the expectation is a function of the last observed price, P_{t-1}.

Brock and Hommes assume that agents pay a cost C to acquire the fundamentalist predictor. The underlying idea is to let them choose whether to buy a "sophisticated" predictor (that requires calculating the fundamental value) or, alternatively, to extrapolate from past realized prices. The belief of the trend-followers is given by:

$$E_{\mathrm{TF},t}(P_{t+1}) = g_{\mathrm{TF}} P_{t-1} \,. \tag{4}$$

The value of the parameter g_{TF} determines the strength of extrapolation of the trend-followers. If $g_{\mathrm{TF}} > 1$, they expect an upward trend in prices and, vice versa, for $0 < g_{\mathrm{TF}} < 1$.

Assuming the supply of the risky asset, S, in Eq. (2) is equal to 0, the equilibrium asset price, P_t, is given by:

$$P_t = \left(\frac{n_{\mathrm{F},t}(1 - g_{\mathrm{F}}) - r}{1 + r}\right) P^* + \left(\frac{n_{\mathrm{F},t}(g_{\mathrm{F}} - g_{\mathrm{TF}}) + g_{\mathrm{TF}}}{1 + r}\right) P_{t-1} \,, \tag{5}$$

where $n_{\mathrm{F},t}$ indicates the fraction of agents in the population using the fundamentalist belief and the remaining $n_{\mathrm{TF},t}(= 1 - n_{\mathrm{F},t})$ using the trend-following one. [7] assumes the evolution of the fractions $n_{\mathrm{F},t}$ is governed by a discrete choice probability model:

$$n_{\mathrm{F},t} = \frac{1}{1 - \exp\left[\beta(U_{\mathrm{TF},t-1} - U_{\mathrm{F},t-1})\right]} \,, \tag{6}$$

where $U_{h,t-1}(h = F, TF)$ is a measure of the fitness of belief h defined as:

$$U_{\mathrm{F},t-1} = \pi_{\mathrm{F},t-1} + \eta U_{\mathrm{F},t-2} - C, \quad \text{and}$$
$$U_{\mathrm{TF},t-1} = \pi_{\mathrm{TF},t-1} + \eta U_{\mathrm{TF},t-2} \,,$$

where $\pi_{h,t-1}$ measures the fitness performance (measured by realized profits or forecast accuracy) of the belief h at time $t-1$ and η is a parameter that determines the memory in the performance measure. C in $U_{\mathrm{F},t-1}$ represents the cost that agents face if they adopt the fundamentalist belief (while the trend-following is available at no cost). The fraction in Eq. (6) depends on the parameter $\beta(> 0)$ that determines the speed at which agents respond to differentials of performance among beliefs. If β is small, agents are very reluctant to switch and require a significantly large difference in fitness to adopt another predictor. On the other hand, when β is large, even small differences of performance cause dramatic changes in the fractions. For a given value of β, if the fundamentalist belief significantly outperforms the trend-following (that is, $U_{\mathrm{F},t-1} \gg U_{\mathrm{TF},t-1}$), then the fraction $n_{\mathrm{F},t-1}$ tends to 1, meaning that most agents in the economy switch to the fundamentalist expectation.

The interesting feature of this model is that it can converge to the RE equilibrium or generate complicated dynamics depending on the value of the parameters. For some combinations of the g_{F} and g_{TF} parameters, the system converges to the RE equilibrium (i. e., the deviation is equal to 0). However, trend-followers can destabilize the economy when their extrapolation rate, g_{TF} is high enough. For small values of β the dynamical system converges to the RE equilibrium. However, for increasing values of β the system experiences a transition toward a non-

fundamental steady state and complicated dynamics (limit cycles and strange attractors) emerge.

In the presence of information cost (to buy the fundamentalist predictor) and evolutionary switching between strategies, the economy might still converge to the RE equilibrium for a large set of parameter values. However, it is also able to generate large fluctuations of the asset price around the fundamental value. Figure 4 shows a time series of the asset price P_t and the fraction of fundamentalists described in Eqs. (5) and (6). The constant fundamental value in this Figure is equal to 25. As it is clear from the picture, the asset price experiences large swings away from the fundamentals that are explained by the increased importance of agents using the trend-following belief. When the mispricing becomes too large, the economy experiences a sudden change of sentiment with most agents switching to the fundamentalist belief. In this sense, the model is more appropriate to explain the boom-bust dynamics of financial markets.

Although the purely deterministic model captures the relevant features of the dynamics of financial markets, adding a stochastic component provides simulated series that better resemble the observed ones (such as Fig. 1). The model can be extended by considering an approximation error term in Eq. (5) that interacts with the dynamics of the model. Figure 5 shows the asset price and the fraction of fundamentalists for a normally distributed error term with standard deviation equal to 2. The asset price shows large and persistent deviations from the fundamental value ($P^* = 25$), in some periods as extreme as reaching 100 while, in other periods, more moderate. Since the dividend process is assumed to be $i.i.d.$, the price can also be interpreted as a valuation (PD) ratio. Comparing the properties of this time series with the one for the S&P500 in Fig. 1, it seems that it is able to capture its main qualitative features. [5] provide empirical evidence of the ability of a similar model to explain the long-run behavior of stock prices.

The model proposed by Brock and Hommes is an example of a literature interested in the (possible) emergence of complicated dynamics in asset pricing models. An early contribution to the deterministic literature is [15]. They assume that there are two types of investors: some extrapolating from past deviations while the other group is more sophisticated and able to evaluate whether the asset is over- or under-valued (and sells or buys more aggressively if the mispricing is large). A third actor in the model is the market maker. The role of the market maker is to aggregate demand and supply and to fix the price of the asset. This mechanism is different from the assumption in Eq. (2). In that case, agents submit their demand function (quantity as a function of price) and the price is set at the value that clears the market. Instead, the market maker receives orders from the agents and moves the price to offset excess demand or supply. This represents a disequilibrium mechanism since market makers use their in-

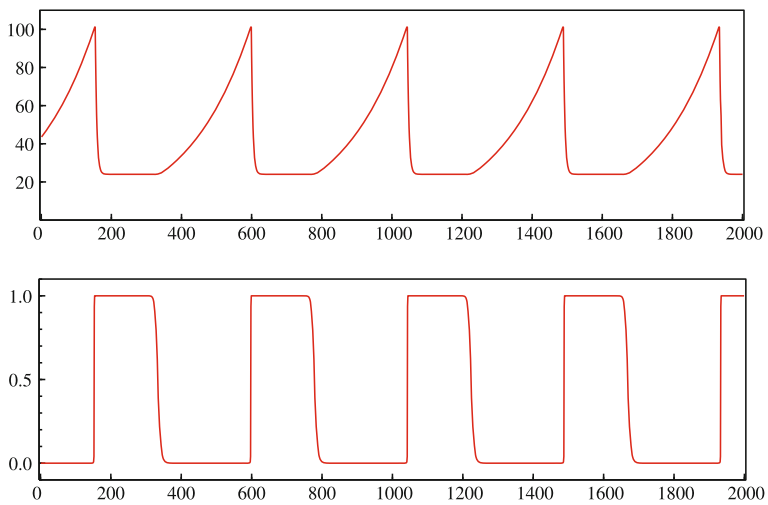

Finance, Agent Based Modeling in, Figure 4

Brock and Hommes model with 2 belief types, fundamentalists and trend-followers. The top plot represents a time series of the asset price and the bottom plot depicts the fraction of fundamentalists, $n_{F,t}$. The parameters of the model: intensity of choice $\beta = 0.5$, the interest rate $r = 0.0083$, the parameter of the fundamentalists $g_F = 0.8$, the parameter of the trend-followers belief $g = 1.014$, the cost of the fundamentalist predictor $C = 0.5$, memory parameter $\eta = 0.99$

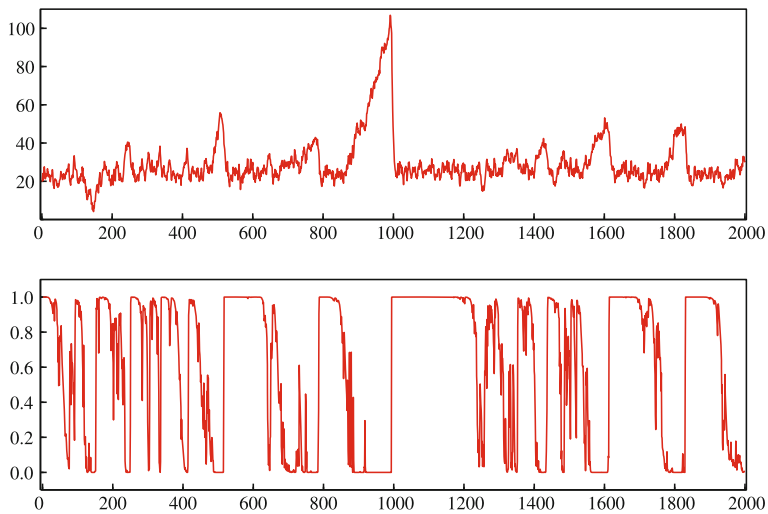

Finance, Agent Based Modeling in, Figure 5
Same model and parameter values used in Fig. 4 with an error term added to Eq. (5) that is normally distributed with mean zero and standard deviation equal to 2

ventory of stocks to provide liquidity in case of excess demand and accumulate stocks in case of excess supply. The results for this model are similar to what was discussed above. The RE equilibrium is obtained when the sophisticated agents dominate the market. However, limit cycles and chaos arise when the trend-following agents are relatively important and the economy fluctuates between periods of increasing asset prices and sudden crashes. Another model that assumes the market maker mechanism is proposed by [9]. In this case, agents form their expectations based on either the fundamental or extrapolative approach. However, the excess demand function of the chartist is assumed to be nonlinear. When the extrapolation rate of the chartist is sufficiently high, the system becomes unstable and limit cycles arise. While these early models assumed that the fractions of agents are fixed, [16] and [17] introduced, in a similar setup, time-variation in those fractions. The driving force for the variation of the fractions is the relative performance of the beliefs (similar to what we discussed above for the model of Brock and Hommes). Some of the more recent models that extend these early contributions are [10,11,12,18,19,24,50], and [51]. A comprehensive survey of the literature is provided in [26].

Stochastic Evolution

An early example of an agent-based model in which individuals switch in a stochastic fashion was proposed by [27]. He uses a slightly different setup compared to the Standard RE Model. In his model the asset is a foreign bond and the agent has to decide whether to invest at home (at the riskless interest rate r) or buy a unit of foreign currency and invest abroad (at the risky interest rate ρ_t, assumed to be normally distributed with mean ρ and variance σ_ρ^2). The price P_{t+1} represents the exchange rate. The only difference with the model described earlier is that in the demand of type h agent in Eq. (1), $E_{h,t}(P_{t+1} + D_{t+1})$ is replaced by $(1+\rho)E_{h,t}(P_{t+1})$. The fundamental value of the asset in this model is assumed to evolve as a random walk, that is, $P_t^* = P_{t-1}^* + \epsilon_t$ where $\epsilon_t \sim N(0, \sigma_\epsilon^2)$.

Similarly to the previous model, there are two types of beliefs: fundamentalists and chartists. The fundamentalist belief is the same as in Eq. (3), while the chartists have belief given by:

$$E_{\mathrm{TF},t}(P_{t+1}) = (1 - g_{\mathrm{TF}})P_t + g_{\mathrm{TF}}P_{t-1} \,.$$

The switching between beliefs in Kirman's model is driven by two mechanism: social interactions and herding. Interaction means that agents meet in pairs and communicate about their beliefs. The result of this communication is that, with a probability $(1-\delta)$, an agent changes her belief to the one of the other agent. In this model, market information (such as prices or dividends) do not play any role in the decision of the agents to adopt the fundamentalist or trend-following beliefs. This is in sharp contrast to the model of [7] where the selection of the belief is endogenous and based on their past performance. In addition to the probability of switching belief due to social interaction, there is a probability ϵ that an agent indepen-

dently changes belief. If we denote by $N_{\mathrm{F},t}$ the number of agents in the population (N is the total number of agents) using the fundamentalist belief at time t, Kirman models the evolution from $N_{\mathrm{F},t-1}$ to $N_{\mathrm{F},t}$ according to a markov chain with the following transition probabilities:

$$\begin{aligned} P(N_{\mathrm{F},t} - N_{\mathrm{F},t-1} = 1) &= \left(1 - \frac{N_{\mathrm{F},t-1}}{N}\right) \\ &\quad \left(\epsilon + (1-\delta)\frac{N_{\mathrm{F},t-1}}{N-1}\right) \\ P(N_{\mathrm{F},t} - N_{\mathrm{F},t-1} = -1) &= \frac{N_{\mathrm{F},t-1}}{N} \\ &\quad \left(\epsilon + (1-\delta)\frac{N - N_{\mathrm{F},t-1}}{N-1}\right) \\ P(N_{\mathrm{F},t} - N_{\mathrm{F},t-1} = 0) &= 1 - P(N_{\mathrm{F},t} - N_{\mathrm{F},t-1} = 1) \\ &\quad - P(N_{\mathrm{F},t} - N_{\mathrm{F},t-1} = -1)\,. \end{aligned}$$

The second part of the opinion formation can be characterized as herding. Kirman assumes that the agents receive a noisy signal, $q_{i,t}$, about the fraction of the population that is fundamentalist:

$$q_{i,t} = \frac{N_{\mathrm{F},t}}{N} + \xi_{i,t}\,,$$

where $\xi_{i,t}$ is distributed as $N(0, \sigma_\xi^2)$ and $i = 1, \ldots, N$. Based on this signal about the average opinion in the economy, agents herd by coordinating in adopting the belief that is more popular. Agent i uses the fundamentalist belief if her signal, $q_{i,t}$, is larger than 0.5 and the trend-following belief otherwise. The fraction of agents using the fundamentalist belief (denoted by $n_{\mathrm{F},t}$) is then given by:

$$n_{\mathrm{F},t} = \frac{1}{N}\sum_{i=1}^{N} I(q_{i,t} \geq 0.5)\,.$$

Given the beliefs of the two types, the fractions and assuming that the supply of foreign currency is proportional to the time varying fundamental value, the equilibrium price of the model is given by:

$$P_t = \frac{n_{\mathrm{F},t} - \gamma}{A} P_t^* - \frac{n_{\mathrm{F},t} g_{\mathrm{F}}}{A} P_{t-1}^* + \frac{(1 - n_{\mathrm{F},t}) g_{\mathrm{TF}}}{A} P_{t-1}\,, \quad (7)$$

where the constants γ and A are functions of the structural parameters of the model.

Figure 6 shows a time series simulated from Kirman's model. The fraction of agents using the fundamentalist belief, $n_{\mathrm{F},t}$, displays significant variation over time, with some periods being close to 1 (most agents are fundamentalists) and other periods close to zero (trend-followers dominate). The resulting price dynamics can be characterized as follows. When fundamentalists dominate the market, the asset price tracks closely the fundamental value and returns volatility is high. On the other hand, when

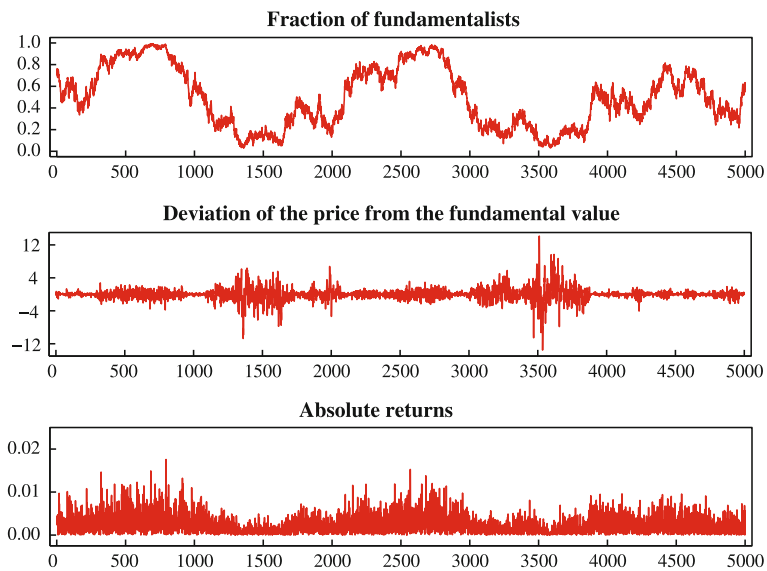

Finance, Agent Based Modeling in, Figure 6

Time series generated from the model proposed by [27] for the following parameter values: $N = 1000$, variance of the $\sigma_\epsilon^2 = 10$, $\rho = 0.00018538$, $r = 0.000133668$, $g_{\mathrm{F}} = 0.6$, $g_{\mathrm{TF}} = 0.475$, $\delta = 0.10$, $\epsilon = 0.000325$, and $\sigma_\xi^2 = 0.43/N$. The top plot shows the fraction of fundamentalist agents in the population, the middle plot the deviation of the market price from the fundamental value, and the bottom plot displays the absolute value of the asset returns

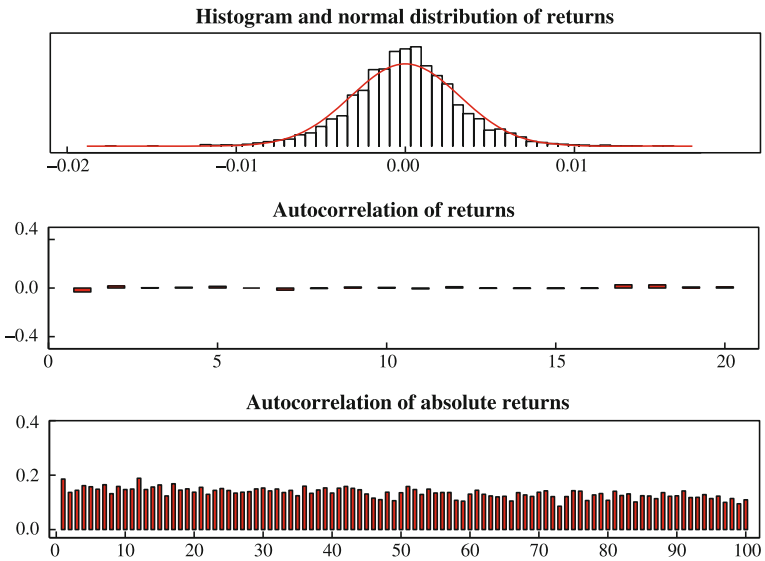

Finance, Agent Based Modeling in, Figure 7

Statistical properties of the time series in Fig. 6. The top plot shows histogram of the series and the parametric distribution under the assumption of normality, and the bottom plots show the autocorrelation of the returns and absolute returns, respectively

trend-followers dominate the market the price tends to deviate significantly from the fundamental and volatility is lower. The time series provide a clear intuition about the ability of the model to account for periods of large deviation from the fundamentals and of persistent volatility.

A main objective of this model is to provide an explanation for the stylized facts of financial returns that were discussed earlier. Figure 7 shows some of the statistical properties for the simulated series. The histogram shows the typical leptokurtic property of financial returns. The distribution of the simulated returns shows a higher concentration of probability mass around zero and in the tails (compared to the normal distribution). The returns autocorrelations are very close to zero and statistically insignificant. However, the absolute returns show significantly positive and slowly-decaying autocorrelations. Hence, the simulated series from Kirman's model are able to replicate the relevant empirical features of daily financial returns.

[41] and [42] propose a model inspired by the opinion formation mechanism of Kirman. The model assumes that agents are either fundamentalists or chartists. In addition, the chartist group is composed of agents that are either optimistic or pessimistic. Agents can switch between the two sub-groups due to herding (following the majority opinion) and also to incorporate the recent trend in asset prices. Instead, the switching between fundamentalist and chartist beliefs is based on the excess profits of the rules. In this aspect, the model allows for a feedback effect from market price to the fractions similarly to [7]. A market maker mechanism aggregates the demand for the risky asset of the agents and determines the prices. Another model based on interacting agents and herding behavior is proposed by [14]. They model the communication among (groups of) agents as a random graph and the returns dynamics closely match the stylized facts of financial returns. [26] and [49] provide extensive overviews of models based on random communication in a population of interacting agents.

Computational Agent-Based Models

Computational agent-based models move one step further compared to analytical models. The setup is very similar: the simple asset pricing model described above, the assumption of heterogeneity of beliefs in the population, and evolutionary pressure to use the best performing predictors. However, computational models do not assume a priori the form of agents' beliefs. Instead, they let agents learn, adapt and explore a large set of strategies and use the best performing ones. In this sense these models allow to investigate the *emergence* of trading strategies and their survival in the market. A relevant question, and an unsettled dispute between academics and practitioners, is the role and importance of technical analysis. Computational agent-based models do not assume (a priori) that trend-following rules are used by agents (as in the analyti-

cal approach), but allow for these rules to emerge from the evolutionary and learning processes. Hence, they can indicate the conditions that lead to the emerge and survival of trend-following rules in the market.

One of the first and most famous example of a computational agent-based model is the Santa Fe Institute (SFI) artificial market proposed by [3]. As mentioned above, the key feature of this and similar models is the way the expectation formation process is represented.

Each agent in the economy is endowed with a set of predictors, in the form of *condition/forecast* rules. These rules are a form of classifier system that identify a state of the world and indicate an action (in this case a forecast of future returns). Each agent in the economy is assumed to rely on a set J of market predictors (classifier rules) that consist of two elements:

Condition a set of bit-strings that characterize different possible states of the market. Each bit represents a state of the world, and the design of the SFI market allows for a set of bits related to *fundamentals* (that relate the asset price to the underlying dividend process) and another set of *technical* bits (that relate the current price to a moving-average of past prices of different length). The role of the bit-strings is to provide the agent with the ability to identify the current state of the market.

Forecast associated with each bit-string j (for $j = 1, \ldots, J$) is a parameter vector (a_j, b_j) that together with the linear forecasting rule $E^j_{i,t}(P_{t+1}+D_{t+1}) = a_j(P_t+D_t)+b_j$ provides agent i with the forecast for next period payoff. The agent then combines the forecast of the H most accurate predictors that are active, based on the observed market condition.

The next building block of the SFI artificial market is the learning process. This is implemented using a Genetic Algorithm (GA) that enables learning in both the condition and the forecast part of the classifier. Agents have a probability p to update their predictors through learning in every period. The frequency of learning (measured by p) plays a relevant role in the resulting dynamics of the model since it identifies how quickly agents adapt to changes in the environment and respond to it. In the learning phase, 15% of the worst performing predictors are dropped, and new rules are generated using a GA with uniform crossover and mutation.

The aim of the model is to test the hypotheses discussed above:

1. Does the SFI market converge to the RE equilibrium?
2. Can the SFI market explain the stylized facts of financial returns?

It turns out that the answer is positive to both questions, depending on the speed of learning parameter p. This parameter plays a crucial role in the resulting dynamics of the SFI market and two regimes can be identified:

Slow-learning in this case the agents are engaged in learning (via the GA) every 1000 periods (on average). The resulting price dynamics shows convergence to the RE equilibrium characterized by agents having homogeneous expectations, negligible trading volume (although some occurs when agents change their beliefs due to learning), and returns that are normal and homoskedastic. What is remarkable is that the convergence to the RE equilibrium is not built-in the model, but it is achieved by the learning and evolutionary process taking place in the SFI market. Another interesting result is that the technical trading bits of the classifier play no role and are never active.

Fast-learning in this experiment the agents update their predictors via learning (on average) every 250 periods. The price dynamics shows the typical features of financial time series, such as alternating periods of high and low volatility, fat tailed distribution, high trading volume, and bubbles and crashes. An interesting result of the fast-learning experiment is the *emergence* of the trend-following rules. The technical trading bits of the predictors are activated and their effect on the asset price spurs even more agents to activate them. In this sense, trend-following beliefs emerge endogenously in the economy and they are not eliminated in the evolution of the system, but survive. This is a quite relevant result also from an empirical point of view. As we mentioned earlier, technical analysis is widely used by practitioners and the SFI market provides an explanation for its emergence (and survival).

[29,30], and [31] have recently proposed an artificial market that is largely inspired by the SFI market. However, LeBaron changed some very relevant assumptions compared to the earlier model. An innovation in this new artificial market is the assumption on the preferences of agents. While the SFI market (and many analytical models) rely on CARA preferences, [29] considers Constant Relative Risk Aversion (CRRA) preferences. In this case, wealth plays a role in the demand of agents (while with CARA does not) and, consequently, allows for differential market impact of agents based on their wealth. Another innovation concerns the expectation formation process. [29] departs from the SFI market assumption of different "*speed of learning*" across agents. Instead, LeBaron assumes that agents have different memory length in eval-

uating strategies. In every period agents assess the profitability of the strategy adopted. However, agents evaluate their strategy using different backtesting periods: some agents test their strategies on only the last few months, while other agents consider the last 20 years. In this sense, they are heterogeneous in their *memory* rather than in the speed of learning. Another feature of this model is that the classifier predictor is replaced by a neural network. The learning and evolution is always characterized by a GA. Despite these innovations, the earlier results of the SFI market are confirmed: a market populated by long-memory agents converges to the RE equilibrium. However, in an economy with agents holding different memory lengths, the asset price series shows the typical features of financial returns (no serial correlation, volatility clustering, fat-tailed distribution, high trading volume, and correlation between volume and volatility).

Another recent artificial stock market model is proposed by [8]. The setup is the simple asset pricing model described in Sect. "The Standard RE Model". Chen and Yeh assume that the expectation of agent i about next period payoff is of the form $E_{i,t}(P_{t+1}+D_{t+1}) = (P_t+\mu)(1+\theta_1 \tanh(\theta_2 f_{i,t}))$. The quantity $f_{i,t}$ characterizes the expectation of the agent and it evolves according to genetic programming. If $f_{i,t}$ is equal to zero the agent believes in the efficiency and rationality of the market, that is, expects the asset price tomorrow to increase by the expected dividend growth rate.

Compared to the SFI market, they adopt a Genetic Programming (GP) approach to model agents' learning and evolution. The model assumes that agents evolve due to two types of pressures: peer-pressure (the agent performance compared to the rest of the population) and self-pressure (own evaluation of the performance). The probability that an agent searches for better forecasting rules depends on both forms of pressure. If agents rank low in terms of performance compared to their peers, then the probability that they will search for other forecasting rules is higher. The population of forecasting rules evolves due to competition with new rules that are generated by applying genetic operators (reproduction, cross-over, mutation) to the existing rules. The rules space evolves independently of the rules adopted by the agents. When an agent decides to search (due to the pressures mentioned above), forecasting rules are randomly selected from the population until a rule is found that outperforms the one currently adopted by the agent. Chen and Yeh show that the price dynamics of the model is consistent with an efficient market. The investigation of the statistical properties of the returns generated by the model shows that the series does not have any linear and nonlinear dependence, although there is some evidence for volatility clustering. Analyzing the type of rules that agents use, they show that only a small fraction of them are actually using forecasting rules that are consistent with an efficient market (in the sense that they believe that $E_{i,t}(P_{t+1} + D_{t+1}) = P_t + \mu$ in which case $f_{i,t}$ is equal to 0). In other words, although a large majority of agents uses rules that imply some form of predictability in asset returns, the aggregation of their beliefs delivers an asset price that looks "*unpredictable*". In this sense they claim that the efficiency (or unpredictability) of the artificial market is an emerging property that results from the aggregation of heterogeneous beliefs in the economy. Another property that emerges from the analysis of this market is the rationality of a representative agent. Chen and Yeh consider the expectation of a representative agent by averaging the expectations across the agents in the economy. The forecasting errors of this "*representative*" expectation indicate that they satisfy a test for rationality: there is no evidence of systematic forecasting errors in the expectation (in statistical terms, the errors are independent).

Evolution and learning (via GA) have received quite a lot of attention in the literature. Other artificial-market models have been proposed in the literature. Some early examples are [4], and [46]. Some more recent examples are [1,2,47,48]. [32] is an extensive survey of the computational agent-based modeling approach.

Other Applications in Finance

The common theme across the models presented above is to show that departing from a representative rational agent is a viable way to explain the empirical behavior of asset prices. The more recent agent-based literature has shifted interest toward nesting this type of models in more realistic market structures. There are two typical assumptions used in the agent-based literature to determine the asset price: a market clearing or a market maker mechanism. Recent progress in the analysis of the micro-structure of financial markets has indicated the increasing importance of alternative trading mechanisms, such as order-driven markets. In these markets, traders decide whether to buy (sell) using a market or limit order. A market order means that the trader is ready to buy (sell) a certain quantity of stocks at the best available price; instead, with limit orders traders fix both a quantity of shares and a price at which they are willing to buy (sell). Limit orders are stored in the book until a matching order arrives to the market. They are different from quote-driven markets, where a market maker provides quotes and liquidity to investors. This has spurred a series of articles that propose agent-

based models in this more realistic market structure. In particular, [13,36,44], and [45] consider an order-driven market where agents submit limit orders. Typically these models make simple behavioral assumptions on the belief formation process and do not consider learning and evolution of agents' expectations (typical of the computational agent-based models). In this sense, these models are closer to the stochastic agent-based approach reviewed in Sect. "Analytical Agent-Based Models". Recently, [33] has proposed a computational model for an order-driven market in which strategies evolve by imitation of the most successful rules.

[13] propose an order-driven market model in which the demand for the risky asset of the agents is determined by a fundamentalist, a chartist, and a noise component. The agents share the same demand function but the weights on the components are different across agents. Simulating the model suggests that the stylized facts of financial returns can be explained when all behavioral components (fundamentalist, chartist, and noise) participate to determine agents' beliefs. An additional feature of this setup is that it accounts for the persistence in the volatility of returns and trading volume. Such a micro-structural model allows also to investigate the effect of some key market design parameters (such as tick size, liquidity, and average life of an order) on the price formation process.

[44] consider a market structure where agents submit limit orders and the price is determined by market clearing of supply (sell orders) and demand (buy orders) schedules. The behavioral assumptions are closely related to the clustering approach of [14]: a probabilistic mechanism governs the formation of clusters and, within a clusters, all agents coordinate in buying or selling the risky asset. Another behavioral assumption introduced in this model concerns the (positive) relation between market volatility and the limit order price. When the volatility is low, agents set the price of their limit order close to yesterday's asset price. However, when the market is experiencing wide swings in prices, agents' set limit prices that are significantly above or below yesterday's price for their orders. The results suggest that the model is able to explain the main stylized facts of financial returns. [45] consider an economy with a similar market structure but more sophisticated assumption on agents' behavior. They assume the population is composed of four types of agents: random traders (with 50% probability to buy or sell), momentum (buy/sell following an increase/decrease in prices), contrarian (act in the opposite direction of momentum traders), and fundamentalists (buy/sell if the price is below/above the fundamental value). They simulate the model in a way that non-random agents do not affect the asset price. The idea is to investigate the survival of these types of agents in the economy without affecting the aggregate properties of the model. They show that, on average, the fraction of wealth of momentum agents decreases while it increases for fundamentalist and contrarian traders.

Another recent paper that models an order-driven market is [36]. Agents can submit market and limit orders. They introduce the behavioral assumption that the agents are all fundamentalists, although they are heterogeneous in their belief of the fundamental value of the asset. They show that simulated series from this simple model follow a leptokurtic distribution and attribute this property to the structure of the market (rather than the behavioral assumptions). The same result is also obtained when random traders are considered. However, they are not able to explain other stylized fact such as the autocorrelation structure of volatility. This paper is interesting because it suggest that some of the stylized facts discussed earlier might not be related to agents' bounded rationality, but rather to the details of the market mechanism that is typically neglected in the agent-based literature.

Another area of application of agent-based modeling is corporate finance. [43] propose an agent-based model to investigate the security issuance preferences of firms and investors. The empirical evidence indicates that there is a clear dominance of debt financing, compared to other instruments, such as equities and convertible debt. This is a puzzle for theoretical models where it is customarily assumed that the payoff structure of the financing instruments are common knowledge. Under this assumption, the price should reflect the different characteristics of the securities and investors should be indifferent among them. Noe et al. consider a model that departs from the assumption of perfect knowledge about the security characteristics, and assume that firms and investors are learning (via a GA) about the profitability of the different alternatives. Debt and equity imply different degrees of risk-sharing between investors and firms: in particular, debt provides lower risk and return, contrary to equities that have a more volatile payoff structure. Investors' learning about the risk-return profile of the different instruments leads to the prevalence of debt on equity or convertible debt. Another conclusion of this model is that learning is imperfect: agents learn to price specific contracts and have difficulties in dealing with contracts that rarely occur in the market.

Future Directions

Agent-based modeling in finance has had a significant impact in shaping the way we understand the working of

financial markets. By introducing realistic behavioral assumptions, agent-based models have demonstrated that they provide a coherent explanation for many empirical findings in finance. In addition, they are also able to provide a framework to explain how aggregate rationality can emerge in a population of bounded rational learning agents.

The strength of the agent-based approach is the ability to specify in greater detail the agents' behavior and the structure of market interactions. Simple agent-based models use realistic assumptions and can be solved analytically. However, they sacrifice the important aspect of the emergence of aggregate pattern based on agents' learning. This can be achieved by computational agent-based models. Since the approach is not bounded by the analytical tractability of the model, very detailed (and potentially more realistic) assumption can be introduced. However, this can represent a weakness of the approach since it might lead to over-parametrized models where it is hard to disentangle the role played by each of the assumptions on the aggregate behavior. In this sense, agent-based modeling should aim at balancing parsimony and realism of agents' description.

As already suggested in Sect. "Other Applications in Finance", the application of agent-based models is not limited to asset pricing issues. Recently, they have been used in corporate finance and market microstructure. This is certainly a trend that will increase in the future since these models are particularly suitable to investigate the interaction of market structure and agents' behavior.

Bibliography

Primary Literature

1. Arifovic J (1996) The Behavior of the Exchange Rate in the Genetic Algorithm and Experimental Economies. J Political Econ 104:510–541
2. Arifovic J, Gencay R (2000) Statistical properties of genetic learning in a model of exchange rate. J Econ Dyn Control 24:981–1006
3. Arthur WB, Holland JH, LeBaron B, Tayler P (1997) Asset pricing under endogeneous expectations in an artificial stock market. In: Lane D, Arthur WB, Durlauf S (ed) The Economy as an Evolving Complex System II. Addison–Wesley, Reading
4. Beltratti A, Margarita S (1992) Evolution of trading strategies among heterogeneous artificial economic agents. In: Wilson SW, Meyer JA, Roitblat HL (ed) From Animals to Animats II. MIT Press, Cambridge, pp 494–501
5. Boswijk HP, Hommes CH, Manzan S (2007) Behavioral heterogeneity in stock prices. J Econ Dyn Control 31:1938–1970
6. Brav A, Heaton JB (2002) Competing Theories of Financial Anomalies. Rev Financ Stud 15:575–606
7. Brock WA, Hommes CH (1998) Heterogeneous beliefs and and routes to chaos in a simple asset pricing model. J Econ Dyn Control 22:1235–1274
8. Chen S-H, Yeh C-H (2002) On the emergence properties of artificial stock markets: the efficient market hypothesis and the rational expectations hypothesis. J Econ Behav Organ 49:217–239
9. Chiarella C (1992) The dynamics of speculative behavior. Ann Oper Res 37:101–123
10. Chiarella C, He XZ (2001) Asset price and wealth dynamics under heterogeneous expectations. Quant Financ 1:509–526
11. Chiarella C, He XZ (2002) Heterogeneous Beliefs, Risk and Learning in a Simple Asset Pricing Model. Comput Econ 19:95–132
12. Chiarella C, He XZ (2003) Heterogeneous beliefs, risk, and learning in a simple asset-pricing model with a market maker. Macroecon Dyn 7:503–536
13. Chiarella C, Iori G (2002) A simulation analysis of the microstructure of double auction markets. Quant Financ 2: 346–353
14. Cont R, Bouchaud JP (2000) Herd behavior and aggregate fluctuations in financial markets. Macroecon Dyn 4:170–196
15. Day R, Huang W (1990) Bulls, bears and market sheep. J Econ Behav Organ 14:299–329
16. de Grauwe P, Dewachter H (1993) A chaotic model of the exchange rate: The role of fundamentalists and chartists. Open Econ Rev 4:351–379
18. de Grauwe P, Grimaldi M (2005) The exchange rate and its fundamentals in a complex world. Rev Int Econ 13:549–575
17. de Grauwe P, Dewachter H, Embrechts M (1993) Exchange Rate Theory – Chaotic Models of Foreign Exchange Markets. Blackwell, Oxford
19. Farmer JD, Joshi S (2002) The price dynamics of common trading strategies. J Econ Behav Organ 49:149–171
20. Frankel JA, Froot KA (1987) Using survey data to test standard propositions regarding exchange rate expectations. Am Econ Rev 77:133–153
21. Frankel JA, Froot KA (1990) Chartists, fundamentalists, and trading in the foreign exchange market. Am Econ Rev 80:181–185
22. Gaunersdorfer A (2000) Endogenous fluctuations in a simple asset pricing model with heterogeneous agents. J Econ Dyn Control 24:799–831
23. Gaunersdorfer A, Hommes CH (2006) A nonlinear structural model for volatility clustering. In: Kirman A, Teyssiere G (ed) Microeconomic models for long memory in economics. Springer, Berlin, pp 265–288
24. Goldbaum D (2005) Market efficiency and learning in an endogenously unstable environment. J Econ Dyn Control 29:953–978
25. Grossman SJ, Stiglitz JE (1980) On the Impossibility of Informationally Efficient Markets. Am Econ Rev 70:393–408
26. Hommes CH (2006) Heterogeneous agent models in economics and finance. In: Judd KL, Tesfatsion L (ed) Handbook of Computational Economics, vol 2: Agent-Based Computational Economics. North-Holland, Amsterdam, pp 1109–1185
27. Kirman A (1991) Epidemics of opinion and speculative bubbles in financial markets. In: Taylor MP (ed) Money and Financial Markets. Blackwell, Oxford, pp 354–368
28. Kirman A, Teyssiere G (2001) Microeconomics models for long-

memory in the volatility of financial time series. Stud Nonlinear Dyn Econ 5:281–302
29. LeBaron B (2001) Empirical regularities from interacting long and short memory investors in an agent based in stock market. IEEE Trans Evol Comput 5:442–455
30. LeBaron B (2001) Evolution and time horizons in an agent based stock market. Macroecon Dyn 5:225–254
31. LeBaron B (2002) Short-memory traders and their impact on group learning in financial markets. In: Proceedings of the National Academy of Science: Colloquium, Washington DC, 99:7201–7206
32. LeBaron B (2006) Agent-based computational finance. In: Judd KL, Tesfatsion L (ed) Handbook of Computational Economics, vol 2: Agent-Based Computational Economics. North-Holland, Amsterdam, pp 1187–1233
33. LeBaron B, Yamamoto R (2007) Long-memory in an order-driven market. Physica A 383:85–89
34. Levy M, Levy H, Solomon S (1994) A microscopic model of the stock market. Econ Lett 45:103–111
35. Levy M, Solomon S, Levy H (2000) Microscopic Simulation of Financial Markets. Academic Press, New York
36. Licalzi M, Pellizzari P (2003) Fundamentalists clashing over the book: a study of order-driven stock markets. Quant Financ 3:470–480
37. Lucas RE (1978) Asset prices in an exchange economy. Econometrica 46:1429–1445
38. Lux T (1995) Herd Behaviour, Bubbles and Crashes. Econ J 105:881–896
39. Lux T (1997) Time variation of second moments from a noise trader/infection model. J Econ Dyn Control 22:1–38
40. Lux T (1998) The socio-economic dynamics of speculative markets: interacting agents, chaos, and the fat tails of return distributions. J Econ Behav Organ 33:143–165
41. Lux T, Marchesi M (1999) Scaling and criticality in a stochastic multi-agent model of interacting agents. Nature 397:498–500
42. Lux T, Marchesi M (2000) Volatility clustering in financial markets: A micro-simulation of interacting agents. Int J Theoret Appl Financ 3:675–702
43. Noe TH, Rebello MJ, Wang J (2003) Corporate financing: an artificial agent-based analysis. J Financ 58:943–973
44. Raberto M, Cincotti S, Focardi SM, Marchesi M (2001) Agent-based simulation of a financial market. Physica A: Stat Mech Appl 299:319–327
45. Raberto M, Cincotti S, Focardi SM, Marchesi M (2003) Traders'Long-Run Wealth in an Artificial Financial Market. Comput Econ 22:255–272
46. Rieck C (1994) Evolutionary simulations of asset trading strategies. In: Hillebrand E, Stender J (ed) Many-Agent Simulation and Artificial Life. IOS Press, Amsterdam
47. Routledge BR (1999) Adaptive Learning in Financial Markets. Rev Financ Stud 12:1165–1202
48. Routledge BR (2001) Genetic algorithm learning to choose and use information. Macroecon Dyn 5:303–325
49. Samanidou E, Zschischang E, Stauffer D, Lux T (2007) Microscopic Models of Financial Markets. In: Schweitzer F (ed) Microscopic Models of Economic Dynamics. Springer, Berlin
50. Westerhoff FH (2003) Expectations driven distortions in the foreign exchange market. J Econ Behav Organ 51:389–412
51. Westerhoff FH (2004) Greed, fear and stock market dynamics. Physica A: Stat Mech Appl 343:635–642

Books and Reviews

Anderson PW, Arrow KJ, Pines D (1989) The Economy as an Evolving Complex System. Addison-Wesley, Reading
Arthur WB, Durlauf SN, Lane DA (1999) The Economy as an Evolving Complex System: vol 2. Addison-Wesley, Reading
Blume LE, Durlauf SN (2005) The Economy as an Evolving Complex System: vol 3. Oxford University Press, Oxford
Judd KL, Tesfatsion L (2006) Handbook of Computational Economics, vol 2: Agent-based Computational Economics. North-Holland, Amsterdam

Finance and Econometrics, Introduction to

Bruce Mizrach
Department of Economics, Rutgers University, New Jersey, USA

Article Outline

Introduction
Econometrics
Agent Based Modeling
Finance
Market Microstructure
Conclusion
Acknowledgments
Bibliography

Introduction

Economics and finance have slowly emerged from the Walrasian, representative agent paradigm exemplified by the research agenda in general equilibrium theory. This program may have reached its pinnacle in the 1970s, with a highly abstract treatment of the existence of a market clearing mechanism. The normative foundation of this research was provided by powerful welfare theorems that demonstrated the optimality of the market allocations. Unfortunately, this abstract world had little economics in it. The models rarely provided empirical implications. Lifetime consumption and portfolio allocation plans were formed in infancy, unemployment was Pareto optimal, and the role for government was largely limited to public goods provision.

The demonstration by Benhabib, Brock, Day, Gale, Grandmont, [1,4,8,9] and others, that even simple math-

ematical models could display highly complex dynamics was the beginning of a new research program in economics. This section on finance and econometrics surveys some of the developments of the last 20 years that were inspired by this research.

Econometrics

Time series econometrics was originally built on the representation theorems for Euclidean spaces. The existence of a Wold decomposition in linear time series led to the widespread use of Box–Jenkins [3] style modeling as an alternative to structural or reduced form models.

A number of stylized facts about the economy emerged that simply could not be explained in this linear world. Rob Engle [2] and Tim Bollerslev [5] showed that volatility was quite persistent, even in markets that appeared to be nearly random walks. In ► GARCH Modeling, Christian Hafner surveys the extensive development in this area.

James Hamilton [10] and Salih Neftci [11] demonstrated that the business cycle was asymmetric and could be well described by a Markov switching model. James Morley ► Macroeconomics, Non-linear Time Series in and Jeremy Piger ► Econometrics: Models of Regime Changes describe the developments in this area. Virtually all the moments, not just the conditional mean, are now thought to be varying over the business cycle. These models help us to understand why recessions are shorter than expansions and why certain variables lead and lag the cycle.

Nearly all the business cycle models involve the use of latent or unobservable state variables. This reflects a reality that policy makers themselves face. We rarely know whether we are in a recession until it is nearly over. These latent variable models are often better described in a Bayesian rather than a classical paradigm. Oleg Korenok ► Bayesian Methods in Non-linear Time Series provides an introduction to the frontier research in this area.

Markets are often drawn towards equilibrium states in the absence of exogenous shocks, and, since the 1940s, this simple idea was reflected in the building of macroeconometric models. In linear models, Engle and Granger [6] formalized this notion in an error correction framework. When the adjustment process is taking place between two variables that are not stationary, we say that they are cointegrated. Escanciano and Escribano extend the error correction framework and cointegration analysis to nonlinear models in ► Econometrics: Non-linear Cointegration.

Because we often know very little about the data generating mechanism for an economy, nonparametric methods have become increasingly popular in the analysis of time series. Cees Diks discusses in ► Nonparametric Tests for Independence methods to analyze both data and the residuals from an econometric model.

Our last two entries look at the data generated by individual consumers and households. Pravan Trivedi ► Microeconometrics surveys the microeconometric literature, and Jeff Wooldridge ► Econometrics: Panel Data Methods examines the tools and techniques useful for analyzing cross-sectional data.

Agent Based Modeling

The neo-classical synthesis in economics was built upon the abstraction of a single optimizing agent. This assumption simplified the model building and allowed for analytical solutions of the standard models. As computational power became cheaper, it became easier to relax these assumptions. Many economists underestimated the complexity of a world in which multiple agents interact in a dynamic setting. Econophysicists, as Bertrand Roehner describes in ► Econophysics, Observational, were not surprised. Roehner is just one of scores of physicists who have brought their tools and perspectives to economics.

Agent based modeling has had a large impact on finance. Financial economics had been led by a Chicago influenced school that saw markets as both rational and efficient. Behavioral finance has eroded the view that people always make optimizing decisions even when large sums of money are at stake. The boundedly rational agents in Sebastiano Manzan's ► Finance, Agent Based Modeling in are prone to speculative bubbles. Markets crash suddenly in agent based computational models and in large scale experimental stock markets.

Finance

The foundation of financial economics is the theory of optimal consumption and saving. The goal of the empirical literature was to identify a set of risk factors that would explain why certain assets have a higher return than others. Ralitsa Petkova ► Financial Economics, The Cross–Section of Stock Returns and the Fama-French Three Factor Model surveys the canonical model of Fama and French [7] and the extensions to this model in the last decade.

With risk averse agents, asset returns are often predictable. Stijn van Nieuwerburgh and Ralph S.J. Koijen ► Financial Economics, Return Predictability and Market Efficiency demonstrate the robustness of this result in a structural model and show that the dividend price ratio does predict future stock returns.

Mototsugu Shintani addresses in ► Financial Forecasting, Sensitive Dependence the concept of predictability from an information theoretic perspective through the use of Lyapunov exponents. The exponents not only tell us which systems display sensitive dependence on initial conditions ("chaos") but also provide a predictive horizon for data generated by the model. Shintani finds that financial data appear to not be chaotic, even though they display local dependence on initial conditions.

Mark Kamstra and Lisa Kramer's entry on ► Financial Economics, Time Variation in the Market Return primarily focus on the equity premium, the substantially higher return in the US and other countries on equities, over default free securities like Treasury bonds. They document its statistical significance and discuss some behavioral explanations. They demonstrate that behavioral moods can influence asset prices.

Terence Mills' ► Financial Economics, Non-linear Time Series in surveys the use of nonlinear time series techniques in finance. Gloria Gonzalez-Rivera and Tae-Hwy Lee look at the ability of nonlinear models to forecast in ► Financial Forecasting, Non-linear Time Series in. They also cover the methodology for assessing forecast improvement. The best forecast may not be the one that predicts the mean most accurately; it may instead be the one that keeps you from large losses.

Our last two papers in this area focus on volatility. Markus Haas and Christian Pigorsch discuss the ubiquitous phenomenon of fat-tailed distributions in asset markets in ► Financial Economics, Fat-Tailed Distributions. They provide evidence on the frequency of extreme events in many different markets, and develop the implications for risk management when the world is not normally distributed. Torben Andersen and Luca Benzoni ► Stochastic Volatility introduce the standard volatility model from the continuous time finance literature. They contrast it with the GARCH model discussed earlier and develop econometric methods for estimating volatility from discretely sampled data.

Market Microstructure

Market microstructure examines the institutional mechanisms by which prices adjust to their fundamental values. The literature has grown with the availability of transactions frequency databases. Clara Vega and Christian Miller ► Market Microstructure survey the topic largely from a theoretical perspective. Because disparate markets are likely to have different mechanisms and regulators, the literature has evolved by instrument. Carol Osler ► Market Microstructure, Foreign Exchange examines the microstructure of the foreign currency market, the largest and most liquid asset market. Bruce Mizrach and Chris Neely ► Treasury Market, Microstructure of the U.S. look at the government bond market in the US as it has evolved into an electronic market. Michael Piwowar ► Corporate and Municipal Bond Market Microstructure in the U.S. looks at two bond markets with a large number of issues that trade only very infrequently. Both the markets which he examines have become substantially more transparent through recent government initiatives.

Conclusion

This section covers a wide range of material from theoretical time series analysis to descriptive modeling of financial markets. The theme of complexity is a unifying one in the sense that the models are generally nonlinear and can produce a wide range of possible outcomes. There is complexity in the data which now evolves at a millisecond frequency. Readers should find a variety of perspectives and directions for future research in a heterogenous but interconnected range of fields.

Acknowledgments

I would like to thank all of the contributors to this section of the encyclopedia.

Bibliography

1. Benhabib J, Day RH (1982) A characterization of erratic dynamics in the overlapping generations models. J Econ Dyn Control 4:37–55
2. Bollerslev TP (1986) Generalized autoregressive conditional heteroscedasticity. J Econ 31:307–327
3. Box G, Jenkins G (1994) Time Series Analysis Forecasting and Control, 3rd ed. Prentice Hall
4. Brock WA (1986) Distinguishing random and deterministic systems. J Econ Theory 40:168–195
5. Engle RF (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation. Econ 50:987–1008
6. Engle RF, Granger CWJ (1987) Co-integration and error correction: representation, estimation, and testing. Econometrica 55:251–276
7. Fama E, French K (1992) The cross-section of expected stock returns. J Finance 47:427–465
8. Gale D (1973) Pure exchange equilibrium of dynamic economic models. J Econ Theory 6:12–36
9. Grandmont JM (1985) On Endogenous Competitive Business Cycles. Econometrica 53:995–1045
10. Hamilton JD (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57:357–384
11. Neftçi SH (1984) Are economic time series asymmetric over the business cycle? J Political Econ 92:307–328

Financial Economics, The Cross-Section of Stock Returns and the Fama-French Three Factor Model

RALITSA PETKOVA
Mays Business School, Texas A&M University, College Station, USA

Article Outline

Glossary
Definition of the Subject
Introduction
The Fama–French Model as a Linear Beta Pricing Model
Explaining the Performance of the Fama–French Model: A Risk-Based Interpretation
Other Risk-Based Interpretations
Future Directions
Bibliography

Glossary

Market capitalization Market capitalization is a measure of the size of a public company. It is equal to the share price times the number of shares outstanding. Small stocks have small market capitalizations, while large stocks have large market capitalizations.

Book-to-market ratio A ratio used to compare a company's book value to its market capitalization value. It is calculated by dividing the latest book value by the latest market value of the company.

Value stocks Value stocks tend to trade at lower prices relative to fundamentals like dividends, earnings, sales and others. These stocks are considered undervalued by value investors. Value stocks usually have high dividend yields, and high book-to-market ratios.

Growth stocks Growth stocks tend to trade at higher prices relative to fundamentals like dividends, earnings, sales and others. Growth stocks usually do not pay dividends and have low book-to-market ratios.

Market beta The market beta is a measure of the systematic risk of a security in comparison to the market as a whole. It measures the tendency of the security return to respond to market movements.

Capital asset pricing model (CAPM) The CAPM describes the relationship between risk and expected return and it is used in the pricing of risky securities. According to the CAPM, the expected return of a security equals the rate on a risk-free security plus a risk premium that increases in the security's market beta.

Definition of the Subject

Different stocks have different expected rates of return and many asset pricing models have been developed to understand why this is the case. According to such models, different assets earn different average returns because they differ in their exposures to systematic risk factors in the economy. Fama and French [12] derive a model in which the systematic risk factors are the market index, and two portfolios related to the size of a company, and its ratio of book value to market value (book-to-market). The size and book-to-market factors are empirically motivated by the observation that small stocks and stocks with high book-to-market ratios (value stocks) earn higher average returns than justified by their exposures to market risk (beta) alone. These observations suggest that size and book-to-market may be proxies for exposures to sources of systematic risk different from the market return.

Introduction

An important class of asset pricing models in finance are linear beta models. They assume that the expected return of an asset in excess of the risk-free rate is a linear function of exposures to systematic sources of risk. Usually, the asset's exposures to common sources of risk in the economy are referred to as betas. In general, linear beta models assume the following form for the unconditional expected excess return on assets:

$$E(R_i) = \gamma_M \beta_{i,M} + \sum \gamma_K \beta_{i,K}, \quad \text{for all} \quad i \tag{1}$$

where $E(R_i)$ is the expected excess return of asset i, γ_M is the market risk premium or the price for bearing market risk, and γ_K is the price of risk for factor K. The model stated above implies that exposures to systematic sources of risk are the only determinants of expected returns. Thus, assets with high betas earn higher expected returns. The betas are the slope coefficients from the following return-generating process:

$$R_{i,t} = \alpha_i + \beta_{i,M} R_{M,t} + \sum \beta_{i,K} K_t + \varepsilon_{i,t}, \quad \text{for all} \quad i \tag{2}$$

where $R_{i,t}$ is the return on asset i in excess of the risk-free rate at the end of period t, $R_{M,t}$ is the excess return on the market portfolio at the end of period t, and K_t is the realization for factor K at the end of period t.

One approach of selecting the pervasive risk factors is based on empirical evidence. For example, many empirical studies document that small stocks have higher average returns than large stocks, and value stock have higher average returns than growth stocks (see [12] for a review). The differences in average returns of these classes of stocks

are statistically and economically significant. If the market sensitivities of small and value stocks were high then their high average returns would be consistent with the Capital Asset Pricing Model (CAPM), which predicts that the market beta is the only determinant of average returns. However, the patterns in returns for these stocks cannot be explained by the CAPM.

In a series of papers, Fama and French [12,13,14] show that a three-factor model performs very well at capturing the size and value effects in average stock returns. The three factors are the excess return on the market portfolio, the return on a portfolio long in value stocks and short in growth stocks, and the return on a portfolio long in small stocks and short in large stocks.

The impressive performance of the Fama–French three-factor model has spurred an enthusiastic debate in the finance literature over what underlying economic interpretation to give to the size and book- to-market factors. One side of the debate favors a risk-based explanation and contends that these factors reflect systematic risks that the static CAPM has failed to capture. For example, if the return distributions of different assets change over time (i. e., expected returns, variances, correlation), then the investment opportunity set available to investors varies over time as well. If individual assets covary with variables that track this variation then the expected returns of these assets will reflect that. Fama and French argue that the factors in their model proxy for such variables.

Another side of the debate favors a non-risk explanation. For example, Lakonishok, Shleifer, and Vishny [22] argue that the book-to-market effect arises since investors over-extrapolate past earnings growth into the future and overvalue companies that have performed well in the past. Namely, investors tend to over-extrapolate recent performance: they overvalue the firms with good recent performance (growth) and undervalue the firms with bad recent performance (value). When the market realizes its mistake, the prices of the former fall, while the prices of the latter rise. Therefore on average, growth firms tend to underperform value firms. Daniel and Titman [9] suggest that stocks characteristics, rather than risks, are priced in the cross-section of average returns. Other authors attribute the success of the size and book-to-market factors to data-snooping and other biases in the data [21,27]. Berk, Green, and Naik [1] and Gomes, Kogan, and Zhang [17] derive models in which problems in the measurement of market beta may explain the Fama–French results.

This article focuses on the risk-based explanation behind the success of the Fama–French three-factor model. If the Fama–French factors are to be explained in the context of a rational asset pricing model, then they should be correlated with variables that characterize time variation in the investment opportunity set. The rest of the article is organized as follows. Section "The Fama–French Model as a Linear Beta Pricing Model" discusses the set-up of the Fama–French model and presents some empirical tests of the model. Section "Explaining the Performance of the Fama–French Model: A Risk-Based Interpretation" argues that the Fama–French factors proxy for fundamental variables that describe variation in the investment opportunity set over time, and presents empirical results. Section "Other Risk-Based Interpretations" presents additional arguments for the relation between the Fama–French factors and more fundamental sources of risk. Section "Future Directions" summarizes and concludes.

The Fama–French Model as a Linear Beta Pricing Model

Model Set-up

Fama and French [12] propose a three-factor linear beta model to explain the empirical performance of small and high book-to-market stocks. The intuition behind the factors they propose is the following.

If small firms earn higher average returns than large firms as a compensation for risk, then the return differential between a portfolios of small firms and a portfolio of large firms would mimic the factor related to size provided the two portfolios have similar exposures to other sources of risk. Similarly, if value firms earn higher average returns than growth firms as a compensation for risk, then the return differential between a portfolio of value firms and a portfolio of growth firms, would mimic the factor related to book-to-market provided the two portfolios have similar exposure to other sources of risk. Fama and French [12] construct two pervasive risk factors in this way that are now commonly used in empirical studies. The composition of these factors is explained below.

In June of each year independent sorts are used to allocate the NYSE, AMEX, and NASDAQ stocks to two size groups and three book-to-market groups. Big stocks are above the median market equity of NYSE firms and small stocks are below. Similarly, low book-to-market stocks are below the 30th percentile of book-to-market for NYSE firms, medium book-to-market stocks are in the middle 40 percent, and high book-to-market stocks are in the top 30 percent. Size is market capitalization at the end of June. Book-to-market is book equity at the last fiscal year end of the prior calendar year divided by market cap as of 6 months before formation. Firms with negative book equity are not considered. At the end of June of each year, six value-weight portfolios are formed, SL, SM, SH, BL, BM,

and BH, as the intersections of the size and book-to-market groups. For example, SL is the value-weight return on the portfolio of stocks that are below the NYSE median in size and in the bottom 30 percent of book-to-market. The portfolios are rebalanced annually. *SMB* in each period is the difference between the equal-weight averages of the returns on the three small stock portfolios and the three big stock portfolios, constructed to be neutral with respect to book-to-market:

$$SMB = (SL + SM + SH)/3 - (BL + BM + BH)/3\,. \quad (3)$$

Similarly, *HML* in each period is the difference between the return on a portfolio of high book-to-market stocks and the return on a portfolio of low book-to-market stocks, constructed to be neutral with respect to size:

$$HML = (SH + BH)/2 - (SL + BL)/2\,. \quad (4)$$

Therefore, the Fama–French three-factor linear model implies that:

$$E(R_i) = \gamma_M \beta_{i,M} + \gamma_{SMB}\beta_{i,SMB} + \gamma_{HML}\beta_{i,HML}, \quad \text{for all} \quad i \quad (5)$$

where $E(R_i)$ is the excess return of asset i, γ_M is the market risk premium, γ_{SMB} is the price of risk for the size factor, and γ_{HML} is the price of risk for the book-to-market factor. The betas are the slope coefficients from the following return-generating process:

$$R_{i,t} = \alpha_i + \beta_{i,M}R_{M,t} + \beta_{i,SMB}R_{SMB,t} + \beta_{i,HML}R_{HML,t} + \varepsilon_{i,t}, \quad \text{for all} \quad i \quad (6)$$

where $R_{i,t}$ is the return on asset i in excess of the risk-free rate at the end of period t, $R_{M,t}$ is the excess return on the market portfolio at the end of period t, $R_{SMB,t}$ is the return on the SMB portfolio at the end of period t, and $R_{HML,t}$ is the return on the *HML* portfolio at the end of period t.

Testing the Fama–French Model and Results

The return-generating process is Eq. (6) applies to the excess return of any asset. The Fama–French model is usually tested on a set of portfolios sorted by book-to-market and size. Similarly to the construction of *HML* and *SMB*, 25 value-weighted portfolios are formed as the intersections of five size and five book-to-market groups. These 25 portfolios are the test assets used most often in testing competing asset-pricing models. These assets represent one of the most challenging set of portfolios in the asset pricing literature.

In this article, monthly data for the period from July of 1963 to December of 2001 is used. The returns on the market portfolio, the risk-free rate, *HML*, and *SMB* are taken from Ken French's web site, as well as the returns on 25 portfolios sorted by size and book-to-market.

To test the Fama–French specification in Eq. (5), the Fama–MacBeth [15] cross-sectional method can be used. In the first pass of this method, a multiple time-series regression as in (6) is estimated for each one of the 25 portfolios mentioned above which provides estimates of the assets' betas with respect to the market return, and the size and book-to-market factors.

Table 1 reports the estimates of the factor loadings computed in the first-pass time-series regression (6) for each portfolio. The table also present joint tests of the significance of the corresponding loadings, computed from a seemingly unrelated regressions (SUR) system. This is done in order to show that the Fama–French factors are relevant in the sense that the 25 portfolios load significantly on them.

The results from Table 1 reveal that within each size quintile, the loadings of the portfolios with respect to *HML* increase monotonically with book-to-market. Within each size group, portfolios in the lowest book-to-market quintile (growth) have negative betas with respect to *HML*, while portfolios in the highest book-to-market quintile (value) have positive betas with respect to *HML*. Further, within each book-to-market quintile, the loadings of the portfolios with respect to *SMB* decrease monotonically with size. Within each book-to-market group, portfolios in the lowest size quintile (small) have positive betas with respect to *SMB*, while portfolios in the highest size quintile (large) have negative betas with respect to *SMB*. The table shows that small and large portfolios, and value and growth portfolios have similar market betas.

Note that only six of the 25 intercepts in Table 1 are significant (although the intercepts are jointly significant). The large R-square statistics show that the excess returns of the 25 portfolios are explained well by the three-factor model. Furthermore the large t-statistics on the size and book-to-market betas show that these factors contribute significantly to the explanatory power of the model.

The second step of the Fama–MacBeth procedure involves relating the average excess returns of the 25 portfolios to their exposures to the risk factors in the model. More specifically, the following cross-sectional relation is estimated

$$\overline{R}_{i,t} = \gamma_0 + \gamma_M\widehat{\beta}_{i,M} + (\gamma_{HML})\widehat{\beta}_{i,HML} + (\gamma_{SMB})\widehat{\beta}_{i,SMB} + e_{i,t}\,. \quad (7)$$

Financial Economics, The Cross-Section of Stock Returns and the Fama-French Three Factor Model, Table 1

Loadings on the Fama–French Factors from Time-Series Regressions

This table reports loadings on the excess market return, R_M, and the Fama–French factors R_{HML} and R_{SMB} computed in time-series regressions for 25 portfolios sorted by size and book-to-market. The corresponding *t*-statistics are also reported and are corrected for autocorrelation and heteroscedasticity using the Newey–West estimator with five lags. The sample period is from July 1963 to December 2001. The intercepts are in percentage form. The last column reports *F*-statistics and their corresponding *p*-values from an SUR system, testing the joint significance of the corresponding loadings. The *p*-values are in percentage form. R^2s from each time-series regression are reported in percentage form

Regression: $R_{i,t} = \alpha_i + \beta_{i,M}R_{M,t} + \beta_{i,HML}R_{HML,t} + \beta_{i,SMB}R_{SMB,t} + \varepsilon_{i,t}$												
	Low	2	3	4	High		Low	2	3	4	High	
	α						t_α					F
Small	−0.38	0.01	0.04	0.18	0.12		−3.40	0.18	0.56	2.84	1.91	2.96
2	−0.17	−0.10	0.08	0.08	−0.00		−2.25	−1.45	1.15	1.28	−0.01	0.01
3	−0.07	−0.00	−0.09	0.01	0.00		−1.03	−0.03	−1.26	0.17	0.06	
4	0.16	0.21	−0.08	0.04	−0.05		1.67	−2.27	−0.99	0.61	−0.54	
Large	0.21	−0.04	−0.02	−0.09	−0.21		3.25	−0.53	−0.27	−1.29	−2.36	
	β_M						t_{β_M}					F
Small	1.04	0.96	0.93	0.92	0.98		44.38	39.40	50.88	46.60	43.39	> 100
2	1.11	1.03	1.00	0.99	1.08		48.84	45.42	46.47	60.69	52.11	< 0.01
3	1.09	1.07	1.03	1.01	1.10		52.59	38.53	32.93	52.70	38.97	
4	1.05	1.11	1.08	1.03	1.17		46.03	36.33	36.86	41.15	36.74	
Large	0.96	1.04	0.99	1.01	1.04		45.08	49.22	36.71	46.18	31.59	
	β_{HML}						$t_{\beta_{HML}}$					F
Small	−0.31	0.09	0.31	0.47	0.69		−5.86	1.79	9.62	14.97	17.10	> 100
2	−0.38	0.18	0.43	0.59	0.76		−8.52	2.96	7.36	13.97	23.28	< 0.01
3	−0.43	0.22	0.52	0.67	0.82		−14.90	3.10	7.39	10.58	15.94	
4	−0.45	0.26	0.51	0.61	0.83		−10.55	3.42	7.43	11.92	16.07	
Large	−0.38	0.14	0.27	0.64	0.85		−10.47	2.58	5.65	11.82	20.56	
	β_{SMB}						$t_{\beta_{SMB}}$					F
Small	1.41	1.33	1.12	1.04	1.09		36.39	24.68	36.50	24.34	25.40	> 100
2	1.00	0.89	0.75	0.70	0.82		27.61	18.51	15.90	25.31	25.68	< 0.01
3	0.72	0.51	0.44	0.38	0.53		24.97	7.68	6.81	8.28	8.87	
4	0.37	0.20	0.16	0.20	0.26		9.26	3.42	2.64	6.70	4.22	
Large	−0.26	−0.24	−0.24	−0.22	−0.08		−9.25	−6.92	−6.12	−6.81	−2.11	
				R^2								
				92.61	94.32	94.89	94.51	94.58				
				95.16	93.99	93.56	93.85	94.62				
				94.88	90.22	89.49	89.69	90.31				
				93.52	88.31	87.65	88.41	85.77				
				93.35	89.79	84.32	87.39	80.60				

The $\widehat{\beta}$ terms are the independent variables in the regression, while the average excess returns of the assets are the dependent variables. If loadings with respect to the Fama–French factors are important determinants of average returns, then there should be a significant price of risk associated with the factors.

Since the betas are estimated from the time-series regression in (6), they represent generated regressors in (7). This is the classical errors-in-variables problem, arising from the two-pass nature of this approach. Following Shanken [33], a correction procedure can be used that accounts for the errors-in-variables problem. Shanken's correction is designed to adjust for the overstated precision of the Fama–MacBeth standard errors. It assumes that the error terms from the time-series regression are independently and identically distributed over time, conditional on the time series of observations for the risk factors. The adjustment also assumes that the risk factors are generated by a stationary process. Jagannathan and Wang [19] argue that if the error terms are heteroscedastic, then the Fama–

Financial Economics, The Cross-Section of Stock Returns and the Fama-French Three Factor Model, Table 2

Cross-Sectional Regressions with the Fama–French Factor Loadings

This table presents Fama–MacBeth cross-sectional regressions using the average excess returns on 25 portfolios sorted by book-to-market and size. The full-sample factor loadings, which are the independent variables in the regressions, are computed in one multiple time-series regression. The coefficients are expressed as percentage per month. The Adjusted R^2 follows Jagannathan and Wang [18] and is reported in percentage form. The first set of *t*-statistics, indicated by FM *t*-stat, stands for the Fama–MacBeth estimate. The second set, indicated by SH *t*-stat, adjusts for errors-in-variables and follows Shanken [33]. The sample period is from July 1963 to December 2001

The Fama–French Three-Factor Model					
	γ_0	γ_M	γ_{HML}	γ_{SMB}	Adj. R^2
Estimate	1.15	−0.65	0.44	0.16	71.00
FM *t*-stat	3.30	−1.60	3.09	1.04	
SH *t*-stat	3.19	−1.55	3.07	1.00	

MacBeth procedure does not necessarily result in smaller standard errors of the cross-sectional coefficients. In light of these two issues, researchers often report both unadjusted and adjusted cross-sectional statistics.

Table 2 reports the estimates of the factor prices of risk computed in the second-pass cross-sectional regression (7). The table also presents the t-statistics for the coefficients, adjusted for errors-in-variables following Shanken [33]. The table shows that the market beta is not an important factor in the cross-section of returns sorted by size and book-to-market.[1] Further, the table reveals that loadings on *HML* represent a significant factor in the cross-section of the 25 portfolios, even after correcting for the sampling error in the loadings. Loadings on *SMB* do not appear to be significant in the cross-section of portfolio returns for this time period. The large R-square of 0.71 shows that the loadings from the Fama–French model explain a significant portion of the cross-sectional variation in the average returns of these portfolios.

It is also helpful to examine the performance of the model visually. This is done by plotting the fitted expected return of each portfolio against its realized average return in Fig. 1. The fitted expected return is computed using the estimated parameter values from the Fama–French model specification. The realized average return is the time-series average of the portfolio return. If the fitted expected return and the realized average return for each portfolio are the same, then they should lie on a 45-degree line through the origin.

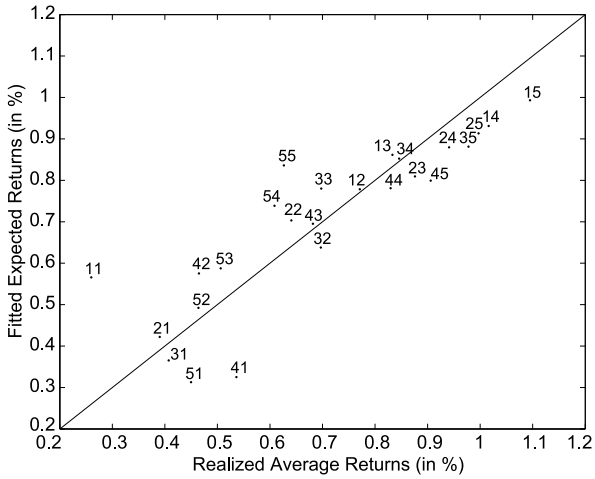

Financial Economics, The Cross-Section of Stock Returns and the Fama-French Three Factor Model, Figure 1

Fitted Expected Returns vs. Average Realized Returns for 1963:07-2001:12.

This figure shows realized average returns (%) on the *horizontal axis* and fitted expected returns (%) on the *vertical axis* for 25 size and book-to-market sorted portfolios. Each two-digit number represents a separate portfolio. The first digit refers to the size quintile (1 being the smallest and 5 the largest), while the second digit refers to the book-to-market quintile (1 being the lowest and 5 the highest). For each portfolio, the realized average return is the time-series average of the portfolio return and the fitted expected return is the fitted value for the expected return from the corresponding model. The straight line is the 45-degree line from the origin

Figure 1 shows the fitted versus realized returns for the 25 portfolios in two different models for the period from July of 1963 to December of 2001. Each two-digit number represents a separate portfolio. The first digit refers to the size quintile of the portfolio (1 being the smallest and 5 the biggest), while the second digit refers to the book-to-market quintile (1 being the lowest and 5 the highest). For example, portfolio 15 has the highest book-to-market value among the portfolios in the smallest size quintile. In other words, it is the smallest value portfolio.

It can be seen form the graph that the model goes a long way toward explaining the value effect: in general, the fitted expected returns on value portfolios (bigger second digit) are higher than the fitted expected returns on growth portfolios (lower second digit). This is consistent with the data on realized average returns for these portfolios. By inspection of Fig. 1, a few portfolios stand out as problematic for the FF model, in terms of distance from

[1] The estimate of the market risk premium tends to be negative. This result is consistent with previous results reported in the literature. Fama and French [11], Jagannathan and Wang [18], and Lettau and Ludvigson [24] report negative estimates for the market risk premium, using monthly or quarterly data.

the 45-degree line, namely the growth portfolios within the smallest and largest size quintiles (11, 41, and 51) and the value portfolios within the largest size quintiles (45, 54, and 55).

In summary, the Fama–French model performs remarkably well at explaining the average return difference between small and large, and value and growth portfolios.

The natural question that arises is what drives the superior performance of the Fama–French model in explaining average stock returns. One possible explanation is that the Fama–French factors *HML* and *SMB* proxy for sources of risk not captured by the return on the market portfolio. This explanation is consistent with a multifactor asset pricing model like the Intertemporal Capital Asset Pricing Model (ICAPM), which states that if investment opportunities change over time, then variables other than the market return will be important factors driving stock returns. Therefore, one possible interpretation of the *HML* and *SMB* portfolios is that they proxy for variables that describe how investment opportunities change over time. The following sections examine the ICAPM explanation behind the performance of the Fama–French model.

Explaining the Performance of the Fama–French Model: A Risk-Based Interpretation

The ICAPM Framework

The analysis in this paper assumes that asset returns are governed by the discrete-time version of the ICAPM of Merton [29]. According to the ICAPM, if investment opportunities change over time, then assets' exposures to these changes are important determinants of average returns in addition to the market beta. Campbell [3] develops a framework to model changes in the investment opportunity set as innovations in state variables that capture uncertainty about investment opportunities in the future. Therefore, the model for the unconditional expected excess returns on assets becomes

$$E(R_i) = \gamma_M \beta_{i,M} + \sum (\gamma_{u^K}) \beta_{i,u^K}, \quad \text{for all} \quad i \tag{8}$$

where $E(R_i)$ is the excess return of asset i, γ_M is the market risk premium, and γ_{u^K} is the price of risk for innovations in state variable K. The betas are the slope coefficients from the following return-generating process:

$$R_{i,t} = \alpha_i + \beta_{i,M} R_{M,t} + \sum (\beta_{i,u^K}) u_t^K + \varepsilon_{i,t}, \quad \text{for all} \quad i \tag{9}$$

where $R_{i,t}$ is the return on asset i in excess of the risk-free rate at the end of period t, $R_{M,t}$ is the excess return on the market portfolio at the end of period t, and u_t^K is the innovation to state variable K at the end of period t. The innovation is the unexpected component of the variable. According to the asset-pricing model, only the unexpected component of the state variable should command a risk premium. Note that the innovations to the state variables are contemporaneous to the excess market returns. This equation captures the idea that the market portfolio and the innovations to the state variables are the relevant risk factors.

It is important to specify a process for the time-series dynamics of the state variables in the model. A vector autoregressive (VAR) approach, for example, specifies the excess market return as the first element of a state vector z_t. The other elements of z_t are state variables that proxy for changes in the investment opportunity set. The assumption is that the demeaned vector z_t follows a first-order VAR:

$$z_t = \mathbf{A} z_{t-1} + \mathbf{u}_t \,. \tag{10}$$

The residuals in the vector $\mathbf{u}_t$ are the innovations terms which are the risk factors in Eq. (2). These innovations are risk factors since they represent the surprise components of the state variables that proxy for changes in the investment opportunity set.

The State Variables of Interest

For the empirical implementation of the model described above, it is necessary to specify the identity of the state variables. Petkova [31] chooses a set of state variables to model the following aspects of the investment opportunity set: the yield curve and the conditional distribution of asset returns. In particular, she chooses the short-term Treasury bill, the term spread, the aggregate dividend yield, and the default spread.

The choice of these state variables is motivated as follows. The ICAPM dictates that the yield curve is an important part of the investment opportunity set. Furthermore, Long [28] points out that the yield curve is important in an economy with a bond market. Therefore, the short-term Treasury bill yield (*RF*) and the term spread (*TERM*) are good candidates that capture variations in the level and the slope of the yield curve. Litterman and Scheinkman [26] show that the two most important factors driving the term structure of interest rates are its level and its slope.

In addition to the yield curve, the conditional distribution of asset returns is a relevant part of the investment opportunity set facing investors in the ICAPM world. There is growing evidence that the conditional distribution of asset returns, as characterized by its mean and variance, changes over time. The time-series literature has identified variables that proxy for variation in the mean and variance

of returns. The aggregate dividend yield (*DIV*), the default spread (*DEF*), and interest rates are among the most common.[2]

The variables described above are good candidates for state variable within the ICAPM. Merton [29] states that stochastic interest rates are important for changing investment opportunities. In addition, the default spread, the dividend yield, and interest rate variables have been used as proxies for time-varying risk premia under changing investment opportunities. Therefore, all these variables are likely to capture the hedging concerns of investors related the changes in interest rates and to variations in risk premia.

As argued in the previous sections of this article, two other variables proposed as candidates for state variables within the ICAPM are the returns on the *HML* and *SMB* portfolios. Fama and French [12] show that these factors capture common variation in portfolio returns that is independent of the market and that carries a different risk premium. The goal of the following section is to show that the FF factors proxy for the state variables described above that have been shown to track time-variation in the market risk premium and the yield curve.

Econometric Approach

First, a vector autoregressive (VAR) process for the vector of state variables is specified. The first element of the vector is the excess return on the market, while the other elements are DIV, $TERM$, DEF, RF, R_{HML}, and R_{SMB}, respectively. For convenience, all variables in the state vector have been demeaned. The first-order VAR is as follows:

$$\begin{Bmatrix} R_{M,t} \\ DIV_t \\ TERM_t \\ DEF_t \\ RF_t \\ R_{HML,t} \\ R_{SMB,t} \end{Bmatrix} = \mathbf{A} \begin{Bmatrix} R_{M,t-1} \\ DIV_{t-1} \\ TERM_{t-1} \\ DEF_{t-1} \\ RF_{t-1} \\ R_{HML,t-1} \\ R_{SMB,t-1} \end{Bmatrix} + \mathbf{u}_t \tag{11}$$

where $\mathbf{u}_t$ represents a vector of innovations for each element in the state vector. From $\mathbf{u}_t$ six surprise series can be extracted, corresponding to the dividend yield, the term spread, the default spread, the one-month T-bill yield, and the FF factors. They are denoted as follows: u^{DIV}, u^{TERM}, u^{DEF}, u^{RF}, u^{HML}, and u^{SMB}, respectively. This VAR represents a joint specification of the dynamics of all candidate state variables within the ICAPM. This specification treats the FF factors as potential candidates for state variables that command separate risk premia from the other variables.

The innovations derived from the VAR model are risk factors in addition to the excess return of the market portfolio. Asset's exposures to these risk factors are important determinants of average returns according to the ICAPM. To test the ICAPM specification, the Fama–MacBeth [15] cross-sectional method can be used as previously discussed. In the first pass of this method, a multiple time-series regression is specified which provides estimates of the assets' loadings with respect to the market return and the innovations in the state variables. More precisely, the following time-series regression is examined for each asset:

$$\begin{aligned} R_{i,t} = \alpha_i + \beta_{i,M} R_{M,t} &+ (\beta_{i,\hat{u}^{DIV}})\hat{u}_t^{DIV} + (\beta_{i,\hat{u}^{TERM}})\hat{u}_t^{TERM} \\ &+ (\beta_{i,\hat{u}^{DEF}})\hat{u}_t^{DEF} + (\beta_{i,\hat{u}^{RF}})\hat{u}_t^{RF} + (\beta_{i,\hat{u}^{HML}})\hat{u}_t^{HML} \\ &+ (\beta_{i,\hat{u}^{SMB}})\hat{u}_t^{SMB} + \varepsilon_{i,t}, \quad \text{for all} \quad i\,. \end{aligned} \tag{12}$$

The $\hat{u}$-terms represent the estimated surprises in the state variables. Note that the innovations terms are generated regressors and they appear on the right-hand side of the equation. However, as pointed out by Pagan [30], the OLS estimates of the parameters' standard errors will still be correct if the generated regressor represents the unanticipated part of a certain variable. On the other hand, if the $\hat{u}$-terms are only noisy proxies for the true surprises in the state variables, then the estimates of the factor loadings in the above regression will be biased downwards. This will likely bias the results against finding a relation between the innovations and asset returns.

The second step of the Fama–MacBeth procedure involves relating the average excess returns of all assets to their exposures to the risk factors in the model. Therefore, the following cross-sectional relation applies

$$\begin{aligned} \overline{R}_{i,t} = \gamma_0 + \gamma_M \widehat{\beta}_{i,M} &+ (\gamma_{\hat{u}^{DIV}})\widehat{\beta}_{i,\hat{u}^{DIV}} + (\gamma_{\hat{u}^{TERM}})\widehat{\beta}_{i,\hat{u}^{TERM}} \\ &+ (\gamma_{\hat{u}^{DEF}})\widehat{\beta}_{i,\hat{u}^{DEF}} + (\gamma_{\hat{u}^{RF}})\widehat{\beta}_{i,\hat{u}^{RF}} + (\gamma_{\hat{u}^{HML}})\widehat{\beta}_{i,\hat{u}^{HML}} \\ &+ (\gamma_{\hat{u}^{SMB}})\widehat{\beta}_{i,\hat{u}^{SMB}} + e_{i,t}, \quad \text{for all} \quad t\,. \end{aligned} \tag{13}$$

Data, Time-Series Analysis, and Results

In this section, monthly data for the period from July of 1963 to December of 2001 is used. The state variables in the context of the ICAPM are the dividend yield of the

[2] The following is only a partial list of papers that document time-variation in the excess market return and the variables they use: Campbell [2], term spread; Campbell and Shiller [4], dividend yield; Fama and Schwert [16], T-bill rate; Fama and French [10], default spread.

value-weighted market index (computed as the sum of dividends over the last 12 months, divided by the level of the index), the difference between the yield of a 10-year and a 1-year government bond (term spread), the difference between the yield of a long-term corporate Baa bond and a long-term government bond (default spread), and the one-month Treasury-bill yield. Data on bond yields is taken from the FRED® database of the Federal Reserve Bank of St. Louis. The T-bill yield and the term spread are used to measure the level and the slope of the yield curve, respectively.

VAR Estimation The state variables are the FF factors and the four predictive variables described above. All of them are included in a first-order VAR system. Campbell [3] emphasizes that it is hard to interpret estimation results for a VAR factor model unless the factors are orthogonalized and scaled in some way. In his paper the innovations to the state variables are orthogonal to the excess market return and to labor income. Following Campbell, the VAR system in Eq. (4) is triangularized in a similar way: the innovation in the excess market return is unaffected, the orthogonalized innovation in *DIV* is the component of the original *DIV* innovation orthogonal to the excess market return, and so on. The orthogonalized innovation to *DIV* is a change in the dividend/price ratio with no change in the market return, therefore it can be interpreted as a shock to the dividend. Similarly, shocks to the term spread, the default spread, the short-term rate, and the FF factors are orthogonal to the contemporaneous stock market return. As in Campbell [3], the innovations are scaled to have the same variance as the innovation in the excess market return.

It is interesting to note that the returns on the FF factors are very highly correlated with their respective innovation series. For example, the correlation between $R_{HML,t}$ and $\hat{u}_t^{HML}$ is 0.90, while the correlation between $R_{SMB,t}$ and $\hat{u}_t^{SMB}$ is 0.92. Therefore, the returns on the *HML* and *SMB* portfolios are good proxies for the innovations associated with those variables.

Relation Between R_{HML} and R_{SMB} and the VAR Innovations As a first step towards testing whether the FF factors proxy for innovations in state variables that track investment opportunities, the joint distribution of R_{HML} and R_{SMB} and innovations to *DIV*, *TERM*, *DEF*, and *RF* is examined. The following time-series regression is analyzed

$$\hat{u}_t = c_0 + c_1 R_{M,t} + c_2 R_{HML,t} + c_3 R_{SMB,t} + \varepsilon_t \qquad (14)$$

for each series of innovations in the state variables. The results for these regressions are presented in Table 3, with the corresponding t-statistics, below the coefficients, corrected for heteroscedasticity and autocorrelation. Innovations in the dividend yield, $\hat{u}_t^{DIV}$, covary negatively and significantly with the return on *HML*. In addition, $\hat{u}_t^{TERM}$ covaries positively and significantly with the *HML* return. These results are robust to the presence of the market factor in the regression. The return on the *HML* portfolio covaries positively and significantly with $\hat{u}_t^{DEF}$, while the return on the *SMB* factor covaries negatively with $\hat{u}_t^{DEF}$ (the corresponding t-statistic is marginally significant). The last regression in Table 3 indicates that the FF factors are not significant determinants of innovations in the T-bill yield. The results in the table remain unchanged if the independent variables in the equation above are the innovations to R_{HML} and R_{SMB} derived from the VAR system. The R-squares in the regressions reported in Table 3 are rather low. This does not imply, however, that the innovations in the state variables cannot potentially price assets as well as the FF factors. It could be the case that only the information in the FF factors correlated with the state variables is relevant for the pricing of risky assets. A similar point is made by Vassalou [34].

As pointed out by FF [10], the values of the term spread signal that expected market returns are low during expansions and high during recessions. In addition, FF document that the term spread very closely tracks the short-

Financial Economics, The Cross-Section of Stock Returns and the Fama-French Three Factor Model, Table 3

Time-Series Regressions Showing the Contemporaneous Relations Between Innovations in State Variables and the Fama–French Factors

This table presents time-series regressions of innovations in the dividend yield ($\hat{u}_t^{DIV}$), term spread ($\hat{u}_t^{TERM}$), default spread ($\hat{u}_t^{DEF}$), and one-month T-bill yield ($\hat{u}_t^{RF}$) on the excess market return, R_M, and the Fama–French factors R_{HML} and R_{SMB}. The innovations to the state variables are computed in a VAR system. The t-statistics are below the coefficients and are corrected for heteroscedasticity and autocorrelation using the Newey–West estimator with five lags. The Adjusted R^2 is reported in percentage form. The sample period is from July 1963 to December 2001

Regression: $\hat{u}_t = c_0 + c_1 R_{M,t} + c_2 R_{HML,t} + c_3 R_{SMB,t} + \varepsilon_t$					
Dep. Variable	c_0	c_1	c_2	c_3	Adj. R^2
$\hat{u}_t^{DIV}$	0.00	−0.08	−0.30	−0.01	3.00
	0.85	−0.70	−2.43	−0.09	
$\hat{u}_t^{TERM}$	−0.00	0.06	0.24	0.03	2.00
	−0.56	0.75	2.30	0.59	
$\hat{u}_t^{DEF}$	−0.00	0.07	0.17	−0.12	2.00
	−0.38	1.11	2.10	−1.92	
$\hat{u}_t^{RF}$	0.00	−0.04	−0.13	0.01	0.00
	0.36	−0.51	−1.36	0.14	

term fluctuations in the business cycle. Therefore, positive shocks to the term premium are associated with bad times in terms of business conditions, while negative shocks are associated with good times. In light of the results documented by Petkova and Zhang [32], that value stocks are riskier than growth stocks in bad times and less risky during good times, the relation between *HML* and shocks to the term spread seems natural.

Another interpretation of the relation between shocks to the term spread and the *HML* portfolio is in the context of cash flow maturities of assets. This point is discussed by Cornell [8] and Campbell and Vuolteenaho [5]. The argument is that growth stocks are high-duration assets, which makes them similar to long-term bonds and more sensitive to innovations in the long end of the term structure. Similarly, value stocks have lower duration than growth stocks, which makes them similar to short-term bonds and more sensitive to shocks to the short end of the yield curve.

Chan and Chen [6] have argued that small firms examined in the literature tend to be marginal firms, that is, they generally have lost market value due to poor performance, they are likely to have high financial leverage and cash flow problems, and they are less likely to survive poor economic conditions. In light of this argument, it is reasonable to assume that small firms will be more sensitive to news about the state of the business cycle. Therefore, it is puzzling that I find no significant relation between *SMB* and surprises to the term spread. Innovations in the term spread seem to be mostly related to *HML*. This observation suggests that the *HML* portfolio might represent risk related to cash flow maturity, captured by unexpected movements in the slope of the term structure.

Innovations in default spread, u_t^{DEF}, stand for changes in forecasts about expected market returns and changes in forecasts about default spread. FF [10] show that the default premium tracks time variation in expected returns that tends to persist beyond the short-term fluctuations in the business cycle. A possible explanation for the negative relation between *SMB* and shocks to the default spread could be that bigger stocks are able to track long-run trends in the business cycle better than the smaller stocks. The result that *HML* is also related to shocks in the default spread is consistent with the interpretation of *HML* as a measure of distress risk. The distress risk interpretation of the book-to-market effect is advocated by FF [11,12,13,14] and Chen and Zhang [7], among others.

In summary, the empirical literature has documented that both value and small stocks tend to be under distress, with high leverage and cash flow uncertainty. The results in this study suggest that the book-to-market factor might be related to asset duration risk, measured by the slope of the term structure, while the size factor might be related to asset distress risk, measured by the default premium.

It is reasonable to test whether the significant relation between the state variables surprises and the FF factors gives rise to the significant explanatory power of *HML* and *SMB* in the cross-section of returns. The next section examines whether *HML* and *SMB* remain significant risk factors in the presence of innovations to the other state variables. The results from the cross-sectional regressions suggest that *HML* and *SMB* lose their explanatory power for the cross-section of returns once accounting for the other variables. This supports an ICAPM explanation behind the empirical success of the FF three-factor model.

Cross-Sectional Regressions

Incremental Explanatory Power of the Fama–French Factors This section examines the pricing performance of the full set of state variables considered before over the period from July 1963 to December 2001. The full set of state variables consists of the dividend yield, the term spread, the default spread, the short-term T-bill yield, and the FF factors. The innovations to these state variables derived from a VAR system are risk factors in the ICAPM model. The objective is to test whether an asset's loadings with respect to these risk factors are important determinants of its average return.

The first specification is

$$\begin{aligned}\overline{R}_{i,t} = \gamma_0 &+ \gamma_{MKT}\widehat{\beta}_{i,MKT} + (\gamma_{\hat{u}^{DIV}})\widehat{\beta}_{i,\hat{u}^{DIV}} \\ &+ (\gamma_{\hat{u}^{TERM}})\widehat{\beta}_{i,\hat{u}^{TERM}} + (\gamma_{\hat{u}^{DEF}})\widehat{\beta}_{i,\hat{u}^{DEF}} \\ &+ (\gamma_{\hat{u}^{RF}})\widehat{\beta}_{i,\hat{u}^{RF}} + (\gamma_{\hat{u}^{HML}})\widehat{\beta}_{i,\hat{u}^{HML}} \\ &+ (\gamma_{\hat{u}^{SMB}})\widehat{\beta}_{i,\hat{u}^{SMB}} + e_{i,t}\,, \end{aligned} \tag{15}$$

where the $\widehat{\beta}$ terms stand for exposures to the corresponding factor, while the γ terms stand for the reward for bearing the risk of that factor. The $\widehat{\beta}$ terms are the independent variables in the regression, while the average excess returns of the assets are the dependent variables. If loadings with respect to innovations in a state variable are important determinants of average returns, then there should be a significant price of risk associated with that state variable.

The results are reported in Table 4. The table shows that assets' exposures to innovations in R_{HML} and R_{SMB} are not significant variables in the cross-section in the presence of betas with respect to surprises in the other state variables. The corresponding t-statistics are 1.40 and 1.56, respectively, under the errors-in-variables correction. Therefore, based on the results presented in Table 4, the hypothesis that innovations in the dividend yield, the

Financial Economics, The Cross-Section of Stock Returns and the Fama-French Three Factor Model, Table 4

Cross-Sectional Regressions Showing the Incremental Explanatory Power of the Fama–French Factor Loadings

This table presents Fama–MacBeth cross-sectional regressions using the average excess returns on 25 portfolios sorted by book-to-market and size. The full-sample factor loadings, which are the independent variables in the regressions, are computed in one multiple time-series regression. The coefficients are expressed as percentage per month. The table presents results for the model including the excess market return, R_M, and innovations in the dividend yield, term spread, default spread, one-month T-bill yield, and the Fama–French factors *HML* and *SMB*. The Adjusted R^2 follows Jagannathan and Wang [18] and is reported in percentage form. The first set of *t*-statistics, indicated by FM *t*-stat, stands for the Fama–MacBeth estimate. The second set, indicated by SH *t*-stat, adjusts for errors-in-variables and follows Shanken [33]. The table examines the sample period from July 1963 to December 2001

The Model with Innovations in All State Variables									
	γ_0	γ_M	$\gamma_{\hat{u}^{DIV}}$	$\gamma_{\hat{u}^{TERM}}$	$\gamma_{\hat{u}^{DEF}}$	$\gamma_{\hat{u}^{RF}}$	$\gamma_{\hat{u}^{HML}}$	$\gamma_{\hat{u}^{SMB}}$	Adj. R^2
Estimate	1.11	−0.57	−0.83	3.87	0.37	−2.90	0.42	0.41	77.26
FM *t*-stat	3.29	−1.45	−0.94	3.53	0.42	−3.33	1.62	1.75	
SH *t*-stat	2.36	−1.10	−0.69	2.56	0.31	−2.44	1.40	1.56	

term spread, the default spread, and the short-term T-bill span the information contained in the FF factors cannot be rejected.

A Model Based on R_M, and Innovations in *DIV*, *TERM*, *DEF*, and *RF* This part examines separately the set of innovations in the variables associated with time-series predictability: the dividend yield, the term spread, the default spread, and the short-term T-bill. The model specification is as follows

$$\begin{aligned} R_{i,t} = \alpha_i + \beta_{i,M} R_{M,t} + (\beta_{i,\hat{u}^{DIV}})\hat{u}_t^{DIV} \\ + (\beta_{i,\hat{u}^{TERM}})\hat{u}_t^{TERM} + (\beta_{i,\hat{u}^{DEF}})\hat{u}_t^{DEF} \\ + (\beta_{i,\hat{u}^{RF}})\hat{u}_t^{RF} + \varepsilon_{i,t}, \quad \text{for all} \quad i \end{aligned} \tag{16}$$

$$\begin{aligned} \overline{R}_{i,t} = \gamma_0 + \gamma_M \widehat{\beta}_{i,M} + (\gamma_{\hat{u}^{DIV}})\widehat{\beta}_{i,\hat{u}^{DIV}} + (\gamma_{\hat{u}^{TERM}})\widehat{\beta}_{i,\hat{u}^{TERM}} \\ + (\gamma_{\hat{u}^{DEF}})\widehat{\beta}_{i,\hat{u}^{DEF}} + (\gamma_{\hat{u}^{RF}})\widehat{\beta}_{i,\hat{u}^{RF}} + e, \quad \text{for all} \quad t \end{aligned} \tag{17}$$

which corresponds to a model in which the relevant risk factors are innovations to predictive variables. The objective is to compare the pricing performance of this model with that of the Fama–French model for the cross-section of returns sorted by book-to-market and size. The specification is motivated by the previous observation that *HML* and *SMB* do not add explanatory power to the set of state variables that are associated with time-series predictability.

Table 5 report the estimates of the factor loadings computed in the first-pass time-series regressions defined in Eq. (16). It also presents joint tests of the significance of the corresponding loadings, computed from a SUR system. This is done in order to show that the innovations factors are relevant in the sense that the 25 portfolios load significantly on them. A similar analysis was performed on the Fama–French model in Sect. "The Fama–French Model as a Linear Beta Pricing Model".

An *F*-test implies that the 25 loadings on innovations to the term spread are jointly significant, with the corresponding *p*-value being 0.47%. Furthermore, portfolios' loadings on $\hat{u}_t^{TERM}$ are related to book-to-market: within each size quintile, the loadings increase monotonically from lower to higher book-to-market quintiles. In fact, the portfolios within the lowest book-to-market quintile have negative sensitivities with respect to $\hat{u}_t^{TERM}$, while the portfolios within the highest book-to-market quintile have positive loadings on $\hat{u}_t^{TERM}$. This pattern resembles very much the one observed in Table 1 for the loadings on R_{HML}.

Similarly, loadings on shocks to default spread are jointly significant in Table 5, with the corresponding *p*-value being 0.24%. Moreover, the slopes on $\hat{u}_t^{DEF}$ are systematically related to size. Within each book-to-market quintile, the loadings increase almost monotonically from negative values for the smaller size quintiles to positive values for the larger size quintiles. This pattern closely resembles the mirror image of the one observed in Table 1 for the loadings on R_{SMB}. The slopes on dividend yield and T-bill innovations do not exhibit any systematic patterns related to size or book-to-market. However, both of these are jointly significant.

Note that the R^2s in the time-series regressions with the innovations factors in Table 5 are smaller than the ones in the regressions with the FF factors in Table 1. This indicates that potential errors-in-variables problems that arise in measuring the factor loadings will be more serious in the case of the innovations terms. Therefore, the results will be potentially biased against finding significant factor

Financial Economics, The Cross-Section of Stock Returns and the Fama-French Three Factor Model, Table 5

Loadings on R_M, $\hat{u}_t^{DIV}$, $\hat{u}_t^{TERM}$, $\hat{u}_t^{DEF}$, and $\hat{u}_t^{RF}$ from Time-Series Regressions

This table reports loadings on the excess market return, R_M, and innovations in the dividend yield ($\hat{u}_t^{DIV}$), term spread ($\hat{u}_t^{TERM}$), default spread ($\hat{u}_t^{DEF}$), and short-term T-bill ($\hat{u}_t^{RF}$) computed in time-series regressions for 25 portfolios sorted by size and book-to-market. The corresponding t-statistics are also reported and are corrected for autocorrelation and heteroscedasticity using the Newey–West estimator with five lags. The sample period is from July 1963 to December 2001. The last column reports F-statistics and their corresponding p-values from an SUR system, testing the joint significance of the corresponding loadings. The p-values are in percentage form. R^2s from each time-series regression are reported in percentage form

Regression: $R_{i,t} = \alpha_i + \beta_{i,M} R_{M,t} + \beta_{i,\hat{u}^{DIV}} \hat{u}_t^{DIV} + \beta_{i,\hat{u}^{TERM}} \hat{u}_t^{TERM} + \beta_{i,\hat{u}^{DEF}} \hat{u}_t^{DEF} + \beta_{i,\hat{u}^{RF}} \hat{u}_t^{RF} + \varepsilon_{i,t}$											
	Low	2	3	4	High		Low	2	3	4	High	
	β_{MKT}						$t_{\beta_{MKT}}$					F
Small	1.44	1.23	1.09	1.01	1.02		24.20	22.74	20.76	19.57	18.87	> 100
2	1.44	1.18	1.04	0.98	1.05		31.33	25.11	22.63	21.90	18.76	< 0.01
3	1.38	1.12	0.98	0.90	0.98		39.96	32.34	22.52	21.58	17.66	
4	1.27	1.08	0.97	0.90	0.99		45.46	29.07	24.02	23.95	19.60	
Large	1.01	0.95	0.85	0.78	0.78		42.69	36.55	26.89	20.47	15.34	
	$\beta_{\hat{u}^{DIV}}$						$t_{\beta_{\hat{u}^{DIV}}}$					F
Small	4.75	0.43	−5.02	−5.61	−7.88		0.76	0.08	−0.89	−1.10	−1.44	2.33
2	3.38	−4.01	−7.66	−6.76	−6.51		0.76	−0.79	−1.55	−1.35	−1.09	0.02
3	7.45	−1.30	−5.91	−8.27	−9.18		2.34	−0.35	−1.16	−1.53	−1.36	
4	8.65	−5.83	−6.17	−8.18	−11.81		2.90	−1.29	−1.21	−1.72	−2.04	
Large	−0.78	−3.49	−1.73	−9.69	−9.50		−0.29	−1.18	−0.47	−1.83	−1.49	
	$\beta_{\hat{u}^{TERM}}$						$t_{\beta_{\hat{u}^{TERM}}}$					F
Small	1.51	1.04	1.69	2.82	8.68		0.26	0.26	0.47	0.79	2.24	1.89
2	−8.21	−2.73	−0.19	1.36	5.16		−1.87	−0.75	0.06	0.46	1.44	0.47
3	−6.34	−3.52	−1.72	2.08	4.39		−1.77	−1.17	−0.55	0.55	1.18	
4	−0.73	−1.51	0.21	0.02	2.13		−0.26	−0.59	0.06	0.01	0.54	
Large	−5.98	−3.26	0.78	−0.90	2.90		−2.22	−1.37	0.31	−0.26	0.74	
	$\beta_{\hat{u}^{DEF}}$						$t_{\beta_{\hat{u}^{DEF}}}$					F
Small	−15.45	−14.54	−6.86	−4.79	−8.58		−2.27	−2.17	−1.39	−1.09	−1.68	1.99
2	−10.03	−5.90	−4.78	0.82	−2.20		−2.04	−1.62	−1.37	0.22	−0.49	0.24
3	−11.17	0.22	1.73	4.03	0.81		−2.75	0.08	0.49	1.15	0.18	
4	−5.80	4.81	4.80	8.03	1.08		−2.10	1.92	1.44	2.50	0.25	
Large	−2.45	3.99	9.12	7.25	2.56		−0.96	1.91	3.85	1.91	0.63	
	$\beta_{\hat{u}^{RF}}$						$t_{\beta_{\hat{u}^{RF}}}$					F
Small	4.07	−2.58	0.07	1.03	2.77		0.77	−0.50	0.01	0.22	0.56	1.76
2	−4.37	−5.20	−6.25	−4.57	0.97		−1.00	−1.19	−1.60	−1.16	0.20	1.08
3	−7.63	−4.40	−6.53	−4.08	0.71		−2.29	−1.38	−2.07	−1.09	0.15	
4	−3.43	0.47	−2.04	−5.74	−3.71		−1.12	0.16	−0.69	−1.61	−0.90	
Large	−3.55	−0.59	4.81	−0.89	0.30		−1.14	−0.22	1.41	−0.25	0.06	
				R^2								
				61.51	60.92	63.41	62.41	59.93				
				73.93	73.95	74.47	71.88	67.96				
				79.81	81.80	77.54	73.58	68.96				
				84.99	86.05	80.32	77.51	69.42				
				87.65	86.11	77.89	67.67	55.88				

loadings on the shocks to the predictive variables. Kan and Zhang [20] emphasize that checking the joint significance of the assets' factor loadings is an important step in detecting useless factors in the cross-section of returns.

Table 6 contains the results for Eq. (17) which correspond to the second pass of the Fama–MacBeth method. The results reveal that the explanatory power of this model is very close to the one for the Fama–French model re-

Financial Economics, The Cross-Section of Stock Returns and the Fama-French Three Factor Model, Table 6

Cross-Sectional Regressions with Loadings on Innovations in State Variables

This table presents Fama–MacBeth cross-sectional regressions using the average excess returns on 25 portfolios sorted by book-to-market and size. The full-sample factor loadings, which are the independent variables in the regressions, are computed in one multiple time-series regression. The coefficients are expressed as percentage per month. The Adjusted R^2 follows Jagannathan and Wang [18] and is reported in percentage form. The first set of *t*-statistics, indicated by FM *t*-stat, stands for the Fama–MacBeth estimate. The second set, indicated by SH *t*-stat, adjusts for errors-in-variables and follows Shanken [33]. The sample period is from July 1963 to December 2001

The Model with R_M and Innovations in *DIV*, *TERM*, *DEF*, and *RF*							
	γ_0	γ_M	$\gamma_{\hat{u}^{DIV}}$	$\gamma_{\hat{u}^{TERM}}$	$\gamma_{\hat{u}^{DEF}}$	$\gamma_{\hat{u}^{RF}}$	Adj. R^2
Estimate	0.64	–0.07	–1.39	4.89	–0.54	–3.22	77.00
FM *t*-stat	1.74	–0.16	–1.56	4.44	–0.58	–3.79	
SH *t*-stat	1.08	–0.11	–0.99	2.79	–0.37	–2.40	

ported previously in Table 2. Figure 2 plots the fitted versus the realized average returns from the model. It can be seen form the graph that the model based on innovation in predictive variables goes a long way toward explaining the value effect: in general, the fitted expected returns on value portfolios (bigger second digit) are higher than the fitted expected returns on growth portfolios (lower second digit). This is consistent with the data on realized average returns for these portfolios. Further, the model with R_M, $\hat{u}^{DIV}$, $\hat{u}^{TERM}$, $\hat{u}^{DEF}$, and $\hat{u}^{RF}$ is more successful at pricing the portfolios that are challenging for the Fama–French model. The realized returns on growth portfolios within the smallest and largest size groups and the value portfolios within the largest size groups are brought closer to the 45-degree line under the model with the four innovations factors.

In summary, this section has shown that the performance of the model based on innovation in predictive variables is very close to the performance of the Fama–French model in the cross-section of average returns sorted by size and book-to-market. This suggest that the Fama–French factors *HML* and *SMB* might proxy for fundamental state variables that describe variation in investment opportunities over time.

Other Risk-Based Interpretations

Liew and Vassalou [25] show that there is a relation between the Fama–French portfolios *HML* and *SMB* and macroeconomic events. They find that not only in the US but also in several other countries, the corresponding

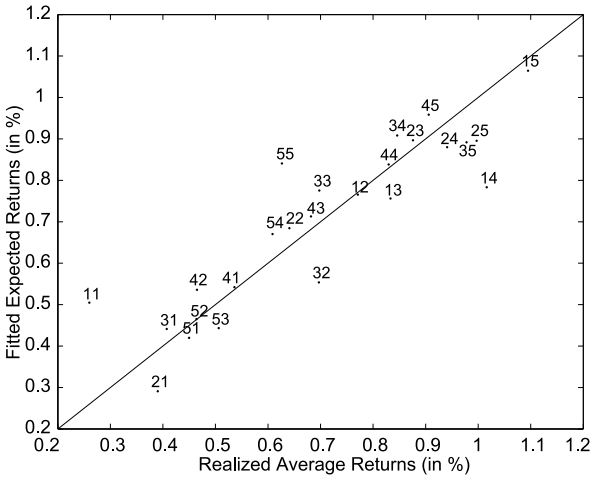

Financial Economics, The Cross-Section of Stock Returns and the Fama-French Three Factor Model, Figure 2

Fitted Expected Returns vs. Average Realized Returns for 1963:07-2001:12.

This figure shows realized average returns (%) on the horizontal axis and fitted expected returns (%) on the vertical axis for 25 size and book-to-market sorted portfolios. Each two-digit number represents a separate portfolio. The first digit refers to the size quintile (1 being the smallest and 5 the largest), while the second digit refers to the book-to-market quintile (1 being the lowest and 5 the highest). For each portfolio, the realized average return is the time-series average of the portfolio return and the fitted expected return is the fitted value for the expected return from the corresponding model. The straight line is the 45-degree line from the origin. The Model with the Excess Market Return and Innovations in the Dividend Yield, Term Spread, Default Spread, and Short-Term T-bill

HML and *SMB* portfolios contain information about future GDP growth. Therefore, the authors conclude that the size and book-to-market factors are related to future macroeconomic growth. This evidence is consistent with interpreting the *HML* and *SMB* factors as proxies for business cycle risk.

Other studies try to relate the difference in average returns between value and growth portfolios to the time-varying nature of the riskiness of those portfolios. Namely, if value stocks are riskier than growth stocks during bad economic times and if the price of bearing risk is higher during those times, then it follows that value stocks should earn higher average returns than growth stocks. Lettau and Ludvigson [24] document that *HML* is indeed sensitive to bad news in bad macroeconomic times.

Petkova and Zhang [32] is another study that looks are the time-varying risk of value and growth portfolios. They find that the market risk of value stocks is high in bad times when the expected premium for risk is high and it is low

in good times when the expected premium for risk is low. What might lead to this time-varying of value and growth stocks? Zhang [35] suggest that the reason might be irreversible investment. He notes that firms with high book-to-market ratios on average will have larger amounts of tangible capital. In addition, it is more costly for firms to reduce than to expand capital. In bad times, firms want to scale down, especially value firms that are less productive than growth firms (Fama and French [13]). Because scaling down is more difficult, value firms are more adversely affected by economic downturns. In good times, growth firms face less flexibility because they tend to invest more. Expanding is less urgent for value firms because their previously unproductive assets have become more productive. In sum, costly reversibility causes value firms to have higher (lower) betas than growth firms in bad (good) times and this contributes to the return differential between these two classes of stocks.

Future Directions

The Fama–French model states that asset returns are driven by three market-wide factors: the excess return on the market portfolio, and the returns on two portfolios related to size (*SMB*) and book-to-market (*HML*). The *HML* and *SMB* portfolios capture the empirical observation that value firms earn higher average returns than growth firms, and small firms earn higher average returns than large firms. The Fama–French model has been very successful at explaining average stock returns, but the exact economic interpretation of the *HML* and *SMB* portfolios has been an issue of debate.

This article examines the risk-based explanation behind the empirical success of the Fama–French model and suggests that the value and size premia arise due to differences in exposure to systematic sources of risk. As mentioned in the introduction, several authors (e. g., Lakonishok, Shleifer, Vishny [22], La Porta, Lakonishok, Shleifer, Vishny [23]), however, claim that the value premium results from irrationality on the side of investors. Namely, investors tend to over-extrapolate recent stock performance: they overvalue the stocks of growth firms and undervalue the stocks of value firms. When the market realizes its mistake, the prices of the former fall, while the prices of the latter rise, resulting in the value premium.

The Fama–French model provides a useful performance benchmark relative to a set of market-wide factors. The results in this article suggest that the Fama–French factors proxy for systematic sources of risk that capture time variation in investment opportunities. However, the debate about the economic interpretation behind the size and value premia is still not settled. Whether they arise as a result of rational compensation for risk or irrational investor behavior is still a matter of controversy.

Bibliography

1. Berk J, Green R, Naik V (1999) Optimal investment, growth options, and security returns. J Finance 54:1553–1608
2. Campbell J (1987) Stock returns and the term structure. J Financial Econ 18:373–399
3. Campbell J (1996) Understanding risk and return. J Political Econ 104:298–345
4. Campbell J, Shiller R (1988) Stock prices, earnings, and expected dividends, J Finance 43:661–676
5. Campbell J, Vuolteenaho T (2004) Bad beta, good beta. Am Econ Rev 5:1249–1275
6. Chan KC, Chen N (1991) Structural and return characteristics of small and large firms. J Finance 46:1467–1484
7. Chen N, Zhang F (1998) Risk and return of value stocks. J Bus 71:501–535
8. Cornell B (1999) Risk, duration, and capital budgeting: New evidence on some old questions. J Bus 72:183–200
9. Daniel K, Titman S (1997) Evidence on the characteristics of cross-sectional variation in stock returns. J Finance 52:1–33
10. Fama E, French K (1989) Business conditions and expected returns on stocks and bonds. J Financial Econ 25:23–49
11. Fama E, French K (1992) The cross-section of expected stock returns. J Finance 47:427–465
12. Fama E, French K (1993) Common risk factors in the returns on bonds and stocks. J Financial Econ 33:3–56
13. Fama E, French K (1995) Size and book-to-market factors in earnings and returns. J Finance 50:131–155
14. Fama E, French K (1996) Multifactor explanations of asset pricing anomalies. J Finance 51:55–84
15. Fama E, MacBeth J (1973) Risk, return, and equilibrium: Empirical tests. J Political Econ 81:607–636
16. Fama E, Schwert W (1977) Asset returns and inflation. J Financial Econ 5:115–146
17. Gomes J, Kogan L, Zhang L (2003) Equilibrium cross-section of returns. J Political Econ 111:693–732
18. Jagannathan R, Wang Z (1996) The conditional CAPM and the cross-section of expected returns. J Finance 51:3–53
19. Jagannathan R, Wang Z (1998) Asymptotic theory for estimating beta pricing models using cross-sectional regressions. J Finance 53:1285–1309
20. Kan R, Zhang C (1999) Two-pass tests of asset pricing models with useless factors. J Finance 54:203–235
21. Kothari SP, Shanken J, Sloan R (1995) Another look at the cross-section of expected returns. J Finance 50:185–224
22. Lakonishok J, Shleifer A, Vishny R (1994) Contrarian investment, extrapolation, and risk. J Finance 49:1541–1578
23. La Porta R, Lakonishok J, Shleifer A, Vishny R (1997) Good news for value stocks: Further evidence on market efficiency. J Finance 52:859–874
24. Lettau M, Ludvigson S (2001) Resurrecting the (C)CAPM: A cross-sectional test when risk premia are time-varying. J Political Econ 109:1238–1287
25. Liew J, Vassalou M (2000) Can book-to-market, size, and momentum be risk factors that predict economic growth? J Financial Econ 57:221–245

26. Litterman R, Sheinkman J (1991) Common factors affecting bond returns. J Fixed Income 1:54–61
27. Lo A, MacKinlay C (1990) Data-snooping biases in tests of financial asset pricing models. Rev Financial Stud 3:431–467
28. Long J (1974) Stock prices, inflation, and the term structure of interest rates. J Financial Econ 1:131–170
29. Merton R (1973) An intertemporal capital asset-pricing model. Econometrica 41:867–887
30. Pagan A (1984) Econometric issues in the analysis of regressions with generated regressors. Int Econ Rev 25:221–247
31. Petkova R (2006) Do the Fama–French factors proxy for innovations in predictive variables? J Finance 61:581–612
32. Petkova R, Zhang L (2005) Is value riskier than growth? J Financial Econ 78:187–202
33. Shanken J (1992) On the estimation of beta-pricing models. Rev Financial Stud 5:1–34
34. Vassalou M (2003) News related to future GDP growth as a risk factor in equity returns. J Financial Econ 68:47–73
35. Zhang L (2005) The value premium. J Finance 60:67–103

Financial Economics, Fat-Tailed Distributions

Markus Haas[1], Christian Pigorsch[2]
[1] Department of Statistics, University of Munich, Munich, Germany
[2] Department of Economics, University of Bonn, Bonn, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Defining Fat-Tailedness
Empirical Evidence About the Tails
Some Specific Distributions
Volatility Clustering and Fat Tails
Application to Value-at-Risk
Future Directions
Bibliography

Glossary

Leptokurtosis A distribution is leptokurtic if it is more peaked in the center and thicker tailed than the normal distribution with the same mean and variance. Occasionally, leptokurtosis is also identified with a moment-based kurtosis measure larger than three, see Sect. "Introduction".

Return Let S_t be the price of a financial asset at time t. Then the *continuous* return, r_t, is $r_t = \log(S_t/S_{t-1})$. The *discrete* return, R_t, is $R_t = S_t/S_{t-1} - 1$. Both are rather similar if $-0.15 < R_t < 0.15$, because $r_t = \log(1 + R_t)$. See Sect. "Introduction".

Tail The (upper) tail, denoted by $\bar{F}(x) = P(X > x)$, characterizes the probability that a random variable X exceeds a certain "large" threshold x. For analytical purposes, "large" is often translated with "as $x \to \infty$". For financial returns, a daily change of 5% is already infinitely large. A Gaussian model essentially excludes such an event.

Tail index The tail index, or *tail exponent*, α, characterizes the rate of tail decay if the tail goes to zero, in essence, like a power function, i. e., $\bar{F}(x) = x^{-\alpha} L(x)$, where L is slowly varying. Moments of order lower (higher) than α are (in)finite.

Definition of the Subject

Have a look at Fig. 1. The top plot shows the daily percentage changes, or *returns*, of the S&P500 index ranging from January 2, 1985 to December 29, 2006, a total of 5,550 daily observations. We will use this data set throughout the article to illustrate some of the concepts and models to be discussed. Two observations are immediate. The first is that both small and large changes come clustered, i. e., there are periods of low and high volatility. The second is that, from time to time, we observe rather large changes which may be hard to reconcile with *the* standard distributional assumption in statistics and econometrics, that is, normality. The most outstanding return certainly occurred on October 19, 1987, the "Black Monday", where the index lost more than 20% of its value, but the phenomenon is chronic. For example, if we fit a normal distribution to the data, the resulting model predicts that we observe an absolute daily change larger than 5% once in approximately 1,860 years, whereas we actually encountered that 13 times during our 22-year sample period. This suggests that, compared to the normal distribution, the distribution of the returns is *fat-tailed*, i. e., the probability of large losses and gains is much higher than would be implied by a time-invariant unconditional Gaussian distribution. The latter is obviously not suitable for describing the booms, busts, bubbles, and bursts of activity which characterize financial markets, and which are apparent in Fig. 1.

The two aforementioned phenomena, i. e., volatility clustering and fat tails, have been detected in almost every financial return series that was subject to statistical analysis since the publication of Mandelbrot's [155] seminal study of cotton price changes, and they are of paramount importance for any individual or institution engaging in the financial markets, as well as for financial economists trying to understand their mode of operation. For exam-

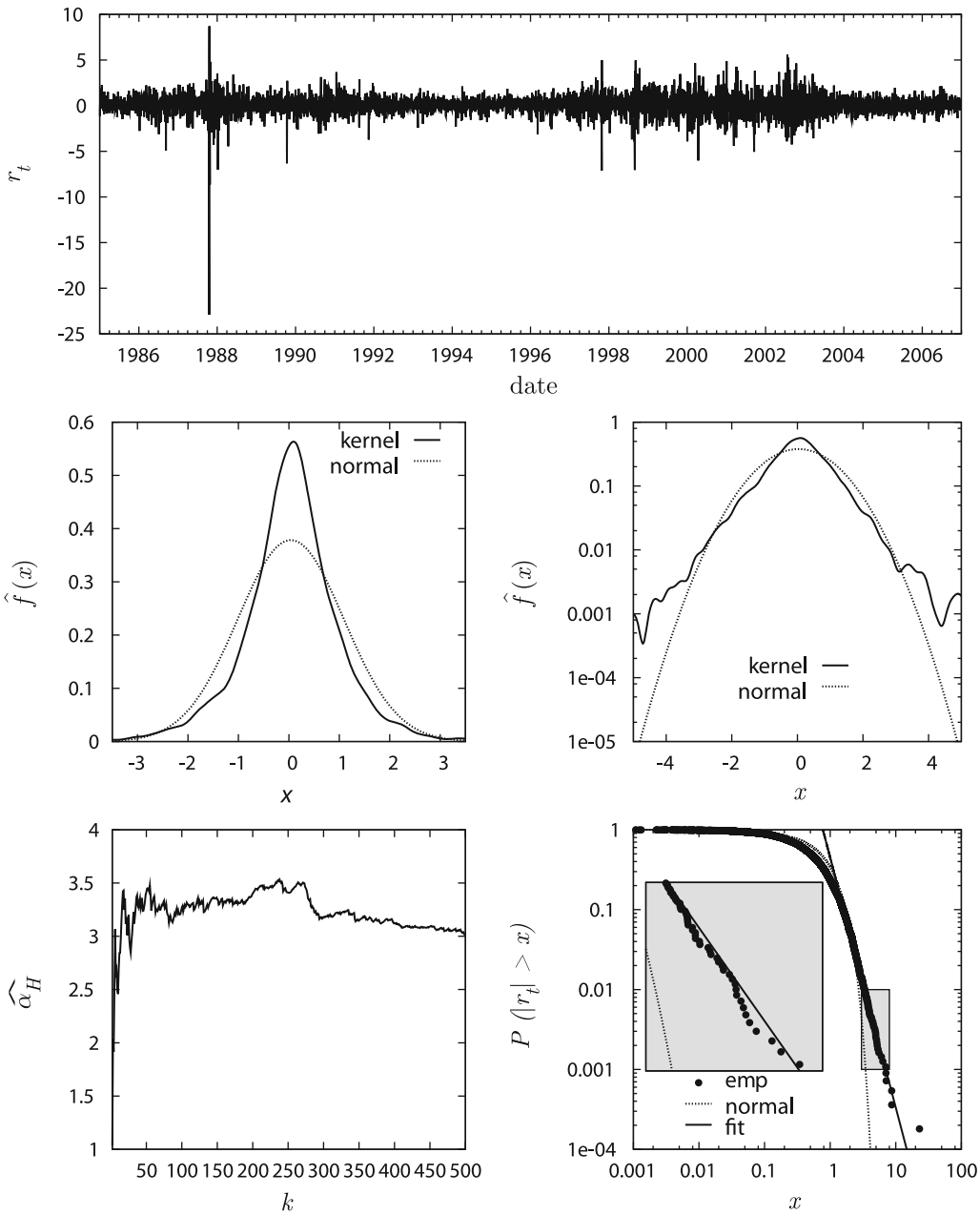

Financial Economics, Fat-Tailed Distributions, Figure 1

The *top plot* shows the S&P500 *percentage* returns, r_t, from January 1985 to December 2006, i. e., $r_t = 100 \times \log(S_t/S_{t-1})$, where S_t is the index level at time t. The left plot of the *middle panel* shows a nonparametric density estimate (*solid*), along with the fitted normal density (*dotted*); the right graph is similar but shows the respective log-densities in order to better visualize the tail regions. The *bottom left plot* represents a *Hill plot* for the S&P500 returns, i. e., it displays $\hat{\alpha}_{k,n}$ defined in (11) for $k \leq 500$. The *bottom right plot* shows the complementary cdf, $\bar{F}(x)$, on a log-log scale, see Sect. "Empirical Evidence About the Tails" for discussion

ple, investors holding significant portions of their wealth in risky assets need a realistic assessment of the likelihood of severe losses. Similarly, economists trying to learn about the relation between risk and return, the pricing of financial derivatives, such as options, and the inherent dynamics of financial markets, can only benefit from building their models on adequate assumptions about the stochastic properties of the variables under study, and they have

to reconcile the predictions of their models with the actual facts.

This article reviews some of the most important concepts and distributional models that are used in empirical finance to capture the (almost) ubiquitous stochastic properties of returns as indicated above. Section "Introduction" defines in a somewhat more precise manner than above the central variable of interest, the return of a financial asset, and gives a brief account of the early history of the problem. Section "Defining Fat-Tailedness" discusses various operationalizations of the term "fat-tailedness", and Sect. "Empirical Evidence About the Tails" summarizes what is or is at least widely believed to be known about the tail characteristics of typical return distributions. Popular parametric distributional models are discussed in Sect. "Some Specific Distributions". The alpha stable model as the archetype of a fat-tailed distribution in finance is considered in detail, as is the generalized hyperbolic distribution, which provides a convenient framework for discussing, as special or limiting cases, many of the important distributions employed in the literature. An empirical comparison using the S&P500 returns is also included. In Sect. "Volatility Clustering and Fat Tails", the relation between the two "stylized facts" mentioned above, i. e., clusters of volatility and fatness of the tails, is highlighted, where we concentrate on the GARCH approach, which has gained outstanding popularity among financial econometricians. This model has the intriguing property of producing fat-tailed marginal distributions even with light-tailed innovation processes, thus emphasizing the role of the market dynamics. In Sect. "Application to Value-at-Risk", we compare both the unconditional parametric distributional models introduced in Sect. "Some Specific Distributions" as well as the GARCH model of Sect. "Volatility Clustering and Fat Tails" on an economic basis by evaluating their ability to accurately measure the Value-at-Risk, which is an important tool in risk management. Finally, Sect. "Future Directions" identifies some open issues.

Introduction

To fix notation, let S_t be the price of an asset at time t, e. g., a stock, a market index, or an exchange rate. The *continuously compounded* or *log* return from time t to time $t+\Delta t$, $r_{t,t+\Delta t}$, is then defined as

$$r_{t,t+\Delta t} = \log S_{t+\Delta t} - \log S_t \,. \tag{1}$$

Often the quantity defined in (1) is also multiplied by 100, so that it can be interpreted in terms of *percentage returns*, see Fig. 1. Moreover, in applications, Δt is usually set equal to one and represents the horizon over which the returns are calculated, e. g., a day, week, or month. In this case, we drop the first subscript and define $r_t := \log S_t - \log S_{t-1}$. The log returns (1) can be additively aggregated over time, i. e.,

$$r_{t,t+\tau} = \sum_{i=1}^{\tau} r_{t+i} \,. \tag{2}$$

Empirical work on the distribution of financial returns is usually based on log returns. In some applications a useful fact is that, over short intervals of time, when returns tend to be small, (1) can also serve as a reasonable approximation to the *discrete* return, $R_{t,t+\Delta t} := S_{t+\Delta t}/S_t - 1 = \exp(r_{t,t+\Delta t}) - 1$. For further discussion of the relationship between continuous and discrete returns and their respective advantages and disadvantages, see, e. g., [46,76].

The seminal work of Mandelbrot [155], to be discussed in Subsect. "Alpha Stable and Related Distributions", is often viewed as the beginning of modern empirical finance. As reported in [74], "[p]rior to the work of Mandelbrot the usual assumption … was that the distribution of price changes in a speculative series is approximately Gaussian or normal". The rationale behind this prevalent view, which was explicitly put forward as early as 1900 by Bachelier [14], was clearly set out in [178]: If the log-price changes (1) from transaction to transaction are independently and identically distributed with finite variance, and if the number of transactions is fairly uniformly distributed in time, then (2) along with the central limit theorem (CLT) implies that the return distribution over longer intervals, such as a day, a week, or a month, approaches a Gaussian shape.

However, it is now generally acknowledged that the distribution of financial returns over horizons shorter than a month is not well described by a normal distribution. In particular, the empirical return distributions, while unimodal and approximately symmetric, are typically found to exhibit considerable *leptokurtosis*, i. e., they are more peaked in the center and have fatter tails than the Gaussian with the same variance. Although this has been occasionally observed in the pre-Mandelbrot literature (e. g., [6]), the first systematic account of this phenomenon appeared in [155] and the follow-up papers by Fama [74,75] and Mandelbrot [156], and it was consistently confirmed since then. The typical shape of the return distribution, as compared to a fitted Gaussian, is illustrated in the middle panel of Fig. 1 for the S&P500 index returns, where a nonparametric kernel density estimator (e. g., [198]) is superimposed on the fitted Gaussian curve (dashed line). Interestingly, this pattern has been detected not only for modern

financial markets but also for those of the eighteenth century [103].

The (location and scale-free) standardized fourth moment, or coefficient of *kurtosis*,

$$\mathbb{K}[X] = \frac{\mathbb{E}\left[(X-\mu)^4\right]}{\sigma^4}\,, \tag{3}$$

where μ and σ are the mean and the standard deviation of the random variable (rv) X, respectively, is sometimes used to assess the degree of leptokurtosis of a given distribution. For the normal distribution, $\mathbb{K} = 3$, and $\mathbb{K} > 3$, referred to as *excess kurtosis*, is taken as an indicator of a leptokurtic shape (e. g., [164], p. 429). For example, the sample analogue of (3) for the S&P500 returns shown in Fig. 1 is 47.9, indicating very strong excess kurtosis. A formal test could be conducted using the fact that, under normality, the sample kurtosis is asymptotically normal with mean 3 and standard deviation $\sqrt{24/T}$ (T being the sample size), but the result can be anticipated.

As is well-known, however, such moment-based summary measures have to be interpreted with care, because a particular moment need not be very informative about a density's shape. We know from Finucan [82] that if two symmetric densities, f and g, have common mean and variance and finite fourth moment, and if g is more peaked and has thicker tails than f, *then* the fourth moment (and hence $\mathbb{K}$) is greater for g than for f, provided the densities cross exactly twice on both sides of the mean. However, the converse of this statement is, of course, not true, and a couple of (mostly somewhat artificial) counterexamples can be found in [16,68,121]. [158] provides some intuition by relating density crossings to moment crossings. For example, (only) if the densities cross more than four times, it may happen that the fourth moment is greater for f, but the sixth and all higher moments are greater for g, reflecting the thicker tails of the latter. Nevertheless, Finucan's result, along with his (in some respects justified) hope that we can view "this pattern as the common explanation of a high observed kurtosis", may serve to argue for a certain degree of usefulness of the kurtosis measure (3), provided the fourth moment is assumed to be finite. However, a nonparametric density estimate will in any case be more informative. Note that the density crossing condition in Finucan's theorem is satisfied for the S&P500 returns in Fig. 1.

Defining Fat-Tailedness

The notion of leptokurtosis as discussed so far is rather vague, and both financial market researchers as well as practitioners, such as risk managers, are interested in a more precise description of the tail behavior of financial variables, i. e., the laws governing the probability of large gains and losses. To this end, we define the *upper tail* of the distribution of a rv X as

$$\bar{F}(x) = P(X > x) = 1 - F(x)\,, \tag{4}$$

where F is the cumulative distribution function (cdf) of X. Consideration of the upper tail is the standard convention in the literature, but it is clear that everything could be phrased just as well in terms of the lower tail.

We are interested in the behavior of (4) as x becomes large. For our benchmark, i. e., the normal distribution with (standardized) density (pdf) $\phi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$, we have (cf. p. 131 in [79])

$$\bar{F}(x) \cong \frac{1}{\sqrt{2\pi}\,x}\exp\left(-\frac{x^2}{2}\right) = \frac{\phi(x)}{x} \quad \text{as } x \to \infty\,, \tag{5}$$

where the notation $f(x) \cong g(x)$ as $x \to \infty$ means that $\lim_{x\to\infty} f(x)/g(x) = 1$. Thus, the tails of the normal tend to zero faster than exponentially, establishing its very light tails.

To appreciate the difference between the general concept of leptokurtosis and the approach that focuses on the tails, consider the class of finite normal mixtures as discussed in Subsect. "Finite Mixtures of Normal Distributions". These are leptokurtic in the sense of peakedness and tailedness (compared to the normal), but are light-tailed according to the tail-based perspective.

While it is universally accepted in the literature that the Gaussian is too light-tailed to be an appropriate model for the distribution of financial returns, there is no complete agreement with respect to the actual shape of the tails. This is not surprising because we cannot reasonably expect to find a model that accurately fits all markets at any time and place. However, the current mainstream opinion is that the probability for the occurrence of large (positive and negative) returns can often appropriately be described by Pareto-type tails. Such tail behavior is also frequently adopted as the definition of fat-tailedness per se, but the terminology in the literature is by no means unique.

A distribution has Pareto-type tails if they decay essentially like a power function as x becomes large, i. e., $\bar{F}$ is regularly varying (at infinity) with index $-\alpha$ (written $\bar{F} \in \mathrm{RV}_{-\alpha}$), meaning that

$$\bar{F}(x) = x^{-\alpha}L(x)\,, \quad \alpha > 0\,, \tag{6}$$

where $L > 0$ is a slowly varying function, which can be interpreted as "slower than any power function" (see [34,188,195] for a technical treatment of regular variation). The defining property of a slowly varying function is $\lim_{x\to\infty} L(tx)/L(x) = 1$ for any $t > 0$, and the aforemen-

tioned interpretation follows from the fact that, for any $\gamma > 0$, we have (cf. [195], p. 18)

$$\lim_{x\to\infty} x^{\gamma} L(x) = \infty\,, \quad \lim_{x\to\infty} x^{-\gamma} L(x) = 0\,. \tag{7}$$

Thus, for large x, the parameter α in (6), called the *tail index* or *tail exponent*, controls the rate of tail decay and provides a measure for the fatness of the tails.

Typical examples of slowly varying functions include $L(x) = c$, a constant, $L(x) = c + o(1)$, or $L(x) = (\log x)^k$, $x > 1$, $k \in \mathbb{R}$. The first case corresponds to strict Pareto tails, while in the second the tails are asymptotically Paretian in the sense that $\bar{F}(x) \cong cx^{-\alpha}$, which includes as important examples in finance the (non-normal) stable Paretian (see (13) in Subsect. "Alpha Stable and Related Distributions") and the Student's t distribution considered in Sect. "The Student t Distribution", where the tail index coincides with the characteristic exponent and the number of degrees of freedom, respectively. As an example for both, the Cauchy distribution with density $f(x) = [\pi(1 + x^2)]^{-1}$ has cdf $F(x) = 0.5 + \pi^{-1}\arctan(x)$. As $\arctan(x) = \sum_0^{\infty}(-1)^i x^{2i+1}/(2i + 1)$ for $|x| < 1$, and $\arctan(x) = \pi/2 - \arctan(1/x)$ for $x > 0$, we have $\bar{F}(x) \cong (\pi x)^{-1}$.

For the distributions mentioned in the previous paragraph, it is known that their moments exist only up to (and excluding) their tail indices, α. This is generally true for rvs with regularly varying tails and follows from (7) along with the well-known connection between moments and tail probabilities, i. e., for a non-negative rv X, and $r > 0$, $\mathbb{E}[X^r] = r\int_0^{\infty} x^{r-1}\bar{F}(x)\mathrm{d}x$ (cf. [95], p. 75). The only possible minor variation is that, depending on L, $\mathbb{E}\left[X^{\alpha}\right]$ may be finite. For example, a rv X with tail $\bar{F}(x) = cx^{-1}(\log x)^{-2}$ has finite mean. The property that moments greater than α do not exist provides further intuition for α as a measure of tail-fatness.

As indicated above, there is no universal agreement in the literature with respect to the definition of fat-tailedness. For example, some authors (e. g., [72,196]) emphasize the class of *subexponential* distributions, which are (although not exclusively) characterized by the property that their tails tend to zero slower than any exponential, i. e., for any $\gamma > 0$, $\lim_{x\to\infty} \mathrm{e}^{\gamma x}\bar{F}(x) = \infty$, implying that the moment generating function does not exist. Clearly a regularly varying distribution is also subexponential, but further members of this class are, for instance, the lognormal as well as the *stretched exponential*, or heavy-tailed Weibull, which has a tail of the form

$$\bar{F}(x) = \exp\left(-x^b\right)\,, \quad 0 < b < 1\,. \tag{8}$$

The stretched exponential has recently been considered by [134,152,153] as an alternative to the Pareto-type distribution (6) for modeling the tails of asset returns. Note that, as opposed to (6), both the lognormal as well as the stretched exponential possess power moments of all orders, although no exponential moment.

In addition, [22] coined the term *semi-heavy tails* for the generalized hyperbolic (GH) distribution, but the label may be employed more generally to refer to distributions with slower tails than the normal but existing moment generating function. The GH, which is now very popular in finance and nests many interesting special cases, will be examined in detail in Subsect. "The Generalized Hyperbolic Distribution".

As will be discussed in Sect. "Empirical Evidence About the Tails", results of extreme value theory (EVT) are often employed to identify the tail shape of return distributions. This has the advantage that it allows one to concentrate fully on the tail behavior, without the need to model the central part of the distribution. To sketch the idea behind this approach, suppose we attempt to classify distributions according to the limiting behavior of their normalized maxima. To this end, let $\{X_i, i \geq 1\}$ be an iid sequence of rvs with common cdf F, $M_n = \max\{X_1, \ldots, X_n\}$, and assume there exist sequences $a_n > 0$, $b_n \in \mathbb{R}$, $n \geq 1$, such that

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) = F^n\left(a_n x + b_n\right) \xrightarrow{n\to\infty} G(x)\,, \tag{9}$$

where G is assumed nondegenerate. To see that normalization is necessary, note that $\lim_{n\to\infty} P(M_n \leq x) = \lim_{n\to\infty} F^n(x) = 0$ for all $x < x_M := \sup\{x : F(x) < 1\} \leq \infty$, so that the limiting distribution is degenerate and of little help. If the above assumptions are satisfied, then, according to the classical Fisher–Tippett theorem of extreme value theory (cf. [188]), the limiting distribution G in (9) must be of the following form:

$$G_{\xi}(x) = \exp\left(-(1 + \xi x)^{-1/\xi}\right)\,, \quad 1 + \xi x > 0\,, \tag{10}$$

which is known as the *generalized extreme value distribution* (GEV), or *von Mises representation of the extreme value distributions* (EV). The latter term can be explained by the fact that (10) actually nests three different types of EV distributions, namely

(i) the Fréchet distribution, denoted by G_{ξ}^{+}, where $\xi > 0$ and $x > -1/\xi$,
(i) the so-called Weibull distribution of EVT, denoted by G_{ξ}^{-}, where $\xi < 0$ and $x < -1/\xi$, and
(iii) the Gumbel distribution, denoted by G_0, which corresponds to the limiting case as $\xi \to 0$, i. e., $G_0(x) = \exp(-\exp(-x))$, where $x \in \mathbb{R}$.

A cdf F belongs to the *maximum domain of attraction* (MDA) of one of the extreme value distributions nested in (10), written $F \in \mathrm{MDA}\left(G_\xi\right)$, if (9) holds, i. e., classifying distributions according to the limiting behavior of their extrema amounts to figuring out the MDAs of the extreme value distributions. It turns out that it is the tail behavior of a distribution F that accounts for the MDA it belongs to. In particular, $F \in \mathrm{MDA}(G_\xi^+)$ if and only if its tail $\bar{F} \in \mathrm{RV}_{-\alpha}$, where $\alpha = 1/\xi$. As an example, for a strict Pareto distribution, i. e., $F(x) = 1 - (u/x)^\alpha$, $x \geq u > 0$, with $a_n = n^{1/\alpha} u/\alpha$ and $b_n = n^{1/\alpha} u$, we have $\lim_{n\to\infty} F^n(a_n x + b_n) = \lim_{n\to\infty}(1 - n^{-1}(1+x/\alpha)^{-\alpha})^n = G_{1/\alpha}^+(x)$. Distributions in $\mathrm{MDA}(G_\xi^-)$ have a finite right endpoint, while, roughly, most of the remaining distributions, such as the normal, the lognormal and (stretched) exponentials, belong to $\mathrm{MDA}(G_0)$. The latter also accommodates a few distributions with finite right endpoint. See [188] for precise conditions. The important case of non-iid rvs is discussed in [136]. A central result is that, rather generally, vis-à-vis an iid sequence with the same marginal cdf, the maxima of stationary sequences converge to the same type of limiting distribution. See [63,167] for an application of this theory to ARCH(1) and GARCH(1,1) processes (see Sect. "Volatility Clustering and Fat Tails"), respectively.

One approach to exploit the above results, referred to as the *method of block maxima*, is to divide a given sample of return data into subsamples of equal length, pick the maximum of each subsample, assume that these have been generated by (10) (enriched with location and scale parameters to account for the unknown a_n and b_n), and find the maximum-likelihood estimate for ξ, location, and scale. Standard tests can then be conducted to assess, e. g., whether $\xi > 0$, i. e., the return distribution has Pareto-type tails. An alternative but related approach, which is based on further theoretical developments and often makes more efficient use of the data, is the *peaks over thresholds* (POT) method. See [72] for a critical discussion of these and alternative techniques.

We finally note that $1 - G_{1/\alpha}^+(\alpha(x-1)) \cong x^{-\alpha}$, while $1 - G_0(x) \cong \exp(-x)$, i. e., for the extremes, we have asymptotically a Pareto and an exponential tail, respectively. This may provide, on a meta-level, a certain rationale for reserving the notion of genuine fat-tailedness for the distributions with regularly varying tails.

Empirical Evidence About the Tails

The first application of power tails in finance appeared in Mandelbrot's [155] study of the log-price changes of cotton. Mandelbrot proposed to model returns with non-normal *alpha stable*, or *stable Paretian*, distributions, the properties of which will be discussed in some detail in Subsect. "Alpha Stable and Related Distributions". For the present discussion, it suffices to note that for this model the tail index α in (6), also referred to as *characteristic exponent* in the context of stable distributions, is restricted to the range $0 < \alpha < 2$, and that much of its theoretical appeal derives from the fact that, due to the generalized CLT, "Mandelbrot's hypothesis can actually be viewed as a generalization of the central-limit theorem arguments of Bachelier and Osborne to the case where the underlying distributions of price changes from transaction to transaction … have infinite variances" [75]. For the cotton price changes, Mandelbrot came up with a tail index of about 1.7, and his work was subsequently complemented by Fama [75] with an analysis of daily returns of the stocks belonging to the Dow Jones Industrial Average. [75] came to the conclusion that Mandelbrot's theory was supported by these data, with an average estimated α close to 1.9.

The findings of Mandelbrot and Fama initiated an extensive discussion about the appropriate distributional model for stock returns, partly because the stable model's implication that the tails are so fat that even the variance is infinite appeared to be too radical to many economists used to working with models built on the assumption of finite second moments. The evidence concerning the stable hypothesis gathered in the course of the debate until the end of the 1980s was not ultimately conclusive, but there were many papers reporting mainly negative results [4,28,36,40,54,67,98,99,109,135,176,180,184].

From the beginning of the 1990s, a number of researchers have attempted to estimate the tail behavior of asset returns directly, i. e., without making specific assumptions about the entire distributional shape. [86,115,142,143] use the method of block maxima (see Sect. "Defining Fat-Tailedness") to identify the maximum domain of attraction of the distribution of stock returns. They conclude that the Fréchet distribution with a tail index $\alpha > 2$ is most likely, implying Pareto-type tails which are thinner than those of the *stable* Paretian.

A second strand of literature assumes a priori the presence of a Pareto-type tail and focuses on the estimation of the tail index α. If, as is often the case, a power tail is deemed adequate, an explicit estimate of α is of great interest both from a practical and an academic viewpoint. For example, investors want to assess the likelihood of large losses of financial assets. This is often done using methods of extreme value theory, which require an accurate estimate of the tail exponent. Such estimates are also important because the properties of statistical tests and other quantities of interest, such as empirical autocorrelation

functions, frequently depend on certain moment conditions (e. g., [144,167]). Clearly the desire to figure out the actual tail shape has also an intrinsic component, as is reflected in the long-standing debate on the stable Paretian hypothesis. People simply wanted to know whether this distribution, with its appealing theoretical properties, is consistent with actual data. Moreover, empirical findings may guide economic theorizing, as they can help both in assessing the validity of certain existing models as well as in suggesting new explanations. Two examples will briefly be discussed at the end of the present section.

Within this second strand of literature, the Hill estimator [106] has become the most popular tool. It is given by

$$\hat{\alpha}_{k,n} = \left(\frac{1}{k-1} \sum_{j=1}^{k-1} \log X_{j,n} - \log X_{k,n} \right)^{-1}, \qquad (11)$$

where $X_{i,n}$ denotes the ith upper order statistic of a sample of length n, i. e., $X_{1,n} \geq X_{2,n} \geq \cdots \geq X_{n,n}$. See [64,72] for various approaches to deriving (11). If the tail is not regularly varying, the Hill estimator does not estimate anything.

A crucial choice to be made when using (11) is the selection of the threshold value k, i. e., the number of order statistics to be included in the estimation. Ideally, only observations from the tail region should be used, but choosing k to small will reduce the precision of the estimator. There exist various methods for picking k optimally in a mean-squared error sense [61,62], but much can be learned by looking at the *Hill plot*, which is obtained by plotting $\hat{\alpha}_{k,n}$ against k. If we can find a range of k-values where the estimate is approximately constant, this can be taken as a hint for where the "true" tail index may be located. As illustrated in [189], however, the Hill plot may not always be so well-behaved, and in this case the semi-automatic methods mentioned above will presumably also be of little help.

The theoretical properties of (11), along with technical conditions, are summarized in [72,189]. Briefly, for iid data generated from a distribution with tail $\bar{F} \in \mathrm{RV}_{-\alpha}$, the Hill estimator has been shown to be consistent [159] and asymptotically normal with standard deviation $\alpha/\sqrt{k}$ [100]. Financial data, however, are usually not iid but exhibit considerable dependencies in higher-order moments (see Sect. "Volatility Clustering and Fat Tails"). In this situation, i. e., with ARCH-type dynamics, (11) will still be consistent [190], but little is known about its asymptotic variance. However, simulations conducted in [123] with an IGARCH model, as defined in Sect. "Volatility Clustering and Fat Tails", indicate that, under such dependencies, the actual standard errors may be seven to eight times larger than those implied by the asymptotic theory for iid variables.

The Hill estimator was introduced into the econometrics literature in the series of articles [107,113,125,126]. [125,126], using weekly observations, compare the tails of exchange rate returns in floating and fixed exchange rate systems, such as the Bretton Woods period and the EMS. They find that for the fixed systems, most tail index estimates are below 2, i. e., consistent with the alpha stable hypothesis, while the estimates are significantly larger than 2 (ranging approximately from 2.5 to 4) for the float. [126] interpret these results in the sense that "a float lets exchange rates adjust more smoothly than any other regime that involves some amount of fixity". Subsequent studies of floating exchange rates using data ranging from weekly [107,110,111] over daily [58,89,144] to very high-frequency [59,61,170] have confirmed the finding of these early papers that the tails are not fat enough to be compatible with the stable Paretian hypothesis, with estimated tail indices usually somewhere in the region 2.5–5. [58] is the first to investigate the tail behavior of the euro against the US dollar, and finds that it is similar both to the German mark in the pre-euro era as well as to the yen and the British pound, with estimated exponents hovering around 3–3.5.

Concerning estimation with data at different time scales, a comparison of the results reported in the literature reveals that the impact on the estimated tail indices appears to be moderate. [59] observe an increase in the estimates when moving from 30-minute to daily returns, but they argue that these changes, due to the greater bias at the lower frequencies, are small enough to be consistent with α being invariant under time aggregation.

Note that if returns were independently distributed, their tail behavior would in fact be unaffected by time aggregation. This is a consequence of (2) along with Feller's (p. 278 in [80]) theorem on the convolution of regularly varying distributions, stating that any finite convolution of a regularly varying cdf $F(x)$ has a regularly varying tail with the same index. Thus, in principle, the tail survives forever, but, as long as the variance is finite, the CLT ensures that in the course of aggregation an increasing probability weight is allocated to the center of the distribution, which becomes closer to a Gaussian shape. The probability of observing a tail event will thus decrease. However, for fat-tailed distributions, the convergence to normality can be rather slow, as reflected in the observation that pronounced non-normalities in financial returns are often observed even at a weekly and (occasionally) monthly frequency. See [41] for an informative discussion of these is-

sues. The fact that returns are, in general, not iid makes the interpretation of the approximate stability of the tail index estimates observed across papers employing different frequencies not so clear-cut, but Feller's theorem may nevertheless provide some insight.

There is also an extensive literature reporting tail index estimates of stock returns, mostly based on daily [2,89,92,112,113,144,145,146,177] and higher frequencies [2,91,92,147,181]. The results are comparable to those for floating exchange rates in that the tenor of this literature, which as a whole covers all major stock markets, is that most stock return series are characterized by a tail index somewhere in the region 2.5–5, and most often below 4. That is, the tails are thinner than expected under the stable Paretian hypothesis, but the finiteness of the third and in particular the fourth moments (and hence kurtosis) may already be questionable. Again, the results appear to be relatively insensitive with respect to the frequency of the data, indicating a rather slow convergence to the normal distribution. Moreover, most authors do not find significant differences between the left and the right tail, although, for stock returns, the point estimates tend to be somewhat lower for the left tail (e. g., [115,145]).

Applications to the bond market appear to be rare, but see [201], who report tail index estimates between 2.5 and 4.5 for 5-minute and 1-hour Bund future returns and somewhat higher values for daily data. [160] compare the tail behaviors of spot and future prices of various commodities (including cotton) and find that, while future prices resemble stock prices with tail indices in the region 2.5–4, spot prices are somewhat fatter tailed with α hovering around 2.5 and, occasionally, smaller than 2.

Summarizing, it is now a widely held view that the distribution of asset returns can typically be described as fat-tailed in the power law sense but with finite variance. Thus, currently there seems to exist a far reaching consensus that the stable Paretian model is not adequate for financial data, but see [162,202] for a different viewpoint. A consequence of the prevalent view is that asset return distributions belong to the Gaussian domain of attraction, but that the convergence appears to be very slow.

To illustrate typical findings as reported above, let us consider the S&P500 returns described in Sect. "Definition of the Subject". A first informal check of the appropriateness of a power law can be obtained by means of a log-log plot of the empirical tail, i. e., if $1 - F(x) = \bar{F}(x) \approx cx^{-\alpha}$ for large x, then a plot of the log of the empirical complementary cdf, $\bar{F}(x)$, against $\log x$ should display a linear behavior in its outer part. For the data at hand, such a plot is shown in the bottom right panel of Fig. 1. Assuming homogeneity across the tails, we pool negative and positive returns by first removing the sample mean and then taking absolute values. We have also multiplied (1) by 100, so that the returns are interpretable in terms of percentage changes. The plot suggests that a power law regime may be starting from approximately the 90% quantile. Included in Fig. 1 is also a regression line ("fit") fitted to the log-tail using the 500 upper (absolute) return observations. This yields, as a rough estimate for the tail index, a slope of $\hat{\alpha} = 3.13$, with a coefficient of determination $R^2 = 0.99$. A Hill plot for $k \leq 500$ in (11) is shown in the bottom left panel of Fig. 1. The estimates are rather stable over the entire region and suggest an α somewhere in the interval $(3, 3.5)$, which is reconcilable with the results in the literature summarized above. A somewhat broader picture can be obtained by considering individual stocks. Here we consider the 176 stocks that were listed in the S&P500 from January 1985 to December 2006. Figure 2 presents, for each $k \leq 500$, the 5%, 50%, and 95% quantiles of the distribution of (11) over the different stocks. The median is close to 3 throughout, and it appears that an estimate in $(2.5, 4.5)$ would be reasonable for most stocks.

At this point, it may be useful to note that the issue is not whether a power law is true in the strict sense but only if it provides a reasonable approximation in the relevant tail region. For example, it might be argued that asset returns actually have finite support, implying finiteness of all moments and hence inappropriateness of a Pareto-type tail. However, as concisely pointed out in [144], "saying that the support of an empirical distribution is bounded says very little about the nature of outlier activity that may occur in the data".

We clearly cannot expect to identify the "true" distribution of financial variables. For example, [153] have demonstrated that by standard techniques of EVT it is virtually impossible, even in rather large samples, to discriminate between a power law and a stretched exponential (8) with a small value of b, thus questioning, for example, the conclusiveness of studies relying on the block maxima method, as referred to above. A similar point was made in [137], who showed by simulation that a three-factor stochastic volatility model, with a marginal distribution known to have all its moments finite, can generate apparent power laws in practically relevant sample sizes. As put forward in [152], "for most practical applications, the relevant question is not to determine what is the true asymptotic tail, but what is the best effective description of the tails in the domain of useful applications".

As is evident in Fig. 1, a power law may (and often does) provide a useful approximation to the tail behavior of actual data, but there is no reason to expect that it

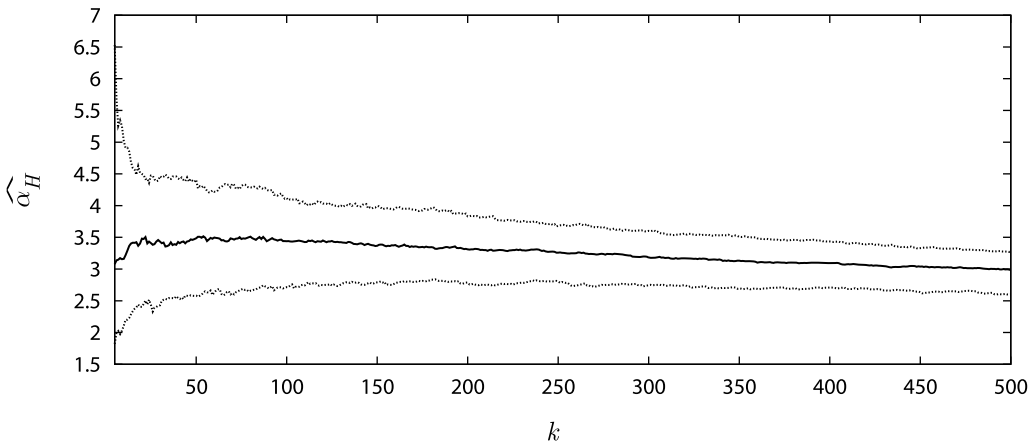

Financial Economics, Fat-Tailed Distributions, Figure 2
Shown are, for $k \leq 500$, the 95%, 50%, and 5% quantiles of the distribution of the Hill estimator $\hat{\alpha}_{k,n}$, as defined in (11), over 176 stocks included in the S&P500 stock index

will appear in every market, and a broad range of heavy and semi-heavy tailed distributions (such as the GH in Subsect. "The Generalized Hyperbolic Distribution") may provide an adequate fit. For instance, [93] investigate the tail behavior of high-frequency returns of one of the most frequently traded stocks on the Paris Stock Exchange (Alcatel) and conclude that the tails decay at an exponential rate, and [119,197] obtain similar results for daily returns of the Nikkei 225 index and various individual US stocks, respectively. As a further illustration, without rigorous statistical testing, Fig. 3 shows the log-log tail plot for daily returns of the German stock market index DAX

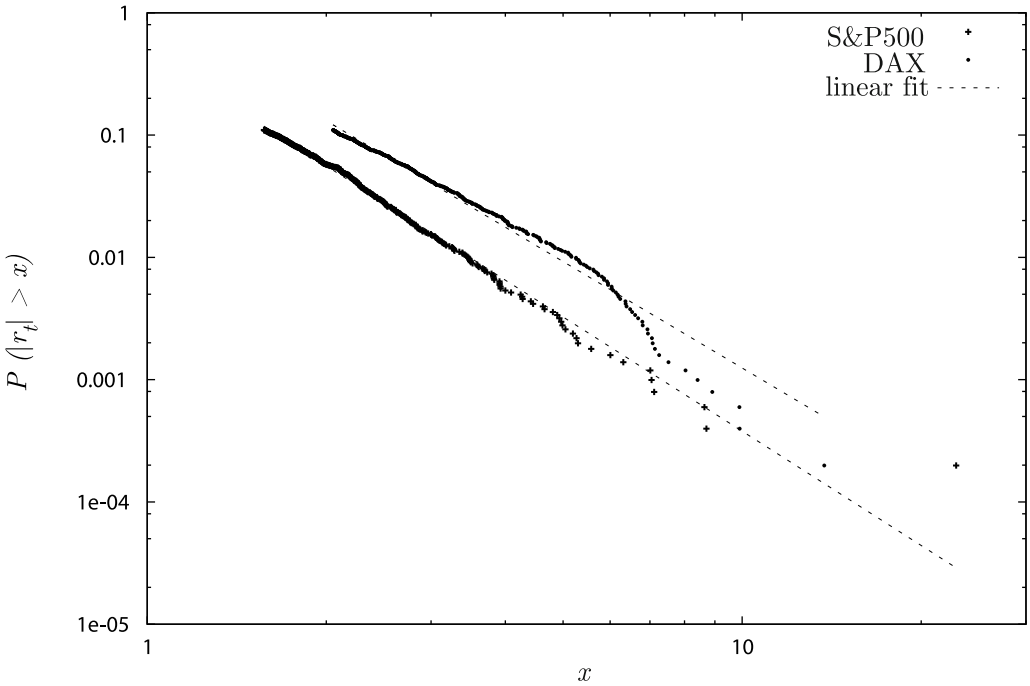

Financial Economics, Fat-Tailed Distributions, Figure 3
The figure shows, on a log-log scale, the complementary cdf, $\bar{F}(x)$, for the largest 500 absolute return observations both for the daily S&P500 returns from January 1985 to December 2006 and the daily DAX returns from July 1987 to July 2007

from July 3, 1987 to July 4, 2007 for the largest 500 out of 5,218 (absolute) return observations, along with a regression-based linear fit. For purposes of comparison, the corresponding figure for the S&P500 has also been reproduced from Fig. 1. While the slopes of the fitted power laws exhibit an astonishing similarity (in fact, the estimated tail index of the DAX is 2.93), it is clear from Fig. 3 that an asymptotic power law, although not necessarily inconsistent with the data, is much less self-evident for the DAX than for the S&P500, due to the apparent curvature particularly in the more extreme tail.

It is finally worthwhile to mention that financial theory in general, although some of its models are built on the *assumption* of a specific distribution, has little to say about the distribution of financial variables. For example, according to the efficient markets paradigm, asset prices change in response to the arrival of relevant new information, and, consequently, the distributional properties of returns will essentially reflect those of the news process. As noted by [148], an exception to this rule is the model of rational bubbles of [35]. [148] show that this class of processes gives rise to an (approximate) power law for the return distribution. However, the structure of the model, i. e., the assumption of rational expectations, restricts the tail exponent to be below unity, which is incompatible with observed tail behaviors.

More recently, prompted by the observation that estimated tail indices are often located in a relatively narrow interval around 3, [83,84,85] have developed a model to explain a hypothesized "inverse cubic law for for the distribution of stock price variations" [91], valid for highly developed economies, i. e., a power law tail with index $\alpha = 3$. This model is based on Zipf's law for the size of large institutional investors and the hypothesis that large price movements are generated by the trades of large market participants via a square-root price impact of volume, V, i. e., $r \cong h\sqrt{V}$, where r is the log return and h is a constant. Putting these together with a model for profit maximizing large funds, which have to balance between trading on a perceived mispricing and the price impact of their actions, leads to a power law distribution of volume with tail index 1.5, which by the square-root price impact function and simple power law accounting then produces the "cubic law". See [78,182] for a discussion of this model and the validity of its assumptions. In a somewhat similar spirit, [161] find strong evidence for *exponentially* decaying tails of daily Indian stock returns and speculate about a general inverse relationship between the stage of development of an economy and the closeness to Gaussianity of its stock markets, but it is clear that this is really just speculation.

Some Specific Distributions

Alpha Stable and Related Distributions

As noted in Sect. "Empirical Evidence About the Tails", the history of heavy tailed distributions in finance has its origin in the *alpha stable* model proposed by Mandelbrot [154,155]. Being the first alternative to the Gaussian law, the alpha stable distribution has a long history in financial economics and econometrics, resulting in a large number of books and review articles.

Apart from its good empirical fit the stable distribution has also some attractive theoretical properties such as the stability property and domains of attraction. The stability property states that the index of stability (or shape parameter) remains the same under scaling and addition of different stable rv with the same shape parameter. The concept of domains of attraction is related to a generalized CLT. More specifically, dropping the assumption of a finite variance in the classical CLT, the domains of attraction states, loosely speaking, that the alpha stable distribution is the only possible limit distribution. For a more detailed discussion of this concept we refer to [169], who also provide an overview over alternative stable schemes. While the fat-tailedness of the alpha stable distributions makes it already an attractive candidate for modeling financial returns, the concept of the domains of attraction provides a further argument for its use in finance, as under the relaxation of the assumption of a finite variance of the continuously arriving return innovations the resulting return distribution at lower frequencies is generally an alpha stable distribution.

Although the alpha stable distribution is well established in financial economics and econometrics, there still exists some confusion about the naming convention and parameterization. Popular terms for the alpha stable distribution are the *stable Paretian*, *Lévy stable* or simply *stable* laws. The parameterization of the distribution in turn varies mostly with its application. For instance, to numerically integrate the characteristic function, it is preferable to have a continuous parameterization in all parameters.

The numerical integration of the alpha stable distributions is important, since with the exception of a few special cases, its pdf is unavailable in closed form. However, the characteristic function of the standard parameterization is given by

$$\mathbb{E}\left[\exp(itX)\right] = \begin{cases} \exp\left(-c^{\alpha}\,|t|^{\alpha}\left(1 - i\beta\,\mathrm{sign}(t)\tan\frac{\pi\alpha}{2}\right) + i\tau t\right) & \alpha \neq 1 \\ \exp\left(-c\,|t|\left(1 + i\beta\frac{2}{\pi}\,\mathrm{sign}(t)\ln(|t|)\right) + i\tau t\right) & \alpha = 1\,, \end{cases} \tag{12}$$

where i is the imaginary unit, $\operatorname{sign}(\cdot)$ denotes the sign function, which is defined as

$$\operatorname{sign}(x) = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0\,, \end{cases}$$

$0 < \alpha \leq 2$ denotes the *shape parameter*, *characteristic exponent* or *index of stability*, $-1 \leq \beta \leq 1$ is the skewness parameter, and $\tau \in \mathbb{R}$ and $c \geq 0$ are the location and scale parameters, respectively.

Figure 4 highlights the impact of the parameters α and β. β controls the skewness of the distribution. The shape parameter α controls the behavior of the tails of the distribution and therefore the degree of leptokurtosis. For $\alpha < 2$ moments only up to (and excluding) α exist, and for $\alpha > 1$ we have $\mathbb{E}[X] = \tau$. In general, for $\alpha \in (0, 1)$ and $\beta = 1$ ($\beta = -1$) the support of the distribution is the set (τ, ∞) (or $(-\infty, \tau)$) rather than the whole real line. In the following we call this stable distribution with $\alpha \in (0, 1)$, $\tau = 0$ and $\beta = 1$ the positive alpha stable distribution.

Moreover, for $\alpha < 2$ the stable law has asymptotic power tails,

$$\bar{F}(x) = P(X > x) \cong c^{\alpha} d\,(1+\beta)\, x^{-\alpha}$$
$$f_S(x, \alpha, \beta, c, \tau) \cong \alpha c^{\alpha} d\,(1+\beta)\, x^{-\alpha+1}$$

with $d = \sin\left(\frac{\pi\alpha}{2}\right) \Gamma(\alpha)/\pi$.

For $\alpha = 2$ the stable law is equivalent to the normal law with variance $2c^2$, for $\alpha = 1$ and $\beta = 0$ the Cauchy distribution is obtained, and for $\alpha = 1/2$, $\beta = 1$ and $\tau = 0$ the stable law is equivalent to the Lévy distribution, with support over the positive real line.

An additional property of the stable laws is that they are closed under convolution for constant α, i. e., for two independent alpha stable rvs $X_1 \sim S(\alpha, \beta_1, c_1, \tau_1)$ and $X_2 \sim S(\alpha, \beta_2, c_2, \tau_2)$ with common shape parameter α we

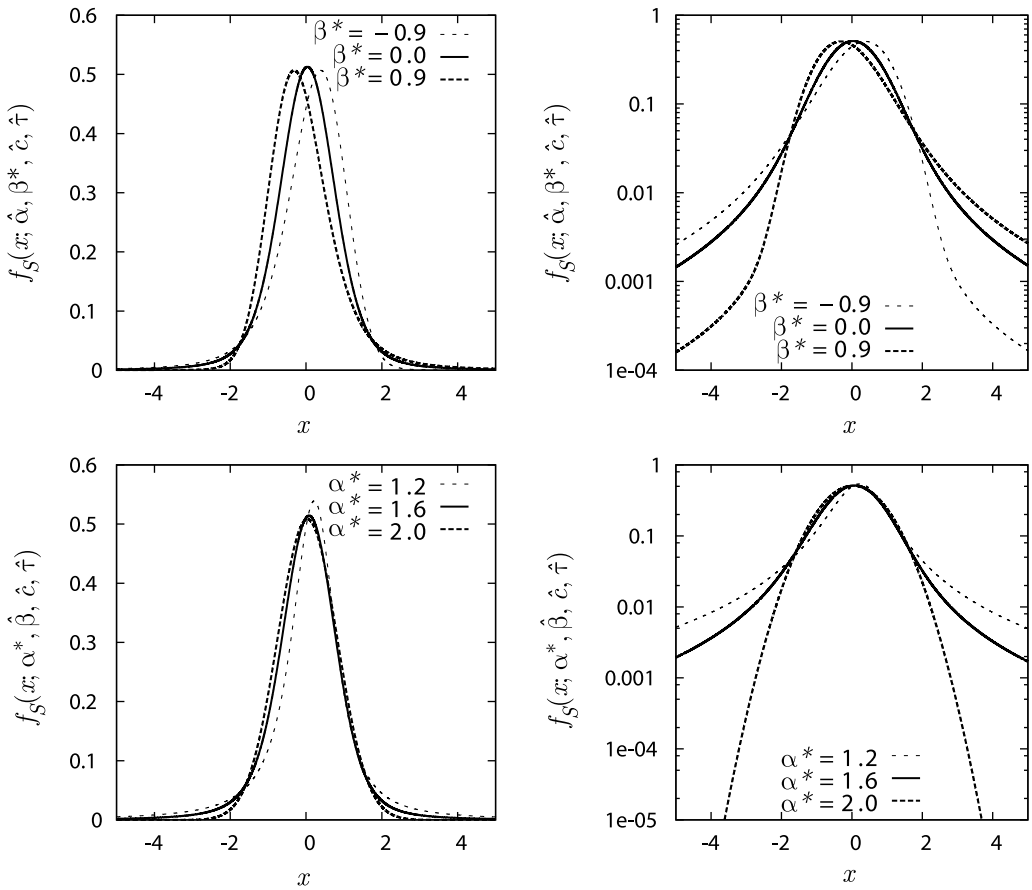

Financial Economics, Fat-Tailed Distributions, Figure 4
Density function (pdf) of the alpha stable distribution for different parameter vectors. The *right panel* plots the log-densities to better visualize the tail behavior. The *upper (lower) section* present the pdf for different values of β (α)

have

$$X_1 + X_2 \sim S\left(\alpha, \frac{\beta_1 c_1^\alpha + \beta_2 c_2^\alpha}{c_1^\alpha + c_2^\alpha}, \left(c_1^\alpha + c_2^\alpha\right)^{1/\alpha}, \tau_1 + \tau_2\right)$$

and

$$aX_1 + b \sim \begin{cases} S\left(\alpha, \operatorname{sign}(a)\beta, |a|\, c, a\tau + b\right) & \alpha \neq 1 \\ S\left(1, \operatorname{sign}(a)\beta, |a|\, c, a\tau + b - \frac{2}{\pi}\beta c a \log|a|\right) & \alpha = 1\,. \end{cases}$$

These results can be extended to n stable rvs. The closedness under convolution immediately implies the infinite divisibility of the stable law. As such every stable law corresponds to a Lévy process. A more detailed analysis of alpha stable processes in the context of Lévy processes is given in [192,193].

The computation and estimation of the alpha stable distribution is complicated by the aforementioned nonexistence of a closed form pdf. As a consequence, a number of different approximations for evaluating the density have been proposed, see e. g. [65,175]. On the basis of these approximations, parameter estimation is facilitated using for example the maximum-likelihood estimator, see [66], or other estimation methods. As maximum-likelihood estimation relies on computationally demanding numerical integration methods, the early literature focused on alternative estimation methods. The most important methods include the quantile estimation suggested by [77,163], which is still heavily applied in order to obtain starting values for more sophisticated estimation procedures, as well as the characteristic function approach proposed by [127,131,186]. However, based on its nice asymptotic properties and presently available computational power, the maximum-likelihood approach is preferable.

Many financial applications also involve the simulation of a return series. In derivative pricing, for example, the computation of an expectation is oftentimes impossible as the financial instrument is generally a highly nonlinear function of asset returns. A common way to alleviate this problem is to apply Monte Carlo integration, which in turn requires quasi rvs drawn from the respective return distribution, i. e. the alpha stable distribution. A useful simulation algorithm for alpha stable rvs is proposed by [49], which is a generalization of the algorithm of [120] to the non-symmetric case. A random variable X distributed according to the stable law, $S(\alpha, \beta, c, \tau)$, can be generated as follows:

1. Draw a rv U, uniformly distributed over the interval $(-\pi/2, \pi/2)$, and an (independent) exponential rv E with unit mean,
2. if $\alpha \neq 1$, compute

$$X = cS\frac{\sin(\alpha(U+B))}{\cos^{1/\alpha}(U)} \cdot \left(\frac{\cos(U - \alpha(U+B))}{E}\right)^{(1-\alpha)/\alpha} + \tau$$

where

$$B := \frac{\arctan\left(\beta \tan\left(\frac{\pi\alpha}{2}\right)\right)}{\alpha}$$

$$S := \left(1 + \beta^2 \tan^2\left(\frac{\pi\alpha}{2}\right)\right)^{1/(2\alpha)}$$

for $\alpha = 1$ compute

$$X = c\frac{2}{\pi}\left(\left(\frac{\pi}{2} + \beta U\right)\tan(U) - \beta \log\left(\frac{\frac{\pi}{2}E\cos(U)}{\frac{\pi}{2} + \beta U}\right)\right) + \frac{2}{\pi}\beta c \log(c) + \tau\,.$$

Interestingly, for $\alpha = 2$ the algorithm collapses to the Box–Muller method [42] to generate normally distributed rvs.

As further discussed in Subsect. "The Generalized Hyperbolic Distribution", the mixing of normal distributions allows one to derive interesting distributions having support over the real line and exhibiting heavier tails than the Gaussian. While generally any distribution with support over the positive real line can be used as the mixing distribution for the variance, transformations of the positive alpha stable distribution are often used in financial modeling.

In this context the symmetric alpha stable distributions have a nice representation. In particular, if $X \sim S(\alpha^*, 0, c, 0)$ and A (independent of X) is an α/α^* positive alpha stable rv with scale parameter $\cos^{\alpha^*/\alpha}\left(\frac{\pi\alpha}{2\alpha^*}\right)$ then

$$Z = A^{1/\alpha^*} X \sim S(\alpha, 0, c, 0)\,.$$

For $\alpha^* = 2$ this property implies that every symmetric alpha stable distribution, i. e. an alpha stable distribution with $\beta = 0$, can be viewed as being conditionally normally distributed, i. e., it can be represented as a continuous variance normal mixture distribution.

Generally, the tail behavior of the resulting mixture distribution is completely determined by the (positive) tails of the variance mixing distribution. In the case of the positive alpha stable distribution this implies that the resulting symmetric stable distribution exhibits very heavy tails, in fact the second moment does not exist. As the literature is controversial on the adequacy of such heavy tails (see Sect. "Empirical Evidence About the Tails"), transformations of the positive alpha stable distribution are oftentimes considered to weight down the tails. The method of

exponential tilting is very useful in this context. In a general setup the exponential tilting of a rv X with respect to a rv Y (defined on the same probability space) defines a new rv $\tilde{X}$, whose pdf can be written as

$$f_{\tilde{X}}(x;\theta) = f_X(x)\frac{\mathbb{E}\left[\exp(\theta Y)|X=x\right]}{\mathbb{E}\left[\exp(\theta Y)\right]},$$

where the parameter θ determines the "degree of dampening". The exponential tilting of a rv X with respect to itself, known as *Esscher transformation*, is widely used in financial economics and mathematics, see e. g. [88]. In this case the resulting pdf is given by

$$\begin{aligned} f_{\tilde{X}}(x;\theta) &= \frac{\exp(\theta x)}{\mathbb{E}\left[\exp(\theta X)\right]} f_X(x) \\ &= \exp(\theta x - K(\theta))\, f_X(x)\,, \end{aligned} \tag{13}$$

with $K(\cdot)$ denoting the cumulant generating function, $K(\theta) := \log\left(\mathbb{E}\left[\exp(\theta X)\right]\right)$.

The *tempered stable* (TS) laws are given by an application of the Esscher transform (13) to a positive alpha stable law. Note that the *Laplace transform* $\mathbb{E}[\exp(-tX)]$, $t \geq 0$, of a positive alpha stable rv is given by $\exp(-\delta(2t)^\alpha)$, where $\delta = c^\alpha/(2^\alpha\cos(\frac{\pi\alpha}{2}))$. Thus, with $\theta = -(1/2)\gamma^{1/\alpha} \leq 0$, the pdf of the tempered stable law is given by

$$\begin{aligned} f_{\mathrm{TS}}(x;\alpha,\delta,\gamma) &= \frac{\exp\left(-\frac{1}{2}\gamma^{1/\alpha}x\right)}{\mathbb{E}\left[\exp\left(-\frac{1}{2}\gamma^{1/\alpha}X\right)\right]} f_{\mathrm{S}}(x;\alpha,1,c(\delta,\alpha),0) \\ &= \exp\left(\delta\gamma - \tfrac{1}{2}\gamma^{1/\alpha}x\right) f_{\mathrm{S}}(x;\alpha,1,c(\delta,\alpha),0) \end{aligned}$$

with $0 < \alpha < 1$, $\delta > 0$, and $\gamma \geq 0$.

A generalization of the TS distribution was proposed by [22]. This class of *modified stable* (MS) laws can be obtained by applying the following transformation

$$f_{\mathrm{MS}}(x,\alpha,\lambda,\delta,\gamma) = c(\alpha,\lambda,\delta,\gamma)\, x^{\lambda+\alpha} f_{\mathrm{TS}}(x;\alpha,\delta,\gamma)\,, \tag{14}$$

with $\lambda \in \mathbb{R}$, $\gamma \vee (-\lambda) > 0$ and $c(\alpha,\lambda,\delta,\gamma)$ is a norming constant. For a more detailed analysis, we refer to [22]. Note that the terms "modified stable" or "tempered stable distribution" are not unique in the literature. Very often the so-called truncated Lévy flights/processes (see for example [56,130,157]) are also called TS processes (or corresponding distributions). These distributions are obtained by applying a smooth downweighting of the large jumps (in terms of the Lévy density).

The MS distribution is a quite flexible distribution defined over the positive real line and nests several important distributions. For instance, for $\alpha = 1/2$, the MS distribution reduces to the *generalized inverse Gaussian* (GIG) distribution, which is of main interest in Subsect. "The Generalized Hyperbolic Distribution".

Note that in contrast to the unrestricted MS distribution, the pdf of the GIG distribution is available in closed form and can be straightforwardly obtained by applying the above transformation. In particular, for $\alpha = 1/2$, the positive alpha stable distribution is the Lévy distribution with closed form pdf given by

$$f_{\mathrm{S}}(x;1/2,1,c,0) = \sqrt{\frac{c}{2\pi}}\frac{\exp\left(-\frac{c}{2x}\right)}{x^{3/2}}\,.$$

Applying the Esscher transformation (13) with $\theta = -(1/2)\gamma^2$ yields the pdf of the inverse Gaussian (or Wald) distribution

$$f_{\mathrm{IG}}(x;\delta,\gamma) = \frac{\delta}{\sqrt{2\pi}}x^{-3/2}\exp\left(\delta\gamma - \left(\delta^2x^{-1} + \gamma^2x\right)/2\right),$$

where $\delta > 0$ and $\gamma \geq 0$. Applying the transformation (14) delivers the pdf of the GIG distribution,

$$\begin{aligned} f_{\mathrm{GIG}}(x;\lambda,\delta,\gamma) = \frac{(\gamma/\delta)^\lambda}{2K_\lambda(\delta\gamma)}x^{\lambda-1} \\ \cdot\exp\left(-\left(\delta^2x^{-1} + \gamma^2x\right)/2\right)\,, \end{aligned} \tag{15}$$

with $K_\lambda(\cdot)$ being the *modified Bessel function of the third kind* and of order $\lambda \in \mathbb{R}$. Note that this function is oftentimes called the modified Bessel function of the second kind or Macdonald function. Nevertheless, one representation is given by

$$K_\lambda(x) = \frac{1}{2}\int_0^\infty y^{\lambda-1}\exp\left(-\frac{1}{2}x\left(y+y^{-1}\right)\right)\mathrm{d}y\,.$$

The parameters of the GIG distribution are restricted to satisfy the following conditions:

$$\begin{aligned} &\delta \geq 0 \text{ and } \gamma > 0 \quad \text{if } \lambda > 0 \\ &\delta > 0 \text{ and } \gamma > 0 \quad \text{if } \lambda = 0 \\ &\delta > 0 \text{ and } \gamma \geq 0 \quad \text{if } \lambda < 0\,. \end{aligned} \tag{16}$$

Importantly, in contrast to the positive alpha stable law, all moments exist and are given by

$$\mathbb{E}\left[X^r\right] = \left(\frac{\delta}{\gamma}\right)^r\frac{K_{\lambda+r}(\delta\gamma)}{K_\lambda(\delta\gamma)} \tag{17}$$

for all $r > 0$. For a very detailed analysis of the GIG distribution we refer to [117]. The GIG distribution nests several positive distributions as special cases and as limit distributions. Since all of these distributions belong to a special class of the generalized hyperbolic distribution, we proceed with a discussion of the latter, thus providing a broad framework for the discussion of many important distributions in finance.

The Generalized Hyperbolic Distribution

The mixing of normal distributions is well suited for financial modeling, as it allows construction of very flexible distributions. For example, the *normal variance-mean mixture*, given by

$$X = \mu + \beta V + \sqrt{V} Z \,, \tag{18}$$

with Z being normally distributed and V a positive random variable, generally exhibits heavier tails than the Gaussian distribution. Moreover, this mixture possesses interesting properties, for an overview see [26]. First, similarly to other mixtures, the normal variance-mean mixture is a conditional Gaussian distribution with conditioning on the volatility states, which is appealing when modeling financial returns. Second, if the mixing distribution, i. e. the distribution of V, is infinitely divisible, then X is likewise infinitely divisible. This implies that there exists a Lévy process with support over the whole real line, which is distributed at time $t = 1$ according to the law of X. As the theoretical properties of Lévy processes are well established in the literature (see, e. g., [24,194]), this result immediately suggests to formulate financial models in terms of the corresponding Lévy process.

Obviously, different choices for the distribution of V result in different distributions of X. However, based on the above property, an infinitely divisible distribution is most appealing. For the MS distributions discussed in Subsect. "Alpha Stable and Related Distributions", infinite divisibility is not yet established, although [22] strongly surmise so. However, for some special cases infinite divisibility has been shown to hold. The most popular is the GIG distribution yielding the *generalized hyperbolic* (GH) distribution for X. The latter distribution was introduced by [17] for modeling the distribution of the size of sand particles. The infinite divisibility of the GIG distribution was shown by [20].

To derive the GH distribution as a normal variance-mean mixture, let $V \sim \mathrm{GIG}(\lambda, \delta, \gamma)$ as in (15), with $\gamma = \sqrt{\alpha^2 - \beta^2}$, and $Z \sim \mathrm{N}(0, 1)$ independent of V. Applying (18) yields the GH distributed rv $X \sim \mathrm{GH}(\lambda, \alpha, \beta, \mu, \delta)$ with pdf

$$\begin{aligned} & f_{\mathrm{GH}}(x; \lambda, \alpha, \beta, \mu, \delta) \\ &= \frac{(\delta\gamma)^{\lambda} (\delta\alpha)^{1/2-\lambda}}{\sqrt{2\pi}\delta K_{\lambda}(\delta\gamma)} \left(1 + \frac{(x-\mu)^2}{\delta^2}\right)^{\lambda/2-1/4} \\ &\quad \cdot K_{\lambda-1/2}\left(\alpha\delta\sqrt{1 + \frac{(x-\mu)^2}{\delta^2}}\right) \exp(\beta(x-\mu)) \end{aligned}$$

for $\mu \in \mathbb{R}$ and

$$\begin{aligned} &\delta \geq 0 \text{ and } |\beta| < \alpha \quad \text{if } \lambda > 0 \\ &\delta > 0 \text{ and } |\beta| < \alpha \quad \text{if } \lambda = 0 \\ &\delta > 0 \text{ and } |\beta| \leq \alpha \quad \text{if } \lambda < 0 \,, \end{aligned}$$

which are the induced parameter restrictions of the GIG distribution given in (16).

Note that, based on the mixture representation (18), the existing algorithms for generating GIG distributed rvs can be used to draw rvs from the GH distribution, see [12, 60].

For $|\beta + u| < \alpha$, the moment generating function of the GH distribution is given by

$$\begin{aligned} \mathbb{E}\left[\exp(uX)\right] &= \exp(\mu u) \left(\frac{\alpha^2 - \beta^2}{\alpha^2 - (\beta+u)^2}\right)^{\lambda/2} \\ &\quad \cdot \frac{K_{\lambda}\left(\delta\sqrt{\alpha^2 - (\beta+u)^2}\right)}{K_{\lambda}\left(\delta\sqrt{\alpha^2 - \beta^2}\right)} \,. \end{aligned} \tag{19}$$

As the moment generating function is infinitely differentiable in the neighborhood of zero, moments of all orders exist and have been derived in [25]. In particular, the mean and the variance of a GH rv X are given by

$$\begin{aligned} \mathbb{E}[X] &= \mu + \frac{\beta\delta}{\sqrt{\alpha^2 - \beta^2}} \frac{K_{\lambda+1}\left(\delta\sqrt{\alpha^2 - \beta^2}\right)}{K_{\lambda}\left(\delta\sqrt{\alpha^2 - \beta^2}\right)} \\ &= \mu + \beta\mathbb{E}[X_{\mathrm{GIG}}] \\ \mathbb{V}[X] &= \frac{\delta}{\sqrt{\alpha^2 - \beta^2}} \frac{K_{\lambda+1}\left(\delta\sqrt{\alpha^2 - \beta^2}\right)}{K_{\lambda}\left(\delta\sqrt{\alpha^2 - \beta^2}\right)} + \frac{\beta^2\delta^2}{\alpha^2 - \beta^2} \\ &\quad \cdot \left(\frac{K_{\lambda+2}\left(\delta\sqrt{\alpha^2 - \beta^2}\right)}{K_{\lambda}\left(\delta\sqrt{\alpha^2 - \beta^2}\right)} - \frac{K_{\lambda+1}^2\left(\delta\sqrt{\alpha^2 - \beta^2}\right)}{K_{\lambda}^2\left(\delta\sqrt{\alpha^2 - \beta^2}\right)}\right) \\ &= \mathbb{E}[X_{\mathrm{GIG}}] + \beta^2\mathbb{V}[X_{\mathrm{GIG}}] \,, \end{aligned}$$

with $X_{\mathrm{GIG}} \sim \mathrm{GIG}(\lambda, \delta, \gamma)$ denoting a GIG distributed rv. Skewness and kurtosis can be derived in a similar way using the third and fourth derivative of the moment generating function (19). However, more information on the tail behavior is given by

$$f_{\mathrm{GH}}(x; \lambda, \alpha, \beta, \mu, \delta) \cong |x|^{\lambda-1} \exp((\mp\alpha + \beta)x) \,,$$

which shows that the GH distribution exhibits semi-heavy tails, see [22].

The moment generating function (19) also shows that the GH distribution is generally not closed under convolution. However, for $\lambda \in \{-1/2, 1/2\}$, the modified Bessel

function of the third kind satisfies

$$K_{-\frac{1}{2}}(x) = K_{\frac{1}{2}}(x) = \sqrt{\frac{\pi}{2x}} \exp(-x) \ ,$$

yielding the following form of the moment generating function for $\lambda = -1/2$

$$\mathbb{E}\left[\exp(uX)|\lambda = -1/2\right] = \exp(\mu u) \frac{\exp\left(\delta\sqrt{\alpha^2 - \beta^2}\right)}{\exp\left(\delta\sqrt{\alpha^2 - (\beta + u)^2}\right)} \ ,$$

which is obviously closed under convolution. This special class of the GH distribution is called normal inverse Gaussian distribution and is discussed in more detail in Subsect. "The Normal Inverse Gaussian Distribution". The closedness under convolution is an attractive property of this distribution as it facilitates forecasting applications.

Another very popular distribution that is nested in the GH distribution is the hyperbolic distribution given by $\lambda = 1$ (see the discussion in Subsect. "The Hyperbolic Distribution"). Its popularity is primarily based on its pdf, which can (except for the norming constant) be expressed in terms of elementary functions allowing for a very fast numerical evaluation. However, given the increasing computer power and the slightly better performance in reproducing the unconditional return distribution, the normal inverse Gaussian distribution is now the most often used subclass.

Interestingly, further well-known distributions can be expressed as limiting cases of the GH distribution, when certain of its parameters approach their extreme values. To this end, the following alternative parametrization of the GH distribution turns out to be useful. Setting $\xi = 1/\sqrt{1 + \delta\sqrt{\alpha^2 - \beta^2}}$ and $\chi = \xi\beta/\alpha$ renders the two parameters invariant under location and scale transformations. This is an immediate result of the following property of the GH distribution. If $X \sim \mathrm{GH}(\lambda, \alpha, \beta, \mu, \delta)$, then

$$a + bX \sim \mathrm{GH}\left(\lambda, \frac{\alpha}{|b|}, \frac{\beta}{|b|}, a + b\mu, \delta|b|\right) \ .$$

Furthermore, the parameters are restricted by $0 \le |\chi| < \xi < 1$, implying that they are located in the so-called *shape triangle* introduced by [27]. Figure 5 highlights the impact of the parameters in the GH distribution in terms of χ, ξ and λ. Obviously, χ controls the skewness and ξ the tailedness of the distribution. The impact of λ is not so clear-cut. The lower panels in Fig. 5 depict the pdfs for different values of λ whereby the first two moments and the values of χ and ξ are kept constant to show the "partial" influence.

As pointed out by [69], the limit distributions can be classified by the values of ξ and χ as well as by the values ϱ and ζ of a second location and scale invariant parametrization of the GH, given by $\varrho = \beta/\alpha$ and $\zeta = \delta\sqrt{\alpha^2 - \beta^2}$, as follows:

- $\xi = 1$ and $-1 \le \chi \le 1$: The resulting limit distributions depend here on the values of λ. Note, that in order to reach the boundary either $\delta \to 0$ or $|\beta| \to \alpha$.
  - For $\lambda > 0$ and $|\beta| \to \alpha$ no limit distribution exists, but for $\delta \to 0$ the GH distribution converges to the distribution of a variance gamma rv (see Subsect. "The Variance Gamma Distribution"). However, note that $|\beta| < \alpha$ implies $|\chi| < \xi$ and so the limit distribution is not valid in the corners. For these cases, the limit distributions are given by $\xi = |\chi|$ and $0 < \xi \le 1$, i. e. the next case.
  - For $\lambda = 0$ there exists no proper limit distribution.
  - For $\lambda < 0$ and $\delta \to 0$ no proper distribution exists but for $\beta \to \pm\alpha$ the limit distribution is given in [185] with pdf

$$\frac{2^{\lambda+1}\left(\delta^2 + (x-\mu)^2\right)^{(\lambda-1/2)/2}}{\sqrt{2\pi}\,\Gamma(-\lambda)\,\delta^{2\lambda}\alpha^{\lambda-1/2}} \cdot K_{\lambda-1/2}\left(\alpha\sqrt{\delta^2 + (x-\mu)^2}\right) \cdot \exp(\pm\alpha(x-\mu)) \ , \tag{20}$$

    which is the limit distribution of the corners, since $\beta = \pm\alpha$ is equivalent to $\chi = \pm\xi$. This distribution was recently called the GH skew t distribution by [1]. Assuming additionally that $\alpha \to 0$ and $\beta = \varrho\alpha \to 0$ with $\varrho \in (-1, 1)$ yields the limit distribution in between

$$\frac{\Gamma(-\lambda + 1/2)}{\sqrt{\pi\delta^2}\,\Gamma(-\lambda)}\left(1 + \frac{(x-\mu)^2}{\delta^2}\right)^{\lambda-1/2} \ ,$$

    which is the scaled and shifted Student's t distribution with -2λ degrees of freedom, expectation μ and variance $4\lambda^2\nu/(\nu - 2)$, for more details see Subsect. "The Student t Distribution".
- $\xi = |\chi|$ and $0 < \xi \le 1$: Except for the upper corner the limit distribution of the right boundary can be derived for

$$\beta = \alpha - \frac{\phi}{2}\,; \quad \alpha \to \infty\,; \quad \delta \to 0\,; \quad \alpha\delta^2 \to \tau$$

  with $\phi > 0$ and is given by the μ-shifted GIG distribution $\mathrm{GIG}\left(\lambda, \sqrt{\tau}, \sqrt{\phi}\right)$. The distribution for the left boundary is the same distribution but mirrored at $x = 0$. Note that the limit behavior does not depend

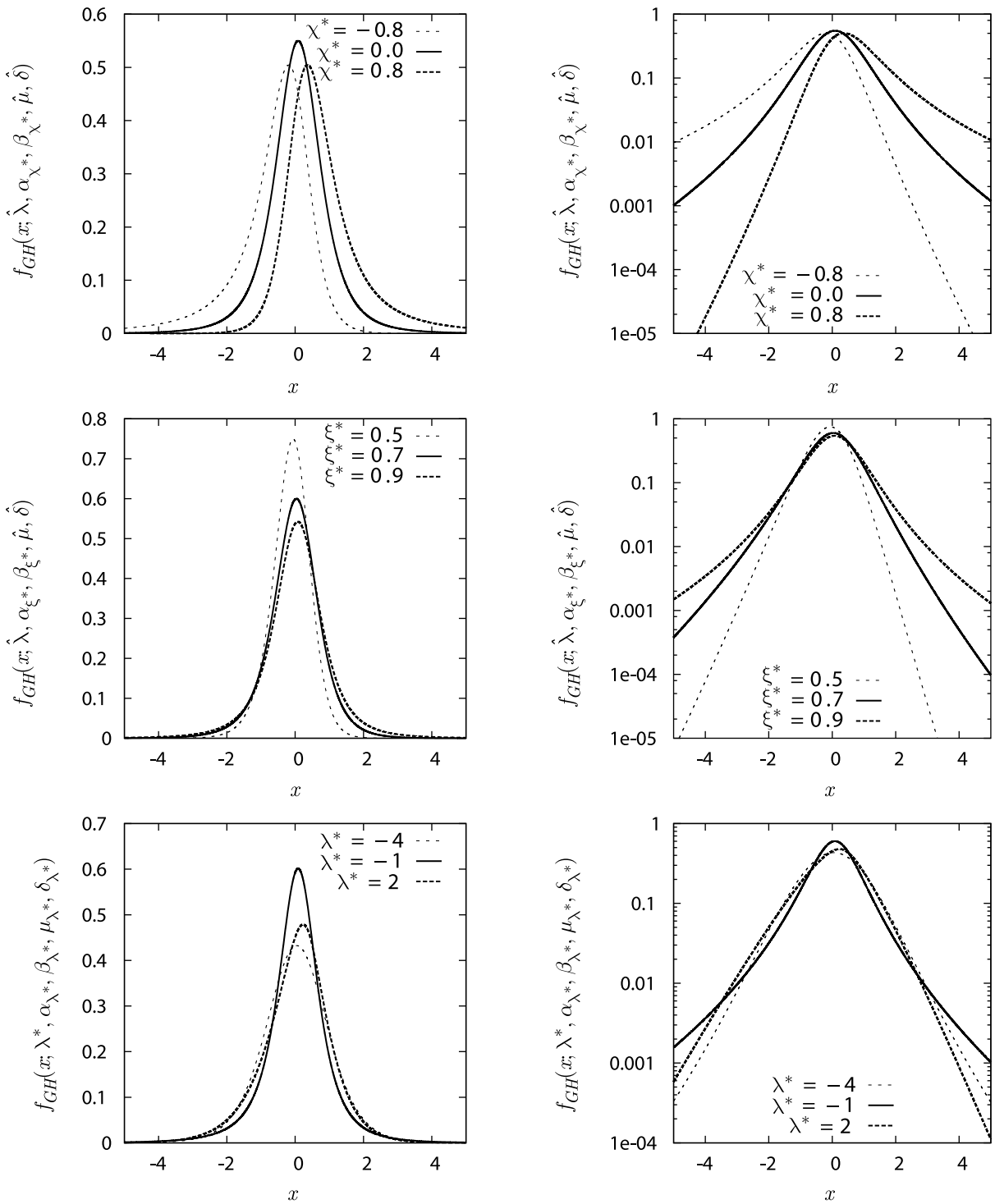

Financial Economics, Fat-Tailed Distributions, Figure 5

Density function (pdf) of the GH distribution for different parameter vectors. The *right panel* plots the log-densities to better visualize the tail behavior. The *upper and middle section* present the pdf for different values of χ and ξ. Note that these correspond to different values of α and β. The *lower panel* highlights the influence of λ if the first two moments, as well as χ and ξ, are held fixed. This implies that α, β, μ and δ have to be adjusted accordingly

on λ. However, to obtain the limit distributions for the left and right upper corners we have to distinguish for different values of λ. Recall that for the regime $\xi = 1$ and $-1 \leq \chi \leq 1$ the derivation was not possible.
  - For $\lambda > 0$ the limit distribution is a gamma distribution.
  - For $\lambda = 0$ no limit distribution exists.
  - For $\lambda < 0$ the limit distribution is the reciprocal gamma distribution.
- $\xi = \chi = 0$: This is the case for $\alpha \to \infty$ or $\delta \to \infty$. If only $\alpha \to \infty$ then the limit distribution is the Dirac measure concentrated in μ. If in addition $\delta \to \infty$ and $\delta/\alpha \to \sigma^2$ then the limit distribution is a normal distribution with mean $\mu + \beta\sigma^2$ and variance σ^2.

As pointed out by [185] applying the unrestricted GH distribution to financial data results in a very flat likelihood function especially for λ. This characteristic is illustrated in Fig. 6, which plots the maximum of the log likelihood for different values of λ using our sample of the S&P500 index returns. This implies that the estimate of λ is generally associated with a high standard deviation. As a consequence, rather than using the GH distribution directly, the finance literature primarily predetermines the value of λ, resulting in specific subclasses of the GH distribution, which are discussed in the sequel. However, it is still interesting to derive the general results in terms of the GH distribution (or the corresponding Lévy process) directly and to restrict only the empirical application to a subclass. For example [191] derived a diffusion process with GH marginal distribution, which is a generalization of the result of [33], who proposed a diffusion process with hyperbolic marginal distribution.

The Hyperbolic Distribution Recall, that the *hyperbolic* (HYP) distribution can be obtained as a special case of the GH distribution by setting $\lambda = 1$. Thus, all properties of the GH law can be applied to the HYP case. For instance the pdf of the HYP distribution is straightforwardly given by (19) setting $\lambda = 1$

$$\begin{aligned} f_{\mathrm{H}}(x;\alpha,\beta,\mu,\delta) &:= f_{\mathrm{GH}}(x;1,\alpha,\beta,\mu,\delta) \\ &= \frac{\sqrt{\alpha^2-\beta^2}}{2\alpha\delta K_1\left(\delta\sqrt{\alpha^2-\beta^2}\right)} \\ &\quad \cdot \exp\left(-\alpha\sqrt{\delta 2+(x-\mu)2}\right. \\ &\quad \left.+\beta(x-\mu)\right) , \end{aligned} \tag{21}$$

where $0 \leq |\beta| < \alpha$ are the shape parameter and $\mu \in \mathbb{R}$ and $\delta > 0$ are the location and scale parameter, respectively.

The distribution was applied to stock return data by [70,71,132] while [33] derived a diffusion model with marginal distribution belonging to the class of HYP distributions.

The Normal Inverse Gaussian Distribution The *normal inverse Gaussian* (NIG) distribution is given by the GH distribution with $\lambda = -1/2$ and has the following pdf

$$\begin{aligned} f_{\mathrm{NIG}}(x;\alpha,\beta,\mu,\delta) &:= f_{\mathrm{GH}}\left(x;-\frac{1}{2},\alpha,\beta,\mu,\delta\right) \\ &= \frac{\alpha\delta K_1\left(\alpha\sqrt{\delta^2+(x-\mu)^2}\right)}{\pi\sqrt{\delta^2+(x-\mu)^2}} \\ &\quad \cdot \exp\left(\delta\gamma+\beta(x-\mu)\right) \end{aligned} \tag{22}$$

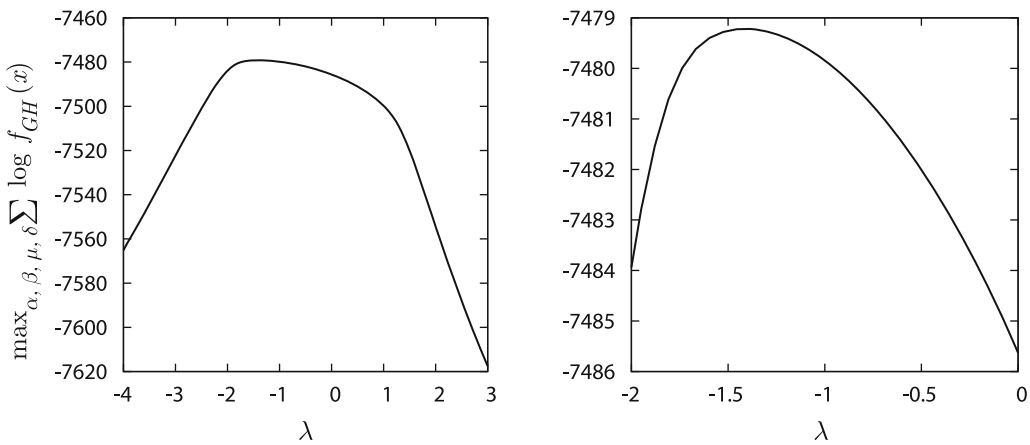

Financial Economics, Fat-Tailed Distributions, Figure 6
Partially maximized log likelihood, estimated maximum log likelihood values of the GH distribution for different values of λ

with $0 < |\beta| \leq \alpha$, $\delta > 0$ and $\mu \in \mathbb{R}$. The moments of a NIG distributed rv can be obtained from the moment generating function of the GH distribution (19) and are given by

$$\mathbb{E}[X] = \mu + \frac{\delta\beta}{\sqrt{\alpha^2 - \beta^2}} \quad \text{and} \quad \mathbb{V}[X] = \frac{\delta\alpha^2}{\sqrt{\alpha^2 - \beta^2}^3}$$
$$\mathbb{S}[X] = 3\frac{\beta}{\alpha\sqrt{\delta\sqrt{\alpha^2 - \beta^2}}} \quad \text{and} \quad \mathbb{K}[X] = 3\frac{\alpha^2 + 4\beta^2}{\delta\alpha^2\sqrt{\alpha^2 - \beta^2}} .$$

This distribution was heavily applied in financial economics for modeling the unconditional as well as the conditional return distribution, see e. g. [18,21,185]; as well as [10,19,114], respectively. Recently, [57] used the NIG distribution for modeling realized variance and found improved forecast performance relative to a Gaussian model. A more realistic modeling of the distributional properties is not only important for risk management or forecasting, but also for statistical inference. For example the efficient method of moments, proposed by [87] requires the availability of a highly accurate auxiliary model, which provide the objective function to estimate a more structural model. Recently, [39] provided such an auxiliary model, which uses the NIG distribution and realized variance measures.

Recall that for $\lambda = -1/2$ the mixing distribution is the inverse Gaussian distribution, which facilitates the generation of rvs. Hence, rvs with NIG distribution can be generated in the following way:

1. Draw a chi-square distributed rv C with one degree of freedom and a uniformly distributed rv over the interval $(0, 1)$ U
2. Compute
$$X_1 = \frac{\delta}{\gamma} + \frac{1}{2\delta\gamma}\left(\frac{\delta C}{\gamma} - \sqrt{4\delta^3 C/\gamma + \delta^2 C^2/\gamma^2}\right)$$
3. If $U < \delta/(\gamma(\delta/\gamma + X_1))$ return X_1 else return $\delta^2/(\gamma^2 X_1)$.

As pointed out by [187] the main difference between the HYP and NIG distribution: "Hyperbolic log densities, being hyperbolas, are strictly concave everywhere. Therefore they cannot form any sharp tips near $x = 0$ without loosing too much mass in the tails … In contrast, NIG log densities are concave only in an interval around $x = 0$, and convex in the tails." Moreover, [19] concludes, "It is, moreover, rather typical that asset returns exhibit tail behavior that is somewhat heavier than log linear, and this further strengthens the case for the NIG in the financial context".

The Student *t* Distribution Next to the alpha stable distribution *Student's t* (t thereafter) distribution has the longest history in financial economics. One reason is that although the non-normality of asset returns is widely accepted, there still exists some discussion on the exact tail behavior. While the alpha stable distribution implies extremely slowly decreasing tails for $\alpha \neq 2$, the t distribution exhibits power tails and existing moments up to (and excluding) ν. As such, the t distribution might be regarded as the strongest competitor to the alpha stable distribution, shedding also more light on the empirical tail behavior of returns. The pdf for the scaled and shifted t distribution is given by

$$f_t(x; \nu, \mu, \sigma) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)\,\sigma} \cdot \left(1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-(\nu+1)/2} \tag{23}$$

for $\nu > 0$, $\sigma > 0$ and $\mu \in \mathbb{R}$. For $\mu = 0$ and $\sigma = 1$ the well-known standard t distribution is obtained. The shifted and scaled t distribution can also be interpreted as a mean-variance mixture (18) with a reciprocal gamma distribution as a mixing distribution. The mean, variance, and kurtosis (3) are given by μ, $\sigma^2\nu/(\nu-2)$, and $3(\nu-2)/(\nu-4)$, provided that $\nu > 1$, $\nu > 2$, and $\nu > 4$, respectively. The tail behavior is

$$f_t(x, \nu, \mu, \sigma) \cong c x^{-\nu-1} .$$

The t distribution is one of the standard non-normal distributions in financial economics, see e. g. [36,38,184]. However, as the unconditional return distribution may exhibit skewness, a skewed version of the t distribution might be more adequate in some cases. In fact, several skewed t distributions were proposed in the literature, for a short overview see [1]. The following special form of the pdf was considered in [81,102]

$$f_{t,\mathrm{FS}}(x; \nu, \mu, \sigma, \beta) = \frac{2\beta}{\beta^2 + 1}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma(\nu/2)\sqrt{\pi\nu}\sigma} \cdot \left(1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\left(\frac{1}{\beta^2}\mathcal{I}(X \geq \mu) + \beta^2\mathcal{I}(x < \mu)\right)\right)^{-\frac{\nu+1}{2}}$$

with $\beta > 0$. For $\beta = 1$ the pdf reduces to the pdf of the usual symmetric scaled and shifted t distribution. Another

skewed t distribution was proposed by [116] with pdf

$$f_{t,\mathrm{JF}}(x,\nu,\mu,\sigma,\beta) = \frac{\Gamma(\nu+\beta)\,2^{1-\nu-\beta}}{\Gamma(\nu/2)\,\Gamma(\nu/2+\beta)\sqrt{\nu+\beta}\sigma} \cdot \left(1+\frac{\frac{x-\mu}{\sigma}}{\sqrt{\nu+\beta+\left(\frac{x-\mu}{\sigma}\right)^2}}\right)^{(\nu+1)/2} \cdot \left(1-\frac{\frac{x-\mu}{\sigma}}{\sqrt{\nu+\beta+\left(\frac{x-\mu}{\sigma}\right)^2}}\right)^{\beta+(\nu+1)/2}$$

for $\beta > -\nu/2$. Again, the usual t distribution can be obtained as a special case for $\beta = 0$. A skewed t distribution in terms of the pdf and cdf of the standard t distribution $f_t(x;\nu,0,1)$ and $F_t(x;\nu,0,1)$ is given by [13,43]

$$f_{t,\mathrm{AC}}(x;\nu,\mu,\sigma,\beta) = \frac{2}{\sigma} f_t\left(\frac{x-\mu}{\sigma},\nu,0,1\right) \cdot F_t\left(\beta\left(\frac{x-\mu}{\sigma}\right)\sqrt{\frac{\nu+1}{\nu+\left(\frac{x-\mu}{\sigma}\right)^2}},\nu+1,0,1\right)$$

for $\beta \in \mathbb{R}$.

Alternatively, a skewed t distribution can also be obtained as a limit distribution of the GH distribution. Recall that for $\lambda < 0$ and $\beta \to \alpha$ the limit distribution is given by (20) as

$$f_{t,\mathrm{GH}}(x;\lambda,\mu,\delta,\alpha) = \frac{2^{\lambda+1}\left(\delta^2+(x-\mu)^2\right)^{(\lambda-1/2)/2}}{\sqrt{2\pi}\,\Gamma(-\lambda)\,\delta^{2\lambda}\alpha^{\lambda-1/2}} \cdot K_{\lambda-1/2}\left(\alpha\sqrt{\delta^2+(x-\mu)^2}\right)\cdot\exp(\alpha(x-\mu)) \,.$$

for $\alpha \in \mathbb{R}$. The symmetric t distribution is obtained for $\alpha \to 0$. The distribution was introduced by [185] and a more detailed examination was recently given in [1].

The Variance Gamma Distribution The *variance gamma* (VG) distribution can be obtained as a mean-variance mixture with gamma mixing distribution. Note that the gamma distribution is obtained in the limit from the GIG distribution for $\lambda > 0$ and $\delta \to 0$. The pdf of the VG distribution is given by

$$f_{\mathrm{VG}}(x;\mu,\alpha,\beta,\lambda) := \lim_{\delta\to 0} f_{\mathrm{GH}}(x;\lambda,\alpha,\beta,\delta,\mu) = \frac{\gamma^{2\lambda}\,|x-\mu|^{\lambda-1/2}\,K_{\lambda-1/2}(\alpha\,|x-\mu|)}{\sqrt{\pi}\,\Gamma(\lambda)\,(2\alpha)^{\lambda-1/2}}\exp\beta(x-\mu) \,. \tag{24}$$

Note, the usual parameterization of the VG distribution

$$f^*_{\mathrm{VG}}(x;\sigma^*,\theta^*,\nu^*,\mu^*) = \frac{2\exp\left(\theta^*(x-\mu^*)/\sigma^{*2}\right)}{\nu^{*1/\nu^*}\sqrt{2\pi\sigma^{*2}}\,\Gamma(1/\nu^*)}\left(\frac{(x-\mu^*)^2}{2\sigma^{*2}/\nu^*+\theta^{*2}}\right)^{\frac{1}{2\nu^*}-\frac{1}{4}} \cdot K_{\frac{1}{\nu^*}-\frac{1}{2}}\left(\frac{\sqrt{(x-\mu^*)^2\left(2\sigma^{*2}/\nu^*+\theta^{*2}\right)}}{\sigma^{*2}}\right)$$

is different from the one used here, however the parameters can be transformed between these representations in the following way

$$\sigma^* = \sqrt{\frac{2\lambda}{\alpha^2-\beta^2}}\,; \quad \theta^* = \frac{2\beta\lambda}{\alpha^2-\beta^2}\,; \quad \nu^* = \frac{1}{\lambda}\,; \quad \mu^* = \mu\,.$$

This distribution was introduced by [149,150,151]. For $\lambda = 1$ (the HYP case) we obtain a skewed, shifted and scaled Laplace distribution with pdf

$$f_{\mathrm{L}}(x;\alpha,\beta,\mu) := \lim_{\delta\to 0} f_{\mathrm{GH}}(x;1,\alpha,\beta,\delta,\mu) = \frac{\alpha^2-\beta^2}{2\alpha}\exp(-\alpha\,|x-\mu|+\beta(x-\mu)) \,.$$

A generalization of the VG distribution to the so-called CGMY distribution was proposed by [48].

The Cauchy Distribution Setting $\lambda = -1/2$, $\beta \to 0$ and $\alpha \to 0$ the GH distribution converges to the *Cauchy* distribution with parameters μ and δ. Since the Cauchy distribution belongs to the class of symmetric alpha stable ($\alpha = 1$) and symmetric t distributions ($\nu = 1$) we refer to Subsect. "Alpha Stable and Related Distributions" and "The Student t Distribution" for a more detailed discussion.

The Normal Distribution For $\alpha \to \infty$, $\beta = 0$ and $\delta = 2\sigma^2$ the GH distribution converges to the *normal* distribution with mean μ and variance σ^2.

Finite Mixtures of Normal Distributions

The density of a (finite) mixture of k normal distributions is given by a linear combination of k Gaussian *component*

densities, i. e.,

$$f_{\mathrm{NM}}(x;\theta) = \sum_{j=1}^{k} \lambda_j \phi\left(x;\mu_j,\sigma_j^2\right) ,$$
$$\phi\left(x;\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) , \tag{25}$$

where $\theta = (\lambda_1,\ldots,\lambda_{k-1},\mu_1,\ldots,\mu_k,\sigma_1^2,\ldots,\sigma_k^2)$, $\lambda_k = 1-\sum_{j=1}^{k-1}\lambda_j$, $\lambda_j > 0$, $\mu_j \in \mathbb{R}$, $\sigma_j^2 > 0$, $j = 1,\ldots,k$, and $(\mu_i,\sigma_i^2) \neq (\mu_j,\sigma_j^2)$ for $i \neq j$. In (25), the λ_j, μ_j, and σ_j^2 are called the *mixing weights*, *component means*, and *component variances*, respectively.

Finite mixtures of normal distributions have been applied as early as 1886 in [174] to model leptokurtic phenomena in astrophysics. A voluminous literature exists, see [165] for an overview. In our discussion, we shall focus on a few aspects relevant for applications in finance. In this context, (25) arises naturally when the component densities are interpreted as different *market regimes*. For example, in a two-component mixture ($k = 2$), the first component, with a relatively high mean and small variance, may be interpreted as the bull market regime, occurring with probability λ_1, whereas the second regime, with a lower expected return and a greater variance, represents the bear market. This (typical) pattern emerges for the S&P500 returns, see Table 1. Clearly (25) can be generalized to accommodate non-normal component densities; e. g., [104] model stock returns using mixtures of generalized error distributions of the form (40). However, it may be argued that in this way much of the original appeal of (25), i. e., within-regime normality along with CLT arguments, is lost.

The moments of (25) can be inferred from those of the normal distribution, with mean and variance given by

$$\mathbb{E}[X] = \sum_{j=1}^{k} \lambda_j \mu_j \text{ , and}$$
$$\mathbb{V}[X] = \sum_{j=1}^{k} \lambda_j \left(\sigma_j^2 + \mu_j^2\right) - \left(\sum_{j=1}^{k} \lambda_j \mu_j\right)^2 , \tag{26}$$

respectively. The class of finite normal mixtures is very flexible in modeling the leptokurtosis and, if existent, skewness of financial data. To illustrate the first property, consider the *scale normal mixture*, where, in (25), $\mu_1 = \mu_2 = \cdots = \mu_k := \mu$. In fact, when applied to financial return data, it is often found that the market regimes differ mainly in their variances, while the component means are rather close in value, and often their differences are not significant statistically. This reflects the observation that excess kurtosis is a much more pronounced (and ubiquitous) property of asset returns than skewness. In the scale mixture case, the density is symmetric, but with higher peaks and thicker tails than the normal with the same mean and variance. To see this, note that $\sum_j(\lambda_j/\sigma_j) >$

Financial Economics, Fat-Tailed Distributions, Table 1
Maximum-likelihood parameter estimates of the iid model

Distribution	Parameters					Loglik
GH	$\hat\lambda$	$\hat\mu$	$\hat\alpha$	$\hat\beta$	$\hat\delta$	−7479.2
	−1.422	0.087	0.322	−0.046	1.152	
	(0.351)	(0.018)	(0.222)	(0.022)	(0.139)	
t_{GH}	$\hat\mu$	$\hat\delta$	$\hat\lambda$	$\hat\alpha$		−7479.7
	0.084	1.271	3.445	−0.041		
	(0.018)	(0.052)	(0.181)	(0.021)		
t_{JF}	$\hat\nu$	$\hat\mu$	$\hat\sigma$	$\hat\beta$		−7480.0
	3.348	0.098	0.684	0.091		
	(0.179)	(0.025)	(0.012)	(0.049)		
t_{AC}	$\hat\nu$	$\hat\mu$	$\hat\sigma$	$\hat\beta$		−7480.1
	3.433	0.130	0.687	−0.123		
	(0.180)	(0.042)	(0.013)	(0.068)		
t_{FS}	$\hat\nu$	$\hat\mu$	$\hat\sigma$	$\hat\beta$		−7480.3
	3.432	0.085	0.684	0.972		
	(0.180)	(0.020)	(0.012)	(0.017)		
Symmetric t	$\hat\nu$	$\hat\mu$	$\hat\sigma$			−7481.7
	3.424	0.056	0.684			
	(0.179)	(0.011)	(0.012)			
NIG	$\hat\mu$	$\hat\alpha$	$\hat\beta$	$\hat\delta$		−7482.0
	0.088	0.784	−0.048	0.805		
	(0.018)	(0.043)	(0.022)	(0.028)		
HYP	$\hat\mu$	$\hat\alpha$	$\hat\beta$	$\hat\delta$		−7499.5
	0.090	1.466	−0.053	0.176		
	(0.018)	(0.028)	(0.023)	(0.043)		
VG	$\hat\mu$	$\hat\alpha$	$\hat\beta$	$\hat\lambda$		−7504.2
	0.092	1.504	−0.054	1.115		
	(0.013)	(0.048)	(0.019)	(0.054)		
Alpha stable	$\hat\alpha$	$\hat\beta$	$\hat c$	$\hat\tau$		−7522.5
	1.657	−0.094	0.555	0.036		
	(0.024)	(0.049)	(0.008)	(0.015)		
Finite mixture ($k = 2$)	$\hat\lambda_1$	$\hat\mu_1$	$\hat\mu_2$	$\hat\sigma_1^2$	$\hat\sigma_2^2$	−7580.8
	0.872	0.063	−0.132	0.544	4.978	
	(0.018)	(0.012)	(0.096)	(0.027)	(0.530)	
Cauchy	$\hat\mu$	$\hat\sigma$				−7956.6
	0.060	0.469				
	(0.010)	(0.008)				
Normal	$\hat\mu$	$\hat\sigma$				−8168.9
	0.039	1.054				
	(0.014)	(0.010)				

Shown are maximum likelihood estimates for iid models with different assumptions about the distribution of the innovations. Standard errors are given in parentheses. "Loglik" is the value of the maximized log likelihood function.

$(\sum_j \lambda_j \sigma_j^2)^{-1/2} \Leftrightarrow (\sum_j \lambda_j \sigma_j^2)^{1/2} > [\sum_j (\lambda_j/\sigma_j)]^{-1}$. But $(\sum_j \lambda_j \sigma_j^2)^{1/2} > \sum_j \lambda_j \sigma_j > [\sum_j (\lambda_j/\sigma_j)]^{-1}$ by Jensen's and the arithmetic-harmonic mean inequality, respectively. This shows $f_{\mathrm{NM}}(\mu;\theta) > \phi(\mu;\mu,\sum_j \lambda_j \sigma_j^2)$, i. e., peakedness. Tailedness follows from the observation that the difference between the mixture and the mean-variance equivalent normal density is asymptotically dominated by the component with the greatest variance. Moreover, the densities of the scale mixture and the mean-variance equivalent Gaussian intersect exactly two times on both sides of the mean, so that the scale mixture satisfies the density crossing condition in Finucan's theorem mentioned in Sect. "Definition of the Subject" and observed in the center panel of Fig. 1. This follows from the fact that, if $a_1, \ldots, a_n$ and $\gamma_1 < \cdots < \gamma_n$ are real constants, and N is the number of real zeros of the function $\varphi(x) = \sum_i a_i e^{\gamma_i x}$, then $W - N$ is a non-negative even integer, where W is the number of sign changes in the sequence $a_1, \ldots, a_n$ [183]. Skewness can be incorporated into the model when the component means are allowed to differ. For example, if, in the two-component mixture, the high-variance component has both a smaller mean and mixing weight, then the distribution will be skewed to the left.

Because of their flexibility and the aforementioned economic interpretation, finite normal mixtures have been frequently used to model the unconditional distribution of asset returns [40,44,129,179], and they have become rather popular since the publication of Hamilton's [101] paper on Markov-switching processes, where the mixing weights are assumed to be time-varying according to a k-state Markov chain; see, e. g., [200] for an early contribution in this direction.

However, although a finite mixture of normals is a rather flexible model, its tails decay eventually in a Gaussian manner, and therefore, according to the discussion in Sect. "Empirical Evidence About the Tails", it may often not be appropriate to model returns at higher frequencies unconditionally. Nevertheless, when incorporated into a GARCH structure (see Sect. "Volatility Clustering and Fat Tails"), it provides a both useful and intuitively appealing framework for modeling the *conditional* distribution of asset returns, as in [5,96,97]. These papers also provide a discussion of alternative interpretations of the mixture model (25), as well as an overview over the extensive literature.

Empirical Comparison

In the following we empirically illustrate the adequacy of the various distributions discussed in the previous sections for modeling the unconditional return distribution. Table 1 presents the estimation results for the S&P500 index assuming iid returns. The log likelihood values clearly indicate the inadequacy of the normal, Cauchy and stable distributions. This is also highlighted in the upper panel of Fig. 7, which clearly shows that the tails of the Cauchy and stable distributions are too heavy, whereas those of the normal distribution are too weak. To distinguish the other distributions in more detail, the lower left panel is an enlarged display of the shadowed box in the upper panel. It illustrates nicely that the two component mixture, VG and HYP distribution exhibit semi-heavy tails, which are probably a little bit to weak for an adequate modeling as is indicated by the log likelihood values. Similarly, the two-component finite normal mixture, although much better than the normal, cannot keep up with most of the other models, presumably due to its essentially Gaussian tails. Although the pdf of the NIG distribution lies somewhere in between the pdfs of the HYP and the different t distributions, the log likelihood value clearly indicates that this distribution is in a statistical sense importantly closer to the t distributions. A further distinction between the other distributions including all kinds of t distributions and the GH distribution is nearly impossible, as can be seen from the lower right plot, which is an enlarged display of the lower left panel. The log likelihood values also do not allow for a clear distinction. Note also that the symmetric t distribution performs unexpectedly well. In particular, its log likelihood is almost indistinguishable from those of the skewed versions. Also note that, for all t distributions, the estimated tail index, ν, is close to 3.5, which is in accordance with the results from semiparametric tail estimation in Sect. "Empirical Evidence About the Tails".

The ranking of the distributions in terms of the log likelihood depends of course heavily on the dataset, and different return series may imply different rankings. However, Table 1 also highlights some less data-dependent results, which are more or less accepted in the literature, e. g., the tails of the Cauchy and stable distributions are too heavy, and those of the HYP and VG are too light for the unconditional distribution. This needs of course no longer be valid in a different modeling setup. Especially in a GARCH framework the conditional distribution don't need to imply such heavy tails because the model itself imposes fatter tails.

In Sect. "Application to Value-at-Risk", the comparison of the models will be continued on the basis of their ability the measure the Value-at-Risk, an important concept in risk management.

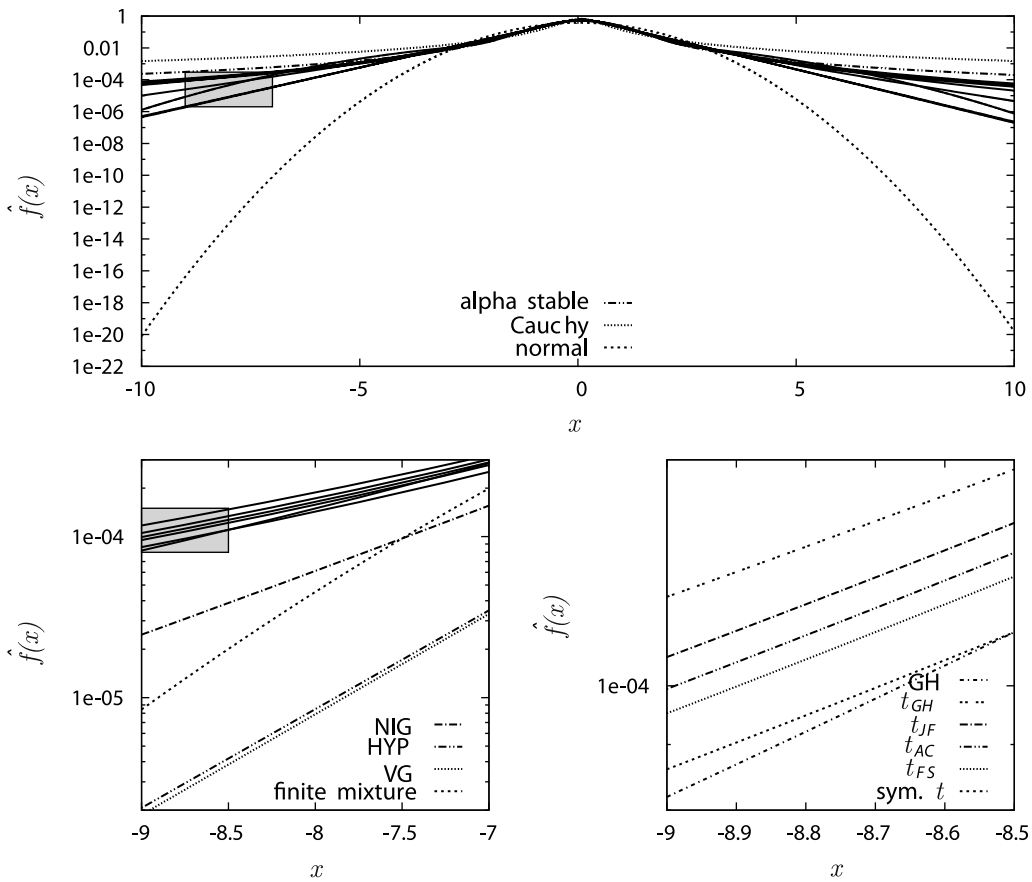

Financial Economics, Fat-Tailed Distributions, Figure 7
Plot of the estimated pdfs of the different return distributions assuming iid returns

Volatility Clustering and Fat Tails

It has long been known that the returns of most financial assets, although close to being unpredictable, exhibit significant dependencies in measures of volatility, such as absolute or squared returns. Moreover, the empirical results based on the recent availability of more precise volatility measures, such as the *realized volatility*, which is defined as the sum over the squared intradaily high-frequency returns (see, e. g., [7] and [23]), also point towards the same direction. In particular, the realized volatility has been found to exhibit strong persistence in its autocorrelation function, which shows a hyperbolic decay indicating the presence of long memory in the volatility process. In fact, this finding as well as other stylized features of the realized volatility have been observed across different data sets and markets and are therefore by now widely acknowledged and established in the literature. For a more detailed and originating discussion on the stylized facts of the high-frequency based volatility measures for stock returns and exchange returns we refer to [8,9], respectively.

The observed dependence of time-varying pattern of the volatility is usually referred to as *volatility clustering*. It is also apparent in the top panel of Fig. 1 and was already observed by Mandelbrot [155], who noted that "large changes tend to be followed by large changes—of either sign—and small changes tend to be followed by small changes". It is now well understood that volatility clustering can explain at least part of the fat-tailedness of the unconditional return distribution, even if the *conditional* distribution is Gaussian. This is also supported by the recent observation that if the returns are scaled by the realized volatility then the distribution of the resulting series is approximately Gaussian (see [9] and [8]). To illustrate, consider a time series $\{\epsilon_t\}$ of the form

$$\epsilon_t = \eta_t \sigma_t \,, \tag{27}$$

where $\{\eta_t\}$ is an iid sequence with mean zero and unit

variance, with η_t being independent of σ_t, so that σ_t^2 is the conditional variance of ϵ_t. With respect to the kurtosis measure $\mathbb{K}$ in (3), it has been observed by [108], and earlier by [31] in a different context, that, as long as σ_t^2 is not constant, Jensen's inequality implies $\mathbb{E}[\epsilon_t^4] = \mathbb{E}[\eta_t^4]\mathbb{E}[\sigma_t^4] > \mathbb{E}[\eta_t^4]\mathbb{E}^2[\sigma_t^2]$, so that the kurtosis of the unconditional distribution exceeds that of the innovation process. Clearly, $\mathbb{K}$ provides only limited information about the actual shape of the distribution, and more meaningful results can be obtained by specifying the dynamics of the conditional variance, σ_t^2. A general useful result [167] for analyzing the tail behavior of processes such as (27) is that, if ξ_t and σ_t are independent non-negative random variables with σ_t regularly varying, i. e., $P(\sigma_t > x) = L(x)x^{-\alpha}$ for some slowly varying L, and $\mathbb{E}[\xi_t^{\alpha+\delta}] < \infty$ for some $\delta > 0$, then $\xi_t\sigma_t$ is likewise regularly varying with tail index α, namely,

$$P(\xi_t\sigma_t > x) \cong \mathbb{E}\left[\xi^\alpha\right] P(\sigma_t > x) \quad \text{as } x \to \infty\,. \tag{28}$$

Arguably the most popular model for the evolution of σ_t^2 in (27) is the generalized autoregressive conditional heteroskedasticity process of orders p and q, or GARCH(p, q), as introduced by [37,73], which specifies the conditional variance as

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2\,. \tag{29}$$

The case $p = 0$ in (29) is referred to as an ARCH(q) process, which is the specification considered in [73]. To make sure that the conditional variance remains positive for all t, appropriate restrictions have to be imposed on the parameters in (29), i. e., α_i, $i = 0, \ldots, q$, and β_i, $i = 1, \ldots, p$. It is clearly sufficient to assume that α_0 is positive and all the other parameters are non-negative, as in [37], but these conditions can be relaxed substantially if $p, q > 0$ and $p + q > 2$ [173].

(27) and (29) is covariance stationary iff

$$P(z) = z^m - \sum_{i=1}^{m} (\alpha_i + \beta_i) z^{m-i} = 0 \Rightarrow |z| < 1\,, \tag{30}$$

where $m = \max\{p, q\}$, and $\alpha_i = 0$ for $i > q$, and $\beta_i = 0$ for $i < p$, which boils down to $\sum_i \alpha_i + \sum_i \beta_i < 1$ in case the non-negativity restrictions of [37] are imposed. The situation $\sum_i \alpha_i + \sum_i \beta_i = 1$ is referred to as an integrated GARCH (IGARCH) model, and in applications it is often found that the sum is just below unity. This indicates a high degree of volatility persistence, but the interpretation of this phenomenon is not so clear-cut [166]. If (30) holds, the unconditional variance of the process defined by (27) and (29) is given by

$$\mathbb{E}\left[\epsilon_t^2\right] = \frac{\alpha_0}{1 - \sum_{i=1}^{q} \alpha_i - \sum_{i=1}^{p} \beta_i}\,. \tag{31}$$

In practice, the GARCH(1,1) specification is of particular importance, and it will be the focus of our discussion too, i. e., we shall concentrate on the model (27) with

$$\sigma_t^2 = \alpha_0 + \left(\alpha_1 \eta_{t-1}^2 + \beta_1\right) \sigma_{t-1}^2\,, \qquad \alpha_0 > 0\,, \quad \alpha_1 > 0\,, \quad 1 > \beta_1 \geq 0\,. \tag{32}$$

The case $\alpha_1 = 0$ corresponds to a model with constant variance, which is of no interest in the current discussion.

An interesting property of the GARCH process is that its unconditional distribution is fat-tailed even with light-tailed (e. g., Gaussian) innovations, i. e., the distributional properties of the returns will not reflect those of the innovation (news) process. This has been known basically since [37,73], who showed that, even with normally distributed innovations, (G)ARCH processes do not have all their moments finite. For example, for the GARCH(1,1) model, [37] showed that, with $m \in \mathbb{N}$, the unconditional $(2m)$th moment of ϵ_t in (27) is finite if and only if

$$\mathbb{E}\left[(\alpha_1 \eta_t^2 + \beta_1)^m\right] < 1\,, \tag{33}$$

which, as long as $\alpha_1 > 0$, will eventually be violated for all practically relevant distributions. The argument in [37] is based on the relation

$$\mathbb{E}\left[\sigma_t^{2m}\right] = \sum_{i=0}^{m} \binom{m}{i} \alpha_0^i \mathbb{E}\left[\left(\alpha_1 \eta_{t-1}^2 + \beta_1\right)^{m-i}\right] \mathbb{E}\left[\sigma_{t-1}^{2(m-i)}\right]\,, \tag{34}$$

which follows from (32). The coefficient of $\mathbb{E}\left[\sigma_{t-1}^{2m}\right]$ on the right-hand side of (34) is just the expression appearing in (33), and consequently the $(2m)$th unconditional moment cannot be finite if this exceeds unity. The heavy-tailedness of the GARCH process is sometimes also exemplified by means of its unconditional kurtosis measure (3), which is finite for the GARCH(1,1) model with Gaussian innovations iff $3\alpha_1^2 + 2\alpha_1\beta_1 + \beta_1^2 < 1$. Writing (34) down for $m = 2$, using (31) and substituting into (3) gives

$$\mathbb{K}[\epsilon_t] = \frac{3\left[1 - (\alpha_1 + \beta_1)^2\right]}{1 - (\alpha_1 + \beta_1)^2 - 2\alpha_1^2} > 3\,,$$

as $\mathbb{E}\left[\epsilon_t^4\right] = 3\mathbb{E}\left[\sigma_t^4\right]$. [73] notes that "[m]any statistical procedures have been designed to be robust to large errors, but … none of this literature has made use of the

fact that temporal clustering of outliers can be used to predict their occurrence and minimize their effects. This is exactly the approach taken by the ARCH model". Conditions for the existence of and expressions for higher-order moments of the GARCH(p, q) model can be found in [50,105,122,139]. The relation between the conditional and unconditional kurtosis of GARCH models was investigated in [15], see also [47] for related results.

A more precise characterization of the tails of GARCH processes has been developed by applying classical results about the tail behavior of solutions of stochastic difference equations as, for example, in [124]. We shall continue to concentrate on the GARCH(1,1) case, which admits relatively explicit results, and which has already been written as a first-order stochastic difference equation in (32). Iterating this,

$$\sigma_t^2 = \sigma_0^2 \prod_{i=1}^{t} \left(\alpha_1 \eta_{t-i}^2 + \beta_1\right) + \alpha_0 \left[1 + \sum_{k=1}^{t-1} \prod_{i=1}^{k} \left(\alpha_1 \eta_{t-i}^2 + \beta_1\right)\right] . \tag{35}$$

Nelson [171] has shown that the GARCH(1,1) process (32) has a strictly stationary solution, given by

$$\sigma_t^2 = \alpha_0 \left[1 + \sum_{k=1}^{\infty} \prod_{i=1}^{k} \left(\alpha_1 \eta_{t-i}^2 + \beta_1\right)\right] , \tag{36}$$

if and only if

$$\mathbb{E}\left[\log\left(\alpha_1 \eta_t^2 + \beta_1\right)\right] < 0 . \tag{37}$$

The keynote of the argument in [171] is the application of the strong law of large numbers to the terms of the form $\prod_{i=1}^{k}(\alpha_1 \eta_{t-i}^2 + \beta_1) = \exp\{\sum_1^k \log\left(\alpha_1 \eta_{t-i}^2 + \beta_1\right)\}$ in (35), revealing that (35) converges almost surely if (37) holds. Note that $\mathbb{E}\left[\log\left(\alpha_1 \eta_t^2 + \beta_1\right)\right] < \log \mathbb{E}\left[\alpha_1 \eta_t^2 + \beta_1\right] = \log\left(\alpha_1 + \beta_1\right)$, i. e., stationary GARCH processes need not be covariance stationary. Using (36) and standard moment inequalities, [171] further established that, in case of stationarity, $\mathbb{E}\left[|\epsilon_t|^p\right]$, $p > 0$, is finite if and only if $\mathbb{E}[(\alpha_1 \eta_t^2 + \beta_1)^{p/2}] < 1$, which generalizes (33) to noninteger moments. It may now be supposed, and, building on the results of [90,124], has indeed been established by [167], that the tails of the marginal distribution of ϵ_t generated by a GARCH(1,1) process decay asymptotically in a Pareto-type fashion, i. e.,

$$P(|\epsilon_t| > x) \cong c x^{-\alpha} \quad \text{as } x \to \infty , \tag{38}$$

where the tail index α is the unique positive solution of the equation

$$h(\alpha) := \mathbb{E}\left[\left(\alpha_1 \eta_t^2 + \beta_1\right)^{\alpha/2}\right] = 1 . \tag{39}$$

This follows from (28) along with the result that the tails of σ_t^2 and σ_t are asymptotically Paretian with tail indices $\alpha/2$ and α, respectively. For a discussion of technical conditions, see [167]. [167] also provides an expression for the constant c in (38), which is difficult to calculate explicitly, however. For the ARCH(1) model with Gaussian innovations, (39) becomes $(2\alpha_1)^{\alpha/2}\, \Gamma\left[(\alpha+1)/2\right]/\sqrt{\pi} = 1$, which has already been obtained by [63] and was foreshadowed in the work of [168]. The results reported above have been generalized in various directions, with qualitatively similar conclusions. The GARCH(p, q) case is treated in [29], while [140,141] consider various extensions of the standard GARCH(1,1) model.

Although the *unconditional* distribution of a GARCH model with Gaussian innovations has genuinely fat tails, it is often found in applications that the tails of empirical return distributions are even fatter than those implied by fitted Gaussian GARCH models, indicating that the *conditional* distribution, i. e., the distribution of η_t in (27), is likewise fat-tailed. Therefore, it has become standard practice to assume that the innovations η_t are also heavy tailed, although it has been questioned whether this is the best modeling strategy [199]. The most popular example of a heavy tailed innovation distribution is certainly the t considered in Subsect. "The Student t Distribution", which was introduced by [38] into the GARCH literature. Some authors have also found it beneficial to let the degrees of freedom parameter ν in (23) be time-varying, thus obtaining time-varying conditional fat-tailedness [45].

In the following, we shall briefly discuss a few GARCH(1,1) estimation results for the S&P500 series in order to compare the tails implied by these models with those from the semiparametric estimation procedures in Sect. "Empirical Evidence About the Tails". As distributions for the innovation process $\{\eta_t\}$, we shall consider the Gaussian, t, and the generalized error distribution (GED), which was introduced by [172] into the GARCH literature, see [128] for a recent contribution and asymmetric extensions. It has earlier been used in an unconditional context by [94] for the S&P500 returns. The density of the GED with mean zero and unit variance is given by

$$f_{\text{GED}}(x;\nu) = \frac{\lambda \nu}{2^{1/\nu+1}\Gamma(1/\nu)} \exp\left(-\frac{|\lambda x|^\nu}{2}\right) , \quad \nu > 0 , \tag{40}$$

where $\lambda = 2^{1/\nu}\sqrt{\Gamma(3/\nu)/\Gamma(1/\nu)}$. Parameter ν in (40) controls the thickness of the tails. For $\nu = 2$, we get the normal distribution, and a leptokurtic shape is obtained for $\nu < 2$. In the latter case, the tails of (40) are therefore thicker than those of the Gaussian, but they are not fat in

the Pareto sense. However, even if one argues for Pareto tails of return distributions, use of (40) may be appropriate as a conditional distribution in GARCH models, because the power law already accompanies the volatility dynamics. To make the estimates of the parameter α_1 in (32) comparable, we also use the unit variance version of the t, which requires multiplying X in (23) by $\sqrt{(\nu-2)/\nu}$. Returns are modeled as $r_t = \mu + \epsilon_t$, where μ is a constant mean and ϵ_t is generated by (27) and (32). Parameter estimates, obtained by maximum-likelihood estimation, are provided in Table 2. In addition to the GARCH parameters in (32) and the shape parameters of the innovation distributions, Table 2 reports the log likelihood values and the implied tail indices, $\hat{\alpha}$, which are obtained by solving (39) numerically. First note that all the GARCH models have considerably higher likelihood values than the iid models in Table 1, which highlights the importance of accounting for conditional heteroskedasticity. We can also conclude that the Gaussian assumption is still inadequate as a conditional distribution in GARCH models, as both the t and the GED achieve significantly higher likelihood values, and their estimated shape parameters indicate pronounced non-normalities. However, the degrees of freedom parameter of the t, ν, is somewhat increased in comparison to Table 1, as part of the leptokurtosis is now explained by the GARCH effects.

Compared to the nonparametric tail estimates obtained in Sect. "Empirical Evidence About the Tails", the tail index implied by the Gaussian GARCH(1,1) model turns out to be somewhat too high, while those of the more flexible models are both between 3 and 4 and therefore more in line with what has been found in Sect. "Empirical Evidence About the Tails". However, for all three models, the confidence intervals for α, as obtained from 1,000 simulations from the respective estimated GARCH processes, are rather wide, so that we cannot conclusively rule out the existence of the unconditional fourth (and even fifth) moment. The width of the confidence intervals reflects the fact that the implied tail indices are very sensitive to small variations in the underlying GARCH parameters. For example, if, in the GARCH model with conditional normality, we replace the estimate $\hat{\beta}_1 = 0.911$ with 0.9, the implied tail index is 7.31, and with $\beta_1 = 0.92$, we get $\alpha = 2.05$, which is close to an infinite variance. The situation is depicted in Fig. 8, showing $h(\alpha)$ in (39) for the different values of β_1. The shape of h follows generally from $h(0) = 1$, $h'(0) < 0$ by (37), $h'' > 0$, i. e., h is convex, and $\lim_{\alpha\to\infty} h(\alpha) = \infty$ as long as $P\left[\left(\alpha_1\eta_t^2 + \beta_1\right) > 1\right] > 0$, so that $h(\alpha) = 1$ has a unique positive solution. Note that both 0.9 and 0.92 are covered by $0.911 \pm 2 \times 0.009$, i. e., a 95% confidence interval for β_1. This shows that the GARCH-implied tail indices are rather noisy.

Alternatively, we may avoid precise assumptions about the distribution of the innovation process $\{\eta_t\}$ and rely on quasi maximum-likelihood results [138]. That is, we estimate the innovations by $\hat{\eta}_t = \hat{\epsilon}_t/\hat{\sigma}_t$, $t = 1, \ldots, 5,550$, where $\{\hat{\sigma}_t\}$ is the sequence of conditional standard deviations implied by the estimated Gaussian GARCH model, and then solve the sample analogue of (39), i. e., $T^{-1}\sum_{t=1}^{T}(\hat{\alpha}_1\hat{\eta}_t^2 + \hat{\beta}_1)^{\alpha/2} = 1$, a procedure theoretically justified in [32]. Doing so, we obtain $\hat{\alpha} = 2.97$, so that we recover the "universal cubic law". However, the 95% confidence interval, calculated from 1,000 GARCH simulations, where the innovation sequences are obtained by sampling with replacement from the $\hat{\eta}_t$-series, is (1.73, 4.80), which is still reconcilable with a finite fourth moment, and even with an infinite second moment. These results clearly underline the caveat brought out by [72] (p. 349), that "[t]here is no free lunch when it comes to [tail index] estimation".

Financial Economics, Fat-Tailed Distributions, Table 2
GARCH parameter estimates

Distribution	$\hat{\mu}$	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\hat{\nu}$	$\hat{\alpha}$	Loglik
Normal	0.059	0.012	0.080	0.911	—	4.70	−7271.7
	(0.011)	(0.002)	(0.008)	(0.009)		(3.20, 7.22)	
GED	0.063	0.007	0.058	0.936	1.291	3.95	−7088.2
	(0.010)	(0.002)	(0.007)	(0.008)	(0.031)	(2.52, 6.95)	
Symmetric t	0.063	0.006	0.051	0.943	6.224	3.79	−7068.1
	(0.010)	(0.002)	(0.006)	(0.007)	(0.507)	(2.38, 5.87)	

Shown are maximum-likelihood estimation results for GARCH(1,1) models, as given by (27) and (32), with different assumptions about the distribution of the innovations η_t in (27). Standard errors for the model parameters and 95% confidence intervals for the implied tail indices, $\hat{\alpha}$, are given in parentheses. "Loglik" is the value of the maximized log likelihood function.

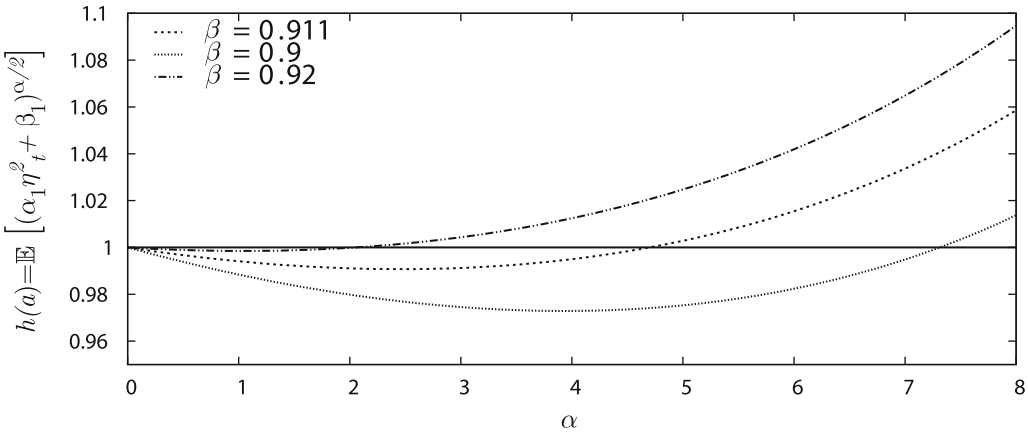

Financial Economics, Fat-Tailed Distributions, Figure 8
The figure displays the function $h(\alpha)$, as defined in (39), for Gaussian η_t, $\alpha = 0.0799$ and various values of β_1. Note that $\hat{\alpha}_1 = 0.0799$ and $\hat{\beta}_1 = 0.911$ are the maximum likelihood estimates for the S&P500 returns, as reported in Table 2

Application to Value-at-Risk

In this section, we compare the models discussed in Sects. "Some Specific Distributions" and "Volatility Clustering and Fat Tails" on an economic basis by employing the Value-at-Risk (VaR) concept, which is a widely used measure to describe the downside risk of a financial position both in industry and in academia [118]. Consider a time series of portfolio returns, r_t, and an associated series of ex-ante VaR measures with target probability ξ, $\mathrm{VaR}_t(\xi)$. The $\mathrm{VaR}_t(\xi)$ implied by a model $\mathcal{M}$ is defined by

$$\Pr_{t-1}^{\mathcal{M}}\left(r_t < -\mathrm{VaR}_t(\xi)\right) = \xi\,, \tag{41}$$

where $\Pr_{t-1}^{\mathcal{M}}(\cdot)$ denotes a probability derived from model $\mathcal{M}$ using the information up to time $t-1$, and the negative sign in (41) is due to the convention of reporting VaR as a positive number. For an appropriately specified model, we expect $100 \times \xi\%$ of the observed return values not to exceed the (negative of the) respective VaR forecast. Thus, to assess the performance of the different models, we examine the percentage shortfall frequencies,

$$U_\xi = 100 \times \frac{x}{T} = 100 \times \hat{\xi}\,, \tag{42}$$

where T denotes the number of forecasts evaluated, x is the observed shortfall frequency, i. e., the number of days for which $r_t < -\mathrm{VaR}_t(\xi)$, and $\hat{\xi} = x/T$ is the empirical shortfall probability. If $\hat{\xi}$ is significantly less (higher) than ξ, then model $\mathcal{M}$ tends to overestimate (underestimate) the risk of the position. In the present application, in order to capture even the more extreme tail region, we focus on the target probabilities $\xi = 0.001$, 0.0025, 0.005, 0.01, 0.025, and 0.05.

To formally test whether a model correctly estimates the risk (according to VaR) inherent in a given financial position, that is, whether the empirical shortfall probability, $\hat{\xi}$, is statistically indistinguishable from the nominal shortfall probability, ξ, we use the likelihood ratio test [133]

$$\mathrm{LRT}_{\mathrm{VaR}} = -2\left\{x \log\frac{\xi}{\hat{\xi}} + (T-x)\log\frac{1-\xi}{1-\hat{\xi}}\right\} \overset{\text{asy}}{\sim} \chi^2(1)\,. \tag{43}$$

On the basis of the first 1,000 return observations, we calculate one-day-ahead VaR measures based on parameter estimates obtained from an expanding data window, where the parameters are updated every day. Thus we get, for each model, 4,550 one-day-ahead out-of-sample VaR measures.

Table 3 reports the realized one-day-ahead percentage shortfall frequencies for the different target probabilities, ξ, as given above. The upper panel of the table shows the results for the unconditional distributions discussed in Sect. "Some Specific Distributions". The results clearly show that the normal distribution strongly *underestimates* ($\hat{\xi} > \xi$) the downside risk for the lower target probabilities, while the Cauchy as well as the alpha stable distributions tend to significantly *overestimate* ($\hat{\xi} < \xi$) the tails. This is in line with what we have observed from the empirical density plots presented in Fig. 7, which, in contrast to the out-of-sample VaR calculations, are based on estimates for the entire sample. Interestingly, the finite normal mixture distribution also tends to overestimate the risk at the lower VaR levels, leading to a rejection of cor-

Financial Economics, Fat-Tailed Distributions, Table 3
Backtesting Value-at-Risk measures

Unconditional Distributional Models						
Distribution	$U_{0.001}$	$U_{0.0025}$	$U_{0.005}$	$U_{0.01}$	$U_{0.025}$	$U_{0.05}$
GH	0.04	0.11**	0.24***	0.73*	2.70	5.89***
t_{GH}	0.07	0.11**	0.22***	0.75*	2.75	5.96***
t_{JF}	0.04	0.11**	0.31**	0.88	2.64	5.32
t_{AC}	0.04	0.11**	0.26**	0.84	2.48	5.16
t_{FS}	0.07	0.13*	0.33**	0.95	2.77	5.38
Symmetric t	0.07	0.15	0.31**	0.92	3.08**	6.35***
NIG	0.07	0.15	0.26**	0.70**	2.35	5.34
HYP	0.13	0.24	0.51	0.95	2.50	5.16
VG	0.13	0.24	0.51	0.92	2.46	5.10
Alpha stable	0.04	0.11**	0.33**	0.75*	2.44	4.90
Finite mixture ($k = 2$)	0.04	0.07***	0.11***	0.37***	2.99**	6.40***
Cauchy	0.00***	0.00***	0.00***	0.00***	0.09***	0.88***
Normal	0.48***	0.64***	0.97***	1.36**	2.44	4.02***

GARCH(1,1) Models						
Distribution	$U_{0.001}$	$U_{0.0025}$	$U_{0.005}$	$U_{0.01}$	$U_{0.025}$	$U_{0.05}$
Normal	0.40***	0.66***	0.92***	1.36**	2.95*	4.57
GED	0.20*	0.33	0.44	0.79	2.48	4.79
Symmetric t	0.11	0.26	0.40	0.92	2.86	5.45

The table shows the realized one-day-ahead percentage shortfall frequencies, U_ξ, for given target probabilities, ξ, as defined in (42). Asterisks *, ** and *** indicate significance at the 10%, 5% and 1% levels, respectively, as obtained from the likelihood ratio test (43).

rect coverage for almost all target probabilities. In contrast, the HYP distribution, whose empirical tails have been very close to those of the normal mixture in-sample (see Fig. 7), nicely reproduces the target probabilities, as does the VG distribution.

Similarly to the log likelihood results presented in Subsect. "Empirical Comparison" the Value-at-Risk evaluation does not allow for a clear distinction between the different t distributions, the GH and the NIG distribution. Similar to the Cauchy and the stable, they all tend to overestimate the more extreme target probabilities, while they imply too large shortfall probabilities at the five percent quantile.

The fact that most unconditional distributional models tend to overestimate the risk at the lower target probabilities may be due to our use of an expanding data window and the impact of the "Black Monday", where the index decreased by more than 20%, at the beginning of our sample period. In this regard, the advantages of accounting for time-varying volatility via a GARCH(1,1) structure may become apparent, as this model allows the more recent observations to have much more impact on the conditional density forecasts.

In fact, by inspection of the results for the GARCH models, as reported in the lower part of Table 3, it turns out that the GARCH(1,1) model with a normal distribution strongly underestimates the empirical shortfall probabilities at all levels except the largest (5%). However, considering a GED or t distribution for the return innovations within the GARCH model provides accurate estimates of downside risks.

To further discriminate between the GARCH processes and the iid models, tests for *conditional* coverage may be useful, which are discussed in the voluminous VaR literature (e. g., [53]).

Finally, we point out that the current application is necessarily of an illustrative nature. In particular, if the data generating process is not constant but evolves slowly over time and/or is subject to abrupt structural breaks, use of a rolling data window will be preferred to an expanding window.

Future Directions

As highlighted in the previous sections, there exists a plethora of different and well-established approaches

for modeling the tails of univariate financial time series. However, on the multivariate level the number of models and distributions is still very limited, although the joint modeling of multiple asset returns is crucial for portfolio risk management and allocation decisions. The problem is then to model the dependencies between financial assets. In the literature, this problem has been tackled, for example, by means of multivariate extensions of the mean-variance mixture (18) [19], multivariate GARCH models [30], regime-switching models [11], and copulas [51]. The problem is particularly intricate if the number of assets to be considered is large, and much work remains to understand and properly model their dependence structure.

It is also worth mentioning that the class of GARCH processes, due to its interesting conditional and unconditional distributional properties, has been adopted, for example, in the signal processing literature [3,52,55], and it is to be expected that it will be applied in other fields in the future.

Bibliography

Primary Literature

1. Aas K, Haff IH (2006) The generalized hyperbolic skew Student's *t*-distribution. J Financial Econom 4:275–309
2. Abhyankar A, Copeland LS, Wong W (1995) Moment condition failure in high frequency financial data: Evidence from the S&P 500. Appl Econom Lett 2:288–290
3. Abramson A, Cohen I (2006) State smoothing in Markov-switching time-frequency GARCH models. IEEE Signal Process Lett 13:377–380
4. Akgiray V, Booth GG (1988) The stable-law model of stock returns. J Bus Econ Stat 6:51–57
5. Alexander C, Lazar E (2006) Normal mixture GARCH(1,1). Applications to exchange rate modelling. J Appl Econom 21:307–336
6. Alexander SS (1961) Price movements in speculative markets: Trends or random walks. Ind Manag Rev 2:7–25
7. Andersen TG, Bollerslev T (1998) Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. Int Econ Rev 39:885–905
8. Andersen TG, Bollerslev T, Diebold FX, Ebens H (2001) The distribution of realized stock return volatility. J Financial Econ 61:43–76
9. Andersen TG, Bollerslev T, Diebold FX, Labys P (2001) The distribution of realized exchange rate volatility. J Am Stat Assoc 96:42–55
10. Andersson J (2001) On the normal inverse Gaussian stochastic volatility model. J Bus Econ Stat 19:44–54
11. Ang A, Chen J (2002) Asymmetric correlations of equity portfolios. J Financial Econ 63:443–494
12. Atkinson AC (1982) The simulation of generalized inverse Gaussian and hyperbolic random variables. SIAM J Sci Stat Comput 3:502–515
13. Azzalini A, Capitanio A (2003) Distributions generated by perturbation of symmetry with emphasis on a multivariate skew *t*-distribution. J R Stat Soc Ser B 65:367–389
14. Bachelier L (1964) Theory of speculation. In: Cootner PH (ed) The random character of stock market prices. MIT Press, Cambridge, pp 17–75
15. Bai X, Russell JR, Tiao GC (2003) Kurtosis of GARCH and stochastic volatility models with non-normal innovations. J Econom 114:349–360
16. Balanda KP, MacGillivray HL (1988) Kurtosis: A critical review. Am Stat 42:111–119
17. Barndorff-Nielsen OE (1977) Exponentially decreasing distributions for the logarithm of particle size. Proc R Soc Lond Ser A 353:401–419
18. Barndorff-Nielsen OE (1988) Processes of normal inverse Gaussian type. Finance Stoch 2:41–68
19. Barndorff-Nielsen OE (1997) Normal inverse Gaussian distributions and stochastic volatility modelling. Scand J Stat 24:1–13
20. Barndorff-Nielsen OE, Halgreen C (1977) Infinite divisibility of the hyperbolic and generalized inverse Gaussian distributions. Probab Theory Relat Fields 38:309–311
21. Barndorff-Nielsen OE, Prause K (2001) Apparent scaling. Finance Stoch 5:103–113
22. Barndorff-Nielsen OE, Shephard N (2001) Normal modified stable processes. Theory Prob Math Stat 65:1–19
23. Barndorff-Nielsen OE, Shephard N (2002) Estimating quadratic variation using realized variance. J Appl Econom 17:457–477
24. Barndorff-Nielsen OE, Shephard N (2007) Financial volatility in continuous time. Cambridge University Press
25. Barndorff-Nielsen OE, Stelzer R (2005) Absolute moments of generalized hyperbolic distributions and approximate scaling of normal inverse Gaussian Lévy processes. Scand J Stat 32:617–637
26. Barndorff-Nielsen OE, Kent J, Sørensen M (1982) Normal variance-mean mixtures and z distributions. Int Stat Rev 50:145–159
27. Barndorff-Nielsen OE, Blæsild P, Jensen JL, Sørensen M (1985) The fascination of sand. In: Atkinson AC, Fienberg SE (eds) A celebration of statistics. Springer, Berlin, pp 57–87
28. Barnea A, Downes DH (1973) A reexamination of the empirical distribution of stock price changes. J Am Stat Assoc 68:348–350
29. Basrak B, Davis RA, Mikosch T (2002) Regular variation of GARCH processes. Stoch Process Appl 99:95–115
30. Bauwens L, Laurent S, Rombouts JVK (2006) Multivariate GARCH models: A survey. J Appl Econom 21:79–109
31. Beale EML, Mallows CL (1959) Scale mixing of symmetric distributions with zero mean. Ann Math Stat 30:1145–1151
32. Berkes I, Horváth L, Kokoszka P (2003) Estimation of the maximal moment exponent of a GARCH(1,1) sequence. Econom Theory 19:565–586
33. Bibby BM, Sørensen M (1997) A hyperbolic diffusion model for stock prices. Finance Stoch 1:25–41
34. Bingham NH, Goldie CM, Teugels JL (1987) Regular variation. Cambridge University Press
35. Blanchard OJ, Watson MW (1982) Bubbles, rational expectations, and financial markets. In: Wachtel P (ed) Crises in the economic and financial structure. Lexington Books, Lexington, pp 295–315

36. Blattberg RC, Gonedes NJ (1974) A comparison of the stable and Student distributions as statistical models for stock prices. J Bus 47:244–280
37. Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. J Econom 31:307–327
38. Bollerslev T (1987) A conditionally heteroskedastic time series model for speculative prices and rates of return. Rev Econ Stat 69:542–547
39. Bollerslev T, Kretschmer U, Pigorsch C, Tauchen G (2007) A discrete-time model for daily S&P500 returns and realized variations: Jumps and leverage effects. J Econom (in press)
40. Boothe P, Glassman D (1987) The statistical distribution of exchange rates. J Int Econ 22:297–319
41. Bouchaud JP, Potters M (2000) Theory of financial risks. From statistical physics to risk management. Cambridge University Press, Cambridge
42. Box GEP, Muller ME (1958) A note on the generation of random normal deviates. Ann Math Stat 29:610–611
43. Branco MD, Dey DK (2001) A general class of multivariate skew-elliptical distributions. J Multivar Anal 79:99–113
44. Broca DS (2004) Mixture distribution models of Indian stock returns: An empirical comparison. Indian J Econ 84:525–535
45. Brooks C, Burke SP, Heravi S, Persand G (2005) Autoregressive conditional kurtosis. J Financial Econom 3:399–421
46. Campbell J, Lo AW, MacKinlay AC (1997) The econometrics of financial markets. Princeton University Press, Princeton
47. Carnero MA, Peña D, Ruiz E (2004) Persistence and kurtosis in GARCH and stochastic volatility models. J Financial Econom 2:319–342
48. Carr P, Geman H, Madan DB, Yor M (2002) The fine structure of asset returns: An empirical investigation. J Bus 75:305–332
49. Chambers JM, Mallows CL, Stuck BW (1976) A method for simulating stable random variables. J Am Stat Assoc 71:340–344
50. Chen M, An HZ (1998) A note on the stationarity and the existence of moments of the GARCH model. Stat Sin 8:505–510
51. Cherubini U, Luciano E, Vecchiato W (2004) Copula methods in finance. Wiley, New York
52. Cheung YM, Xu L (2003) Dual multivariate auto-regressive modeling in state space for temporal signal separation. IEEE Trans Syst Man Cybern B: 33:386–398
53. Christoffersen PF, Pelletier D (2004) Backtesting value-at-risk: A duration-based approach. J Financial Econom 2:84–108
54. Clark PK (1973) A subordinated stochastic process model with finite variance for speculative prices. Econometrica 41:135–155
55. Cohen I (2004) Modeling speech signals in the time-frequency domain using GARCH. Signal Process 84:2453–2459
56. Cont R, Tankov P (2004) Financial modelling with jump processes. Chapman & Hall, Boca Raton
57. Corsi F, Mittnik S, Pigorsch C, Pigorsch U (2008) The volatility of realized volatility. Econom Rev 27:46–78
58. Cotter J (2005) Tail behaviour of the euro. Appl Econ 37:827–840
59. Dacorogna MM, Müller UA, Pictet OV, de Vries CG (2001) Extremal forex returns in extremely large data sets. Extremes 4:105–127
60. Dagpunar JS (1989) An easily implemented generalised inverse Gaussian generator. Commun Stat Simul 18:703–710
61. Danielsson J, de Vries CG (1997) Tail index and quantile estimation with very high frequency data. J Empir Finance 4:241–257
62. Danielsson J, de Haan L, Peng L, de Vries CG (2001) Using a bootstrap method to choose the sample fraction in tail index estimation. J Multivar Anal 76:226–248
63. de Haan L, Resnick SI, Rootzén H, de Vries CG (1989) Extremal behaviour of solutions to a stochastic difference equation with applications to ARCH processes. Stoch Process Appl 32:213–234
64. de Vries CG (1994) Stylized facts of nominal exchange rate returns. In: van der Ploeg F (ed) The handbook of international macroeconomics. Blackwell, Oxford, pp 348–389
65. Doganoglu T, Mittnik S (1998) An approximation procedure for asymmetric stable Paretian densities. Comput Stat 13:463–475
66. DuMouchel WH (1973) On the asymptotic normality of the maximum-likelihood estimate when sampling from a stable distribution. Ann Stat 1:948–957
67. DuMouchel WH (1983) Estimating the stable index α in order to measure tail thickness: A critique. Ann Stat 11:1019–1031
68. Dyson FJ (1943) A note on kurtosis. J R Stat Soc 106:360–361
69. Eberlein E, von Hammerstein EA (2004) Generalized hyperbolic and inverse Gaussian distributions: Limiting cases and approximation of processes. In: Dalang RC, Dozzi M, Russo F (eds) Seminar on stochastic analysis, random fields and applications IV. Birkhäuser, Basel, pp 221–264
70. Eberlein E, Keller U (1995) Hyperbolic distributions in finance. Bernoulli 1:281–299
71. Eberlein E, Keller U, Prause K (1998) New insights into smile, mispricing, and value at risk: The hyperbolic model. J Bus 71:371–405
72. Embrechts P, Klüppelberg C, Mikosch T (1997) Modelling extremal events for insurance and finance. Springer, Berlin
73. Engle RF (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50:987–1006
74. Fama EF (1963) Mandelbrot and the stable Paretian hypothesis. J Bus 36:420–429
75. Fama EF (1965) The behavior of stock market prices. J Bus 38:34–105
76. Fama EF (1976) Foundations of finance. Basic Books, New York
77. Fama EF, Roll R (1971) Parameter estimates for symmetric stable distributions. J Am Stat Assoc 66:331–338
78. Farmer JD, Lillo F (2004) On the origin of power-law tails in price fluctuations. Quantit Finance 4:C7–C11
79. Feller W (1950) An introduction to probability theory and its applications I. Wiley, New York
80. Feller W (1971) An introduction to probability theory and its applications II. Wiley, New York
81. Fernández C, Steel MFJ (1998) On Bayesian modeling of fat tails and skewness. J Am Stat Assoc 93:359–371
82. Finucan HM (1964) A note on kurtosis. J R Stat Soc Ser B 26:111–112
83. Gabaix X, Gopikrishnan P, Plerou V, Stanley E (2007) A unified econophysics explanation for the power-law exponents of stock market activity. Physica A 382:81–88
84. Gabaix X, Gopikrishnan P, Plerou V, Stanley HE (2003) A theory of power-law distributions in financial market fluctuations. Nature 423:267–270
85. Gabaix X, Gopikrishnan P, Plerou V, Stanley HE (2006) Institutional investors and stock market volatility. Quart J Econ 121:461–504

86. Galbraith JW, Zernov S (2004) Circuit breakers and the tail index of equity returns. J Financial Econom 2:109–129
87. Gallant AR, Tauchen G (1996) Which moments to match? Econom Theory 12:657–681
88. Gerber HU, Shiu ESW (1994) Option pricing by Esscher transforms. Trans Soc Actuar 46:99–140
89. Ghose D, Kroner KF (1995) The relationship between GARCH and symmetric stable processes: Finding the source of fat tails in financial data. J Empir Finance 2:225–251
90. Goldie CM (1991) Implicit renewal theory and tails of solutions of random equations. Ann Appl Probab 1:126–166
91. Gopikrishnan P, Meyer M, Amaral LAN, Stanley HE (1998) Inverse cubic law for the distribution of stock price variations. Eur Phys J B 3:139–140
92. Gopikrishnan P, Plerou V, Amaral LAN, Meyer M, Stanley HE (1999) Scaling of the distribution of fluctuations of financial market indices. Phys Rev E 60:5305–5316
93. Gourieroux C, Jasiak J (1998) Truncated maximum likelihood, goodness of fit tests and tail analysis. Working Paper, CREST
94. Gray JB, French DW (1990) Empirical comparisons of distributional models for stock index returns. J Bus Finance Account 17:451–459
95. Gut A (2005) Probability: A graduate course. Springer, New York
96. Haas M, Mittnik S, Paolella MS (2004) Mixed normal conditional heteroskedasticity. J Financial Econom 2:211–250
97. Haas M, Mittnik S, Paolella MS (2004) A new approach to Markov-switching GARCH models. J Financial Econom 2: 493–530
98. Hagerman RL (1978) More evidence on the distribution of security returns. J Finance 33:1213–1221
99. Hall JA, Brorsen W, Irwin SH (1989) The distribution of futures prices: A test of the stable Paretian and mixture of normals hypotheses. J Financial Quantit Anal 24:105–116
100. Hall P (1982) On some simple estimates of an exponent of regular variation. J R Stat Soc Ser B 44:37–42
101. Hamilton JD (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57:357–384
102. Hansen BE (1994) Autoregressive conditional density estimation. Int Econ Rev 35:705–730
103. Harrison P (1996) Similarities in the distribution of stock market price changes between the eighteenth and twentieth centuries. J Bus 71:55–79
104. Hazan A, Landsman Z, Makov UE (2003) Robustness via a mixture of exponential power distributions. Comput Stat Data Anal 42:111–121
105. He C, Teräsvirta T (1999) Fourth moment structure of the GARCH(p, q) process. Econom Theory 15:824–846
106. Hill BM (1975) A simple general approach to inference about the tail of a distribution. Ann Stat 3:1163–1174
107. Hols MCAB, de Vries CG (1991) The limiting distribution of extremal exchange rate returns. J Appl Econom 6:287–302
108. Hsieh DA (1989) Modeling heteroskedasticity in daily foreign-exchange rates. J Bus Econ Stat 7:307–317
109. Hsu DA, Miller RB, Wichern DW (1974) On the stable Paretian behavior of stock-market prices. J Am Stat Assoc 69:108–113
110. Huisman R, Koedijk KG, Kool CJM, Palm F (2001) Tail-index estimates in small samples. J Bus Econ Stat 19:208–216
111. Huisman R, Koedijk KG, Kool CJM, Palm F (2002) The tail-fatness of FX returns reconsidered. Economist 150:299–312
112. Hyung N, de Vries CG (2005) Portfolio diversification effects of downside risk. J Financial Econom 3:107–125
113. Jansen DW, de Vries CG (1991) On the frequency of large stock returns: Putting booms and busts into perspective. Rev Econ Stat 73:18–24
114. Jensen MB, Lunde A (2001) The NIG-S&ARCH model: A fat-tailed, stochastic, and autoregressive conditional heteroskedastic volatility model. Econom J 4:319–342
115. Jondeau E, Rockinger M (2003) Testing for differences in the tails of stock-market returns. J Empir Finance 10:559–581
116. Jones MC, Faddy MJ (2003) A skew extension of the t-distribution, with applications. J R Stat Soc Ser B 65:159–174
117. Jørgensen B (1982) Statistical properties of the generalized inverse gaussian distribution. Springer, Berlin
118. Jorion P (2006) Value at risk: The new benchmark for controlling derivatives risk. McGraw-Hill, New York
119. Kaizoji T, Kaizoji M (2003) Empirical laws of a stock price index and a stochastic model. Adv Complex Syst 6:303–312
120. Kanter M (1975) Stable densities under change of scale and total variation inequalities. J Am Stat Assoc 3:697–707
121. Kaplansky I (1945) A common error concerning kurtosis. J Am Stat Assoc 40:259
122. Karanasos M (1999) The second moment and the autocovariance function of the squared errors of the GARCH model. J Econom 90:63–76
123. Kearns P, Pagan A (1997) Estimating the density tail index for financial time series. Rev Econ Stat 79:171–175
124. Kesten H (1973) Random difference equations and renewal theory for products of random matrices. Acta Math 131:207–248
125. Koedijk KG, Schafgans MMA, de Vries CG (1990) The tail index of exchange rate returns. J Int Econ 29:93–108
126. Koedijk KG, Stork PA, de Vries CG (1992) Differences between foreign exchange rate regimes: The view from the tails. J Int Money Finance 11:462–473
127. Kogon SM, Williams DB (1998) Characteristic function based estimation of stable distribution parameters. In: Adler RJ, Feldman RE, Taqqu MS (eds) A practical guide to heavy tails: Statistical techniques and applications. Birkhäuser, Basel, pp 311–335
128. Komunjer I (2007) Asymmetric power distribution: Theory and applications to risk management. J Appl Econom 22:891–921
129. Kon SJ (1984) Models of stock returns: A comparison. J Finance 39:147–165
130. Koponen I (1995) Analytic approach to the problem of convergence of truncated Lévy flights towards the Gaussian stochastic process. Phys Rev E 52:1197–1199
131. Koutrouvelis IA (1980) Regression-type estimation of the parameters of stable laws. J Am Stat Assoc 75:918–928
132. Küchler U, Neumann K, Sørensen M, Streller A (1999) Stock returns and hyperbolic distributions. Math Comput Model 29:1–15
133. Kupiec PH (1995) Techniques for verifying the accuracy of risk management models. J Deriv 3:73–84
134. Laherrère J, Sornette D (1998) Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. Eur Phys J B 2:525–539
135. Lau AHL, Lau HS, Wingender JR (1990) The distribution of stock returns: New evidence against the stable model. J Bus Econ Stat 8:217–223

136. Leadbetter MR, Lindgren G, Rootzén H (1983) Extremes and related properties of random sequences and processes. Springer, New York
137. LeBaron B (2001) Stochastic volatility as a simple generator of apparent financial power laws and long memory. Quantit Finance 1:621–631
138. Lee SW, Hansen BE (1994) Asymptotic theory for the GARCH(1,1) quasi-maximum likelihood estimator. Econom Theory 10:29–52
139. Ling S, McAleer M (2002) Necessary and sufficient moment conditions for the GARCH(r, s) and asymmetric power GARCH(r, s) models. Econom Theory 18:722–729
140. Liu JC (2006) On the tail behaviors of a family of GARCH processes. Econom Theory 22:852–862
141. Liu JC (2006) On the tail behaviors of Box–Cox transformed threshold GARCH(1,1) processes. Stat Probab Lett 76:1323–1330
142. Longin FM (1996) The asymptotic distribution of extreme stock market returns. J Bus 69:383–408
143. Longin FM (2005) The choice of the distribution of asset returns: How extreme value theory can help? J Bank Finance 29:1017–1035
144. Loretan M, Phillips PCB (1994) Testing the covariance stationarity of heavy-tailed time series. J Empir Finance 1:211–248
145. Lux T (1996) The stable Paretian hypothesis and the frequency of large returns: An examination of major German stocks. Appl Financial Econ 6:463–475
146. Lux T (2000) On moment condition failure in German stock returns: An application of recent advances in extreme value statistics. Empir Econ 25:641–652
147. Lux T (2001) The limiting extremal behaviour of speculative returns: An analysis of intra-daily data from the Frankfurt stock exchange. Appl Financial Econ 11:299–315
148. Lux T, Sornette D (2002) On rational bubbles and fat tails. J Money Credit Bank 34:589–610
149. Madan DB, Carr PP, Chang EC (1998) The variance gamma process and option pricing. Eur Finance Rev 2:79–105
150. Madan DB, Milne F (1991) Option pricing with v. g. martingale components. Math Finance 1:39–55
151. Madan DB, Seneta E (1990) The variance gamma (v. g.) model for share market returns. J Bus 63(4):511–524
152. Malevergne Y, Pisarenko V, Sornette D (2005) Empirical distributions of stock returns: Between the stretched exponential and the power law? Quantit Finance 5:379–401
153. Malevergne Y, Pisarenko V, Sornette D (2006) On the power of generalized extreme value (GEV) and generalized Pareto distribution (GDP) estimators for empirical distributions of stock returns. Appl Financial Econ 16:271–289
154. Mandelbrot B (1963) New methods in statistical economics. J Polit Econ 71:421–440
155. Mandelbrot B (1963) The variation of certain speculative prices. J Bus 36:394–419
156. Mandelbrot B (1967) The variation of some other speculative prices. J Bus 40:393–413
157. Mantegna RN, Stanley HE (1994) tic process with ultraslow convergence to a Gaussian: The truncated Lévy flight. Phys Rev Lett 73:2946–2949
158. Marsaglia G, Marshall AW, Proschan F (1965) Moment crossings as related to density crossings. J R Stat Soc Ser B 27:91–93
159. Mason DM (1982) Laws of large numbers for sums of extreme values. Ann Probab 10:754–764
160. Matia K, Amaral LAN, Goodwin SP, Stanley HE (2002) Different scaling behaviors of commodity spot and future prices. Phys Rev E 66:045103
161. Matia K, Pal M, Salunkay H, Stanley HE (2004) Scale-dependent price fluctuations for the Indian stock market. Europhys Lett 66:909–914
162. McCulloch JH (1997) Measuring tail thickness to estimate the stable index α: A critique. J Bus Econ Stat 15:74–81
163. McCulloch JH (1986) Simple consistent estimators of stable distribution parameters. Commun Stat Simul 15:1109–1136
164. McDonald JB (1996) Probability distributions for financial models. In: Maddala GS, Rao CR (eds) Handbook of statistics 14: Statistical methods in finance. Elsevier, Amsterdam, pp 427–461
165. McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, New York
166. Mikosch T, Stărică C (2004) Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. Rev Econ Stat 86:378–390
167. Mikosch T, Stărică C (2000) Limit theory for the sample autocorrelations and extremes of a GARCH(1,1) process. Ann Stat 28:1427–1451
168. Milhøj A (1985) The moment structure of ARCH processes. Scand J Stat 12:281–292
169. Mittnik S, Rachev ST (1993) Modeling asset retunrs with alternative stable distributions. Econom Rev 12:261–330
170. Mizuno T, Kurihara S, Takayasu M, Takayasu H (2003) Analysis of high-resolution foreign exchange data of USD-JPY for 13 years. Physica A 324:296–302
171. Nelson DB (1990) Stationarity and persistence in the GARCH(1,1) model. Econom Theory 6:318–334
172. Nelson DB (1991) Conditional heteroskedasticity in asset returns: A new approach. Econometrica 59:347–370
173. Nelson DB, Cao CQ (1992) Inequality constraints in the univariate GARCH model. J Bus Econ Stat 10:229–235
174. Newcomb S (1980) A generalized theory of the combination of observations so as to obtain the best result. In: Stigler SM (ed) American contributions to mathematical statistics in the nineteenth century, vol. 2. Arno, New York, pp. 343–366
175. Nolan JP (1997) Numerical calculation of stable densities and distribution functions. Commun Stat Stoch Models 13:759–774
176. Officer RR (1972) The distribution of stock returns. J Am Stat Assoc 67:807–812
177. Omran MF (1997) Moment condition failure in stock returns: UK evidence. Appl Math Finance 4:201–206
178. Osborne MFM (1959) Brownian motion in the stock market. Oper Res 7:145–173
179. Peiró A (1994) The distribution of stock returns: International evidence. Appl Financial Econ 4:431–439
180. Perry PR (1983) More evidence on the nature of the distribution of security returns. J Financial Quantit Anal 18:211–221
181. Plerou V, Gopikrishnan P, Amaral LAN, Meyer M, Stanley HE (1998) Scaling of the distribution of price fluctuations of individual companies. Phys Rev E 60:6519–6529
182. Plerou V, Gopikrishnan P, Gabaix X, Stanley E (2004) On the origin of power-law fluctuations in stock prices. Quantit Finance 4:C11–C15

183. Pólya G, Szegö G (1976) Problems and theorems in analysis II. Springer, Berlin
184. Praetz PD (1972) The distribution of share price changes. J Bus 45:49–55
185. Prause K (1999) The generalized hyperbolic model: Estimation, financial derivatives, and risk measures. Ph D thesis, Albert-Ludwigs-Universität Freiburg i. Br.
186. Press SJ (1972) Estimation in univariate and multivariate stable distributions. J Am Stat Assoc 67:842–846
187. Raible S (2000) Lévy processes in finance: Theory, numerics, and empirical facts. Ph D thesis, Albert-Ludwigs-Universität Freiburg i. Br.
188. Resnick SI (1987) Extreme values, regular variation, and point processes. Springer, New York
189. Resnick SI (1997) Heavy tail modeling and teletraffic data. Ann Stat 25:1805–1849
190. Resnick SI, Stărică C (1998) Tail index estimation for dependent data. Ann Appl Probab 8:1156–1183
191. Rydberg TH (1999) Generalized hyperbolic diffusion processes with applications in finance. Math Finance 9:183–201
192. Samorodnitsky G, Taqqu MS (1994) Stable non-gaussian random processes: Stochastic models with infinite variance. Chapman & Hall, New York
193. Sato KI (1999) Lévy processes and infinitely divisible distributions. Cambridge University Press, Cambridge
194. Schoutens W (2003) Lévy processes in finance: Pricing financial derivatives. Wiley, New York
195. Seneta E (1976) Regularly varying functions. Springer, Berlin
196. Sigman K (1999) Appendix: A primer on heavy-tailed distributions. Queueing Syst 33:261–275
197. Silva AC, Prange RE, Yakovenko VM (2004) Exponential distribution of financial returns at mesoscopic time lags: A new stylized fact. Physica A 344:227–235
198. Silverman BW (1986) Density estimation for statistics and data analysis. Chapman & Hall, New York
199. Tsiakas I (2006) Periodic stochastic volatility and fat tails. J Financial Econom 4:90–135
200. Turner CM, Startz R, Nelson CR (1989) A Markov model of heteroskedasticity, risk, and learning in the stock market. J Financial Econ 25:3–22
201. Werner T, Upper C (2004) Time variation in the tail behavior of bund future returns. J Future Markets 24:387–398
202. Weron R (2001) Levy-stable distributions revisited: Tail index > 2 does not exclude the Levy-stable regime. Int J Mod Phys C 12:209–223

Books and Reviews

Bollerslev T, Engle RF, Nelson DB (1994) ARCH models. In: Engle RF, McFadden DL (eds) Handbook of econometrics IV. Elsevier, Amsterdam, pp 2959–3038
Borak S, Härdle W, Weron R (2005) Stable distributions. In: Cizek P, Härdle W, Weron R (eds) Statistical tools for finance and insurance. Springer, Berlin, pp 21–44
Janicki A, Weron A (1993) Simulation and chaotic behavior of α-stable stochastic processes. Dekker, New York
McCulloch JH (1996) Financial applications of stable distributions. In: Maddala GS, Rao CR (eds) Handbook of statistics 14: Statistical methods in finance. Elsevier, Amsterdam, pp 393–425
Mikosch T (2004) Modeling dependence and tails of financial time series. In: Finkenstädt B, Rootzén H (eds) Extreme values in finance, telecommunications, and the environment. Chapman & Hall, Boca Raton, pp 185–286
Mittnik S, Rachev ST, Paolella MS (1998) Stable paretian modeling in finance: Some empirical and theoretical aspects. In: Adler RJ, Feldman RE, Taqqu MS (eds) A practical guide to heavy tails: Statistical techniques and applications. Birkhäuser, Basel, pp 79–110
Nelsen RB (2006) An introduction to copulas, 2nd edn. Springer, New York
Pagan A (1996) The econometrics of financial markets. J Empir Finance 3:15–102
Palm FC (1996) GARCH models of volatility. In: Maddala GS, Rao CR (eds) Handbook of statistics 14. Elsevier, Amsterdam, pp 209–240
Rachev ST, Mittnik S (2000) Stable paretian models in finance. Wiley, Chichester
Zolotarev VM (1986) One-dimensional stable distributions. AMS, Providence

Financial Economics, Non-linear Time Series in

Terence C. Mills[1], Raphael N. Markellos[1,2]
[1] Department of Economics, Loughborough University, Loughborough, UK
[2] Department of Management Science and Technology, Athens University of Economics and Business, Athens, Greece

Article Outline

Glossary
Definition of the Subject
Introduction
Basic Nonlinear Financial Time Series Models
Future Directions
Bibliography

Glossary

Arbitrage The possibility of producing a riskless profit by exploiting price differences between identical or linked assets.

Market efficiency A market is called efficient when all available information is reflected accurately, instantly and fully in the prices of traded assets. Depending on the definition of the available information set, private, public or that contained in historical prices, market efficiency is considered as strong, semi-strong or weak, respectively. The market price of an asset in an efficient market is an unbiased estimate of its true value. Sys-

tematic excess profits, which cannot be justified on the basis of the underlying risk, are not possible in such a market.

Martingale The term was originally used to describe a particular gambling strategy in which the stake is doubled following a losing bet. In probability theory it refers to a stochastic process that is a mathematical model of 'fair play'. This has been one of the most widely assumed processes for financial prices. It implies that the best forecast for tomorrow's price is simply today's price or, in other words, that the expected difference between any two successive prices is zero. Assuming a positive (negative) expected difference leads to the more general and realistic class of submartingale (supermartingale) processes. The martingale process implies that price differences are serially uncorrelated and that univariate linear time series models of prices have no forecasting value. However, martingales do not preclude the potential usefulness of nonlinear models in predicting the evolution of higher moments, such as the variance. The efficient market hypothesis is often incorrectly equated to the so-called random walk hypothesis, which roughly states that financial prices are martingales.

Option A call (put) option is a contractual agreement which gives the holder the right to buy (sell) a specified quantity of the underlying asset, within a specified period of time, at a price that is agreed when the contract is executed. Options are derivative assets since their value is based upon the variation in the underlying, which is typically the price of some asset such as a stock, commodity, bond, etc. Other basic types of derivatives include futures, forwards and swaps. An option is real, in contrast to financial, when the corresponding right refers to some business decision, such as the right to build a factory.

Portfolio theory The study of how resources should be optimally allocated between alternative investments on the basis of a given time investment horizon and a set of preferences.

Systematic risk Reflects the factors affecting all securities or firms in an economy. It cannot be reduced by diversification and it is also known as market risk. In the context of one of the most popular financial models, the Capital Asset Pricing Model (CAPM), systematic risk is measured by the beta coefficient.

Unsystematic risk This is the part of risk that is unique to a particular security or firm and can be reduced through diversification. This risk cannot be explained on the basis of fluctuations in the market as whole and it is also known as residual or idiosyncratic risk.

Volatility A measure of overall risk for an asset or portfolio which represents the sum of systematic and unsystematic risk. While several different approaches have been proposed for approximating this unobservable variable, the simplest one is based on the annualized standard deviation estimated using a historical sample of daily returns.

Definition of the Subject

Financial economics is the branch of economic science that deals with how groups of agents, such as households, firms, investors, creditors and economies as a whole, allocate and exchange financial resources in the context of markets. A wide variety of problems and applications fall within this broad subject area, including asset pricing, portfolio optimization, market efficiency, capital budgeting, interest and exchange rate modeling, risk management, forecasting and trading, market microstructure and behavioral finance. It is a highly quantitative and empirical discipline which draws its theoretical foundations and tools primarily from economics, mathematics and econometrics. Academic research in this area has flourished over the past century in line with the growing importance of financial markets and assets for the everyday life of corporations and individuals (for a historical overview of financial economics, see [69]). Consequently, at least 6 out of the 39 Nobel prizes in Economics have been awarded for research undertaken in areas related to financial economics. The close relationship between finance and time series analysis became widely apparent when Sir Clive W.J. Granger and Robert F. Engle III jointly received the 2003 Nobel Prize. Their work in time series econometrics has had a profound impact both on academic research and on the practice of finance. In particular, the ARCH model, first proposed by Engle [27] for modeling the variability of inflation, is today one of the most well known and important applications of a nonlinear time series model in finance. We should also acknowledge the Nobel Prize received by Robert C. Merton and Myron S. Scholes in 1993 for their pioneering work in the 1970s on pricing financial derivatives. In particular, they, along with Fischer Black, developed an analytical framework and simple mathematical formulae for pricing derivative assets, such as options and warrants, which have highly nonlinear payoff functions. Their work was the first step in the development of the derivatives industry and the whole risk management culture and practice in finance.

The close link between finance and nonlinear time series analysis is by no means accidental, being a consequence of four main factors. First, financial time series

have always been considered ideal candidates for data-hungry nonlinear models. The fact that organized financial markets and information brokers (e. g., newspapers, data vendors, analysts, etc.) have been around for many years has meant that an abundance of high quality historical data exists. Most of this data is in the form of time series and usually spans several decades, sometimes exceeding a century. Furthermore, asset prices can now be collected at ultra-high frequencies, often less than a minute, so that sample sizes may run into millions of observations. Second, the poor forecasting performance of linear models allied to the prospect of obtaining large financial gains by 'beating the market' on the basis of superior forecasts produced by nonlinear time series models has provided a natural motive for researchers from several disciplines. Third, developments in the natural sciences since the 1980s with respect to chaos theory, nonlinear dynamics and complexity have fueled a 'nonlinearist' movement in finance and have motivated a new research agenda on relevant theories, models and testing procedures for financial time series. Underlying this movement was the concern that the apparent unpredictability of financial time series may simply be due to the inadequacy of standard linear models. Moreover, it was also thought that the irregular fluctuations in financial markets may not be the result of propagated exogenous random shocks but, rather, the outcome of some, hopefully low-dimensional, chaotic system (see the entry by Shintani on ► Financial Forecasting, Sensitive Dependence). Fourth, and most importantly, although the bulk of financial theory and practice is built upon affine models, a wealth of theoretical models and supporting empirical evidence has been published suggesting that the nature of some financial problems may be inherently nonlinear. Two prime examples are the time-varying and asymmetric nature of financial risk and the highly nonlinear relationships that arise in situations involving financial options and other derivatives.

Introduction

Traditionally, theorists and empirical researchers in finance and economics have had rather different views concerning nonlinearity. Theorists have shown some interest in nonlinearities and have used them in a variety of different ways, such as first order conditions, multimodality, ceilings and floors, regime switching, multiple equilibria, peso problems, bandwagon effects, bubbles, prey-predator dynamics, time-varying parameters, asymmetries, discontinuities and jump behavior, non-additivity, non-transitivity, etc. Theories and structural models that have nonlinear elements can be found in most areas of finance and economics (selective reviews with a focus mainly on economics are given by Brock and de Lima [19], Lorenz (see Chaps. 1–3 and 6 in [53]), Mullineux and Peng [67], Rosser [74]; other sources include Chap. 3 and pp. 114–147 in [40], [75]). Prominent examples include the noise-trader models of exchange rate determination [34,35], the target-zone exchange rate models [36,50] and the imperfect knowledge models [37]. Nonlinearities find their natural place in the theory of financial derivatives (for overviews, see [46,62]) and real options [26], where payoff functions and relationships between pricing variables are inherently highly nonlinear. The popularity of nonlinearities is limited by the prevailing equilibrium theory assumptions (convexity and continuity conditions, concavity of utility and production functions, constant returns to scale, intertemporally independent tastes and technology, rational aggregate expectations and behavior, etc.) which invariably lead to linear relationships.

For many years, nonlinearities were not a serious consideration when attempting to build empirical models. Alfred Marshall, one of the great pioneers of mathematical economics, epitomized the culture against nonlinear models when saying that "*natura non facit saltum*", or nature dislikes jumps. Although he contemplated the possibility of multiple equilibria and switching behavior and understood that this situation would entail a tendency for stable and unstable equilibria to alternate, he dismissed it as deriving "*from the sport of imagination rather than the observation of facts*". Correspondingly, in empirical and theoretical finance the mainstream approach has been to transform any nonlinearities to linearized forms using Taylor series expansions which excluded second-and higher-order terms. Since the 1990s, however, there has been a significant turn in favor of nonlinear modeling in finance. In addition to the reasons advanced earlier, this development has also been the result of advances in econometric estimation and of the widespread availability of cheap computer power. Some of the basic nonlinear models and relevant theories that have been used in finance will be discussed in the subsequent section (for a comprehensive review of the linear and nonlinear time series models used in finance, see [22,63]).

Basic Nonlinear Financial Time Series Models

Most of the theoretical and empirical research in financial economics has typically hypothesized that asset price time series are unit root stochastic processes with returns that are serially unpredictable. For many years it was thought that this unpredictability was necessary in order to ensure that financial markets function properly according to

the Efficient Market Hypothesis (EMH; see the reviews by Fama [32,33]). Within this framework, a market is considered efficient with respect to a specific information set, an asset pricing model and a data generating process, respectively. For example, a general condition of efficiency is that market prices fully, correctly and instantaneously reflect the available information set. This is sometimes formalized as the Random Walk Hypothesis (RWH), which predicts that prices follow random walks with price changes that are unforecastable on the basis of past price changes. An even milder condition is that trading on the information set does not allow profits to be made at a level of risk that is inconsistent with the underlying asset pricing model. Although initially the EMH and RWH were thought to be an unavoidable consequence of the widely accepted paradigm of rational expectations, this was later refuted by a series of studies showing that random walk behavior was neither a necessary nor sufficient condition for rationally determined financial prices. Market efficiency has profound practical economic implications insofar as financial prices serve both as ways of integrating and distributing available information and as asset allocation devices.

One of the simplest models of financial prices that can be derived on the basis of unpredictability is the martingale process:

$$p_t = p_{t-1} + \varepsilon_t \tag{1}$$

where p_t is the price of an asset observed at time t and ε_t is the martingale increment or martingale difference. The martingale has the following properties: a) $E(|p_t|) < \infty$ for each t, b) $E(p_t|\Im_s) = p_s$ whenever $s \leqslant t$, where $\Im_s$ is the σ-algebra comprized of events determined by observations over the interval $[0, t]$, so that $\Im_s \subseteq \Im_t$ when $s \leqslant t$. The martingale possesses the Markov property since the differences $\Delta p_t = p_t - p_{t-1} = \varepsilon_t$ are unpredictable on the basis of past differences. By successive backward substitution in (1) we can express the current price as the accumulation of all past errors. In financial terms, errors can be thought to be the result of unexpected information or news. By restricting the differences ε_t to be identically and independently distributed (iid) we obtain what is often called the random walk process. The random walk is a term and assumption which is widely employed in finance. It was first used by Karl Pearson in a letter to Nature in 1905 trying to describe a mosquito infestation in a forest. Soon after, Pearson compared the process to the walk of an intoxicated man, hence the graphical term "drunkard's walk".

By representing the random walk in continuous time with a growth rate μ, as is often useful when dealing with derivatives, we obtain the generalized Wiener process (also called Brownian motion or diffusion):

$$\mathrm{d}p_t = \mu \mathrm{d}t + \sigma \mathrm{d}w_t \tag{2}$$

where $\mathrm{d}w_t$ is a standard normal random variable. The parameters μ and σ are referred to in finance as the drift and volatility of the process, respectively. Another point worth mentioning is that in both discrete and continuous time the analysis is typically undertaken using logarithmically transformed prices. This precludes the paradoxical possibility of obtaining negative prices while, at the same time, regularizing the statistical behavior of the data. Assuming that prices are lognormally distributed means that logarithmic returns are normally distributed and can be calculated as $\log p_t - \log p_{t-1}$ or $\log(p_t/p_{t-1})$. These represent continuously compounded returns and are approximately equal to simple percentage returns.

Random walks, along with continuous-time mathematical finance, were formally introduced in 1900 by Louis Bachelier in his brilliant doctoral dissertation *Théorie de la Spéculation*. Under the supervision of the great Henri Poincaré, who first realized the possibility of chaotic motion, Bachelier developed the mathematical framework of random walks in continuous time in order to describe the unpredictable evolution of stock prices and to build the first option pricing model (biographical details of Bachelier are given in [58]). Random walks were independently discovered by Albert Einstein in 1905 and, of course, have since played a fundamental role in physics and mathematics. They were later rigorously treated, along with forecasting and nonlinear modeling, by Norbert Wiener, the father of cybernetics. Several important deviations from the Bachelierian random walk and normal distribution paradigm were developed several decades later by Benoit Mandelbrot and his co-authors (for an overview see [59], and the references given therein). This research developed around the generalized Central Limit Theorem (CLT), the stable family of distributions, long-term dependence processes, scaling and fractals. Indeed, it is clear that Mandelbrot views his research as similar to that of Bachelier in that both were inspired by finance and both found great applications later in physics or, to use Mandelbrot's words, both were cases of the "unexpected historical primacy of financial economics over physics" (see p. 174 in [59]).

Much of the motivation behind nonlinear time series modeling in finance has to do with certain empirical characteristics, or stylized facts, which have been observed over the years across many financial assets, markets and time periods. Since these characteristics were not always consistent with a linear data generating process, nonlinear models seemed to be a reasonable explanation. In particular,

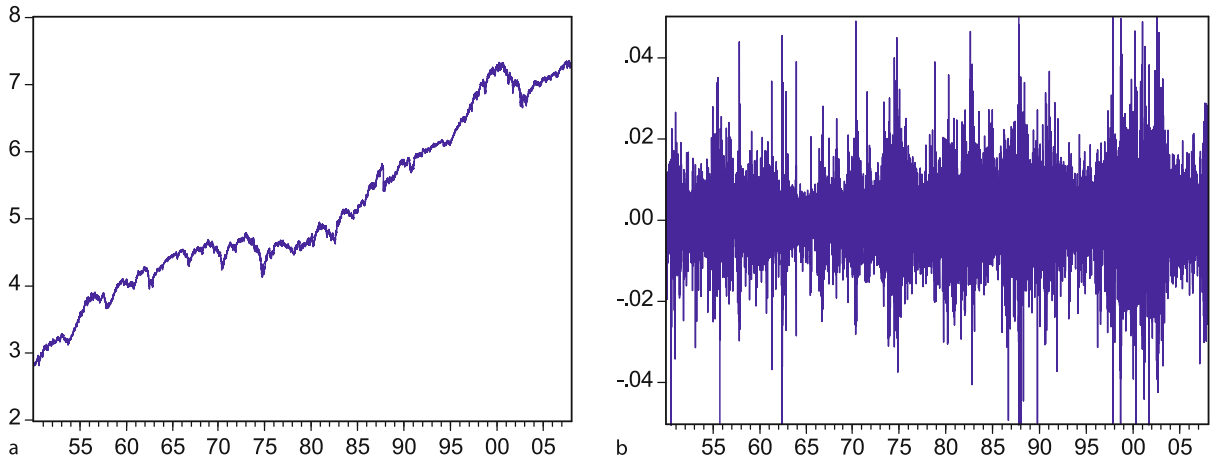

Financial Economics, Non-linear Time Series in, Figure 1
Daily S&P 500 log Index Prices *(left)* and Returns *(right)* (3/1/1950–14/12/2007) (Returns are trimmed to ±5% in order to improve the readability of the graph)

starting with Mandelbrot and others in the 1960s, several empirical studies have reported that financial assets typically have daily returns exhibiting:

- Nonnormality: skewed and leptokurtic (fat-tailed and high-peaked) unconditional distributions.
- Jump behavior: discontinuous variations that result in extreme observations.
- Volatility clustering: large (small) returns in magnitude tend to be followed by large (small) returns of either sign.
- Unpredictability: zero or weak serial autocorrelations in returns.

In order to illustrate these characteristics, we investigate the empirical behavior of daily logarithmic prices and returns (simply referred to as prices and returns hereafter) for the S&P 500 index. The series is publicly available from *Yahoo Finance*. The empirical analysis is undertaken using the econometric software packages *EViews* 5.0 by Quantitative Micro Software and *Time Series Modelling* 4.18 by James Davidson, respectively. The sample consists of 14,582 closing (end of the day) prices covering the period 3/1/1950–14/12/2007. The index is calculated as a weighted average of the common stock prices for the 500 largest firms traded on the New York Stock Exchange (NYSE) and is adjusted for dividends and splits. The S&P 500 is often used as a proxy for the so-called market portfolio and as a measure of the overall performance of the US stock market.

The prices depicted in the left part of Fig. 1 exhibit the upward drifting random walk behavior which is so routinely observed in financial time series. This is consistent with the fact that the series could not be predicted on the basis of past values using a member of the ARIMA class of linear models (as popularized by Box and Jenkins [16]). More specifically, the best fit was offered by an ARIMA(1,1,1) model, although with a disappointingly low R-squared statistic of just 0.64% (absolute t-statistics in brackets):

$$\Delta p_t = \underset{(3.9175)}{0.0003} - \underset{(3.3030)}{0.3013}\,\Delta p_{t-1} + \underset{(4.2664)}{0.3779}\,\varepsilon_{t-1} + \varepsilon_t\,. \quad (3)$$

Such weak linear serial predictabilities are often found at high sampling frequencies and are usually explained by market microstructures. They do not represent true predictabilities but, rather, result from specific market mechanisms and trading systems (see the survey by Biais et al. [12]).

A close examination of the return series, presented in the right part of Fig. 1, suggests the presence of large, discontinuous variations. In particular, we can count 26 daily returns which, in absolute value, exceed five standard deviations. This implies that such extreme events occur with a probability of 0.18% or, on average, almost once every two years (assuming 250 trading days in each calendar year). Under a normal distribution, such 'five-standard deviation' events should be extremely rare, with a probability of occurrence of only 0.00003%, or less than 3 days in every 40,000 years! The fat tails of the return distribution are also reflected in the kurtosis coefficient of 37.3, which is much larger than the value of 3 that corresponds to a normal distribution. In terms of asymmetry, the distribution

is skewed to the left with the relevant coefficient estimated at −1.3.

Clearly, the normal distribution provides a poor approximation of reality here: the distribution of the errors in the random process described by (1) should be allowed to follow some non-Gaussian, fat-tailed and possibly skewed distribution (see the entry by Haas and Pigorsch on ► Financial Economics, Fat-Tailed Distributions). Various distributions having these properties have been proposed, including the Student-t, the mixture of normals, double Weibull, generalized beta, Tukey's $g \times h$, generalized exponential, asymmetric scale gamma, etc. (see [48,71]). Although some of the non-Gaussian distributions that have been proposed have many desirable properties, empirical evidence regarding their appropriateness for describing financial returns has been inconclusive. Moreover, these distributions often bring with them acute mathematical problems in terms of representation, tractability, estimation, mixing and asymptotics. A distribution that has received considerable attention is the stable family (also known as the stable Paretian, Pareto–Lévy, or Lévy flight), which was initially proposed by Mandelbrot [54,55]. Stable distributions are highly flexible, have the normal and Cauchy as special cases, and can represent 'problematic' empirical densities that exhibit asymmetry and leptokurtosis. Furthermore, they are consistent with stochastic behavior that is characterized by discontinuities or jumps. From a theoretical point of view, stable distributions are particularly appealing since they are the limiting class of distributions in the generalized CLT, which applies to scaled sums of iid random variables with infinite variances. Stable distributions also exhibit invariance under addition, a property that is important for financial data, which are usually produced as the result of time aggregation. For a comprehensive discussion of these distributions, see [54,55,64,65,71,76].

In terms of the conditional distribution, it is evident from the graph of returns that the variance is not homogeneous across time, as one would expect for an iid process. In line with this observation, the autocorrelation of squared or absolute returns suggest the presence of strong dependencies in higher moments, something that in turn is indicative of conditional heteroskedasticity (see Fig. 3 below). On the basis of the above, it appears that the simple random walk model is far too restrictive and that the more general martingale process provides a better approximation to the data. Unlike the random walk, the martingale rules out any dependence in the conditional expectation of Δp_{t+1} on the information available at t, while allowing dependencies involving higher conditional moments of Δp_{t+1}. This property of martingales is very useful for explaining clusters in variances, since it allows persistence (correlation) in the conditional variances of returns.

It should be noted that much of the empirical work on nonlinear financial time series has involved modeling time varying variances. This concentration on the variance stems from it being the most widely used measure of risk, which, in orthodox finance theory, is the sole determinant of the expected return of any asset. Knowing the expected return enables the opportunity cost of any investment or asset to be estimated and, ultimately, to have a fair price put on its value by discounting all future revenues against the expected return. Variance was introduced in the path-breaking research of Nobel Laureate Harry Markowitz in the 1950s on investment portfolio selection, which laid the basis for what is known today as modern portfolio theory. The main innovation of Markowitz was that he treated portfolio selection as a tractable, purely quantitative problem of utility maximization under uncertainty, hence the term 'quant analysis'. Markowitz assumed that economic agents face a choice over two-dimensional indifference curves of investment preferences for risk and return. Under some additional assumptions, he obtained a solution to this problem and described the preferences of homogeneous investors in a normative manner using the mean and variance of the probability distribution of single period returns: such investors should optimize their portfolios on the basis of a 'mean-variance' efficiency criterion, which yields the investment with the highest expected return for a given level of return variance.

Let us now turn to some of the processes that have been used to model regularities in variance. For example, consider the GARCH(1,1) process, which has become very popular for modeling the conditional variance, σ_t^2, as a deterministic function of lagged variances and squared errors (see the entry by Hafner on ► GARCH Modeling):

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{4}$$

where the ε_t are, in general, the residuals from a fitted conditional mean equation. This specification corresponds to a single-lagged version of the GARCH(p, q) (Generalized Autoregressive Conditional Heteroskedasticity) model proposed by Bollerslev [13] and can easily be modified to include additional lagged squared errors and variances. The GARCH model is an extension of the ARCH process originally proposed by Engle [27] and has served as the basis for the development of an extensive family of related models. For a review of this huge literature see, among others, [10,14,15,52,79], and Chap. 5 in [63]. Multivariate extensions of GARCH processes have also been proposed, but bring several computational and estimation problems (see [9,21]).

Two alternative approaches to modeling conditional variances in finance are extreme value estimators (see [17]) and realized variance (see [8]). Extreme value estimators depend on opening, closing, high and low prices during the trading day. Although they perform relatively well in terms of efficiency and are easy to estimate, they are quite badly biased. Realized variances are considered to be very accurate and are easily estimated as the sum of squared returns within a fixed time interval. Their limitation is that they require high frequency data at the intradaily level, which can be strongly affected by market microstructures and may not always be readily available. Rather than focusing on just the conditional variance, models have also been proposed for higher moments, such as conditional skewness and kurtosis (e. g., [43,47]).

To illustrate the application of some of the most popular GARCH parameterizations, consider again the S&P 500 return series. Using Maximum Likelihood (ML) estimation with t-student errors, the following 'GARCH(1,1)-in-Mean' (GARCH-M) model was obtained (absolute z-statistics appear in brackets):

$$\Delta p_t = \underset{(10.0994)}{0.0785}\sigma_t + \varepsilon_t$$

$$\sigma_t^2 = \underset{(6.5616)}{5.77 \cdot 10^{-7}} + \underset{(16.3924)}{0.0684}\varepsilon_{t-1}^2 + \underset{(218.0935)}{0.9259}\,\sigma_{t-1}^2 \,.$$

In this model, originally proposed by Engle et al. [30], returns are positively related to the conditional standard deviation, σ_t. This is a particularly useful specification since it is directly consistent with Markowitz's theory about the positive relation between expected return and risk. In particular, the slope coefficient in the conditional mean equation can be interpreted as a relative risk aversion parameter, measuring how investors are compensated by higher returns for bearing higher levels of risk.

It is instructive to show in Fig. 2 both the estimated GARCH(1,1)-M conditional standard deviations and the standardized residuals, ε_t/σ_t. On the one hand, the model clearly produces mean reversion in volatility, which resembles the empirical behavior observed in the original series. Although the estimated conditional variance process appears to be highly persistent, it is nevertheless stationary since the sufficiency condition is satisfied because $\alpha_1 + \beta_1 = 0.0684 + 0.9259 = 0.9943 < 1$. On the other hand, the standardized residuals have a far more homogeneous conditional volatility than the original series and more closely resemble a white noise process. Moreover, the standardized residuals are closer to a normal distribution, with a kurtosis coefficient of 7.7, almost five times smaller than that of the original return series.

Careful inspection of the relationship between returns and conditional variance often reveals an asymmetric relationship. Threshold GARCH (TGARCH) and Exponential GARCH (EGARCH) are two of the specifications often used to model this commonly encountered nonlinearity. These models were estimated using the ML approach and the following conditional variance specifications were obtained.

TGARCH

$$\sigma_t^2 = \underset{(8.5110)}{7.50 \cdot 10^{-7}} + \underset{(6.3766)}{0.0259}\,\varepsilon_{t-1}^2 + \underset{(13.1058)}{0.0865}\,\varepsilon_{t-1}^2\, g + \underset{(224.9279)}{0.9243}\,\sigma_{t-1}^2$$

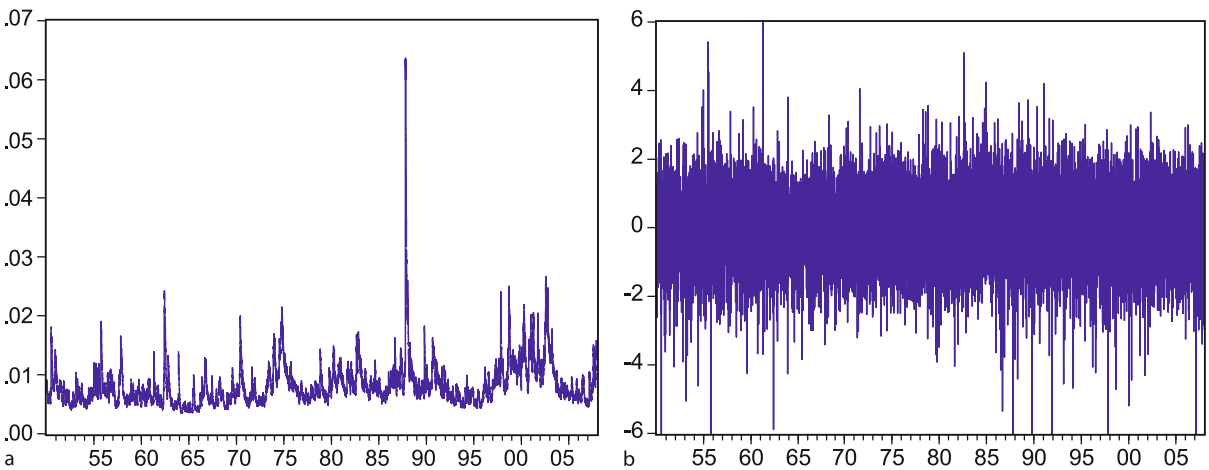

Financial Economics, Non-linear Time Series in, Figure 2
GARCH(1,1)-M standard deviations *(left)* and standardized residuals *(right)* (Residuals are trimmed to ±6 standard deviations in order to improve the readability of the graph)

EGARCH

$$\log(\sigma_t^2) = -\underset{(13.7858)}{0.2221} + \underset{(16.9612)}{0.1209}\,|\varepsilon_{t-1}/\sigma_{t-1}| - \underset{(15.9599)}{0.0690}\,\varepsilon_{t-1}/\sigma_{t-1} + \underset{(703.7161)}{0.9864}\,\log(\sigma_{t-1}^2)\,.$$

In the TGARCH model, the threshold parameter is defined as $g = 1$ if $\varepsilon_{t-1} < 0$ and 0 otherwise. Standard GARCH models, such as the GARCH-M estimated previously, assume that positive and negative errors (or news) have a symmetric effect on volatility. In the TGARCH and EGARCH models, news has an asymmetric effect on volatility depending on its sign. Specifically, in the TGARCH model news will have differential impacts on volatility depending on the signs and sizes of the coefficients on ε_{t-1}^2 and $\varepsilon_{t-1}^2 \cdot g$: good news ($\varepsilon_{t-1} > 0$) has an impact of 0.0259, while bad news ($\varepsilon_{t-1} < 0$) has a stronger impact of $0.0259 + 0.0865 = 0.1124$. Since the coefficient of $\varepsilon_{t-1}^2 \cdot g$ is positive (0.0865), bad news tends to increase volatility, producing what is known as the 'leverage' effect. This was first observed in the 1970s and postulates that negative returns will usually reduce the stock price and market value of the firm, which in turn means an increase in leverage, i. e. a higher debt to equity ratio, and ultimately an increase in volatility. In the EGARCH model, forecasts are guaranteed to be positive since logarithms of the conditional variance are modeled. Since the sign of the coefficient on $\varepsilon_{t-1}/\sigma_{t-1}$ is non-zero and negative we can conclude that the effect of news on volatility is asymmetric and that a leverage effect is present.

Inspection of the autocorrelation functions (ACFs) in Fig. 3 for the returns and absolute returns of the S&P 500, the latter being a proxy for volatility, suggests very different behavior of the two series. While returns have an ACF that is typical of a white noise process, the autocorrelations of the absolute returns die out very slowly and become negative only after 798 lags! It turns out that many financial series have such extremely persistent or long-memory behavior. This phenomenon was first described by Mandelbrot [56,57] in the context of the 'Hurst effect' and was latter defined as fractional Brownian motion (see the relevant review by Brock [18]). Hosking [45] and Granger and Joyeux [39] modeled long-memory by extending the ARIMA class of processes to allow for fractional unit roots (for reviews, see [3,11,73,82]). The ARFIMA(p, d, q) model uses a fractional difference operator based on a binomial series expansion of the parameter d for any value between −0.5 and 0.5:

$$\Delta^d = 1 - dB + \frac{d\,(d-1)}{2!}B^2 - \frac{d\,(d-1)\,(d-2)}{3!}B^3 + \ldots \tag{5}$$

where B is the backshift (or lag) operator with $B^m x_t = x_{t-m}$. In a similar fashion, investigating the existence of long-memory in the conditional variance of the returns could be undertaken in the context of a Fractional GARCH model (see [4]).

In our S&P 500 example, we have shown that nonlinearities enter through the conditional variance process and do so in an asymmetric manner. A natural question to ask is whether nonlinearities also exist in the conditional mean. Consider, for example, a generalization of the linear ARMA process

$$\Delta p_t = f(\Delta p_{t-i}, \varepsilon_{t-i}) + \varepsilon_t \tag{6}$$

where $f()$ is a nonlinear function and $\Delta p_{t-i}, \varepsilon_{t-i}$ are lagged price differences and errors, respectively. A wide

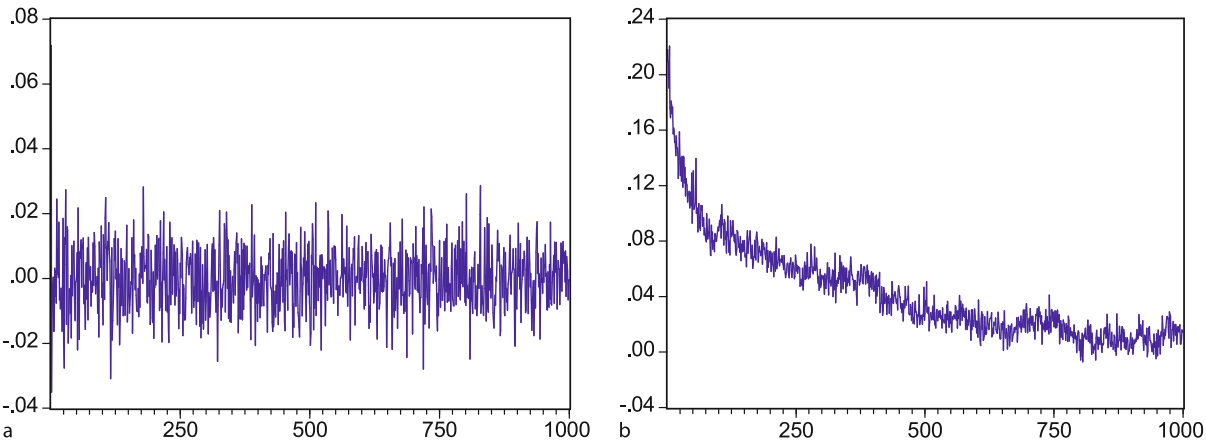

Financial Economics, Non-linear Time Series in, Figure 3
Autocorrelation function of S&P 500 simple *(left)* and absolute returns *(right)*

variety of testing procedures have been proposed for examining the possibility of nonlinearities in the conditional mean process (for reviews see the relevant sections in [40], and [63]). Here we use the BDS test of the null hypothesis of serial independence, which has been widely applied and has been shown to have good power against a variety of nonlinear alternatives (see [20]). The test is inspired by chaos theory and phase space analysis and is based on the concept of the correlation integral. Specifically, the test relies on the property that, for an iid series, the probability of the distance between any two points being no greater than a predefined distance (ε) should be constant. A joint probability can also be calculated for sets comprising multiple pairs of points chosen by moving through consecutive sequences of observations in the sample. The number of consecutive data points used in such a set is called the (embedding) dimension and may be chosen by the user. Brock et al. [20] constructed an asymptotically normally distributed test statistic for the constancy of the distance ε between points. When this test was applied to the residuals from an MA(1)-EGARCH(1,1) model fitted to the S&P 500 returns, it was always insignificant across a variety of dimensions, implying that any nonlinear dependencies in the returns are due solely to GARCH effects.

An agnostic, yet often convenient, way to approximate the unknown nonlinear function (6) is to consider some nonparametric estimator (see [72]). Although several nonparametric estimators have been used with mixed success, one of the most popular is based on the neural network family of models (see [80]). A rich variety of parametric nonlinear functions have also been proposed in finance. A convenient and intuitive way of introducing nonlinearity is to allow 'regime switching' or 'time-variation' in the parameters of the data generating process (for a review see [70]). Three of the most popular approaches in this category are the Markov switching, the Threshold Autoregressive (TAR) and the Smooth Transition (STAR) models. In the first approach (for a popular implementation, see [41,42]), the model parameters switch according to a multiple (typically two) unobserved state Markov process. In TAR models (see [81], for a comprehensive description), nonlinearities are captured using piecewise autoregressive linear models over a number of different states. For example, consider the simple two regime case:

$$x_t = \begin{cases} \omega_1 + \sum_{i=1}^{p} \varphi_{1i} x_{t-i+1} + \sigma_1 \varepsilon_t, & s_{t-d} < c \\ \omega_2 + \sum_{i=1}^{p} \varphi_{2i} x_{t-i+1} + \sigma_2 \varepsilon_t, & s_{t-d} \geqslant c \end{cases} \tag{7}$$

where c is the threshold value, s_t is a threshold variable, d is a delay parameter assumed to be less than or equal to p, and the ε_t are iid standard normal variates assumed to be independent of lagged s_ts. The threshold variable is often determined by a linear combination of the lagged x_ts, in which case we obtain the Self Exciting TAR (SETAR) model. This has become a popular parameterization in finance since it can produce different dynamic behavior across regimes with characteristics such as asymmetry, limit cycles, jumps and time irreversibility (recall the TGARCH model introduced earlier, which has a related specification). STAR models allow a smooth switch between regimes using a smooth transition function. Transition functions that have been considered include the cumulative distribution of the standard normal, the exponential (ESTAR) and the logistic (LSTAR).

It is instructive to see how regime switching can be applied in the context of asset pricing models (for a comprehensive treatment of asset pricing, see [24]). The best known and most influential framework, which builds upon Markowitz's portfolio theory, is the Capital Asset Pricing Model (CAPM) proposed by Sharpe, Lintner, Black and others. The CAPM can be expressed as a single-period equilibrium model:

$$E(r_i) = r_\mathrm{f} + \beta_i \left[E(r_\mathrm{m}) - r_\mathrm{f}\right] \tag{8}$$

where $E(r_i)$ is the expected return on asset i, $E(r_\mathrm{m})$ is the expected return on the market portfolio, r_f is the risk-free interest rate, and the slope β_i is the so-called beta coefficient of asset i, measuring its systematic risk. Empirical implementations and tests of the CAPM are usually based on the 'excess market' and 'market model' regressions, respectively

$$r_{i,t} - r_{\mathrm{f},t} = r_{\mathrm{f},t} + \beta_i \left[r_{\mathrm{m},t} - r_{\mathrm{f},t}\right] + \varepsilon_{i,t} \tag{9}$$

and

$$r_{i,t} = \alpha_i + \beta_i r_{\mathrm{m},t} + \varepsilon_{i,t}\,. \tag{10}$$

The variance of the residuals $\varepsilon_{i,t}$ reflects the unsystematic risk in asset i. In practice the CAPM is typically estimated using ordinary least squares regression with five years of monthly data. A wealth of empirical evidence has been published showing that the basic assumptions of the CAPM regressions with respect to parameter stability and residual iid-ness are strongly refuted (see [60]). In particular, betas have been found to be persistent but unstable over time due to factors such as stock splits, business cycle conditions, market maturity and other political and economic events. In order to demonstrate the modeling of time-varying betas in the CAPM, consider first the simple market model regression for the stock returns of Tiffany & Co (listed on the New York Stock Exchange) against S&P

500 returns:

$$r_t = \underset{(17.5396)}{1.4081}\, r_{\mathrm{m},t} + \varepsilon_t, \quad R^2 = 28.75\% .$$

The regression was estimated using weekly returns from 30/12/1987 to 14/12/2007, a total of 1,044 observations. The R^2 statistic denotes the proportion of total risk that can be explained by the model and which is thus systematic. The beta coefficient is significantly higher than unity, suggesting that the stock is 'aggressive' in that it carries more risk than the market portfolio. Allowing the beta coefficient to switch according to a Markov process produces the following two-regime market model:

$$r_t = \begin{cases} \text{Regime 1: } \underset{(1.9220)}{0.4797} r_{\mathrm{m},t} + \varepsilon_t \\ \text{Regime 2: } \underset{(10.6381)}{1.9434} r_{\mathrm{m},t} + \varepsilon_t \end{cases} \quad R^2 = 41.63\% .$$

The explanatory power of the model has increased significantly and the stock is now characterized by both passive ($\beta = 0.4797 < 1$) and aggressive ($\beta = 1.9434 > 1$) systematic risk behavior regimes. The Markov transition probabilities $P\left(i \mid j\right)$, $j = 1, 2$, were estimated as $P(1|1) = 0.6833$, $P(1|2) = 0.3167$, $P(2|1) = 0.2122$ and $P(2|2) = 0.7878$. The smoothed probabilities for regime 1 are depicted in Fig. 4 and are seen to be rather volatile, so that the returns switch regimes rather frequently. For a discussion of the threshold CAPM see [2].

Another important category of models allows for nonlinear relationships between persistent financial time series. The most popular framework here is that of cointegration, which deals with variables that are individually nonstationary but have some joint stationary representation. For example, consider the linear combination of two unit root ($I(1)$) processes x_t and y_t

$$x_t = a + y_t + \varepsilon_t . \tag{11}$$

In general, ε_t will also be $I(1)$. However, as shown by Engle and Granger [29], if ε_t is actually $I(0)$, then x_t and y_t are said to be (linearly) cointegrated and will have an error-correction representation which, for example, could take the form

$$\Delta x_t = -\gamma \varepsilon_{t-1} + u_t \tag{12}$$

where $-\gamma$ denotes the strength of reversal to the equilibrium cointegrating relationship through the error-correction term, i. e., the lagged residual from the cointegrating regression (11). The finance literature has considered nonlinear generalizations of both the cointegrating regression (11) and the error-correction model (12) (see the entry by Escribano et al. on ► Econometrics: Non-linear Cointegration). Nonlinear error-correction mechanisms can be accommodated rather straightforwardly within the cointegration analysis framework, with the residuals from a linear cointegration relationship entering a nonlinear error-

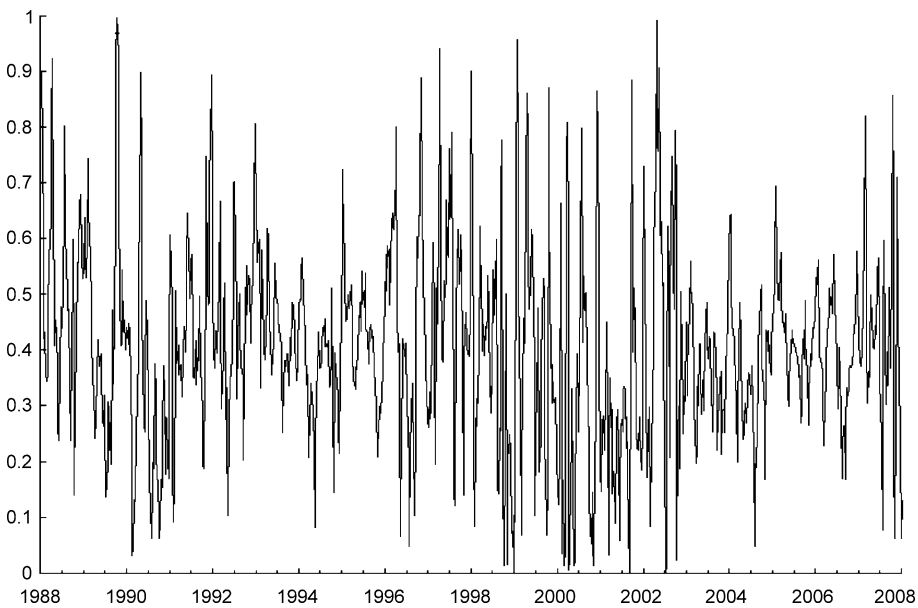

Financial Economics, Non-linear Time Series in, Figure 4
Tiffany stock Markov switching market model smoothed probabilities for Regime 1 of 2

correction model. It has been shown that such nonlinearities may arise simply because of complex relationships between variables (see pp. 59–61 in [40]). Justifications in terms of finance theory have been based on factors such as arbitrage in the presence of transaction costs, heterogeneity among arbitrageurs, existence of partial adjustment models and market segmentation, agents' maximizing or minimizing behavior, constraints on central bank intervention, and intertemporal choice behavior under asymmetric adjustment costs. While almost all the different nonlinear specifications discussed previously have also been applied in error-correction modeling, threshold models hold a prominent position, as they allow large errors from equilibrium, i. e., those above some threshold, to be corrected while small errors are ignored (see, for example, [6]). The use of nonlinearities directly within the cointegrating relationship is not as straightfoward and brings several conceptual and estimation problems (see [63]).

Returning to the bivariate market model setting, it has been found that cointegrating relationships do exist between stock prices and index levels (see [60]). In our example, the logarithms of Tiffany's stock prices are cointegrated with S&P 500 logarithmic price levels. The following asymmetric error correction model was then estimated:

$$r_t = \underset{(3.0329)}{-0.0168}\, \varepsilon_{t-1}\, g + u_t$$

where g is the heavyside function defined previously with $g = 1$ if $\varepsilon_{t-1} < 0$ and 0 otherwise, ε_{t-1} being obtained from the cointegrating regression.

Several studies have shown that empirical characteristics and regularities, such as those discussed previously are very unlikely to remain stable if the sampling frequency of the data changes. For example, we find that if the S&P 500 returns are estimated at an annual frequency using the first available January price, then their distribution becomes approximately Gaussian with skewness and kurtosis coefficients estimated at −0.4 and 2.7, respectively. The annual prices are highly predictable using an ARIMA(2,1,2) process with an impressive adjusted R-squared value of 15.7%. Moreover, standard tests of heteroskedasticity suggest that the variance of annual returns can be assumed to be constant! In contrast, for very high sampling frequencies, say at the intradaily or tick-by-tick level, the data behave in a different manner and are characterized by strong seasonalities, e. g., variances and volumes follow an inverse J shape throughout the trading day (see the review by Goodhart and O'Hara [38], and the discussion in [28]).

Finally, let us now turn our discussion to models in a continuous time setting. As previously mentioned, the analysis of derivatives provides a natural setting for nonlinear modeling since it deals with the pricing of assets with highly nonlinear payoff functions. For example, under the widely used Black–Scholes option pricing model (see [46], for a thorough description), stock prices are lognormally distributed and follow a Wiener process. The Black–Scholes model allows for highly nonlinear relationships between the pricing variables and parameters, as shown in Fig. 5.

Another popular use of continuous time processes is in modeling the autonomous dynamics of processes such as interest rates and the prices of stocks and commodities. A generic stochastic differential equation that can be used to nest alternative models is the following:

$$\mathrm{d}S_t = \mu\left(S_t, t\right)\mathrm{d}t + \sigma\left(S_t, t\right)\mathrm{d}W_t + y\left(S_t, t\right)\mathrm{d}q_t \quad (13)$$

where S_t is the price at time t, $\mathrm{d}W_t$ is a standard Wiener process, $\mu\left(S_t, t\right)$ is the drift, and $\sigma\left(S_t, t\right)$ is the diffusion coefficient. Both the drift and diffusion coefficients are assumed to be functions of time and price, respectively. A jump component is also allowed by incorporating a Poisson process, $\mathrm{d}q_t$, with a constant arrival parameter λ, i. e., $\Pr\{\mathrm{d}q_t = 1\} = \lambda \mathrm{d}t$ and $\Pr\{\mathrm{d}q_t = 0\} = 1 - \lambda \mathrm{d}t$: y is the jump amplitude, also a function of time and price. $\mathrm{d}W_t$, $\mathrm{d}q_t$ and y are assumed to be mutually independent processes. Several nonlinear models can be obtained by combining various assumptions for the components $\mu\left(S_t, t\right)$, $\sigma\left(S_t, t\right)$ and $y\left(S_t, t\right)$. For example, consider the following processes.

Mean Reverting Square-Root Process (MRSRP)

$$\mathrm{d}S_t = \kappa\left(\theta - S_t\right)\mathrm{d}t + \sigma\sqrt{S_t}\mathrm{d}W_t \quad (14)$$

Constant Elasticity of Variance (CEV)

$$\mathrm{d}S_t = \kappa\left(\theta - S_t\right)\mathrm{d}t + \sigma S_t^{\gamma}\mathrm{d}W_t \quad (15)$$

Geometric Wiener Process augmented by Jumps (GWPJ)

$$\mathrm{d}S_t = \left(\mu - \lambda\mu_j\right)S_t\mathrm{d}t + \sigma S_t\mathrm{d}W_t + \left(\mathrm{e}^{y} - 1\right)S_t\mathrm{d}q_t \quad (16)$$

MRSRP augmented by Jumps (MRSRPJ)

$$\mathrm{d}S_t = \kappa\left(\theta - S_t\right)\mathrm{d}t + \sigma\sqrt{S_t}\mathrm{d}W_t + y\mathrm{d}q_t\,. \quad (17)$$

Model (14) has been widely used in modeling interest rates (e. g., [1,23,25]) and stochastic volatility (e. g., [44,68]). Process (16) is often used for representing the dynamics of stock prices and indices (e. g., [61]). Model (17) has been recently employed by several researchers for modeling volatility, because it allows rapid changes in volatility during times of market stress (e. g., [31]). While process (16) has a proportional structure, with μ being the expected return of the asset per unit of time and σ its

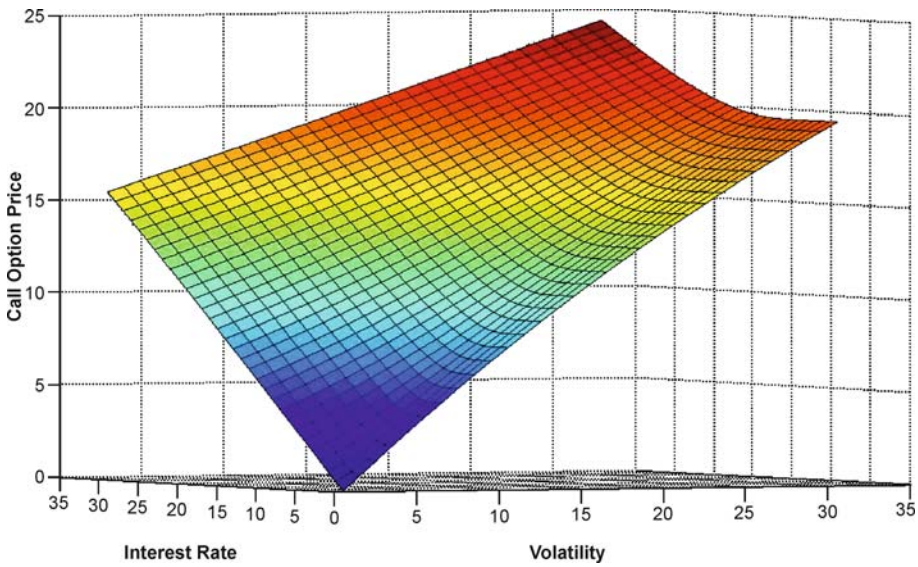

Financial Economics, Non-linear Time Series in, Figure 5
Call option prices, volatility and interest rate in the Black–Scholes model (Call option prices were estimated using the Black–Scholes model assuming a strike price of 50, 1 year time to maturity and a zero dividend yield)

volatility, the other processes have mean reverting drifts. In Eqs. (14), (15) and (17) κ is the speed of mean reversion, θ is the unconditional long-run mean, and σ the volatility of the price process. In Eq. (15), γ is a free parameter to be estimated that determines the dependence of the diffusion component on the current level of S. In Eqs. (16) and (17), λ is the average number of jumps per year and y is the jump size, which can be drawn from a normal or a double exponential distribution (see [49]).

An alternative way of representing the conditional variance is to use a stochastic volatility model, in which volatility is driven by its own noise (see the entry by Andersen and Benzoni on ► Stochastic Volatility). Stochastic volatility models are advantageous in that they are very flexible and have representations in both discrete and continuous time. The square root volatility model (also known as a scalar affine diffusion), proposed by Heston [44], is one of the most popular models in this area and is represented by the stochastic processes

$$\begin{aligned} \mathrm{d}\log(p_t) &= (\mu - 0.5\sigma_t)\,\mathrm{d}t + \sqrt{V_t}\mathrm{d}W_{1t} \\ \mathrm{d}V_t &= (\alpha - \beta\sigma_t)\,\mathrm{d}t + \sigma_V\sqrt{V_t}\mathrm{d}W_{2t} \end{aligned} \tag{18}$$

where V_t is the instantaneous (latent) stochastic volatility, which is assumed to follow a mean reverting square root process. The parameter k measures the speed of mean reversion, while θ is the unconditional long run mean. $\mathrm{d}W_{1t}$ and $\mathrm{d}W_{2t}$ are Brownian motions with instantaneous correlation $\rho\mathrm{d}t$.

Future Directions

The coverage in this essay has, unavoidably, been far from exhaustive. The realm of relevant nonlinear models and theories in finance is extremely rich and is developing fast (a useful review of new developments is [66]). By transcending the representative agent framework and by extending the standard notion of rationality, researchers are now allowing for interactions between heterogeneous groups of investors using agent based models (for an overview of these fascinating developments, see [51] and the entry on ► Finance, Agent Based Modeling in by Manzan). While such approaches can reproduce stylized facts such as volatility clustering and long-term dependencies, it remains to be seen how they can be standardized and applied to the solution of specific problems by academics and practitioners.

Bibliography

1. Aït-Sahalia Y (1999) Transition densities for interest rate and other nonlinear diffusions. J Finance 54:1361–1395
2. Akdeniz L, Altay-Salih A, Caner M (2003) Time varying betas help in asset pricing: The threshold CAPM. Stud Nonlinear Dyn Econom 6:1–16
3. Baillie RT (1996) Long memory processes and fractional integration in econometrics. J Econom 73:5–59
4. Baillie RT, Bollerslev T, Mikkelson HO (1996) Fractionally integrated generalized autoregressive conditional heteroskedasticity. J Econom 74:3–30

5. Bakshi G, Ju N, Yang H (2006) Estimation of continuous-time models with an application to equity volatility dynamics. J Financ Econom 82:227–249
6. Balke NS, Fomby TB (1997) Threshold cointegration. Int Econom Rev 38:627–645
7. Barkley Rosser J Jr (1999) On the complexities of complex economic dynamics. J Econom Perspect 13:169–192
8. Barndorff-Nielsen OE, Graversen SE, Shephard N (2004) Power variation and stochastic volatility: A review and some new results. J Appl Probab 41:133–143
9. Bauwens L, Laurent S, Rombouts JVK (2006) Multivariate GARCH models: A survey. J Appl Econom 21:79–109
10. Bera AK, Higgins ML (1993) On ARCH models: Properties, estimation and testing. J Econom Surv 7:305–366
11. Beran JA (1992) Statistical methods for data with long-range dependence. Stat Sci 7:404–427
12. Biais B, Glosten L, Spatt C (2005) Market microstructure: A survey of microfoundations, empirical results, and policy implications. J Financ Mark 8:217–264
13. Bollerslev T (1986) Generalised autoregressive conditional heteroskedasticity. J Econom 31:307–27
14. Bollerslev T, Chou RY, Kroner KF (1992) ARCH modelling in finance: A review of the theory and empirical evidence. J Econom 52:5–59
15. Bollerslev T, Engle RF, Nelson DB (1994) ARCH Models. In: Engle RF, McFadden DL (eds) Handbook of Econometrics, vol 4. New York, North-Holland, pp 2959–3038
16. Box GEP, Jenkins GM (1976) Time Series Analysis: Forecasting and Control. Rev. Edn., Holden Day, San Francisco
17. Brandt MW, Diebold FX (2006) A no-arbitrage approach to range-based estimation of return covariances and correlations. J Bus 79:61–74
18. Brock WA (1999) Scaling in economics: A reader's guide. Ind Corp Change 8:409–446
19. Brock WA, de Lima PJF (1996) Nonlinear time series, complexity theory, and finance. In: Maddala GS, Rao RS (eds) Handbook of Statistics, vol 14. Elsevier, Amsterdam, pp 317–361
20. Brock WA, Dechert WD, Scheinkman JA, LeBaron B (1996) A test for independence based on the correlation dimension. Econom Rev 15:197–235
21. Brooks C (2006) Multivariate Volatility Models. In: Mills TC, Patterson K (eds) Palgrave Handbook of Econometrics, vol 1. Econometric Theory. Palgrave Macmillan, Basingstoke, pp 765–783
22. Campbell JY, Lo AW, MacKinlay AC (1997) The Econometrics of Financial Markets. Princeton University Press, New Jersey
23. Chan KC, Karolyi A, Longstaff FA, Sanders AB (1992) An empirical comparison of alternative models of the short-term interest rate. J Finance 47:1209–1227
24. Cochrane JH (2005) Asset Pricing. Princeton University Press, Princeton
25. Cox JC, Ingersoll JE, Ross SA (1985) A theory of the term structure of interest rates. Econometrica 53:385–408
26. Dixit AK, Pindyck RS (1994) Investment under Uncertainty. Princeton University Press, Princeton
27. Engle RF (1982) Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. Econometrica 50:987–1008
28. Engle RF (2000) The econometrics of ultra-high frequency data. Econometrica 68:1–22
29. Engle RF, Granger CWJ (1987) Cointegration and error correction: representation, estimation and testing. Econometrica 55:251–276
30. Engle RF, Lilien DM, Robins RP (1987) Estimating time varying risk premia in the term structure: the ARCH-M model. Econometrica 55:391–408
31. Eraker B, Johannes M, Polson N (2003) The impact of jumps in volatility and returns. J Finance 53:1269–1300
32. Fama EF (1991) Efficient capital markets, vol II. J Finance 26:1575–1617
33. Fama EF (1998) Market efficiency, long-term returns, and behavioural finance. J Financ Econom 49:283–306
34. Frankel FA, Froot KA (1987) Using survey data to test propositions regarding exchange rate expectations. Am Econom Rev 77:33–153
35. Frankel FA, Froot KA (1988) Chartists, fundamentalists and the demand for dollars. Greek Econom Rev 10:49–102
36. Froot KA, Obstfeld M (1991) Exchange-rate dynamics under stochastic regime shifts – A unified approach. J Int Econom 31:203–229
37. Goldberg MD, Frydman R (1996) Imperfect knowledge and behaviour in the foreign exchange market. Econom J 106: 869–893
38. Goodhart CAE, O'Hara M (1997) High frequency data in financial markets: Issues and applications. J Empir Finance 4:73–114
39. Granger CWJ, Joyeux R (1980) An introduction to long memory time series models and fractional differencing. J Time Ser Anal 1:15–29
40. Granger CWJ, Teräsvirta T (1993) Modelling Nonlinear Economic Relationships. Oxford University Press, New York
41. Hamilton JD (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57:357–384
42. Hamilton JD (1990) Analysis of time series subject to changes in regime. J Econom 45:39–70
43. Hansen BE (1994) Autoregressive conditional density estimation. Int Econom Rev 35:705–730
44. Heston SL (1993) A closed-form solution for options with stochastic volatility with applications to bond and currency options. Rev Financ Stud 6:327–343
45. Hosking JRM (1981) Fractional differencing. Biometrika 68:165–176
46. Hull JC (2005) Options, Futures and Other Derivatives, 6th edn. Prentice Hall, Upper Saddle River
47. Jondeau E, Rockinger M (2003) Conditional volatility, skewness, and kurtosis: existence, persistence, and comovements. J Econom Dyn Control 27:1699–1737
48. Kon S (1984) Models of stock returns – a comparison. J Finance 39:147–65
49. Kou SG (2002) A jump-diffusion model for option pricing. Management Sci 48:1086–1101
50. Krugman PR (1991) Target zones and exchange-rate dynamics. Q J Econom 106:669–682
51. LeBaron B (2006) Agent-based Computational Finance. In: Tesfatsion L, Judd K (eds) Handbook of Computational Economics. North-Holland, Amsterdam, pp 1187–1232
52. Li WK, Ling S, McAleer M (2002) Recent theoretical results for time series models with GARCH errors. J Econom Surv 16: 245–269
53. Lorenz HW (1989) Nonlinear Dynamical Economics and Chaotic Motion. Springer, New York

54. Mandelbrot BB (1963) New methods in statistical economics. J Political Econom 71:421–440
55. Mandelbrot BB (1963) The variation of certain speculative prices. J Bus 36:394–419
56. Mandelbrot BB (1969) Long-run linearity, locally Gaussian process, H-spectra, and infinite variances. Int Econom Rev 10: 82–111
57. Mandelbrot BB (1972) Statistical methodology for nonperiodic cycles: From the covariance to R/S analysis. Ann Econom Soc Measurement 1/3:259–290
58. Mandelbrot BB (1989) Louis Bachelier. In: The New Palgrave: Finance. Macmillan, London, pp 86–88
59. Mandelbrot BB (1997) Three fractal models in finance: Discontinuity, concentration, risk. Econom Notes 26:171–211
60. Markellos RN, Mills TC (2003) Asset pricing dynamics. Eur J Finance 9:533–556
61. Merton RC (1976) Option prices when underlying stock returns are discontinuous. J Financ Econom 3:125–144
62. Merton RC (1998) Applications of option-pricing theory: Twenty-five years later. Am Econom Rev 88:323–347
63. Mills TC, Markellos RN (2008) The Econometric Modelling of Financial Time Series, 3rd edn. Cambridge University Press, Cambridge
64. Mittnik S, Rachev ST (1993) Modeling asset returns with alternative stable distributions. Econom Rev 12:261–330
65. Mittnik S, Rachev ST (1993) Reply to comments on "Modeling asset returns with alternative stable distributions" and some extensions. Econom Rev 12:347–389
66. Mizrach B (2008) Nonlinear Time Series Analysis. In: Blume L, Durlauf S (eds) The New Palgrave Dictionary of Economics, 2nd edn. Macmillan, London, pp 4611–4616
67. Mullineux A, Peng W (1993) Nonlinear business cycle modelling. J Econom Surv 7:41–83
68. Pan J (2002) The jump-risk premia implicit in options: Evidence from an integrated time-series study. J Financ Econom 63:3–50
69. Poitras G (2000) The Early History of Financial Economics. Edward Elgar, Cheltenham, pp 1478–1776
70. Potter S (1999) Nonlinear time series modelling: An introduction. J Econom Surv 13:505–528
71. Rachev ST, Menn C, Fabozzi FJ (2005) Fat Tailed and Skewed Asset Distributions. Wiley, New York
72. Racine JS, Ullah A (2006) Nonparametric Econometrics. In: Mills TC, Patterson K (eds) Palgrave Handbook of Econometrics, vol 1: Econometric Theory. Palgrave Macmillan, Basingstoke, pp 1001–1034
73. Robinson PM (2003) Long Memory Time Series. In: Robinson PM (ed) Time Series with Long Memory. Oxford University Press, London, pp 4–32
74. Rosser JB Jr (1991) From Catastrophe to Chaos: A General Theory of Economic Discontinuities. Kluwer Academic, Norwell
75. Rostow WW (1993) Nonlinear Dynamics: Implications for economics in historical perspective. In: Day RH, Chen P (eds) Nonlinear Dynamics and Evolutionary Economics. Oxford University Press, Oxford
76. Samorodnitsky G, Taqqu MS (1994) Stable Non-Gaussian Random Processes. Chapman and Hall, New York
77. Schumpeter JA (1939) Business Cycles. McGraw-Hill, New York
78. Schwartz E (1997) The stochastic behavior of commodity prices: Implications for valuation and hedging. J Finance 52:923–973
79. Teräsvirta T (2006) An Introduction to Univariate GARCH Models. In: Andersen TG, Davis RA, Kreiss JP, Mikosch T (eds) Handbook of Financial Time Series. Springer, New York
80. Teräsvirta T, Medeiros MC, Rech G (2006) Building neural network models for time series: A statistical approach. J Forecast 25:49–75
81. Tong H (1990) Nonlinear Time Series: A Dynamical Systems Approach. Oxford University Press, Oxford
82. Velasco C (2006) Semiparametric Estimation of Long-Memory Models. In: Mills TC, Patterson K (eds) Palgrave Handbook of Econometrics, vol 1: Econometric Theory. Palgrave MacMillan, Basingstoke, pp 353–95

Financial Economics, Return Predictability and Market Efficiency

Stijn van Nieuwerburgh[1], Ralph S. J. Koijen[2]
[1] Department of Finance, Stern School of Business, New York University, New York, USA
[2] Department of Finance, Tilburg University, Tilburg, The Netherlands

Article Outline

Glossary
Definition of the Subject
Introduction
Motivating Predictive Regressions
Structural Model
Future Directions
Bibliography

Glossary

Stock return The stock return in this entry refers to the return on the portfolio of all stocks that are traded on the three largest equity markets in the US: the NYSE, NASDAQ, and AMEX. The return is measured as the price of the stock at the end of the year plus the dividends received during the year divided by the price at the beginning of the year. The return of each stock is weighted by its market capitalization when forming the portfolio. The source for the data is CRSP.

Dividend-price ratio and dividend yield The dividend-price ratio of a stock is the ratio of the dividends received during the year divided by the price of the stock at the *end* of the year. The dividend yield, instead, is the ratio of the dividends received during the year divided by the price of the stock at the *beginning* of the year. The stock return is the sum of the dividend yield and the capital gain yield, which measures the ratio of the end-of-year stock price to the beginning-of-year stock price.

Predictability A stock return r_{t+1} is said to be predictable by some variable x_t if the expected return conditional on x_t, $E[r_{t+1} \mid x_t]$, is different from the unconditional expected return, $E[r_{t+1}]$. No predictability means that the best predictor of tomorrow's return is the constant, unconditional average return, i. e., $E[r_{t+1} \mid x_t] = E[r_{t+1}]$. When stock returns are unpredictable, stock prices are said to follow a random walk.

Market model The market model links the return on any asset i, r_{it} to the return on the market portfolio (r_t). Under joint normality of returns, it holds:

$$r_{it} = \alpha_i + \beta_i r_t + \varepsilon_{it} \,, \tag{1}$$

with $E[\varepsilon_{it}] = 0$ and $\mathrm{Var}[\varepsilon_{it}] = \sigma^2_{\varepsilon_i}$, see [16]. The typical assumption in the literature until the 1980s has been that $E[r]$ is constant.

Definition of the Subject

The efficient market hypothesis, due to [21,22] and [23], states that financial markets are efficient with respect to a particular information set when prices aggregate all available information. Testing the efficient market hypothesis requires a "market model" which specifies how information is incorporated into asset prices. Efficiency of markets is then synonymous with the inability of investors to make economic, i. e., risk-adjusted, profits based on this information set [36]. The question of market efficiency and return predictability is of tremendous importance for investors and academics alike. For investors, the presence of return predictability would lead to different optimal asset allocation rules. Failing to make portfolios conditional on this information may lead to substantial welfare losses. For academics, return predictability or the lack thereof has substantial implications for general equilibrium models that are able to accurately describe the risks and returns in financial markets.

Introduction

Until the 1980s, the standard market model assumed constant expected returns. The first empirical evidence, which showed evidence that returns were predictable to some extent, was therefore interpreted as a sign of market inefficiency [25,54]. [56] proposed the alternative explanation of time-varying expected returns. This prompted the question of why aggregate stock market returns would be time varying in equilibrium. [23] provides a summary of this debate.

Recently developed general equilibrium models show that expected returns can indeed be time varying, even if markets are efficient. Time-variation in expected returns can result from time-varying risk aversion [11], long-run consumption risk [5], or time-variation in risk-sharing opportunities, captured by variation in housing collateral [44]. Predictability of stock returns is now, by-and-large, interpreted as evidence of time-varying expected returns rather than market inefficiency.

Motivating Predictive Regressions

Define the gross return on an equity investment between period t and period $t+1$ as

$$R_{t+1} = \frac{P_{t+1} + D_{t+1}}{P_t} \,,$$

where P denotes the stock price and D denotes the dividend. [9] log-linearizes the definition of a return to obtain:

$$r_{t+1} = k + \Delta d_{t+1} + \rho dp_{t+1} - dp_t \,. \tag{2}$$

All lower-case letters denote variables in logs; d_t stands for dividends, p_t stands for the price, $dp_t \equiv d_t - p_t$ is the log dividend–price ratio, and r_t stands for the return. The constants k and $\rho = (1 + \exp(\overline{dp}))^{-1}$ are related to the long-run average log dividend–price ratio $\overline{dp}$. By iterating forward on Eq. (2) and by imposing a transversality condition (i. e., we rule out rational bubbles), one obtains

$$dp_t = \overline{dp} + E_t \sum_{j=1}^{\infty} \rho^{j-1} \left[(r_{t+j} - \overline{r}) - (\Delta d_{t+j} - \overline{d}) \right] . \tag{3}$$

Since this equation holds both ex-post and ex-ante, an expectation operator can be added on the right-hand side. This equation is one of the central tenets of the return predictability literature, the so-called Campbell and Shiller [12,13] equation. It says that, as long as the expected returns and expected dividend growth are stationary, deviations of the dividend–price ratio (dp_t) from its long-term mean ($\overline{dp}$) ought to forecast either future returns, or future dividend growth rates, or both.

This accounting identity has motivated some of the earliest empirical work in return predictability, which regressed returns on the lagged dividend–price ratio, as in Eq. (4):

$$(r_{t+1} - \bar{r}) = \kappa_r (dp_t - \overline{dp}) + \tau^r_{t+1} \,, \tag{4}$$

$$(\Delta d_{t+1} - \overline{d}) = \kappa_d (dp_t - \overline{dp}) + \tau^d_{t+1} \,, \tag{5}$$

$$(dp_{t+1} - \overline{dp}) = \phi (dp_t - \overline{dp}) + \tau^{dp}_{t+1} \,, \tag{6}$$

where $\bar{r}$ is the long-run mean return and τ^r is a mean-zero innovation. The logic of (3) suggests that the dividend–

price ratio could predict future dividend growth rates instead of, or in addition to, future returns. Testing for dividend growth predictability would lead one to estimate Eq. (5), where $\overline{d}$ denotes the long-run mean log dividend growth.

The empirical return predictability literature started out by estimating Eq. (4) with the dividend–price ratio on the right-hand side; see [12,17,24,29,34,53] and [42], among others. It found evidence for return predictability, i. e., $\kappa_r > 0$. This finding was initially interpreted as evidence against the efficient market hypothesis.

Around the same time, [25] and [52] document a negative autocorrelation in long-horizon returns. Good past returns forecast bad future returns. [16] and [18] summarize the evidence based on long-horizon autocorrelations and variance ratios, and conclude that the statistical evidence in favor of mean reversion in long-horizon returns is weak, possibly due to small sample problems. This motivates [4] to use a large cross-section of countries and use a panel approach instead. They in turn document strong evidence in favor of mean-reversion of long-horizon returns with an estimated half-life of 3–3.5 years.

Second, other financial ratios, such as the earnings-price ratio or the book-to-market ratio, or macro-economic variables such as the consumption-wealth ratio, the labor income-to-consumption ratio, or the housing collateral ratio, as well as corporate decisions, and the cross-sectional price of risk have subsequently been shown to predict returns as well; see [3,38,39,43,45,50] and [51], among others.

Third, long-horizon returns are typically found to be more predictable than one-period ahead returns. The coefficient $\kappa_r(H)$ in the H-period regression

$$\sum_{j=1}^{H} r_{t+j} = \kappa_r(H)\, dp_t + \tau^r_{t,t+H} \tag{7}$$

exceeds the coefficient κ_r in the one-period regression. This finding is interpreted as evidence for the fact that the time-varying component in expected returns is quite persistent.

Fourth, these studies conclude that growth rates of fundamentals, such as dividends or earnings, are much less forecastable than returns using financial ratios. This suggests that most of the variation of financial ratios is due to variation in expected returns.

Fifth, predictability of stock returns does not only arise for the US. Studies by [10,26,33], and [2] analyze a large cross-section of countries and find evidence in favor of predictability by financial ratios in some countries, even though the evidence is mixed. More robust results are documented for the predictive ability of term structure variables.

These conclusions regarding predictability of stock returns are controversial because the forecasting relationship of financial ratios and future stock returns exhibits three disconcerting statistical features. First, correct inference is problematic because financial ratios are extremely persistent. The empirical literature typically augments Eq. (4) with an auto-regressive specification for the predictor variable, as in Eq. (6), where $\overline{dp}$ is the long-run mean of the dividend–price ratio. The estimated autoregressive parameter ϕ is near unity and standard tests leave the possibility of a unit root open (i. e., $\phi = 1$). [2,27,46,55] and [58] conclude that the statistical evidence of forecastability is weaker once tests are adjusted for high persistence. [1,2,15,42,57] and [20] derive asymptotic distributions for predictability coefficients under the assumption that the forecasting variable follows a local-to-unit root, yet stationary, process.

Second, financial ratios have poor out-of-sample forecasting power, as shown in [7,31], and [32], but see [35] and [14] for different interpretations of the out-of-sample tests and evidence.

Third, the forecasting relationship of returns and financial ratios exhibits significant instability over time. Figure 1 shows that in rolling 30-year regressions of annual log CRSP value-weighted returns on lagged log dividend–price ratios, the ordinary least squares (OLS) regression coefficient varies between zero and 0.5 and the associated R^2 ranges from close to zero to 30% depending on the subsample.

The figure plots estimation results for the equation $r_{t+1} - \bar{r} = \kappa_r(dp_t - \overline{dp}) + \tau^r_{t+1}$. It shows the estimates for κ_r using 30-year rolling windows. The dashed line in the left panels denote the point estimate plus or minus one standard deviation. The standard errors are asymptotic. The parameters $\bar{r}$ and $\overline{dp}$ are the sample means of log returns r and the log dividend–price ratio dp. The data are annual for 1927–2004.

[60] and [49] report evidence in favor of breaks in the OLS coefficient in the forecasting regression of returns on the lagged dividend–price ratio, while [41] report evidence for structural shifts in $\overline{dp}$. [47] use Bayesian methods to estimate structural breaks in the equity premium.

Empirical Evidence Revisited

Table 1 reviews the empirical evidence using annual value-weighted CRSP log return, dividend growth, and dividend–price ratio data for 1927–2004. In Panel A, the sys-

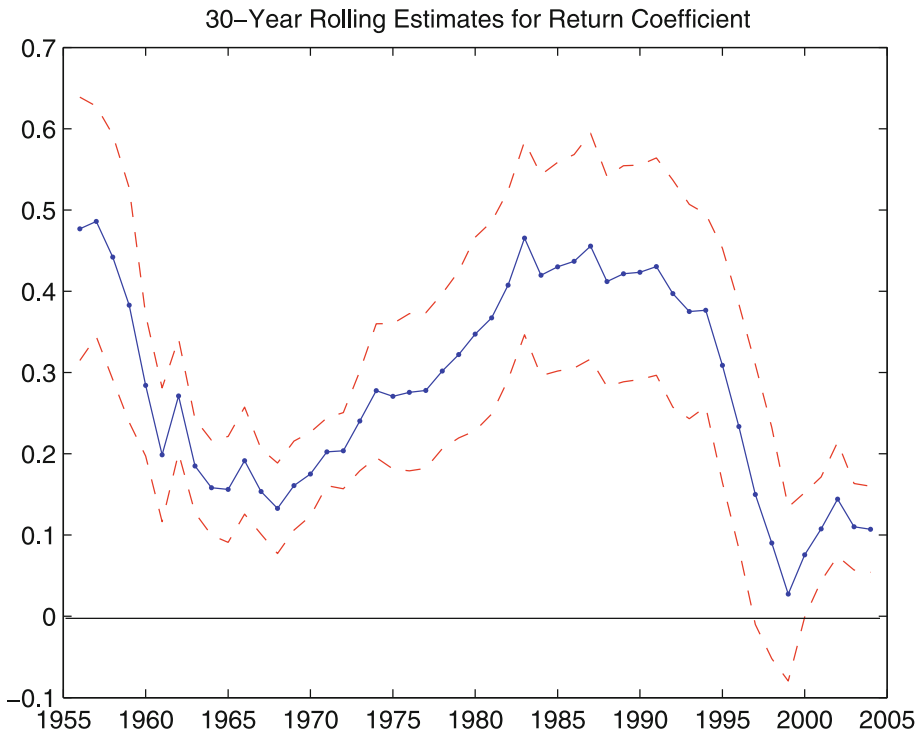

Financial Economics, Return Predictability and Market Efficiency, Figure 1
Parameter Instability in Return Predictability Coefficient

tem of Eqs. (4) and (5) is estimated by GMM. The first row indicates that a higher dividend–price ratio leads to a higher return ($\kappa_r = .094$ in Column 2) and a higher dividend growth rate ($\kappa_d = .005$ in Column 1). The latter coefficient has the wrong sign, but the coefficient is statistically indistinguishable from zero. The asymptotic standard error on the estimate for κ_r is .046. The corresponding asymptotic p-value is 3.6% so that κ_r is statistically different from zero at conventional levels. In other words, the dividend–price ratio seems to predict stock returns, but not dividend growth. A similar result holds if returns in excess of a risk-free rate are used, or real returns instead of nominal returns.

[41] conduct an extensive Monte Carlo analysis to investigate the small-sample properties of estimates for κ_r and κ_d. Consistent with [55], the estimate for κ_r displays an upward small-sample bias. In addition, the standard error on κ_r is understated by the asymptotic standard error. As a result, one can no longer reject the null hypothesis that κ_r is zero. Based on this evidence, one is tempted to conclude that neither returns nor dividend growth are forecastable.

The second and third rows implement the suggestion of [41] to correct the long-run mean dividend–price ratio, $\overline{dp}$, for structural breaks. The data strongly suggest either one break in 1991, or two breaks in 1954 and 1994 in favor of either no breaks or three breaks. This break-adjusted dividend–price ratio is less persistent and less volatile. Its

Financial Economics, Return Predictability and Market Efficiency, Table 1
Return and Dividend Growth Predictability in the Data

	κ_d	κ_r	ϕ	PV violation
Panel A: No Long-Horizon Moments $H = \{1\}$				
No Break	.005	.094	.945	−.046
	(.037)	(.046)	(.052)	
1 Break ('91)	.019	.235	.813	.004
	(.047)	(.055)	(.052)	
2 Breaks ('54, '94)	.124	.455	.694	−.001
	(.073)	(.079)	(.070)	
Panel B: Long-Horizon Moments $H = \{1, 3, 5\}$				
No Break	.021	.068	.990	.189
	(.018)	(.038)	(.032)	
1 Break ('91)	.012	.210	.834	.076
	(.019)	(.043)	(.042)	
2 Breaks ('54, '94)	.080	.409	.697	.100
	(.065)	(.078)	(.060)	

lower persistence alleviates the econometric issues mentioned above.

The second row of Table 1 uses the one-break adjusted dividend–price ratio as a regressor in the return and dividend growth predictability equations. The evidence in favor of return predictability is substantially strengthened. The point estimate for κ_r more than doubles to .235, and is highly significant. In the two-break case in the third row, the point estimate further doubles to 0.455. The small-sample bias in κ_r is negligible relative to the size of the coefficient. The R^2 of the return equation is 10% in the one-break case and even 23% in the two-break case. This compares to 3.8% in the no-break case. Furthermore, rolling regression estimates of κ_r indicate that it is much more stable over time when the break-adjusted dp series is used as a regressor. The dividend growth coefficient κ_d remains statistically indistinguishable from zero. This evidence strengthens the view that returns are predictable and dividend growth is not, and that these findings are not an artefact of statistical issues.

This table reports GMM estimates for the parameters $(\kappa_d, \kappa_r, \phi)$ and their asymptotic standard errors (in parentheses). The results in panel A are for the system with one-year ahead equations for dividend growth and returns ($H = 1$, $N = 0$). The results in panel B are for the system with one-year, three-year and five-year ahead equations for dividend growth and returns ($H = \{1, 3, 5\}$, $N = 2$). The first-stage GMM weighting matrix is the identity matrix. The asymptotic standard errors and p-values are computed using the Newey–West HAC procedure (second stage weighting matrix) with four lags in panel A and $H = 5$ lags in panel B. The last column denotes the present-value constraint violation of the univariate OLS slope estimators: $(1 - \rho\phi^{\text{ols}})^{-1}(\kappa_r^{\text{ols}} - \kappa_d^{\text{ols}})$. It is expressed in the same units as κ_d and κ_r. In panel B this number is the average violation of the three constraints, one constraint at each horizon. The dividend–price ratio in rows 1 and 4 is the unadjusted one. In rows 2 and 5, the dividend–price ratio is adjusted for one break in 1991, and in rows 3 and 6, it is the series adjusted for two breaks in 1954 and 1994. All estimation results are for the annual sample 1927–2004.

Structural Model

What are researchers estimating when they run the return predictability regression (4)? How are the return and dividend growth predictability regressions in (4) and (5) related? To answer these important questions, we set up a simple structural model with time-varying expected returns and expected dividend growth rates. This structural model has the system of Eqs. (4)–(6) as its reduced-form. The main purpose of this model is to show that (i) the dividend–price ratio is a contaminated predictor of returns and dividend growth rates, (ii) that the parameters in (4)–(6) have to satisfy a cross-equation restriction, which we call the *present-value constraint*, and (iii) this restriction enables decomposing the dividend–price ratio into expected returns and expected dividend growth. Similar models can be derived for financial ratios other than the dividend–price ratio (e. g., [61]). [6] show how stock returns and book-to-market ratios are related in a general equilibrium model.

A Present-Value Model

We assume that expected dividend growth, z, and expected returns, x, follow an AR(1) process with autoregressive coefficient ϕ:

$$\Delta d_{t+1} - \overline{d} = z_t + \epsilon_{t+1}\,, \qquad z_{t+1} = \phi z_t + \zeta_{t+1}\,, \quad (8)$$

$$r_{t+1} - \bar{r} = x_t + \eta_{t+1}\,, \qquad x_{t+1} = \phi x_t + \xi_{t+1}\,. \quad (9)$$

The model has three fundamental shocks: an innovation in unexpected dividends ϵ_{t+1}, an innovation in expected dividends ζ_{t+1}, and an innovation in expected returns ξ_{t+1}. We assume that all three errors are serially uncorrelated and have zero cross-covariance at all leads and lags: $\text{Cov}(\epsilon_{t+1}, \zeta_{t+j}) = 0,\ \forall j \neq 1, \text{Cov}(\xi_{t+1}, \zeta_{t+j}) = 0,\ \forall j \neq 1$, and $\text{Cov}(\epsilon_{t+1}, \xi_{t+j}) = 0,\ \forall j$, except for a contemporaneous correlation between expected return and expected dividend growth innovations $\text{Cov}(\zeta_t, \xi_t) = \chi$, and a correlation between expected and unexpected dividend growth innovations $\text{Cov}(\zeta_t, \epsilon_t) = \lambda$. We discuss innovations to unexpected returns η below.

In steady-state, the log dividend–price ratio is a function of the long-run mean return and dividend growth rate $\overline{dp} = \log\big((\bar{r} - \overline{d})/(1 + \overline{d})\big)$. The log dividend–price ratio in (3) can then be written as:

$$dp_t - \overline{dp} = \frac{x_t - z_t}{1 - \rho\phi}\,. \quad (10)$$

The dividend–price ratio is the difference of two AR(1) processes with the same root ϕ, which is again an AR(1) process. I.e., we recover Eq. (6).

The return decomposition in [9] implies that the innovation to unexpected returns follows from the three fundamental shocks (i. e., combine (2) with (8)–(10)):

$$\eta_{t+1} = \frac{-\rho}{1 - \rho\phi}\xi_{t+1} + \frac{\rho}{1 - \rho\phi}\zeta_{t+1} + \epsilon_{t+1}\,. \quad (11)$$

Since both ρ and ϕ are positive and $\rho\phi < 1$, a positive shock to expected returns leads, ceteris paribus, to a neg-

ative contemporaneous return. Likewise, a shock to expected or unexpected dividend growth induces a positive contemporaneous return.

Contaminated Predictor

The first main insight from the structural model is that the demeaned dividend–price ratio in (10) is an imperfect forecaster of both returns and dividend growth. Returns are predicted by x_t (see Eq. (9)), but variation in the dividend–price ratio is not only due to variation in x, but also in expected dividend growth z_t. The same argument applies to dividend growth which is predicted by z_t (see Eq. (8)). This implies that the regressions in the reduced-form model in (4) and (5) suffer from an errors-in-variables problem [24,30,37].

To illustrate the bias, we can link the regression coefficients κ_r and κ_d explicitly to the underlying structural parameters:

$$\kappa_r = \frac{\mathrm{Cov}(r_{t+1}, dp_t)}{\mathrm{Var}(dp_t)} = \frac{(1-\rho\phi)(\sigma_\xi^2 - \chi)}{\sigma_\xi^2 + \sigma_\zeta^2 - 2\chi}\,, \tag{12}$$

$$\kappa_d = \frac{\mathrm{Cov}(\Delta d_{t+1}, dp_t)}{\mathrm{Var}(dp_t)} = \frac{-(1-\rho\phi)(\sigma_\zeta^2 - \chi)}{\sigma_\xi^2 + \sigma_\zeta^2 - 2\chi}\,. \tag{13}$$

If growth rates are constant, i. e., $\chi = 0$ and $\sigma_\zeta = 0$, then the dividend–price ratio is a perfect predictor of returns and $\kappa_r^\star = 1 - \rho\phi$. In all other cases, there is a bias in the return predictability coefficient:

$$\kappa_r^\star - \kappa_r = \frac{(1-\rho\phi)(\sigma_\zeta^2 - \chi)}{\sigma_\xi^2 + \sigma_\zeta^2 - 2\chi}\,. \tag{14}$$

[24] argue that κ_r is downward biased ($\kappa_r^\star - \kappa_r > 0$). In fact, the structural parameters that are implied by the reduced-form model parameters indicate an upward bias. This occurs because the correlation between expected dividend growth and expected returns is sufficiently high.

A similar argument applies to κ_d. [40] construct a variable based on the co-integrating relationship between consumption, dividends from asset wealth, and dividends from human wealth. They show that this variable has strong predictive power for dividend growth, and they show that expected returns and expected growth rates are highly positively correlated. This implies that expected growth rates and expected returns have an offsetting effect on financial ratios, which makes it hard to reliably detect time-varying growth rates using such financial ratios.

Present-Value Constraint

The second main insight from the structural model is that there is a cross-equation restriction on the three innovations $\tau = (\tau^d, \tau^r, \tau^{dp})$ of the reduced-form model (4)–(6). Expressed in terms of the structural parameters, these innovations are:

$$\tau_{t+1}^d = \epsilon_{t+1} + x_t\left(\frac{-\kappa_d}{1-\rho\phi}\right) + z_t\left(\frac{\kappa_r}{1-\rho\phi}\right) \tag{15}$$

$$\tau_{t+1}^r = \epsilon_{t+1} + x_t\left(\frac{-\kappa_d}{1-\rho\phi}\right) + z_t\left(\frac{\kappa_r}{1-\rho\phi}\right) - \rho\left(\frac{\xi_{t+1} - \zeta_{t+1}}{1-\rho\phi}\right) \tag{16}$$

$$\tau_{t+1}^{dp} = \frac{\xi_{t+1} - \zeta_{t+1}}{1-\rho\phi}\,. \tag{17}$$

They imply the present value restriction:

$$\rho\tau_{t+1}^{dp} = \tau_{t+1}^d - \tau_{t+1}^r \;\Leftrightarrow\; \kappa_r - \kappa_d = 1 - \rho\phi\,. \tag{18}$$

Another way to write this restriction is as a restriction on a weighted sum of κ_r and κ_d: Any two equations from the system (4)–(6) implies the third. Evidence that dividend growth is not forecastable is evidence that returns are forecastable: if $\kappa_d = 0$ in Eq. (18), then $\kappa_r > 0$ because $\rho\phi < 1$. If estimating (5) uncovers that a high dividend–price ratio forecasts a higher future dividend growth ($\kappa_d > 0$), as we showed it does, then this strengthens the evidence for return predictability. [19] makes an important and closely related point: That it is important to impose the present-value relationship when testing the null hypothesis of no return predictability. That null ($\kappa_r = 0$) is truly a joint hypothesis, because it implies a negative coefficient in the dividend growth equation ($\kappa_d < 0$). [19], too, finds strong evidence for return predictability.

Returning to Panel A of Table 1, Column 3 backs out the AR(1) coefficient ϕ from the estimated κ_d and κ_r, and from the present-value constraint (18).[1] In the first row, $\phi = .945$, and is statistically undistinguishable from a unit root. This high persistence is a familiar result in the literature. The last column reports the left-hand side and the right-hand side of Eq. (18) for *univariate* OLS regressions of (4)–(6). It shows the violation of the present-value constraint. In the first row, the violation is half as large as the actual point estimate κ_r. The standard OLS point estimates do not satisfy the present-value constraint, which can lead to faulty inference.

However, when we use the break-adjusted dividend–price ratio series in rows 2 and 3, we find that (1) the persistence of the break-adjusted dp ratio is much lower

[1] The linearization parameter ρ is tied to the average dividend–price ratio, and is held fixed at 0.9635.

than the unadjusted series (.81 and .69 versus .95), and (2) the present-value constraint is satisfied by the OLS coefficients.

A similar present-value constraint can be derived for long-horizon return and dividend growth regressions:

$$\kappa_r(H) = \kappa_r \left(\frac{1-\phi^H}{1-\phi} \right)$$
$$\kappa_d(H) = \kappa_d \left(\frac{1-\phi^H}{1-\phi} \right) .$$

Not only are the coefficients on the long-horizon return predictability regressions for all horizons linked to each other (see [8]), all long-horizon regression coefficients in the return equations are also linked to those from the dividend growth equations. I. e., there is one present-value constraint for each horizon H. Imposing these restrictions in a joint estimation procedure improves efficiency.

Panel B of Table 1 shows the results from a joint estimation of 1-year, 3-year, and 5-year cumulative returns and dividend growth rates on the lagged dividend–price ratio. Because of the restrictions, there are only two parameters to be estimated from these six equations. The results are close to those from the one-year system in Panel A, confirming the main message of [8]. The main conclusion remains that returns are strongly predictable, and dividend growth rates are not.

Exploiting Correlation in Innovations

The present-value model implies a restriction on the innovations in returns and the dividend–price ratio (see Eq. (18)). A third main insight from the structural model is that this correlation contains useful information for estimating the structural parameters, and hence for how much return predictability and dividend growth predictability there truly is. [48] show that exploiting the correlation between expected and unexpected stock returns can lead to substantially more accurate estimates. The information in correlations is incorporated by specifying a prior belief about the correlation between expected and unexpected returns, and updating that prior in a Bayesian fashion using observed data. Their method ignores the present-value constraint. The structural parameters in Panel B of Table 1, which impose the present-value constraint, imply that two-thirds of the variability in the price-dividend ratio is due to expected future returns and one-third is due to expected future dividend growth rates.

Likewise, [59] write down a model like (8)–(9) where expected returns and growth rates of dividends are autoregressive, exploiting the present-value constraint. Because the price-dividend ratio is linear in expected returns x and expected dividend growth z (see Eq. (10)), its innovations in (17) can be attributed to either innovations in expected returns or expected growth rates. The present-value constraint enables one to disentangle the information in price-dividend ratios about both expected returns and growth rates, and therefore to undo the contamination coming from correlated innovations. With this decomposition in hand, it is then possible to recover the full time-series of expected returns, x, and expected growth rates, z. [59] show that the resulting processes are strong predictors of realized returns and realized dividend growth rates, respectively. This underscores the importance of specifying a present-value model to address return predictability.

Geometric or Arithmetic Returns

As a final comment, most predictive regressions are estimated using geometric, i. e. log returns, instead of arithmetic, i. e. simple returns. This choice is predominantly motivated by the [12] log-linearization discussed before. Since investors are ultimately interested in arithmetic instead of log returns, [59] specify a process for expected simple returns instead. This is made possible by applying the techniques of linearity-inducing models, recently introduced by [28].

Future Directions

The efficient market hypothesis, which states that markets efficiently aggregate all information, was first interpreted to mean that returns are not predictable. Early evidence of predictability of stock returns by the lagged dividend–price ratio seemed to be evidence against the efficient market hypothesis. However, return predictability and efficient markets are not incompatible because return predictability arises naturally in a world with time-varying expected returns. In the last 15 years, the empirical literature has raised a set of statistical objections to return predictability findings. Meanwhile, the theoretical literature has progressed, seemingly independently, in its pursuit of new ways to build models with time-varying expected returns. Only very recently has it become clear that theory is necessary to understand the empirical facts.

In this entry, we have set up a simple present-value model with time-varying expected returns that generates the regression that is the focus of the empirical literature. The model also features time-varying expected dividend growth. It shows that the dividend–price ratio contains information about both expected returns and expected dividend growth. A regression of returns on the dividend–price ratio may therefore be a poor indicator of the true ex-

tent of return predictability. At the same time, the present-value model provides a solution to this problem: It disentangles the two pieces of information in the price-dividend ratio. This allows us to interpret the standard predictability regressions in a meaningful way. Combining data with the present-value model, we conclude that there is strong evidence for return predictability. We interpret this as evidence for the presence of time-varying expected returns, not evidence against the efficient market hypothesis. The main challenge for the future is to better understand the underlying reasons for this time-variation.

Bibliography

Primary Literature

1. Amihud Y, Hurvich CM (2004) Predictive regressions: A reduced-bias estimation method. Financial Quant Anal 39:813–841
2. Ang A, Bekaert G (2007) Stock return predictability: Is it there? Rev Financial Stud 20(3):651–707
3. Baker M, Wurgler J (2000) The equity share in new issues and aggregate stock returns. J Finance 55:2219–2258
4. Balvers R, Wu Y, Gilliland E (2000) Mean reversion across national stock markets and parametric contrarian investment strategies. J Finance 55:745–772
5. Bansal R, Yaron A (2004) Risks for the long-run: A potential resolution of asset pricing puzzles. J Finance 59(4):1481–1509
6. Berk JB, Green RC, Naik V (1999) Optimal investment, growth options and security returns. J Finance 54:1153–1607
7. Bossaerts P, Hillion P (1999) Implementing statistical criteria to select return forecasting models: What do we learn? Rev Financial Stud 12:405–428
8. Boudoukh J, Richardson M, Whitelaw RF (2007) The myth of long-horizon predictability. Rev Financial Stud (forthcoming)
9. Campbell JY (1991) A variance decomposition for stock returns. Econ J 101:157–179
10. Campbell JY (2003) Consumption-based asset pricing. In: Constantinides G, Harris M, Stulz R (eds) Handbook of the Economics of Finance. North-Holland, Amsterdam (forthcoming)
11. Campbell JY, Cochrane JH (1999) By force of habit: A consumption-based explanation of aggregate stock market behavior. J Political Econ 107:205–251
12. Campbell JY, Shiller RJ (1988) The dividend–price ratio and expectations of future dividends and discount factors. Rev Financial Stud 1:195–227
13. Campbell JY, Shiller RJ (1991) Yield spreads and interest rates: A bird's eye view. Rev Econ Stud 58:495–514
14. Campbell JY, Thompson S (2007) Predicting excess stock returns out of sample: Can anything beat the historical average? Rev Financial Stud (forthcoming)
15. Campbell JY, Yogo M (2002) Efficient tests of stock return predictability. Harvard University (unpublished paper)
16. Campbell JY, Lo AW, MacKinlay C (1997) The Econometrics of Financial Markets. Princeton University Press, Princeton
17. Cochrane JH (1991) Explaining the variance of price-dividend ratios. Rev Financial Stud 5(2):243–280
18. Cochrane JH (2001) Asset Pricing. Princeton University Press, Princeton
19. Cochrane JH (2006) The dog that did not bark: A defense of return predictability. University of Chicago Graduate School of Business (unpublished paper)
20. Eliasz P (2005) Optimal median unbiased estimation of coefficients on highly persistent regressors. Department of Economics, Princeton University (unpublished paper)
21. Fama EF (1965) The behavior of stock market prices. J Bus 38:34–101
22. Fama EF (1970) Efficient capital markets: A review of theory and empirical work. J Finance 25:383–417
23. Fama EF (1991) Efficient markets: II. J Finance 46(5):1575–1618
24. Fama EF, French KR (1988) Dividend yields and expected stock returns. J Financial Econ 22:3–27
25. Fama EF, French KR (1988) Permanent and temporary components of stock prices. J Political Econ 96(2):246–273
26. Ferson WE, Harvey CR (1993) The risk and predictability of international equity returns. Rev Financial Stud 6:527–566
27. Ferson WE, Sarkissian S, Simin TT (2003) Spurious regressions in financial economics? J Finance 58(4):1393–1413
28. Gabaix X (2007) Linearity-generating processes: A modelling tool yielding closed forms for asset prices. MIT (working paper)
29. Goetzman WN, Jorion P (1993) Testing the predictive power of dividend yields. J Finance 48:663–679
30. Goetzman WN, Jorion P (1995) A longer look at dividend yields. J Bus 68:483–508
31. Goyal A, Welch I (2003) Predicting the equity premium with dividend ratios. Manag Sci 49(5):639–654
32. Goyal A, Welch I (2006) A comprehensive look at the empirical performance of the equity premium prediction. Rev Financial Stud (forthcoming)
33. Hjalmarsson E (2004) On the predictability of global stock returns, Yale University (unpublished paper)
34. Hodrick R (1992) Dividend yields and expected stock returns: Alternative procedures for inference and measurement. Rev Financial Stud 5:357–386
35. Inoue A, Kilian L (2004) In-sample or out-of-sample tests of predictability: Which one should we use? Econom Rev 23:371–402
36. Jensen MC (1978) Some anomalous evidence regarding market efficiency. J Financial Econ 6:95–101
37. Kothari S, Shanken J (1992) Stock return variation and expected dividends: A time-series and cross-sectional analysis. J Financial Econ 31:177–210
38. Lamont O (1998) Earnings and expected returns. J Finance 53:1563–87
39. Lettau M, Ludvigson SC (2001) Consumption, aggregate wealth and expected stock returns. J Finance 56(3):815–849
40. Lettau M, Ludvigson SC (2005) Expected returns and expected dividend growth. J Financial Econ 76:583–626
41. Lettau M, Van Nieuwerburgh S (2006) Reconciling the return predictability evidence. Rev Financial Stud (forthcoming)
42. Lewellen JW (2004) Predicting returns with financial ratios. J Financial Econ 74(2):209–235
43. Lustig H, Van Nieuwerburgh S (2005) Housing collateral, consumption insurance and risk premia: An empirical perspective. J Finance 60(3):1167–1219
44. Lustig H, Van Nieuwerburgh S (2006) Can housing collateral explain long-run swings in asset returns? University of California at Los Angeles and New York University (unpublished manuscript)
45. Menzly L, Santos T, Veronesi P (2004) Understanding predictability. J Political Econ 112(1):1–47

46. Nelson CC, Kim MJ (1993) Predictable stock returns: The role of small sample bias. J Finance 43:641–661
47. Pastor L, Stambaugh RF (2001) The equity premium and structural breaks. J Finance 56(4):1207–1239
48. Pastor L, Stambaugh RF (2006) Predictive systems: Living with imperfect predictors, graduate School of Business. University of Chicago Journal of Finance (forthcoming)
49. Paye BS, Timmermann A (2006) Instability of return prediction models. J Empir Finance 13(3):274–315
50. Piazzesi M, Schneider M, Tuzel S (2007) Housing, consumption, and asset pricing. J Financial Econ 83(March):531–569
51. Polk C, Thompson S, Vuolteenaho T (2006) Cross-sectional forecasts of the equity risk premium. J Financial Econ 81: 101–141
52. Poterba JM, Summers LH (1988) Mean reversion in stock returns: Evidence and implications. J Financial Econ 22:27–60
53. Rozeff MS (1984) Dividend yields are equity risk premia. J Portfolio Manag 49:141–160
54. Shiller RJ (1984) Stock prices and social dynamics. Brook Pap Econ Act 2:457–498
55. Stambaugh RF (1999) Predictive regressions. J Financial Econ 54:375–421
56. Summers LH (1986) Does the stock market rationally reflect fundamental values? J Finance 41:591–601
57. Torous W, Volkanov R, Yan S (2004) On predicting returns with nearly integrated explanatory variables. J Bus 77:937–966
58. Valkanov R (2003) Long-horizon regressions: Theoretical results and applications. J Financial Econ 68:201–232
59. van Binsbergen J, Koijen RS (2007) Predictive regressions: A present-value approach. Duke University (working paper)
60. Viceira L (1996) Testing for structural change in the predictability of asset returns. Harvard University (unpublished manuscript)
61. Vuolteenaho T (2000) Understanding the aggregate book-market ratio and its implications to current equity-premium expectations. Harvard University (unpublished paper)

Books and Reviews

Campbell JY, Lo AW, MacKinlay C (1997) The Econometrics of Financial Markets. Princeton University Press, Princeton
Cochrane JH (2005) Asset Pricing. Princeton University Press, Princeton, NJ
Malkiel BG (2004) A Random Walk Down Wall Street. W.W. Norton, New York

Financial Economics, Time Variation in the Market Return

Mark J. Kamstra[1], Lisa A. Kramer[2]
[1] Schulich School of Business, York University, Toronto, Canada
[2] Rotman School of Management, University of Toronto, Toronto, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
Valuation
The Equity Premium Puzzle
Time-Varying Equity Premia: Possible Biological Origins
Future Directions
Bibliography

Glossary

AR(k) An autoregressive process of order k; a time series model allowing for first order dependence; for instance, an AR(1) model is written as $y_t = \alpha + \rho_1 y_{t-1} + \epsilon_t$ where α and ρ are parameters, ρ is typically assumed to be less than 1 in absolute value, and ϵ_t is an innovation term, often assumed to be Gaussian, independent, and identically distributed over t.

ARCH(q) A special case of the GARCH(p, q) model (see below) where $p = 0$.

Basis point A hundredth of one percent.

Bootstrap A computer intensive resampling procedure, where random draws with replacement from an original sample are used, for instance to perform inference.

Discount rate The rate of return used to discount future cashflows, typically calculated as a risk-free rate (e. g. the 90-day US T-bill rate) plus an equity risk premium.

Equity premium puzzle The empirical observation that the ex post equity premium (see entry below) is higher than is indicated by financial theory.

Ex ante equity premium The extra return investors *expect* they will receive for holding risky assets, over and above the return they would receive for holding a risk-free asset like a Treasury bill. "Ex ante" refers to the fact that the expectation is formed in advance.

Ex post equity premium The extra return investors received *after* having held a risky asset for some period of time. The ex post equity premium often differs from the ex ante equity premium due to random events that impact a risky asset's return.

Free cash flows Cash flows that could be withdrawn from a firm without lowering the firm's current rate of growth. Free cash flows are substantially different from accounting earnings and even accounting measures of the cash flow of a firm.

Fundamental valuation The practice of determining a stock's intrinsic value by discounting cash flows to their present value using the required rate of return.

GARCH(p, q) Generalized autoregressive conditional heteroskedasticity of order (p, q), where p is the order of the lagged variance terms and q is the order of the lagged squared error terms; a time series model

allowing for dependence in the conditional variance of a random variable, y. A GARCH(1,1) model is specified as:

$$y_t = \alpha + \epsilon_t; \quad \epsilon_t \sim \left(0, h_t^2\right)$$
$$h_t^2 = \theta + \beta h_{t-1}^2 + \gamma \epsilon_{t-1}^2 ,$$

where α, θ, β, and γ are parameters and ϵ_t is an innovation term.

Market anomalies Empirical regularities in financial market prices or returns that are difficult to reconcile with conventional theories and/or valuation methods.

Markov model A model of a probabilistic process where the random variable can only take on a finite number of different values, typically called states.

Method of moments A technique for estimating parameters (like parameters of the conditional mean and conditional variance) by matching sample moments, then solving the equations for the parameters to be estimated.

SAD Seasonal Affective Disorder, a medical condition by which reduced daylight in the fall and winter leads to seasonal depression for roughly ten percent of the world's population.

Sensation seeking A measure used by psychologists to capture an individual's degree of risk tolerance. High sensation-seeking tendency correlates with low risk tolerance, including tolerance for risk of a financial nature.

Simulated method of moments A modified version of the method of moments (see entry above) that is based on Monte Carlo simulation, used in situations when the computation of analytic solutions is infeasible.

Definition of the Subject

The realized return to any given asset varies over time, occasionally in a dramatic fashion. The value of an asset, its *expected* return, and its volatility, are of great interest to investors and to policy makers. An asset's expected return in excess of the return to a riskless asset (such as a short-term US Treasury bill) is termed the equity premium. The value of the equity premium is central to the valuation of risky assets, and hence a much effort has been devoted to determining the value of the equity premium, whether it varies, and if it varies, how predictable it is. Any evidence of predictable returns is either evidence of a predictably varying equity premium (say, because risk varies predictably) or a challenge to the rationality of markets and the efficient allocation of our society's scarce resources.

In this article, we start by considering the topic of valuation, with emphasis on simulation-based techniques. We consider the valuation of income-generating assets in the context of a constant equity premium, and we also explore the consequences of allowing some time-variation and predictability in the equity premium. Next we consider the equity premium puzzle, discussing a simulation-based technique which allows for precise estimation of the value of the equity premium, and which suggests some constraints on the types of models that should be used for specifying the equity premium process. Finally, we focus on evidence of seasonally varying expected returns in financial markets. We consider evidence that as a whole either presents some challenges to traditional hypotheses of efficient markets, or suggests agents' risk tolerance may vary over time.

Introduction

The pricing of a firm is conceptually straightforward. One approach to valuing a firm is to use historical dividend payments and discount rate data to forecast future payments and discount rates. Restrictions on the dividend and discount rate processes are typically imposed to produce an analytic solution to the fundamental valuation equation (an equation that involves calculating the expectation of an infinite sum of discounted dividends).

Common among many of the available valuation techniques is some form of consideration of multiple scenarios, including good and bad growth and discount rate evolutions, with valuation based on a weighted average of prices from the various scenarios. The valuation technique we focus some attention on, the Donaldson and Kamstra [14] (henceforth DK) methodology, is similar to pricing path-dependent options, as it utilizes Monte Carlo simulation techniques and numerical integration of the possible paths followed by the joint processes of dividend growth and discount rates, explicitly allowing path-dependence of the evolutions. The DK method is very similar in spirit to other approaches in the valuation literature which consider multiple scenarios. One distinguishing feature of the DK methodology we consider is the technique it employs for modeling the discount rate.

Cochrane [9] highlights three interesting approaches for modeling the discount rate: a constant discount rate, a consumption-based discount rate, and a discount rate equal to some variable reference return plus a risk premium. Virtually the entire valuation literature limits its attention to the constant discount rate case, as constant discount rates lead to closed-form solutions to many valuation formulas. DK explore all three methods for modeling the discount rate and find they lead to qualitatively similar results. However, their quantitative results indicate an

overall better fit to the price and return data when using a reference return plus a risk premium. Given DK's findings, we use a discount rate equal to some variable reference return plus a risk premium. In implementing this approach for modeling the discount rate used in valuation, it is simplest to assume a *constant* equity premium is added to the reference rate, in particular since the reference rate is permitted to vary (since it is typically proxied using a variable rate like the three-month US T-bill rate). We do not, however, restrict ourselves to the constant equity premium case.

Using the simulation-based valuation methodology of DK and the method of simulated moments, we explore the evidence for a time-varying equity premium and its implications for a long-standing puzzle in financial economics, the equity premium puzzle of Mehra and Prescott [51]. Over the past century the average annual return to investing in the US stock market has been roughly 6% higher than the return to investing in risk-free US T-bills. Making use of consumption-based asset-pricing models, Mehra and Prescott argue that consumption within the US has not been sufficiently volatile to warrant such a large premium on risky stocks relative to riskless bonds, leading them to describe this large premium as the "equity premium puzzle."

Utilizing simulations of the distribution from which ex post equity premia are drawn, conditional on various possible values for investors' ex ante equity premium and calibrated to S&P 500 dividends and US interest rates, we present statistical tests that show a true ex ante equity premium as low as 2% could easily produce ex post premia of 6%. This result is consistent with the well-known observation that ex post equity premia are observed with error, and a large range of realized equity premia are consistent with any given value of the ex ante equity premium. Examining the marginal and joint distributions of financial statistics like price-dividend ratios and return volatility that arise in the simulations versus actual realizations from the US economy, we argue that the range of ex ante equity premia most consistent with the US market data is very close to 3.5%, and the ex ante equity premium process is very unlikely to be constant over time.

A natural question to ask is why might the equity premium fluctuate over time? There are only two likely explanations: changing risk or changing risk aversion. Evidence from the asset-pricing literature, including [20,37,49], and many others shows that priced risk varies over time. We explore some evidence that risk aversion itself may vary over time, as revealed in what is often termed market anomalies. Market anomalies are variations in expected returns which appear to be incongruous with variations in discount rates or risk. The most stark anomalies have to do with deterministic asset return seasonalities, including seasonalities at the weekly frequency such as the weekend effect (below-average equity returns on Mondays), annual effects like the above-average equity returns typically witnessed in the month of January, and other effects like the lower-than-average equity returns often witnessed following daylight saving time-change weekends, and opposing cyclicality in bond versus equity returns correlated to the length of day (known as the SAD effect). We briefly review some of these outstanding puzzles, focusing our attention on the SAD effect and the daylight saving effect.

Valuation

Overview

We begin our discussion of valuation with a broad survey of the literature, including dividend-based valuation, relative valuation, and accounting-based methods. We introduce dividend-based valuation first.

Fundamental valuation techniques that utilize dividends in a *discrete* time framework include Gordon [25], Hawkins [30], Michaud and Davis [53], Farrell [22], Sorensen and Williamson [73], Rappaport [63], Barsky and DeLong [2], Hurley and Johnson [33], [34], Donaldson and Kamstra [14], and Yao [78]. Invariably these approaches are partial equilibrium solutions to the valuation exercise. Papers that use *continuous* time tools to evaluate the fundamental present value equation include Campbell and Kyle [6], Chiang, Davidson, and Okuney [8], Dong and Hirshleifer [17], and Bakshi and Chen [3]. The Dong and Hirshleifer [17] and Bakshi and Chen [3] papers conduct valuation by assuming dividends are proportional to earnings and then modeling earnings. Continuous time papers in this literature typically start with the representative agent/complete markets economic paradigm. Models are derived from primitive assumptions on markets and preferences, such as the equilibrium condition that there exist no arbitrage opportunities, dividend (cash flow) growth rates follow an Ornstein–Uhlenbeck mean-reverting process, and preferences over consumption are represented by the log utility function. Time-varying stochastic discount rates (i. e. the pricing kernel) fall out of the marginal rate of utility of consumption in these models, and the solution to the fundamental valuation problem is derived with the same tools used to price financial derivatives. A critique of dividend-discounting methods is that dividends are typically smoothed and are set low enough so that the dividend payments can be maintained through economic downturns. Authors such as Hackel and Livnat (see p. 9 in [27]) argue that these sorts of considerations

imply that historical records of dividend payments may therefore be poor indicators of future cash payments to investors.

A distinct valuation approach, popular amongst practitioners, determines the value of inactively traded firms by finding an actively traded firm that has similar risk, profitability, and investment-opportunity characteristics and then multiplying the actively traded firm's price-earnings (P/E) ratio by the inactively traded firm's earnings. This approach to valuation is often referred to as the relative value method or the constant P/E model. References to this sort of approach can be found in textbooks like [4], and journal articles such as [60,62].

There are also several valuation approaches that are based on the book value of equity, abnormal earnings, and free-cash flows. These approaches are linked to dividends and hence to formal fundamental valuation by well-established accounting relationships. They produce price estimates by valuing firm assets and income streams. The most popular of this class of techniques include the residual income and free-cash-flow methods. See [23,57,61] for further information. All of these valuation methods implicitly or explicitly take the present value of the stream of firm-issued dividends to the investor. The motivation for considering accounting relationships is that these accounting measures are not easily manipulated by firms and so should reflect more accurately the ability of firms to generate cashflows and hence allow more accurate assessments of the fundamental value of a firm than techniques based on dividends.

Fundamental Valuation Methods in Detail

Now that we have surveyed the valuation literature in general, we turn to a formal derivation of several *fundamental* valuation techniques. Investor rationality requires that the current market price P_t of a stock which will pay a per share dividend (cash payment) D_{t+1} one period from now and then sell for P_{t+1}, discounting payments received during period t (i. e., from the beginning of period t to the beginning of period $t+1$) at rate r_t, must satisfy Eq. (1):

$$P_t = \mathcal{E}_t \left\{ \frac{P_{t+1} + D_{t+1}}{1 + r_t} \right\} . \tag{1}$$

$\mathcal{E}_t$ is the expectations operator conditional on information available up to the end of period t. Solving Eq. (1) forward under the transversality condition that the expected present value of P_{t+k} goes to zero as k goes to infinity (a "no-bubble" assumption) produces the familiar result that the market price equals the expected present value of future dividends (cash payments); i. e.,

$$P_t = \sum_{k=0}^{\infty} \mathcal{E}_t \left\{ \left(\prod_{i=0}^{k} \left[\frac{1}{1 + r_{t+i}} \right] \right) D_{t+k+1} \right\} . \tag{2}$$

Defining the growth rate of dividends from the beginning of period t to the beginning of period $t+1$ as $g_t^{\mathrm{d}} \equiv (D_{t+1} - D_t)/D_t$ it follows that

$$P_t = D_t \mathcal{E}_t \left\{ \sum_{k=1}^{\infty} \left(\prod_{i=0}^{k} \left[\frac{1 + g_{t+i}^{\mathrm{d}}}{1 + r_{t+i}} \right] \right) \right\} . \tag{3}$$

Equation (3) is the fundamental valuation equation, which is not controversial and can be derived under the law of one price and non-satiation alone, as by Rubinstein [69] and others. Notice that the cash payments D_{t+k} include all cash disbursements from the firm, including cash dividends and share repurchases. Fundamental valuation methods based directly on Eq. (3) are typically called dividend discount models.

Perhaps the most famous valuation estimate based on Eq. (3) comes from the Gordon [25] Growth Model. If dividend growth rates and discount rates are constant, then it is straightforward to derive the Gordon fundamental price estimate from Eq. (3):

$$P_t^{\mathrm{G}} = D_t \left[\frac{1 + g^{\mathrm{d}}}{r - g^{\mathrm{d}}} \right] , \tag{4}$$

where r is the constant discount rate value and g^{d} is the (conditionally) constant growth rate of dividends. To produce the Gordon Growth Model valuation estimate, all we need are estimates of the dividend growth rate and discount rate, which can be obtained in a variety of ways, including the use of historically observed dividends and returns.

Extensions of the Gordon Growth Model exploit the fundamental valuation equation, imposing less stringent assumptions. The simple Gordon Growth Model imposes a constant growth rate on dividends (dividends are expected to grow at the same rate every period) while Hurley and Johnson [33] and [34] and Yao [78] develop Markov models (models that presume a fixed probability of, say, maintaining the dividend payment at current levels, and a probability of raising it, thus incorporating more realistic dividend growth processes). Two examples of these models found in Yao [78] are the Additive Markov Gordon model (Eq. (1) of Yao [78] and the Geometric Markov Gordon model (Eq. (2) of Yao [78]). These models can be interpreted as considering different scenarios for dividend growth for a particular asset, estimating the appropriate

price for the asset under each scenario, and then averaging the prices using as weights the probability of given scenarios being observed.

The Additive Markov Gordon Growth Model is:

$$P_t^{\mathrm{ADD}} = D_t/r + \left[1/r + (1/r)^2\right]\left(q^{\mathrm{u}} - q^{\mathrm{d}}\right)\Delta\,, \tag{5}$$

where r is the average discount rate, q^{u} is the proportion of the time the dividend increases, q^{d} is the proportion of the time the dividend decreases, and $\Delta = \sum_{t=2}^{T}|D_t - D_{t-1}|/(T-1)$ is the average absolute value of the level change in the dividend payment.

The Geometric Markov Gordon Growth Model is:

$$P_t^{\mathrm{GEO}} = D_t\left[\frac{1 + (q^{\mathrm{u}} - q^{\mathrm{d}})\Delta^{\%}}{r - (q^{\mathrm{u}} - q^{\mathrm{d}})\Delta^{\%}}\right]\,, \tag{6}$$

where $\Delta^{\%} = \sum_{t=2}^{T}|(D_t - D_{t-1})/D_{t-1}|/(T-1)$ is the average absolute value of the percentage rate of change in the dividend payment.

The method of DK is also an extension of the Gordon Growth Model, taking the discounted dividend growth model of Eq. (3) and re-writing it as

$$P_t = D_t\sum_{k=0}^{\infty}\mathcal{E}_t\left\{\prod_{i=0}^{k} y_{t+i}\right\}\,, \tag{7}$$

where $y_{t+i} = (1 + g_{t+i}^{\mathrm{d}})/(1 + r_{t+i})$ is the discounted dividend growth rate. Under the DK method, the fundamental price is calculated by forecasting the range of possible evolutions of y_{t+i} up to some distant point in the future, period $t + I$, calculating $PV = D_t\sum_{k=0}^{I}(\prod_{i=0}^{k} y_{t+i})$ for each possible evolution of y_{t+i}, and averaging these values of PV across all the possible evolutions. (The value of I is chosen to produce a very small truncation error. Values of $I = 400$ to 500 for annual data have been found by DK to suffice). In this way, the DK approach mirrors other extensions of the Gordon Growth Model. It is primarily distinguished from other approaches that extend the Gordon Growth Model in two regards. First, more sophisticated time series models, estimated with historical data, are used to generate the different outcomes (scenarios) by application of Monte Carlo simulation. Second, in contrast to typical modeling in which only dividend growth rates vary, the joint evolution of cashflow growth rates and discount rates are explicitly modeled as time-varying.

Among the attractive features of the free-cash-flow and residual income valuation methods is that they avoid the problem of forecasting dividends, by exploiting relationships between accounting data and dividends. It is the practical problem of forecasting dividends to infinity that have led many researchers to explore methods based on accounting data. See, for instance, Penman and Sougiannis [61].

Assume a flat term structure (i. e., a constant discount rate $r_t = r$ for all t) and write

$$P_t = \sum_{k=1}^{\infty}\frac{\mathcal{E}_t\{D_{t+k}\}}{(1+r)^k}\,. \tag{8}$$

The clean-surplus relationship relating dividends to earnings is invoked in order to derive the residual income model:

$$B_{t+k} = B_{t+k-1} + E_{t+k} - D_{t+k}\,, \tag{9}$$

where B_{t+k} is book value and E_{t+k} is earnings per share. Solving for D_{t+k} in Eq. (9) and substituting into Eq. (8) yields

$$P_t = \sum_{k=1}^{\infty}\frac{\mathcal{E}_t\{B_{t+k-1} + E_{t+k} - B_{t+k}\}}{(1+r)^k}\,,$$

or

$$\begin{aligned} P_t &= B_t + \sum_{k=1}^{\infty}\frac{\mathcal{E}_t\{E_{t+k} - r\cdot B_{t+k-1}\}}{(1+r)^k} - \frac{\mathcal{E}_t\{B_{t+\infty}\}}{(1+r)^{\infty}} \\ &= B_t + \sum_{k=1}^{\infty}\frac{\mathcal{E}_t\{E_{t+k} - r\cdot B_{t+k-1}\}}{(1+r)^k}\,, \end{aligned} \tag{10}$$

where $B_{t+\infty}/(1+r)^{\infty}$ is assumed to equal zero. $E_{t+k} - r\cdot B_{t+k-1}$ is termed abnormal earnings.

To derive the free cash flow valuation model, we relate dividends to cash flows with a financial assets relation in place of the clean surplus relation:

$$fa_{t+k} = fa_{t+k-1} + i_{t+k} + c_{t+k} - D_{t+k}\,, \tag{11}$$

where fa_{t+k} is financial assets net of financial obligations, i_{t+k} is interest revenues net of interest expenses, and c_{t+k} is cash flows realized from operating activities net of investments in operating activities, all of which can be positive or negative. A net interest relation is often assumed,

$$i_{t+k} = rfa_{t+k-1}\,. \tag{12}$$

See Fetham and Ohlson [23] for further discussion. Solving for D_{t+k} in Eq. (11) and substituting into Eq. (8), utilizing Eq. (12) and assuming the discounted present value of financial assets fa_{t+k} goes to zero as k increases, yields the free-cash-flow valuation equation:

$$P_t = fa_t + \sum_{k=1}^{\infty}\frac{\mathcal{E}_t\{c_{t+k}\}}{(1+r)^k}\,. \tag{13}$$

More on the Fundamental Valuation Method of Donaldson and Kamstra

A number of approaches can be taken to conduct valuation using the DK model shown in Eq. (7). By imposing a very simple structure for the conditional expectation of discounted dividend growth rate (y_t in Eq. (7)), the expression can be solved analytically, for instance by assuming that the discounted dividend growth rate is a constant. As shown by DK, however, analytic solutions become complex for even simple ARMA models, and with sufficient non-linearity, the analytics can be intractable. For this reason, we present a general solution algorithm based on the DK method of Monte Carlo simulation.

This method simulates y_t into the future and performs a numerical (Monte Carlo) integration to estimate the terms $\{\prod_{k=0}^{i} y_{t+k}\}$ where $y_{t+k} = (1 + g^{\mathrm{d}}_{t+k})/(1 + r_{t+k})$ in the classic case of a dividend-paying firm. A general heuristic is as follows:

Step I: Model y_t, $t = 1, \ldots, T$, as conditionally time-varying, for instance as an AR(k)-GARCH(p, q) process, and use the estimated model to make conditional mean forecasts $\hat{y}_t$, $t = 1, \ldots, T$, and variance forecasts, conditional on data observed only before period t. Ensure that this model is consistent with theory, for instance that the mean level of y is less than one. This mean value can be calibrated to available data, such as the mean annual y value of 0.94 observed in the last 50 years of S&P 500 data. Recall that although analytic solutions are available for simple processes, the algorithm presented here is applicable to virtually arbitrarily non-linear conditional processes for the discounted cash payment rate y.

Step IIa: Simulate discounted cash payment growth rates. That is, produce ys that might be observed in period t given what is known at period $t-1$. To do this for a given period t, simulate a population of J independent possible shocks (say draws from a normal distribution with mean zero and appropriate variance, or bootstrapped from the data) $\epsilon_{t,j}$, $j = 1, \ldots, J$, and add these shocks separately to the conditional mean forecast $\hat{y}_t$ from Step I, producing $y_{t,j} = \hat{y}_t + \epsilon_{t,j}$, $j = 1, \ldots, J$. The result is a simulated cross-section of J possible realizations of y_t standing at time $t-1$, i. e. different paths the economy may take next period.

Step IIb: Use the estimated model from Step I to make the conditional mean forecast $\hat{y}_{t+1,j}$, conditional on only the jth realization for period t (i. e., $y_{t,j}$ and $\epsilon_{t,j}$) and the data known at period $t-1$, to form $y_{t+1,j}$.

Step IIc: Repeat Step IIb to form $y_{t+2,j}, y_{t+3,j}, \ldots, y_{t+I,j}$ for each of the J economies, where I is the number of periods into the future at which the simulation is truncated. Form the perfect foresight present value ($P^*_{t,j}$) for each of the J possible economies:

$$P^*_{t,j} = A_t\Big(y_{t,j} + y_{t,j}y_{t+1,j} + y_{t,j}y_{t+1,j}y_{t+2,j} + \cdots + \prod_{i=0}^{I} y_{t+i,j}\Big); \quad j = 1, \ldots, J\,.$$

Provided I is chosen to be large enough, the truncated terms $\prod_{i=0}^{K} y_{t+i,j}$, $K = I+1, \ldots, \infty$ will be negligible.

Step III: Calculate the DK fundamental price for each $t = 1, \ldots, T$:

$$P^{\mathrm{DK}}_t = \sum_{j=1}^{J} P^*_{t,j}/J\,. \tag{14}$$

These fundamental price estimates P^{DK}_t can be compared to the actual price (if market prices exist) at the beginning of period t to test for bubbles as demonstrated by DK, or if period t is the future, P^{DK}_t is the fundamental price forecast. This procedure is represented diagrammatically in Exhibit 1.

To illustrate the sort of forecasts that can be produced using this technique, we illustrate graphically the S&P 500 index over the past 100 years together with predicted values based on the Gordon Growth Model and the DK method. The free-cash-flow and residual income methods are not easily adapted to forecasting index prices like the S&P 500, and so are omitted here. The type of data depicted in the following figure is described in some detail by Kamstra [39].

Figure 1 has four panels. In the panels, we plot the level of the S&P 500 index (marked with bullets and a solid line) alongside price forecasts from each of the valuation techniques. In Panel A we plot the index together with the basic Gordon Growth Model price forecasts (marked with stars), in Panels B and C we plot the index together with the Additive and Geometric Gordon Growth Models' forecasts (with squares and triangles respectively), and in Panel D we plot the index alongside the DK method's forecasts (marked with diamonds). In each panel the price scale is logarithmic.

We see in Panels A, B, and C that the use of the any of the Gordon models for forming annual forecasts of the S&P 500 index level produces excessively smooth price forecasts. (If we had plotted return volatility, then the market returns would appear excessively volatile in comparison to to forecasted returns). Evidence of periods of inflated market prices relative to the forecasted prices, i. e.,

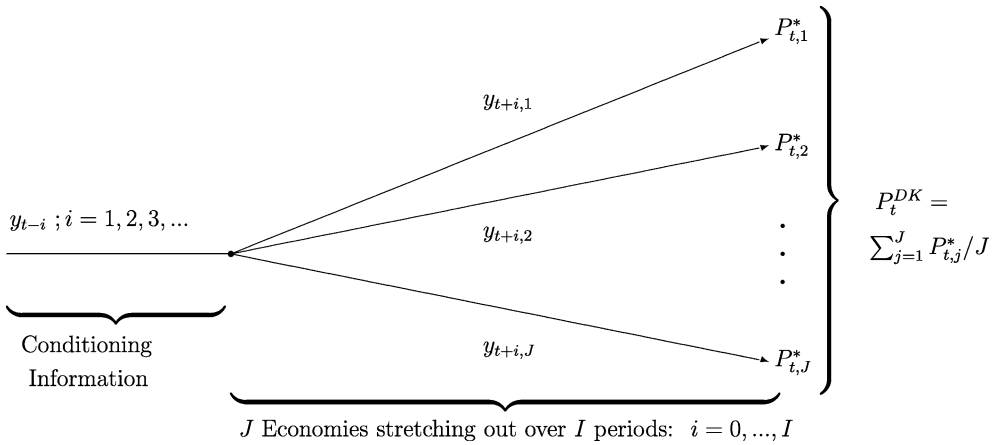

Financial Economics, Time Variation in the Market Return, Exhibit 1
Diagram of DK Monte Carlo integration

evidence of price bubbles, is apparent in the periods covering the 1920s, the 1960s, the last half of the 1980s, and the 1990s. However, if the Gordon models are too simple (since each Gordon-based model ignores the forecastable nature of discount rates and dividend growth rates), then this evidence may be misleading.

In Panel D, we see that the DK model is better able to capture the volatility of the market, including the boom of the 1920s, the 1960s and the 1980s. The relatively better performance of the DK price estimate highlights the importance of accounting for the slow fade rate of dividend growth rates and discount rates, i. e., the autocorrelation of these series. The failure of the DK method to capture the height of the 1990s boom leaves evidence of surprisingly high prices during the late 1990s. If the equity premium fell in the 1990s, as some researchers have speculated (see for instance Pástor and Stambaugh [59]), then all four sets of the plotted fundamental valuation forecasts would be expected to produce forecasts that undershoot actual prices in the 1990s, as all these methods incorporate a constant equity premium. If this premium were set too high, future cashflows would be discounted too aggressively, biasing the valuation methods downward.

The Equity Premium Puzzle

The fact that all four fundamental valuation methods we consider spectacularly fail to capture the price boom of the 1990s, possibly as a result of not allowing a time-varying equity premium, sets the stage to investigate the equity premium puzzle of Mehra and Prescott [51]. The equity premium is the extra return, or premium, that investors demand in order to be compelled to purchase risky stock instead of risk-free debt. We call this premium the ex ante equity premium (denoted π_e), and it is formally defined as the difference between the expected return on risky assets, $\mathcal{E}\{R\}$, and the expected risk-free rate, $\mathcal{E}\{r_f\}$:

$$\pi_e \equiv \mathcal{E}\{R\} - \mathcal{E}\{r_f\} \ . \tag{15}$$

The ex post equity premium is typically estimated using historical equity returns and risk-free rates, as we do not observe the ex ante premium. Define $\overline{R}$ as the average historical annual return on the S&P 500 and $\overline{r}_f$ as the average historical return on US T-bills. A standard approach to calculate ex post equity premium, $\hat{\pi}_e$, is:

$$\hat{\pi}_e \equiv \overline{R} - \overline{r}_f \ . \tag{16}$$

Of course it is unlikely that the stock return we estimate ex post equals investors' anticipated ex ante return. Thus a 6% ex post equity premium in the US data may not be a challenge to economic theory. The question we ask is therefore: if investors' true ex ante premium is X%, what is the probability that the US economy could randomly produce an ex post premium of at least 6%? We can then argue whether or not the 6% ex post premium observed in the US data is consistent with various ex ante premium values, X%, with which standard economic theory may be more compatible. We can also consider key financial statistics and yields from the US economy to investigate if an X% ex ante equity premium could likely be consistent with the combinations that have been observed, such as high Sharpe ratio and low dividend yields, low interest rates and high ex post equity premia, and so on.

Authors have investigated the extent to which ex ante considerations may impact the realized equity premium.

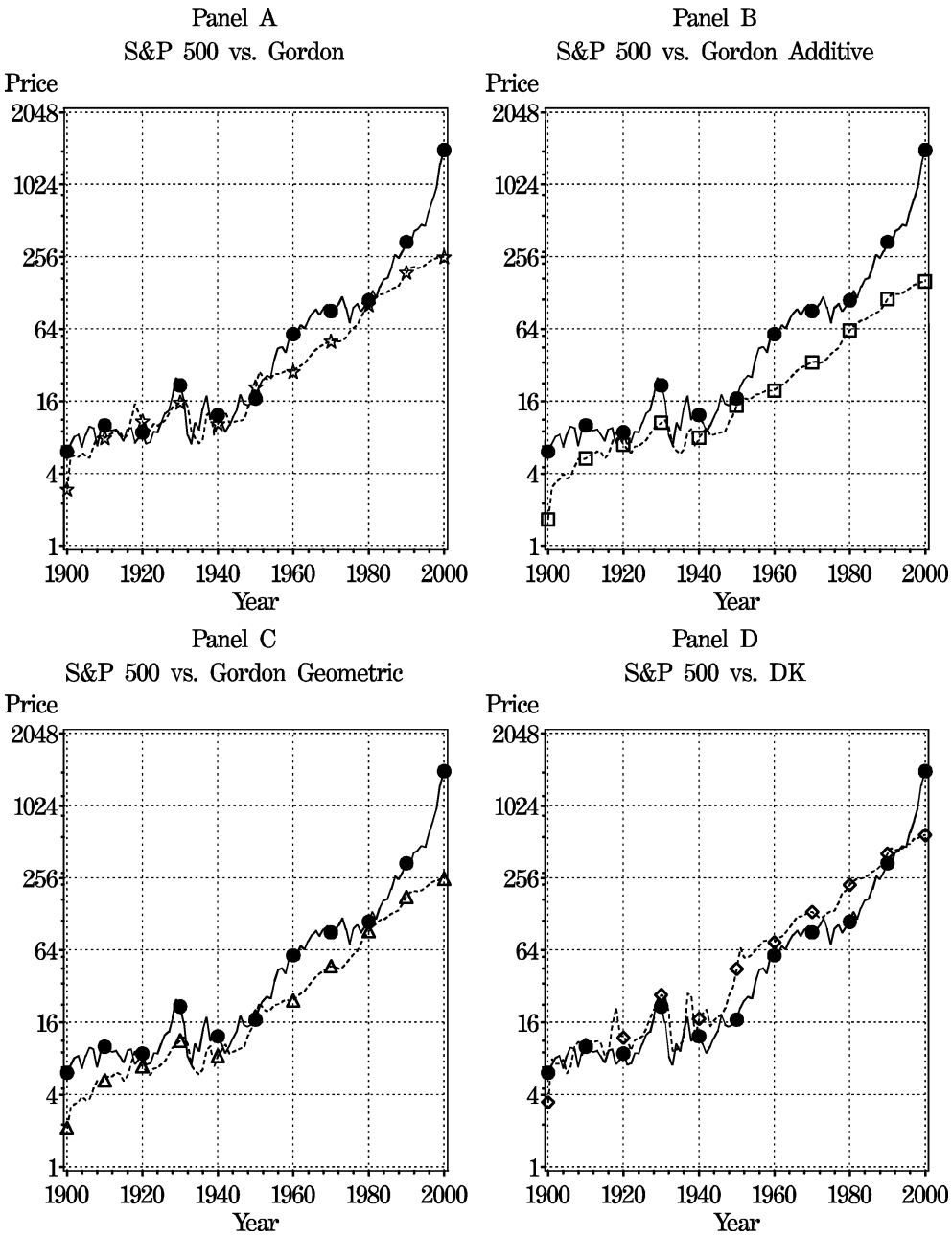

Financial Economics, Time Variation in the Market Return, Figure 1
S&P 500 index level versus price forecasts from four models. S&P 500 index: ●, Gordon Growth price: ★, Additive Gordon Growth price: □, Geometric Gordon Growth price: △, DK price ◇

For example, Rietz [65] investigated the effect that the fear of a serious, but never realized, depression would have on equilibrium asset prices and equity premia. Jorion and Goetzmann [38] take the approach of comparing the US stock market's performance with stock market experiences in many other countries. They find that, while some markets such as the US and Canada have done very well over the past century, other countries have not been so fortunate; average stock market returns from 1921 to 1996 in France, Belgium, and Italy, for example, are all close to zero, while countries such as Spain, Greece, and Romania have experienced negative returns. It is difficult, how-

ever, to conduct statistical tests because, first, the stock indices Jorion and Goetzmann consider are largely contemporaneous and returns from the various indices are not independent. Statistical tests would have to take into account the panel nature of the data and explicitly model covariances across countries. Second, many countries in the comparison pool are difficult to compare directly to the United States in terms of economic history and underlying data generating processes. (Economies like Egypt and Romania, for example may have equity premia generated from data generating processes that differ substantially from that of the US).

There are some recent papers that make use of fundamental information in examining the equity premium. One such paper, Fama and French [21], uses historical dividend yields and other fundamental information to calculate estimates of the equity premium which are smaller than previous estimates. Fama and French obtain point estimates of the ex post equity premium ranging from 2.55% (based on dividend growth rate fundamentals) to 4.78% (based on bias-adjusted earnings growth rate fundamentals), however these estimates have large standard errors. For example, for their point estimate of 4.32% based on non-bias-adjusted earnings growth rates, a 99% confidence interval stretches from approximately −1% to about 9%. Mehra and Prescott's [51] initially troubling estimate of 6% is easily within this confidence interval and is in fact within one standard deviation of the Fama and French point estimate.

Calibrating to economy-wide dividends and discount rates, Donaldson, Kamstra, and Kramer [16] employ simulation methods similar to DK to simulate a distribution of possible price and return outcomes. Comparing these simulated distributions with moments of the actual data then permits them to test various models for the equity premium process. Could a realized equity premium of 6% be consistent with an ex ante equity premium of 2%? Could an ex ante equity premium of 2% have produced the low dividend yields, high ex post equity premia, and high Sharpe ratios observed in the US over the last half century?

A summary of the basic methodology implemented by Donaldson, Kamstra, and Kramer [16], is as follows:

(a) Assume a mean value for the equity premium that investors demand when they first purchase stock (e. g., 2%) and a time series process for the premium, say a deterministic drift downward in the premium of 5 basis points per year, asymptoting no lower than perhaps 1%. This assumed premium is added to the risk-free interest rate to determine the discount rate that an investor would rationally apply to a forecasted dividend stream in order to calculate the present value of dividend-paying stock.
(b) Estimate econometric models for the time-series processes driving dividends and interest rates in the US economy (and, if necessary, for the equity premium process), allowing for autocorrelation and covariation. Then use these models to Monte Carlo simulate a variety of potential paths for US dividends, interest rates, and equity premia. The simulated paths are of course different in each of these simulated economies because different sequences of random innovations are applied to the common stochastic processes in each case. However, the key drivers of the simulated economies themselves are all still identical to those of the US economy since all economies share common stochastic processes fitted to US data.
(c) Given the assumed process for the equity premium investors demand ex ante (which is the same for all simulated economies in a given experiment), use a discounted-dividend model to calculate the fundamental stock returns (and hence ex post equity premia) that arise in each simulated economy. All economies have the same ex ante equity premium process, and yet all economies have different ex post equity premia. Given the returns and ex post equity premia for each economy, as well as the means of the interest rates and dividend growth rates produced for each economy, it is feasible to calculate various other important characteristics, like Sharpe ratios and dividend yields.
(d) Examine the distribution of ex post equity premia, interest rates, dividend growth rates, Sharpe ratios, and dividend yields that arise conditional on various values of the ex ante equity premia. Comparing the performance of the US economy with intersections of the various univariate and multivariate distributions of these quantities and conducting joint hypothesis tests allows the determination of a narrow range of equity premia consistent with the US market data. Note that this is the method of simulated moments, which is well adapted to estimate the ex ante equity premium. The simulated method of moments was developed by McFadden [50] and Pakes and Pollard [58]. Duffie and Singleton [18] and Corradi and Swanson [11] employ simulated method of moments in an asset pricing context.

Further details on the simulation methodology are provided by Donaldson, Kamstra, and Kramer [16]. They make use of annual US stock and Treasury data observed from 1952 through 2004, with the starting year of 1952 motivated by the US Federal Reserve Board's adoption of

a modern monetary policy regime in 1951. The model that generated the data we use to illustrate this simulation methodology is Model 6 of Donaldson, Kamstra, and Kramer [16], a model that allows for trending, autocorrelated, and co-varying dividend growth rates, interest rates and equity premia, as well as for a structural break in the equity premium process. We show later that allowing for trends and structural breaks in the equity premium process is a crucial factor in the model's ability to capture the behavior of the observed US market data.

We focus on the intuition behind the Donaldson, Kamstra, and Kramer technique by looking at bivariate plots of simulated data, conditional on various values of the ex ante equity premium. In every case, the pair of statistics we plot are dependent on each other in some way, allowing us to make interesting conditional statements. Among the bivariate distributions we consider, we will see some that serve primarily to confirm the ability of our simulations to produce the character and diversity of results observed in US markets. Some sets of figures rule out ex ante equity premia below 2.5% while others rule out ex ante equity premia above 4.5%. Viewed collectively, the figures serve to confirm that the range of ex ante equity premia consistent with US market data is in the close vicinity of 3.5%.

Figure 2 contains joint distributions of mean returns and return standard deviations arising in our simulations based on four particular values of the ex ante equity premium (2.5% in Panel A, 3.5% in Panel B, 4.5% in Panel C, and 6% in Panel D). Each panel contains a scatter plot of two thousand points, with each point representing a pair of statistics (mean return and return standard deviation) arising in one of the simulated half-century economies. The combination based on the US realization is shown in each plot with a crosshair (a pair of solid straight lines with the intersection marked by a solid dot). The set of simulated pairs in each panel is surrounded by an ellipse which represents a 95% bivariate confidence bound, based on the asymptotic normality (or log-normality, where appropriate) of the plotted variables. (The 95% confidence ellipsoids are asymptotic approximations based on joint normality of the sample estimates of the moments of the simulated data. All of the sample moment estimates we consider are asymptotically normally distributed, as can be seen by appealing to the appropriate law of large numbers). The confidence ellipse for the 2.5% case is marked with diamonds, the 3.5% case with circles, the 4.5% case with squares, and the 6% case with circled crosses.

Notice that the sample mean for the US economy (the intersection of the crosshairs) lies loosely within cloud of points that depict the set of simulated economies for each ex ante equity premium case. That is, our simulations produce mean returns and return volatility that roughly match the US observed moments of returns, *without our having calibrated to returns*. Notice also that the intersection of the crosshairs is outside (or very nearly outside) the 95% confidence ellipse in all cases except that of the 3.5% ex ante equity premium. (In unreported results that study a finer grid of ex ante equity premium values, we found that only those simulations based on values of the ex ante equity premium between about 2.5% and 4.5% lead to 95% confidence ellipses that encompass the US economy crosshairs. As the value of the ex ante equity premium falls below 2.5% or rises above 4.5%, the confidence ellipse drifts further away from the crosshairs). Based on this set of plots, we can conclude that ex ante equity premia much less than or much greater than 3.5% are inconsistent at the 5% confidence level with the observed mean return and return volatility of S&P 500 returns. χ^2 tests presented in Donaldson, Kamstra, and Kramer [16] confirm this result.

We can easily condense the information contained in these four individual plots into one plot, as shown in Panel A of Fig. 3. The scatterplot of points representing individual simulations are omitted in the condensed plot, but the confidence ellipses themselves (and the symbols used to distinguish between them) are retained. Panel A of Fig. 3 repeats the ellipses shown in Fig. 2, so that again we see that only the 3.5% ex ante equity premium case is well within the confidence ellipse at the 5% significance level. In presenting results for additional bivariate combinations, we follow the same practice, omitting the points that represent individual simulations and using the same set of symbols to distinguish between confidence ellipses based on ex ante equity premia of 2.5%, 3.5%, 4.5%, and 6%.

In Panel B of Fig. 3 we consider the four sets of confidence ellipses for mean return and mean dividend yield combinations. Notice that as we increase the ex ante equity premium, the confidence ellipses shift upward and to the right. Notice also that with higher values of the ex ante equity premium we tend to have more variable dividend yields. That is, the confidence ellipse covers a larger range of dividend yields when the value of the ex ante equity premium is larger. The observed combination of S&P 500 mean return and mean dividend yield, represented by the intersecting crosshairs, lies within the confidence ellipse for the 2.5% and 3.5% cases, very close to the ellipse for the 4.5% case, and far outside the ellipse for the 6% case.

Panel C of Fig. 3 plots confidence ellipses for mean interest rates versus mean ex post equity premia. The intersection of the crosshairs is within all four of the shown confidence ellipses. As we calibrate our model to the US interest rate, and as the ex post equity premium has a large

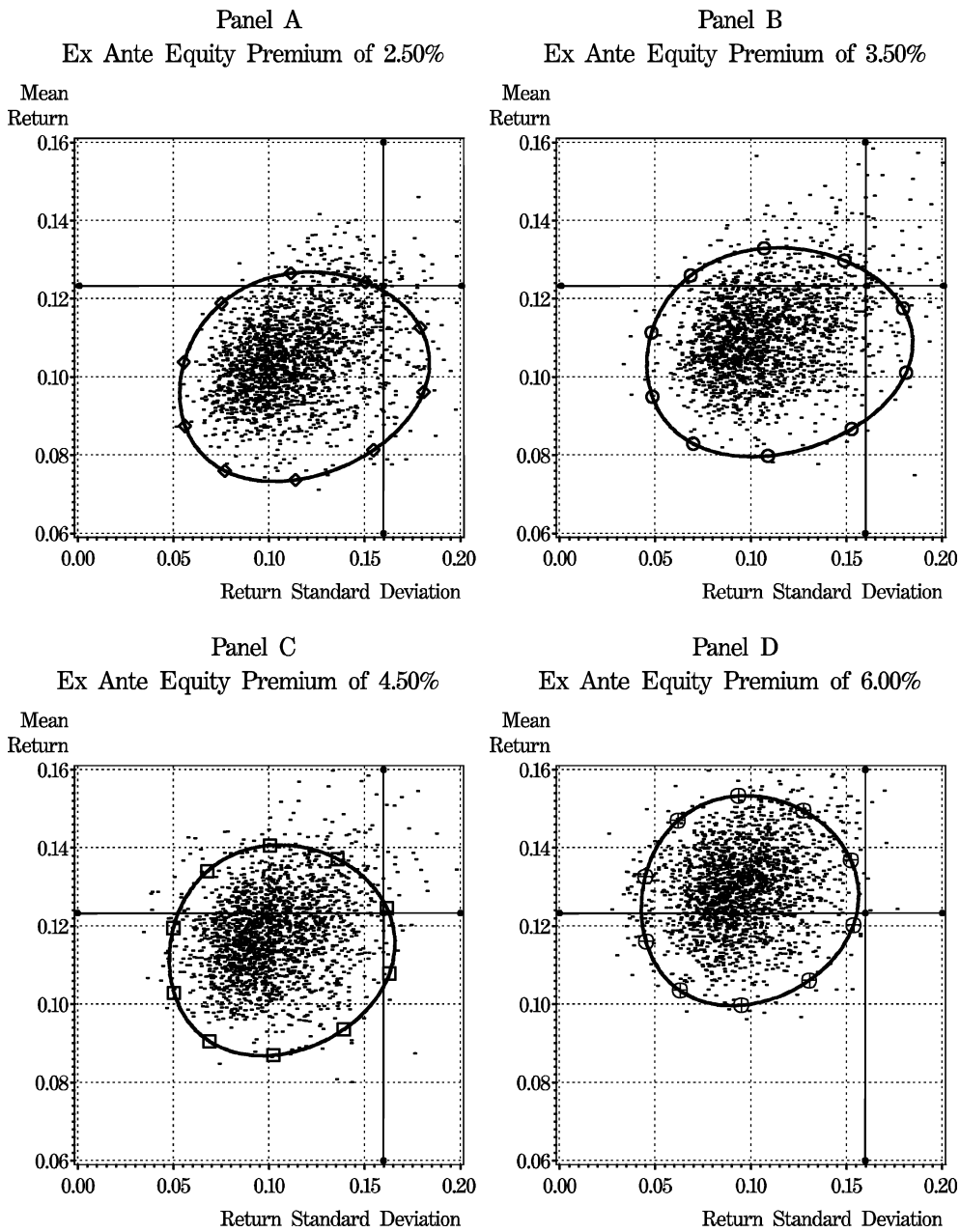

Financial Economics, Time Variation in the Market Return, Figure 2

Bivariate scatterplots of simulated data for a model allowing for trends and structural breaks. The model upon which these scatterplots are based allows for trends and structural breaks in the equity premium process, as well as autocorrelated and co-varying dividend growth rates, interest rates, and equity premia. Observed market data are indicated with crosshairs, and confidence ellipses are marked as follows. Ex ante equity premium of 2.5%: ⋄, Ex ante equity premium of 3.5%: ∘, Ex ante equity premium of 4.5%: □, Ex ante equity premium of 6%: ⊕

variance, it is not surprising that the US experience is consistent with the simulated data from the entire range of ex ante equity premia considered here. This result is merely telling us that the ex post equity premium is not, by itself, particularly helpful in narrowing the possible range for the ex ante equity premium (consistent with the empirical imprecision in measuring the ex post equity premium which has been extensively documented in the literature). Notice as well that the confidence ellipses in Panel C are all negatively sloped: we see high mean interest rates with low eq-

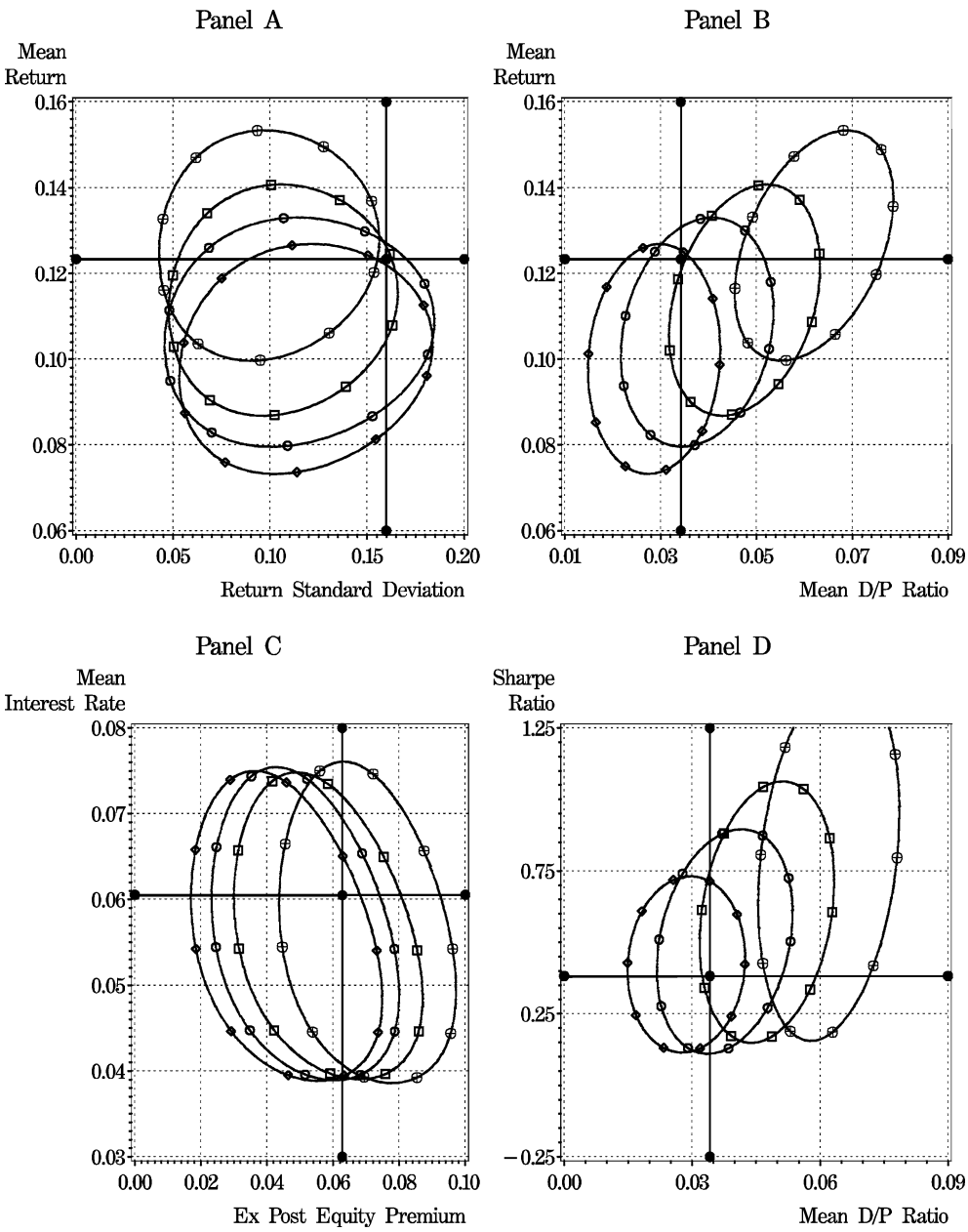

Financial Economics, Time Variation in the Market Return, Figure 3

Confidence ellipses based on simulated data for a model allowing for trends and structural breaks. The model upon which these scatterplots are based allows for trends and structural breaks in the equity premium process, as well as autocorrelated and co-varying dividend growth rates, interest rates, and equity premia. Observed market data are indicated with crosshairs, and confidence ellipses are marked as follows. 2.5% ex post equity premium: ⋄, 3.5% ex post equity premium: ∘, 4.5% ex post equity premium: □, 6% ex post equity premium: ⊕

uity premia and low mean interest rates with high equity premia. Many researchers, including Weil [74], have commented that the flip side of the high equity premium puzzle is the low risk-free rate puzzle. Here we confirm that the dual puzzle arises in our simulated economies as well. It appears that this puzzle is a mechanical artifact coming out of the calculation of the premium. As the ex post equity premium equals the mean return minus the mean interest rate, a decrease in the interest rate, all else held constant, must lead to a higher ex post equity equity premium.

Panel D of Fig. 3 contains the confidence ellipses for the Sharpe ratio (or reward-to-risk ratio, calculated as the average annual difference between the arithmetic return and the risk-free rate divided by the standard deviation of the annual differences) and the mean dividend yield. As the ex ante equity premium is increased from 2.5%, the confidence ellipses shift from being centered on the crosshairs to far to the right of the crosshairs. The US expe-

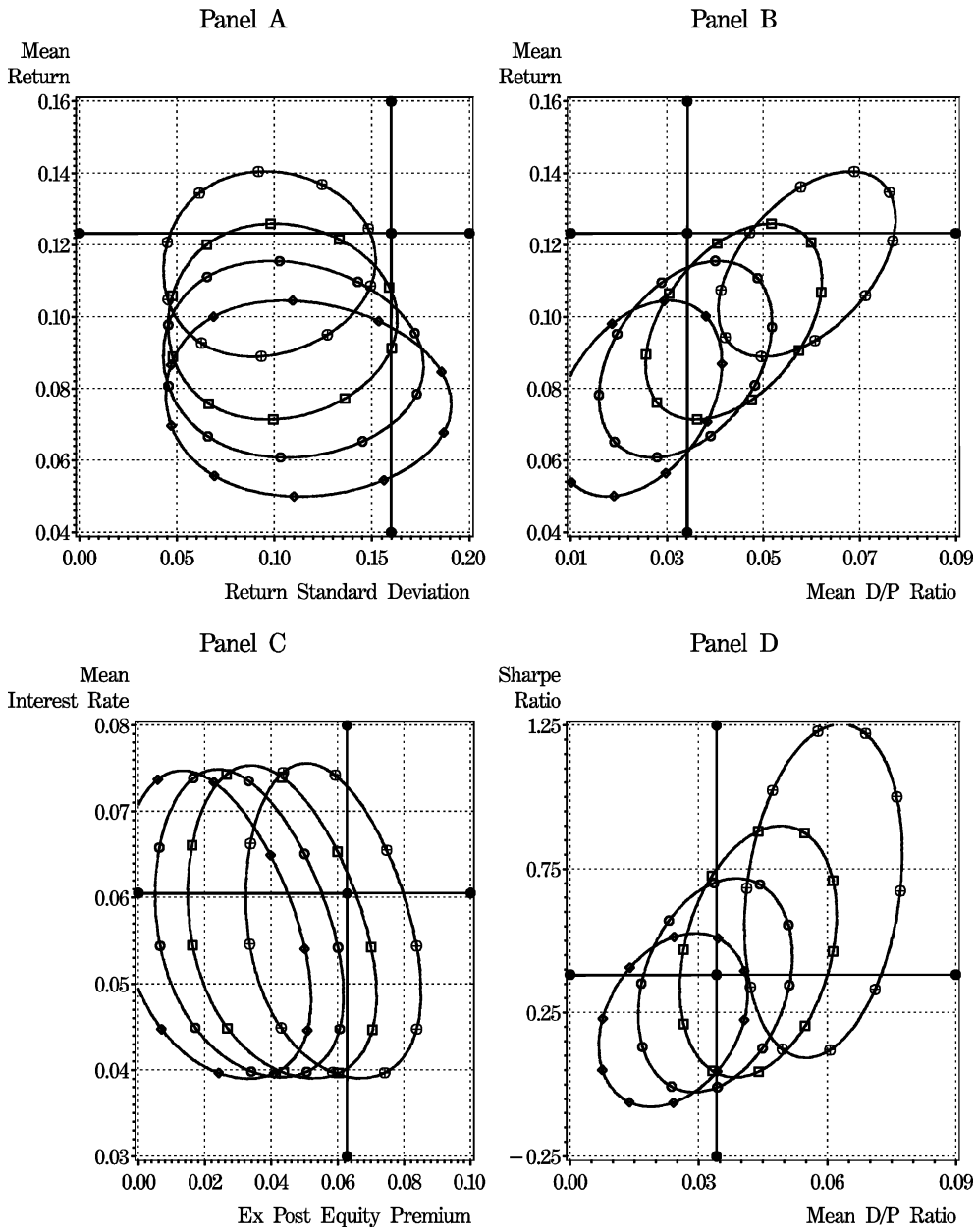

Financial Economics, Time Variation in the Market Return, Figure 4

Confidence ellipses based on simulated data for a restricted model that does not allow for trends and structural breaks. The model upon which these scatterplots are based does not allow for trends or structural breaks in the equity premium process, but does allow for autocorrelated and co-varying dividend growth rates, interest rates, and equity premia. Observed market data are indicated with crosshairs, and confidence ellipses are marked as follows. 2.5% ex post equity premium: ⋄, 3.5% ex post equity premium: ∘, 4.5% ex post equity premium: □, 6% ex post equity premium: ⊕

rience, indicated by the crosshairs at a Sharpe ratio of approximately 0.4 and a mean dividend yield of about 3.5%, is well outside the 95% confidence ellipse for the 6% ex ante equity premium case, suggesting a 6% ex ante equity premium is inconsistent with the jointly observed S&P 500 Sharpe ratio and mean dividend yield. Indeed Fama and French [21] and Jagannathan, McGrattan, and Scherbina [35] make reference to dividend yields to argue that the equity premium may be much smaller than 6%; our analysis gives us a glimpse of just how much smaller it might be.

Overall in Fig. 3, the joint realization of key characteristics of the US market data suggests that the true ex ante equity premium is no lower than 2.5%, no higher than 4.5%, and is most likely near 3.5%. Multivariate χ^2 tests performed by Donaldson, Kamstra, and Kramer [16] indicate a 95% confidence interval of plus-or-minus 50 basis points around 3.5%.

Consider now Fig. 4, which details simulated data from a restricted model that has a time-varying equity premium but no trends or structural breaks. Donaldson, Kamstra, and Kramer [16] study this simplified model and find that it performs poorly relative to the model we consider in Figs. 2 and 3 in terms of its ability to capture the behavior of US market data. Figure 4 shows that with the restricted model, no values of the ex ante equity premium are consistent with the observed US mean return, standard deviation, and dividend yield. That is, the simulation-based mean return and dividend yield ellipses do not contain the US data crosshairs for any value of the ex ante equity premium considered. (χ^2 tests presented in Donaldson, Kamstra, and Kramer [16] strongly support this conclusion). The implication is that it is essential to model trends and structural breaks in the equity premium process in order to accurately capture the dynamics of observed US data. Donaldson, Kamstra, and Kramer show that model failure becomes even more stark if the equity premium is constrained to be constant.

Overall, the evidence in Figs. 3 and 4 does not itself resolve the equity premium puzzle, but evidence in Fig. 3 (based on the model that allows for trends and structural breaks in the equity premium process) does provide a narrow target range of plausible equity premia that economic models should be able to explain. Additionally, the evidence in Figs. 3 and 4 points to a secondary issue ignored in the literature prior to the work of Donaldson, Kamstra, and Kramer [16], that it is crucial to model the equity premium as both time-varying and as having trends and structural breaks. We saw in Fig. 4 that high return volatility, high ex post equity premia, and low dividend yields cannot be explained easily by constant equity premium models. This result has clear implications for valuation: simple techniques that restrict the discount rate to a constant are remarkably inconsistent with the US experience of time-varying equity premia, and serious attention should be paid to modeling a time-varying rate for use in discounting future expected cash flows.

Time-Varying Equity Premia: Possible Biological Origins

To the extent that the simulation techniques considered in the previous section suggest that the equity premium varies over time, it is interesting to consider some empirical evidence of time-varying equity premia. We first survey some examples of high-frequency variations in the equity premium, and then we explore in detail two examples which may arise due to reasons that relate to human biology and/or psychology.

There is a wide range of evidence of high-frequency movement in the equity premium. At the highest frequency, we observe roughly 'U-shaped' intra-day returns (see [29,36,77]), with returns being perhaps somewhat higher during the morning trading period than in the afternoon (see [46]). At the weekly frequency, returns from Friday's close until Monday's close are low and even negative on average, as first identified by Cross [12]. Rogalski [66] found prices rose during Mondays, thus identifying the negative average realizations that followed Fridays as a weekend effect and not a Monday effect. Turning to the monthly domain, Ogden [56] documented a turn of the month effect where returns in the first half of the month are higher than returns in the second half of the month. At the annual frequency, there is the well-known turn-of-the-year effect, first shown by Rozeff and Kinney [68]. Keim [45] showed that half of the year's excess returns for small firms arose in January, and half of the January returns took place in the first five days of the month. Further, Reinganum [64] showed that January returns are higher for small firms whose price performed poorly in the previous year. All of this is consistent with the tax-loss-selling hypothesis whereby investors realize losses at the end of the tax year, leading to higher returns in January after the tax-loss selling ends.

Next we turn our attention to two cases of time-varying equity premia that may arise for reasons related to human physiology. One is Seasonal Affective Disorder (SAD), and another is daylight saving time changes.

Seasonal Affective Disorder

Past research suggests there are seasonal patterns in the equity premium which may arise due to cyclical changes in

the risk tolerance of individual investors over the course of the year related to SAD. The medical condition of SAD, according to Rosenthal [67], is a recurrent depression associated with diminished daylight in the fall, affecting many millions of Americans, as well as peoples from around the world, even those located near the equator. (In a study of 303 patients attending a primary care facility in Vancouver, Schlager, Froom, and Jaffe [70] found that 9% were clinically diagnosed with SAD and another 29% had significant winter depressive symptoms without meeting conditions for major depression. Other studies have found similar magnitudes, though some research has found that prevalence varies by latitude, with more extreme latitudes having a larger proportion of SAD-sufferers.) SAD is classified as a major depressive disorder. The symptoms of SAD include anxiety, periods of sadness, chronic fatigue, difficulty concentrating, lethargy, sleep disturbance, sugar and carbohydrate craving and associated weight gain, loss of interest in sex, and of course, clinical depression. Psychologists have shown that depressed people have less tolerance for risk in general. (See [7,32,82,83]). Psychologists refer to risk tolerance in terms of "sensation seeking" tendency, measured using a scale developed by Zuckerman [80], [81]. Those who tolerate (or seek) high levels of risk tend to score high on the sensation-seeking scale. Differences in sensation-seeking tendencies have been linked to gender (see [5] for example), race (see [31] for instance), age (see, for example, [84]), and other personal characteristics.

Economists and psychologists working together have shown that sensation-seeking tendency translates into tolerance for risk of a specifically financial or economic nature. For instance, Wong and Carducci [76] find that individuals who score low on tests of sensation seeking display greater risk aversion in making financial decisions, including the decision to purchase stocks, bonds, and insurance. Harlow and Brown [28] document the link between sensation seeking and financial risk tolerance by building on results from psychiatry which show that high blood levels of a particular enzyme are associated with depression and a lack of sensation seeking while low levels of the enzyme are associated with a high degree of sensation seeking. Harlow and Brown write, "Individuals with neurochemical activity characterized by lower levels of [the enzyme] and with a higher degree of sensation-seeking are *more willing to accept economic risk* ... Conversely, high levels of this enzyme and a low level of sensation seeking appear to be associated with risk-averse behavior." (pp. 50–51, emphasis added). These findings suggest an individual's level of sensation seeking is indicative of his or her tolerance for financial risk.

Given these relationships, Kamstra, Kramer, and Levi [42] conjecture that during the fall and winter seasons, when a fraction of the population suffers from SAD, the proportion of risk-averse investors rises. Risk-averse investors shun risky stocks in the fall, they argue, which has a negative influence on stock prices and returns. As winter progresses and daylight becomes more plentiful, investors start to recover from their depression and become more willing to hold risky assets, at which time stock prices and returns should be positively influenced.

If the extent or severity of SAD is greater at more extreme latitudes, then the SAD effect on stock returns should be greater in stock markets at high latitudes and less in markets close to the equator. Also, the pattern of returns in the Southern Hemisphere should be the opposite of that in the Northern Hemisphere as are the seasons. Thus, Kamstra, Kramer and Levi [42] study stock market indices for the US, Sweden, Britain, Germany, Canada, New Zealand, Japan, Australia, and South Africa. They regress each country's daily stock returns on a variety of standard control variables plus two variables intended to capture the impact of SAD on returns. The first of these two variables, SAD_t, is a simple function of the length of night at the latitude of the respective market for the fall and winter months for which SAD has been documented to be most severe. The second of these variables, a fall dummy variable denoted Fall_t, is included because the SAD hypothesis implies the expected effect on returns is different before versus after winter solstice. Specifically, when agents initially become more risk averse, they should shun risky assets which should cause prices to be lower than would otherwise be observed, and when agents revert to normal as daylight becomes more plentiful, prices should rebound. The result should be lower returns in the autumn, higher returns in the winter, and thus a high equity premium for investors who hold through the autumn and winter periods. The Fall_t dummy variable is used to capture the lower autumn returns. Both SAD_t and Fall_t are appropriately defined for the Southern Hemisphere countries, accounting for the six month difference in seasons relative to the Northern Hemisphere markets.

Table 1 summarizes the average annual effect due to each of the SAD variables, SAD_t and Fall_t, for each of the international indices Kamstra, Kramer, and Levi [42] study. For comparison, the unconditional average annual return for each index is also provided. Observe that the annualized return due to SAD_t is positive in every country, varying from 5.7 to 17.5 percent. The SAD effect is generally larger the further are the markets from the equator. The negative annualized returns due to Fall_t demonstrate the fact that SAD typically causes returns to be

Financial Economics, Time Variation in the Market Return, Table 1
Average annual percentage return due to SAD variables

Country (Index)	Annual return due to SAD_t	Annual return due to $fall_t$	Unconditional annual return
US (S&P 500)	9.2***	−3.6**	6.3***
Sweden (Veckans Affärar)	13.5**	−6.9**	17.1***
Britain (FTSE 100)	10.3**	−2.3	9.6***
Germany (DAX 30)	8.2*	−4.3**	6.5**
Canada (TSX 300)	13.2***	−4.3**	6.1***
New Zealand (Capital 40)	10.5**	−6.6**	3.3
Japan (NIKKEI 225)	6.9*	−3.7**	9.7***
Australia (All ordinaries)	5.7	0.5	8.8***
South Africa (Datastream global index)	17.5*	−2.1	14.6***

One, two, and three asterisks denote significantly different from zero at the ten, five, and one percent level respectively, based on one-sided tests. Source: Table 3 in [42].

shifted from the fall to the winter. Garrett, Kamstra, and Kramer [24] study seasonally-varying risk aversion in the context of an equilibrium asset pricing model, allowing the price of risk to vary with length of night through the fall and winter seasons. They find the risk premium on equity varies through the seasons in a manner consistent with investors being more risk averse due to SAD in the fall and winter.

Kamstra, Kramer, and Levi [43] show that there is an opposite seasonal pattern in Treasury returns relative to stock returns, consistent with time-varying risk aversion being the underlying force behind the seasonal pattern previously shown to exist in stock returns. If SAD-affected investors are shunning risky stocks in the fall as they become more risk averse, then they should be favoring safe assets at that time, which should lead to an opposite pattern in Treasury returns relative to stock returns. The seasonal cycle in the Treasury market is striking, with a variation of more than 80 basis points between the highest and lowest average monthly returns. The highest Treasury returns are observed when equity returns are lowest, and *vice versa*, which is a previously unknown pattern in Treasury returns.

Kamstra, Kramer, and Levi [43] define a new measure which is linked directly to the clinical incidence of SAD. The new measure uses data on the weekly or monthly onset of and recovery from SAD, obtained from studies of SAD patients in Vancouver and Chicago conducted by medical researchers. Young, Meaden, Fogg, Cherin, and Eastman [79] and Lam [47] document the clinical *onset* of SAD symptoms and *recovery* from SAD symptoms among North Americans known to be affected by SAD. Young et al. study 190 SAD-sufferers in Chicago and find that 74 percent of them are first diagnosed with SAD in the weeks between mid-September and early November. Lam studies 454 SAD patients in Vancouver on a monthly basis and finds, that the peak timing of diagnosis is during the early fall. Lam [47] also studies the timing of clinical remission of SAD and finds it peaks in April, with almost half of all SAD-sufferers first experiencing complete remission in that month. March is the second most common month for subjects to first experience full remission, corresponding to almost 30 percent of subjects. For most SAD patients, the initial onset and full recovery are separated by several months over the fall and winter.

Direct use of Kamstra, Kramer, and Levi's [43] variable (which is an estimate of population-wide SAD onset/recovery based on specific samples of individuals) could impart an error-in-variables problem (see [48]), thus they utilize an instrumented version detailed in the paper, which they call Onset/Recovery, denoted $\hat{OR}_t$. The instrumented SAD measure $\hat{OR}_t$ reflects the change in the proportion of SAD-affected individuals actively suffering from SAD. The measure is defined year-round (unlike the original Kamstra, Kramer, and Levi [42], SAD_t variable, which is defined for only the fall and winter months), taking on positive values in the summer and fall and negative values in the winter and spring. Its value peaks near the fall equinox and reaches a trough near the spring equinox. (The exact monthly values of $\hat{OR}_t$ are reported by Kamstra, Kramer, and Levi [43].) The opposite signs on $\hat{OR}_t$ across the fall and winter seasons should, in principle, permit it to capture the opposite impact on equity or Treasury returns across the seasons, without use of a dummy variable. Kamstra, Kramer, and Levi [43] find that use of $\hat{OR}_t$ as a regressor to explain seasonal patterns in Treasury and equity returns renders the SAD_t and $Fall_t$ (used by Kamstra, Kramer, and Levi [42]) as economically and statisti-

cally insignificant, suggesting the Onset/Recovery variable does a far better job of explaining seasonal variation in returns than the original proxies which are not directly related to the incidence of SAD.

Kamstra, Kramer, and Levi [43] show that the seasonal Treasury and equity return patterns are unlikely to arise from macroeconomic seasonalities, seasonal variation in risk, cross-hedging between equity and Treasury markets, investor sentiment, seasonalities in the Treasury market auction schedule, seasonalities in the Treasury debt supply, seasonalities in the Federal Reserve Board's interest-rate-setting cycle, or peculiarities of the sample period considered. They find that the seasonal cycles in equity and Treasury returns become more pronounced during periods of high market volatility, consistent with time-varying risk aversion among market participants. Furthermore, they apply the White [75] reality test and find that the correlation between returns and the clinical incidence of seasonal depression cannot be easily dismissed as the simple result of data snooping.

DeGennaro, Kamstra, and Kramer [13] and Kamstra, Kramer, and Levi [13] provide further corroborating evidence for the hypothesis that SAD leads to time variation in financial markets by considering (respectively) bid-ask spreads for stocks and the flow of funds in and out of risky and safe mutual funds. In both papers they find strong support for the link between seasonal depression and time-varying risk aversion.

Daylight Saving Time Changes

The second potential biological source of time-varying equity premia we consider arises on the two dates of the year when most of the developed world shifts clocks forward or backward an hour in the name of daylight saving. Psychologists have found that changes in sleep patterns (due to shift work, jet lag, or daylight saving time changes, for example) are associated with increased anxiety, which is suggestive of a link between changes in sleep habits and time-varying risk tolerance. See [26,52], and citations found in [10] and [72] for more details on the link between sleep disruptions and anxiety. In addition to causing heightened anxiety, changes in sleep patterns also inhibit rational decision-making, lower one's information-processing ability, affect judgment, slow reaction time, and reduce problem-solving capabilities. Even a change of one hour can significantly affect behavior.

Kamstra, Kramer, and Levi [40] explore the financial market ramifications of a link between daylight saving time-change-induced disruptions in sleep patterns and individuals' tolerance for risk. They find, consistent with psychology studies that show a gain or loss of an hour's sleep leads to increased anxiety, investors seem to shun risky stock on the trading day following a daylight saving time change. They consider stock market indexes from four countries where the time changes happen on non-overlapping dates, the US, Canada, Britain, and Germany. Based on stock market behavior over the past three decades, the authors find that the magnitude of the average return on spring daylight saving weekends is typically between two to five times that of ordinary weekends, and the effect is even stronger in the fall. Kamstra, Kramer, and Levi [41] show that the effect is not driven by a few extremely negative observations, but rather the entire distribution of returns shifts to the left following daylight saving time changes, consistent with anxious investors selling risky stock.

Future Directions

We divide our discussion in this section into three parts, one for each major topic discussed in the article.

Regarding fundamental valuation, a promising future path is to compare estimates emerging from sophisticated valuation methods to market prices, using the comparison to highlight inconsistencies in the modeling assumptions (such as restrictions on the equity premium used by the model, restrictions on the growth rate imposed for expected cash flows, and the implied values of those quantities that can be inferred from market prices). Even if one believes that markets are efficient and investors are rational, there is still much to be learned from calculating fundamentals using models and examining discrepancies relative to observed market prices.

Regarding the simulation techniques for estimating the equity premium, a promising direction for future research is to exploit these tools to forecast the volatility of stock prices. This may lead to new alternatives to existing option-implied volatility calculations and time-series techniques such as ARCH (for an overview of these methods see [15]). Another fruitful future direction would be to apply the simulation techniques to the valuation of individual companies' stock (as opposed to valuing, say, stock market indexes).

Regarding the topic of time-varying equity premia that may arise for biological reasons, a common feature of both of the examples explored in Sect. "Time-Varying Equity Premia: Possible Biological Origins", SAD and daylight-saving-time-change-induced fluctuations in the risk premium, is that in both cases the empirical evidence is based on aggregate financial market data. There is a recent trend in finance toward documenting phenomena at

the individual level, using data such as individuals' financial asset holdings and trades in their brokerage accounts. (See [1,54,55] for instance). A natural course forward is to build upon the existing aggregate market support for the prevalence of time-varying risk aversion by testing at the individual level whether risk aversion varies through the course of the year due to seasonal depression and during shorter intervals due to changes in sleep patterns. An additional potentially fruitful direction for future research is to integrate into classical asset pricing models the notion that biological factors might impact asset returns through changes in agents' degree of risk aversion. That is, human traits such as seasonal depression may lead to regularities in financial markets that are not mere anomalies; rather they may be perfectly consistent with rational agents making sensible decisions given their changing tolerance for risk. This new line of research would be similar in spirit to the work of Shefrin [71] who considers the way behavioral biases like overconfidence can be incorporated into the pricing kernel in standard asset pricing models. While the behavioral biases Shefrin considers typically involve humans making errors, the biological factors described here might be considered rational due to their involvement of time-varying risk aversion.

Bibliography

Primary Literature

1. Barber B, Odean T (2001) Boys will be boys: Gender, overconfidence, and common stock investment. Q J Econ 116: 261–292
2. Barsky RB, DeLong JB (1993) Why does the stock market fluctuate? Q J Econ 108:291–311
3. Bakshi G, Chen Z (2005) Stock valuation in dynamic economies. J Financ Market 8:115–151
4. Brealey RA, Myers SC, Allen F (2006) Principles of corporate finance, 8th edn. McGraw-Hill Irwin, New York
5. Byrnes JP, Miller DC, Schafer WD (1999) Gender differences in risk taking: A meta-analysis. Psychol Bull 125:367–383
6. Campbell JY, Kyle AS (1993) Smart money, noise trading and stock price behavior. Rev Econ Stud 60:1–34
7. Carton S, Jouvent R, Bungener C, Widlöcher D (1992) Sensation seeking and depressive mood. Pers Individ Differ 13: 843–849
8. Chiang R, Davidson I, Okuney J (1997) Some theoretical and empirical implications regarding the relationship between earnings, dividends and stock prices. J Bank Financ 21:17–35
9. Cochrane JH (2001) Asset pricing. Princeton University Press, Princeton
10. Coren S (1996) Sleep Thieves. Free Press, New York
11. Corradi V, Swanson NR (2005) Bootstrap specification tests for diffusion processes. J Econom 124:117–148
12. Cross F (1973) The behavior of stock prices on Fridays and Mondays. Financ Anal J 29:67–69
13. DeGennaro R, Kamstra MJ, Kramer LK (2005) Seasonal variation in bid-ask spreads. University of Toronto (unpublished manuscript)
14. Donaldson RG, Kamstra MJ (1996) A new dividend forecasting procedure that rejects bubbles in asset prices. Rev Financ Stud 9:333–383
15. Donaldson RG, Kamstra MJ (2005) Volatility forecasts, trading volume, and the ARCH versus option-implied volatility tradeoff. J Financ Res 28:519–538
16. Donaldson RG, Kamstra MJ, Kramer LA (2007) Estimating the ex ante equity premium. University of Toronto Manuscript
17. Dong M, Hirshleifer DA (2005) A generalized earnings-based stock valuation model. Manchester School 73:1–31
18. Duffie D, Singleton KJ (1993) Simulated moments estimation of Markov models of asset prices. Econometrica 61:929–952
19. Engle RF (1982) Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50:987–1007
20. Fama EF, French KR (1993) Common risk factors in the returns on stocks and bonds. J Financ Econ 33:3–56
21. Fama EF, French KR (2002) The equity premium. J Financ 57:637–659
22. Farrell JL (1985) The dividend discount model: A primer. Financ Anal J 41:16–25
23. Feltham GA, Ohlson JA (1995) Valuation and clean surplus accounting for operating and financial activities. Contemp Account Res 11:689–731
24. Garrett I, Kamstra MJ, Kramer LK (2005) Winter blues and time variation in the price of risk. J Empir Financ 12:291–316
25. Gordon M (1962) The Investment, Financing and Valuation of the Corporation. Irwin, Homewood
26. Gordon NP, Cleary PD, Parker CE, Czeisler CA (1986) The prevalence and health impact of shiftwork. Am J Public Health 76:1225–1228
27. Hackel KS, Livnat J (1996) Cash flow and security analysis. Irwin, Chicago
28. Harlow WV, Brown KC (1990) Understanding and assessing financial risk tolerance: A biological perspective. Financ Anal J 6:50–80
29. Harris L (1986) A transaction data study of weekly and intradaily patterns in stock returns. J Financ Econ 16:99–117
30. Hawkins DF (1977) Toward an old theory of equity valuation. Financ Anal J 33:48–53
31. Hersch J (1996) Smoking, seat belts and other risky consumer decisions: differences by gender and race. Managerial Decis Econ 17:471–481
32. Horvath P, Zuckerman M (1993) Sensation seeking, risk appraisal, and risky behavior. Personal Individ Diff 14:41–52
33. Hurley WJ, Johnson LD (1994) A realistic dividend valuation model. Financ Anal J 50:50–54
34. Hurley WJ, Johnson LD (1998) Generalized Markov dividend discount models. J Portf Manag 24:27–31
35. Jagannathan R, McGrattan ER, Scherbina A (2000) The declining US equity premium. Fed Reserve Bank Minneap Q Rev 24:3–19
36. Jain PC, Joh G (1988) The dependence between hourly prices and trading volume. J Financ Quant Analysis 23:269–283
37. Jegadeesh N, Titman S (1993) Returns to buying winners and selling losers: Implications for stock market efficiency. J Financ 48:65–91

38. Jorion P, Goetzmann WN (1999) Global stock markets in the twentieth century. J Financ 54:953–980
39. Kamstra MJ (2003) Pricing firms on the basis of fundamentals. Fed Reserve Bank Atlanta Econ Rev First Quarter:49–70
40. Kamstra MJ, Kramer LA, Levi MD (2000) Losing sleep at the market: the daylight saving anomaly. Am Econ Rev 90: 1005–1011
41. Kamstra MJ, Kramer LA, Levi MD (2002) Losing sleep at the market: The daylight saving anomaly: Reply. Am Econ Rev 92:1257–1263
42. Kamstra MJ, Kramer LA, Levi MD (2003) Winter blues: A SAD stock market cycle. Am Econ Rev 93:324–343
43. Kamstra MJ, Kramer LA, Levi MD (2007) Opposing Seasonalities in Treasury versus Equity Returns. University of Toronto Manuscript
44. Kamstra MJ, Kramer LA, Levi MD, Wermers R (2008) Seasonal asset allocation: evidence from mutual fund flows. University of Toronto Manuscript
45. Keim DB (1983) Size-related anomalies and stock return seasonality: Further Empirical Evidence. J Financ Econ 12:13–32
46. Kramer LA (2002) Intraday stock returns, time-varying risk premia, and diurnal mood variation. University of Toronto Manuscript
47. Lam RW (1998) Seasonal Affective Disorder: Diagnosis and management. Prim Care Psychiatry 4:63–74
48. Levi MD (1973) Errors in the variables bias in the presence of correctly measured variables. Econometrica 41:985–986
49. Liu W (2006) A liquidity-augmented capital asset pricing model. J Financ Econ 82:631–671
50. McFadden D (1989) A method of simulated moments for estimation of discrete response models without numerical integration. Econometrica 47:995–1026
51. Mehra R, Prescott EC (1985) The equity premium: A puzzle. J Monet Econ 15:145–161
52. Mellinger GD, Balter MB, Uhlenhuth EH (1985) Insomnia and its treatment: prevalence and correlates. Arch Gen Psychiatry 42:225–232
53. Michaud RO, Davis PL (1982) Valuation model bias and the scale structure of dividend discount returns. J Financ 37: 563–576
54. Odean T (1998) Are investors reluctant to realize their losses? J Financ 53:1775–1798
55. Odean T (1999) Do investors trade too much? Am Econ Rev 89:1279–1298
56. Ogden JP (1990) Turn-of-month evaluations of liquid profits and stock returns: A common explanation for the monthly and January effects. J Financ 45:1259–1272
57. Ohlson JA (1995) Earnings, book values, and dividends in equity valuation. Contemp Acc Res 11:661–687
58. Pakes A, Pollard D (1989) Simulation and the asymptotics of optimization estimators. Econometrica 57:1027–1057
59. Pástor L', Stambaugh R (2001) The equity premium and structural breaks. J Financ 56:1207–1239
60. Penman SH (1998) Combining earnings and book value in equity valuation. Contemp Acc Res 15:291–324
61. Penman SH, Sougiannis T (1998) A comparison of dividend, cash flow and earnings approaches to equity valuation. Contemp Acc Res 15:343–383
62. Peters DJ (1991) Using PE/Growth ratios to develop a contrarian approach to growth stocks. J Portf Manag 17:49–51
63. Rappaport A (1986) The affordable dividend approach to equity valuation. Financ Anal J 42:52–58
64. Reinganum MR (1983) The anomalous stock market behavior of small firms in January. J Financ Econ 12:89–104
65. Rietz TA (1988) The equity risk premium: A solution. J Monet Econ 22:117–31
66. Rogalski RJ (1984) New findings regarding day-of-the-week returns: A note. J Financ 35:1603–1614
67. Rosenthal NE (1998) Winter Blues: Seasonal Affective Disorder: What is It and How to Overcome It, 2nd edn. Guilford Press, New York
68. Rozeff MS, Kinney WR (1976) Capital market seasonality: The case of stock returns. J Financ Econ 3:379–402
69. Rubinstein M (1976) The valuation of uncertain income streams and the pricing of options. Bell J Econ 7:407–425
70. Schlager D, Froom J, Jaffe A (1995) Winter depression and functional impairment among ambulatory primary care patients. Compr Psychiatry 36:18–24
71. Shefrin H (2005) A Behavioral Approach to Asset Pricing. Academic Press, Oxford
72. Spira AP, Friedman L, Flint A, Sheikh J (2005) Interaction of sleep disturbances and anxiety in later life: perspectives and recommendations for future research. J Geriatr Psychiatry Neurol 18:109–115
73. Sorensen EH, Williamson DA (1985) Some evidence on the value of dividend discount models. Financ Anal J 41:60–69
74. Weil P (1989) The equity premium puzzle and the risk-free rate puzzle. J Monet Econ 24:401–421
75. White H (2000) A reality check for data snooping. Econometrica 68:1097–1126
76. Wong A, Carducci B (1991) Sensation seeking and financial risk taking in everyday money matters. J Bus Psychol 5: 525–530
77. Wood RA, McInish TH, Ord JK (1985) An investigation of transactions data for NYSE stocks. J Financ 40:723–741
78. Yao Y (1997) A trinomial dividend valuation model. J Portf Manag 23:99–103
79. Young MA, Meaden PM, Fogg LF, Cherin EA, Eastman CI (1997) Which environmental variables are related to the onset of Seasonal Affective Disorder? J Abnorm Psychol 106: 554–562
80. Zuckerman M (1976) Sensation seeking and anxiety, traits and states, as determinants of behavior in novel situations. In: Sarason IG, Spielberger CD (eds) Stress and Anxiety, vol 3. Hemisphere, Washington DC
81. Zuckerman M (1983) Biological Bases of Sensation Seeking, Impulsivity and Anxiety. Lawrence Erlbaum Associates, Hillsdale
82. Zuckerman M (1984) Sensation seeking: A comparative approach to a human trait. Behav Brain Sci 7:413–471
83. Zuckerman M, Buchsbaum MS, Murphy DL (1980) Sensation seeking and its biological correlates. Psychol Bull 88: 187–214
84. Zuckerman M, Eysenck S, Eysenck HJ (1978) Sensation seeking in England and America: Cross-cultural, age, and sex comparisons. J Consult Clin Psychol 46:139–149

Books and Reviews

Dimson E (1988) Stock Market Anomalies. Cambridge University Press, Cambridge

Kocherlakota NR (1996) The equity premium: It's still a puzzle. J Econ Lit 34:42–71
Mehra R (2003) The equity premium: Why is it a puzzle? Financ Anal J 59:54–69
Mehra R, Prescott EC (2003) The equity premium in retrospect. In: Constantinides GM, Harris M, Stulz RM (eds) Handbook of the Economics of Finance: Financial Markets and Asset Pricing, vol 1B. North Holland, Amsterdam, pp 889–938
Penman S (2003) Financial Statement Analysis and Security Valuation, 2nd edn. McGraw-Hill/Irwin, New York
Siegel JJ, Thaler RH (1997) Anomalies: The equity premium puzzle. J Econ Perspect 11:191–200
Thaler RH (2003) The Winner's Curse: Paradoxes and Anomalies of Economic Life. Princeton University Press, Princeton

Financial Forecasting, Non-linear Time Series in

Gloria González-Rivera, Tae-Hwy Lee
Department of Economics, University of California, Riverside, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Nonlinear Forecasting Models for the Conditional Mean
Nonlinear Forecasting Models for the Conditional Variance
Forecasting Beyond Mean and Variance
Evaluation of Nonlinear Forecasts
Conclusions
Future Directions
Bibliography

Glossary

Arbitrage pricing theory (APT) the expected return of an asset is a linear function of a set of factors.

Artificial neural network is a nonlinear flexible functional form, connecting inputs to outputs, being capable of approximating a measurable function to any desired level of accuracy provided that sufficient complexity (in terms of number of hidden units) is permitted.

Autoregressive conditional heteroskedasticity (ARCH) the variance of an asset returns is a linear function of the past squared surprises to the asset.

Bagging short for *b*ootstrap *agg*regat*ing*. Bagging is a method of smoothing the predictors' instability by averaging the predictors over bootstrap predictors and thus lowering the sensitivity of the predictors to training samples. A predictor is said to be unstable if perturbing the training sample can cause significant changes in the predictor.

Capital asset pricing model (CAPM) the expected return of an asset is a linear function of the covariance of the asset return with the return of the market portfolio.

Factor model a linear factor model summarizes the dimension of a large system of variables by a set of factors that are linear combinations of the original variables.

Financial forecasting prediction of prices, returns, direction, density or any other characteristic of financial assets such as stocks, bonds, options, interest rates, exchange rates, etc.

Functional coefficient model a model with time-varying and state-dependent coefficients. The number of states can be infinite.

Linearity in mean the process $\{y_t\}$ is linear in mean conditional on X_t if

$$\Pr\left[\mathbb{E}(y_t|X_t) = X_t'\theta^*\right] = 1 \quad \text{for some } \theta^* \in \mathbb{R}^k.$$

Loss (cost) function When a forecast $f_{t,h}$ of a variable Y_{t+h} is made at time t for h periods ahead, the loss (or cost) will arise if a forecast turns out to be different from the actual value. The loss function of the forecast error $e_{t+h} = Y_{t+h} - f_{t,h}$ is denoted as $c_{t+h}(Y_{t+h}, f_{t,h})$, and the function $c_{t+h}(\cdot)$ can change over t and the forecast horizon h.

Markov-switching model features parameters changing in different regimes, but in contrast with the threshold models the change is dictated by a non-observable state variable that is modelled as a hidden Markov chain.

Martingale property tomorrow's asset price is expected to be equal to today's price given some information set

$$\mathbb{E}(p_{t+1}|\mathcal{F}_t) = p_t\ .$$

Nonparametric regression is a data driven technique where a conditional moment of a random variable is specified as an unknown function of the data and estimated by means of a kernel or any other weighting scheme on the data.

Random field a scalar random field is defined as a function $m(\omega, x) : \Omega \times A \to R$ such that $m(\omega, x)$ is a random variable for each $x \in A$ where $A \subseteq R^k$.

Sieves the sieves or approximating spaces are approximations to an unknown function, that are dense in the original function space. Sieves can be constructed us-

ing linear spans of power series, e. g., Fourier series, splines, or many other basis functions such as artificial neural network (ANN), and various polynomials (Hermite, Laguerre, etc.).

Smooth transition models threshold model with the indicator function replaced by a smooth monotonically increasing differentiable function such as a probability distribution function.

Threshold model a nonlinear model with time-varying coefficients specified by using an indicator which takes a non-zero value when a state variable falls on a specified partition of a set of states, and zero otherwise. The number of partitions is finite.

Varying cross-sectional rank (VCR) of asset i is the proportion of assets that have a return less than or equal to the return of firm i at time t

$$z_{i,t} \equiv M^{-1} \sum_{j=1}^{M} \mathbf{1}(y_{j,t} \leq y_{i,t})$$

Volatility Volatility in financial economics is often measured by the conditional variance (e. g., ARCH) or the conditional range. It is important for any decision making under uncertainty such as portfolio allocation, option pricing, risk management.

Definition of the Subject

Financial Forecasting

Financial forecasting is concerned with the prediction of prices of financial assets such as stocks, bonds, options, interest rates, exchange rates, etc. Though many agents in the economy, i. e. investors, money managers, investment banks, hedge funds, etc. are interested in the forecasting of financial prices per se, the importance of financial forecasting derives primarily from the role of financial markets within the macro economy. The development of financial instruments and financial institutions contribute to the growth and stability of the overall economy. Because of this interconnection between financial markets and the real economy, financial forecasting is also intimately linked to macroeconomic forecasting, which is concerned with the prediction of macroeconomic aggregates such as growth of the gross domestic product, consumption growth, inflation rates, commodities prices, etc. Financial forecasting and macroeconomic forecasting share many of the techniques and statistical models that will be explained in detail in this article.

In financial forecasting a major object of study is the return to a financial asset, mostly calculated as the continuously compounded return, i. e., $y_t = \log p_t - \log p_{t-1}$ where p_t is the price of the asset at time t. Nowadays financial forecasters use sophisticated techniques that combine the advances in modern finance theory, pioneered by Markowitz [113], with the advances in time series econometrics, in particular the development of nonlinear models for conditional moments and conditional quantiles of asset returns.

The aim of finance theory is to provide models for expected returns taking into account the uncertainty of the future asset payoffs. In general, financial models are concerned with investors' decisions under uncertainty. For instance the portfolio allocation problem deals with the allocation of wealth among different assets that carry different levels of risk. The implementation of these theories relies on econometric techniques that aim to estimate financial models and testing them against the data. Financial econometrics is the branch of econometrics that provides model-based statistical inference for financial variables, and therefore financial forecasting will provide their corresponding model-based predictions. However there are also econometric developments that inform the construction of ad hoc time series models that are valuable on describing the stylized facts of financial data.

Since returns $\{y_t\}$ are random variables, the aim of financial forecasting is to forecast conditional moments, quantiles, and eventually the conditional distribution of these variables. Most of the time our interest will be centered on expected returns and volatility as these two moments are crucial components on portfolio allocation problems, option valuation, and risk management, but it is also possible to forecast quantiles of a random variable, and therefore to forecast the expected probability density function. Density forecasting is the most complete forecast as it embeds all the information on the financial variable of interest. Financial forecasting is also concerned with other financial variables like durations between trades and directions of price changes. In these cases, it is also possible to construct conditional duration models and conditional probit models that are the basis for forecasting durations and market timing.

Critical to the understanding of the methodological development in financial forecasting is the statistical concept of *martingale*, which historically has its roots in the games of chance also associated with the beginnings of probability theory in the XVI century. Borrowing from the concept of fair game, financial prices are said to enjoy the *martingale property* if tomorrow's price is expected to be equal to today's price given some information set; in other words tomorrow's price has an equal chance to either move up or move down, and thus the best forecast must be the current price. The martingale property is writ-

ten as

$$\mathbb{E}(p_{t+1}|\mathcal{F}_t) = p_t$$

where $\mathbb{E}$ is the expectation operator and the information set $\mathcal{F}_t \equiv \{p_t, p_{t-1}, p_{t-2}, \dots\}$ is the collection of past and current prices, though it may also include other variables known at time t such as volume. From a forecasting point of view, the martingale model implies that changes in financial prices $(p_{t+1} - p_t)$ are not predictable.

The most restrictive form of the martingale property, proposed by Bachelier [6] in his theory of speculation is the model (in logarithms)

$$\log p_{t+1} = \mu_t + \log p_t + \varepsilon_{t+1} ,$$

where $\mu_t = \mu$ is a constant drift and ε_{t+1} is an identically and independently distributed (i.i.d.) error that is assumed to be normally distributed with zero mean and constant variance σ^2. This model is also known as a random walk model. Since the return is the percentage change in prices, i. e. $y_t = \log p_t - \log p_{t-1}$, an equivalent model for asset returns is

$$y_{t+1} = \mu_t + \varepsilon_{t+1} .$$

Then, taking conditional expectations, we find that $\mathbb{E}(y_{t+1}|\mathcal{F}_t) = \mu_t$. If the conditional mean return is not time-varying, $\mu_t = \mu$, then the returns are not forecastable based on past price information. In addition and given the assumptions on the error term, returns are independent and identically distributed random variables. These two properties, a constant drift and an i.i.d error term, are too restrictive and they rule out the possibility of any predictability in asset returns. A less restrictive and more plausible version is obtained when the i.i.d assumption is relaxed. The error term may be heteroscedastic so that returns have different (unconditional or conditional) variances and consequently they are not identically distributed, and/or the error term, though uncorrelated, may exhibit dependence in higher moments and in this case the returns are not independent random variables.

The advent of modern finance theory brings the notion of systematic risk, associated with return variances and covariances, into asset pricing. Though these theories were developed to explain the cross-sectional variability of financial returns, they also helped many years later with the construction of time series models for financial returns. Arguably, the two most important asset pricing models in modern finance theory are the Capital Asset Pricing Model (CAPM) proposed by Sharpe [137] and Lintner [103] and the Arbitrage Pricing Theory (APT) proposed by Ross [131]. Both models claim that the expected return to an asset is a linear function of risk; in CAPM risk is related to the covariance of the asset return with the return to the market portfolio, and in APT risk is measured as exposure to a set of factors, which may include the market portfolio among others. The original version of CAPM, based on the assumption of normally distributed returns, is written as

$$\mathbb{E}(y_i) = y_f + \beta_{im}\left[\mathbb{E}(y_m) - y_f\right] ,$$

where y_f is the risk-free rate, y_m is the return to the market portfolio, and β_{im} is the risk of asset i defined as

$$\beta_{im} = \frac{\mathrm{cov}(y_i, y_m)}{\mathrm{var}(y_m)} = \frac{\sigma_{im}}{\sigma_m^2} .$$

This model has a time series version known as the conditional CAPM [17] that it may be useful for forecasting purposes. For asset i and given an information set as $\mathcal{F}_t = \{y_{i,t}, y_{i,t-1}, \dots; y_{m,t}, y_{m,t-1}, \dots\}$, the expected return is a linear function of a time-varying beta

$$\mathbb{E}(y_{i,t+1}|\mathcal{F}_t) = y_f + \beta_{im,t}\left[\mathbb{E}(y_{m,t+1}|\mathcal{F}_t) - y_f\right]$$

where $\beta_{im,t} = \frac{\mathrm{cov}(y_{i,t+1}, y_{m,t+1}|\mathcal{F}_t)}{\mathrm{var}(y_{m,t+1}|\mathcal{F}_t)} = \frac{\sigma_{im,t}}{\sigma_{m,t}^2}$. From this type of models is evident that we need to model the conditional second moments of returns jointly with the conditional mean. A general finding of this type of models is that when there is high volatility, expected returns are high, and hence forecasting volatility becomes important for the forecasting of expected returns. In the same spirit, the APT models have also conditional versions that exploit the information contained in past returns. A K-factor APT model is written as

$$y_t = c + B' f_t + \varepsilon_t ,$$

where f_t is a $K \times 1$ vector of factors and B is a $K \times 1$ vector of sensitivities to the factors. If the factors have time-varying second moments, it is possible to specify an APT model with a factor structure in the time-varying covariance matrix of asset returns [48], which in turn can be exploited for forecasting purposes.

The conditional CAPM and conditional APT models are fine examples on how finance theory provides a base to specify time-series models for financial returns. However there are other time series specifications, more *ad hoc* in nature, that claim that financial prices are nonlinear functions – not necessarily related to time-varying second moments – of the information set and by that, they impose some departures from the martingale property. In this

case it is possible to observe some predictability in asset prices. This is the subject of nonlinear financial forecasting. We begin with a precise definition of linearity versus nonlinearity.

Linearity and Nonlinearity

Lee, White, and Granger [99] are the first who precisely define the concept of "linearity". Let $\{Z_t\}$ be a stochastic process, and partition Z_t as $Z_t = (y_t\ X_t')'$, where (for simplicity) y_t is a scalar and X_t is a $k \times 1$ vector. X_t may (but need not necessarily) contain a constant and lagged values of y_t. LWG define that the process $\{y_t\}$ is *linear in mean conditional on* X_t if

$$\Pr\left[\mathbb{E}(y_t|X_t) = X_t'\theta^*\right] = 1 \quad \text{for some } \theta^* \in \mathbb{R}^k .$$

In the context of forecasting, Granger and Lee [71] define linearity as follows. Define $\mu_{t+h} = \mathbb{E}(y_{t+h}|\mathcal{F}_t)$ being the optimum least squares h-step forecast of y_{t+h} made at time t. μ_{t+h} will generally be a nonlinear function of the contents of $\mathcal{F}_t$. Denote m_{t+h} the optimum *linear* forecast of y_{t+h} made at time t be the best forecast that is constrained to be a linear combination of the contents of $X_t \in \mathcal{F}_t$. Granger and Lee [71] define that $\{y_t\}$ is said to be *linear in conditional mean* if μ_{t+h} is linear in X_t, i. e., $\Pr\left[\mu_{t+h} = m_{t+h}\right] = 1$ for all t and for all h. Under this definition the focus is the conditional mean and thus a process exhibiting autoregressive conditional heteroskedasticity (ARCH) [44] may nevertheless exhibit linearity of this sort because ARCH does not refer to the conditional mean. This is appropriate whenever we are concerned with the adequacy of linear models for forecasting the conditional mean returns. See [161], Section 2, for a more rigorous treatment on the definitions of linearity and nonlinearity.

This definition may be extended with some caution to the concept of linearity in higher moments and quantiles, but the definition may depend on the focus or interest of the researcher. Let $\varepsilon_{t+h} = y_{t+h} - \mu_{t+h}$ and $\sigma^2_{t+h} = \mathbb{E}(\varepsilon^2_{t+h}|\mathcal{F}_t)$. If we consider the ARCH and GARCH as linear models, we say $\left\{\sigma^2_{t+h}\right\}$ is linear in conditional variance if σ^2_{t+h} is a linear function of lagged ε^2_{t-j} and σ^2_{t-j} for some h or for all h. Alternatively, $\sigma^2_{t+h} = \mathbb{E}(\varepsilon^2_{t+h}|\mathcal{F}_t)$ is said to be linear in conditional variance if σ^2_{t+h} is a linear function of $x_t \in \mathcal{F}_t$ for some h or for all h. Similarly, we may consider linearity in conditional quantiles. The issue of linearity versus nonlinearity is most relevant for the conditional mean. It is more relevant whether a certain specification is correct or incorrect (rather than linear or nonlinear) for higher order conditional moments or quantiles.

Introduction

There exists a nontrivial gap between martingale difference and serial uncorrelatedness. The former implies the latter, but not vice versa. Consider a stationary time series $\{y_t\}$. Often, serial dependence of $\{y_t\}$ is described by its autocorrelation function $\rho(j)$, or by its standardized spectral density

$$h(\omega) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\rho(j)\mathrm{e}^{-ij\omega}, \quad \omega \in [-\pi, \pi] .$$

Both $h(\omega)$ and $\rho(j)$ are the Fourier transform of each other, containing the same information of serial correlations of $\{y_t\}$. A problem with using $h(\omega)$ and $\rho(j)$ is that they cannot capture nonlinear time series that have zero autocorrelation but are not serially independent. Nonlinear MA and Bilinear series are good examples:

$$\begin{aligned} \text{Nonlinear MA}: &\quad Y_t = be_{t-1}e_{t-2} + e_t , \\ \text{Bilinear}: &\quad Y_t = be_{t-1}Y_{t-2} + e_t . \end{aligned}$$

These processes are serially uncorrelated, but they are predictable using the past information. Hong and Lee [86] note that the autocorrelation function, the variance ratios, and the power spectrum can easily miss these processes. Misleading conclusions in favor of the martingale hypothesis could be reached when these test statistics are insignificant. It is therefore important and interesting to explore whether there exists a gap between serial uncorrelatedness and martingale difference behavior for financial forecasting, and if so, whether the neglected nonlinearity in conditional mean can be explored to forecast financial asset returns.

In the forthcoming sections, we will present, without being exhaustive, nonlinear time series models for financial returns, which are the basis for nonlinear forecasting. In Sect. "Nonlinear Forecasting Models for the Conditional Mean", we review nonlinear models for the conditional mean of returns. A general representation is $y_{t+1} = \mu(y_t, y_{t-1}, \dots) + \varepsilon_{t+1}$ with $\mu(\cdot)$ a nonlinear function of the information set. If $\mathbb{E}(y_{t+1}|y_t, y_{t-1}, \dots) = \mu(y_t, y_{t-1}, \dots)$, then there is a departure from the martingale hypothesis, and past price information will be relevant to predict tomorrow's return. In Sect. "Nonlinear Forecasting Models for the Conditional Variance", we review models for the conditional variance of returns. For instance, a model like $y_{t+1} = \mu + u_{t+1}\sigma_{t+1}$ with time-varying conditional variance $\sigma^2_{t+1} = \mathbb{E}((y_{t+1} - \mu)^2|\mathcal{F}_t)$ and i.i.d. error u_{t+1}, is still a martingale-difference for returns but it represents a departure from the

independence assumption. The conditional mean return may not be predictable but the conditional variance of the return will be. In addition, as we have seen modeling time-varying variances and covariances will be very useful for the implementation of conditional CAPM and APT models.

Nonlinear Forecasting Models for the Conditional Mean

We consider models to forecast the expected price changes of financial assets and we restrict the loss function of the forecast error to be the mean squared forecast error (MSFE). Under this loss, the optimal forecast is $\mu_{t+h} = \mathbb{E}(y_{t+h}|\mathcal{F}_t)$. Other loss functions may also be used but it will be necessary to forecast other aspects of the forecast density. For example, under a mean absolute error loss function the optimal forecast is the conditional median.

There is evidence for μ_{t+h} being time-varying. Simple linear autoregressive polynomials in lagged price changes are not sufficient to model μ_{t+h} and nonlinear specifications are needed. These can be classified into parametric and nonparametric. Examples of parametric models are autoregressive bilinear and threshold models. Examples of nonparametric models are artificial neural network, kernel and nearest neighbor regression models.

It will be impossible to have an exhaustive review of the many nonlinear specifications. However, as discussed in White [161] and Chen [25], some nonlinear models are universal approximators. For example, the sieves or approximating spaces are proven to approximate very well unknown functions and they can be constructed using linear spans of power series, Fourier series, splines, or many other basis functions such as artificial neural network (ANN), Hermite polynomials as used in e. g., [56] for modelling semi-nonparametric density, and Laguerre polynomials used in [119] for modelling the yield curve. Diebold and Li [36] and Huang, Lee, and Li [89] use the Nelson–Siegel model in forecasting yields and inflation.

We review parametric nonlinear models like threshold model, smooth transition model, Markov switching model, and random fields model; nonparametric models like local linear, local polynomial, local exponential, and functional coefficient models; and nonlinear models based on sieves like ANN and various polynomials approximations. For other nonlinear specifications we recommend some books on nonlinear time series models such as Fan and Yao [52], Gao [57], and Tsay [153]. We begin with a very simple nonlinear model.

A Simple Nonlinear Model with Dummy Variables

Goyal and Welch [66] forecast the equity premium on the S&P 500 index – index return minus T-bill rate – using many predictors such as stock-related variables (e. g., dividend-yield, earning-price ratio, book-to-market ratio, corporate issuing activity, etc.), interest-rate-related variables (e. g., treasury bills, long-term yield, corporate bond returns, inflation, investment to capital ratio), and ex ante consumption, wealth, income ratio (modified from [101]). They find that these predictors have better performance in bad times, such as the Great Depression (1930–33), the oil-shock period (1973–75), and the tech bubble-crash period (1999–2001). Also, they argue that it is reasonable to impose a lower bound (e. g., zero or 2%) on the equity premium because no investor is interested in (say) a negative premium.

Campbell and Thompson [23], inspired by the out-of-sample forecasting of Goyal and Welch [66], argue that if we impose some restrictions on the signs of the predictors' coefficients and excess return forecasts, some predictors can beat the historical average equity premium. Similarly to Goyal and Welch [66], they also use a rich set of forecasting variables – valuation ratios (e. g., dividend price ratio, earning price ratio, and book to market ratio), real return on equity, nominal interest rates and inflation, and equity share of new issues and consumption-wealth ratio. They impose two restrictions – the first one is to restrict the predictors' coefficients to have the theoretically expected sign and to set wrong-signed coefficients to zero, and the second one is to rule out a negative equity premium forecast. They show that the effectiveness of these theoretically-inspired restrictions almost always improve the out-of sample performance of the predictive regressions. This is an example where "shrinkage" works, that is to reduce the forecast error variance at the cost of a higher forecast bias but with an overall smaller mean squared forecast error (the sum of error variance and the forecast squared bias).

The results from Goyal and Welch [66] and Campbell and Thompson [23] support a simple form of nonlinearity that can be generalized to threshold models or time-varying coefficient models, which we consider next.

Threshold Models

Many financial and macroeconomic time series exhibit different characteristics over time depending upon the state of the economy. For instance, we observe bull and bear stock markets, high volatility versus low volatility periods, recessions versus expansions, credit crunch versus excess liquidity, etc. If these different regimes are present

in economic time series data, econometric specifications should go beyond linear models as these assume that there is only a single structure or regime over time. Nonlinear time series specifications that allow for the possibility of different regimes, also known as state-dependent models, include several types of models: threshold, smooth transition, and regime-switching models.

Threshold autoregressive (TAR) models [148,149] assume that the dynamics of the process is explained by an autoregression in each of the n regimes dictated by a conditioning or threshold variable. For a process $\{y_t\}$, a general specification of a TAR model is

$$y_t = \sum_{j=1}^{n} \left[\phi_o^{(j)} + \sum_{i=1}^{p_j} \phi_i^{(j)} y_{t-i} + \varepsilon_t^{(j)} \right] \mathbf{1}(r_{j-1} < x_t \leq r_j).$$

There are n regimes, in each one there is an autoregressive process of order p_j with different autoregressive parameters $\phi_i^{(j)}$, the threshold variable is x_t with r_j thresholds and $r_o = -\infty$ and $r_n = +\infty$, and the error term is assumed i.i.d. with zero mean and different variance across regimes $\varepsilon_t^{(j)} \sim$ i.i.d. $\left(0, \sigma_j^2\right)$, or more generally $\varepsilon_t^{(j)}$ is assumed to be a martingale difference. When the threshold variable is the lagged dependent variable itself y_{t-d}, the model is known as self-exciting threshold autoregressive (SETAR) model. The SETAR model has been applied to the modelling of exchange rates, industrial production indexes, and gross national product (GNP) growth, among other economic data sets. The most popular specifications within economic time series tend to find two, at most three regimes. For instance, Boero and Marrocu [18] compare a two and three-regime SETAR models with a linear AR with GARCH disturbances for the euro exchange rates. On the overall forecasting sample, the linear model performs better than the SETAR models but there is some improvement in the predictive performance of the SETAR model when conditioning on the regime.

Smooth Transition Models

In the SETAR specification, the number of regimes is discrete and finite. It is also possible to model a *continuum* of regimes as in the Smooth Transition Autoregressive (STAR) models [144]. A typical specification is

$$y_t = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i} + \left(\theta_0 + \sum_{i=1}^{p} \theta_i y_{t-i} \right) F(y_{t-d}) + \varepsilon_t$$

where $F(y_{t-d})$ is the transition function that is continuous and in most cases is either a logistic function or an exponential,

$$F(y_{t-d}) = \left[1 + \exp\left(-\gamma \left(y_{t-d} - r\right)\right)\right]^{-1}$$
$$F(y_{t-d}) = 1 - \left[\exp\left(-\gamma \left(y_{t-d} - r\right)^2\right)\right]$$

This model can be understood as many autoregressive regimes dictated by the values of the function $F(y_{t-d})$, or alternatively as an autoregression where the autoregressive parameters change smoothly over time. When $F(y_{t-d})$ is logistic and $\gamma \to \infty$, the STAR model collapses to a threshold model SETAR with two regimes. One important characteristic of these models, SETAR and STAR, is that the process can be stationary within some regimes and non-stationary within others moving between explosive and contractionary stages.

Since the estimation of these models can be demanding, the first question to solve is whether the nonlinearity is granted by the data. A test for linearity is imperative before engaging in the estimation of nonlinear specifications. An LM test that has power against the two alternatives specifications SETAR and STAR is proposed by Luukkonen et al. [110] and it consists of running two regressions: under the null hypothesis of linearity, a linear autoregression of order p is estimated in order to calculate the sum of squared residuals, SSE_0; the second is an auxiliary regression

$$y_t = \beta_0 + \sum_{i=1}^{p} \beta_i y_{t-i} + \sum_{i=1}^{p} \sum_{j=1}^{p} \psi_{ij} y_{t-i} y_{t-j} + \sum_{i=1}^{p} \sum_{j=1}^{p} \zeta_{ij} y_{t-i} y_{t-j}^2 + \sum_{i=1}^{p} \sum_{j=1}^{p} \xi_{ij} y_{t-i} y_{t-j}^3 + u_t$$

from which we calculate the sum of squared residuals, SSE_1. The test is constructed as $\chi^2 = T(SSE_0 - SSE_1)/SSE_0$ that under the null hypothesis of linearity is chi-squared distributed with $p(p+1)/2 + 2p^2$ degrees of freedom. There are other tests in the literature, for instance Hansen [80] proposes a likelihood ratio test that has a non-standard distribution, which is approximated by implementing a bootstrap procedure. Tsay [151] proposes a test based on arranged regressions with respect to the increasing order of the threshold variable and by doing this the testing problem is transformed into a change-point problem.

If linearity is rejected, we proceed with the estimation of the nonlinear specification. In the case of the SETAR model, if we fix the values of the delay parameter d and the thresholds r_j, the model reduces to n linear regressions for which least squares estimation is straightforward.

Tsay [151] proposes a conditional least squares (CLS) estimator. For simplicity of exposition suppose that there are two regimes in the data and the model to estimate is

$$y_t = \left[\phi_o^{(1)} + \sum_{i=1}^{p_1} \phi_i^{(1)} y_{t-i}\right] \mathbf{1}(y_{t-d} \le r) + \left[\phi_o^{(2)} + \sum_{i=1}^{p_2} \phi_i^{(2)} y_{t-i}\right] \mathbf{1}(y_{t-d} > r) + \varepsilon_t$$

Since r and d are fixed, we can apply least squares estimation to the model and to obtain the LS estimates for the parameters ϕ_i's. With the LS residual $\hat{\varepsilon}_t$, we obtain the total sum of squares $S(r,d) = \sum_t \hat{\varepsilon}_t^2$. The CLS estimates of r and d are obtained from $(\hat{r}, \hat{d}) = \arg\min S(r,d)$.

For the STAR model, it is also necessary to specify a priori the functional form of $F(y_{t-d})$. Teräsvirta [144] proposes a modeling cycle consisting of three stages: specification, estimation, and evaluation. In general, the specification stage consists of sequence of null hypothesis to be tested within a linearized version of the STAR model. Parameter estimation is carried out by nonlinear least squares or maximum likelihood. The evaluation stage mainly consists of testing for no error autocorrelation, no remaining nonlinearity, and parameter constancy, among other tests.

Teräsvirta and Anderson [146] find strong nonlinearity in the industrial production indexes of most of the OECD countries. The preferred model is the logistic STAR with two regimes, recessions and expansions. The dynamics in each regime are country dependent. For instance, in USA they find that the economy tends to move from recessions into expansions very aggressively but it will take a large negative shock to move rapidly from an expansion into a recession. Other references for applications of these models to financial series are found in [28,73,94].

For forecasting with STAR models, see Lundbergh and Teräsvirta [109]. It is easy to construct the one-step-ahead forecast but the multi-step-ahead forecast is a complex problem. For instance, for the 2-regime threshold model, the one-step-ahead forecast is constructed as the conditional mean of the process given some information set

$$\mathbb{E}(y_{t+1}|\mathcal{F}_t;\theta) = \left[\phi_o^{(1)} + \sum_{i=1}^{p_1} \phi_i^{(1)} y_{t+1-i}\right] \mathbf{1}(y_{t+1-d} \le r) + \left[\phi_o^{(2)} + \sum_{i=1}^{p_2} \phi_i^{(2)} y_{t+1-i}\right] \mathbf{1}(y_{t+1-d} > r)$$

provided that $y_{t+1-i}, y_{t+1-d} \in \mathcal{F}_t$. However, a multi-step-ahead forecast will be a function of variables that being dated at a future date do not belong to the information set; in this case the solution requires the use of numerical integration techniques or simulation/bootstrap procedures. See Granger and Teräsvirta [72], Chapter 9, and Teräsvirta [145] for more details on numerical methods for multi-step forecasts.

Markov-Switching Models

A Markov-switching (MS) model [76,77] also features changes in regime, but in contrast with the SETAR models the change is dictated by a non-observable state variable that is modelled as a Markov chain. For instance, a first order autoregressive Markov switching model is specified as

$$y_t = c_{s_t} + \phi_{s_t} y_{t-1} + \varepsilon_t$$

where $s_t = 1, 2, \ldots, N$ is the unobserved state variable that is modelled as an N-state Markov chain with transition probabilities $p_{ij} = P(s_t = j | s_{t-1} = i)$, and $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$ or more generally ε_t is a martingale difference. Conditioning in a given state and an information set $\mathcal{F}_t$, the process $\{y_t\}$ is linear but unconditionally the process is nonlinear. The conditional forecast is $\mathbb{E}(y_{t+1}|s_{t+1} = j, \mathcal{F}_t; \theta) = c_j + \phi_j y_t$ and the unconditional forecast based on observable variables is the sum of the conditional forecasts for each state weighted by the probability of being in that state,

$$\mathbb{E}(y_{t+1}|\mathcal{F}_t;\theta) = \sum_{j=1}^{N} P(s_{t+1} = j|\mathcal{F}_t;\theta)\mathbb{E}(y_{t+1}|s_{t+1} = j, \mathcal{F}_t;\theta)\,.$$

The parameter vector $\theta = (c_1 \ldots c_N, \phi_1 \ldots \phi_N, \sigma^2)'$ as well as the transition probabilities p_{ij} can be estimated by maximum likelihood.

MS models have been applying to the modeling of foreign exchange rates with mixed success. Engel and Hamilton [43] fit a two-state MS for the Dollar and find that there are long swings and by that they reject the random walk behavior in the exchange rate. Marsh [114] estimates a two-state MS for the Deutschemark, the Pound Sterling, and the Japanese Yen. Though the model approximates the characteristics of the data well, the forecasting performance is poor when measured by the profit/losses generated by a set of trading rules based on the predictions of the MS model. On the contrary, Dueker and Neely [40] find that for the same exchange rate a MS model with three states variables – in the scale factor of the variance of a Student-t error, in the kurtosis of the error, and in

the expected return – produces out-of-sample excess returns that are slightly superior to those generated by common trading rules. For stock returns, there is evidence that MS models perform relatively well on describing two states in the mean (high/low returns) and two states in the variance (stable/volatile periods) of returns [111]. In addition, Perez-Quiros and Timmermann [124] propose that the error term should be modelled as a mixture of Gaussian and Student-t distributions to capture the outliers commonly found in stock returns. This model provides some gains in predictive accuracy mainly for small firms returns. For interest rates in USA, Germany, and United Kingdom, Ang and Bekaert [5] find that a two-state MS model that incorporates information on international short rate and on term spread is able to predict better than an univariate MS model. Additionally they find that in USA the classification of regimes correlates well with the business cycles.

SETAR, STAR, and MS models are successful specifications to approximate the characteristics of financial and macroeconomic data. However, good in-sample performance does not imply necessarily a good out-of-sample performance, mainly when compared to simple linear ARMA models. The success of nonlinear models depends on how prominent the nonlinearity is in the data. We should not expect a nonlinear model to perform better than a linear model when the contribution of the nonlinearity to the overall specification of the model is very small. As it is argued in Granger and Teräsvirta [72], the prediction errors generated by a nonlinear model will be smaller only when the nonlinear feature modelled in-sample is also present in the forecasting sample.

A State Dependent Mixture Model Based on Cross-sectional Ranks

In the previous section, we have dealt with nonlinear time series models that only incorporate time series information. González-Rivera, Lee, and Mishra [63] propose a nonlinear model that combines time series with cross sectional information. They propose the modelling of expected returns based on the joint dynamics of a sharp jump in the cross-sectional rank and the realized returns. They analyze the marginal probability distribution of a jump in the cross-sectional rank within the context of a duration model, and the probability of the asset return conditional on a jump specifying different dynamics depending on whether or not a jump has taken place. The resulting model for expected returns is a mixture of normal distributions weighted by the probability of jumping.

Let $y_{i,t}$ be the return of firm i at time t, and $\{y_{i,t}\}_{i=1}^{M}$ be the collection of asset returns of the M firms that constitute the *market* at time t. For each time t, the asset returns are ordered from the smallest to the largest, and define $z_{i,t}$, the *Varying Cross-sectional Rank* (VCR) of firm i within the market, as the proportion of firms that have a return less than or equal to the return of firm i. We write

$$z_{i,t} \equiv M^{-1} \sum_{j=1}^{M} \mathbf{1}(y_{j,t} \leq y_{i,t}) \,, \tag{1}$$

where $\mathbf{1}(\cdot)$ is the indicator function, and for M large, $z_{i,t} \in (0, 1]$. Since the rank is a highly dependent variable, it is assumed that small movements in the asset ranking will not contain significant information and that most likely large movements in ranking will be the result of news in the overall market and/or of news concerning a particular asset. Focusing on large rank movements, we define, at time t, a sharp jump as a binary variable that takes the value one when there is a minimum (upward or downward) movement of 0.5 in the ranking of asset i, and zero otherwise:

$$J_{i,t} \equiv \mathbf{1}(|z_{i,t} - z_{i,t-1}| \geq 0.5) \,. \tag{2}$$

A jump of this magnitude brings the asset return above or below the median of the cross-sectional distribution of returns. Note that this notion of jumps differs from the more traditional meaning of the word in the context of continuous-time modelling of the univariate return process. A jump in the cross-sectional rank implicitly depends on numerous univariate return processes.

The analytical problem now consists in modeling the joint distribution of the return $y_{i,t}$ and the jump $J_{i,t}$, i. e. $f(y_{i,t}, J_{i,t}|\mathcal{F}_{t-1})$ where $\mathcal{F}_{t-1}$ is the information set up to time $t-1$. Since $f(y_{i,t}, J_{i,t}|\mathcal{F}_{t-1}) = f_1(J_{i,t}|\mathcal{F}_{t-1}) f_2(y_{i,t}|J_{i,t}, \mathcal{F}_{t-1})$, the analysis focuses first on the modelling of the marginal distribution of the jump, and subsequently on the modelling of the conditional distribution of the return.

Since $J_{i,t}$ is a Bernoulli variable, the marginal distribution of the jump is $f_1(J_{i,t}|\mathcal{F}_{t-1}) = p_{i,t}^{J_{i,t}}(1 - p_{i,t})^{(1-J_{i,t})}$ where $p_{i,t} \equiv \Pr(J_{i,t} = 1|\mathcal{F}_{t-1})$ is the conditional probability of a jump in the cross-sectional ranks. The modelling of $p_{i,t}$ is performed within the context of a dynamic duration model specified in calendar time as in Hamilton and Jordà [79]. The calendar time approach is necessary because asset returns are reported in calendar time (days, weeks, etc.) and it has the advantage of incorporating any other available information also reported in calendar time.

It is easy to see that the probability of jumping and duration must have an inverse relationship. If the probability

of jumping is high, the expected duration must be short, and vice versa. Let $\Psi_{N(t)}$ be the expected duration. The expected duration until the next jump in the cross-sectional rank is given by $\Psi_{N(t)} = \sum_{j=1}^{\infty} j(1-p_t)^{j-1}p_t = p_t^{-1}$. Note that $\sum_{j=0}^{\infty}(1-p_t)^j = p_t^{-1}$. Differentiating with respect to p_t yields $\sum_{j=0}^{\infty} -j(1-p_t)^{j-1} = -p_t^{-2}$. Multiplying by $-p_t$ gives $\sum_{j=0}^{\infty} j(1-p_t)^{j-1}p_t = p_t^{-1}$ and thus $\sum_{j=1}^{\infty} j(1-p_t)^{j-1}p_t = p_t^{-1}$. Consequently, to model $p_{i,t}$, it suffices to model the expected duration and compute its inverse. Following Hamilton and Jordà [79], an autoregressive conditional hazard (ACH) model is specified. The ACH model is a calendar-time version of the autoregressive conditional duration (ACD) of Engle and Russell [49]. In both ACD and ACH models, the expected duration is a linear function of lag durations. However as the ACD model is set up in event time, there are some difficulties on how to introduce information that arrives between events. This is not the case in the ACH model because the set-up is in calendar time. In the ACD model, the forecasting object is the expected time between events; in the ACH model, the objective is to forecast the probability that the event will happen tomorrow given the information known up to today. A general ACH model is specified as

$$\Psi_{N(t)} = \sum_{j=1}^{m} \alpha_j D_{N(t)-j} + \sum_{j=1}^{r} \beta_j \Psi_{N(t)-j} \,. \tag{3}$$

Since p_t is a probability, it must be bounded between zero and one. This implies that the conditional duration must have a lower bound of one. Furthermore, working in calendar time it is possible to incorporate information that becomes available between jumps and can affect the probability of a jump in future periods. The conditional hazard rate is specified as

$$p_t = [\Psi_{N(t-1)} + \delta' X_{t-1}]^{-1} \,, \tag{4}$$

where X_{t-1} is a vector of relevant calendar time variables such as past VCRs and past returns. This completes the marginal distribution of the jump $f_1(J_{i,t}|\mathcal{F}_{t-1}) = p_{i,t}^{J_{i,t}}(1-p_{i,t})^{(1-J_{i,t})}$.

On modelling $f_2(y_t|J_t, \mathcal{F}_{t-1}; \theta_2)$, it is assumed that the return to asset i may behave differently depending upon the occurrence of a jump. The modelling of two potential different states (whether a jump has occurred or not) will permit to differentiate whether the conditional expected return is driven by active or/and passive movements in the asset ranking in conjunction with its own return dynamics. A priori, different dynamics are possible in these two states. A general specification is

$$f_2(y_t|J_t, \mathcal{F}_{t-1}; \theta_2) = \begin{cases} N(\mu_{1,t}, \sigma_{1,t}^2) & \text{if } J_t = 1 \\ N(\mu_{0,t}, \sigma_{0,t}^2) & \text{if } J_t = 0 \end{cases} \,, \tag{5}$$

where $\mu_{j,t}$ is the conditional mean and $\sigma_{j,t}^2$ the conditional variance in each state ($j = 1, 0$). Whether these two states are present in the data is an empirical question and it should be answered through statistical testing.

Combining the models for the marginal density of the jump and the conditional density of the returns, the estimation can be conducted with maximum likelihood techniques. For a sample $\{y_t, J_t\}_{t=1}^T$, the joint log-likelihood function is

$$\begin{aligned} &\sum_{t=1}^{T} \ln f(y_t, J_t|\mathcal{F}_{t-1}; \theta) \\ &= \sum_{t=1}^{T} \ln f_1(J_t|\mathcal{F}_{t-1}; \theta_1) + \sum_{t=1}^{T} \ln f_2(y_t|J_t, \mathcal{F}_{t-1}; \theta_2) \,. \end{aligned}$$

Let us call $\mathcal{L}_1(\theta_1) = \sum_{t=1}^T \ln f_1(J_t|\mathcal{F}_{t-1}; \theta_1)$ and $\mathcal{L}_2(\theta_2) = \sum_{t=1}^T \ln f_2(y_t|J_t, \mathcal{F}_{t-1}; \theta_2)$. The maximization of the joint log-likelihood function can be achieved by maximizing $\mathcal{L}_1(\theta_1)$ and $\mathcal{L}_2(\theta_2)$ separately without loss of efficiency by assuming that the parameter vectors θ_1 and θ_2 are "variation free" in the sense of Engle et al. [45].

The log-likelihood function $\mathcal{L}_1(\theta_1) = \sum_{t=1}^T \ln f_1(J_t|\mathcal{F}_{t-1}; \theta_1)$ is

$$\mathcal{L}_1(\theta_1) = \sum_{t=1}^{T} \left[J_t \ln p_t(\theta_1) + (1-J_t)\ln(1-p_t(\theta_1)) \right] , \tag{6}$$

where θ_1 includes all parameters in the conditional duration model.

The log-likelihood function $\mathcal{L}_2(\theta_2) = \sum_{t=1}^T \ln f_2(y_t|J_t, \mathcal{F}_{t-1}; \theta_2)$ is

$$\begin{aligned} \mathcal{L}_2(\theta_2) = \sum_{t=1}^{T} \ln \Bigg[&\frac{J_t}{\sqrt{2\pi\sigma_{1,t}^2}} \exp\left\{ -\frac{1}{2}\left(\frac{y_t - \mu_{1,t}}{\sigma_{1,t}}\right)^2 \right\} \\ &+ \frac{1-J_t}{\sqrt{2\pi\sigma_{0,t}^2}} \exp\left\{ -\frac{1}{2}\left(\frac{y_t - \mu_{0,t}}{\sigma_{0,t}}\right)^2 \right\} \Bigg] , \end{aligned}$$

where θ_2 includes all parameters in the conditional means and conditional variances under both regimes.

If the two proposed states are granted in the data, the marginal density function of the asset return must be

a mixture of two normal density functions where the mixture weights are given by the probability of jumping p_t:

$$\begin{aligned} g(y_t|\mathcal{F}_{t-1};\theta) &\equiv \sum_{J_t=0}^{1} f(y_t, J_t|\mathcal{F}_{t-1};\theta) \\ &= \sum_{J_t=0}^{1} f_1(J_t|\mathcal{F}_{t-1};\theta_1) f_2(y_t|J_t, \mathcal{F}_{t-1};\theta_2) \\ &= p_t \cdot f_2(y_t|J_t = 1, \mathcal{F}_{t-1};\theta_2) \\ &\quad + (1-p_t)\cdot f_2(y_t|J_t = 0, \mathcal{F}_{t-1};\theta_2), \end{aligned} \tag{7}$$

as $f_1(J_t|\mathcal{F}_{t-1};\theta_1) = p_t^{J_t}(1-p_t)^{(1-J_t)}$. Therefore, the one-step ahead forecast of the return is

$$\begin{aligned} &\mathbb{E}(y_{t+1}|\mathcal{F}_t;\theta) \\ &= \int y_{t+1}\cdot g(y_{t+1}|\mathcal{F}_t;\theta)\mathrm{d}y_{t+1} \\ &= p_{t+1}(\theta_1)\cdot \mu_{1,t+1}(\theta_2) + (1-p_{t+1}(\theta_1))\cdot \mu_{0,t+1}(\theta_2)\,. \end{aligned} \tag{8}$$

The expected return is a function of the probability of jumping p_t, which is a nonlinear function of the information set as shown in (4). Hence the expected returns are nonlinear functions of the information set, even in a simple case where $\mu_{1,t}$ and $\mu_{0,t}$ are linear.

This model was estimated for the returns of the constituents of the SP500 index from 1990 to 2000, and its performance was assessed in an out-of-sample exercise from 2001 to 2005 within the context of several trading strategies. Based on the one-step-ahead forecast of the mixture model, a proposed trading strategy called VCR-Mixture Trading Rule is shown to be a superior rule because of its ability to generate large risk-adjusted mean returns when compared to other technical and model-based trading rules. The VCR-Mixture Trading Rule is implemented by computing for each firm in the SP500 index the one-step ahead forecast of the return as in (8). Based on the forecasted returns $\{\hat{y}_{i,t+1}(\hat{\theta}_t)\}_{t=R}^{T-1}$, the investor predicts the VCR of all assets in relation to the overall market, that is,

$$\hat{z}_{i,t+1} = M^{-1}\sum_{j=1}^{M} \mathbf{1}(\hat{y}_{j,t+1} \le \hat{y}_{i,t+1}), \qquad t = R,\ldots,T-1\,, \tag{9}$$

and buys the top K performing assets if their forecasted return is above the risk-free rate. In every subsequent out-of-sample period ($t = R,\ldots,T-1$), the investor revises her portfolio, selling the assets that fall out of the top performers and buying the ones that rise to the top, and she computes the one-period portfolio return

$$\pi_{t+1} = K^{-1}\sum_{j=1}^{M} y_{j,t+1}\cdot \mathbf{1}\left(\hat{z}_{j,t+1} \ge z_{t+1}^{K}\right), \qquad t = R,\ldots,T-1\,, \tag{10}$$

where z_{t+1}^{K} is the cutoff cross-sectional rank to select the K best performing stocks such that $\sum_{j=1}^{M}\mathbf{1}\left(\hat{z}_{j,t+1} \ge z_{t+1}^{K}\right) = K$. In the analysis of González-Rivera, Lee, and Mishra [63] a portfolio is formed with the top 1% ($K = 5$ stocks) performers in the SP500 index. Every asset in the portfolio is weighted equally. The evaluation criterion is to compute the "mean trading return" over the forecasting period

$$MTR = P^{-1}\sum_{t=R}^{T-1}\pi_{t+1}\,.$$

It is also possible to correct MTR according to the level of risk of the chosen portfolio. For instance, the traditional Sharpe ratio will provide the excess return per unit of risk measured by the standard deviation of the selected portfolio

$$SR = P^{-1}\sum_{t=R}^{T-1}\frac{(\pi_{t+1} - r_{f,t+1})}{\sigma_{t+1}^{\pi}(\hat{\theta}_t)}\,,$$

where $r_{f,t+1}$ is the risk free rate. The VCR-Mixture Trading Rule produces a weekly MTR of 0.243% (63.295% cumulative return over 260 weeks), equivalent to a yearly compounded return of 13.45%, that is significantly more than the next most favorable rule, which is the Buy-and-Hold-the-Market Trading Rule with a weekly mean return of −0.019%, equivalent to a yearly return of −1.00%. To assess the return-risk trade off, we implement the Sharpe ratio. The largest SR (mean return per unit of standard deviation) is provided by the VCR-Mixture rule with a weekly return of 0.151% (8.11% yearly compounded return per unit of standard deviation), which is lower than the mean return provided by the same rule under the MTR criterion, but still a dominant return when compared to the mean returns provided by the Buy-and-Hold-the-Market Trading Rule.

Random Fields

Hamilton [78] proposed a flexible parametric regression model where the conditional mean has a linear parametric component and a potential nonlinear component

represented by an isotropic Gaussian random field. The model has a nonparametric flavor because no functional form is assumed but, nevertheless, the estimation is fully parametric.

A scalar random field is defined as a function $m(\omega, x) : \Omega \times A \rightarrow R$ such that $m(\omega, x)$ is a random variable for each $x \in A$ where $A \subseteq R^k$. A random field is also denoted as $m(x)$. If $m(x)$ is a system of random variables with finite dimensional Gaussian distributions, then the scalar random field is said to be Gaussian and it is completely determined by its mean function $\mu(x) = \mathbb{E}\left[m(x)\right]$ and its covariance function with typical element $C(x, z) = \mathbb{E}\left[(m(x) - \mu(x))(m(z) - \mu(z))\right]$ for any $x, z \in A$. The random field is said to be homogeneous or stationary if $\mu(x) = \mu$ and the covariance function depends only on the difference vector $x - z$ and we should write $C(x, z) = C(x - z)$. Furthermore, the random field is said to be isotropic if the covariance function depends on $d(x, z)$, where $d(\cdot)$ is a scalar measure of distance. In this situation we write $C(x, z) = C(d(x, z))$.

The specification suggested by Hamilton [78] can be represented as

$$y_t = \beta_0 + x_t'\beta_1 + \lambda m(g \odot x_t) + \epsilon_t \,, \tag{11}$$

for $y_t \in R$ and $x_t \in R^k$, both stationary and ergodic processes. The conditional mean has a linear component given by $\beta_0 + x_t'\beta_1$ and a nonlinear component given by $\lambda m(g \odot x_t)$, where $m(z)$, for any choice of z, represents a realization of a Gaussian and homogenous random field with a moving average representation; x_t could be predetermined or exogenous and is independent of $m(\cdot)$, and ϵ_t is a sequence of independent and identically distributed $N(0, \sigma^2)$ variates independent of both $m(\cdot)$ and x_t as well as of lagged values of x_t. The scalar parameter λ represents the contribution of the nonlinear part to the conditional mean, the vector $g \in R^k_{0,+}$ drives the curvature of the conditional mean, and the symbol $\odot$ denotes element-by-element multiplication.

Let H_k be the covariance (correlation) function of the random field $m(\cdot)$ with typical element defined as $H_k(x, z) = \mathbb{E}\left[m(x)m(z)\right]$. Hamilton [78] proved that the covariance function depends solely upon the Euclidean distance between x and z, rendering the random field isotropic. For any x and $z \in R^k$, the correlation between $m(x)$ and $m(z)$ is given by the ratio of the volume of the overlap of k-dimensional unit spheroids centered at x and z to the volume of a single k-dimensional unit spheroid. If the Euclidean distance between x and z is greater than two, the correlation between $m(x)$ and $m(z)$ will be equal to zero. The general expression of the correlation function is

$$H_k(h) = \begin{cases} G_{k-1}(h, 1)/G_{k-1}(0, 1) & \text{if } h \le 1 \\ 0 & \text{if } h > 1 \end{cases} , \tag{12}$$
$$G_k(h, r) = \int_h^r (r^2 - w^2)^{k/2} \mathrm{d}w \,,$$

where $h \equiv \frac{1}{2} d_{L_2}(x, z)$, and $d_{L_2}(x, z) \equiv \left[(x-z)'(x-z)\right]^{1/2}$ is the Euclidean distance between x and z.

Within the specification (11), Dahl and González-Rivera [33] provided alternative representations of the random field that permit the construction of Lagrange multiplier tests for neglected nonlinearity, which circumvent the problem of unidentified nuisance parameters under the null of linearity and, at the same time, they are robust to the specification of the covariance function associated with the random field. They modified the Hamilton framework in two directions. First, the random field is specified in the L_1 norm instead of the L_2 norm, and secondly they considered random fields that may not have a simple moving average representation. The advantage of the L_1 norm, which is exploited in the testing problem, is that this distance measure is a linear function of the nuisance parameters, in contrast to the L_2 norm which is a nonlinear function. Logically, Dahl and González-Rivera proceeded in an opposite fashion to Hamilton. Whereas Hamilton first proposed a moving average representation of the random field, and secondly, he derived its corresponding covariance function, Dahl and González-Rivera first proposed a covariance function, and secondly they inquire whether there is a random field associated with it. The proposed covariance function is

$$C_k(h^*) = \begin{cases} (1 - h^*)^{2k} & \text{if } h^* \le 1 \\ 0 & \text{if } h^* > 1 \end{cases} , \tag{13}$$

where $h^* \equiv \frac{1}{2} d_{L_1}(x, z) = \frac{1}{2}|x - z|'1$. The function (13) is a permissible covariance, that is, it satisfies the positive semidefiniteness condition, which is $q'C_k q \ge 0$ for all $q \ne 0_T$. Furthermore, there is a random field associated with it according to the Khinchin's theorem (1934) and Bochner's theorem (1959). The basic argument is that the class of functions which are covariance functions of homogenous random fields coincides with the class of positive semidefinite functions. Hence, (13) being a positive semidefinite function must be the covariance function of a homogenous random field.

The estimation of these models is carried out by maximum likelihood. From model (11), we can write $y \sim N(X\beta, \lambda^2 C_k + \sigma^2 I_T)$ where $y = (y_1, y_2, \ldots, y_T)'$, $X_1 = (x_1', x_2', \ldots, x_T')'$, $X = (\mathbf{1} : X_1)$, $\beta = (\beta_0, \beta_1')'$, $\epsilon =$

$(\epsilon_1, \epsilon_2, \ldots, \epsilon_T)'$ and σ^2 is the variance of ϵ_t. C_k is a generic covariance function associated with the random field, which could be equal to the Hamilton spherical covariance function in (12), or to the covariance in (13). The log-likelihood function corresponding to this model is

$$\ell(\beta, \lambda^2, g, \sigma^2) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\log|\lambda^2 C_k + \sigma^2 I_T| - \frac{1}{2}(y - X\beta)'(\lambda^2 C_k + \sigma^2 I_T)^{-1}(y - X\beta). \tag{14}$$

The flexible regression model has been applied successfully to detect nonlinearity in the quarterly growth rate of the US real GNP [34] and in the Industrial Production Index of sixteen OECD countries [33]. This technology is able to mimic the characteristics of the actual US business cycle. The cycle is dissected according to measures of duration, amplitude, cumulation and excess cumulation of the contraction and expansion phases. In contrast to Harding and Pagan [82] who find that nonlinear models are not uniformly superior to linear ones, the flexible regression model represents a clear improvement over linear models, and it seems to capture just the right shape of the expansion phase as opposed to Hamilton [76] and Durland and McCurdy [41] models, which tend to overestimate the cumulation measure in the expansion phase. It is found that the expansion phase must have at least two subphases: an aggressive early expansion after the trough, and a moderate/slow late expansion before the peak implying the existence of an inflexion point that we date approximately around one-third into the duration of the expansion phase. This shape lends support to parametric models of the growth rate that allow for three regimes [136], as opposed to models with just two regimes (contractions and expansions). For the Industrial Production Index, testing for nonlinearity within the flexible regression framework brings similar conclusions to those in Teräsvirta and Anderson [146], who propose parametric STAR models for industrial production data. However, the tests proposed in Dahl and González-Rivera [33], which have superior performance to detect smooth transition dynamics, seem to indicate that linearity cannot be rejected in the industrial production indexes of Japan, Austria, Belgium and Sweden as opposed to the findings of Teräsvirta and Anderson.

Nonlinear Factor Models

For the last ten years forecasting using a data-rich environment has been one of the most researched topic in economics and finance, see [140,141]. In this literature, factor models are used to reduce the dimension of the data but mostly they are linear models. Bai and Ng (BN) [7] introduce a nonlinear factor model with a quadratic principal component model as a special case. First consider a simple factor model

$$x_{it} = \lambda_i' F_t + e_{it}\,. \tag{15}$$

By the method of principal component, the elements of $\mathbf{f}_t$ are linear combinations of elements of $\mathbf{x}_t$. The factors are estimated by minimizing the sum of squared residuals of the linear model, $x_{it} = \lambda_i F_t + e_{it}$.

The factor model in (15) assumes a linear link function between the predictor $\mathbf{x}_t$ and the latent factors F_t. BN consider a more flexible approach by a nonlinear link function $g(\cdot)$ such that

$$g(x_{it}) = \phi_i' J_t + v_{it}\,,$$

where J_t are the common factors, and ϕ_i is the vector of factor loadings. BN consider $g(x_{it})$ to be x_{it} augmented by some or all of the unique cross-products of the elements of $\{x_{it}\}_{i=1}^N$. The second-order factor model is then $x_{it}^* = \phi_i' J_t + v_{it}$ where x_{it}^* is an $N^* \times 1$ vector. Estimation of J_t then proceeds by the usual method of principal components. BN consider $x_{it}^* = \{x_{it}\, x_{it}^2\}_{i=1}^N$ with $N^* = 2N$, which they call the SPC (squared principal components).

Once the factors are estimated, the forecasting equation for y_{t+h} would be

$$y_{t+h} = (1\hat{F}_t')\boldsymbol{\gamma} + \varepsilon_t\,.$$

The forecasting equation remains linear whatever the link function g is. An alternative way of capturing nonlinearity is to augment the forecasting equation to include functions of the factors

$$y_{t+h} = (1\hat{F}_t')\boldsymbol{\gamma} + a(\hat{F}_t) + \varepsilon_t\,,$$

where $a(\cdot)$ is nonlinear. A simple case when $a(\cdot)$ is quadratic is referred to as PC2 (squared factors) in BN.

BN note that the PC2 is conceptually distinct from SPC. While the PC2 forecasting model allows the volatility of factors estimated by linear principal components to have predictive power for y, the SPC model allows the factors to be possibly nonlinear functions of the predictors while maintaining a linear relation between the factors and y. Ludvigson and Ng [108] found that the square of the first factor estimated from a set of financial factors (i. e., volatility of the first factor) is significant in the regression model for the mean excess returns. In contrast, factors estimated from the second moment of data (i. e., volatility factors) are much weaker predictors of excess returns.

Artificial Neural Network Models

Consider an augmented single hidden layer feedforward neural network model $f(x_t, \theta)$ in which the network output y_t is determined given input x_t as

$$\begin{aligned} y_t &= f(x_t, \theta) + \varepsilon_t \\ &= x_t\beta + \sum_{j=1}^{q} \delta_j \psi(x_t\gamma_j) + \varepsilon_t \end{aligned}$$

where $\theta = (\beta'\gamma'\delta')'$, β is a conformable column vector of connection strength from the input layer to the output layer; γ_j is a conformable column vector of connection strength from the input layer to the hidden units, $j = 1, \ldots, q$; δ_j is a (scalar) connection strength from the hidden unit j to the output unit, $j = 1, \ldots, q$; and ψ is a squashing function (e. g., the logistic squasher) or a radial basis function. Input units x send signals to intermediate hidden units, then each of hidden unit produces an activation ψ that then sends signals toward the output unit. The integer q denotes the number of hidden units added to the affine (linear) network. When $q = 0$, we have a two layer *affine* network $y_t = x_t\beta + \varepsilon_t$. Hornick, Stinchcombe and White [88] show that neural network is a nonlinear flexible functional form being capable of approximating any Borel measurable function to any desired level of accuracy provided sufficiently many hidden units are available. Stinchcombe and White [138] show that this result holds for any $\psi(\cdot)$ belonging to the class of "generically comprehensively revealing" functions. These functions are "comprehensively revealing" in the sense that they can reveal arbitrary model misspecifications $\mathbb{E}(y_t|x_t) \neq f(x_t, \theta^*)$ with non-zero probability and they are "generic" in the sense that almost any choice for γ will reveal the misspecification.

We build an artificial neural network (ANN) model based on a test for neglected nonlinearity likely to have power against a range of alternatives. See White [158] and Lee, White, and Granger [99] on the neural network test and its comparison with other specification tests. The neural network test is based on a test function $h(x_t)$ chosen as the activations of 'phantom' hidden units $\psi(x_t\Gamma_j)$, $j = 1, \ldots, q$, where Γ_j are random column vectors independent of x_t. That is,

$$\mathbb{E}[\psi(x_t\Gamma_j)\varepsilon_t^*|\Gamma_j] = \mathbb{E}[\psi(x_t\Gamma_j)\varepsilon_t^*] = 0 \quad j = 1, \ldots, q\,, \tag{16}$$

under H_0, so that

$$\mathbb{E}(\Psi_t\varepsilon_t^*) = 0\,, \tag{17}$$

where $\Psi_t = (\psi(x_t\Gamma_1), \ldots, \psi(x_t\Gamma_q))'$ is a phantom hidden unit activation vector. Evidence of correlation of ε_t^* with Ψ_t is evidence against the null hypothesis that y_t is linear in mean. If correlation exists, augmenting the linear network by including an additional hidden unit with activations $\psi(x_t\Gamma_j)$ would permit an improvement in network performance. Thus the tests are based on sample correlation of affine network errors with phantom hidden unit activations,

$$n^{-1}\sum_{t=1}^{n}\Psi_t\hat{\varepsilon}_t = n^{-1}\sum_{t=1}^{n}\Psi_t(y_t - x_t\hat{\beta})\,. \tag{18}$$

Under suitable regularity conditions it follows from the central limit theorem that $n^{-1/2}\sum_{t=1}^{n}\Psi_t\hat{\varepsilon}_t \xrightarrow{d} N(0, W^*)$ as $n \to \infty$, and if one has a consistent estimator for its asymptotic covariance matrix, say $\hat{W}_n$, then an asymptotic chi-square statistic can be formed as

$$\left(n^{-1/2}\sum_{t=1}^{n}\Psi_t\hat{\varepsilon}_t\right)'\hat{W}_n^{-1}\left(n^{-1/2}\sum_{t=1}^{n}\Psi_t\hat{\varepsilon}_t\right) \xrightarrow{d} \chi^2(q)\,. \tag{19}$$

Elements of Ψ_t tend to be collinear with X_t and with themselves. Thus LWG conduct a test on $q^* < q$ principal components of Ψ_t not collinear with x_t, denoted Ψ_t^*. This test is to determine whether or not there exists some advantage to be gained by adding hidden units to the affine network. We can estimate $\hat{W}_n$ robust to the conditional heteroskedasticity, or we may use with the empirical null distribution of the statistic computed by a bootstrap procedure that is robust to the conditional heteroskedasticity, e. g., wild bootstrap.

Estimation of an ANN model may be tedious and sometimes results in unreliable estimates. Recently, White [161] proposes a simple algorithm called QuickNet, a form of "relaxed greedy algorithm" because QuickNet searches for a single best additional hidden unit based on a sequence of OLS regressions, that may be analogous to the least angular regressions (LARS) of Efron, Hastie, Johnstone, and Tibshirani [42]. The simplicity of the QuickNet algorithm achieves the benefits of using a forecasting model that is nonlinear in the predictors while mitigating the other computational challenges to the use of nonlinear forecasting methods. See White [161], Section 5, for more details on QuickNet, and for other issues of controlling for overfit and the selection of the random parameter vectors Γ_j independent of x_t.

Campbell, Lo, and MacKinlay [22], Section 12.4, provide a review of these models. White [161] reviews

the research frontier in ANN models. Trippi and Turban [150] review the applications of ANNs to finance and investment.

Functional Coefficient Models

A functional coefficient model is introduced by Cai, Fan, and Yao [24] (CFY), with time-varying and state-dependent coefficients. It can be viewed as a special case of Priestley's [127] state-dependent model, but it includes the models of Tong [149], Chen and Tsay [26] and regime-switching models as special cases. Let $\{(y_t, s_t)'\}_{t=1}^n$ be a stationary process, where y_t and s_t are scalar variables. Also let $X_t \equiv (1, y_{t-1}, \ldots, y_{t-d})'$. We assume

$$\mathbb{E}(y_t|\mathcal{F}_{t-1}) = a_0(s_t) + \sum_{j=1}^{d} a_j(s_t) y_{t-j} ,$$

where the $\{a_j(s_t)\}$ are the autoregressive coefficients depending on s_t, which may be chosen as a function of X_t or something else. Intuitively, the functional coefficient model is an AR process with time-varying autoregressive coefficients. The coefficient functions $\{a_j(s_t)\}$ can be estimated by local linear regression. At each point s, we approximate $a_j(s_t)$ locally by a linear function $a_j(s_t) \approx a_j + b_j(s_t - s)$, $j = 0, 1, \ldots, d$, for s_t near s, where a_j and b_j are constants. The local linear estimator at point s is then given by $\hat{a}_j(s) = \hat{a}_j$, where $\{(\hat{a}_j, \hat{b}_j)\}_{j=0}^d$ minimizes the sum of local weighted squares $\sum_{t=1}^n [y_t - \mathbb{E}(y_t|\mathcal{F}_{t-1})]^2 K_h(s_t - s)$, with $K_h(\cdot) \equiv K(\cdot/h)/h$ for a given kernel function $K(\cdot)$ and bandwidth $h \equiv h_n \to 0$ as $n \to \infty$. CFY [24], p. 944, suggest to select h using a modified multi-fold "leave-one-out-type" cross-validation based on MSFE.

It is important to choose an appropriate smooth variable s_t. Knowledge on data or economic theory may be helpful. When no prior information is available, s_t may be chosen as a function of explanatory vector X_t or using such data-driven methods as AIC and cross-validation. See Fan, Yao and Cai [52] for further discussion on the choice of s_t. For exchange rate changes, Hong and Lee [85] choose s_t as the difference between the exchange rate at time $t-1$ and the moving average of the most recent L periods of exchange rates at time $t-1$. The moving average is a proxy for the local trend at time $t-1$. Intuitively, this choice of s_t is expected to reveal useful information on the direction of changes.

To justify the use of the functional coefficient model, CFY [24] suggest a goodness-of-fit test for an AR(d) model against a functional coefficient model. The null hypothesis of AR(d) can be stated as

$$\mathbb{H}_0 : a_j(s_t) = \beta_j, \quad j = 0, 1, \ldots, d ,$$

where β_j is the autoregressive coefficient in AR(d). Under $\mathbb{H}_0$, $\{y_t\}$ is linear in mean conditional on X_t. Under the alternative to $\mathbb{H}_0$, the autoregressive coefficients depend on s_t and the AR(d) model suffers from "neglected nonlinearity". To test $\mathbb{H}_0$, CFY compares the residual sum of squares (RSS) under $\mathbb{H}_0$

$$RSS_0 \equiv \sum_{t=1}^n \hat{\varepsilon}_t^2 = \sum_{t=1}^n \left[Y_t - \hat{\beta}_0 - \sum_{j=1}^d \hat{\beta}_j Y_{t-j} \right]^2$$

with the RSS under the alternative

$$RSS_1 \equiv \sum_{t=1}^n \tilde{\varepsilon}_t^2 = \sum_{t=1}^n \left[Y_t - \hat{a}_0(s_t) - \sum_{j=1}^d \hat{a}_j(s_t) Y_{t-j} \right]^2 .$$

The test statistic is $T_n = (RSS_0 - RSS_1)/RSS_1$. We reject $\mathbb{H}_0$ for large values of T_n. CFY suggest the following bootstrap method to obtain the p-value of T_n: (i) generate the bootstrap residuals $\{\varepsilon_t^b\}_{t=1}^n$ from the centered residuals $\tilde{\varepsilon}_t - \bar{\varepsilon}$ where $\bar{\varepsilon} \equiv n^{-1} \sum_{t=1}^n \tilde{\varepsilon}_t$ and define $y_t^b \equiv X_t' \hat{\beta} + \varepsilon_t^b$, where $\hat{\beta}$ is the OLS estimator for AR(d); (ii) calculate the bootstrap statistic T_n^b using the bootstrap sample $\{y_t^b, X_t', s_t\}_{t=1}^n$; (iii) repeat steps (i) and (ii) B times ($b = 1, \ldots, B$) and approximate the bootstrap p-value of T_n by $B^{-1} \sum_{b=1}^B \mathbf{1}(T_n^b \geq T_n)$. See Hong and Lee [85] for empirical application of the functional coefficient model to forecasting foreign exchange rates.

Nonparametric Regression

Let $\{y_t, x_t\}$, $t = 1, \ldots, n$, be stochastic processes, where y_t is a scalar and $x_t = (x_{t1}, \ldots, x_{tk})$ is a $1 \times k$ vector which may contain the lagged values of y_t. Consider the regression model

$$y_t = m(x_t) + u_t$$

where $m(x_t) = \mathbb{E}\left(y_t|x_t\right)$ is the true but unknown regression function and u_t is the error term such that $\mathbb{E}(u_t|x_t) = 0$.

If $m(x_t) = g(x_t, \delta)$ is a correctly specified family of parametric regression functions then $y_t = g(x_t, \delta) + u_t$ is a correct model and, in this case, one can construct a consistent least squares (LS) estimator of $m(x_t)$ given by $g(x_t, \hat{\delta})$, where $\hat{\delta}$ is the LS estimator of the parameter δ.

In general, if the parametric regression $g(x_t, \delta)$ is incorrect or the form of $m(x_t)$ is unknown then $g(x_t, \hat{\delta})$ may not be a consistent estimator of $m(x_t)$. For this case, an alternative approach to estimate the unknown $m(x_t)$ is to use the consistent nonparametric kernel regression estimator which is essentially a local constant LS (LCLS) es-

timator. To obtain this estimator take a Taylor series expansion of $m(x_t)$ around x so that

$$\begin{aligned} y_t &= m(x_t) + u_t \\ &= m(x) + e_t \end{aligned}$$

where $e_t = (x_t - x)m^{(1)}(x) + \frac{1}{2}(x_t - x)^2 m^{(2)}(x) + \cdots + u_t$ and $m^{(s)}(x)$ represents the sth derivative of $m(x)$ at $x_t = x$. The LCLS estimator can then be derived by minimizing

$$\sum_{t=1}^{n} e_t^2 K_{tx} = \sum_{t=1}^{n} (y_t - m(x))^2 K_{tx}$$

with respect to constant $m(x)$, where $K_{tx} = K\left(\frac{x_t - x}{h}\right)$ is a decreasing function of the distances of the regressor vector x_t from the point $x = (x_1, \ldots, x_k)$, and $h \to 0$ as $n \to \infty$ is the window width (smoothing parameter) which determines how rapidly the weights decrease as the distance of x_t from x increases. The LCLS estimator so estimated is

$$\hat{m}(x) = \frac{\sum_{t=1}^{n} y_t K_{tx}}{\sum_{t=1}^{n} K_{tx}} = (\mathbf{i}'\mathbf{K}(x)\mathbf{i})^{-1}\, \mathbf{i}'\mathbf{K}(x)\mathbf{y}$$

where $\mathbf{K}(x)$ is the $n \times n$ diagonal matrix with the diagonal elements K_{tx} $(t = 1, \ldots, n)$, $\mathbf{i}$ is an $n \times 1$ column vector of unit elements, and $\mathbf{y}$ is an $n \times 1$ vector with elements y_t $(t = 1, \ldots, n)$. The estimator $\hat{m}(x)$ is due to Nadaraya [118] and Watson [155] (NW) who derived this in an alternative way. Generally $\hat{m}(x)$ is calculated at the data points x_t, in which case we can write the leave-one out estimator as

$$\hat{m}(x) = \frac{\sum_{t'=1, t' \neq t}^{n} y_{t'} K_{t't}}{\sum_{t'=1, t' \neq t}^{n} K_{t't}},$$

where $K_{t't} = K\frac{x_{t'} - x_t}{h}$. The assumption that $h \to 0$ as $n \to \infty$ gives $x_t - x = O(h) \to 0$ and hence $\mathbb{E}e_t \to 0$ as $n \to \infty$. Thus the estimator $\hat{m}(x)$ will be consistent under certain smoothing conditions on h, K, and $m(x)$. In small samples however $\mathbb{E}e_t \neq 0$ so $\hat{m}(x)$ will be a biased estimator, see [122] for details on asymptotic and small sample properties.

An estimator which has a better small sample bias and hence the mean square error (MSE) behavior is the local linear LS (LLLS) estimator. In the LLLS estimator we take a first order Taylor-Series expansion of $m(x_t)$ around x so that

$$\begin{aligned} y_t &= m(x_t) + u_t = m(x) + (x_t - x)m^{(1)}(x) + v_t \\ &= \alpha(x) + x_t\beta(x) + v_t \\ &= X_t\delta(x) + v_t \end{aligned}$$

where $X_t = (1\ x_t)$ and $\delta(x) = [\alpha(x)\ \beta(x)']'$ with $\alpha(x) = m(x) - x\beta(x)$ and $\beta(x) = m^{(1)}(x)$. The LLLS estimator of $\delta(x)$ is then obtained by minimizing

$$\sum_{t=1}^{n} v_t^2 K_{tx} = \sum_{t=1}^{n} (y_t - X_t\delta(x))^2 K_{tx}$$

sand it is given by

$$\tilde{\delta}(x) = (\mathbf{X}'\mathbf{K}(x)\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}(x)\mathbf{y}\,. \tag{20}$$

where $\mathbf{X}$ is an $n \times (k+1)$ matrix with the tth row X_t $(t = 1, \ldots, n)$.

The LLLS estimator of $\alpha(x)$ and $\beta(x)$ can be calculated as $\tilde{\alpha}(x) = (1\ 0)\tilde{\delta}(x)$ and $\tilde{\beta}(x) = (0\ 1)\tilde{\delta}(x)$. This gives

$$\tilde{m}(x) = (1\ x)\tilde{\delta}(x) = \tilde{\alpha}(x) + x\tilde{\beta}(x)\,.$$

Obviously when $X = \mathbf{i}$, $\tilde{\delta}(x)$ reduces to the NW's LCLS estimator $\hat{m}(x)$. An extension of the LLLS is the local polynomial LS (LPLS) estimators, see [50].

In fact one can obtain the local estimators of a general nonlinear model $g(x_t, \delta)$ by minimizing

$$\sum_{t=1}^{n} [y_t - g(x_t, \delta(x))]^2 K_{tx}$$

with respect to $\delta(x)$. For $g(x_t, \delta(x)) = X_t\delta(x)$ we get the LLLS in (20). Further when $h = \infty$, $K_{tx} = K(0)$ is a constant so that the minimization of $K(0)\sum[y_t - g(x_t, \delta(x))]^2$ is the same as the minimization of $\sum[y_t - g(x_t, \delta)]^2$, that is the local LS becomes the global LS estimator $\hat{\delta}$.

The LLLS estimator in (20) can also be interpreted as the estimator of the functional coefficient (varying coefficient) linear regression model

$$\begin{aligned} y_t &= m(x_t) + u_t \\ &= X_t\delta(x_t) + u_t \end{aligned}$$

where $\delta(x_t)$ is approximated locally by a constant $\delta(x_t) \simeq \delta(x)$. The minimization of $\sum u_t^2 K_{tx}$ with respect to $\delta(x)$ then gives the LLLS estimator in (20), which can be interpreted as the LC varying coefficient estimator. An extension of this is to consider the linear approximation $\delta(x_t) \simeq \delta(x) + D(x)(x_t - x)'$ where $D(x) = \frac{\partial\delta(x_t)}{\partial x_t'}$ evaluated at $x_t = x$. In this case

$$\begin{aligned} y_t &= m(x_t) + u_t = X_t\delta(x_t) + u_t \\ &\simeq X_t\delta(x) + X_t D(x)(x_t - x)' + u_t \\ &= X_t\delta(x) + [(x_t - x) \otimes X_t]vecD(x) + u_t \\ &= X_t^x\delta^x(x) + u_t \end{aligned}$$

where $X_t^x = [X_t \;\; (x_t - x) \otimes X_t]$ and $\delta^x(x) = [\delta(x)' \; (vecD(x))']'$. The LL varying coefficient estimator of $\delta^x(x)$ can then be obtained by minimizing

$$\sum_{t=1}^{n} [y_t - X_t^x \delta^x(x)]^2 K_{tx}$$

with respect to $\delta^x(x)$ as

$$\dot{\delta}^x(x) = (\mathbf{X}^{x\prime}\mathbf{K}(x)\mathbf{X}^x)^{-1}\mathbf{X}^{x\prime}\mathbf{K}(x)\mathbf{y}\,. \tag{21}$$

From this $\dot{\delta}(x) = (\mathbf{I}\;0)\dot{\delta}^x(x)$, and hence

$$\dot{m}(x) = (1\;x\;0)\dot{\delta}^x(x) = (1\;x)\dot{\delta}(x)\,.$$

The above idea can be extended to the situations where $\xi_t = (x_t\;z_t)$ such that

$$\mathbb{E}(y_t|\xi_t) = m(\xi_t) = m(x_t, z_t) = X_t\delta(z_t)\,,$$

where the coefficients are varying with respect to only a subset of ξ_t; z_t is $1 \times l$ and ξ_t is $1 \times p$, $p = k + l$. Examples of these include functional coefficient autoregressive models of Chen and Tsay [26] and CFY [24], random coefficient models of Raj and Ullah [128], smooth transition autoregressive models of Granger and Teräsvirta [72], and threshold autoregressive models of Tong [149].

To estimate $\delta(z_t)$ we can again do a local constant approximation $\delta(z_t) \simeq \delta(z)$ and then minimize $\sum[y_t - X_t\delta(z)]^2 K_{tz}$ with respect to $\delta(z)$, where $K_{tz} = K(\frac{z_t - z}{h})$. This gives the LC varying coefficient estimator

$$\tilde{\delta}(z) = (\mathbf{X}'\mathbf{K}(z)\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}(z)\mathbf{y} \tag{22}$$

where $\mathbf{K}(z)$ is a diagonal matrix of $K_{tz}, t = 1, \ldots, n$. When $z = x$, (22) reduces to the LLLS estimator $\tilde{\delta}(x)$ in (20).

CFY [24] consider a local linear approximation $\delta(z_t) \simeq \delta(z) + D(z)(z_t - z)'$. The LL varying coefficient estimator of CFY is then obtained by minimizing

$$\begin{aligned}
&\sum_{t=1}^{n} [y_t - X_t\delta(z_t)]^2 K_{tz} \\
&\quad = \sum_{t=1}^{n} [y_t - X_t\delta(z) - [(z_t - z) \otimes X_t] vecD(z)]^2 K_{tz} \\
&\quad = \sum_{t=1}^{n} [y_t - X_t^z \delta^z(z)]^2 K_{tz}
\end{aligned}$$

with respect to $\delta^z(z) = [\delta(z)' \;\; (vecD(z))']'$ where $X_t^z = [X_t \;\; (z_t - z) \otimes X_t]$. This gives

$$\ddot{\delta}^z(z) = (\mathbf{X}^{z\prime}\mathbf{K}(z)\mathbf{X}^z)^{-1}\mathbf{X}^{z\prime}\mathbf{K}(z)\mathbf{y}\,, \tag{23}$$

and $\ddot{\delta}(z) = (\mathbf{I}\;0)\ddot{\delta}^z(z)$. Hence

$$\ddot{m}(\xi) = (1\;x\;0)\ddot{\delta}^z(z) = (1\;x)\ddot{\delta}(z)\,.$$

For the asymptotic properties of these varying coefficient estimators, see CFY [24]. When $z = x$, (23) reduces to the LL varying coefficient estimator $\dot{\delta}^x(x)$ in (21). See Lee and Ullah [98] for more discussion of these models and issues of testing nonlinearity.

Regime Switching Autoregressive Model Between Unit Root and Stationary Root

To avoid the usual dichotomy between unit-root nonstationarity and stationarity, we may consider models that permit two regimes of unit root nonstationarity and stationarity.

One model is the Innovation Regime-Switching (IRS) model of Kuan, Huang, and Tsay [96]. Intuitively, it may be implausible to believe that all random shocks exert only one effect (permanent or transitory) on future financial asset prices in a long time span. This intuition underpins the models that allow for breaks, stochastic unit root, or regime switching. As an alternative, Kuan, Huang, and Tsay [96] propose the IRS model that permits the random shock in each period to be permanent or transitory, depending on a switching mechanism, and hence admits distinct dynamics (unit-root nonstationarity or stationarity) in different periods. Under the IRS framework, standard unit-root models and stationarity models are just two extreme cases. By applying the IRS model to real exchange rate, they circumvent the difficulties arising from unit-root (or stationarity) testing. They allow the data to speak for themselves, rather than putting them in the straitjacket of unit-root nonstationarity or stationarity. Huang and Kuan [90] re-examine long-run PPP based on the IRS model and their empirical study on US/UK real exchange rates shows that there are both temporary and permanent influences on the real exchange rate such that approximately 42% of the shocks in the long run are more likely to have a permanent effect. They also found that transitory shocks dominate in the fixed-rate regimes, yet permanent shocks play a more important role during the floating regimes. Thus, the long-run PPP is rejected due to the presence of a significant amount of permanent shocks, but there are still long periods of time in which the deviations from long-run PPP are only transitory.

Another model is a threshold unit root (TUR) model or threshold integrated moving average (TIMA) model of Gonzalo and Martíneza [65]. Based on this model they examine whether large and small shocks have different long-

run effects, as well as whether one of them is purely transitory. They develop a new nonlinear permanent – transitory decomposition, that is applied to US stock prices to analyze the quality of the stock market.

Comparison of these two models with the linear autoregressive model with a unit root or a stationary AR model for the out-of-sample forecasting remains to be examined empirically.

Bagging Nonlinear Forecasts

To improve on unstable forecasts, bootstrap aggregating or bagging is introduced by Breiman [19]. Lee and Yang [100] show how bagging works for binary and quantile predictions. Lee and Yang [100] attributed part of the success of the bagging predictors to the small sample estimation uncertainties. Therefore, a question that may arise is that whether the good performance of bagging predictors critically depends on algorithms we employ in nonlinear estimation.

They find that bagging improves the forecasting performance of predictors on highly nonlinear regression models – e. g., artificial neural network models, especially when the sample size is limited. It is usually hard to choose the number of hidden nodes and the number of inputs (or lags), and to estimate the large number of parameters in an ANN model. Therefore, a neural network model generate poor predictions in a small sample. In such cases, bagging can do a valuable job to improve the forecasting performance as shown in [100], confirming the result of Breiman [20]. A bagging predictor is a combined predictor formed over a set of training sets to smooth out the "instability" caused by parameter estimation uncertainty and model uncertainty. A predictor is said to be "unstable" if a small change in the training set will lead to a significant change in the predictor [20].

As bagging would be valuable in nonlinear forecasting, in this section, we will show how a bagging predictor may improve the predicting performance of its underlying predictor. Let

$$\mathcal{D}_t \equiv \{(Y_s, \mathbf{X}_{s-1})\}_{s=t-R+1}^{t} \quad (t = R, \ldots, T)$$

be a training set at time t and let $\varphi(\mathbf{X}_t, \mathcal{D}_t)$ be a forecast of Y_{t+1} or of the binary variable $G_{t+1} \equiv \mathbf{1}(Y_{t+1} \geq 0)$ using this training set $\mathcal{D}_t$ and the explanatory variable vector $\mathbf{X}_t$. The optimal forecast $\varphi(\mathbf{X}_t, \mathcal{D}_t)$ for Y_{t+1} will be the conditional mean of Y_{t+1} given $\mathbf{X}_t$ under the squared error loss function, or the conditional quantile of Y_{t+1} on $\mathbf{X}_t$ if the loss is a tick function. Below we also consider the binary forecast for $G_{t+1} \equiv \mathbf{1}(Y_{t+1} \geq 0)$.

Suppose each training set $\mathcal{D}_t$ consists of R observations generated from the underlying probability distribution $\mathbf{P}$. The forecast $\{\varphi(\mathbf{X}_t, \mathcal{D}_t)\}_{t=R}^{T}$ can be improved if more training sets were able to be generated from $\mathbf{P}$ and the forecast can be formed from averaging the multiple forecasts obtained from the multiple training sets. Ideally, if $\mathbf{P}$ were known and multiple training sets $\mathcal{D}_t^{(j)}$ $(j = 1, \ldots, J)$ may be drawn from $\mathbf{P}$, an ensemble aggregating predictor $\varphi_A(\mathbf{X}_t)$ can be constructed by the weighted averaging of $\varphi(\mathbf{X}_t, \mathcal{D}_t^{(j)})$ over j, i. e.,

$$\varphi_A(\mathbf{X}_t) \equiv \mathbb{E}_{\mathcal{D}_t}\varphi(\mathbf{X}_t, \mathcal{D}_t) \equiv \sum_{j=1}^{J} w_{j,t}\varphi(\mathbf{X}_t, \mathcal{D}_t^{(j)}) ,$$

where $\mathbb{E}_{\mathcal{D}_t}(\cdot)$ denotes the expectation over $\mathbf{P}$, $w_{j,t}$ is the weight function with $\sum_{j=1}^{J} w_{j,t} = 1$, and the subscript A in φ_A denotes "aggregation".

Lee and Yang [100] show that the ensemble aggregating predictor $\varphi_A(X_t)$ has not a larger expected loss than the original predictor $\varphi(X_t, \mathcal{D}_t)$. For any convex loss function $c(\cdot)$ on the forecast error z_{t+1}, we will have

$$\mathbb{E}_{\mathcal{D}_t, Y_{t+1}, \mathbf{X}_t} c(z_{t+1}) \geq \mathbb{E}_{Y_{t+1}, \mathbf{X}_t} c(\mathbb{E}_{\mathcal{D}_t}(z_{t+1})),$$

where $\mathbb{E}_{\mathcal{D}_t}(z_{t+1})$ is the aggregating forecast error, and $\mathbb{E}_{\mathcal{D}_t, Y_{t+1}, \mathbf{X}_t}(\cdot) \equiv \mathbb{E}_{\mathbf{X}_t}[\mathbb{E}_{Y_{t+1}|\mathbf{X}_t}\{\mathbb{E}_{\mathcal{D}_t}(\cdot)|X_t\}]$ denotes the expectation $\mathbb{E}_{\mathcal{D}_t}(\cdot)$ taken over $\mathbf{P}$ (i. e., averaging over the multiple training sets generated from $\mathbf{P}$), then taking an expectation of Y_{t+1} conditioning on X_t, and then taking an expectation of X_t. Similarly we define the notation $\mathbb{E}_{Y_{t+1}, \mathbf{X}_t}(\cdot) \equiv \mathbb{E}_{\mathbf{X}_t}[\mathbb{E}_{Y_{t+1}|\mathbf{X}_t}(\cdot)|X_t]$. Therefore, the aggregating predictor will always have no larger expected cost than the original predictor for a convex loss function $\varphi(X_t, D_t)$. The examples of the convex loss function includes the squared error loss and a tick (or check) loss $\rho_\alpha(z) \equiv [\alpha - \mathbf{1}(z < 0)]z$.

How much this aggregating predictor can improve depends on the distance between $\mathbb{E}_{\mathcal{D}_t, Y_{t+1}, \mathbf{X}_t} c(z_{t+1})$ and $\mathbb{E}_{Y_{t+1}, \mathbf{X}_t} c(\mathbb{E}_{\mathcal{D}_t}(z_{t+1}))$. We can define this distance by $\Delta \equiv \mathbb{E}_{\mathcal{D}_t, Y_{t+1}, \mathbf{X}_t} c(z_{t+1}) - \mathbb{E}_{Y_{t+1}, \mathbf{X}_t} c(\mathbb{E}_{\mathcal{D}_t}(z_{t+1}))$. Therefore, the effectiveness of the aggregating predictor depends on the *convexity* of the cost function. The more convex is the cost function, the more effective this aggregating predictor can be. If the loss function is the squared error loss, then it can be shown that $\Delta = \mathbb{V}_{\mathcal{D}_t}\left[\varphi(\mathbf{X}_t, \mathcal{D}_t)\right]$ is the variance of the predictor, which measures the "instability" of the predictor. See Lee and Yang [100], Proposition 1, and Breiman [20]. If the loss is the tick function, the effectiveness of bagging is also different for different quantile predictions: bagging works better for tail-quantile predictions than for mid-quantile predictions.

In practice, however, $\mathbf{P}$ is not known. In that case we may estimate $\mathbf{P}$ by its empirical distribution, $\hat{\mathbf{P}}(\mathcal{D}_t)$, for a given $\mathcal{D}_t$. Then, from the empirical distribution $\hat{\mathbf{P}}(\mathcal{D}_t)$, multiple training sets may be drawn by the bootstrap method. Bagging predictors, $\varphi^B(\mathbf{X}_t, \mathcal{D}_t^*)$, can then be computed by taking weighted average of the predictors trained over a set of bootstrap training sets. More specifically, the bagging predictor $\varphi^B(\mathbf{X}_t, \mathcal{D}_t^*)$ can be obtained in the following steps:

1. Given a training set of data at time t, $\mathcal{D}_t \equiv \{(Y_s, \mathbf{X}_{s-1})\}_{s=t-R+1}^{t}$, construct the jth bootstrap sample $\mathcal{D}_t^{*(j)} \equiv \{(Y_s^{*(j)}, \mathbf{X}_{s-1}^{*(j)})\}_{s=t-R+1}^{t}$, $j = 1, \ldots, J$, according to the empirical distribution of $\hat{\mathbf{P}}(\mathcal{D}_t)$ of $\mathcal{D}_t$.
2. Train the model (estimate parameters) from the jth bootstrapped sample $\mathcal{D}_t^{*(j)}$.
3. Compute the bootstrap predictor $\varphi^{*(j)}(\mathbf{X}_t, \mathcal{D}_t^{*(j)})$ from the jth bootstrapped sample $\mathcal{D}_t^{*(j)}$.
4. Finally, for mean and quantile forecast, the bagging predictor $\varphi^B(\mathbf{X}_t, \mathcal{D}_t^*)$ can be constructed by averaging over J bootstrap predictors

$$\varphi^B(\mathbf{X}_t, \mathcal{D}_t^*) \equiv \sum_{j=1}^{J} \hat{w}_{j,t} \varphi^{*(j)}(\mathbf{X}_t, \mathcal{D}_t^{*(j)}) ;$$

and for binary forecast, the bagging binary predictor $\varphi^B(\mathbf{X}_t, \mathcal{D}_t^*)$ can be constructed by majority voting over J bootstrap predictors:

$$\varphi^B(\mathbf{X}_t, \mathcal{D}_t^*) \equiv \mathbf{1}\left(\sum_{j=1}^{J} \hat{w}_{j,t} \varphi^{*(j)}(\mathbf{X}_t, \mathcal{D}_t^{*(j)}) > 1/2\right)$$

with $\sum_{j=1}^{J} \hat{w}_{j,t} = 1$ in both cases.

One concern of applying bagging to time series is whether a bootstrap can provide a sound simulation sample for dependent data, for which the bootstrap is required to be consistent. It has been shown that some bootstrap procedure (such as moving block bootstrap) can provide consistent densities for moment estimators and quantile estimators. See, e. g., Fitzenberger [54].

Nonlinear Forecasting Models for the Conditional Variance

Nonlinear Parametric Models for Volatility

Volatility models are of paramount importance in financial economics. Issues such as portfolio allocation, option pricing, risk management, and generally any decision making under uncertainty rely on the understanding and forecasting of volatility. This is one of the most active ares of research in time series econometrics. Important surveys as in Bollerslev, Chou, and Kroner [15], Bera and Higgins [13], Bollerslev, Engle, and Nelson [16], Poon and Granger [125], and Bauwens, Laurent, and Rombouts [12] attest to the variety of issues in volatility research. The motivation for the introduction of the first generation of volatility models namely the ARCH models of Engle [44] was to account for clusters of activity and fat-tail behavior of financial data. Subsequent models accounted for more complex issues. Among others and without being exclusive, we should mention issues related to asymmetric responses of volatility to news, probability distribution of the standardized innovations, i.i.d. behavior of the standardized innovation, persistence of the volatility process, linkages with continuous time models, intraday data and unevenly spaced observations, seasonality and noise in intraday data. The consequence of this research agenda has been a vast array of specifications for the volatility process.

Suppose that the return series $\{y_t\}_{t=1}^{T+1}$ of a financial asset follows the stochastic process $y_{t+1} = \mu_{t+1} + \varepsilon_{t+1}$, where $\mathbb{E}(y_{t+1}|\mathcal{F}_t) = \mu_{t+1}(\theta)$ and $\mathbb{E}(\varepsilon_{t+1}^2|\mathcal{F}_t) = \sigma_{t+1}^2(\theta)$ given the information set $\mathcal{F}_t$ (σ-field) at time t. Let $z_{t+1} \equiv \varepsilon_{t+1}/\sigma_{t+1}$ have the conditional normal distribution with zero conditional mean and unit conditional variance. Volatility models can be classified in three categories: MA family, ARCH family, and stochastic volatility (SV) family.

The simplest method to forecast volatility is to calculate a historical moving average variance, denoted as MA(m), or an exponential weighted moving average (EWMA):

MA(m)	$\sigma_t^2 = \frac{1}{m}\sum_{j=1}^{m}(y_{t-j} - \hat{\mu}_t^m)^2, \quad \hat{\mu}_t^m = \frac{1}{m}\sum_{j=1}^{m} y_{t-j}$
EWMA	$\sigma_t^2 = (1-\lambda)\sum_{j=1}^{t-1}\lambda^{j-1}(y_{t-j} - \hat{\mu}_t)^2,$ $\hat{\mu}_t = \frac{1}{t-1}\sum_{j=1}^{t-1} y_{t-j}$

In the EWMA specification, a common practice is to fix the λ parameter, for instance $\lambda = 0.94$ [129]. For these two MA family models, there are not parameters to estimate.

Second, the ARCH family is very extensive with many variations on the original model ARCH(p) of Engle [44]. Some representative models are: GARCH model of Bollerslev [14]; Threshold GARCH (T-GARCH) of Glosten et al. [60]; Exponential GARCH (E-GARCH) of Nelson [120]; quadratic GARCH models (Q-GARCH) as in Sentana [135]; Absolute GARCH (ABS-GARCH) of

Taylor [143] and Schwert [134] and Smooth Transition GARCH (ST-GARCH) of González-Rivera [61].

ARCH(p)	$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2$
GARCH	$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha\varepsilon_{t-1}^2$
I-GARCH	$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha\varepsilon_{t-1}^2,\ \alpha + \beta = 1$
T-GARCH	$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha\varepsilon_{t-1}^2 + \gamma\varepsilon_{t-1}^2 \mathbf{1}(\varepsilon_{t-1} \geq 0)$
ST-GARCH	$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha\varepsilon_{t-1}^2 + \gamma\varepsilon_{t-1}^2 F(\varepsilon_{t-1}, \delta)$ with $F(\varepsilon_{t-1}, \delta) = [1 + \exp(\delta\varepsilon_{t-1})]^{-1} - 0.5$
E-GARCH	$\ln\sigma_t^2 = \omega + \beta \ln\sigma_{t-1}^2 + \alpha[\lvert z_{t-1}\rvert - cz_{t-1}]$
Q-GARCH	$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha(\varepsilon_{t-1} + \gamma)^2$
ABS-GARCH	$\sigma_t = \omega + \beta\sigma_{t-1} + \alpha\lvert\varepsilon_{t-1}\rvert$

The EWMA specification can be viewed as an integrated GARCH model with $\omega = 0, \alpha = \lambda$, and $\beta = 1 - \lambda$. In the T-GARCH model, the parameter γ allows for possible asymmetric effects of positive and negative innovations. In Q-GARCH models, the parameter γ measures the extent of the asymmetry in the news impact curve. For the ST-GARCH model, the parameter γ measures the asymmetric effect of positive and negative shocks, and the parameter $\delta > 0$ measures the smoothness of the transition between regimes, with a higher value of δ making ST-GARCH closer to T-GARCH.

Third, the stationary SV model of Taylor [143] with η_t is i.i.d. N $(0, \sigma_\eta^2)$ and ξ_t is i.i.d. N$(0, \pi^2/2)$ is a representative member of the SV family.

SV	$\sigma_t^2 = \exp(0.5h_t), \quad \ln(y_t^2) = -1.27 + h_t + \xi_t,$ $h_t = \gamma + \phi h_{t-1} + \eta_t.$

With so many models, the natural question becomes which one to choose. There is not a universal answer to this question. The best model depends upon the objectives of the user. Thus, given an objective function, we search for the model(s) with the best predictive ability controlling for possible biases due to "data snooping" [105]. To compare the relative performance of volatility models, it is customary to choose either a statistical loss function or an economic loss function.

The preferred statistical loss functions are based on moments of forecast errors (mean-error, mean-squared error, mean absolute error, etc.). The best model will minimize a function of the forecast errors. The volatility forecast is often compared to a measure of realized volatility. With financial data, the common practice has been to take squared returns as a measure of realized volatility. However, this practice is questionable. Andersen and Bollerslev [2] argued that this measure is a noisy estimate, and proposed the use of the intra-day (at each five minutes interval) squared returns to calculate the daily realized volatility. This measure requires intra-day data, which is subject to the variation introduced by the bid-ask spread and the irregular spacing of the price quotes.

Some other authors have evaluated the performance of volatility models with criteria based on economic loss functions. For example, West, Edison, and Cho [157] considered the problem of portfolio allocation based on models that maximize the utility function of the investor. Engle, Kane, and Noh [46] and Noh, Engle, and Kane [121] considered different volatility forecasts to maximize the trading profits in buying/selling options. Lopez [107] considered probability scoring rules that were tailored to a forecast user's decision problem and confirmed that the choice of loss function directly affected the forecast evaluation of different models. Brooks and Persand [21] evaluated volatility forecasting in a financial risk management setting in terms of Value-at-Risk (VaR). The common feature to these branches of the volatility literature is that none of these has controlled for forecast dependence across models and the inherent biases due to data-snooping.

Controlling for model dependence [160], González-Rivera, Lee, and Mishra [62] evaluate fifteen volatility models for the daily returns to the SP500 index according to their out-of-sample forecasting ability. The forecast evaluation is based, among others, on two economic loss functions: an option pricing formula and a utility function; and a statistical loss function: a goodness-of-fit based on a Value-at-Risk (VaR) calculation. For option pricing, volatility is the only component that is not observable and it needs to be estimated. The loss function assess the difference between the actual price of a call option and the estimated price, which is a function of the estimated volatility of the stock. The second economic loss function refers to the problem of wealth allocation. An investor wishes to maximize her utility allocating wealth between a risky asset and a risk-free asset. The loss function assesses the performance of the volatility estimates according to the level of utility they generate. The statistical function based on the goodness-of-fit of a VaR calculation is important for risk management. The main objective of VaR is to calculate extreme losses within a given probability of occurrence, and the estimation of the volatility is central to the VaR measure. The preferred models depend very strongly upon the loss function chosen by the user. González-Rivera, Lee, and Mishra [62] find that, for option pricing, simple models such as the exponential weighted moving average (EWMA) proposed by Riskmetrics [64] performed as well as any GARCH model. For an utility loss function, an asymmetric quadratic GARCH model is the most pre-

ferred. For VaR calculations, a stochastic volatility model dominates all other models.

Nonparametric Models for Volatility

Ziegelmann [163] considers the kernel smoothing techniques that free the traditional parametric volatility estimators from the constraints related to their specific models. He applies the nonparametric local 'exponential' estimator to estimate conditional volatility functions, ensuring its nonnegativity. Its asymptotic properties are established and compared with those for the local linear estimator for the volatility model of Fan and Yao [51]. Long, Su, and Ullah [106] extend this idea to semiparametric multivariate GARCH and show that there may exist substantial out-of-sample forecasting gain over the parametric models. This gain accounts for the presence of nonlinearity in the conditional variance-covariance that is neglected in parametric linear models.

Forecasting Volatility Using High Frequency Data

Using high-frequency data, quadratic variation may be estimated using realized volatility (RV). Andersen, Bollerslev, Diebold, and Labys [3] and Barndorff-Nielsen and Shephard [11] establish that RV, defined as the sum of squared intraday returns of small intervals, is an asymptotically unbiased estimator of the unobserved quadratic variation as the interval length approaches zero. Besides the use of high frequency information in volatility estimation, volatility forecasting using high frequency information has been addressed as well. In an application to volatility prediction, Ghysels, Santa-Clara, and Valkanov [58] investigate the predictive power of various regressors (lagged realized volatility, squared return, realized power, and daily range) for future volatility forecasting. They find that the best predictor is realized power (sum of absolute intraday returns), and more interestingly, direct use of intraday squared returns in mixed data sampling (MIDAS) regressions does not necessarily lead to better volatility forecasts.

Andersen, Bollerslev, Diebold, and Labys [4] represent another approach to forecasting volatility using RV. The model they propose is a fractional integrated AR model: ARFI(5, d) for logarithmic RV's obtained from foreign exchange rates data of 30-minute frequency and demonstrate the superior predictive power of their model.

Alternatively, Corsi [32] proposes the heterogeneous autoregressive (HAR) model of RV, which is able to reproduce long memory. McAleer and Medeiros [115] propose a new model that is a multiple regime smooth transition (ST) extension of the HAR model, which is specifically designed to model the behavior of the volatility inherent in financial time series. The model is able to describe simultaneously long memory as well as sign and size asymmetries. They apply the model to several Dow Jones Industrial Average index stocks using transaction level data from the Trades and Quotes database that covers ten years of data, and find strong support for long memory and both sign and size asymmetries. Furthermore, they show that the multiple regime smooth transition HAR model, when combined with the linear HAR model, is flexible for the purpose of forecasting volatility.

Forecasting Beyond Mean and Variance

In the previous section, we have surveyed the major developments in nonlinear time series, mainly modeling the conditional mean and the conditional variance of financial returns. However it is not clear yet that any of those nonlinear models may generate profits after accounting for various market frictions and transactions costs. Therefore, some research efforts have been directed to investigate other aspects of the conditional density of returns such as higher moments, quantiles, directions, intervals, and the density itself. In this section, we provide a brief survey on forecasting these other features.

Forecasting Quantiles

The optimal forecast of a time series model depends on the specification of the loss function. A symmetric quadratic loss function is the most prevalent in applications due to its simplicity. Under symmetric quadratic loss, the optimal forecast is simply the conditional mean. An asymmetric loss function implies a more complicated forecast that depends on the distribution of the forecast error as well as the loss function itself [67].

Consider a stochastic process $Z_t \equiv (Y_t, X_t')'$ where Y_t is the variable of interest and X_t is a vector of other variables. Suppose there are $T+1\,(\equiv R+P)$ observations. We use the observations available at time t, $R \le t < T+1$, to generate P forecasts using each model. For each time t in the prediction period, we use either a rolling sample $\{Z_{t-R+1}, \ldots, Z_t\}$ of size R or the whole past sample $\{Z_1, \ldots, Z_t\}$ to estimate model parameters $\hat{\beta}_t$. We can then generate a sequence of one-step-ahead forecasts $\{f(Z_t, \hat{\beta}_t)\}_{t=R}^{T}$.

Suppose that there is a decision maker who takes an one-step point forecast $f_{t,1} \equiv f(Z_t, \hat{\beta}_t)$ of Y_{t+1} and uses it in some relevant decision. The one-step forecast error $e_{t+1} \equiv Y_{t+1} - f_{t,1}$ will result in a cost of $c(e_{t+1})$, where the function $c(e)$ will increase as e increases in size, but not necessarily symmetrically or continuously. The optimal forecast $f_{t,1}^*$ will be chosen to produce the forecast er-

rors that minimize the expected loss

$$\min_{f_{t,1}} \int_{-\infty}^{\infty} c(y - f_{t,1}) \mathrm{d}F_t(y) \,,$$

where $F_t(y) \equiv \Pr(Y_{t+1} \leq y | I_t)$ is the conditional distribution function, with I_t being some proper information set at time t that includes Z_{t-j}, $j \geq 0$. The corresponding optimal forecast error will be

$$e^*_{t+1} = Y_{t+1} - f^*_{t,1} \,.$$

Then the optimal forecast would satisfy

$$\frac{\partial}{\partial f_{t,1}} \int_{-\infty}^{\infty} c(y - f^*_{t,1}) \mathrm{d}F_t(y) = 0 \,.$$

When we interchange the operations of differentiation and integration,

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial f_{t,1}} c(y - f^*_{t,1}) \mathrm{d}F_t(y) \equiv \mathbb{E}\left(\frac{\partial}{\partial f_{t,1}} c(Y_{t+1} - f^*_{t,1}) | I_t \right)$$

Based on the "generalized forecast error", $g_{t+1} \equiv \frac{\partial}{\partial f_{t,1}} c(Y_{t+1} - f^*_{t,1})$, the condition for forecast optimality is:

$$H_0 : \mathbb{E}\left(g_{t+1} | I_t\right) = 0 \quad a.s. \,,$$

that is a martingale difference (MD) property of the generalized forecast error. This forms the optimality condition of the forecasts and gives an appropriate regression function corresponding to the specified loss function $c(\cdot)$.

To see this we consider the following two examples. First, when the loss function is the squared error loss

$$c(Y_{t+1} - f_{t,1}) = (Y_{t+1} - f_{t,1})^2 \,,$$

the generalized forecast error will be $g_{t+1} \equiv \frac{\partial}{\partial f_t} c(Y_{t+1} - f^*_{t,1}) = -2e^*_{t+1}$ and thus $\mathbb{E}\left(e^*_{t+1} | I_t\right) = 0$ *a.s.*, which implies that the optimal forecast

$$f^*_{t,1} = \mathbb{E}\left(Y_{t+1} | I_t\right)$$

is the conditional mean. Next, when the loss is the check function, $c(e) = \left[\alpha - \mathbf{1}(e < 0)\right] \cdot e \equiv \rho_\alpha(e_{t+1})$, the optimal forecast $f_{t,1}$, for given $\alpha \in (0, 1)$, minimizing

$$\min_{f_{t,1}} \mathbb{E}\left[c(Y_{t+1} - f_{t,1}) | I_t\right]$$

can be shown to satisfy

$$\mathbb{E}\left[\alpha - \mathbf{1}(Y_{t+1} < f^*_{t,1}) | I_t\right] = 0 \quad a.s.$$

Hence, $g_{t+1} \equiv \alpha - \mathbf{1}(Y_{t+1} < f^*_{t,1})$ is the generalized forecast error. Therefore,

$$\alpha = \mathbb{E}\left[\mathbf{1}(Y_{t+1} < f^*_{t,1}) | I_t\right] = \Pr(Y_{t+1} \leq f^*_{t,1} | I_t) \,,$$

and the optimal forecast $f^*_{t,1} = q^\alpha\left(Y_{t+1} | I_t\right) \equiv q^\alpha_t$ is the conditional α-quantile.

Forecasting conditional quantiles are of paramount importance for risk management, which nowdays is key activity in financial institutions due to the increasing financial fragility in emerging markets and the extensive use of derivative products over the last decade. A risk measurement methodology called Value-at-Risk (VaR) has received a great attention from both regulatory and academic fronts. During a short span of time, numerous papers have studied various aspects of the VaR methodology. Bao, Lee, and Saltoglu [8] examine the relative out-of-sample predictive performance of various VaR models.

An interesting VaR model is the CaViaR (conditional autoregressive Value-at-Risk) model suggested by Engle and Manganelli [47]. They estimate the VaR from a quantile regression rather than inverting a conditional distribution. The idea is similar to the GARCH modeling in that VaR is modeled autoregressively

$$q_t\left(\alpha\right) = a_0 + a_1 q_{t-1}\left(\alpha\right) + h(x_t | \theta) \,,$$

where $x_t \in \mathcal{F}_{t-1}$, θ is a parameter vector, and $h(\cdot)$ is a function to explain the VaR model. Depending on the specification of $h(\cdot)$, the CaViaR model may be

$$q_t\left(\alpha\right) = a_0 + a_1 q_{t-1}\left(\alpha\right) + a_2 |r_{t-1}| \,,$$

$$q_t\left(\alpha\right) = a_0 + a_1 q_{t-1}\left(\alpha\right) + a_2 |r_{t-1}| + a_3 |r_{t-1}| \cdot \mathbf{1}(r_{t-1} < 0),$$

where the second model allow nonlinearity (asymmetry) similarly to the asymmetric GARCH models.

Bao, Lee, and Saltoglu [8] compare various VaR models. Their results show that the CaViaR quantile regression models of Engle and Manganelli [47] have shown some success in predicting the VaR risk measure for various periods of time, and it is generally more stable than the models that invert a distribution function.

Forecasting Directions

It is well known that, while financial returns $\{Y_t\}$ may not be predictable, their variance, sign, and quantiles may be predictable. Christofferson and Diebold [27] show that binary variable $G_{t+1} \equiv \mathbf{1}(Y_{t+1} > 0)$, where $\mathbf{1}(\cdot)$ takes the value of 1 if the statement in the parenthesis is true, and 0 otherwise, is predictable when some conditional moments are time varying, Hong and Lee [86], Hong and

Chung [85], Linton and Whang [104], Lee and Yang [100] among many others find some evidence that the directions of stock returns and foreign exchange rate changes are predictable.

Lee and Yang [100] also show that forecasting quantiles and forecasting binary (directional) forecasts are related, in that the former may lead to the latter. As noted by Powell [126], using the fact that for any monotonic function $h(\cdot)$, $q_t^\alpha(h(Y_{t+1})|\mathbf{X}_t) = h(q_t^\alpha(Y_{t+1}|\mathbf{X}_t))$, which follows immediately from observing that $\Pr(Y_{t+1} < y|\mathbf{X}_t) = \Pr[h(Y_{t+1}) < h(y)|\mathbf{X}_t]$, and noting that the indicator function is monotonic, $q_t^\alpha(G_{t+1}|\mathbf{X}_t) = q_t^\alpha(\mathbf{1}(Y_{t+1} > 0)|\mathbf{X}_t) = \mathbf{1}(q_t^\alpha(Y_{t+1}|\mathbf{X}_t) > 0)$. Therefore, predictability of conditional quantiles of financial returns may imply predictability of conditional direction.

Probability Forecasts

Diebold and Rudebush [38] consider the probability forecasts for the turning points of the business cycle. They measure the accuracy of predicted probabilities, that is the average distance between the predicted probabilities and observed realization (as measured by a zero-one dummy variable). Suppose there are $T+1\,(\equiv R+P)$ observations. We use the observations available at time $t\,(R \le t < T+1)$, to estimate a model. We then have time series of $P = T - R + 1$ probability forecasts $\{p_{t+1}\}_{t=R}^T$ where p_t is the predicted probability of the occurrence of an event (e. g., business cycle turning point) in the next period $t+1$. Let $\{d_{t+1}\}_{t=R}^T$ be the corresponding realization with $d_t = 1$ if a business cycle turning point (or any defined event) occurs in period t and $d_t = 0$ otherwise. The loss function analogous to the squared error is the Brier's score based on quadratic probability score (QPS):

$$QPS = P^{-1}\sum_{t=R}^{T} 2(p_t - d_t)^2 \,.$$

The QPS ranges from 0 to 2, with 0 for perfect accuracy. As noted by Diebold and Rudebush [38], the use of the symmetric loss function may not be appropriate as a forecaster may be penalized more heavily for missing a call (making a type II error) than for signaling a false alarm (making a type I error). Another loss function is given by the log probability score (LPS)

$$LPS = -P^{-1}\sum_{t=R}^{T} \ln\left(p_t^{d_t}(1-p_t)^{(1-d_t)}\right) \,,$$

which is similar to the loss for the interval forecast. A large mistake is penalized more heavily under LPS than under QPS. More loss functions are discussed in Diebold and Rudebush [38].

Another loss function useful in this context is the Kuipers score (KS), which is defined by

$$KS = \text{Hit Rate} - \text{False Alarm Rate} \,,$$

where Hit Rate is the fraction of the bad events that were correctly predicted as good events (power, or 1− probability of type II error), and False Alarm Rate is the fraction of good events that had been incorrectly predicted as bad events (probability of type I error).

Forecasting Interval

Suppose Y_t is a stationary series. Let the one-period ahead conditional interval forecast made at time t from a model be denoted as

$$J_{t,1}(\alpha) = (L_{t,1}(\alpha), U_{t,1}(\alpha)), \quad t = R, \ldots, T \,,$$

where $L_{t,1}(\alpha)$ and $U_{t,1}(\alpha)$ are the lower and upper limits of the ex ante interval forecast for time $t+1$ made at time t with the coverage probability α. Define the indicator variable $X_{t+1}(\alpha) = \mathbf{1}[Y_{t+1} \in J_{t,1}(\alpha)]$. The sequence $\{X_{t+1}(\alpha)\}_{t=R}^T$ is i.i.d. Bernoulli (α). The optimal interval forecast would satisfy $\mathbb{E}(X_{t+1}(\alpha)|I_t) = \alpha$, so that $\{X_{t+1}(\alpha) - \alpha\}$ will be an MD. A better model has a larger expected Bernoulli log-likelihood

$$\mathbb{E}\alpha^{X_{t+1}(\alpha)}(1-\alpha)^{[1-X_{t+1}(\alpha)]} \,.$$

Hence, we can choose a model for interval forecasts with the largest out-of-sample mean of the predictive log-likelihood, which is defined by

$$P^{-1}\sum_{t=R}^{T} \ln\left(\alpha^{x_{t+1}(\alpha)}(1-\alpha)^{[1-x_{t+1}(\alpha)]}\right) \,.$$

Evaluation of Nonlinear Forecasts

In order to evaluate the possible superior predictive ability of nonlinear models, we need to compare competing models in terms of a certain loss function. The literature has recently been exploding on this issue. Examples are Granger and Newbold [69], Diebold and Mariano [37], West [156], White [160], Hansen [81], Romano and Wolf [130], Giacomini and White [59], etc. In different perspective, to test the optimality of a given model, Patton and Timmermann [123] examine various testable properties that should hold for an optimal forecast.

Loss Functions

The loss function (or cost function) is a crucial ingredient for the evaluation of nonlinear forecasts. When a forecast $f_{t,h}$ of a variable Y_{t+h} is made at time t for h periods ahead, the loss (or cost) will arise if a forecast turns out to be different from the actual value. The loss function of the forecast error $e_{t+h} = Y_{t+h} - f_{t,h}$ is denoted as $c(Y_{t+h}, f_{t,h})$. The loss function can depend on the time of prediction and so it can be $c_{t+h}(Y_{t+h}, f_{t,h})$. If the loss function is not changing with time and does not depend on the value of the variable Y_{t+h}, the loss can be written simply as a function of the error only, $c_{t+h}(Y_{t+h}, f_{t,h}) = c(e_{t+h})$.

Granger [67] discusses the following required properties for a loss function: (i) $c(0) = 0$ (no error and no loss), (ii) $\min_e c(e) = 0$, so that $c(e) \geq 0$, and (iii) $c(e)$ is monotonically nondecreasing as e moves away from zero so that $c(e_1) \geq c(e_2)$ if $e_1 > e_2 > 0$ and if $e_1 < e_2 < 0$.

When $c_1(e), c_2(e)$ are both loss functions, Granger [67] shows that further examples of loss functions can be generated: $c(e) = ac_1(e) + bc_2(e), a \geq 0, b \geq 0$ will be a loss function. $c(e) = c_1(e)^a c_2(e)^b$, $a > 0, b > 0$ will be a loss function. $c(e) = 1(e > 0)c_1(e) + 1(e < 0)c_2(e)$ will be a loss function. If $h(\cdot)$ is a positive monotonic nondecreasing function with $h(0)$ finite, then $c(e) = h(c_1(e)) - h(0)$ is a loss function.

Granger [68] notes that an expected loss (a risk measure) of financial return Y_{t+1} that has a conditional predictive distribution $F_t(y) \equiv \Pr(Y_{t+1} \leq y | I_t)$ with $\mathbf{X}_t \in I_t$ may be written as

$$\mathbb{E}c(e) = A_1 \int_0^\infty |y-f|^p dF_t(y) + A_2 \int_{-\infty}^0 |y-f|^p \mathrm{d}F_t(y),$$

with A_1, A_2 both > 0 and some $\theta > 0$. Considering the symmetric case $A_1 = A_2$, one has a class of volatility measures $V_p = \mathbb{E}\left[|y - f|^p\right]$, which includes the variance with $p = 2$, and mean absolute deviation with $p = 1$.

Ding, Granger, and Engle [39] study the time series and distributional properties of these measures empirically and show that the absolute deviations are found to have some particular properties such as the longest memory. Granger remarks that given that the financial returns are known to come from a long tail distribution, $p = 1$ may be more preferable.

Another problem raised by Granger is how to choose optimal L_p-norm in empirical works, to minimize $\mathbb{E}[|\varepsilon_t|^p]$ for some p to estimate the regression model $Y_t = \mathbb{E}(Y_t|X_t;\beta) + \varepsilon_t$. As the asymptotic covariance matrix of $\hat{\beta}$ depends on p, the most appropriate value of p can be chosen to minimize the covariance matrix. In particular, Granger [68] refers to a trio of papers [84,116,117] who find that the optimal $p = 1$ from Laplace and Cauchy distribution, $p = 2$ for Gaussian and $p = \infty$ (min/max estimator) for a rectangular distribution. Granger [68] also notes that in terms of the kurtosis κ, Harter [84] suggests to use $p = 1$ for $\kappa > 3.8$; $p = 2$ for $2.2 \leq \kappa \leq 3.8$; and $p = 3$ for $\kappa < 2.2$. In finance, the kurtosis of returns can be thought of as being well over 4 and so $p = 1$ is preferred.

Forecast Optimality

Optimal forecast of a time series model extensively depends on the specification of the loss function. Symmetric quadratic loss function is the most prevalent in applications due to its simplicity. The optimal forecast under quadratic loss is simply the conditional mean, but an asymmetric loss function implies a more complicated forecast that depends on the distribution of the forecast error as well as the loss function itself [67], as the expected loss function if formulated with the expectation taken with respect to the conditional distribution. Specification of the loss function defines the model under consideration.

Consider a stochastic process $Z_t \equiv (Y_t, X_t')'$ where Y_t is the variable of interest and X_t is a vector of other variables. Suppose there are $T + 1 (\equiv R + P)$ observations. We use the observations available at time t, $R \leq t < T + 1$, to generate P forecasts using each model. For each time t in the prediction period, we use either a rolling sample $\{Z_{t-R+1}, \ldots, Z_t\}$ of size R or the whole past sample $\{Z_1, \ldots, Z_t\}$ to estimate model parameters $\hat{\beta}_t$. We can then generate a sequence of one-step-ahead forecasts $\{f(Z_t, \hat{\beta}_t)\}_{t=R}^T$.

Suppose that there is a decision maker who takes an one-step point forecast $f_{t,1} \equiv f(Z_t, \hat{\beta}_t)$ of Y_{t+1} and uses it in some relevant decision. The one-step forecast error $e_{t+1} \equiv Y_{t+1} - f_{t,1}$ will result in a cost of $c(e_{t+1})$, where the function $c(e)$ will increase as e increases in size, but not necessarily symmetrically or continuously. The optimal forecast $f_{t,1}^*$ will be chosen to produce the forecast errors that minimize the expected loss

$$\min_{f_{t,1}} \int_{-\infty}^{\infty} c(y - f_{t,1}) \mathrm{d}F_t(y) \,,$$

where $F_t(y) \equiv \Pr(Y_{t+1} \leq y | I_t)$ is the conditional distribution function, with I_t being some proper information set at time t that includes Z_{t-j}, $j \geq 0$. The corresponding optimal forecast error will be

$$e_{t+1}^* = Y_{t+1} - f_{t,1}^*.$$

Then the optimal forecast would satisfy

$$\frac{\partial}{\partial f_{t,1}} \int_{-\infty}^{\infty} c(y - f_{t,1}^*) \mathrm{d}F_t(y) = 0 \,.$$

When we may interchange the operations of differentiation and integration,

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial f_{t,1}} c(y - f_{t,1}^*) \mathrm{d}F_t(y) \equiv \mathbb{E}\left(\frac{\partial}{\partial f_{t,1}} c(Y_{t+1} - f_{t,1}^*) | I_t \right)$$

the "generalized forecast error", $g_{t+1} \equiv \frac{\partial}{\partial f_{t,1}} c(Y_{t+1} - f_{t,1}^*)$, forms the condition of forecast optimality:

$$H_0 : \mathbb{E}\left(g_{t+1} | I_t\right) = 0 \quad a.s.,$$

that is a martingale difference (MD) property of the generalized forecast error. This forms the optimality condition of the forecasts and gives an appropriate regression function corresponding to the specified loss function $c(\cdot)$.

Forecast Evaluation of Nonlinear Transformations

Granger [67] note that it is implausible to use the same loss function for forecasting Y_{t+h} and for forecasting $h_{t+1} = h(Y_{t+h})$ where $h(\cdot)$ is some function, such as the log or the square, if one is interested in forecasting volatility. Suppose the loss functions $c_1(\cdot), c_2(\cdot)$ are used for forecasting Y_{t+h} and for forecasting $h(Y_{t+h})$, respectively. Let $e_{t+1} \equiv Y_{t+1} - f_{t,1}$ will result in a cost of $c_1(e_{t+1})$, for which the optimal forecast $f_{t,1}^*$ will be chosen from $\min_{f_{t,1}} \int_{-\infty}^{\infty} c_1(y - f_{t,1}) \mathrm{d}F_t(y)$, where $F_t(y) \equiv \Pr(Y_{t+1} \leq y | I_t)$. Let $\varepsilon_{t+1} \equiv h_{t+1} - h_{t,1}$ will result in a cost of $c_2(\varepsilon_{t+1})$, for which the optimal forecast $h_{t,1}^*$ will be chosen from $\min_{h_{t,1}} \int_{-\infty}^{\infty} c_2(h - h_{t,1}) \mathrm{d}H_t(h)$, where $H_t(h) \equiv \Pr(h_{t+1} \leq h | I_t)$. Then the optimal forecasts for Y and h would respectively satisfy

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial f_{t,1}} c_1(y - f_{t,1}^*) \mathrm{d}F_t(y) = 0 \,,$$
$$\int_{-\infty}^{\infty} \frac{\partial}{\partial h_{t,1}} c_2(h - h_{t,1}^*) \mathrm{d}H_t(h) = 0 \,.$$

It is easy to see that the optimality condition for $f_{t,1}^*$ does not imply the optimality condition for $h_{t,1}^*$ in general. Under some strong conditions on the functional forms of the transformation $h(\cdot)$ and of the two loss functions $c_1(\cdot), c_2(\cdot)$, the above two conditions may coincide. Granger [67] remarks that it would be strange behavior to use the same loss function for Y and $h(Y)$. We leave this for further analysis in a future research.

Density Forecast Evaluation

Most of the classical finance theories, such as asset pricing, portfolio selection and option valuation, aim to model the surrounding uncertainty via a parametric distribution function. For example, extracting information about market participants' expectations from option prices can be considered another form of density forecasting exercise [92]. Moreover, there has also been increasing interest in evaluating forecasting models of inflation, unemployment and output in terms of density forecasts [29]. While evaluating each density forecast model has become versatile since Diebold et al. [35], there has been much less effort in comparing alternative density forecast models.

Given the recent empirical evidence on volatility clustering and asymmetry and fat-tailedness in financial return series, relative adequacy of a given model among alternative models would be useful measure of evaluating forecast models. Deciding on which distribution and/or volatility specification to use for a particular asset is a common task even for finance practitioners. For example, despite the existence of many volatility specifications, a consensus on which model is most appropriate has yet to be reached. As argued in Poon and Granger [125], most of the (volatility) forecasting studies do not produce very conclusive results because only a subset of alternative models are compared, with a potential bias towards the method developed by the authors. Poon and Granger [125] argue that lack of a uniform forecast evaluation technique makes volatility forecasting a difficult task. They wrote (p. 507), "… it seems clear that one form of study that is included is conducted just to support a viewpoint that a particular method is useful. It might not have been submitted for publication if the required result had not been reached. This is one of the obvious weaknesses of a comparison such as this; the papers being prepared for different reasons, use different data sets, many kinds of assets, various intervals between readings, and a variety of evaluation techniques".

Following Diebold et al. [35], it has become common practice to evaluate the adequacy of a forecast model based on the probability integral transform (PIT) of the process with respect to the model's density forecast. If the density forecast model is correctly specified, the PIT follows an i.i.d. uniform distribution on the unit interval and, equivalently, its inverse normal transform follows an i.i.d. normal distribution. We can therefore evaluate a density forecast model by examining the departure of the transformed PIT from this property (i.i.d. and normality). The departure can be quantified by the Kullback-Leibler [97] information criterion, or KLIC, which is the expected logarithmic

value of the likelihood ratio (LR) of the transformed PIT and the i.i.d. normal variate. Thus the LR statistic measures the distance of a candidate model to the unknown true model.

Consider a financial return series $\{y_t\}_{t=1}^T$. This observed data on a univariate series is a realization of a stochastic process $\mathbf{Y}^T \equiv \{Y_\tau : \Omega \to \mathbb{R}, \tau = 1, 2, \ldots, T\}$ on a complete probability space $(\Omega, \mathcal{F}_T, P_0^T)$, where $\Omega = \mathbb{R}^T \equiv \times_{\tau=1}^T \mathbb{R}$ and $\mathcal{F}_T = \mathcal{B}(\mathbb{R}^T)$ is the Borel σ-field generated by the open sets of $\mathbb{R}^T$, and the *joint* probability measure $P_0^T(B) \equiv P_0[\mathbf{Y}^T \in B]$, $B \in \mathcal{B}(\mathbb{R}^T)$ completely describes the stochastic process. A sample of size T is denoted as $\mathbf{y}^T \equiv (y_1, \ldots, y_T)'$.

Let σ-finite measure ν^T on $\mathcal{B}(\mathbb{R}^T)$ be given. Assume $P_0^T(B)$ is absolutely continuous with respect to ν^T for all $T = 1, 2, \ldots$, so that there exists a measurable Radon–Nikodým density $g^T(\mathbf{y}^T) = \mathrm{d}P_0^T/\mathrm{d}\nu^T$, unique up to a set of zero measure-ν^T.

Following White [159], we define a probability model $\mathcal{P}$ as a collection of distinct probability measures on the measurable space $(\Omega, \mathcal{F}_T)$. A probability model $\mathcal{P}$ is said to be correctly specified for $\mathbf{Y}^T$ if $\mathcal{P}$ contains P_0^T. Our goal is to evaluate and compare a set of parametric probability models $\{P_\theta^T\}$, where $P_\theta^T(B) \equiv P_\theta[Y^T \in B]$. Suppose there exists a measurable Radon–Nikodým density $f^T(\mathbf{y}^T) = \mathrm{d}P_\theta^T/\mathrm{d}\nu^T$ for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\theta}$ is a finite-dimensional vector of parameters and is assumed to be identified on $\boldsymbol{\Theta}$, a compact subset of $\mathbb{R}^k$. See Theorem 2.6 in White [159].

In the context of forecasting, instead of the joint density $g^T(\mathbf{y}^T)$, we consider forecasting the *conditional* density of $\mathbf{Y}^t$, given the information $\mathcal{F}_{t-1}$ generated by $\mathbf{Y}^{t-1}$. Let $\varphi_t\left(y_t\right) \equiv \varphi_t(y_t|\mathcal{F}_{t-1}) \equiv g^t(\mathbf{y}^t)/g^{t-1}(\mathbf{y}^{t-1})$ for $t = 2, 3, \ldots$ and $\varphi_1\left(y_1\right) \equiv \varphi_1(y_1|\mathcal{F}_0) \equiv g^1(\mathbf{y}^1) = g^1(y_1)$. Thus the goal is to forecast the (true, unknown) conditional density $\varphi_t\left(y_t\right)$.

For this, we use an one-step-ahead conditional density forecast model $\psi_t\left(y_t; \boldsymbol{\theta}\right) \equiv \psi_t(y_t|\mathcal{F}_{t-1}; \boldsymbol{\theta}) \equiv f^t(\mathbf{y}^t)/f^{t-1}(\mathbf{y}^{t-1})$ for $t = 2, 3, \ldots$ and $\psi_1\left(y_1\right) \equiv \psi_1(y_1|\mathcal{F}_0) \equiv f^1(\mathbf{y}^1) = f^1(y_1)$. If $\psi_t(y_t; \boldsymbol{\theta}_0) = \varphi_t(y_t)$ almost surely for some $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, then the one-step-ahead density forecast is correctly specified, and it is said to be optimal because it dominates all other density forecasts for any loss functions as discussed in the previous section (see [35,67,70]).

In practice, it is rarely the case that we can find an optimal model. As it is very likely that "the true distribution is in fact too complicated to be represented by a simple mathematical function" [133], all the models proposed by different researchers can be possibly misspecified and thereby we regard each model as an approximation to the truth. Our task is then to investigate which density forecast model can approximate the true conditional density most closely. We have to first define a metric to measure the distance of a given model to the truth, and then compare different models in terms of this distance.

The adequacy of a density forecast model can be measured by the conditional Kullback-Leibler [97] Information Criterion (KLIC) divergence measure between two conditional densities,

$$\mathbb{I}_t\left(\varphi : \psi, \boldsymbol{\theta}\right) = \mathbb{E}_{\varphi_t}[\ln \varphi_t\left(y_t\right) - \ln \psi_t\left(y_t; \boldsymbol{\theta}\right)] ,$$

where the expectation is with respect to the true conditional density $\varphi_t\left(\cdot|\mathcal{F}_{t-1}\right)$, $\mathbb{E}_{\varphi_t} \ln \varphi_t\left(y_t|\mathcal{F}_{t-1}\right) < \infty$, and $\mathbb{E}_{\varphi_t} \ln \psi_t\left(y_t|\mathcal{F}_{t-1}; \boldsymbol{\theta}\right) < \infty$. Following White [159], we define the distance between a density model and the true density as the minimum of the KLIC

$$\mathbb{I}_t\left(\varphi : \psi, \boldsymbol{\theta}_{t-1}^*\right) = \mathbb{E}_{\varphi_t}\left[\ln \varphi_t\left(y_t\right) - \ln \psi_t\left(y_t; \boldsymbol{\theta}_{t-1}^*\right)\right] ,$$

where $\boldsymbol{\theta}_{t-1}^* = \arg\min \mathbb{I}_t\left(\varphi : \psi, \boldsymbol{\theta}\right)$ is the pseudo-true value of $\boldsymbol{\theta}$ [133]. We assume that $\boldsymbol{\theta}_{t-1}^*$ is an interior point of $\boldsymbol{\Theta}$. The smaller this distance is, the closer the density forecast $\psi_t\left(\cdot|\mathcal{F}_{t-1}; \boldsymbol{\theta}_{t-1}^*\right)$ is to the true density $\varphi_t\left(\cdot|\mathcal{F}_{t-1}\right)$.

However, $\mathbb{I}_t\left(\varphi : \psi, \boldsymbol{\theta}_{t-1}^*\right)$ is unknown since $\boldsymbol{\theta}_{t-1}^*$ is not observable. We need to estimate $\boldsymbol{\theta}_{t-1}^*$. If our purpose is to compare the out-of-sample predictive abilities among competing density forecast models, we split the data into two parts, one for estimation and the other for out-of-sample validation. At each period t in the out-of-sample period ($t = R + 1, \ldots, T$), we estimate the unknown parameter vector $\boldsymbol{\theta}_{t-1}^*$ and denote the estimate as $\hat{\boldsymbol{\theta}}_{t-1}$. Using $\{\hat{\boldsymbol{\theta}}_{t-1}\}_{t=R+1}^T$, we can obtain the out-of-sample estimate of $\mathbb{I}_t\left(\varphi : \psi, \boldsymbol{\theta}_{t-1}^*\right)$ by

$$\mathbb{I}_P(\varphi : \psi) \equiv \frac{1}{P} \sum_{t=R+1}^{T} \ln[\varphi_t(y_t)/\psi_t(y_t; \hat{\boldsymbol{\theta}}_{t-1})]$$

where $P = T - R$ is the size of the out-of-sample period. Note that

$$\mathbb{I}_P(\varphi : \psi) = \frac{1}{P} \sum_{t=R+1}^{T} \ln\left[\varphi_t(y_t)/\psi_t\left(y_t; \boldsymbol{\theta}_{t-1}^*\right)\right] + \frac{1}{P} \sum_{t=R+1}^{T} \ln[\psi_t\left(y_t; \boldsymbol{\theta}_{t-1}^*\right)/\psi_t(y_t; \hat{\boldsymbol{\theta}}_{t-1})] ,$$

where the first term in $\mathbb{I}_P(\varphi : \psi)$ measures model uncertainty (the distance between the optimal density $\varphi_t(y_t)$ and the model $\psi_t\left(y_t; \boldsymbol{\theta}_{t-1}^*\right)$) and the second term mea-

sures parameter estimation uncertainty due to the distance between $\boldsymbol{\theta}^*_{t-1}$ and $\hat{\boldsymbol{\theta}}_{t-1}$.

Since the KLIC measure takes on a smaller value when a model is closer to the truth, we can regard it as a loss function and use $\mathbb{I}_P(\varphi : \psi)$ to formulate the loss-differential. The out-of-sample average of the loss-differential between model 1 and model 2 is

$$\mathbb{I}_P(\varphi : \psi^1) - \mathbb{I}_P(\varphi : \psi^2) = \frac{1}{P} \sum_{t=R+1}^{T} \ln \left[\psi_t^2 \left(y_t; \hat{\boldsymbol{\theta}}_{t-1}^2 \right) / \psi_t^1 \left(y_t; \hat{\boldsymbol{\theta}}_{t-1}^1 \right) \right] ,$$

which is the ratio of the two predictive log-likelihood functions. With treating model 1 as a benchmark model (for model selection) or as the model under the null hypothesis (for hypothesis testing), $\mathbb{I}_P(\varphi : \psi^1) - \mathbb{I}_P(\varphi : \psi^2)$ can be considered as a loss function to minimize. To sum up, the KLIC differential can serve as a *loss* function for density forecast evaluation as discussed in Bao, Lee, and Saltoglu [10]. See Corradi and Swanson [31] for the related ideas using different loss functions.

Using the KLIC divergence measure to characterize the extent of misspecification of a forecast model, Bao, Lee, and Saltoglu [10], in an empirical study with the S&P500 and NASDAQ daily return series, find strong evidence for rejecting the Normal-GARCH benchmark model, in favor of the models that can capture skewness in the conditional distribution and asymmetry and long-memory in the conditional variance. Also, Bao and Lee [8] investigate the nonlinear predictability of stock returns when the density forecasts are evaluated/compared instead of the conditional mean point forecasts. The conditional mean models they use for the daily closing S&P500 index returns include the martingale difference model, the linear ARMA models, the STAR and SETAR models, the ANN model, and the polynomial model. Their empirical findings suggest that the out-of-sample predictive abilities of nonlinear models for stock returns are asymmetric in the sense that the right tails of the return series are predictable via many of the nonlinear models while we find no such evidence for the left tails or the entire distribution.

Conclusions

In this article we have selectively reviewed the state-of-the-art in nonlinear time series models that are useful in forecasting financial variables. Overall financial returns are difficult to forecast, and this may just be a reflection of the efficiency of the markets on processing information. The success of nonlinear time series on producing better forecasts than linear models depends on how persistent the nonlinearities are in the data. We should note that though many of the methodological developments are concerned with the specification of the conditional mean and conditional variance, there is an active area of research investigating other aspects of the conditional density – quantiles, directions, intervals – that seem to be promising from a forecasting point of view.

For a more extensive coverage to complement this review, the readers may find the following additional references useful. Campbell, Lo, and MacKinlay [22], Chapter 12, provides a brief but excellent summary of nonlinear time series models for the conditional mean and conditional variance as well and various methods such as ANN and nonparametric methods. Similarly, the interested readers may also refer to the books and monographs of Granger and Teräsvirta [72], Franses and van Dijk [55], Fan and Yao [52], Tsay [153], Gao [57], and some book chapters such as Stock [139], Tsay [152], Teräsvirta [145], and White [161].

Future Directions

Methodological developments in nonlinear time series have happened without much guidance from economic theory. Nonlinear models are for most part ad hoc specifications that, from a forecasting point of view, are validated according to some statistical loss function. Though we have surveyed some articles that employ some economic rationale to evaluate the model and/or the forecast – bull/bear cycles, utility function, profit/loss function –, there is still a vacuum on understanding why, how, and when nonlinearities may show up in the data.

From a methodological point of view, future developments will focus on multivariate nonlinear time series models and their associated statistical inference. Nonlinear VAR-type models for the conditional mean and high-dimensional multivariate volatility models are still in their infancy. Dynamic specification testing in a multivariate setting is paramount to the construction of a multivariate forecast and though multivariate predictive densities are inherently difficult to evaluate, they are most important in financial economics.

Another area of future research will deal with the econometrics of a data-rich environment. The advent of large databases begs the introduction of new techniques and methodologies that permits the reduction of the many dimensions of a data set to a parsimonious but highly informative set of variables. In this sense, criteria on how to combine information and how to combine models to produce more accurate forecasts are highly desirable.

Finally, there are some incipient developments on defining new stochastic processes where the random variables that form the process are of a symbolic nature, i. e. intervals, boxplots, histograms, etc. Though the mathematics of these processes are rather complex, future developments in this area will bring exciting results for the area of forecasting.

Bibliography

1. Ait-Sahalia Y, Hansen LP (2009) Handbook of Financial Econometrics. Elsevier Science, Amsterdam
2. Andersen TG, Bollerslev T (1998) Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts. Int Econ Rev 39(4):885–905
3. Andersen TG, Bollerslev T, Diebold FX, Labys P (2001) The Distribution of Realized Exchange Rate Volatility. J Amer Stat Assoc 96:42–55
4. Andersen TG, Bollerslev T, Diebold FX, Labys P (2003) Modeling and Forecasting Realized Volatility. Econometrica 71: 579–625
5. Ang A, Bekaert G (2002) Regime Switccheds in Interest Rates. J Bus Econ Stat 20:163–182
6. Bachelier L (1900) Theory of Speculation. In: Cootner P (ed) The Random Character of Stock Market Prices. MIT Press, Cambridge. (1964) reprint
7. Bai J, Ng S (2007) Forecasting Economic Time Series Using Targeted Predictors. Working Paper, New York University and Columbia University
8. Bao Y, Lee TH (2006) Asymmetric Predictive Abilities of Nonlinear Models for Stock Returns: Evidence from Density Forecast Comparison. Adv Econ 20 B:41–62
9. Bao Y, Lee TH, Saltoglu B (2006) Evaluating Predictive Performance of Value-at-Risk Models in Emerging Markets: A Reality Check. J Forecast 25(2):101–128
10. Bao Y, Lee TH, Saltoglu B (2007) Comparing Density Forecast Models. J Forecast 26(3):203–225
11. Barndorff-Nielsen OE, Shephard N (2002) Econometric Analysis of Realised Volatility and Its Use in Estimating Stochastic Volatility Models. J Royal Stat Soc B 64:853–223
12. Bauwens L, Laurent S, Rombouts JVK (2006) Multivariate GARCH Models: A Survey. J Appl Econ 21:79–109
13. Bera AK, Higgins ML (1993) ARCH Models: Properties, Estimation, and Testing. J Econ Surv 7:305–366
14. Bollerslev T (1986) Generalized Autoregressive Conditional Heteroskedasticity. J Econ 31:307–327
15. Bollerslev T, Chou RY, Kroner KF (1992) ARCH Models in Finance. J Econ 52:5–59
16. Bollerslev T, Engle RF, Nelson DB (1994) ARCH Models. In: Engle RF, McFadden DL (eds) Handbook of Econometrics, vol 4. Elsevier Science, Amsterdam
17. Bollerslev T, Engle RF, Wooldridge J (1988) A Capital Asset Pricing Model with Time Varying Covariances. J Political Econ 96:116–131
18. Boero G, Marrocu E (2004) The Performance of SETAR Models: A Regime Conditional Evaluation of Point, Interval, and Density Forecasts. Int J Forecast 20:305–320
19. Breiman L (1996) Bagging Predictors. Machine Learning 24:123–140
20. Breiman L (1996) Heuristics of Instability and Stabilization in Model Selection. Ann Stat 24(6):2350–2383
21. Brooks C, Persand G (2003) Volatility Forecasting for Risk Management. J Forecast 22(1):1–22
22. Campbell JY, Lo AW, MacKinlay AC (1997) The Econometrics of Financial Markets. Princeton University Press, New Jersey
23. Campbell JY, Thompson SB (2007) Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average? Harvard Institute of Economic Research, Discussion Paper No. 2084
24. Cai Z, Fan J, Yao Q (2000) Functional-coefficient Regression Models for Nonlinear Time Series. J Amer Stat Assoc 95: 941–956
25. Chen X (2006) Large Sample Sieve Estimation of Semi-Nonparametric Models. In: Heckman JJ, Leamer EE (eds) Handbook of Econometrics, vol 6. Elsevier Science, Amsterdam, Chapter 76
26. Chen R, Tsay RS (1993) Functional-coefficient Autoregressive Models. J Amer Stat Assoc 88:298–308
27. Christofferson PF, Diebold FX (2006) Financial Asset Returns, Direction-of-Change Forecasting, and Volatility Dynamics. Manag Sci 52:1273–1287
28. Clements MP, Franses PH, Swanson NR (2004) Forecasting Economic and Financial Time-Series with Non-linear Models. Int J Forecast 20:169–183
29. Clements MP, Smith J (2000) Evaluating the forecast densities of linear and non-linear models: applications to output growth and unemployment. J Forecast 19:255–276
30. Cleveland WS (1979) Robust Locally Weighted Regression and Smoothing Scatter Plots. J Amer Stat Assoc 74: 829–836
31. Corradi V, Swanson NR (2006) Predictive Density Evaluation. In: Granger CWJ, Elliot G, Timmerman A (eds) Handbook of Economic Forecasting. Elsevier, Amsterdam, pp 197–284
32. Corsi F (2004) A Simple Long Memory Model of Realized Volatility. Working Paper, University of Lugano
33. Dahl CM, González-Rivera G (2003) Testing for Neglected Nonlinearity in Regression Models based on the Theory of Random Fields. J Econ 114:141–164
34. Dahl CM, González-Rivera G (2003) Identifying Nonlinear Components by Random Fields in the US GNP Growth. Implications for the Shape of the Business Cycle. Stud Nonlinear Dyn Econ 7(1):art2
35. Diebold FX, Gunther TA, Tay AS (1998) Evaluating Density Forecasts with Applications to Financial Risk Management. Int Econ Rev 39:863–883
36. Diebold FX, Li C (2006) Forecasting the Term Structure of Government Bond Yields. J Econom 130:337–364
37. Diebold FX, Mariano R (1995) Comparing predictive accuracy. J Bus Econ Stat 13:253–265
38. Diebold FX, Rudebusch GD (1989) Scoring the Leading Indicators. J Bus 62(3):369–391
39. Ding Z, Granger CWJ, Engle RF (1993) A Long Memory Property of Stock Market Returns and a New Model. J Empir Finance 1:83–106
40. Dueker M, Neely CJ (2007) Can Markov Switching Models Predict Excess Foreign Exchange Returns? J Bank Finance 31:279–296
41. Durland JM, McCurdy TH (1994) Duration-Dependent Transi-

tions in a Markov Model of US GNP Growth. J Bus Econ Stat 12:279–288
42. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least Angle Regression. Ann Stat 32(2):407–499
43. Engel C, Hamilton JD (1990) Long Swings in the Dollar: Are they in the Data and Do Markets Know it? Amer Econ Rev 80(4):689–713
44. Engle RF (1982) Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of UK Inflation. Econometrica 50:987–1008
45. Engle RF, Hendry DF, Richard J-F (1983) Exogeneity. Econometrica 51:277–304
46. Engle RF, Kane A, Noh J (1997) Index-Option Pricing with Stochastic Volatility and the Value of Accurate Variance Forecasts. Rev Deriv Res 1:139–157
47. Engle RF, Manganelli S (2004) CaViaR: Conditional autoregressive Value at Risk by regression quantiles. J Bus Econ Stat 22(4):367–381
48. Engle RF, Ng VK, Rothschild M (1990) Asset Pricing with a Factor ARCH Covariance Structure: Empirical Estimates for Treasury Bills. J Econ 45:213–238
49. Engle RF, Russell JR (1998) Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data. Econometrica 66:1127–1162
50. Fan J, Gijbels I (1996) Local Polynomial Modelling and Its Applications. Chapman and Hall, London
51. Fan J, Yao Q (1998) Efficient estimation of conditional variance functions in stochastic regression. Biometrika 85: 645–660
52. Fan J, Yao Q (2003) Nonlinear Time Series. Springer, New York
53. Fan J, Yao Q, Cai Z (2003) Adaptive varying-coefficient linear models. J Royal Stat Soc B 65:57–80
54. Fitzenberger B (1997) The Moving Blocks Bootstrap and Robust Inference for Linear Least Squares and Quantile Regressions. J Econ 82:235–287
55. Franses PH, van Dijk D (2000) Nonlinear Time Series Models in Empirical Finance. Cambridge University Press, Cambridge
56. Gallant AR, Nychka DW (1987) Semi-nonparametric maximum likelihood estimation. Econometrica 55:363–390
57. Gao J (2007) Nonlinear Time Series: Semiparametric and Nonparametric Methods. Chapman and Hall, Boca Raton
58. Ghysels E, Santa-Clara P, Valkanov R (2006) Predicting Volatility: How to Get Most out of Returns Data Sampled at Different Frequencies. J Econ 131:59–95
59. Giacomini R, White H (2006) Tests of Conditional Predictive Ability. Econometrica 74:1545–1578
60. Glosten LR, Jaganathan R, Runkle D (1993) On the Relationship between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. J Finance 48:1779–1801
61. González-Rivera G (1998) Smooth-Transition GARCH Models. Stud Nonlinear Dyn Econ 3(2):61–78
62. González-Rivera G, Lee TH, Mishra S (2004) Forecasting Volatility: A Reality Check Based on Option Pricing, Utility Function, Value-at-Risk, and Predictive Likelihood. Int J Forecast 20(4):629–645
63. González-Rivera G, Lee TH, Mishra S (2008) Jumps in Cross-Sectional Rank and Expected Returns: A Mixture Model. J Appl Econ; forthcoming
64. González-Rivera G, Lee TH, Yoldas E (2007) Optimality of the Riskmetrics VaR Model. Finance Res Lett 4:137–145
65. Gonzalo J, Martíneza O (2006) Large shocks vs. small shocks. (Or does size matter? May be so). J Econ 135:311–347
66. Goyal A, Welch I (2006) A Comprehensive Look at The Empirical Performance of Equity Premium Prediction. Working Paper, Emory and Brown, forthcoming in Rev Financ Stud
67. Granger CWJ (1999) Outline of Forecast Theory Using Generalized Cost Functions. Span Econ Rev 1:161–173
68. Granger CWJ (2002) Some Comments on Risk. J Appl Econ 17:447–456
69. Granger CWJ, Newbold P (1986) Forecasting Economic Time Series, 2nd edn. Academic Press, San Diego
70. Granger CWJ, Pesaran MH (2000) A Decision Theoretic Approach to Forecasting Evaluation. In: Chan WS, Li WK, Tong H (eds) Statistics and Finance: An Interface. Imperial College Press, London
71. Granger CWJ, Lee TH (1999) The Effect of Aggregation on Nonlinearity. Econ Rev 18(3):259–269
72. Granger CWJ, Teräsvirta T (1993) Modelling Nonlinear Economic Relationships. Oxford University Press, New York
73. Guidolin M, Timmermann A (2006) An Econometric Model of Nonlinear Dynamics in the Joint Distribution of Stock and Bond Returns. J Appl Econ 21:1–22
74. Haggan V, Ozaki T (1981) Modeling Nonlinear Vibrations Using an Amplitude-dependent Autoregressive Time Series Model. Biometrika 68:189–196
75. Hamilton JD (1994) Time Series Analysis. Princeton University Press, New Jersey
76. Hamilton JD (1989) A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. Econometrica 57:357–384
77. Hamilton JD (1996) Specification Testing in Markov-Switching Time Series Models. J Econ 70:127–157
78. Hamilton JD (2001) A Parametric Approach to Flexible Nonlinear Inference. Econometrica 69:537–573
79. Hamilton JD, Jordà O (2002) A Model of the Federal Funds Target. J Political Econ 110:1135–1167
80. Hansen BE (1996) Inference when a Nuisance Parameter is not Identified under the Null Hypothesis. Econometrica 64: 413–430
81. Hansen PR (2005) A test for superior predictive ability. J Bus Econ Stat 23:365–380
82. Harding D, Pagan A (2002) Dissecting the Cycle: A Methodological Investigation. J Monet Econ 49:365–381
83. Härdle W, Tsybakov A (1997) Local polynomial estimators of the volatility function in nonparametric autoregression. J Econ 81:233–242
84. Harter HL (1977) Nonuniqueness of Least Absolute Values Regression. Commun Stat – Theor Methods A6:829–838
85. Hong Y, Chung J (2003) Are the Directions of Stock Price Changes Predictable? Statistical Theory and Evidence. Working Paper, Department of Economics, Cornell University
86. Hong Y, Lee TH (2003) Inference on Predictability of Foreign Exchange Rates via Generalized Spectrum and Nonlinear Time Series Models. Rev Econ Stat 85(4):1048–1062
87. Hong Y, Lee TH (2003b) Diagnostic Checking for Adequacy of Nonlinear Time Series Models. Econ Theor 19(6):1065–1121
88. Hornik K, Stinchcombe M, White H (1989) Multi-Layer Feedforward Networks Are Universal Approximators. Neural Netw 2:359–366
89. Huang H, Tae-Hwy L, Canlin L (2007) Forecasting Output Growth and Inflation: How to Use Information in the Yield

Curve. Working Paper, University of California, Riverside, Department of Economics
90. Huang YL, Kuan CM (2007) Re-examining Long-Run PPP under an Innovation Regime Switching Framework. Academia Sinica, Taipei
91. Inoue A, Kilian L (2008) How Useful is Bagging in Forecasting Economic Time Series? A Case Study of US CPI Inflation, forthcoming. J Amer Stat Assoc 103(482):511–522
92. Jackwerth JC, Rubinstein M (1996) Recovering probability distributions from option prices. J Finance 51:1611–1631
93. Judd KL (1998) Numerical Methods in Economics. MIT Press, Cambridge
94. Kanas A (2003) Non-linear Forecasts of Stock Returns. J Forecast 22:299–315
95. Koenker R, Bassett Jr G (1978) Regression Quantiles. Econometrica 46(1):33–50
96. Kuan CM, Huang YL, Tsayn RS (2005) An unobserved component model with switching permanent and transitory innovations. J Bus Econ Stat 23:443–454
97. Kullback L, Leibler RA (1951) On Information and Sufficiency. Ann Math Stat 22:79–86
98. Lee TH, Ullah A (2001) Nonparametric Bootstrap Tests for Neglected Nonlinearity in Time Series Regression Models. J Nonparametric Stat 13:425–451
99. Lee TH, White H, Granger CWJ (1993) Testing for Neglected Nonlinearity in Time Series Models: A Comparison of Neural Network Methods and Alternative Tests. J Econ 56:269–290
100. Lee TH, Yang Y (2006) Bagging Binary and Quantile Predictors for Time Series. J Econ 135:465–497
101. Lettau M, Ludvigson S (2001) Consumption, Aggregate Wealth, and Expected Stock Returns. J Finance 56:815–849
102. Lewellen J (2004) Predicting Returns with Financial Ratios. J Financial Econ 74:209–235
103. Lintner J (1965) Security Prices, Risk and Maximal Gains from Diversification. J Finance 20:587–615
104. Linton O, Whang YJ (2007) A Quantilogram Approach to Evaluating Directional Predictability. J Econom 141:250-282
105. Lo AW, MacKinlay AC (1999) A Non-Random Walk Down Wall Street. Princeton University Press, Princeton
106. Long X, Su L, Ullah A (2007) Estimation and Forecasting of Dynamic Conditional Covariance: A Semiparametric Multivariate Model. Working Paper, Department of Economics, UC Riverside
107. Lopez JA (2001) Evaluating the Predictive Accuracy of Volatility Models. J Forecast 20:87–109
108. Ludvigson S, Ng S (2007) The Empirical Risk Return Relation: A Factor Analysis Approach. J Financ Econ 83:171–222
109. Lundbergh S, Teräsvirta T (2002) Forecasting with smooth transition autoregressive models. In: Clements MP, Hendry DF (eds) A Companion to Economic Forecasting. Blackwell, Oxford, Chapter 21
110. Luukkonen R, Saikkonen P, Teräsvirta T (1988) Testing Linearity in Univariate Time Series Models. Scand J Stat 15:161–175
111. Maheu JM, McCurdy TH (2000) Identifying Bull and Bear Markets in Stock Returns. J Bus Econ Stat 18:100–112
112. Manski CF (1975) Maximum Score Estimation of the Stochastic Utility Model of Choice. J Econ 3(3):205–228
113. Markowitz H (1959) Portfolio Selection: Efficient Diversification of Investments. John Wiley, New York
114. Marsh IW (2000) High-frequency Markov Switching Models in the Foreign Exchange Market. J Forecast 19:123–134
115. McAleer M, Medeiros MC (2007) A multiple regime smooth transition heterogeneous autoregressive model for long memory and asymmetries. J Econ; forthcoming
116. Money AH, Affleck-Graves JF, Hart ML, Barr GDI (1982) The Linear Regression Model and the Choice of p. Commun Stat – Simul Comput 11(1):89–109
117. Nyguist H (1983) The Optimal L_p-norm Estimation in Linear Regression Models. Commun Stat – Theor Methods 12: 2511–2524
118. Nadaraya ÉA (1964) On Estimating Regression. Theor Probab Appl 9:141–142
119. Nelson CR, Siegel AF (1987) Parsimonious Modeling of Yield Curves. J Bus 60:473–489
120. Nelson DB (1991) Conditional Heteroscedasticity in Asset Returns: A New Approach. Econometrica 59(2):347–370
121. Noh J, Engle RF, Kane A (1994) Forecasting Volatility and Option Prices of the S&P 500 Index. J Deriv 17–30
122. Pagan AR, Ullah A (1999) Nonparametric Econometrics. Cambridge University Press, Cambridge
123. Patton AJ, Timmermann A (2007) Testing Forecast Optimality Under Unknown Loss. J Amer Stat Assoc 102(480): 1172–1184
124. Perez-Quiros G, Timmermann A (2001) Business Cycle Asymmetries in Stock Returns: Evidence form Higher Order Moments and Conditional Densities. J Econ 103:259–306
125. Poon S, Granger CWJ (2003) Forecasting volatility in financial markets. J Econ Lit 41:478–539
126. Powell JL (1986) Censored Regression Quantiles. J Econ 32:143–155
127. Priestley MB (1980) State-dependent Models: A General Approach to Nonlinear Time Series Analysis. J Time Ser Anal 1:47–71
128. Raj B, Ullah A (1981) Econometrics: A Varying Coefficients Approach. Croom Helm, London
129. Riskmetrics (1995) Technical Manual, 3rd edn. New York
130. Romano JP, Wolf M (2005) Stepwise multiple testing as formalized data snooping. Econometrica 73:1237–1282
131. Ross S (1976) The Arbitrage Theory of Capital Asset Pricing. J Econ Theor 13:341–360
132. Ruppert D, Wand MP (1994) Multivariate Locally Weighted Least Squares Regression. Ann Stat 22:1346–1370
133. Sawa T (1978) Information Criteria for Discriminating among Alternative Regression Models. Econometrica 46:1273–1291
134. Schwert GW (1990) Stock Volatility and the Crash of '87. Rev Financ Stud 3(1):77–102
135. Sentana E (1995) Quadratic ARCH models. Rev Econ Stud 62(4):639–661
136. Sichel DE (1994) Inventories and the Three Phases of the Business Cycle. J Bus Econ Stat 12:269–277
137. Sharpe W (1964) Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. J Finance 19:425–442
138. Stinchcombe M, White H (1998) Consistent Specification Testing with Nuisanse Parameters Present only under the Alternative. Econ Theor 14:295–325
139. Stock JH (2001) Forecasting Economic Time Series. In: Baltagi BP (ed) A Companion to Theoretical Econometrics. Blackwell, Oxford, Chapter 27
140. Stock JH, Watson MW (2002) Forecasting Using Principal Components from a Large Number of Predictors. J Amer Stat Assoc 97:1167–1179
141. Stock JH, Watson MW (2006) Forecasting Using Many Predic-

tors. In: Elliott G, Granger CWJ, Timmermann A (eds) Handbook of Economic Forecasting, vol 1. Elsevier, Amsterdam
142. Stone CJ (1977) Consistent Nonparametric Regression. Ann Stat 5:595–645
143. Taylor SJ (1986) Modelling Financial Time Series. Wiley, New York
144. Teräsvirta T (1994) Specification, Estimation and Evaluation of Smooth Transition Autoregressive Models. J Amer Stat Assoc 89:208–218
145. Teräsvirta T (2006) Forecasting economic variables with nonlinear models. In: Elliott G, Granger CWJ, Timmermann A (eds) Handbook of Economic Forecasting, vol 1. Elsevier, Amsterdam, pp 413–457
146. Teräsvirta T, Anderson H (1992) Characterizing Nonlinearities in Business Cycles using Smooth Transition Autoregressive Models. J Appl Econ 7:119–139
147. Teräsvirta T, Lin CF, Granger CWJ (1993) Power of the Neural Network Linearity Test. J Time Ser Analysis 14:209–220
148. Tong H (1983) Threshold Models in Nonlinear Time Series Analysis. Springer, New York
149. Tong H (1990) Nonlinear Time Series: A Dynamical Systems Approach. Oxford University Press, Oxford
150. Trippi R, Turban E (1992) Neural Networks in Finance and Investing: Using Artificial Intelligence to Improve Real World Performance. McGraw-Hill, New York
151. Tsay RS (1998) Testing and Modeling Multivariate Threshold Models. J Amer Stat Assoc 93:1188–1202
152. Tsay RS (2002) Nonlinear Models and Forecasting. In: Clements MP, Hendry DF (eds) A Companion to Economic Forecasting. Blackwell, Oxford, Chapter 20
153. Tsay RS (2005) Analysis of Financial Time Series, 2nd edn. Wiley, New York
154. Varian HR (1975) A Bayesian Approach to Real Estate Assessment. In: Fienberg SE, Zellner A (eds) Studies in Bayesian Econometrics and Statistics in Honor of L.J. Savage. North Holland, Amsterdam, pp 195–208
155. Watson GS (1964) Smooth Regression Analysis. Sankhya Series A 26:359–372
156. West KD (1996) Asymptotic inference about predictive ability. Econometrica 64:1067–1084
157. West KD, Edison HJ, Cho D (1993) A Utility Based Comparison of Some Models of Exchange Rate Volatility. J Int Econ 35: 23–45
158. White H (1989) An Additional Hidden Unit Tests for Neglected Nonlinearity in Multilayer Feedforward Networks. In: Proceedings of the International Joint Conference on Neural Networks, Washington, DC. IEEE Press, New York, II, pp 451–455
159. White H (1994) Estimation, Inference, and Specification Analysis. Cambridge University Press, Cambridge
160. White H (2000) A Reality Check for Data Snooping. Econometrica 68(5):1097–1126
161. White H (2006) Approximate Nonlinear Forecasting Methods. In: Elliott G, Granger CWJ, Timmermann A (eds) Handbook of Economic Forecasting, vol 1. Elsevier, Amsterdam, chapter 9
162. Zellner A (1986) Bayesian Estimation and Prediction Using Asymmetric Loss Functions. J Amer Stat Assoc 81:446–451
163. Ziegelmann FA (2002) Nonparametric estimation of volatility functions: the local exponential estimator. Econ Theor 18:985–991

Financial Forecasting, Sensitive Dependence

Mototsugu Shintani
Vanderbilt University, Nashville, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Lyapunov Exponent and Forecastability
Nonparametric Estimation of the Global Lyapunov Exponent
Four Local Measures of Sensitive Dependence
Forecasting Financial Asset Returns and Sensitive Dependence
Future Directions
Bibliography

Glossary

Global Lyapunov exponent A global stability measure of the nonlinear dynamic system. It is a long-run average of the exponential growth rate of infinitesimally small initial deviation and is uniquely determined in the ergodic and stationary case. In this sense, this initial value sensitivity measure does not depend on the initial value. A system with positive Lyapunov exponents is considered chaotic for both deterministic and stochastic cases.

Local Lyapunov exponent A local stability measure based on a short-run average of the exponential growth rate of infinitesimally small initial deviations. Unlike the global Lyapunov exponent, this initial value sensitivity measure depends both on the initial value and the horizon for the average calculation. A smaller local Lyapunov exponent implies a better performance at the point of forecast.

Noise amplification In a stochastic system with the additive noise, the effect of shocks can either grow, remain, or die out with the forecast horizon. If the system is nonlinear, this effect depends both on the initial value and size of the shock. For a chaotic system, the degree of noise amplification is so high that it makes the forecast almost identical to the iid forecast within the next few steps ahead.

Nonlinear impulse response function In a stochastic system with the additive noise, the effect of shocks on the variable in subsequent periods can be summarized in impulse response functions. If the system is

linear, the impulse response does not depend on the initial value and its shape is proportional to the size of shocks. If the system is nonlinear, however, the impulse response depends on the initial value, or the history, and its shape is no longer proportional to the size of shocks.

Definition of the Subject

Empirical studies show that there are at least some components in future asset returns that are predictable using information that is currently available. When the linear time series models are employed in prediction, the accuracy of the forecast does not depend on the current return or the initial condition. In contrast, with nonlinear time series models, properties of the forecast error depend on the initial value or the history. The effect of the difference in initial values in a stable nonlinear model, however, usually dies out quickly as the forecast horizon increases. For both deterministic and stochastic cases, the dynamic system is chaos if a small difference in the initial value is amplified at an exponential rate. In a chaotic nonlinear model, the reliability of the forecast can decrease dramatically even for a moderate forecast horizon. Thus, the knowledge of the sensitive dependence on initial conditions in a particular financial time series offers practically useful information on its forecastability. The most frequently used measure of initial value sensitivity is the largest Lyapunov exponent, defined as the long-run average growth rate of the difference between two nearby trajectories. It is a global initial value sensitivity measure in the sense that it contains the information on the global dynamic property of the whole system. The dynamic properties around a single point in the system can be also described using other local measures. Both global and local measures of the sensitive dependence on initial conditions can be estimated nonparametrically from data without specifying the functional form of the nonlinear autoregressive model.

Introduction

When the asset market is efficient, all the information contained in the history of the asset price is already reflected in the current price of the asset. Mathematically, the conditional mean of asset returns becomes independent of the conditioning information set, and thus price changes must be unpredictable (a martingale property). A convenient model to have such a characteristic is a random walk model with independent and identically distributed (iid) increments given by

$$\ln P_t - \ln P_{t-1} = x_t$$

for $t = 0, 1, 2, \ldots$, where P_t is the asset price and x_t is an iid random variable with mean μ_x and variance σ_x^2. When $\mu_x = 0$, the model becomes a random walk without drift, otherwise, it is a random walk with drift μ_x.

Chaos is a nonlinear deterministic process that can generate a random-like fluctuation. In principle, if a purely deterministic model, instead of a random walk process, is used to describe the dynamics of the asset return x_t, all future asset returns should be completely predictable. However, in the case of chaos, a small perturbation can make the performance of a few steps ahead forecast almost identical to that of a random walk forecast. A leading example is the tent map:

$$x_t = 1 - |2x_{t-1} - 1|$$

with some initial value x_0 between 0 and 1. This map almost surely has the uniform distribution $U(0, 1)$ as its natural measure, defined as the distribution of a typical trajectory of x_t. This dynamic system thus provides aperiodic trajectory or random-like fluctuation of x_t as the number of iteration increases. By introducing a randomness in the initial value x_0, marginal distribution of x_t approaches the natural measure. This property, referred to as ergodicity, implies that the temporal average of any smooth function of a trajectory x_t, $M^{-1}\sum_{t=0}^{M-1} h(x_t)$, converges to a mathematical expectation $E[h(x_t)] = \int_{-\infty}^{\infty} h(x)\pi(x)\mathrm{d}x$ as M tends to infinity, where the marginal distribution of x_t is expressed in terms of the probability density function (pdf) $\pi(x)$. The natural measure $U(0, 1)$ is also a stationary or invariant distribution since the marginal distribution of x_t for any $t \geq 1$, is $U(0, 1)$ whenever initial value x_0 follows $U(0, 1)$. In this case, the mean μ_x and variance σ_x^2 are 1/2 and 1/12, respectively. Furthermore, almost all the trajectories are second-order white noise in the sense that they have a flat spectrum and zero autocorrelation at all leads and lags.

Therefore, the knowledge of the marginal distribution $\pi(x)$ or spectrum of asset returns cannot be directly used to distinguish between the case of a random walk combined with an iid random variable and the case of the tent map generating the returns. Yet, the two cases have significantly different implications on the predictability of asset returns, at least for the extremely short horizon. When the initial value x_0 is given, using $\mu_x = 1/2$ as a one-period-ahead forecast at $t = 0$ provides a minimum mean square forecast error (MSFE) of $\sigma_x^2 = 1/2$ for the former case. In contrast, using $1 - |2x_0 - 1|$ as the forecast gives zero forecast error for the latter case. With a very tiny perturbation, however, the MSFE of the multiple-period-ahead forecast for the latter case quickly approaches $\sigma_x^2 = 1/2$, which is identical to that of the random walk case.

Another example is the logistic map:

$$x_t = 4x_{t-1}(1 - x_{t-1})$$

with some initial value x_0 between 0 and 1. This map again provides chaotic fluctuation with the natural measure almost surely given by beta distribution $B(1/2, 1/2)$. Provided the same distribution as its stationary or invariant distribution, the mean μ_x and variance σ_x^2 are 1/2 and 1/8, respectively (see [37] for the invariant distribution of the logistic map in general). Again, the random walk model combined with an iid random variable with the same marginal distribution $B(1/2, 1/2)$ is not distinguishable from the logistic map based only on the marginal distribution nor spectra. But the two have very different predictive implications.

The key feature of chaos that is not observed in the iid random variable is the sensitivity of the trajectories to the choice of initial values. This sensitivity can be measured by the Lyapunov exponent which is defined as the average rate of divergence (or convergence) of two nearby trajectories. Indeed, the positivity of the Lyapunov exponent in a bounded dissipative nonlinear system is a widely used formal definition of chaos. To derive this measure for the two examples above, first consider a one-dimensional general nonlinear system

$$x_t = f(x_{t-1})$$

where $f: R \to R$ is a continuously differentiable map, with two initial values $x_0 = \overline{x}_0$ and $x_0^* = \overline{x}_0 + \delta$ where δ represents infinitesimal difference in the initial condition. When the distance between two trajectories $\{x_t\}_{t=0}^{\infty}$ and $\{x_t^*\}_{t=0}^{\infty}$ after M steps is measured by the exponential growth rate of δ using $x_M^* - x_M = \delta \exp(M\lambda_M(\overline{x}_0))$, the average of the growth rate in each iteration is given by

$$\lambda_M(\overline{x}_0) = \frac{1}{M} \ln \left| \frac{x_M^* - x_M}{\delta} \right| .$$

Further, let $f^{(M)}$ be the M-fold composition of f. Then from the first order term in the Taylor series expansion of $x_M^* - x_M = f^{(M)}(\overline{x}_0 + \delta) - f^{(M)}(\overline{x}_0)$ around $\overline{x}_0$, combined with the chain rule applied to $df^{(M)}(x)/dx|_{x=\overline{x}_0}$ yields $[f'(x_0)f'(x_1)\cdots f'(x_{M-1})]\delta = [\prod_{t=1}^{M} f'(x_{t-1})]\delta$. Thus, the product $\prod_{t=1}^{M} f'(x_{t-1})$ is the amplifying factor to the initial difference after M periods. Substituting this approximation result into the average growth rate formula yields $\lambda_M(\overline{x}_0) = M^{-1} \sum_{t=1}^{M} \ln |f'(x_{t-1})|$. This measure is called a *local Lyapunov exponent* (of order M) and in general depends on both $\overline{x}_0$ and M (see Fig. 1).

Next, consider the case M tending to infinity. If x_t is ergodic and stationary, $M^{-1} \sum_{t=1}^{M} \ln |f'(x_{t-1})|$ converges

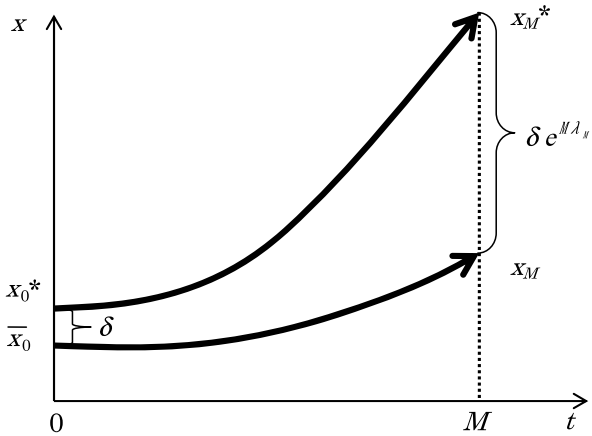

Financial Forecasting, Sensitive Dependence, Figure 1
Lyapunov exponent is an exponential growth rate

to $E[\ln |f'(x_{t-1})|] = \int_{-\infty}^{\infty} \ln |f'(x)|\, \pi(x) dx$, which does not depend on $\overline{x}_0$. Thus, a *global Lyapunov exponent*, or simply a *Lyapunov exponent*, of a one-dimensional system is defined as $\lambda = \lim_{M\to\infty} \lambda_M(\overline{x}_0)$ or

$$\lambda = \lim_{M\to\infty} \frac{1}{M} \sum_{t=1}^{M} \ln \left| f'(x_{t-1}) \right| .$$

According to this definition, the computation of the Lyapunov exponent of the tent map is straightforward. Since the tent map is $x_t = 2x_{t-1}$ for $0 \le x_{t-1} \le 1/2$ and $x_t = 2 - 2x_{t-1}$ for $1/2 < x_{t-1} \le 1$, its first derivative $f'(x_{t-1})$ is 2 for $0 \le x_{t-1} \le 1/2$ and -2 for $1/2 < x_{t-1} \le 1$. Using the uniform distribution as its stationary distribution, we have

$$\lambda = \int_0^{1/2} \ln |2|\, dx + \int_{1/2}^{1} \ln |-2|\, dx = \ln 2\ (\approx 0.693) .$$

Similarly, for the logistic map $x_t = ax_{t-1}(1 - x_{t-1})$ with $a = 4$,

$$\lambda = \int_0^1 \frac{\ln |4 - 8x|}{\pi \sqrt{x(1-x)}} dx = \ln 4 - \ln 2 = \ln 2 .$$

Thus, both the tent map and the logistic map with $a = 4$ have a common positive Lyapunov exponent. The value $\ln 2$ implies that, on average, the effect of an initial deviation doubles each time of iteration. Such a rapid rate of divergence is the source of the fact that their trajectories become unpredictable very quickly. Chaos is thus characterized by sensitive dependence on initial conditions measured by a positive Lyapunov exponent.

Let us next consider a simple linear difference equation $x_t = \rho x_{t-1}$ with $|\rho| < 1$. Since its first derivative

$f'(x)$ is a constant ρ, not only the Lyapunov exponent but also the local Lyapunov exponent $\lambda_M(\overline{x}_0)$ does not depend on the initial value $\overline{x}_0$. For example, when $\rho = 0.5$, $\lambda = \lambda_M(\overline{x}_0) = -\ln 2$ (≈ -0.693). The logistic map $x_t = ax_{t-1}(1 - x_{t-1})$, can be either chaotic or stable depending on the choice of a. When $a = 1.5$, all the trajectories converge to a point mass at $x_t = 1/3$, where the first derivative is 1/2 thus $\lambda = -\ln 2$. For these two examples, the system has a common negative Lyapunov exponent. In this case, the effect of the initial condition is short-lived and the system is not sensitive to initial conditions. The value $-\ln 2$ implies that, on average, the effect of initial deviation reduces by one half each time of iteration.

Knowing the Lyapunov exponent of the asset returns, or their transformation, thus offers a useful information regarding the predictability of a financial market. In particular, for a system with sensitive dependence (namely, the one with a positive Lyapunov exponent), the performance of a multiple step forecast can worsen quickly as the forecast horizon increases if there are (i) a small uncertainty about the current value at the time of forecast (observation noise) and/or (ii) a small additive noise in the system (system noise).

Lyapunov Exponent and Forecastability

Lyapunov Spectrum

As a global measure of initial value sensitivity in a multi-dimensional system, the largest Lyapunov exponent and Lyapunov spectrum will first be introduced. For the p-dimensional deterministic nonlinear system,

$$x_t = f\left(x_{t-1}, \ldots, x_{t-p}\right) ,$$

where $f\colon R^p \to R$ is continuously differentiable, the (global) largest Lyapunov exponent of the system is defined as

$$\lambda = \lim_{M\to\infty} \frac{1}{2M} \ln \left|\nu_1(\mathsf{T}'_M \mathsf{T}_M)\right|$$

where $\nu_1(\mathsf{T}'_M \mathsf{T}_M)$ is the largest eigenvalue of $\mathsf{T}'_M \mathsf{T}_M$, and $\mathsf{T}_M = J_{M-1} \cdot J_{M-2} \cdot \cdots \cdot J_0$. Here J_{t-1}'s are Jacobian matrices defined as

$$J_{t-1} = \begin{bmatrix} \Delta f_1(\mathsf{X}_{t-1}) & \Delta f_2(\mathsf{X}_{t-1}) & \cdots & \Delta f_{p-1}(\mathsf{X}_{t-1}) & \Delta f_p(\mathsf{X}_{t-1}) \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

for $t = 1, \ldots, M$, where $\Delta f_j(\mathsf{X}_{t-1}) = \partial f(\mathsf{X}_{t-1})/\partial x_{t-j}$, for $j = 1, \ldots, p$, are partial derivatives of the conditional mean function evaluated at $\mathsf{X}_{t-1} = (x_{t-1}, \ldots, x_{t-p})'$.

Using an analogy to the one-dimensional case, the local Lyapunov exponent can be defined similarly by $\lambda_M(\mathbf{x}) = (2M)^{-1} \ln \left|\nu_1(\mathsf{T}'_M \mathsf{T}_M)\right|$ with initial value $\mathbf{x} = (\overline{x}_0, \overline{x}_{-1}, \ldots, \overline{x}_{-p+1})'$. Note that $(2M)^{-1} \ln \left|\nu_1(\mathsf{T}'_M \mathsf{T}_M)\right|$ reduces to the sum of absolute derivatives in logs used for the one-dimensional case since $(2M)^{-1} \sum_{t=1}^{M} \ln[f'(x_{t-1})^2] = M^{-1} \sum_{t=1}^{M} \ln|f'(x_{t-1})|$.

In the multi-dimensional case, the whole spectrum of Lyapunov exponents can be also considered using ith Lyapunov exponent λ_i, for $i = 1, \ldots, p$, defined by replacing the largest eigenvalue ν_1 with the ith largest eigenvalue ν_i. A set of all Lyapunov exponents is called the Lyapunov spectrum. Geometrically, each Lyapunov exponent represents the rate of growth (or contraction) of the corresponding principal axis of a growing (or shrinking) ellipsoid. An attracting set of a dynamic system, or simply the attractor, is defined as the set to which x_t approaches in the limit. The attractor can be a point, a curve, a manifold, or more complicated set. The Lyapunov spectrum contains information on the type of the attractor. For example, a system with all negative Lyapunov exponents has an equilibrium point as an attracting set. To understand this claim, let $\mathbf{x}_{\mathrm{EQ}}$ be an equilibrium point and consider a small initial deviation $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)'$ from $\mathbf{x}_{\mathrm{EQ}}$. By the linearization of $f\colon R^p \to R$ at $\mathbf{x}_{\mathrm{EQ}}$, the deviation from $\mathbf{x}_{\mathrm{EQ}}$ after M periods is approximated by $c_1\eta_1 \exp\{\widetilde{\lambda}_1 M\} + \cdots + c_p\eta_p \exp\{\widetilde{\lambda}_p M\}$, where $\widetilde{\lambda}_i$'s and η_i's are the eigenvalues and eigenvectors of J_{EQ}, respectively, where J_{EQ} is J_{t-1} evaluated at $\mathsf{X}_{t-1} = \mathbf{x}_{\mathrm{EQ}}$, and c_i's are scalar constants. The real part of $\widetilde{\lambda}_i$, denoted by $\mathrm{Re}[\widetilde{\lambda}_i]$, represents the rate of growth (contraction) around the equilibrium point $\mathbf{x}_{\mathrm{EQ}}$ along the direction of η_i if $\mathrm{Re}[\widetilde{\lambda}_i]$ is positive (negative). Thus if $\mathrm{Re}[\widetilde{\lambda}_i] < 0$ for all $i = 1, \ldots, p$, $\mathbf{x}_{\mathrm{EQ}}$ is asymptotically stable and is an attractor. Otherwise, $\mathbf{x}_{\mathrm{EQ}}$ is either unstable with $\mathrm{Re}[\widetilde{\lambda}_i] > 0$ for all i, or a saddle point with $\mathrm{Re}[\widetilde{\lambda}_i] > 0$ for some i, provided that none of $\mathrm{Re}[\widetilde{\lambda}_i]$ is zero. In this simple case, ith Lyapunov exponent λ_i corresponds to $\mathrm{Re}[\widetilde{\lambda}_i]$.

Among all the Lyapunov exponents, the largest Lyapunov exponent λ_1, or simply λ, is a key measure to distinguish chaos from other stable systems. By using an analogy to the equilibrium point example, $\lambda > 0$ implies the expansion in the direction of η_1. An attractor requires that the sum of all the Lyapunov exponents be negative since contraction on the whole must be stronger than the expansion. When this condition is met with some positive λ_i's, the system is said to have a strange attractor. Chaos is thus excluded if the largest Lyapunov exponent is not positive.

Financial Forecasting, Sensitive Dependence, Table 1
Lyapunov spectrum and attractors

Attractor	Point	Closed curve	k-torus	Strange attractor
Steady state	equilibrium point	limit cycle (periodic)	k-periodic	chaotic
Dimension	0	1	k (integer)	noninteger
Lyapunov exponents	$\lambda_i < 0\ (i = 1, \ldots, p)$	$\lambda_1 = 0$ $\lambda_i < 0\ (i = 2, \ldots, p)$	$\lambda_1 = \cdots = \lambda_k = 0$ $\lambda_i < 0\ (i = k+1, \ldots, p)$	$\lambda_1 > 0$

A system with a zero largest Lyapunov exponent implies that the average distance of two orbits (along some directions) is same as their initial deviation, the property often referred to as Lyapunov stability. A zero largest Lyapunov exponent and strictly negative remaining Lyapunov exponents lead to a system with a limit cycle. If only the first two (k) largest Lyapunov exponents are zero, the system has a two-torus (k-torus) attractor. The types of attractors and their relationship to the signs of Lyapunov exponents are summarized in Table 1.

Entropy and Dimension

In a deterministic system with initial value sensitivity, the information on how quickly trajectories separate on the whole has a crucial implication in the predictability. This is because if two different trajectories which are initially indistinguishable become distinguishable after a finite number of steps, the knowledge of the current state is useful in forecasting only up to a finite times ahead. Kolmogorov entropy of the system measures the rate at which information is produced and has a close relationship to the Lyapunov exponents. In general, the sum of all positive Lyapunov exponents provides an upper bound to Kolmogorov entropy, which contains the information on how quickly trajectories separate on the whole. Under some conditions, both the entropy and the sum become identical (see [22,56]). This fact can intuitively be understood as follows. Suppose a system with k positive Lyapunov exponents and an attractor of size L. Here, the size of an attractor roughly refers to the range of an invariant distribution of an attractor, which becomes unpredictable as a result of magnified small initial deviation of size d. Note that the length of the first k principal axes after M steps of iteration is proportional to $\exp(M \sum_{i=1}^{k} \lambda_i)$. From $d \exp(M \sum_{i=1}^{k} \lambda_i) = L$, the expected time M to reach the size of attractor is given by $(1/\sum_{i=1}^{k} \lambda_i) \ln(L/d)$. This result implies that the larger $\sum_{i=1}^{k} \lambda_i$ becomes, the shorter the period during which the path is predictable.

Lyapunov exponents are also closely related to the notion of dimension designed to classify the type of attractors. An equilibrium point has zero dimension. A limit cycle is one-dimensional since it resembles an interval in a neighborhood of any point. A k-torus is k-dimensional since it locally resembles an open subset of R^k. However, the neighborhood of any point of a strange attractor does not resemble any Euclidean space and does not have integer dimension. Among many possibilities of introducing a non-integer dimension, one can consider the Lyapunov dimension, or Kaplan–Yorke dimension, defined as

$$D_L = k + \frac{1}{|\lambda_{k+1}|} \sum_{i=1}^{k} \lambda_i$$

where λ_i is the ith Lyapunov exponent and k is the largest integer for which $\sum_{i=1}^{k} \lambda_i \geq 0$. This definition provides the dimension of zero for an equilibrium point, one for a limit cycle, and k for a k-torus. For a chaotic example, suppose a three-dimensional system with a positive Lyapunov exponent ($\lambda_1 = \lambda_+ > 0$), a zero Lyapunov exponent ($\lambda_2 = 0$), and a negative Lyapunov exponent ($\lambda_3 = \lambda_- < 0$). The Lyapunov dimension D_L is then given by $2 + \lambda_+/|\lambda_-|$ which is a fraction that lies strictly between 2 and 3 since an attractor should satisfy $\lambda_+ + \lambda_- < 0$. Likewise, in general, the Lyapunov dimension D_L will be a fraction between k and $k+1$ since $\sum_{i=1}^{k} \lambda_i \leq |\lambda_{k+1}|$ always holds by the definition of k (see Fig. 2).

Since the Lyapunov spectrum contains richer information than the largest Lyapunov exponent alone, several empirical studies reported the Lyapunov spectrum [18], or the transformation such as Kolmogorov entropy and Lyapunov dimension [1,2] of financial time series. However, one must be careful on the interpretation of these quantities since their properties under noisy environment is not rigorously established. In addition, it should be noted that some other forms of the entropy and the dimension can be computed without estimating each Lyapunov exponent separately. For example, [36] recommended using an approximation to Kolmogorov entropy, given by

$$K_2 = \lim_{\substack{\delta \to 0 \\ p \to \infty}} \ln\left(\frac{C^p(\delta)}{C^{p+1}(\delta)}\right)$$

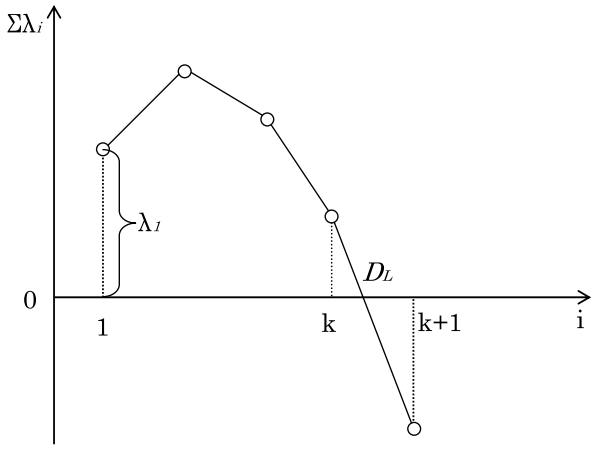

Financial Forecasting, Sensitive Dependence, Figure 2
Lyapunov dimension

where $C^p(\delta)$ is the correlation integral defined by

$$C^p(\delta) = \lim_{T\to\infty} \sharp\{(t,s)|\, \|\mathsf{X}_t - \mathsf{X}_s\| < \delta\}/T^2$$

where $\mathsf{X}_t = (x_t, \ldots, x_{t-p+1})'$ and $\|\cdot\|$ is a vector norm. The approximation given by K_2 provides a lower bound of Kolmogorov entropy (see [22]). The correlation dimension, a type of dimension, can also be defined as

$$D_C = \lim_{\delta\to 0} \frac{\ln C^p(\delta)}{\ln \delta} .$$

Both the K_2 entropy and correlation dimension can be estimated by replacing $C^p(\delta)$ with its sample analogue. In applications to financial time series, these two measures are computed in [30] and [50]. Finally, note that the correlation integral has been used as the basis of the BDS test, a well-established nonlinear dependence test frequently used in economic application, developed by [9]. Formally, the test statistic relies on the sample analogue of $C^p(\delta) - [C^1(\delta)]^p$ and follows normal distribution under the null hypothesis of iid randomness. The BDS test appears to have a good power against the alternative of linear or nonlinear dependence including some low-dimensional chaotic process. Thus, the BDS test is useful in providing the indirect evidence of sensitive dependence and can be complementarily used along with a more direct test based on Lyapunov exponents (see [5] for an example on the comparison between the two approaches).

System Noise and Noisy Chaos

Unlike the data generated from a purely deterministic system, economic and financial data are more likely to be contaminated by noise. There are two main types of random noise used to extend the deterministic model to the stochastic model in the analysis of initial value sensitivity: observation noise and system noise. In the case of the observation noise, or measurement noise, observables are given as the sum of stochastic noise and the unobservables generated from the deterministic model. In contrast, with the system noise, or dynamic noise, observables are generated directly from a nonlinear autoregressive (AR) model. In practice, it is often convenient to introduce the system noise in the additive manner. Theoretically, system noise can make the system to have a unique stationary distribution. Note that for the examples of tent map and logistic map, aperiodic trajectory, or random-like fluctuation, could not be obtained with some choice of initial condition with measure zero. In general, the deterministic system can have infinitely many stationary distributions. However, typically, the presence of additive noise can exclude all degenerate marginal distributions. Furthermore, additive system noise is convenient to generalize the use of the Lyapunov exponents, originally defined in the deterministic system as a measure of sensitive dependence, to the case of a stochastic system.

To see this point, first, consider the following simple linear system with an additive system noise. Adding an iid stochastic error term ε_t, with $E(\varepsilon_t) = 0$ and $E(\varepsilon_t^2) = \sigma^2$, in the previously introduced linear difference equation leads to a linear AR model of order one,

$$x_t = \rho x_{t-1} + \varepsilon_t .$$

The model has a stationary distribution if $|\rho| < 1$. Even if the error term is present, since $f'(x_{t-1}) = \rho$, a one-dimensional Lyapunov exponent can be computed as $\lambda = \ln|\rho| < 0$, the value identical to the case of the deterministic linear difference equation. Thus, the stationarity condition $|\rho| < 1$ in the linear model always implies a negative Lyapunov exponent, while a unit root process $\rho = 1$ implies zero Lyapunov exponent.

Next, consider the introduction of a system noise to a nonlinear system. A general (stationary) nonlinear AR model of order one is defined as

$$x_t = f(x_{t-1}) + \varepsilon_t$$

where $f\colon R \to R$ is a smooth function. For a known unique stationary marginal distribution $\pi(x)$, Lyapunov exponent can be computed as $E[\ln|f'(x_{t-1})|] = \int_{-\infty}^{\infty} \ln|f'(x)|\,\pi(x)\mathrm{d}x$. Thus, by using an analogy of the definition of deterministic chaos, *noisy chaos* can be defined as a stationary nonlinear AR model with a positive Lyapunov exponent. Even if an analytical solution is

not available, the value of Lyapunov exponent is typically obtained numerically or by simulation. Similarly, for the multidimensional nonlinear AR model,

$$x_t = f(x_{t-1}, \ldots, x_{t-p}) + \varepsilon_t ,$$

(noisy) chaos can be defined by a positive largest Lyapunov exponent computed from the Jacobian and the stationary joint distribution of $\mathsf{X}_{t-1} = (x_{t-1}, \ldots, x_{t-p})'$. Furthermore, as long as the process has a stationary distribution, for both the chaotic and non-chaotic case, M-period ahead least squares predictor $f_M(\mathbf{x}) \equiv E[x_{t+M}|\mathsf{X}_t = \mathbf{x}]$ and its conditional MSFE $\sigma_M^2(\mathbf{x}) \equiv E[\{x_{t+M} - f_M(\mathbf{x})\}^2|\mathsf{X}_t = \mathbf{x}]$ depend on the initial condition $\mathbf{x} = (\overline{x}_0, \overline{x}_{-1}, \ldots, \overline{x}_{-p+1})'$ but do not depend on the timing of forecast t.

Noise Amplification

The next issue involves the prediction in the stochastic dynamic system. When additive noise is present in the nonlinear system, the amplification of noise can depend on the initial values and is not necessarily monotonic in horizon. This feature is not unique to the chaotic model but holds for general nonlinear models. However, a small noise is expected to be amplified rapidly in time if the nonlinear system is chaotic.

To understand the process of noise amplification, consider the previously introduced linear AR model of order one with a non-zero coefficient ρ and an initial condition $x_0 = \overline{x}_0$. Then, at the period M,

$$\begin{aligned} x_M &= \rho\{\rho x_{M-2} + \varepsilon_{M-1}\} + \varepsilon_M \\ &= \rho^2 x_{M-2} + \rho\varepsilon_{M-1} + \varepsilon_M \\ &= \rho^M \overline{x}_0 + \varepsilon_M + \cdots + \rho^{M-1}\varepsilon_1 . \end{aligned}$$

Since $\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_M\}$ are not predictable at period 0, the least square M-period ahead predictor is $\rho^M \overline{x}_0$ with its MSFE σ_M^2 given by $\mu_M \sigma^2$ where

$$\mu_M = 1 + \cdots + \rho^{2(M-1)} = 1 + \sum_{j=1}^{M-1} \rho^{2j}$$

is a monotonically increasing proportional factor that does not depend on $\overline{x}_0$. Since $\mu_M > 1$, MSFE is strictly greater than the variance of the noise for all M. However, for a stationary process with $|\rho| < 1$, increments in such a noise amplification become smaller and μ_M converges to $1/(1-\rho^2)$ as M tends to infinity. Thus, eventually, the MSFE converges to the unconditional variance $\sigma_x^2 = \sigma^2/(1-\rho^2)$. In a special case with $\rho = 0$, when the asset price have iid increments, the proportional factor becomes 1 for all M giving its MSFE $\sigma_M^2 = \sigma_x^2 = \sigma^2$ for all M.

Suppose, instead, a general nonlinear AR model of order one with an initial condition $x_0 = \overline{x}_0$. In addition, let $|\varepsilon_t| \le \zeta$ almost surely, where $\zeta > 0$ is a small constant. By Taylor series expansion, for $M \ge 1$,

$$\begin{aligned} x_M &= f\{f(x_{M-2}) + \varepsilon_{M-1}\} + \varepsilon_M \\ &= f^{(2)}(x_{M-2}) + f'\{f(x_{M-2})\}\varepsilon_{M-1} + \varepsilon_M + O(\zeta^2) . \end{aligned}$$

Using the fact that $x_{M-2} = f^{(M-2)}(\overline{x}_0) + O(\zeta)$, and repeating applications of Taylor series expansion,

$$\begin{aligned} x_M &= f^{(2)}(x_{M-2}) + f'\{f^{(M-1)}(\overline{x}_0)\}\varepsilon_{M-1} + \varepsilon_M + O(\zeta^2) \\ &= f^{(M)}(\overline{x}_0) + \varepsilon_M + f'\{f^{(M-1)}(\overline{x}_0)\}\varepsilon_{M-1} + \cdots \\ &\quad + \prod_{k=1}^{M-1} f'\{f^{(k)}(\overline{x}_0)\}\varepsilon_1 + O(\zeta^2) . \end{aligned}$$

Thus the least square M-period ahead predictor is $f^{(M)}(\overline{x}_0)$ with its conditional MSFE given by

$$\sigma_M^2(\overline{x}_0) = \mu_M(\overline{x}_0)\sigma^2 + O(\zeta^3)$$

where

$$\mu_M(\overline{x}_0) = 1 + \sum_{j=1}^{M-1} \left[\prod_{k=j}^{M-1} f'\{f^{(k)}(\overline{x}_0)\} \right]^2 .$$

A comparison of μ_M for the linear model and $\mu_M(\overline{x}_0)$ for the nonlinear model provides some important features of the nonlinear prediction. First, unlike the linear case, the proportional factor now depends not only on the forecast horizon M but also on the initial condition $\overline{x}_0$. Thus, in general, performance of the nonlinear prediction depends on where you are.

Second, $\mu_M(\overline{x}_0)$ does not need to be monotonically increasing with M in nonlinear case. The formula for $\mu_M(\overline{x}_0)$ can be rewritten as

$$\mu_{M+1}(\overline{x}_0) = 1 + \mu_M(\overline{x}_0) f'\{f^{(M)}(\overline{x}_0)\}^2 .$$

Thus, $\mu_{M+1}(\overline{x}_0) < \mu_M(\overline{x}_0)$ is possible when $f'\{f^{(M)}(\overline{x}_0)\}^2 < 1 - 1/\mu_M(\overline{x}_0)$. Therefore, with some initial value and M, the $(M+1)$-period ahead MSFE can be smaller than the M-period ahead MSFE.

Third, and most importantly, unlike the stationary linear model, which imposes the restriction $|\rho| < 1$, $|f'(x)| > 1$ is possible for a large range of values of x in the nonlinear model even if it has a bounded and stationary distribution. In such a case, $\mu_M(\overline{x}_0)$ can grow rapidly

for the moderate or short forecast horizon M. The rapid noise amplification makes the long-horizon forecast very unreliable especially when the model is chaotic. To see this point, it is convenient to rewrite the proportional factor $\mu_M(\overline{x}_0)$ in terms of the local Lyapunov exponent as

$$\mu_M(\overline{x}_0) = 1 + \sum_{j=1}^{M-1} \exp\left\{2(M-j)\lambda_{M-j}(f^{(j)}(\overline{x}_0))\right\} .$$

When the local Lyapunov exponent is positive, the proportional factor grows at an exponential rate as M grows. Recall that in the case of iid forecast (random walk forecast in terms of price level), the MSFE σ_M^2 becomes σ_x^2. Likewise, for the chaotic case with infinitesimally small σ^2, the MSFE σ_M^2 reaches σ_x^2 only after a few steps even if the MSFE is close to zero for the one-step ahead forecast. Thus, the global Lyapunov exponent or other local measures of sensitive dependence contain important information on the predictability in the nonlinear time series framework.

Nonparametric Estimation of the Global Lyapunov Exponent

Local Linear Regression

The measures of initial value sensitivity can be computed from the observed data. Since the Lyapunov exponent is by definition the average growth rate of initial deviations between two trajectories, it can be directly computed by finding pairs of neighbors and then averaging growth rates of the subsequent deviations of such pairs [77]. This 'direct' method, however, provides a biased estimator when there is a random component in the system [51]. A modified regression method proposed by [63] is considered more robust to the presence of measurement noise but not necessarily when the system noise is present. A natural approach to compute the Lyapunov exponent in the nonlinear AR model framework is to rely on the estimation of the nonlinear conditional mean function $f\colon R^p \to R$. For example, based on an argument similar to the deterministic case, the noisy logistic map, $x_t = ax_{t-1}(1-x_{t-1}) + \varepsilon_t$, can be either chaotic or stable depending on the value of the parameter a. The first derivative $f'(x) = a - 2ax$ can be evaluated at each data point once an estimate of a is provided. Thus, the parametric approach in the estimation of Lyapunov exponents has been considered in some cases (e. g., [7]). In practice, however, information on the functional form is rarely available and the nonparametric approach is a reasonable alternative. In principle, any nonparametric estimator can be used to estimate the function f and its partial derivatives in the nonlinear AR model,

$$x_t = f\left(x_{t-1}, \ldots, x_{t-p}\right) + \varepsilon_t$$

where f is smooth and ε_t is a martingale difference sequence with $E[\varepsilon_t | x_{t-1}, x_{t-2}, \ldots] = 0$ and $E[\varepsilon_t^2 | x_{t-1}, x_{t-2}, \ldots] = \sigma^2(x_{t-1}, \ldots, x_{t-p}) = \sigma^2(\mathbf{x})$. To simplify the discussion, here, the one based on a particular type of the kernel regression estimator is explained in detail. Methods based on other types of nonparametric estimators will be later mentioned briefly (see, for example, [27], on the nonparametric approach in time series analysis).

The local linear estimator of the conditional mean function and its first partial derivatives at a point $\mathbf{x}$ can be obtained by minimizing the weighted least squares criterion $\sum_{t=1}^{T}(x_t - \beta_0 - \beta_1'(\mathsf{X}_{t-1} - \mathbf{x}))^2 K_H(\mathsf{X}_{t-1} - \mathbf{x})$, where H is the $d \times d$ bandwidth matrix, K is d-variate kernel function such that $\int K(u)du = 1$, and $K_H(u) = |H|^{-1/2} K(H^{-1/2}u)$. For example, the standard p-variate normal density

$$K(u) = \frac{1}{2\pi^{-p/2}} \exp(-||u||^2/2)$$

with H given by hI_p where h is a scalar bandwidth and I_p is an identity matrix of order p, can be used in the estimation. The solution to the minimization problem is given by $\widehat{\beta}(x) = (\mathsf{X}_x' \mathbf{W}_x \mathsf{X}_x)^{-1} \mathsf{X}_x' \mathbf{W}_x \mathbf{Y}$ where

$$\mathsf{X}_x = \begin{bmatrix} 1 & (\mathsf{X}_0 - \mathbf{x})' \\ \vdots & \vdots \\ 1 & (\mathsf{X}_{T-1} - \mathbf{x})' \end{bmatrix} ,$$

$\mathbf{Y} = (x_1, \ldots, x_T)'$ and $W_x = \operatorname{diag}\{K_H(\mathsf{X}_0 - \mathbf{x}), \ldots, K_H(\mathsf{X}_{T-1} - \mathbf{x})\}$. The local linear estimator of the nonlinear function $f(\mathbf{x})$ and its first derivatives $(\partial f)/(\partial x_{t-j})(\mathbf{x})$ for $j = 1, \ldots, p$ are given by $\widehat{\beta}_0(\mathbf{x}) = \widehat{f}(\mathbf{x})$ and

$$\widehat{\beta}_1(\mathbf{x}) = \begin{bmatrix} \widehat{\beta}_{11}(\mathbf{x}) \\ \vdots \\ \widehat{\beta}_{1p}(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \Delta\widehat{f}_1(\mathbf{x}) \\ \vdots \\ \Delta\widehat{f}_p(\mathbf{x}) \end{bmatrix} ,$$

respectively. [22] and [21] proposed a method, known as the 'Jacobian' method, to estimate the Lyapunov exponent by substituting $\Delta f_i(\mathbf{x})$ in the Jacobian formula by its nonparametric estimator $\Delta\widehat{f}_i(\mathbf{x})$. It should be noted that, in general, the "sample size" T used for estimating Jacobian $\widehat{J}_t$ and the "block length" M, which is the number of evaluation points used for estimating the Lyapunov exponent, can be different. Formally, the Lyapunov exponent estima-

tor of λ is given by

$$\widehat{\lambda}_M = \frac{1}{2M} \ln \nu_1 \left(\widehat{\mathsf{T}}'_M \widehat{\mathsf{T}}_M \right),$$

$$\widehat{\mathsf{T}}_M = \prod_{t=1}^{M} \widehat{\mathsf{J}}_{M-t} = \widehat{\mathsf{J}}_{M-1} \cdot \widehat{\mathsf{J}}_{M-2} \cdot \cdots \cdot \widehat{\mathsf{J}}_0 ,$$

where

$$\widehat{\mathsf{J}}_{t-1} = \begin{bmatrix} \Delta\widehat{f}_1(\mathsf{X}_{t-1}) & \Delta\widehat{f}_2(\mathsf{X}_{t-1}) & \cdots & \Delta\widehat{f}_{p-1}(\mathsf{X}_{t-1}) & \Delta\widehat{f}_p(\mathsf{X}_{t-1}) \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix},$$

for $t = 0, 1, \ldots, M-1$, where $\Delta\widehat{f}_j(\mathbf{x})$ is a nonparametric estimator of $\Delta f_j(\mathbf{x}) = \partial f(\mathbf{x})/\partial x_{t-j}$ for $j = 1, \ldots, p$.

As an estimator for the global Lyapunov exponent, setting $M = T$ gives the maximum number of Jacobians and thus the most accurate estimation can be expected. Theoretically, however, it is often convenient to have a block length M smaller than T. For a fixed M, with T tends to infinity, $\widehat{\lambda}_M$ is a consistent estimator of the local Lyapunov exponent with initial value $\mathbf{x} = (\overline{x}_0, \overline{x}_{-1}, \ldots, \overline{x}_{-p+1})'$ (see [48]). In case both M and T increase with M/T tends to zero, $\widehat{\lambda}_M$ is still a consistent estimator of the global Lyapunov exponent.

Statistical Inference on the Sign of Lyapunov Exponent

Since the positive Lyapunov exponent is the condition that distinguishes the chaotic process from the stable system without high initial value sensitivity, conducting the inference regarding the sign of the Lyapunov exponent is often of practical interest. For such inference, a consistent standard error formula for $\widehat{\lambda}_M$ is available. Under the condition that M grows at a sufficiently slow rate, a standard error can be computed by $\sqrt{\widehat{\Phi}/M}$ where

$$\widehat{\Phi} = \sum_{j=-M+1}^{M-1} w(j/S_M)\widehat{\gamma}(j)$$

$$\text{with} \quad \widehat{\gamma}(j) = \frac{1}{M} \sum_{t=|j|+1}^{M} \widehat{\eta}_t \widehat{\eta}_{t-|j|} ,$$

$$\widehat{\eta}_t = \widehat{\xi}_t - \widehat{\lambda}_M \quad \text{with} \quad \widehat{\xi}_t = \frac{1}{2} \ln \left(\frac{\nu_1 \left(\widehat{\mathsf{T}}'_t \widehat{\mathsf{T}}_t \right)}{\nu_1 \left(\widehat{\mathsf{T}}'_{t-1} \widehat{\mathsf{T}}_{t-1} \right)} \right)$$

$$\text{for} \quad t \geq 2 \quad \text{and} \quad \widehat{\xi}_1 = \frac{1}{2} \ln \nu_1 \left(\widehat{\mathsf{T}}'_1 \widehat{\mathsf{T}}_1 \right) ,$$

where $w(u)$ and S_M denote a kernel function and a lag truncation parameter, respectively (see [67,68,74]). An example of $w(u)$ is the triangular (Bartlett) kernel given by $w(u) = 1 - |u|$ for $|u| < 1$ and $w(u) = 0$, otherwise. The lag truncation parameter S_M should grow at a rate slower than the rate of M.

The procedure above relies on the asymptotic normality of the Lyapunov exponent estimator. Therefore, if the number of Jacobians, M, is not large, an approximation by the normal distribution may not be appropriate. An alternative approach to computing the standard error is to use the resample methods, such as bootstrapping or subsampling. See [32,35] and [79] for the applications of resampling methods to the evaluation of the global Lyapunov exponent estimates.

Consistent Lag Selection

Performance of the nonparametric Lyapunov exponent estimator is often influenced by the choice of lag length p in the nonlinear AR model when the true lag is not known in practice. To see this point, artificial data is generated from a noisy logistic map with an additive system error given by

$$x_t = a x_{t-1}(1 - x_{t-1}) + \sigma(x_{t-1})\varepsilon_t$$

where $\varepsilon_t \sim$ iid $U(-1/2, 1/2)$ and $\sigma(x_{t-1}) = 0.5 \times \min\{a x_{t-1}(1 - x_{t-1}), 1 - a x_{t-1}(1 - x_{t-1})\}$. Note that the conditional heteroskedasticity function $\sigma(x)$ here ensures that the process x_t is restricted to the unit interval $[0, 1]$. When $a = 4.0$, the system has a positive Lyapunov exponent 0.699. Figure 3 shows an example of a sample path from a deterministic logistic map (left) and a noisy logistic map with the current specification of an error term (right). When $a = 1.5$, the system has a negative Lyapunov exponent -0.699. Table 2 reports the mean and median of nonparametric estimates of Lyapunov exponents using the lags from 1 to 4, $M = T = 50$, based on 1,000 replications.

The simulation results show that overfitting has relatively small effect when the true Lyapunov exponent is positive. On the other hand, in case of negative Lyapunov

Financial Forecasting, Sensitive Dependence, Table 2
Lyapunov exponent estimates when $T = 50$: logistic map

		$p = 1$	$p = 2$	$p = 3$	$p = 4$
Logistic map with $a = 4.0$ (true $\lambda = 0.699$)	Mean	0.694	0.706	0.713	0.720
	Median	0.696	0.704	0.710	0.715
Logistic map with $a = 1.5$ (true $\lambda = -0.699$)	Mean	−0.560	−0.046	0.115	0.179
	Median	−0.661	−0.152	0.060	0.149

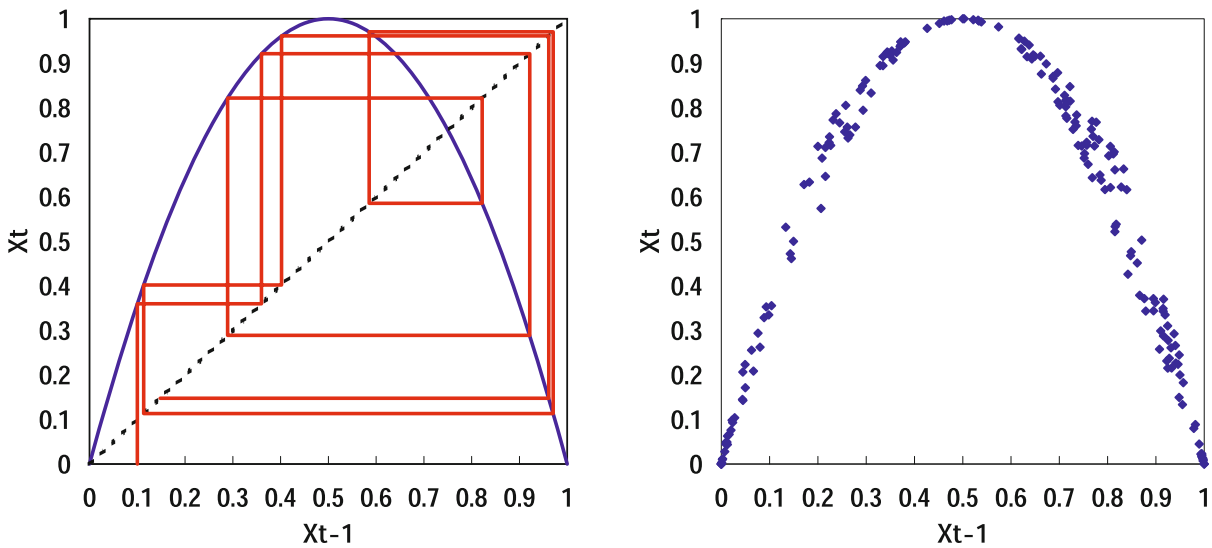

Financial Forecasting, Sensitive Dependence, Figure 3
Logistic map and noisy logistic map

exponent, the upward bias caused by including redundant lags in the nonparametric regression can result in positive Lyapunov exponent estimates. Therefore, when the true lag length of the system is not known, lag selection procedure will be an important part of the analysis of sensitive dependence.

There are several alternative criteria that are designed to select lag length p in the nonparametric kernel autoregressions. With respect to lag selection in the nonparametric analysis of chaos, [15] suggested minimizing the cross-validation (CV) defined by

$$\widehat{CV}(p) = T^{-1}\sum_{t=1}^{T}\left\{x_t - \widehat{f}_{-(t-1)}(\mathrm{X}_{t-1})\right\}^2 W^2(\mathrm{X}_{t-1})$$

where $\widehat{f}_{-(t-1)}(\mathrm{X}_{t-1})$ is the leave-one-out estimator evaluated at X_{t-1} and $W^2(\mathbf{x})$ is a weight function. [70] suggested minimizing the nonparametric version of the final prediction error (FPE) defined by

$$\widehat{\mathrm{FPE}}(p) = T^{-1}\sum_{t=1}^{T}\left\{x_t - \widehat{f}(\mathrm{X}_{t-1})\right\}^2 W^2(\mathrm{X}_{t-1}) + \frac{2}{Th^p}K(0)^p T^{-1}\sum_{t=1}^{T}\{x_t - \widehat{f}(\mathrm{X}_{t-1})\}^2 \cdot W^2(\mathrm{X}_{t-1})/\widehat{\pi}(\mathrm{X}_{t-1})$$

where $\widehat{\pi}(\mathbf{x})$ is a nonparametric joint density estimator at $\mathbf{x}$. [71] proposed a modification to the FPE to prevent overfitting in a finite sample with a multiplicative correction term $\{1 + p(T - p + 1)\}^{-4/(p+4)}$. All three nonparametric criteria, the CV, FPE, and the corrected version of the FPE (CFPE) are proved to be consistent lag selection criteria so that the probability of selecting the correct p converges to one as T increases. Table 3 reports frequencies of selected lags based on these criteria among 1,000 iterations.

The simulation results show that all the lag selection criteria perform reasonably well when the data is generated from a noisy logistic map.

While a noisy logistic map has the nonlinear AR(1) form, it should be informative to examine the performance of the procedures when the true process is the AR model of a higher lag order. [15] considered a nonlinear AR(2) model of the form,

$$x_t = 1 - 1.4x_{t-1}^2 + 0.3x_{t-2} + \varepsilon_t$$

Financial Forecasting, Sensitive Dependence, Table 3
Frequencies of selected lags when $T = 50$: logistic map

		$p = 1$	$p = 2$	$p = 3$	$p = 4$
Logistic map with $a = 4.0$ (true $\lambda = 0.699$)	CV	0.989	0.011	0.000	0.000
	FPE	0.998	0.002	0.000	0.000
	CFPE	1.000	0.000	0.000	0.000
Logistic map with $a = 1.5$ (true $\lambda = -0.699$)	CV	0.697	0.168	0.080	0.055
	FPE	0.890	0.085	0.017	0.008
	CFPE	0.989	0.011	0.000	0.000

Financial Forecasting, Sensitive Dependence, Table 4
Lyapunov exponent estimates when $T = 50$: Hénon map

		$p = 1$	$p = 2$	$p = 3$	$p = 4$
Hénon map (true $\lambda = 0.409$)	Mean	0.411	0.419	0.424	0.431
	Median	0.407	0.423	0.427	0.425

Financial Forecasting, Sensitive Dependence, Table 5
Frequencies of selected lags when $T = 50$: Hénon map

		$p = 1$	$p = 2$	$p = 3$	$p = 4$
Hénon map (true $\lambda = 0.409$)	CV	0.006	0.740	0.250	0.004
	FPE	0.028	0.717	0.253	0.002
	CFPE	0.043	0.762	0.194	0.001

where $\varepsilon_t \sim$ iid $U(-0.01, 0.01)$. This is a noisy Hénon map with a positive Lyapunov exponent, $\lambda = 0.409$. Table 4 shows the mean and median of 1,000 nonparametric estimates of Lyapunov exponents using the lags from 1 to 4, $M = T = 50$, when the data is artificially generated from this higher order noisy chaos process.

As in the finding from a chaotic logistic map example, estimates do not seem to be very sensitive to the choice of lags. The results on the lag selection criteria are provided in Table 5.

The table shows that frequencies of selecting the true lag ($p = 2$) becomes less than in the case of the chaotic logistic map in Table 3. However, the performance of CV improves when it is compared to the case of stable logistic map.

The results from this small-scale simulation exercise show that when the true lag length is not known, combining the automatic lag selection method with Lyapunov exponent estimation is recommended in practice.

Other Nonparametric Estimators

In addition to the class of kernel regression estimators, which includes Nadaraya–Watson, local linear or local polynomial estimators, other estimators have also been employed in the estimation of the Lyapunov exponent. With the kernel regression method, Jacobians are evaluated using a local approximation to the nonlinear function at the lagged point X_{t-1}. Another example of the local smoothing method used in Lyapunov exponent estimation is the local thin-plate splines suggested by [51,54]. The local estimation method, however, is subject to the data sparseness problem in the high-dimensional system. Alternatively, Jacobians can be evaluated using a global approximation to the unknown function. As a global estimation method, a global spline function may be used to smooth all the available sample. However, the most frequently used global method in Lyapunov exponent estimation in practice is the neural nets ([2,18,68], among others). A single hidden-layer, feedforward neural network is given by

$$f(\mathsf{X}_{t-1}) = \beta_0 + \sum_{j=1}^{k} \beta_j \psi(a_j' \mathsf{X}_{t-1} + b_j)$$

where ψ is an activation function (most commonly a logistic distribution function) and k is a number of hidden units. The neural network estimator $\widehat{f}$ can be obtained by minimizing the (nonlinear) least square criterion. Jacobians are then evaluated using the analytical first derivative of neural net function. Compared to other functional approximations, the neural net form is less sensitive to increasing lag length, p. Thus, it has a merit in terms of the effective sample size.

Four Local Measures of Sensitive Dependence

Local Lyapunov Exponent

The global Lyapunov exponent measures the initial value sensitivity of long horizon forecast. For the ergodic and stationary case, this initial value sensitivity measure does not depend on the initial value. By definition, the global Lyapunov exponent is the limit of the local Lyapunov exponent when its order M tends to infinity. Unlike the global Lyapunov exponent, the local Lyapunov exponent is a function of an initial value and thus the initial value sensitivity of the short-term forecast depends on where you are. In this sense, local measures of sensitive dependence contain more detailed information on the predictability in the nonlinear dynamic system.

Recall that both the deterministic tent map and the logistic map with $a = 4.0$ have a common positive Lyapunov exponent 0.693. Thus in terms of long-horizon predictability, two processes have exactly the same degree of initial value sensitivity. Yet, in terms of short term forecast, it is possible that predictability at the same point differs among two processes.

The sign of local Lyapunov exponents of the single process can also be different in some range of initial values. Figure 4 shows the local Lyapunov exponents of the deterministic logistic map with $a = 4.0$ for different values of M. Consistent with the definition, as M grows, it approaches to a flat line at the value of 0.693. However, when M is finite, there is a range of initial values associated with a negative local Lyapunov exponent. Within such a range of initial values, sensitive dependence is low

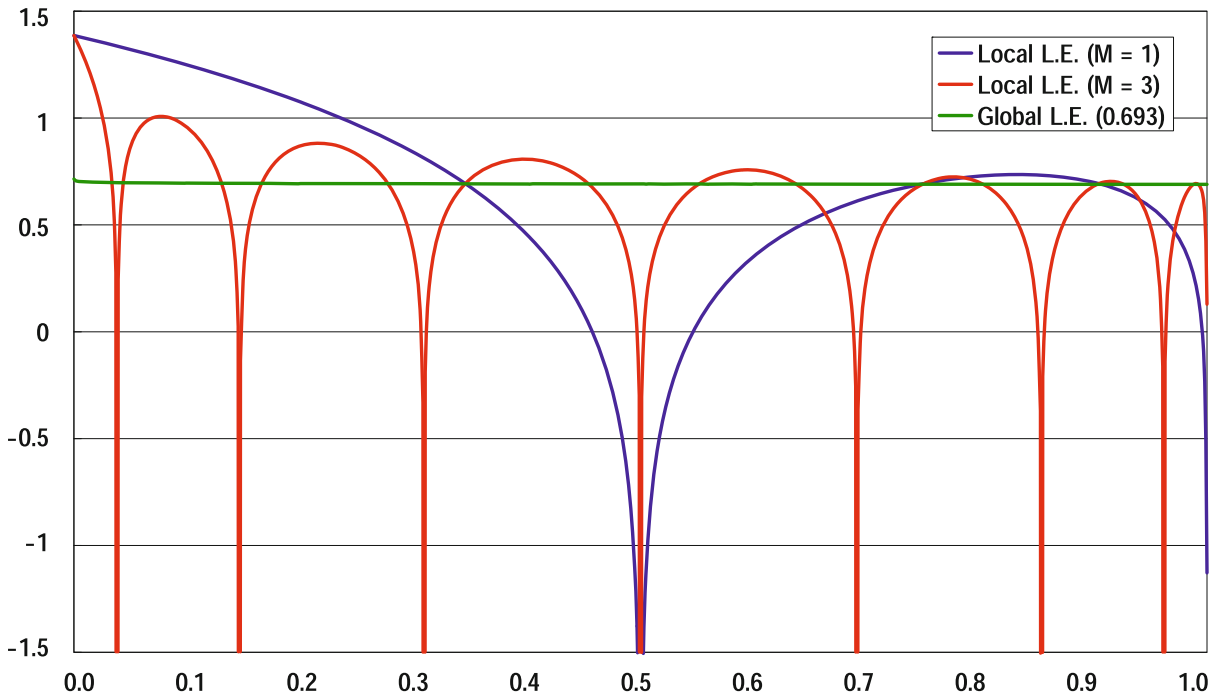

Financial Forecasting, Sensitive Dependence, Figure 4
Local and global Lyapunov exponents of logistic map

and predictability is high even if it is a globally chaotic process.

Analysis of local Lyapunov exponent is also valid in the presence of noise. Studies by [4,48,78], among others, investigate the properties of the local Lyapunov exponent in a noisy system.

The local Lyapunov exponent can be estimated nonparametrically from data using the following procedure. First, obtain the nonparametric Jacobian estimate $\widehat{J}_{t-1}$ for each t using a full sample, as in the case of global Lyapunov exponent estimation. Second, choose a single horizon M of interest. Third, choose the p-dimensional initial value $\mathbf{x} = (x_{t^*}, x_{t^*-1}, \ldots, x_{t^*-p+1})'$ from the data subsequence $\{x_t\}_{t=-p+1}^{T-M}$. Finally, the local Lyapunov exponent estimator at $\mathbf{x}$ is given by $\widehat{\lambda}_M(\mathbf{x}) = (2M)^{-1} \ln \nu_1(\widehat{\mathrm{T}}'_M \widehat{\mathrm{T}}_M)$ where $\widehat{\mathrm{T}}_M = \prod_{t=t^*}^{t^*+M} \widehat{J}_{M-t}$.

While the local Lyapunov exponent is a simple and straightforward local measure of the sensitive dependence, three other useful local measures will be introduced below.

Nonlinear Impulse Response Function

The impulse response function (IRF) is a widely used measure of the persistence effect of shocks in the analysis of economic time series. Here, it is useful to view the IRF as the difference between the two expected future paths: one with and the other without a shock occurred at the current period. When the shock, or the initial deviation, is very small, the notion of impulse responses is thus closely related to the concept of sensitive dependence on initial conditions. To verify this claim, a simple example of a one-dimensional linear IRF is first provided below, followed by the generalization of the IRF to the case of nonlinear time-series model.

For a linear AR model of order one, $x_t = \rho x_{t-1} + \varepsilon_t$, the M-period ahead IRF to a unit shock is defined as

$$\mathrm{IRF}_M = \rho^M.$$

Let $\{x_t^*\}_{t=0}^{\infty}$ be a sample path that contains a single unit shock whereas $\{x_t\}_{t=0}^{\infty}$ is a sample path without any shock. Also let $x_0 = \overline{x}_0$ be an initial condition for the latter path. Then, this linear IRF can be interpreted in two ways. One interpretation is the sequence of the responses to a shock defined to increase one unit of x_0 at time 0 ($x_1 = \rho\overline{x}_0, x_1^* = \rho(\overline{x}_0 + 1), x_1^* - x_1 = \rho, \ldots, x_M^* - x_M = \rho^M$). In this case, the initial value of x_t^* is given as $x_0^* = \overline{x}_0 + 1$, so the shock can be simply viewed as the deviation of two paths at the initial condition. The other interpretation is the sequence of the responses to a shock defined to increase one unit of x_1 at time 1 ($x_1 = \rho\overline{x}_0, x_1^* = \rho\overline{x}_0 + 1$,

$x_1^* - x_1 = 1, \ldots, x_{M+1}^* - x_{M+1} = \rho^M$). In contrast to the first case, two paths have a common initial condition $x_0^* = x_0 = \overline{x}_0$, but the second path is perturbed as if a shock of $\varepsilon_1 = 1$ is realized at time 1 through the dynamic system of $x_t = \rho x_{t-1} + \varepsilon_t$. In either interpretation, however, IRF_M is the difference between x_t^* and x_t at exactly M-period after the shock has occurred and the IRF does not depend on the initial condition $\overline{x}_0$. In addition, the shape of IRF is preserved even if we replace the unit shock with a shock of size δ. The IRF becomes $\rho^M \delta$ and thus the IRF to a unit shock can be considered as a ratio of $\rho^M \delta$ to δ or the normalized IRF.

In the linear framework, the choice between the two interpretations does not matter in practice since the two cases yield exactly the same IRF. However, for nonlinear models, two alternative interpretations lead to different definitions of the IRF. Depending on the objective of the analysis, one may use the former version [31] or the latter version [42,58] of the nonlinear IRFs. The M-period ahead nonlinear impulse response based on the first interpretation considered by [31] is defined as

$$\begin{aligned}\mathrm{IRF}_M(\delta, \mathbf{x}) &= E\left[x_{t+M-1} | \mathsf{X}_{t-1} = \mathbf{x}^*\right] \\ &\quad - E\left[x_{t+M-1} | \mathsf{X}_{t-1} = \mathbf{x}\right] \\ &= E\left[x_M | \mathsf{X}_0 = \mathbf{x}^*\right] - E\left[x_M | \mathsf{X}_0 = \mathbf{x}\right] \\ &= f_M(\mathbf{x}^*) - f_M(\mathbf{x})\end{aligned}$$

where $\mathsf{X}_{t-1} = (x_{t-1}, \ldots, x_{t-p})'$, $\mathbf{x}^* = (\overline{x}_0 + \delta, \overline{x}_{-1}, \ldots, \overline{x}_{-p+1})'$ and $\mathbf{x} = (\overline{x}_0, \overline{x}_{-1}, \ldots, \overline{x}_{-p+1})'$. Unlike the linear IRF, the nonlinear IRF depends on the size of shock δ and the initial condition (or the history) $\mathsf{X}_0 = \mathbf{x}$. Interestingly, the partial derivative $\Delta f_{M,1}(\mathbf{x}) = \partial f_M(\mathbf{x}) / \partial x_{t-1}$ corresponds to normalized IRF (proportional to the nonlinear IRF) for small δ since

$$\begin{aligned}&\lim_{\delta \to 0} \frac{\mathrm{IRF}_M(\delta, \mathbf{x})}{\delta} \\ &= \lim_{\delta \to 0} \frac{f_M(\overline{x}_0 + \delta, \overline{x}_{-1}, \ldots, \overline{x}_{-p+1})}{\delta} \\ &\qquad - \frac{f(\overline{x}_0, \overline{x}_{-1}, \ldots, \overline{x}_{-p+1})}{\delta} = \Delta f_{M,1}(\mathbf{x}) .\end{aligned}$$

In the one-dimensional case, the IRF simplifies to

$$\begin{aligned}\mathrm{IRF}_M(\delta, \overline{x}_0) &= E\left[x_{t+M-1} | x_{t-1} = \overline{x}_0 + \delta\right] \\ &\quad - E\left[x_{t+M-1} | x_{t-1} = \overline{x}_0\right] \\ &= E\left[x_M | x_0 = \overline{x}_0 + \delta\right] - E\left[x_M | x_0 = \overline{x}_0\right] \\ &= f_M(\overline{x}_0 + \delta) - f_M(\overline{x}_0) .\end{aligned}$$

The first derivative $f_M'(x)$, thus corresponds to the IRF to an infinitesimally small deviation since

$$\lim_{\delta \to 0} \frac{\mathrm{IRF}_M(\delta, \overline{x}_0)}{\delta} = \lim_{\delta \to 0} \frac{f_M(\overline{x}_0 + \delta) - f(\overline{x}_0)}{\delta} = f_M'(\overline{x}_0).$$

Recall that $\lambda_M(\overline{x}_0) = M^{-1} \ln |\prod_{t=1}^M f'(x_{t-1})|$. If $\prod_{t=1}^M f'(x_{t-1})$ can be approximated by $f_M'(\overline{x}_0)$, both normalized IRF and the local Lyapunov exponent contain the same information regarding the initial value sensitivity.

Next, based on the second interpretation, IRF can be alternatively defined as

$$\begin{aligned}\mathrm{IRF}_M^*(\delta, \mathbf{x}) &= E\left[x_{t+M-1} | x_t = f(\mathbf{x}) + \delta, \mathsf{X}_{t-1} = \mathbf{x}\right] \\ &\quad - E(x_{t+M-1} | \mathsf{X}_{t-1} = \mathbf{x}) \\ &= E\left[x_M | x_1 = f(\mathbf{x}) + \delta, \mathsf{X}_0 = \mathbf{x}\right] \\ &\quad - E(x_M | \mathsf{X}_0 = \mathbf{x}) \\ &= f_{M-1}(\mathbf{x}^*) - f_M(\mathbf{x})\end{aligned}$$

where $\mathsf{X}_{t-1} = (x_{t-1}, \ldots, x_{t-p})'$ and $\mathbf{x} = (\overline{x}_0, \overline{x}_{-1}, \ldots, \overline{x}_{-p+1})'$ and $\mathbf{x}^* = (f(\mathbf{x}) + \delta, \overline{x}_0, \overline{x}_{-1}, \ldots, \overline{x}_{-p+2})'$. This version of nonlinear IRF is sometimes referred to as the generalized impulse response function [42,58]. Using the fact that

$$f_M(\mathbf{x}) = f_{M-1}\left(f(\mathbf{x}), \overline{x}_0, \overline{x}_{-1}, \ldots, \overline{x}_{-p+2}\right) ,$$

the equivalence of the partial derivative $\Delta f_{M-1,1}(\mathbf{x}) = \partial f_{M-1}(\mathbf{x}) / \partial x_{t-1}$ and the small deviation IRF can be also shown as

$$\begin{aligned}&\lim_{\delta \to 0} \frac{\mathrm{IRF}_M^*(\delta, \mathbf{x})}{\delta} \\ &= \lim_{\delta \to 0} \frac{f_{M-1}\left(f(\mathbf{x}) + \delta, \overline{x}_0, \overline{x}_{-1}, \ldots, \overline{x}_{-p+2}\right)}{\delta} \\ &\quad - \frac{f_{M-1}\left(f(\mathbf{x}), \overline{x}_0, \overline{x}_{-1}, \ldots, \overline{x}_{-p+2}\right)}{\delta} \\ &= \Delta f_{M-1,1}(f(\mathbf{x}), \overline{x}_0, \overline{x}_{-1}, \ldots, \overline{x}_{-p+2}) .\end{aligned}$$

In the one dimensional case, the IRF formula reduces to

$$\begin{aligned}\mathrm{IRF}_M^*(\delta, \overline{x}_0) &= E\left[x_{t+M-1} | x_t = f(\overline{x}_0) + \delta, x_{t-1} = \overline{x}_0\right] \\ &\quad - E(x_{t+M-1} | x_{t-1} = \overline{x}_0) \\ &= E\left[x_M | x_1 = f(\overline{x}_0) + \delta, x_0 = \overline{x}_0\right] \\ &\quad - E(x_M | x_0 = \overline{x}_0) \\ &= E\left[x_{M-1} | x_0 = f(\overline{x}_0) + \delta\right] \\ &\quad - E(x_M | x_0 = \overline{x}_0) \\ &= f_{M-1}(f(\overline{x}_0) + \delta) - f_M(\overline{x}_0) \\ &= f_{M-1}(f(\overline{x}_0) + \delta) - f_{M-1}(f(\overline{x}_0)) .\end{aligned}$$

Similarly, the small deviation IRF is given by

$$\lim_{\delta\to 0}\frac{\mathrm{IRF}^*_M(\delta,\overline{x}_0)}{\delta} = \lim_{\delta\to 0}\frac{f_{M-1}(f(\overline{x}_0)+\delta)-f_{M-1}(f(\overline{x}_0))}{\delta} = f'_{M-1}(f(\overline{x}_0))\,.$$

The nonlinear impulse response function can be estimated nonparametrically without specifying the functional form by an analogy to Lyapunov exponent estimation (see [72] and [66]). Instead of minimizing $\sum_{t=1}^{T}(x_t-\beta_0-\beta_1'(\mathsf{X}_{t-1}-\mathbf{x}))^2K_H(\mathsf{X}_{t-1}-\mathbf{x})$, the local linear estimator of M-period ahead predictor $f_M(\mathbf{x})$ and its partial derivatives $(\partial f_M)/(\partial x_{t-j})(\mathbf{x})$ for $j=1,\ldots,p$ can be obtained by minimizing, $\sum_{t=1}^{T-M+1}(x_{t+M-1}-\beta_{M,0}-\beta'_{M,1}(\mathsf{X}_{t-1}-\mathbf{x}))^2K_H(\mathsf{X}_{t-1}-\mathbf{x})$, or $\widehat{\beta}_{M,0}(\mathbf{x})=\widehat{f}_M(\mathbf{x})$ and

$$\widehat{\beta}_{M,1}(\mathbf{x})=\begin{bmatrix}\widehat{\beta}_{M,11}(\mathbf{x})\\ \vdots\\ \widehat{\beta}_{M,1p}(\mathbf{x})\end{bmatrix}=\begin{bmatrix}\Delta\widehat{f}_{M,1}(\mathbf{x})\\ \vdots\\ \Delta\widehat{f}_{M,p}(\mathbf{x})\end{bmatrix},$$

respectively, where $\widehat{\beta}_M(x)=(\widehat{\beta}_{M,0}(\mathbf{x}),\widehat{\beta}_{M,1}(\mathbf{x})')'=(\mathsf{X}_x'\mathbf{W}_x\mathsf{X}_x)^{-1}\mathsf{X}_x'\mathbf{W}_x\mathbf{Y}$,

$$\mathsf{X}_x=\begin{bmatrix}1 & (\mathsf{X}_0-\mathbf{x})'\\ \vdots & \vdots\\ 1 & (\mathsf{X}_{T-M}-\mathbf{x})'\end{bmatrix},$$

$\mathbf{Y}=(x_M,\ldots,x_T)'$ and $W_x=\mathrm{diag}\,\{K_H(\mathsf{X}_0-\mathbf{x}),\ldots,K_H(\mathsf{X}_{T-M}-\mathbf{x})\}$. The local linear estimator of the IRF is then given by

$$\widehat{\mathrm{IRF}}_M(\delta,\mathbf{x})=\widehat{f}_M(\mathbf{x}^*)-\widehat{f}_M(\mathbf{x})$$

where $\mathbf{x}^*=(\overline{x}_0+\delta,\overline{x}_{-1},\ldots,\overline{x}_{-p+1})'$ and $\mathbf{x}=(\overline{x}_0,\overline{x}_{-1},\ldots,\overline{x}_{-p+1})'$. Similarly, the estimator of the alternative IRF is given by

$$\widehat{\mathrm{IRF}}^*_M(\delta,\mathbf{x})=\widehat{f}_{M-1}(\mathbf{x}^*)-\widehat{f}_M(\mathbf{x})$$

where $\mathbf{x}^*=(\widehat{f}(\mathbf{x})+\delta,\overline{x}_0,\overline{x}_{-1},\ldots,\overline{x}_{-p+2})'$ and $\mathbf{x}=(\overline{x}_0,\overline{x}_{-1},\ldots,\overline{x}_{-p+1})'$. When $\mathbf{x}$ and δ are given, computing nonparametric IRFs for a sequence of M provide a useful information on the persistence of deviation without specifying the autoregressive function. However, instead of reporting IRFs for many possible combinations of $\mathbf{x}$ and δ, one can also compute the small deviation IRF based on the nonparametric estimate of the first partial derivative at $\mathbf{x}$. The local linear estimator of the small deviation IRF is given by $\Delta\widehat{f}_{M,1}(\mathbf{x})$ for the first version, and $\Delta\widehat{f}_{M-1,1}(\widehat{f}(\mathbf{x}),\overline{x}_0,\overline{x}_{-1},\ldots,\overline{x}_{-p+2})$ for the second version, respectively. A large change in the value of derivatives with increasing M represents the sensitive dependence on initial conditions.

Yao and Tong's Variance Decomposition

The initial value sensitivity of the system with dynamic noise also has an implication in the presence of additional observation noise. Suppose that current observation is subject to a measurement error, a rounding error, or when only preliminary estimates of aggregate economic variables announced by the statistical agency are available. When the true current position deviates slightly from $\mathbf{x}=(\overline{x}_0,\overline{x}_{-1},\ldots,\overline{x}_{-p+1})'$ by $\boldsymbol{\delta}=(\delta_1,\ldots,\delta_p)'$, the performance of the same predictor may be measured by $E[\{x_{t+M}-f_M(\mathbf{x})\}^2|\mathsf{X}_t=\mathbf{x}+\boldsymbol{\delta}]$. Under a certain condition, this MSFE can be decomposed as follows:

$$\begin{aligned}E\left[\{x_{t+M}-f_M(\mathbf{x})\}^2|\mathsf{X}_t=\mathbf{x}+\boldsymbol{\delta}\right]&=\sigma_M^2(\mathbf{x}+\boldsymbol{\delta})+\{f_M(\mathbf{x}+\boldsymbol{\delta})-f_M(\mathbf{x})\}^2\\&=\sigma_M^2(\mathbf{x}+\boldsymbol{\delta})+\{\boldsymbol{\delta}'\Delta f_M(\mathbf{x})\}^2+o(||\boldsymbol{\delta}||^2)\end{aligned}$$

where $\Delta f_M(\mathbf{x})=(\Delta f_{M,1}(\mathbf{x}),\ldots,\Delta f_{M,p}(\mathbf{x}))'$, $\Delta f_{M,j}(\mathbf{x})=(\partial f_M)/(\partial x_{t-j})(\mathbf{x})$ for $j=1,\ldots,p$. This decomposition shows two dominant components in the MSFE. The first component represents the prediction error caused by the randomness in the system at point $\mathbf{x}+\boldsymbol{\delta}$. This component will be absent in the case where there is no dynamic noise ε_t in the system. The second component represents the difference caused by the deviation $\boldsymbol{\delta}$ from the initial point $\mathbf{x}$. When the non-zero deviation $\boldsymbol{\delta}$ is much smaller than σ, the standard deviation of ε_t, the first component $\sigma_M^2(\mathbf{x}+\boldsymbol{\delta})=\sigma_M^2(\mathbf{x})+O(||\boldsymbol{\delta}||)$ is the dominant term because the second component $\{\boldsymbol{\delta}'\Delta f_M(\mathbf{x})\}^2$ is of order $O(||\boldsymbol{\delta}||^2)$. However, for a nonlinear system with a very small error ε_t, the contribution of the second term can become nonnegligible. Thus, [80] considered $\Delta f_M(\mathbf{x})$ as a measure of sensitivity to initial conditions for the M-period ahead forecast (They referred to the M-step Lyapunov-like index).

If $f_M(\mathbf{x})$ is replaced by a mean square consistent estimator $\widehat{f}_M(\mathbf{x})$, such as a local linear estimator $\widehat{\beta}_{M,0}(\mathbf{x})$,

$$\begin{aligned}\lim_{T\to\infty}E\left[\{x_{T+M}-\widehat{f}_M(\mathbf{x})\}^2|\mathsf{X}_T=\mathbf{x}+\boldsymbol{\delta}\right]&\\=\sigma_M^2(\mathbf{x}+\boldsymbol{\delta})+\{\boldsymbol{\delta}'\Delta f_M(\mathbf{x})\}^2+o(||\boldsymbol{\delta}||^2)\,.&\end{aligned}$$

Thus the decomposition is still valid. For the estimation of the sensitivity measure $\Delta f_M(\mathbf{x})$, the local linear estimator $\widehat{\beta}_{M,0}(\mathbf{x})=\Delta\widehat{f}_M(\mathbf{x})$ can be used. In practice, it

is convenient to consider a norm version of the measure $\mathrm{LI}_M(\mathbf{x}) = ||\Delta f_M(\mathbf{x})||$ and report its estimator

$$\widehat{\mathrm{LI}}_M(\mathbf{x}) = ||\Delta \widehat{f}_M(\mathbf{x})||$$

evaluated at various $\mathbf{x}$. In a one-dimensional case, they are $\mathrm{LI}_M(\overline{x}_0) = |f'_M(\overline{x}_0)|$ and $\widehat{\mathrm{LI}}_M(\overline{x}_0) = |\widehat{f'}_M(\overline{x}_0)|$, respectively. Note that $\mathrm{LI}_M(\mathbf{x})$ is related to the derivative of the normalized nonlinear impulse response function $\mathrm{IRF}_M(\delta, \mathbf{x})$. Recall that, in the one-dimensional case, a normalized IRF to infinitesimal shocks becomes the first derivative. Thus, $\mathrm{LI}_M(\overline{x}_0)$ is the absolute value of the estimator of the corresponding IRF. In the multidimensional case, IRF to small shocks becomes the partial derivative with respect to the first components. If shocks are also given to other initial values in IRF, computing the norm of the estimator of all IRFs yields $\mathrm{LI}_M(\mathbf{x})$.

This sensitivity measure is also related to the local Lyapunov exponent. In the one-dimensional case, with a fixed M, the local Lyapunov exponent can be written as $\lambda_M(\overline{x}_0) = M^{-1} \ln |\prod_{t=1}^{M} f'(x_{t-1})|$. If the contribution of ε_t is very small, $\mathrm{d}f^{(M)}(x_0)/\mathrm{d}x \approx \prod_{t=1}^{M} f'(x_{t-1})$ and then the estimator $\widehat{f'}_M(x_0)$ becomes an estimator of $\mathrm{d}f^{(M)}(x_0)/\mathrm{d}x$. Thus $\lambda_M(\overline{x}_0)$ can be also estimated by $M^{-1} \ln \widehat{\mathrm{LI}}_M(\overline{x}_0)$.

Information Matrix

The last measure of the initial value sensitivity is the one based on the distance between two distributions of M-steps ahead forecast, conditional on two nearby initial values $\mathbf{x} = (\overline{x}_0, \overline{x}_{-1}, \ldots, \overline{x}_{-p+1})'$ and $\mathbf{x} + \boldsymbol{\delta}$ where $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)'$. Let $\pi_M(y|\mathbf{x})$ and $\Delta\pi_M(y|\mathbf{x})$ be the conditional density function of x_M given $\mathsf{X}_0 = \mathbf{x}$ and a $p \times 1$ vector of its partial derivatives. [81] suggested using Kullback–Leibler information to measure the distance, which is given by

$$K_M(\delta, \mathbf{x}) = \int_{-\infty}^{+\infty} \left\{\pi_M(y|\mathbf{x} + \boldsymbol{\delta}) - \pi_M(y|\mathbf{x})\right\} \cdot \ln \left\{\pi_M(y|\mathbf{x} + \boldsymbol{\delta})/\pi_M(y|\mathbf{x})\right\} \mathrm{d}y \,.$$

Assuming the smoothness of conditional distribution and interchangeability of integration and differentiation, Taylor series expansion around $\mathbf{x}$ for small $\boldsymbol{\delta}$ yields

$$K_M(\delta, \mathbf{x}) = \boldsymbol{\delta}' I_M(\mathbf{x}) \boldsymbol{\delta} + o(||\boldsymbol{\delta}||^2)$$

where

$$I_M(\mathbf{x}) = \int_{-\infty}^{+\infty} \Delta\pi_M(y|\mathbf{x}) \Delta\pi_M(y|\mathbf{x})'/\pi_M(y|\mathbf{x}) \mathrm{d}y \,.$$

If initial value $\mathbf{x}$ is treated as a parameter vector of the distribution, $I_M(\mathbf{x})$ is the Fisher's information matrix, which represents the information on $\mathbf{x}$ contained in x_M. This quantity can be used as an initial value sensitivity measure since more information on $\mathbf{x}$ implies more sensitivity of distribution of x_M to the initial condition $\mathbf{x}$. This information matrix measure and the M-step Lyapunov-like index are related via the following inequality when the system is one-dimensional,

$$I_M(\overline{x}_0) \geq \frac{\mathrm{LI}_M^2(\overline{x}_0)}{\sigma_M^2(\overline{x}_0)} \,.$$

Thus, for a given M-step Lyapunov-like index, a larger conditional MSFE implies more sensitivity. In addition, because that $\lambda_M(\overline{x}_0) \approx M^{-1} \ln \mathrm{LI}_M(\overline{x}_0)$ and $\sigma_M^2(\overline{x}_0) \approx \sigma^2[1 + \sum_{j=1}^{M-1} \exp\{2(M-j)\lambda_{M-j}(f^{(j)}(\overline{x}_0))\}]$,

$$\ln I_M(\overline{x}_0) \geq 2M\lambda_M(\overline{x}_0) - \ln\left[1 + \sum_{j=1}^{M-1} \exp\left\{2(M-j)\lambda_{M-j}(f^{(j)}(\overline{x}_0))\right\}\right] - \ln \sigma^2$$

holds approximately.

As an alternative to Kullback–Leibler distance, [28] considered L_2-distance given by

$$D_M(\delta, \mathbf{x}) = \int_{-\infty}^{+\infty} \{\pi_M(y|\mathbf{x} + \boldsymbol{\delta}) - \pi_M(y|\mathbf{x})\}^2 \mathrm{d}y \,.$$

Because of a similar argument, for small $\boldsymbol{\delta}$, $D_M(\delta, \mathbf{x})$ can be approximated by

$$D_M(\delta, \mathbf{x}) = \boldsymbol{\delta}' J_M(\mathbf{x}) \boldsymbol{\delta} + o(||\boldsymbol{\delta}||^2)$$

where

$$J_M(\mathbf{x}) = \int_{-\infty}^{+\infty} \Delta\pi_M(y|\mathbf{x}) \Delta\pi_M(y|\mathbf{x})' \mathrm{d}y \,.$$

Note that $J_M(\mathbf{x})$ cannot be interpreted as Fisher's information but can still be used as a sensitivity measure.

Both $I_M(\mathbf{x})$ and $J_M(\mathbf{x})$ can be estimated nonparametrically. Consider the minimization problem of $\sum_{t=1}^{T-M+1} (\kappa_h(x_{t+M-1} - y) - \beta_{M,0} - \beta'_{M,1}(\mathsf{X}_{t-1} - \mathbf{x}))^2 K_H(\mathsf{X}_{t-1} - \mathbf{x})$ where $\kappa_h(u) = \kappa(u/h)/h$, h is the bandwidth and κ is a univariate kernel function, instead of minimizing $\sum_{t=1}^{T-M+1} (x_{t+M-1} - \beta_{M,0} - \beta'_{M,1}(\mathsf{X}_{t-1} - \mathbf{x}))^2 K_H(\mathsf{X}_{t-1} - \mathbf{x})$. Then, $\widehat{\beta}_{M,0}(\mathbf{x}, y) = \widehat{\pi}_M(y|\mathbf{x})$ and $\widehat{\beta}_{M,1}(\mathbf{x}, y) = \Delta\widehat{\pi}_M(y|\mathbf{x})$, where $\widehat{\beta}_M(\mathbf{x}, y) = (\widehat{\beta}_{M,0}(\mathbf{x}, y), \widehat{\beta}_{M,1}(\mathbf{x}, y)')' = (\mathsf{X}'_x \mathbf{W}_x \mathsf{X}_x)^{-1} \mathsf{X}'_x \mathbf{W}_x \mathbf{Y}_y$,

$$\mathsf{X}_x = \begin{bmatrix} 1 & (\mathsf{X}_0 - \mathbf{x})' \\ \vdots & \vdots \\ 1 & (\mathsf{X}_{T-M} - \mathbf{x})' \end{bmatrix},$$

$\mathbf{Y}_y = \{\kappa_h(x_M - y), \ldots, \kappa_h(x_T - y)\}'$ and $W_x = \mathrm{diag}\{K_H(\mathrm{X}_0 - \mathbf{x}), \ldots, K_H(\mathrm{X}_{T-M} - \mathbf{x})\}$. Then the estimators of $I_M(\mathbf{x})$ and $J_M(\mathbf{x})$ are given by

$$\widehat{I}_M(\mathbf{x}) = \int_{-\infty}^{+\infty} \Delta\widehat{\pi}_M(y|\mathbf{x})\Delta\widehat{\pi}_M(y|\mathbf{x})'/\widehat{\pi}_M(y|\mathbf{x})\mathrm{d}y$$

and

$$\widehat{J}_M(\mathbf{x}) = \int_{-\infty}^{+\infty} \Delta\widehat{\pi}_M(y|\mathbf{x})\Delta\widehat{\pi}_M(y|\mathbf{x})'\mathrm{d}y\,,$$

respectively.

Forecasting Financial Asset Returns and Sensitive Dependence

Nonlinear Forecasting of Asset Returns

In this subsection, a quick review of the general issues of forecasting financial asset returns is first provided, then the empirical results on the nonlinear forecasting based on nonparametric methods are summarized.

In the past, the random walk model was considered as the most appropriate model to describe the dynamics of asset prices in practice (see [24]). However, after decades of investigation, more evidence on some predictable components of asset returns has been documented in the literature. Although the evidence is often not very strong, several studies report the positive serial dependence for relatively short horizon stock returns. For example, [47] show that first-order autocorrelation of weekly returns on the Center for Research in Security Prices (CRSP) index is as high as 30 percent and significant when an equal-weighted index is used, but is somewhat less when a value-weighted index is used ([12] provide similar evidence for the daily return). The conditional mean of stock returns may not depend only on the past returns but also on other economic variables, including dividend yields, price earnings ratio, short and long interest rates, industrial production and inflation rate. A comprehensive statistical analysis to evaluate the 1-month-ahead out-of-sample forecast of 1 month excess returns by these predictors is conducted by [55]. Some results on the long-horizon predictability in stock returns, based on lagged returns (e. g., [25] and [57]) and other economic variables such as dividend yields or dividend-price ratios (e. g., [26] and [13]) are also available. This evidence on long-horizon forecasts, however, is still controversial because the standard statistical inference procedure may not be reliable in case when the correlation coefficient is computed from a small number of nonoverlapping observations [62] or when the predictor is very persistent in the forecasting regression [73].

The question is whether the introduction of nonlinear structure helps improve the forecasting performance of future asset returns. When the nonlinear condition mean function is unspecified, the neural network method has often been employed as a reliable nonparametric method in predicting the returns. For IBM daily stock returns, [75] found no improvement in out-of-sample predictability based on the neural network model. For daily returns of the Dow Jones Industrial Average (DJIA) index, [33] estimated a nonlinear AR model using the same method. He, in contrast, showed that MSFE reduction over a benchmark linear AR model could be as large as 12.3 percent for the 10-day-ahead out-of-sample forecast. The role of economic fundamentals as predictors can be also investigated under the nonlinear framework. Using a model selection procedure similar to the one employed by [55], some evidence of MSFE improvement from neural network-based forecast of excess returns was provided in [60] and [59] but no encouraging evidence was found in similar studies by [61] and [49]. In practice, 'noise traders' or 'chartists' may predict prices using some technical trading rules (TTRs) rather than using economic fundamentals. For example, a simple TTR based on the moving average can generate a buy signal when the current asset price level P_t is above $n^{-1}\sum_{i=1}^{n} P_{t-i+1}$ for some positive integer n and a sell signal when it is below. [11] found some evidence on the nonlinearity in the conditional mean of DJIA returns conditional on buy-sell signals. [33] further considered including past buy-sell signals as predictors in the neural network model and found that an improvement in MSFE over the linear AR model was even larger than the case when only lagged returns are used as a predictor in the neural network model. One useful nonlinear model is the functional coefficient AR model where the AR coefficient can depend on time or some variables. For example, as in [39], the AR coefficient can be a function of buy-sell signals. [44] claimed that a functional coefficient AR model with a coefficient as a function of the moving average of squared returns well described the serial correlation feature of stock returns.

This moderate but increasing evidence of nonlinear forecastability applies not only to the stock market but also to the foreign exchange market. In the past, [52] could not find any reasonable linear model that could out-perform the random walk model in an out-of-sample forecast of foreign exchange rates. The nonlinear AR model was estimated nonparametrically by [19] but no forecasting improvement over the random walk model could be found in their analysis. However, many follow-up studies, including [34,39,43,76], provided some evidence on forecastability with nonlinear AR models es-

timated using neural networks or other nonparametric methods.

One important and robust empirical fact is that much higher positive serial correlation is typically observed for the volatility measures such as the absolute returns, $|x_t|$, and their power transformation, $|x_t|^\alpha$ for $\alpha > 0$, than for the returns, x_t ([20,69]). This observation is often referred to as a volatility clustering. As a result, forecasting volatility has been much more successful than forecasting returns themselves. The most commonly used approach in forecasting volatility is to describe the conditional variance of asset returns using the class of ARCH and GARCH models ([8,23]). The volatility of stock returns is also known to respond more strongly to negative shocks in returns than positive ones. This 'leverage effect' often motivates the introduction of nonlinear structure in volatility modeling such as the EGARCH model of [53]. Instead of estimating the unknown parameter in a specified ARCH model, the nonparametric method can be also employed to estimate the possibly nonlinear ARCH model in forecasting (see [46]). The better forecastability of market direction (or market timing), sign (x_t), than that of returns, has also been documented in the literature. Examples are [55] for the stock market and [39,43], and [16] for the foreign exchange market. Since the return, x_t, can be decomposed into a product of the two components, $|x_t| \times \text{sign}\,(x_t)$, one may think the strong linear or nonlinear forecastability of the volatility and the sign of returns should lead to forecastability of the returns as well. Interestingly, however, [17] theoretically showed that the serial dependence of asset return volatilities and that of return signs did not necessarily imply the serial dependence of returns.

In summary, a growing number of recent studies show some evidence of linear and nonlinear forecastability of asset returns, and stronger evidence of forecastability of their nonlinear transformations, such as the squared returns, absolute returns and the sign of returns. In this sense, the nonlinearity seems to be playing a non-negligible role in explaining the dynamic behavior of asset prices.

Initial Value Sensitivity in Financial Data

Theoretically, when investors have heterogeneous expectations about the future prices, asset price dynamics can be chaotic with a positive Lyapunov exponent [10]. A comprehensive list on earlier empirical work related to the sensitive dependence and chaos in financial data is provided in [2,6]. Many early studies employed either the BDS test or a dimension estimator and provided the indirect evidence on sensitive dependence and chaos. For example, [64] applied the BDS test to weekly returns on the value-weighted CRSP portfolio and rejected iid randomness. [41] further examined weekly value-weighted and equally weighted CRSP portfolio returns, as well as Standard & Poor 500 (S&P 500) index returns for various frequencies, and found strong evidence against iid. Similar findings are also reported for the daily foreign exchange rate returns in [40]. For financial variables, high-frequency data or tick data is often available to researchers. Earlier examples of studies on chaos using high-frequency data include [50], who found some evidence of low-dimensional chaos based on the correlation dimension and K_2 entropy of 20-second S&P 500 index returns, with a number of observations as large as 19,027. Estimation results on Lyapunov exponents for high-frequency stock returns are also available. In addition to the BDS test, [1] and [2] employed the neural network method and found negative Lyapunov exponents in 1- and 5-minute returns of cash series of S&P 500, UK Financial Times Stock Exchange-100 (FTSE-100) index, Deutscher Aktienindex (DAX), the Nikkei 225 Stock Average, and of futures series of S&P 500 and FTSE-100. Using the resampling procedure of [32], [65] obtained significantly negative Lyapunov exponents for daily stock returns of the Austrian Traded Index (ATX). For the foreign exchange market, [18] estimated Lyapunov exponents of the Canadian, German, Italian and Japanese monthly spot exchange rates using neural nets and found some mixed result regarding their sign.

By using the absolute returns or their power transformation instead of using returns themselves, sensitive dependence of volatility on initial conditions may be examined nonparametrically. [68] used neural nets and estimated Lyapunov exponents of higher order daily returns of the DJIA index. Figure 5 shows their global Lyapunov exponent estimates for simple returns, squared returns and absolute returns. For all cases, Lyapunov exponents are significantly negative but the values of absolute returns are always larger than that of simple returns. While some estimates are close to zero, the observation of the monotonically increasing Lyapunov exponent with increasing p, for daily and absolute returns, resembles the simulation results of the previous section implying the upward bias when the true Lyapunov exponent is negative.

For the exchange rate market, [29] applied [63]'s method to absolute changes and their power transformation of Canadian and German nominal exchange rates and did not reject the null hypothesis of chaos.

For a local measure of initial value sensitivity, [68] also reported the median values of 145 estimates of local Lyapunov exponents for DJIA returns, in addition

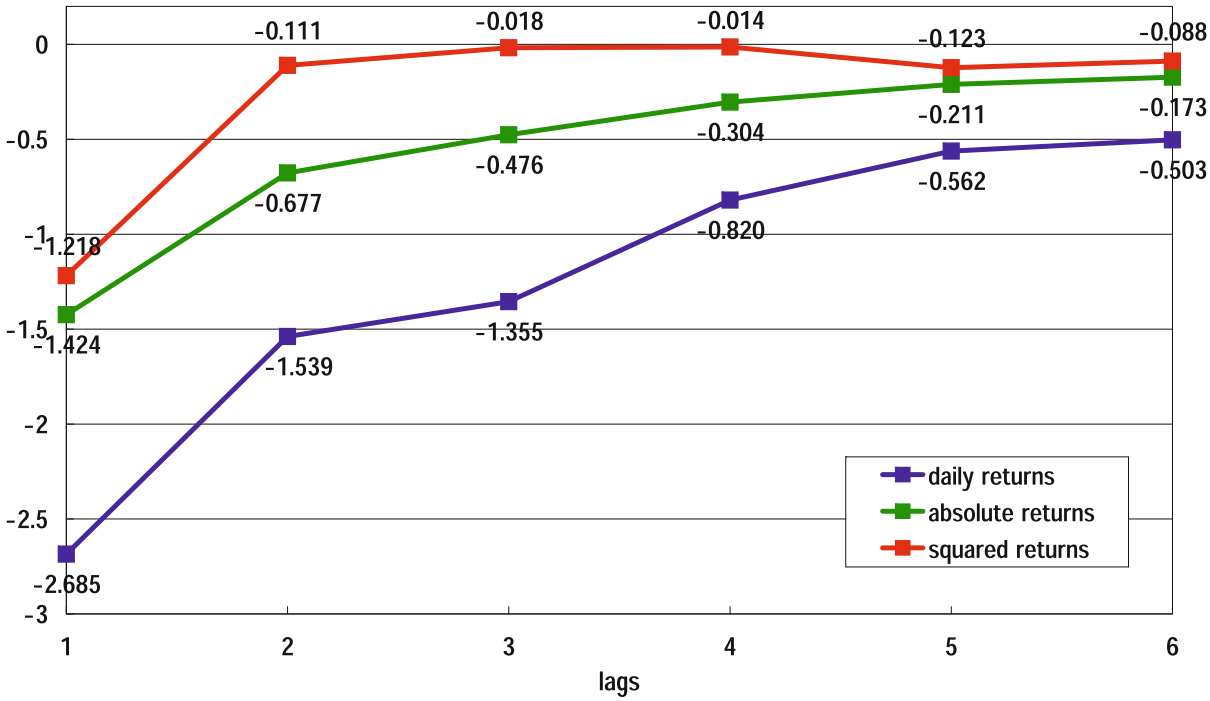

Financial Forecasting, Sensitive Dependence, Figure 5
Global Lyapunov exponents of stock returns

to the global Lyapunov exponents. [45] reported the nonlinear impulse response functions of yen/dollar and deutschemark/dollar exchange rate returns based on parametrically estimated GARCH model. [14] reported a Lyapunov-like index of [80] for the simple returns and absolute returns of CRSP data used in [64]. From Fig. 6, they concluded that (i) the first half of the CRSP series is more volatile than its second half, suggesting that the market

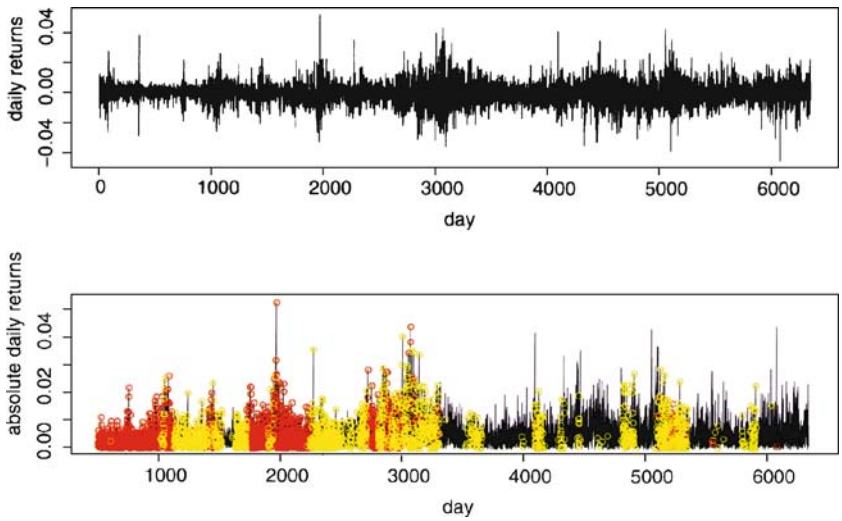

Financial Forecasting, Sensitive Dependence, Figure 6
Upper **panel displays the time series plot of the CRSP daily returns.** ***Lower*** **panel shows the absolute CRSP daily returns with data coloured** ***red*** **whenever their Lyapunov-like indices are above the third quartile of the indices, and data coloured** ***yellow*** **if their indices are between the median and the third quartile**

becomes more mature with time, and (ii) volatile periods tend to form clusters. [65] reported the information matrix measure of local sensitive dependence computed from ATX data based on the parametric estimation of ARCH and GARCH models, in addition to nonparametric estimates of the global Lyapunov exponent.

On the whole, empirical studies on global and local sensitivity measures suggested less sensitive dependence than the chaotic model would predict, but some sensitivity of short-term forecastability on initial conditions.

Future Directions

Some of the possible directions of future research topics are in order.

The first direction is to search for the economic theory behind the initial value sensitivity if detected in the data. The statistical procedures introduced here are basically data description and the empirical results obtained by this approach are not directly connected to underlying economic or finance theory. Theories, such as the one developed by [10], can predict complex behavior of asset prices but direct estimation of the model are typically not possible. Thus, for most cases, the model is evaluated by matching the actual data with the one generated from the model in simulation. Thus direct implication to the sensitive dependence measure would provide a more convincing argument for the importance of knowing the structure. [38] may be considered as one attempt in this direction.

The second direction is to develop better procedures in estimating the initial value sensitivity with the improved accuracy in the environment of a relatively small sample size. In the Jacobian method of estimating the Lyapunov exponent, the conditional mean function has been estimated either parametrically and nonparametrically. A fully nonparametric approach, however, is known to suffer from a high dimensionality problem. A semiparametric approach, such as the one for an additive AR model, is likely to be useful in this context but has not been used in the initial value sensitivity estimation.

The third direction is towards further analysis based on high-frequency data, which has become more commonly available in empirical finance. Much progress has been made in the statistical theory on the realized volatility computed from such data, and forecasting volatility of asset returns based on the realized volatility has been empirically successful (see, e. g., [3]). However, so far, this approach has not been used in detecting the initial value sensitivity in volatility. In addition, realized volatility is known to suffer from market microstructure noise when sampling frequency increases. Given the fact that the initial value sensitivity measures can be considered in the framework of the nonlinear AR models, namely, the stochastic environment in the presence of noise, it is of interest in investigating the robustness of the procedure to the market microstructure noise when applied to high-frequency returns.

Bibliography

1. Abhyankar A, Copeland LS, Wong W (1995) Nonlinear dynamics in real-time equity market indices: evidence from the United Kingdom. Econ J 105:864–880
2. Abhyankar A, Copeland LS, Wong W (1997) Uncovering nonlinear structure in real-time stock-market indexes: The S&P 500, the DAX, the Nikkei 225, and the FTSE-100. J Bus Econ Stat 15:1–14
3. Andersen TB, Bollerslev T, Diebold FX, Labys P (2003) Modeling and forecasting realized volatility. Econometrica 71:579–625
4. Bailey BA, Ellner S, Nychka DW (1997) Chaos with confidence: Asymptotics and applications of local Lyapunov exponents. In: Cutler CD, Kaplan DT (eds) Fields Institute Communications, vol 11. American Mathematical Society, Providence, pp 115–133
5. Barnett WA, Gallant AR, Hinich MJ, Jungeilges J, Kaplan D, Jensen MJ (1995) Robustness of nonlinearity and chaos tests to measurement error, inference method, and sample size. J Econ Behav Organ 27:301–320
6. Barnett WA, Serletis A (2000) Martingales, nonlinearity, and chaos. J Econ Dyn Control 24:703–724
7. Bask M, de Luna X (2002) Characterizing the degree of stability of non-linear dynamic models. Stud Nonlinear Dyn Econom 6:3
8. Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. J Econ 31:307–327
9. Brock WA, Dechert WD, Scheinkman JA, LeBaron B (1996) A test for independence based on the correlation dimension. Econ Rev 15(3):197–235
10. Brock WA, Hommes CH (1998) Heterogeneous beliefs and routes to chaos in a simple asset pricing model. J Econ Dyn Control 22:1235–1274
11. Brock WA, Lakonishok J, LeBaron B (1992) Simple technical trading rules and the stochastic properties of stock returns. J Financ 47:1731–1764
12. Campbell JY, Lo AW, MacKinlay AC (1997) The Econometrics of Financial Markets. Princeton University Press, Princeton
13. Campbell JY, Shiller R (1988) The dividend-price ratio and expectations of future dividends and discount factors. Rev Financ Stud 1:195–228
14. Chan KS, Tong H (2001) Chaos: A Statistical Perspective. Springer, New York
15. Cheng B, Tong H (1992) On consistent nonparametric order determination and chaos. J Royal Stat Soc B 54:427–449
16. Cheung YW, Chinn MD, Pascual AG (2005) Empirical exchange rate models of the nineties: Are any fit to survive? J Int Money Financ 24:1150–1175
17. Christoffersen PF, Diebold FX (2006) Financial asset returns, direction-of-change forecasting, and volatility dynamics. Management Sci 52:1273–1287

18. Dechert WD, Gençay R (1992) Lyapunov exponents as a nonparametric diagnostic for stability analysis. J Appl Econom 7:S41–S60
19. Diebold FX, Nason JA (1990) Nonparametric exchange rate prediction? J Int Econ 28:315–332
20. Ding Z, Granger CWJ, Engle RF (1993) A long memory property of stock market returns and a new model. J Empir Financ 1: 83–106
21. Eckmann JP, Kamphorst SO, Ruelle D, Ciliberto S (1986) Liapunov exponents from time series. Phys Rev A 34:4971–4979
22. Eckmann JP, Ruelle D (1985) Ergodic theory of chaos and strange attractors. Rev Mod Phys 57:617–656
23. Engle RF (1982) Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. Econometrica 50:987–1008
24. Fama E (1970) Efficient capital markets: Review of theory and empirical work. J Financ 25:383–417
25. Fama E, French K (1988) Permanent and temporary components of stock prices. J Political Econ 96:246–273
26. Fama E, French K (1988) Dividend yields and expected stock returns. J Financ Econ 22:3–5
27. Fan J, Yao Q (2003) Nonlinear Time Series: Nonparametric and Parametric Methods. Springer, New York
28. Fan J, Yao Q, Tong H (1996) Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. Biometrika 83:189–206
29. Fernándes-Rodríguez F, Sosvilla-Rivero S, Andrada-Félix J (2005) Testing chaotic dynamics via Lyapunov exponents. J Appl Econom 20:911–930
30. Frank M, Stengos T (1989) Measuring the strangeness of gold and silver rates of return. Rev Econ Stud 56:553–567
31. Gallant AR, Rossi PE, Tauchen G (1993) Nonlinear dynamic structures. Econometrica 61:871–907
32. Gençay R (1996) A statistical framework for testing chaotic dynamics via Lyapunov exponents. Physica D 89:261–266
33. Gençay R (1998) The predictability of security returns with simple technical trading rules. J Empir Financ 5:347–359
34. Gençay R (1999) Linear, non-linear and essential foreign exchange rate prediction with simple technical trading rules. J Int Econ 47:91–107
35. Giannerini S, Rosa R (2001) New resampling method to assess the accuracy of the maximal Lyapunov exponent estimation. Physica D 155:101–111
36. Grassberger P, Procaccia I (1983) Estimation of the Kolmogorov entropy from a chaotic signal. Phys Rev A 28:2591–2593
37. Hall P, Wolff RCL (1995) Properties of invariant distributions and Lyapunov exponents for chaotic logistic maps. J Royal Stat Soc B 57:439–452
38. Hommes CH, Manzan S (2006) Comments on Testing for nonlinear structure and chaos in economic time series. J Macroecon 28:169–174
39. Hong Y, Lee TH (2003) Inference on predictability of foreign exchange rates via generalized spectrum and nonlinear time series models. Rev Econ Stat 85:1048–1062
40. Hsieh DA (1989) Testing for nonlinear dependence in daily foreign exchange rates. J Bus 62:339–368
41. Hsieh DA (1991) Chaos and nonlinear dynamics: application to financial markets. J Financ 46:1839–1877
42. Koop G, Pesaran MH, Potter SM (1996) Impulse response analysis in nonlinear multivariate models. J Econom 74:119–147
43. Kuan CM, Liu T (1995) Forecasting exchange rates using feedforward and recurrent neural networks. J Appl Econom 10:347–364
44. LeBaron B (1992) Some relation between the volatility and serial correlations in stock market returns. J Bus 65:199–219
45. Lin WL (1997) Impulse response function for conditional volatility in GARCH models. J Bus Econ Stat 15:15–25
46. Linton OB (2008) Semiparametric and nonparametric ARCH modelling. In: Anderson TG, Davis RA, Kreiss JP, Mikosch T (ed) Handbook of Financial Time Series. Springer, Berlin
47. Lo AW, MacKinlay AC (1988) Stock market prices do not follow random walks: evidence from a simple specification test. Rev Financ Stud 1:41–66
48. Lu ZQ, Smith RL (1997) Estimating local Lyapunov exponents. In: Cutler CD, Kaplan DT (eds) Fields Institute Communications, vol 11. American Mathematical Society, Providence, pp 135–151
49. Maasoumi E, Racine J (2002) Entropy and predictability of stock market returns. J Econom 107:291–312
50. Mayfield ES, Mizrach B (1992) On determining the dimension of real time stock price data. J Bus Econ Stat 10:367–374
51. McCaffrey DF, Ellner S, Gallant AR, Nychka DW (1992) Estimating the Lyapunov exponent of a chaotic system with nonparametric regression. J Am Stat Assoc 87:682–695
52. Meese R, Rogoff K (1983) Exchange rate models of the seventies. Do they fit out of sample? J Int Econ 14:3–24
53. Nelson DB (1990) Conditional heteroskedasticity in asset returns: A new approach. Econometrica 59:347–370
54. Nychka D, Ellner S, Gallant AR, McCaffrey D (1992) Finding chaos in noisy system. J Royal Stat Soc B 54:399–426
55. Pesaran MH, Timmermann A (1995) Predictability of stock returns: robustness and economic significance. J Financ 50: 1201–1228
56. Pesin JB (1977) Characteristic Liapunov exponents and smooth ergodic theory. Russ Math Surv 32:55–114
57. Poterba JM, Summers LH (1988) Mean reversion in stock prices: evidence and implications. J Financ Econ 22:27–59
58. Potter SM (2000) Nonlinear impulse response functions. J Econ Dyn Control 24:1425–1446
59. Qi M (1999) Nonlinear predictability of stock returns using financial and economic variables. J Bus Econ Stat 17:419–429
60. Qi M, Maddala GS (1999) Economic factors and the stock market: a new perspective. J Forecast 18:151–166
61. Racine J (2001) On the nonlinear predictability of stock returns using financial and economic variables. J Bus Econ Stat 19: 380–382
62. Richardson M, Stock JH (1989) Drawing inferences from statistics based on multiyear asset returns. J Financ Econ 25: 323–348
63. Rosenstein MT, Collins JJ, De Luca CJ (1993) A practical method for calculating largest Lyapunov exponents from small data sets. Physica D 65:117–134
64. Scheinkman JA, LeBaron B (1989) Nonlinear dynamics and stock returns. J Bus 62:311–337
65. Schittenkopf C, Dorffner G, Dockner EJ (2000) On nonlinear, stochastic dynamics in economic and financial time series. Stud Nonlinear Dyn Econom 4:101–121
66. Shintani M (2006) A nonparametric measure of convergence towards purchasing power parity. J Appl Econom 21:589–604
67. Shintani M, Linton O (2003) Is there chaos in the world economy? A nonparametric test using consistent standard errors. Int Econ Rev 44:331–358

68. Shintani M, Linton O (2004) Nonparametric neural network estimation of Lyapunov exponents and a direct test for chaos. J Econom 120:1–33
69. Taylor SJ (1986) Modelling Financial Time Series. Wiley, New York
70. Tjostheim D, Auestad BH (1994) Nonparametric identification of nonlinear time series: Selecting significant lags. J Am Stat Assoc 89:1410–1419
71. Tschernig R, Yang L (2000) Nonparametric lag selection for time series. J Time Ser Analysis 21:457–487
72. Tschernig R, Yang L (2000) Nonparametric estimation of generalized impulse response functions. Michigan State University, unpublished
73. Valkanov R (2003) Long-horizon regressions: theoretical results and applications. J Financ Econom 68:201–232
74. Whang YJ, Linton O (1999) The asymptotic distribution of nonparametric estimates of the Lyapunov exponent for stochastic time series. J Econom 91:1–42
75. White H (1988) Economic prediction using neural networks: the case of IBM stock returns. Proceedings of the IEEE International Conference on Neural Networks 2. The Institute of Electrical and Electronics Engineers, San Diego, pp 451–458
76. White H, Racine J (2001) Statistical inference, the bootstrap, and neural-network modeling with application to foreign exchange rates. IEEE Trans Neural Netw 12:657–673
77. Wolf A, Swift JB, Swinney HL, Vastano JA (1985) Determining Lyapunov exponents from a time series. Physica D 16:285–317
78. Wolff RCL (1992) Local Lyapunov exponent: Looking closely at chaos. J Royal Stat Soc B 54:353–371
79. Wolff R, Yao Q, Tong H (2004) Statistical tests for Lyapunov exponents of deterministic systems. Stud Nonlinear Dyn Econom 8:10
80. Yao Q, Tong H (1994) Quantifying the influence of initial values on non-linear prediction. J Royal Stat Soc Ser B 56:701–725
81. Yao Q, Tong H (1994) On prediction and chaos in stochastic systems. Philos Trans Royal Soc Lond A 348:357–369

Finite Dimensional Controllability

Lionel Rosier
Institut Elie Cartan, Vandoeuvre-lès-Nancy, France

Article Outline

Glossary
Definition of the Subject
Introduction
Control Systems
Linear Systems
Linearization Principle
High Order Tests
Controllability and Observability
Controllability and Stabilizability
Flatness
Future Directions
Bibliography

Glossary

Control system A **control system** is a dynamical system incorporating a **control input** designed to achieve a control objective. It is **finite dimensional** if the phase space (e. g. a vector space or a manifold) is of finite dimension. A **continuous-time** control system takes the form $\mathrm{d}x/\mathrm{d}t = f(x, u)$, $x \in X$, $u \in U$ and $t \in \mathbb{R}$ denoting respectively the **state**, the **input**, and the **continuous time**. A **discrete-time** system assumes the form $x_{k+1} = f(x_k, u_k)$, where $k \in \mathbb{Z}$ is the **discrete time**.

Open/closed loop A control system is said to be in **open loop** form when the input u is any function of time, and in **closed loop** form when the input u is a function of the state only, i. e., it takes the more restrictive form $u = h(x(t))$, where $h\colon X \to U$ is a given function called a **feedback law**.

Controllability A control system is **controllable** if any pair of states may be connected by a trajectory of the system corresponding to an appropriate choice of the control input.

Stabilizability A control system is **asymptotically stabilizable** around an equilibrium point if there exists a feedback law such that the corresponding closed loop system is asymptotically stable at the equilibrium point.

Output function An **output function** is any function of the state.

Observability A control system given together with an output function is said to be **observable** if two different states give rise to two different outputs for a convenient choice of the input function.

Flatness An output function is said to be **flat** if the state and the input can be expressed as functions of the output and of a finite number of its derivatives.

Definition of the Subject

A control system is controllable if any state can be steered to another one in the phase space by an appropriate choice of the control input. While the stabilization issue has been addressed since the XIXth century (Watt's steam engine governor providing a famous instance of a stabilization mechanism at the beginning of the English Industrial Revolution), the controllability issue has been addressed for the first time by Kalman in the 1960s [20]. The controllability is a basic mathematical property which characterizes the degrees of freedom available when we try to control a system. It is strongly connected to other control concepts: optimal control, observability, and stabilizability. While the controllability of linear finite-dimensional systems is well understood since Kalman's seminal papers,

the situation is more tricky for nonlinear systems. For the later, concepts borrowed from differential geometry (e. g. Lie brackets, holonomy, ...) come into play, and the study of their controllability is still a field of active research.

The controllability of finite dimensional systems is a basic concept in control theory, as well as a notion involved in many applications, such as spatial dynamics (with e. g. spatial rendezvous), airplane autopilot, industrial robots, quantic chemistry.

Introduction

A very familiar example of a controllable finite dimensional system is given by a car that one attempts to park at some place in a parking. The phase space is roughly the three dimensional Euclidean space $\mathbb{R}^3$, a state being composed of the two coordinates of the center of mass together with the angle formed by some axis linked to the car with the (fixed) abscissa axis. The driver may act on the angle of the wheels and on their velocity, which may thus be taken as control inputs. In general, the presence of obstacles (e. g. other cars) impose to change the phase space $\mathbb{R}^3$ to a subset of it. The controllability issue is roughly how to combine changes of direction and of velocity to drive the car from a position to another one. Note that the system is controllable, even if the number of control inputs (2) is less than the number of independent coordinates (3). This is an important property resting upon the many connections between the coordinates of the state. While in nature each motion is generally controlled by an input (think of the muscles in an arm), the control theory focuses on the study of systems in which an input living in a space of low dimension (typically, one) is sufficient to control the coordinates of a state living in a space of high dimension.

The article is outlined as follows. In Sect. "Control Systems", we introduce the basic concepts (controllability, stabilizability) used thereafter. In the next section, we review the linear theory, recalling the Kalman and Hautus tests for the controllability of a time invariant system, and the Gramian test for a time dependent system. Section "Linearization Principle" is devoted to the linearization principle, which allows to deduce the controllability of a nonlinear system from the controllability of its linearization along a trajectory. The focus in Sect. "High Order Tests" is on nonlinear systems for which the linear test fails, i. e., the linearized system fails to be controllable. High order conditions based upon Lie brackets ensuring controllability will be given, first for systems without drift, and next for systems with a drift. Section "Controllability and Observability" explores the connections between controllability and observability, while Sect. "Controllability and Stabilizability" shows how to derive stabilization results from the controllability property. A final section on the flatness, a new theory used in many applications to design explicit control inputs, is followed by some thoughts on future directions.

Control Systems

A **finite dimensional (continuous-time) control system** is a differential equation of the form

$$\dot{x} = f(x, u) \tag{1}$$

where $x \in X$ is the **state**, $u \in U$ is the **input**, $f\colon X \times U \to U$ is a smooth (typically real analytic) nonlinear function, $\dot{x} = \mathrm{d}x/\mathrm{d}t$, and X and U denote finite dimensional manifolds. For the sake of simplicity, we shall assume here that $X \subset \mathbb{R}^n$ and $U \subset \mathbb{R}^m$ are open sets. Sometimes, we impose U to be bounded (or to be a compact set) to force the control input to be bounded. Given some control input $u \in L^\infty(I; U)$, i. e. a measurable essentially bounded function $u\colon I \to U$, a **solution** of (1) is a locally Lipschitz continuous function $x(\cdot)\colon J \to X$, where $J \subset I$, such that

$$\dot{x}(t) = f(x(t), u(t)) \qquad \text{for almost every } t \in J\,. \tag{2}$$

Note that $J \subset I$ only, that is, x needs not exist on all I, as it may escape to infinity in finite time. In general, u is piecewise smooth, so that (2) holds actually for all t except for finitely many values. The basic problem of the controllability is the issue whether, given an initial state $x_0 \in X$, a terminal state $x_T \in X$, and a control time $T > 0$, one may design a control input $u \in L^\infty([0, T]; U)$ such that the solution of the system

$$\begin{cases} \dot{x}(t) = f(x(t)\,, u(t)) \\ x(0) = x_0 \end{cases} \tag{3}$$

satisfies $x(T) = x_T$.

An **equilibrium position** for (1) is a point $\overline{x} \in X$ such that there exists a value $\overline{u} \in U$ (typically, 0) such that $f(\overline{x}, \overline{u}) = 0$.

An **asymptotically stabilizing feedback law** is a function $k\colon X \to U$ with $k(\overline{x}) = \overline{u}$, such that the closed loop system

$$\dot{x} = f(x, k(x)) \tag{4}$$

obtained by plugging the control input

$$u(t) := k(x(t)) \tag{5}$$

into (1), is **locally asymptotically stable** at $\overline{x}$. Recall that it means that (i) the equilibrium point is **stable**: for any $\varepsilon > 0$, there exists some $\delta > 0$ such that any solution of (4) starting from a point x_0 such that $|x_0 - \overline{x}| < \delta$ at $t = 0$, is defined on $\mathbb{R}^+$ and satisfies $|x(t)| \leq \varepsilon$ for all $t \geq 0$; (ii) the equilibrium point is **attractive**. For some $\delta > 0$ as in (i), we have also that $x(t) \to 0$ as $t \to \infty$ whenever $|x_0| < \delta$.

The feedback laws considered here will be **continuous**, and we shall mean by a solution of (4) any function $x(t)$ satisfying (4) for all t. Notice that the solutions of the Cauchy problems exist (locally in time) by virtue of Peano theorem.

A control system is **asymptotically stabilizable** around an equilibrium position if an asymptotically stabilizing feedback law as above does exist. In the following, we shall also consider time-varying systems, i. e. systems of the form

$$\dot{x} = f(x, t, u) \tag{6}$$

where $f\colon X \times I \times U \to X$ is smooth and $I \subset \mathbb{R}$ denotes some interval. The controllability and stabilizability concepts extend in a natural way to that setting. Time-varying feedback laws

$$u(t) = k(x(t), t) \tag{7}$$

where $k\colon X \times \mathbb{R} \to U$ is a smooth (generally time periodic) function, prove to be useful in situations where the classical static stabilization defined above fails.

Linear Systems

Time Invariant Linear Systems

A **time invariant linear system** is a system of the form

$$\dot{x} = Ax + Bu \tag{8}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ denote some time invariant matrices. This corresponds to the situation where the function f in (1) is linear in both the state and the input. Here, $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$, and we consider square integrable inputs $u \in L^2([0, T]; \mathbb{R}^m)$. We say that (8) is **controllable in time** T if for any pair of states $x_0, x_T \in \mathbb{R}^n$, one may construct an input $u \in L^2([0, T]; \mathbb{R}^m)$ that steers (8) from x_0 to x_T. Recall that the solution of (8) emanating from x_0 at $t = 0$ is given by Duhamel formula

$$x(t) = \mathrm{e}^{tA}x_0 + \int_0^t \mathrm{e}^{(t-s)A}Bu(s)\,\mathrm{d}s\,.$$

Let us introduce the $n \times nm$ matrix $\mathcal{R}(A, B) = (B|AB|A^2B|\cdots|A^{n-1}B)$ obtained by gathering together the matrices $B, AB, A^2B, \ldots, A^{n-1}B$. Then we have the following rank condition due to Kalman [20] for the controllability of a linear system.

Theorem 1 *(Kalman) The linear system (8) is controllable in time T if and only if* $\operatorname{rank}\mathcal{R}(A, B) = n$.

We notice that the controllability of a linear system does not depend of the control time T, and that the control input may actually be chosen very smooth (e. g. in $C^\infty([0, T]; \mathbb{R}^n)$).

Example 1 Consider a pendulum to which is applied a torque as a control input. A simplified model is then given by the following linear system

$$\dot{x}_1 = x_2 \tag{9}$$

$$\dot{x}_2 = -x_1 + u \tag{10}$$

where x_1, x_2 and u stand respectively for the angle with the vertical, the angular velocity, and the torque. Here, $n = 2$, $m = 1$, and

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \text{ and } B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{11}$$

As $\operatorname{rank}(B, AB) = 2$, we infer from Kalman rank test that (9)–(10) is controllable.

When (8) fails to be controllable, it may be important (e. g. when studying the stabilizability) to identify the uncontrollable part of (8). Assume that (8) is not controllable, and let $r = \operatorname{rank}\mathcal{R}(A, B) < n$. It may be seen that the reachable space from the origin, that is the set

$$\mathcal{R} = \left\{ x_T \in \mathbb{R}^n; \exists T > 0, \exists u \in L^2\left([0, T]; \mathbb{R}^n\right), \int_0^T \mathrm{e}^{(T-s)A}Bu(s)\,\mathrm{d}s = x_T \right\}, \tag{12}$$

coincides with the space spanned by the columns of the matrix $\mathcal{R}(A, B)$:

$$\mathcal{R} = \{\mathcal{R}(A, B)V\,; V \in \mathbb{R}^{\mathrm{nm}}\} \tag{13}$$

In particular, $\dim\mathcal{R} = \operatorname{rank}\mathcal{R}(A, B) = r < n$. Let $\mathbf{e} = \{\mathrm{e}_1, \ldots, \mathrm{e}_n\}$ be the canonical basis of $\mathbb{R}^n$, and let $\{f_1, \ldots, f_r\}$ be a basis of $\mathcal{R}$, that we complete in a basis $\mathbf{f} = \{f_1, \ldots, f_n\}$ of $\mathbb{R}^n$. If x (resp. $\tilde{x}$) denotes the vector of the coordinates of a point in the basis $\mathbf{e}$ (resp. $\mathbf{f}$), then $x = T\tilde{x}$ where $T = (f_1, f_2, \ldots, f_n) \in \mathbb{R}^{n\times n}$. In the new coordinates $\tilde{x}$, (8) may be written

$$\dot{\tilde{x}} = \tilde{A}\tilde{x} + \tilde{B}u \tag{14}$$

where $\tilde{A} := T^{-1}AT$ and $\tilde{B} := T^{-1}B$ read

$$\tilde{A} = \begin{pmatrix} \tilde{A}_1 & \tilde{A}_2 \\ 0 & \tilde{A}_3 \end{pmatrix} \quad \tilde{B} = \begin{pmatrix} \tilde{B}_1 \\ 0 \end{pmatrix} \tag{15}$$

with $\tilde{A}_1 \in \mathbb{R}^{r\times r}$, $\tilde{B}_1 \in \mathbb{R}^{r\times m}$. Writing $\tilde{x} = (\tilde{x}_1, \tilde{x}_2) \in \mathbb{R}^r \times \mathbb{R}^{n-r}$, we have

$$\dot{\tilde{x}}_1 = \tilde{A}_1\tilde{x}_1 + \tilde{A}_2\tilde{x}_2 + \tilde{B}_1 u \tag{16}$$

$$\dot{\tilde{x}}_2 = \tilde{A}_3\tilde{x}_2 \tag{17}$$

and it may be proved that

$$\text{rank } \mathcal{R}(\tilde{A}_1, \tilde{B}_1) = r\,. \tag{18}$$

This is the **Kalman controllability decomposition**. By (18), the dynamics of $\tilde{x}_1$ is well controlled. Actually, a solution of (18) evaluated at $t = T$ assumes the form

$$\begin{aligned}\tilde{x}_1(T) &= \mathrm{e}^{T\tilde{A}_1}\tilde{x}_1(0) + \int_0^T \mathrm{e}^{(T-s)\tilde{A}_1}\tilde{A}_2\tilde{x}_2(s)\,\mathrm{d}s \\ &\quad + \int_0^T \mathrm{e}^{(T-s)\tilde{A}_1}\tilde{B}_1 u(s)\mathrm{d}s \\ &= \mathrm{e}^{T\tilde{A}_1}\overline{x}_1 + \int_0^T \mathrm{e}^{(T-s)\tilde{A}_1}\tilde{B}_1 u(s)\mathrm{d}s\end{aligned}$$

if we set $\overline{x}_1 = \tilde{x}_1(0) + \int_0^T \mathrm{e}^{-s\tilde{A}_1}\tilde{A}_2\tilde{x}_2(s)\mathrm{d}s$. Hence $\tilde{x}_1(T)$ may be given any value in $\mathbb{R}^r$. On the other hand, no control input is present in (17). Thus $\tilde{x}_2$ stands for the uncontrolled part of the dynamics of (8).

Another test based upon a spectral analysis has been furnished by Hautus in [16].

Theorem 2 *(Hautus) The control system (8) is controllable in time T if and only if* $\text{rank}(\lambda I - A, B) = n$ *for all* $\lambda \in \mathbb{C}$.

Notice that in Hautus test we may restrict ourselves to the complex numbers λ which are eigenvalues of A, for otherwise $\text{rank}(\lambda I - A) = n$.

Time-Varying Linear Systems

Let us now turn to the controllability issue for a time-varying linear system

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) \tag{19}$$

where $A \in L^\infty([0, T]\,; \mathbb{R}^{n\times n})$, $B \in L^\infty([0, T]\,; \mathbb{R}^{n\times m})$ denote time-varying matrices. Such a system arises in a natural way when linearizing a control system (1) along a trajectory. The input $u(t)$ is any function in $L^\infty([0, T]\,; \mathbb{R}^m)$ and a solution of (19) is any locally Lipschitz continuous function satisfying (19) almost everywhere.

We define the **fundamental solution** ϕ associated with A as follows. Pick any $s \in [0, T]$, and let $M\colon [0, T] \to \mathbb{R}^{n\times n}$ denote the solution of the system

$$\begin{aligned}\dot{M}(t) &= A(t)M(t) \\ M(s) &= I\end{aligned} \tag{20}$$

Then $\phi(t, s) := M(t)$. Notice that $\phi(t, s) = \mathrm{e}^{(t-s)A}$ when A is constant.

The solution x of (20) starting from x_0 at time t_0 reads then

$$x(t) = \phi(t, t_0)x_0 + \int_{t_0}^t \phi(t, s)B(s)u(s)\mathrm{d}s\,.$$

The **controllability Gramian** of (20) is the matrix

$$G = \int_0^T \phi(T, t)B(t)B^*(t)\phi^*(T, t)\mathrm{d}t$$

where * denotes transpose. Note that $G \in \mathbb{R}^{n\times n}$, and that G is a nonnegative symmetric matrix. Then the following result holds.

Theorem 3 *(Gramian test) The system (19) is controllable on* $[0, T]$ *if and only if the Gramian G is invertible.*

Note that the Gramian test provides a third criterion to test whether a time invariant linear system (8) is controllable or not.

Corollary 4 *(8) is controllable in time T if and only if the Gramian*

$$G = \int_0^T e^{(T-t)A}BB^*e^{(T-t)A^*}\,dt$$

is invertible.

As the value of the control time T plays no role according to Kalman test, it follows that the Gramian G is invertible for *all* $T > 0$ whenever it is invertible for *one* $T > 0$.

If (19) is controllable, then an explicit control input steering (19) from x_0 to x_T is given by

$$\overline{u}(t) = B^*(t)\phi^*(T, t)G^{-1}(x_T - \phi(T, 0)x_0)\,. \tag{21}$$

A remarkable property of the control input $\overline{u}$ is that $\overline{u}$ minimizes the control cost

$$E(u) = \int_0^T |u(t)|^2\mathrm{d}t$$

among all the control inputs $u \in L^\infty([0, T]\,; \mathbb{R}^m)$ (or $u \in L^2([0, T]\,; \mathbb{R}^m)$) steering (19) from x_0 to x_T. Actually a little more can be said.

Proposition 5 *If $u \in L^2([0,T];\mathbb{R}^m)$ is such that the solution x of (19) emanating from x_0 at $t=0$ reaches x_T at time T, and if $u \neq \overline{u}$, then*

$$E(\overline{u}) < E(u).$$

The main drawback of the Gramian test is that the knowledge of the fundamental solution $\phi(t,s)$ and the computation of an integral term are both required. In the situation where $A(t)$ and $B(t)$ are smooth functions of time, a criterion based only upon derivatives in time is also available. Assume that $A \in C^\infty([0,T];\mathbb{R}^{n\times n})$ and that $B \in C^\infty([0,T];\mathbb{R}^{n\times m})$, and define a sequence of functions $B_i \in C^\infty([0,T];\mathbb{R}^{n\times m})$ by induction on i by

$$\begin{aligned} B_0 &= B \\ B_i &= AB_{i-1} - \frac{\mathrm{d}B_{i-1}}{\mathrm{d}t}\,. \end{aligned} \tag{22}$$

Then the following result holds (see, e. g., [11]).

Theorem 6 *Assume that there exists a time $\overline{t} \in [0,T]$ such that*

$$\operatorname{span}(B_i(\overline{t})v\,;v \in \mathbb{R}^m, i \geq 0) = \mathbb{R}^n\,. \tag{23}$$

Then (19) is controllable.

The converse of Theorem 6 is true when A and B are real analytic functions of time. More precisely, we have the following result.

Theorem 7 *If A and B are real analytic on $[0,T]$, then (20) is controllable if and only if for all $\overline{t} \in [0,T]$*

$$\operatorname{span}(B_i(\overline{t})v\,;v \in \mathbb{R}^m\,, i \geq 0) = \mathbb{R}^n\,.$$

Clearly, Theorem 7 is not valid when A and B are merely of class C^∞. (Take $n=m=1$, $A(t)=0$, $B(t)=\exp(-t^{-1})$ and $\overline{t}=0$.)

Linear Control Systems in Infinite Dimension

Let us end this section with some comments concerning the extensions of the above controllability tests to control systems in infinite dimension (see [30] for more details). Let us consider a control system of the form

$$\dot{x} = Ax + Bu \tag{24}$$

where $A\colon D(A) \subset X \to X$ is an (unbounded) operator generating a strongly continuous semigroup $(S(t))_{t\geq 0}$ on a (complex) Hilbert space X, and $B\colon U \to X$ is a bounded operator, U denoting another Hilbert space.

Definition 8 We shall say that (24) is

- **Exactly controllable in time** T if for any $x_0, x_T \in X$ there exists $u \in L^2([0,T];U)$ such that the solution x of (24) emanating from x_0 at $t=0$ satisfies $x(T)=x_T$;
- **Null controllable in time** T if for any $x_0 \in X$ there exists $u \in L^2([0,T];U)$ such that the solution x of (24) emanating from x_0 at $t=0$ satisfies $x(T)=0$;
- **Approximatively controllable in time** T if for any $x_0, x_T \in X$ and any $\varepsilon > 0$ there exists $u \in L^2([0,T];U)$ such that the solution x of (24) emanating from x_0 at $t=0$ satisfies $||x(T)-x_T|| < \varepsilon$ ($||\cdot||$ denoting the norm in X).

This setting is convenient for a partial differential equation with an internal control $Bu := gu(t)$, where $g = g(x)$ is such that $gU \subset X$. Let us review the above controllability tests.

- Kalman rank test, which is based on a computation of the dimension of the reachable space, possesses some extension giving the approximate (not exact!) controllability of (24) (See [13], Theorem 3.16). More interestingly, a Kalman-type condition has been introduced in [23] to investigate the null controllability of a system of coupled parabolic equations.
- Hautus test admits the following extension due to Liu [25]: (24) is (exactly) controllable in time T if and only if there exists some constant $\delta > 0$ such that

$$||(A^* - \lambda I)z||^2 + ||B^*z||^2 \geq \delta||z||^2 \qquad \forall z \in D(A^*)\,, \forall \lambda \in \mathbb{C}\,. \tag{25}$$

  In (25), A^* (resp. B^*) denotes the adjoint of the operator A (resp. B), and $||\cdot||$ denotes the norm in X.
- The Gramian test admits the following extension, due to Dolecky–Russell [14] and J.L. Lions [24]: (24) is exactly controllable in time T if and only if there exists some constant $\delta > 0$ such that

$$\int_0^T ||B^*S^*(t)x_0||^2 \mathrm{d}t \geq \delta||x_0||^2 \quad \forall x_0 \in X\,.$$

Linearization Principle

Assume given a smooth nonlinear control system

$$\dot{x} = f(x,u) \tag{26}$$

where $f\colon X \times U \to \mathbb{R}^n$ is a smooth map (i. e. of class C^∞) and $X \subset \mathbb{R}^n$, $U \subset \mathbb{R}^m$ denote some open sets. Assume also given a reference trajectory $(\overline{x},\overline{u})\colon [0,T] \to X \times U$

where $\overline{u} \in L^\infty([0,T];U)$ is the control input and $\overline{x}$ solves (26) for $u \equiv \overline{u}$. We introduce the linearized system along the reference trajectory defined as

$$\dot{y} = A(t)y + B(t)v \tag{27}$$

where

$$\begin{aligned} A(t) &= \frac{\partial f}{\partial x}(\overline{x}(t), \overline{u}(t)) \in \mathbb{R}^{n\times n}\,, \\ B(t) &= \frac{\partial f}{\partial u}(\overline{x}(t), \overline{u}(t)) \in \mathbb{R}^{n\times m} \end{aligned} \tag{28}$$

and $y \in \mathbb{R}^n, v \in \mathbb{R}^m$. (27) is formally derived from (26) by letting $x = \overline{x} + y$, $u = \overline{u} + v$ and observing that

$$\begin{aligned} \dot{y} &= \dot{x} - \dot{\overline{x}} = f(\overline{x} + y, \overline{u} + v) - f(\overline{x}, \overline{u}) \\ &\approx \frac{\partial f}{\partial x}(\overline{x}, \overline{u})y + \frac{\partial f}{\partial u}(\overline{x}, \overline{u})v\,. \end{aligned}$$

Notice that if $(\overline{x}_0, \overline{u}_0)$ is an equilibrium point of f (i. e. $f(\overline{x}_0, \overline{u}_0) = 0$) and $\overline{x}(t) \equiv x_0$, $\overline{u}(t) \equiv u_0$, then $A(t) = A = \frac{\partial f}{\partial x}(\overline{x}_0, \overline{u}_0)$ and $B(t) = B = \frac{\partial f}{\partial u}(\overline{x}_0, \overline{u}_0)$ take constant values. Equation (27) is in that case the time invariant linear system

$$\dot{y} = Ay + Bv\,. \tag{29}$$

Let $x_0, x_T \in X$. We seek for a trajectory x of (26) connecting x_0 to x_T when x_0 (resp. x_T) is close to $\overline{x}(0)$ (resp. $\overline{x}(T)$). In addition, we will impose that the trajectory (x, u) be uniformly close to the reference trajectory $(\overline{x}, \overline{u})$. We are led to the following

Definition 9 The system (26) is said to be **controllable along** $(\overline{x}, \overline{u})$ if for each $\varepsilon > 0$, there exists some $\delta > 0$ such that for each $x_0, x_T \in X$ with $||x_0 - \overline{x}(0)|| < \delta$, $||x_T - \overline{x}(T)|| < \delta$, there exists a control input $u \in L^\infty([0,T];U)$ such that the solution of (26) starting from x_0 at $t = 0$ satisfies $x(T) = x_T$ and

$$\sup_{t\in[0,T]} \left(|x(t) - \overline{x}(t)| + |u(t) - \overline{u}(t)|\right) \le \varepsilon\,.$$

We are in a position to state the linearization principle.

Theorem 10 *Let $(\overline{x}, \overline{u})$ be a trajectory of (26). If the linearized system (27) along $(\overline{x}, \overline{u})$ is controllable, then the system (26) is controllable along the trajectory $(\overline{x}, \overline{u})$.*

When the reference trajectory is stationary, we obtain the following result.

Corollary 11 *Let (x_0, u_0) be such that $f(x_0, u_0) = 0$. If the linearized system (29) is controllable, then the system (26) is controllable along the stationary trajectory $(\overline{x}, \overline{u}) = (x_0, u_0)$.*

Notice that the converse of Corollary 11 is not true. (Consider the system $\dot{x} = u^3$ with $n = m = 1$, $x_0 = u_0 = 0$, which is controllable at the origin as it may be seen by performing the change of inputs $v = u^3$.) Often, the nonlinear part of f plays a crucial role in the controllability of (26). This will be explained in the next section using Lie algebraic techniques. Another way to use the nonlinear contribution in f is to consider a linearization of (26) along a convenient (not stationary) trajectory. We consider a system introduced by Brockett in [6] to exhibit an obstruction to stabilizability.

Example 2 The Brockett's system reads

$$\dot{x}_1 = u_1 \tag{30}$$

$$\dot{x}_2 = u_2 \tag{31}$$

$$\dot{x}_3 = x_1u_2 - x_2u_1\,. \tag{32}$$

Its linearization along $(x_0, u_0) = (0, 0)$, which reads

$$\begin{aligned} \dot{y}_1 &= v_1 \\ \dot{y}_2 &= v_2 \\ \dot{y}_3 &= 0\,, \end{aligned}$$

is not controllable, by virtue of the Kalman rank condition. We may however construct a smooth closed trajectory such that the linearization of (30)–(32) along it is controllable. Pick any time $T > 0$ and let $\overline{u}_1(t) = \cos(\pi t/T)$, $\overline{u}_2(t) = 0$ for $t \in [0,T]$ and $x_0 = 0$. Let $\overline{x}$ denote the corresponding solution of (30)–(32). Notice that $\overline{x}_1(t) = (T/\pi)\sin(\pi t/T)$ (hence $\overline{x}_1(T) = 0$), and $\overline{x}_2 = \overline{x}_3 \equiv 0$. The linearization of (30)–(32) along $(\overline{x}, \overline{u})$ is (27) with

$$\begin{aligned} A(t) &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -\cos(\pi t/T) & 0 \end{pmatrix}, \\ B(t) &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & (T/\pi)\sin(\pi t/T) \end{pmatrix}. \end{aligned} \tag{33}$$

Notice that A and B are real analytic, so that we may apply Theorem 8 to check whether (27) is controllable or not on $[0,T]$. Simple computations give

$$B_1(t) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & -2\cos(\pi t/T) \end{pmatrix}.$$

Clearly,

$$\operatorname{span}\left(B_0(0)\begin{pmatrix} 1 \\ 0 \end{pmatrix}, B_0(0)\begin{pmatrix} 0 \\ 1 \end{pmatrix}, B_1(0)\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) = \mathbb{R}^3$$

hence (27) is controllable. We infer from Theorem 10 that (30)–(32) is controllable along $(\overline{x}, \overline{u})$.

Notice that if we want to prove the controllability around an equilibrium point as above for Brockett's system, we have to design a reference control input $\overline{u}$ so that

$$x_0 = \overline{x}(0) = \overline{x}(T) \,. \tag{34}$$

When f is odd with respect to the control, i. e.

$$f(x, -u) = -f(x, u) \,, \tag{35}$$

then (34) is automatically satisfied whenever $\overline{u}$ fulfills

$$\overline{u}(t) = -\overline{u}(T - t) \quad \forall t \in [0, T] \,. \tag{36}$$

Indeed, it follows from (35)–(36) that the solution $\overline{x}$ to (26) starting from x_0 at $t = 0$ satisfies

$$\overline{x}(t) = \overline{x}(T - t) \quad \forall t \in [0, T] \,. \tag{37}$$

In particular, $\overline{x}(T) = \overline{x}(0) = x_0$. Of course, the control inputs of interest are those for which the linearized system (27) is controllable, and the latter property is "generically" satisfied for a controllable system. The above construction of the reference trajectory is due to J.-M. Coron, and is referred to as the **return method**. A precursor to that method is [34]. Beside giving interesting results for the stabilization of finite dimensional systems (see below Sect. "Controllability and Stabilizability"), the return method has also been successfully applied for the control of some important partial differential equations arising in Fluid Mechanics (see [11]).

High Order Tests

In this section, we shall derive new controllability tests for systems for which the linearization principle is inconclusive; that is, the linearization at an equilibrium point fails to be controllable. To simplify the exposition, we shall limit ourselves to systems **affine in the control**, i. e. systems of the form

$$\dot{x} = f_0(x) + u_1 f_1(x) + \cdots + u_m f_m(x), \quad x \in \mathbb{R}^n, |u|_\infty \le \delta \tag{38}$$

where $|u|_\infty := \sup_{1 \le i \le m} |u_i|$ and $\delta > 0$ denotes a fixed number. We assume that $f_0(0) = 0$ and that $f_i \in C^\infty(\mathbb{R}^n, \mathbb{R}^n)$ for each $i \in \{0, \ldots, m\}$. To state the results we need to introduce a few notations. Let $v = (v_1, \ldots, v_n)$ and $w = (w_1, \ldots, w_n)$ be two vector fields of class C^∞ on $\mathbb{R}^n$. The **Lie bracket** of v and w, denoted by $[v, w]$, is the vector field

$$[v, w] = \frac{\partial w}{\partial x} v - \frac{\partial v}{\partial x} w$$

where $\partial w / \partial x$ is the Jacobian matrix $(\partial w_i / \partial x_j)_{i,j=1,\ldots,n}$, and the vector $v(x) = (v_1(x), \ldots, v_n(x))$ is identified to the column $\begin{pmatrix} v_1(x) \\ \vdots \\ v_n(x) \end{pmatrix}$. As $[v, w]$ is still a smooth vector field, we may bracket it with v, or w, etc. Vector fields like $[v, [v, w]]$, $[[v, w], [v, [v, w]]]$, etc. are termed **iterated Lie brackets** of v, w.

The Lie bracketing of vector fields is an operation satisfying the two following properties (easy to check)

1. **Anticommutativity:** $[w, v] = -[v, w]$;
2. **Jacobi identity:** $[f, [g, h]] + [g, [h, f]] + [h, [f, g]] = 0$.

The **Lie algebra** generated by $f_1, \ldots, f_m$, denoted $\mathrm{Lie}(f_1, \ldots, f_m)$, is the smallest vector subspace v of $C^\infty(\mathbb{R}^n, \mathbb{R}^n)$ which contains $f_1, \ldots, f_m$ and which is closed under Lie bracketing (i. e. $v, w \in V \Rightarrow [v, w] \in V$). It is easily to see that $\mathrm{Lie}(f_1, \ldots, f_m)$ is the vector space spanned by all the iterated Lie brackets of $f_1, \ldots, f_m$. (Actually, using the anticommutativity and Jacobi identity, we may restrict ourselves to iterated Lie brackets of the form $[f_{i_1}, [f_{i_2}, [\ldots [f_{i_{p-1}}, f_{i_p}] \ldots]]]$ to span $\mathrm{Lie}(f_1, \ldots, f_m)$.) For any $x \in \mathbb{R}^n$, we set

$$\mathrm{Lie}(f_1, \ldots, f_m)(x) = \{g(x); g \in \mathrm{Lie}(f_1, \ldots, f_m)\} \subset \mathbb{R}^n \,.$$

Example 3

Let us consider the system

$$\dot{x} = u_1 f_1(x) + u_2 f_2(x) = u_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + u_2 \begin{pmatrix} 0 \\ 1 \\ x_1 \end{pmatrix} \tag{39}$$

where $x = (x_1, x_2, x_3) \in \mathbb{R}^3$. Clearly, the linearized system at $(x_0, u_0) = (0, 0)$, which reduces to $\dot{x} = u_1 f_1(0) + u_2 f_2(0)$, is not controllable. However, we shall see later that the system (39) is controllable. First, each point $(\pm\varepsilon, 0, 0)$ (resp. $(0, \pm\varepsilon, 0)$) may be reached from the origin by letting $u(t) = (\pm 1, 0)$ (resp. $u(t) = (0, \pm 1)$) on the time interval $[0, \varepsilon]$. More interestingly, any point $(0, 0, \varepsilon^2)$ may also be reached from the origin in a time $T = O(\varepsilon)$, even though $(0, 0, 1) \notin \mathrm{span}(f_1(0), f_2(0))$. To prove the last claim, let us introduce for $i = 1, 2$ the flow map ϕ_i^t defined by $\phi_i^t(x_0) = x(t)$, where $x(\cdot)$ solves

$$\dot{x} = f_i(x) \,, \quad x(0) = x_0 \,.$$

Then, it is easy to see that

$$\phi_2^{-\varepsilon} \phi_1^{-\varepsilon} \phi_2^{\varepsilon} \phi_1^{\varepsilon}(0) = (0, 0, \varepsilon^2) \,.$$

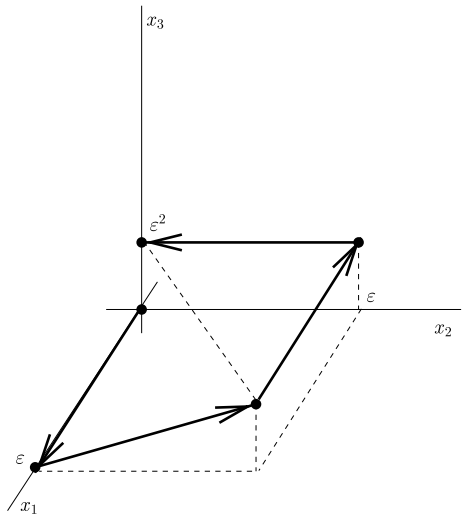

Finite Dimensional Controllability, Figure 1
Trajectory from $x_0 = (0, 0, 0)$ to $x_T = (0, 0, \varepsilon^2)$

It means that the control

$$u(t) = \begin{cases} (1,0) & \text{if } 0 \le t < \varepsilon \\ (0,1) & \text{if } \varepsilon \le t < 2\varepsilon \\ (-1,0) & \text{if } 2\varepsilon \le t < 3\varepsilon \\ (0,-1) & \text{if } 3\varepsilon \le t < 4\varepsilon \end{cases}$$

steers the solution of (39) from $x_0 = (0, 0, 0)$ at $t = 0$ to $x_T = (0, 0, \varepsilon^2)$ at $T = 4\varepsilon$. (See Fig. 1.) More generally, if f_1 and f_2 denote arbitrary smooth vector fields and ϕ_1^t, ϕ_2^t denote the corresponding flow maps, we have that

$$\phi_2^{-\varepsilon}\phi_1^{-\varepsilon}\phi_2^{\varepsilon}\phi_1^{\varepsilon}(0) = \varepsilon^2[f_1, f_2](0) + O(\varepsilon^3)\,. \tag{40}$$

Thus, it is possible to reach points in the direction of the Lie bracket $[f_1, f_2](0)$. However, the process is not very efficient: in the direction of the Lie bracket $[f_1, f_2](0)$, only points located at a distance from the origin of order ε^2 may be reached in a time of order ε.

Let us proceed to the controllability properties of affine systems. We first review the results for systems without drift (i. e. $f_0 = 0$), and next consider the general (only partially understood) situation where a drift exists.

Affine Systems Without Drift

We assume that $f_0 = 0$, so that (38) takes the form

$$\dot{x} = u_1 f_1(x) + \cdots + u_m f_m(x)\,. \tag{41}$$

When $m < n$, the linearized system at any stationary trajectory $(\overline{x}, 0)$, which reads

$$\dot{y} = (f_1(\overline{x}) \cdots f_m(\overline{x}))v\,, \tag{42}$$

cannot be controllable. High order tests are therefore very useful in this setting. We adopt the following

Definition 12 The system (41) is said to be **controllable** if for any $T > 0$ and any $x_0, x_T \in \mathbb{R}^n$, there exists a control $u \in L^\infty([0, T]\,; \mathbb{R}^m)$ such that the solution $x(t)$ of (41) emanating from x_0 at $t = 0$ reaches x_T at time $t = T$.

Note that the controllability may be supplemented with the words: local, with small controls, in small time, in large time, etc. An important instance, the small-time local controllability, is introduced below in Definition 14.

The following result has been obtained by Rashevski [28] and Chow [8].

Theorem 13 *(Rashevski–Chow) Assume that $f_i \in C^\infty(\mathbb{R}^n, \mathbb{R}^n)$ for all $i \in \{1, \ldots, m\}$. If*

$$\mathrm{Lie}(f_1, \ldots, f_m)(x) = \mathbb{R}^n \qquad \forall x \in \mathbb{R}^n\,, \tag{43}$$

then (41) is controllable.

Example 3 continued (39) is controllable, since $[f_1, f_2] = (0, 0, 1)$ gives

$$\mathrm{span}(f_1(x), f_2(x), [f_1, f_2](x)) = \mathbb{R}^3 \quad \forall x \in \mathbb{R}^3\,. \tag{44}$$

Example 2 continued Brockett's system is also controllable, since $[f_1, f_2] = (0, 0, 2)$ gives (43).

Notice that Theorem 13 is almost sharp. Indeed, it has been proved by Sussmann–Jurdjevic in [36] that (43) has to be satisfied for a controllable system (41) with real analytic vector fields (i. e. $f_i \in C^\omega(\mathbb{R}^n, \mathbb{R}^n)$ for each i).

Affine Systems with Drifts

We consider now a control system with a drift

$$\dot{x} = f_0(x) + u_1 f_1(x) + \cdots + u_m f_m(x) \tag{45}$$

where $f_i \in C^\infty(\mathbb{R}^n, \mathbb{R}^n)$ for any $i \in \{0, \ldots, m\}$, and $f_0(0) = 0$ but $f_0 \not\equiv 0$. As no control is associated with the vector field f_0, one expects that the latter will occupy a singular place in the controllability tests.

Let $\rho > 0$ be a given number. We assume that the control input u is a measurable function taking its values in the compact set $U_\rho = [-\rho, \rho]^m$. To emphasize the dependence in u, we shall sometimes denote by $x(t, u)$ the solution of (45) such that $x(0, u) = 0$. The **attainable set at time** $T > 0$ is the set

$$\mathcal{A}_\rho(T) = \{x(T, u); u(t) \in U_\rho \forall t \in [0, T]\}\,.$$

Definition 14 We shall say that the control system (45) is

- **accessible** from the origin if the interior of $\mathcal{A}_\rho(T)$ is nonempty for any $T > 0$;
- **small time locally controllable (STLC)** at the origin if for any $T > 0$ there exists a number $\delta > 0$ such that for any pair of states (x_0, x_T) with $|x_0| < \delta$, $|x_T| < \delta$, there exists a control input $u \in L^\infty([0, T]\,; U_\rho)$ such that the solution x of

$$\dot{x} = f_0(x) + u_1 f_1(x) + \cdots + u_m f_m(x), \quad x(0) = x_0 \quad (46)$$

satisfies $x(T) = x_T$.

(Notice that in the definition of the small time local controllability, certain authors assume $x_0 = 0$ in above definition, or require the existence of $\delta > 0$ for any $T > 0$ and any $\rho > 0$.)

The accessibility property turns out to be easy to characterize. As a STLC system is clearly accessible, the accessibility property is often considered as the first property to test before investigating the controllability of a given affine system.

The following result provides an accessibility test based upon the rank at the origin of the Lie algebra generated by all the vectors fields involved in the control system.

Theorem 15 *(Hermann–Nagano) If*

$$\dim \mathrm{Lie}(f_0, f_1, \ldots, f_m)(0) = n \quad (47)$$

then the system (45) is accessible from the origin. Conversely, if (45) is accessible from the origin and the vector fields $f_0, f_1, \ldots, f_m$ are real analytic, then (47) has to be satisfied.

Example 4 Pick any $k \in \mathbb{N}^*$ and consider as in [22] the system

$$\begin{cases} \dot{x}_1 = u, & |u| \le 1\,, \\ \dot{x}_2 = (x_1)^k \end{cases} \quad (48)$$

so that $n = 2$, $m = 1$, and $f_0(x) = (0, (x_1)^k)$, $f_1(x) = (1, 0)$. Setting

$$(\mathrm{ad}\, f_1, f_0) = [f_1, f_0]$$
$$(\mathrm{ad}^{i+1} f_1, f_0) = [f_1, (\mathrm{ad}^i f_1, f_0)] \qquad \forall i \ge 1$$

we obtain at once that

$$(\mathrm{ad}^k f_1, f_0)(x) = (0, k\,!)$$

hence

$$\mathrm{Lie}(f_0, f_1)(0) = \mathbb{R}^2\,.$$

It follows from Theorem 15 that (48) is accessible from the origin. On the other hand, it is clear that (48) is not STLC at the origin when k is even, since $\dot{x}_2 = (x_1)^k \ge 0$, hence x_2 is nondecreasing. Using a controllability test due to Hermes [17], it may be shown that (48) is STLC at the origin if and only if k is odd. Note that the linearized system at the origin fails to be controllable whenever $k \ge 2$.

Let us consider some affine system (45) which is accessible from the origin. We exclude the trivial situation when the linearization principle may be applied, and seek for a Lie algebraic condition ensuring that (45) is STLC. If h is a given iterated Lie bracket of $f_0, f_1, \ldots, f_m$, we let $\delta_i(h)$ denote the number of occurrences of f_i in the definition of h. For instance, if $m = 3$ and

$$h = [[f_1, [f_1, f_0]], [f_1, f_2]] \quad (49)$$

then

$$\delta_0(h) = 1\,, \delta_1(h) = 3\,, \delta_2(h) = 1\,, \delta_3(h) = 0\,.$$

Notice that the fields $f_0, f_1, \ldots, f_m$ are considered as indeterminates when computing the $\delta_i(h)$'s, their effective values as vector fields being ignored.

Let S_m denote the usual symmetric group, i. e. S_m is the group of all the permutations of the set $\{1, \ldots, m\}$. If $\psi \in S_m$ and h is an iterated Lie bracket of $f_0, f_1, \ldots, f_m$, we denote by h^ψ the iterated Lie bracket obtained by replacing, in the definition of h, f_i by $f_{\psi(i)}$ for each $i \in \{1, \ldots, m\}$. For instance, if $\psi = (1\,2\,3)$ and h is as in (49), then

$$h^\psi = [[f_2, [f_2, f_0]], [f_2, f_3]]\,.$$

Finally, we set

$$\sigma(h) = \sum_{\psi \in S_m} h^\psi\,.$$

With h as in (49), and $S_3 = \{\mathrm{id}_{\{1,2,3\}}, (1\,2\,3), (1\,3\,2), (1\,2), (1\,3), (2\,3)\}$, we obtain

$$\begin{aligned}\sigma(h) = {} & [[f_1, [f_1, f_0]], [f_1, f_2]] + [[f_2, [f_2, f_0]], [f_2, f_3]] \\ & + [[f_3, [f_3, f_0]], [f_3, f_1]] + [[f_2, [f_2, f_0]], [f_2, f_1]] \\ & + [[f_3, [f_3, f_0]], [f_3, f_2]] + [[f_1, [f_1, f_0]], [f_1, f_3]]\,.\end{aligned}$$

We need the following

Definition 16 Let $\theta \in [0, 1]$. We say that the system (45) satisfies the condition $S(\theta)$ if, for any iterated Lie bracket h with $\delta_0(h)$ odd and $\delta_i(h)$ even for all $i \in \{1, \ldots, m\}$, the

vector $\sigma(h)(0)$ belongs to the vector space spanned by the vectors $g(0)$ where g is an iterated Lie bracket satisfying

$$\theta\delta_0(g) + \sum_{i=1}^{m} \delta_i(g) < \theta\delta_0(h) + \sum_{i=1}^{m} \delta_i(h) .$$

Then we have the following result proved in [35].

Theorem 17 *(Sussmann) If the condition $S(\theta)$ is satisfied for some $\theta \in [0, 1]$, then the system (45) is small time locally controllable at the origin.*

When $\theta = 0$, Sussmann's theorem is nothing else than Hermes' theorem. Sussmann's theorem, which is in itself very useful in Engineering (see, e. g., [5]), has been extended in [1,4,18].

Controllability and Observability

Assume given a control system

$$\dot{x} = f(x, u) \tag{50}$$

together with an output function

$$y = h(x) \tag{51}$$

where $f\colon X \times U \to \mathbb{R}^n$ and $h\colon X \to Y$ are smooth functions, and $X \subset \mathbb{R}^n$, $U \subset \mathbb{R}^m$ and $Y \subset \mathbb{R}^p$ denote some open sets. y typically stands for a (partial) measurement of the state, e. g. the p first coordinates, where $p < n$. Often, only y is available, and for the stabilization of (50) we should consider an output feedback law of the form $u = k(y)$. We shall say that (50)–(51) is **observable** on the interval $[0, T]$ if, for any pair $x_0, \tilde{x}_0$ of points in X, one may find a control input $u \in L^\infty([0, T]\,; U)$ such that if x (resp. $\tilde{x}$) denotes the solution of (50) emanating from x_0 (resp. $\tilde{x}_0$) at time $t = 0$, and y (resp. $\tilde{y}$) denotes the corresponding output function, we have

$$y(t) \neq \tilde{y}(t) \quad \text{for some } t \in [0, T] . \tag{52}$$

For a time invariant linear control system and a linear output function

$$\dot{x} = Ax + Bu \tag{53}$$

$$y = Cx \tag{54}$$

with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, the output is found to be

$$y(t) = Cx(t) = Ce^{tA}x_0 + C\int_0^t e^{(t-s)A}Bu(s)\mathrm{d}s . \tag{55}$$

In particular,

$$y(t) - \tilde{y}(t) = Ce^{tA}(x_0 - \tilde{x}_0) \tag{56}$$

and B does not play any role. Therefore, (53)–(54) is observable in time T if and only if the only state $\overline{x} \in \mathbb{R}^n$ such that $Ce^{tA}\overline{x} = 0$ for any $t \in [0, T]$ is $\overline{x} = 0$. Differentiating in time and applying Cayley–Hamilton theorem, we obtain the following result.

Theorem 18 *System (53)–(54) is observable in time T if and only if* rank $\mathcal{O}(A, C) = n$, *where the **observability matrix** $\mathcal{O}(A, C) \in \mathbb{R}^{np\times n}$ is defined by*

$$\mathcal{O}(A, C) = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix} .$$

Noticing that

$$\mathcal{R}(A, B)^* = \mathcal{O}(A^*, B^*)$$

and introducing the adjoint system

$$\dot{\tilde{x}} = A^*\tilde{x} \tag{57}$$

$$\tilde{y} = B^*\tilde{x} \tag{58}$$

we arrive to the following duality principle.

Theorem 19 *A time invariant linear system is controllable if and only if its adjoint system is observable.*

Notice that the observability of the adjoint system is easily shown to be equivalent to the existence of a constant $\delta > 0$ such that

$$\int_0^T ||B^*e^{tA^*}x_0||^2\mathrm{d}t \geq \delta||x_0||^2 \quad \forall x_0 \in \mathbb{R}^n . \tag{59}$$

As it has been pointed out in Sect. "Linear Systems", the equivalence between the controllability of a system and the observability of its adjoint, expressed in the form (59), is still true in infinite dimension, and provides a very useful way to investigate the controllability of a partial differential equation.

Finally, using the duality principle, we see that any controllability test gives rise to an observability test, and vice-versa.

Controllability and Stabilizability

In this section we shall explore the connections between the controllability and the stabilizability of a system. Let us begin with a linear control system

$$\dot{x} = Ax + Bu\,. \tag{60}$$

Performing a linear change of coordinates if needed, we may assume that (60) has the block structure given by the Kalman controllability decomposition

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} A_1 & A_2 \\ 0 & A_3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} B_1 \\ 0 \end{pmatrix} u \tag{61}$$

where $A_1 \in \mathbb{R}^{r\times r}$, $B_1 \in \mathbb{R}^{r\times m}$, and where $\operatorname{rank}\mathcal{R}(A_1, B_1) = r$. For any square matrix $M \in \mathbb{C}^{n\times n}$ we denote by $\sigma(M)$ its spectrum, i. e.

$$\sigma(M) = \{\lambda \in \mathbb{C}\,; \det(M - \lambda I) = 0\}\,.$$

Let us note $\mathbb{C}_- := \{\lambda \in \mathbb{C}\,; \operatorname{Re}\lambda < 0\}$. Then the asymptotic stabilizability of (60) may be characterized as follows.

Theorem 20 *There exists a (continuous) asymptotically stabilizing feedback law $u = k(x)$ with $k(0) = 0$ for (60) if and only if $\sigma(A_3) \subset \mathbb{C}_-$. If it is the case, then for any family $S = (\lambda_i)_{1\le i\le r}$ of elements of $\mathbb{C}_-$ invariant by conjugation, there exists a linear asymptotically stabilizing feedback law $u(x) = Kx$, with $K \in \mathbb{R}^{m\times n}$, such that*

$$\sigma(A + BK) = \sigma(A_3) \cup S\,. \tag{62}$$

The property (62), which shows to what extend the spectrum of the closed loop system can be assigned, is referred to as the **pole shifting theorem**.

As a direct consequence of Theorem 20, we obtain the following result.

Corollary 21 *A time invariant linear system which is controllable is also asymptotically stabilizable.*

It is natural to ask whether Corollary 21 is still true for a nonlinear system, i. e. if a controllable system is necessarily asymptotically stabilizable. A general result cannot be obtained, as is shown by Brockett's system.

Example 2 continued Brockett's system (30)–(32) may be written $\dot{x} = u_1 f_1(x) + u_2 f_2(x) = f(x, u)$, and it follows from Theorem 13 that (30)–(32) is controllable around $(x_0, u_0) = (0, 0)$. Let us now recall a necessary condition for stabilizability due to R. Brockett [6].

Brockett third condition for stabilizability. Let $f \in C(\mathbb{R}^n \times \mathbb{R}^m\,; \mathbb{R}^n)$ with $f(x_0, u_0) = 0$. If the control system $\dot{x} = f(x, u)$ can be locally asymptotically stabilized at x_0 by means of a continuous feedback law u satisfying $u(x_0) = u_0$, then the image by f of any open neighborhood of $(x_0, u_0) \in \mathbb{R}^n \times \mathbb{R}^m$ contains an open neighborhood of $0 \in \mathbb{R}^n$.

Here, for any neighborhood V of (x_0, u_0) in $\mathbb{R}^5$, $f(V)$ does not cover an open neighborhood of 0 in $\mathbb{R}^3$, since $(0, 0, \varepsilon) \notin f(V)$ for any $\varepsilon \in \mathbb{R}$. According to Brockett's condition, the system (30)–(32) is not asymptotically stabilizable at the origin.

Thus a controllable system may fail to be asymptotically stabilizable by a continuous feedback law $u = k(x)$ due to topological obstructions. It turns out that this phenomenon does not occur when the phase space is the plane, as it has been demonstrated by Kawski in [21].

Theorem 22 *(Kawski) Let f_0, f_1 be real analytic vector fields on $\mathbb{R}^2$ with $f_0(0) = 0$ and $f_1(0) \neq 0$. Assume that the system*

$$\dot{x} = f_0(x) + u f_1(x)\,, \qquad u \in \mathbb{R} \tag{63}$$

is small time locally controllable at the origin. Then it is also asymptotically stabilizable at the origin by a Hölder continuous feedback law $u = k(x)$.

In larger dimension ($n \ge 3$), a way to go round the topological obstruction is to consider a time-varying feedback law $u = k(x, t)$. It has been first observed by Sontag and Sussmann in [33] that for one-dimensional state and input ($n = m = 1$), the controllability implies the asymptotic stabilizability by means of a time-varying feedback law. This kind of stabilizability was later established by Samson in [31] for Brockett's system ($n = 3$ and $m = 2$). Finally, using the return method, Coron proved that the implication

Controllability $\Rightarrow$ Asymptotic Stabilizability by Time-Varying Feedback

was a principle verified in most cases. To state precise results, we consider affine systems

$$\dot{x} = f_0(x) + \sum_{i=1}^{m} u_i f_i(x)\,, \qquad x \in \mathbb{R}^n$$

and distinguish again two cases: (1) $f_0 \equiv 0$ (no drift); (2) $f_0 \not\equiv 0$ (a drift).

(1) System without drift

Theorem 23 *(Coron [9]) Assume that (43) holds for the system (41). Pick any number $T > 0$. Then there exists a feedback law $u = (u_1, \ldots, u_m) \in C^\infty(\mathbb{R}^n \times \mathbb{R}, \mathbb{R}^m)$ such that*

$$u(0, t) = 0 \quad \forall t \in \mathbb{R} \tag{64}$$

$$u(x, t + T) = u(x, t) \quad \forall (x, t) \in \mathbb{R}^n \times \mathbb{R} \tag{65}$$

and such that 0 is globally asymptotically stable for the system

$$\dot{x} = \sum_{i=1}^{m} u_i(x,t) f_i(x) .$$

(2) System with a drift

Theorem 24 *(Coron [10]) Assume that the system (45) satisfies the condition $S(\theta)$ for some $\theta \in [0,1]$. Assume also that $n \notin \{2,3\}$ and that*

$$\mathrm{Lie}(f_0, f_1, \ldots, f_m)(0) = \mathbb{R}^n .$$

Pick any $T > 0$. Then there exists a feedback law $u = (u_1, \ldots, u_m) \in C^0(\mathbb{R}^n \times \mathbb{R}, \mathbb{R}^m)$ with $u \in C^\infty((\mathbb{R}^n \setminus \{0\}) \times \mathbb{R}, \mathbb{R}^m)$ such that (64)–(65) hold and such that 0 is locally asymptotically stable for the system

$$\dot{x} = f_0(x) + \sum_{i=1}^{m} u_i(x,t) f_i(x) .$$

Flatness

While powerful criterions enable us to decide whether a control system is controllable or not, most of them do not provide any indication on how to design an explicit control input steering the system from a point to another one. Fortunately, there exists a large class of systems, the so-called **flat systems**, for which explicit control inputs may easily be found. The flatness theory has been introduced by Fliess, Levine, Martin, and Rouchon in [15], and since then it has attracted the interest of many researchers thanks to its numerous applications in Engineering. Here, we only sketch the main ideas, referring the interested reader to [27] for a comprehensive introduction to the subject.

Let us consider a smooth control system

$$\dot{x} = f(x,u) , \quad x \in \mathbb{R}^n , u \in \mathbb{R}^m$$

given together with an output $y \in \mathbb{R}^m$ depending on x, u, and a finite number of derivatives of u

$$y = h(x, u, \dot{u}, \ldots, u^{(r)}) .$$

Following [15], we shall say that y is a **flat output** if the components of y are differentially independent, and both x and u may be expressed as functions of y and of a finite number of its derivatives

$$x = k(y, \dot{y}, \ldots, y^{(p)}) \tag{66}$$

$$u = l(y, \dot{y}, \ldots, y^{(q)}) . \tag{67}$$

In (66)–(67), p and q denote some nonnegative integers, and k and l denote some smooth functions. Since the state x and the input u are parameterized by the flat output y, to solve the controllability problem

$$\dot{x} = f(x,u) \tag{68}$$

$$x(0) = x_0 , \quad x(T) = x_T \tag{69}$$

it is sufficient to pick any function $y \in C^{\max(p,q)}([0,T]; \mathbb{R}^m)$ such that

$$k(y, \dot{y}, \ldots, y^{(p)})(0) = x(0) = x_0 \tag{70}$$

$$k(y, \dot{y}, \ldots, y^{(p)})(T) = x(T) = x_T . \tag{71}$$

The constraints (70)–(71) are generally very easy to satisfy. The desired control input u is then given by (67). Let us show how this program may be carried out on two simple examples.

Example 5 For the simple integrator

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= u \end{aligned}$$

the output $y = x_1$ is flat, for $x_2 = \dot{y}$ and $u = \ddot{y}$. The output $z = x_2$ is not flat, as $x_1 = \int z(t)\,\mathrm{d}t$. To steer the system from $x_0 = (0,0)$ to $x_T = (1,0)$ in time T, we only have to pick a function $y \in C^2([0,T]; \mathbb{R})$ such that

$$y(0) = 0 , \quad \dot{y}(0) = 0 , \quad y(T) = 1 , \quad \text{and} \quad \dot{y}(T) = 0 .$$

Clearly, $y(t) = t^2(2T - t)^2/T^4$ is convenient.

Example 6 Consider now the nonlinear system

$$\dot{x}_1 = u_1 \tag{72}$$

$$\dot{x}_2 = u_2 \tag{73}$$

$$\dot{x}_3 = x_2 u_1 . \tag{74}$$

Eliminating the input u_1 in (72)–(74) yields $\dot{x}_3 = x_2 \dot{x}_1$, so that x_2 may be expressed as a function of x_1, x_3 and their derivatives. The same is true for u_2, thanks to (73). Let us pick

$$y = (y_1, y_2) = (x_1, x_3) .$$

We claim that y is a flat output. Indeed,

$$(x_1, x_2, x_3) = (y_1, \frac{\dot{y}_2}{\dot{y}_1}, y_2) , \tag{75}$$

$$(u_1, u_2) = (\dot{y}_1, \frac{\ddot{y}_2 \dot{y}_1 - \dot{y}_2 \ddot{y}_1}{(\dot{y}_1)^2}) . \tag{76}$$

Pick $x_0 = (0,0,0)$ and $x_T = (0,0,1)$. Notice that, by the mean value theorem, $\dot{y}_1$ has to vanish somewhere, say at $t = \bar{t}$. We shall construct y_1 in such a way that $\dot{y}_1$ vanishes *only* at $t = \bar{t}$. For x_2 not to be singular at $\bar{t}$, we have to impose the condition

$$\dot{y}_2(\bar{t}) = 0\,.$$

If y_1 and y_2 are both analytic near $\bar{t}$, we notice that $x_2 = \dot{y}_2/\dot{y}_1$ is analytic near $\bar{t}$, hence $u_2 = \dot{x}_2$ is well-defined (and analytic) near $\bar{t}$. To steer the solution of (72)–(74) from $x_0 = (0,0,0)$ to $x_T = (0,0,1)$, it is then sufficient to pick a function $y = (y_1, y_2) \in C^\omega([0,T];\mathbb{R}^2)$ such that

$$\begin{aligned}
&y_1(0) = y_2(0) = 0, \quad \dot{y}_2(0) = 0, \quad \text{and } \dot{y}_1(0) \neq 0\,,\\
&y_1(T) = 0, \quad y_2(T) = 1, \quad \dot{y}_2(T) = 0, \quad \text{and } \dot{y}_1(T) \neq 0\,.\\
&\dot{y}_1\left(\frac{T}{2}\right) = 0, \quad \dot{y}_2\left(\frac{T}{2}\right) = 0, \quad \ddot{y}_1\left(\frac{T}{2}\right) \neq 0\\
&\text{and } \dot{y}_1(t) \neq 0 \text{ for } t \neq \frac{T}{2}\,.
\end{aligned}$$

Clearly,

$$(y_1(t), y_2(t)) = \left(t(T-t), \frac{t^4}{4} - \frac{Tt^3}{2} + \frac{T^2t^2}{4}\right)$$

is convenient.

Future Directions

Lie algebraic techniques have been used to provide powerful controllability tests for affine systems. However, there is still an important gap between the known necessary conditions (e. g. the Legendre Clebsh condition or its extensions) and the sufficient conditions (e. g. the $S(\theta)$ condition) for the small time local controllability. On the other hand, it has been noticed by Kawski in ([22], Example 6.1) that certain systems can be controlled on small time intervals $[0, T]$ only by using faster switching control variations, the number of switchings tending to infinity as T tends to zero. As for switched systems [2], it is not clear whether purely algebraic conditions be sufficient to characterize the controllability of systems with drift.

Of great interest for applications is the development of methods providing explicit control inputs, or control inputs computed in real time with the aid of some numerical schemes, both for the motion planning and the stabilization issue. The flatness theory seems to be a very promising method, and it has been successfully applied in Engineering. An active research is devoted to filling the gap between necessary and sufficient conditions for the existence of flat outputs, and to extending the theory to partial differential equations.

The control of nonlinear partial differential equations may sometimes be reduced to the control of a family of finite-dimensional systems by means of the Galerkin method [3]. On the other hand, the spatial discretization of partial differential equations by means of finite differences, finite elements, or spectral methods leads in a natural way to the control of finite dimensional systems (see, e. g., [38]). Of great importance is the *uniform boundedness* with respect to the discretization parameter of the $L^2(0,T)$-norms of the control inputs associated with the finite dimensional approximations.

Geometric ideas borrowed from the control of finite dimensional systems (e. g. the return method, the power series expansion, the quasistatic deformation) have been applied to prove the controllability of certain nonlinear partial differential equations whose linearization fails to be controllable. (See [11] for a survey of these techniques.) The Korteweg–de Vries equation provides an interesting example of a partial differential equation whose linearization fails to be controllable for certain lengths of the space domain [29]. However, for these critical lengths the reachable space proves to be of finite codimension, and it may be proved that the full equation is controllable by using the nonlinear term in order to reach the missing directions [7,12].

Bibliography

Primary Literature

1. Agrachev AA, Gamkrelidze RV (1993) Local controllability and semigroups of diffeomorphisms. Acta Appl Math 32:1–57
2. Agrachev A, Liberzon D (2001) Lie-algebraic stability criteria for switched systems. SIAM J Control Optim 40(1):253–269
3. Agrachev A, Sarychev A (2005) Navier–Stokes equations: controllability by means of low modes forcing. J Math Fluid Mech 7(1):108–152
4. Bianchini RM, Stefani G (1986) Sufficient conditions of local controllability. In: Proc. 25th IEEE–Conf. Decision & Control. Athens
5. Bonnard B (1992) Contrôle de l'altitude d'un satellite rigide. RAIRO Autom Syst Anal Control 16:85–93
6. Brockett RW (1983) Asymptotic stability and feedback stabilization. In: Brockett RW, Millman RS, Sussmann HJ (eds) Differential Geometric Control Theory. Progr. Math., vol 27. Birkhäuser, Basel, pp 181–191
7. Cerpa E, Crépeau E (2007) Boundary controllability for the nonlinear Korteweg–de Vries equation on any critical domain
8. Chow WL (1940) Über Systeme von linearen partiellen Differentialgleichungen erster Ordnung. Math Ann 117:98–105
9. Coron JM (1992) Global asymptotic stabilization for controllable systems without drift. Math Control Signals Syst 5: 295–312

10. Coron JM (1995) Stabilization in finite time of locally controllable systems by means of continuous time-varying feedback laws. SIAM J Control Optim 33:804–833
11. Coron JM (2007) Control and nonlinearity, mathematical surveys and monographs. American Mathematical Society, Providence
12. Coron JM, Crépeau E (2004) Exact boundary controllability of a nonlinear KdV equation with critical lengths. J Eur Math Soc 6(3):367–398
13. Curtain RF, Pritchard AJ (1978) Infinite Dimensional Linear Systems Theory. Springer, New York
14. Dolecky S, Russell DL (1977) A general theory of observation and control. SIAM J Control Optim 15:185–220
15. Fliess M, Lévine J, Martin PH, Rouchon P (1995) Flatness and defect of nonlinear systems: introductory theory and examples. Intern J Control 31:1327–1361
16. Hautus MLJ (1969) Controllability and observability conditions for linear autonomous systems. Ned Akad Wetenschappen Proc Ser A 72:443–448
17. Hermes H (1982) Control systems which generate decomposable Lie algebras. J Differ Equ 44:166–187
18. Hermes H, Kawski M (1986) Local controllability of a single-input, affine system. In: Proc. 7th Int. Conf. Nonlinear Analysis. Dallas
19. Hirschorn R, Lewis AD (2004) High-order variations for families of vector fields. SIAM J Control Optim 43:301–324
20. Kalman RE (1960) Contribution to the theory of optimal control. Bol Soc Mat Mex 5:102–119
21. Kawski M (1989) Stabilization of nonlinear systems in the plane. Syst Control Lett 12:169–175
22. Kawski M (1990) High-order small time local controllability. In: Sussmann HJ (ed) Nonlinear controllability and optimal control. Textbooks Pure Appl. Math., vol 113. Dekker, New York, pp 431–467
23. Khodja FA, Benabdallah A, Dupaix C, Kostin I (2005) Null-controllability of some systems of parabolic type by one control force. ESAIM Control Optim Calc Var 11(3):426–448
24. Lions JL (1988) Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués. In: Recherches en Mathématiques Appliquées, Tome 1, vol 8. Masson, Paris
25. Liu K (1997) Locally distributed control and damping for the conservative systems. SIAM J Control Optim 35:1574–1590
26. Lobry C (1970) Contrôlabilité des systèmes non linéaires. SIAM J Control 8:573–605
27. Martin PH, Murray RM, Rouchon P (2002) Flat systems. Mathematical Control Theory, Part 1,2 (Trieste 2001). In: ICTP Lect. Notes, vol VIII. Abdus Salam Int Cent Theoret Phys, Trieste
28. Rashevski PK (1938) About connecting two points of complete nonholonomic space by admissible curve. Uch Zapiski ped inst Libknexta 2:83–94
29. Rosier L (1997) Exact boundary controllability for the Korteweg–de Vries equation on a bounded domain. ESAIM: Control Optim Calc Var 2:33–55
30. Rosier L (2007) A Survey of controllability and stabilization results for partial differential equations. J Eur Syst Autom (JESA) 41(3–4):365–412
31. Samson C (1991) Velocity and torque feedback control of a nonholonomic cart. In: de Witt C (ed) Proceedings of the International workshop on nonlinear and adaptative control: Issues in robotics, Grenoble, France nov. 1990. Lecture Notes in Control and Information Sciences, vol 162. Springer, Berlin, pp 125–151
32. Silverman LM, Meadows HE (1967) Controllability and observability in time-variable linear systems. SIAM J Control 5:64–73
33. Sontag ED, Sussmannn H (1980) Remarks on continuous feedbacks. In: Proc. IEEE Conf. Decision and Control, (Albuquerque 1980). pp 916–921
34. Stefani G (1985) Local controllability of nonlinear systems: An example. Syst Control Lett 6:123–125
35. Sussmann H (1987) A general theorem on local controllability. SIAM J Control Optim 25:158–194
36. Sussmann H, Jurdjevic V (1972) Controllability of nonlinear systems. J Differ Equ 12:95–116
37. Tret'yak AI (1991) On odd-order necessary conditions for optimality in a time-optimal control problem for systems linear in the control. Math USSR Sbornik 79:47–63
38. Zuazua E (2005) Propagation, observation and control of waves approximated by finite difference methods. SIAM Rev 47(2):197–243

Books and Reviews

Agrachev AA, Sachkov YL (2004) Control theory from the geometric viewpoint. In: Encyclopaedia of Mathematical Sciences, vol 87. Springer, Berlin
Bacciotti A (1992) Local stabilization of nonlinear control systems. In: Ser. Adv. Math. Appl. Sci, vol 8. World Scientific, River Edge
Bacciotti A, Rosier L (2005) Liapunov functions and stability in control theory. Springer, Berlin
Isidori A (1989) Nonlinear control systems. Springer, Berlin
Nijmeijer H, van der Schaft AJ (1990) Nonlinear dynamical control systems. Springer, New York
Sastry S (1999) Nonlinear systems. Springer, New York
Sontag E (1990) Mathematical control theory. Springer, New York
Zabczyk J (1992) Mathematical control theory: An introduction. Birkhäuser, Boston

Firing Squad Synchronization Problem in Cellular Automata

Hiroshi Umeo
Electro-Communication, University of Osaka, Osaka, Japan

Article Outline

Glossary
Definition of the Subject
Introduction
Firing Squad Synchronization Problem
Variants of the Firing Squad Synchronization Problem
Firing Squad Synchronization Problem on Two-dimensional Arrays
Summary and Future Directions
Bibliography

Glossary

Cellular automaton A cellular automaton is a discrete computational model studied in mathematics, computer science, economics, biology, physics and chemistry etc. It consists of a regular array of cells, each cell is a finite state automaton. The array can be in any finite number of dimensions. Time (step) is also discrete, and the state of a cell at time t (≥ 1) is a function of the states of a finite number of cells (called its neighborhood) at time $t-1$. Each cell has a same rule set for updating its next state, based on the states in the neighborhood. At every step the rules are applied to the whole array synchronously, yielding a new configuration.

Time-space diagram A time-space diagram is frequently used to represent signal propagations in one-dimensional cellular space. Usually, the time is drawn on the vertical axis and the space on the horizontal axis. The trajectories of individual signals in propagation are expressed in this diagram by sloping lines. The slope of the line represents the propagation speed of the signal. Time-space-diagrams that show the position of individual signals in time and in space are very useful for understanding cellular algorithms, signal propagations and crossings in the cellular space.

Definition of the Subject

The firing squad synchronization problem (**FSSP** for short) is formalized in terms of the model of cellular automata. Figure 1 shows a finite one-dimensional cellular array consisting of n cells, denoted by C_i, where $1 \leq i \leq n$. All cells (except the end cells) are identical finite state automata. The array operates in lock-step mode such that the next state of each cell (except the end cells) is determined by both its own present state and the present states of its right and left neighbors. All cells (*soldiers*), except the left end cell, are initially in the *quiescent* state at time $t = 0$ and have the property whereby the next state of a quiescent cell having quiescent neighbors is the quiescent state. At time $t = 0$ the left end cell (*general*) is in the *fire-when-ready* state, which is an initiation signal to the array.

The firing squad synchronization problem is stated as follows: Given an array of n identical cellular automata, including a *general* on the left end which is activated at time $t = 0$, one wants to give the description (state set and next-state transition function) of the automata so that, *at some future time*, all of the cells will *simultaneously* and, *for the first time*, enter a special *firing* state. The set of states and the next-state transition function must be independent of n. Without loss of generality, it is assumed that $n \geq 2$. The tricky part of the problem is that the same kind of soldiers having a fixed number of states must be synchronized, regardless of the length n of the array. The problem itself is interesting as a mathematical puzzle and a good example of recursive divide-and-conquer strategy operating in parallel. It has been referred to as *achieving macro-synchronization given in micro-synchronization systems* and realizing a *global synchronization using only local information exchange* [10].

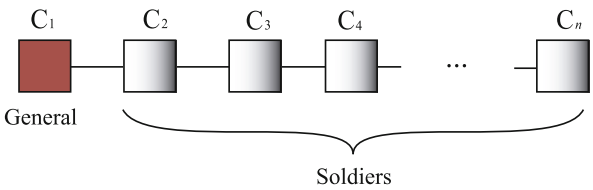

Firing Squad Synchronization Problem in Cellular Automata, Figure 1
One-dimensional cellular automaton

Introduction

Cellular automata are considered to be a nice model of complex systems in which an infinite one-dimensional array of finite state machines (cells) updates itself in synchronous manner according to a uniform local rule. A comprehensive study is made for a synchronization problem that gives a finite-state protocol for synchronizing a large scale of cellular automata. Synchronization of general network is a computing primitive of parallel and distributed computations. The synchronization in cellular automata has been known as *firing squad synchronization problem* since its development, in which it was originally proposed by J. Myhill in Moore [29] to synchronize all parts of self-reproducing cellular automata. The problem has been studied extensively for more than 40 years [1–80].

The present article firstly examines the state transition rule sets for the famous firing squad synchronization algorithms that give a finite-state protocol for synchronizing large-scale cellular automata, focusing on the fundamental synchronization algorithms operating in optimum steps on one-dimensional cellular arrays. The algorithms discussed herein are the Goto's first algorithm [12], the eight-state Balzer's algorithm [1], the seven-state Gerken's algorithm [9], the six-state Mazoyer's algorithm [25], the 16-state Waksman's algorithm [74] and a number of revised versions thereof. In addition, the article constructs a survey of current optimum-time synchronization algorithms and compares their transition rule sets with respect

to the number of internal states of each finite state automaton, the number of transition rules realizing the synchronization, and the number of state-changes on the array. It also presents herein a survey and a comparison of the quantitative and qualitative aspects of the optimum-time synchronization algorithms developed thus far for one-dimensional cellular arrays. Then, it provides several variants of the firing squad synchronization problems including fault-tolerant synchronization protocols, one-bit communication protocols, non-optimum-time algorithms, and partial solutions etc. Finally, a survey on two-dimensional firing squad synchronization algorithms is presented. Several new results and viewpoints are also given.

Firing Squad Synchronization Problem

A Brief History of the Developments of Firing Squad Synchronization Algorithms

The problem known as the *firing squad synchronization problem* was devised in 1957 by J. Myhill, and first appeared in print in a paper by E.F. Moore [29]. This problem has been widely circulated, and has attracted much attention. The firing squad synchronization problem first arose in connection with the need to simultaneously turn on all parts of a self-reproducing machine. The problem was first solved by J. McCarthy and M. Minsky [28] who presented a non-optimum-time synchronization scheme that operates in $3n+O(1)$ steps for synchronizing n cells. In 1962, the first optimum-time, i. e. $(2n-2)$-step, synchronization algorithm was presented by Goto [12], with each cell having several thousands of states. Waksman [74] presented a 16-state optimum-time synchronization algorithm. Afterward, Balzer [1] and Gerken [9] developed an eight-state algorithm and a seven-state algorithm, respectively, thus decreasing the number of states required for the synchronization. In 1987, Mazoyer [25] developed a six-state synchronization algorithm which, at present, is the algorithm having the fewest states.

Firing Squad Synchronization Algorithm

Section "Firing Squad Synchronization Algorithm" briefly sketches the design scheme for the firing squad synchronization algorithm according to Waksman [74] in which the first transition rule set was presented. It is quoted from Waksman [74].

> The code book of the state transitions of machines is so arranged to cause the array to progressively divide itself into 2^k equal parts, where k is an integer and an increasing function of time. The end machines in each partition assume a special state so that when the last partition occurs, all the machines have for both neighbors machines at this state. This is made the only condition for any machine to assume terminal state.

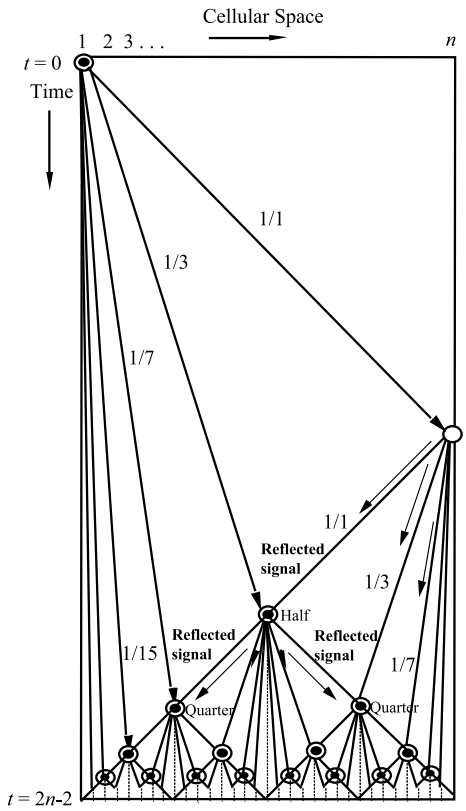

Firing Squad Synchronization Problem in Cellular Automata, Figure 2
Time-space diagram for Waksman's optimum-time firing squad synchronization algorithm

Figure 2 is a time-space diagram for the Waksman's optimum-step firing squad synchronization algorithm. The general at time $t = 0$ emits an infinite number of signals which propagate at $1/(2^{k+1}-1)$ speed, where k is positive integer. These signals meet with a reflected signal at half point, quarter points, ..., etc., denoted by $\odot$ in Fig. 2. It is noted that these cells indicated by $\odot$ are synchronized. By increasing the number of synchronized cells exponentially, eventually all of the cells are synchronized.

Complexity Measures and Properties in Firing Squad Synchronization Algorithms

Time Complexity Any solution to the firing squad synchronization problem can easily be shown to require

$(2n-2)$-steps for synchronizing n cells, since signals on the array can propagate no faster than one cell per step, and the time from the general's instruction until the synchronization must be at least $2n-2$. See Balzer [1], Goto [12] and Waksman [74] for a proof. The next two theorems show the optimum-time complexity for synchronizing n cells on one-dimensional arrays.

Theorem 1 ([1,12,74]) *Synchronization of n cells in less than $(2n-2)$-steps is impossible.*

Theorem 2 ([1,12,74]) *Synchronization of n cells at exactly $(2n-2)$-steps is possible.*

Number of States The following three distinct states: the *quiescent* state, the *general* state, and the *firing* state, are required in order to define any cellular automaton that can solve the firing squad synchronization problem. The boundary state for C_0 and C_{n+1} is not generally counted as an internal state. Balzer [1] implemented a search strategy in order to prove that there exists no four-state solution. He showed that no four-state optimum-time solution exists. Sanders [40] studied a similar problem on a parallel computer and showed that the Balzer's backtrack heuristic was not correct, rendering the proof incomplete and gave a proof based on a computer simulation for the non-existence of four-state solution. Balzer [1] also showed that there exist no five-state optimum-time solution satisfying special conditions. It is noted that the Balzer's special conditions do not hold for the Mazoyer's six-state solution with the fewest states known at present. The question that remains is: "What is the minimum number of states for an optimum-time solution of the problem?" At present, that number is *five* or *six*. Section "Partial Solutions" gives some 4- and 5-state partial solutions that can synchronize infinite cells, but not all.

Theorem 3 ([1,40]) *There is no four-state CA that can synchronize n cells.*

Berthiaume, Bittner, Perković, Settle and Simon [2] considered the state lower bound on ring-connected cellular automata. It is shown that there exists no three-state solution and no four-state symmetric solution for rings.

Theorem 4 ([2]) *There is no four-state symmetric optimum-time solution for ring-connected cellular automata.*

Number of Transition Rules Any k-state transition table for the synchronization has at most $(k-1)k^2$ entries in $(k-1)$ matrices of size $k \times k$. The number of transition rules reflects the complexity of synchronization algorithms.

Transition Rule Sets for Optimum-Time Firing Squad Synchronization Algorithms

Section "Transition Rule Sets for Optimum-Time Firing Squad Synchronization Algorithms" implements most of the transition rule sets for the synchronization algorithms above mentioned on a computer and check whether these rule sets yield successful firing configurations at exactly $t = 2n-2$ steps for any n such that $2 \le n \le 10,000$.

Waksman's 16-State Algorithm Waksman [74] proposed a 16-state firing squad synchronization algorithm, which, together with an unpublished algorithm by Goto [12], is referred to as the first-in-the-world optimum-time synchronization algorithm. Waksman presented the first set of transition rules described in terms of a state transition table that is defined on the following state set $\mathcal{D}$ consisting of 16 states such that $\mathcal{D} = \{Q, T, P_0, P_1, B_0, B_1, R_0, R_1, A_{000}, A_{001}, A_{010}, A_{011}, A_{100}, A_{101}, A_{110}, A_{111}\}$, where Q is a quiescent state, T is a firing state, P_0 and P_1 are prefiring states, B_0 and B_1 are states for signals propagating at various speeds, R_0 and R_1 are trigger states which cause the B_0 and B_1 states move in the left or right direction and A_{ijk}, $i, j, k \in \{0, 1\}$ are control states which generate the state R_0 or R_1 either with a unit delay or without any delay. The state P_0 also acts as an initial general.

USN Transition Rule Set Cellular automata researchers have reported that some errors are included in the Waksman's transition table. A computer simulation made in Umeo, Sogabe and Nomura [64] reveals this to be true. They corrected some errors included in the original Waksman's transition rule set. The correction procedures can be found in Umeo, Sogabe and Nomura [64]. This subsection gives a complete list of the transition rules which yield successful synchronizations for any n. Figure 3 is the complete list, which consists of 202 transition rules. The list is referred to as the *USN transition rule set*. In the correction, a ninety-three percent reduction in the number of transition rules is realized compared to the Waksman's original list. The computer simulation based on the table of Fig. 3 gives the following observation. Figure 4 shows snapshots of the Waksman's 16-state optimum-time synchronization algorithm on 21 cells.

Observation 3.1 ([54,64]) *The set of rules given in Fig. 3 is the smallest transition rule set for Waksman's optimum-time firing squad synchronization algorithm.*

Balzer's Eight-State Algorithm Balzer [1] constructed an eight-state, 182-rule synchronization algorithm and

Q State				
1:	Q	Q	Q →	Q
2:	Q	Q	B0 →	Q
3:	Q	Q	B1 →	Q
4:	Q	Q	R0 →	R0
5:	Q	Q	R1 →	Q
6:	Q	Q	P0 →	A000
7:	Q	Q	P1 →	A100
8:	Q	Q	A000 →	A001
9:	Q	Q	A001 →	A000
10:	Q	Q	A100 →	A101
11:	Q	Q	A101 →	A100
12:	Q	Q	A010 →	R0
13:	Q	Q	A110 →	Q
14:	Q	Q	* →	Q
15:	B0	Q	Q →	Q
16:	B0	Q	B0 →	Q
17:	B0	Q	R0 →	R0
18:	B0	Q	P1 →	A100
19:	B0	Q	A000 →	A001
20:	B0	Q	A101 →	A100
21:	B0	Q	A010 →	R0
22:	B0	Q	A110 →	Q
23:	B1	Q	Q →	Q
24:	B1	Q	B1 →	Q
25:	B1	Q	R0 →	R0
26:	B1	Q	R1 →	Q
27:	B1	Q	P0 →	A000
28:	B1	Q	A001 →	A000
29:	B1	Q	A100 →	A101
30:	R0	Q	Q →	Q
31:	R0	Q	B1 →	Q
32:	R0	Q	P0 →	A000
33:	R0	Q	A000 →	A001
34:	R0	Q	A001 →	A000
35:	R0	Q	A011 →	Q
36:	R1	Q	Q →	R1
37:	R1	Q	B0 →	R1
38:	R1	Q	B1 →	R1
39:	P0	Q	Q →	A010
40:	P0	Q	B1 →	A010
41:	P0	Q	R1 →	A010
42:	P0	Q	* →	P1
43:	P1	Q	Q →	A110
44:	P1	Q	B0 →	A110
45:	A000	Q	Q →	R1
46:	A000	Q	B0 →	R1
47:	A001	Q	R1 →	Q
48:	A100	Q	Q →	Q
49:	A100	Q	B0 →	Q
50:	A010	Q	Q →	A011
51:	A010	Q	B0 →	A011
52:	A010	Q	R1 →	A011
53:	A010	Q	* →	P0
54:	A011	Q	Q →	A010
55:	A011	Q	B1 →	A010
56:	A011	Q	R1 →	A010
57:	A011	Q	* →	P1
58:	A110	Q	Q →	A111
59:	A110	Q	B1 →	A111
60:	A111	Q	Q →	A110
61:	A111	Q	B0 →	A110
B0 State				
62:	Q	B0	Q →	B0
63:	Q	B0	R0 →	R0
64:	Q	B0	P0 →	B0
65:	Q	B0	P1 →	B0
66:	Q	B0	A001 →	P1
67:	Q	B0	A100 →	P1
68:	B1	B0	P0 →	B0
69:	B1	B0	P1 →	B0
70:	R1	B0	Q →	R1
71:	R1	B0	P0 →	R1
72:	R1	B0	P1 →	R1
73:	P0	B0	Q →	B0
74:	P0	B0	B1 →	B0
75:	P0	B0	R0 →	R0
76:	P0	B0	P0 →	P0
77:	P0	B0	P1 →	P0
78:	P0	B0	A100 →	P1
79:	P0	B0	A011 →	B0
80:	P1	B0	Q →	B0
81:	P1	B0	B1 →	B0
82:	P1	B0	R0 →	R0
83:	P1	B0	P0 →	P0
84:	P1	B0	A100 →	P1
85:	A001	B0	P0 →	B0
86:	A011	B0	Q →	P1
87:	A110	B0	Q →	P1
88:	A110	B0	P0 →	P1
89:	A110	B0	P1 →	P1
B1 State				
90:	Q	B1	Q →	B1
91:	Q	B1	B0 →	B1
92:	Q	B1	R0 →	Q
93:	Q	B1	R1 →	B1
94:	Q	B1	A000 →	P0
95:	Q	B1	A101 →	P0
96:	B0	B1	Q →	B1
97:	B0	B1	R0 →	Q
98:	B0	B1	A000 →	P0
99:	B0	B1	A101 →	P0
100:	R0	B1	Q →	B1
101:	R0	B1	A000 →	P0
102:	R1	B1	Q →	Q
103:	R1	B1	B0 →	Q
104:	A010	B1	Q →	P0
105:	A010	B1	B0 →	P0
106:	A010	B1	R1 →	P0
107:	A111	B1	Q →	P0
108:	A111	B1	B0 →	P0
R0 State				
109:	Q	R0	Q →	Q
110:	Q	R0	B1 →	Q
111:	Q	R0	A111 →	Q
112:	B0	R0	Q →	B1
113:	B1	R0	Q →	B0
114:	P0	R0	B1 →	B0
115:	P1	R0	B1 →	B0
116:	P1	R0	A111 →	B0
R1 State				
117:	Q	R1	Q →	Q
118:	Q	R1	B0 →	B1
119:	Q	R1	B1 →	B0
120:	B1	R1	Q →	Q
121:	B1	R1	P0 →	B0
122:	B1	R1	P1 →	B0
123:	A101	R1	Q →	Q
124:	A101	R1	P1 →	B0
P0 State				
125:	Q	P0	Q →	P0
126:	Q	P0	P0 →	P0
127:	Q	P0	* →	P0
128:	B0	P0	B0 →	P0
129:	B0	P0	P0 →	P0
130:	B0	P0	* →	P0
131:	R1	P0	R0 →	P0
132:	R1	P0	P0 →	P0
133:	R1	P0	* →	P0
134:	P0	P0	Q →	P0
135:	P0	P0	B0 →	P0
136:	P0	P0	R0 →	P0
137:	P0	P0	P0 →	T
138:	P0	P0	P1 →	T
139:	P0	P0	A010 →	P0
140:	P0	P0	* →	T
141:	P1	P0	P0 →	T
142:	P1	P0	P1 →	T
143:	P1	P0	* →	T
144:	A000	P0	P0 →	P0
145:	A000	P0	A010 →	P0
146:	A000	P0	* →	P0
147:	*	P0	Q →	P0
148:	*	P0	B0 →	P0
149:	*	P0	R0 →	P0
150:	*	P0	P0 →	T
151:	*	P0	P1 →	T
152:	*	P0	A010 →	P0
P1 State				
153:	Q	P1	Q →	P1
154:	Q	P1	P1 →	P1
155:	Q	P1	* →	P1
156:	B0	P1	B0 →	P1
157:	B0	P1	P1 →	P1
158:	B0	P1	* →	P1
159:	R1	P1	R0 →	P1
160:	R1	P1	P1 →	P1
161:	R1	P1	* →	P1
162:	P0	P1	P0 →	T
163:	P0	P1	P1 →	T
164:	P0	P1	* →	T
165:	P1	P1	Q →	P1
166:	P1	P1	B0 →	P1
167:	P1	P1	R0 →	P1
168:	P1	P1	P0 →	T
169:	P1	P1	P1 →	T
170:	P1	P1	A110 →	P1
171:	P1	P1	* →	T
172:	A100	P1	P1 →	P1
173:	A100	P1	A110 →	P1
174:	A100	P1	* →	P1
A000 State				
175:	Q	A000	Q →	Q
176:	Q	A000	P0 →	B0
177:	B1	A000	Q →	Q
178:	B1	A000	P0 →	B0
A001 State				
179:	Q	A001	Q →	Q
180:	Q	A001	B0 →	Q
181:	B0	A001	Q →	Q
182:	B0	A001	B0 →	Q
A100 State				
183:	Q	A100	Q →	R1
184:	Q	A100	P1 →	R1
185:	B0	A100	Q →	P1
186:	B0	A100	P1 →	P1
A101 State				
187:	Q	A101	R1 →	Q
188:	B1	A101	R1 →	P0
A010 State				
189:	Q	A010	Q →	Q
190:	Q	A010	B1 →	Q
191:	P0	A010	Q →	B0
192:	P0	A010	B1 →	B0
A011 State				
193:	Q	A011	Q →	Q
194:	Q	A011	B0 →	Q
195:	B0	A011	Q →	Q
196:	B0	A011	B0 →	Q
A110 State				
197:	Q	A110	Q →	R0
198:	Q	A110	B0 →	P1
199:	P1	A110	Q →	R0
200:	P1	A110	B0 →	P1
A111 State				
201:	R0	A111	Q →	Q
202:	R0	A111	B1 →	P0

Firing Squad Synchronization Problem in Cellular Automata, Figure 3
USN transition table consisting of 202 rules that realize Waksman's synchronization algorithm. The symbol ∗ represents the boundary state

the structure of which is completely identical to that of Waksman [74]. A computer examination made by Umeo, Hisaoka and Sogabe [54] revealed no errors, however, 17 rules were found to be redundant. Figure 5 gives a list of transition rules for Balzer's algorithm and snapshots for synchronization operations on 28 cells. Those redundant rules are indicated by shaded squares. In the transition table, the symbols "M", "L", "F" and "X" represent the general, quiescent, firing and boundary states, respectively. Noguchi [34] also constructed an eight-state, 119-rule optimum-time synchronization algorithm.

Gerken's Seven-State Algorithm Gerken [9] constructed a seven-state, 118-rule synchronization algorithm. In the computer examination, no errors were found, however, 13 rules were found to be redundant. Figure 6 gives a list of the transition rules for Gerken's algorithm and snapshots for synchronization operations on 28 cells. The

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
0	P0	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q
1	P0	A01	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q
2	P0	B0	A01	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q
3	P0	B0	Q	A01	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q
4	P0	B0	R0	Q	A01	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q
5	P0	R0	B1	Q	Q	A01	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q
6	P0	B0	B1	Q	R0	Q	A01	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q
7	P0	B0	B1	R0	Q	Q	Q	A01	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q
8	P0	B0	Q	B0	Q	Q	R0	Q	A01	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q
9	P0	B0	Q	B0	Q	R0	Q	Q	Q	A01	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q
10	P0	B0	Q	B0	R0	Q	Q	Q	R0	Q	A01	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q
11	P0	B0	Q	R0	B1	Q	Q	R0	Q	Q	Q	A01	Q	Q	Q	Q	Q	Q	Q	Q	Q
12	P0	B0	R0	Q	B1	Q	R0	Q	Q	Q	R0	Q	A01	Q	Q	Q	Q	Q	Q	Q	Q
13	P0	R0	B1	Q	B1	R0	Q	Q	Q	R0	Q	Q	Q	A01	Q	Q	Q	Q	Q	Q	Q
14	P0	B0	B1	Q	Q	B0	Q	Q	R0	Q	Q	Q	R0	Q	A01	Q	Q	Q	Q	Q	Q
15	P0	B0	B1	Q	Q	B0	Q	R0	Q	Q	Q	R0	Q	Q	Q	A01	Q	Q	Q	Q	Q
16	P0	B0	B1	Q	Q	B0	R0	Q	Q	Q	R0	Q	Q	Q	R0	Q	A01	Q	Q	Q	Q
17	P0	B0	B1	Q	Q	R0	B1	Q	Q	R0	Q	Q	Q	R0	Q	Q	Q	A01	Q	Q	Q
18	P0	B0	B1	Q	R0	Q	B1	Q	R0	Q	Q	Q	R0	Q	Q	Q	R0	Q	A01	Q	Q
19	P0	B0	B1	R0	Q	Q	B1	R0	Q	Q	Q	R0	Q	Q	Q	R0	Q	Q	Q	A01	Q
20	P0	B0	Q	B0	Q	Q	Q	B0	Q	Q	R0	Q	Q	Q	R0	Q	Q	Q	R0	Q	P0
21	P0	B0	Q	B0	Q	Q	Q	B0	Q	R0	Q	Q	Q	R0	Q	Q	Q	R0	Q	A00	P0
22	P0	B0	Q	B0	Q	Q	Q	B0	R0	Q	Q	Q	R0	Q	Q	Q	R0	Q	A00	B0	P0
23	P0	B0	Q	B0	Q	Q	Q	R0	B1	Q	Q	R0	Q	Q	Q	R0	Q	A00	Q	B0	P0
24	P0	B0	Q	B0	Q	Q	R0	Q	B1	Q	R0	Q	Q	Q	R0	Q	A00	Q	R1	B0	P0
25	P0	B0	Q	B0	Q	R0	Q	Q	B1	R0	Q	Q	Q	R0	Q	A00	Q	Q	B1	R1	P0
26	P0	B0	Q	B0	R0	Q	Q	Q	Q	B0	Q	Q	R0	Q	A00	Q	R1	Q	B1	B0	P0
27	P0	B0	Q	R0	B1	Q	Q	Q	Q	B0	Q	R0	Q	A00	Q	Q	Q	R1	B1	B0	P0
28	P0	B0	R0	Q	B1	Q	Q	Q	Q	B0	R0	Q	A00	Q	R1	Q	Q	B0	Q	B0	P0
29	P0	R0	B1	Q	B1	Q	Q	Q	Q	R0	B1	A00	Q	Q	Q	R1	Q	B0	Q	B0	P0
30	P0	B0	B1	Q	B1	Q	Q	Q	R0	Q	P0	Q	R1	Q	Q	Q	R1	B0	Q	B0	P0
31	P0	B0	B1	Q	B1	Q	Q	R0	Q	A00	P0	A01	Q	R1	Q	Q	B1	R1	Q	B0	P0
32	P0	B0	B1	Q	B1	Q	R0	Q	A00	B0	P0	B0	A01	Q	R1	Q	B1	Q	R1	B0	P0
33	P0	B0	B1	Q	B1	R0	Q	A00	Q	B0	P0	B0	Q	A01	Q	R1	B1	Q	B1	R1	P0
34	P0	B0	B1	Q	Q	B0	A00	Q	R1	B0	P0	B0	R0	Q	A01	B0	Q	Q	B1	B0	P0
35	P0	B0	B1	Q	Q	P1	Q	Q	B1	R1	P0	R0	B1	Q	Q	P1	Q	Q	B1	B0	P0
36	P0	B0	B1	Q	A10	P1	A11	Q	B1	B0	P0	B0	B1	Q	A10	P1	A11	Q	B1	B0	P0
37	P0	B0	B1	A10	R1	P1	R0	A11	B1	B0	P0	B0	B1	A10	R1	P1	R0	A11	B1	B0	P0
38	P0	B0	P0	P0	B0	P1	B0	P0	P0	B0	P0	B0	P0	P0	B0	P1	B0	P0	P0	B0	P0
39	P0	P0	P0	P0	P0	P1	P0	P0	P0	P0	P0	P0	P0	P0	P0	P1	P0	P0	P0	P0	P0
40	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T

Firing Squad Synchronization Problem in Cellular Automata, Figure 4

Snapshots of the Waksman's 16-state optimum-time synchronization algorithm on 21 cells

13 redundant rules are marked by shaded squares in the table. The symbols ">", "/", "…" and "#" represent the general, quiescent, firing and boundary states, respectively. The symbol "…" is replaced by "F" in the configuration (right) at time $t = 54$.

Mazoyer's Six-State Algorithm Mazoyer [25] proposed a six-state, 120-rule synchronization algorithm, the structure of which differs greatly from the previous three algorithms discussed above. The computer examination revealed no errors and only one redundant rule. Figure 7 presents a list of transition rules for Mazoyer's algorithm and snapshots of configurations on 28 cells. In the transition table, the letters "G", "L", "F" and "X" represent the general, quiescent, firing and boundary states, respectively.

Goto's Algorithm The first synchronization algorithm presented by Goto [12] was not published as a journal paper. According to Prof. Goto, the original note Goto [12] is now unavailable, and the only existing material that treats the algorithm is Goto [13]. The Goto's study presents one figure (Fig. 3.8 in Goto [13]) demonstrating how the algorithm works on 13 cells with a very short description in Japanese. Umeo [50] reconstructed the Goto's algorithm based on this figure. Mazoyer [27] also reconstructed this algorithm again. The algorithm that Umeo [50] reconstructed is a non-recursive algorithm consisting of a marking phase and a $3n$-step synchronization phase. In the first phase, by printing a special marker in the cellular space, the entire cellular space is divided into many smaller subspaces, the lengths of which increase exponentially with a common ratio of two, that is 2^j, for any integer j such that $1 \leq j \leq \lfloor \log_2 n \rfloor - 1$. The marking is made from both the left and right ends. In the second phase, each subspace is synchronized using a well-known conventional $3n$-step simple synchronization algorithm. A time-space diagram of the reconstructed algorithm is shown in Fig. 8.

Gerken's 155-State Algorithm Gerken [9] constructed two kinds of optimum-time synchronization algorithms. One seven-state algorithm has been discussed in the previous subsection, and the other is a 155-state algorithm having $\Theta(n \log n)$ state-change complexity. The transition table given in Gerken [9] is described in terms of two-layer construction with 32 states and 347 rules. An expansion of the transition table into a single-layer format yields a 155-state table consisting of 2371 rules. Figure 9 shows a configuration on 28 cells.

State Change Complexity

Vollmar [73] introduced a state-change complexity in order to measure the efficiency of cellular algorithms and showed that $\Omega(n \log n)$ state-changes are required for the synchronization of n cells in $(2n - 2)$ steps.

Theorem 5 ([73]) *$\Omega(n \log n)$ state-change is necessary for synchronizing n cells in $(2n - 2)$ steps.*

Theorem 6 ([9,54]) *Each optimum-time synchronization algorithm developed by Balzer [1], Gerken [9], Ma-*

A	Right State							
Left State	A	B	C	L	M	Q	R	X
A	A			A		A		
B	C			C	R	C		
C	C			C	Q	C		
L	L					L		
M	B			B		B		
Q	A			A	Q	A		
R	L					L		
X								

B	Right State							
Left State	A	B	C	L	M	Q	R	X
A								
B	Q	B	B	R	A	Q	R	
C	Q	B	B	R	A	Q	R	
L								
M	L		C			C		
Q								
R	Q	B	B		A	Q		
X								

C	Right State							
Left State	A	B	C	L	M	Q	R	X
A								
B		C	R	R		M	C	
C	M	C	R	R		M	C	
L	L					L		
M	L		C	C	M	M	C	
Q								
R		C	B	B		M	C	
X								

L	Right State							
Left State	A	B	C	L	M	Q	R	X
A				A		Q		
B			R		Q			
C	C		R	C	Q	C		M
L	L			L		L		L
M	C		C	C	M	C		M
Q	L			L		L		
R			B	A	A	Q		
X								

Q	Right State							
Left State	A	B	C	L	M	Q	R	X
A	Q			Q		Q		
B					R	M	R	
C	M			M	M	M	R	
L	Q			Q	Q	Q		
M					M			
Q	L			A	Q	L		
R	L			A	Q	L		
X								

M	Right State							
Left State	A	B	C	L	M	Q	R	X
A		M	M		M			M
B					M		M	M
C					M	M	M	M
L	M			M		M		
M	M	M	M	M	F	M		F
Q		M	M		M			M
R				M	M		M	M
X	M	M	M	M	F	M		F

R	Right State							
Left State	A	B	C	L	M	Q	R	X
A								
B			R		Q	Q	R	
C		C	R	C	Q	Q	R	
L								
M	C			B	M	C		
Q	L			A	Q	L		
R		B	R	B	Q	Q	R	
X					M	M	R	

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
0	M	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
1	M	C	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
2	M	C	C	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
3	M	C	R	C	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
4	M	C	R	B	C	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
5	M	C	C	B	R	C	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
6	M	C	C	R	R	B	C	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
7	M	C	C	R	B	B	R	C	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
8	M	C	C	C	B	R	R	B	C	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
9	M	C	R	C	R	R	B	B	R	C	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
10	M	C	R	C	R	B	B	R	R	B	C	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
11	M	C	R	C	C	B	R	R	B	B	R	C	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
12	M	C	R	B	C	R	R	B	B	R	R	B	C	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
13	M	C	C	B	C	R	B	B	R	R	B	B	R	C	L	L	L	L	L	L	L	L	L	L	L	L	L	L
14	M	C	C	B	C	C	B	R	R	B	B	R	R	B	C	L	L	L	L	L	L	L	L	L	L	L	L	L
15	M	C	C	B	R	C	R	R	B	B	R	R	B	B	R	C	L	L	L	L	L	L	L	L	L	L	L	L
16	M	C	C	R	R	C	R	B	B	R	R	B	B	R	B	C	L	L	L	L	L	L	L	L	L	L	L	L
17	M	C	C	R	R	C	C	B	R	R	B	B	R	R	B	B	R	C	L	L	L	L	L	L	L	L	L	L
18	M	C	C	R	R	B	C	R	R	B	B	R	R	B	B	R	R	B	C	L	L	L	L	L	L	L	L	L
19	M	C	C	R	B	B	C	R	B	B	R	R	B	B	R	R	B	B	R	C	L	L	L	L	L	L	L	L
20	M	C	C	C	B	B	C	C	B	R	R	B	B	R	R	B	B	R	R	B	C	L	L	L	L	L	L	L
21	M	C	R	C	B	B	R	C	R	R	B	B	R	R	B	B	R	R	B	B	R	C	L	L	L	L	L	L
22	M	C	R	C	B	R	R	C	R	B	B	R	R	B	B	R	R	B	B	R	R	B	C	L	L	L	L	L
23	M	C	R	C	R	R	R	C	C	B	R	R	B	B	R	R	B	B	R	R	B	B	R	C	L	L	L	L
24	M	C	R	C	R	R	R	B	C	R	R	B	B	R	R	B	B	R	R	B	B	R	R	B	C	L	L	L
25	M	C	R	C	R	R	B	B	C	R	B	B	R	R	B	B	R	R	B	B	R	R	B	B	R	C	L	L
26	M	C	R	C	R	B	B	B	C	C	B	R	R	B	B	R	R	B	B	R	R	B	B	R	R	B	C	L
27	M	C	R	C	C	B	B	B	R	C	R	R	B	B	R	R	B	B	R	R	B	B	R	R	B	B	R	M
28	M	C	R	B	C	B	B	R	R	C	R	B	B	R	R	B	B	R	R	B	B	R	R	B	B	R	Q	M
29	M	C	C	B	C	B	R	R	R	C	C	B	R	R	B	B	R	R	B	B	R	R	B	B	R	Q	Q	M
30	M	C	C	B	C	R	R	R	R	B	C	R	R	B	B	R	R	B	B	R	R	B	B	R	Q	L	Q	M
31	M	C	C	B	C	R	R	R	B	B	C	R	B	B	R	R	B	B	R	R	B	B	R	Q	A	L	Q	M
32	M	C	C	B	C	R	R	B	B	B	C	C	B	R	R	B	B	R	R	B	B	R	Q	L	A	Q	Q	M
33	M	C	C	B	C	R	B	B	B	B	R	C	R	R	B	B	R	R	B	B	R	Q	A	L	L	Q	Q	M
34	M	C	C	B	C	C	B	B	B	R	R	C	R	B	B	R	R	B	B	R	Q	L	A	A	L	Q	Q	M
35	M	C	C	B	R	C	B	B	R	R	R	C	C	B	R	R	B	B	R	Q	A	L	L	A	Q	Q	Q	M
36	M	C	C	R	R	C	B	R	R	R	R	B	C	R	R	B	B	R	Q	L	A	A	L	L	Q	L	Q	M
37	M	C	C	R	R	C	R	R	R	R	B	B	C	R	B	B	R	Q	A	L	L	A	A	L	Q	L	Q	M
38	M	C	C	R	R	C	R	R	R	B	B	B	C	C	B	R	Q	L	A	A	L	L	A	Q	Q	L	Q	M
39	M	C	C	R	R	C	R	R	B	B	B	B	R	C	R	Q	A	L	L	A	A	L	L	Q	A	L	Q	M
40	M	C	C	R	R	C	R	B	B	B	B	R	R	C	Q	L	A	A	L	L	A	A	L	Q	A	Q	Q	M
41	M	C	C	R	R	C	C	B	B	B	R	R	R	M	M	L	L	A	A	L	L	A	Q	Q	A	Q	Q	M
42	M	C	C	R	R	B	C	B	B	R	R	R	Q	M	M	C	L	L	A	A	L	L	Q	L	A	Q	Q	M
43	M	C	C	R	B	B	C	B	R	R	R	Q	Q	M	M	C	C	L	L	A	A	L	Q	L	L	Q	Q	M
44	M	C	C	C	B	B	C	R	R	R	Q	L	Q	M	M	C	R	C	L	L	A	Q	Q	L	L	Q	Q	M
45	M	C	R	C	B	B	C	R	R	Q	A	L	Q	M	M	C	R	B	C	L	L	Q	A	L	L	Q	Q	M
46	M	C	R	C	B	B	C	R	Q	L	A	Q	Q	M	M	C	C	B	R	C	L	Q	A	A	L	Q	Q	M
47	M	C	R	C	B	B	C	Q	A	L	L	Q	Q	M	M	C	C	R	R	B	C	Q	A	A	Q	Q	Q	M
48	M	C	R	C	B	B	M	M	A	A	L	Q	Q	M	M	C	C	R	B	B	M	M	A	A	Q	L	Q	M
49	M	C	R	C	B	A	M	M	B	A	Q	Q	Q	M	M	C	C	C	B	A	M	M	B	A	Q	L	Q	M
50	M	C	R	C	Q	R	M	M	L	C	Q	L	Q	M	M	C	R	C	Q	R	M	M	L	C	Q	L	Q	M
51	M	C	R	M	R	Q	M	M	C	L	M	L	Q	M	M	C	R	M	R	Q	M	M	C	L	M	L	Q	M
52	M	C	Q	M	C	Q	M	M	C	Q	M	C	Q	M	M	C	Q	M	C	Q	M	M	C	Q	M	C	Q	M
53	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M
54	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F

Firing Squad Synchronization Problem in Cellular Automata, Figure 5
Transition table for the Balzer's eight-state protocol (*left*) and its snapshots for synchronization operations on 28 cells (*right*)

zoyer [25] *and Waksman* [74] *has an* $O(n^2)$ *state-change complexity, respectively.*

Theorem 7 ([9]) *Gerken's 155-state synchronization algorithm has a* $\Theta(n \log n)$ *state-change complexity.*

It has been shown that any $3n$-step *thread-like* synchronization algorithm has a $\Theta(n \log n)$ state-change complexity and such a $3n$-step thread-like synchronization algorithm can be used for subspace synchronization in the Goto's time-optimum synchronization algorithm. Umeo, Hisaoka and Sogabe [54] has shown that:

Theorem 8 ([54]) *Goto's time-optimum synchronization algorithm as reconstructed by Umeo* [50] *has* $\Theta(n \log n)$ *state-change complexity.*

Figure 10 shows a comparison of the state-change complexities in several optimum-time synchronization algorithms.

/	Right State						
Left State	/	>	]	\	<	[	#
/	/			/	/	/	/
>	]			]		]	<
]	>			>	>		[
\	\	\		\	[	<	
<	/			/	/		
[	/					/	
#							

]	Right State						
Left State	/	>	]	\	<	[	#
/							
>	]	]		]		<	
]							
\	/	/		]		<	
<							
[	]	]		]		<	
#	]	]		]		<	

<	Right State						
Left State	/	>	]	\	<	[	#
/	<			<		<	<
>	]	···		]		]	···
]							
\	\	<		/		/	<
<							
[	/	<		/		/	<
#		···					

\	Right State						
Left State	/	>	]	\	<	[	#
/	/	\		\	[	<	
>	]	\		\		<	
]	>		\	\	[		
\	/	\	\	\	[	<	
<	/						
[							
#				\	>		]

>	Right State						
Left State	/	>	]	\	<	[	#
/	\			\	>	[	
>							
]	\			\	>	[	
\	\			\	>	[	
<	>			>		···	···
[							
#	>			>	>	···	

[	Right State						
Left State	/	>	]	\	<	[	#
/	[			[		[	[
>							
]	>			>		>	>
\	\			[		[	[
<	\			[		[	[
[							
#							

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
0	>	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
1	>	]	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
2	>	]	>	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
3	>	]	^	]	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
4	>	]	^	/	>	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
5	>	]	>	^	^	]	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
6	>	]	>	^	^	/	>	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
7	>	]	>	^	/	^	^	]	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
8	>	]	>	]	^	^	^	/	>	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
9	>	]	^	]	^	^	/	^	^	]	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
10	>	]	^	]	^	/	^	^	^	/	>	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
11	>	]	^	]	>	^	^	^	/	^	^	]	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
12	>	]	^	/	>	^	^	/	^	^	^	/	>	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/
13	>	]	>	^	>	^	/	^	^	^	/	^	^	]	/	/	/	/	/	/	/	/	/	/	/	/	/	/
14	>	]	>	^	>	]	^	^	^	/	^	^	^	/	>	/	/	/	/	/	/	/	/	/	/	/	/	/
15	>	]	>	^	^	]	^	^	/	^	^	^	/	^	^	]	/	/	/	/	/	/	/	/	/	/	/	/
16	>	]	>	^	^	]	^	/	^	^	^	/	^	^	^	/	>	/	/	/	/	/	/	/	/	/	/	/
17	>	]	>	^	^	]	>	^	^	^	/	^	^	^	/	^	^	]	/	/	/	/	/	/	/	/	/	/
18	>	]	>	^	^	/	>	^	^	/	^	^	^	/	^	^	^	/	>	/	/	/	/	/	/	/	/	/
19	>	]	>	^	/	^	>	^	/	^	^	^	/	^	^	^	/	^	^	]	/	/	/	/	/	/	/	/
20	>	]	>	]	^	^	>	]	^	^	^	/	^	^	^	/	^	^	^	/	>	/	/	/	/	/	/	/
21	>	]	^	]	^	^	^	]	^	^	/	^	^	^	/	^	^	^	/	^	^	]	/	/	/	/	/	/
22	>	]	^	]	^	^	^	]	^	/	^	^	^	/	^	^	^	/	^	^	^	/	>	/	/	/	/	/
23	>	]	^	]	^	^	^	]	>	^	^	^	/	^	^	^	/	^	^	^	/	^	^	]	/	/	/	/
24	>	]	^	]	^	^	^	/	>	^	^	/	^	^	^	/	^	^	^	/	^	^	^	/	>	/	/	/
25	>	]	^	]	^	^	/	^	>	^	/	^	^	^	/	^	^	^	/	^	^	^	/	^	^	]	/	/
26	>	]	^	]	^	/	^	^	>	]	^	^	^	/	^	^	^	/	^	^	^	/	^	^	^	/	>	/
27	>	]	^	]	>	^	^	^	^	]	^	^	/	^	^	^	/	^	^	^	/	^	^	^	/	^	^	<
28	>	]	^	/	>	^	^	^	^	]	^	/	^	^	^	/	^	^	^	/	^	^	^	/	^	^	[	<
29	>	]	>	^	>	^	^	^	^	]	>	^	^	^	/	^	^	^	/	^	^	^	/	^	^	<	[	<
30	>	]	>	^	>	^	^	^	^	/	>	^	^	/	^	^	^	/	^	^	^	/	^	^	[	/	[	<
31	>	]	>	^	>	^	^	^	/	^	>	^	/	^	^	^	/	^	^	^	/	^	^	<	^	/	[	<
32	>	]	>	^	>	^	^	/	^	^	>	]	^	^	^	/	^	^	^	/	^	^	[	/	/	<	[	<
33	>	]	>	^	>	^	/	^	^	^	^	]	^	^	/	^	^	^	/	^	^	<	^	/	/	<	[	<
34	>	]	>	^	>	]	^	^	^	^	^	]	^	/	^	^	^	/	^	^	[	/	/	/	^	/	<	[
35	>	]	>	^	^	]	^	^	^	^	^	]	>	^	^	^	/	^	^	<	^	/	/	/	[	<	[	<
36	>	]	>	^	^	]	^	^	^	^	^	/	>	^	^	/	^	^	[	/	/	^	/	/	[	/	[	<
37	>	]	>	^	^	]	^	^	^	^	/	^	>	^	/	^	^	<	^	/	/	/	^	/	[	/	[	<
38	>	]	>	^	^	]	^	^	^	/	^	^	>	]	^	^	[	/	/	^	/	/	/	<	[	/	[	<
39	>	]	>	^	^	]	^	^	/	^	^	^	^	]	^	<	^	/	/	/	^	/	/	<	^	/	[	<
40	>	]	>	^	^	]	^	/	^	^	^	^	^	]	[	/	/	/	^	/	/	/	^	/	<	/	<	[	<
41	>	]	>	^	^	]	>	^	^	^	^	^	^	<	>	/	/	/	/	^	/	/	/	[	<	/	<	[	<
42	>	]	>	^	^	/	>	^	^	^	^	^	[	<	>	]	/	/	/	/	^	/	/	/	[	/	/	<	[	<
43	>	]	>	^	/	^	>	^	^	^	^	<	[	<	>	]	>	/	/	/	^	/	[	/	/	<	[	<
44	>	]	>	]	^	^	>	^	^	^	[	/	[	<	>	]	^	]	/	/	/	<	[	/	/	<	[	<
45	>	]	^	]	^	^	>	^	^	<	^	/	[	<	>	]	^	/	>	/	/	<	^	/	/	<	[	<
46	>	]	^	]	^	^	>	^	[	/	/	<	[	<	>	]	>	^	^	]	/	<	/	^	/	<	[	<
47	>	]	^	]	^	^	>	<	^	/	/	<	[	<	>	]	>	^	^	/	>	<	/	/	[	<	[	<
48	>	]	^	]	^	^	[	]	/	^	/	<	[	<	>	]	>	^	/	^	[	]	/	/	[	/	[	<
49	>	]	^	]	^	<	[	]	>	/	[	<	[	<	>	]	>	]	^	<	[	]	>	/	[	/	[	<
50	>	]	^	]	[	/	[	]	^	]	[	/	[	<	>	]	^	]	[	/	[	]	^	]	[	/	[	<
51	>	]	^	<	>	/	[	]	^	<	>	/	[	<	>	]	^	<	>	/	[	]	^	<	>	/	[	<
52	>	]	[	<	>	]	[	]	[	<	>	]	[	<	>	]	[	<	>	]	[	]	[	<	>	]	[	<
53	>	<	>	<	>	<	>	<	>	<	>	<	>	<	>	<	>	<	>	<	>	<	>	<	>	<	>	<
54	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F

Firing Squad Synchronization Problem in Cellular Automata, Figure 6
Transition table for the Gerken's seven-state protocol (*left*) and snapshots for synchronization operations on 28 cells (*right*)

A Comparison of Quantitative Aspects of Optimum-Time Synchronization Algorithms

Section "A Comparison of Quantitative Aspects of Optimum-Time Synchronization Algorithms" presents a table based on a quantitative comparison of optimum-time synchronization algorithms and their transition tables discussed above with respect to the number of internal states of each finite state automaton, the number of transition rules realizing the synchronization, and the number of state-changes on the array.

One-Sided vs. Two-Sided Recursive Algorithms

Firing squad synchronization algorithms have been designed on the basis of parallel divide-and-conquer strategy that calls itself recursively in parallel. Those recursive calls are implemented by generating many *Generals* that

A	Right State					
Left State	A	B	C	G	L	X
A	A	B	C	B	A	F
B		G	C	C	G	C
C	A				A	
G			C	C		C
L	A	L	G			
X	F		G			

B	Right State					
Left State	A	B	C	G	L	X
A	B	B	L		G	
B	A	B	C	B	G	
C	A			L	L	L
G	C		B	G	C	G
L	G	B	L	B		
X						

C	Right State					
Left State	A	B	C	G	L	X
A		B		B	B	B
B			C	G	C	G
C	A	B	C	B	C	
G		B		B	B	B
L	A	G	C	G	C	
X						

G	Right State					
Left State	A	B	C	G	L	X
A		G	G		B	
B		G	G	G	B	G
C		G	G	A	A	A
G		G	G	F	B	F
L	G	G	G			
X		G	G	F	A	

L	Right State					
Left State	A	B	C	G	L	X
A	L	L	L	C	G	C
B	L	L	L	L	L	L
C	L	L	L	G	A	G
G	L	L	L	A	C	A
L		L	L	L	L	L
X					L	

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
0	G	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
1	A	C	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
2	G	B	A	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
3	G	C	G	G	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
4	G	B	A	B	C	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
5	G	C	G	L	C	A	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
6	G	B	A	L	A	A	G	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
7	G	C	G	L	A	B	B	C	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
8	G	B	A	L	L	B	C	C	A	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
9	G	C	G	G	L	L	C	A	A	G	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
10	G	B	A	B	C	L	A	A	B	B	C	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
11	G	C	G	L	C	L	A	B	B	C	C	A	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
12	G	B	A	L	C	L	L	B	C	C	A	A	G	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
13	G	C	G	L	C	A	L	L	C	A	A	B	B	C	L	L	L	L	L	L	L	L	L	L	L	L	L	L
14	G	B	A	L	A	A	G	L	A	A	B	B	C	C	A	L	L	L	L	L	L	L	L	L	L	L	L	L
15	G	C	G	L	A	B	B	L	A	B	B	C	C	A	A	G	L	L	L	L	L	L	L	L	L	L	L	L
16	G	B	A	L	L	B	G	L	L	B	C	C	A	A	B	B	C	L	L	L	L	L	L	L	L	L	L	L
17	G	C	G	G	L	B	B	C	L	L	C	A	A	B	B	C	C	A	L	L	L	L	L	L	L	L	L	L
18	G	B	A	B	L	B	C	C	A	L	A	A	B	B	C	C	A	A	G	L	L	L	L	L	L	L	L	L
19	G	C	G	G	L	L	C	A	A	L	A	B	B	C	C	A	A	B	B	C	L	L	L	L	L	L	L	L
20	G	B	A	B	C	L	A	A	A	L	L	B	C	C	A	A	B	B	C	C	A	L	L	L	L	L	L	L
21	G	C	G	L	C	L	A	A	A	G	L	L	C	A	A	B	B	C	C	A	A	G	L	L	L	L	L	L
22	G	B	A	L	C	L	A	A	B	B	C	L	A	A	B	B	C	C	A	A	B	B	C	L	L	L	L	L
23	G	C	G	L	C	L	A	B	B	C	C	L	A	B	B	C	C	A	A	B	B	C	C	A	L	L	L	L
24	G	B	A	L	C	L	L	B	C	C	C	L	L	B	C	C	A	A	B	B	C	C	A	A	G	L	L	L
25	G	C	G	L	C	A	L	L	C	C	C	A	L	L	C	A	A	B	B	C	C	A	A	B	B	C	L	L
26	G	B	A	L	A	A	G	L	C	C	A	A	G	L	A	A	B	B	C	C	A	A	B	B	C	C	A	L
27	G	C	G	L	A	B	B	L	C	A	A	B	B	L	A	B	B	C	C	A	A	B	B	C	C	A	A	C
28	G	B	A	L	L	B	G	L	A	A	B	B	G	L	L	B	C	C	A	A	B	B	C	C	A	A	C	B
29	G	C	G	G	L	B	B	L	A	B	B	B	B	C	L	L	C	A	A	B	B	C	C	A	A	C	B	L
30	G	B	A	B	L	B	G	L	L	B	B	B	C	C	A	L	A	A	B	B	C	C	A	A	C	B	L	L
31	G	C	G	G	L	B	B	C	L	B	B	C	C	A	A	L	A	B	B	C	C	A	A	C	B	L	L	L
32	G	B	A	B	L	B	C	C	L	B	C	C	A	A	A	L	L	B	C	C	A	A	C	B	L	L	L	L
33	G	C	G	G	L	L	C	C	L	L	C	A	A	A	A	G	L	L	C	A	A	C	B	L	L	L	L	L
34	G	B	A	B	C	L	C	C	A	L	A	A	A	A	B	B	C	L	A	A	C	B	L	L	L	L	L	L
35	G	C	G	L	C	L	C	A	A	L	A	A	A	B	B	C	C	L	A	C	B	L	L	L	L	L	L	L
36	G	B	A	L	C	L	A	A	A	L	A	A	B	B	C	C	C	L	G	B	L	L	L	L	L	L	L	L
37	G	C	G	L	C	L	A	A	A	L	A	B	B	C	C	C	C	G	G	C	L	L	L	L	L	L	L	L
38	G	B	A	L	C	L	A	A	A	L	L	B	C	C	C	C	B	A	G	B	A	L	L	L	L	L	L	L
39	G	C	G	L	C	L	A	A	A	G	L	L	C	C	C	B	A	C	G	C	G	G	L	L	L	L	L	L
40	G	B	A	L	C	L	A	A	B	B	C	L	C	C	B	A	C	B	G	B	A	B	C	L	L	L	L	L
41	G	C	G	L	C	L	A	B	B	C	C	L	C	B	A	C	B	L	G	C	G	L	C	A	L	L	L	L
42	G	B	A	L	C	L	L	B	C	C	C	L	G	A	C	B	L	L	G	B	A	L	A	A	G	L	L	L
43	G	C	G	L	C	A	L	L	C	C	C	G	G	C	B	L	L	L	G	C	G	L	A	B	B	C	L	L
44	G	B	A	L	A	A	G	L	C	C	B	A	G	B	L	L	L	L	G	B	A	L	L	B	C	C	A	L
45	G	C	G	L	A	B	B	L	C	B	A	C	G	C	L	L	L	L	G	C	G	G	L	L	C	A	A	C
46	G	B	A	L	L	B	G	L	G	A	C	B	G	B	A	L	L	L	G	B	A	B	C	L	A	A	C	B
47	G	C	G	G	L	B	B	A	G	C	B	L	G	C	G	G	L	L	G	C	G	L	C	L	A	C	B	L
48	G	B	A	B	L	B	A	C	G	B	L	L	G	B	A	B	C	L	G	B	A	L	C	L	G	B	L	L
49	G	C	G	G	L	G	C	B	G	C	L	L	G	C	G	L	C	G	G	C	G	L	C	G	G	C	L	L
50	G	B	A	B	A	G	B	L	G	B	A	L	G	B	A	L	G	A	G	B	A	L	G	A	G	B	A	L
51	G	C	G	B	C	G	C	L	G	C	G	C	G	C	G	C	G	C	G	C	G	C	G	C	G	C	G	C
52	G	B	G	B	G	G	B	G	G	B	G	B	G	B	G	B	G	B	G	B	G	B	G	B	G	B	G	B
53	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
54	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F

Firing Squad Synchronization Problem in Cellular Automata, Figure 7
Transition table for the Mazoyer's six-state protocol (*left*) and its snapshots of configurations on 28 cells (*right*)

work for synchronizing divided small areas in the cellular space. Initially a *General* G_0 located at the left end works for synchronizing the whole cellular space consisting of n cells. In Fig. 11 (left), G_1 synchronizes the subspace between G_1 and the right end of the array. The ith *General* G_i, $i = 2, 3, \ldots,$ works for synchronizing the cellular space between G_{i-1} and G_i, respectively. Thus, all of the *Generals* generated by G_0 are located at the left end of the divided cellular spaces to be synchronized. On the other hand, in Fig. 11 (right), the *General* G_0 generates *General* G_i, $i = 1, 2, 3, \ldots,$. Each G_i, $i = 1, 2, 3, \ldots,$ synchronizes the divided space between G_i and G_{i+1}, respectively. In addition, G_i, $i = 2, 3, \ldots,$ does the same operations as G_0. Thus, in Fig. 11 (right) one can find *Generals* located at either end of the subspace for which they are responsible. If all of the recursive calls for the synchronization are issued by *Generals* located at one (both two) end(s) of partitioned cellular spaces for which the *General* works, the synchronization algorithm is said to have *one-sided* (*two-sided*) recursive property, respectively. A synchronization algorithm with the one-sided (two-sided) recursive property is referred to as one-sided (two-sided) recursive synchro-

Firing Squad Synchronization Problem in Cellular Automata, Table 1
Quantitative comparison of transition rule sets for optimum-time firing squad synchronization algorithms. The * symbol shows the correction and reduction of transition rules made in Umeo, Hisaoka and Sogabe [54]. The ** symbol indicates the number of states and rules obtained after the expansion of the original two-layer construction

Algorithm	# of states	# of transition rules	State change complexity
Goto [12]	many thousands	–	$\Theta(n \log n)$
Waksman [74]	16	202* (3216)	$O(n^2)$
Balzer [1]	8	165* (182)	$O(n^2)$
Noguchi [34]	8	119	$O(n^2)$
Gerken [9]	7	105* (118)	$O(n^2)$
Mazoyer [25]	6	119* (120)	$O(n^2)$
Gerken [9]	155**(32)	2371**(347)	$\Theta(n \log n)$

Firing Squad Synchronization Problem in Cellular Automata, Table 2
A qualitative comparison of optimum-time firing squad synchronization algorithms

Algorithm	One-/two-sided	Recursive/non-recursive	# of signals
Goto [12]	–	non-recursive	finite
Waksman [74]	two-sided	recursive	infinite
Balzer [1]	two-sided	recursive	infinite
Noguchi [34]	two-sided	recursive	infinite
Gerken [9]	two-sided	recursive	infinite
Mazoyer [25]	one-sided	recursive	infinite
Gerken [9]	two-sided	recursive	finite

nization algorithm. Figure 11 illustrates a time-space diagram for one-sided (Fig. 11 (left)) and two-sided (Fig. 11 (right)) recursive synchronization algorithms both operating in optimum $2n - 2$ steps. It is noted that optimum-time synchronization algorithms developed by Balzer [1], Gerken [9], Noguchi [34] and Waksman [74] are two-sided ones and an algorithm proposed by Mazoyer [25] is an only synchronization algorithm with the one-sided recursive property.

Observation 3.2 ([54]) *Optimum-time synchronization algorithms developed by Balzer [1], Gerken [9], Noguchi [34] and Waksman [74] are two-sided ones. The algorithm proposed by Mazoyer [25] is a one-sided one.*

A more general design scheme for one-sided recursive optimum-time synchronization algorithms can be found in Mazoyer [24].

Recursive vs. Non-recursive algorithms

As is shown in the previous section, the optimum-time synchronization algorithms developed by Balzer [1], Gerken [9], Mazoyer [25], Noguchi [34] and Waksman [74] are recursive ones. On the other hand, it is noted that overall structure of the reconstructed Goto's algorithm is a non-recursive one, however divided subspaces are synchronized by using a recursive $3n + O(1)$-step synchronization algorithm.

Number of Signals

Waksman [74] devised an efficient way to cause a general cell to generate infinite signals at propagating speeds of $1/1, 1/3, 1/7, .., 1/(2^k - 1)$, where k is any natural number. These signals play an important role in dividing the array into two, four, eight, ..., equal parts synchronously. The same set of signals is used in Balzer [1]. Gerken [9] had a similar idea in the construction of his seven-state algorithm. Thus *infinite* set of signals with different propagation speed is used in the first three algorithms. On the other hand, *finite* sets of signals with propagating speed $\{1/5, 1/2, 1/1\}$ and $\{1/3, 1/2, 3/5, 1/1\}$ are made use of in Gerken's 155-state algorithm and the reconstructed Goto's algorithm, respectively.

A Comparison of Qualitative Aspects of Optimum-Time Synchronization Algorithms

Section "A Comparison of Qualitative Aspects of Optimum-Time Synchronization Algorithms" presents a table based on a qualitative comparison of optimum-time synchronization algorithms with respect to one/two-sided re-

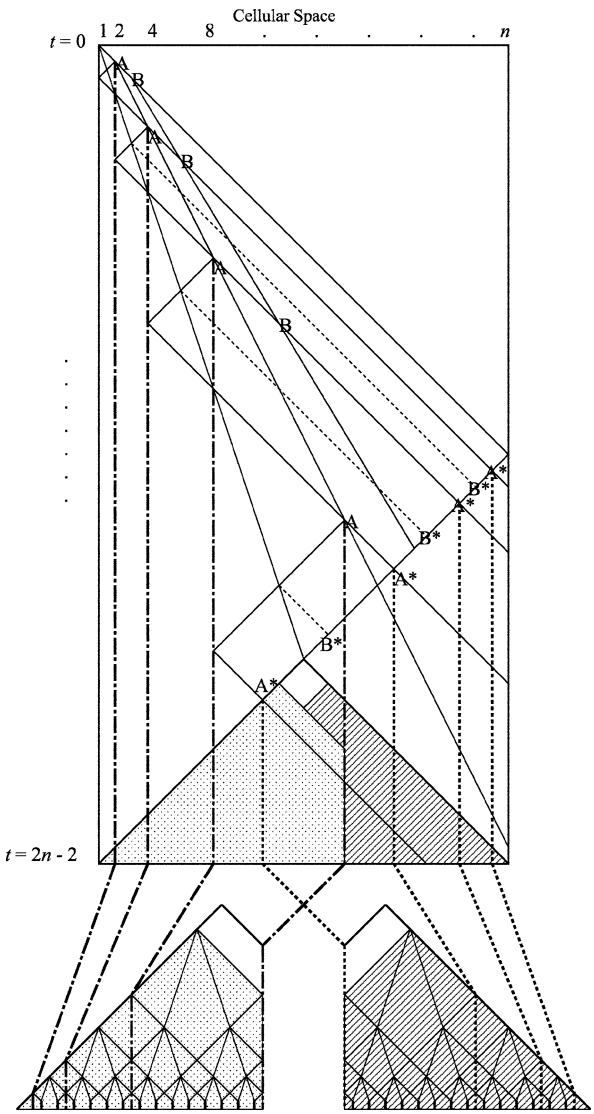

Firing Squad Synchronization Problem in Cellular Automata, Figure 8
Time-space diagram for Goto's algorithm as reconstructed by Umeo [50]

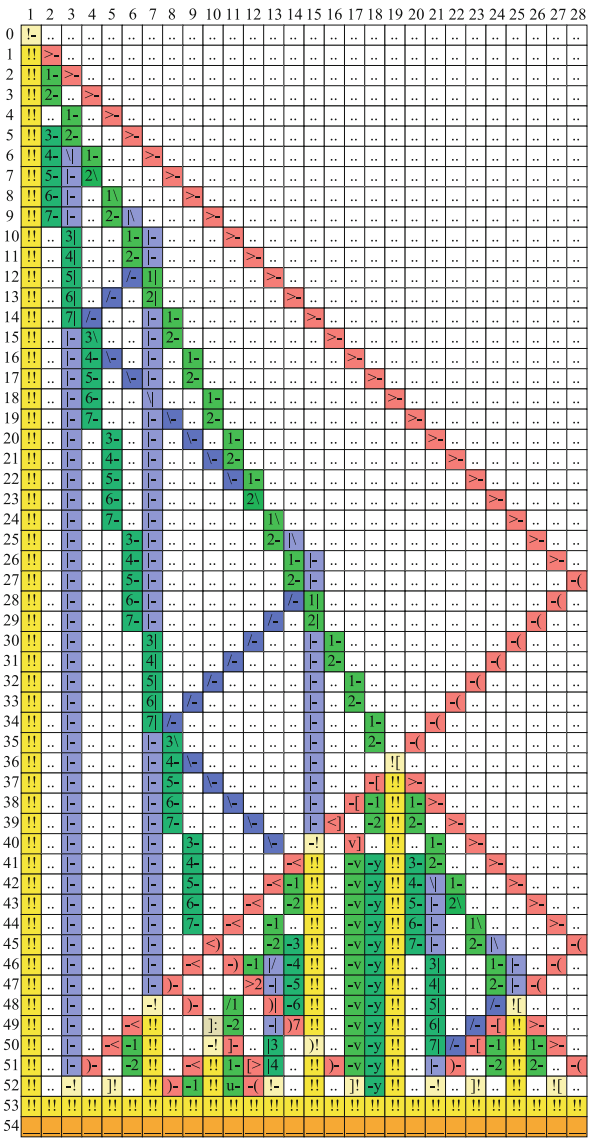

Firing Squad Synchronization Problem in Cellular Automata, Figure 9
Snapshots of the Gerken's 155-state algorithm on 28 cells

cursive properties and the number of signals being used for simultaneous space divisions.

Variants of the Firing Squad Synchronization Problem

Generalized Firing Squad Synchronization Problem

Section "Generalized Firing Squad Synchronization Problem" considers a *generalized* firing squad synchronization problem which allows the general to be located anywhere on the array. It has been shown to be impossible to synchronize any array of length n less than $n - 2 + \max(k, n - k + 1)$ steps, where the general is located on C_k. Moore and Langdon [30], Szwerinski [45] and Varshavsky, Marakhovsky and Peschansky [69] developed a generalized optimum-time synchronization algorithm with 17, 10 and 10 internal states, respectively, that can synchronize any array of length n at exactly $n - 2 + \max(k, n - k + 1)$ steps. Recently, Settle and Simon [43] and Umeo, Hisaoka, Michisaka, Nishioka and Maeda [56]

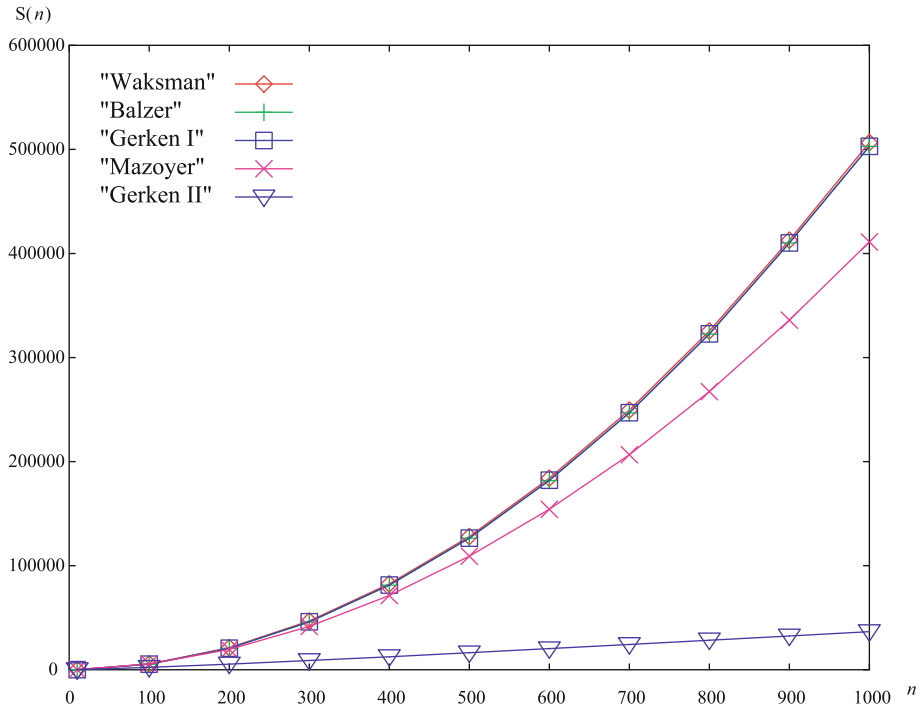

Firing Squad Synchronization Problem in Cellular Automata, Figure 10
A comparison of state-change complexities in optimum-time synchronization algorithms

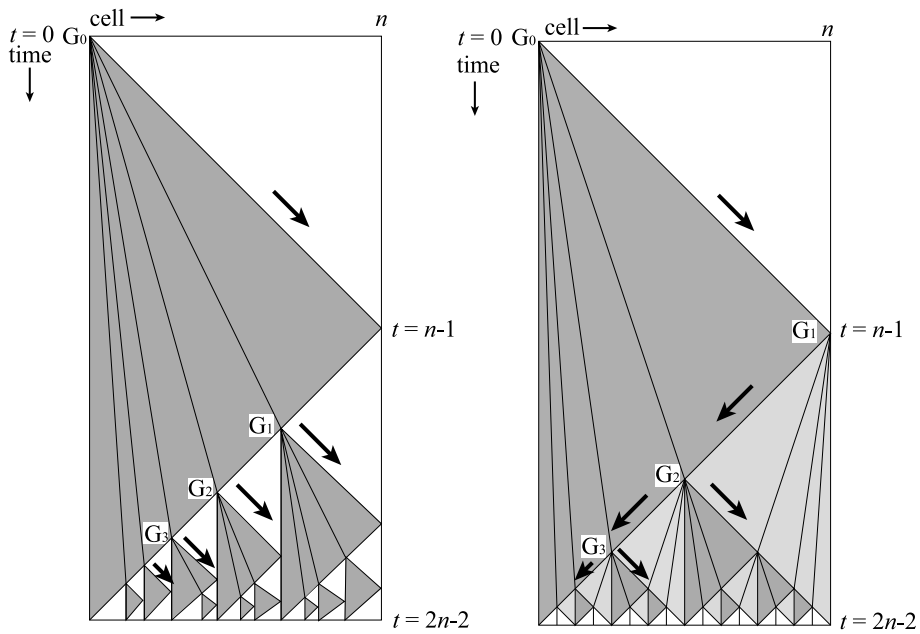

Firing Squad Synchronization Problem in Cellular Automata, Figure 11
One-sided recursive synchronization scheme (*left*) and two-sided recursive synchronization scheme (*right*)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
0	A	A	A	A	A	A	A	B	A	A	A	A	A	A	A	A	A	A	A	A	A	A
1	A	A	A	A	A	A	B	B	B	A	A	A	A	A	A	A	A	A	A	A	A	A
2	A	A	A	A	A	B	D>	A	D<	B	A	A	A	A	A	A	A	A	A	A	A	A
3	A	A	A	A	B	D>	A	A	A	D<	B	A	A	A	A	A	A	A	A	A	A	A
4	A	A	A	B	D>	A	D>	A	D<	A	D<	B	A	A	A	A	A	A	A	A	A	A
5	A	A	B	D>	A	D>	A	A	A	D<	A	D<	B	A	A	A	A	A	A	A	A	A
6	A	B	D>	A	D>	A	D>	A	D<	A	D<	A	D<	B	A	A	A	A	A	A	A	A
7	R	D>	A	D>	A	D>	A	A	A	D<	A	D<	A	D<	B	A	A	A	A	A	A	A
8	R	C>	D>	A	D>	A	D>	A	D<	A	D<	A	D<	A	D<	B	A	A	A	A	A	A
9	R	D<	C>	D>	A	D>	A	A	A	D<	A	D<	A	D<	A	D<	B	A	A	A	A	A
10	R	E2	D>	C>	D>	A	D>	A	D<	A	D<	A	D<	A	D<	A	D<	B	A	A	A	A
11	R	E2	A	D<	C>	D>	A	A	A	D<	A	D<	A	D<	A	D<	A	D<	B	A	A	A
12	R	E2	D<	A	D>	C>	D>	A	D<	A	D<	A	D<	A	D<	A	D<	A	D<	B	A	A
13	R	D<	E1	A	A	D<	C>	A	A	D<	A	D<	A	D<	A	D<	A	D<	A	D<	B	A
14	R	E2	E1	A	D<	A	D>	C>	D<	A	D<	A	D<	A	D<	A	D<	A	D<	A	D<	R
15	R	E2	E1	D<	A	A	A	E2	D>	D<	A	D<	A	D<	A	D<	A	D<	A	D<	C<	R
16	R	E2	A	E2	A	A	A	E2	D<	D<	D<	A	D<	A	D<	A	D<	A	D<	C<	D>	R
17	R	E2	A	E2	A	A	A	D<	E1	C>	D>	D<	A	D<	A	D<	A	D<	C<	D<	E2	R
18	R	E2	A	E2	A	A	D<	A	E1	A	C>	D<	D<	A	D<	A	D<	C<	D>	A	E2	R
19	R	E2	A	E2	A	D<	A	A	E1	D<	A	C>	D>	D<	A	D<	C<	D<	A	D>	E2	R
20	R	E2	A	E2	D<	A	A	A	A	E2	A	A	C>	D<	D<	C<	D>	A	A	E1	D>	R
21	R	E2	A	D<	E1	A	A	A	A	E2	A	D<	A	C>	C<	D<	A	D>	A	E1	E2	R
22	R	E2	D<	A	E1	A	A	A	A	E2	D<	A	A	C<	D>	A	A	A	D>	E1	E2	R
23	R	D<	E1	A	E1	A	A	A	A	D<	E1	A	C<	D<	A	D>	A	A	E2	A	E2	R
24	R	E2	E1	A	E1	A	A	A	D<	A	E1	C<	D>	A	A	A	D>	A	E2	A	E2	R
25	R	E2	E1	A	E1	A	A	D<	A	A	R	R	A	D>	A	A	A	D>	E2	A	E2	R
26	R	E2	E1	A	E1	A	D<	A	A	C<	R	R	C>	A	D>	A	A	E1	D>	A	E2	R
27	R	E2	E1	A	E1	D<	A	A	C<	D>	R	R	D<	C>	A	D>	A	E1	A	D>	E2	R
28	R	E2	E1	A	A	E2	A	C<	D<	E2	R	R	E2	D>	C>	A	D>	E1	A	E1	D>	R
29	R	E2	E1	A	A	E2	C<	D>	A	E2	R	R	E2	A	D<	C>	E2	A	A	E1	E2	R
30	R	E2	E1	A	A	R	D<	A	D>	E2	R	R	E2	D<	A	D>	R	A	A	E1	E2	R
31	R	E2	E1	A	C<	R	C>	A	E1	D>	R	R	D<	E1	A	C<	R	C>	A	E1	E2	R
32	R	E2	E1	C<	D>	R	D<	C>	E1	E2	R	R	E2	E1	C<	D>	R	D<	C>	E1	E2	R
33	R	E2	R	R	E2	R	E2	R	R	E2	R	R	E2	R	R	E2	R	E2	R	R	E2	R
34	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
35	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F

Firing Squad Synchronization Problem in Cellular Automata, Figure 12

Snapshots of the three Russian's 10-state generalized optimum-time synchronization algorithm on 22 cells

State Q

L \ R	Q	P	R	Z	M	*
Q	Q	P	Q	Q		Q
P	P	P				P
R	Q		Q			
Z	Q			Q		Q
M						
*	Q	P	Q	Q		

State Z

L \ R	Q	P	R	Z	M	*
Q			Q	P	Q	
P		Z		Z		
R	Q		Q	Q		Q
Z	P	Z	Q	F	Q	F
M	Q			Q	Q	Q
*		Z	Q	F	Q	

State P

L \ R	Q	P	R	Z	M	*
Q	Z	Z	R	R		
P	Z		Z	Z		
R	R	Z	Z			Z
Z	R	Z		Z		Z
M						
*	Z		Z	Z		

State M

L \ R	Q	P	R	Z	M	*
Q						
P						
R			R	Z		
Z			Z			
M						
*						

State R

L \ R	Q	P	R	Z	M	*
Q			R	P	Z	
P			M	R	M	
R	R	M			M	
Z	P	R			R	
M	Z	M	M	R	M	
*						

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	P	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q
1	Z	P	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q
2	Z	R	P	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q
3	Q	R	R	P	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q
4	Q	R	M	R	P	Q	Q	Q	Q	Q	Q	Q	Q	Q
5	Q	Z	R	M	R	P	Q	Q	Q	Q	Q	Q	Q	Q
6	Q	Q	R	R	M	R	P	Q	Q	Q	Q	Q	Q	Q
7	Q	Q	R	M	R	M	R	P	Q	Q	Q	Q	Q	Q
8	Q	Q	Z	R	M	R	M	R	P	Q	Q	Q	Q	Q
9	Q	Q	Q	R	R	M	R	M	R	P	Q	Q	Q	Q
10	Q	Q	Q	R	M	R	M	R	M	R	P	Q	Q	Q
11	Q	Q	Q	Z	R	M	R	M	R	M	R	P	Q	Q
12	Q	Q	Q	Q	R	R	M	R	M	R	M	R	P	Q
13	Q	Q	Q	Q	R	M	R	M	R	M	R	M	R	P
14	Q	Q	Q	Q	Z	R	M	R	M	R	M	R	M	Z
15	Q	Q	Q	Q	Q	R	R	M	R	M	R	M	Z	Q
16	Q	Q	Q	Q	Q	R	M	R	M	R	M	Z	Q	Q
17	Q	Q	Q	Q	Q	Z	R	M	R	M	Z	Q	Q	Q
18	Q	Q	Q	Q	Q	Q	R	R	M	Z	Q	Q	Q	Q
19	Q	Q	Q	Q	Q	Q	R	M	Z	Q	Q	Q	Q	Q
20	Q	Q	Q	Q	Q	Q	Z	Z	Q	Q	Q	Q	Q	Q
21	Q	Q	Q	Q	Q	Q	P	P	Q	Q	Q	Q	Q	Q
22	Q	Q	Q	Q	Q	P	Z	Z	P	Q	Q	Q	Q	Q
23	Q	Q	Q	Q	P	R	Z	Z	R	P	Q	Q	Q	Q
24	Q	Q	Q	P	R	R	Q	Q	R	R	P	Q	Q	Q
25	Q	Q	P	R	M	R	Q	Q	R	M	R	P	Q	Q
26	Q	P	R	M	R	Z	Q	Q	Z	R	M	R	P	Q
27	P	R	M	R	R	Q	Q	Q	Q	R	R	M	R	P
28	Z	M	R	M	R	Q	Q	Q	Q	R	M	R	M	Z
29	Q	Z	M	R	Z	Q	Q	Q	Q	Z	R	M	Z	Q
30	Q	Q	Z	R	Q	Q	Q	Q	Q	Q	R	Z	Q	Q
31	Q	Q	Q	P	Q	Q	Q	Q	Q	Q	P	Q	Q	Q
32	Q	Q	P	Z	P	Q	Q	Q	Q	P	Z	P	Q	Q
33	Q	P	R	Z	R	P	Q	Q	P	R	Z	R	P	Q
34	P	R	R	Q	R	R	P	P	R	R	Q	R	R	P
35	Z	M	R	Q	R	M	Z	Z	M	R	Q	R	M	Z
36	Q	Z	Z	Q	Z	Z	Q	Q	Z	Z	Q	Z	Z	Q
37	Q	P	P	Q	P	P	Q	Q	P	P	Q	P	P	Q
38	P	Z	Z	P	Z	Z	P	P	Z	Z	P	Z	Z	P
39	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z
40	F	F	F	F	F	F	F	F	F	F	F	F	F	F

Firing Squad Synchronization Problem in Cellular Automata, Figure 13

Transition table for symmetric six-state protocol (*left*) and snapshots for synchronization algorithm on 14 cells

have proposed a 9-state generalized synchronization algorithm operating in optimum-step. Figure 12 shows snapshots for synchronization configurations based on the rule set of Varshavsky, Marakhovsky and Peschansky [69].

Theorem 9 ([30,43,45,56,69]) *There exists a cellular automaton that can synchronize any one-dimensional array of length n in optimum $n - 2 + \max(k, n - k + 1)$ steps, where the general is located on the kth cell from left end.*

Non-Optimum-Time 3*n*-Step Synchronization Algorithms

Non-optimum-time $3n$-step algorithm is a simple and straightforward one that exploits a parallel divide-and-conquer strategy based on an efficient use of 1/1- and 1/3-speed of signals. Minsky and MacCarthy [28] gave an idea for designing the $3n$-step synchronization algorithm, and Fischer [8] implemented the $3n$-step algorithm, yielding a 15-state implementation, respectively. Yunès [76] developed two seven-state synchronization algorithms, thus decreasing the number of internal states of each cellular automaton. This section presents a new symmetric six-state $3n$-step firing squad synchronization algorithm developed in Umeo, Maeda and Hongyo [61]. The number *six* is the smallest one known at present in the class of $3n$-step synchronization algorithms. Figure 13 shows the 6-state transition table and snapshots for synchronization on 14 cells. In the transition table, the symbols "P", "Q", "F" and "$\star$" represent the general, quiescent, firing and boundary states, respectively. Yunès [79] also developed a symmetric 6-state $3n$-step solution.

Theorem 10 ([61,79]) *There exists a symmetric 6-state cellular automaton that can synchronize any n cells in $3n + O(\log n)$ steps.*

State Q

L \ R	Q	P	R	M	Z	*
Q	Q	P	Q	M	Q	Q
P	P	P				P
R	Q		Q			Q
M	M					M
Z	Q				Q	Q
*	Q	P	Q	M	Q	

State Z

L \ R	Q	P	R	M	Z	*
Q	P	Q	Q	Q	P	
P	Q	Z			Z	Q
R	Q		Q		Q	Q
M	Q			Q	Q	Q
Z	P	Z	Q	Q	F	F
*		Q	Q	Q	F	

State M

L \ R	Q	P	R	M	Z	*
Q	P	M	R	M		R
P	M			M	Z	Z
R	R		R	R	Z	Z
M	M	M	R	M	Z	Z
Z		Z	Z	Z		
*	R	Z	Z	Z		

State R

L \ R	Q	P	R	M	Z	*
Q			R	Z	P	
P			M	M	R	
R	R	M		M		
M	Z	M	M	M	R	Z
Z	P	R		R		
*				Z		

State P

L \ R	Q	P	R	M	Z	*
Q	Z	Z	R		R	
P	Z		Z		Z	
R	R	Z	Z			Z
M				P	R	
Z	R	Z		R	Z	Z
*			Z		Z	

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	Q	Q	Q	Q	M	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q
1	Q	Q	Q	M	P	M	Q	Q	Q	Q	Q	Q	Q	Q	Q
2	Q	Q	M	M	P	M	M	Q	Q	Q	Q	Q	Q	Q	Q
3	Q	M	M	M	P	M	M	M	Q	Q	Q	Q	Q	Q	Q
4	M	M	M	M	P	M	M	M	M	Q	Q	Q	Q	Q	Q
5	Z	M	M	M	P	M	M	M	M	M	Q	Q	Q	Q	Q
6	Q	Z	M	M	P	M	M	M	M	M	M	Q	Q	Q	Q
7	Q	Q	Z	M	P	M	M	M	M	M	M	M	Q	Q	Q
8	Q	Q	Q	Z	P	M	M	M	M	M	M	M	M	Q	Q
9	Q	Q	Q	Q	R	M	M	M	M	M	M	M	M	M	Q
10	Q	Q	Q	Q	Z	R	M	M	M	M	M	M	M	M	M
11	Q	Q	Q	Q	Q	R	R	M	M	M	M	M	M	M	Z
12	Q	Q	Q	Q	Q	R	M	R	M	M	M	M	M	Z	Q
13	Q	Q	Q	Q	Q	Z	R	M	R	M	M	M	Z	Q	Q
14	Q	Q	Q	Q	Q	Q	R	R	M	R	M	Z	Q	Q	Q
15	Q	Q	Q	Q	Q	Q	R	M	R	M	Z	Q	Q	Q	Q
16	Q	Q	Q	Q	Q	Q	Z	R	M	Z	Q	Q	Q	Q	Q
17	Q	Q	Q	Q	Q	Q	Q	R	Z	Q	Q	Q	Q	Q	Q
18	Q	Q	Q	Q	Q	Q	Q	P	Q	Q	Q	Q	Q	Q	Q
19	Q	Q	Q	Q	Q	Q	P	Z	P	Q	Q	Q	Q	Q	Q
20	Q	Q	Q	Q	Q	P	R	Z	R	P	Q	Q	Q	Q	Q
21	Q	Q	Q	Q	P	R	R	Q	R	R	P	Q	Q	Q	Q
22	Q	Q	Q	P	R	M	R	Q	R	M	R	P	Q	Q	Q
23	Q	Q	P	R	M	R	Z	Q	Z	R	M	R	P	Q	Q
24	Q	P	R	M	R	R	Q	Q	Q	R	R	M	R	P	Q
25	P	R	M	R	M	R	Q	Q	Q	R	M	R	M	R	P
26	Z	M	R	M	R	Z	Q	Q	Q	Z	R	M	R	M	Z
27	Q	Z	M	R	R	Q	Q	Q	Q	Q	R	R	M	Z	Q
28	Q	Q	Z	M	R	Q	Q	Q	Q	Q	R	M	Z	Q	Q
29	Q	Q	Q	Z	Z	Q	Q	Q	Q	Q	Z	Z	Q	Q	Q
30	Q	Q	Q	P	P	Q	Q	Q	Q	Q	P	P	Q	Q	Q
31	Q	Q	P	Z	Z	P	Q	Q	Q	P	Z	Z	P	Q	Q
32	Q	P	R	Z	Z	R	P	Q	P	R	Z	Z	R	P	Q
33	P	R	R	Q	Q	R	R	P	R	R	Q	Q	R	R	P
34	Z	M	R	Q	Q	R	M	Z	M	R	Q	Q	R	M	Z
35	Q	Z	Z	Q	Q	Z	Z	Q	Z	Z	Q	Q	Z	Z	Q
36	Q	P	P	Q	Q	P	P	Q	P	P	Q	Q	P	P	Q
37	P	Z	Z	P	P	Z	Z	P	Z	Z	P	P	Z	Z	P
38	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z
39	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F

Firing Squad Synchronization Problem in Cellular Automata, Figure 14

Transition table for generalized symmetric six-state protocol (*left*) and snapshots for synchronization algorithm on 15 cells with a *General* on C_5

A non-trivial, new symmetric six-state $3n$-step generalized firing squad synchronization algorithm is also presented in Umeo, Maeda and Hongyo [61]. Figure 14 gives a list of transition rules for the 6-state generalized synchronization algorithm and snapshots of configurations on 15 cells. The symbol "M" is the general state.

Theorem 11 ([61]) *There exists a symmetric 6-state cellular automaton that can solve the generalized firing squad synchronization problem in* $\max(k, n-k+1) + 2n + O(\log n)$ *steps.*

In addition, a state-change complexity is studied in $3n$-step firing squad synchronization algorithms. It has been shown that the six-state algorithms presented above have $O(n^2)$ state-change complexity, on the other hand, the *thread-like* $3n$-step algorithms developed so far have $O(n \log n)$ state-change complexity. Here, the following table presents a quantitative comparison of the $3n$-step synchronization algorithms developed so far.

Delayed Firing Squad Synchronization Algorithm

This section introduces a *freezing-thawing* technique that yields a delayed synchronization algorithm for one-dimensional arrays. The technique is very useful in the design of time-efficient synchronization algorithms for one- and two-dimensional arrays in Umeo [52], Yunès [77] and Umeo and Uchino [65]. A similar technique was used by Romani [37] in the tree synchronization. The technique is stated as in the following theorem.

Theorem 12 ([52]) *Let t_0, t_1, t_2 and Δt be any integer such that $t_0 \geq 0$, $t_0 \leq t_1 \leq t_0 + n - 1$, $t_1 \leq t_2$ and $\Delta t = t_2 - t_1$. It is assumed that a usual optimum-time synchronization operation is started at time $t = t_0$ by generating a special signal at the left end of one-dimensional array and the right end cell of the array receives another special signals from outside at time $t = t_1$ and t_2, respectively. Then, there exists a one-dimensional cellular automaton that can synchronize the array of length n at time $t = t_0 + 2n - 2 + \Delta t$.*

The array operates as follows:

1. Start an optimum-time firing squad synchronization algorithm at time $t = t_0$ at the left end of the array. A 1/1-speed signal is propagated towards the right direction to wake-up cells in quiescent state. The signal is referred to as *wake-up signal*. A *freezing* signal is given from outside at time $t = t_1$ at the right end of the array. The signal is propagated in the left direction at its maximum speed, that is, 1 cell per 1 step, and freezes the configuration progressively. Any cell that receives the freezing signal from its right neighbor has to stop its state-change and transmits the freezing signal to its left neighbor. The frozen cell keeps its state as long as no thawing signal will arrive.
2. A special signal supplied with outside at time $t = t_2$ is used as a *thawing* signal that thaws the frozen configuration. The thawing signal forces the frozen cell to resume its state-change procedures immediately. See Fig. 15 (left). The signal is also transmitted toward the left end at speed 1/1.

The readers can see how those three signals work. The entire configuration can be freezed during Δt steps and the synchronization on the array is delayed for Δt steps. It is easily seen that the freezing signal can be replaced by the reflected signal of the wake-up signal, that is generated at

Firing Squad Synchronization Problem in Cellular Automata, Table 3
A comparison of 3*n*-step firing squad synchronization algorithms

Algorithm	# States	# Rules	Time complexity	State-change complexity	Generals's position	Type	Notes	Ref.
Minsky and MacCarthy	13	–	$3n + \theta_n \log n + c$	$O(n \log n)$	left	thread	$0 \le \theta_n < 1$	[28]
Fischer	15	–	$3n - 4$	$O(n \log n)$	left	thread	–	[8]
Yunès	7	105	$3n \pm 2\theta_n \log n + c$	$O(n \log n)$	left	thread	$0 \le \theta_n < 1$	[76]
Yunès	7	107	$3n \pm 2\theta_n \log n + c$	$O(n \log n)$	left	thread	$0 \le \theta_n < 1$	[76]
Settle and Simon	6	134	$3n + 1$	$O(n^2)$	right	plane	–	[43]
Settle and Simon	7	127	$2n - 2 + k$	$O(n^2)$	arbitrary	plane	–	[43]
Umeo et al.	6	78	$3n + O(\log n)$	$O(n^2)$	left	plane	–	[61]
Umeo et al.	6	115	$\max(k, n - k + 1) + 2n + O(\log n)$	$O(n^2)$	arbitrary	plane	–	[61]
Umeo and Yanagihara	5	67	$3n - 3$	$O(n^2)$	left/right	plane	$n = 2^k$, $k = 1, 2, \ldots$	[67]
Yunès	6	105	$3n + \lceil \log n \rceil - 3$	$O(n \log n)$	left	thread		[79]

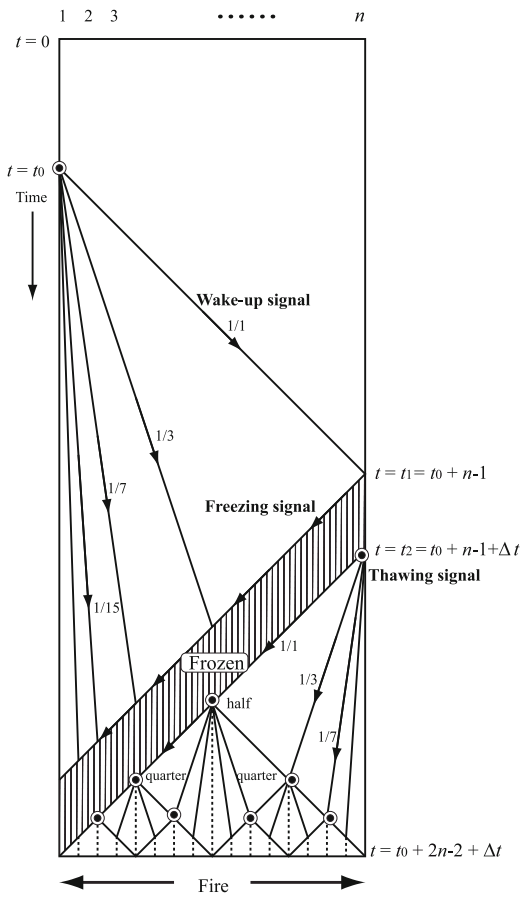

	1	2	3	4	5	6	7	8	9	10	11
0	M	L	L	L	L	L	L	L	L	L	L
1	M	C	L	L	L	L	L	L	L	L	L
2	M	C	C	L	L	L	L	L	L	L	L
3	M	C	R	C	L	L	L	L	L	L	L
4	M	C	R	B	C	L	L	L	L	L	L
5	M	C	C	B	R	C	L	L	L	L	L
6	M	C	C	R	R	B	C	L	L	L	L
7	M	C	C	R	B	B	R	C	L	L	L
8	M	C	C	C	B	R	R	B	C	L	L
9	M	C	R	C	R	R	B	B	R	C	L
10	M	C	R	C	R	B	B	R	R	B	F5
11	M	C	R	C	C	B	R	R	B	FB	F4
12	M	C	R	B	C	R	R	B	FB	FB	F3
13	M	C	C	B	C	R	B	FB	FB	FB	F2
14	M	C	C	B	C	C	FB	FB	FB	FB	F1
15	M	C	C	B	R	FC	FB	FB	FB	FB	M
16	M	C	C	R	FR	FC	FB	FB	FB	A	M
17	M	C	C	FR	FR	FC	FB	FB	Q	R	M
18	M	C	FC	FR	FR	FC	FB	Q	R	Q	M
19	M	FC	FC	FR	FR	FC	Q	R	L	Q	M
20	FM	FC	FC	FR	FR	M	R	A	Q	Q	M
21	FM	FC	FC	FR	Q	M	C	L	Q	Q	M
22	FM	FC	FC	Q	Q	M	C	C	Q	Q	M
23	FM	FC	M	M	Q	M	C	M	M	Q	M
24	FM	M	M	M	M	M	M	M	M	M	M
25	F	F	F	F	F	F	F	F	F	F	F

Firing Squad Synchronization Problem in Cellular Automata, Figure 15
Time-space diagram for delayed firing squad synchronization scheme based on the *freezing-thawing* technique (*left*) and a delayed (for $\Delta t = 5$ steps) configuration in Balzer's optimum-time firing squad synchronization algorithm on $n = 11$ cells (*right*)

the right end cell at time $t = t_0 + n - 1$. See Fig. 15. The scheme is referred to as *freezing-thawing* technique.

Fault-Tolerant Firing Squad Synchronization Problem

Consider a one-dimensional array of cells, some of which are defective. At time $t = 0$, the left end cell C_1 is in the *fire-when-ready* state, which is the initialization signal for the array. The *fault-tolerant* firing squad synchronization problem for cellular automata with *defective* cells is to determine a description of cells that ensures all *intact* cells enter the *fire* state at exactly the same time and for the first time. The fault-tolerant firing squad synchronization problem has been studied in Kutrib and Vollmar [22,23], Umeo [52] and Yunès [77].

Cellular Automata with Defective Cells

- **Intact and Defective Cells:** Each cell has its own self-diagnosis circuit that diagnoses itself before its operation. A consecutive defective (intact) cells are referred to as a *defective* (*intact*) segment, respectively. Any defective and intact cell can detect whether its neighbor cells are defective or not. Cellular arrays are assumed to have an intact segment at its left and right ends. New defections do not occur during the operational lifetime on any cell.
- **Signal Propagation in a Defective Segment:** It is assumed that any cell in defective segment can only transmit the signal to its right or left neighbor depending on the direction in which it comes to the defective segment. The speed of the signal in any defective segment is fixed to 1/1, that is, one cell per one step. In defective segments, both the information carried by the signal and the direction in which the signal is propagated are preserved without any modifications. Thus, one can see that any defective segment has two one-way pipelines that can transfer one state at 1/1 speed in either direction. Note that from a standard viewpoint of state transition of usual CA each cell in a defective segment can change its internal states in a specific manner.

The array consists of p defective segments and $(p + 1)$ intact segments, where they are denoted by I_i and D_j, respectively and p is any positive integer. Let n_i and m_j be number of cells on the ith intact and jth defective segments, where i and j be any integer such that $1 \le i \le p + 1$ and $1 \le j \le p$. Let n be the number of cells of the array such that $n = (n_1+m_1)+(n_2+m_2)+, \ldots, +(n_p+m_p)+n_{p+1}$.

Fault-Tolerant Firing Squad Synchronization Algorithms Umeo [52] studied the synchronization algorithms for such arrays that there are locally more intact cells than defective ones, i. e., $n_i \ge m_i$ for any i such that $1 \le i \le p$. First, consider the case $p = 1$ where the array has one defective segment and $n_1 \ge m_1$. Figure 16 illustrates a simple synchronization scheme. The fault-tolerant synchronization algorithm for one defective segment is stated as follows:

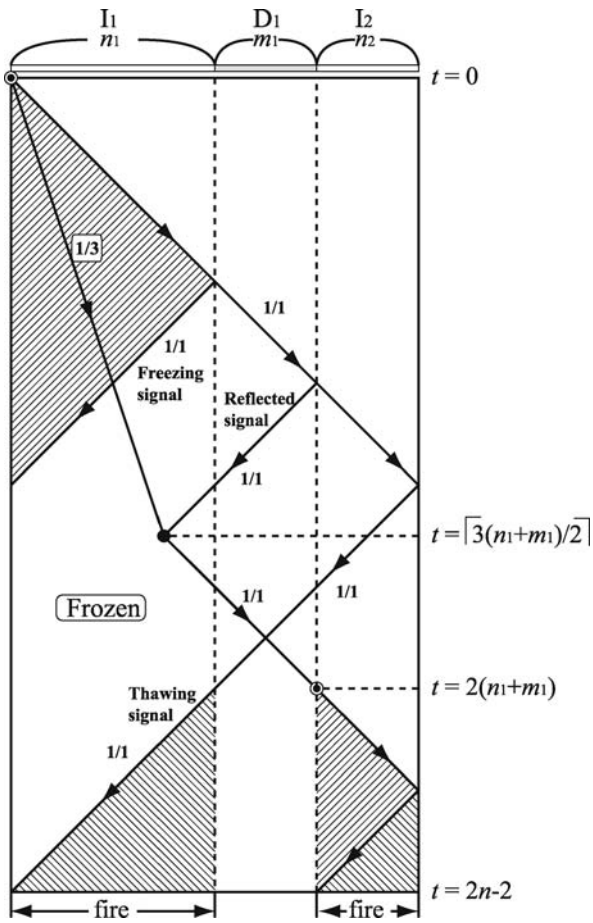

Firing Squad Synchronization Problem in Cellular Automata, Figure 16
Time-space diagram for optimum-time firing squad synchronization algorithm with one defective segment

Theorem 13 ([52]) *Let M be any cellular array of length n with one defective and two intact segments such that $n_1 \ge m_1$, where n_1 and m_1 denote the number of cells on the first intact and defective segments, respectively. Then, M is synchronizable in $2n - 2$ optimum-time.*

The synchronization scheme above can be generalized to arrays with multiple defective segments more than two. Figure 17 shows the synchronization scheme for a cellu-

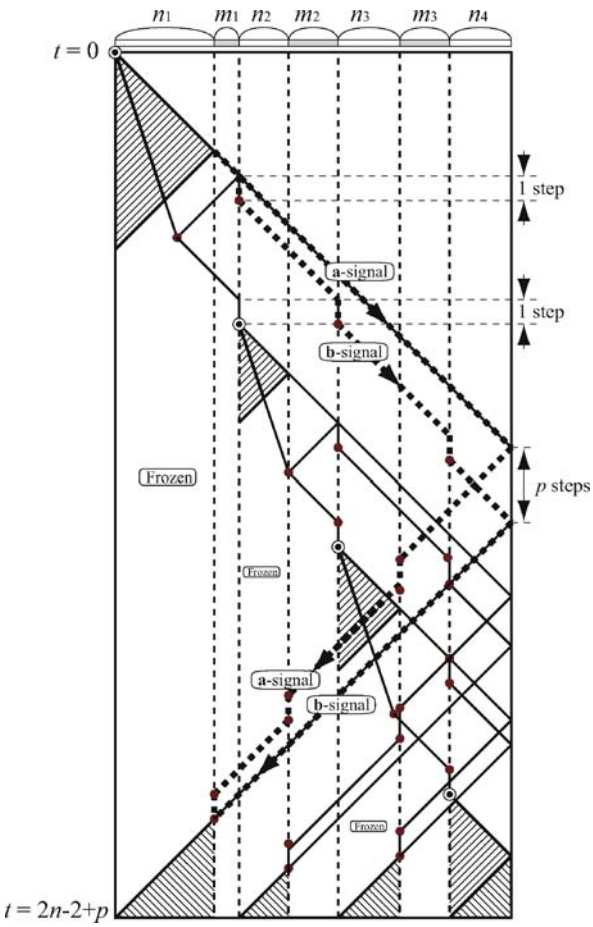

Firing Squad Synchronization Problem in Cellular Automata, Figure 17
Time-space diagram for optimum-time firing squad synchronization algorithm with three defective segments

lar array with three defective segments. Details of the algorithm can be found in Umeo [52].

Theorem 14 ([52]) *Let p be any positive integer and M be any cellular array of length n with p defective segments, where $n_i \geq m_i$ and $n_i + m_i \geq p - i$, for any i such that $1 \leq i \leq p$. Then, M is synchronizable in $2n - 2 + p$ steps.*

Partial Solutions

The original firing squad synchronization problem is defined to synchronize all cells of one-dimensional array. In this section, consider a *partial FSSP solution* that can synchronize an infinite number of cells, but not all. The first partial solution was given by Umeo and Yanagihara [67]. They proposed a five-state solution that can synchronize any one-dimensional cellular array of length $n = 2^k$ in $3n - 3$ steps for any positive integer k. Figure 18 shows the five-state transition table consisting of 67 rules and its snapshots for $n = 8$ and 16. In the transition table, the symbols "R", "Q", "F" and "*" represent the general, quiescent, firing and boundary states, respectively.

Theorem 15 ([67]) *There exists a 5-state cellular automaton that can synchronize any array of length $n = 2^k$ in $3n - 3$ steps, where k is any positive integer.*

Surprisingly, Yunès [80] and Umeo, Kamikawa and Yunès [58] proposed four-state synchronization protocols which are based on an algebraic property of Wolfram's two-state cellular automata.

Theorem 16 ([58,80]) *There exists a 4-state cellular automaton that can synchronize any array of length $n = 2^k$ in non-optimum $2n - 1$ steps, where k is any positive integer.*

Figure 19 shows the four-state transition table given by Yunès [80] which is based on Wolfram's Rule 60. It consists of 32 rules. Snapshots for $n = 16$ cells are also illustrated in Fig. 19. In the transition table, the symbols "G", "Q", "F" and "*" represent the general, quiescent, firing and boundary states, respectively. Figure 20 shows the four-state transition table given by Umeo, Kamikawa and Yunès [58] which is based on Wolfram's Rule 150. It has 32 transition rules. Snapshots for $n = 16$ are given in the figure. In the transition table, the symbols "G", "Q", "F" and "*" represent the general, quiescent, firing and boundary states, respectively.

Umeo, Kamikawa and Yunès [58] proposed a different, but looking-similar, 4-state protocol based on Wolfram's Rule 150. Figure 21 shows the four-state transition table consisting of 37 rules and its snapshots for $n = 9$ and 17. The four-state protocol has some desirable properties. Note that the algorithm operates in optimum-step and its transition rule is symmetric. The state of a general can be either "G" or "A". Its initial position can be at the left or right end.

Theorem 17 ([58]) *There exists a symmetric 4-state cellular automaton that can synchronize any array of length $n = 2^k + 1$ in $2n - 2$ optimum-steps, where k is any positive integer.*

Yunès [80] has given a state lower bound for the partial solution:

Theorem 18 ([80]) *There is no 3-state partial solution.*

Thus, the 4-state partial solutions given above are optimum in state-number complexity in partial solutions.

State Q

Left\Right	Q	R	L	S	*
Q	Q	Q	L	Q	Q
R	R	R			R
L	Q		L	S	Q
S	Q	S			
*	Q	Q	L	Q	

State L

Left\Right	Q	R	L	S	*
Q	L	S	Q	Q	
R	Q	Q	R	R	Q
L	L		L	L	
S	R	F			F
*			R	R	

State R

Left\Right	Q	R	L	S	*
Q	R	R	Q	L	
R	Q	R	L		L
L	S		Q	F	
S	Q	R	L		L
*	S		Q	F	

State S

Left\Right	Q	R	L	S	*
Q	Q	S	L	Q	
R	R			F	
L	S			S	
S	Q	S	F		
*	Q	S	F		

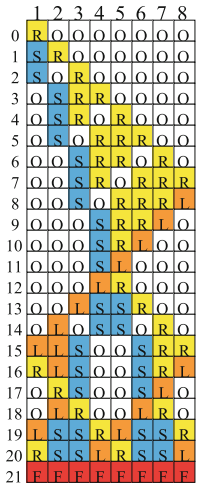

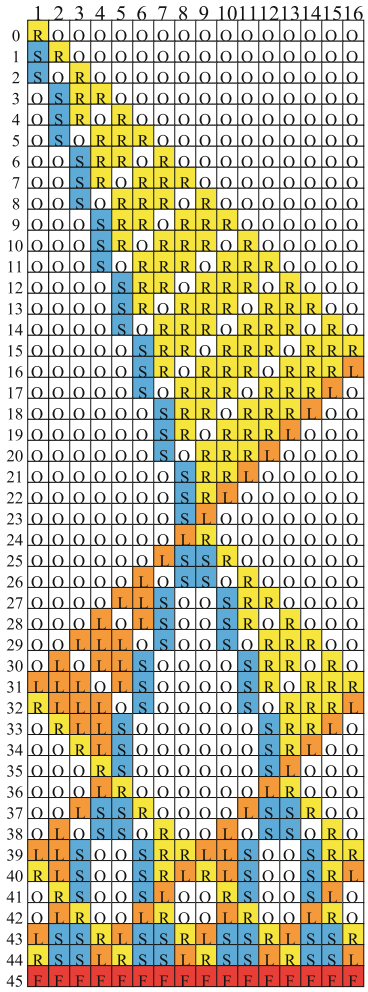

Firing Squad Synchronization Problem in Cellular Automata, Figure 18
Transition table for the five-state protocol (*left*) and its snapshots of configurations on 8 and 16 cells (*right*)

Synchronization Algorithm for One-Bit Communication Cellular Automata

In the study of cellular automata, the amount of bit-information exchanged at one step between neighboring cells has been assumed to be O(1)-bit. An O(1)-bit communication CA is a *conventional* cellular automaton in which the number of communication bits exchanged at one step between neighboring cells is assumed to be O(1)-bit, however, such an inter-cell bit-information exchange has been hidden behind the definition of conventional automata-theoretic finite state description. On the other hand, the 1-bit inter-cell communication model is a new cellular automaton in which inter-cell communication is restricted to 1-bit data, referred to as the 1-bit CA model (CA_{1-bit}). The number of internal states of the CA_{1-bit} is assumed to be finite in the usual sense. The next state of each cell is determined by the present state of that cell and two binary 1-bit inputs from its left and right neighbor cells. Thus, the CA_{1-bit} can be thought of as one of the most powerless and the simplest models in a variety of CA's. A precise definition of the CA_{1-bit} can be found in Umeo [51] and Umeo and Kamikawa [57]. Mazoyer [26] and Nishimura and Umeo [32] each designed an optimum-time synchronization algorithm on the CA_{1-bit} based on Balzer's algorithm and Waksman's algorithm, respectively. Figure 22 shows a configuration of the 1-bit synchronization algorithm on 15 cells that is based on the design of Nishimura and Umeo [32]. Each cell has 78 internal states and 208 transition rules. The small black triangles ► and ◄ indicate a 1-signal transfer in the right or left direction, respectively, between neigh-

Q		Right State			
		Q	R	G	*
Left State	Q	Q	R	Q	Q
	R	Q	R		
	G	G		G	G
	*				

G		Right State			
		Q	R	G	*
Left State	Q	G		G	
	R	G	F	G	
	G	Q	R	Q	R
	*	G	F	G	

R		Right State			
		Q	R	G	*
Left State	Q	R	Q	R	R
	R	R	Q	R	R
	G		G	F	F
	*				

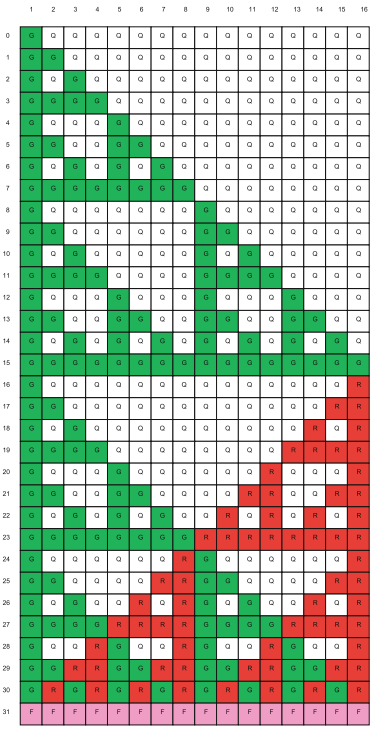

Firing Squad Synchronization Problem in Cellular Automata, Figure 19

Transition table for the four-state protocol based on Wolfram's Rule 60 (*left*) and its snapshots of configurations on 16 cells (*right*)

Q		Right State			
		Q	G	R	*
Left State	Q	Q	G	R	Q
	G	G	Q		G
	R	R		Q	
	*				

G		Right State			
		Q	G	R	*
Left State	Q	G	Q		
	G	Q	G	R	R
	R	G	G	F	
	*	G	G	F	

R		Right State			
		Q	G	R	*
Left State	Q	R	R	Q	R
	G		F	G	F
	R	Q	R	R	R
	*				

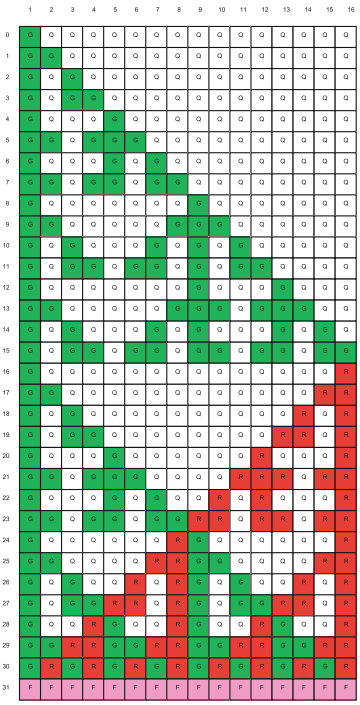

Firing Squad Synchronization Problem in Cellular Automata, Figure 20

Transition table for the four-state protocol based on Wolfram's Rule 150 (*left*) and its snapshots of configurations on 16 cells (*right*)

boring cells. A symbol in a cell shows an internal state of the cell.

Theorem 19 ([26,32]) *There exists a CA_{1-bit} that can synchronize n cells in optimum $2n-2$ steps.*

Synchronization Algorithms for Multi-Bit Communication Cellular Automata

Section "Synchronization Algorithms for Multi-Bit Communication Cellular Automata" studies a trade-off between internal states and communication bits in firing squad synchronization protocols for k-bit communication-restricted cellular automata ($CA_{k-\text{bit}}$) and propose several time-optimum state-efficient bit-transfer-based synchronization protocols. It is shown that there exists a 1-state $CA_{5-\text{bit}}$ that can synchronize any n cells in $2n-2$ optimum-step. The result is interesting, since one knows that there exists no 4-state synchronization algorithm on *conventional* O(1)*-bit communication* cellular automata. A bit-transfer complexity is also introduced to measure the efficiency of synchronization protocols. It is shown that $\Omega(n \log n)$ bit-transfer is a lower-bound for synchronizing n cells in $(2n-2)$ steps. In addition, each optimum-time/non-optimum-time synchronization protocols presented has an $O(n^2)$ bit-transfer complexity, respectively. Most of the results presented here are from Umeo, Yanagihara and Kanazawa [68]. A computational relation between the conventional CA and $CA_{k-\text{bit}}$ is stated as follows:

Lemma 20 ([57]) *Let N be any s-state conventional cellular automaton with time complexity T(n). Then, there exists a $CA_{1-\text{bit}}$ which can simulate N in kT(n) steps, where k is a positive constant integer such that $k = \lceil \log_2 s \rceil$.*

Lemma 21 ([68]) *Let N be any s-state conventional cellular automaton. Then, there exists an s-state $CA_{k-\text{bit}}$ which can simulate N in real time, where k is a positive integer such that $k = \lceil \log_2 s \rceil$.*

The following theorems show a trade-off between internal states and communication bits, and present state-efficient synchronization protocols $\mathcal{P}_i, 1 \le i \le 5$. In some sense, no internal state is necessary to synchronize the whole array, as is shown in Theorem 27. The protocol design is based on the 6-state Mazoyer's algorithm given in Mazoyer [25].

Q		Right State			
		Q	G	A	*
Left State	Q	Q	G	A	Q
	G	G	A	Q	A
	A	A	Q	G	G
	*	Q	A	G	

G		Right State			
		Q	G	A	*
Left State	Q	A		Q	A
	G		F		F
	A	Q		G	G
	*	A	F	G	

A		Right State			
		Q	G	A	*
Left State	Q	G	Q		G
	G	Q	A		A
	A			F	F
	*	G	A	F	

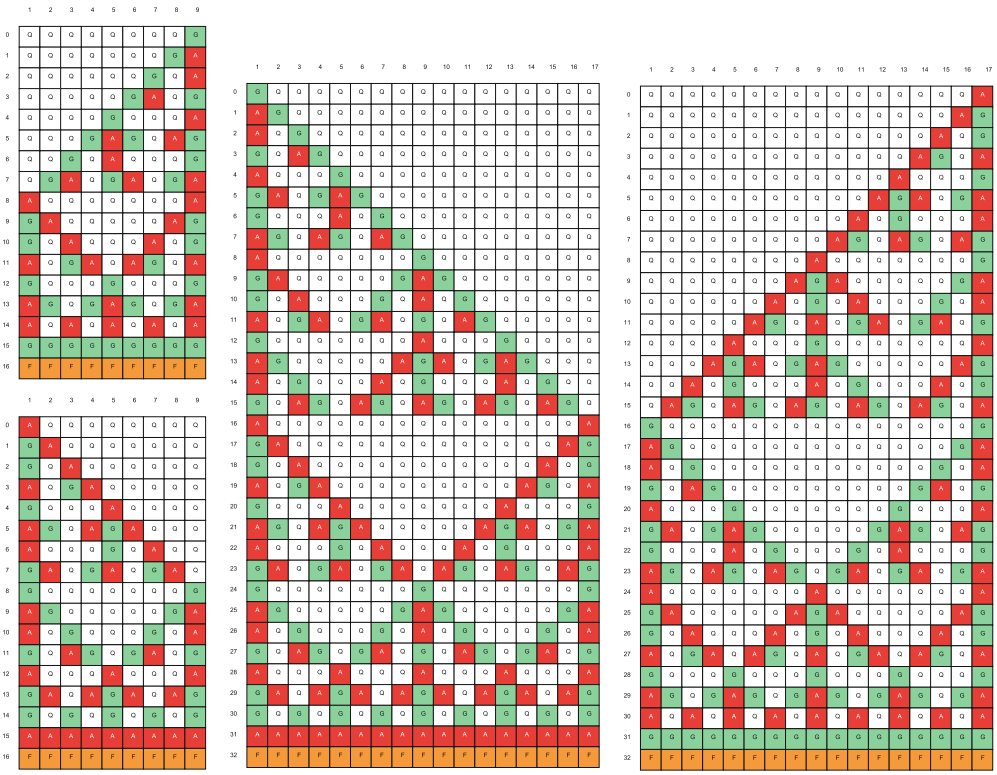

Firing Squad Synchronization Problem in Cellular Automata, Figure 21
Transition table for the four-state protocol (*left*) and its snapshots of configurations on 9 and 17 cells (*right*)

Firing Squad Synchronization Problem in Cellular Automata, Table 4
A comparison of optimum-time/non-optimum-time firing squad synchronization protocols for multi-bit communication cellular automata

Synchronization protocol	Communication bits transferred	# of states	# of Transition rules	Time complexity	One-/Two-sided recursiveness
$\mathcal{P}_1$	1	54	207	$2n-1$	One-sided
$\mathcal{P}_2$	2	6	60	$2n-2$	One-sided
$\mathcal{P}_3$	3	4	76	$2n-2$	One-sided
$\mathcal{P}_4$	4	3	87	$2n-2$	One-sided
$\mathcal{P}_4^*$	4	2	88	$2n-2$	One-sided
$\mathcal{P}_5$	5	1	114	$2n-2$	One-sided
Mazoyer [26]	1	58	–	$2n-2$	Two-sided
Mazoyer [26]	12	3	–	$2n-2$	–
Nishimura and Umeo [32]	1	78	208	$2n-2$	Two-sided

Theorem 22 ([68]) *There exists a 54-state $CA_{1-\text{bit}}$ with protocol $\mathcal{P}_1$ that can synchronize any n cells in $2n-1$ optimum-step.*

Theorem 23 ([68]) *There exists a 6-state $CA_{2-\text{bit}}$ with protocol $\mathcal{P}_2$ that can synchronize any n cells in $2n-2$ optimum-step.*

Theorem 24 ([68]) *There exists a 4-state $CA_{3-\text{bit}}$ with protocol $\mathcal{P}_3$ that can synchronize any n cells in $2n-2$ optimum-step.*

Figure 23 illustrates snapshots of the 6-state $(2n-2)$-step synchronization protocol $\mathcal{P}_2$ operating on $CA_{2-\text{bit}}$ (left) and for the 4-state protocol $\mathcal{P}_3$ operating on $CA_{3-\text{bit}}$ with 24 cells (right).

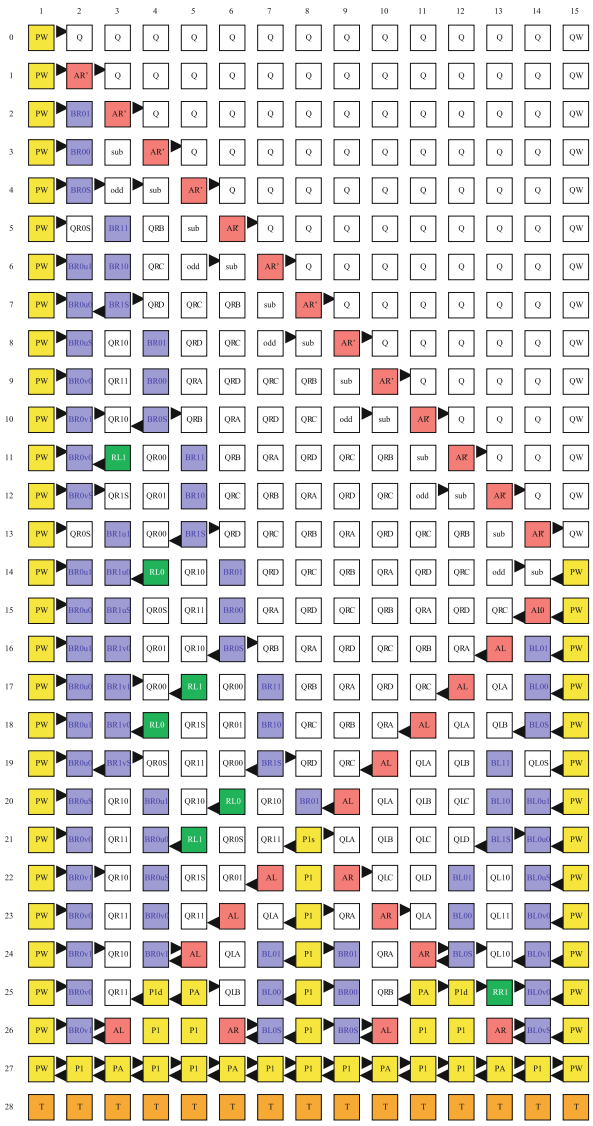

Firing Squad Synchronization Problem in Cellular Automata, Figure 22
A configuration of optimum-time synchronization algorithm with 1-bit inter-cell communication on 15 cells

Theorem 25 ([68]) *There exists a 3-state $CA_{4-\text{bit}}$ with protocol $\mathcal{P}_4$ that can synchronize any n cells in $2n-2$ optimum-step.*

Theorem 26 ([68]) *There exists a 2-state $CA_{4-\text{bit}}$ with protocol $\mathcal{P}_4^*$ that can synchronize any n cells in $2n-2$ optimum-step.*

Theorem 27 ([68]) *There exists a 1-state $CA_{5-\text{bit}}$ with protocol $\mathcal{P}_5$ that can synchronize any n cells in $2n-2$ optimum-step.*

Figures 24 and 25 illustrate snapshots of the 3-state $(2n-2)$-step synchronization protocol $\mathcal{P}_4$ operating on $\text{CA}_{4-\text{bit}}$ (left) and for the 1-state protocol $\mathcal{P}_5$ operating on $\text{CA}_{5-\text{bit}}$ (right).

Let $BT(n)$ be total number of bits transferred which is needed for synchronizing n cells on $\text{CA}_{k-\text{bit}}$. By using the similar technique developed by Vollmar [73], a lower-bound on bit-transfer complexity can be established for synchronizing n cells on $\text{CA}_{k-\text{bit}}$ in a way such that $BT(n) = \Omega(n \log n)$. In addition, it is shown that each synchronization protocol $\mathcal{P}_i, 1 \le i \le 5$ and $\mathcal{P}_4^*$ presented above has an $O(n^2)$ bit-transfer complexity, respectively.

Theorem 28 ([68]) *$\Omega(n \log n)$ bit-transfer is a lower bound for synchronizing n cells on $CA_{k-\text{bit}}$ in $(2n-2)$ steps.*

Theorem 29 ([68]) *Each optimum-time/non-optimum-time synchronization protocols $\mathcal{P}_i, 1 \le i \le 5$ and $\mathcal{P}_4^*$ has an $O(n^2)$ bit-transfer complexity, respectively.*

Firing Squad Synchronization Problem on Two-Dimensional Arrays

Figure 26 shows a finite two-dimensional (2-D) cellular array consisting of $m \times n$ cells. The array operates in lock-step mode in such a way that the next state of each cell (except border cells) is determined by both its own present state and the present states of its north, south, east and west neighbors. All cells (*soldiers*), except the north-west corner cell (*general*), are initially in the quiescent state at time $t = 0$ with the property that the next state of a quiescent cell with quiescent neighbors is the quiescent state again. At time $t = 0$, the north-west corner cell C_{11} is in the *fire-when-ready* state, which is the initiation signal for the synchronization. The firing squad synchronization problem is to determine a description (state set and next-state function) of cells that ensures all cells enter the *fire* state at exactly the same time and for the first time. The set of states and transition function must be independent of m and n.

Several synchronization algorithms on 2-D arrays have been proposed by Beyer [3], Grasselli [14], Shinahr [44], Szwerinski [45], Umeo, Maeda and Fujiwara [59], and Umeo, Hisaoka and Akiguchi [53]. Most of the synchronization algorithms for 2-D arrays are based on mapping schemes which map synchronized configurations for 1-D arrays onto 2-D arrays. Section "Firing Squad Synchronization Problem on Two-Dimensional Arrays" presents

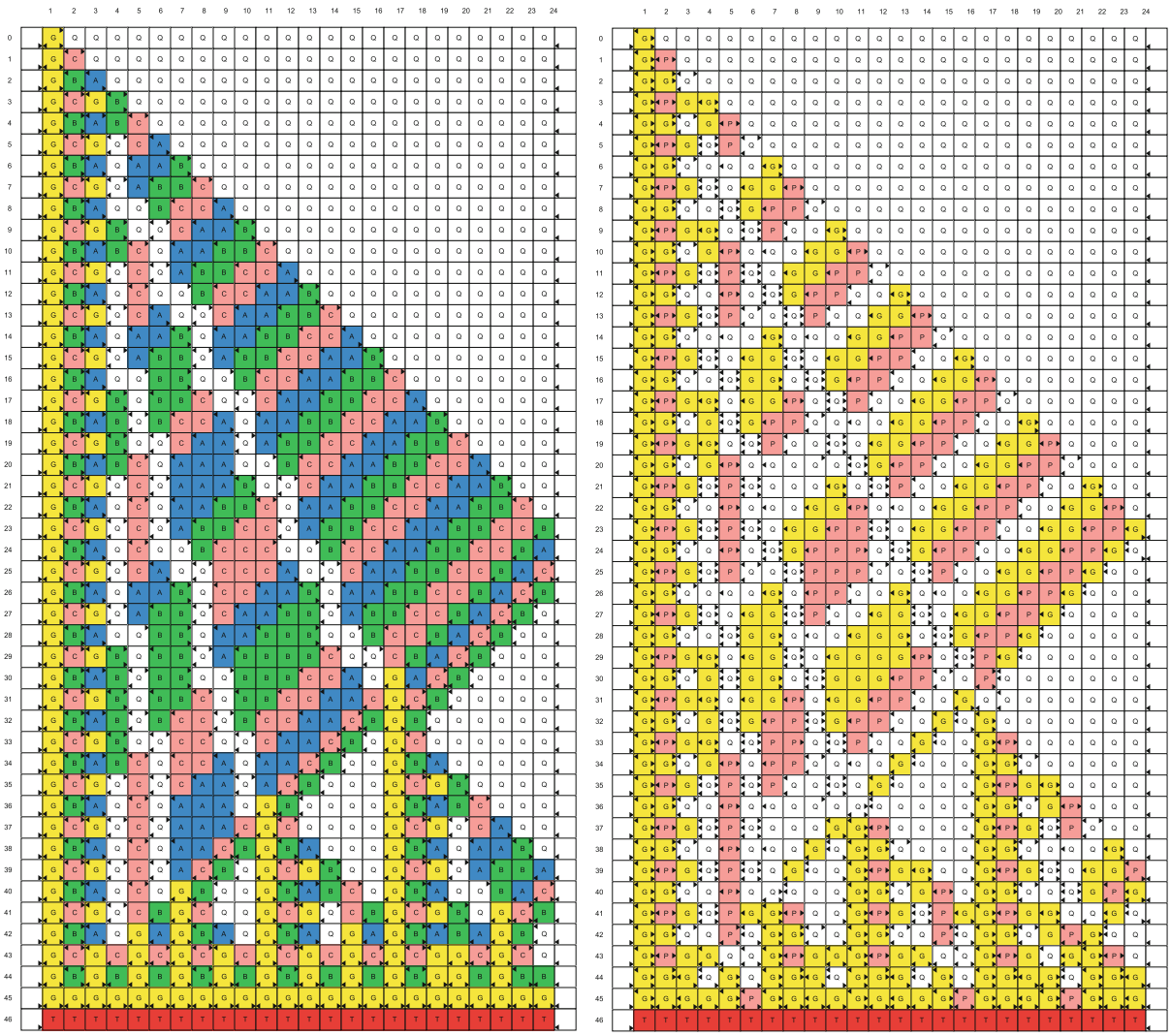

Firing Squad Synchronization Problem in Cellular Automata, Figure 23
Snapshots for the 6-state protocol $\mathcal{P}_{\in}$ operating on CA_{2-bit} with 24 cells (*left*) and for the 4-state protocol $\mathcal{P}_3$ operating on CA_{3-bit} with 24 cells (*right*)

such several mapping schemes that yield time-efficient 2-D synchronization algorithms.

Orthogonal Mapping: A Simple Linear-Time Algorithm

In this section, a very simple synchronization algorithm is provided for 2-D arrays. The overall of the algorithm is as follows:

1. First, *synchronize* the first column cells using a usual optimum-step 1-D algorithm with a general at one end, thus requiring $2m - 2$ steps.
2. Then, *start the row synchronization operation* on each row simultaneously. Additional $2n - 2$ steps are required for the row synchronization. Totally, its time complexity is $2(m + n) - 4$ steps.

The implementation is referred to as *orthogonal mapping*. It is shown that $s + 2$ states are enough for the implementation of the algorithm above, where s is the number of internal states of the 1-D base algorithm. Figure 27 shows snapshots of the 8-state synchronization algorithm running on a rectangular array of size 4×6.

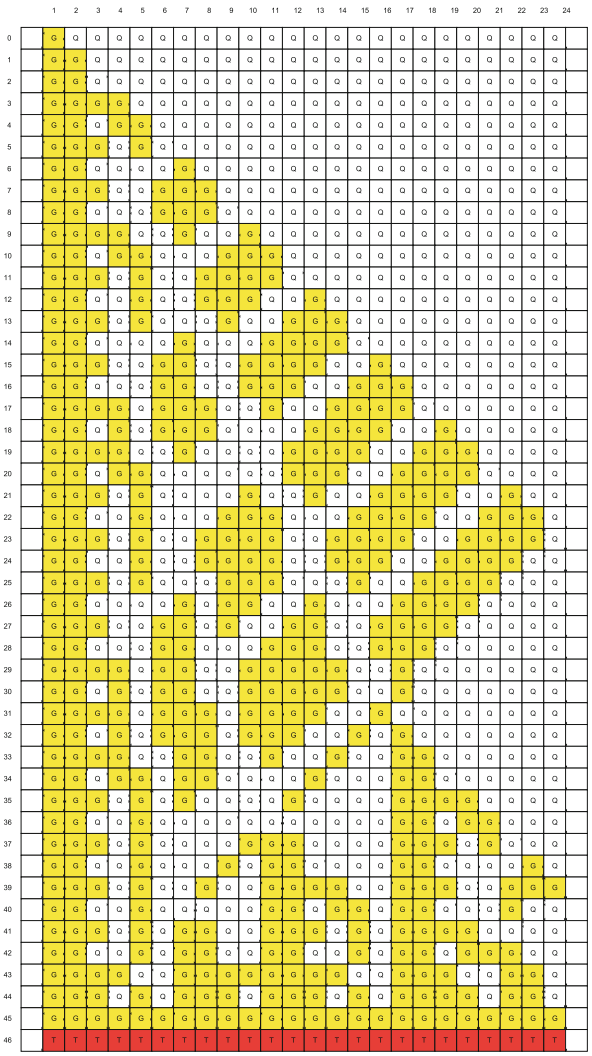

Firing Squad Synchronization Problem in Cellular Automata, Figure 24
Snapshots for 3-state ($2n - 2$)-step synchronization protocol $\mathcal{P}_4$ operating on CA_{4-bit} with 24 cells

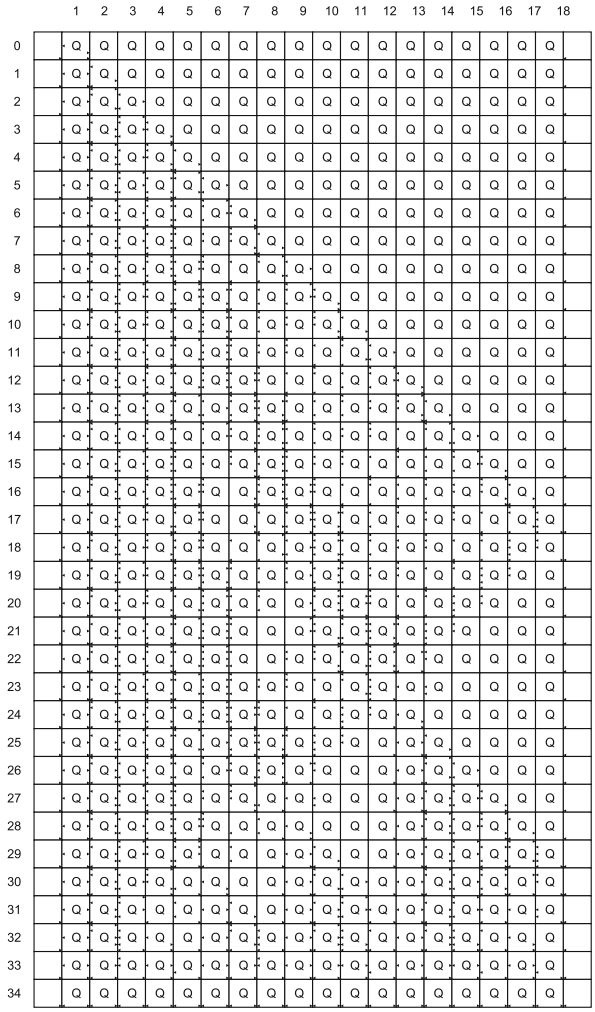

Firing Squad Synchronization Problem in Cellular Automata, Figure 25
Snapshots for 1-state ($2n - 2$)-step synchronization protocol $\mathcal{P}_5$ operating on CA_{5-bit} with 18 cells

Theorem 30 *There exists an $(s + 2)$-state protocol for synchronizing any $m \times n$ rectangular arrays in $2(m + n) - 4$ steps, where s is number of states of any optimum-time 1-D synchronization protocol.*

L-shaped Mapping: Shinar's Optimum-Time Algorithm

The first optimum-time 2-D synchronization algorithm was developed by Shinar [44] and Beyer [3]. The rectangular array of size $m \times n$ is regarded as $\min(m, n)$ rotated (90° in clockwise direction) L-shaped 1-D arrays, where they are synchronized independently using the generalized firing squad synchronization algorithm. The configuration of the generalized synchronization on 1-D array can be mapped on 2-D array. See Fig. 28. Thus, an $m \times n$ array synchronization problem is reduced to independent $\min(m, n)$ 1-D generalized synchronization problems such that: $\mathcal{P}(m, m + n - 1)$, $\mathcal{P}(m - 1, m + n - 3)$, …, $\mathcal{P}(1, n - m + 1)$ in the case $m \leq$ n and $\mathcal{P}(m, m + n - 1)$, $\mathcal{P}(m - 1, m + n - 3)$, …, $\mathcal{P}(m - n + 1, m - n + 1)$ in the case $m > n$, where $\mathcal{P}(k, \ell)$ means the 1-D generalized synchronization problem for ℓ cells with a general on the kth cell from left end. Beyer [3] and Shinahr [44] have shown that an optimum-time complexity for syn-

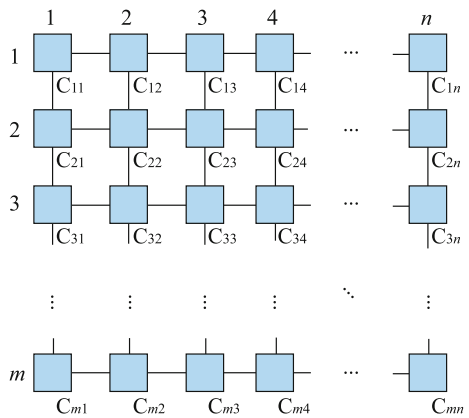

Firing Squad Synchronization Problem in Cellular Automata, Figure 26
A two-dimensional cellular automaton

chronizing any $m \times n$ arrays is $m + n + \max(m, n) - 3$ steps. Shinahr [44] has also given a 28-state implementation.

Theorem 31 ([3,44]) *There exists an optimum-time 28-state protocol for synchronizing any $m \times n$ rectangular arrays in $m + n + \max(m, n) - 3$ steps.*

Diagonal Mapping I: Six-State Linear-Time Algorithm

The proposal is a simple and state-efficient mapping scheme that enables us to embed any 1-D firing squad synchronization algorithm with a general at one end onto two-dimensional arrays without introducing additional states. Consider a 2-D array of size $m \times n$, where $m, n \geq 2$. Firstly, divide mn cells on the array into $m + n - 1$ groups g_k, $1 \leq k \leq m + n - 1$, defined as follows.

$$g_k = \{C_{i,j} | (i-1) + (j-1) = k - 1\}, \quad \text{i. e.,}$$
$$g_1 = \{C_{1,1}\}, \quad g_2 = \{C_{1,2}, C_{2,1}\},$$
$$g_3 = \{C_{1,3}, C_{2,2}, C_{3,1}\}, \ \ldots, \ g_{m+n-1} = \{C_{m,n}\}.$$

Figure 29 shows the division of the two-dimensional array of size $m \times n$ into $m + n - 1$ groups. Let M be any 1-D CA that synchronizes ℓ cells in $T(\ell)$ steps. Assume that M has $m + n - 1$ cells, denoted by C_i, where $1 \leq i \leq m + n - 1$. Then, consider a one-to-one correspondence between the ith group g_i and the ith cell C_i on M such that $g_i \leftrightarrow C_i$, where $1 \leq i \leq m + n - 1$. One can construct a 2-D CA N such that all cells in g_i simulate the ith cell C_i in real-time and N can synchronize any $m \times n$ arrays at time $t = T(m + n - 1)$ if and only if M synchronizes 1-D arrays of length $m + n - 1$ at time $t = T(m + n - 1)$. It is noted that the set of internal states of N constructed is the same as M. Thus an $m \times n$ 2-D array synchronization problem is reduced to one 1-D synchronization problem with the general at the left end. The algorithm obtained is slightly slower than the optimum ones, but the number of internal states is considerably smaller. Figure 30 shows snapshots of the proposed 6-state linear-time firing squad synchronization algorithm on rectangular arrays. For the details of the construction of the transition rule set, see Umeo, Maeda, Hisaoka and Teraoka [60].

Theorem 32 ([60]) *Let A be any s-state firing synchronization algorithm operating in $T(\ell)$ steps on 1-D ℓ cells. Then, there exists a 2-D s-state cellular automaton that can synchronize any $m \times n$ rectangular array in $T(m + n - 1)$ steps.*

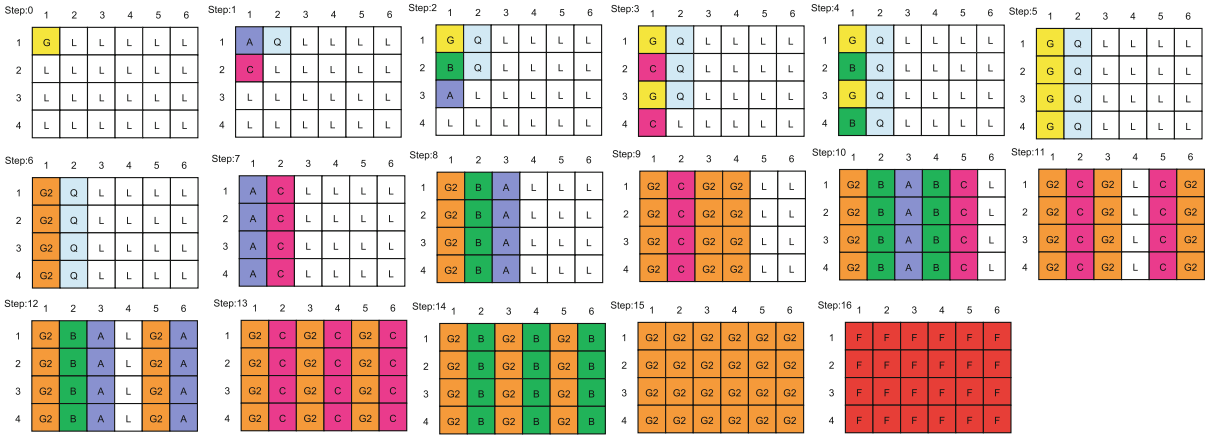

Firing Squad Synchronization Problem in Cellular Automata, Figure 27
Snapshots of the synchronization process on 4 × 6 array

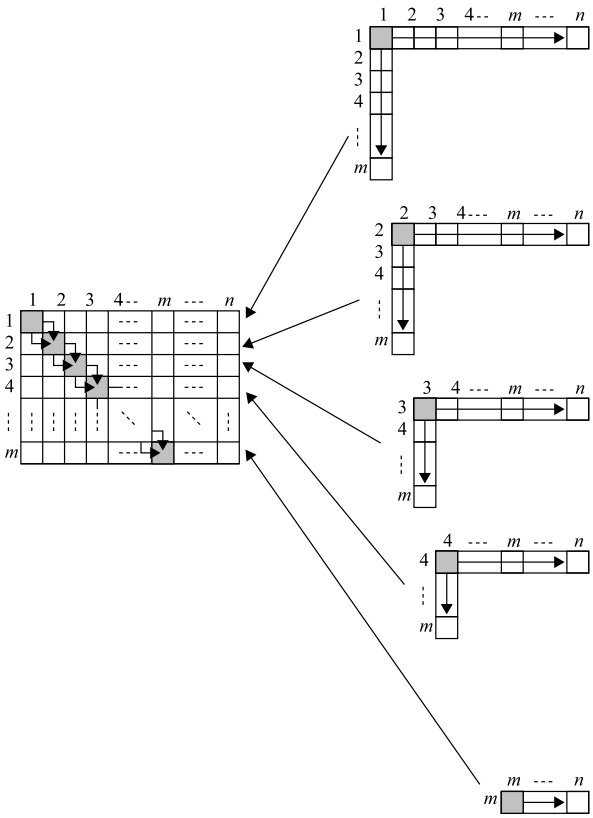

Firing Squad Synchronization Problem in Cellular Automata, Figure 28
An optimum-time synchronization scheme for rectangular arrays

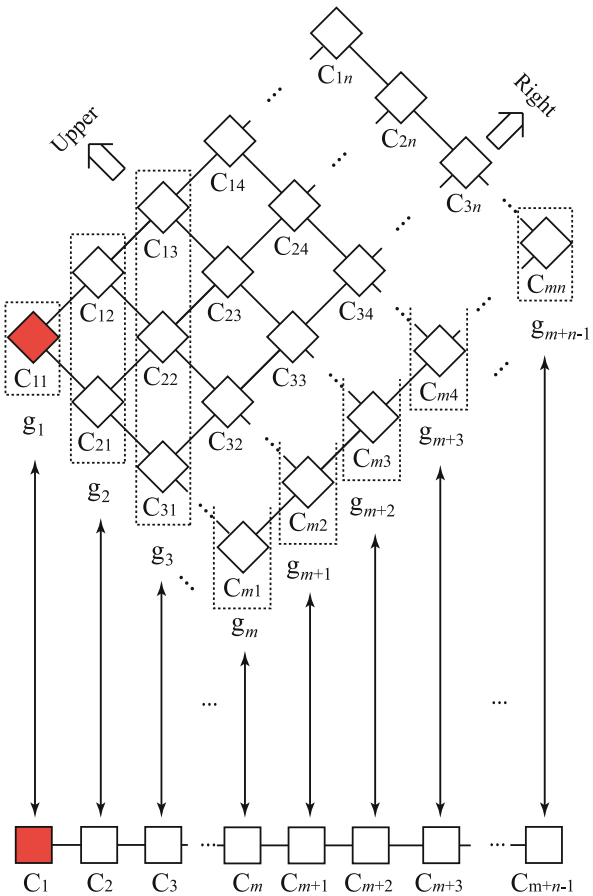

Firing Squad Synchronization Problem in Cellular Automata, Figure 29
A correspondence between 1-D and 2-D arrays

Theorem 33 ([60]) *There exists a 6-state 2-D CA that can synchronize any $m \times n$ rectangular array in $2(m+n)-4$ steps.*

Theorem 34 ([60]) *There exists a 6-state 2-D CA that can synchronize any $m \times n$ rectangular array containing isolated rectangular holes in $2(m+n)-4$ steps.*

Theorem 35 ([60]) *There exists a 6-state firing squad synchronization algorithm that can synchronize any 3-D $m \times n \times \ell$ solid arrays in $2(m+n+\ell)-6$ steps.*

Diagonal Mapping II: Twelve-State Time-Optimum Algorithm

The second diagonal mapping scheme in this section enables us to embed a special class of 1-D generalized synchronization algorithm onto two-dimensional arrays without introducing additional states. An $m \times n$ 2-D array synchronization problem is reduced to one 1-D generalized synchronization problem: $\mathcal{P}(m, m+n-1)$. Divide mn cells into $m+n-1$ groups g_k defined as follows, where k is any integer such that $-(m-1) \le k \le n-1$:

$$g_k = \{C_{i,j} | j - i = k\}, \quad -(m-1) \le k \le n-1 .$$

Figure 31 shows the correspondence between 1-D and 2-D arrays.

Property $\mathcal{A}$: Let S_i^t denote the state of C_i at step t. It is said that a generalized firing algorithm has a *property* $\mathcal{A}$, where any state S_i^t appearing *in the area* $\mathcal{A}$ can be computed from its left and right neighbor states S_{i-1}^{t-1} and S_{i+1}^{t-1} but it never depends on its own previous state S_i^{t-1}. Figure 32 shows the area $\mathcal{A}$ in the time-space diagram for the generalized optimum-step firing squad synchronization algorithm.

Any one-dimensional generalized firing squad synchronization algorithm with the property $\mathcal{A}$ can be easily embedded onto two-dimensional arrays without introducing additional states.

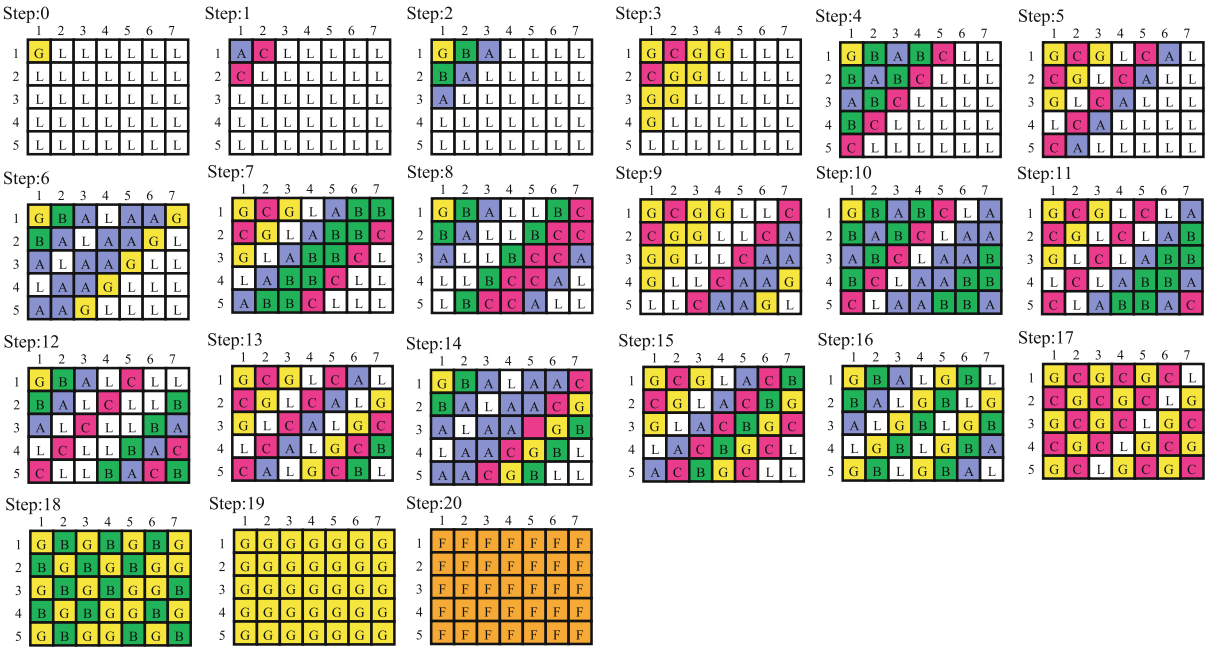

Firing Squad Synchronization Problem in Cellular Automata, Figure 30
Snapshots of the proposed 6-state linear-time firing squad synchronization algorithm on rectangular arrays

Theorem 36 ([53]) *Let M be any s-state generalized synchronization algorithm with the property $\mathcal{A}$ operating in $T(k, \ell)$ steps on 1-D ℓ cells with a general on the kth cell from the left end. Then, based on M, one can construct a 2-D s-state cellular automaton that can synchronize any $m \times n$ rectangular array in $T(m, m + n - 1)$ steps.*

It has been shown in Umeo, Hisaoka and Akiguchi [53] that there exists a 12-state implementation of the generalized optimum-time synchronization algorithms having the property $\mathcal{A}$. Then, one can get a 12-state optimum-time synchronization algorithm for rectangular arrays. Figure 33 shows snapshots of the proposed 12-state optimum-time firing squad synchronization algorithm operating on a 7×9 array.

Theorem 37 ([53]) *There exists a 12-state firing squad synchronization algorithm that can synchronize any $m \times n$ rectangular array in optimum $m + n + \max(m, n) - 3$ steps.*

Rotated L-Shaped Mapping: Time-Optimum Algorithm

In this section we present an optimum-time synchronization algorithm based on a rotated L-shaped mapping. The synchronization scheme is quite different from previous designs. The scheme uses the freezing-thawing technique. Without loss of generality, it is assumed that $m \leq$ n.

A rectangular array of size $m \times n$ is regarded as m *rotated* (90° in counterclockwise direction) *L-shaped* 1-D arrays. Each L-shaped array is denoted by $\mathrm{L}_i, 1 \leq i \leq m$. See Fig 34. Each L_i consists of three segments of length i, $n - m$, and i, respectively. Each segment can be synchronized by the freezing-thawing technique. Synchronization operations for $\mathrm{L}_i, 1 \leq i \leq m$ are as follows: Figure 35 shows a time-space diagram for synchronizing L_i. The wake-up signals for the three segments of L_i are generated at time $t = m + 2(m - i) - 1, 3m - i - 2$, and $n + 2(m - i) - 1$, respectively. Synchronization operations on each segments are delayed for $\Delta t_{i_j}, 1 \leq j \leq 3$ such that:

$$\Delta t_{i_j} = \begin{cases} 2(n - m) & j = 1 \\ i & j = 2 \\ n - m & j = 3 \end{cases} . \tag{1}$$

The synchronization for the first segment of L_i is started at time $t = m + 2(m - i) - 1$ and its operations are delayed for $\Delta t = \Delta t_{i_1} = 2(n - m)$ steps. Now letting $t_0 = m + 2(m - i) - 1, \Delta t = \Delta t_{i_1} = 2(n - m)$ in freezing-thawing technique, the first segment of L_i can be synchronized at time $t = t_0 + 2i - 2 + \Delta t = m + 2n - 3$. In a similar way, the second and the third segments can be synchronized at time $t = m + 2n - 3$. Thus, L_i can be synchronized at time $t = m + 2n - 3$. Figure 36 shows some snapshots of the synchronization process operating

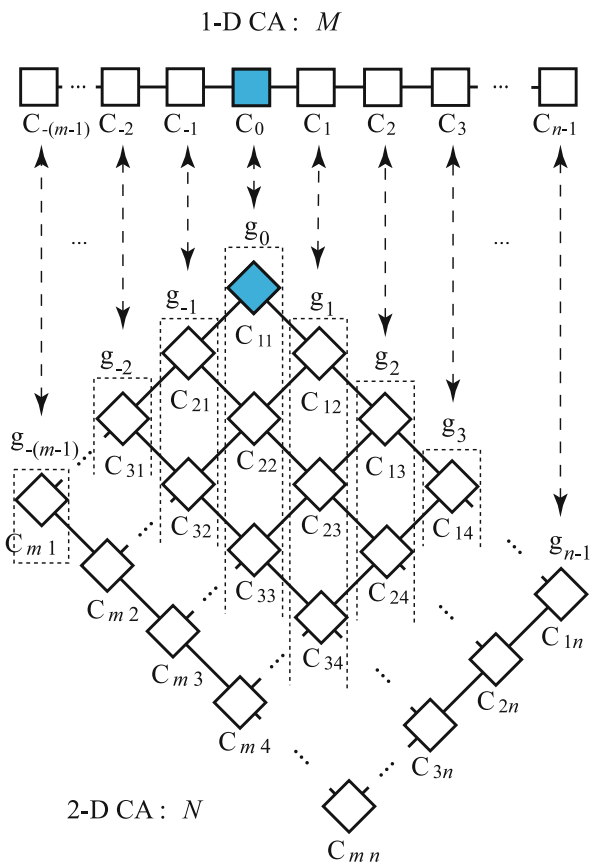

Firing Squad Synchronization Problem in Cellular Automata, Figure 31
Correspondence between 1-D and 2-D cellular arrays

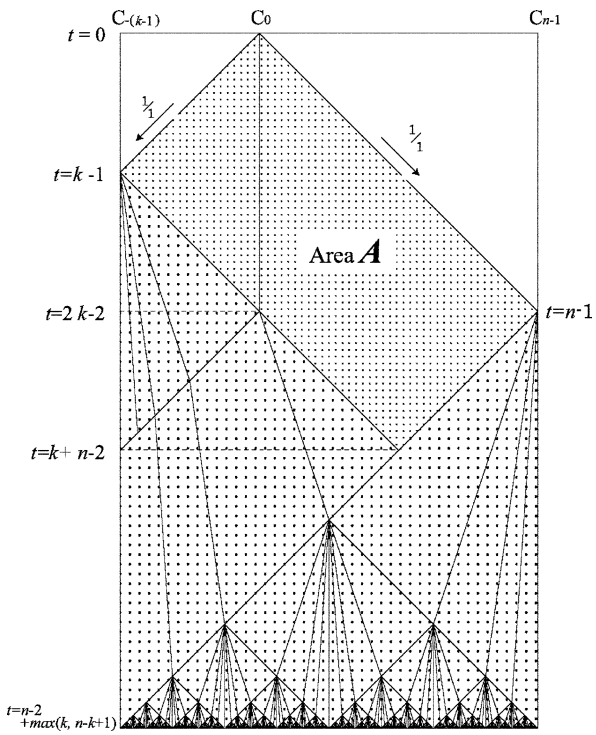

Firing Squad Synchronization Problem in Cellular Automata, Figure 32
Time-space diagram for generalized optimum-step firing squad synchronization algorithm

in optimum-steps on 5×8 array. Now the next theorem can be established.

Theorem 38 ([65]) *The algorithm above can synchronize any $m \times n$ rectangular array in optimum $m + n + \max(m, n) - 3$ steps.*

Frame Mapping: Time-Optimum Algorithm

Section "Frame Mapping: Time-Optimum Algorithm" presents an optimum-time 2-D synchronization algorithm based on frame mapping. Without loss of generality, it is assumed that $m \leq$ n. A rectangular array of size $m \times n$ is regarded as consisting of rectangle-shaped *frames* of width 1. See Fig. 37. Each frame $L_i, 0 \leq i \leq \lceil m/2 \rceil - 1$, is divided into six segments and these six segments are synchronized using the freezing-thawing technique. The length of each segment of L_i is $m - 2i$, $m - 2i$, $n - m$, $m - 2i$, $m - 2i$, and $n - m$, respectively. Figure 38 shows a time-space diagram of the synchronization operations for the outermost frame L_0. Synchronization operations on jth segment of L_0 are delayed for Δt_{0_j} steps, $1 \leq j \leq 6$, such that:

$$\Delta t_{0_j} = \begin{cases} 2(n-m) & j = 1 \\ 2(n-m) & j = 2 \\ m & j = 3 \\ n-m & j = 4 \\ n-m & j = 5 \\ m & j = 6 \end{cases} . \tag{2}$$

Using the freezing-thawing technique, L_0 can be synchronized at time $t = m + 2n - 3$. The synchronization operation for L_i, $i \geq 1$ can be done similarly. Note that the ith frame is of size $(m - 2i) \times (n - 2i)$. Let T_i be steps required for synchronizing the ith frame with the synchronization operations given above starting at time $t = 0$. Then, $T_i = m + 2n - 3 - 6i = T_0 - 6i$, for any i such that $0 \leq i \leq \lceil m/2 \rceil - 1$. Thus, $T_i - T_{i-1} = 6$. Therefore, the starting time for synchronizing each frame is delayed for 6 steps so that synchronization operations for each frame can be finished simultaneously. In order to start the

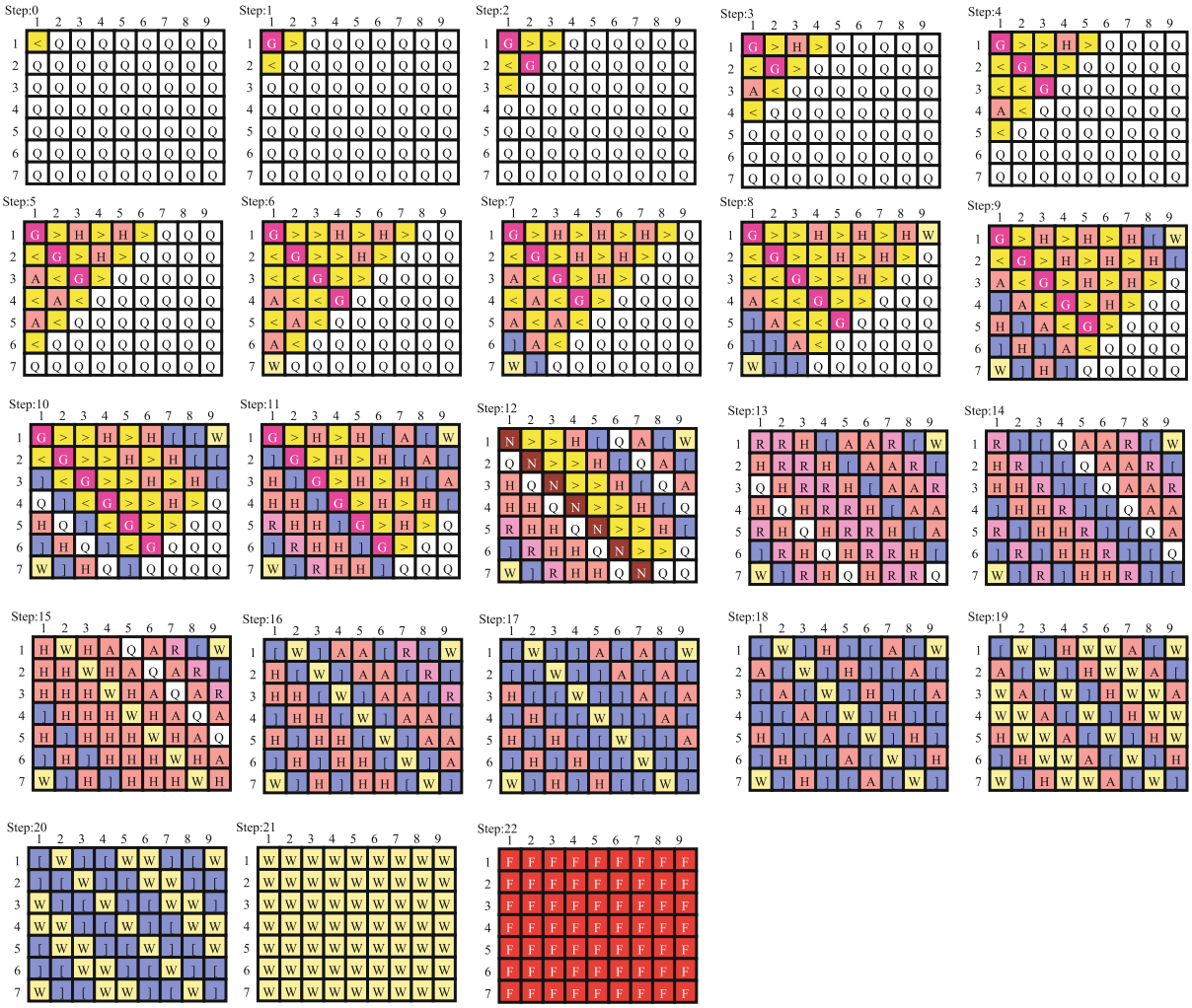

Firing Squad Synchronization Problem in Cellular Automata, Figure 33
Snapshots of the proposed 12-state optimum-time firing squad synchronization algorithm on rectangular arrays

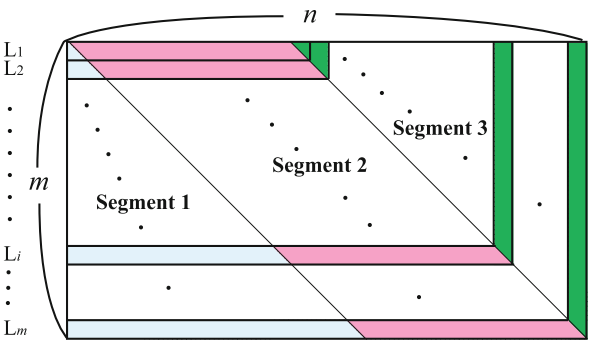

Firing Squad Synchronization Problem in Cellular Automata, Figure 34
A two-dimensional array of size $m \times n$ is regarded as consisting of m rotated (90° in counterclockwise direction) L-shaped 1-D array

operation progressively, an activating signal that travels in the diagonal direction is given to each north-west corner of each L_i at time $t = 6i$. In this way all of the frames can be synchronized. Figure 39 illustrates some snapshots of the synchronization process operating in optimum-steps on 5×8 array.

Theorem 39 ([66]) *The algorithm based on frame mapping can synchronize any $m \times n$ rectangular array in $m + n + \max(m, n) - 3$ optimum steps.*

Generalized Firing Squad Synchronization Algorithms for Two-Dimensional Rectangular Cellular Automata

Szwerinski [45] presented a first optimum-time generalized firing squad synchronization algorithm for rectangular arrays. Umeo, Hisaoka, Teraoka and Maeda [55]

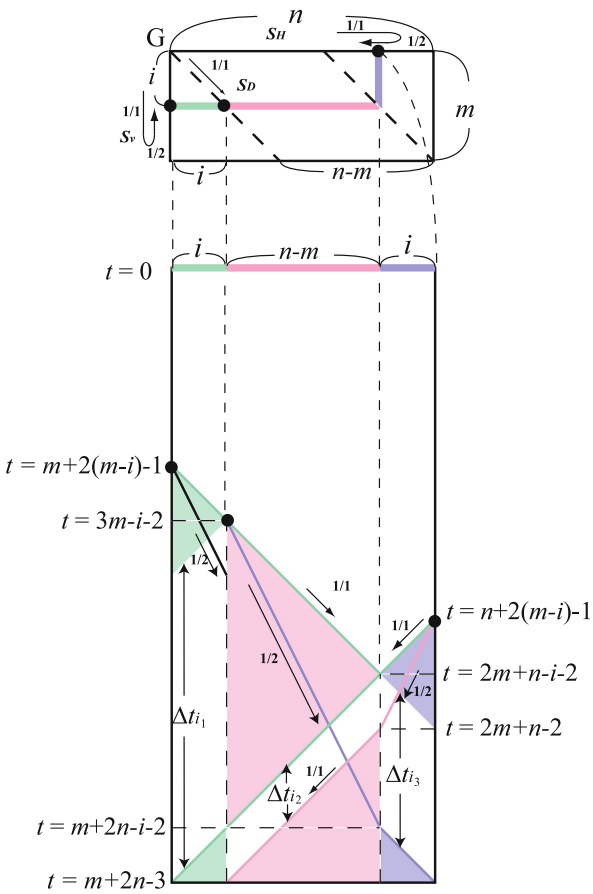

Firing Squad Synchronization Problem in Cellular Automata, Figure 35
Time-space diagram for synchronizing L_i

proposed a new optimum-time generalized synchronization algorithm. Section "Generalized Firing Squad Synchronization Algorithms for Two-Dimensional Rectangular Cellular Automata" presents the generalized optimum-time algorithm based on Umeo et al. [55]. The algorithm can synchronize any rectangular arrays of size $m \times n$ with the general being located at any position (r, s) on the array in $m + n + \max(m, n) - \min(r, m - r + 1) - \min(s, n - s + 1) - 1$ optimum steps, where $1 \le r \le m, 1 \le s \le n$. The following theorem is a useful technique for delaying the generalized synchronization on 1-D arrays.

Theorem 40 ([55]) *Let A be any 1-D cellular automaton that runs a generalized $T(s, n)$-step synchronization algorithm on n cells with a general on $C_s (1 \le s \le n)$ and t_0, t_1, t_2, ℓ be any integer satisfying the following conditions such that $\Delta t = t_2 - t_1 = 2\ell$, $\ell \ge 1$, $t_2 > t_1 \ge t_0 \ge 0$ and $t_1 + t_2 - 2t_0 \le 2T(s, n) - 2\max(s, n - s + 1) + 2$. It is assumed that three special signals are given to cell C_s at step $t = t_0$, t_1, and t_2. These signals play an important role of initializing the generalized synchronization process, starting the delayed operation, and stopping the delayed operation, respectively. Then, one can construct a cellular automaton B that synchronizes the array at time $t = t_0 + \ell + T(s, n)$. In the case where $T(s, n)$ is an optimum time complexity such that $T(s, n) = n - 2 + \max(s, n - s + 1)$, the constructed B synchronizes at time $t = t_0 + \ell + n - 2 + \max(s, n - s + 1)$. Thus the synchronization operation is delayed for ℓ steps. Figure 40 is a time-space diagram for the delayed optimum-time generalized synchronization algorithm operating on n cells. In the darker area of the diagram, each cell simulates the operation of A at speed 1/2 by repeating a simulate-one-step of A then keep-the-state operations alternatively at each step.*

Consider the case where $m \le n$, $1 \le r < \lceil m/2 \rceil$ and $1 \le s < \lceil n/2 \rceil$. An array of size $m \times n$ is regarded as consisting of m rows of length n. An optimum-time generalized synchronization algorithm with a general at $C_{i,s}$ $(1 \le i \le m)$ is used for the synchronization of the ith row. The operation is referred to as *row-synchronization*. To synchronize all rows simultaneously, an efficient timing control scheme shown in Fig. 41 is developed. Figure 41 is a time-space diagram for giving special signals to each cell on the s-column. These special signals act as a timing $t = t_0$, t_1 and t_2 stated in Theorem 40. For example, the row-synchronization on the yth $(r \le y \le m)$ row is started at time $t = t_0 = t_1 = y - r$ and the process is delayed from time $t = t_1 = y - r$ to $t = t_2 = 2m - y - r$, shown in the darker area in Fig. 41. Thus $\ell = m - y$. Based on Theorem 40 the yth row is synchronized at time $t = m + 2n - r - s - 1$. Figure 42 is a time-space diagram for the row-synchronization on the xth and yth row, where $1 \le x < r$ and $r \le y \le m$. The readers can see how all rows are synchronized at time $t = m + 2n - r - s - 1$.

Thus, the following theorem can be developed.

Theorem 41 ([55]) *In the row-synchronization process, all of the rows can be synchronized simultaneously at time $t = m + 2n - r - s - 1$ in the case $m \le n$, $1 \le r < \lceil m/2 \rceil$ and $1 \le s < \lceil n/2 \rceil$.*

In the column-synchronization process, all of the cells take a firing state prior to the row-synchronization, but the column-synchronization fails to synchronize. Symmetrically, the following theorem holds in the case where the array is longer than is wide.

Theorem 42 ([55]) *In the column-synchronization process, all of the columns can be synchronized simultane-*

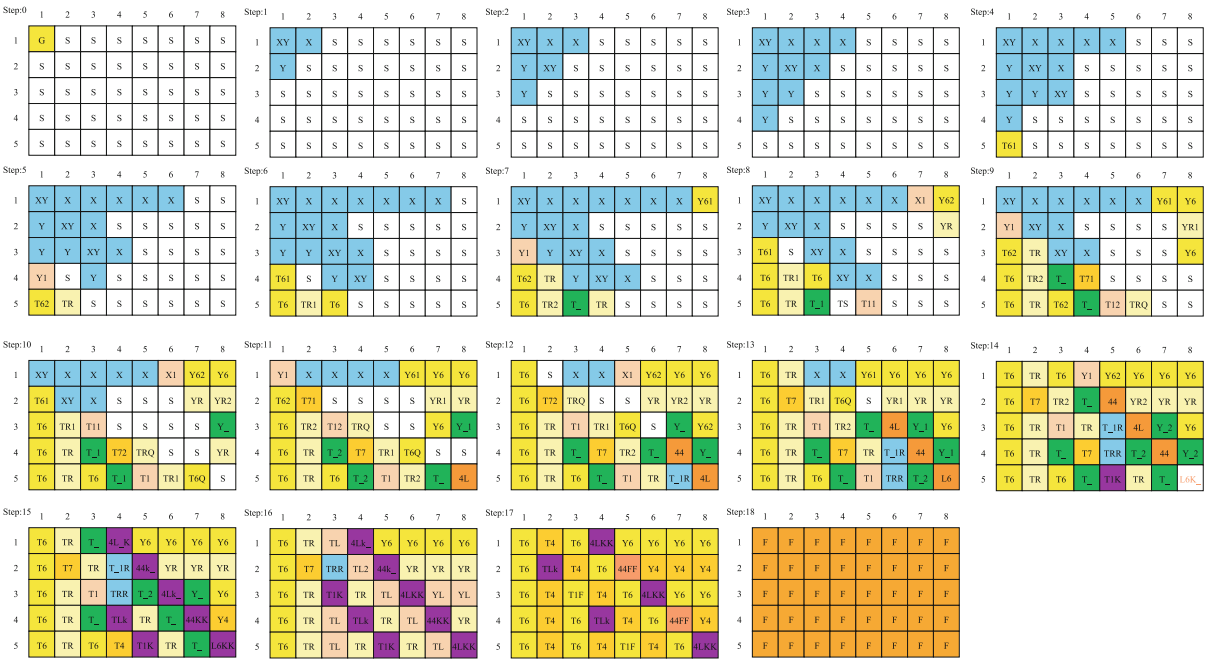

Firing Squad Synchronization Problem in Cellular Automata, Figure 36
Snapshots of the synchronization process on 5 × 8 array

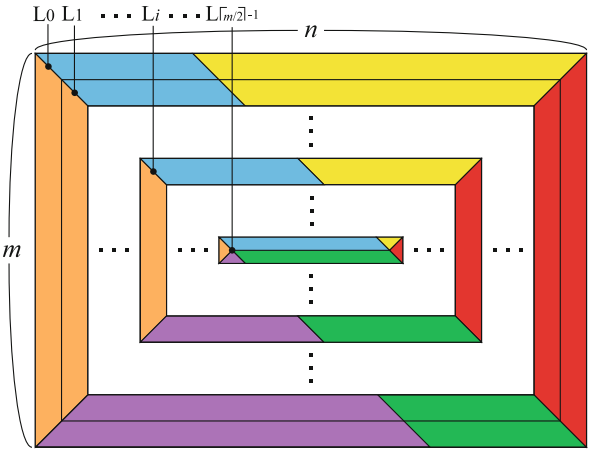

Firing Squad Synchronization Problem in Cellular Automata, Figure 37
A two-dimensional array of size $m \times n$ is regarded as consisting of $\lceil m/2 \rceil$ frames

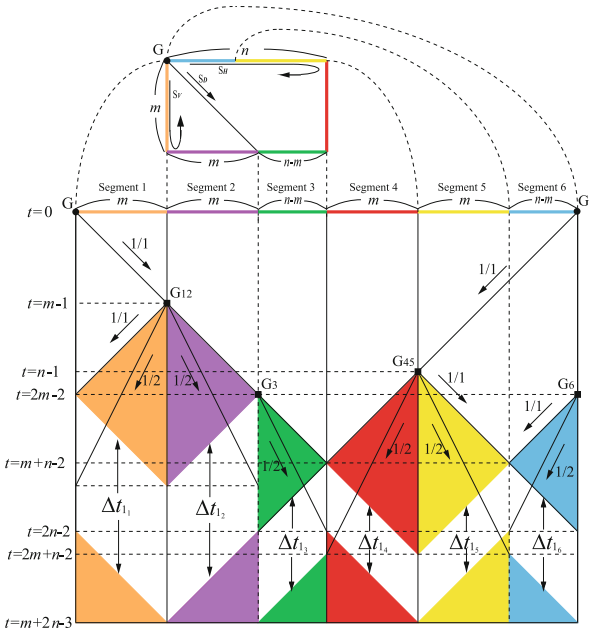

Firing Squad Synchronization Problem in Cellular Automata, Figure 38
Time-space diagram for synchronizing L_0

ously at time $t = n + 2m - r - s - 1$ *in the case* $m \geq n$, $1 \leq r < \lceil m/2 \rceil$ *and* $1 \leq s < \lceil n/2 \rceil$.

To synchronize the array in optimum-time, the array performs both row- and column-synchronization operations. Each cell should take a firing state on each layer at two different times. The first one is false and it should be ignored. The second one is the right firing state. By combining the Theorems 41, 42, the following theorem can be established.

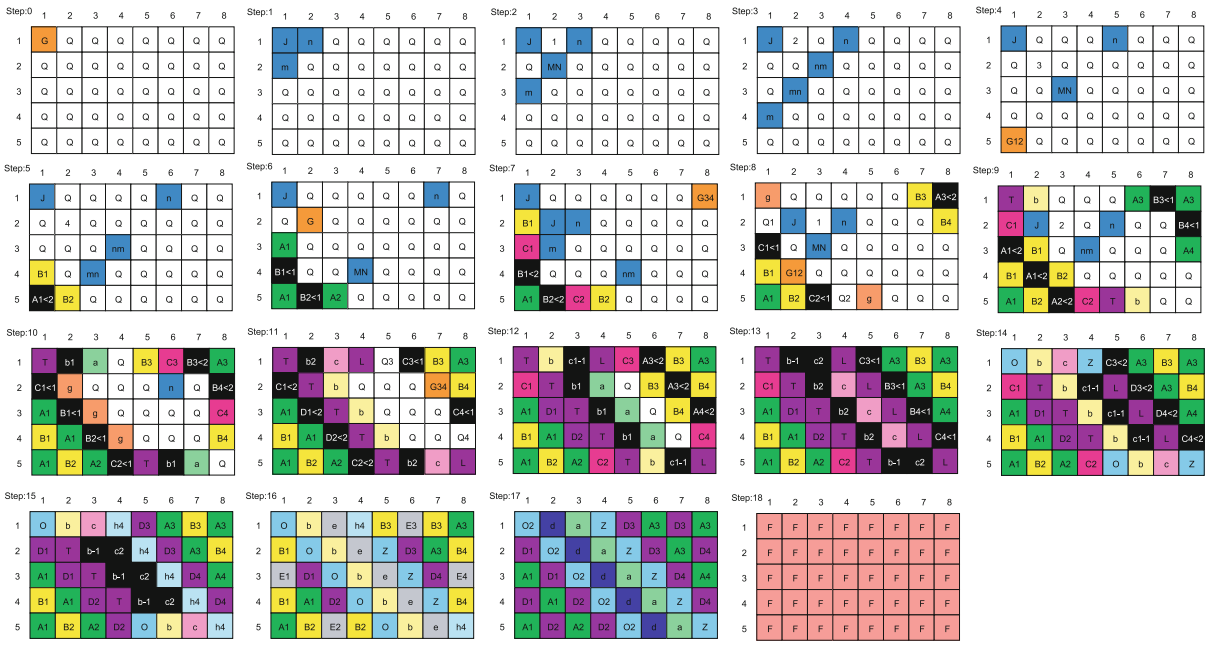

Firing Squad Synchronization Problem in Cellular Automata, Figure 39
Snapshots of the synchronization process on 5 × 8 array

Theorem 43 ([55]) *The scheme given above can synchronize any $m \times n$ array in $m + n + \max(m, n) - \min(r, m - r+1) - \min(s, n-s+1) - 1$ optimum steps, where (r, s) is the general's initial position such that $1 \le r \le m, 1 \le s \le n$.*

The number of internal states of an automaton realizing the Szwerinski's algorithm is 25,600. Umeo, Maeda, Hisaoka and Teraoka [60] presented a 14-state implementation for a non-optimum-time generalized algorithm. The 2-D generalized synchronization algorithm is $\max(r+s, m+n-r-s+2) - \max(m, n) + \min(r, m-r+1) + \min(s, n-s+1) - 3$ steps larger than the optimum algorithm proposed by Szwerinski [45]. However, the number of internal states required to yield the synchronizing condition is the smallest known at present. Snapshots of the 14-state generalized synchronization algorithm running on a rectangular array of size 6×8 with the general at $C_{2,3}$ are shown in Fig. 43.

Theorem 44 ([60]) *There exists a 14-state 2-D CA that can synchronize any $m \times n$ rectangular array in $m + n + \max(r + s, m + n - r - s + 2) - 4$ steps with the general at an arbitrary initial position (r, s).*

Firing Squad Synchronization Algorithms for Two-Dimensional Square Cellular Automata

An optimum-time square synchronization algorithm has been proposed by Shinahr [44]. The algorithm operates as follows: By dividing the entire square array into n *rotated L-shaped* 1-D arrays such that the length of the ith L is $2n - 2i + 1$ $(1 \le i \le n)$, one treats the square synchronization as n independent 1-D synchronizations with the general located at the center cell. On the ith L, a general is generated at $C_{i,i}$ at time $t = 2i - 2$, and the general initiates the horizontal and vertical synchronizations on the row and column arrays via an optimum-time synchronization algorithm. The array can be synchronized in optimum time $t = 2i - 2 + 2(n - i + 1) - 2 = 2n - 2$. It has been shown in Umeo, Maeda and Fujiwara [59] that 9 states are sufficient for the optimum-time square synchronization. The implementation is based on Mazoyer's 6-state algorithm. Figure 44 shows snapshots of configurations of the 9-state synchronization algorithm running on a square of size 8×8.

Theorem 45 ([44,59]) *There exists a 9-state 2-D CA that can synchronize any $n \times n$ square array in $2n - 2$ steps.*

Firing Squad Synchronization Algorithms for Two-Dimensional One-Bit Communication Cellular Automata

The firing squad synchronization problem for 2-D one-bit communication cellular automata has been studied by Torre, Napoli and Parente [49], Umeo, Michisaka and Kamikawa [62], Gruska, Torre, and Parente [17],

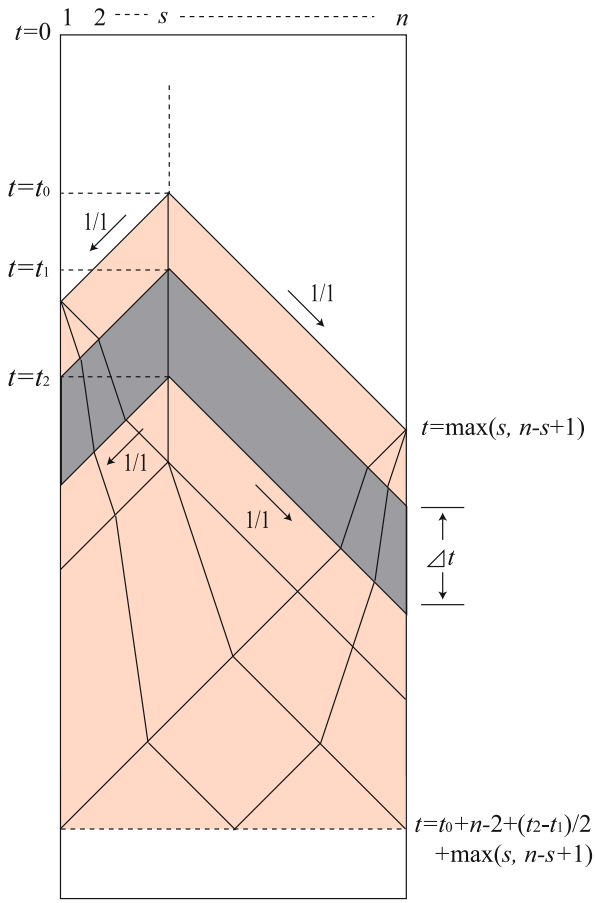

Firing Squad Synchronization Problem in Cellular Automata, Figure 40
Time-space diagram for delayed optimum-time generalized synchronization algorithm

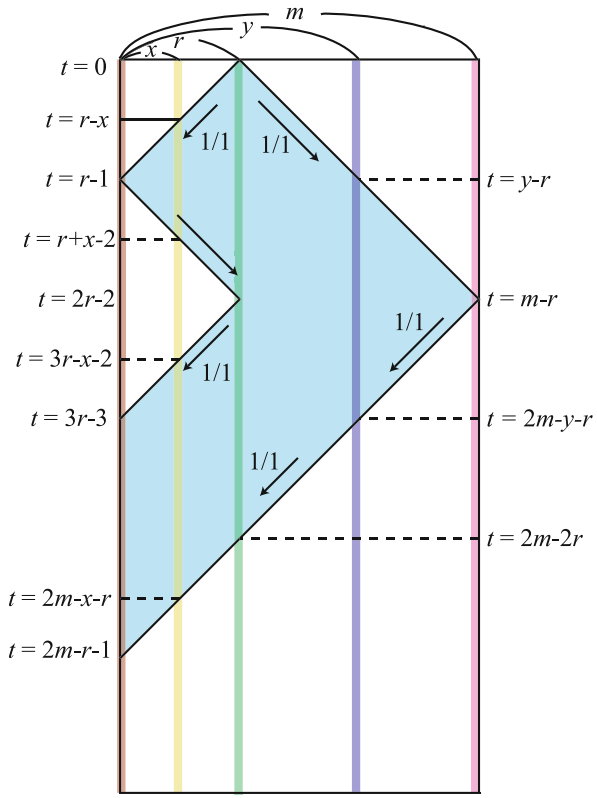

Firing Squad Synchronization Problem in Cellular Automata, Figure 41
Time-space diagram for delaying row synchronization

and Umeo, Michisaka, Kamikawa and Kanazawa [63]. Section "Firing Squad Synchronization Algorithms for Two-Dimensional One-Bit Communication Cellular Automata" presents two implementations of the square and rectangular synchronization algorithms on CA_{1-bit}. The first one is for square arrays given in Umeo, Michisaka, Kamikawa and Kanazawa [63]. It runs in $(2n-1)$ steps on $n \times n$ square arrays. The proposed implementation is one step slower than optimum-time for the O(1)-bit communication model. The total numbers of internal states and transition rules of the CA_{1-bit} are 127 and 405, respectively. Figure 45 shows snapshots of configurations of the 127-state implementation running on a square of size 8×8. Gruska, Torre, and Parente [17] presented an optimum-time algorithm.

Theorem 46 ([17]) *There exists a 2-D CA_{1-bit} that can synchronize any $n \times n$ square arrays in $2n-2$ steps.*

Umeo, Michisaka, Kamikawa and Kanazawa [63] has also implemented the rectangular synchronization algorithm for 2-D CA_{1-bit}. The total numbers of internal states and transition rules of the CA_{1-bit} are 862 and 2217, respectively. Figure 46 shows snapshots of the synchronization process on a 5×8 rectangular array.

Theorem 47 ([63]) *There exists a 2-D CA_{1-bit} that can synchronize any $m \times n$ rectangular arrays in $m+n+\max(m,n)$ steps.*

Summary and Future Directions

The present article has examined via computer the state transition rule sets for which optimum-time synchronization algorithms were designed over the past forty years. The smallest transition rule sets for the well-known firing squad synchronization algorithms are useful and important for researchers who might have interests in those transition rule sets that realize the classical optimum-time firing algorithms quoted frequently in the literatures. It has also presented a survey and a comparison of the quantitative and qualitative aspects of the optimum-time

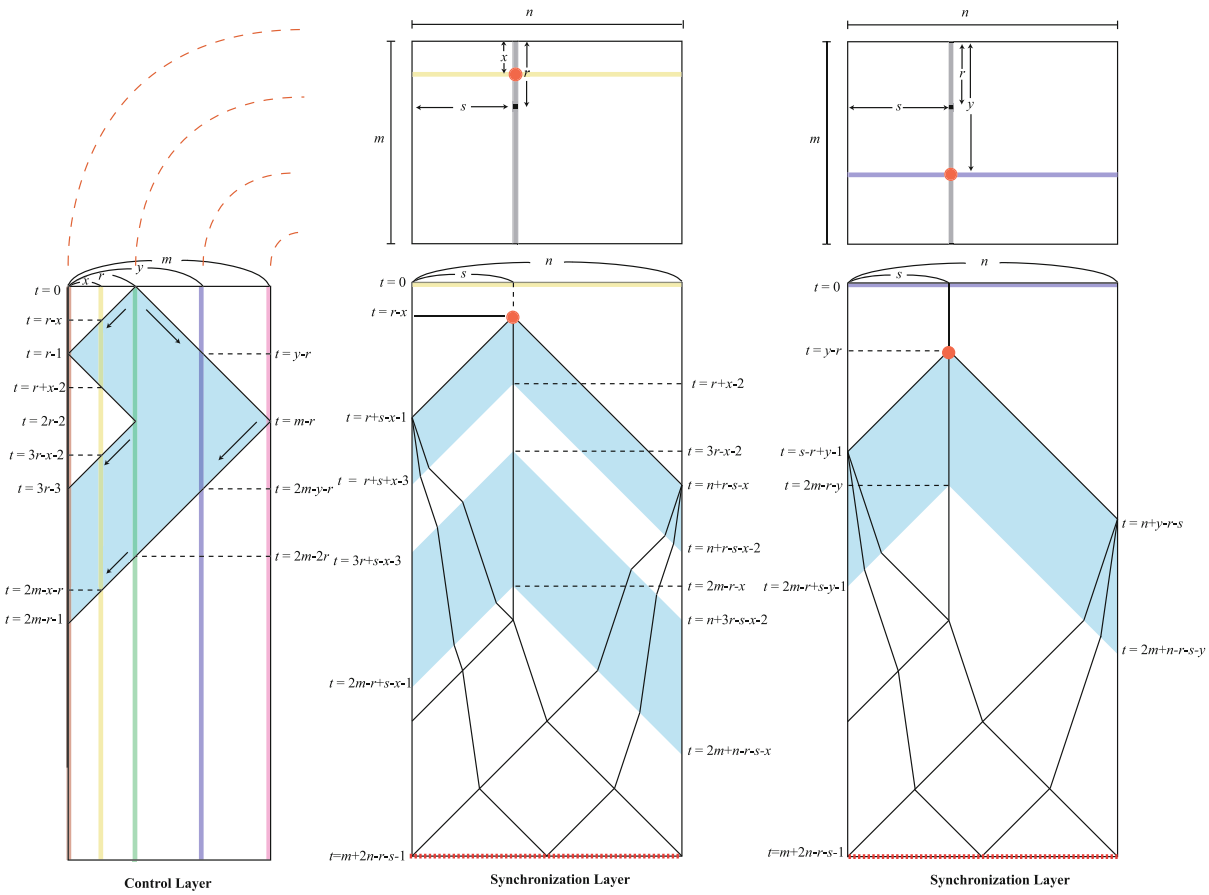

Firing Squad Synchronization Problem in Cellular Automata, Figure 42
Time-space diagram for the row-synchronization on the *x*th and *y*th row of a rectangular array of size $m \times n$, where $m \leq n$, $1 \leq r < \lceil m/2 \rceil$, $1 \leq s < \lceil n/2 \rceil$, $1 \leq x < r$ and $r \leq y \leq m$

synchronization algorithms developed thus far for one-dimensional cellular arrays. It has studied several variants of the firing squad synchronization problems including fault-tolerant synchronization protocols, 4- and 5-state partial solutions, one-bit communication protocols and non-optimum-time algorithms etc. Finally a survey on two-dimensional firing squad synchronization algorithms has been given. Several new results and new viewpoints have been provided. The question: "What is the minimum number of states for an optimum-time solution of the problem?" still remains open. A new approach should be required.

Bibliography

Primary Literature

1. Balzer R (1967) An 8-state minimal time solution to the firing squad synchronization problem. Inf Control 10:22–42
2. Berthiaume A, Bittner T, Perković L, Settle A, Simon J (2004) Bounding the firing synchronization problem on a ring. Theor Comput Sci 320:213–228
3. Beyer WT (1969) Recognition of topological invariants by iterative arrays. Ph.D. Thesis, MIT, pp 144
4. Burns JE, Lynch NA (1987) The Byzantine firing squad problem. In: Advances in Computing Research: Parallel and Distributed Computing. JAI press, 4:147–161
5. Coan BA, Dolev D, Dwork C, Stockmeyer L (1989) The distributed firing squad problem. SIAM J Comput 18(5):990–1012
6. Culik K II (1989) Variations of the firing squad synchronization problem and applications. Inf Process Lett 30:153–157
7. Culik K II, Dube S (1991) An efficient solution of the firing mob problem. Theor Comput Sci 91:57–69
8. Fischer PC (1965) Generation of primes by a one-dimensional real-time iterative array. J ACM 12(3):388–394
9. Gerken HD (1987) über Synchronisations-Probleme bei Zellularautomaten. Diplomarbeit, Institut für Theoretische Informatik, Technische Universität Braunschweig
10. Goldstein D, Kobayashi K (2004) On the complexity of network synchronization. In: Proc. of ISSAC 2004. LNCS, vol 3341. Springer, Berlin, pp 496–507

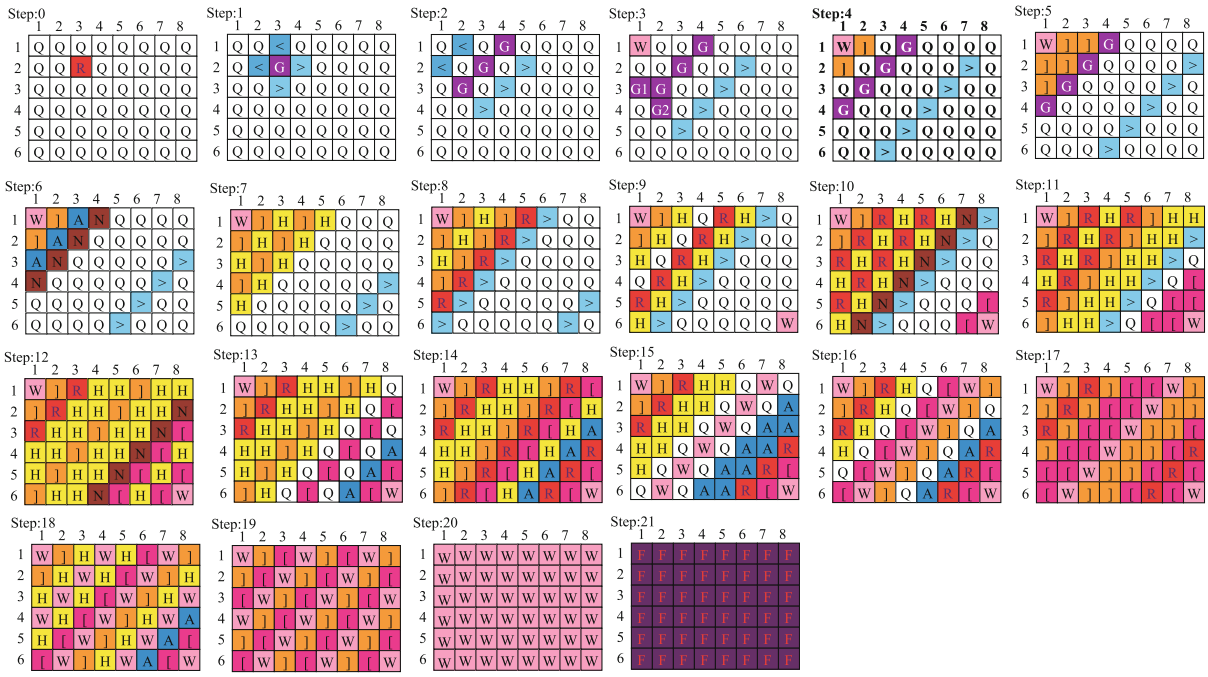

Firing Squad Synchronization Problem in Cellular Automata, Figure 43

Snapshots of the 14-state linear-time generalized firing squad synchronization algorithm on rectangular arrays

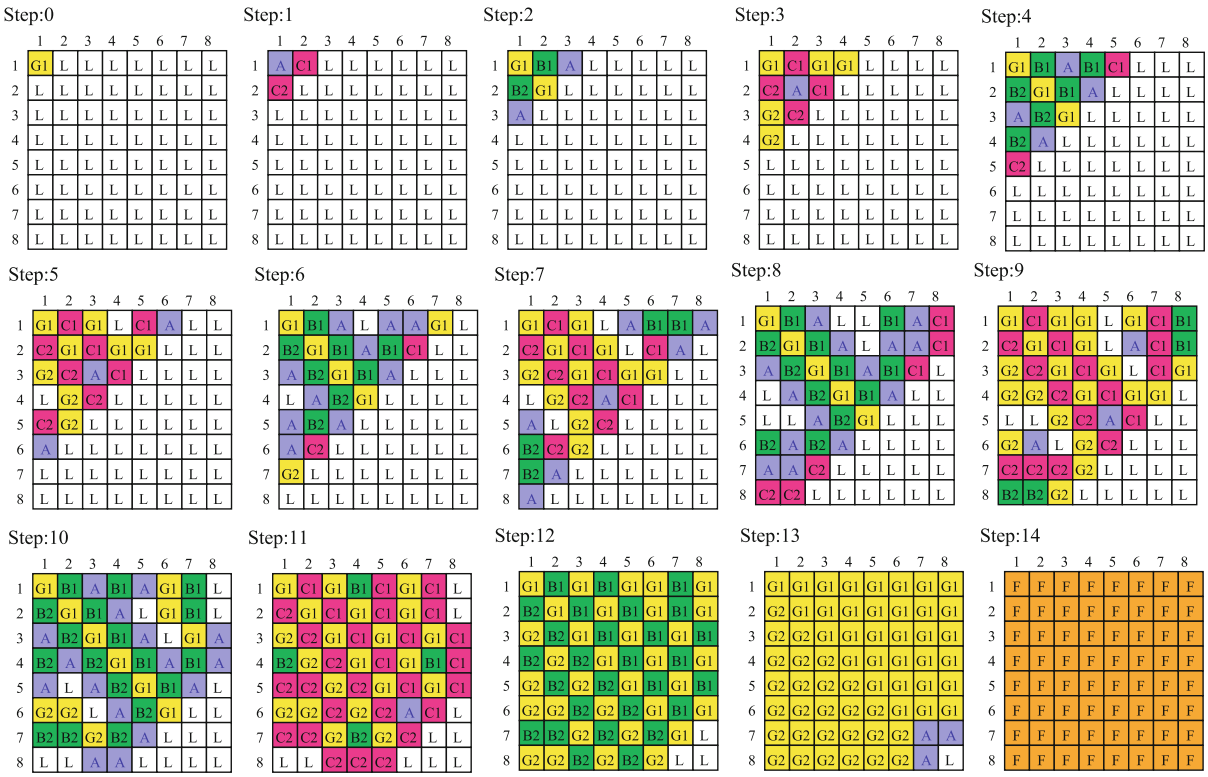

Firing Squad Synchronization Problem in Cellular Automata, Figure 44

A configuration of a 9-state implementation of optimum-time firing on square arrays

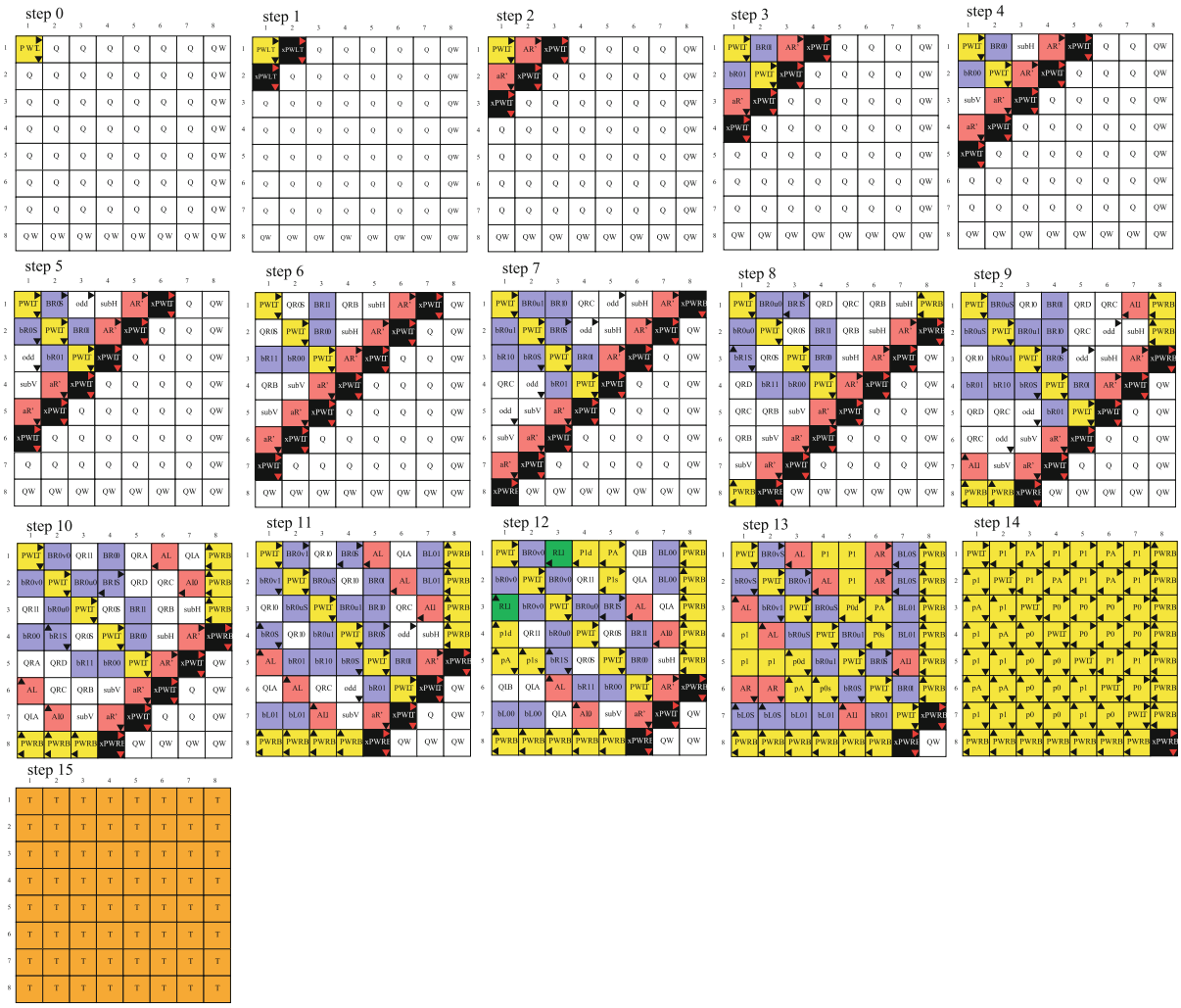

Firing Squad Synchronization Problem in Cellular Automata, Figure 45
Snapshots of the (2*n* — 1)-step square synchronization algorithm with the general on the northwest corner

11. Goldstein D, Meyer N (2002) The wake up and report problem is asymptotically time-equivalent to the firing squad synchronization problem. In: Proc. of 13th Annual ACM-SIAM Symp. on Discrete Algorithms, pp 578–587
12. Goto E (1962) A minimal time solution of the firing squad problem. Dittoed course notes for Applied Mathematics 298, Harvard University, pp 52–59, with an illustration in color
13. Goto E (1966) Some puzzles on automata. In: Kitagawa T (ed) Toward computer sciences. Kyouritsu, pp 67–91 (in Japanese)
14. Grasselli A (1975) Synchronization of cellular arrays: The firing squad problem in two dimensions. Inf Control 28:113–124
15. Grefenstette JJ (1983) Network structure abd the firing squad synchronization problem. J Comput Syst Sci 26:139–152
16. Grigorieff S (2006) Synchronization of a bounded degree graph of cellular automata with non uniform delays in time $\Delta\lfloor\log_m(\delta)\rfloor$. Theor Comput Sci 356(1):170–185
17. Gruska J, Torre SL, Parente M (2007) The firing squad synchronization problem on squares, toruses and rings. Int J Found Comput Sci, 18(3):637–654
18. Herman GT, Liu WH, Rowland S, Walker A (1974) Synchronization of growing cellular systems. Inf Control 25:103–122
19. Imai K, Morita K (1996) Firing squad synchronization problem in reversible cellular automata. Theor Comput Sci 165: 475–482
20. Jiang T (1992) The synchronization of nonuniform networks of finite automata. Inf Comput 97:234–261
21. Kobayashi K (1977) The firing squad synchronization problem for two-dimensional arrays. Inf Control 34:177–197
22. Kutrib M, Vollmar R (1991) Minimal time synchronization in restricted defective cellular automata. J Inf Process Cybern ELK 27:179–196
23. Kutrib M, Vollmar R (1995) The firing squad synchronization problem in defective cellular automata. IEICE Trans Inf Syst E78-D(7):895–900

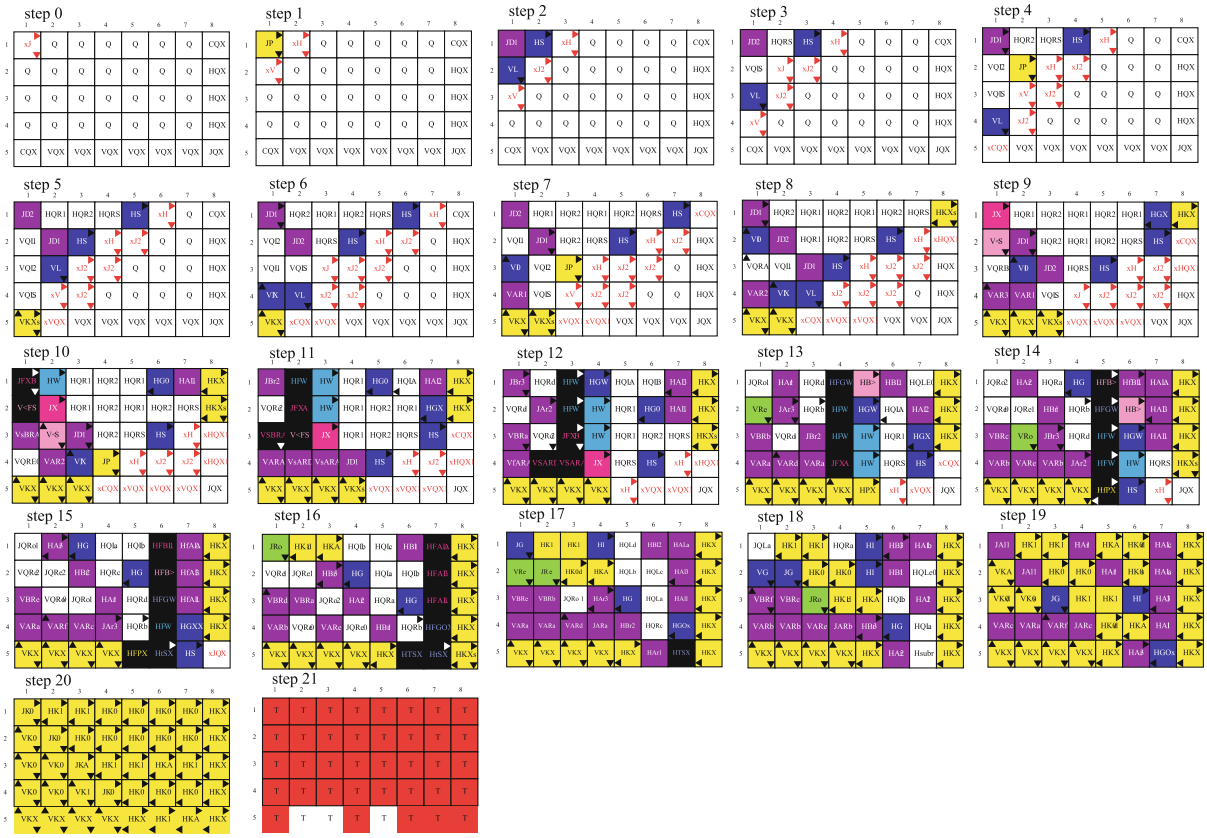

Firing Squad Synchronization Problem in Cellular Automata, Figure 46
Snapshots of the proposed rectangular firing squad synchronization algorithm with the general at the northwest corner

24. Mazoyer J (1986) An overview of the firing squad synchronization problem. Lecture Notes on Computer Science, vol 316. Springer, Berlin
25. Mazoyer J (1987) A six-state minimal time solution to the firing squad synchronization problem. Theor Comput Sci 50:183–238
26. Mazoyer J (1996) On optimal solutions to the firing squad synchronization problem. Theor Comput Sci 168:367–404
27. Mazoyer J (1997) A minimal-time solution to the FSSP without recursive call to itself and with bounded slope of signals. Draft, p 8
28. Minsky M (1967) Computation: Finite and infinite machines. Prentice Hall, Englewood Cliffs
29. Moore EF (1964) The firing squad synchronization problem. In: Moore EF (ed) Sequential Machines, Selected Papers, Addison-Wesley, Reading MA
30. Moore FR, Langdon GG (1968) A generalized firing squad problem. Inf Control 12:212–220
31. Nguyen HB, Hamacher VC (1974) Pattern synchronization in two-dimensional cellular space. Inf Control 26:12–23
32. Nishimura J, Umeo H (2005) An optimum-time synchronization protocol for 1-bit-communication cellular automaton. In: Proc. of the 9th World Multi-Conference on Systemics, Cybernetics and Informatics, vol X, pp 232–237
33. Nishitani Y, Honda N (1981) The firing squad synchronization problem. Theor Comput Sci 14:39–61
34. Noguchi K (2004) Simple 8-state minimal time solution to the firing squad synchronization problem. Theor Comput Sci 314:303–334
35. Roka Z (1995) The firing squad synchronization problem on Cayley graphs. In: Proc. of MFCS'95. LNCS, vol 969. Springer, Berlin, pp 402–411
36. Romani F (1976) Cellular automata synchronization. Inform Sci 10:299–318
37. Romani F (1977) On the fast synchronization of tree connected networks. Inform Sci, 12:229–244
38. Romani F (1978) The parallelism principle: speeding up the cellular automata synchronization. Inf Control 36:245–255
39. Rosenstiehl P, Fiksel JR, Holliger A (1972) Intelligent graphs: networks of finite automata capable of solving graph problems. In: Graph Theory and Computing, Academic Press, New York, pp 219–265
40. Sanders P (1994) Massively parallel search for transition-tables of polyautomata. In: Jesshope C, Jossifov V, Wilhelmi W (eds) Proc. of the VI International Workshop on Parallel Processing by Cellular Automata and Arrays. Akademie, Berlin, pp 99–108
41. Schmid H, Worsch T (2004) The firing squad synchronization

problem with many generals for one-dimensional. In: CA Proc. of IFIP World Congress, pp 111–124
42. Settle A, Simon J (1998) Improved bounds for the firing synchronization problem. In: SIROCCO 5: Proc. of the 5th International Colloquium on Structural Information and Communication Complexity. Carleton Scientific, Ontario, pp 66–81
43. Settle A, Simon J (2002) Smaller solutions for the firing squad. Theor Comput Sci 276:83–109
44. Shinahr I (1974) Two- and three-dimensional firing squad synchronization problems. Inf Control 24:163–180
45. Szwerinski H (1982) Time-optimum solution of the firing-squad-synchronization-problem for *n*-dimensional rectangles with the general at an arbitrary position. Theor Comput Sci 19:305–320
46. Torre SL, Napoli M, Parente D (1996) Synchronization of one-way connected processors. Complex Syst 10:239–255
47. Torre SL, Napoli M, Parente D (1998) Synchronization of a line of identical processors at a given time. Fundamenta Informaticae 34:103–128
48. Torre SL, Napoli M, Parente M (2000) A compositional approach to synchronize two dimensional networks of processors. Theor Inf Appl 34:549–564
49. Torre SL, Napoli M, Parente M (2001) Firing squad synchronization problem on bidimensional cellular automata with communication constraints. In: Proc. of MCU 2001. LNCS,vol 2055. Springer, Berlin, pp 264–275
50. Umeo H (1996) A note on firing squad synchronization algorithms-A reconstruction of Goto's first-in-the-world optimum-time firing squad synchronization algorithm. In: Kutrib M, Worsch T (eds) Proc. of Cellular Automata Workshop, pp 65
51. Umeo H (2001) Linear-time recognition of connectivity of binary images on 1-bit inter-cell communication cellular automaton. Parallel Computing, 27:587–599
52. Umeo H (2004) A simple design of time-efficient firing squad synchronization algorithms with fault-tolerance. IEICE Trans Inf Syst E87-D(3):733–739
53. Umeo H, Hisaoka M, Akiguchi S (2005) A twelve-state optimum-time synchronization algorithm for two-dimensional rectangular arrays. In: Proc. of the 4th International Conference on Unconventional Computation: UC 2005. LNCS, vol 3699. Springer, Berlin, pp 214–223
54. Umeo H, Hisaoka M, Sogabe T (2005) A survey on optimum-time firing squad synchronization algorithms for one-dimensional cellular automata. Int J Unconv Comput 1: 403–426
55. Umeo H, Hisaoka M, Teraoka M, Maeda M (2005) Several new generalized linear- and optimum-time synchronization algorithms for two-dimensional rectangular arrays. In: Margenstern M (ed) Proc. of the 4th International Conference on Machines, Computations and Universality: MCU 2004. LNCS, vol 3354. Springer, Berlin, pp 223–232
56. Umeo H, Hisaoka M, Michisaka K, Nishioka K, Maeda M (2002) Some new generalized synchronization algorithms and their implementations for large scale cellular automata. In: Proc. of the 3rd International Conference on Unconventional Models of Computation: UMC 2002. LNCS, vol 2509. Springer, Berlin, pp 276–286
57. Umeo H, Kamikawa N (2003) Real-time generation of primes by a 1-bit-communication cellular cutomaton. Fundamenta Informaticae 58:421–435
58. Umeo H, Kamikawa N, Yunès JB (2008) A family of smallest symmetrical four-state firing squad synchronization protocols for one-dimensional ring celluar automata. Luniver Press, Beckington, pp 174–186
59. Umeo H, Maeda M, Fujiwara N (2002) An efficient mapping scheme for embedding any one-dimensional firing squad synchronization algorithm onto two-dimensional arrays. In: Proc. of the 5th International Conference on Cellular Automata for Research and Industry. LNCS, vol 2493. Springer, Berlin, pp 69–81
60. Umeo H, Maeda M, Hisaoka M, Teraoka M (2006) A state-efficient mapping scheme for designing two-dimensional firing squad synchronization algorithms. Fundamenta Informaticae 74(4):603–623
61. Umeo H, Maeda M, Hongyo K (2006) A design of symmetrical six-state 3*n*-step firing squad synchronization algorithms and their implementations. In: Proc. of 7th International Conference on Cellular Automata for Research and Industry, ACRI2006. LNCS, vol 4173. Springer, Berlin, pp 157–168
62. Umeo H, Michisaka K, Kamikawa N (2003) A synchronization problem on 1-bit-communication cellular automata. In: Proc. of International Conference on Computational Science-ICCS2003. LNCS, vol 2657. Springer, Berlin, pp 492–500
63. Umeo H, Michisaka K, Kamikawa N, Kanazawa M (2007) State-efficient one-bit-communication solutions for some classical cellular automata problems. Fundamenta Informaticae 78:449–465
64. Umeo H, Sogabe T, Nomura Y (2000) Correction, optimization and verification of transition rule set for Waksman's firing squad synchronization algorithm. In: Proc. of the Fourth Intern. Conference on Cellular Automata for Research and Industry. Springer, London, pp 152–160
65. Umeo H, Uchino H (2007) A new time-optimum synchronization algorithm for two-dimensional cellular arrays. In: Quesada-Arenciba A, Rodriguez JC, Moreno-Diaz R Jr, R Moreno-Diaz (eds) Proc. of Intern. Conf. on Computer Aided Systems Theory, EUROCAST 2007. LNCS, vol 4739. Springer, Berlin, pp 604–611
66. Umeo H, Yamawaki T, Shimizu N, Uchino H (2007) Modeling and simulation of global synchronization processes for large-scale-of two-dimensional cellular arrays. In: Proc. of Intern. Conf. on Modeling and Simulation, AMS 2007, pp 139–144
67. Umeo H, Yanagihara T (2007) A smallest five-state solution to the firing squad synchronization problem. In: Proc. of 5th Intern. Conf. on Machines, Computations, and Universality, MCU 2007. LNCS, vol 4664. Springer, Berlin, pp 292–302
68. Umeo H, Yanagihara T, Kanazawa M (2006) State-efficient firing squad synchronization protocols for communication-restricted cellular automata. In: Proc. of 7th International Conference on Cellular Automata for Research and Industry, ACRI2006. LNCS, vol 4173. Springer, Berlin, pp 169–181
69. Varshavsky VI, Marakhovsky VB, Peschansky VA (1970) Synchronization of Interacting Automata. Math Syst Theor 4(3):212–230
70. Vivien H (1996) A quasi-optimal time for synchronization two interacting finite automata. J Algebra Comput 6(2):261–267
71. Vivien H (2005) Cellular automata: a geometrical approach. Draft, pp 412
72. Vollmar R (1979) Algorithmen in Zellularautomaten, Teubner, Stuttgart, pp 192

73. Vollmar R (1982) Some remarks about the "Efficiency" of polyautomata. Int J Theor Phys 21(12):1007–1015
74. Waksman A (1966) An optimum solution to the firing squad synchronization problem. Inf Control 9:66–78
75. Worsch T (2000) Linear time language recognition on cellular automata with restricted communication. In: Gonnet GH, Panario D, Viola A (eds) Proc. of LATIN 2000:Theoretical Informatics. LNCS, vol 1776. Springer, Berlin, pp 417–426
76. Yunès JB (1994) Seven-state solution to the firing squad synchronization problem. Theor Comput Sci 127:313–332
77. Yunès JB (2006) Fault tolerant solutions to the firing squad synchronization problem in linear cellular automata. J Cell Autom 1:253–268
78. Yunès JB (2007) Simple new algorithms which solve the firing squad synchronization problem: A 7-states 4*n*-steps solution. In: Proc. of 5th Intern. Conf. on Machines, Computations, and Universality, MCU 2007. LNCS, vol 4664. Springer, Berlin, pp 316–324
79. Yunès JB (2008) An intrinsically non minimal-time Minsky-like 6-states solution to the firing squad synchronization problem. Theor Inf Appl 42(1):55–68
80. Yunès JB (2007) A 4-states algebraic solution to linear cellular automata synchronization. submitted to Information Processing Letters

Books and Reviews

Delorme M, Mozoyer J (eds) (1999) Cellular Automata. Kluwer Academic Publishers, Dordrecht
Ilachinski A (2001) Cellular Automata – A Discrete Universe –. World Scientific, Singapore
Wolfram S (1994) Cellular Automata and Complexity. Addison-Wesley Publishing Company, Reading
Wolfram S (2002) A New Kind of Science. Wolfram Media, Champaign

Fluctuations, Importance of: Complexity in the View of Stochastic Processes

Rudolf Friedrich[1], Joachim Peinke[2], M. Reza Rahimi Tabar[3]
[1] Institute for Theoretical Physics, University of Münster, Münster, Germany
[2] Institute of Physics, Carl-von-Ossietzky University Oldenburg, Oldenburg, Germany
[3] Dept. of Physics, Sharif University of Technology, Theran, Iran

Article Outline

Glossary
Definition of the Subject
Introduction
Stochastic Processes
Stochastic Time Series Analysis
Applications: Processes in Time
Applications: Processes in Scale
Future Directions
Further Reading
Acknowledgment
Bibliography

Glossary

Complexity in time Complex structures may be characterized by non regular time behavior of a describing variable $q \in \mathbf{R}^d$. Thus the challenge is to understand or to model the time dependence of $q(t)$, which may be achieved by a differential equation $\frac{\mathrm{d}}{\mathrm{d}t} q(t) = \dots$ or by the discrete dynamics $q(t+\tau) = f(q(t), \dots)$ fixing the evolution in the future. Of special interest are not only nonlinear equations leading to chaotic dynamics but also those which include general noise terms, too.

Complexity in space Complex structures may be characterized by their spatial disorder. The disorder on a selected scale l may be measured at the location x by some scale dependent quantities, $q(l, x)$, like wavelets, increments and so on. The challenge is to understand or to model the features of the disorder variable $q(l, x)$ on different scales l. If the moments of q show power behavior $\langle q(l, x)^n \rangle \propto l^{\zeta(n)}$ the complex structures are called fractals. Well known examples of spatial complex structures are turbulence or financial market data. In the first case the complexity of velocity fluctuations over different distances l are investigated, in the second case the complexity of price changes over different time steps (time scale) are of interest.

Fokker–Planck equation The evolution of a variable $\mathbf{x}(t)$ from $\mathbf{x}'$ at t' to $\mathbf{x}$ at t, with $t' > t$, is described in a statistical manner by the conditional probability distribution $p(\mathbf{x}, t|\mathbf{x}', t')$. The conditional probability is subject to a Fokker–Planck equation, also known as second Kolmogorov equation, if

$$\begin{aligned}\frac{\partial}{\partial t} p(\mathbf{x}, t \,|\, \mathbf{x}', t') = \\ -\sum_{i=1}^{d} \frac{\partial}{\partial x_i} D_i^{(1)}(\mathbf{x}, t) p(\mathbf{x}, t \,|\, \mathbf{x}', t') \\ + \frac{1}{2} \sum_{i,j=1}^{d} \frac{\partial^2}{\partial x_i \partial x_j} D_{ij}^{(2)}(\mathbf{x}, t) p(\mathbf{x}, t \,|\, \mathbf{x}', t') \,,\end{aligned}$$

holds. Here $\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$ are the drift vector and the diffusion matrix, respectively.

Kramers–Moyal coefficients Knowing for a stochastic process the conditional probability distribution $p(\mathbf{x}(\mathbf{t}), t|\mathbf{x}', t')$, for all t and t' the Kamers–Moyal coefficients can be estimated as nth order moments of the conditional probability distribution. In this way also the drift and diffusion coefficient of the Fokker–Planck equation can be obtained form the empirically measured conditional probability distributions.

Langevin equation The time evolution of a variable $\mathbf{x}(t)$ is described by Langevin equation if for $\mathbf{x}(t)$ it holds:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}(\mathbf{t}) = \mathbf{D}^{(1)}(\mathbf{x}, \mathbf{t}) \cdot \tau + \sqrt{\mathbf{D}^{(2)}(\mathbf{x}, \mathbf{t})} \cdot \boldsymbol{\Gamma}(\mathbf{t}_i) \,.$$

Using Itô's interpretation the deterministic part of the differential equation is equal to the drift term, the noise amplitude is equal to the square root of the diffusion term of a corresponding Fokker–Planck equation. Note, for vanishing noise a purely deterministic dynamics is included in this description.

Stochastic process in scale For the description of complex system with spatial or scale disorder usually a measure of disorder on different scales $q(l, x)$ is used. A stochastic process in scale is now a description of the l evolution of $q(l, x)$ by means of stochastic equations. As a special case the single event $q(l, x)$ follows a Langevin equation, whereas the probability $p(q(l))$ follows a Fokker–Planck equation.

Definition of the Subject

Measurements of time signals of complex systems of the inanimate and the animate world like turbulent fluid motions, traffic flow or human brain activity yield fluctuating time series. In recent years, methods have been devised which allow for a detailed analysis of such data. In particular methods for parameter free estimations of the underlying stochastic equations have been proposed. The present article gives an overview on the achievements obtained so far for analyzing stochastic data and describes results obtained for a variety of complex systems ranging from electrical nonlinear circuits, fluid turbulence, to traffic flow and financial market data. The systems will be divided into two classes, namely systems with complexity in time and systems with complexity in scale.

Introduction

The central theme of the present article is exhibited in Fig. 1. Given a fluctuating, sequentially measured set of experimental data one can pose the question whether it is possible to determine underlying trends and to assess the characteristics of the fluctuations generating the experimental traces. This question becomes especially important for nonlinear systems, which can only partly be analyzed by the evaluation of the powerspectra obtained from a Fourier representation of the data.

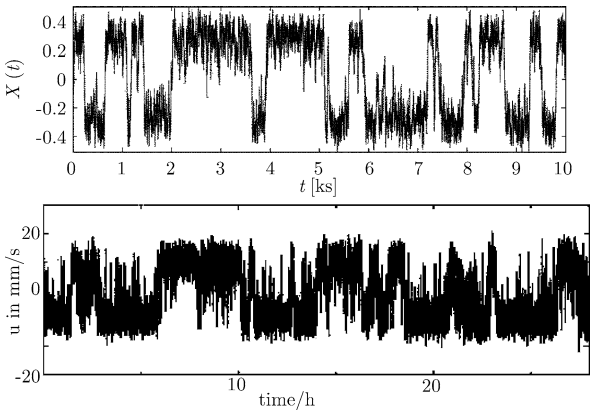

Fluctuations, Importance of: Complexity in the View of Stochastic Processes, Figure 1
Stochastic time series, data generated numerically and measured in a Rayleigh–Bénard experiment: Is it possible to disentangle turbulent trends from chances?

In recent years it has become evident that for a wide class of stochastic processes the posed question can be answered in an affirmative way. A common method has been developed which can deal with stochastic processes (Langevin processes, Lévy processes) in time as well as in scale. In the first case, one faces the analysis of temporal disorder, whereas in the second case one considers scale disorder, which is an inherent feature of turbulent fluid motion and, quite interestingly, can also be detected in financial time series. This scale disorder is often linked to fractal scaling behavior and can be treated by a stochastic ansatz in a more general way.

In the present article we shall give an overview on the developed methods for analyzing stochastic data in time and scale. Furthermore, we list complex systems ranging from electrical nonlinear circuits, fluid turbulence, finance to biological data like heart beat data or data of human tremor, for which a successful application of the data analysis method has been performed. Furthermore, we shall focus on results obtained from some exemplary applications of the method to electronics, traffic flow, turbulence, and finance.

Complexity in Time

Complex systems are composed of a large number of subsystems behaving in a collective manner. In systems far

from equilibrium collectivity arises due to selforganization [1,2,3]. It results in the formation of temporal, spatial, spatio-temporal and functional structures.

The investigation of systems like lasers, hydrodynamic instabilities and chemical reactions has shown that selforganization can be described in terms of order parameters $u_i(t)$ which obey a set of stochastic differential equations of the form

$$\mathrm{d}u_i = N_i[u_1, .., u_n]\mathrm{d}t + \sum_j g_{ij}(u_1, .., u_n)\mathrm{d}W_j \,, \quad (1)$$

where W_j are independent Wiener processes. Although the state vector $\mathbf{q}(t)$ of the complex system under consideration is high dimensional, its long time behavior is entirely governed by the dynamics of typically few order parameters:

$$\mathbf{q}(t) = \mathbf{Q}(u_1, \ldots u_n) \,. \quad (2)$$

This fact allows one to perform a macroscopic treatment of complex systems on the basis of the order parameter concept [1,2,3].

For hydrodynamic instabilities in the laminar flow regime like Rayleigh–Bénard convection or the Taylor–Couette experiment thermal fluctuations are usually small and can be neglected. However, in nonlinear optics and, especially, in biological systems the impact of noise has been shown to be of great importance. In principle, the order parameter equations (1) can be derived from basic equations characterizing the system under consideration close to instabilities [1,2]. However if the basic equations are not available, as is the case e. g. for systems considered in biology or medicine, the order parameter concept yields a top-down approach to complexity [3].

In this top-down approach the analysis of experimental time series becomes a central issue. Methods of nonlinear time series analysis (c.f. the monograph of Kantz and Schreiber [4]) have been widely applied to analyze complex systems. However, the developed methods aim at the understanding of deterministic systems and can only be successful if the stochastic forces are weak. Apparently, these methods have to be extended to include stochastic processes.

Complexity in Scale

In the case of selfsimilar structures complexity is commonly investigated by a local measure $q(l, x)$ characterizing the structure on the scale l at x. Selfsimilarity means that in a certain range of l the processes

$$q(l, x) \,, \quad \lambda^{\xi} q(\lambda l, \lambda^{\gamma} x) \quad (3)$$

should have the same statistics. More precisely, the probability distribution of the quantity q takes the form

$$f(q, l) = \frac{1}{l^{\xi}} F\left(\frac{q}{l^{\xi}}\right) \quad (4)$$

with a universal function $F(Q)$. Furthermore, the moments exhibt scaling behavior

$$\langle q^n(l)\rangle = \int \mathrm{d}q\, q^n \frac{1}{l^{\xi}} F\left(\frac{q}{l^{\xi}}\right) = Q_n l^{n\xi} \quad (5)$$

Such type of behavior has been termed *fractal* scaling behavior.

There are many experimental examples of systems, like turbulent fields or surface roughness, just to mention two, that such a simple picture of a complex structure is only a rough, first approximation. In fact, especially for turbulence where $q(l, x)$ is taken as a velocity increment, it has been argued that *multifractal behavior* is more appropriate, where the nth order moments scale according to

$$\langle q^n(l)\rangle = Q_n l^{\zeta(n)} \,, \quad (6)$$

where the scaling indices $\zeta(n)$ are nonlinear functions of the order n:

$$\zeta(n) = n\xi_0 + n^2\xi_1 + n^3\xi_2 + \cdots \,. \quad (7)$$

Such a behavior can formally be obtained by the assumption that the probability distribution $f(q, l)$ has the following form

$$f(q, l) = \int \mathrm{d}\xi\, p(\xi, l) \frac{1}{l^{\xi}} F\left(\frac{q}{l^{\xi}}\right) \,. \quad (8)$$

This formula is based on the assumption that in a turbulent flow regions with different scaling indices ξ exist, where $p(\xi, l)$ gives a measure of the scaling indices ξ at scale l. The major shortcoming of the fractal and multifractal approach to complexity in scale is the fact that it only addresses the statistics of the measure $q(l, x)$ at a single scale l. In general one has to expect dependencies of the measures $q(l, x)$ and $q(l', x)$ from different scales. Thus the question, which we will address in the following, can be posed, are there methods, which lead to a more comprehensive characterization of the scale disorder by general joint statistics

$$f(q_N, l_N; q_{N-1}, l_{N-1}; \ldots; q_1, l_1; q_0, l_0) \quad (9)$$

of the local measure q at multiple scales l_i.

Stochastic Data Analysis

Processes in time and scale can be analyzed in a similar way, if we generalize the time process given in (1) to a stochastic process equation evolving in scale

$$q(l + \mathrm{d}l) = q(l) + N(q, l)\mathrm{d}l + g(q, l)\mathrm{d}W(l) \tag{10}$$

where $\mathrm{d}W$ belongs to a random process. The aim of data analysis is to disentangle deterministic dynamics and the impact of fluctuations. Loosely speaking, this amounts to detect *trends* and *chances* in data series of complex systems.

A complete analysis of experimental data, which is generated by the interplay of deterministic dynamics and dynamical noise, has to address the following issues

- Identification of the order parameters
- Extracting the deterministic dynamics
- Evaluating the properties of fluctuations

The outline of the present article is as follows. First, we shall summarize the description of stochastic processes focusing mainly on Markovian processes. Second, we discuss the approach developed to analyze stochastic processes. The last parts are devoted to applications of the data analysis method to processes in time and processes in scale.

Stochastic Processes

In the following we consider the class of systems which are described by a multivariate state vector $\mathbf{X}(t)$ contained in a d-dimensional state space $\{\mathbf{x}\}$. The evolution of the state vector $\mathbf{X}(t)$ is assumed to be governed by a deterministic part and by noise:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{X}(t) = \mathbf{N}\big(\mathbf{X}(t), t\big) + \mathbf{F}\big(\mathbf{X}(t), t\big)\,. \tag{11}$$

$\mathbf{N}$ denotes a nonlinear function depending on the stochastic variable $\mathbf{X}(t)$ and additionally, may explicitly depend on time t (Note, time t can also be considered as a general variable and replaced for example by a scale variable l like in (86)). Because the function $\mathbf{N}$ can be nonlinear, also systems exhibiting chaotic time evolution in the deterministic case are included in the class of stochastic processes (11).

The second part, $\mathbf{F}(\mathbf{X}(t), t)$, fluctuates on a fast time scale. We assume that the d components F_i can be represented in the form

$$F_i\big(\mathbf{X}(t), t\big) = \sum_{j=1}^{d} g_{ij}\big(\mathbf{X}(t), t\big)\Gamma_j(t)\,. \tag{12}$$

The quantities $\Gamma_j(t)$ are considered to be random functions, whose statistical characteristics are well-defined. It is evident that these properties significantly determine the dynamical behavior of the state vector $\mathbf{X}(t)$. Formally, our approach also includes purely deterministic processes taking $F = 0$.

Discrete Time Evolution

It is convenient to consider the temporal evolution (11) of the state vector $\mathbf{X}(t)$ on a time scale, which is large compared to the time scale of the fluctuations $\Gamma_j(t)$.

As we shall briefly indicate below, a stochastic process related to the evolution Equation (11) can be modeled by stochastic evolution laws relating the state vectors $\mathbf{X}(t)$ at times $t_i, t_{i+1} = t_i + \tau, t_{i+2} = t_i + 2\tau, \ldots$ for small but finite values of τ.

In the present article we shall deal with the class of proper Langevin processes and generalized Langevin processes, which are defined by the following discrete time evolutions.

a) Proper Langevin Equations: White Noise The discrete time evolution of a proper Langevin process is given by

$$\mathbf{X}(t_{i+1}) = \mathbf{X}(t_i) + \mathbf{N}(\mathbf{X}(t_i), t_i)\cdot\tau + g(\mathbf{X}(t_i), t_i)\cdot\sqrt{\tau}\boldsymbol{\eta}(t_i) \tag{13}$$

where the stochastic increment $\boldsymbol{\eta}(t_i)$ is a fluctuating quantity characterized by a Gaussian distribution with zero mean, $\langle\boldsymbol{\eta}(t_i)\rangle = 0$

$$h(\boldsymbol{\eta}) = \frac{1}{(\sqrt{2\pi})^d}\mathrm{e}^{-\frac{\boldsymbol{\eta}^2}{2}} = \frac{1}{(\sqrt{2\pi})^d}\mathrm{e}^{-\sum_{\alpha=1}^{d}\frac{\eta_\alpha^2}{2}} \tag{14}$$

Furthermore, the increments are statistically independent for different times

$$\langle\boldsymbol{\eta}(t_i)\boldsymbol{\eta}(t_j)\rangle = \delta_{ij} \tag{15}$$

b) Generalized Langevin Equations: Lévy Noise A more general class is formed by the discrete time evolution laws [5,6],

$$\mathbf{X}(t_{i+1}) = \mathbf{X}(t_i) + \mathbf{N}(\mathbf{X}(t_i), t_i)\cdot\tau + g(\mathbf{X}(t_i), t_i)\cdot\tau^{1/\alpha}\boldsymbol{\eta}_{\alpha,\beta}(t_i) \tag{16}$$

where the increment $\boldsymbol{\eta}_{\alpha,\beta}(t_i)$ is a fluctuating quantity distributed according to the Lévy stable law characterized by the Lévy parameters α, β, [7].

As is well-known, only the Fourier-transform of this distribution can be given:

$$\begin{aligned} h_{\alpha,\beta} &= \frac{1}{2\pi}\int \mathrm{d}k\, Z(k,\alpha,\beta)\,\mathrm{e}^{-ikx} \\ Z(k,\alpha,\beta) &= \mathrm{e}^{-i|k|^{\alpha}(1-i\beta\,\mathrm{sign}(k)\Phi)} \\ \Phi &= \tan\pi\frac{\alpha}{2} \quad \alpha \neq 1\,, \\ \Phi &= -\frac{2}{\pi}\ln|k| \quad \alpha = 1\,. \end{aligned} \tag{17}$$

The Gaussian distribution is contained in the class of Lévy stable distributions ($\alpha = 2$, $\beta = 0$). Formally, α can be taken from the interval $0 < \alpha \leq 2$. However, for applications it seems reasonable to choose $1 < \alpha \leq 2$ in order that the first moment of the noise source exists.

The consideration of this type of statistics for the noise variables η is based on the central limit theorem, as discussed in a subsection below.

The discrete Langevin (13) and generalized Langevin Equations (16) have to be considered in the limit $\tau \to 0$. They are the basis of all further treatments. A central point is that if one assumes the noise sources to be independent of the variable $\mathbf{X}(t_i)$ the discrete time evolution equations define a Markov process, whose generator, i. e. the conditional probability distribution or short time propagator can be established on the basis of (13), (16).

In the following we shall discuss, how the discrete time processes can be related to the stochastic evolution equation (11).

Discrete Time Approximation of Stochastic Evolution Equations In order to motivate the discrete time approximations (13), (16) we integrate the evolution law (11) over a finite but small time increment τ:

$$\begin{aligned} \mathbf{X}(t+\tau) &= \mathbf{X}(t) + \int_t^{t+\tau} \mathrm{d}t'\, \mathbf{N}(\mathbf{X}(t'),t') \\ &\quad + \int_t^{t+\tau} \mathrm{d}t'\, g(\mathbf{X}(t'),t')\boldsymbol{\Gamma}(t') \\ &\approx \mathbf{X}(t) + \tau\mathbf{N}(\mathbf{X}(t),t) \\ &\quad + \int_t^{t+\tau} \mathrm{d}t'\, g(\mathbf{X}(t'),t')\boldsymbol{\Gamma}(t')\,. \end{aligned} \tag{18}$$

The time interval τ is chosen to be larger than the time scale of the fluctuations of $\Gamma_j(t)$. It involves the rapidly fluctuating quantities $\Gamma_j(t)$ and is denoted as a stochastic integral [8,9,10,11].

If we assume the matrix g to be independent on time t and state vector $\mathbf{X}(t)$, we arrive at the integrals

$$\mathrm{d}\mathbf{W}(t,\tau) = \int_t^{t+\tau} \mathrm{d}t'\, \boldsymbol{\Gamma}(t')\,. \tag{19}$$

These are the quantities, for which a statistical characterization can be given. We shall pursue this problem in the next subsection.

However, looking at (18) we encounter the difficulty that the integrals over the noise forces may involve functions of the state vector within the time interval $(t, t+\tau)$. The interpretation of such integrals for wildly fluctuating, stochastic quantities $\boldsymbol{\Gamma}(t)$ is difficult. The simplest way is to formulate an interpretation of these terms leading to different interpretations of the stochastic evolution equation (11). We formulate the widely used definitions due to Itô and Stratonovich.

In the *Itô* sense, the integral is interpreted as

$$\begin{aligned} \int_t^{t+\tau} \mathrm{d}t'\, g(\mathbf{X}(t'),t')\boldsymbol{\Gamma}(t') &= g(\mathbf{X}(t),t)\int_t^{t+\tau} \mathrm{d}t'\, \boldsymbol{\Gamma}(t') \\ &= g(\mathbf{X}(t),t)\,\mathrm{d}\mathbf{W}(t,\tau)\,. \end{aligned} \tag{20}$$

The *Stratonovich* definition is

$$\begin{aligned} &\int_t^{t+\tau} \mathrm{d}t'\, g(\mathbf{X}(t'),t')\Gamma(t') \\ &\quad = g\left(\frac{\mathbf{X}(t+\tau)+\mathbf{X}(t)}{2}, t+\frac{\tau}{2}\right)\int_t^{t+\tau} \mathrm{d}t'\, \boldsymbol{\Gamma}(t') \\ &\quad = g\left(\frac{\mathbf{X}(t+\tau)+\mathbf{X}(t)}{2}, t+\frac{\tau}{2}\right)\mathrm{d}\mathbf{W}(t,\tau)\,. \end{aligned} \tag{21}$$

Since from experiments one obtains probability distributions of stochastic processes which are related to stochastic Langevin equations, we are free to choose a certain interpretation of the process. In the following we shall always adopt the Itô interpretation. In this case, the drift vector $\mathbf{D}^1(\mathbf{x},t) = \mathbf{N}(\mathbf{x},t)$ coincides with the nonlinear vector field $\mathbf{N}(\mathbf{x},t)$.

Limit Theorems, Wiener Process, Lévy Process In the following we shall discuss possibilities to characterize the stochastic integrals

$$W(t+\tau) - W(t) = \int_t^{t+\tau} \mathrm{d}t'\, \Gamma(t') \tag{22}$$

$\Gamma(t)$ is a rapidly fluctuating quantity of zero mean. In order to characterize the properties of this force one can resort to the limit theorems of statistical mechanics [7].

The central limit theorem states that if the quantities Γ_j, $j = 1,\ldots,n$ are statistically independent variables of zero mean and variance σ^2 then the sum

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\Gamma_j = \eta \tag{23}$$

tends to a Gaussian random variable with variance σ^2 for large values of n. The limiting probability distribution $h(\eta)$ is then a Gaussian distribution with variance σ^2:

$$h(\eta) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{\eta^2}{2\sigma^2}} . \tag{24}$$

As is well known, there is a generalization of the central limit theorem, which applies to random variables whose second moment does not exist. It states that the distribution of the sum over identically distributed random variables Γ_j

$$\frac{1}{n^{1/\alpha}} \sum_{j=1}^{n} \Gamma_j = \eta_{\alpha,\beta} \tag{25}$$

tends to a random variable $\eta_{\alpha,\beta}$, which is distributed according to the Lévy-stable distribution $h_{\alpha,\beta}(\eta)$. The Lévy stable distributions can only be given by their Fourier – transforms, cf. Eq. (17).

In order to evaluate the integral (22) using the limit theorems, it is convenient to represent the stochastic force $\Gamma(t)$ as a sum over N_τ δ-kicks occuring at discrete times t_j

$$\Gamma(t) = \sum_j \Gamma_j (\Delta t)^{1/\alpha} \delta(t - t_j) . \tag{26}$$

Thereby, Δt is the time difference between the occurrence of two kicks. Then, we obtain

$$\begin{aligned} \mathrm{d}W(t,\tau) &= \sum_{j,t_j\in\tau} \Gamma_j (\Delta t)^{1/\alpha} \\ &= (N_\tau \Delta t)^{1/\alpha} \frac{1}{N_\tau^{1/\alpha}} \sum_{j,t_j\in\tau} \Gamma_j = \tau^{1/\alpha} \eta(t) . \end{aligned} \tag{27}$$

An application of the central limit theorem shows that if the quantities Γ_j are identically distributed independet variables the integral

$$\frac{1}{\tau^{1/\alpha}} \int_t^{t+\tau} = \eta(t) \tag{28}$$

can be considered to be a random variable $\eta(t)$ which in the limit $N_\tau \to \infty$ tends to a stable random variable.

Thus, for the case $\alpha = 2$, i. e. for the case where the second moments of the random kicks exist, the stochastic variable $\mathrm{d}W(t,\tau)$ can be represented by the increments

$$\mathrm{d}W(t,\tau) = \sqrt{\tau}\eta(t) \tag{29}$$

where $\eta(t)$ is a Gaussian distributed random variable.

For the more general case, $\mathrm{d}W(t,\tau)$ is a stochastic variable

$$\mathrm{d}W(t,\tau) = \tau^{1/\alpha} \eta_{\alpha,\beta}(t) \tag{30}$$

where the distribution of $\eta_{\alpha,\beta}$ is the Lévy distribution (17).

Statistical Description of Stochastic Processes

In the previous subsection we have discussed processes described by stochastic equations. In the present subsection we shall summarize the corresponding statistical description. Such a description is achieved by introducing suitable statistical averages. We shall denote these averages by the brackets $\langle \ldots \rangle$. For stationary processes the averages can be viewed as a time average. For nonstationary processes averages are defined as ensemble averages, i. e. averages over an ensemble of experimental (or numerical) realizations of the stochastic process (11). For stationary processes in time, one usually deals with time averages. For processes in scale, the average is an ensemble average.

Probability Distributions The set of stochastic evolution equations (11), or it's finite time representations (13), (16) define a Markov process. We consider the joint probability density

$$f(\mathbf{x}_n, t_n; \ldots; \mathbf{x}_1, t_1; \mathbf{x}_0, t_0) \tag{31}$$

which is related to the joint probability, to find the system at times t_i in the volume ΔV_i in phase space. If we take times t_i which are separated by the small time increment $\tau = t_{i+1} - t_i$, then the probability density can be related to the discrete time representation of the stochastic process (13), (16) according to

$$\begin{aligned} &f(\mathbf{x}_n, t_n; \ldots; \mathbf{x}_1, t_1; \mathbf{x}_0, t_0) \\ &\quad = \langle \delta(\mathbf{x}_n - \mathbf{X}(t_n)) \ldots \delta(\mathbf{x}_0 - \mathbf{X}(t_0)) \rangle , \end{aligned} \tag{32}$$

where the brackets indicate the statistical average, which may be a time average (for stationary processes) or an ensemble average.

Markov Processes An important subclass of stochastic processes are Markov processes. For these processes the joint probability distribution $f(\mathbf{x}_n, t_n; \ldots; \mathbf{x}_1, t_1; \mathbf{x}_0, t_0)$ can be constructed from the knowledge of the conditional probability distributions

$$p(\mathbf{x}_{i+1}, t_{i+1} | \mathbf{x}_i, t_i) = \frac{f(\mathbf{x}_{i+1}, t_{i+1}; \mathbf{x}_i, t_i)}{f(\mathbf{x}_i, t_i)} \tag{33}$$

according to

$$\begin{aligned} &f(\mathbf{x}_n, t_n; \ldots; \mathbf{x}_1, t_1; \mathbf{x}_0, t_0) \\ &\quad = p(\mathbf{x}_n, t_n | \mathbf{x}_{n-1}, t_{n-1}) \ldots p(\mathbf{x}_1, t_1 | \mathbf{x}_0, t_0) f(\mathbf{x}_0, t_0) , \end{aligned} \tag{34}$$

Here the Markov property of a process for multiple conditioned probabilities

$$p(\mathbf{x}_i, t_i \,|\, \mathbf{x}_{i-1}, t_{i-1}; \ldots; \mathbf{x}_0, t_0) = p(\mathbf{x}_i, t_i \,|\, \mathbf{x}_{i-1}, t_{i-1}) \tag{35}$$

is used. As a consequence, the knowledge of the transition probabilities together with the initial probability distribution $f(\mathbf{x}_0, t_0)$ suffices to define the N-times probability distribution.

It is straightforward to prove the Chapman–Kolmogorov equation

$$p(\mathbf{x}_j, t_j \,|\, \mathbf{x}_i, t_i) = \int \mathrm{d}\mathbf{x}_k \, p(\mathbf{x}_j, t_j \,|\, \mathbf{x}_k, t_k)\, p(\mathbf{x}_k, t_k \,|\, \mathbf{x}_i, t_i). \tag{36}$$

This relation is valid for all times $t_i < t_k < t_j$. In the following we shall show that the transition probabilities $p(\mathbf{x}_{j+1}, t+\tau|\mathbf{x}_j, t)$ can be determined for small time differences τ. This defines the so-called short time propagators.

Short Time Propagator of Langevin Processes It is straightforward to determine the short time propagator from the finite time approximation (13) of the Langevin equation. We shall denote these propagators by $p(\mathbf{x}_{j+1}, t+\tau|\mathbf{x}_j, t)$, in contrast to the finite time propagators (33), for which the time interval $t_{i+1} - t_i$ is large compared to τ.

We first consider the case of Gaussian noise. The variables $\boldsymbol{\eta}(t_i)$ are Gaussian random vectors with probability distribution

$$h[\boldsymbol{\eta}] = \frac{1}{\sqrt{(2\pi)^d}} \exp\left[-\frac{\boldsymbol{\eta}\cdot\boldsymbol{\eta}}{2}\right]. \tag{37}$$

The finite time intepretation of the Langevin equation can be rewritten in the form

$$\boldsymbol{\eta}(t_i) = \frac{1}{\tau^{1/2}}[g(\mathbf{X}(t_i), t_i)]^{-1}[\mathbf{X}(t_{i+1}) - \mathbf{X}(t_i) - \tau\mathbf{N}(\mathbf{X}(t_i))]. \tag{38}$$

This relation, in turn, defines the transition probability distribution

$$p(\mathbf{x}_{i+1}, t_{i+1} \,|\, \mathbf{x}_i, t_i)\,\mathrm{d}\mathbf{x}_{i+1} = h[\boldsymbol{\eta} = \boldsymbol{\eta}(t_i)]J(\mathbf{x}_i, t_i)\,\mathrm{d}\mathbf{x}_{i+1}, \tag{39}$$

where J is the determinant of the Jacobian

$$J_{\alpha\beta} = \frac{\partial\boldsymbol{\eta}_\alpha(t_i)}{\partial\mathbf{x}_{i+1,\beta}}, \tag{40}$$

and $[g]^{-1}$ denotes the inverse of the matrix g (which is assumed to exist).

For the following it will be convenient to define the so-called diffusion matrix $D^{(2)}(\mathbf{x}_i, t_i)$

$$D^{(2)}(\mathbf{x}_i, t_i) = g^T(\mathbf{x}_i, t_i)\, g(\mathbf{x}_i, t_i) \tag{41}$$

We are now able to explicitly state the short time propagator of the process (13):

$$p(\mathbf{x}_i+\tau, t_{i+1}|\mathbf{x}_i, t_i) = \frac{1}{\sqrt{(2\pi\tau)^d \operatorname{Det}[D^{(2)}]}} \mathrm{e}^{-S(\mathbf{x}_{i+1},\mathbf{x}_i,t_i,\tau)} \tag{42}$$

We have defined the quantity $S(\mathbf{x}_{i+1}, \mathbf{x}_i, t_i, \tau)$ according to

$$S(\mathbf{x}_{i+1}, \mathbf{x}_i, t_i, \tau) = \tau\left[\frac{\mathbf{x}_{i+1}-\mathbf{x}_i}{\tau} - \mathbf{D}^{(1)}(\mathbf{x}_i, t)\right] \cdot [D^{(2)}(\mathbf{x}_i, t_i)]^{-1}\left[\frac{\mathbf{x}_{i+1}-\mathbf{x}_i}{\tau} - \mathbf{D}^{(1)}(\mathbf{x}_i, t_i)\right]. \tag{43}$$

As we see, the short time propagator, which yields the transition probability density from state $\mathbf{x}_i$ to state $\mathbf{x}_{i+1}$ in the finite but small time interval τ is a normal distribution.

Short Time Propagator of Lévy Processes It is now straightforward to determine the short time propagator for Lévy processes. We have to replace the Gaussian distribution by the (multivariate) Lévy distribution $h_{\alpha,\beta}(\boldsymbol{\eta})$. As a consequence, we obtain the conditional probability, i. e. the short time propagator, for Lévy processes:

$$p(\mathbf{x}_{i+1}, t_i+\tau \,|\, \mathbf{x}_i, t_i) = \frac{1}{\operatorname{Det}[g(\mathbf{x}_i, t_i)]} h_{\alpha,\beta} \cdot \left\{\frac{1}{\tau^{1/\alpha}}[g(\mathbf{x}(t_i), t_i)]^{-1}[\mathbf{x}(t_{i+1}) - \mathbf{x}(t_i) - \tau\mathbf{N}(\mathbf{x}(t_i))]\right\} \tag{44}$$

Joint Probability Distribution and Markovian Properties Due to statistical independence of the random vectors $\boldsymbol{\eta}(t_i)$, $\boldsymbol{\eta}(t_j)$ for $i \neq j$ we obtain the joint probability distribution as a product of the distributions $h(\boldsymbol{\eta})$:

$$h(\boldsymbol{\eta}_N, \ldots, \boldsymbol{\eta}_0) = h(\boldsymbol{\eta}_N)\, h(\boldsymbol{\eta}_{N-1}) \ldots h(\boldsymbol{\eta}_0) \tag{45}$$

Furthermore, we observe that under the assumption that the random vector $\boldsymbol{\eta}(t_i)$ is independent on the variables $\mathbf{X}(t_j)$ for all $j \leq i$ we can construct the N-time probability distribution

$$f(\mathbf{x}_n, t_n; \ldots; \mathbf{x}_1, t_1; \mathbf{x}_0, t_0) = p(\mathbf{x}_n, t_n \,|\, \mathbf{x}_{n-1}, t_{n-1}) \ldots p(\mathbf{x}_1, t_1 \,|\, \mathbf{x}_0, t_0) f(\mathbf{x}_0, t_0) \tag{46}$$

However, this is the definition of a Markov chain. Thereby, the transition probabilities are the short time propagators, i. e. the representation (46) is valid in the short time limit $\tau \to 0$. The probability distribution (46) is the discrete approximation of the *path integral representation* of the stochastic process under consideration [8].

Let us summarize: The statistical description of the Langevin equation based on the n-times joint probability distribution leads to the representation in terms of the conditional probability distribution. This representation is the definition of a Markov process.

Due to the assumptions on the statistics of the fluctuating forces different processes arise. If the fluctuating forces are assumed to be Gaussian the short time propagator is Gaussian and, as a consequence, solely defined by the drift vector and the diffusion matrix.

If the fluctuating forces are assumed to be Lévy distributed, more complicated short time propagators arise.

Let us add the following remarks:

a) The assumption of Gaussianity of the statistics is not necessary. One can consider fluctuating forces with non-Gaussian probability distributions. In this case the probability distributions have to be characterized by higher order moments, or, more explicitly, by its cumulants. At this point we remind the reader that for non-Gaussian distributions, infinitely many cumulants exist.

b) The Markovian property, i. e. the fact that the propagator $p(\mathbf{x}_i, t_i|\mathbf{x}_{i-1}t_{i-1})$ does not depend on states $\mathbf{x}_k$ at times $t_k < t_{i-1}$ is usually violated for physical systems due to the fact that the noise sources become correlated for small time differences T_{mar}. This point already has been emphasized in the famous work of A. Einstein on Brownian motion, which is one of the first works on stochastic processes [12]. This time scale is denoted as Markov–Einstein scale. It seems to be a highly interesting quantity especially for nonequilibrium systems like turbulence [13,14] and earthquake signals [15].

Finite Time Propagators

Up to now, we have considered the short time propagators $p(\mathbf{x}_i, t_i|\mathbf{x}_{i-1}, t_{i-1})$ for inifinitesimal time differences $t_i - t_{i-1} = \tau$. However, one is interested in the conditional probability distributions for finite time intervals, $p(\mathbf{x}, t|\mathbf{x}', t')$, $t - t' \gg \tau$.

Fokker–Planck Equation The conditional probability distribution $p(\mathbf{x}, t|\mathbf{x}', t')$, $t - t' \gg \tau$ can be obtained from the solution of the Fokker–Planck equation (also known as second Kolmogorov equation [16]):

$$\frac{\partial}{\partial t} p(\mathbf{x}, t \,|\, \mathbf{x}', t') = -\sum_{i=1}^{d} \frac{\partial}{\partial x_i} D_i^{(1)}(\mathbf{x}, t)\, p(\mathbf{x}, t \,|\, \mathbf{x}', t') + \frac{1}{2} \sum_{i,j=1}^{d} \frac{\partial^2}{\partial x_i \partial x_j} D_{ij}^{(2)}(\mathbf{x}, t)\, p(\mathbf{x}, t \,|\, \mathbf{x}', t') \,, \quad (47)$$

$\mathbf{D}^{(1)}$ and $\mathbf{D}^{(2)}$ are drift vector and diffusion matrix.

Under consideration of Itô's definitions of stochastic integrals the coefficients $\mathbf{D}^{(1)}$, $\mathbf{D}^{(2)}$ of the Fokker–Planck equation (47) and the functionals $\mathbf{N}$, $\mathbf{g}$ of the Langevin equation (11), (12) are related by

$$D_i^{(1)}(\mathbf{x}, t) = N_i(\mathbf{x}, t) \,, \quad (48)$$

$$D_{ij}^{(2)}(\mathbf{x}, t) = \sum_{l=1}^{d} g_{il}(\mathbf{x}, t)\, g_{jl}(\mathbf{x}, t) \,. \quad (49)$$

They are defined according to

$$D_i^{(1)}(\mathbf{x}, t) = \lim_{\tau \to 0} \frac{1}{\tau} \langle X_i(t+\tau) - x_i \rangle|_{\mathbf{X}(t)=\mathbf{x}} = \lim_{\tau \to 0} \frac{1}{\tau} \int \mathrm{d}\mathbf{x}'\, p(\mathbf{x}', t+\tau \,|\, \mathbf{x}, t)\big(x_i' - x_i\big) \quad (50)$$

$$D_{ij}^{(2)}(\mathbf{x}, t) = \lim_{\tau \to 0} \frac{1}{\tau} \langle (X_i(t+\tau) - x_i)(X_j(t+\tau) - x_j) \rangle|_{\mathbf{X}(t)=\mathbf{x}} = \lim_{\tau \to 0} \frac{1}{\tau} \int \mathrm{d}\mathbf{x}'\, p(\mathbf{x}', t+\tau|\mathbf{x}, t)\big(x_i' - x_i\big)\big(x_j' - x_j\big) \,. \quad (51)$$

These expressions demonstrate that drift vector and diffusion matrix can be determined as the first and second moments of the conditional probability distributions $p(\mathbf{x}', t+\tau|\mathbf{x}, t)$ in the small time limit.

Fractional Fokker–Planck Equations The finite time propagators or conditional probability distributions of stochastic processes containing Lévy-noise lead to fractional diffusion equations. For a discussion of this topic we refer the reader to [5,6].

Master Equation The most general equation specifying a Markov process for a continuous state vector $\mathbf{X}(t)$ takes the form

$$\frac{\partial}{\partial t} p(\mathbf{x}, t \,|\, \mathbf{x}_0, t_0) = \int \mathrm{d}\mathbf{x}'\, w(\mathbf{x}, \mathbf{x}', t)\, p(\mathbf{x}', t \,|\, \mathbf{x}_0, t_0) - \int \mathrm{d}\mathbf{x}'\, w(\mathbf{x}', \mathbf{x}, t)\, p(\mathbf{x}, t \,|\, \mathbf{x}_0, t_0) \quad (52)$$

here w denote transition probabilities.

Measurement Noise

We can now go a step ahead and include measurement noise. Due to measurement noise, the observed state vector, which we shall now denote by $\mathbf{Y}(t_i)$, is related to the stochastic variable $\mathbf{X}(t_i)$ by an additional noise term $\epsilon(t_i)$:

$$\mathbf{Y}(t_i) = \mathbf{X}(t_i) + \epsilon(t_i) \quad (53)$$

We assume that the stochastic variables $\epsilon(t_i)$ have zero mean, are statistically independent and obey the probability density $h(\epsilon)$. Then the probability distribution for the variable $\mathbf{Y}_{t_i}$ is given by

$$g(\mathbf{y}_n, t_n; \ldots; \mathbf{y}_0, t_0) = \int \ldots \int \mathrm{d}\boldsymbol{\epsilon}_n \ldots \mathrm{d}\boldsymbol{\epsilon}_0 \cdot f(\mathbf{y}_n - \epsilon_n, t_n; \ldots; \mathbf{y}_0 - \epsilon_0, t_0)\, h(\epsilon_n) \ldots h(\epsilon_0)\,. \tag{54}$$

The short time propagator recently has been determined for the Ornstein–Uhlenbeck process, a process with linear drift term and constant diffusion. Analysis of data sets spoilt by measurement noise is currently under investigation [17,18].

Stochastic Time Series Analysis

The ultimate goal of nonlinear time series analysis applied to deterministic systems is to extract the underlying nonlinear dynamical system directly from measured time series in the form of a system of differential equations, cf. [4]. The role played by dynamic fluctuations has not been fully appreciated. Mostly, noise has been considered as a random variable additively superimposed on a trajectory generated by a deterministic dynamical system. Noise has been usually considered as extrinsic or measurement noise. The problem of dynamical noise, i. e. fluctuations which interfere with the deterministic dynamical evolution, has not been addressed in full details.

The natural extension of the nonlinear time series analysis to (continuous) Markov processes is the estimation of *short time propagators* from time series. During recent years, it has become evident that such an approach is feasible. In fact, noise may help in the estimation of the deterministic ingredients of the dynamics. Due to dynamical noise the system explores a larger part of phase space and thus measurements of signals yield considerably more information about the dynamics as compared to the purely deterministic case, where the trajectories fastly converge to attractors providing only limited information.

The analysis of data set's of stochastic systems exhibiting Markov properties has to be performed along the following lines:

- Disentangling Measurement and Dynamical Noise
- Evaluating Markovian Properties
- Determination of Short Time Propagators
- Reconstruction of Data

Since the methods for disentangling measurement and dynamical noise are currently under intense investigation, see Subsect. "Measurement Noise", our focus is on the three remaining issues.

Evaluating Markovian Properties

In principle it is a difficult task to decide on Markovian properties by an inspection of experimental data. The main point is that Markovian properties usually are violated for small time increments τ, as it already has been pointed out above and in [12]. There are at least two reasons for this fact.

First, the dynamical noise sources become correlated at small time differences. If we consider Gaussian noise sources one usually observes an exponential decay of correlations

$$\langle \Gamma_i(t)\Gamma_j(t')\rangle = \delta_{ij}\,\frac{\mathrm{e}^{-|t-t'|/T_{\mathrm{mar}}}}{T_{\mathrm{mar}}} \tag{55}$$

Markovian properties can only expected to hold for time increments $\tau > T_{\mathrm{mar}}$.

Second, measurement noise can spoil Markovian properties [19].

Thus, the estimation of the Markovian time scale T_{mar} is a necessary step for stochastic data analysis. Several methods have been proposed to test Markov properties.

Direct Evaluation of Markovian Properties A direct way is to use the definition of a Markov process (35) and to consider the higher order conditional probability distributions

$$p(\mathbf{x}_3, t_3 \,|\, \mathbf{x}_2 t_2; \mathbf{x}_1, t_1) = \frac{f(\mathbf{x}_3, t_3; \mathbf{x}_2 t_2; \mathbf{x}_1, t_1)}{f(\mathbf{x}_2 t_2; \mathbf{x}_1, t_1)} = p(\mathbf{x}_3, t_3 \,|\, \mathbf{x}_2 t_2) \tag{56}$$

This procedure is feasible if large data sets are available. Due to the different conditioning both probabilities are typically based on different number of events. As an appropriate method to statistically show the similarity of (56) the Wilcoxon Test has been proposed, for details see [20,21].

In principle, higher order conditional probability distributions should be considered in a similar way. However, the validity of relation (56) is a strong hint for Markovianity.

Evaluation of Chapman–Kolmogorov Equation An indirect way is to use the Chapman–Kolmogorov equation (36), whose validity is a necessary condition for Markovianity. The method is based on a comparison between the conditional pdf

$$p(\mathbf{x}_k, t_k \,|\, \mathbf{x}_i, t_i) \tag{57}$$

taken from experiment and the one calculated by the Chapman–Kolmogorov equation

$$\tilde{p}(\mathbf{x}_k, t_k|\mathbf{x}_i, t_i) = \int \mathrm{d}\mathbf{x}_j \, p(\mathbf{x}_k, t_k|\mathbf{x}_j, t_j) \, p(\mathbf{x}_j, t_j|\mathbf{x}_i, t_i) \tag{58}$$

where t_j is an intermediate time $t_i < t_j < t_k$. A refined method can be based on an iteration of the Chapman–Kolmogorov equation, i. e. considering several intermediate times. If the Chapman–Kolmogorov equation is not fulfilled, deviations are enhanced by each iteration.

Direct Estimation of Stochastic Forces Probably the most direct way is the determination of the stochastic forces from data. If the drift vector field $\mathbf{D}^{(1)}(\mathbf{x}, t)$ has been established, as discussed below, the fluctuating forces can be estimated according to

$$g(\mathbf{x}, t)\boldsymbol{\Gamma}(t) = \frac{\mathrm{d}\mathbf{x} - \tau \mathbf{D}^{(1)}(\mathbf{x}, t)}{\sqrt{\tau}} \, . \tag{59}$$

The correlations of this force can then be examined directly, see also [22] and Subsect. "Noisy Circuits".

Differentiating Between Stochastic Process and Noise Data Looking at the joint statistics of increments extracted from given data, it could be shown that the nesting of increments and the resulting statistics can be used to differentiate between noise – like data sets and those resulting from stochastic time processes [24].

Estimating the Short Time Propagator

A crucial step in the stochastic analysis is the assessment of the short time propagator of the continuous Markov process. This gives access to the deterministic part of the dynamics as well as to the fluctuations.

Gaussian Noise We shall first consider the case of Gaussian white noise. As we have already indicated, drift vector and diffusion matrix are defined as the first and second moments of the conditional probability distribution or short time propagator, Eq. (50).

We shall now describe an operational approach, which allows one to estimate drift vector and diffusion matrix from data and has been successfully applied to a variety of stochastic processes. We shall discuss the case, where averages are taken with respect to an ensemble of experimental realizations of the stochastic process under consideration in order to include nonstationary processes. Replacing the ensemble averages by time averages for statistically stationary processes is straightforward.

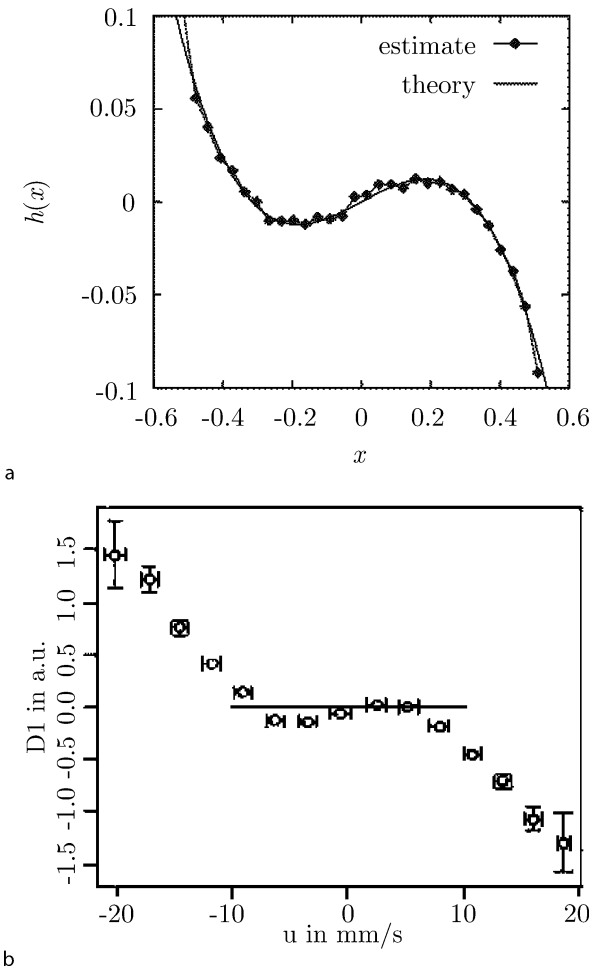

Fluctuations, Importance of: Complexity in the View of Stochastic Processes, Figure 2
Estimated drift terms from the numerical and experimental data of Fig. 1. In part a also the exact function of the numerical model is shown as solid curve, in part b the line $D^{(1)} = 0$ is shown to visualize the multiple fixed points; after [23]

For illustration we show in Fig. 2 the estimated functions of the drift terms obtained from the analysis of the data of Fig. 1.

The procedure is as follows:

- The data is represented in a d-dimensional phase space
- The phase space is partitioned into a set of finite, but small d-dimensional volume elements
- For each bin (denoted by α), located at the point $\mathbf{x}_\alpha$ of the partition we consider the quantity

$$\mathbf{x}(t_j + \tau) = \mathbf{x}(t_j) + \mathbf{D}^{(1)}(\mathbf{x}(t_j), t_j)\tau + \sqrt{\tau} g(\mathbf{x}(t_j)\boldsymbol{\Gamma}(t_j) \tag{60}$$

Thereby, the considered points $\mathbf{x}(t_i)$ are taken from the bin located at $\mathbf{x}_\alpha$. Since we consider time dependent processes this has to be done for each time step t_j separately

- **Estimation of the drift vector:**
  The drift vector assigned to the bin located at $\mathbf{x}_\alpha$ is determined as the small τ- limit

$$\mathbf{D}^{(1)}(\mathbf{x}, t) = \lim_{\tau\to 0} \frac{1}{\tau}\mathbf{M}(\mathbf{x}, t, \tau) \tag{61}$$

of the conditional moment

$$\mathbf{M}^{(1)}(\mathbf{x}_\alpha, t_j, \tau) = \frac{1}{N_\alpha}\sum_{\mathbf{x}(t_j)\in\alpha}[\mathbf{x}(t_j+\tau)-\mathbf{x}(t_j)] \tag{62}$$

The sum is over all N_α points contained in the bin α.
Proof: The drift vector assigned to the bin α located at $\mathbf{x}_\alpha$ is approximated by the conditional expectation value

$$\mathbf{M}^{(1)}(\mathbf{x}_\alpha, t, \tau) = \tau\frac{1}{N_\alpha}\sum_{\mathbf{x}(t_j)\in\alpha}\mathbf{D}^{(1)}(\mathbf{x}_j, t_j) + \sqrt{\tau}\frac{1}{N_\alpha}\sum_{\mathbf{x}(t_j)\in\alpha} g(\mathbf{x}(t_j)\boldsymbol{\Gamma}(t_j) \tag{63}$$

Thereby, the sum is over all points $\mathbf{x}(t_j)$ located in the bin assigned to $\mathbf{x}_\alpha$. Assuming that $\mathbf{D}^{(1)}(\mathbf{x}, t)$ and $g(\mathbf{x}, t)$ do not vary significantly over the bin, the second contribution drops out since

$$\frac{1}{N_\alpha}\sum_{\mathbf{x}_j\in\alpha}\boldsymbol{\Gamma}(t_j) \to 0 \tag{64}$$

□

- **Estimation of the Diffusion matrix**
  The diffusion matrix can be estimated by the small τ-limit

$$\mathbf{D}^{(2)}(\mathbf{x}, t) = \lim_{\tau\to 0}\frac{1}{\tau}\mathbf{M}^{(2)}(\mathbf{x}, t, \tau) \tag{65}$$

of the conditional second moment

$$\mathbf{M}^{(2)}(\mathbf{x}_\alpha, t, \tau) = \frac{1}{N_\alpha}\sum_j\{[\mathbf{x}(t_j+\tau)-\mathbf{x}(t_j)] - \tau\mathbf{D}^{(1)}(\mathbf{x}_j, t_j)\}^2 \tag{66}$$

Proof: We consider the quantity

$$\mathbf{M}^{(2)}(\mathbf{x}_\alpha, t, \tau) = \tau\frac{1}{N_\alpha}\sum_{\mathbf{x}(t_k)\in\alpha}\sum_k g(\mathbf{x}(t_k), t_k)\boldsymbol{\Gamma}(t_k)g(\mathbf{x}(t_j)\boldsymbol{\Gamma}(t_j) \tag{67}$$

If the bin size is small comparable to the scale, where the matrix $g(\mathbf{x}, t)$ varies significantly, we can replace $g(\mathbf{x}(t_k), t_k)$ by $g(\mathbf{x}_\alpha, t_k)$ such that

$$\begin{aligned}\mathbf{M}^{(2)}(\mathbf{x}_\alpha, t, \tau) &= g(\mathbf{x}_\alpha, t_k)\left[\frac{1}{N_\alpha}\sum_{\mathbf{x}_j\in\alpha}\sum_{\mathbf{x}_k\in\alpha}\boldsymbol{\Gamma}(t_k)\boldsymbol{\Gamma}(t_j)\right]g^T(\mathbf{x}_\alpha, t_k)\\ &= \tau g(\mathbf{x}_\alpha, t_k)g^T(\mathbf{x}_\alpha, t_k)\end{aligned} \tag{68}$$

Thereby, we have used the assumption of the statistical independence of the fluctuations

$$\frac{1}{N_\alpha}\sum_{\mathbf{x}(t_j)\in\alpha}\sum_{\mathbf{x}(t_k)\in\alpha}\Gamma(t_k)\Gamma(t_j) = \delta_{kj}E \tag{69}$$

□

- **Higher order cumulants**
  In a similar way one may estimate higher cumulants M^n, which in the small time limit converge to the so-called Kramers–Moyal coefficients. The estimation of these quantities allows to answer the question whether the noise sources actually are Gaussian distributed random variables

Technical Aspects The above procedure of estimating drift vector and diffusion matrix explicitly shows the properties, which limit the accuracy of the determined quantities.

First of all, the bin size influences the results. The bin size should allow for a resonable number of events such that the sums converge, however, it should be reasonable small in order to allow for a accurate resolution of drift vector and diffusion matrix.

Second, the data should allow for the estimation of the conditional moments in the limit $\lim_{\tau\to 0}$ [25,26]. Here, a finite Markov–Einstein coherence length may cause problems. Furthermore, measurement noise can spoil the possibility of performing this limit.

From the investigation of the Fokker–Planck equation much is known on the τ dependence of the conditional moments. This may be used for further improved estimations, as has been discussed in [27]. Furthermore, as we shall discuss below, extended estimation procedures have been devised, which overcome the problems related with the small τ-limit.

Lévy Processes A procedure to analyze Lévy processes along the same lines has been proposed in [28]. An important point here is the determination of the Lévy parameter α.

Selfconsistency After determining the drift vector and the characteristics of the noise sources from data, it is straightforward to synthetically generate data sets by iterating the corresponding stochastic evolution equations. Subsequently, their statistical properties can be compared with the properties of the real world data. This yields a selfconsistent check of the obtained results.

Estimation of Drift and Diffusion from Sparsely Sampled Time Series

As we have discussed, the results from an analysis of data sets can be reconsidered selfconsistently. This fact can be used to extend the procedure to data sets with insufficient amount of data or sparsely sampled time series, for which the estimation of conditional moments $M^{(i)}(\mathbf{x}, t, \tau)$ and the subsequent limiting procedure $\tau \to 0$ can not be performed accurately.

In this case, one may proceed as follows. In a first step one obtains a zeroth order approximation of drift vector $\mathbf{D}^{(1)}(\mathbf{x})$ and diffusion matrix $\mathbf{D}^{(2)}(\mathbf{x})$. Based on this estimate one performes, in a second step, a suitable ansatz for the drift vector and the diffusion matrix containing a set of free parameters σ

$$\mathbf{D}^{(1)}(\mathbf{x}, \sigma), \mathbf{D}^{(2)}(\mathbf{x}, \sigma) \tag{70}$$

defining a class of Langevin-equations. Each Langevin equation defines a joint probability distribution

$$f(\mathbf{x}_n, t_n; \ldots \mathbf{x}_1, t_1; \sigma) \tag{71}$$

This joint probability distribution can be compared with the experimental one, $f(\mathbf{x}_n, t_n; \ldots \mathbf{x}_1, t_1; \text{exp})$. The best representative of the class of Langevin equations for the reconstruction of experimental data is then obtained by minimizing a suitably defined distance between the two distributions:

$$\text{Dist}\{f(\mathbf{x}_n, t_n; \ldots \mathbf{x}_1, t_1; \sigma) - f(\mathbf{x}_n, t_n; \ldots \mathbf{x}_1, t_1; \text{exp})\} = \text{Min} \tag{72}$$

A reasonable choice is the so-called Kullback–Leibler distance between two distributions, defined as

$$K = \int \mathrm{d}f\,(\mathbf{x}_n, t_n; \ldots; \mathbf{x}_1, t_1; \text{exp}) \cdot \ln \frac{f(\mathbf{x}_n, t_n; \ldots; \mathbf{x}_1, t_1; \text{exp})}{f(\mathbf{x}_n, t_n; \ldots; \mathbf{x}_1, t_1; \sigma)} \tag{73}$$

Recently, it has been shown how the iteration procedure can be obtained from maximum likelihood arguments. For more details, we refer the reader to [29,30]. A technical question concerns the determination of the minimum. In [31] the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm for constraint problems has been used for the solution of the optimization problem.

Applications: Processes in Time

The method outlined in the previous section has been used for revealing nonlinear deterministic behavior in a variety of problems ranging from physics, meteorology, biology to medicine. In most of these cases, alternative procedures with strong emphasis on deterministic features have only been partly successful due to their inappropriate treatment of dynamical fluctuations. The following list (with some exemplary citations) gives an overview on the investigated phenomena, which range from technical applications over many particle systems to biological and geophysical systems.

- Chatter in cutting processes [32,33]
- Identification of bifurcations towards drifting solitary structures in a gas-discharge system [34,35,36,37,38]
- Electric circuits [17,39,40]
- Wind energy convertors [41,42,43,44]
- Traffic flow [45]
- Inverse grading of granular flows [46]
- Heart rythms [47,48,49,50]
- Tremor data [39]
- Epileptic brain dynamics [51]
- Meteorological data like El NINO [52,53,54]
- Earth quake prediction [15]

The main advantage of the stochastic data analysis method is its independence on modeling assumptions. It is purely data driven and is based on the mathematical features of Markov processes. As mentioned above these properties can be verified and validated selfconsistently.

Before we proceed to consider some exemplary applications we would like to add the following comment. The described analysis method cleans data from dynamical and measurement noise and provides the drift vector field, i. e. one obtains the underlying deterministic dynamical system. In turn, this system can be analyzed by the methods of nonlinear time series analysis: One can determine proper embedding, Ljapunov-exponents, dimensions, fixed points, stable and unstable limit cycles etc. [55]. We want to point out that the determination of these quantities from uncleaned data usually is flawed by the presence of dynamical noise.

Synthetic Data: Potential Systems, Limit Cycles, Chaos

The above method of disentangling drift and dynamical noise has been tested for synthetically generated data. The investigated systems include the pitchfork bifurcation, the Van der Pol oscillator as an example of a nonlinear oscillator, as well as the noisy Lorenz equations as an example of a system exhibiting chaos [56,57]. Furthermore, it has been shown how one can analyze processes which additionally contain a time periodic forcing [58]. This is of high interest for analyzing systems exhibiting the phenomena of stochastic resonance. Quite recently, stochastic systems with time delay have been considered [59,60].

The results of these investigations may be summarized as follows: Provided there is enough data and the data is well sampled, it is possible to extract the underlying deterministic dynamics and the strength of the fluctuations accurately. Figure 3 summarizes what can be achieved for the example of the noisy Lorenz model. For a detailed discussion we refer the reader to [57].

Noisy Circuits

Next, we present the application of the method to data sets from experimental systems. As first example, a chaotic electric circuit has been chosen. Its dynamics is formed by a damped oscillator with nonlinear energy support and additional dynamic noise terms. In this case, well defined electric quantities are measured for which the dynamic equations are known. The measured time series are analyzed according to the numerical algorithm described above. Afterwards, the numerically determined results and the expected results according to the system's equations are compared.

The dynamic equations of the electric circuit are given by the following equations, where the deterministic part is known as Shinriki oscillator [61]:

$$\dot{X}_1 = \left(-\frac{1}{R_{NIC}C_1} - \frac{1}{R_1C_1}\right)X_1 - \frac{f(X_1 - X_2)}{C_1} + \frac{1}{R_{NIC}C_1}\Gamma(t) \tag{74}$$

$$= g_1(X_1, X_2) + h_1\Gamma(t) \tag{75}$$

$$\dot{X}_2 = \frac{f(X_1 - X_2)}{C_2} - \frac{1}{R_3C_2}X_3 = g_2(X_1, X_2, X_3) \tag{76}$$

$$\dot{X}_3 = -\frac{R_3}{L}(X_2 - X_3) = g_3(X_2, X_3) \tag{77}$$

X_1, X_2 and X_3 denote voltage terms, R_i are values of resistors, L and C stand for inductivity and capacity values. The function $f(X_1 - X_2)$ denotes the characteristic of the nonlinear element. The quantities X_i, characterizing the stochastic variable of the Shinriki oscillator with dynamical noise, were measured by means of a 12 bit A/D converter. Our analysis is based on the measurement of 100.000 data points [39]. The attractor of the noise free dynamics is shown in Fig. 4.

The measured 3-dimensional time series were analyzed as outlined above. The determined deterministic dynamics – expressed by the deterministic part of the evolution equations – corresponds to a vector field in the

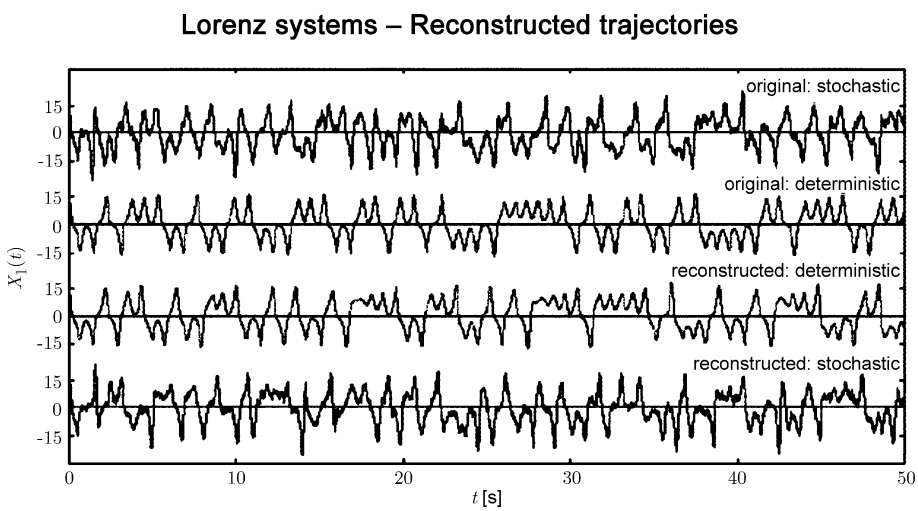

Fluctuations, Importance of: Complexity in the View of Stochastic Processes, Figure 3
Time series of the stochastic Lorenz equation from top to bottom: a) Original time series b) Deterministic time series c) Time series obtained from an integration of the reconstructed vector field d) Reconstructed time series including noise. For details cf. [57]

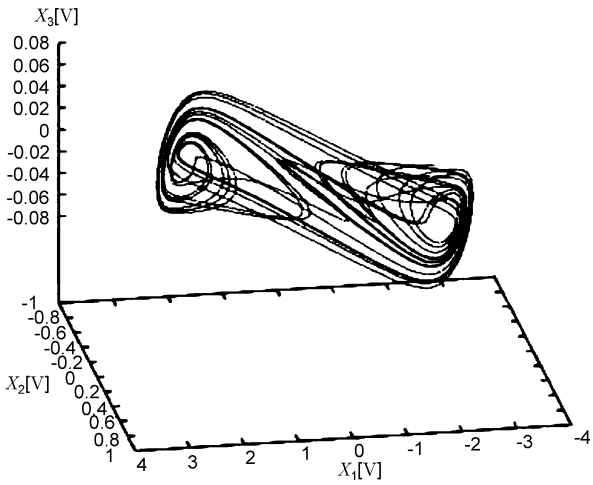

Fluctuations, Importance of: Complexity in the View of Stochastic Processes, Figure 4
Trajectory for the Shinriki oscillator in the phase space without noise; after [40]

three dimensional state space. For presentation of the results, cuts of lower dimension have been generated. Part a) of Fig. 5 illustrates the vector field $(g_1(x_1, x_2), g_2(x_1, x_2, x_3 = 0))$ of the reconstructed deterministic parts affiliated with (75), (76). Furthermore, the one-dimensional curve $g_1(x_1, x_2 = 0)$ is drawn in part b). Additionally to the numerically determined results found by data analysis the expected vector field and curve (75), (76) are shown for comparison. A good agreement can be recognized.

Based on the reconstructed deterministic functions it is possible to reconstruct also the noisy part from the data set, see Subsect. "Direct Estimation of Stochastic Forces". This has been performed for the three dimensionally embedded data, as well for the case of two dimensional embedding. From these reconstructed noise data, the autocorrelation was estimated. As shown in Fig. 6 correlated noise is obtained for wrong embedding indicating the violation of Markovian properties. In fact such an approach can be used to verify the correct embedding of nonlinear noisy dynamical systems. We emphasize that, provided sufficient data is available, this check of correct embedding can also be performed locally in phase space to find out where for the corresponding deterministic system crossing of trajectories take place. This procedure can be utilized to find the correct local embedding of data.

The electronic circuit of the Shinriki oscillator has also been investigated with two further perturbations. In [17] the reconstruction of the deterministic dynamics by the presence of additional measurement noise has been ad-

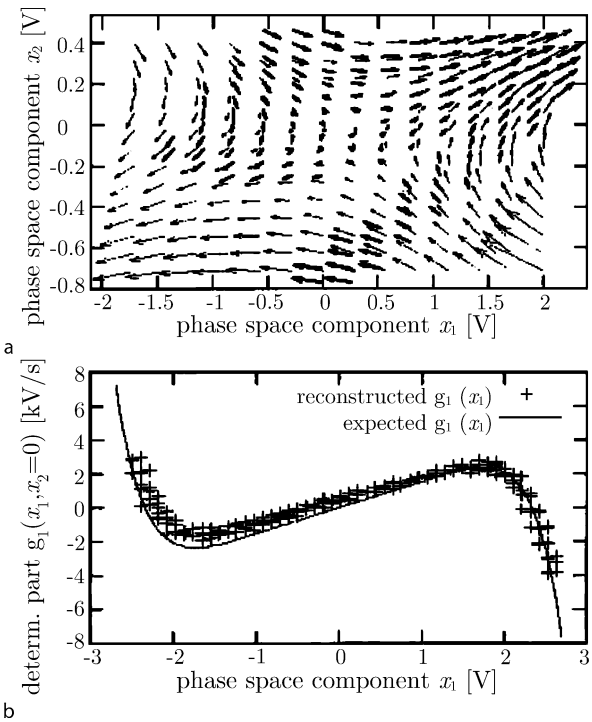

Fluctuations, Importance of: Complexity in the View of Stochastic Processes, Figure 5
Cuts of the function $D^{(1)}(x)$ reconstructed from experimental data of the electric circuit in comparison with the expected functions according to the known differential (Eq. (75), (76)). In part a the cut $g_1(X_1, X_2)$, $g_2(X_1, X_2, X_3 = 0)$ is shown as a two-dimensional vector field. *Thick arrows* represent values determined by data analysis, *thin arrows* represent the theoretically expected values. In areas of the state space where the trajectory did not show up during the measurement no estimated values for the functions are obtained. Figure b shows the one dimensional cut $g_1(X_1, X_2 = 0)$. *Crosses* represent values estimated numerically by data analysis. Additionally, the affiliated theoretically curve is printed as well; after [39]

dressed. In [40] the Langevin noise has been exchanged by a high frequency periodic source, as shown in Fig. 7. Even for this case reasonable (correct) estimations of the deterministic part can be achieved.

Many Particle Physics – Traffic Flow

Far from equilibrium interacting many particle systems exhibit collective macroscopic phenomena like pattern formation, which can be described by the concept of order parameters. In the following we shall exemplify for the case of traffic flow how complex behavior can be analyzed by means of the proposed method leading to stochastic equations for the macroscopic description of the system.

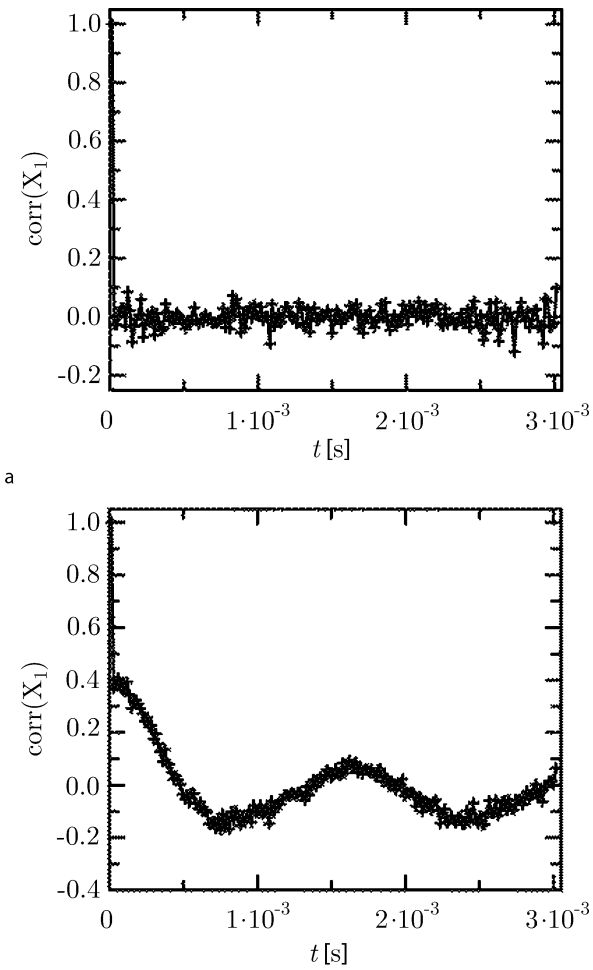

Fluctuations, Importance of: Complexity in the View of Stochastic Processes, Figure 6
Autocorrelation function of reconstructed dynamical noise, a correctly three dimensional embedded showing δ-correlated noise, b the projected dynamics in the twodimensional phase space $X_1(t)$ and $X_3(t)$, showing finite time correlations; after [17]

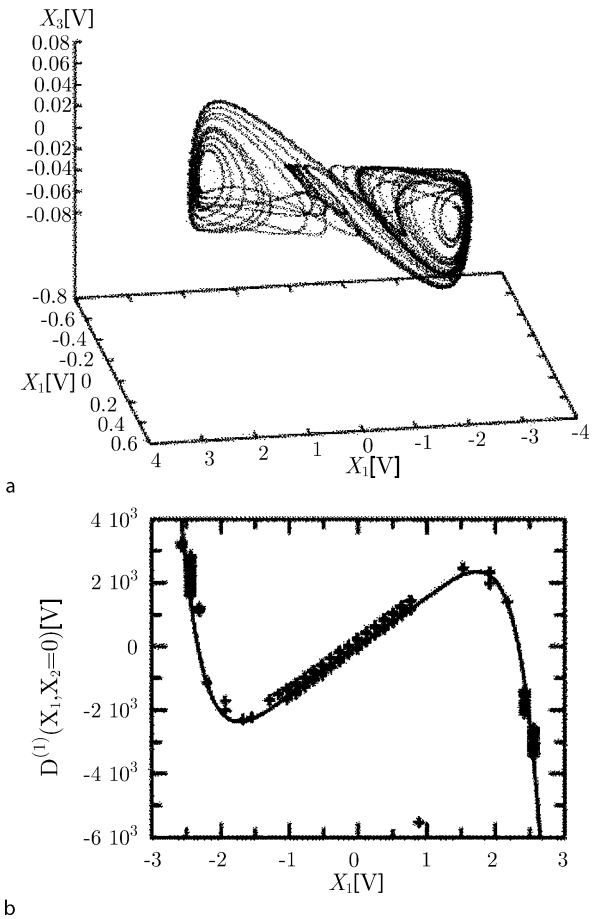

Fluctuations, Importance of: Complexity in the View of Stochastic Processes, Figure 7
a Trajectory for the Shinriki oscillator in the phase space with a sinusoidal force. b The corresponding trajectory in the phase space, compare Fig. 5b; after [40]

Traffic flow certainly is a collective phenomena, for which a huge amount of data is available. Measured quantities are velocity v and current $q = \rho v$ of cars passing a fixed line on the highway. Theoretical models of traffic flow are based on the so-called fundamental diagram, which is a type of material law for traffic flow relating current and velocity of the traffic flow

$$q = Q(v) \tag{78}$$

The special form of this relation has been much under debate in recent years.

It is tempting to describe the dynamics by the following set of stochastic difference equation

$$\begin{aligned} v_{N+1} &= G(v_N, q_N) + \xi_N \\ q_{N+1} &= F(v_N, q_N) + \eta_N \end{aligned} \tag{79}$$

Here, v_{N+1}, $q_N + 1$ are velocity and current of the $N + 1$ car traversing the line. ξ_N and η_N are noise terms with zero mean, which may depend on the variables u and q. The drift vector field $\mathbf{D}^1 = [D_1^1(v, q), D_2^1(v, q)]$ has been determined in [45]. We point out that the meaning of the drift vector field does not depend on the assumption of ideal noise sources, as has been discussed above Subsect. "Estimating the Short Time Propagator" and Subsect. "Noisy Circuits". The obtained drift vector field is depicted in Fig. 8.

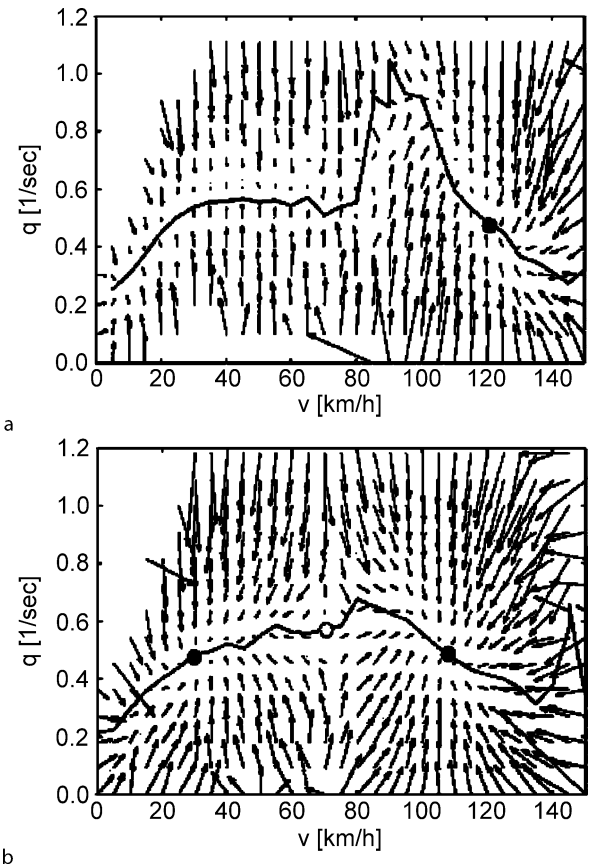

Fluctuations, Importance of: Complexity in the View of Stochastic Processes, Figure 8
Deterministic dynamics of the twodimensional traffic states (q, v) given by the drift vector for a right lane of freeway with vans and b all three lanes. Bold dots indicate stable fixed points and open dots saddle points, respectively; after [45]

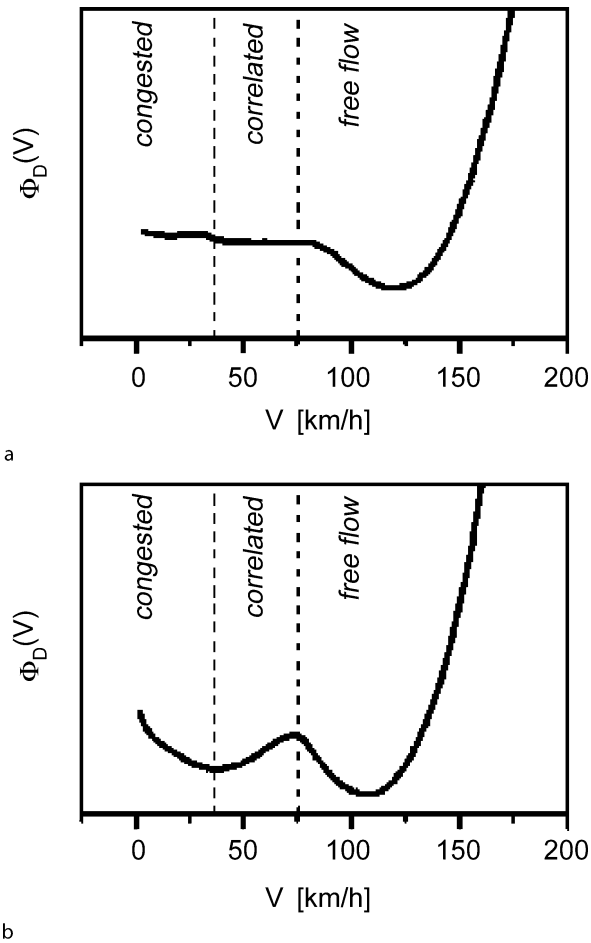

Fluctuations, Importance of: Complexity in the View of Stochastic Processes, Figure 9
Corresponding potentials for the deterministic dynamics given by the drift coefficients of a one dimensional projection of the results in Fig. 8 [45]

The phase space yields interesting behavior. For the traffic data involving all three lanes there are three fixed points, two sinks separated by a saddle point. Furthermore, the arrows representing the drift vector field indicate the existence of an invariant manifold,

$$q = Q(v) \tag{80}$$

which has to be interpreted as the above mentioned fundamental diagram of traffic flow. It is interesting to see differences caused by separation of the traffic dynamics into that of cars and that of vans. This has been roughly achieved by considering different lanes of the highway. In Figs. 8a and 9a the dynamics of the right lane caused by vans is shown. It can clearly be seen that up to a speed of about 80 km/h a meta stable plateau is present corresponding to a quasi interaction free dynamics of vans. Up to now the information contained in the noise terms has not led to clear new insights.

For the discussion of further examples we refer the reader to the literature. In particular, we want to mention the analysis of segregational dynamics of single particles with different sizes in avalanches [46], which can be treated along similar lines .

Applications: Processes in Scale

In the following we shall consider complex behavior in scale as discussed in the introduction by the method of stochastic processes.

In order to statistically describe scale complexity in a comprehensive way, one has to study joint probability

distributions

$$f(q_N, l_N; q_{N-1}, l_{N-1}; \ldots; q_1, l_1; q_0, l_0) \tag{81}$$

of the local measure q at multiple scales l_i. To grasp relations across several scales (like given by "coherent" structures) all q_i tuples are taken at common locations x. In the case of statistical independency of q at different scales l_i, this joint probability density factorizes:

$$f(q_N, l_N; q_{N-1}, l_{N-1}; \ldots; q_1, l_1; q_0, l_0) = f(q_N, l_N) \ldots f(q_0, l_0) \,. \tag{82}$$

In this case multifractal scaling (6) is a sufficient characterization of systems with scale complexity, provided the scaling property is given. If there is no factorization, one can use the joint probability distribution Eq. (81) to define a stochastic process in scale. Thus, one can identify the scale l with time t and try to obtain a representation of the spatial disorder in the form of a stochastic process similar to a stochastic process in time (see Subsect. "Statistical Description of Stochastic Processes" and Subsect. "Finite Time Propagators").

For such problems our method has been used as an alternative description of multifractal behavior. The present method has the advantage to relate the random variables across different scales by introducing a conditional probability distribution or, in fact, a two scale probability distribution. Scaling properties are no prerequisite. Provided that the Markovian property holds, which can be tested experimentally, a complete statistical characterization of the process across scale is obtained. The method has been used to characterize the complexity of data sets for the following systems (with some exemplary citations):

- Turbulent flows [20,62,63,64]
- Passive scalar in turbulent flows [21]
- Financial data [65,66,67,68]
- Surface roughness [69,70,71,72,73,74]
- Earthquakes [15]
- Cosmic background radiation [75]

The stochastic analysis of scale dependent complex system aims to achieve a n-scale characterization, from which the question arose, whether it will be possible to derive from these stochastic processes methods for generating synthetic time series.

Based on the knowledge of the n-scale statistics a way to estimate the n-point statistics has been proposed, which enables to generate synthetic time series with the same n-scale statistics [76]. It is even possible to extend given data sets, an interesting subject for forecasting. Other approaches has been proposed in [70,77].

Turbulence and Financial Market Data

In the following we present results from investigations of fully developed turbulence and from data of the financial market. The complexity of turbulent fields still has not been understood in detail. Although the basic equations, namely the Navier Stokes equations, are known for more than 150 years, a general solution of these equations for high Reynolds numbers, i. e. for turbulence, is not known. Even with the use of powerful computers no rigorous solutions can be obtained. Thus for a long time there has been the challenge to understand at least the complexity of an idealized turbulent situation, which is taken to be isotropic and homogeneous. The main problem is to formulate statistical laws based on the treatment of the deterministic evolution laws of fluid dynamics. A first approach is due to Kolmogorov [78], who formulated a phenomenological theory characterizing properties of turbulent fluid motions by statistical quantities. The central observable of Kolmogorov's theory is the so-called longitudinal velocity increment (which we label here with q) of a turbulent velocity field $\mathbf{u}(\mathbf{x}, t)$ defined according to:

$$\mathbf{q}_{\mathbf{x}}(\mathbf{l}, t) = \frac{\mathbf{l}}{l} \cdot [\mathbf{u}(\mathbf{x} + \mathbf{l}, t) - \mathbf{u}(\mathbf{x}, t)] \,. \tag{83}$$

A statistical description is given in terms of the probability distribution

$$f(q, \mathbf{l}, t, \mathbf{x}) = \langle \delta(q - q_{\mathbf{x}}(\mathbf{l}, t)) \rangle \,. \tag{84}$$

For stationary, homogeneous and isotropic turbulence, this probability distribution is independent of the reference point $\mathbf{x}$, time t, and, due to isotropy, only depends on the scale $l = |\mathbf{l}|$. As a consequence, the central statistical quantity is the probability distribution $f(q, l)$. Turbulent fields have been considered from the viewpoint of selfsimilarity addressing scaling behavior of the probability distribution $f(q, l)$ and their nth order moments, the so-called structure functions $\langle q^n \rangle$. Multifractal scaling properties, mentioned already in Eq. (7), of the velocity increments for turbulence are identical to the well known intermittency problem[1], which manifests itself in the occurrence of heavy tailed statistics, i. e. an unexpected high probability

[1] Here it should be noted that the term "intermittency" is used frequently in physics for different phenomena, and may cause confusions. This turbulent intermittency is not equal to the intermittency of chaos. There are also different intermittency phenomena introduced for turbulence. There is this intermittency due to the nonlinear scaling, there is the intermittency of switches between turbulent and laminar flows for non local isotropic fully developed turbulent flows, there is the intermittency due to the statistics of small scale turbulence which we discuss here as heavy tailed statistics.

of extreme events. Kolmogorov and Oboukhov proposed the so-called intermittency correction [79]

$$\langle q(l,x)^n \rangle = l^{\xi_n} \text{ with } \xi_n = \frac{n}{3} - \mu \frac{n(n-3)}{18} \text{ and } n > 2 \tag{85}$$

with $0.25 < \mu < 0.5$ (for further details see [80]). The form of ξ_n has been heavily debated during the last decades. For isotropic turbulence the central issue is to reveal the mechanism, which leads to this anomalous statistics (see [80,81]).

A completely different point of view of the properties of turbulent fields is gained by interpreting the probability distribution as the distribution of a random process in scale l [62,63]. It is tempting to hypothesize the existence of a stochastic process, see Subsect. "Stochastic Data Analysis"

$$q(l + \mathrm{d}l) = q(l) + N(q,l)\mathrm{d}l + g(q,l)\mathrm{d}W(l)\,, \tag{86}$$

where $\mathrm{d}W(l)$ is an increment of a random process. This type of stochastic equation indicates how the velocity increment of a snapshot of the turbulent field changes as a function of scale l. In this respect, the process $q(l)$ can be considered to be a stochastic cascade process in "*time l*".

This concept of complexity in scale can be carried over to other systems like the roughness of surfaces or financial data. In the latter case the scale variable l is replaced by the time distance or time scale τ.

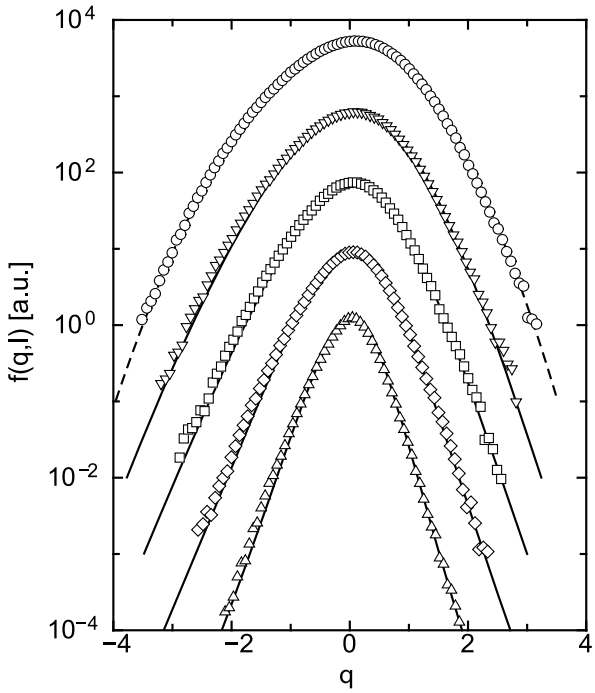

Fluctuations, Importance of: Complexity in the View of Stochastic Processes, Figure 10
Comparison of the numerical solution of the Fokker–Planck equation (*solid lines*) for the pdfs $f(q(x), l)$ with the pdfs obtained directly from the experimental data (bold symbols). The scales l are (from top to bottom): $l = L_0$, $0.6L_0$, $0.35L_0$, $0.2L_0$ and $0.1L_0$. The distribution at the largest scale L_0 was parametrized (*dashed line*) and used as initial condition for the Fokker–Planck equation (L_0 is the correlation length of the turbulent velocity signal). The pdfs have been shifted in vertical direction for clarity of presentation and all pdfs have been normalized to their standard deviations; after [20]

Anomalous Statistics

A direct consequence of multifractal scaling related with nonlinear behavior of the scaling exponents ξ_n is the fact that the shape of the probability distribution $f(q,l)$ has to change its shape as a function of scale. A selfsimilar behavior of the form (4) would lead to fractal scaling behavior, as outlined in Eq. (5). Using experimental data from a turbulent flow this change of the shape of the pdf becomes obvious. In Fig. 10 we present $f(q,l)$ for a data set measured in the central line of a free jet with Reynolds number of $2.7 \cdot 10^4$, see [20]. Note for large scales ($l \approx L_0$) the distributions become nearly Gaussian. As the scale is decreased the probability densities become more and more heavy tailed.

Quite astonishingly, the anomalous statistical features of data from the financial market are similar to the just discussed intermittency of turbulence [82]. The following analysis is based on a data set $Y(t)$, which consists of 1 472 241 quotes for US dollar-German Mark exchange rates from the years 1992 and 1993. Many of the features we will discuss here are also found in other financial data like for instance quotes of stocks, see [83]. A central issue is the understanding of the statistics of price changes over a certain time interval τ which determines losses and gains. The changes of a time series of quotations $Y(t)$ are commonly measured by returns $r(\tau,t) := Y(t+\tau)/Y(t)$, logarithmic returns or by increments $q(\tau,t) := Y(t+\tau) - Y(t)$ [84]. The moments of these quantities often exhibit power law behavior similar to the just discussed Kolmogorov scaling for turbulence, cf. [85,86,87]. For the probability distributions one additionally observes an increasing tendency to heavy tailed probability distributions for small τ (see Fig. 11). This represents the high frequency dynamics of the financial market. The identification of the underlying process leading to these heavy tailed probability density functions of price

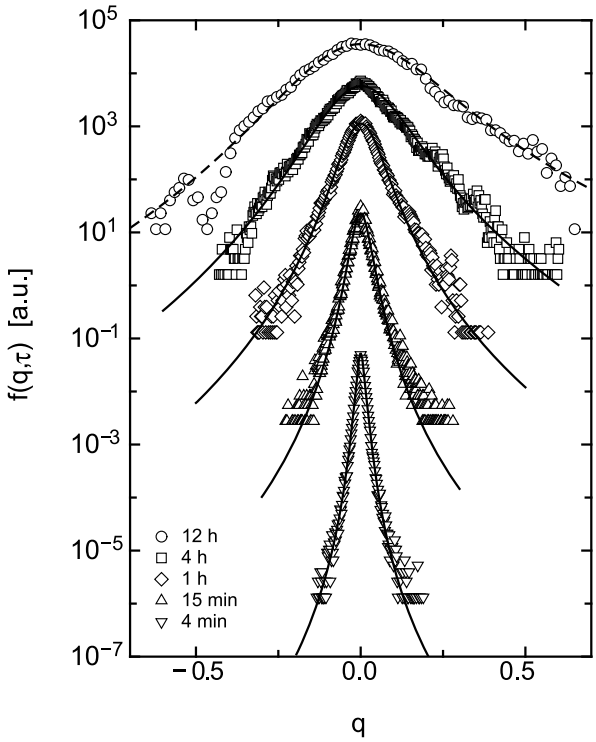

Fluctuations, Importance of: Complexity in the View of Stochastic Processes, Figure 11
Probability densities (*pdf*) $f(q(t), \tau)$ of the price changes $q(\tau, t) = Y(t+\tau) - Y(t)$ for the time delays $\tau = 5120, 10240, 20480, 40960\,s$ (from bottom to top). Symbols: results obtained from the analysis of middle prices of bit-ask quotes for the US dollar-German Mark exchange rates from October 1st, 1992 until September 30th, 1993. *Full lines*: results from a numerical iteration of the Fokker–Planck equation (95); the probability distribution for $\tau = 40960s$ (*dashed lines*) was taken as the initial condition. The pdfs have been shifted in vertical direction for clarity of presentation and all pdfs have been normalized to their standard deviations; after [66]

changes is a prominent puzzle (see [86,87,88,89,90,91]), like it is for turbulence.

Stochastic Cascade Process for Scale

The occurrence of the heavy tailed probability distributions on small scales will be discussed as a consequence of stochastic process evolving in scale, using the above mentioned methods. Guided by the finding that the statistics changes with scale, as shown in Figs. 10 and 11, we consider the evolution of the quantity $q(l, x)$, or $q(\tau, t)$ with the scale variable l or τ, respectively.

For a single fixed scale l we get the scale dependent disorder by the statistics of $q(l, x)$. The complete stochastic information of the disorder on all length scales is given by the joint probability density function

$$f(q_1, \ldots, q_n)\,, \tag{87}$$

where we set $q_i = q(l_i, x)$. Without loss of generality we take $l_i < l_{i+1}$. This joint probability may be seen in analogy to joint probabilities of statistical mechanics (thermodynamics), describing in the most general way the occupation probabilities of the microstates of n particles, where q is a six-dimensional phase state vector (space and momentum).

Next, the question is, whether it is possible to simplify the joint probability by conditional probabilities:

$$f(q_1, \ldots, q_n) = p(q_1|q_2, \ldots, q_n) \cdot p(q_2|q_3, \ldots, q_n) \ldots p(q_{n-1}|q_n)\, f(q_n)\,, \tag{88}$$

where the multiple conditioned probabilities are given by

$$p(q_i \mid q_{i+1}, \ldots, q_n) = p(q_i \mid q_{i+1})\,. \tag{89}$$

Eq. (89) is nothing else than the condition for a Markov process evolving from state q_{i+1} to the state q_i, i. e. from scale l_{i+1} to l_i as it has been introduced above, see Eq. (35).

A "single particle approximation" would correspond to the representation:

$$f(q_1, \ldots, q_n) = f(q_1) f(q_2) \ldots f(q_n)\,. \tag{90}$$

According to Eqs. (89) and (90), Eq. (91) holds if $p(q_i|q_{i+1}) = f(q_i)$. Only for this case, the knowledge of $f(q_i)$ is sufficient to characterize the complexity of the whole system. Otherwise an analysis of the single scale probability distribution $f(q_i)$ is an incomplete description of the complexity of the whole system. This is the deficiency of the approach characterizing complex structures by means of fractality or multiaffinity (cf. for multiaffinity [92], for turbulence [80,81], for financial market [85]). The scaling analysis of moments as indicated for turbulence in Eq. (85) provides a complete knowledge of any joint n-scale probability density only if Eq. (90) is valid.

These remarks underline the necessity to investigate these conditional probabilities, which can be done in a straightforward manner from given experimental or numerical data. For the case of turbulence as well as for financial data we see that $p(q_i|q_{i+1})$ does not coincide with $f(q_i)$, as it is shown for turbulence data in Fig. 12. If $p(q_i|q_{i+1}) = f(q_i)$ no dependency on q_{i+1} could be detected.

The next point of interest is whether the Markov properties are fulfilled. Therefore doubly conditioned probabil-

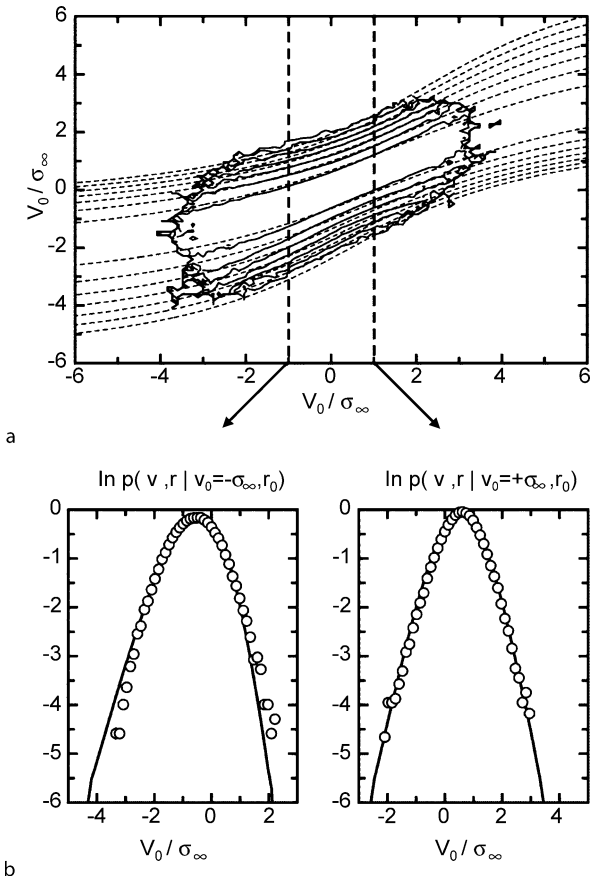

Fluctuations, Importance of: Complexity in the View of Stochastic Processes, Figure 12

Comparison of the numerical solution of the Fokker–Planck equation for the conditional pdf $p(q, l|q_0, l_0)$ denoted in this figure as $p(v, r|v_0, r_0)$ with the experimental data. a: Contour plots of $p(v, r|v_0, r_0)$ for $r_0 = L$ and $r = 0.6L$, where L denotes the integral length. *Dashed lines*: numerical solution of (94), *solid lines*: experimental data. b and c: Cuts through $p(v, r|v_0, r_0)$ for $v_0 = +\sigma_\infty$ and $v_0 = -\sigma_\infty$ respectively. *Open symbols*: experimental data, solid lines: numerical solution of the Fokker–Planck equation; after [20]

ities were extracted from data and compared to the single conditioned ones. For financial data as well as for turbulence we found evidence that the Markov property is fulfilled if the step size is larger than a Markov–Einstein length [20,66]. For turbulence it could be shown that the Markov–Einstein length coincides with the Taylor length marking the small scale end of the inertial range [14]. (The extensive discussion of the analysis of financial and turbulent data can be found in [20,65,66,93]).

Based on the fact that the multiconditioned probabilities are equal to the single conditioned probabilities, and taking this as a verification of Markovian properties we can proceed according to Subsect. "Estimating the Short Time Propagator" and estimate from given data sets the stochastic equations underlying the cascade process. The evolution of the conditional probability density $p(q, l|q_0, l_0)$ starting from a selected scale l_0 follows

$$-l\frac{\partial}{\partial l}\,p(q, l|q_0, l_0) = \sum_{k=1}^{\infty}\frac{1}{k!}\left(-\frac{\partial}{\partial q}\right)^k D^{(k)}(q, l)\,p(q, l|q_0, l_0)\,. \tag{91}$$

(The minus sign on the left side is introduced, because we consider processes running to smaller scales l, furthermore we multiply the stochastic equation by l, which leads to a new parametrization of the cascade by the variable $\ln(1/l)$, a simplification for a process with scaling law behavior of its moments.) This equation is known as the Kramers–Moyal expansion [8]. As outlined in Subsect. "Finite Time Propagators" and Subsect. "Estimating the Short Time Propagator", the Kramers–Moyal coefficients $D^{(k)}(q, l)$ are now defined as the limit $\Delta l \to 0$ of the conditional moments $M^{(k)}(q, l, \Delta l)$:

$$D^{(k)}(q, l) = \lim_{\Delta l\to 0} M^{(k)}(q, l, \Delta l)\,, \tag{92}$$

$$M^{(k)}(q, l, \Delta l) := \frac{l}{\Delta l}\int_{-\infty}^{+\infty}\left(\tilde{q} - q\right)^k p\left(\tilde{q}, l - \Delta l\,|\,q, l\right)\mathrm{d}\,\tilde{q}. \tag{93}$$

Thus, for the estimation of the $D^{(k)}$ coefficients it is only necessary to estimate the conditional probabilities $p(\tilde{q}, l - \Delta l|q, l)$. For a general stochastic process, all Kramers–Moyal coefficients are different from zero. According to Pawula's theorem, however, the Kramers–Moyal expansion stops after the second term, provided that the fourth order coefficient $D^{(4)}(q, l)$ vanishes. In that case the Kramers–Moyal expansion is reduced to a Fokker–Planck equation:

$$-l\frac{\partial}{\partial l}\,p(q, l|q_0, l_0) = \left\{-\frac{\partial}{\partial q}D^{(1)}(q, l) + \frac{1}{2}\frac{\partial^2}{\partial q^2}D^{(2)}(q, l)\right\}p(q, l|q_0, l_0)\,. \tag{94}$$

$D^{(1)}$ is denoted as drift term, $D^{(2)}$ as diffusion term now for the cascade process. The probability density function $f(q, l)$ has to obey the same equation:

$$-l\frac{\partial}{\partial l}\,f(q, l) = \left\{-\frac{\partial}{\partial q}D^{(1)}(q, l) + \frac{1}{2}\frac{\partial^2}{\partial q^2}D^{(2)}(q, l)\right\}f(q, l)\,. \tag{95}$$

The novel point of our analysis here is that knowing the evolution equation (91), the n-increment statistics $f(q_1, \ldots, q_n)$ can be retrieved as well. Definitely, information like scaling behavior of the moments of $q(l, x)$ can also be extracted from the knowledge of the process equations. Multiplying (91) by q^n and successively integrating over q, an equation for the moments is obtained:

$$-l\frac{\partial}{\partial l}\langle q^n\rangle = \sum_{k=1}^{n}\left(-\frac{\partial}{\partial q}\right)^k \frac{n!}{k!(n-k)!}\langle D^{(k)}(q,l)q^{n-k}\rangle\,. \tag{96}$$

Scaling, i. e. multi affinity as described in Eq. (7), is obtained if $D^{(k)}(q, l) \propto q^k$, see [71,94].

We summarize: By the described procedure we were able to reconstruct stochastic processes in scale directly from given data. Knowing these processes one can perform numerical solutions in order to obtain a selfconsistent check of the procedure(see [20,66]). In Figs. 10, 11 and 12 the numerical solutions are shown by solid (dashed) curves. The heavy tailed structure of the single scale probabilities as well as the conditional probabilities are well described by this approach based on a Fokker–Planck equation. Further improvements can be achieved by optimization procedures mentioned in Subsect. "Estimation of Drift and Diffusion from Sparsely Sampled Time Series".

New Insights

The fact that the complexity of financial market data as well as turbulent data can be expressed by a Markovian process in scale, has the consequence that the conditioned probabilities involving only two different scales are sufficient for the general n scale statistics. This indicates that three or four point correlation are sufficient for the formulation of n-point statistics.

The finding of a finite length scale above which the Markov properties are fulfilled [14] has lead to a new interpretation of the Taylor length for turbulence, which so far had no specific physical meaning.

For financial, as well as, for the turbulent data it has been found that the diffusion term is quadratic in the state space of the scale resolved variable. With respect to the corresponding Langevin equation, the *multiplicative nature of the noise* term becomes evident, which causes heavy tailed probability densities and multifractal scaling. The scale dependency of drift and diffusion terms corresponds to a non-stationary process in scale variables τ and l, respectively. From this point we conclude that a Levy statistics for one fixed scale, i. e. for the statistics of $q(l, x)$ for fixed l can not be an adequate class for the statistical description of financial and turbulent data.

Comparing the maximum of the distributions at small scales in Figs. 10 and 11 one finds a less sharp tip for the turbulence data. This finding is in accordance with a comparably larger additive contribution in the diffusion term $D^{(2)} = a + bq^2$ for turbulence data. Knowing that $D^{(2)}$ has an additive term and quadratic q dependence it is clear that for small q values, i. e. for the tips of the distribution, the additive term dominates. Taking this result in combination with the Langevin equation, we see that for small q values Gaussian noise is present, which leads to a *Gaussian tip* of the probability distribution, as found in Fig. 10.

A further consequence of the additive term in $D^{(2)}$ is that the structure functions as given by Eq. (96) are not independent and a general scaling solution does not exist. This fact has been confirmed by optimizing the coefficients [31]. This *nonscaling behavior of turbulence* seems to be present also for higher Reynolds numbers. It has been found that the additive terms becomes smaller but still remains relevant [96]. The cascade processes encoded in the functions $D^{(1)}$ and $D^{(2)}$ seem to depend on the Reynolds number, which might indicate that turbulence is less universal as commonly thought.

As has been outlined above it is straight forward to extend the analysis to higher dimensions. For the case of *longitudinal and transversal increments* a symmetry for these two different directions of the complex velocity field has been found in the way that the cascade process runs with different speeds. The factor is 3/2 [64,96] and again is not in accordance with the proposed multifractal scaling property of turbulence.

The investigation of the advection of passive scalar in a turbulent flow along the present method has revealed the interesting result that the Markovian properties are fulfilled but that higher order Kramers–Moyal coefficients are nonnegligible [21]. This indicates that for passive scalars non-Gaussian noise is present, which can be attributed to the existence of shock like structures in the distribution of the passive scalar.

Future Directions

The description of complex systems on the basis of stochastic processes, which include nonlinear dynamics, seems to by a promising approach for the future. The challenge will be to extend the understanding to more complicated processes, like Levy processes, processes with no white noise or higher dimensional processes, just to mention some. As it has been shown in this contribution, for these cases it should be possible to derive from precise

mathematical results general methods of data series analysis, too.

Besides the further improve of the method, we are convinced that there is still a wide range of further applications. Advanced sensor techniques enables scientists to collect huge data sets measured with high precision. Based on the stochastic approach we have presented here it is not any more the question to put much efforts into noise reduction, but in contrary the involved noise can help to derive a better characterization and thus a better understanding of the system considered. Thus there seem to be many applications in the inanimate and the animate world, ranging form technical applications over socio-econo systems to biomedical applications. An interesting feature will be the extraction of of higher correlation aspects, like the question of the cause and effect chain, which may be unfolded by asymmetric determinism and noise terms reconstructed from data.

Further Reading

For further reading we suggest the publications [1,2,3,4,8,9,10,11].

Acknowledgment

The scientific results reported in this review have been worked out in close collaboration with many colleagues and students. We mention St. Barth, F. Böttcher, F. Ghasemi, I. Grabec, J. Gradisek, M. Haase, A. Kittel, D. Kleinhans, St. Lück, A. Nawroth, Chr. Renner, M. Siefert, and S. Siegert.

Bibliography

1. Haken H (1983) Synergetics, An Introduction. Springer, Berlin
2. Haken H (1987) Advanced Synergetics. Springer, Berlin
3. Haken H (2000) Information and Self-Organization: A Macroscopic Approach to Complex Systems. Springer, Berlin
4. Kantz H, Schreiber T (1997) Nonlinear Time Series Analysis. Cambridge University Press, Cambridge
5. Yanovsky VV, Chechkin AV, Schertzer D, Tur AV (2000) Physica A 282:13
6. Schertzer D, Larchevéque M, Duan J, Yanovsky VV, Lovejoy S (2001) J Math Phys 42:200
7. Gnedenko BV, Kolmogorov AN (1954) Limit distributions of sums of independent random variables. Addison-Wesley, Cambridge
8. Risken H (1989) The Fokker-Planck Equation. Springer, Berlin
9. Gardiner CW (1983) Handbook of Stochastic Methods. Springer, Berlin
10. van Kampen NG (1981) Stochastic processes in physics and chemistry. North-Holland Publishing Company, Amsterdam
11. Hänggi P, Thomas H (1982) Stochastic processes: time evolution, symmetries and linear response. Phys Rep 88:207
12. Einstein A (1905) Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. Ann Phys 17:549
13. Friedrich R, Zeller J, Peinke J (1998) A Note in Three Point Statistics of Velocity Increments in Turbulence. Europhys Lett 41:153
14. Lück S, Renner Ch, Peinke J, Friedrich R (2006) The Markov Einstein coherence length a new meaning for the Taylor length in turbulence. Phys Lett A 359:335
15. Tabar MRR, Sahimi M, Ghasemi F, Kaviani K, Allamehzadeh M, Peinke J, Mokhtari M, Vesaghi M, Niry MD, Bahraminasab A, Tabatabai S, Fayazbakhsh S, Akbari M (2007) Short-Term Prediction of Mediumand Large-Size Earthquakes Based on Markov and Extended Self-Similarity Analysis of Seismic Data. In: Bhattacharyya P, Chakrabarti BK (eds) Modelling Critical and Catastrophic Phenomena in Geoscience. Lecture Notes in Physics, vol 705. Springer, Berlin, pp 281–301
16. Kolmogorov AN (1931) Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. Math Ann 140:415
17. Siefert M, Kittel A, Friedrich R, Peinke J (2003) On a quantitative method to analyze dynamical and measurement noise. Europhys Lett 61:466
18. Böttcher F, Peinke J, Kleinhans D, Friedrich R, Lind PG, Haase M (2006) On the proper reconstruction of complex dynamical systems spoilt by strong measurement noise. Phys Rev Lett 97:090603
19. Kleinhans D, Friedrich R, Wächter M, Peinke J (2007) Markov properties under the influence of measurement noise. Phys Rev E 76:041109
20. Renner C, Peinke J, Friedrich R (2001) Experimental indications for Markov properties of small scale turbulence. J Fluid Mech 433:383
21. Tutkun M, Mydlarski L (2004) Markovian properties of passive scalar increments in grid-generated turbulence. New J Phys 6:49
22. Marcq P, Naert A (2001) A Langevin equation for turbulent velocity increments. Phys Fluids 13:2590
23. Langner M, Peinke J, Rauh A (2008) A Langevin analysis with application to a Rayleigh-Bénard convection experiment. Exp Fluids (submitted)
24. Wächter M, Kouzmitchev A, Peinke J (2004) Increment definitions for sale-dependent analysis of stochastic data. Phys Rev E 70:055103(R)
25. Ragwitz M, Kantz H (2001) Indispensable finite time corrections for Fokker-Planck equations from time series. Phys Rev Lett 87:254501
26. Ragwitz M, Kantz H (2002) Comment on: Indispensable finite time correlations for Fokker-Planck equations from time series data-Reply. Phys Rev Lett 89:149402
27. Friedrich R, Renner C, Siefert M, Peinke J (2002) Comment on: Indispensable finite time correlations for Fokker-Planck equations from time series data. Phys Rev Lett 89:149401
28. Siegert S, Friedrich R (2001) Modeling nonlinear Lévy processes by data analysis. Phys Rev E 64:041107
29. Kleinhans D, Friedrich R, Nawroth AP, Peinke J (2005) An iterative procedure for the estimation of drift and diffusion coefficients of Langevin processes. Phys Lett A 346:42
30. Kleinhans D, Friedrich R (2007) Note on Maximum Likeli-

hood estimation of drift and diffusion functions. Phys Lett A 368:194
31. Nawroth AP, Peinke J, Kleinhans D, Friedrich R (2007) Improved estimation of Fokker-Planck equations through optimisation. Phys Rev E 76:056102
32. Gradisek J, Grabec I, Siegert S, Friedrich R (2002). Stochastic dynamics of metal cutting: Bifurcation phenomena in turning. Mech Syst Signal Process 16(5):831
33. Gradisek J, Siegert S, Friedrich R, Grabec I (2002) Qualitative and quantitative analysis of stochastic processes based on measured data-I. Theory and applications to synthetic data. J Sound Vib 252(3):545
34. Purwins HG, Amiranashvili S (2007) Selbstorganisierte Strukturen im Strom. Phys J 6(2):21
35. Bödeker HU, Röttger M, Liehr AW, Frank TD, Friedrich R, Purwins HG (2003) Noise-covered drift bifurcation of dissipative solitons in planar gas-discharge systems. Phys Rev E 67: 056220
36. Purwins HG, Bödeker HU, Liehr AW (2005) In: Akhmediev N, Ankiewicz A (eds) Dissipative Solitons. Springer, Berlin
37. Bödeker HU, Liehr AW, Frank TD, Friedrich R, Purwins HG (2004) Measuring the interaction law of dissipative solitions. New J Phys 6:62
38. Liehr AW, Bödeker HU, Röttger M, Frank TD, Friedrich R, Purwins HG (2003) Drift bifurcation detection for dissipative solitons. New J Phys 5:89
39. Friedrich R, Siegert S, Peinke J, Lück S, Siefert M, Lindemann M, Raethjen J, Deuschl G, Pfister G (2000) Extracting model equations from experimental data. Phys Lett A 271:217
40. Siefert M Peinke J (2004) Reconstruction of the Deterministic Dynamics of Stochastic systems. Int J Bifurc Chaos 14:2005
41. Anahua E, Lange M, Böttcher F, Barth S, Peinke J (2004) Stochastic Analysis of the Power Output for a Wind Turbine. DEWEK 2004, Wilhelmshaven, 20–21 October 2004
42. Anahua E, Barth S, Peinke J (2006) Characterization of the wind turbine power performance curve by stochastic modeling. EWEC 2006, BL3.307, Athens, February 27–March 2
43. Anahua E, Barth S, Peinke J (2007) Characterisation of the power curve for wind turbines by stochastic modeling. In: Peinke J, Schaumann P, Barth S (eds) Wind Energy – Proceedings of the Euromech Colloquium. Springer, Berlin, p 173–177
44. Anahua E, Barth S, Peinke J (2008) Markovian Power Curves for Wind Turbines. Wind Energy 11:219
45. Kriso S, Friedrich R, Peinke J, Wagner P (2002) Reconstruction of dynamical equations for traffic flow. Phys Lett A 299:287
46. Kern M, Buser O, Peinke J, Siefert M, Vulliet L (2005) Stochastic analysis of single particle segregational dynamics. Phys Lett A 336:428
47. Kuusela T (2004) Stochastic heart-rate model can reveal pathologic cardiac dynamics. Phys Rev E 69:031916
48. Ghasemi F, Peinke J, Reza Rahimi Tabar M, Muhammed S (2006) Statistical properties of the interbeat interval cascade in human subjects. Int J Mod Phys C 17:571
49. Ghasemi F, Sahimi M, Peinke J, Reza Rahimi Tabar M (2006) Analysis of Non-stationary Data for Heart-rate Fluctuations in Terms of Drift and Diffusion Coefficients. J Biological Phys 32:117
50. Tabar MRR, Ghasemi F, Peinke J, Friedrich R, Kaviani K, Taghavi F, Sadghi S, Bijani G, Sahimi M (2006) New computational approaches to analysis of interbeat intervals in human subjects. Comput Sci Eng 8:54
51. Prusseit J, Lehnertz K (2007) Stochastic Qualifiers of Epileptic Brain Dynamics. Phys Rev Lett 98:138103
52. Sura P, Gille ST (2003) Interpreting wind-driven Southern Ocean variability in a stochastic framework. J Marine Res 61:313
53. Sura P (2003) Stochastic Analysis of Southern and Pacific Ocean Sea Surface Winds. J Atmospheric Sci 60:654
54. Egger J, Jonsson T (2002) Dynamic models for islandic meteorological data sets. Tellus A 51(1):1
55. Letz T, Peinke J, Kittel A (2008) How to characterize chaotic time series distorted by interacting dynamical noise. Preprint
56. Siegert S, Friedrich R, Peinke J (1998) Analysis of data sets of stochastic systems. Phys Lett A 234:275–280
57. Gradisek J, Siegert S, Friedrich R, Grabec I (2000) Analysis of time series from stochastic processes. Phys Rev E 62:3146
58. Gradisek J, Friedrich R, Govekar E, Grabec I (2002) Analysis of data from periodically forced stochastic processes. Phys Lett A 294:234
59. Frank TD, Beek PJ, Friedrich R (2004) Identifying noise sources of time-delayed feedback systems. Phys Lett A 328:219
60. Patanarapeelert K, Frank TD, Friedrich R, Beek PJ, Tang IM (2006) A data analysis method for identifying deterministic components of stable and unstable time-delayed systems with colored noise. Phys Lett A 360:190
61. Shinriki M, Yamamoto M, Mori S (1981) Multimode Oscillations in a Modified Van-der-Pol Oscillator Containing a Positive Nonlinear Conductance. Proc IEEE 69:394
62. Friedrich R, Peinke J (1997). Statistical properties of a turbulent cascade. Physica D 102:147
63. Friedrich R, Peinke J (1997) Description of a turbulent cascade by a Fokker-Planck equation. Phys Rev Lett 78:863
64. Siefert M, Peinke J (2006) Joint multi-scale statistics of longitudinal and transversal increments in small-scale wake turbulence. J Turbul 7:1
65. Friedrich R, Peinke J, Renner C (2000) How to quantify deterministic and random influences on the statistics of the foreign exchange market. Phys Rev Lett 84:5224
66. Renner C, Peinke J, Friedrich R (2001) Markov properties of high frequency exchange rate data. Physica A 298:499–520
67. Ghasemi F, Sahimi M, Peinke J, Friedrich R, Reza Jafari G, Reza Rahimi Tabar M (2007) Analysis of Nonstationary Stochastic Processes with Application to the Fluctuations in the Oil Price. Phys Rev E (Rapid Commun) 75:060102
68. Farahpour F, Eskandari Z, Bahraminasab A, Jafari GR, Ghasemi F, Reza Rahimi Tabar M, Muhammad Sahimi (2007) An Effective Langevin Equation for the Stock Market Indices in Approach of Markov Length Scale. Physica A 385:601
69. Wächter M, Riess F, Kantz H, Peinke J (2003) Stochastic analysis of raod surface roughness. Europhys Lett 64:579
70. Jafari GR, Fazeli SM, Ghasemi F, Vaez Allaei SM, Reza Rahimi Tabar M, Iraji zad A, Kavei G (2003) Stochastic Analysis and Regeneration of Rough Surfaces. Phys Rev Lett 91:226101
71. Friedrich R, Galla T, Naert A, Peinke J, Schimmel T (1998) Disordered Structures Analyzed by the Theory of Markov Processes. In: Parisi J, Müller S, Zimmermann W (eds) A Perspective Look at Nonlinear Media. Lecture Notes in Physics, vol 503. Springer, Berlin
72. Waechter M, Riess F, Schimmel T, Wendt U, Peinke J (2004) Stochastic analysis of different rough surfaces. Eur Phys J B 41:259
73. Sangpour P, Akhavan O, Moshfegh AZ, Jafari GR, Reza Rahimi

Tabar M (2005) Controlling Surface Statistical Properties Using Bias Voltage: Atomic force microscopy and stochastic analysis. Phys Rev B 71:155423
74. Jafari GR, Reza Rahimi Tabar M, Iraji zad A, Kavei G (2007) Etched Glass Surfaces, Atomic Force Microscopy and Stochastic Analysis. J Phys A 375:239
75. Ghasemi F, Bahraminasab A, Sadegh Movahed M, Rahvar S, Sreenivasan KR, Reza Rahimi Tabar M (2006) Characteristic Angular Scales of Cosmic Microwave Background Radiation. J Stat Mech P11008
76. Nawroth AP, Peinke J (2006) Multiscale reconstruction of time series. Phys Lett A 360:234
77. Ghasemi F, Peinke J, Sahimi M, Reza Rahimi Tabar M (2005) Regeneration of Stochastic Processes: An Inverse Method. Eur Phys J B 47:411
78. Kolmogorov AN (1941) Dissipation of energy in locally isotropic turbulence. Dokl Akad Nauk SSSR 32:19
79. Kolmogorov AN (1962) A refinement of previous hypotheses concerning the local structure of turbulence in a viscous incompressible fluid at high Reynolds number. J Fluid Mech 13:82
80. Frisch U (1995) Turbulence. Cambridge University Press, Cambridge
81. Sreenivasan KR, Antonia RA (1997) The phenomenology of small-scale turbulence. Annu Rev Fluid Mech 29:435–472
82. Ghashghaie S, Breymann W, Peinke J, Talkner P, Dodge Y (1996) Turbulent Cascades in Foreign Exchange Markets. Nature 381:767–770
83. Nawroth AP, Peinke J (2006) Small scale behavior of financial data. Eur Phys J B 50:147
84. Karth M, Peinke J (2002) Stochastic modelling of fat-tailed probabilities of foreign exchange rates. Complexity 8:34
85. Bouchaud JP, Potters M, Meyer M (2000) Apparent multifractality in financial time series. Eur Phys J B 13:595–599
86. Bouchaud JP (2001) Power laws in economics and finance: some ideas from physics Quant Finance 1:105–112
87. Mandelbrot BB (2001) Scaling in financial prices: I. Tails and dependence. II. Multifractals and the star equation. Quant Finance 1:113–130
88. Embrechts P, Klüppelberg C, Mikosch T (2003) Modelling extremal events. Springer, Berlin
89. Mantegna RN, Stanley HE (1995) Nature 376:46–49
90. McCauley J (2000) The Futility of Utility: how market dynamics marginalize Adam Smith. Physica A 285:506–538
91. Muzy JF, Sornette D, Delour J, Areneodo A (2001) Multifractal returns and hierarchical portfolio theory. Quant Finance 1: 131–148
92. Viscek T (1992) Fractal Growth Phenomena. World Scientific, Singapore
93. Renner C, Peinke J, Friedrich R (2000) Markov properties of high frequency exchange rate data. Int J Theor Appl Finance 3:415
94. Davoudi J, Reza Rahimi Tabar M (1999) Theoretical Model for Kramers-Moyal's description of Turbulence Cascade. Phys Rev Lett 82:1680
95. Renner C, Peinke J, Friedrich R, Chanal O, Chabaud B (2002) Universality of small scale turbulence. Phys Rev Lett 89: 124502
96. Siefert M, Peinke J (2004) Different cascade speeds for longitudinal and transverse velocity increments of small-scale turbulence. Phys Rev E 70:015302R

Fluctuation Theorems, Brownian Motors and Thermodynamics of Small Systems

Felix Ritort
Department de Fisica Fonamental, Faculty of Physics, Universitat de Barcelona, Barcelona, Spain

Article Outline

Glossary
Definition of the Subject
Introduction
Molecular Motors
Non-equilibrium Thermodynamics of Small Systems
Fluctuation Theorems
Future Directions
Bibliography

Glossary

Trajectory or path A crucial concept in the statistical description of small system is that of a trajectory or path. A path is the time sequence of configurations followed by the system as it is driven to a non-equilibrium state by the action of an external perturbation.

Control parameter External perturbations are usually described in terms of the control parameter λ. These are a set of external parameters (e. g. an electric field, magnetic field, optical force, ...) that can be experimentally controlled and do not fluctuate. Experimentally, control parameters are produced by macroscopic systems that are used to manipulate the small system under study and which are insensitive to thermal fluctuations (but that produce other sorts of uncontrolled instrumental noises and drift effects).

Single molecule experiments (SME) Recent technological developments have provided the tools to design and build scientific instruments of high enough sensitivity and precision to manipulate and visualize individual molecules and measure microscopic forces. Using SME it is possible to manipulate molecules one at a time and measure distributions describing molecular properties, characterize the kinetics of bio-molecular reactions, and detect molecular intermediates. SME provide the additional information about thermodynamics and kinetics of bio-molecular processes. This complements information obtained in traditional bulk assays. In SME it is also possible to measure small energies and detect large Brownian deviations in

bio-molecular reactions, thereby offering new methods and systems to scrutinize the basic foundations of statistical mechanics. Common single molecule experimental techniques are: atomic-force microscopy, laser optical tweezers, magnetic tweezers and single-molecule fluorescence.

Free energy The natural or spontaneous evolution of any thermodynamic process is determined by the free energy. The free energy in thermodynamics is the equivalent of the mechanical energy in classical mechanics. Spontaneous transformations take place by a decrease of the free energy in the system. In addition, mechanical work must be exerted by an external agent upon the system to increase its free energy. For reversible processes the amount of work is equal to the free energy change. However, in general, processes are irreversible and the work must be always larger than the free energy difference (a statement of the second law of thermodynamics). Free energies in small systems are typically expressed in either work (pN·nm) or energy units (kJ/mol, kcal/mol or $k_B T$ where k_B is the Boltzmann constant and T is a reference temperature – usually 298 K or 25 degrees Celsius). The conversion factors are ($T = 298\,\mathrm{K}$): $1\,k_B T = 4.11\,\mathrm{pN}\cdot\mathrm{nm} = 4.1110^{-21}\,\mathrm{J}$, $1\,k_B T = 0.6\,\mathrm{kcal/mol} = 2.4\,\mathrm{kJ/mol}$.

ATP Acronym for adenosine triphosphate, the molecule that carries the energy necessary to sustain life processes. ATP is made of one adenosine base weakly bonded to three phosphate groups. Upon conversion (by hydrolysis) to ADP (adenosine diphosphate) and inorganic phosphate or AMP (adenosine monophosphate) and pirophosphate (P-P), ATP delivers a considerable amount of free energy (in the range 8–12 kcal/mol, depending on buffer conditions). By coupling to other reactions, ATP hydrolysis supplies the energy necessary to carry out unfavorable transformations.

RNA RNA (ribonucleic acid) is a very important player in molecular biology that shows biological functions in between those attributed to DNA and proteins. For the biophysicist and the statistical physicist RNA is also a fascinating molecule. Primarily found in nature in single stranded form, RNA folds into a three dimensional structure mainly stabilized by stacking interactions and hydrogen bonds between complementary bases (A-U,G-C). Full complementarity between different RNA segments is often impossible so, at difference with DNA, RNA structure includes also mismatches between bases as well other structural defects (bulges, loops, junctions, ...). In addition to Watson–Crick base pairing, RNA forms a compact structure through specific interactions mediated by magnesium ions that bring together distal RNA segments.

Definition of the Subject

The thermodynamics of small systems describes energy exchange processes between a system and its environment in the low energy range of a few $k_B T$ where Brownian fluctuations are dominant [1]. The main goal of this discipline is to identify the building blocks of a general theory describing energy fluctuations in non-equilibrium processes occurring in systems ranging from condensed matter physics to biophysics.

Thermodynamics, a scientific discipline inherited from the 18th century, is facing new challenges in the description of non-equilibrium small (sometimes also called mesoscopic) systems. Thermodynamics is a discipline built in order to explain and interpret energetic processes occurring in macroscopic systems made out of a large number of molecules on the order of the Avogadro number. The subsequent development of statistical mechanics has provided a solid probabilistic basis to thermodynamics and increased its predictive power at the same time. The development of statistical mechanics goes together with the establishment of the molecular hypothesis. Matter is made out of interacting molecules in motion. Heat, energy and work are measurable quantities that depend on the motion of molecules. The laws of thermodynamics operate at all scales.

However, thermodynamics, a science inherited in the 18th century from the times of the industrial revolution, has been inspired by motors and steam engines that proved to be indispensable during that time. It is fair then to question the relevance and applicability of all this knowledge when scientists immerse into the realm of the very small, far from the initial context that inspired Carnot and others.

What are the novel features of thermodynamics when applied to small (also called mesoscopic) systems? Is it necessary to revisit some of the main concepts that we learned from standard thermodynamics? How are energy dissipation and efficiency related for non-equilibrium small systems where energy fluctuations are dominant? Finally, what are the implications in quantum systems already governed by quantum fluctuations? Answering such questions is one of the main goals of this new discipline.

Introduction

The *non-equilibrium thermodynamics of small systems* is becoming quite popular among statistical physicists who

recognize there new aspects of thermodynamics where large Brownian fluctuations are of pivotal importance as compared to fluctuations in macroscopic (or large) systems. In macroscopic systems, fluctuations represent just small deviations respect to the average behavior. For example, an ideal gas of N molecules in thermal contact with a bath at temperature T has an average total kinetic energy of $(3/2)Nk_BT$. However, the total energy is not a conserved quantity but fluctuates, its spectrum being a Gaussian distribution of variance $(3/2)N(k_BT)^2$ according to the law of equipartition. Therefore, relative deviations of the energy are on the order $1/\sqrt{N}$ respect to the average value. For macroscopic systems such deviations are very small: for $N = 10^{12}$ (this is the typical number of molecules in a 1ml test tube at nanomolar concentration) relative deviations are on the order of 10^{-6}, hence experimentally unobservable by calorimetry methods. For a few molecules, $N \sim \mathcal{O}(1)$, relative deviations are on the same order. Fluctuations are then measurable by direct observation of individual molecules.

Small systems share the property that energy fluctuations are much larger than $\sim \sqrt{E}$ (the prediction by the law of large numbers) where E is the average total energy. Large deviations from average values are normally observed in mesoscopic systems where non-equilibrium fluctuations are governed by a few degrees of freedom. Examples abound in physics and biology: the Brownian motion of a micron-size silica bead captured in an optical trap; the unfolding of a bio-molecule (e. g. a nucleic acid hairpin or a protein); the movement of molecular motors inside the cell; the cooperative rearrangement of a nano-sized region containing a few molecules inside an amorphous material such as a glass.

As a rule of thumb we can say that small systems are those where the typical energy content of the system is a few times k_BT, maybe from 1 to 1000 but not much more. As often happens, there is no well defined frontier separating the small-size regime from the large-size regime. The name thermodynamics of small systems was first coined by T. L. Hill [2] who showed the importance of the statistical ensemble in thermodynamic relations. A main result of statistical mechanics is the independence of the equation of state on the statistical ensemble in the thermodynamic limit. Such independence breaks down in small systems due to the contribution of fluctuations which depend on the type of statistical ensemble considered. In biology, the most important aspect of these tiny machines is that they operate far from equilibrium; its consequences and importance in their biological function are still unknown. The combination of small size and non-equilibrium behavior is the playground for the striking behavior observed in condensed matter physics and biophysics.

Prominent in the field is the study of the so called work and heat fluctuations in systems driven to a non-equilibrium state. Fluctuation theorems are mathematical relations that quantify the relative probability of trajectories that release and absorb a given amount of work/heat to and from the environment. Taken individually, the work and heat along these trajectories can violate some of the inequalities of thermodynamics, leading to what is commonly referred as *transient violations of the second law*. This name has raised strong objections among some groups of physicists. Of course the second law remains inviolate. The name just stresses the fact that Brownian fluctuations are big enough for such deviations from the average value to be observed. For macroscopic systems these trajectories are known to be irrelevant and unobservable, however at the level of small system sizes, when the energies involved are of order of several times k_BT, these trajectories become important. Although thermodynamic inequalities are known to describe the behavior of average values, it is important to explore the implications and relevance of these deviations in our understanding of energy transformation processes at the molecular level.

The quantitative experimental observation and measurement of large energy fluctuations has become possible only recently with the development of new micro-manipulation tools. Particularly important are the application of single-molecule techniques to explore physical theories in systems out of equilibrium. The use of new micro-manipulation tools in the exploration of the behavior of tiny objects (such as bio-molecules and motors) embedded in a thermal environment opens the possibility to investigate how these systems exchange energy with their environment. This question is of great interest both at a fundamental and practical level. From a fundamental point of view, the comprehension of how bio-molecules operating very far from equilibrium are so efficient raises the question whether such tiny systems exploit rare and large deviations from their average behavior by rectifying thermal fluctuations from the bath. From a practical point of view, this might help in the design of efficient nanomotors in the future.

In the next sections we overview a selection of topics in this fascinating area of research. We will start by discussing molecular motors in Sect. "Molecular Motors". In Sect. "Non-equilibrium Thermodynamics of Small Systems" we explain what are the main types of non-equilibrium states and introduce the microscopic definitions of heat and work. In Subsect. "The Jarzynski Equality" we

derive the Jarzynski equality by using the master equation approach. This part is a bit technical and the reader who is not interested in the details can jump to Eq. (17) and continue reading from there. Finally, in Sect. "Fluctuation Theorems" we discuss fluctuation theorems and review some of their applications to condensed matter physics and biophysics. The article finishes with a discussion of future directions in research.

Molecular Motors

Molecular motors are proteins that use the energy extracted from the hydrolysis of ATP to exert mechanical work (Fig. 1) [3]. The mechanism by which motors utilize the energy stored in the high energy bonds of the ATP molecules to perform mechanical work is based on two hypothesized mechanisms: 1) Power stroke generation or 2) Brownian ratchet mechanism. In the first mechanism the release of the pyrophosphate during the ATP hydrolysis cycle is tightly coupled to the generation of force which drives the motor. In the second mechanism, the motor diffuses reversibly along the substrate. Movement is then produced by the hydrolysis of ATP that induces a conformational change in the protein. This change is then rectified by thermal fluctuations that induce the release of ADP. By steady repetition of this mechanochemical cycle (one ATP molecule is hydrolyzed per cycle) the motor carries out important cellular functions. Motors are characterized by the so called processivity or number of turnover cycles the motor does until detaching from the substrate. Processivities of molecular motors can vary by several orders of magnitude depending on the type of motor and the presence of other regulating factors. For example, the muscle myosin II motors work in large assemblies, each myosin having a processivity around 1, meaning that each myosin performs one mechanochemical cycle on average before detaching from the substrate. In the other extreme of the scale there are DNA polymerases in eukaryotes which show processivities that range from 1 (adding approximately one nucleotide before detaching) up to sev-

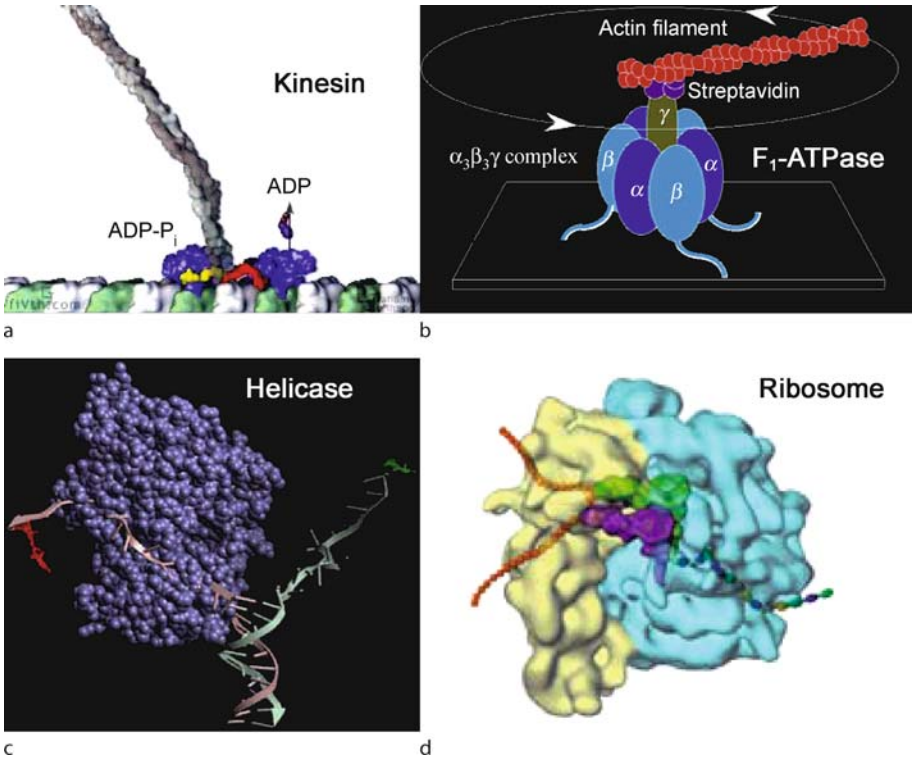

Fluctuation Theorems, Brownian Motors and Thermodynamics of Small Systems, Figure 1
Examples of molecular machines: a Kinesin walking a long micro-tubule and transporting a cargo. b F_1-ATP synthase is the proton pump responsible of producing ATP in the mitochondria of eukaryotic cells. c Helicases are forerunners of the DNA polymerase that unwind DNA by transforming dsDNA into two strands of ssDNA. d The ribosome is one among the largest molecular machines inside the cytoplasm of the cell in charge of manufacturing proteins

eral thousands or even millions. However, in the presence of sliding clamps (proteins with the shape of a doughnut that encircle the DNA and tightly bind DNA polymerases) processivities go up to 10^9. Molecular motors are magnificent objects from the point of view of their efficiency. If we define the efficiency rate as the ratio between the useful work performed by the motor and the energy released in the hydrolysis of one ATP molecule in one mechanochemical cycle, then typical values for the efficiencies are around several tens per cent, reaching the value of 90% in some cases (like in the rotary motor F_1-ATPase). For example, out of the $20\,k_B T$ obtained from the hydrolysis of one molecule of ATP, kinesin exerts a mechanical work of $12\,k_B T$ at every step, having an efficiency of around 60%. Such large efficiencies are rarely found in macroscopic systems (motors of cars have efficiencies below 20%) meaning that molecular motors have been designed by evolution to efficiently operate in a highly noisy environment. Molecular motors are expected to be essential constituents of future nanodevices.

What is the relation between molecular motors and the non-equilibrium thermodynamics of small systems? It is a well established fact that the typical amounts of energy obtained from chemical sources (e. g. ATP or GTP hydrolysis) used by most molecular machines are a few kcal per mol (at $T \sim 300$ K this corresponds to a few units of $k_B T$, $1\,k_B T \simeq 0.6$ kcal/mol). Let us consider the example of RNA transcription. The process by which RNA nucleotides (A,U,G,C) are added to the newly synthesized RNA strand during the transcription process involves the hydrolysis of the different nucleoside-phosphate complexes as they are added to the 3' end of the growing chain. The overall process by which one base is added to the newly synthesized strand is a highly favorable reaction (mainly driven by the hydrolysis of the pyrophosphate) with a free energy release, ΔG, in the range between 7 and 12 kcal/mol mainly depending on the magnesium concentration in the environment. Effectively this is an irreversible process that generates an amount of available energy between 10 and $20\,k_B T$ at room temperature (~ 300 K, $1 k_B T \simeq 0.6$ kcal/mol) per base pair added. This energy would be lost to the environment in the form of heat were it not for the fact that a big part of the energy is used by the RNA polymerase to locally unwind the double DNA helix and pull apart the two DNA strands to produce a bubble a few bases long of denatured DNA. This bubble is then used by the DNA/RNA/polymerase ternary complex as a substrate to polymerize the RNA. As transcription proceeds the bubble moves downstream together with the RNA polymerase and the RNA transcript is synthesized.

For this process to occur, the RNA polymerase must move against the Stokes friction produced by water as well as other roadblocks that hamper its motion. In particular, the RNA polymerase must exert force and torque on the DNA. Typical forces to unzip DNA are on the order of 15 pN meaning that the minimum mechanical work necessary to unzip one base pair is around 15 pN times 12 Angstroms (the typical extension gained after pulling apart two bases at the fork of a DNA hairpin), which is equal to $18\,\mathrm{pN}\cdot\mathrm{nm}$ or equivalently $4.4\,k_B T$ ($1\,k_B T \simeq 4.11\mathrm{pN}\cdot\mathrm{nm}$ at room temperature). We can define the efficiency of the RNA motor as the ratio between the mechanical work needed to unzip one base pair and the amount of energy obtained from hydrolysis upon the addition of a nucleotide (however this is not the only way to define mechanical efficiencies, e. g. see [4,5]). The efficiency of the transcription process is then about 40 per cent, a quite remarkable feat if we compare this number with the ones obtained in man made machines (cars have efficiencies below 20 per cent). The motion of a single RNA polymerase has been studied in several prokaryotic systems using optical and magnetic tweezers [6,7]. In these experiments a DNA/polymerase complex is tethered between a trapped bead and an streptavidin coated immobilized bead or surface. To initiate transcription nucleotides are allowed to flow inside the chamber and the elongation of the transcript can be followed in real time while force is applied on the tether. The extension of the RNA transcript as a function of time reveals a complex intermittent motion of the polymerase with pauses (temporary stops), arrests (permanent stops) and even backtracking events [8,9].

Non-equilibrium Thermodynamics of Small Systems

Non-equilibrium States

An important concept in thermodynamics is the state variable. State variables are those that, once determined, uniquely specify the thermodynamic state of the system. Examples are the temperature, the pressure, the volume and the mass of the different components in a given system. To specify the state variables of a system it is common to put the system in contact with a bath. The bath is any set of sources (of energy, volume, mass, etc.) large enough to remain unaffected by the interaction with the system under study. The bath ensures that a system can reach a given temperature, pressure, volume and mass concentrations of the different components when put in thermal contact with the bath (i. e. with all the relevant sources). Equilibrium states are then generated by putting the system in contact with a bath and waiting until the system proper-

ties relax to the equilibrium values. Under such conditions the system properties do not change with time and the average heat/work/mass exchanged between the system and the bath is zero.

Non-equilibrium states can be produced in many ways, either by continuously changing the parameters of the bath or by preparing the system in an initial non-equilibrium state that slowly relaxes toward equilibrium. In general a non-equilibrium state is produced whenever the system properties change with time and/or the net heat/work/mass exchanged by the system and the bath is non zero. We can distinguish at least three different classes of non-equilibrium states:

- **Non-equilibrium transient state (NETS)** The system is initially prepared in an equilibrium state and later driven out of equilibrium by switching on an external perturbation. The system quickly returns to a new equilibrium state once the external perturbation stops changing.
  A classic example of NETS is the case of a protein in its initial native state that is mechanically pulled (e. g. using AFM) by exerting force on the ends of the molecule. The protein is initially folded and in thermal equilibrium with the surrounding aqueous solvent. Upon pulling the protein is driven away from equilibrium into a transient state until it finally settles into the unfolded and extended equilibrium state. Another example of a NETS is a bead immersed in water and trapped in an optical well generated by a focused laser beam. When the trap is moved to a nearby new position (e. g. by moving the laser beams) the bead is driven to a NETS. After some time the bead reaches equilibrium again at the new position of the trap. In another experiment the trap is suddenly put in motion at a speed v so the bead is transiently driven away from its equilibrium average position until it settles into a non-equilibrium steady-state (NESS, see below) characterized by the speed of the trap. The average position of the bead lags behind the position of the center of the trap.
- **Non-equilibrium steady-state (NESS)** The system is driven by external forces (either time dependent or non-conservative) in a stationary non-equilibrium state where its properties do not change with time. The steady state cannot be described by the Boltzmann–Gibbs distribution and the average net heat that is dissipated by the system (equal to the entropy production of the bath) is positive.
  A classic example of a NESS is an electrical circuit made out of a battery and a resistance. The current flows through the resistance and the chemical energy stored in the battery is dissipated to the environment in the form of heat; the average dissipated power, $\mathcal{P}_{\rm dis} = VI$, is equal to the power supplied by the battery. Another example is a sheared fluid between two plates or coverslips and one of them is moved relative to the other at a constant velocity v. To sustain such state a mechanical power equal to $\mathcal{P} \propto \eta v^2$ has to be exerted upon the moving plate where η is the viscosity of water. The mechanical work produced is then dissipated in the form of heat through the viscous friction between contiguous fluid layers. Further examples of NESS are chemical reactions in metabolic pathways that are sustained by activated carrier molecules such as ATP. In such cases, hydrolysis of ATP is strongly coupled to specific oxidative reactions. For example, ionic channels use ATP hydrolysis to transport protons against the electromotive force.
- **Non-equilibrium aging state (NEAS)** The system is initially prepared in a non-equilibrium state and put in contact with the sources. However, at difference with NETS, the system fails to reach thermal equilibrium in observable or laboratory time scales. In this case the system is in a non-stationary slowly relaxing non-equilibrium state called *aging state* and characterized by a very small entropy production of the sources. In the aging state two-time correlations decay slower as the system becomes older. Two-time correlation functions depend on both times and not just on their difference. The classic example of a NEAS is a supercooled liquid cooled below its glass transition temperature [10]. The liquid solidifies into an amorphous slowly relaxing state characterized by huge relaxational times and anomalous low frequency response. Other systems are colloids that can be prepared in a NEAS by the sudden reduction/increase of the volume fraction of the colloidal particles or by putting the system under a strain/stress [11].

The classes of non-equilibrium states previously described do not make distinctions whether the system is macroscopic or small. In small systems, however, it is common to speak about the control parameter to emphasize the importance of the constraints imposed by the bath that are externally controlled and do not fluctuate. The control parameter (λ) represents a value (in general, a set of values) that defines the state of the bath. Its value determines the equilibrium properties of the system, e. g. the equation of state. In macroscopic systems it is unnecessary to discern which value is externally controlled because fluctuations are small and all equilibrium ensembles give the same equivalent thermodynamic description, i. e. the same

equation of state. Differences arise when taking into account fluctuations. The non-equilibrium behavior of small systems is then strongly dependent on the specific non-equilibrium protocol. Figure 2 shows a representation of a few examples of NESS and control parameters.

Microscopic Definitions of Work and Heat

Microscopic definitions of work and heat can be given using Markov processes. Let us consider a general system described by an energy function $E(C)$ where C is a generic configuration in contact with a bath at temperature T. For instance, in a gas of N molecules, C would stand for their positions and momenta. The dynamics are assumed to be discrete in time with elementary time-step Δt. A trajectory or path of the system is characterized by the sequence of configurations

$$\Gamma \equiv \{C_k; 0 \leq k \leq M\} \equiv \{C_0, C_1, C_2, \ldots, C_M\} \qquad (1)$$

where k is the index for the discrete time step and M is the total number of time steps. The time corresponding to step k is then given by $t = k\Delta t$ with $t = 0\,(k = 0)$ and $t_f(k = M/\Delta t)$ denoting the initial and final times respectively. The continuous-time limit is recovered by taking $\Delta t \to 0$, $M \to \infty$ with t_f finite.

Now we will treat the case where the system is perturbed in a prescribed way. Because the dynamics is stochastic, it will generate an ensemble of trajectories when the same experiment is repeated many times. In addition to the configuration C, and in order to characterize the perturbation protocol, we need to specify the temporal sequence of values $\{\lambda_k; 0 \leq k \leq M\}$. The control parameter λ shifts the energy levels of the system according to the relation,

$$E_\lambda(C) = E(C) - \lambda A(C) \qquad (2)$$

where $A(C)$ is the observable coupled to the external perturbation λ (e. g., if λ is a magnetic or gravitational field then A stands for the magnetization or the height of the center of mass respectively).

We now consider the variation of energy along a given path $\Delta E(\Gamma) = E_{\lambda_f}(C_f) - E_{\lambda_0}(C_0)$ where C_0, C_f are the initial and final configurations for that path and λ_0, λ_f are the initial and final values of the control parameter as defined by the protocol (path independent). From Eq. (2) the

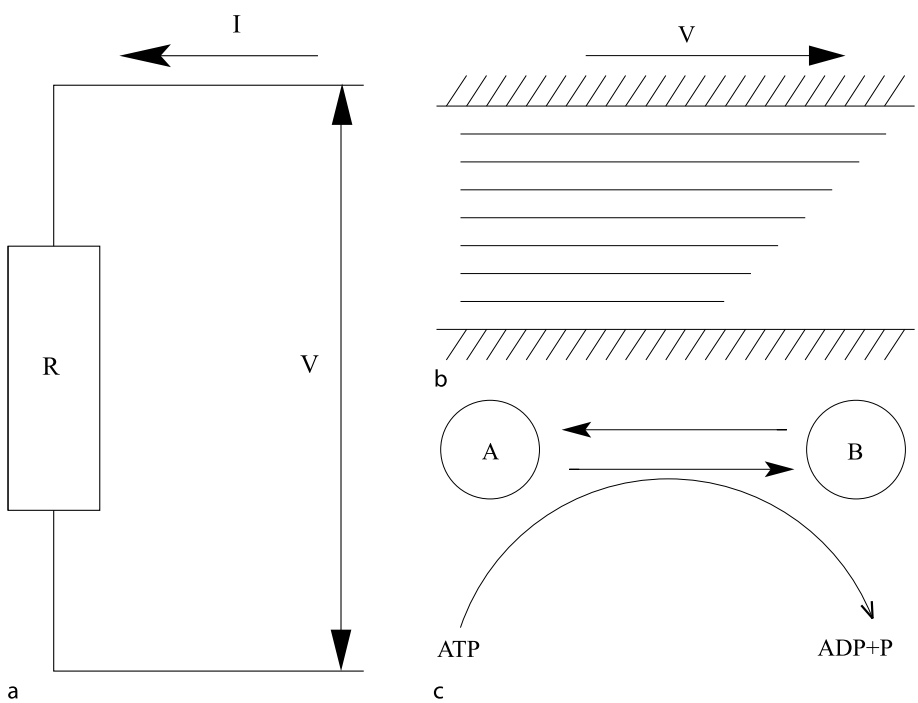

Fluctuation Theorems, Brownian Motors and Thermodynamics of Small Systems, Figure 2
Examples of NESS: a An electric current *I* flowing through a resistance *R* and maintained by a voltage source or control parameter *V*. b A fluid sheared between two plates that move at speed *v* (the control parameter) relative to each other. c A chemical reaction *A* → *B* coupled to ATP hydrolysis. The control parameters are the concentrations of ATP and ADP

energy variation is given by,

$$\begin{aligned}\Delta E(\Gamma) &= E_{\lambda_M}(C_M) - E_{\lambda_0}(C_0)\\ &= \sum_{k=0}^{M-1}\left(E_{\lambda_{k+1}}(C_{k+1}) - E_{\lambda_k}(C_{k+1})\right)\\ &\quad - \sum_{k=0}^{M-1}\left(E_{\lambda_k}(C_k) - E_{\lambda_k}(C_{k+1})\right)\end{aligned} \tag{3}$$

with $\Delta\lambda_k = \lambda_{k+1} - \lambda_k$. This decomposition identifies work and heat by using the first law of thermodynamics, $\Delta E = W - Q$. The first term in Eq. (3) is identified as the work W exerted upon the system whereas the second corresponds to the heat Q transferred from the system to the bath,

$$W(\Gamma) = \sum_{k=0}^{M-1}\left(E_{\lambda_{k+1}}(C_{k+1}) - E_{\lambda_k}(C_{k+1})\right) \tag{4}$$

$$Q(\Gamma) = \sum_{k=0}^{M-1}\left(E_{\lambda_k}(C_k) - E_{\lambda_k}(C_{k+1})\right) . \tag{5}$$

We concentrate our attention on the work exerted upon the system along a given path Γ. Inserting (2) in (4) we get

$$\begin{aligned}W(\Gamma) &= \sum_{k=0}^{M-1}\left(\frac{\partial E_\lambda(C_k)}{\partial\lambda}\right)_{\lambda=\lambda_k}\Delta\lambda_k\\ &= -\sum_{k=0}^{M-1}A(C_k)\Delta\lambda_k \equiv -\int_0^t \mathrm{d}s\dot{\lambda}(s)A(C(s))\mathrm{d}s\end{aligned} \tag{6}$$

where we have applied the continuous-time limit in the last term in the r.h.s. of (6).

As the path is stochastic the work is a fluctuating quantity that can be characterized by its probability distribution $\mathcal{P}(W)$ defined as,

$$\mathcal{P}(W) = \sum_\Gamma P(\Gamma)\delta(W - W(\Gamma)) \tag{7}$$

where Γ stands for the path and $P(\Gamma)$ indicates the probability of that path. The importance of $\mathcal{P}(W)$ relies upon the fact that it is a quantity that is experimentally measurable and therefore is suitable to quantitatively characterize work fluctuations with the aid of recently developed micro-manipulation tools.

The Jarzynski Equality

In 1997 Chris Jarzynski derived a remarkable equality describing work fluctuations in non-equilibrium isothermal systems [12]. This relation was somewhat unexpected because it related the free energy change in a reversible process with exponential averages of the work measured in irreversible processes. The equality applies to all non-equilibrium systems under general assumptions of local detailed balance and ergodicity. In what follows we show a derivation of the equality based on a master equation approach. The following calculation intends to show the simplicity of the algebraic math used in the derivation of this general result. The reader not interested in math details can jump directly to Eq. (17) and continue reading from there.

Let us consider a system in contact with a thermal bath at temperature T that is initially in thermal equilibrium. The system is then driven to a NETS under the action of an external perturbation described by the temporal sequence of values $\{\lambda_k; 0 \le k \le M\}$.

We consider the ensemble of all possible trajectories that start from an initial state characterized by the distribution $P_{\lambda_0}(C)$. Dynamics of the system are then given by the set of probabilities $P_{\lambda_k}(C)$ for the system to be found at configuration C at time-step k. These probabilities satisfy a master equation. For a Markov process the time evolution of the $P_\lambda(C)$ depends on the quantities, $\mathcal{W}_{\lambda_k}(C \to C')$, defined as the transition probability per unit time to go from configuration C to C' at time-step k. The $\mathcal{W}$'s are assumed to lead to an ergodic dynamics (where any pair of configurations are always connected by at least one trajectory, i. e. dynamics is *irreducible*) and satisfy the detailed balance condition,

$$\begin{aligned}\frac{\mathcal{W}_{\lambda_k}(C \to C')}{\mathcal{W}_{\lambda_k}(C' \to C)} &= \frac{P^{\mathrm{eq}}_{\lambda_k}(C')}{P^{\mathrm{eq}}_{\lambda_k}(C)}\\ &= \exp\left(-\beta(E_{\lambda_k}(C') - E_{\lambda_k}(C))\right)\end{aligned} \tag{8}$$

where $\beta = 1/k_{\mathrm{B}}T$, k_{B} is the Boltzmann constant and T is the temperature. The $P^{\mathrm{eq}}_{\lambda_k}(C)$ is the Boltzmann–Gibbs distribution,

$$P^{\mathrm{eq}}_\lambda(C) = \exp(-\beta E_\lambda(C_0))/Z_\lambda\,; \qquad Z_\lambda = \sum_C \exp(\beta E_\lambda(C) \tag{9}$$

where $Z_\lambda = \exp(-\beta F_\lambda)$ is the partition function and F_λ is the free energy.

The energy function $E_{\lambda_k}(C)$ is given in (2). Under very general conditions these dynamics guarantee that the system reaches a stationary state where configurations are populated according to the Boltzmann–Gibbs weight. The solution to the master equation gives the time evolution for the system.

For a generic path-dependent observable $\mathcal{A}(\Gamma)$, the ensemble average value is given by,

$$\langle \mathcal{A} \rangle = \sum_{\Gamma} P(\Gamma)\mathcal{A}(\Gamma) \,. \tag{10}$$

Using the fact that the dynamics are Markovian together with the definition (1) we can write,

$$P(\Gamma) = P^{\mathrm{eq}}_{\lambda_0}(C_0) \prod_{k=0}^{M-1} \mathcal{W}_{\lambda_k}(C_k \to C_{k+1}) \tag{11}$$

where the system initially starts in equilibrium at λ_0. By inserting (11) into (10) we obtain,

$$\langle \mathcal{A} \rangle = \sum_{\Gamma} \mathcal{A}(\Gamma) P^{\mathrm{eq}}_{\lambda_0}(C_0) \prod_{k=0}^{M-1} \mathcal{W}_{\lambda_k}(C_k \to C_{k+1}) \,. \tag{12}$$

Using the detailed balance condition (8) this expression reduces to,

$$\langle \mathcal{A} \rangle = \sum_{\Gamma} P^{\mathrm{eq}}_{\lambda_0}(C_0)\mathcal{A}(\Gamma) \prod_{k=0}^{M-1} \Big[\mathcal{W}_{\lambda_k}(C_{k+1} \to C_k) \exp\left[-\beta(E_{\lambda_k}(C_{k+1}) - E_{\lambda_k}(C_k))\right] \Big] \tag{13}$$

$$= \sum_{\Gamma} \mathcal{A}(\Gamma) P^{\mathrm{eq}}_{\lambda_0}(C_0) \exp\left[-\beta \sum_{k=0}^{M-1} \left(E_{\lambda_k}(C_{k+1}) - E_{\lambda_k}(C_k)\right) \right] \prod_{k=0}^{M-1} \mathcal{W}_{\lambda_k}(C_{k+1} \to C_k) \,. \tag{14}$$

This equation can not be worked out further for a general observable $\mathcal{A}$. However, let us consider the observable, $\mathcal{A}(\Gamma) = \exp(-W(\Gamma))$, where $W(\Gamma)$ stands for the work defined in (4). By inserting this expression in (14) we obtain the Jarzynski equaltity:

$$\begin{aligned} &\langle \exp(-W) \rangle \\ &= \frac{1}{Z_{\lambda_0}} \sum_{\Gamma} \prod_{k=0}^{M-1} \mathcal{W}_{\lambda_k}(C_{k+1} \to C_k) \exp(-\beta E_{\lambda_M}(C_0)) \\ &= \frac{Z_{\lambda_M}}{Z_{\lambda_0}} = \exp\left(-\beta F_{\lambda_M} - F_{\lambda_0}\right) = \exp(-\beta \Delta F) \end{aligned} \tag{15}$$

where we have applied a telescopic sum and used (9). To carry out the telescopic sums we first summed over $C_0, C_1 \ldots$ by applying the normalization condition on the transition probabilities,

$$\sum_{C'} \mathcal{W}_k(C \to C') = 1 \,. \tag{16}$$

The second law of thermodynamics, $\langle W \rangle \geq \Delta F$, also follows naturally as a particular case of (15) by using the convexity inequality, $\langle \exp(x) \rangle \geq \exp\langle x \rangle$. The Jarzynski equality is often written in the form,

$$\langle \exp(-W_{\mathrm{diss}}) \rangle = 1 \tag{17}$$

where $W_{\mathrm{diss}} = W - \Delta F$ is called the dissipated work. The second law of thermodynamics puts bounds on the minimum amount of average work performed on the system: although W may strongly fluctuate from path to path its mean value (averaged over an infinite number of repeated experiments, i. e. the first moment of $P(W)$) is always greater than the reversible or quasi-static work, W_{rev}, which is also equal to the free energy difference ΔF between the initial and final equilibrium states. The reversible work is the value of the work that is obtained for protocols that are adiabatic or quasi-static, i. e. the control parameters are changed infinitely slowly. The difference between the actual work and the reversible work then corresponds to the dissipated work, $W_{\mathrm{diss}} = W - \Delta F$. The second law establishes that in average a positive amount of heat is irreversibly lost to the environment, $\langle W_{\mathrm{diss}} \rangle \geq 0$ (Fig. 3). The amount of dissipation in irreversible processes is then related to the asymmetry between the phase space densities obtained when the process is run forward and backward in time [13].

Gaussian work distributions exactly satisfy (17) provided the average dissipated work $\langle W_{\mathrm{diss}} \rangle$ and the variance of the work, σ_W^2, are related by $\sigma_W^2 = 2k_{\mathrm{B}}T\langle W_{\mathrm{diss}} \rangle$. A fluctuation-dissipation parameter R can be introduced to quantify deviations from the Gaussian behavior [14],

$$R = \frac{\sigma_W^2}{2k_{\mathrm{B}}T\langle W_{\mathrm{diss}} \rangle} \,. \tag{18}$$

For Gaussian work distributions $R = 1$, corresponding to non-equilibrium processes in the linear regime where the fluctuation-dissipation theorem holds.

The validity of the Jarzynski equality extends to deterministic dynamics (e. g. Hamiltonian or thermostated). In Hamiltonian dynamics the set of phase space points then behaves as an incompressible fluid, a consequence of the Liouville theorem. The case of Hamiltonian dynamics was originally addressed by Jarzynski in his original derivation of the non-equilibrium work relation [12]. The

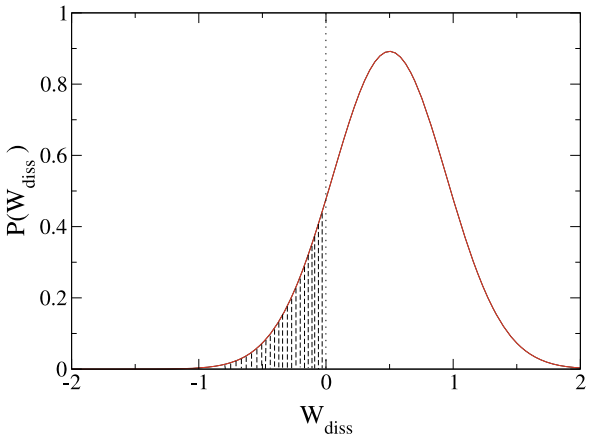

Fluctuation Theorems, Brownian Motors and Thermodynamics of Small Systems, Figure 3
Probability distribution of the dissipated work: According to the second law of thermodynamics the average dissipated work is always positive. However, because of fluctuations, the dissipated work of some paths can be negative (*shaded area*). These paths are sometimes referred to as "transient violations of the second law"

stochastic case has been analyzed also for general Markov processes by Crooks [15,16] and for Langevin dynamics by Kurchan [17] and Seifert [18]. Equation (15) has appeared in the past in the literature in the form of a generalized fluctuation-dissipation relation proposed by Bochkov and Kuzovlev [19] which is mathematically identical to the Jarzynski equality [12]. Related results to the Jarzynski's equality can be also traced back also in the free-energy perturbation identity derived by Zwanzig [20] and the Kirkwood formula [21].

Fluctuation Theorems

Since the beginning of the 90's some exact results in statistical mechanics have provided a mathematical description of energy fluctuations (in the form of heat and work) for non-equilibrium systems. This new class of results go under the name of fluctuation theorems (FTs) and provide a solid theoretical basis to quantify energy fluctuations in non-equilibrium systems. FTs describe energy fluctuations in systems while they execute transitions between different types of states. For these fluctuations to be observable the system has to be small enough and/or operate over short periods of time, otherwise the measured properties approach the macroscopic limit where fluctuations get masked by the dominant average behavior. Most fluctuation theorems are of the form,

$$\frac{P(+S)}{P(-S)} = \exp\left(\frac{S}{k_B}\right) , \tag{19}$$

where S has the dimensions of an entropy that may represent heat and/or work produced during a given time interval. The precise mathematical form of relations such as (19) (for instance, the precise definition of S or whether they are valid at finite time intervals or just in the limit where the time interval goes to infinity) depends on the particular non-equilibrium conditions (e. g. whether the systems starts in an equilibrium Gibbs state, or whether the system is in a non-equilibrium steady state, or whether the system executes transitions between steady states, etc.).

Generally speaking, FTs relate the amounts of work or heat exchanged between the system and its environment for a given non-equilibrium process and its corresponding time-reversed process. The time-reversed process is defined as follows. Let us consider a given non-equilibrium process (we call it forward, denoted by F) characterized by the protocol $\lambda_F(t)$ of duration t_f. In the reverse process (denoted by R) the system starts at $t = 0$ in a stationary state at the value $\lambda_F(t_f)$ and the control parameter is varied for the same duration t_f as in the forward process according to the protocol $\lambda_R(t) = \lambda_F(t_f - t)$. FTs depend on the type of initial state and the particular type of dynamics (deterministic versus stochastic) or thermostated conditions.

Despite of the fact that most of these theorems are treated as distinct they are in fact closely related. The main hypothesis for all theorems is the validity of some form of microscopic reversibility or local detailed balance (see however [22,23,24] for some controversy). Major classes of FTs include the transient FT (TFT) and the steady state FT (SSFT):

- The transient FT (TFT) In the TFT the system initially starts in an equilibrium (Boltzmann–Gibbs) state and is driven away from equilibrium by the action of time-dependent forces that derive from a time-dependent potential $V_{\lambda(t)}$. The potential depends on time through the value of the control parameter $\lambda(t)$. At any time during the process the system in an unknown transient non-equilibrium state. If the value of λ is kept fixed then the system relaxes into a new equilibrium state. The TFT was introduced by Evans and Searles [25] in thermos-tatted systems and later extended by Crooks to Markov processes [15].
- The steady state FT (SSFT) In the SSFT the system is in a non-equilibrium steady state where it exchanges net heat and work with the environment. The existence of the SSFT was numerically anticipated by Evans and collaborators for thermostated systems [26] and demonstrated for deterministic Anosov systems by Gallavotti and Cohen [27]. The entropy production S by the system (equal to the heat exchanged by the system divided

by the temperature of the environment) satisfies the relation (19) in the asymptotic limit of large times $t \to \infty$ and for bounded energy fluctuations, $\sigma = \frac{|S|}{t} < \sigma^*$ where σ^* is a model dependent quantity. Other classes of SSFTs include stochastic Langevin dynamics [17], Markov chains [28,29] or the case where the system initially starts in a steady state [30] and executes transitions between different steady states [31,32].

Particularly relevant to the single molecule context is the FT by Crooks [15,16] which relates the work distributions measured along the forward (F) and reverse (R) paths,

$$\frac{P_{\mathrm{F}}(W)}{P_{\mathrm{R}}(-W)} = \exp\left(\frac{W - \Delta F}{k_{\mathrm{B}} T}\right), \tag{20}$$

where $P_{\mathrm{F}}(W), P_{\mathrm{R}}(-W)$ are the work distributions along the F and R processes respectively, and ΔF is the free energy difference between the equilibrium states corresponding to the final value of the control parameter $\lambda_f = \lambda(t_f)$ and the initial one $\lambda_i = \lambda(0)$. A particular result of (20) is the Jarzynski equality [12] described in Subsect. "The Jarzynski Equality" that is obtained from (20) by rewriting it as $P_{\mathrm{R}}(-W) = \exp\left(\frac{-W+\Delta G}{k_{\mathrm{B}} T}\right) P_{\mathrm{F}}(W)$ and integrating both sides of the equation between $W = -\infty$ and $W = \infty$. Because of the normalization property of probability distributions, the left hand side is equal to 1 and the Jarzynski equality reads,

$$\begin{aligned} &\left\langle \exp\left(-\frac{W}{k_{\mathrm{B}} T}\right)\right\rangle_{\mathrm{F}} = \exp\left(-\frac{\Delta F}{k_{\mathrm{B}} T}\right) \text{ or} \\ &\Delta F = -k_{\mathrm{B}} T \log\left(\left\langle \exp\left(-\frac{W}{k_{\mathrm{B}} T}\right)\right\rangle_{\mathrm{F}}\right), \end{aligned} \tag{21}$$

where $< \ldots >_{\mathrm{F}}$ denotes an average over an infinite number of paths, all generated by a given forward protocol $\lambda_{\mathrm{F}}(t)$.

Experimental Tests and Free Energy Recovery

Various categories of FTs have been introduced and experimentally validated. The first experimental tests of FTs were carried out by Ciliberto and coworkers for the Gallavoti–Cohen FT in Rayleigh–Bernard convection [33] and turbulent flows [34]. Later on FTs were tested for beads trapped in an optical potential and moved through water at low Reynolds numbers. The motion of the bead is then well described by a Langevin equation that includes a friction (non-conservative) force, a confining (conservative) potential and a source of stochastic noise. Experiments have been carried out by Evans and collaborators who have tested the validity of the TFT [35,36], and by Liphardt and collaborators for a bead executing transitions between different steady states [37]. The validity of the TFT has been also recently tested for non-Gaussian optical trap potentials [38].

The Jarzynski equality and the FT by Crooks can be used to recover equilibrium free-energy differences between different molecular states by using non-equilibrium measurements in single molecule experiments [39,40,41]. In particular, by using the Jarzynski equality (21) it is possible to extract the value of ΔF from repeated measurements of the mechanical work along many trajectories. The idea is to repeat non-equilibrium experiments many times and evaluate the exponential average in the r.h.s of (21) to extract the work corresponding to the reversible process. Would it then not be easier to directly measure the work for a reversible process? Unfortunately many processes cannot be carried in quasistatic conditions (either simulations or experiments) and therefore, alternative methods are required to determine free-energy differences. There are practical difficulties in the applicability of (21) as the number of trajectories included in the exponential average must be actually infinite. This is unrealizable in practice as non-equilibrium experiments can be performed only a finite number of times and the finiteness of the number of trajectories introduces a bias. It is known that the number of trajectories required to evaluate the Jarzynski equality grows exponentially with the average value of the dissipated work. The dependence of the bias and error with the number of pulls has been estimated in some cases [42,43]. In general this dependence can be quite complicated as it depends on the behavior of the low-work tails of the distribution $P(W)$ which are difficult to analyze in general.

In 2002, the Jarzynski equality was experimentally tested by pulling the P5ab RNA hairpin, a derivative of the *Tetrahymena Termophila* L21 ribozyme [44] using optical tweezers. However, in that case the molecule was pulled not too far from equilibrium. The Jarzynski equality and related identities for athermal systems have been recently put under scrutiny in other systems [45,46,47]. The Jarzynski equality and the FT by Crooks have inspired several theoretical papers discussing other related exact results [48,49,50,51,52,53], free-energy recovery from numerical simulations [54,55,56,57,58,59], bias and error estimates for free-energy differences [42,43,60,61,62,63], enzyme kinetics [64,65] or solvable models [66,67,68,69,70]. In addition, analytical studies on small systems thermodynamics show that work/heat distributions display non-Gaussian tails describing large and rare deviations from the average and/or most probable behavior [71,72,73,74]. These theoretical studies open the way to investigate the

possible relevance of these large deviations in other non-equilibrium systems in condensed matter physics.

The FT by Crooks can be applied and tested by measuring the unfolding and refolding work distributions in single molecule pulling experiments (Fig. 4a). For example, let us consider the case of a molecule (e. g. a DNA or RNA hairpin or a protein) initially in thermal equilibrium in the folded (F) or native state. By applying mechanical force (e. g. using atomic force microscopy or optical tweezers) the molecule can be mechanically unfolded and the conformation of the molecule changed from the native to the unfolded (U) state. The unfolding event is observed by the presence of a rip in the force-extension curve of the molecule (Fig. 4b). During the unfolding process the tip of the cantilever or the bead in the trap exerts a mechanical work on the molecule that is given by,

$$W = \int_{x_0}^{x_f} F dx \tag{22}$$

where x_0, x_f are the initial and final extension of the molecule. In (22) we are assuming that the molecular extension x is the externally controlled parameter (i. e. $\lambda \equiv x$) which is not necessarily the case. However the corrections introduced by using (22) are shown to be often small. The work (22) done upon the molecule along a given path corresponds to the area below the force-extension curve that is limited by the initial and final extensions, x_0 and x_f (Fig. 4b). Because the unfolding of the

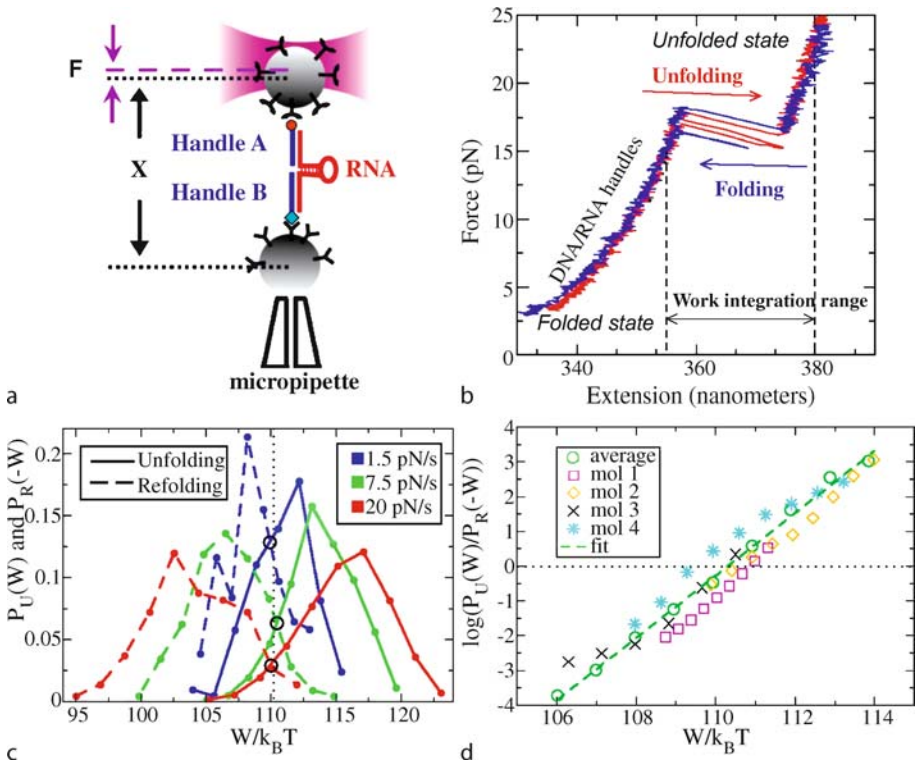

Fluctuation Theorems, Brownian Motors and Thermodynamics of Small Systems, Figure 4

Experimental measurement of work fluctuations in small systems using single molecules. a Experimental setup in RNA force pulling experiments using optical tweezers. An RNA hairpin of a few tens of base pairs is inserted between two molecular handles (A and B) that are attached to two beads. One bead is trapped in the optical well, the other bead is immobilized on the tip of a micropipette. As the optical trap is moved relative to the micropipette the RNA molecule is stretched. b Measured force-extension cycle showing a force rip around 15pN characteristic of the unfolding/refolding of the RNA molecule. The work along a given unfolding/refolding force-extension curve corresponds to the area below the curve integrated along a given range of the molecular extension (indicated by the two *vertical dashed lines*). c Work distributions for the unfolding and refolding process measured at three different pulling speeds. According to the FT by Crooks all curves should cross at a given value of the work, $W = \Delta F$, independently of the pulling speed. In this case $\Delta F \simeq 110 k_B T$, a number that includes also the stretching contributions from the hybrid handles and the ssRNA. d Experimental verification of the FT by Crooks for four different tethered molecules (all with identical sequence). The *black dashed line* is the best fit for all curves and has slope equal to 0.9 ± 0.1. Results taken from [75]

molecule is a stochastic (i. e. random) process, the value of the force at which the molecule unfolds changes from experiment to experiment and so does the value of the mechanical work required to unfold the molecule. Upon repetition of the experiment many times a distribution of unfolding work values for the molecule to go from the folded (F) to the unfolded (U) state is obtained, $P_{\mathrm{F}\to\mathrm{U}}(W)$. A related work distribution can be obtained if we reverse the pulling process by releasing the molecular extension at the same speed at which the molecule was previously pulled, to allow the molecule to go from the unfolded (U) to the folded (F) state. In that case the molecule refolds by performing mechanical work on the cantilever or the optical trap. Upon repetition of the folding process many times the work distribution, $P_{\mathrm{U}\to\mathrm{F}}(W)$ can be also measured. The unfolding and refolding work distributions can then be measured in stretching/releasing cycles (Fig. 4c). From (20) we observe that $P_{\mathrm{F}}(\Delta F) = P_{\mathrm{R}}(-\Delta F)$ so the forward and reverse work probability distributions cross each other at $W = \Delta F$. In Fig. 4c we observe that both distributions cross each other at a value (ΔF) that is independent of the pulling speed as expected. Figure 4d shows the experimental verification of (20).

The FT by Crooks has been tested in different types of RNA molecules and the method has been shown capable of recovering free-energies under strong non-equilibrium conditions [75]. The work probability distributions were measured along the unfolding and refolding pathways for a three-way junction RNA molecule and found to strongly deviate from a Gaussian distribution [75]. These experimental results pave the way for other related studies, for example in molecular dynamics simulations [76].

These kind of studies will expand in the future to cover more complex cases and other non-equilibrium situations such as the free-energy recovery of folding free energies in native states in proteins or free energies in misfolded structures and intermediate states in RNA molecules and proteins. Ultimately FTs, when combined with SME, will offer an excellent opportunity to characterize and understand the possible biological relevance of large deviations and extremal fluctuations in molecular systems.

Future Directions

The experimental and theoretical study of non-equilibrium small systems offers exciting possibilities for the statistical physicist and the biophysicist. This discipline aims to describe the novel properties observed in bio-molecules and molecular machines operating far from equilibrium, such as the folding of a nucleic acid or a protein or the trans-location motion of a molecular motor.

We are just starting to have a glance about how these small objects exchange energy with their environment. It is a well known fact in molecular biology and biochemistry that biological function at the molecular level is tightly related to structure. It might not be surprising that the link between molecular structure and biological function is encoded in the low frequency region of the spectrum of non-equilibrium energy fluctuations (the spectrum of energy fluctuations extending far at the most extreme tails of the distribution). It is difficult to imagine how bio-molecular processes, often carrying a lot of information, can operate solely from high frequency events describing the motion of a few number of atoms. Rather, these should somehow rely on the low frequency cooperative motion between different and distant parts of the molecule. Investigating fluctuations in non-equilibrium systems calls for a deeper theoretical understanding of large deviation functions in non-equilibrium systems as well as more systematic and accurate experiments identifying sources of large energy fluctuations in biological systems.

We are at the dawn of an interdisciplinary scientific discipline that will bring together scientists with expertises coming from very different branches of knowledge. This merging process might culminate with the future engineering of artificial mesoscopic structures capable of reproducing and even improving the behavior of the biological ones.

Bibliography

Primary Literature

1. Bustamante C, Liphardt J, Ritort F (2005) The nonequilibrium thermodynamics of small systems. Phys Today 58:43–48
2. Hill TL (1994) Thermodynamics of small systems. Dover Publications, New York
3. Spudich A (2002) How molecular motors work. Nature 372: 515–518
4. Wang H, Oster G (2002) The Stokes efficiency for molecular motors and its applications. Europhys Lett 57:134–140
5. Bustamante C, Chemla YR, Forde NR, Izhaky D (2004) Mechanical processes in biochemistry. Annu Rev Biochem 73:705–748
6. Yin H, Wang MD, Svoboda K, Landick R, Block SM, Gelles J (1995) Transcription against an applied force. Science 270: 1653–1657
7. Wang MD, Schnitzer MJ, Yin H, Landick R, Gelles J, Block SM (1998) Force and velocity measured for single molecules of RNA polymerase. Science 282:902–907
8. Davenport RJ, Wuite GJ, Landick R, Bustamante C (2000) Single-molecule study of transcriptional pausing and arrest by E. coli RNA polymerase. Science 287:2497–2500
9. Forde NR, Izhaky D, Woodcock GR, Wuite GJL, Bustamante C (2002) Using mechanical force to probe the mechanism of pausing and arrest during continuous elongation by Es-

cherichia coli RNA polymerase. Proceedings of the National Academy of Sciences 99:11682–11687
10. Ediger MD, Angell CA, Nagel SR (1996) Supercooled liquids and glasses. J Phys Chem 100:13200–13212
11. Cipelletti L, Ramos L (2005) Slow dynamics in glassy soft matter. J Phys 17:R253–R285
12. Jarzynski C (1997) Non-equilibrium equality for free-energy differences. Phys Rev Lett 78:2690–2693
13. Kawai R, Parrondo JMR, den Broeck CV (2007) Dissipation: the phase-space perspective. Phys Rev Lett 98:080602
14. Ritort F, Bustamante C, Tinoco I (2002) A two-state kinetic model for the unfolding of single molecules by mechanical force. Proceedings of the National Academy of Sciences 99:13544–13548
15. Crooks GE (1999) Entropy production fluctuation theorem and the non-equilibrium work relation for free-energy differences. Phys Rev E 60:2721–2726
16. Crooks GE (2000) Path-ensemble averages in systems driven far from equilibrium. Phys Rev E 61:2361–2366
17. Kurchan J (1998) Fluctuation theorem for stochastic dynamics. J Phys A 31:3719–3729
18. Seifert U (2005) Entropy production along a stochastic trajectory and an integral fluctuation theorem. Phys Rev Lett 95:040602
19. Bochkov GN, Kuzovlev JE (1981) Non-linear fluctuation relations and stochastic models in non-equilibrium thermodynamics. I. Generalized fluctuation-dissipation theorem. Physica A 106:443–479
20. Zwanzig RW (1954) High-temperature equation of state by a perturbation method. i. non-polar gases. J Chem Phys 22: 1420–1426
21. Kirkwood JG (1935) Statistical mechanics of fluid mixtures. J Chem Phys 3:300–313
22. Cohen EGD, Mauzerall D (2004) A note on the Jarzynski equality. J Stat Mech P07006
23. Jarzynski C (2004) Non-equilibrium work theorem for a system strongly coupled to a thermal environment. J Stat Mech P09005
24. Astumian RD (2006) The unreasonable effectiveness of equilibrium-like theory for interpreting non-equilibrium experiments. Am J Phys 74:683
25. Evans DJ, Searles DJ (1994) Equilibrium micro-states which generate second law violating steady-states. Phys Rev E 50:1645–1648
26. Cohen EGD, Evans DJ, Morriss GP (1993) Probability of second law violations in shearing steady states. Phys Rev Lett 71: 2401–2404
27. Gallavotti G, Cohen EGD (1995) Dynamical ensembles in nonequilibrium statistical mechanics. Phys Rev Lett 74:2694–2697
28. Lebowitz JL, Spohn H (1999) A Gallavotti-Cohen type symmetry in the large deviation functional for stochastic dynamics. J Stat Phys 95:333–365
29. Maes C (1999) The fluctuation theorem as a Gibbs property. J Stat Phys 95:367–392
30. Oono Y, Paniconi M (1998) Steady state thermodynamics. Prog Theor Phys Suppl 130:29–44
31. Hatano T, Sasa S (2005) Steady-state thermodynamics of Langevin systems. Phys Rev Lett 86:3463–3466
32. Speck T, Seifert U (2005) Integral fluctuation theorem for the housekeeping heat. J Phys A 38:L581–L588
33. Ciliberto S, Laroche C (1998) An experimental test of the Gallavotti–Cohen fluctuation theorem. J Phys IV (France) 8:215–220
34. Ciliberto S, Garnier N, Hernandez S, Lacpatia C, Pinton JF, Ruiz-Chavarria G (2004) Experimental test of the Gallavotti–Cohen fluctuation theorem in turbulent flows. Physica A 340: 240–250
35. Wang GM, Sevick EM, Mittag E, Searles DJ, Evans DJ (2002) Experimental demonstration of violations of the second law of thermodynamics for small systems and short timescales. Phys Rev Lett 89:050601
36. Carberry DM, Reid JC, Wang GM, Sevick EM, Searles DJ, Evans DJ (2004) Fluctuations and irreversibility: An experimental demonstration of a second-law-like theorem using a colloidal particle held in an optical trap. Phys Rev Lett 92:140601
37. Trepagnier EH, Jarzynski C, Ritort F, Crooks GE, Bustamante C, Liphardt J (2004) Experimental test of Hatano and Sasa's nonequilibrium steady state equality. Proceedings of the National Academy of Sciences 101:15038–15041
38. Blickle V, Speck T, Helden L, Seifert U, Bechinger C (2005) Thermodynamics of a colloidal particle in a time-dependent nonharmonic potential. Phys Rev Lett 93:158105
39. Hummer G, Szabo A (2001) Free-energy reconstruction from non-equilibrium single-molecule experiments. Proc Nat Acad Sci 98:3658–3661
40. Jarzynski C (2001) How does a system respond when driven away from thermal equilibrium? Proc Nat Acad Sci 98: 3636–3638
41. Hummer G, Szabo A (2005) Free-energy surfaces from single-molecule force spectroscopy. Acc Chem Res 38:504–513
42. Zuckermann DM, Woolf TB (2002) Theory of systematic computational error in free energy differences. Phys Rev Lett 89:180602
43. Gore J, Ritort F, Bustamante C (2003) Bias and error in estimates of equilibrium free-energy differences from non-equilibrium measurements. Proc Nat Acad Sci 100:12564–12569
44. Liphardt J, Dumont S, Smith SB, Tinoco I, Bustamante C (2002) Equilibrium information from non-equilibrium measurements in an experimental test of the Jarzynski equality. Science 296:1833–1835
45. Schuler S, Speck T, Tierz C, Wrachtrup J, Seifert U (2005) Experimental test of the fluctuation theorem for a driven two-level system with time-dependent rates. Phys Rev Lett 94:180602
46. Douarche F, Ciliberto S, Petrosyan A (2005) An experimental test of the Jarzynski equality in a mechanical experiment. Europhys Lett 70:593–598
47. Douarche F, Ciliberto S, Petrosyan A (2005) Estimate of the free energy difference in mechanical systems from work fluctuations: experiments and models. J Stat Mech (Theor. Exp.) P09011
48. Van Zon R, Cohen EGD (2003) Extension of the fluctuation theorem. Phys Rev Lett 91:110601
49. Van Zon R, Cohen EGD (2003) Stationary and transient work-fluctuation theorems for a dragged Brownian particle. Phys Rev E 67:046102
50. Seifert U (2004) Fluctuation theorem for birth-death or chemical master equations with time-dependent rates. J Phys A 37:L517–L521
51. Jarzynski C, Wojcik DK (2004) Classical and quantum fluctuation theorems for heat exchange. Phys Rev Lett 92:230602
52. Seifert U (2005) Entropy production along a stochastic tra-

jectory and an integral fluctuation theorem. Phys Rev Lett 95:040602
53. Reid JC, Sevick EM, Evans DJ (2005) A unified description of two theorems in non-equilibrium statistical mechanics: The fluctuation theorem and the work relation. Europhys Lett 72: 726–730
54. Hummer G (2001) Fast-growth thermodynamic integration: Error and efficiency analysis. J Chem Phys 114:7330–7337
55. Isralewitz B, Gao M, Schulten K (2001) Steered molecular dynamics and mechanical functions of proteins. Curr Opin Struct Biol 11:224–230
56. Jensen MO, Park S, Tajkhorshid E, Schulten K (2002) Energetics of glycerol conduction through aquaglyceroporin Glpf. Proc Nat Acad Sci 99:6731–6736
57. Park S, Khalili-Araghi F, Tajkhorshid E, Schulten K (2003) Free-energy calculation from steered molecular dynamics simulations using Jarzynski's equality. J Phys Chem B 119: 3559–3566
58. Andriocioaei I, Dinner AR, Karplus M (2003) Self-guided enhanced sampling methods for thermodynamic averages. J Chem Phys 118:1074–1084
59. Park S, Schulten K (2004) Calculating potentials of mean force from steered molecular dynamics simulation. J Chem Phys 13:5946–5961
60. Bennett CH (1976) Efficient estimation of free-energy differences from Monte Carlo data. J Comput Phys 22:245–268
61. Wood RH, Mühlbauer WCF, Thompson PT (1991) Systematic errors in free energy perturbation calculations due to a finite sample of configuration space: sample-size hysteresis. J Phys Chem 95:6670–6675
62. Hendrix DA, Jarzynski C (2001) A "fast growth" method of computing free-energy differences. J Chem Phys 114:5974–5981
63. Shirts MR, Bair E, Hooker G, Pande VS (2003) Equilibrium free energies from non-equilibrium measurements using maximum-likelihood methods. Phys Rev Lett 91:140601
64. Qian H (2005) Cycle kinetics, steady state thermodynamics and motors – a paradigm for living matter physics. J Phys (Condensed Matter) 17:S3783–3794
65. Min W, Jiang L, Yu J, Kou SC, Qian H, Xie XS (2005) Non-equilibrium steady state of a nanometric biochemical system: Determining the thermodynamic driving force from single enzyme turnover time traces. Nanoletters 5:2373–2378
66. Mazonka O, Jarzynski C (1999) Exactly solvable model illustrating far-from-equilibrium predictions. Preprint arXiv:cond-mat/9912121
67. Lhua RC, Grossberg AY (2005) Practical applicability of the Jarzynski relation in statistical mechanics: a pedagogical example. J Phys Chem B 109:6805–6811
68. Bena I, Van den Broeck C, Kawai R (2005) Jarzynski equality for the Jepsen gas. Europhys Lett 71:879–885
69. Speck T, Seifert U (2005) Dissipated work in driven harmonic diffusive systems: general solution and application to stretching rouse polymers. Eur Phys J B 43:521–527
70. Cleuren B, Van den Broeck C, Kawai R (2007) Fluctuation and dissipation of work in a Joule experiment. Phys Rev Lett 98:080602
71. Van Zon R, Cohen EGD (2004) Extended heat-fluctuation theorems for a system with deterministic and stochastic forces. Phys Rev E 69:056121
72. Ritort F (2004) Work and heat fluctuations in two-state systems: a trajectory thermodynamics formalism. J Stat Mech (Theor. Exp.) P10016
73. Imparato A, Peliti L (2005) Work distribution and path integrals in mean-field systems. Europhys Lett 70:740–746
74. Imparato A, Peliti L (2005) Work probability distribution in systems driven out of equilibrium. Phys Rev E 72:046114
75. Collin D, Ritort F, Jarzynski C, Smith SB, Tinoco I, Bustamante C (2005) Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies. Nature 437:231–234
76. Kosztin I, Barz B, Janosi L (2006) Calculating potentials of mean force and diffusion coefficients from non-equilibrium processes without Jarzynski equality. J Chem Phys 124:064106

Books and Reviews

Comptes Rendus Physique (2007) Work, dissipation and fluctuations in non-equilibrium physics, vol 8(5–6). Elsevier

Duplantier B, Dalibard J, Rivasseau V (eds)(2004) Bose-Einstein Condensation and Entropy 2, Birkhäuser, Basel. Contributions by Maes C, On the origin and use of fluctuation relations on the entropy, p 145–191 and Ritort F, Work fluctuations, transient violations of the second law and free-energy recovery methods, pp 193–226. Available also at: http://www.ffn.ub.es/ritort/publications.html

Evans DJ, Searles D (2002) The fluctuation theorem. Adv Phys 51:1529–1585

Garbaczewski P, Olkiewicz R (eds)(2002) Dynamics of dissipation. Springer, Berlin. Contribution by Jarzynski C, What is the microscopic response of a system driven far from equilibrium, pp 63–82

Mathews CK, van Holde KE, Ahern KG (2000) Biochemistry. Addison-Wesley, Longman, SF

Ritort F (2008) Nonequilibrium fluctuations in small systems: From physics to biology. In: Rice SA (ed) Advances in Chemical Physics, vol 137. Wiley, Hoboken, NJ, pp 31–123

Fluid Dynamics, Pattern Formation

Michael Bestehorn
Brandenburg University of Technology,
Cottbus, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
The Basic Equations of Fluid Dynamics
Surface Waves
Instabilities
Order Parameter Equations
Conserved Order Parameter Fields
Future Directions
Bibliography

Glossary

Order parameter(s) (Field) variable(s) characterizing the spatio-temporal state of a system. Other state variables such as velocity, temperature, density, etc. can be computed if the order parameter(s) are known.

Control parameter(s) Parameters which are fixed and can be tuned from outside of the system under consideration.

Critical point, threshold, onset The points in control parameter space where new and qualitatively different solutions bifurcate from (usually simpler) ones.

Slaving principle Stated by H. Haken in 1975, the slaving principle allows for a huge reduction of degrees of freedom close to a critical point. It states that a very large number of linearly damped modes are slaved to and therefore completely determined by the few modes that grow in the vicinity of the critical point. The amplitudes of the growing modes are also called order parameters.

Natural patterns Spatial patterns showing a certain periodic (near) order, but also defects, grain boundaries etc.

Turing patterns Natural patterns that have a certain typical length scale and that show relaxation to a stationary state in the long term. Typical ingredients of Turing patterns are stripes, hexagons and squares.

Swift–Hohenberg equation Derived by Swift and Hohenberg in 1977 and nowadays established as the standard form for a scalar, real-valued order parameter equation showing Turing patterns at onset.

Coarsening The monotonic increase of the typical length scale of a structure in time. Often connected to spinodal decomposition, for example, of a binary mixture of non-mixing components such as water and oil. Small oil droplets in the beginning merge and finally form a large oil drop on the water surface. Coarsening slows down if the length scale increases.

Definition of the Subject

The state of a fluid is described by its velocity, density, pressure, and temperature. All these variables depend in general on space and time. Pattern formation refers to the situation where one or more of these variables are organized within a certain spatial and/or temporal order. This order has macroscopic length and time scales, that is, characteristic lengths and times are much larger than those of the atoms or molecules which constitute the fluid. Therefore a continuous description is appropriate.

Macroscopic fluid patterns may be encountered in nature as well as in technological applications for a large variety of different systems. Far from being complete, we mention some examples:

- Water waves caused by wind or by sea quakes and land slides (Tsunamis).
- Localized excitations of the surface of a fluid (solitons), such as that seen on shallow water channels.
- Shear instabilities in clouds or in multi-layer systems such as the Kelvin–Helmholtz instability or the Rayleigh–Taylor-instability.
- Surface deflections in the form of holes or drops of thin fluid films in coating or wetting processes.
- Convection instabilities in laboratory experiments, but also in the atmosphere, in the earth's interior or in stars.
- Creation and controlled growth of ordered structures in (nano-) technological applications.
- Biological applications: Behavior of liquid films on leaves or of the tear film on the cornea of the eye. Dynamics of thin blood layers, blood clotting.
- Films on the walls of combustion cells.
- Lubrication films in mechanical machines.

Fluid patterns may occur due to several mechanisms. One can distinguish between two main cases: Patterns excited and organized by some external forces or disturbances (such as Tsunamis) and those formed by instabilities. The latter may show the aspects of self-organization and will be the focus of the present contribution.

Introduction

Since the first observations of Michael Faraday almost 180 years ago [42] (Fig. 1), pattern formation in liquids or gases (*fluids*) has been subject to innumerable experimental [19,80,90], theoretical [32,34,43] and, later on, numerical work [5,26,73]. After the famous experiments by Henri Bénard around 1901 [17], convection in a single or later in multi-component fluids came into the focus of interest. The first theoretical studies were made by Lord Rayleigh [78]. Theoretical computations up to the early 1960s were restricted on the linearized basic equations and could explain the existence of critical points in parameter space as well as the observed length scales of the structures found experimentally [18,72]. In the meantime, Alan Turing [89] showed in his famous paper of 1952 that similar patterns could emerge out of equilibrium in reaction-diffusion systems. It took almost 40 years for an experimental confirmation using the so-called CIMA reaction [31,71].

With the appearance and rapid development of computers, the field gained further momentum from the new discipline of nonlinear dynamics and nonlinear system

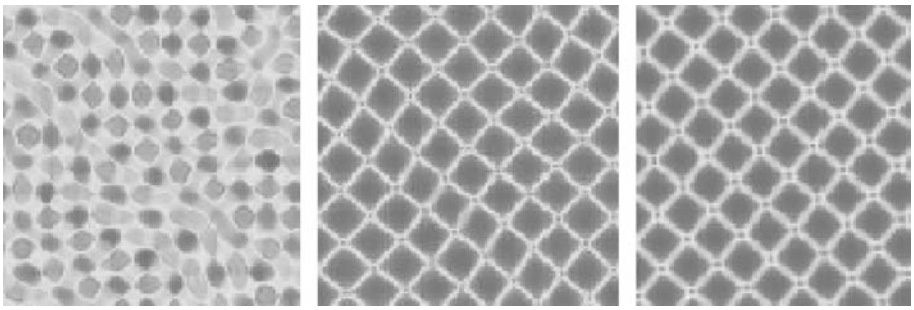

Fluid Dynamics, Pattern Formation, Figure 1
Michael Faraday observed surface patterns on a liquid horizontal layer if the whole layer vibrated vertically with a certain amplitude and frequency. Very often, regular squares are found, as shown in the time series as a numerical result of the shallow water equations (see Sect. "Surface Waves")

theory [4,45,48,50]. Early computations in 1963 by Edward Lorenz of a system of three coupled ordinary differential equations derived by a crudely truncated mode expansion of the Navier–Stokes equations revealed the first chaotic attractor of a dissipative system [60]. Though the chaotic behavior seen in the Lorenz equations does not originate from hydrodynamic equations and has nothing to do with irregular fluid behavior, the Lorenz model now stands on its own as a paradigm for a relatively simple system showing low dimensional chaos [87].

Patterns that emerge from an instability roughly pass through two phases. As long as amplitudes (or order parameters) are small, the behavior is often determined by the linear parts of the system and exponential growth of a certain part of the mode spectrum is found. In the second phase, nonlinearities come into play and may lead to saturation and selection of certain mode configurations, seen then as regular structures in configuration space (Fig. 2). The full mathematical description of hydrodynamic systems has been well known for a long time. Fluid motion is described by the Euler or Navier–Stokes equations, temperature fields by the heat equation and chemical concentrations by some nonlinear reaction-diffusion equations. The location and spatio-temporal evolution of surfaces or interfaces can be computed by the kinematic boundary conditions if the velocity of the fluid near the interface is known. All these equations can be coupled and provided with suitable boundary and initial conditions, resulting in rather complicated systems of nonlinear partial differential equations. Even today in the age of supercomputers, their further treatment, especially in three spatial dimensions, remains a challenge.

On the other hand, directly solving the basic equations, can be considered merely as another experiment. For these reasons and to get a deeper insight into the physics behind pattern formation, other methods have been devised. Very often one of the three spatial dimension is distinguished, either for physical reasons or simply due to the geometry of the system. A good example is surface waves on a water layer. Here, the behavior of the solutions in the vertical direction (z) is very different from those in the horizontal ones. For shallow water waves (wave length long compared to the layer depth) the velocities are more or less independent on z, where in the other limit of deep water waves, fluid motion takes only place along a small layer under the surface and decreases exponentially with depth. In both cases one may reduce the dimension of the basic problem by an expansion with respect to simple functions for the vertical dependence of the variables [33]. An analogue method can be applied describing thin film surface patterns [70]. Also for convection cells, the vertical dimension plays a special role and the solution can be projected onto a few modes near the critical point [25,73].

Another concept that reduces the number of dependent variables and equations is that of order parameters. The notion of the "order parameter" goes back to Landau [58] and refers originally to a variable that measures the order of a certain system. Rather a variable than a parameter, the order parameter normally depends on time and, in theories describing the formation of natural patterns, also on space [66]. Thus, the order parameter equation (abbreviated: OPE) is a partial differential equation with certain nonlinear terms that become important for pattern selection and saturation.

Theoretical methods developed by the Haken school [47,48,50] starting in the 1970s allow for a systematic derivation of the OPEs (sometimes also called "generalized Ginzburg–Landau equations") for a great variety of nonequilibrium and open systems from physics, chemistry and biology. The key idea is to find a reduced

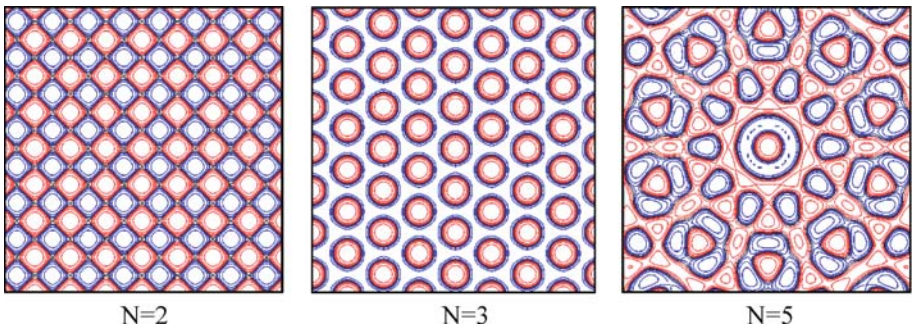

Fluid Dynamics, Pattern Formation, Figure 2
The composition of plane waves with the same wave number but different orientation in 2D space results in regular patterns. For two modes ($N = 2$) one sees squares, for $N = 3$ hexagons and for $N > 3$ quasi periodic structures in space or Penrose tilings [74] are found

description in terms of relevant or active modes close to a certain bifurcation point. The amplitudes of these active modes, the order parameters, now generalized to a nonequilibrium, pattern forming system, obey unified and simplified equations, namely the OPEs. It turns out that the structure of these equations depends not so much on the particular system under consideration as on the type of bifurcation. To each type of bifurcation a special "normal form" of OPE is related [35]. In deriving the OPEs, the slaving principle [50] allows us to eliminate a huge number of slaved variables and express them by the active ones.

This contribution is concerned mainly with structures in fluids that originate from self-organized processes. It tries to bring together direct numerical solutions of hydrodynamic equations with the modern concepts of pattern formation. After introducing the basic equations (Sect. "The Basic Equations of Fluid Dynamics") of fluid dynamics, it presents a short section on waves and descriptions reduced by geometrical reasons. Several types of instabilities are discussed in Sect. "Instabilities", together with computer solutions for the different cases. Section "Order Parameter Equations" presents different types of two-dimensional order parameter equations. Finally, Sect. "Conserved Order Parameter Fields" is devoted to the special case of conserved order parameters.

The Basic Equations of Fluid Dynamics

Let the state of a fluid be described by its velocity, its density, its pressure, and its temperature field

$$\vec{v}(\vec{r},t)\,, \quad \rho(\vec{r},t)\,, \quad p(\vec{r},t)\,, \quad T(\vec{r},t)\,. \tag{1}$$

In this section we wish to specify the basic hydrodynamic equations that rule the spatio-temporal behavior of these seven variables. They have to be completed by suitable boundary conditions (abbreviated: b.c.) which we shall present later with the particular systems under consideration.

Continuity Equation

The conservation of mass yields the continuity equation

$$\partial_t\rho + \mathrm{div}(\rho\vec{v}) = \partial_t\rho + (\vec{v}\cdot\nabla)\rho + \rho\ \mathrm{div}\,\vec{v} = 0\,. \tag{2}$$

In most cases, liquids are difficult to compress. One can usually assume that a volume element does not change its density while it moves with the fluid (Lagrangian description)

$$\partial_t\rho + (\vec{v}\cdot\nabla)\rho = 0\,.$$

From (2) one finds the condition of incompressibility

$$\mathrm{div}\,\vec{v}(\vec{r},t) = 0\,, \tag{3}$$

or, in other words, the velocity field is free of sources and sinks. Equation (3) can be satisfied by the ansatz

$$\vec{v}(\vec{r},t) = \mathrm{curl}\,\vec{A}(\vec{r},t) \tag{4}$$

where $\vec{A}$ plays the role of a vector potential. In (4) one can use the particular decomposition [5,32]

$$\vec{v}(\vec{r},t) = \mathrm{curl}(\Phi\hat{e_z}) + \mathrm{curl}\,\mathrm{curl}(\Psi\hat{e_z}) = \begin{pmatrix} \partial_y\Phi + \partial_z\partial_x\Psi \\ -\partial_x\Phi + \partial_z\partial_y\Psi \\ -\Delta_2\Psi \end{pmatrix} \tag{5}$$

with the two independent scalar functions $\Phi(\vec{r},t)$ and $\Psi(\vec{r},t)$ and $\Delta_2 = \partial^2_{xx} + \partial^2_{yy}$ as the 2D-Laplacian.

If the velocity field is irrotational, that is without vortices ($\operatorname{curl}\vec{v} = 0$), it can be derived from a scalar potential

$$\vec{v} = \operatorname{grad}\phi\,. \tag{6}$$

For incompressible and irrotational flows, hydrodynamics is reduced to a boundary value problem, since the potential must fulfill the Laplace equation

$$\operatorname{div}\vec{v} = \Delta\phi = 0 \tag{7}$$

and the velocity field is solely determined by its boundary conditions.

Euler Equations

For a perfect fluid, a fluid with no viscosity, one derives the Euler equations from the law of conservation of momentum [56]. They read

$$\rho(\vec{r},t)\left[\partial_t\vec{v}(\vec{r},t) + (\vec{v}(\vec{r},t)\cdot\nabla)\,\vec{v}(\vec{r},t)\right] = -\operatorname{grad}\; p(\vec{r},t) + \vec{f}(\vec{r},t)\,, \tag{8}$$

where $\vec{f}$ denotes external volume forces. Together with a state equation of the form

$$p = p(\rho, T)\,, \tag{9}$$

the continuity equation (2) and the temperature equations (to be shown below) (Subsect. "Transport Equations") constitute the basic set for the seven state variables (1).

Incompressible Fluids For an incompressible fluid, a state equation of the form (9) makes no sense since pressure will not change with density. So p can be eliminated by forming the curl of (8)

$$\partial_t\vec{\Omega} = \operatorname{curl}(\vec{v}\times\vec{\Omega}) + \frac{1}{\rho}\operatorname{curl}\vec{f} \tag{10}$$

where

$$\vec{\Omega} = \operatorname{curl}\vec{v} \tag{11}$$

denotes the vorticity. If p must be known, it can be computed from the divergence of (8) which yields

$$\nabla^2 p = \rho\left\{-\operatorname{Tr}\left[(\nabla\circ\vec{v})(\nabla\circ\vec{v})\right] + \operatorname{div}\vec{f}\,\right\}\,, \tag{12}$$

where $\circ$ is the dyadic product and $\operatorname{Tr}[\ldots]$ the trace.

Incompressible Irrotational Fluids If, in addition, the flow is free of vortices, one may integrate the Euler equations and find the theorem of Bernoulli

$$\partial_t\phi = -\frac{1}{\rho}(p+U) - \frac{1}{2}(\nabla\phi)^2 \tag{13}$$

where U is the potential to $\vec{f}$ ($\vec{f}$ must be irrotational, too). For stationary solutions, the velocity potential ϕ is found from (7) and (13) can be used to determine the pressure.

Navier–Stokes Equations

Compressible Fluids In real fluids, shear stresses are a result of friction. They must be added to the balance of momentum and yield the Navier–Stokes equations. For a compressible Newtonian fluid they read

$$\rho\left[\partial_t\vec{v} + (\vec{v}\cdot\nabla)\,\vec{v}\right] = -\operatorname{grad}p + \vec{f} + \eta\,\Delta\vec{v} + \left(\zeta + \frac{\eta}{3}\right)\operatorname{grad}\operatorname{div}\vec{v} \tag{14}$$

where η denotes the first and ζ the second viscosity [59].

Incompressible Fluids The Navier–Stokes equations for incompressible fluids are simpler:

$$\rho\left[\partial_t\vec{v} + (\vec{v}\cdot\nabla)\,\vec{v}\right] = -\nabla p + \vec{f} + \eta\,\Delta\vec{v}\,. \tag{15}$$

Again, pressure can be eliminated by forming the curl. Taking the ansatz (5), the z-components of the curl and of the curl of the curl of (15), we have

$$\{\nu\;\Delta - \partial_t\}\,\Delta_2\Phi(\vec{r},t) = \left[\operatorname{curl}((\vec{v}\cdot\nabla)\vec{v})\right]_z + \frac{1}{\rho}\left(\partial_x f_y - \partial_y f_x\right) \tag{16a}$$

$$\{\nu\;\Delta - \partial_t\}\,\Delta\Delta_2\Psi(\vec{r},t) = \left[\operatorname{curl}\operatorname{curl}((\vec{v}\cdot\nabla)\vec{v})\right]_z + \frac{1}{\rho}\left(\Delta_2 f_z - \partial_x\partial_z f_x - \partial_y\partial_z f_y\right)\,. \tag{16b}$$

Here we have introduced the kinematic viscosity $\nu = \eta/\rho$. We note that this decomposition is of particular interest if $f_x = f_y = 0$ as is the case in convection problems of a plane layer.

Incompressible Fluids with a Small Reynolds Number For some applications it is convenient to use the Navier–Stokes equations in dimensionless form. With scaling with respect to a characteristic length L and velocity V_0:

$$\vec{r} = L\cdot\vec{r}\,'\,,\quad t = (L/V_0)\cdot t'\,,\quad \vec{v} = V_0\cdot\vec{v}\,'\,,\quad p' = \frac{L}{\eta V_0}\cdot p\,,$$

(15) turns into

$$R_e \left[\partial_{t'}\vec{v}\,' + (\vec{v}\,' \cdot \nabla')\,\vec{v}\,'\right] = -\nabla' p' + \Delta'\vec{v}\,' \,. \tag{17}$$

(We assumed a potential for $\vec{f}$ which can be confined into p.) The dimensionless quantity

$$R_e = \frac{LV_0}{\nu} \tag{18}$$

is the Reynolds number. If $R_e \ll 1$, the left hand side of (17) can be neglected and the Navier–Stokes equations become linear (primes omitted):

$$\Delta\vec{v} = \nabla p \,. \tag{19}$$

This is the Stokes equation, in which no time derivative of $\vec{v}$ occurs. Thus, as known from over-damped motion, the velocity field directly follows the pressure gradients.

Transport Equations

Scalar fields such as temperature or concentration of a mixture that may diffuse into and be transported with the fluid are ruled by the transport equation. Let $S(\vec{r}, t)$ be the scalar field, then the transport equation reads

$$\partial_t S + (\vec{v} \cdot \nabla)\, S = D_s \Delta S \,, \tag{20}$$

where D_s is the appropriate diffusion coefficient.

Surface Waves

The only elastic forces in fluids are those coming from volume changes and may exist, therefore, only in compressible fluids. They give rise to longitudinal compression waves which usually have small amplitudes and behave linearly in a good approximation. A linear wave equation can be derived with a (space dependent) sound speed [56].

A transversal wave which is also possible in incompressible fluids can be formed along a deformable interface. Gravity and, for small wavelengths, surface tension provide the stabilizing mechanism of a flat surface, around which oscillations (gravity waves) may occur. If their amplitudes are big enough, nonlinearities may play an essential role for surface waves, as is clearly seen by solitons [40,64] and wave breaking [39]. For this reason we shall discuss only surface waves in this section.

Gravity Waves

If one assumes an irrotational flow of a perfect and incompressible fluid on a flat substrate and with a free, deformable surface (Fig. 3), then the velocity is determined by the Laplace equation (7) which must be accomplished by boundary conditions at $z = 0$

$$v_z\big|_{z=0} = \partial_z\phi\big|_{z=0} = 0 \tag{21}$$

and at $z = h(x, y, t)$

$$\partial_t\phi\big|_{z=h} = -gh - p(h)/\rho - \frac{1}{2}\,(\nabla\phi)^2 \tag{22}$$

where g denotes the gravitational acceleration. Equation (22) is nothing other than the Bernoulli equation (13) evaluated at the surface. The surface itself is determined by the so-called kinematic boundary condition that reads (see Fig. 3, right frame).

$$\begin{aligned}\partial_t h &= v_z\big|_{z=h} - v_x\big|_{z=h}\partial_x h - v_y\big|_{z=h}\partial_y h \\ &= \partial_z\phi\big|_{z=h} - (\partial_x h)\,(\partial_x\phi)_{z=h} - (\partial_y h)(\partial_y\phi)_{z=h} \,.\end{aligned} \tag{23}$$

Shallow Water Equations

For shallow water waves, one can introduce the small parameter

$$\delta = \mathrm{d}/\ell \tag{24}$$

which is the ratio of the water depth d and a typical horizontal scale (such as a wavelength) ℓ. Then (7) can be

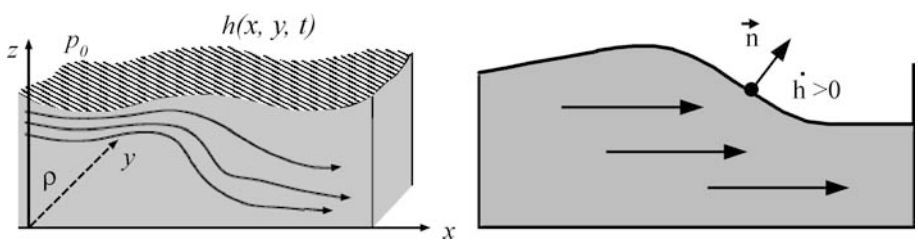

Fluid Dynamics, Pattern Formation, Figure 3
***Left*: an (incompressible) fluid with a free and deformable surface located at $z = h(x, y, t)$, on which a constant external pressure p_0 is applied. *Right*: The location of a certain point of the surface changes if the fluid is in motion**

solved iteratively; the result is a power series in δ^2 [33,39]:

$$\phi(\vec{r},t) = \Phi(x,y,t) + \delta^2\left[-\frac{z^2}{2}\Delta_2\Phi(x,y,t) + \varphi^{(1)}(x,y,t)\right] + O(\delta^4) \tag{25}$$

with an arbitrary function φ_1. Inserting (25) into (23) and (22) yields up to the lowest order in δ the *shallow water equations*

$$\partial_t h = -h\Delta_2\Phi - (\partial_x h)(\partial_x\Phi) - (\partial_y h)(\partial_y\Phi) \tag{26a}$$

$$\partial_t\Phi = -gh - p(h)/\rho - \frac{1}{2}(\partial_x\Phi)^2 - \frac{1}{2}(\partial_y\Phi)^2 . \tag{26b}$$

This is the first example of how to derive a two-dimensional system starting from three-dimensional fluid motion. Equations (26) constitute a closed system of partial differential equations for the evolution of the two functions $h(x,y,t)$ and $\Phi(x,y,t)$. Using (25), one immediately finds from the latter the velocity field (up to the order δ^2).

Numerical Solutions

Figure 4 shows numerical solutions of the shallow water equations (left frame in one dimension, right frame in two dimensions). In one dimension, one sees clearly traveling surface waves which may run around due to the periodic boundary conditions in x. On the other hand, one can recognize a second wave with a smaller amplitude going to the left hand side. Both waves seem to penetrate each other without further interaction. The reason seems to be the smallness of the amplitude which results in a more or less linear behavior. In the two-dimensional frame, a snapshot of the temporal evolution of the surface is presented. The initial condition was chosen randomly. For numerical stability reasons, an additional damping of the form $\tilde{\nu}\Delta_2\Phi$ was added to the right hand side of (26b) which filters out the short wave lengths. This could be justified phenomenologically by friction and leads in the long term to a fluid at rest, if only gravity acts.

Instabilities

Laplace Pressure and Disjoining Pressure To discuss Eqs. (26) further, we must elaborate a little on the dependence of the surface pressure on the depth $h(x,y,t)$ and on its curvature $-\Delta h$.

The length scale of the surface structures is proportional to the depth of the fluid layer. If the films are very thin, we expect to have scales in the range or even well below the capillary length $a = \sqrt{\Gamma/g\rho}$ where Γ denotes the surface tension. Then one has to take into account the additional pressure which originates from the curvature of the surface, the so-called *Laplace pressure* [59] $-\Gamma\Delta_2 h$. Thus we substitute in (26b)

$$p(h) = p_1(h) - \Gamma\Delta_2 h , \tag{27}$$

with a function p_1 (the disjoining pressure) that will be specified later [38,91].

Linear Stability Analysis of the Flat Surface To see if the flat film $h = h_0$ is stable against small perturbations, one may perform a linear stability analysis. Inserting

$$(h - h_0, \Phi) = (a, b)\exp(\lambda t + ikx)$$

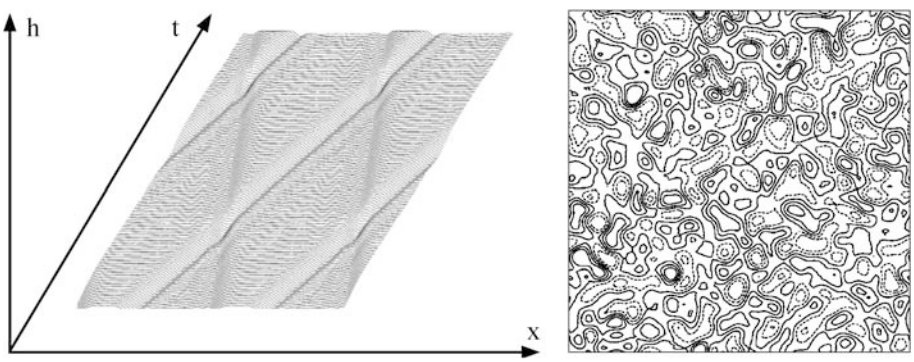

Fluid Dynamics, Pattern Formation, Figure 4
Numerical solutions of the shallow water equations, *left frame* shows a temporal evolution in one dimension, *right frame* a snapshot in two dimensions. *Dashed contour lines* mark troughs, *solid ones* correspond to peaks of the sea

into (26) yields, after linearization with respect to a, b, a linear eigenvalue problem with the solvability condition

$$\lambda_{12}(k) = -\frac{\tilde{\nu} k^2}{2} \pm i|k| \sqrt{h_0\big(g + p_1' + \Gamma k^2\big)/\rho - \tilde{\nu}^2 k^2/4}\,, \quad (28)$$

where

$$p_1' = \left.\frac{\mathrm{d}p_1}{\mathrm{d}h}\right|_{h=h_0}.$$

We assume that the artificial viscosity is small, $\tilde{\nu}^2 \ll \Gamma/\rho$. An instability occurs first at $k = 0$ if the expression under the integral can be negative, that is, for $p_1' + \rho g < 0$. This corresponds to the region of initial thickness h_0 where the generalized pressure

$$p_1(h) + \rho g h \quad (29)$$

has a negative slope. For that case, the real part of λ_1 starts at $k = 0$ at zero with positive slope, has a maximum at $k = k_c$ and decreases again to the value $-\nu k^2/2$. We shall revisit this instability in the next section and call it there, in a more systematic classification, a type II instability. How can (29) have a negative slope for a certain range of h_0? It is obvious that one has to assume that the pressure p_1 depends in some nonlinear nonmonotonic fashion on the value of h (Fig. 5). As we shall see later, this can be the case for very thin films where van der Waals forces between the solid support and the free surface come into play [38,52,91]. But also, in thicker films, this should be possible in non-isothermal situations, where the surface temperature, and therefore the surface tension, changes with the vertical coordinate (Marangoni effect, see Sect. "Instabilities", Fig. 8). If we take (for instance) as a model the polynomial

$$p_1 = c \cdot h \cdot (h - h_1) \cdot (h - h_2)\,, \quad c > 0\,, \quad (30)$$

then the flat surface is unstable for h between the two spinodals

$$h_- < h < h_+$$

with $h_\pm$ being the roots of

$$3h^2 - 2h(h_1 + h_2) + h_1 h_2 + \rho g/c = 0\,.$$

Figure 6 shows a numerically determined time series of a random dot initial condition. The mean thickness h_0 was chosen in the unstable region. The formation shows traveling waves in the linear phase, followed by coarsening to a large scale structure, in this case one big region of depression, or a hole. This hole becomes more and more symmetric while the velocity decays due to the friction term. Finally, a steady state of a big circular hole remains.

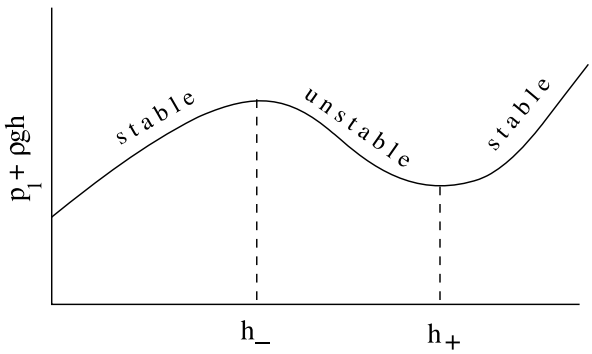

Fluid Dynamics, Pattern Formation, Figure 5
If the pressure depends on *h* and has a certain region with a negative slope, the flat film is unstable in this region and pattern formation sets in

Parametric Excitation of a Thin Bistable Fluid Layer
One way to replace the energy lost by damping (to "open" the system) is to accelerate the whole layer periodically in the vertical direction. This was done first in an experiment by Michael Faraday in 1831 [42]. He obtained regular surface patterns normally in the form of squares, see Fig. 1.

Faraday patterns can be seen as a solution of the shallow water equations if the gravity constant g is modulated harmonically [7]

$$g(t) = g_0 + g_1 \cos \omega t\,. \quad (31)$$

A linear stability analysis leads to a Mathieu equation [1]. The flat film is unstable if frequency and amplitude fall into certain domains, the so-called *Arnold tongues*. There, one usually finds squares for not-too-supercritical values.

We conclude this section by showing a numerical solution of (26) with parameters as in Fig. 6, but now with an additional periodic excitation (Fig. 7). Coarsening is still present, but now oscillating drops emerge in the form of stars. No time stable structure is found in the long time limit.

Instabilities

Mechanisms of Instability in Fluids

We start with the specific case of a plane layer of a viscous fluid with a vertically applied, constant temperature gradient β (Fig. 11), where

$$\beta = (T_1 - T_0)/d \quad (32)$$

and T_0, T_1 are the temperatures at the lower, upper side

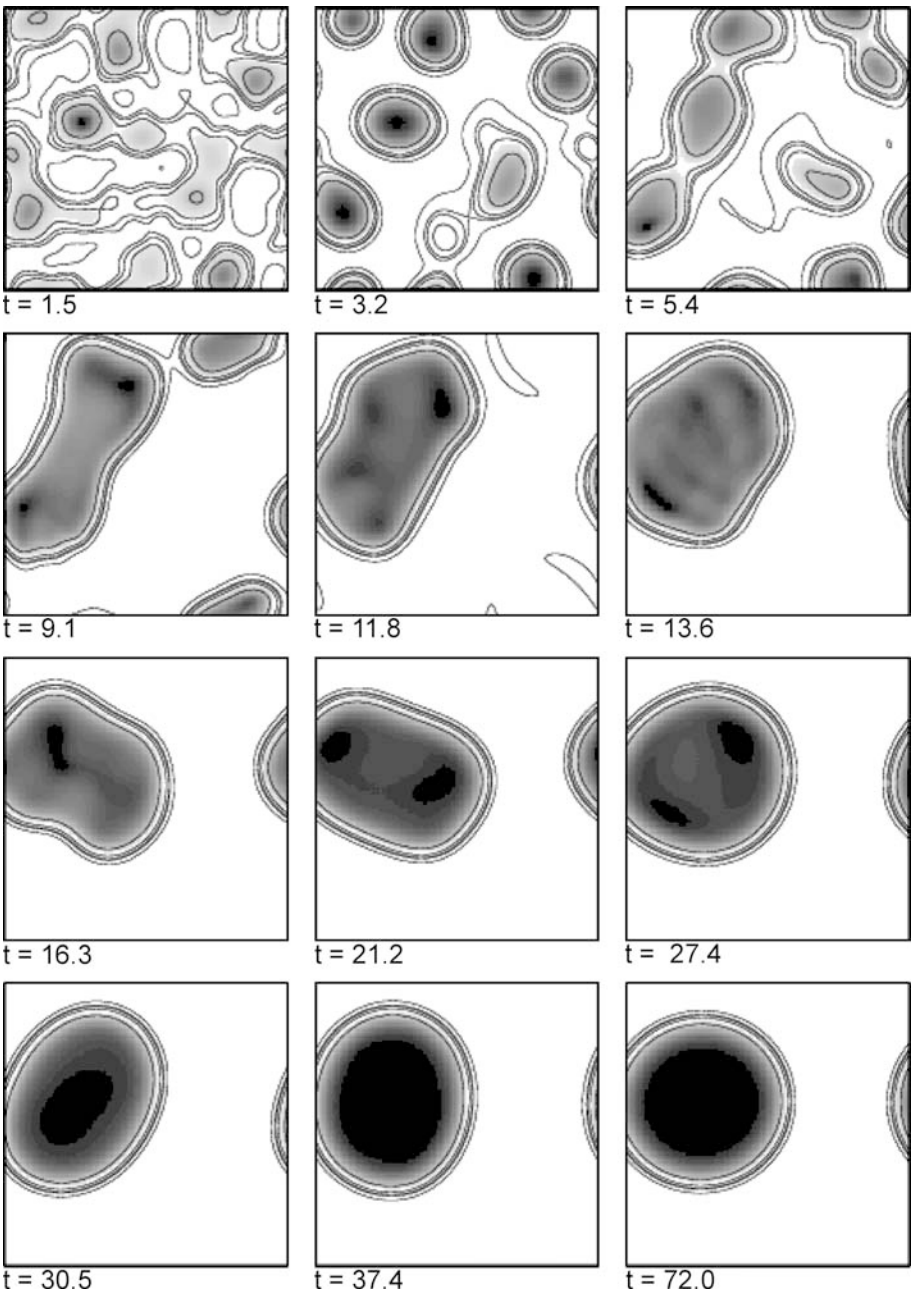

Fluid Dynamics, Pattern Formation, Figure 6
Time series from a numerical solution of (26) with artificial damping and bistable pressure according to (30). Coarsening dominates the nonlinear evolution and eventually a stationary circular region of surface depression (a hole) remains. Periodic boundary conditions in both horizontal directions have been used

of the layer. We assume that a motionless stationary state exists as a (stable or unstable) solution of (15) and of an equation such as (20) for the temperature. The temperature and pressure distribution of that state can then be computed from (15, 20) by putting $\vec{v}$ and all time derivatives to zero:

$$\nabla p^0 = \vec{f} \tag{33a}$$

$$\Delta T^0 = 0\,. \tag{33b}$$

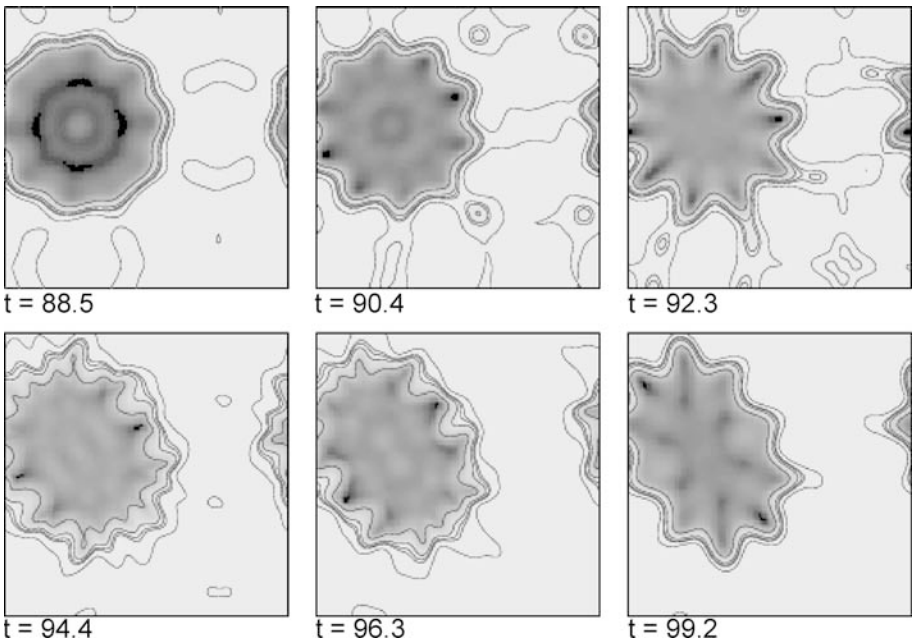

Fluid Dynamics, Pattern Formation, Figure 7
Continuation of the series of Fig. 6, but with additional parametric excitation according to (30) switched on at $t = 72$. Instead of stationary patterns pulsating stars are found

If an external force is provided by buoyancy, we may align the z-axis of the coordinate system along $\vec{f}$ which yields

$$\vec{f} = -g\,\rho(\vec{r})\,\hat{e}_z\,,$$

where g is the gravitational acceleration. Equation (33a) can be solved only if ρ does not depend on x and y. If one assumes that the density depends on temperature

$$\rho = \rho(T^0) \tag{34}$$

then T^0 can also depend only on z. Thus one finds from (33b)

$$T^0(z) = a + bz\,. \tag{35}$$

Taking a linear relation for (34)

$$\rho(T) = \rho_0[1 - \alpha(T - T_0)] \tag{36}$$

with the heat expansion coefficient $\alpha \equiv -\rho_0^{-1}\,\mathrm{d}\rho/\mathrm{d}T$ and ρ_0 as the density at the reference temperature T_0, one may integrate (33a) and find for the pressure of the motionless state (hydrostatic pressure)

$$p^0(z) = -g\int \rho\mathrm{d}z = -g\rho_0\left(z - \frac{1}{2}\alpha\beta z^2\right) \tag{37}$$

where we put $a = T_0$ and $b = \beta$, in agreement with (32).

A linear stability analysis [32] shows that the motionless, nonequilibrium state (35) can become unstable if the temperature gradient β exceeds a certain critical value, depending on the fluid properties and the geometry of the layer. There are two different mechanisms, if the fluid layer is heated from below:

(1) *Buoyancy*: Hot fluid particles (volume elements) near the bottom are lighter than colder ones and want to rise. Colder particles near the top want to sink. If the stabilizing forces of thermal conduction and friction in the fluid are exceeded by the externally applied temperature gradient, patterned fluid motion sets in.

(2) *Surface tension*: If the upper surface of the fluid is free, that is, in contact with the ambient air, tangential surface tension normally increases with decreasing surface temperature (Fig. 8a–c). If a fluid particle near the surface moves by fluctuations, say, to the right, then warmer fluid is pulled up from the bottom, increasing the surface temperature locally. Due to laterally increasing surface tension with respect to the neighbored points, even more hot fluid is pumped up from the bottom and the fluid starts to move. This is called the Marangoni effect and works even without gravity, that is, in space experiments.

In both cases, the typical length of the structures which bifurcate from the motionless state is of the order of the layer depth. These instabilities are sometimes called *small scale instabilities*.

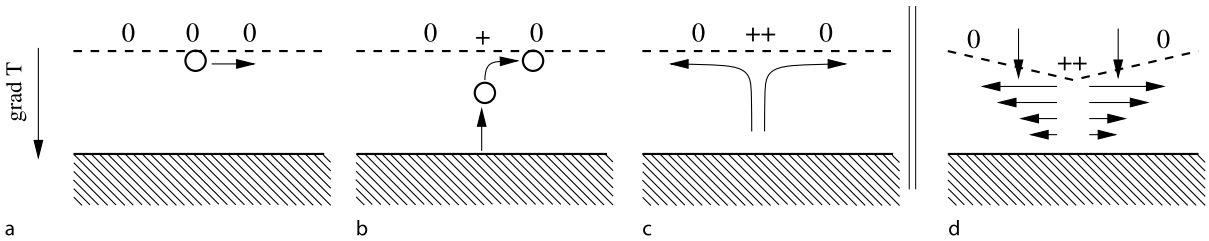

Fluid Dynamics, Pattern Formation, Figure 8
a–c The Marangoni effect may destabilize a fluid layer at rest and may generate a (regular) fluid motion. The surface remains flat (to a good approximation). If the surface is deformable d, a large scale instability may occur as a consequence of the Marangoni effect and mass conservation. For both instabilities, it is sufficient to assume the surface tension as a linear function of temperature. (+/-/0) denote relative temperatures

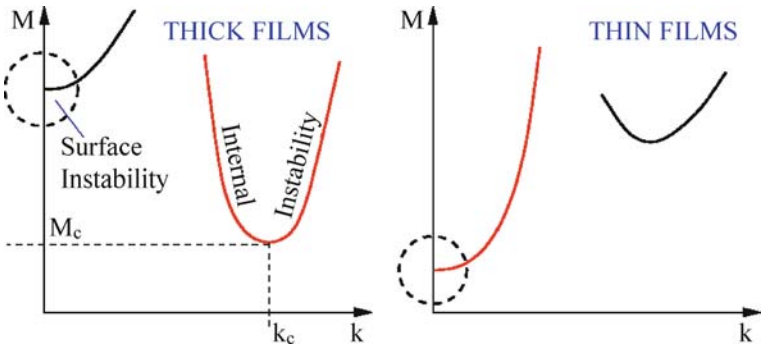

Fluid Dynamics, Pattern Formation, Figure 9
The two cases "thick films" and "thin films" are defined by the instability that comes first when the temperature gradient is increased. The two instabilities differ in the wavelength Λ (wave number $k = 2\pi/\Lambda$) of the growing structures

In the situation described above, the surface can be assumed to be flat and undeformable. Of course this is only an approximation, but valid for not-too-thin fluid layers and parameters not too far from threshold. If, on the other hand, the thickness of the fluid layer is less than a certain value which is on the order of 10^{-4} m for common silicone oils, another mechanism comes to the foreground. This mechanism is based on

Surface Deformation. If the surface is locally depressed by an arbitrary fluctuation, the depressed part is heated up due to the vertical temperature gradient. A lateral surface tension gradient is formed which pulls the liquid outside the depressed region (Fig. 8d). Since the continuity equation must hold, the surface becomes even more depressed and an instability occurs. The same mechanism leads to the growth of elevated parts of the surface, under which fluid is pumped in from adjacent regions [44,65,70]. The deformation mode belongs to the so-called *large-scale instability*. This means that the fastest growing modes have a wavelength that is very large compared to the layer depth. It is the depth of the layer which distinguishes which instability occurs first if the temperature gradient is increased from the sub-critical region (Fig. 9).

In ultra-thin films (depth of few 100 nm or less), other mechanisms are possible. Van der Waals forces between the free surface and the solid substrate then become important. They have a potential and can be expressed in the pressure by an extra term, *disjoining pressure*, as already shown in Sect. "Surface Waves". If that pressure increases with decreasing layer depth, fluid is pressed out of depressed regions and pumped into elevated regions and an instability occurs, even for isothermal cases (Fig. 10).

Pattern Formation – Examples

What happens if the critical value for the temperature gradient is exceeded? Since the famous experiments of Henri Bénard [17] in the beginning of the 20th century, one knows that the fluid starts to move in form of hexagons if the surface is free and the layer is "thick" (Fig. 11). These kinds of experiments were repeated many times under ex-

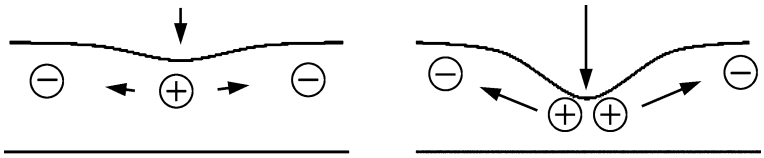

Fluid Dynamics, Pattern Formation, Figure 10
In ultra-thin films [79,85,92], van der Waals forces between free surface and solid substrate may destabilize a plane fluid layer even without an external temperature gradient (+/— denote relative values of the disjoining pressure)

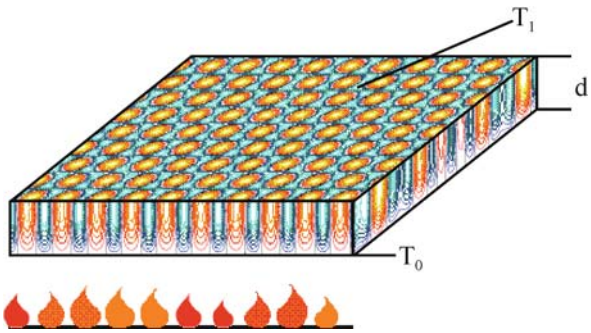

Fluid Dynamics, Pattern Formation, Figure 11
Hexagonal motion of a fluid heated from below, found by computer solution. Shown are contour lines of the temperature field (after [5])

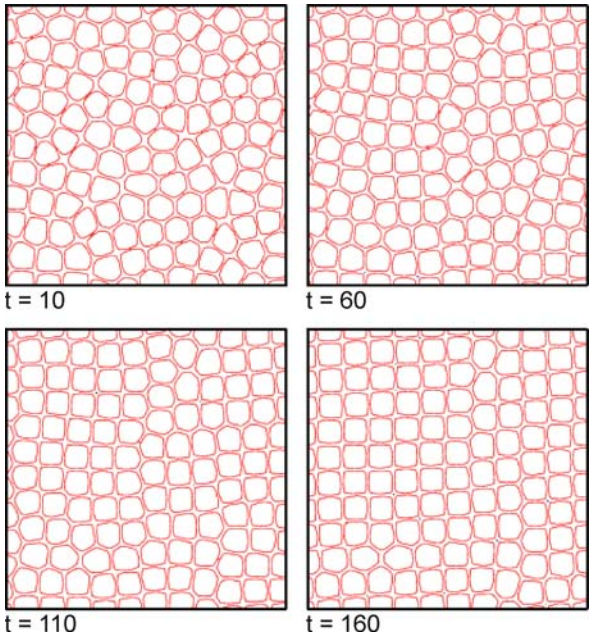

Fluid Dynamics, Pattern Formation, Figure 12
Regular squares as a secondary instability of hexagons. Numerical solution of the basic equations [6]

cellent conditions, for free and closed surfaces, with different fluids, even under micro gravity conditions [55,83].

Surprisingly, a secondary instability takes place for a larger external temperature gradient, which was not known before 1995, almost 100 years after Bénard. This instability shows the occurrence of rather regular squares and was discovered by Eckert and Thess in Dresden, Germany [6,41,67] and, in the meantime but independently, by Schatz and Swinney in Austin, Texas [80,81] (Fig. 12).

If the fluid is covered by a good thermal conductor (a sapphire plate, for instance), hexagons are not the typically found structure at onset, but rather stripes or rolls are encountered [90]. This can be understood in the frame of reduced order parameter equations by simple symmetry arguments. We shall discuss this in more detail in Sect. "Order Parameter Equations". For small Prandtl numbers (the ratio between viscosity and thermal diffusivity of the fluid) more complicated and time dependent patterns are found in the form of spirals (Fig. 13) [15,61,73].

The initial growth of patterns, with a certain horizontal length scale of the order of the depth of the fluid layer, is typical for pattern formation in thick films. In the long term, these structures can be stationary or time dependent, depending on several control and fluid parameters (temperature gradient, material properties, etc.). On the other hand, the spatio-temporal behavior is completely different for thin and ultra-thin films. Here one finds, after a rather short initial phase, the formation of larger and larger structures, known as *coarsening*. Eventually, the dynamics converge to a stationary state that consists of a single elevation (drop) or suppression (hole) on the surface (Fig. 14 left panel). This development can be interrupted by rupture of the film. Rupture is obtained if the surface touches the substrate and the thickness reaches zero in certain domains. Rupture can be avoided by introducing a repelling disjoining pressure acting for a very small depth. In this situation, a completely dry region cannot exist but the substrate is, rather, covered by a so-called (ultra thin) precursor film [38,70], already proposed by Hardy in 1919 [51].

If, in addition, horizontal forces are applied, that is, by inclining the fluid layer, interesting studies of falling

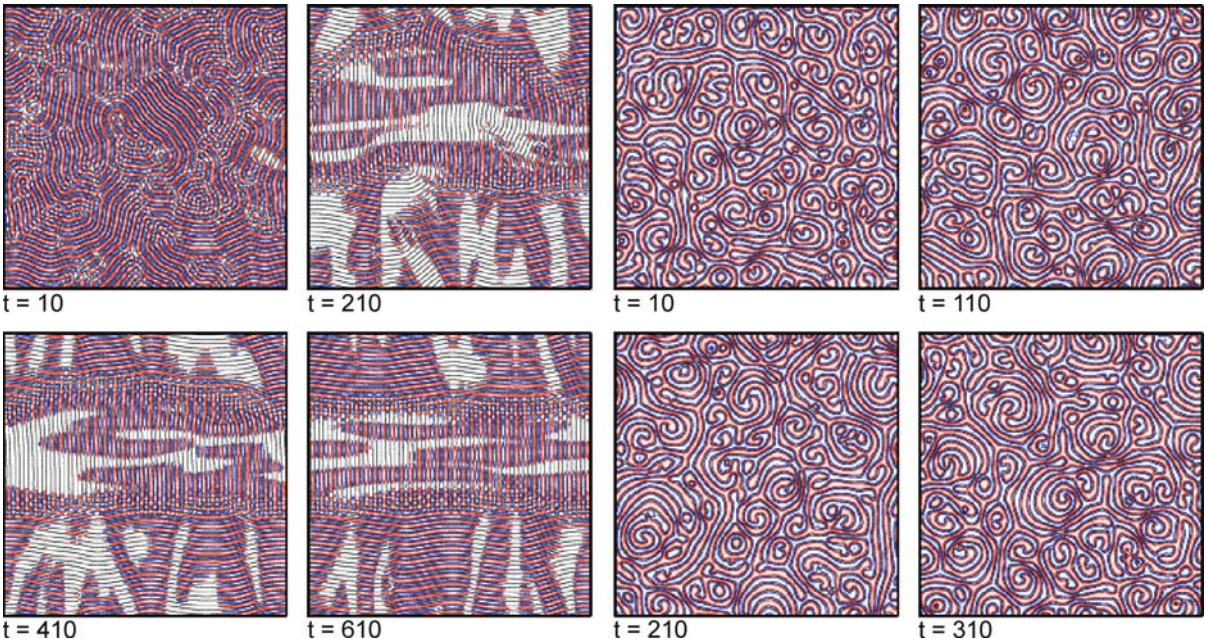

Fluid Dynamics, Pattern Formation, Figure 13
***Left*: Rolls for high Prandtl number (Pr) fluids; *Right*: spirals for low Pr are found if the surface is covered by a good thermal conductor**

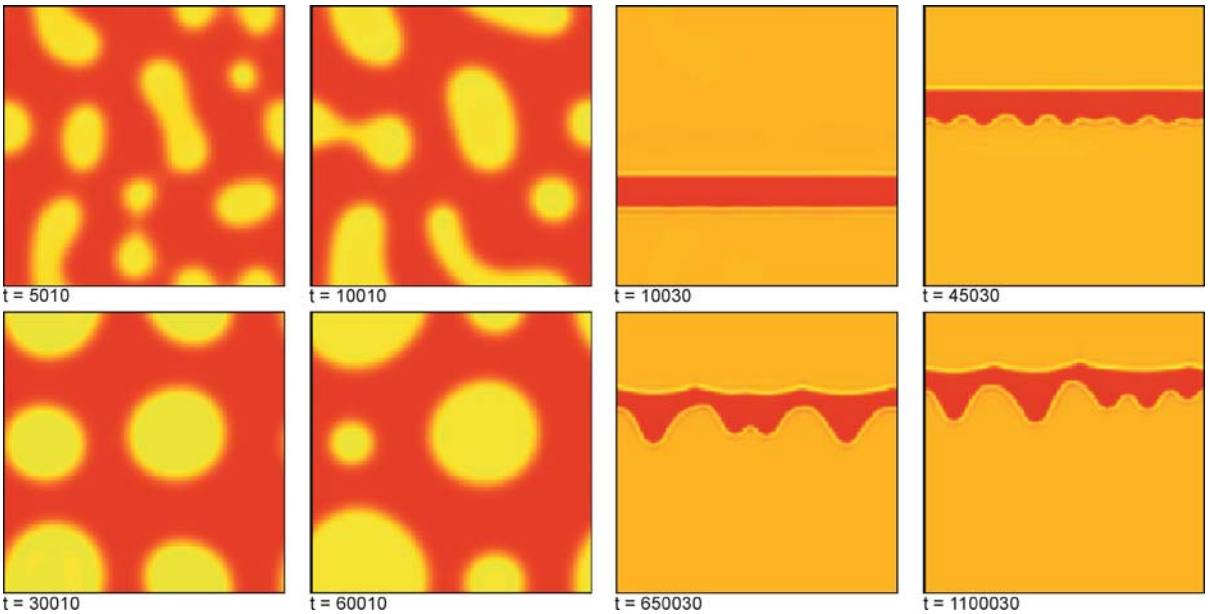

Fluid Dynamics, Pattern Formation, Figure 14
Numerical solution of the thin film equation (see Sect. "Conserved Order Parameter Fields"), *red*: elevation, yellow: suppression. *Left*: Coarsening is the typical spatial behavior for a thin film. Finally, a stationary solution consisting of one single hole would survive. *Right*: If the layer is inclined, the motion of fronts and the development of front instabilities can be examined. From [16]

films and front instabilities can be made in the frame of the thin film equation [13,82]. A typical example is shown in Fig. 14, right panel.

Types of Instabilities

Different types of instabilities can be classified according to their linear behavior at onset. Consider a mode having the complex eigenvalue

$$\lambda(k^2) = i\omega(k^2) + \sigma(k^2) \tag{38}$$

with real valued frequency ω and real valued growth rate σ. Due to rotation symmetry with respect to the horizontal coordinates, all values depend only on the modulus of the wave vector of the unstable mode (assumed as a plane wave in horizontal direction).

According to [36], we use the following notions:

Type $\mathrm{III_s}$ "s" denotes stationary or monotonic and refers to the temporal behavior of the unstable mode close to onset. The type number specifies the spatial behavior of the modes. Type III means slowly varying or even constant in space ($k \approx 0$). The spatial structure beyond instability is then mainly dominated by the geometry and boundary conditions of the system under consideration. For (38) this means

$$\omega = 0 \quad \text{and} \quad \left.\frac{\mathrm{d}\sigma}{\mathrm{d}k}\right|_{k=0} = 0\,,$$

See Fig. 15. A typical example for a type $\mathrm{III_s}$ instability is the real Ginzburg–Landau equation. A computer solution clearly showing the spatially (and temporally) slowly varying behavior can be seen in Fig. 16.

Type $\mathrm{III_o}$ "o" stands for oscillatory and denotes a nonvanishing imaginary part of (38) at threshold. This type includes Hopf-instabilities which have the same slow spatial behavior as $\mathrm{III_s}$. In (38) we have

$$\omega \neq 0 \quad \text{and} \quad \left.\frac{\mathrm{d}\sigma}{\mathrm{d}k}\right|_{k=0} = 0\,.$$

For this kind of instability one needs at least two coupled diffusion equations. It is often encountered in reaction diffusion systems, as for instance the "Brusselator" [76,77].

Type $\mathrm{I_s}$ The short scale pattern forming instabilities shown in Figs. 11–13 with periodicity in space, $k_c \neq 0$ are of this type, see Fig. 15 middle frame. Again one needs at least two coupled diffusion equations to obtain such an instability. For the eigenvalue,

$$\omega = 0 \quad \text{and} \quad \left.\frac{\mathrm{d}\sigma}{\mathrm{d}k}\right|_{k=k_c} = 0 \quad \text{with} \quad k_c \neq 0$$

holds. Sometimes these kinds of patterns are called *Turing structures* or *Turing instabilities*, after the seminal work of Alan Turing, who predicted this patterns in skin, scales, or hair coating of certain animals [89] (Fig. 17). For more details and pattern formation in biology see [63].

Type $\mathrm{I_o}$ Denotes oscillating Turing structures, sometimes also called *wave instabilities*. The eigenvalue λ then has the form

$$\omega \neq 0 \quad \text{and} \quad \left.\frac{\mathrm{d}\sigma}{\mathrm{d}k}\right|_{k=k_c} = 0 \quad \text{with} \quad k_c \neq 0\,.$$

For this instability, the system must be described by at least three coupled diffusion equations. In fluid mechanics, this kind of instability can be encountered in binary mixtures and give rise to a very complicated, in general chaotic, spatio-temporal behavior just at onset [9].

Type $\mathrm{II_s}$ This type is realized in the surface patterns of thin films, Fig. 14. Here, λ depends on k as shown in

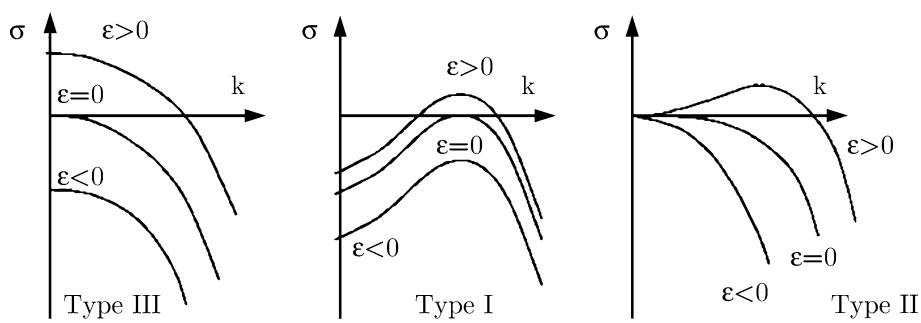

Fluid Dynamics, Pattern Formation, Figure 15
Schematic drawing of the real part of the eigenvalue (38) as function of the wave vector for the three types of instabilities

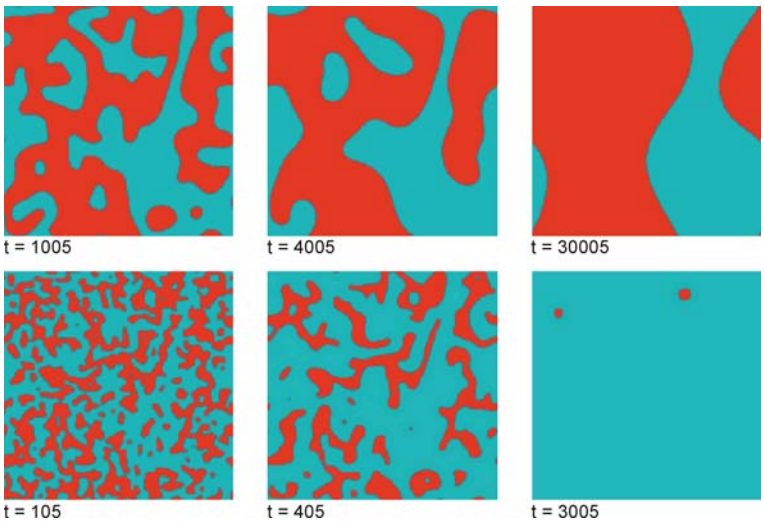

Fluid Dynamics, Pattern Formation, Figure 16
Numerical solution (time series) of the real Ginzburg–Landau equation. (39) which shows a III_s instability. The Ginzburg–Landau equation can be considered as a simple model for the magnetization of a ferro magnet. Then the two rows show the spatio-temporal evolution of the magnetization, *Top*: without external field, *Bottom*: with external field

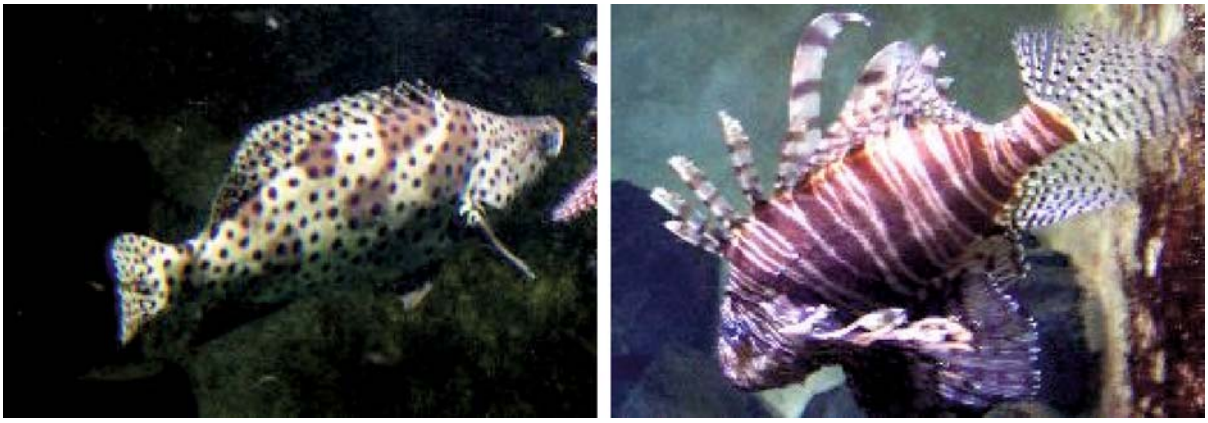

Fluid Dynamics, Pattern Formation, Figure 17
After a theory by A. Turing the painting on skin, scales, or coats of animals is organized by a nonequilibrium chemical reaction during the embryonal phase. *Left*: regular spots arranged in a hexagonal manner on the panther fish, *Right*: stripes with defects on the lion fish (pictures taken by the author in the Berlin Zoo)

the right frame of Fig. 15. One has

$$\omega = 0 \quad \text{and} \quad \left.\frac{\mathrm{d}\sigma}{\mathrm{d}k}\right|_{k=k_c} = 0 \quad \text{with} \quad k_c \neq 0$$

and in addition

$$\sigma(k = 0) = 0 \,.$$

From the last condition one sees that modes with $k = 0$, that is, those which are homogeneous in space, are marginally stable, meaning neither stable nor unstable. One may then add a constant to the order parameter (the mode amplitude)

$$\xi' = \xi + \text{const} \,,$$

where ξ' is still a solution of the linear part of the order parameter equation. This property usually has its origin in a symmetry of the basic problem. We shall discuss this instability type in Sect. "Conserved Order Parameter Fields" on thin films. There, the symmetry corresponds to a global shift of the surface in vertical direction.

Type II$_0$ The same as II$_s$ but with an additional imaginary part $\omega \neq 0$. We mention this type only for the sake of completeness; there will be no further examples in this contribution.

Order Parameter Equations

Order Parameters

In this section, we wish to describe pattern formation in the weakly nonlinear regime. We shall mainly restrict ourselves to the case of monotonic (non-oscillatory) instabilities. For further references see [10,12,36,47,50]. Close to a bifurcation point to a new state, it is natural to expand nonlinearities with respect to small deviations from the old, unstable state. These deviations can be written as a composition of certain Galerkin functions or modes; the amplitudes of these modes are called *order parameters*. If the order parameters are functions only of time, the dynamics given by the order parameter equations (ordinary differential equations, abbreviated: ODE) are perfect patterns, for instance parallel stripes, squares (two order parameters) or hexagons (three order parameters). Natural patterns having defects and grain boundaries, as for instance the structures shown in Figs. 11, 12 and 13 can also be described in this frame. One then has to make the additional assumption that the order parameters also vary (slowly) in space and are ruled by partial differential equations (abbreviated: PDE).

The Ginzburg–Landau Equation

A prominent (and historically the first) example of such a PDE order parameter equation is the Ginzburg–Landau equation [3,58]. In one spatial dimension it has the normal form

$$\partial_t \xi(x,t) = \varepsilon\, \xi(x,t) + q_0^2\, \partial_{xx}^2 \xi(x,t) - c_3 |\xi(x,t)|^2 \xi(x,t) \tag{39}$$

and describes the spatio-temporal evolution of the complex order parameter field ξ. If ξ is the mode amplitude of a roll structure with a certain wave number, for example, the critical one, then stripes with defects are obtained if ξ varies (slowly) in space. If c_3 and q_0 are real valued, (39) is called the real Ginzburg–Landau equation. For complex values of the coefficients, an incomparably richer and much more complicated spatio-temporal behavior of the order parameter is encountered, for details we refer to [3].

In the theory of nonequilibrium pattern formation, writing down an equation such as (39) is far from being purely phenomenological. It can be derived rather systematically from the basic hydrodynamic equations [48,66]. To give an idea of that, we do it briefly for the (two-dimensional) case of convection (the reader who is not interested in technical details can skip the rest of this section).

Starting point are the Eqs. (16) and (20) where in the latter, S stands for temperature T.

Scaling of independent $(\vec{r}, t)$ and dependent $(\vec{v}, T)$ variables allows the reduction of the numbers of parameters:

$$\vec{r} = \vec{\tilde{r}} \cdot d\,, \quad t = \tilde{t} \cdot (d^2/\kappa)\,, \quad \vec{v} = \vec{\tilde{v}} \cdot (\kappa/d)\,, \quad T = \tilde{T} \cdot \beta \cdot \mathrm{d}, \tag{40}$$

with the constant depth d and the externally applied temperature gradient (32). Note that if the liquid is heated from below, $\beta < 0$. Introducing the deviation Θ from the thermally conducting state

$$T(\vec{r}, t) = T^0(z) + \Theta(\vec{r}, t) = T_0 + \beta z + \Theta(\vec{r}, t) \tag{41}$$

transforms (20) into

$$\{\Delta - \partial_t\}\, \Theta(\vec{r}, t) = -\Delta_2 \Psi(\vec{r}, t) + (\vec{v} \cdot \nabla)\Theta(\vec{r}, t) \tag{42}$$

and (16) into

$$\left\{\Delta - \frac{1}{\mathrm{Pr}}\partial_t\right\} \Delta_2 \Phi(\vec{r}, t) = -\frac{1}{\mathrm{Pr}}\left[\mathrm{curl}((\vec{v} \cdot \nabla)\vec{v})\right]_z \tag{43a}$$

$$\left\{\Delta - \frac{1}{\mathrm{Pr}}\partial_t\right\} \Delta\Delta_2 \Psi(\vec{r}, t) = -R\, \Delta_2 \Theta(\vec{r}, t) - \frac{1}{\mathrm{Pr}}\left[\mathrm{curl\,curl}((\vec{v} \cdot \nabla)\vec{v})\right]_z . \tag{43b}$$

Two dimensionless numbers occurred. One is the material dependent *Prandtl number*

$$\mathrm{Pr} = \frac{\nu}{\kappa}\,, \tag{44}$$

which measures the ratio of the diffusion times of heat and momentum. The other one is called the *Rayleigh number* and turns out to be

$$R = -\frac{\beta g \alpha d^4}{\nu \kappa} \tag{45}$$

with α defined in (36). The system (42), (43) constitutes the basic equations for the three scalar fields Φ, Ψ, and Θ which describe convective motion and temperature of a plane fluid layer with a flat and undeformable surface onto a plane substrate. This can be further simplified by taking the large Prandtl number limes $1/\mathrm{Pr} = 0$ (good for fluids with high viscosity, oils, etc.). Then Φ vanishes ev-

erywhere and only two equations are left:

$$\Delta^2\Psi(\vec{r},t) = -R\Theta(\vec{r},t) \tag{46a}$$

$$\{\Delta - \partial_t\}\,\Theta(\vec{r},t) = -\Delta_2\Psi(\vec{r},t) + (\vec{v}\cdot\nabla)\Theta(\vec{r},t)\,. \tag{46b}$$

A general nonlinear (2D) solution of Eqs. (46) may be expressed by

$$\begin{bmatrix} \Psi(x,z,t) \\ \Theta(x,z,t) \end{bmatrix} = \sum_\ell \int_{-\infty}^{\infty} \mathrm{d}k\,\xi_\ell(k,t) \begin{bmatrix} f_\ell(k^2,z) \\ g_\ell(k^2,z) \end{bmatrix} \mathrm{e}^{-ikx} \tag{47}$$

and

$$\xi_\ell(k,t) = \xi_\ell^*(-k,t)$$

where f and g are eigenfunctions of the ODE eigenvalue problem

$$\begin{aligned} (d_z^2 - k^2)^2 f_\ell + R\, g_\ell &= 0 \\ \left(d_z^2 - k^2 - \lambda_\ell(k^2)\right) g_\ell - k^2 f_\ell &= 0\,. \end{aligned} \tag{48}$$

Here, ℓ labels the different eigenfunctions. Equation (48) is obtained by inserting (47) with $\xi_\ell \sim \exp(\lambda t)$ into (46) and keeping only linear terms. The functions f_ℓ and g_ℓ can be calculated numerically by a finite difference method in vertical direction where suitable boundary conditions must be implemented.

Inserting (47) into (46) yields, after multiplication with the adjoint function $g_\ell^+ \exp(ikx)$ and integration over the spatial coordinates, the system:

$$\begin{aligned} \partial_t\xi_\ell(k,t) &= \lambda_\ell(k^2)\xi_\ell(k,t) \\ &- \sum_{\ell'\ell''} \int_{-\infty}^{\infty} \mathrm{d}k'\mathrm{d}k'' c_{\ell\ell'\ell''}(kk'k'')\xi_{\ell'}(k',t)\xi_{\ell''}(k'',t) \\ &\qquad \delta(k-k'-k'')\,, \end{aligned} \tag{49}$$

where the coefficients c are matrix elements that can be computed directly from the basic equations for any given set of control parameters:

$$\begin{aligned} &c_{\ell\ell'\ell''}(kk'k'') \\ &\equiv k'^2 \int_0^1 \mathrm{d}z g_\ell^+(k^2,z) f_{\ell'}(k'^2,z)\partial_z g_{\ell''}(k''^2,z) \\ &\quad - k'k'' \int_0^1 \mathrm{d}z g_\ell^+(k^2,z) g_{\ell'}(k'^2,z)\partial_z f_{\ell''}(k''^2,z)\,. \end{aligned} \tag{50}$$

Here we are still at the same level of complexity; the infinitely many degrees of freedom intrinsic in the basic partial differential equations are expressed by an infinite number of mode amplitudes $\xi_\ell(k,t)$. To eliminate the fast damped modes by the linearly growing ones, we divide the eigenmodes into two groups:

$$\lambda_l \longrightarrow \begin{cases} \lambda_u(k^2) \approx 0 \Longrightarrow \xi_u(k,t), \\ \qquad\qquad |k| \approx k_\mathrm{c}, u = \ell = 1 \\ \lambda_s(k^2) \ll 0 \Longrightarrow \xi_s(k,t), \\ \qquad s = \ell > 1 \text{ or } s = \ell = 1 \text{ but } |k| \neq k_\mathrm{c}\,. \end{cases} \tag{51}$$

In the following we may therefore substitute the index ℓ by u (unstable) or s (stable), depending on the values of ℓ and $|k|$. Now we express the amplitudes of the enslaved modes invoking an adiabatic elimination (k_c denotes the wave vector that maximizes λ_u). In this case, the dynamics of the enslaved modes are neglected, they follow instantaneously to the order parameters. This is a special case of the slaving principle of *synergetics*, which can be used in many other disciplines beyond hydrodynamics [47].

The remaining equations for the order parameters ξ_u, the amplitude equations, read (here and in the following we suppress the index u at ξ and λ):

$$\begin{aligned} \partial_t\xi(k,t) &= \lambda(k^2)\xi(k,t) \\ &+ \int \mathrm{d}k'\mathrm{d}k''\mathrm{d}k''' B(k,k',k'',k''')\xi(k',t)\xi(k'',t) \\ &\qquad \xi(k''',t)\delta(k-k'-k''-k''') \end{aligned} \tag{52}$$

where $|k|, |k'|, |k''|, |k'''| \approx |k_\mathrm{c}|$. Note that there are no quadratic expressions in ξ. This is because $k - k' - k''$ cannot vanish if all wave numbers have the same (nonzero) absolute value. In three spatial dimensions this is different. Three k-vectors can then form a resonant triangle, which is the reason why stable hexagons may occur.

The Landau coefficient B is directly related to the matrix elements (50):

$$\begin{aligned} B(k,k',k'',k''') &= \sum_s \frac{1}{\lambda_s((k''+k''')^2)} \\ &c_{suu}(k''+k''',k',k'')\big[c_{uus}(k,k',k''+k''') \\ &\qquad + c_{usu}(k,k''+k''',k')\big] \end{aligned}$$

where the indices u and s are defined in (51).

To arrive at the Ginzburg–Landau equation, one must transform back to real space. If we express the δ-function in (52) as

$$\delta(k-k'-k''-k''') = \frac{1}{2\pi}\int \mathrm{d}x\, \mathrm{e}^{i(k-k'-k''-k''')x}$$

and assume, that the coefficient B does not depend much on k (it can be evaluated at $k = \pm k_c$), the cubic part of (52) takes the form

$$\frac{\tilde{B}}{2\pi}\int \mathrm{d}x \mathrm{e}^{ikx}\int \mathrm{d}k'\xi(k',t)\,\mathrm{e}^{-ik'x}\int \mathrm{d}k''\xi(k'',t)\,\mathrm{e}^{-ik''x}\int \mathrm{d}k'''\xi(k''',t)\,\mathrm{e}^{-ik'''x} = \frac{\tilde{B}}{2\pi}\int \mathrm{d}x\ \mathrm{e}^{ikx}\Psi^3(x,t) \quad (53)$$

where we have introduced the Fourier transform

$$\Psi(x,t) = \int \mathrm{d}k\ \xi(k,t)\ \mathrm{e}^{-ikx}\,. \quad (54)$$

Inserting (53) into (52), multiplying with $\mathrm{e}^{-ik\tilde{x}}$ and integrating over k yields the order parameter equation in real space

$$\partial_t\Psi(\tilde{x},t) = \int \mathrm{d}k\lambda(k)\xi(k,t)\mathrm{e}^{-ik\tilde{x}} + \tilde{B}\ \Psi^3(\tilde{x},t)\,. \quad (55)$$

If we replace the k^2-dependence of λ under the integral by $-\partial^2_{\tilde{x}\tilde{x}}$, we may pull λ out of the integral and write (55) in the form

$$\partial_t\Psi(x,t) = \lambda(-\partial^2_{xx})\ \Psi(x,t) + \tilde{B}\ \Psi^3(x,t)\,. \quad (56)$$

The function $\Psi(x,t)$ can also be called an "order parameter", though it is not slowly varying in space compared to the small scale structure of the rolls, an idea which we shall work out in the following section. One big advantage can already be seen: the reduction of the number of space dimensions by one. We started with the hydrodynamic equations in two dimensions and get an order parameter equation in only one spatial dimension.

To find the form of the Ginzburg–Landau equation, we must introduce a slowly varying order parameter. This is done by recalling that the Fourier transform of Ψ is mainly excited around $k = \pm k_c$. Then it is natural to make a "rotating wave approximation" with respect to x of the form

$$\Psi(x,t) = \xi(x,t)\ \mathrm{e}^{ik_cx} + \xi^*(x,t)\ \mathrm{e}^{-ik_cx}\,. \quad (57)$$

Inserting this into (56), multiplying by e^{-ik_cx} and integrating with respect to x over one period $2\pi/k_c$ yields with the assumption of constant (slowly varying) ξ in this period

$$\partial_t\xi(x,t) = \lambda(-(\partial_x + ik_c)^2)\ \xi(x,t) + 3\tilde{B}\ |\xi(x,t)|^2\xi(x,t).$$

The last approximation is concerned with the evaluation of the eigenvalue in form of a differential operator. Close to k_c, it has the form of a parabola, see Fig. 15 middle frame. Thus we may approximate

$$\lambda(k^2) = \varepsilon - q^2\left(k^2 - k_c^2\right)^2 \quad (58)$$

and also

$$\begin{aligned}\lambda(-(\partial_x + ik_c)^2) &= \varepsilon - q^2((\partial_x + ik_c)^2 + k_c^2)^2\\ &= \varepsilon - q^2(\partial^2_{xx} + 2ik_c\partial_x)^2\\ &\approx \varepsilon + 4q^2k_c^2\partial^2_{xx}\,.\end{aligned} \quad (59)$$

For the last conversion, we neglect higher derivatives, which is justified due to the slowly varying spatial dependence of ξ. After scaling of ξ and the additional assumption $\tilde{B} < 0$ we finally have derived the Ginzburg–Landau equation (39).

The Swift–Hohenberg Equation

In two spatial dimensions, the drawback of the Ginzburg–Landau equation is its lack of rotational symmetry. Therefore, it is better to pass on the rotating wave approximation (57) and to consider instead the fully space-dependent function Ψ as an order parameter, but now in two spatial dimensions. The resulting evolution equation in its lowest nonlinear approximation is the Swift–Hohenberg equation [88]

$$\dot{\Psi}(\vec{x},t) = \left[\varepsilon - (1+\Delta_2)^2\right]\Psi(\vec{x},t) - \Psi^3(\vec{x},t)\,, \quad (60)$$

which we shall derive now.

Non-local Order Parameter Equations To this end we go back to (52) and write it down in two dimensions, now including the quadratic terms ($\vec{k} = (k_x, k_y)$):

$$\begin{aligned}\partial_t\xi(\vec{k},t) = \lambda(k^2)\xi(\vec{k},t) &+ \int \mathrm{d}^2\vec{k}'\mathrm{d}^2\vec{k}''A(\vec{k},\vec{k}',\vec{k}'')\\ &\cdot \xi(\vec{k}',t)\xi(\vec{k}'',t)\delta(\vec{k}-\vec{k}'-\vec{k}'') + \int \mathrm{d}^2\vec{k}'\mathrm{d}^2\vec{k}''\mathrm{d}^2\vec{k}'''\\ &\cdot B(\vec{k},\vec{k}',\vec{k}'',\vec{k}''')\xi(\vec{k}',t)\xi(\vec{k}'',t)\xi(\vec{k}''',t)\\ &\cdot \delta(\vec{k}-\vec{k}'-\vec{k}''-\vec{k}''')\,.\end{aligned} \quad (61)$$

Introducing the (2D) Fourier transform ($\vec{x} = (x, y)$),

$$\Psi(\vec{x},t) = \int \mathrm{d}^2\vec{k}\ \xi(\vec{k},t)\ \mathrm{e}^{-i\vec{k}\vec{x}}\,. \quad (62)$$

and transforming (61) to real space yields the integro-differential equation

$$\begin{aligned}\partial_t \Psi(\vec{x},t) &= \lambda(\Delta)\Psi(\vec{x},t)\\ &+ \iint \mathrm{d}^2\vec{x}'\,\mathrm{d}^2\vec{x}''\,G^{(2)}(\vec{x}-\vec{x}',\vec{x}-\vec{x}'')\Psi(\vec{x}',t)\Psi(\vec{x}'',t)\\ &+ \iiint \mathrm{d}^2\vec{x}'\mathrm{d}^2\vec{x}''d^2\vec{x}'''\,G^{(3)}(\vec{x}-\vec{x}',\vec{x}-\vec{x}'',\vec{x}-\vec{x}''')\\ &\qquad \Psi(\vec{x}',t)\Psi(\vec{x}'',t)\Psi(\vec{x}''',t)\end{aligned} \tag{63}$$

where the kernels are computed by the Fourier transforms:

$$\begin{aligned}&G^{(2)}(\vec{x},\vec{x}')\\ &\quad= \frac{1}{16\pi^4}\int \mathrm{d}^2\vec{k}\mathrm{d}^2\vec{k}'A(\vec{k}+\vec{k}',\vec{k},\vec{k}')\mathrm{e}^{-i\vec{k}\vec{x}}\,\mathrm{e}^{-i\vec{k}'\vec{x}'},\\ &G^{(3)}(\vec{x},\vec{x}',\vec{x}''')\\ &\quad= \frac{1}{64\pi^6}\int \mathrm{d}^2\vec{k}\mathrm{d}^2\vec{k}'\mathrm{d}^2\vec{k}''B\Big(\vec{k}+\vec{k}'+\vec{k}'',\vec{k},\vec{k}',\vec{k}''\Big)\\ &\qquad\qquad \cdot\mathrm{e}^{-i\vec{k}\vec{x}}\,\mathrm{e}^{-i\vec{k}'\vec{x}'}\,\mathrm{e}^{-i\vec{k}''\vec{x}''}\,.\end{aligned} \tag{64}$$

Gradient Expansion Although Eq. (63) has a rather general form, its further numerical treatment is not practicable, at least not in two dimensions. Each integral must be approximated somehow as a sum over mesh points. The cubic coefficients would result in a 6-fold sum with, if N is the number of mesh points, N^6 summands, which is, if N is around the size of 100, rather hopeless.

On the other hand, the excitation of ξ mainly close to k_c, in two dimensions on a (narrow) ring in Fourier space with radius k_c, makes it natural to expand Ψ under the integrals around $\vec{x}$. This works well if the kernels (64) have a finite (small) range with significant contribution only for $|\vec{x}-\vec{x}'| < \Lambda$ with $\Lambda = 2\pi/k_c$.

To save space we demonstrate the method only for the quadratic term of (63) and in one spatial dimension. A Taylor expansion of Ψ leads to

$$\iint \mathrm{d}x'\,\mathrm{d}x''G^{(2)}(x-x',x-x'')\sum_{m,n=0}^{\infty}\frac{1}{m!n!}\frac{\partial^m\Psi}{\partial x^m}\frac{\partial^n\Psi}{\partial x^n}(x-x')^m(x-x'')^n\,,$$

where the derivatives must be evaluated at x. They can be written in front of the integrals, yielding

$$\sum_{m,n=0}^{\infty} g^{(2)}_{mn}\frac{\partial^m\Psi}{\partial x^m}\frac{\partial^n\Psi}{\partial x^n} \tag{65}$$

with the moments

$$g^{(2)}_{mn} = \frac{1}{m!n!}\iint \mathrm{d}x_1\,\mathrm{d}x_2\,G^{(2)}(x_1,x_2)\,x_1^m x_2^n\,.$$

A similar expression can be found for the cubic coefficient. A series of the form (65) is called *gradient expansion*. In this way, a local order parameter equation results, but which now has infinitely many nonlinear terms. It reads

$$\begin{aligned}\partial_t\Psi = \lambda(\Delta)\Psi &+ \sum_{m,n=0}^{\infty} g^{(2)}_{mn}\frac{\partial^m\Psi}{\partial x^m}\frac{\partial^n\Psi}{\partial x^n}\\ &+ \sum_{\ell,m,n=0}^{\infty} g^{(3)}_{\ell mn}\frac{\partial^\ell\Psi}{\partial x^\ell}\frac{\partial^m\Psi}{\partial x^m}\frac{\partial^n\Psi}{\partial x^n}\end{aligned} \tag{66}$$

with

$$g^{(3)}_{\ell mn} = \frac{1}{\ell!m!n!}\iiint \mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}x_3\,G^{(3)}(x_1,x_2,x_3)x_1^\ell x_2^m x_3^n.$$

For more details see [10].

Swift–Hohenberg–Haken Equation The series in (66) will converge rapidly if the kernels have a short range. Here we consider only the extreme case of δ-shaped kernels, now in two dimensions:

$$\begin{aligned}G^{(2)}(\vec{x}_1,\vec{x}_2) &= A\cdot\delta(\vec{x}_1)\delta(\vec{x}_2)\,,\\ G^{(3)}(\vec{x}_1,\vec{x}_2,\vec{x}_3) &= B\cdot\delta(\vec{x}_1)\delta(\vec{x}_2)\delta(\vec{x}_3)\,.\end{aligned}$$

All coefficients vanish, except $g^{(2)}_{00}$ and $g^{(3)}_{000}$. Then (66) simplifies to

$$\partial_t\Psi(\vec{x},t) = \lambda(\Delta)\Psi(\vec{x},t) + A\,\Psi^2(\vec{x},t) + B\,\Psi^3(\vec{x},t). \tag{67}$$

For the linear part we again use the expansion (58) and replace k^2 by $-\Delta$. After rescaling of length, time and Ψ, (67) turns into the canonical form

$$\dot{\Psi}(\vec{x},t) = \left[\varepsilon-(1+\Delta_2)^2\right]\Psi(\vec{x},t) + a\,\Psi^2(\vec{x},t) - \Psi^3(\vec{x},t) \tag{68}$$

with

$$a = \frac{A}{\sqrt{-B}}\,.$$

Equation (68) is the Swift–Hohenberg–Haken equation derived first using the theoretical methods of synergetics by Haken [11,49].

Numerical solutions of (68) with $a = 0$ are shown in Fig. 18. Stripes as known from convection, but also from Turing instabilities, can be clearly seen. If $|a|$ exceeds a certain value which depends on $\sqrt{\varepsilon}$, hexagonal structures are found which agree qualitatively with those obtained in Bénard–Marangoni convection (Fig. 19). It can be shown that the symmetry break $z \to -z$ caused by the different vertical boundary conditions on top and bottom of the fluid gives rise to a (positive) quadratic coefficient. In the Swift–Hohenberg equation, this violates the symmetry $\Psi \to -\Psi$ and may stabilize two different sorts of hexagons, namely the already mentioned ℓ- and g-hexagons. The first ones are found for large enough positive a, the latter for negative a.

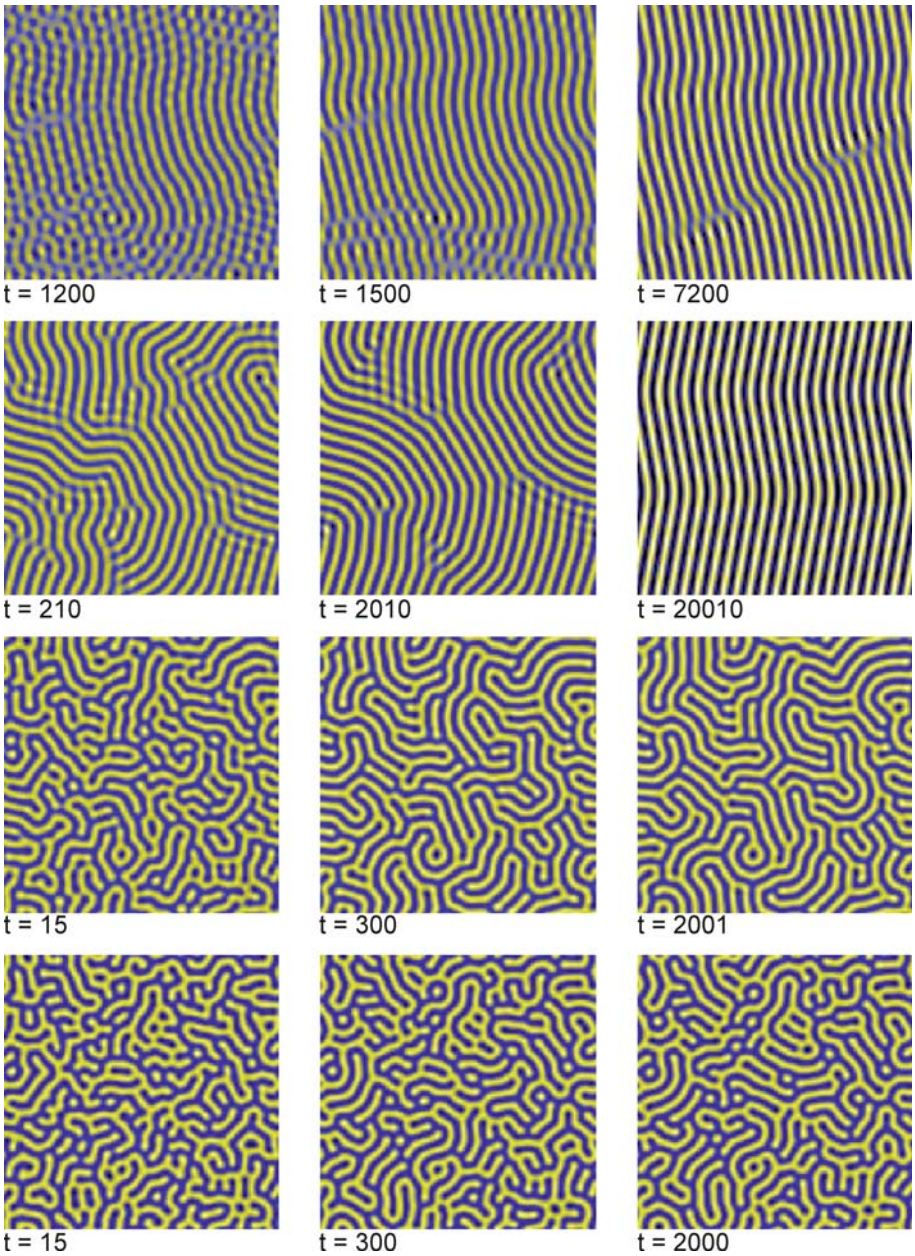

Fluid Dynamics, Pattern Formation, Figure 18
Computer solutions of the Swift–Hohenberg Eq. (60) for several $\varepsilon = 0.01, 0.1, 1.0, 2.0$ (*top to bottom*). The evolution time scales with $1/\varepsilon$, the number of defects increases with ε

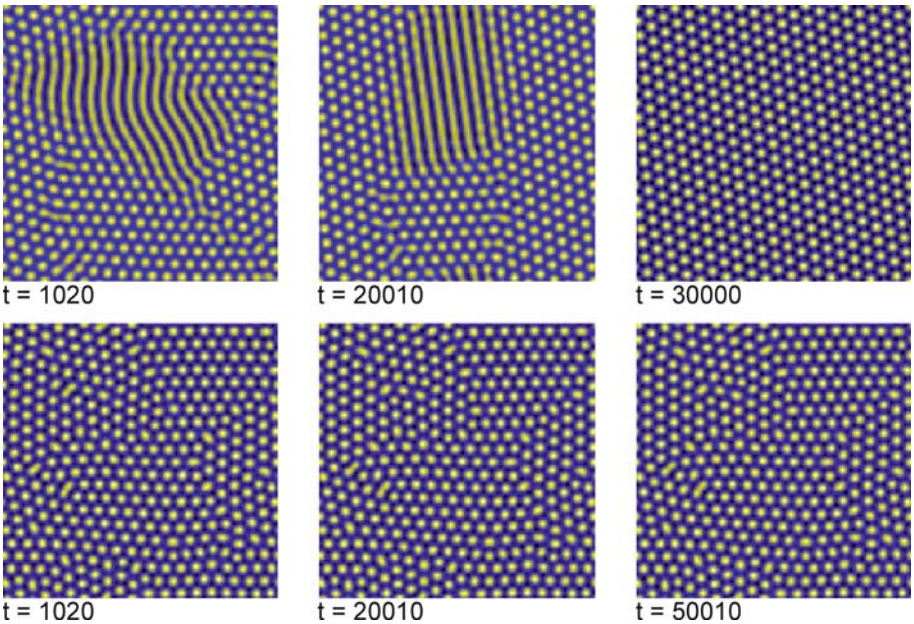

Fluid Dynamics, Pattern Formation, Figure 19
Evolution of a random dot initial condition from (68) with $\varepsilon = 0.1$, $a = 0.26$ (*top*) and $a = 1.3$ (*bottom*). For a in the bistable region, top row, stripes and hexagons coexist for a long time until hexagons win. *Bottom*: for rather large a hexagons are formed soon showing many defects and grain boundaries. The defects survive for quite a long time

The Swift–Hohenberg equation can be considered as normal form of type I_s instabilities. The bifurcation scenario is general (Fig. 20): hexagons are the generic form at onset if symmetry breaking (quadratic) terms occur, which is normal. Even very small symmetry breaking effects lead to hexagons, although their stability region will decrease and finally shrink to the critical point $\varepsilon = 0$ if $a \to 0$. Well above threshold, stripes are expected – or squares.

Squares A linear stability analysis of the Swift–Hohenberg Eq. (68) shows that squares are always unstable in favor of rolls (or hexagons). Therefore there exists no stable square pattern as a solution. This can be changed including higher order terms in the gradient expansion (66), for details see [14]. In this spirit, the equation

$$\partial_t \Psi = \varepsilon\, \Psi - (\Delta + 1)^2 \Psi - b\, \Psi^3 - c\, \Psi\, \Delta^2(\Psi^2) \quad (69)$$

has a stable square solution for $-32c/9 < b < 0$. In Fig. 21 we present numerical solutions of (69) for two different values of the parameter b.

Regular squares are found in convection experiments with two poorly conducting top and bottom plates [28] or in binary mixtures with a certain mean concentration [62]. If the fluid viscosity is strongly temperature dependent (non-boussinesq effects), squares are also preferred, as shown in [27]. For all these cases, an equation of the form (69) can be approximately derived close to onset.

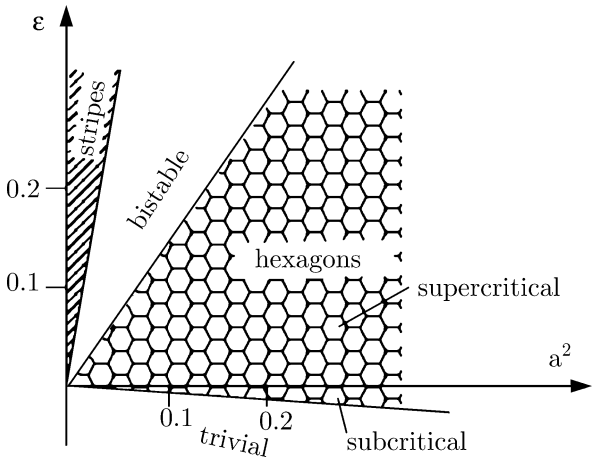

Fluid Dynamics, Pattern Formation, Figure 20
Stability regions in the parameter plane of Eq. (68). Hexagons bifurcate subcritically from the trivial solution $\Psi = 0$. As a secondary instability, stripes emerge. The transition hexagons-stripes as well as trivial sol-hexagons both show hysteresis

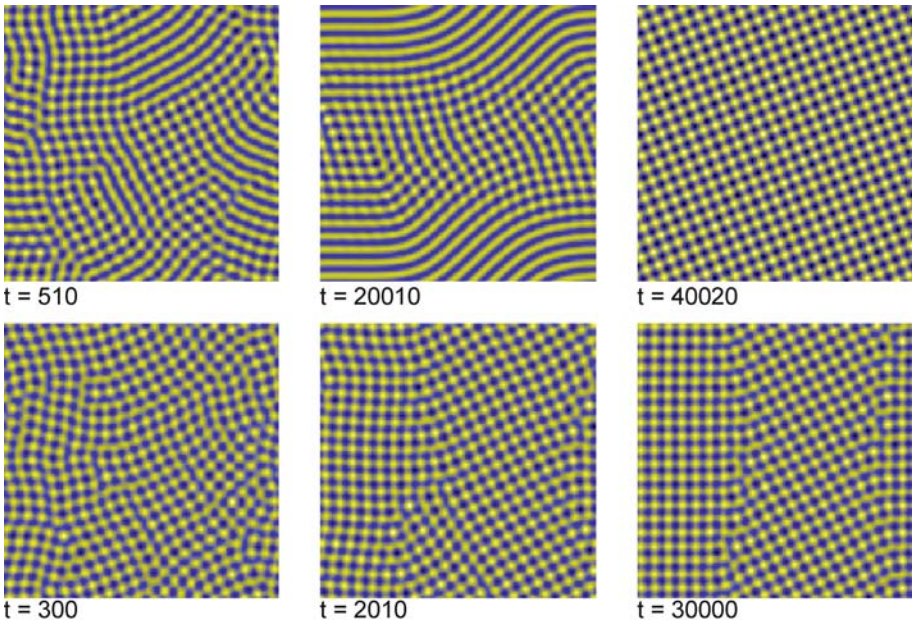

Fluid Dynamics, Pattern Formation, Figure 21
Numerical solutions of (69) for $\varepsilon = 0.1$, $c = 1/16$ and $b = 0$ *(top)*, $b = -0.1$ *(bottom)*. For $b = 0$ both squares and stripes are stable. After a longer time squares win the competition. *Bottom*: clearly in the square region of parameter space. Squares are formed soon having many defects and grain boundaries. Finally, a rather regular square pattern evolves

Conserved Order Parameter Fields

In the previous section, the OPE had the general form

$$\partial_t \xi(\vec{r}, t) = F(\xi, \nabla\xi, \Delta\xi) \tag{70}$$

with no further restrictions (except of boundary conditions) for the order parameter field ξ. However, there are many cases where the physical meaning of the order parameter is that of a density belonging to a conserved quantity such as total mass, volume or charge. Let ξ be such a density; then the mean value

$$M = \langle F \rangle \equiv \frac{1}{V} \int_V \mathrm{d}^3\vec{r}\, F(\xi, \nabla\xi, \Delta\xi) \tag{71}$$

should vanish, if $\langle \xi \rangle$ is a conserved quantity in the constant volume V. Then F can be written as

$$F(\xi, \nabla\xi, \Delta\xi) = -\operatorname{div} \vec{j}(\vec{r}, t) \tag{72}$$

if the total flow of the current density $\vec{j}$ through the surface A of V vanishes

$$\oint_{A(V)} \mathrm{d}^2\vec{f} \cdot \vec{j}(\vec{r}, t) = 0\,. \tag{73}$$

With (72), Eq. (70) takes the form of a continuity equation. In this section we wish to consider OPEs that fulfil (72) and (73).

Thin Films

Consider a fluid with a free and deformable surface located at $z = h(x, y, t)$ as already shown in Fig. 3. If the fluid is incompressible and there is no flow through the sidewalls, the total volume of the fluid layer

$$A \cdot \langle h \rangle = \int_A \mathrm{d}x\, \mathrm{d}y\, h(x, y, t) \tag{74}$$

is a conserved quantity, where A is the base area of the layer. As a consequence, the evolution equation for h must have the form

$$\partial_t h = -\operatorname{div} \vec{j} = -\partial_x j_x - \partial_y j_y\,. \tag{75}$$

Comparing (75) with the kinematic boundary conditions (23) and taking $v_z\big|_{z=h}$ from the integral of the incompressibility condition (3)

$$v_z\big|_{z=h} = -\int_0^h \mathrm{d}z\, (\partial_x v_x + \partial_y v_y) + v_z\big|_{z=0}$$

one finds with $v_z\big|_{z=0} = 0$

$$\vec{j} = \int_0^h \mathrm{d}z\, \vec{v}_H\,, \tag{76}$$

where $\vec{v}_H$ denotes the two horizontal velocity components.

The Lubrication Approximation To close the Eqs. (75), (76), it is necessary to compute $\vec{v}_H$ as a function of h. For thin films, the Reynolds number is small and the Stokes equation (19) determines the fluid velocity to a good approximation. Using scaling [70]

$$x = \tilde{x} \cdot \ell \,, \quad y = \tilde{y} \cdot \ell \,, \quad z = \tilde{z} \cdot \mathrm{d} \,, \quad t = \tilde{t} \cdot \tau \,, \quad h = \tilde{h} \cdot \mathrm{d}, \tag{77}$$

(19) turns into

$$\left(\delta^2\left(\partial^2_{\tilde{x}\tilde{x}} + \partial^2_{\tilde{y}\tilde{y}}\right) + \partial^2_{\tilde{z}\tilde{z}}\right)\vec{\tilde{v}}_H = \tilde{\nabla}_2 \tilde{P} \tag{78a}$$

$$\delta^2\left(\delta^2\left(\partial^2_{\tilde{x}\tilde{x}} + \partial^2_{\tilde{y}\tilde{y}}\right) + \partial^2_{\tilde{z}\tilde{z}}\right)\tilde{v}_z = \partial_{\tilde{z}} \tilde{P} \,. \tag{78b}$$

with the 2D-gradient $\nabla_2 = (\partial_x, \partial_y)$. In (78) we have introduced the dimensionless velocity and pressure

$$\vec{v}_H = \vec{\tilde{v}}_H \cdot \frac{\ell}{\tau} \,, \quad v_z = \tilde{v}_z \cdot \frac{d}{\tau} \,, \quad P = \tilde{P} \cdot \frac{\eta}{\delta^2 \tau}$$

and $\delta = d/\ell$ as a small parameter already defined in (24). In the limit $\delta \to 0$ it follows from (78b)

$$\partial_{\tilde{z}} \tilde{P} = 0 \quad \text{or} \quad \tilde{P} = \tilde{P}(\tilde{x}, \tilde{y}) \,.$$

Thus one can integrate (78a) twice over $\tilde{z}$ and finds with the no-slip condition $\vec{\tilde{v}}_H(0) = 0$

$$\vec{\tilde{v}}_H(\tilde{x}, \tilde{y}, \tilde{z}) = \vec{f}(\tilde{x}, \tilde{y}) \cdot \tilde{z} + \frac{1}{2}(\tilde{\nabla}_2 \tilde{P}(\tilde{x}, \tilde{y})) \cdot \tilde{z}^2 \tag{79}$$

with a function $\vec{f}(\tilde{x}, \tilde{y})$ which can be determined by the boundary conditions. To this end we consider an inhomogeneous surface tension (caused, for example, by a temperature gradient) at the free surface, which yields the condition

$$\eta \, \partial_z \vec{v}_H|_{z=h} = \nabla_2 \Gamma|_{z=h} \,.$$

Inserting (79) there one finds

$$\vec{f} = \tilde{\nabla}_2 \tilde{\Gamma} - (\tilde{\nabla}_2 \tilde{P}) \cdot \tilde{h}$$

with the non-dimensional surface tension

$$\tilde{\Gamma} = \Gamma \, \frac{\tau d}{\eta \ell^2} \,.$$

Inserting everything into (76) and integrating by $\tilde{z}$ finally yields (all tildes omitted)

$$\partial_t h = -\nabla_2 \cdot \left[-\frac{h^3}{3} \nabla_2 P + \frac{h^2}{2} \nabla_2 \Gamma \right] \,. \tag{80}$$

This is the basic equation for the evolution of the surface of a thin film in the so-called lubrication approximation [68]. Equation (80) is sometimes denoted as the *thin film equation* [70,92].

The Disjoining Pressure for Ultra–thin Films Gravitation and surface tension can be included into the pressure P as already outlined in Sect. "Surface Waves". They both stabilize the flat film. On the other hand, an instability mechanism is encountered in very thin (ultra-thin) films where the thickness is some 100 nm or even less [52,79,84]. Then, van der Waals forces between free surface and solid substrate can no longer be neglected [52]. For an attractive force between surface and substrate one has

$$d_h P < 0 \,.$$

But there can also exist a repelling van der Waals force with $d_h P > 0$ which stabilizes the flat surface. Attractive and repelling forces have different ranges. Usually, the repelling force is short range, the attractive one long range. Then, the initially "thick" film can be unstable due to attraction but rupture is avoided by repulsion. In this way completely dry regions cannot exist but the substrate always remains covered by an extremely thin film (some nm), called *precursor film*, Fig. 22 [51].

The complete expression for such an attractive/repulsive disjoining pressure including gravity and surface tension would be (Fig. 23)

$$P(h) = \frac{A_3}{h^3} - \frac{A_9}{h^9} + Gh - q \, \Delta_2 \, h \tag{81}$$

where A_3 and A_9 are material parameters, the Hamaker constants, and

$$G = \frac{d^3 g \tau}{\ell^2 \nu} \,, \quad q = \Gamma \, \frac{\tau d^3}{\ell^4 \eta}$$

denote the dimensionless gravitation number and the surface tension, respectively.

Spinodal Dewetting – Numerical Results If a thin liquid film is exposed to a non- or partially wetting substrate, a small perturbance is sufficient to destabilize the flat surface. The fluid then bubbles and many small drops are formed. This phenomenon can be seen for instance if rain falls on a waxed cloth or on a well polished car roof. Such a process is called *spinodal dewetting* and refers to the unstable region of Fig. 23, [13,86]. As already explained in Sect. "Instabilities", an instability of the flat film occurs in the region where P has a negative slope. This instability is of type II, as is shown in Fig. 23, right frame, and has the growth rate (dispersion relation)

$$\lambda = \frac{1}{3} h_0^3 \left(-D(h_0)k^2 - k^4\right) \tag{82}$$

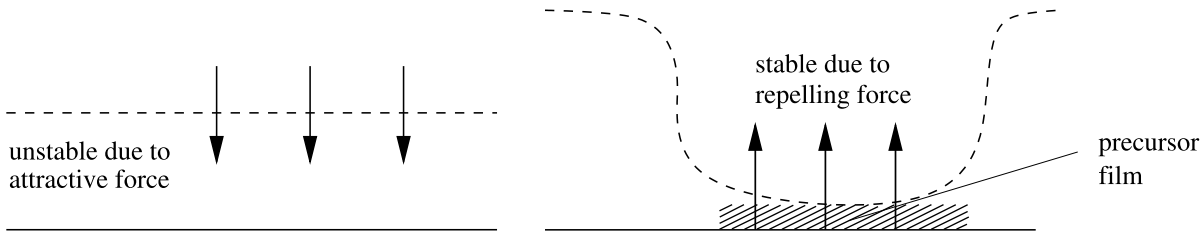

Fluid Dynamics, Pattern Formation, Figure 22
***Left*: Thin flat films are unstable due to an attractive, long range van der Waals force between the free surface and the solid substrate of the fluid. *Right*: If the film is extremely thin (some nm), a repelling short range force acts as a stabilizer and the precursor film remains intact instead of rupturing**

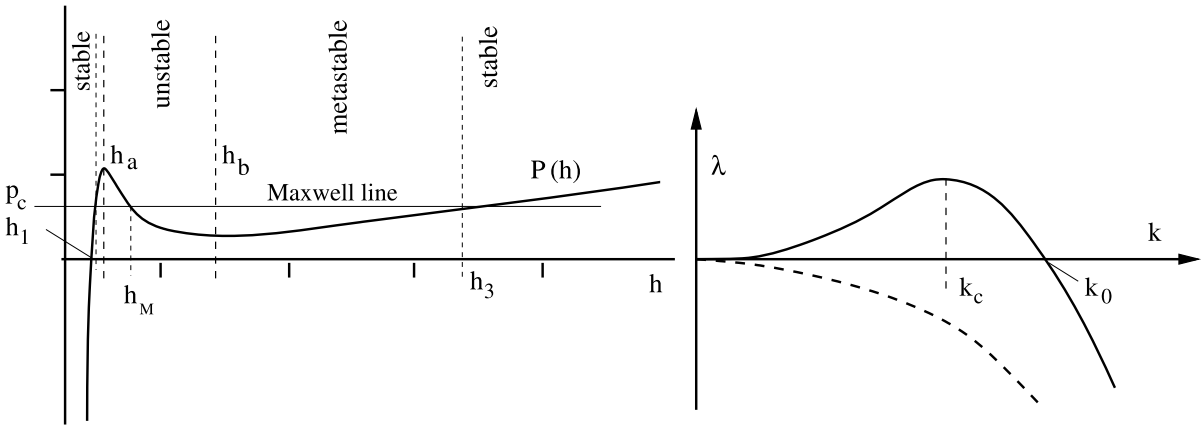

Fluid Dynamics, Pattern Formation, Figure 23
***Left*: The disjoining pressure for a film with uniform thickness h including gravitation, $A_3 = 3, A_9 = 1, G = 0.1$. The region of unstable films is bounded by h_a and h_b. The critical pressure (depth) P_c (h_M) where drops turn into holes is determined by a Maxwell construction. *Right*: Growth rates of periodic disturbances of the plane surface with wave number k. The solid line corresponds to a film with a mean thickness in the unstable regime. Waves having a wave number $0 < k < k_0$ grow exponentially, the mode with $k = k_c$ has the largest growth rate (most dangerous mode). The instability is of type II**

with the "diffusion coefficient"

$$D(h) = d_h P \,.$$

(Here we restrict our further study to fluids with a uniform surface tension. For non-isothermal films with $\nabla_2 \Gamma \neq 0$ we refer to [16,69]). Next we wish to present numerical solutions of the fully nonlinear Eq. (80) with (81). To this end we used the parameters of Fig. 23 and several initial depths h_0. As initial condition a random distribution around the average depth h_0 was chosen.

In the early stage of the evolution (top line of Fig. 24), structures having a length scale of the critical wave length $\Lambda = 2\pi/k_c$, occur, were k_c is the wave number of the fastest growing mode

$$k_c = \sqrt{-\frac{D}{2}} \,.$$

This can be called "linear phase" since the amplitudes are still small and nonlinearities play no important role. The structure grows on the typical time scale

$$\tau = \lambda^{-1}(k_c) = \frac{12}{h_0^3 D^2} = \frac{12}{h_0^3}\left(P'(h_0)\right)^{-2} \,,$$

which is inverse to the square of the slope of the disjoining pressure. This is the reason why pattern formation in thicker films takes much longer (right column in Fig. 24). As a consequence, the small scale (linear phase) structures are overlayed by holes created by certain seeds. After the linear phase, the position of h_0 with respect to the Maxwell point h_M (Fig. 23, left frame) is decisive. If $h_0 > h_M$, holes are formed, for $h_0 < h_M$, one finds drops. If $h_0 \approx h_M$, maze-like patterns are obtained in form of bent, rather irregular stripes (Fig. 24, middle column). In a last, strongly nonlinear phase, coarsening is observed. The final stationary structure (long term) is often a single entity in the form

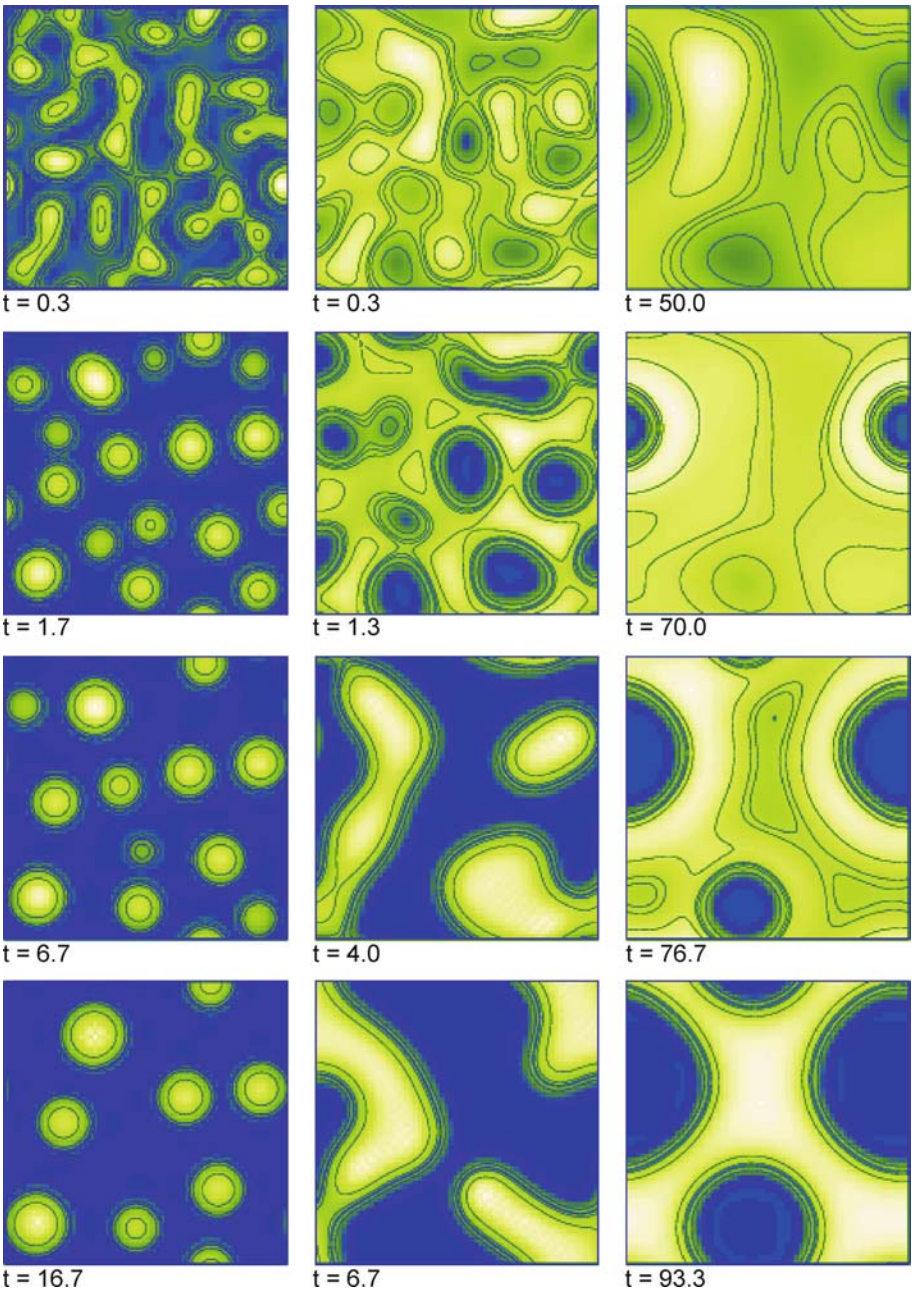

Fluid Dynamics, Pattern Formation, Figure 24
Time series found by numerical integration of (80) for $h_0 = 1.2$ *(left column)*, 1.862 *(middle)*, and 2.8 *(right)*. Light areas correspond to elevated regions of the surface (from [8])

of one big drop or hole. The whole spatio-temporal evolution is transient and can be formulated as a gradient dynamics. The potential plays the role of a generalized free energy reaching its minimum in the steady end state [16].

The flat film is unstable with respect to infinitesimal disturbances if h_0 is in the region between h_a and h_b. On the other hand, two meta-stable domains exist, where the flat film is stable, although the free energy could be lowered by pattern formation. Then, a finite disturbance is necessary, which can be caused by seeds coming, for instance, from impurities. Such a process is called *nucleation* and can be seen in the right column of Fig. 24. There, the seeds

were provided by the random dot initial conditions and two holes are formed. Both processes (nucleation and wetting) converge in this region and it is a question of time scales which one emerges first. In experiments, the formation of holes by nucleation is seen quite often. The reason is that for a Lennard–Jones like disjoining pressure as (81), the meta-stable hole region is much larger compared to that of drops (Fig. 23, left frame).

Phase Field Models

In solidification processes, phase fields are introduced as additional variables to describe the state, here liquid or solid, of the system [57]. Phase fields depend on space and time and governing equations for the phase field variables must be stated or derived. If the phase field obeys an equation of the form of (70), it is called Model A, according to a classification given by Hohenberg and Halperin [46].

Model B Here, we are more interested in phase field equations belonging to Model B. The phase field (we call it Φ) is conserved and a continuity equation

$$\partial_t \Phi = -\operatorname{div} \vec{j} \tag{83}$$

must hold. As in nonequilibrium thermodynamics [30] one assumes that the current density $\vec{j}$ is proportional to a generalized force $\vec{f}$

$$\vec{j} = Q(\Phi) \cdot \vec{f} \tag{84}$$

where Q stands for a non-negative mobility, which is normally a function of the phase field itself, but may also explicitly depend on space coordinates. If the force can be derived from a potential P (pressure)

$$\vec{f} = -\nabla P(\Phi) \tag{85}$$

which in turn can be written as functional derivative of another (thermodynamic) potential (free energy) F

$$P = \frac{\delta F}{\delta \Phi}\,, \tag{86}$$

we finally obtain a closed equation for (83) of the form

$$\partial_t \Phi = \operatorname{div}\left[Q(\Phi) \nabla \frac{\delta F}{\delta \Phi} \right]. \tag{87}$$

With (87) it is easy to show that $\mathrm{d}_t F \leq 0$.

The Cahn–Hilliard Equation As known from the Ginzburg–Landau equation, one may expand the free energy with respect to powers of the phase field. The surface term $(\nabla \Phi)^2$ penalizes phase field variations with respect to space by an increase of F:

$$F[\Phi] = \int_V d^3\vec{r} \Big[\frac{D}{2} (\nabla \Phi)^2 + a_0 \Phi + \frac{a_1}{2} \Phi^2 + \frac{a_2}{3} \Phi^3 + \frac{a_3}{4} \Phi^4 + \ldots \Big]. \tag{88}$$

Substituting this into (87) yields

$$\partial_t \Phi = \operatorname{div}\big[Q(\Phi) \nabla \big(-D \Delta \Phi + a_0 + a_1 \Phi + a_2 \Phi^2 + a_3 \Phi^3 \big) \big]. \tag{89}$$

We further assume $a_2 = 0$ (this can be always obtained by a simple shift of Φ) and $a_1 < 0,\ a_3 > 0$. If we restrict us to the case of a constant mobility, we arrive from (89) after a suitable scaling at the Cahn–Hilliard Eq. [29]

$$\partial_t \Phi = -\Delta \Phi - \Delta^2 \Phi + \Delta(\Phi^3)\,. \tag{90}$$

Equation (90) can be considered as a simple model for a conserved order parameter. A family of stationary solutions of (90) is given by $\Phi = \Phi_0 =$ constant. A linear stability analysis shows that these solutions are type II unstable if $\Phi_0^2 < \frac{1}{3}$ holds. Since (90) belongs to Model B, an infinitesimal disturbance can grow only in a way that keeps the mean value of $\Phi = \Phi_0$ constant. Therefore, spatially structured solutions are expected (Fig. 25).

The density of the free energy of a homogeneous solution reads

$$f(\Phi) = -\frac{\Phi^2}{2} + \frac{\Phi^4}{4} \tag{91}$$

and has its minima at $\Phi_m = \pm 1$. From Fig. 25 it becomes clear that the stationary pattern forming solutions are located between these two minima independently from the mean Φ_0. If the mean value is increased, the regions with $\Phi \approx 1$ grow at the cost of the regions with $\Phi \approx -1$ and vice versa. Taking (90) as a simple model for the phase transition from liquid to gas, the phase field defines the state of aggregation. The density can then be found from the linear relation

$$\rho(\vec{r}, t) = \frac{1}{2}(\rho_f - \rho_g)\, \Phi(\vec{r}, t) + \frac{1}{2}(\rho_f + \rho_g) \tag{92}$$

with ρ_g (ρ_f) as the density of the gaseous (liquid) state. Regions where $\Phi \approx -1$ are gaseous, those with $\Phi \approx +1$ liquid.

Equation (90) has no free parameters. On the other hand, the mean $\Phi_0 = \langle \Phi \rangle$ is a conserved quantity which influences the dynamics of pattern formation qualitatively

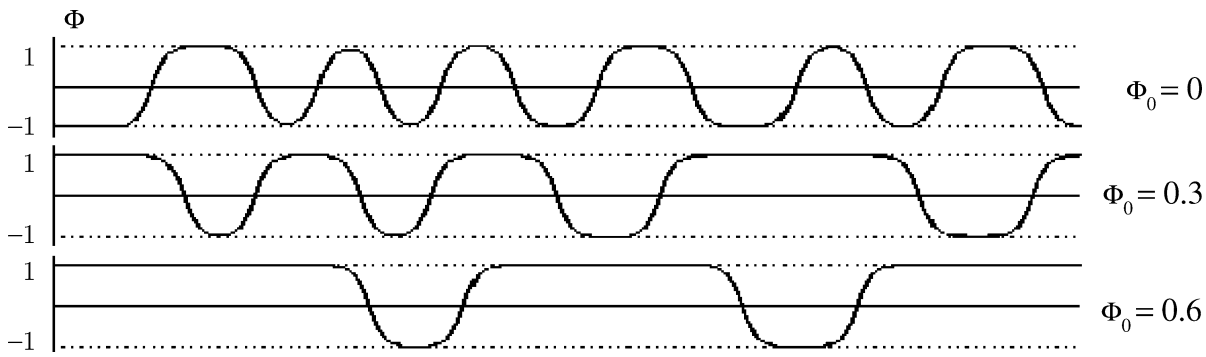

Fluid Dynamics, Pattern Formation, Figure 25
Stationary solutions of the 1D Cahn–Hilliard equation for several mean values Φ_0

and which can be considered as a control parameter. Integrating (92) over the volume, it turns out that Φ_0 is linked to the total mass M of the gas/liquid system

$$M = \frac{1}{2}(\rho_f - \rho_g)\,\Phi_0 \cdot V + \frac{1}{2}(\rho_f + \rho_g) \cdot V\,.$$

The stable homogeneous solutions $\Phi_0^2 > 1/3$ correspond to a pure gas phase ($\Phi_0 < 0$, small total mass), or to a pure liquid phase ($\Phi_0 > 0$, large total mass). In the unstable regime $\Phi_0^2 < 1/3$ the (homogeneous) system has a medium density; this corresponds either to an oversaturated vapor atmosphere ($\Phi_0 < 0$) or to a liquid with a temperature above its boiling point. In both cases, an infinitesimally small disturbance is sufficient to trigger pattern formation in the form of phase separation. In the first case, one observes drops in the gas atmosphere, in the latter, bubbles in the liquid. Figure 26 shows a numerical simulation of (90) in three dimensions.

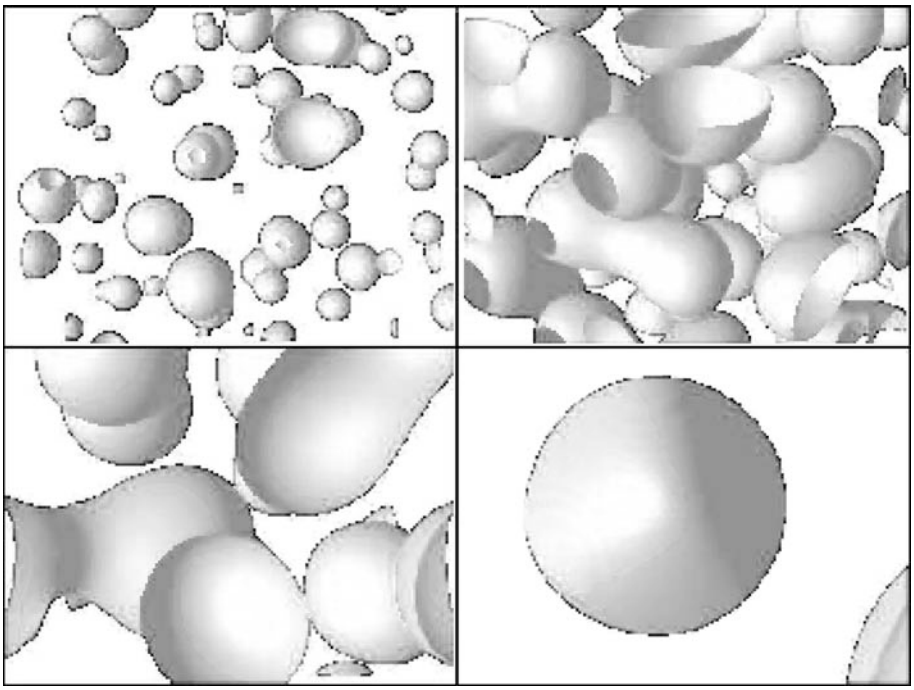

Fluid Dynamics, Pattern Formation, Figure 26
Numerical solution of the Cahn–Hilliard Eq (90) in three room dimensions. The time series *(top left to bottom right)* shows how liquid drops are formed in an oversaturated gas atmosphere. Finally they merge to one big drop by coarsening, a typical dynamic for a type II instability (from [7])

The Fluid Density as Phase Field

Writing down an equation such as (87) and an expansion such as (88) seems to be rather ad hoc. However, for pure fluids it is evident to use the density itself as the phase field, if one is interested in the liquid/gas phase transition. Then, the continuity equation (2) may serve as a phase field equation in lieu of (87). Consequently, the fluid can no longer be considered incompressible.

The Model The Navier–Stokes equations for a compressible fluid (14) must be extended by a force term coming from spatial variations of the phase field (density). They read [21,53]

$$\rho\left[\partial_t\vec{v} + (\vec{v}\cdot\nabla)\,\vec{v}\right] = -\operatorname{grad} p + \vec{f} + \eta\,\Delta\vec{v} + \left(\zeta + \frac{\eta}{3}\right)\operatorname{grad}\operatorname{div}\vec{v} + \mathcal{K}\rho\ \operatorname{grad}\Delta\rho\,. \tag{93}$$

The extra term at the end of (93) was first used by Korteweg in 1901 and is sometimes called Korteweg stress [54]. For (93) we assumed constant material parameters η, ζ, and $\mathcal{K}$. Using the methods of thermodynamics, the pressure is related to the free energy density f [2,37]

$$p(\rho) = \rho\,\frac{\partial f(\rho)}{\partial\rho} - f(\rho) \tag{94}$$

and the free energy as a functional of ρ reads

$$F[\rho] = \int_V \mathrm{d}^3\vec{r}\,\left[\frac{\mathcal{K}}{2}(\nabla\rho)^2 + f(\rho)\right]\,, \tag{95}$$

according to (88). Equations (93) with (94) and (2) form a closed system for the variables $\vec{v}$ and ρ. Wetting properties and contact angles at the walls depend on the boundary conditions $\rho = \rho_w$ along the wall [75]. The choice $\rho_w = \rho_f$ corresponds to a completely wetting (hydrophilic) material, $\rho_w = \rho_g$ to a non-wetting (hydrophobic) boundary. The boundary condition for $\vec{v}$ can either be no-slip (along a wall), no flux, or periodic. It is straightforward to include evaporation and condensation effects into the model, which is studied in [20].

Note that now the free energy (95) is not needed for determining the evolution of the phase field by (*ad-hoc*) gradient dynamics. However, it can be shown that the free energy decreases monotonically in time.

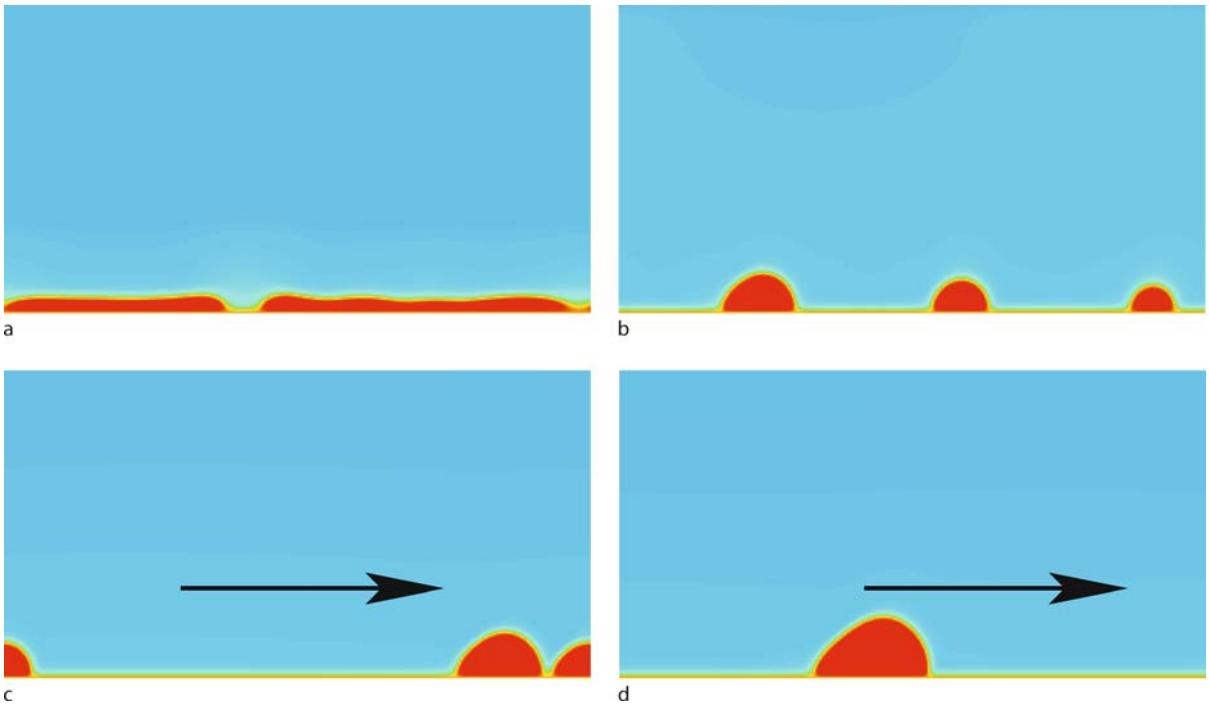

Fluid Dynamics, Pattern Formation, Figure 27
Transition from a flat unstable liquid layer to a drop running down on an inclined substrate (arrows) under gravity effects. Numerical simulation [22] of (93) with (96) and the material parameters for water/vapor from [24] and $\rho_w = 0.5\rho_f$

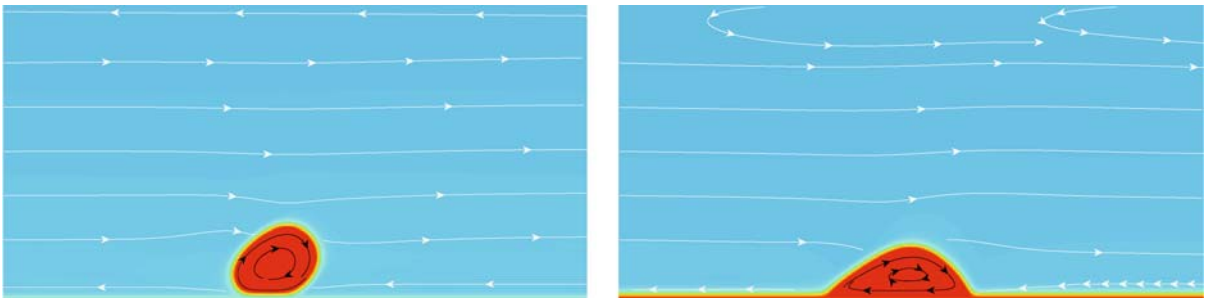

Fluid Dynamics, Pattern Formation, Figure 28
Two final states showing a drop sliding down the inclined substrate with $\rho_w = 0.1\rho_f$ (left, almost hydrophobic) and $\rho_w = 0.8\rho_f$ (*right*, almost hydrophilic). The flow in the gas and in the liquid is indicated by *small arrows* [23]

Results Again, we wish to consider the formation of one state of aggregation on the background of the other. To account for the two stable states, liquid and gaseous, we take (for sake of simplicity we assume $\rho_g \approx 0$)

$$f(\rho) = \gamma\rho^2(\rho - \rho_f)^2 \,. \tag{96}$$

where γ is a positive material constant. In Fig. 27 we show results of the breakup of a flat liquid film aligned along a rigid bottom plate. The layer is inclined by an angle φ and under vertical gravitation. Thus, an external force density of the form

$$\vec{f} = \rho g \begin{pmatrix} \sin\varphi \\ -\cos\varphi \end{pmatrix}$$

occurs in (93). The bottom material is assumed to be partially wetting ($\rho_w = 0.5\rho_f$) and the initial film is unstable under these conditions. Periodic disturbances grow along the fluid's surface. After rupture, bubbles separate and travel from left to right due to downhill force. Figure 28 shows final states of a sliding drop for two boundary values ρ_w. Clearly, the contact angles are different [22].

The phase field description goes far beyond the one based on the thin film equation of Sect. "Thin Films", since there the treatment was restricted to small contact angles and rupture was excluded from the beginning.

Future Directions

There is a huge number of applications in science, industry, and technology where the methods and models outlined in the present article can be used and developed further. In the field of patterns not formed by self-organized processes, but rather by external events such as tidal waves, storm surges or Tsunamis, a reduced and simplified description as discussed in Sects. "Surface Waves", "Order Parameter Equations" should allow for a better understanding of the underlying mechanisms and their effects. Highly involved problems, as for instance the flow, temperature, and concentration fields inside a combustion cell, could be tackled by such models, extended in a suitable way.

Self-organized fluid patterns (Sect. "Instabilities") are the focus of attention in many actual fields of quite different disciplines and scales. The conditions that lead to the creation and stabilization of hurricanes are not yet completely known. The rather high probability of the occurrence of freak waves in the open sea still waits for an explanation. On a planetary scale, convection problems are encountered in the interior of planets and stars and may give rise to the spontaneous formation of a magnetic field. Another problem of great interest for the geophysicist is that of a fluid (such as oil) in a porous medium. The equations for that case differ only a little from that discussed in Sect. "The Basic Equations of Fluid Dynamics" and could therefore be treated in the same spirit.

Understanding the mechanisms responsible for pattern formation can also help to control systems to avoid the occurrence of spatial patterns. In this way, the quality of products obtained from industrial processes, such as coating or solidification (crystal growth), might be improved.

On the micro-scale, fluid problems in general ruled by the Stokes equations discussed in Sect. "Conserved Order Parameter Fields" form a major issue, founding the new discipline of micro-fluidics. But even on the nanoscale, there are new applications in view. The self-organized growth of structures could be a promising tool in the conception and construction of nano-circuits.

An extension of the treatment to complex fluids such as mixtures and emulsions, or to non-Newtonian fluids using the phase field approach (Sect. "Conserved Order Parameter Fields"), is desirable. These fluids are important for biological applications.

Bibliography

Primary Literature

1. Abramowitz M, Stegun IA (1965) Handbook of mathematical functions. Dover, New York
2. Anderson DM, Mc Fadden GB (1997) A diffuse-interface description of internal waves in a near-critical fluid. Phys Fluids 9:1870–1879
3. Aranson IS, Kramer L (2002) The world of the complex Ginzburg–Landau equation. Rev Mod Phys 74:99–143
4. Argyris JH, Faust G, Haase M (1994) An exploration of chaos: An introduction for natural scientists and engineers. North-Holland, Amsterdam
5. Bestehorn M (1993) Phase and amplitude instabilities for Bénard–Marangoni convection in fluid layers with large aspect ratio. Phys Rev 48:3622–3634
6. Bestehorn M (1996) Square patterns in Bénard–Marangoni convection. Phys Rev Lett 76:46–49
7. Bestehorn M (2006) Hydrodynamik und Strukturbildung. Springer, Berlin
8. Bestehorn M (2007) Convection in thick and thin fluid layers with a free interface. Eur Phys J Spec Top 146:391–405
9. Bestehorn M, Colinet P (2000) Bénard–Marangoni–convection of a binary mixture as an example of an oscillatory bifurcation under strong symmetry-breaking effects. Phys D 145:84
10. Bestehorn M, Friedrich R (1999) Rotationally invariant order parameter equations for natural patterns in nonequilibrium systems. Phys Rev E 59:2642–2652
11. Bestehorn M, Haken H (1983) A calculation of transient solutions describing roll and hexagon formation in the convection instability. Phys Lett A 99:265–268
12. Bestehorn M, Haken H (1990) Traveling waves and pulses in a two-dimensional large-aspect-ratio system. Phys Rev A 42:7195–7203
13. Bestehorn M, Neuffer K (2001) Surface patterns of laterally extended thin liquid films in three dimensions. Phys Rev Lett 87:046101
14. Bestehorn M, Pérez-García C (1992) Study of a model of thermal convection in cylindrical containers. Phys D 61:67–76
15. Bestehorn M, Neufeld M, Friedrich R, Haken H (1994) Comment on spiral-pattern formation in Rayleigh–Bénard convection. Phys Rev E 50:625
16. Bestehorn M, Pototsky A, Thiele U (2003) 3D Large scale Marangoni convection in liquid films. Eur Phys J B 33:457
17. Bénard H (1901) Les tourbillons cellulaires dans une nappe liquide. Ann Chim Phys 23:62–143
18. Block MJ (1956) Surface tension as the cause of Bénard cells. Nat Lond 176:650–651
19. Bodenschatz E, Pesch W, Ahlers G (2000) Recent developments in Rayleigh–Bénard convection. Ann Rev Fluid Mech 32: 709–778
20. Borcia R, Bestehorn M (2005) Phase-filed simulations for evaporation with convection in liquid-vapor systems. Eur Phys J B 44:101–108
21. Borcia R, Bestehorn M (2007) Phase-field simulations for drops and bubbles. Phys Rev E 75:056309
22. Borcia R, Borcia I, Bestehorn M (2008) Phys Rev E (submitted)
23. Borcia R, Borcia I, Bestehorn M (2008) Static and dynamic contact angles. Eur Phys J Spec Top (in print)
24. Burelbach JP, Bankoff SG, Davis SH (1988) Nonlinear stability of evaporating/condensing liquid films. J Fluid Mech 195:463
25. Busse FH (1967) The stability of finite amplitude cellular convection. J Fluid Mech 30:625–649
26. Busse FH (1989) In: Peltier WR (ed) Fundamentals of thermal convection. Taylor, pp 23–95
27. Busse FH, Frick H (1985) Square-pattern convection in fluids with strongly temperature-dependent viscosity. J Fluid Mech 150:451–465
28. Busse FH, Riahi N (1980) Nonlinear convection in a layer with nearly insulating boundaries. J Fluid Mech 96:243–256
29. Cahn JW, Hilliard JE (1958) Free energy of a nonuniform system. J Chem Phys 28:258–267
30. Callen HB (1985) Thermodynamics and an introduction to thermostatistics. Wiley, New York
31. Castets V, Dulos E, Boissonade J, De Kepper P (1990) Experimental evidence of a sustained standing Turing-type nonequilibrium chemical pattern. Phys Rev Lett 64:2953–2956
32. Chandrasekhar S (1961) Hydrodynamic and hydromagnetic stability. Dover, New York
33. Cohen IM, Kundu PK (2004) Fluid Mechanics. Academic, Amsterdam
34. Colinet P, Legros JC, Velarde MG (2001) Nonlinear dynamics of surface tension driven instabilities. Wiley, Berlin
35. Cross MC (1988) Theoretical methods in pattern formation in physics, chemistry and biology. In: Garrido L (ed) Far from equilibrium phase transitions. Springer, Berlin
36. Cross MC, Hohenberg PC (1993) Pattern formation outside of equilibrium. Rev Mod Phys 65:851–1112
37. Davis HT, Scriven LE (1982) Stress and structures in fluid interfaces. Adv Chem Phys 49:357
38. de Gennes PG (1985) Wetting: statics and dynamics. Rev Mod Phys 57:827–863
39. Dean RG, Dalrymple RA (2000) Water wave mechanics for engineers and scientists. World Science, Singapore
40. Drazin PG, Johnson RS, Crighton DG (1989) Solitons: An Introduction. Cambridge University Press, Cambridge
41. Eckert K, Bestehorn M, Thess A (1998) Square cells in surface tension driven Bénard convection: experiment and theory. J Fluid Mech 356:155–197
42. Faraday M (1831) On a peculiar class of acoustical figures and on certain forms assumed by groups of particles upon vibrating elastic surfaces. Philos Trans R Soc Lond 121:299
43. Getling AV (1998) Rayleigh–benard convection: structures and dynamics. World Scientific
44. Golovin AA, Nepomnyashchy AA, Pismen LM (1997) Nonlinear evolution and secondary instabilities of Marangoni convection. J Fluid Mech 341:317–341
45. Guckenheimer J, Holmes P (2002) Nonlinear oscillations, dynamical systems, and bifurcations of vector fields. Springer, Berlin
46. Hohenberg PC, Halperin BI (1977) Theory of dynamic critical phenomena. Rev Mod Phys 49:435–47
47. See the up to now 69 volumes of the Springer Series of Synergetics, especially the monographs by Haken H (1983) Synergetics. An Introduction.; Advanced Synergetics.; Information and Self-Organization (2006).
48. Haken H (1975) Cooperative phenomena in systems far from thermal equilibrium and in nonphysical systems. Rev Mod Phys 47:67–121
49. Haken H (1983) Advanced Synergetics: Instability hierarchies of self-organizing systems and devices. Springer, Berlin

50. Haken H (2004) Synergetics introduction and advanced topics. Springer, Berlin
51. Hardy W (1919) The spreading of fluids on glass. Philos Mag 38:49–55
52. Israelachvili JN (1992) Intermolecular and surface forces. Academic Press, London
53. Jasnow D, Vinals J (1996) Coarse-grained description of thermo-capillary flow. Phys Fluids 8:660–669
54. Korteweg DJ (1901) Sur la forme que prennent les équation du mouvements des fluides. Arch Néerl Sci Exact Nat Ser II 6:1
55. Koschmieder EL (1993) Bénard cells and Taylor vortices. Cambridge University Press, Cambridge
56. Lai WM, Rubin D, Krempl E (1993) Introduction to continuum mechanics. Pergamon, Oxford
57. Langer JS (1980) Instabilities and pattern formation in crystal growth. Rev Mod Phys 52:1–28
58. Landau LD, Lifshitz EM (1996) Statistical physics. In: Course of theoretical physics, vol 5. Heinemann, Butterworth
59. Landau LD, Lifshitz EM (2004) Fluid dynamics. In: Course of theoretical physics, vol 6. Heinemann, Butterworth
60. Lorenz EN (1963) Deterministic nonperiodic flow. J Atmos Sci 20:130–141
61. Morris SW, Bodenschatz E, Cannell DS, Ahlers G (1993) Spiral defect chaos in large aspect ratio Rayleigh–Bénard convection. Phys Rev Lett 71:2026–2029
62. Moses E, Steinberg V (1986) Competing patterns in a convective binary mixture. Phys Rev Lett 57:2018–2021
63. Murray JD (1993) Mathematical biology. Springer, Berlin
64. Nekorkin VI, Velarde MG (2002) Synergetic phenomena in active lattices. Patterns, waves, solitons, chaos. Springer, Berlin
65. Nepomnyashchy AA, Velarde MG, Colinet P (2002) Interfacial phenomena and convection. Chapman & Hall, Boca Raton
66. Newell AC, Whitehead JA (1969) Finite bandwidth, finite amplitude convection. J Fluid Mech 38:279–303
67. Nitschke-Eckert K, Thess A (1995) Secondary instability in surface tension driven Bénard convection. Phys Rev E 52: 5772–5775
68. Ockendon H, Ockendon JR (1995) Viscous flow. Cambridge University Press, Cambridge
69. Oron A (2000) Nonlinear dynamics of three-dimensional long-wave Marangoni instability in thin liquid films. Phys Fluids 12:1633–1645
70. Oron A, Davis SH, Bankoff SG (1997) Long-scale evolution of thin liquid films. Rev Mod Phys 69:931–980
71. Ouyang Q, Swinney HL (1991) Transition from a uniform state to hexagonal and striped Turing patterns. Nature 352:610–612
72. Palm E (1960) On the tendency towards hexagonal cells in steady convection. J Fluid Mech 19:183–192
73. Pesch W (1996) Complex spatiotemporal convection patterns. Chaos 6:348–357
74. Penrose R (1974) Role of aesthetics in pure and applied research. Bull Inst Math Appl 10:266
75. Pismen LM, Pomeau Y (2000) Disjoining potential and spreading of thin liquid layers in the Diffuse-interface model coupled to hydrodynamics. Phys Rev E 62:2480–2492
76. Prigogine I, Levever R (1968) Symmetry breaking instabilities in dissipative systems II. J Chem Phys 48:1695–1700
77. Prigogine I, Nicolis G (1967) On symmetry-breaking instabilities in dissipative systems. J Chem Phys 46:3542–3550
78. Rayleigh Lord (1915) On convection currents in a horizontal layer of fluid. Phil Mag 32:462–468
79. Reiter G, Sharma A, Casoli A, David M-O, Khanna R, Auroy P (1999) Thin film instability induced by long-range forces. Langmuir 15:2551–2558
80. Schatz MF, Neitzel GP (2001) Experiments on thermocapillary instabilities. Ann Rev Fluid Mech 33:93–127
81. Schatz MF, Van Hook SJ, Mc Cormick WD, Swift JB, Swinney HL (1999) Time-independent square patterns in surface-tension-driven Bénard convection. Phys Fluids 11:2577–2582
82. Scheid B, Oron A, Colinet P, Thiele U, Legros JC (2002) Nonlinear evolution of nonuniformly heated falling liquid films. Phys Fluids 14:4130–4151
83. Schwabe D (2006) Marangoni instabilities in small circular containers under microgravity. Exp Fluids 40:942–950
84. Sharma A, Khanna R (1998) Pattern formation in unstable thin liquid films. Phys Rev Lett 81:3463–3466
85. Sharma A, Khanna R (1999) Pattern formation in unstable thin liquid films under the influence of antagonistic short and long-range forces. J Chem Phys 110:4929–4936
86. Seemann R, Herminghaus S, Jacobs K (2001) Dewetting patterns and molecular forces: A reconciliation. Phys Rev Lett 86:5534–553
87. Sparrow C (1982) The Lorenz equations. Springer, Berlin
88. Swift JB, Hohenberg PC (1977) Hydrodynamic fluctuations at the convective instability. Phys Rev A 15:319; Chem Phys 48:1695–1700
89. Turing AM (1952) The chemical basis of morphogenesis. Phil Trans R Soc Lond B237:37
90. Van Dyke M (1982) An album of fluid Motion. Parabolic, Stanford
91. Van Oss CJ, Chaudhury MK, Good RJ (1988) Interfacial Lifshitz–van der Waals and polar interactions in macroscopic systems. Chem Rev 88:927–941
92. Vrij A (1966) Possible mechanism for the spontaneous rupture of thin, free liquid films. Disc Faraday Soc 42:23–33

Books and Reviews

Emmerich H (2003) The diffusive interface approach in material science. Springer, Berlin

Fletcher CAJ (1988) Computational techniques for fluid dynamics, vol 1,2. Springer, Berlin

Manneville P (1990) Dissipative structures and weak turbulence. Academic Press, London

Platten JK, Legros JC (1984) Convection in liquids. Springer, Berlin

Pismen LM (2006) Patterns and interfaces in dissipative dynamics. Springer, Berlin

Simanovskii IB, Nepomnyashchy AA (1993) Convective instabilities in systems with interface. Gordon Preach

Schlichting H, Gersten K (2000) Boundary-layer theory. Springer, Berlin

Fluid Dynamics, Turbulence

Rudolf Friedrich[1], Joachim Peinke[2]

[1] Institute for Theoretical Physics, University of Münster, Münster, Germany

[2] Institute of Physics, Carl-von-Ossietzky University Oldenburg, Oldenburg, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
The Basic Hydrodynamic Equations
Vortex Solutions of the Navier–Stokes Equation
Patterns, Chaos, and Turbulence
Turbulence: Determinism and Stochasticity
Reynolds Equation and Turbulence Modeling
The Fine Structure of Turbulence
Phenomenological Theories of Turbulence
Multiscale Analysis of Turbulent Fields
Lagrangian Fluid Dynamics
Future Directions
Acknowledgments
Bibliography

Glossary

Basic equation of fluid dynamics Fluid motion is mathematically treated on the basis of a continuum theory. The fundamental evolution equations are the Euler equation for ideal fluids and the Navier–Stokes equation for Newtonian fluids.

Vortex motions Numerical calculations of turbulent flow fields show that the flows are dominated by coherent structures in form of vortex sheets or tube-like vortices. The question, why vorticity tends to be condensed in localized objects, is one of the central issues of fluid dynamics. Regarding two-dimensional flows there are attempts to approximate fluid motion by a collection of point vortices. This allows one to investigate properties of flows on the basis of a finite dimensional (Hamiltonian) dynamical system.

Turbulence modeling and large eddy simulations The evolution equation for the average velocity field of turbulent flows contains the Reynolds stresses, whose origin are the turbulent pulsations. Turbulence modeling consists of relating the Reynolds stresses to averaged quantities of the fluid motion. This allows one to perform numerical computations of large-scale flows without resolving the turbulent fine structure.

Phenomenological theories of the fine structure of turbulence Phenomenological theories play an important role in physics, and are quite often formulated before a microscopic understanding of the physical problem has been achieved. Phenomenological theories have been developed for the fine structure of turbulence. Of great importance is the theory of Kolmogorov, which he formulated in the year 1941 and refined in 1962. The so-called K41 and K62 theories focus on the self-similar behavior of statistical properties of velocity increments, i. e. the velocity difference between two points with a spatial distance r. Recently, phenomenological theories have been developed that consider the joint probabilities of velocity increments at different scales. It is expected that multiple scale analysis of turbulence will provide new insights into the spatio-temporal complexity of turbulence.

Turbulent cascades Fluid motions are dissipative systems. Stationary flows can only be maintained by a constant energy input in the form of shear flows or body forces. Usually, the length and time scales related to the energy input are widely separated from the ones on which energy is dissipated. A consequence is the establishment of an energy transport across scales. It is believed that this energy transport is local in scale leading to the so-called energy cascades. These cascades are related to the emergence of scaling behavior. There is a direct energy cascade in three dimensions from large to small scales and an inverse cascade of energy from small scales to large scales in two-dimensional flows.

Analytical theories of turbulence Analytical theories of turbulence try to assess the experimental results on turbulent flows directly from a statistical treatment of the basic fluid dynamical equations. Analytical theories rely on renormalized perturbation expansions and use methods from quantum field theory and renormalization group methods. However, no generally accepted theory has emerged so far.

Definition of the Subject

Fluid flows are open systems far from equilibrium. Fluid motion is sustained by energy injected at a certain scale, the so-called integral scale and is dissipated by viscosity mainly in small-scale structures. If the integral scale and the dissipative scale are widely separated and the motions on the integral scale are sufficiently strong, the fluid develops a range of spatio-temporal structures. In three-dimensional flows these structures steadily decay into smaller structures and are generated by the instability of larger structures. This leads to a cascading process which transports energy across scales. Turbulence appears if the fluid motion is driven far away from equilibrium. It develops through sequences of instabilities and processes of selforganization. From this respect, turbulence is a highly ordered phenomenon, whose spatio-temporal complexity, however, has still to be explored.

Introduction

Turbulence is one of the outstanding problems in the field of nonlinear dynamics and complex systems. Although the basic equations of ideal fluid dynamics were formulated by L. Euler 250 years ago and the equations for viscous flows, the so-called Navier–Stokes equation, were established about 150 years ago [13], only a few analytical solutions have been found so far, because of the inherent nonlinear character of fluid flows. Furthermore, fluid motions are systems far from equilibrium. Their maintenance requires a constant input of energy, which is transformed by the viscous flow into heat. A measure for the distance from equilibrium, which corresponds to vanishing fluid velocity, is the Reynolds number,

$$\mathrm{Re} = \frac{UL}{\nu} \tag{1}$$

where U is a characteristic velocity, L a characteristic length scale, and ν is the kinematic viscosity. Flows with Reynolds numbers larger than $\mathrm{Re} = 1000$ usually are turbulent. A turbulent field generated in a free jet experiment is exhibited in Fig. 1. By increasing the Reynolds number one observes the occurrence of various types of instabilities resulting in time-dependent and chaotic patterns making these systems paradigms of self-organization. Whereas the flows generated by the first few instabilities can be treated satisfactorily the transitions and properties of flows at higher Reynolds numbers are by far less understood. This lack of understanding hinders the scientific development in various fields, ranging from astrophysics, engineering to the life sciences.

Basic research on turbulence has always stimulated and contributed to the formulation of new scientific concepts like self-organization and pattern formation, chaos, and the theory of fractals. As a classical nonlinear field theory the description of fluid motion has advanced the mathematical understanding of infinite dimensional nonlinear dynamical systems and the development of efficient computational tools. It is expected that combined experimental and theoretical efforts will lead to a satisfactory understanding of high Reynolds number flows in the near future.

The Basic Hydrodynamic Equations

Fluid motions are described in terms of a continuum theory. The basic ingredients of continuum theories are balance equations for a density $h(\mathbf{x}, t)$ of a physical quantity like mass or momentum defined at location $\mathbf{x}$ and time t:

$$\frac{\partial}{\partial t} h + \nabla \cdot [\mathbf{u}h + \mathbf{j}_h] = q\,. \tag{2}$$

Here, $\mathbf{u}$ denotes the fluid velocity, $\mathbf{j}_h$ the current corresponding to the density h and q denotes a source term [11,47].

The balance equation for the mass density ρ reads

$$\frac{\partial}{\partial t} \rho + \nabla \cdot \rho\mathbf{u} = 0\,. \tag{3}$$

Since mass is conserved, the source term vanishes identically, $q = 0$.

Incompressible fluid motions are characterized by the condition

$$\nabla \cdot \mathbf{u} = 0\,. \tag{4}$$

In the present review we shall mainly focus on incompressible fluids.

The balance equation for the density of momentum, $\rho\mathbf{u}(\mathbf{x}, t)$, takes the form

$$\frac{\partial}{\partial t} \rho u_i + \sum_j \frac{\partial}{\partial x_j} u_j \rho u_i = -\frac{\partial}{\partial x_i} p + \sum_j \frac{\partial}{\partial x_j} \sigma_{ij} + f_i\,, \tag{5}$$

where the momentum current $\mathbf{j}_h$ is expressed by pressure p and the viscous stress tensor σ_{ij}. External forces are summarized in f_i. A complete description requires the formulation of boundary conditions for the velocity field.

It is straightforward to derive the balance equation for the density of the kinetic energy, $\rho\mathbf{u}^2(\mathbf{x}, t)/2$, from the con-

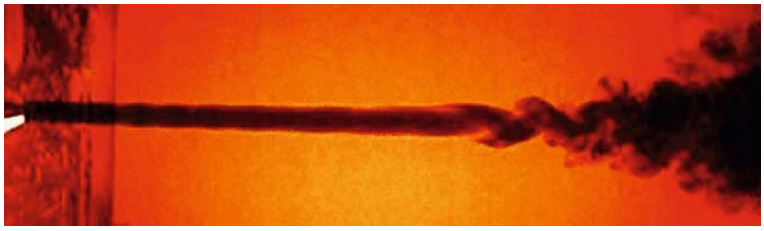

Fluid Dynamics, Turbulence, Figure 1
Development of turbulent structures in a free jet experiment

servation law of momentum (5) for incompressible flows:

$$\frac{\partial}{\partial t}\frac{\rho}{2}\mathbf{u}^2 + \sum_j \frac{\partial}{\partial x_j}\left\{u_j\left[\frac{\rho}{2}\mathbf{u}^2 + p\right] - \sum_i \sigma_{ji}u_i\right\} = -\sum_{ij}\sigma_{ij}\frac{\partial u_i}{\partial x_j} + \sum_j u_j f_j\,. \quad (6)$$

This equation shows that energy is conserved provided the viscous stresses σ_{ij} vanish.

Ideal Fluids: Euler's Equation

For ideal fluid motions, the kinetic energy is conserved provided external forces are absent. The balance Eq. (6) shows that in this case the viscous stresses σ_{ij} have to vanish leading to the Euler equation for incompressible fluid motions:

$$\left[\frac{\partial}{\partial t} + \mathbf{u}\cdot\nabla\right]\mathbf{u} = -\frac{1}{\rho}\nabla p + \frac{1}{\rho}\mathbf{f}\,. \quad (7)$$

The dynamics of ideal fluid motion is restricted by Kelvin's theorem. The circulation $\oint \mathbf{u}(\mathbf{x},t)\,\mathrm{d}\mathbf{r}$ along closed curves going with the flow remains constant [11,47].

Newtonian Fluids: Navier–Stokes Equation

Newtonian fluids are characterized by the presence of viscous stresses. They are assumed to be proportional to the strain matrix S_{ij}

$$\sigma_{ij} = \nu\rho S_{ij} = \nu\rho\frac{1}{2}\left[\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right]\,. \quad (8)$$

Assuming isotropic material properties of the fluid as well as incompressibility one obtains the Navier–Stokes equation:

$$\left[\frac{\partial}{\partial t} + \mathbf{u}\cdot\nabla\right]\mathbf{u} = \nu\Delta\mathbf{u} - \nabla p + \mathbf{f}\,. \quad (9)$$

The kinematic viscosity ν characterizes different fluids. The local energy dissipation rate, denoted by ϵ is obtained from the balance equation of the density of kinetic energy, (6):

$$\epsilon = \frac{\nu}{2}\sum_{ij}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_i}{\partial x_j}\right)^2\,. \quad (10)$$

This quantity plays a crucial role for the understanding of turbulent fluid motions.

At first glance, the Navier–Stokes equation seems to be underdetermined due to the appearance of the gradient pressure term. However, as a result of incompressibility, the pressure is uniquely defined by the Poisson equation in connection with suitable boundary conditions:

$$\Delta p = -\sum_{ij}\frac{\partial u_i}{\partial x_j}\frac{\partial u_j}{\partial x_i}\,. \quad (11)$$

The pressure can be determined with the help of Green's function G of the Laplacian,

$$\Delta G(\mathbf{x}-\mathbf{x}') = -\delta(\mathbf{x}-\mathbf{x}')\,, \quad (12)$$

and yields the pressure as a functional of the velocity field:

$$p = \int \mathrm{d}\mathbf{x}'\, G(\mathbf{x}-\mathbf{x}')\sum_{ij}\frac{\partial u_i}{\partial x_j'}\frac{\partial u_j}{\partial x_i'}\,. \quad (13)$$

This clearly demonstrates that incompressible fluid motions are governed by nonlinear, nonlocal interactions. The gradient pressure term can be regarded as a Lagrange parameter which guaranties the incompressibility of fluid motion.

Vorticity Formulation of Incompressible Fluid Dynamics It is possible to formulate the basic fluid dynamic equations using the vorticity $\boldsymbol{\omega}(\mathbf{x},t) = \nabla\times\mathbf{u}(\mathbf{x},t)$. Provided the vorticity is known, one can obtain the velocity field by the analogue of Maxwell's equation of magnetostatics:

$$\nabla\times\mathbf{u} = \boldsymbol{\omega}\,, \quad \nabla\cdot\mathbf{u} = 0\,. \quad (14)$$

The velocity field is determined by the analogy to Biot–Savart's law

$$\mathbf{u}(\mathbf{x},t) = \int \mathrm{d}\mathbf{x}'\,\boldsymbol{\omega}(\mathbf{x}',t)\times\mathbf{K}(\mathbf{x}-\mathbf{x}') + \nabla\Phi\,, \quad (15)$$

where $\mathbf{K}$ is related to Green's function $G(\mathbf{x})$ of the Laplacian $\mathbf{K}(\mathbf{x}) = \nabla G(\mathbf{x})$. The potential Φ has to fulfill $\Delta\Phi = 0$.

It is straightforward to derive an evolution equation for the vorticity:

$$\frac{\partial}{\partial t}\boldsymbol{\omega} + \mathbf{u}\cdot\nabla\boldsymbol{\omega} = \boldsymbol{\omega}\cdot\nabla\mathbf{u} + \nu\Delta\boldsymbol{\omega} + \mathbf{f}_\omega\,. \quad (16)$$

Here, an important difference between two- and three-dimensional fluid motion becomes evident. For two-dimensional flows the vorticity only has a component perpendicular to the motion and the so-called vortex stretching term $\boldsymbol{\omega}\cdot\nabla\mathbf{u}$ vanishes identically.

Lagrangian Formulation of Incompressible Fluid Dynamics Up to now, we have treated fluid dynamics from the Eulerian point of view by considering fields defined at a fixed spatial location. There is an alternative approach to the description of fluid motion. This so-called Lagrangian treatment is based on the introduction of the Lagrangian velocity $\mathbf{U}(\mathbf{y}, t)$ and the Lagrangian path $\mathbf{X}(\mathbf{y}, t)$ of a point moving with the flow starting at time $t = 0$ at the location $\mathbf{y}$. For obvious reasons the quantity $\mathbf{X}(\mathbf{y}, t)$ is also denoted as *Lagrangian map*. The inverse map is denoted as $\mathbf{y}(\mathbf{x}, t)$. The basic fluid dynamical equations can also be formulated in this Lagrangian picture.

As an example we formulate the evolution equation for the Lagrangian vorticity for two-dimensional incompressible flows. To this end we introduce the Lagrangian vorticity $\Omega(\mathbf{y}, t) = \boldsymbol{\omega}(\mathbf{X}(\mathbf{y}, t), t)$. The first equation defines the Lagrangian path, the second the evolution of the Lagrangian vorticity.

$$\frac{\partial}{\partial t}\mathbf{X}(\mathbf{y}, t) = \int \mathrm{d}\mathbf{y}'\, \Omega(\mathbf{y}', t)\{\mathbf{e}_z \times \mathbf{K}[\mathbf{X}(\mathbf{y}, t) - \mathbf{X}(\mathbf{y}', t)]\}\,,$$
$$\frac{\partial}{\partial t}\Omega(\mathbf{y}, t) = \nu\left[\sum_{ijkl}\frac{\partial Y_l}{\partial x_i}\frac{\partial Y_k}{\partial x_i}\frac{\partial^2}{\partial y_l \partial y_k} + \sum_{li}\frac{\partial^2 Y_l}{\partial x_i \partial x_i}\frac{\partial}{\partial y_l}\right]\Omega(\mathbf{y}, t)\,. \tag{17}$$

For two-dimensional flows the gradient of Green's function of the Laplacian takes the form $\mathbf{K}(\mathbf{x}) = \frac{\mathbf{x}}{2\pi|\mathbf{x}|^2}$. It is immediately obvious that in the ideal fluid case, the two-dimensional vorticity is conserved along Lagrangian trajectories. A similar formulation exists for three-dimensional flows, where, however, the evolution equation for vorticity contains the vortex stretching term.

Recently, the Lagrangian formulation of fluid dynamics has become important due to the possibility to measure the path of passive tracer particles [46,63,69]. Its importance for the description of turbulence has already been emphasized by Taylor [82] and Richardson [77].

Existence and Smoothness Results

Although the Euler and the Navier–Stokes equations are of fundamental interest for various fields ranging from astrophysics to applied mechanics and engineering, their mathematical properties still remain puzzling. Especially for three-dimensional flows results on the existence (or nonexistence) and smoothness of solutions could not yet be obtained. This topic is one of the millennium problems formulated by the Clay Mathematics Institute. As an introduction to the subject we refer the reader to the webpage of the Clay institute with an outline of the mathematical problem due to Fefferman [90], as well as the two monographs [16,20].

Vortex Solutions of the Navier–Stokes Equation

Figure 2 shows a volume rendering of the absolute value of vorticity above a certain threshold obtained from a direct numerical solution of the vorticity equation. The field is characterized by the presence of elongated vortex structures [39]. Whereas fully developed turbulent flows tend to be dominated by vortex-like objects it seems that modest turbulent flows are characterized by the presence of sheet-like structures. There are several exact solutions of the Navier–Stokes equation, which seem to be related with the vortex structures observed in fully developed turbulence. They can be investigated using symmetry arguments and methods from group theory [31].

Axisymmetric Vortices: Lamb–Oseen Vortex

An axisymmetric vorticity distribution $\boldsymbol{\omega} = \Omega(r)\mathbf{e}_z$ generates a purely azimuthal velocity field. This fact has the consequence that in the vorticity equation both, the convective term $\mathbf{u} \cdot \nabla\boldsymbol{\omega}$ and the vortex stretching term $\boldsymbol{\omega} \cdot \nabla\mathbf{u}$, identically vanish, leading to an equation which is the radially symmetric heat equation

$$\frac{\partial}{\partial t}\Omega = \nu\frac{1}{r}\frac{\partial}{\partial r}r\frac{\partial}{\partial r}\Omega\,. \tag{18}$$

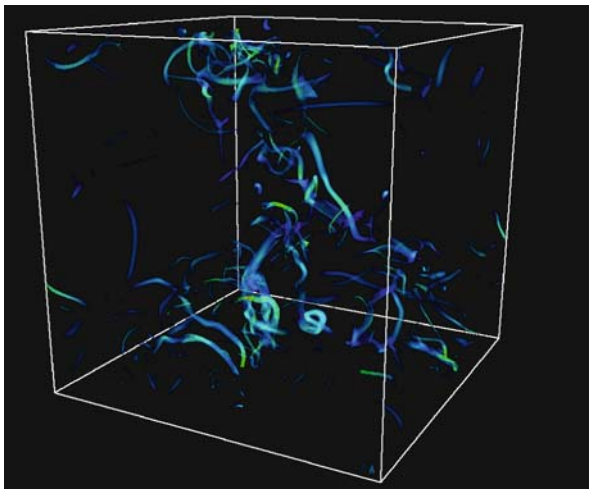

Fluid Dynamics, Turbulence, Figure 2
Direct numerical calculation of the vorticity field of a turbulent fluid motion: Absolute value of vorticity above a given threshold. After [88]

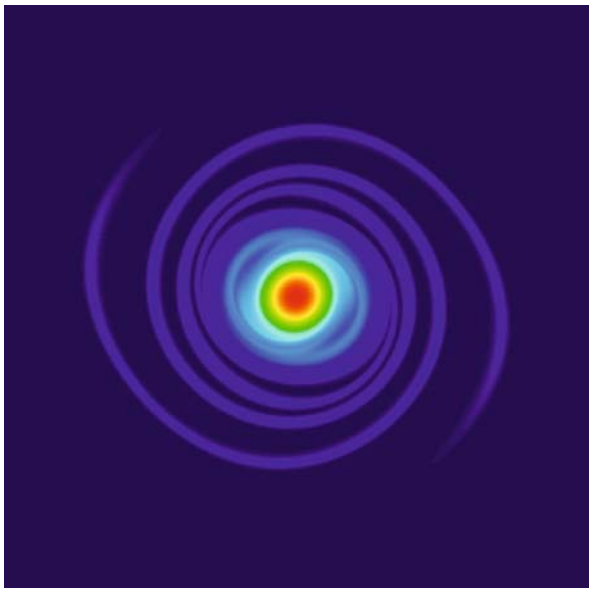

Fluid Dynamics, Turbulence, Figure 3
Vorticity field of a two-armed Lundgren spiral

A solution is the so-called Lamb–Oseen vortex,

$$\Omega = \frac{\Gamma}{\pi r_B^2} e^{-r^2/r_B^2}, \quad r_B^2 = 4\nu t. \tag{19}$$

The corresponding velocity field is azimuthal and has the form

$$v_\varphi(r) = \frac{\Gamma}{2\pi r}\left[1 - e^{-r^2/r_B^2}\right]. \tag{20}$$

It decays like $\Gamma/2\pi r$ for large distances. Because of viscosity the velocity field at the origin vanishes identically.

The Lundgren Spiral

Another vortex solution in the form of a spiral vorticity distribution has been considered by Lundgren [53]. It is generated by a localized central, strong vorticity distribution whose velocity field drags surrounding, weaker vorticity distributions into spiral arms. In this way it is also possible to generate multiple-armed spirals. This process is depicted in Fig. 3. Recently the Lundgren spiral has been generated experimentally by the group of Petitjeans [12].

Stretched Vortices

A highly interesting class of solutions has been found by Lundgren. He considered the velocity field

$$\mathbf{u}(\mathbf{x}, t) = \frac{a(t)}{2}[-x, -y, 2z] + [w_1(x, y, t), w_2(x, y, t), 0] \tag{21}$$

and showed that the two-dimensional velocity field $\mathbf{w}(x, y, t)$ can be obtained from the velocity field $\mathbf{W}(\xi_1, \xi_2, \tau)$ by the *Lundgren transformation*

$$\mathbf{w}(x, y, t) = A(t)\mathbf{W}[A(t)x, A(t)y, \tau(t)]$$
$$A(t) = e^{\int_0^t dt' a(t')}, \quad \tau(t) = \int_0^t dt' A(t')^2. \tag{22}$$

Thereby, the field $\mathbf{W}(\xi_1, \xi_2, \tau)$ obeys the two-dimensional Navier–Stokes equation.

In the case of a time constant a, the decaying *Lamb–Oseen vortex*, Eq. (19) is changed into the *Burgers vortex*, where r_B becomes constant, $r_B^2 = \frac{4\nu}{a}$.

The corresponding two-dimensional velocity field is an azimuthal field, which, for large values of r decays like $\frac{1}{r}$. In the limit $\nu \to 0$ the vorticity field approaches a delta-distribution.

Vorticity Alignment

Figure 2 exhibits the vorticity field obtained from a numerical simulation of the Navier–Stokes equation. Exhibited is a volume rendering of the absolute value of vorticity above a given value and it is quite evident that Burgers-like vortices play a major role in the spatio-temporal organization of turbulence. Consequently, the Burgers vortices have been denoted as the *sinews of turbulence* [60]. Although the emergence of vortex-like objects as organization centers of turbulence has not yet been fully clarified, it is clear that it is related to the phenomenon of vorticity alignment [28]. It has been emphasized that locally the vorticity vector is predominantly aligned to the eigenvector of the intermediate eigenvalue λ_2 of the strain matrix S, Eq. (8). This matrix is symmetric and has three real eigenvalues $\lambda_1 \le \lambda_2 \le \lambda_3$. Vorticity alignment is still investigated intensively [36].

Vorticity alignment has also played a major role in the discussion of the possibility of finite time singularities in Euler flows [30,32], a question which is fundamentally related to the question of the existence of solutions of the Euler and Navier–Stokes equations [16,20].

Modeling Turbulent Fields by Random Vortex Distributions

There have been several attempts to model the fine structure of turbulent fields by statistically distributed vortex solutions of the Navier–Stokes equations. Townsend [84] used a random arrangement of Burgers vortices. As already mentioned Lundgren [53,54] considered the above-mentioned spiral structures in a strain field. He showed

that suitable space-time average over a decaying Lundgren spiral leads to an energy spectrum predicted by Kolmogorov's phenomenological theory of turbulence. A cascade interpretation of Lundgren's model has been given by Gilbert [29]. Kambe [37] discussed randomly arranged Burgers vortices in a strain field and was able to model intermittency effects of Eulerian velocity increments.

Patterns, Chaos, and Turbulence

Pattern Formation and Routes to Chaos in Fluid Dynamics

Experiments on fluid dynamics in confined geometries like the Rayleigh–Bénard system or the Taylor–Couette experiment exhibit a variety of instabilities leading from stationary patterns to time periodic structures and to chaotic motions. For an overview with further references on fluid instabilities we refer the reader to the monograph of Manneville [55]. These flows are characterized by temporal complexity, however the flow structures remain spatially coherent. The scale of energy injection and the scale of energy dissipation usually are not widely separated in these systems. This has the consequence that only few degrees of freedom are excited. Such types of flows can be successfully treated on the basis of the slaving principle of synergetics [34,35]. In mathematical terms, these types of flows are related to the existence of center or inertial manifolds [83] in phase space. This allows explanation of the various *routes to chaos* or *routes to turbulence* observed in fluid motions, especially in confined geometries, on the basis of the theory of low-dimensional dynamical systems.

Point Vortex Motion

The *Lagrangian map* $\mathbf{X}(\mathbf{y}, t)$ of a two-dimensional ideal fluid motion is determined by the solution of the integro-differential Eq. (17), where the Lagrangian vorticity $\Omega(\mathbf{y}, t)$ is temporally constant. This integrodifferential equation can be reduced to a finite set of ordinary differential equations by considering fields with strongly localized vorticity

$$\Omega(\mathbf{y}, t) = \sum_j \Gamma_j \delta(\mathbf{y} - \mathbf{y}_j) \,. \tag{23}$$

Introducing the notation $\mathbf{X}(\mathbf{y}_j, t) = \mathbf{x}_j(t)$ we obtain the set of differential equations for the positions $\mathbf{x}_j(t)$ of the point vortices:

$$\dot{\mathbf{x}}_i = \sum_{j \neq i} \frac{\Gamma_j}{2\pi} \mathbf{e}_z \times \frac{\mathbf{x}_i - \mathbf{x}_j}{|\mathbf{x}_i - \mathbf{x}_j|^2} \,. \tag{24}$$

This set of equations for the vortex positions was already known by Kirchhoff [41]. Since then, there have been many studies of this problem [3,4,56,64]. We mention that the N-vortex problem of two-dimensional fluid motion shares many properties with the N-body problem of classical mechanics.

The Lagrangian motion of an arbitrary point $\mathbf{X}(\mathbf{y}, t)$ can be determined by the solution of the nonautonomous differential equation

$$\dot{\mathbf{X}}(\mathbf{y}, t) = \sum_{i=1}^{N} \frac{\Gamma_i}{2\pi} \mathbf{e}_z \times \frac{\mathbf{X}(\mathbf{y}, t) - \mathbf{x}_i(t)}{|\mathbf{X}(\mathbf{y}, t) - \mathbf{x}_i(t)|^2} \,, \tag{25}$$

where the point vortex positions $\mathbf{x}_i(t)$ are given by the evolution equations (24). The point vortex approximation, thus, allows one to study *mixing in two- dimensional flows* on the basis of sets of ordinary differential equations [3,70,81].

Since the point vortex system is obtained from the two-dimensional Euler equation it is evident that the kinetic energy is conserved. In fact, introducing the Hamilton function

$$\begin{aligned} H &= -\frac{1}{4\pi} \sum_{i \neq j} \Gamma_i \ln |\mathbf{x}_i - \mathbf{x}_j| \Gamma_j \\ &= \frac{1}{2} \sum_{i \neq j} \Gamma_i G(\mathbf{x}_i, \mathbf{x}_j) \Gamma_j \end{aligned} \tag{26}$$

we can rewrite the evolution equations in Hamiltonian form

$$\Gamma_i \dot{\mathbf{x}}_i = \mathbf{e}_z \times \nabla H \,. \tag{27}$$

There are further conserved quantities due to symmetry. The *center of vorticity*,

$$\mathbf{R} = \sum_i \Gamma_i \mathbf{x}_i \,, \tag{28}$$

as well as the quantity

$$I = \sum_i \Gamma_i \mathbf{x}_i^2 \tag{29}$$

are time constants.

The advantage of considering the motion of point vortices in two- dimensional ideal flows is the applicability of methods and notions from the theory of finite Hamiltonian systems. Explicitly, one can investigate the integrability of point vortex motions, i. e. the formation of time periodic as well as quasiperiodic vortex motions and the emergence of chaotic motions. Because of the existence of

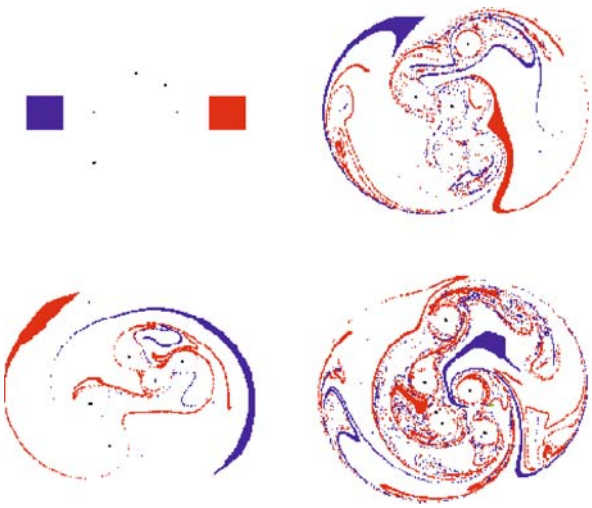

Fluid Dynamics, Turbulence, Figure 4
Mixing induced by the motion of four point vortices

the integrals of motions (26), (28), (29) the motion of two vortices as well as three point vortices is integrable leading to quasiperiodic motions in time. Furthermore, it has been shown that in the four vortex problem, in addition to quasiperiodic motions, chaotic motion is possible. We refer the reader to the review articles of Aref [3,4] and the monograph of Newton [64]. Also the investigations of mixing by point vortex motion has revealed interesting insights into the geometry and the complexity of the mixing process. An example of the missing process by four point vortices is exhibited in Fig. 4. Similar concepts are expected to be applicable for general two- and three- dimensional turbulent flows.

Onsager's Statistical Theory of Two-Dimensional Turbulence

Onsager [68] recognized the importance of two-dimensional vortex models for turbulent flows. Since the dynamics is Hamiltonian, one can perform a statistical treatment along the lines of equilibrium statistical mechanics. One can determine the corresponding probability distributions. If we focus onto the canonical ensemble, i. e. a point vortex system in connection with a heat bath, the probability distribution reads

$$f(\mathbf{x}_i) = Z^{-1}(\beta)\mathrm{e}^{-\beta\left[\frac{1}{2}\sum_{ij}\Gamma_i G(\mathbf{x}_i,\mathbf{x}_j)\Gamma_j\right]} = Z^{-1}(\beta)\mathrm{e}^{-\beta H} . \tag{30}$$

Onsager recognized that this probability distribution is normalizable both for negative as well as for positive values of β, which in statistical mechanics is related to temperature. A consequence of the existence of these *negative temperature states* is the tendency of the vortices to form large-scale flow structures. This property can be seen quite easily by recognizing that the probability distribution is the stationary probability distribution of the set of Langevin equations

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}_i = -\beta\nabla_{\mathbf{x}_i}H + \eta_j . \tag{31}$$

Here, η_j represents Gaussian white noise. For negative temperatures the force between two point vortices is attractive for $\Gamma_i\Gamma_j > 0$ and repulsive in the other case. This leads to the formation of large-scale flows. We refer the reader to the recent review [18].

Extension to Three Dimensions

The extension of the philosophy of point vortex motion to three dimensions leads one to consider vortex filaments. The location of a single vortex filament is given by the local induction equation for the position $\mathbf{X}(s,t)$ of the filament as a function of arclength s and time t. This equation is again obtained from Biot–Savart's law assuming a filamentary vorticity distribution:

$$\frac{\partial}{\partial t}\mathbf{X}(s,t) = \frac{\partial}{\partial s}\mathbf{X}(s,t) \times \frac{\partial^2}{\partial s^2}\mathbf{X}(s,t) \tag{32}$$

It has been shown that the single filament equation is integrable. For a discussion we refer the reader to [64].

Turbulence: Determinism and Stochasticity

It is evident that an understanding of the characteristics of turbulent fluid motions has to be based on the deterministic dynamics generated by the basic fluid dynamical equations in combination with methods of statistical physics. This has led to the field of *Statistical Hydrodynamics*, a topic which has been treated extensively by the Russian school founded by Kolmogorov [61,62]. A good overview can be found in [26,50].

Statistical Averages

In *Statistical Hydrodynamics* the Eulerian velocity field and related fields are treated as *random fields* in a probabilistic sense. To this end one has to define suitable averages, where usually *ensemble averages* are chosen. They are defined by specifying the statistics of the initial flow field by probability distributions $f(\mathbf{u}_1,\mathbf{x}_1,t=0)$, $f(\mathbf{u}_1,\mathbf{x}_1,t=0;\ldots;\mathbf{u}_N,\mathbf{x}_N,t=0)$ of the velocities $\mathbf{u}_i$ at positions $\mathbf{x}_i$ at initial time $t=0$. The transition to a continuum of points requires one to consider a probability density functional $F[\mathbf{u}(\mathbf{x}),t=0]$.

In practice, instead of ensemble averages, time averages are taken, provided the flow is stationary in a statistical sense.

The corresponding probability distributions at time t, which specify the temporal evolution of the considered statistical ensemble are given by

$$f(\mathbf{u}, \mathbf{x}, t) = \langle \delta(\mathbf{u} - \mathbf{u}(\mathbf{x}, t; \mathbf{u}_0)) \rangle \,. \tag{33}$$

Here, $\mathbf{u}(\mathbf{x}, t; \mathbf{u}_0)$ is the solution of the fluid dynamic equation with the initial value of the velocity field at point $\mathbf{x}$: $\mathbf{u}(\mathbf{x}, t = 0; \mathbf{u}_0) = \mathbf{u}_0$. The brackets denote an ensemble average. Joint probability distributions and the probability functional $F(\mathbf{u}(\mathbf{x}, t))$ are defined accordingly. Frequently, the characteristic functions, defined as the Fourier transform of the probability distributions are used. The characteristic functional

$$Z(\alpha) = \langle \mathrm{e}^{i \int \mathrm{d}t \int \mathrm{d}^3\mathbf{x} \boldsymbol{\alpha}(\mathbf{x}, t) \cdot \mathbf{u}(\mathbf{x}, t)} \rangle \tag{34}$$

is the functional Fourier transform of the probability functional. The ultimate goal of statistical hydrodynamics is the determination of this functional, since it contains all information on the various correlation functions of the Eulerian velocity fields of the statistical ensemble.

Similar probability distributions can be defined for the Lagrangian description of fluid dynamics.

Hierarchy of Moment Equations

The temporal evolution of moments of the velocity field are determined by the basic fluid dynamical equations. For instance, an equation for the moment $\langle u_i(\mathbf{x}, t) u_j(\mathbf{x}', t) \rangle$ can be obtained in a straightforward manner from the Navier–Stokes equation

$$\begin{aligned}
&\frac{\partial}{\partial t} \langle u_i(\mathbf{x}, t) u_j(\mathbf{x}', t) \rangle + \sum_k \frac{\partial}{\partial x_k} \langle u_k(\mathbf{x}, t) u_i(\mathbf{x}, t) u_j(\mathbf{x}', t) \rangle \\
&\qquad + \sum_k \frac{\partial}{\partial x'_k} \langle u_k(\mathbf{x}', t) u_i(\mathbf{x}, t) u_j(\mathbf{x}', t) \rangle \\
&= -\frac{\partial}{\partial x_i} \langle p(\mathbf{x}, t) u_j(\mathbf{x}', t) \rangle - \frac{\partial}{\partial x'_j} \langle p(\mathbf{x}', t) u_i(\mathbf{x}, t) \rangle \\
&\qquad + \nu [\Delta_x + \Delta_{x'}] \langle u_i(\mathbf{x}, t) u_j(\mathbf{x}', t) \rangle \,.
\end{aligned} \tag{35}$$

Similar equations can be formulated for the higher-order moments $\langle u_i(\mathbf{x}_1, t_1) u_j(\mathbf{x}_2, t_2) \ldots u_j(\mathbf{x}_N, t_N) \rangle$ using the Navier–Stokes equation. The equation for the average flow field $\langle \mathbf{u}(\mathbf{x}, t) \rangle$ has been considered by O. Reynolds [76] (see the discussion below). The chain of evolution equations for the higher-order correlation functions are the so-called Friedmann–Keller equations.

The moment equations are not closed. The evolution equation containing the Nth order moment contains the $(N+1)$-th order moment. This is the famous closure problem of turbulence. It is an immediate consequence of the nonlinearity of the Navier–Stokes equation. A mathematical discussion of the closure problem is given in [27].

Evolution Equations for Probability Distributions

It is straightforward to derive evolution equations for the probability distributions from the Navier–Stokes equation. This has been emphasized by Lundgren [52] and Ulinich and Lyubimov [86]. The evolution equation for $f(\mathbf{u}, \mathbf{x}, t)$ reads

$$\begin{aligned}
&\left[\frac{\partial}{\partial t} + \mathbf{u} \cdot \nabla_{\mathbf{x}} \right] f(\mathbf{u}, \mathbf{x}, t) \\
&= -\nabla_{\mathbf{u}} \cdot \int \mathrm{d}\mathbf{u}' \, \mathrm{d}\mathbf{x}' \mathbf{K}(\mathbf{x} - \mathbf{x}') (\mathbf{u}' \cdot \nabla_{\mathbf{x}'})^2 f(\mathbf{u}', \mathbf{x}', t; \mathbf{u}, \mathbf{x}, t) \\
&\qquad - \nabla_{\mathbf{u}} \cdot \int \mathrm{d}\mathbf{u}' \, \mathrm{d}\mathbf{x}' \delta(\mathbf{x} - \mathbf{x}') \Delta_{\mathbf{x}'} \mathbf{u}' f(\mathbf{u}', \mathbf{x}', t; \mathbf{u}, \mathbf{x}, t) \,.
\end{aligned} \tag{36}$$

(Here, $\mathbf{K}(\mathbf{x})$ is the gradient of Green's function $G(\mathbf{x})$ of the Laplacian.) The dynamics of the single-point probability distribution $f(\mathbf{u}, \mathbf{x}, t)$ is coupled to the two-point probability distribution $f(\mathbf{u}', \mathbf{x}', t; \mathbf{u}, \mathbf{x}, t)$ by the nonlocal pressure term and the dissipative term. Similar equations can be obtained for higher-order probability distributions. Again, the hierarchy shows clearly the closure problem of turbulence theory. The hierarchy is of considerable interest for the so-called Lagrangian pdf (probability density function) model approach advocated by S.B. Pope [72,73]. In this approach, the terms involving the two-point pdf's are modeled leading to a description of the turbulent velocity in terms of a stochastic process.

For the case of Burgers equation, which is the Navier–Stokes equation without the pressure term, it has been possible to solve the corresponding Lundgren hierarchy without approximation for a certain external forcing [17].

Functional Equations

The Lundgren hierarchy of evolution equations arises due to the fact that the N-point probability distributions contain incomplete information on the evolution of the fluid continuum. A closed evolution equation is obtained for the characteristic functional (34). This equation is the famous Hopf functional equation [38], which forms a concise formulation of the statistical treatment of fluid motions. For a more detailed treatment we refer the reader to the monograph of Monin and Yaglom [62].

Path-Integral Formulation

As has been emphasized by Martin, Siggia and Rose [58], each classical field theory can be represented in terms of the path integral formalism. This is essentially true for a fluid motion driven by a fluctuating force, which is Gaussian and δ-correlated in time. The generating (MSR) functional has the following path integral representation

$$Z(\alpha, \hat{\alpha}) = \int D\mathbf{u} D\hat{\mathbf{u}} \mathrm{e}^{S[\mathbf{u},\hat{\mathbf{u}}] + i \int \mathrm{d}\mathbf{x} \int \mathrm{d}t \boldsymbol{\alpha}(\mathbf{x},t)\cdot\mathbf{u}(\mathbf{x},t) + \hat{\boldsymbol{\alpha}}(\mathbf{x},t)\cdot\hat{\mathbf{u}}(\mathbf{x},t)} . \tag{37}$$

The functional $Z(\alpha, \hat{\alpha} = 0)$ is just the Hopf characteristic functional (34). The MSR-action is defined according to

$$\begin{aligned} S = i \int \mathrm{d}\mathbf{x} \int \mathrm{d}t \hat{\mathbf{u}}(\mathbf{x}, t) \\ \cdot \left\{ \dot{\mathbf{u}}(\mathbf{x}, t) + \mathbf{u}(\mathbf{x}, t) \cdot \nabla \mathbf{u}(\mathbf{x}, t) - \nu \Delta \mathbf{u}(\mathbf{x}, t) + \nabla p(\mathbf{x}, t) \right\} \\ - \frac{1}{2} \int \mathrm{d}t \int \mathrm{d}\mathbf{x}' \int \mathrm{d}\mathbf{x} \hat{\mathbf{u}}(\mathbf{x}, t) Q(\mathbf{x} - \mathbf{x}') \hat{\mathbf{u}}(\mathbf{x}', t) . \end{aligned} \tag{38}$$

The MSR-formalism is a convenient starting point for an analytical determination of correlation functions of the velocity field. A naive perturbation expansion of this functional yields the diagrammatic representation of the series by Wyld [89]. A renormalized perturbation expansion leads to the so-called direct interaction approximation (DIA) of Kraichnan and related analytical approximations. For an overview we refer the reader to the monographs of Lesieur [49] and McComb [59]. Also the recent work by V. L'vov and I. Procaccia [45] is based on this approach.

Reynolds Equation and Turbulence Modeling

O. Reynolds [76] suggested to decompose a turbulent flow field $\mathbf{u}(\mathbf{x}, t)$ into a mean flow $\langle \mathbf{u} \rangle$ and turbulent pulsations $\mathbf{w}(\mathbf{x}, t)$:

$$\mathbf{u}(\mathbf{x}, t) = \langle \mathbf{u}(\mathbf{x}, t) \rangle + \mathbf{w}(\mathbf{x}, t) . \tag{39}$$

By averaging the Navier–Stokes equation one ends up with the famous Reynolds's equation, which is the first equation of the hierarchy of moment equations (35):

$$\left[\frac{\partial}{\partial t} + \langle \mathbf{u} \rangle \cdot \nabla \right] \langle \mathbf{u} \rangle = -\nabla p - \sum_{ij} \frac{\partial^2}{\partial x_i \partial x_j} \langle w_i w_j \rangle + \nu \Delta \langle \mathbf{u} \rangle + \langle \mathbf{f} \rangle . \tag{40}$$

This equation contains the so-called Reynolds stress tensor $\langle w_i w_j \rangle$, which cannot be neglected since the turbulent pulsations $\mathbf{w}$ can be larger than the averaged velocity $\langle \mathbf{u} \rangle$. The turbulent pulsations $\mathbf{w}$ are closely linked to the *fine structures of turbulence*.

If the Reynolds stress tensor is known as a functional of the mean velocity field, the Reynolds equation is a closed evolution equation determining average flow properties. This makes numerical computations of the average flow quantities rather efficient since the *fine structures* or the *small-scale flow* have not to be resolved by numerical schemes.

The Reynolds stress tensor has been the subject of various investigations. Since a general theory determining the Reynolds stress tensor is lacking, engineers have developed the area of *turbulence modeling* in order to overcome the closure problem. Famous turbulence models are eddy viscosity models, which replace the Reynolds stress tensor by an effective damping term modeling the energy flux from the averaged flow into the turbulent pulsations, or so-called $K - \epsilon$ models, which are based on the evolution equation for the local energy dissipation rate ϵ of the turbulent pulsations.

Turbulence modeling has to fulfill the *requirement of physical realizability* and should, for example, not lead to the development of negative kinetic energies. Furthermore, symmetry arguments should be taken into account [65].

Turbulence modeling has led to the development of the field of *Large Eddy Simulations* (LES), which has provided a variety of numerical, also commercially available schemes for calculating the large-scale flows of applied fluid dynamical problems. The LES-approach, however, is limited by the fact that the properties of the Reynolds stresses cannot yet be derived from a physical treatment of the small-scale turbulent pulsations of flows and the obtained numerical results have to be met with caution. For details we refer the reader to Piquet [71], Jovanovic [40], and the reviews of Métais and Leschziner in [50]. Of considerable interest are the two articles of Johansson and Oberlack in [66].

The Fine Structure of Turbulence

The fine structure of turbulence essentially influences the large-scale flows via the Reynolds stresses. Therefore, the investigation of the fine structure of turbulent flows is a central theme in turbulence research. It is commonly believed that the statistical characteristics of the turbulent fine structures are universal. A point emphasized by U. Frisch [26] is that in the fine structures symmetries of

the Euler equation of fluid dynamics are restored in a statistical sense. The symmetries of the Euler equations are translational symmetry, isotropy, and rescaling symmetry of space, time, and velocity. Although each of these symmetries are broken by the turbulent flows, the symmetries are restored for averaged quantities. If this hypothesis is true, then the turbulent fine structure is related with a universal state, which is called *fully developed stationary, homogeneous, and isotropic turbulence*.

Increments

The fine-scale structure of turbulence is evaluated by the introduction of the so-called velocity increment $\mathbf{v}_{\mathbf{x}}(\mathbf{r}, t)$

$$\mathbf{v}_{\mathbf{x}}(\mathbf{r}, t) = \mathbf{u}(\mathbf{x} + \mathbf{r}, t) - \mathbf{u}(\mathbf{x}, t) . \tag{41}$$

Velocity increments are defined with respect to a reference point $\mathbf{x}$. A mean flow is eliminated by the definition of increments. Furthermore, one may consider velocity increments with respect to a moving reference point $\mathbf{x} = \mathbf{X}(\mathbf{y}, t)$. In the following we shall consider such moving increments.

The corresponding evolution equation for the incompressible velocity increment can easily be established using the definition (41) and the Navier–Stokes equation:

$$\begin{aligned}\left[\frac{\partial}{\partial t} + \mathbf{v}(\mathbf{r}, t) \cdot \nabla_{\mathbf{r}}\right]\mathbf{v}(\mathbf{r}, t) &= -\nabla_r p(\mathbf{r}, t) + \nu \Delta_r \mathbf{v}(\mathbf{r}, t) \\ &\quad - \left[-\nabla_r p(\mathbf{r}, t) + \nu \Delta_r \mathbf{v}(\mathbf{r}, t)\right]_{\mathbf{r}=0} , \\ \nabla \cdot \mathbf{v}_{\mathbf{x}}(\mathbf{r}, t) &= 0 .\end{aligned} \tag{42}$$

Length and Time Scales in Turbulent Flows

The Integral Scale The integral scales are measures for a spatial distance L or a time interval T, across which the turbulent fluctuations become uncorrelated. The integral length scale is based on the velocity-velocity correlation function

$$\langle u(x + r, t) u(x, t)\rangle = \langle u(x, t) u(x, t)\rangle F\left(\frac{r}{L}\right) , \tag{43}$$

which decays to zero as a function of the distance r. The integral length scale L is defined by the integral

$$L = \int_0^\infty \mathrm{d}r F\left(\frac{r}{L}\right) . \tag{44}$$

Similarly, one can define temporal integral scales based on the decay of temporal correlations $\langle u(x, t + \tau) u(x, t)\rangle = \langle u(x, t) u(x, t)\rangle F\left(\frac{\tau}{T}\right)$ leading to the definition of the integral time scale T. An integral velocity scale can be formed quite naturally by the ratio

$$U_{\mathrm{int}} = L/T . \tag{45}$$

The velocity field $u(x, t)$ describes the turbulent pulsations. A mean flow has already been subtracted. The correlation functions (43) actually are tensors. For flows which are nonisotropic in the statistical sense, it might be necessary to introduce different integral length scales.

The Kolmogorov Scales The Navier–Stokes equation involves the kinematic viscosity ν as well as a measure of the excitation of turbulence, the mean local energy dissipation ϵ. It is convenient to form length and time scales from these two quantities and obtain the so-called Kolmogorov scales

$$\eta = \left(\frac{\nu^3}{\epsilon}\right)^{1/4} , \quad \tau_\eta = \left(\frac{\nu}{\epsilon}\right)^{1/2} , \quad u_\eta = \frac{\eta}{\tau_\eta} = (\epsilon\nu)^{1/4} . \tag{46}$$

On the basis of these quantities one can define a Reynolds number which turns out to be unity, $\mathrm{Re} = \frac{\eta u_\eta}{\nu} = 1$. Therefore, the Kolmogorov scales have to be related with small scale motions, which, due to dissipation, can be considered to be laminar on these scales.

Relation Between the Integral and the Kolmogorov Length Scale One may find a relation between the integral length scale L and the Kolmogorov length scale η. This relation is based on the observation that the local energy dissipation rate ϵ can be dimensionally expressed in terms of the velocity at the integral scale, U_{int}, Eq. (45) via

$$\epsilon \approx \frac{U_{\mathrm{int}}^3}{L} . \tag{47}$$

This relation allows one to determine the ratio of the intergal scale L and the Kolmogorov scale η leading to a Reynolds number dependence

$$\frac{L}{\eta} \approx \mathrm{Re}^{3/4} = \left(\frac{U_{\mathrm{int}} L}{\nu}\right)^{3/4} . \tag{48}$$

The ratio increases with the Reynolds number. Typical length scales estimated from experimental data of a grid flow are shown in Fig. 5. It is seen that the estimate (48) can be confirmed experimentally.

From a dynamical point of view the quantity $(\frac{L}{\eta})^3 \approx \mathrm{Re}^{9/4}$ is a measure of the number of active degrees of freedom of the fluid motion.

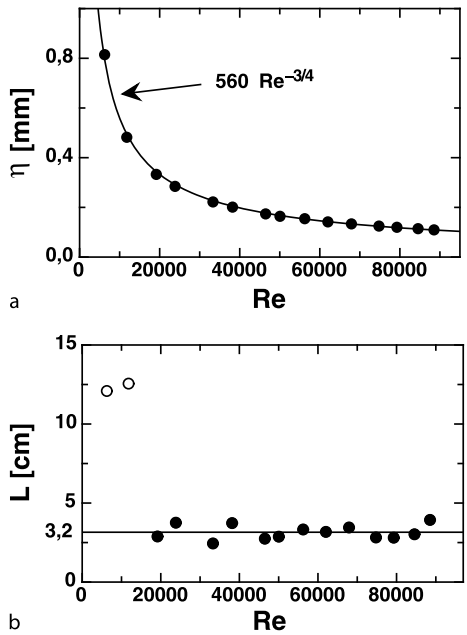

Fluid Dynamics, Turbulence, Figure 5
Reynolds number dependence of the Kolmogorov scale (a) and the integral length scale (b) estimated from data of a grid experiment, [51]

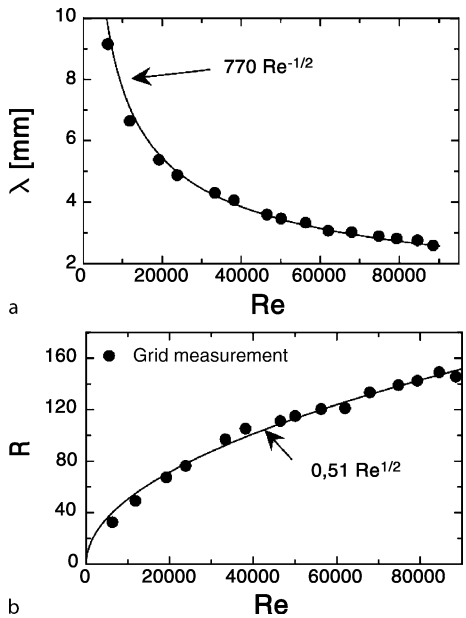

Fluid Dynamics, Turbulence, Figure 6
Reynolds number dependence of the Taylor scale (a) and the Taylor length-based Reynold number from data of a grid experiment, [51]

The Taylor Length and the Taylor-Based Reynolds Number A third length scale has been used to characterize a turbulent flow field. This is the so-called Taylor length, which frequently is denoted also as Taylor microscale, λ.

It can be derived from the velocity u (here u denotes again the fluctuating part of a measured velocity signal) and the derivative $\partial u/\partial x$ as

$$\lambda^2 = \lim \frac{\langle u^2 \rangle}{\langle (\partial u/\partial x)^2 \rangle} . \tag{49}$$

Aronson and Löfdahl [6] suggested to estimate the Taylor length using the velocity increment $v(r) = u(x+r) - u(x)$:

$$\lambda^2 = \lim_{r \to 0} \frac{\langle u(x)^2 \rangle \, r^2}{\langle v(r)^2 \rangle} . \tag{50}$$

For the Taylor length it is expected that $\frac{L}{\lambda} \approx \mathrm{Re}^{1/2}$ holds, which is also found experimentally see Fig. 6a.

On the basis of this Taylor length it is also possible to define a new Reynolds number by the ratio of velocity times length scale and kinematic viscosity. The Taylor length-based Reynolds number R_λ should scale with the square of the usual Reynolds number, see Fig. 6b.

Statistics of Increments: Structure Functions

A great amount of work in turbulence research has been devoted to the so-called structure functions. In general, structure functions are equal time moments of the velocity increments:

$$\langle v_i(\mathbf{r}, t) v_j(\mathbf{r}, t) \rangle , \quad \langle v_i(\mathbf{r}, t) v_j(\mathbf{r}, t) v_k(\mathbf{r}, t) \rangle . \tag{51}$$

For stationary, homogeneous, and isotropic turbulence the tensorial quantities can be considerably reduced by symmetry arguments. For homogeneous and isotropic turbulence the second- and third-order moments can be related to the so-called longitudinal structure functions, i. e. the moments of the component of the velocity increment in the direction of $\mathbf{r}$, $\mathbf{e}_r \cdot \mathbf{v}(\mathbf{r}, t)$. They are defined according to

$$S^N(r) = \left\langle \left(\frac{\mathbf{r}}{r} \cdot \mathbf{v}(\mathbf{r}, t) \right)^N \right\rangle . \tag{52}$$

For small values of r, the structure functions have to behave like $S^N \approx r^N$, since for small r the velocity increment can be expanded in a Taylor series.

The structure functions of different orders are related through the Navier–Stokes dynamics. The evaluation of the structure functions from the resulting hierarchy of evolution equations is one of the major theoretical challenges in turbulence research. However, it seems that

so far only the lowest-order equations relating second- and third-order structure (the so-called Kolmogorov's -4/5 law) functions have been exploited rigorously.

Kolmogorov's -4/5 Law

Kolmogorov (c. f. [26,62]) showed that the mean kinetic energy in an *eddy of scale r*, $\langle \mathbf{v_x}(\mathbf{r},t)^2\rangle$, in a homogeneous, isotropic turbulent field is given by the equation

$$\frac{\partial}{\partial t}\left\langle \frac{\mathbf{v}_x(\mathbf{r},t)^2}{2}\right\rangle + \nabla_r \cdot \left\langle \mathbf{v}_x(\mathbf{r},t)\frac{\mathbf{v}_x(\mathbf{r},t)^2}{2}\right\rangle = \nu\Delta\langle \mathbf{v_x}(\mathbf{r},t)^2\rangle - \langle\epsilon\rangle\,, \quad (53)$$

where $\langle\epsilon\rangle$ denotes the mean local energy dissipation rate

$$\langle\epsilon\rangle = \frac{\nu}{2}\sum_{ij}\left\langle\left(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right)^2\right\rangle. \quad (54)$$

We formally consider the *turbulent limit* $\nu \to 0$. In this limit the mean local energy dissipation rate $\langle\epsilon\rangle$ has to be constant. Thus, one can estimate that the gradients have to behave like

$$\frac{\partial u_i}{\partial x_j} \approx \frac{1}{\sqrt{\nu}}\,. \quad (55)$$

The relation (53) can be obtained as a balance equation for the density of the mean kinetic energy from the evolution equation for the velocity increment (42) by scalar multiplication with **v**, and subsequent averaging. The correlations involving the pressure term drop out due to homogeneity.

The resulting equation reads:

$$S^3(r) - 6\nu\frac{\mathrm{d}}{\mathrm{d}r}S^2(r) = -\frac{4}{5}\langle\epsilon\rangle r\,. \quad (56)$$

This equation relates the second- and third-order structure functions and, at first glance, seems to be underdetermined. However, for small values of r we can neglect the third-order structure function, since $S^3(r) \approx r^3$. The second-order structure function is then given by

$$S^2(r) = \frac{\langle\epsilon\rangle}{15\nu}r^2\,. \quad (57)$$

The range of validity of this law is denoted as the *dissipative range* and is close to the Kolmogorov length and smaller. It is dominated by viscous structures.

In the second regime, whose existence is inferred by the -4/5th law, the third-order structure function dominates, leading to

$$S^3(r) = -\frac{4}{5}\langle\epsilon\rangle r\,. \quad (58)$$

This defines the so-called *inertial range* and is located between the Taylor length and the integral length. For data from a free jet experiment with high Reynolds numbers we show in Fig. 7 the third-order structure function in different presentations, which clearly demonstrate the existence of the inertial range.

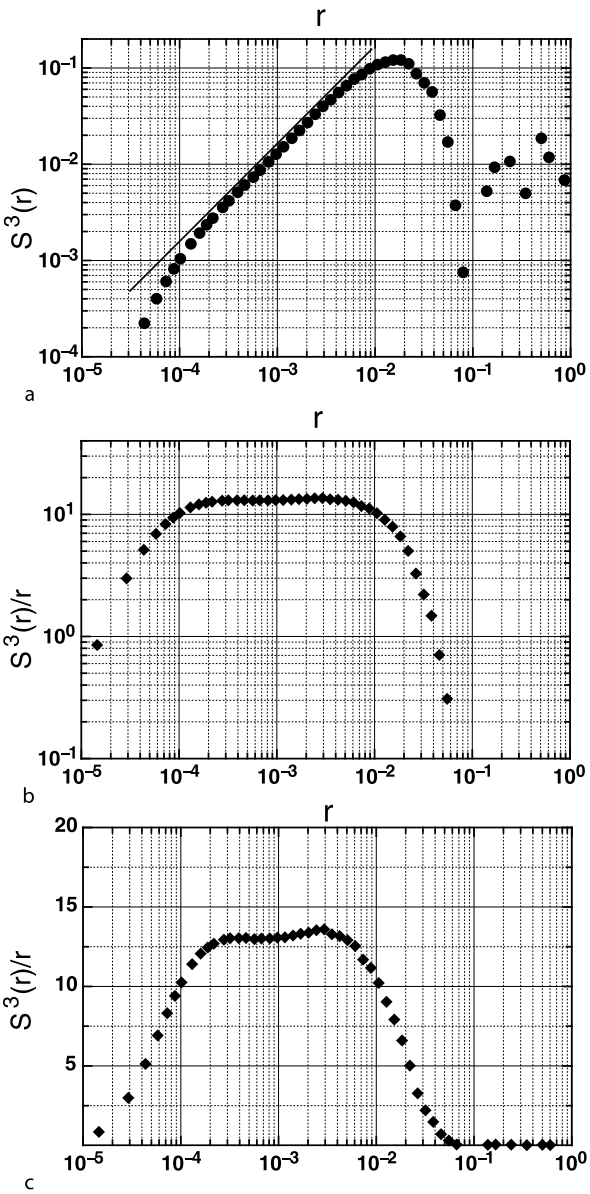

Fluid Dynamics, Turbulence, Figure 7
Third-order structure function $S^3(r)$ (a) of cryogenic free jet measurements with Re = 210.000 and the compensated structure functions $S^3(r)/r$ in a semi logarithmic plot (b) and in a linear plot (c), showing in more detail the quality of the present scaling behavior. (Note here we used the absolute values for $S^3(r)$, thus the values are all positive). [9]

Phenomenological Theories of Turbulence

Kolmogorov's Theory K41: Selfsimilarity in the Inertial Range

The inertial range is characterized by the decay of eddies, which can be assumed to be self-similar. As a consequence, the probability distribution of the longitudinal velocity increment v at scale r, $f(v, r)$, should have the form

$$f(v, r) = \frac{1}{\sqrt{\langle v(r)^2\rangle}} F\left(\frac{v}{\sqrt{\langle v(r)^2\rangle}}\right) , \tag{59}$$

where $F(\xi)$ is a universal function in the range of scales $\lambda < r < L$. As a consequence, the nth order moments should have the form

$$\begin{aligned}\langle v^n(r)\rangle &= \int \mathrm{d}v\, v^n \frac{1}{\sqrt{\langle v(r)^2\rangle}} F\left(\frac{v}{\sqrt{\langle v(r)^2\rangle}}\right) \\ &= \langle v(r)^2\rangle^{n/2} \int \mathrm{d}w\, w^n F(w) = \langle v(r)^2\rangle^{n/2} V_n .\end{aligned} \tag{60}$$

Since the r-dependence of the 3rd-order moment is known in the limit of high Reynolds number,

$$\langle v^3(r)\rangle = V_3 \langle\epsilon\rangle r = \langle v(r)^2\rangle^{3/2} V_3 , \tag{61}$$

we obtain

$$\langle v(r)^2\rangle = K\,(\langle\epsilon\rangle r)^{2/3} . \tag{62}$$

The K41 assumption of self-similarity of the velocity increment statistics in the inertial range leads to the following fractal scaling behavior of the nth order moments

$$\langle |v(r)|^n\rangle = K_n\,(\langle\epsilon\rangle r)^{n/3} . \tag{63}$$

The corresponding probability distribution is entirely determined by the constants K_n. The scaling exponents $\zeta_n = N/3$ are linear functions of n.

Failure of K41: Intermittency

Kolmogorov's hypothesis on the self-similarity of the statistics in the inertial range has been tested experimentally as well as numerically. Figure 8 exhibits the probability distribution of the scaled variable $\tilde{v}$

$$\tilde{v} = \frac{v}{\sigma(r)} \quad \text{with } \sigma(r) = \sqrt{\langle v(r)^2\rangle} . \tag{64}$$

Because of K41 theory all probability distributions should collapse for values of r taken from the inertial range. Both, experimental and numerical results show clear deviations from this behavior. Although the scaling exponents of the second-order structure function is close to the K41 value 2/3, a characteristic change of shape of the probability distribution can be detected. The change of the pdf with scale r in the inertial range is a signature of the phenomenon called *intermittency*. Consequently, the structure functions do not scale as suggested by the theory of Kolmogorov. This has been experimentally documented by [1] and has been discussed by several groups.

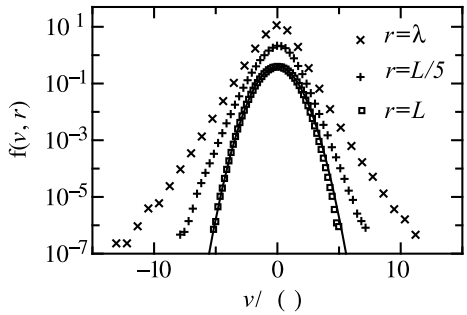

Fluid Dynamics, Turbulence, Figure 8
Probability density functions (pdf) for velocity increments on three different length scales $r = L, L/5, L/30 \approx \lambda$. The pdfs are shifted along the ordinate for a better representation. On the largest scale L a Gaussian distribution is fitted for comparison. Towards smaller scales the deviations from the Gaussian form become obvious ($R_\lambda = 180$). For further details see [79]

Kolmogorov's Theory K62

In a famous note in the volume on hydrodynamics in his Course on Theoretical Physics Landau [47] remarked that the formula of the K41 theory contains the mean value of the local energy dissipation rate, $\langle\epsilon\rangle$. However, this quantity is a strongly fluctuating quantity in space, due to the strong spatial variations of the velocity gradient, as can be seen from Fig. 9. According to Landau these fluctuations should show up in the structure functions. Instead of the K41-result (63) he suggested use of the following representation

$$\langle |v(r)|^n\rangle = \tilde{K}_n \langle \epsilon_r^{n/3}\rangle r^{n/3} \tag{65}$$

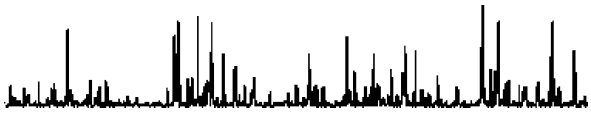

Fluid Dynamics, Turbulence, Figure 9
Spatial distribution of the local energy dissipation rate. The quantity strongly fluctuates in space. Data is obtained from the experiment [79]

where the quantity ϵ_r denotes the local energy dissipation rate averaged over a sphere of radius r. This reasoning leads to an extension of the K41 formula for the probability distribution in the form

$$f(v,r) = \int \mathrm{d}\epsilon_r p\left(\epsilon_r, \frac{r}{L}\right) \frac{1}{(\epsilon r)^{2/3}} F\left(\frac{v}{(\epsilon_r r)^{2/3}}\right) . \quad (66)$$

In 1962 Kolmogorov [44] suggested use of a lognormal distribution for the local energy dissipation rate ϵ_r, whose variance is $\mu \ln(L/r)$. As a result, the structure functions scale like

$$\langle |v(r)|^n \rangle = \tilde{K}_n r^{n/3} (L/r)^{\mu n(n-3)} . \quad (67)$$

The experimental value for μ can be obtained from a fit of the K62 formula to experimental data. However, we note that the formula can only be valid for small values of n, since the scaling exponents ζ_n have to be a monotonously increasing function of the order n of the structure function.

The Multifractal Model

The multifractal model was introduced by U. Frisch and G. Parisi. For a detailed description we refer the reader to [26]. The basic idea is to view a turbulent field to be composed of regions where the velocity increment field is assumed to be characterized by a scaling index h:

$$v(r,t) = \beta \left(\frac{r}{L}\right)^h . \quad (68)$$

The structure functions are then given by

$$S^N(r) = \beta_N \int \mathrm{d}h P(h,r) \left(\frac{r}{L}\right)^{Nh} , \quad (69)$$

where $P(r,h)\mathrm{d}h$ is the probability to find increments for a certain scale r with scaling exponent h. Assuming self-similarity $P(h,r)$ should have the form

$$P(h,r) = \left(\frac{r}{L}\right)^{3-D(h)} . \quad (70)$$

Consequently, we have

$$S^N(r) = \beta_N \int \mathrm{d}h \left(\frac{r}{L}\right)^{Nh+3-D(h)} \approx r^{\zeta_N} , \quad (71)$$

where the evaluation of the integral with the method of steepest descend yields

$$\zeta_N = \mathrm{Min}[Nh + 3 - D(h)] . \quad (72)$$

The scaling indices ζ_N are related to the dimension $D(h)$ via a Legendre transform. Recently, an extension of the multifractal model has been presented by Chevillard et al. [10]. This approach is essentially based on the representation of the probability distribution due to [8]. They succeeded to obtain a model for the symmetric part of the probability function of the longitudinal velocity increment, which is valid both in the integral as well as the dissipative scale. This gives a reasonable approximation to the experimentally determined probability distribution $f(v,r)$.

As we have seen, the statistics of the longitudinal velocity increment for a single scale r can be modeled in various ways in order to describe the deviations from the fractal scaling behavior predicted by the phenomenological theory of Kolmogorov formulated in 1941. The major shortcoming of these approaches, however, is the fact that they contain no information on the joint statistics of the velocity fields at different scales and times. However, due to the presumed energy cascade, the velocity increments on different scales have to be correlated.

Multiscale Analysis of Turbulent Fields

The spatial correlation of the velocity of turbulent fields has been examined in two ways. From a field theoretic point of view, pursued by V. L'vov and I. Procaccia, the existence of so-called fusion rules have been hypothesized. For this approach we refer the reader to the survey [45] and the work cited therein. A phenomenological approach has been performed by Friedrich and Peinke [22,23]. In this approach notions from the theory of stochastic processes have been used in order to characterize multiscale statistics of velocity increments. Relations to the fusion rule approach has been discussed in [15]. In the following we shall discuss the phenomenological approach to multiscale statistics of turbulence.

Statistics Across Scales

In order to address the spatial signatures of the cascading process underlying stationary, homogeneous, and isotropic turbulence it is necessary to consider the probability distributions for velocity increments at different scales:

$$f^N(\mathbf{v}_1, r_1; \mathbf{v}_2, r_2; \ldots; \mathbf{v}_N, r_N) . \quad (73)$$

Thereby, the quantities $\mathbf{v}_i$ are velocity increments with respect to a common point of reference, $\mathbf{v}_i(\mathbf{r}_i, t) = \mathbf{u}(\mathbf{x} + \mathbf{r}_i, t) - \mathbf{u}(\mathbf{x}, t)$ and different distances $\mathbf{r}_i$.

In the following we shall consider the longitudinal velocity increments and locations $\mathbf{r}_i$ positioned along a straight line. These quantities are easily accessible from experimental data.

In case that the longitudinal velocity increments v_i of different scales r_i are statistically independent, the N-point probability distribution simply factorizes:

$$f^N(v_1, r_1; v_2, r_2; \ldots; v_N, r_N) = f^1(v_1, r_1) f^1(v_2, r_2) \ldots f^1(v_N, r_N) . \quad (74)$$

Because of the cascading process, which involves the dynamics of velocity increments at different scales, this relationship cannot hold true. However, provided the cascade is generated locally by the nonlinear interaction of velocity increments at neighboring scales, one may expect that the two scale probability distributions or the conditional probability distribution

$$p(v_1, r_1 | v_2, r_2) = \frac{f^2(v_1, r_1; v_2, r_2)}{f^1(v_2, r_2)} \quad (75)$$

contains the most important information on the N-scale probability distribution (73).

Markovian Properties

One may wonder whether the knowledge of the conditional probability distribution suffices to reconstruct the N-scale probability distribution (73) in the form

$$f^N(v_1, r_1; \ldots; v_N, r_N) = p(v_1, r_1 | v_2, r_2) \ldots p(v_{N-1}, r_{N-1} | v_N, r_N) \cdot f(v_N, r_N) . \quad (76)$$

In this case, the probability distribution f^N defines a Markov chain.

The question, whether Markovian properties in scale exist for fully developed turbulence, has been pursued in several ways. First of all, a necessary condition for the existence of Markovian properties is the validity of the Chapman–Kolmogorov equation:

$$p(v_1, r_1 | v_3, r_3) = \int \mathrm{d}v_2 \, p(v_1, r_1 | v_2, r_2) p(v_2, r_2 | v_3, r_3) . \quad (77)$$

This approach has been pursued in [24]. Second, one can validate the Markovian properties by a direct inspection of the conditional probability distribution (75). Defining the conditional probability distribution

$$p(v_1, r_1 | v_2, r_2; v_3, r_3) = \frac{f^3(v_1, r_1; v_2, r_2; v_3, r_3)}{f^2(v_2, r_2; v_3, r_3)} , \quad (78)$$

the Markovian property can be assessed by comparing the conditional probability distributions

$$p(v_1, r_1 | v_2, r_2; v_3, r_3) = p(v_1, r_1 | v_2, r_2) . \quad (79)$$

Summarizing the outcomes of the experimental investigations one can state that the Markovian property can be empirically validated provided the differences of the scales $r_1 - r_2, r_2 - r_3$ are not too small. This statement can be made even more precise. As has been shown in [51], the Markovian property breaks down provided the scale differences are smaller than the Taylor microscale [see Eq. (49)]. This finding attributes a statistical definition to the Taylor length scale. Because of the memory effect we have now called this length Markov–Einstein coherence length L_{mar} [51].

Estimation of the Conditional Probability Distribution

Markov processes are defined through their conditional probability distributions. If one considers the statistics on scales larger compared to the Markov–Einstein length, one may perform the limit $L_{\mathrm{mar}} \to 0$ and consider the process to be continuous in scale r, the conditional probability distribution obeys a Fokker–Planck equation of the form

$$-\frac{\partial}{\partial r} p(v, r | v_0 r_0) = \left[-\frac{\partial}{\partial v} D^1(v, r) + \frac{\partial^2}{\partial v^2} D^2(v, r) \right] p(v, r | v_0 r_0) , \quad (80)$$

where the statistics are determined by the drift function $D^1(v, r)$ and the diffusion function $D^2(v, r)$. We remind the reader that in the definition of $p(v, r | v_0, r_0)$ we have used $r_0 > r$, which leads to the minus sign in the Fokker–Planck equation.

The drift function and the diffusion function have been determined empirically using methods of data analysis of stochastic processes, for further details see [74]. We also refer the reader to [57].

Path-Integral Representation of the N-Scale Probability Distribution

Since the joint N-point probability distribution can be constructed by the conditional probability function due to the Markovian property one can write down a path integral formula for the turbulent cascade in the form

$$F[v(r)] = Z^{-1} \exp \left[-\int \mathrm{d}r' \frac{\left[-\frac{\mathrm{d}v(r')}{\mathrm{d}r'} - D^1(v(r'), r') \right]^2}{D^2(v(r'), r')} \right] . \quad (81)$$

This probability distribution is the analog to the Gibbs distribution describing the statistics of systems in thermodynamic equilibrium.

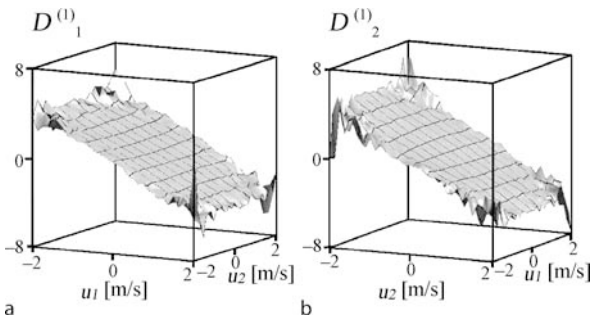

Fluid Dynamics, Turbulence, Figure 10
The u_1 and u_2 dependence of the drift vector for the scale $r = L/4$. a The drift coefficient $D_1^{(1)}$, b the drift coefficient $D_2^{(1)}$. Note that the vertical axis is rotated for better comparison between a and b. Both coefficients are linear functions in u [79]

Statistics of Longitudinal and Transversal Components

Recently, an analysis of the joint statistics of the longitudinal and transversal components of the velocity increments has been performed [78,79]. The result is a Fokker–Planck equation of the form

$$-r\frac{\partial}{\partial r}p(\mathbf{u}, r|\mathbf{u}_0, r_0) = \left\{-\frac{\partial}{\partial u_i}D_i^{(1)}(\mathbf{u}, r) + \frac{\partial^2}{\partial u_i \partial u_j}D_{ij}^{(2)}(\mathbf{u}, t)\right\} \cdot p(\mathbf{u}, r|\mathbf{u}_0, r_0)\,. \quad (82)$$

Here, $\mathbf{u} = (u_1, u_2)$ denotes the increment vector of the longitudinal and transversal components, respectively. As mentioned above and worked out in more detail in another contribution to this Encyclopedia of Complexity and System Science [25] both drift and diffusion terms $D^{(1)}, D^{(2)}$ can be estimated directly from given data. Typical results are shown in Figs. 10 and 11. The drift terms turn out to be linear,

$$D_i^{(1)}(\mathbf{u}, r) = d_i(r)u_i\,. \quad (83)$$

Furthermore, the diffusion matrix can be approximated by low-order polynomials

$$D_{ij}^{(2)}(\mathbf{u}, r) = d_{ij}(r) + \sum_k d_{ij;k}u_k + \sum_{kl} d_{ij;kl}(r)u_k u_l\,. \quad (84)$$

Knowing the drift and diffusion coefficient it is possible to solve the Fokker–Planck equation numerically, which can be taken as a selfconsistent verification of the estimation procedure. In Fig. 12 it is shown that the numerical solution of the Fokker–Planck equation reproduces the probability distribution directly obtained from the data quite well.

It is quite interesting to notice that the resulting moment equations reproduce Karman's equation, which is a relation between the longitudinal and transversal velocity increments

$$\langle u_2(r)^2\rangle = \frac{1}{2r}\frac{\mathrm{d}}{\mathrm{d}r}r^2\langle u_1(r)^2\rangle\,. \quad (85)$$

Karman's equation is a direct consequence of the incompressibility of the fluid motion. It can be interpreted as a low-order Taylor expansion of $\langle u_2(r)^2\rangle = \langle u_1(\frac{3}{2}r)^2\rangle$ which leads to the proposition that the complex structures of longitudinal and transversal velocity increments mainly differ in a different scale parametrization of the cascade, $r \to \frac{3}{2}r$.

Using these findings the different scaling exponents found for longitudinal and transversal increments can be explained. In Fig. 13 the longitudinal and transversal

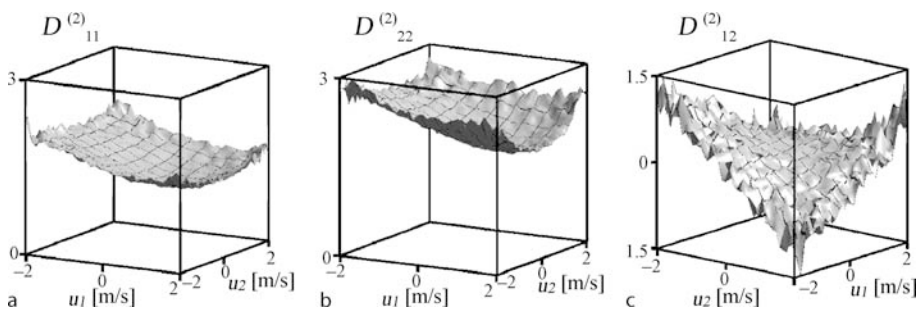

Fluid Dynamics, Turbulence, Figure 11
The u_1 and u_2-dependence of the diffusion matrix for the scale $r = L/4$. a The coefficient $D_{11}^{(2)}$, b the coefficient $D_{22}^{(2)}$. It can be seen that the diagonal coefficients are not constant but have a parabolic form, which is more pronounced for the $D_{22}^{(2)}$ coefficient (multiplicative noise). Both coefficients are symmetric under reflection with respect to $u_2 \to -u_2$, but not for $u_1 \to -u_1$. c The saddle-formed off-diagonal coefficient $D_{12}^{(2)}$ [79]

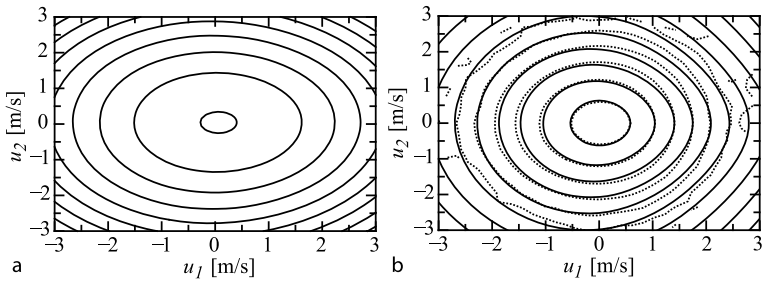

Fluid Dynamics, Turbulence, Figure 12

Solution of the Fokker–Planck equation. a Contour plot of the initial condition in logarithmic scale. The simulation starts at the integral length $r = L$ with a Gaussian distribution fitted to the data. b The contour plots in logarithm scale of the simulated probability distribution on the scale $r = 2\lambda$. The distance between the contour lines is chosen in logarithmic scale and corresponds to a factor 10. *Dashed lines* indicate the probability distribution calculated directly from data, the *full lines* are the simulation ones. The simulation reproduces well the properties of the data [79]

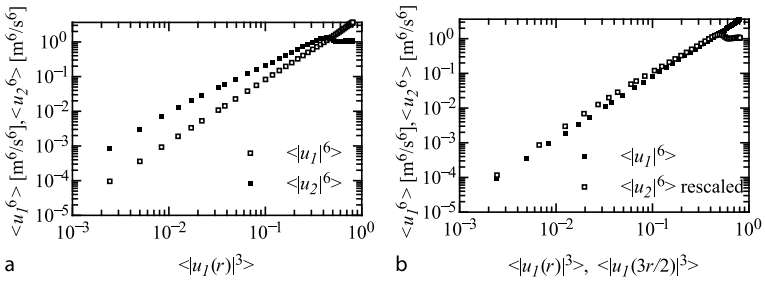

Fluid Dynamics, Turbulence, Figure 13

The sixth-order longitudinal (*black squares*) and transversal (*white squares*) structure function in a ESS representation and b ESST representation [79]

structure function are shown as a function of the third-order structure function, which should reveal the scaling behavior, too, as the third-order structure function is expected to depend linearly on r. (This presentation is called ESS – extended self-similarity [7]). Most interestingly, the difference in the scaling exponents vanishes if the 3/2 rescaling is applied, see Fig. 13b. The result indicates that proper scaling behavior is not detected and can only be valid approximately.

The Markov analysis of turbulent velocity fields yields a closed phenomenological description of the spatial complexity of turbulence. We note that a similar multiscale analysis has been performed for a turbulent passive scalar field [85].

Lagrangian Fluid Dynamics

Because of experimental progress in the detection of the Lagrangian path of passive tracer particles [46,63] interest in the formulation of stochastic processes describing Lagrangian motion in turbulent fluids has been renewed. One of the first descriptions of such a process is due to Obukhov [67], who suggested to model the Lagrangian acceleration as a white noise process. He formulated the corresponding evolution equation for the probability distribution $f(\mathbf{x}, \mathbf{u}, t)$ specifying the Lagrangian path in form of a Fokker–Planck equation. The results, which can be derived from the Fokker–Planck equation indicate scaling behavior for the Lagrangian velocity increments and the distance traveled by the particle during time t:

$$\begin{aligned} \langle(\mathbf{u}(t) - \mathbf{u}(0))^2\rangle &= c\langle\epsilon\rangle t\,, \\ \langle(\mathbf{x}(t) - \mathbf{x}(0))^2\rangle &= d(\langle\epsilon\rangle t)^3\,. \end{aligned} \tag{86}$$

This scaling behavior is inferred from dimensional considerations along the theory of Kolmogorov (K41). Therefore, it is not a surprise that the experimentally observed probability distributions deviate from the Gaussian shape required by the Oboukhov model [63]. A modification of Oboukhov's model has been suggested in [21]. Recently, Lagrangian particle statistics has been modeled on the basis of a simple vortex model [87], reproducing qualitatively

the intermittent characteristics of Lagrangian velocity increments [63].

The investigations of the Lagrangian statistics of turbulent flows have benefited a lot from the investigation of passive tracers in disordered flows. Especially the treatment of the Kraichnan model has led to considerable insights into the statistics of particles in flows. We refer the reader to the review article [19].

Future Directions

Although a fundamental understanding is still lacking turbulence research has always led to new concepts and scientific ideas, which substantially have influenced the development of modern science. This will especially hold true for the future, where one may expect a major breakthrough due to combined efforts of experimental, numerical, and analytical work. The following points for future direction of research are only a rather subjective listing of the authors.

Fine Structure of Turbulence: A question which should be investigated experimentally in more detail is whether and how the fine structure is influenced by the mechanism of the generation of turbulence. This will also involve turbulent flows close to walls or in pipes. From the experimental side high precision measurements are required.

Geometrical and Topological Aspects of Turbulence: There is a variational formulation of ideal hydrodynamics, emphasized by V.I. Arnold [5]. We expect that research based on topological and geometrical reasoning will yield further important insights into the spatio-temporal organization of fluid flow.

Statistical Properties of the Lagrangian Map: Topological and geometrical aspects are intimately related with the properties of the Lagrangian map. A further analysis of the trajectories of several Lagrangian particles is currently under consideration. A highly interesting question is whether the stochastic processes of the particle motion in the inertial range can be assessed.

Further Reading

For further reading we suggest the monographs and proceedings [5,13,14,26,33,61,62,73,80].

Acknowledgments

We thank all members of our working groups for fruitful collaboration. We thank T. Christen and M. Wilczek for crictically reading the manuscript.

Bibliography

1. Anselmet F, Gagne Y, Hopfinger EJ, Antonia RA (1984) High order structure functions in turbulent shear flows. J Fluid Mech 140:63
2. Antonia RA, Ould-Rouis M, Zhu Y, Anselmet F (1997) Fourth-order moments of longitudinal- and transverse-velocity structure functions. Europhys Lett 37:85
3. Aref H (1983) Integrable, chaotic, and turbulent vortex motion in two-dimensional flows. Ann Rev Fluid Mech 15:345
4. Aref H (2007) Vortices and polynomials. Fluid Dyn Res 39:5
5. Arnold VI, Khesin BA (1999) Topological methods in hydrodynamics. Springer, Berlin
6. Aronson D, Löfdahl LL (1993) The plane wake of a cylinder. Measurements and inferences on the turbulence modelling. Phys Fluids A 5:1433
7. Benzi R, Ciliberto S, Baudet C, Chavarria GR, Tripiccione R (1993) Extended self-similarity in the dissipation range of fully-developed turbulence. Europhys Lett 24:275
8. Castaing B, Gagne Y, Hopfinger EJ (1990) Velocity probability density functions of high Reynolds number turbulence. Phys D 46:177
9. Data provided by Chabaud B, Chanal O, CNRS Gernoble, France
10. Chevillard L, Roux SG, Leveque E, Mordant N, Pinton J-F, Arneodo A (2006) Phys D 218:77
11. Chorin AJ, Marsden GE (2000) A mathematical introduction to fluid mechanics. Springer, Berlin
12. Cuypers Y, Maurel A, Petitjeans P (2003) Vortex burst as a source of turbulence. Phys Rev Lett 91:194502
13. Darrigol O (2005) Worlds of Flow. Oxford University Press, Oxford
14. Davidson PA (2004) Turbulence. Oxford University Press, Oxford
15. Davoudi J, Tabar MRR (1999) Theoretical Model for Kramers-Moyal's description of Turbulence Cascade. Phys Rev Lett 82:1680
16. Doering CR, Gibbon JD (1995) Applied Analysis of the Navier-Stokes Equations. Cambridge University Press, Cambridge
17. Eule S, Friedrich R (2005) A note on a random driven Burgers equation. Phys Lett A 351:238
18. Eyink GL, Sreenivasan KR (2006) Onsager and the theory of hydrodynamic turbulence. Rev Mod Phys 78:87
19. Falkovich G, Gawedzki K, Vergassola M (2001) Particles and fields in fluid turbulence. Rev Mod Phys 73:797
20. Foias C, Rosa R, Manley O, Temam R (2001) Navier–Stokes Equation and Turbulence. Cambridge University Press, Cambridge
21. Friedrich R (2003) Statistics of Lagrangian velocities in turbulent flows. Phys Rev Lett 90:084501
22. Friedrich R, Peinke J (1997) Description of a turbulent cascade by a Fokker–Planck equation. Phys Rev Lett 78:863
23. Friedrich R, Peinke J (1997) Statistical properties of a turbulent cascade. Phys D 102:147
24. Friedrich R, Zeller J, Peinke J (1998) A Note in Three Point Statistics of Velocity Increments in Turbulence. Europhys Lett 41:153
25. Friedrich R, Peinke J, Tabar MRR (2008) Importance of Fluctuations: Complexity in the View of Stochastic Processes within this issue. Springer, Berlin
26. Frisch U (1995) Turbulence. The legacy of Kolmogorov. Cambridge University Press, Cambridge

27. Fursikov AV (1999) The closure problem of the Friedmann–Keller infinite chain of moment equations, corresponding to the Navier–Stokes system. In: Gyr A, Kinzelbach W, Tsinober A (eds) Fundamental problematic issues in turbulence. Birkhäuser, Basel
28. Galanti B, Gibbon JD, Heritage M (1997) Vorticity alignment results for the three-dimensional Euler and Navier–Stokes equations. Nonlinearity 10:1675
29. Gilbert AD (1993) A cascade interpretation of Lundgrens stretched spiral vortex model for turbulent fine structure. Phys Fluids A5:2831
30. Grafke T, Homann H, Dreher J, Grauer R (2008) Numerical simulations of possible finite time singularities in the incompressible Euler equations: comparison of numerical methods. Phys D 237:1932
31. Grassi V, Leo R a, Soliani G, Tempesta P (2000) Vortices and invariant surfaces generated by symmetries for the 3D Navier–Stokes equations. Physica A 286:79
32. Grauer R, Marliani C, Germaschewski K (1998) Adaptive mesh refinement for singular solutions of the incompressible Euler equations. Phys Rev Lett 84:4850
33. Gyr A, Kinzelbach W, Tsinober A (1999) Fundamental problematic issues in turbulence. Birkhäuser, Basel
34. Haken H (1983) Synergetics, An Introduction. Springer, Berlin
35. Haken H (1987) Advanced Synergetics. Springer, Berlin
36. Hamlington PE, Schumacher J, Dahm WJA (2008) Local and nonlocal strain rate fields and vorticity alignment in turbulent flows. Phys Rev E 77:026303
37. Hatakeyama N, Kambe T (1997) Statistical laws of random strained vortices in turbulence. Phys Rev Lett 79:1257
38. Hopf E (1957) Statistical hydromechanics and functional calculus. J Rat Mech Anal 1:87
39. Jimenez J, Wray AA, Saffman PG, Rogallo RS (1993) The structure of intense vorticity in isotropic turbulence. J Fluid Mech 255:65
40. Jovanovic J (2004) The statistical dynamics of turbulence. Springer, Berlin
41. Kirchhoff G (1883) Vorlesungen über mathematische Physik, vol 1, 3rd edn. Teubner, Leipzig
42. Kolmogorov AN (1941) Dissipation of energy in locally isotropic turbulence. Dokl Akad Nauk SSSR 32:19
43. Kolmogorov AN (1941) The local structure of turbulence in incompressible viscous fluid for very large Reynold's numbers. Dokl Akad Nauk SSSR 30:301
44. Kolmogorov AN (1962) A refinement of previous hypotheses concerning the local structure of turbulence in a viscous incompressible fluid at high Reynolds number. J Fluid Mech 13:82
45. L'vov V, Procaccia I Hydrodynamic Turbulence: a 19th Century Problem with a Challenge for the 21st Centurs. arXiv:chao-dyn 96006015
46. La Porta A, Voth G, Crawford AM, Alexander J, Bodenschatz E (2000) Fluid particle acceleration in fully developped turbulence. Nature 409:1017; La Porta A, Voth G, Moisy F, Bodenschatz E (2001) Using cavitation to measure statistics of low-pressure events in large Reynolds-number turbulence. Phys Fluids 12:1485
47. Landau L, Lifshitz EM (1981) Lehrbuch der Theoretischen Physik, Bd. 6, Hydrodynamik. Akademie, Berlin
48. Langner M, Peinke J, Rauh A () Langevin analysis with application to a Rayleigh–Benard convection experiment. Int J Nonlinear Dyn Syst Chaos (in press)
49. Lesieur M (1997) Turbulence in fluids. Kluwer Academic Publishers, Dordrecht
50. Lesieur M, Yaglom A, David F (2000) New trends in turbulence, LesHouches Summer School. Springer, Berlin
51. Lück S, Renner C, Peinke J, Friedrich R (2006) The Markov–Einstein coherence length – a new meaning for the Taylor length in turbulence. Phys Lett A 359:335
52. Lundgren TS (1969) Distribution functions in the statistical theory of turbulence. Phys Fluids 10:969
53. Lundgren TS (1982) Strained spiral vortex model for turbulent fine structure. Phys Fluids 25:2193
54. Lundgren TS (1993) A small-scale turbulence model. Phys Fluids A5:1472
55. Manneville P (2004) Instabilities, chaos and turbulence. Imperial College Press, London
56. Marchioro C, Pulvirenti M (1984) Vortex methods in two-dimensional fluid dynamics. Lecture Notes in Physics. Springer, Berlin
57. Marcq P, Naert AA (2001) Langevin equation for turbulent velocity increments. Phys Fluids 13:2590
58. Martin PC, Siggia ED, Rose HA (1973) Statistical dynamics of classical systems. Phys Rev A8:423
59. McComb WD (1990) The physics of fluid turbulence. Clarendon Press, Oxford
60. Moffat HK, Kida S, Ohkitani K (1994) Stretched vortices—the sinews of turbulence. J Fluid Mech 259:231
61. Monin AS, Yaglom AM (1971) Statistical Fluid Mechanics: Mechanics of Turbulence, vol 1. MIT Press, Cambridge, MA
62. Monin AS, Yaglom AM (1975) Statistical Fluid Mechanics: Mechanics of Turbulence, vol 2. MIT Press, Cambridge, MA
63. Mordant N, Metz P, Michel O, Pinton J-F (2001) Measurement of Lagrangian velocity in fully developed turbulence. Phys Rev Lett 87:214501
64. Newton PK (2001) The N-Vortex problem. Springer, New York
65. Oberlack M (2000) Symmetrie, Invarianz und Selbstähnlichkeit in der Turbulenz. Shaker, Aachen
66. Oberlack M, Busse FH (2002) Theories of turbulence. Springer, Wien
67. Obukhov AM (1959) Description of turbulence in terms of Lagrangian variables. Adv Geophys 6:113
68. Onsager L (1949) Statistical hydrodynamics. Suppl Nuovo Cimento 6:279
69. Ott S, Mann J (2000) An experimental investigation of the relative diffusion of particle pairs in three-dimensional turbulent flow. J Fluid Mech 402:207
70. Ottino JM (1989) The Kinematics of Mixing: Stretching, Chaos, and Transport. Cambridge University Press, Cambridge
71. Piquet J (1999) Turbulent flows, models and physics. Springer, Berlin
72. Pope SB (1985) Lagrangian PDF methods for turbulent flows. Annu Rev Fluid Mech 26:23; Pope SB (1994) Pdf methods for turbulent reactive flows. Prog Energy Combust Sci 11:119
73. Pope SB (2000) Turbulent flows. Cambridge University Press, Cambridge
74. Renner C, Peinke J, Friedrich R (2001) Experimental indications for Markov properties of small scale turbulence. J Fluid Mech 433:383

75. Renner C, Peinke J, Friedrich R, Chanal O, Chabaud B (2002) Universality of small scale turbulence. Phys Rev Lett 89:124502
76. Reynolds O (1883) Philos Trans R Soc Lond 174:935
77. Richardson LF (1926) Atmospheric diffusion shown on a distance-neighbour graph. Proc R Soc Lond A 110:709
78. Siefert M, Peinke J (2004) Different cascade speeds for longitudinal and transverse velocity increments of small-scale turbulence. Phys Rev E 70:015302R
79. Siefert M, Peinke J (2006) Joint multi-scale statistics of longitudinal and transversal increments in small-scale wake turbulence. J Turbul 7:1
80. Sreenivasan KR, Antonia RA (1997) The phenomenology of small-scale turbulence. Annu Rev Fluid Mech 29:435–472
81. Sturman R, Ottino JM, Wiggins S (2006) The mathematical foundations of mixing. In: Cambridge Monographs of Applied and Computational Mathematics, No 22
82. Taylor GI (1921) Diffusion by continuous movement. Proc Lond Math Soc Ser 20(2):196
83. Temam R (2007) Infinite dimensional dynamical systems in mechanics and physics. Springer, Heidelberg
84. Townsend AA (1951) On the fine structure of turbulence. Proc R Soc Lond A 208:534
85. Tutkun M, Mydlarski L (2004) Markovian properties of passive scalar increments in grid-generated turbulence. New J Phys 6:49
86. Ulinich FR, Ya Lyubimov B (1969) Statistical theory of turbulence of an incompressible fluid at large Reynolds numbers. Zh Exper Teor Fiz 55:951
87. Wilczek M, Jenko F, Friedrich R Lagrangian particle statistics in turbulent flows from a simple vortex model. Phys Rev E (to appear)
88. Calculation performed by Wilczek M (Münster)
89. Wyld HW (1961) Formulation of the theory of turbulence in incompressible fluids. Ann Phys 14:143
90. http://www.claymath.org/millenium/

Food Webs

Jennifer A. Dunne[1,2]
[1] Santa Fe Institute, Santa Fe, USA
[2] Pacific Ecoinformatics and Computational Ecology Lab, Berkeley, USA

Article Outline

Glossary
Definition of the Subject
Introduction: Food Web Concepts and Data
Early Food Web Structure Research
Food Web Properties
Food Webs Compared to Other Networks
Models of Food Web Structure
Structural Robustness of Food Webs
Food Web Dynamics
Ecological Networks
Future Directions
Bibliography

Glossary

Connectance(C) The proportion of possible links in a food web that actually occur. There are many algorithms for calculating connectance. The simplest and most widely used algorithm, sometimes referred to as "directed connectance," is links per species^2 (L/S^2), where S^2 represents all possible directed feeding interactions among S species, and L is the total number of actual feeding links. Connectance ranges from ~ 0.03 to 0.3 in food webs, with a mean of ~ 0.10 to 0.15.

Consumer-resource interactions A generic way of referring to a wide variety of feeding interactions, such as predator-prey, herbivore-plant or parasite-host interactions. Similarly, "consumer" refers generically to anything that consumes or preys on something else, and "resource" refers to anything that is consumed or preyed upon. Many taxa are both consumers and resources within a particular food web.

Food web The network of feeding interactions among diverse co-occurring species in a particular habitat.

Trophic species (S) Defined within the context of a particular food web, a trophic species is comprised of a set of taxa that share the same set of consumers and resources. A particular trophic species is represented by a single node in the network, and that node is topologically distinct from all other nodes. "Trophic species" is a convention introduced to minimize bias due to uneven resolution in food web data and to focus analysis and modeling on functionally distinct network components. S is used to denote the number of trophic species in a food web. The terms "trophic species," "species," and "taxa" will be used somewhat interchangeably throughout this article to refer to nodes in a food web. "Original species" will be used specifically to denote the taxa found in the original dataset, prior to trophic species aggregation.

Definition of the Subject

Food webs refer to the networks of feeding ("trophic") interactions among species that co-occur within particular habitats. Research on food webs is one of the few subdisciplines within ecology that seeks to quantify and analyze direct and indirect interactions among diverse species, rather than focusing on particular types of taxa. Food webs ideally represent whole communities including plants, bacteria, fungi, invertebrates and vertebrates. Feeding links represent transfers of biomass and encompass

a variety of trophic strategies including detritivory, herbivory, predation, cannibalism and parasitism. At the base of every food web are one or more types of autotrophs, organisms such as plants or chemoautotrophic bacteria, which produce complex organic compounds from an external energy source (e. g., light) and simple inorganic carbon molecules (e. g., CO_2). Food webs also have a detrital component—non-living particulate organic material that comes from the body tissues of organisms. Feeding-mediated transfers of organic material, which ultimately trace back to autotrophs or detritus via food chains of varying lengths, provide the energy, organic carbon and nutrients necessary to fuel metabolism in all other organisms, referred to as heterotrophs.

While food webs have been a topic of interest in ecology for many decades, some aspects of contemporary food web research fall within the scope of the broader cross-disciplinary research agenda focused on complex, "real-world" networks, both biotic and abiotic [2,83,101]. Using the language of graph theory and the framework of network analysis, species are represented by vertices (nodes) and feeding links are represented by edges (links) between vertices. As with any other network, the structure and dynamics of food webs can be quantified, analyzed and modeled. Links in food webs are generally considered directed, since biomass flows from a resource species to a consumer species ($A \rightarrow B$). However, trophic links are sometimes treated as undirected, since any given trophic interaction alters the population and biomass dynamics of both the consumer and resource species ($A \leftrightarrow B$). The types of questions explored in food web research range from "Do food webs from different habitats display universal topological characteristics, and how does their structure compare to that of other types of networks?" to "What factors promote different aspects of stability of complex food webs and their components given internal dynamics and external perturbations?" Two fundamental measures used to characterize food webs are S, the number of species or nodes in a web, and C, connectance—the proportion of possible feeding links that are actually realized in a web ($C = L/S^2$, where L is the number of observed directed feeding links, and S^2 is the number of possible directed feeding interactions among S taxa).

This article focuses on research that falls at the intersection of food webs and complex networks, with an emphasis on network structure augmented by a brief discussion of dynamics. This is a subset of a wide variety of ecological research that has been conducted on feeding interactions and food webs. Refer to the "Books and Reviews" in the bibliography for more information about a broader range of research related to food webs.

Introduction: Food Web Concepts and Data

The concept of food chains (e. g., grass is eaten by grasshoppers which are eaten by mice which are eaten by owls; $A \rightarrow B \rightarrow C \rightarrow D$) goes back at least several hundred years, as evidenced by two terrestrial and aquatic food chains briefly described by Carl Linnaeus in 1749 [42]. The earliest description of a food web may be the mostly detrital-based feeding interactions observed by Charles Darwin in 1832 on the island of St. Paul, which had only two bird species (Darwin 1939, as reported by Egerton [42]):

> By the side of many of these *[tern]* nests a small flying-fish was placed; which, I suppose, had been brought by the male bird for its partner … quickly a large and active crab (Craspus), which inhabits the crevices of the rock, stole the fish from the side of the nest, as soon as we had disturbed the birds. Not a single plant, not even a lichen, grows on this island; yet it is inhabited by several insects and spiders. The following list completes, I believe, the terrestrial fauna: a species of Feronia and an acarus, which must have come here as parasites on the birds; a small brown moth, belonging to a genus that feeds on feathers; a staphylinus (Quedius) and a woodlouse from beneath the dung; and lastly, numerous spiders, which I suppose prey on these small attendants on, and scavengers of the waterfowl.

The earliest known diagrams of generalized food chains and food webs appeared in the late 1800s, and diagrams of specific food webs, began appearing in the early 1900s, for example the network of insect predators and parasites on cotton-feeding weevils ("the boll weevil complex," [87]). By the late 1920s, diagrams and descriptions of terrestrial and marine food webs were becoming more common (e. g., Fig. 1 from [103], see also [48,104]). Charles Elton introduced the terms "food chain" and "food cycle" in his classic early textbook, *Animal Ecology* [43]. By the time Eugene Odum published a later classic textbook, *Fundamentals of Ecology* [84], the term "food web" was starting to replace "food cycle."

From the 1920s to the 1980s, dozens of system-specific food web diagrams and descriptions were published, as well as some webs that were more stylized (e. g., [60]) and that quantified link flows or species biomasses. In 1977, Joel Cohen published the first comparative studies of empirical food web network structure using up to 30 food webs collected from the literature [23,24]. To standardize the data, he transformed the diagrams and descriptions of webs in the literature into binary matrices with m rows and n columns [24]. Each column is headed by the number of

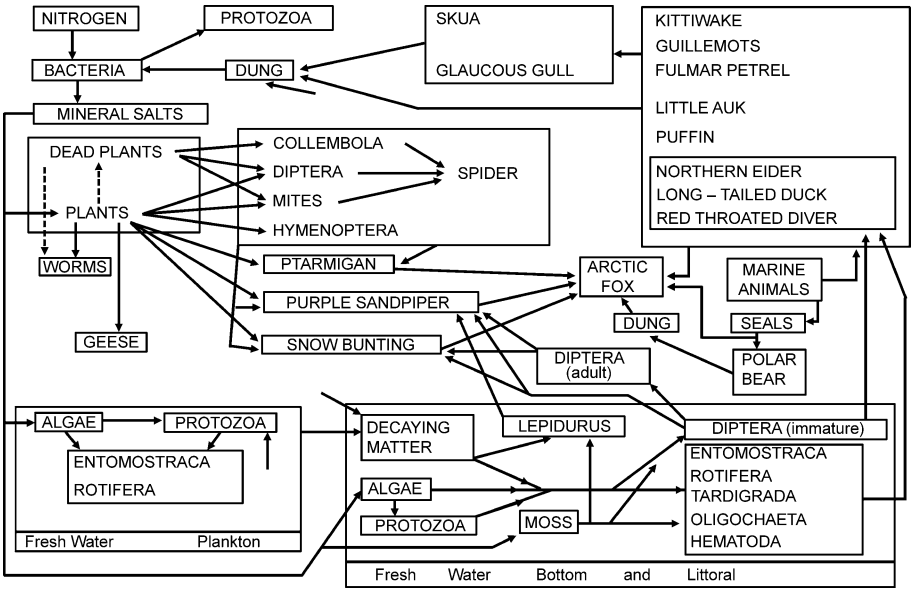

Food Webs, Figure 1
A diagram of a terrestrial Arctic food web, with a focus on nitrogen cycling, for Bear Island, published in 1923 [103]

one of the consumer taxa in a particular web, and each row is headed by the number of one of the resource taxa for that web. If w_{ij} represents the entry in the ith row and the jth column, it equals 1 if consumer j eats resource i or 0 if j does not eat i. This matrix-based representation of data is still often used, particularly in a full S by S format (where S is the number of taxa in the web), but for larger datasets a compressed two- or three-column notation for observed links is more efficient (Fig. 2).

By the mid-1980s, those 30 initial webs had expanded into a 113-web catalog [30] which included webs mostly culled from the literature, dating back to the 1923 Bear Island food web ([103], Fig. 1). However, it was apparent that there were many problems with the data. Most of the 113 food webs had very low diversity compared to the biodiversity known to be present in ecosystems, with a range of only 5 to 48 species in the original datasets and 3 to 48 trophic species. This low diversity was largely due to very uneven resolution and inclusion of taxa in most of these webs. The webs were put together in many different ways and for various purposes that did not include comparative, quantitative research. Many types of organisms were aggregated, underrepresented, or missing altogether, and in a few cases animal taxa had no food chains connecting them to basal species. In addition, cannibalistic links were purged when the webs were compiled into the 113-web catalog. To many ecologists, these food webs looked like little more than idiosyncratic cartoons of much richer and more complex species interactions found in natural systems, and they appeared to be an extremely unsound foundation on which to build understanding and theory [86,92].

Another catalog of "small" webs emerged in the late 1980s, a set of 60 insect-dominated webs with 2 to 87 original species (mean = 22) and 2 to 54 trophic species (mean = 12) [102]. Unlike the 113-web catalog, these webs are highly taxonomically resolved, mostly to the species level. However, they are still small due to their focus, in most cases, on insect interactions in ephemeral microhabitats such as phytotelmata (i. e., plant-held aquatic systems such as water in tree holes or pitcher plants) and singular detrital sources (e. g., dung paddies, rotting logs, animal carcasses). Thus, while the 113-web catalog presented food webs for communities at fairly broad temporal and spatial scales, but with low and uneven resolution, the 60-web catalog presented highly resolved but very small spatial and temporal slices of broader communities. These two very different catalogs were compiled into ECOWeB, the "Ecologists Co-Operative Web Bank," a machine readable database of food webs that was made available by Joel Cohen in 1989 [26]. The two catalogs, both separately and together as ECOWeB, were used in many studies of regularities in food web network structure, as discussed in the next Sect. "Early Food Web Structure Research".

A new level of detail, resolution and comprehensiveness in whole-community food web characterization was presented in two seminal papers in 1991. Gary Polis [92] published an enormous array of data for taxa found in the

a

	3	4	5	6
1	1	1	0	0
2	1	1	0	1
3	1	0	1	0
4	0	0	0	1
5	0	0	1	1
6	0	0	1	1

b

	1	2	3	4	5	6
1	0	0	1	1	0	0
2	0	0	1	1	0	1
3	0	0	1	0	1	0
4	0	0	0	0	0	1
5	0	0	0	0	1	1
6	0	0	0	0	1	1

c

3	1
3	2
3	3
4	1
4	2
5	3
5	5
5	6
6	2
6	4
6	5
6	6

d

3	1	3
4	1	2
5	3	
5	5	6
6	2	
6	4	6

Food Webs, Figure 2

Examples of formats for standardized notation of binary food web data. A hypothetical web with 6 taxa and 12 links is used. *Numbers* 1–6 correspond to the different taxa. a Partial matrix format: the 1s or 0s inside the matrix denote the presence or absence of a feeding link between a consumer (whose numbers 3–6 head columns) and a resource (whose numbers 1–6 head rows); b Full matrix format: similar to a, but all 6 taxa are listed at the heads of columns and rows; c Two-column format: a consumer's number appears in the *first column*, and one of its resource's numbers appears in the *second column*; d Three-column format: similar to c, but where there is a third number, the second and third numbers refer to a range of resource taxa. In this hypothetical web, taxa numbers 1 and 2 are basal taxa (i. e., taxa that do not feed on other taxa—autotrophs or detritus), and taxa numbers 3, 5, and 6 have cannibalistic links to themselves

Coachella Valley desert (California). Over two decades, he collected taxonomic and trophic information on at least 174 vascular plant species, 138 vertebrate species, 55 spider species, thousands of insect species including parasitoids, and unknown numbers of microorganisms, acari, and nematodes. He did not compile a complete food web including all of that information, but instead reported a number of detailed subwebs (e. g., a soil web, a scorpion-focused web, a carnivore web, etc.), each of which was more diverse than most of the ECOWeB webs. On the basis of the subwebs and a simplified, aggregated 30-taxa web of the whole community, he concluded that "*… most catalogued webs are oversimplified caricatures of actual communities …* [they are]*grossly incomplete representations of communities in terms of both diversity and trophic connections.*"

At about the same time, Neo Martinez [63] published a detailed food web for Little Rock Lake (Wisconsin) that he compiled explicitly to test food web theory and patterns (see Sect. "Early Food Web Structure Research"). By piecing together diversity and trophic information from multiple investigators actively studying various types of taxa in the lake, he was able to put together a relatively complete and highly resolved food web of 182 taxa, most identified to the genus, species, or ontogentic life-stage level, including fishes, copepods, cladocera, rotifers, diptera and other insects, mollusks, worms, porifera, algae, and cyanobacteria. In later publications, Martinez modified the original dataset slightly into one with 181 taxa. The 181 taxa web aggregates into a 92 trophic-species web, with nearly 1000 links among the taxa (Fig. 3). This dataset, and the accompanying analysis, set a new standard for food web empiricism and analysis. It still stands as the best whole-community food web compiled, in terms of even, detailed, comprehensive resolution.

Since 2000, the use of the ECOWeB database for comparative analysis and modeling has mostly given way to a focus on a smaller set of more recently published food webs [10,37,39,99,110]. These webs, available through www.foodwebs.org or from individual researchers, are compiled for particular, broad-scale habitats such as St. Mark's Estuary [22], Little Rock Lake [63], the island of St. Martin [46], and the Northeast U.S. Marine Shelf [61]. Most of the food webs used in contemporary comparative research are still problematic—while they generally are more diverse and/or evenly resolved than the earlier webs, most could still be resolved more highly and evenly. Among several issues, organisms such as parasites are usually left out (but see [51,59,67,74]), microorganisms are either missing or highly aggregated, and there is still a tendency to resolve vertebrates more highly than lower level organisms. An important part of future food web research is the compilation of more inclusive, evenly resolved, and well-defined datasets. Meanwhile, the careful selection and justification of datasets to analyze is an important part of current research that all too often is ignored.

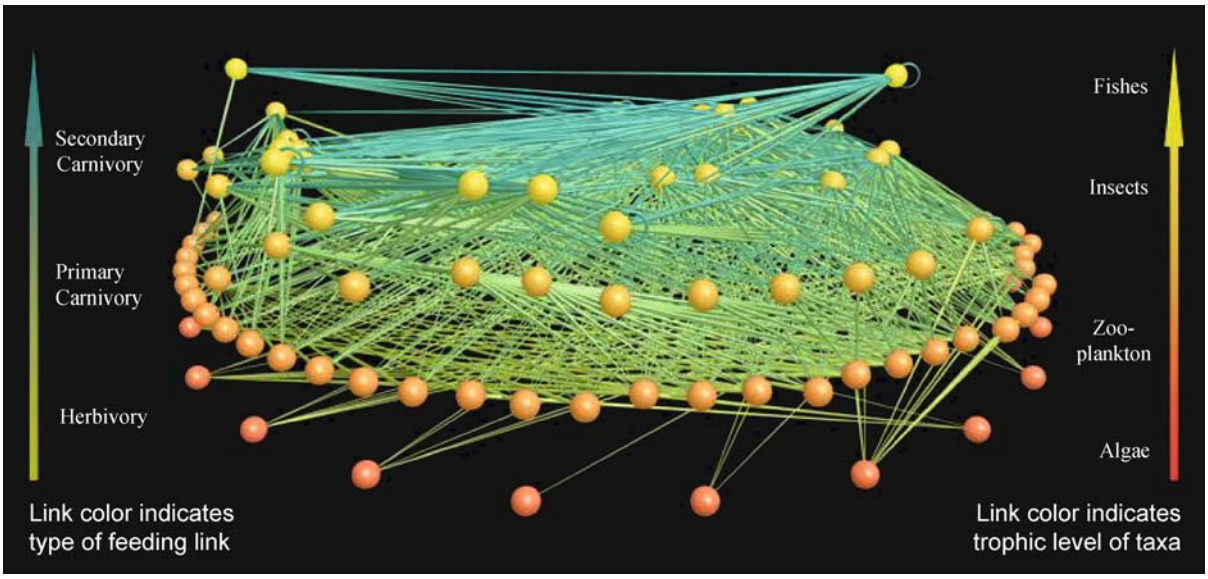

Food Webs, Figure 3
Food web of Little Rock Lake, Wisconsin [63]. 997 feeding links among 92 trophic species are shown. Image produced with FoodWeb3D, written by R.J. Williams, available at the Pacific Ecoinformatics and Computational Ecology Lab (www.foodwebs.org)

How exactly are food web data collected? In general, the approach is to compile as complete a species list as possible for a site, and then to determine the diets of each species present at that site. However, researchers have taken a number of different approaches to compiling food webs. In some cases, researchers base their food webs on observations they make themselves in the field. For example, ecologists in New Zealand have characterized the structure of stream food webs by taking samples from particular patches in the streams, identifying the species present in those samples, taking several individuals of each species present, and identifying their diets through gut-content analysis [106]. In other cases, researchers compile food web data by consulting with experts and conducting literature searches. For example, Martinez [63] compiled the Little Rock Lake (WI) food web by drawing on the expertise of more than a dozen biologists who were specialists on various types of taxa and who had been working at Little Rock Lake for many years. Combinations of these two approaches can also come into play—for example, a researcher might compile a relatively complete species list through field-based observations and sampling, and then assign trophic habits to those taxa through a combination of observation, consulting with experts, and searching the literature and online databases.

It is important to note that most of the webs used for comparative research can be considered "cumulative" webs. Contemporary food web data range from time- and space-averaged or "cumulative" (e. g., [63]) to more finely resolved in time (e. g., seasonal webs—[6]) and/or space (e. g., patch-scale webs—[106]; microhabitat webs—[94]). The generally implicit assumption underlying cumulative food web data is that the set of species in question co-exist within a habitat and individuals of those species have the opportunity over some span of time and space to interact directly. To the degree possible, such webs document who eats whom among all species within a macrohabitat, such as a lake or meadow, over multiple seasons or years, including interactions that are low frequency or represent a small proportion of consumption. Such cumulative webs are used widely for comparative research to look at whether there are regularities in food web structure across habitat (see Sect. "Food Webs Compared to Other Networks" and Sect. "Models of Food Web Structure"). More narrowly defined webs at finer scales of time or space, or that utilize strict evidence standards (e. g., recording links only through gut content sampling), have been useful for characterizing how such constraints influence perceived structure within habitats [105,106], but are not used as much to look for cross-system regularities in trophic network structure.

Early Food Web Structure Research

The earliest comparative studies of food web structure were published by Joel Cohen in 1977. Using data from the first 30-web catalog, one study focused on the ratio of predators to prey in food webs [23], and the other in-

vestigated whether food webs could be represented by single dimension interval graphs [24], a topic which continues to be of interest today (see Sect. "Food Webs Compared to Other Networks"). In both cases, he found regularities—(1) a ratio of prey to predators of $\sim 3/4$ regardless of the size of the web, and (2) most of the webs are interval, such that all species in a food web can be placed in a fixed order on a line such that each predator's set of prey forms a single contiguous segment of that line. The prey-predator ratio paper proved to be the first salvo in a quickly growing set of papers that suggested that a variety of food web properties were "scale-invariant." In its strong sense, scale invariance means that certain properties have constant values as the size (S) of food webs change. In its weak sense, scale-invariance refers to properties not changing systematically with changing S. Other scale-invariant patterns identified include constant proportions of top species (Top, species with no predators), intermediate species (Int, species with both predators and prey), and basal species (Bas, species with no prey), collectively called "species scaling laws" [12], and constant proportions of T-I, I-B, T-B, and I-I links between T, I, and B species, collectively called "link scaling laws" [27]. Other general properties of food webs were thought to include: food chains are short [31,43,50,89]; cycling/looping is rare (e. g., A $\rightarrow$ B $\rightarrow$ C $\rightarrow$ A; [28]); compartments, or subwebs with many internal links that have few links to other subwebs, are rare [91]; omnivory, or feeding at more than one trophic level, is uncommon [90]; and webs tend to be interval, with instances of intervality decreasing as S increases [24,29,116]. Most of these patterns were reported for the 113-web catalog [31], and some of the regularities were also documented in a subset the 60 insect-dominated webs [102].

Another, related prominent line of early comparative food web research was inspired by Bob May's work from the early 1970s showing that simple, abstract communities of interacting species will tend to transition sharply from local stability to instability as the complexity of the system increases—in particular as the number of species (S), the connectance (C) *or* the average interaction strength (i) increase beyond critical values [69,70]. He formalized this as a criterion that ecological communities near equilibrium will tend to be stable if $i(SC)^{1/2} < 1$. This mathematical analysis flew in the face of the intuition of many ecologists (e. g., [44,50,62,84]) who felt that increased complexity (in terms of greater numbers of species and links between them) in ecosystems gives rise to stability.

May's criterion and the general question of how diversity is maintained in communities provided a framework within which to analyze some readily accessible empirical data, namely the numbers of links and species in food webs. Assuming that average interaction strength (i) is constant, May's criterion suggests that communities can be stable given increasing diversity (S) as long as connectance (C) decreases. This can be empirically demonstrated using food web data in three similar ways, by showing that 1) C hyperbolically declines as S increases, so that the product SC remains constant, 2) the ratio of links to species (L/S), also referred to as link or linkage density, remains constant as S increases, or 3) L plotted as a function of S on a log-log graph, producing a power-law relation of the form $L = \alpha S^{\beta}$, displays an exponent of $\beta = 1$ (the slope of the regression) indicating a linear relationship between L and S. These relationships were demonstrated across food webs in a variety of studies (see detailed review in [36]), culminating with support from the 113-web catalog and the 60 insect-dominated web catalog. Cohen and colleagues identified the "link-species scaling law" of $L/S \approx 2$ using the 113 web catalog (i. e., there are two links per species on average in any given food web, regardless of its size) [28,30], and SC was reported as "roughly independent of species number" in a subset of the 60 insect-dominated webs [102].

However, these early conclusions about patterns of food web structure began to crumble with the advent of improved data and new analysis methods that focused on the issues of species aggregation, sampling effort, and sampling consistency [36]. Even before there was access to improved data, Tom Schoener [93] set the stage for critiques of the conventional paradigm in his Ecological Society of America MacArthur Award lecture, in which he explored the ramifications of a simple conceptual model based on notions of "generality" (what Schoener referred to as "generalization") and "vulnerability." He adopted the basic notion underlying the "link-species scaling law": that how many different taxa something can eat is constrained, which results in the number of resource taxa per consumer taxon (generality) holding relatively steady with increasing S. However, he further hypothesized that the ability of resource taxa to defend again consumers is also constrained, such that the number of consumer taxa per resource taxon (vulnerability) should increase with increasing S. A major consequence of this conceptual model is that total links per species (L/S, which includes links to resources and consumers) and most other food web properties should display scale-dependence, not scale-invariance. A statistical reanalysis of a subset of the 113-web catalog supported this contention as well as the basic assumptions of his conceptual model about generality and vulnerability.

Shortly thereafter, more comprehensive, detailed datasets, like the ones for Coachella Valley [92] and Little

Rock Lake [63], began to appear in the literature. These and other new datasets provided direct empirical counterpoints to many of the prevailing notions about food webs: their connectance and links per species were much higher than expected from the "link-species scaling law," food chains could be quite long, omnivory and cannibalism and looping could be quite frequent, etc. In addition, analyzes such as the one by Martinez [63], in which he systematically aggregated the Little Rock Lake food web taxa and links to generate small webs that looked like the earlier data, demonstrated that "*most published food web patterns appear to be artifacts of poorly resolved data*." Comparative studies incorporating newly available data further undermined the whole notion of "scale invariance" of most properties, particularly L/S(e. g., [65,66]).

For many researchers, the array of issues brought to light by the improved data and more sophisticated analyzes was enough for them to turn their back on structural food web research. A few hardy researchers sought to build new theory on top of the improved data. "Constant connectance" was suggested as an alternative hypothesis to constant L/S (the "link-species scaling law"), based on a comparative analysis of the relationship of L to S across a subset of available food webs including Little Rock Lake [64]. The mathematical difference between constant C and constant L/S can be simply stated using a log-log graph of links as a function of species (Fig. 4). If a power law exists of the form $L = \alpha S^{\beta}$, in the case of the link-species scaling law $\beta = 1$, which means that $L = \alpha S, L/S = \alpha$, indicating constant L/S. In the case of constant connectance, $\beta = 2$ and thus $L = \alpha S^2$, $L/S^2 = \alpha$, indicating constant C (L/S^2). Constant connectance means that L/S increases as a fixed proportion of S. One ecological interpretation of constant connectance is that consumers are likely to exploit an approximately constant fraction of available prey species, so as diversity increases, links per species increases [108].

Given the $L = \alpha S^{\beta}$ framework, $\beta = 2$ was reported for a set of 15 webs derived from an English pond [108], and $\beta = 1.9$ for a set of 50 Adirondack lakes [65], suggesting connectance may be constant across webs within a habitat or type of habitat. Across habitats, the picture is less clear. While $\beta = 2$ was reported for a small subset of the 5 most "credible" food webs then available from different habitats [64], several analyzes of both the old ECOWeB data and the more reliable newer data suggest that the exponent lies somewhere between 1 and 2, suggesting that C declines non-linearly with S (Fig. 4, [27,30,36,64,79,93]). For example, Schoener's reanalysis of the 113-web catalog suggested that $\beta = 1.5$, indicating that $L^{2/3}$ is proportional to S. A recent analysis of 19 recent trophic-species food webs with S of 25 to 172 also reported $\beta = 1.5$, with much scatter in the data (Fig. 4).

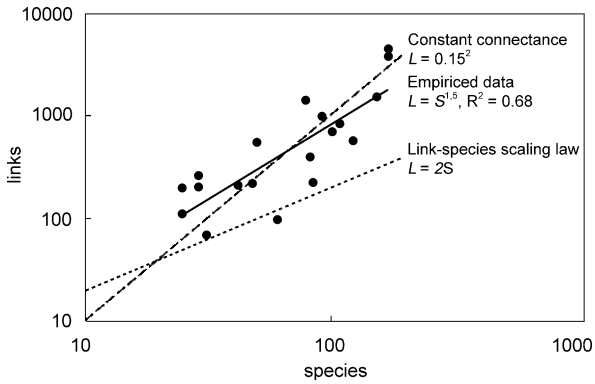

Food Webs, Figure 4
The relationship of links to species for 19 trophic-species food webs from a variety of habitats (*black circles*). The *solid line* shows the log-log regression for the empirical data, the *dashed line* shows the prediction for constant connectance, and the *dotted line* shows the prediction for the link-species scaling law (reproduced from [36], Fig. 1)

A recent analysis has provided a possible mechanistic basis for the observed constrained variation in C (~ 0.03 to 0.3 in cumulative community webs) as well as the scaling of C with S implied by β intermediate between 1 and 2 [10]. A simple diet breadth model based on optimal foraging theory predicts both of these patterns across food webs as an emergent consequence of individual foraging behavior of consumers. In particular, a contingency model of optimal foraging is used to predict mean diet breadth for S animal species in a food web, based on three parameters for an individual of species j: (1) net energy gain from consumption of an individual of species i, (2) the encounter rate of individuals of species i, and (3) the handling time spent attacking an individual of species i. This allows estimation of C for the animal portion of food webs, once data aggregation and cumulative sampling, well-known features of empirical datasets, are taken into account. The model does a good job of predicting values of C observed in empirical food webs and associated patterns of C across food webs.

Food Web Properties

Food webs have been characterized by a variety of properties or metrics, several of which have been mentioned previously (Sect. "Early Food Web Structure Research"). Many of these properties are quantifiable just using the basic network structure ("topology") of feeding interactions. These types of topological properties have been used to

evaluate simple models of food web structure (Sect. "Food Web Properties"). Any number of properties can be calculated on a given network—ecologists tend to focus on properties that are meaningful within the context of ecological research, although other properties such as path length (**Path**) and clustering coefficient (**Cl**) have been borrowed from network research [109]. Examples of several types of food web network structure properties, with common abbreviations and definitions, follow.

Fundamental Properties: These properties characterize very simple, overall attributes of food web network structure.

S: number of nodes in a food web
***L*:** number of links in a food web
***L*/*S*:** links per species
***C*, or *L*/*S*^{2}:** connectance, or the proportion of possible links that are realized

Types of Taxa: These properties characterize what proportion or percentage of taxa within a food web fall into particular topologically defined roles.

Bas: percentage of basal taxa (taxa without resources)
Int: percentage of intermediate taxa (taxa with both consumers and resources)
Top: percentage of top taxa (taxa with no consumers)
Herb: percentage of herbivores plus detritivores (taxa that feed on autotrophs or detritus)
Can: percentage of cannibals (taxa that feed on their own taxa)
Omn: percentage of omnivores (taxa that feed that feed on taxa at different trophic levels
Loop: percentage of taxa that are in loops, food chains in which a taxon occur twice (e. g., A → B → C → A)

Network Structure: These properties characterize other attributes of network structure, based on how links are distributed among taxa.

TL: trophic level averaged across taxa. Trophic level represents how many steps energy must take to get from an energy source to a taxon. Basal taxa have TL = 1, and obligate herbivores are TL = 2. TL can be calculated using many different algorithms that take into account multiple food chains that can connect higher level organisms to basal taxa (Williams and Martinez 2004).
ChLen: mean food chain length, averaged over all species
ChSD: standard deviation of ChLen
ChNum: log number of food chains
LinkSD: normalized standard deviation of links (# links per taxon)
GenSD: normalized standard deviation of generality (# resources per taxon)
VulSD: normalized standard deviation of vulnerability (# consumers per taxon)
MaxSim: mean across taxa of the maximum trophic similarity of each taxon to other taxa
Ddiet: the level of diet discontinuity—the proportion of triplets of taxa with an irreducible gap in feeding links over the number of possible triplets [19]—a local estimate of intervality
Cl: clustering coefficient (probability that two taxa linked to the same taxon are linked)
Path: characteristic path length, the mean shortest set of links (where links are treated as undirected) between species pairs

The previous properties (most of which are described in [110] and [39]) each provide a single metric that characterizes some aspect of food web structure. There are other properties, such as **Degree Distribution**, which are not single-number properties. "Degree" refers to the number of links that connect to a particular node, and the degree distribution of a network describes (in the format of a function or a graph) the total number of nodes in a network that have a given degree for each level of degree (Subsect. "Degree Distribution"). In food web analysis, **LinkSD**, **GenSD**, and **VulSD** characterize the variability of different aspects of degree distribution. Many food web structure properties are correlated with each other, and vary in predictable ways with *S* and/or *C*. This provides opportunities for topological modeling that are discussed below (Sect. "Models of Food Web Structure").

In addition to these types of metrics based on networks with unweighted links and nodes, it is possible to calculate a variety of metrics for food webs with nodes and/or links that are weighted by measures such as biomass, numerical abundance, frequency, interaction strength, or body size [11,33,67,81]. However, few food web datasets are "enriched" with such quantitative data and it remains to be seen whether such approaches are primarily a tool for richer description of particular ecosystems or whether they can give rise to novel generalities, models and predictions. One potential generality was suggested by a study of interaction strengths in seven soil food webs, where interaction strength reflects the size of the effects of species on each other's dynamics near equilibrium. Interaction strengths appear to be organized such that long loops contain many weak links, a pattern which enhances stability of complex food webs [81].

Food Webs Compared to Other Networks

Small-World Properties

How does the structure of food webs compare to that of other kinds of networks? One common way that various networks have been compared is in terms of whether they are "small-world" networks. Small-world networks are characterized by two of the properties described previously, characteristic path length (**Path**) and clustering coefficient (**Cl**) [109]. Most real-world networks appear to have high clustering, like what is seen on some types of regular spatial lattices (such as a planar triangular lattice, where many of a node's neighbors are neighbors of one another), but have short path lengths, like what is seen on "random graphs" (i. e., networks in which links are distributed randomly). Food webs do display short path lengths that are similar to what is seen in random webs (Table 1, [16,37,78,113]). On average, taxa are about two links from other taxa in a food web ("two degrees of separation"), and path length decreases with increasing connectance [113].

However, clustering tends to be quite low in many food webs, closer to the clustering expected on a random network (Table 1). This relatively low clustering in food webs appears consistent with their small size compared to most other kinds of networks studied, since the ratio of clustering in empirical versus comparable random networks increases linearly with the size of the network (Fig. 5).

Degree Distribution

In addition to small-world properties, many real-world networks appear to display power-law degree distributions [2]. Whereas regular graphs have the same number of links per node, and random graphs display a Poisson degree distribution, many empirical networks, both biotic and abiotic, display a highly skewed power-law ("scale-free") degree distribution, where most nodes have few links and a few nodes have many links. However, some empirical networks display less-skewed distributions such as exponential distributions [4]. Most empirical food webs display exponential or uniform degree distributions, not power-law distributions [16,37], and it has been suggested that normalized degree distributions in food webs follow universal functional forms [16] although there is a quite a bit of scatter when a wide range of data are considered (Fig. 6, [37]). Variable degree distributions, like what is seen in individual food webs, could result from simple mechanisms. For example, exponential and uniform food web degree distributions are generated by a model that combines (1) random immigration to local webs from a randomly linked regional set of taxa, and (2) random extinctions in the local webs [5]. The general lack of power-

Food Webs, Table 1

Topological properties of empirical and random food webs, listed in order of increasing connectance. *Path* refers to characteristic path length, and *Cl* refers to the clustering coefficient. $Path_r$ and Cl_r refer to the mean *D* and *Cl* for 100 random webs with the same *S* and *C*. Modified from [37] Table 1

Food Web	S	$C(L/S^2)$	L/S	$Path$	$Path_r$	Cl	Cl_r	Cl/Cl_r
Grassland	61	0.026	1.59	3.74	3.63	0.11	0.03	3.7
Scotch Broom	85	0.031	2.62	3.11	2.82	0.12	0.04	3.0
Ythan Estuary 1	124	0.038	4.67	2.34	2.39	0.15	0.04	3.8
Ythan Estuary 2	83	0.057	4.76	2.20	2.19	0.16	0.06	2.7
El Verde Rainforest	155	0.063	9.74	2.20	1.95	0.12	0.07	1.4
Canton Creek	102	0.067	6.83	2.27	2.01	0.02	0.07	0.3
Stony Stream	109	0.070	7.61	2.31	1.96	0.03	0.07	0.4
Chesapeake Bay	31	0.071	2.19	2.65	2.40	0.09	0.09	1.0
St. Marks Seagrass	48	0.096	4.60	2.04	1.94	0.14	0.11	1.3
St. Martin Island	42	0.116	4.88	1.88	1.85	0.14	0.13	1.1
Little Rock Lake	92	0.118	10.84	1.89	1.77	0.25	0.12	2.1
Lake Tahoe	172	0.131	22.59	1.81	1.74	0.14	0.13	1.1
Mirror Lake	172	0.146	25.13	1.76	1.72	0.14	0.15	0.9
Bridge Brook Lake	25	0.171	4.28	1.85	1.68	0.16	0.19	0.8
Coachella Valley	29	0.312	9.03	1.42	1.43	0.43	0.32	1.3
Skipwith Pond	25	0.315	7.88	1.33	1.41	0.33	0.33	1.0

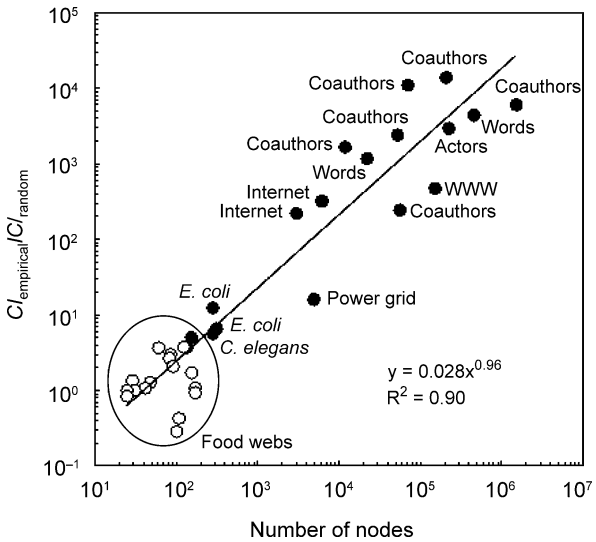

Food Webs, Figure 5
Trends in clustering coefficient across networks. The ratio of clustering in empirical networks ($Cl_{empirical}$) to clustering in random networks with the same number of nodes and links (Cl_{random}) is shown as a function of the size of the network (number of nodes). Reproduced from [37], Fig. 1

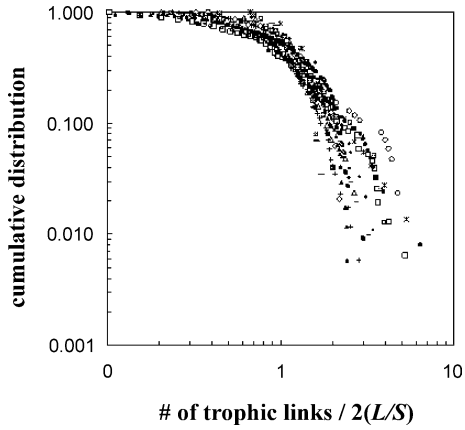

Food Webs, Figure 6
Log-log overlay plot of the cumulative distributions of links per species in 16 food webs. The link data are normalized by the average number of links/species in each web. If the distributions followed a power law, the data would tend to follow a straight line. Instead, they follow a roughly exponential shape. Reproduced from [37], Fig. 3

law degree distributions in food webs may result partly from the small size and large connectance of such networks, which limits the potential for highly skewed distributions. Many of the networks displaying power-law degree distributions are much larger and much more sparsely connected than food webs.

Other Properties

Assortative mixing, or the tendency of nodes with similar degree to be linked to each other, appears to be a pervasive phenomenon in a variety of social networks [82]. However, other kinds of networks, including technological and biological networks, tend to show disassortative mixing, where nodes with high degree tend to link to nodes with low degree. Biological networks, and particularly two food webs examined, show strong disassortativity [82]. Some of this may relate to a finite-size effect in systems like food webs that have limits on how many links are recorded between pairs of nodes. However, in food webs it may also result from the stabilizing effects of having feeding specialists linked to feeding generalists, as has been suggested for plant-animal pollination and frugivory (fruit-eating) networks ([7], Sect. "Ecological Networks").

Another aspect of structure that has been directly compared across several types of networks including food webs are "motifs," defined as "recurring, significant patterns of interconnections" [77]. A variety of networks (transcriptional gene regulation, neuron connectivity, food webs, two types of electronic circuits, the World Wide Web) were scanned for all possible subgraphs that could be constructed out of 3 or 4 nodes (13 and 199 possible subgraphs, respectively). Subgraphs that appeared significantly more often in empirical webs than in their randomized counterparts (i. e., networks with the same number of nodes and links, and the same degree for each node, but with links otherwise randomly distributed) were identified. For the seven food webs examined, there were two "consensus motifs" shared by most of the webs—a three-node food chain, and a four-species diamond where a predator has two prey, which in turn prey on the same species (Fig. 7). The four-node motif was shared by two other types of networks (neuron connectivity, one type of electronic circuit), and nothing shared the three-node chain. The WWW and food web networks appear most dissimilar to other types of networks (and to each other) in terms of significant motifs.

Complex networks can be decomposed into minimum spanning trees (MST). A MST is a simplified version of a network created by removing links to minimize the distance between nodes and some destination. For example, a food web can be turned into MST by adding an "environment" node that all basal taxa link to, tracing the shortest food chain from each species to the environment node, and removing links that do not appear in the shortest chains. Given this algorithm, a MST removes links that

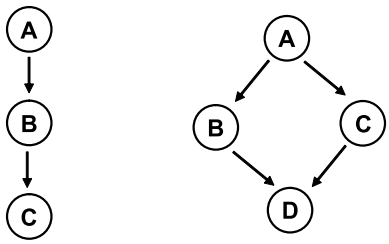

Food Webs, Figure 7
The two 3 or 4-node network motifs found to occur significantly more often than expected in most of seven food webs examined. There is one significant 3-node motif (out of 13 possible motifs), a food chain of the form A eats B eats C. There is one significant 4-node motif (out of 199 possible motifs), a trophic diamond ("bi-parallel") of the form A eats B and C, which both eat D

occur in loops and retains a basic backbone that has a tree-like structure. In a MST, the quantity A_i is defined as the number of nodes in a subtree rooted at node i, and can be regarded as the transportation rate through that node. C_i is defined as the integral of A_i (i. e., the sum of A_i for all nodes rooted at node i) and can be regarded as the transportation cost at node i. These properties can be used to plot C_i versus A_i for each node in a networks, or to plot whole-system C_o versus A_o across multiple networks, to identify whether scaling relationships of the form $C(A) \propto A^n$ are present, indicating self-similarity in the structure of the MST (see [18] for review). In a food web MST, the most efficient configuration is a star, where every species links directly to the environment node, resulting in an exponent of 1, and the least efficient configuration is a single chain, where resources have to pass through each species in a line, resulting in an exponent of 2. It has been suggested that food webs display a universal exponent of 1.13 [18,45], reflecting an invariant functional food web property relating to very efficient resource transportation within an ecosystem. However, analyzes based on a larger set of webs (17 webs versus 7) suggest that exponents for C_i as a function of A_i range from 1.09 to 1.26 and are thus not universal, that the exponents are quite sensitive to small changes in food web structure, and that the observed range of exponent values would be similarly constrained in any network with only 3 levels, as is seen in food web MSTs [15].

Models of Food Web Structure

An important area of research on food webs has been whether their observed structure, which often appears quite complex, can emerge from simple rules or models. As with other kinds of "real-world" networks, models that assign links among nodes randomly, according to fixed probabilities, fail to reproduce the network structure of empirically observed food webs [24,28,110]. Instead, several models that combine stochastic elements with simple link assignment rules have been proposed to generate and predict the network structure of empirical food webs.

The models share a basic formulation [110]. There are two empirically quantifiable parameters: (1) S, the number of trophic species in a food web, and (2) C, the connectance of a food web, defined as links per species squared, L/S^2. Thus, S specifies the number of nodes in a network, and C specifies the number of links in a network with S nodes. Each species is assigned a "niche value" n_i drawn randomly and uniformly from the interval [0,1]. The models differ in the rules used to distribute links among species. The link distribution rules follow in the order the models were introduced in the literature:

Cascade Model (Cohen and Newman [28]): Each species has the fixed probability $P = 2CS/(S-1)$ of consuming species with niche values less than its own. This creates a food web with hierarchical feeding, since it does not allow feeding on taxa with the same niche value (cannibalism) or taxa with higher niche values (looping/cycling). This formulation [110] is a modified version of the original cascade model that allows L/S, equivalent to the CS term in the probability statement above, to vary as a tunable parameter, rather than be fixed as a constant [28].

Niche Model (Williams and Martinez [110], Fig. 8): Each species consumes all species within a segment of the [0,1] interval whose size r_i is calculated using the feeding range width algorithm described below. The r_i's center c_i is set at a random value drawn uniformly from the interval $[r_i/2, n_i]$ or $[r_i/2, 1 - r_i/2]$ if $n_i > 1 - r_i/2$, which places c_i equal to or lower than the niche value n_i and keeps the r_i segment within [0,1]. The c_i rule relaxes the strict feeding hierarchy of the cascade model and allows for the possibility of cannibalism and looping. Also, the r_i rule ensures that species feed on a contiguous range of species, necessarily creating interval graphs (i. e., species can be lined up along a single interval such that all of their resource species are located in contiguous segments along the interval).

Feeding range width algorithm: The value of $r_i = xn_i$, where $0 < x < 1$ is randomly drawn from the probability density function $p(x) = \beta(1-x)^{b-1}$ (the beta distribution), where $\beta = (1/2C) - 1$ to obtain a C close to the desired C.

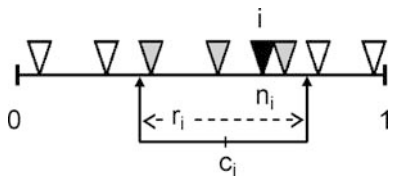

Food Webs, Figure 8
Graphical representation of the niche model: Species *i* feeds on 4 taxa including itself and one with a higher niche value

Nested-Hierarchy Model (Cattin et al. [19]): Like the niche model, the number of prey items for each species is drawn randomly from a beta distribution that constrains C close to a target value. Once the number of prey items for each species is set, those links are assigned in a multistep process. First, a link is randomly assigned from species i to a species j with a lower n_i. If j is fed on by other species, the next feeding links for i are selected randomly from the pool of prey species fed on by a set of consumer species defined as follows: they share at least one prey species and at least one of them feeds on j. If more links need to be distributed, they are then randomly assigned to species without predators and with niche values $< n_i$, and finally to those with niche value $\geq n_i$. These rules were chosen to relax the contiguity rule of the niche model and to allow for trophic habit overlap among taxa in a manner which the authors suggest evokes phylogenetic constraints.

Generalized Cascade Model (Stouffer et al. [99]): Species i feeds on species j if $n_j \leq n_i$ with a probability drawn from the interval [0,1] using the beta or an exponential distribution. This model combines the beta distribution introduced in the niche model with the hierarchical, non-contiguous feeding of the cascade model.

These models have been evaluated with respect to their relative fit to empirical data in a variety of ways. In a series of six papers published from 1985 to 1990 with the common title "A stochastic theory of community food webs," the cascade model was proposed as a means of explaining "*the phenomenology of observed food web structure, using a minimum of hypotheses*" [31]. This was not the first simple model proposed for generating food web structure [25,88,89,116], but it was the most well-developed model. Cohen and colleagues also examined several model variations, most of which performed poorly. While the cascade model appeared to generate structures that qualitatively fit general patterns in the data from the 113-web catalog, subsequent statistical analyzes suggested that the fit between the model and that early data was poor [93,96,97]. Once improved data began to emerge, it became clear that some of the basic assumptions built in to the cascade model, such as no looping and minimal overlap and no clustering of feeding habits across taxa, are violated by common features of multi-species interactions.

The niche model was introduced in 2000, along with a new approach to analysis: numerical simulations to compare statistically the ability of the niche model, the cascade model, and one type of random network model to fit empirical food web data [110]. Because of stochastic variation in how species and links are distributed in any particular model web, analysis begins with the generation of hundreds to thousands of model webs with the same S and similar C as an empirical food web of interest. Model webs that fall within 3% of the target C are retained. Model-generated webs occasionally contain species with no links to other species, or species that are trophically identical. Either those webs are thrown out, or those species are eliminated and replaced, until every model web has no disconnected or identical species. Also, each model web must contain at least one basal species. These requirements ensure that model webs can be sensibly comparable to empirical trophic-species webs.

Once a set of model webs are generated with the same S and C as an empirical web, model means and standard deviations are calculated for each food web property of interest, which can then be compared to empirical values. Raw error, the difference between the value of an empirical property and a model mean for that property, is normalized by dividing it by the standard deviation of the property's simulated distribution. This approach allows assessment not only of whether a model over- or under-estimates empirical properties as indicated by the raw error, but also to what degree a model's mean deviates from the empirical value. Normalized errors within ± 2 are considered to indicate a good fit between the model prediction and the empirical value. This approach has also been used to analyze network motifs [77] (Subsect. "Other Properties").

The initial niche model analyzes examined seven more recent, diverse food webs ($S = 24$ to 92) and up to 12 network structure properties for each web [110]. The random model (links are distributed randomly among nodes) performs poorly, with an average normalized error (ANE) of 27.1 (SD = 202). The cascade model performs better, with ANE of -3.0 (SD = 14.1). The niche model performs an order of magnitude better than the cascade model, with ANE of 0.22 (SD = 1.8). Only the niche model falls within ± 2 ANE considered to show a good fit to the data. Not surprisingly, there is variability in how all three models fit different food webs and properties. For example, the niche

model generally overestimates food-chain length. Specific mismatches are generally attributable either to limitations of the models or biases in the data [110]. A separate test of the niche and cascade models with three marine food webs, a type of habitat not included in the original analysis, obtained similar results [39]. These analyzes demonstrate that the structure of food webs is far from random and that simple link distribution rules can yield apparently complex network structure, similar to that observed in empirical data. In addition, the analyzes suggest that food webs from a variety of habitats share a fundamentally similar network structure, and that the structure is scale-dependent in predictable ways with S and C.

The nested-hierarchy model [19] and generalized cascade model [99], variants of the niche model, do not appear to improve on the niche model, and in fact may be worse at representing several aspects of empirical network structure. Although the nested-hierarchy model breaks the intervality of the niche model and uses a complicated-sounding set of link distribution rules to try to mimic phylogenetic constraints on trophic structure, it "*generates webs characterized by the same universal distributions of numbers of prey, predators, and links*" as the niche model [99]. Both the niche and nested-hierarchy models have a beta distribution at their core. The beta distribution is reasonably approximated by an exponential distribution for $C < 0.12$ [99], and thus reproduces the exponential degree distributions observed in many empirical webs, particularly those with average or less-than-average C [37]. The generalized cascade model was proposed as a simplified model that would return the same distributions of taxa and links as the niche and nested-hierarchy models. It is defined using only two criteria: (1) taxa form a totally ordered set—this is fulfilled by the arrangement of taxa along a single "niche" interval or dimension, and (2) each species has an exponentially decaying probability of preying on a given fraction of species with lower niche values [99].

Although the generalized cascade model does capture a central tendency of successful food web models, only some food web properties are derivable from link distributions (e. g., Top, Bas, Can, VulSD, GenSD, Clus). There are a variety of food web structure properties of interest that are not derivable from degree distributions (e. g., Loop, Omn, Herb, TL, food-chain statistics, intervality statistics). The accurate representation of these types of properties may depend on additional factors, for example the contiguous feeding ranges specified by the niche model but absent from the cascade, nested-hierarchy, and generalized cascade models. While it is known that empirical food webs are not interval, until recently it was not clear how non-interval they are. Intervality is a brittle property that is broken by a single gap in a single feeding range (i. e., a single missing link in a food web), and trying to arrange species in a food web into their most interval ordering is a computationally challenging problem. A more robust measure of intervality in food webs has been developed, in conjunction with the use of simulated annealing to estimate the most interval ordering of empirical food webs [100]. This analysis suggests that complex food webs "*do exhibit a strong bias toward contiguity of prey, that is, toward intervality*" when compared to several alternative "null" models, including the generalized cascade model. Thus, the intervality assumption of the niche model, initially critiqued as a flaw of the model [19], helps to produce a better fit to empirical data than the non-interval alternate models.

Structural Robustness of Food Webs

A series of papers have examined the response of a variety of networks including the Internet and WWW web pages [1] and metabolic and protein networks [52,53] to the simulated loss of nodes. In each case, the networks, all of which display highly skewed power-law degree distributions, appear very sensitive to the targeted loss of highly-connected nodes but relatively robust to random loss of nodes. When highly-connected nodes are removed from scale-free networks, the average path length tends to increase rapidly, and the networks quickly fragment into isolated clusters. In essence, paths of information flow in highly skewed networks are easily disrupted by loss of nodes that are directly connected to an unusually large number of other nodes. In contrast, random networks with much less skewed Poisson degree distributions display similar responses to targeted loss of highly-connected nodes versus random node loss [101].

Within ecology, species deletions on small ($S < 14$) hypothetical food web networks as well as a subset of the 113-web catalog have been used to examine the reliability of network flow, or the probability that sources (producers) are connected to sinks (consumers) in food webs [54]. The structure of the empirical webs appears to conform to reliable flow patterns identified using the hypothetical webs, but that result is based on low diversity, poorly resolved data. The use of more highly resolved data with node knock-out algorithms to simulate the loss of species allows assessment of potential secondary extinctions in complex empirical food webs. Secondary extinctions result when the removal of taxa results in one or more consumers losing all of their resource taxa. Even though most food webs do not have power-law degree distributions, they show similar patterns of robustness to other networks: re-

moval of highly-connected species results in much higher rates of secondary extinctions than random loss of species ([38,39,95], Fig. 9). In addition, loss of high-degree species results in more rapid fragmentation of the webs [95]. Protecting basal taxa from primary removal increases the robustness of the web (i. e., fewer secondary extinctions occur) ([38], Fig. 9). While removing species with few links generally results in few secondary extinctions, in a quarter of the food webs examined, removing low-degree species results in secondary extinctions comparable to or greater than what is seen with removal of high-degree species [38]. This tends to occur in webs with relatively high *C*.

Beyond differential impacts of various sequences of species loss in food webs, food web 'structural robustness' can be defined as the fraction of primary species loss that induces some level of species loss (primary + secondary extinctions) for a particular trophic-species web. Analysis of R_{50} (i. e., what proportion of species have to be removed to achieve $\geq$ 50% total species loss) across multiple food webs shows that robustness increases approximately logarithmically with increasing connectance ([38,39], Fig. 9, 10). In essence, from a topological perspective food webs with more densely interconnected taxa are better protected from species loss, since it takes greater species loss for consumers to lose all of their resources.

It is also potentially important from a conservation perspective to identify particular species likely to result in the greatest number of secondary extinctions through their loss. The loss of a particular highly-connected species may or may not result in secondary extinctions. One way to identify critical taxa is to reduce the topological structure of empirical food webs into linear pathways that define the essential chains of energy delivery in each web. A particular species can be said to "dominate" other

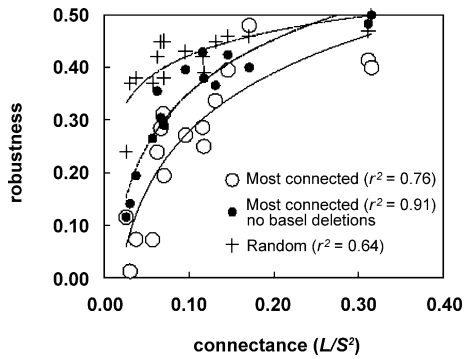

Food Webs, Figure 10

The proportion of primary species removals required to induce a total loss (primary removals plus secondary extinctions) of 50% of the species in each of 16 food webs ("robustness," see the shorter red line of Fig. 9 for visual representation) as a function of the connectance of each web. Logarithmic fits to the three data sets are shown, with a *solid line* for the most connected deletion order, a *long dashed line* for the most connected with basal species preserved deletion order, and a *short dashed* line for random deletion order. The maximum possible y value is 0.50. The equations for the fits are: $y = 0.162\ln(x) + 0.651$ for most connected species removals, $y = 0.148\ln(x) + 0.691$ for most connected species removals with basal species preserved, and $y = 0.067\ln(x) + 0.571$ for random species removals. Reproduced from [38], Fig. 2

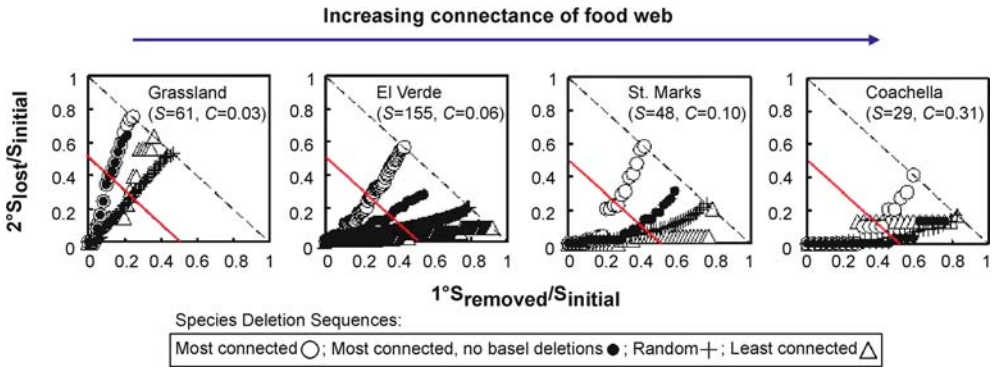

Food Webs, Figure 9

Secondary extinctions resulting from primary species loss in 4 food webs ordered by increasing connectance (*C*). The *y-axis* shows the cumulative secondary extinctions as a fraction of initial S, and the *x-axis* shows the primary removals of species as a fraction of initial *S*. 95% error bars for the random removals fall within the size of the symbols and are not shown. For the most connected (*circles*), least connected (*triangles*), and random removal (*plus symbols*) sequences, the data series end at the *black diagonal dashed line*, where primary removals plus secondary extinctions equal *S* and the web disappears. For the most connected species removals with basal species preserved (*black dots*), the data points end when only basal species remain. The shorter *red diagonal lines* show the points at which 50% of species are lost through combined primary removals and secondary extinctions ("robustness" or R_{50})

species if it passes energy to them along a chain in the dominator tree. The higher the number of species that a particular species dominates, the greater the secondary extinctions that may result from its removal [3]. This approach has the advantage of going beyond assessment of direct interactions to include indirect interactions.

As in food webs, the order of pollinator loss has an effect on potential plant extinction patterns in plant-pollinator networks [75] (see Sect. "Ecological Networks"). Loss of plant diversity associated with targeted removal of highly-connected pollinators is not as extreme as comparable secondary extinctions in food webs, which may be due to pollinator redundancy and the nested topology of those networks.

While the order in which species go locally extinct clearly affects the potential for secondary extinctions in ecosystems, the focus on high-degree, random, or even dominator species does not provide insight on ecologically plausible species loss scenarios, whether the focus is on human perturbations or natural dynamics. The issue of what realistic natural extinction sequences might look like has been explored using a set of pelagic-focused food webs for 50 Adirondack lakes [49] with up to 75 species [98]. The geographic nestedness of species composition across the lakes is used to derive an ecologically plausible extinction sequence scenario, with the most restricted taxa the most likely to go extinct. This sequence is corroborated by the pH tolerances of the species. Species removal simulations show that the food webs are highly robust in terms of secondary extinctions to the "realistic" extinction order and highly sensitive to the reverse order. This suggests that nested geographical distribution patterns coupled with local food web interaction patterns appear to buffer effects of likely species losses. This highlights important aspects of community organization that may help to minimize biodiversity loss in the face of a naturally changing environment. However, anthropogenic disturbances may disrupt the inherent buffering of how taxa are organized geographically and trophically, reducing the robustness of ecosystems.

Food Web Dynamics

Analysis of the topology of food webs has proven very useful for exploring basic patterns and generalities of "who eats whom" in ecosystems. This approach seeks to identify "*the most universal, high-level, persistent elements of organization*" [35] in trophic networks, and to leverage understanding of such organization for thinking about ecosystem robustness. However, food webs are inherently dynamical systems, since feeding interactions involve biomass flows among species whose "stocks" can be characterized by numbers of individuals and/or aggregate population biomass. All of these stocks and flows change through time in response to direct and indirect trophic and other types of interactions. Determining the interplay among network structure, network dynamics, and various aspects of stability such as persistence, robustness, and resilience in complex "real-world" networks is one of the great current challenges in network research [101]. It is particularly important in the study of ecosystems, since they face a variety of anthropogenic perturbations such as climate change, habitat loss, and invasions, and since humans depend on them for a variety of "ecosystem services" such as supply of clean water and pollination of crops [34].

Because it is nearly impossible to compile detailed, long-term empirical data for dynamics of more than two interacting species, most research on species interaction dynamics relies on analytical or simulation modeling. Most modeling studies of trophic dynamics have focused narrowly on predator-prey or parasite-host interactions. However, as the previous sections should make clear, in natural ecosystems such interaction dyads are embedded in diverse, complex networks, where many additional taxa and their direct and indirect interactions can play important roles for the stability of focal species as well as the stability or persistence of the broader community. Moving beyond the two-species population dynamics modeling paradigm, there is a strong tradition of research that looks at interactions among 3–8 species, exploring dynamics and simple variations in structure in slightly more complex systems (see reviews in [40,55]). However, these interaction modules still present a drastic simplification of the diversity and structure of natural ecosystems. Other dynamical approaches have focused on higher diversity model systems [69], but ignore network structure in order to conduct analytically tractable analyzes.

Researchers are increasingly integrating dynamics with complex food web structure in modeling studies that move beyond small modules. The Lotka–Volterra cascade model [20,21,32] was an early incarnation of this type of integration. As its name suggests, the Lotka–Volterra cascade model runs classic L-V dynamics, including a nonsaturating linear functional response, on sets of species interactions structured according to the cascade model [28]. The cascade model was also used to generate the structural framework for a dynamical food web model with a linear functional response [58] used to study the effects of prey-switching on ecosystem stability. Improving on aspects of biological realism of both dynamics and structure, a bioenergetic dynamical model with nonlinear functional responses [119] was used in conjunction with em-

pirically-defined food web structure among 29 species to simulate the biomass dynamics of a marine fisheries food web [117,118]. This type of nonlinear bioenergetic dynamical modeling approach has been integrated with niche model network structure and used to study more complex networks [13,14,68,112]. A variety of types of dynamics are observed in these non-linear models, including equilibrium, limit cycle, and chaotic dynamics, which may or may not be persistent over short or long time scales (Fig. 11). Other approaches model ecological and evolutionary dynamics to assemble species into networks, rather than imposing a particular structure on them. These models, which typically employ an enormous amount of parameters, are evaluated as to whether they generate plausible persistence, diversity, and network structure (see review by [72]). All of these approaches are generally used to examine stability, characterized in a diversity of ways, in ecosystems with complex structure and dynamics [71,85].

While it is basically impossible to empirically validate models of integrated structure and dynamics for complex ecological networks, in some situations it is possible to draw interesting connections between models and data at more aggregated levels. This provides opportunities to move beyond the merely heuristic role that such models generally play. For example, nonlinear bioenergetic models of population dynamics parametrized by biological rates allometrically scaled to populations' average body masses have been run on various types of model food web structures [14]. This approach has allowed the comparison of trends in two different measures of food web stability, and how they relate to consumer-resource body-size ratios and to initial network structure. One measure of stability is the fraction of original species that display persistent dynamics, i. e., what fraction of species do not go extinct in the model when it is run over many time steps ("species persistence"). Another measure of stability is how variable the densities of all of the persistent species are ("population stability")—greater variability across all the species indicates decreased stability in terms of population dynamics.

Brose and colleagues [14] ran the model using different hypothetical consumer-resource body-size ratios that range from 10^{-2} (consumers are 100 times smaller than their resources) to 10^5 (consumers are 100,000 times larger than their resources) (Fig. 12). Species persistence increases dramatically with increasing body-size ratios, until inflection points are reached at which persistence shifts to high levels ($\sim$ 0.80) of persistence (Fig. 12a). However, population stability decreases with increasing body-size ratios until inflection points are reached that show the lowest stability, and then increases again beyond those points (Fig. 12b). In both cases, the inflection points correspond to empirically observed consumer-resource body-size ratios, both for webs parametrized to represent invertebrate dominated webs, and for webs parametrized to represent ectotherm vertebrate dominated webs. Thus, across thousands of observations, invertebrate consumer-resource body size ratios are $\sim 10^1$, and ectotherm vertebrate consumer-resource body size ratios are $\sim 10^2$, which correspond to the model's inflection points for species persistence and population stability (Fig. 12). It is interesting

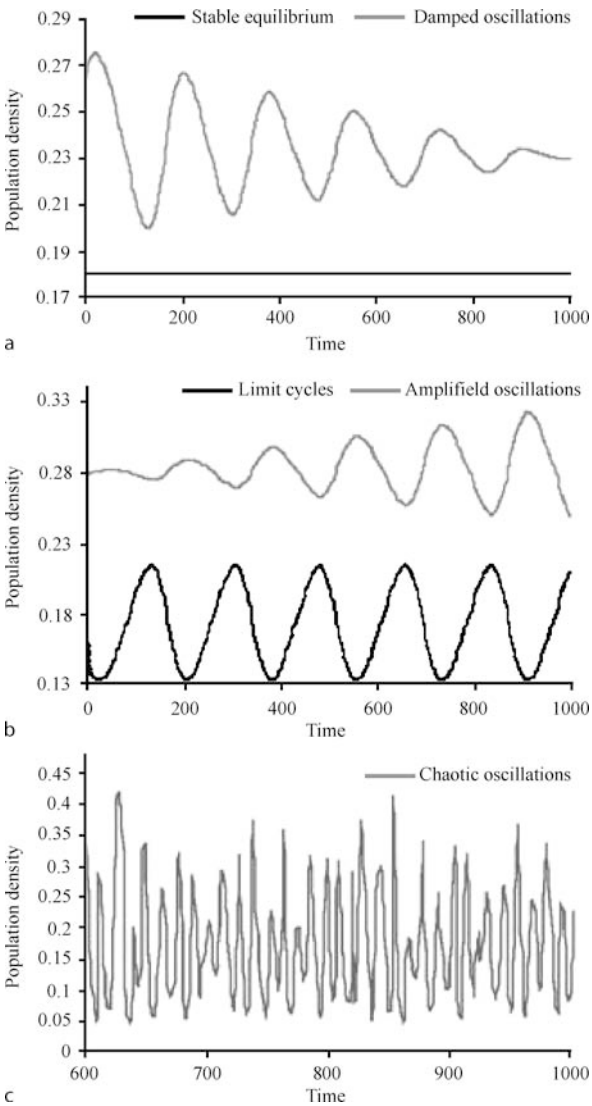

Food Webs, Figure 11

5 different types of population dynamics shown as time series of population density (from [40], Fig. 10.1). The types of dynamics shown include a stable equilibrium, damped oscillations, limit cycles, amplified oscillations, and chaotic oscillations

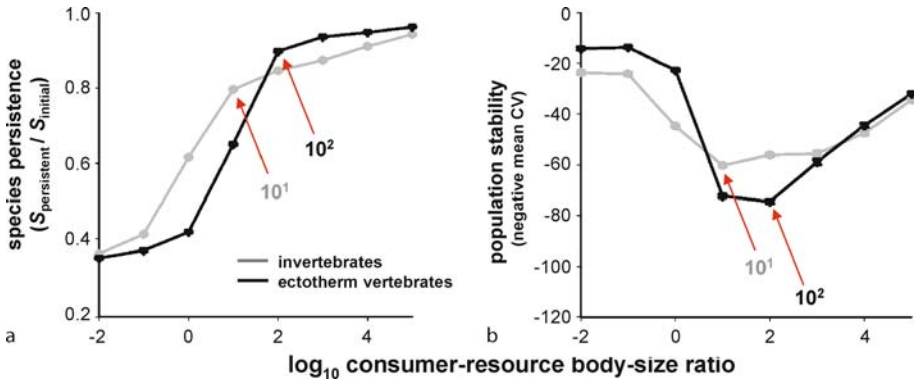

Food Webs, Figure 12

a shows the fraction of species that display persistent dynamics as a function of consumer-resource body-size ratios for model food webs parametrized for invertebrates (*gray line*) and ectotherm vertebrates (*black line*). The inflection points for shifts to high-persistence dynamics are indicated by *red arrows* for both curves, and those inflection points correspond to empirically observed consumer-resource body size ratios for invertebrate dominated webs (10^1—consumers are on average 10 times larger than their resources) and ectotherm vertebrate dominated webs (10^2—consumers are on average 100 times larger than their resources). b shows results for population stability, the mean of how variable species population biomasses are in persistent webs. In this case, the inflection points for shifts to low population stability are indicated by red arrows, and those inflection points also correspond to the empirically observed body-size ratios for consumers and resources. Figure adapted from [14]

to note that high species persistence is coupled to low population stability—i. e., an aspect of increased stability of the whole system (species persistence) is linked to an aspect of decreased stability of components of that system (population stability). It is also interesting to note that in this formulation, using initial cascade versus niche model structure had little impact on species persistence or population stability [14], although other formulations show increased persistence when dynamics are initiated with niche model versus other structures [68]. How structure influences dynamics, and vice-versa, is an open question.

Ecological Networks

This article has focused on food webs, which generally concern classic predator-herbivore-primary producer feeding interactions. However, the basic concept of food webs can be extended to a broader framework of "ecological networks" that is more inclusive of different components of ecosystem biomass flow, and that takes into consideration different kinds of species interactions that are not classic "predator-prey" interactions. Three examples are mentioned here. First, parasites have typically been given short shrift in traditional food webs, although exceptions exist (e. g., [51,67,74]). This is changing as it becomes clear that parasites are ubiquitous, often have significant impacts on predator-prey dynamics, and may be the dominant trophic habitat in most food webs, potentially altering our understanding of structure and dynamics [59]. The dynamical models described previously have been parametrized with more conventional, non-parasite consumers in mind. An interesting open question is how altering dynamical model parameters such as metabolic rate, functional response, and consumer-resource body size ratios to reflect parasite characteristics will affect our understanding of food web stability.

Second, the role of detritus, or dead organic matter, in food webs has yet to be adequately resolved in either structural or dynamical approaches. Detritus has typically been included as one or several separate nodes in many binary-link and flow-weighted food webs. In some cases, it is treated as an additional "primary producer," while in other cases both primary producers and detritivores connect to it. Researchers must think much more carefully about how to include detritus in all kinds of ecological studies [80], given that it plays a fundamental role in most ecosystems and has particular characteristics that differ from other food web nodes: it is non-living organic matter, all species contribute to detrital pools, it is a major resource for many species, and the forms it takes are extremely heterogeneous (e. g., suspended organic matter in water columns; fecal material; rotting trees; dead animal bodies; small bits of plants and molted cuticle, skin, and hair mixed in soil; etc.).

Third, there are many interactions that species participate in that go beyond strictly trophic interactions. Plant-animal mutualistic networks, particularly pollination and seed dispersal or "frugivory" networks, have re-

ceived the most attention thus far. They are characterized as "bipartite" (two-level) graphs, with links from animals to plants, but no links among plants or among animals [7,9,56,57,73,107]. While both pollination and seed dispersal do involve a trophic interaction, with animals gaining nutrition from plants during the interactions, unlike in classic predator-prey relationships a positive benefit is conferred upon both partners in the interaction. The evolutionary and ecological dynamics of such mutualistic relationships place unique constraints on the network structure of these interactions and the dynamical stability of such networks. For example, plant-animal mutualistic networks are highly nested and thus asymmetric, such that generalist plants and generalist animals tend to interact among themselves, but specialist species (whether plants or animals) also tend to interact with the most generalist species [7,107]. When simple dynamics are run on these types of "coevolutionary" bipartite networks, it appears that the asymmetric structure enhances long-term species coexistence and thus biodiversity maintenance [9].

Future Directions

Food web research of all kinds has expanded greatly over the last several years, and there are many opportunities for exciting new work at the intersection of ecology and network structure and dynamics. In terms of empiricism, there is still a paucity of detailed, evenly resolved community food webs in every habitat type. Current theory, models, and applications need to be tested against more diverse, more complete, and more highly quantified data. In addition, there are many types of datasets that could be compiled which would support novel research. For example, certain kinds of fossil assemblages may allow the compilation of detailed paleo food webs, which in turn could allow examination of questions about how and why food web structure does or does not change over deep time or in response to major extinction events. Another example is data illustrating the assembly of food [41] webs in particular habitats over ecological time. In particular, areas undergoing rapid successional dynamics would be excellent candidates, such as an area covered by volcanic lava flows, a field exposed by a retreating glacier, a hillside denuded by an earth slide, or a forest burned in a large fire. This type of data would allow empirically-based research on the topological dynamics of food webs. Another empirical frontier is the integration of multiple kinds of ecological interaction data into networks with multiple kinds of links—for example, networks that combine mutualistic interactions such as pollination and antagonistic interactions such as predator-prey relationships. In addition, more spatially explicit food web data can be compiled across microhabitats or linked macrohabitats [8]. Most current food web data is effectively aspatial even though trophic interactions occur within a spatial context. More could also be done to collect food web data based on specific instances of trophic interactions. This was done for the insects that live inside the stems of grasses in British fields. The web includes multiple grass species, grass herbivores, their parasitoids, hyper-parasitoids, and hyper-hyper parasitoids [67]. Dissection of over 160,000 grass stems allowed detailed quantification of the frequency with which the species ($S = 77$ insect plus 10 grass species) and different trophic interactions ($L = 126$) were observed.

Better empiricism will support improved and novel analysis, modeling, and theory development and testing. For example, while food webs appear fundamentally different in some ways from other kinds of "real-world" networks (e. g., they don't display power-law degree distributions), they also appear to share a common core network structure that is scale-dependent with species richness and connectance in predicable ways, as suggested by the success of the niche and related models. Some of the disparity with other kinds of networks, and the shared structure across food webs, may be explicable through finite-size effects or other methodological or empirical constraints or artifacts. However, aspects of these patterns may reflect attributes of ecosystems that relate to particular ecological, evolutionary, or thermodynamic mechanisms underlying how species are organized in complex bioenergetic networks of feeding interactions. Untangling artifacts from attributes [63] and determining potential mechanisms underlying robust phenomenological patterns (e. g., [10]) is an important area of ongoing and future research. As a part of this, there is much work to be done to continue to integrate structure and dynamics of complex ecological networks. This is critical for gaining a more comprehensive understanding of the conditions that underlie and promote different aspects of stability, at different levels of organization, in response to external perturbations and to endogenous short- and long-term dynamics.

As the empiricism, analysis and modeling of food web structure and dynamics improves, food web network research can play a more central and critical role in conservation and management [76]. It is increasingly apparent that an ecological network perspective, which encompasses direct and indirect effects among interacting taxa, is critical for understanding, predicting, and managing the impacts of species loss and invasion, habitat conversion, and climate change. Far too often, critical issues of ecosystem management have been decided on extremely limited knowledge of one or a very few taxa. For example, this

has been an ongoing problem in fisheries science. The narrow focus of most research driving fisheries management decisions has resulted in overly optimistic assessments of sustainable fishing levels. Coupled with climate stressors, over-fishing appears to be driving steep, rapid declines in diversity of common predator target species, and probably many other kinds of associated taxa [114]. Until we acknowledge that species of interest to humans are embedded within complex networks of interactions that can produce unexpected effects through the interplay of direct and indirect effects, we will continue to experience negative outcomes from our management decisions [118]. An important part of minimizing and mitigating human impacts on ecosystems also involves research that explicitly integrates human and natural dynamics. Network research provides a natural framework for analyzing and modeling the complex ways in which humans interact with and impact the world's ecosystems, whether through local foraging or large-scale commercial harvesting driven by global economic markets.

These and other related research directions will depend on efficient management of increasingly dispersed and diversely formatted ecological and environmental data. Ecoinformatic tools—the technologies and practices for gathering, synthesizing, analyzing, visualizing, storing, retrieving and otherwise managing ecological knowledge and information—are playing an increasingly important role in the study of complex ecosystems, including food web research [47]. Indeed, ecology provides an excellent testbed for developing, implementing, and testing new information technologies in a biocomplexity research context (e. g., Semantic Prototypes in Research Ecoinformatics/SPiRE: spire.umbc.edu/us/; Science Environment for Ecological Knowledge/SEEK: seek.ecoinformatics.org; Webs on the Web/WoW: www.foodwebs.org). Synergistic ties between ecology, physics, computer science and other disciplines will dramatically increase the efficacy of research that takes advantage of such interdisciplinary approaches, as is currently happening in food web and related research.

Bibliography

Primary Literature

1. Albert R, Jeong H, Barabási AL (2000) Error and attack tolerance of complex networks. Nature 406:378–382
2. Albert R, Barabási AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
3. Allesina S, Bodini A (2004) Who dominates whom in the ecosystem? Energy flow bottlenecks and cascading extinctions. J Theor Biol 230:351–358
4. Amaral LAN, Scala A, Berthelemy M, Stanley HE (2000) Classes of small-world networks. Proc Nat Acad Sci USA 97:11149–11152
5. Arii K, Parrott L (2004) Emergence of non-random structure in local food webs generated from randomly structured regional webs. J Theor Biol 227:327–333
6. Baird D, Ulanowicz RE (1989) The seasonal dynamics of the Chesapeake Bay ecosystem. Ecol Monogr 59:329–364
7. Bascompte J, Jordano P, Melián CJ, Olesen JM (2003) The nested assembly of plant-animal mutualistic networks. Proc Natl Acad Sci USA 100:9383–9387
8. Bascompte J, Melián CJ, Sala E (2005) Interaction strength combinations and the overfishing of a marine food web. Proc Natl Acad Sci 102:5443–5447
9. Bascompte J, Jordano P, Olesen JM (2006) Asymmetric coevolutionary networks facilitate biodiversity maintenance. Science 312:431–433
10. Beckerman AP, Petchey OL, Warren PH (2006) Foraging biology predicts food web complexity. Proc Natl Acad Sci USA 103:13745–13749
11. Bersier L-F, Banašek-Richter C, Cattin M-F (2002) Quantitative descriptors of food web matrices. Ecology 83:2394–2407
12. Briand F, Cohen JE (1984) Community food webs have scale-invariant structure. Nature 398:330–334
13. Brose U, Williams RJ, Martinez ND (2003) Comment on "Foraging adaptation and the relationship between food-web complexity and stability". Science 301:918b
14. Brose U, Williams RJ, Martinez ND (2006) Allometric scaling enhances stability in complex food webs. Ecol Lett 9:1228–1236
15. Camacho J, Arenas A (2005) Universal scaling in food-web structure? Nature 435:E3-E4
16. Camacho J, Guimerà R, Amaral LAN (2002) Robust patterns in food web structure. Phys Rev Lett 88:228102
17. Camacho J, Guimerà R, Amaral LAN (2002) Analytical solution of a model for complex food webs. Phys Rev Lett E 65:030901
18. Cartoza CC, Garlaschelli D, Caldarelli G (2006) Graph theory and food webs. In: Pascual M, Dunne JA (eds) Ecological Networks: Linking Structure to Dynamics in Food Webs. Oxford University Press, New York, pp 93–117
19. Cattin M-F, Bersier L-F, Banašek-Richter C, Baltensperger M, Gabriel J-P (2004) Phylogenetic constraints and adaptation explain food-web structure. Nature 427:835–839
20. Chen X, Cohen JE (2001) Global stability, local stability and permanence in model food webs. J Th Biol 212:223–235
21. Chen X, Cohen JE (2001) Transient dynamics and food web complexity in the Lotka–Volterra cascade model. Proc Roy Soc Lond B 268:869–877
22. Christian RR, Luczkovich JJ (1999) Organizing and understanding a winter's seagrass foodweb network through effective trophic levels. Ecol Model 117:99–124
23. Cohen JE (1977) Ratio of prey to predators in community food webs. Nature 270:165–167
24. Cohen JE (1977) Food webs and the dimensionality of trophic niche space. Proc Natl Acad Sci USA 74:4533–4563
25. Cohen JE (1978) Food Webs and Niche Space. Princeton University Press, NJ
26. Cohen JE (1989) Ecologists Co-operative Web Bank (ECOWeB™). Version 1.0. Machine Readable Data Base of Food Webs. Rockefeller University, NY

27. Cohen JE, Briand F (1984) Trophic links of community food webs. Proc Natl Acad Sci USA 81:4105–4109
28. Cohen JE, Newman CM (1985) A stochastic theory of community food webs: I. Models and aggregated data. Proc R Soc Lond B 224:421–448
29. Cohen JE, Palka ZJ (1990) A stochastic theory of community food webs: V. Intervality and triangulation in the trophic niche overlap graph. Am Nat 135:435–463
30. Cohen JE, Briand F, Newman CM (1986) A stochastic theory of community food webs: III. Predicted and observed length of food chains. Proc R Soc Lond B 228:317–353
31. Cohen JE, Briand F, Newman CM (1990) Community Food Webs: Data and Theory. Springer, Berlin
32. Cohen JE, Luczak T, Newman CM, Zhou Z-M (1990) Stochastic structure and non-linear dynamics of food webs: qualitative stability in a Lotka–Volterra cascade model. Proc R Soc Lond B 240:607–627
33. Cohen JE, Jonsson T, Carpenter SR (2003) Ecological community description using the food web, species abundance, and body size. Proc Natl Acad Sci USA 100:1781–1786
34. Daily GC (ed) (1997) Nature's services: Societal dependence on natural ecosystems. Island Press, Washington DC
35. Doyle J, Csete M (2007) Rules of engagement. Nature 446:860
36. Dunne JA (2006) The network structure of food webs. In: Pascual M, Dunne JA (eds) Ecological Networks: Linking Structure to Dynamics in Food Webs. Oxford University Press, New York, pp 27–86
37. Dunne JA, Williams RJ, Martinez ND (2002) Food web structure and network theory: the role of connectance and size. Proc Natl Acad Sci USA 99:12917–12922
38. Dunne JA, Williams RJ, Martinez ND (2002) Network structure and biodiversity loss in food webs: robustness increases with connectance. Ecol Lett 5:558–567
39. Dunne JA, Williams RJ, Martinez ND (2004) Network structure and robustness of marine food webs. Mar Ecol Prog Ser 273:291–302
40. Dunne JA, Brose U, Williams RJ, Martinez ND (2005) Modeling food-web dynamics: complexity-stability implications. In: Belgrano A, Scharler U, Dunne JA, Ulanowicz RE (eds) Aquatic Food Webs: An Ecosystem Approach. Oxford University Press, New York, pp 117–129
41. Dunne JA, Williams RJ, Martinez ND, Wood RA, Erwing DE (2008) Compilation and network analyses of Cambrian food webs. PLoS Biology 5:e102. doi:10.1371/journal.pbio.0060102
42. Egerton FN (2007) Understanding food chains and food webs, 1700–1970. Bull Ecol Soc Am 88(1):50–69
43. Elton CS (1927) Animal Ecology. Sidgwick and Jackson, London
44. Elton CS (1958) Ecology of Invasions by Animals and Plants. Chapman & Hall, London
45. Garlaschelli D, Caldarelli G, Pietronero L (2003) Universal scaling relations in food webs. Nature 423:165–168
46. Goldwasser L, Roughgarden JA (1993) Construction of a large Caribbean food web. Ecology 74:1216–1233
47. Green JL, Hastings A, Arzberger P, Ayala F, Cottingham KL, Cuddington K, Davis F, Dunne JA, Fortin M-J, Gerber L, Neubert M (2005) Complexity in ecology and conservation: mathematical, statistical, and computational challenges. Bioscience 55:501–510
48. Hardy AC (1924) The herring in relation to its animate environment. Part 1. The food and feeding habits of the herring with special reference to the East Coast of England. Fish Investig Ser II 7:1–53
49. Havens K (1992) Scale and structure in natural food webs. Science 257:1107–1109
50. Hutchinson GE (1959) Homage to Santa Rosalia, or why are there so many kinds of animals? Am Nat 93:145–159
51. Huxham M, Beany S, Raffaelli D (1996) Do parasites reduce the chances of triangulation in a real food web? Oikos 76:284–300
52. Jeong H, Tombor B, Albert R, Oltvia ZN, Barabási A-L (2000) The large-scale organization of metabolic networks. Nature 407:651–654
53. Jeong H, Mason SP, Barabási A-L, Oltavia ZN (2001) Lethality and centrality in protein networks. Nature 411:41
54. Jordán F, Molnár I (1999) Reliable flows and preferred patterns in food webs. Ecol Ecol Res 1:591–609
55. Jordán F, Scheuring I (2004) Network ecology: topological constraints on ecosystem dynamics. Phys Life Rev 1:139–229
56. Jordano P (1987) Patterns of mutualistic interactions in pollination and seed dispersal: connectance, dependence asymmetries, and coevolution. Am Nat 129:657–677
57. Jordano P, Bascompte J, Olesen JM (2003) Invariant properties in co-evolutionary networks of plant-animal interactions. Ecol Lett 6:69–81
58. Kondoh M (2003) Foraging adaptation and the relationship between food-web complexity and stability. Science 299:1388–1391
59. Lafferty KD, Dobson AP, Kurls AM (2006) Parasites dominate food web links. Proc Nat Acad Sci USA 103:11211–11216
60. Lindeman RL (1942) The trophic-dynamic aspect of ecology. Ecology 23:399–418
61. Link J (2002) Does food web theory work for marine ecosystems? Mar Ecol Prog Ser 230:1–9
62. MacArthur RH (1955) Fluctuation of animal populations and a measure of community stability. Ecology 36:533–536
63. Martinez ND (1991) Artifacts or attributes? Effects of resolution on the Little Rock Lake food web. Ecol Monogr 61:367–392
64. Martinez ND (1992) Constant connectance in community food webs. Am Nat 139:1208–1218
65. Martinez ND (1993) Effect of scale on food web structure. Science 260:242–243
66. Martinez ND (1994) Scale-dependent constraints on food-web structure. Am Nat 144:935–953
67. Martinez ND, Hawkins BA, Dawah HA, Feifarek BP (1999) Effects of sampling effort on characterization of food-web structure. Ecology 80:1044–1055
68. Martinez ND, Williams RJ, Dunne JA (2006) Diversity, complexity, and persistence in large model ecosystems. In: Pascual M, Dunne JA (eds) Ecological Networks: Linking Structure to Dynamics in Food Webs. Oxford University Press, New York, pp 163–185
69. May RM (1972) Will a large complex system be stable? Nature 238:413–414
70. May RM (1973) Stability and Complexity in Model Ecosystems. Princeton University Press, Princeton. Reprinted in 2001 as a "Princeton Landmarks in Biology" edition
71. McCann KS (2000) The diversity-stability debate. Nature 405:228–233
72. McKane AJ, Drossel B (2006) Models of food-web evolution. In: Pascual M, Dunne JA (eds) Ecological Networks: Linking

Structure to Dynamics in Food Webs. Oxford University Press, New York, pp 223–243
73. Memmott J (1999) The structure of a plant-pollinator network. Ecol Lett 2:276–280
74. Memmott J, Martinez ND, Cohen JE (2000) Predators, parasitoids and pathogens: species richness, trophic generality and body sizes in a natural food web. J Anim Ecol 69:1–15
75. Memmott J, Waser NM, Price MV (2004) Tolerance of pollination networks to species extinctions. Proc Royal Soc Lond Series B 271:2605–2611
76. Memmott J, Alonso D, Berlow EL, Dobson A, Dunne JA, Sole R, Weitz J (2006) Biodiversity loss and ecological network structure. In: Pascual M, Dunne JA (eds) Ecological Networks: Linking Structure to Dynamics in Food Webs. Oxford University Press, New York, pp 325–347
77. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298:763–764
78. Montoya JM, Solé RV (2002) Small world patterns in food webs. J Theor Biol 214:405–412
79. Montoya JM, Solé RV (2003) Topological properties of food webs: from real data to community assembly models. Oikos 102:614–622
80. Moore JC, Berlow EL, Coleman DC, de Ruiter PC, Dong Q, Hastings A, Collin Johnson N, McCann KS, Melville K, Morin PJ, Nadelhoffer K, Rosemond AD, Post DM, Sabo JL, Scow KM, Vanni MJ, Wall DH (2004) Detritus, trophic dynamics and biodiversity. Ecol Lett 7:584–600
81. Neutel AM, Heesterbeek JAP, de Ruiter PC (2002) Stability in real food webs: weak links in long loops. Science 296:1120–1123
82. Newman MEJ (2002) Assortative mixing in networks. Phys Rev Lett 89:208701
83. Newman M, Barabasi A-L, Watts DJ (eds) (2006) The Structure and Dynamics of Networks. Princeton University Press, Princeton
84. Odum E (1953) Fundamentals of Ecology. Saunders, Philadelphia
85. Pascual M, Dunne JA (eds) (2006) Ecological Networks: Linking Structure to Dynamics in Food Webs. Oxford University Press, New York
86. Paine RT (1988) Food webs: road maps of interactions or grist for theoretical development? Ecology 69:1648–1654
87. Pierce WD, Cushamn RA, Hood CE (1912) The insect enemies of the cotton boll weevil. US Dept Agric Bull 100:9–99
88. Pimm SL (1982) Food Webs. Chapman and Hall, London. Reprinted in 2002 as a 2nd edition by University of Chicago Press
89. Pimm SL (1984) The complexity and stability of ecosystems. Nature 307:321–326
90. Pimm SL, Lawton JH (1978) On feeding on more than one trophic level. Nature 275:542–544
91. Pimm SL, Lawton JH (1980) Are food webs divided into compartments? J Anim Ecol 49:879–898
92. Polis GA (1991) Complex desert food webs: an empirical critique of food web theory. Am Nat 138:123–155
93. Schoener TW (1989) Food webs from the small to the large. Ecology 70:1559–1589
94. Schoenly K, Beaver R, Heumier T (1991) On the trophic relations of insects: a food web approach. Am Nat 137:597–638
95. Solé RV, Montoya JM (2001) Complexity and fragility in ecological networks. Proc R Soc Lond B 268:2039–2045
96. Solow AR (1996) On the goodness of fit of the cascade model. Ecology 77:1294–1297
97. Solow AR, Beet AR (1998) On lumping species in food webs. Ecology 79:2013–2018
98. Srinivasan UT, Dunne JA, Harte H, Martinez ND (2007) Response of complex food webs to realistic extinction sequences. Ecology 88:671–682
99. Stouffer DB, Camacho J, Guimera R, Ng CA, Amaral LAN (2005) Quantitative patterns in the structure of model and empirical food webs. Ecology 86:1301–1311
100. Stouffer DB, Camacho J, Amaral LAN (2006) A robust measure of food web intervality. Proc Nat Acad Sci 103:19015–19020
101. Strogatz SH (2001) Exploring complex networks. Nature 410:268–275
102. Sugihara G, Schoenly K, Trombla A (1989) Scale invariance in food web properties. Science 245:48–52
103. Summerhayes VS, Elton CS (1923) Contributions to the ecology of Spitzbergen and Bear Island. J Ecol 11:214–286
104. Summerhayes VS, Elton CS (1928) Further contributions to the ecology of Spitzbergen and Bear Island. J Ecol 16:193–268
105. Thompson RM, Townsend CR (1999) The effect of seasonal variation on the community structure and food-web attributes of two streams: implications for food-web science. Oikos 87:75–88
106. Thompson RM, Townsend CR (2005) Food web topology varies with spatial scale in a patchy environment. Ecology 86:1916–1925
107. Vásquez DP, Aizen MA (2004) Asymmetric specialization: a pervasive feature of plant-pollinator interactions. Ecology 85:1251–1257
108. Warren PH (1989) Spatial and temporal variation in the structure of a freshwater food web. Oikos 55:299–311
109. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442
110. Williams RJ, Martinez ND (2000) Simple rules yield complex food webs. Nature 404:180–183
111. Williams RJ, Martinez ND (2004) Trophic levels in complex food webs: theory and data. Am Nat 163:458–468
112. Williams RJ, Martinez ND (2004) Diversity, complexity, and persistence in large model ecosystems. Santa Fe Institute Working Paper 04-07-022
113. Williams RJ, Berlow EL, Dunne JA, Barabási AL, Martinez ND (2002) Two degrees of separation in complex food webs. Proc Natl Acad Sci USA 99:12913–12916
114. Worm B, Sandow M, Oschlies A, Lotze HK, Myers RA (2005) Global patterns of predator diversity in the open oceans. Science 309:1365–1369
115. Yodzis P (1980) The connectance of real ecosystems. Nature 284:544–545
116. Yodzis P (1984) The structure of assembled communities II. J Theor Biol 107:115–126
117. Yodzis P (1998) Local trophodynamics and the interaction of marine mammals and fisheries in the Benguela ecosystem. J Anim Ecol 67:635–658
118. Yodzis P (2000) Diffuse effects in food webs. Ecology 81:261–266
119. Yodzis P, Innes S (1992) Body-size and consumer-resource dynamics. Am Nat 139:1151–1173.

Books and Reviews

Belgrano A, Scharler U, Dunne JA, Ulanowicz RE (eds) (2005) Aquatic Food Webs: An Ecosystem Approach. Oxford University Press, Oxford

Berlow EL, Neutel A-M, Cohen JE, De Ruiter P, Ebenman B, Emmerson M, Fox JW, Jansen VAA, Jones JI, Kokkoris GD, Logofet DO, McKane AJ, Montoya J, Petchey OL (2004) Interaction strengths in food webs: issues and opportunities. J Animal Ecol 73:585–598

Borer ET, Anderson K, Blanchette CA, Broitman B, Cooper SD, Halpern BS (2002) Topological approaches to food web analyses: a few modifications may improve our insights. Oikos 99:397–401

Christensen V, Pauly D (1993) Trophic Models of Aquatic Ecosystems. ICLARM, Manila

Cohen JE, Beaver RA, Cousins SH, De Angelis DL, et al (1993) Improving food webs. Ecology 74:252–258

Cohen JE, Briand F, Newman CM (1990) Community Food Webs: Data and Theory. Springer, Berlin

DeAngelis DL, Post WM, Sugihara G (eds) (1983) Current Trends in Food Web Theory. ORNL-5983, Oak Ridge Natl Laboratory

Drossel B, McKane AJ (2003) Modelling food webs. In: Bornholt S, Schuster HG (eds) Handbook of Graphs and Networks: From the Genome to the Internet. Wiley-VCH, Berlin

Hall SJ, Raffaelli DG (1993) Food webs: theory and reality. Advances in Ecological Research 24:187–239

Lawton JH (1989) Food webs. In Cherett JM, (ed) Ecological Concepts. Blackwell Scientific, Oxford

Lawton JH, Warren PH (1988) Static and dynamic explanations for patterns in food webs. Trends in Ecology and Evolution 3:242–245

Martinez ND (1995) Unifying ecological subdisciplines with ecosystem food webs. In Jones CG, Lawton JH, (eds) Linking Species and Ecosystems. Chapman and Hall, New York

Martinez ND, Dunne JA (1998) Time, space, and beyond: scale issues in food-web research. In Peterson D, Parker VT, (eds) Ecological Scale: Theory and Applications. Columbia University Press, New York

May RM (1983) The structure of food webs. Nature 301:566–568

May RM (2006) Network structure and the biology of populations. Trends Ecol Evol 21:394–399

Montoya JM, Pimm SL, Sole RV (2006) Ecological networks and their fragility. Nature 442:259–264

Moore J, de Ruiter P, Wolters V (eds) (2005) Dynamic Food Webs: Multispecies Assemblages, Ecosystem Development and Environmental Change. Academic Press, Elsevier, Amsterdam

Pimm SL, Lawton JH, Cohen JE (1991) Food web patterns and their consequences. Nature 350:669–674

Polis GA, Winemiller KO, (eds) (1996) Food Webs: Integration of Patterns & Dynamics. Chapman and Hall

Polis GA, Power ME, Huxel GR, (eds) (2003) Food Webs at the Landscape Level. University of Chicago Press

Post DM (2002) The long and short of food-chain length. Trends Ecol Evol 17:269–277

Strong DR (ed) (1988) Food web theory: a ladder for picking strawberries. Special Feature. Ecology 69:1647–1676

Warren PH (1994) Making connections in food webs. Trends Ecol Evol 9:136–141

Woodward G, Ebenman B, Emmerson M, Montoya JM, Olesen JM, Valido A, Warren PH (2005) Body size in ecological networks. Trends Ecol Evol 20:402–409

Foraging Robots

Alan FT Winfield
University of the West of England, Bristol, UK

Article Outline

Glossary
Definition of the Subject
Introduction
An Abstract Model of Robot Foraging
Single Robot Foraging
Multi-Robot (Collective) Foraging
Future Directions
Acknowledgments
Bibliography

Glossary

Autonomy In robotics autonomy conventionally refers to the degree to which a robot is able to make its own decisions about which actions to take next. Thus a fully autonomous robot would be capable of carrying out its entire mission or function without human control or intervention. A semi-autonomous robot would have a degree of autonomy but require some human supervision.

Behavior-based control Behavior-based control describes a class of robot control systems characterized by a set of conceptually independent task achieving modules, or behaviors. All task achieving modules are able to access the robot's sensors and when a particular module becomes active it is able to temporarily take control of the robot's actuators [2].

Braitenberg vehicle In robotics a Braitenberg vehicle is a conceptual mobile robot in which simple sensors are connected directly to drive wheels. Thus if, for instance, a front-left-side sensor is connected to the right-side drive wheel and vice-versa, then if the sensors are light sensitive the robot will automatically steer towards a light source [11].

Finite state machine In the context of this article a finite state machine (FSM) is a model of robot behavior which has a fixed number of states. Each state represents a particular set of actions or behaviors. The robot can be in only one of these states at any given instant in time and transitions between states may be triggered by either external or internal events.

Odometry Odometry refers to the technique of self-localization in which a robot measures how far it has

traveled by, for instance, counting the revolutions of its wheels. Odometry suffers the problem that wheel-slip leads to cumulative errors so odometric position estimates are generally inaccurate and of limited value unless combined with other localization techniques.

Robot In this article the terms *robot* and *mobile robot* are used interchangeably. A mobile robot is a man-made device or vehicle capable of (1) sensing its environment and (2) purposefully moving through and acting upon or within that environment. A robot may be fully autonomous, semi-autonomous or tele-operated.

Swarm intelligence The term swarm intelligence describes the purposeful collective behaviors observed in nature, most dramatically in social insects. Swarm intelligence is the study of those collective behaviors, in both natural and artificial systems of multiple agents, and how they emerge from the local interactions of the agents with each other and with their environment [8,19].

Tele-operation A robot is said to be tele-operated if it is remotely controlled by a human operator.

Definition of the Subject

Foraging robots are mobile robots capable of searching for and, when found, transporting objects to one or more collection points. Foraging robots may be single robots operating individually, or multiple robots operating collectively. Single foraging robots may be remotely tele-operated or semi-autonomous; multiple foraging robots are more likely to be fully autonomous systems. In robotics foraging is important for several reasons: firstly, it is a metaphor for a broad class of problems integrating exploration, navigation and object identification, manipulation and transport; secondly, in multi-robot systems foraging is a canonical problem for the study of robot-robot cooperation, and thirdly, many actual or potential real-world applications for robotics are instances of foraging robots, for instance cleaning, harvesting, search and rescue, landmine clearance or planetary exploration.

Introduction

Foraging is a benchmark problem for robotics, especially for multi-robot systems. It is a powerful benchmark problem for several reasons: (1) sophisticated foraging observed in social insects, recently becoming well understood, provides both inspiration and system level models for artificial systems. (2) Foraging is a complex task involving the coordination of several – each also difficult – tasks including efficient exploration (searching) for food or prey, physical collection (harvesting) of food or prey almost certainly requiring physical manipulation, transport of the food or prey, homing or navigation whilst carrying the food or prey back to a nest site, and deposition of the food item in the nest before returning to foraging. (3) Effective foraging requires cooperation between individuals involving either communication to signal to others where food or prey may be found (e. g. pheromone trails, or direction giving) and/or cooperative transport of food items too large for a single individual to transport.

There are, at the time of writing, very few types of foraging robots successfully employed in real-world applications. Most foraging robots are to be found in research laboratories or, if they are aimed at real-world applications, are at the stage of prototype or proof-of-concept. The reason for this is that foraging is a complex task which requires a range of competencies to be tightly integrated within the physical robot and, although the principles of robot foraging are now becoming established, many of the sub-system technologies required for foraging robots remain very challenging. In particular, sensing and situational awareness; power and energy autonomy; actuation, locomotion and safe navigation in unknown physical environments and proof of safety and dependability all remain difficult problems in robotics.

This article therefore focuses on describing and defining the principles of robot foraging. The majority of examples will necessarily be of laboratory systems not aimed at solving real-world applications but designed to model, illuminate and demonstrate those principles. The article proceeds as follows. Section "An Abstract Model of Robot Foraging" describes an abstract model of robot foraging, using a finite state machine (FSM) description to define the discrete sub-tasks, or states, that constitute foraging. The FSM method will be used throughout this article. The section then develops a taxonomy of robot foraging. Section "Single Robot Foraging" describes the essential design features that are a requirement of any foraging robot, whether operating singly or in a multi-robot team, and the technologies currently available to implement those features; the section then outlines a number of examples of single-robot foraging, including robots that are commercially available. Section "Multi-Robot (Collective) Foraging" then describes the development and state-of-the-art in multi-robot (collective) foraging; strategies for cooperation are described including information sharing, cooperative transport and division of labor (task allocation), the section then reviews approaches to the mathematical modeling of multi-robot foraging. The article concludes in Sect. "Future Directions" with a discussion of future directions in robot foraging and an outline of the technical challenges that remain to be solved.

An Abstract Model of Robot Foraging

Foraging, by humans or animals, is the act of searching (widely) for and collecting (or capturing) food for storage or consumption. Robot foraging is defined more broadly as searching for and collecting *any* objects, then returning those objects to a collection point. Of course if the robot(s) are searching for energy resources for themselves then robot foraging will have precisely the same meaning as human or animal foraging. In their definitive review paper on cooperative mobile robotics Cao et al. state simply "In foraging, a group of robots must pick up objects scattered in the environment" [14]. Østergaard et al. define foraging as "a two-step repetitive process in which (1) robots search a designated region of space for certain objects, and (2) once found these objects are brought to a goal region using some form of navigation" [54].

Figure 1 shows a Finite State Machine (FSM) representation of a foraging robot (or robots). In the model the robot is in always in one of four states: *searching*, *grabbing*, *homing* or *depositing*. Implied in this model is, firstly, that the environment or search space contains more than one of the target objects; secondly, that there is a single collection point (hence this model is sometimes referred to as central-place foraging), and thirdly, that the process continues indefinitely. The four states are defined as follows.

1. Searching. In this state the robot is physically moving through the search space using its sensors to locate and recognize the target items. At this level of abstraction we do not need to state how the robot searches: it could, for instance, wander at random, or it could employ a systematic strategy such as moving alternately left and right in a search pattern. The fact that the robot has to search at all follows from the pragmatic real-world assumptions that either the robot's sensors are of short range and/or the items are hidden (behind occluding obstacles for instance); in either event we must assume that the robot cannot find items simply by staying in one place and scanning the whole environment with its sensors. Object identification or recognition could require one of a wide range of sensors and techniques. When the robot finds an item it changes state from searching to grabbing. If the robot fails to find the target item then it remains in the searching state forever; searching is therefore the 'default' state.
2. Grabbing. In this state the robot physically captures and grabs the item ready to transport it back to the home region. Here we assume that the item is capable of being grabbed and conveyed by a single robot (the case of larger items that require cooperative transport by more than one robot will be covered later in this article). As soon as the item has been grabbed the robot will change state to homing.
3. Homing. In this state the robot must move, with its collected object, to a home or nest region. Homing clearly requires a number of stages, firstly, determination of the position of the home region relative to where the robot is now, secondly, orientation toward that position and, thirdly, navigation to the home region. Again there are a number of strategies for homing: one would be to re-trace the robot's path back to the home region using, for instance, odometry or by following a marker trail; another would be to home in on a beacon with a long range beacon sensor. When the robot has successfully reached the home region it will change state to depositing.
4. Depositing. In this state the robot deposits or delivers the item in the home region, and then immediately changes state to searching and hence resumes its search.

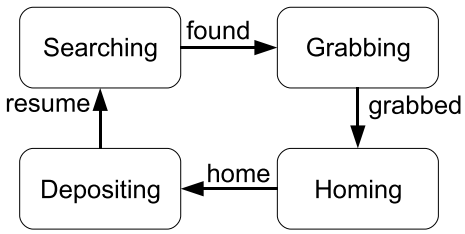

Foraging Robots, Figure 1
Finite State Machine for Basic Foraging

There are clearly numerous variations on this basic foraging model. Some are simplifications: for instance if a robot is searching for one or a known fixed number of objects then the process will not loop indefinitely. Real robots do not have infinite energy and so a model of practical foraging would need to take account of energy management. However, many variations entail either complexity within one or more of the four basic states (consider, for instance, objects that actively evade capture – a predator-prey model of foraging), or complexity in the interaction or cooperation between robots in multi-robot foraging. Thus the basic model stands as a powerful top-level abstraction.

A Taxonomy of Robot Foraging

Oster and Wilson classify the foraging strategies of social insects into five types, summarized in Table 1 [53]. Hölldobler and Wilson describe a more comprehensive taxonomy of insect foraging as a combination of strategies for (1) hunting, (2) retrieval and (3) defense [30]. However, since we will not be concerned in this article with

Foraging Robots, Table 1
Oster and Wilson's classification of insect foraging

Type	Description
I	Solitary insects find and retrieve prey singly
II	As I except that solitary foragers signal the location of food to other insects
III	Foragers depart the nest and follow 'trunk trails' before branching off to search unmarked terrain
IV	As II except that a group of insects assaults or retrieves the prey en-masse
V	Multiple insects forage as groups

defensive robot(s), then Oster and Wilson's classification is more than sufficient as a basis for consideration of robot foraging.

In robotics several taxonomies have been proposed for multi-robot systems. Dudek et al. define seven taxonomic axes: collective size; communications [range, topology and bandwidth]; collective reconfigurability; processing ability and collective composition [21]. Here collective size may be: single robot, pair of robots, limited (in relation to the size of the environment) or infinite (number of robots $N_r \gg 1$); communications range may be: none (i. e. robots do not communicate directly), near (robots have limited range communication) or infinite (any robot may communicate with any other). Collective reconfigurability refers to spatial organization and may be: static (robots are in a fixed formation); coordinated (robots may coordinate to alter their formation) or dynamic (spatial organization may change arbitrarily). Processing ability refers to the computational model of individuals, here Dudek et al. make the distinction between the general purpose computer which most practical robots will have, or simpler models including the finite state machine. Collective composition may be: identical (robots are both physically and functionally identical), homogeneous or heterogeneous. Dudek et al. makes the distinction – highly relevant to foraging robots – between tasks that are *traditionally single-agent*, tasks that are *traditionally multi-agent*, tasks that *require* multiple agents, or tasks that *may benefit* from multiple agents.

In contrast to Dudek's taxonomy which is based upon the characteristics of the robot(s), Balch characterizes tasks and rewards [3]. Balch's task taxonomy is particularly relevant to robot foraging because it leads naturally to the definition of performance metrics. Balch articulates six task axes: time; criteria; subject of action; resource limits; group movement and platform capabilities. Time and criteria are linked; time may be: limited (task performance is determined by how much can be achieved in the fixed time); minimum (task performance is measured as time taken to complete the task); unlimited time, or synchronized (robots must synchronize their actions). Criteria refers to how performance is optimized over time; it may be finite (performance is summed over a finite number of time steps); average (performance is averaged over all time) or discounted (future performance is discounted geometrically). Subject of action may be: object- or robot-based, depending upon whether the movement or positioning of objects or robots, respectively, is important. Balch's fourth criterion is again relevant to foraging: resource limits which may be: limited (external resources, i. e. objects to be foraged, are limited); energy (energy consumption must be minimized); internally competitive (one robot's success reduces the likelihood of success of another), or externally competitive (if, for instance, one robot team competes against another). See also [24] for a formal analysis and taxonomy of task allocation.

Østergaard et al. [54] develop a simple taxonomy of foraging by defining eight characteristics each of which has two values:

1. number of robots: single or multiple;
2. number of sinks (collection points for foraged items): single or multiple;
3. number of source areas (of objects to be collected): single or multiple;
4. search space: unbounded or constrained;
5. number of types of object to be collected: single or multiple;
6. object placement: in fixed areas or randomly scattered;
7. robots: homogeneous or heterogeneous and
8. communication: none or with.

This taxonomy maps more closely (but not fully) onto the insect foraging taxonomy of Table 1, but fails to capture task performance criteria, nor does it specify the strategy for either searching for, physically collecting or retrieving objects. Tables 2 and 3 propose a more comprehensive taxonomy for robot foraging that incorporates the robot-centric and task/performance oriented features of Dudek et al. and Balch, respectively, with the environmental features of Østergaard et al., whilst mapping onto the insect foraging classification of Oster and Wilson. The four major axes are Environment, Robot(s), Performance and Strategy. Each major axis has several minor axes and each of these can take the values enumerated in the third column of Tables 2 and 3. The majority of the values are self-explanatory, those that are not are annotated. Table 3 suggests a mapping of Oster and Wilson's classification onto robot foraging strategies.

Foraging Robots, Table 2
A taxonomy of robot foraging, part A

Major Axis	Minor Axis	Value	Notes
Environment	Search space	Unbounded	
		Constrained	
	Source areas	Single limited	Fixed number of objects
		Single unlimited	Objects 're-grow'
		Multiple	
	Sinks	Single	Home, nest or collection point
		Multiple	
	Object types	Single static	One type of static object, food or 'prey'
		Multiple static	
		Single active	One type of prey which evades capture
	Object placement	Fixed known locations	
		Uniform distribution	
		Clustered	
Robot(s)	Number	Single	
		Multiple	
	Type	Homogeneous	
		Heterogeneous	
	Object sensing	Limited	Short-range sensing
		Unlimited	Unlimited-range sensing
	Localization	None	
		Relative	
		Absolute	
	Communications	None	
		Near	
		Infinite	
	Power	Limited	Robot can run out of energy
		Forage	Robot forages for own energy
		Unlimited	

Following Balch [3], we can formalize successful object collection and retrieval as follows:

$$F(O_i, t) = \begin{cases} 1 & \text{if object } O_i \text{ is in a sink at time } t \\ 0 & \text{otherwise .} \end{cases} \quad (1)$$

If the foraging task is performance time limited (Performance time = fixed) and the objective is to maximize the number of objects foraged within fixed time T, then we may define a performance metric for the number of objects collected in time T,

$$P = \sum_{i=1}^{N_o} F(O_i, t_0 + T) \quad (2)$$

where N_o is the number of objects available for collection and t_0 is the start time. A metric for the number of objects foraged per second is clearly, $P_t = P/T$. P as defined here is independent of the number of robots. In order to measure the performance improvement of multi-robot foraging, for example the benefit gained by search or homing with trail following, recruitment or coordination (assuming the task can be completed by a single robot, grabbing = single and transport = single), then we may define the performance of a single robot P_s as defined in Eq. 2 and use this a baseline for the normalized performance P_m of a multi-robot system,

$$P_m = \frac{P}{N_r} \quad (3)$$

where N_r is the total number of robots. The efficiency of multi-robot foraging is then the ratio P_m/P_s.

Consider now that we wish instead to minimize the energy cost of foraging (Performance energy = minimum). If the energy cost of foraging object i is E_i, then we may define a performance metric for the number of objects for-

Foraging Robots, Table 3
A taxonomy of robot foraging, part B

Major Axis	Minor Axis	Value	Notes
Performance	Time	Fixed	Objects foraged per second
		Minimum	Minimize time to forage
		Unlimited	
	Energy	Fixed	Objects foraged per Joule
		Minimum	Minimize energy used
		Unlimited	
Strategy	Search	Random wander	
		Geometrical pattern	
		Trail following	Type III
		Follow other robots	
		In teams	Type V
	Grabbing	Single	
		Cooperative	Type IV
	Transport	Single	
		Cooperative	Type IV
	Homing	Self-navigation	
		Home on beacon	
		Follow trail	
	Recruitment	None	Type I
		Direct	Type II
		Indirect	
	Coordination	None	Type I
		Self-organized	Types II-V
		Master slave	
		Central control	

aged per Joule of energy,

$$P_e = \frac{N_o}{\sum_{i=1}^{N_o} E_i} \tag{4}$$

then seek the foraging strategy that achieves the highest value for P_e.

Single Robot Foraging

The design of any foraging robot, whether operating alone or as part of a multi-robot team, will necessarily follow a similar basic pattern. The robot will require one or more *sensors*, with which it can both sense its environment for safe navigation and detect the objects or food-items it seeks; *actuators* for both locomotion through the environment and for physically collecting, holding then depositing its prey, and a *control system* to provide the robot with – at the very least – a set of basic reflex behaviors. Since robots are machines that perform work, which requires energy, then *power management* is important; if, for instance, the robot is foraging for its own energy then balancing its energy needs with the energy cost of foraging is clearly critical. Normally, a *communication* transceiver is also a requirement, either to allow remote tele-operation or monitoring or, in the case of multi-robot collective foraging, for robot-robot communications. A foraging robot is therefore a complex set of interconnected sub-systems and, although its system-level structure may follow a standard pattern, the shape and form of the robot will vary significantly depending upon its intended environment and application.

This section will review approaches and techniques for sensing, actuation, communications and control, within the context of robot foraging and with reference to research which focuses on advancing specific capabilities within each of these domains of interest. Then a number of examples of single robot foraging are given, including real-world applications.

Sensing

Obstacle Avoidance and Path Planning There are many sensors available to designers of foraging robots and

a comprehensive review can be found in [22]. A foraging robot will typically require short or medium range proximity sensors for obstacle avoidance, such as infra-red return-signal-intensity or ultrasonic- or laser-based time-of-flight systems. The most versatile and widely used device is the 2D or 3D scanning laser range finder which can provide the robot with a set of radial distance measurements and hence allow the robot to plan a safe path through obstacles [64].

Localization All but the simplest foraging robots will also require sensors for localization, that is to enable the robot to estimate its own position in the environment. If external reference signals are available such as fixed beacons so that a robot can use radio trilateration to fix its position relative to those beacons, or a satellite navigation system such as the Global Positioning System (GPS), then localization is relatively straightforward. If no external infrastructure is available then a robot will typically make use of several sensors including odometry, an inertial measurement unit (IMU) and a magnetic compass, often combining the data from all of these sensors, including laser scanning data, to form an estimate of its position. Simultaneous Localization and Mapping (SLAM) is a well-known stochastic approach which typically employs Kalman filters to allow a robot (or a team of robots) to both fix their position relative to observed landmarks and map those landmarks with increasing confidence as the robot(s) move through the environment [18].

Object Detection Vision is often the sensor of choice for object detection in laboratory experiments in foraging robots. If, for instance, the object of interest has a distinct color which stands out in the environment then standard image processing techniques can be used to detect then steer towards the object [31]. However, if the environment is visually cluttered, unknown or poorly illuminated then vision becomes problematical. Alternative approaches to object detection include, for instance, artificial odor sensors: Hayes et al. demonstrated a multi-robot approach to localization of an odor source [28]. An artificial whisker modeled on the Rat mystacial vibrissae has recently been demonstrated [56], such a sensor could be of particular value in dusty or smoky environments.

Actuation

Locomotion The means of physical locomotion for a foraging robot can take many forms and clearly depends on the environment in which the robot is intended to operate. Ground robots typically use wheels, tracks or legs, although wheels are predominantly employed in proof-of-concept or demonstrator foraging robots. An introduction to the technology of robot mobility can be found in [63]. Flying robots (unmanned air vehicles – UAVs) are either fixed- or rotary-wing; for recent examples of work towards teams of flying robots see [13] (fixed-wing) and [51] (rotary-wing). Underwater robots (unmanned underwater vehicles – UUVs) generally use the same principles for propulsion as submersible remotely operated vehicles (ROVs), [70]. Whatever the means of locomotion important principles which apply to all foraging robots are that robot(s) must be able to (1) move with sufficient stability for the object detection sensors to be able to operate effectively and (2) position themselves with sufficient precision and stability to allow the object to be physically grabbed. These factors place high demands on a foraging robot's physical locomotion system, especially if the robot is required to operate in soft or unstable terrain.

Object Manipulation The manipulation required of a foraging robot is clearly dependent on the form of the object and the way the object presents itself to the robot as it approaches. The majority of foraging experiments or demonstrations have simplified the problem of object manipulation by using objects that are, for instance, always the right way up (metal pucks or wooden sticks protruding from holes) so that a simple gripper mounted on the front of the robot is able to grasp the objects with reasonable reliability. However, in general a foraging robot would require the versatility of a robot arm (multi-axis manipulator) and general purpose gripper (hand) such that – with appropriate vision sensing – the robot can pick up the object regardless of its shape and orientation. This technology is well developed in tele-operated robots used for remote inspection and handling of dangerous materials or devices, see [62,66].

Communications

Communications is of fundamental importance to robot foraging. Only in the simplest case of a single robot foraging autonomously would communications be unnecessary. For single robot teleoperation radio communication between operator and robot is clearly an essential requirement. In multi-robot foraging robot-robot communication is frequently employed to improve multi-robot performance; all six axes of strategy in the taxonomy of Table 3: search, grabbing, transport, homing, recruitment and coordination may require some form of robot-robot communication. Arai et al. point out the important distinction between *explicit* and *implicit* communication [1].

Explicit Communication Explicit communication applies when robots need to exchange information directly. The physical medium of communication is frequently (but not necessarily) radio, and wireless local area network (WLAN) technology is highly appropriate to terrestrial multi-robot systems, not least because a spatially distributed team of wireless networked robots naturally forms an *ad-hoc* network, which – providing the team maintains sufficient connectivity – allows any robot to communicate with any other via multiple hops, [69]. A method for linking wireless connectivity to locomotion in order to maintain connectivity is described in [52]; work that falls within the framework of *situated* communications proposed by Støy. Situated communication pertains when "both the physical properties of the signal that transfers the message and the content of the message contribute to its meaning" [65].

Implicit Communication Implicit communication applies when robots communicate not directly but via the environment, also known as *stigmergic* communications. Thus one robot changes the environment and another senses the change and alters its behavior accordingly. Beckers et al., in one of the first demonstrations of self-organized multi-robot puck clustering, show that stigmergic communication alone can give rise to the desired overall group behavior [6]. However, in their study on multi-robot communication, Balch and Arkin show that while stigmergy may be sufficient to complete the task, direct communication can increase efficiency [4]. Trail following, in which a robot follows a short-lived trail left by other(s), is an example of implicit communication [59,60].

Control

From a control perspective the simplicity of the finite state machine for basic foraging, in Fig. 1, is deceptive. In principle, a very simple foraging robot could be built with basic hard-wired reflex actions such as obstacle avoidance and taxis toward the attractor object; such a robot is known as a Braitenberg vehicle, after his landmark work [11]. However, even simple foraging requires a complex set of competencies that would be impractical to implement except as a program on one or more embedded computers (microprocessors) in the robot. There are clearly many ways of building such a control program, but in the field of mobile robotics a number of robot control architectures have been defined. Such architectures mean that robot designers can approach the design of the control system in a principled way.

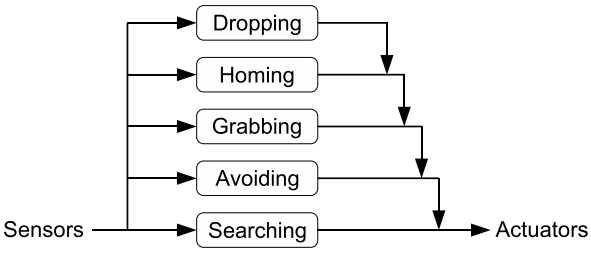

Foraging Robots, Figure 2
Subsumption control architecture for basic foraging

A widely adopted robot control architecture, first proposed and developed by Brooks, is the layered *subsumption* architecture known generically as behavior-based control [12]. Behavior-based control is particularly relevant to foraging robots since, like foraging, it is biologically inspired. In particular, as Arkin describes in [2], the principles of behavior-based control draw upon ethology – the study of animal behavior in the natural environment. Essentially behavior-based control replaces the functional modularity of earlier robot control architectures with task achieving modules, or behaviors. Matarić uses Brooks' behavior language (BL) to implement a set of basic behaviors for multi-robot foraging, as described in more detail below in Sect. Multi-Robot (Collective) Foraging, [46,47]. Refer to [14] for a comprehensive review of group control architectures for multi-robot systems.

Figure 2 shows the subsumption architecture for basic foraging (from Fig. 1), with the addition of *avoidance* for safely avoiding obstacles (including other robots in the case of multi-robot foraging). Each behavior runs in parallel and, when activated suppresses the output of the layer(s) below to take control of the robot's actuators.

Examples of Single Robot Foraging

A Soda-Can Collecting Robot Possibly the first demonstration of autonomous single-robot foraging is Connell's soda-can collecting robot *Herbert*, [15]. Herbert's task was to wander safely through an office environment while searching for empty soda-cans; upon finding a soda-can Herbert would need to carefully grab the can with its hand and 2 degrees-of-freedom arm, then return to a waste basket to deposit it before resuming the search. Herbert therefore represents an implementation of exactly the basic foraging model of Fig. 1 and 2. However, two of the behaviors are not so straightforward. Both searching and homing require the robot to be able to navigate safely through a cluttered and unstructured 'real-world' environment, while grabbing is equally complex given the precision required

to safely reach and grab the soda-can. Thus Herbert's control system required around 40 low-level behaviors in order to realize foraging.

A Robot Predator Arguably the first attempt to build a robot capable of actively predating for its own energy is the *Slugbot* of Holland and co-workers, [26,33]. The Slugbot (Fig. 3) solved the difficult problems of finding and collecting slugs in an energy efficient manner by means of, firstly, a long but light articulated arm which allows the robot to scan (in spiral fashion) a large area of ground for slugs without having to physically move the whole robot (which is much more costly in energy). Secondly, the special purpose gripper at the end of the arm is equipped with a camera which, by means of reflected red light and appropriate vision processing, is able to reliably detect and collect the slugs. An evolution of the Slugbot, the *Ecobot*, uses microbial fuel cell (MFC) technology to generate electrical energy directly from unrefined biomass [49].

Real-World Foraging Robots Autonomous crop harvesting is an obvious real-world application of single-robot foraging. The Demeter system [57] has successfully demonstrated automated harvesting of cereal crops. Demeter uses a combination of GPS for coarse navigation and vision to sense the crop-line and hence fine-tune the harvester's steering to achieve a straight and even cut of the crop. The vision processing is challenging because it has to cope with a wide range of lighting conditions including – in conditions of bright sunlight – shadows cast onto the crop line by the harvester itself. In the field of automated agriculture a number of proof-of-concept robot harvesters have been demonstrated for cucumber, tomato and other fruits [34,35].

Robot lawn mowers and vacuum cleaners can similarly be regarded as simple forms of foraging robot and are notable because they are the only form of autonomous foraging robot in commercial production; in both cases the search task is simple because the grass, or dirt are not discrete objects to be found. The search problem for robot lawn movers and vacuum cleaners thus becomes the problem of energy efficient strategies for (1) safely covering the whole search space while avoiding obstacles and (2) homing and docking to a re-charging station. Robot lawn mowers typically require a wire to be installed at the perimeter of the lawn, thus delimiting the robot's working area, see [29] for a survey of commercial robot lawn mowers. A short account of the development of a vacuum cleaning robot is given in [58].

Although technically an *off-world* application, the planetary rover may be regarded as an instance of single-robot foraging in which the objects of interest (geological samples) are collected and analyzed within the robot. Autonomous sample-return robots would be true foragers [61]. The proof-of-concept robot astrobiologist *Zoë* forages – in effect – for evidence of life [67].

Multi-Robot (Collective) Foraging

Foraging is clearly a task that lends itself to multi-robot systems and, even if the task can be accomplished by a single robot, foraging should – with careful design of strategies for cooperation – benefit from multiple robots. Swarm intelligence is the study of natural and artificial systems of multiple agents in which there is no centralized or hierarchical command or control. Instead, global swarm behaviors emerge as a result of local interactions between the agents and each other, and between agents and the environment, [8]. Swarm robotics is concerned with the design of artificial robot swarms based upon the principles of swarm intelligence, thus control is completely distributed and robots, typically, must choose actions on the basis only of local sensing and communications, [7,16]. Swarm robotics is thus a sub-set of multi-robot systems and, in the taxonomy of Table 2 the strategy: coordination = self-organized.

Foraging Robots, Figure 3
The Slugbot: a proof-of-concept robot predator

Foraging is therefore a benchmark problem within swarm robotics, not least because of the strong cross-over between the study of self-organization in social insects and their artificial counterparts within swarm intelligence [19]. This section will therefore focus on examples of multi-robot foraging from within the field of swarm robotics. Three strategies for cooperation will be outlined: information sharing, physical cooperation and division of labor. The section will conclude with an outline of the problem of mathematical modeling of swarms of foraging robots.

Without Cooperation

Balch and co-workers describe the winners of the 'Office Cleanup Event' of the 1994 AAAI Mobile Robot Competition: a multi-robot trash-collecting team [5]. The robots were equipped with a vision system for recognition and distance estimation of trash items (primarily soda cans) and differentiation between trash items, wastebaskets and other robots. The robots did not communicate, but employed a collective strategy in which robots generate a strong repulsive force if they see each other while searching, and a weaker (but sufficient for avoidance) repulsive force while in other states; this had the effect of causing the robots to spread-out and hence search the environment more efficiently. Interestingly, Balch et al. found that the high density of trash in the competition favored a 'sit-and-spin' strategy to scan for trash items rather than the random wander approach of the original design. The FSM was essentially the same schema as shown in Fig. 1 except that since there could be a number of wastebaskets at unknown locations then 'homing' becomes 'search for nearest wastebasket'.

Strategies for Cooperation

Information Sharing Matarić and Marjanovic provide what is believed to be the first description of a multi-robot foraging experiment using real (laboratory) robots in which there is no centralized control [47]. They describe a system of 20 identical 12″ 4-wheeled robots, equipped with: a two-pronged forklift for picking up, carrying and stacking metal pucks; proximity and bump sensors; radio transceivers for data communication and a sonar-based global positioning system. Matarić and Marjanovic extend the basic five state foraging model (wandering, grabbing, homing, dropping and avoiding), to introduce information sharing as follows. If a robot finds a puck it will grab it but also broadcast a radio message to tell other robots it has found a puck. Meanwhile, if another robot in the locale hears this message it will first enter state *tracking* to home in on the source of the message, then state *searching* – a more localized form of wandering. The robot will return to wandering if it finds no puck within some time out period. Furthermore, while in state *tracking* a robot will also transmit a radio signal. If nearby robots hear this signal they will switch from wandering into *following* to pursue the tracking robot. Thus the tracking robot actively recruits additional robots as it seeks the original successful robot (a form of secondary swarming, [48]); when the tracking robot switches to searching its recruits will do the same. Figure 4 shows a simplified FSM. Within the taxonomy of Table 3 Strategy : recruitment = direct and indirect.

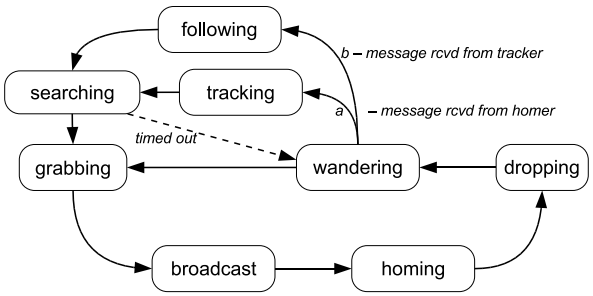

Foraging Robots, Figure 4
Finite State Machine for multi-robot foraging with recruitment – adapted from [47]

Physical Cooperation

Cooperative Grabbing Consider the case of multi-robot foraging in which the object to be collected cannot be grabbed by a single robot working alone, in Table 3 this is Strategy: grabbing = cooperative. Ijspeert et al. describe an experiment in collaborative stick-pulling in which two robots must work together to pull a stick out of a hole [32,44]. Each *Khephera* robot is equipped with a gripper capable of grabbing and lifting the stick, but the hole containing the stick is too deep for one robot to be able to pull the stick out alone; one robot must pull the stick half-way then wait for another robot to grab the stick and lift it clear of the hole, see Fig. 5. Ijspeert and co-workers describe an elegant minimalist strategy which requires no direct communication between robots. If one robot finds a stick it will lift it and wait. If another finds the same stick it will also lift it, on sensing the force on the stick from the second robot the first robot will let go, hence allowing the second to complete the operation.

Cooperative Transport Now consider the the situation in which the object to be collected is too large to be transported by a single robot, in Table 3 Strategy: transport =

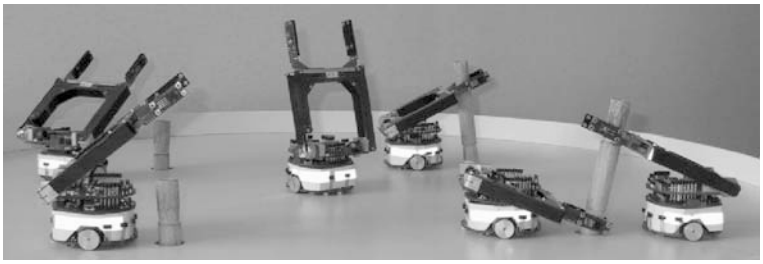

Foraging Robots, Figure 5
Cooperative grabbing: Khephera robots engaged in collective stick-pulling. With kind permission of A. Martinoli

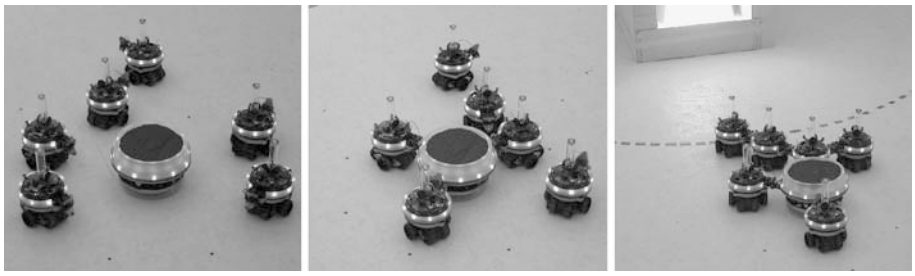

Foraging Robots, Figure 6
Cooperative transport by s-bots. (*Left*) s-bots approach the attractor object, (*middle*) s-bots start to grab the object, (*right*) s-bots collectively drag the object toward a beacon. With kind permission of M. Dorigo

cooperative. Parker describes the ALLIANCE group control architecture applied to an example of cooperative box-pushing by two robots [55].

Arguably the most accomplished demonstration of cooperative multi-robot foraging to date is within the *swarm-bot* project of Dorigo and co-workers [20]. The *s-bot* is a modular robot equipped with both a gripper and a gripping ring, which allows one robot to grip another [50]. Importantly, the robot is able to rotate its wheelbase independently of the gripping ring so that robots can grip each other at any arbitrary point on the circumference of the grip ring but then rotate and align their wheels in order to be able to move as a single unit (a *swarm-bot*). Groß et al. describe cooperative transport which uses visual signaling [27]. s-bots are attracted to the (large) object to be collected by its ring of red LEDs. The s-bot's LEDs are blue, but when an s-bot finds and grabs the attractor object it switches its LEDs to red. This increases the red light intensity to attract further s-bots which may grab either the object, or arbitrarily a robot already holding the object. The s-bots are then able to align and collectively move the object.

Division of Labor In multi-robot foraging it is well know that overall performance (measured, for instance, as the number of objects foraged per robot in a given time interval), does not increase monotonically with increasing team size because of interference between robots (overcrowding), [4,25,38]. Division of labor in ant colonies has been well studied and in particular a response threshold model is described in [9] and [10]; in essence a threshold model means that an individual will engage in a task when the level of some task-associated stimulus exceeds its threshold.

For threshold-based multi-robot foraging with division of labor Fig. 7 shows a generalized finite state machine for each robot. In this foraging model the robot will not search endlessly. If the robot fails to find a food-item because, for instance, its searching time exceeds a maximum search time threshold T_s, or its energy level falls below a minimum energy threshold, then it will abandon its search and return home without food, shown as *failure*. Conversely *success* means food was found, grabbed and deposited. Note, however, that a robot might see a food-item but fail to grab it because, for instance, of competition with another robot for the same food-item. The robot now also has a *resting* state during which time it remains in the nest conserving energy. The robot will stop resting and leave home which might be according to some threshold criterion, such as its resting time exceeding the maximum rest time threshold T_r, or the overall nest energy falling below a given threshold.

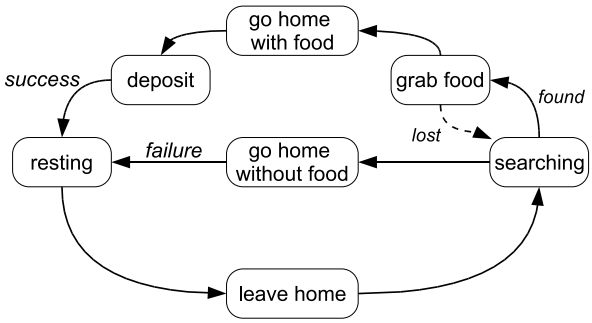

Foraging Robots, Figure 7
Finite State Machine for foraging with division of labor

Let us consider the special case of multi-robot foraging in which robots are foraging for their own energy. For an individual robot foraging costs energy, whereas resting conserves energy. We can formally express this as follows. Each robot consumes energy at A units per second while searching or retrieving and B units per second while resting, where $A > B$. Each discrete food item collected by a robot provides C units of energy to the swarm. The average food item retrieval time, is a function of the number of foraging robots x, and the density of food items in the environment, ρ, thus $t = f(x, \rho)$.

If there are N robots in the swarm, E_c is the energy consumed and E_r the energy retrieved, per second, by the swarm then

$$E_c = Ax + B(N - x) \tag{5}$$

$$E_r = Cx/t = \frac{Cx}{f(x, \rho)} \,. \tag{6}$$

The average energy income to the swarm, per second, is clearly the difference between the energy retrieved and the energy consumed,

$$E = E_r - E_c = \left(\frac{C}{f(x, \rho)} - (A - B) \right) x - BN \tag{7}$$

Equation 7 shows that maximizing the energy income to the swarm requires either increasing the number of foragers x or decreasing the average retrieval time $f(x, \rho)$. However, if we assume that the density of robots in the foraging area is high enough that interference between robots will occur then, for constant ρ, increasing x will increase $f(x, \rho)$. Therefore, for a given food density ρ there must be an optimal number of foragers x^*.

Krieger ad Billeter adopt a threshold-based approach to the allocation of robots to either foraging or resting; in their scheme each robot is allocated a fixed but randomly chosen activation threshold [36]. While waiting in the nest each robot listens to a periodic radio broadcast indicating the nest-energy level E; when the nest-energy level falls below the robot's personal activation threshold then it leaves the nest and searches for food. It will continue to search until either its search is successful, or it runs out of energy and returns home; if its search is successful and it finds another food-item the robot will record its position (using odometry). On returning home the robot will radio its energy consumption thus allowing the nest to update its overall net energy. Krieger and Billeter show that team sizes of 3 or 6 robots perform better than 1 robot foraging alone, but larger teams of 9 or 12 robots perform less well. Additionally, they test a recruitment mechanism in which a robot signals to another robot waiting in the nest to follow it to the food source, in tandem. Krieger's approach is, strictly speaking, not fully distributed in that the nest is continuously tracking the average energy income E; the nest is – in effect – acting as a central coordinator.

Based upon the work of [17] on individual adaptation and division of labor in ants, Labella et al. describe a fully distributed approach that allows the swarm to self-organize to automatically find the optimal value x^* [37]. They propose a simple adaptive mechanism to change the ratio of foragers to resters by adjusting the probability of leaving home based upon successful retrieval of food. With reference to Fig. 7 the mechanism works as follows. Each robot will *leave home*, i. e. change state from resting to searching, with probability P_l. Each time the robot makes the *success* transition from deposit to resting, it increments its P_l value by a constant Δ multiplied by the number of consecutive successes, up to a maximum value $P_{\max}$. Conversely, if the robot's searching time is up, the transition *failure* in Fig. 7, it will decrement its P_l by Δ times the number of consecutive failures, down to minimum $P_{\min}$. Interestingly, trials with laboratory robots show that the same robots self-select as foragers or resters – the algorithm exploits minor mechanical differences that mean that some robots are better suited as foragers.

Recently Liu et al. have extended this fully distributed approach by introducing two additional adaptation rules [43]. As in the case of Labella et al. individual robots use internal cues (successful object retrieval), but Liu adds environmental cues (collisions with team mates while searching), and social cues (team mate success in object retrieval), to dynamically vary the time spent foraging or resting. Furthermore, Liu investigates the performance of a number of different adaptation strategies based on combinations of these three cues. The three cues increment or decrement the searching time and resting time thresholds T_s and T_r as follows (note that adjusting T_r is equivalent

Foraging Robots, Table 4
Foraging swarm strategy – cue combinations

	Internal cues	Social cues	Environment cues
S_1 (baseline)	×	×	×
S_2	✓	×	×
S_3	✓	✓	×
S_4	✓	✓	✓

to changing the probability of leaving the nest P_l):

1. Internal cues. If a robot successfully finds food it will reduce its own rest time T_r; conversely if the robot fails to find food it will increase its own rest time T_r.
2. Environment cues. If a robot collides with another robot while searching, it will reduce its T_s and increase its T_r times.
3. Social cues. When a robot returns to the nest it will communicate its food retrieval success or failure to the other robots in the nest. A successful retrieval will cause the other robots in the nest to increase their T_s and reduce their T_r times. Conversely failure will cause the other robots in the nest to reduce their T_s and increase their T_r times.

In order to evaluate the relative effect of these cues three different strategies are tested, against a baseline strategy of no cooperation. The strategy/cue combinations are detailed in Table 4.

Figures 8 and 9, from [43], show the number of active foragers and the instantaneous net swarm energy, respectively, for a swarm of eight robots. In both plots the food density in the environment is changed at time $t = 5000$ and again at time $t = 10000$ seconds. Figure 8 shows the swarm's ability to automatically adapt the number of active foragers in response to each of the step changes in food density. The baseline strategy S_1 shows of course that all eight robots are actively foraging continuously; S_2–S_4 however require fewer active foragers and strategies with social and environmental cues, S_3 and S_4, clearly show the best performance. Notice, firstly that the additional of social cues – communication between robots – significantly improves the rate at which the system can adapt the ratio of foragers to resters and, secondly, that the addition of environmental cues – collisions with other robots – brings only a marginal improvement. The rates of change of net swarm energy in Fig. 9 tell a similar story. Interestingly, however, we see very similar gradients for S_2–S_4 when the food density is high (on the RHS of the plot), but when the food density is medium or poor the rate of increase in net energy of strategies S_3 and S_4 is significantly better than S_2. This result interestingly suggests that foraging robots benefit more from cooperation when food is scarce, than when food is plentiful.

Mathematical Modeling

A multi-robot system of foraging robots is typically a stochastic non-linear dynamical system and therefore challenging to mathematically model, but without such models any claims about the correctness of foraging algorithms are weak. Experiments in computer simulation or with real-robots (which provide in effect an 'embodied' simulation) allow limited exploration of the parameter space and can at best only provide weak inductive proof of correctness. Mathematical models on the other hand, allow analysis of the whole parameter space and discovery of optimal parameters. Ultimately, in real-world applications, validation of a foraging robot system for safety and dependability will require a range of formal approaches including mathematical modeling.

Martinoli and coworkers proposed a *microscopic* approach to study collective behavior of a swarm of robots engaged in cluster aggregation [45] and collaborative stick-pulling [32], in which a robot's interactions with other robots and the environment are modeled as a series of stochastic events, with probabilities determined by simple geometric considerations and systematic experiments with one or two real robots.

Lerman, Martinoli and co-workers have also developed the *macroscopic* approach, as widely used in physics, chemistry, biology and the social sciences, to directly describe the collective behavior of the robotic swarm. A class of macroscopic models have been used to study the effect of interference in a swarm of foraging robots [38] and collaborative stick-pulling [39,44]. A review of macroscopic models is given in [41]. More recently, Lerman et al. [40] successfully expanded the macroscopic probabilistic model to study dynamic task allocation in a group of robots engaged in a puck collecting task, in which the robots need to decide whether to pick up red or green pucks based on observed local information.

A Macroscopic Mathematical Model of Multi-Robot Foraging with Division of Labor Recently Liu et al. have applied the macroscopic approach to develop a mathematical model for foraging with division of labor (as described above in Section "Division of Labor"), [42]. The finite state machine of Fig. 7 is extended in order to describe the probabilistic behavior of the whole swarm, resulting in a probabilistic finite state machine (PFSM). In Fig. 10 each state represents the average number of robots

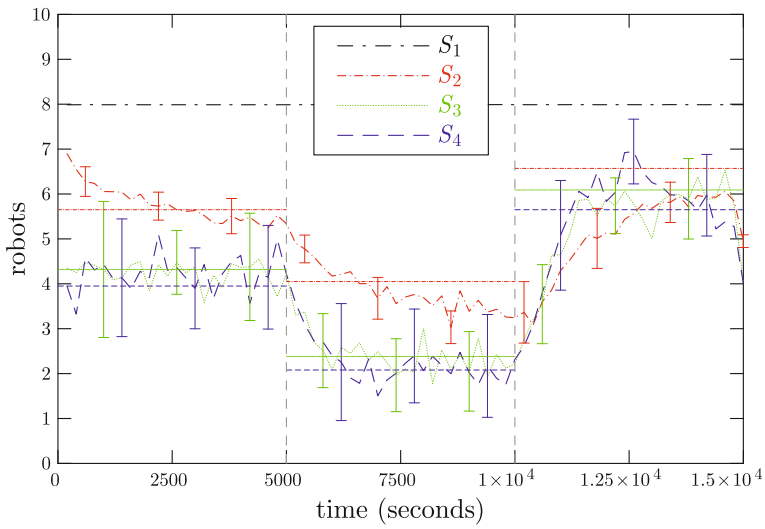

Foraging Robots, Figure 8

Number of foraging robots x in a foraging swarm of $N = 8$ robots with self-organized division of labor. S_1 is the baseline (no cooperation strategy); S_2, S_3 and S_4 are three different cooperation strategies (see Table 4). Food density changes from 0.03 (medium) to 0.015 (poor) at $t = 5000$, then from 0.015 (poor) to 0.045 (rich) at $t = 10000$. Each plot is the average of 10 runs

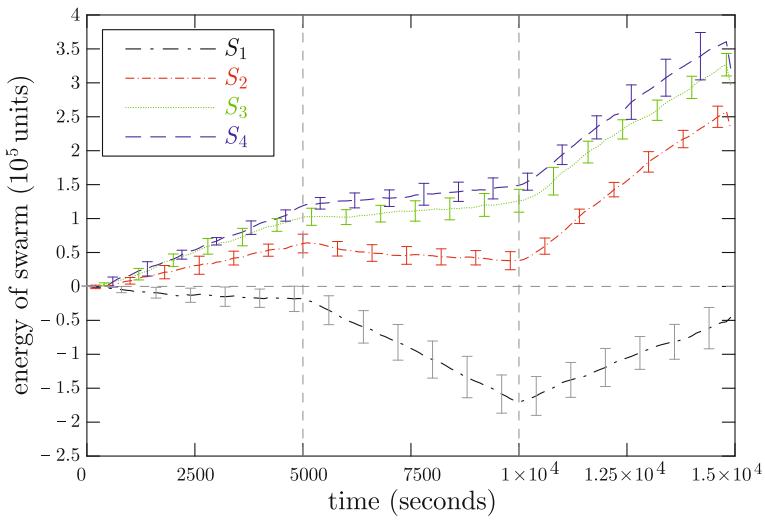

Foraging Robots, Figure 9

Instantaneous net energy E of a foraging swarm with self-organized division of labor. S_1 is the baseline (no cooperation strategy); S_2, S_3 and S_4 are three different cooperation strategies (see Table 4). Food density changes from 0.03 (medium) to 0.015 (poor) at $t = 5000$, then from 0.015 (poor) to 0.045 (rich) at $t = 10000$. Each plot is the average of 10 runs

in that state. The five basic states are S for *searching*, H for *homing*, G for *grabbing*, D for *depositing* and R for *resting*, and the average number of robots in each of these states is respectively N_S, N_H, N_G, N_D and N_R. τ_S, τ_H, τ_G, τ_D and τ_R represent the average times a robot will spend in each state before moving to the next state.

In each time step a robot in state S has probability γ_f of finding a food-item and moving to state G, in which it will move towards the target food-item until it is close enough to grab it using the gripper. Once the robot successfully grabs the food-item it will move to state D, in which the robot moves back to the 'nest' carrying the food-item and deposits it. After the robot has unloaded the food-item it will rest in state R, for τ_R seconds and then move to S to resume *searching*. Meanwhile, if the robot in state S fails to find a food-item within time τ_S, it will move to state H, and

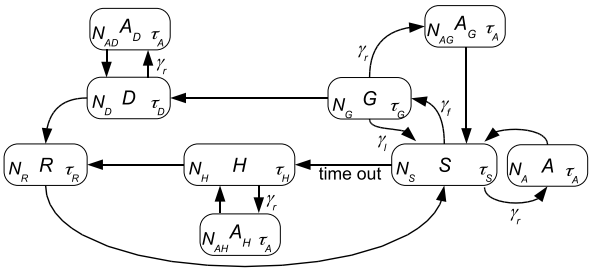

Foraging Robots, Figure 10
Probabilistic Finite State Machine (PFSM) for foraging with division of labor

return to the 'nest' to save energy or minimize interference with other robots. Because of competition among robots more than one robot may see the same food-item and thus move towards it at the same time; clearly only one of them can grab it, a robot in state G therefore has probability γ_l to lose sight of the food-item if it has already been grabbed by another robot, which in turn drives the robot back to state S to resume its search.

In foraging interference between robots because of overcrowding, competition for food-items or simply random collisions is a key aspect of the dynamics of foraging. Thus collision avoidance is modeled as follows. Robots in states S, G, D and H will move to *avoidance* states A, A_G, A_D and A_H respectively with probability γ_r, as shown in Fig. 10. The avoidance behavior then takes τ_A seconds to complete before the robot moves back to its previous state.

Constructing the mathematical model requires two further steps. Firstly, writing down a set of difference equations (DEs) describing the change in the average number of robots in each state from one time step to the next and, secondly, estimating the state transition probabilities. Expressing the PFSM as a set of DEs is relatively straightforward. For instance, the change in the average number of robots N_A in state A from time step k to $k+1$ is given as:

$$N_A(k+1) = N_A(k) + \gamma_r N_S(k) - \gamma_r N_S(k - T_A) \quad (8)$$

where $\gamma_r N_S(k)$ is the number of robots that move from the search to the avoidance state A and $\gamma_r N_S(k - T_A)$ is the number of robots that return to S from state A after time T_A (note T_A is τ_A discretized for time step duration Δt). The full set of DEs is given in [42]. Clearly, the total number of robots in the swarm remains constant from one time step to the next,

$$\begin{aligned} N =& N_S(k) + N_R(k) + N_G(k) + N_D(k) + N_H(k) \\ &+ N_A(k) + N_{A_H}(k) + N_{A_G}(k) + N_{A_D}(k) \end{aligned} \quad (9)$$

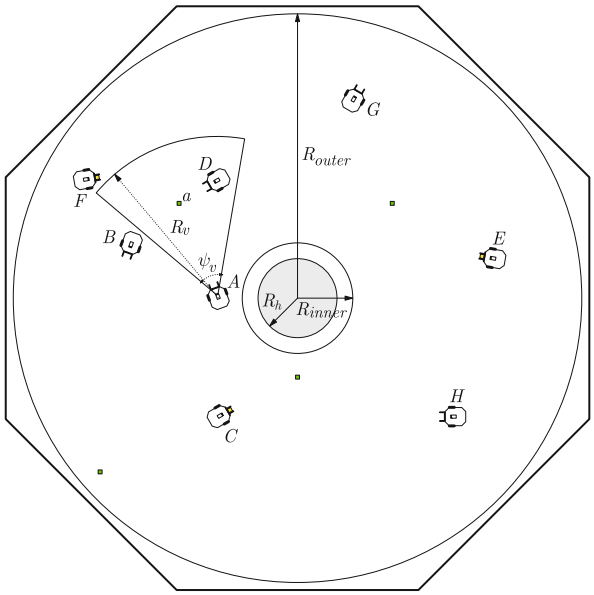

Foraging Robots, Figure 11
Foraging environment showing 8 robots labeled *A* — *H*. The nest region is the *grey circle* with radius R_h at the center. Robot *A* is shown with its arc of vision in which it can sense food items; robots *C*, *E* and *F* have grabbed food items and are in the process of returning to the nest to deposit these. Food items, shown as *small squares*, 'grow' in order to maintain uniform density within the annular region between *circles* with radius R_{inner} and R_{outer}

Estimating state transition probabilities can be challenging but if we simplify the environment by placing the 'nest' region at the center of a circular environment in which the food growing area is bounded by two concentric rings in a bounded arena, as shown in Fig. 11, then a purely geometrical approach can be used to estimate γ_f, γ_r and γ_l together with the average times for grabbing, depositing and homing τ_G, τ_D and τ_H. Clearly τ_R and τ_S are the design parameters we seek to optimize, while τ_A is determined by the physical design of the robot and its sensors.

Figure 12, from [42], plots the average number of robots, from both simulation and the mathematical model, in states *searching*, *resting* and *homing* for the swarm with $\tau_r = 80$. The average number of robots in each state predicted by the probabilistic model quickly settles to a constant value. In contrast, but as one would expect, the average number of robots from simulation oscillates over time but stays near the value predicted by the model.

Figure 13 compares the predicted value of net swarm energy from the mathematical model, with the measured value from simulation, for resting time parameter τ_r increasing from 0 to 200s. The two curves show, firstly a good match between measured and predicted curves

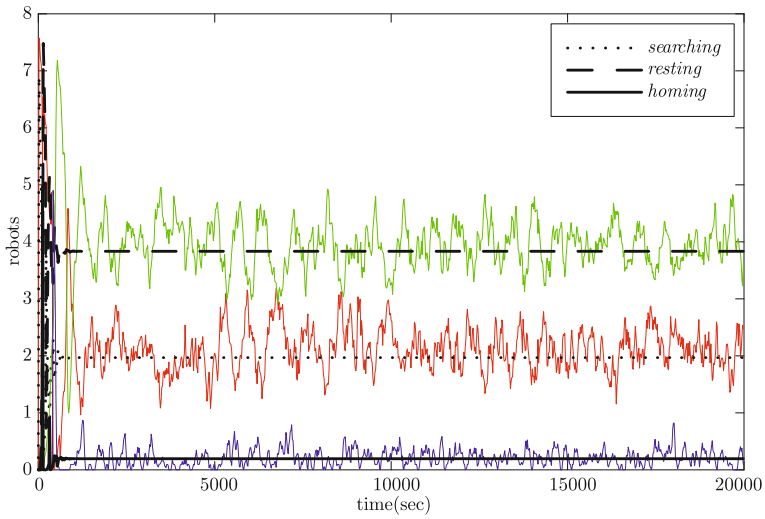

Foraging Robots, Figure 12
The number of robots in states *searching*, *resting* and *homing* for the swarm with $\tau_r = 80$ seconds. The horizontal black dashed lines are predicted by the mathematical model; colored graphs show the instantaneous number of robots measured from simulation

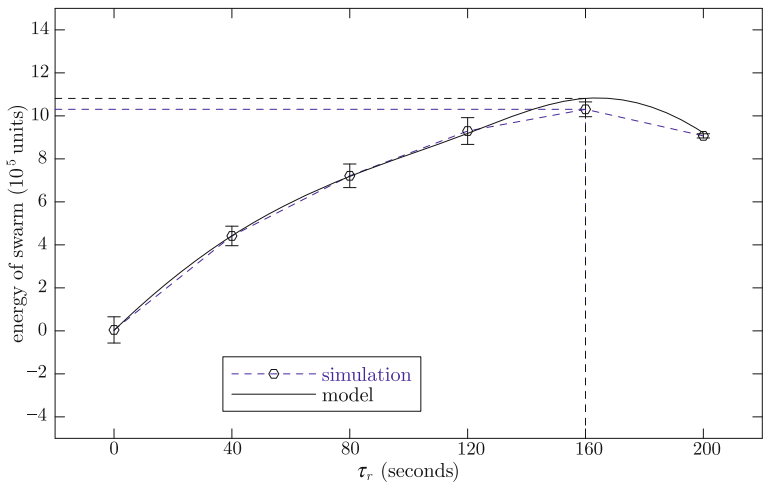

Foraging Robots, Figure 13
The net energy of the swarm for different values of the resting time parameter τ_r. The black curve is the prediction of the mathematical model; the *dashed curve* with error bars is measured from simulation

therefore validating the mathematical model and, secondly, that there is indeed an optimal value for τ_r (at about 160 seconds). We thus have confirmation that a mathematical model can be used to analyze the effect of individual parameters on the overall performance of collective foraging.

Future Directions

This article has defined robot foraging, set out a taxonomy and described both the development and state-of-the-art in robot foraging. Although the principles of robot foraging are well understood, the engineering realization of those principles remains a research problem. Consider multi-robot cooperative robot foraging. Separate aspects have been thoroughly researched and demonstrated, and a number of exemplars have been described in this article. However, to date there has been no demonstration of autonomous multi-robot foraging which integrates self-organized cooperative search, object manipulation and transport in unknown or unstructured real-world environments. Such a demonstration would be a precursor to

a number of compelling real-world applications including search and rescue, toxic waste cleanup or foraging for recycling of materials.

The future directions for foraging robots lie along two separate axes. One axis is the continuing investigation and discovery of foraging algorithms – especially those which seek to mimic biologically inspired principles of self-organization. The other axis is the real-world application of foraging robots and it is here that many key challenges and future directions are to be found. Foraging robot teams are complex systems and the key challenges are in *systems integration and engineering*, which would need to address:

1. Principled design and test methodologies for self-organized multi-robot foraging robot systems.
2. Rigorous methodologies and tools for the specification, analysis and modeling of multi-robot foraging robot systems.
3. Agreed metrics and quantitative benchmarks to allow comparative evaluation of different approaches and systems.
4. Tools and methodologies for provable multi-robot foraging stability, safety and dependability [23,68].

Acknowledgments

The author is indebted to both Wenguo Liu and Guy Théraulaz for case studies, advice and discussion during the preparation of this article.

Bibliography

Primary Literature

1. Arai T, Pagello E, Parker L (2002) Guest editorial: Advances in multirobot systems. IEEE Trans Robotics Autom 18:655–661
2. Arkin RC (1998) Behaviour-Based Robotics. MIT Press, Cambridge
3. Balch T (2002) Taxonomies of multirobot task and reward. In: Balch T, Parker LE (eds) Robot Teams. A K Peters, Wellesley, pp 23–35
4. Balch T, Arkin RC (1994) Communication in reactive multiagent robotic systems. Auton Robots 1:1–25
5. Balch T, Boone G, Collins T, Forbes H, MacKenzie D, Santamaria J (1995) Io, Ganymede and Callisto: A multiagent robot trash-collecting team. AI Magazine 16(2):39–53
6. Beckers R, Holland OE, Deneubourg JL (1994) From local actions to global tasks: Stigmergy and collective robotics. In: Artificial Life IV. MIT Press, Cambridge, pp 181–189
7. Beni G (2005) From swarm intelligence to swarm robotics. In: Şahin E, Spears W (eds) Swarm Robotics Workshop: State-of-the-art Survey, number 3342. Springer, Berlin, pp 1–9
8. Bonabeau E, Dorigo M, Theraulaz G (1999) Swarm Intelligence – From Natural to Artificial Systems. Oxford Univ Press, Oxford
9. Bonabeau E, Theraulaz C, Deneubourg JL (1996) Quantitative study of the fixed threshold model for the regulation of division of labor in insect societies. In: Proceedings of the Royal Society of London, Series B Biological Sciences 263:1565–1569
10. Bonabeau E, Theraulaz G, Deneubourg JL (1998) Fixed response thresholds and the regulation of division of labour in insect societies. Bull Math Biol 60:753–807
11. Braitenberg V (1984) Vehicles – Experiments in Synthetic Psychology. MIT Press, Cambridge
12. Brooks RA (1986) A robust layered control system for a mobile robot. J Robotics Autom 2:14–23
13. Bryson MT, Sukkarieh S (2007) Decentralised trajectory control for multi-UAV SLAM. In: 4th International Symposium on Mechatronics and its Applications (ISMA '07), Sharjah, United Arab Emirates, March 2007
14. Cao YU, Fukunaga AS, Kahng AB (1997) Cooperative mobile robotics: Antecedents and directions. Auton Robots 4:1–23
15. Connell JH (1990) Minimalist Mobile Robotics: A colony-style architecture for an artificial creature. Academic Press, San Diego
16. Şahin E (2005) Swarm robotics: From sources of inspiration to domains of application. In: Şahin E, Spears W (eds) Swarm Robotics Workshop: State-of-the-art Survey, number 3342. Lecture Notes in Computer Science. Springer, Berlin, pp 10–20
17. Deneubourg JL, Goss S, Pasteels JM, Fresneau D, Lachaud JP (1987) Self-organization mechanisms in ant societies (ii): learning in foraging and division of labour. Experientia Suppl 54:177–196
18. Dissanayake MWMG, Newman PM, Durrant-Whyte HF, Clark S, Csorba M (2001) A solution to the simultaneous localization and map building (SLAM) problem. IEEE Trans Robotics Autom 17(3):229–241
19. Dorigo M, Birattari M (2007) Swarm intelligence. Scholarpedia 2(9):1462
20. Dorigo M, Tuci E, Groß T, Trianni V, Labella TH, Nouyan S, Ampatzis C (2005) The SWARM-BOT project. In: Şahin E, Spears W (eds) Swarm Robotics Workshop: State-of-the-art Survey, number 3342. Lecture Notes in Computer Science. Springer, Berlin, pp 31–44
21. Dudek G, Jenkin M, Milios E, Wilkes D (1996) A taxonomy for multi-agent robotics. Auton Robots 3:375–397
22. Everett HR (1995) Sensors for mobile robots: Theory and applications. AK Peters, Wellesley
23. Gazi V, Passino KM (2004) Stability analysis of social foraging swarms. IEEE Trans Syst Man Cybernetics – Part B: Cybernetics 34(1):539–557
24. Gerkey BP, Matarić MJ (2004) A formal analysis and taxonomy of task allocation in multi-robot systems. Int J Robot Res 23(9):939–954
25. Goldberg D, Matarić MJ (1997) Interference as a tool for designing and evaluating multi-robot controllers. In: Proc 14th National conference on Artificial Intelligence (AAAI-97), Providence, July 1997. MIT Press, Cambridge, pp 637–642
26. Greenman J, Holland OE, Kelly I, Kendall K, McFarland D, Melhuish CR (2003) Towards robot autonomy in the natural world: A robot in predator's clothing. Mechatronics 13(3):195–228
27. Groß R, Tuci E, Dorigo M, Bonani M, Mondada F (2006) Object transport by modular robots that self-assemble. In: Proc IEEE International Conference on Robotics and Automation, Orlando, May 2006, pp 2558–2564
28. Hayes AT, Martinoli A, Goodman RMF (2002) Distributed odor

source localization. IEEE Sensors, Special Issue on Artificial Olfaction 2(3):260–271
29. Hicks RW, Hall EL (2000) A survey of robot lawn mowers. In: Casasent DP (ed) Proc SPIE Intelligent Robots and Computer Vision XIX: Algorithms, Techniques, and Active Vision, vol 4197. SPIE, Bellingham, pp 262–269
30. Hölldobler B, Wilson EO (1990) The Ants. Harvard University Press, Cambridge
31. Horn BKP (1986) Robot Vision. MIT Press, Cambridge
32. Ijspeert AJ, Martinoli A, Billard A, Gambardella LM (2001) Collaboration through the exploitation of local interactions in autonomous collective robotics: The stick pulling experiment. Auton Robots 11(2):149–171
33. Kelly I, Holland OE, Melhuish CR (2000) Slugbot: a robotic predator in the natural world. In: 5th Symposium on Artificial Life and Robotics (AROB2000), Oita, January 2000
34. Kondo N, Monta M, Shibano Y, Mohri K (1993) Basic mechanism of robot adapted to physical properties of tomato plant. In: Proc International Conference for Agricultural Machinery and Process Engineering, Seoul, October 1993, vol 3, pp 840–849. The Korean Society for Agricultural Machinery
35. Kondo N, Nakamura M, Monta M, Shibano Y, Mohri K, Arima S (1994) Visual sensor for cucumber harvesting robot. In: Proceedings of the Food Processing Automation Conference, Orlando, February 1994, pp 461–470
36. Krieger M, Billeter JB (2000) The call of duty: Self-organised task allocation in a population of up to twelve mobile robots. J Robotics Auton Syst 30:65–84
37. Labella TH, Dorigo M, Deneubourg JL (2006) Division of labour in a group of robots inspired by ants' foraging behaviour. ACM Trans Auton Adapt Syst 1(1):4–25
38. Lerman K (2002) Mathematical model of foraging in a group of robots: Effect of interference. Auton Robots 13(2):127–141
39. Lerman K, Galstyan A, Martinoli A, Ijspeert AJ (2002) A macroscopic analytical model of collaboration in distributed robotic systems. Artif Life 7:375–393
40. Lerman K, Jones C, Galstyan A, Matarić MJ (2006) Analysis of dynamic task allocation in multi-robot systems. Int J Robot Res 25(3):225–242
41. Lerman K, Martinoli A, Galstyan A (2005) A review of probabilistic macroscopic models for swarm robotic systems. In: Şahin E, Spears W (eds) Swarm Robotics Workshop: State-of-the-art Survey, number 3342. Springer, Berlin, pp 143–152
42. Liu W, Winfield AFT, Sa J (2007) Modelling swarm robotic systems: A case study in collective foraging. In: Towards Autonomous Robotic Systems (TAROS 07), pp 25–32, Aberystwyth, September 2007
43. Liu W, Winfield AFT, Sa J, Chen J, Dou L (2007) Towards energy optimisation: Emergent task allocation in a swarm of foraging robots. Adapt Behav 15(3):289–305
44. Martinoli A, Easton K, Agassounon W (2004) Modeling swarm robotic systems: A case study in collaborative distributed manipulation. Int J Robot Res, Special Issue on Experimental Robotics 23(4):415–436
45. Martinoli A, Ijspeert AJ, Gambardella LM (1999) A probabilistic model for understanding and comparing collective aggregation mechanisms. In: Proc Euro Conf on Artificial Life ECAL'99, Lausanne, September 1999, pp 575–584
46. Matarić MJ (1992) Designing emergent behaviours: From local interactions to collective intelligence. In: From Animals To Animats (SAB-92). MIT Press, Cambridge, pp 432–441
47. Matarić MJ, Marjanovic MJ (1993) Synthesizing complex behaviors by composing simple primitives. In: Proc Self Organization and Life, From Simple Rules to Global Complexity, European Conference on Artificial Life (ECAL-93), pp 698–707, Brussels, May (1993)
48. Melhuish C (1999) Employing secondary swarming with small scale robots: a biologically inspired collective approach. In: Proc of the 2nd Int Conf on Climbing & Walking Robots CLAWAR, Portsmouth, September 1999
49. Melhuish C, Ieropoulos I, Greenman J, Horsfield I (2006) Energetically autonomous robots: Food for thought. Auton Robots 21(3):187–198
50. Mondada F, Gambardella LM, Floreano D, Nolfi S, Deneubourg JL, Dorigo M (2005) The cooperation of Swarm-bots: Physical interactions in collective robotics. IEEE Robotics Autom Mag 12(2):21–28
51. De Nardi R, Holland OE (2007) Ultraswarm: A further step towards a flock of miniature helicopters. In: Second International Workshop on Swarm Robotics at SAB (2006), vol 4433. Springer,Heidelberg, pp 116–128
52. Nembrini J, Winfield AFT, Melhuish C (2002) Minimalist coherent swarming of wireless networked autonomous mobile robots. In: From Animals to Animats SAB'02. MIT Press, Cambridge, pp 373–382
53. Oster GF, Wilson EO (1978) Caste and Ecology in the Social Insects. Princeton University Press, Priceton
54. Østergaard EH, Sukhatme GS, Matarić MJ (2001) Emergent bucket brigading: A simple mechanism for improving performance in multi-robot constrained-space foraging tasks. In: Proc Int Conf on Autonomous Agents, Montreal, May 2001
55. Parker LE (1994) ALLIANCE: an architecture for fault tolerant, cooperative control of heterogeneous mobile robots. In: Proc IEEE/RSJ International Conference on Intelligent Robots and Systems, Munich, September 1994, pp 776–783
56. Pearson MJ, Pipe AG, Melhuish CR, Mitchinson B, Prescott TJ (2007) Whiskerbot: A robotic active touch system modeled on the rat whisker sensory system. Adapt Behav 15:223–240
57. Pilarski T, Happold M, Pangels H, Ollis M, Fitzpatrick K, Stentz A (1999) The Demeter system for automated harvesting. In: Proceedings of the 8th International Topical Meeting on Robotics and Remote Systems, Pittsberg, April 1999
58. Rooks B (2001) Robots reach the home floor. Ind Robot 28(1):27–28
59. Russell A (1993) Mobile robot guidance using a short-lived heat trail. Robotica 11:427–431
60. Russell A (1995) Laying and sensing odor markings as a strategy for assisting mobile robot navigation tasks. IEEE Robotics Autom Mag 2(3):3–9
61. Schenker PS, Huntsberger TL, Pirjanian P, Baumgartner ET, Tunstel E (2003) Planetary rover developments supporting mars exploration, sample return and future human-robotic colonization. Auton Robots 14(2–3):103–126
62. Schilling T (ed) (2000) Telerobotic Applications. Professional Engineering Publishing, London
63. Siegwart RY, Nourbakhsh IR (2004) Introduction to Autonomous Mobile Robots (Intelligent Robotics and Autonomous Agents). Bradford Books, Cambridge
64. Spero DJ, Jarvis RA (2002) Path planning for a mobile robot in a rough terrain environment. In: Third International Workshop on Robot Motion and Control, pp 417–422, Bukowy Dworek, November 2001

65. Støy K (2001) Using situated communication in distributed autonomous mobile robotics. In: 7th Scandinavian Conf on AI. Odense, February 2001, pp 44–52
66. Vertut J, Coiffet P (1986) Teleoperation and Robotics. Prentice Hall, Englewood Cliffs
67. Wettergreen D, Cabrol N, Baskaran V, Calderón F, Heys S, Jonak D, Lüders A, Pane D, Smith T, Teza J, Tompkins P, Villa D, Williams C, Wagner M (2005) Second experiments in the robotic investigation of life in the Atacama desert of Chile. In: Proc International Symposium on Artificial Intelligence, Robotics and Automation in Space, Munich, September 2005
68. Winfield AFT, Harper CJ, Nembrini J (2005) Towards dependable swarms and a new discipline of swarm engineering. In: Şahin E, Spears W (eds) Swarm Robotics Workshop: State-of-the-art Survey, number 3342. Springer, Berlin, pp 126–142
69. Winfield AFT, Holland OE (2000) The application of wireless local area network technology to the control of mobile robots. Microprocess Microsyst 23:597–607
70. Yuh J (1996) Underwater Robots. Kluwer Academic, Boston

Books and Reviews

Balch T, Parker LE (eds) (2002) Robot Teams: From diversity to polymorphism. AK Peters, Wellesley
Bekey GA (2005) Autonomous Robots: From Biological Inspiration to Implementation and Control. MIT Press, Cambridge
Brooks RA (1999) Cambrian Intelligence: The Early History of the New AI. MIT Press, Cambridge
Melhuish CR (2001) Strategies for Collective Minimalist Robotics. Professional Engineering Publishing, London
Nehmzow U (2003) Mobile Robotics: A Practical Introduction. Springer, New York

Fractal Geometry, A Brief Introduction to

Armin Bunde[1], Shlomo Havlin[2]
[1] Institut für Theoretische Physik, Gießen, Germany
[2] Institute of Theoretical Physics, Bar-Ilan-University, Ramat Gan, Israel

Article Outline

Definition of the Subject
Deterministic Fractals
Random Fractal Models
How to Measure the Fractal Dimension
Self-Affine Fractals
Long-Term Correlated Records
Long-Term Correlations in Financial Markets and Seismic Activity
Multifractal Records
Acknowledgments
Bibliography

Definition of the Subject

In this chapter we present some definitions related to the fractal concept as well as several methods for calculating the fractal dimension and other relevant exponents. The purpose is to introduce the reader to the basic properties of fractals and self-affine structures so that this book will be self contained. We do not give references to most of the original works, but, we refer mostly to books and reviews on fractal geometry where the original references can be found.

Fractal geometry is a mathematical tool for dealing with complex systems that have no characteristic length scale. A well-known example is the shape of a coastline. When we see two pictures of a coastline on two different scales, with 1 cm corresponding for example to 0.1 km or 10 km, we cannot tell which scale belongs to which picture: both look the same, and this features characterizes also many other geographical patterns like rivers, cracks, mountains, and clouds. This means that the coastline is scale invariant or, equivalently, has no characteristic length scale. Another example are financial records. When looking at a daily, monthly or annual record, one cannot tell the difference. They all look the same.

Scale-invariant systems are usually characterized by noninteger ("fractal") dimensions. The notion of noninteger dimensions and several basic properties of fractal objects were studied as long ago as the last century by Georg Cantor, Giuseppe Peano, and David Hilbert, and in the beginning of this century by Helge von Koch, Waclaw Sierpinski, Gaston Julia, and Felix Hausdorff. Even earlier traces of this concept can be found in the study of arithmetic-geometric averages by Carl Friedrich Gauss about 200 years ago and in the artwork of Albrecht Dürer (see Fig. 1) about 500 years ago. Georg Friedrich Lichtenberg discovered, about 230 years ago, fractal discharge patterns. He was the first to describe the observed self-similarity of the patterns: A part looks like the whole. Benoit Mandelbrot [1] showed the relevance of fractal geometry to many systems in nature and presented many important features of fractals. For further books and reviews on fractals see [2,3,4,5,6,7,8,9,10,11,12,13,14,15,16].

Before introducing the concept of fractal dimension, we should like to remind the reader of the concept of dimension in regular systems. It is well known that in regular systems (with uniform density) such as long wires, large thin plates, or large filled cubes, the dimension d characterizes how the mass $M(L)$ changes with the linear size L

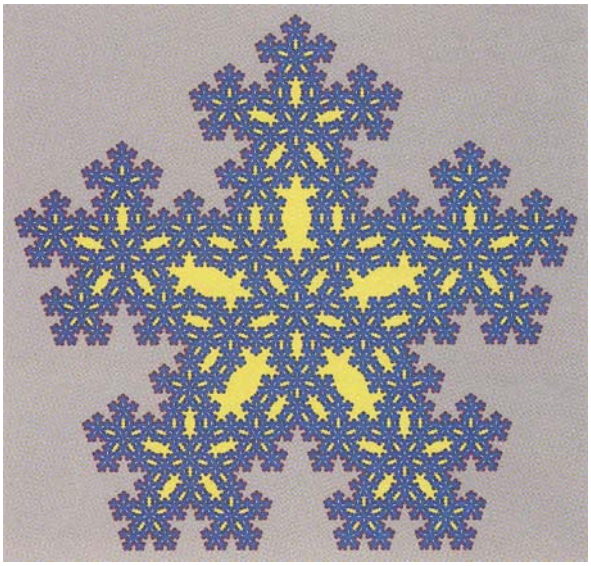

Fractal Geometry, A Brief Introduction to, Figure 1
The Dürer pentagon after five iterations. For the generating rule, see Fig. 8. The Dürer pentagon is in *blue*, its external perimeter is in *red*, Courtesy of M. Meyer

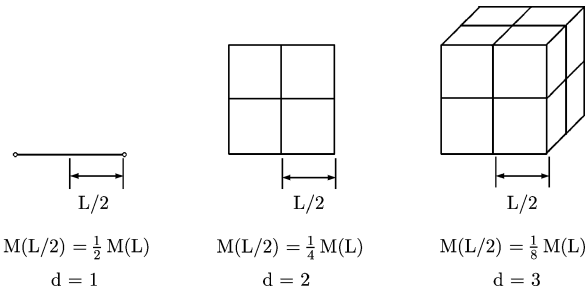

Fractal Geometry, A Brief Introduction to, Figure 2
Examples of regular systems with dimensions $d = 1$, $d = 2$, and $d = 3$

of the system. If we consider a smaller part of the system of linear size bL ($b < 1$), then $M(bL)$ is decreased by a factor of b^d, i. e.,

$$M(bL) = b^d M(L) \,. \tag{1}$$

The solution of the functional equation (1) is simply $M(L) = AL^d$. For the long wire the mass changes linearly with b, i. e., $d = 1$. For the thin plates we obtain $d = 2$, and for the cubes $d = 3$; see Fig. 2.

Next we consider fractal objects. Here we distinguish between deterministic and random fractals. Deterministic fractals are generated iteratively in a deterministic way, while random fractals are generated using a stochastic process. Although fractal structures in nature are random, it is useful to study deterministic fractals where the fractal properties can be determined exactly. By studying deterministic fractals one can gain also insight into the fractal properties of random fractals, which usually cannot be treated rigorously.

Deterministic Fractals

In this section, we describe several examples of deterministic fractals and use them to introduce useful fractal concepts such as fractal and chemical dimension, self similarity, ramification, and fractal substructures (minimum path, external perimeter, backbone, and red bonds).

The Koch Curve

One of the most common deterministic fractals is the Koch curve. Figure 3 shows the first $n = 4$ iterations of this fractal curve. By each iteration the length of the curve is increased by a factor of 4/3. The mathematical fractal is defined in the limit of infinite iterations, $n \to \infty$, where the total length of the curve approaches infinity.

The dimension of the curve can be obtained as for regular objects. From Fig. 3 we notice that, if we decrease the linear size by a factor of $b = 1/3$, the total length (mass) of the curve decreases by a factor of 1/4, i. e.,

$$M\left(\tfrac{1}{3}\,L\right) = \tfrac{1}{4}\,M(L) \,. \tag{2}$$

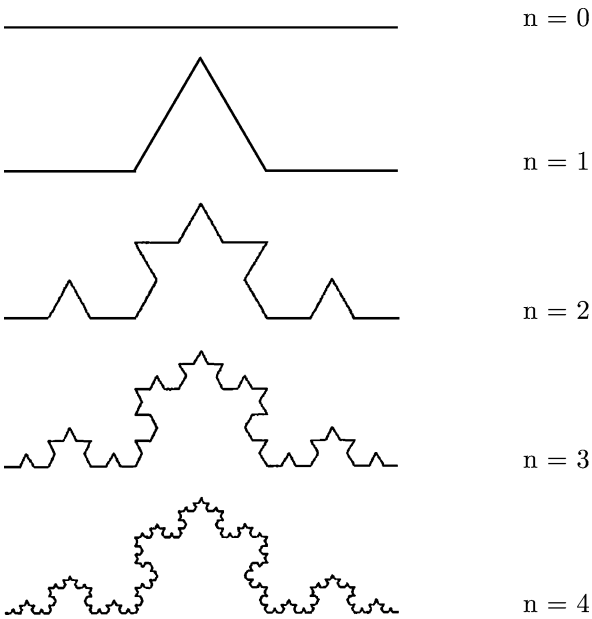

Fractal Geometry, A Brief Introduction to, Figure 3
The first iterations of the Koch curve. The fractal dimension of the Koch curve is $d_f = \log 4/\log 3$

This feature is very different from regular curves, where the length of the object decreases proportional to the linear scale. In order to satisfy Eqs. (1) and (2) we are led to introduce a *noninteger* dimension, satisfying $1/4 = (1/3)^d$, i. e., $d = \log 4/\log 3$. For such non-integer dimensions Mandelbrot coined the name "fractal dimension" and those objects described by a fractal dimension are called fractals. Thus, to include fractal structures, Eq. (1) is generalized by

$$M(bL) = b^{d_f} M(L)\,, \tag{3}$$

and

$$M(L) = AL^{d_f}\,, \tag{4}$$

where d_f is the fractal dimension.

When generating the Koch curve and calculating d_f, we observe the striking property of fractals – the property of *self-similarity*. If we examine the Koch curve, we notice that there is a central object in the figure that is reminiscent of a snowman. To the right and left of this central snowman there are two other snowmen, each being an exact reproduction, only smaller by a factor of 1/3. Each of the smaller snowmen has again still smaller copies (by 1/3) of itself to the right and to the left, etc. Now, if we take any such triplet of snowmen (consisting of $1/3^m$ of the curve), for any m, and magnify it by 3^m, we will obtain exactly the original Koch curve. This property of self-similarity or scale invariance is the basic feature of all deterministic and random fractals: if we take a part of a fractal and magnify it by the same magnification factor in all directions, the magnified picture cannot be distinguished from the original.

For the Koch curve as well as for all deterministic fractals generated iteratively, Eqs. (3) and (4) are of course valid only for length scales L below the linear size L_0 of the whole curve (see Fig. 3). If the number of iterations n is finite, then Eqs. (3) and (4) are valid only above a lower cut off length $L_{\min}$, $L_{\min} = L_0/3^n$ for the Koch curve. Hence, for a finite number of iterations, there exist two cut-off length scales in the system, an upper cut-off $L_{\max} = L_0$ representing the total linear size of the fractal, and a lower cut-off $L_{\min}$. This feature of having two characteristic cut-off lengths is shared by all fractals in nature.

An interesting modification of the Koch curve is shown in Fig. 4, which demonstrates that the *chemical distance* is an important concept for describing structural properties of fractals (for a review see, for example, [16] and Chap. 2 in [13]). The chemical distance ℓ is defined as shortest path on the fractal between two sites of the fractal. In analogy to the fractal dimension d_f that characterizes how the mass of a fractal scales with (air) distance L, we introduce the *chemical dimension* d_ℓ in order to characterize how the mass scales with the chemical distance ℓ,

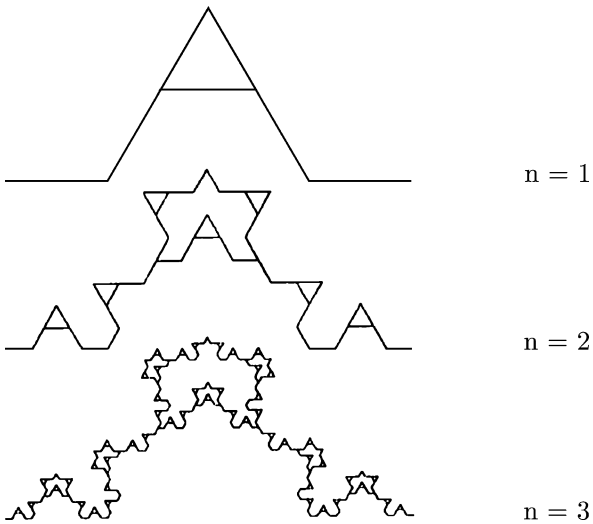

Fractal Geometry, A Brief Introduction to, Figure 4
The first iterations of a modified Koch curve, which has a nontrivial chemical distance metric

$$M(b\ell) = b^{d_\ell} M(\ell)\,, \quad \text{or} \quad M(\ell) = B\ell^{d_\ell}\,. \tag{5}$$

From Fig. 4 we see that if we reduce ℓ by a factor of 5, the mass of the fractal within the reduced chemical distance is reduced by a factor of 7, i. e., $M(1/5\,\ell) = 1/7\,M(\ell)$, yielding $d_\ell = \log 7/\log 5 \cong 1.209$. Note that the chemical dimension is smaller than the fractal dimension $d_f = \log 7/\log 4 \cong 1.404$, which follows from $M(1/4\,L) = 1/7\,M(L)$.

The structure of the shortest path between two sites represents an interesting fractal by itself. By definition, the length of the path is the chemical distance ℓ, and the *fractal dimension of the shortest path*, $d_{\min}$, characterizes how ℓ scales with (air) distance L. Using Eqs. (4) and (5), we obtain

$$\ell \sim L^{d_f/d_\ell} \equiv L^{d_{\min}}\,, \tag{6}$$

from which follows $d_{\min} = d_f/d_\ell$. For our example we find that $d_{\min} = \log 5/\log 4 \cong 1.161$. For the Koch curve, as well as for any linear fractal, one simply has $d_\ell = 1$ and hence $d_{\min} = d_f$. Since, by definition, $d_{\min} \geq 1$, it follows that $d_\ell \leq d_f$ for all fractals.

The Sierpinski Gasket, Carpet, and Sponge

Next we discuss the Sierpinski fractal family: the "gasket", the "carpet", and the "sponge".

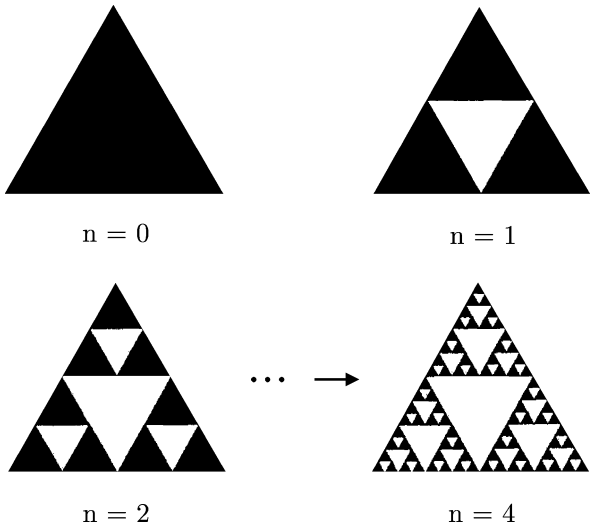

Fractal Geometry, A Brief Introduction to, Figure 5
The Sierpinski gasket. The fractal dimension of the Sierpinski gasket is $d_f = \log 3/\log 2$

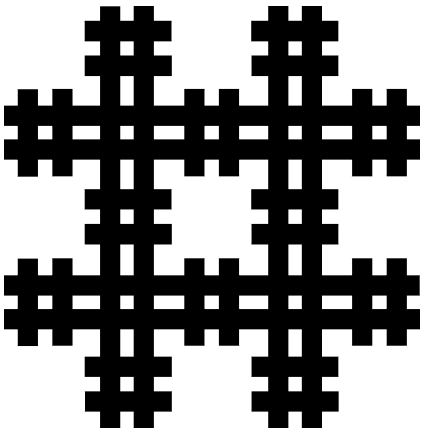

Fractal Geometry, A Brief Introduction to, Figure 6
A Sierpinski carpet with $n = 5$ and $k = 9$. The fractal dimension of this structure is $d_f = \log 16/\log 5$

The Sierpinski Gasket The Sierpinski gasket is generated by dividing a full triangle into four smaller triangles and removing the central triangle (see Fig. 5). In the following iterations, this procedure is repeated by dividing each of the remaining triangles into four smaller triangles and removing the central triangles.

To obtain the fractal dimension, we consider the mass of the gasket within a linear size L and compare it with the mass within $1/2\,L$. Since $M(1/2\,L) = 1/3\,M(L)$, we have $d_f = \log 3/\log 2 \cong 1.585$. It is easy to see that $d_\ell = d_f$ and $d_{\min} = 1$.

The Sierpinski Carpet The Sierpinski carpet is generated in close analogy to the Sierpinski gasket.

Instead of starting with a full triangle, we start with a full square, which we divide into n^2 equal squares. Out of these squares we choose k squares and remove them. In the next iteration, we repeat this procedure by dividing each of the small squares left into n^2 smaller squares and removing those k squares that are located at the same positions as in the first iteration. This procedure is repeated again and again.

Figure 6 shows the Sierpinski carpet for $n = 5$ and the specific choice of $k = 9$. It is clear that the k squares can be chosen in many different ways, and the fractal structures will all look very different. However, since $M(1/n\,L) = 1/(n^2 - k)\,M(L)$ it follows that $d_f = \log(n^2 - k)/\log n$, irrespective of the way the k squares are chosen. Similarly to the gasket, we have $d_\ell = d_f$ and hence $d_{\min} = 1$.

In contrast, the *external perimeter* ("hull", see also Fig. 1) of the carpet and its fractal dimension d_h depend strongly on the way the squares are chosen. The hull consists of those sites of the cluster, which are adjacent to empty sites and are connected with infinity via empty sites. In our example, see Fig. 6, the hull is a fractal with the fractal dimension $d_h = \log 9/\log 5 \cong 1.365$. On the other hand, if a Sierpinski gasket is constructed with the $k = 9$ squares chosen from the center, the external perimeter stays smooth and $d_h = 1$.

Although the rules for generating the Sierpinski gasket and carpet are quite similar, the resulting fractal structures belong to two different classes, to *finitely ramified* and *infinitely ramified* fractals. A fractal is called finitely ramified if any bounded subset of the fractal can be isolated by cutting a *finite* number of bonds or sites. The Sierpinski gasket and the Koch curve are finitely ramified, while the Sierpinski carpet is infinitely ramified. For finitely ramified fractals like the Sierpinski gasket many physical properties, such as conductivity and vibrational excitations, can be calculated exactly. These exact solutions help to provide insight onto the anomalous behavior of physical properties on fractals, as was shown in Chap. 3 in [13].

The Sierpinski Sponge The Sierpinski sponge shown in Fig. 7 is constructed by starting from a cube, subdividing it into $3 \times 3 \times 3 = 27$ smaller cubes, and taking out the central small cube and its six nearest neighbor cubes. Each of the remaining 20 small cubes is processed in the same way, and the whole procedure is iterated ad infinitum. After each iteration, the volume of the sponge is reduced by a factor of 20/27, while the total surface area increases. In the limit of infinite iterations, the surface area is infinite,

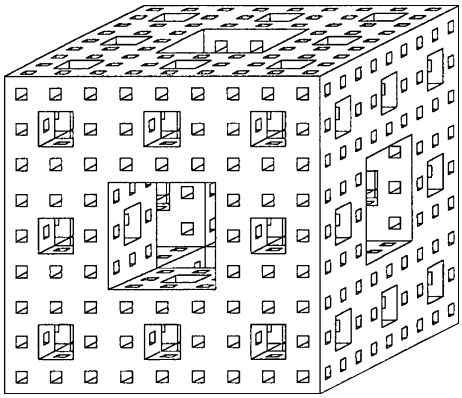

Fractal Geometry, A Brief Introduction to, Figure 7
The Sierpinski sponge (third iteration). The fractal dimension of the Sierpinski sponge is $d_f = \log 20/\log 3$

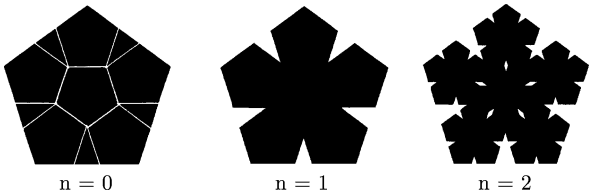

Fractal Geometry, A Brief Introduction to, Figure 8
The first iterations of the Dürer pentagon. The fractal dimension of the Dürer pentagon is $d_f = \log 6/\log(1+g)$

while the volume vanishes. Since $M(1/3\,L) = 1/20\,M(L)$, the fractal dimension is $d_f = \log 20/\log 3 \cong 2.727$. We leave it to the reader to prove that both the fractal dimension d_h of the external surface and the chemical dimension d_ℓ is the same as the fractal dimension d_f.

Modification of the Sierpinski sponge, in analogy to the modifications of the carpet can lead to fractals, where the fractal dimension of the hull, d_h, differs from d_f.

The Dürer Pentagon

Five-hundred years ago the artist Albrecht Dürer designed a fractal based on regular pentagons, where in each iteration each pentagon is divided into six smaller pentagons and five isosceles triangles, and the triangles are removed (see Fig. 8). In each triangle, the ratio of the larger side to the smaller side is the famous *proportio divina* or golden ratio, $g \equiv 1/(2\cos 72°) \equiv (1+\sqrt{5})/2$. Hence, in each iteration the sides of the pentagons are reduced by $1+g$. Since $M(L/(1+g)) = 1/6\,M(L)$, the fractal dimension of the Dürer pentagon is $d_f = \log 6/\log(1+g) \cong 1.862$. The external perimeter of the fractal (see Fig. 1) forms a fractal curve with $d_h = \log 4/\log(1+g)$.

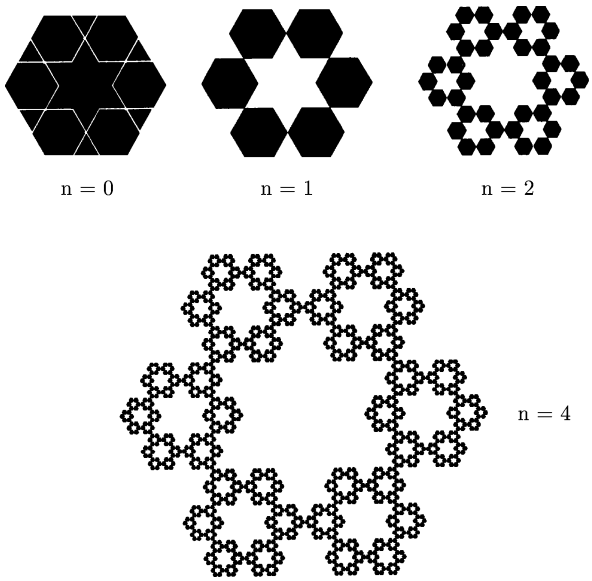

Fractal Geometry, A Brief Introduction to, Figure 9
The first iterations of the David fractal. The fractal dimension of the David fractal is $d_f = \log 6/\log 3$

A nice modification of the Dürer pentagon is a fractal based on regular hexagons, where in each iteration one hexagon is divided into six smaller hexagons, six equilateral triangles, and a David-star in the center, and the triangles and the David-star are removed (see Fig. 9). We leave it as an exercise to the reader to show that $d_f = \log 6/\log 3$ and $d_h = \log 4/\log 3$.

The Cantor Set

Cantor sets are examples of disconnected fractals (*fractal dust*). The simplest set is the triadic Cantor set (see Fig. 10). We divide a unit interval $[0,1]$ into three equal intervals and remove the central one. In each following iteration, each of the remaining intervals is treated in this way. In the limit of $n = \infty$ iterations one obtains a set of points. Since $M(1/3\,L) = 1/2\,M(L)$, we have $d_f = \log 2/\log 3 \cong 0.631$, which is smaller than one.

In chaotic systems, strange fractal attractors occur. The simplest strange attractor is the Cantor set. It occurs, for example, when considering the one-dimensional *logistic map*

$$x_{t+1} = \lambda x_t(1-x_t)\,. \tag{7}$$

The index $t = 0, 1, 2, \ldots$ represents a discrete time. For $0 \le \lambda \le 4$ and x_0 between 0 and 1, the trajectories x_t are bounded between 0 and 1. The dynamical behavior of x_t for $t \to \infty$ depends on the parameter λ. Below $\lambda_1 = 3$,

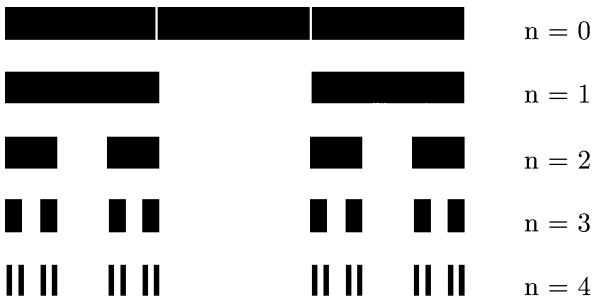

Fractal Geometry, A Brief Introduction to, Figure 10
The first iterations of the triadic Cantor set. The fractal dimension of this Cantor set is $d_f = \log 2/\log 3$

only one stable fixed-point exists to which x_t is attracted. At λ_1, this fixed-point becomes unstable and bifurcates into two new stable fixed-points. At large times, the trajectories move alternately between both fixed-points, and the motion is periodic with period 2. At $\lambda_2 = 1 + \sqrt{6} \cong 3.449$ each of the two fixed-points bifurcates into two new stable fix points and the motion becomes periodic with period 4. As λ is increased, further bifurcation points λ_n occur, with periods of 2^n between λ_n and λ_{n+1}.

For large n, the differences between λ_{n+1} and λ_n become smaller and smaller, according to the law $\lambda_{n+1} - \lambda_n = (\lambda_n - \lambda_{n-1})/\delta$, where $\delta \cong 4.6692$ is the so-called Feigenbaum constant. The Feigenbaum constant is "universal", since it applies to all nonlinear "single-hump" maps with a quadratic maximum [17].

At $\lambda_\infty \cong 3.569\,945\,6$, an infinite period occurs, where the trajectories x_t move in a "chaotic" way between the infinite attractor points. These attractor points define the strange attractor, which forms a Cantor set with a fractal dimension $d_f \cong 0.538$ [18]. For a further discussion of strange attractors and chaotic dynamics we refer to [3,8,9].

The Mandelbrot–Given Fractal

This fractal was suggested as a model for percolation clusters and its substructures (see Sect. 3.4 and Chap. 2 in [13]). Figure 11 shows the first three generations of the Mandelbrot–Given fractal [19]. At each generation, each segment of length a is replaced by 8 segments of length $a/3$. Accordingly, the fractal dimension is $d_f = \log 8/\log 3 \cong 1.893$, which is very close to $d_f = 91/46 \cong 1.896$ for percolation in two dimensions. It is easy to verify that $d_\ell = d_f$, and therefore $d_{\min} = 1$. The structure contains loops, branches, and dangling ends of all length scales.

Imagine applying a voltage difference between two sites at opposite edges of a metallic Mandelbrot–Given fractal: the *backbone* of the fractal consists of those bonds which carry the electric current. The *dangling ends* are those parts of the cluster which carry no current and are connected to the backbone by a single bond only. The *red bonds* (or singly connected bonds) are those bonds that carry the total current; when they are cut the current flow stops. The *blobs*, finally, are those parts of the backbone that remain after the red bonds have been removed.

The backbone of this fractal can be obtained easily by eliminating the dangling ends when generating the fractal (see Fig. 12). It is easy to see that the fractal dimension of the backbone is $d_B = \log 6/\log 3 \cong 1.63$. The red bonds are all located along the x axis of the figure and form a Cantor set with the fractal dimension $d_{red} = \log 2/\log 3 \cong 0.63$.

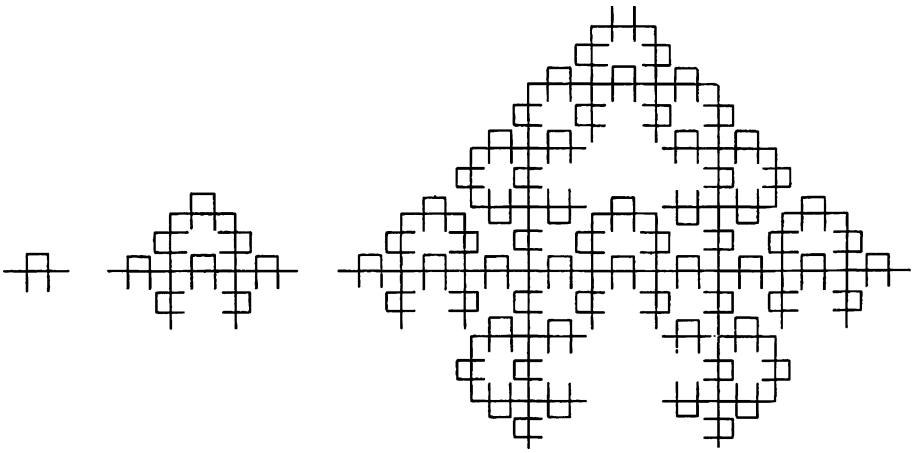

Fractal Geometry, A Brief Introduction to, Figure 11
Three generations of the Mandelbrot–Given fractal. The fractal dimension of the Mandelbrot–Given fractal is $d_f = \log 8/\log 3$

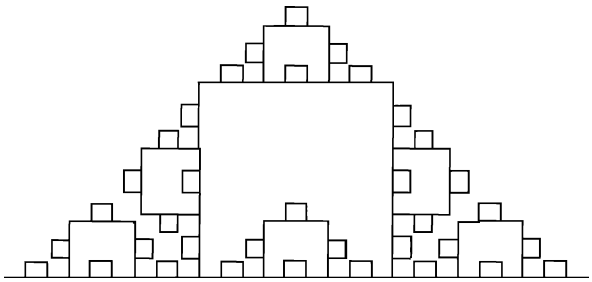

Fractal Geometry, A Brief Introduction to, Figure 12
The backbone of the Mandelbrot–Given fractal, with the red bonds shown in *bold*

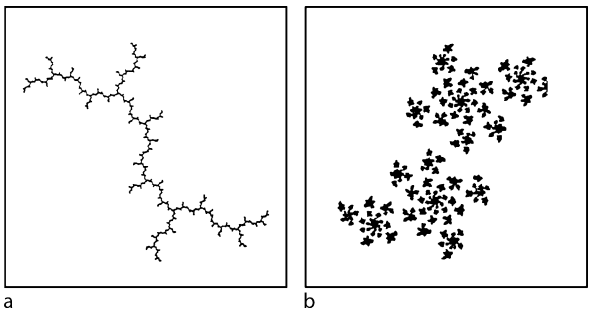

Fractal Geometry, A Brief Introduction to, Figure 13
Julia sets for a $c = i$ and b $c = 0.11031 - i0.67037$. After [9]

Julia Sets and the Mandelbrot Set

A complex version of the logistic map (7) is

$$z_{t+1} = z_t^2 + c \,, \tag{8}$$

where both the trajectories z_t and the constant c are complex numbers. The question is: if a certain c-value is given, for example $c = -1.5652 - i1.03225$, for which initial values z_0 are the trajectories z_t bounded? The set of those values forms the *filled-in Julia set*, and the boundary points of them form the Julia set.

To clarify these definitions, consider the simple case $c = 0$. For $|z_0| > 1$, z_t tends to infinity, while for $|z_0| < 1$, z_t tends to zero. Accordingly, the filled-in Julia set is the set of all points $|z_0| \leq 1$, the Julia set is the set of all points $|z_0| = 1$.

In general, points on the Julia set form a chaotic motion on the set, while points outside the Julia set move away from the set. Accordingly, the Julia set can be regarded as a "repeller" with respect to Eq. (8). To generate the Julia set, it is thus practical to use the inverted transformation

$$z_t = \pm\sqrt{z_{t+1} - c} \,, \tag{9}$$

start with an arbitrarily large value for $t + 1$, and go backward in time. By going backward in time, even points far away from the Julia set are attracted by the Julia set.

For obtaining the Julia set for a given value of c, one starts with some arbitrary value for z_{t+1}, for example, $z_{t+1} = 2$. To obtain z_t, we use Eq. (9), and determine the sign randomly. This procedure is continued to obtain z_{t-1}, z_{t-2}, etc. By disregarding the initial points, e. g., the first 1000 points, one obtains a good approximation of the Julia set.

The Julia sets can be connected (Fig. 13a) or disconnected (Fig. 13b) like the Cantor sets. The self-similarity of the pictures is easy to see. The set of c values that yield connected Julia sets forms the famous Mandelbrot set. It has been shown by Douady and Hubbard [20] that the Mandelbrot set is identical to that set of c values for which z_t converges starting from the initial point $z_0 = 0$. For a detailed discussion with beautiful pictures see [10] and Chaps. 13 and 14 in [3].

Random Fractal Models

In this section we present several random fractal models that are widely used to mimic fractal systems in nature. We begin with perhaps the simplest fractal model, the random walk.

Random Walks

Imagine a random walker on a square lattice or a simple cubic lattice. In one unit of time, the random walker advances one step of length a to a randomly chosen nearest neighbor site. Let us assume that the walker is unwinding a wire, which he connects to each site along his way. The length (mass) M of the wire that connects the random walker with his starting point is proportional to the number of steps n (Fig. 14) performed by the walker.

Since for a random walk in any d-dimensional space the mean end-to-end distance R is proportional to $n^{1/2}$ (for

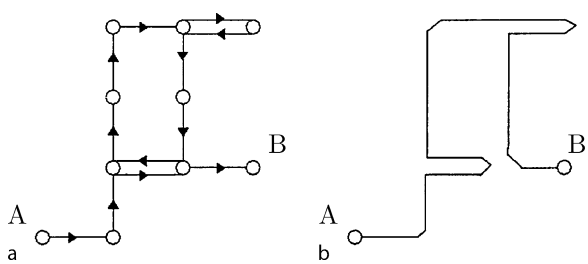

Fractal Geometry, A Brief Introduction to, Figure 14
a A normal random walk with loops. b A random walk without loops

a simple derivation see e. g., Chap. 3 in [13], it follows that $M \sim R^2$. Thus Eq. (4) implies that the fractal dimension of the structure formed by this wire is $d_f = 2$, for all lattices.

The resulting structure has loops, since the walker can return to the same site. We expect the chemical dimension d_ℓ to be 2 in $d = 2$ and to decrease with increasing d, since. Loops become less relevant. For $d \geq 4$ we have $d_\ell = 1$. If we assume, however, that there is no contact between sections of the wire connected to the same site (Fig. 14b), the structure is by definition linear, i. e., $d_\ell = 1$ for all d. For more details on random walks and its relation to Brownian motion, see Chap. 5 in [15] and [21].

Self-Avoiding Walks

Self-avoiding walks (SAWs) are defined as the subset of all nonintersecting random walk configurations. An example is shown in Fig. 15a. As was found by Flory in 1944 [22], the end-to-end distance of SAWs scales with the number of steps n as

$$R \sim n^\nu \,, \tag{10}$$

with $\nu = 3/(d+2)$ for $d \leq 4$ and $\nu = 1/2$ for $d > 4$. Since n is proportional to the mass of the chain, it follows from Eq. (4) that $d_f = 1/\nu$. Self-avoiding walks serve as models for polymers in solution, see [23].

Subsets of SAWs do not necessarily have the same fractal dimension. Examples are the kinetic growth walk (KGW) [24] and the smart growth walk (SGW) [25], sometimes also called the "true" or "intelligent" self-avoiding walk. In the KGW, a random walker can only step on those sites that have not been visited before. Asymptotically, after many steps n, the KGW has the same fractal dimension as SAWs. In $d = 2$, however, the asymptotic regime is difficult to reach numerically, since the random walker can be trapped with high probability (see Fig. 15b). A related structure is the hull of a random walk in $d = 2$. It has been conjectured by Mandelbrot [1] that the fractal dimension of the hull is $d_h = 4/3$, see also [26].

In the SGW, the random walker avoids traps by stepping only at those sites from which he can reach infinity. The structure formed by the SGW is more compact and characterized by $d_f = 7/4$ in $d = 2$ [25]. Related structures with the same fractal dimension are the hull of percolation clusters (see also Sect. "Percolation") and diffusion fronts (for a detailed discussion of both systems see also Chaps. 2 and 7 in [13]).

Kinetic Aggregation

The simplest model of a fractal generated by diffusion of particles is the diffusion-limited aggregation (DLA)

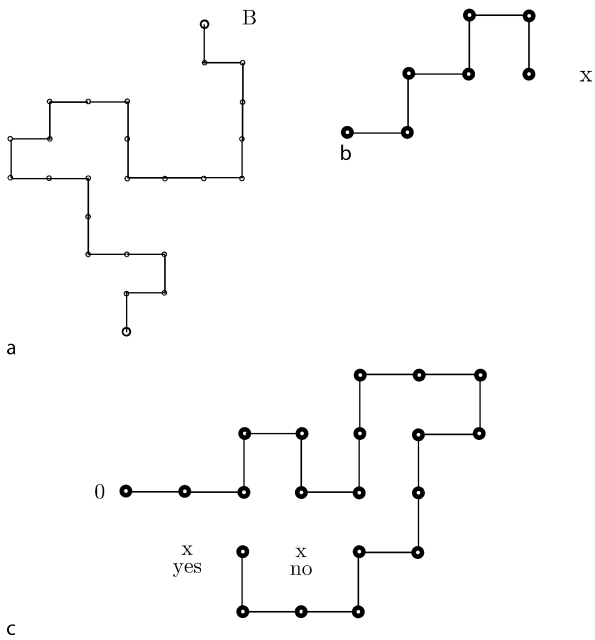

Fractal Geometry, A Brief Introduction to, Figure 15
a A typical self-avoiding walk. b A kinetic growth walk after 8 steps. The available sites are marked by *crosses*. c A smart growth walk after 19 steps. The only available site is marked by "yes"

model, which was introduced by Witten and Sander in 1981 [27]. In the lattice version of the model, a seed particle is fixed at the origin of a given lattice and a second particle is released from a circle around the origin. This particle performs a random walk on the lattice. When it comes to a nearest neighbor site of the seed, it sticks and a cluster (aggregate) of two particles is formed. Next, a third particle is released from the circle and performs a random walk. When it reaches a neighboring site of the aggregate, it sticks and becomes part of the cluster. This procedure is repeated many times until a cluster of the desired number of sites is generated. For saving computational time it is convenient to eliminate particles that have diffused too far away from the cluster (see Fig. 16).

In the continuum (off-lattice) version of the model, the particles have a certain radius a and are not restricted to diffusing on lattice sites. At each time step, the length ($\leq a$) and the direction of the step are chosen randomly. The diffusing particle sticks to the cluster, when its center comes within a distance a of the cluster perimeter. It was found numerically that for off-lattice DLA, $d_f = 1.71 \pm 0.01$ in $d = 2$ and $d_f = 2.5 \pm 0.1$ in $d = 3$ [28,29]. These results may be compared with the mean field result $d_f = (d^2+1)/(d+1)$ [30]. For a renormalization group approach, see [31] and references

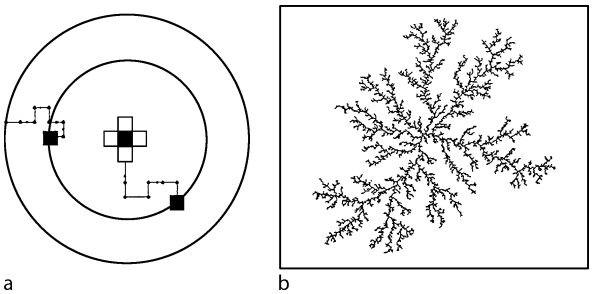

Fractal Geometry, A Brief Introduction to, Figure 16
a Generation of a DLA cluster. The inner release radius is usually a little larger than the maximum distance of a cluster site from the center, the outer absorbing radius is typically 10 times this distance. b A typical off-lattice DLA cluster of 10,000 particles

therein. The chemical dimension d_ℓ is found to be equal to d_f [32].

Diffusion-limited aggregation serves as an archetype for a large number of fractal realizations in nature, including viscous fingering, dielectric breakdown, chemical dissolution, electrodeposition, dendritic and snowflake growth, and the growth of bacterial colonies. For a detailed discussion of the applications of DLA we refer to [5,13], and [29]. Models for the complex structure of DLA have been developed by Mandelbrot [33] and Schwarzer et al. [34].

A somewhat related model for aggregation is the cluster-cluster aggregation (CCA) [35]. In CCA, one starts from a very low concentration of particles diffusing on a lattice. When two particles meet, they form a cluster of two, which can also diffuse. When the cluster meets another particle or another cluster, a larger cluster is formed. In this way, larger and larger aggregates are formed. The structures are less compact than DLA, with $d_f \cong 1.4$ in $d = 2$ and $d_f \cong 1.8$ in $d = 3$. CCA seems to be a good model for smoke aggregates in air and for gold colloids. For a discussion see Chap. 8 in [13].

Percolation

Consider a square lattice, where each site is occupied randomly with probability p or empty with probability $1 - p$. At low concentration p, the occupied sites are either isolated or form small clusters (Fig. 17a). Two occupied sites belong to the same cluster, if they are connected by a path of nearest neighbor occupied sites. When p is increased, the average size of the clusters increases. At a critical concentration p_c (also called the percolation threshold) a large cluster appears which connects opposite edges of the lattice (Fig. 17b). This cluster is called the *infinite* cluster, since its size diverges when the size of the lattice is increased to infinity. When p is increased further, the density of the infinite cluster increases, since more and more sites become part of the infinite cluster, and the average size of the *finite* clusters decreases (Fig. 17c).

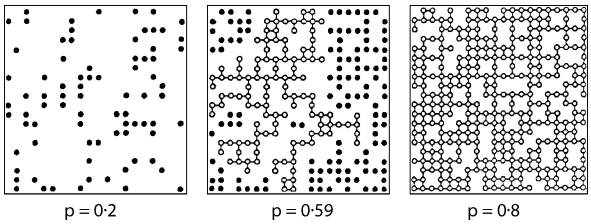

Fractal Geometry, A Brief Introduction to, Figure 17
Square lattice of size 20 × 20. Sites have been randomly occupied with probability p ($p = 0.20, 0.59, 0.80$). Sites belonging to finite clusters are marked by *full circles*, while sites on the infinite cluster are marked by *open circles*

The percolation transition is characterized by the geometrical properties of the clusters near p_c. The probability P_∞ that a site belongs to the infinite cluster is zero below p_c and increases above p_c as

$$P_\infty \sim (p - p_c)^\beta \ . \tag{11}$$

The linear size of the *finite* clusters, below and above p_c, is characterized by the *correlation length* ξ. The correlation length is defined as the mean distance between two sites on the same finite cluster and represents the characteristic length scale in percolation. When p approaches p_c, ξ increases as

$$\xi \sim |p - p_c|^{-\nu} \ , \tag{12}$$

with the same exponent ν below and above the threshold. While p_c depends explicitly on the type of the lattice (e. g., $p_c \cong 0.593$ for the square lattice and 1/2 for the triangular lattice), the *critical exponents* β and ν are universal and depend only on the dimension d of the lattice, but not on the type of the lattice.

Near p_c, on length scales smaller than ξ, both the infinite cluster and the finite clusters are self-similar. Above p_c, on length scales larger than ξ, the infinite cluster can be regarded as an homogeneous system which is composed of many unit cells of size ξ. Mathematically, this can be summarized as

$$M(r) \sim \begin{cases} r^{d_f} \ , & r \ll \xi \ , \\ r^d \ , & r \gg \xi \ . \end{cases} \tag{13}$$

The fractal dimension d_f can be related to β and ν:

$$d_f = d - \frac{\beta}{\nu} \ . \tag{14}$$

Since β and ν are universal exponents, d_f is also universal. One obtains $d_f = 91/48$ in $d = 2$ and $d_f \cong 2.5$ in $d = 3$.

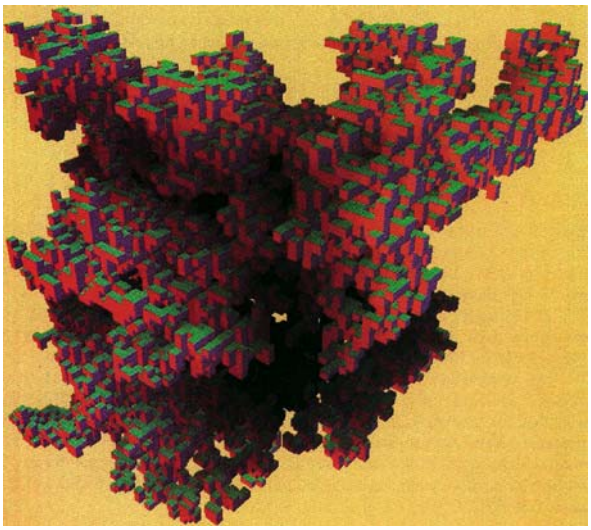

Fractal Geometry, A Brief Introduction to, Figure 18
A large percolation cluster in $d = 3$. The colors mark the topological distance from an arbitrary center of the cluster in the middle of the page. Courtesy of M. Meyer

The chemical dimension d_ℓ is smaller than d_f, $d_\ell \cong 1.15$ in $d = 2$ and $d_\ell \cong 1.33$ in $d = 3$. A large percolation cluster in $d = 3$ is shown in Fig. 18.

Interestingly, a percolation cluster is composed of several fractal sub-structures such as the backbone, dangling ends, blobs, External perimeter, and the red bonds, which are all described by different fractal dimensions.

The percolation model has found applications in physics, chemistry, and biology, where occupied and empty sites may represent very different physical, chemical, or biological properties. Examples are the physics of two component systems (the random resistor, magnetic or superconducting networks), the polymerization process in chemistry, and the spreading of epidemics and forest fires. For reviews with a comprehensive list of references, see Chaps. 2 and 3 of [13] and [36,37,38].

How to Measure the Fractal Dimension

One of the most important "practical" problems is to determine the fractal dimension d_f of either a computer generated fractal or a digitized fractal picture. Here we sketch the two most useful methods: the "sandbox" method and the "box counting" method.

The Sandbox Method

To determine d_f, we first choose one site (or one pixel) of the fractal as the origin for n circles of radii $R_1 < R_2 < \cdots < R_n$, where R_n is smaller than the radius R of the fractal, and count the number of points (pixels) $M_1(R_i)$ within each circle i. (Sometimes, it is more convenient to choose n squares of side length $L_1 \ldots L_n$ instead of the circles.) We repeat this procedure by choosing randomly many other (altogether m) pixels as origins for the n circles and determine the corresponding number of points $M_j(R_i)$, $j = 2, 3, \ldots, m$ within each circle (see Fig. 19a). We obtain the mean number of points $M(R_i)$ within each circle by averaging, $M(R_i) = 1/m \sum_{j=1}^{m} M_j(R_i)$, and plot $M(R_i)$ versus R_i in a double logarithmic plot. The slope of the curve, for large values of R_i, determines the fractal dimension.

In order to avoid boundary effects, the radii must be smaller than the radius of the fractal, and the centers of the circles must be chosen well inside the fractal, so that the largest circles will be well within the fractal. In order to obtain good statistics, one has either to take a very large fractal cluster with many centers of circles or many realizations of the same fractal.

The Box Counting Method

We draw a grid on the fractal that consists of N_1^2 squares, and determine the number of squares $S(N_1)$ that are needed to cover the fractal (see Fig. 19b). Next we choose finer and finer grids with $N_1^2 < N_2^2 < N_3^2 < \cdots < N_m^2$ squares and calculate the corresponding numbers of squares $S(N_1) \ldots S(N_m)$ needed to cover the fractal. Since $S(N)$ scales as

$$S(N) \sim N^{-d_f} , \tag{15}$$

we obtain the fractal dimension by plotting $S(N)$ versus $1/N$ in a double logarithmic plot. The asymptotic slope, for large N, gives d_f.

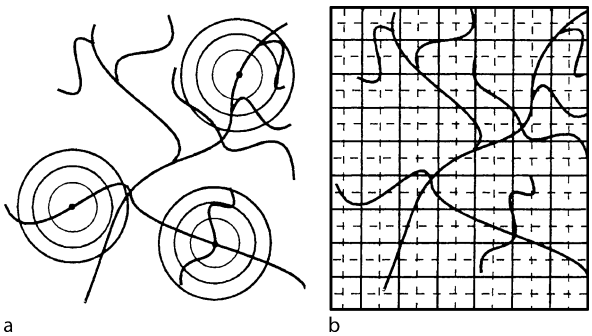

Fractal Geometry, A Brief Introduction to, Figure 19
Illustrations for determining the fractal dimension: a the sandbox method, b the box counting method

Of course, the finest grid size must be larger than the pixel size, so that many pixels can fall into the smallest square. To improve statistics, one should average $S(N)$ over many realizations of the fractal. For applying this method to identify self-similarity in real networks, see Song et al. [39].

Self-Affine Fractals

The fractal structures we have discussed in the previous sections are self-similar: if we cut a small piece out of a fractal and magnify it isotropically to the size of the original, both the original and the magnification look the same. By magnifying isotropically, we have rescaled the x, y, and z axis by the same factor.

There exist, however, systems that are invariant only under *anisotropic* magnifications. These systems are called *self-affine* [1]. A simple model for a self-affine fractal is shown in Fig. 20. The structure is invariant under the anisotropic magnification $x \to 4x$, $y \to 2y$. If we cut a small piece out of the original picture (in the limit of $n \to \infty$ iterations), and rescale the x axis by a factor of four and the y axis by a factor of two, we will obtain exactly the original structure. In other words, if we describe the form of the curve in Fig. 20 by the function $F(x)$, this function satisfies the equation $F(4x) = 2F(x) = 4^{1/2}F(x)$.

In general, if a self-affine curve is scale invariant under the transformation $x \to bx$, $y \to ay$, we have

$$F(bx) = aF(x) \equiv b^H F(x)\,, \tag{16}$$

where the exponent $H = \log a/\log b$ is called the Hurst exponent [1]. The solution of the functional equation (16) is simply $F(x) = Ax^H$. In the example of Fig. 20, $H = 1/2$.

Next we consider random self-affine structures, which are used as models for random surfaces. The simplest structure is generated by a one-dimensional random walk, where the abscissa is the time axis and the ordinate is the displacement $Z(t) = \sum_{i=1}^{t} e_i$ of the walker from its starting point. Here, $e_i = \pm 1$ is the unit step made by the random walker at time t. Since different steps of the random walker are uncorrelated, $\langle e_i e_j \rangle = \delta_{ij}$, it follows that the root mean square displacement $F(t) \equiv \langle Z^2(t)\rangle^{1/2} = t^{1/2}$, and the Hurst exponent of the structure is $H = 1/2$.

Next we assume that different steps i and j are correlated in such a way that $\langle e_i e_j \rangle = b|i-j|^{-\gamma}$, $1 > \gamma \geq 0$. To see how the Hurst exponent depends on γ, we have to evaluate again $\langle Z^2(t)\rangle = \sum_{i,j}^{t} \langle e_i e_j \rangle$. For calculating the double sum it is convenient to introduce the Fourier transform of e_i, $e_\omega = (1/\Omega)^{1/2} \sum_{l=1}^{\Omega} e_l \exp(-i\omega l)$, where Ω is the number of sites in the system. It is easy to verify that $\langle Z^2(t)\rangle$ can be expressed in terms of the power spectrum $\langle e_\omega e_{-\omega}\rangle$ [40]:

$$\langle Z^2(t)\rangle = \frac{1}{\Omega} \sum_{\omega} \langle e_\omega e_{-\omega}\rangle |f(\omega, t)|^2\,, \tag{17a}$$

where

$$f(\omega, t) \equiv \frac{\mathrm{e}^{-i\omega(t+1)} - 1}{\mathrm{e}^{-i\omega} - 1}\,.$$

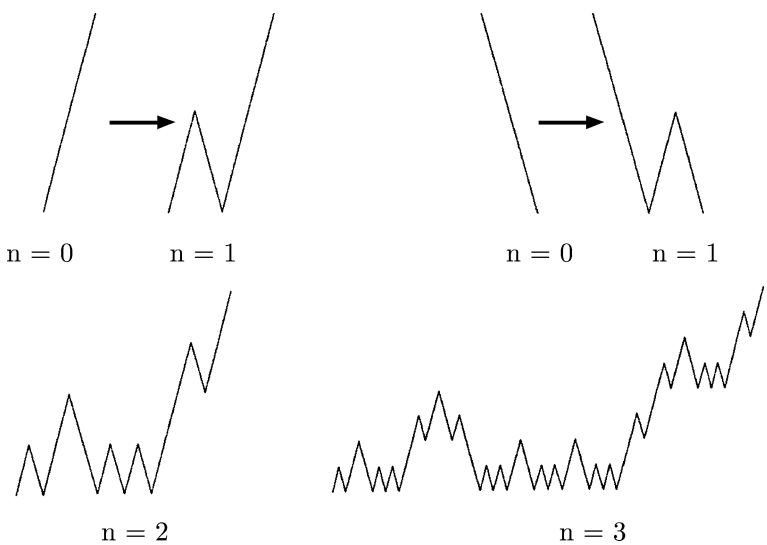

Fractal Geometry, A Brief Introduction to, Figure 20
A simple deterministic model of a self-affine fractal

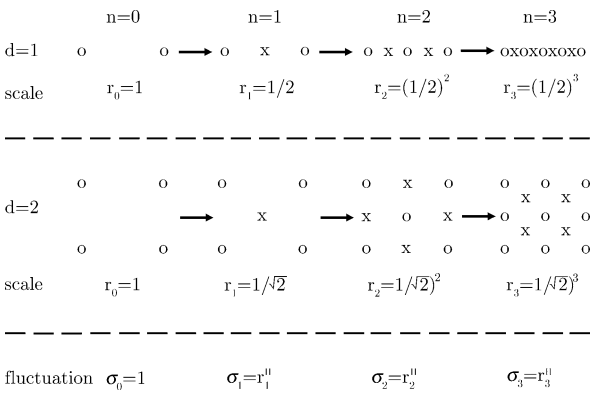

Fractal Geometry, A Brief Introduction to, Figure 21
Illustration of the successive random addition method in $d = 1$ and $d = 2$. The *circles* mark those points that have been considered already in the earlier iterations, the *crosses* mark the new midpoints added at the present iteration. At each iteration n, first the Z values of the midpoints are determined by linear interpolation from the neighboring points, and then random displacements of variance σ_n are added to all Z values

Since the power spectrum scales as

$$\langle e_\omega e_{-\omega} \rangle \sim \omega^{-(1-\gamma)} , \tag{17b}$$

the integration of (17a) yields, for large t,

$$\langle Z^2(t) \rangle \sim t^{2-\gamma} . \tag{17c}$$

Therefore, the Hurst exponent is $H = (2-\gamma)/2$. According to Eq. (17c), for $0 < \gamma < 1$, $\langle Z^2(t) \rangle$ increases faster in time than the uncorrelated random walk. The long-range correlated random walks were called *fractional Brownian motion* by Mandelbrot [1].

There exist several methods to generate correlated random surfaces. We shall describe the *successive random additions* method [41], which iteratively generates the self-affine function $Z(x)$ in the unit interval $0 \le x \le 1$. An alternative method that is detailed in the chapter of Jan Kantelhardt is the Fourier-filtering technique and its variants.

In the $n = 0$ iteration, we start at the edges $x = 0$ and $x = 1$ of the unit interval and choose the values of $Z(0)$ and $Z(1)$ from a distribution with zero mean and variance $\sigma_0^2 = 1$ (see Fig. 21). In the $n = 1$ iteration, we choose the midpoint $x = 1/2$ and determine $Z(1/2)$ by linear interpolation, i. e., $Z(1/2) = (Z(0) + Z(1))/2$, and add to all so-far calculated Z values ($Z(0)$, $Z(1/2)$, and $Z(1)$) random displacements from the same distribution as before, but with a variance $\sigma_1 = (1/2)^H$ (see Fig. 21). In the $n = 2$ iteration we again first choose the midpoints ($x = 1/4$ and $x = 3/4$), determine their Z values by linear interpolation, and add to all so-far calculated Z values random displacements from the same distribution as before, but with a variance $\sigma_2 = (1/2)^{2H}$. In general, in the nth iteration, one first interpolates the Z values of the midpoints and then adds random displacements to all existing Z values, with variance $\sigma_n = (1/2)^{nH}$. The procedure is repeated until the required resolution of the surface is obtained. Figure 22 shows the graphs of three random surfaces generated this way, with $H = 0.2$, $H = 0.5$, and $H = 0.8$.

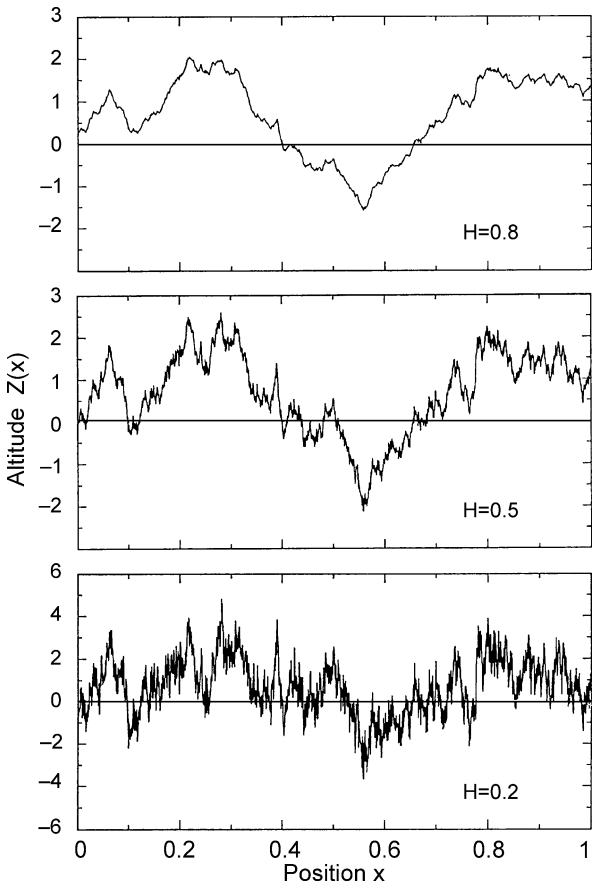

Fractal Geometry, A Brief Introduction to, Figure 22
Correlated random walks with $H = 0.2$, 0.5, and 0.8, generated by the successive random addition method in $d = 1$

The generalization of the successive random addition method to two dimensions is straightforward (see Fig. 21). We consider a function $Z(x, y)$ on the unit square $0 \le x, y \le 1$. In the $n = 0$ iteration, we start with the four corners $(x, y) = (0, 0), (1, 0), (1, 1), (0, 1)$ of the unit square and choose their Z values from a distribution with zero mean and variance $\sigma_0^2 = 1$ (see Fig. 21). In the $n = 1$ iteration, we choose the midpoint at $(x, y) = (1/2, 1/2)$ and determine $Z(1/2, 1/2)$ by linear interpolation, i. e., $Z(1/2, 1/2) = (Z(0, 0) + Z(0, 1) + Z(1, 1) + Z(1, 0))/4$. Then we add to all so far calculated Z-values ($Z(0, 0)$, $Z(0, 1)$,

$Z(1,0)$, $Z(1,1)$ and $Z(1/2,1/2)$) random displacements from the same distribution as before, but with a variance $\sigma_1 = (1/\sqrt{2})^H$ (see Fig. 21). In the $n = 2$ iteration we again choose the midpoints of the five sites $(0,1/2)$, $(1/2,0)$, $(1/2,1)$ and $(1,1/2)$, determine their Z value by linear interpolation, and add to all so far calculated Z values random displacements from the same distribution as before, but with a variance $\sigma_2 = (1/\sqrt{2})^{2H}$. This procedure is repeated again and again, until the required resolution of the surface is obtained.

At the end of this section we like to note that self-similar or self-affine fractal structures with features similar to those fractal models discussed above can be found in nature on all, astronomic as well as microscopic, length scales. Examples include clusters of galaxies (the fractal dimension of the mass distribution is about 1.2 [42]), the crater landscape of the moon, the distribution of earthquakes (see Chap. 2 in [15]), and the structure of coastlines, rivers, mountains, and clouds. Fractal cracks (see, for example, Chap. 5 in [13]) occur on length scales ranging from 10^3 km (like the San Andreas fault) to micrometers (like fractures in solid materials) [44].

Many naturally growing plants show fractal structures, examples range from trees and the roots of trees to cauliflower and broccoli. The patterns of blood vessels in the human body, the kidney, the lung, and some types of nerve cells have fractal features (see Chap. 3 in [15]). In materials sciences, fractals appear in polymers, gels, ionic glasses, aggregates, electro-deposition, rough interfaces and surfaces (see [13] and Chaps. 4 and 6 in [15]), as well as in fine particle systems [43]. In all these structures there is no characteristic length scale in the system besides the physical upper and lower cut-offs.

The occurrence of self-similar or self-affine fractals is not limited to structures in real space as we will discuss in the next section.

Long-Term Correlated Records

Long-range dependencies as described in the previous section do not only occur in surfaces. Of great interest is long-term memory in climate, physiology and financial markets, the examples range from river floods [45,46,47,48,49], temperatures [50,51,52,53,54], and wind fields [55] to market volatilities [56], heart-beat intervals [57,58] and internet traffic [59].

Consider a record x_i of discrete numbers, where the index i runs from 1 to N. x_i may be daily or annual temperatures, daily or annual river flows, or any other set of data consisting of N successive data points. We are interested in the fluctuations of the data around their (sometimes seasonal) mean value. Without loss of generality, we assume that the mean of the data is zero and the variance equal to one. In analogy to the previous section, we call the data long-term correlated, when the corresponding autocorrelation function $C_x(s)$ decays by a power law,

$$C_x(s) = \langle x_i x_{i+s} \rangle \equiv \frac{1}{N-s} \sum_{i=1}^{N-s} x_1 x_{i+s} \sim s^{-\gamma} , \quad (18)$$

where γ denotes the correlation exponent, $0 < \gamma < 1$. Such correlations are named 'long-term' since the mean correlation time $T = \int_0^\infty C_x(s)\,\mathrm{d}s$ diverges in the limit of infinitely long series where $N \to \infty$. If the x_i are uncorrelated, $C_x(s) = 0$ for $s > 0$. More generally, if correlations exist up to a certain correlation time s_x, then $C(s) > 0$ for $s < s_x$ and $C(s) = 0$ for $s > s_x$.

Figure 23 shows parts of an uncorrelated (*left*) and a long-term correlated (*right*) record, with $\gamma = 0.4$; both series have been generated by the computer. The red line is the moving average over 30 data points. For the uncorrelated data, the moving average is close to zero, while for the long-term correlated data set, the moving average can have large deviations from the mean, forming some kind of mountain valley structure. This structure is a consequence of the power-law persistence. The mountains and valleys in Fig. 23b look as if they had been generated by external trends, and one might be inclined to draw a trend-line and to extrapolate the line into the near future for some kind of prognosis. But since the data are trend-free, only a short-term prognosis utilizing the persistence can be made, and not a longer-term prognosis, which often is the aim of such a regression analysis.

Alternatively, in analogy to what we described above for self-affine surfaces, one can divide the data set in K_s equidistant windows of length s and determine in each window ν the squared sum

$$F_\nu^2(s) = \left(\sum_{i=1}^{s} x_i \right)^2 \quad (19a)$$

and detect how the average of this quantity over all windows, $F^2(s) = 1/K_s \sum_{\nu=1}^{K_s} F_\nu^2(s)$, scales with the window size s. For long-term correlated data one can show that $F^2(s)$ scales as $\langle Z^2(t) \rangle$ in the previous section, i. e.

$$F^2(s) \sim s^{2\alpha} , \quad (19b)$$

where $\alpha = 1 - \gamma/2$. This relation represents an alternative way to determine the correlation exponent γ.

Since trends resemble long-term correlations and vice versa, there is a general problem to distinguish between

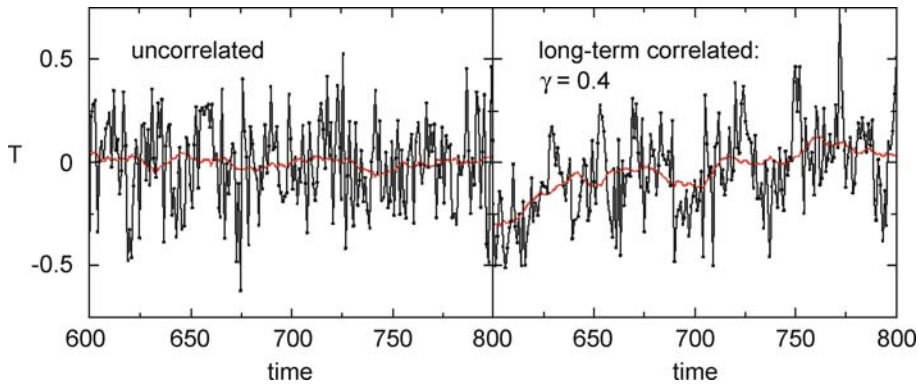

Fractal Geometry, A Brief Introduction to, Figure 23
Comparison of an uncorrelated and a long-term correlated record with $\gamma = 0.4$. The *full line* is the moving average over 30 data points

trends and long-term persistence. In recent years, several methods have been developed, mostly based on the hierarchical detrended fluctuation analysis (DFAn) where long-term correlations in the presence of smooth polynomial trends of order $n-1$ can be detected [57,58,60] (see also ► Fractal and Multifractal Time Series). In DFAn, one considers the cumulated sum ("profile") of the x_i and divides the N data points of the profile into equidistant windows of fixed length s. Then one determines, in each window, the best fit of the profile by an nth order polynomial and determines the variance around the fit. Finally, one averages these variances to obtain the mean variance $F^2_{(n)}$ and the corresponding mean standard deviation (mean fluctuation) $F_{(n)}(s)$. One can show that for long-term correlated trend-free data $F_{(n)}(s)$ scales with the window size s as $F(s)$ in Eq. (19b), i. e., $F_{(n)}(s) \sim s^{\alpha}$, with $\alpha = 1 - \gamma/2$, irrespective of the order of the polynomial n. For short-term correlated records (including the case $\gamma \geq 1$), the exponent is 1/2 for s above s_x. It is easy to verify that trends of order $k-1$ in the original data are eliminated in $F_{(k)}(s)$ but contribute to $F_{(k-1)}$, $F_{(k-2)}$ etc., and this allows one to determine the correlation exponent γ in the presence of trends. For example, in the case of a linear trend, DFA0 and DFA1 (where $F_{(0)}(s)$ and $F_{(1)}(s)$ are determined) are affected by the trend and will exaggerate the asymptotic exponents α, while DFA2, DFA3 etc. (where $F_{(2)}(s)$ and $F_{(3)}(s)$ etc. is determined) are not affected by the trend and will show, in a double logarithmic plot, the same value of α, which then gives immediately the correlation exponent γ. When γ is known this way, one can try to detect the trend, but there is no unique treatment available. In recent papers [61,62,63], different kinds of analysis have been elaborated and applied to estimate trends in the temperature records of the Northern Hemisphere and Siberian locations.

Climate Records

Figure 24 shows representative results of the DFAn analysis, for temperature, precipitation and run-off data. For continental temperatures, the exponent α is around 0.65, while for island stations and sea surface temperatures the exponent is considerably higher. There is no crossover towards uncorrelated behavior at larger time scales. For the precipitation data, the exponent is close to 0.55, not being significantly larger than for uncorrelated records.

Figure 25 shows a summary of the exponent α for a large number of climate records. It is interesting to note that while the distribution of α-values is quite broad for run-off, sea-surface temperature, and precipitation records, the distribution is quite narrow, located around $\alpha = 0.65$ for continental atmospheric temperature records. For the island records, the exponent is larger. The quite universal exponent $\alpha = 0.65$ for continental stations can be used as an efficient test bed for climate models [62,64,65].

The time window accessible by DFAn is typically 1/4 of the length of the record. For instrumental records, the time window is thus restricted to about 50 years. For extending this limit, one has to take reconstructed records or model data, which range up to 2000y. Both have, of course, large uncertainties, but it is remarkable that exactly the same kind of long-term correlations can be found in these data, thus extending the time scale where long-term memory exists to at least 500y [61,62].

Clustering of Extreme Events

Next we consider the consequences of long-term memory on the occurrence of rare events. Understanding (and predicting) the occurrence of extreme events is one of the

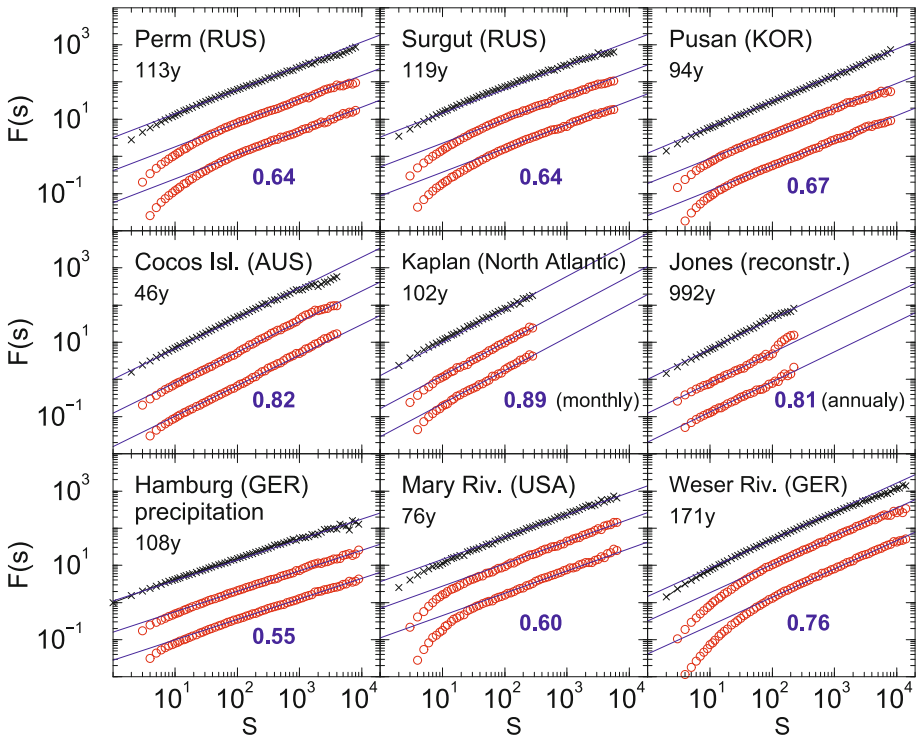

Fractal Geometry, A Brief Introduction to, Figure 24

DFA*n* analysis of six temperature records, one precipitation record and two run-off records. The *black curves* are the DFA0 results, while the *upper red curves* refer to DFA1 and the *lower red curves* to DFA2. The *blue numbers* denote the asymptotic slopes of the curves

major challenges in science (see, e. g., [68]). An important quantity here is the time interval between successive extreme events (see Fig. 26), and by understanding the statistics of these return intervals one aims to better understand the occurrence of extreme events.

Since extreme events are, by definition, very rare and the statistics of their return intervals poor, one usually studies also the return intervals between less extreme events, where the data are above some threshold q and where the statistics is better, and hopes to find some general "scaling" relations between the return intervals at low and high thresholds, which then allows one to extrapolate the results to very large, extreme thresholds (see Fig. 26).

For uncorrelated data, the return intervals are independent of each other and their probability density function (pdf) is a simple exponential, $P_q(r) = (1/R_q) \times \exp(-r/R_q)$. In this case, all relevant quantities can be derived from the knowledge of the mean return interval R_q. Since the return intervals are uncorrelated, a sequential ordering cannot occur. There are many cases, however, where some kind of ordering has been observed where the hazardous events cluster, for example in the floods in Central Europe during the middle ages or in the historic water levels of the Nile river which are shown in Fig. 26 for 663y. Even by eye one can realize that the events are not distributed randomly but are arranged in clusters. A similar clustering was observed for extreme floods, winter storms, and avalanches in Central Europe (see, e. g., Figs. 4.4, 4.7, 4.10, and 4.13 in [69], Fig. 66 in [70], and Fig. 2 in [71]). The reason for this clustering is the long-term memory.

Figure 27 shows $P_q(r)$ for long-term correlated records with $\alpha = 0.4$ (corresponding to $\gamma = 0.8$), for three values of the mean return interval R_q (which is easily obtained from the threshold q and independent of the correlations). The pdf is plotted in a scaled way, i. e., $R_q P_q(r)$ as a function of r/R_q. The figure shows that all three curves collapse. Accordingly, when we know the functional form of the pdf for one value of R_q, we can easily deduce its functional form also for very large R_q values which due to its poor statistics cannot be obtained directly from the data. This scaling is a very important property, since it allows one to make predictions also for rare events which otherwise are not accessible with meaningful statistics. When the data

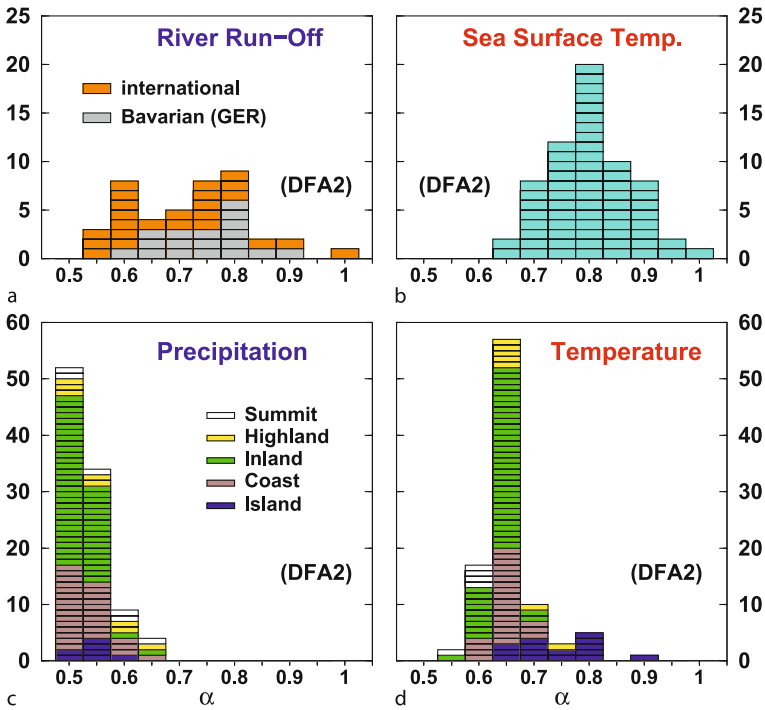

Fractal Geometry, A Brief Introduction to, Figure 25
Distribution of fluctuation exponents α for several kinds of climate records (from [53,66,67])

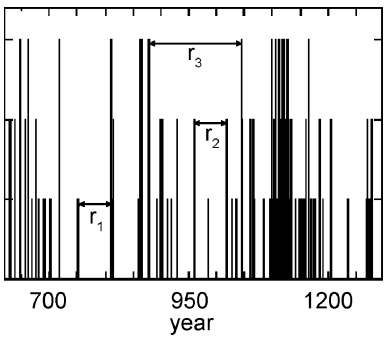

Fractal Geometry, A Brief Introduction to, Figure 26
Illustration of the return intervals for three equidistant threshold values q_1, q_2, q_3 for the water levels of the Nile at Roda (near Cairo, Egypt). One return interval for each threshold (quantile) q is indicated by *arrows*

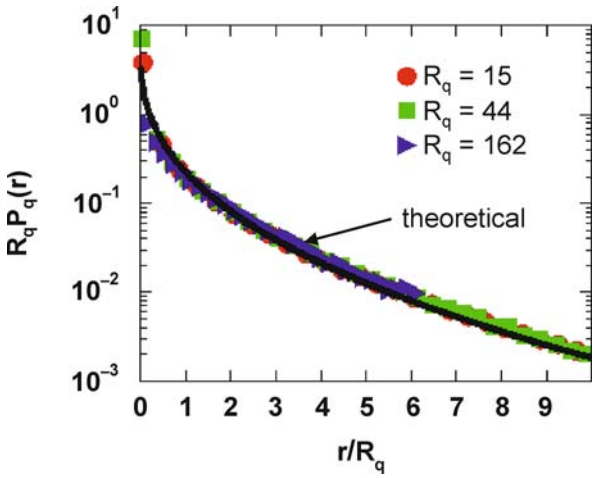

Fractal Geometry, A Brief Introduction to, Figure 27
Probability density function of the return intervals in long-term correlated data, for three different return periods R_q, plotted in a scaled way. The *full line* is a stretched exponential, with exponent $\gamma = 0.4$ (after [73])

are shuffled, the long-term correlations are destroyed and the pdf becomes a simple exponential.

The functional form of the pdf is a quite natural extension of the uncorrelated case. The figure suggests that

$$\ln P_q(r) \sim -(r/R_q)^{\gamma} \tag{20}$$

i. e. simple stretched exponential behavior [72,73]. For γ approaching 1, the long-term correlations tend to vanish and we obtain the simple exponential behavior characteristic for uncorrelated processes. For r well below R_q, however, there are deviations from the pure stretched exponential behavior. Closer inspection of the data shows that

for $r/R_q \ll$ the decay of the pdf is characterized by a power law, with the exponent $\gamma - 1$. This overall behavior does not depend crucially on the way the original data are distributed. In the cases shown here, the data had a Gaussian distribution, but similar results have been obtained also for exponential, power-law and log-normal distributions [74]. Indeed, the characteristic stretched exponential behavior of the pdf can also be seen in long historic and reconstructed records [73].

The form of the pdf indicates that return intervals both well below and well above their average value are considerably more frequent for long-term correlated data than for uncorrelated data. The distribution does not quantify, however, if the return intervals themselves are arranged in a correlated or in an uncorrelated fashion, and if clustering of rare events may be induced by long-term correlations.

To study this question, [73] and [74] have evaluated the autocorrelation function of the return intervals in synthetic long-term correlated records. They found that also the return intervals are arranged in a long-term correlated fashion, with the same exponent as the original data. Accordingly, a large return interval is more likely to be followed by a large one than by a short one, and a small return interval is more likely to be followed by a small one than by a large one, and this leads to clustering of events above some threshold q, including extreme events.

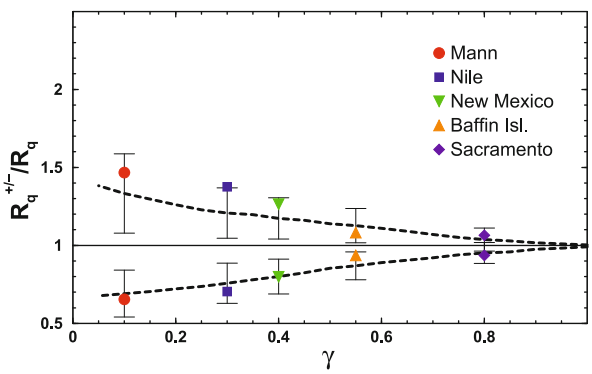

Fractal Geometry, A Brief Introduction to, Figure 28
Mean of the (conditional) return intervals that either follow a return interval below the median (*lower dashed line*) or above the median (*upper dashed line*), as a function of the correlation exponent γ, for five long reconstructed and natural climate records. The theoretical curves are compared with the corresponding values of the climate records (from *right* to *left*): The reconstructed run-offs of the Sacramento River, the reconstructed temperatures of Baffin Island, the reconstructed precipitation record of New Mexico, the historic water levels of the Nile and one of the reconstructed temperature records of the Northern hemisphere (Mann record) (after [73])

As a consequence of the long-term memory, the probability of finding a certain return interval depends on the preceding interval. This effect can be easily seen in synthetic data sets generated numerically, but not so well in climate records where the statistics is comparatively poor. To improve the statistics, we now only distinguish between two kinds of return intervals, "small" ones (below the median) and "large" ones (above the median), and determine the mean R_q^+ and R_q^- of those return intervals following a large (+) or a small (−) return interval. Due to scaling, R_q^+/R_q and R_q^-/R_q are independent of q. Figure 28 shows both quantities (calculated numerically for long-term correlated Gaussian data) as a function of the correlation exponent γ. The lower dashed line is R_q^-/R_q, the upper dashed line is R_q^+/R_q. In the limit of vanishing long-term memory, for $\gamma = 1$, both quantities coincide, as expected. Figure 28 also shows R_q^+/R_q and R_q^-/R_q for five climate records with different values of γ. One can see that the data agree very well, within the error bars, with the theoretical curves.

Long-Term Correlations in Financial Markets and Seismic Activity

The characteristic behavior of the return intervals, i. e. long-term correlations and stretched exponential decay, can also be observed in financial markets and seismic activity. It is well known (see, e. g. [56]) that the volatility of stocks and exchange rates is long-term correlated. Figure 29 shows that, as expected from the foregoing, also the return intervals between daily volatilities are long-term correlated, with roughly the same exponent γ as the original data [75]. It has further been shown [75] that also the pdfs show the characteristic behavior predicted above.

A further example where long-term correlations seem to play an important role, are earthquakes in certain bounded areas (e. g. California) in time regimes where the seismic activity is (quasi) stationary. It has been discovered recently by [76] that the magnitudes of earthquakes in Northern and Southern California, from 1995 until 1998, are long-term correlated with an exponent around $\gamma = 0.4$, and that also the return intervals between the earthquakes are long-term correlated with the same exponent. For the given exponential distribution of the earthquake magnitudes (following the Gutenberg–Richter law), the long-term correlations lead to a characteristic dependence on the scaled variable r/R_q which can explain, without any fit parameter, the previous results on the pdf of the return intervals by [77].

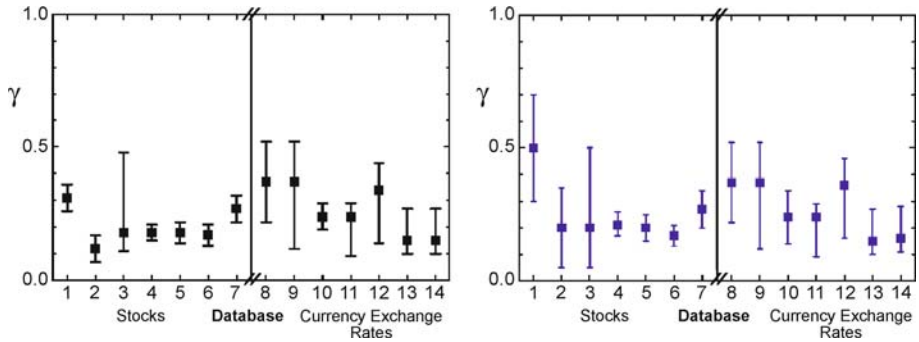

Fractal Geometry, A Brief Introduction to, Figure 29

Long-term correlation exponent γ for the daily volatility (*left*) and the corresponding return intervals (*right*). The studied commodities are (from *left* to *right*), the S&P 500 index, six stocks (IBM, DuPont, AT&T, Kodak, General Electric, Coca-Cola) and seven currency exchange rates (US$ vs. Japanese Yen, British Pound vs. Swiss Franc, US$ vs. Swedish Krona, Danish Krone vs. Australian $, Danish Krone vs. Norwegian Krone, US$ vs. Canadian $ and US$ vs. South African $). Courtesy of Lev Muchnik

Multifractal Records

Many records do not exhibit a simple monofractal scaling behavior, which can be accounted for by a single scaling exponent. In some cases, there exist crossover (time-) scales s_x separating regimes with different scaling exponents, e. g. long-term correlations on small scales below s_x and another type of correlations or uncorrelated behavior on larger scales above s_x. In other cases, the scaling behavior is more complicated, and different scaling exponents are required for different parts of the series. In even more complicated cases, such different scaling behavior can be observed for many interwoven fractal subsets of the time series. In this case a multitude of scaling exponents is required for a full description of the scaling behavior, and a multifractal analysis must be applied (see, e. g., [78,79] and literature therein).

To see this, it is meaningful to extend Eqs. (19a) and (19b) by considering the more general average

$$F^q(s) = \frac{1}{K_s} \sum_{\nu=1}^{K_s} \left[F_\nu^2(s)\right]^{q/2} \tag{21}$$

with q between $-\infty$ and $+\infty$. For $q \ll -1$ the small fluctuations will dominate the sum, while for $q \gg 1$ the large fluctuations are dominant. It is reasonable to assume that the q-dependent average scales with s as

$$F^q(s) \sim s^{q\beta(q)} , \tag{22}$$

with $\beta(2) = \alpha$. Equation (22) generalizes Eq. (19b). If $\beta(q)$ is independent of q, we have $(F^q(s))^{1/q} \sim s^\alpha$, independent of q, and both large and small fluctuations scale the same. In this case, a single exponent is sufficient to characterize the record, which then is referred to as *monofractal*. If $\beta(q)$ is not identical to α, we have a *multifractal* [1,4,12]. In this case, the dependence of $\beta(q)$ on q characterizes the record. Instead of $\beta(q)$ one considers frequently the spectrum $f(\omega)$ that one obtains by Legendre transform of $q\beta(q)$, $\omega = \mathrm{d}(q\beta(q))/\mathrm{d}q$, $f(\omega) = q\omega - q\beta(q) + 1$. In the monofractal limit we have $f(\omega) = 1$.

For generating multifractal data sets, one considers mostly multiplicative random cascade processes, described, e. g., in [3,4]. In this process, the data set is obtained in an iterative way, where the length of the record doubles in each iteration. It is possible to generate random cascades with vanishing autocorrelation function ($C_x(s) = 0$ for $s \geq 1$) or with algebraically decaying autocorrelation functions ($C_x(s) \sim s^{-\gamma}$). Here we focus on a multiplicative random cascade with vanishing autocorrelation function, which is particularly interesting since it can be used as a model for the arithmetic returns $(P_i - P_{i-1})/P_i$ of daily stock closing prices P_i [80]. In the zeroth iteration $n = 0$, the data set (x_i) consists of one value, $x_1^{(n=0)} = 1$. In the nth iteration, the data $x_i^{(n)}$ consist of 2^n values that are obtained from

$$x_{2l-1}^{(n)} = x_l^{(n-1)} m_{2l-1}^{(n)}$$

and

$$x_{2l}^{(n)} = x_l^{(n-1)} m_{2l}^{(n)} , \tag{23}$$

where the multipliers m are independent and identically distributed (i.i.d.) random numbers with zero mean and unit variance. The resulting pdf is symmetric with log-normal tails, with vanishing correlation function $C_x(s)$ for $s \geq 1$.

It has been shown that in this case, the pdf of the return intervals decays by a power-law

$$P_q(r) \sim \left(\frac{r}{R_q}\right)^{-\delta(q)} , \tag{24}$$

where the exponent δ depends explicitly on R_q and seems to converge to a limiting curve for large data sets. Despite of the vanishing autocorrelation function of the original record, the autocorrelation function of the return intervals decays by a power law with a threshold-dependent exponent [80]. Obviously, these long-term correlations have been induced by the nonlinear correlations in the multifractal data set. Extracting the return interval sequence from a data set is a nonlinear operation, and thus the return intervals are influenced by the nonlinear correlations in the original data set. Accordingly, the return intervals in data sets without linear correlations are sensitive indicators for nonlinear correlations in the data records. The power-law dependence of $P_q(r)$ can be used for an improved risk estimation. Both power-law dependencies can be observed in economic and physiological records that are known to be multifractal [81].

Acknowledgments

We like to thank all our coworkers in this field, in particular Eva Koscielny-Bunde, Mikhail Bogachev, Jan Kantelhardt, Jan Eichner, Diego Rybski, Sabine Lennartz, Lev Muchnik, Kazuko Yamasaki, John Schellnhuber and Hans von Storch.

Bibliography

1. Mandelbrot BB (1977) Fractals: Form, chance and dimension. Freeman, San Francisco; Mandelbrot BB (1982) The fractal geometry of nature. Freeman, San Francisco
2. Jones H (1991) Part 1: 7 chapters on fractal geometry including applications to growth, image synthesis, and neutral net. In: Crilly T, Earschaw RA, Jones H (eds) Fractals and chaos. Springer, New York
3. Peitgen H-O, Jürgens H, Saupe D (1992) Chaos and fractals. Springer, New York
4. Feder J (1988) Fractals. Plenum, New York
5. Vicsek T (1989) Fractal growth phenomena. World Scientific, Singapore
6. Avnir D (1992) The fractal approach to heterogeneous chemistry. Wiley, New York
7. Barnsley M (1988) Fractals everywhere. Academic Press, San Diego
8. Takayasu H (1990) Fractals in the physical sciences. Manchester University Press, Manchester
9. Schuster HG (1984) Deterministic chaos – An introduction. Physik Verlag, Weinheim
10. Peitgen H-O, Richter PH (1986) The beauty of fractals. Springer, Heidelberg
11. Stanley HE, Ostrowsky N (1990) Correlations and connectivity: Geometric aspects of physics, chemistry and biology. Kluwer, Dordrecht
12. Peitgen H-O, Jürgens H, Saupe D (1991) Chaos and fractals. Springer, Heidelberg
13. Bunde A, Havlin S (1996) Fractals and disordered systems. Springer, Heidelberg
14. Gouyet J-F (1992) Physique et structures fractales. Masson, Paris
15. Bunde A, Havlin S (1995) Fractals in science. Springer, Heidelberg
16. Havlin S, Ben-Avraham D (1987) Diffusion in disordered media. Adv Phys 36:695; Ben-Avraham D, Havlin S (2000) Diffusion and reactions in fractals and disordered systems. Cambridge University Press, Cambridge
17. Feigenbaum M (1978) Quantitative universality for a class of non-linear transformations. J Stat Phys 19:25
18. Grassberger P (1981) On the Hausdorff dimension of fractal attractors. J Stat Phys 26:173
19. Mandelbrot BB, Given J (1984) Physical properties of a new fractal model of percolation clusters. Phys Rev Lett 52:1853
20. Douady A, Hubbard JH (1982) Itération des polynômes quadratiques complex. CRAS Paris 294:123
21. Weiss GH (1994) Random walks. North Holland, Amsterdam
22. Flory PJ (1971) Principles of polymer chemistry. Cornell University Press, New York
23. De Gennes PG (1979) Scaling concepts in polymer physics. Cornell University Press, Ithaca
24. Majid I, Jan N, Coniglio A, Stanley HE (1984) Kinetic growth walk: A new model for linear polymers. Phys Rev Lett 52:1257; Havlin S, Trus B, Stanley HE (1984) Cluster-growth model for branched polymers that are "chemically linear". Phys Rev Lett 53:1288; Kremer K, Lyklema JW (1985) Kinetic growth models. Phys Rev Lett 55:2091
25. Ziff RM, Cummings PT, Stell G (1984) Generation of percolation cluster perimeters by a random walk. J Phys A 17:3009; Bunde A, Gouyet JF (1984) On scaling relations in growth models for percolation clusters and diffusion fronts. J Phys A 18:L285; Weinrib A, Trugman S (1985) A new kinetic walk and percolation perimeters. Phys Rev B 31:2993; Kremer K, Lyklema JW (1985) Monte Carlo series analysis of irreversible self-avoiding walks. Part I: The indefinitely-growing self-avoiding walk (IGSAW). J Phys A 18:1515; Saleur H, Duplantier B (1987) Exact determination of the percolation hull exponent in two dimensions. Phys Rev Lett 58:2325
26. Arapaki E, Argyrakis P, Bunde A (2004) Diffusion-driven spreading phenomena: The structure of the hull of the visited territory. Phys Rev E 69:031101
27. Witten TA, Sander LM (1981) Diffusion-limited aggregation, a kinetic critical phenomenon. Phys Rev Lett 47:1400
28. Meakin P (1983) Diffusion-controlled cluster formation in two, three, and four dimensions. Phys Rev A 27:604,1495
29. Meakin P (1988) In: Domb C, Lebowitz J (eds) Phase transitions and critical phenomena, vol 12. Academic Press, New York, p 335
30. Muthukumar M (1983) Mean-field theory for diffusion-limited cluster formation. Phys Rev Lett 50:839; Tokuyama M, Kawasaki K (1984) Fractal dimensions for diffusion-limited aggregation. Phys Lett A 100:337
31. Pietronero L (1992) Fractals in physics: Applications and theoretical developments. Physica A 191:85

32. Meakin P, Majid I, Havlin S, Stanley HE (1984) Topological properties of diffusion limited aggregation and cluster-cluster aggregation. Physica A 17:L975
33. Mandelbrot BB (1992) Plane DLA is not self-similar; is it a fractal that becomes increasingly compact as it grows? Physica A 191:95; see also: Mandelbrot BB, Vicsek T (1989) Directed recursive models for fractal growth. J Phys A 22:L377
34. Schwarzer S, Lee J, Bunde A, Havlin S, Roman HE, Stanley HE (1990) Minimum growth probability of diffusion-limited aggregates. Phys Rev Lett 65:603
35. Meakin P (1983) Formation of fractal clusters and networks by irreversible diffusion-limited aggregation. Phys Rev Lett 51:1119; Kolb M (1984) Unified description of static and dynamic scaling for kinetic cluster formation. Phys Rev Lett 53:1653
36. Stauffer D, Aharony A (1992) Introduction to percolation theory. Taylor and Francis, London
37. Kesten H (1982) Percolation theory for mathematicians. Birkhauser, Boston
38. Grimmet GR (1989) Percolation. Springer, New York
39. Song C, Havlin S, Makse H (2005) Self-similarity of complex networks. Nature 433:392
40. Havlin S, Blumberg-Selinger R, Schwartz M, Stanley HE, Bunde A (1988) Random multiplicative processes and transport in structures with correlated spatial disorder. Phys Rev Lett 61:1438
41. Voss RF (1985) In: Earshaw RA (ed) Fundamental algorithms in computer graphics. Springer, Berlin, p 805
42. Coleman PH, Pietronero L (1992) The fractal structure of the universe. Phys Rep 213:311
43. Kaye BH (1989) A random walk through fractal dimensions. Verlag Chemie, Weinheim
44. Turcotte DL (1997) Fractals and chaos in geology and geophysics. Cambridge University Press, Cambridge
45. Hurst HE, Black RP, Simaika YM (1965) Long-term storage: An experimental study. Constable, London
46. Mandelbrot BB, Wallis JR (1969) Some long-run properties of geophysical records. Wat Resour Res 5:321–340
47. Koscielny-Bunde E, Kantelhardt JW, Braun P, Bunde A, Havlin S (2006) Long-term persistence and multifractality of river runoff records: Detrended fluctuation studies. Hydrol J 322:120–137
48. Mudelsee M (2007) Long memory of rivers from spatial aggregation. Wat Resour Res 43:W01202
49. Livina VL, Ashkenazy Y, Braun P, Monetti A, Bunde A, Havlin S (2003) Nonlinear volatility of river flux fluctuations. Phys Rev E 67:042101
50. Koscielny-Bunde E, Bunde A, Havlin S, Roman HE, Goldreich Y, Schellnhuber H-J (1998) Indication of a universal persistence law governing athmospheric variability. Phys Rev Lett 81: 729–732
51. Pelletier JD, Turcotte DL (1999) Self-affine time series: Application and models. Adv Geophys 40:91
52. Talkner P, Weber RO (2000) Power spectrum and detrended fluctuation analysis: Application to daily temperatures. Phys Rev E 62:150–160
53. Eichner JF, Koscielny-Bunde E, Bunde A, Havlin S, Schellnhuber H-J (2003) Power-law persistence and trends in the atmosphere: A detailed study of long temperature records. Phys Rev E 68:046133
54. Király A, Bartos I, Jánosi IM (2006) Correlation properties of daily temperature anormalies over land. Tellus 58A(5):593–600
55. Santhanam MS, Kantz H (2005) Long-range correlations and rare events in boundary layer wind fields. Physica A 345: 713–721
56. Liu YH, Cizeau P, Meyer M, Peng C-K, Stanley HE (1997) Correlations in economic time series. Physica A 245:437; Liu YH, Gopikrishnan P, Cizeau P, Meyer M, Peng C-K, Stanley HE (1999) Statistical properties of the volatility of price fluctuations. Phys Rev E 60:1390
57. Peng C-K, Mietus J, Hausdorff JM, Havlin S, Stanley HE, Goldberger AL (1993) Long-range anticorrelations and non-gaussian behavior of the heartbeat. Phys Rev Lett 70:1343–1346
58. Bunde A, Havlin S, Kantelhardt JW, Penzel T, Peter J-H, Voigt K (2000) Correlated and uncorrelated regions in heart-rate fluctuations during sleep. Phys Rev Lett 85:3736
59. Leland WE, Taqqu MS, Willinger W, Wilson DV (1994) On the self-similar nature of Ethernet traffic. IEEE/Transactions ACM Netw 2:1–15
60. Kantelhardt JW, Koscielny-Bunde E, Rego HA, Bunde A, Havlin S (2001) Detecting long-range correlations with detrended fluctuation analysis. Physica A 295:441
61. Rybski D, Bunde A, Havlin S, Von Storch H (2006) Long-term persistence in climate and the detection problem. Geophys Res Lett 33(6):L06718
62. Rybski D, Bunde A (2008) On the detection of trends in long-term correlated records. Physica A
63. Giese E, Mossig I, Rybski D, Bunde A (2007) Long-term analysis of air temperature trends in Central Asia. Erdkunde 61(2): 186–202
64. Govindan RB, Vjushin D, Brenner S, Bunde A, Havlin S, Schellnhuber H-J (2002) Global climate models violate scaling of the observed atmospheric variability. Phys Rev Lett 89:028501
65. Vjushin D, Zhidkov I, Brenner S, Havlin S, Bunde A (2004) Volcanic forcing improves atmosphere-ocean coupled general circulation model scaling performance. Geophys Res Lett 31:L10206
66. Monetti A, Havlin S, Bunde A (2003) Long-term persistence in the sea surface temperature fluctuations. Physica A 320: 581–589
67. Kantelhardt JW, Koscielny-Bunde E, Rybski D, Braun P, Bunde A, Havlin S (2006) Long-term persistence and multifractality of precipitation and river runoff records. Geophys J Res Atmosph 111:1106
68. Bunde A, Kropp J, Schellnhuber H-J (2002) The science of disasters – climate disruptions, heart attacks, and market crashes. Springer, Berlin
69. Pfisterer C (1998) Wetternachhersage, 500 Jahre Klimavariationen und Naturkatastrophen 1496–1995. Verlag Paul Haupt, Bern
70. Glaser R (2001) Klimageschichte Mitteleuropas. Wissenschaftliche Buchgesellschaft, Darmstadt
71. Mudelsee M, Börngen M, Tetzlaff G, Grünwald U (2003) No upward trends in the occurrence of extreme floods in Central Europe. Nature 425:166
72. Bunde A, Eichner J, Havlin S, Kantelhardt JW (2003) The effect of long-term correlations on the return periods of rare events. Physica A 330:1
73. Bunde A, Eichner J, Havlin S, Kantelhardt JW (2005) Long-term memory: A natural mechanism for the clustering of extreme events and anomalous residual times in climate records. Phys Rev Lett 94:048701

74. Eichner J, Kantelhardt JW, Bunde A, Havlin S (2006) Extreme value statistics in records with long-term persistence. Phys Rev E 73:016130
75. Yamasaki K, Muchnik L, Havlin S, Bunde A, Stanley HE (2005) Scaling and memory in volatility return intervals in financial markets. PNAS 102:26 9424–9428
76. Lennartz S, Livina VN, Bunde A, Havlin S (2008) Long-term memory in earthquakes and the distribution of interoccurence times. Europ Phys Lett 81:69001
77. Corral A (2004) Long-term clustering, scaling, and universality in the temporal occurrence of earthquakes. Phys Rev Lett 92:108501
78. Stanley HE, Meakin P (1988) Multifractal phenomena in physics and chemistry. Nature 355:405
79. Ivanov PC, Goldberger AL, Havlin S, Rosenblum MG, Struzik Z, Stanley HE (1999) Multifractality in human heartbeat dynamics. Nature 399:461
80. Bogachev MI, Eichner JF, Bunde A (2007) Effect of nonlinear correlations on the statistics of return intervals in multifractal data sets. Phys Rev Lett 99:240601
81. Bogachev MI, Bunde A (2008) Memory effects in the statistics of interoccurrence times between large returns in financial records. Phys Rev E 78:036114; Bogachev MI, Bunde A (2008) Improving risk extimation in multifractal records: Applications to physiology and financing. Preprint

Fractal Growth Processes

Leonard M. Sander
Physics Department and Michigan Center for Theoretical Physics, The University of Michigan, Ann Arbor, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Fractals and Multifractals
Aggregation Models
Conformal Mapping
Harmonic Measure
Scaling Theories
Future Directions
Bibliography

Glossary

Fractal A fractal is a geometric object which is self-similar and characterized by an effective dimension which is not an integer.

Multifractal A multifractal measure is a non-negative real function) defined on a *support* (a geometric region) which has a spectrum of scaling exponents.

Diffusion-limited aggregation Diffusion-limited aggregation (DLA) is a discrete model for the irreversible growth of a cluster. The rules of the model involve a sequence of random walkers that are incorporated into a growing aggregate when they wander into contact with one of the previously aggregated walkers.

Dielectric breakdown model The dielectric breakdown model (DBM) is a generalization of DLA where the probability to grow is proportional to a power of the diffusive flux onto the aggregate. If the power is unity, the model is equivalent to DLA: in this version it is called Laplacian growth.

Harmonic measure If a geometric object is thought of as an isolated grounded conductor of fixed charge, the distribution of electric field on its surface is the harmonic measure. The harmonic measure of a DLA cluster is the distribution of growth probability on the surface.

Definition of the Subject

Fractal growth processes are a class of phenomena which produce self-similar, disordered objects in the course of development far from equilibrium. The most famous of these processes involve growth which can be modeled on the large scale by the diffusion-limited aggregation algorithm of Witten and Sander [1]. DLA is a paradigm for pattern formation modeling which has been very influential.

The algorithm describes growth limited by diffusion: many natural processes fall in this category, and the salient characteristics of clusters produced by the DLA algorithm are observed in a large number of systems such as electrodeposition clusters, viscous fingering patterns, colonies of bacteria, dielectric breakdown patterns, and patterns of blood vessels.

At the same time the DLA algorithm poses a very rich problem in mathematical physics. A full "solution" of the DLA problem, in the sense of a scaling theory that can predict the important features of computer simulations is still lacking. However, the problem shows many features that resemble thermal continuous phase transitions, and a number of partially successful approaches have been given. There are deep connections between DLA in two dimensions and theories such as Loewner evolution that use conformal maps.

Introduction

In the 1970s Benoit Mandelbrot [2] developed the idea of *fractal geometry* to unify a number of earlier studies of irregular shapes and natural processes. Mandelbrot focused on a particular set of such objects and shapes, those that are *self-similar*, i. e., where a part of the object is identical

(either in shape, or for an ensemble of shapes, in distribution) to a larger piece. He dubbed these *fractals*. He noted the surprising ubiquity of self-similar shapes in nature.

Of particular interest to Mandelbrot and his collaborators were incipient percolation clusters [3]. These are the shapes generated when, for example, a lattice is diluted by cutting bonds at random until a cluster of connected bonds just reaches across a large lattice. This model has obvious applications in physical processes such as transport in random media. The model has a non-trivial mapping to a statistical model [4] and can be treated by the methods of equilibrium statistical physics. It is likely that percolation processes account for quite a few observations of fractals in nature.

In 1981 Tom Witten and the present author observed that a completely different type of process surprisingly appears to make fractals [1,5]. These are unstable, irreversible, growth processes which we called *diffusion-limited aggregation* (DLA). DLA is a kinetic process which is not related to equilibrium statistical physics, but rather defined by growth rules. The rules idealize growth limited by diffusion: in the model there are random walking "particles" which attach irreversibly to a single cluster made up of previously aggregated particles. As we will see, quite a few natural processes can be described by DLA rules, and DLA-like clusters are reasonably common in nature. The Witten–Sander paper and subsequent developments unleashed a large amount of activity – the original work has been cited more than 2700 times as of this writing. The literature in this area up to 1998 was reviewed in a very comprehensive manner by T. Vicsek [6] and P. Meakin [7]. See also the chapter by the present author in [8]. For non-technical reviews see [9,10,11].

There are three major areas where self-similar shapes arising from non-equilibrium processes have been studied. The first is the related to the original DLA algorithm. The model may be seen as an idealization of solidification of an amorphous substance. The study of this simple-seeming model is quite unexpectedly rich, and quite difficult to treat theoretically. It will be our focus in this article. We will review the early work, but emphasize developments since [7].

Meakin [12] and Kolb, et al. [13] generalized DLA to consider cluster-cluster or colloid aggregation. In this process particles aggregate when they come into contact, but the clusters so formed are mobile, and themselves aggregate by attaching to each other. This is an idealization of colloid or coagulated aerosol formation. This model also produces fractals but this is not really a surprise: at each stage, clusters of similar size are aggregating, and the result is an approximately hierarchical object. This model is quite important in applications in colloid science. The interested reader should consult [6,14].

A third line of work arose from studies of ballistic aggregation [15] and the Eden model [16]. In the former case particles attach to an aggregate after moving in straight paths, and in the latter, particles are simply added to the surface of an aggregate at any available site, with equal probability. These models give rise to non-fractal clusters with random rough surfaces. The surfaces have scaling properties which are often characterized at the continuum level by a stochastic partial differential equation proposed by Kardar, Parisi, and Zhang [17]. For accounts of this area the reader can consult [18,19,20].

The remainder of this article is organized as follows: we first briefly review fractals and multifractals. Then we give details about DLA and related models, numerical methods, and applications of the models. The major development that has fueled a remarkable revival of interest in this area was the work of Hastings and Levitov [21] who related two-dimensional DLA to a certain class of conformal maps. Developments of this theme is the subject of the subsequent section. Then we discuss the question of the distribution of growth probabilities on the cluster surface, and finally turn to scaling theories of DLA.

Fractals and Multifractals

The kind of object that we deal with in this article is highly irregular; see Fig. 1. We think of the object as being made up of a large number, N, of units, and to be of overall linear dimension R. In growth processes the objects are formed by adding the units according to some dynamics. We will call the units "particles", and take their size to be a.

Such patterns need not be merely random, but can have well-defined scaling properties in the regime $a \ll R$. The picture is of a geometric object in two dimensions, but the concepts we introduce here are often applied to more abstract spaces. For example, a strange attractor in phase space often has the kind of scaling properties described here.

In order to characterize the geometry of such complex objects, we first cover the points in question with a set of $n(l)$ "boxes" of fixed size, l such that $a \ll l \ll R$. Clearly, for a smooth curve the product $ln(l)$ approaches a limit (the length of the curve) as $l \to 0$. This is a number of order R. For a planar region with smooth boundaries $l^2 n(l)$ approaches the area, of order R^2. The objects of ordinary geometry in d dimensions have measures given by the limit of $l^d n(l)$. For an object with many scales (a fractal), in general none of these relations hold. Rather, the product $l^D n(l)$ approaches a limit with D not necessarily an inte-

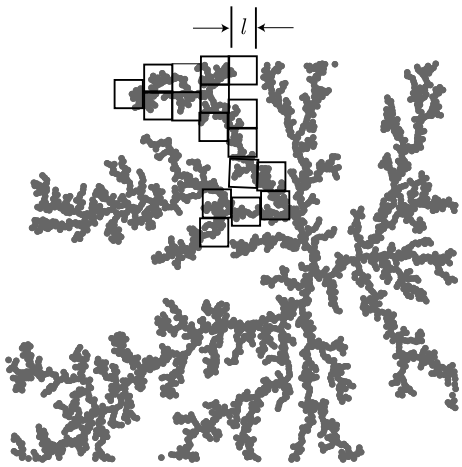

Fractal Growth Processes, Figure 1
A partial covering of a pattern with boxes. Smaller boxes reveal smaller scales of the pattern. For a pattern with many scales (like this one) there is a non-trivial scaling between the box size, *l* and the number of boxes

ger; D is called the (similarity) fractal dimension. Said another way, we define the fractal dimension by:

$$n(l) \propto (R/l)^D \,. \tag{1}$$

For many simple examples of mathematical objects with non-integer values of D see [2,6].

For an infinite fractal there are no characteristic lengths. For a finite size object there is a characteristic length, the overall scale, R. This can be taken to be any measure of the size such as the radius of gyration or the extremal radius – all such lengths must be proportional.

It is useful to generalize this definition in two ways. First we consider not only a geometric object, but also a *measure*, that is a non-negative function μ defined on the points of the object such that $\int \mathrm{d}\mu = 1$. For the geometry, we take the measure to be uniform on the points. However, for the case of growing fractals, we could also consider the growth probability at a point. As we will see, this is very non-uniform for DLA. Second, we define a sequence of generalized dimensions. If we are interested in geometry, we denote the mass of the object covered by box i by p_i. For an arbitrary measure, we define:

$$p_i = \int \mathrm{d}\mu \,, \tag{2}$$

where the integral is over the box labeled i. Then, following [22,23] we define a partition function for the p_i:

$$\chi(q) = \sum_{i=1}^{n} p_i^q \,, \tag{3}$$

where q is a real number. For an object with well-defined scaling properties we often find that χ scales with l in the following way as $l/R \to 0$:

$$\begin{aligned} \chi(q) \propto (R/l)^{-\tau(q)} &\equiv (R/l)^{-(q-1)D_q} \,; \\ \tau(q) &= (q-1)D_q \,. \end{aligned} \tag{4}$$

Objects with this property are called fractals if all the D_q are the same. Otherwise they are *multifractals*.

Some of the D_q have special significance. The similarity (or box-counting) dimension mentioned above is D_0 since in this case $\chi = n$. If we take the limit $q \to 1$ we have the information dimension of dynamical systems theory:

$$D_1 = \left.\frac{\mathrm{d}\tau}{\mathrm{d}q}\right|_{q=1} = \sum_i p_i \frac{\ln p_i}{\ln l} \,. \tag{5}$$

D_2 is called the correlation dimension since p_i^2 measures the probability that two points are close together, i. e. the number of pairs within distance l. This interpretation gives rise to a popular way to measure D_2. If we suppose that the structure is homogeneous, then the number of pairs of points can be found by focusing on any point, and drawing a d-dimensional disk of radius r around it. The number of other points in the disk will scale as r^{D_2}. For DLA, it is common to simply count the number of points within radius r of the origin, or, alternatively, the dependence of some mean radius, R on the total number of aggregated particles, N, that is $N \propto R^{D_2}$ This method is closely related to the Grassberger–Procaccia correlation integral [24].

For a simple fractal all of the D_q are the same, and we use the symbol D. If the generalized dimensions differ, then we have a multifractal. Multifractals were introduced by Mandelbrot [25] in the context of turbulence. In the context of fractal growth processes, the clusters themselves are usually simple fractals. They are the support for a multifractal measure, the growth probability.

We need another bit of formalism. It is useful to look at the fractal measure and note how the p_i scale with l/R. Suppose we assume a power-law form, $p_i \propto (l/R)^\alpha$, where there are different values of α for different parts of the measure. Also, suppose that we make a histogram of the α, and look at the parts of the support on which we have the same scaling. It is natural to adopt a form like Eq. (1) for the size of these sets, $(l/R)^{-f(\alpha)}$. (It is natural to think of $f(\alpha)$ as the fractal dimension of the set on which the measure has exponent α, but this is not quite right because f can be negative due to ensemble averaging.) Then it is not hard to show [23] that $\alpha, f(\alpha)$ are related to $q, \tau(q)$ by a Legendre transform:

$$f(\alpha) = q\frac{\mathrm{d}\tau}{\mathrm{d}q} - \tau \,; \quad \alpha = \frac{\mathrm{d}\tau}{\mathrm{d}q} \,. \tag{6}$$

Aggregation Models

In this section we review aggregation models of the DLA type, and their relationship to certain continuum processes. We look at practical methods of simulation, and exhibit a few applications to physical and biological processes.

DLA

The original DLA algorithm [1,5] was quite simple: on a lattice, declare that the point at the origin is the first member of the cluster. Then launch a random walker from a distant point and allow it to wander until it arrives at a neighboring site to the origin, and attach it to the cluster, i. e., freeze its position. Then launch another walker and let it attach to one of the two previous points, and so on. The name, diffusion-limited aggregation, refers to the fact that random walkers, i. e., diffusing particles, control the growth. DLA is a simplified view of a common physical process, growth limited by diffusion.

It became evident that for large clusters the overall shape is dependent on the lattice type [26,27], that is, DLA clusters are deformed by lattice anisotropy. This is an interesting subject [26,28,29,30,31] but most modern work is on DLA clusters without anisotropy, *off-lattice* clusters. The off-lattice algorithm is similar to the original one: instead of a random walk on a lattice the particle is considered to have radius a. For each step of the walk the particle moves its center from the current position to a random point on its perimeter. If it overlaps a particle of the current cluster, it is backed up until it just touches the cluster, and frozen at that point. Then another walker is launched. A reasonably large cluster grown in two dimensions is shown in Fig. 2, along with a smaller three-dimensional example. Most of the work on DLA has been for two dimensions, but dimensions up to 8 have been considered [32].

Patterns like the one in Fig. 2 have been analyzed for fractal properties. Careful measurements of both D_0 and D_2 (using the method, mentioned above, of fitting $n(r)$ to r^{D_2}) give the same fractal dimension, D=1.71 [32,33,34]; see Fig. 3. There is some disagreement about the next digit. There have been suggestions that DLA is a mass multifractal [35], but most authors now agree that all of the D_q are the same for the mass distribution. For three dimensions $D \approx 2.5$ [31,32], and for four dimensions $D \approx 3.4$ [32].

However, some authors [36,37] have claimed that plane DLA is not a self-similar fractal at all, and that the fractal dimension will drift towards 2 as the number of particles increases. More recent work based on conformal maps [38] has cast doubt on this. We will return to this point below.

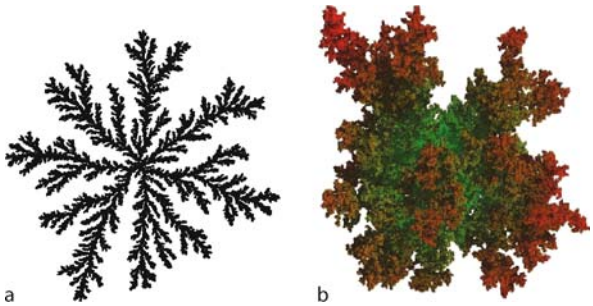

Fractal Growth Processes, Figure 2
a A DLA cluster of 50,000,000 particles produced with an off-lattice algorithm by E. Somfai. b A three-dimensional off-lattice cluster. Figure courtesy of R. Ball

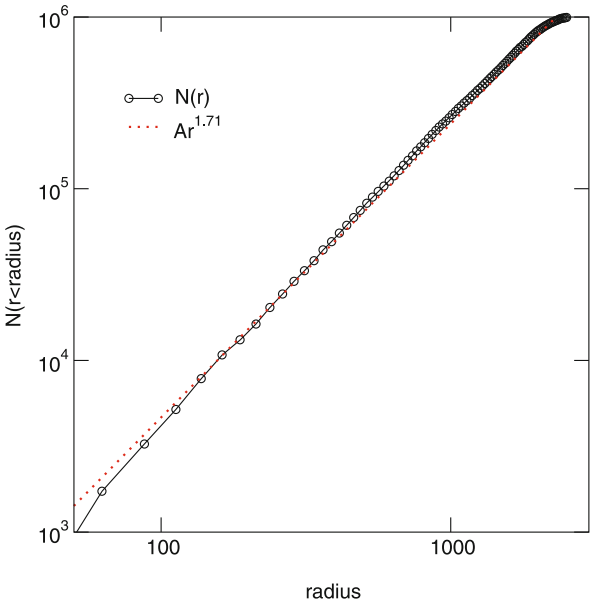

Fractal Growth Processes, Figure 3
The number of particles inside of radius *r* for a large DLA cluster. This plot gives an estimate of D_2

Plane DLA clusters can be grown in restricted geometries, see Fig. 4. The shape of such clusters is an interesting problem in pattern formation [39,40,41,42]. It was long thought that DLA grown in a channel had a different fractal dimension than in radial geometry [43,44,45]. However, more careful work has shown that the dimensions are the same in the two cases [34].

Laplacian Growth and DBM

Suppose we ask how the cluster of Fig. 2 gets to be rough. A simple answer is that if we already have a rough shape, it is quite difficult for a random walker to penetrate a nar-

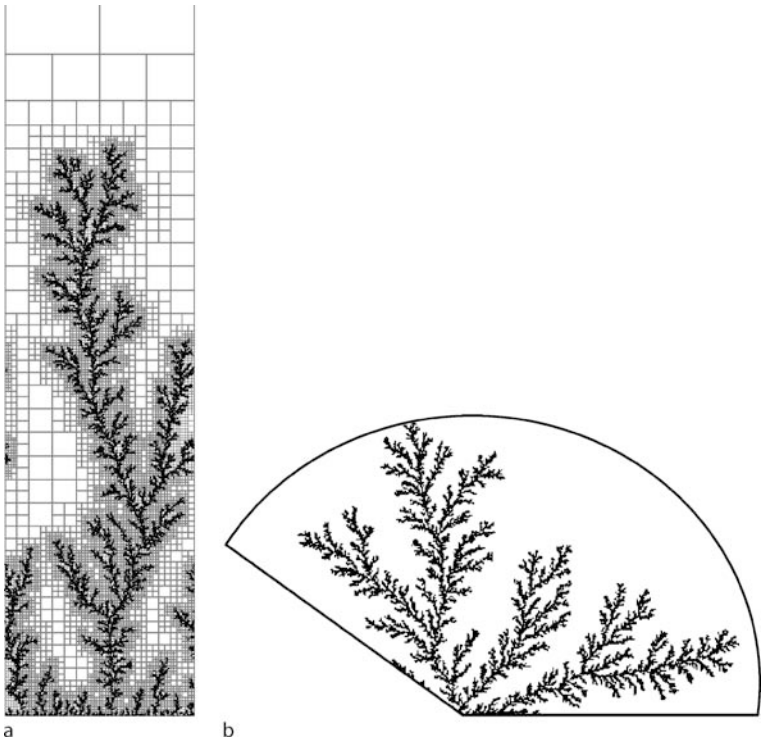

Fractal Growth Processes, Figure 4
DLA grown in a channel (a) and a wedge (b). The boxes in a are the hierarchical maps of the Sect. "Numerical Methods". Figures due to E. Somfai

row channel. (Just how difficult is treated in the Sect. "Harmonic Measure", below) The channels don't fill in, and the shape might be preserved. But we can also ask why a a smooth outline, e. g. a disk, does not continue to grow smoothly. In fact, it is easy to test that any initial condition is soon forgotten in the growth [5]. If we start with a smooth shape it roughens immediately because of a *growth instability* intrinsic to diffusion-limited growth. This instability was discovered by Mullins and Sekerka [46] who used a statement of the problem of diffusion-limited growth in continuum terms: this is known as the Stefan problem (see [47,48]), and is the standard way to idealize crystallization in the diffusion-limited case.

The Stefan problem goes as follows: suppose that we have a density $\phi(\mathbf{r}, t)$ of particles that diffuse until they reach the growing cluster where they deposit. Then we have:

$$\frac{\partial \phi}{\partial t} = \nu \nabla^2 \phi \tag{7}$$

$$\frac{\partial \phi}{\partial n} \propto v_n \tag{8}$$

That is, ϕ should obey the diffusion equation; ν is the diffusion constant. The normal growth velocity, v_n, of the interface is proportional to the flux onto the surface, $\partial\phi/\partial n$. However the term $\partial\phi/\partial t$ is of order $v\partial\phi/\partial x$, where v is a typical growth velocity. Now $|\nabla^2\phi| \approx (v/D)|\partial\phi/\partial n|$. In the DLA case we launch one particle at a time, so that the velocity goes to zero. Hence Eq. (7) reduces to the Laplace equation,

$$\nabla^2 \phi = 0 \tag{9}$$

Since the cluster absorbs the particles, we should think of it as having $\phi = 0$ on the surface. We are to solve an *electrostatics* problem: the cluster is a grounded conductor with fixed electric flux far away. We grow by an amount proportional to the electric field at each point on the surface. This is called the quasi-static or *Laplacian* growth regime for deterministic growth. A linear stability analysis of these equations gives the Mullins–Sekerka instability [46]. The qualitative reason for the instability is that near the tips of the cluster the contours of ϕ are compressed so that $\partial\phi/\partial n$, the growth rate, is large. Thus tips grow unstably. We expect DLA to have a growth instability.

However, we can turn the argument, and use these observations to give a restatement of the DLA algorithm in continuum terms: we calculate the electric field on the sur-

face of the aggregate, and interpret Eq. (8) as giving the distribution of the *growth probability*, p, at a point on the surface. We add a particle with this probability distribution, recalculate the potential using Eq. (9) and continue. This is called Laplacian growth. Simulations of Laplacian growth yield the same sort of clusters as the original discrete algorithm. (Some authors use the term Laplacian growth in a different way, to denote deterministic growth according to the Stefan model without surface tension [49].)

DLA is thus closely related to one of the classic problems of mathematical physics, dendritic crystal growth in the quasistatic regime. However, it is not quite the same for several reasons: DLA is dominated by noise, whereas the Stefan problem is deterministic. Also, the boundary conditions are different [48]: for a crystal, if we interpret u as $T - T_{\mathrm{m}}$, where T is the temperature, and T_{m} the melting temperature, we have $\phi = 0$ only on a flat surface. On a curved surface we need $\phi \propto \gamma\kappa$ where γ is the surface stiffness, and κ is the curvature. The surface tension acts as a regularization which prevents the Mullins–Sekerka instability from producing sharp cusps [50]. In DLA the regularization is provided by the finite particle size. And, of course, crystals have anisotropy in the surface tension.

There is another classic problem very similar to this, that of viscous fingering in Hele–Shaw flow [48]. This is the description of the displacement of an incompressible viscous fluid by an inviscid one: the "bubble" of inviscid fluid plays the role of the cluster, ϕ is the difference of pressures in the viscous fluid and the bubble, and the Laplace equation is the direct result of incompressibility, $\nabla \cdot \mathbf{v} = 0$ and D'Arcy's law, $\mathbf{v} = k\nabla\phi$ where k is the permeability, and $\mathbf{v}$ the fluid velocity [51].

These considerations led Niemeyer, Pietronero, and Weismann [52] to a clever generalization of Laplacian growth. They were interested in dielectric breakdown with ϕ representing a real electrostatic potential. This is known to be a threshold process so that we expect that the breakdown probability is non-linear in $\partial\phi/\partial n$. To generalize they chose:

$$p \propto \left(\frac{\partial\phi}{\partial n}\right)^{\eta} , \tag{10}$$

where η is a non-negative real number.

There are some interesting special cases for this model. For $\eta = 0$ each growth site is equally likely to be used. This is the Eden model [16]. For $\eta = 1$ we have the Laplacian growth version of DLA, and for larger η we get a higher probability to grow at the tips so that the aggregates are more spread out (as in a real dielectric breakdown pattern like atmospheric lightning). There is a remarkable fact which was suggested numerically in [53] and confirmed more recently [54,55]: for $\eta > 4$ the aggregate prefers growth at tips so much that it becomes essentially linear and non-fractal. DBM patterns for small numbers of particles are shown in Fig. 5.

A large number of variants of the basic model have been proposed such as having the random walkers perform Lévy flights, having a variable particle size, imposing a drift on the random walkers, imposing anisotropy in attachment, etc. For references on these and other variants, see [7]. There are two qualitative results from these studies that are worth mentioning here: In the presence of drift, DLA clusters cross over to (non-fractal) Eden-like clusters [56,57]. And, anisotropy deforms the shape of the clusters on the large scale, as we mentioned above.

Numerical Methods

In the foregoing we have talked about the algorithm for DLA, but we have not described what is actually done to compute the positions of 50,000,000 particles. This is a daunting computational task, and considerable ingenu-

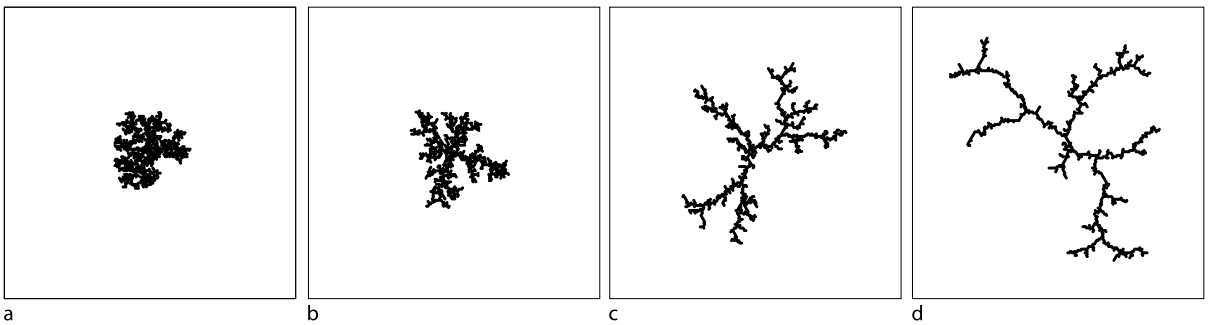

Fractal Growth Processes, Figure 5
DBM patterns for 1000 particles on a triangular lattice. a $\eta = 0.5$ b $\eta = 1$. This is a small DLA cluster. c $\eta = 2$. d $\eta = 3$

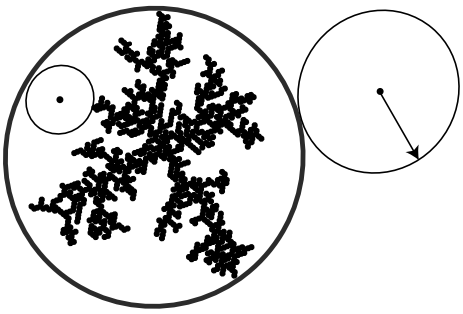

Fractal Growth Processes, Figure 6
A cluster is surrounded by a starting circle (*gray*). A random walker outside the circle can safely take steps of length of the *black circle*. If the walker is at any point, it can walk on a circle equal in size to the distance to the nearest point of the cluster. However, finding that circle appears to require a time-consuming search

ity has been devoted to making efficient algorithms. The techniques are quite interesting in their own right, and have been applied to other types of simulation.

The first observation to make is that the algorithm is an idealization of growth due to particles wandering into contact with the aggregate from far away. However, it is not necessary to start the particles far away: they arrive at the aggregate with uniform probability on a circle which just circumscribes all of the presently aggregated particles; thus we can start the particles there. As the cluster grows, the starting circle grows. This was already done in the original simulations [1,5].

However, particles can wander in and then out of the starting circle without attaching. These walkers must be followed, and could take a long time to find the cluster again. However, since there is no matter outside, it is possible to speed up the algorithm by noting that if the random walker takes *large steps* it will still have the same probability distribution, provided that it cannot encounter any matter. For example, if it walks onto the circumference of a circle that just reaches the starting circle, it will have the correct distribution. The radius of this circle is easy to find. This observation, due to P. Meakin, is the key to what follows: see Fig. 6.

We should note that the most efficient way to deal with particles that wander away is to return the particle to the starting circle in one step using the Green's function for a point charge outside an absorbing circle. A useful algorithm to do this is given in [58]. However, the idea of taking big steps is still a good one because there is a good deal of empty space *inside* the starting circle. If we could take steps in this empty space (see Fig. 6) we could again speed up the algorithm. The trick is to efficiently find the largest circle centered on the random walker that has no point of the aggregate within it.

One could imagine simply doing a spiral search from the current walker position. This technique has actually been used in a completely different setting, that of Kinetic Monte Carlo simulations of surface growth in materials science [59]. For the case of DLA an even more efficient method, the method of hierarchical maps, was devised [26]. It was extended and applied to higher dimensions and off-lattice in [32], and is now the standard method.

One version of the idea is illustrated in Fig. 4a for the case of growth in a channel. What is shown is an adaptively refined square mesh. The cluster is covered with a square – a map. The square is subdivided into four smaller squares, and each is further divided, but only if the cluster is closer to it than half of the side of the square. The subdivision continues only up to a predefined maximum depth so that the smallest maps are a few particle diameters. All particles of the cluster will be in one of the smallest maps: a list of the particles is attached to these maps.

As the cluster grows the maps are updated. Each time a particle is added to a previously empty smallest map, the neighboring maps (on all levels) are checked to see whether they satisfy the rule. If not, they are subdivided until they do. When a walker lands, we find the smallest map containing the point. If this map is not at the maximum depth, then the particle is far away from any matter, and half the side of the map is a lower estimate of the walker's distance from the cluster. If, on the other hand, the particle lands in a map of maximum depth, then it is close to the cluster. The particle lists of the map and of the neighboring smallest size maps can be checked to calculate the exact distance from the cluster. Either way, the particle is enclosed in an empty circle of known radius, and can be brought to the perimeter of the circle in one step. Note that updating the map means that there is only a search for the cluster if we are in the smallest map. Empirically, the computational time, T for an N particle cluster obeys $T \sim N^{1.1}$, and the memory is linear in N. A more recent version of the algorithm for three-dimensional growth uses a covering with balls rather than cubes [31].

For simulations of the Laplacian growth version of the situation is quite different, and simulations are much slower. The reason is that a literal interpretation of the algorithm requires that the Laplace equation be solved each time a particle is added, for example, by the relaxation or boundary-integral method. This is how corresponding simulations for viscous fingering are done [60]. For DBM clusters the growth step requires taking a power of the electric field at the surface.

It is possible to make DBM clusters using random walkers [61], but the method is rather subtle. It involves estimating the growth probability at a point by figuring out the "age" of a site, i. e., the time between attachments. This can be converted into an estimate of the growth probability. This algorithm is quite fast.

There is another class of methods involving conformal maps. These are slow methods: the practical limit of the size of clusters that can be grown this way is about 50,000. However, this method calculates the growth probability as it goes along, and thus is of great interest. Conformal mapping is discussed in the Sect. "Loewner Evolution and the Hastings–Levitov Scheme", below.

Selected Applications

Probably the most important impact of the DLA model has been the realization that diffusion-limited growth naturally gives rise to tenuous, branched objects. Of course, in the context of crystallization, this was understood in the centuries-long quest to understand the shape of snowflakes. However, applications of DLA gave rise to a unification of this with the study of many other physical applications. Broad surveys are given in [6,7,10]. Here we will concentrate on a few illustrative examples.

Rapid crystallization in a random environment is the most direct application of DLA scheme. One particularly accessible example is the growth of islands on surfaces in molecular beam epitaxy experiments. In the proper growth conditions it is easy to see patterns dominated by the Mullins–Sekerka instability. These are often referred to with the unlovely phrase "fractal-like". An example is given in Fig. 7a. There are many examples of this type, e. g. [62]. A related example is the electrodeposition of metal ions from solution. For overpotential situations DLA patterns are often observed [44,63,64,65].

There are also examples of Laplacian growth. A case of this type was discovered by Matsushita and collaborators [66], and exploited in a series of very interesting studies by the group of Ben-Jacob [67] and others. This is the growth of colonies of bacteria on hard agar plates in conditions of low nutrient supply. In this case, bacteria movement is suppressed (by the hard agar) and the limiting step in colony growth is the diffusion of nutrients to the colony. Thus we have an almost literal realization of Laplacian growth, and, indeed, colonies do look like DLA clusters in these conditions; see Fig. 7b. The detailed study of this system has led to very interesting insights into the biophysics of bacteria: these are far from our subject here, and the reader should consult [67].

We remarked above that noisy viscous fingering patterns are similar to DLA clusters, but not the same in detail: in viscous fingering the surface tension boundary condition is different from that of DLA which involves discrete particles. We should note that for viscous fingering patterns in a channel, the asymptotic state is not a disorderly cluster, but rather a single finger that fills half the channel [48] because surface tension smooths the finger. In radial growth this is not true: the Mullins–Sekerka instability gives rise to a disorderly pattern [51] which looks very much like a DLA cluster, see Fig. 8. Even for growth in a channel, if there is sufficient noise, patterns look rather

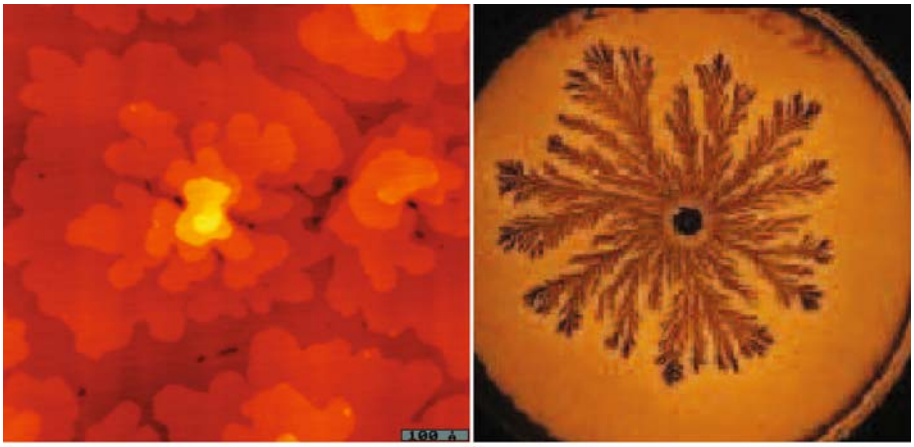

a b

Fractal Growth Processes, Figure 7

a A scanning tunneling microscope picture of Rh metal islands. The *color scale* indicates height, and the figure is about 500 Å across. Figure courtesy of R. Clarke. **b** A bacteria colony on a Petri dish. The figure is a few centimeters across. Figure courtesy of E. Ben-Jacob

Fractal Growth Processes, Figure 8
A radial viscous fingering pattern. Courtesy of M. Moore and E. Sharon

like those in Fig. 4a. In fact, Tang [68] used a version of DLA which allowed him to reduce the noise (see below) and introduce surface tension to do simulations of viscous fingering.

We should ask if this resemblance is more than a mere coincidence. It is the case that the measured fractal dimension of patterns like the one in Fig. 8 is close to 1.71, as for DLA. On this basis there are several claims in the literature that the large-scale structure of the patterns is identical [9,51,69]. Many authors have disagreed and given ingenious arguments about how to verify this. For example, some have claimed that viscous patterns become two-dimensional at large scales [70], or that viscous fingering patterns are most like DBM patterns with $\eta \approx 1.2$ [71]. Most recently a measurement of the growth probability of a large viscous fingering pattern was found to agree with that of DLA [72]. On this basis, these authors claim that DLA and viscous fingering are in the same universality class. In our opinion, this subject is still open.

Conformal Mapping

For pattern formation in two dimensions the use of analytic function theory and conformal mapping methods allows a new look at growth processes. The idea is to think of a pattern in the z plane as the image of a simple reference shape, e. g. the unit circle in the w plane, under a time-dependent analytic function, $z = F_t(w)$. More precisely, we think of the region outside of the pattern as the image of the region outside of the reference shape. By the Riemann mapping theorem the map exists and is unique if we set a boundary condition such as $F(w) \to r_0 w$ as $w \to \infty$. We will also use the inverse map, $w = G(z) = F^{-1}(z)$.

For Laplacian growth processes this idea is particularly interesting since the growth rules depend on solving the Laplace equation outside the cluster. We recall that the Laplace equation is conformally invariant; that is, it retains its form under a conformal transformation. Thus we can solve in the w plane, and transform the solution: if $\nabla^2\phi(w) = 0$, and $\phi = 0$ on the unit circle, and we take ϕ to be the real part of a complex potential Φ in the w plane, then Re $\Phi(G(z))$ solves the Laplace equation in the z plane with Re $\Phi = 0$ on the cluster boundary. Thus we can solve for the potential outside the unit circle in w-space (which is easy): $\Phi = \ln(w)$. Then if we map to z space we have the solution outside of the cluster:

$$\Phi(z) = \ln G(z) \,. \tag{11}$$

Note that the constant r_0 has the interpretation of a mean radius since $\Phi \to \ln(z/r_0)$ as $z \to \infty$. In fact, r_0 is the radius of the disk that gives the same potential as the cluster far away.

This has another consequence: the electric field (i. e. the growth probability) is uniform around the unit circle in the w plane. This means that equal intervals on the unit circle in w space map into equal regions of growth probability in the z plane. The map contains not only the shape of the cluster (the image of $|w| = 1$) but also information about the growth probability: using Eq. (11) we have:

$$|\nabla\Phi| = |G'| = \frac{1}{|F'|} \,. \tag{12}$$

The problem remains to construct the maps G or F for a given cluster. Somfai, et al. gave a direct method [38]: for a given cluster release a large number, M, of random walkers and record where they hit, say at points z_j. We know from the previous paragraph that the images of these points are spaced roughly equally on the unit circle in the w plane. That is, if we start somewhere on the cluster and number the landing positions sequentially around the cluster surface, the images in the w-plane, $w_j = r_j e^{i\theta_j}$, are given by $\theta_j = 2\pi j/M$, $r_j = 1$. Thus we have the boundary values of the map and by analytic continuation we can construct the entire map. In fact, if we represent F by a Laurent series:

$$F(w) = r_0 w + A_0 + \sum_{j=1}^{\infty} \frac{A_j}{w^j} \,, \tag{13}$$

it is easy to see that the Fourier coefficients of the function $F(\theta_j)$ are the A_j.

Unfortunately, for DLA this method only gives a map of the highly probable points on the surface of the cluster. The inner, frozen, regions are very badly sampled by random walkers, so the generated map will not represent these regions.

Loewner Evolution and the Hastings–Levitov Scheme

A more useful approach to finding F is to impose a dynamics on the map which gives rise to the growth of the cluster. This is very closely related to Loewner evolution, where a curve in the z plane is generated by a map that obeys an equation of motion:

$$\frac{\mathrm{d}G_t(z)}{\mathrm{d}t} = \frac{2}{G_t(z) - \xi(t)} \,. \tag{14}$$

The map is to the upper half plane from the upper half plane minus the set of singularities, $G = \xi$. (For an elementary discussion and references see [73].) If $\xi(t)$ is a stochastic process then many interesting statistical objects such as percolation clusters can be generated.

For DLA a similar approach was presented by Hastings and Levitov [21,30]. In this case the evolution is discrete, and is designed to represent the addition of particles to a cluster. The process is iterative: suppose we know the map for N particles, F_N. Then we want to add a "bump" corresponding to a new particle. This is accomplished by adding the bump of area λ in the w-plane on the surface of the unit circle at angle θ. There are various explicit functions that generate bumps; for the most popular example see [21]. This is a function that depends on a parameter which gives the aspect ratio of the bump.

Let us call the resulting transformation $f_{\lambda,\theta}$. If we use F_N to transform the unit circle *with a bump* we get a cluster with an extra bump in the z-plane: that is, we have added a particle to the cluster and $F_{N+1} = F_N \circ f$. The scheme is represented in Fig. 9.

There are two matters that need to be dealt with. First, we need to pick θ. Since the probability to grow is uniform on the circle in w-space, we take θ to be a random variable uniformly distributed between 0 and 2π. Also, we need the bump in z-space to have a fixed area, λ_0. That means that the bump in w-space needs to be adjusted because conformal transformations stretch lengths by $|F'|$. A first guess for the area λ is:

$$\lambda_{N+1} = \frac{\lambda_0}{|F'_N(\mathrm{e}^{i\theta_{N+1}})|^2} \,. \tag{15}$$

However, this is just a first guess. The stretching of lengths varies over the cluster, and there can be some regions where the approximation of Eq. (15) is not adequate. In this case an iterative procedure is necessary to get the area right [30,34]. The transformation itself is given by:

$$F_N = f_{\lambda_1,\theta_1} \circ f_{\lambda_2,\theta_2} \circ \cdots \circ f_{\lambda_N,\theta_N} \,. \tag{16}$$

All of the information needed to specify the mapping is contained in the list λ_j, θ_j, $1 \le j \le N$. An example of a cluster made this way is shown in Fig. 10.

If we choose θ uniformly in w-space, we have chosen points with the harmonic measure, $|F'|^{-1}$ and we make DLA clusters. To simulate DBM clusters with $\eta \neq 1$ we must choose the angles non-uniformly. Hastings [54] has shown how to do this: since a uniform distribution gives a distribution according to $|F'|^{-1}$ (see Eq. (12)) we have to pick angles with probability $|F'|^{1-\eta}$ in order to grow with probability $|\nabla\phi|^{\eta}$. This can be done with a Metropolis algorithm with $p(\theta_{N+1}) = |F'_N(\mathrm{e}^{i\theta})|^{1-\eta}$ playing the role of a Boltzmann factor.

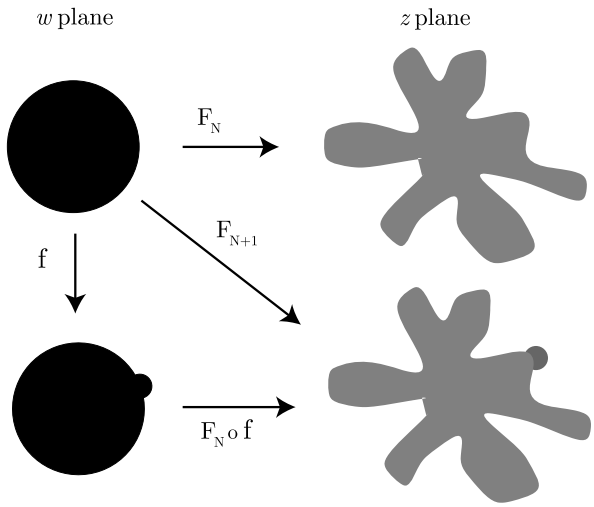

Fractal Growth Processes, Figure 9
The Hastings–Levitov scheme for fractal growth. At each stage a "bump" of the proper size is added to the unit circle. The composition of the bump map, ϕ, with the map at stage N gives the map at stage $N + 1$

Applications of the Hastings–Levitov Method: Scaling and Crossovers

The Hastings–Levitov method is a new numerical algorithm for DLA, but not a very efficient one. Constructing the composed map takes of order N steps so that the algorithm is of order N^2. One may wonder what has been gained. The essential point is that new understanding arises from considering the objects that are produced in the course of the computation. For example, Davidovitch and collaborators [74] showed that averages of λ over the

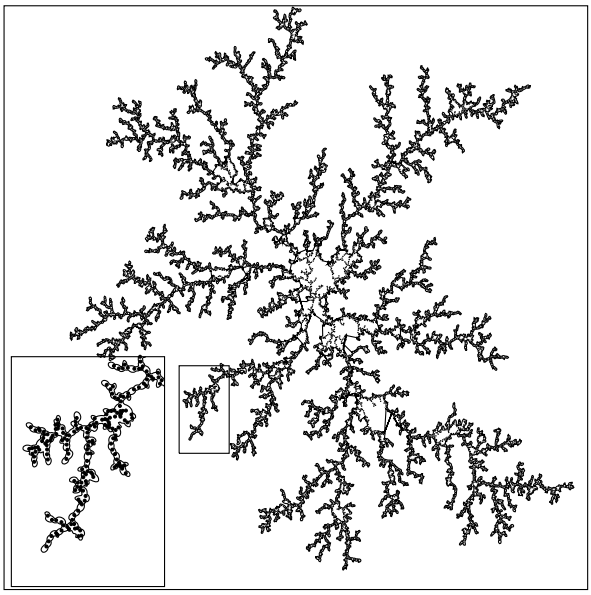

Fractal Growth Processes, Figure 10
A DLA cluster made by iterated conformal maps. Courtesy of E. Somfai

cluster boundary are related to the generalized dimensions of the growth probability measure, cf. Eq. (4). Further, the Laurent coefficients of the map, Eq. (13), are meaningful in themselves and have interesting scaling properties. We have seen above that r_0 is the radius of a disk with the same capacitance as the cluster. It scales as $N^{1/D}$, as we expect. The constant, A_0 gives the wandering of the center of the cluster from its original position. Its scaling can be worked out in terms of D and the generalized dimension, D_2.

The other coefficients, A_j, are related to the moments of the charge density on the cluster surface. We expect them to scale in the same way as r_0 for the following reason: there is an elementary theorem in complex analysis [75] for any univalent map that says, in our notation:

$$\pi r_0^2 = S_N + \pi \sum_{k=1} k|A_k|^2 \,. \tag{17}$$

Here S_N is the area of the cluster. However, this is just the area of N particles, and is linear in N. Therefore, in leading order, the sum must cancel the $N^{2/D}$ dependence of r_0^2. The simplest way this can occur is if every term in the sum goes as $N^{2/D}$. This seemed to be incorrect according to the data in [74]: they found that for for the first few k the scaling exponents of the $\langle |A_k|^2 \rangle$ were smaller than $2/D$.

However, later work [38] showed that the asymptotic scaling of *all* of the coefficients is the same. The apparent difference in the exponents is due to a slow *crossover*. This effect also seems to resolve a long-standing controversy about the asymptotic scaling behavior of DLA clusters, namely the anomalous behavior of the penetration depth of random walkers as the clusters grow.

The anomaly was pointed out numerically by Plischke and Racz [76] soon after the DLA algorithm was introduced. These authors showed numerically that the width of the region where deposition occurred, ξ, (a measure of penetration of random walkers into the cluster) seemed to grow more slowly with N than the mean radius of deposition, R_{dep}. However for a simple fractal all of the characteristic lengths must scale in the same way, $R \propto N^{1/D}$. Mandelbrot and collaborators [36,37,77] used this and other numerical evidence to suggest that DLA would not be a simple fractal for large N. However, Meakin and Sander [78] gave numerical evidence that the anomaly in the scaling of ξ is not due to a different exponent, but is a crossover.

The controversy is resolved [38] by noting that the penetration depth can be estimated from the Laurent coefficients of f:

$$\xi^2 = \frac{1}{2} \sum_{k=1} |A_k|^2 \,. \tag{18}$$

It is easy to see [38] that this version of the penetration depth can be interpreted as the rms deviation of the position of the end of a field line from those for a disk of radius r_0. Thus an anomaly in the scaling of the A_k is related to the effect discovered in [76]. Further, a slow crossover in A_k for small k is reasonable geometrically since this is a slow crossover in low moments of the charge. These low moments might be expected to have intrinsically slow dynamics.

In [38,79,80] strong numerical evidence was given for the crossover and for universal asymptotic scaling of the A_k. Indeed, many quantities with the interpretation of a length fit an expression of the form:

$$N^{1/D}\left(1 + \frac{C}{N^{\nu}}\right) \,, \tag{19}$$

where the subdominant correction to scaling is characterized by an exponent $\nu = 0.33 \pm 0.06$. This observation verifies the crossover of ξ. The asymptotic value of the penetration depth is measured to be $\xi/R_{\text{dep}} \to 0.12$. The same value is found in three dimensions [81].

In fact, the crossover probably accounts for the anomalies reported in [36,37,77]. Further, a careful analysis of numerical data shows that the reported *multiscaling* [82] (an apparent dependence of D on the distance from the center of the cluster) is also a crossover effect. These in-

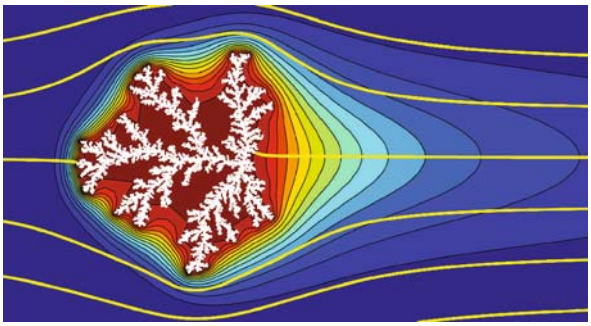

Fractal Growth Processes, Figure 11
An aggregate grown by the advection-diffusion mechanism, Eq. (20) for Pe=10. Courtesy of M. Bazant

sights lead to the view that DLA clusters are simple fractals up to a slow crossover. The origin of the crossover exponent, $\nu \approx 1/3$, is not understood. In three dimensions a similar crossover exponent was found by different techniques [81] with a value of 0.22 ± 0.03.

Conformal Mapping for Other Discrete Growth Processes

The Hastings–Levitov scheme depends on the conformal invariance of the Laplace equation. Bazant and collaborators [83,84] pointed out that the same techniques can be used for more general growth processes. For example, consider growth of an aggregate in a flowing fluid. Now there are two relevant fields, the particle concentration, c, and the velocity of the fluid, $\mathbf{v} = \nabla\psi$ (for potential flow). The current associated with c can be written $\mathbf{j} = \mu c\mathbf{v} - \nu\nabla c$, where μ is a mobility and ν a diffusion coefficient. For steady incompressible flow we have, after rescaling:

$$\mathrm{Pe}\nabla\psi \cdot \nabla c = \nabla^2 c\,; \quad \nabla^2\psi = 0\,. \tag{20}$$

Here Pe is the Péclet number, UL/ν, where U is the value of the velocity far from the aggregate and L its initial size. These equations are conformally invariant, and the problem of flow past a disk in the w plane is easily solved. In [83] growth was accomplished by choosing bumps with probability distribution $\partial c/\partial n$ using the method described above. This scheme solves the flow equation past the complex growing aggregate and adds particles according to their flux at the surface. An example of a pattern of this type is given in Fig. 11.

It has long been known that it is possible to treat quasi-static fracture as a growth process similar to DLA [85,86,87]. This is due to the fact that the Lamé equation of elasticity is similar in form to the Laplace equation and the condition for breaking a region of the order of a process zone is related to the stress at the current crack, i. e., boundary values of derivatives of the displacement field. Recent work has exploited this similarity to give yet another application of conformal mapping. For example, for Mode III fracture the quasi-static elastic equation reduces to the Laplace equation, and it is only necessary to replace the growth probability and the boundary conditions in order to use the method of iterated conformal maps [88]. For Mode I and Mode II fracture it is necessary to solve for two analytic functions, but this can also be done [89].

Harmonic Measure

The distribution of the boundary values of the normal derivatives of a potential on a electrode of complex shape is called the problem of the harmonic measure. For DLA it is equivalent to the distribution of growth probabilities on the surface of the cluster, or, in other terms, the penetration of random walkers into the cluster. For other complex shapes the problem is still interesting. Its practical significance is that of the study of electrodes [90] or catalytic surfaces. In some cases the harmonic measure has deep relationships to conformal field theory.

For the case of the harmonic measure we can interpret the variable α of Eq. (6) as the singularity strength of the electric field near a sharp tip. This is seen as follows: it is well known [91] that near the apex of a wedge-shaped conductor the electric field diverges as $r^{\pi/\beta-1}$ where r is the distance from the tip and the exterior angle of the wedge is β. For example, near a square corner with $\beta = 3\pi/2$ there is a $r^{-1/3}$ divergence. Now the quantity p_i is the integral over a box of size l. Thus a sequence of boxes centered on the tip will give a power-law $l^{\pi/\beta} = l^{\alpha}$. Smaller α means stronger divergence.

For fractals that can be treated with the methods of conformal field theory, a good deal is known about the harmonic measure. For example, for a percolation cluster at p_c the harmonic measure is completely understood [92]. The D_q are given by a formula:

$$D_q = \frac{1}{2} + \frac{5}{\sqrt{24q+1}+5}\,, \quad q \geq -\frac{1}{24}\,. \tag{21}$$

The $f(\alpha)$ spectrum is easy to compute from this. This formula in good accord with the numerical results of Meakin and collaborators [93] who sampled the measure by firing many random walkers at a cluster. There is an interesting feature of this formula: $D_0 = 4/3$ is the dimension of the support of the measure. This is *less* than the dimension of a percolation hull, 7/4. There is a large part of the surface of a percolation cluster which is inaccessible to random walk-

ers – the interior surfaces are cut off from the exterior by narrow necks whose widths vanish in the scaling regime.

For DLA the harmonic measure is much less well understood. Some results are known. For example, Makarov [94] proved that $D_1 = 1$ under very general circumstances for a two-dimensional harmonic measure. Halsey [95] used Green's functions for electrostatics to prove the following for DBM models with parameter η:

$$\tau(\eta + 2) - \tau(\eta) = D \,, \tag{22}$$

Here D is the fractal dimension of the cluster. For DLA $\eta = 1$, and $\tau(1) = 0$. Thus $\tau(3) = D = 1.71$.

We introduced the function $f(\alpha)$ as the Legendre transform of $D(q)$. There is an argument due to Turkevich and Sher [96] which allows us to see a feature of $f(\alpha)$ directly by giving an estimate of the singularity associated with the most active tip of the growing cluster. Note that the growth rate of the extremal radius of the cluster is related to the fractal dimension because $R_{\text{ext}} \propto N^{1/D}$. Suppose we imagine adding one particle per unit time and recall that $p_{\text{tip}} \propto (l/R_{\text{ext}})^{\alpha_{\text{tip}}}$. Then:

$$\frac{\mathrm{d}R_{\text{ext}}}{\mathrm{d}t} = \frac{\mathrm{d}R_{\text{ext}}}{\mathrm{d}N}\,\frac{\mathrm{d}N}{\mathrm{d}t} \propto R_{\text{ext}}^{-\alpha_{\text{tip}}}$$

$$D = 1 + \alpha_{\text{tip}} \approx 1 + \alpha_{\min} \,. \tag{23}$$

Since the singularity at the tip is close to being the most active one, we have an estimate of the minimum value of α.

There have been a very large number of numerical investigations of $D(q)$, $f(\alpha)$ for two dimensional DLA; see [7] for a comprehensive list. They proceed either by launching many random walkers, e. g. [93] or by solving the Laplace equation [97], or, most recently, by using the Hastings–Levitov method [98]. The general features of the results are shown in Fig. 12. There is fairly good agreement about the left hand side of the curve which corresponds to large probabilities. The intercept for small α is close to the Turkevich–Sher relation above. The maximum of the curve corresponds to $\mathrm{d}f/\mathrm{d}\alpha = q = 0$; thus it is the dimension of the support of the measure. This seems to be quite close to D so that the whole surface is accessible to random walkers.

There is very little agreement about the right hand side of the curve. It arises from regions with small probabilities which are very hard to estimate. The most reliable current method is that of [98] where conformal maps are manipulated to get at probabilities as small as 10^{-70}, quite beyond the reach of other techniques. Unfortunately, these computations are for rather small clusters ($N \approx 50,000$) which are the largest ones that can be made by the Hastings–Levitov method. A particularly active controversy relates to the value of $\alpha_{\max}$, if it exists at all; that is, the question is whether there is a maximum value of the slope $\mathrm{d}\tau/\mathrm{d}q$. The authors of [98] find $\alpha_{\max} \approx 20$.

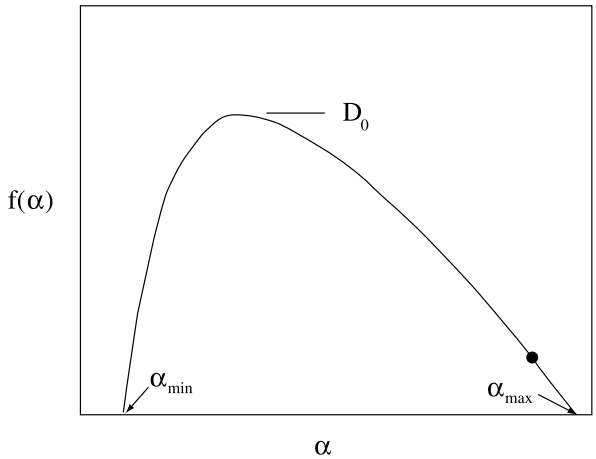

Fractal Growth Processes, Figure 12
A sketch of the $f(\alpha)$ curve. Some authors find a maximum slope at a position like that marked by *the dot* so that the curve ends there. The curve is extended to the real axis with a *straight line*. This is referred to as a phase transition

Scaling Theories

Our understanding of non-equilibrium fractal growth processes is not very satisfactory compared to that of equilibrium processes. A long-term goal of many groups has been to find a "theory of DLA" which has the nice features of the renormalization theory of critical phenomena. The result of a good deal of work is that we have theories that give the general features of DLA, but they do not explain things in satisfactory detail.

We should note that a mere estimate of the fractal dimension, D, is not what we have in mind. There are several ad hoc estimates in the literature that give reasonable values of D [41,99]. We seek, rather, a theory that allows a good understanding of the fixed point of fractal growth with a description of relevant and irrelevant operators, crossovers, etc. There have been an number of such attempts. In the first section below we describe a semi-numerical scheme which sheds light on the fixed point. Then we look at several attempts at ab initio theory.

Scaling of Noise

The first question one might ask about DLA growth is the role of noise. It might seem that for a very large cluster the noise of individual particle arrivals should average

out. However, this is not the case. Clusters fluctuate on all scales for large N.

Further, the noise-free case seems to be unstable. We can see this by asking about the related question of the growth of viscous fingers in two dimensions. As was remarked long ago [51] this sort of growth is always unstable. Numerical simulations of viscous fingering [69] show that any asymmetric initial condition develops into a pattern with a fractal dimension close to that of DLA. DLA organizes its noise to a fixed point exactly as turbulence does.

Several authors [79,81,100] have looked at the idea that the noise (measured in some way) flows to a fixed value as clusters grow large. For example, we could characterize the shape fluctuations by measuring the scaled variance of the number of particles necessary to grow to a fixed extremal radius:

$$\left.\frac{\delta N}{N}\right|_{R_{\text{ext}}} = \sqrt{A^*}\,. \tag{24}$$

This is easy to measure for large clusters. In two dimensions $A^* = .0036$ [79]. In [100] it was argued that DLA will be at its fixed point if if one unit of growth acts as a coarse graining of DLA on finer length scales. This amounts to a kind of real-space renormalization.

For the original DLA model the scaled noise to grow one unit of length is of order unity. We can reduce it to the fixed point value by *noise reduction*. The slow approach to the fixed point governed by the exponent ν can be interpreted as the drift of the noise to its proper value as N grows.

Noise reduction was introduced into lattice models by Tang [68]. He kept a counter on each site and only grew there if a certain number of random walkers, m, hit that point. For off-lattice clusters a similar reduction is obtained if shallow bumps of height A are added to the cluster by letting the particles overlap as they stick. We must add $m = 1/A$ particles to advance the growth by one particle diameter [79]. For the Hastings–Levitov method it is equivalent to use a bump map with a small aspect ratio [30].

In either case if we examine a number of sites whose mean advance is one unit, we will find $\delta N/N = \sqrt{A}$. We should expect that if we tune the input value of A to A^* we will get to asymptotic behavior quickly. This is what is observed in two dimensions [79] and in three dimensions [81]. The amplitude of the crossover, the parameter C in Eq. (19), is smallest for A near the fixed point. In Fig. 13 we show two clusters, a small one grown with noise reduction, and a much larger one with $A = 1$. They both have the same value of ξ/R_{dep}, near the asymptotic value of 0.12.

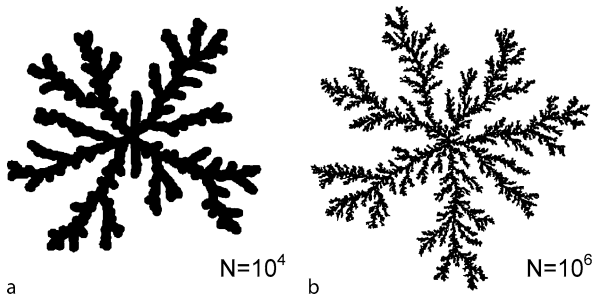

Fractal Growth Processes, Figure 13
A small noise-reduced cluster and a larger one with no noise reduction. Noise reduction of the proper size accelerates the approach to the fixed point. From [79]

Attempts at Theoretical Description

The last section described a semi-empirical approach to DLA. The reader may wonder why the techniques of phase transition theory are not simply applicable here. One problem is that one of the favorite methods in equilibrium theory, the ε-expansion, cannot be used because DLA has no upper critical dimension [5,101].

To be more precise, in other cluster theories such as percolation there is a dimension, d_c such that if $d > d_c$ the fractal dimension of the cluster does not change. For example, percolation clusters are 4 dimensional for all dimensions above 6. For DLA this is not true because if d is much bigger than D then random walkers will *penetrate and fill up the cluster* so that its dimension would increase. This results [5] from the fact that for a random walker the number of sites visited in radius R is $N_w \propto R^2$. In the same region there are R^D sites of the cluster, so the density of cluster sites goes as R^{D-d}. The mean number of intersections is $R^{D-d}R^2$. Therefore if $D + 2 < d$ there are a vanishing number of intersections, and the cluster will fill up. Thus we must have $D \geq d - 2$. A related argument [101] sharpens the bound to $D \geq d - 1$: the fractal dimension increases without limit as d increases.

Halsey and collaborators have given a theoretical description based on branch competition [102,103]. This method has been used to give an estimate of D_q for positive q [104]. The idea is to think of two branches that are born from a single site. Then the probability to stick to the first or second is called p_1, p_2 where $p_1 + p_2 = p_b$ is the total probability to stick to that branch. Similarly there are numbers of particles, $n_1 + n_2 = n_b$. Define $x = p_1/p_b$, $y = n_1/n_b$. The two variables x, y regulate the

branch competition. On the average $p_1 = \mathrm{d}n_1/\mathrm{d}n_b$ so we have:

$$n_b \frac{\mathrm{d}y}{\mathrm{d}n_b} = x - y \,. \tag{25}$$

The equation of motion for x is assumed to be of similar form:

$$n_b \frac{\mathrm{d}x}{\mathrm{d}n_b} = g(x, y) \,. \tag{26}$$

Thus the competition is reduced to a two-dimensional dynamical system with an unstable fixed point at $x = y = 1/2$ corresponding to the point when the two branches start with equal numbers. The fixed point is unstable because one branch will eventually screen the other. If g is known, then the unstable manifold of this fixed point describes the growth of the dominant branch, and it turns out, by counting the number of particles that attach to the dominant branch, that the eigenvalue of the unstable manifold is $1/D$. The starting conditions for the growth are taken to result from microscopic processes that distribute points randomly near the fixed point.

The problem remains to find g. This was done several ways: numerically, by doing a large number of simulations of branches that start equally, or in terms of a complicated self-consistent equation [103]. The result is a fractal dimension of 1.66, and a multifractal spectrum that agrees pretty well with direct simulations [104].

Another approach is due to Pietronero and collaborators [105]. It is called the method of fixed scale transformations. It is a real-space method where a small system at one scale is solved essentially exactly, and the behavior at the next coarse-grained scale estimated by assuming that there is a scale-invariant dynamics and estimating the parameters from the fixed-scale solution. The method is much more general than a theory of DLA: in [105] it is applied to directed percolation, the Eden model, sandpile models, and DBM. For DLA the fractal dimension calculated is about 1.6. The rescaled noise (cf. the previous section) comes out to be of order unity rather than the small value, 0.0036, quoted above [106] .

The most recent attempt at a fundamental theory is due to Ball and Somfai [71,107]. The idea depends on a mapping from DLA to an instance of the DBM which has different boundary conditions on the growing tip. The scaling of the noise and the multifractal spectrum (for small α) are successfully predicted.

Future Directions

The DLA model is 27 years old as of this writing. Every year (including last year) there have been about 100 references to the paper. Needless to say, this author has only read a small fraction of them. Space and time prevented presenting here even the interesting ones that I am familiar with.

For example, there is a remarkable literature associated with the viscous-fingering problem without surface tension which seems, on the one hand, to describe some facets of experiments [108] and on the other to have deep relationships with the theory of 2d quantum gravity [109,110]. Where this line of work will lead is a fascinating question. There are other examples: I hope that those whose work I have not covered will not feel slighted. There is simply too much going on.

A direction that should be pursued is to use the ingenious techniques that have been developed for the DLA problem for problems in different areas; [59,83] are examples of this.

It is clear that this field is as lively as ever after 27 years, and will certainly hold more surprises.

Bibliography

Primary Literature

1. Witten TA, Sander LM (1981) Diffusion-limited aggregation, a kinetic critical phenomenon. Phys Rev Lett 47:1400
2. Mandelbrot BB (1982) The Fractal Geometry of Nature. Freeman, San Francisco
3. Stauffer D, Aharony A (1994) Introduction to percolation theory. Taylor & Francis, London
4. Fortuin CM, Kasteleyn PW (1972) On the random-cluster model: I introduction and relation to other models. Physica 57(4):536–564
5. Witten TA, Sander LM (1983) Diffusion-limited aggregation. Phys Rev B 27:5686
6. Vicsek T (1992) Fractal Growth Phenomena, 2nd edn. World Scientific, Singapore
7. Meakin P (1998) Fractals, scaling, and growth far from equilibrium. Cambridge University Press, Cambridge
8. Godreche G (1991) Solids far from equilibrium. Cambridge, Cambridge, New York
9. Sander LM (1986) Fractal growth-processes. Nature 322(6082):789–793
10. Sander LM (2000) Diffusion limited aggregation, a kinetic critical phenomenon? Contemporary Physics 41:203–218
11. Halsey TC (2000) Diffusion-limited aggregation: A model for pattern formation. Physics Today 53(4):36–41
12. Meakin P (1983) Formation of fractal clusters and networks by irreversible diffusion-limited aggregation. Phys Rev Lett 51(13):1119–1122
13. Kolb M, Botet R, Jullien R (1983) Scaling of kinetically growing clusters. Phys Rev Lett 51(13):1123–1126
14. Meakin P (1988) Fractal aggregates. Adv Colloid Interface Sci 28(4):249–331
15. Meakin P, Ramanlal P, Sander LM, Ball RC (1986) Ballistic deposition on surfaces. Phys Rev A 34(6):5091–5103

16. Eden M (1961) A two-dimensional growth model. In: Neyman J (ed) Proceedings of the 4th Berkeley symposium on mathematics, statistics, and probability. University of California Press, Berkeley
17. Kardar M, Parisi G, Zhang Y (1986) Dynamic scaling of growing interfaces. Phys Rev Lett 56:889
18. Barabasi A, Stanley HE (1995) Fractal Concepts in Surface Growth. Cambridge, Cambridge, New York
19. Family F, Vicsek T (1992) The Dynamics of Growing Interfaces. World Scientific, Singapore
20. Halpin-Healy T, Zhang Y-C (1995) Kinetic roughening phenomena, stochastic growth, directed polymers and all that. Aspects of multidisciplinary statistical mechanics. Phys Rep 254(4–6):215–414
21. Hastings MB, Levitov LS (1998) Laplacian growth as one-dimensional turbulence. Physica D 116:244–252
22. Hentschel HGE, Procaccia I (1983) The infinite number of generalized dimensions of fractals and strange attractors. Physica D 8(3):435–444
23. Halsey TC, Jensen MH, Kadanoff LP, Procaccia I, Shraiman BI (1986) Fractal measures and their singularities – the characterization of strange sets. Phys Rev A 33(2):1141–1151
24. Grassberger P, Procaccia I (1983) Measuring the strangeness of strange attractors. Physica D 9:189–208
25. Mandelbrot BB (1974) Intermittant turbulence in self-similar cascades; divergence of high moments and dimension of the carrier. Fluid J Mech 62:331–358
26. Ball RC, Brady RM (1985) Large-scale lattice effect in diffusion-limited aggregation. Phys J A-Math Gen 18(13):L809
27. Ball RC, Brady RM, Rossi G, Thompson BR (1985) Anisotropy and cluster growth by diffusion-limited aggregation. Phys Rev Lett 55(13):1406– 1409
28. Meakin P, Ball RC, Ramanlal P, Sander LM (1987) Structure of large two-dimensional square-lattice diffusion-limited aggregates – approach to asymptotic-behavior. Phys Rev A 35(12):5233–5239
29. Eckmann JP, Meakin P, Procaccia I, Zeitak R (1990) Asymptotic shape of diffusion-limited aggregates with anisotropy. Phys Rev Lett 65(1):52–55
30. Stepanov MG, Levitov LS (2001) Laplacian growth with separately controlled noise and anisotropy. Phys Rev E 63:061102
31. Goold NR, Somfai E, Ball RC (2005) Anisotropic diffusion limited aggregation in three dimensions: Universality and nonuniversality. Phys Rev E 72(3):031403
32. Tolman S, Meakin P (1989) Off-lattice and hypercubic-lattice models for diffusion-limited aggregation in dimensionalities 2–8. Phys Rev A 40:428–437
33. Ossadnik P (1991) Multiscaling analysis of large-scale off-lattice DLA. Physica A 176:454–462
34. Somfai E, Ball RC, DeVita JP, Sander LM (2003) Diffusion-limited aggregation in channel geometry. Phys Rev E 68: 020401(R)
35. Vicsek T, Family F, Meakin P (1990) Multifractal geometry of diffusion-limited aggregates. Europhys Lett 12(3): 217–222
36. Mandelbrot BB, Kaufman H, Vespignani A, Yekutieli I, Lam CH (1995) Deviations from self-similarity in plane DLA and the infinite drift scenario. Europhys Lett 29(8):599–604
37. Mandelbrot BB, Vespignani A, Kaufman H (1995) Crosscut analysis of large radial DLA – departures from self-similarity and lacunarity effects. Europhys Lett 32(3):199–204
38. Somfai E, Sander LM, Ball RC (1999) Scaling and crossovers in diffusion limited aggregation. Phys Rev Lett 83:5523–5526
39. Arneodo A, Elezgaray J, Tabard M, Tallet F (1996) Statistical analysis of off-lattice diffusion-limited aggregates in channel and sector geometries. Phys Rev E 53(6):6200–6223(B)
40. Tu YH, Levine H (1995) Mean-field theory of the morphology transition in stochastic diffusion-limited growth. Phys Rev E 52(5):5134–5141
41. Kessler DA, Olami Z, Oz J, Procaccia I, Somfai E, Sander LM (1998) Diffusion-limited aggregation and viscous fingering in a wedge: Evidence for a critical angle. Phys Rev E 57(6): 6913–6916
42. Sander LM, Somfai E (2005) Random walks, diffusion limited aggregation in a wedge, and average conformal maps. Chaos 15:026109
43. Meakin P, Family F (1986) Diverging length scales in diffusion-limited aggregation. Phys Rev A 34(3):2558–2560
44. Argoul F, Arneodo A, Grasseau G, Swinney HL (1988) Self-similarity of diffusion-limited aggregates and electrodeposition clusters. Phys Rev Lett 61(22):2558–2561
45. Kol B, Aharony A (2001) Diffusion-limited aggregation as markovian process: Site-sticking conditions. Phys Rev E 63(4): 046117
46. Mullins WW, Sekerka RF (1963) Morphological stability of a particle growing by diffusion or heat flow. J Appl Phys 34:323
47. Langer JS (1980) Instabilities and pattern formation in crystal growth. Rev Mod Phys 52:1
48. Pelcé P (2004) New visions on form and growth: fingered growth, dendrites, and flames. In: Thèorie des formes de croissance. Oxford University Press, Oxford
49. Mineev-Weinstein MB, Dawson SP (1994) Class of nonsingular exact-solutions for laplacian pattern-formation. Phys Rev E 50(1):R24–R27
50. Shraiman B, Bensimon D (1984) Singularities in nonlocal interface dynamics. Phys Rev A 30:2840–2842
51. Paterson L (1984) Diffusion-limited aggregation and 2-fluid displacements in porous-media. Phys Rev Lett 52(18): 1621–1624
52. Niemeyer L, Pietronero L, Wiesmann HJ (1984) Fractal dimension of dielectric breakdown. Phys Rev Lett 52:1033–1036
53. Sanchez A, Guinea F, Sander LM, Hakim V, Louis E (1993) Growth and forms of Laplacian aggregates. Phys Rev E 48:1296–1304
54. Hastings MB (2001) Fractal to nonfractal phase transition in the dielectric breakdown model. Phys Rev Lett 87:175502
55. Hastings MB (2001) Growth exponents with 3.99 walkers. Phys Rev E 64:046104
56. Meakin P (1983) Effects of particle drift on diffusion-limited aggregation. Phys Rev B 28(9):5221–5224
57. Nauenberg M, Richter R, Sander LM (1983) Crossover in diffusion-limited aggregation. Phys Rev B 28(3):1649–1651
58. Sander E, Sander LM, Ziff R (1994) Fractals and fractal correlations. Comput Phys 8:420
59. DeVita JP, Sander LM, Smereka P (2005) Multiscale kinetic monte carlo algorithm for simulating epitaxial growth. Phys Rev B 72(20):205421
60. Hou TY, Lowengrub JS, Shelley MJ (1994) Removing the stiffness from interfacial flow with surface-tension. J Comput Phys 114(2):312–338
61. Somfai E, Goold NR, Ball RC, DeVita JP, Sander LM (2004)

Growth by random walker sampling and scaling of the dielectric breakdown model. Phys Rev E 70:051403
62. Radnoczi G, Vicsek T, Sander LM, Grier D (1987) Growth of fractal crystals in amorphous $GeSe_2$ films. Phys Rev A 35(9):4012–4015
63. Brady RM, Ball RC (1984) Fractal growth of copper electrodeposits. Nature 309(5965):225–229
64. Grier D, Ben-Jacob E, Clarke R, Sander LM (1986) Morphology and microstructure in electrochemical deposition of zinc. Phys Rev Lett 56(12):1264–1267
65. Sawada Y, Dougherty A, Gollub JP (1986) Dendritic and fractal patterns in electrolytic metal deposits. Phys Rev Lett 56(12):1260–1263
66. Fujikawa H, Matsushita M (1989) Fractal growth of bacillus-subtilis on agar plates. Phys J Soc Jpn 58(11):3875–3878
67. Ben-Jacob E, Cohen I, Levine H (2000) Cooperative self-organization of microorganisms. Adv Phys 49(4):395–554
68. Tang C (1985) Diffusion-limited aggregation and the Saffman–Taylor problem. Phys Rev A 31(3):1977–1979
69. Sander LM, Ramanlal P, Ben-Jacob E (1985) Diffusion-limited aggregation as a deterministic growth process. Phys Rev A 32:3160–3163
70. Barra F, Davidovitch B, Levermann A, Procaccia I (2001) Laplacian growth and diffusion limited aggregation: Different universality classes. Phys Rev Lett 87:134501
71. Ball RC, Somfai E (2002) Theory of diffusion controlled growth. Phys Rev Lett 89:135503
72. Mathiesen J, Procaccia I, Swinney HL, Thrasher M (2006) The universality class of diffusion-limited aggregation and viscous-limited aggregation. Europhys Lett 76(2):257–263
73. Gruzberg IA, Kadanoff LP (2004) The Loewner equation: Maps and shapes. J Stat Phys 114(5–6):1183–1198
74. Davidovitch B, Hentschel HGE, Olami Z, Procaccia I, Sander LM, Somfai E (1999) Diffusion limited aggregation and iterated conformal maps. Phys Rev E 59:1368–1378
75. Duren PL (1983) Univalent Functions. Springer, New York
76. Plischke M, Rácz Z (1984) Active zone of growing clusters: Diffusion-limited aggregation and the Eden model. Phys Rev Lett 53:415–418
77. Mandelbrot BB, Kol B, Aharony A (2002) Angular gaps in radial diffusion-limited aggregation: Two fractal dimensions and nontransient deviations from linear self-similarity. Phys Rev Lett 88:055501
78. Meakin P, Sander LM (1985) Comment on "Active zone of growing clusters: Diffusion-limited aggregation and the Eden model". Phys Rev Lett 54:2053–2053
79. Ball RC, Bowler NE, Sander LM, Somfai E (2002) Off-lattice noise reduction and the ultimate scaling of diffusion-limited aggregation in two dimensions. Phys Rev E 66:026109
80. Somfai E, Ball RC, Bowler NE, Sander LM (2003) Correction to scaling analysis of diffusion-limited aggregation. Physica A 325(1–2):19–25
81. Bowler NE, Ball RC (2005) Off-lattice noise reduced diffusion-limited aggregation in three dimensions. Phys Rev E 71(1):011403
82. Amitrano C, Coniglio A, Meakin P, Zannetti A (1991) Multiscaling in diffusion-limited aggregation. Phys Rev B 44:4974–4977
83. Bazant MZ, Choi J, Davidovitch B (2003) Dynamics of conformal maps for a class of non-laplacian growth phenomena. Phys Rev Lett 91(4):045503
84. Bazant MZ (2004) Conformal mapping of some non-harmonic functions in transport theory. Proceedings of the Royal Society of London Series A – Mathematical Physical and Engineering Sciences 460(2045):1433–1452
85. Louis E, Guinea F (1987) The fractal nature of fracture. Europhys Lett 3(8):871–877
86. Pla O, Guinea F, Louis E, Li G, Sander LM, Yan H, Meakin P (1990) Crossover between different growth regimes in crack formation. Phys Rev A 42(6):3670–3673
87. Yan H, Li G, Sander LM (1989) Fracture growth in 2d elastic networks with Born model. Europhys Lett 10(1):7–13
88. Barra F, Hentschel HGE, Levermann A, Procaccia I (2002) Quasistatic fractures in brittle media and iterated conformal maps. Phys Rev E 65(4)
89. Barra F, Levermann A, Procaccia I (2002) Quasistatic brittle fracture in inhomogeneous media and iterated conformal maps: Modes I, II, and III. Phys Rev E 66(6):066122
90. Halsey TC, Leibig M (1992) The double-layer impedance at a rough-surface – theoretical results. Ann Phys 219(1):109–147
91. Jackson JD (1999) Classical electrodynamics, 3rd edn. Wiley, New York
92. Duplantier B (1999) Harmonic measure exponents for two-dimensional percolation. Phys Rev Lett 82(20):3940–3943
93. Meakin P, Coniglio A, Stanley HE, Witten TA (1986) Scaling properties for the surfaces of fractal and nonfractal objects – an infinite hierarchy of critical exponents. Phys Rev A 34(4):3325–3340
94. Makarov NG (1985) On the distortion of boundary sets under conformal-mappings. P Lond Math Soc 51:369–384
95. Halsey TC (1987) Some consequences of an equation of motion for diffusive growth. Phys Rev Lett 59:2067–2070
96. Turkevich LA, Scher H (1985) Occupancy-probability scaling in diffusion-limited aggregation. Phys Rev Lett 55(9):1026–1029
97. Ball RC, Spivack OR (1990) The interpretation and measurement of the f(alpha) spectrum of a multifractal measure. J Phys A 23:5295–5307
98. Jensen MH, Levermann A, Mathiesen J, Procaccia I (2002) Multifractal structure of the harmonic measure of diffusion-limited aggregates. Phys Rev E 65:046109
99. Ball RC (1986) Diffusion limited aggregation and its response to anisotropy. Physica A: Stat Theor Phys 140(1–2):62–69
100. Barker PW, Ball RC (1990) Real-space renormalization of diffusion-limited aggregation. Phys Rev A 42(10):6289–6292
101. Ball RC, Witten TA (1984) Causality bound on the density of aggregates. Phys Rev A 29(5):2966–2967
102. Halsey TC, Leibig M (1992) Theory of branched growth. Phys Rev A 46:7793–7809
103. Halsey TC (1994) Diffusion-limited aggregation as branched growth. Phys Rev Lett 72(8):1228–1231
104. Halsey TC, Duplantier B, Honda K (1997) Multifractal dimensions and their fluctuations in diffusion-limited aggregation. Phys Rev Lett 78(9):1719–1722
105. Erzan A, Pietronero L, Vespignani A (1995) The fixed-scale transformation approach to fractal growth. Rev Mod Phys 67(3):545–604
106. Cafiero R, Pietronero L, Vespignani A (1993) Persistence of screening and self-criticality in the scale-invariant dynamics of diffusion-limited aggregation. Phys Rev Lett 70(25):3939–3942

107. Ball RC, Somfai E (2003) Diffusion-controlled growth: Theory and closure approximations. Phys Rev E 67(2):021401, part 1
108. Ristroph L, Thrasher M, Mineev-Weinstein MB, Swinney HL (2006) Fjords in viscous fingering: Selection of width and opening angle. Phys Rev E 74(1):015201, part 2
109. Mineev-Weinstein MB, Wiegmann PB, Zabrodin A (2000) Integrable structure of interface dynamics. Phys Rev Lett 84(22):5106–5109
110. Abanov A, Mineev-Weinstein MB, Zabrodin A (2007) Self-similarity in Laplacian growth. Physica D – Nonlinear Phenomena 235(1–2):62–71

Books and Reviews

Vicsek T (1992) Fractal Growth Phenomena, 2nd edn. World Scientific, Singapore
Meakin P (1998) Fractals, scaling, and growth far from equilibrium. Cambridge University Press, Cambridge
Godreche G (1991) Solids far from equilibrium. Cambridge, Cambridge, New York
Sander LM (2000) Diffusion limited aggregation, a kinetic critical phenomenon? Contemporary Physics 41:203–218
Halsey TC (2000) Diffusion-limited aggregation: A model for pattern formation. Physics Today 53(4)

Fractal and Multifractal Scaling of Electrical Conduction in Random Resistor Networks

Sidney Redner
Center for Polymer Studies and Department of Physics, Boston University, Boston, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Solving Resistor Networks
Conduction Near the Percolation Threshold
Voltage Distribution in Random Networks
Random Walks and Resistor Networks
Future Directions
Bibliography

Glossary

Conductance (G) The relation between the current I in an electrical network and the applied voltage V: $I = GV$.

Conductance exponent (t) The relation between the conductance G and the resistor (or conductor) concentration p near the percolation threshold: $G \sim (p-p_c)^t$.

Effective medium theory (EMT) A theory to calculate the conductance of a heterogeneous system that is based on a homogenization procedure.

Fractal A geometrical object that is invariant at any scale of magnification or reduction.

Multifractal A generalization of a fractal in which different subsets of an object have different scaling behaviors.

Percolation Connectivity of a random porous network.

Percolation threshold p_c The transition between a connected and disconnected network as the density of links is varied.

Random resistor network A percolation network in which the connections consist of electrical resistors that are present with probability p and absent with probability $1-p$.

Definition of the Subject

Consider an arbitrary network of nodes connected by links, each of which is a resistor with a specified electrical resistance. Suppose that this network is connected to the leads of a battery. Two natural scenarios are: (a) the "busbar geometry" (Fig. 1), in which the network is connected to two parallel lines (in two dimensions), plates (in three dimensions), etc., and the battery is connected across the two plates, and (b) the "two-point geometry", in which a battery is connected to two distinct nodes, so that a current I injected at a one node and the same current withdrawn from the other node. In both cases, a basic question is: what is the nature of the current flow through the network?

There are many reasons why current flows in resistor networks have been the focus of more than a century of research. First, understanding currents in networks is one of the earliest subjects in electrical engineering. Second, the development of this topic has been characterized by beautiful mathematical advancements, such as Kirchhoff's formal solution for current flows in networks in terms of tree matrices [52], symmetry arguments to determine the electrical conductance of continuous two-component media [10,29,48,69,74], clever geometrical methods to simplify networks [33,35,61,62], and the use of integral transform methods to solve node voltages on regular networks [6,16,94,95]. Third, the nodes voltages of a network through which a steady electrical current flows are *harmonic* [26]; that is, the voltage at a given node is a suitably-weighted average of the voltages at neighboring nodes. This same harmonicity also occurs in the probability distribution of random walks. Consequently, there are deep connections between the probability distribution of

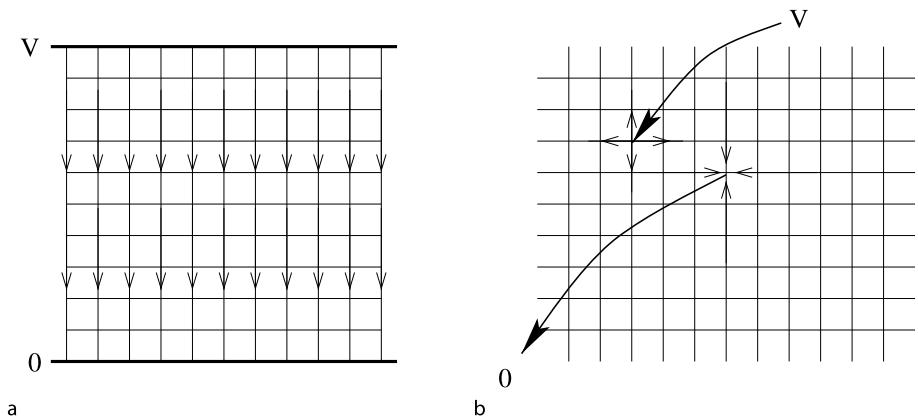

Fractal and Multifractal Scaling of Electrical Conduction in Random Resistor Networks, Figure 1
Resistor networks in a the bus-bar geometry, and b the two-point geometry

random walks on a given network and the node voltages on the same network [26].

Another important theme in the subject of resistor networks is the essential role played by *randomness* on current-carrying properties. When the randomness is weak, *effective medium theory* [10,53,54,55,57,74] is appropriate to characterize how the randomness affects the conductance. When the randomness is strong, as embodied by a network consisting of a random mixture of resistors and insulators, this *random resistor network* undergoes a transition between a conducting phase and an insulating phase when the resistor concentration passes through a percolation threshold [54]. The feature underlying this phase change is that for a small density of resistors, the network consists of disconnected clusters. However, when the resistor density passes through the percolation threshold, a macroscopic cluster of resistors spans the system through which current can flow. Percolation phenomenology has motivated theoretical developments, such as scaling, critical point exponents, and multifractals that have advanced our understanding of electrical conduction in random resistor networks.

This article begins with an introduction to electrical current flows in networks. Next, we briefly discuss analytical methods to solve the conductance of an arbitrary resistor network. We then turn to basic results related to percolation: namely, the conduction properties of a large random resistor network as the fraction of resistors is varied. We will focus on how the conductance of such a network vanishes as the percolation threshold is approached from above. Next, we investigate the more microscopic current *distribution* within each resistor of a large network. At the percolation threshold, this distribution is *multifractal* in that all moments of this distribution have independent scaling properties. We will discuss the meaning of multifractal scaling and its implications for current flows in networks, especially the largest current in the network. Finally, we discuss the relation between resistor networks and random walks and show how the classic phenomena of recurrence and transience of random walks are simply related to the conductance of a corresponding electrical network.

The subject of current flows on resistor networks is a vast subject, with extensive literature in physics, mathematics, and engineering journals. This review has the modest goal of providing an overview, from my own myopic perspective, on some of the basic properties of random resistor networks near the percolation threshold. Thus many important topics are simply not mentioned and the reference list is incomplete because of space limitations. The reader is encouraged to consult the review articles listed in the reference list to obtain a more complete perspective.

Introduction

In an elementary electromagnetism course, the following classic problem has been assigned to many generations of physics and engineering students: consider an infinite square lattice in which each bond is a 1 ohm resistor; equivalently, the conductance of each resistor (the inverse resistance) also equals 1. There are perfect electrical connections at all vertices where four resistors meet. A current I is injected at one point and the same current I is extracted at a nearest-neighbor lattice point. What is the electrical resistance between the input and output? A more challenging question is: what is the resistance between two diagonal points, or between two arbitrary points? As we

shall discuss, the latter questions can be solved elegantly using Fourier transform methods.

For the resistance between neighboring points, superposition provides a simple solution. Decompose the current source and sink into its two constituents. For a current source I, symmetry tells us that a current $I/4$ flows from the source along each resistor joined to this input. Similarly, for a current sink $-I$, a current $I/4$ flows into the sink along each adjoining resistor. For the source/sink combination, superposition tells us that a current $I/2$ flows along the resistor directly between the source and sink. Since the total current is I, a current of $I/2$ flows indirectly from source to sink via the rest of the lattice. Because the direct and indirect currents between the input and output points are the same, the resistance of the direct resistor and the resistance of rest of the lattice are the same, and thus both equal to 1. Finally, since these two elements are connected in parallel, the resistance of the infinite lattice between the source and the sink equals 1/2 (conductance 2). As we shall see in Sect. "Effective Medium Theory", this argument is the basis for constructing an effective medium theory for the conductance of a random network.

More generally, suppose that currents I_i are injected at each node of a lattice network (normally many of these currents are zero and there would be both positive and negative currents in the steady state). Let V_i denote the voltage at node i. Then by Kirchhoff's law, the currents and voltages are related by

$$I_i = \sum_j g_{ij}(V_i - V_j)\,, \tag{1}$$

where g_{ij} is the conductance of link ij, and the sum runs over all links ij. This equation simply states that the current flowing into a node by an external current source equals the current flowing out of the node along the adjoining resistors. The right-hand side of Eq. (1) is a *discrete Laplacian* operator. Partly for this reason, Kirchhoff's law has a natural connection to random walks. At nodes where the external current is zero, the node voltages in Eq. (1) satisfy

$$V_i = \frac{\sum_j g_{ij} V_j}{\sum_j g_{ij}} \to \frac{1}{z}\sum_j V_j\,. \tag{2}$$

The last step applies if all the conductances are identical; here z is the coordination number of the network. Thus for steady current flow, the voltage at each unforced node equals the weighted average of the voltages at the neighboring sites. This condition defines V_i as a *harmonic function* with respect to the weight function g_{ij}.

An important general question is the role of spatial disorder on current flows in networks. One important example is the *random resistor network*, where the resistors of a lattice are either present with probability p or absent with probability $1-p$ [54]. Here the analysis tools for regular lattice networks are no longer applicable, and one must turn to qualitative and numerical approaches to understand the current-carrying properties of the system. A major goal of this article is to outline the essential role that spatial disorder has on the current-carrying properties of a resistor network by such approaches.

A final issue that we will discuss is the deep relation between resistor networks and random walks [26,63]. Consider a resistor network in which the positive terminal of a battery (voltage $V = 1$) is connected to a set of boundary nodes, defined to be $\mathcal{B}_+$), and where a disjoint set of boundary nodes $\mathcal{B}_-$ are at $V = 0$. Now suppose that a random walk hops between nodes of the same geometrical network in which the probability of hopping from node i to node j in a single step is $g_{ij}/\sum_k g_{ik}$, where k is one of the neighbors of i and the boundary sets are absorbing. For this random walk, we can ask: what is the probability F_i for a walk to eventually be absorbed on $\mathcal{B}_+$ when it starts at node i? We shall show in Sect. "Random Walks and Resistor Networks" that F_i satisfies Eq. (2): $F_i = \sum_j g_{ij}F_j/\sum_j g_{ij}$! We then exploit this connection to provide insights about random walks in terms of known results about resistor networks and vice versa.

Solving Resistor Networks

Fourier Transform

The translational invariance of an infinite lattice resistor network with identical bond conductances $g_{ij} = 1$ cries out for applying Fourier transform methods to determine node voltages. Let's study the problem mentioned previously: what is the voltage at any node of the network when a unit current enters at some point? Our discussion is specifically for the square lattice; the extension to other lattices is straightforward.

For the square lattice, we label each site i by its x, y coordinates. When a unit current is injected at $\mathbf{r_0} = (x_0, y_0)$, Eq. (1) becomes

$$\begin{aligned} -\delta_{x,x_0}\,\delta_{y,y_0} = {} & V(x+1,y) + V(x-1,y) \\ & + V(x,y+1) + V(x,y-1) - 4V(x,y)\,, \end{aligned} \tag{3}$$

which clearly exposes the second difference operator of the discrete Laplacian. To find the node voltages, we define $V(\mathbf{k}) = \sum_{\mathbf{r}} V(\mathbf{r})\,\mathrm{e}^{i\mathbf{k}\cdot\mathbf{r}}$ and then we Fourier transform Eq. (3) to convert this infinite set of difference equations

into the single algebraic equation

$$V(\mathbf{k}) = \frac{e^{i\mathbf{k}\cdot\mathbf{r_0}}}{4 - 2(\cos k_x + \cos k_y)} \,. \tag{4}$$

Now we calculate $V(\mathbf{r})$ by inverting the Fourier transform

$$V(\mathbf{r}) = \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \frac{e^{-i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r_0})}}{4 - 2(\cos k_x + \cos k_y)} \, d\mathbf{k} \,. \tag{5}$$

Formally, at least, the solution is trivial. However, the integral in the inverse Fourier transform, known as a Watson integral [96], is non-trivial, but considerable understanding has gradually been developed for evaluating this type of integral [6,16,94,95,96].

For a unit input current at the origin and a unit sink of current at $\mathbf{r_0}$, the resistance between these two points is $V(0) - V(\mathbf{r_0})$, and Eq. (5) gives

$$\begin{aligned} R &= V(0) - V(\mathbf{r_0}) \\ &= \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \frac{(1 - e^{i\mathbf{k}\cdot\mathbf{r_0}})}{4 - 2(\cos k_x + \cos k_y)} \, d\mathbf{k} \,. \end{aligned} \tag{6}$$

Tables for the values of R for a set of closely-separated input and output points are given in [6,94]. As some specific examples, for $\mathbf{r_0} = (1, 0)$, $R = 1/2$, thus reproducing the symmetry argument result. For two points separated by a diagonal, $\mathbf{r_0} = (1, 1)$, $R = 2/\pi$. For $\mathbf{r_0} = (2, 0)$, $R = 2 - 4/\pi$. Finally, for two points separated by a knight's move, $\mathbf{r_0} = (2, 1)$, $R = 4/\pi - 1/2$.

Direct Matrix Solution

Another way to solve Eq. (1), is to recast Kirchhoff's law as the matrix equation

$$I_i = \sum_{j=1}^{N} G_{ij} V_j \,, \quad i = 1, 2, \ldots, N \tag{7}$$

where the elements of the *conductance matrix* are:

$$G_{ij} = \begin{cases} \sum_{k\neq i} g_{ik} \,, & i = j \\ -g_{ij} \,, & i \neq j \,. \end{cases}$$

The conductance matrix is an example of a *tree matrix*, as $\mathbf{G}$ has the property that the sum of any row or any column equals zero. An important consequence of this tree property is that all cofactors of G are identical and are equal to the *spanning tree polynomial* [43]. This polynomial is obtained by enumerating all possible tree graphs (graphs with no closed loops) on the original electrical network that includes each node of the network. The weight of each spanning tree is simply the product of the conductances for each bond in the tree.

Inverting Eq. (7), one obtains the voltage V_i at each node i in terms of the external currents I_j ($j = 1, 2, \ldots, N$) and the conductances g_{ij}. Thus the two-point resistance R_{ij} between two arbitrary (not necessarily connected) nodes i and j is then given by $R_{ij} = (V_i - V_j)/I$, where the network is subject to a specified external current; for example, for the two-point geometry, $I_i = 1$, $I_j = -1$, and $I_k = 0$ for $k \neq i, j$. Formally, the two-point resistance can be written as [99]

$$R_{ij} = \frac{|G^{(ij)}|}{|G^{(j)}|} \,, \tag{8}$$

where $|G^{(j)}|$ is the determinant of the conductance matrix with the jth row and column removed and $|G^{(ij)}|$ is the determinant with the ith and jth rows and columns removed. There is a simple geometric interpretation for this conductance matrix inversion. The numerator is just the spanning tree polynomial for the original network, while the denominator is the spanning tree polynomial for the network with the additional constraint that nodes i and j are identified as a single point. This result provides a concrete prescription to compute the conductance of an arbitrary network. While useful for small networks, this method is prohibitively inefficient for larger networks because the number of spanning trees grows exponentially with network size.

Potts Model Connection

The matrix solution of the resistance has an alternative and elegant formulation in terms of the spin correlation function of the q-state Potts model of ferromagnetism in the $q \to 0$ limit [88,99]. This connection between a statistical mechanical model in a seemingly unphysical limit and an enumerative geometrical problem is one of the unexpected charms of statistical physics. Another such example is the n-vector model, in which ferromagnetically interacting spins "live" in an n-dimensional spin space. In the limit $n \to 0$ [20], the spin correlation functions of this model are directly related to all self-avoiding walk configurations.

In the q-state Potts model, each site i of a lattice is occupied by a spin s_i that can assume one of q discrete values. The Hamiltonian of the system is

$$\mathcal{H} = -\sum_{i,j} J \delta_{s_i, s_j},$$

where the sum is over all nearest-neighbor interacting spin pairs, and δ_{s_i,s_j} is the Kronecker delta function ($\delta_{s_i,s_j} = 1$ if $s_i = s_j$ and $\delta_{s_i,s_j} = 0$ otherwise). Neighboring aligned spin pairs have energy $-J$, while spin pairs in different states have energy zero. One can view the spins as pointing from the center to a vertex of a q-simplex, and the interaction energy is proportional to the dot product of two interacting spins.

The partition function of a system of N spins is

$$\mathcal{Z}_N = \sum_{\{s\}} e^{\beta \sum_{i,j} J \delta_{s_i,s_j}} , \tag{9}$$

where the sum is over all 2^N spin states $\{s\}$. To make the connection to resistor networks, notice that: (i) the exponential factor associated with each link ij in the partition function takes the values 1 or $e^{\beta J}$, and (ii) the exponential of the sum can be written as the product

$$\mathcal{Z}_N = \sum_{\{s_i\}} \prod_{i,j} (1 + v\delta_{s_i,s_j}) , \tag{10}$$

with $v = \tanh \beta J$. We now make a high-temperature (small-v) expansion by multiplying out the product in (10) to generate all possible graphs on the lattice, in which each bond carries a weight $v\delta_{s_i,s_j}$. Summing over all states, the spins in each disjoint cluster must be in the same state, and the last sum over the common state of all spins leads to each cluster being weighted by a factor of q. The partition function then becomes

$$\mathcal{Z}_N = \sum_{\text{graphs}} q^{N_c} v^{N_b} , \tag{11}$$

where N_c is the number of distinct clusters and N_b is the total number of bonds in the graph.

It was shown by Kasteleyn and Fortuin [47] that the limit $q = 1$ corresponds to the percolation problem when one chooses $v = p/(1-p)$, where p is the bond occupation probability in percolation. Even more striking [34], if one chooses $v = \alpha q^{1/2}$, where α is a constant, then $\lim_{q\to 0} \mathcal{Z}_N/q^{(N+1)/2} = \alpha^{N-1} T_N$, where T_N is again the spanning tree polynomial; in the case where all interactions between neighboring spins have the same strength, then the polynomial reduces to the number of spanning trees on the lattice. It is because of this connection to spanning trees that the resistor network and Potts model are intimately connected [99]. In a similar vein, one can show that the correlation function between two spins at nodes i and j in the Potts model is simply related to the conductance between these same two nodes when the interactions J_{ij} between the spins at nodes i and j are equal to the conductances g_{ij} between these same two nodes in the corresponding resistor network [99].

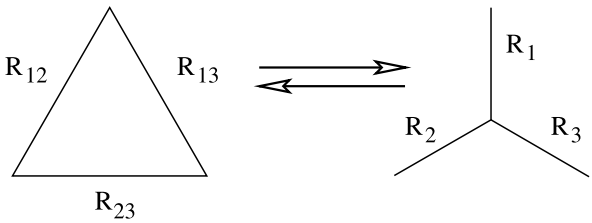

Fractal and Multifractal Scaling of Electrical Conduction in Random Resistor Networks, Figure 2
Illustration of the Δ-Y and Y-Δ transforms

Δ-Y and Y-Δ Transforms

In elementary courses on circuit theory, one learns how to combine resistors in series and parallel to reduce the complexity of an electrical circuit. For two resistors with resistances R_1 and R_2 in series, the net resistance is $R = R_1 + R_2$, while for resistors in parallel, the net resistance is $R = \left(R_1^{-1} + R_2^{-2}\right)^{-1}$. These rules provide the resistance of a network that contains only series and parallel connections. What happens if the network is more complicated? One useful way to simplify such a network is by the Δ-Y and Y-Δ transforms that was apparently first discovered by Kennelly in 1899 [50] and applied extensively since then [33,35,61,62,81].

The basic idea of the Δ-Y transform is illustrated in Fig. 2. Any triangular arrangement of resistors R_{12}, R_{13}, and R_{23} within a larger circuit can be replaced by a star, with resistances R_1, R_2, and R_3, such that all resistances between any two points among the three vertices in the triangle and the star are the same. The conditions that all two-point resistances are the same are:

$$(R_1 + R_2) = \left[R_{12}^{-1} + (R_{13} + R_{23})^{-1}\right]^{-1} \equiv a_{12} \quad + \text{cyclic permutations} .$$

Solving for R_1, R_2, and R_3 gives $R_1 = \frac{1}{2}(a_{12} - a_{23} + a_{13})$ + cyclic permutations; the explicit result in terms of the R_{ij} is:

$$R_1 = \frac{R_{12}R_{13}}{R_{12} + R_{13} + R_{23}} \quad + \text{cyclic permutations}, \tag{12}$$

as well as the companion result for the conductances $G_i = R_i^{-1}$:

$$G_1 = \frac{G_{12}G_{13} + G_{12}G_{23} + G_{13}G_{23}}{G_{23}} + \text{cyclic permutations}.$$

These relations allow one to replace any triangle by a star to reduce an electrical network.

However, sometimes we need to replace a star by a triangle to simplify a network. To construct the inverse Y-Δ

transform, notice that the Δ-Y transform gives the resistance ratios $R_1/R_2 = R_{13}/R_{23}$ + cyclic permutations, from which $R_{13} = R_{12}(R_3/R_2)$ and $R_{23} = R_{12}(R_3/R_1)$. Substituting these last two results in Eq. (12), we eliminate R_{13} and R_{23} and thus solve for R_{12} in terms of the R_i:

$$R_{12} = \frac{R_1R_2 + R_1R_3 + R_2R_3}{R_3} + \text{cyclic permutations}, \quad (13)$$

and similarly for $G_{ij} = R_{ij}^{-1}$. To appreciate the utility of the Δ-Y and Y-Δ transforms, the reader is invited to apply them on the Wheatstone bridge.

When employed judiciously and repeatedly, these transforms systematically reduce planar lattice circuits to a single bond, and thus provide a powerful approach to calculate the conductance of large networks near the percolation threshold. We will return to this aspect of the problem in Sect. "Conductance Exponent".

Effective Medium Theory

Effective medium theory (EMT) determines the macroscopic conductance of a random resistor network by a homogenization procedure [10,53,54,55,57,74] that is reminiscent of the Curie–Weiss effective field theory of magnetism. The basic idea in EMT is to replace the random network by an effective homogeneous medium in which the conductance of each resistor is determined self-consistently to optimally match the conductances of the original and homogenized systems. EMT is quite versatile and has been applied, for example, to estimate the dielectric constant of dielectric composites and the conductance of conducting composites. Here we focus on the conductance of random resistor networks, in which each resistor (with conductance g_0) is present with probability p and absent with probability $1-p$. The goal is to determine the conductance as a function of p.

To implement EMT, we first replace the random network by an effective homogeneous medium in which each bond has the same conductance g_m (Fig. 3). If a voltage is applied across this effective medium, there will be a potential drop V_m and a current $I_m = g_m V_m$ across each bond. The next step in EMT is to assign one bond in the effective medium a conductance g and adjust the external voltage to maintain a fixed total current I passing through the network. Now an additional current δi passes through the conductor g. Consequently, a current $-\delta i$ must flow through one terminal of g to the other terminal via the remainder of the network (Fig. 3). This current perturbation leads to an additional voltage drop δV across g. Thus the current-voltage relations for the marked bond and the remainder of the network are

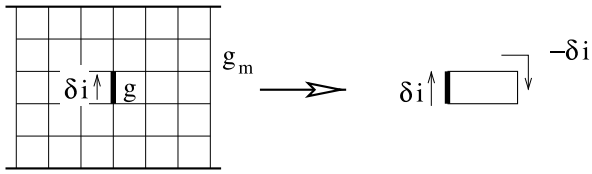

Fractal and Multifractal Scaling of Electrical Conduction in Random Resistor Networks, Figure 3
Illustration of EMT. (*left*) The homogenized network with conductances g_m and one bond with conductance g. (*right*) The equivalent circuit to the lattice

$$\begin{aligned} I_m + \delta i &= g(V_m + \delta V) \\ -\delta i &= G_{ab}\delta V\,, \end{aligned} \quad (14)$$

where G_{ab} is the conductance of the rest of the lattice between the terminals of the conductor g.

The last step in EMT is to require that the mean value δV averaged over the probability distribution of individual bond conductances is zero. Thus the effective medium "matches" the current-carrying properties of the original network. Solving Eq. (14) for δV, and using the probability distribution $P(g) = p\delta(g-g_0) + (1-p)\delta(g)$ appropriate for the random resistor network, we obtain

$$\langle \delta V \rangle = V_m \left[\frac{(g_m - g_0)p}{(G_{ab} + g_0)} + \frac{g_m(1-p)}{G_{ab}} \right] = 0\,. \quad (15)$$

It is now convenient to write $G_{ab} = \alpha g_m$, where α is a lattice-dependent constant of the order of one. With this definition, Eq. (15) simplifies to

$$g_m = g_o \frac{p(1+\alpha) - 1}{\alpha}\,. \quad (16)$$

The value of α – the proportionality constant for the conductance of the initial lattice with a single bond removed – can usually be determined by a symmetry argument of the type presented in Sect. "Introduction to Current Flows". For example, for the triangular lattice (coordination number 6), the conductance $G_{ab} = 2g_m$ and $\alpha = 2$. For the hypercubic lattice in d dimensions (coordination number $z = 2^d$), $G_{ab} = ((z-2)/2)g_m$.

The main features of the effective conductance g_m that arises from EMT are: (i) the conductance vanishes at a lattice-dependent percolation threshold $p_c = 1/(1+\alpha)$; for the hypercubic lattice $\alpha = (z-2)/2$ and the percolation threshold $p_c = 2/z = 2^{1-d}$ (fortuitously reproducing the exact percolation threshold in two dimensions); (ii) the conductance varies linearly with p and vanishes linearly in $p - p_c$ as p approaches p_c from above. The linearity

of the effective conductance away from the percolation threshold accords with numerical and experimental results. However, EMT fails near the percolation threshold, where large fluctuations arise that invalidate the underlying assumptions of EMT. In this regime, alternative methods are needed to estimate the conductance.

Conduction Near the Percolation Threshold

Scaling Behavior

EMT provides a qualitative but crude picture of the current-carrying properties of a random resistor network. While EMT accounts for the existence of a percolation transition, it also predicts a linear dependence of the conductance on p. However, near the percolation threshold it is well known that the conductance varies non-linearly in $p - p_c$ near p_c [85]. This non-linearity defines the conductance exponent t by

$$G \sim (p - p_c)^t \qquad p \downarrow p_c \,, \tag{17}$$

and much research on random resistor networks [85] has been performed to determine this exponent. The conductance exponent generically depends only on the spatial dimension of the network and not on any other details (a notable exception, however, is when link resistances are broadly distributed, see [40,93]). This *universality* is one of the central tenets of the theory of critical phenomena [64,83]. For percolation, the mechanism underlying universality is the absence of a characteristic length scale; as illustrated in Fig. 4, clusters on all length scales exist when a network is close to the percolation threshold.

The scale of the largest cluster defines the correlation length ξ by $\xi \sim (p_c - p)^{-\nu}$ as $p \to p_c$. The divergence in ξ also applies for $p > p_c$ by defining the correlation length as the typical size of finite clusters only (Fig. 4), thus eliminating the infinite percolating cluster from consideration. At the percolation threshold, clusters on all length scales exist, and the absence of a characteristic length implies that the singularity in the conductance should not depend on microscopic variables. The only parameter remaining upon which the conductance exponent t can depend upon is the spatial dimension d [64,83]. As typifies critical phenomena, the conductance exponent has a constant value in all spatial dimensions $d > d_c$, where d_c is the upper critical dimension which equals 6 for percolation [22]. Above this critical dimension, mean-field theory (not to be confused with EMT) gives the correct values of critical exponents.

While there does not yet exist a complete theory for the dimension dependence of the conductance exponent below the critical dimension, a crude but useful *nodes, links, and blobs* picture of the infinite cluster [21,82,84] provides partial information. The basic idea of this picture is that

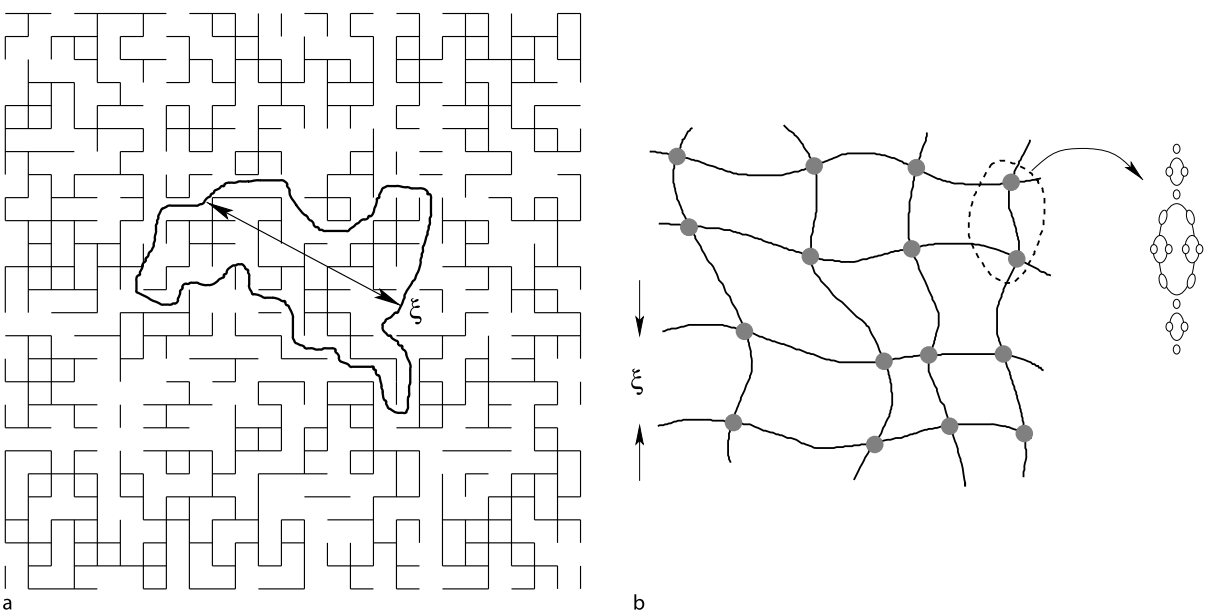

Fractal and Multifractal Scaling of Electrical Conduction in Random Resistor Networks, Figure 4
(*left*) Realization of bond percolation on a 25 × 25 square lattice at p=0.505. (*right*) Schematic picture of the nodes (shaded circles), links and blobs picture of percolation for $p \gtrsim p_c$

for $p \gtrsim p_c$, a large system has an irregular network-like topology that consists of quasi-linear chains that are separated by the correlation length ξ (Fig. 4). For a macroscopic sample of linear dimension L with a bus bar-geometry, the percolating cluster above p_c then consists of $(L/\xi)^{d-1}$ statistically identical chains in parallel, in which each chain consists of L/ξ macrolinks in series, and the macrolinks consists of nested blob-like structures.

The conductance of a macrolink is expected to vanish as $(p - p_c)^{\zeta}$, with ζ a new unknown exponent. Although a theory for the conductance of a single macrolink, and even a precise definition of a macrolink, is still lacking, the nodes, links, and blobs picture provides a starting point for understanding the dimension dependence of the conductance exponent. Using the rules for combining parallel and series conductances, the conductance of a large resistor network of linear dimension L is then

$$G(p, L) \sim \left(\frac{L}{\xi}\right)^{d-1} \frac{(p - p_c)^{\zeta}}{L/\xi} \sim L^{d-2}\,(p - p_c)^{(d-2)\nu+\zeta} \,. \tag{18}$$

In the limit of large spatial dimension, we expect that a macrolink is merely a random walk between nodes. Since the spatial separation between nodes is ξ, the number of bonds in the macrolink, and hence its resistance, scales as ξ^2 [92]. Using the mean-field result $\xi \sim (p - p_c)^{-1/2}$, the resistance of the macrolink scales as $(p - p_c)^{-1}$ and thus the exponent $\zeta = 1$. Using the mean-field exponents $\nu = 1/2$ and $\zeta = 1$ at the upper critical dimension of $d_c = 6$, we then infer the mean-field value of the conductance exponent $t = 3$ [22,91,92].

Scaling also determines the conductance of a finite-size system of linear dimension L *exactly at* the percolation threshold. Although the correlation length formally diverges when $p - p_c = 0$, ξ is limited by L in a finite system of linear dimension L. Thus the only variable upon which the conductance can depend is L itself. Equivalently, deviations in $p - p_c$ that are smaller than $L^{-1/\nu}$ cannot influence critical behavior because ξ can never exceed L. Thus to determine the dependence of a singular observable for a finite-size system at p_c, we may replace $(p - p_c)$ by $L^{-1/\nu}$. By this prescription, the conductance at p_c of a large finite-size system of linear dimension L becomes

$$G(p_c, L) \sim L^{d-2}(L^{-1/\nu})^{(d-2)\nu+\zeta} \sim L^{-\zeta/\nu} \,. \tag{19}$$

In this *finite-size scaling* [85], we fix the occupation probability to be exactly at p_c and study the dependence of an observable on L to determine percolation exponents. This approach provides a convenient and more accurate method to determine the conductance exponent compared to studying the dependence of the conductance of a large system as a function of $p - p_c$.

Conductance Exponent

In percolation and in the random resistor network, much effort has been devoted to computing the exponents that characterize basic physical observables – such as the correlation length ξ and the conductance G – to high precision. There are several reasons for this focus on exponents. First, because of the universality hypothesis, exponents are a meaningful quantifier of phase transitions. Second, various observables near a phase transition can sometimes be related by a scaling argument that leads to a corresponding exponent relation. Such relations may provide a decisive test of a theory that can be checked numerically. Finally, there is the intellectual challenge of developing accurate numerical methods to determine critical exponents. The best such methods have become quite sophisticated in their execution.

A seminal contribution was the "theorists' experiment" of Last and Thouless [58] in which they punched holes at random in a conducting sheet of paper and measured the conductance of the sheet as a function of the area fraction of conducting material. They found that the conductance vanished faster than linearly with $(p - p_c)$; here p corresponds to the area fraction of the conductor. Until this experiment, there was a sentiment that the conductance should be related to the fraction of material in the percolating cluster [30] – the percolation probability $P(p)$ – a quantity that vanished slower than linearly with $(p - p_c)$. The reason for this disparity is that in a resistor network, much of the percolating cluster consists of *dangling ends* – bonds that carry no current – and thus make no contribution to the conductance. A natural geometrical quantity that ought to be related to the conductance is the fraction of bonds $B(p)$ in the *conducting backbone* – the subset of the percolating cluster without dangling ends. However, a clear relation between the conductivity and a geometrical property of the backbone has not yet been established.

Analytically, there are primary two methods that have been developed to compute the conductance exponent: the renormalization group [44,86,87,89] and low-density series expansions [1,2,32]. In the real-space version of the renormalization group, the evolution of conductance distribution under length rescaling is determined, while the momentum-space version involves a diagrammatic implementation of this length rescaling in momentum space. The latter is a perturbative approach away from mean-field theory in the variable $6 - d$ that become exact as $d \to 6$.

Considerable effort has been devoted to determining the conductance exponent by numerical and algorithmic methods. Typically, the conductance is computed for networks of various linear dimensions L at $p = p_c$, and the conductance exponent is extracted from the L dependence of the conductance, which should vanish as $L^{-\zeta/\nu}$. An exact approach, but computationally impractical for large networks, is Gauss elimination to invert the conductance matrix [79]. A simple approximate method is Gauss relaxation [59,68,80,90,97] (and its more efficient variant of Gauss–Seidel relaxation [71]). This method uses Eq. (2) as the basis for an iteration scheme, in which the voltage V_i at node i at the nth update step is computed from (2) using the values of V_j at the $(n-1)^{\text{st}}$ update in the right-hand side of this equation. However, one can do much better by the conjugate gradient algorithm [27] and speeding up this method still further by Fourier acceleration methods [7].

Another computational approach is based on the node elimination method, in which the Δ-Y and Y-Δ transforms are used to successively eliminate bonds from the network and ultimately reduce a large network to a single bond [33,35,62]. In a different vein, the transfer matrix method has proved to be extremely accurate and efficient [24,25,70,100]. The method is based on building up the network one bond at a time and immediately calculating the conductance of the network after each bond addition. This method is most useful when applied to very long strips of transverse dimension L so that a single realization gives an accurate value for the conductance.

As a result of these investigations, as well as by series expansions for the conductance, the following exponents have been found. For $d = 2$, where most of the computational effort has been applied, the best estimate [70] for the exponent t (using $\zeta = t$ in $d = 2$ only) is $t = 1.299 \pm 0.002$. One reason for the focus on two dimensions is that early estimates for t were tantalizingly close to the correlation length exponent ν that is now known to exactly equal 4/3 [23]. Another such connection was the Alexander–Orbach conjecture [5], which predicted $t = 91/72 = 1.2638\ldots$, but again is incompatible with the best numerical estimate for t. In $d = 3$, the best available numerical estimate for t appears to be $t = 2.003 \pm 0.047$ [12,36], while the low concentration series method gives an equally precise result of $t = 2.02 \pm 0.05$ [1,2]. These estimates are just compatible with the rigorous bound that $t \leq 2$ in $d = 3$ [37,38]. In greater than three dimensions, these series expansions give $t = 2.40 \pm 0.03$ for $d = 4$ and $t = 2.74 \pm 0.03$ for $d = 5$, and the dimension dependence is consistent with $t = 3$ when d reaches 6.

Voltage Distribution in Random Networks

Multifractal Scaling

While much research has been devoted to understanding the critical behavior of the conductance, it was realized that the *distribution* of voltages across each resistor of the network was quite rich and exhibited *multifractal* scaling [17,19,72,73]. Multifractality is a generalization of fractal scaling in which the distribution of an observable is sufficiently broad that different moments of the distribution scale independently. Such multifractal scaling arises in phenomena as diverse as turbulence [45,66], localization [13], and diffusion-limited aggregation [41,42]. All these diverse examples showed scaling properties that were much richer than first anticipated.

To make the discussion of multifractality concrete, consider the example of the Maxwell–Boltzmann velocity distribution of a one-dimensional ideal gas

$$P(v) = \sqrt{\frac{m}{2\pi k_{\mathrm{B}} T}}\, \mathrm{e}^{-mv^2/2k_{\mathrm{B}}T} \equiv \frac{1}{\sqrt{2\pi v_{th}^2}}\, \mathrm{e}^{-v^2/2v_{th}^2} ,$$

where k_{B} is Boltzmann's constant, m is the particle mass, T is the temperature, and $v_{th} = \sqrt{k_{\mathrm{B}}T/m}$ is the characteristic thermal velocity. The even integer moments of the velocity distribution are

$$\langle (v^2)^n \rangle \propto \left(v_{th}^2\right)^n \equiv \left(v_{th}^2\right)^{p(n)} .$$

Thus a single velocity scale, v_{th}, characterizes all positive moments of the velocity distribution. Alternatively, the exponent $p(n)$ is linear in n. This linear dependence of successive moment exponents characterizes single-parameter scaling. The new feature of multifractal scaling is that a wide range of scales characterizes the voltage distribution (Fig. 5). As a consequence, the moment exponent $p(n)$ is a non-linear function of n.

One motivation for studying the voltage distribution is its relation to basic aspects of electrical conduction. If a voltage $V = 1$ is applied across a resistor network, then the conductance G and the total current flow I are equal: $I = G$. Consider now the power dissipated through the network $P = IV = GV^2 \to G$. We may also compute the dissipated power by adding up these losses in each resistor to give

$$P = G = \sum_{ij} g_{ij} V_{ij}^2 \to \sum_{ij} V_{ij}^2 = \sum_V V^2\, N(V) . \quad (20)$$

Here $g_{ij} = 1$ is the conductance of resistor ij, and V_{ij} is the corresponding voltage drop across this bond. In the last equality, $N(V)$ is the number of resistors with a voltage

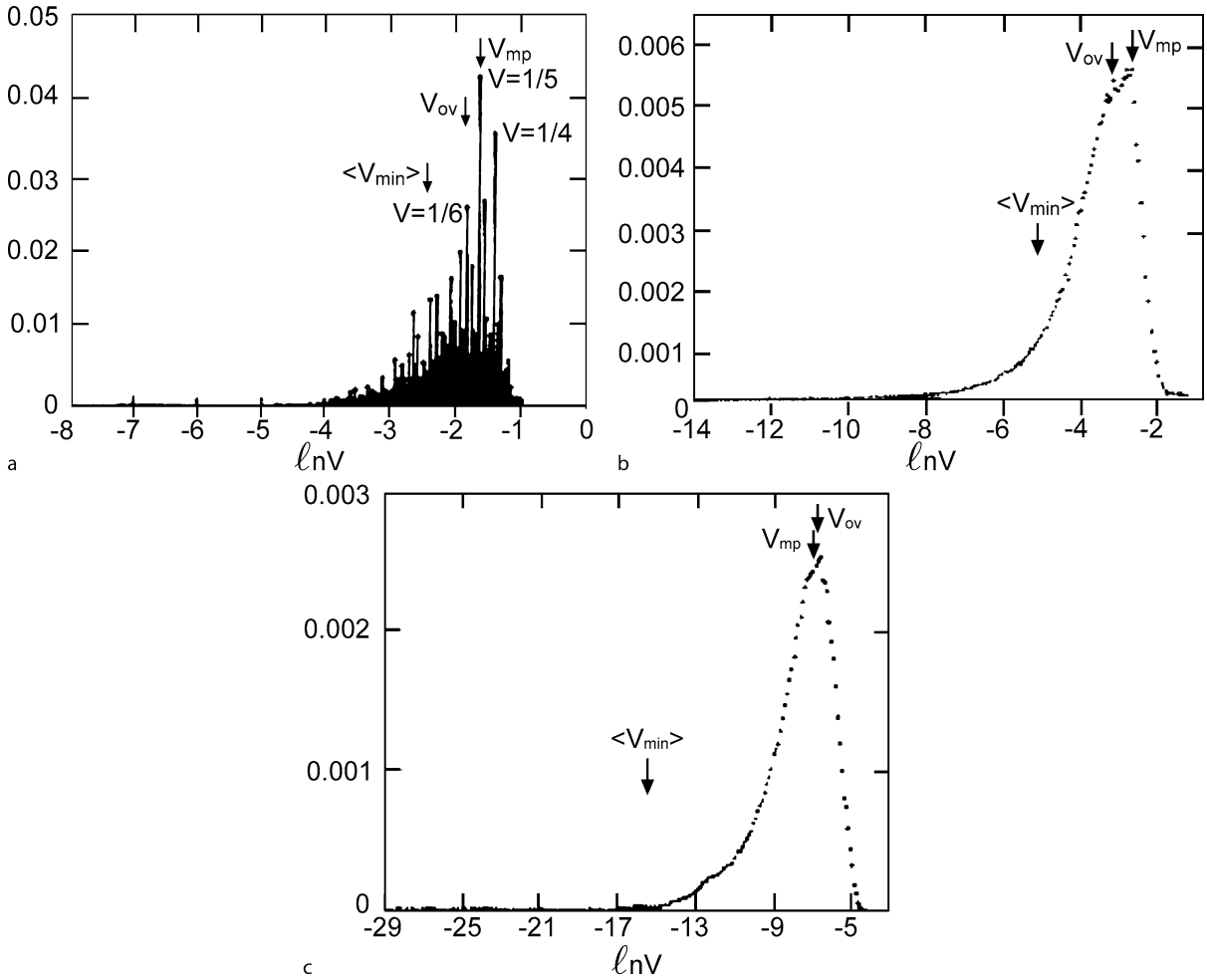

Fractal and Multifractal Scaling of Electrical Conduction in Random Resistor Networks, Figure 5

The voltage distribution on an $L \times L$ square lattice random resistor network at the percolation threshold for a $L = 4$ (exact), b $L = 10$, and c $L = 130$. The latter two plots are based on simulation data. For $L = 4$, a number of peaks, that correspond to simple rational fractions of the unit potential drop, are indicated. Also shown are the average voltage over all realizations, V_{av}, the most probable voltage, V_{mp}, and the average of the minimum voltage in each realization, $\langle V_{min} \rangle$. [Reprinted from Ref. [19]]

drop V. Thus the conductance is just the second moment of the distribution of voltage drops across each bond in the network.

From the statistical physics perspective it is natural to study other moments of the voltage distribution and the voltage distribution itself. Analogous to the velocity distribution, we define the family of exponents $p(k)$ for the scaling dependence of the voltage distribution at $p = p_c$ by

$$\mathcal{M}(k) \equiv \sum_V N(V) V^k \sim L^{-p(k)/\nu} . \tag{21}$$

Since $\mathcal{M}(2)$ is just the network conductance, $p(2) = \zeta$. Other moments of the voltage distribution also have simple interpretations. For example, $\langle V^4 \rangle$ is related to the magnitude of the noise in the network [9,73], while $\langle V^k \rangle$ for $k \to \infty$ weights the bonds with the highest currents, or the "hottest" bonds of the network, most strongly, and they help understand the dynamics of fuse networks of failure [4,18]. On the other hand, negative moments weight low-current bonds more strongly and emphasize the low-voltage tail of the distribution. For example, $\mathcal{M}(-1)$ characterizes hydrodynamic dispersion [56], in which passive tracer particles disperse in a network due to a multiplicity of network paths. In hydrodynamics dispersion, the transit time across each bond is proportional to the inverse of the current in the bond, while the probability for tracer to enter a bond is proportional to the enter-

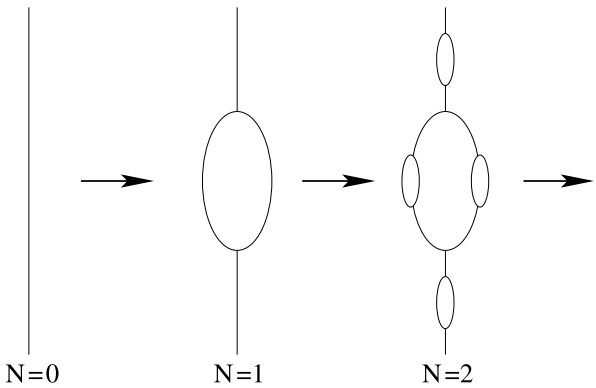

Fractal and Multifractal Scaling of Electrical Conduction in Random Resistor Networks, Figure 6
The first few iterations of a hierarchical model

ing current. As a result, the kth moment of the transit time distribution varies as $\mathcal{M}(-k+1)$, so that the quantity that quantifies dispersion, $\langle t^2\rangle - \langle t\rangle^2$, scales as $\mathcal{M}(-1)$.

A simple fractal model [3,11,67] of the conducting backbone (Fig. 6) illustrates the multifractal scaling of the voltage distribution near the percolation threshold [17]. To obtain the Nth-order structure, each bond in the $(N-1)^{\text{st}}$ iteration is replaced by the first-order structure. The resulting fractal has a hierarchical embedding of links and blobs that captures the basic geometry of the percolating backbone. Between successive generations, the length scale changes by a factor of 3, while the number of bonds changes by a factor of 4. Defining the fractal dimension d_f as the scaling relation between mass ($M = 4^N$) and the length scale ($\ell = 3^N$) via $M \sim \ell^{d_f}$, gives a fractal dimension $d_f = \ln 4/\ln 3$.

Now let's determine the distribution of voltage drops across the bonds. If a unit voltage is applied at the opposite ends of a first-order structure ($N = 1$) and each bond is a 1 ohm resistor, then the two resistors in the central bubble each have a voltage drop of 1/5, while the two resistors at the ends have a voltage drop 2/5. In an Nth-order hierarchy, the voltage of any resistor is the product of these two factors, with number of times each factor occurs dependent on the level of embedding of a resistor within the blobs. It is a simple exercise to show that the voltage distribution is [17]

$$N(V(j)) = 2^N \binom{N}{j}, \tag{22}$$

where the voltage $V(j)$ can take the values $2^j/5^N$ (with $j = 0, 1, \ldots, N$). Because j varies logarithmically in V, the voltage distribution is log binomial [75]. Using this distribution in Eq. (21), the moments of the voltage distribution are

$$\mathcal{M}(k) = \left[\frac{2(1+2^k)}{5^k}\right]^N . \tag{23}$$

In particular, the average voltage, $\mathcal{M}(1)/\mathcal{M}(0) \equiv V_{\text{av}}$ equals $((3/2)/5)^N$, which is very different from the most probable voltage, $V_{\text{mp}} = (\sqrt{2}/5)^N$ as $N \to \infty$. The underlying multiplicativity of the bond voltages is the ultimate source of the large disparity between the average and most probable values.

To calculate the moment exponent $p(k)$, we first need to relate the iteration index N to a physical length scale. For percolation, the appropriate relation is based on Coniglio's theorem [15], which is a simple but profound statement about the structure of the percolating cluster. This theorem states that the number of singly-connected bonds in a system of linear dimension L, $\mathcal{N}_s$, varies as $L^{1/\nu}$. Singly-connected bonds are those that would disconnect the network if they were cut. An equivalent form of the theorem is $\mathcal{N}_s = \partial p'/\partial p$, where p' is the probability that a spanning cluster exists in the system. This relation reflects the fact that when p is decreased slightly, p' changes only if a singly-connected bond happens to be deleted.

In the Nth-order hierarchy, the number of such singly-connected links is simply 2^N. Equating these two gives an effective linear dimension, $L = 2^{N\nu}$. Using this relation in (23), the moment exponent $p(k)$ is

$$p(k) = k - 1 + \left[k \ln(5/4) - \ln(1+2^{-k})\right] / \ln 2 . \tag{24}$$

Because each $p(k)$ is independent, the moments of the voltage distribution are characterized by an infinite set of exponents. Equation (24) is in excellent agreement with numerical data for the voltage distribution in two-dimensional random resistor networks at the percolation threshold [19]. A similar multifractal behavior was also found for the voltage distribution of the resistor network at the percolation threshold in three dimensions [8].

Maximum Voltage

An important aspect of the voltage distribution, both because of its peculiar scaling properties [27] and its application to breakdown problems [18,27], is the maximum voltage in a network. The salient features of this maximum voltage are: (i) logarithmic scaling as a function of system size [14,27,28,60,65], and (ii) non-monotonic dependence on the resistor concentration p [46]. The former property is a consequence of the expected size of the largest defect

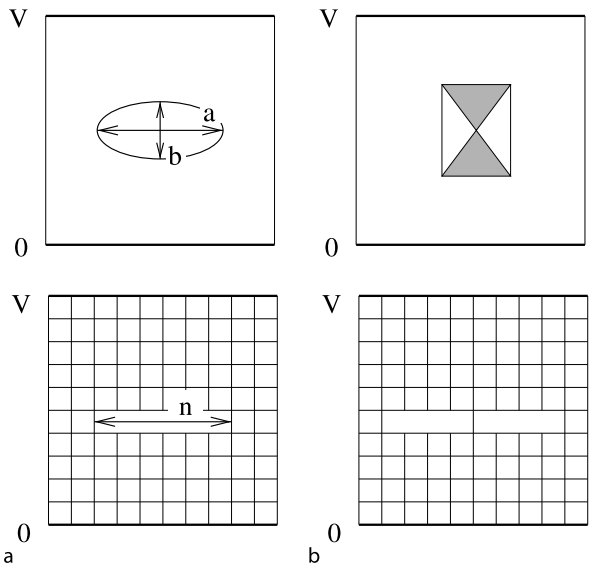

Fractal and Multifractal Scaling of Electrical Conduction in Random Resistor Networks, Figure 7

Defect configurations in two dimensions. a An ellipse and its square lattice counterpart, b a funnel, with the region of good conductor shown shaded, and a 2-slit configuration on the square lattice

in the network that gives maximal local currents. Here, we use the terms maximum local voltage and maximum local current interchangeably because they are equivalent.

To find the maximal current, we first need to identify the optimal defects that lead to large local currents. A natural candidate is an ellipse [27,28] with major and minor axes a and b (continuum), or its discrete analog of a linear crack (hyperplanar crack in greater than two dimensions) in which n resistors are missing (Fig. 7). Because current has to detour around the defect, the local current at the ends of the defect is magnified. For the continuum problem, the current at the tip of the ellipse is $I_{\text{tip}} = I_0(1+a/b)$, where I_0 is the current in the unperturbed system [27]. For the maximum current in the lattice system, one must integrate the continuum current over a one lattice spacing and identify a/b with n [60]. This approach gives the maximal current at the tip of a crack $I_{\max} \propto (1+n^{1/2})$ in two dimensions and as $I_{\max} \propto (1+n^{1/2(d-1)})$ in d dimensions.

Next, we need to find the size of the largest defect, which is an extreme-value statistics exercise [39]. For a linear crack, each broken bond occurs with probability $1-p$, so that the probability for a crack of length n is $(1-p)^n \equiv \mathrm{e}^{-an}$, with $a = -\ln(1-p)$. In a network of volume L^d, we estimate the size of the largest defect by $L^d \int_{n_{\max}}^{\infty} \mathrm{e}^{-an}\,\mathrm{d}x = 1$; that is, there exists of the order of one defect of size $n_{\max}$ or larger in the network [39]. This estimate gives $n_{\max}$ varying as $\ln L$. Combining this result with the current at the tip of a crack of length n, the largest current in a system of linear dimension L scales as $(\ln L)^{1/2(d-1)}$.

A more thorough analysis shows, however, that a single crack is not quite optimal. For a continuum two-component network with conductors of resistance 1 with probability p and with resistance $r > 1$ with probability $1-p$, the configuration that maximizes the local current is a funnel [14,65]. For a funnel of linear dimension ℓ, the maximum current at the apex of the funnel is proportional to $\ell^{1-\nu}$, where $\nu = (4/\pi)\tan^{-1}(r^{-1/2})$ [14,65]. The probability to find a funnel of linear dimension ℓ now scales as $\mathrm{e}^{-b\ell^2}$ (exponentially in its area), with b a constant. By the same extreme statistics reasoning given above, the size of the largest funnel in a system of linear dimension L then scales as $(\ln L)^{1/2}$, and the largest expected current correspondingly scales as $(\ln L)^{(1-\nu)/2}$. In the limit $r \to \infty$, where one component is an insulator, the optimal discrete configuration in two dimensions becomes two parallel slits, each of length n, between which a single resistor remains [60]. For this two-slit configuration, the maximum current is proportional to n in two dimensions, rather than $n^{1/2}$ for the single crack. Thus the maximal current in a system of linear dimension L scales as $\ln L$ rather than as a fractional power of $\ln L$.

The p dependence of the maximum voltage is intriguing because it is non-monotonic. As p, the fraction of occupied bonds, decreases from 1, less total current flows (for a fixed overall voltage drop) because the conductance is decreasing, while local current in a funnel is enhanced because such defects grow larger. The competition between these two effects leads to $V_{\max}$ attaining its peak at p_{peak} above the percolation threshold that only slowly approaches p_c as $L \to \infty$. An experimental manifestation of this non-monotonicity in $V_{\max}$ occurred in a resistor-diode network [77], where the network reproducibly burned (solder connections melting and smoking) when $p \simeq 0.77$, compared to a percolation threshold of $p_c \simeq 0.58$. Although the directionality constraint imposed by diodes enhances funneling, similar behavior should occur in a random resistor network.

The non-monotonic p dependence of $V_{\max}$ can be understood within the quasi-one-dimensional "bubble" model [46] that captures the interplay between local funneling and overall current reduction as p decreases (Fig. 8). Although this system looks one-dimensional, it can be engineered to reproduce the percolation properties of a system in greater than one dimension by choosing the length L to scale exponentially with the width w. The prob-

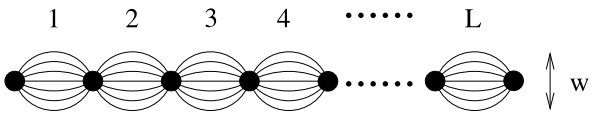

Fractal and Multifractal Scaling of Electrical Conduction in Random Resistor Networks, Figure 8
The bubble model: a chain of L bubbles in series, each consisting of w bonds in parallel. Each bond is independently present with probability p

ability for a spanning path in this structure is

$$p' = \left[1-(1-p)^w\right]^L \to \exp[-L\,\mathrm{e}^{-pw}] \qquad L, w \to \infty, \quad (25)$$

which suddenly changes from 0 to 1 – indicative of percolation – at a threshold that lies strictly within (0,1) as $L \to \infty$ and $L \sim \mathrm{e}^w$. In what follows, we take $L = 2^w$, which gives $p_c = 1/2$.

To determine the effect of bottlenecking, we appeal to the statement of Coniglio's theorem [15], $\partial p'/\partial p$ equals the average number of singly-connected bonds in the system. Evaluating $\partial p'/\partial p$ in Eq. (25) at the percolation threshold of $p_c = 1/2$ gives

$$\frac{\partial p'}{\partial p} = w + \mathcal{O}(\mathrm{e}^{-w}) \sim \ln L\,. \quad (26)$$

Thus at p_c there are $w \sim \ln L$ bottlenecks. However, current focusing due to bottlenecks is substantially diluted because the conductance, and hence the total current through the network, is small at p_c. What is needed is a single bottleneck of width 1. One such bottleneck ensures the total current flow is still substantial, while the narrowing to width 1 endures that the focusing effect of the bottleneck is maximally effective.

Clearly, a single bottleneck of width 1 occurs above the percolation threshold. Thus let's determine when a such an isolated bottleneck of width 1 first appears as a function of p. The probability that a single non-empty bubble contains at least two bonds is $(1-q^w-wpq^{w-1})/(1-q^w)$, where $q = 1-p$. Then the probability $P_1(p)$ that the width of the narrowest bottleneck has width 1 in a chain of L bubbles is

$$P_1(p) = 1 - \left(1 - \frac{wpq^{w-1}}{1-q^w}\right)^L \sim 1 - \exp\left(-Lw\frac{p}{q(1-q^w)}\,\mathrm{e}^{-pw}\right)\,. \quad (27)$$

The subtracted term is the probability that L non-empty bubbles contain at least two bonds, and then $P_1(p)$ is the complement of this quantity. As p decreases from 1, $P_1(p)$ sharply increases from 0 to 1 when the argument of the outer exponential becomes of the order of 1; this change occurs at $\hat{p} \sim p_c + \mathcal{O}(\ln(\ln L)/\ln L)$. At this point, a bottleneck of width 1 first appears and therefore $V_{\max}$ also occurs for this value of p.

Random Walks and Resistor Networks

The Basic Relation

We now discuss how the voltages at each node in a resistor network and the resistance of the network are directly related to *first-passage* properties of random walks [31,63,76,98]. To develop this connection, consider a random walk on a finite network that can hop between nearest-neighbor sites i to j with probability p_{ij} in a single step. We divide the boundary points of the network into two disjoint classes, $\mathcal{B}_+$ and $\mathcal{B}_-$, that we are free to choose; a typical situation is the geometry shown in Fig. 9. We now ask: starting at an arbitrary point i, what is the probability that the walk *eventually* reaches the boundary set $\mathcal{B}_+$ without first reaching any node in $\mathcal{B}_-$? This quantity is termed the exit probability $\mathcal{E}_+(i)$ (with an analogous definition for the exit probability $\mathcal{E}_-(i) = 1 - \mathcal{E}_+(i)$ to $\mathcal{B}_-$).

We obtain the exit probability $\mathcal{E}_+(i)$ by summing the probabilities for all walk trajectories that start at i and reach a site in $\mathcal{B}_+$ without touching any site in $\mathcal{B}_-$ (and similarly for $\mathcal{E}_-(i)$). Thus

$$\mathcal{E}_\pm(i) = \sum_{p_\pm} \mathcal{P}_{p_\pm}(i)\,, \quad (28)$$

where $\mathcal{P}_{p_\pm}(i)$ denotes the probability of a path from i to $\mathcal{B}_\pm$ that avoids $\mathcal{B}_\mp$. The sum over all these restricted paths

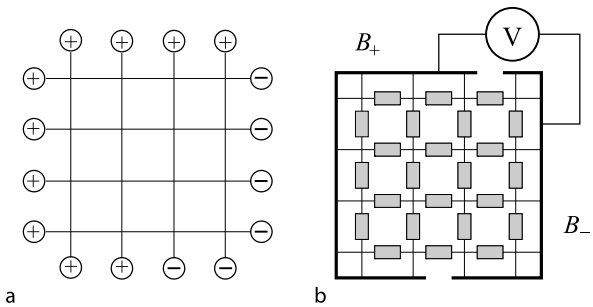

Fractal and Multifractal Scaling of Electrical Conduction in Random Resistor Networks, Figure 9
a Lattice network with boundary sites $\mathcal{B}_+$ or $\mathcal{B}_-$. b Corresponding resistor network in which each rectangle is a 1 ohm resistor. The sites in $\mathcal{B}_+$ are all fixed at potential $V = 1$, and sites in $\mathcal{B}_-$ are all grounded

can be decomposed into the outcome after one step, when the walk reaches some intermediate site j, and the sum over all path remainders from j to $\mathcal{B}_\pm$. This decomposition gives

$$\mathcal{E}_\pm(i) = \sum_j p_{ij}\, \mathcal{E}_\pm(j)\,. \tag{29}$$

Thus $\mathcal{E}_\pm(i)$ is a harmonic function because it equals a weighted average of $\mathcal{E}_\pm$ at neighboring points, with weighting function p_{ij}. This is exactly the same relation obeyed by the node voltages in Eq. (2) for the corresponding resistor network when we identify the single-step hopping probabilities p_{ij} with the conductances $g_{ij} / \sum_j g_{ij}$. We thus have the following equivalence:

- Let the boundary sets $\mathcal{B}_+$ and $\mathcal{B}_-$ in a resistor network be fixed at voltages 1 and 0 respectively, with g_{ij} the conductance of the bond between sites i and j. Then the voltage at any interior site i coincides with the probability for a random walk, which starts at i, to reach $\mathcal{B}_+$ before reaching $\mathcal{B}_-$, when the hopping probability from i to j is $p_{ij} = g_{ij} / \sum_j g_{ij}$.

If all the bond conductances are the same – corresponding to single - step hopping probabilities in the equivalent random walk being identical – then Eq. (29) is just the discrete Laplace equation. We can then exploit this correspondence between conductances and hopping probabilities to infer non-trivial results about random walks and about resistor networks from basic electrostatics. This correspondence can also be extended in a natural way to general random walks with a spatially-varying bias and diffusion coefficient, and to continuous media.

The consequences of this equivalence between random walks and resistor networks is profound. As an example [76], consider a diffusing particle that is initially at distance r_0 from the center of a sphere of radius $a < r_0$ in otherwise empty d-dimensional space. By the correspondence with electrostatics, the probability that this particle eventually hits the sphere is simply the electrostatic potential at r_0 , $\mathcal{E}_-(r_0) = (a/r_0)^{d-2}$!

Network Resistance and Pólya's Theorem

An important extension of the relation between exit probability and node voltages is to infinite resistor networks. This extension provides a simple connection between the classic recurrence/transience transition of random walks on a given network [31,63,76,98] and the electrical resistance of this same network [26]. Consider a symmetric random walk on a regular lattice in d spatial dimensions. Suppose that the walk starts at the origin at $t = 0$. What is the probability that the walk *eventually* returns to its starting point? The answer is strikingly simple:

- For $d \le 2$, a random walk is *certain* to eventually return to the origin. This property is known as *recurrence*.
- For $d > 2$, there is a non-zero probability that the random walk will *never* return to the origin. This property is known as *transience*.

Let's now derive the transience and recurrence properties of random walks in terms of the equivalent resistor network problem. Suppose that the voltage V at the boundary sites $\mathcal{B}_+$ is set to one. Then by Kirchhoff's law, the total current entering the network is

$$I = \sum_j (1 - V_j) g_{+j} = \sum_j (1 - V_j) p_{+j} \sum_k g_{+k}\,. \tag{30}$$

Here g_{+j} is the conductance of the resistor between $\mathcal{B}_+$ and a neighboring site j, and $p_{+j} = g_{+j} / \sum_j g_{+j}$. Because the voltage V_j also equals the probability for the corresponding random walk to reach $\mathcal{B}_+$ without reaching $\mathcal{B}_-$, the term $V_j\, p_{+j}$ is just the probability that a random walk starts at $\mathcal{B}_+$, makes a single step to one of the sites j adjacent to $\mathcal{B}_+$ (with hopping probability p_{ij}), and then returns to $\mathcal{B}_+$ without reaching $\mathcal{B}_-$. We therefore deduce that

$$\begin{aligned} I &= \sum_j (1 - V_j) g_{+j} \\ &= \sum_k g_{+k} \sum_j (1 - V_j) p_{+j} \\ &= \sum_k g_{+k} \times (1 - \text{return probability}) \\ &= \sum_k g_{+k} \times \text{ escape probability }\,. \end{aligned} \tag{31}$$

Here "escape" means that the random walk reaches the set $\mathcal{B}_-$ without returning to a node in $\mathcal{B}_+$.

On the other hand, the current and the voltage drop across the network are related to the conductance G between the two boundary sets by $I = GV = G$. From this fact, Eq. (31) gives the fundamental result

$$\text{escape probability } \equiv P_{\text{escape}} = \frac{G}{\sum_k g_{+k}}\,. \tag{32}$$

Suppose now that a current I is injected at a single point of an infinite network, with outflow at infinity (Fig. 10). Thus the probability for a random walk to never return to its

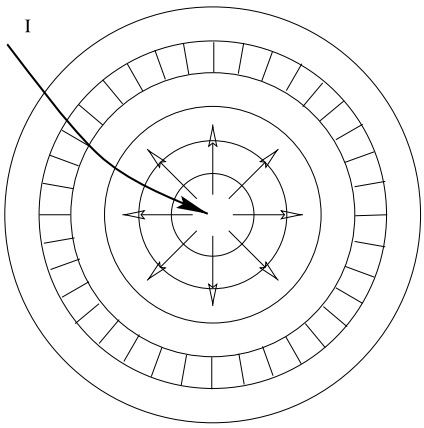

Fractal and Multifractal Scaling of Electrical Conduction in Random Resistor Networks, Figure 10
Decomposition of a conducting medium into concentric shells, each of which consists of fixed-conductance blocks. A current *I* is injected at the origin and flows radially outward through the medium

starting point, is simply proportional to the conductance G from this starting point to infinity of the same network. Thus a subtle feature of random walks, namely, the escape probability, is directly related to currents and voltages in an equivalent resistor network.

Part of the reason why this connection is so useful is that the conductance of the infinite network for various spatial dimensions can be easily determined, while a direct calculation of the return probability for a random walk is more difficult. In one dimension, the conductance of an infinitely long chain of identical resistors is clearly zero. Thus $P_{\text{escape}} = 0$ or, equivalently, $P_{\text{return}} = 1$. Thus a random walk in one dimension is recurrent. As alluded to at the outset of Sect. "Introduction to Current Flows", the conductance between one point and infinity in an infinite resistor lattice in general spatial dimension is somewhat challenging. However, to merely determine the recurrence or transience of a random walk, we only need to know if the return probability is zero or greater than zero. Such a simple question can be answered by a crude physical estimate of the network conductance.

To estimate the conductance from one point to infinity, we replace the discrete lattice by a continuum medium of constant conductance. We then estimate the conductance of the infinite medium by decomposing it into a series of concentric shells of fixed thickness dr. A shell at radius r can be regarded as a parallel array of r^{d-1} volume elements, each of which has a fixed conductance. The conductance of one such shell is proportional to its surface area, and the overall resistance is the sum of these shell resistances. This reasoning gives

$$R \sim \int^{\infty} R_{\text{shell}}(r)\mathrm{d}r \sim \int^{\infty} \frac{\mathrm{d}r}{r^{d-1}} = \begin{cases} \infty & \text{for } d \le 2 \\ (P_{\text{escape}} \sum_j g_{+j})^{-1} & \text{for } d > 2\,. \end{cases} \tag{33}$$

The above estimate gives an easy solution to the recurrence/transience transition of random walks. For $d \le 2$, the conductance to infinity is zero because there are an insufficient number of independent paths from the origin to infinity. Correspondingly, the escape probability is zero and the random walk is recurrent. The case $d = 2$ is more delicate because the integral in Eq. (33) diverges only logarithmically at the upper limit. Nevertheless, the conductance to infinity is still zero and the corresponding random walk is recurrent (but just barely). For $d > 2$, the conductance between a single point and infinity in an infinite homogeneous resistor network is non zero and therefore the escape probability of the corresponding random walk is also non zero – the walk is now transient.

There are many amusing ramifications of the recurrence of random walks and we mention two such properties. First, for $d \le 2$, even though a random walk eventually returns to its starting point, the mean time for this event is infinite! This divergence stems from a power-law tail in the time dependence of the first-passage probability [31,76], namely, the probability that a random walk returns to the origin for the first time. Another striking aspect of recurrence is that because a random walk returns to its starting point with certainty, it necessarily returns an infinite number of times.

Future Directions

There is a good general understanding of the conductance of resistor networks, both far from the percolation threshold, where effective medium theory applies, and close to percolation, where the conductance G vanishes as $(p - p_c)^t$. Many advancements in numerical techniques have been developed to determine the conductance accurately and thereby obtain precise values for the conductance exponent, especially in two dimensions. In spite of this progress, we still do not yet have the right way, if it exists at all, to link the geometry of the percolation cluster or the conducting backbone to the conductivity itself. Furthermore, many exponents of two-dimensional percolation are known exactly. Is it possible that the exact approaches developed to determine percolation exponents can be extended to give the exact conductance exponent?

Finally, there are aspects about conduction in random networks that are worth highlighting. The first falls under the rubric of *directed percolation* [51]. Here each link in a network has an intrinsic directionality that allows current to flow in one direction only – a resistor and diode in series. Links are also globally oriented; on the square lattice for example, current can flow rightward and upward. A qualitative understanding of directed percolation and directed conduction has been achieved that parallels that of isotropic percolation. However, there is one facet of directed conduction that is barely explored. Namely, the state of the network (the bonds that are forward biased) must be determined self consistently from the current flows. This type of non-linearity is much more serious when the circuit elements are randomly oriented. These questions about the coupling between the state of the network and its conductance are central when the circuit elements are intrinsically non-linear [49,78]. This is a topic that seems ripe for new developments.

Bibliography

1. Adler J (1985) Conductance Exponents From the Analysis of Series Expansions for Random Resistor Networks. J Phys A Math Gen 18:307–314
2. Adler J, Meir Y, Aharony A, Harris AB, Klein L (1990) Low-Concentration Series in General Dimension. J Stat Phys 58:511–538
3. Aharony A, Feder J (eds) (1989) Fractals in Physics. Phys D 38:1–398
4. Alava MJ, Nukala PKVV, Zapperi S (2006) Statistical Models for Fracture. Adv Phys 55:349–476
5. Alexander S, Orbach R (1982) Density of States of Fractals: Fractons. J Phys Lett 43:L625–L631
6. Atkinson D, van Steenwijk FJ (1999) Infinite Resistive Lattice. Am J Phys 67:486–492
7. Batrouni GG, Hansen A, Nelkin M (1986) Fourier Acceleration of Relaxation Processes in Disordered Systems. Phys Rev Lett 57:1336–1339
8. Batrouni GG, Hansen A, Larson B (1996) Current Distribution in the Three-Dimensional Random Resistor Network at the Percolation Threshold. Phys Rev E 53:2292–2297
9. Blumenfeld R, Meir Y, Aharony A, Harris AB (1987) Resistance Fluctuations in Randomly Diluted Networks. Phys Rev B 35:3524–3535
10. Bruggeman DAG (1935) Berechnung verschiedener physikalischer Konstanten von heterogenen Substanzen. I. Dielektrizitätskonstanten und Leitfähigkeiten der Mischkörper aus isotropen Substanzen. Ann Phys (Leipzig) 24:636–679. [Engl Trans: Computation of Different Physical Constants of Heterogeneous Substances. I. Dielectric Constants and Conductivenesses of the Mixing Bodies from Isotropic Substances.]
11. Bunde A, Havlin S (eds) (1991) Fractals and Disordered Systems. Springer, Berlin
12. Byshkin MS, Turkin AA (2005) A new method for the calculation of the conductance of inhomogeneous systems. J Phys A Math Gen 38:5057–5067
13. Castellani C, Peliti L (1986) Multifractal Wavefunction at the Localisation Threshold. J Phys A Math Gen 19:L429–L432
14. Chan SK, Machta J, Guyer RA (1989) Large Currents in Random Resistor Networks. Phys Rev B 39:9236–9239
15. Coniglio A (1981) Thermal Phase Transition of the Dilute s-State Potts and n-Vector Models at the Percolation Threshold. Phys Rev Lett 46:250–253
16. Cserti J (2000) Application of the lattice Green's function of calculating the resistance of an infinite network of resistors. Am J Phys 68:896–906
17. de Arcangelis L, Redner S, Coniglio A (1985) Anomalous Voltage Distribution of Random Resistor Networks and a New Model for the Backbone at the Percolation Threshold. Phys Rev B 3:4725–4727
18. de Arcangelis L, Redner S, Herrmann HJ (1985) A Random Fuse Model for Breaking Processes. J Phys 46:L585–L590
19. de Arcangelis L, Redner S, Coniglio A (1986) Multiscaling Approach in Random Resistor and Random Superconducting Networks. Phys Rev B 34:4656–4673
20. de Gennes PG (1972) Exponents for the Excluded vol Problem as Derived by the Wilson Method. Phys Lett A 38:339–340
21. de Gennes PG (1976) La Notion de Percolation: Un Concept Unificateur. La Recherche 7:919–927
22. de Gennes PG (1976) On a Relation Between Percolation Theory and the Elasticity of Gels. J Phys Lett 37:L1–L3
23. den Nijs M (1979) A Relation Between the Temperature Exponents of the Eight-Vertex and q-state Potts Model. J Phys A Math Gen 12:1857–1868
24. Derrida B, Vannimenus J (1982) Transfer–Matrix Approach to Random Resistor Networks. J Phys A: Math Gen 15:L557–L564
25. Derrida B, Zabolitzky JG, Vannimenus J, Stauffer D (1984) A Transfer Matrix Program to Calculate the Conductance of Random Resistor Networks. J Stat Phys 36:31–42
26. Doyle PG, Snell JL (1984) Random Walks and Electric Networks. The Carus Mathematical Monograph, Series 22. The Mathematical Association of America, USA
27. Duxbury PM, Beale PD, Leath PL (1986) Size Effects of Electrical Breakdown in Quenched Random Media. Phys Rev Lett 57:1052–1055
28. Duxbury PM, Leath PL, Beale PD (1987) Breakdown Properties of Quenched Random Systems: The Random-Fuse Network. Phys Rev B 36:367–380
29. Dykhne AM (1970) Conductivity of a Two-Dimensional Two-Phase System. Zh Eksp Teor Fiz 59:110–115 [Engl Transl: (1971) Sov Phys-JETP 32:63–65]
30. Eggarter TP, Cohen MH (1970) Simple Model for Density of States and Mobility of an Electron in a Gas of Hard-Core Scatterers. Phys Rev Lett 25:807–810
31. Feller W (1968) An Introduction to Probability Theory and Its Applications, vol 1. Wiley, New York
32. Fisch R, Harris AB (1978) Critical Behavior of Random Resistor Networks Near the Percolation Threshold. Phys Rev B 18: 416–420
33. Fogelholm R (1980) The Conductance of Large Percolation Network Samples. J Phys C 13:L571–L574
34. Fortuin CM, Kasteleyn PW (1972) On the Random Cluster Model. I. Introduction and Relation to Other Models. Phys 57:536–564
35. Frank DJ, Lobb CJ (1988) Highly Efficient Algorithm for Percolative Transport Studies in Two Dimensions. Phys Rev B 37:302–307

36. Gingold DB, Lobb CJ (1990) Percolative Conduction in Three Dimensions. Phys Rev B 42:8220–8224
37. Golden K (1989) Convexity in Random Resistor Networks. In: Kohn RV, Milton GW (eds) Random Media and Composites. SIAM, Philadelphia, pp 149–170
38. Golden K (1990) Convexity and Exponent Inequalities for Conduction Near Percolation. Phys Rev Lett 65:2923–2926
39. Gumbel EJ (1958) Statistics of Extremes. Columbia University Press, New York
40. Halperin BI, Feng S, Sen PN (1985) Differences Between Lattice and Continuum Percolation Transport Exponents. Phys Rev Lett 54:2391–2394
41. Halsey TC, Jensen MH, Kadanoff LP, Procaccia I, Shraiman BI (1986) Fractal Measures and Their Singularities: The Characterization of Strange Sets. Phys Rev A 33:1141–1151
42. Halsey TC, Meakin P, Procaccia I (1986) Scaling Structure of the Surface Layer of Diffusion-Limited Aggregates. Phys Rev Lett 56:854–857
43. Harary F (1969) Graph Theory. Addison Wesley, Reading, MA
44. Harris AB, Kim S, Lubensky TC (1984) ε Expansion for the Conductance of a Random Resistor Network. Phys Rev Lett 53:743–746
45. Hentschel HGE, Procaccia I (1983) The Infinite Number of Generalized Dimensions of Fractals and Strange Attractors. Phys D 8:435–444
46. Kahng B, Batrouni GG, Redner S (1987) Logarithmic Voltage Anomalies in Random Resistor Networks. J Phys A: Math Gen 20:L827–834
47. Kasteleyn PW, Fortuin CM (1969) Phase Transitions in Lattice Systems with Random Local Properties. J Phys Soc Japan (Suppl) 26:11–14
48. Keller JB (1964) A Theorem on the Conductance of a Composite Medium. J Math Phys 5:548–549
49. Kenkel SW, Straley JP (1982) Percolation Theory of Nonlinear Circuit Elements. Phys Rev Lett 49:767–770
50. Kennelly AE (1899) The Equivalence of Triangles and Three-Pointed Stars in Conducting Networks. Electr World Eng 34:413–414
51. Kinzel W (1983) Directed Percolation in Percolation Structures and Processes. In: Deutscher G, Zallen R, Adler J, Hilger A, Bristol UK, Redner S (eds) Annals of the Israel Physical Society, vol 5, Percolation and Conduction in Random Resistor-Diode Networks, ibid, pp 447–475
52. Kirchhoff G (1847) Über die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Verteilung galvanischer Ströme geführt wird. Ann Phys Chem 72:497–508. [English translation by O'Toole JB, Kirchhoff G (1958) On the solution of the equations obtained from the investigation of the linear distribution of galvanic currents. IRE Trans Circuit Theory CT5:4–8.]
53. Kirkpatrick S (1971) Classical Transport in Disordered Media: Scaling and Effective-Medium Theories. Phys Rev Lett 27:1722–1725
54. Kirkpatrick S (1973) Percolation and Conduction. Rev Mod Phys 45:574–588
55. Koplik J (1981) On the Effective Medium Theory of Random Linear Networks. J Phys C 14:4821–4837
56. Koplik J, Redner S, Wilkinson D (1988) Transport and Dispersion in Random Networks with Percolation Disorder. Phys Rev A 37:2619–2636
57. Landauer R (1952) The Electrical Resistance of Binary Metallic Mixtures. J Appl Phys 23:779–784
58. Last BL, Thouless DJ (1971) Percolation Theory and Electrical Conductance. Phys Rev Lett 27:1719–1721
59. Li PS, Strieder W (1982) Critical Exponents for Conduction in a Honeycomb Random Site Lattice. J Phys C 15:L1235–L1238; Also in: Li PS, Strieder W (1982) Monte Carlo Simulation of the Conductance of the Two-Dimensional Triangular Site Network. J Phys C 15:6591–6595
60. Li YS, Duxbury PM (1987) Size and Location of the Largest Current in a Random Resistor Network. Phys Rev B 36:5411–5419
61. Lobb CJ, Frank DJ (1979) A Large-Cell Renormalisation Group Calculation of the Percolation Conduction Critical Exponent. J Phys C 12:L827–L830
62. Lobb CJ, Frank DJ (1982) Percolative Conduction and the Alexander–Orbach Conjecture in Two Dimensions. Phys Rev B 30:4090–4092
63. Lovasz L (1993) Random Walks on Graphs: A Survey. In: Miklós D, Sós VT, Szönyi T (eds) Combinatorics, Paul Erdös is Eighty, vol 2. János Bolyai Mathematical Society, Budapest 2, pp 1–46
64. Ma SK (1976) Modern Theory of Critical Phenomena. WA Benjamin, Reading, MA
65. Machta J, Guyer RA (1987) Largest Current in a Random Resistor Network. Phys Rev B 36:2142–2146
66. Mandelbrot BB (1974) Intermittent Turbulence in Self-Similar Cascades: Divergence of High Moments and Dimension of the Carrier. J Fluid Mech 62:331–358
67. Mandelbrot BB (1982) The Fractal Geometry of Nature. WH Freeman, San Francisco
68. Mitescu CD, Allain A, Guyon E, Clerc J (1982) Electrical Conductance of Finite-Size Percolation Networks. J Phys A: Math Gen 15:2523–2532
69. Nevard J, Keller JB (1985) Reciprocal Relations for Effective Conductivities of Anisotropic Media. J Math Phys 26: 2761–2765
70. Normand JM, Herrmann HJ, Hajjar M (1988) Precise Calculation of the Dynamical Exponent of Two-Dimensional Percolation. J Stat Phys 52:441–446
71. Press W, Teukolsky S, Vetterling W, Flannery B (1992) Numerical Recipes in Fortran 90. The Art of Parallel Scientific Computing. Cambridge University Press, New York
72. Rammal R, Tannous C, Breton P, Tremblay A-MS (1985) Flicker (1/f) Noise in Percolation Networks: A New Hierarchy of Exponents Phys Rev Lett 54:1718–1721
73. Rammal R, Tannous C, Tremblay A-MS (1985) 1/f Noise in Random Resistor Networks: Fractals and Percolating Systems. Phys Rev A 31:2662–2671
74. Rayleigh JW (1892) On the Influence of Obstacles Arranged in Rectangular Order upon the Properties of a Medium. Philos Mag 34:481–502
75. Redner S (1990) Random Multiplicative Processes: An Elementary Tutorial. Am J Phys 58:267–272
76. Redner S (2001) A Guide to First-Passage Processes. Cambridge University Press, New York
77. Redner S, Brooks JS (1982) Analog Experiments and Computer Simulations for Directed Conductance. J Phys A Math Gen 15:L605–L610
78. Roux S, Herrmann HJ (1987) Disorder-Induced Nonlinear Conductivity. Europhys Lett 4:1227–1231
79. Sahimi M, Hughes BD, Scriven LE, Davis HT (1983) Critical

Exponent of Percolation Conductance by Finite-Size Scaling. J Phys C 16:L521–L527
80. Sarychev AK, Vinogradoff AP (1981) Drop Model of Infinite Cluster for 2d Percolation. J Phys C 14:L487–L490
81. Senturia SB, Wedlock BD (1975) Electronic Circuits and Applications. Wiley, New York, pp 75
82. Skal AS, Shklovskii BI (1975) Topology of the Infinite Cluster of The Percolation Theory and its Relationship to the Theory of Hopping Conduction. Fiz Tekh Poluprov 8:1586–1589 [Engl. transl.: Sov Phys-Semicond 8:1029–1032]
83. Stanley HE (1971) Introduction to Phase Transition and Critical Phenomena. Oxford University Press, Oxford, UK
84. Stanley HE (1977) Cluster Shapes at the Percolation Threshold: An Effective Cluster Dimensionality and its Connection with Critical-Point Exponents. J Phys A Math Gen 10: L211–L220
85. Stauffer D, Aharony A (1994) Introduction to Percolation Theory, 2nd edn. Taylor & Francis, London, Bristol, PA
86. Stenull O, Janssen HK, Oerding K (1999) Critical Exponents for Diluted Resistor Networks. Phys Rev E 59:4919–4930
87. Stephen M (1978) Mean-Field Theory and Critical Exponents for a Random Resistor Network. Phys Rev B 17:4444–4453
88. Stephen MJ (1976) Percolation problems and the Potts model. Phys Lett A 56:149–150
89. Stinchcombe RB, Watson BP (1976) Renormalization Group Approach for Percolation Conductance. J Phys C 9:3221–3247
90. Straley JP (1977) Critical Exponents for the Conductance of Random Resistor Lattices. Phys Rev B 15:5733–5737
91. Straley JP (1982) Random Resistor Tree in an Applied Field. J Phys C 10:3009–3014
92. Straley JP (1982) Threshold Behaviour of Random Resistor Networks: A Synthesis of Theoretical Approaches. J Phys C 10:2333–2341
93. Trugman SA, Weinrib A (1985) Percolation with a Threshold at Zero: A New Universality Class. Phys Rev B 31:2974–2980
94. van der Pol B, Bremmer H (1955) Operational Calculus Based on the Two-Sided Laplace Integral. Cambridge University Press, Cambridge, UK
95. Venezian G (1994) On the resistance between two points on a grid. Am J Phys 62:1000–1004
96. Watson GN (1939) Three Triple Integrals. Oxford Ser 2. Quart J Math 10:266–276
97. Webman I, Jortner J, Cohen MH (1975) Numerical Simulation of Electrical Conductance in Microscopically Inhomogeneous Materials. Phys Rev B 11:2885–2892
98. Weiss GH (1994) Aspects and Applications of the Random Walk. Elsevier Science Publishing Co, New York
99. Wu FY (1982) The Potts Model. Rev Mod Phys 54:235–268
100. Zabolitsky JG (1982) Monte Carlo Evidence Against the Alexander–Orbach Conjecture for Percolation Conductance. Phys Rev B 30:4077–4079

Fractal and Multifractal Time Series

Jan W. Kantelhardt
Institute of Physics,
Martin-Luther-University Halle-Wittenberg,
Halle, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Fractal and Multifractal Time Series
Methods for Stationary Fractal Time Series Analysis
Methods for Non-stationary Fractal Time Series Analysis
Methods for Multifractal Time Series Analysis
Statistics of Extreme Events in Fractal Time Series
Simple Models for Fractal and Multifractal Time Series
Future Directions
Acknowledgment
Bibliography

Glossary

Time series One dimensional array of numbers (x_i), $i = 1, \ldots, N$, representing values of an observable x usually measured equidistant (or nearly equidistant) in time.

Complex system A system consisting of many non-linearly interacting components. It cannot be split into simpler sub-systems without tampering with the dynamical properties.

Scaling law A power law with a scaling exponent (e. g. α) describing the behavior of a quantity F (e. g., fluctuation, spectral power) as function of a scale parameter s (e. g., time scale, frequency) at least asymptotically: $F(s) \sim s^{\alpha}$. The power law should be valid for a large range of s values, e. g., at least for one order of magnitude.

Fractal system A system characterized by a scaling law with a fractal, i. e., non-integer exponent. Fractal systems are self-similar, i. e., a magnification of a small part is statistically equivalent to the whole.

Self-affine system Generalization of a fractal system, where different magnifications s and $s' = s^H$ have to be used for different directions in order to obtain a statistically equivalent magnification. The exponent H is called Hurst exponent. Self-affine time series and time series becoming self-affine upon integration are commonly denoted as fractal using a less strict terminology.

Multifractal system A system characterized by scaling laws with an infinite number of different fractal exponents. The scaling laws must be valid for the same range of the scale parameter.

Crossover Change point in a scaling law, where one scaling exponent applies for small scale parameters and another scaling exponent applies for large scale pa-

rameters. The center of the crossover is denoted by its characteristic scale parameter $s_\times$ in this article.

Persistence In a persistent time series, a large value is usually (i. e., with high statistical preference) followed by a large value and a small value is followed by a small value. A fractal scaling law holds at least for a limited range of scales.

Short-term correlations Correlations that decay sufficiently fast that they can be described by a characteristic correlation time scale; e. g., exponentially decaying correlations. A crossover to uncorrelated behavior is observed on larger scales.

Long-term correlations Correlations that decay sufficiently slow that a characteristic correlation time scale cannot be defined; e. g., power-law correlations with an exponent between 0 and 1. Power-law scaling is observed on large time scales and asymptotically. The term long-range correlations should be used if the data is not a time series.

Non-stationarities If the mean or the standard deviation of the data values change with time, the weak definition of stationarity is violated. The strong definition of stationarity requires that all moments remain constant, i. e., the distribution density of the values does not change with time. Non-stationarities like monotonous, periodic, or step-like trends are often caused by external effects. In a more general sense, changes in the dynamics of the system also represent non-stationarities.

Definition of the Subject

Data series generated by complex systems exhibit fluctuations on a wide range of time scales and/or broad distributions of the values. In both equilibrium and non-equilibrium situations, the natural fluctuations are often found to follow a scaling relation over several orders of magnitude. Such scaling laws allow for a characterization of the data and the generating complex system by fractal (or multifractal) scaling exponents, which can serve as characteristic fingerprints of the systems in comparisons with other systems and with models. Fractal scaling behavior has been observed, e. g., in many data series from experimental physics, geophysics, medicine, physiology, and even social sciences. Although the underlying causes of the observed fractal scaling are often not known in detail, the fractal or multifractal characterization can be used for generating surrogate (test) data, modeling the time series, and deriving predictions regarding extreme events or future behavior. The main application, however, is still the characterization of different states or phases of the complex system based on the observed scaling behavior. For example, the health status and different physiological states of the human cardiovascular system are represented by the fractal scaling behavior of the time series of intervals between successive heartbeats, and the coarsening dynamics in metal alloys are represented by the fractal scaling of the time-dependent speckle intensities observed in coherent X-ray spectroscopy.

In order to observe fractal and multifractal scaling behavior in time series, several tools have been developed. Besides older techniques assuming stationary data, there are more recently established methods differentiating truly fractal dynamics from fake scaling behavior caused by non-stationarities in the data. In addition, short-term and long-term correlations have to be clearly distinguished to show fractal scaling behavior unambiguously. This article describes several methods originating from statistical physics and applied mathematics, which have been used for fractal and multifractal time series analysis in stationary and non-stationary data.

Introduction

The characterization and understanding of complex systems is a difficult task, since they cannot be split into simpler subsystems without tampering with the dynamical properties. One approach in studying such systems is the recording of long *time series* of several selected variables (observables), which reflect the state of the system in a dimensionally reduced representation. Some systems are characterized by periodic or nearly periodic behavior, which might be caused by oscillatory components or closed-loop regulation chains. However, in truly complex systems such periodic components are usually not limited to one or two characteristic frequencies or frequency bands. They rather extend over a wide spectrum, and fluctuations on many time scales as well as broad distributions of the values are found. Often no specific lower frequency limit – or, equivalently, upper characteristic time scale – can be observed. In these cases, the dynamics can be characterized by *scaling laws* which are valid over a wide (possibly even unlimited) range of time scales or frequencies; at least over orders of magnitude. Such dynamics are usually denoted as *fractal* or *multifractal*, depending on the question if they are characterized by one scaling exponent or by a multitude of scaling exponents.

The first scientist who applied fractal analysis to natural time series was Benoit B. Mandelbrot [1,2,3], who included early approaches by H.E. Hurst regarding hydrological systems [4,5]. For extensive introductions describing fractal scaling in complex systems, we refer to [6,7,8,

9,10,11,12,13]. In the last decade, fractal and multifractal scaling behavior has been reported in many natural time series generated by complex systems, including

- Geophysics time series (recordings of temperature, precipitation, water runoff, ozone levels, wind speed, seismic events, vegetational patterns, and climate dynamics),
- Medical and physiological time series (recordings of heartbeat, respiration, blood pressure, blood flow, nerve spike intervals, human gait, glucose levels, and gene expression data),
- DNA sequences (they are not actually *time* series),
- Astrophysical time series (X-ray light sources and sunspot numbers),
- Technical time series (internet traffic, highway traffic, and neutronic power from a reactor),
- Social time series (finance and economy, language characteristics, fatalities in conflicts), as well as
- Physics data (also going beyond *time* series), e. g., surface roughness, chaotic spectra of atoms, and photon correlation spectroscopy recordings.

If one finds that a complex system is characterized by fractal (or multifractal) dynamics with particular scaling exponents, this finding will help in obtaining predictions on the future behavior of the system and on its reaction to external perturbations or changes in the boundary conditions. Phase transitions in the regulation behavior of a complex system are often associated with changes in their fractal dynamics, allowing for a detection of such transitions (or the corresponding states) by fractal analysis. One example for a successful application of this approach is the human cardiovascular system, where the fractality of heartbeat interval time series was shown to reflect certain cardiac impairments as well as sleep stages [14,15]. In addition, one can test and iteratively improve models of the system until they reproduce the observed scaling behavior. One example for such an approach is climate modeling, where the models were shown to need input from volcanos and solar radiation in order to reproduce the long-term correlated (fractal) scaling behavior [16] previously found in observational temperature data [17].

Fractal (or multifractal) scaling behavior certainly cannot be assumed a priori, but has to be established. Hence, there is a need for refined analysis techniques, which help to differentiate truly fractal dynamics from fake scaling behavior caused, e. g., by non-stationarities in the data. If conventional statistical methods are applied for the analysis of time series representing the dynamics of a complex system [18,19], there are two major problems. (i) The number of data series and their durations (lengths) are usually very limited, making it difficult to extract significant information on the dynamics of the system in a reliable way. (ii) If the length of the data is extended using computer-based recording techniques or historical (proxy) data, non-stationarities in the signals tend to be superimposed upon the intrinsic fluctuation properties and measurement noise. Non-stationarities are caused by external or internal effects that lead to either continuous or sudden changes in the average values, standard deviations or regulation mechanism. They are a major problem for the characterization of the dynamics, in particular for finding the scaling properties of given data.

Fractal and Multifractal Time Series

Fractality, Self-Affinity, and Scaling

The topic of this article is the fractality (and/or multifractality) of time series. Since fractals and multifractals in general are discussed in many other articles of the encyclopedia, the concept is not thoroughly explained here. In particular, we refer to the articles ► Fractal Geometry, A Brief Introduction to and ► Fractals and Multifractals, Introduction to for the formalism describing fractal and multifractal structures, respectively.

In a strict sense, most time series are one dimensional, since the values of the considered observable are measured in homogeneous time intervals. Hence, unless there are missing values, the fractal dimension of the support is $D(0) = 1$. However, there are rare cases where most of the values of a time series are very small or even zero, causing a dimension $D(0) < 1$ of the support. In these cases, one has to be very careful in selecting appropriate analysis techniques, since many of the methods presented in this article are not accurate for such data; the wavelet transform modulus maxima technique (see Subsect. "Wavelet Transform Modulus Maxima (WTMM) Method") is the most advanced applicable method.

Even if the fractal dimension of support is one, the information dimension $D(1)$ and the correlation dimension $D(2)$ can be studied. As we will see in Subsect. "The Structure Function Approach and Singularity Spectra", $D(2)$ is in fact explicitly related to all exponents studied in monofractal time series analysis. However, usually a slightly different approach is employed based on the notion of self-affinity instead of (multi-) fractality. Here, one takes into account that the time axis and the axis of the measured values $x(t)$ are not equivalent. Hence, a rescaling of time t by a factor a may require rescaling of the series values $x(t)$ by a different factor a^H in order to obtain a statistically similar (i. e., self-similar) picture. In this case the

scaling relation

$$x(t) \to a^H x(at) \tag{1}$$

holds for an arbitrary factor a, describing the data as self-affine (see, e. g., [6]). The *Hurst* exponent H (after the water engineer H.E. Hurst [4]) characterizes the type of self affinity. Figure 1a shows several examples of self-affine time series with different H. The trace of a random walk (Brownian motion, third line in Fig. 1a), for example, is characterized by $H = 0.5$, implying that the position axis must be rescaled by a factor of two if the time axis is rescaled by a factor of four. Note that self-affine series are often denoted as fractal even though they are not fractal in the strict sense. In this article the term "fractal" will be used in the more general sense including all data, where a Hurst exponent H can be reasonably defined.

The scaling behavior of self-affine data can also be characterized by looking at their mean-square displacement. Since the mean-square displacement of a random walker is known to increase linearly in time, $\langle x^2(t)\rangle \sim t$, deviations from this law will indicate the presence of self-affine scaling. As we will see in Subsect. "Fluctuation Analysis (FA)", one can thus retrieve the Hurst (or self-affinity) exponent H by studying the scaling behavior of the mean-square displacement, or the mean-square fluctuations $\langle x^2(t)\rangle \sim t^{2H}$.

Persistence, Long- and Short-Term Correlations

Self-affine data are persistent in the sense that a large value is usually (i. e., with high statistical preference) followed by a large value and a small value is followed by a small value. For the trace of a random walk, persistence on all time scales is trivial, since a later position is just a former one plus some random increment(s). The persistence holds for all time scales, where the self-affinity relation (1) holds. However, the degree of persistence can also vary on different time scales. Weather is a typical example: while the weather tomorrow or in one week is probably similar to the weather today (due to a stable general weather condition), persistence is much harder to be seen on longer time scales.

Considering the increments $\Delta x_i = x_i - x_{i-1}$ of a self-affine series, (x_i), $i = 1, \ldots, N$ with N values measured equidistant in time, one finds that the Δx_i can be either persistent, independent, or anti-persistent. Examples for all cases are shown in Fig. 1b. In our example of the random walk with $H = 0.5$ (third line in the figure), the increments (steps) are fully independent of each other. Persistent and anti-persistent increments, where a positive increment is likely to be followed by another positive or negative increment, respectively, are also leading to persistent integrated series $x_i = \sum_{j=1}^{i} \Delta x_j$.

For stationary data with constant mean and standard deviation the auto-covariance function of the increments,

$$C(s) = \left\langle \Delta x_i \Delta x_{i+s} \right\rangle = \frac{1}{N-s} \sum_{i=1}^{N-s} \Delta x_i \Delta x_{i+s} \tag{2}$$

can be studied to determine the degree of persistence. If $C(s)$ is divided by the variance $\langle (\Delta x_i)^2 \rangle$, it becomes the auto-correlation function; both are identical if the data are

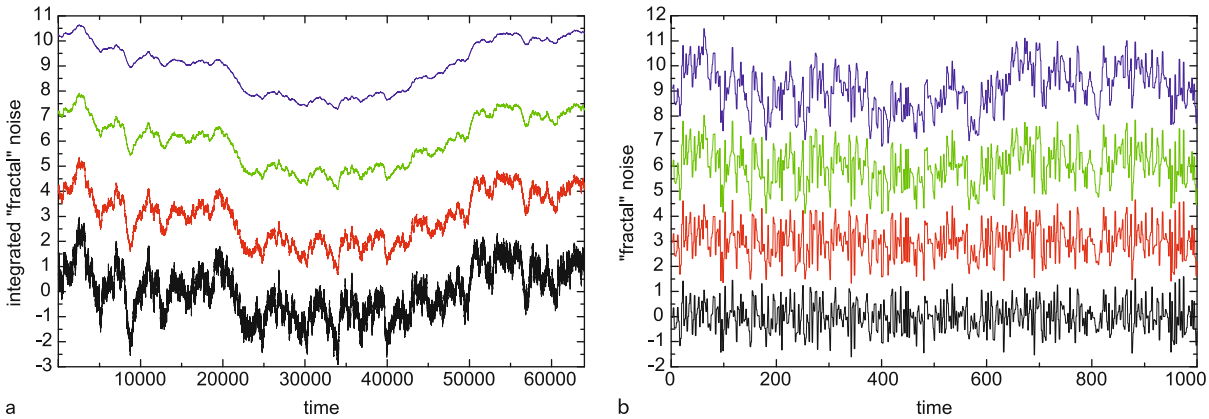

Fractal and Multifractal Time Series, Figure 1

a Examples of self-affine series x_i characterized by different Hurst exponents H = 0.9, 0.7, 0.5, 0.3 (from *top* to *bottom*). The data has been generated by Fourier filtering using the same seed for the random number generator. b Differentiated series Δx_i of the data from a; the Δx_i are characterized by positive long-term correlations (persistence) with γ = 0.2 and 0.6 (*first* and *second line*), uncorrelated behavior (*third line*), and anti-correlations (*bottom line*), respectively

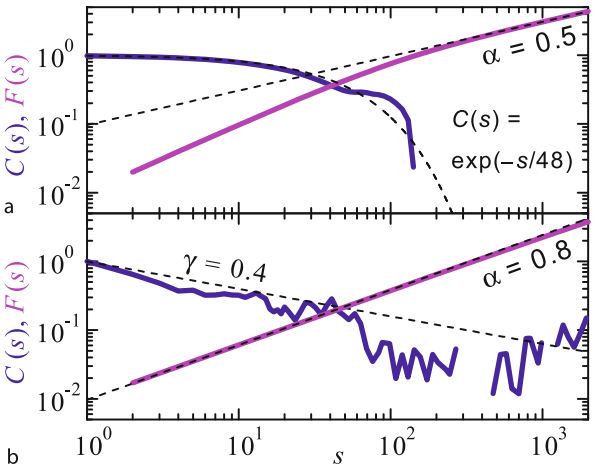

Fractal and Multifractal Time Series, Figure 2
Comparison of the autocorrelation functions $C(s)$ (decreasing functions) and fluctuation functions $F_2(s)$ (increasing functions) for short-term correlated data (*top panel*) and long-term correlated data ($\gamma = 0.4$, *bottom panel*). The asymptotic slope $H \approx \alpha = 0.5$ of $F_2(s)$ clearly indicates missing long-term correlations, while $H \approx \alpha = 1 - \gamma/2 > 0.5$ indicates long-term correlations. The difference is much harder to observe in $C(s)$, where statistical fluctuations and negative values start occurring above $s \approx 100$. The data have been generated by an AR process Eq. (4) and Fourier filtering, respectively. The *dashed lines* indicate the theoretical curves

normalized with unit variance. If the Δx_i are uncorrelated (as for the random walk), $C(s)$ is zero for $s > 0$. Short-range correlations of the increments Δx_i are usually described by $C(s)$ declining exponentially,

$$C(s) \sim \exp(-s/t_\times) \tag{3}$$

with a characteristic decay time $t_\times$. Such behavior is typical for increments generated by an auto-regressive (AR) process

$$\Delta x_i = c\Delta x_{i-1} + \varepsilon_i \tag{4}$$

with random uncorrelated offsets ε_i and $c = \exp(-1/t_\times)$. Figure 2a shows the auto-correlation function for one configuration of an AR process with $t_\times = 48$.

For so-called *long-range correlations* $\int_0^\infty C(s)\mathrm{d}s$ diverges in the limit of infinitely long series ($N \to \infty$). In practice, this means that $t_\times$ cannot be defined because it increases with increasing N. For example, $C(s)$ declines as a power-law

$$C(s) \propto s^{-\gamma} \tag{5}$$

with an exponent $0 < \gamma < 1$. Figure 2b shows $C(s)$ for one configuration with $\gamma = 0.4$. This type of behavior can be modeled by the Fourier filtering technique (see Subsect. "Fourier Filtering"). Long-term correlated, i. e. persistent, behavior of the Δx_i leads to self-affine scaling behavior of the x_i, characterized by $H = 1 - \gamma/2$, as will be shown below.

Crossovers and Non-Stationarities in Time Series

Short-term correlated increments Δx_i characterized by a finite characteristic correlation decay time $t_\times$ lead to a crossover in the scaling behavior of the integrated series $x_i = \sum_{j=1}^{i} \Delta x_j$, see Fig. 2a for an example. Since the position of the crossover might be numerically different from $t_\times$, we denote it by $s_\times$ here. Time series with a crossover are not self-affine and there is no unique Hurst exponent H characterizing them. While $H > 0.5$ is observed on small time scales (indicating correlations in the increments), the asymptotic behavior (for large time scales $s \gg t_\times$ and $\gg s_\times$) is always characterized by $H = 0.5$, since all correlations have decayed. Many natural recordings are characterized by pronounced short-term correlations in addition to scaling long-term correlations. For example, there are short-term correlations due to particular general weather situations in temperature data and due to respirational effects in heartbeat data. Crossovers in the scaling behavior of complex time series can also be caused by different regulation mechanisms on fast and slow time scales. Fluctuations of river runoff, for example, show different scaling behavior on time scales below and above approximately one year.

Non-stationarities can also cause crossovers in the scaling behavior of data if they are not properly taken into account. In the most strict sense, non-stationarities are variations in the mean or the standard deviation of the data (violating weak stationarity) or the distribution of the data values (violating strong stationarity). Non-stationarities like monotonous, periodic or step-like trends are often caused by external effects, e. g., by the greenhouse warming and seasonal variations for temperature records, different levels of activity in long-term physiological data, or unstable light sources in photon correlation spectroscopy. Another example for non-stationary data is a record consisting of segments with strong fluctuations alternating with segments with weak fluctuations. Such behavior will cause a crossover in scaling at the time scale corresponding to the typical duration of the homogeneous segments. Different mechanisms of regulation during different time segments – like, e. g., different heartbeat regulation during different sleep stages at night – can also cause crossovers; they are regarded as non-stationarities here, too. Hence, if crossovers in the scaling behavior of

data are observed, more detailed studies are needed to find out the cause of the crossovers. One can try to obtain homogenous data by splitting the original series and employing methods that are at least insensitive to monotonous (polynomially shaped) trends.

To characterize a complex system based on time series, trends and fluctuations are usually studied separately (see, e. g., [20] for a discussion). Strong trends in data can lead to a false detection of long-range statistical persistence if only one (non-detrending) method is used or if the results are not carefully interpreted. Using several advanced techniques of scaling time series analysis (as described in Sect. "Methods for Non-stationary Fractal Time Series Analysis") crossovers due to trends can be distinguished from crossovers due to different regulation mechanisms on fast and slow time scales. The techniques can thus assist in gaining insight into the scaling behavior of the natural variability as well as into the kind of trends of the considered time series.

It has to be stressed that crossovers in scaling behavior must not be confused with multifractality. Even though several scaling exponents are needed, they are not applicable for the same regime (i. e., the same range of time scales). Real multifractality, on the other hand, is characterized by different scaling behavior of different moments over the full range of time scales (see next section).

Multifractal Time Series

Many records do not exhibit a simple monofractal scaling behavior, which can be accounted for by a single scaling exponent. As discussed in the previous section, there might exist crossover (time-) scales $s_\times$ separating regimes with different scaling exponents. In other cases, the scaling behavior is more complicated, and different scaling exponents are required for different parts of the series. In even more complicated cases, such different scaling behavior can be observed for many interwoven fractal subsets of the time series. In this case a multitude of scaling exponents is required for a full description of the scaling behavior in the same range of time scales, and a multifractal analysis must be applied.

Two general types of multifractality in time series can be distinguished: (i) Multifractality due to a broad probability distribution (density function) for the values of the time series, e. g. a Levy distribution. In this case the multifractality cannot be removed by shuffling the series. (ii) Multifractality due to different long-term correlations of the small and large fluctuations. In this case the probability density function of the values can be a regular distribution with finite moments, e. g., a Gaussian distribution. The corresponding shuffled series will exhibit non-multifractal scaling, since all long-range correlations are destroyed by the shuffling procedure. Randomly shuffling the order of the values in the time series is the easiest way of generating surrogate data; however, there are more advanced alternatives (see Sect. "Simple Models for Fractal and Multifractal Time Series"). If both kinds of multifractality are present, the shuffled series will show weaker multifractality than the original series.

A multifractal analysis of time series will also reveal higher order correlations. Multifractal scaling can be observed if, e. g., three or four-point correlations scale differently from the standard two-point correlations studied by classical autocorrelation analysis (Eq. (2)). In addition, multifractal scaling is observed if the scaling behavior of small and large fluctuations is different. For example, extreme events might be more or less correlated than typical events.

Methods for Stationary Fractal Time Series Analysis

In this section we describe four traditional approaches for the fractal analysis of stationary time series, see [21,22,23] for comparative studies. The main focus is on the determination of the scaling exponents H or γ, defined in Eqs. (1) and (5), respectively, and linked by $H = 1 - \gamma/2$ in long-term persistent data. Methods taking non-stationarities into account will be discussed in the next chapter.

Autocorrelation Function Analysis

We consider a record (x_i) of $i = 1, \ldots, N$ equidistant measurements. In most applications, the index i will correspond to the time of the measurements. We are interested in the correlation of the values x_i and x_{i+s} for different time lags, i. e. correlations over different time scales s. In order to remove a constant offset in the data, the mean $\langle x \rangle = \frac{1}{N}\sum_{i=1}^{N} x_i$ is usually subtracted, $\tilde{x}_i \equiv x_i - \langle x \rangle$. Alternatively, the correlation properties of increments $\tilde{x}_i = \Delta x_i = x_i - x_{i-1}$ of the original series can be studied (see also Subsect. "Persistence, Long- and Short-Term Correlations"). Quantitatively, correlations between $\tilde{x}$-values separated by s steps are defined by the (auto-) covariance function $C(s) = \langle \tilde{x}_i \tilde{x}_{i+s} \rangle$ or the (auto-) correlation function $C(s)/\langle \tilde{x}_i^2 \rangle$, see also Eq. (2).

As already mentioned in Subsect. "Persistence, Long- and Short-Term Correlations", the $\tilde{x}_i$ are short-term correlated if $C(s)$ declines exponentially, $C(s) \sim \exp(-s/t_\times)$, and long-term correlated if $C(s)$ declines as a power-law $C(s) \propto s^{-\gamma}$ with a correlation exponent $0 < \gamma < 1$ (see Eqs. (3) and (5), respectively). As illustrated by the two examples shown in Fig. 2, a direct calculation of $C(s)$ is

usually not appropriate due to noise superimposed on the data $\tilde{x}_i$ and due to underlying non-stationarities of unknown origin. Non-stationarities make the definition of $C(s)$ problematic, because the average $\langle x \rangle$ is not well-defined. Furthermore, $C(s)$ strongly fluctuates around zero on large scales s (see Fig. 2b), making it impossible to find the correct correlation exponent γ. Thus, one has to determine the value of γ indirectly.

Spectral Analysis

If the time series is stationary, we can apply standard spectral analysis techniques (Fourier transform) and calculate the power spectrum $S(f)$ of the time series $(\tilde{x}_i)$ as a function of the frequency f to determine self-affine scaling behavior [24]. For long-term correlated data characterized by the correlation exponent γ, we have

$$S(f) \sim f^{-\beta} \quad \text{with } \beta = 1 - \gamma \,. \tag{6}$$

The spectral exponent β and the correlation exponent γ can thus be obtained by fitting a power-law to a double logarithmic plot of the power spectrum $S(f)$. An example is shown in Fig. 3. The relation (6) can be derived from the Wiener–Khinchin theorem (see, e. g., [25]). If, instead of $\tilde{x}_i = \Delta x_i$ the integrated runoff time series is Fourier transformed, i. e., $\tilde{x}_i = x_i \;\; X = \sum_{j=1}^{i} \Delta x_j$, the resulting power spectrum scales as $S(f) \sim f^{-2-\beta}$.

Spectral analysis, however, does not yield more reliable results than auto-correlation analysis unless a logarithmic binning procedure is applied to the double logarithmic plot of $S(f)$ [21], see also Fig. 3. I. e., the average of $\log S(f)$ is calculated in successive, logarithmically wide bands from $a^n f_0$ to $a^{n+1} f_0$, where f_0 is the minimum frequency, $a > 1$ is a factor (e. g., $a = 1.1$), and the index n is counting the bins. Spectral analysis also requires stationarity of the data.

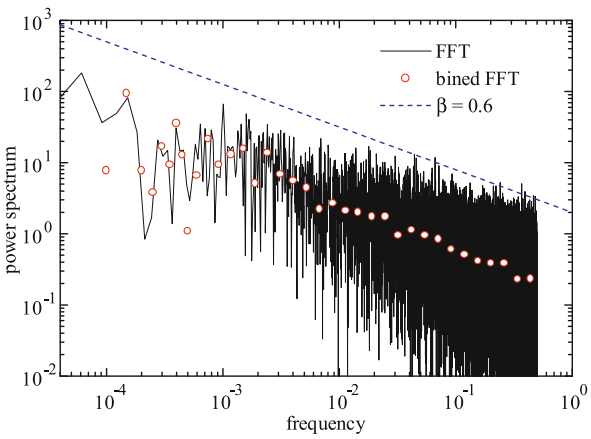

Fractal and Multifractal Time Series, Figure 3
Spectral analysis of a fractal time series characterized by long-term correlations with $\gamma = 0.4$ ($\beta = 0.6$). The expected scaling behavior (*dashed line* indicating the slope $-\beta$) is observed only after binning of the spectrum (*circles*). The data has been generated by Fourier filtering

Hurst's Rescaled-Range Analysis

The first method for the analysis of long-term persistence in time series based on random walk theory has been proposed by the water construction engineer Harold Edwin Hurst (1880–1978), who developed it while working in Egypt. His so-called rescaled range analysis (R/S analysis) [1,2,4,5,6] begins with splitting of the time series $(\tilde{x}_i)$ into non-overlapping segments ν of size (time scale) s (first step), yielding $N_s = \text{int}(N/s)$ segments altogether. In the second step, the *profile* (integrated data) is calculated in each segment $\nu = 0, \ldots, N_s - 1$,

$$\begin{aligned} Y_\nu(j) &= \sum_{i=1}^{j} (\tilde{x}_{\nu s+i} - \langle \tilde{x}_{\nu s+i} \rangle_s) \\ &= \sum_{i=1}^{j} \tilde{x}_{\nu s+i} - \frac{j}{s} \sum_{i=1}^{s} \tilde{x}_{\nu s+i} \,. \end{aligned} \tag{7}$$

By the subtraction of the local averages, piecewise constant trends in the data are eliminated. In the third step, the differences between minimum and maximum value (*ranges*) $R_\nu(s)$ and the standard deviations $S_\nu(s)$ in each segment are calculated,

$$\begin{aligned} R_\nu(s) &= \max_{j=1}^{s} Y_\nu(j) - \min_{j=1}^{s} Y_\nu(j), \\ S_\nu(s) &= \sqrt{\frac{1}{s} \sum_{j=1}^{s} Y_\nu^2(j)} \,. \end{aligned} \tag{8}$$

Finally, the rescaled range is averaged over all segments to obtain the fluctuation function $F(s)$,

$$F_{RS}(s) = \frac{1}{N_s} \sum_{\nu=0}^{N_s-1} \frac{R_\nu(s)}{S_\nu(s)} \sim s^H \quad \text{for } s \gg 1 \,, \tag{9}$$

where H is the Hurst exponent already introduced in Eq. (1). One can show [1,24] that H is related to β and γ by $2H \approx 1 + \beta = 2 - \gamma$ (see also Eqs. (6) and (14)). Note that $0 < \gamma < 1$, so that the right part of the equation does not hold unless $0.5 < H < 1$. The relationship does *not* hold in general for multifractal data. Note also that H actually characterizes the self-affinity of the profile function (7), while β and γ refer to the original data.

The values of H, that can be obtained by Hurst's rescaled range analysis, are limited to $0 < H < 2$, and significant inaccuracies are to be expected close to the bounds. Since H can be increased or decreased by one if the data is integrated ($\tilde{x}_j \to \sum_{i=1}^{j} \tilde{x}_i$) or differentiated ($\tilde{x}_i \to \tilde{x}_i - \tilde{x}_{i-1}$), respectively, one can always find a way to calculate H by rescaled range analysis provided the data is stationary. While values $H < 1/2$ indicate long-term anti-correlated behavior of the data $\tilde{x}_i$, $H > 1/2$ indicates long-term positively correlated behavior. For power-law correlations decaying faster than $1/s$, we have $H = 1/2$ for large s values, like for uncorrelated data.

Compared with spectral analysis, Hurst's rescaled range analysis yields smoother curves with less effort (no binning procedure is necessary) and works also for data with piecewise constant trends.

Fluctuation Analysis (FA)

The standard fluctuation analysis (FA) [8,26] is also based on random walk theory. For a time series $(\tilde{x}_i)$, $i = 1, \ldots, N$, with zero mean, we consider the global profile, i. e., the cumulative sum (cf. Eq. (7))

$$Y(j) = \sum_{i=1}^{j} \tilde{x}_i \,, \quad j = 0, 1, 2, \ldots, N \,, \tag{10}$$

and study how the fluctuations of the profile, in a given time window of size s, increase with s. The procedure is illustrated in Fig. 4 for two values of s. We can consider the profile $Y(j)$ as the position of a random walker on a linear chain after j steps. The random walker starts at the origin and performs, in the ith step, a jump of length $\tilde{x}_i$ to the bottom, if $\tilde{x}_i$ is positive, and to the top, if $\tilde{x}_i$ is negative.

To find how the square-fluctuations of the profile scale with s, we first divide the record of N elements into $N_s = \mathrm{int}(N/s)$ non-overlapping segments of size s starting from the beginning (see Fig. 4) and another N_s non-overlapping segments of size s starting from the end of the considered series. This way neither data at the end nor at the beginning of the record is neglected. Then we determine the fluctuations in each segment ν.

In the standard FA, we obtain the fluctuations just from the values of the profile at both endpoints of each segment $\nu = 1, \ldots, N_s$,

$$F^2_{\mathrm{FA}}(\nu, s) = [Y(\nu s) - Y((\nu 1)s)]^2 \,, \tag{11}$$

(see Fig. 4) and analogous for $\nu = N_{s+1}, \ldots, 2N_s$,

$$F^2_{\mathrm{FA}}(\nu, s) = [Y(N-(\nu-N_s)s) - Y(N-(\nu-1-N_s)s)]^2 . \tag{12}$$

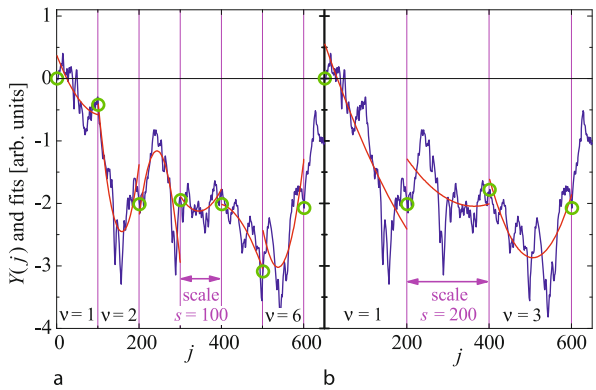

Fractal and Multifractal Time Series, Figure 4
Illustration of the fluctuation analysis (FA) and the detrended fluctuation analysis (DFA). For two segment durations (time scales) $s = 100$ (a) and 200 (b), the profiles $Y(j)$ (*blue lines*; defined in Eq. (11), the values used for fluctuation analysis in Eq. (12) (*green circles*), and least-square quadratic fits to the profiles (*red lines*) are shown

Then we average $F^2_{\mathrm{FA}}(\nu, s)$ over all subsequences to obtain the mean fluctuation $F_2(s)$,

$$F_2(s) = \left[\frac{1}{2N_s} \sum_{\nu=1}^{2N_s} F^2_{\mathrm{FA}}(\nu, s) \right]^{1/2} \sim s^{\alpha} \,. \tag{13}$$

By definition, $F_2(s)$ can be viewed as the root-mean-square displacement of the random walker on the chain, after s steps (the reason for the index 2 will become clear later). For uncorrelated x_i values, we obtain Fick's diffusion law $F_2(s) \sim s^{1/2}$. For the relevant case of long-term correlations, in which $C(s)$ follows the power-law behavior of Eq. (5), $F_2(s)$ increases by a power law,

$$F_2(s) \sim s^{\alpha} \quad \text{with } \alpha \approx H \,, \tag{14}$$

where the fluctuation exponent α is identical with the Hurst exponent H for mono-fractal data and related to γ and β by

$$2\alpha = 1 + \beta = 2 - \gamma \,. \tag{15}$$

The typical behavior of $F_2(s)$ for short-term correlated and long-term correlated data is illustrated in Fig. 2. The relation (15) can be derived straightforwardly by inserting Eqs. (10), (2), and (5) into Eq. (11) and separating sums over products $\tilde{x}_i \tilde{x}_j$ with identical and different i and j, respectively.

The range of the α values that can be studied by standard FA is limited to $0 < \alpha < 1$, again with significant inaccuracies close to the bounds. Regarding integration or

differentiation of the data, the same rules apply as listed for H in the previous subsection. The results of FA become statistically unreliable for scales s larger than one tenth of the length of the data, i. e. the analysis should be limited by $s < N/10$.

Methods for Non-stationary Fractal Time Series Analysis

Wavelet Analysis

The origins of wavelet analysis come from signal theory, where frequency decompositions of time series were studied [27,28]. Like the Fourier transform, the wavelet transform of a signal $x(t)$ is a convolution integral to be replaced by a summation in case of a discrete time series $(\tilde{x}_i)$, $i = 1, \ldots, N$,

$$L_\psi(\tau, s) = \frac{1}{s}\int_{-\infty}^{\infty} x(t)\psi[(t-\tau)/s]\,\mathrm{d}t = \frac{1}{s}\sum_{i=1}^{N} \tilde{x}_i\,\psi[(i-\tau)/s]\,. \tag{16}$$

Here, $\psi(t)$ is a so-called mother wavelet, from which all daughter wavelets $\psi_{\tau,s}(t) = \psi((t-\tau)/s)$ evolve by shifting and stretching of the time axis. The wavelet coefficients $L_\psi(\tau, s)$ thus depend on both time position τ and scale s. Hence, the local frequency decomposition of the signal is described with a time resolution appropriate for the considered frequency $f = 1/s$ (i. e., inverse time scale).

All wavelets $\psi(t)$ must have zero mean. They are often chosen to be orthogonal to polynomial trends, so that the analysis method becomes insensitive to possible trends in the data. Simple examples are derivatives of a Gaussian, $\psi^{(n)}_{\text{Gauss}}(t) = \frac{\mathrm{d}^n}{\mathrm{d}t^n}\exp(-x^2/2)$, like the Mexican hat wavelet $-\psi^{(2)}_{\text{Gauss}}$ and the Haar wavelet, $\psi^{(1)}_{\text{Haar}}(t) = +1$ if $0 \le t < 1$, -1 if $1 \le t < 2$, and 0 otherwise. It is straightforward to construct higher order Haar wavelets that are orthogonal to linear, quadratic and cubic trends, e. g., $\psi^{(2)}_{\text{Haar}}(t) = 1$ for $t \in [0,1) \cup [2,3)$, -2 for $t \in [1,2)$, and 0 otherwise, or $\psi^{(3)}_{\text{Haar}}(t) = 1$ for $t \in [0,1)$, -3 for $t \in [1,2)$, $+3$ for $t \in [2,3)$, -1 for $t \in [3,4)$, and 0 otherwise.

Discrete Wavelet Transform (WT) Approach

A detrending fractal analysis of time series can be easily implemented by considering Haar wavelet coefficients of the profile $Y(j)$, Eq. (10) [17,30]. In this case the convolution (16) corresponds to the addition and subtraction of mean values of $Y(j)$ within segments of size s. Hence, defining $\bar{Y}_\nu(s) = \frac{1}{s}\sum_{j=1}^{s} Y(\nu s + j)$, the coefficients can be written as

$$F_{\text{WT1}}(\nu, s) \equiv L_{\psi^{(0)}_{\text{Haar}}}(\nu s, s) = \bar{Y}_\nu(s) - \bar{Y}_{\nu+1}(s)\,, \tag{17}$$

$$F_{\text{WT2}}(\nu, s) \equiv L_{\psi^{(1)}_{\text{Haar}}}(\nu s, s) = \bar{Y}_\nu(s) - 2\bar{Y}_{\nu+1}(s) + \bar{Y}_{\nu+2}(s)\,, \tag{18}$$

and

$$F_{\text{WT3}}(\nu, s) \equiv L_{\psi^{(2)}_{\text{Haar}}}(\nu s, s) = \bar{Y}_\nu(s) - 3\bar{Y}_{\nu+1}(s) + 3\bar{Y}_{\nu+2}(s) - \bar{Y}_{\nu+3}(s) \tag{19}$$

for constant, linear and quadratic detrending, respectively. The generalization for higher orders of detrending is obvious. The resulting mean-square fluctuations $F^2_{\text{WT}n}(\nu, s)$ are averaged over all ν to obtain the mean fluctuation $F_2(s)$, see Eq. (13). Figure 5 shows typical results for WT analysis of long-term correlated, short-term correlated and uncorrelated data.

Regarding trend-elimination, wavelet transform WT0 corresponds to standard FA (see Subsect. "Fluctuation Analysis (FA)"), and only constant trends in the profile are eliminated. WT1 is similar to Hurst's rescaled range analysis (see Subsect. "Hurst's Rescaled-Range Analysis"): linear trends in the profile and constant trends in the data are eliminated, and the range of the fluctuation exponent

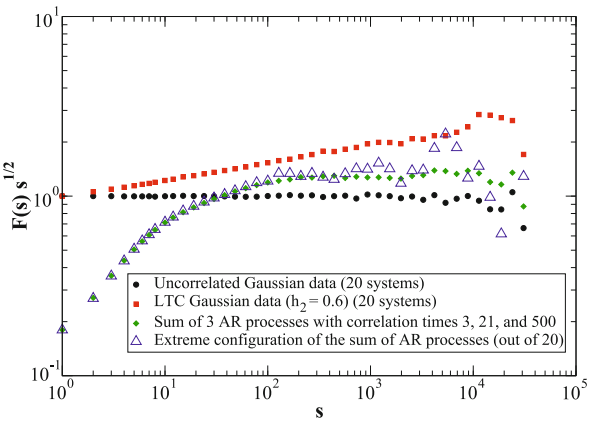

Fractal and Multifractal Time Series, Figure 5
Application of discrete wavelet transform (WT) analysis on uncorrelated data (*black circles*), long-term correlated data ($\gamma = 0.8$, $\alpha = 0.6$, *red squares*), and short-term correlated data (summation of three AR processes, *green diamonds*). Averages of $F_2(s)$ averaged over 20 series with $N = 2^{16}$ points and divided by $s^{1/2}$ are shown, so that a horizontal line corresponds to uncorrelated behavior. The *blue open triangles* show the result for one selected extreme configuration, where it is hard to decide about the existence of long-term correlations (figure after [29])

$\alpha \approx H$ is up to 2. In general, WTn determines the fluctuations from the nth derivative, this way eliminating trends described by $(n-1)$st-order polynomials in the data. The results become statistically unreliable for scales s larger than one tenth of the length of the data, just as for FA.

Detrended Fluctuation Analysis (DFA)

In the last 14 years *Detrended Fluctuation Analysis* (DFA), originally introduced by Peng et al. [31], has been established as an important method to reliably detect long-range (auto-) correlations in non-stationary time series. The method is based on random walk theory and basically represents a linear detrending version of FA (see Subsect. "Fluctuation Analysis (FA)"). DFA was later generalized for higher order detrending [15], separate analysis of sign and magnitude series [32] (see Subsect. "Sign and Magnitude (Volatility) DFA"), multifractal analysis [33] (see Subsect. "Multifractal Detrended Fluctuation Analysis (MFDFA)"), and data with more than one dimension [34]. Its features have been studied in many articles [35,36,37,38,39,40]. In addition, several comparisons of DFA with other methods for stationary and non-stationary time-series analysis have been published, see, e. g., [21,23,41,42] and in particular [22], where DFA is compared with many other established methods for short data sets, and [43], where it is compared with recently suggested improved methods. Altogether, there are about 600 papers applying DFA (till September 2008). In most cases positive auto-correlations were reported leaving only a few exceptions with anti-correlations, see, e. g., [44,45,46].

Like in the FA method, one first calculates the global profile according to Eq. (10) and divides the profile into $N_s = \mathrm{int}(N/s)$ non-overlapping segments of size s starting from the beginning and another N_s segments starting from the end of the considered series. DFA explicitly deals with monotonous trends in a detrending procedure. This is done by estimating a polynomial trend $y^m_{\nu,s}(j)$ within each segment ν by least-square fitting and subtracting this trend from the original profile ('detrending'),

$$\tilde{Y}_s(j) = Y(j) - y^m_{\nu,s}(j) \,. \tag{20}$$

The degree of the polynomial can be varied in order to eliminate constant ($m = 0$), linear ($m = 1$), quadratic ($m = 2$) or higher order trends of the profile function [15]. Conventionally the DFA is named after the order of the fitting polynomial (DFA0, DFA1, DFA2, ...). In DFAm, trends of order m in the profile $Y(j)$ and of order $m-1$ in the original record $\tilde{x}_i$ are eliminated. The variance of the detrended profile $\tilde{Y}_s(j)$ in each segment ν yields the mean-

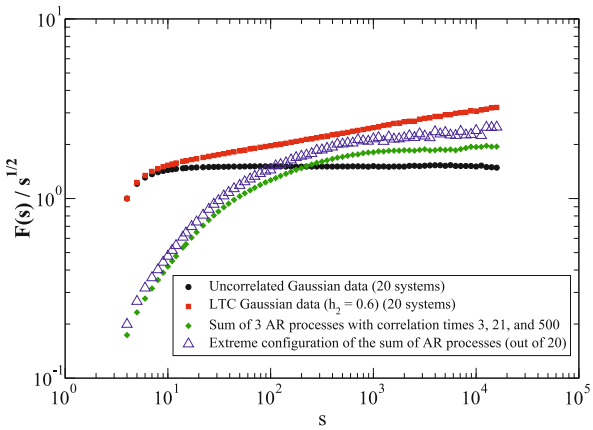

Fractal and Multifractal Time Series, Figure 6
Application of Detrended Fluctuation Analysis (DFA) on the data already studied in Fig. 5 (figure after [29])

square fluctuations,

$$F^2_{\mathrm{DFA}m}(\nu, s) = \frac{1}{s} \sum_{j=1}^{s} \tilde{Y}^2_s(j) \,. \tag{21}$$

As for FA and discrete wavelet analysis, the $F^2_{\mathrm{DFA}m}(\nu, s)$ are averaged over all segments ν to obtain the mean fluctuations $F_2(s)$, see Eq. (13). Calculating $F_2(s)$ for many s, the fluctuation scaling exponent α can be determined just as with FA, see Eq. (14). Figure 6 shows typical results for DFA of the same long-term correlated, short-term correlated and uncorrelated data studied already in Fig. 5.

We note that in studies that include averaging over many records (or one record cut into many separate pieces by the elimination of some unreliable intermediate data points) the averaging procedure (13) must be performed for all data. Taking the square root is always the final step after all averaging is finished. It is not appropriate to calculate $F_2(s)$ for parts of the data and then average the $F_2(s)$ values, since such a procedure will bias the results towards smaller scaling exponents on large time scales.

If $F_2(s)$ increases for increasing s by $F_2(s) \sim s^\alpha$ with $0.5 < \alpha < 1$, one finds that the scaling exponent $\alpha \approx H$ is related to the correlation exponent γ by $\alpha = 1 - \gamma/2$ (see Eq. (15)). A value of $\alpha = 0.5$ thus indicates that there are no (or only short-range) correlations. If $\alpha > 0.5$ for all scales s, the data are long-term correlated. The higher α, the stronger the correlations in the signal are. $\alpha > 1$ indicates a non-stationary local average of the data; in this case, FA fails and yields only $\alpha = 1$. The case $\alpha < 0.5$ corresponds to long-term anti-correlations, meaning that large values are most likely to be followed by small values and vice versa. α values below 0 are not possible. Since the

maximum value for α in DFAm is $m+1$, higher detrending orders should be used for very non-stationary data with large α. Like in FA and Hurst's analysis, α will decrease or increase by one upon additional differentiation or integration of the data, respectively.

Small deviations from the scaling law (14), i. e. deviations from a straight line in a double logarithmic plot, occur for small scales s, in particular for DFAm with large detrending order m. These deviations are intrinsic to the usual DFA method, since the scaling behavior is only approached asymptotically. The deviations limit the capability of DFA to determine the correct correlation behavior in very short records and in the regime of small s. DFA6, e. g., is only defined for $s \geq 8$, and significant deviations from the scaling law $F_2(s) \sim s^{\alpha}$ occur even up to $s \approx 30$. They will lead to an over-estimation of the fluctuation exponent α, if the regime of small s is used in a fitting procedure. An approach for correction of this systematic artefact in DFA is described in [35].

The number of independent segments of length s is larger in DFA than in WT, and the fluctuations in FA are larger than in DFA. Hence, the analysis has to be based on s values lower than $s_{\max} = N/4$ for DFA compared with $s_{\max} = N/10$ for FA and WT. The accuracy of scaling exponents α determined by DFA was recently studied as a function of the length N of the data [43] (fitting range $s \in [10, N/2]$ was used). The results show that statistical standard errors of α (one standard deviation) are approximately 0.1 for $N = 500$, 0.05 for $N = 3000$, and reach 0.03 for $N = 10{,}000$. Findings of long-term correlations with $\alpha = 0.6$ in data with only 500 points are thus not significant.

A generalization of DFA for two-dimensional data (or even higher dimensions d) was recently suggested [34]. The generalization works well when tested with synthetic surfaces including fractional Brownian surfaces and multifractal surfaces. In the 2D procedure, a double cumulative sum (profile) is calculated by summing over both directional indices analogous with Eq. (10), $Y(k,l) = \sum_{i=1}^{k} \sum_{j=1}^{l} \tilde{x}_{i,j}$. This surface is partitioned into squares of size $s \times s$ with indices ν and μ, in which polynomials like $y^2_{\nu,\mu,s}(i,j) = ai^2 + bj^2 + cij + di + ej + f$ are fitted. The fluctuation function $F_2(s)$ is again obtained by calculating the variance of the profile from the fits.

Detection of Trends and Crossovers with DFA

Frequently, the correlations of recorded data do not follow the same scaling law for all time scales s, but one or sometimes even more crossovers between different scaling regimes are observed (see Subsect. "Crossovers and Non-stationarities in Time Series"). Time series with a well-defined crossover at $s_\times$ and vanishing correlations above $s_\times$ are most easily generated by Fourier filtering (see Subsect. "Fourier Filtering"). The power spectrum $S(f)$ of an uncorrelated random series is multiplied by $(f/f_\times)^{-\beta}$ with $\beta = 2\alpha - 1$ for frequencies $f > f_\times = 1/s_\times$ only. The series obtained by inverse Fourier transform of this modified power spectrum exhibits power-law correlations on time scales $s < s_\times$ only, while the behavior becomes uncorrelated on larger time scales $s > s_\times$.

The crossover from $F_2(s) \sim s^{\alpha}$ to $F_2(s) \sim s^{1/2}$ is clearly visible in double logarithmic plots of the DFA fluctuation function for such short-term correlated data. However, it occurs at times $s_\times^{(m)}$ that are different from the original $s_\times$ used for the generation of the data and that depend on the detrending order m. This systematic deviation is most significant in the DFAm with higher m. Extensive numerical simulations (see Fig. 3 in [35]) show that the ratios of $s_\times^{(m)}/s_\times$ are 1.6, 2.6, 3.6, 4.5, and 5.4 for DFA1, DFA2, …, DFA5, with an error bar of approximately 0.1. Note, however, that the precise value of this ratio will depend on the method used for fitting the crossover times $s_\times^{(m)}$ (and the method used for generating the data if generated data is analyzed). If results for different orders of DFA shall be compared, an observed crossover $s_\times^{(m)}$ can be systematically corrected dividing by the ratio for the corresponding DFAm. If several orders of DFA are used in the procedure, several estimates for the real $s_\times$ will be obtained, which can be checked for consistency or used for an error approximation. A real crossover can thus be well distinguished from the effects of non-stationarities in the data, which lead to a different dependence of an apparent crossover on m.

The procedure is also required if the characteristic time scale of short-term correlations shall be studied with DFA. If consistent (corrected) $s_\times$ values are obtained based on DFAm with different m, the existence of a real characteristic correlation time scale is positively confirmed. Note that lower detrending orders are advantageous in this case, since the observed crossover time scale $s_\times^{(m)}$ might become quite large and nearly reach one forth of the total series length ($N/4$), where the results become statistically inaccurate.

We would like to note that studies showing scaling long-term correlations should not be based on DFA or variants of this method alone in most applications. In particular, if it is not clear whether a given time series is indeed long-term correlated or just short-term correlated with a fairly large crossover time scale, results of DFA should be compared with other methods. For example, one can employ wavelet methods (see, e. g., Subsect. "Discrete

Wavelet Transform (WT) Approach"). Another option is to remove short-term correlations by considering averaged series for comparison. For a time series with daily observations and possible short-term correlations up to two years, for example, one might consider the series of two-year averages and apply DFA together with FA, binned power spectra analysis, and/or wavelet analysis. Only if these methods still indicate long-term correlations, one can be sure that the data are indeed long-term correlated.

As discussed in Subsect. "Crossovers and Non-stationarities in Time Series", records from real measurements are often affected by non-stationarities, and in particular by trends. They have to be well distinguished from the intrinsic fluctuations of the system. To investigate the effect of trends on the DFAm fluctuation functions, one can generate artificial series $(\tilde{x}_i)$ with smooth monotonous trends by adding polynomials of different power p to the original record (x_i),

$$\tilde{x}_i = x_i + Ax^p \quad \text{with} \quad x = i/N\,. \tag{22}$$

For the DFAm, such trends in the data can lead to an artificial crossover in the scaling behavior of $F_2(s)$, i. e., the slope α is strongly increased for large time scales s. The position of this artificial crossover depends on the strength A and the power p of the trend. Evidently, no artificial crossover is observed, if the detrending order m is larger than p and p is integer. The order p of the trends in the data can be determined easily by applying the different DFAm. If p is larger than m or p is not an integer, an artificial crossover is observed, the slope α_{trend} in the large s regime strongly depends on m, and the position of the artificial crossover also depends strongly on m. The artificial crossover can thus be clearly distinguished from real crossovers in the correlation behavior, which result in identical slopes α and rather similar crossover positions for all detrending orders m. For more extensive studies of trends with non-integer powers we refer to [35,36]. The effects of periodic trends are also studied in [35].

If the functional form of the trend in given data is not known a priori, the fluctuation function $F_2(s)$ should be calculated for several orders m of the fitting polynomial. If m is too low, $F_2(s)$ will show a pronounced crossover to a regime with larger slope for large scales s [35,36]. The maximum slope of $\log F_2(s)$ versus $\log s$ is $m+1$. The crossover will move to larger scales s or disappear when m is increased, unless it is a real crossover not due to trends. Hence, one can find m such that detrending is sufficient. However, m should not be larger than necessary, because shifts of the observed crossover time scales and deviations on short scales s increase with increasing m.

Sign and Magnitude (Volatility) DFA

To study the origin of long-term fractal correlations in a time series, the series can be split into two parts which are analyzed separately. It is particularly useful to split the series of increments, $\Delta x_i = x_i - x_{i-1}$, $i = 1, \ldots, N$, into a series of signs $\tilde{x}_i = s_i = \text{sign}\Delta x_i$ and a series of magnitudes $\tilde{x}_i = m_i = |\Delta x_i|$ [32,47,48]. There is an extensive interest in the magnitude time series in economics [49,50]. These data, usually called volatility, represent the absolute variations in stock (or commodity) prices and are used as a measure quantifying the risk of investments. While the actual prices are only short-term correlated, long-term correlations have been observed in volatility series [49,50].

Time series having identical distributions and long-range correlation properties can exhibit quite different temporal organizations of the magnitude and sign sub-series. The DFA method can be applied independently to both of these series. Since in particular the signs are often rather strongly anti-correlated and DFA will give incorrect results if α is too close to zero, one often studies integrated sign and magnitude series. As mentioned above, integration $\tilde{x}_i \to \sum_{j=1}^{i} \tilde{x}_j$ increases α by one.

Most published results report short-term anti-correlations and no long-term correlations in the sign series, i. e., $\alpha_{\text{sign}} < 1/2$ for the non-integrated signs s_i (or $\alpha_{\text{sign}} < 3/2$ for the integrated signs) on low time scales and $\alpha_{\text{sign}} \to 1/2$ asymptotically for large s. The magnitude series, on the other hand, are usually either uncorrelated $\alpha_{\text{magn}} = 1/2$ (or 3/2) or positively long-term correlated $\alpha_{\text{magn}} > 1/2$ (or 3/2). It has been suggested that findings of $\alpha_{\text{magn}} > 1/2$ are related with nonlinear properties of the data and in particular multifractality [32,47,48], if $\alpha < 1.5$ in standard DFA. Specifically, the results suggest that the correlation exponent of the magnitude series is a monotonically increasing function of the multifractal spectrum (i. e., the singularity spectrum) width of the original series (see Subsect. "The Structure Function Approach and Singularity Spectra"). On the other hand, the sign series mainly relates to linear properties of the original series. At small time scales $s < 16$ the standard α is approximately the average of α_{sign} and α_{magn}, if integrated sign and magnitude series are analyzed. For $\alpha > 1.5$ in the original series, the integrated magnitude and sign series have approximately the same two-point scaling exponents [47]. An analytical treatment is presented in [48].

Further Detrending Approaches

A possible drawback of the DFA method is the occurrence of abrupt jumps in the detrended profile $\tilde{Y}_s(j)$ (Eq. (20)) at the boundaries between the segments, since the fit-

ting polynomials in neighboring segments are not related. A possible way to avoid these jumps would be the calculation of $F_2(s)$ based on polynomial fits in overlapping windows. However, this is rather time consuming due to the polynomial fit in each segment and is consequently not done in most applications. To overcome the problem of jumps several modifications and extensions of the FA and DFA methods have been suggested in recent years. These methods include

- The detrended moving average technique [51,52,53], which we denote by the backward moving average (BMA) technique (following [54]),
- The centered moving average (CMA) method [54], an essentially improved version of BMA,
- The modified detrended fluctuation analysis (MDFA) [55], which is essentially a mixture of old FA and DFA,
- The continuous DFA (CDFA) technique [56,57], which is particularly useful for the detection of crossovers,
- The Fourier DFA [58],
- A variant of DFA based on empirical mode decomposition (EMD) [59],
- A variant of DFA based on singular value decomposition (SVD) [60,61], and
- A variant of DFA based on high-pass filtering [62].

Detrended moving average techniques will be thoroughly described and discussed in the next section. A study comparing DFA with CMA and MDFA can be found in [43]. For studies comparing DFA and BMA, see [63,64]; note that [64] also discusses CMA.

The method we denote as *modified detrended fluctuation analysis* (MDFA) [55], eliminates trends similar to the DFA method. A polynomial is fitted to the profile function $Y(j)$ in each segment ν and the deviation between the profile function and the polynomial fit is calculated, $\tilde{Y}_s(j) = Y(j) - y^p_{\nu,s}(j)$ (Eq. (20)). To estimate correlations in the data, this method uses a derivative of $\tilde{Y}_s(j)$, obtained for each segment ν, by $\Delta\tilde{Y}_s(j) = \tilde{Y}_s(j+s/2) - \tilde{Y}_s(j)$. Hence, the fluctuation function (compare with Eqs. (13) and (21)) is calculated as follows:

$$F_2(s) = \left[\frac{1}{N}\sum_{j=1}^{N}\left(\tilde{Y}_s(j+s/2) - \tilde{Y}_s(j)\right)^2\right]^{1/2} . \qquad (23)$$

As in case of DFA, MDFA can easily be generalized to remove higher order trends in the data. Since the fitting polynomials in adjacent segments are not related, $\tilde{Y}_s(j)$ shows abrupt jumps on their boundaries as well. This leads to fluctuations of $F_2(s)$ for large segment sizes s and limits the maximum usable scale to $s < N/4$ as for DFA. The detection of crossovers in the data, however, is more exact with MDFA (compared with DFA), since no correction of the estimated crossover time scales seems to be needed [43].

The *Fourier-detrended fluctuation analysis* [58] aims to eliminate slow oscillatory trends which are found especially in weather and climate series due to seasonal influences. The character of these trends can be rather periodic and regular or irregular, and their influence on the detection of long-range correlations by means of DFA was systematically studied previously [35]. Among other things it has been shown that slowly varying periodic trends disturb the scaling behavior of the results much stronger than quickly oscillating trends and thus have to be removed prior to the analysis. In the case of periodic and regular oscillations, e. g., in temperature fluctuations, one simply removes the low frequency seasonal trend by subtracting the daily mean temperatures from the data. Another way, which the Fourier-detrended fluctuation analysis suggests, is to filter out the relevant frequencies in the signals' Fourier spectrum before applying DFA to the filtered signal. Nevertheless, this method faces several difficulties especially its limitation to periodic and regular trends and the need for a priori knowledge of the interfering frequency band.

To study correlations in data with quasi-periodic or irregular oscillating trends, *empirical mode decomposition* (EMD) was suggested [59]. The EMD algorithm breaks down the signal into its intrinsic mode functions (IMFs) which can be used to distinguish between fluctuations and background. The background, estimated by a quasi-periodic fit containing the dominating frequencies of a sufficiently large number of IMFs, is subtracted from the data, yielding a slightly better scaling behavior in the DFA curves. However, we believe that the method might be too complicated for wide-spread applications.

Another method which was shown to minimize the effect of periodic and quasi-periodic trends is based on *singular value decomposition* (SVD) [60,61]. In this approach, one first embeds the original signal in a matrix whose dimension has to be much larger than the number of frequency components of the periodic or quasi-periodic trends obtained in the power spectrum. Applying SVD yields a diagonal matrix which can be manipulated by setting the dominant eigenvalues (associated with the trends) to zero. The filtered matrix finally leads to the filtered data, and it has been shown that subsequent application of DFA determines the expected scaling behavior if the embedding dimension is sufficiently large. None the

less, the performance of this rather complex method seems to decrease for larger values of the scaling exponent. Furthermore SVD-DFA assumes that trends are deterministic and narrow banded.

The detrending procedure in DFA (Eq. (20)) can be regarded as a scale-dependent high-pass filter since (low-frequency) fluctuations exceeding a specific scale s are eliminated. Therefore, it has been suggested to obtain the detrended profile $\tilde{Y}_s(j)$ for each scale s directly by applying digital high-pass filters [62]. In particular, Butterworth, Chebyshev-I, Chebyshev-II, and an elliptical filter were suggested. While the elliptical filter showed the best performance in detecting long-range correlations in artificial data, the Chebyshev-II filter was found to be problematic. Additionally, in order to avoid a time shift between filtered and original profile, the average of the directly filtered signal and the time reversed filtered signal is considered. The effects of these complicated filters on the scaling behavior are, however, not fully understood.

Finally, a continuous DFA method has been suggested in the context of studying heartbeat data during sleep [56,57]. The method compares unnormalized fluctuation functions $F_2(s)$ for increasing length of the data. I. e., one starts with a very short recording and subsequently adds more points of data. The method is particularly suitable for the detection of change points in the data, e. g., physiological transitions between different activity or sleep stages. Since the main objective of the method is not the study of scaling behavior, we do not discuss it in detail here.

Centered Moving Average (CMA) Analysis

Particular attractive modifications of DFA are the *detrended moving average* (DMA) methods, where running averages replace the polynomial fits. The first suggested version, the *backward moving average* (BMA) method [51,52,53], however, suffers from severe problems, because an artificial time shift of s between the original signal and the moving average is introduced. This time shift leads to an additional contribution to the detrended profile $\tilde{Y}_s(j)$, which causes a larger fluctuation function $F_2(s)$ in particular for small scales in the case of long-term correlated data. Hence, the scaling exponent α is systematically underestimated [63]. In addition, the BMA method preforms even worse for data with trends [64], and its slope is limited by $\alpha < 1$ just as for the non-detrending method FA.

It was soon recognized that the intrinsic error of BMA can be overcome by eliminating the artificial time shift. This leads to the *centered moving average* (CMA) method [54], where $\tilde{Y}_s(j)$ is calculated as

$$\tilde{Y}_s(j) = Y(j) - \frac{1}{s} \sum_{i=-(s-1)/2}^{(s-1)/2} Y(j+i)\,, \tag{24}$$

replacing Eq. (20) while Eq. (21) and the rest of the DFA procedure described in Subsect. "Detrended Fluctuation Analysis (DFA)" stay the same. Unlike DFA, the CMA method cannot easily be generalized to remove linear and higher order trends in the data.

It was recently proposed [43] that the scaling behavior of the CMA method is more stable than for DFA1 and MDFA1, suggesting that CMA could be used for reliable computation of α even for scales $s < 10$ (without correction of any systematic deviations needed in DFA for this regime) and up to $s_{\max} = N/2$. The standard errors in determining the scaling exponent α by fitting straight lines to the double logarithmic plots of $F_2(s)$ have been studied in [43]; they are comparable with DFA1 (see end of Subsect. "Detrended Fluctuation Analysis (DFA)").

Regarding the determination of crossovers, CMA is comparable to DFA1. Ultimately, the CMA seems to be a good alternative to DFA1 when analyzing the scaling properties in short data sets without trends. Nevertheless for data with possible unknown trends we recommend the application of standard DFA with several different detrending polynomial orders in order to distinguish real crossovers from artificial crossovers due to trends. In addition, an independent approach (e. g., wavelet analysis) should be used to confirm findings of long-term correlations (see also Subsect. "Detection of Trends and Crossovers with DFA").

Methods for Multifractal Time Series Analysis

This section describes the multifractal characterization of time series, for an introduction, see Subsect. "Multifractal Time Series". The simplest type of multifractal analysis is based upon the standard partition function multifractal formalism, which has been developed for the multifractal characterization of normalized, stationary measures [6,12,65,66]. Unfortunately, this standard formalism does not give correct results for non-stationary time series that are affected by trends or that cannot be normalized. Thus, in the early 1990s an improved multifractal formalism has been developed, the wavelet transform modulus maxima (WTMM) method [67,68,69,70,71], which is based on wavelet analysis and involves tracing the maxima lines in the continuous wavelet transform over all scales. An important alternative is the multifractal DFA (MFDFA) algorithm [33], which does not require the

modulus maxima procedure, and hence involves little more effort in programming than the conventional DFA. For studies comparing methods for detrending multifractal analysis (MFDFA and WTMM, see [33,72,73].

The Structure Function Approach and Singularity Spectra

In the general multifractal formalism, one considers a normalized measure $\mu(t)$, $t \in [0,1]$, and defines the box probabilities $\tilde{\mu}_s(t) = \int_{t-s/2}^{t+s/2} \mu(t')\,\mathrm{d}t'$ in neighborhoods of (scale) length $s \ll 1$ around t. The multifractal approach is then introduced by the partition function

$$Z_q(s) = \sum_{\nu=0}^{1/s-1} \tilde{\mu}_s^q[(\nu+1/2)s] \sim s^{\tau(q)} \quad \text{for } s \ll 1\,, \tag{25}$$

where $\tau(q)$ is the Renyi scaling exponent and q is a real parameter that can take positive as well as negative values. Note that $\tau(q)$ is sometimes defined with opposite sign (see, e. g., [6]). A record is called monofractal (or self-affine), when the Renyi scaling exponent $\tau(q)$ depends linearly on q; otherwise it is called multifractal. The generalized multifractal dimensions $D(q)$ (see also Subsect. "Multifractal Time Series") are related to $\tau(q)$ by $D(q) = \tau(q)/(q-1)$, such that the fractal dimension of the support is $D(0) = -\tau(0)$ and the correlation dimension is $D(2) = \tau(2)$.

In time series, a discrete version has to be used, and the considered data (x_i), $i = 1,\ldots,N$ may usually include negative values. Hence, setting $N_s = \mathrm{int}(N/s)$ and $X(\nu,s) = \sum_{i=1}^{s} x_{\nu s+i}$ for $\nu = 0,\ldots,N_s-1$ we can define [6,12],

$$Z_q(s) = \sum_{\nu=0}^{N_s-1} |X(\nu,s)|^q \sim s^{\tau(q)} \quad \text{for } s > 1\,. \tag{26}$$

Inserting the profile $Y(j)$ and $F_{\mathrm{FA}}(\nu,s)$ from Eqs. (10) and (11), respectively, we obtain

$$\begin{aligned} Z_q(s) &= \sum_{\nu=0}^{N_s-1} \left\{[Y((\nu+1)s) - Y(\nu s)]^2\right\}^{q/2} \\ &= \sum_{\nu=1}^{N_s} |F_{\mathrm{FA}}(\nu,s)|\,. \end{aligned} \tag{27}$$

Comparing Eq. (27) with (13), we see that this multifractal approach can be considered as a generalized version of the fluctuation analysis (FA) method, where the exponent 2 is replaced by q. In particular we find (disregarding the summation over the second partition of the time series)

$$F_2(s) \sim \left[\frac{1}{N_s} Z_2(s)\right]^{1/2} \sim s^{[1+\tau(2)]/2} \Rightarrow 2\alpha = 1 + \tau(2) = 1 + D(2). \tag{28}$$

We thus see that all methods for (mono-)fractal time analysis (discussed in Sect. "Methods for Stationary Fractal Time Series Analysis"and Sect. "Methods for Non-stationary Fractal Time Series Analysis") in fact study the correlation dimension $D(2) = 2\alpha - 1 = \beta = 1 - \gamma$ (see Eq. (15)).

It is straightforward to define a generalized (multifractal) Hurst exponent $h(q)$ for the scaling behavior of the qth moments of the fluctuations [65,66],

$$F_q(s) = \left[\frac{1}{N_s} Z_2(s)\right]^{1/q} \sim s^{[1+\tau(q)]/q} = s^{h(q)} \Rightarrow h(q) = \frac{1+\tau(q)}{q} \tag{29}$$

with $h(2) = \alpha \approx H$. In the following, we will use only $h(2)$ for the standard fluctuation exponent (denoted by α in the previous chapters), and reserve the letter α for the Hölder exponent.

Another way to characterize a multifractal series is the singularity spectrum $f(\alpha)$, that is related to $\tau(q)$ via a Legendre transform [6,12],

$$\alpha = \frac{\mathrm{d}}{\mathrm{d}q}\tau(q) \quad \text{and} \quad f(\alpha) = q\alpha - \tau(q)\,. \tag{30}$$

Here, α is the singularity strength or Hölder exponent (see also ► Fractals and Multifractals, Introduction to in the encyclopedia), while $f(\alpha)$ denotes the dimension of the subset of the series that is characterized by α. Note that α is *not* the fluctuation scaling exponent in this section, although the same letter is traditionally used for both. Using Eq. (29), we can directly relate α and $f(\alpha)$ to $h(q)$,

$$\alpha = h(q) + qh'(q) \quad \text{and} \quad f(\alpha) = q[\alpha - h(q)] + 1\,. \tag{31}$$

Wavelet Transform Modulus Maxima (WTMM) Method

The wavelet transform modulus maxima (WTMM) method [67,68,69,70,71] is a well-known method to investigate the multifractal scaling properties of fractal and self-affine objects in the presence of non-stationarities. For applications, see e. g. [74,75]. It is based upon the wavelet transform with continuous basis functions as defined in Subsect. "Wavelet Analysis", Eq. (16). Note that in this

case the series $\tilde{x}_i$ are analyzed directly instead of the profile $Y(j)$ defined in Eq. (10). Using wavelets orthogonal to mth order polynomials, the corresponding trends are eliminated.

Instead of averaging over all wavelet coefficients $L_\psi(\tau, s)$, one averages, within the modulo-maxima method, only the local maxima of $|L_\psi(\tau, s)|$. First, one determines for a given scale s, the positions τ_j of the local maxima of $|W(\tau, s)|$ as a function of τ, so that $|L_\psi(\tau_j - 1, s)| < |L_\psi(\tau_j, s)| \geq |L_\psi(\tau_j + 1, s)|$ for $j = 1, \ldots, j_{\max}$. This maxima procedure is demonstrated in Fig. 7. Then one sums up the qth power of the maxima,

$$Z(q, s) = \sum_{j=1}^{j_{\max}} |L_\psi(\tau_j, s)|^q \,. \tag{32}$$

The reason for the maxima procedure is that the absolute wavelet coefficients $|L_\psi(\tau, s)|$ can become arbitrarily small. The analyzing wavelet $\psi(x)$ must always have positive values for some x and negative values for other x, since it has to be orthogonal to possible constant trends. Hence there are always positive and negative terms in the sum (16), and these terms might cancel. If that happens, $|L_\psi(\tau, s)|$ can become close to zero. Since such small terms would spoil the calculation of negative moments in Eq. (32), they have to be eliminated by the maxima procedure.

In fluctuation analysis, on the other hand, the calculation of the variances $F^2(\nu, s)$, e. g. in Eq. (11), involves only positive terms under the summation. The variances cannot become arbitrarily small, and hence no maximum procedure is required for series with compact support. In addition, the variances will increase if the segment length s is increased, because the fit will usually be worse for a longer segment. In the WTMM method, in contrast, the absolute wavelet coefficients $|L_\psi(\tau, s)|$ need not increase with increasing scale s, even if only the local maxima are considered. The values $|L_\psi(\tau, s)|$ might become smaller for increasing s since just more (positive and negative) terms are included in the summation (16), and these might cancel even better. Thus, an additional supremum procedure has been introduced in the WTMM method in order to keep the dependence of $Z(q, s)$ on s monotonous. If, for a given scale s, a maximum at a certain position τ_j happens to be smaller than a maximum at $\tau'_j \approx \tau_j$ for a lower scale $s' < s$, then $L_\psi(\tau_j, s)$ is replaced by $L_\psi(\tau'_j, s')$ in Eq. (32).

Often, scaling behavior is observed for $Z(q, s)$, and scaling exponents $\hat{\tau}(q)$ can be defined that describe how $Z(q, s)$ scales with s,

$$Z(q, s) \sim s^{\hat{\tau}(q)} \,. \tag{33}$$

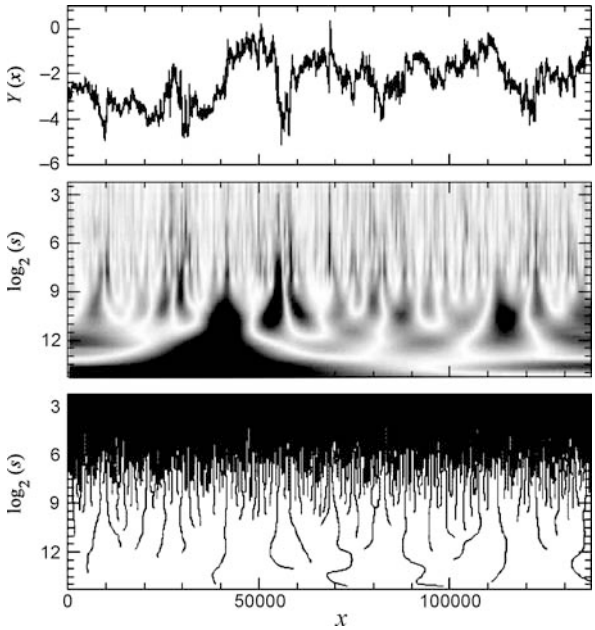

Fractal and Multifractal Time Series, Figure 7
Example of the wavelet transform modulus maxima (WTMM) method, showing the original data (*top*), its continuous wavelet transform (*gray scale* coded amplitude of wavelet coefficients, *middle*), and the extracted maxima lines (*bottom*) (figure taken from [68])

The exponents $\hat{\tau}(q)$ characterize the multifractal properties of the series under investigation, and theoretically they are identical with the $\tau(q)$ defined in Eq. (26) [67,68,69,71] and related to $h(q)$ by Eq. (29).

Multifractal Detrended Fluctuation Analysis (MFDFA)

The multifractal DFA (MFDFA) procedure consists of five steps [33]. The first three steps are essentially identical to the conventional DFA procedure (see Subsect. "Detrended Fluctuation Analysis (DFA)" and Fig. 4). Let us assume that $(\tilde{x}_i)$ is a series of length N, and that this series is of compact support. The support can be defined as the set of the indices j with nonzero values $\tilde{x}_j$, and it is compact if $\tilde{x}_j = 0$ for an insignificant fraction of the series only. The value of $\tilde{x}_j = 0$ is interpreted as having no value at this j. Note that we are *not* discussing the fractal or multifractal features of the plot of the time series in a two-dimensional graph (see also the discussion in Subsect. "Fractality, Self-Affinity, and Scaling"), but analyzing time series as one-dimensional structures with values assigned to each point. Since real time series always have finite length N, we explicitly want to determine the multifractality of finite series, and we are not discussing the limit for $N \to \infty$ here

(see also Subsect. "The Structure Function Approach and Singularity Spectra").

Step 1 Calculate the profile $Y(j)$, Eq. (10), by integrating the time series.

Step 2 Divide the profile $Y(j)$ into $N_s = \mathrm{int}(N/s)$ non-overlapping segments of equal length s. Since the length N of the series is often not a multiple of the considered time scale s, the same procedure can be repeated starting from the opposite end. Thereby, $2N_s$ segments are obtained altogether.

Step 3 Calculate the local trend for each of the $2N_s$ segments by a least-square fit of the profile. Then determine the variance by Eqs. (20) and (21) for each segment $\nu = 1, \ldots, 2N_s$. Again, linear, quadratic, cubic, or higher order polynomials can be used in the fitting procedure, and the corresponding methods are thus called MFDFA1, MFDFA2, MFDFA3, ...) [33]. In (MF-)DFAm [mth order (MF-)DFA] trends of order m in the profile (or, equivalently, of order $m-1$ in the original series) are eliminated. Thus a comparison of the results for different orders of DFA allows one to estimate the type of the polynomial trend in the time series [35,36].

Step 4 Average over all segments to obtain the qth order fluctuation function

$$F_q(s) = \left\{ \frac{1}{2N_s} \sum_{\nu=1}^{2N_s} \left[F^2_{\mathrm{DFA}m}(\nu, s) \right]^{q/2} \right\}^{1/q} . \qquad (34)$$

This is the generalization of Eq. (13) suggested by the relations derived in Subsect. "The Structure Function Approach and Singularity Spectra". For $q = 2$, the standard DFA procedure is retrieved. One is interested in how the generalized q dependent fluctuation functions $F_q(s)$ depend on the time scale s for different values of q. Hence, we must repeat steps 2 to 4 for several time scales s. It is apparent that $F_q(s)$ will increase with increasing s. Of course, $F_q(s)$ depends on the order m. By construction, $F_q(s)$ is only defined for $s \geq m + 2$.

Step 5 Determine the scaling behavior of the fluctuation functions by analyzing log-log plots $F_q(s)$ versus s for each value of q. If the series $\tilde{x}_i$ are long-range power-law correlated, $F_q(s)$ increases, for large values of s, as a power-law,

$$F_q(s) \sim s^{h(q)} \quad \text{with} \quad h(q) = \frac{1 + \tau(q)}{q} . \qquad (35)$$

For very large scales, $s > N/4$, $F_q(s)$ becomes statistically unreliable because the number of segments N_s for the averaging procedure in step 4 becomes very small. Thus, scales $s > N/4$ should be excluded from the fitting procedure determining $h(q)$. Besides that, systematic deviations from the scaling behavior in Eq. (35), which can be corrected, occur for small scales $s \approx 10$.

The value of $h(0)$, which corresponds to the limit $h(q)$ for $q \to 0$, cannot be determined directly using the averaging procedure in Eq. (34) because of the diverging exponent. Instead, a logarithmic averaging procedure has to be employed,

$$F_0(s) = \exp\left\{ \frac{1}{4N_s} \sum_{\nu=1}^{2N_s} \ln\left[F^2(\nu, s) \right] \right\} \sim s^{h(0)} . \qquad (36)$$

Note that $h(0)$ cannot be defined for time series with fractal support, where $h(q)$ diverges for $q \to 0$.

For monofractal time series with compact support, $h(q)$ is independent of q, since the scaling behavior of the variances $F^2_{\mathrm{DFA}m}(\nu, s)$ is identical for all segments ν, and the averaging procedure in Eq. (34) will give just this identical scaling behavior for all values of q. Only if small and large fluctuations scale differently, there will be a significant dependence of $h(q)$ on q. If we consider positive values of q, the segments ν with large variance $F^2(\nu, s)$ (i. e. large deviations from the corresponding fit) will dominate the average $F_q(s)$. Thus, for positive values of q, $h(q)$ describes the scaling behavior of the segments with large fluctuations. On the contrary, for negative values of q, the segments ν with small variance $F^2_{\mathrm{DFA}m}(\nu, s)$ will dominate the average $F_q(s)$. Hence, for negative values of q, $h(q)$ describes the scaling behavior of the segments with small fluctuations. Figure 8 shows typical results obtained for $F_q(s)$ in the MFDFA procedure.

Usually the large fluctuations are characterized by a smaller scaling exponent $h(q)$ for multifractal series than the small fluctuations. This can be understood from the following arguments. For the maximum scale $s = N$ the fluctuation function $F_q(s)$ is independent of q, since the sum in Eq. (34) runs over only two identical segments. For smaller scales $s \ll N$ the averaging procedure runs over several segments, and the average value $F_q(s)$ will be dominated by the $F^2(\nu, s)$ from the segments with small (large) fluctuations if $q < 0$ ($q > 0$). Thus, for $s \ll N$, $F_q(s)$ with $q < 0$ will be smaller than $F_q(s)$ with $q > 0$, while both become equal for $s = N$. Hence, if we assume an homogeneous scaling behavior of $F_q(s)$ following Eq. (35), the slope $h(q)$ in a log-log plot of $F_q(s)$ with $q < 0$ versus s must be larger than the corresponding slope for $F_q(s)$ with $q > 0$. Thus, $h(q)$ for $q < 0$ will usually be larger than $h(q)$ for $q > 0$.

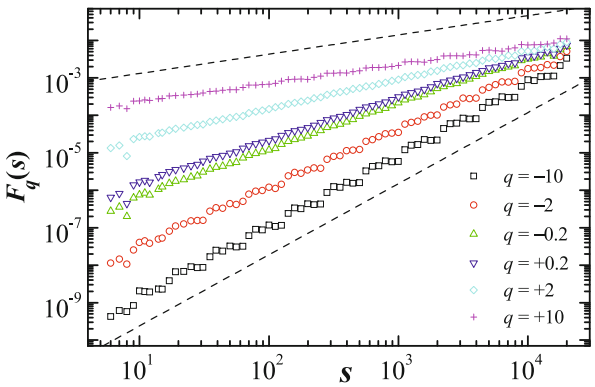

Fractal and Multifractal Time Series, Figure 8
Multifractal detrended fluctuation analysis (MFDFA) of data from the binomial multifractal model (see Subsect. "The Extended Binomial Multifractal Model") with $a = 0.75$. $F_q(s)$ is plotted versus s for the q values given in the legend; the slopes of the curves correspond to the values of $h(q)$. The *dashed lines* have the slopes of the theoretical slopes $h(\pm\infty)$ from Eq. (42). 100 configurations have been averaged

However, the MFDFA method can only determine *positive* generalized Hurst exponents $h(q)$, and it already becomes inaccurate for strongly anti-correlated signals when $h(q)$ is close to zero. In such cases, a modified (MF-)DFA technique has to be used. The most simple way to analyze such data is to integrate the time series before the MFDFA procedure. Following the MFDFA procedure as described above, we obtain a generalized fluctuation functions described by a scaling law with $\tilde{h}(q) = h(q) + 1$. The scaling behavior can thus be accurately determined even for $h(q)$ which are smaller than zero for some values of q.

The accuracy of $h(q)$ determined by MFDFA certainly depends on the length N of the data. For $q = \pm 10$ and data with $N = 10{,}000$ and $100{,}000$, systematic and statistical error bars (standard deviations) up to $\Delta h(q) \approx \pm 0.1$ and $\approx \pm 0.05$ should be expected, respectively [33]. A difference of $h(-10) - h(+10) = 0.2$, corresponding to an even larger width $\Delta\alpha$ of the singularity spectrum $f(\alpha)$ defined in Eq. (30) is thus not significant unless the data was longer than $N = 10{,}000$ points. Hence, one has to be very careful when concluding multifractal properties from differences in $h(q)$.

As already mentioned in the introduction, two types of multifractality in time series can be distinguished. Both of them require a multitude of scaling exponents for small and large fluctuations: (i) Multifractality of a time series can be due to a broad probability density function for the values of the time series, and (ii) multifractality can also be due to different long-range correlations for small and large fluctuations. The most easy way to distinguish between these two types is by analyzing also the corresponding randomly shuffled series [33]. In the shuffling procedure the values are put into random order, and thus all correlations are destroyed. Hence the shuffled series from multifractals of type (ii) will exhibit simple random behavior, $h_{\text{shuf}}(q) = 0.5$, i. e. non-multifractal scaling. For multifractals of type (i), on the contrary, the original $h(q)$ dependence is not changed, $h(q) = h_{\text{shuf}}(q)$, since the multifractality is due to the probability density, which is not affected by the shuffling procedure. If both kinds of multifractality are present in a given series, the shuffled series will show weaker multifractality than the original one.

Comparison of WTMM and MFDFA

The MFDFA results turn out to be slightly more reliable than the WTMM results [33,72,73]. In particular, the MFDFA has slight advantages for negative q values and short series. In the other cases the results of the two methods are rather equivalent. Besides that, the main advantage of the MFDFA method compared with the WTMM method lies in the simplicity of the MFDFA method. However, contrary to WTMM, MFDFA is restricted to studies of data with full one-dimensional support, while WTMM is not. Both WTMM and MFDFA have been generalized for higher dimensional data, see [34] for higher dimensional MFDFA and, e. g., [71] for higher dimensional WTMM. Studies of other generalizations of detrending methods like the discrete WT approach (see Subsect. "Discrete Wavelet Transform (WT) Approach") and the CMA method (see Subsect. "Centered Moving Average (CMA) Analysis") are currently under investigation [76].

Statistics of Extreme Events in Fractal Time Series

The statistics of return intervals between well defined extremal events is a powerful tool to characterize the temporal scaling properties of observed time series and to derive quantities for the estimation of the risk for hazardous events like floods, very high temperatures, or earthquakes. It was shown recently that long-term correlations represent a natural mechanism for the clustering of the hazardous events [77]. In this section we will discuss the most important consequences of long-term correlations and fractal scaling of time series upon the statistics of extreme events [77,78,79,80,81]. Corresponding work regarding multifractal data [82] is not discussed here.

Return Intervals Between Extreme Events

To study the statistics of return intervals we consider again a time series (x_i), $i = 1, \ldots, N$ with fractal scaling behav-

ior, sampled homogeneously and normalized to zero mean and unit variance. For describing the reoccurrence of rare events exceeding a certain threshold q, we investigate the return intervals r_q between these events, see Fig. 9. The average return interval $R_q = \langle r_q \rangle$ increases as a function of the threshold q (see, e. g. [83]). It is known that for uncorrelated records ('white noise'), the return intervals are also uncorrelated and distributed according to the Poisson distribution, $P_q(r) = \frac{1}{R_q} \exp(-r/R_q)$. For fractal (long-term correlated) data with auto-correlations following Eq. (5), we obtain a *stretched* exponential [77,78,79,80,84],

$$P_q(r) = \frac{a_\gamma}{R_q} \exp\left[-b_\gamma (r/R_q)^\gamma\right] . \tag{37}$$

This behavior is shown in Fig. 10. The exponent γ is the correlation exponent from $C(s)$, and the parameters a_γ and b_γ are independent of q. They can be determined from the normalization conditions for $P_q(r)$, i. e., $\int P_q(r)\,\mathrm{d}r = 1$ and $\int rP_q(r)\,\mathrm{d}r = R_q$. The form of the distribution (37) indicates that return intervals both well below and well above their average value R_q (which is independent of γ) are considerably more frequent for long-term correlated than for uncorrelated data. It has to be noted that there are deviations from the stretched exponential law (37) for very small r (discretization effects and an additional power-law regime) and for very large r (finite size effects), see Fig. 10. The extent of the deviations from Eq. (37) depends on the distribution of the values x_i of the time series. For a discussion of these effects, see [80].

Equation (37) does not quantify, however, if the return intervals themselves are arranged in a correlated or in an uncorrelated fashion, and if clustering of rare events may be induced by long-term correlations. To study this question, one has to evaluate the auto-covariance function $C_r(s) = \langle r_q(l) r_q(l+s)\rangle - R_q^2$ of the return intervals. The results for model data suggests that also the return intervals are long-term power-law correlated, with the same exponent γ as the original record. Accordingly, large and small return intervals are not arranged in a random fashion but are expected to form clusters. As a consequence, the probability of finding a certain return interval r depends on the value of the preceding interval r_0, and this effect has to be taken into account in predictions and risk estimations [77,80].

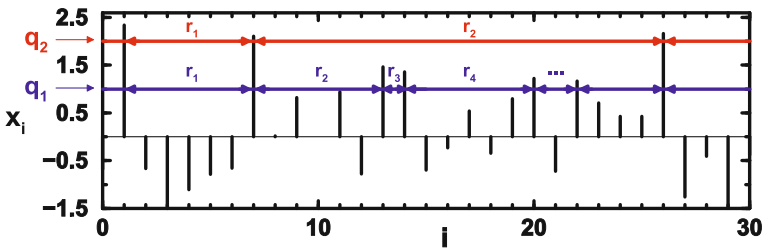

Fractal and Multifractal Time Series, Figure 9
Illustration for the definition of return intervals r_q between extreme events above two quantiles (thresholds) q_1 and q_2 (figure by Jan Eichner)

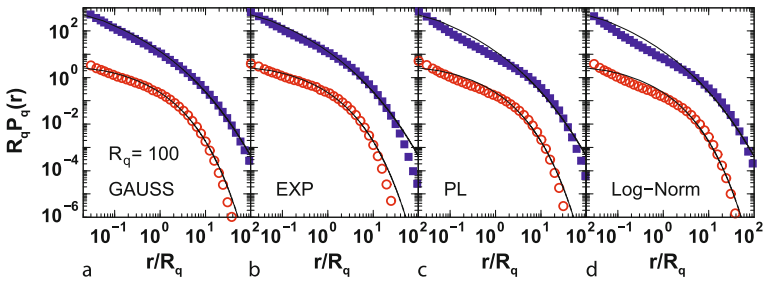

Fractal and Multifractal Time Series, Figure 10
Normalized rescaled distribution density functions $R_qP_q(r)$ of r values with $R_q = 100$ as a function of r/R_q for long-term correlated data with $\gamma = 0.4$ (*open symbols*) and $\gamma = 0.2$ (*filled symbols*; we multiplied the data for the filled symbols by a factor 100 to avoid overlapping curves). In a the original data were Gaussian distributed, in b exponentially distributed, in c power-law distributed with power −5.5, and in d log-normally distributed. All four figures follow quite well stretched exponential curves (*solid lines*) over several decades. For small r/R_q values a power-law regime seems to dominate, while on large scales deviations from the stretched exponential behavior are due to finite-size effects (figure by Jan Eichner)

The conditional distribution function $P_q(r|r_0)$ is a basic quantity, from which the relevant quantities in risk estimations can be derived [83]. For example, the first moment of $P_q(r|r_0)$ is the average value $R_q(r_0)$ of those return intervals that directly follow r_0. By definition, $R_q(r_0)$ is the expected waiting time to the next event, when the two events before were separated by r_0. The more general quantity is the expected waiting time $\tau_q(x|r_0)$ to the next event, when the time x has elapsed. For $x = 0$, $\tau_q(0|r_0)$ is identical to $R_q(r_0)$. In general, $\tau_q(x|r_0)$ is related to $P_q(r|r_0)$ by

$$\tau_q(x|r_0) = \int_x^\infty (r-x)P_q(r|r_0)\mathrm{d}r \Big/ \int_x^\infty P_q(r|r_0)\mathrm{d}r. \quad (38)$$

For uncorrelated records, $\tau_q(x|r_0)/R_q = 1$ (except for discreteness effects that lead to $\tau_q(x|r_0)/R_q > 1$ for $x > 0$, see [85]). Due to the scaling of $P_q(r|r_0)$, also $\tau_q(x|r_0)/R_q$ scales with r_0/R_q and x/R_q. Small and large return intervals are more likely to be followed by small and large ones, respectively, and hence $\tau_q(0|r_0)/R_q = R_q(r_0)/R_q$ is well below (above) one for r_0/R_q well below (above) one. With increasing x, the expected residual time to the next event increases. Note that only for an infinite long-term correlated record, the value of $\tau_q(x|r_0)$ will increase indefinitely with x and r_0. For real (finite) records, there exists a maximum return interval which limits the values of x, r_0 and $\tau_q(x|r_0)$.

Distribution of Extreme Events

In this section we describe how the presence of fractal long-term correlations affects the statistics of the extreme events, i. e., maxima within time segments of fixed duration R, see Fig. 11 for illustration. By definition, extreme events are rare occurrences of extraordinary nature, such as, e. g. floods, very high temperatures, or earthquakes. In hydrological engineering such conventional extreme value statistics are commonly applied to decide what building projects are required to protect river-side areas against typical floods that occur, for example, once in 100 years. Most of these results are based on statistically independent values x_i and hold only in the limit $R \to \infty$. However, both of these assumptions are not strictly fulfilled for correlated fractal scaling data.

In classical extreme value statistics one assumes that records (x_i) consist of i.i.d. data, described by density distributions $P(x)$, which can be, e. g., a Gaussian or an exponential distribution. One is interested in the distribution density function $P_R(m)$ of the maxima (m_j) determined in segments of length R in the original series (x_i), see Fig. 11. Note that all maxima are also elements of the original data.

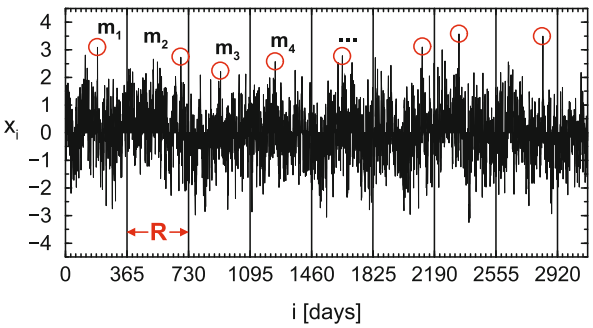

Fractal and Multifractal Time Series, Figure 11
Illustration for the definition of maxima m_R within periods of $R = 365$ values (figure by Jan Eichner)

The corresponding integrated maxima distribution $G_R(m)$ is defined as

$$G_R(m) = 1 - E_R(m) = \int_{-\infty}^m P_R(m')\mathrm{d}m' . \quad (39)$$

Since $G_R(m)$ is the probability of finding a maximum smaller than m, $E_R(m)$ denotes the probability of finding a maximum that exceeds m. One of the main results of traditional extreme value statistics states that for independently and identically distributed (i.i.d.) data (x_i) with Gaussian or exponential distribution density function $P(x)$ the integrated distribution $G_R(m)$ converges to a double exponential (Fisher–Tippet–Gumbel) distribution (often labeled as Type I) [86,87,88,89,90], i. e.,

$$G_R(m) \to G\left(\frac{m-u}{\alpha}\right) = \exp\left[-\exp\left(-\frac{m-u}{\alpha}\right)\right] \quad (40)$$

for $R \to \infty$, where α is the scale parameter and u the location parameter. By the method of moments those parameters are given by $\alpha = \sqrt{6}/\pi\, \sigma_R$ and $u = m_R - n_e\alpha$ with the Euler constant $n_e = 0.577216$ [89,91,92,93]. Here m_R and σ_R denote the (R-dependent) mean maximum and the standard deviation, respectively. Note that different asymptotics will be reached for broader distributions of data (x_i) that belong to other domains of attraction [89]. For example, for data following a power-law distribution (or Pareto distribution), $P(x) = (x/x_0)^{-k}$, $G_R(m)$ converges to a Fréchet distribution, often labeled as Type II. For data following a distribution with finite upper endpoint, for example the uniform distribution $P(x) = 1$ for $0 \le x \le 1$, $G_R(m)$ converges to a Weibull distribution, often labeled as Type III. We do not consider the latter two types of asymptotics here.

Numerical studies of fractal model data have recently shown that the distribution $P(x)$ of the original data has a much stronger effect upon the convergence towards the

Gumbel distribution than the long-term correlations in the data. Long-term correlations just slightly delay the convergence of $G_R(m)$ towards the Gumbel distribution (40). This can be observed very clearly in a plot of the integrated and scaled distribution $G_R(m)$ on logarithmic scale [81].

Furthermore, it was found numerically that (i) the maxima series (m_j) exhibit long-term correlations similar to those of the original data (x_i), and most notably (ii) the maxima distribution as well as the mean maxima significantly depend on the history, in particular on the previous maximum [81]. The last item implies that conditional mean maxima and conditional maxima distributions should be considered for improved extreme event predictions.

Simple Models for Fractal and Multifractal Time Series

Fourier Filtering

Fractal scaling with long-term correlations can be introduced most easily into time series by the Fourier-filtering technique, see, e. g., [94,95,96]. The Fourier filtering technique is not limited to the generation of long-term correlated data characterized by a power-law auto-correlation function $C(s) \sim x^{-\gamma}$ with $0 < \gamma < 1$. All values of the scaling exponents $\alpha = h(2) \approx H$ or $\beta = 2\alpha - 1$ can be obtained, even those that cannot be found directly by the fractal analysis techniques described in Sect. "Methods for Stationary Fractal Time Series Analysis" and Sect. "Methods for Non-stationary Fractal Time Series Analysis" (e. g. $\alpha < 0$). Note, however, that Fourier filtering will always yield Gaussian distributed data values and that no non-linear or multifractal properties can be achieved (see also Subsect. "Multifractal Time Series", Subsect. "Sign and Magnitude (Volatility) DFA", and Sect. "Methods for Multifractal Time Series Analysis"). In Subsect. "Detection of Trends and Crossovers with DFA", we have briefly described a modification of Fourier filtering for obtaining reliable short-term correlated data.

For the generation of data characterized by fractal scaling with $\beta = 2\alpha - 1$ [94,95] we start with uncorrelated Gaussian distributed random numbers x_i from an i.i.d. generator. Transforming a series of such numbers into frequency space with discrete Fourier transform or FFT (fast Fourier transform, for suitable series lengths N) yields a flat power spectrum, since random numbers correspond to white noise. Multiplying the (complex) Fourier coefficients by $f^{-\beta/2}$, where $f \propto 1/s$ is the frequency, will rescale the power spectrum $S(f)$ to follow Eq. (6), as expected for time series with fractal scaling. After transforming back to the time domain (using inverse Fourier transform or inverse FFT) we will thus obtain the desired long-term correlated data $\tilde{x}_i$. The final step is the normalization of this data.

The Fourier filtering method can be improved using modified Bessel functions instead of the simple factors $f^{-\beta/2}$ in modifying the Fourier coefficients [96]. This way problems with the divergence of the autocorrelation function $C(s)$ at $s = 0$ can be avoided.

An alternative method to the Fourier filtering technique, the random midpoint displacement method, is based on the construction of self-affine surfaces by an iterative procedure, see, e. g. [6]. Starting with one interval with constant values, the intervals are iterative split in the middle and the midpoint is displaced by a random offset. The amplitude of this offset is scaled according to the length of the interval. Since the method generates a self-affine surface x_i characterized by a Hurst exponent H, the differentiated series Δx_i can be used as long-term correlated or anti-correlated random numbers. Note, however, that the correlations do not persist for the whole length of the data generated this way. Another option is the use of wavelet synthesis, the reverse of wavelet analysis described in Subsect. "Wavelet Analysis". In that method, the scaling law is introduced by setting the magnitudes of the wavelet coefficients according to the corresponding time scale s.

The Schmitz–Schreiber Method

When long-term correlations in random numbers are introduced by the Fourier-filtering technique (see previous section), the original distribution $P(x)$ of the time series values x_i is always modified such that it becomes closer to a Gaussian. Hence, no series (x_i) with broad distributions of the values *and* fractal scaling can be generated. In these cases an iterative algorithm introduced by Schreiber and Schmitz [98,99] must be applied.

The algorithm consists of the following steps: First one creates a Gaussian distributed long-term correlated data set with the desired correlation exponent γ by standard Fourier-filtering [96]. The power spectrum $S_G(f) = F_G(f)F_G^*(f)$ of this data set is considered as reference spectrum (where f denotes the frequency in Fourier space and the $F_G(f)$ are the complex Fourier coefficients). Next one creates an uncorrelated sequence of random numbers (x_i^{ref}), following a desired distribution $P(x)$. The (complex) Fourier transform $F(f)$ of the (x_i^{ref}) is now divided by its absolute value and multiplied by the square root of the reference spectrum,

$$F_{\text{new}}(f) = \frac{F(f)\sqrt{S_G(f)}}{|F(f)|} \,. \tag{41}$$

After the Fourier back-transformation of $F_{\text{new}}(f)$, the new sequence (x_i^{new}) has the desired correlations (i. e. the desired γ), but the shape of the distribution has changed towards a (more or less) Gaussian distribution. In order to enforce the desired distribution, we exchange the (x_i^{new}) by the (x_i^{ref}), such that the largest value of the new set is replaced by the largest value of the reference set, the second largest of the new set by the second largest of the reference set and so on. After this the new sequence has the desired distribution and is clearly correlated. However, due to the exchange algorithm the perfect long-term correlations of the new data sequence were slightly altered again. So the procedure is repeated: the new sequence is Fourier transformed followed by spectrum adjustment, and the exchange algorithm is applied to the Fourier back-transformed data set. These steps are repeated several times, until the desired quality (or the best possible quality) of the spectrum of the new data series is achieved.

The Extended Binomial Multifractal Model

The multifractal cascade model [6,33,65] is a standard model for multifractal data, which is often applied, e. g., in hydrology [97]. In the model, a record x_i of length $N = 2^{n_{\max}}$ is constructed recursively as follows. In generation $n = 0$, the record elements are constant, i. e. $x_i = 1$ for all $i = 1, \ldots, N$. In the first step of the cascade (generation $n = 1$), the first half of the series is multiplied by a factor a and the second half of the series is multiplied by a factor b. This yields $x_i = a$ for $i = 1, \ldots, N/2$ and $x_i = b$ for $i = N/2 + 1, \ldots, N$. The parameters a and b are between zero and one, $0 < a < b < 1$. One need not restrict the model to $b = 1 - a$ as is often done in the literature [6]. In the second step (generation $n = 2$), we apply the process of step 1 to the two subseries, yielding $x_i = a^2$ for $i = 1, \ldots, N/4$, $x_i = ab$ for $i = N/4 + 1, \ldots, N/2$, $x_i = ba = ab$ for $i = N/2 + 1, \ldots, 3N/4$, and $x_i = b^2$ for $i = 3N/4 + 1, \ldots, N$. In general, in step $n + 1$, each subseries of step n is divided into two subseries of equal length, and the first half of the x_i is multiplied by a while the second half is multiplied by b. For example, in generation $n = 3$ the values in the eight subseries are $a^3, a^2b, a^2b, ab^2, a^2b, ab^2, ab^2, b^3$. After $n_{\max}$ steps, the final generation has been reached, where all subseries have length 1 and no more splitting is possible. We note that the final record can be written as $x_i = a^{n_{\max} - n(i-1)} b^{n(i-1)}$, where $n(i)$ is the number of digits 1 in the binary representation of the index i, e. g. $n(13) = 3$, since 13 corresponds to binary 1101.

For this multiplicative cascade model, the formula for $\tau(q)$ has been derived earlier [6,33,65]. The result is $\tau(q) = [-\ln(a^q + b^q) + q \ln(a + b)]/\ln 2$ or

$$h(q) = \frac{1}{q} - \frac{\ln(a^q + b^q)}{q \ln 2} + \frac{\ln(a + b)}{\ln 2} \,. \tag{42}$$

It is easy to see that $h(1) = 1$ for all values of a and b. Thus, in this form the model is limited to cases where $h(1)$, which is the exponent Hurst defined originally in the R/S method, is equal to one.

In order to generalize this multifractal cascade process such that any value of $h(1)$ is possible, one can subtract the offset $\Delta h = \ln(a + b)/\ln(2)$ from $h(q)$ [100]. The constant offset Δh corresponds to additional long-term correlations incorporated in the multiplicative cascade model. For generating records without this offset, we rescale the power spectrum. First, we transform (FFT) the simple multiplicative cascade data into the frequency domain. Then, we multiply all Fourier coefficients by $f^{-\Delta h}$, where f is the frequency. This way, the slope β of the power spectra $S(f) \sim f^{-\beta}$ is decreased from $\beta = 2h(2) - 1 = [2\ln(a + b) - \ln(a^2 + b^2)]/\ln 2$ into $\beta' = 2[h(2) - \Delta h] - 1 = -\ln(a^2 + b^2)/\ln 2$. Finally, backward FFT is employed to transform the signal back into the time domain.

The Bi-fractal Model

In some cases a simple bi-fractal model is already sufficient for modeling apparently multifractal data [101]. For bi-fractal records the Renyi exponents $\tau(q)$ are characterized by two distinct slopes α_1 and α_2,

$$\tau(q) = \begin{cases} q\alpha_1 - 1 & q \le q_\times \\ q\alpha_2 + q_\times(\alpha_1 - \alpha_2) - 1 & q > q_\times \end{cases} \tag{43}$$

or

$$\tau(q) = \begin{cases} q\alpha_1 + q_\times(\alpha_2 - \alpha_1) - 1 & q \le q_\times \\ q\alpha_2 - 1 & q > q_\times \end{cases} \,. \tag{44}$$

If this behavior is translated into the $h(q)$ picture using Eq. (29), we obtain that $h(q)$ exhibits a plateau from $q = -\infty$ up to a certain $q_\times$ and decays hyperbolically for $q > q_\times$,

$$h(q) = \begin{cases} \alpha_1 & q \le q_\times \\ q_\times(\alpha_1 - \alpha_2)\frac{1}{q} + \alpha_2 & q > q_\times \end{cases} \,, \tag{45}$$

or vice versa,

$$h(q) = \begin{cases} q_\times(\alpha_2 - \alpha_1)\frac{1}{q} + \alpha_1 & q \le q_\times \\ \alpha_2 & q > q_\times \end{cases} \,. \tag{46}$$

Both versions of this bi-fractal model require three parameters. The multifractal spectrum is degenerated to two single points, thus its width can be defined as $\Delta\alpha = \alpha_1 - \alpha_2$.

Future Directions

The most straightforward future direction is to analyze more types of time series from other complex systems than those listed in Sect. "Introduction" to check for the presence of fractal scaling and in particular long-term correlations. Such applications may include (i) data that are not recorded as a function of time but as a function of another parameter and (ii) higher dimensional data. In particular, the inter-relationship between fractal time series and spatially fractal structures can be studied. Studies of *fields* with fractal scaling in time and space have already been performed in Geophysics. In some cases studying new types of data will require dealing with more difficult types of non-stationarities and transient behavior, making further development of the methods necessary. In many studies, detrending methods have not been applied yet. However, discovering fractal scaling in more and more systems cannot be an aim on its own.

Up to now, the reasons for observed fractal or multifractal scaling are not clear in most applications. It is thus highly desirable to study causes for fractal and multifractal correlations in time series, which is a difficult task, of course. One approach might be based on modeling and comparing the fractal aspects of real and modeled time series by applying the methods described in this article. The fractal or multifractal characterization can thus be helpful in improving the models. For many applications, practically usable models which display fractal or transient fractal scaling still have to be developed. One example for a model explaining fractal scaling might be a precipitation, storage and runoff model, in which the fractal scaling of runoff time series could be explained by fractional integration of rainfall in soil, groundwater reservoirs, or river networks characterized by a fractal structure. Also studies regarding the inter-relationship between fractal scaling and complex networks, representing the structure of a complex system, are desirable. This way one could gain an interpretation of the causes for fractal behavior.

Another direction of future research is regarding the linear and especially non-linear inter-relationships between several time series. There is great need for improved methods characterizing cross-correlations and similar statistical inter-relationships between several non-stationary time series. Most methods available so far are reserved to stationary data, which is, however, hardly found in natural recordings. An even more ambitious aim is the (time-dependent) characterization of a larger network of signals. In such a network, the signals themselves would represent the nodes, while the (possibly directed) inter-relationships between each pair represent the links (or bonds) between the nodes. The properties of both nodes and links can vary with time or change abruptly, when the represented complex system goes through a phase transition.

Finally, more work will have to be invested in studying the practical consequences of fractal scaling in time series. Studies should particularly focus on predictions of future values and behavior of time series and whole complex systems. This is very relevant, not only in hydrology and climate research, where a clear distinguishing of trends and natural fluctuations is crucial, but also for predicting dangerous medical events on-line in patients based on the continuous recording of time series.

Acknowledgment

We thank Ronny Bartsch, Amir Bashan, Mikhail Bogachev, Armin Bunde, Jan Eichner, Shlomo Havlin, Diego Rybski, Aicko Schumann, and Stephan Zschiegner for helpful discussions and contribution. This work has been supported by the Deutsche Forschungsgemeinschaft (grant KA 1676/3) and the European Union (STREP project DAPHNet, grant 018474-2).

Bibliography

1. Mandelbrot BB, van Ness JW (1968) Fractional Brownian motions, fractional noises and applications. SIAM Review 10: 422
2. Mandelbrot BB, Wallis JR (1969) Some long-run properties of geophysical records. Water Resour Res 5:321–340
3. Mandelbrot BB (1999) Multifractals and 1/f noise: wild self-affinity in physics. Springer, Berlin
4. Hurst HE (1951) Long-term storage capacity of reservoirs. Tran Amer Soc Civ Eng 116:770
5. Hurst HE, Black RP, Simaika YM (1965) Long-term storage: an experimental study. Constable, London
6. Feder J (1988) Fractals. Plenum Press, New York
7. Barnsley MF (1993) Fractals everywhere. Academic Press, San Diego
8. Bunde A, Havlin S (1994) Fractals in science. Springer, Berlin
9. Jorgenssen PET (2000) Analysis and probability: Wavelets, signals, fractals. Springer, Berlin
10. Bunde A, Kropp J, Schellnhuber HJ (2002) The science of disasters – climate disruptions, heart attacks, and market crashes. Springer, Berlin
11. Kantz H, Schreiber T (2003) Nonlinear time series analysis. Cambridge University Press, Cambridge
12. Peitgen HO, Jürgens H, Saupe D (2004) Chaos and fractals. Springer, Berlin
13. Sornette D (2004) Critical phenomena in natural sciences. Springer, Berlin
14. Peng CK, Mietus J, Hausdorff JM, Havlin S, Stanley HE, Goldberger AL (1993) Long-range anti-correlations and non-Gaussian behaviour of the heartbeat. Phys Rev Lett 70: 1343

15. Bunde A, Havlin S, Kantelhardt JW, Penzel T, Peter JH, Voigt K (2000) Correlated and uncorrelated regions in heart-rate fluctuations during sleep. Phys Rev Lett 85:3736
16. Vyushin D, Zhidkov I, Havlin S, Bunde A, Brenner S (2004) Volcanic forcing improves atmosphere-ocean coupled general circulation model scaling performance. Geophys Res Lett 31:L10206
17. Koscielny-Bunde E, Bunde A, Havlin S, Roman HE, Goldreich Y, Schellnhuber HJ (1998) Indication of a universal persistence law governing atmospheric variability. Phys Rev Lett 81:729
18. Box GEP, Jenkins GM, Reinsel GC (1994) Time-series analysis. Prentice Hall, New Jersey
19. Chatfield C (2003) The analysis of time series. An introduction. Taylor & Francis, Boca Raton
20. Schmitt DT, Schulz M (2006) Analyzing memory effects of complex systems from time series. Phys Rev E 73:056204
21. Taqqu MS, Teverovsky V, Willinger W (1995) Estimators for long-range dependence: An empirical study. Fractals 3:785
22. Delignieresa D, Ramdania S, Lemoinea L, Torrea K, Fortesb M, Ninot G (2006) Fractal analyses for 'short' time series: A reassessment of classical methods. J Math Psychol 50:525
23. Mielniczuk J, Wojdyllo P (2007) Estimation of Hurst exponent revisited. Comp Stat Data Anal 51:4510
24. Hunt GA (1951) Random Fourier transforms. Trans Amer Math Soc 71:38
25. Rangarajan G, Ding M (2000) Integrated approach to the assessment of long range correlation in time series data. Phys Rev E 61:4991
26. Peng CK, Buldyrev SV, Goldberger AL, Havlin S, Sciortino F, Simons M, Stanley HE (1992) Long-range correlations in nucleotide sequences. Nature 356:168
27. Goupillaud P, Grossmann A, Morlet J (1984) Cycle-octave and related transforms in seismic signal analysis. Geoexploration 23:85
28. Daubechies I (1988) Orthogonal bases of compactly supported wavelets. Commun Pure Appl Math 41:909
29. Bogachev M, Schumann AY, Kantelhardt JW, Bunde A (2009) On distinguishing long-term and short-term memory in finite data. Physica A, to be published
30. Kantelhardt JW, Roman HE, Greiner M (1995) Discrete wavelet approach to multifractality. Physica A 220:219
31. Peng C-K, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL (1994) Mosaic organization of DNA nucleotides. Phys Rev E 49:1685
32. Ashkenazy Y, Ivanov PC, Havlin S, Peng CK, Goldberger AL, Stanley HE (2001) Magnitude and sign correlations in heartbeat fluctuations. Phys Rev Lett 86:1900
33. Kantelhardt JW, Zschiegner SA, Bunde A, Havlin S, Koscielny-Bunde E, Stanley HE (2002) Multifractal detrended fluctuation analysis of non-stationary time series. Physica A 316:87
34. Gu GF, Zhou WX (2006) Detrended fluctuation analysis for fractals and multifractals in higher dimensions. Phys Rev E 74:061104
35. Kantelhardt JW, Koscielny-Bunde E, Rego HHA, Havlin S, Bunde A (2001) Detecting long-range correlations with detrended fluctuation analysis. Physica A 295:441
36. Hu K, Ivanov PC, Chen Z, Carpena P, Stanley HE (2001) Effect of trends on detrended fluctuation analysis. Phys Rev E 64:011114
37. Chen Z, Ivanov PC, Hu K, Stanley HE (2002) Effect of nonstationarities on detrended fluctuation analysis. Phys Rev E 65:041107
38. Chen Z, Hu K, Carpena P, Bernaola-Galvan P, Stanley HE, Ivanov PC (2005) Effect of nonlinear filters on detrended fluctuation analysis. Phys Rev E 71:011104
39. Grau-Carles P (2006) Bootstrap testing for detrended fluctuation analysis. Physics A 360:89
40. Nagarajan R (2006) Effect of coarse-graining on detrended fluctuation analysis. Physica A 363:226
41. Heneghan C, McDarby G (2000) Establishing the relation between detrended fluctuation analysis and power spectral density analysis for stochastic processes. Phys Rev E 62: 6103
42. Weron R (2002) Estimating long-range dependence: finite sample properties and confidence intervals. Physica A 312:285
43. Bashan A, Bartsch R, Kantelhardt JW, Havlin S (2008) Comparison of detrending methods for fluctuation analysis. Physica A 387:580
44. Bahar S, Kantelhardt JW, Neiman A, Rego HHA, Russell DF, Wilkens L, Bunde A, Moss F (2001) Long range temporal anticorrelations in paddlefish electro-receptors. Europhys Lett 56:454
45. Bartsch R, Henning T, Heinen A, Heinrichs S, Maass P (2005) Statistical analysis of fluctuations in the ECG morphology. Physica A 354:415
46. Santhanam MS, Bandyopadhyay JN, Angom D (2006) Quantum spectrum as a time series: fluctuation measures. Phys Rev E 73:015201
47. Ashkenazy Y, Havlin S, Ivanov PC, Peng CK, Schulte-Frohlinde V, Stanley HE (2003) Magnitude and sign scaling in power-law correlated time series. Physica A 323:19
48. Kalisky T, Ashkenazy Y, Havlin S (2005) Volatility of linear and nonlinear time series. Phys Rev E 72:011913
49. Mantegna RN, Stanley HE (2000) An introduction to econophysics – correlations and complexity in finance. Cambridge Univ Press, Cambridge
50. Bouchaud JP, Potters M (2003) Theory of financial risks: from statistical physics to risk management. Cambridge Univ Press, Cambridge
51. Alessio E, Carbone A, Castelli G, Frappietro V (2002) Second-order moving average and scaling of stochastic time series. Europ Phys J B 27:197
52. Carbone A, Castelli G, Stanley HE (2004) Analysis of clusters formed by the moving average of a long-range correlated time series. Phys Rev E 69:026105
53. Carbone A, Castelli G, Stanley HE (2004) Time-dependent Hurst exponent in financial time series. Physica A 344:267
54. Alvarez-Ramirez J, Rodriguez E, Echeverría JC (2005) Detrending fluctuation analysis based on moving average filtering. Physica A 354:199
55. Kiyono K, Struzik ZR, Aoyagi N, Togo F, Yamamoto Y (2005) Phase transition in a healthy human heart rate. Phys Rev Lett 95:058101
56. Staudacher M, Telser S, Amann A, Hinterhuber H, Ritsch-Marte M (2005) A new method for change-point detection developed for on-line analysis of the heart beat variability during sleep. Physica A349:582
57. Telser S, Staudacher M, Hennig B, Ploner Y, Amann A, Hinter-

huber H, Ritsch-Marte M (2007) Temporally resolved fluctuation analysis of sleep-ECG. J Biol Phys 33:190
58. Chianca CV, Ticona A, Penna TJP (2005) Fourier-detrended fluctuation analysis. Physica A 357:447
59. Jánosi IM, Müller R (2005) Empirical mode decomposition and correlation properties of long daily ozone records. Phys Rev E 71:056126
60. Nagarajan R, Kavasseri RG (2005) Minimizing the effect of trends on detrended fluctuation analysis of long-range correlated noise. Physica A 354:182
61. Nagarajan R (2006) Reliable scaling exponent estimation of long-range correlated noise in the presence of random spikes. Physica A 366:1
62. Rodriguez E, Echeverria JC, Alvarez-Ramirez J (2007) Detrending fluctuation analysis based on high-pass filtering. Physica A 375:699
63. Grech D, Mazur Z (2005) Statistical properties of old and new techniques in detrended analysis of time series. Acta Phys Pol B 36:2403
64. Xu L, Ivanov PC, Hu K, Chen Z, Carbone A, Stanley HE (2005) Quantifying signals with power-law correlations: a comparative study of detrended fluctuation analysis and detrended moving average techniques. Phys Rev E 71:051101
65. Barabási AL, Vicsek T (1991) Multifractality of self-affine fractals. Phys Rev A 44:2730
66. Bacry E, Delour J, Muzy JF (2001) Multifractal random walk. Phys Rev E 64:026103
67. Muzy JF, Bacry E, Arneodo A (1991) Wavelets and multifractal formalism for singular signals: Application to turbulence data. Phys Rev Lett 67:3515
68. Muzy JF, Bacry E, Arneodo A (1994) The multifractal formalism revisited with wavelets. Int J Bifurcat Chaos 4:245
69. Arneodo A, Bacry E, Graves PV, Muzy JF (1995) Characterizing long-range correlations in DNA sequences from wavelet analysis. Phys Rev Lett 74:3293
70. Arneodo A, Manneville S, Muzy JF (1998) Towards log-normal statistics in high Reynolds number turbulence. Eur Phys J B 1:129
71. Arneodo A, Audit B, Decoster N, Muzy JF, Vaillant C (2002) Wavelet based multifractal formalism: applications to DNA sequences, satellite images of the cloud structure, and stock market data. In: Bunde A, Kropp J, Schellnhuber HJ (eds) The science of disaster: climate disruptions, market crashes, and heart attacks. Springer, Berlin
72. Kantelhardt JW, Rybski D, Zschiegner SA, Braun P, Koscielny-Bunde E, Livina V, Havlin S, Bunde A (2003) Multifractality of river runoff and precipitation: comparison of fluctuation analysis and wavelet methods. Physica A 330:240
73. Oswiecimka P, Kwapien J, Drozdz S (2006) Wavelet versus detrended fluctuation analysis of multifractal structures. Phys Rev E 74:016103
74. Ivanov PC, Amaral LAN, Goldberger AL, Havlin S, Rosenblum MG, Struzik ZR, Stanley HE (1999) Multifractality in human heartbeat dynamics. Nature 399:461
75. Amaral LAN, Ivanov PC, Aoyagi N, Hidaka I, Tomono S, Goldberger AL, Stanley HE, Yamamoto Y (2001) Behavioral-independence features of complex heartbeat dynamics. Phys Rev Lett 86:6026
76. Bogachev M, Schumann AY, Kantelhardt JW (2008) (in preparation)
77. Bunde A, Eichner JF, Kantelhardt JW, Havlin S (2005) Long-term memory: A natural mechanism for the clustering of extreme events and anomalous residual times in climate records. Phys Rev Lett 94:048701
78. Bunde A, Eichner JF, Kantelhardt JW, Havlin S (2003) The effect of long-term correlations on the return periods of rare events. Physica A 330:1
79. Altmann EG, Kantz H (2005) Recurrence time analysis, long-term correlations, and extreme events. Phys Rev E 71: 056106
80. Eichner JF, Kantelhardt JW, Bunde A, Havlin S (2007) Statistics of return intervals in long-term correlated records. Phys Rev E 75:011128
81. Eichner JF, Kantelhardt JW, Bunde A, Havlin S (2006) Extreme value statistics in records with long-term persistence. Phys Rev E 73:016130
82. Bogachev MI, Eichner JF, Bunde A (2007) Effect of nonlinear correlations on the statistics of return intervals in multifractal data sets. Phys Rev Lett 99:240601
83. Storch HV, Zwiers FW (2001) Statistical analysis in climate research. Cambridge Univ Press, Cambridge
84. Newell GF, Rosenblatt M (1962) Ann Math Statist 33:1306
85. Sornette D, Knopoff L (1997) The paradox of the expected time until the next earthquake. Bull Seism Soc Am 87:789
86. Fisher RA, Tippett LHC (1928) Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proc Camb Phi Soc 24:180
87. Gumbel EJ (1958) Statistics of extremes. Columbia University Press, New York
88. Galambos J (1978) The asymptotic theory of extreme order statistics. Wiley, New York
89. Leadbetter MR, Lindgren G, Rootzen H (1983) Extremes and related properties of random sequences and processes. Springer, New York
90. Galambos J, Lechner J, Simin E (1994) Extreme value theory and applications. Kluwer, Dordrecht
91. te Chow V (1964) Handbook of applied hydrology. McGraw-Hill, New York
92. Raudkivi AJ (1979) Hydrology. Pergamon Press, Oxford
93. Rasmussen PF, Gautam N (2003) Alternative PWM-estimators of the Gumbel distribution. J Hydrol 280:265
94. Mandelbrot BB (1971) A fast fractional Gaussian noise generator. Water Resour Res 7:543
95. Voss RF (1985) In: Earnshaw RA (ed) Fundamental algorithms in computer graphics. Springer, Berlin
96. Makse HA, Havlin S, Schwartz M, Stanley HE (1996) Method for generating long-range correlations for large systems. Phys Rev E 53:5445
97. Rodriguez-Iturbe I, Rinaldo A (1997) Fractal river basins – change and self-organization. Cambridge Univ Press, Cambridge
98. Schreiber T, Schmitz A (1996) Improved surrogate data for nonlinearity tests. Phys Rev Lett 77:635
99. Schreiber T, Schmitz A (2000) Surrogate time series. Physica D 142:346
100. Koscielny-Bunde E, Kantelhardt JW, Braun P, Bunde A, Havlin S (2006) Long-term persistence and multifractality of river runoff records. J Hydrol 322:120
101. Kantelhardt JW, Koscielny-Bunde E, Rybski D, Braun P, Bunde A, Havlin S (2006) Long-term persistence and multifractality of precipitation and river runoff records. J Geophys Res Atmosph 111:D01106

Fractals in Biology

SERGEY V. BULDYREV
Department of Physics, Yeshiva University,
New York, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Self-similar Branching Structures
Fractal Metabolic Rates
Physical Models of Biological Fractals
Diffusion Limited Aggregation and Bacterial Colonies
Measuring Fractal Dimension of Real Biological Fractals
Percolation and Forest Fires
Critical Point and Long-Range Correlations
Lévy Flight Foraging
Dynamic Fractals
Fractals and Time Series
SOC and Biological Evolution
Fractal Features of DNA Sequences
Future Directions
Bibliography

Glossary

Allometric laws An allometric law describes the relationship between two attributes of living organisms y and x, and is usually expressed as a power-law: $y \sim x^{\alpha}$, where α is the scaling exponent of the law. For example, x can represent total body mass M and y can represent the mass of a brain m_b. In this case $m_b \sim M^{3/4}$. Another example of an allometric law: $B \sim M^{3/4}$ where B is metabolic rate and M is body mass. Allometric laws can be also found in ecology: the number of different species N found in a habitat of area A scales as $N \sim A^{1/4}$.

Radial distribution function Radial distribution function $g(r)$ describes how the average density of points of a set behaves as function of distance r from a point of this set. For an empirical set of N data points, the distances between all pair of points are computed and the number of pairs $N_p(r)$ such that their distance is less then r is found. Then $M(r) = 2N_p(r)/N$ gives the average number of the neighbors (mass) of the set within a distance r. For a certain distance bin $r_1 < r < r_2$, we define $g[(r_2 + r_1)/2] = [M(r_2) - M(r_1)]/[V_d(r_2) - V_d(r_1)]$, where $V_d(r) = 2\pi^{d/2} r^d/[d\Gamma(d/2)]$ is the volume/area/length of a d-dimensional sphere/circle/interval of radius r.

Fractal set We define a fractal set with the fractal dimension $0 < d_f < d$ as a set for which $M(r) \sim r^{d_f}$ for $r \to \infty$. Accordingly, for such a set $g(r)$ decreases as a power law of the distance $g(r) \sim r^{-\chi}$, where $\chi = d - d_f$.

Correlation function For a superposition of a fractal set and a set with a finite density defined as $\rho = \lim_{r\to\infty} M(r)/V_d(r)$, the correlation function is defined as $h(r) \equiv g(r) - \rho$.

Long-range power law correlations The set of points has long-range power law correlations (LRPLC) if $h(r) \sim r^{-\chi}$ for $r \to \infty$ with $0 < \chi < d$. LRPLC indicate the presence of a fractal set with fractal dimension $d_f = d - \chi$ superposed with a uniform set.

Critical point Critical point is defined as a point in the system parameter space (e. g. temperature, $T = T_c$, and pressure $P = P_c$), near which the system acquires LRPLC

$$h(r) \sim \frac{1}{r^{d-2+\eta}} \exp(r/\xi) , \tag{1}$$

where ξ is the correlation length which diverges near the critical point as $\sim |T - T_c|^{-\nu}$. Here $\eta > 0$ and $\nu > 0$ are critical exponents which depend on the few system characteristics such as dimensionality of space. Accordingly, the system is characterized by fractal density fluctuations with $d_f = 2 - \eta$.

Self-organized criticality Self-organized criticality (SOC) is a term which describes a system for which the critical behavior characterized by a large correlation length is achieved for a wide range of parameters and thus does not require special tuning. This usually occurs when a critical point corresponds to an infinite value of a system parameter, such as a ratio of the characteristic time of the stress build up and a characteristic time of the stress release.

Morphogenesis Morphogenesis is a branch of developmental biology concerned with the shapes of organs and the entire organisms. Several types of molecules are particularly important during morphogenesis. Morphogens are soluble molecules that can diffuse and carry signals that control cell differentiation decisions in a concentration-dependent fashion. Morphogens typically act through binding to specific protein receptors. An important class of molecules involved in morphogenesis are transcription factor proteins that determine the fate of cells by interacting with DNA. The morphogenesis of the branching fractal-like structures such as lungs involves a dozen of morpho-

genes. The mechanism for keeping self-similarity of the branches at different levels of branching hierarchy is not yet fully understood. The experiments with transgenic mice with certain genes knocked-out produce mice without limbs and lungs or without terminal buds.

Definition of the Subject

Fractals occur in a wide range of biological applications:

1) In morphology when the shape of an organism (tree) or an organ (vertebrate lung) has a self-similar brunching structure which can be approximated by a fractal set (Sect. "Self-Similar Branching Structures").
2) In allometry when the allometric power laws can be deduced from the fractal nature of the circulatory system (Sect. "Fractal Metabolic Rates").
3) In ecology when a colony or a habitat acquire fractal shapes due to some SOC processes such as diffusion limited aggregation (DLA) or percolation which describes forest fires (Sects. "Physical Models of Biological Fractals"–"Percolation and Forest Fires").
4) In epidemiology when some of the features of the epidemics is described by percolation which in turn leads to fractal behavior (Sect. "Percolation and Forest Fires").
5) In behavioral sciences, when a trajectory of foraging animal acquires fractal features (Sect. "Lévy Flight Foraging").
6) In population dynamics, when the population size fluctuates chaotically (Sect. "Dynamic Fractals").
7) In physiology, when time series have LRPLC (Sect. "Fractals and Time Series").
8) In evolution theory, which may have some features described by SOC (Sect. "SOC and Biological Evolution").
9) In bioinformatics when a DNA sequence has a LRPLC or a network describing protein interactions has a self-similar fractal behavior (Sect. "Fractal Features of DNA Sequences").

Fractal geometry along with Euclidian geometry became a part of general culture which any scientist must be familiar with. Fractals often originate in the theory of complex systems describing the behavior of many interacting elements and therefore have a great number of biological applications. Complex systems have a general tendency for self-organization and complex pattern formation. Some of these patterns have certain nontrivial symmetries, for example fractals are characterized by scale invariance i. e. they look similarly on different magnification. Fractals are characterized by their fractal dimension, which is specific for each model and therefore may shed light on the origin of a particular biological phenomenon. In Sect. "Diffusion Limited Aggregation and Bacterial Colonies", we discuss the techniques for measuring fractal dimension and their limitations.

Introduction

The fact that simple objects of Euclidian geometry such as straight lines, circles, cubes, and spheres are not sufficient to describe complex biological shapes has been known for centuries. Physicists were always accused by biologists for introducing a "spherical cow". Nevertheless people from antiquity to our days were fascinated by finding simple mathematical regularities which can describe the anatomy and physiology of leaving creatures. Five centuries ago, Leonardo da Vinci observed that "the branches of a tree at every stage of its height when put together are equal in thickness to the trunk" [107]. Another famous but still poorly understood phenomenon is the emergence of the Fibonacci numbers in certain types of pine cones and composite flowers [31,134].

In the middle of the seventies a new concept of fractal geometry was introduced by Mandelbrot [79]. This concept was readily accepted for analysis of the complex shapes in the biological world. However, after initial splash of enthusiasm, [14,40,41,74,75,76,88,122] the application of fractals in biology significantly dwindled and today the general consensus is that a "fractal cow" is often not much better than a "spherical cow". Nature is always more complex than mathematical abstractions.

Strictly speaking the fractal is an object whose mass, M, grows as a fractional power law of its linear dimension, L,

$$M \sim L^{d_f} , \tag{2}$$

where d_f is a non-integer quantity called fractal dimension. Simple examples of fractal objects of various fractal dimensions are given by iterative self-similar constructs such as a Cantor set, a Koch curve, a Sierpinski gasket, and a Menger sponge [94]. In all this constructs a next iteration of the object is created by arranging p exact copies of the previous iteration of the object (Fig. 1) in such a way that the linear size of the next iteration is q times larger than the linear size of the previous iteration. Thus the mass, M_n, and the length, L_n, of the nth iteration scale as

$$\begin{aligned} M_n &= p^n M_0 \\ L_n &= q^n L_0 , \end{aligned} \tag{3}$$

where M_0 and L_0 are mass and length of the zero order

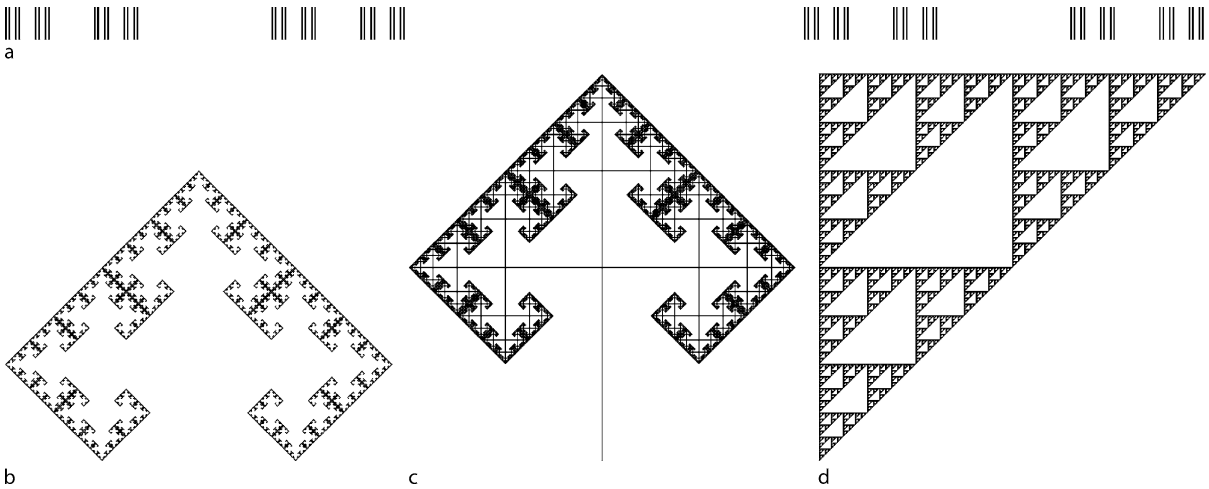

Fractals in Biology, Figure 1

a Cantor set ($p = 2$, $q = 3$, $d_f = \ln 2/\ln 3 \approx 0.63$) is an example of fractal "dust" with fractal dimension less then 1; b Fractal tree ($p = 3$, $q = 2$, $d_f = \ln 3/\ln 2 \approx 1.58$) with branches removed so that only the terminal points (leaves) can be seen. c The same tree with the trunks added. Both trees are produced by recursive combining of the three smaller trees: one serving as the top of the tree and the other two rotated by 90° clockwise and counter-clockwise serving as two brunches joined with the top at the middle. In c a vertical segment representing a trunk of the length equal to the diagonal of the branches is added at each recursive step. Mathematically, the fractal dimensions of sets b and c are the same, because the mass of the tree with trunks (number of black pixels) for the system of linear size 2^n grows as $3^{n+1} - 2^{n+1}$, while for the tree without trunks mass scales simply as 3^n. In the limit $n \to \infty$ this leads to the same fractal dimension ln 3/ ln 2. However, visual inspection suggests that the tree with branches has a larger fractal dimension. This is in accord with the box counting method, which produces a higher value of the slope for the finite tree with the trunks. The slope slowly converges to the theoretical value as the number of recursive steps increases. d Sierpinski gasket has the same fractal dimension as the fractal tree ($p = 3$, $q = 2$) but has totally different topology and visual appearance than the tree

iteration. Excluding n from Eq. (3), we get

$$M_n = M_0 \left(\frac{L_n}{L_0}\right)^{d_f} , \tag{4}$$

where

$$d_f = \ln p/\ln q \tag{5}$$

can be identified as fractal dimension. The above-described objects have the property of self-similarity, the previous iteration magnified q times looks exactly like the next iteration once we neglect coarse-graining on the lowest iteration level, which can be assumed to be infinitely small. An interesting feature of such fractals is the power law distribution of their parts. For example, the cumulative distribution of distances L between the points of Cantor set and the branch lengths of the fractal tree (Fig. 1) follows a power law:

$$P(L > x) \sim x^{-d_f} . \tag{6}$$

Thus the emergence of power law distributions and other power law dependencies is often associated with fractals and often these power law regularities not necessarily related to geometrical fractals are loosely referred as fractal properties.

Self-similar Branching Structures

Perfect deterministic fractals described above are never observed in nature, where all shapes are subjects to random variations. The natural examples closest to the deterministic fractals are structures of trees, certain plants such as a cauliflower, lungs and a cardiovascular system. The control of branching morphogenesis [85,114,138] involves determining when and where a branch will occur, how long the tube grows before branching again, and at what angle the branch will form. The development of different organs (such as salivary gland, mammary gland, kidney, and lung) creates branching patterns easily distinguished from each other. Moreover, during the development of a particular organ, the form of branching often changes, depending on the place or time when the branching occurs. Morphogenesis is controlled by complex molecular interactions of morphogenes.

Let us illustrate the idea of calculating fractal dimension of a branching object such as a cauliflower [51,64]. If we assume (which is not quite correct, see Fig. 2) that each branch of a cauliflower give rise to exactly p branches of the next generation exactly q times smaller than the original branch applying Eq. (5), we get $d_f = \ln q/\ln p$. Note

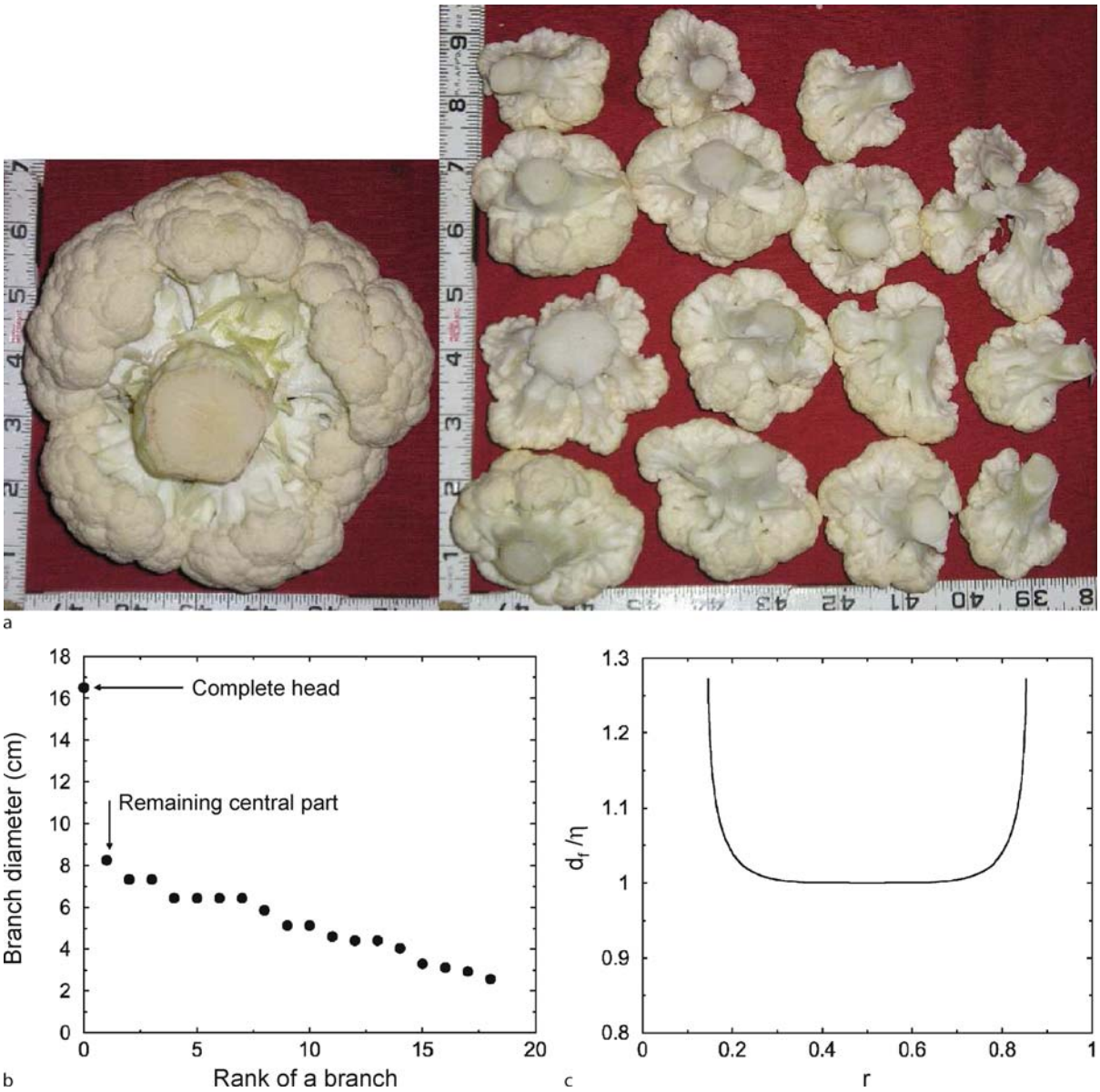

Fractals in Biology, Figure 2

a Cauliflower anatomy. A complete head of a cauliflower (*left*) and after it is taken apart (*right*). We remove 17 branches until the diameter of the remaining part becomes half of the diameter of the original head. b Diameters of the branches presented in a. c Estimation of the fractal dimension for asymmetric lungs ($\eta = 3$) and trees ($\eta = 2$) with branching parameter r using Eq. (7). For $r > 0.855$ or $r < 0.145$ the fractal dimension is not defined due to the "infrared catastrophe": the number of large branches diverge in the limit of an infinite tree

that this is not the fractal dimension of the cauliflower itself, but of its skeleton in which each brunch is represented by the elements of its daughter branches, with an addition of a straight line connecting the daughter branches as in the example of a fractal tree (Fig. 1b, c). Because this addition does not change the fractal dimension formula, the fractal dimension of the skeleton is equal to the fractal dimension of the surface of a cauliflower, which can be represented as the set of the terminal branches. As a physical object, the cauliflower is not a fractal but a three-dimensional body so that the mass of a branch of length L scales as L^3. This is because the diameter of each branch is proportional to the length of the branch.

The distribution of lengths of the collection of all the branches of a cauliflower is also a power law given by a simple idea that there are $N_n = \sum_{k=0}^{n-1} p^k \approx p^n/(p-1)$ branches of length larger than $L_n = L_0 q^{-n}$. Excluding n gives $N_n(L > L_n) \sim L_n^{-d_f}$. In reality, however (Fig. 2a, b), the branching structure of a cauliflower is more complex. Simple measurements show that there are about $p = 18$ branches of sizes $L_k = L_0 \approx L_0[0.5 - 0.2(k-1)]$ for $k = 1, 2, \ldots, p$, where L_0 is the diameter of the complete head of the cauliflower. (We keep removing branches until the diameter of the central remaining part is equal to the half diameter of the complete head and we assume that it is similar to the rest of the branches). Subsequent generations (we count at least eight) obey similar rules, however p has a tendency to decrease with the generation number. To find fractal dimension of a cauliflower we will count the number of branches $N(L)$ larger than certain size L. As in case of equal branches, this number scales as L^{-d_f}. Calculations, similar to those presented in [77] show that in this case the fractal dimension is equal to

$$d_f = -\frac{\mu}{\sigma^2}\left(\sqrt{1 - \ln p \frac{2\sigma^2}{\mu^2}} - 1\right) , \tag{7}$$

where μ and σ are the mean and the standard deviation of $\ln q_k \equiv \ln(L_0/L_k)$. For the particular cauliflower shown in Fig. 2a the measurements presented in Fig. 2b give $d_f = 2.75$, which is very close to the estimate $d_f = 2.8$, of [64]. The physiological reason for such a peculiar branching pattern of cauliflower is not known. It is probably designed to store energy for a quick production of a dense inflorescence.

For a tree, [142] the sum of the cross-section areas of the two daughter branches according to Leonardo is equal to the cross-section area of the mother branch. This can be understood because the number of capillary bundles going from the mother to the daughters is conserved. Accordingly, we have a relation between the daughter branch diameters d_1 and d_2 and the mother branch diameter d_0,

$$d_1^\eta + d_2^\eta = d_0^\eta , \tag{8}$$

where $\eta = 2$. We can assume that the branch diameter of the largest daughter branch scales as $d_1^\eta = r d_0^\eta$, where r is asymmetry ratio maintaining for each generation of branches. In case of a symmetric tree $d_1 = d_2$, $r = 1/2$. If we assume that the branch length $L = sd$, where s is a constant which is the same for any branch of the tree, then we can relate our tree to a fractal model with $p = 2$ and $q = 2^{1/\eta}$. Using Eq. (5) we get $d_f = \eta = 2$, i. e. the tree skeleton or the surface of the terminal branches is an object of fractal dimension two embedded in the three-dimensional space. This is quite natural because all the leaves whose number is proportional to the number of terminal branches must be exposed to the sunlight and the easiest way to achieve this is to place all the leaves on the surface of a sphere which is a two-dimensional object.

For an asymmetric tree the fractal dimension can be computed using Eq. (7) with

$$\mu = |\ln[r(1-r)]/2\eta| \tag{9}$$

and

$$\sigma = |\ln[r/(1-r)]/2\eta| . \tag{10}$$

This value is slightly larger than 2 for a wide range of r (Fig. 2c). This property may be related to a tendency of a tree to maximize the surface of it leaves, which must not be necessarily exposed to the direct sunlight but can suffice on the light reflected by the outer leaves.

For a lung [56,66,77,102,117], the flow is not a capillary but can be assumed viscous. According to flow conservation, the sum of the air flows of the daughter branches must be equal to the flow of the mother branch:

$$Q_1 + Q_2 = Q_0 . \tag{11}$$

For the Poiseuille flow, $Q \sim \Delta P d^4/L$, where ΔP is the pressure drop, which is supposed to be the same for the airways of all sizes. Assuming that the lung maintains the ratio $s = L/d$ in all generations of branches, we conclude that the diameters of the daughter and mother branches must satisfy Eq. (8) with $\eta = 4 - 1 = 3$. Accordingly, for a symmetrically branching lung $d_f = \eta = 3$, which means that the surface of the alveoli which is proportional to the number of terminal airways scales as L^3, i. e. it is a space filling object. Again this prediction is quite reasonable because nature tends to maximize the gas exchange area so that it completely fills the volume of the lung.

In reality, the flow in the large airways is turbulent, and the parameter η of lungs of different species varies between 2 and 3 [77]. Also the lungs are known to be asymmetric and the ratio $r = Q_1/Q_0 \neq 1/2$ changes from one generation of the airways to the next [77]. However, Eq. (7) with μ and σ given by Eqs. (9) and (10) shows that the fractal dimension of an asymmetric tree remains very close to 3 for a wide range of $0.146 < r < 0.854$ (Fig. 2c). The fact that the estimated fractal dimension is slightly larger than 3 does not contradict common sense because the branching stops when the airway diameter becomes smaller than some critical cutoff. Other implications of the lung asymmetry are discussed in [77]. An interesting idea

was proposed in [117], according to which the dependence of the branching pattern on the generation of the airways can be derived from optimization principles and give rise to the complex value of the fractal dimension.

An interesting property of the crackle sound produced by diseased lungs is that during inflation the resistance to airflow of the small airways decreases in discrete jumps. Airways do not open individually, but in a sequence of bursts or avalanches involving many airways; both the size of these jumps and the time intervals between jumps follow power-law distributions [128]. These avalanches are not related to the SOC as one might expect, but their power law distributions directly follow from the branching structure of lungs (see [3] and references therein).

Fractal Metabolic Rates

An important question in biology is the scaling of the metabolic rate with respect to the body mass (Kleiber's Law). It turns out that for almost all species of animals, which differ in terms of mass by 21 orders of magnitudes, the metabolic rate B scales as $B \sim M^{3/4}$ [28,112,140]. The scatter plot $\ln B$ vs $\ln M$ is a narrow cloud concentrated around the straight line with a slope 3/4 (Fig. 3). This is one of the many examples of the allometric laws which describe the dependence of various biological parameters on the body mass or population size [1,9,87,105,106].

A simple argument based on the idea that the thermal energy loss and hence the metabolic rate should be proportional to the surface area predicts however that

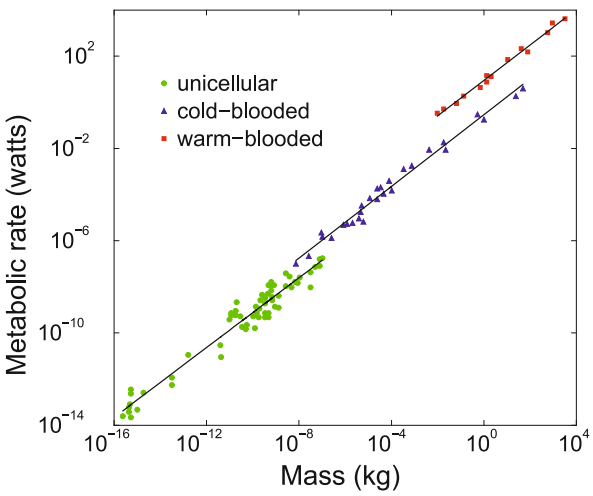

Fractals in Biology, Figure 3
Dependence of metabolic rate on body mass for different types of organisms (Kleiber's Law, after [112]). The least square fit lines have slopes 0.76 ± 0.01

$B \sim L^2 \sim M^{2/3}$. Postulating that the metabolic rate is proportional to some effective metabolic area leads to a strange prediction that this area should have a fractal dimension of 9/4.

Many attempts has been made to explain this interesting fact [10,32,142,143]. One particular attempt [142] was based on the ideas of energy optimization and the fractal organization of the cardiovascular system. However, the arguments were crucially dependent on such details as turbulent vs laminar flow, the elasticity of the blood vessels and the pulsatory nature of the cardiovascular system. The fact that this derivation does not work for species with different types of circulatory systems suggests that there might be a completely different and quite general explanation of this phenomenon.

Recently [10], as the theory of networks has became a hot subject, a general explanation of the metabolic rates was provided using a mathematical theorem that the total flow Q (e. g. of the blood) in the most efficient supply network must scale as $Q \sim BL/u$, where L is the linear size of the network, B is the total consumption rate (e. g. metabolic rate), and u is the linear size of each consumption unit (e. g. cell). The authors argue that the total flow is proportional to the total amount of the liquid in the network, which must scale as the body mass M. On the other hand, $L \sim M^{1/3}$. If one assumes that u is independent of the body mass, then $B \sim M^{2/3}$, which is identical to the simple but incorrect prediction based on the surface area. In order to make ends meet, the authors postulate that some combination of parameters must be independent of the body size, from where it follows that $u \sim M^{1/12}$. If one identifies a consumption unit with the cell, it seems that the cell size must scale as $L^{1/4}$. Thus the cells of a whale ($L = 30\,\mathrm{m}$) must be about 12 times larger than those of a C. elegance ($L = 1\,\mathrm{mm}$), which more or less consistent with the empirical data [110] for slowly dividing cells such as neurons in mammals. However, the issue of the metabolic scaling is still not fully resolved [11,33,144]. It is likely that the universal explanation of Kleiber's law is impossible. Moreover, recently it was observed that Kleiber's law does not hold for plants [105].

An interesting example of an allometric law which may have some fractal implications is the scaling of the brain size with the body mass $M_\mathrm{b} \sim M^{3/4}$ [1]. Assuming that the mass of the brain is proportional to the mass of the neurons in the body, we can conclude that if the average mass of a neuron does not depend on the body mass, neurons must form a fractal set with the fractal dimension 9/4. However, if we assume that the mass of a neuron must scale with the body mass as $M^{1/12}$, as it follows from [10,110], we can conclude that the number of neu-

rons in the body scales simply as $M^{2/3} \sim L^2$, which means that the number of neurons is proportional to the surface area of the body. The latter conclusion is physiologically more plausible than the former, since it is obvious that the neurons are more likely to be located near the surface of an organism.

From the point of view of comparative zoology, the universal scaling of the brain mass is not as important as the deviations from it. It is useful [1] to characterize an organism by the ratio of its actual brain mass to its expected brain mass which is defined as $E_b = AM^{3/4}$, where A is a constant measured by the intercept of the log–log graph of brain mass versus body mass. Homo sapiens have the largest $M_b/E_b = 8$, several times larger than those of gorillas and chimpanzees.

Physical Models of Biological Fractals

The fractals discussed in Sect. "Self-Similar Branching Structures", which are based on the explicit rules of construction and self-similarity, are useful concepts for analyzing branching structures in the living organisms whose development and growth are programmed according to these rules. However, in nature, there are many instances when fractal shapes emerge just from general physical principles without a special fractal "blueprint" [26,34,58,124,125,131]. One of such examples particularly relevant to biology is the diffusion limited aggregation (DLA) [145]. Another phenomenon, described by the same equations as DLA and thus producing similar shapes is viscous fingering [17,26,131]. Other examples are a random walk (RW) [42,58,104,141], a self-avoiding walk (SAW) [34,58], and a percolation cluster [26,58,127].

DLA explains the growth of ramified inorganic aggregates which sometimes can be found in rock cracks. The aggregates grow because certain ions or molecules deposit on the surface of the aggregate. These ions come to the surface due to diffusion which is equivalent on the microscopic level to Brownian motion of individual ions. The Brownian trajectories can be modeled as random walks in which the direction of each next step is independent of the previous steps.

The Brownian trajectory itself has fractal properties expressed by the Einstein formula:

$$r^2 = 2dtD\,, \tag{12}$$

where r^2 is the average square displacement of the Brownian particle during time t, D is the diffusion coefficient and d is the dimensionality of the embedding space. The number of steps, n, of the random walk is proportional to time of the Brownian motion $t = n\tau$, where τ is the average duration of the random walk step. The diffusion coefficient can be expressed as $D = r^2(m)/(2dm\tau)$, where $r^2(m)$ is the average square displacement during m time steps. Accordingly,

$$r(n) = \sqrt{p}\,r(m)\,, \tag{13}$$

which shows that the random walk of n steps consists of $p = n/m$ copies of the one-step walks arranged in space in such a way that the average linear size of this arrangement is $q = \sqrt{n/m}$ times larger than the size of the m-step walk (Fig. 4). Applying Eq. (5), we get $d_f = 2$ which does not depend on the embedding space. It is the same in one-dimensional, two-dimensional, and three-dimensional space. Note that Brownian trajectory is self-similar only in a statistical sense: each of the n concatenated copies of a m-step trajectory are different from each other, but their average square displacements are the same. Our eye can easily catch this statistical self-similarity. A random walk is a cloudy object of an elongated shape, its inertia ellipsoid is characterized by the average ratios of the squares of its axis: 12.1/2.71/1 [18].

The Brownian trajectory itself has a relevance in biology, since it may describe the changing of the electrical potential of a neuron [46], spreading of a colony [68], the foraging trajectory of a bacteria or an animal as well as the motion of proteins and other molecules in the cell. The self-avoiding walk (SAW) which has a smaller fractal dimension ($d_{f,SAW} = 4/3$ in $d = 2$ and $d_{f,SAW} \approx 1.7$ in

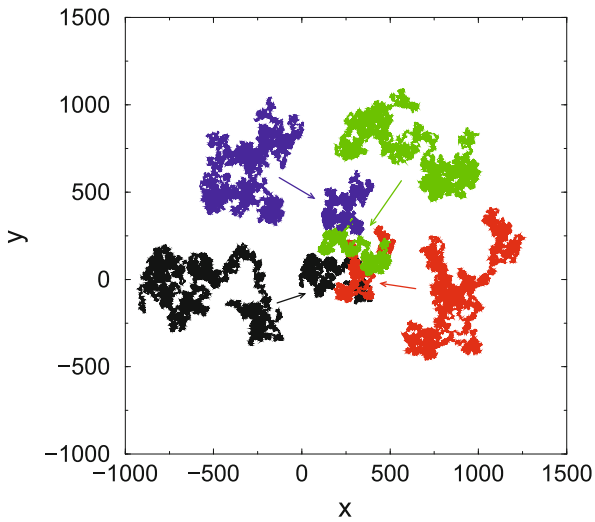

Fractals in Biology, Figure 4
A random walk of $n = 2^{16}$ steps is surrounded by its $p = 4$ parts of $m = n/p = 2^{14}$ steps magnified by factor of $q = \sqrt{4} = 2$. One can see that the shapes and sizes of each of the magnified parts are similar to the shape and size of the whole

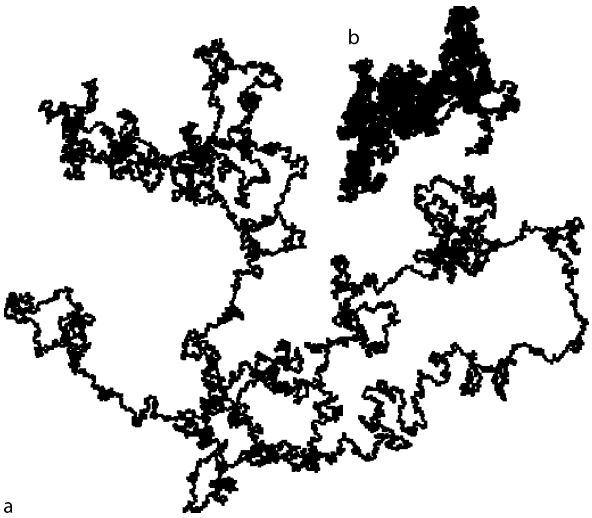

Fractals in Biology, Figure 5
A self-avoiding walk (a) of $n = 10^4$ in comparison with a random walk (b) of the same number of steps, both in $d = 3$. Both walks are produced by molecular dynamic simulations of the bead-on-a-string model of polymers. The hard core diameter of monomers in the SAW is equal to the bond length ℓ, while for the RW it is zero. Comparing their fractal dimensions $d_{f,SAW} \approx 1.7$, and $d_{f,RW} = 2$, one can predict that the average size (radius of inertia) of SAW must be $n^{1/d_{f,SAW}-1/d_{f,RW}} \approx 2.3$ times larger than that of RW. Indeed the average radius of inertia of SAW and RW are 98ℓ and 40ℓ, respectively. The fact that their complex shapes resemble leaving creatures was noticed by P. G. de Gennes in a cartoon published in his book "Scaling Concepts in Polymer Physics" [34]

$d = 3$) is a model of a polymer (Fig. 5) in a good solvent, such as for example a random coil conformation of a protein. Thus a SAW provides another example of a fractal object which has certain relevance in molecular biology. The fractal properties of SAWs are well established in the works of Nobel Laureates Flory and de Gennes [34,44,63].

Diffusion Limited Aggregation and Bacterial Colonies

The properties of a Brownian trajectory explain the ramified structure of the DLA cluster. Since the Brownian trajectory is not straight, it is very difficult for it to penetrate deep into the fiords of a DLA cluster. With a much higher probability it hits the tips of the branches. Thus DLA cluster (Fig. 6a) has a tree-like structure with usually 5 main branches in 2 dimensions. The analytical determination of its fractal dimension is one of the most challenging questions in modern mathematics. In computer simulations it is defined by measuring the number of aggregated particles versus the radius of gyration of the aggregate (Fig. 6c). The fractal dimension thus found is approximately 1.71

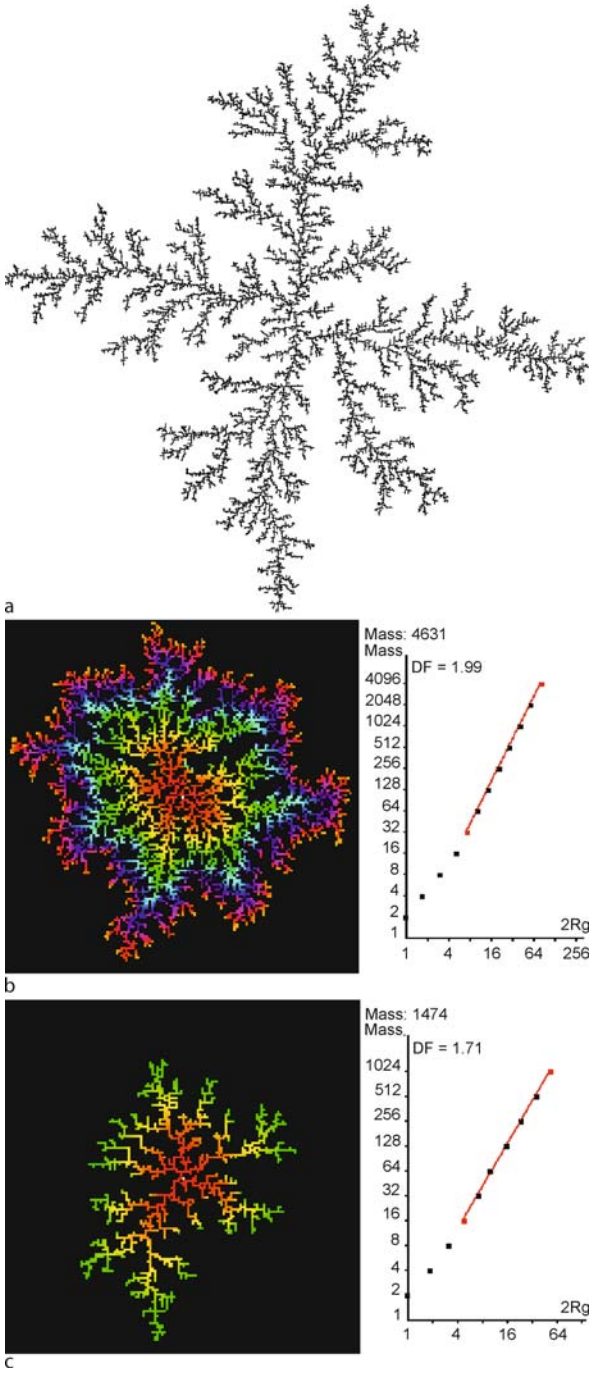

Fractals in Biology, Figure 6
a A DLA cluster of $n = 2^{14}$ particles produced by the aggregation of the random walks on the square lattice (*top*). A comparison of a small BD cluster (b) and a DLA cluster (c). The color in b and c indicates the deposition time of a particle. The slopes of the log-log graphs of the mass of the growing aggregates versus their gyration radius give the values of their fractal dimensions in the limit of large mass

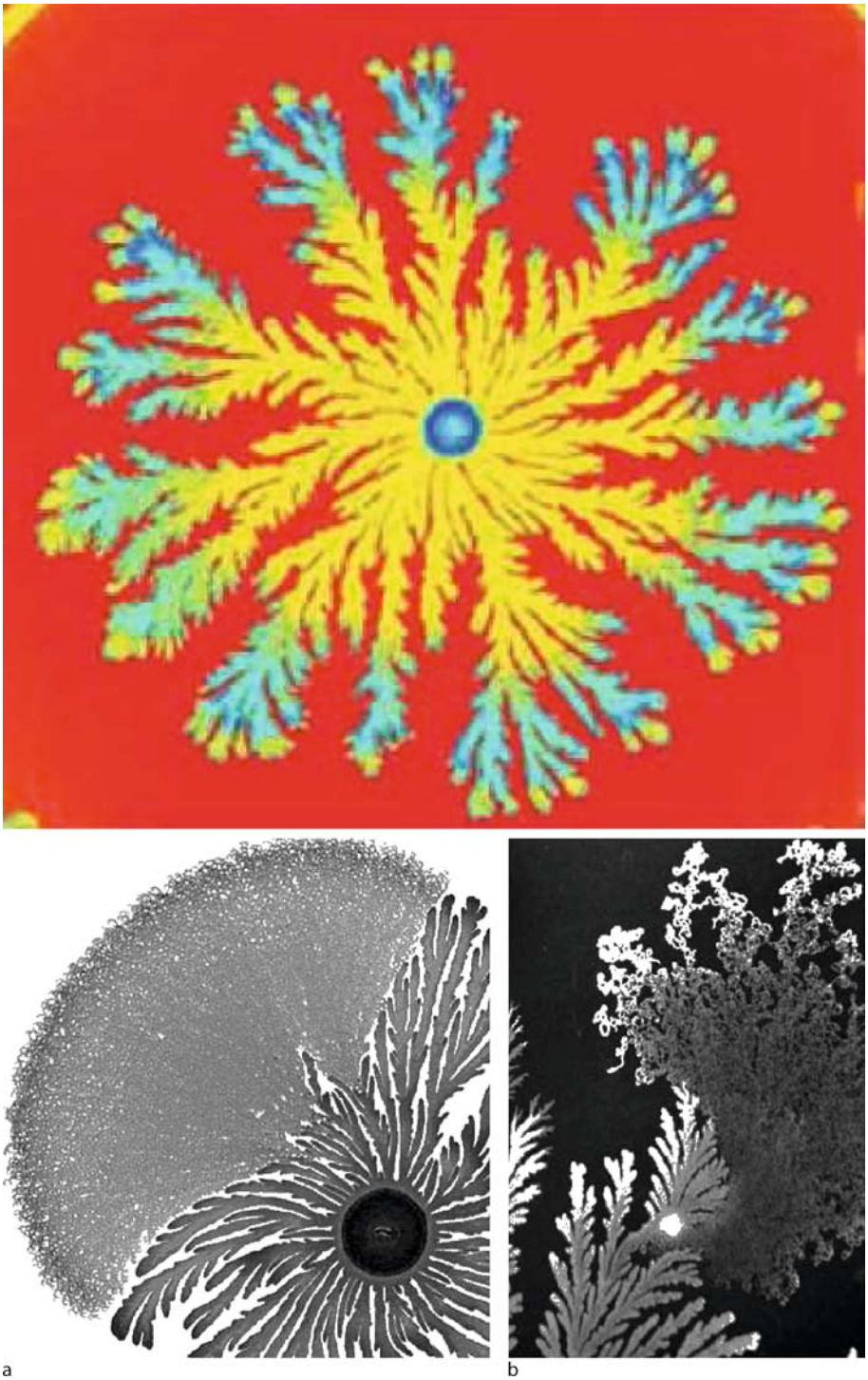

Fractals in Biology, Figure 7

A typical DLA-like colony grown in a Petri dish with a highly viscous substrate (*top*). A change in morphology from DLA-like colony growth to a swimming chiral pattern presumably due to a cell morphotype transition (*bottom*) (from [15])

in two dimensions and 2.50 in three dimensions [34]. It seems that as the aggregate grows larger the fractal dimension slightly increases.

Note that if the deposition is made by particles moving along straight lines (ballistic deposition) the aggregate changes its morphology and becomes almost compact and circular with small fluctuation on the boundaries and small holes in the interior (Fig. 6b). The fractal dimension of the ballistic aggregate coincides with the dimension of embedding space. Ballistic deposition (BD) belongs to the same universality class as the Eden model, the first model to describe the growth of cancer. In this model, each cell on the surface of the cluster can produce an offspring with equal probability. This cell does not diffuse but occupies one of the empty spaces neighboring to the parent cell [13,83].

DLA is supposed to be a good model for growth of bacterial colonies in the regime when nutrients are coming via diffusion in a viscous media from the exterior. Under different conditions bacteria colonies observe various transition in their morphology. If the nutrient supply is greater, the colonies increase their fractal dimension and start to resemble BD.

An interesting phenomenon is observed upon changing the viscosity of the substrate [15]. If the viscosity is small it is profitable for bacteria to swim in order to get to the regions with high nutrient concentrations. If the viscosity is large it is more profitable to grow in a static colony which is supplied by the diffusion of the nutrient from the outside. It seems that the way how the colony grows is inherited in the bacteria gene expression. When a bacteria grown at high viscosity is planted into a low viscosity substrate, its descendants continue to grow in a DLA pattern until a transition in a gene-expression happens in a sufficiently large number of neighboring bacteria which change their morphotype to a swimming one (Fig. 7). All the descendants of these bacteria start to swim and thus quickly take over the slower growing morphotype. Conversely, when a swimming bacteria is planted into a viscous media its descendants continue to swim until a reverse transition of a morphotype happens and the descendants of the bacteria with this new morphotype start a DLA-like growing colony. The morphotype transition can be also induced by fungi. Thus, it is likely that although bacteria are unicellular organisms, they exchange chemical signals similarly to the cells in the multicellular organisms, which undergo a complex process of cell differentiation during organism development (morphogenesis).

There is evidence that the tree roots also follow the DLA pattern, growing in the direction of diffusing nutrients [43]. Coral reefs [80] whose life depends on the diffusion of oxygen and nutrients also may to a certain degree follow DLA or BD patterns. Another interesting conjecture is that neuronal dendrites grow in vivo obeying the same mechanism [29]. In this case, they follow signaling chemicals released by other cells. It was also conjectured that even fingers of vertebrate organisms may branch in a DLA controlled fashion, resembling the pattern of viscous fingering which are observed when a liquid of low viscosity is pushed into a liquid of higher viscosity.

Measuring Fractal Dimension of Real Biological Fractals

There were attempts to measure fractal dimension of growing neurons [29] and other biological objects from photographs using a box-counting method and a circle method developed for empirical determination of fractal dimension. The circle method is the simplified version of the radial distribution function analysis described in the definition section. It consists of placing a center of a circle or a sphere of a variable radius R at each point of an object and counting the average number of other points of this object $M(R)$ found inside these circles or spheres. The slope of $\ln M(R)$ vs. $\ln R$ gives an estimate of the fractal dimension.

The box counting method consists of placing a square grid of a given spacing ℓ on a photograph and counting number of boxes $n(\ell)$ needed to cover all the points belonging to a fractal set under investigation. Each box containing at least one point of a fractal set is likely to contain on average $N(\ell) = \ell^{d_f}$ other points of this set, and $n(\ell)N(\ell) = N$, where N is the total number of the points. Thus for a fractal image $n(\ell) \sim N/\ell^{d_f} \sim \ell^{-d_f}$. Accordingly the fractal dimension of the image can be estimated by the slope of a graph of $\ln n(\ell)$ vs. $\ln \ell$.

The problem is that real biological objects do not have many orders of self-similarity, and also, as we see above, only a skeleton or a perimeter of a real object has fractal properties. The box counting method usually produces curvy lines on a log–log paper with at best one decade of approximately constant slope (Fig. 8). What is most disappointing is that almost any image analyzed in this fashion produces a graph of similar quality. For example, an Einstein cartoon presented in [93] has the same fractal dimension as some of the growing neurons. Therefore, experimental determination of the fractal dimension from the photographs is no longer in use in scientific publications. However, in the 1980s and early 1990s when the computer scanning of images became popular and the fractals were at the highest point of their career, these exercises were frequent.

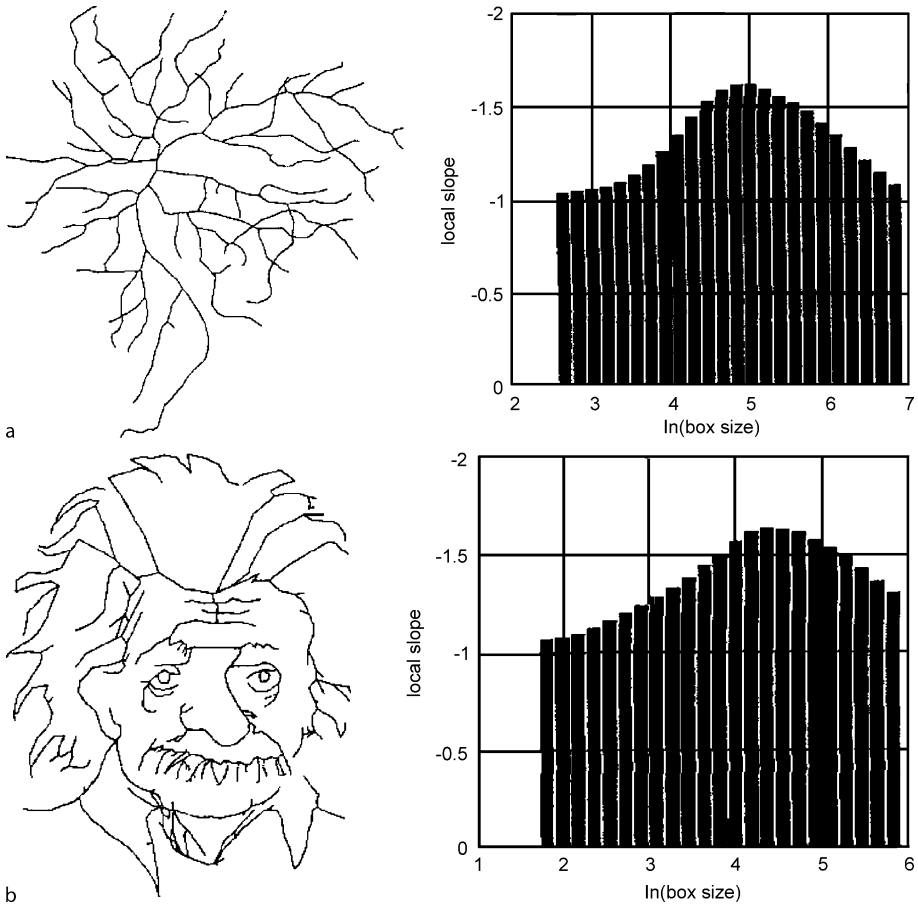

Fractals in Biology, Figure 8

An image of a neuron and its box-counting analysis (a) together with the analogous analysis of an Einstein cartoon (b). The log–log plot of the box-counting data has a shape of a curve with the changing slope. *Bars* represent the local slope of this graph, which may be interpreted as some effective fractal dimension. However, in both graphs the slopes change dramatically from 1 to almost 1.7 and it is obvious that any comparison to the DLA growth mechanism based on these graphs is invalid (from [93])

For example, R. Voss [137] analyzed fractal dimensions of Chinese graphics of various historical periods. He found that the fractal dimension of the drawings fluctuated from century to century and was about 1.3 at the time of the highest achievements of this technique. It was argued that the perimeters of many natural objects such as perimeters of the percolation clusters and random walks have similar fractal dimension; hence the works of the artists who were able to reproduce this feature were the most pleasant for our eyes.

Recently, the box counting method was used to analyze complex networks such as protein interaction networks (PIN). In this case the size of the box was defined as the maximal number of edges needed to connect any two nodes belonging to the box. The box-counting method appears to be a powerful tool for analyzing self-similarity of the network structure [120,121].

Percolation and Forest Fires

It was also conjectured that biological habitats of certain species may have fractal properties. However it is not clear whether we have a true self-similarity or just an apparent mosaic pattern which is due to complex topography and geology of the area. If there is no reasonable theoretical model of an ecological process, which displays true fractal properties the assertions of the fractality of a habitat are pointless. One such model, which produces fractal clusters is percolation [26,127]. Suppose a lightning hits a random tree in the forest. If the forest is sufficiently dry, and the inflammable trees are close enough the fire will spread from a tree to a tree and can burn the entire forest. If the inflammable trees form a connected cluster, defined as such that its any two trees are connected with a path along which the fire can spread, all the trees in this cluster will

be burnt down. Percolation theory guarantees that there is a critical density p_c of inflammable trees below which all the connected clusters are finite and the fire will naturally extinguish burning only an infinitesimally small portion of the forest. In contrast, above this threshold there exists a giant cluster which constitutes a finite portion of the forest so that the fire started by a random lightning will on average destroy a finite portion of the forest. Exactly at the critical threshold, the giant cluster is a fractal with a fractal dimension $d_f = 91/48 \approx 1.89$. The probability to burn a finite cluster of mass S follows a power law $P(S) \sim S^{-d/d_f}$. Above and below the percolation threshold, this distribution is truncated by an exponential factor which specifies that clusters of linear dimensions exceeding a characteristic size are exponentially rare. The structure of the clusters which do not exceed this characteristic scale is also fractal, so if their mass S is plotted versus their linear size R (e. g. radius of inertia) it follows a power law $S \sim R^{d_f}$

In the natural environment the thunderstorms happen regularly and they produce fractal patterns of burned down patches. However, if the density of trees is small, the patches are finite and the fires are confined. As the density the trees reaches the critical threshold, the next thunderstorm is likely to destroy the giant cluster of the forest, which will produce a bare patch of a fractal shape spreading over the entire forest. No major forest fire will happen until the new forest covers this patch creating a new giant cluster, because the remaining disconnected groves are of finite size.

There is an evidence that the forest is the system which drives itself to a critical point [30,78,111] (Fig. 9). Suppose that each year there is certain number of lightning strokes per unit area n_l. The average number of trees in a cluster is $\langle S \rangle = a|p - p_c|^{-\gamma}$, where $\gamma = 43/18$ is one of the critical exponents describing percolation, a is a constant, and p is the number of inflammable trees per unit area. Thus the number of trees destroyed annually per unit area is $n_l a|p_c - p]^{-\gamma}$. On the other hand, since the trees are growing, the number of trees per unit area increases annually by n_t, which is the parameter of the ecosystem. At equilibrium, $n_l a|p_c - p]^{-\gamma} = n_t$. Accordingly, the equilibrium density of trees must reach $p_e = p_c - (a n_l/n_t)^{1/\gamma}$. For very low n_l, p_e will be almost equal to p_c, and the chance that in the next forest fire a giant cluster which spans the entire ecosystem will be burned is very high. It can be shown that for a given chance c of hitting such a cluster in a random lightning stroke, the density of trees must reach $p = p_c - f(c)L^{-1/\nu}$, where L is the linear size of the forest, $\nu = 4/3$ is the correlation critical exponent and $f(c) > 0$ is a logarithmically growing function of c. Accordingly, if $n_t/n_l > bL^{\gamma/\nu}$, where b is some constant, the chance of getting a devastating forest fire is close to 100%. We have here a paradoxical situation: the more frequent are the forest fires, the least dangerous they are. This implies that the fire fighters should not extinguish small forest fires which will be contained by themselves. Rather they should annually cut a certain fraction of trees to decrease n_t.

As we see, the forest fires can be regarded as a self-organized critical (SOC) system which drive themselves towards criticality. As in many SOC systems, here there are two processes one is by several orders of magnitudes faster than the other. In this case they are the tree growing process and the lightning striking process. The model reaches the critical point if a tuning parameter $n_t/n_l \to \infty$ and in addition $n_t \to 0$ and $L \to \infty$, which are quite reasonable assumptions. In a regular critical system a tuning parameter (e. g. temperature) must be in the vicinity of a specific finite value. In a SOC system the tuning parameter must be just large enough.

There is an evidence, that the forest fires follow a power law distribution [109]. One can also speculate that the areas burned down by the previous fires shape fractal habitats for light loving and fire-resistant trees such as pines. The attempts to measure a fractal dimension of such habitats from aerial photographs are dubious due to the limitations discussed above and also because by adjusting a color threshold one can produce fractal-like clusters in almost any image. This artifact is itself a trivial consequence of the percolation theory.

The frequently reported Zipf's law for the sizes of colonies of various species including the areas and populations of the cities [146] usually arises not from fractality of the habitats but from preferential attachment growth in which the old colonies grow in proportion to their present population, while the new colonies may form with a small probability. The preferential attachment model is a very simple mechanism which can create a power-law distribution of colony sizes $P(S) \sim S^{-2-\phi}$, where ϕ is a small correction which is proportional to the probability of formation of new colonies [25]. Other examples of power laws in biology [67], such as distributions of clusters in the metabolic networks, the distribution of families of proteins, etc. are also most likely come from the preferential attachment or also, in some cases, can arise as artifacts of specifically selected similarity threshold, which brings a network under consideration to a critical point of the percolation theory.

The epidemic spreading can also be described by a percolation model. In this case the contagious disease spreads from a person to a person as the forest fire spreads from

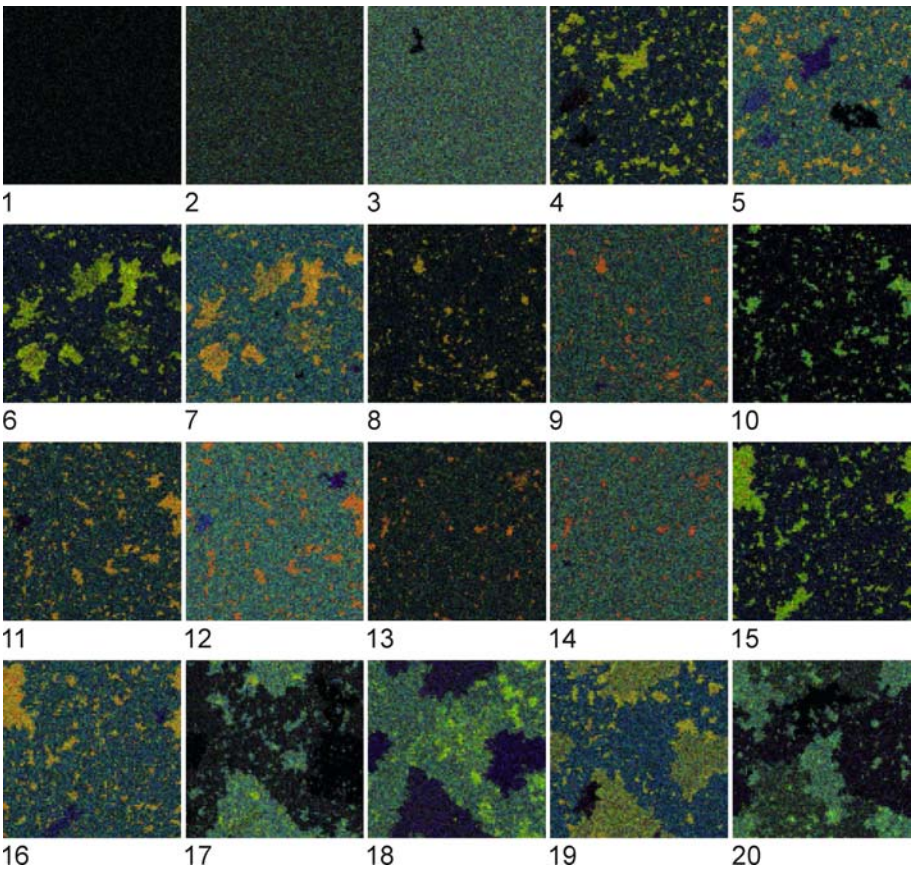

Fractals in Biology, Figure 9

A sequence of frames of a forest fire model. Each tree occupies a site on 500 × 500 square lattice. At each time step a tree (a colored site) is planted at a randomly chosen empty site (*black*). Each 10,000 time steps a lightning strikes a randomly chosen tree and the forest fire eliminates a connected cluster of trees. The frames are separated by 60,000 time steps. The color code indicates the age of the trees from *blue* (young) to *red* (old). The initial state of the system is an empty lattice. As the concentration of trees reaches percolation threshold (frame 3) a small finite cluster is burned. However it does not sufficiently decrease the concentration of trees and it continues to build up until a devastating forest fire occurs between frames 3 and 4, with only few green groves left. Between frames 4 and 5 several lightnings hit these groves and they are burned down, while surrounding patch of the old fire continues to be populated by the new trees. Between frame 5 and 6 a new devastating forest fire occurs. At the end of the movie huge intermittent forest fires produce gigantic patches of dense and rare groves of various ages

a tree to a tree. Analogously to the forest fire model, a person who caught the disease dies or recovers and becomes immune to the disease, so he or she cannot catch it again. This model is called susceptible-infective-removed (SIR) [62]. The main difference is that the epidemics spread not on a two-dimensional plane but on the network describing contacts among people [61,86]. This network is usually supposed to be scale free, i. e. the number of contacts different people have (degree) is distributed according to a an inverse power law [2,71]. As the epidemic spreads the number of connections in the susceptible population depletes [115] so the susceptible population comes to a percolation threshold after which the epidemic stops. This model explains for example why the Black Death epidemic stopped in the late 14th century after killing about one third of the total European population.

Critical Point and Long-Range Correlations

Percolation critical threshold [26,127] is an example of critical phenomena the most well known of which is the liquid-gas critical point [124]. As one heats a sample of any liquid occupying a certain fraction of a closed rigid container, part of the liquid evaporates and the pressure in the container increases so that the sample of liquid remains at equilibrium with its vapor. However, at certain

temperature the visible boundary between the liquid at the bottom and the gas on the top becomes fuzzy and eventually the system becomes completely nontransparent as milk. This phenomena is called critical opalescence and the temperature, pressure, and density at which it happens is called a critical point of this liquid. For water, the critical point is $T_c = 374\,C^\circ$, $P_c = 220\,\text{atm}$ and $\rho_c = 330\,\text{kg/m}^3$. As the temperature goes above the critical point the system becomes transparent again, but the phase boundary disappears: liquid and gas cannot coexist above the critical temperature. They form a single phase, supercritical fluid, which has certain properties of both gas (high compressibility) and liquid (high density and slow diffusivity).

At the critical point the system consists of regions of high density (liquid like) and low density (gas like) of all sizes from the size of a single molecule to several microns across which are larger than the wave length of visible light $\approx 0.5\,\mu$m. These giant density fluctuations scatter visible light causing the critical opalescence. The characteristic linear size of the density fluctuations is called the correlation length ξ which diverges at critical temperature as $\xi \sim |T - T_c|^{-\nu}$. The shapes of these density fluctuations are self-similar, fractal-like. The system can be represented as a superposition of a uniform set with the density corresponding to the average density of the system and a fractal set with the fractal dimension d_f. The density correlation function $h(r)$ decreases for $r \to \infty$ as $r^{-\chi}\exp(-r/\xi)$, where $\chi \equiv d - 2 + \eta = d - d_f$ and $d = 1, 2, 3, \ldots$ is the dimension of space in which the critical behavior is observed. The exponential cutoff sets the upper limit of the fractal density fluctuations to be equal to the correlation length. The lower limit of the fractal behavior of the density fluctuations is one molecule. As the system approaches the critical point, the range of fractal behavior increases and can reach at least three orders of magnitude in the narrow vicinity of the critical point. The intensity of light scattered by density fluctuations at certain angle can be expressed via the Fourier transform of the density correlation function $S(\mathbf{f}) \sim \int h(\mathbf{r})\exp(\mathrm{i}\mathbf{r}\cdot\mathbf{f})\mathrm{d}\mathbf{r}$, where the integral is taken over d-dimensional space. This experimentally observed quantity $S(f)$ is called the structure factor. It is a powerful tool to study the fractal properties of matter since it can uncover the presence of a fractal set even if it is superposed with a set of uniform density. If the density correlation function has a power law behavior $h(r) \sim r^{-\chi}$, the structure factor also has a power law behavior $S(f) \sim f^{\chi - d}$. For the true fractals with $d_f < d$, the average density over the entire embedding space is zero, so the correlation function coincides with the average density of points of the set at certain distance from a given point, $h(r) \sim r^{-d+1}dM(r)/dr \sim r^{d_f - d}$, where $M(r) \sim r^{d_f}$ is the mass of the fractal within the radius r from a given point. Thus, for a fractal set of points with fractal dimension d_f, the structure factor has a simple power law form $S(f) \sim f^{-d_f}$. Therefore whenever $S(f) \sim f^{-\beta}$, $\beta < d$ is identified with the fractal dimension and the system is said to have fractal density fluctuations even if the average density of the system is not zero.

There are several methods of detecting spatial correlations in ecology [45,108] including box-counting [101]. In addition one can study the density correlation function $h(r)$ of certain species on the surface of the Earth defined as $h(r) = N(r)/2\pi r\Delta r - N/A$, where $N(r)$ is the average number of the representatives of a species within a circular rim of radius r and width Δr from a given representative, N is the total population, and A is the total area. For certain species, the correlation function may follow about one order of magnitude of a fractal (power-law) behavior [9,54,101]. One can speculate that there is some effective attraction between the individuals such as cooperation (mimicking the van der Walls forces between molecules) and a tendency to spread over the larger territory (mimicking the thermal motion of the particles). The interplay of these two tendencies may produce a fractal behavior like in the vicinity of the critical point.

However, there is no reason of why the ecological system although complex and chaotic [81] should drive itself to a critical point. Whether or not there is a fractal behavior, the density correlation function is a useful way to describe the spatial distribution of the population. It can be applied not only in ecology but also in physiology and anatomy to describe the distribution of cells, for example, neurons in the cortex [22].

Lévy Flight Foraging

A different set of mathematical models which produce fractal patterns and may be relevant in ecology is the Lévy flight and Lévy walk models [118,147]. The Lévy flight model is a generalization of a random walk, in which the distribution of the steps (flights) of length ℓ follows a power law $P(\ell) \sim \ell^{-\mu}$ with $\mu < 3$. Such distributions do not have a finite variance. The probability density of the landing points of the Lévy flight converges not to a Gaussian as for a normal random walk with a finite step variance but to a Lévy stable distribution with the parameter $\alpha \equiv \mu - 1$. The landing points of a Lévy flight form fractal dust similar to a Cantor set with the fractal dimension $d_f = \alpha$.

It was conjectured that certain animals may follow Lévy flights during foraging [103,116,132,133]. It is a mathematical theorem [23,24] that in case when targets

are scarce but there is a high probability that a new target could be discovered in the vicinity of the previously found target, the optimal strategy for a forager is to perform a Lévy flight with $\mu = 2 + \phi$, where ϕ is a small correction which depends on the density of the target sites, the radius of sight of the forager and the probability to find a new target in the vicinity of the old one. In this type of "inverse square" foraging strategy a balance is reached between finding new rich random target areas and returning back to the previously visited area which although depleted, may still provide some food necessary for survival. The original report [132] of the distribution of the flight times of wandering albatross was in agreement with the theory.

However, recent work [39] using much longer flight time records and more reliable analysis showed that the distribution of flight times is better described by a Poisson distribution corresponding to regular random walk rather then by a power law. Several other reports of the Lévy flight foraging are also found dubious. The theory [23] however predicts that the inverse square law of foraging is optimal only in case of scarce sources distribution. Thus if the harvest is good and the food is plenty there is no reason to discover the new target locations and the regular random walk strategy becomes the most efficient. Subsequent observations show that power laws exist in some other marine predator search behavior [119]. In order to find a definitive answer, the study must be repeated for the course of many successive years characterized by different harvests.

Once the miniature electronic devices for tracing animals are becoming cheaper and more efficient, a new branch of electronic ecology is emerging with the goal of quantifying migratory patterns and foraging strategies of various species in various environments. Whether or not the foraging patterns are fractal, this novel approach will help to establish better conservational policy with scientifically sound borders of wild-life reservations. Recent observations indicate that human mobility patterns might also possess Lévy flight properties [49].

Dynamic Fractals

It is not necessary that a fractal is a real geometrical object embedded in the regular three-dimensional space. Fractals can be found in the properties of the time series describing the behavior of the biological objects. One classical example of a biological phenomenon which under certain condition may display fractal properties is the famous logistic map [38,82,95,123] based on the ideas of a great British economist and demographer Robert Malthus. Suppose that there is a population N_t of a certain species enclosed in a finite habitat (e. g. island) at time t. Here t is an integer index which may denote year. Suppose that at time $t + 1$ the population becomes

$$N_{t+1} = bN_t - dN_t^2 ,$$

where b is the natural birth rate and d is the death rate caused by the competition for limited resources. In the most primitive model, the animals are treated as particles randomly distributed over area A annihilating at each time step if the distance between them is less than r. In this case $d = \pi r^2/A$. The normalized population $x_t = dN_t/b$ obeys a recursive relation with a single parameter b:

$$x_{t+1} = bx_t(1 - x_t) .$$

The behavior of the population is quite different for different b. For $b \leq 1$ the population dies out as $t \to \infty$. For $1 < b \leq b_0 = 3$ the population converges to a stable size. If $b_{n-1} < b \leq b_n$, the population repetitively visits 2^n values $0 < x_1, \ldots, x_{2^n} < 1$ called attractors as $t \to \infty$. The bifurcation points $3 = b_0 < b_1 < b_2 < \cdots < b_n < b_\infty \approx 3.569945672$ converge to a critical value b_∞, at which the set of population sizes becomes a fractal with fractal dimension $d_f \approx 0.52$ [50]. This fractal set resembles a Cantor set confined between 0 and 1. For $b_\infty < b < 4$, the behavior is extremely complex. At certain values b chaos emerges and the behavior of the population become unpredictable, i. e. exponentially sensitive to the initial conditions. At some intervals of parameter b, the predictable behavior with a finite attractor set is restored. For $b > 4$, the population inevitably dies out. Although the set of attractors becomes truly fractal only at certain values of the birth rate, and the particular value of the fractal dimension does not have any biological meaning, the logistic map has a paradigmatic value in the studies of population dynamics with an essentially Malthusian take home massage: excessive birth rate leads to disastrous consequences such as famines coming at unpredictable times, and in the case of Homo sapiens to devastating wars and revolutions.

Fractals and Time Series

As we can see in the previous section, even simple systems characterized by nonlinear feedbacks may display complex temporal behavior, which often becomes chaotic and sometimes to fractal. Obviously, such features must be present in the behavior of the nervous system, in particular in the human brain which is probably the most complex system known to contemporary science. Nevertheless, the source of fractal behavior can be sometimes trivial with-

out evidence of any cognitive ability. An a example of such a trivial fractal behavior is the random firing of a neuron [46] which integrates through its dendritic synapses inhibitory and excitatory signals from its neighbors. The action potential of such a neuron can be viewed as performing a one-dimensional random walk going down if an inhibitory signal comes from a synapse or going up if an excitatory signal comes form a different synapse. As soon as the action potential reaches a firing threshold the neuron fires, its action potential drops to an original value and the random walk starts again. Thus the time intervals between the firing spikes of such a neuron are distributed as the returning times of a one-dimensional random walk to an origin. It is well known [42,58,104,141], that the probability density $P(\Delta t)$ of the random walk returns scales as $(\Delta t)^{-\mu}$ with $\mu = 3/2$. Accordingly, the spikes on the time axis form a fractal dust with the fractal dimension $d_f = \mu - 1 = 1/2$.

A useful way to study the correlations in the time series is to compute its autocorrelation function and its Fourier transform which is called power spectrum, analogous to the structure factor for the spatial correlation analysis. Due to the property of the Fourier transform to convert a convolution into a product, the power spectrum is also equal to the square of the Fourier transform of the original time series. Accordingly it has a simple physical meaning telling how much energy is carried in a certain frequency range. In case of a completely uncorrelated signal, the power spectrum is completely flat, which means that all frequencies carry the same energy as in white light which is the mixture of all the rainbow colors of different frequencies. Accordingly a signal which has a flat power spectrum is called white noise. If the autocorrelation function $C(t)$ decays for $t \to \infty$ as $C(t) \sim t^{-\chi}$, where $0 < \chi < 1$, the power spectrum $S(f)$ of this time series diverges as $f^{-\beta}$, with $\beta = 1 - \chi$. Thus the LRPLC in the time series can be detected by studying the power spectrum. There are alternative ways of detecting LRPLC in time series, such as Hurst analysis [41], detrended fluctuation analysis (DFA) [19,98] and wavelet analysis [4]. These methods are useful for studying short time series for which the power spectrum is too noisy. They measure the Hurst exponent $\alpha = (1 + \beta)/2 = 1 - \chi/2$.

It can be shown [123] that for a time series which is equal to zero everywhere except at points $t_1, t_2, \ldots, t_n, \ldots$, at which it is equal to unity and these points form a fractal set with fractal dimension d_f, the time autocorrelation function $C(t)$ decreases for $t \to \infty$ as $C(t) \sim t^{-\chi}$, where $\chi = 1 - d_f$. Therefore the power spectrum $S(f)$ of this time series diverges for $f \to 0$ as $f^{-\beta}$, with $\beta = 1 - \chi = d_f$. Accordingly, for the random walk model of neuron firing, the power spectrum is characterized by $\beta = 1/2$.

In the more general case, the distribution of the intervals $\Delta t_n = t_n - t_{n-1}$ can decay as a power law $P(\Delta t) \sim (\Delta t)^{-\mu}$, with $1 < \mu < 3$. For $\mu < 2$, the set $t_1, t_2, \ldots, t_n$ is a fractal, and the power spectrum decreases as power law $S(f) \sim f^{-d_f} = f^{-\mu+1}$. For $2 < \mu < 3$, the set $t_1, t_2, \ldots, t_n$ has a finite density, with $d_f = D = 1$. However, the power spectrum and the correlation function maintain their power law behavior for $\mu < 3$. This behavior indicates that although the time series itself is uniform, the temporal fluctuations remain fractal. In this case, the exponent β characterizing the low frequency limit is given by $\beta = 3 - \mu$. The maximal value of $\beta = 1$ is achieved when $\mu = 2$. This type of signal is called $1/f$-noise or "red" noise. If $\mu \geq 3$, $\beta = 0$ in the limit of low frequencies and we again recover white noise.

The physical meaning of $1/f$ noise is that the temporal correlations are of infinite range. For the majority of the processes in nature, temporal correlations decay exponentially $C(t) \sim \exp(-t/\tau)$, where the characteristic memory span τ is called relaxation or correlation time. For a time series with a finite correlation time τ, the power spectrum has a Lorentzian shape, namely it stays constant for $f < 1/\tau$ and decreases as $1/f^2$ for $f > 1/\tau$. The signal in which $S(f) \sim 1/f^2$ is called brown noise, because it describes the time behavior of the one-dimensional Brownian motion. Thus for the majority of the natural processes, the time spectrum has a crossover from a white noise at low frequencies to a brown noise at high frequencies. The relatively unusual case when $S(f) \sim f^{-\beta}$, where $0 < \beta < 1$ is called fractal noise because as we see above it describes the behavior of the fractal time series. $1/f$-noise is the special type of the fractal noise corresponding to the maximal value of β, which can be achieved in the limit of low frequencies.

R. Voss and J. Clarke [135,136] had analyzed the music written by different composers and found that it follows $1/f$ noise over at least three orders of magnitude. It means that music does not have a characteristic time-scale. There is an evidence that physiological process such as heart-beat, gate, breath and sleeping patterns as well as certain types of human activity such as sending e-mails has certain fractal features [6,16,20,27,47,55,57,72,73,97,99,100,113,126,129]. A. Goldberger [99,126] suggested that music is pleasant for us because it mimics the fractal features of our physiology. It has to be pointed out that in all these physiological time series there is no clear power law behavior expanding over many orders of magnitude. There is also no simple clear explanation of the origins of fractality. One possible mechanism could be due

the distribution of the return times of a random walk, which has been used to explain the sleeping patterns and response to the e-mails.

SOC and Biological Evolution

Self-organized criticality (SOC) [8,92] describes the behavior of the systems far form equilibrium, the general feature of which is a slow increase of strain which is interrupted by an avalanche-like stress release. These avalanches are distributed in a power law fashion. The power spectrum of an activity at a given spatial site is described by a fractal noise. One of the most successful application of SOC is the explanation of the power-law distribution of the magnitudes of the earthquakes (Gutenberg–Richter's law) [89,130].

The simple physical models of self-organized criticality are invasion percolation [26,127], sand-pile model [8], and Bak–Sneppen model of biological evolution [7]. In the one-dimensional Bak–Sneppen model, an ecosystem is represented by a linear chain of the pray-predator relationships in with each species is represented by a site on a straight line surrounded by its predator (a site to the right) and a pray (a site to the left). Each site is characterized by its fitness f which at the beginning is uniformly distributed between 0 and 1. At each time step, a site with the lowest fitness becomes extinct and is replaced by a mutated species with a new fitness randomly taken from a uniform distribution between 0 and 1. The fitnesses of its two neighbors (predator and prey) are also changed at random. After certain equilibration time, the fitness of almost all the species except a few becomes larger than a certain critical value f_c. These few active species with low fitness which can spontaneously mutate form a fractal set on the pray-predator line. The activity of each site can be represented by a time series of mutations shown as spikes corresponding to the times of individual mutations at this site. The power spectrum of this time series indicates the presence of the fractal noise. At a steady state the minimal fitness value which spontaneously mutates fluctuates below f_c with a small probability $P(\epsilon)$ comes into the interval between $f_c - \epsilon$ and f_c. The distribution of the first return times Δt_ϵ to a given ϵ-vicinity of f_c follows a power law with an ϵ-dependent exponential cutoff. Since each time step corresponds to a mutation, the time interval for which f stays below $f_c - \epsilon$ corresponds to an avalanche of mutations caused by a mutation of a very stable species with $f > f_c - \epsilon$. Accordingly one can speculate, that evolution goes as a punctuated equilibrium so that an extinction of a stable species causes a gigantic extinction of many other species which hitherto have been well adopted. The problem with this SOC model is the definition of a time step. In order to develop a realistic model of evolution one needs to assume that the real time needed for a spontaneous mutation of species with different fitness dramatically increases with f, going for example as $\exp(fA)$, where A is a large value. Unfortunately, it is impossible to verify the predictions of this model because paleontological records do not provide us with a sufficient statistics.

Fractal Features of DNA Sequences

DNA molecules are probably the largest molecules in nature [139]. Each strand of DNA in large human chromosomes consist of about 10^8 monomers or base-pairs (bp) which are adenine (A), cytosine (C), guanine (G), and thymine (T). The length of this molecule if stretched would reach several centimeters. The geometrical packing of DNA in a cell resembles a self-similar structure with at least 6 levels of packing: a turn of the double helix (10 bp), a nucleosome (200 bp), a unit of 30 nm fiber (6 nucleosomes), a loop domain ($\approx$ 100 units of 30 nm fiber), a turn of a metaphase chromosome ($\approx$ 100 loop domains), a metaphase chromosome ($\approx$ 100 turns). The packing principle is quite similar to the organization of the information in the library: letters form lines, lines form pages, pages form books, books are placed on the shelves, shelves form bookcases, bookcases form rows, and rows are placed in different rooms. This structure however is not a rigorous fractal, because packing of the units on different levels follows different organization principles.

The DNA sequence treated as a sequence of letters also has certain fractal properties [4,5,19,69,70,96]. This sequence can be transformed into a numerical sequence by several mapping rules: for example A rule, in which A is replaced by 1 and C, T, G are replaced by 0, or SW rule in which strongly bonded bp (C and G) are replaced by 1 and weakly bonded bp (A and T) are replaced by -1. Purine-Pyrimidine (RY) mapping rule (A,G: +1; C,T:-1) and KM mapping rule (A,C:+1; G,T -1) are also possible. Power spectra of such sequences display large regions of approximate power law behavior in the range from $f = 10^{-2}$ to $f = 10^{-8}$. For SW mapping rule we has almost perfect $1/f$ noise in the region of low frequencies (Fig. 10). This is not surprising because chromosomes are organized in a large CG rich patches followed by AT rich patches called isochors which extend over millions of bp. The changing slope of the power spectra for different frequency ranges clearly indicates that DNA sequences are also not rigorous fractals but rather a mosaic structures with different organization principles on different length-scales [60,98]. A possible relation between the frac-

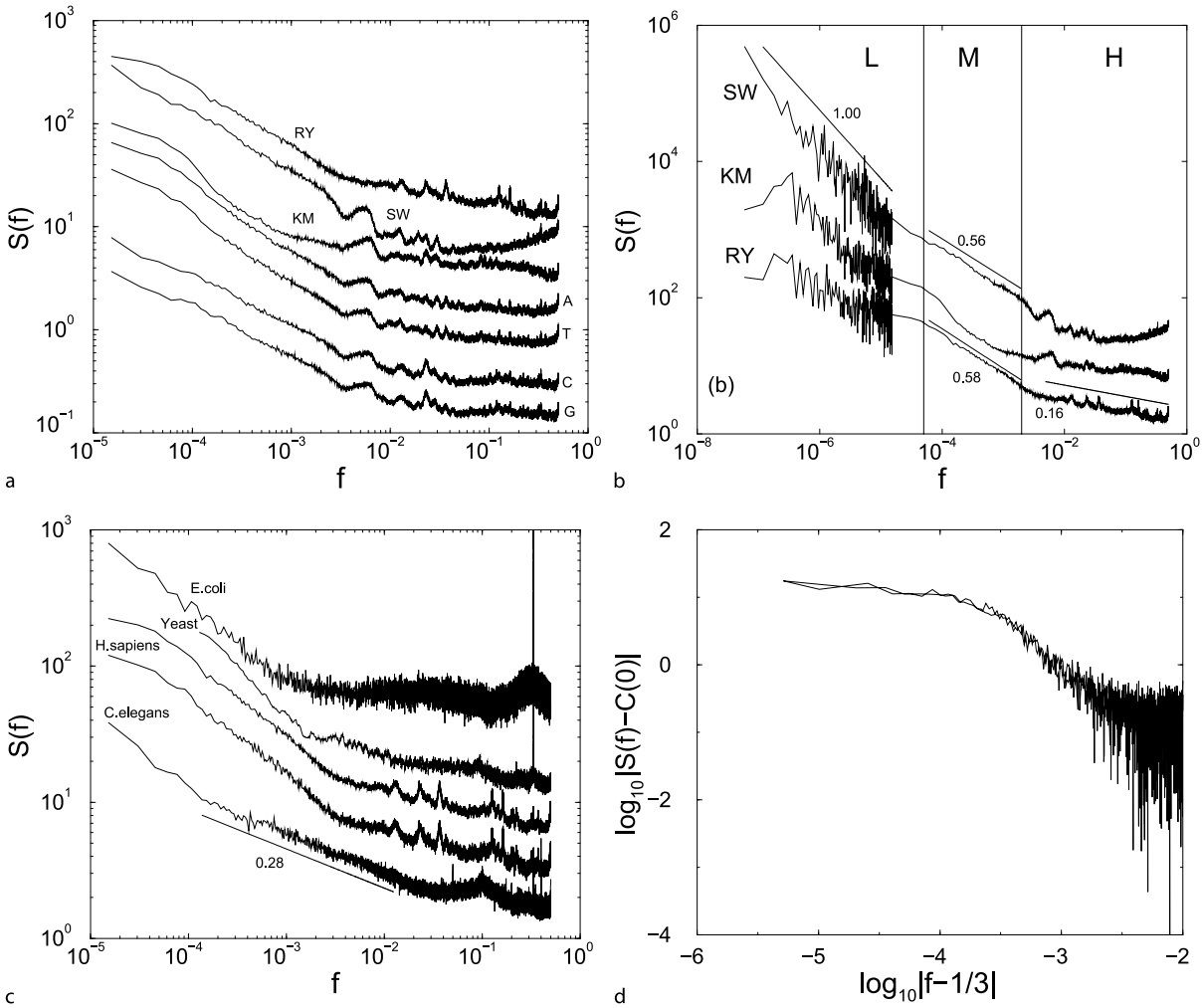

Fractals in Biology, Figure 10

a Power spectra for seven different mapping rules computed for the Homo sapiens chromosome XIV, genomic contig NT_026437. The result is obtained by averaging 1330 power spectra computed by fast Fourier transform for non-overlapping segments of length $N = 2^{16} = 65536$. **b** Power spectra for SW, RY, and KM mapping rules for the same contig extended to the low frequency region characterizing extremely long range correlations. The extension is obtained by extracting low frequencies from the power spectra computed by FFT with $N = 2^{24} \approx 16 \times 10^6$ bp. Three distinct correlation regimes can be identified. High frequency regime ($f < 0.003$) is characterized by small sharp peaks. Medium frequency regime ($0.5 \cdot 10^{-5} < f < 0.003$) is characterized by approximate power-law behavior for RY and SW mapping rules with exponent $\beta_M = 0.57$. Low frequency regime ($f < 0.5 \cdot 10^{-5}$) is characterized by $\beta = 1.00$ for SW rule. The high frequency regime for RY rule can be approximated by $\beta_H = 0.16$ in agreement with the data of Fig. 11. **c** RY Power spectra for the entire genome of *E. coli* (bacteria), *S. cerevisae* (yeast) chromosome IV, *H. sapiens* (human) chromosome XIV and the largest contig (NT_032977.6) on the chromosome I; and *C. elegans* (worm) chromosome X. It can be clearly seen that the high frequency peaks for the two different human chromosomes are exactly the same, while they are totally different from the high frequency peaks for other organisms. These high frequency peaks are associated with the interspersed repeats. One can also notice the presence of enormous peaks for $f = 1/3$ in *E. coli* and yeast, indicating that their genomes do not have introns, so that the lengths of coding segments are very large. The *C. elegans* data can be very well approximated by power law correlations $S(f) \sim f^{-0.28}$ for $10^{-4} < f < 10^{-2}$. **d** Log–log plot of the RY power spectrum for *E. coli* with subtracted white noise level versus $|f - 1/3|$. It shows a typical behavior for a signal with finite correlation length, indicating that the distribution of the coding segments in *E. coli* has finite average square length of approximately 3×10^3 bp

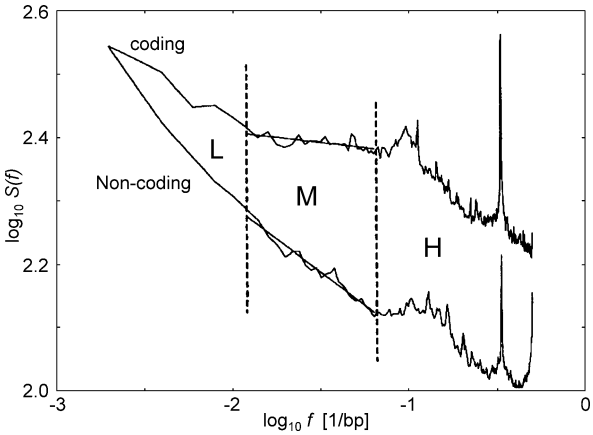

Fractals in Biology, Figure 11

RY Power spectra averaged over all eukaryotic sequences longer than 512 bp, obtained by FFT with window size 512. *Upper curve* is average over 29,453 coding sequences; *lower curve* is average over 33,301 noncoding sequences. For clarity, the power spectra are shifted vertically by arbitrary quantities. The *straight lines* are least squares fits for second decade (Region M). The values of β_M for coding and noncoding DNA obtained from the slopes of the fits are 0.03 and 0.21, respectively (from [21])

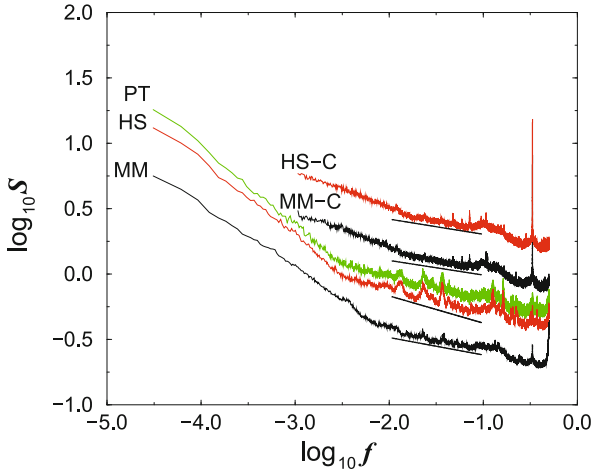

Fractals in Biology, Figure 12

Comparison of the correlation properties of coding and non-coding DNA of different mammals. Shown RY power spectra averaged over all complimentary DNA sequences of *Homo sapiens* (HS-C) and *Mus musculus* (mouse) (MM-C). The complimentary DNA sequences are obtained from messenger RNA by reverse transcriptase and thus lack non-coding elements. They are characterized by huge peaks at $f = 1/3$, corresponding to the inverse codon length 3 bp. The power spectra for human and mouse are almost indistinguishable. Also shown RY power spectra of large continuously sequenced segments of chromosomes (contigs) of about 10^7 bp long for mouse (MM), human (HS) and chimpanzee (PT, Pan troglodytes). Their power spectra have different high frequency peaks absent in the coding DNA power spectra: a peak at $f = 1/2$, corresponding to simple repeats and several large peaks in the range from $f = 1/3$ to $f = 1/100$ corresponding to interspersed repeats. Notice that the magnitudes of these peaks are similar for humans and chimpanzees (although for humans they are slightly larger, especially the peak at 80 bp corresponding to the long interspersed repeats) and much larger than those of mouse. This means that mouse has much smaller number of the interspersed repeats than primates. On the other hand, mouse has much larger fraction of dimeric simple repeats indicated by the peak at $f = 1/2$

tal correlations of DNA sequences and packing of DNA molecules was suggested in [52,53].

An intriguing property of the DNA of multicellular organisms is that an overwhelming portion of it (97% in case of humans) is not used for coding proteins. It is interesting that the percent of non-coding DNA increases with the complexity of an organism. Bacteria practically do not have non-coding DNA, and Yeast has only 30% of it. The coding sequences form genes ($\approx 10^4$ bp) which carry information for one protein. The genes of multicellular organisms are broken by many noncoding intervening sequences (introns). Only exons which are short ($\approx 10^2$ bp) coding sequences located between introns ($\approx 10^3$ bp) are eventually translated into a protein. The genes themselves are separated by very long intergenic sequences $\approx 10^5$ bp. Thus the coding structure of DNA resembles a Cantor set.

The purpose and properties of coding DNA are well understood. Each three consequent bp form a codon, which is translated into one amino acid of the protein sequence. Accordingly, the power spectrum computed for the coding DNA has a characteristic peak at $f = 1/3$ corresponding to the inverse codon length (Figs. 11 and 12). Coding DNA is highly conserved and the power spectra of coding DNA of different organisms are very similar and in case of mammals are indistinguishable (Fig. 12).

The properties of noncoding DNA are very different. Non coding DNA contains a lot of useful information including protein binding sites controlling gene transcription and expression and other regulatory sequences. However its overwhelming fraction lacks any known purpose. This "junk" DNA is full of simple repeats such as CACACACA… as well as the interspersed repeats or retroposons which are virus-like sequences inserting themselves in great number of copies into the intergenic DNA. It is a great challenge of molecular biology and genetics to understand the meaning of non-coding DNA and even to learn how to manipulate it. It would be very interesting to create transgenic animals without non-coding DNA and test if their phenotype will differ from that of the wild type species. The non-coding DNA even for very closely

related species can significantly differ (Fig. 12). The length of simple repeats varies even for close relatives. That is why simple repeats are used in forensic studies.

The power spectra of non-coding DNA significantly differ from those of coding DNA. Non-coding DNA does not have a peak at $f = 1/3$. The presence of simple repeats make non-coding DNA more correlated than coding DNA on a scale from 10 to 100 bp [21] (Fig. 11). This difference observed in 1992 [70,96] lead to the hypothesis that noncoding DNA is governed by some mutation-duplication stochastic process which creates long-range (fractal) correlations, while the coding DNA lacks long-range correlations because it is highly conserved. Almost any mutation which happens in coding DNA alter the sequence of the corresponding protein and thus may negatively affect its function and lead to a non-viable or less fit organism. Researchers have proposed using the difference in the long-range correlations to find the coding sequences in the sea of non-coding DNA [91]. However this method appears to be unreliable and today the non-coding sequences are found with much greater accuracy by the bioinformatics methods based on sequence similarity to the known proteins.

Bioinformatics has developed powerful methods for comparing DNA of different species. Even a few hundred bp stretch of the mitochondrial DNA of a Neanderthal man can tell that Neanderthals diverged from humans about 500 000 years ago [65]. The power spectra and other correlation methods such as DFA or wavelet analysis does not have such an accuracy. Never the less, power spectra of the large stretches of DNA carry important information on the evolutionary processes in the DNA. They are similar for different chromosomes of the same organism, but differ even for closely related species (Fig. 12). In this sense they can play a role similar to the role played by X-ray or Raman spectra for chemical substances. A quick look at them can tell a human and a mouse and even a human and a monkey apart. Especially interesting is the difference in peak heights produced by different interspersed repeats. The height of these peaks is proportional to the number of the copies of the interspersed repeats. According to this simple criterion, the main difference between humans and chimps is the insertion of hundreds of thousands of extra copies of interspersed repeats into human DNA [59].

Future Directions

All the above examples show that there are no rigorous fractals in biology. First of all, there are always a lower and an upper cutoff of the fractal behavior. For example, a polymer in a good solvent which is probably the most rigorous of all real fractals in biology has the lower cutoff corresponding to the persistence length comprised of a few monomers and the upper cutoff corresponding to the length of the entire polymer or to the size of the compartment in which it is confined.

For the hierarchical structures such as trees and lungs, in addition to the obvious lower and upper cutoffs, the branching pattern changes from one generation to the next. In the DNA, the packing principles employ different mechanisms on the different levels of packing. In bacterial colonies, the lower cutoff is due to some tendency of bacteria to clump together analogous to surface tension, and the upper cutoff is due to the finite concentration of the nutrients, which originally are uniformly distributed on the Petri dish. In the ideal DLA case, the concentration of the diffusing particles is infinitesimally small, so at any given moment of time there is only one particle in the vicinity of the aggregate.

The temporal physiological series are also not exactly self-similar, but are strongly affected by the daily schedule with a characteristic frequency of 24 h and shorter overtones. Some of these signals can be described with help of multifractals.

Fractals in ecology are limited by the topographical features of the land which may be also fractal due to complex geological processes, so it is very difficult to distinguish whether certain features are caused by biology or geology. The measurements of the fractal dimension are hampered by the lack of statistics, the noise in the image or signal, and by the crossovers due to intrinsic features of the system which is differently organized on different length scales. So very often the mosaic organization of the system with patches of several fixed length scales can be mistakenly identified with the fractal behavior. Thus the use of fractal dimension or Hurst exponent for diagnostics or distinguishing some parts of the system from one another has a limited value. After all, the fractal dimension is only one number which is usually obtained from a slope of a least square linear fit of a log–log graph over a subjectively identified range of values. Other features of the power spectrum, such as peaks at certain characteristic frequencies may have more biological information than fractal dimension. Moreover, the presence of certain fractal behavior may originate from simple physical principles, while the deviation from it may indicate the presence of a nontrivial biological phenomenon.

On the other hand, fractal geometry is an important concept which can be used for qualitative understanding of the mechanisms behind certain biological processes. The use of similar organization principles on different length scales is a remarkable feature, which is certainly em-

ployed by nature to design the shape and behavior of living organisms.

Though fractals themselves may have a limited value in biology, the theory of complex systems in which they often emerge continues to be a leading approach in understanding life. One of the most impelling challenges of modern interdisciplinary science which involves biology, chemistry, physics, mathematics, computer science, and bioinformatics is to build a comprehensive theory of morphogenesis. Such a great mind as de Gennes turned to this subject in his late years [35,36,37]. In these studies the theory of complex networks [2,12,48,67,90] which describe interactions of biomolecules with many complex positive and negative feedbacks will take a leading part.

Challenges of the same magnitude face researches in neuroscience, behavioral science, and ecology. Only complex interdisciplinary approach involving the specialists in the theory of complex systems may lead to new breakthroughs in this field.

Bibliography

1. Aiello L, Dean C (1990) An Introduction to Human Evolutionary Anatomy. Academic Press, London
2. Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
3. Alencar AM, Buldyrev SV, Majumdar A, Stanley HE, Suki B (2003) Perimeter growth of a branched structure: application to crackle sounds in the lung. Phys Rev E 68:11909
4. Arneodo A, Bacry E, Graves PV, Muzy JF (1995) Characterizing long-range correlations in DNA sequences from wavelet analysis. Phys Rev Lett 74:3293–3296
5. Arneodo A, D'Aubenton-Carafa Y, Audit B, Bacry E, Muzy JF, Thermes C (1998) What can we learn with wavelets about DNA sequences. Physica A 249:439–448
6. Ashkenazy Y, Hausdorff JM, Ivanov PC, Stanley HE (2002) A stochastic model of human gait dynamics. Physica A 316:662–670
7. Bak P, Sneppen K (1993) Punctuated equilibrium and criticality in a simple model of evolution. Phys Rev Lett 71:4083–4086
8. Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality: an explanation of 1/*f* noise. Phys Rev Lett 59:381–384
9. Banavar JR, Green JL, Harte J, Maritan A (1999) Finite size scaling in ecology. Phys Rev Lett 83:4212–4214
10. Banavar JR, Damuth J, Maritan A, Rinaldo A (2002) Supply-demand balance and metabolic scaling. Proc Natl Acad Sci USA 99:10506–10509 (2002)
11. Banavar JR, Damuth J, Maritan A, Rinaldo A (2003) Allometric cascades. Nature 421:713–714
12. Barabási A-L (2005) The origin of bursts and heavy tails in human dynamics. Nature 435:207–211
13. Barabási A-L, Stanley HS (1995) Fractal Concepts in Surface Growth. Cambridge University Press, Cambridge
14. Bassingthwaighte JB, Liebovitch L, West BJ (1994) Fractal Physiology. Oxford University Press, Oxford
15. Ben Jacob E, Aharonov Y, Shapira Y (2005) Bacteria harnessing complexity. Biofilms 1:239–263
16. Bernaola-Galvan P, Ivanov PC, Amaral LAN, Stanley HE (2001) Scale invariance in the nonstationarity of human heart rate. Phys Rev Lett 87:168105
17. Bhaskar KR, Turner BS, Garik P, Bradley JD, Bansil R, Stanley HE, LaMont JT (1992) Viscous fingering of HCl through gastric mucin. Nature 360:458–461
18. Bishop M, Michels JPJ (1985) The shape of ring polymers. J Chem Phys 82:1059–1061
19. Buldyrev SV (2006) Power Law Correlations in DNA Sequences. In: Koonin EV, Karev G, Wolf Yu (eds) Power Laws, Scale-Free Networks and Genome Biology. Springer, Berlin, pp 123–164
20. Buldyrev SV, Goldberger AL, Havlin S, Peng C-K, Stanley HE (1994) Fractals in Biology and Medicine: From DNA to the Heartbeat. In: Bunde A, Havlin S (eds) Fractals in Science. Springer, Berlin, pp 48–87
21. Buldyrev SV, Goldberger AL, Havlin S, Mantegna RN, Matsa ME, Peng C-K, Simons M, Stanley HE (1995) Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. Phys Rev E 51:5084–5091
22. Buldyrev SV, Cruz L, Gomez-Isla T, Havlin S, Stanley HE, Urbanc B, Hyman BT (2000) Description of microcolumnar ensembles in association cortex and their disruption in alzheimer and lewy body dementia. Proc Natl Acad Sci USA 97:5039–5043
23. Buldyrev SV, Havlin S, Kazakov AY, da Luz MGE, Raposo EP, Stanley HE, Viswanathan GM (2001) Average time spent by levy flights and walks on an interval with absorbing boundaries. Phys Rev E 64:041108
24. Buldyrev SV, Gitterman M, Havlin S, Kazakov AY, da Luz MGE, Raposo EP, Stanley HE, Viswanathan GM (2001) Properties of Levy fights on an interval with absorbing boundaries. Physica A 302:148–161
25. Buldyrev SV, Pammolli F, Riccaboni M, Yamasaki K, Fu D-F, Matia K, Stanley HE (2007) Generalized preferential attachment model for business firms growth rates II. Eur Phys J B 57: 131–138
26. Bunde A, Havlin S (eds) (1996) Fractals and Disordered Systems, 2nd edn. Springer, New York
27. Bunde A, Havlin S, Kantelhardt J, Penzel T, Peter J-H, Voigt K (2000) Correlated and uncorrelated regions in heart-rate fluctuations during sleep. Phys Rev Lett 85:3736–3739
28. Calder WA 3rd (1984) Size, Function and Life History. Harvard University Press, Cambridge
29. Caserta F, Eldred WD, Fernández E, Hausman RE, Stanford LR, Bulderev SV, Schwarzer S, Stanley HE (1995) Determination of fractal dimension of physiologically characterized neurons in two and three dimensions. J Neurosci Meth 56:133–144
30. Clar S, Drossel B, Schwabl F (1996) Self-organized criticality in forest-fire models and elsewhere. Review article. J Phys C 8:6803
31. D'Arcy WT, John TB (ed) (1992) On Growth and Form. Cambridge University Press, Cambridge
32. Darveau C-A, Suarez RK, Andrews RD, Hochachka PW (2002) Allometric cascade as a unifying principle of body mass effects on metabolism. Nature 417:166–170
33. Darveau C-A, Suarez RK, Andrews RD, Hochachka PW (2003) Allometric cascades – Reply. Nature 421:714–714
34. De Gennes PG (1979) Scaling Concepts in Polymer Physics. Cornell University Press, Ithaca

35. De Gennes PG (2004) Organization of a primitive memory: Olfaction. Proc Natl Acad Sci USA 101:15778–15781
36. De Gennes PG (2007) Collective neuronal growth and self organization of axons. Proc Natl Acad Sci USA 104:4904–4906
37. De Gennes PG, Puech PH, Brochard-Wyart F (2003) Adhesion induced by mobile stickers: A list of scenarios. Langmuir 19:7112–7119
38. Devaney RL (1989) An Introduction to Chaotic Dynamical Systems, 2nd edn. Addison-Wesley, Redwood City
39. Edwards AM, Phillips RA, Watkins NW, Freeman MP, Murphy EJ, Afanasyev V, Buldyrev SV, da Luz MGE, Raposo EP, Stanley HE, Viswanathan GM (2007) Revisiting Levy flight search patterns of wandering albatrosses, bumblebees and deer. Nature 449:1044–1047
40. Falconer K (2003) Fractal Geometry: Mathematical Foundations and Applications. Wiley, Hoboken
41. Feder J (1988) Fractals. Plenum, New York
42. Feller W (1970) An introduction to ptobability theory and its applications, vol 1–2. Wiley, New York
43. Fleury V, Gouyet J-F, Leonetti M (eds) (2001) Branching in Nature: Dynamics and Morphogenesis. Springer, Berlin
44. Flory PJ (1955) Principles of Polymer Chemistry. Cornell University Press, Ithaca
45. Fortin MJ, Dale MRT (2005) Spatial analysis: a guide for ecologists. Cambridge Univ Press, Cambridge
46. Gerstein GL, Mandelbrot B (1964) Random walk models for the spike activity of a single neuron. Biophys J 4:41–68
47. Goldberger AL, Amaral LAN, Hausdorff JM, Ivanov PC, Peng C-K, Stanley HE (2002) Fractal dynamics in physiology: Alterations with disease and aging. Proc Natl Acad Sci USA 99:2466–2472
48. Gonzalez MC, Barabási A-L (2007) Complex networks – From data to models. Nature Phys 3:224–225
49. Gonzalez MC, Hidalgo CA, Barabási A-L (2008) Understanding individual human mobility patterns. Nature 453:779–782
50. Grassberger P, Procaccia I (1983) Measuring the strangeness of strange attractors. Physica D 9:189–208
51. Grey F, Kjems JK (1989) Aggregates, broccoli and cauliflower. Physica D 38:154–159
52. Grosberg A, Rabin Y, Havlin S, Neer A (1993) Crumpled globule model of three-dimensional structure of DNA. Europhys Lett 23:373–378
53. Grosberg A, Rabin Y, Havlin S, Neer A (1993) Self-similarity in the structure of DNA: why are introns needed? Biofizika 38:75–83
54. Halley JM, Hartley S, Kallimanis AS, Kunin WE, Lennon JJ, Sgardelis SP (2004) Uses and abuses of fractal methodology in ecology. Ecol Lett 7:254–271
55. Hausdorff JM, Ashkenazy Y, Peng CK, Ivanov PC, Stanley HE, Goldberger AL (2001) When human walking becomes random walking: Fractal analysis and modeling of gait rhythm fluctuations. Physica A 302:138–147
56. Horsfield K, Thurlbeck A (1981) Relation between diameter and flow in branches of the bronchial tree. Bull Math Biol 43:681–691
57. Hu K, Ivanov PC, Hilton MF, Chen Z, Ayers RT, Stanley HE, Shea SA (2004) Endogenous circadian rhythm in an index of cardiac vulnerability independent of changes in behavior. Proc Natl Acad Sci USA 101:18223–18227
58. Hughes BD (1995) Random Walks and Random Environments, vol 1: Random Walks. Clarendon Press, Oxford
59. Hwu RH, Roberts JW, Dawidson EH, et al. (1986) Insertion and/or deletion of many repeated DNA sequences in human and higher apes evolution. Proc Natl Acad Sci 83:3875–3879
60. Karlin S, Brandel V (1993) Patchiness and correlations in DNA sequences. Science 259:677–680
61. Kenah E, Robins MJ (2007) Second look at the spread of epidemics on networks. Phys Rev E 76:036113
62. Kermack WO, McKendrick AG (1927) Contribution to the Mathematical Theory of Epidemics. Proc R Soc A 115:700–721
63. Khokhlov AR, Grosberg AY (2002) Statistical physics of macromolecules. AIP, Woodbury
64. Kim S-H (2005) Fractal structure of a white cauliower. J Korean Phys Soc 46:474–477
65. Kings M, Geisert H, Schmitz RW, Krainitzki H, Pääbo S (1999) DNA sequence of the mitochondrial hypervariable region II from the neandertal type specimen. Proc Natl Acad Sci USA 96:5581–5585
66. Kitaoka H, Suki B (1997) Branching design of the bronchial tree based on a diameter-flow relationship. J Appl Physiol 82:968–976
67. Koonin EV, Karev G, Wolf Yu (eds) (2006) Power Laws, Scale-Free Networks and Genome Biology. Springer, Berlin
68. Larralde H, Trunfio PA, Havlin S, Stanley HS, Weiss GH (1992) Territory covered by N diffusing particles. Nature 355: 423–426
69. Li W (1997) The study of correlation structures of DNA sequences: a critical review. Comput Chem 21:257–271
70. Li W, Kaneko K (1992) Long-range correlation and partial 1/*F*-alpha spectrum in a noncoding DNA-sequence. Europhys Lett 17:655–660
71. Liljeros F, Edling CR, Amaral LAN, Stanley HE, Aberg Y (2001) The web of human sexual contacts. Nature 411:907–908
72. Lo C-C, Amaral LAN, Havlin S, Ivanov PC, Penzel T, Peter J-H, Stanley HE (2002) Dynamics of sleep-wake transitions during sleep. Europhys Lett 57:625–631
73. Lo C-C, Chou T, Penzel T, Scammell T, Strecker RE, Stanley HE, Ivanov PC (2004) Common scale-invariant pattern of sleep-wake transitions across mammalian species. Proc Natl Acad Sci USA 101:17545–17548
74. Losa GA, Merlini D, Nonnenmacher TF, Weibel ER (eds) (1998) Fractals in Biology and Medicine, vol II. Birkhäuser Publishing, Berlin
75. Losa GA, Merlini D, Nonnenmacher TF, Weibel ER (eds) (2002) Fractals in Biology and Medicine, vol III. Birkhäuser Publishing, Basel
76. Losa GA, Merlini D, Nonnenmacher TF, Weibel ER (eds) (2005) Fractals in Biology and Medicine, vol IV. Birkhäuser Publishing, Basel
77. Majumdar A, Alencar AM, Buldyrev SV, Hantos Z, Lutchen KR, Stanley HE, Suki B (2005) Relating airway diameter distributions to regular branching asymmetry in the lung. Phys Rev Lett 95:168101
78. Malamud BD, Morein G, Turcotte DL (1998) Forest fires: an example of self-organized critical behavior. Science 281: 1840–1841
79. Mandelbrot BB (1982) The Fractal Geometry of Nature. Freeman WH and Co., New York
80. Mark DM (1984) Fractal dimension of a coral-reef at ecological scales – a discussion. Mar Ecol Prog Ser 14:293–294
81. May RM (1975) In: Cody ML, Diamond JM (eds) Ecology

and Evolution of Communities. Belknap Press, Cambridge, pp 81–120
82. May RM (1976) Simple mathematical models with very complicated dynamics. Nature 261:459–467
83. Meakin P (1998) Fractals, Scaling and Growth Far from Equilibrium. Cambridge University Press, Cambridge
84. Menshutin AY, Shchur LN, Vinokur VM (2007) Probing surface characteristics of diffusion-limited-aggregation clusters with particles of variable size. Phys Rev E 75:010401
85. Metzger RJ, Krasnow MA (1999) Genetic Control of Branching Morphogenesis Science 284:1635–1639
86. Newman MEJ (2002) Spread of epidemic disease on networks. Phys Rev E 66:016128
87. Niklas KJ (2007) Sizing up life and death. PNAS 104:15589–15590
88. Nonnenmacher TF, Losa GA, Weibel ER (eds) (1994) Fractals in Biology and Medicine, vol I. Birkhäuser Publishing, Basel
89. Olami Z, Feder HJS, Christensen K (1992) Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes. Phys Rev Lett 68:1244–1247
90. Oliveira JG, Barabási A-L (2005) Human dynamics: Darwin and Einstein correspondence patterns. Nature 437:1251–1251
91. Ossadnik SM, Buldyrev SV, Goldberger AL, Havlin S, Mantegna RN, Peng C-K, Simons M, Stanley HE (1994) Correlation approach to identify coding regions in DNA sequences. Biophys J 67:64–70
92. Paczuski M, Maslov S, Bak P (1996) Avalanche dynamics in evolution, growth, and depinning models. Phys Rev E 53: 414–343
93. Panico J, Sterling P (1995) Retinal neurons and vessels are not fractal but space-filling. J Comparat Neurol 361:479–490
94. Peitgen H-O, Saupe D (ed) (1988) The Science of Fractal Images. Springer, Berlin
95. Peitgen H-O, Jiirgens H, Saupe D (1992) Chaos and Fractals. Springer, New York
96. Peng C-K, Buldyrev SV, Goldberger A, Havlin S, Sciortino F, Simons M, Stanley HE (1992) Long-range correlations in nucleotide sequences. Nature 356:168–171
97. Peng C-K, Mietus J, Hausdorff JM, Havlin S, Stanley HE, Goldberger AL (1993) Long-range anticorrelations and non-gaussian behavior of the heartbeat. Phys Rev Lett 70:1343–1347
98. Peng C-K, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL (1994) Mosaic organization of DNA nucleotides. Phys Rev E 49:1685–1689
99. Peng C-K, Buldyrev SV, Hausdorff JM, Havlin S, Mietus JE, Simons M, Stanley HE, Goldberger AL (1994) Non-equilibrium dynamics as an indispensable characteristic of a healthy biological system. Integr Physiol Behav Sci 29:283–293
100. Peng C-K, Havlin S, Stanley HE, Goldberger AL (1995) Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos 5:82–87
101. Pocock MJO, Hartley S, Telfer MG, Preston CD, Kunin WE (2006) Ecological correlates of range structure in rare and scarce British plants. J Ecol 94:581–596
102. Ramchandani R, Bates JHT, Shen X, Suki B, Tepper RS (2001) Airway branching morphology of mature and immature rabbit lungs. J Appl Physiol 90:1584–1592
103. Ramos-Fernandez G, Meteos JL, Miramontes O, Cocho G, Larralde H, Ayala-Orozco B (2004) Lévy walk patterns in the foraging movements of spider monkeys (Ateles geoffroyi). Behav Ecol Sociobiol 55:223–230
104. Redner S (2001) A Guide to First-Passage Processes. Cambridge University Press, Cambridge
105. Reich PB, Tjoelker MG, Machado J-L, Oleksyn J (2006) Universal scaling of respiratory metabolism, size and nitrogen in plants. Nature 439:457–461
106. Reiss MJ (2006) Allometry of Growth and Reproduction. Cambridge University Press, Cambridge
107. Richter JP (ed) (1970) The Notebooks of Leonardo da Vinci. Dover Publications, New York
108. Rosenzweig ML (1995) Species Diversity in Space and Time. Cambridge University Press, Cambridge
109. Santullia A, Telescaa L (2005) Time-clustering analysis of forest-fire sequences in southern Italy Rosa Lasaponaraa. Chaos Solitons Fractals 24:139–149
110. Savage VM, Allen AP, Brown JH, Gillooly JF, Herman AB, Woodruff WH, West GB (2007) Scaling of number, size, and metabolic rate of cells with body size in mammals. Proc Natl Acad Sci 104:4718–4723
111. Schenk K, Drossel B, Schwabl F (2002) The self-organized critical forest-fire model on large scales. Phys Rev E 65:026135
112. Schmidt-Nielsen K (1984) Scaling: Why is Animal Size so Important? Cambridge University Press, Cambridge
113. Schulte-Frohlinde V, Ashkenazy Y, Ivanov PC, Glass L, Goldberger AL, Stanley HE (2001) Noise effects on the complex patterns of abnormal heartbeats. Phys Rev Lett 87:068104
114. Scott FG (2006) Developmental Biology, 8th edn. Sinauer Associates, Sunderland
115. Shao J, Buldyrev SV, Cohen R, Kitsak M, Havlin S, Stanley HE (2008) Fractal boundaries of complex networks. Preprint
116. Shlesinger MF (1986) In: Stanley HE, Ostrowsky N (eds) On Growth and Form. Nijhoff, Dordrecht
117. Shlesinger MF, West BJ (1991) Complex fractal dimension of the bronchial tree. Phys Rev Lett 67:2106–2109
118. Shlesinger MF, Zaslavsky G, Frisch U (eds) (1995) Lévy Flights and Related Topics in Physics. Springer, Berlin
119. Sims DW et al (2008) Scaling laws in marine predator search behavior. Nature 451:1098–1102
120. Song C, Havlin H, Makse H (2005) Self-similarity of complex networks. Nature 433:392–395
121. Song C, Havlin S, Makse HA (2006) Origins of fractality in the growth of complex networks. Nature Phys 2:275–281
122. Sornette D (2003) Critical Phenomena in Natural Sciences. Chaos, Fractals, Selforganization and Disorder: Concepts and Tools, 2nd edn. Springer, Berlin
123. Sprott JC (2003) Chaos and Time-Series Analysis. Oxford University Press
124. Stanley HE (1971) Introduction to Phase Transitions and Critical Phenomena. Oxford University Press, New York
125. Stanley HE, Ostrowsky N (eds) (1986) On Growth and Form. Nijhoff, Dordrecht
126. Stanley HE, Buldyrev SV, Goldberger AL, Goldberger ZD, Havlin S, Mantegna RN, Ossadnik SM, Peng C-K, Simons M (1994) Statistical mechanics in biology – how ubiquitous are long-range correlations. Physica A 205:214–253
127. Stauffer D, Aharony A (1992) Introduction to percolation theory. Taylor & Francis, Philadelphia
128. Suki B, Barabási A-L, Hantos Z, Petak F, Stanley HE (1994) Avalanches and power law behaviour in lung inflation. Nature 368:615–618
129. Suki B, Alencar AM, Frey U, Ivanov PC, Buldyrev SV, Majumdar A, Stanley HE, Dawson CA, Krenz GS, Mishima M (2003) Fluc-

tuations, noise, and scaling in the cardio-pulmonary system. Fluct Noise Lett 3:R1–R25
130. Turcotte DL (1997) Fractals and Chaos in Geology and Geophysics. Cambridge University Press, Cambridge
131. Viscek T, Shlesinger MF, Matsushita M (eds) (1994) Fractals in Natural Sciences. World Scientific, New York
132. Viswanathan GM, Afanasyev V, Buldyrev SV, Murphy EJ, Prince PA, Stanley HE (1996) Levy flight search patterns of wandering albatrosses. Nature 381:413–415
133. Viswanathan GM, Buldyrev SV, Havlin S, da Luz MGE, Raposo E, Stanley HE (1999) Optimizing the success of random searches. Nature 401:911–914
134. Vogel H (1979) A better way to construct the sunflower head. Math Biosci 44:179–189
135. Voss RF, Clarke J (1975) 1/*f* noise in music and speech. Nature 258:317–318
136. Voss RF, Clarke J (1978) 1/*f* noise in music: Music from 1/*f* noise. J Acoust Soc Am 63:258–263
137. Voss RF, Wyatt J (1993) Multifractals and the Local Connected Fractal Dimension: Classification of Early Chinese Landscape Paintings. In: Crilly AJ, Earnshaw RA, Jones H (eds) Applications of Fractals and Chaos. Springer, Berlin
138. Warburton D, Schwarz M, Tefft D, Flores-Delgado F, Anderson KD, Cardoso WV (2000) The molecular basis of lung morphogenesis. Mech Devel 92:55–81
139. Watson JD, Gilman M, Witkowski J, Zoller M (1992) Recombinant DNA. Scientific American Books, New York
140. Weibel ER (2000) Symmorphosis: On Form and Function in Shaping Life. Harvard University Press, Cambridge
141. Weiss GH (1994) Aspects and applications of the random walk. North-Holland, New York
142. West GB, Brown JH, Enquist BJ (1997) A general model for the origin of allometric scaling laws in biology. Science 276: 122–126
143. West GB, Woodruff WH, Brown JH (2002) Allometric scaling of metabolic rate from molecules and mitochondria to cells and mammals. Proc Natl Acad Sci 99:2473–2478
144. West GB, Savage VM, Gillooly J, Enquist BJ, Woodruff WH, Brown JH (2003) Why does metabolic rate scale with body size? Nature 421:713–713
145. Witten TA, Sander LM (1981) Phys Rev Lett 47:1400–1404
146. Zipf GK (1949) Human Behavior and the Principle of Least–Effort. Addison-Wesley, Cambridge
147. Zolotarev VM, Uchaikin VV (1999) Chance and Stability: Stable Distributions and their Applications. VSP BV, Utrecht

Fractals and Economics

Misako Takayasu[1], Hideki Takayasu[2]
[1] Tokyo Institute of Technology, Tokyo, Japan
[2] Sony Computer Science Laboratories Inc, Tokyo, Japan

Article Outline

Glossary
Definition of the Subject
Introduction
Examples in Economics
Basic Models of Power Laws
Market Models
Income Distribution Models
Future Directions
Bibliography

Glossary

Fractal An adjective or a noun representing complex configurations having scale-free characteristics or self-similar properties. Mathematically, any fractal can be characterized by a power law distribution.

Power law distribution For this distribution the probability density is given by a power law, $p(r) = c \cdot r^{-\alpha-1}$, where c and α are positive constants.

Foreign exchange market A free market of currencies, exchanging money in one currency for other, such as purchasing a United States dollar (USD) with Japanese yen (JPY). The major banks of the world are trading 24 hours and it is the largest market in the world.

Definition of the Subject

Market price fluctuation was the very first example of fractals, and since then many examples of fractals have been found in the field of Economics. Fractals are everywhere in economics. In this article the main attention is focused on real world examples of fractals in the field of economics, especially market properties, income distributions, money flow, sales data and network structures. Basic mathematics and physics models of power law distributions are reviewed so that readers can start reading without any special knowledge.

Introduction

Fractal is the scientific word coined by B.B. Mandelbrot in 1975 from the Latin word *fractus*, meaning "fractured" [25]. However, *fractal* does not directly mean fracture itself. As an image of a fractal Fig. 1 shows a photo of fractured pieces of plaster fallen on a hard floor. There are several large pieces, many middle size pieces and countless fine pieces. If you have a microscope and observe a part of floor carefully then you will find in your vision several large pieces, many small pieces and countless fine pieces, again in the microscopic world. Such scale-invariant nature is the heart of the fractal. There is no explicit definition on the word fractal, it generally means a complicated scale-invariant configuration.

Scale-invariance can be defined mathematically [42]. Let $P(\geq r)$ denote the probability that the diameter of

Fractals and Economics, Figure 1
Fractured pieces of plaster fallen on a hard floor (provided by H. Inaoka)

a randomly chosen fractured piece is larger than r, then this distribution is called scale-invariant if this function satisfies the following proportional relation for any positive scale factor λ in a considering scale range:

$$P(\geq \lambda r) \propto P(\geq r)\,. \tag{1}$$

The proportional factor should be a function of λ, so we can re-write Eq. (1) as

$$P(\geq \lambda r) = C(\lambda)P(\geq r)\,. \tag{2}$$

Assuming that $P(\geq r)$ is a differentiable function, and differentiate Eq. (2) by λ, and then let $\lambda = 1$.

$$rP'(\geq r) = C'(1)P(\geq r) \tag{3}$$

As $C'(1)$ is a constant this differential equation is readily integrated as

$$P(\geq r) = c_0 r^{C'(1)}\,. \tag{4}$$

$P(\geq r)$ is a cumulative distribution and it is a non-increasing function in general, the exponent $C'(1)$ can be replaced by $-\alpha$ where α is a positive constant. Namely, from the scale-invariance with the assumption of differentiability we have the following power law:

$$P(\geq r) = c_0 r^{-\alpha}\,. \tag{5}$$

The reversed logic also holds, namely for any power law distribution there is a fractal configuration or a scale-invariant state.

In the case of real impact fracture, the size distribution of pieces is experimentally obtained by repeating sieves of various sizes, and it is empirically well-known that a fractured piece's diameter follows a power law with the exponent about $\alpha = 2$ independent of the details about the material or the way of impact [14]. This law is one of the most stubborn physical laws in nature as it is known to hold from 10^{-6} m to 10^5 m, from glass pieces around us to asteroids. From theoretical viewpoint this phenomenon is known to be described by a scale-free dynamics of crack propagation and the universal properties of the exponent value are well understood [19].

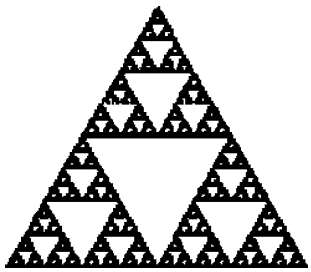

Fractals and Economics, Figure 2
Sierpinski gasket

Usually fractal is considered geometric concept introducing the quantity fractal dimension or the concept of self-similarity. However, in economics there are very few geometric objects, so, the concept of fractals in economics are mostly used in the sense of power law distributions.

It should be noted that any geometrical fractal object accompanies a power law distribution even a deterministic fractal such as Sierpinski gasket. Figure 2 shows Sierpinski gasket which is usually characterized by the fractal dimension D given by

$$D = \frac{\log 3}{\log 2}\,. \tag{6}$$

Paying attention to the distribution of length r of white triangles in this figure, it is easy to show that the probability that a randomly chosen white triangle's side is larger than r, $P(\geq r)$, follows the power law,

$$P(\geq r) \propto r^{-\alpha}\,, \quad \alpha = D = \frac{\log 3}{\log 2}\,. \tag{7}$$

Here, the power law exponent of distribution equals to the fractal dimension; however, such coincidence occurs only when the considering distribution is for a length distribution. For example, in Sierpinski gasket the area s of white triangles follow the power law,

$$P(\geq s) \propto s^{-\alpha}\,, \quad \alpha = \frac{\log 3}{\log 4}\,. \tag{8}$$

The fractal dimension is applicable only for geometric fractals, however, power law distributions are applicable for any fractal phenomena including shapeless quantities. In such cases the power law exponent is the most important quantity for quantitative characterization of fractals.

According to Mandelbrot's own review on his life the concept of fractal was inspired when he was studying economics data [26]. At that time he found two basic properties in the time series data of daily prices of New York cotton market [24]:

(A) Geometrical similarity between large scale chart and an expanded chart.
(B) Power law distribution of price changes in a unit time interval, which is independent of the time scale of the unit.

He thought such scale invariance in both shape and distribution is a quite general property, not only in price charts but also in nature at large. His inspiration was correct and the concept of fractals spread over physics first and then over almost all fields of science. In the history of science it is a rare event that a concept originally born in economics has been spread widely to all area of sciences.

Basic mathematical properties of cumulative distribution can be summarized as follows (here we consider distribution of non-negative quantity for simplicity):

1. $P(\geq 0) = 1$, $P(\geq \infty) = 0$.
2. $P(\geq r)$ is a non-increasing function of r.
3. The probability density is given as $p(r) \equiv -\frac{\mathrm{d}}{\mathrm{d}r}P(\geq r)$. As for power law distributions there are three peculiar characteristics:
4. Difficulty in normalization. Assuming that $P(\geq r) = c_0 r^{-\alpha}$ for all in the range $0 \leq r < \infty$, then the normalization factor c_0 must be 0 considering the limit of $r \to 0$. To avoid this difficulty it is generally assumed that the power law does not hold in the vicinity of $r = 0$. In the case of observing distribution from real data there are naturally lower and upper bounds, so this difficulty should be necessary only for theoretical treatment.
5. Divergence of moments. As for moments defined by $\langle r^n \rangle \equiv \int_0^\infty r^n p(r)\mathrm{d}r$, $\langle r^n \rangle = \infty$ for $n \geq \alpha$. In the special case of $2 \geq \alpha > 0$ the basic statistical quantity, the variance, diverges, $\sigma^2 \equiv \langle r^2 \rangle - \langle r \rangle^2 = \infty$. In the case of $1 \geq \alpha > 0$ even the average can not be defined as $\langle r \rangle = \infty$.
6. Stationary or non-stationary? In view of the data analysis, the above characteristics of diverging moments is likely to cause a wrong conclusion that the phenomenon is non-stationary by observing its averaged value. For example, assume that we observe k samples $\{r_1, r_2, \ldots, r_k\}$ independently from the power law distribution with the exponent, $1 \geq \alpha > 0$. Then, the sample average, $\langle r \rangle_k \equiv \frac{1}{k}\{r_1 + r_2 + \cdots + r_k\}$, is shown to diverge as, $\langle r \rangle_k \propto k^{1/\alpha}$. Such tendency of monotonic increase of averaged quantity might be regarded as a result of non-stationarity, however, this is simply a general property of a power law distribution. The best way to avoid such confusion is to observe the distribution directly from the data.

Other than the power law distribution there is another important statistical quantity in the study of fractals, that is, the autocorrelation. For given time series, $\{x(t)\}$, the autocorrelation is defined as,

$$C(T) \equiv \frac{\langle x(t+T)x(t) \rangle - \langle x(t) \rangle^2}{\langle x(t)^2 \rangle - \langle x(t) \rangle^2}, \tag{9}$$

where $\langle \cdots \rangle$ denotes an average over realizations. The autocorrelation can be defined only for stationary time series with finite variance, in which any statistical quantities do not depend on the location of the origin of time axis.

For any case, the autocorrelation satisfies the following basic properties,

1. $C(0) = 1$ and $C(\infty) = 0$
2. $|C(T)| \leq 1$ for any $T \geq 0$.
3. The Wiener–Khinchin theorem holds, $C(T) = \int_0^\infty S(f) \cos 2\pi f \mathrm{d}f$, where $S(f)$ is the power spectrum defined by $S(f) \equiv \langle \widehat{x}(f)\widehat{x}(-f) \rangle$, with the Fourier transform, $\widehat{x}(f) \equiv \int x(t)\mathrm{e}^{2\pi i f t}\mathrm{d}t$.

In the case that the autocorrelation function is characterized by a power law, $C(T) \propto T^{-\beta}$, $\beta > 0$, then the time series $\{x(t)\}$ is said to have a fractal property, in the sense that the autocorrelation function is scale-independent for any scale-factor, $\lambda > 0$, $C(\lambda T) \propto C(T)$. In the case $1 > \beta > 0$ the corresponding power spectrum is given as $S(f) \propto f^{-1+\beta}$.

The power spectrum can be applied to any time series including non-stationary situations. A simple way of telling non-stationary situation is to check the power law exponent of $S(f) \propto f^{-1+\beta}$ in the vicinity of $f = 0$, for $0 > \beta$ the time series is non-stationary.

Three basic examples of fractal time series are the followings:

1. White noise. In the case that $\{x(t)\}$ is a stationary independent noise, the autocorrelation is given by the Kronecker's function, $C(T) = \delta_T$, where

$$\delta_T = \begin{cases} 1, & T = 0 \\ 0, & T \neq 0. \end{cases}$$

The corresponding power spectrum is $S(f) \propto f^0$. This case is called white noise from an analogy that superposition of all frequency lights with the same amplitude make a colorless white light. White noise is a plausible model of random phenomena in general including economic activities.

2. Random walk. This is defined by summation of a white noise, $X(t) = X(0) + \sum_{s=0}^{t} x(s)$, and the power spectrum is given by $S(f) \propto f^{-2}$. In this case the autocorrelation function can not be defined because the data is non-stationary. Random walks are quite generic models widely used from Brownian motions of colloid to market prices. The graph of a random walk has a fractal property such that an expansion of any part of the graph looks similar to the whole graph.
3. The $1/f$ noise. The boundary of stationary and non-stationary states is given by the so-called $1/f$ noise, $S(f) \propto f^{-1}$. This type of power spectrum is also widely observed in various fields of sciences from electrical circuit noise [16] to information traffics in the Internet [53]. The graph of this $1/f$ noise also has the fractal property.

Examples in Economics

In this chapter fractals observed in real economic activities are reviewed. Mathematical models derived from these empirical findings will be summarized in the next chapter.

As mentioned in the previous chapter the very first example of a fractal was the price fluctuation of the New York cotton market analyzed by Mandelbrot with the daily data for a period of more than a hundred years [24]. This research attracted much attention at that time, however, there was no other good market data available for scientific analysis, and no intensive follow-up research was done until the 1990s. Instead of earnest scientific data analysis artificial mathematical models of market prices based on random walk theory became popular by the name of Financial Technology during the years 1960–1980.

Fractal properties of market prices are confirmed with huge amount of high resolution market data since the 1990s [26,43,44]. This is due to informationization of financial markets in which transaction orders are processed by computers and detail information is recorded automatically, while until the 1980s many people gathered at a market and prices are determined by shouting and screaming which could not be recorded. Now there are more than 100 financial market providers in the world and the number of transacted items exceeds one million. Namely, millions of prices in financial markets are changing with time scale in seconds, and you can access any market price at real time if you have a financial provider's terminal on your desk via the Internet.

Among these millions of items one of the most representative financial markets is the US Dollar-Japanese Yen (USD-JPY) market. In this market Dollar and Yen are exchanged among dealers of major international banks. Unlike the case of stock markets there is no physical trading place, but major international banks are linked by computer networks and orders are emitted from each dealer's terminal and transactions are done at an electronic broking system. Such a broking system and the computer networks are provided by financial provider companies like Reuters.

The foreign exchange markets are open 24 hours and deals are done whenever buy- and sell-orders meet. The minimum unit of a deal is one million USD (called a bar), and about three million bars are traded everyday in the whole foreign exchange markets in which more than 100 kinds of currencies are exchanged continuously. The total amount of money flow is about 100 times bigger than the total amount of daily world trade, so it is believed that most of deals are done not for the real world's needs, but they are based on speculative strategy or risk hedge, that is, to get profit by buying at a low price and selling at a high price, or to avoid financial loss by selling decreasing currency.

In Fig. 3 the price of one US Dollar paid by Japanese Yen in the foreign exchange markets is shown for 13 years [30]. The total number of data points is about 20 million, that is, about 10 thousand per day or the averaged transaction interval is seven seconds. A magnified part of the top figure for one year is shown in the second figure. The third figure is the enlargement of one month in the second figure. The bottom figure is again a part of the third figure, here the width is one day. It seems that at least the top three figures look quite similar. This is one of the fractal properties of market price (A) introduced in the previous chapter. This geometrical fractal property can be found in any market, so that this is a very universal market property.

However, it should be noted that this geometrical fractal property breaks down for very short time scale as typically shown in Fig. 4. In this figure the abscissa is 10 minutes range and we can observe each transaction separately. Obviously the price up down is more zigzag and more discrete than the large scale continuous market fluctuations shown in Fig. 3. In the case of USD-JPY market the time scale that this breakdown of scale invariance occurs typically at time scale of several hours.

The distribution of rate change in a unit time (one minute) is shown in Fig. 5. Here, there are two plots of cu-

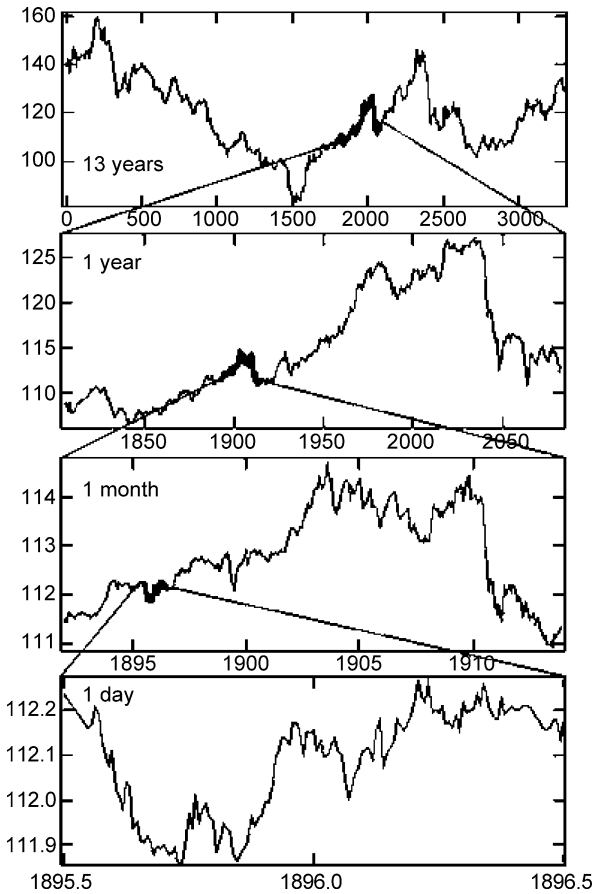

Fractals and Economics, Figure 3
Dollar-Yen rate for 13 years (*Top*). *Dark areas* are enlarged in the following figure [30]

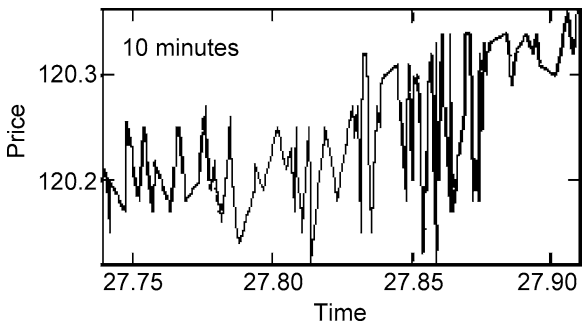

Fractals and Economics, Figure 4
Market price changes in 10 minutes

mulative distributions, $P(> \Delta x)$ for positive rate changes and $P(> |\Delta x|)$ for negative rate changes, which are almost identical meaning that the up-down symmetry of rate changes is nearly perfect. In this log–log plot the estimated power law distribution's exponent is 2.5. In the

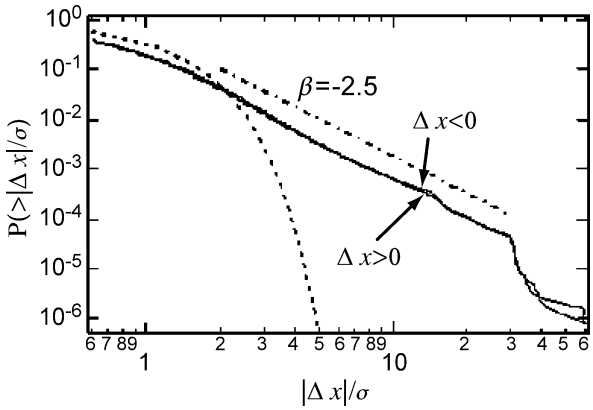

Fractals and Economics, Figure 5
Log–log plot of cumulative distribution of rate change [30]

original finding of Mandelbrot, (B) in the previous chapter, the reported exponent value is about 1.7 for cotton prices. In the case of stock markets power laws are confirmed universally for all items, however, the power exponents are not universal, taking value from near one to near five, typically around three [15]. Also the exponent values change in time year by year.

In order to demonstrate the importance of large fluctuations, Fig. 6 shows a comparison of three market prices. The top figure is the original rate changes for a week. The middle figure is produced from the same data, but it is consisted of rate changes of which absolute values are larger than 2σ, that is, about 5 % of all the data. In the bottom curve such large rate changes are omitted and the residue of 95 % of small changes makes the fluctuations. As known from these figures the middle figure is much closer to the original market price changes. Namely, the contribution from the power law tails of price change distribution is very large for macro-scale market prices.

Power law distribution of market price changes is also a quite general property which can be confirmed for any market. Up-down symmetry also holds universally in short time scale in general, however, for larger unit time the distribution of price changes gradually deforms and for very large unit time the distribution becomes closer to a Gaussian distribution. It should be noted that in special cases of market crashes or bubbles or hyper-inflations the up-down symmetry breaks down and the power law distribution is also likely to be deformed.

The autocorrelation of the time sequence of price changes generally decays quickly to zero, sometimes accompanied by a negative correlation in a very short time. This result implies that the market price changes are apparently approximated by white noise, and market prices

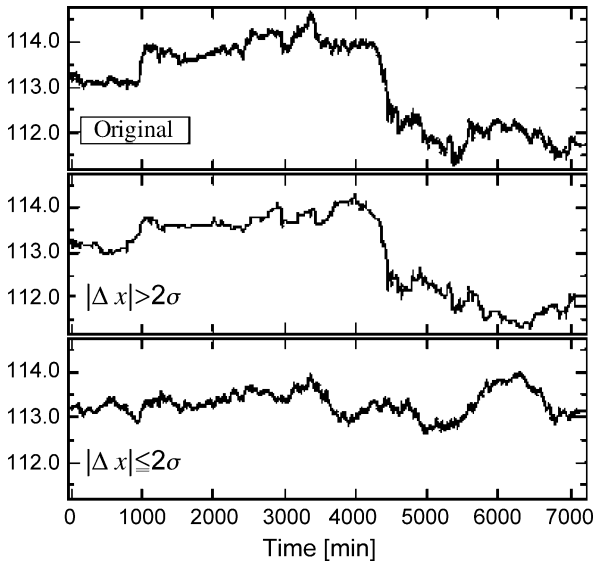

Fractals and Economics, Figure 6
USD-JPY exchange rate for a week (*top*) Rate changes smaller than 2σ are neglected (*middle*) Rate changes larger than 2σ are neglected (*bottom*)

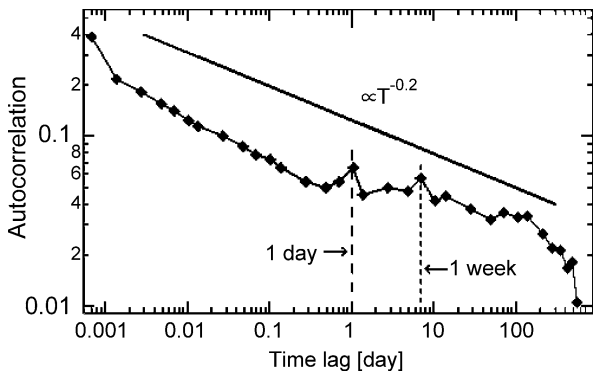

Fractals and Economics, Figure 7
Autocorrelation of volatility [30]

are known to follow nearly a random walk as a result. However, market price is not a simple random walk. In Fig. 7 the autocorrelation of volatility, which is defined by the square of price change, is shown in log–log scale. In the case of a simple random walk this autocorrelation should also decay quickly. The actual volatility autocorrelation nearly satisfies a power law implying that the volatility time series has a fractal clustering property. (See also Fig. 31 representing an example of price change clustering.)

Another fractal nature of markets can be found in the intervals of transactions. As shown in Fig. 8 the transac-

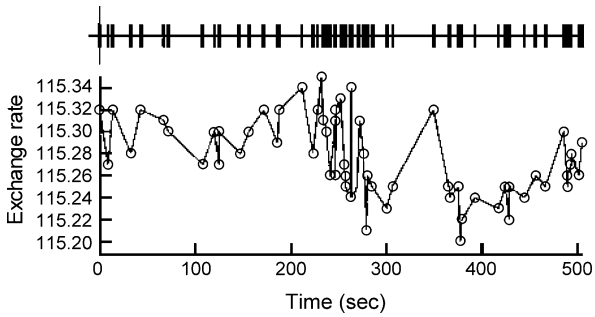

Fractals and Economics, Figure 8
Clustering of transaction intervals

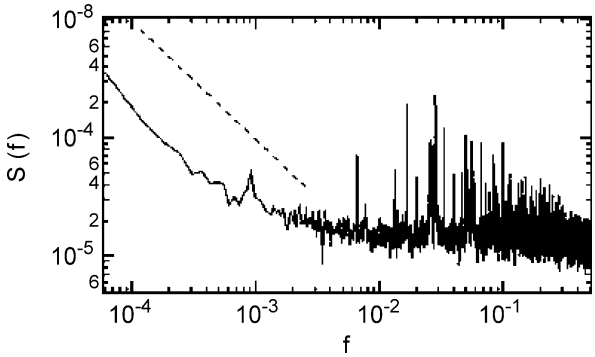

Fractals and Economics, Figure 9
Power spectrum of transaction intervals [50]

tion intervals fluctuate a lot in very short time scale. It is known that the intervals make clusters, namely, shorter intervals tend to gather. To characterize such clustering effect we can make a time sequence consisted of 0 and 1, where 0 denotes no deal was done at that time, and 1 denotes a deal was done. The corresponding power spectrum follows a $1/f$ power spectrum as shown in Fig. 9 [50].

Fractal properties are found not only in financial markets. Company's income distribution is known to follow also a power law [35]. A company's income is roughly given by subtraction of incoming money flow minus outgoing money flow, which can take both positive and negative values. There are about six million companies in Japan and Fig. 10 shows the cumulative distribution of annual income of these companies. Clearly we have a power law distribution of income I with the exponent very close to -1 in the middle size range, so-called the Zipf's law,

$$P(>I) \propto I^{-\beta}\,, \quad \beta = 1\,. \tag{10}$$

Although in each year every company's income fluctuates, and some percentage of companies disappear or are newly born, this power law is known to hold for more than 30

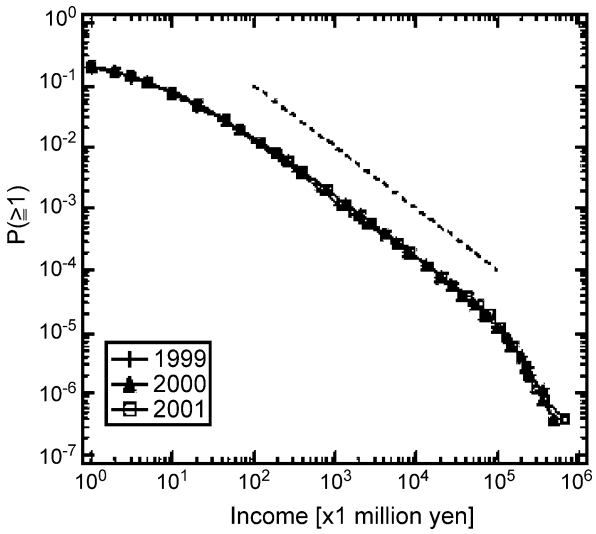

Fractals and Economics, Figure 10
Income distribution of companies in Japan

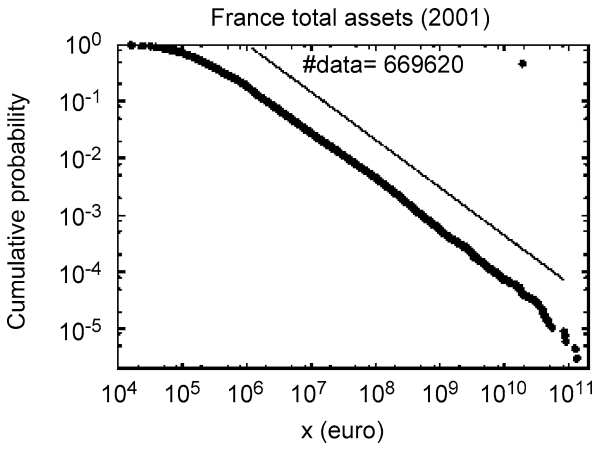

Fractals and Economics, Figure 11
Income distribution of companies in France [13]

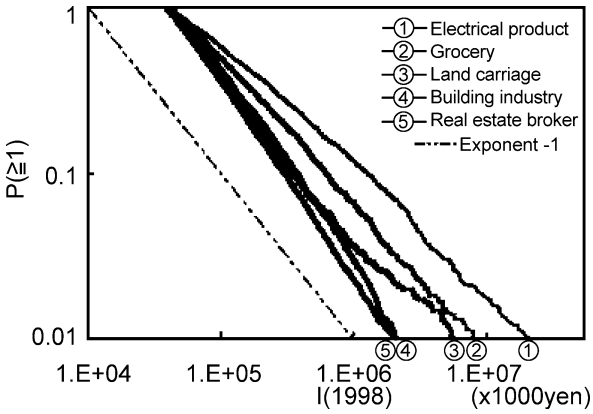

Fractals and Economics, Figure 12
Income distribution of companies in each category [29]

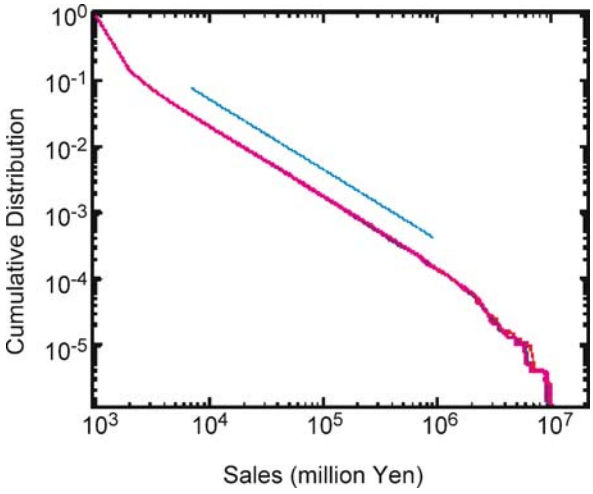

Fractals and Economics, Figure 13
The distribution of whole sales [34]

years. Similar power laws are confirmed in various countries, the case of France is plotted in Fig. 11 [13].

Observing more details by categorizing the companies, it is found that the income distribution in each job category follows nearly a power law with the exponent depending on the job category as shown in Fig. 12 [29]. The implication of this phenomenon will be discussed in Sect. "Income Distribution Models".

A company's size can also be viewed by the amount of whole sale or the number of employee. In Figs. 13 and 14 distributions of these quantities are plotted [34]. In both cases clear power laws are confirmed. The size distribution of debts of bankrupted companies is also known to follow a power law as shown Fig. 15 [12].

A power law distribution can also be found in personal income. Figure 16 shows the personal income distribution in Japan in a log–log plot [1]. The distribution is clearly separated into two parts. The majority of people's incomes are well approximated by a log-normal distribution (the left top part of the graph), and the top few percent of people's income distribution is nicely characterized by a power law (the linear line in the left part of the graph). The majority of people are getting salaries from companies. This type of composite of two distributions is well-known from the pioneering study by Pareto about 100 years ago and it holds in various countries [8,22].

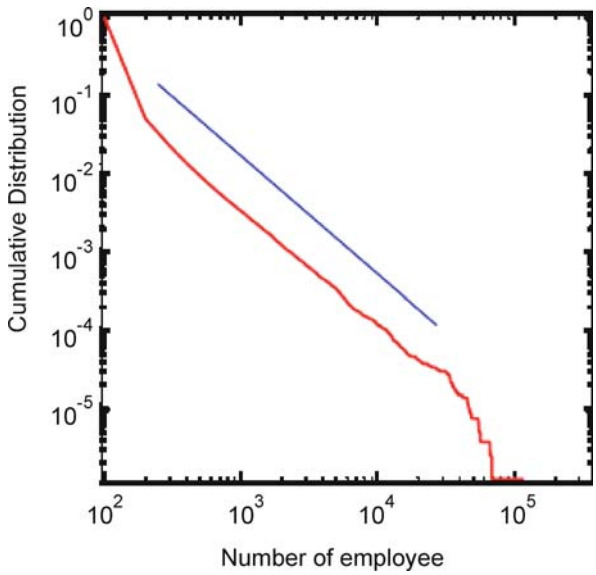

Fractals and Economics, Figure 14
The distribution of employee numbers [34]

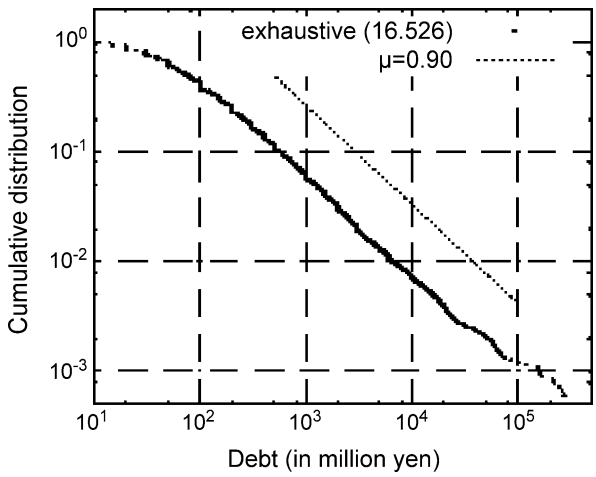

Fractals and Economics, Figure 15
The size distribution of debts of bankrupted companies [12]

A typical value of the power exponent is about two, significantly larger than the income distribution of companies. However, the exponent of the power law seems to be not universal and the value changes county by country or year by year. There is a tendency that the exponent is smaller, meaning more rich people, when the economy is improving [40].

Another fractal in economics can be found in a network of economic agents such as banks' money transfer network. As a daily activity banks transfer money to other

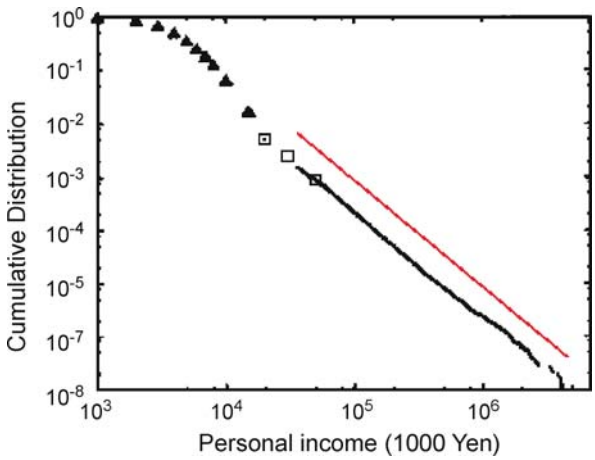

Fractals and Economics, Figure 16
Personal income distribution in Japan [1]

banks for various reasons. In Japan all of these interbank money transfers are done via a special computer network provided by the Bank of Japan. Detailed data of actual money transfer among banks are recorded and analyzed for the basic study.

The total amount of money flow among banks in a day is about 30×10^{12} yen with the number of transactions about 10 000. Figure 17 shows the distribution of the amount of money at a transaction. The range is not wide enough but we can find a power law with an exponent about 1.3 [20].

The number of banks is about 600, so the daily transaction number is only a few percent of the theoretically

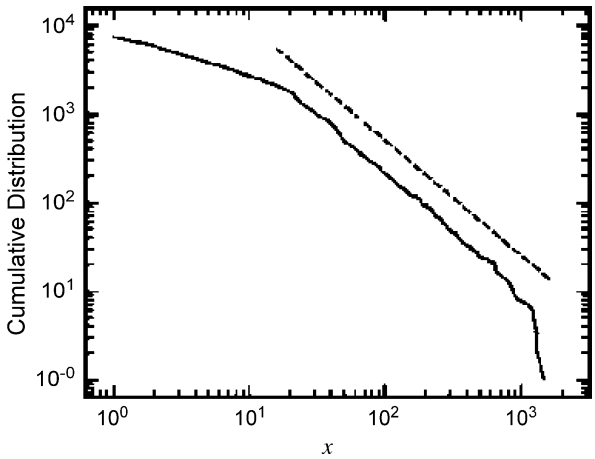

Fractals and Economics, Figure 17
The distribution of the amount of transferred money [21]

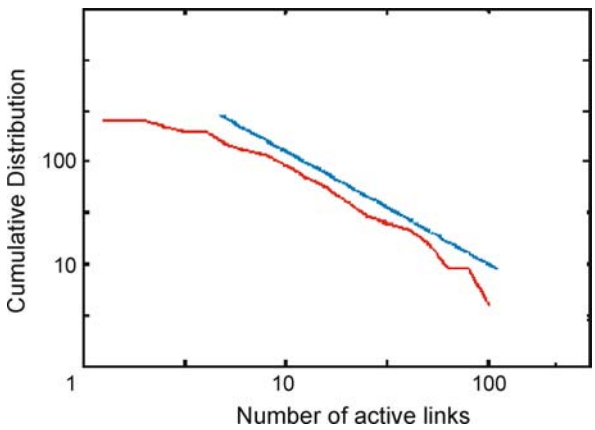

Fractals and Economics, Figure 18
The number distribution of active links per site [20]

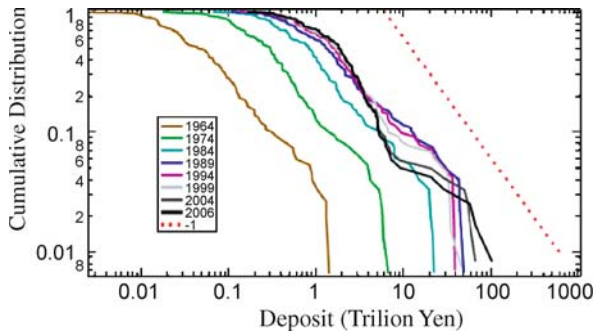

Fractals and Economics, Figure 20
Distribution of total deposit for Japanese banks [57] Power law breaks down from 1999

possible combinations. It is confirmed that there are many pairs of banks which never transact directly. We can define active links between banks for pairs with the averaged number of transaction larger than one per day. By this criterion the number of links becomes about 2000, that is, about 0.5 percent of all possible link numbers. Compared with the complete network, the actual network topologies are much more sparse.

In Fig. 18 the number distribution of active links per site are plotted in log–log plot [21]. As is known from this graph, there is an intermediate range in which the link number distribution follows a power law. In the terminology of recent complex network study, this property is called the scale-free network [5]. The scale-free network structure among these intermediate banks is shown in Fig. 19.

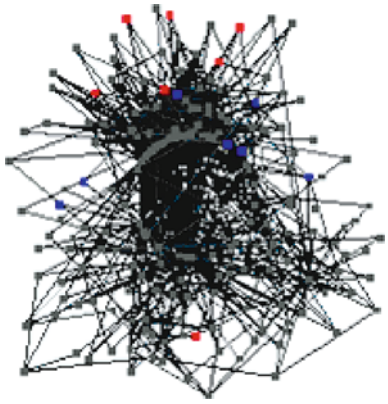

Fractals and Economics, Figure 19
Scale-free network of intermediate banks [20]

There are about 10 banks with large link numbers which deviate from the power law, also small link number banks with link number less than four are out of the power law. Such small banks are known to make a satellite structure that many banks linked to one large link number banks. It is yet to clarify why intermediate banks make fractal network, and also to clarify the role of large banks and small banks which are out of the fractal configuration.

In relation with the banks, there are fractal properties other than cash flow and the transaction network. The distribution of the whole amount of deposit of Japanese bank is approximated by a power law as shown in Fig. 20 [57]. In recent years large banks merged making a few mega banks and the distribution is a little deformed. Historically there were more than 6000 banks in Japan, however, now we have about 600 as mentioned. It is very rare that a bank disappears, instead banks are merged or absorbed. The number distribution of banks which are historically behind a present bank is plotted in Fig. 21, again a power law can be confirmed.

Other than the example of the bank network, network structures are very important generally in economics. In production process from materials, through various parts to final products the network structure is recently studied in view of complex network analysis [18]. Trade networks among companies can also be described by network terminology. Recently, network characterization quantities such as link numbers (Fig. 22), degrees of authority, and Pageranks are found to follow power laws from real trade data for nearly a million of companies in Japan [34].

Still more power laws in economics can be found in sales data. A recent study on the distribution of expenditure at convenience stores in one shopping trip shows a clear power law distribution with the exponent close to two as shown in Fig. 23 [33]. Also, book sales, movie hits,

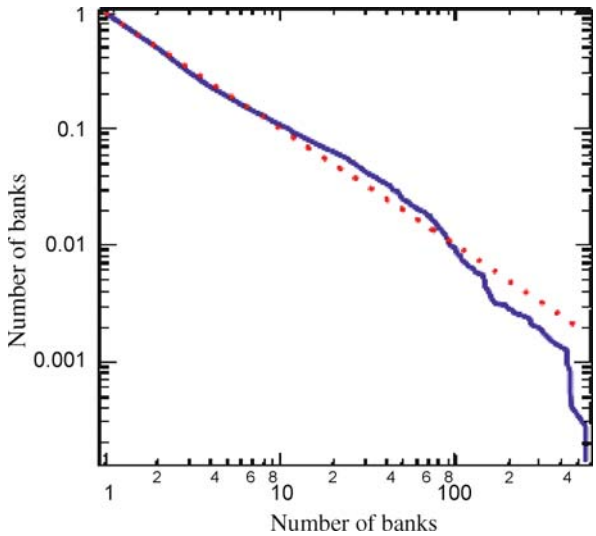

Fractals and Economics, Figure 21
Distribution of bank numbers historically behind a present bank [57]

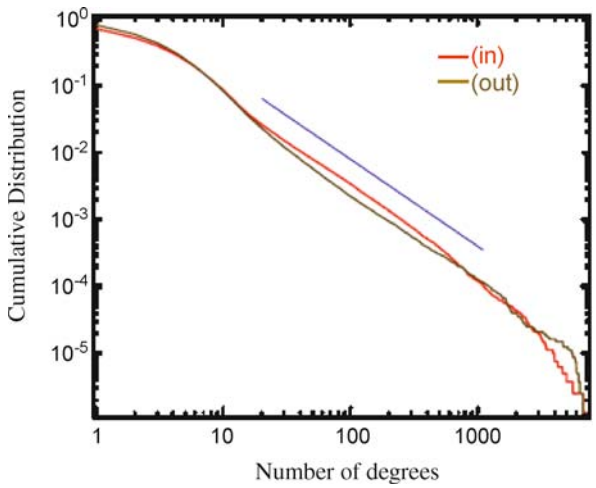

Fractals and Economics, Figure 22
Distribution of in-degrees and out-degrees in Japanese company network [34]

news paper sales are known to be approximated by power laws [39].

Viewing all these data in economics, we may say that fractals are everywhere in economics. In order to understand why fractals appear so frequently, we firstly need to make simple toy models of fractals which can be analyzed completely, and then, based on such basic models we can make more realistic models which can be directly comparable with real data. At that level of study we will be able to predict or control the complex real world economy.

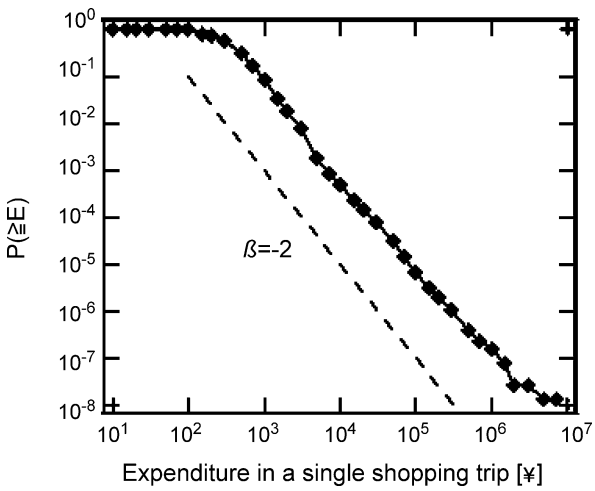

Fractals and Economics, Figure 23
Distribution of expenditure in one shopping trip [33]

Basic Models of Power Laws

In this chapter we introduce general mathematical and physical models which produce power law distributions. By solving these simple and basic cases we can deepen our understanding of the underlying mechanism of fractals or power law distributions in economics.

Transformation of Basic Distributions

A power law distribution can be easily produced by variable transformation from basic distributions.

1. Let x be a stochastic variable following a uniform distribution in the range $(0, 1]$, then, $y \equiv x^{-1/\alpha}$ satisfies a power law, $P(> y) = y^{-\alpha}$ for $y \geq 1$. This is a useful transformation in case of numerical simulation using random variable following power laws.
2. Let x be a stochastic variable following an exponential distribution, $P(> x) = \mathrm{e}^{-x}$, for positive x, then, $y \equiv \mathrm{e}^{x/\alpha}$ satisfies a power law, $P(> y) \propto y^{-\alpha}$. As exponential distributions occur frequently in random process such as the Poisson process, or energy distribution in thermal equilibrium, this simple exponential variable transformation can make it a power law.

Superposition of Basic Distributions

A power law distribution can also be easily produced by superposition of basic distributions.

Let x be a Gaussian distribution with the probability density given by

$$p_R(x) = \frac{\sqrt{R}}{\sqrt{2\pi}} \mathrm{e}^{-\frac{R}{2}x^2} , \tag{11}$$

and R be a χ^2 distribution with degrees of freedom α,

$$w(R) = \frac{\left(\frac{1}{2}\right)^{\alpha/2}}{\Gamma\left(\frac{\alpha}{2}\right)} R^{\frac{\alpha}{2}-1} \mathrm{e}^{-\frac{R}{2}} . \tag{12}$$

Then, the superposition of Gaussian distribution, Eq. (11), with the weight given by Eq. (12) becomes the T-distribution having power law tails:

$$\begin{aligned} p(x) &= \int_0^\infty W(R) p_R(x) \mathrm{d}R \\ &= \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{\alpha}{2}\right)} \frac{1}{(1+x^2)^{\frac{\alpha+1}{2}}} \propto |x|^{-\alpha-1} , \end{aligned} \tag{13}$$

which is $P(> |x|) \propto |x|^{-\alpha}$ in cumulative distribution. In the special case that R, the inverse of variance of the normal distribution, distributes exponentially, the value of α is 2. Similar super-position can be considered for any basic distributions and power law distributions can be produced by such superposition.

Stable Distributions

Assume that stochastic variables, $x_1, x_2, \ldots, x_n$, are independent and follow the same distribution, $p(x)$, then consider the following normalized summation;

$$X_n \equiv \frac{x_1 + x_2 + \cdots + x_n - \mu_n}{n^{1/\alpha}} . \tag{14}$$

If there exists $\alpha > 0$ and μ_n, such that the distribution of X_n is identical to $p(x)$, then, the distribution belongs to one of the Levy stable distributions [10]. The parameter α is called the characteristic exponent which takes a value in the range (0, 2]. The stable distribution is characterized by four continuous parameters, the characteristic exponent, an asymmetry parameter which takes a value in [−1, 1], the scale factor which takes a positive value and the location parameter which takes any real number. Here, we introduce just a simple case of symmetric distribution around the origin with the unit scale factor. The probability density is then given as

$$p(x;\alpha) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{-i\rho x} \mathrm{e}^{-|\rho|^\alpha} \mathrm{d}\rho . \tag{15}$$

For large $|x|$ the cumulative distribution follows the power law, $P(> x;\alpha) \propto |x|^{-\alpha}$ except the case of $\alpha = 2$. The stable distribution with $\alpha = 2$ is the Gaussian distribution.

The most important property of the stable distribution is the generalized central limit theorem: If the distribution of sum of any independent identically distributed random variables like X_n in Eq. (14) converges in the limit of $n \to \infty$ for some value of α, then the limit distribution is a stable distribution with the characteristic exponent α. For any distribution with finite variance, the ordinary central limit theory holds, that is, the special case of $\alpha = 2$. For any infinite variance distribution the limit distribution is $\alpha \neq 2$ with a power law tail. Namely, a power law realizes simply by summing up infinitely many stochastic variables with diverging variance.

Entropy Approaches

Let x_0 be a positive constant and consider a probability density $p(x)$ defined in the interval $[x_0, \infty)$, the entropy of this distribution is given by

$$S \equiv -\int_{x_0}^{\infty} p(x) \log p(x) \mathrm{d}x . \tag{16}$$

Here, we find a distribution that maximizes the entropy with a constraint such that the expectation of logarithm of x is a constant, $\langle \log x \rangle = M$. Then, applying the variational principle to the following function,

$$\begin{aligned} L \equiv -\int_{x_0}^{\infty} p(x) \log p(x) \mathrm{d}x - \lambda_1 \left(\int_{x_0}^{\infty} p(x) \mathrm{d}x - 1 \right) \\ + \lambda_2 \left(\int_{x_0}^{\infty} p(x) \log x \mathrm{d}x - M \right) \end{aligned} \tag{17}$$

the power law is obtained,

$$P(\geq x) = \left(\frac{x}{x_0} \right)^{-\frac{1}{M - \log x_0}} . \tag{18}$$

In other words, a power law distribution maximizes the entropy in the situation where products are conserved. To be more precise, consider two time dependent random variables interacting each other satisfying the relation, $x_1(t) \cdot x_2(t) = x_1(t') \cdot x_2(t')$, then the equilibrium distribution follows a power law.

Another entropy approach to the power laws is to generalize the entropy by the following form [56],

$$S_q \equiv \frac{1 - \int_{x_0}^{\infty} p(x)^q \mathrm{d}x}{q-1} , \tag{19}$$

where q is a real number. This function is called the q-entropy and the ordinary entropy, Eq. (15), recovers in the

limit of $q \to 1$. Maximizing the q-entropy keeping the variance constant, so-called a q-Gaussian distribution is obtained, which has the same functional form with the T-distribution, Eq. (12), with the exponent α given by

$$\alpha = \frac{q-3}{1-q} . \tag{20}$$

This generalized entropy formulation is often applied to nonlinear systems having long correlations, in which power law distributions play the central role.

Random Multiplicative Process

Stochastic time evolution described by the following formulation is called the multiplicative process,

$$x(t+1) = b(t)x(t) + f(t) , \tag{21}$$

where $b(t)$ and $f(t)$ are both independent random variables [17]. In the case that $b(t)$ is a constant, the distribution of $x(t)$ depends on the distribution of $f(t)$, for example, if $f(t)$ follows a Gaussian distribution, then the distribution of $x(t)$ is also a Gaussian. However, in the case that $b(t)$ fluctuates randomly, the resulting distribution of $x(t)$ is known to follows a power law independent of $f(t)$,

$$P(> x) \propto |x|^{-\alpha} , \tag{22}$$

where the exponent α is determined by solving the following equation [48],

$$\langle |b(t)|^{\alpha} \rangle = 1 . \tag{23}$$

This steady distribution exists when $\langle \log |b(t)| \rangle < 0$ and $f(t)$ is not identically 0. As a special case that $b(t) = 0$ with a finite probability, then a steady state exists. It is proved rigorously that there exists only one steady state, and starting from any initial distribution the system converges to the power law steady state.

In the case $\langle \log |b(t)| \rangle \geq 0$ there is no statistically steady state, intuitively the value of $|b(t)|$ is so large that $x(t)$ is likely to diverge. Also in the case $f(t)$ is identically 0 there is no steady state as known from Eq. (21) that $\log |x(t)|$ follows a simple random walk with random noise term, $\log |b(t)|$.

The reason why this random multiplicative process produces a power law can be understood easily by considering a special case that $b(t) = b > 1$ with probability 0.5 and $b(t) = 0$ otherwise, with a constant value of $f(t) = 1$. In such a situation the value of $x(t)$ is $1 + b + b^2 + \cdots + b^K$ with probability $(0.5)^K$. From this we can directly evaluate the distribution of $x(t)$,

$$\begin{aligned} &P\left(\geq \frac{b^{K+1}-1}{b-1}\right) = 2^{-K+1} \quad \text{i. e.} \\ &P(\geq x) = 4(1+(b-1)x)^{-\alpha} , \quad \alpha = \frac{\log 2}{\log b} . \end{aligned} \tag{24}$$

As is known from this discussion, the mechanism of this power law is deeply related to the above mentioned transformation of exponential distribution in Sect. "Transformation of Basic Distributions".

The power law distribution of a random multiplicative process can also be confirmed experimentally by an electrical circuit in which resistivity fluctuates randomly [38]. In an ordinary electrical circuit the voltage fluctuations in thermal equilibrium is nearly Gaussian, however, for a circuit with random resistivity a power law distribution holds.

Aggregation with Injection

Assume the situation that many particles are moving randomly and when two particles collide they coalesce making a particle with mass conserved. Without any injection of particles the system converges to the trivial state that only one particle remains. In the presence of continuous injection of small mass particles there exists a non-trivial statistically steady state in which mass distribution follows a power law [41]. Actually, the mass distribution of aerosol in the atmosphere is known to follow a power law in general [11].

The above system of aggregation with injection can be described by the following model. Let j be the discrete space, and $x_j(t)$ be the mass on site j at time t, then choose one site and let the particle move to another site and particles on the visited site merge, then add small mass particles to all sites, this process can be mathematically given as,

$$x_j(t+1) = \begin{cases} x_j(t) + x_k(t) + f_j(t) , & \text{prob} = 1/N \\ x_j(t) + f_j(t) , & \text{prob} = (N-2)/N \\ f_j(t) , & \text{prob} = 1/N \end{cases} \tag{25}$$

where N is the total number of sites and $f_j(t)$ is the injected mass to the site j.

The characteristic function, $Z(\rho, t) \equiv \langle e^{-\rho x_j(t)} \rangle$, which is the Laplace transform of the probability density,

satisfies the following equation by assuming uniformity,

$$Z(\rho, t+1) = \left\{ \frac{N-2}{N} Z(\rho, t)^2 + \frac{1}{N} Z(\rho, t) + \frac{1}{N} \right\} \langle e^{-\rho f_j(t)} \rangle . \tag{26}$$

The steady state solution in the vicinity of $\rho = 0$ is obtained as

$$Z(\rho) = 1 - \sqrt{\langle f \rangle} \rho^{1/2} + \cdots . \tag{27}$$

From this behavior the following power law steady distribution is obtained.

$$P(\geq x) \propto x^{-\alpha} , \quad \alpha = \frac{1}{2} . \tag{28}$$

By introducing a collision coefficient depending on the size of particles power laws with various values of exponents realized in the steady state of such aggregation with injection system [46].

Critical Point of a Branching Process

Consider the situation that a branch grows and splits with probability q or stops growing with probability $1 - q$ as shown in Fig. 24. What is the size distribution of the branch? This problem can be solved in the following way. Let $p(r)$ be the probability of finding a branch of size r, then the next relation holds.

$$p(r+1) = q \sum_{s=1}^{r-1} p(s) p(r-s) . \tag{29}$$

Multiplying y^{r+1} and summing up by r from 0 to ∞, a closed equation of the generating function, $M(y)$, is obtained,

$$M(y) - 1 + q = q \cdot y \cdot M(y)^2 , \quad M(y) \equiv \sum_{r=0}^{\infty} y^r p(r) . \tag{30}$$

Solving this quadratic equation and expanding in terms of y, we have the probability density,

$$p(r) \propto r^{-3/2} e^{-Q(q) r} , \quad Q(q) \equiv \log 4q(1-q) . \tag{31}$$

For $q < 0.5$ the probability decays exponentially for large r, in this case all branches has a finite size. At $q = 0.5$ the branch size follows the power law, $P(\geq r) \propto r^{-1/2}$, and the average size of branch becomes infinity. For

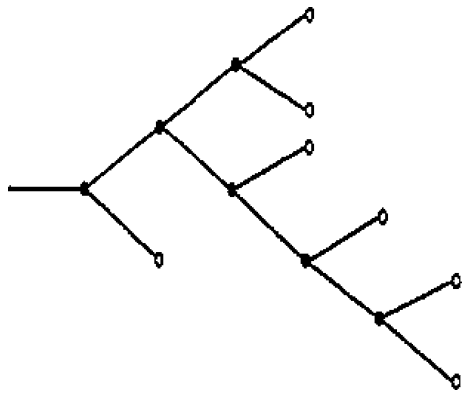

Fractals and Economics, Figure 24
Branching process (from *left* to *right*)

$q > 0.5$ there is a finite probability that a branch grows infinitely. The probability of having an infinite branch, $p(\infty) = 1 - M(1)$, is given as,

$$p(\infty) = \frac{2q - 1 + \sqrt{1 - 4q(1-q)}}{2q} , \tag{32}$$

which grows monotonically from zero to one in the range $q = [0.5, 1]$. It should be noted that the power law distribution realizes at the critical point between the finite-size phase and the infinite-size phase [42].

Compared with the preceding model of aggregation with injection, Eq. (28), the mass distribution is the same as the branch size distribution at the critical point in Eq. (31). This coincidence is not an accident, but it is known that aggregation with injection automatically chooses the critical point parameter. Aggregation and branching are reversed process and the steady occurrence of aggregation implies that branching numbers keep a constant value on average and this requires the critical point condition. This type of critical behaviors is called the self-organized criticality and examples are found in various fields [4].

Finite Portion Transport

Here, a kind of mixture of aggregation and branching is considered. Assume that conserved quantities are distributed in N-sites. At each time step choose one site randomly, and transport a finite portion, $\theta x_j(t)$, to another randomly chosen site, where θ is a parameter in the range $[0, 1]$.

$$x_j(t+1) = (1-\theta) x_j(t) , \quad x_k(t+1) = x_k(t) + \theta x_j(t) . \tag{33}$$

It is known that for small positive θ the statistically steady distribution x is well approximated by a Gaussian like the case of thermal fluctuations. For θ close to 1 the fluctuation of x is very large and its distribution is close to a power law. In the limit θ goes to 1 and the distribution converges to Eq. (28), the aggregation with injection case. For intermediate values of θ the distribution accompanies a fat tail between Gaussian and a power law [49].

Fractal Tiling

A fractal tiling is introduced as the final basic model. Figure 25 shows an example of fractal tiling of a plane by squares. Like this case Euclidean space is covered by various sizes of simple shapes like squares, triangles, circles etc. The area size distribution of squares in Fig. 25 follows the power law,

$$P(\geq x) \propto x^{-\alpha}\,, \quad \alpha = 1/2\,. \tag{34}$$

Generalizing this model in d-dimensional space, the distribution of d-dimensional volume x is characterized by a power law distribution with an exponent, $\alpha = (d-1)/d$, therefore, the Zipf's law which is the case of $\alpha = 1$ realizes in the limit of $d = \infty$. The fracture size distribution measured in mass introduced in the beginning of this article corresponds to the case of $d = 3$.

A classical example of fractal tiling is the Apollonian gasket, that is, a plane is covered totally by infinite number of circles which are tangent each other. For a given river pattern like Fig. 26 the basin area distribution follows a power law with exponent about $\alpha = 0.4$ [45]. Although these are very simple geometric models, simple models may sometimes help our intuitive understanding of fractal phenomena in economics.

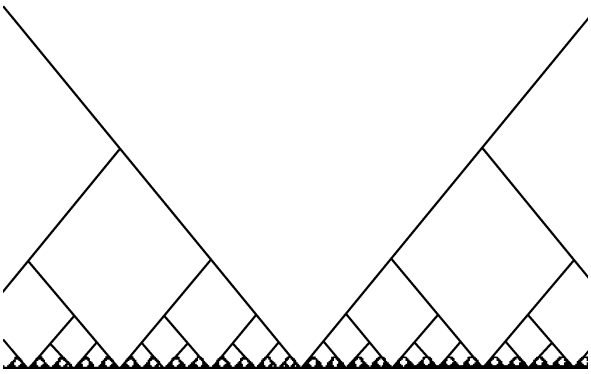

Fractals and Economics, Figure 25
An example of fractal tiling

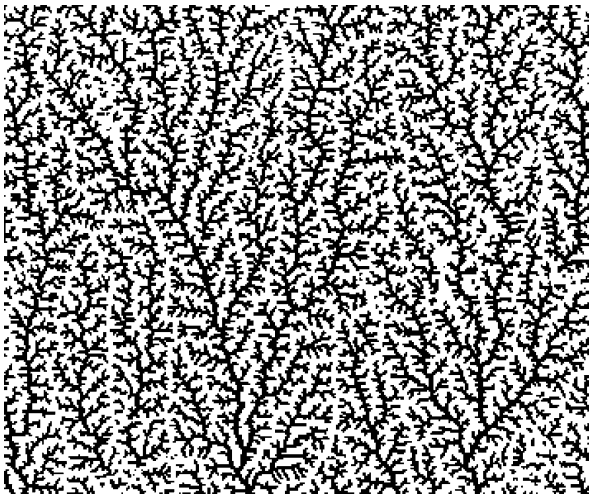

Fractals and Economics, Figure 26
Fractal tiling by river patterns [45]

Market Models

In this chapter market price models are reviewed in view of fractals. There are two approaches for construction of market models. One is modeling the time sequences directly by some stochastic model, and the other is modeling markets by agent models which are artificial markets in computer consisted of programmed dealers.

The first market price model was proposed by Bachelier in 1900 written as his Ph.D thesis [3], that is, five years before the model of Einstein's random walk model of colloid particles. His idea was forgotten for nearly 50 years. In 1950's Markowitz developed the portfolio theory based on a random walk model of market prices [28]. The theory of option prices by Black and Scholes was introduced in the 1970s, which is also based on random walk model of market prices, or to be more precise a logarithm of market prices in continuum description [7].

In 1982 Engle introduced a modification of the simple random walk model, the ARCH model, which is the abbreviation of auto-regressive conditional heteroscedasticity [9]. This model is formulated for market price difference as,

$$\Delta x(t) = \sigma(t) f(t)\,, \tag{35}$$

where $f(t)$ is a random variable following a Gaussian distribution with 0 mean and variance unity, the local variance $\sigma(t)$ is given as

$$\sigma(t)^2 = c_0 + \sum_{j=1}^{k} c_k (\Delta x(t-k))^2\,, \tag{36}$$

with adjustable positive parameters, $\{c_0, c_1, \ldots, c_k\}$. By the effect of this modulation on variance, the distribution of price difference becomes superposition of Gaussian distribution with various values of variance, and the distribution becomes closer to a power law. Also, volatility clustering occurs automatically so that the volatility autocorrelation becomes longer.

There are many variants of ARCH models, such as GARCH and IGARCH, but all of them are based on purely probabilistic modeling, and the probability of prices going up and that of going down are identical.

Another type of market price model has been proposed from physics view point [53]. The model is called the PUCK model, an abbreviation of potentials of unbalanced complex kinetics, which assumes the existence of market's time-dependent potential force, $U_{\mathrm{M}}(x;t)$, and the time evolution of market price is given by the following set of equations;

$$x(t+1)-x(t) = -\frac{\mathrm{d}}{\mathrm{d}x} U_{\mathrm{M}}(x;t)\bigg|_{x=x(t)-x_{\mathrm{M}}(t)} + f(t)\,, \quad (37)$$

$$U_{\mathrm{M}}(x;t) \equiv \frac{b(t)}{M-1}\frac{x^2}{2}\,, \quad (38)$$

where M is the number of moving average needed to define the center of potential force,

$$x_{\mathrm{M}}(t) \equiv \frac{1}{M}\sum_{k=0}^{M-1} x(t-k)\,. \quad (39)$$

In this model $f(t)$ is the external noise and $b(t)$ is the curvature of quadratic potential which changes with time. When $b(t) = 0$ the model is identical to the simple random walk model. When $b(t) > 0$ the market prices are attracted to the moving averaged price, $x_{\mathrm{M}}(t)$, the market is stable, and when $b(t) < 0$ prices are repelled from $x_{\mathrm{M}}(t)$ so that the price fluctuation is large and the market is unstable. For $b(t) < -2$ the price motion becomes an exponential function of time, which can describe singular behavior such as bubbles and crashes very nicely.

In the simplest case of $M = 2$ the time evolution equation becomes,

$$\Delta x(t+1) = -\frac{b(t)}{2}\Delta x(t) + f(t)\,. \quad (40)$$

As is known from this functional form in the case $b(t)$ fluctuates randomly, the distribution of price difference follows a power law as mentioned in the previous Sect. "Random Multiplicative Process", Random multiplicative process. Especially the PUCK model derives the ARCH model by introducing a random nonlinear potential function [54]. The value of $b(t)$ can be estimated from the data and most of known empirical statistical laws including fractal properties are fulfilled as a result [55].

The peculiar difference of this model compared with financial technology models is that directional prediction is possible in some sense. Actually, from the data it is known that $b(t)$ changes slowly in time, and for non-zero $b(t)$ the autocorrelation is not zero implying that the up-down statistics in the near future is not symmetric. Moreover in the case of $b(t) < -2$ the price motion show an exponential dynamical growth hence predictable.

As introduced in Sect. "Examples in Economics" the tick interval fluctuations can be characterized by the $1/f$ power spectrum. This power law can be explained by a model called the self-modulation model [52]. Let Δt_j be the jth tick interval, and we assume that the tick interval can be approximated by the following random process,

$$\Delta t_{j+1} = \mu_j \frac{1}{K}\sum_{k=0}^{K-1} \Delta t_{j-k} + g_j\,, \quad (41)$$

where μ_j is a positive random number following an exponential distribution with the mean value 1, and K is an integer which means the number of moving average, g_j is a positive random variable. Due to the moving average term in Eq. (41) the tick interval automatically make clusters as shown in Fig. 27, and the corresponding power spectrum is proved to be proportional to $1/f$ as typically represented in Fig. 28.

The market data of tick intervals are tested whether Eq. (41) really works or not. In Fig. 29 the cumulative probability of estimated value of μ_j from market data is plotted where the moving average size is determined by the physical time of 150 seconds and 400 seconds. As known from this figure, the distribution fits very nicely with the exponential distribution when the moving average size is 150 seconds. This result implies that dealers in the market are mostly paying attention to the latest transaction for about a few minutes only. And the dealers' clocks in their

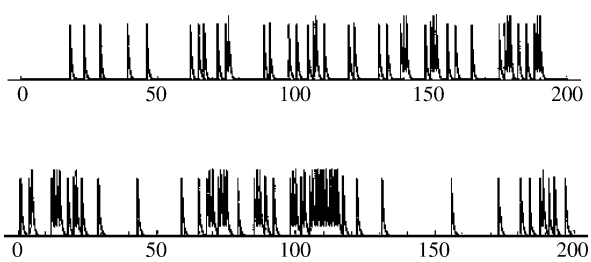

Fractals and Economics, Figure 27
Tick intervals of Poisson process (*top*) and the self-modulation process (*bottom*) [52]

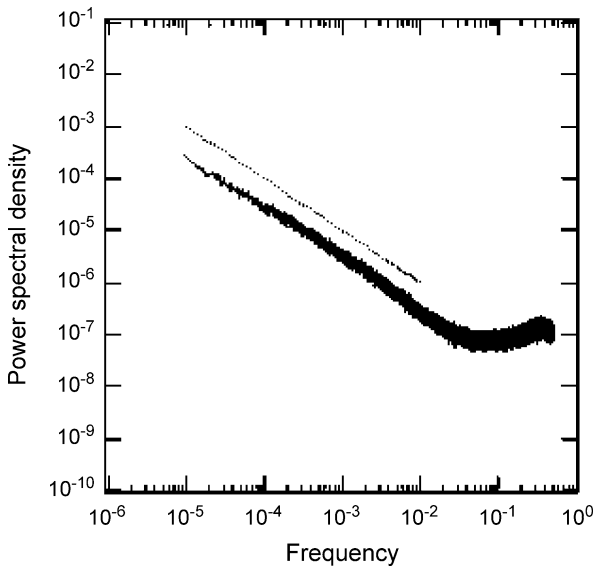

Fractals and Economics, Figure 28
The power spectrum of the self-modulation process [52]

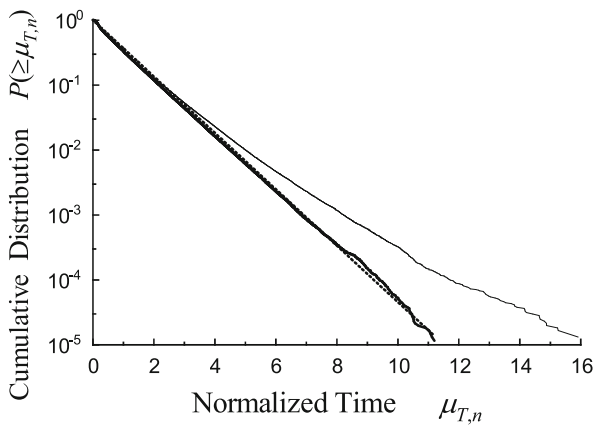

Fractals and Economics, Figure 29
The distribution of normalized time interval [50]

minds move quicker if the market becomes busier. By this self-modulation effect transactions in markets automatically make a fractal configuration.

Next, we introduce a dealer model approach to the market [47]. In any financial market dealers' final goal is to gain profit from the market. To this end dealers try to buy at the lowest price and to sell at the highest price. Assume that there are N dealers at a market, and let the jth dealer's buying and selling prices in their mind $B_j(t)$ and $S_j(t)$. For each dealer the inequality, $B_j(t) < S_j(t)$, always holds. We pay attention to the maximum price of $\{B_j(t)\}$ called the best bid, and to the minimum price of $\{S_j(t)\}$ called the best ask. Transactions occur in the market if there exists a pair of dealers, j and k, who give the best bid and best ask respectively, and they fulfill the following condition,

$$B_j(t) \geq S_k(t)\,. \tag{42}$$

In the model the market price is given by the mean value of these two prices.

As a simple situation we consider a deterministic time evolution rule for these dealers. For all dealers the spread, $S_j(t) - B_j(t)$, is set to be a constant L. Each dealer has a position, either a seller or a buyer. When the jth dealer's position is a seller the selling price in mind, $S_j(t)$, decreases every time step until he can actually sell. Similar dynamics is applied to a buyer with the opposite direction of motion. In addition we assume that all dealers shift their prices in mind proportional to a market price change. When this proportional coefficient is positive, the dealer is categorized as a trend-follower. If this coefficient is negative, the dealer is called a contrarian. These rules are summarized by the following time evolution equations.

$$B_j(t+1) = B_j(t) + a_j S_j + b_j \Delta x(t)\,, \tag{43}$$

where S_j takes either $+1$ or -1 meaning the buyer position or seller position, respectively, $\Delta x(t)$ gives the latest market price change, $\{a_j\}$ are positive numbers given initially, $\{b_j\}$ are also parameters given initially.

Figure 30 shows an example of market price evolution in the case of three dealers. It should be noted that although the system is deterministic, namely, the future price is determined uniquely by the set of initial values, resulting market price fluctuates almost randomly even in the minimum case of three dealers. The case of $N = 2$

Fractals and Economics, Figure 30
Price evolution of a market with deterministic three dealers

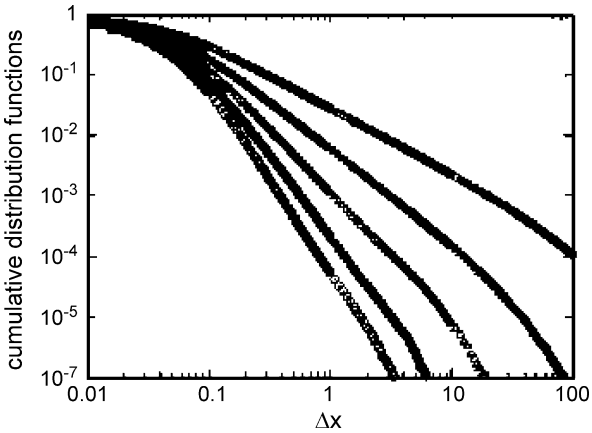

Fractals and Economics, Figure 31
Cumulative distribution of a dealer model for different values of *b*. For weaker trend-follow the slope is steeper [38]

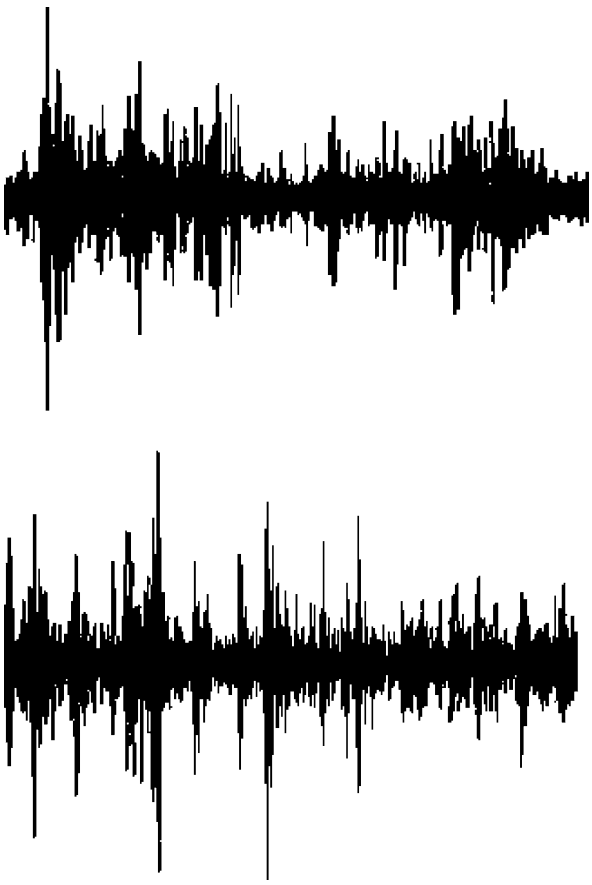

Fractals and Economics, Figure 32
Price difference time series for a real market (*top*) and a dealer model (*bottom*)

gives only periodic time evolution as expected, while for $N \geq 3$ the system can produce market price fluctuations similar to the real market price fluctuations, for example, the fractal properties of price chart and the power law distribution of price difference are realized.

In the case that the value of $\{b_j\}$ are identical for all dealers, b, then the distribution of market price difference follows a power law where the exponent is controllable by this trend-follow parameter, b as shown in Fig. 31 [37]. The volatility clustering is also observed automatically for large dealer number case as shown in Fig. 32 (bottom) which looks quite similar to a real price difference time series Fig. 32 (top).

By adding a few features to this basic dealer model it is now possible to reproduce almost all statistical characteristics of market, such as tick-interval fluctuations, abnormal diffusions etc. [58]. In this sense the study of market behaviors are now available by computer simulations based on the dealer model. Experiments on the market is either impossible or very difficult for a real market, however, in an artificial market we can repeat occurrence of bubbles and crashes any times, so that we might be able to find a way to avoid catastrophic market behaviors by numerical simulation.

Income Distribution Models

Let us start with a famous historical problem, the St. Petersburg Paradox, as a model of income. This paradox was named after Daniel Bernoulli's paper written when he was staying in the Russian city, Saint Petersburg, in 1738 [6]. This paradox treats a simple lottery as described in the following, which is deeply related to the infinite expected value problem in probability theory and also it has been attracting a lot of economists' interest in relation with the essential concept in economics, the utility [2].

Assume that you enjoy a game of chance, you pay a fixed fee, X dollars, to enter, and then you toss a fair coin repeatedly until a tail firstly appears. You win 2^n dollars where n is the number of heads. What is the fair price of the entrance fee, X?

Mathematically a fair price should be equal to the expectation value, therefore, it should be given as,

$$X = \sum_{n=0}^{\infty} 2^n \cdot \frac{1}{2^{n+1}} = \infty \,. \tag{44}$$

This mathematical answer implies that even X is one million dollars this lottery is generous enough and you should buy because expectation is infinity. But, would you dare to buy this lottery, in which you will win only one dollar with probability 0.5, and two dollars with probability 0.25, ...?

Bernoulli's answer to this paradox is to introduce the human feeling of value, or utility, which is proportional to the logarithm of price, for example. Based on this expected utility hypothesis the fair value of X is given as follows,

$$X = \sum_{n=0}^{\infty} \frac{U(2^n)}{2^{n+1}} = \sum_{n=0}^{\infty} \frac{\log(2^n)}{2^{n+1}} = 1 + \log 2 \approx 1.69\,, \quad (45)$$

where the utility function, $U(x) = 1 + \log x$, is normalized to satisfy $U(1) = 1$. This result implies that the appropriate entry fee X should be about two dollars.

The idea of utility was highly developed in economics for description of human behavior, in the way that human preference is determined by maximal point of utility function, the physics concept of the variational principle applied to human action. Recently, in the field of behavioral finance which emerged from psychology the actual observation of human behaviors about money is the main task and the St. Petersburg paradox is attracting attention [36].

Although Bernoulli's solution may explain the human behavior, the fee $X = 2$ is obviously so small that the bookmaker of this lottery will bankrupt immediately if the entrance fee is actually fixed as two dollars and if a lot of people actually buy it. The paradox is still a paradox.

To clarify what is the problem we calculate the distribution of income of a gambler. As an income is 2^n with probability 2^{-n-1}, the cumulative distribution of income is readily obtained as,

$$P(\geq x) \propto 1/x\,. \quad (46)$$

This is the power law which we observed for income distribution of companies in Sect. "Examples in Economics".

The key of this lottery is the mechanism that the prize money doubles at each time a head appears and the coin toss stops when a tail appears. By denoting the number of coin toss by t, we can introduce a stochastic process or a new lottery which is very much related to the St. Petersburg lottery.

$$x(t+1) = b(t)x(t) + 1\,, \quad (47)$$

where $b(t)$ is 2 with probability 0.5 and is 0 otherwise. As introduced in Sect. "Random Multiplicative Process", this problem is solved easily and it is confirmed that the steady state cumulative distribution of $x(t)$ also follows Eq. (46). The difference between the St. Petersburg lottery and the new lottery Eq. (47) is the way of payment of entrance fee. In the case of St. Petersburg lottery the entrance fee X is paid in advance, while in the case of new lottery you have to add one dollar each time you toss a coin. This new lottery is fair from both the gambler side and the bookmaker side because the expectation of income is given by $\langle x(t)\rangle = t$ and the amount of paid fee is also t.

Now we introduce a company's income model by generalizing this new fair lottery in the following way,

$$I(t+1) = b(t)I(t) + f(t)\,, \quad (48)$$

where $I(t)$ denotes the annual income of a company, $b(t)$ represents the growth rate which is given randomly from a growth rate distribution $g(b)$, and $f(t)$ is a random noise. Readily from the results of Sect. "Random Multiplicative Process", we have a condition to satisfy the empirical relation, Eq. (10),

$$\langle b(t)\rangle = \int bg(b) = 1\,. \quad (49)$$

This relation is confirmed to hold approximately in actual company data [32].

In order to explain the job category dependence of the company's income distribution already shown in Fig. 12, we plot the comparison of exponents in Fig. 33. Empirically estimated exponents are plotted in the ordinate and the solutions of the following equation calculated in each job category are plotted in the abscissa,

$$\langle b(t)^{\beta}\rangle = 1\,. \quad (50)$$

The data points are roughly on a straight line demonstrating that the simple growth model of Eq. (48) seems to be meaningful.

An implication of this result is that if a job category is expanding, namely, $\langle b(t)\rangle > 1$, then the power law exponent determined by Eq. (50) is smaller than 1. On the other hand if a job category is shrinking, we have an exponent that is larger than 1.

This type of company's income model can be generalized to take into account the effect of company's size

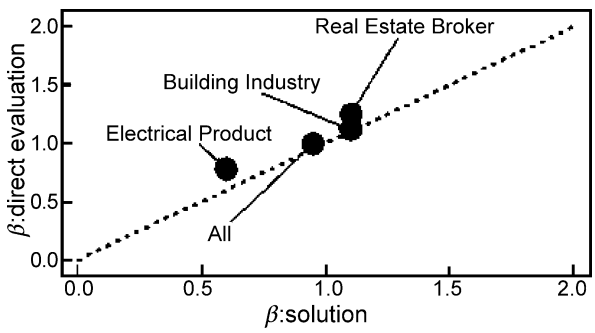

Fractals and Economics, Figure 33
Theoretical predicted exponent value vs. observed value [29]

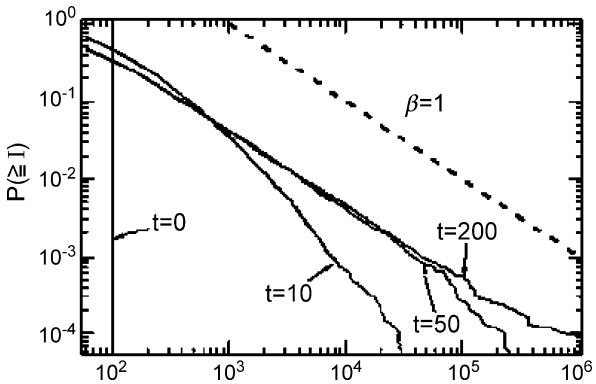

Fractals and Economics, Figure 34
Numerical simulation of income distribution evolution of Japanese companies [32]

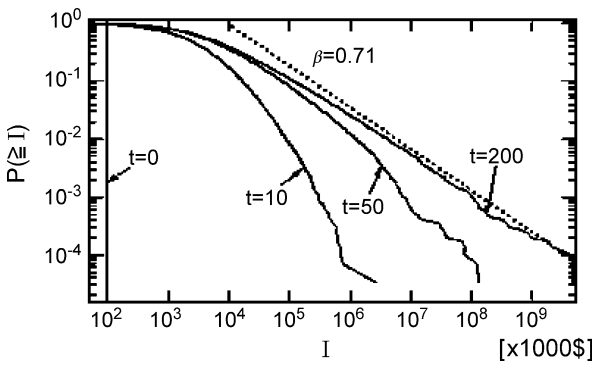

Fractals and Economics, Figure 35
Numerical simulation of income distribution evolution of USA companies [32]

dependence on the distribution of growth rate. Also, the magnitude of the random force term can be estimated from the probability of occurrence of negative income. Then, assuming that the present growth rate distribution continues we can perform a numerical simulation of company's income distribution starting from a uniform distribution as shown in Fig. 34 for Japan and in Fig. 35 for USA. It is shown that in the case of Japan, the company size distribution converges to the power law with exponent −1 in 20 years, while in the case of USA the steady power law's slope is about −0.7 and it takes about 100 years to converge [31]. According to this result extremely large companies with size about 10 times bigger than the present biggest company will appear in USA in this century. Of course the growth rate distribution will change faster than this prediction, however, this model can tell the qualitative direction and the speed of change in very macroscopic economical conditions.

Other than this simple random multiplicative model approach there are various approaches to explain empirical facts about company's statistics assuming a hierarchical structure of organization, for example [23].

Future Directions

Fractal properties generally appear in almost any huge data in economics. As for financial market models, empirical fractal laws are reproduced and the frontier of study is now at the level of practical applications. However, there are more than a million markets in the world and little is known about their interaction. More research on market interaction will be promising. Company data so far analyzed show various fractal properties as introduced in Sect. "Examples in Economics", however, they are just a few cross-sections of global economics. Especially, companies' interaction data are inevitable to analyze the underlying network structures. Not only money flow data it will be very important to observe material flow data in manufacturing and consumption processes. From the viewpoint of environmental study, such material flow network will be of special importance in the near future. Detail sales data analysis is a new topic and progress is expected.

Bibliography

Primary Literature

1. Aoyama H, Nagahara Y, Okazaki MP, Souma W, Takayasu H, Takayasu M (2000) Pareto's law for income of individuals and debt of bankrupt companies. Fractals 8:293–300
2. Aumann RJ (1977) The St. Petersburg paradox: A discussion of some recent comments. J Econ Theory 14:443–445 http://en.wikipedia.org/wiki/Robert_Aumann
3. Bachelier L (1900) Theory of Speculation. In: Cootner PH (ed) The Random Character of Stock Market Prices. MIT Press, Cambridge (translated in English)
4. Bak P (1996) How Nature Works. In: The Science of Self-Organized Criticality. Springer, New York
5. Barabási AL, Réka A (1999) Emergence of scaling in random networks. Science 286:509–512; http://arxiv.org/abs/cond-mat/9910332
6. Bernoulli D (1738) Exposition of a New Theory on the Measurement of Risk; Translation in: (1954) Econometrica 22:22–36
7. Black F, Scholes M (1973) The Pricing of Options and Corporate Liabilities. J Political Econ 81:637–654
8. Brenner YS, Kaelble H, Thomas M (1991) Income Distribution in Historical Perspective. Cambridge University Press, Cambridge
9. Engle RF (1982) Autoregressive Conditional Heteroskedasticity With Estimates of the Variance of UK Inflation. Econometrica 50:987–1008
10. Feller W (1971) An Introduction to Probability Theory and Its Applications, 2nd edn, vol 2. Wiley, New York

11. Friedlander SK (1977) Smoke, dust and haze: Fundamentals of aerosol behavior. Wiley-Interscience, New York
12. Fujiwara Y (2004) Zipf Law in Firms Bankruptcy. Phys A 337:219–230
13. Fujiwara Y, Guilmi CD, Aoyama H, Gallegati M, Souma W (2004) Do Pareto-Zipf and Gibrat laws hold true? An analysis with European firms. Phys A 335:197–216
14. Gilvarry JJ, Bergstrom BH (1961) Fracture of Brittle Solids. II Distribution Function for Fragment Size in Single Fracture (Experimental). J Appl Phys 32:400–410
15. Gopikrishnan P, Meyer M, Amaral LAN, Stanley HE (1998) Inverse Cubic Law for the Distribution of Stock Price Variations. Eur Phys J B 3:139–143
16. Handel PH (1975) 1/*f* Noise-An "Infrared" Phenomenon. Phys Rev Lett 34:1492–1495
17. Havlin S, Selinger RB, Schwartz M, Stanley HE, Bunde A (1988) Random Multiplicative Processes and Transport in Structures with Correlated Spatial Disorder. Phys Rev Lett 61: 1438–1441
18. Hidalgo CA, Klinger RB, Barabasi AL, Hausmann R (2007) The Product Space Conditions the Development of Nations. Science 317:482–487
19. Inaoka H, Toyosawa E, Takayasu H (1997) Aspect Ratio Dependence of Impact Fragmentation. Phys Rev Lett 78:3455–3458
20. Inaoka H, Ninomiya T, Taniguchi K, Shimizu T, Takayasu H (2004) Fractal Network derived from banking transaction – An analysis of network structures formed by financial institutions. Bank of Japan Working Paper. http://www.boj.or.jp/en/ronbun/04/data/wp04e04.pdf
21. Inaoka H, Takayasu H, Shimizu T, Ninomiya T, Taniguchi K (2004) Self-similarity of bank banking network. Phys A 339: 621–634
22. Klass OS, Biham O, Levy M, Malcai O, Solomon S (2006) The Forbes 400 and the Pareto wealth distribution. Econ Lett 90:290–295
23. Lee Y, Amaral LAN, Canning D, Meyer M, Stanley HE (1998) Universal Features in the Growth Dynamics of Complex Organizations. Phys Rev Lett 81:3275–3278
24. Mandelbrot BB (1963) The variation of certain speculative prices. J Bus 36:394–419
25. Mandelbrot BB (1982) The Fractal Geometry of Nature. W.H. Freeman, New York
26. Mandelbrot BB (2004) The (mis)behavior of markets. Basic Books, New York
27. Mantegna RN, Stanley HE (2000) An Introduction to Economics: Correlations and Complexity in Finance. Cambridge Univ Press, Cambridge
28. Markowitz HM (1952) Portfolio Selection. J Finance 7:77–91
29. Mizuno T, Katori M, Takayasu H, Takayasu M (2001) Statistical laws in the income of Japanese companies. In: Empirical Science of Financial Fluctuations. Springer, Tokyo, pp 321–330
30. Mizuno T, Kurihara S, Takayasu M, Takayasu H (2003) Analysis of high-resolution foreign exchange data of USD-JPY for 13 years. Phys A 324:296–302
31. Mizuno T, Kurihara S, Takayasu M, Takayasu H (2003) Investment strategy based on a company growth model. In: Takayasu H (ed) Application of Econophysics. Springer, Tokyo, pp 256–261
32. Mizuno T, Takayasu M, Takayasu H (2004) The mean-field approximation model of company's income growth. Phys A 332:403–411
33. Mizuno T, Toriyama M, Terano T, Takayasu M (2008) Pareto law of the expenditure of a person in convenience stores. Phys A 387:3931–3935
34. Ohnishi T, Takayasu H, Takayasu M (in preparation)
35. Okuyama K, Takayasu M, Takayasu H (1999) Zipf's law in income distribution of companies. Phys A 269:125–131. http://www.ingentaconnect.com/content/els/03784371;jsessionid=5e5wq937wfsqu.victoria
36. Rieger MO, Wang M (2006) Cumulative prospect theory and the St. Petersburg paradox. Econ Theory 28:665–679
37. Sato AH, Takayasu H (1998) Dynamic numerical models of stock market price: from microscopic determinism to macroscopic randomness. Phys A 250:231–252
38. Sato AH, Takayasu H, Sawada Y (2000) Power law fluctuation generator based on analog electrical circuit. Fractals 8: 219–225
39. Sinha S, Pan RK (2008) How a "Hit" is Born: The Emergence of Popularity from the Dynamics of Collective Choice. http://arxiv.org/PS_cache/arxiv/pdf/0704/0704.2955v1.pdf
40. Souma W (2001) Universal structure of the personal income distribution. Fractals 9:463–470; http://www.nslij-genetics.org/j/fractals.html
41. Takayasu H (1989) Steady-state distribution of generalized aggregation system with injection. Phys Rev Lett 63: 2563–2566
42. Takayasu H (1990) Fractals in the physical sciences. Manchester University Press, Manchester
43. Takayasu H (ed) (2002) Empirical Science of Financial Fluctuations–The Advent of Econophysics. Springer, Tokyo
44. Takayasu H (ed) (2003) Application of Econophysics. Springer, Tokyo
45. Takayasu H, Inaoka H (1992) New type of self-organized criticality in a model of erosion. Phys Rev Lett 68:966–969
46. Takayasu H, Takayasu M, Provata A, Huber G (1991) Statistical properties of aggregation with injection. J Stat Phys 65: 725–745
47. Takayasu H, Miura H, Hirabayashi T, Hamada K (1992) Statistical properties of deterministic threshold elements–The case of market price. Phys A 184:127–134
48. Takayasu H, Sato AH, Takayasu M (1997) Stable infinite variance fluctuations in randomly amplified Langevin systems. Phys Rev Lett 79:966–969
49. Takayasu M, Taguchi Y, Takayasu H (1994) Non-Gaussian distribution in random transport dynamics. Inter J Mod Phys B 8:3887–3961
50. Takayasu M (2003) Self-modulation processes in financial market. In: Takayasu H (ed) Application of Econophysics. Springer, Tokyo, pp 155–160
51. Takayasu M (2005) Dynamics Complexity in Internet Traffic. In: Kocarev K, Vatty G (eds) Complex Dynamics in Communication Networks. Springer, New York, pp 329–359
52. Takayasu M, Takayasu H (2003) Self-modulation processes and resulting generic 1/*f* fluctuations. Phys A 324:101–107
53. Takayasu M, Mizuno T, Takayasu H (2006) Potentials force observed in market dynamics. Phys A 370:91–97
54. Takayasu M, Mizuno T, Takayasu H (2007) Theoretical analysis of potential forces in markets. Phys A 383:115–119
55. Takayasu M, Mizuno T, Watanabe K, Takayasu H (preprint)
56. Tsallis C (1988) Possible generalization of Boltzmann–Gibbs statistics. J Stat Phys 52:479–487; http://en.wikipedia.org/wiki/Boltzmann_entropy

57. Ueno H, Mizuno T, Takayasu M (2007) Analysis of Japanese bank's historical network. Phys A 383:164–168
58. Yamada K, Takayasu H, Takayasu M (2007) Characterization of foreign exchange market using the threshold-dealer-model. Phys A 382:340–346

Books and Reviews

Takayasu H (2006) Practical Fruits of Econophysics. Springer, Tokyo
Chatterjee A, Chakrabarti BK (2007) Econophysics of Markets and Business Networks (New Economic Windows). Springer, New York

Fractals in Geology and Geophysics

Donald L. Turcotte
Department of Geology,
University of California,
Davis, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Drainage Networks
Fragmentation
Earthquakes
Volcanic Eruptions
Landslides
Floods
Self-Affine Fractals
Topography
Earth's Magnetic Field
Future Directions
Bibliography

Glossary

Fractal A collection of objects that have a power-law dependence of number on size.

Fractal dimension The power-law exponent in a fractal distribution.

Definition of the Subject

The scale invariance of geological phenomena is one of the first concepts taught to a student of geology. When a photograph of a geological feature is taken, it is essential to include an object that defines the scale, for example, a coin or a person. It was in this context that Mandelbrot [7] introduced the concept of fractals. The length of a rocky coastline is obtained using a measuring rod with a specified length. Because of scale invariance, the length of the coastline increases as the length of the measuring rod decreases according to a power law. It is not possible to obtain a specific value for the length of a coastline due to small indentations down to a scale of millimeters or less.

A fractal distribution requires that the number of objects N with a linear size greater than r has an inverse power-law dependence on r so that

$$N = \frac{C}{r^D} \tag{1}$$

where C is a constant and the power D is the fractal dimension. This power-law scaling is the only distribution that is scale invariant. However, the power-law dependence cannot be used to define a statistical distribution because the integral of the distribution diverges to infinity either for large values or small values of r. Thus fractal distributions never appear in compilations of statistical distributions. A variety of statistical distributions have power-law behavior either at large scales or small scales, but not both. An example is the Pareto distribution.

Many geological phenomena are scale invariant. Examples include the frequency-size distributions of fragments, faults, earthquakes, volcanic eruptions, and landslides. Stream networks and landforms exhibit scale invariance. In terms of these applications there must always be upper and lower cutoffs to the applicability of a fractal distribution. As a specific application consider earthquakes on the Earth. The number of earthquakes has a power-law dependence on the size of the rupture over a wide range of sizes. But the largest earthquake cannot exceed the size of the Earth, say 10^4 km. Also, the smallest earthquake cannot be smaller than the grain size of rocks, say 1 mm. But this range of scales is 10^{10}. Actual earthquakes appear to satisfy fractal scaling over the range 1 m to 10^3 km.

An example of fractal scaling is the number-area distribution of lakes [10], this example is illustrated in Fig. 1. Excellent agreement with the fractal relation given in Eq. (1) is obtained taking $D = 1.90$. The linear dimension r is taken to be the square root of the area A and the power-law (fractal) scaling extends from $r = 100$ m to $r = 300$ km.

Introduction

Fractal scaling evolved primarily as an empirical means of correlating data. A number of examples are given below. More recently a theoretical basis has evolved for the applicability of fractal distributions. The foundation of this basis is the concept of self-organized criticality. A number of simple computational models have been shown to

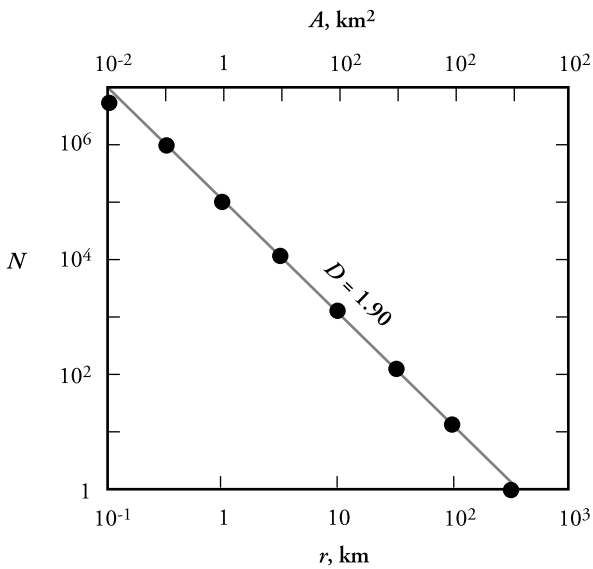

Fractals in Geology and Geophysics, Figure 1
Dependence of the cumulative number of lakes *N* with areas greater than *A* as a function of *A*. Also shown is the linear dimension *r* which is taken to be the square root of *A*. The straight-line correlation is with Eq. (1) taking the fractal dimension $D = 1.90$

yield fractal distributions. Examples include the sand-pile model, the forest-fire model, and the slider-block model.

Drainage Networks

Drainage networks are a universal feature of landscapes on the Earth. Small streams merge to form larger streams, large streams merge to form rivers, and so forth. Strahler [16] quantified stream networks by introducing an ordering system. When two like-order streams of order i merge they form a stream of order $i + 1$. Thus two $i = 1$ streams merge to form a $i = 2$ stream, two $i = 2$ streams merge to from a $i = 3$ stream and so forth. A bifurcation ratio R_b is defined by

$$R_b = \frac{N_i}{N_{i+1}} \tag{2}$$

where N_i is the number of streams of order i. A length order ratio R_r is defined by

$$R_r = \frac{r_{i+1}}{r_i} \tag{3}$$

where r_i is the mean length of streams of order i. Empirically both R_b and R_r are found to be nearly constant for a range of stream orders in a drainage basin. From Eq. (1) the fractal dimension of a drainage basin

$$D = \frac{\ln(N_i/N_{i+1})}{\ln(r_{i+1}/r_i)} = \frac{\ln R_b}{\ln R_r} \tag{4}$$

Typically $R_b = 4.6$, $R_r = 2.2$, and the corresponding fractal dimension is $D = 1.9$. This scale invariant scaling of drainage networks was recognized some 20 years before the concept of fractals was introduced in 1967.

A major advance in the quantification of stream networks was made by Tokunaga [17]. This author was the first to recognize the importance of side branching, that is some $i = 1$ streams intersect $i = 2$, $i = 3$, and all higher-order streams. Similarly, $i = 2$ streams intersect $i = 3$ and higher- order streams and so forth. A fully self-similar, side-branching topology was developed. Applications to drainage networks have been summarized by Peckham [11] and Pelletier [13].

Fragmentation

An important application of power-law (fractal) scaling is to fragmentation. In many examples the frequency-mass distributions of fragments are fractal. Explosive fragmentation of rocks (for example in mining) give fractal distributions. At the largest scale the frequency size distribution of the tectonic plates of plate tectonics are reasonably well approximated by a power-law distribution. Fault gouge is generated by the grinding process due to earthquakes on a fault. The frequency-mass distribution of the gouge fragments is fractal. Grinding (comminution) processes are common in tectonics. Thus it is not surprising that fractal distributions are ubiquitous in geology.

As a specific example consider the frequency-mass distribution of asteroids. Direct measurements give a fractal distribution. Since asteroids are responsible for the impact craters on the moon, it is not surprising that the frequency-area distribution of lunar craters is also fractal.

Using evidence from the moon and a fractal extrapolation it is estimated that on average, a 1m diameter meteorite impacts the earth every year, that a 100 m diameter meteorite impacts every 10,000 years, and that a 10 km diameter meteorite impacts the earth every 100,000,000 years. The classic impact crater is Meteor Crater in Arizona, it is over 1 km wide and 200 m deep. Meteor Crater formed about 50,000 years ago and it is estimated that the impacting meteorite had a diameter of 30 m. The largest impact to occur in the 20th century was the June 30, 1908 Tunguska event in central Siberia. The impact was observed globally and destroyed over 1000 km^2 of forest. It is believed that this event was the result of a 30 m diameter meteorite that exploded in the atmosphere.

One of the major global extinctions occurred at the Cretaceous/Tertiary boundary 65 million years ago. Some 65% of the existing species were destroyed including dinosaurs. This extinction is attributed to a massive impact

at the Chicxulub site on the Yucatan Peninsula, Mexico. It is estimated that the impacting meteorite had a 10 km diameter. In addition to the damage done by impacts there is evidence that impacts on the oceans have created massive tsunamis. The fractal power-law scaling can be used to quantify the risk of future impacts.

Earthquakes

Earthquakes universally satisfy several scaling laws. The most famous of these is Gutenberg–Richter frequency-magnitude scaling. The magnitude M of an earthquake is an empirical measure of the size of an earthquake. If the magnitude is increased by one unit it is observed that the cumulative number of earthquakes greater than the specified magnitude is reduced by a factor of 10.

For the entire earth, on average, there is 1 magnitude 8 earthquake per year, 10 magnitude 7 earthquakes per year, and 100 magnitude 6 earthquakes per year. When magnitude is converted to rupture area a fractal relation is obtained. The numbers of earthquakes that occur in a specified region and time interval have a power-law dependence on the rupture area.

The validity of this fractal scaling has important implications for probabilistic seismic risk assessment. The number of small earthquakes that occur in a region can be extrapolated to estimate the risk of larger earthquakes [1]. As an example consider southern California. On average there are 30 magnitude 4 or larger earthquakes per year. Using the fractal scaling it is estimated that the expected intervals between magnitude 6 earthquakes will be 3 years, between magnitude 7 earthquakes will be 30 years, and between magnitude 8 earthquakes will be 300 years.

The fractal scaling of earthquakes illustrate a useful aspect of fractal distributions. The fractal distribution requires two parameters. The first parameter, the fractal dimension D (known as the b-value in seismology), gives the dependence of number on size (magnitude). For earthquakes the fractal dimension is almost constant independent of the tectonic setting. The second parameter gives the level of activity. For example, this can be the number of earthquakes greater than a specified magnitude in a region. This level of activity varies widely and is an accepted measure of seismic risk. The level is essentially zero in states like Minnesota and is a maximum in California.

Volcanic Eruptions

There is good evidence that the frequency-volume statistics of volcanic eruptions are also fractal [9]. Although it is difficult to quantify the volumes of magma and ash associated with older eruptions, the observations suggest that an eruption with a volume of 1 km^3 would be expected each 10 years, 10 km^3 each 100 years, and 100 km^3 each 1000 years. For example, the 1991 Mount Pinatubo, Philippines eruption had an estimated volume of about 5 km^3. The most violent eruption in the last 200 years was the 1815 Tambora, Indonesia eruption with an estimated volume of 150 km^3. This eruption influenced the global climate in 1816 which was known as the year without a summer. It is estimated that the Long Valley, California eruption with an age of about 760,000 years had a volume of about 600 km^3 and the Yellowstone eruptions of about 600,000 years ago had a volume of about 2000 km^3.

Although the validity of the power-law (fractal) extrapolation of volcanic eruption volumes to long periods in the past can be questioned, the extrapolation does give some indication of the risk of future eruptions to global climate. There is no doubt that the large eruptions that are known to have occurred on time scales of 10^5 to 10^6 years would have a catastrophic impact on global agricultural production.

Landslides

Landslides are a complex natural phenomenon that constitutes a serious natural hazard in many countries. Landslides also play a major role in the evolution of landforms. Landslides are generally associated with a trigger, such as an earthquake, a rapid snowmelt, or a large storm. The landslide event can include a single landslide or many thousands. The frequency-area distribution of a landslide event quantifies the number of landslides that occur at different sizes. It is generally accepted that the number of large landslides with area A has a power-law dependence on A with an exponent in the range 1.3 to 1.5 [5].

Unlike earthquakes, a complete statistical distribution can be defined for landslides. A universal fit to an inverse-gamma distribution has been found for a number of event inventories. This distribution has a power-law (fractal) behavior for large landslides and an exponential cut-off for small landslides. The most probable landslides have areas of about 40 m^2. Very few small landslides are generated.

As a specific example we consider the 11,111 landslides generated by the magnitude 6.7 Northridge (California) earthquake on January 17, 1994. The total area of the landslides was 23.8 km^2 and the area of the largest landslide was 0.26 km^2. The inventory of landslide areas had a good power-law dependence on area for areas greater than 10^3 m^2 (10^{-3} km^2). The number of landslides generated by earthquakes have a strong dependence on earthquake magnitude. Typically earthquakes with magnitudes M less than 4 do not generate any landslides [6].

Floods

Floods are a major hazard to many cities and estimates of flood hazards have serious economic implications. The standard measure of the flood hazard is the 100-year flood. This is quantified as the river discharge Q_{100} expected during a 100 year period. Since there is seldom a long enough history to establish Q_{100} directly, it is necessary to extrapolate smaller floods that occur more often.

One extrapolation approach is to assume that flood discharges are fractal (power-law) [3,19]. This scale invariant distribution can be expressed in terms of the ratio F of the peak discharge over a 10 year interval to the peak discharge over a 1 year interval, $F = Q_{10}/Q_1$. With self-similarity the parameter F is also the ratio of the 100 year peak discharge to the 10 year peak discharge, $F = Q_{100}/Q_{10}$. Values of F have a strong dependence on climate. In temperate climates such as the northeastern and northwestern US values are typically in the range $F = 2$–3. In arid and tropical climates such as the southwestern and southeastern US values are typically in the range $F = 4$–6.

The applicability of fractal concepts to flood forecasting is certainly controversial. In 1982, the US government adopted the log-Pearson type 3 (LP3) distribution for the legal definition of the flood hazard [20]. The LP3 is a thin-tailed distribution relative to the thicker tailed power-law (fractal) distribution. Thus the forecast 100 year flood using LP3 is considerably smaller than the forecast using the fractal approach. This difference is illustrated by considering the great 1993 Mississippi River flood. Considering data at the Keukuk, Iowa gauging station [4] this flood was found to be a typical 100 year flood using the power-law (fractal) analysis and a 1000 to 10,000 year flood using the federal LP3 formulation. Concepts of self-similarity argue for the applicability of fractal concepts for flood-frequency forecasting. This applicability also has important implications for erosion. Erosion will be dominated by the very largest floods.

Self-Affine Fractals

Mandelbrot and Van Ness [8] extended the concept of fractals to time series. Examples of time series in geology and geophysics include global temperature, the strength of the Earth's magnetic field, and the discharge rate in a river. After periodicities and trends have been removed, the remaining values are the stochastic (noise) component of the time series. The standard approach to quantifying the noise component is to carry out a Fourier transform on the time series [2]. The power-spectral density coefficients S_i are proportional to the squares of the Fourier coefficients. The time series is a self-affine fractal if the power-spectral density coefficients have an inverse power-law dependence on frequency f_i, that is

$$S_i = \frac{C}{f_i^{\beta}} \tag{5}$$

where C is a constant and β is the power-law exponent.

For a Gaussian white noise the values in the time series are selected randomly from a Gaussian distribution. Adjacent values are not correlated with each other. In this case the spectrum is flat and the power spectral density coefficients are not a function of frequency, $\beta = 0$. The classic example of a self-affine fractal is a Brownian walk. A Brownian walk is obtained by taking the running sum of a Gaussian white noise. In this case we have $\beta = 2$. Another important self-affine time series is a red (or pink) noise with power spectral density coefficients proportional to $1/f$, that is $\beta = 1$. We will see that the variability in the Earth's magnetic field is well approximated by a $1/f$ noise.

Self-affine fractal time series in the range $\beta = 0$ to 1 are known as fractional Gaussian noises. These noises are stationary and the standard deviation is a constant independent of the length of the time series. Self-affine time series with β larger than 1 are known as fractional Brownian walk. These motions are not stationary and the standard deviation increases as a power of the length of the time series, there is a drift. For a Brownian walk the standard deviation increases with the square root of the length of the time series.

Topography

The height of topography along linear tracks can be considered to be a continuous time series. In this case we consider the wave number k_i (1/wave length) instead of frequency. Topography is a self-affine fractal if

$$S_i = \frac{C}{k_i^{\beta}} \tag{6}$$

Spectral expansions of global topography have been carried out, an example [15] is given in Fig. 2. Excellent agreement with the fractal relation given in Eq. (6) is obtained taking $\beta = 2$, topography is well approximated by a Brownian walk. It has also shown that this fractal behavior of topography is found for the moon, Venus, and Mars [18].

Earth's Magnetic Field

Paleomagnetic studies have given the strength and polarity of the Earth's magnetic field as a function of time over millions of years. These studies have also shown that the field has experienced a sequence of reversals.

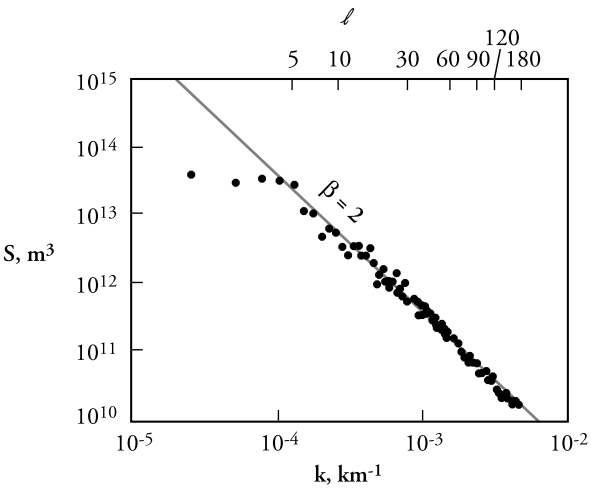

Fractals in Geology and Geophysics, Figure 2
Power spectral density S as a function of wave number k for a spherical harmonic expansion of the Earth's topography (degree l). The straight-line correlation is with Eq. (6) taking $\beta = 2$, a Brownian walk

Spectral studies of the absolute amplitude of the field have been shown that it is a self-affine fractal [12,14]. The power-spectral density is proportional to one over the frequency, it is a $1/f$ noise. When the fluctuations of the $1/f$ noise take the magnitude to zero the polarity of the field reverses. The predicted distribution of polarity intervals is fractal and is in good agreement with the observed polarity intervals.

Future Directions

There is no question that fractals are a useful empirical tool. They provide a rational means for the extrapolation and interpolation of observations. The wide applicability of power-law (fractal) distributions is generally accepted, but does this applicability have a more fundamental basis? Fractality appears to be fundamentally related to chaotic behavior and to numerical simulations exhibiting self-organized criticality. The entire area of fractals, chaos, self-organized criticality, and complexity remains extremely active, and it is impossible to predict with certainty what the future holds.

Bibliography

Primary Literature

1. Kossobokov VG, Keilis-Borok VI, Turcotte DL, Malamud BD (2000) Implications of a statistical physics approach for earthquake hazard assessment and forecasting. Pure Appl Geophys 157:2323
2. Malamud BD, Turcotte DL (1999) Self-affine time series: I. Generation and analyses. Adv Geophys 40:1
3. Malamud BD, Turcotte DL (2006) The applicability of power-law frequency statistics to floods. J Hydrol 332:168
4. Malamud BD, Turcotte DL, Barton CC (1996) The 1993 Mississippi river flood: A one hundred or a one thousand year event? Env Eng Geosci 2:479
5. Malamud BD, Turcotte DL, Guzzetti F, Reichenbach P (2004) Landslide inventories and their statistical properties. Earth Surf Process Landf 29:687
6. Malamud BD, Turcotte DL, Guzzetti F, Reichenbach P (2004) Landslides, earthquakes, and erosion. Earth Planet Sci Lett 229:45
7. Mandelbrot BB (1967) How long is the coast of Britain? Statistical self-similarity and fractional dimension. Science 156:636
8. Mandelbrot BB, Van Ness JW (1968) Fractional Brownian motions, fractional noises and applications. SIAM Rev 10:422
9. McClelland L et al (1989) Global Volcanism 1975-1985. Prentice-Hall, Englewood Cliffs
10. Meybeck M (1995) Global distribution of lakes. In: Lerman A, Imboden DM, Gat JR (eds) Physics and Chemistry of Lakes, 2nd edn. Springer, Berlin, pp 1-35
11. Peckham SD (1989) New results for self-similar trees with applications to river networks. Water Resour Res 31:1023
12. Pelletier JD (1999) Paleointensity variations of Earth's magnetic field and their relationship with polarity reversals. Phys Earth Planet Int 110:115
13. Pelletier JD (1999) Self-organization and scaling relationships of evolving river networks. J Geophys Res 104:7259
14. Pelletier JD, Turcotte DL (1999) Self-affine time series: II. Applications and models. Adv Geophys 40:91
15. Rapp RH (1989) The decay of the spectrum of the gravitational potential and the topography of the Earth. Geophys J Int 99:449
16. Strahler AN (1957) Quantitative analysis of watershed geomorphology. Trans Am Geophys Un 38:913
17. Tokunaga E (1978) Consideration on the composition of drainage networks and their evolution. Geogr Rep Tokyo Metro Univ 13:1
18. Turcotte DL (1987) A fractal interpretation of topography and geoid spectra on the earth, moon, Venus, and Mars. J Geophys Res 92:E597
19. Turcotte DL (1994) Fractal theory and the estimation of extreme floods. J Res Natl Inst Stand Technol 99:377
20. US Water Resources Council (1982) Guidelines for Determining Flood Flow Frequency. Bulletin 17B. US Geological Survey, Reston

Books and Reviews

Feder J (1988) Fractals. Plenum Press, New York
Korvin G (1992) Fractal Models in the Earth Sciences. Elsevier, Amsterdam
Mandelbrot BB (1982) The Fractal Geometry of Nature. Freeman, San Francisco
Turcotte DL (1997) Fractals and Chaos in Geology and Geophysics, 2nd edn. Cambridge University Press, Cambridge

Fractals Meet Chaos

Tony Crilly
Middlesex University, London, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Dynamical Systems
Curves and Dimension
Chaos Comes of Age
The Advent of Fractals
The Merger
Future Directions
Bibliography

Glossary

Dimension The traditional meaning of dimension in modern mathematics is "topological dimension" and is an extension of the classical Greek meaning. In modern concepts of this dimension can be defined in terms of a separable metric space. For the practicalities of Fractals and Chaos the notion of dimension can be limited to subsets of Euclidean n-space where n is an integer. The newly arrived "fractal dimension" is *metrically* based and can take on fractional values. Just as for topological dimension.

As for topological dimension itself there is a profusion of different (but related) concepts of metrical dimension. These are widely used in the study of fractals, the ones of principal interest being:

- Hausdorff dimension (more fully, Hausdorff–Besicovitch dimension),
- Box dimension (often referred to as Minkowski–Bouligand dimension),
- Correlation dimension (due to A. Rényi, P. Grassberger and I. Procaccia).

Other types of metric dimension are also possible. There is "divider dimension" (based on ideas of an English mathematician/meteorologist L. F. Richardson in the 1920s); the "Kaplan–Yorke dimension" (1979) derived from Lyapunov exponents, known also as the "Lyapunov dimension"; "packing dimension" introduced by Tricot (1982). In addition there is an overall *general* dimension due to A. Rényi (1970) which admits box dimension, correlation dimension and information dimension as special cases. With many of the concepts of dimension there are upper and lower refinements, for example, the separate *upper* and *lower* box dimensions. Key references to the vast (and highly technical) subject of mathematical dimension include [31,32,33,60,73,92,93].

Hausdorff dimension (Hausdorff–Besicovitch dimension). In the study of fractals, the most sophisticated concept of dimension is Hausdorff dimension, developed in the 1920s.

The following definition of Hausdorff dimension is given for a subset A of the real number line. This is readily generalized to subsets of the plane, Euclidean 3-space and Euclidean n-space, and more abstractly to separable metric spaces by taking neighborhoods as disks instead of intervals. Let $\{U_i\}$ be an *r-covering* of A, (a covering of A where the width of *all* intervals U_i, satisfies $w(U_i) \leq r$). The measure m_r is defined by

$$m_r(A) = \inf\left(\sum_{i=1}^{\infty} w(U_i)\right) ,$$

where the infimum (or greatest of the minimum values) is taken over all *r-coverings* of A.

The Hausdorff dimension D_{H} of A is:

$$D_{\mathrm{H}} = \lim_{r\to 0} m_r(A) ,$$

provided the limit exists. The subset $E = \{1/n\colon n = 1, 2, , 3, \ldots\}$ of the unit interval has $D_{\mathrm{H}} = 0$ (the Hausdorff dimension of a countable set is always zero). The Hausdorff dimension is the basis of "fractal dimension" but because it takes into account intervals of unequal widths it may be difficult to calculate in practice.

Box dimension (or Minkowski–Bouligand dimension, known also as capacity dimension, cover dimension, grid dimension). The box counting dimension is a more direct and practical method for computing dimension in the case of fractals. To define it, we again confine our attention to the real number line in the knowledge that box dimension is readily extended to subsets of more general spaces.

As before, let $\{U_i\}$ be an *r-covering* of A, and let $N_r(A)$ be the least number of sets in such a covering. The box dimension D_{B} of A is defined by:

$$D_{\mathrm{B}} = \lim_{r\to 0} \frac{\log N_r(A)}{\log 1/r} .$$

The box dimension of the subset $E = \{1/n\colon n = 1, 2, 3, \ldots\}$ of the unit interval can be calculated to give $D_{\mathrm{B}} = 0.5$.

In general Hausdorff and Box dimensions are related to each other by the inequality $D_B \geq D_H$, as happens in the above example. The relationship between D_H and D_B is investigated in [49]. For compact, self-similar fractal sets $D_B = D_H$ but there are fractal sets for which $D_B > D_H$ [76]. Though Hausdorff dimension and Box dimension have similar properties, Box dimension is only finitely additive, while Hausdorff dimension is countably additive.

Correlation dimension For a set of n points A on the real number line, let $P(r)$ be the probability that two different points of A chosen at random are closer than r apart. For a large number of points n, the graph of $v = \log P(r)$ against $u = \log r$ is approximated to its slope for small values of r, and theoretically, a straight line. The correlation dimension D_C is defined as its slope for small values of r, that is,

$$D_C = \lim_{r \to 0} \frac{dv}{du} .$$

The correlation dimension involves the separation of points into "boxes", whereas the box dimension merely counts the boxes that cover A.

If P_i is the probability of a point of A being in box i (approximately n_i/n where n_i is the number of points in box i and n is the totality of points in A) then an alternative definition of correlation dimension is

$$D_C = \lim_{r \to 0} \frac{\log \sum_i P_i^2}{\log r} .$$

Attractor A point set in phase space which "attracts" trajectories in its vicinity. More formally, a bounded set A in phase space is called an attractor for the solution $x(t)$ of a differential equation if

- $x(0) \in A \Rightarrow x(t) \in A \;\; for\ all \;\; t$. Thus, an attractor is invariant under the dynamics (trajectories which start in A remain in A).
- There is a neighborhood $U \supset A$ such that any trajectory starting in U is attracted to A (the trajectory gets closer and closer to A).
- If $B \subset A$ and if B satisfies the above two properties then $B = A$.

An attractor is therefore the minimal set of points A which attracts all orbits starting at some point in a neighborhood of A.

Orbit is a sequence of points $\{x_i\} = x_0, x_1, x_2, \ldots$ defined by an iteration $x_n = f^n(x_0)$. If n is a positive it is called a forwards orbit, and if n is negative a backwards orbit. If $x_0 = x_n$ for some finite value n, the orbit is periodic. In this case, the smallest value of n for which this is true is called the period of the orbit.

For an invertible function f, a point x is homoclinic to a if

$$\lim f^n(x) = \lim f^{-n}(x) = a \quad \text{as} \;\; n \to \infty .$$

and in this case the orbit $\{f^n(x_0)\}$ is called a homoclinic orbit – the orbit which converges to the same saddle point a forwards or backwards. This term was introduced by Poincaré.

The terminology "orbit" may be regarded as applied to the solution of a difference equation, in a similar way the solution of a differential equation $x(t)$ is termed a trajectory. Orbit is the term used for discrete dynamical system and trajectory for the continuous time case.

Basin of attraction If a is an attractive fixed point of a function f, its basin of attraction $B(a)$ is the subset of points defined by

$$B(a) = \{x\colon f^k(x) \to a\,, \text{ as } k \to \infty\} .$$

It is the subset containing all the initial points of orbits attracted to a. The basins of attraction may have a complicated structure. An important example applies to the case where a is a point in the complex plane C.

Julia set A set J_f is the boundary between the basins of attraction of a function f. For example, in the case where $z = \pm 1$ are attracting points (solutions of $z^2 - 1 = 0$), the Julia set of the "Newton–Fourier" function $f(z) = z - ((z^2 - 1)/2z)$ is the set of complex numbers which lie along the imaginary axis $x = 0$ (as proved by Schröder and Cayley in the 1870s). The case of the Julia set involved with the solutions of $z^3 - 1 = 0$ was beyond these pioneers and is fractal in nature. An alternative definition for a Julia set is the closure of the subset of the complex plane whose orbits of f tend to infinity.

Definition of the Subject

Though "Chaos" and "Fractals" are yoked together to form "Fractals and Chaos" they have had separate lines of development. And though the names are modern, the mathematical ideas which lie behind them have taken more than a century to gain the prominence they enjoy today. Chaos carries an applied connotation and is linked to differential equations which model physical phenomena. Fractals is directly linked to subsets of Euclidean space which have a fractional dimension, which may be obtained by the iteration of functions.

This brief survey seeks to highlight some of the significant points in the history of both of these subjects. There are brief academic histories of the field. A history of

Chaos has been shown attention [4,46], while an account of the early history of the iteration of complex functions (up to Julia and Fatou) is given in [3]. A broad survey of the whole field of fractals is given in [18,50]. Accounts of Chaos and Fractals tend to concentrate on one side at the expense of the other. Indeed, it is only quite recently that the two subjects have been seen to have a substantial bearing on each other [72].

This account treats a "prehistory" and a "modern period". In the prehistory period, before about 1960, topics which contributed to the modern theory are not so prominent. There is a lack of impetus to create a new field. In the modern period there is a greater sense of continuity, driven on by the popular interest in the subject. The modern theory coincided with a rapid increase in the power of computers unavailable to the early workers in the prehistory period. Scientists, mathematicians, and a wider public could now "see" the beautiful geometrical shapes displayed before them [25].

The whole theory of "fractals and chaos" necessarily involves nonlinearity. It is a mathematical theory based on the properties of processes which are assumed to be modeled by nonlinear differential equations and nonlinear functions. Chaos shows itself when solutions to these differential equations become unstable. The study of stability is rooted in the mathematics of the nineteenth century. Fractals are derived from the geometric study of curves and sets of points generally, and from abstract iterative schemes. The modern theory of fractals is the outcome of explorations by mathematicians and scientists in the 1960s and 1970s, though, as we shall see, it too has an extensive prehistory.

Recently, there has been an explosion of published work in Fractals and Chaos. Just in Chaos theory alone a bibliography of six hundred articles and books compiled by 1982, grew to over seven thousand by the end of 1990 [97]. This is most likely an excessive underestimate. Fractals and Chaos has since grown into a wide-ranging and variegated theory which is rapidly developing. It has widespread applications in such areas as

Astronomy the motions of planets and galaxies

Biology population dynamics chemistry chemical reactions

Economics time series, analysis of financial markets

Engineering capsize of ships, analysis of road traffic flow

Geography measurement of coastlines, growth of cities, weather forecasting

Graphic art analysis of early Chinese Landscape Paintings, fractal geometry of music

Medicine dynamics of the brain, psychology, heart rhythms.

There are many works which address the application of Chaos and Fractals to a broad sweep of subjects. In particular, see [13,19,27,59,63,64,87,96].

Introduction

"Fractals and Chaos" is the popular name for the subject which burst onto the scientific scene in the 1970s and 1980s and drew together practical scientists and mathematicians. The subject reached the popular ear and perhaps most importantly, its eye. Computers were becoming very powerful and capable of producing remarkable visual images. Quite elementary mathematics was capable of creating beautiful pictures that the world had never seen before.

The popular ear was captivated – momentarily at least – by the neologisms created for the theory. "Chaos" was one, but "fractals", the "horseshoe", and the superb "strange attractors" became an essential part of the scientific vocabulary. In addition, these were being passed around by those who dwelled far from the scientific front. It seemed all could take part, from scientific and mathematical researchers to lay scientists and amateur mathematicians. All could at least experience the excitement the new subject offered. Fractals and Chaos also carried implications for the philosophy of science.

The widespread interest in the subject owes much to articles and books written for the popular market. J. Gleick's *Chaos* was at the top of the heap in this respect and became a best seller. He advised his readership that chaos theory was one of the great discoveries of the twentieth century and quoted scientists who placed chaos theory alongside the revolutions of Relativity and Quantum Mechanics. Gleick claimed that "chaos [theory] eliminates the Laplacean fantasy of deterministic predictability". In a somewhat speculative flourish, he reasoned: "Of the three theories, the revolution in chaos applies to the universe we see and touch, to objects at human scale. Everyday experience and real pictures of the world become legitimate targets for inquiry. There has long been a feeling, not always expressed openly, that theoretical physics has strayed far from human intuition about the world [38]."

More properly chaos is "deterministic chaos". The equations and functions used to model a dynamical system are stated exactly. The contradictory nature of chaos is that the mathematical solutions appeared to be random. On the one hand the situation is deterministic, but on the other there did not seem to be any order in the solutions. Chaos theory resolves this difficulty by conceptualizing the notion of orbits, trajectories phase space, attractors, and fractals. What appeared paradoxical seen through the

old lenses offered explanation through a more embracing mathematical theory.

In the scientific literature the modern Chaos is presented through "dynamical systems", and this terminology gives us a clue to its antecedents. Dynamical systems conveys the idea of a physical system. It is clear that mechanical problems, for example, those involving motion are genuine dynamical systems because they evolve in space through time. Thus the pendulum swings backwards and forwards in time, planets trace out orbits, and the beat of a heart occurs in time.

The differential equations which describe physical dynamical system give rise to chaos, but how do fractals enter the scene? In short, the trajectories in the phase space which describes the physical system through a process of spreading and folding pass through neighborhoods of the attractor set infinitely often, and on magnification reveal them running closer and closer together. This is the fine structure of the attractor. Measuring the metric dimension of this set results in a fraction and it is there that the connection with fractals is made. But this is running ahead of the story.

Dynamical Systems

The source of "Chaos" lies in the analysis of physical systems and goes back to the eighteenth century and the work of the Swiss mathematician Leonhard Euler. The most prolific mathematician of all time, Euler (whose birth tercentenary occurred in 2007) was amongst natural philosophers who made a study of differential equations in order to solve practical mechanical and astronomical problems. Chief among the problems is the problem of fluid flow which occurs in hydrodynamics.

Sensitive Initial Conditions

The key characteristic of "chaotic solutions" is their sensitivity to initial conditions: two sets of initial conditions close together can generate very different solution trajectories, which after a long time has elapsed will bear very little relation to each other. Twins growing up in the same household will have a similar life for the childhood years but their lives may diverge completely in the fullness of time. Another image used in conjunction with chaos is the so-called "butterfly effect" – the metaphor that the difference between a butterfly flapping its wings in the southern hemisphere (or not) is the difference between fine weather and hurricanes in Europe. The butterfly effect notion most likely got its name from the lecture E. Lorenz gave in Washington in 1972 entitled "Predictability: Does the Flap of a Butterfly's wings in Brazil Set off a Tornado in Texas?" [54]. An implication of chaos theory is that prediction in the long term is impossible for we can never know for certain whether the "causal" butterfly really did flap its wings.

The sensitivity of a system to initial conditions, the hallmark of what makes a chaotic solution to a differential equation is derived from the stability of a system. Writing in 1873, the mathematical physicist James Clerk Maxwell alluded to this sensitivity in a letter to man of science Francis Galton:

> Much light may be thrown on some of these questions [of mechanical systems] by the consideration of stability and instability. When the state of things is such that an infinitely small variation of the present state will alter only by an infinitely small quantity the state at some future time, the condition of the system, whether at rest or in motion, is said to be stable; but when an infinitely small variation in the present state may bring about a finite difference in the state of the system in a finite time, the condition is said to be unstable [44].

H. Poincaré too, was well aware that small divergences in initial conditions could result in great differences in future outcomes, and said so in his discourse on *Science and Method*: "It may happen that slight differences in the initial conditions produce very great differences in the final phenomena; a slight error in the former would make an enormous error in the latter. Prediction becomes impossible." As an example of this he looked at the task of weather forecasting:

> Why have the meteorologists such difficulty in predicting the weather with any certainty? Why do the rains, the tempests seem to us to come by chance, so that many persons find it quite natural to pray for rain or shine, when they would think it ridiculous to pray for an eclipse? We see that great perturbations generally happen in regions where the atmosphere is in unstable equilibrium. … one tenth of a degree more or less at any point, and the cyclone bursts here and not there, and spreads its ravages over countries it would have spared. … Here again we find the same contrast between a very slight cause, unappreciable to the observer, and important effects, which are sometimes tremendous disasters [75].

So here is the answer to one conundrum – Poincaré is the true author of the butterfly effect! Weather forecasting is ultimately a problem of fluid flow in this case air flow.

Fluid Motion and Turbulence

The study of the motion of fluids predates Poincaré and Lorenz. Euler published "General Principles of the Motion of Fluids" in 1755 in which he wrote down a set of partial differential equations to describe the motion of non-viscous fluids. These were improved by the French engineer C. Navier when in 1821 he published a paper which took viscosity into account. From a different starting point in 1845 the British mathematical physicist G. G. Stokes derived the same equations in a paper entitled "On the Theories of the Internal Friction of Fluids in Motion."

These are the Navier–Stokes equations, a set of nonlinear partial differential equations which generally apply to fluid motion and which are studied by the mathematical physicist. In modern vector notation (in the case of unforced incompressible flow) they are

$$\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v}.\nabla\mathbf{v} = -\frac{1}{\rho}\nabla p + \frac{\mu}{\rho}\nabla^2\mathbf{v}\,,$$

where $\mathbf{v}$ is velocity, p is pressure, and ρ and μ are density and viscosity constants. It is the nonlinearity of the Navier–Stokes equations which makes them intractable, the nonlinearity manifested by the "products of terms" like $\mathbf{v}.\nabla\mathbf{v}$ which occur in them. They have been studied intensively since the nineteenth century particularly in special forms obtained by making simplifying assumptions.

L. F. Richardson, who enters the subject of Fractals and Chaos at different points made attempts to solve nonlinear differential equations by numerical methods. In the 1920s, Richardson adapted words from Gulliver's Travels in one of his well-known refrains on turbulence: "big whirls have little whirls that feed on their velocity, and little whirls have lesser whirls and so on to viscosity – in the molecular sense" [79]. This numerical work was hampered by the lack of computing power. The only computers available in the 1920s were human ones, and the traditions of using paid "human computers" was still in operation. Richardson visualized an orchestra of human computers harmoniously carrying out the vast array of calculations under the baton of a mathematician. There were glimmers of all this changing, and during the 1920s the appearance of electrical devices gave an impetus to the mathematical study of both numerical analysis and the study of nonlinear differential equations. John von Neumann for one saw the need for the electronic devices as an aid to mathematics. Notwithstanding this development, the problem of fluid flow posed intrinsic mathematical difficulties.

In 1932, Horace Lamb addressed the British Association for the Advancement of Science, with a prophetic statement dashed with his impish touch of humor: "I am an old man now, and when I die and go to Heaven there are two matters on which I hope for enlightenment. One is quantum electro-dynamics, and the other is the turbulent motion of fluids. And about the former I am really rather optimistic." Forty years on Werner Heisenberg continued in the same vein. He certainly knew about quantum theory, having invented it, but chose relativity as the competing theory with turbulence for his own lamb's tale. On his death bed it is said he compared quantum theory with turbulence and is reputed to have singled out turbulence as of the greater difficulty.

In 1941, A. Kolmogorov, the many-sided Russian mathematician published two papers on problems of turbulence caused by the jet engine and astronomy. Kolmogorov also made contributions to probability and topology though these two papers are the ones rated highly by fluid dynamicists and physicists. In 1946 he was appointed to head the Turbulence Laboratory of the Academic Institute of Theoretical Geophysics. With Kolmolgov, an influential school of mathematicians that included L. S. Pontryagin, A. A. Andronov, D. V. Anosov and V. I. Arnol'd became active in Russia in the field of dynamical systems [24].

Qualitative Differential Equations and Topology

Poincaré's qualitative study of differential equations pioneered the idea of viewing the solution of differential equations as curves rather than functions and replacing the local with the global. Poincaré's viewpoint was revolutionary. As he explained it in *Science and Method*: "formerly an equation was considered solved only when its solution had been expressed by aid of a finite number of known functions; but that is possible scarcely once in a hundred times. What we always can do, or rather what we should always seek to do, is to solve the problem *qualitatively* [his italics] so to speak; that is to say; seek to know the general form of the curve [trajectory] which represents the unknown function."

For the case of the plane, $m = 2$, for instance, what do the solutions to differential equations look like across the whole plane, viewing them as trajectories $x(t)$ starting at an initial point? This contrasts with the traditional view of solving differential equations whereby specific functions are sought which satisfy initial and boundary conditions.

Poincaré's attack on the three body problem (the motion of the moon, earth, sun, is an example) was stimulated by a prize offered in 1885 to commemorate the sixtieth birthday of King Oscar II of Sweden. The problem set was for an n-body problem but Poincaré's essay on the re-

stricted three-body problem was judged so significant that he was awarded the prize in January 1889. Before publication, Poincaré went over his working and found a significant error. It proved a profound error. According to Barrow-Green his description of doubly asymptotic trajectories is the first mathematical description of chaotic motion in a dynamical system [6,7]. Poincaré introduced the "Poincaré section" the surface section that transversely intersects the trajectory in phase space and defines a sequence of points on it. In the case of the damped pendulum, for example, a sequence of points is obtained as the spiral trajectory in phase space hits the section. This idea brings a physical dynamical systems problem in conjunction with topology, a theme developed further by Birkhoff in 1927 [1,10].

Poincaré also introduced the idea of phase space but his study mainly revolves around periodic trajectories and not the non periodic ones typical of chaos. Poincaré said of periodic orbits, that they were "the only gap through which may attempt to penetrate, into a place where up to now was reputed to be unreachable". The concentration on periodic trajectories has a parallel in nearby pure mathematics – where irregular curves designated as "monsters" were ignored in favor of "normal curves".

G. D. Birkhoff learned from Poincaré. It was said that, apart from J. Hadamard, no other mathematician knew Poincaré's work as well as Birkhoff did. Like his mentor, Birkhoff adopted phase space as his template, and emphasised periodic solutions and treated conservative systems and not dissipative systems (such as the damped pendulum which loses energy). Birkhoff spent the major part of his career, between 1912 and 1945 contributing to the theory of dynamical systems. His aim was to provide a qualitative theory which characterized equilibrium points in connection with their stability. A dynamical system was defined by a set of n differential equations $\mathrm{d}x_i/\mathrm{d}t = F_i(x_1, \ldots, x_n)$ defined locally. An interpretation of these could be a situation in chemistry where $x_1(t), \ldots, x_n(t)$ might be the concentrations of n chemical reactants at time t where the initial values $x_1(0), \ldots, x_n(0)$ are given, though applications were not Birkhoff's main concern.

In 1941, towards the end of his career, Birkhoff reappraised the field in "Some unsolved problems of theoretical dynamics" Later M. Morse discussed these and pointed out that Birkhoff listed the same problems he had considered in 1920. In 1920 Birkhoff had written about dynamical systems in terms of the "general analysis" propounded by E. H. Moore in Chicago where he had been a student. In the essay of 1941 modern topological language was used:

> As was first realized about fifty years ago by the great French mathematician, Henri Poincaré, the study of dynamical systems (such as the solar system) leads directly to extraordinary diverse and important problems in point-set theory, topology and the theory of functions of real variables.

The idea was to describe the phase space by an abstract topological space. In his talk of 1941, Birkhoff continued:

> The kind of abstract space which it seems best to employ is a compact metric space. The individual points represent "states of motion", and each curve of motion represents a complete motion of the abstract dynamical system [11].

Using Birkhoff's reappraisal, Morse set out future goals: " 'Conservative flows' are to be studied both in the topological and the statistical sense, and abstract variational theory is to enter. There is no doubt about the challenge of the field, and the need for a powerful and varied attack [68]."

Duffing, Van der Pol and Radar

It is significant that chaos theory was first derived from practical problems. Poincaré was an early pioneer with a problem in astronomy but other applications shortly arrived. C. Duffing (1918) introduced a second order nonlinear differential equation which described a mechanical oscillating device. In a simple form of it (with zero forcing term):

$$\mathrm{d}^2x/\mathrm{d}t^2 + \alpha \mathrm{d}x/\mathrm{d}t + (\beta x^3 + \gamma x) = 0\,,$$

Duffing's equation exhibits chaotic solutions. B. van der Pol working at the Radio Scientific Research group at the Philips Laboratory at Eindhoven, described "irregular noise" in an electronic diode. Van der Pol's equations of the form (with right hand side forcing term):

$$\mathrm{d}^2x/\mathrm{d}t^2 + k(x^2 - 1)\mathrm{d}x/\mathrm{d}t + x = A\cos\Omega t$$

described "relaxational" oscillations or arrhythmic beats of an electrical circuit. Such "relaxational" oscillations are of the type which also occur in the beating of the heart. Richardson noted the transition from periodic solutions to van der Pol's equation (1926) to unstable solutions. Both Duffing's equation and Van der Pol's equation play an important part in chaos theory.

In 1938 the English mathematician Mary Cartwright answered a call for help from the British Department of Scientific and Industrial Research. A solution to the differential equations connected with the new radar technology

was wanted, and van der Pol's equation was relevant. It was in the connection this equation that she collaborated with J. E. Littlewood in the 1940s and made discoveries which presaged modern Chaos.

The mathematical difficulty was caused by the equation being nonlinear but otherwise the equation appear nondescript. Yet in mathematics, the nondescript can yield surprises. A corner of mathematics was discovered that Cartwright described as "a curious branch of mathematics developed by different people from different standpoints – straight mechanics, radio oscillations, pure mathematics and servo-mechanisms of automatic control theory [65]" The mathematician and physicist Freeman Dyson wrote of this work and its practical significance:

> Cartwright had been working with Littlewood on the solutions of the equation, which describe the output of a nonlinear radio amplifier when the input is a pure sine-wave. The whole development of radio in World War II depended on high power amplifiers, and it was a matter of life and death to have amplifiers that did what they were supposed to do. The soldiers were plagued with amplifiers that misbehaved, and blamed the manufacturers for their erratic behavior. Cartwright and Littlewood discovered that the manufacturers were not to blame. The equation itself was to blame. They discovered that as you raise the gain [the ratio of output to input] of the amplifier, the solutions of the equation become more and more irregular. At low power the solution has the same period as the input, but as the power increases you see solutions with double the period, and finally you have solutions that are not periodic at all [30].

The story is now familiar: there is the phenomenon of period doubling of solutions followed by chaotic solutions as the gain of the amplifier is raised still higher. A further contribution to the theory of this equation was made by N. Levinson of the Massachusetts Institute of Technology in the United States.

Curves and Dimension

The subject of "Fractals" is a more recent development. First inklings of them appeared in "Analysis situs" in the latter part of the nineteenth century when the young subject of topology was gaining ground. Questions were being asked about the properties of sets of points in Euclidean spaces, the nature of curves, and the meaning of dimension itself.

Crinkly Curves

In 1872, K. Weierstrass introduced the famous function defined by a convergent series:

$$f(x) = \sum_{k=0}^{\infty} b^k \cos(a^k \pi x) \quad (a > 1,\ \ 0 < b < 1,\ \ ab > 1),$$

which was continuous everywhere but differentiable nowhere. It was the original "crinkly curve" but in 1904, H. von Koch produced a much simpler one based only on elementary geometry. While the idea of a function being continuous but not differentiable could be traced back to A-M Ampère at the beginning of the nineteenth century, von Koch's construction has similarities with examples produced by B. Bolzano. Von Koch's curve has become a classic half-way between a "monster curves" and regularity – an example of a curve of infinite length which encloses a finite area, as well as being an iconic fractal.

In retrospect G. Cantor's middle third set (also discovered by H.J.S. Smith (1875) a professor at Oxford), is also a fractal. It is a totally disconnected and uncountable set, with the curiosity that after subtractions of the middle-third segments from the unit interval the ultimate Cantor set has same cardinality as the original unit interval.

At the beginning of the twentieth century searching questions about the theory of curves were being asked. The very basic question "what is a curve" had been brought to life by G. Peano by his space filling curve which is defined in accordance with Jordan's definition of a curve but fills out a "two-dimensional square." Clearly the theory of dimension needed serious attention for how could an ostensibly two-dimensional "filled-in square" be a curve?

The Iteration of Functions

A principal source of fractals is obtained by the iteration of functions, what is now called a discrete dynamical system, or a system of symbolic dynamics [15,23].

One of the earliest forays in this field was made by the English mathematician A. Cayley, but he was not the first as D. S. Alexander has pointed out. F.W. K. E. (Ernst) Schröder anticipated him, and may even have been the inspiration for Cayley's attraction to the problem.

Is there a link between the two mathematicians? Schröder studied at Heidelberg and where L. O. Hesse was his doctoral adviser. Hesse who contributed to algebra, geometry, and invariant theory and was known to Cayley. Nowadays Schröder is known for his contributions to logic but in 1871 he published "Ueber iterite Functionen" in the newly founded *Mathematische Annalen*. The principal objective of this journal was the publication of articles

on invariant theory then a preoccupation of the English and German schools of mathematics, and Cayley was at the forefront of this research.

So did Cayley know of Schröder's work? Cayley had an appetite for all things (pure) mathematical and had acquired encyclopedic knowledge of the field, just about possible in the 1870s. In 1871 Cayley published an article ("Note on the theory of invariants") in the same volume and issue of the *Mathematische Annalen* in which Schröder's article appeared, and actually published three articles in the volume. In this period of his life, when he was in his fifties he allowed his interests full rein and became a veritable magpie of mathematics. He covered a wide range of topics, too many to list here, but his contributions were invariably short. Moreover he dropped invariant theory at this point and only resumed in 1878 when J. J. Sylvester reawakened his interest. There was time to rediscover the four color problem and perhaps Schröder's work on the iteration of functions. It was a field where Schroeder had previously "encountered very few collaborators." The custom of English mathematicians citing previous work of others began in the 1880s and Cayley was one of the first to do this. The fact that Cayley did not cite Schröder is not significant.

In February 1879, Cayley wrote to Sir William Thomson (the later Lord Kelvin) about the Newton–Fourier method of finding the root of an equation, named after I. Newton (c.1669) and J. Fourier (1818) that dealt with the real variable version of the method. This achieved a degree of significance in Cayley's mind, for a few days later he wrote to another colleague about it:

> I have a beautiful question which is bothering me – the extension of the Newton–Fourier method of approximation to imaginary values: it is very easy and pretty for a quadric equation, but I do not yet see how it comes out for a cubic. The general notion is that the plane must be divided into regions; such that starting with a point P in one of these say the A-region … [his ellipsis], the whole series of derived points $P_1, P_2, P_3, \ldots$ up is P_∞ (which will be the point A) lies in this [planar] region; … and so for the B. and C. regions. But I do not yet see how to find the bounding curves [of these regions] [21].

So Cayley's regions are the modern basins of attraction for the point A, the bounding curves now known as Julia sets. He tried out the idea before the Cambridge Philosophical Society, and by the beginning of March had done enough to send the problem for publication:

> In connexion herewith, throwing aside the restrictions as to reality, we have what I call the Newton–Fourier Imaginary Problem, as follows.
>
> Take $f(u)$ a given rational and integral function [a polynomial] of u, with real or imaginary coefficients; z, a given real or imaginary value, and from this derive z_1 by the formula
>
> $$z_1 = z - \frac{f(z)}{f'(z)}$$
>
> and thence $z_1, z_2, z_3, \ldots$ each from the preceding one by the like formula. … The solution is easy and elegant in the case of a quadric equation: but the next succeeding case of the cubic equation appears to present considerable difficulty [17].

Cayley's connection with German mathematicians was close in the 1870s, and later on in 1879 (and in 1880) he went on tours of Germany, and visited mathematicians. No doubt he took the problem with him, and it was one he returned to periodically.

Both Cayley and Schröder solved this problem for the roots of $z^2 = 1$ but their methods differ. Cayley's is geometrical whereas Schröder's was analytical. There are only two fixed points $z = \pm 1$ and the boundary (Julia set) between the two basins of attraction of the root finding function is the "y-axis." The algorithm for finding the complex roots of the cubic equation $z^3 = 1$ has three stable fixed points at $z = 1, z = \exp(2\pi i/3), z = \exp(-2\pi i/3)$ and with the iteration $z \to z - (z^3 - 1)/3z^2$, three domains, basins of attraction exist with highly interlaced boundaries. The problem of determining the Julia set for the quadratic was straightforward but the cubic seemed impossible [42]. For good reason, with the aid of modern computing machinery the Julia set in the case of the cubic is an intricately laced trefoil. But some progress was made before computers entered the field. After WWI G. Julia and P. Fatou published lengthy papers on iteration and later C. L. Siegel studied the field [35,48,85].

Topological Dimension

The concept of dimension has been an enduring study for millennia, and fractals has prolonged the centrality of this concept in mathematics. The ordinary meaning of dimensions one, two, three dimensions applied to line, plane, solid of the Greeks, was initially extended to n-dimensions. Since the 1870s mathematicians ascribed newer meanings to the term.

At the beginning of the twentieth century, Topology was in its infancy, and emerging from "analysis situs." An impetus was the Hausdorff's definition of a topological space in terms of neighborhoods in 1914. The evolution of topological dimension was developed in the hands

of such mathematicians as L. E. J. Brouwer, H. Poincaré, K. Menger, P. Urysohn, F. Hausdorff but in the new surroundings, the concepts of curve and dimension proved elusive.

Poincaré made several attempts to define dimension. One was in terms of group theory, one he described as a "dynamical theory." This was unsatisfactory for why should dimension depend on the idea of a group. He gave another definition that he described as a "statical theory." Accordingly in papers of 1903, and 1912, a topological notion of n-dimensions (where n a natural number) was based on the notion of a cut, and Poincaré wrote:

> If to *divide* a continuum it suffices to consider as cuts a certain number of elements all distinguishable from one another, we say this continuum is of *one dimension*; if, on the contrary, to divide a continuum it is necessary to consider as cuts a system of elements themselves forming one or several continua, we shall say that this continuum is of *several dimensions*. If to divide a continuum C, cuts which form one or several continua of one dimension suffice, we shall say that C is a continuum of *two dimensions*; if cuts which form one or several continua of at most two dimensions suffice, we shall say that C is a continuum of *three dimensions*; and so on [74].

To illustrate the pitfalls of this game of cat and mouse where "definition" attempts to capture the right notion, we see this definition yielded some curious results – the dimension of a double cone ostensibly of two dimensions turns out to be of one dimension, since one can delete the zero dimensional point where the two ends of the cone meet.

Curves were equally difficult to pin down. Menger used a physical metaphor to get at the pure notion of a curve:

> We can think of a curve as being represented by fine wires, surfaces as produced from thin metal sheets, bodies as if they were made of wood. Then we see that in order to separate a point in the surface from points in a neighborhood or from other surfaces, we have to cut the surfaces along continuous lines with a scissors. In order to extract a point in a body from its neighborhood we have to saw our way through whole surfaces. On the other hand in order to excise a point in a curve from its neighborhood irrespective of how twisted or tangled the curve may be, it suffices to pinch at discrete points with tweezers. This fact, that is independent of the particular form of curves or surfaces we consider, equips us with a strong conceptual description [22].

Menger set out his ideas about basic notions in mathematical papers and in the books *Dimensionstheorie* (1928) and *Kurventheorie* (1932). In *Dimensionstheorie* he gave an inductive definition of dimension, on the implicit understanding that dimension only made sense for n an integer (≥ -1):

> A space is called at most n-dimensional, if every point is contained in arbitrarily small neighborhoods with an most $(n-1)$-dimensional boundaries. A space that is not at most $(n-1)$-dimensional we call at least n-dimensional. … A Space is called n-dimensional, if it is both at most n-dimensional and also at least n-dimensional, in other words, if every point is contained in arbitrarily small neighborhoods with at most $(n-1)$-dimensional boundaries, but at least one point is not contained in arbitrarily small boundaries with less than $(n-1)$-dimensional boundaries. … The empty set and only this is (–1)-dimensional (and at most (–1)-dimensional. A space that for no natural number n is n-dimensional we call infinite dimensional.

Different notions of topological dimension defined in the 1920s and 1930s, were $\operatorname{ind} X$ (the small inductive dimension), $\operatorname{Ind} X$ (the large inductive dimension), and $\dim X$ (the Lebesgue covering dimension). Researchers investigated the various inequalities between these and the properties of abstract topological spaces which ensured all these notions coincided. By 1950, the theory of topological dimension was still in its infancy [20,47,66].

Metric Dimension

For the purposes of fractals, it is the *metric* definitions of dimension which are fundamental.

For instance, how can we define the dimension of the Cantor's "middle third" set which takes into account its metric structure? What about the iconic fractal known as the von Koch curve snow flake curve (1904)? These sets pose interesting questions. Pictorially von Koch's curve is made up of small 1 dimensional lines and we might be persuaded that it too should be 1 dimensional. But the real von Koch curve is defined as a limiting curve and for this there are differences between it and a line. Between any two points on a 1 dimensional line there is a finite distance but between any two points on the Koch curve the distance is infinite. This suggests its dimensionality is greater than 1 while at the same time it does not fill out a 2 dimensional region. The Sierpinski curve (or gasket) is another example, but perhaps more spectacular is the "sponge curve". In

Kurventheorie (1932) Menger gave what is now known as the Menger Sponge, the set obtained from the unit cube by successively deleting sub-blocks in the same way as Cantor's middle third set is obtained from the unit interval.

A colorful approach to defining metric dimension is provided by L. F. Richardson. Richardson (1951) took on the practical measurement of coastlines from maps. To measure the coastline, Richardson used dividers set a distance l apart. The length of the coastline will be $L = \sum l$ after walking the dividers around the coast. Richardson was a practical man, and he found by plotting L against l, the purely empirical result that $L \propto l^{-\alpha}$ for a constant α that depends on the chosen coastline. So, for a given country we get an approximate length $L = cl^{-\alpha}$ but the smaller the dividers are set apart the longer becomes the coastline!

The phenomenon of the coastline length being "divider dependent" explains the discrepancy of 227 kilometers in the measurement by of the border between Spain and Portugal (987 km stated by Spain, and 1214 km stated by Portugal). The length of the Australian coastline turns out to be 9400 km if one divider space represents 2000 km, and is 14,420 km if the divider space represent 100 km. Richardson cautions "to speak simply of the "length" of a coast is therefore to make an unwarranted assumption. When a man says that he "walked 10 miles along the coast," he usually means that he walked 10 miles [on a smooth curve] *near* the coast." Richardson goes into incredible numerical data and his work is of an empirical nature, but it paid off, and he was able to combine two branches of science, the empirical and the mathematical. Moreover here we have a link with chaos: Richardson described the "sensitivity to initial conditions" where a small difference in the setting of the dividers can result in a large difference in the "length" of the coastline.

The constant α is a characteristic of the coastline or frontier whose value is dependent on its place in the range between smoothness and jaggedness. So if the frontier is a straight line, α would be zero, and increases the more irregular the coast line. In the case of Australia, α was found to be about 0.13 and for the very irregular west coast of Britain, α was found to be about 0.25. The value $1 + \alpha$ anticipates the mathematical concept known as "divider dimension". So, for example, the divider dimension of the west coast of Britain would be 1.25. The divider dimension idea is fruitful – it can be applied to Brownian motion, a mathematical theory set out by Norbert Wiener in the 1920s but it is not the main one when applied to fractal dimension.

"Fractal dimension" means Hausdorff dimension, or as we already noted, the Hausdorff–Besicovitch dimension. Generally Hausdorff dimension is difficult to calculate, but for the self-similar sets and curves such as Cantor's middle-third set, the Sierpinski curve, the von Koch curve, Menger's sponge, it is equivalent to calculating the box dimension or Minkowski–Bouligand dimension (see also [61]). The middle third set has fractal dimension $D_{\mathrm{H}} = \log 2/\log 3 = 0.63\ldots$, and the Sierpinski curve has $D_{\mathrm{H}} = \log 3/\log 2 = 1.58\ldots$, the von Koch curve $D_{\mathrm{H}} = \log 4/\log 3 = 1.26\ldots$, and Menger's sponge embedded in three-dimensional Euclidean space has Hausdorff dimension $D_{\mathrm{H}} = \log 20/\log 3 = 2.72\ldots$.

Chaos Comes of Age

The modern theory of Chaos occurred around the end of the 1950s. S. Smale, Y. Ueda, E. N. Lorenz made discoveries which ushered in the new age. The subject became extremely popular in the early 1970s and the "avant-garde" looked back to this period as the beginning of chaos theory. They contributed some of the most often cited papers in further developments.

Physical Systems

In 1961 Y. Ueda, a third year undergraduate student in Japan discovered a curious phenomenon in connection with the single "Van der Pol" type nonlinear equation

$$\mathrm{d}^2x/\mathrm{d}t^2 + k(\gamma x^2 - 1)\mathrm{d}x/\mathrm{d}t + x^3 = \beta \cos \lambda t\,,$$

another example of an equation used to model an oscillator. With the parameter values set at $k = 0.2, \gamma = 8, \beta = 0.35$ Ueda found the dynamics was "chaotic" – though of course he did not use that term. With an analogue computer, of a type then used to solve differential equations, the attractor in phase space appeared as a "shattered egg." Solutions for many other values of the parameters revealed orderly behavior, but what was special about the values 0.2, 8, 0.35? Ueda had no idea he had stumbled on a major discovery. Forty years later he reminisced on the higher principles of the scientific quest: "but while I was toiling alone in my laboratory," he said, "I was never trying to pursue such a grandiose dream as making a revolutionary new discovery, not did I ever anticipate writing a memoir about it. I was simply trying to find an answer to a persistent question, faithfully trying to follow the lead of my own perception of a problem [2]."

In 1963, E. N. Lorenz published a landmark paper (with at least 5000 citations to date) but it was one published in a journal off the beaten track for mathematicians. Lorenz investigated a simple model for atmospheric convection along the lines of the Rayleigh–Bernard convection model. To see what Lorenz actually did we quote the abstract to his original paper:

Finite systems of deterministic ordinary nonlinear differential equations may be designed to represent forced dissipative hydrodynamical flow. Solutions of these equations can be identified with trajectories in phase space. For those solutions with bounded solutions, it is found that non periodic solutions are ordinarily unstable with respect to small modifications, so that slightly differing initial states can evolve into considerably different states. Systems with bounded solutions are shown to possess bounded numerical solutions. A simple system representing cellular convection is solved numerically. All of the solutions are found to be unstable, and almost all of them are non periodic [53].

The oft quoted equations (with simplifying assumptions, and a truncated version of the Navier–Stokes equations) used to model Rayleigh–Bernard convection are the nonlinear equations [89]:

$$\begin{aligned} \mathrm{d}x/\mathrm{d}t &= \sigma(y-x) \\ \mathrm{d}y/\mathrm{d}t &= rx - y - xz \\ \mathrm{d}z/\mathrm{d}t &= xy - bz\,, \end{aligned}$$

where σ is called the Prandtl number, r is the Rayleigh number and b is a geometrically determined parameter. From the qualitative viewpoint, the solution of these equations, $(x_1(t), x_2(t), x_3(t))$ with initial values $(x_1(0), x_2(0), x_3(0))$ traces out a trajectory in 3-dimensional phase space. Lorenz solved the equations numerically by a forward difference procedure based on the time honored Runge–Kutta method which set up an iterative scheme of difference equations. Lorenz discovered that the case where the parameters have specific values of $\sigma = 10, r = 28, b = 8/3$ gave chaotic trajectories which wind around the famous Lorenz attractor in the phase space – and by accident, he discovered the butterfly effect. The significance of Lorenz's work was the discovery of Chaos in low dimensional systems – the equations described dynamical behavior of a kind seen by only a few – including Ueda in Japan. One can imagine his surprise and delight, for though working in Meteorology, Lorenz had been a graduate student of G. D. Birkhoff at Harvard. Once the chink in the mathematical fabric was made, the same kind of mathematical behavior of chaos was discovered in other systems of differential equations.

The Lorenz attractor became the subject of intensive research. Once it was discovered by mathematicians – not many mathematicians read the meteorology journals – it excited much attention and helped to launch the chaos craze. But one outstanding problem remained: did the Lorenz attractor actually exist or was its presence due to the accumulation of numerical errors in the approximate methods used to solve the differential equations. Computing power was in its infancy and Lorenz made his calculations on a fairly primitive string and sealing-wax Royal McBee LGP-30 machine.

Some believed that the numerical evidence was sufficient for the existence of a Lorenz attractor but this did not satisfy everyone. This question of actual existence resisted all attempts at its solution because there were no mathematical tools to solve the equations explicitly. The problem was eventually cracked by W. Tucker in 1999 then a postgraduate student at the University of Uppsala [94]. Tucker's proof that the Lorenz attractor actually exists involved a rigorous computer algorithm in conjunction with a rigorous set of bounds on the possible numerical errors which could occur. It is very technical [95].

Other sets of differential equations which exemplified the chaos phenomena, one even more basic than Lorenz's is due to O. Rössler, equations which modeled chemical reactions:

$$\begin{aligned} \mathrm{d}x/\mathrm{d}t &= -y - z \\ \mathrm{d}y/\mathrm{d}t &= x + \alpha y \\ \mathrm{d}z/\mathrm{d}t &= \alpha - \mu z + xz\,. \end{aligned}$$

Rössler discovered chaos for the parameter values $\alpha = 0.2$ and $\mu = 5.7$ [77]. This is the simplest system yet found, for it has only one nonlinear term xz compared with two in the case of Lorenz's equations. Other examples of chaotic solutions are found in the dripping faucet experiment caried out by Shaw (1984) [84].

Strange Attractors

An attractor is a set of points in phase space with the property that a trajectory with an initial point near it will be attracted to it – and if the trajectory touches the attractor it is trapped to stay within it. But what makes an attractor "strange"?

The simple pendulum swings to and fro. This is a dissipative system as the loss of energy causes the bob to come to rest. Whatever the initial point stating point of the bob, the trajectory in phase space will spiral down to the origin. The attractor in this case is simply the point at the origin; the rest point where displacement is nil and velocity is nil. This point is said to attract all solutions of the differential equations which models the pendulum. This single point attractor is hardly strange.

If the pendulum is configured to swing to and fro with a fixed amplitude, the attractor in phase space will be a circle. In this case the system conserves energy and the whole

system is called a conservative system. If the pendulum is slightly perturbed it will return to the previous amplitude, and in this sense the circle or a limit cycle will have attracted the displaced trajectory. Neither is the circle strange.

The case of the *driven* pendulum is different. In the driven pendulum the anchor point of the pendulum oscillates with a constant amplitude a and constant drive frequency f. The equation of motion of the angular displacement from the vertical θ with damping parameter q can be described as a second order nonlinear (because of the presence of the $\sin\theta$ term in the differential equation):

$$\mathrm{d}^2\theta/\mathrm{d}t^2 + (1/q)\,\mathrm{d}\theta/\mathrm{d}t + \sin\theta = a\cos ft\,.$$

Alternatively this motion can be described by three simultaneous equations:

$$\begin{aligned}\mathrm{d}w/\mathrm{d}t &= -(1/q)w - \sin\theta + a\cos\phi\\ \mathrm{d}\theta/\mathrm{d}t &= w\\ \mathrm{d}\phi/\mathrm{d}t &= f\,,\end{aligned}$$

where ϕ is the phase of the drive term. The three variables (w, θ, ϕ) describe the motion of the driven pendulum in three–dimensional phase space and its shape will depend on the values of the parameters (a, f, q). For some values, just as for the Lorenz attractor, the motion will be chaotic [5].

For an attractor to be strange, it should be fractal in structure. The notion of strange attractor is due to D. Ruelle and F. Takens in papers of 1971 [82]. In their work there is no mention of fractals, simply because fractals had not risen to prominence at that time. Strange attractors for Ruelle and Takens were infinite sets of points in phase space corresponding to points of a physical dynamic system which appeared to have a complicated evolution – they were just very weird sets of points. But the principle was gained. Dynamical systems, like the gusts in wind turbulence, are in principle modeled by deterministic differential equations but now their solution trajectories seemed random. In classical physics mechanical processes were supposed to be as uncomplicated as the pendulum where the attractor was a single point or a circular limit cycle. The seemingly random processes that now appeared offered the mathematician and physicist a considerable challenge.

The name "strange attractor" caught on and quickly captured the scientific and popular imagination. Ruelle asked Takens if he had dreamed up the name, and he replied: "Did you ever ask God whether he created this damned universe? … I don't remember anything … I often create without remembering it …" and Ruelle wrote the "creation of strange attractors thus seems to be surrounded by clouds and thunder. Anyway, the name is beautiful, and I well suited to these astonishing objects, of which we understand so little [80]." Ruelle wrote "These systems of curves [arising in the study of turbulence], these clouds of points, sometimes evoke galaxies or fireworks, other times quite weird and disturbing blossomings. There is a whole world of forms still to be explored, and harmonies still to be discovered [45,91]."

Initially a strange attractor was a set which was extraordinary in some sense and there were naturally attempts to pin this down. In particular distinctions between "strange attractors" and "chaotic behavior" were drawn out [14]. Accordingly, a strange attractor is an attractor which is *not* (i) a finite set of points (ii) a closed curve (iii) a smooth or piecewise smooth surface or a volume bounded by such a surface. "Chaotic behavior" relates to the behavior of trajectories on the points of the attractor where nearby orbits around the attractor diverge with time. A strange nonchaotic attractor is one (with the above exclusions) where the trajectories are not chaotic, that is, there is an absence of sensitivity to initial conditions.

Pure Dynamical Systems

Chaos had its origins in real physical problems and the consequent differential equations, but the way had been set by Poincaré and Birkhoff for the entry of pure mathematicians. S. Smale gained his doctorate from the University of Michigan in 1956 supervised by the topologist R. Bott, stepped into this role. In seeking out fresh fields for research he surveyed Birkhoff's *Collected Papers* and read the papers of Levinson on the van der Pol equation..

On a visit to Rio de Janeiro in late 1959, Smale focused on general dynamical systems. What better place to do research than on the beach:

> My work was mostly scribbling down ideas and trying to see how arguments could be sequenced. Also I would sketch crude diagrams of geometric objects flowing through space and try to link the pictures with formal deductions. Deeply involved in this kind of thinking and writing on a pad of paper, the distraction of the beach did not bother me. Moreover, one could take time off from the research to swim [88].

He got into hotter water to the north when it was brought to the attention of government officials that national research grants were being spent at the seaside. They seemed unaware that a mathematical researcher is always working. Smale certainly had a capacity for hard concentrated work.

Receiving a letter from N. Levinson, about the Van der Pol equation, he reported:

> I worked day and night to try to resolve the challenge to my beliefs that letter posed. It was necessary to translate Levinson's analytic argument into my own geometric way of thinking. At least in my own case, understanding mathematics doesn't come from reading or even listening. It comes from rethinking what I see or hear. I must redo the mathematics in the context of my particular background. ... In any case I eventually convinced myself that Levinson was correct, and that my conjecture was wrong. Chaos was already implicit in the analyzes of Cartwright and Littlewood! The paradox was resolved, I had guessed wrongly. But while learning that, I discovered the horseshoe [8].

The famous "horseshoe" mapping allowed him to construct a proof of the Poincaré conjecture in dimensions greater than or equal to five, and for this, he was awarded a Field's medal in 1966.

Iteration provides an example of a dynamical system, of a kind involving discrete time intervals. These were the systems suggested by Schröder and Cayley, and take on a pure mathematical guise with no physical experimenting to act as a guide. In the late 1950s, a research group in Toulouse led by I. Gumowski and his student C. Mira pursued an exploration of nonlinear systems from this point of view. Their approach was inductive, and started with the simplest example of nonlinear maps of the plane which were then iterated. Functions chosen for exploration were the family,

$$x \to y - F(x)$$
$$y \to x + F(y - F(x)) ,$$

for various rational functions $F(x)$. The pictures in the plane were noted for their spectacular "chaos esthétique."

The connection of chaos with the iteration of functions was pioneered by Gumowski and Mira, O. M. Sharkovsky (in 1964), Smale (in 1967) [86], and N. Metropolis, M. Stein, P. Stein (in 1973) [67]. M. Feigenbaum (in the late 1970s) studied the quadratic logistic map $x \to \lambda x(1-x)$ and its iterates and discovered period doubling and the connection with physical phenomena. For the value $\lambda = 3.5699456\ldots$ the orbit is non periodic and the attractor is the Cantor set so it is a strange attractor.

Feigenbaum studied one dimensional iterative maps of a general type $x \to \lambda f(x)$ for a general f and discovered properties independent of the form of the recursion function. The Feigenbaum number, the bifurcation rate δ with value $\delta = 4.6692\ldots$ is the limit of the parameter intervals in which period doubling takes place before the onset of chaos. The intervals are shortened at each period doubling by the inverse of the Feigenbaum number, a universal value for many functions [36,37]. In the case of fluid flow, it was discovered that the transition from smooth flow to turbulence is linked with period doubling.

The iteration of functions, a very simple process, was brought to the attention of the wider scientific public by R. May. In his oft quoted piece on the one dimension quadratic mapping, he stressed the importance it held for mathematical education:

> I would therefore urge that people be introduced to, say, equation $x \to \lambda x(1-x)$ early in their mathematical education. This equation can be studied phenomenologically by iterating it on a calculator, or even by hand. Its study does not involve as much conceptual sophistication as does elementary calculus. Such study would greatly enrich the student's intuition about nonlinear systems.
>
> Not only in research, but also in the everyday world of politics and economics, we would all be better off if more people realized that simple nonlinear systems do not necessarily possess simple dynamical properties [62].

The example of $x \to \lambda x(1-x)$ provides another link with the mathematical past, which amply demonstrates that Fractals and Chaos is not merely a child of the sixties but is joined to the mainstream of mathematics which slowly evolves over centuries. The Schwarzian derivative $y = f(x)$ defined by

$$S(y) = \frac{\mathrm{d}x}{\mathrm{d}y}\frac{\mathrm{d}^3 y}{\mathrm{d}x^3} - \frac{3}{2}\left(\frac{\mathrm{d}x}{\mathrm{d}y}\frac{\mathrm{d}^2 y}{\mathrm{d}x^2}\right)^2 ,$$

is connected with complex analysis and also introduced into invariant theory (where it was called a differentiant) of a hundred years previously. It was introduced by D. Singer in 1978 in the context of one dimensional dynamical systems [26].

While the quadratic map is one dimensional, we can go into two dimensions with pairs of nonlinear difference equations. A beautiful example is the one $(x, y) \to (y - x^2, a + bx)$ or in the equivalent form $(x, y) \to (1+y-ax^2, bx)$. In this case, for some values of the parameters ellipses are generated, but in others we gain strange attractors. This is the Hénon scheme$b = 0.3$, and a near 1.4 the attractor has an infinity of attracting points, and being a fractal set it is a strange attractor [81]. A Poincaré

section of this attractor is a Cantor set, further evidence for the set being a strange attractor.

The Hénon map is one of the earliest examples illustrating chaos. It has Jacobian (which measures area) b for all points in the plane, so if $|b| < 1$ the Hénon mapping of the plane contracts area by a constant factor b for any point in the plane [41]. The effect of iterating this mapping is to stretch and fold its image in the manner of the Smale horseshoe mapping. The Hénon attractor is the prototype strange attractor. It arises in the case of a diffeomorphism of the plane which stretches and folds an open set U with the property that $f(\bar{U}) \subset U$.

The Advent of Fractals

The term was coined by B. B. Mandelbrot in the 1970s and now Mandelbrot is regarded as the "father of fractals" [55,56]. His first forays were to study fractals which were invariant under linear transformations. His uncle (Szolem Mandelbrojt, a professor of Mathematics and Mechanics at the Collège de France and later at the French Académie des Sciences) encouraged him to read the original papers by G. Julia and P. Fatou as a young man.

Euclid's *Elements*

Fractals are revolutionary because they challenge one of the sacred texts of mathematics. Euclid's *Elements* had reigned over mathematics for well over two thousand years and enjoys the distinction of still be referred to by working mathematicians. Up to the nineteenth century it was the essential fare of mathematics taught in schools. In Britain the *Elements* exerted the same authority as the Bible which was the other book committed to memory by generations of school pupils. Towards the end of the nineteenth century its central position in the curriculum was challenged, and it was much for the same reasons that Mandelbrot challenged it in the 1980s.

Praised for its logical structure, learning Euclid's deductive proofs by heart was simply not the way to teach an understanding of geometry. In France it had been displaced at the beginning of the century but in Britain and many other countries the attack on its centrality came at the end of the century. Ironically one of the main upholders of the sovereignty of Euclid and a staunch defender of the sacred book was Cayley. He was just the man who a few years before had shown himself to be more in the Mandelbrot mode, who threw aside "the restrictions as to reality" in his investigation of primitive symbolic dynamic systems. And there is a second irony, for Cayley was a man who loved nature and likened the beauty of the natural world to the terrain of mathematics itself – and in practical terms used his mountaineering exploits as a way of discovering new ideas in the subject.

A century later Mandelbrot's criticism came but from a different direction but containing enough phrases which would have gained a nod of agreement from those who mounted their opposition all those years before. Mandelbrot's opening paragraph launched the attack:

> Why is geometry often described as "cold" and "dry"? One reason lies in its inability to describe the shape of a cloud, a mountain, a coastline, or a tree. Clouds are nor spheres, mountains are not cones, coastlines are not circles, and bark is not smooth, nor does lightning travel in a straight line [57].

Mandelbrot was setting out an ambitious claim, and behind it was the way he used for conducting mathematical research. He advocated a different methodology from the Euclidean style updated by the Bourbakian mantra of "definition, theorem, proof". In fact he describes himself as an exile, driven from France by the Bourbaki school. Jean Dieudonné, a leading Bourbakian was at the opposite end of the spectrum in his mathematical style. Dieudonné's objective, as set out in a graduate text *Foundations of Modern Analysis* (1960) was to "train the student in the use of the most fundamental mathematical tool of our time – the axiomatic method with which he [*sic*] will have had very little contact, if any at all, during his undergraduate years". Echoing the attitude of the nineteenth century geometer Jacob Steiner, who eschewed diagrams of any kind as inimical to the property development of the subject, Dieudonné appealed only to axiomatic methods, and to make the point in his book, he deliberately avoided introducing any diagrams and appeal to "geometric intuition"[28].

What Mandelbrot advocated was the methodology of the physician's and lawyer's "casebook." He noted "this term has no counterpart in science, and I suggest we appropriate it". This was rather novel and presents a counterpoint to the image that the Bourbakian ideal of great theorems with nicely turned out proofs are what is required above all else. If mathematics is presented as a completed house built by a great mathematician, as the Bourkadians could suggest, questions are only to be found in the higher reaches of the subjects, that demand years of study to reach them.

In his autobiography, Mandelbrot claimed "in pure mathematics, my main contribution has not been to provide proofs, but to ask new questions – usually very hard ones – suggested by physics and pictures". His questions spring from elementary objects in mathematics, but ones

which start off in the world. He summarizes his long career in his autobiography:

> My whole career became one long, ardent pursuit of the concept of roughness. The roughness of clusters in the physics of disorder, of turbulent flows, of exotic noises, of chaotic dynamical systems, of the distribution of galaxies, of coastlines, of stock-prize charts, and of mathematical constructions [58].

The "monster" examples which previously existed as instances of known theorems (and therefore were of little interest in themselves from the novelty point of view) or put aside because they were not examples, were now brought into the limelight and put on to the dissecting table to be investigated.

What is a Fractal?

In his essay (1975) Mandelbrot said "I stopped short of giving a mathematical definition, because I felt this notion – like a good wine – demanded a bit of ageing before being 'bottled'." [71]. He knew of Hausdorff dimension and developed an intuitive understanding for it, but postponed a definition. A little later he adopted a working definition, and in his *Fractal Geometry of Nature* (1982) he gave his manifesto for fractals.

What is Mandelbrot's subsequent working definition of a fractal? It is simply a point set for which the Hausdorff dimension exceeds its topological dimension, $D_H > D_T$. Examples are Cantor middle third set. where $D_H = \log 2/\log 3 = 0.63\ldots$ with $D_T = 0$, and the von Koch Curve where $D_H = \log 4/\log 3 = 1.26\ldots$ with $D_T = 1$.

Mandelbrot's definition gives rise to a programme of research similar in nature to one carried out by Menger and Urysohn in the 1920s. Just as they asked "what is a curve?" and "what is dimension?" it could now be asked "what is a fractal exactly?" and "what is fractal dimension?" The whole game of topological dimension was to be played over for metric dimension. Mandelbrot's definition of a fractal fired a first shot in answering one question but, just as there were exceptions to C. Jordan's definitions of curve like Peano's famous space filling curve of the 1890s, exceptions were found to Mandelbrot's preliminary definition. It was a regret, for example, that the example of the "devil's staircase" (characterized in terms of a continuous weight function defined on the Cantor middle third set) which looked like a fractal did not conform to the working definition of a fractal since in this case $D_H = D_T = 1$.

There are, however, many examples of obvious fractals, in nature and in mathematics. The pathological curves of the late nineteenth century which suffered this fate of not conforming to definitions of "normal curves" were brought back to life. The pathologies, such as those invented by Brouwer and Cantor had been put in a cupboard and locked away. An instance of this is Brouwer's decomposition of the plane into three "countries" so that the boundary points touch each other (not just meeting at one point). Brouwer's decomposition is not unlike the shape of a fractal. Cantor's examples were equally pathological.

Mandelbrot made his own discoveries, but when he opened the cupboard of "monster curves" he liked what he saw. Otto Rössler, with justification called Mandelbrot "Cantor's time-displaced younger friend". As Freeman Dyson wrote: "Now, as Mandelbrot points out, ..., Nature has played a joke on the mathematicians. The nineteenth century mathematicians may have been lacking in imagination [in limiting themselves to Euclid and Newton], but nature has not. The same pathological structures [the gallery of monster curves] that the mathematicians invented to break loose from nineteenth century naturalism turn out to be inherent in familiar objects all around us [38]."

The Mandelbrot Set

Just how did Mandelbrot arrive at the famous Mandelbrot set? With the freedom of being funded by an IBM fellowship, he went further into the study of $p_c(z) = z^2 + c$ and experimented. The basement of the Thomas J. Watson Research Center in the early 1980s held a brand new VAX main frame computer and a Tektronix cathode ray tube for viewing results.

The quadratic map $p_c(z) = z^2 + c$ is the simplest of all the nonlinear rational transformations. The key element is to consider z, c as complex numbers. The Mandelbrot set M in the plane is defined as

$$M = \{c \in C\colon \text{ the Julia set } J_c \text{ is connected}\}$$

or, alternatively, in terms of the sequence $\{p_c^k(0)\}$

$$M = \{c \in C\colon \{p_c^k(0)\} \text{ is bounded}\}\,.$$

It was an exciting event when Mandelbrot first glimpsed the set, which was initially called an M-set. First observed on the Tektronix cathode ray tube, was the messy outline a fault in the primitive equipment? He and his colleagues tried again with a more powerful computer but found "the mess had failed to vanish" and even that the image showed signs of being more systematic. Mandebrot reported "we promptly took a much closer look. Many specks of dirt duly vanished after we zoomed in. But some

specks failed to vanish; in fact, they proved to resolve into complex structures endowed with 'sprouts' very similar to those of the whole set M. Peter Moldave and I could not contain our excitement." Yet concrete mathematical properties were hard to come by. One immediate result did follow quickly, when the M-set, now termed the Mandelbrot set was proved to be a connected set [29].

The Merger

When chaos theory was being discussed in the 1970s, strange attractors were regarded as weird subsets of Euclidean space. To be sure their importance was recognized but their nature was only vaguely understood. Towards the end of the 1970s, prompted by the advent of fractals, attractors could be viewed in a new light. But Chaos which arose in physical situations did not give rise to ordinary self-similar fractals, like the Koch curve, but to more complicated sets. The outcome of Chaos is apparent randomness while the key properties of ordinary fractals is regularity.

Measuring Strange Attractors

Once strange attractors were brought to light, the next stage was to measure them. This perspective was put by practitioners in the 1980s:

> Nonlinear physics presents us with a perplexing variety of complicated fractal objects and strange sets. Notable examples include strange attractors for chaotic dynamical systems. … Naturally one wishes to characterize the objects and described the events occurring on them. For example, in dynamical systems theory one is often interested in a strange attractor … [43].

At the same time other researchers put a similar point of view, and an explanation for making the calculations. The way to characterize strange attractors was by the natural measure of a fractal – its fractal dimension – and this became the way to proceed:

> The determination of the fractal dimension d of strange attractors has become a standard diagnostic tool in the analysis of dynamical systems. The dimension roughly speaking measures the number of degrees of freedom that are relevant to the dynamics. Most work on dimension has concerned maps, such as the Hénon map, or systems given by a few coupled ordinary differential equations, such as the Lorenz and Rössler models. For any chaotic system described by differential equations, d must be greater than 2, but d could be much larger for a system described by partial differential equations, such as the Navier–Stokes equations [12].

As an example, the fractal dimension of the original Lorenz attractor derived from the meteorology differential equations was found to be approximately 2.05 [34].

Newer Concepts of Dimension

Just as for metric free topological dimension itself, there are a myriad of different concepts of metrically based dimension. What was wanted were measures which could be used in practice. What "dimensions" would shine light on the onset of chaotic solutions to a set of differential equations ?

One was the Lyapunov dimension drawing on the work of the Russian mathematician Alexander Lyapunov, and investigated by J. L. Kaplan and J. A. Yorke (1979). The Lyapunov dimension of an attractor A embedded in an Euclidean space of dimension n is defined by:

$$D_{\mathrm{L}} = k + \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{|\lambda_{k+1}|} ,$$

where $\lambda_1, \lambda_2, \ldots, \lambda_k$ are Lyapunov exponents and k is the maximum integer for which $\sum_{i=1}^{k} \lambda_i \geq 0$. In the case where $k = 1$ which occurs for two-dimensional chaotic maps, for instance,

$$D_{\mathrm{L}} = 1 - \frac{\lambda_1}{\lambda_2}$$

and because in this case $\lambda_2 < 0$ the value of the Lyapunov dimension will be sandwiched between 1 and the topological dimension which is 2. In the case of the Hénon mapping of the plane, the Lyapunov dimension of the attractor is approximately 1.25.

In a different direction, a generalized dimension D_q was introduced by A. Rényi in the 1950s [78]. Working in probability and information theory, he defined a spectrum of dimensions:

$$D_q = \lim_{r \to 0} \frac{1}{q-1} \frac{\log \sum_i P_i^q}{\log r}$$

depending on the value of q.

The special cases are worth noting:

- $D_0 = D_{\mathrm{B}}$ the box dimension
- $D_1 =$ "information dimension" related to C. Shannon's entropy [39,40]
- $D_2 = D_{\mathrm{C}}$ the correlation dimension.

The correlation dimension is a very practical measure and can be applied to experimental data which has been presented visually, as well as other shapes, such as photographs where the calculation involves counting points. It has been used for applications in hydrodynamics, business, lasers, astronomy, signal analysis. It can be calculated directly for time series where the "phase space" points are derived from lags or time-delays. In two dimensions this would be a sequence of points $(x(n), x(n+l))$ where l is the chosen lag. A successful computation depends on the number of points chosen. For example, it is possible to estimate the fractal dimension of the Hénon attractor to within 6% of its supposed value with 500 data points [70].

There is a sequence of inequalities between the various dimensions of a set A from the embedding space of topological dimension $D_E = n$ to the topological dimension D_T of A [90]:

$$D_\mathrm{T} \leq \ldots \leq D_2 \leq D_1 \leq D_\mathrm{H} \leq D_0 \leq n\,.$$

Historically these questions of inequalities between the various types of dimension are of the same type as had been asked about topological dimension fifty years before. Just as it was shown that the "coincidence theorem"

$$\operatorname{ind} X = \operatorname{Ind} X = \dim X$$

holds for well behaved spaces, such as compact metric spaces, so it is true that

$$D_\mathrm{L} = D_\mathrm{B} = D_1$$

for ordinary point sets such as single points and limit cycles. For the Lorenz attractor D_L agrees with the value of D_B and that of D_2 [69]. But each of the equalities fails for some fractal sets; the study of fractal dimension of attractors is covered in [34,83].

Multifractals

An ordinary fractal is one where the generalized Rényi dimension D_q is independent of q and returns a single value for the whole fractal. This occurs for the standard fractals such as the von Koch curve. A multifractal is a set in which D_q is dependent on q and a continuous spectrum of values results.

While an ordinary self-similar fractal are useful for expository purposes (and are beautifully symmetric shapes) the attractors found in physics are not uniformly self-similar. The resulting object is called a multifractal because it is *multi*-dimensional. The values of D_q is called its spectrum of the multifractal. Practical examples usually require their strange attractors to be modeled by multifractals.

Future Directions

The range of written material on Fractals and Chaos has exploded since the 1970s. A vast number of expository articles have been lodged in such journals as *Nature* and *New Scientist* and technical articles in *Physica D* and *Nonlinearity* [9].

Complexity Theory and Chaos Theory are relatively new sciences that can revolutionize the way we see our world. Stephen Hawking has said, "Complexity will be the science of the 21st century." There is even a relationship between fractals with the yet unproved Riemann hypothesis [51].

Many problems remain. Here it is sufficient to mention one representative of the genre. Fluid dynamicists are broadly in agreement that fluids are accurately modeled by the nonlinear Navier–Stokes equations. These are based on Newtonian principles and are deterministic, but theoretically the solution of these equations is largely unknown territory. A proof of the global regularity of the solutions represents a formidable challenge. The current view is that there is a dichotomy between laminar flow (like flow in the upper atmosphere) being smooth while turbulent flow (like flow near the earth's surface) is violent. Yet even this "turbulent" flow could be regular but of a complicated kind. A substantial advance in the theory will be rewarded with one of the Clay Institute's million dollar prizes. One expert is not optimistic the prize will be claimed in the near future [52].

The implication of Chaos dependent as it is on the sensitivity of initial conditions, suggests that forecasting some physical processes is theoretically impossible. Long range weather forecasting falls into this mould since it is predicated on knowing the weather conditions *exactly* at some point in time. There will inevitably be inaccuracies so exactness appears to be an impossibility.

No doubt "Chaos and (multi)Fractals" is here to stay. Rössler wrote of Chaos as the key to understanding Nature: "hairs and noodles and spaghettis and dough and taffy form an irresistible, disentanglable mess. The world of causality is thereby caricatured and, paradoxically, faithfully represented [2]". Meanwhile, the challenge for scientists and mathematicians remains [16].

Bibliography

Primary Literature

1. Abraham RH (1985) In pursuit of Birkhoff's chaotic attractor. In: Pnevmatikos SN (ed) Singularities and Dynamical Systems. North Holland, Amsterdam, pp 303–312
2. Abraham RH, Ueda Y (eds) (2000) The Chaos Avant-Garde:

memories of the early days of chaos theory. World Scientific, River Edge
3. Alexander DS (1994) A History of Complex Dynamics; from Schröder to Fatou and Julia. Vieweg, Braunschweig
4. Aubin D, Dahan Dalmedico A (2002) Writing the History of Dynamical Systems and Chaos: *Longue Durée* and revolution, Disciplines and Cultures. Historia Mathematica 29(3):235–362
5. Baker GL, Gollub JP (1996) Chaotic Dynamics. Cambridge University Press, Cambridge
6. Barrow-Green J (1997) Poincaré and the three body problem. American Mathematical Society, Providence; London Mathematical Society, London
7. Barrow-Green J (2005) Henri Poincaré, Memoir on the Three-Body Problem (1890). In: Grattan-Guinness I (ed) Landmark Writings in Western Mathematics 1640–1940. Elsevier, Amsterdam, pp 627–638
8. Batterson S (2000) Stephen Smale: the mathematician who broke the dimension barrier. American Mathematical Society, Providence
9. Berry MV, Percival IC, Weiss NO (1987) Dynamical Chaos. Proc Royal Soc London 413(1844):1–199
10. Birkhoff GD (1927) Dynamical Systems. American Mathematical Society, New York
11. Birkhoff GD (1941) Some unsolved problems of theoretical dynamics. Science 94:598–600
12. Brandstater A, Swinney HL (1986) Strange attractors in weakly turbulent Couette-Tayler flow. In: Ott E et al (eds) Coping with Chaos. Wiley Interscience, New York, pp 142–155
13. Briggs J (1992) Fractals, the patterns of chaos: discovering a new aesthetic of art, science, and nature. Thames and Hudson, London
14. Brindley J, Kapitaniak T, El Naschie MS (1991) Analytical conditions for strange chaotic and nonchaotic attractors of the quasiperiodically forced van der Pol equation. Physica D 51:28–38
15. Brolin H (1965) Invariant Sets Under Iteration of Rational Functions. Ark Mat 6:103–144
16. Campbell DK, Ecke R, Hyman JM (eds) (1992) Nonlinear science: the next decade. MIT Press, Cambridge
17. Cayley A (1879) The Newton–Fourier Imaginary Problem. Am J Math 2:97
18. Chabert J-L (1990) Un demi-siècle de fractales: 1870–1920. Historia Mathematica 17:339–365
19. Crilly AJ, Earnshaw RA, Jones H (eds) (1993) Applications of fractals and chaos: the shape of things. Springer, Berlin
20. Crilly T (1999) The emergence of topological dimension theory. In: James IM (ed) History of Topology. Elsevier, Amsterdam, pp 1–24
21. Crilly T (2006) Arthur Cayley: Mathematician Laureate of the Victorian Age. Johns Hopkins University Press, Baltimore
22. Crilly T, Moran A (2002) Commentary on Menger's Work on Curve Theory and Topology. In: Schweizer, B. et al Karl Menger, Selecta. Springer, Vienna
23. Curry J, Garnett L, Sullivan D (1983) On the iteration of rational functions: Computer experiments with Newton's method. Commun Math Phys 91:267–277
24. Dahan-Dalmedico A, Gouzevitch I (2004) Early developments in nonlinear science in Soviet Russia: The Andronov School at Gor'kiy. Sci Context 17:235–265
25. Devaney RL (1984) Julia sets and bifurcation diagrams for exponential maps. Bull Am Math Soc 11:167–171
26. Devaney RL (2003) An Introduction to Chaotic Dynamical Systems, 2nd edn. Westview Press, Boulder
27. Diacu F, Holmes P (1996) Celestial Encounters: the Origins of Chaos and Stability. Princeton University Press, Princeton
28. Dieudonné J (1960) Foundations of Modern Analysis. Academic Press, New York
29. Douady A, Hubbard J (1982) Iteration des polynomes quadratiques complexes. Comptes Rendus Acad Sci Paris Sér 1 Math 294:123–126
30. Dyson F (1997) 'Nature's Numbers' by Ian Stewart. Math Intell 19(2):65–67
31. Edgar G (1993) Classics on Fractals. Addison-Wesley, Reading
32. Falconer K (1990) Fractal Geometry: Mathematical Foundations and Applications. Wiley, Chichester
33. Falconer K (1997) Techniques in Fractal Geometry. Wiley, Chichester
34. Farmer JD, Ott E, Yorke JA (1983) The Dimension of Chaotic Attractors. Physica D 7:153–180
35. Fatou P (1919/20) Sur les equations fonctionelles. Bull Soc Math France 47:161–271; 48:33–94, 208–314
36. Feigenbaum MJ (1978) Quantitative Universality for a Class of Nonlinear Transformations. J Stat Phys 19:25–52
37. Feigenbaum MJ (1979) The Universal Metric Properties of Nonlinear Transformations. J Stat Phys 21:669–706
38. Gleick J (1988) Chaos. Sphere Books, London
39. Grassberger P (1983) Generalised dimensions of strange attractors. Phys Lett A 97:227–230
40. Grassberger P, Procaccia I (1983) Measuring the strangeness of strange attractors. Physica D 9:189–208
41. Hénon M (1976) A Two dimensional Mapping with a Strange Attractor. Commun Math Phys 50:69–77
42. Haeseler F, Peitgen H-O, Saupe D (1984) Cayley's Problem and Julia Sets. Math Intell 6:11–20
43. Halsey TC, Jensen MH, Kadanoff LP, Procaccia Shraiman I (1986) Fractal measures and their singularities: the characterization of strange sets. Phys Rev A 33:1141–1151
44. Harman PM (1998) The Natural Philosophy of James Clerk Maxwell. Cambridge University Press, Cambridge
45. Hofstadter DR (1985) Metamathematical Themas: questing for the essence of mind and pattern. Penguin, London
46. Holmes P (2005) Ninety plus years of nonlinear dynamics: More is different and less is more. Int J Bifurc Chaos Appl Sci Eng 15(9):2703–2716
47. Hurewicz W, Wallman H (1941) Dimension Theory. Princeton University Press, Princeton
48. Julia G (1918) Sur l'iteration des functions rationnelles. J Math Pures Appl 8:47–245
49. Kazim Z (2002) The Hausdorff and box dimension of fractals with disjoint projections in two dimensions. Glasgow Math J 44:117–123
50. Lapidus ML, van Frankenhuijsen M (2004) Fractal geometry and Applications: A Jubilee of Benoît Mandelbrot, Parts, 1, 2. American Mathematical Society, Providence
51. Lapidus ML, van Frankenhuysen (2000) Fractal geometry and Number Theory: Complex dimensions of fractal strings and zeros of zeta functions. Birkhäuser, Boston
52. Li YC (2007) On the True Nature of Turbulence. Math Intell 29(1):45–48
53. Lorenz EN (1963) Deterministic Nonperiodic Flow. J Atmospheric Sci 20:130–141

54. Lorenz EN (1993) The Essence of Chaos. University of Washington Press, Seattle
55. Mandelbrot BB (1975) Les objets fractals, forme, hazard et dimension. Flammarion, Paris
56. Mandelbrot BB (1980) Fractal aspects of the iteration of $z \to \lambda(1-z)$ for complex λ, z. Ann New York Acad Sci 357:249–259
57. Mandelbrot BB (1982) The Fractal Geometry of Nature. Freeman, San Francisco
58. Mandelbrot BB (2002) A maverick's apprenticeship. Imperial College Press, London
59. Mandelbrot BB, Hudson RL (2004) The (Mis)Behavior of Markets: A Fractal View of Risk, Ruin, and Reward. Basic Books, New York
60. Mattila P (1995) Geometry of Sets and Measures in Euclidean Spaces. Cambridge University Press, Cambridge
61. Mauldin RD, Williams SC (1986) On the Hausdorff dimension of some graphs. Trans Am Math Soc 298:793–803
62. May RM (1976) Simple Mathematical models with very complicated dynamics. Nature 261:459–467
63. May RM (1987) Chaos and the dynamics of biological populations. Proc Royal Soc Ser A 413(1844):27–44
64. May RM (2001) Stability and Complexity in Model Ecosystems, 2nd edn, with new introduction. Princeton University Press, Princeton
65. McMurran S, Tattersall J (1999) Mary Cartwright 1900–1998. Notices Am Math Soc 46(2):214–220
66. Menger K (1943) What is dimension? Am Math Mon 50:2–7
67. Metropolis, Stein ML, Stein P (1973) On finite limit sets for transformations on the unit interval. J Comb Theory 15:25–44
68. Morse M (1946) George David Birkhoff and his mathematical work. Bull Am Math Soc 52(5, Part 1):357–391
69. Nese JM, Dutton JA, Wells R (1987) Calculated Attractor Dimensions for low-Order Spectral Models. J Atmospheric Sci 44(15):1950–1972
70. Ott E, Sauer T, Yorke JA (1994) Coping with Chaos. Wiley Interscience, New York
71. Peitgen H-O, Richter PH (1986) The Beauty of Fractals. Springer, Berlin
72. Peitgen H-O, Jürgens H, Saupe D (1992) Chaos and fractals. Springer, New York
73. Pesin YB (1997) Dimension Theory in Dynamical Systems: contemporary views and applications. University of Chicago Press, Chicago
74. Poincaré H (1903) L'Espace et ses trois dimensions. Rev Métaphys Morale 11:281–301
75. Poincaré H, Halsted GB (tr) (1946) Science and Method. In: Cattell JM (ed) Foundations of Science. The Science Press, Lancaster
76. Przytycki F, Urbański M (1989) On Hausdorff dimension of some fractal sets. Studia Mathematica 93:155–186
77. Rössler OE (1976) An Equation for Continuous Chaos. Phys Lett A 57:397–398
78. Rényi A (1970) Probability Theory. North Holland, Amsterdam
79. Richardson LF (1993) Collected Papers, 2 vols. Cambridge University Press, Cambridge
80. Ruelle D (1980) Strange Attractors. Math Intell 2:126–137
81. Ruelle D (2006) What is a Strange Attractor? Notices Am Math Soc 53(7):764–765
82. Ruelle D, Takens F (1971) On the nature of turbulence. Commun Math Phys 20:167–192; 23:343–344
83. Russell DA, Hanson JD, Ott E (1980) Dimension of strange attractors. Phys Rev Lett 45:1175–1178
84. Shaw R (1984) The Dripping Faucet as a Model Chaotic System. Aerial Press, Santa Cruz
85. Siegel CL (1942) Iteration of Analytic Functions. Ann Math 43:607–612
86. Smale S (1967) Differentiable Dynamical Systems. Bull Am Math Soc 73:747–817
87. Smale S (1980) The Mathematics of Time: essays on dynamical systems, economic processes, and related topics. Springer, New York
88. Smale S (1998) Chaos: Finding a horseshoe on the Beaches of Rio. Math Intell 20:39–44
89. Sparrow C (1982) The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors. Springer, New York
90. Sprott JC (2003) Chaos and Time-Series Analysis. Oxford University Press, Oxford
91. Takens F (1980) Detecting Strange Attractors in Turbulence. In: Rand DA, Young L-S (eds) Dynamical Systems and Turbulence. Springer Lecture Notes in Mathematics, vol 898. Springer, New York, pp 366–381
92. Theiler J (1990) Estimating fractal dimensions. J Opt Soc Am A 7(6):1055–1073
93. Tricot C (1982) Two definitions of fractional dimension. Math Proc Cambridge Philos Soc 91:57–74
94. Tucker W (1999) The Lorenz Attractor exists. Comptes Rendus Acad Sci Paris Sér 1, Mathematique 328:1197–1202
95. Viana M (2000) What's New on Lorenz Strange Attractors. Math Intell 22(3):6–19
96. Winfree AT (1983) Sudden Cardiac death- a problem for topology. Sci Am 248:118–131
97. Zhang S-Y (compiler) (1991) Bibliography on Chaos. World Scientific, Singapore

Books and Reviews

Abraham RH, Shaw CD (1992) Dynamics, The Geometry of Behavior. Addison-Wesley, Redwood City

Barnsley MF, Devaney R, Mandelbrot BB, Peitgen H-O, Saupe D, Voss R (1988) The Science of Fractal Images. Springer, Berlin

Barnsley MF, Rising H (1993) Fractals Everywhere. Academic Press, Boston

Çambel AB (1993) Applied Complexity Theory: A Paradigm for Complexity. Academic Press, San Diego

Crilly AJ, Earnshaw RA, Jones H (eds) (1991) Fractals and Chaos. Springer, Berlin

Cvitanovic P (1989) Universality in Chaos, 2nd edn. Adam Hilger, Bristol

Elliott EW, Kiel LD (eds) (1997) Chaos Theory in the Social Sciences: foundations and applications. University of Michigan Press, Ann Arbor

Gilmore R, Lefranc M (2002) The Topology of Chaos: Alice in Stretch and Squeezeland. Wiley Interscience, New York

Glass L, MacKey MM (1988) From Clocks to Chaos. Princeton University Press, Princeton

Holden A (ed) (1986) Chaos. Manchester University Press, Manchester

Kellert SH (1993) In the Wake of Chaos. University of Chicago Press, Chicago

Lauwerier H (1991) Fractals: Endlessly Repeated Geometrical Figures. Princeton University Press, Princeton

Mullin T (ed) (1994) The Nature of Chaos. Oxford University Press, Oxford
Ott E (2002) Chaos in Dynamical Systems. Cambridge University Press, Cambridge
Parker B (1996) Chaos in the Cosmos: The Stunning Complexity of the Universe. Plenum, New York, London
Peitgen H-O, Jürgens, Saupe D, Zahlten C (1990) Fractals: An animated discussion with Edward Lorenz and Benoit Mandelbrot. A VHS film in color (63 mins). Freeman, New York
Prigogine I, Stengers I (1985) Order out of Chaos: man's new dialogue with Nature. Fontana, London
Ruelle D (1993) Chance and Chaos. Penguin, London
Schroeder M (1991) Fractals, Chaos, Power Laws. Freeman, New York
Smith P (1998) Explaining Chaos. Cambridge University Press, Cambridge
Thompson JMT, Stewart HB (1988) Nonlinear Dynamics and Chaos. Wiley, New York

Fractals and Multifractals, Introduction to

Daniel ben-Avraham[1], Shlomo Havlin[2]
[1] Clarkson University, Potsdam, USA
[2] Bar-Ilan University, Ramat Gan, Israel

Fractals generalize Euclidean geometrical objects to non-integer dimensions and allow us, for the first time, to delve into the study of complex systems, disorder, and chaos. In the words of B. Mandelbrot: "Clouds are not spheres, mountains are not cones, coastlines are not circles, bark is not smooth, nor does lightning travel in a straight line," [1]. Indeed, much has changed in our perception of nature, and today it is hard to conceive of natural phenomena that are *not* fractal, in the same way that it is hard to conceive of everyday life dynamical systems that are not non-linear. The discovery of fractals over three decades ago signaled a profound shift in the way we understand and analyze the physical world around us.

Not only does fractal geometry model complex disordered objects such as clouds and mountains, coastlines and lightning, but it also finds a beautiful new symmetry in the midst of all this complexity – invariance under dilation of space – and it is this self-similarity symmetry that lends fractals their tremendous power and analytical appeal (► Fractals and Multifractals, Introduction to). The same symmetry plays a significant role in critical phase transitions, and one of the earliest applications of fractals has been to the study of the paradigmatic model of percolation (► Fractals and Percolation). Soon after their advent, researchers systematically used fractals to explain a host of anomalous phenomena (anomalous with respect to the expectations from regular Euclidean objects) in condensed matter and solid state physics (► Fractal Structures in Condensed Matter Physics).

One of the most influential models of nonequilibrium growth, Diffusion Limited Aggregation (DLA), produces a fractal object that is still studied today for its fascinating connections to the Laplace equation and electrical discharge, conformal mapping, patterns of bacterial growth, and viscous fingering (► Fractal Growth Processes). Fractals and fractal scaling also arise in regular systems with superposed disorder, for example when resistances from a random distribution are assigned to the bonds of a regular lattice (► Fractal and Multifractal Scaling of Electrical Conduction in Random Resistor Networks). More recently, interest has grown in the study of large, stochastic complex nets, which seem infinite-dimensional at the outset, but even so can be found to have fractal geometry (► Fractal and Transfractal Scale-Free Networks). Having realized that the geometry of everyday life objects around us is more likely fractal than not, it was only natural to ask about the physics of phase transitions in disordered media, for example, for the magnetization of spins in the Ising model, placed on the nodes of fractals (► Phase Transitions on Fractals and Networks).

Early on, interest developed in the physics of transport in fractal substrates. It was quickly discovered that fractal media diffusion, the most elementary mode of transport, does not obey Fick's law but is, rather, anomalous (► Anomalous Diffusion on Fractal Networks), leading to entirely new ways of viewing and analyzing transport, most notably among them Lévy flights, continuum random walks, and fractional diffusion equations (► Levy Statistics and Anomalous Transport: Levy Flights and Subdiffusion). Because of the anomalous diffusion of reactants in fractal disordered media, so are the kinetics of reactions among them anomalous (► Reaction Kinetics in Fractals). The graph Laplacian in fractals has an anomalous spectrum that displays a peculiar scaling, having a profound effect on dynamics in general (► Dynamics on Fractals).

Fractals refer not only to geometrical objects but, more broadly, to any kind of phenomena possessing scaling that exhibits dilation symmetry, or scale invariance, often characterized by the appearance of a power-law distribution. A whole suite of techniques has evolved to analyze this type of fractal scaling (► Fractal and Multifractal Time Series). Fractal phenomena of this type (and occasionally fractal objects) find applications in several diverse fields of interest, such as finances and economics (► Fractals and Economics), geology (► Fractals in Geology and Geophysics), the analysis of DNA sequences (► Fractals and Wavelets: What Can We Learn on Transcription and

Replication from Wavelet-Based Multifractal Analysis of DNA Sequences?), biology (► Fractals in Biology), and even in path integrals in the quantum theory of spacetime (► Fractals in the Quantum Theory of Spacetime).

Fractals and the modern study of nonlinear systems, strange attractors, and chaos have infused and enriched one another from their very incipience (► Fractals Meet Chaos). Together they provide us with a fundamentally new way of understanding the everyday world around us.

Despite our best intentions, this Section on Fractals and Multifractals remains incomplete, as it fails to include numerous important subjects such as the practical use of fractals for image compression, their ubiquity in medicine, astronomy, and their prominence in the study of real systems of porous materials, electrodeposition, molecular surfaces, colloids, and polymers, to name just a few. We hope that some, if not all of these subjects will be included in future editions.

Bibliography

1. Mandelbrot BB (1982) The Fractal Geometry of Nature. Freeman, San Francisco

Fractals and Percolation

Yakov M. Strelniker[1], Shlomo Havlin[1], Armin Bunde[2]
[1] Department of Physics, Bar–Ilan University, Ramat–Gan, Israel
[2] Institut für Theoretische Physik III, Justus-Liebig-Universität, Giessen, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Percolation
Percolation Clusters as Fractals
Anomalous Transport on Percolation Clusters: Diffusion and Conductivity
Networks
Summary and Future Directions
Bibliography

Glossary

Percolation In the traditional meaning, percolation concerns the movement and filtering of fluids through porous materials. In this chapter, percolation is the subject of physical and mathematical models of porous media that describe the formation of a long-range connectivity in random systems and phase transitions. The most common percolation model is a lattice, where each site is occupied randomly with a probability p or empty with probability $1 - p$. At low p values, there is no connectivity between the edges of the lattice. Above some concentration p_c, the *percolation threshold*, connectivity appears between the edges. Percolation represents a geometric critical phenomena where p is the analogue of temperature in thermal phase transitions.

Fractal A fractal is a structure which can be subdivided into parts, where the shape of each part is similar to that of the original structure. This property of fractals is called self-similarity, and it was first recognized by G.C. Lichtenberg more than 200 years ago. Random fractals represent models for a large variety of structures in nature, among them porous media, colloids, aggregates, flashes, etc. The concepts of self-similarity and fractal dimensions are used to characterize percolation clusters. Self-similarity is strongly related to renormalization properties used in critical phenomena, in general, and in percolation phase transition properties.

Definition of the Subject

Percolation theory is useful for characterizing many disordered systems. Percolation is a pure random process of choosing sites to be randomly occupied or empty with certain probabilities. However, the topology obtained in such processes has a rich structure related to fractals. The structural properties of percolation clusters have become clearer thanks to the development of fractal geometry since the 1980s.

Introduction

Percolation represents the simplest model of a phase transition [1,8,13,14,26,27,30,48,49,61,64,65,68]. Assume a regular lattice (grid) where each site (or bond) is occupied with probability p or empty with probability $1 - p$. At a critical threshold, p_c, a long-range connectivity first appears: p_c is called the percolation threshold (see Fig. 1). Occupied and empty sites (or bonds) may stand for very different physical properties. For example, occupied sites may represent electrical conductors, empty sites may represent insulators, and electrical current may flow only through nearest-neighbor conducting sites. Below p_c, the grid represents an isolator since there is no conducting path between two adjacent bars of the lattice, while above p_c, conducting paths start to occur and the grid becomes a conductor. One can also consider percolation as a model

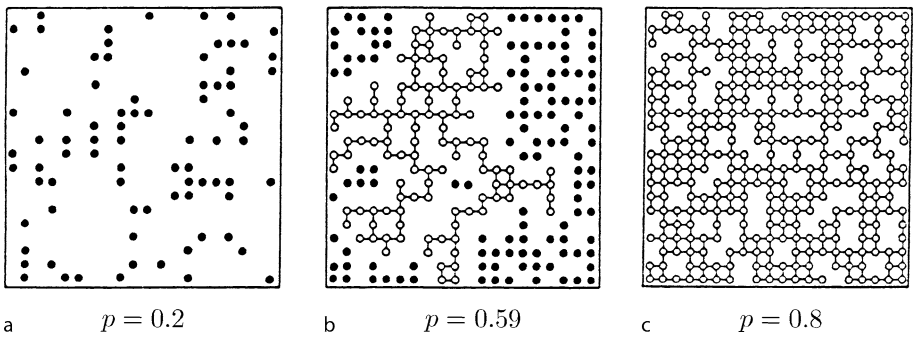

Fractals and Percolation, Figure 1
Square lattice of size 20 × 20. Sites have been randomly occupied with probability *p* (*p* = 0.20, 0.59, 0.80). Sites belonging to finite clusters are marked by *full circles*, while sites on the infinite cluster are marked by *open circles*

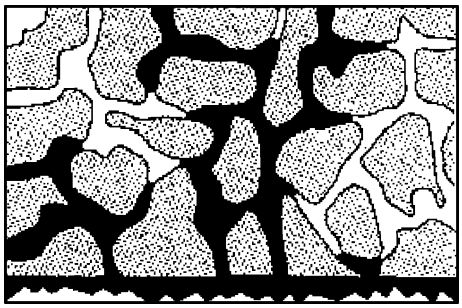

Fractals and Percolation, Figure 2
Invasion percolation through porous media

for liquid filtration (i. e., invasion *percolation* (see Fig. 2), which is the source of this terminology) through porous media.

A possible application of bond percolation in chemistry is the polymerization process [25,31,44], where small branching molecules can form large molecules by activating more and more bonds between them. If the activation probability p is above the critical concentration, a network of chemical bonds spanning the whole system can be formed, while below p_c only macromolecules of finite size can be generated. This process is called a *sol-gel* transition. An example of this *gelation* process is the boiling of an egg, which at room temperature is liquid but, upon heating, becomes a solid-like *gel*.

An example from biology concerns the spreading of an epidemic [35]. In its simplest form, an epidemic starts with one sick individual which can infect its nearest neighbors with probability p in one time step. After one time step, it dies, and the infected neighbors in turn can infect their (so far) uninfected neighbors, and the process is continued. Here the critical concentration separates a phase at low p where the epidemic always dies out after a finite number of time steps, from a phase where the epidemic can continue forever. The same process can be used as a model for forest fires [52,60,64,71], with the infection probability replaced by the probability that a burning tree can ignite its nearest-neighbor trees in the next time step. In addition to these simple examples, percolation concepts have been found useful for describing a large number of disordered systems in physics and chemistry.

The first study introducing the concept of percolation is attributable to Flory and Stockmayer about 65 years ago, when studying the gelation process [32]. The name percolation was proposed by Broadbent and Hammersley in 1957 when they were studying the spreading of fluids in random media [15]. They also introduced relevant geometrical and probabilistic concepts. The developments of phase transition theory in the following years, in particular the series expansion method by Domb [27] and renormalization group theory by Wilson, Fisher and Kadanoff [51,65], very much stimulated research activities into the geometric percolation transition.

At the percolation threshold, the conducting (as well as insulating) clusters are self-similar (see Fig. 3) and, therefore, can be described by fractal geometry [53], where various fractal dimensions are introduced to quantify the clusters and their physical properties.

Percolation

As above (see Sect. "Introduction"), consider a square lattice, where each site is occupied randomly with probability p (see Fig. 1). For simplicity, let us assume that the occupied sites are electrical conductors and the empty sites represent insulators. At low concentration p, the occupied sites either are isolated or form small clusters (Fig. 1a). Two occupied sites belong to the same cluster if they are connected by a path of nearest-neighbor occupied sites

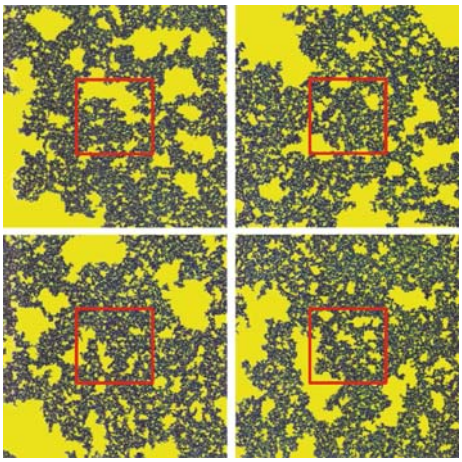

Fractals and Percolation, Figure 3
Self-similarity of the *random* percolation cluster at the critical concentration; courtesy of M. Meyer

and a current can flow between them. When p is increased, the average size of the clusters increases. At a critical concentration p_c (also called the *percolation threshold*), a large cluster appears which connects opposite edges of the lattice (Fig. 1). This cluster is called the *infinite* cluster, since its size diverges when the size of the lattice is increased to infinity. When p is increased further, the density of the infinite cluster increases, since more and more sites become part of the infinite cluster, and the average size of the *finite* clusters decreases (Fig. 1c).

The *percolation threshold* separates two different phases and, therefore, the *percolation transition* is a *geometrical phase transition*, which is characterized by the geometric features of large clusters in the neighborhood of p_c. At low values of p, only small clusters of occupied sites exist. When the concentration p is increased, the average size of the clusters increases. At the critical concentration p_c, a large cluster appears which connects opposite edges of the lattice. Accordingly, the average size of the *finite* clusters which do not belong to the infinite cluster decreases. At $p = 1$, trivially, all sites belong to the infinite cluster.

Similar to *site percolation*, it is possible to consider *bond percolation* when the bonds between sites are randomly occupied. An example of bond percolation in physics is a *random resistor network*, where the metallic wires in a regular network are cut at random. If sites are occupied with probability p and bonds are occupied with probability q, we speak of *site-bond percolation*. Two occupied sites belong to the same cluster if they are connected by a path of nearest-neighbor occupied sites with occupied bonds in between.

The definitions of site and bond percolation on a square lattice can easily be generalized to any lattice in d-dimensions. In general, in a given lattice, a bond has more nearest neighbors than a site. Thus, large clusters of bonds can be formed more effectively than large clusters of sites, and therefore, on a given lattice, the percolation threshold for bonds is smaller than the percolation threshold for sites (see Table 1).

A natural example of percolation, is *continuum percolation*, where the positions of the two components of a random mixture are not restricted to the discrete sites of a regular lattice [9,73]. As a simple example, consider a sheet of conductive material, with circular holes punched randomly in it (Swiss cheese model, see Fig. 4). The relevant quantity now is the fraction p of remaining conductive material.

Fractals and Percolation, Table 1
Percolation thresholds for the Cayley tree and several two- and three-dimensional lattices (see Refs. [8,14,41,64,75] and references therein)

Lattice	Percolation of	
	Sites	Bonds
Triangular	1/2	$2\sin(\pi/18)$
Square	0.5927460	1/2
Honeycomb	0.6962	$1 - 2\sin(\pi/18)$
Face Centered Cubic	0.198	0.119
Body Centered Cubic	0.245	0.1803
Simple Cubic (1^{st}nn)	0.31161	0.248814
Simple Cubic (2^{nd}nn)	0.137	–
Simple Cubic (3^{rd}nn)	0.097	–
Cayley Tree	$1/(z-1)$	$1/(z-1)$

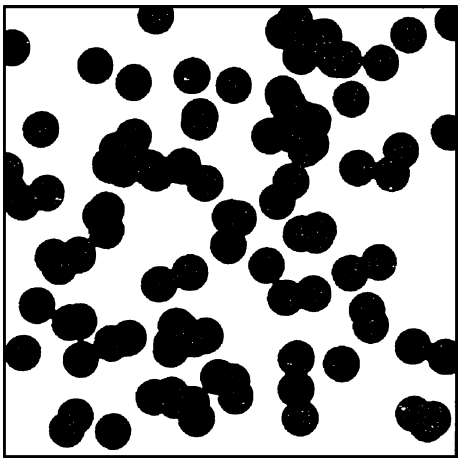

Fractals and Percolation, Figure 4
Continuum percolation: Swiss cheese model

Hopping Percolation

Above, we have discussed traditional percolation with only two values of local conductivities, 0 and 1 (insulator–conductor) or ∞ and 1 (superconductor–normal conductor). However, quantum systems should be treated by hopping conductivity, which can be described by an exponential function representing the local conductivities (between ith and jth sites): $\sigma_{ij} \sim \exp\left(-\kappa x_{ij}\right)$. Here κ can be interpreted as the dimensionless mean hopping distance or as the degree of disorder (the smaller the density of the deposited grains, the larger κ becomes), and x_{ij} is a random number taken from a uniform distribution in the range (0,1) [70].

In contrast to the traditional bond (or site) percolation model, in which the system is either a metal or an insulator, in the hopping percolation model the system always conducts some current. However, there are two regimes of such percolation [70]. A regime with many conducting paths which is not sensitive to the removal of a single bond (*weak disorder* $L/\kappa^{\nu} > 1$, where L is size of the system and ν is percolation critical exponent) and a regime with a single or only a few dominating conducting paths which is very sensitive to the removal of a specific single bond with the highest current (*strong disorder* $L/\kappa^{\nu} \ll 1$). In the strong disorder regime, the trajectories along which the highest current flows (analogous to the spanning cluster at criticality in the traditional percolation network, see Fig. 5) can be distinguished and a single bond can determine the transport properties of the entire macroscopic system.

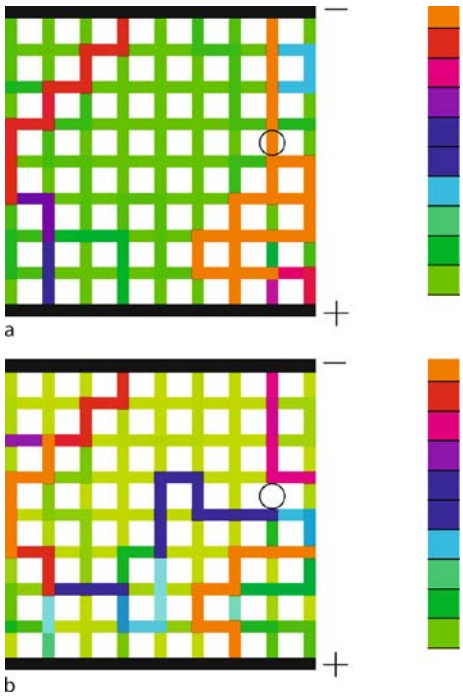

Fractals and Percolation, Figure 5
A color density plot of the current distribution in a bond-percolating lattice for which voltage is applied in the vertical direction for strong disorder with $\kappa = 10$. The current between the sites is shown by the different colors (*orange* corresponds to the highest value, *green* to the lowest). a The location of the resistor, on which the value of the local current is maximal, is shown by a *circle*. b The current distribution after removing the above resistor. This removal results in a significant change of the current trajectories

Percolation as a Critical Phenomenon

In percolation, the concentration p of occupied sites plays the same role as temperature in thermal phase transitions. The percolation transition is a geometrical phase transition where the critical concentration p_c separates a phase of finite clusters ($p < p_c$) from a phase where an infinite cluster is present ($p > p_c$).

An important quantity is the probability P_∞ that a site (or a bond) belongs to the infinite cluster. For $p < p_c$, only finite clusters exist, and $P_\infty = 0$. For $p > p_c$, P_∞ increases with p by a power law

$$P_\infty \sim (p - p_c)^{\beta} \ . \tag{1}$$

P_∞ can be identified as the *order parameter* similar to magnetization, $m(T) \sim (T_c - T)^{\beta}$, in magnetic materials. With decreasing temperature, T, more elementary magnetic moments (spins) become aligned in the same direction, and the system becomes more ordered.

The linear size of the *finite* clusters, below and above p_c, is characterized by *correlation length* ξ. Correlation length is defined as the mean distance between two sites on the same finite cluster. When p approaches p_c, ξ increases as

$$\xi \sim |p - p_c|^{-\nu} \ , \tag{2}$$

with the same exponent ν below and above the threshold. The mean number of sites (mass) of a finite cluster also diverges,

$$S \sim |p - p_c|^{-\gamma} \ , \tag{3}$$

again with the same exponent γ above and below p_c. Analogous to S in magnetic systems is the susceptibility χ (see Fig. 6 and Table 2).

The exponents β, ν, and γ describe the critical behavior of typical quantities associated with the percolation transition, and are called *critical exponents*. The exponents

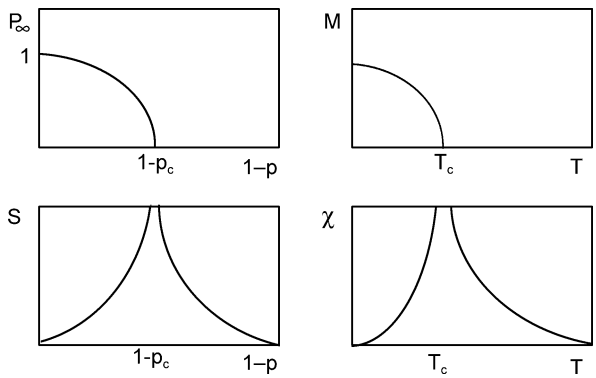

Fractals and Percolation, Figure 6
P_∞ and S compared with magnetization M and susceptibility χ

Fractals and Percolation, Table 2
Exact and best estimate values for the critical exponents for percolation (see Refs. [8,14,41,64] and references therein)

Percolation	$d = 2$	$d = 3$	$d \geq 6$
Order parameter P_∞: β	5/36	0.417 ± 0.003	1
Correlation length ξ: ν	4/3	0.875 ± 0.008	1/2
Mean cluster size S: γ	43/18	1.795 ± 0.005	1

are universal and do not depend on the structural details of the lattice (e. g., square or triangular) nor on the type of percolation (site, bond, or continuum), but depend only on the dimension d of the lattice.

This universality property is a general feature of phase transitions, where the order parameter vanishes continuously at the critical point (second order phase transition).

In Table 2, the values of the critical exponents β, ν, and γ for percolation in two, three, and six dimensions. The exponents considered here describe the geometrical properties of the percolation transition. The physical properties associated with this transition also show power-law behavior near p_c and are characterized by critical exponents. Examples include the conductivity in a random resistor or random superconducting network and the spreading velocity of an epidemic disease near the critical infection probability. It is believed that the "dynamical" exponents cannot be generally related to the geometric exponents discussed above.

Note that all quantities described above are defined in the thermodynamic limit of large systems. In a finite system, for example, P_∞, is not strictly zero below p_c.

Percolation Clusters as Fractals

As first noticed by Stanley [66], the structure of percolation clusters (when the length scale is smaller than ξ) can be well described by the fractal concept [53]. Fractal geometry is a mathematical tool for dealing with complex structures that have no characteristic length scale. Scale-invariant systems are usually characterized by noninteger ("fractal") dimensions. This terminology is associated with B. Mandelbrot [53] (though some notion of noninteger dimensions and several basic properties of fractal objects were studied earlier by G. Cantor, G. Peano, D. Hilbert, H. von Koch, W. Sierpinski, G. Julia, F. Hausdorff, C. F. Gauss, and A. Dürer).

Fractal Dimension d_f

In regular systems (with uniform density) such as long wires, large thin plates, or large filled cubes, the dimension d characterizes how the mass $M(L)$ changes with the linear size L of the system. If we consider a smaller part of a system of linear size bL ($b < 1$), then $M(bL)$ is decreased by a factor of b^d, i. e.,

$$M(bL) = b^d M(L) \,. \tag{4}$$

The solution of the functional Eq. (4) is simply $M(L) = AL^d$. For a long wire, mass changes linearly with b, i. e., $d = 1$. For the thin plates, we obtain $d = 2$, and for cubes $d = 3$; see Fig. 7.

Mandelbrot coined the name "fractal dimension", and those objects described by a fractal dimension are called fractals. Thus, to include fractal structures, (4) we can generalize

$$M(bL) = b^{d_f} M(L) \,, \tag{5}$$

and

$$M(L) = AL^{d_f} \,, \tag{6}$$

where d_f is the fractal dimension and can be a noninteger. Below, we present two examples of dealing with d_f: (i) the deterministic Sierpinski Gasket and (ii) random percolation clusters and criticality.

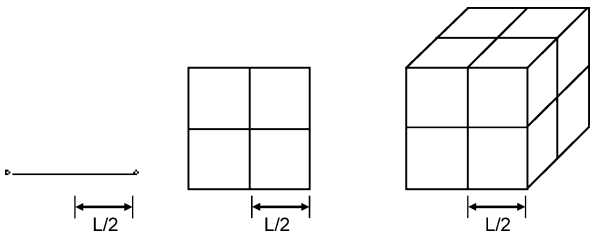

Fractals and Percolation, Figure 7
Examples of regular systems with dimensions $d = 1$, $d = 2$, and $d = 3$

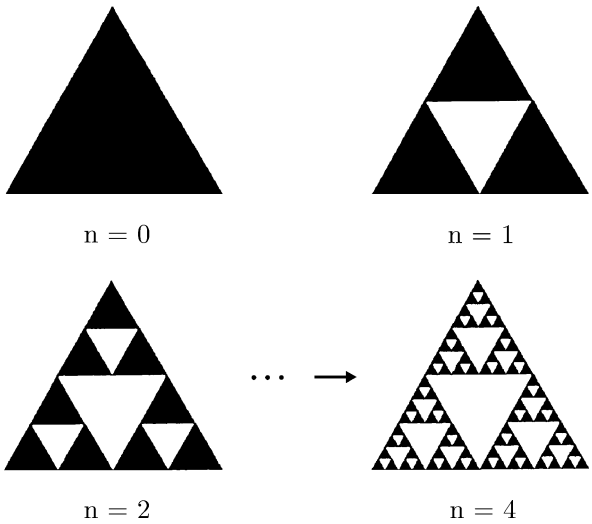

Fractals and Percolation, Figure 8
2D Sierpinski gasket. Generation and self-similarity

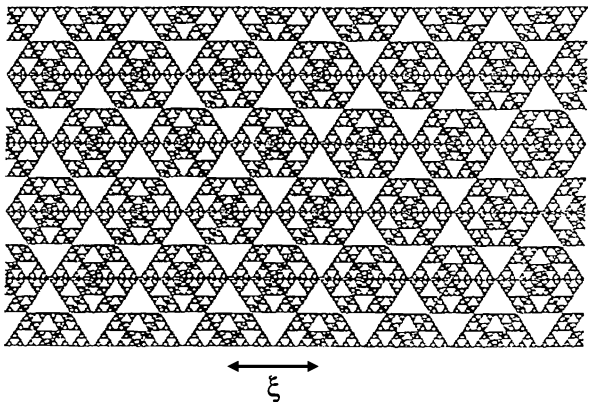

Fractals and Percolation, Figure 9
Lattice composed of Sierpinski gasket cells of size ξ

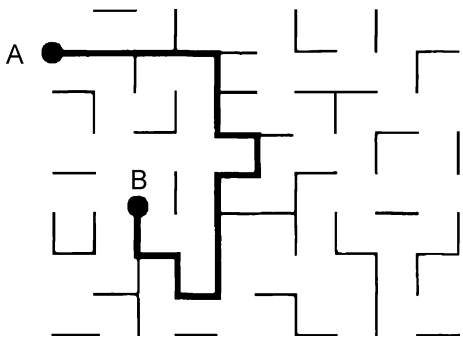

Fractals and Percolation, Figure 10
Shortest path between two sites A and B on a percolation cluster

Sierpinski Gasket This fractal is generated by dividing a full triangle into four smaller triangles and removing the central triangle (see Fig. 8). In subsequent iterations, this procedure is repeated by dividing each of the remaining triangles into four smaller triangles and removing the central triangles. To obtain the fractal dimension, we consider the mass of the gasket within a linear size L and compare it with the mass within $\frac{1}{2}L$. Since $M(\frac{1}{2}L) = \frac{1}{3}M(L)$, we have $d_f = \log 3/\log 2 \cong 1.585$.

Percolation Fractal

We assume that at p_c ($\xi = \infty$) the clusters are fractals. Thus for $p > p_c$, we expect length scales smaller than ξ to have critical properties and therefore a fractal structure. For length scales larger than ξ, one expects a homogeneous system which is composed of many unit cells of size ξ:

$$M(r) \sim \begin{cases} r^{d_f}\,, & r \ll \xi\,, \\ r^{d}\,, & r \gg \xi\,. \end{cases} \tag{7}$$

For a demonstration of this feature see Fig. 9.

One can relate the fractal dimension d_f of percolation clusters to the exponents β and ν. The probability that an arbitrary site within a circle of radius r smaller than ξ belongs to the infinite cluster, is the ratio between the number of sites on the infinite cluster and the total number of sites,

$$P_\infty \sim r^{d_f}/r^{d}\,, \qquad r < \xi\,. \tag{8}$$

This equation is certainly correct for $r = a\xi$, where a is an arbitrary constant smaller than 1. Substituting $r = a\xi$ in (8) yields $P_\infty \sim \xi^{d_f}/\xi^{d}$. Both sides are powers of $p - p_c$. Substituting Eqs. (1) and (2) into the latter one obtains [8,28,41,49,64],

$$d_f = d - \beta/\nu\,. \tag{9}$$

Thus, the fractal dimension of the infinite cluster at p_c is not a new independent exponent but depends on β and ν. Since β and ν are universal exponents, d_f is also universal.

Shortest Path Dimensions, $d_{\min}$ and d_ℓ

The fractal dimension, however, is not sufficient to fully characterize a percolation cluster, since two clusters with very different topologies may have the same fractal dimension d_f. As an additional characterization of a fractal, one can consider, e. g., the shortest path between two arbitrary sites A and B on the cluster (see Figs. 10, 11) [3,16,35,42,55,58].

The structure formed by the sites of this path is also self-similar and is described by a fractal dimension

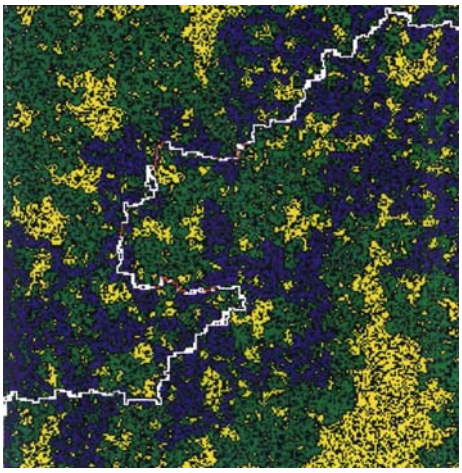

Fractals and Percolation, Figure 11
Percolation system at critical concentration in a 510 × 510 square lattice. The finite clusters are in *yellow*. Substructures of the infinite percolation cluster are shown in different colors: the shortest path between two points at opposite sites of the system is shown in *white*, the single connected sites ("red" sites) in *red*, the loops in blue and the dangling ends in *green*; courtesy of S. Schwarzer

$d_{\min}$ [46,67]. Accordingly, the length ℓ of the path, which is often called the "chemical distance", scales with the "Euclidean distance" r between A and B as

$$\ell \sim r^{d_{\min}} . \tag{10}$$

The inverse relation

$$r \sim \ell^{1/d_{\min}} \equiv \ell^{\tilde{\nu}} \tag{11}$$

tells how r scales with ℓ.

Closely related to $d_{\min}$ and d_f is the "chemical" dimension d_ℓ, which describes how the cluster mass M within the chemical distance ℓ from a given site scales with ℓ,

$$M(\ell) \sim \ell^{d_\ell} . \tag{12}$$

While the fractal dimension d_f characterizes how the mass of a cluster scales with the Euclidean distance r, the graph dimension d_ℓ characterizes how the mass scales with the chemical distance ℓ. Combining Eqs. (7), (10) and (12) we obtain the relation between $d_{\min}$, d_ℓ and d_f

$$d_\ell = d_f/d_{\min} . \tag{13}$$

To measure d_f, an arbitrary site is chosen on the cluster and one determines the number $M(r)$ of sites within a distance r from this site. To measure d_ℓ, an arbitrary site is chosen on the cluster at criticality and one determines the number $M(\ell)$ of sites which are connected to this site by a shortest path with length less than or equal to ℓ. Finally, to measure $d_{\min}$, two arbitrary sites are chosen on the cluster and one determines the length $\ell(r)$ of the shortest path connecting them. As for $M(r)$, averages must be performed for $M(\ell)$ and $r(\ell)$ over many realizations. In regular Euclidean lattices, both d_ℓ and d_f coincide with the Euclidean space dimension d and $d_{\min} = 1$.

The chemical dimension d_ℓ (or $d_{\min} = 1/\tilde{\nu}$) is an important tool for distinguishing between different fractal structures which may have a similar fractal dimension. In $d = 3$, for example, DLA (diffusion limited aggregation) clusters and percolation clusters have approximately the same fractal dimension $d_f \cong 2.5$, but have different $\tilde{\nu}$: $\tilde{\nu} = 1$ for DLA [54] but $\tilde{\nu} \cong 0.73$ for percolation [36,46,63,69].

While d_f has been related to the (known) critical exponents, (9), no such relation has been found for d_{min} or d_ℓ. The values of d_ℓ or $d_{\min}$ are known only from approximate methods, mainly numerical simulations (see also Refs. [17,37,57,76]).

Fractal Substructures

The fractal dimensions d_f and d_ℓ are not the only exponents characterizing a percolation cluster at p_c. A percolation cluster is composed of several fractal sub-structures, which are described by other exponents. Imagine applying a voltage difference between two sites at opposite edges of a metallic percolation cluster: The *backbone* of the cluster consists of those sites (or bonds) which carry the electric current. The *dangling ends* are those parts of the cluster which carry no current and are connected to the backbone by a single site only. The *red bonds* (or singly connected bonds) [22,66] are those bonds that carry the total current; when they are cut the current flow stops. In analogy to red bonds we can define anti-red bonds [34]. If an anti-red bond is added to a nonconducting percolation system below p_c, the current will be able to flow in the system. The *blobs*, finally, are those parts of the backbone that remain after the red bonds have been removed.

Further fractal substructures of the cluster are the *external perimeter* (which is also called the *hull*), the *skeleton* and the *elastic backbone*. The hull consists of those sites of the cluster which are adjacent to empty sites and are connected with infinity via empty sites. In contrast, the *total* perimeter also includes the holes in the cluster. The external perimeter is an important model for random fractal interfaces. The skeleton is defined as the union of *all* shortest paths from a given site to all sites at a chemical distance ℓ [43]. The elastic backbone is the union of all shortest paths between two sites [47].

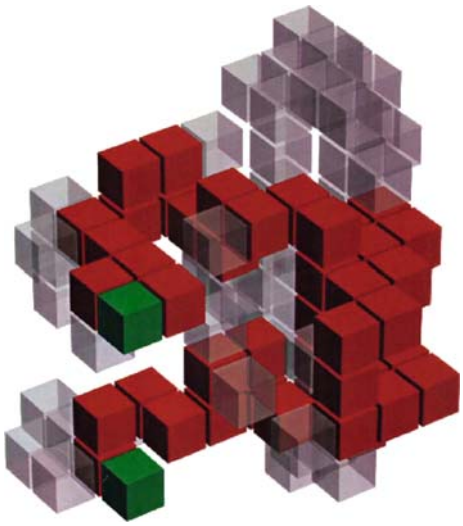

Fractals and Percolation, Figure 12
Percolation cluster at the critical concentration in a simple cubic lattice. The backbone between two (*green*) cluster sites is shown in *red*, the *gray* sites represent the dangling ends; courtesy of M. Porto

The fractal dimension d_B of the backbone is smaller than the fractal dimension d_f of the cluster (see Table 3). This reflects the fact that most of the mass of the cluster is concentrated in the dangling ends, which is seen clearly in Fig. 12. The value of the fractal dimension of the backbone is known only from numerical simulations [45,59]. Note, also, that the graph dimension d_ℓ^B of the backbone is smaller than that of percolation. In contrast, $\tilde{\nu}$ is the *same* for both backbone and percolation cluster, indicating the more universal nature of $\tilde{\nu}$. This can be understood by recalling that every two sites on a percolation cluster are located on the corresponding backbone.

The fractal dimensions of the red bonds d_{red} and the hull d_h are known from exact analytical arguments. It has been proven by Coniglio [22,23] that the mean number of red bonds varies with p as

$$n_{red} \sim (p - p_c)^{-1} \sim \xi^{1/\nu} \sim r^{1/\nu} , \tag{14}$$

and the fractal dimension of the red bonds is therefore $d_{red} = 1/\nu$. The fractal dimension of the skeleton is very close to $d_{min} = 1/\tilde{\nu}$, supporting the assumption that percolation clusters at criticality are finitely ramified [43].

The hull of the cluster in $d = 2$ has the fractal dimension $d_h = 7/4$, which was first found numerically by Sapoval, Rosso, and Gouyet [63] and proven rigorously by Saleur and Duplantier [62]. If the hull is defined slightly differently and next-nearest neighbors of the perimeter are regarded as connected, many "fjords" are removed from the hull. According to Grossmann and Aharony [38], the fractal dimension of this modified hull is close to 4/3, the fractal dimension of self-avoiding random walks in $d = 2$. In three dimensions, in contrast, the mass of the hull seems to be proportional to the mass of the cluster, and both have the same fractal dimension.

Fractals and Percolation, Table 3
Fractal dimensions of the substructures composing percolation clusters (see Ref. [8,14,41,59,64] and references therein). For fractal dimensions in $d = 4$ and $d = 5$ see Ref. [57]

Fractal dimensions	Space dimension		
	$d = 2$	$d = 3$	$d \geq 6$
d_f	91/48	2.524 ± 0.008	4
d_ℓ	1.678 ± 0.005	1.84 ± 0.02	2
d_{min}	1.13 ± 0.004	1.374 ± 0.004	2
d_{red}	3/4	1.143 ± 0.01	2
d_h	7/4	2.548 ± 0.014	4
d_B	1.64 ± 0.02	1.87 ± 0.04	2
d_ℓ^B	1.43 ± 0.02	1.34 ± 0.03	1

In Table 3 we summarize the values of the fractal dimension d_f and the graph dimension d_ℓ of the percolation cluster and its fractal substructures.

Anomalous Transport on Percolation Clusters: Diffusion and Conductivity

Due to self-similarity, transport quantities are significantly modified on fractal substrates. This can be seen in two representative examples:

(1) total resistance or the conductivity,
(2) mean square displacement and the probability density of random walks.

Consider a metallic network of size L^d. At opposite faces of the network are metallic bars with a voltage difference between them.

If we vary the linear size L of the system, the total resistance R varies as

$$R \sim \sigma^{-1} L / L^{d-1} , \tag{15}$$

where $\sigma \sim L^0 = \text{const}$ is the conductivity (inverse to the resistivity, $\sigma = \rho^{-1}$) of the metal. Since σ does not depend on L, (15) states that the total resistance of the network depends on its linear size L via the power law $R \sim L^{2-d} \equiv L^{\tilde{\zeta}}$, which defines the resistance exponent $\tilde{\zeta}$, here $\tilde{\zeta} = 2 - d$.

The idea that transport properties of percolation systems can be efficiently studied by means of diffusion was suggested by de Gennes [24] (see also Kopelman [50]).

The diffusion process can be modeled by random walkers, which can jump randomly between nearest-neighbor occupied sites in the lattice. For such a random walker moving in a disordered environment, including bottlenecks, loops, and dead ends, de Gennes coined the term *ant in the labyrinth*.

By calculating the mean square displacement of the walker, one obtains the diffusion constant, which according to Einstein is proportional to dc conductivity. Not only are the conductivity and diffusion exponents above p_c related; also related are the exponents characterizing the size dependence of the dc conductivity and the time dependence of the mean square displacement of the random walker. Since it is numerically more efficient to calculate the relevant transport quantities by simulating random walks than to determine the conductivity directly from Kirchhoff's equations, the study of random walks has improved our knowledge not only of diffusion but also of the transport process in percolation in general [5,6,10,12,18,39,40,56,74].

Due to the presence of large holes, bottlenecks, and dangling ends in the fractal, the motion of a random walker is slowed down. Fick's law for the mean square displacement ($\langle r^2(t)\rangle = a^2 t$) is no longer valid. Instead, the mean square displacement is described by a more general power law,

$$\langle r^2(t)\rangle \sim t^{2/d_w} , \tag{16}$$

where the new exponent d_w ("diffusion exponent" or "fractal dimension of the random walk") is always greater than 2.

Both the resistance exponent $\tilde{\zeta}$ and the exponent d_w can be related by the Einstein equation

$$\sigma = (e^2 n/k_B T) D , \tag{17}$$

which relates the dc conductivity σ of the system to the diffusion constant $D = \lim_{t\to\infty}\langle r^2(t)\rangle/2dt$ of the random walk. In Eq. (17), e and n denote charge and density of the mobile particles, respectively.

Simple scaling arguments can now be used to relate d_w to $\tilde{\zeta}$ and $\tilde{\mu}$. Since n is proportional to the density of the substrate, $n \sim L^{d_f-d}$, the right-hand side of Eq. (17) is proportional to $L^{d_f-d}\, t^{2/d_w-1}$. The left-hand side of Eq. (17) is proportional to $L^{-\tilde{\mu}}$. Since the time a random walker takes to travel a distance L scales as L^{d_w}, we find $L^{-\tilde{\mu}} \sim L^{d_f-d+2-d_w}$, from which the Einstein relation [4]

$$d_w = d_f - d + 2 + \tilde{\mu} = d_f + \tilde{\zeta} \tag{18}$$

follows. For example, for the Sierpinski gasket $d_f = \log 3/\log 2$, and $\tilde{\zeta} = \log(5/3)/\log(2/4)$, therefore $d_w = \log 5/\log 2$.

In general, determining d_w for random fractals is not easy. An exception is topologically linear fractal structures ($d_\ell = 1$), which can be considered as nonintersecting paths. Along a path (in ℓ-space), diffusion is normal and $\langle \ell^2(t)\rangle = t$. Since $\ell \sim r^{d_f}$, the mean square displacement in r-space scales as $\langle r^2\rangle \sim t^{1/d_f}$, leading to $d_w = 2d_f$ in this case. In percolation, d_w cannot be calculated exactly, but upper and lower bounds can be derived which are very close to each other in $d \geq 3$ dimensions. A good estimate is $d_w \cong 3d_f/2$ (Alexander–Orbach conjecture [4]).

The long-term behavior of the mean square displacement of a random walker on an infinite percolation cluster is characterized by the diffusion constant D. It is easy to see that D is related to the diffusion constant D' of the whole percolation system: above p_c, the dc conductivity of the percolation system increases as $\sigma \sim (p-p_c)^{\mu}$, so due to the Einstein relation, Eq. (17), the diffusion constant D' must also increase in this way. The mean square displacement (and hence D') is obtained by averaging over all possible starting points of a particle in the percolation system. It is clear that only those particles which start on the infinite cluster can travel from one side of the system to the other, and thus contribute to D'. Particles that start on a finite cluster cannot leave the cluster, and thus do not contribute to D'. Hence D' is related to D by $D' = DP_\infty$, implying

$$D \sim (p-p_c)^{\mu-\beta} \sim \xi^{-(\mu-\beta)/\nu} . \tag{19}$$

Combining (16) and (19), the mean square displacement on the infinite cluster can be written as [7,33,72]

$$\langle r^2(t)\rangle \sim \begin{cases} t^{2/d_w} & \text{if } t \ll t_\xi , \\ (p-p_c)^{\mu-\beta}\, t & \text{if } t \gg t_\xi , \end{cases} \tag{20}$$

where

$$t_\xi \sim \xi^{d_w} \tag{21}$$

describes the time scale the random walker needs, on average, to explore the fractal regime in the cluster. As $\xi \sim (p-p_c)^{-\nu}$ is the only length scale here, t_ξ is the only relevant time scale, and we can bridge the short time regime and the long time regime by a scaling function $f(t/t_\xi)$,

$$\langle r^2(t)\rangle = t^{2/d_w} f(t/t_\xi) . \tag{22}$$

To satisfy (20)–(21), we require $f(x) \sim x^0$ for $x \ll 1$ and $f(x) \sim x^{1-2/d_w}$ for $x \gg 1$. The first relation trivially satisfies (20)–(21). The second relation gives $D = \lim_{t\to\infty}\langle r^2(t)\rangle/2dt \sim t_\xi^{2/d_w-1}$, which in connection

with (19) and (21) yields a relation between d_w and μ [7, 33,72],

$$d_w = 2 + (\mu - \beta)/\nu \,. \tag{23}$$

Comparing (18) and (23) one can express the exponent $\tilde{\mu}$ by μ,

$$\tilde{\mu} = \mu/\nu \,. \tag{24}$$

Networks

Networks are defined as nodes connected by links, called graphs in mathematics. Many real world system can be describe as networks. Perhaps the best known example of a network is the Internet, where computers (nodes) around the globe are connected by cables (links) in such a way that an email message can travel from one computer to another by traveling along only a few links. Social relations between people can be represented by a social network [2]; nodes represent people and links represent their relations. One important property of a network is the "small world" phenomena. The shortest path (minimum number of hops) between any two nodes is very small, of the order of $\log N$ or smaller, where N is the number of nodes in the network [11,19,29]. The lattices discussed in Sect. "Percolation" are also networks, where the sites of the lattice are the nodes and the bonds represent the links. In this case, the number of links per node is fixed but, in general, the number of links per node can be taken from any degree distribution, $P(k)$. In lattices, due to spatial constraints, the distances between nodes is large and scales as $N^{1/d}$, where d is the dimension of the lattice. Since many networks have no spatial constraints, it follows that such networks can be regarded as embedded in infinite dimension, $d = \infty$, justifying a very small distance, of order $\log N$. In networks which are not constrained to geographical space, there is no Euclidean distance and the distance metric is only the shortest path ℓ defined in Sect. "Percolation Clusters as Fractals".

We will show in this chapter that ideas from percolation and fractals can be applied to obtain useful results in networks which are not embedded in space. The main difference compared to lattices is that the condition for percolation is no longer the spanning property, but rather the property of having a cluster containing number of nodes of order N, where N is the total original number of nodes in the network. Such a component, if it exists, is termed the *giant component*. The condition for the existence of a giant component above the percolation threshold, and its absence below the threshold, applies also to lattices, and therefore can be regarded as more general than the spanning property. An interesting property of percolation, called *universality*, is that the behavior at and close to the critical point depends only on the dimensionality of the lattice, and not on the microscopic connection details of the lattice. This behavior is characterized by a set of critical exponents that are the same for all two-dimensional lattices, square, triangular or hexagonal, and for either site or bond percolation. However, a different set of critical exponents will be obtained for a lattice of another dimension. Furthermore, above some critical dimension ($d_c = 6$ for percolation in d-dimensional lattices), known as the upper critical dimension, the critical behavior remains the same for all $d \geq d_c$. This is due to the insignificance of loops in high dimensions, and thus usually allows for easy determination of the critical exponents for high dimensions, using the "infinite dimensional" or "mean field" approach. Erdős and Rényi (ER) studied an ensemble of networks with N nodes and $2M$ links that randomly connect pairs of nodes. They found that $p_c = 1/\langle k \rangle = N/2M$. Percolation on ER networks or on infinite dimensional lattices, as well as on Cayley trees, has the same critical exponents, due to the fact that their topology is the same and no spatial constraint is imposed on the networks. For ER networks, as for lattices, in $d \geq 6$ the size S of the percolation cluster at p_c, scales with N as, $S \sim N^{2/3}$ [11,20].

The value of 2/3 can be related to the upper critical dimension, $d_c = 6$, and to the fractal dimension of percolation clusters $d_f = 4$ for $d = 6$. Since $N = L^6$ and $S = L^4$, it follows that $S \sim N^{2/3}$.

In recent years it was realized [2] that $P(k)$ for many real networks is very broad and, in many cases, is best represented by a power law, $P(k) \sim k^{-\gamma}$. Networks with a power law degree distribution are called *scale free (SF) network*. Heterogeneity of the degrees may affect critical behavior, even above the upper critical dimension. The heterogeneity of the degrees can be regarded as a breakdown of translational symmetry that exists in lattices, ER networks and Cayley trees. In these cases, each node has a typical number of neighbors, while in scale free networks the variation between node degrees is very large. A general result for p_c for any random network with a given degree distribution is [21]

$$p_c = \frac{1}{\kappa - 1} \,, \quad \kappa \equiv \frac{\langle k^2 \rangle}{\langle k \rangle} \,.$$

This result yields that for $\gamma < 3$, $\langle k^2 \rangle \to \infty$ and therefore $p_c \to 0$. That is, no finite percolation threshold exists. Thus, even if most of the nodes of the Internet are removed, those which are left can still communicate. This explains the puzzle of why viruses and worms stay a long time in the Internet even if many people use antivirus soft-

ware. It also explains why in order to effectively immunize populations, one needs to immunize most of the people.

The percolation critical exponents for SF networks are still mean-field or infinite-dimensional in the sense of the insignificance of loops. However, they are different from those of the standard mean field percolation. Indeed, for scale free networks [2], the size of the spanning percolation cluster is [20],

$$S \sim \begin{cases} N^{(\gamma-2)/(\gamma-1)}, & 3 < \gamma < 4 \\ N^{2/3}, & \gamma > 4 \end{cases} \tag{25}$$

As shown above, for $\gamma < 3$, there is no percolation threshold, and therefore no spanning percolation cluster. Note that SF networks can be regarded as a generalization of ER networks, since for $\gamma > 4$ one obtains the ER network results.

Summary and Future Directions

The percolation problem and its numerous modifications can be useful in describing several physical, chemical, and biological processes, such as the spreading of epidemics or forest fires, gelation processes, and the invasion of water into oil in porous media, which is relevant for the process of recovering oil from porous rocks. In some cases, modification changes the universality class of a percolation. We begin with an example in which the universality class does not change. We showed that a random process such as percolation can lead naturally to fractal structures. This may be one of the reasons why fractals occur frequently in nature.

Bibliography

Primary Literature

1. Aharony A (1986) In: Grinstein G, Mazenko G (eds) Directions in condensed matter physics. World Scientific, Singapore
2. Albert R, Barabasi A-L (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47
3. Alexandrowicz Z (1980) Phys Lett A 80:284
4. Alexander S, Orbach R (1982) J Phys Lett 43:L625
5. Alexander S, Bernasconi J, Schneider WR, Orbach R (1981) Rev Mod Phys 53:175; Alexander S (1983) In: Deutscher G, Zallen R, Adler J (eds) Percolation Structures and Processes. Adam Hilger, Bristol, p 149
6. Avnir D (ed) (1989) The fractal approach to heterogeneos chemistry. Wiley, Chichester
7. Ben-Avraham D, Havlin S (1982) J Phys A 15:L691; Havlin S, Ben-Avraham D, Sompolinsky H (1983) Phys Rev A 27: 1730
8. Ben-Avraham D, Havlin S (2000) Diffusion and reactions in fractals and disordered systems. Cambridge University Press, Cambridge
9. Benguigui L (1984) Phys Rev Lett 53:2028
10. Blumen A, Klafter J, Zumofen G (1986) In: Zschokke I (ed) Optical spectroscopy of glasses. Reidel, Dordrecht, pp 199–265
11. Bollobás B (1985) Random graphs. Academic Press, London
12. Bouchaud JP, Georges A (1990) Phys Rep 195:127
13. Bunde A (1986) Adv Solid State Phys 26:113
14. Bunde A, Havlin S (eds) (1996) Fractals and disordered systems, 2nd edn. Springer, Berlin; Bunde A, Havlin S (eds) (1995) Fractals in Science, 2nd edn. Springer, Berlin
15. Broadbent SR, Hammersley JM (1957) Proc Camb Phil Soc 53:629
16. Cardey JL, Grassberger P (1985) J Phys A 18:L267
17. Cardy J (1998) J Phys A 31:L105
18. Clerc JP, Giraud G, Laugier JM, Luck JM (1990) Adv Phys 39:191
19. Cohen R, Havlin S (2003) Scale-free networks are ultrasmall. Phys Rev Lett 90:058701
20. Cohen R, Havlin S (2008) Complex networks: Structure, stability and function. Cambridge University Press, Cambridge
21. Cohen R, Erez K, Ben-Avraham D, Havlin S (2000) Resilience of the internet to random breakdowns. Phys Rev Lett 85:4626
22. Coniglio A (1982) J Phys A 15:3829
23. Coniglio A (1982) Phys Rev Lett 46:250
24. de Gennes PG (1976) La Recherche 7:919
25. de Gennes PG (1979) Scaling concepts in polymer physics. Cornell University Press, Ithaca
26. Deutscher G, Zallen R, Adler J (eds) (1983) A collection of review articles: percolation structures and processes. Adam Hilger, Bristol
27. Domb C (1983) In: Deutscher G, Zallen R, Adler J (eds) Percolation structures and processes. Adam Hilger, Bristol; Domb C, Stoll E, Schneider T (1980) Contemp Phys 21: 577
28. Elam WT, Kerstein AR, Rehr JJ (1984) Phys Rev Lett 52:1515
29. Erdős P, Rényi A (1959) On random graphs. Publicationes Mathematicae 6:290; (1960) Publ Math Inst Hung Acad Sci 5:17
30. Essam JW (1980) Rep Prog Phys 43:843
31. Family F, Landau D (eds) (1984) Kinetics of aggregation and gelation. North Holland, Amsterdam; For a review on gelation see: Kolb M, Axelos MAV (1990) In: Stanley HE, Ostrowsky N (eds) Correlatios and Connectivity: Geometric Aspects of Physics, Chemistry and Biology. Kluwer, Dordrecht, p 225
32. Flory PJ (1971) Principles of polymer chemistry. Cornell University, New York; Flory PJ (1941) J Am Chem Soc 63:3083–3091–3096; Stockmayer WH (1943) J Chem Phys 11:45
33. Gefen Y, Aharony A, Alexander S (1983) Phys Rev Lett 50:77
34. Gouyet JF (1992) Phys A 191:301
35. Grassberger P (1986) Math Biosci 62:157; (1985) J Phys A 18:L215; (1986) J Phys A 19:1681
36. Grassberger P (1992) J Phys A 25:5867
37. Grassberger P (1999) J Phys A 32:6233
38. Grossman T, Aharony A (1987) J Phys A 20:L1193
39. Haus JW, Kehr KW (1987) Phys Rep 150:263
40. Havlin S, Ben-Avraham D (1987) Adv Phys 36:695
41. Havlin S, Ben-Avraham D (1987) Diffusion in random media. Adv Phys 36:659
42. Havlin S, Nossal R (1984) Topological properties of percolation clusters. J Phys A 17:L427
43. Havlin S, Nossal R, Trus B, Weiss GH (1984) J Stat Phys A 17:L957
44. Herrmann HJ (1986) Phys Rep 136:153
45. Herrmann HJ, Stanley HE (1984) Phys Rev Lett 53:1121; Hong DC, Stanley HE (1984) J Phys A 16:L475
46. Herrmann HJ, Stanley HE (1988) J Phys A 21:L829
47. Herrmann HJ, Hong DC, Stanley HE (1984) J Phys A 17:L261

48. Kesten H (1982) Percolation theory for mathematicians. Birkhauser, Boston (A mathematical approach); Grimmet GR (1989) Percolation. Springer, New York
49. Kirkpatrick S (1979) In: Maynard R, Toulouse G (eds) Le Houches Summer School on Ill Condensed Matter. North Holland, Amsterdam
50. Kopelman R (1976) In: Fong FK (ed) Topics in applied physics, vol 15. Springer, Heidelberg
51. Ma SK (1976) Modern theory of critical phenomena. Benjamin, Reading
52. Mackay G, Jan N (1984) J Phys A 17:L757
53. Mandelbrot BB (1982) The fractal geometry of nature. Freeman, San Francisco; Mandelbrot BB (1977) Fractals: Form, Chance and Dimension. Freeman, San Francisco
54. Meakin P, Majid I, Havlin S, Stanley HE (1984) J Phys A 17:L975
55. Middlemiss KM, Whittington SG, Gaunt DC (1980) J Phys A 13:1835
56. Montroll EW, Shlesinger MF (1984) In: Lebowitz JL, Montroll EW (eds) Nonequilibrium phenomena II: from stochastics to hydrodynamics. Studies in Statistical Mechanics, vol 2. North-Holland, Amsterdam
57. Paul G, Ziff RM, Stanley HE (2001) Phys Rev E 64:26115
58. Pike R, Stanley HE (1981) J Phys A 14:L169
59. Porto M, Bunde A, Havlin S, Roman HE (1997) Phys Rev E 56:1667
60. Ritzenberg AL, Cohen RI (1984) Phys Rev B 30:4036
61. Sahimi M (1993) Application of percolation theory. Taylor Francis, London
62. Saleur H, Duplantier B (1987) Phys Rev Lett 58:2325
63. Sapoval B, Rosso M, Gouyet JF (1985) J Phys Lett 46:L149
64. Stauffer D, Aharony A (1994) Introduction to percolation theory, 2nd edn. Taylor Francis, London
65. Stanley HE (1971) Introduction to phase transition and critical phrenomena. Oxford University, Oxford
66. Stanley HE (1977) J Phys A 10:L211
67. Stanley HE (1984) J Stat Phys 36:843
68. Turcotte DL (1992) Fractals and chaos. In: Geology and geophysics. Cambridge University Press, Cambridge
69. Toulouse G (1974) Nuovo Cimento B 23:234
70. Tyc S, Halperin BI (1989) Phys Rev B 39:R877; Strelniker YM, Berkovits R, Frydman A, Havlin S (2004) Phys Rev E 69:R065105; Strelniker YM, Havlin S, Berkovits R, Frydman A (2005) Phys Rev E 72:016121; Strelniker YM (2006) Phys Rev B 73:153407
71. von Niessen W, Blumen A (1988) Canadian J For Res 18:805
72. Webman I (1991) Phys Rev Lett 47:1496
73. Webman I, Jortner J, Cohen MH (1976) Phys Rev B 14:4737
74. Weiss GH, Rubin RJ (1983) Adv Chem Phys 52:363; Weiss GH (1994) Aspects and applications of the random walk. North Holland, Amsterdam
75. Ziff RM (1992) Phys Rev Lett 69:2670
76. Ziff RM (1999) J Phys A 32:L457

Books and Reviews

Bak P (1996) How nature works: The science of self organized criticality. Copernicus, New York
Barabási A-L (2003) Linked: How everything is connected to everything else and what it means for business, science and everyday life. Plume
Bergman DJ, Stroud D (1992) Solid State Phys 46:147–269
Dorogovtsev SN, Mendes JFF (2003) Evolution of networks: From biological nets to the internet and www (physics). Oxford University Press, Oxford
Eglash R (1999) African fractals: Modern computing and indigenous design. Rutgers University Press, New Brunswick, NJ
Feder J (1988) Fractals. Plenum, New York
Gleick J (1997) Chaos. Penguin Books, New York
Gould H, Tobochnik J (1988) An introduction to computer simulation methods. In: Application to physical systems. Addison-Wesley, Reading, MA
Meakin P (1998) Fractals, scaling and growth far from equilibrium. Cambridge University Press, Cambridge
Pastor-Satorras R, Vespignani A (2004) Evolution and structure of the internet: A statistical physics approach. Cambridge University Press, Cambridge
Peitgen HO, Jurgens H, Saupe D (1992) Chaos and fractals. Springer, New York
Peng G, Decheng T (1990) The fractal nature of a fracture surface. J Physics A 14:3257–3261
Pikovsky A, Rosenblum M, Kurths J, Chirikov B, Cvitanovic P, Moss F, Swinney H (2003) Synchronization: A universal concept in nonlinear sciences. Cambridge University Press, Cambridge
Vicsek T (1992) Fractal growth phenomena. World Scientific, Singapore

Fractals in the Quantum Theory of Spacetime

Laurent Nottale
CNRS, Paris Observatory and Paris Diderot University, Paris, France

Article Outline

Glossary
Definition of the Subject
Introduction
Foundations of Scale Relativity Theory
Scale Laws
From Fractal Space to Nonrelativistic Quantum Mechanics
From Fractal Space-Time to Relativistic Quantum Mechanics
Gauge Fields as Manifestations of Fractal Geometry
Future Directions
Bibliography

Glossary

Fractality In the context of the present article, the geometric property of being structured over all (or many) scales, involving explicit scale dependence which may go up to scale divergence.

Spacetime Inter-relational level of description of the set of all positions and instants (events) and of their transformations. The events are defined with respect to a given reference system (i. e., in a relative way), but a spacetime is characterized by invariant relations which are valid in all reference systems, such as, e. g., the metric invariant. In the generalization to a fractal space-time, the events become explicitly dependent on resolution.

Relativity The property of physical quantities according to which they can be defined only in terms of relationships, not in an abolute way. These quantities depend on the state of the reference system, itself defined in a relative way, i. e., with respect to other coordinate systems.

Covariance Invariance of the form of equations under general coordinate transformations.

Geodesics Curves in a space (more generally in a spacetime) which minimize the proper time. In a geometric spacetime theory, the motion equation is given by a geodesic equation.

Quantum Mechanics Fundamental axiomatic theory of elementary particle, nuclear, atomic, molecular, etc. physical phenomena, according to which the state of a physical system is described by a wave function whose square modulus yields the probability density of the variables, and which is solution of a Schrödinger equation constructed from a correspondence principle (among other postulates).

Definition of the Subject

The question of the foundation of quantum mechanics from first principles remains one of the main open problems of modern physics. In its current form, it is an axiomatic theory of an algebraic nature founded upon a set of postulates, rules and derived principles. This is to be compared with Einstein's theory of gravitation, which is founded on the principle of relativity and, as such, is of an essentially geometric nature. In its framework, gravitation is understood as a very manifestation of the curvature of a Riemannian space-time.

It is therefore relevant to question the nature of the quantum space-time and to ask for a possible refoundation of the quantum theory upon geometric first principles. In this context, it has been suggested that the quantum laws and properties could actually be manifestations of a fractal and nondifferentiable geometry of space-time [52,69,71], coming under the principle of scale relativity [53,54]. This principle extends, to scale transformations of the reference system, the theories of relativity (which have been, up to now, applied to transformations of position, orientation and motion).

Such an approach allows one to recover the main tools and equations of standard quantum mechanics, but also to suggest generalizations, in particular toward high energies, since it leads to the conclusion that the Planck length scale could be a minimum scale in nature, unreachable and invariant under dilations [53].

But it has another radical consequence. Namely, it allows the possibility of a new form of macroscopic quantum-type behavior for a large class of complex systems, namely those whose behavior is characterized by Newtonian dynamics, fractal stochastic fluctuations over a large range of scales, and small scale irreversibility. In this case the equation of motion may take the form of a Schrödinger equation, which yields peaks of probability density according to the symmetry, field and limit conditions. These peaks may be interpreted as a tendency for the system to form structures [57], in terms of a macroscopic constant which is no longer $\hbar$, therefore possibly leading to a new theory of self-organization.

It is remarkable that, under such a relativistic view, the question of complexity may be posed in an original way. Namely, in a fully relativistic theory there is no intrinsic complexity, since the various physical properties of an 'object' are expected to vanish in the proper system of coordinates linked to the object. Therefore, in such a framework the apparent complexity of a system comes from the complexity of the change of reference frames from the proper frame to the observer (or measurement apparatus) reference frame. This does not mean that the complexity can always be reduced, since this change can itself be infinitely complex, as it is the case in the situation described here of a fractal and nondifferentiable space-time.

Introduction

There have been many attempts during the 20th century at understanding the quantum behavior in terms of differentiable manifolds. The failure of these attempts indicates that a possible 'quantum geometry' should be of a completely new nature. Moreover, following the lessons of Einstein's construction of a geometric theory of gravitation, it seems clear that any geometric property to be attributed to space-time itself, and not only to particular objects or systems embedded in space-time, must necessarily be universal.

Fortunately, the founders of quantum theory have brought to light a universal and fundamental behavior of the quantum realm, in opposition to the classical world; namely, the explicit dependence of the measurement re-

sults on the apparatus resolution described by the Heisenberg uncertainty relations. This leads one to contemplate the possibility that the space-time underlying the quantum mechanical laws can be explicitly dependent on the scale of observation [52,69,71]. Now the concept of a scale-dependent geometry (at the level of objects and media) has already been introduced and developed by Benoit Mandelbrot, who coined the word 'fractal' in 1975 to describe it. But here we consider a complementary program that uses fractal geometry, not only for describing 'objects' (that remain embedded in an Euclidean space), but also for intrinsically describing the geometry of space-time itself.

A preliminary work toward such a goal may consist of introducing the fractal geometry in Einstein's equations of general relativity at the level of the source terms. This would amount to giving a better description of the density of matter in the Universe accounting for its hierarchical organization and fractality over many scales (although possibly not all scales), then to solve Einstein's field equations for such a scale dependent momentum-energy tensor. A full implementation of this approach remains a challenge to cosmology.

But a more direct connection of the fractal geometry with fundamental physics comes from its use in describing not only the distribution of matter in space, but also the geometry of space-time itself. Such a goal may be considered as the continuation of Einstein's program of generalization of the geometric description of space-time. In the new fractal space-time theory, [27,52,54,69,71], the essence of quantum physics is a manifestation of the nondifferentiable and fractal geometry of space-time.

Another line of thought leading to the same suggestion comes, not from relativity and space-time theories, but from quantum mechanics itself. Indeed, it has been discovered by Feynman [30] that the typical quantum mechanical paths (i. e., those that contribute in a main way to the path integral) are nondifferentiable and fractal. Namely, Feynman has proved that, although a mean velocity can be defined for them, no mean-square velocity exists at any point, since it is given by $\langle v^2 \rangle \propto \delta t^{-1}$. One now recognizes in this expression the behavior of a curve of fractal dimension $D_F = 2$ [1].

Based on these premises, the reverse proposal, according to which the laws of quantum physics find their very origin in the fractal geometry of space-time, has been developed along three different and complementary approaches.

Ord and co-workers [50,71,72,73], extending the Feynman chessboard model, have worked in terms of probabilistic models in the framework of the statistical mechanics of binary random walks.

El Naschie has suggested to give up not only the differentiability, but also the continuity of space-time. This leads him to work in terms of a 'Cantorian' space-time [27,28,29], and to therefore use in a preferential way the mathematical tool of number theory (see a more detailed review of these two approaches in Ref. [45]).

The scale relativity approach [52,54,56,63,69] which is the subject of the present article, is, on the contrary, founded on a fundamentally continuous geometry of space-time which therefore includes the differentiable and nondifferentiable cases, constrained by the principle of relativity applied to both motion and scale.

Other applications of fractals to the quantum theory of space-time have been proposed in the framework of a possible quantum gravity theory. They are of another nature than those considered in the present article, since they are applicable only in the framework of the quantum theory instead of deriving it from the geometry, and they concern only very small scales on the magnitude of the Planck scale. We send the interested reader to Kröger's review paper on "Fractal geometry in quantum mechanics, field theory and spin systems", and to references therein [45].

In the present article, we summarize the steps by which one recovers, in the scale relativity and fractal space-time framework, the main tools and postulates of quantum mechanics and of gauge field theories. A more detailed account can be found in Refs. [22,23,54,56,63,67,68], including possible applications of the theory to various sciences.

Foundations of Scale Relativity Theory

The theory of scale relativity is based on giving up the hypothesis of manifold differentiability. In this framework, the coordinate transformations are continuous but can be nondifferentiable. This has several consequences [54], leading to the following preliminary steps of construction of the theory:

(1) One can prove the following theorem [5,17,18,54,56]: a continuous and nondifferentiable curve is fractal in a general meaning, namely, its length is explicitly dependent on a scale variable ε, i. e., $\mathcal{L} = \mathcal{L}(\varepsilon)$, and it diverges, $\mathcal{L} \to \infty$, when $\varepsilon \to 0$. This theorem can be readily extended to a continuous and nondifferentiable manifold, which is therefore fractal not as an hypothesis, but as a consequence of giving up an hypothesis (that of differentiability).

(2) The fractality of space-time [52,54,69,71] involves the scale dependence of the reference frames. One therefore adds a new variable ε which characterizes the 'state of scale' to the usual variables defining the coordinate system.

In particular, the coordinates themselves become functions of these scale variables, i. e., $X = X(\varepsilon)$.

(3) The scale variables ε can never be defined in an absolute way, but only in a relative way. Namely, only their ratio $\rho = \varepsilon'/\varepsilon$ has a physical meaning. In experimental situations, these scale variables amount to the resolution of the measurement apparatus (it may be defined as standard errors, intervals, pixel size, etc.). In a theoretical analysis, they are the space and time differential elements themselves. This universal behavior extends to the principle of relativity in such a way that it also applies to the transformations (dilations and contractions) of these resolution variables [52,53,54].

Scale Laws

Fractal Coordinate and Differential Dilation Operator

Consider a variable length measured on a fractal curve, and (more generally) a non-differentiable (fractal) curvilinear coordinate $\mathcal{L}(s, \varepsilon)$, that depends on some parameter s which characterizes the position on the curve (it may be, e. g., a time coordinate), and on the resolution ε. Such a coordinate generalizes the concept of curvilinear coordinates introduced for curved Riemannian space-times in Einstein's general relativity [54] to nondifferentiable and fractal space-times.

Such a scale-dependent fractal length $\mathcal{L}(s, \varepsilon)$ remains finite and differentiable when $\varepsilon \neq 0$; namely, one can define a slope for any resolution ε, being aware that this slope is itself a scale-dependent fractal function. It is only at the limit $\varepsilon \to 0$ that the length is infinite and the slope undefined, i. e., that nondifferentiability manifests itself.

Therefore the laws of dependence of this length upon position and scale may be written in terms of a double differential calculus, i. e., it can be the solution of differential equations involving the derivatives of $\mathcal{L}$ with respect to both s and ε.

As a preliminary step, one needs to establish the relevant form of the scale variables and the way they intervene in scale differential equations. For this purpose, let us apply an infinitesimal dilation $\mathrm{d}\rho$ to the resolution, which is therefore transformed as $\varepsilon \to \varepsilon' = \varepsilon(1 + \mathrm{d}\rho)$. The dependence on position is omitted at this stage in order to simplify the notation. By applying this transformation to a fractal coordinate $\mathcal{L}$, one obtains, to the first order in the differential element,

$$\begin{aligned}\mathcal{L}(\varepsilon') &= \mathcal{L}(\varepsilon + \varepsilon\,\mathrm{d}\rho) = \mathcal{L}(\varepsilon) + \frac{\partial \mathcal{L}(\varepsilon)}{\partial \varepsilon}\,\varepsilon\,\mathrm{d}\rho \\ &= (1 + \tilde{D}\mathrm{d}\rho)\,\mathcal{L}(\varepsilon)\,, \end{aligned} \tag{1}$$

where $\tilde{D}$ is, by definition, the dilation operator.

Since $\mathrm{d}\varepsilon/\varepsilon = \mathrm{d}\ln\varepsilon$, the identification of the two last members of Eq. (1) yields

$$\tilde{D} = \varepsilon\,\frac{\partial}{\partial \varepsilon} = \frac{\partial}{\partial \ln \varepsilon}\,. \tag{2}$$

This form of the infinitesimal dilation operator shows that the natural variable for the resolution is $\ln\varepsilon$, and that the expected new differential equations will indeed involve quantities such as $\partial\mathcal{L}(s, \varepsilon)/\partial\ln\varepsilon$. This theoretical result agrees and explains the current knowledge according to which most measurement devices (of light, sound, etc.), including their physiological counterparts (eye, ear, etc.) respond according to the logarithm of the intensity (e. g., magnitudes, decibels, etc.).

Self-similar Fractals as Solutions of a First Order Scale Differential Equation

Let us start by writing the simplest possible differential equation of scale, and then solve it. We shall subsequently verify that the solutions obtained comply with the principle of relativity. As we shall see, this very simple approach already yields a fundamental result: it gives a foundation and an understanding from first principles for self-similar fractal laws, which have been shown by Mandelbrot and many others to be a general description of a large number of natural phenomena, in particular biological ones (see, e. g., [48,49,70], other volumes of these series and references therein). In addition, the obtained laws, which combine fractal and scale-independent behaviours, are the equivalent for scales of what inertial laws are for motion [49]. Since they serve as a fundamental basis of description for all the subsequent theoretical constructions, we shall now describe their derivation in detail.

The simplest differential equation of explicit scale dependence which one can write is of first order and states that the variation of $\mathcal{L}$ under an infinitesimal scale transformation $\mathrm{d}\ln\varepsilon$ depends only on $\mathcal{L}$ itself. Basing ourselves on the previous derivation of the form of the dilation operator, we thus write

$$\frac{\partial \mathcal{L}(s, \varepsilon)}{\partial \ln \varepsilon} = \beta(\mathcal{L})\,. \tag{3}$$

The function β is a priori unknown. However, still looking for the simplest form of such an equation, we expand $\beta(\mathcal{L})$ in powers of $\mathcal{L}$, namely we write $\beta(\mathcal{L}) = a + b\mathcal{L} + \ldots$. Disregarding for the moment the s dependence, we obtain, to the first order, the following linear equation in which a and b are constants:

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\ln\varepsilon} = a + b\mathcal{L}\,. \tag{4}$$

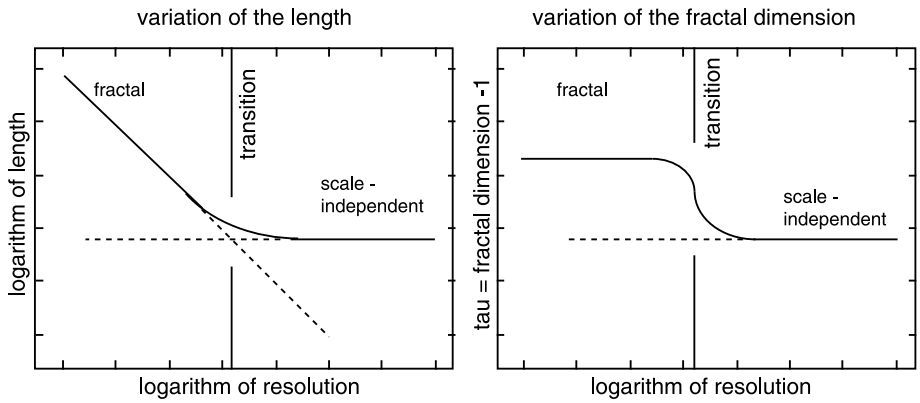

Fractals in the Quantum Theory of Spacetime, Figure 1
Scale dependence of the length (*left*) and of the effective fractal dimension $D_F = \tau_F + 1$ (*right*) in the case of "inertial" scale laws (which are solutions of the simplest, first order scale differential equation). Toward the small scale one gets a scale-invariant law with constant fractal dimension, while the explicit scale-dependence is lost at scales larger than a transition scale λ

In order to find the solution of this equation, let us change the names of the constants as $\tau_F = -b$ and $\mathcal{L}_0 = a/\tau_F$, so that $a + b\mathcal{L} = -\tau_F(\mathcal{L} - \mathcal{L}_0)$. We obtain the equation

$$\frac{d\mathcal{L}}{\mathcal{L} - \mathcal{L}_0} = -\tau_F \, d\ln \varepsilon \,. \tag{5}$$

Its solution is (see Fig. 1)

$$\mathcal{L}(\varepsilon) = \mathcal{L}_0 \left\{ 1 + \left(\frac{\lambda}{\varepsilon} \right)^{\tau_F} \right\} , \tag{6}$$

where λ is an integration constant. This solution corresponds to a length measured on a fractal curve up to a given point. One can now generalize it to a variable length that also depends on the position characterized by the parameter s. One obtains

$$\mathcal{L}(s, \varepsilon) = \mathcal{L}_0(s) \left\{ 1 + \zeta(s) \left(\frac{\lambda}{\varepsilon} \right)^{\tau_F} \right\} , \tag{7}$$

in which, in the most general case, the exponent τ_F may itself be a variable depending on the position.

The same kind of result is obtained for the projections on a given axis of such a fractal length [54]. Let $X(s, \varepsilon)$ be one of these projections, it reads

$$X(s, \varepsilon) = x(s) \left\{ 1 + \zeta_x(s) \left(\frac{\lambda}{\varepsilon} \right)^{\tau_F} \right\} . \tag{8}$$

In this case $\zeta_x(s)$ becomes a highly fluctuating function which may be described by a stochastic variable, as can be seen in Fig. 2.

The important point here and for what follows is that the solution obtained is the sum of two terms (a classical-like, "differentiable part" and a nondifferentiable "fractal part") which is explicitly scale-dependent and tends to infinity when $\varepsilon \to 0$ [22,54]. By differentiating these two parts in the above projection, we obtain the differential

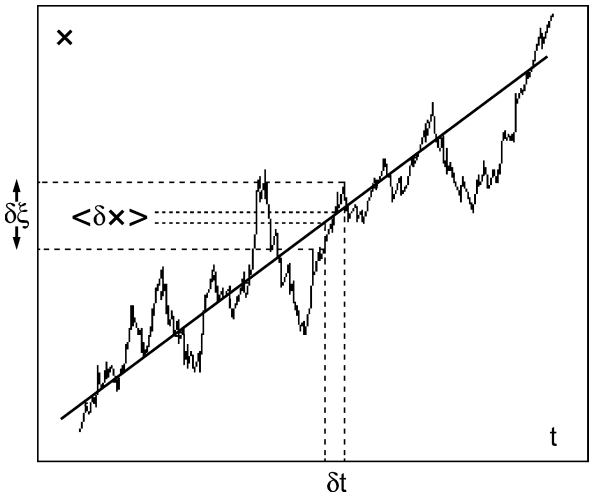

Fractals in the Quantum Theory of Spacetime, Figure 2
A fractal function. An example of such a fractal function is given by the projections of a fractal curve on Cartesian coordinates, as a function of a continuous and monotonous parameter (here the time t) which marks the position on the curve. The figure also exhibits the relation between space and time differential elements for such a fractal function, and compares the differentiable part δx and nondifferentiable part $\delta\xi$ of the space elementary displacement $\delta X = \delta x + \delta\xi$. While the differentiable coordinate variation $\delta x = \langle \delta X \rangle$ is of the same order as the time differential δt, the fractal fluctuation becomes much larger than δt when $\delta t \ll T$, where T is a transition time scale, and it depends on the fractal dimension D_F as $\delta\xi \propto \delta t^{1/D_F}$. Therefore the two contributions to the full differential displacement are related by the fractal law $\delta\xi^{D_F} \propto \delta x$, since δx and δt are differential elements of the same order

formulation of this essential result,

$$\mathrm{d}X = \mathrm{d}x + \mathrm{d}\xi \,, \tag{9}$$

where $\mathrm{d}x$ is a classical differential element, while $\mathrm{d}\xi$ is a differential element of fractional order (see Fig. 2, in which the parameter s that characterizes the position on the fractal curve has been taken to be the time t). This relation plays a fundamental role in the subsequent developments of the theory.

Consider the case when τ_F is constant. In the asymptotic small scale regime, $\varepsilon \ll \lambda$, one obtains a power-law dependence on resolution which reads

$$\mathcal{L}(s,\varepsilon) = \mathcal{L}_0(s)\left(\frac{\lambda}{\varepsilon}\right)^{\tau_\mathrm{F}} . \tag{10}$$

In this expression we recognize the standard form of a self-similar fractal behavior with constant fractal dimension $D_\mathrm{F} = 1 + \tau_\mathrm{F}$, which has already been found to yield a fair description of many physical and biological systems [49]. Here the topological dimension is $D_\mathrm{T} = 1$, since we deal with a length, but this can be easily generalized to surfaces ($D_\mathrm{T} = 2$), volumes ($D_\mathrm{T} = 3$), etc., according to the general relation $D_\mathrm{F} = D_\mathrm{T} + \tau_\mathrm{F}$. The new feature here is that this result has been derived from a theoretical analysis based on first principles, instead of being postulated or deduced from a fit of observational data.

It should be noted that in the above expressions, the resolution is a length interval, $\varepsilon = \delta X$ defined along the fractal curve (or one of its projected coordinate). But one may also travel on the curve and measure its length on constant time intervals, then change the time scale. In this case the resolution ε is a time interval, $\varepsilon = \delta t$. Since they are related by the fundamental relation (see Fig. 2)

$$\delta X^{D_\mathrm{F}} \sim \delta t \,, \tag{11}$$

the fractal length depends on the time resolution as

$$X(s,\delta t) = X_0(s) \times \left(\frac{T}{\delta t}\right)^{1-1/D_\mathrm{F}} . \tag{12}$$

An example of the use of such a relation is Feynman's result according to which the mean square value of the velocity of a quantum mechanical particle is proportional to δt^{-1} (see p. 176 in [30]), which corresponds to a fractal dimension $D_\mathrm{F} = 2$, as later recovered by Abbott and Wise [1] by using a space resolution.

More generally (in the usual case when $\varepsilon = \delta X$), following Mandelbrot, the scale exponent $\tau_\mathrm{F} = D_\mathrm{F} - D_\mathrm{T}$ can be defined as the slope of the $(\ln\varepsilon, \ln\mathcal{L})$ curve, namely

$$\tau_\mathrm{F} = \frac{\mathrm{d}\ln\mathcal{L}}{\mathrm{d}\ln(\lambda/\varepsilon)} . \tag{13}$$

For a self-similar fractal such as that described by the fractal part of the above solution, this definition yields a constant value which is the exponent in Eq. (10). However, one can anticipate on the following, and use this definition to compute an "effective" or "local" fractal dimension, now variable, from the complete solution that includes the differentiable and the nondifferentiable parts, and therefore a transition to effective scale independence. Differentiating the logarithm of Eq. (16) yields an effective exponent given by

$$\tau_\mathrm{eff} = \frac{\tau_\mathrm{F}}{1 + (\varepsilon/\lambda)^{\tau_\mathrm{F}}} . \tag{14}$$

The effective fractal dimension $D_\mathrm{F} = 1 + \tau_\mathrm{F}$ therefore jumps from the nonfractal value $D_\mathrm{F} = D_\mathrm{T} = 1$ to its constant asymptotic value at the transition scale λ (see right part of Fig. 1).

Galilean Relativity of Scales

We can now check that the fractal part of such a law is compatible with the principle of relativity extended to scale transformations of the resolutions (i. e., with the principle of scale relativity). It reads $\mathcal{L} = \mathcal{L}_0(\lambda/\varepsilon)^{\tau_\mathrm{F}}$ (Eq. 10), and it is therefore a law involving two variables ($\ln\mathcal{L}$ and τ_F) as a function of one parameter (ε) which, according to the relativistic view, characterizes the state of scale of the system (its relativity is apparent in the fact that we need another scale λ to define it by their ratio). More generally, all the following statements remain true for the complete scale law including the transition to scale-independence, by making the replacement of $\mathcal{L}$ by $\mathcal{L} - \mathcal{L}_0$. Note that, to be complete, we anticipate on what follows and consider a priori τ_F to be a variable, even if, in the simple law first considered here, it takes a constant value.

Let us take the logarithm of Eq. (10). It yields $\ln(\mathcal{L}/\mathcal{L}_0) = \tau_\mathrm{F}\ln(\lambda/\varepsilon)$. The two quantities $\ln\mathcal{L}$ and τ_F then transform under a finite scale transformation $\varepsilon \to \varepsilon' = \rho\,\varepsilon$ as

$$\ln\frac{\mathcal{L}(\varepsilon')}{\mathcal{L}_0} = \ln\frac{\mathcal{L}(\varepsilon)}{\mathcal{L}_0} - \tau_\mathrm{F}\ln\rho \,, \tag{15}$$

and to be complete,

$$\tau'_\mathrm{F} = \tau_\mathrm{F} \,. \tag{16}$$

These transformations have exactly the same mathematical structure as the Galilean group of motion transformation (applied here to scale rather than motion), which reads

$$x' = x - t\,v, \quad t' = t \,. \tag{17}$$

This is confirmed by the dilation composition law, $\varepsilon \to \varepsilon' \to \varepsilon''$, which reads

$$\ln \frac{\varepsilon''}{\varepsilon} = \ln \frac{\varepsilon'}{\varepsilon} + \ln \frac{\varepsilon''}{\varepsilon'} , \tag{18}$$

and is therefore similar to the law of composition of velocities between three reference systems K, K' and K'',

$$V''(K''/K) = V(K'/K) + V'(K''/K') . \tag{19}$$

Since the Galileo group of motion transformations is known to be the simplest group that implements the principle of relativity, the same is true for scale transformations.

It is important to realize that this is more than a simple analogy: the same physical problem is set in both cases, and is therefore solved under similar mathematical structures (since the logarithm transforms what would have been a multiplicative group into an additive group). Indeed, in both cases, it is equivalent to finding the transformation law of a position variable (X for motion in a Cartesian system of coordinates, $\ln \mathcal{L}$ for scales in a fractal system of coordinates) under a change of the state of the coordinate system (change of velocity V for motion and of resolution $\ln \rho$ for scale), knowing that these state variables are defined only in a relative way. Namely, V is the relative velocity between the reference systems K and K', and ρ is the relative scale: note that ε and ε' have indeed disappeared in the transformation law, only their ratio remains. This remark establishes the status of resolutions as (relative) "scale velocities" and of the scale exponent τ_F as a "scale time".

Recall finally that since the Galilean group of motion is only a limiting case of the more general Lorentz group, a similar generalization is expected in the case of scale transformations, which we shall briefly consider in Sect. "Special Scale-Relativity".

Breaking of Scale Invariance

The standard self-similar fractal laws can be derived from the scale relativity approach. However, it is important to note that Eq. (16) provides us with another fundamental result, as shown in Fig. 1. Namely, it also contains a spontaneous breaking of the scale symmetry. Indeed, it is characterized by the existence of a transition from a fractal to a non-fractal behavior at scales larger than some transition scale λ. The existence of such a breaking of scale invariance is also a fundamental feature of many natural systems, which remains, in most cases, misunderstood.

The advantage of the way it is derived here is that it appears as a natural, spontaneous, but only effective symmetry breaking, since it does not affect the underlying scale symmetry. Indeed, the obtained solution is the sum of two terms, the scale-independent contribution (differentiable part), and the explicitly scale-dependent and divergent contribution (fractal part). At large scales the scaling part becomes dominated by the classical part, but it is still underlying even though it is hidden. There is therefore an apparent symmetry breaking (see Fig. 1), though the underlying scale symmetry actually remains unbroken.

The origin of this transition is, once again, to be found in relativity (namely, in the relativity of position and motion). Indeed, if one starts from a strictly scale-invariant law without any transition, $\mathcal{L} = \mathcal{L}_0(\lambda/\varepsilon)^{\tau_F}$, then adds a translation in standard position space ($\mathcal{L} \to \mathcal{L} + \mathcal{L}_1$), one obtains

$$\mathcal{L}' = \mathcal{L}_1 + \mathcal{L}_0 \left(\frac{\lambda}{\varepsilon}\right)^{\tau_F} = \mathcal{L}_1 \left\{1 + \left(\frac{\lambda_1}{\varepsilon}\right)^{\tau_F}\right\} . \tag{20}$$

Therefore one recovers the broken solution (that corresponds to the constant $a \neq 0$ in the initial scale differential equation). This solution is now asymptotically scale-dependent (in a scale-invariant way) only at small scales, and becomes independent of scale at large scales, beyond some relative transition λ_1 which is partly determined by the translation itself.

Multiple Scale Transitions

Multiple transitions can be obtained by a simple generalization of the above result [58]. Still considering a perturbative approach and taking the Taylor expansion of the differential equation $\mathrm{d}\mathcal{L}/\mathrm{d}\ln\varepsilon = \beta(\mathcal{L})$, but now to the second order of the expansion, one obtains the equation

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\ln\varepsilon} = a + b\mathcal{L} + c\mathcal{L}^2 + \cdots . \tag{21}$$

One of its solutions, which generalizes that of Eq. (5), describes a scaling behavior which is broken toward both the small and large scales, as observed in most real fractal systems,

$$\mathcal{L} = \mathcal{L}_0 \left(\frac{1 + (\lambda_0/\varepsilon)^{\tau_F}}{1 + (\lambda_1/\varepsilon)^{\tau_F}}\right) . \tag{22}$$

Due to the non-linearity of the β function, there are now two transition scales in such a law. Indeed,

*when $\varepsilon < \lambda_1 < \lambda_0$, one has $(\lambda_0/\varepsilon) \gg 1$ and $(\lambda_1/\varepsilon) \gg 1$, so that $\mathcal{L} = \mathcal{L}_0(\lambda_0/\lambda_1)^{\tau_F} \approx$ cst, independent of scale;

*when $\lambda_1 < \varepsilon < \lambda_0$, one has $(\lambda_0/\varepsilon) \gg 1$ but $(\lambda_1/\varepsilon) \ll 1$, so that the denominator disappears, and one recovers the previous pure scaling law $\mathcal{L} = \mathcal{L}_0\ (\lambda_0/\varepsilon)^{\tau_F}$;

*when $\lambda_1 < \lambda_0 < \varepsilon$, one has $(\lambda_0/\varepsilon) \ll 1$ and $(\lambda_1/\varepsilon) \ll 1$, so that $\mathcal{L} = \mathcal{L}_0 =$ cst, independent of scale.

Scale Relativity Versus Scale Invariance

Let us briefly be more specific about the way in which the scale relativity viewpoint differs from scaling or simple scale invariance. In the standard concept of scale invariance, one considers scale transformations of the coordinate,

$$X \to X' = q \times X \,, \tag{23}$$

then one looks for the effect of such a transformation on some function $f(X)$. It is scaling when

$$f(qX) = q^{\alpha} \times f(X) \,. \tag{24}$$

The scale relativity approach involves a more profound level of description, since the coordinate X is now explicitly resolution-dependent, i. e. $X = X(\varepsilon)$. Therefore we now look for a scale transformation of the resolution,

$$\varepsilon \to \varepsilon' = \rho\, \varepsilon \,, \tag{25}$$

which implies a scale transformation of the position variable that takes in the self-similar case the form

$$X(\rho\, \varepsilon) = \rho^{-\tau_{\mathrm{F}}} X(\varepsilon) \,. \tag{26}$$

But now the scale factor on the variable has a physical meaning which goes beyond a trivial change of units. It corresponds to a coordinate measured at two different resolutions on a fractal curve of fractal dimension $D_{\mathrm{F}} = 1 + \tau_{\mathrm{F}}$, and one can obtain a scaling function of a fractal coordinate,

$$f(\rho^{-\tau_{\mathrm{F}}} X) = \rho^{-\alpha \tau_{\mathrm{F}}} \times f(X) \,. \tag{27}$$

In other words, there are now three levels of transformation in the scale relativity framework (the resolution, the variable, and its function) instead of only two in the usual conception of scaling.

Generalized Scale Laws

Discrete Scale Invariance, Complex Dimension, and Log-Periodic Behavior Fluctuations with respect to pure scale invariance are potentially important, namely the log-periodic correction to power laws that is provided, e. g., by complex exponents or complex fractal dimensions. It has been shown that such a behavior provides a very satisfactory and possibly predictive model of the time evolution of many critical systems, including earthquakes and market crashes ([79] and references therein). More recently, it has been applied to the analysis of major event chronology of the evolutionary tree of life [14,65,66], of human development [11] and of the main economic crisis of western and precolumbian civilizations [36,37,40,65].

One can recover log-periodic corrections to self-similar power laws through the requirement of covariance (i. e., of form invariance of equations) applied to scale differential equations [58]. Consider a scale-dependent function $\mathcal{L}(\varepsilon)$. In the applications to temporal evolution quoted above, the scale variable is identified with the time interval $|t - t_c|$, where t_c is the date of a crisis. Assume that $\mathcal{L}$ satisfies a first order differential equation,

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\ln\varepsilon} - \nu\mathcal{L} = 0 \,, \tag{28}$$

whose solution is a pure power law $\mathcal{L}(\varepsilon) \propto \varepsilon^{\nu}$ (cf Sect. "Self-Similar Fractals as Solutions of a First Order Scale Differential Equation"). Now looking for corrections to this law, one remarks that simply incorporating a complex value of the exponent ν would lead to large log-periodic fluctuations rather than to a controllable correction to the power law. So let us assume that the right-hand side of Eq. (28) actually differs from zero:

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\ln\varepsilon} - \nu\mathcal{L} = \chi \,. \tag{29}$$

We can now apply the scale covariance principle and require that the new function χ be a solution of an equation which keeps the same form as the initial equation,

$$\frac{\mathrm{d}\chi}{\mathrm{d}\ln\varepsilon} - \nu'\chi = 0 \,. \tag{30}$$

Setting $\nu' = \nu + \eta$, we find that $\mathcal{L}$ must be a solution of a second-order equation:

$$\frac{\mathrm{d}^2\mathcal{L}}{(\mathrm{d}\ln\varepsilon)^2} - (2\nu + \eta)\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\ln\varepsilon} + \nu(\nu + \eta)\mathcal{L} = 0 \,. \tag{31}$$

The solution reads $\mathcal{L}(\varepsilon) = a\varepsilon^{\nu}(1 + b\varepsilon^{\eta})$, and finally, the choice of an imaginary exponent $\eta = i\omega$ yields a solution whose real part includes a log-periodic correction:

$$\mathcal{L}(\varepsilon) = a\varepsilon^{\nu}\,[1 + b\cos(\omega\ln\varepsilon)] \,. \tag{32}$$

As previously recalled in Sect. "Breaking of Scale Invariance," adding a constant term (a translation) provides a transition to scale independence at large scales (see Fig. 3).

Lagrangian Approach to Scale Laws In order to obtain physically relevant generalizations of the above simplest (scale-invariant) laws, a Lagrangian approach can be used in scale space analogously to using it to derive the laws of

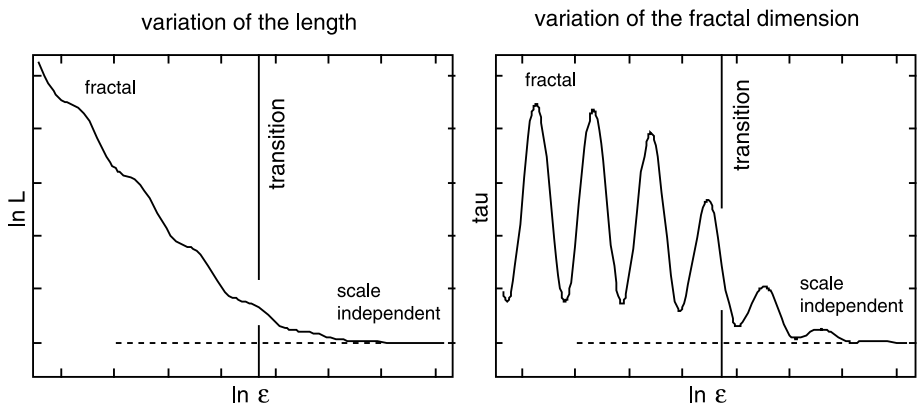

Fractals in the Quantum Theory of Spacetime, Figure 3
Scale dependence of the length $\mathcal{L}$ of a fractal curve and of the effective "scale time" $\tau_F = D_F - D_T$ (fractal dimension minus topological dimension) in the case of a log-periodic behavior with fractal/non-fractal transition at scale λ, which reads $\mathcal{L}(\varepsilon) = \mathcal{L}_0\{1 + (\lambda/\varepsilon)^\nu \exp[b\cos(\omega \ln(\varepsilon/\lambda))]\}$

motion, leading to reversal of the definition and meaning of the variables [58].

This reversal is analogous to that achieved by Galileo concerning the laws of motion. Indeed, from the Aristotle viewpoint, "time is the measure of motion". In the same way, the fractal dimension, in its standard (Mandelbrot's) acception, is defined from the topological measure of the fractal object (length of a curve, area of a surface, etc.) and resolution, namely (see Eq. 13)

$$t = \frac{x}{v} \quad \leftrightarrow \quad \tau_F = D_F - D_T = \frac{d\ln\mathcal{L}}{d\ln(\lambda/\varepsilon)} \,. \tag{33}$$

In the case mainly considered here, when $\mathcal{L}$ represents a length (i. e., more generally, a fractal coordinate), the topological dimension is $D_T = 1$ so that $\tau_F = D_F - 1$. With Galileo, time becomes a primary variable, and the velocity is deduced from space and time, which are therefore treated on the same footing, in terms of a space-time (even though the Galilean space-time remains degenerate because of the implicitly assumed infinite velocity of light).

In analogy, the scale exponent $\tau_F = D_F - 1$ becomes in this new representation a primary variable that plays, for scale laws, the same role as played by time in motion laws (it is called "djinn" in some publications which therefore introduce a five-dimensional 'space-time-djinn' combining the four fractal fluctuations and the scale time).

Carrying on the analogy, in the same way that the velocity is the derivative of position with respect to time, $v = dx/dt$, we expect the derivative of $\ln\mathcal{L}$ with respect to scale time τ_F to be a "scale velocity". Consider as reference the self-similar case, that reads $\ln\mathcal{L} = \tau_F \ln(\lambda/\varepsilon)$. Deriving with respect to τ_F, now considered as a variable, yields $d\ln\mathcal{L}/d\tau_F = \ln(\lambda/\varepsilon)$, i. e., the logarithm of resolution. By extension, one assumes that this scale velocity provides a new general definition of resolution even in more general situations, namely,

$$\mathbb{V} = \ln\left(\frac{\lambda}{\varepsilon}\right) = \frac{d\ln\mathcal{L}}{d\tau_F} \,. \tag{34}$$

One can now introduce a scale Lagrange function $\widetilde{L}(\ln\mathcal{L}, \mathbb{V}, \tau_F)$, from which a scale action is constructed:

$$\widetilde{S} = \int_{\tau_1}^{\tau_2} \widetilde{L}(\ln\mathcal{L}, \mathbb{V}, \tau_F) d\tau_F \,. \tag{35}$$

The application of the action principle yields a scale Euler–Lagrange equation which reads

$$\frac{d}{d\tau_F}\frac{\partial\widetilde{L}}{\partial\mathbb{V}} = \frac{\partial\widetilde{L}}{\partial\ln\mathcal{L}} \,. \tag{36}$$

One can now verify that in the free case, i. e., in the absence of any "scale force" (i. e., $\partial\tilde{L}/\partial\ln\mathcal{L} = 0$), one recovers the standard fractal laws derived hereabove. Indeed, in this case the Euler–Lagrange equation becomes

$$\partial\tilde{L}/\partial\mathbb{V} = \text{const} \quad \Rightarrow \quad \mathbb{V} = \text{const.} \tag{37}$$

which is the equivalent for scale of what inertia is for motion. Still in analogy with motion laws, the simplest possible form for the Lagrange function is a quadratic dependence on the scale velocity, (i. e., $\tilde{L} \propto \mathbb{V}^2$). The constancy of $\mathbb{V} = \ln(\lambda/\varepsilon)$ means that it is independent of the scale time τ_F. Equation (34) can therefore be integrated to give the usual power law behavior, $\mathcal{L} = \mathcal{L}_0(\lambda/\varepsilon)^{\tau_F}$, as expected.

But this reversed viewpoint also has several advantages which allow a full implementation of the principle of scale relativity:

(i) The scale time τ_F is given the status of a fifth dimension and the logarithm of the resolution $\mathbb{V} = \ln(\lambda/\varepsilon)$, is given the status of a scale velocity (see Eq. 34). This is in accordance with its scale-relativistic definition, in which it characterizes the state of scale of the reference system, in the same way as the velocity $v = \mathrm{d}x/\mathrm{d}t$ characterizes its state of motion.

(ii) This allows one to generalize the formalism to the case of four independent space-time resolutions, $\mathbb{V}^\mu = \ln(\lambda^\mu/\varepsilon^\mu) = \mathrm{d}\ln\mathcal{L}^\mu/\mathrm{d}\tau_F$.

(iii) Scale laws more general than the simplest self-similar ones can be derived from more general scale Lagrangians [57,58] involving "scale accelerations" $\mathbb{\Gamma} = \mathrm{d}^2\ln\mathcal{L}/\mathrm{d}\tau_F^2 = \mathrm{d}\ln(\lambda/\varepsilon)/\mathrm{d}\tau_F$, as we shall see in what follows.

Note however that there is also a shortcoming in this approach. Contrary to the case of motion laws, in which time is always flowing toward the future (except possibly in elementary particle physics at very small time scales), the variation of the scale time may be non-monotonic, as exemplified by the previous case of log-periodicity. Therefore this Lagrangian approach is restricted to monotonous variations of the fractal dimension, or, more generally, to scale intervals on which it varies in a monotonous way.

Scale Dynamics The previous discussion indicates that the scale invariant behavior corresponds to freedom (i. e. scale force-free behavior) in the framework of a scale physics. However, in the same way as there are forces in nature that imply departure from inertial, rectilinear uniform motion, we expect most natural fractal systems to also present distortions in their scale behavior with respect to pure scale invariance. This implies taking non-linearity in the scale space into account. Such distorsions may be, as a first step, attributed to the effect of the dynamics of scale ("scale dynamics"), i. e., of a "scale field", but it must be clear from the very beginning of the description that they are of geometric nature (in analogy with the Newtonian interpretation of gravitation as the result of a force, which has later been understood from Einstein's general relativity theory as a manifestation of the curved geometry of space-time).

In this case the Lagrange scale-equation takes the form of Newton's equation of dynamics,

$$F = \mu\,\frac{\mathrm{d}^2\ln\mathcal{L}}{\mathrm{d}\tau_F^2}\,, \tag{38}$$

where μ is a "scale mass", which measures how the system resists to the scale force, and where $\mathbb{\Gamma} = \mathrm{d}^2\ln\mathcal{L}/\mathrm{d}\tau_F^2 = \mathrm{d}\ln(\lambda/\varepsilon)/\mathrm{d}\tau_F$ is the scale acceleration.

In this framework one can therefore attempt to define generic, scale-dynamical behaviours which could be common to very different systems, as corresponding to a given form of the scale force.

Constant Scale Force A typical example is the case of a constant scale force. Setting $G = F/\mu$, the potential reads $\varphi = G\ln\mathcal{L}$, analogous to the potential of a constant force f in space, which is $\varphi = -fx$, since the force is $-\partial\varphi/\partial x = f$. The scale differential equation is

$$\frac{\mathrm{d}^2\ln\mathcal{L}}{\mathrm{d}\tau_F^2} = G\,. \tag{39}$$

It can be easily integrated. A first integration yields $\mathrm{d}\ln\mathcal{L}/\mathrm{d}\tau_F = G\tau_F + \mathbb{V}_0$, where $\mathbb{V}_0$ is a constant. Then a second integration yields a parabolic solution (which is the equivalent for scale laws of parabolic motion in a constant field),

$$\mathbb{V} = \mathbb{V}_0 + G\tau_F\,;\;\; \ln\mathcal{L} = \ln\mathcal{L}_0 + \mathbb{V}_0\tau_F + \frac{1}{2}G\,\tau_F^2\,, \tag{40}$$

where $\mathbb{V} = \mathrm{d}\ln\mathcal{L}/\mathrm{d}\tau_F = \ln(\lambda/\varepsilon)$.

However the physical meaning of this result is not clear under this form. This is due to the fact that, while in the case of motion laws we search for the evolution of the system with time, in the case of scale laws we search for the dependence of the system on resolution, which is the directly measured observable. Since the reference scale λ is arbitrary, the variables can be re-defined in such a way that $\mathbb{V}_0 = 0$, i. e., $\lambda = \lambda_0$. Indeed, from Eq. (40) one gets $\tau_F = (\mathbb{V} - \mathbb{V}_0)/G = [\ln(\lambda/\varepsilon) - \ln(\lambda/\lambda_0)]/G = \ln(\lambda_0/\varepsilon)/G$. Then one obtains

$$\tau_F = \frac{1}{G}\ln\left(\frac{\lambda_0}{\varepsilon}\right), \qquad \ln\left(\frac{\mathcal{L}}{\mathcal{L}_0}\right) = \frac{1}{2G}\ln^2\left(\frac{\lambda_0}{\varepsilon}\right)\,. \tag{41}$$

The scale time τ_F becomes a linear function of resolution (the same being true, as a consequence, of the fractal dimension $D_F = 1 + \tau_F$), and the $(\ln\mathcal{L}, \ln\varepsilon)$ relation is now parabolic instead of linear (see Fig. 4 and compare to Fig. 1). Note that, as in previous cases, we have considered here only the small scale asymptotic behavior, and that we can once again easily generalize this result by including a transition to scale-independence at large scale. This is simply achieved by replacing $\mathcal{L}$ by $(\mathcal{L} - \mathcal{L}_0)$ in every equations.

There are several physical situations where, after careful examination of the data, the power-law models were clearly rejected since no constant slope could be defined in

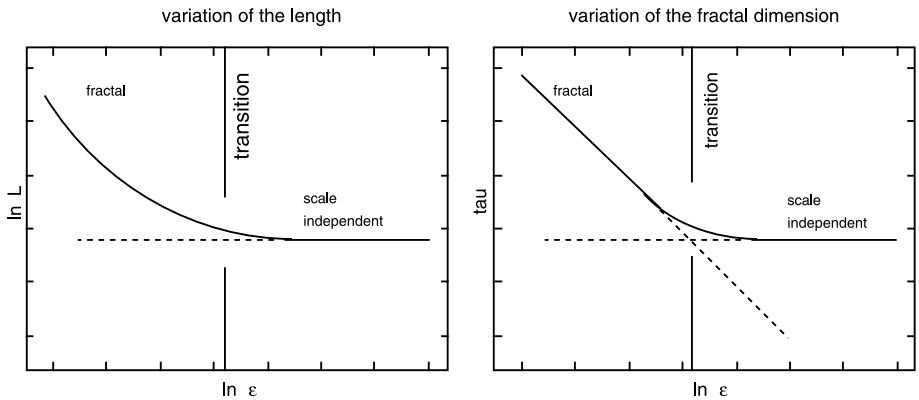

Fractals in the Quantum Theory of Spacetime, Figure 4
Scale dependence of the length of a fractal curve ln $\mathcal{L}$ and of its effective fractal dimension ($D_F = D_T + \tau_F$, where D_T is the topological dimension) in the case of a constant scale force, with an additional fractal to non-fractal transition

the $(\log \mathcal{L}, \log \varepsilon)$ plane. In the several cases where a clear curvature appears in this plane, e. g., turbulence [26], sandpiles [9], fractured surfaces in solid mechanics [10], the physics could come under such a scale-dynamical description. In these cases it might be of interest to identify and study the scale force responsible for the scale distortion (i. e., for the deviation from standard scaling).

Special Scale-Relativity

Let us close this section about the derivation of scale laws of increasing complexity by coming back to the question of finding the general laws of scale transformations that meet the principle of scale relativity [53]. It has been shown in Sect. "Galilean Relativity of Scales" that the standard self-similar fractal laws come under a Galilean group of scale transformations. However, the Galilean relativity group is known, for motion laws, to be only a degenerate form of the Lorentz group. It has been proved that a similar result holds for scale laws [53,54].

The problem of finding the linear transformation laws of fields in a scale transformation $\mathbb{V} = \ln \rho$ $(\varepsilon \to \varepsilon')$ amounts to finding four quantities, $a(\mathbb{V})$, $b(\mathbb{V})$, $c(\mathbb{V})$, and $d(\mathbb{V})$, such that

$$\begin{aligned} \ln \frac{\mathcal{L}'}{\mathcal{L}_0} &= a(\mathbb{V}) \ln \frac{\mathcal{L}}{\mathcal{L}_0} + b(\mathbb{V})\, \tau_F \,, \\ \tau_F{}' &= c(\mathbb{V}) \ln \frac{\mathcal{L}}{\mathcal{L}_0} + d(\mathbb{V}) \tau_F \,. \end{aligned} \tag{42}$$

Set in this way, it immediately appears that the current 'scale-invariant' scale transformation law of the standard form (Eq. 8), given by $a = 1, b = \mathbb{V}, c = 0$ and $d = 1$, corresponds to a Galilean group.

This is also clear from the law of composition of dilatations, $\varepsilon \to \varepsilon' \to \varepsilon''$, which has a simple additive form,

$$\mathbb{V}'' = \mathbb{V} + \mathbb{V}' \,. \tag{43}$$

However the general solution to the 'special relativity problem' (namely, find a, b, c and d from the principle of relativity) is the Lorentz group [47,53]. This result has led to the suggestion of replacing the standard law of dilatation, $\varepsilon \to \varepsilon' = \varrho \times \varepsilon$ by a new Lorentzian relation, namely, for $\varepsilon < \lambda_0$ and $\varepsilon' < \lambda_0$

$$\ln \frac{\varepsilon'}{\lambda_0} = \frac{\ln(\varepsilon/\lambda_0) + \ln \varrho}{1 + \ln \varrho \, \ln(\varepsilon/\lambda_0)/\ln^2(\boldsymbol{\Lambda}/\lambda_0)} \,. \tag{44}$$

This relation introduces a fundamental length scale $\boldsymbol{\Lambda}$, which is naturally identified, towards the small scales, with the Planck length (currently $1.6160(11) \times 10^{-35}$ m) [53],

$$\boldsymbol{\Lambda} = l_{\mathbb{P}} = (\hbar G/c^3)^{1/2} \,, \tag{45}$$

and toward the large scales (for $\varepsilon > \lambda_0$ and $\varepsilon' > \lambda_0$) with the scale of the cosmological constant, $\mathbb{L} = \Lambda^{-1/2}$ (see Chap. 7.1 in [54]).

As one can see from Eq. (44), if one starts from the scale $\varepsilon = \boldsymbol{\Lambda}$ and apply any dilatation or contraction ϱ, one obtains again the scale $\varepsilon' = \boldsymbol{\Lambda}$, whatever the initial value of λ_0. In other words, $\boldsymbol{\Lambda}$ can be interpreted as a limiting lower (or upper) length-scale, which is impassable and invariant under dilatations and contractions.

Concerning the length measured along a fractal coordinate which was previously scale-dependent as $\ln(\mathcal{L}/\mathcal{L}_0) = \tau_0 \ln(\lambda_0/\varepsilon)$ for $\varepsilon < \lambda_0$, it becomes in the new framework, in the simplified case when one starts from the

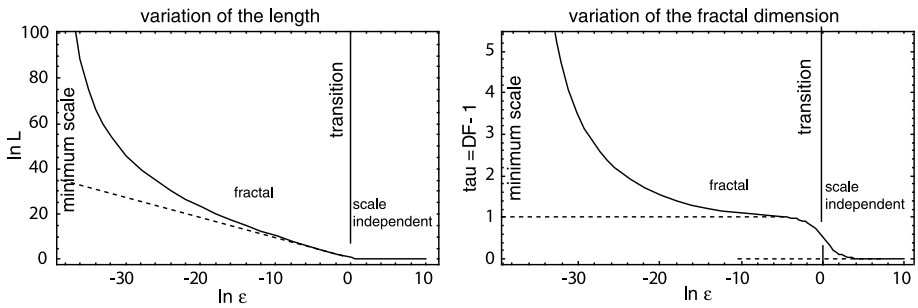

Fractals in the Quantum Theory of Spacetime, Figure 5

Scale dependence of the logarithm of the length and of the effective fractal dimension, $D_F = 1 + \tau_F$, in the case of scale-relativistic Lorentzian scale laws including a transition to scale independence toward large scales. The constant $\mathbb{C}$ has been taken to be $\mathbb{C} = 4\pi^2 \approx 39.478$, which is a fundamental value for scale ratios in elementary particle physics in the scale relativity framework [53,54,63], while the effective fractal dimension jumps from $D_F = 1$ to $D_F = 2$ at the transition, then increases without any limit toward small scales

reference scale $\mathcal{L}_0$ (see Fig. 5)

$$\ln \frac{\mathcal{L}}{\mathcal{L}_0} = \frac{\tau_0 \ln(\lambda_0/\varepsilon)}{\sqrt{1 - \ln^2(\lambda_0/\varepsilon)/\ln^2(\lambda_0/\boldsymbol{\Lambda})}} \,. \tag{46}$$

The main new feature of scale relativity with respect to the previous fractal or scale-invariant approaches is that the scale exponent τ_F and the fractal dimension $D_F = 1 + \tau_F$, which were previously constant ($D_F = 2$, $\tau_F = 1$), are now explicitly varying with scale (see Fig. 5), following the law (given once again in the simplified case when we start from the reference scale $\mathcal{L}_0$):

$$\tau_F(\varepsilon) = \frac{\tau_0}{\sqrt{1 - \ln^2(\lambda_0/\varepsilon)/\ln^2(\lambda_0/\boldsymbol{\Lambda})}} \,. \tag{47}$$

Under this form, the scale covariance is explicit, since one keeps a power law form for the length variation, $\mathcal{L} = \mathcal{L}_0(\lambda/\varepsilon)^{\tau_F(\varepsilon)}$, but now in terms of a variable fractal dimension.

For a more complete development of special relativity, including its implications regarding new conservative quantities and applications in elementary particle physics and cosmology, see [53,54,56,63].

The question of the nature of space-time geometry at the Planck scale is a subject of intense work (see, e. g., [3,46] and references therein). This is a central question for practically all theoretical attempts, including noncommutative geometry [15,16], supersymmetry, and superstring theories [35,75] – for which the compactification scale is close to the Planck scale – and particularly for the theory of quantum gravity. Indeed, the development of loop quantum gravity by Rovelli and Smolin [76] led to the conclusion that the Planck scale could be a quantized minimal scale in Nature, involving also a quantization of surfaces and volumes [77].

Over the last years, there has also been significant research effort aimed at the development of a 'Doubly-Special-Relativity' [2] (see a review in [3]), according to which the laws of physics involve a fundamental velocity scale c and a fundamental minimum length scale L_p, identified with the Planck length.

The concept of a new relativity in which the Planck length-scale would become a minimum invariant length is exactly the founding idea of the special scale relativity theory [53], which has been incorporated in other attempts of extended relativity theories [12,13]. But, despite the similarity of aim and analysis, the main difference between the 'Doubly-Special-Relativity' approach and the scale relativity one is that the question of defining an invariant length-scale is considered in the scale relativity/fractal space-time theory as coming under a relativity of scales. Therefore the new group to be constructed is a multiplicative group, that becomes additive only when working with the logarithms of scale ratios, which are definitely the physically relevant scale variables, as one can show by applying the Gell-Mann-Levy method to the construction of the dilation operator (see Sect. "Fractal Coordinate and Differential Dilation Operator").

From Fractal Space to Nonrelativistic Quantum Mechanics

The first step in the construction of a theory of the quantum space-time from fractal and nondifferentiable geometry, which has been described in the previous sections, has consisted of finding the laws of explicit scale dependence at a given "point" or "instant" (under their new fractal definition).

The next step, which will now be considered, amounts to writing the equation of motion in such a fractal space(-

time) in terms of a geodesic equation. As we shall see, after integration this equation takes the form of a Schrödinger equation (and of the Klein-Gordon and Dirac equations in the relativistic case). This result, first obtained in Ref. [54], has later been confirmed by many subsequent physical [22,25,56,57] and mathematical works, in particular by Cresson and Ben Adda [6,7,17,19] and Jumarie [41,42,43,44], including attempts of generalizations using the tool of the fractional integro-differential calculus [7,21,44].

In what follows, we consider only the simplest case of fractal laws, namely, those characterized by a constant fractal dimension. The various generalized scale laws considered in the previous section lead to new possible generalizations of quantum mechanics [56,63].

Critical Fractal Dimension 2

Moreover, we simplify again the description by considering only the case $D_F = 2$. Indeed, the nondifferentiability and fractality of space implies that the paths are random walks of the Markovian type, which corresponds to such a fractal dimension. This choice is also justified by Feynman's result [30], according to which the typical paths of quantum particles (those which contribute mainly to the path integral) are nondifferentiable and of fractal dimension $D_F = 2$ [1]. The case $D_F \neq 2$, which yields generalizations to standard quantum mechanics has also been studied in detail (see [56,63] and references therein). This study shows that $D_F = 2$ plays a critical role in the theory, since it suppresses the explicit scale dependence in the motion (Schrödinger) equation – but this dependence remains hidden and reappears through, e. g., the Heisenberg relations and the explicit dependence of measurement results on the resolution of the measurement apparatus.

Let us start from the result of the previous section, according to which the solution of a first order scale differential equation reads for $D_F = 2$, after differentiation and reintroduction of the indices,

$$\mathrm{d}X^\mu = \mathrm{d}x^\mu + \mathrm{d}\xi^\mu = v^\mu \mathrm{d}s + \zeta^\mu \sqrt{\lambda_c\, \mathrm{d}s}\,, \tag{48}$$

where λ_c is a length scale which must be introduced for dimensional reasons and which, as we shall see, generalizes the Compton length. The ζ^μ are dimensionless highly fluctuating functions. Due to their highly erratic character, we can replace them by stochastic variables such that $\langle\zeta^\mu\rangle = 0$, $\langle(\zeta^0)^2\rangle = -1$ and $\langle(\zeta^k)^2\rangle = 1$ (k =1 to 3). The mean is taken here on a purely mathematic probability law which can be fully general, since the final result does not depend on its choice.

Metric of a Fractal Space-Time

Now one can also write the fractal fluctuations in terms of the coordinate differentials, $\mathrm{d}\xi^\mu = \zeta^\mu \sqrt{\lambda^\mu\, \mathrm{d}x^\mu}$. The identification of this expression with that of Eq. (3) leads one to recover the Einstein-de Broglie length and time scales,

$$\lambda_x = \frac{\lambda_c}{\mathrm{d}x/\mathrm{d}s} = \frac{\hbar}{p_x}\,, \quad \tau = \frac{\lambda_c}{\mathrm{d}t/\mathrm{d}s} = \frac{\hbar}{E}\,. \tag{49}$$

Let us now assume that the large scale (classical) behavior is given by the Riemannian metric potentials $g_{\mu\nu}(x, y, z, t)$. The invariant proper time $\mathrm{d}S$ along a geodesic in terms of the complete differential elements $\mathrm{d}X^\mu = \mathrm{d}x^\mu + \mathrm{d}\xi^\mu$

$$\mathrm{d}S^2 = g_{\mu\nu}\mathrm{d}X^\mu \mathrm{d}X^\nu = g_{\mu\nu}(\mathrm{d}x^\mu + \mathrm{d}\xi^\mu)(\mathrm{d}x^\nu + \mathrm{d}\xi^\nu). \tag{50}$$

Now replacing the $\mathrm{d}\xi$'s by their expression, one obtains a fractal metric [54,68]. Its two-dimensional and diagonal expression, neglecting the terms of the zero mean (in order to simplify its writing) reads

$$\begin{aligned}\mathrm{d}S^2 =& g_{00}(x, t)\left(1 + \zeta_0^2\, \frac{\tau_F}{\mathrm{d}t}\right) c^2 \mathrm{d}t^2 \\ &- g_{11}(x, t)\left(1 + \zeta_1^2\, \frac{\lambda_x}{\mathrm{d}x}\right) \mathrm{d}x^2\,.\end{aligned} \tag{51}$$

We therefore obtain generalized fractal metric potentials which are divergent and explicitly dependent on the coordinate differential elements [52,54]. Another equivalent way to understand this metric consists in remarking that it is no longer only quadratic in the space-time differental elements, but that it also contains them in a linear way.

As a consequence, the curvature is also explicitly scale-dependent and divergent when the scale intervals tend to zero. This property ensures the fundamentally non-Riemannian character of a fractal space-time, as well as the possibility to characterize it in an intrinsic way. Indeed, such a characterization, which is a necessary condition for defining a space in a genuine way, can be easily made by measuring the curvature at smaller and smaller scales. While the curvature vanishes by definition toward the small scales in Gauss-Riemann geometry, a fractal space can be characterized from the interior by the verification of the divergence toward small scales of curvature, and therefore of physical quantities like energy and momentum.

Now the expression of this divergence is nothing but the Heisenberg relations themselves, which therefore acquire in this framework the status of a fundamental geometric test of the fractality of space-time [52,54,69].

Geodesics of a Fractal Space-Time

The next step in such a geometric approach consists of the identifying the wave-particles with fractal space-time geodesics. Any measurement is interpreted as a selection of the geodesics bundle linked to the interaction with the measurement apparatus (that depends on its resolution) and/or to the information known about it (for example, the which-way-information in a two-slit experiment [56].

The three main consequences of nondifferentiability are:

(i) The number of fractal geodesics is infinite. This leads one to adopt a generalized statistical fluid-like description where the velocity $V^\mu(s)$ is replaced by a scale-dependent velocity field $V^\mu[X^\mu(s, \mathrm{d}s), s, \mathrm{d}s]$.

(ii) There is a breaking of the reflexion invariance of the differential element $\mathrm{d}s$. Indeed, in terms of fractal functions $f(s, \mathrm{d}s)$, two derivatives are defined,

$$\begin{aligned} X'_+(s, \mathrm{d}s) &= \frac{X(s+\mathrm{d}s, \mathrm{d}s) - X(s, \mathrm{d}s)}{\mathrm{d}s}\,, \\ X'_-(s, \mathrm{d}s) &= \frac{X(s, \mathrm{d}s) - X(s-\mathrm{d}s, \mathrm{d}s)}{\mathrm{d}s}\,, \end{aligned} \tag{52}$$

which transform into each other under the reflection ($\mathrm{d}s \leftrightarrow -\mathrm{d}s$), and which have a priori no reason to be equal. This leads to a fundamental two-valuedness of the velocity field.

(iii) The geodesics are themselves fractal curves of fractal dimension $D_\mathrm{F} = 2$ [30].

This means that one defines two divergent fractal velocity fields, $V_+[x(s, \mathrm{d}s), s, \mathrm{d}s] = v_+[x(s), s] + w_+[x(s, \mathrm{d}s), s, \mathrm{d}s]$ and $V_-[x(s, \mathrm{d}s), s, \mathrm{d}s] = v_-[x(s), s] + w_-[x(s, \mathrm{d}s), s, \mathrm{d}s]$, which can be decomposed in terms of differentiable parts v_+ and v_-, and of fractal parts w_+ and w_-. Note that, contrary to other attempts such as Nelson's stochastic quantum mechanics which introduces forward and backward velocities [51] (and which has been later disproved [34,80]), the two velocities are here both forward, since they do not correspond to a reversal of the time coordinate, but of the time differential element now considered as an independent variable.

More generally, we define two differentiable parts of derivatives $\mathrm{d}_+/\mathrm{d}s$ and $\mathrm{d}_-/\mathrm{d}s$, which, when they are applied to x^μ, yield the differential parts of the velocity fields, $v^\mu_+ = \mathrm{d}_+x^\mu/\mathrm{d}s$ and $v^\mu_- = \mathrm{d}_-x^\mu/\mathrm{d}s$.

Covariant Total Derivative

Let us first consider the non-relativistic case. It corresponds to a three-dimensional fractal space, without fractal time, in which the invariant $\mathrm{d}s$ is therefore identified with the time differential element $\mathrm{d}t$. One describes the elementary displacements $\mathrm{d}X^k$, where $k = 1, 2, 3$, on the geodesics of a nondifferentiable fractal space in terms of the sum of the two terms (omitting the indices for simplicity) $\mathrm{d}X_\pm = d_\pm x + \mathrm{d}\xi_\pm$, where $\mathrm{d}x$ represents the differentiable part and $\mathrm{d}\xi$ the fractal (nondifferentiable) part, defined as

$$\mathrm{d}_\pm x = v_\pm \mathrm{d}t\,, \quad \mathrm{d}\xi_\pm = \zeta_\pm \sqrt{2\mathcal{D}}\,\mathrm{d}t^{1/2}\,. \tag{53}$$

Here $\zeta_\pm$ are stochastic dimensionless variables such that $\langle\zeta_\pm\rangle = 0$ and $\langle\zeta^2_\pm\rangle = 1$, and $\mathcal{D}$ is a parameter that generalizes the Compton scale (namely, $\mathcal{D} = \hbar/2m$ in the case of standard quantum mechanics) up to the fundamental constant $c/2$. The two time derivatives are then combined in terms of a complex total time derivative operator [54],

$$\frac{\widehat{\mathrm{d}}}{\mathrm{d}t} = \frac{1}{2}\left(\frac{\mathrm{d}_+}{\mathrm{d}t} + \frac{\mathrm{d}_-}{\mathrm{d}t}\right) - \frac{i}{2}\left(\frac{\mathrm{d}_+}{\mathrm{d}t} - \frac{\mathrm{d}_-}{\mathrm{d}t}\right). \tag{54}$$

Applying this operator to the differentiable part of the position vector yields a complex velocity

$$\mathcal{V} = \frac{\widehat{\mathrm{d}}}{\mathrm{d}t}x(t) = V - iU = \frac{v_+ + v_-}{2} - i\,\frac{v_+ - v_-}{2}\,. \tag{55}$$

In order to find the expression for the complex time derivative operator, let us first calculate the derivative of a scalar function f. Since the fractal dimension is 2, one needs to go to the second order of expansion. For one variable it reads

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial X}\frac{\mathrm{d}X}{\mathrm{d}t} + \frac{1}{2}\frac{\partial^2 f}{\partial X^2}\frac{\mathrm{d}X^2}{\mathrm{d}t}\,. \tag{56}$$

Generalizing this process to three dimensions is straightforward.

Let us now take the stochastic mean of this expression, i. e., we take the mean on the stochastic variables $\zeta_\pm$ which appear in the definition of the fractal fluctuation $\mathrm{d}\xi_\pm$. By definition, since $\mathrm{d}X = \mathrm{d}x + \mathrm{d}\xi$ and $\langle\mathrm{d}\xi\rangle = 0$, we have $\langle\mathrm{d}X\rangle = \mathrm{d}x$, so that the second term is reduced (in 3 dimensions) to $v.\nabla f$. Now concerning the term $\mathrm{d}X^2/\mathrm{d}t$, it is infinitesimal and therefore it would not be taken into account in the standard differentiable case. But in the nondifferentiable case considered here, the mean square fluctuation is non-vanishing and of order $\mathrm{d}t$, namely, $\langle\mathrm{d}\xi^2\rangle = 2\mathcal{D}\mathrm{d}t$, so that the last term of Eq. (56) amounts in three dimensions to a Laplacian operator. One obtains, respectively for the (+) and (−) processes,

$$\frac{\mathrm{d}_\pm f}{\mathrm{d}t} = \left(\frac{\partial}{\partial t} + v_\pm.\nabla \pm \mathcal{D}\Delta\right) f\,. \tag{57}$$

Finally, by combining these two derivatives in terms of the complex derivative of Eq. (54), it reads [54]

$$\frac{\widehat{\mathrm{d}}}{\mathrm{d}t} = \frac{\partial}{\partial t} + \mathcal{V}.\nabla - i\mathcal{D}\Delta. \tag{58}$$

Under this form, this expression is not fully covariant [74], since it involves derivatives of the second order, so that its Leibniz rule is a linear combination of the first and second order Leibniz rules. By introducing the velocity operator [61]

$$\widehat{\mathcal{V}} = \mathcal{V} - i\mathcal{D}\nabla, \tag{59}$$

it may be given a fully covariant expression,

$$\frac{\widehat{\mathrm{d}}}{\mathrm{d}t} = \frac{\partial}{\partial t} + \widehat{\mathcal{V}}.\nabla\,. \tag{60}$$

Under this form it satisfies the first order Leibniz rule for partial derivatives.

We shall now see that $\widehat{\mathrm{d}}/\mathrm{d}t$ plays the role of a "covariant derivative operator" (in analogy with the covariant derivative of general relativity). Namely, one may write in its terms the equation of physics in a nondifferentiable space under a strongly covariant form identical to the differentiable case.

Complex Action and Momentum

The steps of construction of classical mechanics can now be followed, but in terms of complex and scale dependent quantities. One defines a Lagrange function that keeps its usual form, $\mathcal{L}(x, \mathcal{V}, t)$, but which is now complex. Then one defines a generalized complex action

$$S = \int_{t_1}^{t_2} \mathcal{L}(x, \mathcal{V}, t)\mathrm{d}t\,. \tag{61}$$

Generalized Euler–Lagrange equations that keep their standard form in terms of the new complex variables can be derived from this action [22,54], namely

$$\frac{\widehat{\mathrm{d}}}{\mathrm{d}t}\frac{\partial L}{\partial \mathcal{V}} - \frac{\partial L}{\partial x} = 0\,. \tag{62}$$

From the homogeneity of space and Noether's theorem, one defines a generalized complex momentum given by the same form as in classical mechanics as

$$\mathcal{P} = \frac{\partial \mathcal{L}}{\partial \mathcal{V}}\,. \tag{63}$$

If the action is now considered as a function of the upper limit of integration in Eq. (61), the variation of the action from a trajectory to another nearby trajectory yields a generalization of another well-known relation of classical mechanics,

$$\mathcal{P} = \nabla S\,. \tag{64}$$

Motion Equation

As an example, consider the case of a single particle in an external scalar field with potential energy ϕ (but the method can be applied to any situation described by a Lagrange function). The Lagrange function, $L = \frac{1}{2}mv^2 - \phi$, is generalized as $\mathcal{L}(x, \mathcal{V}, t) = \frac{1}{2}m\mathcal{V}^2 - \phi$. The Euler–Lagrange equations then keep the form of Newton's fundamental equation of dynamics $F = m\,\mathrm{d}v/\mathrm{d}t$, namely,

$$m\frac{\widehat{\mathrm{d}}}{\mathrm{d}t}\mathcal{V} = -\nabla\phi\,, \tag{65}$$

which is now written in terms of complex variables and complex operators.

In the case when there is no external field ($\phi = 0$), the covariance is explicit, since Eq. (65) takes the free form of the equation of inertial motion, i. e., of a geodesic equation,

$$\frac{\widehat{\mathrm{d}}}{\mathrm{d}t}\mathcal{V} = 0\,. \tag{66}$$

This is analogous to Einstein's general relativity, where the equivalence principle leads to the covariant equation of the motion of a free particle in the form of an inertial motion (geodesic) equation $Du_\mu/\mathrm{d}s = 0$, written in terms of the general-relativistic covariant derivative D, of the four-vector u_μ and of the proper time differential $\mathrm{d}s$.

The covariance induced by the effects of the nondifferentiable geometry leads to an analogous transformation of the equation of motions, which, as we show below, become after integration the Schrödinger equation, which can therefore be considered as the integral of a geodesic equation in a fractal space.

In the one-particle case the complex momentum $\mathcal{P}$ reads

$$\mathcal{P} = m\mathcal{V}\,, \tag{67}$$

so that from Eq. (64) the complex velocity $\mathcal{V}$ appears as a gradient, namely the gradient of the complex action

$$\mathcal{V} = \nabla S/m\,. \tag{68}$$

Wave Function

Up to now the various concepts and variables used were of a classical type (space, geodesics, velocity fields), even if

they were generalized to the fractal and nondifferentiable, explicitly scale-dependent case whose essence is fundamentally not classical.

We shall now make essential changes of variable, that transform this apparently classical-like tool to quantum mechanical tools (without any hidden parameter or new degree of freedom). The complex wave function ψ is introduced as simply another expression for the complex action S, by making the transformation

$$\psi = \mathrm{e}^{iS/S_0} \,. \tag{69}$$

Note that, despite its apparent form, this expression involves a phase and a modulus since S is complex. The factor S_0 has the dimension of an action (i. e., an angular momentum) and must be introduced because S is dimensioned while the phase should be dimensionless. When this formalism is applied to standard quantum mechanics, S_0 is nothing but the fundamental constant $\hbar$. As a consequence, since

$$S = -iS_0 \ln \psi \,, \tag{70}$$

one finds that the function ψ is related to the complex velocity appearing in Eq. (68) as follows

$$\mathcal{V} = -i\,\frac{S_0}{m}\,\nabla \ln \psi \,. \tag{71}$$

This expression is the fundamental relation that connects the two description tools while giving the meaning of the wave function in the new framework. Namely, it is defined here as a velocity potential for the velocity field of the infinite family of geodesics of the fractal space. Because of nondifferentiability, the set of geodesics that defines a 'particle' in this framework is fundamentally non-local. It can easily be generalized to a multiple particle situation (in particular to entangled states) which are described by a single wave function ψ, from which the various velocity fields of the subsets of the geodesic bundle are derived as $\mathcal{V}_k = -i\,(S_0/m_k)\,\nabla_k \ln \psi$, where k is an index for each particle. The indistinguishability of identical particles naturally follows from the fact that the 'particles' are identified with the geodesics themselves, i. e., with an infinite ensemble of purely geometric curves. In this description there is no longer any point-mass with 'internal' properties which would follow a 'trajectory', since the various properties of the particle – energy, momentum, mass, spin, charge (see next sections) – can be derived from the geometric properties of the geodesic fluid itself.

Correspondence Principle

Since we have $\mathcal{P} = -iS_0\nabla \ln \psi = -iS_0(\nabla\psi)/\psi$, we obtain the equality [54]

$$\mathcal{P}\psi = -i\hbar\nabla\psi \tag{72}$$

in the standard quantum mechanical case $S_0 = \hbar$, which establishes a correspondence between the classical momentum p, which is the real part of the complex momentum in the classical limit, and the operator $-i\hbar\nabla$.

This result is generalizable to other variables, in particular to the Hamiltonian. Indeed, a strongly covariant form of the Hamiltonian can be obtained by using the fully covariant form of the covariant derivative operator given by Eq. (60). With this tool, the expression of the relation between the complex action and the complex Lagrange function reads

$$\mathcal{L} = \frac{\widehat{\mathrm{d}}\,S}{\mathrm{d}t} = \frac{\partial S}{\mathrm{d}t} + \widehat{\mathcal{V}}.\nabla S \,. \tag{73}$$

Since $\mathcal{P} = \nabla S$ and $\mathcal{H} = -\partial S/\partial t$, one obtains for the generalized complex Hamilton function the same form it has in classical mechanics, namely [63,67],

$$\mathcal{H} = \widehat{\mathcal{V}}.\mathcal{P} - \mathcal{L} \,. \tag{74}$$

After expanding the velocity operator, one obtains $\mathcal{H} = \mathcal{V}.\mathcal{P} - i\mathcal{D}\nabla.\mathcal{P} - \mathcal{L}$, which includes an additional term [74], whose origin is now understood as an expression of nondifferentiability and strong covariance.

Schrödinger Equation and Compton Relation

The next step of the construction amounts to writing the fundamental equation of dynamics Eq. (65) in terms of the function ψ. It takes the form

$$iS_0\,\frac{\widehat{\mathrm{d}}}{\mathrm{d}t}(\nabla \ln \psi) = \nabla\phi \,. \tag{75}$$

As we shall now see, this equation can be integrated in a general way in the form of a Schrödinger equation. Replacing $\widehat{\mathrm{d}}/\mathrm{d}t$ and $\mathcal{V}$ by their expressions yields

$$\nabla\Phi = iS_0\left[\frac{\partial}{\partial t}\nabla \ln \psi - i\left\{\frac{S_0}{m}(\nabla \ln \psi.\nabla)(\nabla \ln \psi) + \mathcal{D}\Delta(\nabla \ln \psi)\right\}\right] . \tag{76}$$

This equation may be simplified thanks to the identity [54],

$$\nabla\left(\frac{\Delta\psi}{\psi}\right) = 2(\nabla \ln \psi.\nabla)(\nabla \ln \psi) + \Delta(\nabla \ln \psi) \,. \tag{77}$$

We recognize, in the right-hand side of Eq. (77), the two terms of Eq. (76), which were respectively a factor of S_0/m and $\mathcal{D}$. This leads to the definition of the wave function as

$$\psi = \mathrm{e}^{iS/2m\mathcal{D}} , \tag{78}$$

which means that the arbitrary parameter S_0 (which is identified with the constant $\hbar$ in standard QM) is now linked to the fractal fluctuation parameter by the relation

$$S_0 = 2m\mathcal{D} . \tag{79}$$

This relation (which can actually be proved instead of simply being set as a simplifying choice, see [62,67]) is actually a generalization of the Compton relation, since the geometric parameter $\mathcal{D} = \langle \mathrm{d}\xi^2 \rangle / 2\mathrm{d}t$ can be written in terms of a length scale as $\mathcal{D} = \lambda c/2$, so that, when $S_0 = \hbar$, it becomes $\lambda = \hbar/mc$. But a geometric meaning is now given to the Compton length (and therefore to the inertial mass of the particle) in the fractal space-time framework.

The fundamental equation of dynamics now reads

$$\nabla\phi = 2im\mathcal{D}\left[\frac{\partial}{\partial t}\nabla \ln\psi - i\left\{2\mathcal{D}(\nabla \ln\psi.\nabla)(\nabla \ln\psi) + \mathcal{D}\Delta(\nabla \ln\psi)\right\}\right] . \tag{80}$$

Using the above remarkable identity and the fact that $\partial/\partial t$ and ∇ commute, it becomes

$$-\frac{\nabla\phi}{m} = -2\mathcal{D}\nabla\left\{i\frac{\partial}{\partial t}\ln\psi + \mathcal{D}\frac{\Delta\psi}{\psi}\right\} . \tag{81}$$

The full equation becomes a gradient,

$$\nabla\left\{\frac{\phi}{m} - 2\mathcal{D}\nabla\left(\frac{i\,\partial\psi/\partial t + \mathcal{D}\Delta\psi}{\psi}\right)\right\} = 0 \tag{82}$$

and it can be easily integrated to finally obtain a generalized Schrödinger equation [54]

$$\mathcal{D}^2\Delta\psi + i\mathcal{D}\frac{\partial}{\partial t}\psi - \frac{\phi}{2m}\psi = 0 , \tag{83}$$

up to an arbitrary phase factor which may be set to zero by a suitable choice of the ψ phase. One recovers the standard Schrödinger equation of quantum mechanics for the particular case when $\mathcal{D} = \hbar/2m$.

Von Neumann's and Born's Postulates

In the framework described here, "particles" are identified with the various geometric properties of fractal space(-time) geodesics. In such an interpretation, a measurement (and more generally any knowledge about the system) amounts to selecting the sub-set of the geodesics family which only contains the geodesics with the geometric properties corresponding to the measurement result. Therefore, just after the measurement, the system is in the state given by the measurement result, which is precisely the von Neumann postulate of quantum mechanics.

The Born postulate can also be inferred from the scale-relativity construction [22,62,67]. Indeed, the probability for the particle to be found at a given position must be proportional to the density of the geodesics fluid at this point. The velocity and the density of the fluid are expected to be solutions of a Euler and continuity system of four equations, with four unknowns, (ρ, V_x, V_y, V_z).

Now, by separating the real and imaginary parts of the Schrödinger equation, setting $\psi = \sqrt{P} \times \mathrm{e}^{i\theta}$ and using a mixed representation (P, V), where $V = \{V_x, V_y, V_z\}$, one precisely obtains such a standard system of fluid dynamics equations, namely,

$$\begin{aligned} &\left(\frac{\partial}{\partial t} + V\cdot\nabla\right)V = -\nabla\left(\phi - 2\mathcal{D}^2\frac{\Delta\sqrt{P}}{\sqrt{P}}\right), \\ &\frac{\partial P}{\partial t} + \mathrm{div}(PV) = 0 . \end{aligned} \tag{84}$$

This allows one to unequivocally identify $P = |\psi|^2$ with the probability density of the geodesics and therefore with the probability of presence of the 'particle'. Moreover,

$$Q = -2\mathcal{D}^2\frac{\Delta\sqrt{P}}{\sqrt{P}} \tag{85}$$

can be interpreted as the new potential which is expected to emerge from the fractal geometry, in analogy with the identification of the gravitational field as a manifestation of the curved geometry in Einstein's general relativity. This result is supported by numerical simulations, in which the probability density is obtained directly from the distribution of geodesics without writing the Schrödinger equation [39,63].

Nondifferentiable Wave Function

In more recent works, instead of taking only the differentiable part of the velocity field into account, one constructs the covariant derivative and the wave function in terms of the full velocity field, including its divergent nondifferentiable part of zero mean [59,62]. As we shall briefly see now, this still leads to the standard form of the Schrödinger equation. This means that, in the scale relativity framework, one expects the Schrödinger equation to have fractal and nondifferentiable solutions. This result

agrees with a similar conclusion by Berry [8] and Hall [38], but it is considered here as a direct manifestation of the nondifferentiability of space itself.

Consider the full complex velocity field, including its differentiable and nondifferentiable parts,

$$\tilde{\mathcal{V}} = \mathcal{V} + \mathcal{W} = \left(\frac{v_+ + v_-}{2} - i\,\frac{v_+ - v_-}{2}\right) + \left(\frac{w_+ + w_-}{2} - i\,\frac{w_+ - w_-}{2}\right) . \quad (86)$$

It is related to a full complex action $\tilde{S}$ and a full wave function $\tilde{\psi}$ as

$$\tilde{\mathcal{V}} = \mathcal{V} + \mathcal{W} = \nabla\tilde{S}/m = -2i\mathcal{D}\nabla\ln\tilde{\psi} . \quad (87)$$

Under the standard point of view, the complex fluctuation $\mathcal{W}$ is infinite and therefore $\nabla\ln\tilde{\psi}$ is undefined, so that Eq. (87) would be meaningless. In the scale relativity approach, on the contrary, this equation keeps a mathematical and physical meaning, in terms of fractal functions, which are explicitly dependent on the scale interval $\mathrm{d}t$ and divergent when $\mathrm{d}t \to 0$.

After some calculations [59,62], one finds that the covariant derivative built from the total process (including the differentiable and nondifferentiable divergent terms) is finally

$$\frac{\widehat{\mathrm{d}}}{\mathrm{d}t} = \frac{\partial}{\partial t} + \tilde{\mathcal{V}}.\nabla - i\mathcal{D}\Delta . \quad (88)$$

The subsequent steps of the derivation of the Schrödinger equation are unchanged (now in terms of scale-dependent fractal functions), so that one obtains

$$\mathcal{D}^2\Delta\tilde{\psi} + i\mathcal{D}\,\frac{\partial\tilde{\psi}}{\partial t} - \frac{\phi}{2m}\tilde{\psi} = 0 , \quad (89)$$

where $\tilde{\psi}$ can now be a nondifferentiable and fractal function. The research of such a behavior in laboratory experiments is an interesting new challenge for quantum physics.

One may finally stress the fact that this result is obtained by accounting for all the terms, differentiable ($\mathrm{d}x$) and nondifferentiable ($\mathrm{d}\xi$), in the description of the elementary displacements in a nondifferentiable space, and that it does not depend at all on the probability distribution of the stochastic variables $\mathrm{d}\xi$, about which no hypothesis is needed. This means that the description is fully general, and that the effect on motion of a nondifferentiable space(-time) amounts to a fundamental indeterminism, i. e., to a total loss of information about the past path which will in all cases lead to the QM description.

From Fractal Space-Time to Relativistic Quantum Mechanics

All these results can be generalized to relativistic quantum mechanics, which corresponds in the scale relativity framework to a full fractal space-time. This yields, as a first step, the Klein–Gordon equation [22,55,56].

Then an account of the new two-valuedness of the velocity allows one to suggest a geometric origin for the spin and to obtain the Dirac equation [22]. Indeed, the total derivative of a physical quantity also involves partial derivatives with respect to the space variables, $\partial/\partial x^\mu$. From the very definition of derivatives, the discrete symmetry under the reflection $\mathrm{d}x^\mu \leftrightarrow -\mathrm{d}x^\mu$ is also broken. Since, at this level of description, one should also account for parity as in the standard quantum theory, this leads to introduce a bi-quaternionic velocity field [22], in terms of which the Dirac bispinor wave function can be constructed.

The successive steps that lead to the Dirac equation naturally generalize the Schrödinger case. One introduces a biquaternionic generalization of the covariant derivative that keeps the same form as in the complex case, namely,

$$\frac{\widehat{\mathrm{d}}}{\mathrm{d}s} = \mathcal{V}^\nu\partial_\nu + i\,\frac{\lambda}{2}\,\partial^\nu\partial_\nu , \quad (90)$$

where $\lambda = 2\mathcal{D}/c$. The biquaternionic velocity field is related to the biquaternionic (i. e., bispinorial) wave function, by

$$\mathcal{V}_\mu = i\,\frac{S_0}{m}\,\psi^{-1}\partial_\mu\psi . \quad (91)$$

This is the relativistic expression of the fundamental relation between the scale relativity tools and the quantum mechanical tools of description. It gives a geometric interpretation to the wave function, which is, in this framework, a manifestation of the geodesic fluid and of its associated fractal velocity field. The covariance principle allows us to write the equation of motion under the form of a geodesic differential equation,

$$\frac{\widehat{\mathrm{d}}\,\mathcal{V}_\mu}{\mathrm{d}s} = 0 . \quad (92)$$

After some calculations, this equation can be integrated and factorized, and one finally derives the Dirac equation [22],

$$\frac{1}{c}\frac{\partial\psi}{\partial t} = -\alpha^k\frac{\partial\psi}{\partial x^k} - i\frac{mc}{\hbar}\beta\psi . \quad (93)$$

Finally it is easy to recover the Pauli equation and Pauli spinors as nonrelativistic approximations of the Dirac equation and Dirac bispinors [23].

Gauge Fields as Manifestations of Fractal Geometry

General Scale Transformations and Gauge Fields

Finally, let us briefly recall the main steps of applying of the scale relativity principles to the foundation of gauge theories, in the Abelian [55,56] and non-Abelian [63,68] cases.

This application is based on a general description of the internal fractal structures of the "particle" (identified with the geodesics of a nondifferentiable space-time) in terms of scale variables $\eta_{\alpha\beta}(x, y, z, t) = \varrho_{\alpha\beta}\, \varepsilon_\alpha\, \varepsilon_\beta$ whose true nature is tensorial, since it involves resolutions that may be different for the four space-time coordinates and may be correlated. This resolution tensor (similar to a covariance error matrix) generalizes the single resolution variable ε. Moreover, one considers a more profound level of description in which the scale variables may now be functions of the coordinates. Namely, the internal structures of the geodesics may vary from place to place and during the time evolution, in agreement with the non-absolute character of the scale space. This generalization amounts to the construction of a 'general scale relativity' theory.

We assume, for simplicity of the writing, that the two tensorial indices can be gathered under one common index. We therefore write the scale variables under the simplified form $\eta_{\alpha_1\alpha_2} = \eta_\alpha$, $\alpha = 1$ to $N = n(n+1)/2$, where n is the number of space-time dimensions ($n = 3$, $N = 6$ for fractal space, $n = 4$, $N = 10$ for fractal space-time and $n = 5$, $N = 15$ in the special scale relativity case where one treats the djinn (scale-time τ_F) as a fifth dimension [53]).

Let us consider infinitesimal scale transformations. The transformation law on the η_α can be written in a linear way as

$$\eta'_\alpha = \eta_\alpha + \delta\eta_\alpha = (\delta_{\alpha\beta} + \delta\theta_{\alpha\beta})\, \eta^\beta \,, \tag{94}$$

where $\delta_{\alpha\beta}$ is the Kronecker symbol. Let us now assume that the η_α's are functions of the standard space-time coordinates. This leads one to define a new scale-covariant derivative by writing the total variation of the resolution variables as the sum of the inertial variation, described by the covariant derivative, and of the new geometric contribution, namely,

$$\mathrm{d}\eta_\alpha = D\eta_\alpha - \eta^\beta\, \delta\theta_{\alpha\beta} = D\eta_\alpha - \eta^\beta\, W^\mu_{\alpha\beta}\, \mathrm{d}x_\mu \,. \tag{95}$$

This covariant derivative is similar to that of general relativity, i. e., it amounts to the subtraction of the new geometric part in order to only keep the inertial part (for which the motion equation will therefore take a geodesical, free-like form). This is different from the case of the previous quantum-covariant derivative, which includes the effects of nondifferentiability by adding new terms in the total derivative.

In this new situation we are led to introduce "gauge field potentials" $W^\mu_{\alpha\beta}$ that enter naturally in the geometrical framework of Eq. (95). These potentials are linked to the scale transformations as follows:

$$\delta\theta_{\alpha\beta} = W^\mu_{\alpha\beta}\, \mathrm{d}x_\mu \,. \tag{96}$$

But one should keep in mind, when using this expression, that these potentials find their origin in a covariant derivative process and are therefore not gradients.

Generalized Charges

After having written the transformation law of the basic variables (the η_α's), one now needs to describe how various physical quantities transform under these η_α transformations. These new transformation laws are expected to depend on the nature of the objects they transform (e. g., vectors, tensors, spinors, etc.), which implies a jump to group representations.

We anticipate the existence of charges (which are fully constructed herebelow) by generalizing the relation (91) to multiplets between the velocity field and the wave function. In this case the multivalued velocity becomes a biquaternionic matrix,

$$\mathcal{V}^\mu_{jk} = i\lambda\, \psi_j^{-1} \partial^\mu \psi_k \,. \tag{97}$$

The biquaternionic, and therefore non-commutative, nature of the wave function [15], which is equivalent to Dirac bispinors, plays an essential role here. Indeed, the general structure of Yang–Mills theories and the correct construction of non-Abelian charges can be obtained thanks to this result [68].

The action also becomes a tensorial biquaternionic quantity,

$$\mathrm{d}S_{jk} = \mathrm{d}S_{jk}\left(x^\mu, \mathcal{V}^\mu_{jk}, \eta_\alpha\right) , \tag{98}$$

and, in the absence of a field (free particle) it is linked to the generalized velocity (and therefore to the spinor multiplet) by the relation

$$\partial^\mu S_{jk} = -mc\, \mathcal{V}^\mu_{jk} = -i\hbar\, \psi_j^{-1} \partial^\mu \psi_k \,. \tag{99}$$

Now, in the presence of a field (i. e., when the second-order effects of the fractal geometry appearing in the right hand side of Eq. (95) are included), using the complete expression for $\partial^\mu \eta_\alpha$,

$$\partial^\mu \eta_\alpha = D^\mu \eta_\alpha - W^\mu_{\alpha\beta}\, \eta^\beta \,, \tag{100}$$

one obtains a non-Abelian relation,

$$\partial^{\mu} S_{jk} = D^{\mu} S_{jk} - \eta^{\beta} \frac{\partial S_{jk}}{\partial \eta_{\alpha}} W^{\mu}_{\alpha\beta} \,. \tag{101}$$

This finally leads to the definition of a general group of scale transformations whose generators are

$$T^{\alpha\beta} = \eta^{\beta} \partial^{\alpha} \tag{102}$$

(where we use the compact notation $\partial^{\alpha} = \partial/\partial\eta_{\alpha}$), yielding the generalized charges,

$$\frac{\tilde{g}}{c} t^{\alpha\beta}_{jk} = \eta^{\beta} \frac{\partial S_{jk}}{\partial \eta_{\alpha}} \,. \tag{103}$$

This unified group is submitted to a unitarity condition, since, when it is applied to the wave functions, $\psi\psi^{\dagger}$ must be conserved. Knowing that α, β represent two indices each, this is a large group – at least SO(10) – that contains the electroweak theory [33,78,81] and the standard model $U(1) \times SU(2) \times SU(3)$ and its simplest grand unified extension SU(5) [31,32] as a subset (see [53,54] for solutions in the special scale relativity framework to the problems encountered by SU(5) GUTs).

As it is shown in more detail in Ref. [68], the various ingredients of Yang–Mills theories (gauge covariant derivative, gauge invariance, charges, potentials, fields, etc.) may subsequently be recovered in such a framework, but they now have a first principle and geometric scale-relativistic foundation.

Future Directions

In this contribution, we have recalled the main steps that lead to a new foundation of quantum mechanics and of gauge fields on the principle of relativity itself (once it includes scale transformations of the reference system), and on the generalized geometry of space-time which is naturally associated with such a principle, namely, nondifferentiability and fractality.

For this purpose, two covariant derivatives have been constructed, which account for the nondifferentiable and fractal geometry of space-time, and which allow one to write the equations of motion as geodesics equations. After a change of variable, these equations finally take the form of the quantum mechanical and quantum field equations.

Let us conclude by listing some original features of the scale relativity approach which could lead to experimental tests of the theory and/or to new experiments in the future [63,67]: (i) nondifferentiable and fractal solutions of the Schrödinger equation; (ii) zero particle interference in a Young slit experiment; (iii) possible breaking of the Born postulate for a possible effective kinetic energy operator $\widehat{T} \neq -(\hbar^2/2m)\Delta$ [67]; (iv) underlying quantum behavior in the classical domain, at scales far larger than the de Broglie scale [67]; (v) macroscopic systems described by a Schrödinger-type mechanics based on a generalized macroscopic parameter $\mathcal{D} \neq \hbar/2m$ (see Chap. 7.2 in [54] and [24,57]); (vi) applications to cosmology [60]; (vii) applications to life sciences and other sciences [4,64,65], etc.

Bibliography

Primary Literature

1. Abbott LF, Wise MB (1981) Am J Phys 49:37
2. Amelino-Camelia G (2001) Phys Lett (B)510:255
3. Amelino-Camelia G (2002) Int J Mod Phys (D)11:1643
4. Auffray C, Nottale L (2007) Progr Biophys Mol Bio 97:79
5. Ben Adda F, Cresson J (2000) CR Acad Sci Paris 330:261
6. Ben Adda F, Cresson J (2004) Chaos Solit Fractals 19:1323
7. Ben Adda F, Cresson J (2005) Appl Math Comput 161:323
8. Berry MV (1996) J Phys A: Math Gen 29:6617
9. Cafiero R, Loreto V, Pietronero L, Vespignani A, Zapperi S (1995) Europhys Lett 29:111
10. Carpinteri A, Chiaia B (1996) Chaos Solit Fractals 7:1343
11. Cash R, Chaline J, Nottale L, Grou P (2002) CR Biologies 325:585
12. Castro C (1997) Found Ph ys Lett 10:273
13. Castro C, Granik A (2000) Chaos Solit Fractals 11:2167
14. Chaline J, Nottale L, Grou P (1999) C R Acad Sci Paris 328:717
15. Connes A (1994) Noncommutative Geometry. Academic Press, New York
16. Connes A, Douglas MR, Schwarz A J High Energy Phys 02:003 (hep-th/9711162)
17. Cresson J (2001) Mémoire d'habilitation à diriger des recherches. Université de Franche-Comté, Besançon
18. Cresson J (2002) Chaos Solit Fractals 14:553
19. Cresson J (2003) J Math Phys 44:4907
20. Cresson J (2006) Int J Geometric Methods in Mod Phys 3(7)
21. Cresson J (2007) J Math Phys 48:033504
22. Célérier MN, Nottale L (2004) J Phys A: Math Gen 37:931
23. Célérier MN, Nottale L (2006) J Phys A: Math Gen 39:12565
24. da Rocha D, Nottale L (2003) Chaos Solit Fractals 16:565
25. Dubois D (2000) In: Proceedings of CASYS'1999, 3rd International Conference on Computing Anticipatory Systems, Liège, Belgium, Am. Institute of Physics Conference Proceedings 517:417
26. Dubrulle B, Graner F, Sornette D (eds) (1997) In: Dubrulle B, Graner F, Sornette D (eds) Scale invariance and beyond, Proceedings of Les Houches school, EDP Sciences, Les Ullis/Springer, Berlin, New York, p 275
27. El Naschie MS (1992) Chaos Solit Fractals 2:211
28. El Naschie MS Chaos Solit Fractals 11:2391
29. El Naschie MS, Rössler O, Prigogine I (eds) (1995) Quantum mechanics, diffusion and chaotic fractals. Pergamon, New York
30. Feynman RP, Hibbs AR (1965) Quantum mechanics and path integrals. MacGraw-Hill, New York
31. Georgi H, Glashow SL (1974) Phys Rev Lett 32:438
32. Georgi H, Quinn HR, Weinberg S (1974) Phys Rev Lett 33:451
33. Glashow SL (1961) Nucl Phys 22:579
34. Grabert H, Hänggi P, Talkner P (1979) Phys Rev A(19):2440

35. Green MB, Schwarz JH, Witten E (1987) Superstring Theory, vol 2. Cambridge University Press,
36. Grou P (1987) L'aventure économique. L'Harmattan, Paris
37. Grou P, Nottale L, Chaline J (2004) In: Zona Arqueologica, Miscelanea en homenaje a Emiliano Aguirre, IV Arqueologia, 230, Museo Arquelogico Regional, Madrid
38. Hall MJW (2004) J Phys A: Math Gen 37:9549
39. Hermann R (1997) J Phys A: Math Gen 30:3967
40. Johansen A, Sornette D (2001) Physica A(294):465
41. Jumarie G (2001) Int J Mod Phys A(16):5061
42. Jumarie G (2006) Chaos Solit Fractals 28:1285
43. Jumarie G (2006) Comput Math 51:1367
44. Jumarie G (2007) Phys Lett A 363:5
45. Kröger H (2000) Phys Rep 323:81
46. Laperashvili LV, Ryzhikh DA (2001) arXiv: hep-th/0110127 (Institute for Theoretical and Experimental Physics, Moscow)
47. Levy-Leblond JM (1976) Am J Phys 44:271
48. Losa G, Merlini D, Nonnenmacher T, Weibel E (eds) Fractals in biology and medicine. vol 3. Proceedings of Fractal 2000 Third International Symposium, Birkhäuser
49. Mandelbrot B (1982) The fractal geometry of nature. Freeman, San Francisco
50. McKeon DGC, Ord GN (1992) Phys Rev Lett 69:3
51. Nelson E (1966) Phys Rev 150:1079
52. Nottale L (1989) Int J Mod Phys A(4):5047
53. Nottale L (1992) Int J Mod Phys A(7):4899
54. Nottale L (1993) Fractal space-time and microphysics: Towards a theory of scale relativity. World Scientific, Singapore
55. Nottale L (1994) In: Relativity in general, (Spanish Relativity Meeting (1993)), edited Alonso JD, Paramo ML (eds), Editions Frontières, Paris, p 121
56. Nottale L (1996) Chaos Solit Fractals 7:877
57. Nottale L (1997) Astron Astrophys 327:867
58. Nottale L (1997) In: Scale invariance and beyond, Proceedings of Les Houches school, Dubrulle B, Graner F, Sornette D (eds) EDP Sciences, Les Ullis/Springer, Berlin, New York, p 249
59. Nottale L (1999) Chaos Solit Fractals 10:459
60. Nottale L (2003) Chaos Solit Fractals 16:539
61. Nottale L (2004) American Institute of Physics Conference Proceedings 718:68
62. Nottale L (2008) Proceedings of 7th International Colloquium on Clifford Algebra and their applications, 19–29 May 2005, Toulouse, Advances in Applied Clifford Algebras (in press)
63. Nottale L (2008) The theory of scale relativity. (submitted)
64. Nottale L, Auffray C (2007) Progr Biophys Mol Bio 97:115
65. Nottale L, Chaline J, Grou P (2000) Les arbres de l'évolution: Univers, Vie, Sociétés. Hachette, Paris, 379 pp
66. Nottale L, Chaline J, Grou P (2002) In: Fractals in biology and medicine, vol 3. Proceedings of Fractal (2000) Third International Symposium, Losa G, Merlini D, Nonnenmacher T, Weibel E (eds), Birkhäuser, p 247
67. Nottale L, Célérier MN (2008) J Phys A 40:14471
68. Nottale L, Célérier MN, Lehner T (2006) J Math Phys 47:032303
69. Nottale L, Schneider J (1984) J Math Phys 25:1296
70. Novak M (ed) (1998) Fractals and beyond: Complexities in the sciences, Proceedings of the Fractal 98 conference, World Scientific
71. Ord GN (1983) J Phys A: Math Gen 16:1869
72. Ord GN (1996) Ann Phys 250:51
73. Ord GN, Galtieri JA (2002) Phys Rev Lett 1989:250403
74. Pissondes JC (1999) J Phys A: Math Gen 32:2871
75. Polchinski J (1998) String theories. Cambridge University Press, Cambridge
76. Rovelli C, Smolin L (1988) Phys Rev Lett 61:1155
77. Rovelli C, Smolin L (1995) Phys Rev D(52):5743
78. Salam A (1968) Elementary particle theory. Svartholm N (ed). Almquist & Wiksells, Stockholm
79. Sornette D (1998) Phys Rep 297:239
80. Wang MS, Liang WK (1993) Phys Rev D(48):1875
81. Weinberg S (1967) Phys Rev Lett 19:1264

Books and Reviews

Georgi H (1999) Lie Algebras in particle physics. Perseus books, Reading, Massachusetts
Landau L, Lifchitz E (1970) Theoretical physics, 10 volumes, Mir, Moscow
Lichtenberg AJ, Lieberman MA (1983) Regular and stochastic motion. Springer, New York
Lorentz HA, Einstein A, Minkowski H, Weyl H (1923) The principle of relativity. Dover, New York
Mandelbrot B (1975) Les objets fractals. Flammarion, Paris
Misner CW, Thorne KS, Wheeler JA (1973) Gravitation. Freeman, San Francisco
Peebles J (1980) The large-scale structure of the universe. Princeton University Press, Princeton
Rovelli C (2004) Quantum gravity. Cambridge Universty Press, Cambridge
Weinberg S (1972) Gravitation and cosmology. Wiley, New York

Fractal Structures in Condensed Matter Physics

Tsuneyoshi Nakayama
Toyota Physical and Chemical Research Institute, Nagakute, Japan

Article Outline

Glossary
Definition of the Subject
Introduction
Determining Fractal Dimensions
Polymer Chains in Solvents
Aggregates and Flocs
Aerogels
Dynamical Properties of Fractal Structures
Spectral Density of States and Spectral Dimensions
Future Directions
Bibliography

Glossary

Anomalous diffusion It is well known that the mean-square displacement $\langle r^2(t)\rangle$ of a diffusing particle on a uniform system is proportional to the time t such

as $\langle r^2(t)\rangle \sim t$. This is called *normal diffusion*. Particles on fractal networks diffuse more slowly compared with the case of normal diffusion. This slow diffusion called *anomalous diffusion* follows the relation given by $\langle r^2(t)\rangle \sim t^a$, where the condition $0 < a < 1$ always holds.

Brownian motion Einstein published the important paper in 1905 opening the way to investigate the movement of small particles suspended in a stationary liquid, the so-called Brownian motion, which stimulated J. Perrin in 1909 to pursue his experimental work confirming the atomic nature of matter. The trail of a random walker provides an instructive example for understanding the meaning of random fractal structures.

Fractons Fractons, excitations on fractal elastic-networks, were named by S. Alexander and R. Orbach in 1982. Fractons manifest not only static properties of fractal structures but also their dynamic properties. These modes show unique characteristics such as strongly localized nature with the localization length of the order of wavelength.

Spectral density of states The spectral density of states of ordinary elastic networks are expressed by the Debye spectral density of states given by $D(\omega) \sim \omega^{d-1}$, where d is the Euclidean dimensionality. The spectral density of states of fractal networks is given by $D(\omega) \sim \omega^{d_s-1}$, where d_s is called the spectral or fracton dimension of the system.

Spectral dimension This exponent characterizes the spectral density of states for vibrational modes excited on fractal networks. The spectral dimension constitutes the dynamic exponent of fractal networks together with the conductivity exponent and the exponent of anomalous diffusion.

Definition of the Subject

The idea of fractals is based on *self-similarity*, which is a symmetry property of a system characterized by invariance under an isotropic scale-transformation on certain length scales. The term *scale-invariance* has the implication that objects look the same on different scales of observations. While the underlying concept of fractals is quite simple, the concept is used for an extremely broad range of topics, providing a simple description of highly complex structures found in nature. The term *fractal* was first introduced by Benoit B. Mandelbrot in 1975, who gave a definition on fractals in a simple manner "*A fractal is a shape made of parts similar to the whole in some way*". Thus far, the concept of fractals has been extensively used to understand the behaviors of many complex systems or has been applied from physics, chemistry, and biology for applied sciences and technological purposes. Examples of fractal structures in condensed matter physics are numerous such as polymers, colloidal aggregations, porous media, rough surfaces, crystal growth, spin configurations of diluted magnets, and others. The critical phenomena of phase transitions are another example where self-similarity plays a crucial role. Several books have been published on fractals and reviews concerned with special topics on fractals have appeared.

Length, area, and volume are special cases of ordinary Euclidean *measures*. For example, length is the measure of a one-dimensional (1d) object, area the measure of a two-dimensional (2d) object, and volume the measure of a three-dimensional (3d) object. Let us employ a physical quantity (observable) as the measure to define dimensions for Euclidean systems, for example, a total mass $M(r)$ of a fractal object of the size r. For this, the following relation should hold

$$r \propto M(r)^{1/d} , \tag{1}$$

where d is the Euclidean dimensionality. Note that Euclidean spaces are the simplest scale-invariant systems. We extend this idea to introduce dimensions for self-similar fractal structures. Consider a set of particles with unit mass m randomly distributed on a d-dimensional Euclidean space called the *embedding space* of the system. Draw a sphere of radius r and denote the total mass of particles included in the sphere by $M(r)$. Provided that the following relation holds *in the meaning of statistical average* such as

$$r \propto \langle M(r)\rangle^{1/D_f} , \tag{2}$$

where $\langle \ldots \rangle$ denotes the ensemble-average over different spheres of radius r, we call D_f the *similarity dimension*. It is necessary, of course, that D_f is smaller than the embedding Euclidean dimension d. The definition of dimension as a *statistical quantity* is quite useful to specify the characteristic of a self-similar object if we could choose a suitable measure.

There are many definitions to allocate dimensions. Sometimes these take the same value as each other and sometimes not. The *capacity dimension* is based on the coverage procedure. As an example, the length of a curved line L is given by the product of the number N of straight-line segment of length r needed to step along the curve from one end to the other such as $L(r) = N(r)r$. While, the area $S(r)$ or the volume $V(r)$ of arbitrary objects can be measured by covering it with squares or cubes of linear

size r. The identical relation,

$$M(r) \propto N(r) r^d \tag{3}$$

should hold for the total mass $M(r)$ as measure, for example. If this relation does not change as $r \to 0$, we have the relation $N(r) \propto r^{-d}$. We can extend the idea to define the dimensions of fractal structures such as

$$N(r) \propto r^{-D_f} , \tag{4}$$

from which the *capacity dimension* D_f is given by

$$D_f := \lim_{r \to 0} \frac{\ln N(r)}{\ln(1/r)} . \tag{5}$$

The definition of D_f can be rendered in the following implicit form

$$\lim_{r \to 0} N(r) r^{D_f} = \text{const} . \tag{6}$$

Equation (5) brings out a key property of the *Hausdorff dimension* [10], where the product $N(r)r^{D_f}$ remains finite as $r \to 0$. If D_f is altered even by an infinitesimal amount, this product will diverge either to zero or to infinity. The Hausdorff dimension coincides with the capacity dimension for many fractal structures, although the Hausdorff dimension is defined less than or equal to the capacity dimension. Hereafter, we refer to the capacity dimension or the Hausdorff dimension mentioned above as the *fractal dimension*.

Introduction

Fractal structures are classified into two categories; *deterministic* fractals and *random* fractals. In condensed matter physics, we encounter many examples of random fractals. The most important characteristic of random fractals is the spatial and/or sample-to-sample fluctuations in their properties. We must discuss their characteristics by averaging over a large ensemble. The nature of deterministic fractals can be easily understood from some examples. An instructive example is the *Mandelbrot–Given fractal* [12], which can be constructed by starting with a structure with eight line segments as shown in Fig. 1a (the first stage of the Mandelbrot–Given fractal). In the second stage, each line segment of the initial structure is replaced by the initial structure itself (Fig. 1b). This process is repeated indefinitely. The Mandelbrot–Given fractal possesses an obvious dilatational symmetry, as seen from Fig. 1c, i. e., when we magnify a part of the structure, the enlarged portion looks just like the original one. Let us apply (5) to determine D_f of the Mandelbrot–Given fractal. The Mandelbrot–Given fractal is composed of 8 parts of size 1/3, hence, $N(1/3) = 8$, $N((1/3)^2) = 8^2$, and so on. We thus have a relation of the form $N(r) \propto r^{-\ln_3 8}$, which gives the fractal dimension $D_f = \ln_3 8 = 1.89278\ldots$. The Mandelbrot–Given fractal has many analogous features with *percolation networks* (see Sect. "Dynamical Properties of Fractal Structures"), a typical random fractal, such that the fractal dimension of a 2d percolation network is $D_f = 91/48 = 1.895833\ldots$, which is very close to that of the Mandelbrot–Given fractal.

The geometric characteristics of random fractals can be understood by considering two extreme cases of random structures. Figure 2a represents the case in which particles are randomly but *homogeneously* distributed in a d-dimensional box of size L, where d represents ordinary Euclidean dimensionality of the embedding space. If we divide this box into smaller boxes of size l, the *mass density* of the ith box is

$$\rho_i(l) = \frac{M_i(l)}{l^d} , \tag{7}$$

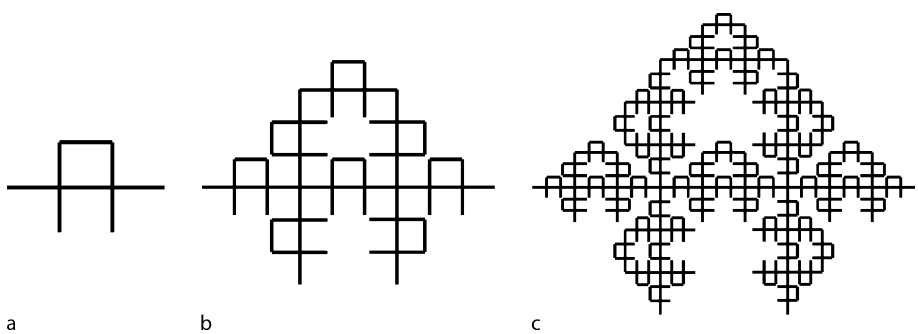

Fractal Structures in Condensed Matter Physics, Figure 1
Mandelbrot–Given fractal. a The initial structure with eight line segments, b the object obtained by replacing each line segment of the initial structure by the initial structure itself (the second stage), and c the third stage of the Mandelbrot–Given fractal obtained by replacing each line segment of the second-sage structure by the initial structure

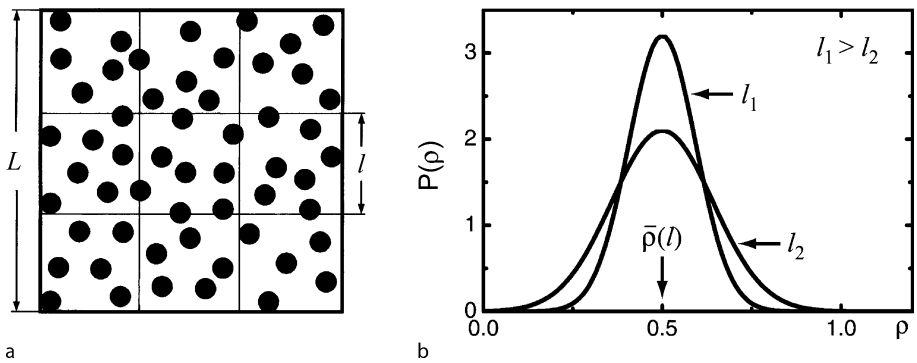

Fractal Structures in Condensed Matter Physics, Figure 2
a Homogeneous random structure in which particles are randomly but homogeneously distributed, and b the distribution functions of local densities ρ, where $\overline{\rho}(l)$ is the average mass density independent of l

where $M_i(l)$ represents the total mass (measure) inside box i. Since this quantity depends on the box i, we plot the distribution function $P(\rho)$, from which curves like those in Fig. 2b may be obtained for two box sizes l_1 and $(l_2 < l_2)$. We see that the central peak position of the distribution function $P(\rho)$ is the same for each case. This means that the average mass density yields

$$\overline{\rho}(l) = \frac{\langle M_i(l)\rangle_i}{l^d}$$

becomes constant, indicating that $\langle M_i(l)\rangle_i \propto l^d$. The above is equivalent to

$$\overline{\rho} = \frac{m}{\overline{a}^d}\,, \tag{8}$$

where $\overline{a}$ is the average distance (characteristic length-scale) between particles and the mass of a single particle. This indicates that there exists a single length scale $\overline{a}$ characterizing the random system given in Fig. 2a.

The other type of random structure is shown in Fig. 3a, where particle positions are correlated with each other and $\rho_i(l)$ greatly fluctuates from box to box, as shown in Fig. 3b. The relation $\langle M_i(l)\rangle_i \propto l^d$ may not hold at all for this type of structure. Assuming the fractality for this system, namely, if the power law $\langle M_i(l)\rangle_i \propto l_{\mathrm{f}}^D$ holds, the average mass density becomes

$$\bar{\rho}(l) = \frac{\langle M_i(l)\rangle_i}{l^d} \propto l^{D_{\mathrm{f}}-d}\,, \tag{9}$$

where $\rho_i(l) = 0$ is excluded. In the case $D_{\mathrm{f}} < d$, $\bar{\rho}(l)$ depends on l and decreases with increasing l. Thus, there is no characteristic length scale for the type of random structure shown in Fig. 3a. If (9) holds with $D_{\mathrm{f}} < d$, so that $\langle M_i(l)\rangle_i$ is proportional to l_{f}^D, the structure is said to be *fractal*. It is important to note that there is no *characteristic length scale* for the type of random fractal structure shown in Fig. 3b. Thus, we can extend the idea of self-similarity not only for deterministic self-similar structures, but also

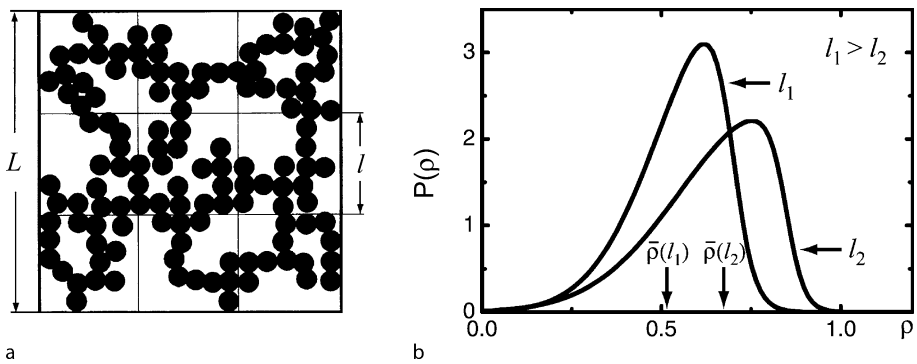

Fractal Structures in Condensed Matter Physics, Figure 3
a Correlated random fractal structure in which particles are randomly distributed, but correlated with each other, and b the distribution functions of local densities ρ with finite values, where the average mass densities depend on l

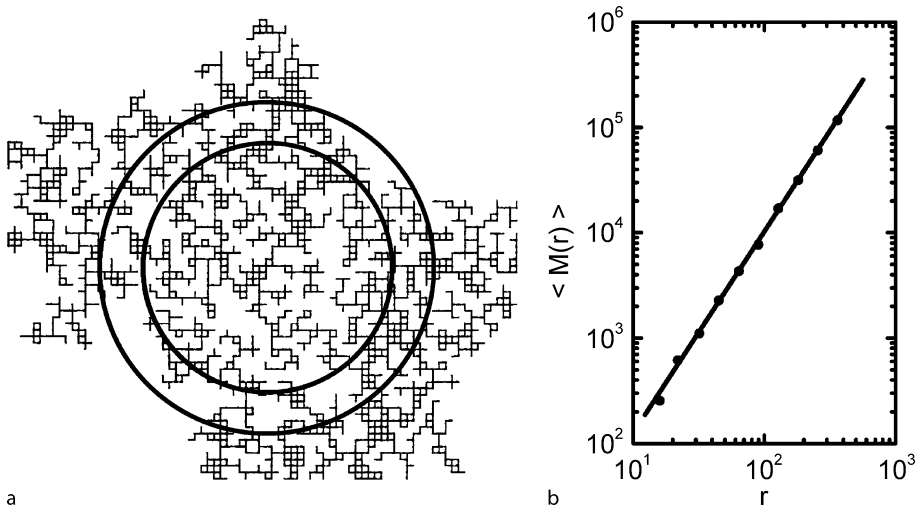

Fractal Structures in Condensed Matter Physics, Figure 4
a 2d site-percolation network and circles with different radii. b The power law relation holds between *r* and the number of particles in the sphere of radius *r*, indicating the fractal dimension of the 2d network is $D_f = 1.89\ldots = 91/48$

for random and disordered structures, the so-called *random fractals*, in the meaning of statistical average.

The percolation network made by putting particles or bonds on a lattice with the probability p is a typical example of random fractals. The theory of percolation was initiated in 1957 by S.R. Broadbent and J.M. Hammersley [5] in connection with the diffusion of gases through porous media. Since their work, it has been widely accepted that the percolation theory describes a large number of physical and chemical phenomena such as gelation processes, transport in amorphous materials, hopping conduction in doped semiconductors, the quantum Hall effect, and many other applications. In addition, it forms the basis for studies of the flow of liquids or gases through porous media. Percolating networks thus serve as a model which helps us to understand physical properties of complex fractal structures.

For both deterministic and random fractals, it is remarkable that no characteristic length scale exists, and this is a key feature of fractal structures. In other words, fractals are defined to be objects invariant under isotropic scale transformations, i. e., uniform dilatation of the system in every spatial direction. In contrast, there exist systems which are invariant under *anisotropic* transformations. These are called *self-affine fractals*.

Determining Fractal Dimensions

There are several methods to determine fractal dimensions D_f of complex structures encountered in condensed matter physics. The following methods for obtaining the fractal dimension D_f are known to be quite efficient.

Coverage Method

The idea of coverage in the definition of the capacity dimension (see (5)) can be applied to obtain the fractal dimension D_f of material surfaces. An example is the fractality of rough surfaces or inner surfaces of porous media. The fractal nature is probed by changing the sizes of adsorbed molecules on solid surfaces. Power laws are verified by plotting the total number of adsorbed molecules versus their size r. The area of a surface can be estimated with the aid of molecules weakly adsorbed by van der Waals forces. Gas molecules are adsorbed on empty sites until the surface is uniformly covered with a layer one molecule thick. Provided that the radius r of one adsorbed molecule and the number of adsorbed molecules $N(r)$ are known, the surface area S obtained by molecules is given by

$$S(r) \propto N(r)r^2 \,. \tag{10}$$

If the surface of the adsorbate is perfectly smooth, we expect the measured area to be independent of the radius r of the probe molecules, which indicates the power law

$$N(r) \propto r^{-2} \,. \tag{11}$$

However, if the surface of the adsorbate is rough or contains pores that are small compared with r, less of the surface area S is accessible with increasing size r. For a fractal

surface with fractal dimension D_f, (11) gives the relation

$$N(r) \propto r^{-D_f} . \tag{12}$$

Box-Counting Method

Consider as an example a set of particles distributed in a space. First, we divide the space into small boxes of size r and count the number of boxes containing more than one particle, which we denote by $N(r)$. From the definition of the capacity dimension (4), the number of particle

$$N(r) \propto r^{-D_f} . \tag{13}$$

For homogeneous objects distributed in a d-dimensional space, the number of boxes of size r becomes, of course

$$N(r) \propto r^{-d} .$$

Correlation Function

The fractal dimension D_f can be obtained via the correlation function, which is the fundamental statistical quantity observed by means of X-ray, light, and neutron scattering experiments. These techniques are available to bulk materials (not surface), and is widely used in condensed matter physics. Let $\rho(\boldsymbol{r})$ be the number density of atoms at position $\boldsymbol{r}$. The density-density correlation function $G(\boldsymbol{r}, \boldsymbol{r}')$ is defined by

$$G(\boldsymbol{r}, \boldsymbol{r}') = \langle \rho(\boldsymbol{r})\rho(\boldsymbol{r}') \rangle , \tag{14}$$

where $\langle \ldots \rangle$ denotes an ensemble average. This gives the correlation of the number-density fluctuation. Provided that the distribution is isotropic, the correlation function becomes a function of only one variable, the radial distance $r = |\boldsymbol{r} - \boldsymbol{r}'|$, which is defined in spherical coordinates. Because of the translational invariance of the system on average, $\boldsymbol{r}'$ can be fixed at the coordinate origin $\boldsymbol{r}' = 0$. We can write the correlation function as

$$G(\boldsymbol{r}) = \langle \rho(\boldsymbol{r})\rho(0) \rangle . \tag{15}$$

The quantity $\langle \rho(\boldsymbol{r})\rho(0) \rangle$ is proportional to the probability that a particle exists at a distance r from another particle. This probability is proportional to the particle density $\rho(r)$ within a sphere of radius r. Since $\rho(r) \propto r^{D_f - d}$ for a fractal distribution, the correlation function becomes

$$G(\boldsymbol{r}) \propto r^{D_f - d} , \tag{16}$$

where D_f and d are the fractal and the embedded Euclidean dimensions, respectively. This relation is often used directly to determine D_f for random fractal structures.

The scattering intensity in an actual experiment is proportional to the structure factor $S(q)$, which is the Fourier transform of the correlation function $G(r)$. The structure factor is calculated from (16) as

$$S(q) = \frac{1}{V} \int_V G(r) e^{i\boldsymbol{q}\cdot\boldsymbol{r}} \mathrm{d}r \propto q^{-D_f} \tag{17}$$

where V is the volume of the system. Here $\mathrm{d}r$ is the d-dimensional volume element. Using this relation, we can determine the fractal dimension D_f from the data obtained by scattering experiments.

When applying these methods to obtain the fractal dimension D_f, we need to take care over the following point. Any fractal structures found in nature must have upper and lower length-limits for their fractality. There usually exists a *crossover* from homogeneous to fractal. Fractal properties should be observed only between these limits.

We describe in the succeeding Sections several examples of fractal structures encountered in condensed matter physics.

Polymer Chains in Solvents

Since the concept of fractal was coined by B.B. Mandelbrot in 1975, scientists reinterpreted random complex structures found in condensed matter physics in terms of fractals. They found that a lot of objects are classified as fractal structures. We show at first from polymer physics an instructive example exhibiting the fractal structure. That is an early work by P.J. Flory in 1949 on the relationship between the mean-square end-to-end distance of a polymer chain $\langle r^2 \rangle$ and the degree of polymerization N. Consider a dilute solution of separate coils in a solvent, where the total length of a flexible polymer chain with a monomer length a is Na. The simplest idealization views the polymer chain in analogy with a *Brownian motion* of a random walker. The walk is made by a succession of N steps from the origin $\boldsymbol{r} = 0$ to the end point $\boldsymbol{r}$. According to the *central limit theorem* of the probability theory, the probability to find a walker at $\boldsymbol{r}$ after N steps ($N \gg 1$) follows the *diffusion equation* and we have the expression for the probability to find a particle after N steps at $\boldsymbol{r}$

$$P_N(\boldsymbol{r}) = (2\pi N a^2/3)^{-3/2} \exp(-3r^2/2Na^2) , \tag{18}$$

where the prefactor arises from the normalization of $P_N(\boldsymbol{r})$. The mean squared distance calculated from $P_N(\boldsymbol{r})$ becomes

$$\langle \boldsymbol{r}^2 \rangle = \int r^2 P_N(\boldsymbol{r}) \mathrm{d}^3\boldsymbol{r} = Na^2 . \tag{19}$$

Then, the mean-average end-to-end distance of a polymer chain yields $R = \langle r^2 \rangle^{1/2} = N^{1/2} a$. Since the number

of polymerization N corresponds to the total mass M of a polymer chain, the use of (19) leads to the relation such as $M(R) \sim R^2$. The mass $M(R)$ can be considered as a measure of a polymer chain, the fractal dimension of this ideal chain as well as the trace of Brown motion becomes $D_f = 2$ for any d-dimensional embedding space.

The entropy of the idealized chain of the length $L = Na$ is obtained from (18) as

$$S(r) = S(0) - \frac{3r^2}{2R^2}\,, \tag{20}$$

from which the free energy $F_{el} = U - TS$ is obtained as

$$F_{el}(r) = F_{el}(0) + \frac{3k_B T r^2}{2R^2}\,. \tag{21}$$

Here U is assumed to be independent of distinct configurations of polymer chains. This is an elastic energy of an ideal chain due to entropy where F_{el} decreases as $N \to$ large. P.J. Flory added the repulsive energy term due to monomer-monomer interactions, the so-called *excluded volume effect*. This has an analogy with *self-avoiding random walk*. The contribution to the free energy is obtained by the virial expansion into the power series on the concentration $c_{int} = N/r^d$. According to the mean field theory on the repulsive term $F_{int} \propto c_{int}^2$, we have the total free-energy F such as

$$\frac{F}{k_B T} = \frac{3r^2}{2Na^2} + \frac{v(T)N^2}{r^d}\,, \tag{22}$$

where $v(T)$ is the excluded volume parameter. We can obtain a minimum of $F(r)$ at $r = R$ by differentiating $F(r)$ with respect to r such that

$$M(R) \propto R^{\frac{d+2}{3}}\,. \tag{23}$$

Here the number of polymerization N corresponds to the total mass $M(R)$ of a polymer chain. Thus, we have the fractal dimension $D_f = (d+2)/3$, in particular, $D_f = 5/3 = 1.666\ldots$ for a polymer chain in a solvent.

Aggregates and Flocs

The structures of a wide variety of flocculated colloids in suspension (called aggregates or flocs) can be described in terms of *fractals*. A colloidal suspension is a fluid containing small charged particles that are kept apart by Coulomb repulsion and kept afloat by Brownian motion. A change in the particle-particle interaction can be induced by varying the chemical composition of the solution and in this manner an aggregation process can be initiated. Aggregation processes are classified into two simple types: diffusion-limited aggregation (DLA) and diffusion-limited cluster-cluster aggregation (DLCA), where a DLA is due to the cluster-particle coalescence and a DLCA to the cluster-cluster flocculation. In most cases, actual aggregates involve a complex interplay between a variety of flocculation processes. The pioneering work was done by M.V. Smoluchowski in 1906, who formulated a kinetic theory for the irreversible aggregation of particles into clusters and further clusters combining with clusters. The inclusion of cluster-cluster aggregation makes this process distinct from the DLA process due to particle-cluster interaction. There are two distinct limiting regimes of the irreversible colloidal aggregation process: the diffusion-limited CCA (DLCA) in dilute solutions and the reaction-limited CCA (RLCA) in dense solutions. The DLCA is due to the fast process determined by the time for the clusters to encounter each other by diffusion, and the RLCA is due to the slow process since the cluster-cluster repulsion has to dominate thermal activation.

Much of our understanding on the mechanism forming aggregates or flocs has been mainly due to computer simulations. The first simulation was carried out by Vold in 1963 [23], who used the ballistic aggregation model and found that the number of particles $N(r)$ within a distance r measured from the first seed particle is given by $N(r) \sim r^{2.3}$. Though this relation surely exhibits the scaling form of (2), the applicability of this model for real systems was doubted in later years. The researches on fractal aggregates has been developed from a simulation model on DLA introduced by T.A. Witten and L.M. Sander in 1981 [26] and on the DLCA model proposed by P. Meakin in 1983 [14] and M. Kolb et al. in 1983 [11], independently. The DLA has been used to describe diverse phenomena forming fractal patterns such as electro-depositions, surface corrosions and dielectric breakdowns. In the simplest version of the DLA model for irreversible colloidal aggregation, a particle is located at an initial site $\boldsymbol{r} = 0$ as a seed for cluster formation. Another particle starts a random walk from a randomly chosen site in the spherical shell of radius r with width $dr(\ll r)$ and center $\boldsymbol{r} = 0$. As a first step, a random walk is continued until the particle contacts the seed. The cluster composed of two particles is then formed. Note that the finite-size of particles is the very reason of dendrite structures of DLA. This procedure is repeated many times, in each of which the radius r of the starting spherical shell should be much larger than the gyration radius of the cluster. If the number of particles contained in the DLA cluster is huge (typically $10^4 \sim 10^8$), the cluster generated by this process is highly branched, and forms fractal structures in the meaning of statistical average. The fractality arises from the fact that the faster growing parts of the cluster shield the other parts, which there-

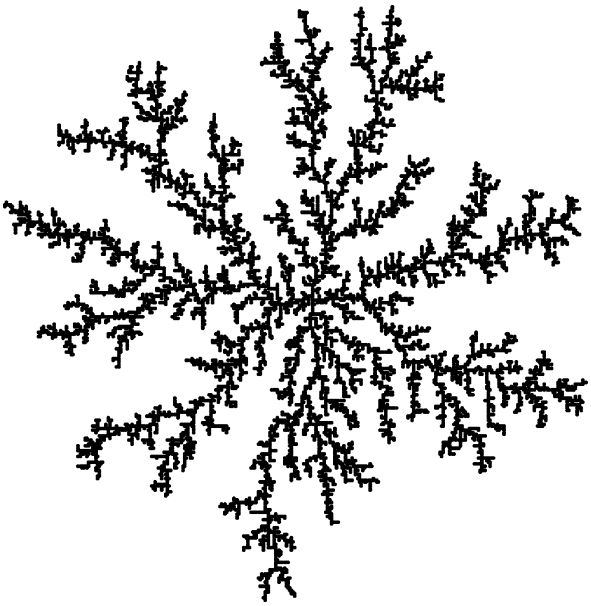

Fractal Structures in Condensed Matter Physics, Figure 5
Simulated results of a 2d diffusion-limited aggregation (DLA). The number of particles contained in this DLA cluster is 10^4

fore become less accessible to incoming particles. An arriving random walker is far more likely to attach to one of the tips of the cluster. Thus, the essence of the fractal-pattern formation arises surely from nonlinear process. Figure 5 illustrates a simulated result for a 2d DLA cluster obtained by the procedure mentioned above. The number of particles N inside a sphere of radius L ($\ll$ the gyration radius of the cluster) follows the scaling law given by

$$N \propto L^{D_f} , \tag{24}$$

where the fractal dimension takes a value of $D_f \approx 1.71$ for the 2d DLA cluster and $D_f \approx 2.5$ for the 3d DLA cluster without an underlying lattice. Note that these fractal dimensions are sensitive to the embedding lattice structure. The reason for this open structure is that a wandering molecule will settle preferentially near one of the tips of the fractal, rather than inside a cluster. Thus, different sites have different growth probabilities, which are high near the tips and decrease with increasing depth inside a cluster.

One of the most extensively studied DLA processes is the growth of metallic forms by electrochemical deposition. The scaling properties of electrodeposited metals were pointed out by R.M. Brady and R.C. Ball in 1984 for copper electrodepositions. The confirmation of the fractality for zinc metal leaves was made by M. Matsushita et al. in 1984. In their experiments [13], zinc metal leaves are grown two-dimensionally by electrodeposition. The structures clearly recover the pattern obtained by computer simulations for the DLA model proposed by T.A. Witten and L.M. Sander in 1981. Figure 6 shows a typical zinc

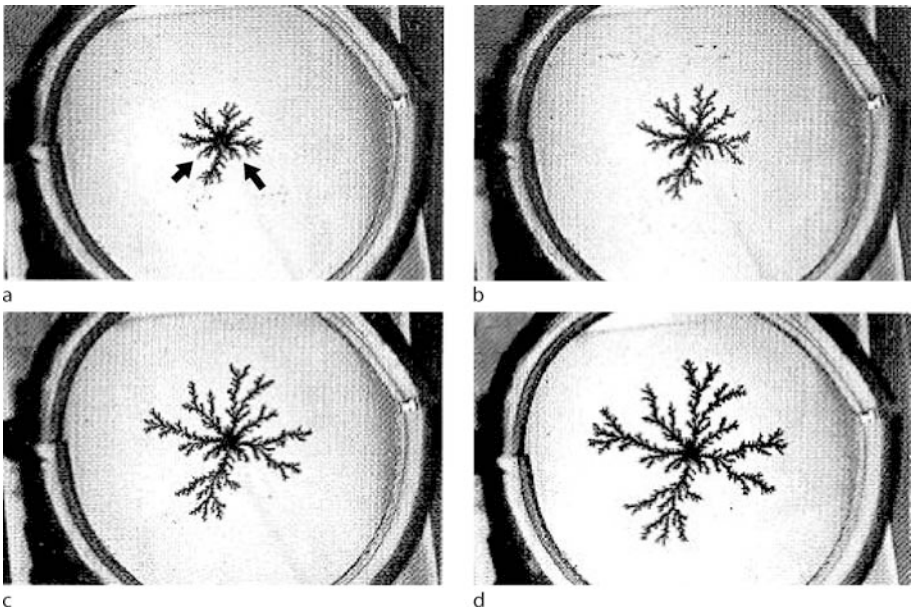

Fractal Structures in Condensed Matter Physics, Figure 6
The fractal structures of zinc metal leaves grown by electrodeposition. Photographs a–d were taken 3,5,9, and 15 min after initiating the electrolysis, respectively. After [13]

dendrite that was deposited on the cathode in one of these experiments. The fractal dimensionality $D_f = 1.66 \pm 0.33$ was obtained by computing the density-density correlation function $G(r)$ for patterns grown at applied voltages of less than 8V.

The fractality of uniformly sized gold-colloid aggregates according to the DLCA was experimentally demonstrated by D.A. Weitz in 1984 [25]. They used transmission-electron micrographs to determine the fractal dimension of this systems to be $D_f = 1.75$. They also performed quasi-elastic light-scattering experiments to investigate the dynamic characteristics of DLCA of aqueous gold colloids. They confirmed the scaling behaviors for the dependence of the mean cluster size on both time and initial concentration.

These works were performed *consciously* to examine the fractality of aggregates. There had been earlier works exhibiting the mass-size scaling relationship for actual aggregates. J.M. Beeckmans [2] pointed out in 1963 the power law behaviors by analyzing the data for aerosol and precipitated smokes in the literature (1922–1961). He used in his paper the term "aggregates-within-aggregates", implying the fractality of aggregates. However, the data available at that stage were not adequate and scattered. Therefore, this work did not provide decisive results on the fractal dimensions of aggregates. There were smarter experiments by N. Tambo and Y. Watanabe in 1967 [20], which precisely determined fractal dimensions of flocs formed in an aqueous solution. These were performed without being aware of the concept of fractals. Original works were published in Japanese. The English versions of these works were published in 1979 [21]. We discuss these works below.

Flocs generated in aqueous solutions have been the subject of numerous studies ranging from basic to applied sciences. In particular, the settling process of flocs formed in water incorporating kaolin colloids is relevant to water and wastewater treatment. The papers by N. Tambo and Y. Watanabe pioneered the discussion on the so-called *fractal approach* to floc structures; they performed their own settling experiments to clarifying the size dependences of mass densities for clay-aluminum flocs by using Stokes' law $u_r \propto \Delta\rho(r)r^2$ where $\Delta\rho$ is the difference between the densities of water and flocs taking so-small values $\Delta\rho \sim 0.01$–$0.001\ \mathrm{g/cm^3}$. Thus, the settling velocities u_r are very slow of the order of 0.001 m/sec for flocs of sizes $r \sim 0.1$ mm, which enabled them to perform precise measurements. Since flocs are very fragile aggregates, they made the settling experiments with special cautions on convection and turbulence, and by careful and intensive experiments of flocculation conditions. They confirmed from thousands of pieces of data the scaling relationship between settling velocities u_r and sizes of aggregates such as $u_r \propto r^b$. From the analysis of these data, they found the scaling relation between effective mass densities and sizes of flocs such as $\Delta\rho(r) \propto r^{-c}$, where the exponents c were found to take values from 1.25 to 1.00 depending on the aluminum-ion concentration, showing that the fractal dimensions become $D_f = 1.75$ to 2.00 with increasing aluminum-ion concentration. This is because the repulsive force between charged clay-particles is screened, and van der Waals attractive force dominates between the pair of particles. It is remarkable that these fractal dimensions D_f show excellent agreement with those determined for actual DLCA and RLCA clusters in the 1980s by using various experimental and computer simulation methods. Thus, they had found that the size dependences of mass densities of flocs are controlled by the aluminum-ion concentration dosed/suspended particle concentration, which they named the ALT ratio. These correspond

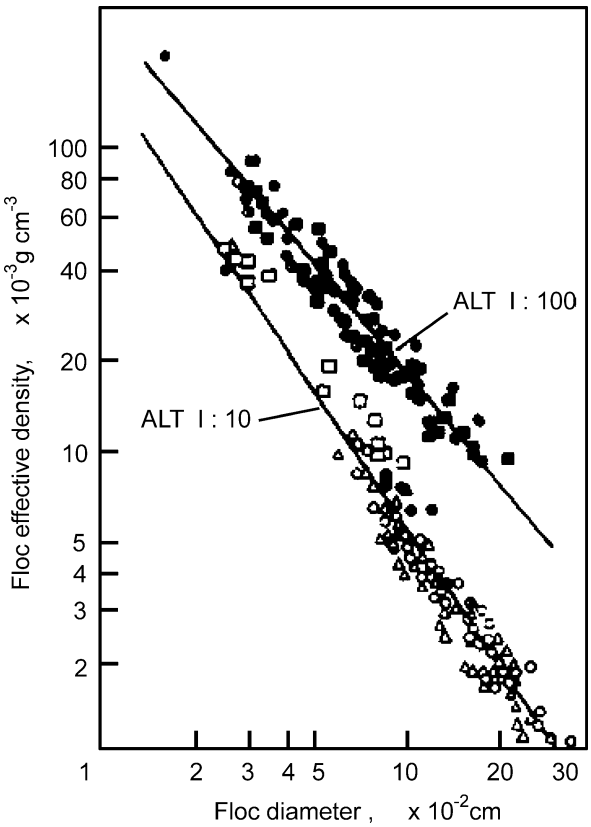

Fractal Structures in Condensed Matter Physics, Figure 7
Observed scaling relations between floc densities and their diameters where aluminum chloride is used as coagulants. After [21]

to the transition from DLCA (established now taking the value of $D_f \approx 1.78$ from computer simulations) process to the RLCA one (established at present from computer simulations as $D_f \approx 2.11$). The ALT ratio has since the publication of the paper been used in practice as a criterion for the coagulation to produce flocs with better settling properties and less sludge volume. We show their experimental data in Fig. 7, which demonstrate clearly that flocs (aggregates) are fractal.

Aerogels

Silica aerogels are extremely light materials with porosities as high as 98% and take fractal structures. The initial step in the preparation is the hydrolysis of an alkoxysilane $Si(OR)_4$, where R is CH_3 or C_2H_5. The hydrolysis produces silicon hydroxide $Si(OH)_4$ groups which polycondense into siloxane bonds –Si–O–Si–, and small particles start to grow in the solution. These particles bind to each other by *diffusion-limited cluster-cluster aggregation* (DLCA) (see Sect. "Aggregates and Flocs") until eventually they produce a disordered network filling the reaction volume. After suitable aging, if the solvent is extracted above the critical point, the open porous structure of the network is preserved and decimeter-size monolithic blocks with a range of densities from 50 to 500 kg/m^3 can be obtained. As a consequence, aerogels exhibit unusual physical properties, making them suitable for a number of practical applications, such as Cerenkov radiation detectors, supports for catalysis, or thermal insulators.

Silica aerogels possess two different length scales. One is the radius r of primary particles. The other length is the correlation length of the gel. At intermediate length scales, lying between these two length scales, the clusters possess a fractal structure and at larger length scales the gel is a homogeneous porous glass. Aerogels have a very low thermal conductivity, solid-like elasticity, and very large internal surfaces.

In elastic neutron scattering experiments, the scattering differential cross-section measures the Fourier components of spatial fluctuations in the mass density. For aerogels, the differential cross-section is the product of three factors, and is expressed by

$$\frac{d\sigma}{d\Omega} = Af^2(q)S(q)C(q) + B\,. \tag{25}$$

Here A is a coefficient proportional to the particle concentration and $f(q)$ is the primary-particle form factor. The structure factor $S(q)$ describes the correlation between particles in a cluster and $C(q)$ accounts for cluster-cluster correlations. The incoherent background is expressed by B. The structure factor $S(q)$ is proportional to the spatial Fourier transform of the density-density correlation function defined by (16), and is given by (17). Since the structure of the aerogel is fractal up to the correlation length ξ of the system and homogeneous for larger scales, the correlation function $G(r)$ is expressed by (25) for $r \ll \xi$ and $G(r) = \text{Const.}$ for $r \gg \xi$. Corresponding to this, the structure factor $S(q)$ is given by (17) for $q\xi \gg 1$, while $S(q)$ is independent of q for $q\xi \ll 1$. The wavenumber regime for which $S(q)$ becomes a constant is called the Guinier regime. The value of D_f can be deduced from the slope of the observed intensity versus momentum transfer ($q\xi \gg 1$) in a double logarithmic plot. For very large q, there exists a regime called the *Porod regime* in which the scattering intensity is proportional to q^{-4}.

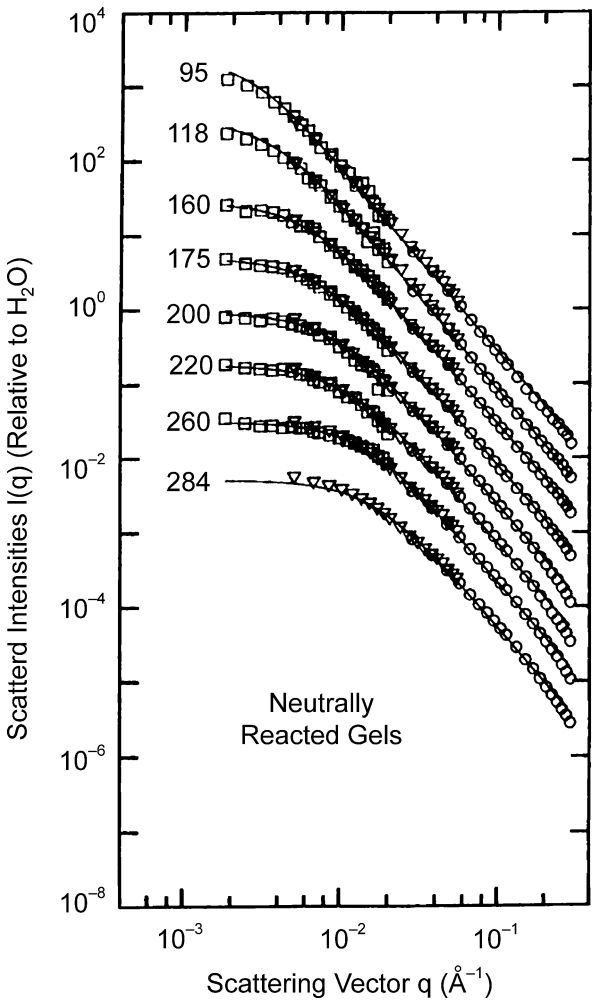

Fractal Structures in Condensed Matter Physics, Figure 8
Scattered intensities for eight neutrally reacted samples. Curves are labeled with ρ in kg/m^3. After [22]

The results in Fig. 8 by R. Vacher et al. [22] are from small-angle neutron scattering experiments on silica aerogels. The various curves are labeled by the macroscopic density ρ of the corresponding sample in Fig. 8. For example, 95 refers to a neutrally reacted sample with $\rho = 95\,\mathrm{kg/m^3}$. Solid lines represent best fits. They are presented even in the particle regime $q > 0.15\,\text{Å}^{-1}$ to emphasize that the fits do not apply in the region, particularly for the denser samples. Remarkably, D_f is independent of sample density to within experimental accuracy: $D_\mathrm{f} = 2.40 \pm 0.03$ for samples 95 to 360. The departure of $S(q)$ from the q^{-D_f} dependence at large q indicates the presence of particles with gyration radii of a few Å.

Dynamical Properties of Fractal Structures

The dynamics of fractal objects is deeply related to the time-scale problems such as diffusion, vibration and transport on fractal support. For the diffusion of a particle on any d-dimensional ordinary Euclidean space, it is well known that the mean-square displacement $\langle r^2(t)\rangle$ is proportional to the time such as $\langle r^2(t)\rangle \propto t$ for any Euclidean dimension d (see also (19)). This is called *normal diffusion*. While, on fractal supports, a particle more slowly diffuses, and the mean-square displacement follows the power law

$$\langle r^2(t)\rangle \propto t^{2/d_\mathrm{w}} , \tag{26}$$

where d_w is termed the exponent of *anomalous diffusion*. The exponents is expressed as $d_\mathrm{w} = 2 + \theta$ with a positive $\theta > 0$ (see (31)), implying that the diffusion becomes slower compared with the case of normal diffusion. This is because the inequality $2/d_\mathrm{w} < 1$ always holds. This slow diffusions on fractal supports are called *anomalous diffusion*.

The scaling relation between the length-scale and the time-scale can be easily extended to the problem of atomic vibrations of elastic fractal-networks. This is because various types of equations governing dynamics can be mapped onto the diffusion equation. This implies that both equations are governed by the same eigenvalue problem, namely, the replacement of eigenvalues $\omega \to \omega^2$ between the diffusion equation and the equation of atomic vibrations is justified. Thus, the basic properties of vibrations of fractal networks, such as the density of states, the dispersion relation and the localization/delocalization property, can be derived from the same arguments for diffusion on fractal networks. The dispersion relation between the frequency ω and the wavelength $\Lambda(\omega)$ is obtained from (26) by using the reciprocal relation $t \to \omega^{-2}$ (here the diffusion problem is mapped onto the vibrational one) and $\langle r^2(t)\rangle \to \Lambda(\omega)^{-2}$. Thus we obtain the dispersion relation for vibrational excitations on fractal networks such as

$$\omega \propto \Lambda(\omega)^{d_\mathrm{w}/2} . \tag{27}$$

If $d_\mathrm{w} = 2$, we have the ordinary dispersion relation $\omega \propto \Lambda(\omega)$ for elastic waves excited on homogeneous systems.

Consider the diffusion of a random walker on a *percolating fractal network*. How does $\langle r^2(t)\rangle$ behave in the case of fractal percolating networks? For this, P.G. de Gennes in 1976 [7] posed the problem called an ant in the labyrinth. Y. Gefen et al. in 1983 [9] gave a fundamental description of this problem in terms of a *scaling argument*. D. Ben-Avraham and S. Havlin in 1982 [3] investigated this problem in terms of Monte Carlo simulations. The work by Y. Gefen [9] triggered further developments in the dynamics of fractal systems, where the *spectral (or fracton)* dimension d_s is a key dimension for describing the dynamics of fractal networks, in addition to the fractal dimension D_f. The fractal dimension D_f characterizes how the geometrical distribution of a static structure depends on its length scale, whereas the spectral dimension d_s plays a central role in characterizing dynamic quantities on fractal networks. These dynamical properties are described in a unified way by introducing a new dynamic exponent called the *spectral or fracton* dimension defined by

$$d_\mathrm{s} = \frac{2D_\mathrm{f}}{d_\mathrm{w}} . \tag{28}$$

The term *fracton*, coined by S. Alexander and R. Orbach in 1982 [1], denotes vibrational modes peculiar to fractal structures. The characteristics of fracton modes cover a rich variety of physical implications. These modes are strongly localized in space and their localization length is of the order of their wavelengths.

We give below the explicit form of the exponent of anomalous diffusion d_w by illustrating percolation fractal networks. The mean-square displacement $\langle r^2(t)\rangle$ after a sufficiently long time t should follow the anomalous diffusion described by (26). For a finite network with a size ξ, the mean-square distance at sufficiently large time becomes $\langle r^2(t)\rangle \approx \xi^2$, so we have the diffusion coefficient for anomalous diffusion from (26) such as

$$D \propto \xi^{2-d_\mathrm{w}} . \tag{29}$$

For percolating networks, the diffusion constant D in the vicinity of the critical percolation density p_c behaves

$$D \propto (p - p_\mathrm{c})^{t-\beta} \propto \xi^{-(t-\beta)/\nu} , \tag{30}$$

where t is called the *conductivity exponent* defined by $\sigma_{dc} \sim (p - p_c)^t$, β the exponent for the percolation order parameter defined by $S(p) \propto (p - p_c)^\beta$, and ν the exponent for the correlation length defined by $\xi \propto |p - p_c|^{-\nu}$, respectively. Comparing (29) and (30), we have the relation between exponents such as

$$d_w = 2 + \frac{t - \beta}{\nu} = 2 + \theta \,. \tag{31}$$

Due to the condition $t > \beta$, and hence $\theta > 0$, implying that the diffusion becomes slow compared with the case of normal diffusion. This slow diffusion is called anomalous diffusion.

Spectral Density of States and Spectral Dimensions

The spectral density of states of atomic vibrations is the most fundamental quantity describing the dynamic properties of homogeneous or fractal systems such as specific heats, heat transport, scattering of waves and others. The simplest derivation of the spectral density of states (abbreviated, SDOS) of a homogeneous elastic system is given below. The density of states at ω is defined as the number of modes per particle, which is expressed by

$$D(\Delta\omega) = \frac{1}{\Delta\omega L^d} \,, \tag{32}$$

where $\Delta\omega$ is the frequency interval between adjacent eigenfrequencies close to ω and L is the linear size of the system. In the lowest frequency region, $\Delta\omega$ is the lowest eigenfrequency which depends on the size L. The relation between the frequency $\Delta\omega$ and L is obtained from the well-known linear dispersion relationship $\omega = \upsilon k$, where υ is the velocity of phonons (quantized elastic waves) such that

$$\Delta\omega = \frac{2\pi\upsilon}{\lambda} \propto \frac{1}{L} \,. \tag{33}$$

The substitution of (33) into (32) yields

$$D(\Delta\omega) \propto \Delta\omega^{d-1} \,. \tag{34}$$

Since this relation holds for any length scale L due to the scale-invariance property of homogeneous systems, we can replace the frequency $\Delta\omega$ by an arbitrary ω. Therefore, we obtain the conventional Debye density of states as

$$D(\omega) \propto \omega^{d-1} \,. \tag{35}$$

It should be noted that this derivation is based on the scale invariance of the system, suggesting that we can derive the SDOS for fractal networks in the same line with this treatment. Consider the SDOS of a fractal structure of size L with fractal dimension D_f. The density of states per particle at the lowest frequency $\Delta\omega$ for this system is, as in the case of (32), written as

$$D(\Delta\omega) \propto \frac{1}{L_f^D \Delta\omega} \,. \tag{36}$$

Assuming that the dispersion relation for $\Delta\omega$ corresponding to (33) is

$$\Delta\omega \propto L^{-z} \,, \tag{37}$$

we can eliminate L from (36) and obtain

$$D(\Delta\omega) \propto \Delta\omega^{D_f/z-1} \,. \tag{38}$$

The exponent z of the dispersion relation (37) is evaluated from the exponent of anomalous diffusion d_w. Considering the mapping correspondence between diffusion and atomic vibrations, we can replace $\langle r^2(t)\rangle$ and t by L^2 and $1/\Delta\omega^2$, respectively. Equation (26) can then be read as

$$L \propto \Delta\omega^{-2/d_w} \,. \tag{39}$$

The comparison of (28),(37) and (39) leads to

$$z = \frac{d_w}{2} = \frac{D_f}{d_s} \,. \tag{40}$$

Since the system has a scale-invariant fractal (self-similar) structure $\Delta\omega$, can be replaced by an arbitrary frequency ω. Hence, from (38) and (40) the SDOS for fractal networks is found to be

$$D(\omega) \propto \omega^{d_s-1} \,, \tag{41}$$

and the dispersion relation (39) becomes

$$\omega \propto L(\omega)^{-D_f/d_s} \,. \tag{42}$$

For percolating networks, the spectral dimension is obtained from (40)

$$d_s = \frac{2D_f}{2+\theta} = \frac{2\upsilon D_f}{2\upsilon + \mu - \beta} \,. \tag{43}$$

This exponent d_s is called the *fracton dimension* after S. Alexander and R. Orbach [1] or the *spectral dimension* after R. Rammal and G. Toulouse [17], hereafter we use the term *spectral dimension* for d_s. S. Alexander and R. Orbach [1] estimated the values of d_s for percolating networks on d-dimensional Euclidean lattices from the known values of the exponents D_f, υ, μ and β. They

pointed out that, while these exponents depend largely on d, the spectral dimension (fracton) dimension d_s does not.

The spectral dimension d_s can be obtained from the value of the conductivity exponent t or vice versa. In the case of percolating networks, the conductivity exponent t is related to d_s through (43), which means that the conductivity $\sigma_{dc} \sim (p - p_c)^t$ is also characterized by the spectral dimension d_s. In this sense, the spectral dimension d_s is an intrinsic exponent related to the dynamics of fractal systems. We can determine the precise values of d_s from the numerical calculations of the spectral density of states of percolation fractal networks.

The fracton SDOS for 2d, 3d, and 4d bond percolation networks at the percolation threshold $p = p_c$ are given in Fig. 9a and b, which were calculated by K. Yakubo and T. Nakayama in 1989. These were obtained by large-scale computer simulations [27]. At $p = p_c$, the correlation length diverges as $\xi \propto |p - p_c|^{-\nu}$ and the network has a fractal structure at any length scale. Therefore, fracton SDOS should be recovered in the wide frequency range $\omega_L \ll \omega \ll \omega_D$, where ω_D is the Debye cutoff frequency and ω_L is the lower cutoff determined by the system size. The SDOSs and the integrated SDOSs per atom are shown by the filled squares for a 2d bond percolation (abbreviated, BP) network at $p_c = 0.5$. The lowest frequency ω_L is quite small ($\omega \sim 10^{-5}$ for the 2d systems) as seen from the results in Fig. 9 because of the large sizes of the systems. The spectral dimension d_s is obtained as $d_s = 1.33 \pm 0.11$ from Fig. 9a, whereas data in Fig. 9b give the more precise value $d_s = 1.325 \pm 0.002$. The SDOS and the integrated SDOS for 3d BP networks at $p_c = 0.249$ are given in Fig. 9a and b by the filled triangles (middle). The spectral dimension d_s is obtained as $d_s = 1.31 \pm 0.02$ from Fig. 9a and $d_s = 1.317 \pm 0.003$ from Fig. 9b. The SDOS and the integrated SDOS of 4d BP networks at $p_c = 0.160$.

A typical mode pattern of a fracton on a 2d percolation network is shown in Fig. 10a, where the eigenmode belongs to the angular frequency $\omega = 0.04997$. To bring out the details more clearly, Fig. 10b by K. Yakubo and T. Nakayama [28] shows cross-sections of this fracton mode along the line drawn in Fig. 10a. Filled and open circles represent occupied and vacant sites in the percolation network, respectively. We see that the fracton core (the largest amplitude) possesses very clear boundaries for the edges of the excitation, with an almost step-like character and a long tail in the direction of the weak segments. It should be noted that displacements of atoms in dead ends (weakly connected portions in the percolation network) move in phase, and fall off sharply at their edges.

The spectral dimension can be obtained exactly for deterministic fractals. In the case of the d-dimensional Sierpinski gasket, the spectral dimension is given by [17]

$$d_s = \frac{2\log(d+1)}{\log(d+3)} .$$

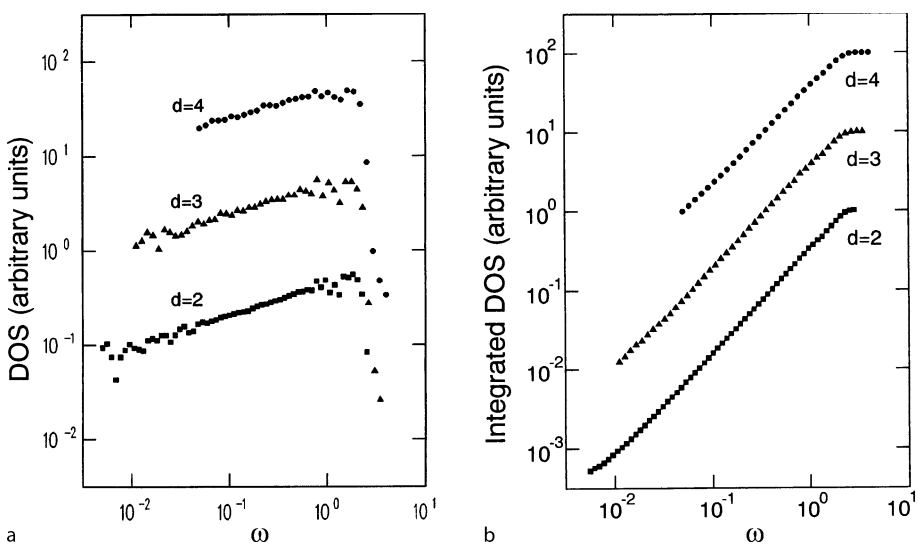

Fractal Structures in Condensed Matter Physics, Figure 9

a Spectral densities of states (SDOS) per atom for 2d, 3d, and 4d BP networks at $p = p_c$. The angular frequency ω is defined with mass units $m = 1$ and force constant $K_{ij} = 1$. The networks are formed on 1100 × 1100 (2d), 100 × 100 × 100 (3d), and 30 × 30 × 30 × 30 (4d) lattices with periodic boundary conditions, respectively. b Integrated densities of states for the same

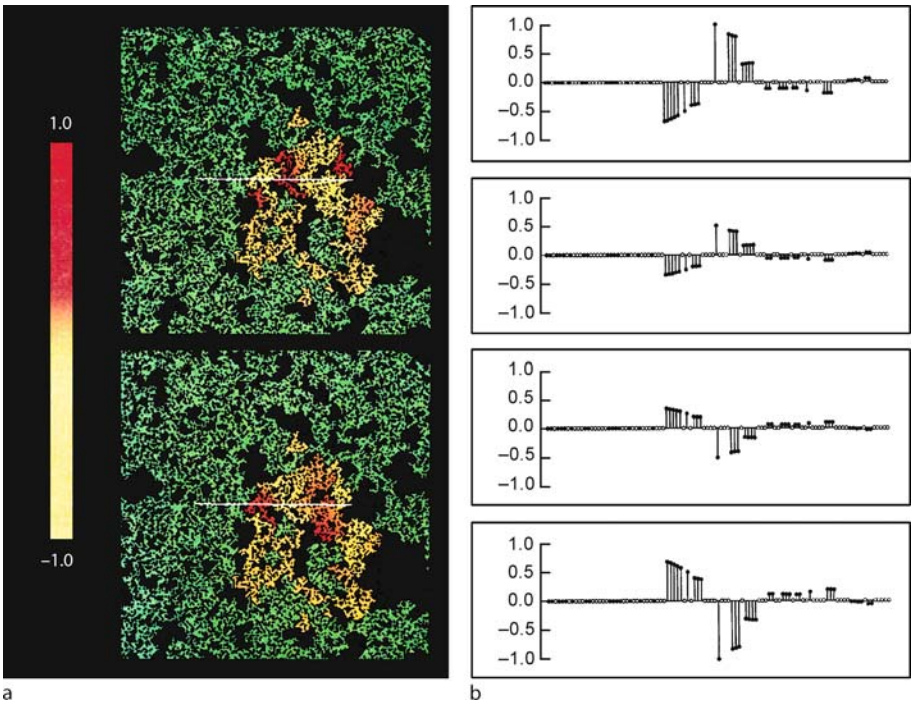

Fractal Structures in Condensed Matter Physics, Figure 10
a Typical fracton mode ($\omega = 0.04997$) on a 2d network. *Bright region* represents the large amplitude portion of the mode. b Cross-section of the fracton mode shown in a along the *white line*. The four figures are snapshots at different times. After [28]

We see from this that the upper bound for a Sierpinski gasket is $d_s = 2$ as $d \to \infty$. The spectral dimension for the Mandelbrot–Given fractal depicted is also calculated analytically as

$$d_s = \frac{2\log 8}{\log 22} = 1.345\ldots\ .$$

This value is close to those for percolating networks mentioned above, in addition to the fact that the fractal dimension $d_s = \log 8/\log 3$ of the Mandelbrot–Given fractal is close to $D_f = 91/48$ for 2d percolating networks and that the Mandelbrot–Given fractal has a structure with nodes, links, and blobs as in the case of percolating networks.

For real systems, E. Courtens et al. in 1988 [6] observed fracton excitations in aerogels by means of inelastic light scattering.

Future Directions

The significance of fractal researches in sciences is that the very idea of fractals opposes *reductionism*. Modern physics has developed by making efforts to elucidate the physical mechanisms of smaller and smaller structures such as molecules, atoms, and elementary particles. An example in condensed matter physics is the band theory of electrons in solids. Energy spectra of electrons can be obtained by incorporating group theory based on the translational and rotational symmetry of the systems. The use of this mathematical tool greatly simplifies the treatment of systems composed of 10^{22} atoms. If the energy spectrum of a unit cell *molecule* is solved, the whole energy spectrum of the solid can be computed by applying the group theory. In this context, the problem of an ordered solid is reduced to that of a unit cell. Weakly disordered systems can be handled by regarding impurities as a small perturbation to the corresponding ordered systems. However, a different approach is required for elucidating the physical properties of strongly disordered/complex systems with correlations, or of medium-scale objects, for which it is difficult to find an easily identifiable small parameter that would allow a perturbative analysis. For such systems, the concept of fractals plays an important role in building pictures of the realm of nature.

Our established knowledge on fractals is mainly due to experimental observations or computer simulations. The researches are at the phenomenological level stage, not at the intrinsic level, except for a few examples. Concerning future directions of the researches on fractals in condensed

matter physics apart from such a question as "What kinds of fractal structures are involved in condensed matter?", we should consider two directions: one is the very basic aspect such as the problem "Why are there numerous examples showing fractal structures in nature/condensed matter?" However, this type of question is hard. The kinetic growth-mechanisms of fractal systems have a rich variety of applications from the basic to applied sciences and attract much attention as one of the important subjects in non-equilibrium statistical physics and nonlinear physics. Network formation in society is one example where the kinetic growth is relevant. However, many aspects related to the mechanisms of network formations remain puzzling because arguments are at the phenomenological stage. If we compare with researches on Brownian motion as an example, the DLA researches need to advance to the stage of Einstein's intrinsic theory [8], or that of Smoluchowski [18] and H. Nyquist [15]. It is notable that the DLA is a stochastic version of the Hele–Shaw problem, the flow in composite fluids with high and low viscosities: the particles diffuse in the DLA, while the fluid pressure diffuses in Hele–Shaw flow [19]. These are deeply related to each other and involve *many open questions* for basic physics and mathematical physics.

Concerning the opposite direction, one of the important issues in fractal research is to explore practical uses of fractal structures. In fact, the characteristics of fractals are applied to many cases such as the formation of tailor-made nano-scale fractal structures, fractal-shaped antennae with much reduced sizes compared with those of ordinary antennae, and fractal molecules sensitive to frequencies in the infrared region of light.

Deep insights into fractal physics in condensed matter will open the door to new sciences and its application to technologies in the near future.

Bibliography

Primary Literature

1. Alexander S, Orbach R (1982) Density of states: Fractons. J Phys Lett 43:L625–631
2. Beeckmans JM (1963) The density of aggregated solid aerosol particles. Ann Occup Hyg 7:299–305
3. Ben-Avraham D, Havlin S (1982) Diffusion on percolation at criticality. J Phys A 15:L691–697
4. Brady RM, Ball RC (1984) Fractal growth of copper electro-deposits. Nature 309:225–229
5. Broadbent SR, Hammersley JM (1957) Percolation processes I: Crystals and mazes. Proc Cambridge Philos Soc 53:629–641
6. Courtens E, Vacher R, Pelous J, Woignier T (1988) Observation of fractons in silica aerogels. Europhys Lett 6:L691–697
7. de Gennes PG (1976) La percolation: un concept unificateur. Recherche 7:919–927
8. Einstein A (1905) Über die von der molekularkinetischen Theorie der Waerme geforderte Bewegung von in Ruhenden Fluessigkeiten Suspendierten Teilchen. Ann Phys 17:549–560
9. Gefen Y, Aharony A, Alexander S (1983) Anomalous diffusion on percolating clusters. Phys Rev Lett 50:70–73
10. Hausdorff F (1919) Dimension und Aeusseres Mass. Math Ann 79:157–179
11. Kolb M, Botel R, Jullien R (1983) Scaling of kinetically growing clusters. Phys Rev Lett 51:1123–1126
12. Mandelbrot BB, Given JA (1984) Physical properties of a new fractal model of percolation clusters. Phys Rev Lett 52: 1853–1856
13. Matsushita M, et al (1984) Fractal structures of zinc metal leaves grown by electro-deposition. Phys Rev Lett 53:286–289
14. Meakin P (1983) Formation of fractal clusters and networks by irreversible diffusion-limited aggregation. Phys Rev Lett 51:1119–1122
15. Nyquist H (1928) Termal agitation nof electric charge in conductors. Phys Rev 32:110–113
16. Perrin J (1909) Movement brownien et realite moleculaire. Ann Chim Phys 19:5–104
17. Rammal R, Toulouze G (1983) Random walks on fractal structures and percolation clusters. J Phys Lett 44:L13–L22
18. Smoluchowski MV (1906) Zur Kinematischen Theorie der Brownschen Molekular Bewegung und der Suspensionen. Ann Phys 21:756–780
19. Saffman PG, Taylor GI (1959) The penetration of a fluid into a porous medium or hele-shaw cell containing a more viscous fluid. Proc Roy Soc Lond Ser A 245:312–329
20. Tambo N, Watanabe Y (1967) Study on the density profiles of aluminium flocs I (in japanese). Suidou Kyoukai Zasshi 397:2–10; ibid (1968) Study on the density profiles of aluminium flocs II (in japanese) 410:14–17
21. Tambo N, Watanabe Y (1979) Physical characteristics of flocs I: The floc density function and aluminium floc. Water Res 13:409–419
22. Vacher R, Woignier T, Pelous J, Courtens E (1988) Structure and self-similarity of silica aerogels. Phys Rev B 37:6500–6503
23. Vold MJ (1963) Computer simulation of floc formation in a colloidal suspension. J Colloid Sci 18:684–695
24. Weitz DA, Oliveria M (1984) Fractal structures formed by kinetic aggregation of aqueous cold colloids. Phys Rev Lett 52: 1433–1436
25. Weitz DA, Huang JS, Lin MY, Sung J (1984) Dynamics of diffusion-limited kinetics aggregation. Phys Rev Lett 53:1657–1660
26. Witten TA, Sander LM (1981) Diffusion-limited aggregation, a kinetic critical phenomenon. Phys Rev Lett 47:1400–1403
27. Yakubo K, Nakayama T (1989) Direct observation of localized fractons excited on percolating nets. J Phys Soc Jpn 58: 1504–1507
28. Yakubo K, Nakayama T (1989) Fracton dynamics of percolating elastic networks: energy spectrum and localized nature. Phys Rev B 40:517–523

Books and Reviews

Barabasi AL, Stanley HE (1995) Fractal concepts in surface growth. Cambridge University Press, Cambridge

Ben-Avraham D, Havlin S (2000) Diffusion and reactions in fractals and disordered systems. Cambridge University Press, Cambridge

Bunde A, Havlin S (1996) Fractals and disordered systems. Springer, New York
de Gennes PG (1979) Scaling concepts in polymer physics. Cornell University Press, Itahca
Falconer KJ (1989) Fractal geometry: Mathematical foundations and applications. Wiley, New York
Feder J (1988) Fractals. Plenum, New York
Flory PJ (1969) Statistical mechanics of chain molecules. Inter Science, New York
Halsey TC (2000) Diffusion-limited aggregation: A model for pattern formation. Phys Today 11:36–41
Kadanoff LP (1976) Domb C, Green MS (eds) Phase transitions and critical phenomena 5A. Academic Press, New York
Kirkpatrick S (1973) Percolation and conduction. Rev Mod Phys 45:574–588
Mandelbrot BB (1979) Fractals: Form, chance and dimension. Freeman, San Francisco
Mandelbrot BB (1982) The fractal geometry of nature. Freeman, San Francisco
Meakin P (1988) Fractal aggregates. Adv Colloid Interface Sci 28:249–331
Meakin P (1998) Fractals, scaling and growth far from equilibrium. Cambridge University Press, Cambridge
Nakayama T, Yakubo K, Orbach R (1994) Dynamical properties of fractal networks: Scaling, numerical simulations, and physical realizations. Rev Mod Phys 66:381–443
Nakayama T, Yakubo K (2003) Fractal concepts in condensed matter. Springer, Heidelberg
Sahimi M (1994) Applications of percolation theory. Taylor and Francis, London
Schroeder M (1991) Fractals, chaos, power laws. W.H. Freeman, New York
Stauffer D, Aharony A (1992) Introduction to percolation theory, 2nd edn. Taylor and Francis, London
Vicsek T (1992) Fractal growth phenomena, 2nd edn. World Scientific, Singapore

Fractals and Wavelets: What Can We Learn on Transcription and Replication from Wavelet-Based Multifractal Analysis of DNA Sequences?

Alain Arneodo[1], Benjamin Audit[1], Edward-Benedict Brodie of Brodie[1], Samuel Nicolay[2], Marie Touchon[3,5], Yves d'Aubenton-Carafa[4], Maxime Huvet[4], Claude Thermes[4]

[1] Laboratoire Joliot–Curie and Laboratoire de Physique, ENS-Lyon CNRS, Lyon Cedex, France
[2] Institut de Mathématique, Université de Liège, Liège, Belgium
[3] Génétique des Génomes Bactériens, Institut Pasteur, CNRS, Paris, France
[4] Centre de Génétique Moléculaire, CNRS, Gif-sur-Yvette, France
[5] Atelier de Bioinformatique, Université Pierre et Marie Curie, Paris, France

Article Outline

Glossary
Definition of the Subject
Introduction
A Wavelet-Based Multifractal Formalism: The Wavelet Transform Modulus Maxima Method
Bifractality of Human DNA Strand-Asymmetry Profiles Results from Transcription
From the Detection of Relication Origins Using the Wavelet Transform Microscope to the Modeling of Replication in Mammalian Genomes
A Wavelet-Based Methodology to Disentangle Transcription- and Replication-Associated Strand Asymmetries Reveals a Remarkable Gene Organization in the Human Genome
Future Directions
Acknowledgments
Bibliography

Glossary

Fractal Fractals are complex mathematical objects that are invariant with respect to dilations (**self-similarity**) and therefore do not possess a characteristic length scale. Fractal objects display scale-invariance properties that can either fluctuate from point to point (**multifractal**) or be homogeneous (**monofractal**). Mathematically, these properties should hold over all scales. However, in the real world, there are necessarily lower and upper bounds over which self-similarity applies.

Wavelet transform The continuous wavelet transform (WT) is a mathematical technique introduced in the early 1980s to perform time-frequency analysis. The WT has been early recognized as a mathematical microscope that is well adapted to characterize the scale-invariance properties of fractal objects and to reveal the hierarchy that governs the spatial distribution of the singularities of multifractal measures and functions. More specifically, the WT is a space-scale analysis which consists in expanding signals in terms of wavelets that are constructed from a single function, the analyzing wavelet, by means of translations and dilations.

Wavelet transform modulus maxima method The WTMM method provides a unified statistical (thermodynamic) description of multifractal distributions including measures and functions. This method relies on the computation of partition functions from the wavelet transform skeleton defined by the wavelet transform modulus maxima (WTMM). This skeleton provides an adaptive space-scale partition of the fractal distribution under study, from which one can extract the $D(h)$ singularity spectrum as the equivalent of a thermodynamic potential (entropy). With some appropriate choice of the analyzing wavelet, one can show that the WTMM method provides a natural generalization of the classical box-counting and structure function techniques.

Compositional strand asymmetry The DNA double helix is made of two strands that are maintained together by hydrogen bonds involved in the base-pairing between Adenine (resp. Guanine) on one strand and Thymine (resp. Cytosine) on the other strand. Under no-strand bias conditions, i. e. when mutation rates are identical on the two strands, in other words when the two strands are strictly equivalent, one expects equimolarities of adenine and thymine and of guanine and cytosine on each DNA strand, a property named Chargaff's second parity rule. Compositional strand asymmetry refers to deviations from this rule which can be assessed by measuring departure from intrastrand equimolarities. Note that two major biological processes, **transcription** and **replication**, both requiring the opening of the double helix, actually break the symmetry between the two DNA strands and can thus be at the origin of compositional strand asymmetries.

Eukaryote Organisms whose cells contain a nucleus, the structure containing the genetic material arranged into chromosomes. Eukaryotes constitute one of the three domains of life, the two others, called prokaryotes (without nucleus), being the eubacteria and the archaebacteria.

Transcription Transcription is the process whereby the DNA sequence of a gene is enzymatically copied into a complementary messenger RNA. In a following step, **translation** takes place where each messenger RNA serves as a template to the biosynthesis of a specific protein.

Replication DNA replication is the process of making an identical copy of a double-stranded DNA molecule. DNA replication is an essential cellular function responsible for the accurate transmission of genetic information though successive cell generations. This process starts with the binding of initiating proteins to a DNA locus called **origin of replication**. The recruitment of additional factors initiates the bi-directional progression of two replication forks along the chromosome. In eukaryotic cells, this binding event happens at a multitude of replication origins along each chromosome from which replication propagates until two converging forks collide at a **terminus of replication**.

Chromatin Chromatin is the compound of DNA and proteins that forms the chromosomes in living cells. In eukaryotic cells, chromatin is located in the nucleus.

Histones Histones are a major family of proteins found in eukaryotic chromatin. The wrapping of DNA around a core of 8 histones forms a **nucleosome**, the first step of eukaryotic DNA compaction.

Definition of the Subject

The continuous wavelet transform (WT) is a mathematical technique introduced in signal analysis in the early 1980s [1,2]. Since then, it has been the subject of considerable theoretical developments and practical applications in a wide variety of fields. The WT has been early recognized as a mathematical microscope that is well adapted to reveal the hierarchy that governs the spatial distribution of singularities of multifractal measures [3,4,5]. What makes the WT of fundamental use in the present study is that its singularity scanning ability equally applies to singular functions than to singular measures [3,4,5,6,7,8,9,10,11]. This has led Alain Arneodo and his collaborators [12,13,14,15,16] to elaborate a unified thermodynamic description of multifractal distributions including measures and functions, the so-called Wavelet Transform Modulus Maxima (WTMM) method. By using wavelets instead of boxes, one can take advantage of the freedom in the choice of these "generalized oscillating boxes" to get rid of possible (smooth) polynomial behavior that might either mask singularities or perturb the estimation of their strength h (Hölder exponent), remedying in this way for one of the main failures of the classical multifractal methods (e. g. the box-counting algorithms in the case of measures and the structure function method in the case of functions [12,13,15,16]). The other fundamental advantage of using wavelets is that the skeleton defined by the WTMM [10,11], provides an adaptative space-scale partitioning from which one can extract the $D(h)$ singularity spectrum via the Legendre transform of the scaling exponents $\tau(q)$ (q real, positive as well as negative) of some partition functions defined from the WT skeleton. We refer the reader to Bacry et al. [13], Jaffard [17,18] for rigorous

mathematical results and to Hentschel [19] for the theoretical treatment of random multifractal functions.

Applications of the WTMM method to 1D signals have already provided insights into a wide variety of problems [20], e. g., the validation of the log-normal cascade phenomenology of fully developed turbulence [21,22,23,24] and of high-resolution temporal rainfall [25,26], the characterization and the understanding of long-range correlations in DNA sequences [27,28,29,30], the demonstration of the existence of causal cascade of information from large to small scales in financial time series [31,32], the use of the multifractal formalism to discriminate between healthy and sick heartbeat dynamics [33,34], the discovery of a Fibonacci structural ordering in 1D cuts of diffusion limited aggregates (DLA) [35,36,37,38]. The canonical WTMM method has been further generalized from 1D to 2D with the specific goal to achieve multifractal analysis of rough surfaces with fractal dimensions D_F anywhere between 2 and 3 [39,40,41]. The 2D WTMM method has been successfully applied to characterize the intermittent nature of satellite images of the cloud structure [42,43], to perform a morphological analysis of the anisotropic structure of atomic hydrogen (H_I) density in Galactic spiral arms [44] and to assist in the diagnosis in digitized mammograms [45]. We refer the reader to Arneodo et al. [46] for a review of the 2D WTMM methodology, from the theoretical concepts to experimental applications. In a recent work, Kestener and Arneodo [47] have further extended the WTMM method to 3D analysis. After some convincing test applications to synthetic 3D monofractal Brownian fields and to 3D multifractal realizations of singular cascade measures as well as their random function counterpart obtained by fractional integration, the 3D WTMM method has been applied to dissipation and enstrophy 3D numerical data issued from direct numerical simulations (DNS) of isotropic turbulence. The results so-obtained have revealed that the multifractal spatial structure of both dissipation and enstrophy fields are likely to be well described by a multiplicative cascade process clearly non-conservative. This contrasts with the conclusions of previous box-counting analysis [48] that failed to estimate correctly the corresponding multifractal spectra because of their intrinsic inability to master non-conservative singular cascade measures [47].

For many years, the multifractal description has been mainly devoted to scalar measures and functions. However, in physics as well as in other fundamental and applied sciences, fractals appear not only as deterministic or random scalar fields but also as vector-valued deterministic or random fields. Very recently, Kestener and Arneodo [49,50] have combined singular value decomposition techniques and WT analysis to generalize the multifractal formalism to vector-valued random fields. The so-called Tensorial Wavelet Transform Modulus Maxima (TWTMM) method has been applied to turbulent velocity and vorticity fields generated in $(256)^3$ DNS of the incompressible Navier–Stokes equations. This study reveals the existence of an intimate relationship $D_{\mathbf{v}}(h+1) = D_{\boldsymbol{\omega}}(h)$ between the singularity spectra of these two vector fields that are found significantly more intermittent that previously estimated from longitudinal and transverse velocity increment statistics. Furthermore, thanks to the singular value decomposition, the TWTMM method looks very promising for future simultaneous multifractal and structural (vorticity sheets, vorticity filaments) analysis of turbulent flows [49,50].

Introduction

The possible relevance of scale invariance and fractal concepts to the structural complexity of genomic sequences has been the subject of considerable increasing interest [20,51,52]. During the past fifteen years or so, there has been intense discussion about the existence, the nature and the origin of the long-range correlations (LRC) observed in DNA sequences. Different techniques including mutual information functions [53,54], auto-correlation functions [55,56], power-spectra [54,57,58], "DNA walk" representation [52,59], Zipf analysis [60,61] and entropies [62,63], were used for the statistical analysis of DNA sequences. For years there has been some permanent debate on rather struggling questions like the fact that the reported LRC might be just an artifact of the compositional heterogeneity of the genome organization [20,27,52,55,56,64,65,66,67]. Another controversial issue is whether or not LRC properties are different for protein-coding (exonic) and non-coding (intronic, intergenic) sequences [20,27,52,54,55,56,57,58,59,61,68]. Actually, there were many objective reasons for this somehow controversial situation. Most of the pioneering investigations of LRC in DNA sequences were performed using different techniques that all consisted in measuring power-law behavior of some characteristic quantity, e. g., the fractal dimension of the DNA walk, the scaling exponent of the correlation function or the power-law exponent of the power spectrum. Therefore, in practice, they all faced the same difficulties, namely finite-size effects due to the finiteness of the sequence [69,70,71] and statistical convergence issue that required some precautions when averaging over many sequences [52,65]. But beyond these practical problems, there was also a more fundamental restriction since the measurement of a unique exponent characterizing the

global scaling properties of a sequence failed to resolve multifractality [27], and thus provided very poor information upon the nature of the underlying LRC (if they were any). Actually, it can be shown that for a homogeneous (monofractal) DNA sequence, the scaling exponents estimated with the techniques previously mentioned, can all be expressed as a function of the so-called Hurst or roughness exponent H of the corresponding DNA walk landscape [20,27,52]. $H = 1/2$ corresponds to classical Brownian, i. e. uncorrelated random walk. For any other value of H, the steps (increments) are either positively correlated ($H > 1/2$: Persistent random walk) or anti-correlated ($H < 1/2$: Anti-persistent random walk).

One of the main obstacles to LRC analysis in DNA sequences is the genuine mosaic structure of these sequences which are well known to be formed of "patches" of different underlying composition [72,73,74]. When using the "DNA walk" representation, these patches appear as trends in the DNA walk landscapes that are likely to break scale-invariance [20,52,59,64,65,66,67,75,76]. Most of the techniques, e. g. the variance method, used for characterizing the presence of LRC are not well adapted to study non-stationary sequences. There have been some phenomenological attempts to differentiate local patchiness from LRC using ad hoc methods such as the so-called "min-max method" [59] and the "detrended fluctuation analysis" [77]. In previous works [27,28], the WT has been emphasized as a well suited technique to overcome this difficulty. By considering analyzing wavelets that make the WT microscope blind to low-frequency trends, any bias in the DNA walk can be removed and the existence of power-law correlations with specific scale invariance properties can be revealed accurately. In [78], from a systematic WT analysis of human exons, CDSs and introns, LRC were found in non-coding sequences as well as in coding regions somehow hidden in their inner codon structure. These results made rather questionable the model based on genome plasticity proposed at that time to account for the reported absence of LRC in coding sequences [27,28,52,54,59,68]. More recently, some structural interpretation of these LRC has emerged from a comparative multifractal analysis of DNA sequences using structural coding tables based on nucleosome positioning data [29,30]. The application of the WTMM method has revealed that the corresponding DNA chain bending profiles are monofractal (homogeneous) and that there exists two LRC regimes. In the 10–200 bp range, LRC are observed for eukaryotic sequences as quantified by a Hurst exponent value $H \simeq 0.6$ (but not for eubacterial sequences for which $H = 0.5$) as the signature of the nucleosomal structure. These LRC were shown to favor the autonomous formation of small (a few hundred bps) 2D DNA loops and in turn the propensity of eukaryotic DNA to interact with histones to form nucleosomes [79,80]. In addition, these LRC might induce some local hyperdiffusion of these loops which would be a very attractive interpretation of the nucleosomal repositioning dynamics. Over larger distances ($\gtrsim 200$ bp), stronger LRC with $H \simeq 0.8$ seem to exist in any sequence [29,30]. These LRC are actually observed in the *S. cerevisiae* nucleosome positioning data [81] suggesting that they are involved in the nucleosome organization in the so-called 30 nm chromatin fiber [82]. The fact that this second regime of LRC is also present in eubacterial sequences shows that it is likely to be a possible key to the understanding of the structure and dynamics of both eukaryotic and prokaryotic chromatin fibers. In regards to their potential role in regulating the hierarchical structure and dynamics of chromatin, the recent report [83] of sequence-induced LRC effects on the conformations of naked DNA molecules deposited onto mica surface under 2D thermodynamic equilibrium observed by Atomic Force Microscopy (AFM) is a definite experimental breakthrough.

Our purpose here is to take advantage of the availability of fully sequenced genomes to generalize the application of the WTMM method to genome-wide multifractal sequence analysis when using codings that have a clear functional meaning. According to the second parity rule [84,85], under no strand-bias conditions, each genomic DNA strand should present equimolarities of adenines A and thymines T and of guanines G and cytosines C [86,87]. Deviations from intrastrand equimolarities have been extensively studied during the past decade and the observed skews have been attributed to asymmetries intrinsic to the replication and transcription processes that both require the opening of the double helix. Actually, during these processes mutational events can affect the two strands differently and an asymmetry can result if one strand undergoes different mutations, or is repaired differently than the other strand. The existence of transcription and/or replication associated strand asymmetries has been mainly established for prokaryote, organelle and virus genomes [88,89,90,91,92,93,94]. For a long time the existence of compositional biases in eukaryotic genomes has been unclear and it is only recently that (i) the statistical analysis of eukaryotic gene introns have revealed the presence of transcription-coupled strand asymmetries [95,96,97] and (ii) the genome wide multiscale analysis of mammalian genomes has clearly shown some departure from intrastrand equimolarities in intergenic regions and further confirmed the existence of replication-associated strand asymmetries [98,99,100]. In this

manuscript, we will review recent results obtained when using the WT microscope to explore the scale invariance properties of the TA and GC skew profiles in the 22 human autosomes [98,99,100]. These results will enlighten the richness of information that can be extracted from these functional codings of DNA sequences including the prediction of 1012 putative human replication origins. In particular, this study will reveal a remarkable human gene organization driven by the coordination of transcription and replication [101].

A Wavelet-Based Multifractal Formalism

The Continuous Wavelet Transform

The WT is a space-scale analysis which consists in expanding signals in terms of *wavelets* which are constructed from a single function, the *analyzing wavelet* ψ, by means of translations and dilations. The WT of a real-valued function f is defined as [1,2]:

$$T_\psi[f](x_0, a) = \frac{1}{a}\int_{-\infty}^{+\infty} f(x)\,\psi\left(\frac{x - x_0}{a}\right)\mathrm{d}x\,, \tag{1}$$

where x_0 is the space parameter and a (> 0) the scale parameter. The analyzing wavelet ψ is generally chosen to be well localized in both space and frequency. Usually ψ is required to be of zero mean for the WT to be invertible. But for the particular purpose of singularity tracking that is of interest here, we will further require ψ to be orthogonal to low-order polynomials in Fig. 1 [7,8,9,10,11,12,13,14,15,16]:

$$\int_{-\infty}^{+\infty} x^m \psi(x)\mathrm{d}x\,, \quad 0 \le m < n_\psi\,. \tag{2}$$

As originally pointed out by Mallat and collaborators [10,11], for the specific purpose of analyzing the regularity of a function, one can get rid of the redundancy of the WT by concentrating on the WT skeleton defined by its modulus maxima only. These maxima are defined, at each scale a, as the local maxima of $|T_\psi[f](x, a)|$ considered as a function of x. As illustrated in Figs. 2e, 2f, these WTMM are disposed on connected curves in the space-scale (or time-scale) half-plane, called *maxima lines*. Let us define $\mathcal{L}(a_0)$ as the set of all the maxima lines that exist at the scale a_0 and which contain maxima at any scale $a \le a_0$. An important feature of these maxima lines, when analyzing singular functions, is that there is at least one maxima line pointing towards each singularity [10,11,16].

Scanning Singularities with the Wavelet Transform Modulus Maxima

The strength of the singularity of a function f at point x_0 is given by the *Hölder* exponent, i. e., the largest exponent such that there exists a polynomial $P_n(x - x_0)$ of order $n < h(x_0)$ and a constant $C > 0$, so that for any point x in a neighborhood of x_0, one has [7,8,9,10,11,13,16]:

$$|f(x) - P_n(x - x_0)| \le C\,|x - x_0|^h\,. \tag{3}$$

If f is n times continuously differentiable at the point x_0, then one can use for the polynomial $P_n\,(x - x_0)$, the order-n Taylor series of f at x_0 and thus prove that $h(x_0) > n$. Thus $h(x_0)$ measures how irregular the function f is at the point x_0. The higher the exponent $h(x_0)$, the more regular the function f.

The main interest in using the WT for analyzing the regularity of a function lies in its ability to be blind to polynomial behavior by an appropriate choice of the analyzing wavelet ψ. Indeed, let us assume that according to Eq. (3), f has, at the point x_0, a local scaling (Hölder) exponent $h(x_0)$; then, assuming that the singularity is not oscillating [11,102,103], one can easily prove that the local behavior of f is mirrored by the WT which locally behaves like [7,8,9,10,11,12,13,14,15,16,17,18]:

$$T_\psi[f](x_0, a) \sim a^{h(x_0)}\,, \quad a \to 0^+\,, \tag{4}$$

provided $n_\psi > h(x_0)$, where n_ψ is the number of vanishing moments of ψ (Eq. (2)). Therefore one can extract the exponent $h(x_0)$ as the slope of a log-log plot of the WT amplitude versus the scale a. On the contrary, if one chooses $n_\psi < h(x_0)$, the WT still behaves as a power-law but with a scaling exponent which is n_ψ:

$$T_\psi[f](x_0, a) \sim a^{n_\psi}\,, \quad a \to 0^+\,. \tag{5}$$

Thus, around a given point x_0, the faster the WT decreases when the scale goes to zero, the more regular f is around that point. In particular, if $f \in C^\infty$ at x_0 ($h(x_0) = +\infty$), then the WT scaling exponent is given by n_ψ, i. e. a value which is dependent on the shape of the analyzing wavelet. According to this observation, one can hope to detect the points where f is smooth by just checking the scaling behavior of the WT when increasing the order n_ψ of the analyzing wavelet [12,13,14,15,16].

Remark 1 A very important point (at least for practical purpose) raised by Mallat and Hwang [10] is that the local scaling exponent $h(x_0)$ can be equally estimated by looking at the value of the WT modulus along a maxima line converging towards the point x_0. Indeed one can prove that both Eqs. (4) and (5) still hold when following a maxima line from large down to small scales [10,11].

The Wavelet Transform Modulus Maxima Method

As originally defined by Parisi and Frisch [104], the multifractal formalism of multi-affine functions amounts to compute the so-called *singularity spectrum* $D(h)$ defined as the Hausdorff dimension of the set where the Hölder exponent is equal to h [12,13,16]:

$$D(h) = \dim_H \{x,\ h(x) = h\}\ , \tag{6}$$

where h can take, a priori, positive as well as negative real values (e. g., the Dirac distribution $\delta(x)$ corresponds to the Hölder exponent $h(0) = -1$) [17].

A natural way of performing a multifractal analysis of fractal functions consists in generalizing the "classical" multifractal formalism [105,106,107,108,109] using wavelets instead of boxes. By taking advantage of the freedom in the choice of the "generalized oscillating boxes" that are the wavelets, one can hope to get rid of possible smooth behavior that could mask singularities or perturb the estimation of their strength h. But the major difficulty with respect to box-counting techniques [48,106,110,111,112] for singular measures, consists in defining a covering of the support of the singular part of the function with our set of wavelets of different sizes. As emphasized in [12,13,14,15,16], the branching structure of the WT skeletons of fractal functions in the (x, a) half-plane enlightens the hierarchical organization of their singularities (Figs. 2e, 2f). The WT skeleton can thus be used as a guide to position, at a considered scale a, the oscillating boxes in order to obtain a partition of the singularities of f. The wavelet transform modulus maxima (WTMM) method amounts to compute the following partition function in terms of WTMM coefficients [12,13,14,15,16]:

$$Z(q, a) = \sum_{l \in \mathcal{L}(a)} \Big(\sup_{\substack{(x,a') \in l \\ a' \le a}} \big| T_\psi[f](x, a') \big| \Big)^q\ , \tag{7}$$

where $q \in \mathbb{R}$ and the sup can be regarded as a way to define a scale adaptative "Hausdorff-like" partition. Now from the deep analogy that links the multifractal formalism to thermodynamics [12,113], one can define the exponent $\tau(q)$ from the power-law behavior of the partition function:

$$Z(q, a) \sim a^{\tau(q)}\ , \quad a \to 0^+\ , \tag{8}$$

where q and $\tau(q)$ play respectively the role of the inverse temperature and the free energy. The main result of this wavelet-based multifractal formalism is that in place of the energy and the entropy (i. e. the variables conjugated to q and τ), one has h, the Hölder exponent, and $D(h)$, the singularity spectrum. This means that the singularity spectrum of f can be determined from the Legendre transform of the partition function scaling exponent $\tau(q)$ [13,17,18]:

$$D(h) = \min_q (qh - \tau(q))\ . \tag{9}$$

From the properties of the Legendre transform, it is easy to see that *homogeneous* monofractal functions that involve singularities of unique Hölder exponent $h = \partial\tau/\partial q$, are characterized by a $\tau(q)$ spectrum which is a *linear* function of q. On the contrary, a *nonlinear* $\tau(q)$ curve is the signature of nonhomogeneous functions that exhibit *multifractal* properties, in the sense that the Hölder exponent $h(x)$ is a fluctuating quantity that depends upon the spatial position x.

Defining Our Battery of Analyzing Wavelets

There are almost as many analyzing wavelets as applications of the continuous WT [3,4,5,12,13,14,15,16]. In the present work, we will mainly used the class of analyzing wavelets defined by the successive derivatives of the Gaussian function:

$$g^{(N)}(x) = \frac{\mathrm{d}^N}{\mathrm{d}x^N} \mathrm{e}^{-x^2/2}\ , \tag{10}$$

for which $n_\psi = N$ and more specifically $g^{(1)}$ and $g^{(2)}$ that are illustrated in Figs. 1a, 1b.

Remark 2 The WT of a signal f with $g^{(N)}$ (Eq. (10)) takes the following simple expression:

$$\begin{aligned} T_{g^{(N)}}[f](x, a) &= \frac{1}{a} \int_{-\infty}^{+\infty} f(y) g^{(N)}\left(\frac{y - x}{a}\right) \mathrm{d}y\ , \\ &= a^N \frac{\mathrm{d}^N}{\mathrm{d}x^N} T_{g^{(0)}}[f](x, a)\ . \end{aligned} \tag{11}$$

Equation (11) shows that the WT computed with $g^{(N)}$ at scale a is nothing but the Nth derivative of the signal $f(x)$ smoothed by a dilated version $g^{(0)}(x/a)$ of the Gaussian function. This property is at the heart of various applications of the WT microscope as a very efficient multi-scale singularity tracking technique [20].

With the specific goal of disentangling the contributions to the nucleotide composition strand asymmetry coming respectively from transcription and replication processes, we will use in Sect. "A Wavelet-Based Methodology to Disentangle Transcription- and Replication-Associated Strand Asymmetries Reveals a Remarkable Gene Organization in the Human Genome", an adapted analyzing wavelet of the following form (Fig. 1c) [101,114]:

$$\begin{aligned} \phi_R(x) &= -\left(x - \frac{1}{2}\right)\ , \quad \text{for } x \in \left[-\frac{1}{2}, \frac{1}{2}\right] \\ &= 0 \quad \text{elsewhere}\ . \end{aligned} \tag{12}$$

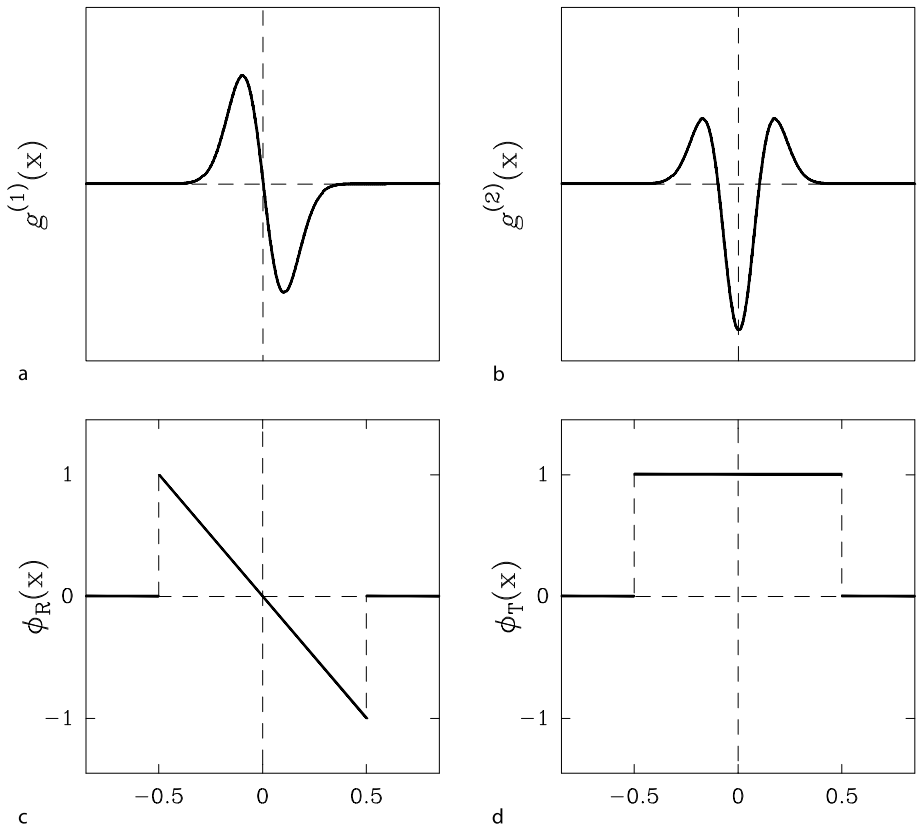

Fractals and Wavelets: What Can We Learn on Transcription and Replication …?, Figure 1
Set of analyzing wavelets $\psi(x)$ that can be used in Eq. (1). a $g^{(1)}$ and b $g^{(2)}$ as defined in Eq. (10). c ϕ_R as defined in Eq. (12), that will be used in Sect. "A Wavelet-Based Methodology to Disentangle Transcription- and Replication-Associated Strand Asymmetries Reveals a Remarkable Gene Organization in the Human Genome" to detect replication domains. d Box function ϕ_T that will be used in Sect. "A Wavelet-Based Methodology to Disentangle Transcription- and Replication-Associated Strand Asymmetries Reveals a Remarkable Gene Organization in the Human Genome" to model step-like skew profiles induced by transcription

By performing multi-scale pattern recognition in the (space, scale) half-plane with this analyzing wavelet, we will be able to define replication domains bordered by putative replication origins in the human genome and more generally in mammalian genomes [101,114].

Test Applications of the WTMM Method on Monofractal and Multifractal Synthetic Random Signals

This section is devoted to test applications of the WTMM method to random functions generated either by *additive* models like fractional Brownian motions [115] or by *multiplicative* models like random $\mathcal{W}$-cascades on wavelet dyadic trees [21,22,116,117]. For each model, we first wavelet transform 1000 realizations of length $L = 65\,536$ with the first order ($n_\psi = 1$) analyzing wavelet $g^{(1)}$. From the WT skeletons defined by the WTMM, we compute the mean partition function (Eq. (7)) from which we extract the annealed $\tau(q)$ (Eq. (8)) and, in turn, $D(h)$ (Eq. (9)) multifractal spectra. We systematically test the robustness of our estimates with respect to some change of the shape of the analyzing wavelet, in particular when increasing the number n_ψ of zero moments, going from $g^{(1)}$ to $g^{(2)}$ (Eq. (10)).

Fractional Brownian Signals Since its introduction by Mandelbrot and van Ness [115], the fractional Brownian motion (fBm) B_H has become a very popular model in signal and image processing [16,20,39]. In 1D, fBm has proved useful for modeling various physical phenomena with long-range dependence, e. g., "$1/f$" noises. The fBm exhibits a power spectral density $S(k) \sim 1/k^\beta$, where the spectral exponent $\beta = 2H + 1$ is related to the Hurst exponent H. fBm has been extensively used as test stochas-

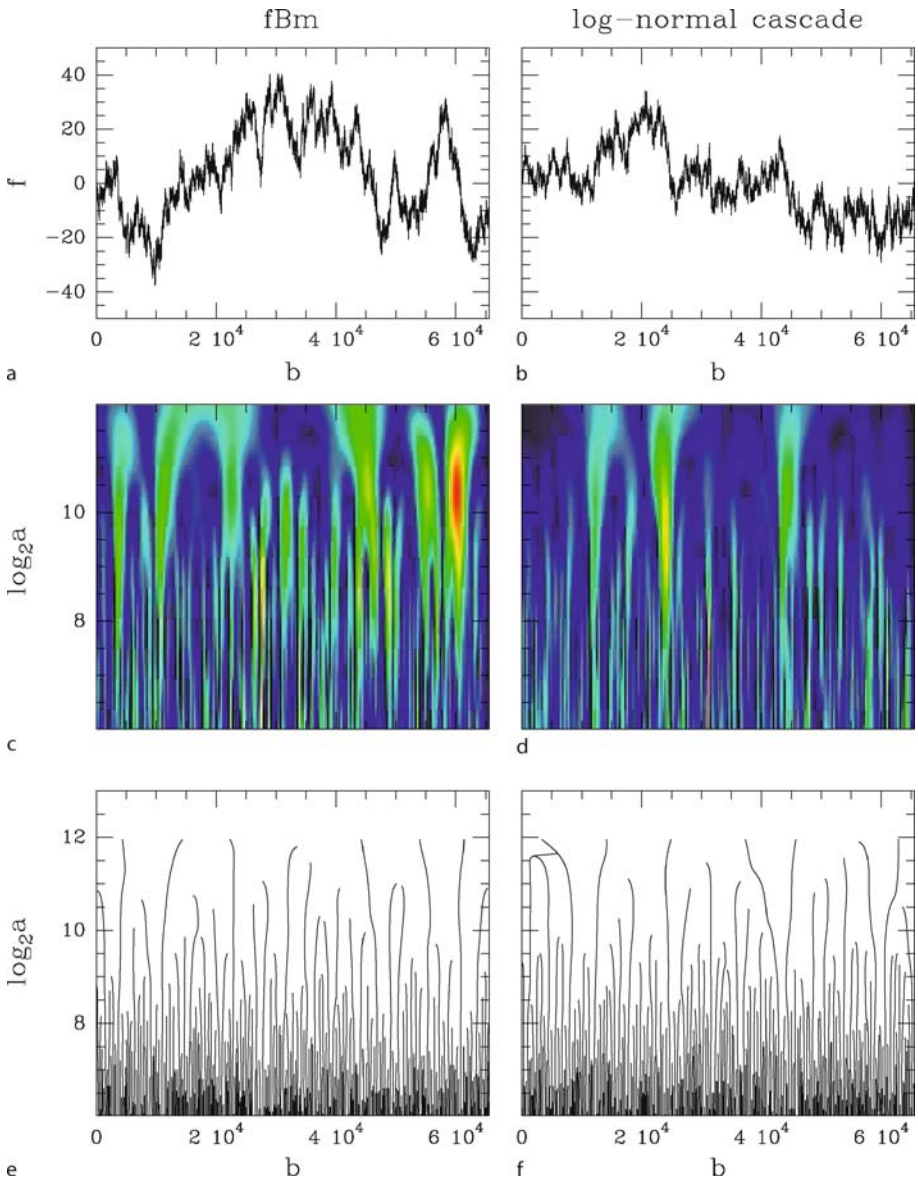

Fractals and Wavelets: What Can We Learn on Transcription and Replication …?, Figure 2

WT of monofractal and multifractal stochastic signals. *Fractional Brownian motion*: a a realization of $B_{1/3}$ ($L = 65\,536$); c WT of $B_{1/3}$ as coded, independently at each scale a, using 256 colors from *black* ($|T_\psi| = 0$) to *red* ($\max_b |T_\psi|$); e WT skeleton defined by the set of all the maxima lines. *Log-normal random $\mathcal{W}$-cascades*: b a realization of the log-normal $\mathcal{W}$-cascade model ($L = 65\,536$) with the following parameter values $m = -0.355 \ln 2$ and $\sigma^2 = 0.02 \ln 2$ (see [116]); d WT of the realization in b represented with the same color coding as in c; f WT skeleton. The analyzing wavelet is $g^{(1)}$ (see Fig. 1a)

tic signals for Hurst exponent measurements. In Figs. 2, 3 and 4, we report the results of a statistical analysis of fBm's using the WTMM method [12,13,14,15,16]. We mainly concentrate on $B_{1/3}$ since it has a $k^{-5/3}$ power-spectrum similar to the spectrum of the multifractal stochastic signal we will study next. Actually, our goal is to demonstrate that, where the power spectrum analysis fails, the WTMM method succeeds in discriminating unambiguously between these two fractal signals. The numerical signals were generated by filtering uniformly generated pseudo-random noise in Fourier space in order to have the required $k^{-5/3}$ spectral density. A $B_{1/3}$ fractional Brownian

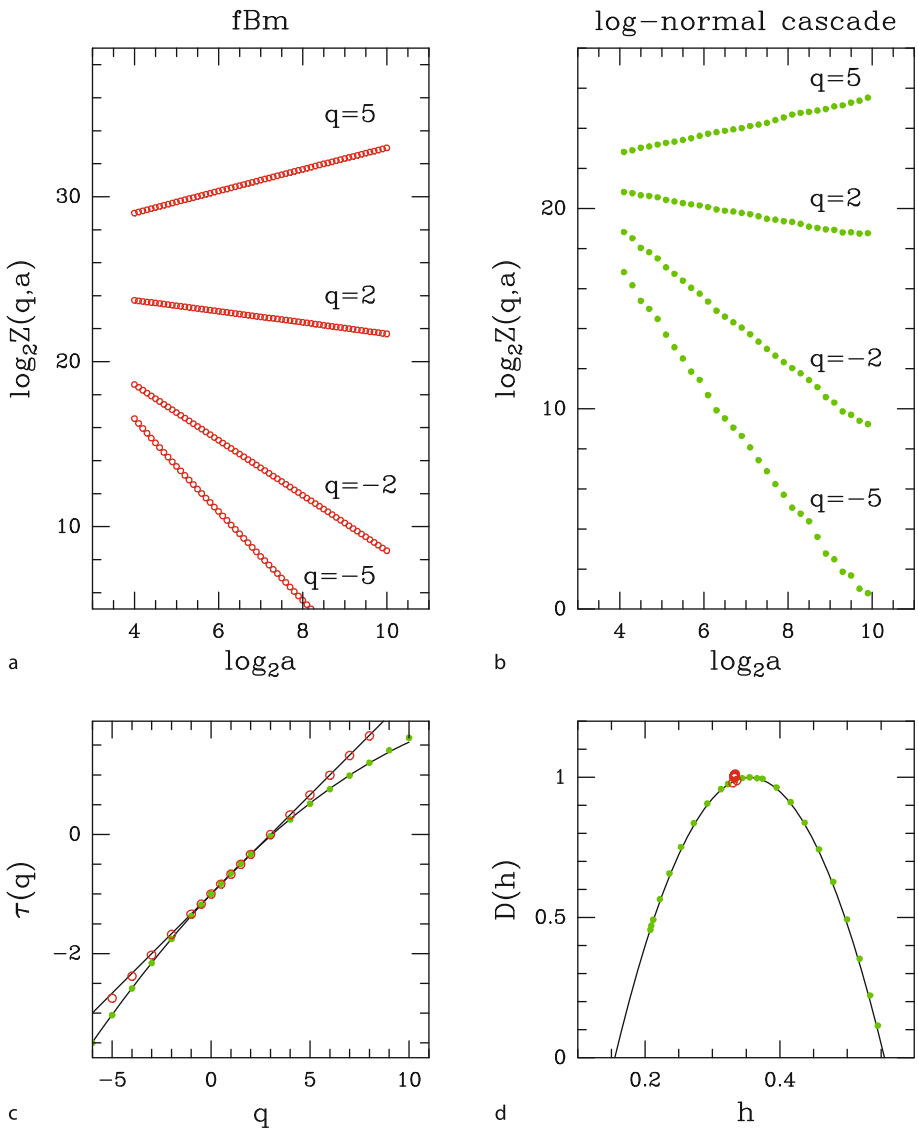

Fractals and Wavelets: What Can We Learn on Transcription and Replication …?, Figure 3

Determination of the $\tau(q)$ and $D(h)$ multifractal spectra of fBm $B_{1/3}$ (*red circles*) and log-normal random $\mathcal{W}$-cascades (*green dots*) using the WTMM method. a $\log_2 Z(q, a)$ vs. $\log_2 a$: $B_{1/3}$. b $\log_2 Z(q, a)$ vs. $\log_2 a$: Log-normal $\mathcal{W}$-cascades with the same parameters as in Fig. 2b. c $\tau(q)$ vs. q; the *solid lines* correspond respectively to the theoretical spectra (13) and (16). d $D(h)$ vs. h; the *solid lines* correspond respectively to the theoretical predictions (14) and (17). The analyzing wavelet is $g^{(1)}$. The reported results correspond to annealed averaging over 1000 realizations of $L = 65\,536$

trail is shown in Fig. 2a. Figure 2c illustrates the WT coded, independently at each scale a, using 256 colors. The analyzing wavelet is $g^{(1)}$ ($n_\psi = 1$). Figure 3a displays some plots of $\log_2 Z(q, a)$ versus $\log_2(a)$ for different values of q, where the partition function $Z(q, a)$ has been computed on the WTMM skeleton shown in Fig. 2e, according to the definition (Eq. (7)). Using a linear regression fit, we then obtain the slopes $\tau(q)$ of these graphs. As shown in Fig. 3c, when plotted versus q, the data for the exponents $\tau(q)$ consistently fall on a straight line that is remarkably fitted by the theoretical prediction:

$$\tau(q) = qH - 1\,, \tag{13}$$

with $H = 1/3$. From the Legendre transform of this linear $\tau(q)$ (Eq. (9)), one gets a $D(h)$ singularity spectrum that

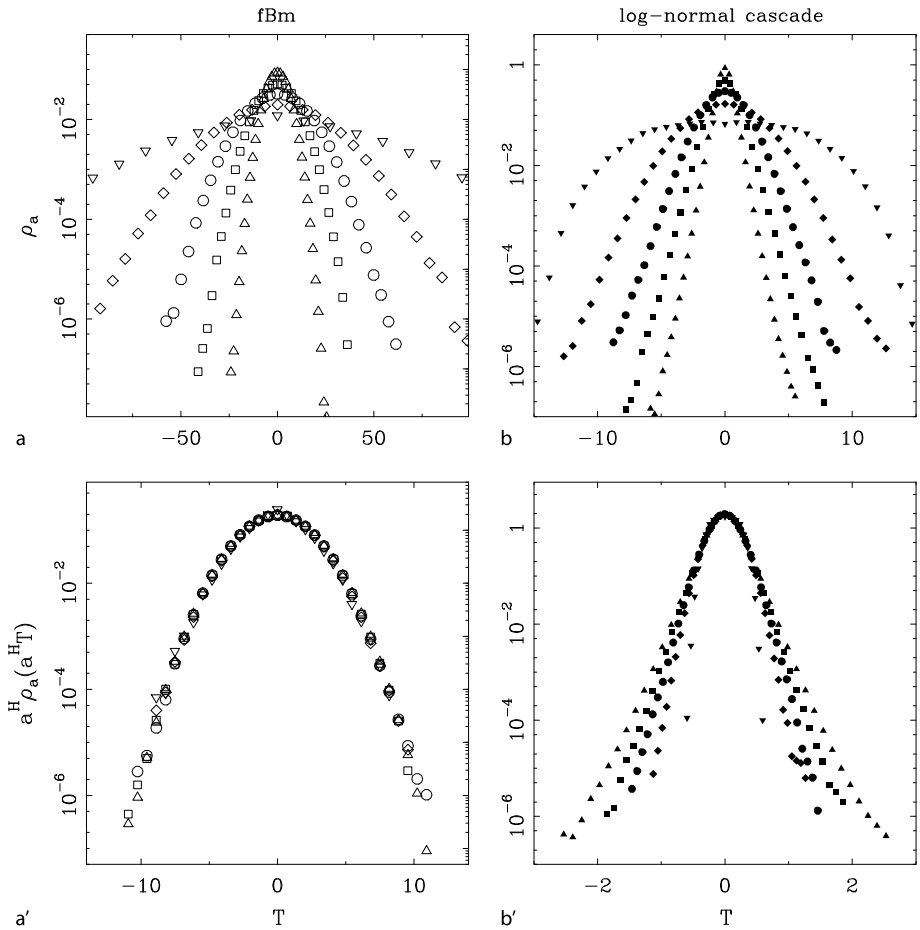

Fractals and Wavelets: What Can We Learn on Transcription and Replication …?, Figure 4
Probability distribution functions of wavelet coefficient values of fBm $B_{1/3}$ (*open symbols*) and log-normal random $\mathcal{W}$-cascades (*filled symbols*) with the same parameters as in Fig. 2b. a ρ_a vs. $T_{g^{(1)}}$ for the set of scales $a = 10$ (△), 50 (□), 100 (○), 1000 (◇), 9000 (▽); a′ $a^H \rho_a(a^H T_{g^{(1)}})$ vs. $T_{g^{(1)}}$ with $H = 1/3$; The symbols have the same meaning as in a. b ρ_a vs. $T_{g^{(1)}}$ for the set of scales $a = 10$ (▲), 50 (■), 100 (●), 1000 (◆), 9000 (▼); (b′) $a^H \rho_a(a^H T_{g^{(1)}})$ vs. $T_{g^{(1)}}$ with $H = -m/\ln 2 = 0.355$. The analyzing wavelet is $g^{(1)}$ (Fig. 1a)

reduces to a single point:

$$D(h) = 1 \quad \text{if } h = H\,, \qquad = -\infty \quad \text{if } h \neq H\,. \tag{14}$$

Thus, as expected theoretically [16,115], one finds that the fBm $B_{1/3}$ is a nowhere differentiable homogeneous fractal signal with a unique Hölder exponent $h = H = 1/3$. Note that similar good estimates are obtained when using analyzing wavelets of different order (e. g. $g^{(2)}$), and this whatever the value of the index H of the fBm [12,13,14,15,16].

Within the perspective of confirming the monofractality of fBm's, we have studied the probability density function (pdf) of wavelet coefficient values $\rho_a(T_{g^{(1)}}(., a))$, as computed at a fixed scale a in the fractal scaling range. According to the monofractal scaling properties, one expects these pdfs to satisfy the self-similarity relationship [20,27,28]:

$$a^H \rho_a(a^H T) = \rho(T)\,, \tag{15}$$

where $\rho(T)$ is a "universal" pdf (actually the pdf obtained at scale $a = 1$) that does not depend on the scale parameter a. As shown in Figs. 4a, 4a′ for $B_{1/3}$, when plotting $a^H \rho_a(a^H T)$ vs. T, all the ρ_a curves corresponding to different scales (Fig. 4a) remarkably collapse on a unique curve when using a unique exponent $H = 1/3$ (Fig. 4a′). Furthermore the so-obtained universal curve cannot be distinguished from a parabola in semi-log representation as the signature of the monofractal Gaussian statistics of fBm fluctuations [16,20,27].

Random $\mathcal{W}$-Cascades Multiplicative cascade models have enjoyed increasing interest in recent years as the paradigm of multifractal objects [16,19,48,105,107,108,118]. The notion of cascade actually refers to a self-similar process whose properties are defined multiplicatively from coarse to fine scales. In that respect, it occupies a central place in the statistical theory of turbulence [48,104]. Originally, the concept of self-similar cascades was introduced to model multifractal measures (e. g. dissipation or enstrophy) [48]. It has been recently generalized to the construction of scale-invariant signals (e. g. longitudinal velocity, pressure, temperature) using orthogonal wavelet basis [116,119]. Instead of redistributing the measure over sub-intervals with multiplicative weights, one allocates the wavelet coefficients in a multiplicative way on the dyadic grid. This method has been implemented to generate multifractal functions (with weights W) from a given deterministic or probabilistic multiplicative process. Along the line of the modeling of fully developed turbulent signals by log-infinitely divisible multiplicative processes [120,121], we will mainly concentrate here on the log-normal $\mathcal{W}$-cascades in order to calibrate the WTMM method. If m and σ^2 are respectively the mean and the variance of $\ln W$ (where W is a multiplicative random variable with log-normal probability distribution), then, as shown in [116], a straightforward computation leads to the following $\tau(q)$ spectrum:

$$\begin{aligned}\tau(q) &= -\log_2\langle W^q\rangle - 1\,, \quad \forall q \in \mathbb{R} \\ &= -\frac{\sigma^2}{2\ln 2}q^2 - \frac{m}{\ln 2}q - 1\,,\end{aligned} \tag{16}$$

where $\langle\ldots\rangle$ means ensemble average. The corresponding $D(h)$ singularity spectrum is obtained by Legendre transforming $\tau(q)$ (Eq. (9)):

$$D(h) = -\frac{(h + m/\ln 2)^2}{2\sigma^2/\ln 2} + 1\,. \tag{17}$$

According to the convergence criteria established in [116], m and σ^2 have to satisfy the conditions: $m < 0$ and $|m|/\sigma > \sqrt{2\ln 2}$. Moreover, by solving $D(h) = 0$, one gets the following bounds for the support of the $D(h)$ singularity spectrum: $h_{\min} = -m/\ln 2 - (\sqrt{2}\sigma)/\sqrt{\ln 2}$ and $h_{\max} = -m/\ln 2 + (\sqrt{2}\sigma)/\sqrt{\ln 2}$.

In Fig. 2b is illustrated a realization of a log-normal $\mathcal{W}$-cascade for the parameter values $m = -0.355\ln 2$ and $\sigma^2 = 0.02\ln 2$. The corresponding WT and WT skeleton as computed with $g^{(1)}$ are shown in Figs. 2d and 2f respectively. The results of the application of the WTMM method are reported in Fig. 3. As shown in Fig. 3b, when plotted versus the scale parameter a in a logarithmic representation, the annealed average of the partition functions $Z(q, a)$ displays a well defined scaling behavior over a range of scales of about 5 octaves. Note that scaling of quite good quality is found for a rather wide range of q values: $-5 \le q \le 10$. When processing to a linear regression fit of the data over the first four octaves, one gets the $\tau(q)$ spectrum shown in Fig. 3c. This spectrum is clearly a nonlinear function of q, the hallmark of multifractal scaling. Moreover, the numerical data are in remarkable agreement with the theoretical quadratic prediction (Eq. (16)). Similar quantitative agreement is observed on the $D(h)$ singularity spectrum in Fig. 3d which displays a single humped parabola shape that characterizes intermittent fluctuations corresponding to Hölder exponents values ranging from $h_{\min} = 0.155$ to $h_{\max} = 0.555$. Unfortunately, to capture the strongest and the weakest singularities, one needs to compute the $\tau(q)$ spectrum for very large values of $|q|$. This requires the processing of many more realizations of the considered log-normal random $\mathcal{W}$-cascade. The multifractal nature of log-normal $\mathcal{W}$-cascade realizations is confirmed in Figs. 4b, 4b′ where the self-similarity relationship (Eq. (15)) is shown not to apply. Actually there does not exist a H value allowing to superimpose onto a single curve the WT pdfs computed at different scales.

The test applications reported in this section demonstrate the ability of the WTMM method to resolve multifractal scaling of 1D signals, a hopeless task for classical power spectrum analysis. They were used on purpose to calibrate and to test the reliability of our methodology, and of the corresponding numerical tools, with respect to finite-size effects and statistical convergence.

Bifractality of Human DNA Strand-Asymmetry Profiles Results from Transcription

During genome evolution, mutations do not occur at random as illustrated by the diversity of the nucleotide substitution rate values [122,123,124,125]. This non-randomness is considered as a by-product of the various DNA mutation and repair processes that can affect each of the two DNA strands differently. Asymmetries of substitution rates coupled to transcription have been mainly observed in prokaryotes [88,89,91], with only preliminary results in eukaryotes. In the human genome, excess of T was observed in a set of gene introns [126] and some large-scale asymmetry was observed in human sequences but they were attributed to replication [127]. Only recently, a comparative analysis of mammalian sequences demonstrated a transcription-coupled excess of G+T over A+C in the coding strand [95,96,97]. In contrast to the substitution

biases observed in bacteria presenting an excess of C→T transitions, these asymmetries are characterized by an excess of purine (A→G) transitions relatively to pyrimidine (T→C) transitions. These might be a by-product of the transcription-coupled repair mechanism acting on uncorrected substitution errors during replication [128]. In this section, we report the results of a genome-wide multifractal analysis of strand-asymmetry DNA walk profiles in the human genome [129]. This study is based on the computation of the TA and GC skews in non-overlapping 1 kbp windows:

$$S_{\mathrm{TA}} = \frac{n_{\mathrm{T}} - n_{\mathrm{A}}}{n_{\mathrm{T}} + n_{\mathrm{A}}}\,, \quad S_{\mathrm{GC}} = \frac{n_{\mathrm{G}} - n_{\mathrm{C}}}{n_{\mathrm{G}} + n_{\mathrm{C}}}\,, \tag{18}$$

where n_{A}, n_{C}, n_{G} and n_{T} are respectively the numbers of A, C, G and T in the windows. Because of the observed correlation between the TA and GC skews, we also considered the total skew

$$S = S_{\mathrm{TA}} + S_{\mathrm{GC}}\,. \tag{19}$$

From the skews $S_{\mathrm{TA}}(n)$, $S_{\mathrm{GC}}(n)$ and $S(n)$, obtained along the sequences, where n is the position (in kbp units) from the origin, we also computed the cumulative skew profiles (or skew walk profiles):

$$\Sigma_{\mathrm{TA}}(n) = \sum_{j=1}^{n} S_{\mathrm{TA}}(j)\,, \quad \Sigma_{\mathrm{GC}}(n) = \sum_{j=1}^{n} S_{\mathrm{GC}}(j)\,, \tag{20}$$

and

$$\Sigma(n) = \sum_{j=1}^{n} S(j)\,. \tag{21}$$

Our goal is to show that the skew DNA walks of the 22 human autosomes display an unexpected (with respect to previous monofractal diagnosis [27,28,29,30]) bifractal scaling behavior in the range 10 to 40 kbp as the signature of the presence of transcription-induced jumps in the LRC noisy S profiles. Sequences and gene annotation data ("refGene") were retrieved from the UCSC Genome Browser (May 2004). We used RepeatMasker to exclude repetitive elements that might have been inserted recently and would not reflect long-term evolutionary patterns.

Revealing the Bifractality of Human Skew DNA Walks with the WTMM Method

As an illustration of our wavelet-based methodology, we show in Fig. 5 the S skew profile of a fragment of human chromosome 6 (Fig. 5a), the corresponding skew DNA walk (Fig. 5b) and its space-scale wavelet decomposition using the Mexican hat analyzing wavelet $g^{(2)}$ (Fig. 1b). When computing $Z(q, a)$ (Eq. (7)) from the WT skeletons of the skew DNA walks Σ of the 22 human autosomes, we get convincing power-law behavior for $-1.5 \le q \le 3$ (data not shown). In Fig. 6a are reported the $\tau(q)$ exponents obtained using a linear regression fit of $\ln Z(q, a)$ vs. $\ln a$ over the range of scales $10\,\mathrm{kbp} \le a \le 40\,\mathrm{kbp}$. All the data points remarkably fall on two straight lines $\tau_1(q) = 0.78q - 1$ and $\tau_2(q) = q - 1$ which strongly suggests the presence of two types of singularities $h_1 = 0.78$ and $h_2 = 1$, respectively on two sets S_1 and S_2 with the same Haussdorf dimension $D = -\tau_1(0) = -\tau_2(0) = 1$, as

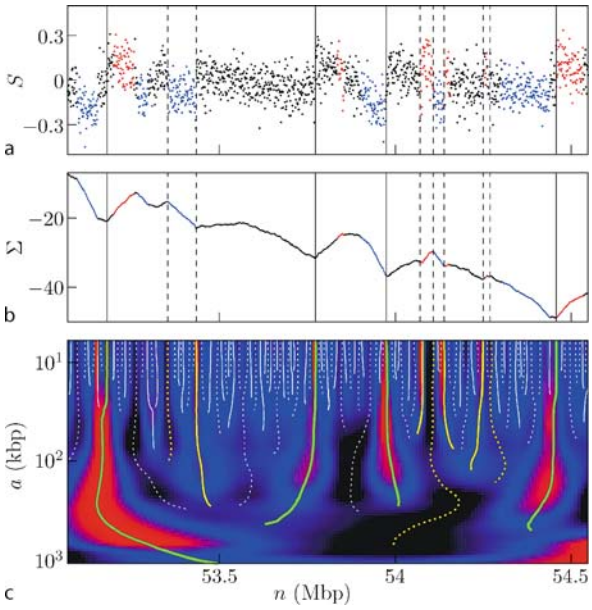

Fractals and Wavelets: What Can We Learn on Transcription and Replication …?, Figure 5

a Skew profile $S(n)$ (Eq. (19)) of a repeat-masked fragment of human chromosome 6; *red* (resp. *blue*) 1 kbp window points correspond to (+) genes (resp. (−) genes) lying on the Watson (resp. Crick) strand; *black points* to intergenic regions. b Cumulated skew profile $\Sigma(n)$ (Eq. (21)). c WT of Σ; $T_{g^{(2)}}(n, a)$ is coded from *black* (min) to *red* (max); the WT skeleton defined by the maxima lines is shown in *solid* (resp. *dashed*) lines corresponding to *positive* (resp. *negative*) WT values. For illustration *yellow solid* (resp. *dashed*) *maxima lines* are shown to point to the positions of 2 *upward* (resp. 2 *downward*) *jumps* in S (*vertical dashed lines* in a and b) that coincide with gene transcription starts (resp. ends). In *green* are shown maxima lines that persist above $a \geq 200$ kbp and that point to sharp upward jumps in S (*vertical solid lines* in a and b) that are likely to be the locations of putative replication origins (see Sect. "From the Detection of Relication Origins Using the Wavelet Transform Microscope to the Modeling of Replication in Mammalian Genomes") [98,100]; note that 3 out of those 4 jumps are co-located with transcription start sites [129]

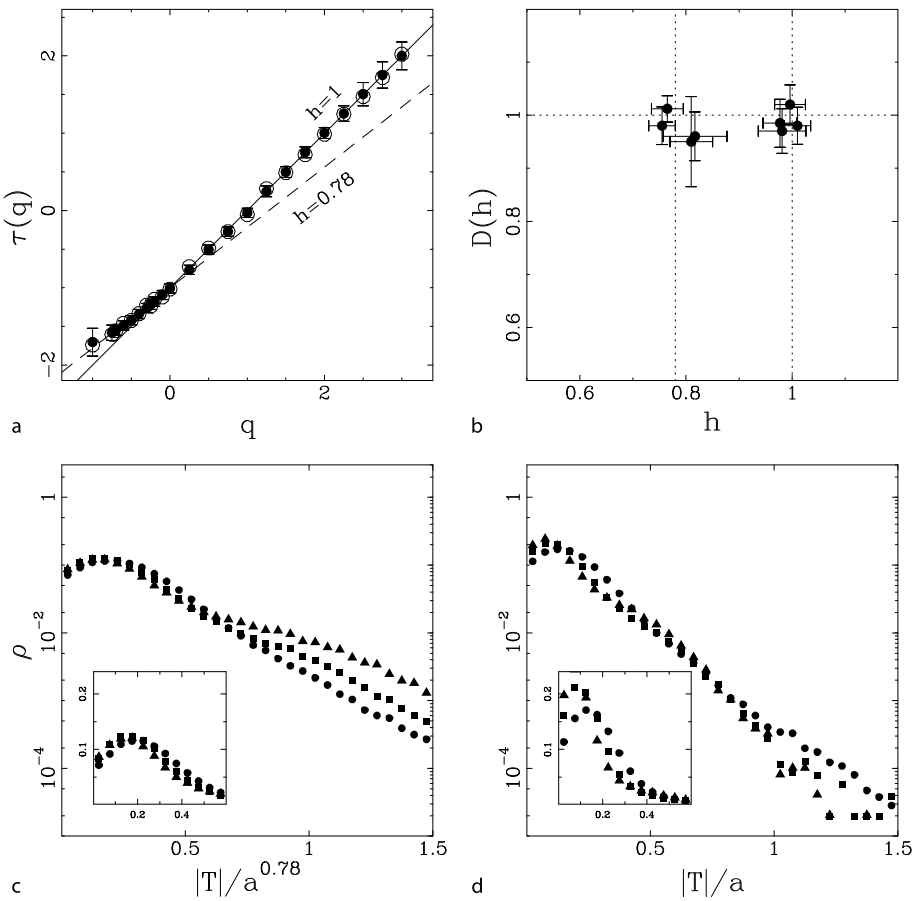

Fractals and Wavelets: What Can We Learn on Transcription and Replication …?, Figure 6
Multifractal analysis of $\Sigma(n)$ of the 22 human (*filled symbols*) and 19 mouse (*open circle*) autosomes using the WTMM method with $g^{(2)}$ over the range 10 kbp $\leq a \leq$ 40 kbp [129]. a $\tau(q)$ vs. q. b $D(h)$ vs. h. c WTMM pdf: ρ is plotted versus $|T|/a^H$ where $H = h_1 = 0.78$, in semi-log representation; the *inset* is an enlargement of the pdf central part in linear representation. d Same as in c but with $H = h_2 = 1$. In c and d, the symbols correspond to scales $a = 10$ (●), 20 (■) and 40 kbp (▲)

confirmed when computing the $D(h)$ singularity spectrum in Fig. 6b. This observation means that $Z(q, a)$ can be split in two parts [12,16]:

$$Z(q, a) = C_1(q)a^{qh_1-1} + C_2(q)a^{qh_2-1}\,, \tag{22}$$

where $C_1(q)$ and $C_2(q)$ are prefactors that depend on q. Since $h_1 < h_2$, in the limit $a \mapsto 0^+$, the partition function is expected to behave like $Z(q, a) \sim C_1(q)a^{qh_1-1}$ for $q > 0$ and like $Z(q, a) \sim C_2(q)a^{qh_2-1}$ for $q < 0$, with a so-called phase transition [12,16] at the critical value $q_c = 0$. Surprisingly, it is the contribution of the weakest singularities $h_2 = 1$ that controls the scaling behavior of $Z(q, a)$ for $q > 0$ while the strongest ones $h_1 = 0.78$ actually dominate for $q < 0$ (Fig. 6a). This inverted behavior originates from finite (1 kbp) resolution which prevents the observation of the predicted scaling behavior in the limit $a \mapsto 0^+$. The prefactors $C_1(q)$ and $C_2(q)$ in Eq. (22) are sensitive to (i) the number of maxima lines in the WT skeleton along which the WTMM behave as a^{h_1} or a^{h_2} and (ii) the relative amplitude of these WTMM. Over the range of scales used to estimate $\tau(q)$, the WTMM along the maxima lines pointing (at small scale) to $h_2 = 1$ singularities are significantly larger than those along the maxima lines associated to $h_1 = 0.78$ (see Figs. 6c, 6d). This implies that the larger $q > 0$, the stronger the inequality $C_2(q) \gg C_1(q)$ and the more pronounced the relative contribution of the second term in the r.h.s. of Eq. (22). On the opposite for $q < 0$, $C_1(q) \gg C_2(q)$ which explains that the strongest singularities $h_1 = 0.78$ now control the scaling behavior of $Z(q, a)$ over the explored range of scales.

In Figs. 6c, 6d are shown the WTMM pdfs computed at scales $a = 10$, 20 and 40 kbp after rescaling by a^{h_1} and

a^{h_2} respectively. We note that there does not exist a value of H such that all the pdfs collapse on a single curve as expected from Eq. (15) for monofractal DNA walks. Consistently with the $\tau(q)$ data in Fig. 6a and with the inverted scaling behavior discussed above, when using the two exponents $h_1 = 0.78$ and $h_2 = 1$, one succeeds in superimposing respectively the central (bump) part (Fig. 6c) and the tail (Fig. 6d) of the rescaled WTMM pdfs. This corroborates the bifractal nature of the skew DNA walks that display two competing scale-invariant components of Hölder exponents: (i) $h_1 = 0.78$ corresponds to LRC homogeneous fluctuations previously observed over the range 200 bp $\lesssim a \lesssim$ 20 kbp in DNA walks generated with structural codings [29,30] and (ii) $h_2 = 1$ is associated to convex $\vee$ and concave $\wedge$ shapes in the DNA walks Σ indicating the presence of discontinuities in the derivative of Σ, i. e., of jumps in S (Figs. 5a, 5b). At a given scale a, according to Eq. (11), a large value of the WTMM in Fig. 5c corresponds to a strong derivative of the smoothed S profile and the maxima line to which it belongs is likely to point to a jump location in S. This is particularly the case for the colored maxima lines in Fig. 5c: Upward (resp. downward) jumps (Fig. 5a) are so-identified by the maxima lines corresponding to positive (resp. negative) values of the WT.

Transcription-Induced Step-like Skew Profiles in the Human Genome

In order to identify the origin of the jumps observed in the skew profiles, we have performed a systematic investigation of the skews observed along 14 854 intron containing genes [96,97]. In Fig. 7 are reported the mean values of S_{TA} and S_{GC} skews for all genes as a function of the distance to the 5′- or 3′- end. At the 5′ gene extremities (Fig. 7a), a sharp transition of both skews is observed from about zero values in the intergenic regions to finite positive values in transcribed regions ranging between 4 and 6% for $\bar{S}_{TA}$ and between 3 and 5% for $\bar{S}_{GC}$. At the gene 3′- extremities (Fig. 7b), the TA and GC skews also exhibit transitions from significantly large values in transcribed regions to very small values in untranscribed regions. However, in comparison to the steep transitions observed at 5′- ends, the 3′- end profiles present a slightly smoother transition pattern extending over ∼5 kbp and including regions downstream of the 3′- end likely reflecting the fact that transcription continues to some extent downstream of the polyadenylation site. In pluricellular organisms, mutations responsible for the observed biases are expected to have mostly occurred in germ-line cells. It could happen that gene 3′- ends annotated in the databank differ from the poly-A sites effectively used in the germ-line cells. Such differences would then lead to some broadening of the skew profiles.

From Skew Multifractal Analysis to Gene Detection

In Fig. 8 are reported the results of a statistical analysis of the jump amplitudes in human S profiles [129]. For maxima lines that extend above $a^* = 10$ kbp in the WT skeleton (see Fig. 5c), the histograms obtained for upward and downward variations are quite similar, especially

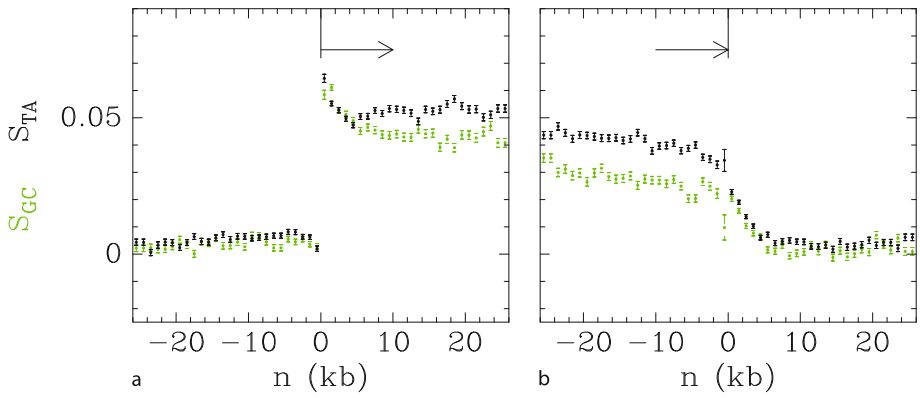

Fractals and Wavelets: What Can We Learn on Transcription and Replication …?, Figure 7
TA (●) and GC (*green* ●) skew profiles in the regions surrounding 5′ and 3′ gene extremities [96]. S_{TA} and S_{GC} were calculated in 1 kbp windows starting from each gene extremities in both directions. In abscissa is reported the distance (n) of each 1 kbp window to the indicated gene extremity; zero values of abscissa correspond to 5′- (a) or 3′- (b) gene extremities. In ordinate is reported the mean value of the skews over our set of 14 854 intron-containing genes for all 1 kbp windows at the corresponding abscissa. *Error bars* represent the standard error of the means

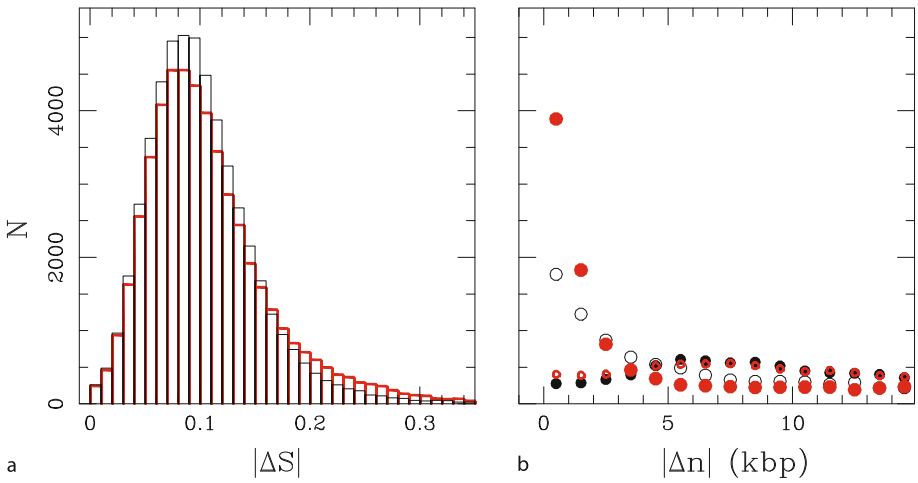

Fractals and Wavelets: What Can We Learn on Transcription and Replication . . . ?, Figure 8
Statistical analysis of skew variations at the singularity positions determined at scale 1 kbp from the maxima lines that exist at scales $a \geq 10$ kbp in the WT skeletons of the 22 human autosomes [129]. For each singularity, we computed the variation amplitudes $\Delta S = \bar{S}(3') - \bar{S}(5')$ over two adjacent 5 kbp windows, respectively in the 3′ and 5′ directions and the distances Δn to the closest TSS (resp. TES). a Histograms $N(|\Delta S|)$ for upward ($\Delta S > 0$, *red*) and downward ($\Delta S < 0$, *black*) skew variations. b Histograms of the distances Δn of upward (*red*) or downward (*black*) jumps with $|\Delta S| \geq 0.1$ to the closest TSS (●, *red* ●) and TES (○, *red* ○)

their tails that are likely to correspond to jumps in the S profiles (Fig. 8a). When computing the distance between upward or downward jumps ($|\Delta S| \geq 0.1$) to the closest transcription start (TSS) or end (TES) sites (Fig. 8b), we reveal that the number of upward jumps in close proximity ($|\Delta n| \lesssim 3$ kpb) to TSS over-exceeds the number of such jumps close to TES. Similarly, downward jumps are preferentially located at TES. These observations are consistent with the step-like shape of skew profiles induced by transcription: $S > 0$ (resp. $S < 0$) is constant along a (+) (resp. (−)) genes and $S = 0$ in the intergenic regions (Fig. 7) [96]. Since a step-like pattern is edged by one upward and one downward jump, the set of human genes that are significantly biased is expected to contribute to an even number of $\Delta S > 0$ and $\Delta S < 0$ jumps when exploring the range of scales $10 \lesssim a \lesssim 40$ kbp, typical of human gene size. Note that in Fig. 8a, the number of sharp upward jumps actually slightly exceeds the number of sharp downward jumps, consistently with the experimental observation that whereas TSS are well defined, TES may extend over 5 kbp resulting in smoother downward skew transitions (Fig. 7b). This TES particularity also explains the excess of upward jumps found close to TSS as compared to the number of downward jumps close to TES (Fig. 8b).

In Fig. 9a, we report the analysis of the distance of TSS to the closest upward jump [129]. For a given upward jump amplitude, the number of TSS with a jump within $|\Delta n|$ increases faster than expected (as compared to the number found for randomized jump positions) up to $|\Delta n| \simeq 2$ kbp. This indicates that the probability to find an upward jump within a gene promoter region is significantly larger than elsewhere. For example, out of 20 023 TSS, 36% (7228) are delineated within 2 kbp by a jump with $\Delta S > 0.1$. This provides a very reasonable estimate for the number of genes expressed in germline cells as compared to the 31.9% recently experimentally found to be bound to Pol II in human embryonic stem cells [130].

Combining the previous results presented in Figs. 8b and 9a, we report in Fig. 9b an estimate of the efficiency/coverage relationship by plotting the proportion of upward jumps ($\Delta S > \Delta S^*$) lying in TSS proximity as a function of the number of so-delineated TSS [129]. For a given proximity threshold $|\Delta n|$, increasing ΔS^* results in a decrease of the number of delineated TSS, characteristic of the right tail of the gene bias pdf. Concomitant to this decrease, we observe an increase of the efficiency up to a maximal value corresponding to some optimal value for ΔS^*. For $|\Delta n| < 2$ kbp, we reach a maximal efficiency of 60% for $\Delta S^* = 0.225$; 1403 out of 2342 upward jumps delineate a TSS. Given the fact that the actual number of human genes is estimated to be significantly larger ($\sim 30\,000$) than the number provided by refGene, a large part of the the 40% (939) of upward jumps that have not been associated to a refGene could be explained by this

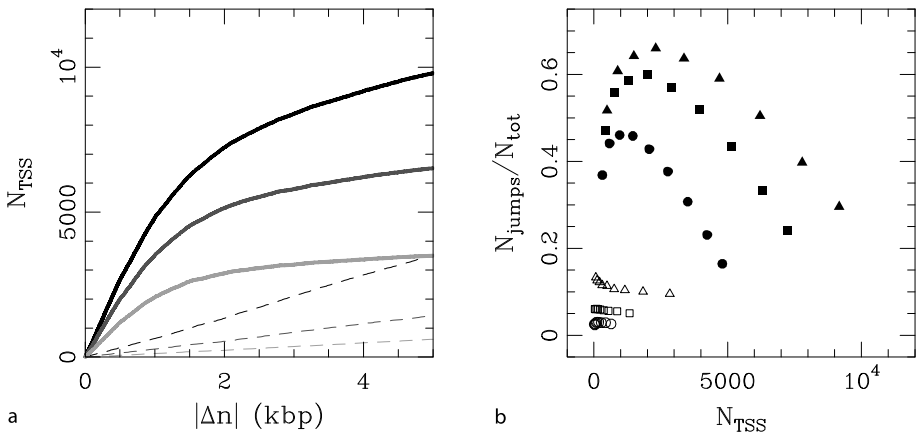

Fractals and Wavelets: What Can We Learn on Transcription and Replication …?, Figure 9
a Number of TSS with an upward jump within $|\Delta n|$ (*abscissa*) for jump amplitudes $\Delta S > 0.1$ (*black*), 0.15 (*dark gray*) and 0.2 (*light gray*). *Solid lines* correspond to true jump positions while *dashed lines* to the same analysis when jump positions were randomly drawn along each chromosome [129]. b Among the $N_{tot}(\Delta S^*)$ upward jumps of amplitude larger than some threshold ΔS^*, we plot the proportion of those that are found within 1 kbp (●), 2 kbp (■) or 4 kbp (▲) of the closest TSS vs. the number N_{TSS} of the so-delineated TSS. Curves were obtained by varying ΔS^* from 0.1 to 0.3 (from *right* to *left*). *Open symbols* correspond to similar analyzes performed on random upward jump and TSS positions

limited coverage. In other words, jumps with sufficiently high amplitude are very good candidates for the location of highly-biased gene promoters. Let us point that out of the above 1403 (resp. 2342) upward jumps, 496 (resp. 624) jumps are still observed at scale $a^* = 200$ kbp. We will see in the next section that these jumps are likely to also correspond to replication origins underlying the fact that large upward jumps actually result from the cooperative contributions of both transcription- and replication- associated biases [98,99,100,101]. The observation that 80% (496/624) of the predicted replication origins are co-located with TSS enlightens the existence of a remarkable gene organization at replication origins [101].

To summarize, we have demonstrated the bifractal character of skew DNA walks in the human genome. When using the WT microscope to explore (repeat-masked) scales ranging from 10 to 40 kbp, we have identified two competing homogeneous scale-invariant components characterized by Hölder exponents $h_1 = 0.78$ and $h_2 = 1$ that respectively correspond to LRC colored noise and sharp jumps in the original DNA compositional asymmetry profiles. Remarkably, the so-identified upward (resp. downward) jumps are mainly found at the TSS (resp. TES) of human genes with high transcription bias and thus very likely highly expressed. As illustrated in Fig. 6a, similar bifractal properties are also observed when investigating the 19 mouse autosomes. This suggests that the results reported in this section are general features of mammalian genomes [129].

From the Detection of Relication Origins Using the Wavelet Transform Microscope to the Modeling of Replication in Mammalian Genomes

DNA replication is an essential genomic function responsible for the accurate transmission of genetic information through successive cell generations. According to the so-called "replicon" paradigm derived from prokaryotes [131], this process starts with the binding of some "initiator" protein to a specific "replicator" DNA sequence called *origin of replication*. The recruitment of additional factors initiate the bi-directional progression of two divergent replication forks along the chromosome. One strand is replicated continuously (leading strand), while the other strand is replicated in discrete steps towards the origin (lagging strand). In eukaryotic cells, this event is initiated at a number of replication origins and propagates until two converging forks collide at a *terminus of replication* [132]. The initiation of different replication origins is coupled to the cell cycle but there is a definite flexibility in the usage of the replication origins at different developmental stages [133,134,135,136,137]. Also, it can be strongly influenced by the distance and timing of activation of neighboring replication origins, by the transcriptional activity and by the local chromatin structure [133,134,135,137]. Actually, sequence requirements for a replication origin vary significantly between different eukaryotic organisms. In the unicellular eukaryote *Saccharomyces cerevisiae*, the replication ori-

gins spread over 100–150 bp and present some highly conserved motifs [132]. However, among eukaryotes, *S. cerevisiae* seems to be the exception that remains faithful to the replicon model. In the fission yeast *Schizosaccharomyces pombe*, there is no clear consensus sequence and the replication origins spread over at least 800 to 1000 bp [132]. In multicellular organisms, the nature of initiation sites of DNA replication is even more complex. Metazoan replication origins are rather poorly defined and initiation may occur at multiple sites distributed over a thousand of base pairs [138]. The initiation of replication at random and closely spaced sites was repeatedly observed in *Drosophila* and *Xenopus* early embryo cells, presumably to allow for extremely rapid S phase, suggesting that any DNA sequence can function as a replicator [136,139,140]. A developmental change occurs around midblastula transition that coincides with some remodeling of the chromatin structure, transcription ability and selection of preferential initiation sites [136,140]. Thus, although it is clear that some sites consistently act as replication origins in most eukaryotic cells, the mechanisms that select these sites and the sequences that determine their location remain elusive in many cell types [141,142]. As recently proposed by many authors [143,144,145], the need to fulfill specific requirements that result from cell diversification may have led multicellular eukaryotes to develop various epigenetic controls over the replication origin selection rather than to conserve specific replication sequence. This might explain that only very few replication origins have been identified so far in multicellular eukaryotes, namely around 20 in metazoa and only about 10 in human [146]. Along the line of this epigenetic interpretation, one might wonder what can be learned about eukaryotic DNA replication from DNA sequence analysis.

Replication Induced Factory-Roof Skew Profiles in Mammalian Genomes

The existence of replication associated strand asymmetries has been mainly established in bacterial genomes [87,90,92,93,94]. S_{GC} and S_{TA} skews abruptly switch sign (over few kbp) from negative to positive values at the replication origin and in the opposite direction from positive to negative values at the replication terminus. This step-like profile is characteristic of the replicon model [131] (see Fig. 13, left panel). In eukaryotes, the existence of compositional biases is unclear and most attempts to detect the replication origins from strand compositional asymmetry have been inconclusive. Several studies have failed to show compositional biases related to replication, and analysis of nucleotide substitutions in the region of the *β-globin* replication origin in primates does not support the existence of mutational bias between the leading and the lagging strands [92,147,148]. Other studies have led to rather opposite results. For instance, strand asymmetries associated with replication have been observed in the subtelomeric regions of *Saccharomyces cerevisiae* chromosomes, supporting the existence of replication-coupled asymmetric mutational pressure in this organism [149].

As shown in Fig. 10a for the TOP1 replication origin [146], most of the known replication origins in the human genome correspond to rather sharp (over several kbp) transitions from negative to positive S (S_{TA} as well as S_{GC}) skew values that clearly emerge from the noisy background. But when examining the behavior of the skews at larger distances from the origin, one does not observe a step-like pattern with upward and downward jumps at the origin and termination positions, respectively, as expected for the bacterial replicon model (Fig. 13, left panel). Surprisingly, on both sides of the upward jump, the noisy S profile decreases steadily in the 5′ to 3′ direction without clear evidence of pronounced downward jumps. As shown in Figs. 10b–10d, sharp upward jumps of amplitude $\Delta S \gtrsim 15\%$, similar to the ones observed for the known replication origins (Fig. 10a), seem to exist also at many other locations along the human chromosomes. But the most striking feature is the fact that in between two neighboring major upward jumps, not only the noisy S profile does not present any comparable downward sharp transition, but it displays a remarkable decreasing linear behavior. At chromosome scale, we thus get jagged S profiles that have the aspect of "factory roofs" [98,100,146]. Note that the jagged S profiles shown in Figs. 10a–10d look somehow disordered because of the extreme variability in the distance between two successive upward jumps, from spacing $\sim$ 50–100 kbp ($\sim$ 100–200 kbp for the native sequences) mainly in GC rich regions (Fig. 10d), up to 1–2 Mbp ($\sim$ 2–3 Mbp for native sequences) (Fig. 10c) in agreement with recent experimental studies [150] that have shown that mammalian replicons are heterogeneous in size with an average size $\sim$ 500 kbp, the largest ones being as large as a few Mbp. But what is important to notice is that some of these segments between two successive skew upward jumps are entirely intergenic (Figs. 10a, 10c), clearly illustrating the particular profile of a strand bias resulting solely from replication [98,100,146]. In most other cases, we observe the superimposition of this replication profile and of the step-like profiles of (+) and (−) genes (Fig. 7), appearing as upward and downward blocks standing out from the replication pattern (Fig. 10c). Importantly, as illustrated in Figs. 10e, 10f, the factory-roof pattern is not specific to human se-

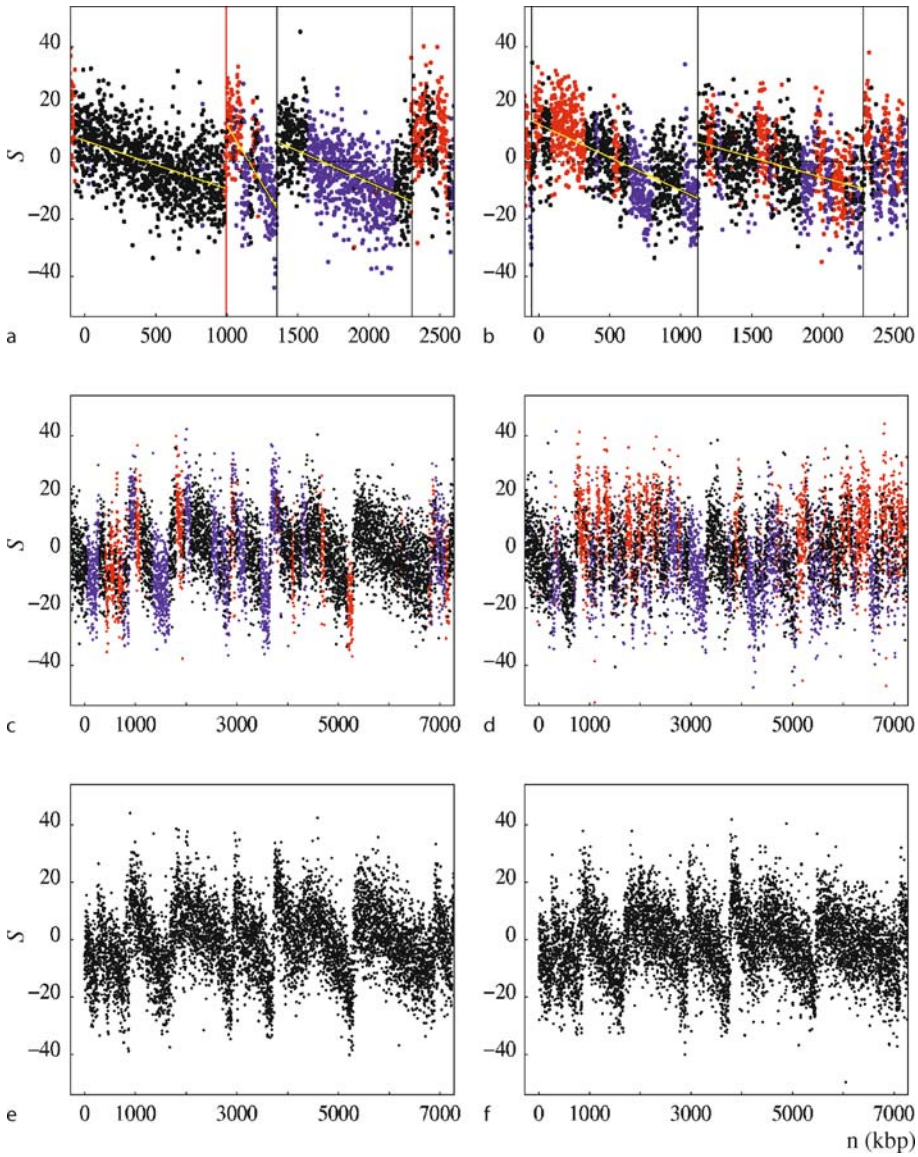

Fractals and Wavelets: What Can We Learn on Transcription and Replication …?, Figure 10
***S* profiles along mammalian genome fragments [100,146]. a Fragment of human chromosome 20 including the TOP1 origin (*red vertical line*). b and c Human chromosome 4 and chromosome 9 fragments, respectively, with low GC content (36%). d Human chromosome 22 fragment with larger GC content (48%). In a and b, *vertical lines* correspond to selected putative origins (see Subsect. "Detecting Replication Origins from the Skew WT Skeleton"); *yellow lines* are linear fits of the *S* values between successive putative origins. *Black intergenic regions*; *red*, (+) genes; *blue*, (—) genes. Note the fully intergenic regions upstream of TOP1 in a and from positions 5290–6850 kbp in c. e Fragment of mouse chromosome 4 homologous to the human fragment shown in c. f Fragment of dog chromosome 5 syntenic to the human fragment shown in c. In e and f, genes are not represented**

quences but is also observed in numerous regions of the mouse and dog genomes [100]. Hence, the presence of strand asymmetry in regions that have strongly diverged during evolution further supports the existence of compostional bias associated with replication in mammalian germ-line cells [98,100,146].

Detecting Replication Origins from the Skew WT Skeleton

We have shown in Fig. 10a that experimentally determined human replication origins coincide with large-amplitude upward transitions in noisy skew profiles. The

corresponding ΔS ranges between 14% and 38%, owing to possible different replication initiation efficiencies and/or different contributions of transcriptional biases (Sect. "Bifractality of Human DNA Strand-Asymmetry Profiles Results from Transcription"). Along the line of the jump detection methodology described in Sect. "Bifractality of Human DNA Strand-Asymmetry Profiles Results from Transcription", we have checked that upward jumps observed in the skew S at these known replication origins correspond to maxima lines in the WT skeleton that extend to rather large scales $a > a^* = 200$ kbp. This observation has led us to select the maxima lines that exist above $a^* = 200$ kbp, i. e. a scale which is smaller than the typical replicon size and larger than the typical gene size [98, 100]. In this way, we not only reduce the effect of the noise but we also reduce the contribution of the upward (5′ extremity) and backward (3′ extremity) jumps associated to the step-like skew pattern induced by transcription only (Sect. "Bifractality of Human DNA Strand-Asymmetry Profiles Results from Transcription"), to the benefit of maintaining a good sensitivity to replication induced jumps. The detected jump locations are estimated as the positions at scale 20 kbp of the so-selected maxima lines. According to Eq. (11), upward (resp. downward) jumps are identified by the maxima lines corresponding to positive (resp. negative) values of the WT as illustrated in Fig. 5c by the green solid (resp. dashed) maxima lines. When applying this methodology to the total skew S along the repeat-masked DNA sequences of the 22 human autosomal chromosomes, 2415 upward jumps are detected and, as expected, a similar number (namely 2686) of downward jumps. In Fig. 11a are reported the histograms of the amplitude $|\Delta S|$ of the so-identified upward ($\Delta S > 0$) and downward ($\Delta S < 0$) jumps respectively. These histograms no longer superimpose as previously observed at smaller scales in Fig. 8a, the former being significantly shifted to larger $|\Delta S|$ values. When plotting $N(|\Delta S| > \Delta S^*)$ versus ΔS^* in Fig. 11b, we can see that the number of large amplitude upward jumps overexceeds the number of large amplitude downward jumps. These results confirm that most of the sharp upward transitions in the S profiles in Fig. 10 have no sharp downward transition counterpart [98,100]. This excess likely results from the fact that, contrasting with the prokaryote replicon model (Fig. 13, left panel) where downward jumps result from precisely positioned replication terminations, in mammals termination appears not to occur at specific positions but to be randomly distributed. Accordingly the small number of downward jumps with large $|\Delta S|$ is likely to result from transcription (Fig. 5) and not from replication. These jumps are probably due to highly biased genes that also generate a small number of large-amplitude upward jumps, giving rise to false-positive candidate replication origins. In that respect, the number of large downward jumps can be taken as an estimation of the number of false positives. In a first step, we have retained as acceptable a proportion of 33% of false positives. As shown in Fig. 11b, this value results from the selection of upward and downward jumps of amplitude $|\Delta S| \geq 12.5\%$, corresponding to a ratio of upward over downward jumps

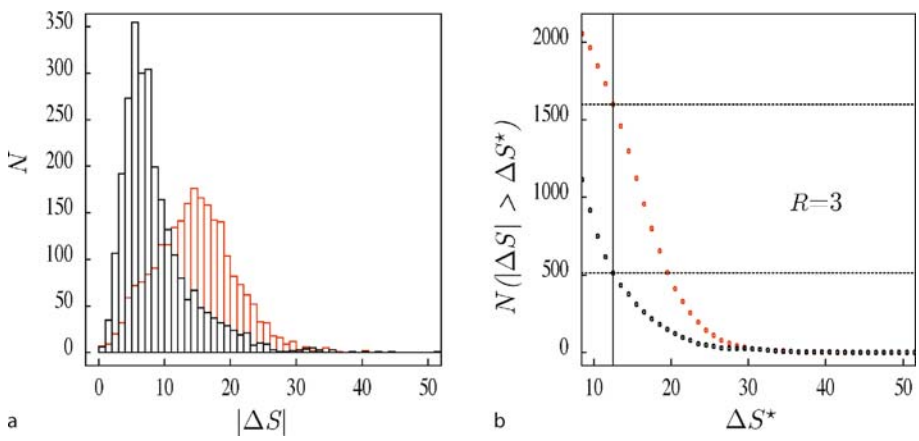

Fractals and Wavelets: What Can We Learn on Transcription and Replication …?, Figure 11
Statistical analysis of the sharp jumps detected in the S profiles of the 22 human autosomal chromosomes by the WT microscope at scale $a^* = 200$ kbp for repeat-masked sequences [98,100]. $|\Delta S| = |\bar{S}(3') - \bar{S}(5')|$, where the averages were computed over the two adjacent 20 kbp windows, respectively, in the 3′ and 5′ direction from the detected jump location. a Histograms $N(|\Delta S|)$ of $|\Delta S|$ values. b $N(|\Delta S| > \Delta S^*)$ vs. ΔS^*. In a and b, the *black* (resp. *red*) *line* corresponds to downward $\Delta S < 0$ (resp. upward $\Delta S > 0$) jumps. $R = 3$ corresponds to the ratio of upward over downward jumps presenting an amplitude $|\Delta S| \geq 12.5\%$ (see text)

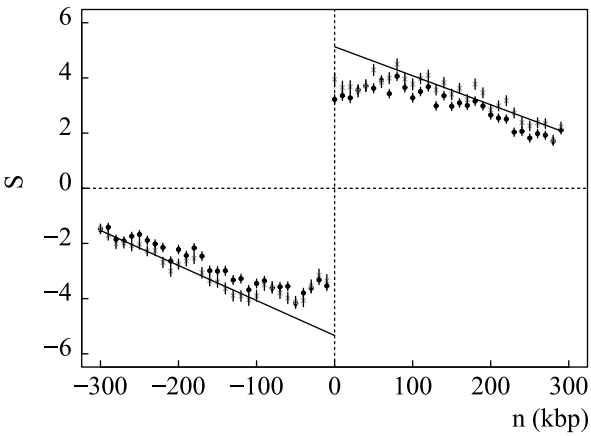

Fractals and Wavelets: What Can We Learn on Transcription and Replication …?, Figure 12

Mean skew profile of intergenic regions around putative replication origins [100]. The skew *S* was calculated in 1 kbp windows (Watson strand) around the position (±300 kbp without repeats) of the 1012 detected upward jumps; 5′ and 3′ transcript extremities were extended by 0.5 and 2 kbp, respectively (•), or by 10 kbp at both ends (∗). The abscissa represents the distance (in kbp) to the corresponding origin; the ordinate represents the skews calculated for the windows situated in intergenic regions (mean values for all discontinuities and for 10 consecutive 1 kbp window positions). The skews are given in percent (*vertical bars*, SEM). The *lines* correspond to linear fits of the values of the skew (∗) for $n < -100$ kbp and $n > 100$ kbp

$R = 3$. Let us notice that the value of this ratio is highly variable along the chromosome [146] and significantly larger than 1 for G+C $\lesssim$ 42%.

In a final step, we have decided [98,100,146] to retain as putative replication origins upward jumps with $|\Delta S| \geq 12.5\%$ detected in regions with G+C $\leq$ 42%. This selection leads to a set of 1012 candidates among which our estimate of the proportion of true replication origins is 79% ($R = 4.76$). In Fig. 12 is shown the mean skew profile calculated in intergenic windows on both sides of the 1012 putative replication origins [100]. This mean skew profile presents a rather sharp transition from negative to positive values when crossing the origin position. To avoid any bias in the skew values that could result from incompletely annotated gene extremities (e. g. 5′ and 3′ UTRs), we have removed 10-kbp sequences at both ends of all annotated transcripts. As shown in Fig. 12, the removal of these intergenic sequences does not significantly modifies the mean skew profile, indicating that the observed values do not result from transcription. On both sides of the jump, we observe a linear decrease of the bias with some flattening of the profile close to the transition point. Note that, due to (i) the potential presence of signals implicated in replication initiation and (ii) the possible existence of dispersed origins [151], one might question the meaningfulness of this flattening that leads to a significant underestimate of the jump amplitude. Furthermore, according to our detection methodology, the numerical uncertainty on

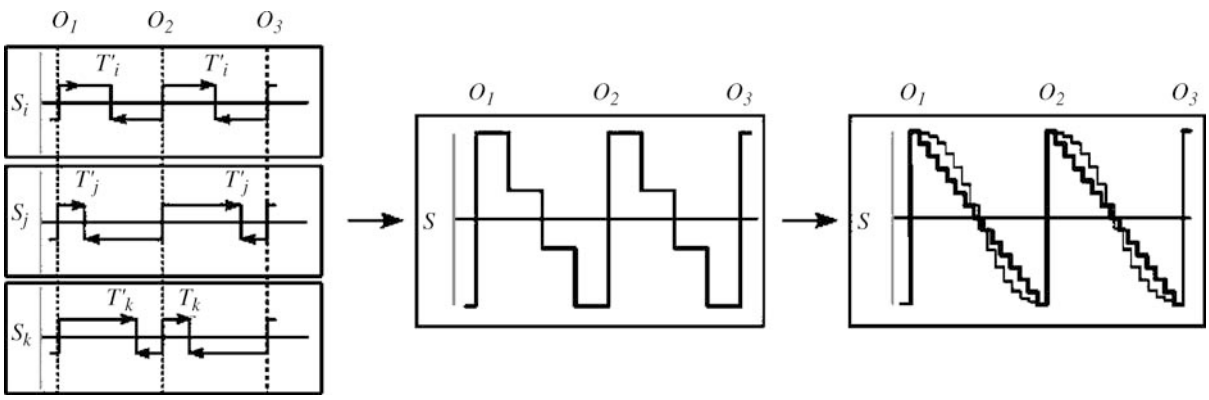

Fractals and Wavelets: What Can We Learn on Transcription and Replication …?, Figure 13

Model of replication termination [98,100]. Schematic representation of the skew profiles associated with three replication origins O_1, O_2, and O_3; we suppose that these replication origins are adjacent, bidirectional origins with similar replication efficiency. The abscissa represents the sequence position; the ordinate represents the *S* value (arbitrary units). *Upward* (or *downward*) steps correspond to origin (or termination) positions. For convenience, the termination sites are symmetric relative to O_2. (*Left*) Three different termination positions T_i, T_j, and T_k, leading to elementary skew profiles S_i, S_j, and S_k as predicted by the replicon model [146]. (*Center*) Superposition of these three profiles. (*Right*) Superposition of a large number of elementary profiles leading to the final factory-roof pattern. In the simple model, termination occurs with equal probability on both sides of the origins, leading to the linear profile (*thick line*). In the alternative model, replication termination is more likely to occur at lower rates close to the origins, leading to a flattening of the profile (*gray line*)

the putative origin position estimate may also contribute to this flattening. As illustrated in Fig. 12, when extrapolating the linear behavior observed at distances > 100 kbp from the jump, one gets a skew of 5.3%, i. e. a value consistent with the skew measured in intergenic regions around the six experimentally known replication origins namely 7.0 ± 0.5%. Overall, the detection of sharp upward jumps in the skew profiles with characteristics similar to those of experimentally determined replication origins and with no downward counterpart further supports the existence, in human chromosomes, of replication-associated strand asymmetries, leading to the identification of numerous putative replication origins active in germ-line cells.

A Model of Replication in Mammalian Genomes

Following the observation of jagged skew profiles similar to factory roofs in Subsect. "Replication Induced Factory-Roof Skew Profiles in Mammalian Genomes", and the quantitative confirmation of the existence of such (piecewise linear) profiles in the neighborhood of 1012 putative origins in Fig. 12, we have proposed, in Touchon et al. [100] and Brodie of Brodie et al. [98], a rather crude model for replication in the human genome that relies on the hypothesis that the replication origins are quite well positioned while the terminations are randomly distributed. Although some replication terminations have been found at specific sites in *S. cerevisiae* and to some extent in *Schizosaccharomyces pombe* [152], they occur randomly between active origins in *Xenopus* egg extracts [153, 154]. Our results indicate that this property can be extended to replication in human germ-line cells. As illustrated in Fig. 13, replication termination is likely to rely on the existence of numerous potential termination sites distributed along the sequence. For each termination site (used in a small proportion of cell cycles), strand asymmetries associated with replication will generate a step-like skew profile with a downward jump at the position of termination and upward jumps at the positions of the adjacent origins (as in bacteria). Various termination positions will thus correspond to classical replicon-like skew profiles (Fig. 13, left panel). Addition of these profiles will generate the intermediate profile (Fig. 13, central panel). In a simple picture, we can reasonably suppose that termination occurs with constant probability at any position on the sequence. This behavior can, for example, result from the binding of some termination factor at any position between successive origins, leading to a homogeneous distribution of termination sites during successive cell cycles. The final skew profile is then a linear segment decreasing between successive origins (Fig. 13, right panel). Let us point out that firing of replication origins during time interval of the *S* phase [155] might result in some flattening of the skew profile at the origins as sketched in Fig. 13 (right panel, gray curve). In the present state, our results [98,100,146] support the hypothesis of random replication termination in human, and more generally in mammalian cells (Fig. 10), but further analyzes will be necessary to determine what scenario is precisely at work.

A Wavelet-Based Methodology to Disentangle Transcription- and Replication-Associated Strand Asymmetries Reveals a Remarkable Gene Organization in the Human Genome

During the duplication of eukaryotic genomes that occurs during the *S* phase of the cell cycle, the different replication origins are not all activated simultaneously [132,135, 138,150,155,156]. Recent technical developments in genomic clone microarrays have led to a novel way of detecting the temporal order of DNA replication [155,156]. The arrays are used to estimate *replication timing ratios* i. e. ratios between the average amount of DNA in the *S* phase at a locus along the genome and the usual amount of DNA present in the *G*1 phase for that locus. These ratios should vary between 2 (throughout the *S* phase, the amount of DNA for the earliest replicating regions is twice the amount during *G*1 phase) and 1 (the latest replicating regions are not duplicated until the end of *S* phase). This approach has been successfully used to generate genome-wide maps of replication timing for *S. cerevisiae* [157], *Drosophila melanogaster* [137] and human [158]. Very recently, two new analyzes of human chromosome 6 [156] and 22 [155] have improved replication timing resolution from 1 Mbp down to ∼ 100 kbp using arrays of overlapping tile path clones. In this section, we report on a very promising first step towards the experimental confirmation of the thousand putative replication origins described in Sect. "From the Detection of Relication Origins Using the Wavelet Transform Microscope to the Modeling of Replication in Mammalian Genomes". The strategy will consist in mapping them on the recent high-resolution timing data [156] and in checking that these regions replicate earlier than their surrounding [114]. But to provide a convincing experimental test, we need as a prerequisite to extract the contribution of the compositional skew specific to replication.

Disentangling Transcription- and Replication-Associated Strand Asymmetries

The first step to detect putative replication domains consists in developing a multi-scale pattern recognition

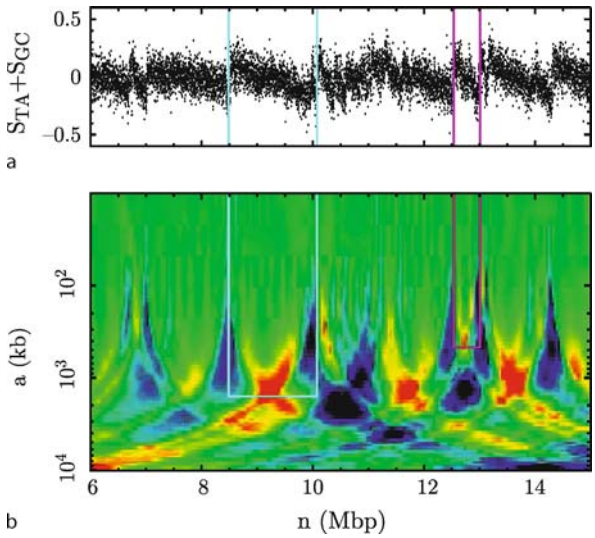

Fractals and Wavelets: What Can We Learn on Transcription and Replication …?, Figure 14

Wavelet-based analysis of genomic sequences. a Skew profile *S* of a 9 Mbp repeat-masked fragment of human chromosome 21. b WT of *S* using φ_R (Fig. 1c); $T_{\phi_R}[S](n,a)$ is color-coded from *dark-blue* (min; negative values) to *red* (max; positive values) through *green* (null values). *Light-blue* and *purple lines* illustrate the detection of two replication domains of significantly different sizes. Note that in b, *blue cone-shape areas* signing upward jumps point at small scale (*top*) towards the putative replication origins and that the vertical positions of the WT maxima (*red areas*) corresponding to the two indicated replication domains match the distance between the putative replication origins (1.6 Mbp and 470 kbp respectively)

methodology based on the WT of the strand compositional asymmetry S using as analyzing wavelet $\phi_R(x)$ (Eq. (12)) that is adapted to perform an objective segmentation of factory-roof skew profiles (Fig. 1c). As illustrated in Fig. 14, the space-scale location of significant maxima values in the 2D WT decomposition (red areas in Fig. 14b) indicates the middle position (spatial location) of candidate replication domains whose size is given by the scale location. In order to avoid false positives, we then check that there does exist a well-defined upward jump at each domain extremity. These jumps appear in Fig. 14b as blue cone-shape areas pointing at small scale to the jumps positions where are located the putative replication origins. Note that because the analyzing wavelet is of zero mean (Eq. (2)), the WT decomposition is insensitive to (global) asymmetry offset.

But as discussed in Sect. "Bifractality of Human DNA Strand-Asymmetry Profiles Results from Transcription", the overall observed skew S also contains some contribution induced by transcription that generates step-like blocks corresponding to (+) and (−) genes [96,97,129]. Hence, when superimposing the replication serrated and transcription step-like skew profiles, we get the following theoretical skew profile in a replication domain [114]:

$$\begin{aligned} S(x') &= S_R(x') + S_T(x') \\ &= -2\delta \times \left(x' - \frac{1}{2}\right) + \sum_{\text{gene}} c_g \chi_g(x') \,, \end{aligned} \tag{23}$$

where position x' within the domain has been rescaled between 0 and 1, $\delta > 0$ is the replication bias, χ_g is the characteristic function for the gth gene (1 when x' points within the gene and 0 elsewhere) and c_g is its transcriptional bias calculated on the Watson strand (likely to be positive for (+) genes and negative for (−) genes). The objective is thus to detect human replication domains by delineating, in the noisy S profile obtained at 1 kbp resolution (Fig. 15a), all chromosomal loci where S is well fitted by the theoretical skew profile Eq. (23).

In order to enforce strong compatibility with the mammalian replicon model (Subsect. "A Model of Replication in Mammalian Genomes"), we will only retain the domains the most likely to be bordered by putative replication origins, namely those that are delimited by upward jumps corresponding to a transition from a negative S value $< -3\%$ to a positive S value $> +3\%$. Also, for each domain so-identified, we will use a least-square fitting procedure to estimate the replication bias δ, and each of the gene transcription bias c_g. The resulting χ^2 value will then be used to select the candidate domains where the noisy S profile is well described by Eq. (23). As illustrated in Fig. 15 for a fragment of human chromosome 6 that contains 4 adjacent replication domains (Fig. 15a), this method provides a very efficient way of disentangling the step-like transcription skew component (Fig. 15b) from the serrated component induced by replication (Fig. 15c). Applying this procedure to the 22 human autosomes, we delineated 678 replication domains of mean length $\langle L \rangle = 1.2 \pm 0.6$ Mbp, spanning 28.3% of the genome and predicted 1060 replication origins.

DNA Replication Timing Data Corroborate *in silico* Human Replication Origin Predictions

Chromosome 22 being rather atypical in gene and GC contents, we mainly report here on the correlation analysis [114] between nucleotide compositional skew and timing data for chromosome 6 which is more representative of the whole human genome. Note that timing data for clones completely included in another clone have been removed after checking for timing ratio value consistency leaving 1648 data points. The timing ratio value at each

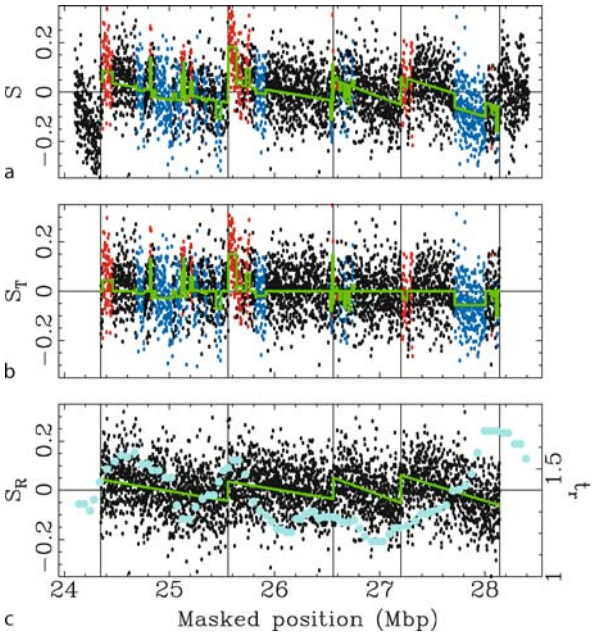

Fractals and Wavelets: What Can We Learn on Transcription and Replication …?, Figure 15

a Skew profile *S* of a 4.3 Mbp repeat-masked fragment of human chromosome 6 [114]; each point corresponds to a 1 kbp window: *Red*, (+) genes; *blue*, (−) genes; *black*, intergenic regions (the color was defined by majority rule); the estimated skew profile (Eq. (23)) is shown in *green*; *vertical lines* correspond to the locations of 5 putative replication origins that delimit 4 adjacent domains identified by the wavelet-based methodology. b Transcription-associated skew S_T obtained by subtracting the estimated replication-associated profile (*green lines* in c) from the original *S* profile in a; the estimated transcription step-like profile (second term on the *rhs* of Eq. (23)) is shown in *green*. c Replication-associated skew S_R obtained by subtracting the estimated transcription step-like profile (*green lines* in b) from the original *S* profile in a; the estimated replication serrated profile (first term in the *rhs* of Eq. (23)) is shown in *green*; the *light-blue dots* correspond to high-resolution t_r data

point has been chosen as the median over the 4 closest data points to remove noisy fluctuations resulting from clone heterogeneity (clone length 100 ± 51 kbp and distance between successive clone mid-points 104 ± 89 kbp), so that the spatial resolution is rather inhomogeneous ~ 300 kbp. Note that using asynchronous cells also results in some smoothing of the data, possibly masking local maxima.

Our wavelet-based methodology has identified 54 replication domains in human chromosome 6 [114]; these domains are bordered by 83 putative replication origins among which 25 are common to two adjacent domains. Four of these contiguous domains are shown in Fig. 15. In Fig. 15c, on top of the replication skew profile S_R, are reported for comparison the high-resolution timing ratio t_r data from [156]. The histogram of t_r values obtained at the 83 putative origin locations displays a maximum at $t_r \simeq \langle t_r \rangle \simeq 1.5$ (data not shown) and confirms what is observed in Fig. 15c, namely that a majority of the predicted origins are rather early replicating with $t_r \gtrsim 1.4$. This contrasts with the rather low t_r ($\simeq$1.2) values observed in domain central regions (Fig. 15c). But there is an even more striking feature in the replication timing profile in Fig. 15c: 4 among the 5 predicted origins correspond, relatively to the experimental resolution, to local maxima of the t_r profile. As shown in Fig. 16a, the average t_r profile around the 83 putative replication origins decreases regularly on both sides of the origins over a few (4–6) hundreds kbp confirming statistically that domain borders replicate earlier than their left and right surroundings which is consistent with these regions being true replication origins mostly active early in *S* phase. In fact, when averaging over the top 20 origins with a well-defined local maximum in the t_r profile, $\langle t_r \rangle$ displays a faster decrease on both sides of the origin and a higher maximum value ~ 1.55 corresponding to the earliest replicating origins. On the opposite, when averaging t_r profiles over the top 10 late replicating origins, we get, as expected, a rather flat mean profile ($t_r \sim 1.2$) (Fig. 16a). Interestingly, these origins are located in rather wide regions of very low GC content ($\lesssim 34\%$, not shown) correlating with chromosomal G banding patterns predominantly composed of GC-poor isochores [159,160]. This illustrates how the statistical contribution of rather flat profiles observed around late replicating origins may significantly affect the overall mean t_r profile. Individual inspection of the 38 replication domains with $L \geq 1$ Mbp shows that, in those domains that are bordered by early replicating origins ($t_r \gtrsim 1.4 - 1.5$), the replication timing ratio t_r and the absolute value of the replication skew $|S_R|$ turn out to be strongly correlated. This is quantified in Fig. 16b by the histogram of the Pearson's correlation coefficient values that is clearly shifted towards positive values with a maximum at ~ 0.4. Altogether the results of this comparative analysis provide the first experimental verification of *in silico* replication origins predictions: The detected putative replication domains are bordered by replication origins mostly active in the early *S* phase, whereas the central regions replicate more likely in late *S* phase.

Gene Organization in the Detected Replication Domains

Most of the 1060 putative replication origins that border the detected replication domains are intergenic (77%)

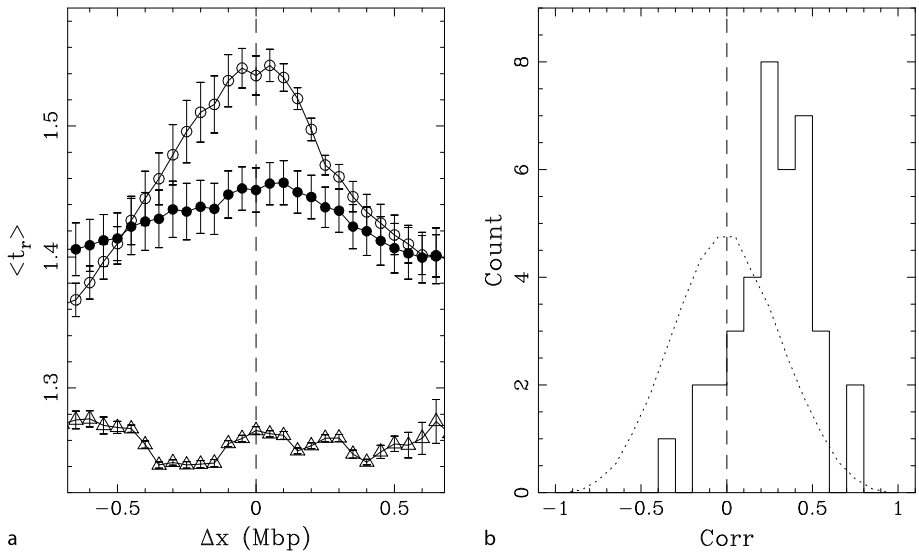

Fractals and Wavelets: What Can We Learn on Transcription and Replication …?, Figure 16

a Average replication timing ratio (±SEM) determined around the 83 putative replication origins (●), 20 origins with well-defined local maxima (○) and 10 late replicating origins (△). Δx is the native distance to the origins in Mbp units [114]. b Histogram of Pearson's correlation coefficient values between t_r and the absolute value of S_R over the 38 predicted domains of length $L \geq 1$ Mbp. The *dotted line* corresponds to the expected histogram computed with the correlation coefficients between t_r and $|S|$ profiles over independent windows randomly positioned along chromosome 6 and with the same length distribution as the 38 detected domains

and are located near to a gene promoter more often than would be expected by chance (data not shown) [101]. The replication domains contain approximately equal numbers of genes oriented in each direction (1511 (+) genes and 1507 (−) genes). Gene distributions in the 5′ halves of domains contain more (+) genes than (−) genes, regardless of the total number of genes located in the half-domains (Fig. 17b). Symmetrically, the 3′ halves contain more (−) genes than (+) genes (Fig. 17b). 32.7% of half-domains contain one gene, and 50.9% contain more than one gene. For convenience, (+) genes in the 5′ halves and (−) genes in the 3′ halves are defined as R+ genes (Fig. 17a): their transcription is, in most cases, oriented in the same direction as the putative replication fork progression (genes transcribed in the opposite direction are defined as R− genes). The 678 replication domains contain significantly more R+ genes (2041) than R− genes (977). Within 50 kbp of putative replication origins, the mean density of R+ genes is 8.2 times greater than that of R− genes. This asymmetry weakens progressively with the distance from the putative origins, up to ∼ 250 kbp (Fig. 17b). A similar asymmetric pattern is observed when the domains containing duplicated genes are eliminated from the analysis, whereas control domains obtained after randomization of domain positions present similar R+ and R− gene density distributions (Supplementary in [101]). The mean length of the R+ genes near the putative origins is significantly greater (∼ 160 kbp) than that of the R− genes (∼ 50 kbp), however both tend towards similar values (∼ 70 kbp) at the center of the domain (Fig. 17c). Within 50 kbp of the putative origins, the ratio between the numbers of base pairs transcribed in the R+ and R− directions is 23.7; this ratio falls to ∼ 1 at the domain centers (Fig. 17d). In Fig. 17e are reported the results of the analysis of the breadth of expression, N_t (number of tissues in which a gene is expressed) of genes located within the detected domains [101]. As measured by EST data (similar results are obtained by SAGE or microarray data [101]), N_t is found to decrease significantly from the extremities to the center in a symmetrical manner in the 5′ and 3′ half-domains (Fig. 17e). Thus, genes located near the putative replications origins tend to be widely expressed whereas those located far from them are mostly tissue-specific.

To summarize, the results reported in this section provide the first demonstration of quantitative relationships in the human genome between gene expression, orientation and distance from putative replication origins [101]. A possible key to the understanding of this complex architecture is the coordination between replication and transcription [101]. The putative replication origins would mostly be active early in the *S* phase in most tissues. Their activity could result from particular genomic context in-

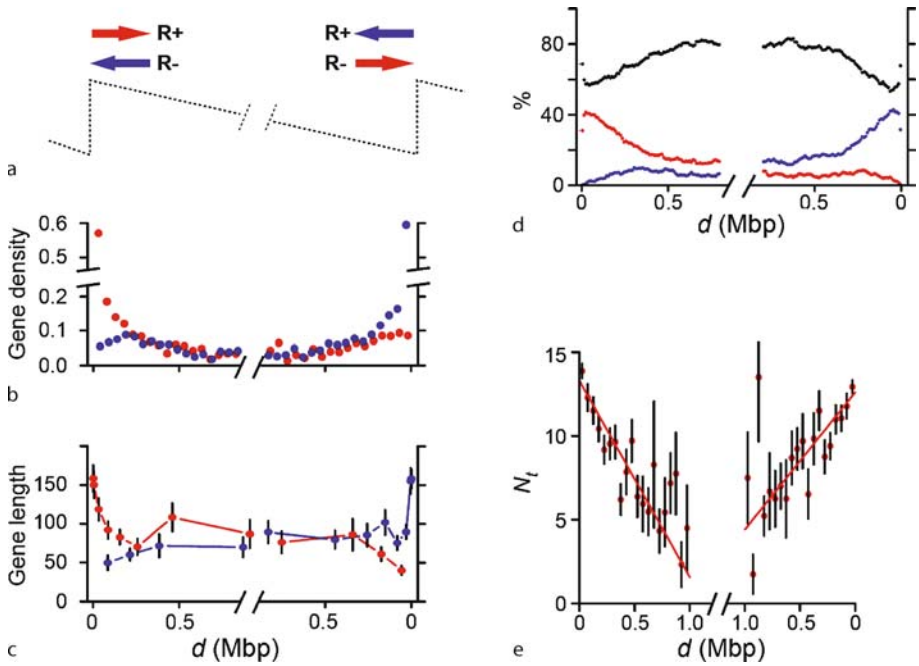

Fractals and Wavelets: What Can We Learn on Transcription and Replication …?, Figure 17

Analysis of the genes located in the identified replication domains [101]. a ***Arrows*** **indicate the R+ orientation, i. e. the same orientation as the most frequent direction of putative replication fork progression; R— orientation (opposed direction);** ***red*****, (+) genes;** ***blue*****, (—) genes. b Gene density. The density is defined as the number of 5′ ends (for (+) genes) or of 3′ ends (for (—) genes) in 50-kbp adjacent windows, divided by the number of corresponding domains. In abscissa, the distance,** ***d*****, in Mbp, to the closest domain extremity. c Mean gene length. Genes are ranked by their distance,** ***d*****, from the closest domain extremity, grouped by sets of 150 genes, and the mean length (kbp) is computed for each set. d Relative number of base pairs transcribed in the + direction (*red*), — direction (*blue*) and non-transcribed (*black*) determined in 10-kbp adjacent sequence windows. e Mean expression breadth using EST data [101]**

volving transcription factor binding sites and/or from the transcription of their neighboring housekeeping genes. This activity could also be associated with an open chromatin structure, permissive to early replication and gene expression in most tissues [161,162,163,164]. This open conformation could extend along the first gene, possibly promoting the expression of further genes. This effect would progressively weaken with the distance from the putative replication origin, leading to the observed decrease in expression breadth. This model is consistent with a number of data showing that in metazoans, ORC and RNA polymerase II colocalize at transcriptional promoter regions [165], and that replication origins are determined by epigenetic information such as transcription factor binding sites and/or transcription [166,167,168,169]. It is also consistent with studies in *Drosophila* and humans that report correlation between early replication timing and increased probability of expression [137,155,156,165,170]. Furthermore, near the putative origins bordering the replication domains, transcription is preferentially oriented in the same direction as replication fork progression. This co-orientation is likely to reduce head-on collisions between the replication and transcription machineries, which may induce deleterious recombination events either directly or via stalling of the replication fork [171,172]. In bacteria, co-orientation of transcription and replication has been observed for essential genes, and has been associated with a reduction in head-on collisions between DNA and RNA polymerases [173]. It is noteworthy that in human replication domains such co-orientation usually occurs in widely-expressed genes located near putative replication origins. Near domain centers, head-on collisions may occur in 50% of replication cycles, regardless of the transcription orientation, since there is no preferential orientation of the replication fork progression in these regions. However, in most cell types, there should be few head-on collisions due to the low density and expression breadth of the corresponding genes. Selective pressure to reduce head-on collisions may thus have contributed to the simultaneous and coordinated organization of gene orientation and expression breadth along the detected replication domains [101].

Future Directions

From a statistical multifractal analysis of nucleotide strand asymmetries in mammalian genomes, we have revealed the existence of jumps in the noisy skew profiles resulting from asymmetries intrinsic to the transcription and replication processes [98,100]. This discovery has led us to extend our 1D WTMM methodology to an adapted multi-scale pattern recognition strategy in order to detect putative replication domains bordered by replication origins [101,114]. The results reported in this manuscript show that directly from the DNA sequence, we have been able to reveal the existence in the human genome (and very likely in all mammalian genomes), of regions bordered by early replicating origins in which gene position, orientation and expression breadth present a high level of organization, possibly mediated by the chromatin structure.

These results open new perspectives in DNA sequence analysis, chromatin modeling as well as in experiment. From a bioinformatic and modeling point of view, we plan to study the lexical and structural characteristics of our set of putative origins. In particular we will search for conserved sequence motifs in these replication initiation zones. Using a sequence-dependent model of DNA-histones interactions, we will develop physical studies of nucleosome formation and diffusion along the DNA fiber around the putative replication origins. These bioinformatic and physical studies, performed for the first time on a large number of replication origins, should shed light on the processes at work during the recognition of the replication initiation zone by the replication machinery. From an experimental point of view, our study raises new opportunities for future experiments. The first one concerns the experimental validation of the predicted replication origins (e. g. by molecular combing of DNA molecules [174]), which will allow us to determine precisely the existence of replication origins in given genome regions. Large scale study of all candidate origins is in current progress in the laboratory of O. Hyrien (École Normale Supérieure, Paris). The second experimental project consists in using Atomic Force Microscopy (AFM) [175] and Surface Plasmon Resonance Microscopy (SPRM) [176] to visualize and study the structural and mechanical properties of the DNA double helix, the nucleosomal string and the 30 nm chromatin fiber around the predicted replication origins. This work is in current progress in the experimental group of F. Argoul at the Laboratoire Joliot–Curie (ENS, Lyon) [83]. Finally the third experimental perspective concerns in situ studies of replication origins. Using fluorescence techniques (FISH chromosome painting [177]), we plan to study the distributions and dynamics of origins in the cell nucleus, as well as chromosome domains potentially associated with territories and their possible relation to nuclear matrix attachment sites. This study is likely to provide evidence of chromatin rosette patterns as suggested in [146]. This study is under progress in the molecular biology experimental group of F. Mongelard at the Laboratoire Joliot–Curie.

Acknowledgments

We thank O. Hyrien, F. Mongelard and C. Moskalenko for interesting discussions. This work was supported by the Action Concertée Incitative Informatique, Mathématiques, Physique en Biologie Moléculaire 2004 under the project "ReplicOr", the Agence Nationale de la Recherche under the project "HUGOREP" and the program "Emergence" of the Conseil Régional Rhônes-Alpes and by the Programme d'Actions Intégrées Tournesol.

Bibliography

Primary Literature

1. Goupillaud P, Grossmann A, Morlet J (1984) Cycle-octave and related transforms in seismic signal analysis. Geoexploration 23:85–102
2. Grossmann A, Morlet J (1984) Decomposition of Hardy functions into square integrable wavelets of constant shape. SIAM J Math Anal 15:723–736
3. Arneodo A, Argoul F, Bacry E, Elezgaray J, Freysz E, Grasseau G, Muzy J-F, Pouligny B (1992) Wavelet transform of fractals. In: Meyer Y (ed) Wavelets and applications. Springer, Berlin, pp 286–352
4. Arneodo A, Argoul F, Elezgaray J, Grasseau G (1989) Wavelet transform analysis of fractals: Application to nonequilibrium phase transitions. In: Turchetti G (ed) Nonlinear dynamics. World Scientific, Singapore, pp 130–180
5. Arneodo A, Grasseau G, Holschneider M (1988) Wavelet transform of multifractals. Phys Rev Lett 61:2281–2284
6. Holschneider M (1988) On the wavelet transform of fractal objects. J Stat Phys 50:963–993
7. Holschneider M, Tchamitchian P (1990) Régularité locale de la fonction non-différentiable de Riemann. In: Lemarié PG (ed) Les ondelettes en 1989. Springer, Berlin, pp 102–124
8. Jaffard S (1989) Hölder exponents at given points and wavelet coefficients. C R Acad Sci Paris Sér. I 308:79–81
9. Jaffard S (1991) Pointwise smoothness, two-microlocalization and wavelet coefficients. Publ Mat 35:155–168
10. Mallat S, Hwang W (1992) Singularity detection and processing with wavelets. IEEE Trans Info Theory 38:617–643
11. Mallat S, Zhong S (1992) Characterization of signals from multiscale edges. IEEE Trans Patt Recog Mach Intell 14:710–732
12. Arneodo A, Bacry E, Muzy J-F (1995) The thermodynamics of fractals revisited with wavelets. Physica A 213:232–275
13. Bacry E, Muzy J-F, Arneodo A (1993) Singularity spectrum of fractal signals from wavelet analysis: Exact results. J Stat Phys 70:635–674

14. Muzy J-F, Bacry E, Arneodo A (1991) Wavelets and multifractal formalism for singular signals: Application to turbulence data. Phys Rev Lett 67:3515–3518
15. Muzy J-F, Bacry E, Arneodo A (1993) Multifractal formalism for fractal signals: The structure-function approach versus the wavelet-transform modulus-maxima method. Phys Rev E 47:875–884
16. Muzy J-F, Bacry E, Arneodo A (1994) The multifractal formalism revisited with wavelets. Int J Bifurc Chaos 4:245–302
17. Jaffard S (1997) Multifractal formalism for functions part I: Results valid for all functions. SIAM J Math Anal 28:944–970
18. Jaffard S (1997) Multifractal formalism for functions part II: Self-similar functions. SIAM J Math Anal 28:971–998
19. Hentschel HGE (1994) Stochastic multifractality and universal scaling distributions. Phys Rev E 50:243–261
20. Arneodo A, Audit B, Decoster N, Muzy J-F, Vaillant C (2002) Wavelet based multifractal formalism: Application to DNA sequences, satellite images of the cloud structure and stock market data. In: Bunde A, Kropp J, Schellnhuber HJ (eds) The science of disasters: Climate disruptions, heart attacks, and market crashes. Springer, Berlin, pp 26–102
21. Arneodo A, Manneville S, Muzy J-F (1998) Towards log-normal statistics in high Reynolds number turbulence. Eur Phys J B 1:129–140
22. Arneodo A, Manneville S, Muzy J-F, Roux SG (1999) Revealing a lognormal cascading process in turbulent velocity statistics with wavelet analysis. Phil Trans R Soc Lond A 357:2415–2438
23. Delour J, Muzy J-F, Arneodo A (2001) Intermittency of 1D velocity spatial profiles in turbulence: A magnitude cumulant analysis. Eur Phys J B 23:243–248
24. Roux S, Muzy J-F, Arneodo A (1999) Detecting vorticity filaments using wavelet analysis: About the statistical contribution of vorticity filaments to intermittency in swirling turbulent flows. Eur Phys J B 8:301–322
25. Venugopal V, Roux SG, Foufoula-Georgiou E, Arneodo A (2006) Revisiting multifractality of high-resolution temporal rainfall using a wavelet-based formalism. Water Resour Res 42:W06D14
26. Venugopal V, Roux SG, Foufoula-Georgiou E, Arneodo A (2006) Scaling behavior of high resolution temporal rainfall: New insights from a wavelet-based cumulant analysis. Phys Lett A 348:335–345
27. Arneodo A, d'Aubenton-Carafa Y, Bacry E, Graves PV, Muzy J-F, Thermes C (1996) Wavelet based fractal analysis of DNA sequences. Physica D 96:291–320
28. Arneodo A, Bacry E, Graves PV, Muzy J-F (1995) Characterizing long-range correlations in DNA sequences from wavelet analysis. Phys Rev Lett 74:3293–3296
29. Audit B, Thermes C, Vaillant C, d'Aubenton Carafa Y, Muzy J-F, Arneodo A (2001) Long-range correlations in genomic DNA: A signature of the nucleosomal structure. Phys Rev Lett 86:2471–2474
30. Audit B, Vaillant C, Arneodo A, d'Aubenton-Carafa Y, Thermes C (2002) Long-range correlations between DNA bending sites: Relation to the structure and dynamics of nucleosomes. J Mol Biol 316:903–918
31. Arneodo A, Muzy J-F, Sornette D (1998) "Direct" causal cascade in the stock market. Eur Phys J B 2:277–282
32. Muzy J-F, Sornette D, Delour J, Arneodo A (2001) Multifractal returns and hierarchical portfolio theory. Quant Finance 1:131–148
33. Ivanov PC, Amaral LA, Goldberger AL, Havlin S, Rosenblum MG, Struzik ZR, Stanley HE (1999) Multifractality in human heartbeat dynamics. Nature 399:461–465
34. Ivanov PC, Rosenblum MG, Peng CK, Mietus J, Havlin S, Stanley HE, Goldberger AL (1996) Scaling behavior of heartbeat intervals obtained by wavelet-based time-series analysis. Nature 383:323–327
35. Arneodo A, Argoul F, Bacry E, Muzy J-F, Tabard M (1992) Golden mean arithmetic in the fractal branching of diffusion-limited aggregates. Phys Rev Lett 68:3456–3459
36. Arneodo A, Argoul F, Muzy J-F, Tabard M (1992) Structural 5-fold symmetry in the fractal morphology of diffusion-limited aggregates. Physica A 188:217–242
37. Arneodo A, Argoul F, Muzy J-F, Tabard M (1992) Uncovering Fibonacci sequences in the fractal morphology of diffusion-limited aggregates. Phys Lett A 171:31–36
38. Kuhn A, Argoul F, Muzy J-F, Arneodo A (1994) Structural-analysis of electroless deposits in the diffusion-limited regime. Phys Rev Lett 73:2998–3001
39. Arneodo A, Decoster N, Roux SG (2000) A wavelet-based method for multifractal image analysis, I. Methodology and test applications on isotropic and anisotropic random rough surfaces. Eur Phys J B 15:567–600
40. Arrault J, Arneodo A, Davis A, Marshak A (1997) Wavelet based multifractal analysis of rough surfaces: Application to cloud models and satellite data. Phys Rev Lett 79:75–78
41. Decoster N, Roux SG, Arneodo A (2000) A wavelet-based method for multifractal image analysis, II. Applications to synthetic multifractal rough surfaces. Eur Phys J B 15: 739–764
42. Arneodo A, Decoster N, Roux SG (1999) Intermittency, lognormal statistics, and multifractal cascade process in high-resolution satellite images of cloud structure. Phys Rev Lett 83:1255–1258
43. Roux SG, Arneodo A, Decoster N (2000) A wavelet-based method for multifractal image analysis, III. Applications to high-resolution satellite images of cloud structure. Eur Phys J B 15:765–786
44. Khalil A, Joncas G, Nekka F, Kestener P, Arneodo A (2006) Morphological analysis of H_I features, II. Wavelet-based multifractal formalism. Astrophys J Suppl Ser 165:512–550
45. Kestener P, Lina J-M, Saint-Jean P, Arneodo A (2001) Wavelet-based multifractal formalism to assist in diagnosis in digitized mammograms. Image Anal Stereol 20:169–174
46. Arneodo A, Decoster N, Kestener P, Roux SG (2003) A wavelet-based method for multifractal image analysis: From theoretical concepts to experimental applications. Adv Imaging Electr Phys 126:1–92
47. Kestener P, Arneodo A (2003) Three-dimensional wavelet-based multifractal method: The need for revisiting the multifractal description of turbulence dissipation data. Phys Rev Lett 91:194501
48. Meneveau C, Sreenivasan KR (1991) The multifractal nature of turbulent energy-dissipation. J Fluid Mech 224:429–484
49. Kestener P, Arneodo A (2004) Generalizing the wavelet-based multifractal formalism to random vector fields: Application to three-dimensional turbulence velocity and vorticity data. Phys Rev Lett 93:044501
50. Kestener P, Arneodo A (2007) A multifractal formalism for vector-valued random fields based on wavelet analysis: Application to turbulent velocity and vorticity 3D nu-

merical data. Stoch Environ Res Risk Assess. doi:10.1007/s00477-007-0121-6
51. Li WT, Marr TG, Kaneko K (1994) Understanding long-range correlations in DNA-sequences. Physica D 75:392–416
52. Stanley HE, Buldyrev SV, Goldberger AL, Havlin S, Ossadnik SM, Peng C-K, Simons M (1993) Fractal landscapes in biological systems. Fractals 1:283–301
53. Li W (1990) Mutual information functions versus correlation-functions. J Stat Phys 60:823–837
54. Li W (1992) Generating non trivial long-range correlations and 1/*f* spectra by replication and mutation. Int J Bifurc Chaos 2:137–154
55. Azbel' MY (1995) Universality in a DNA statistical structure. Phys Rev Lett 75:168–171
56. Herzel H, Große I (1995) Measuring correlations in symbol sequences. Physica A 216:518–542
57. Voss RF (1992) Evolution of long-range fractal correlations and 1/*f* noise in DNA base sequences. Phys Rev Lett 68: 3805–3808
58. Voss RF (1994) Long-range fractal correlations in DNA introns and exons. Fractals 2:1–6
59. Peng C-K, Buldyrev SV, Goldberger AL, Havlin S, Sciortino F, Simons M, Stanley HE (1992) Long-range correlations in nucleotide sequences. Nature 356:168–170
60. Havlin S, Buldyrev SV, Goldberger AL, Mantegna RN, Peng C-K, Simons M, Stanley HE (1995) Statistical and linguistic features of DNA sequences. Fractals 3:269–284
61. Mantegna RN, Buldyrev SV, Goldberger AL, Havlin S, Peng C-K, Simons M, Stanley HE (1995) Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. Phys Rev E 52:2939–2950
62. Herzel H, Ebeling W, Schmitt A (1994) Entropies of biosequences: The role of repeats. Phys Rev E 50:5061–5071
63. Li W (1997) The measure of compositional heterogeneity in DNA sequences is related to measures of complexity. Complexity 3:33–37
64. Borštnik B, Pumpernik D, Lukman D (1993) Analysis of apparent $1/f^{\alpha}$ spectrum in DNA sequences. Europhys Lett 23: 389–394
65. Chatzidimitriou-Dreismann CA, Larhammar D (1993) Long-range correlations in DNA. Nature 361:212–213
66. Nee S (1992) Uncorrelated DNA walks. Nature 357:450
67. Viswanathan GM, Buldyrev SV, Havlin S, Stanley HE (1998) Long-range correlation measures for quantifying patchiness: Deviations from uniform power-law scaling in genomic DNA. Physica A 249:581–586
68. Buldyrev SV, Goldberger AL, Havlin S, Mantegna RN, Matsa ME, Peng C-K, Simons M, Stanley HE (1995) Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. Phys Rev E 51:5084–5091
69. Berthelsen CL, Glazier JA, Raghavachari S (1994) Effective multifractal spectrum of a random walk. Phys Rev E 49: 1860–1864
70. Li W (1997) The study of correlation structures of DNA sequences: A critical review. Comput Chem 21:257–271
71. Peng C-K, Buldyrev SV, Goldberger AL, Havlin S, Simons M, Stanley HE (1993) Finite-size effects on long-range correlations: Implications for analyzing DNA sequences. Phys Rev E 47:3730–3733
72. Bernardi G (2000) Isochores and the evolutionary genomics of vertebrates. Gene 241:3–17
73. Gardiner K (1996) Base composition and gene distribution: Critical patterns in mammalian genome organization. Trends Genet 12:519–524
74. Li W, Stolovitzky G, Bernaola-Galván P, Oliver JL (1998) Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes. Genome Res 8: 916–928
75. Karlin S, Brendel V (1993) Patchiness and correlations in DNA sequences. Science 259:677–680
76. Larhammar D, Chatzidimitriou-Dreismann CA (1993) Biological origins of long-range correlations and compositional variations in DNA. Nucleic Acids Res 21:5167–5170
77. Peng C-K, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL (1994) Mosaic organization of DNA nucleotides. Phys Rev E 49:1685–1689
78. Arneodo A, d'Aubenton-Carafa Y, Audit B, Bacry E, Muzy J-F, Thermes C (1998) Nucleotide composition effects on the long-range correlations in human genes. Eur Phys J B 1: 259–263
79. Vaillant C, Audit B, Arneodo A (2005) Thermodynamics of DNA loops with long-range correlated structural disorder. Phys Rev Lett 95:068101
80. Vaillant C, Audit B, Thermes C, Arneodo A (2006) Formation and positioning of nucleosomes: effect of sequence-dependent long-range correlated structural disorder. Eur Phys J E 19:263–277
81. Yuan G-C, Liu Y-J, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ (2005) Genome-scale identification of nucleosome positions in S. cerevisiae. Science 309:626–630
82. Vaillant C, Audit B, Arneodo A (2007) Experiments confirm the influence of genome long-range correlations on nucleosome positioning. Phys Rev Lett 99:218103
83. Moukhtar J, Fontaine E, Faivre-Moskalenko C, Arneodo A (2007) Probing persistence in DNA curvature properties with atomic force microscopy. Phys Rev Lett 98:178101
84. Chargaff E (1951) Structure and function of nucleic acids as cell constituents. Fed Proc 10:654–659
85. Rudner R, Karkas JD, Chargaff E (1968) Separation of B. subtilis DNA into complementary strands, 3. Direct analysis. Proc Natl Acad Sci USA 60:921–922
86. Fickett JW, Torney DC, Wolf DR (1992) Base compositional structure of genomes. Genomics 13:1056–1064
87. Lobry JR (1995) Properties of a general model of DNA evolution under no-strand-bias conditions. J Mol Evol 40:326–330
88. Beletskii A, Grigoriev A, Joyce S, Bhagwat AS (2000) Mutations induced by bacteriophage T7 RNA polymerase and their effects on the composition of the T7 genome. J Mol Biol 300:1057–1065
89. Francino MP, Ochman H (2001) Deamination as the basis of strand-asymmetric evolution in transcribed Escherichia coli sequences. Mol Biol Evol 18:1147–1150
90. Frank AC, Lobry JR (1999) Asymmetric substitution patterns: A review of possible underlying mutational or selective mechanisms. Gene 238:65–77
91. Freeman JM, Plasterer TN, Smith TF, Mohr SC (1998) Patterns of genome organization in bacteria. Science 279:1827
92. Mrázek J, Karlin S (1998) Strand compositional asymmetry in bacterial and large viral genomes. Proc Natl Acad Sci USA 95:3720–3725
93. Rocha EP, Danchin A, Viari A (1999) Universal replication biases in bacteria. Mol Microbiol 32:11–16

94. Tillier ER, Collins RA (2000) The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. J Mol Evol 50:249–257
95. Green P, Ewing B, Miller W, Thomas PJ, Green ED (2003) Transcription-associated mutational asymmetry in mammalian evolution. Nat Genet 33:514–517
96. Touchon M, Nicolay S, Arneodo A, d'Aubenton-Carafa Y, Thermes C (2003) Transcription-coupled TA and GC strand asymmetries in the human genome. FEBS Lett 555:579–582
97. Touchon M, Arneodo A, d'Aubenton-Carafa Y, Thermes C (2004) Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. Nucleic Acids Res 32:4969–4978
98. Brodie of Brodie E-B, Nicolay S, Touchon M, Audit B, d'Aubenton-Carafa Y, Thermes C, Arneodo A (2005) From DNA sequence analysis to modeling replication in the human genome. Phys Rev Lett 94:248103
99. Nicolay S, Argoul F, Touchon M, d'Aubenton-Carafa Y, Thermes C, Arneodo A (2004) Low frequency rhythms in human DNA sequences: A key to the organization of gene location and orientation? Phys Rev Lett 93:108101
100. Touchon M, Nicolay S, Audit B, Brodie of Brodie E-B, d'Aubenton-Carafa Y, Arneodo A, Thermes C (2005) Replication-associated strand asymmetries in mammalian genomes: Toward detection of replication origins. Proc Natl Acad Sci USA 102:9836–9841
101. Huvet M, Nicolay S, Touchon M, Audit B, d'Aubenton-Carafa Y, Arneodo A, Thermes C (2007) Human gene organization driven by the coordination of replication and transcription. Genome Res 17:1278–1285
102. Arneodo A, Bacry E, Jaffard S, Muzy J-F (1997) Oscillating singularities on Cantor sets: A grand-canonical multifractal formalism. J Stat Phys 87:179–209
103. Arneodo A, Bacry E, Jaffard S, Muzy J-F (1998) Singularity spectrum of multifractal functions involving oscillating singularities. J Fourier Anal Appl 4:159–174
104. Parisi G, Frisch U (1985) Fully developed turbulence and intermittency. In: Ghil M, Benzi R, Parisi G (eds) Turbulence and predictability in geophysical fluid dynamics and climate dynamics. Proc of Int School. North-Holland, Amsterdam, pp 84–88
105. Collet P, Lebowitz J, Porzio A (1987) The dimension spectrum of some dynamical systems. J Stat Phys 47:609–644
106. Grassberger P, Badii R, Politi A (1988) Scaling laws for invariant measures on hyperbolic and non hyperbolic attractors. J Stat Phys 51:135–178
107. Halsey TC, Jensen MH, Kadanoff LP, Procaccia I, Shraiman BI (1986) Fractal measures and their singularities: The characterization of strange sets. Phys Rev A 33:1141–1151
108. Paladin G, Vulpiani A (1987) Anomalous scaling laws in multifractal objects. Phys Rep 156:147–225
109. Rand D (1989) The singularity spectrum for hyperbolic Cantor sets and attractors. Ergod Th Dyn Syst 9:527–541
110. Argoul F, Arneodo A, Elezgaray J, Grasseau G (1990) Wavelet analysis of the self-similarity of diffusion-limited aggregates and electrodeposition clusters. Phys Rev A 41:5537–5560
111. Farmer JD, Ott E, Yorke JA (1983) The dimension of chaotic attractors. Physica D 7:153–180
112. Grassberger P, Procaccia I (1983) Measuring the strangeness of strange attractors. Physica D 9:189–208
113. Bohr T, Tèl T (1988) The thermodynamics of fractals. In: Hao BL (ed) Direction in chaos, vol 2. World Scientific, Singapore, pp 194–237
114. Audit B, Nicolay S, Huvet M, Touchon M, d'Aubenton-Carafa Y, Thermes C, Arneodo A (2007) DNA replication timing data corroborate in silico human replication origin predictions. Phys Rev Lett 99:248102
115. Mandelbrot BB, van Ness JW (1968) Fractional Brownian motions, fractal noises and applications. SIAM Rev 10:422–437
116. Arneodo A, Bacry E, Muzy JF (1998) Random cascades on wavelet dyadic trees. J Math Phys 39:4142–4164
117. Benzi R, Biferale L, Crisanti A, Paladin G, Vergassola M, Vulpiani A (1993) A random process for the construction of multiaffine fields. Physica D 65:352–358
118. Mandelbrot BB (1974) Intermittent turbulence in self-similar cascades: Divergence of high moments and dimension of the carrier. J Fluid Mech 62:331–358
119. Arneodo A, Bacry E, Manneville S, Muzy JF (1998) Analysis of random cascades using space-scale correlation functions. Phys Rev Lett 80:708–711
120. Castaing B, Dubrulle B (1995) Fully-developed turbulence – A unifying point-of-view. J Phys II France 5:895–899
121. Novikov EA (1994) Infinitely divisible distributions in turbulence. Phys Rev E 50:3303–3305
122. Gojobori T, Li WH, Graur D (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. J Mol Evol 18:360–369
123. Li WH, Wu CI, Luo CC (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 21:58–71
124. Petrov DA, Hartl DL (1999) Patterns of nucleotide substitution in Drosophila and mammalian genomes. Proc Natl Acad Sci USA 96:1475–1479
125. Zhang Z, Gerstein M (2003) Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. Nucleic Acids Res 31:5338–5348
126. Duret L (2002) Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev 12:640–649
127. Shioiri C, Takahata N (2001) Skew of mononucleotide frequencies, relative abundance of dinucleotides, and DNA strand asymmetry. J Mol Evol 53:364–376
128. Svejstrup JQ (2002) Mechanisms of transcription-coupled DNA repair. Nat Rev Mol Cell Biol 3:21–29
129. Nicolay S, Brodie of Brodie E-B, Touchon M, Audit B, d'Aubenton-Carafa Y, Thermes C, Arneodo A (2007) Bifractality of human DNA strand-asymmetry profiles results from transcription. Phys Rev E 75:032902
130. Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, Kumar RM, Chevalier B, Johnstone SE, Cole MF, ichi Isono K, Koseki H, Fuchikami T, Abe K, Murray HL, Zucker JP, Yuan B, Bell GW, Herbolsheimer E, Hannett NM, Sun K, Odom DT, Otte AP, Volkert TL, Bartel DP, Melton DA, Gifford DK, Jaenisch R, Young RA (2006) Control of developmental regulators by polycomb in human embryonic stem cells. Cell 125:301–313
131. Jacob F, Brenner S, Cuzin F (1963) On the regulation of DNA replication in bacteria. Cold Spring Harb Symp Quant Biol 28:329–342
132. Bell SP, Dutta A (2002) DNA replication in eukaryotic cells. Annu Rev Biochem 71:333–374
133. Anglana M, Apiou F, Bensimon A, Debatisse M (2003) Dynamics of DNA replication in mammalian somatic cells: Nu-

cleotide pool modulates origin choice and interorigin spacing. Cell 114:385–394
134. Fisher D, Méchali M (2003) Vertebrate HoxB gene expression requires DNA replication. EMBO J 22:3737–3748
135. Gerbi SA, Bielinsky AK (2002) DNA replication and chromatin. Curr Opin Genet Dev 12:243–248
136. Hyrien O, Méchali M (1993) Chromosomal replication initiates and terminates at random sequences but at regular intervals in the ribosomal DNA of Xenopus early embryos. EMBO J 12:4511–4520
137. Schübeler D, Scalzo D, Kooperberg C, van Steensel B, Delrow J, Groudine M (2002) Genome-wide DNA replication profile for Drosophila melanogaster: A link between transcription and replication timing. Nat Genet 32:438–442
138. Gilbert DM (2001) Making sense of eukaryotic DNA replication origins. Science 294:96–100
139. Coverley D, Laskey RA (1994) Regulation of eukaryotic DNA replication. Annu Rev Biochem 63:745–776
140. Sasaki T, Sawado T, Yamaguchi M, Shinomiya T (1999) Specification of regions of DNA replication initiation during embryogenesis in the 65-kilobase DNApolalpha-dE2F locus of Drosophila melanogaster. Mol Cell Biol 19:547–555
141. Bogan JA, Natale DA, Depamphilis ML (2000) Initiation of eukaryotic DNA replication: Conservative or liberal? J Cell Physiol 184:139–150
142. Gilbert DM (2004) In search of the holy replicator. Nat Rev Mol Cell Biol 5:848–855
143. Demeret C, Vassetzky Y, Méchali M (2001) Chromatin remodeling and DNA replication: From nucleosomes to loop domains. Oncogene 20:3086–3093
144. McNairn AJ, Gilbert DM (2003) Epigenomic replication: linking epigenetics to DNA replication. Bioessays 25:647–656
145. Méchali M (2001) DNA replication origins: From sequence specificity to epigenetics. Nat Rev Genet 2:640–645
146. Arneodo A, d'Aubenton-Carafa Y, Audit B, Brodie of Brodie E-B, Nicolay S, St-Jean P, Thermes C, Touchon M, Vaillant C (2007) DNA in chromatin: From genome-wide sequence analysis to the modeling of replication in mammals. Adv Chem Phys 135:203–252
147. Bulmer M (1991) Strand symmetry of mutation rates in the beta-globin region. J Mol Evol 33:305–310
148. Francino MP, Ochman H (2000) Strand symmetry around the beta-globin origin of replication in primates. Mol Biol Evol 17:416–422
149. Gierlik A, Kowalczuk M, Mackiewicz P, Dudek MR, Cebrat S (2000) Is there replication-associated mutational pressure in the Saccharomyces cerevisiae genome? J Theor Biol 202: 305–314
150. Berezney R, Dubey DD, Huberman JA (2000) Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci. Chromosoma 108:471–484
151. Vassilev LT, Burhans WC, DePamphilis ML (1990) Mapping an origin of DNA replication at a single-copy locus in exponentially proliferating mammalian cells. Mol Cell Biol 10: 4685–4689
152. Codlin S, Dalgaard JZ (2003) Complex mechanism of site-specific DNA replication termination in fission yeast. EMBO J 22:3431–3440
153. Little RD, Platt TH, Schildkraut CL (1993) Initiation and termination of DNA replication in human rRNA genes. Mol Cell Biol 13:6600–6613
154. Santamaria D, Viguera E, Martinez-Robles ML, Hyrien O, Hernandez P, Krimer DB, Schvartzman JB (2000) Bi-directional replication and random termination. Nucleic Acids Res 28:2099–2107
155. White EJ, Emanuelsson O, Scalzo D, Royce T, Kosak S, Oakeley EJ, Weissman S, Gerstein M, Groudine M, Snyder M, Schübeler D (2004) DNA replication-timing analysis of human chromosome 22 at high resolution and different developmental states. Proc Natl Acad Sci USA 101:17771–17776
156. Woodfine K, Beare DM, Ichimura K, Debernardi S, Mungall AJ, Fiegler H, Collins VP, Carter NP, Dunham I (2005) Replication timing of human chromosome 6. Cell Cycle 4:172–176
157. Raghuraman MK, Winzeler EA, Collingwood D, Hunt S, Wodicka L, Conway A, Lockhart DJ, Davis RW, Brewer BJ, Fangman WL (2001) Replication dynamics of the yeast genome. Science 294:115–121
158. Watanabe Y, Fujiyama A, Ichiba Y, Hattori M, Yada T, Sakaki Y, Ikemura T (2002) Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: Disease-related genes in timing-switch regions. Hum Mol Genet 11:13–21
159. Costantini M, Clay O, Federico C, Saccone S, Auletta F, Bernardi G (2007) Human chromosomal bands: Nested structure, high-definition map and molecular basis. Chromosoma 116:29–40
160. Schmegner C, Hameister H, Vogel W, Assum G (2007) Isochores and replication time zones: A perfect match. Cytogenet Genome Res 116:167–172
161. Chakalova L, Debrand E, Mitchell JA, Osborne CS, Fraser P (2005) Replication and transcription: shaping the landscape of the genome. Nat Rev Genet 6:669–677
162. Gilbert N, Boyle S, Fiegler H, Woodfine K, Carter NP, Bickmore WA (2004) Chromatin architecture of the human genome: Gene-rich domains are enriched in open chromatin fibers. Cell 118:555–566
163. Hurst LD, Pál C, Lercher MJ (2004) The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet 5:299–310
164. Sproul D, Gilbert N, Bickmore WA (2005) The role of chromatin structure in regulating the expression of clustered genes. Nat Rev Genet 6:775–781
165. MacAlpine DM, Rodriguez HK, Bell SP (2004) Coordination of replication and transcription along a Drosophila chromosome. Genes Dev 18:3094–3105
166. Danis E, Brodolin K, Menut S, Maiorano D, Girard-Reydet C, Méchali M (2004) Specification of a DNA replication origin by a transcription complex. Nat Cell Biol 6:721–730
167. DePamphilis ML (2005) Cell cycle dependent regulation of the origin recognition complex. Cell Cycle 4:70–79
168. Ghosh M, Liu G, Randall G, Bevington J, Leffak M (2004) Transcription factor binding and induced transcription alter chromosomal c-myc replicator activity. Mol Cell Biol 24: 10193–10207
169. Lin CM, Fu H, Martinovsky M, Bouhassira E, Aladjem MI (2003) Dynamic alterations of replication timing in mammalian cells. Curr Biol 13:1019–1028
170. Jeon Y, Bekiranov S, Karnani N, Kapranov P, Ghosh S, MacAlpine D, Lee C, Hwang DS, Gingeras TR, Dutta A (2005) Temporal profile of replication of human chromosomes. Proc Natl Acad Sci USA 102:6419–6424
171. Deshpande AM, Newlon CS (1996) DNA replication fork pause sites dependent on transcription. Science 272:1030–1033

172. Takeuchi Y, Horiuchi T, Kobayashi T (2003) Transcription-dependent recombination and the role of fork collision in yeast rDNA. Genes Dev 17:1497–1506
173. Rocha EPC, Danchin A (2003) Essentiality, not expressiveness, drives gene-strand bias in bacteria. Nat Genet 34:377–378
174. Herrick J, Stanislawski P, Hyrien O, Bensimon A (2000) Replication fork density increases during DNA synthesis in X. laevis egg extracts. J Mol Biol 300:1133–1142
175. Zlatanova J, Leuba SH (2003) Chromatin fibers, one-at-a-time. J Mol Biol 331:1–19
176. Tassius C, Moskalenko C, Minard P, Desmadril M, Elezgaray J, Argoul F (2004) Probing the dynamics of a confined enzyme by surface plasmon resonance. Physica A 342:402–409
177. Müller WG, Rieder D, Kreth G, Cremer C, Trajanoski Z, McNally JG (2004) Generic features of tertiary chromatin structure as detected in natural chromosomes. Mol Cell Biol 24: 9359–9370

Books and Reviews

Fractals

Aharony A, Feder J (eds) (1989) Fractals in Physics, Essays in Honour of BB Mandelbrot. Physica D 38. North-Holland, Amsterdam
Avnir D (ed) (1988) The fractal approach to heterogeneous chemistry: surfaces, colloids, polymers. Wiley, New-York
Barabàsi AL, Stanley HE (1995) Fractals concepts in surface growth. Cambridge University Press, Cambridge
Ben Avraham D, Havlin S (2000) Diffusion and reactions in fractals and disordered systems. Cambridge University Press, Cambridge
Bouchaud J-P, Potters M (1997) Théorie des risques financiers. Cambridge University Press, Cambridge
Bunde A, Havlin S (eds) (1991) Fractals and disordered systems. Springer, Berlin
Bunde A, Havlin S (eds) (1994) Fractals in science. Springer, Berlin
Bunde A, Kropp J, Schellnhuber HJ (eds) (2002) The science of disasters: Climate disruptions, heart attacks and market crashes. Springer, Berlin
Family F, Meakin P, Sapoval B, Wood R (eds) (1995) Fractal aspects of materials. Material Research Society Symposium Proceedings, vol 367. MRS, Pittsburg
Family F, Vicsek T (1991) Dynamics of fractal surfaces. World Scientific, Singapore
Feder J (1988) Fractals. Pergamon, New-York
Frisch U (1995) Turbulence. Cambridge University Press, Cambridge
Mandelbrot BB (1982) The Fractal Geometry of Nature. Freeman, San Francisco
Mantegna RN, Stanley HE (2000) An introduction to econophysics. Cambridge University Press, Cambridge
Meakin P (1998) Fractals, scaling and growth far from equilibrium. Cambridge University Press, Cambridge
Peitgen HO, Jürgens H, Saupe D (1992) Chaos and fractals: New frontiers of science. Springer, New York
Peitgen HO, Saupe D (eds) (1987) The science of fractal images. Springer, New-York
Pietronero L, Tosatti E (eds) (1986) Fractals in physics. North-Holland, Amsterdam
Stanley HE, Osbrowski N (eds) (1986) On growth and form: Fractal and non-fractal patterns in physics. Martinus Nijhof, Dordrecht
Stanley HE, Ostrowski N (eds) (1988) Random fluctuations and pattern growth. Kluwer, Dordrecht
Vicsek T (1989) Fractal growth phenomena. World Scientific, Singapore
Vicsek T, Schlesinger M, Matsuchita M (eds) (1994) Fractals in natural science. World Scientific, Singapore
West BJ (1990) Fractal physiology and chaos in medicine. World Scientific, Singapore
West BJ, Deering W (1994) Fractal physiology for physicists: Levy statistics. Phys Rep 246:1–100
Wilkinson GG, Kanellopoulos J, Megier J (eds) (1995) Fractals in geoscience and remote sensing, image understanding research senes, vol 1. ECSC-EC-EAEC, Brussels

Wavelets

Abry P (1997) Ondelettes et turbulences. Diderot Éditeur, Art et Sciences, Paris
Arneodo A, Argoul F, Bacry E, Elezgaray J, Muzy J-F (1995) Ondelettes, multifractales et turbulences: de l'ADN aux croissances cristallines. Diderot Éditeur, Art et Sciences, Paris
Chui CK (1992) An introduction to wavelets. Academic Press, Boston
Combes J-M, Grossmann A, Tchamitchian P (eds) (1989) Wavelets. Springer, Berlin
Daubechies I (1992) Ten lectures on wavelets. SIAM, Philadelphia
Erlebacher G, Hussaini MY, Jameson LM (eds) (1996) Wavelets: Theory and applications. Oxford University Press, Oxford
Farge M, Hunt JCR, Vassilicos JC (eds) (1993) Wavelets, fractals and Fourier. Clarendon Press, Oxford
Flandrin P (1993) Temps-Fréquence. Hermès, Paris
Holschneider M (1996) Wavelets: An analysis tool. Oxford University Press, Oxford
Jaffard S, Meyer Y, Ryan RD (eds) (2001) Wavelets: Tools for science and technology. SIAM, Philadelphia
Lemarie PG (ed) (1990) Les ondelettes en 1989. Springer, Berlin
Mallat S (1998) A wavelet tour in signal processing. Academic Press, New-York
Meyer Y (1990) Ondelettes. Herman, Paris
Meyer Y (ed) (1992) Wavelets and applications. Springer, Berlin
Meyer Y, Roques S (eds) (1993) Progress in wavelets analysis and applications. Éditions Frontières, Gif-sur-Yvette
Ruskai MB, Beylkin G, Coifman R, Daubechies I, Mallat S, Meyer Y, Raphael L (eds) (1992) Wavelets and their applications. Jones and Barlett, Boston
Silverman BW, Vassilicos JC (eds) (2000) Wavelets: The key to intermittent information? Oxford University Press, Oxford
Torresani B (1998) Analyse continue par ondelettes. Éditions de Physique, Les Ulis

DNA and Chromatin

Alberts B, Watson J (1994) Molecular biology of the cell, 3rd edn. Garland Publishing, New-York
Calladine CR, Drew HR (1999) Understanding DNA. Academic Press, San Diego
Graur D, Li WH (1999) Fundamentals of molecular evolution. Sinauer Associates, Sunderland
Hartl DL, Jones EW (2001) Genetics: Analysis of genes and genomes. Jones and Bartlett, Sudbury
Kolchanov NA, Lim HA (1994) Computer analysis of genetic macromolecules: Structure, function and evolution. World Scientific, Singapore

Kornberg A, Baker TA (1992) DNA Replication. WH Freeman, New-York
Lewin B (1994) Genes V. Oxford University Press, Oxford
Sudbery P (1998) Human molecular genetics. Addison Wesley, Singapore
Van Holde, KE (1988) Chromatin. Springer, New-York
Watson JD, Gilman M, Witkowski J, Zoller M (1992) Recombinant DNA. Freeman, New-York
Wolfe AP (1998) Chromatin structure and function, 3rd edn. Academic Press, London

Fractal and Transfractal Scale-Free Networks

Hernán D. Rozenfeld, Lazaros K. Gallos, Chaoming Song, Hernán A. Makse
Levich Institute and Physics Department,
City College of New York, New York, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Fractality in Real-World Networks
Models: Deterministic Fractal and Transfractal Networks
Properties of Fractal and Transfractal Networks
Future Directions
Acknowledgments
Appendix: The Box Covering Algorithms
Bibliography

Glossary

Degree of a node Number of edges incident to the node.
Scale-free network Network that exhibits a wide (usually power-law) distribution of the degrees.
Small-world network Network for which the diameter increases logarithmically with the number of nodes.
Distance The length (measured in number of links) of the shortest path between two nodes.
Box Group of nodes. In a *connected box* there exists a path within the box between any pair of nodes. Otherwise, the box is *disconnected*.
Box diameter The longest distance in a box.

Definition of the Subject

The explosion in the study of complex networks during the last decade has offered a unique view in the structure and behavior of a wide range of systems, spanning many different disciplines [1]. The importance of complex networks lies mainly in their simplicity, since they can represent practically any system with interactions in a unified way by stripping complicated details and retaining the main features of the system. The resulting networks include only *nodes*, representing the interacting agents and *links*, representing interactions. The term 'interactions' is used loosely to describe any possible way that causes two nodes to form a link. Examples can be real physical links, such as the wires connecting computers in the Internet or roads connecting cities, or alternatively they may be virtual links, such as links in WWW homepages or acquaintances in societies, where there is no physical medium actually connecting the nodes.

The field was pioneered by the famous mathematician P. Erdős many decades ago, when he greatly advanced graph theory [27]. The theory of networks would have perhaps remained a problem of mathematical beauty, if it was not for the discovery that a huge number of everyday life systems share many common features and can thus be described through a unified theory. The remarkable diversity of these systems incorporates artificially or man-made technological networks such as the Internet and the World Wide Web (WWW), social networks such as social acquaintances or sexual contacts, biological networks of natural origin, such as the network of protein interactions of Yeast [1,36], and a rich variety of other systems, such as proximity of words in literature [48], items that are bought by the same people [16] or the way modules are connected to create a piece of software, among many others.

The advances in our understanding of networks, combined with the increasing availability of many databases, allows us to analyze and gain deeper insight into the main characteristics of these complex systems. A large number of complex networks share the *scale-free* property [1,28], indicating the presence of few highly connected nodes (usually called hubs) and a large number of nodes with small degree. This feature alone has a great impact on the analysis of complex networks and has introduced a new way of understanding these systems. This property carries important implications in many everyday life problems, such as the way a disease spreads in communities of individuals, or the resilience and tolerance of networks under random and intentional attacks [19,20,21,31,59].

Although the scale-free property holds an undisputed importance, it has been shown to not completely determine the global structure of networks [6]. In fact, two networks that obey the same distribution of the degrees may dramatically differ in other fundamental structural properties, such as in correlations between degrees or in the average distance between nodes. Another fundamental prop-

erty, which is the focus of this article, is the presence of self-similarity or fractality. In simpler terms, we want to know whether a subsection of the network looks much the same as the whole [8,14,29,66]. In spite of the fact that in regular fractal objects the distinction between self-similarity and fractality is absent, in network theory we can distinguish the two terms: in a *fractal network* the number of boxes of a given size that are needed to completely cover the network scales with the box size as a power law, while a *self-similar network* is defined as a network whose degree distribution remains invariant under renormalization of the network (details on the renormalization process will be provided later). This essential result allows us to better understand the origin of important structural properties of networks such as the power-law degree distribution [35,62,63].

Introduction

Self-similarity is a property of fractal structures, a concept introduced by Mandelbrot and one of the fundamental mathematical results of the 20th century [29,45,66]. The importance of fractal geometry stems from the fact that these structures were recognized in numerous examples in Nature, from the coexistence of liquid/gas at the critical point of evaporation of water [11,39,65], to snowflakes, to the tortuous coastline of the Norwegian fjords, to the behavior of many complex systems such as economic data, or the complex patterns of human agglomeration [29,66].

Typically, real world scale-free networks exhibit the small-world property [1], which implies that the number of nodes increases exponentially with the diameter of the network, rather than the power-law behavior expected for self-similar structures. For this reason complex networks were believed to *not* be length-scale invariant or self-similar.

In 2005, C. Song, S. Havlin and H. Makse presented an approach to analyze complex networks, that reveals their self-similarity [62]. This result is achieved by the application of a renormalization procedure which coarse-grains the system into boxes containing nodes within a given size [62,64]. As a result, a power-law relation between the number of boxes needed to cover the network and the size of the box is found, defining a finite self-similar exponent. These fundamental properties, which are shown for the WWW, cellular and protein-protein interaction networks, help to understand the emergence of the scale-free property in complex networks. They suggest a common self-organization dynamics of diverse networks at different scales into a critical state and in turn bring together previously unrelated fields: the statistical physics of complex networks with renormalization group, fractals and critical phenomena.

Fractality in Real-World Networks

The study of real complex networks has revealed that many of them share some fundamental common properties. Of great importance is the form of the degree distribution for these networks, which is unexpectedly wide. This means that the degree of a node may assume values that span many decades. Thus, although the majority of nodes have a relatively small degree, there is a finite probability that a few nodes will have degree of the order of thousands or even millions. Networks that exhibit such a wide distribution $P(k)$ are known as *scale-free* networks, where the term refers to the absence of a characteristic scale in the degree k. This distribution very often obeys a power-law form with a degree exponent γ, usually in the range $2 < \gamma < 4$ [2],

$$P(k) \sim k^{-\gamma} \,. \tag{1}$$

A more generic property, that is usually inherent in scale-free networks but applies equally well to other types of networks, such as in Erdős–Rényi random graphs, is the *small-world* feature. Originally discovered in sociological studies [47], it is the generalization of the famous 'six degrees of separation' and refers to the very small network diameter. Indeed, in small-world networks a very small number of steps is required to reach a given node starting from any other node. Mathematically this is expressed by the slow (logarithmic) increase of the average diameter of the network, $\bar{\ell}$, with the total number of nodes N, $\bar{\ell} \sim \ln N$, where ℓ is the *shortest* distance between two nodes and defines the distance metric in complex networks [2,12,27,67], namely,

$$N \sim \mathrm{e}^{\bar{\ell}/\ell_0} \,, \tag{2}$$

where ℓ_0 is a characteristic length.

These network characteristics have been shown to apply in many empirical studies of diverse systems [1,2,28]. The simple knowledge that a network has the scale-free and/or small-world property already enables us to qualitatively recognize many of its basic properties. However, structures that have the same degree exponents may still differ in other aspects [6]. For example, a question of fundamental importance is whether scale-free networks are also self-similar or fractals. The illustrations of scale-free networks (see, e. g,. Figs. 1 and 2b) seem to resemble traditional fractal objects. Despite this similarity, Eq. (2) definitely appears to contradict a basic property of fractality: fast increase of the diameter with the system size. More-

over, a fractal object should be self-similar or invariant under a scale transformation, which is again not clear in the case of scale-free networks where the scale has necessarily limited range. So, how is it even possible that fractal scale-free networks exist? In the following, we will see how these seemingly contradictory aspects can be reconciled.

Fractality and Self-Similarity

The classical theory of self-similarity requires a power-law relation between the number of nodes N and the diameter of a fractal object ℓ [8,14]. The fractal dimension can be calculated using either *box-counting* or *cluster-growing* techniques [66]. In the first method the network is covered with N_B boxes of linear size ℓ_B. The fractal dimension or box dimension d_B is then given by [29]:

$$N_B \sim \ell_B^{-d_B} \,. \tag{3}$$

In the second method, instead of covering the network with boxes, a random seed node is chosen and nodes centered at the seed are grown so that they are separated by a maximum distance ℓ. The procedure is then repeated by choosing many seed nodes at random and the average "mass" of the resulting clusters, $\langle M_c \rangle$ (defined as the number of nodes in the cluster) is calculated as a function of ℓ to obtain the following scaling:

$$\langle M_c \rangle \sim \ell^{d_f}, \tag{4}$$

defining the fractal cluster dimension d_f [29]. If we use Eq. (4) for a small-world network, then Eq. (2) readily implies that $d_f = \infty$. In other words, these networks cannot be characterized by a finite fractal dimension, and should be regarded as infinite-dimensional objects. If this were true, though, local properties in a part of the network would not be able to represent the whole system. Still, it is also well established that the scale-free nature is similar in different parts of the network. Moreover, a graphical representation of real-world networks allows us to see that those systems seem to be built by attaching (following some rule) copies of itself.

The answer lies in the inherent inhomogeneity of the network. In the classical case of a *homogeneous* system (such as a fractal percolation cluster) the degree distribution is very narrow and the two methods described above are fully equivalent, because of this local neighborhood invariance. Indeed, all boxes in the box-covering method are statistically similar with each other as well as with the

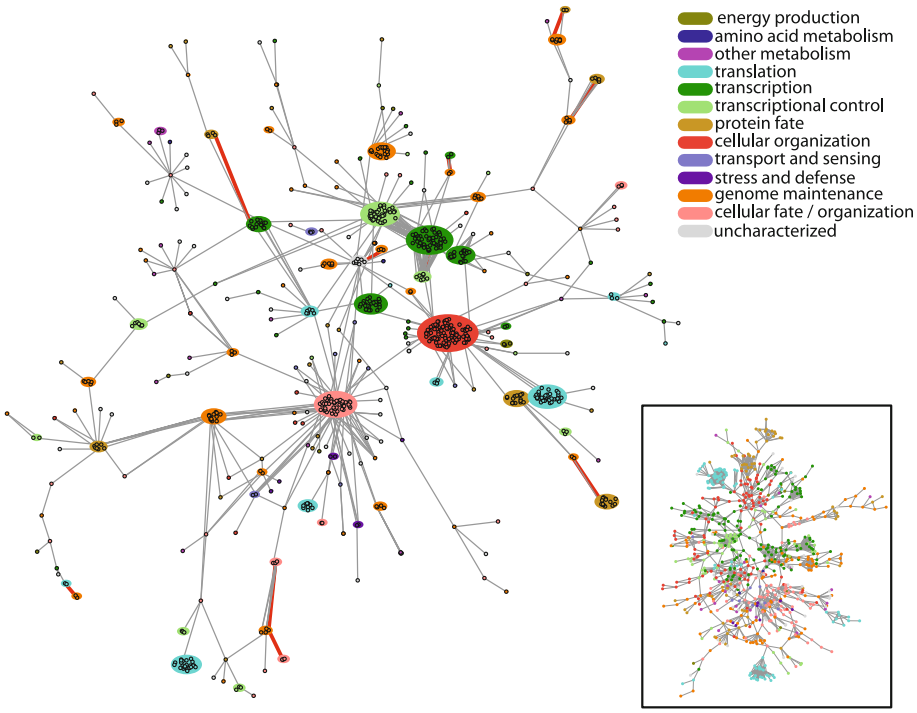

Fractal and Transfractal Scale-Free Networks, Figure 1
Representation of the Protein Interaction Network of Yeast. The colors show different subgroups of proteins that participate in different functionality classes [36]

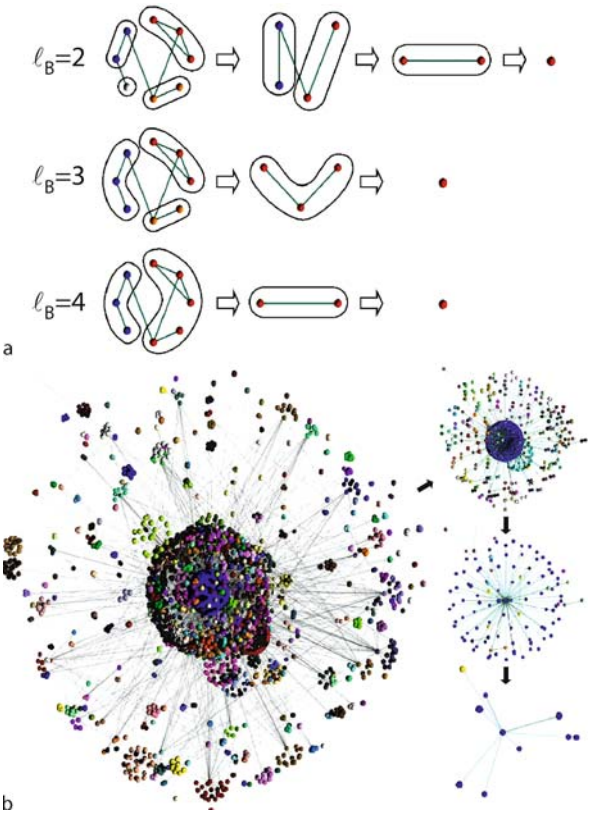

Fractal and Transfractal Scale-Free Networks, Figure 2
The renormalization procedure for complex networks. a Demonstration of the method for different ℓ_B and different stages in a network demo. The first column depicts the original network. The system is tiled with boxes of size ℓ_B (different colors correspond to different boxes). All nodes in a box are connected by a minimum distance smaller than the given ℓ_B. For instance, in the case of $\ell_B = 2$, one identifies four boxes which contain the nodes depicted with colors red, orange, white, and blue, each containing 3, 2, 1, and 2 nodes, respectively. Then each box is replaced by a single node; two renormalized nodes are connected if there is at least one link between the unrenormalized boxes. Thus we obtain the network shown in the second column. The resulting number of boxes needed to tile the network, $N_B(\ell_B)$, is plotted in Fig. 3 vs. ℓ_B to obtain d_B as in Eq. (3). The renormalization procedure is applied again and repeated until the network is reduced to a single node (third and fourth columns for different ℓ_B). b Three stages in the renormalization scheme applied to the entire WWW. We fix the box size to $\ell_B = 3$ and apply the renormalization for four stages. This corresponds, for instance, to the sequence for the network demo depicted in the second row in part a of this figure. We color the nodes in the web according to the boxes to which they belong

boxes grown when using the cluster-growing technique, so that Eq. (4) can be derived from Eq. (3) and $d_B = d_f$.

In *inhomogeneous* systems, though, the local environment can vary significantly. In this case, Eqs. (3) and (4) are no longer equivalent. If we focus on the box-covering technique then we want to cover the entire network with the minimum possible number of boxes $N_B(\ell_B)$, where the distance between any two nodes that belong in a box is smaller than ℓ_B. An example is shown in Fig. 2a using a simple 8-node network. After we repeat this procedure for different values of ℓ_B we can plot N_B vs. ℓ_B.

When the box-covering method is applied to real large-scale networks, such as the WWW [2] (http://www.nd.edu/~networks), the network of protein interaction of *H. sapiens* and *E. coli* [25,68] and several cellular networks [38,52], then they follow Eq. (3) with a clear power-law, indicating the *fractal* nature of these systems (Figs. 3a,b,c). On the other hand when the method is applied to other real world networks such as the Internet [24] or the Barabási–Albert network [7], they do not satisfy Eq. (3), which manifests that these networks are *not* fractal.

The reason behind the discrepancy in the fractality of homogeneous and inhomogeneous systems can be better clarified studying the mass of the boxes. For a given ℓ_B value, the average mass of a box $\langle M_B(\ell_B)\rangle$ is

$$\langle M_B(\ell_B)\rangle \equiv \frac{N}{N_B(\ell_B)} \sim \ell_B^{d_B} , \tag{5}$$

as also verified in Fig. 3 for several real-world networks. On the other hand, the average performed in the cluster growing method (averaging over single boxes without tiling the system) gives rise to an exponential growth of the mass

$$\langle M_c(\ell)\rangle \sim e^{\ell/\ell_1} , \tag{6}$$

in accordance with the small-world effect, Eq. (2). Correspondingly, the probability distribution of the mass of the boxes M_B using box-covering is very broad, while the cluster-growing technique leads to a narrow probability distribution of M_c.

The topology of scale-free networks is dominated by several highly connected hubs—the nodes with the largest degree—implying that most of the nodes are connected to the hubs via one or very few steps. Therefore, the average performed in the cluster growing method is biased; the hubs are overrepresented in Eq. (6) since almost every node is a neighbor of a hub, and there is always a very large probability of including the same hubs in all clusters. On the other hand, the box covering method is a global tiling of the system providing a flat average over all the nodes, i. e. each part of the network is covered with an equal probability. Once a hub (or any node) is covered, it cannot be covered again.

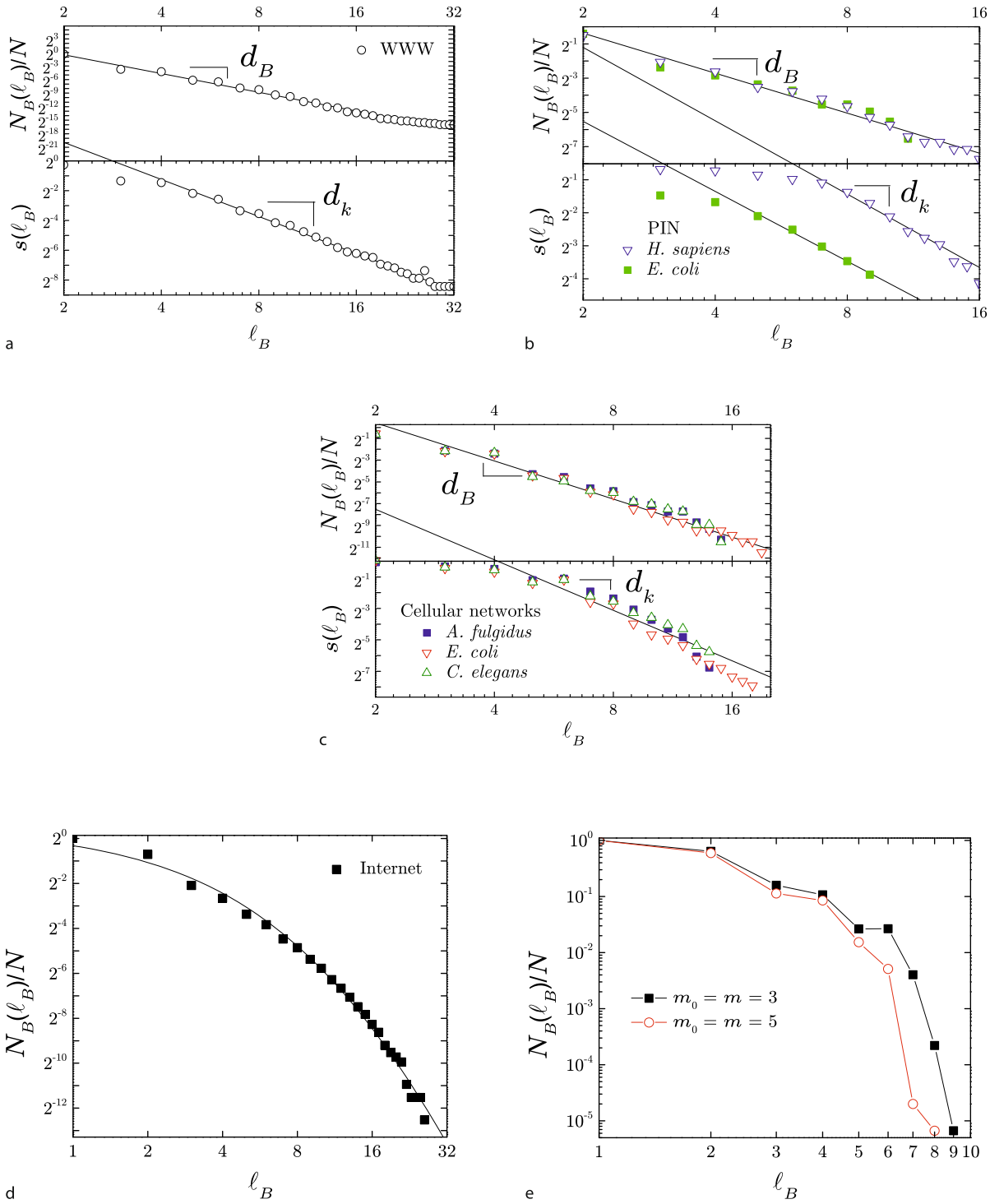

Fractal and Transfractal Scale-Free Networks, Figure 3

Self-similar scaling in complex networks. a *Upper panel*: Log-log plot of the N_B vs. ℓ_B revealing the self-similarity of the WWW according to Eq. (3). *Lower panel*: The scaling of $s(\ell_B)$ vs. ℓ_B according to Eq. (9). **b** Same as a but for two protein interaction networks: *H. sapiens* and *E. coli*. Results are analogous to b but with different scaling exponents. **c** Same as a for the cellular networks of *A. fulgidus*, *E. coli* and *C. elegans*. **d** Internet. Log-log plot of $N_B(\ell_B)$. The solid line shows that the internet [24] is *not* a fractal network since it does not follow the power-law relation of Eq. (5). **e** Same as **d** for the Barabási–Albert model network [7] with $m = 3$ and $m = 5$

In conclusion, we can state that the two dominant methods that are routinely used for calculations of fractality and give rise to Eqs. (3) and (4) are not equivalent in scale-free networks, but rather highlight different aspects: box covering reveals the self-similarity, while cluster growth reveals the small-world effect. The apparent contradiction is due to the hubs being used many times in the latter method.

Scale-free networks can be classified into three groups: (i) pure fractal, (ii) pure small-world and (iii) a mixture between fractal and small-world. (i) A fractal network satisfies Eq. (3) at all scales, meaning that for any value of ℓ_B, the number of boxes always follows a power-law (examples are shown in Fig. 3a,b,c). (ii) When a network is a pure small-world, it never satisfies Eq. (3). Instead, N_B follows an exponential decay with ℓ_B and the network cannot be regarded as fractal. Figures 3d and 3e show two examples of pure small-world networks. (iii) In the case of mixture between fractal and small-world, Eq. (3) is satisfied up to some cut-off value of ℓ_B, above which the fractality breaks down and the small-world property emerges. The small-world property is reflected in the plot of N_B vs. ℓ_B as an exponential cut-off for large ℓ_B.

We can also understand the coexistence of the small-world property and the fractality through a more intuitive approach. In a pure fractal network the length of a path between any pair of nodes scales as a power-law with the number of nodes in the network. Therefore, the diameter L also follows a power-law, $L \sim N^{1/d_B}$. If one adds a few shortcuts (links between randomly chosen nodes), many paths in the network are drastically shortened and the small-world property emerges as $L \sim \mathrm{Log}N$. In spite of this fact, for shorter scales, $\ell_B \ll L$, the network still behaves as a fractal. In this sense, we can say that globally the network is small-world, but locally (for short scales) the network behaves as a fractal. As more shortcuts are added, the cut-off in a plot of N_B vs. ℓ_B appears for smaller ℓ_B, until the network becomes a pure small-world for which all paths lengths increase logarithmically with N.

The reasons why certain networks have evolved towards a fractal or non-fractal structure will be described later, together with models and examples that provide additional insight into the processes involved.

Renormalization

Renormalization is one of the most important techniques in modern Statistical Physics [17,39,58]. The idea behind this procedure is to continuously create smaller replicas of a given object, retaining at the same time the essential structural features, and hoping that the coarse-grained copies will be more amenable to analytic treatment.

The idea for renormalizing the network emerges naturally from the concept of fractality described above. If a network is self-similar, then it will look more or less the same under different scales. The way to observe these different length-scales is based on renormalization principles, while the criterion to decide on whether a renormalized structure retains its form is the invariance of the main structural features, expressed mainly through the degree distribution.

The method works as follows. Start by fixing the value of ℓ_B and applying the box-covering algorithm in order to cover the entire network with boxes (see Appendix). In the renormalized network each box is replaced by a single node and two nodes are connected if there existed at least one connection between the two corresponding boxes in the original network. The resulting structure represents the first stage of the renormalized network. We can apply the same procedure to this new network, as well, resulting in the second renormalization stage network, and so on until we are left with a single node.

The second column of the panels in Fig. 2a shows this step in the renormalization procedure for the schematic network, while Fig. 2b shows the results for the same procedure applied to the entire WWW for $\ell_B = 3$.

The renormalized network gives rise to a new probability distribution of links, $P(k')$ (we use a prime $'$ to denote quantities in the renormalized network). This distribution remains invariant under the renormalization:

$$P(k) \to P(k') \sim (k')^{-\gamma} . \tag{7}$$

Fig. 4 supports the validity of this scale transformation by showing a data collapse of all distributions with the same γ according to (7) for the WWW.

Here, we present the basic scaling relations that characterize renormalizable networks. The degree k' of each node in the renormalized network can be seen to scale with the largest degree k in the corresponding original box as

$$k \to k' = s(\ell_B)k . \tag{8}$$

This equation defines the scaling transformation in the connectivity distribution. Empirically, it was found that the scaling factor $s(< 1)$ scales with ℓ_B with a new exponent, d_k, as $s(\ell_B) \sim \ell_B^{-d_k}$, so that

$$k' \sim \ell_B^{-d_k} k , \tag{9}$$

This scaling is verified for many networks, as shown in Fig. 3.

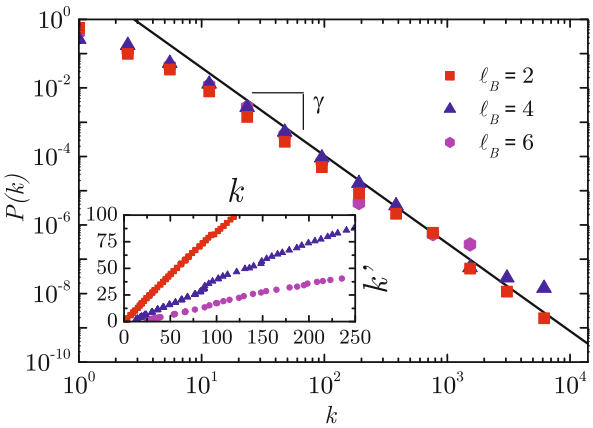

Fractal and Transfractal Scale-Free Networks, Figure 4
Invariance of the degree distribution of the WWW under the renormalization for different box sizes, ℓ_B. We show the data collapse of the degree distributions demonstrating the self-similarity at different scales. The *inset* shows the scaling of $k' = s(\ell_B)k$ for different ℓ_B, from where we obtain the scaling factor $s(\ell_B)$. Moreover, renormalization for a fixed box size ($\ell_B = 3$) is applied, until the network is reduced to a few nodes. It was found that $P(k)$ is invariant under these multiple renormalizations procedures

The exponents γ, d_B, and d_k are not all independent from each other. The proof starts from the density balance equation $n(k)\mathrm{d}k = n'(k')\mathrm{d}k'$, where $n(k) = NP(k)$ is the number of nodes with degree k and $n'(k') = N'P(k')$ is the number of nodes with degree k' after the renormalization (N' is the total number of nodes in the renormalized network). Substituting Eq. (8) leads to $N' = s^{\gamma-1}N$. Since the total number of nodes in the renormalized network is the number of boxes needed to cover the unrenormalized network at any given ℓ_B we have the identity $N' = N_B(\ell_B)$. Finally, from Eqs. (3) and (9) one obtains the relation between the three indexes

$$\gamma = 1 + \frac{d_B}{d_k} . \tag{10}$$

The use of Eq. (10) yields the same γ exponent as that obtained in the direct calculation of the degree distribution. The significance of this result is that the scale-free properties characterized by γ can be related to a more fundamental length-scale invariant property, characterized by the two new indexes d_B and d_k.

We have seen, thus, that concepts introduced originally for the study of critical phenomena in statistical physics, are also valid in the characterization of a different class of phenomena: the topology of complex networks. A large number of scale-free networks are fractals and an even larger number remain invariant under a scale-transformation. The influence of these features on the network properties will be delayed until the sixth chapter, after we introduce some algorithms for efficient numerical calculations and two theoretical models that give rise to fractal networks.

Models: Deterministic Fractal and Transfractal Networks

The first model of a scale-free fractal network was presented in 1979 when N. Berker and S. Ostlund [9] proposed a hierarchical network that served as an exotic example where renormalization group techniques yield exact results, including the percolation phase transition and the $q \to 1$ limit of the Potts model. Unfortunately, in those days the importance of the power-law degree distribution and the concept of fractal and non-fractal complex networks were not known. Much work has been done on these types of hierarchical networks. For example, in 1984, M. Kaufman and R. Griffiths made use of Berker and Ostlund's model to study the percolation phase transition and its percolation exponents [22,37,40,41].

Since the late 90s, when the importance of the power-law degree distribution was first shown [1] and after the finding of C. Song, S. Havlin and H. Makse [62], many hierarchical networks that describe fractality in complex networks have been proposed. These artificial models are of great importance since they provide insight into the origins and fundamental properties that give rise to the fractality and non-fractality of networks.

The Song–Havlin–Makse Model

The correlations between degrees in a network [46,49,50,54] are quantified through the probability $P(k_1, k_2)$ that a node of degree k_1 is connected to another node of degree k_2. In Fig. 5 we can see the degree correlation profile $R(k_1, k_2) = P(k_1, k_2)/P_r(k_1, k_2)$ of the cellular metabolic network of E. *coli* [38] (known to be a fractal network) and for the Internet at the router level [15] (a non-fractal network), where $P_r(k_1, k_2)$ is obtained by randomly swapping the links without modifying the degree distribution.

Figure 5 shows a dramatic difference between the two networks. The network of E. *coli*, that is a fractal network, presents an anti-correlation of the degrees (or dissasortativity [49,50]), which means that mostly high degree nodes are linked to low degree nodes. This property leads to fractal networks. On the other hand, the Internet exhibits a high correlation between degrees leading to a non-fractal network.

With this idea in mind, in 2006 C. Song, S. Havlin and H. Makse presented a model that elucidates the way

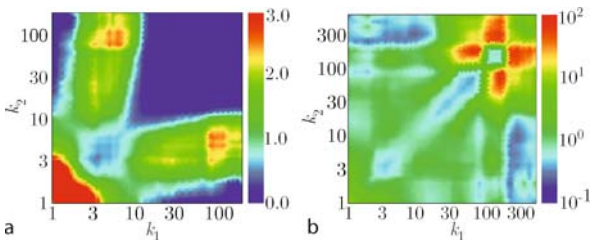

Fractal and Transfractal Scale-Free Networks, Figure 5
Degree correlation profile for a the cellular metabolic network of E. *coli*, and b the Internet at the router level

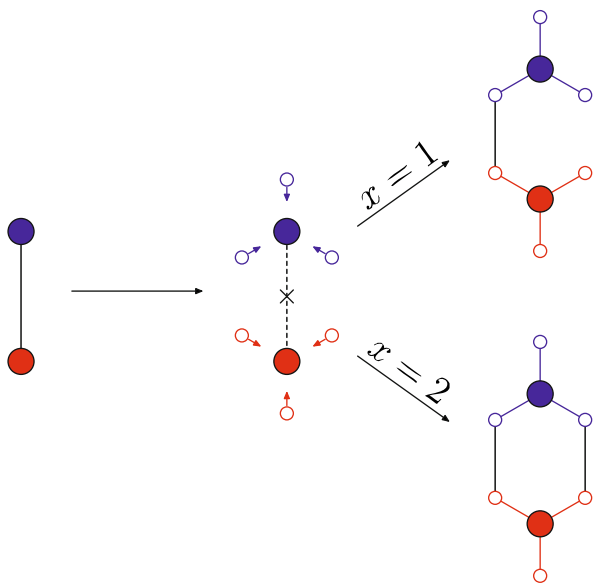

Fractal and Transfractal Scale-Free Networks, Figure 6
The model grows from a small network, usually two nodes connected to each other. During each step and for every link in the system, each endpoint of a link produces *m* offspring nodes (in this drawing $m = 3$). In this case, with probability $e = 1$ the original link is removed and *x* new links between randomly selected nodes of the new generation are added. Notice that the case of $x = 1$ results in a tree structure, while loops appear for $x > 1$

new nodes must be connected to the old ones in order to build a fractal, a non-fractal network, or a mixture between fractal and non-fractal network [63]. This model shows that, indeed, the correlations between degrees of the nodes are a determinant factor for the fractality of a network. This model was later extended [32] to allow loops in the network, while preserving the self-similarity and fractality properties.

The algorithm is as follows (see Fig. 6): In generation $n = 0$, start with two nodes connected by one link. Then, generation $n + 1$ is obtained recursively by attaching m new nodes to the endpoints of each link l of generation n. In addition, with probability e remove link l and add x new links connecting pairs of new nodes attached to the endpoints of l.

The degree distribution, diameter and fractal dimension can be easily calculated. For example, if $e = 1$ (pure fractal network), the degree distribution follows a power-law $P(k) \sim k^{-\gamma}$ with exponent $\gamma = 1 + \log(2m + x)/\log m$ and the fractal dimension is $d_B = \log(2m + x)/\log m$. The diameter L scales, in this case, as power of the number of nodes as $L \sim N^{1/d_B}$ [63,64]. Later, in Sect. "Properties of Fractal and Transfractal Networks", several topological properties are shown for this model network.

(u, v)-Flowers

In 2006, H. Rozenfeld, S. Havlin and D. ben-Avraham proposed a new family of recursive *deterministic* scale-free networks, the (u, v)-*flowers*, that generalize both, the original scale-free model of Berker and Ostlund [9] and the pseudo-fractal network of Dorogovstev, Goltsev and Mendes [26] and that, by appropriately varying its two parameters u and v, leads to either fractal networks or non-fractal networks [56,57]. The algorithm to build the (u, v)-*flowers* is the following: In generation $n = 1$ one starts with a cycle graph (a ring) consisting of $u + v \equiv w$ links and nodes (other choices are possible). Then, generation $n + 1$ is obtained recursively by replacing each link by two parallel paths of u and v links long. Without loss of generality, $u \leq v$. Examples of $(1, 3)$- and $(2, 2)$-*flowers* are shown in Fig. 7. The DGM network corresponds to the special case of $u = 1$ and $v = 2$ and the Berker and Ostlund model corresponds to $u = 2$ and $v = 2$.

An essential property of the (u, v)-*flowers* is that they are self-similar, as evident from an equivalent method of construction: to produce generation $n + 1$, make $w = u + v$ copies of the net in generation n and join them at the hubs.

The number of links of a (u, v)-flower of generation n is

$$M_n = (u + v)^n = w^n \,, \tag{11}$$

and the number of nodes is

$$N_n = \left(\frac{w-2}{w-1}\right) w^n + \left(\frac{w}{w-1}\right) . \tag{12}$$

The degree distribution of the (u, v)-*flowers* can also be easily obtained since by construction, (u, v)-*flowers* have only nodes of degree $k = 2^m$, $m = 1, 2, \ldots, n$. As in the DGM case, (u, v)-*flowers* follow a scale-free degree distri-

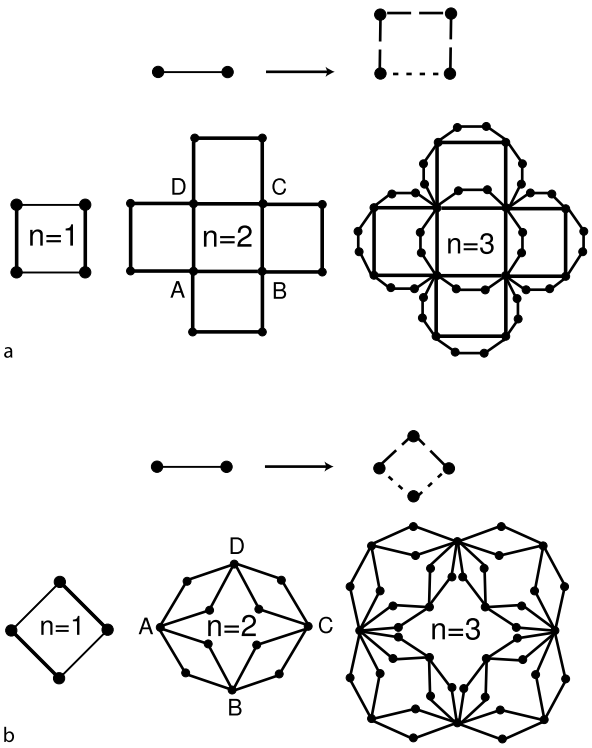

Fractal and Transfractal Scale-Free Networks, Figure 7
(u, v)-*flowers* with $u + v = 4(\gamma = 3)$. a $u = 1$ (*dotted line*) and $v = 3$ (*broken line*). b $u = 2$ and $v = 2$. The graphs may also be iterated by joining four replicas of generation n at the hubs A and B, for a, or A and C, for b

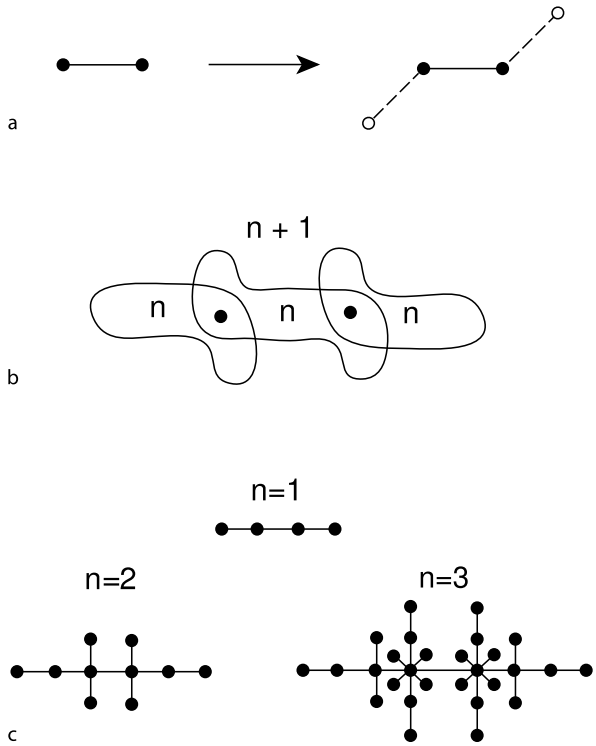

Fractal and Transfractal Scale-Free Networks, Figure 8
The (1, 2)-tree. a Each link in generation n is replaced by a chain of $u = 1$ links, to which ends one attaches chains of $v/2 = 1$ links. b Alternative method of construction highlighting self-similarity: $u + v = 3$ replicas of generation n are joined at the hubs. c Generations $n = 1, 2, 3$

bution, $P(k) \sim k^{-\gamma}$, of degree exponent

$$\gamma = 1 + \frac{\ln(u+v)}{\ln 2} \,. \tag{13}$$

Recursive scale-free *trees* may be defined in analogy to the flower nets. If v is even, one obtains generation $n + 1$ of a (u, v)-tree by replacing every link in generation n with a chain of u links, and attaching to each of its endpoints chains of $v/2$ links. Figure 8 shows how this works for the $(1, 2)$-tree. If v is odd, attach to the endpoints (of the chain of u links) chains of length $(v \pm 1)/2$. The trees may be also constructed by successively joining w replicas at the appropriate hubs, and they too are self-similar. They share many of the fundamental scaling properties with (u, v)-*flowers*: Their degree distribution is also scale-free, with the same degree exponent as (u, v)-*flowers*.

The self-similarity of (u, v)-*flowers*, coupled with the fact that different replicas meet at a *single* node, makes them amenable to exact analysis by renormalization techniques. The lack of loops, in the case of (u, v)-trees, further simplifies their analysis [9,13,56,57].

Dimensionality of the (u, v)-Flowers There is a vast difference between (u, v)-nets with $u = 1$ and $u > 1$. If $u = 1$ the diameter L_n of the nth generation flower scales linearly with n. For example, L_n for the $(1, 2)$-flower [26] and $L_n = 2n$ for the $(1, 3)$-flower. It is easy to see that the diameter of the $(1, v)$-flower, for v odd, is $L_n = (v-1)n + (3-v)/2$, and, in general one can show that $L_n \sim (v-1)n$.

For $u > 1$, however, the diameter grows as a power of n. For example, for the $(2, 2)$-flower we find $L_n = 2^n$, and, more generally, the diameter satisfies $L_n \sim u^n$. To summarize,

$$L_n \sim \begin{cases} (v-1)n & u = 1 \,, \\ u^n & u > 1 \,, \end{cases} \quad \text{flowers} \,. \tag{14}$$

Similar results are quite obvious for the case of (u, v)-trees, where

$$L_n \sim \begin{cases} vn & u = 1 \,, \\ u^n & u > 1 \,, \end{cases} \quad \text{trees} \,. \tag{15}$$

Since $N_n \sim (u+v)^n$ (Eq. (12)), we can recast these relations as

$$L \sim \begin{cases} \ln N & u = 1\,, \\ N^{\ln u/\ln(u+v)} & u > 1\,. \end{cases} \tag{16}$$

Thus, (u, v)-nets are *small world* only in the case of $u = 1$. For $u > 1$, the diameter increases as a power of N, just as in *finite*-dimensional objects, and the nets are in fact *fractal*.

For $u > 1$, the change of mass upon the rescaling of length by a factor b is

$$N(bL) = b^{d_\mathrm{B}} N(L)\,, \tag{17}$$

where d_B is the fractal dimension [8]. In this case, $N(uL) = (u+v)N(L)$, so

$$d_\mathrm{B} = \frac{\ln(u+v)}{\ln u}\,, \quad u > 1\,. \tag{18}$$

Transfinite Fractals Small world nets, such as $(1, v)$-nets, are *infinite*-dimensional. Indeed, their mass (N, or M) increases faster than any power (dimension) of their diameter. Also, note that a naive application of (4) to $u \to 1$ yields $d_\mathrm{f} \to \infty$. In the case of $(1, v)$-nets one can use their weak self-similarity to define a new measure of dimensionality, $\tilde{d}_\mathrm{f}$, characterizing how mass scales with diameter:

$$N(L+\ell) = \mathrm{e}^{\ell \tilde{d}_\mathrm{f}} N(L)\,. \tag{19}$$

Instead of a multiplicative rescaling of length, $L \mapsto bL$, a slower additive mapping, $L \mapsto L + \ell$, that reflects the small world property is considered. Because the exponent $\tilde{d}_\mathrm{f}$ usefully distinguishes between different graphs of infinite dimensionality, $\tilde{d}_\mathrm{f}$ has been termed the *transfinite* fractal dimension of the network. Accordingly, objects that are self-similar and have infinite dimension (but finite transfinite dimension), such as the $(1, v)$-nets, are termed transfinite fractals, or *transfractals*, for short.

For $(1, v)$-nets, we see that upon 'zooming in' one generation level the mass increases by a factor of $w = 1 + v$, while the diameter grows from L to $L + v - 1$ (for flowers), or to $L + v$ (trees). Hence their transfractal dimension is

$$\tilde{d}_\mathrm{f} = \begin{cases} \frac{\ln(1+v)}{v} & (1, v)\text{-trees}\,, \\ \frac{\ln(1+v)}{v-1} & (1, v)\text{-flowers}\,. \end{cases} \tag{20}$$

There is some arbitrariness in the selection of e as the base of the exponential in the definition (19). However the base is inconsequential for the sake of comparison between dimensionalities of different objects. Also, *scaling relations* between various transfinite exponents hold, irrespective of the choice of base: consider the scaling relation of Eq. (10) valid for fractal scale-free nets of degree exponent γ [62,63]. For example, in the fractal (u, v)-nets (with $u > 1$) renormalization reduces lengths by a factor $b = u$ and all degrees are reduced by a factor of 2, so $b^{d_\mathrm{k}} = 2$. Thus $d_\mathrm{k} = \ln 2/\ln u$, and since $d_\mathrm{B} = \ln(u+v)/\ln u$ and $\gamma = 1 + \ln(u+v)/\ln 2$, as discussed above, the relation (10) is indeed satisfied.

For transfractals, renormalization reduces distances by an *additive* length, ℓ, and we express the self-similarity manifest in the degree distribution as

$$P'(k) = \mathrm{e}^{\ell \tilde{d}_\mathrm{k}} P(\mathrm{e}^{-\ell \tilde{d}_\mathrm{k}} k)\,, \tag{21}$$

where $\tilde{d}_\mathrm{k}$ is the transfinite exponent analogous to d_k. Renormalization of the transfractal $(1, v)$-nets reduces the link lengths by $\ell = v - 1$ (for flowers), or $\ell = v$ (trees), while all degrees are halved. Thus,

$$\tilde{d}_\mathrm{k} = \begin{cases} \frac{\ln 2}{v} & (1, v)\text{-trees}\,, \\ \frac{\ln 2}{v-1} & (1, v)\text{-flowers}\,. \end{cases}$$

Along with (20), this result confirms that the scaling relation

$$\gamma = 1 + \frac{\tilde{d}_\mathrm{f}}{\tilde{d}_\mathrm{k}} \tag{22}$$

is valid also for transfractals, and regardless of the choice of base. A general proof of this relation is practically identical to the proof of (10) [62], merely replacing fractal with transfractal scaling throughout the argument.

For scale-free transfractals, following $m = L/\ell$ renormalizations the diameter and mass reduce to order one, and the scaling (19) implies $L \sim m\ell$, $N \sim \mathrm{e}^{m\ell\tilde{d}_\mathrm{f}}$, so that

$$L \sim \frac{1}{\tilde{d}_\mathrm{f}} \ln N\,,$$

in accordance with their small world property. At the same time the scaling (21) implies $K \sim \mathrm{e}^{m\ell\tilde{d}_\mathrm{k}}$, or $K \sim N^{\tilde{d}_\mathrm{k}/\tilde{d}_\mathrm{f}}$. Using the scaling relation (22), we rederive $K \sim N^{1/(\gamma-1)}$, which is indeed valid for scale-free nets *in general*, be they fractal or transfractal.

Properties of Fractal and Transfractal Networks

The existence of fractality in complex networks immediately calls for the question of what is the importance of such a structure in terms of network properties. In general, most of the relevant applications seem to be modified to a larger or lesser extent, so that fractal networks can be considered to form a separate network sub-class, sharing the main properties resulting from the wide distribution of regular scale-free networks, but at the same time bearing novel properties. Moreover, from a practical point of

view a fractal network can be usually more amenable to analytic treatment.

In this section we summarize some of the applications that seem to distinguish fractal from non-fractal networks.

Modularity

Modularity is a property closely related to fractality. Although this term does not have a unique well-defined definition we can claim that modularity refers to the existence of areas in the network where groups of nodes share some common characteristics, such as preferentially connecting within this area (the 'module') rather than to the rest of the network. The isolation of modules into distinct areas is a complicated task and in most cases there are many possible ways (and algorithms) to partition a network into modules.

Although networks with significant degree of modularity are not necessarily fractals, practically all fractal networks are highly modular in structure. Modularity naturally emerges from the effective 'repulsion' between hubs.

Since the hubs are not directly connected to each other, they usually dominate their neighborhood and can be considered as the 'center of mass' for a given module. The nodes surrounding hubs are usually assigned to this module.

The renormalization property of self-similar networks is very useful for estimating how modular a given network is, and especially for how this property is modified under varying scales of observation. We can use a simple definition for modularity M, based on the idea that the number of links connecting nodes within a module, L_i^{in}, is higher than the number of link connecting nodes in different modules, L_i^{out}.

For this purpose, the boxes that result from the box-covering method at a given length-scale ℓ_{B} are identified as the network modules for this scale. This partitioning assumes that the minimization of the number of boxes corresponds to an increase of modularity, taking advantage of the idea that all nodes within a box can reach each other within less than ℓ_{B} steps. This constraint tends to assign the largest possible number of nodes in a given neighborhood within the same box, resulting in an optimized modularity function.

A definition of the modularity function M that takes advantage of the special features of the renormalization process is, thus, the following [32]:

$$M(\ell_{\text{B}}) = \frac{1}{N_{\text{B}}} \sum_{i=1}^{N_{\text{B}}} \frac{L_i^{\text{in}}}{L_i^{\text{out}}} , \tag{23}$$

where the sum is over all the boxes.

The value of M through Eq. (23) for a given ℓ_{B} value is of small usefulness on its own, though. We can gather more information on the network structure if we measure M for different values of ℓ_{B}. If the dependence of M on ℓ_{B} has the form of a power-law, as if often the case in practice, then we can define the modularity exponent d_M through

$$M(\ell_{\text{B}}) \sim \ell_{\text{B}}^{d_M} . \tag{24}$$

The exponent d_M carries the important information of how modularity scales with the length, and separates modular from non-modular networks. The value of d_M is easy to compute in a d-dimensional lattice, since the number of links within any module scales with its bulk, as $L_i^{\text{in}} \sim \ell_{\text{B}}^d$ and the number of links outside the module scale with the length of its interface, i. e. $L_i^{\text{out}} \sim \ell_{\text{B}}^{d-1}$. So, the resulting scaling is $M \sim \ell_{\text{B}}$ i. e. $d_M = 1$. This is also the borderline value that separates non-modular structures ($d_M < 1$) from modular ones ($d_M > 1$).

For the Song–Havlin–Makse fractal model introduced in the previous section, a module can be identified as the neighborhood around a central hub. In the simplest version with $x = 1$, the network is a tree, with well-defined modules. Larger values of x mean that a larger number of links are connecting different modules, creating more loops and 'blurring' the discreteness of the modules, so that we can vary the degree of modularity in the network. For this model, it is also possible to analytically calculate the value of the exponent d_M.

During the growth process at step t, the diameter in the network model increases multiplicatively as $L(t+1) = 3L(t)$. The number of links within a module grows with $2m + x$ (each node on the side of one link gives rise to m new links and x extra links connect the new nodes), while the number of links pointing out of a module is by definition proportional to x. Thus, the modularity $M(\ell_{\text{B}})$ of a network is proportional to $(2m + x)/x$. Equation (24) can then be used to calculate d_M for the model:

$$\frac{2m + x}{x} \sim 3^{d_M} , \tag{25}$$

which finally yields

$$d_M = \frac{\ln\left(2\frac{m}{x} + 1\right)}{\ln 3} . \tag{26}$$

So, in this model the important quantity that determines the degree of modularity in the system is the ratio of the growth parameters m/x.

Most of the real-life networks that have been measured display some sort of modular character, i. e. $d_M > 1$, although many of them have values very close to 1. Only in

a few cases we have observed exponents $d_M < 1$. Most interesting is, though, the case of d_M values much larger than 1, where a large degree of modularity is observed and this trend is more pronounced for larger length-scales.

The importance of modularity as described above can be demonstrated in biological networks. There, it has been suggested that the boxes may correspond to functional modules and in protein interaction networks, for example, there may be an evolution drive of the system behind the development of its modular structure.

Robustness

Shortly after the discovery of the scale-free property, the first important application of their structure was perhaps their extreme resilience to removal of random nodes [3,10,19,21,57,60,61]. At the same time such a network was found to be quite vulnerable to an intentional attack, where nodes are removed according to decreasing order of their degree [20,30]. The resilience of a network is usually quantified through the size of the largest remaining connected cluster $S_{\max}(p)$, when a fraction p of the nodes has been removed according to a given strategy. At a critical point p_c where this size becomes equal to $S_{\max}(p_c) \simeq 0$, we consider that the network has been completely disintegrated. For the random removal case, this threshold is $p_c \simeq 1$, i. e. practically all nodes need to be destroyed. In striking contrast, for intentional attacks p_c is in general of the order of only a few percent, although the exact value depends on the system details.

Fractality in networks considerably strengthens the robustness against intentional attacks, compared to non-fractal networks with the same degree exponent γ. In Fig. 9 the comparison between two such networks clearly shows that the critical fraction p_c increases almost 4 times from $p_c \simeq 0.02$ (non-fractal topology) to $p_c \simeq 0.09$ (fractal topology). These networks have the same γ exponent, the same number of links, number of nodes, number of loops and the same clustering coefficient, differing only in whether hubs are directly connected to each other. The fractal property, thus, provides a way of increasing resistance against the network collapse, in the case of a targeted attack.

The main reason behind this behavior is the dispersion of hubs in the network. A hub is usually a central node that helps other nodes to connect to the main body of the system. When the hubs are directly connected to each other, this central core is easy to destroy in a targeted attack leading to a rapid collapse of the network. On the contrary, isolating the hubs into different areas helps the network to retain connectivity for longer time, since destroying the hubs

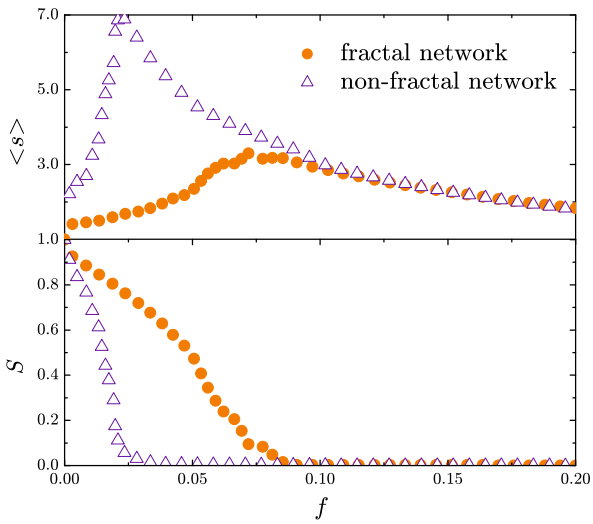

Fractal and Transfractal Scale-Free Networks, Figure 9
Vulnerability under intentional attack of a non-fractal Song–Makse–Havlin network (for $e = 0$) and a fractal Song–Makse–Havlin network (for $e = 1$). The plot shows the relative size of the largest cluster, S, and the average size of the remaining isolated clusters, $\langle s \rangle$ as a function of the removal fraction f of the largest hubs for both networks

now is not similarly catastrophic, with most of the nodes finding alternative paths through other connections.

The advantage of increased robustness derived from the combination of modular and fractal network character, may provide valuable hints on why most biological networks have evolved towards a fractal architecture (better chance of survival against lethal attacks).

Degree Correlations

We have already mentioned the importance of hub-hub correlations or anti-correlations in fractality. Generalizing this idea to nodes of any degree, we can ask what is the joint degree probability $P(k_1, k_2)$ that a randomly chosen link connects two nodes with degree k_1 and k_2, respectively. Obviously, this is a meaningful question only for networks with a wide degree distribution, otherwise the answer is more or less trivial with all nodes having similar degrees. A similar and perhaps more useful quantity is the conditional degree probability $P(k_1|k_2)$, defined as the probability that a random link from a node having degree k_2 points to a node with degree k_1. In general, the following balance condition is satisfied

$$k_2 P(k_1|k_2) P(k_2) = k_1 P(k_2|k_1) P(k_1) \,. \tag{27}$$

It is quite straightforward to calculate $P(k_1|k_2)$ for completely uncorrelated networks. In this case, $P(k_1|k_2)$ does

not depend on k_2, and the probability to chose a node with degree k_1 becomes simply $P(k_1|k_2) = k_1 P(k_1)/\langle k_1 \rangle$. In the case where degree-degree correlations are present, though, the calculation of this function is very difficult, even when restricting ourselves to a direct numerical evaluation, due to the emergence of huge fluctuations.

We can still estimate this function, though, using again the self-similarity principle. If we consider that the function $P(k_1, k_2)$ remains invariant under the network renormalization scheme described above, then it is possible to show that [33]

$$P(k_1, k_2) \sim k_1^{-(\gamma-1)} k_2^{-\epsilon} \quad (k_1 > k_2) \,, \tag{28}$$

and similarly

$$P(k_1|k_2) \sim k_1^{-(\gamma-1)} k_2^{-(\epsilon-\gamma+1)} \quad (k_1 > k_2) \,, \tag{29}$$

In the above equations we have also introduced the correlation exponent ϵ, which characterizes the degree of correlations in a network. For example, the case of uncorrelated networks is described by the value $\epsilon = \gamma - 1$.

The exponent ϵ can be measured quite accurately using an appropriate quantity. For this purpose, we can introduce a measure such as

$$E_b(k) \equiv \frac{\int_{bk}^{\infty} P(k|k_2) \mathrm{d}k_2}{\int_{bk}^{\infty} P(k) \mathrm{d}k} \,, \tag{30}$$

which estimates the probability that a node with degree k has neighbors with degree larger than bk, and b is an arbitrary parameter that has been shown not to influence the results. It is easy to show that

$$E_b(k) \sim \frac{k^{1-\epsilon}}{k^{1-\gamma}} = k^{-(\epsilon-\gamma)} \,. \tag{31}$$

This relation allows us to estimate ϵ for a given network, after calculating the quantity $E_b(k)$ as a function of k.

The above discussion can be equally applied to both fractal and non-fractal networks. If we restrict ourselves to fractal networks, then we can develop our theory a bit further. If we consider the probability $\mathcal{E}(\ell_B)$ that the largest degree node in each box is connected directly with the other largest degree nodes in other boxes (after optimally covering the network), then this quantity scales as a power law with ℓ_B:

$$\mathcal{E}(\ell_B) \sim \ell_B^{-d_e} \,, \tag{32}$$

where d_e is a new exponent describing the probability of hub-hub connection [63]. The exponent ϵ, which describes correlations over any degree, is related to d_e, which refers to correlations between hubs only. The resulting relation is

$$\epsilon = 2 + \frac{d_e}{d_k} = 2 + (\gamma - 1)\frac{d_e}{d_B} \,. \tag{33}$$

For an infinite fractal dimension $d_B \to \infty$, which is the onset of non-fractal networks that cannot be described by the above arguments, we have the limiting case of $\epsilon = 2$. This value separates fractal from non-fractal networks, so that fractality is indicated by $\epsilon > 2$. Also, we have seen that the line $\epsilon = \gamma - 1$ describes networks for which correlations are minimal. Measurements of many real-life networks have verified the above statements, where networks with $\epsilon > 2$ having been clearly characterized as fractals with alternate methods. All non-fractal networks have values of $\epsilon < 2$ and the distance from the $\epsilon = \gamma - 1$ line determines how much stronger or weaker the correlations are, compared to the uncorrelated case.

In short, using the self-similarity principle makes it possible to gain a lot of insight on network correlations, a notoriously difficult task otherwise. Furthermore, the study of correlations can be reduced to the calculation of a single exponent ϵ, which is though capable of delivering a wealth of information on the network topological properties.

Diffusion and Resistance

Scale-free networks have been described as objects of infinite dimensionality. For a regular structure this statement would suggest that one can simply use the known diffusion laws for $d = \infty$. Diffusion on scale-free structures, however, is much harder to study, mainly due to the lack of translational symmetry in the system and different local environments. Although exact results are still not available, the scaling theory on fractal networks provides the tools to better understand processes, such as diffusion and electric resistance.

In the following, we describe diffusion through the average first-passage time T_{AB}, which is the average time for a diffusing particle to travel from node A to node B. At the same time, assuming that each link in the network has an electrical resistance of 1 unit, we can describe the electrical properties through the resistance between the two nodes A and B, R_{AB}.

The connection between diffusion (first-passage time) and electric networks has long been established in homogeneous systems. This connection is usually expressed through the Einstein relation [8]. The Einstein relation is of great importance because it connects a *static quantity* R_{AB} with a *dynamic quantity* T_{AB}. In other words, the be-

havior of a diffusing particle can be inferred by simply having knowledge of a static topological property of the network.

In any renormalizable network the scaling of T and R follows the form:

$$\frac{T'}{T} = \ell_{\mathrm{B}}^{-d_w} \,, \quad \frac{R'}{R} = \ell_{\mathrm{B}}^{-\zeta} \,, \tag{34}$$

where $T'(R')$ and $T(R)$ are the first-passage time (resistance) for the renormalized and original networks, respectively. The dynamical exponents d_w and ζ characterize the scaling in any lattice or network that remains invariant under renormalization. The Einstein relation relates these two exponents through the dimensionality of the substrate d_{B}, according to:

$$d_w = \zeta + d_{\mathrm{B}} \,. \tag{35}$$

The validity of this relation in inhomogeneous complex networks, however, is not yet clear. Still, in fractal and transfractal networks there are many cases where this relation has been proved to be valid, hinting towards a wider applicability. For example, in [13,56] it has been shown that the Einstein Relation [8] in (u, v)-*flowers* and (u, v)-trees is valid for any u and v, that is for both fractal and transfractal networks. In general, in terms of the scaling theory we can study diffusion and resistance (or conductance) in a similar manner [32].

Because of the highly inhomogeneous character of the structure, though, we are interested in how these quantities behave as a function of the end-node degrees k_1 and k_2 when they are separated by a given distance ℓ. Thus, we are looking for the full dependence of $T(\ell; k_1, k_2)$ and $R(\ell; k_1, k_2)$. Obviously, for lattices or networks with narrow degree distribution there is no degree dependence and those results should be a function of ℓ only.

For self-similar networks, we can rewrite Eq. (34) above as

$$\frac{T'}{T} = \left(\frac{N'}{N}\right)^{d_w/d_{\mathrm{B}}} \,, \quad \frac{R'}{R} = \left(\frac{N'}{N}\right)^{\zeta/d_{\mathrm{B}}} \,, \tag{36}$$

where we have taken into account Eq. (3). This approach offers the practical advantage that the variation of N'/N is larger than the variation of ℓ_{B}, so that the exponents calculation can be more accurate. To calculate these exponents, we fix the box size ℓ_{B} and we measure the diffusion time T and resistance R between any two points in a network before and after renormalization. If for every such pair we plot the corresponding times and resistances in T' vs. T and R' vs. R plots, as shown in Fig. 10, then all these points fall in a narrow area, suggesting a constant value for the ratio T'/T over the entire network. Repeating this procedure

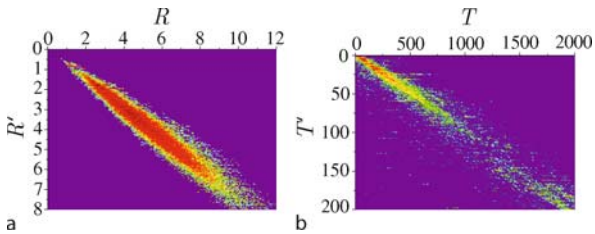

Fractal and Transfractal Scale-Free Networks, Figure 10
Typical behavior of the probability distributions for the resistance R vs. R' and the diffusion time T vs. T', respectively, for a given ℓ_{B} value. Similar plots for other ℓ_{B} values verify that the ratios of these quantities during a renormalization stage are roughly constant for all pairs of nodes in a given biological network

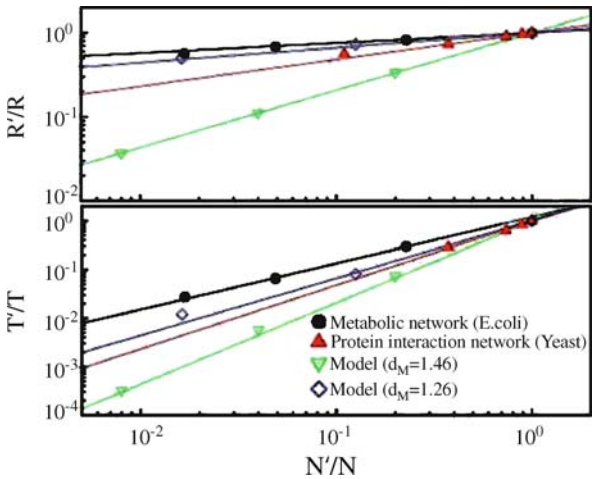

Fractal and Transfractal Scale-Free Networks, Figure 11
Average value of the ratio of resistances R/R' and diffusion times T/T', as measured for different ℓ_{B} values (each point corresponds to a different value of ℓ_{B}). Results are presented for both biological networks, and two fractal network models with different d_M values. The slopes of the curves correspond to the exponents ζ/d_{B} (*top panel*) and d_w/d_{B} (*bottom panel*)

for different ℓ_{B} values yields other ratio values. The plot of these ratios vs. N'/N (Fig. 11) finally exhibits a power-law dependence, verifying Eq. (36). We can then easily calculate the exponents d_w and ζ from the slopes in the plot, since the d_{B} exponent is already known through the standard box-covering methods. It has been shown that the results for many different networks are consistent, within statistical error, with the Einstein relation [32,56].

The dependence on the degrees k_1, k_2 and the distance ℓ can also be calculated in a scaling form using the self-similarity properties of fractal networks. After renormalization, a node with degree k in a given network, will have a degree $k' = \ell_{\mathrm{B}}^{-d_k} k$ according to Eq. (9). At the

same time all distances ℓ are scaled down according to $\ell' = \ell/\ell_B$. This means that Eqs. (36) can be written as

$$R'(\ell'; k_1', k_2') = \ell_B^{-\zeta} R(\ell; k_1, k_2) \tag{37}$$

$$T'(\ell'; k_1', k_2') = \ell_B^{-d_w} T(\ell; k_1, k_2) \,. \tag{38}$$

Substituting the renormalized quantities we get:

$$R'(\ell_B^{-1}\ell; \ell_B^{-d_k} k_1, \ell_B^{-d_k} k_2) = \ell_B^{-\zeta} R(\ell; k_1, k_2) \,. \tag{39}$$

The above equation holds for all values of ℓ_B, so we can select this quantity to be $\ell_B = k_2^{1/d_k}$. This constraint allows us to reduce the number of variables in the equation, with the final result:

$$R\left(\frac{\ell}{k_2^{\frac{1}{d_k}}}; \frac{k_1}{k_2}, 1\right) = k_2^{\frac{-\zeta}{d_k}} R(\ell_B; k_1, k_2) \,. \tag{40}$$

This equation suggests a scaling for the resistance R:

$$R(\ell; k_1, k_2) = k_2^{\frac{\zeta}{d_k}} f_R\left(\frac{\ell}{k_2^{\frac{1}{d_k}}}, \frac{k_1}{k_2}\right) , \tag{41}$$

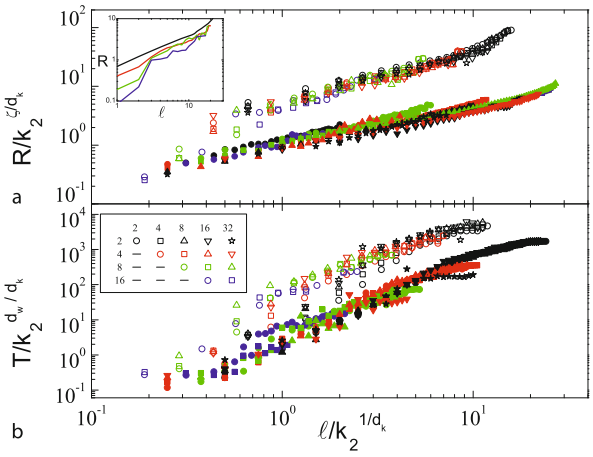

Fractal and Transfractal Scale-Free Networks, Figure 12
Rescaling of a the resistance and b the diffusion time according to Eqs. (41) and (42) for the protein interaction network of yeast (upper symbols) and the Song–Havlin–Makse model for $e = 1$ (lower filled symbols). The data for PIN have been vertically shifted upwards by one decade for clarity. Each symbol corresponds to a fixed ratio k_1/k_2 and the different colors denote a different value for k_1. *Inset:* Resistance R as a function of distance ℓ, before rescaling, for constant ratio $k_1/k_2 = 1$ and different k_1 values

where $f_R()$ is an undetermined function. All the above arguments can be repeated for the diffusion time, with a similar expression:

$$T(\ell; k_1, k_2) = k_2^{\frac{d_w}{d_k}} f_T\left(\frac{\ell}{k_2^{\frac{1}{d_k}}}, \frac{k_1}{k_2}\right) , \tag{42}$$

where the form of the right-hand function may be different. The final result for the scaling form is Eqs. (41) and (42), which is also supported by the numerical data collapse in Fig. 12. Notice that in the case of homogeneous networks, where there is almost no k-dependence, the unknown functions in the rhs reduce to the forms $f_R(x, 1) = x^\zeta$, $f_T(x, 1) = x^{d_w}$, leading to the well-established classical relations $R \sim \ell^\zeta$ and $T \sim \ell^{d_w}$.

Future Directions

Fractal networks combine features met in fractal geometry and in network theory. As such, they present many unique aspects. Many of their properties have been well-studied and understood, but there is still a great amount of open and unexplored questions remaining to be studied.

Concerning the structural aspects of fractal networks, we have described that in most networks the degree distribution $P(k)$, the joint degree distribution $P(k_1, k_2)$ and a number of other quantities remain invariant under renormalization. Are there any quantities that are not invariable, and what would their importance be?

Of central importance is the relation of topological features with functionality. The optimal network covering leads to the partitioning of the network into boxes. Do these boxes carry a message other than nodes proximity? For example, the boxes could be used as an alternative definition for separated communities, and fractal methods could be used as a novel method for community detection in networks [4,5,18,51,53].

The networks that we have presented are all static, with no temporal component, and time evolution has been ignored in all our discussions above. Clearly, biological networks, the WWW, and other networks have grown (and continue to grow) from some earlier simpler state to their present fractal form. Has fractality always been there or has it emerged as an intermediate stage obeying certain evolutionary drive forces? Is fractality a stable condition or growing networks will eventually fall into a non-fractal form?

Finally, we want to know what is the inherent reason behind fractality. Of course, we have already described how hub-hub anti-correlations can give rise to fractal networks. However, can this be directly related to some un-

derlying mechanism, so that we gain some information on the process? In general, in Biology we already have some idea on the advantages of adopting a fractal structure. Still, the question remains: why fractality exists in certain networks and not in others? Why both fractal and non-fractal networks are needed? It seems that we will be able to increase our knowledge for the network evolutionary mechanisms through fractality studies.

In conclusion, a deeper understanding of the self-similarity, fractality and transfractality of complex networks will help us analyze and better understand many fundamental properties of real-world networks.

Acknowledgments

We acknowledge support from the National Science Foundation.

Appendix: The Box Covering Algorithms

The estimation of the fractal dimension and the self-similar features in networks have become standard properties in the study of real-world systems. For this reason, in the last three years many box covering algorithms have been proposed [64,69]. This section presents four of the main algorithms, along with a brief discussion on the advantages and disadvantages that they offer.

Recalling the original definition of box covering by Hausdorff [14,29,55], for a given network G and box size ℓ_B, a box is a set of nodes where all distances ℓ_{ij} between any two nodes i and j in the box are smaller than ℓ_B. The minimum number of boxes required to cover the entire network G is denoted by N_B. For $\ell_B = 1$, each box encloses only 1 node and therefore, N_B is equal to the size of the network N. On the other hand, $N_B = 1$ for $\ell_B \geq \ell_B^{max}$, where ℓ_B^{max} is the diameter of the network plus one.

The ultimate goal of a box-covering algorithm is to find the *minimum* number of boxes $N_B(\ell_B)$ for any ℓ_B. It has been shown that this problem belongs to the family of NP-hard problems [34], which means that the solution cannot be achieved in polynomial time. In other words, for a relatively large network size, there is no algorithm that can provide an exact solution in a reasonably short amount of time. This limitation requires treating the box covering problem with approximations, using for example optimization algorithms.

The Greedy Coloring Algorithm

The box-covering problem can be mapped into another NP-hard problem [34]: the graph coloring problem.

An algorithm that approximates well the optimal solution of this problem was presented in [64]. For an arbitrary value of ℓ_B, first construct a dual network G', in which two nodes are connected if the distance between them in G (the original network) is greater or equal than ℓ_B. Figure 13 shows an example of a network G which yields such a dual network G' for $\ell_B = 3$ (upper row of Fig. 13).

Vertex coloring is a well-known procedure, where labels (or colors) are assigned to each vertex of a network, so that no edge connects two identically colored vertices. It is clear that such a coloring in G' gives rise to a natural box covering in the original network G, in the sense that vertices of the same color will necessarily form a box since the distance between them must be less than ℓ_B. Accordingly, the minimum number of boxes $N_B(G)$ is equal to the minimum required number of colors (or the chromatic number) in the dual network G', $\chi(G')$.

In simpler terms, (a) if the distance between two nodes in G is greater than ℓ_B these two neighbors cannot belong in the same box. According to the construction of G', these two nodes will be connected in G' and thus they cannot have the same color. Since they have a different color they will not belong in the same box in G. (b) On the contrary, if the distance between two nodes in G is less than ℓ_B it is possible that these nodes belong in the same box. In G' these two nodes will not be connected and it is allowed

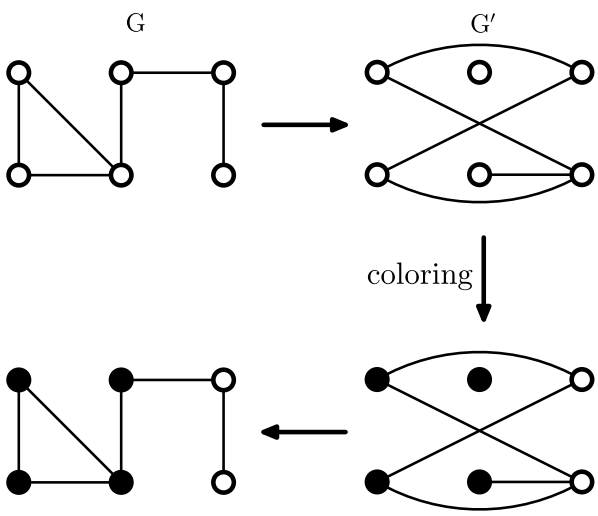

Fractal and Transfractal Scale-Free Networks, Figure 13
Illustration of the solution for the network covering problem via mapping to the graph coloring problem. Starting from *G* (*upper left panel*) we construct the dual network *G′* (*upper right panel*) for a given box size (here $\ell_B = 3$), where two nodes are connected if they are at a distance $\ell \geq \ell_B$. We use a greedy algorithm for vertex coloring in *G′*, which is then used to determine the box covering in *G*, as shown in the plot

for these two nodes to carry the same color, i. e. they may belong to the same box in G, (whether these nodes will actually be connected depends on the exact implementation of the coloring algorithm).

The algorithm that follows both constructs the dual network G' and assigns the proper node colors for all ℓ_B values in one go. For this implementation a two-dimensional matrix $c_{i\ell}$ of size $N \times \ell_B^{max}$ is needed, whose values represent the color of node i for a given box size $\ell = \ell_B$.

1. Assign a unique id from 1 to N to all network nodes, without assigning any colors yet.
2. For all ℓ_B values, assign a color value 0 to the node with id=1, i. e. $c_{1\ell} = 0$.
3. Set the id value $i = 2$. Repeat the following until $i = N$.
   (a) Calculate the distance ℓ_{ij} from i to all the nodes in the network with id j less than i.
   (b) Set $\ell_B = 1$
   (c) Select one of the unused colors $c_{j\ell_{ij}}$ from all nodes $j < i$ for which $\ell_{ij} \geq \ell_B$. This is the color $c_{i\ell_B}$ of node i for the given ℓ_B value.
   (d) Increase ℓ_B by one and repeat (c) until $\ell_B = \ell_B^{max}$.
   (e) Increase i by 1.

The results of the greedy algorithm may depend on the original coloring sequence. The quality of this algorithm was investigated by randomly reshuffling the coloring sequence and applying the greedy algorithm several times and in different models [64]. The result was that the probability distribution of the number of boxes N_B (for all box sizes ℓ_B) is a narrow Gaussian distribution, which indicates that almost any implementation of the algorithm yields a solution close to the optimal.

Strictly speaking, the calculation of the fractal dimension d_B through the relation $N_B \sim \ell_B^{-d_B}$ is valid only for the minimum possible value of N_B, for any given ℓ_B value, so any box covering algorithm must aim to find this minimum N_B. Although there is no rule to determine when this minimum value has been actually reached (since this would require an exact solution of the NP-hard coloring problem) it has been shown [23] that the greedy coloring algorithm can, in many cases, identify a coloring sequence which yields the optimal solution.

Burning Algorithms

This section presents three box covering algorithms based on more traditional breadth-first search algorithm.

A box is defined as *compact* when it includes the maximum possible number of nodes, i. e. when there do not exist any other network nodes that could be included in this box. A *connected* box means that any node in the box can be reached from any other node in this box, without having to leave this box. Equivalently, a *disconnected* box denotes a box where certain nodes can be reached by other nodes in the box only by visiting nodes outside this box. For a demonstration of these definitions see Fig. 14.

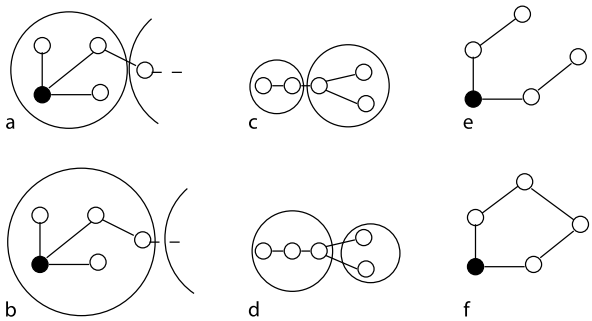

Fractal and Transfractal Scale-Free Networks, Figure 14
Our definitions for a box that is a non-compact for $\ell_B = 3$, i. e. could include more nodes, b compact, c connected, and d disconnected (the nodes in the *right box* are not connected in the box). e For this box, the values $\ell_B = 5$ and $r_B = 2$ verify the relation $\ell_B = 2r_B + 1$. f One of the pathological cases where this relation is not valid, since $\ell_B = 3$ and $r_B = 2$

Burning with the Diameter ℓ_B, and the Compact-Box-Burning (CBB) Algorithm

The basic idea of the CBB algorithm for the generation of a box is to start from a given box center and then expand the box so that it includes the maximum possible number of nodes, satisfying at the same time the maximum distance between nodes in the box ℓ_B. The CBB algorithm is as follows (see Fig. 15):

1. Initially, mark all nodes as uncovered.
2. Construct the set C of all yet uncovered nodes.

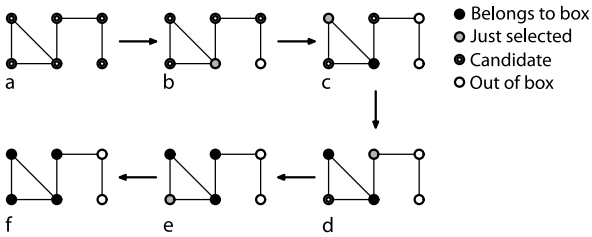

Fractal and Transfractal Scale-Free Networks, Figure 15
Illustration of the CBB algorithm for $\ell_B = 3$. a Initially, all nodes are candidates for the box. b A random node is chosen, and nodes at a distance further than ℓ_B from this node are no longer candidates. c The node chosen in b becomes part of the box and another candidate node is chosen. The above process is then repeated until the box is complete

3. Choose a random node p from the set of uncovered nodes C and remove it from C.
4. Remove from C all nodes i whose distance from p is $\ell_{pi} \geq \ell_B$, since by definition they will not belong in the same box.
5. Repeat steps (3) and (4) until the candidate set is empty.
6. Repeat from step (2) until all the network has been covered.

Random Box Burning

In 2006, J. S. Kim et al. presented a simple algorithm for the calculation of fractal dimension in networks [42,43,44]:

1. Pick a randomly chosen node in the network as a seed of the box.
2. Search using breath-first search algorithm until distance l_B from the seed. Assign all newly burned nodes to the new box. If no new node is found, discard and start from (1) again.
3. Repeat (1) and (2) until all nodes have a box assigned.

This Random Box Burning algorithm has the advantage of being a fast and simple method. However, at the same time there is no inherent optimization employed during the network coverage. Thus, this simple Monte-Carlo method is almost certain that will yield a solution far from the optimal and one needs to implement many different realizations and only retain the smallest number of boxes found out of all these realizations.

Burning with the Radius r_B, and the Maximum-Excluded-Mass-Burning (MEMB) Algorithm

A box of size ℓ_B includes nodes where the distance between any pair of nodes is less than ℓ_B. It is possible, though, to grow a box from a given central node, so that all nodes in the box are within distance less than a given box radius r_B (the maximum distance from a central node). This way, one can still recover the same fractal properties of a network. For the original definition of the box, ℓ_B corresponds to the box diameter (maximum distance between any two nodes in the box) plus one. Thus, ℓ_B and r_B are connected through the simple relation $\ell_B = 2r_B + 1$. In general this relation is exact for loopless configurations, but in general there may exist cases where this equation is not exact (Fig. 14).

The MEMB algorithm always yields the optimal solution for non scale-free homogeneous networks, since the choice of the central node is not important. However, in inhomogeneous networks with wide-tailed degree distribution, such as scale-free networks, this algorithm fails to achieve an optimal solution because of the presence of hubs.

The MEMB, as a difference from the Random Box Burning and the CBB, attempts to locate some *optimal central* nodes which act as the burning origins for the boxes. It contains as a special case the choice of the hubs as centers of the boxes, but it also allows for low-degree nodes to be burning centers, which sometimes is convenient for finding a solution closer to the optimal.

In the following algorithm we use the basic idea of box optimization, in which each box covers the maximum possible number of nodes. For a given burning radius r_B, we define the *excluded mass* of a node as the number of uncovered nodes within a chemical distance less than r_B. First, calculate the excluded mass for all the uncovered nodes. Then, seek to cover the network with boxes of maximum excluded mass. The details of this algorithm are as follows (see Fig. 17):

1. Initially, all nodes are marked as uncovered and non-centers.
2. For all non-center nodes (including the already covered nodes) calculate the excluded mass, and select the node p with the maximum excluded mass as the next center.
3. Mark all the nodes with chemical distance less than r_B from p as covered.
4. Repeat steps (2) and (3) until all nodes are either covered or centers.

Notice that the excluded mass has to be updated in each step because it is possible that it has been modified during this step. A box center can also be an already covered node, since it may lead to a larger box mass. After the above procedure, the number of selected centers coincides with the number of boxes N_B that completely cover the network.

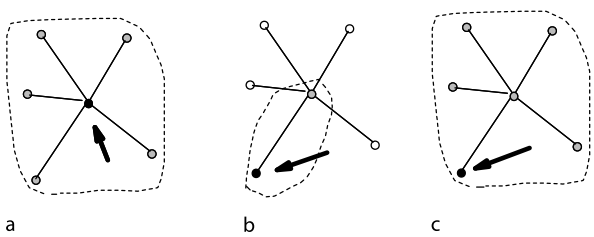

Fractal and Transfractal Scale-Free Networks, Figure 16
Burning with the radius r_B from a a hub node or b a non-hub node results in very different network coverage. In a we need just one box of $r_B = 1$ while in b 5 boxes are needed to cover the same network. This is an intrinsic problem when burning with the radius. c Burning with the maximum distance ℓ_B (in this case $\ell_B = 2r_B + 1 = 3$) we avoid this situation, since independently of the starting point we would still obtain $N_B = 1$

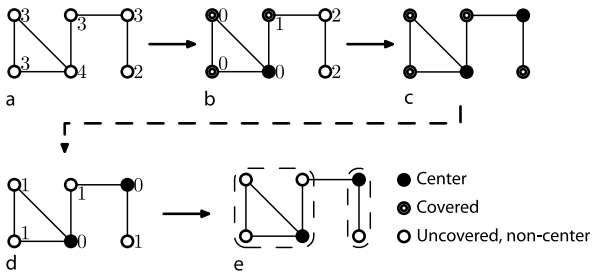

Fractal and Transfractal Scale-Free Networks, Figure 17
Illustration of the MEMB algorithm for $r_B = 1$. *Upper row*: Calculation of the box centers a We calculate the excluded mass for each node. b The node with maximum mass becomes a center and the excluded masses are recalculated. c A new center is chosen. Now, the entire network is covered with these two centers. *Bottom row*: Calculation of the boxes d Each box includes initially only the center. Starting from the centers we calculate the distance of each network node to the closest center. e We assign each node to its nearest box

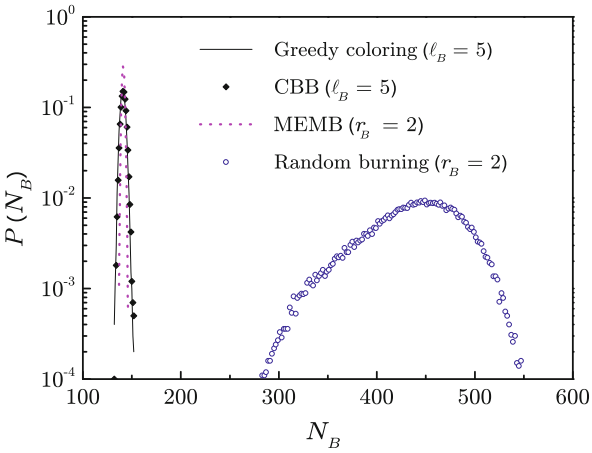

Fractal and Transfractal Scale-Free Networks, Figure 18
Comparison of the distribution of N_B for 10^4 realizations of the four network covering methods presented in this paper. Notice that three of these methods yield very similar results with narrow distributions and comparable minimum values, while the random burning algorithm fails to reach a value close to this minimum (and yields a broad distribution)

However, the non-center nodes have not yet been assigned to a given box. This is performed in the next step:

1. Give a unique box id to every center node.
2. For all nodes calculate the "central distance", which is the chemical distance to its nearest center. The central distance has to be less than r_B, and the center identification algorithm above guarantees that there will always exist such a center. Obviously, all center nodes have a central distance equal to 0.
3. Sort the non-center nodes in a list according to increasing central distance.
4. For each non-center node i, at least one of its neighbors has a central distance less than its own. Assign to i the same id with this neighbor. If there exist several such neighbors, randomly select an id from these neighbors. Remove i from the list.
5. Repeat step (4) according to the sequence from the list in step (3) for all non-center nodes.

Comparison Between Algorithms

The choice of the algorithm to be used for a problem depends on the details of the problem itself. If connected boxes are a requirement, MEMB is the most appropriate algorithm; but if one is only interested in obtaining the fractal dimension of a network, the greedy-coloring or the random box burning are more suitable since they are the fastest algorithms.

As explained previously, any algorithm should intend to find the optimal solution, that is, find the minimum number of boxes that cover the network. Figure 18 shows the performance of each algorithm. The greedy-coloring, the CBB and MEMB algorithms exhibit a narrow distribution of the number of boxes, showing evidence that they cover the network with a number of boxes that is close to the optimal solution. Instead, the Random Box Burning returns a wider distribution and its average is far above the average of the other algorithms. Because of the great ease and speed with which this technique can be implemented, it would be useful to show that the average number of covering boxes is overestimated by a fixed proportionality constant. In that case, despite the error, the predicted number of boxes would still yield the correct scaling and fractal dimension.

Bibliography

1. Albert R, Barabási A-L (2002) Rev Mod Phys 74:47; Barabási A-L (2003) Linked: how everything is connected to everything else and what it means. Plume, New York; Newman MEJ (2003) SIAM Rev 45:167; Dorogovtsev SN, Mendes JFF (2002) Adv Phys 51:1079; Dorogovtsev SN, Mendes JFF (2003) Evolution of networks: from biological nets to the internet and WWW. Oxford University Press, Oxford; Bornholdt S, Schuster HG (2003) Handbook of graphs and networks. Wiley-VCH, Berlin; Pastor-Satorras R, Vespignani A (2004) Evolution and structure of the internet. Cambridge University Press, Cambridge; Amaral LAN, Ottino JM (2004) Complex networks – augmenting the framework for the study of complex systems. Eur Phys J B 38: 147–162
2. Albert R, Jeong H, Barabási A-L (1999) Diameter of the world wide web. Nature 401:130–131

3. Albert R, Jeong H, Barabási AL (2000) Nature 406:p378
4. Bagrow JP, Bollt EM (2005) Phys Rev E 72:046108
5. Bagrow JP (2008) Stat Mech P05001
6. Bagrow JP, Bollt EM, Skufca JD (2008) Europhys Lett 81:68004
7. Barabási A-L, Albert R (1999) Sience 286:509
8. ben-Avraham D, Havlin S (2000) Diffusion and reactions in fractals and disordered systems. Cambridge University Press, Cambridge
9. Berker AN, Ostlund S (1979) J Phys C 12:4961
10. Beygelzimer A, Grinstein G, Linsker R, Rish I (2005) Physica A Stat Mech Appl 357:593–612
11. Binney JJ, Dowrick NJ, Fisher AJ, Newman MEJ (1992) The theory of critical phenomena: an introduction to the renormalization group. Oxford University Press, Oxford
12. Bollobás B (1985) Random graphs. Academic Press, London
13. Bollt E, ben-Avraham D (2005) New J Phys 7:26
14. Bunde A, Havlin S (1996) Percolation I and Percolation II. In: Bunde A, Havlin S (eds) Fractals and disordered systems, 2nd edn. Springer, Heidelberg
15. Burch H, Chewick W (1999) Mapping the internet. IEEE Comput 32:97–98
16. Butler D (2006) Nature 444:528
17. Cardy J (1996) Scaling and renormalization in statistical physics. Cambridge University Press, Cambridge
18. Clauset A, Newman MEJ, Moore C (2004) Phys Rev E 70:066111
19. Cohen R, Erez K, ben-Avraham D, Havlin S (2000) Phys Rev Lett 85:4626
20. Cohen R, Erez K, ben-Avraham D, Havlin S (2001) Phys Rev Lett 86:3682
21. Cohen R, ben-Avraham D, Havlin S (2002) Phys Rev E 66:036113
22. Comellas F Complex networks: deterministic models physics and theoretical computer science. In: Gazeau J-P, Nesetril J, Rovan B (eds) From Numbers and Languages to (Quantum) Cryptography. 7 NATO Security through Science Series: Information and Communication Security. IOS Press, Amsterdam. pp 275–293. 348 pags. ISBN 1-58603-706-4
23. Cormen TH, Leiserson CE, Rivest RL, Stein C (2001) Introduction to algorithms. MIT Press, Cambridge
24. Data from SCAN project. The Mbone. http://www.isi.edu/scan/scan.html Accessed 2000
25. Database of Interacting Proteins (DIP) http://dip.doe-mbi.ucla.edu Accessed 2008
26. Dorogovtsev SN, Goltsev AV, Mendes JFF (2002) Phys Rev E 65:066122
27. Erdős P, Rényi A (1960) On the evolution of random graphs. Publ Math Inst Hung Acad Sci 5:17–61
28. Faloutsos M, Faloutsos P, Faloutsos C (1999) Comput Commun Rev 29:251–262
29. Feder J (1988) Fractals. Plenum Press, New York
30. Gallos LK, Argyrakis P, Bunde A, Cohen R, Havlin S (2004) Physica A 344:504–509
31. Gallos LK, Cohen R, Argyrakis P, Bunde A, Havlin S (2005) Phys Rev Lett 94:188701
32. Gallos LK, Song C, Havlin S, Makse HA (2007) PNAS 104:7746
33. Gallos LK, Song C, Makse HA (2008) Phys Rev Lett 100:248701
34. Garey M, Johnson D (1979) Computers and intractability: a guide to the theory of NP-completeness. W.H. Freeman, New York
35. Goh K-I, Salvi G, Kahng B, Kim D (2006) Phys Rev Lett 96:018701
36. Han J-DJ et al (2004) Nature 430:88–93
37. Hinczewski M, Berker AN (2006) Phys Rev E 73:066126
38. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási A-L (2000) Nature 407:651–654
39. Kadanoff LP (2000) Statistical physics: statics, dynamics and renormalization. World Scientific Publishing Company, Singapore
40. Kaufman M, Griffiths RB (1981) Phys Rev B 24:496(R)
41. Kaufman M, Griffiths RB (1984) Phys Rev B 24:244
42. Kim JS, Goh K-I, Salvi G, Oh E, Kahng B, Kim D (2007) Phys Rev E 75:016110
43. Kim JS, Goh K-I, Kahng B, Kim D (2007) Chaos 17:026116
44. Kim JS, Goh K-I, Kahng B, Kim D (2007) New J Phys 9:177
45. Mandelbrot B (1982) The fractal geometry of nature. W.H. Freeman and Company, New York
46. Maslov S, Sneppen K (2002) Science 296:910–913
47. Milgram S (1967) Psychol Today 2:60
48. Motter AE, de Moura APS, Lai Y-C, Dasgupta P (2002) Phys Rev E 65:065102
49. Newman MEJ (2002) Phys Rev Lett 89:208701
50. Newman MEJ (2003) Phys Rev E 67:026126
51. Newman MEJ, Girvan M (2004) Phys Rev E 69:026113
52. Overbeek R et al (2000) Nucl Acid Res 28:123–125
53. Palla G, Barabási A-L, Vicsek T (2007) Nature 446:664–667
54. Pastor-Satorras R, Vázquez A, Vespignani A (2001) Phys Rev Lett 87:258701
55. Peitgen HO, Jurgens H, Saupe D (1993) Chaos and fractals: new frontiers of science. Springer, New York
56. Rozenfeld H, Havlin S, ben-Avraham D (2007) New J Phys 9:175
57. Rozenfeld H, ben-Avraham D (2007) Phys Rev E 75:061102
58. Salmhofer M (1999) Renormalization: an introduction. Springer, Berlin
59. Schwartz N, Cohen R, ben-Avraham D, Barabasi A-L, Havlin S (2002) Phys Rev E 66:015104
60. Serrano MA, Boguna M (2006) Phys Rev Lett 97:088701
61. Serrano MA, Boguna M (2006) Phys Rev E 74:056115
62. Song C, Havlin S, Makse HA (2005) Nature 433:392
63. Song C, Havlin S, Makse HA (2006) Nature Phys 2:275
64. Song C, Gallos LK, Havlin S, Makse HA (2007) J Stat Mech P03006
65. Stanley HE (1971) Introduction to phase transitions and critical phenomena. Oxford University Press, Oxford
66. Vicsek T (1992) Fractal growth phenomena, 2nd edn. World Scientific, Singapore Part IV
67. Watts DJ, Strogatz SH (1998) Collective dynamics of "small-world" networks. Nature 393:440–442
68. Xenarios I et al (2000) Nucl Acids Res 28:289–291
69. Zhoua W-X, Jianga Z-Q, Sornette D (2007) Physica A 375: 741–752

Freeway Traffic Management and Control

A. Hegyi[1], T. Bellemans[2], B. De Schutter[1]

[1] TU Delft, Delft, The Netherlands

[2] Hasselt University, Diepenbeek, Belgium

Article Outline

Glossary

Acronyms and Abbreviations
Definition of the Subject
Introduction
Sensor Technologies
Traffic Flow Modeling
Freeway Traffic Control Measures
Network-Oriented Traffic Control Systems
Future Directions
Conclusion
Bibliography

Glossary

Feedforward control The block diagram of a feedforward control structure is shown in Fig. 1 [4]. The behavior of process **P** can be influenced by the control inputs. As a result the outputs (measurements or observations) show a given behavior. The controller **C** determines the control inputs in order to reach a given desired behavior of the outputs, taking into account the disturbances that act on the process. In the feedforward structure the controller **C** translates the desired behavior and the measured disturbances into control actions for the process.

The term feedforward refers to the fact that the direction of the information flow in the system contains no loops, i. e., it propagates only "forward".

The main advantages of a feedforward controller are that the complete system is stable if the controller and the process are stable, and that its design is in general simple.

Feedback control In Fig. 2 the feedback control structure is shown [4]. In contrast to the feedforward control structure, here the behavior of the outputs is coupled back to the controller (hence the name feedback). This structure is also often referred to as "closed-loop" control.

The main advantages of a feedback controller over a feedforward controller are that (1) it may have a quicker response (resulting in better performance), (2) it may correct undesired offsets in the output, (3) it may suppress unmeasurable disturbances that are observable through the output only, and (4) it may stabilize an unstable system.

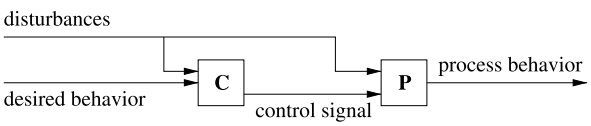

Freeway Traffic Management and Control, Figure 1
The feedforward control structure

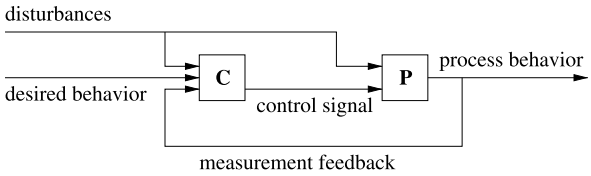

Freeway Traffic Management and Control, Figure 2
The feedback control structure

Optimal control Optimal control is a control methodology that formulates a control problem in terms of a *performance function*, also called an *objective function* [73]. This function expresses the performance of the system over a given period of time, and the goal of the controller is to find the control signals that result in optimal performance. Depending on the mathematical description of the control problem there exist several methods for the optimization of the control input including analytic and numerical approaches. Optimal control can be considered as a feedforward control approach.

Model predictive control Model predictive control (MPC) is an extension of the optimal control framework [15,79]. In Fig. 3 the block diagram of MPC is shown.

In MPC, at each time step k the optimal control signal is computed (by numerical optimization) over a prediction horizon of N_p steps. A control horizon $N_c (< N_p)$ can be selected to reduce the number of variables and to improve the stability of the system. Beyond the control horizon the control signal is usually taken to be constant. From the resulting optimal control signal only the first sample of the computed control signal is applied to the process. In the next time step $k + 1$, a new optimization is performed with a prediction horizon that is shifted one time step ahead, and of the resulting control signal again only the first sample is applied, and so on. This scheme, called rolling horizon, allows for updating the state from measurements, or even for updating the model in every iteration step.

In other words, MPC is equivalent to optimal control extended with feedback. The advantage of updating the state through feedback is that this results in a controller that has a low sensitivity to prediction errors. Regularly updating the prediction model results in an adaptive control system, which could be useful in situations where the model significantly changes, such as in case of incidents or changing weather conditions.

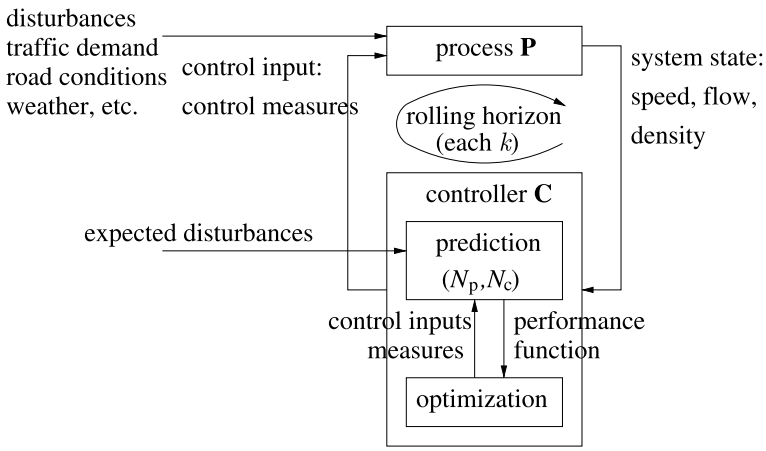

Freeway Traffic Management and Control, Figure 3
The model predictive control (MPC) structure

Acronyms and Abbreviations

MPC Model Predictive Control
OD Origin-Destination
ADAS Advanced Driver Assistance Systems
AHS Automated Highway System
IVHS Intelligent Vehicle/Highway System

Definition of the Subject

The goal of this chapter is to provide an overview of dynamic traffic control techniques described in the literature and applied in practice. *Dynamic traffic control* is the term to indicate a collection of tools, procedures, and methods that are used to intervene in traffic in order to improve the traffic flow on the short term, i. e., ranging from minutes to hours. The nature of the improvement may include increased safety, higher traffic flows, shorter travel times, more stable traffic flows, more reliable travel times, or reduced emissions and noise production.

The tools used for this purpose are in general changeable signs (including traffic signals, dynamic speed limit signs, and changeable message signs), radio broadcast messages, or human traffic controllers at the location of interest. Moreover, currently the possibilities of assisting, informing, and guiding drivers via in-car systems are also being explored.

The term *dynamic traffic management* includes besides dynamic traffic control also the management of emergency services and non-automated procedures (such as the implementation of predefined traffic control scenarios during special events), typically performed in traffic control centers. However, in this chapter the focus is on automatic control methods. Furthermore, this chapter deals exclusively with dynamic freeway traffic control techniques. Given the differences in traffic operation (e. g., higher speed limits) and in traffic infrastructure (e. g., intersections versus on-ramps and off-ramps), the control measures that can be implemented for urban and for freeway traffic differ. The interested reader is referred to [28,35,90] for an overview of urban traffic control.

Introduction

The number of vehicles and the need for transportation is continuously growing, and nowadays cities around the world face serious traffic congestion problems: Almost every weekday morning and evening during rush hours the capacity of many main roads is exceeded. Traffic jams do not only cause considerable costs due to unproductive time losses; they also augment the probability of accidents and have a negative impact on the environment (air pollution, lost fuel) and on the quality of life (health problems, noise, stress).

One solution to the ever growing traffic congestion problem is to extend the road network. Extending the freeway infrastructure is rather expensive, and in many countries this option is currently not considered to be a viable solution. Moreover, in densely populated areas building new roads is sometimes even unfeasible due to spacial limitations. Furthermore, there are often also other socio-economic objectives to be achieved, such as environmental objectives, which are considered alongside the objective of reducing congestion. Dynamic traffic control is an alternative that aims at increasing the safety and efficiency of the existing traffic networks without the necessity of adding new road infrastructure.

Since the 1960s traffic control is applied on freeway systems. However, during the last decades there have been developments in traffic science, traffic technology, control theory, and in the typical traffic patterns that all have consequences for the most appropriate traffic control approach. These developments will be discussed in the next sections.

The Need for Network-Oriented Automatic Traffic Control: Developments

The increasing complexity of the congested traffic patterns and the increasing availability of traffic control measures motivate the increasing usage of automatic traffic control systems and the increasing interest in network-oriented control over the last decades. The interest in network-oriented control from the practitioners' point of view is also motivated from policies aiming at socioeconomic goals, such as the efficient transport over important network corridors.

The fact that the length, the duration, and the number of traffic jams continues to grow has consequences for traffic control. When there are more locations with congestion, the available control measures have to solve more problems, which implies a higher complexity. Since nowadays the chances are higher that a vehicle encounters more than one traffic jam during one trip, the traffic control measures influencing a vehicle in one traffic jam will also influence the other jam(s) that the vehicle encounters. Therefore, the spatial interrelations between traffic situations at different locations in the network get stronger, and consequently the interrelations between the traffic control measures at different locations in the network also get stronger. These interrelations may differ per situation (and depend on, e. g., network topology, traffic demand, etc.) and the control measures may even counteract each other. For the various traffic management agencies or local governments that are responsible for different parts of the traffic network, this means that there is a need for a stronger cooperation and agreement on how the common network goals should be achieved. Similarly, for the automatic control methods, coordinated control strategies are required in these cases, to ensure that all available control measures serve the same objective, or at least that they do not counteract each other.

Another development is that freeways are equipped with more and more traffic control measures. The increasing number of control measures augments the controllability of the freeways. However, with this development the number of possible combinations of control measures also increases drastically, which in its turn increases the complexity of the dynamic traffic management problem.

On modern freeways often a large amount of data is available on-line and off-line. This data can serve as a basis for choices about appropriate control measures given the actual and expected traffic situation. However, the available data is currently not fully utilized either by traffic control center operators, whose actions are typically based on heuristic reasoning, or by automatic control measures, which mostly use local data only. Traffic data may also contain information about the current disturbances of the network (incidents, weather influences, unexpected demands) and information about the traffic system at a network level (about route choice and origin-destination relationships). The origin-destination (OD) matrix describes the traffic demand (vehicles per hour) appearing at each origin in a traffic network towards each destination in the network. An OD matrix may be time-varying, and can be calculated at different levels of temporal aggregation, e. g., hourly, peak or interpeak, 24 hour, etc. Methods have been developed to estimate such OD relationships from traffic measurements, and to estimate the traffic state (e. g., speeds, flows and densities) in networks that are incompletely equipped with detectors [43,124]. The methods can be used to supply a traffic control system with more accurate data, leading to better control actions.

These developments motivate the application of automatic control systems that can handle complex traffic scenarios, multiple control measures, and a large amount of data, and that can benefit from the network-oriented information by selecting appropriate control measures for given OD patterns and disturbances.

Regardless of these developments the three effects that cause the majority of suboptimal traffic network performances – in terms of travel time – have remained the same. Therefore, the primary goal of traffic control was and is to resolve the following effects/issues:

- **Bottlenecks.** Typical bottlenecks are freeway sections with an on-ramp, bridges, tunnels, curves, grades, and merge and diverge areas. The performance degradation typically originates from the phenomenon that the maximum achievable outflow from a traffic jam created at a bottleneck is often lower than the capacity of the road. This phenomenon is often called the capacity drop. A special case of a bottleneck is an upstream propagating jam that is growing at the tail by the incoming vehicles and resolving at the head by the leaving vehicles. A moving jam can be a serious bottleneck as it could reduce the maximum outflow to 70% of the capacity [59], while the capacities of the other bottle-

necks are in the range of 85–100% [37,59]. Dynamic traffic control measures may help to prevent a traffic breakdown at a bottleneck, or to improve the flow when a breakdown has occurred.

- **Suboptimal route choice.** In a dynamically changing network with jams, incidents, and road works the driver may not always be informed sufficiently well to make the optimal route choice. Furthermore, even if each individual driver has found the quickest (or in general, least costly) route to his or her destination, it may not lead to optimal performance at the network level, as known from the famous example of Braess [11]. Systems that influence the drivers' route choice may contribute to a better performance for the users, the network, or even both.
- **Blocking.** The tail of a traffic jam on a freeway may propagate so far upstream that it blocks traffic on a route that is not leading over the bottleneck that has caused the jam. A typical case is when a traffic jam created on the freeway at an on-ramp propagates back to an upstream off-ramp and blocks the traffic that wants to leave via the off-ramp. Figure 4 illustrates a situation where off-ramp traffic is blocked by a jam originating from the downstream on-ramp. All control measures that can limit the length of a traffic jam may in principle be applied to prevent blocking.

Automatic traffic control strategies try to optimize traffic network performance. A simplified, idealized description of the operation of traffic in the network links is given by what is known in traffic theory as the fundamental diagram [81]. The fundamental diagram describes steady-state traffic operation on a homogeneous freeway (i. e., the spatial gradients of speed, flow and density are equal to zero) as illustrated in Fig. 5. For low traffic densities, the relation between traffic density and traffic flow is nearly linear. For traffic densities smaller than the critical density ρ_{crit}, the traffic flow on the freeway increases with increasing traffic density (Fig. 5), despite the fact that the average speed decreases with increasing traffic density. Traffic operation is in a stable regime for traffic densities lower than the critical density. The maximal flow that can be achieved

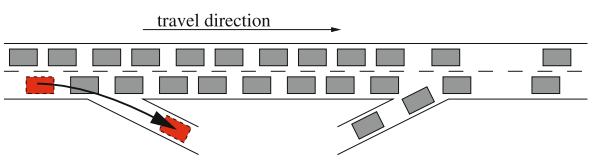

Freeway Traffic Management and Control, Figure 4
Congestion caused by high on-ramp demand could also result in the blocking of an upstream off-ramp

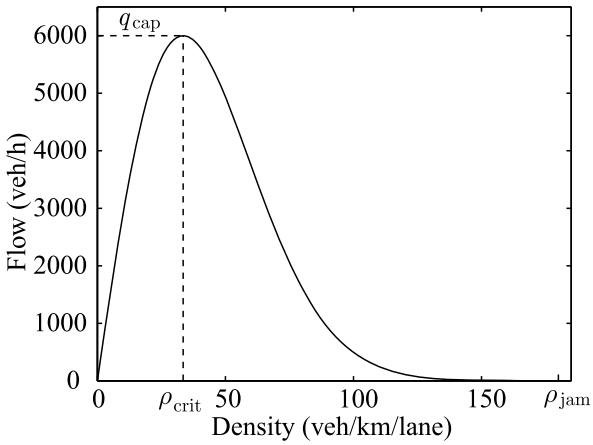

Freeway Traffic Management and Control, Figure 5
A flow-density fundamental diagram for a three lane freeway. As long as the traffic density on the freeway is smaller than the critical density ρ_{crit}, the traffic flow on the freeway increases with increasing traffic density. If the traffic density reaches the critical density, the flow is maximal and equal to the freeway capacity q_{cap}. If the traffic density further increases, the traffic flow on the freeway starts to decrease with increasing traffic density until the traffic stalls at the jam density ρ_{jam}

on the freeway, the capacity q_{cap}, is reached for a traffic density equal to the critical density and the resulting average speed of the vehicles is called the critical speed. If the critical density is exceeded, the average speed continues to decrease and the traffic flow decreases with increasing density. For traffic densities higher than the critical density, congestion sets in and an unstable traffic regime results. Typical values of ρ_{crit} and q_{cap} for a three-lane highway are 33.5 vehicles per kilometer and per lane and 6000 vehicles per hour respectively [94].

Based on the discussion of the fundamental diagram presented above, it can be observed that automatic control strategies can try to prevent or to reduce congestion by steering the state of traffic operation towards the stable region of operation. Here the fundamental diagram is merely presented as a seminal approach to traffic state analysis. However, in the literature alternative approaches to traffic state analysis have been reported. e. g., Kerner [60] has proposed the three phase theory, introducing the concepts "wide moving jams" and "synchronized traffic". The work by Treiber et al. [117] and Lee et al. [70] has added additional traffic states to the analysis such as "oscillating congested traffic" and "homogeneous congested traffic".

Automatic Traffic Control

Dynamic traffic management systems typically operate according to the feedback control concept known from con-

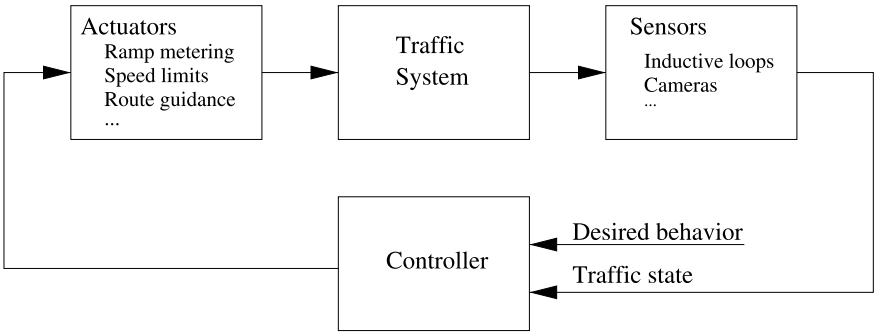

Freeway Traffic Management and Control, Figure 6
Schematic representation of the dynamic traffic management control loop. Based on the measurements provided by the sensors the controller determines the control signals sent to the actuators. Since the control loop is closed, the deviations from the desired traffic system behavior are observed and appropriate control actions are taken

trol systems theory, as shown in Fig. 6. The traffic sensors provide information about the current traffic state, such as speed, flow, density, or occupancy. The controller determines appropriate control signals that are sent to the actuators (depending on the system the changes in the control signal may be implemented instantly or may need to be phased in). The reaction of the traffic system is measured by the sensors again, which closes the control loop. If the new measurements show a deviation from the desired behavior (caused, e. g., by unforeseen disturbances), the new control signals are adapted accordingly. Note that there also exist traffic control systems that have a feedforward structure, e. g., the demand-capacity ramp metering approach that will be discussed in Sect. "Ramp Metering".

We define an "appropriate control" signal in terms of a control objective. From the network operator's point of view typical objectives are:

- **Efficiency.** Efficiency is often expressed in terms of throughput or travel time. This objective is shared by the network operators and the individual drivers. Nevertheless, situations may arise when the network operator and the individual driver have conflicting interests, e. g., minimizing the total travel time in a network (network optimum) may be conflicting with individually minimizing the travel times (user optimum). This will be discussed in more detail in Sect. "Route Guidance".
- **Safety.** In traffic control, safety may be a direct goal of the control or it may be a constraint (boundary condition) that should be satisfied. For example, dynamic speed limits or variable message signs may reduce the speed limit or give a warning under adverse weather conditions or poor visibility conditions in order to improve safety. Other systems may have other primary goals, such as improving the flow, and in these cases the control systems are often still required to be safety-neutral compared with the situation without control. There may also be an interaction between safety and efficiency, which has to be taken into account in the design of the control system. This interaction may be related to the following three processes. First, a safer traffic system in general results in fewer accidents and therefore more often higher flows may occur. Second, if congestion is prevented by an appropriate control method safety may be increased due to the more homogeneous flows. Third, lower speeds and densities in general positively influence safety. More specifically, Brownfield et al. [13] observed that for freeway sites, the accident rate in congested conditions was nearly twice the rate in uncongested conditions. However, the proportion of accidents that were serious or fatal was lower in congested conditions than in uncongested conditions. Hence, depending on the network operator's definition of safety, safety and efficiency may be conflicting or non-conflicting objectives.
- **Network reliability.** Even if not every traffic jam can be prevented, it is valuable for drivers when the travel time to their destinations is predictable, since good arrival time estimations make departure time choices easier. Therefore, improving the network reliability/predictability serves the economic efficiency of the network and improves driver comfort. Traffic control in general can improve reliability (predictability) by aiming at the realization of predicted travel times, or the reverse, by predicting realizable travel times, or both. Furthermore, network reliability can be improved by measures that aim at synchronization of the traffic demand and the capacity supply of the network, and at a better distribution traffic of flows over the

network. For a more elaborate discussion on network reliability we refer the interested reader to [6,21,78,113,114].

- **Low fuel consumption, low air and noise pollution.** In general, congestion contributes to less smooth journeys (more deceleration-acceleration movements), which increases emissions. In or near urban areas the environmental effects of traffic may be considered more important than, efficiency, for example, which can result in a different trade-off between the two objectives. An example of such a trade-off is between travel speed and air pollution [3]. The typical measure for these purposes is speed limitation.

Another important aspect of a traffic control system are the constraints due to physical, technical, or policy-based limitations. Such constraints may include minimum and maximum ramp metering rates, maximum on-ramp queue length, minimum and maximum dynamic speed limit values, etc. The automatic traffic controller is required to cope with these constraints.

Chapter Overview

The remainder of this chapter is organized as follows. We discuss the sensor technologies used in the context of freeway traffic control in Sect. "Sensor Technologies". In Sect. "Traffic Flow Modeling" we address traffic flow models, which play an important role in the design and evaluation of traffic control strategies. Next, the most frequently used freeway control measures are discussed in Sect. "Freeway Traffic Control Measures". While in Sect. "Freeway Traffic Control Measures" the focus is on the individual control measures, in Sect. "Network-Oriented Traffic Control Systems" we discuss the approaches to combine and to integrate several control measures for network-oriented control. We conclude in Sect. "Future Directions" by considering the new developments that are expected to play a role in future freeway traffic control systems.

Sensor Technologies

In order to implement traffic responsive freeway control, traffic measurements need to be collected at different locations throughout the freeway network. This section first deals with the most common traffic variables and traffic sensors to collect them. Next, the need for traffic demand and traffic routing data, and the way these data can be obtained using common traffic measurements, is briefly addressed. New data collection technologies, that will play an important role in future freeway control systems are discussed in Sect. "Future Directions".

Measurements

Traditionally the following traffic variables are measured to determine the traffic state on a freeway [81]: the traffic flow or the traffic intensity on the freeway, the average speed of the vehicles, the traffic density, the occupancy level of the freeway, the time headways (and in some cases the distance headways) and the speed variance. Note that occupancy is defined as the relative time (percentage) that the traffic sensor is occupied by a vehicle. In practice, it is often used as a surrogate measure for traffic density since it is directly related to density (as long as the average vehicle length is constant) and can be measured more easily than density.

Depending on the application and on the traffic measurement system, several levels of detail can be distinguished. The traffic variables can either be measured for every freeway lane separately or a value averaged over all lanes of the freeway can be obtained. Some measurement systems allow for a classification of the vehicles in categories based on their size (e. g., trucks versus cars) and provide the traffic variables per category. Furthermore, instantaneous values of the measured traffic variables can be provided or values averaged over a time period. The period over which the measurements are averaged can range from seconds over minutes to hours. As a rule of thumb, one can assume that the higher the level of detail of the data collected, the higher the cost of the measurement system involved.

In real-life situations, the measurements that are provided by the traffic sensors will contain measurement errors. These errors include incidental missing values, incidental faulty measurements, biased measurements, and missing values over a period of time. Given the importance of the traffic measurements in the dynamic traffic management control loop (Fig. 6), these errors need to be detected. Depending on the application, the controller may or may not be able to deal with erroneous or missing values. Techniques to estimate missing values that have been reported in the literature include reference days [24], multiple imputation [77], time series analysis [24], and Kalman and particle filtering [43,124].

For the purpose of freeway traffic control the most commonly measured traffic variables are the traffic flow, the speed, and the occupancy of the freeway traffic. The choice of these variables is influenced by their importance in traffic theory as well as by the ease by which they can be measured with most common traffic detector technologies.

There exists a wide variety of technologies [58] to measure traffic variables such as, pneumatic sensors, inductive

loops, cameras, ultrasonic sensors, microwave sensors, active and passive infrared sensors, passive acoustic arrays, and magnetometers.

Inductive loops are the most widespread detection systems to date and were introduced as traffic detection systems in the 60s [85]. The main advantages of inductive loops are their wide application range, the flexible design, and the availability of the common traffic variables. The main disadvantages of inductive loop technology are the sensitivity to wear and tear due to physical stress on the loops induced by traffic, the susceptibility of the loops to damage by road maintenance works, and the special installation and maintenance requirements (such as lane closure during maintenance) [58].

Traffic detection using video cameras emerged during the 80s and is a non-intrusive technology that is becoming more and more popular [85]. A fixed camera is mounted above the freeway and its images are sent to a video processing unit that extracts the desired variables using image processing algorithms. Camera traffic detection technology can provide the common traffic variables, is less prone to wear and tear by traffic, and generally requires less lane closures for maintenance and reconfiguration. The main disadvantages of traffic cameras are the higher upfront cost of the installation compared to inductive loops and the dependence on visibility conditions (such as fog, heavy snow, sunlight shining directly into the camera could heavily impair the quality of the images taken by the traffic cameras). Other factors that adversely affect the detection accuracy are vibrations caused by wind and traffic, lack of contrast between vehicle and road color, and varying lighting conditions (such as during dusk and dawn) [33,63,64].

In addition to the registration of the traditional traffic variables, video camera technology is also applied to register travel times on corridors or to obtain information regarding the routes followed throughout the network. This information can be obtained by tracking the vehicles at strategic locations (such as at large junctions or at the entrances and the exits of the area under consideration) using automated license plate recognition algorithms, for example. These systems consist of video image processing units connected to video cameras monitoring the traffic. Often, these systems are implemented at sites where the hardware to register the vehicles is already available (e. g., automated toll booths) [19,111].

Estimation

In order to coordinate or to integrate traffic control measures, the spatial aspect of the traffic network needs to be taken into account such that the impact of control measures on distant parts of the traffic network can be accounted for. However, the traffic sensors discussed above are traffic sensors that are localized in space, and as a consequence, they only yield information on the evolution of the traffic state on the freeway through time and at the sensor locations. Implementing traffic detectors very densely on the freeway in order to register the traffic states for every freeway section would be inconvenient and costly. However, data fusion techniques (such as extended Kalman filtering or particle filtering) allow one to combine traffic measurements scattered over the traffic network in order to obtain network traffic state estimation [43,124,126]. Traffic state estimation and prediction can be used in the implementation of control measures as is illustrated by simulation in [9].

The similarities between traffic flows and flows of compressible fluids have since long been documented in the literature [75,102] (see also Sect. "Traffic Flow Modeling" below). However, when it comes to the routing of traffic flows in networks, a major difference emerges; while the particles in a fluid have no predetermined destination, the vehicles traveling through a traffic network are traveling from a particular origin to a particular destination. Hence, the destination of the vehicles constrains the alternative routes that can be chosen. It is clear that route guidance control measures have an impact on the routing process. However, since travel times are typically an important factor in the routing process of well-informed travelers, other traffic measures, in combination with the traffic demands, influence the routing behavior as well. In order to assess the impact of traffic measures on the traffic states in the traffic network, the traffic demands (OD matrices) and the routing process need to be modeled. As the traffic demand cannot be measured directly, it needs to be estimated using the traffic measurements in the traffic network. Several techniques to estimate the traffic demands (OD matrices) that correspond to the measured traffic states in the traffic network have been developed. For an overview of the literature, the interested reader is referred to [76]. The route choice process, which influences the impact of control measures through rerouting effects, is also the subject of research [57].

Traffic Flow Modeling

Traffic flow models can be classified according to various criteria such as *area of application*, *level of detail*, *deterministic versus stochastic* [49].

An example of the application of traffic flow models for the design of traffic control measures is model predictive control, which makes use of an internal prediction model

in order to find the best traffic control measures to be applied to the real traffic process. Since these models are operated in real-time, and are often used to evaluate several control scenarios, they should allow for fast execution on a computer.

For the assessment of traffic control strategies often a simulation model is used instead of (or before) a real-world test. Simulation has several advantages. Above all, simulation is cheaper and faster, and it does not require real human drivers as test subjects. It also provides an environment where the unpredictable disturbances of a field test, such as weather influences, traffic demand variations, and incidents, can be excluded, or if necessary simulations can be repeated under exactly the same disturbance scenario.

Since none of the available traffic models perfectly describes the real traffic behavior, one has to keep in mind the intended application, when making the choice between the available traffic flow models. As Papageorgiou [89] argues for macroscopic traffic flow models an important criterion is that the model should have sufficient descriptive power to reproduce all important phenomena for the intended application. Similar arguments are also used by Kerner [60] but for different phenomena.

Traffic models can also be classified according to the level of detail with which they describe the traffic process:

- **Microscopic** models describe the behavior of individual vehicles. Important aspects of microscopic models are the so-called *car following* and *lane changing* behavior. Car following and lane changing behavior is generally described as a function of the distance to and (relative) speed of the surrounding vehicles, and the desired speed. Since the vehicles are modeled individually in microscopic traffic models, it is easy to assign different characteristics to each vehicle. These characteristics can be related to the driving style of the driver (aggressive, patient), vehicle type (car, truck), its destination, and route choice.
  A special type of microscopic traffic models are the *cellular-automaton* models [60,80,130] in which the freeway is discretized into cells of about 7.5 m length. Each cell can contain only one vehicle and the traffic dynamics is described by a probabilistic model of the hopping behavior of the vehicles through the cells.
  In general, it is difficult to calibrate microscopic models with real traffic data, due to the large number of parameters in this type of models and the poor availability of appropriate traffic data.
  We refer the interested reader to [2] for an extensive comparison of commercial microscopic simulation models and to [49] for a more theoretical overview.
- **Mesoscopic** models do not track individual vehicles, but describe the behavior of individual vehicles in probabilistic terms. Examples of mesoscopic models are headway distribution models [12] and gas-kinetic models [48]. Typically, these models are not used for traffic control.
- **Macroscopic** models use a high level of aggregation without distinguishing between individual vehicle behavior. Instead, traffic is described in aggregate terms as average speed, average flow, and average density.
  Macroscopic traffic flow modeling started when Lighthill and Whitham [75] presented in 1955 a model based on the analogy between traffic flows and flows in rivers. Independently of Lighthill and Whitham one year later Richards [102] published a similar model. Therefore, this model is usually referred to as the Lighthill–Whitham–Richards (LWR) model.
  Since 1955 a large variety of macroscopic traffic flow models have evolved from the LWR model with differences in the order of the model, the phenomena that they (re)produce (such as capacity drop, stop-and-go waves, and other congestion phenomena or patterns), and the effects of heterogeneous traffic (cars and trucks, etc.) [26,44,48,99].
  Another approach has been followed by Kerner [60] who developed a qualitative traffic flow theory based on empirical observation. This theory distinguishes three so-called traffic phases: free-flow, synchronized flow, and jammed traffic, and describes the transition between these phases qualitatively in probabilistic terms.

A last classification that is relevant in the context of traffic control is whether the model is deterministic or stochastic. Deterministic models define a relationship between model inputs, variables, and outputs that typically describe the average behavior of traffic. Stochastic models describe traffic behavior in terms of relationships between random variables, e. g., random reaction time of drivers, randomness in equilibrium speed-density (or car following) relationships, route choice, etc. These stochastic effects can reproduce phenomena such as the creation of traffic jams by random fluctuations in traffic flows [109], and can be used for the stochastic evaluation of traffic control approaches. Another application of stochastic traffic flow models is in the area of state estimation, which is an essential part of control approaches such as optimal control or model predictive control [43,124].

Freeway Traffic Control Measures

In this section we give an overview of control measures that are used or could be used to improve traffic perfor-

mance. We focus on control measures that are currently applied, or could be applied in the near future, such as ramp metering, speed limits, and route guidance. For each control measure we present the principle of operation including the control approaches, and the existing field tests and simulation results. At the end of this section some other traffic control measures are presented that may also be used to improve the performance of traffic systems.

Ramp Metering

Principle of Operation Ramp metering is one of the most investigated and applied freeway traffic control measures. A ramp metering set-up is implemented as a traffic signal that is placed at the on-ramp of a freeway as shown in Fig. 7. The required metering rate is implemented by appropriately choosing the phase lengths of the traffic signal. Several ramp metering implementations can be distinguished [20], e. g., single-lane with one vehicle per green ramp metering, single-lane with multiple vehicles per green ramp metering (bulk metering), and dual-lane ramp metering.

Ramp metering can be used in two modes: the *traffic spreading mode* and the *traffic restricting mode*. In the traffic spreading mode ramp metering smooths the merging process of on-ramp traffic by breaking the platoons and by spreading the on-ramp traffic demand over time as observed by Elefteriadou [31]. This mitigates the shock waves that can occur under high traffic density conditions. In this application the metering rate equals the average arrival rate of the vehicles.

Freeway Traffic Management and Control, Figure 7
Ramp metering at the freeway A13 in Delft, The Netherlands. One car may pass per green phase. To prevent red-light running the control is enforced

Restrictive ramp metering can be used for three different purposes:

- **Prevention of breakdowns.** When traffic is dense, ramp metering can prevent a traffic breakdown on the freeway by adjusting the metering rate such that the density on the freeway remains below the critical value. Preventing a traffic breakdown has not only the advantage that it results in a higher flow, but also that it prevents the creation of a jam that could block the off-ramp upstream the on-ramp (as shown in Fig. 4). These effects are studied in detail by Papageorgiou and Kotsialos [92]. Daganzo [27] has quantified the role of ramp metering in avoiding the activation of freeway gridlocks.
- **Influencing route choice.** Ramp metering can be implemented to influence the traffic demand and traffic routing. The impact of ramp metering on the traffic state and on the travel times is taken into account by the drivers in their routing behavior [132]. Banks [5] has described a theory to apply ramp metering to influence traffic routing to avoid freeway bottlenecks. Based on a similar idea Middelham [83] has performed a synthetic study on the route choice effects of ramp metering.
- **Localization of traffic jams.** According to Kerner [60] ramp metering can prevent the backpropagation of traffic jams and shock waves occurring at on-ramps. This could be beneficial on the network level since it could localize the traffic jam, and it could also be beneficial to the traffic throughput.

The control strategies that have been developed for restrictive ramp metering can be classified as static or dynamic, fixed-time or traffic-responsive, and local or coordinated.

Fixed-time strategies use (time-dependent) fixed metering rates that are determined off-line based on historical demands. This approach was first suggested by Wattleworth [127], and was extended to a dynamic traffic model by Papageorgiou [87]. The disadvantage of fixed-time strategies is that they do not take into account the day-to-day variations in the traffic demand or the variations in the demand during a period with a constant metering rate, which may result in underutilization of the freeway or inability to prevent congestion.

Traffic-responsive strategies solve these issues by adjusting on-line the metering rate as a function of the prevailing traffic conditions. These strategies also aim at the same objectives as the fixed-time strategies, but use direct traffic measurements instead of historical data to prevent or to reduce congestion. One of the best known strategies

is the *demand-capacity* strategy [91]:

$$q_{\text{ramp}}(k) = \begin{cases} q_{\text{cap}} - q_{\text{in}}(k-1) & \text{if} \quad o_{\text{meas}}(k-1) \le o_{\text{cr}} \\ q_{\text{r,min}} & \text{otherwise} \end{cases}$$

with $q_{\text{ramp}}(k)$ the admitted ramp flow at time step k, q_{cap} the downstream freeway capacity, $q_{\text{in}}(k)$ the freeway flow measured upstream of the on-ramp at time step k, $q_{\text{r,min}}$ the minimal on-ramp flow during congestion, $o_{\text{meas}}(k)$ the occupancy downstream the on-ramp at time step k, and o_{cr} is the critical occupancy (at which the flow is maximal). Since the traffic state on the freeway cannot be determined based on the measurement of the traffic flow alone, the downstream occupancy is measured in order to determine whether congestion is present ($o_{\text{meas}}(k-1) > o_{\text{cr}}$) or not.

A similar strategy is occupancy-based ramp metering, where the upstream traffic flow measurement from demand-capacity ramp metering is replaced by an occupancy measurement. The measured occupancies are related to traffic flows based on historical measurements. Next, the demand-capacity approach described above can be applied [18]. A common disadvantage of both demand-capacity formulations is that they have an (open-loop) feedforward structure, which is known to perform poorly under unknown disturbances and cannot guarantee a zero offset in the output under steady-state conditions.

A better approach is to use a (closed-loop) feedback structure, because it allows for controller formulations that can reject disturbances and have zero steady-state error. ALINEA [96] is such a ramp metering strategy and its control law is defined as follows:

$$q_{\text{ramp}}(k) = q_{\text{ramp}}(k-1) + K\left(\hat{o} - o_{\text{meas}}(k)\right) \ ,$$

where $q_{\text{ramp}}(k)$ is the metered on-ramp flow at time step k, K is a positive constant, $\hat{o}$ is a set-point for the occupancy, and $o_{\text{meas}}(k)$ is the measured occupancy on the freeway downstream of the on-ramp at time step k. ALINEA tries to maintain the occupancy on the freeway equal to a set-point $\hat{o}$, which is chosen in the region of stable operation. Given the probabilistic nature of traffic operation, the set-point $\hat{o}$ is often chosen somewhat smaller than the critical occupancy in order to guarantee free-flow traffic operation.

More advanced ramp metering strategies are the traffic-responsive coordinated strategies such as METALINE [95], FLOW [55], or methods that use optimal control [66] or model predictive control [7].

The ramp metering strategies discussed above attempt to conserve free-flow traffic on the freeway. However, given the probabilistic nature of traffic operation, congestion can set in at lower or higher densities than the critical density. Based on these insights, Kerner [60,61] defined a *congested-pattern control approach* to ramp metering called ANCONA. The basic idea of ANCONA is to allow congestion to set in, but to keep congested traffic conditions to the minimum level possible. Once congestion sets in, ANCONA tries to reestablish free-flow conditions on the freeway by reducing the on-ramp metering rate. Kerner claims that, by allowing congestion to set in, ANCONA utilizes the available freeway capacity better. The control rule of ANCONA is given by [60]:

$$q_{\text{ramp}}(k) = \begin{cases} q_1 & \text{if} \quad v_{\text{det}}(k) \le v_{\text{cong}} \\ q_2 & \text{if} \quad v_{\text{det}}(k) > v_{\text{cong}} \end{cases} \ ,$$

where $q_{\text{ramp}}(k)$ is the on-ramp flow at time step k, q_1 and q_2 are heuristically determined constant flows with $q_1 < q_2$, $v_{\text{det}}(k)$ is the traffic speed on the freeway just upstream of the on-ramp at time step k, and v_{cong} is the speed threshold that separates the free and the synchronized (locally congested) traffic flow phases.

Field Tests and Simulation Studies Several field tests and simulation studies have shown the effectiveness of ramp metering. In Paris on the Boulevard Périphérique and in Amsterdam several ramp metering strategies have been tested [97]. The demand-capacity, occupancy, and ALINEA strategies were applied in the field tests at a single ramp in Paris. It was found that ALINEA was clearly superior to the other two in all the performance measures (total time spent, total traveled distance, mean speed, mean congestion duration). At the Boulevard Périphérique in Paris the multi-variable (coordinated) feedback strategy METALINE was also applied and was compared with the local feedback strategy ALINEA. Both strategies resulted in approximately the same performance improvement [95]. One of the largest field tests was conducted in the Twin Cities metropolitan area of Minnesota. In this area 430 operational ramp meters were shut down to evaluate their effectiveness. The results of this test show that ramp metering not only serves the purposes of improving traffic flow and traffic safety, but also improves travel time reliability [16,72].

A number of studies have simulated ramp metering for different transportation networks and traffic scenarios, with different control approaches, and with the use of microscopic and macroscopic traffic flow models [39,45,66,94,96,112]. Generally the total network travel time is considered as the performance measure and is improved by about 0.39–30% when using ramp metering.

The validation of ANCONA versus ALINEA is performed by simulation by Kerner [61], who found that ANCONA in some cases can lead to higher flows.

Also note that Kerner has criticized modeling approaches to simulations of freeway traffic control strategies in [62] (see also [98] for some comments).

Dynamic Speed Limits

Dynamic speed limits are used to reduce the maximum speed on freeways according to given performance, safety, or environmental criteria. An example of a dynamic speed limit gantry is shown in Fig. 8.

Principle of Operation The working principle of speed limit systems can be categorized based on their intended effects: improving safety, improving traffic flow, or their environmental effects, such as reducing noise or air pollution.

It is generally accepted that speed reduction on freeways leads to improved safety [106,116,128]. Lower speeds in general are associated with lower crash rates and with a lower impact in case of a collision. If the environmental conditions or traffic conditions are such that the posted maximum speeds are considered to be unsafe, the speed limit can be lowered to match the given conditions. Dynamic speed limits may function as a warning that an incident or jam is present ahead.

In the literature, basically two approaches to dynamic speed limit control can be found for flow improvement. The first emphasizes the homogenization effect [1,38,69,109,122], whereas the second is more focused on preventing traffic breakdown by reducing the flow by means of speed limits [23,42,71].

Freeway Traffic Management and Control, Figure 8
A variable speed limit gantry on the A13 freeway near Overschie, The Netherlands. In this particular case the maximum speed limit is 80 km/h due to environmental reasons, and the limit may drop to 70 km/h or 50 km/h in case of a downstream jam

- The basic idea of homogenization is that speed limits can reduce the speed (and/or density) differences, by which a more stable (and safer) flow can be achieved. The homogenizing approach typically uses speed limits that are above the critical speed (i. e., the speed that corresponds to the maximal flow). So, these speed limits do not limit the traffic flow, but only slightly reduce the average speed (and slightly increase the density). In theory this approach can increase the time to breakdown slightly [109], but it cannot suppress or resolve shock waves. An extended overview of speed limit systems that aim at reducing speed differentials is given by Wilkie [128].
- The traffic breakdown prevention approach focuses more on preventing too high densities, and also allows speed limits that are lower than the critical speed in order to limit the inflow to these areas. By resolving the high-density areas (bottlenecks) higher flow can be achieved in contrast to the homogenization approach [42].

Currently, the main purpose of most of the existing practical dynamic speed limit systems is to increase safety by lowering the speed limits in potentially dangerous situations, such as upstream of congested areas or during adverse weather conditions [106,116,128]. Although these systems primarily aim at safety, in general they also have a positive effect on the flow, due to the fact that preventing accidents results in a higher flow. There are also some examples of practical systems that are designed with the purpose of flow improvement [101,104] – with varying success. These practical systems in general use a switching scheme based on traffic conditions, weather conditions, visibility conditions, or pavement conditions [100,129].

Several control methodologies are used in the literature to find a control law for speed control, such as multi-layer control [74], sliding-mode control [71], and optimal control [1]. In [29] optimal control is approximated by a neural network in a rolling horizon framework. Other authors use, or simplify their control law to, a control logic where the switching between the speed limit values is based on traffic volume, speed, or density [38,69,71,109,122]. We refer the interested reader for further reading about the various control methodologies to the references at the end of this chapter.

Some authors recognize the importance of anticipation in the speed control scheme. A pseudo-anticipative scheme is used in [71] by switching between speed limits based on the density of the neighboring downstream seg-

ment. Explicit predictions are used in [1,29] and this is the only approach that results in a significant flow improvement. The heuristic algorithm proposed in [128] also contains anticipation to shock waves being formed.

Another concept of dynamic speed limits is their use in combination with ramp metering to prevent a breakdown on the freeway at the on-ramp and to prevent the ramp queue to propagate back to the urban network [41] by taking over the flow limitation function from the ramp metering when the ramp queue has reached its limit.

Field Tests and Simulation Studies Field data evaluations show that in general homogenization results in a more stable and safer traffic flow, but no significant improvement of traffic volume is expected nor measured [104,122]. Since the introduction of speed control on the M25 in the United Kingdom, an increase of flow of 1.5% per year is reported for the morning peaks, but no improvement is found in the afternoon peaks [101].

The effect of dynamic speed limits on traffic behavior strongly depends on whether the speed limits are enforced or not, and on whether the speed limits are advisory or mandatory, which also determines the suitability for a certain application. Most application oriented studies [110,122,128] enforce speed limits, except for [69]. Enforcement is usually accepted by the drivers if the speed limit system leads to a more stable traffic flow.

Route Guidance

Principle of Operation Route guidance systems assist drivers in choosing their route when alternative routes exist to their destination. The systems typically display traffic information such as congestion length, the delay on the alternative routes, or the travel time to the next common point on the alternative routes (an example is given in Fig. 9). Recently, in-car navigation system manufacturers have shown interest in providing route advice taking the traffic jams and travel times on the alternative routes into account.

In route guidance the notions of *system optimum* and *user equilibrium* (or *user optimum*) play an important role. The system optimum is achieved when the vehicles are guided such that the total costs of all drivers (typically the total travel time) is minimized. However, the system optimum does not necessarily minimize the travel time (or some generalized cost measure) for each individual driver. So, some drivers may select another route that has a shorter individual travel time (lower cost). The traffic network is in user equilibrium when on each utilized route the costs are equal, and on routes that are not utilized the cost is higher than that on the utilized routes. This means that no driver has the possibility to find another route that reduces his or her individual cost.

Freeway Traffic Management and Control, Figure 9
A route guidance system in The Netherlands showing traffic jams lengths on alternative routes to Schiphol Airport (Photo courtesy of Peek Traffic B.V.)

The cost function is typically defined as the travel time, either as the *predicted travel time* or as the *instantaneous travel time*. The predicted travel time is the time that the driver will experience when he or she drives along the given route, while the instantaneous travel time is the travel time determined based on the current speeds on the route. In a dynamic setting the speeds in the network may change during a trip, and consequently the instantaneous travel time may be different from the predicted travel time.

Papageorgiou and Messmer [93] have developed a theoretical framework for route guidance in traffic networks. Three different traffic control problems are formulated: an optimal control problem to achieve the system optimum (minimize the total time that is spent in the network), an optimal control problem to achieve a user optimum (equalize travel times), and a feedback control problem to achieve a user optimum (equalize travel times). The resulting control strategies are demonstrated on a test network with six pairs of alternative routes. The feedback control strategy is tested with instantaneous travel times and results in a user equilibrium for most alternative routes, and the resulting total time spent by the vehicles in the network is very close to the system optimum.

Wang et al. [125] combine the advantages of a feedback approach (relatively simple, robust, fast) and predicted travel times. The resulting *predictive feedback* controller is compared with optimal control and with a feedback controller based on instantaneous travel times.

When the disturbances are known the simulation results show that the predictive feedback results in nearly optimal splitting rates, and is clearly superior to the feedback based on instantaneous travel times. The robustness of the feedback approach is shown for several cases: incorrectly predicted demand, an (unpredictable) incident, and an incorrect compliance rate.

Field Tests and Simulation Studies In several studies it is assumed that the turning rates can be directly manipulated by route guidance messages [93,125]. In the case of in-car systems it is plausible that by giving direct route advice to individual drivers the splitting rates can be influenced sufficiently. However, in the case of route guidance by variable message signs the displayed messages do not directly determine the splitting rates: The drivers make their own decisions. Therefore, empirical studies about the reaction of drivers to dynamic route information messages, and the effectiveness of route guidance can provide useful information.

Kraan et al. [68] present an extensive evaluation of the impact on network performance of variable message signs on the freeway network around Amsterdam, The Netherlands. Several performance indicators are compared before and after the installation of 14 new variable message signs (of which seven are used as incident management signs and seven as dynamic route information signs). The performance indicators, such as the total traveled distance, the total congestion length and duration, and the instantaneous travel time delay are compared for alternative routes and for most locations a small but statistically significant improvement is found. The day-to-day standard deviation of these indicators decreased after the installation of the variable message signs, which indicates that the travel times have become more reliable.

Another field test is reported by Diakaki et al. [30] in which a combination of route guidance, ramp metering, and urban traffic control is applied to the M8 corridor network in Glasgow, UK. The applied control methodology resulted in an increased network throughput and in a reduced average travel time.

Other Control Measures

Besides ramp metering, dynamic speed limits, and route guidance, there are also other dynamic traffic control measures that can potentially improve the traffic performance. In this section we describe a selection of such measures, and describe in which situations they are useful (cf. [84]).

- **Peak lanes.** During peak hours the hard shoulder lane of a freeway (which is normally used only by vehicles in emergency) is opened for traffic. Whether the lane is opened or closed is communicated by variable message signs showing a green arrow or a red cross. Due to the extra lane the capacity of the road is increased, which could prevent congestion. The disadvantage of using the emergency lane as a normal lane is that the safety may be reduced. For this reason, often extra conditions ensuring safety are required, such as the creation of emergency refuges adjacent to the hard shoulder lane, or the requirement that emergency services should be able to access the incident location over or through the guard rail. Furthermore, there may be CCTV surveillance or vehicle patrols to detect incidents early. This traffic control measure is useful where the additional capacity prevents congestion and the downstream infrastructure can accommodate the increased traffic flow.
- **Dedicated lanes.** During congestion the shoulder lane may be opened for dedicated vehicles, such as public transport, freight transport, or high occupancy vehicles (with more than two passengers). This reduces the hindrance that congestion causes to these vehicles. Furthermore, public transport can be made more reliable and thus more attractive by this measure. A dedicated freight transport lane increases the stability and homogeneity of the traffic flow.
- **Tidal flow.** Tidal flow allows one to use a freeway lane in the one or the other direction. Depending on the direction of the highest traffic demand the direction of operation is determined. This direction is communicated by a variable message sign showing a red cross or a green arrow. This traffic control measure is useful when the traffic demand is typically not high in both directions simultaneously.
- **The "keep your lane" directive.** When the "keep your lane" directive is displayed, the drivers are not allowed (not recommended) to change lanes. This results in less disturbances in the freeway traffic flow, which may prevent congestion. This traffic control measure is useful when the traffic flow is nearly unstable (close to the critical density) and may be a good alternative to homogenizing speed limits.

Network-Oriented Traffic Control Systems

The integration of traffic control measures in freeway networks is essential in order to ensure that the control actions taken at different locations in the network reinforce rather than counteract or even cancel each other. While in the previous section individual traffic control measures were discussed along with the most prevalent local con-

trol strategies, this section explicitly considers the integration of several control measures in a freeway network context.

Although the focus in this chapter is on automatic control systems, it must be noted that in practice in traffic control centers there is also often a human controller with "oversight" of the system as a safeguard against problems with the system and in view of the complexity of the control problem.

In network-oriented traffic control two ingredients play an important role: *coordination* and *prediction*. Since in a dense network the effect of a local control measure could also influence the traffic flows in more distant parts of the network, the control measures should be coordinated such that they serve the same objectives. Taking into account the effects of control measures on distant parts of the network often also involves prediction, due to the fact that the effect of a control measure has a delay that is at least the travel time between the two control measures in the downstream direction, and at least the propagation time of shock waves in the upstream direction. An advantage of control systems that use explicit predictions is that by anticipating on predictable future events the control system can also *prevent* problems instead of only *reacting* to them. However, it must be noted that while all network-oriented control approaches apply some form of coordination, many approaches do not explicitly make use of predictions.

Network-oriented traffic control has several advantages compared to local control since it ensures that local traffic problems are solved with the aim of achieving an improvement on the network level. For example, solving a local traffic jam only can have as consequence that the vehicles run faster into another (downstream) jam, whereas still the same amount of vehicles have to pass the downstream bottleneck (with a given capacity). In such a case, the average travel time on the network level will still be the same, regardless of whether or not the jam is solved. However, a global approach would take into account both jams and, if possible, solve both of them.

Furthermore, network-oriented control approaches can utilize network-related historical information. For example, if dynamic OD data is available, control on the network level can take advantage of the predictions of the flows in the network. Local controllers are not able to optimize the network performance even if the dynamic OD data is available, because the effect of the control actions on downstream areas is not taken into account.

In the literature basically three approaches exist for coordinating traffic control measures: model-based optimal control methods, knowledge-based methods, and methods that use simple feedback or switching logic, for which the parameters are optimized. In some approaches different methods are combined in a hierarchical control structure. We discuss these approaches in the following subsections.

Model-Based Control Methods

Model-based traffic control techniques use a traffic flow model for predicting the future behavior of the traffic system based on

- the current state of traffic,
- the expected traffic demand on the network level, possibly including OD relationships and external influences, such as weather conditions,
- the planned traffic control measures.

Since the first two items cannot be influenced (except for the possibility that based on real-time congestion information people cancel their planned trip, change the departure time, or travel via another modality), the future performance of the traffic system is optimized by selecting an appropriate scenario for the traffic control measures. Methods that use *optimal control* or *model predictive control* explicitly take the complex nonlinear nature of traffic into account. For example, they take into account the fact that the effect of ramp metering on distant on-ramps will be delayed by the (time-varying) travel time between the two on-ramps. In general, the other existing methods (such as knowledge-based methods, or control parameter optimization) do not explicitly take this kind of delay into account. Furthermore, other advantages of the model-based methods are that traffic demand predictions can be utilized, constraints on the ramp metering rate and the ramp queues can be included easily, and a user-supplied objective function can be optimized.

Optimal control has been successfully applied in simulation studies to integrated control of ramp metering and freeway-to-freeway control [66], to route guidance [47] and to integration of ramp metering and route guidance [66,67]. In [66,67] the integrated controller performed better than route guidance or ramp metering alone.

The model predictive control (MPC) approach is an extension of the optimal control, which uses a rolling horizon framework. This results in a closed-loop (feedback) controller, which has the advantage that it can handle demand prediction errors and disturbances (such as incidents). MPC is computationally more efficient than optimal control due to the shorter prediction and control horizons, and it can be made adaptive by updating the prediction model on-line.

MPC-based control has been applied in simulations to coordinated ramp metering [8], to integrated control of ramp metering and dynamic speed limits [41], and to integrated control of ramp metering and route guidance [57,121]. An illustration of the ability of MPC-based traffic control to deal with a model mismatch was given in [9]. In [7], it was illustrated by simulation of a simple proof-of-concept network that MPC can be implemented to account for the rerouting behavior of vehicles due to changing travel times caused by applying ramp metering.

Knowledge-Based Methods

Knowledge-based traffic control methods typically describe the knowledge about the traffic system in combination with the control system in terms that are comprehensible for humans. Given the current traffic situation the knowledge-based system generates a solution (control measure) via reasoning mechanisms. A typical motivation for these systems is to help traffic control center operators to find good (not necessarily the best) combinations of control measures. The operators often suffer from cognitive overload by the large number of possible actions (control measures) or by time pressure in case of incidents. The possibility for the operators to track the reasoning path of the knowledge-based system makes these systems attractive and more convincing.

An example of a knowledge-based system is the TRYS system [25,46,86], which uses knowledge about the physical structure of the network, the typical traffic problems, and about effects of the available control measures. The TRYS system has been installed in traffic control centers in Madrid and Barcelona, Spain.

Another knowledge-based system is the freeway incident management system [36] developed in Massachusetts, which assists in the management of non-recurrent congestion. The system contains a knowledge base and a reasoning mechanism to guide the traffic operators through the appropriate questions to manage incidents. Besides incident detection and verification, the system assists in notifying the necessary agencies (such as, ambulance, clean-up forces, towing company) and in applying the appropriate diversion measures. The potential benefits (reduced travel times by appropriate diversion) are illustrated by a case study on the Massachusetts Turnpike. The knowledge-based expert system called freeway real-time expert-system demonstration [103,131] has similar functionalities and is illustrated by applying it to a section of the Riverside Freeway (SR-91) in Orange County, California.

Control Parameter Optimization Methods

Allessandri and Di Febbraro [1] follow another approach: A relatively simple control law is used for speed limit control and ramp metering, and the parameters of the control law are found by simulating a large number of scenarios and optimizing the average performance. In [1] a dynamic speed limit switching scheme is developed. The speed limits switch between approximately 70 km/h and 90 km/h, and the switching is based on the density of the segment to be controlled and two thresholds (to switch up and to switch down). The switching scheme uses a hysteresis loop to prevent too frequent switching. Optimizing the thresholds for several objectives resulted in a slight increase of the average throughput, a decrease of the sum of squared densities – which can be considered as a measure of inhomogeneity (since a non-uniform distribution of vehicles over a freeway stretch results in a higher sum of squared densities) – and a small decrease of the total time spent by the vehicles in the network.

Hierarchical Control

The increasing number of traffic control measures that need to be controlled in a network-control context, as well as their interactions, drastically increases the computational complexity of computing the optimal control signals. Hierarchical control was introduced by some authors in order to tackle this problem [22,65,88]. In hierarchical control the controlled process is partitioned in several subprocesses, and the control task is performed by a high-level controller and several low-level controllers. The high-level controller determines centrally the set-points or trajectories representing the desired behavior of the subprocesses. The low-level controllers are used to steer the subprocesses according to the set-points or trajectories supplied by the high-level controller. Usually, the high-level controller operates at a slower time scale than the low-level controllers. Hierarchical systems do not only enable coordination of control for large networks, but they also provide high reliability and robustness [50].

Future Directions

Although there is a large interest in developing freeway traffic control systems, there is by no means a consensus about the most suitable approaches or methods. One of the reasons is that traffic phenomena, such as traffic breakdown and jam resolution, are not perfectly understood [60] and different views lead to different approaches. In addition, technological developments such as advanced

sensor technologies and intelligent vehicles open new possibilities that enable or require new control approaches.

Advanced Sensor Technologies

Given the complexity of traffic state estimation, traffic demand estimation, and the collection of routing information based on conventional traffic measurement data, new data collection methods are being investigated.

Instead of registering vehicles at certain locations using hardware on the freeway network, floating car data can be collected. The collection of floating car data, where individual vehicles are tracked during their journey through the network, provides valuable route choice and traffic demand information. The evolution in mobile computing and in mobile communication has enabled the incorporation of these technologies in the field of traffic data collection, allowing more detailed and more cost-effective data collection. In contrast to the traditional data collection methods that were discussed in Sect. "Measurements", this section deals with two data collection methodologies that are enabled by mobile computing and communication.

Cell phone service providers collect data regarding the base station each cell phone connects to and the time instant the connection is initiated. Since cell phones regularly connect to their current base station and since the location of these base stations is known, information about the journey of the cell phone can be extracted from the service provider's database. By monitoring a large number of cell phones, and more in particular their hands-off processes when hopping from one base station to the next, an impression of the traffic speeds and the travel times can be obtained [108].

The global positioning system (GPS) is well-suited for tracking probe vehicles through space and time in order to obtain route information and travel times [14,115]. With the further miniaturization of electronics, the processing power available in mainstream navigation units and mobile data communication facilities (e. g., GPRS) the cost of instrumenting fleets of probe vehicles decreases. For example, fleets of taxis, buses, and trucks can be used as probe vehicles as they are often readily equipped with GPS and data communication technology. When dealing with probe vehicles, care must be taken to ensure that the number of probe vehicles is large enough in order to be able to accurately determine the traffic state [56].

Although the technologies presented above are readily available and have been used in the past, their structural deployment as a source for traffic measurements for dynamic traffic control systems still needs to break through. Some issues that may determine whether floating car data becomes a viable option for large-scale data collection are the accuracy of the data obtained, privacy concerns related to registering the whereabouts of individuals, operational communication and computation costs, and standardized mobile or in-vehicle availability of communication and GPS functionality.

Intelligent Vehicles and Traffic Control

We now discuss recent and future developments in connection with intelligent vehicles that can further improve the performance of traffic management and control systems by offering better and more accurate ways to collect traffic data and to apply traffic control measures.

Advanced Driver Assistance Systems The increasing demand for safer passenger cars has stimulated the development of advanced driver assistance systems (ADAS). An ADAS is a control system that uses environment sensors to improve comfort and traffic safety by assisting the driver. Some examples of ADAS are cruise control, forward collision warning, lane departure warning, parking systems, and pre-crash systems for belt-pretensioning [10]. Although traffic management is not the primary goal of ADAS, they can contribute to a better traffic performance [120], either in a more passive way by avoiding incidents and by providing smoother traffic flows, or in an active way by coordination and communication with neighboring vehicles and roadside infrastructure.

The increasing market penetration and use of ADAS and of other in-car navigation, telecommunication, and information systems offer an excellent opportunity to implement a next level of traffic control and management, which shifts away from the road-side traffic management toward a vehicle-oriented traffic management. In this context both inter-vehicle management and road-side/vehicle traffic management and interaction can be considered. The goal is to use the additional measures and control handles offered by intelligent vehicles and to develop control and management methods to substantially improve traffic performance in terms of safety, throughput, reliability, environment, and robustness.

Some examples of new traffic control measures that are made possible by intelligent vehicles are cooperative adaptive cruise control [119] (allowing one to control intervehicle distances), intelligent speed adaptation [17] (allowing one to better and more dynamically control vehicle speeds), and route guidance [82] (where the traffic control centers could on the one hand get data about planned routes and destinations, and on the other hand also send real-time information and control data to the on-board

route planners, for instance, to warn about current and predicted congestion and possibly also to spread the traffic flows more evenly over the network).

These individual ADAS-based traffic control measures could be integrated with roadside traffic control measures such as ramp metering, traffic signals, lanes closures, shoulder lane openings, etc. The actual control strategy could then also make use of a model-based control approach such as MPC.

Cooperative Vehicle-Infrastructure Systems The new intelligent-vehicle technologies allow communication and coordination between vehicles and the roadside infrastructure and among vehicles themselves. This results in cooperative vehicle-infrastructure systems, which can also be seen as a first step towards fully automated highway systems, which will be discussed below. CVIS (Cooperative Vehicle-Infrastructure Systems) [52] is a European research project that aims to design, develop, and test technologies that allow communication between the cars and with the roadside infrastructure, which improves road safety and efficiency, and reduces environmental impact. This project allows drivers to influence the traffic control system directly and also to get information about the quickest route to their destination, speed limits on the road, as well as warning messages via wireless technologies.

Automated Highway Systems ADAS and cooperative vehicle-infrastructure systems can even be extended several steps further towards complete automation. Indeed, one approach to augment the throughput on highways is to implement a fully automated system called Automated Highway System (AHS) or Intelligent Vehicle/Highway System (IVHS) [40,123], in which cars travel on the highway in platoons with small distances (say, 2 m) between vehicles within the platoon, and much larger distances (say, 30–60 m) between different platoons. Due to the very short intra-platoon distances this approach requires automated distance-keeping since human drivers cannot react fast enough to guarantee adequate safety. So in AHS every vehicle contains an automated system that can take over the driver's responsibilities in steering, braking, and throttle control. Due to the short spacing between the vehicles within the platoons, the throughput of the highway can increase, allowing it to carry as much as twice or three times as many vehicles as in the present situation. The other major advantages of the platooning system are increased safety and fuel efficiency. Safety is increased by the automation and close coordination between the vehicles, and is enhanced by the small relative speed between the cars in the platoon. Because the cars in the platoon travel together at the same speed, a small distance apart, even high accelerations and decelerations cannot cause a severe collision between the cars (due to the small relative speeds). The short spacing between the vehicles also produces a significant reduction in aerodynamic drag for the vehicles, which leads to improvements in fuel economy and emissions reductions.

Automated platooning has been investigated very thoroughly within the PATH program [54,105]. Related programs are the Japanese Dolphin framework [118] and the Auto21 Collaborative Driving System framework [51, 53].

Although certain authors argue that only full automation can achieve significant capacity increases on highways and thus reduce the occurrences of traffic congestion [123], AHS do not appear to be feasible on the short term. The AHS approach requires major investments to be made by both the traffic authority and the constructors and owners of the vehicles. Since few decisions are left to the driver, and since the AHS assumes almost complete control over the vehicles, which drive at high speeds and at short distances from each other, a strong psychological resistance to this traffic congestion policy is to be expected. Another important question is how the transition of the current highway system to an AHS-based system should occur, and – once it has been installed – what has to be done with vehicles that are not yet equipped for AHS. So before such systems can be implemented, many financial, legislative, political and organizational issues still have to be resolved [34].

Conclusion

In this chapter we have presented an overview of freeway traffic control theory and practice. In this context we have discussed traffic measurements and estimation, individual traffic control measures, and the approaches behind them that relate the control signals to the given traffic situation. The trend of the ever-increasing traffic demands and the appearance of new control technologies have led to the new field of network-oriented traffic control systems. Although there have been many interesting publications about the theory and practice of integrated traffic control, several challenges remain, such as the integration of traffic state estimation and dynamic OD information in the control approaches.

The lively research in freeway traffic control shows that this field is still practically relevant and theoretically challenging. Facing these challenges can be expected to lead to new freeway traffic control approaches in theory and

practice resulting in higher freeway performance in terms of efficiency, reliability, safety and environmental effects. Furthermore, future developments in the field of in-car systems and advanced sensor technologies are expected to enable new traffic management approaches that may measure and control traffic in more detail and with higher performance.

Bibliography

Primary Literature

1. Alessandri A, Di Febbraro A, Ferrara A, Punta E (1999) Nonlinear optimization for freeway control using variable-speed signaling. IEEE Trans Veh Technol 48(6):2042–2052
2. Algers S, Bernauer E, Boero M, Breheret L, Di Taranto C, Dougherty M, Fox K, Gabard J-F (2000) SMARTEST – Final report for publication. Technical report, ITS, University of Leeds. http://www.its.leeds.ac.uk/projects/smartest. Accessed on 22 July 2008
3. André M, Hammarström U (2000) Driving speeds in europe for pollutant emissions estimation. Transp Res Part D 5(5):321–335
4. Åström KJ, Wittenmark B (1997) Computer Controlled Systems, 3rd edn. Prentice Hall, Upper Saddle River
5. Banks JH (2005) Metering ramps to divert traffic around bottlenecks: Some elementary theory. Transp Res Rec 1925:12–19
6. Bell MGH (2000) A game theory approach to measuring the performance reliability of transport networks. Transp Res Part B 34(6):533–545
7. Bellemans T (2003) Traffic Control on Motorways. Ph.D. Thesis, Katholieke Universiteit Leuven, Leuven. ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/bellemans/PhD/03-82.pdf. Accessed on 22 July 2008
8. Bellemans T, De Schutter B, De Moor B (2006) Model predictive control for ramp metering of motorway traffic: A case study. Control Eng Practice 14:757–767
9. Bellemans T, De Schutter B, Wets G, De Moor B (2006) Model predictive control for ramp metering combined with extended kalman filter-based traffic state estimation. In: Proceedings of the 2006 IEEE Intelligent Transportation Systems Conference (ITSC 2006), Toronto, Canada, pp 406–411
10. Bishop R (2005) Intelligent vehicle technology and trends. In: Walker J (ed) Artech House ITS Library. Artech House, Boston
11. Braess D (1968) Über ein Paradoxon aus der Verkehrsplanung. Unternehmensforsch 12:258–268
12. Branston D (1976) Models of single lane time headway distributions. Transp Sci 10:125–148
13. Brownfield J, Graham A, Eveleigh H, Maunsell F, Ward H, Robertson S, Allsop R (2003) Congestion and accident risk. In: Technical report, Road Safety Research Report no. 44. Department for Transport, London
14. Byon Y-J, Shalaby A, Abdulhai B (2006) Travel time collection and traffic monitoring via GPS technologies. In: Proceedings of the 2006 IEEE Intelligent Transportation Systems Conference (ITSC 2006), Toronto, Canada, pp 677–682
15. Camacho EF, Bordons C (1995) Model Predictive Control in the Process Industry. Springer, Berlin
16. Cambridge Systematics, Inc. (2001) Twin Cities Ramp Meter Evaluation – Final Report. Cambridge Systematics, Inc., Oakland. Prepared for the Minnesota Department of Transportation
17. Carsten O, Tate F (2001) Intelligent speed adaptation: The best collision avoidance system? In: Proceedings of the 17th International Technological Conference on the Enhanced Safety of Vehicles, Amsterdam, Netherlands, pp 1–10
18. Carvell JD, Balke K Jr, Ullman J, Fitzpatrick K, Nowlin L, Brehmer C (1997) Freeway management handbook. Technical report. Federal Highway Administration, Department of Transport (FHWA, DOT), Washington DC, Report No. FHWA-SA-97-064
19. Castello P, Coelho C, Ninno ED (1999) Traffic monitoring in motorways by real-time number plate recognition. In: Proceedings of the 10th International Conference on Image Analysis and Processing, Venice, Italy, pp 1129–1131
20. Chaudhary NA, Messer CJ (2000) Ramp metering technology and practice. Technical Report 2121-1. Texas Transportation Institute, The Texas A&M University System, College Station
21. Chen A, Yang H, Lo HK, Tang WH (2002) Capacity reliability of a road network: An assessment methodology and numerical results. Transp Res Part B 36(3):225–252
22. Chen OJ, Hotz AF, Ben-Akiva ME (1997) Development and evaluation of a dynamic ramp metering control model. In: Proceedings of the 8th IFAC/IFIP/IFORS Symposium on Transportation Systems, Chania, Greece, June 1997, pp 1162–1168
23. Chien C-C, Zhang Y, Ioannou PA (1997) Traffic density control for automated highway systems. Automatica 33(7):1273–1285
24. Cools M, Moons E, Wets G (2007) Investigating the effect of holidays on daily traffic counts: A time series approach. Transp Res Record 2019:22–31
25. Cuena J, Hernández J, Molina M (1995) Knowledge-based models for adaptive traffic management systems. Transp Res Part C 3(5):311–337
26. Daganzo CF (1994) The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. Transp Res Part B 28B(4):269–287
27. Daganzo CF (1996) The nature of freeway gridlock and how to prevent it. In: Lesort JB (ed) Proceedings of the 13th International Symposium of Transportation and Traffic Theory, Lyon, France
28. Daganzo CF (1997) Fundamentals of Transportation and Traffic Operations. Pergamon, Kidlington
29. Di Febbraro A, Parisini T, Sacone S, Zoppoli R (2001) Neural approximations for feedback optimal control of freeway systems. IEEE Trans Veh Technol 50(1):302–312
30. Diakaki C, Papageorgiou M, McLean T (2000) Integrated traffic-responsive urban corridor control strategy in Glasgow, Scotland. Transp Res Rec 1727:101–111
31. Elefteriadou L (1997) Freeway merging operations: A probabilistic approach. In: Proceedings of the 8th International Federation of Automatic Control (IFAC) Symposium on Transportation Systems, Chania, Greece, pp 1351–1356
32. Federal Highway Administration, US Department of Transportation (1995) Detection Technology for IVHS – Final Report Addendum. Technical report. Federal Highway Administration, US Department of Transportation, Washington, Contract DTFH61-91-C-00076, Report No. FHWA-RD-96-109
33. Federal Highway Administration, US Department of Trans-

portation (1995) Detection Technology for IVHS – Task L Final Report. Technical report. Federal Highway Administration, US Department of Transportation, Washington, Contract DTFH61-91-C-00076, Report No. FHWA_RD-95-100
34. Fenton RE (1994) IVHS/AHS: Driving into the future. IEEE Control Syst Mag 14(6):13–20
35. Gartner NH, Improta G (eds) (1995) Urban Traffic Networks – Dynamic Flow Modeling and Control. Springer, Berlin
36. Gupta A, Maslanka VJ, Spring GS (1992) Development of prototype knowledge-based expert system for managing congestion on massachusets turnpike. Transp Res Rec (1358): 60–66
37. Hall FL, Agyemang-Duah K (1991) Freeway capacity drop and the definition of capacity. Transp Res Rec (1320):91–98
38. Hardman EJ (1996) Motorway speed control strategies using SISTM. In: Proceedings of the 8th International Conference on Road Traffic Monitoring and Control. IEE Conference Publication, no 422. London, 23–25 April 1996, pp 169–172
39. Hasan M, Jha M, Ben-Akiva M (2002) Evaluation of ramp control algorithms using microscopic traffic simulation. Transp Res Part C 10:229–256
40. Hedrick JK, Tomizuka M, Varaiya P (1994) Control issues in automated highway systems. IEEE Control Syst Mag 14(6):21–32
41. Hegyi A, De Schutter B, Hellendoorn J (2005) Model predictive control for optimal coordination of ramp metering and variable speed limits. Transp Res Part C 13(3):185–209
42. Hegyi A, De Schutter B, Hellendoorn J (2005) Optimal coordination of variable speed limits to suppress shock waves. IEEE Trans Intell Transp Syst 6(1):102–112
43. Hegyi A, Girimonte D, Babuška R, De Schutter B (2006) A comparison of filter configurations for freeway traffic state estimation. In: Proceedings of the International IEEE Conference on Intelligent Transportation Systems 2006, Toronto, Canada, 17–20 September 2006, pp 1029–1034
44. Helbing D (1997) Verkehrsdynamik – Neue physikalische Modellierungskonzepte. Springer, Berlin
45. Hellinga B, Van Aerde M (1995) Examining the potential of using ramp metering as a component of an ATMS. Transp Res Record 1494:75–83
46. Hernández J, Cuena J, Molina M (1999) Real-time traffic management through knowledge-based models: The TRYS approach. In: ERUDIT tutorial on Intelligent Traffic Management Models. Helsinki, Finland, 3 August 1999. http://www.erudit.de/erudit/events/tc-c/tut990803.htm. Accessed on 23 July 2008
47. Hoogendoorn SP (1997) Optimal control of dynamic route information panels. In: Proceedings of the 4th World Congress on Intelligent Transportation Systems, IFAC Transportation Systems, Chania, Greece, pp 399–404
48. Hoogendoorn SP, Bovy PHL (2000) Continuum modelling of multiclass traffic flow. Transp Res Part B 34:123–146
49. Hoogendoorn SP, Bovy PHL (2001) State-of-the-art of vehicular traffic flow modelling. J Syst Control Eng – Proc Inst Mech Eng, Part I 215(14):283–303
50. Hotz A, Much C, Goblick T, Corbet E, Waxman A, Ashok K, Ben-Akiva M, Koutsopoulos H (1992) A distributed, hierarchical system architecture for advanced traffic management systems and advanced traffic information systems. In: Proceedings of the Second Annual Meeting of IVHS AMERICA, Newport Beach
51. http://www.auto21.ca/index.php. Accessed on 23 July 2008
52. http://www.cvisproject.org/. Accessed on 23 July 2008
53. http://www.damas.ift.ulaval.ca/projets/auto21/en/index.html. Accessed on 23 July 2008
54. http://www.path.berkeley.edu/. Accessed on 23 July 2008
55. Jacobson L, Henry K, Mehyar O (1989) Real-time metering algorithm for centralized control. Transp Res Record 1232: 17–26
56. Jiang G, Gang L, Cai Z (2006) Impact of probe vehicles sample size on link travel time estimation. In: Proceedings of the 2006 IEEE Intelligent Transportation Systems Conference (ITSC 2006), Toronto, Canada, 2006, pp 891–896
57. Karimi A, Hegyi A, De Schutter B, Hellendoorn H, Middelham F (2004) Integration of dynamic route guidance and freeway ramp metering using model predictive control. In: Proceedings of the 2004 American Control Conference (ACC 2004), Boston, MA, USA, 30 June–2 July 2004, pp 5533–5538
58. Kell JH, Fullerton IJ, Mills MK (1990) Traffic Detector Handbook, Number FHWA-IP-90-002, 2nd edn. Federal Highway Administration, US Department of Transportation, Washington DC
59. Kerner BS (2002) Empirical features of congested patterns at highway bottlenecks. Transp Res Record (1802):145–154
60. Kerner BS (2004) The Physics of Traffic. Understanding Complex Systems. Springer, Berlin
61. Kerner BS (2007) Control of spatiotemporal congested traffic patterns at highway bottlenecks. IEEE Trans Intell Transp Syst 8(2):308–320
62. Kerner BS (2007) On-ramp metering based on three-phase traffic theory. Traffic Eng Control 48(1):28–35
63. Klein LA, Kelley MR, Mills MK (1997) Evaluation of overhead and in-ground vehicle detector technologies for traffic flow measurement. J Test Evaluation (JTEVA) 25(2):215–224
64. Klein LA, Mills MK, Gibson DRP (2006) Traffic Detector Handbook, Number FHWA-HRT-06-108, 3rd edn. Federal Highway Administration, US Department of Transportation, Washington DC
65. Kotsialos A, Papageorgiou M (2005) A hierarchical ramp metering control scheme for freeway networks. In: Proceedings of the American Control Conference, Portland, OR, USA, 8–10 June 2005, pp 2257–2262
66. Kotsialos A, Papageorgiou M, Middelham F (2001) Optimal coordinated ramp metering with advanced motorway optimal control. In: Proceedings of the 80th Annual Meeting of the Transportation Research Board, vol 3125. Washington DC
67. Kotsialos A, Papageorgiou M, Mangeas M, Haj-Salem H (2002) Coordinated and integrated control of motorway networks via non-linear optimal control. Transp Res C 10(1):65–84
68. Kraan M, van der Zijpp N, Tutert B, Vonk T, van Megen D (1999) Evaluating networkwide effects of variable message signs in the netherlands. Transp Res Record 1689:60–67
69. Kühne RD (1991) Freeway control using a dynamic traffic flow model and vehicle reidentification techniques. Transp Res Record 1320:251–259
70. Lee HY, Lee H-W, Kim D (1999) Empirical phase diagram of traffic flow on highways with on-ramps. In: Helbing D, Herrmann HJ, Schreckenberg M, Wolf DE (eds) Traffic and Granular Flow '99. Springer, Berlin
71. Lenz H, Sollacher R, Lang M (2001) Standing waves and the influence of speed limits. In: Proceedings of the European Control Conference 2001, Porto, Portugal, pp 1228–1232
72. Levinson D, Zhang L (2006) Ramp meters on trial: Evidence

from the twin cities metering holiday. Transp Res Part A 40A:810–828
73. Lewis FL (1992) Applied Optimal Control and Estimation. Prentice Hall, Upper Saddle River
74. Li PY, Horowitz R, Alvarez L, Frankel J, Robertson AM (1995) Traffic flow stabilization. In: Proceedings of the American Control Conference, Seattle, Washington, June 1995, pp 144–149
75. Lighthill MJ, Whitham GB (1955) On kinematic waves, II. A theory of traffic flow on long crowded roads. Proc Royal Soc 229A(1178):317–345
76. Lin P-W, Chang G-L (2007) A generalized model and solution algorithm for estimation of the dynamic freeway origin-destination matrix. Transp Res Part B, 41B:554–572
77. Little R, Rubin D (1987) Handbook of Transportation Science. Wiley, New York
78. Lo HK, Luo XW, Siu BWY (2006) Degradable transport network: Travel time budget of travelers with heterogeneous risk aversion. Transp Res Part B 40(9):792–806
79. Maciejowski JM (2002) Predictive Control with Constraints. Prentice Hall, Harlow
80. Maerivoet S, De Moor B (2005) Cellular automata models of road traffic. Phys Rep 419(1):1–64
81. May AD (1990) Traffic Flow Fundamentals. Prentice Hall, Englewood Cliffs
82. McDonald M, Hounsell NB, Njoze SR (1995) Strategies for route guidance systems taking account of driver response. In: Proceedings of 6th Vehicle Navigation and Information Systems Conference, Seattle, WA, July 1995, pp 328–333
83. Middelham F (1999) A synthetic study of the network effects of ramp metering. Technical report. Transport Research Centre (AVV), Dutch Ministry of Transport, Public Works and Water Management, Rotterdam
84. Middelham F (2003) State of practice in dynamic traffic management in The Netherlands. In: Proceedings of the 10th IFAC Symposium on Control in Transportation Systems (CTS 2003), Tokyo, Japan, August 2003
85. Middleton D, Gopalakrishna D, Raman M (2002) Advances in traffic data collection and management. Technical Report BAT-02-006. Texas Transportation Institute, Cambridge Systematics Inc., Washington. http://www.itsdocs.fhwa.dot.gov/JPODOCS/REPTS_TE/13766.html. Accessed on 23 July 2008
86. Molina M, Hernández J, Cuena J (1998) A structure of problem-solving methods for real-time decision support in traffic control. Int J Human-Computer Stud 49:577–600
87. Papageorgiou M (1980) A new approach to time-of-day control based on a dynamic freeway traffic model. Transp Res Part B, 14B:349–360
88. Papageorgiou M (1983) A hierarchical control system for freeway traffic. Transp Res Part B, 17B(3):251–261
89. Papageorgiou M (1998) Some remarks on macroscopic traffic flow modelling. Transp Res Part A 32(5):323–329
90. Papageorgiou M (ed) (1991) Concise Encyclopedia of Traffic & Transportation Systems. Pergamon
91. Papageorgiou M, Kotsialos A (2000) Freeway ramp metering: An overview. In: Proceedings of the 3rd Annual IEEE Conference on Intelligent Transportation Systems (ITSC 2000), Dearborn, Michigan, USA, October 2000, pp 228–239
92. Papageorgiou M, Kotsialos A (2002) Freeway ramp metering: An overview. IEEE Trans Intell Transp Syst 3(4):271–280
93. Papageorgiou M, Messmer A (1991) Dynamic network traffic assignment and route guidance via feedback regulation. Transp Res Rec 1306:49–58
94. Papageorgiou M, Blosseville J-M, Hadj-Salem H (1990) Modelling and real-time control of traffic flow on the southern part of Boulevard Périphérique in Paris: Part I: Modelling. Transp Res Part A, 24A(5):345–359
95. Papageorgiou M, Blosseville J-M, Hadj-Salem H (1990) Modelling and real-time control of traffic flow on the southern part of Boulevard Périphérique in Paris: Part II: Coordinated on-ramp metering. Transp Res Part A, 24A(5):361–370
96. Papageorgiou M, Hadj-Salem H, Blosseville J-M (1991) ALINEA: A local feedback control law for on-ramp metering. Transp Res Rec (1320):58–64
97. Papageorgiou M, Hadj-Salem H, Middelham F (1997) ALINEA local ramp metering – summary of field results. Transp Res Rec 1603:90–98
98. Papageorgiou M, Wang Y, Kosmatopoulos E, Papamichail I (2007) ALINEA maximizes motorway throughput – An answer to flawed criticism. Traffic Eng Control 48(6):271–276
99. Payne HJ (1971) Models of freeway traffic and control. Simul Counc Proc 1:51–61
100. Rees I (1995) Orbital decongestant. Highways 63(5):17–18
101. Rees T, Harbord B, Dixon C, Abou-Rhame N (2004) Speed-control and incident detection on the M25 controlled motorway (summary of results 1995–2002). Technical Report PPR033, TRL (UK's Transport Research Laboratory), Wokingham
102. Richards PI (1956) Shock waves on the highway. Oper Res 4:42–51
103. Ritchie SG (1990) A knowledge-based decision support architecture for advanced traffic management. Transp Res Part A, 24A(1):27–37
104. Schik P (2003) Einfluss von Streckenbeeinflussungsanlagen auf die Kapazität von Autobahnabschnitten sowie die Stabilität des Verkehrsflusses. Ph.D. Thesis, Universität Stuttgart
105. Shladover SE, Desoer CA, Hedrick JK, Tomizuka M, Walrand J, Zhang WB, McMahon DH, Peng H, Sheikholesham S, McKeown N (1991) Automatic vehicle control developments in the PATH program. IEEE Trans Veh Technol 40(1):114–130
106. Sisiopiku VP (2001) Variable speed control: Technologies and practice. In: Proceedings of the 11th Annual Meeting of ITS America, Miami, pp 1–11
107. Slotine JE, Li W (1991) Applied Nonlinear Control. Prentice Hall, New Jersey
108. Smith BL, Zhang H, Fontaine MD, Green MW (2004) Wireless location technology based traffic monitoring: Critical assessment and evaluation of an early-generation system. J Transp Eng 130(5):576–584
109. Smulders S (1990) Control of freeway traffic flow by variable speed signs. Transp Res Part B, 24B(2):111–132
110. Smulders SA, Helleman DE (1998) Variable speed control: State-of-the-art and synthesis. In: Road Transport Information and Control, number 454 in Conference Publication, IEE, 21–23 April 1998, pp 155–159
111. Soriguera F, Thorson L, Robuste F (2007) Travel time measurement using toll infrastructure. In: Proceedings of the 86th Annual Meeting of the Transportation Research Board, CDROM paper 07-1389.pdf
112. Taylor CJ, Young PC, Chotai A, Whittaker J (1998) Nonminimal state space approach to multivariable ramp metering control of motorway bottlenecks. IEE Proc – Control Theory Appl 146(6):568–574

113. Taylor MAP (1999) Dense network traffic models, travel time reliability and traffic management. I: General introduction. J Adv Transp 33(2):218–233
114. Taylor MAP (1999) Dense network traffic models, travel time reliability and traffic management. II: Application to network reliability. J Adv Transp 33(2):235–251
115. Taylor MAP, Woolley JE, Zito R (2000) Integration of the global positioning system and geographical information systems for traffic congestion studies. Transp Res Part C, 8C:257–285
116. Transportation Research Board (1998) Managing speed: Review of current practice for setting and enforcing speed limits. In: TRB Special Report 254, Transportation Research Board, Committee for Guidance on Setting and Enforcing Speed Limits. National Academy Press, Washington
117. Treiber M, Hennecke A, Helbing D (2000) Congested traffic states in empirical observations and microscopic simulation. Phys Rev E 62:1805–1824
118. Tsugawa S, Kato S, Tokuda K, Matsui T, Fujii H (2000) An architecture for cooperative driving of automated vehicles. In: Proceedings of the IEEE Intelligent Transportation Symposium, Dearborn, MI, USA 2000, pp 422–427
119. Vahidi A, Eskandarian A (2003) Research advances in intelligent collision avoidance and adaptive cruise control. IEEE Trans Intell Transp Syst 4(3):143–153
120. van Arem B, van Driel CJG, Visser R (2006) The impacts of cooperative adaptive cruise control on traffic-flow characteristics. IEEE Trans Intell Transp Syst (4):429–436
121. van den Berg M, Bellemans T, De Schutter B, De Moor B, Hellendoorn J (2005) Control of traffic with anticipative ramp metering. In: Proceedings of the 84th Annual Meeting of the Transportation Research Board, Washington DC, 2005 CDROM paper 05-0252, pp 166–174
122. van den Hoogen E, Smulders S (1994) Control by variable speed signs: Results of the Dutch experiment. In: Proceedings of the 7th International Conference on Road Traffic Monitoring and Control, IEE Conference Publication No. 391, London, England, 26–28 April 1994, pp 145–149
123. Varaiya P (1993) Smart cars on smart roads: Problems of control. IEEE Trans Autom Control 38(2):195–207
124. Wang Y, Papageorgiou M (2005) Real-time freeway traffic state estimation based on extended Kalman filter: A general approach. Transp Res Part B 39(2):141–167
125. Wang Y, Papageorgiou M, Messmer A (2003) A predictive feedback routing control strategy for freeway network traffic. In: Proceedings of the 82nd Annual Meeting of the Transportation Research Board, Washington DC, January 2003
126. Wang Y, Papageorgiou M, Messmer A (2006) RENAISSANCE A unified macroscopic model-based approach to real-time freeway network traffic surveillance. Transp Res Part C, 14C:190–212
127. Wattleworth JA (1965) Peak-period analysis and control of a freeway system. Highw Res Rec 157:1–21
128. Wilkie JK (1997) Using variable speed limit signs to mitigate speed differentials upstream of reduced flow locations. Technical report. Department of Civil Engineering, Texas A& M University, College Station, Prepared for CVEN 677 Advanced Surface Transportation Systems, Report No. SWUTC/97/72840-00003-2
129. Williams B (1996) Highway control. IEE Rev 42(5):191–194
130. Wolf DE (1999) Cellular automata for traffic simulations. Physica A 263:438–451
131. Zhang H, Ritchie SG (1994) Real-time decision-support system for freeway management and control. J Comput Civ Eng 8(1):35–51
132. Zhang L (2007) Traffic diversion effect of ramp metering at the individual and system levels. In: Proceedings of the 86th Annual Meeting of the Transportation Research Board, Washington DC, January 2007, CDROM paper 07-2087

Books and Reviews

We refer the interested reader to the following references in the various fields that have been discussed in this chapter:

Control: General introduction [4], optimal control [73], model predictive control [15,79], nonlinear control [107],

Traffic flow modeling: General overviews [48,49], cell transmission model [26], Kerner's three-phase theory [60], microscopic simulation models [2], cellular automata [80],

Ramp metering: Overviews of ramp metering strategies [18,91], field test and simulation studies [39],

Speed limit systems: Overviews of practical speed limit systems [106,128],

Intelligent vehicles: Overview [10],

Sensor technologies: Overviews [32,33,63,64]

Functional Genomics for Characterization of Genome Sequences

Adam M. Deutschbauer[1], Lars M. Steinmetz[2]
[1] Lawrence Berkeley National Laboratory, Berkeley, USA
[2] European Molecular Biology Laboratory, Heidelberg, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Computational and Comparative Genomics
Transcription
Genetics Analysis
Functional Genomics and Complex Traits
Future Directions
Acknowledgments
Bibliography

Glossary

Tiling microarray A microarray containing probes representing the entire genome in an unbiased, uniform pattern. The resolution of a tiling microarray is determined by the length of probes and the amount of overlap between adjacent probes to the genome.

Non-coding RNA (ncRNA) RNA that is not translated into protein. The term ncRNA encompasses different classes of RNA including small nucleolar RNA (snoRNA), small inhibitory RNA (siRNA), ribosomal RNA (rRNA), transfer RNA (tRNA) and microRNA (miRNA).

Pyrosequencing A DNA sequencing method that utilizes enzymatic reactions and light detection as a readout for base incorporation.

Complex trait A heritable trait conditioned by multiple genetic and/or environmental factors.

RNA interference (RNAi) The process by which double strand RNA (dsRNA) targets and degrades a complementary transcript.

Epigenetics A class of heritable traits that are stable over multiple cell divisions but are not associated with DNA sequence changes.

Phylogenetics The study of evolutionary relationships among organisms.

ChIP-chip Shorthand terminology for chromatin-immunoprecipitation with microarray detection; a technique to identify regions of DNA bound by a protein of interest. Following immunoprecipitation of a protein-DNA hybrid, the DNA is released from the protein, labeled, and detected on a DNA microarray.

Genotyping The process of measuring the genetic differences (genotype) between individuals within a population.

Linkage disequilibrium The non-random association of two or more alleles caused primarily by population structure and the absence of recombination in a region of the chromosome.

Epistasis A type of genetic interaction between two or more loci in which the phenotype of the mutant combination deviates from the expectation based on the phenotype of the single mutants.

Chromatin The tightly bundled complex of DNA and nucleosomes (made up of histone proteins) that packages the nuclear genome in eukaryotic cells.

Definition of the Subject

Complete genome sequences have been determined for hundreds of organisms ranging in complexity from bacteria to human. However, sequence data alone is currently of limited use for identifying the functional elements of a genome and for elucidating how these elements interact to control physiological processes.

The field of functional genomics aims to meet these challenges using the sequence data as a blueprint. In a broad sense, functional genomics is defined as the large-scale experimental study of gene function and interactions. In this article, we introduce the reader to the techniques and challenges in functional genomics in the context of an ever more common case scenario: given a complete genome sequence, how does one figure out what the sequence means? By highlighting the relevant literature, we go through the current state of the art as to what one would do to arrive at function given a genome sequence. We focus primarily on nucleic acid based technologies though connect in the Future Directions to similar analyzes at the protein and metabolite level.

Introduction

The field of functional genomics is relatively new compared to other molecular biology disciplines such as biochemistry, classical genetics, and cell biology. The technological advances and milestones that laid the groundwork for the completion of genome sequences have occurred primarily over the last 25 years (Fig. 1). In particular, advances related to DNA sequencing deserve special attention as sequence data is the backbone of functional genomics and the ability to generate large amounts of this data has paved the way for many of the techniques described in this article. Foremost among these advances were the development of chain termination based DNA sequencing by Sanger and colleagues [99] and the automation of DNA sequencing by Hood and coworkers [109]. The ability to automate sequencing ushered in a new era of biology, one in which the completion of genomes was technically feasible.

The drive to sequence complete genomes was initiated in the late 1980's with the establishment of the Human Genome Project. However, given that the size of the human genome is immense at 3 billion nucleotides and the optimal techniques for completing a genome sequence were unknown at the time, it was necessary to develop whole genome sequencing in model organisms. In 1995, the first genome of a free living organism, the bacterium *Haemophilus influenzae*, was published [35]. In subsequent years, the genomes of the model genetic organisms *Saccharomyces cerevisiae* (yeast) [45], *Caenorhabditis elegans* (worm) (*C. elegans* Sequencing Consortium 1998), *Drosophila melanogaster* (fruit fly) [2], and mouse [123] were completed by international consortia. Paramount to the success of these projects was the development of automation and robotics for handling large numbers of reactions and software for the analysis and assembly of vast amounts of sequence data. The lessons learned by sequencing model organisms laid the technological foundation for the completion of the draft human sequence

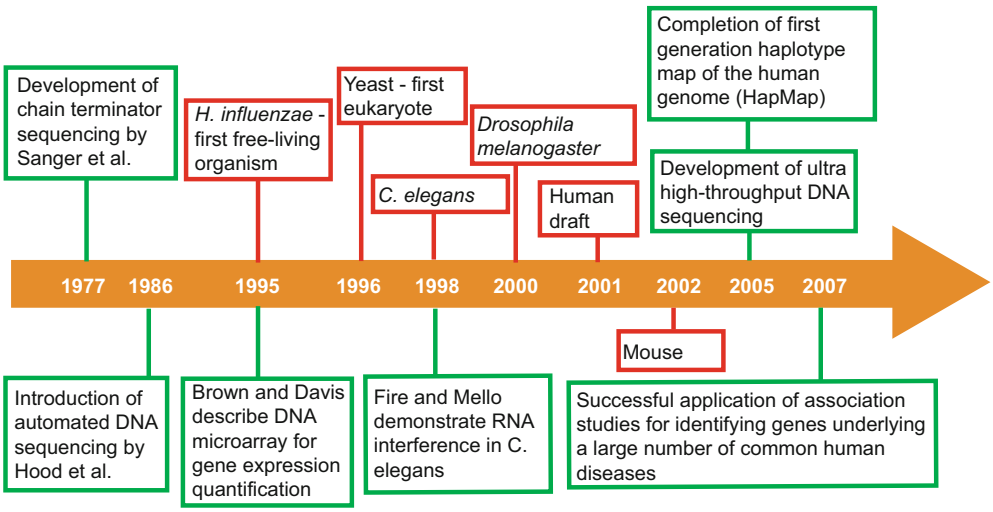

Functional Genomics for Characterization of Genome Sequences, Figure 1
Landmarks in functional genomics. The timeline indicates select technological breakthroughs (in *green boxes*) in functional genomics and highlights the genome completion date of important model organisms (in *red boxes*). References are in parentheses

in 2001 [73]. In some respects, the finishing of the human genome marked the end of the first generation of DNA sequencing. Today, advances in sequencing technology promise a second generation of DNA sequencing, one in which the majority of the earth's living (and some extinct) organisms will have complete genome sequences in the near future (see Future Directions).

The initial analysis of the early genomes was greeted with two harsh realizations. First, current algorithms for gene finding, partly based on the availability of many complete genome sequences, were not yet available. The result was that many predicted genes were later demonstrated not to be genes and conversely, many currently predicted genes were missed by these early, somewhat arbitrary gene prediction methods. Second, the majority of predicted genes identified from the sequence alone had no known function. For example, over 60% of the ~6000 yeast genes identified from the genome sequence were uncharacterized in 1996 despite 20 years of active molecular genetics research [45]. To fill this void, higher-throughput, so called "post-genomic" techniques aimed at the elucidation of gene function at a genome-scale were developed. The subsequent application of these post-genomic techniques to model organisms marked the beginning of modern functional genomics. Today, functional genomic tools such as microarrays for gene expression [103] and whole-genome reverse genetics have increased the percentage of characterized yeast genes to over 80% [92].

Here we describe some major topics in functional genomics in the context of how one would analyze a newly completed genome sequence today. We begin with a description of *in silico* comparative genomic approaches for finding genes, assigning putative gene function, and inferring regulatory elements. We then describe microarray-based gene expression profiling and high-throughput mutant analysis before ending with a discussion of intraspecific genetic variation. In each instance we highlight recent literature to illustrate the utility and shortcomings of different techniques. To avoid organism-bias and to reflect the universal challenges associated with assigning gene function based on DNA sequence data, we discuss literature from a range of organisms, spanning bacteria to human. Finally, in the Future Directions, we link our discussion of functional genomics to future challenges including data integration, genetic interactions, and the impact of technology.

Computational and Comparative Genomics

A newly completed genome sequence is simply a file containing millions (or even billions) of A, T, C, and G nucleotides. Taken in isolation, the task of identifying functional elements from this file of nucleotides can seem like an insurmountable problem. While the field of functional genomics aims to uncover the functionality of the genome through experiment, one can learn a great deal about a new genome sequence using purely computational tools. Inherent in the development, application, and success of computational genomic analyzes are the wealth of already completed genome sequences with

which one can compare a genome sequence for similarities and differences (http://www.ncbi.nlm.nih.gov/sites/entrez?db=Genome). This field of research, comparative genomics, has two broad aims: to identify functional elements such as genes and regulatory sequences and to infer the evolutionary mechanisms that gave rise to different species [51]. Here we concentrate on computational and comparative genomics in the context of finding functional DNA elements and direct the reader to additional reading on the use of comparative genomics to infer evolutionary relationships in prokaryotes [1], yeast [26], *Drosophila* [17], and human [14].

Concepts

The premise of comparative genomics is relatively straightforward. DNA sequences that are shared between two distinct species are likely to be functional. From an evolutionary perspective, these are genomic regions that are conserved since the species' last common ancestor and represent genomic regions that are more intolerant to mutation. Conversely, since mutations are assumed to accumulate in genomic regions that are under no functional constraint, sequences that are significantly diverged since the last common ancestor represent nonfunctional regions. In practice, identifying conserved (functional) versus nonconserved (nonfunctional) genomic regions is a steep challenge due to a myriad of complicating factors including non-uniform mutation rates both within [125] and between [72] genomes and the fact that the premise is not always true; even highly diverged regions between species can retain similar functions [34]. Furthermore, the success of a comparative genomic analysis is dependent on the marriage of the right biological question and the use of an appropriate phylogenetic range of organisms. For instance, one would have a difficult time identifying conserved gene regulatory elements by comparing worm and human orthologous genes. Worm and human diverged too long ago to make such an analysis fruitful. Conversely, in order to detect positively selected genes in the human lineage a comparison between human and other closely related primate genomes is most appropriate (for example [29]).

Applications

The protein-coding gene is a principle functional unit of the genome. Therefore, computationally predicting genes is usually the first task in any genome-level analysis. While software exists to predict genes using single genomes and gene structure modeling, usually a comparative approach to gene finding is utilized based on the knowledge that genes are more likely to be conserved in related species [82] . In the yeast *Saccharomyces cerevisiae*, the initial prediction of the number of genes was over 6000 [45]. By comparing the genome sequence of *S. cerevisiae* to that of other related *Saccharomyces* species, an additional ∼50 small protein-coding genes were predicted to exist based on high sequence conservation while ∼500 of the originally annotated genes were deemed not to be protein-coding genes based on frameshift and nonsense mutations in the orthologs of closely related species [18,68]. Similar approaches are being applied in mammals, where high-genome complexity and extensive gene splicing make the task of gene identification challenging. In one example, gene prediction software [70] based on the sequence conservation between genes was used to predict over 1000 genes in mouse and human [48].

There is growing appreciation for the important regulatory and structural functions of non-coding RNA (ncRNA). Unfortunately, the ability to computationally identify ncRNA is complicated by the absence of defined sequence signatures similar to what is available for protein-coding genes [27]. However, if ncRNA are functionally relevant, then they should be under evolutionary constraint and conserved at least structurally across related species. Using this rationale, programs to detect ncRNA have been developed using comparisons across species [95]. For instance, an algorithm that combines RNA structure prediction and comparative sequence analysis identified both known and novel conserved ncRNA when applied to the genomes of human, mouse, rat, zebrafish, and pufferfish [122].

Genes by themselves are static stretches of DNA which encode information on the chromosome. It is the regulation of these genes that guides development, maintains homeostasis, and contributes significantly to the diversity in nature. Therefore, the identification of the non-coding sequences that transcription factors bind to regulate gene expression is an active area of research. However, the challenges associated with computationally identifying functional regulatory elements are substantial. Most importantly, the DNA sequences that regulate transcription are often small and degenerate. Therefore, the identification of these elements is often hampered by a large number of false positive results. Even more critical than for identifying protein coding regions, identifying regulatory elements requires comparative genomics across an appropriate range of phylogenetic distances. For example, a search for common regulatory motifs in humans via phylogenetic comparisons with orthologous sequences in mouse, rat, and dog identified 105 previously unknown promoter regulatory motifs and 106 motifs present in 3' untranslated re-

gions [126]. As more genome sequences become available, the 'phylogenetic footprinting' approach for identifying regulatory elements should increase in power. In a broad sense, the promise of all comparative genomics techniques as described here is dependent on more genome sequences than are currently available. This is a principle reason why genome sequencing will stay at the forefront of genetics research in the near future.

Summary

What have we learned about our genome using computational and comparative genomics? Without setting foot in the laboratory, we have successfully predicted many protein-coding genes, ncRNAs, and regulatory elements, via comparison with previously sequenced genomes. Furthermore, using sequence similarity tools such as BLAST, we could assign putative biochemical and biological functions to many genes by homology to genes of known function. However, despite this progress, the majority of genes in most newly sequenced species remain partially or completely uncharacterized at a functional level. It must be emphasized that computational predictions are dependent on the availability of data and thus can have limited power (for example, see [28] for review of computational prediction versus experimental validation of transcription factor binding sites). Conservation of DNA across an evolutionary distance does not necessarily mean that the sequences are functional [86] and conversely, the lack of sequence conservation does not necessarily imply a lack of function [34]. Therefore, computational predictions must be interpreted carefully. Finally, simply identifying the locations of putative functional elements on the chromosome provides little information on the dynamic functionality of the genome. Experimental functional genomic techniques aim to fill these voids. In the remaining sections we will describe analyzes at different levels with the goal of enriching this view of genome function.

Transcription

Cells regulate the transcription of RNA from genes to control biological processes. Therefore, a thorough understanding of transcriptional architecture is a chief requirement for the elucidation of organism physiology. Taken together, the transcription of a single gene is a highly complicated process involving genomic DNA, transcription factors, RNA polymerases, and epigenetic modifications such as chromatin dynamics. The functional genomics of transcription aims to analyze each of these phenomena in the context of the entire genome. Here we discuss whole-genome techniques to dissect transcription and their applications to transcript identification, RNA quantification, gene regulation by transcription factor binding localization, and epigenetics.

Before entering into the functional genomics of gene transcription, it is necessary to briefly introduce the primary technology used in the majority of these studies: the microarray. The microarray is a surface (typically glass) with thousands or even millions of unique immobilized molecules (most commonly DNA) systematically arrayed as 'probes'. DNA microarrays come in two primary types: 'spotted' arrays printed with PCR products (or long oligonucleotides) and 'oligo' arrays with short (usually 25 to 70 nucleotide) *in situ* synthesized probes. Regardless of their construction, one hybridizes a complex mixture of usually fluorescently labeled target nucleic acid to the microarray. The abundance of each nucleic acid species in the target sample is then monitored by the fluorescent intensity of the corresponding probe on the microarray.

Transcript Mapping Using Tiling Microarrays

The identification of transcripts is a logical starting point for any global study of transcription. Using unbiased whole-genome tiling arrays that interrogate every base of the genome, one can get a global view of the transcriptional architecture of an organism [7]. The application of this technique to various organisms including *Arabidopsis* [127], yeast [22], and human [30] has revealed a number of fundamental insights both at the global and single-gene level. At the single-gene level, genome tiling arrays provide a high-resolution view of the expression architecture of individual genes (see Fig. 2 for examples). In particular, the start/stop sites and exon/intron boundaries of individual transcripts can be localized to within a handful of base pairs. Importantly, these insights enable the reannotation of the original computational predictions of gene structure [22,132] and demonstrate how experimental methods can be used to systematically update original genome-wide predictions based on computation.

More importantly than the insights into single genes, tiling arrays are revolutionizing our global perspective on transcription [67]. First, there exists considerably more transcription than had previously been thought. Transcriptional elements (e.g. genes) are not arranged as beads-on-a-string along the genome. Strikingly, it appears that nearly every base in the yeast genome is transcribed, often from both strands [22]. In humans, 93% of genomic regions examined in detail (encompassing 1% of the total human genome) appear to be transcribed in an least one condition or tissue [30]. Second, in all organisms examined with tiling arrays to date, the vast major-

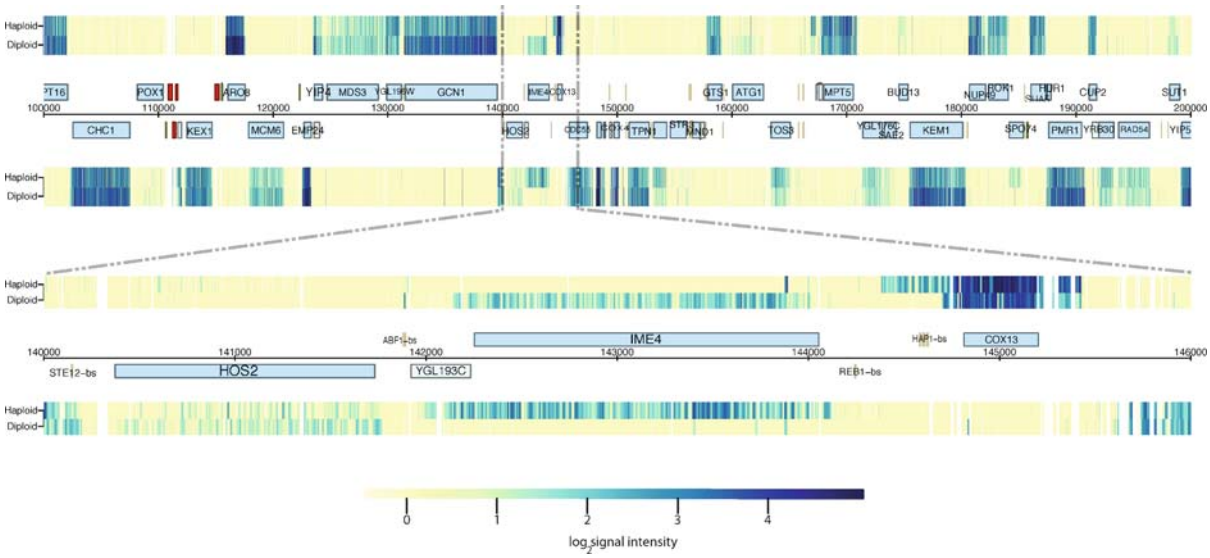

Functional Genomics for Characterization of Genome Sequences, Figure 2

Transcript mapping using tiling microarrays. The top panel shows data for a 100 kb region from a diploid and a haploid yeast strain. Probes on the array are 25 bases long and are tiled every 8 bases on each strand of the genome, the forward strand shown on top and the reverse strand on the bottom. Gene features, such as open reading frames, as annotated in the yeast genome database are shown by boxes. Expression signal is plotted with a color gradient for probes across the region and allows expressed transcripts to be detected. The bottom panel shows a close-up of the region around *IME4*. The Ime4 protein is expressed only in diploids cells. In haploids, an *IME4* antisense transcript inhibits the expression of Ime4 protein by transcriptional interference [56]

ity of this novel transcription does not appear to encode protein, rather the transcription is likely to be ncRNA. Often this ncRNA is in the form of antisense transcription that either physically overlaps the transcription of genes on the opposite strand (the sense strand) forming *cis*-encoded sense/antisense transcript pairs, or matches to sequence transcribed elsewhere in the genome forming *trans*-encoded sense/antisense pairs . Lastly, in humans, primary transcripts appear to be extensively overlapping (or 'interleaved') which may provide a mechanism for increasing the number of possible human proteins [30]. Overall, these findings imply the potential inhibitory function of antisense transcription, the regulatory roles of small ncRNA, and challenge the view that each human gene is its own modular entity.

There are several clues that suggest that natural antisense transcription (in *cis* or in *trans*) is an important mode of gene regulation. First, genome-wide assessments demonstrate that antisense transcription is prevalent in most (if not all) organisms. Second, complementary sense and antisense transcripts are often conserved across evolution [16]. Given the appreciation for the extent of antisense transcription, a challenge today is to decipher the mechanism(s) through which antisense transcripts contribute to gene regulation [74]. One possibility is that that two transcripts encoded in *cis* with opposite orientations collide during RNA polymerase-based elongation [94]. This model provides a mechanism to modulate gene expression from the sense strand by altering the relative strengths of the two promoters. Such a transcriptional interference mechanism has recently been proposed to contribute to the entry of yeast into sporulation [56] (See Fig. 2 for details). Alternatively, antisense transcription might exert its effect through double-stranded RNA pathways such as RNAi [9]. Often these short *trans* encoded antisense transcripts are found to be expressed in alternate tissues and developmental stages than their sense targets, suggesting that some miRNAs may help maintain and define cell types by dampening the expression of unwanted transcripts [13].

The unbiased view of transcription derived from tiling microarrays has profound implications. First, the notion that the non protein-coding portion of genomes is 'junk' requires reassessment. The extent of (and presumably regulation of) transcription in these regions suggests that they serve a functional role. Second, many mechanisms of gene regulation are likely to exist. In addition to the antisense transcription described above, other modes of regulation including complex intercalated transcription, chimeric transcripts, and the action of microRNAs on

chromatin for example underscore the complexity of transcriptional architecture. Lastly, it has been proposed that the aberrant lack of correlation between organism complexity and gene number can be partially reconciled with an increase in ncRNA regulation in more complex organisms [114]. This view contends that an expansion of *trans*-acting regulatory ncRNA (together with an increase in *cis*-acting regulatory DNA) provides a substantial increase in genome information content.

Gene Expression (mRNA Quantification)

The most common functional genomic experiment is the (relative) quantification of mRNA levels using microarrays, typically referred to as gene expression or 'transcriptome' profiling. The reasons for this are twofold. The first is that this type of experiment is relatively straightforward to perform. All one needs is a microarray (many of which can be purchased commercially), the infrastructure to process the microarray, and an RNA sample. The second reason why gene expression experiments are popular is that they can provide substantial insights into transcription and its regulation, particularly when a large number of experiments are performed (the rationale is identical to the increasing power of comparative genomics with more complete genome sequences).

What can gene expression experiments tell us about the function of the genome? First, gene expression experiments can be used to infer the functionality of unknown genes under the hypothesis that genes with similar expression patterns are more likely to have similar functions [11]. Likewise, mutants with similar expression pattern are likely mutant for genes of similar function [57]. Second, gene expression data are useful for identifying regulatory sites that control transcription. This is accomplished by a computational analysis of the genomic region surrounding a set of genes identified as co-regulated. Lastly, gene expression experiments can provide insight into the gene regulatory network of an organism. For example, an expression compendium of over 100 diverse conditions in *E. coli* was sufficient to identify 1079 regulatory interactions including a novel relationship between iron transport and central carbon metabolism [31].

Gene Regulation

Gene expression data and comparative genomic approaches are insufficient in isolation to identify all of the genomic regulatory elements that control transcription. To identify these elements, multiple experiment-based approaches have been developed. The in vivo ChIP-chip method involves the cross-linking of a transcription factor of interest (often epitope-tagged) to the DNA. Following immunoprecipitation of the transcription factor-DNA hybrid, the DNA is released from the protein, labeled, and detected on a DNA microarray. The consensus sequence bound by the transcription factor is identified by a comparative sequence analysis of all 'hits' on the microarray. In one example, ChIP-chip was used to associate the binding of the Rap1 transcription factor to 294 loci in the yeast genome, the majority of which were intergenic regions with a high probability of containing promoters [78]. On a more global level, Harbison et al. analyzed the genomic binding of 203 yeast transcription factors in rich media and across a range of diverse conditions [50]. The results of this work provided the first comprehensive regulatory map for a eukaryotic genome and demonstrated the dynamic nature of transcriptional regulation in the response to environmental change.

There are a number of design and analysis considerations that go into a successful ChIP-chip experiment [12]. One parameter of note is the density of the microarray used in the experiment. In general, a tiling microarray is superior to a conventional microarray because all genome coordinates are covered, and multiple, overlapping probes can interrogate a single genomic region. The principle advantage is that an increase in microarray signal across a number of overlapping probes provides both increased statistical significance and an increased ability to identify the actual binding sequences. However, even with tiling microarrays it is often difficult to determine the cutoff significance value for which microarray probes are enriched in the experiment (and hence, which genomic locations the transcription factor binds to). Another disadvantage of ChIP-chip is that tiling microarrays for large genomes are currently expensive (because multiple microarrays are necessary to adequately tile large genomes). Therefore, other approaches for the detection of DNA from chromatin immunoprecipitation experiments have been explored. One promising avenue is the use of massively parallel sequencing technology (see Future Directions) for the detection of DNA (ChIP-seq). ChIP-seq requires no microarray and provides a direct count of DNA sequences contained in the immunoprecipitated sample. Results obtained using human STAT1 [96] and NRSF [64] indicate that ChIP-seq can detect weakly bound sites that are likely to be missed by microarray based methods.

A promoter array is an in vitro method for identifying transcription factor binding sites [5]. Here, an oligonucleotide microarray with double stranded probes is directly bound by a labeled target protein. A primary advantage of a promoter array over ChIP-chip or ChIP-seq is that all combinations of short sequences (for example,

10mers in [5]) can be assayed for binding associations of different affinities. Additionally, the actual identification of the binding site is more straightforward than the computational analysis required for ChIP-chip or ChIP-seq experiments. Conversely, as an entirely in vitro assay, some positive promoter array results may be physiologically irrelevant.

Epigenetics

In the context of transcription, the regulation of chromatin represents the primary mode of epigenetic inheritance. In eukaryotic genomes, DNA ins packaged into chromatin, which consists of DNA wound around protein (histone) complexes known as nucleosomes. The positioning and regulation (via the modification of their histone subunits) of nucleosomes along the chromosome play a crucial role in gene expression, by 'exposing' regions of the genome for transcriptional activity. Given its importance to gene regulation, the localization and dynamics of nucleosome positioning are an active area of investigation. An early genome-level investigation of nucleosome positions showed that active promoters and their associated transcription factor binding sites are largely devoid of nucleosomes [129]. A second study demonstrated that, during the cell cycle, the occupancy of nucleosomes is reduced at the specific time when the gene is transcribed [54]. These results indicate that chromosome accessibility via nucleosome positioning is an important mechanism for regulating the expression state of genes. Using nucleosome localization data, Segal and colleagues developed a nucleosome-DNA interaction model suggesting that nucleosome positioning is an intrinsic feature of eukaryotic genomes [106]. Interestingly, this model takes advantage of the tendency of DNA to bend sharply around the nucleosome at periodic intervals. More recently, tiling micoarrays were used to construct a comprehensive, high-resolution "atlas" of over 70,000 nucleosome sites in the yeast genome [77].

In addition to the physical localization of nucleosomes, the modification state of the individual histone subunits is thought to be linked to the expression state of the cell. To test the hypothesis that histone methylation marks the expression state of differentiated cells in mammals, Mikkelsen et al. analyzed the genomic pattern of histone methylation in mouse pluripotent cells (capable of differentiating into multiple lineages) and terminal differentiated cells using ChIP-seq [83]. The authors demonstrate that specific histone methylations, such as those on lysine positions 4 and 27 of histone subunit H3, discriminate the expression state of genes and hence provide insight into the lineage potential of different cell types. It is anticipated that future genome-wide epigenetic investigations of the kind described above will provide additional insight into the importance of epigenetics in gene regulation and disease.

Summary

Given that microarray design and construction requires only DNA sequence data, we can learn a lot about the architecture and regulation of transcription of our genome using the approaches described above. However, there are limitations to using transcription data alone to dissect the functionality of an entire genome. Most importantly, neither the presence of transcription, the expression level of genes, nor the identification of regulatory elements tell us much about what biological processes genes are actually involved in. For instance, the majority of genes that are differentially expressed under stress conditions in yeast display no obvious phenotype under the same condition when disrupted [42]. Taking into consideration these limitations, additional genome-wide approaches are necessary to more directly elucidate gene function.

Genetics Analysis

The field of genetics is largely the analysis of mutations. By studying the phenotypic effects of mutations, one can gain insight into the biological processes to which a gene contributes. Classically, in the absence of complete genome sequences, geneticists had to map interesting mutations to the causative gene. While this 'forward' genetic approach is still relevant, today more often mutagenesis approaches are being applied that take advantage of the complete genome sequence of an organism. In this article, we will primarily focus on these 'reverse' genetic techniques.

The mutations discussed below come in two main flavors: heritable and transient. Heritable mutations are permanent changes to the DNA sequence and are induced using techniques such as targeted gene 'knockouts' by homologous recombination and random mutagenesis by a mobile element (such as a transposon). Transient techniques are largely based on the ability of certain molecules to inhibit the expression of specific genes and consequently 'turning them down' for as long as the molecules are active. The most important of these techniques is RNA interference (RNAi). In RNAi, a short double strand RNA molecule degrades its specific complementary transcript. Therefore, RNAi mimics a loss-of-function mutation. The term 'transient' here can be misleading as the effects of the mutagenesis can last for multiple cell divisions. However, because the genomic DNA sequence is left untouched, the

effects of RNAi are not permanent. Additionally, transient perturbations can be introduced via small molecule perturbations, a field referred to as "chemical genomics". In yeast, chemical genomics has proved successful in identifying the targets of small molecules and the mode of action of small molecules[43,90].

Systematic Analysis of Single-Gene Mutations

Essential genes are a special class whose disruption results in a loss of organismal viability. In a number of microorganisms, comprehensive surveys have been undertaken to determine the catalog of essential genes in order to determine which core functions the cell needs to survive [44] and to identify potential drug targets [98]. Approximately 300 genes are essential in the genomes of the bacteria *E. coli* [3] and *B. subtilis* [69] (less than 10% of the total gene complement in each organism). Of these ~300 genes, about one-half are orthologous (derived from a common ancestor) between these distantly related bacteria demonstrating the fundamental importance of these genes across evolution. In yeast, ~18% of the genome is essential [42]. Reflecting the critical role of these genes in fundamental eukaryotic cellular processes, the essential gene class is more likely to have homologs in other organisms than the nonessential gene class [42].

As described above, only a small fraction of the total genome complement is essential for viability. For the nonessential gene class, the systematic determination of mutant phenotypes provides insight into all of the biological processes intrinsic to the organism, not just viability. For a large number of microorganisms, genome-wide approaches have been applied to determine which genes are 'conditionally essential' (required and presumably functional under specific conditions). In bacteria, techniques for the parallel analysis of mutant pools such as signature tagged mutagenesis [52] and transposon site hybridization with microarrays [100] are successful at identifying bacterial genes required for pathogenesis in a host. The conditionally essential gene sets identified from these studies represent attractive targets for novel antibiotics.

In 2002, an international consortium of yeast researchers completed a systematic knockout library of the entire yeast genome [42]. An important aspect to this project was the incorporation of unique molecular barcodes into each deletion strain that enabled pooling and analysis of the entire collection in a single microarray hybridization. Since its completion, the yeast knockout library has been profiled in hundreds of conditions both in pools with microarrays and as individual strains using miniaturized growth assays on plates and in liquid media [104]. The results of these studies have provided tantalizing clues into the putative function of hundreds of uncharacterized yeast genes. For example, an unbiased survey of genes required for sporulation identified virtually all known regulators of the process in addition to hundreds of genes previously unknown to be involved in sporulation [25]. In addition, the evolutionary conservation of many genes permits the reinterpretation of the yeast results across all eukaryotes, including human. For instance, a comprehensive screen for yeast mitochondrial genes was used to identify candidate human disease genes based on sequence homology and disease linkage map intervals [111].

The systematic examination of mutant phenotypes in multicellular eukaryotes is more difficult and costly than that in microorganisms. Using RNAi, however, an increasing number of genome-wide mutagenesis studies have demonstrated the utility and power of reverse genetics in worm [65], fly [37], and human [6,88]. The worm, where the capability of dsRNA to silence complementary transcripts was first demonstrated in 1998 [33], led the genome-wide RNAi initiative due to the early completion of its genome sequence and the ease with which dsRNA can be delivered to the whole organism (by feeding). By targeting 86% of the ~19,000 worm genes by RNAi, Kamath and colleagues identified mutant phenotypes for 1722 genes, two-thirds of which had not been previously associated with a phenotype [65].

In contrast to worm RNAi which is performed against the entire organism, genome-wide RNAi efforts in Drosophila and human have been applied to cultured cells, primarily with well known signaling pathways as the focus. A number of lessons emerge from these studies. First, the number of unknown regulators of 'well' characterized pathways is large. For example, Friedman and Perrimon identified hundreds of potential regulators of the receptor tyrosine kinase/extracellular regulated kinase (RTK/ERK) pathway using a quantitative RNAi screen in Drosophila [37]. Second, the limitations of RNAi targeting efficiency in humans can be largely reduced by using multiple, unique inhibitory RNA molecules against a single transcript [6]. Lastly, molecular barcodes incorporated into the human RNAi libraries enable the pooling and parallel analysis of thousands of cultured cell lines using a barcode microarray [6,88]. Importantly, the use of molecular barcodes will facilitate the systematic profiling of the human RNAi collection across a wide range of cell types, cellular states, and drug conditions using methods analogous to the yeast deletion system.

Given the relatively recent awareness of the extent and importance of ncRNA, the studies described above all fo-

cus on protein-coding genes. However, a future challenge is the functional dissection of complex transcription including ncRNA across a number of species. At this point, it is unclear how to systematically dissect the function of ncRNA and what the expected outcomes are. Similar to its application to protein-coding genes, mutagenesis is an attractive method to assess the functional role of ncRNA. However, only 2 of 50 ncRNA transcripts in yeast exhibited a mutant phenotype when disrupted under rich media conditions [22]. In contrast, over one-third of protein-coding yeast genes are either essential or required for optimal growth under the same condition [24]. One interpretation of these results is that ncRNA is not as important to the cell as protein-coding genes. An alternative explanation is that this ncRNA plays an important cellular role but that more complex arrangements of mutations are necessary to observe phenotypes for these transcripts (see below).

Genetic Interactions

Single gene mutant studies of the kind described above are limited in that many genes will not display a phenotype under any tested condition. Therefore, despite extensive profiling of a single-gene mutant library across a wide number of conditions, a sizeable fraction of the genome will escape functional annotation. This limitation of single gene mutant studies can be partially reconciled by the fact that the genome encodes a network of interacting proteins, RNA, metabolites, and other macromolecules that control complex physiological processes. One implication of this cellular complexity is genetic redundancy at the gene or

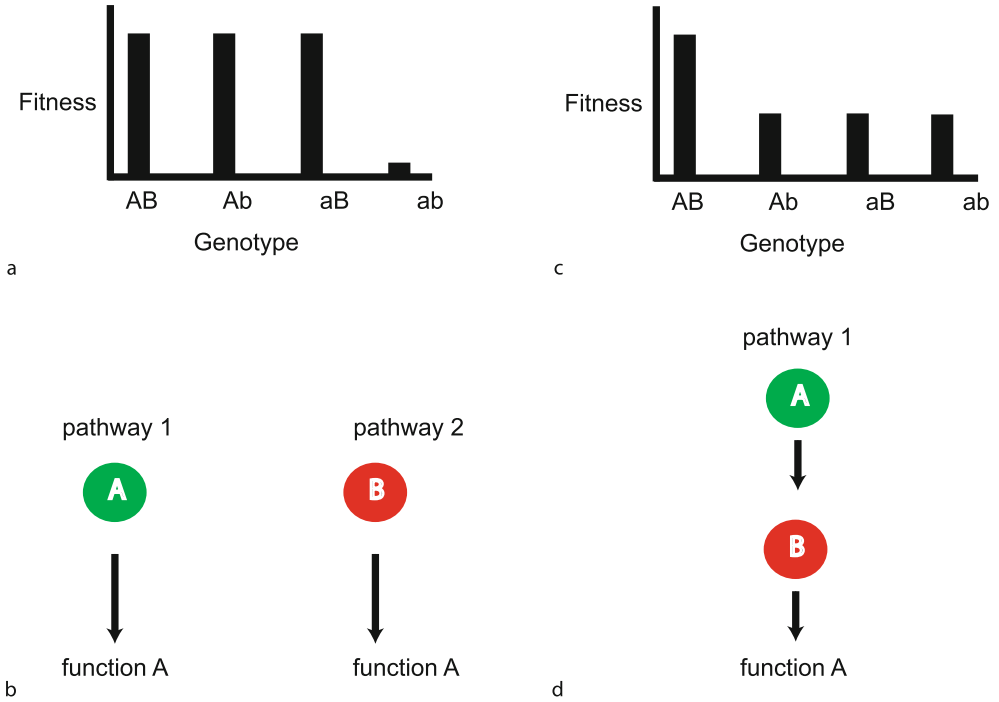

Functional Genomics for Characterization of Genome Sequences, Figure 3

Physical interpretation of genetic interactions. a Relative fitness of two genes, A and B, in combination. Uppercase letter refers to the functional (wildtype) copy of the gene. Lowercase letter refers to a complete loss of function (null) mutation in that gene. Single mutations of both A (genotype aB) and B (Ab) have no effect on fitness. A strain harboring mutations in both A and B (ab) exhibits a severe fitness defect. This deviation is referred to as synergistic epistasis. Synthetic lethality is a special instance of synergistic epistasis in which the double mutant is lethal. b The most common physical explanation for synergistic epistasis is that genes A and B act in separate pathways with redundant functions (function A). The loss of either A and B is compensated by the other, intact gene. c Using the same terminology in a, single gene mutants in both A and B have moderate effects on fitness. The theoretical expectation for genes acting in independent pathways is that the fitness of the double mutant (ab) should be the product of the individual fitness effects. Therefore, the fitness of the double mutant should be less than the fitness of the single mutants. In this example, however, the fitness of the double mutant is equal to the fitness of the single mutants. This deviation is referred to as antagonistic (or alleviating) epistasis. d Antagonistic epistasis occurs when two genes act in the same pathway. The loss of either gene singly abolishes the function of the pathway; therefore the loss of both genes would have an identical effect on fitness

pathway level. A mutated gene with an intact, functional copy elsewhere in the genome will not exhibit a mutant phenotype under most conditions. A second implication of cellular complexity is that the network structure itself has evolved robustness against mutations in many of its constituent components. Robustness against mutation is evident in the metabolic network of *E. coli*, where disruptions in many genes are managed by alternative routes through metabolism [61].

One attractive option for unraveling the complexity of these cellular networks and for assigning putative functions to genes is to analyze strains harboring multiple mutations. In instances where the phenotype of the mutant combination deviates from the expected phenotype of combining the single mutants, a genetic interaction is said to exist between the two genes (Fig. 3). Given the potential of genetic interactions to reveal novel functional linkages between genes, researchers have started to systematically identify these interactions at the genome-level. In yeast, techniques have been developed to probe a query mutation against all systematically deleted genes using both automated analysis of individual double mutants [115] and using molecular barcodes, mutant pools, and microarrays [89]. These findings indicate that the average yeast gene has a genetic interaction with ~34 other nonessential genes. Furthermore, gene pairs that exhibit a genetic interaction are more likely to be functionally related providing evidence that these interactions are biologically meaningful and can be used to assign putative gene functions [115].

Summary

At the current stage, our genome is computationally annotated, transcripts have been mapped using tiling arrays, we have profiled a number of conditions using gene expression microarrays, and we have started to undertake systematic loss-of-function screens to uncover phenotypes for genes. This pattern of experimentation describes an increasingly common scenario for how new genome sequences are analyzed today. What did the mutant profiling screens add? Most importantly, at both the single and double mutant level, mutant phenotypes seems to be more direct indicators of gene function than gene expression data. Therefore, despite the cost and difficulty, it is imperative that systematic mutant collections are tested for a number of model organisms. Additionally, given the effort to generate these collections, novel phenotyping technology needs to be developed such that the value of each mutant collection is maximized. Nevertheless, the experiments described thus far have focused on elucidating genome function in the context of a single representative individual of a species. Natural populations within species, however, exhibit substantial phenotypic diversity due to the presence of genetic variation in the population (polymorphisms) or environmental effects. Consequently, the precise functionality of a gene in a species can be dependent on the individual that is sampled. In the next section, we therefore describe the nature of intraspecific variation and how the tools of functional genomics are being used to tackle this problem.

Functional Genomics and Complex Traits

Most intraspecific phenotypic variation is conditioned by multiple genes and the environment. Furthermore, these genes can interact with each other and the environment both additively and epistatically to determine phenotype. For these reasons, this class of traits is often referred to as complex (referring to the underlying genetics) or quantitative (referring to the effect of each contributing gene or quantitative trait locus to the overall phenotype). Finding the genetic variation that contributes to quantitative phenotypic variation is a major challenge in contemporary genetics. The susceptibility to many common human diseases including type II diabetes, bipolar disorder, and heart disease is a genetically complex, quantitative trait. Here we present the challenge of complex genetics and discuss how the techniques of transcript profiling and mutant analysis are being applied to address this issue both in human and in model organisms.

Genotyping, Linkage, and Association

A primary goal of quantitative genetics is to identify the genetic variation that contributes to differences in observable traits among the population. Most genetic variation within species is single base pair differences commonly referred to as single nucleotide polymorphisms (SNPs). Therefore, it is necessary that one is able to genotype SNPs within a population and relate this information to phenotypic measurements. SNPs that correlate with a particular phenotypic value (for instance, tall height in humans) are potentially causative. In practice, there are two fundamental approaches for mapping SNPs that contribute to phenotypic variation: mapping based approaches such as linkage analysis and association studies and candidate gene approaches. Following a discussion of modern genotyping techniques, we will first describe the experimental design and application of linkage and association.

There are numerous techniques for genotyping individuals within a population. One can directly sequence a known SNP of interest using terminator chemistry or pyrosequencing. Additionally, single nucleotide extension

assays can be used to genotype known SNPs. The advantages of these techniques are the accuracy of the SNP calls and the simplicity of doing the experiment. The primary disadvantage is the cost and labor required to genotype thousands of SNPs from a single individual. Given the availability and necessity of dense SNP (marker) maps in humans and other model organisms, it is imperative that genotyping techniques be both rapid (so that numerous individuals can be genotyped) and massively parallel (so that thousands of SNPs can be genotyped at once). Microarray-based approaches for genotyping fit these criteria and are available from a number of commercial suppliers [32].

Linkage methods rely on known family history (in human) or controlled laboratory crosses (in model organisms) to map causative genetic variation. The basic premise behind linkage is that the causative allele will segregate preferentially among members of the family sharing a certain phenotypic value and not in those without this phenotype. In model organisms, the advantages of linkage techniques include experimental control and the capability to detect genetic loci with major effects on the phenotype. In one classic example, linkage mapping led to the discovery of fw2.2, a major effect QTL that contributes significantly to the size of tomatoes [36]. In humans, the primary disadvantages of linkage mapping are small family sizes and the absence of a detailed genealogy for most populations. The population of Iceland, in which an isolated population kept detailed records of family history, is an exception; linkage methodology applied to this population has led to the discovery of allelic variants that contribute to a number of diseases including type II diabetes [46].

Given the lack of family history for most human populations, the majority of human mapping efforts are currently focused on association. In association studies, a case group (with the phenotype of interest) and an appropriately chosen control are genotyped for a large number of SNPs. In association studies, the case and control groups rarely contain family members, rather the participants are chosen from broader populations. The goal of association studies is to identify SNPs that are statistically overrepresented in the case group compared to the control group. These SNPs represent candidate chromosomal regions that may contain causative genetic variation. Technological advances underlie the recent success of human association studies. First, a haplotype map of the human genome [60] has reduced the number of SNPs that need to genotyped (due to linkage disequilibrium) to achieve genome-wide coverage. Second, hundreds of thousands of SNPs can be genotyped in parallel using microarray-based approaches. With these resources in hand, researchers have used association-based methods to identify genetic variants associated with a number of multigenic diseases including type II diabetes [101,105,131] and bipolar disorder [124].

The successful application of linkage and association studies described above emphasizes the complexity of quantitative traits. As such, a number of fundamental questions remain to be answered: how many genes condition a quantitative trait, what are the effect sizes of each of the genes, is the causative variation more likely to be at the regulatory or protein level, and finally, how does genetic variation interact at the molecular level to determine phenotype? In a recurrent theme during the history of molecular biology, the use of model organisms is necessary to uncover fundamental principles of biology. In addition to novel insights into the genetic architecture of quantitative traits, it is anticipated that the techniques used to dissect quantitative variation in model organisms will be applied to the same class of traits in humans. Below we briefly discuss some research developments in model organism quantitative genetics (by primarily candidate gene approaches) using the functional genomic techniques described in this article as a framework.

Application of Transcription

Differences in transcript levels contribute to observable phenotypic variation (for review, [97]). Consequently, researchers have looked into the genetics of gene expression variation in a number of species including yeast [10], mouse, and human [102]. Because each gene and its transcript level can be considered an individual trait, the genetic analysis of gene expression is a powerful model for QTL studies as thousands of traits can be considered simultaneously. A number of lessons emerge from these studies. Early work in yeast suggests that, for the majority of genes, the underlying genetics controlling gene expression is multigenic and complex [10]. Additionally, the majority of loci that affect gene expression are distant acting factors, not local regulatory sequences linked to the gene under investigation. Finally, the majority of causative distant acting factors do not appear to be transcription factors [130]. This argues that the genetic network that contributes to gene expression variation is affected by many global biological processes, not just those directly related to gene transcription.

In mouse and human, the genetic analysis of gene expression is aiding in the identification of QTLs (by proposing candidate genes) and providing insight into disease mechanisms. Schadt et al. examined the transcript levels of 111 mice derived from a cross between two laboratory

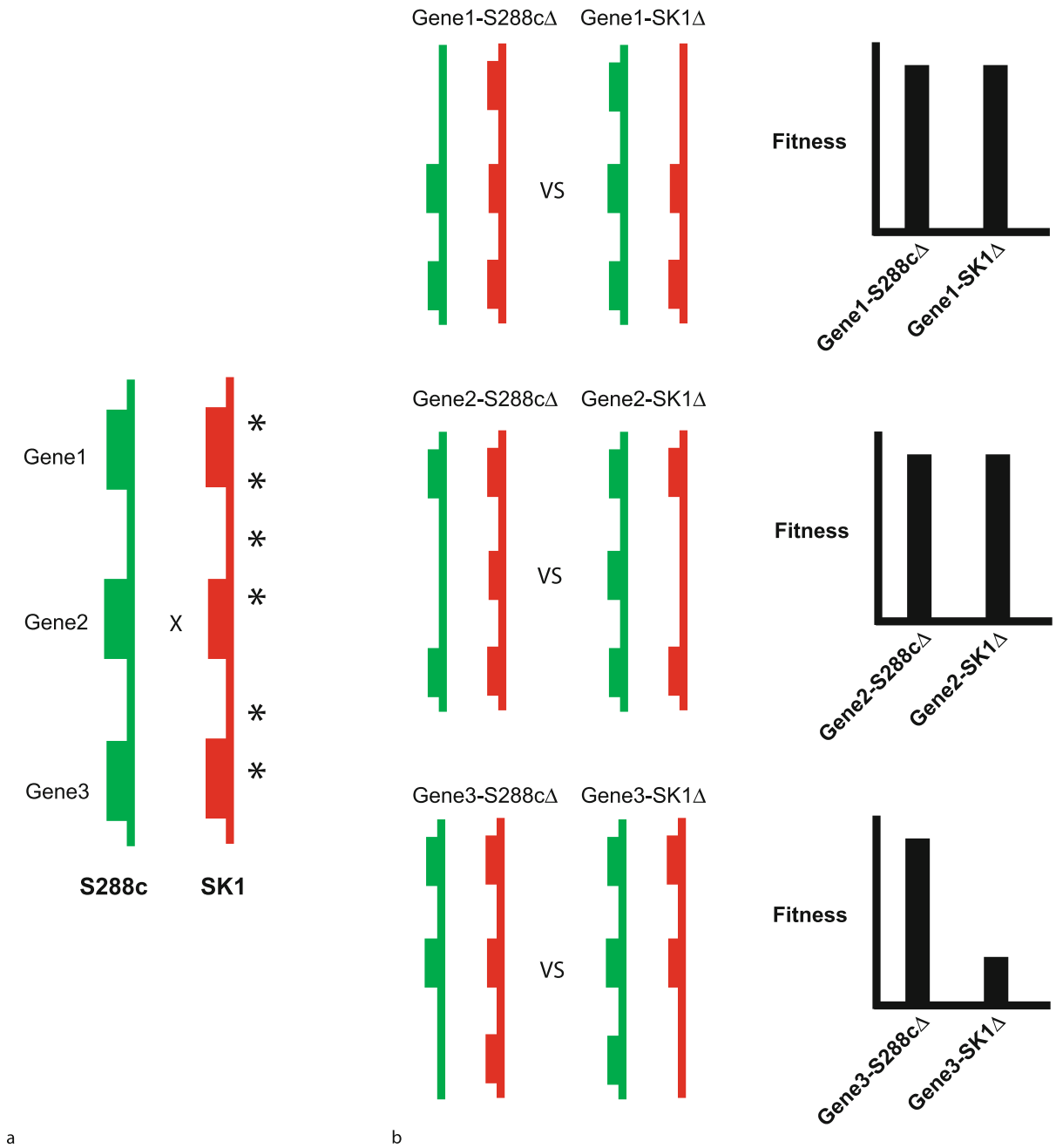

Functional Genomics for Characterization of Genome Sequences, Figure 4

Reciprocal hemizygosity analysis. a Reciprocal hemizygosity analysis (RHA) is the quantitative profiling of hemizygous (one allele is deleted), hybrid diploid strains. Illustrated is a hybrid diploid yeast strain derived from a cross between two divergent strains; shown here for S288c (green) and SK1 (red). The asterisks along the SK1 chromosome represent single nucleotide polymorphisms (SNPs) between the two strains. b In RHA, two reciprocal deletion strains are compared directly for their contribution to the phenotype (in the example here, fitness). The two strains are completely isogenic except each carries a single gene from a different genetic background. For Gene1 and Gene2, the reciprocal deletion strains harbor identical phenotypes. One can conclude that the SK1 and S288c versions of these alleles contribute equally to fitness in the S288c/SK1 hybrid background. For Gene3, the strain with a deletion in the SK1 allele has a significantly lower fitness than the S288c deletion strain. The interpretation is that the SK1 allele has a larger effect on fitness than the S288c allele in the hybrid background. The SNPs in and around Gene3 are candidates for the causative genetic variation

strains segregating for obesity [102]. Over 2000 genes with transcript level differences between the parents significantly linked to quantitative trait loci which may influence gene expression levels or contribute to obesity. Among the progeny two distinct obesity gene expression subtypes could be defined. Interestingly, the two subtypes are under the genetic control of different QTLs suggesting that multiple molecular mechanisms can contribute to what appears to be a single phenotype (obesity). These results demonstrate that the genetic dissection of gene expression in mammals can classify disease subtypes, identify novel interactions between loci, and identify potential therapeutic targets. However, it is important to note that not all phenotypic variation will be the result of transcript level differences. The importance of other molecular mechanisms, including structural mutations in proteins, implies that multiple, complementary techniques are required to fully understand intraspecific phenotypic variation.

Application of Genetics Analysis

Mutants can provide substantial insight into the genes underlying phenotypic variation because they can mirror the effects of certain types of allelic variation present in the natural population. The analysis of mutants can be used as a verification tool for the identification of the causative locus or as a technique for generating a list of candidate genes for future investigation. In mouse, large forward genetic approaches are being applied to isolate interesting mutants that have phenotypes relevant to human health including obesity, diabetes, and heart disease [85]. While this approach is promising, it is not amenable to high-throughput and the mapping of mutations is still time-consuming. Therefore, reverse-genetic efforts to systematically mutate every gene in the mouse genome, although a monumental task, are likely to provide better coverage for the entire genome in the long run [19].

Mutant screening for QTL detection has limitations similar to those of loss-of-function phenotype screens in model systems, in that redundant gene products are missed, more complex interactions not tested, and expectations are unclear when natural, causative SNPs do not elicit the same effects as the laboratory induced mutations. For these reasons, the power of mutant analysis is increased when it is applied so that a mutant is not compared to a wildtype, but instead a mutation enables a comparison between one allele and another. This technique has been established in yeast by reciprocal hemizygosity analysis (Fig. 4). After linkage mapping, reciprocal hemizygosity analysis was applied to identify the causative genes in each QTL region for the complex traits of high-temperature growth and sporulation efficiency [4,23,112]. These studies highlight the complexity of quantitative traits including the presence of multiple causative genes in a single mapped interval and the non-intuitive genetic interactions between the causative genes. In addition, the yeast work suggests that regulatory and protein coding polymorphisms in both candidate and non-candidate genes contribute to quantitative phenotypic variation. One implication of these results is that new techniques for the high-resolution dissection of QTLs are required in multicellular model organisms to truly understand the genetic basis of phenotype variation.

Summary

We argue that a complete understanding of gene function requires an understanding of how gene function varies (often subtly) among individuals within a population, not just a single representative isolate. To fulfill this promise, two substantial challenges need to be addressed. The first challenge is the identification of the genetic variants that influence any particular phenotype. In model organisms, genome-wide techniques based on linkage/association combined with transcription and mutation analysis are proving successful. Further application of mutant approaches requires substantial infrastructure to generate libraries of strains and reagents for the systematic identification of interesting phenotypes but have promise for circumventing the need for linkage mapping (Fig. 4). The second challenge is to unravel the complexity of genetic and environmental interactions at a molecular level. This will first be accomplished in model systems although it is anticipated that the lessons learned from these studies will be applicable to humans. Functional genomic techniques (combined with non-nucleic based approaches, see **a**) provide the necessary global views into gene function and hence are perhaps most promising for dissecting the genetic interactions underlying phenotypic variation.

Future Directions

Technology and Implications

The relationship between technology and biological insight is circular: biology drives technology development and new technology enables new biology. For example, the desire to sequence human genomes has driven the efforts to develop novel, ultra high-throughput sequencing technology. In turn, functional genomic techniques such as automated DNA sequencing and the DNA microarray have revealed numerous unexpected insights into the biology and functioning of genomes. No doubt, the data de-

rived from novel technology, such as ultra high-throughput sequencing of millions of humans, will revolutionize our view of living systems (and so on). Given the importance of technology in functional genomics (and biology in general), we discuss here novel techniques and their implications for discovery.

Ultra high-throughput sequencing encompass methods capable of sequencing millions of molecules in parallel in reaction volumes far lower than used for conventional Sanger sequencing [80,107]. Given the central role of DNA sequence data in functional genomics, this new generation of technology has important implications for genome-level research and comparative genomics in particular. The increased throughput of DNA sequencing will allow researchers to sample more of the natural variation present on earth. It is anticipated that the genome sequences for most commonly known multicellular organisms will be completed in the next 20 years. Furthermore, novel sequencing technologies may have utility in the analysis of extinct organisms as well. In one example, partial Neanderthal genome sequences were obtained using template derived from a fossil bone [47,87]. Such analyzes of extinct organisms can provide insights into the evolutionary past that is not possible using the DNA sequence data from contemporary species. Lastly, the environmental sequences (metagenome) of largely uncultivatable microbial communities are reshaping the way we view ecosystems [120], biogeochemical cycles [118], and even human health [117].

Above all, novel DNA sequencing technologies have the potential to revolutionize human health by cost-effectively providing the genome sequence of many individuals within the human population. The realization of such promise would usher in a new era of personalized medicine based on one's genome sequence. The identification of all polymorphisms among millions of individuals would increase the power of association studies for finding common disease genes (particularly those with small effect). Additionally, the resequencing of tumor samples will aid in the identification of the responsible mutations and perhaps suggest the appropriate therapeutic treatment.

As described in this review, our ability to sequence genomes far exceeds our ability to systematically determine genome function. Therefore, the development of ultra high-throughput DNA sequencing emphasizes the need for novel functional genomic techniques to decipher gene function. For the analysis of gene expression, we believe two major trends will dominate. First, microarrays will trend towards higher density, multiplexing, and custom design. Higher density of probes on microarrays will enable higher-resolution measurements of transcriptional architecture. The multiplexing of arrays will lower costs per experiment by reducing the amount of necessary reagents. Custom *in situ* synthesized microarrays will allow researchers to keep experimentation up to the pace of genome sequencing. The second major trend in gene expression promises to be the use of ultra high-throughput sequencing as a 'digital' readout of gene expression by directly counting the number of individual RNA molecules in a sample (even at the single cell level). The advantages of this approach are that no microarray is needed, data analysis is more straightforward, and most cellular RNA is assayed in an unbiased manner.

Despite the cost and effort, we anticipate that genome-wide resources (such as defined mutant libraries for all genes) will become available for a number of organisms. Such resources offer the potential to systematically determine gene function on a number of levels from genetic to biochemical. However, we believe the full potential of these resources can only be realized with concurrent advances in experimentation such as robotic automation, assay miniaturization, and labeling/detection systems for monitoring phenotypes and individual molecules (for reviews, see [55,81]). Most importantly, increasing accuracy and throughput of experimentation while simultaneously reducing cost will enable the application of novel technology in the individual laboratory and hence spread the impact of these approaches [110].

Data Integration and Systems Biology

In order to understand cellular function one also needs to examine the dynamics and interactions of biological data types beyond DNA, such as proteins, metabolites, and lipids. Like for DNA, high-throughput technologies have been developed to increase the speed and decrease the cost of obtaining this information, and methods have been developed to integrate diverse data types. These developments characterize the fields of proteomics, metabolomics, modeling, and systems biology, to name but a few. While each of these topics are subjects of other chapters in this book, we touch upon them briefly here to provide a context in which all genomic DNA data needs to be analyzed today. The key is to incorporate all of the available information together, to gain a picture of biological function greater than the sum of the individual parts.

Cells are enormously complex and each of the thousands of biochemical reactions carried out by genes is part of a large network of connected reactions – encoded by information in the genome but regulated by cues from the environment. Conceptually, disturbances in cells, like what happens during any type of disease, can be viewed

as a disturbance in the network, and in order to predict the effect of disturbances we need to identify all molecular components, know where they localize, with whom they interact, and how their activity is regulated.

Among biological entities other than DNA, proteins have been most amenable to high-throughput study (often employing genetic manipulation) and their global characterization is perhaps furthest along. High-throughput technologies have been developed to analyze the expression of proteins [41], their localization [58], phosphorylation patterns [21], protein-DNA interactions [50], and protein-protein interactions [39,40,53,62,71,119]. Any dataset that defines a relationship between proteins can be used for the purpose of making a network. Thus, interaction linkages have been defined based on physical protein-protein interactions [39,40,53,62,71,119], expression regulation [76,113], mutant phenotypes [111,115], phylogenetic profiles [91], literature mining [79], and orthology transfer of interaction evidence across species [128]. Not surprisingly, we now know at least some of the putative interrelationships among most of the proteome of yeast (and other organisms are following) [38]. While many of the interactions are suggestive and need to be confirmed, already they provide a context to functionally characterize proteins and enable an understanding of the whole that goes beyond knowledge from the components in isolation.

Because individual large scale datasets are often incomplete, more information is obtained by integrating datasets. Integrating heterogeneous but complementary interaction data types has improved the accuracy and the coverage in detecting protein associations [121] and has been implemented globally [20,49,59,63,75,84,116]. In one example, integrated protein networks have been constructed for mitochondria, where they have enabled functional predictions to be made for hundreds of previously uncharacterized components, providing a survey of systems properties and enabling the prediction of disease candidate genes [93,108].

Currently, we are still far away from being able to develop any complete models of eukaryotic cells. In order to identify and understand the function of biological networks we need to move from static to dynamic models, from defining the parts to understanding the biological processes in the context of the entire system. Knowledge about systems biology can then fuel the next series of applications, to enable the design and construction of biological systems that can (for example) process information, generate energy, fabricate materials, provide food, or maintain and enhance human health and our environment. For these reasons, as well as for the insights obtained from cracking the code itself, understanding the blueprint of the genome using functional genomics is a fascinating quest, and will remain so for some time to come.

Acknowledgments

We thank Zhenyu Xu for preparing Fig. 2 and funding from the National Institutes of Health to AMD and LMS and the Deutsche Forschungsgemeinschaft to LMS.

Bibliography

1. Abby S, Daubin V (2007) Comparative genomics and the evolution of prokaryotes. Trends Microbiol 15:135–141
2. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, et al. (2000) The genome sequence of Drosophila melanogaster. Science 287:2185–2195
3. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol 2:2006 0008
4. Ben-Ari G, Zenvirth D, Sherman A, David L, Klutstein M, Lavi U, Hillel J, Simchen G (2006) Four linked genes participate in controlling sporulation efficiency in budding yeast. PLoS Genet 2:e195
5. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, Bulyk ML (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol 24:1429–1435
6. Berns K, Hijmans EM, Mullenders J, Brummelkamp TR, Velds A, Heimerikx M, Kerkhoven RM, Madiredjo M, Nijkamp W, Weigelt B, Agami R, Ge W, Cavet G, Linsley PS, Beijersbergen RL, Bernards R (2004) A large-scale RNAi screen in human cells identifies new components of the p53 pathway. Nature 428:431–437
7. Bertone P, Gerstein M, Snyder M (2005) Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. Chromosome Res 13:259–274
8. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M (2004) Global identification of human transcribed sequences with genome tiling arrays. Science 306:2242–2246

9. Borsani O, Zhu J, Verslues PE, Sunkar R, Zhu JK (2005) Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis. Cell 123: 1279–1291
10. Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. Science 296:752–755
11. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA 97:262–267
12. Buck MJ, Lieb JD (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics 83:349–360
13. Bushati N, Cohen SM (2007) microRNA Functions. Annu Rev Cell Dev Biol 23:175–205
14. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, Civello D, Adams MD, Cargill M, Clark AG (2005) Natural selection on protein-coding genes in the human genome. Nature 437:1153–1157
15. C. elegans Sequencing Consortium (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 282:2012–2018
16. Chen J, Sun M, Hurst LD, Carmichael GG, Rowley JD (2005) Genome-wide analysis of coordinate expression and evolution of human cis-encoded sense-antisense transcripts. Trends Genet 21:326–329
17. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN, Pollard DA, Sackton TB, Larracuente AM, Singh ND, Abad JP, Abt DN, Adryan B, Aguade M, Akashi H, Anderson WW, Aquadro CF, Ardell DH, Arguello R, Artieri CG, Barbash DA, Barker D, Barsanti P, Batterham P, Batzoglou S, Begun D, Bhutkar A, Blanco E, Bosak SA, Bradley RK, Brand AD, Brent MR, Brooks AN, Brown RH, Butlin RK, Caggese C, Calvi BR, Bernardo de Carvalho A, Caspi A, Castrezana S, Celniker SE, Chang JL, Chapple C, Chatterji S, Chinwalla A, Civetta A, Clifton SW, Comeron JM, Costello JC, Coyne JA, Daub J, David RG, Delcher AL, Delehaunty K, Do CB, Ebling H, Edwards K, Eickbush T, Evans JD, Filipski A, Findeiss S, Freyhult E, Fulton L, Fulton R, Garcia AC, Gardiner A, Garfield DA, Garvin BE, Gibson G, Gilbert D, Gnerre S, Godfrey J, Good R, Gotea V, Gravely B, Greenberg AJ, Griffiths-Jones S, Gross S, Guigo R, Gustafson EA, Haerty W, Hahn MW, Halligan DL, Halpern AL, Halter GM, Han MV, Heger A, Hillier L, Hinrichs AS, Holmes I, Hoskins RA, Hubisz MJ, Hultmark D, Huntley MA, Jaffe DB, Jagadeeshan S, et al. (2007) Evolution of genes and genomes on the Drosophila phylogeny. Nature 450:203–218
18. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science 301:71–76
19. Collins FS, Rossant J, Wurst W (2007) A mouse for all reasons. Cell 128:9–13
20. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO (2004) Integrating high-throughput and computational data elucidates bacterial networks. Nature 429:92–96
21. Cox J, Mann M (2007) Is proteomics the new genomics? Cell 130:395–398
22. David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM (2006) A high-resolution map of transcription in the yeast genome. Proc Natl Acad Sci USA 103:5320–5325
23. Deutschbauer AM, Davis RW (2005) Quantitative trait loci mapped to single-nucleotide resolution in yeast. Nat Genet 37:1333–1340
24. Deutschbauer AM, Jaramillo DF, Proctor M, Kumm J, Hillenmeyer ME, Davis RW, Nislow C, Giaever G (2005) Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. Genetics 169:1915–1925
25. Deutschbauer AM, Williams RM, Chu AM, Davis RW (2002) Parallel phenotypic analysis of sporulation and postgermination growth in Saccharomyces cerevisiae. Proc Natl Acad Sci USA 99:15530–15535
26. Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuveglise C, Talla E, Goffard N, Frangeul L, Aigle M, Anthouard V, Babour A, Barbe V, Barnay S, Blanchin S, Beckerich JM, Beyne E, Bleykasten C, Boisrame A, Boyer J, Cattolico L, Confanioleri F, De Daruvar A, Despons L, Fabre E, Fairhead C, Ferry-Dumazet H, Groppi A, Hantraye F, Hennequin C, Jauniaux N, Joyet P, Kachouri R, Kerrest A, Koszul R, Lemaire M, Lesur I, Ma L, Muller H, Nicaud JM, Nikolski M, Oztas S, Ozier-Kalogeropoulos O, Pellenz S, Potier S, Richard GF, Straub ML, Suleau A, Swennen D, Tekaia F, Wesolowski-Louvel M, Westhof E, Wirth B, Zeniou-Meyer M, Zivanovic I, Bolotin-Fukuhara M, Thierry A, Bouchier C, Caudron B, Scarpelli C, Gaillardin C, Weissenbach J, Wincker P, Souciet JL (2004) Genome evolution in yeasts. Nature 430:35–44
27. Eddy SR (2002) Computational genomics of noncoding RNA genes. Cell 109:137–140
28. Elnitski L, Jin VX, Farnham PJ, Jones SJ (2006) Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. Genome Res 16: 1455–1464
29. Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, Kitano T, Monaco AP, Paabo S (2002) Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418:869–872
30. ENCODE Consortium (2007a) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799–816
31. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol 5:e8
32. Fan JB, Chee MS, Gunderson KL (2006) Highly parallel genomic assays. Nat Rev Genet 7:632–644
33. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature 391: 806–811
34. Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS (2006) Conservation of RET regulatory function from human to zebrafish without sequence similarity. Science 312:276–279
35. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269:496–512
36. Frary A, Nesbitt TC, Grandillo S, Knaap E, Cong B, Liu J, Meller J, Elber R, Alpert KB, Tanksley SD (2000) fw2.2: a quantitative

trait locus key to the evolution of tomato fruit size. Science 289:85–88
37. Friedman A, Perrimon N (2006) A functional RNAi screen for regulators of receptor tyrosine kinase and ERK signalling. Nature 444:230–234
38. Gagneur J, David L, Steinmetz LM (2006) Capturing cellular machines by systematic screens of protein complexes. Trends Microbiol 14:336–339
39. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440:631–636
40. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415:141–147
41. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS (2003) Global analysis of protein expression in yeast. Nature 425:737–741
42. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, Arkin AP, Astromoff A, El-Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian KD, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Guldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kotter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo C, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Ross-Macdonald P, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sookhai-Mahadeo S, Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, Wang CY, Ward TR, Wilhelmy J, Winzeler EA, Yang Y, Yen G, Youngman E, Yu K, Bussey H, Boeke JD, Snyder M, Philippsen P, Davis RW, Johnston M (2002) Functional profiling of the Saccharomyces cerevisiae genome. Nature 418:387–391
43. Giaever G, Flaherty P, Kumm J, Proctor M, Nislow C, Jaramillo DF, Chu AM, Jordan MI, Arkin AP, Davis RW (2004) Chemogenomic profiling: identifying the functional interactions of small molecules in yeast. Proc Natl Acad Sci USA 101:793–798
44. Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, Hutchison CA, Smith HO, Venter JC (2006) Essential genes of a minimal bacterium. Proc Natl Acad Sci USA 103:425–430
45. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996) Life with 6000 genes. Science 274:546, 563–567
46. Grant SF, Thorleifsson G, Reynisdottir I, Benediktsson R, Manolescu A, Sainz J, Helgason A, Stefansson H, Emilsson V, Helgadottir A, Styrkarsdottir U, Magnusson KP, Walters GB, Palsdottir E, Jonsdottir T, Gudmundsdottir T, Gylfason A, Saemundsdottir J, Wilensky RL, Reilly MP, Rader DJ, Bagger Y, Christiansen C, Gudnason V, Sigurdsson G, Thorsteinsdottir U, Gulcher JR, Kong A, Stefansson K (2006) Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat Genet 38:320–323
47. Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Du L, Egholm M, Rothberg JM, Paunovic M, Paabo S (2006) Analysis of one million base pairs of Neanderthal DNA. Nature 444:330–336
48. Guigo R, Dermitzakis ET, Agarwal P, Ponting CP, Parra G, Reymond A, Abril JF, Keibler E, Lyle R, Ucla C, Antonarakis SE, Brent MR (2003) Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. Proc Natl Acad Sci USA 100: 1140–1145
49. Gunsalus KC, Ge H, Schetter AJ, Goldberg DS, Han JD, Hao T, Berriz GF, Bertin N, Huang J, Chuang LS, Li N, Mani R, Hyman AA, Sonnichsen B, Echeverri CJ, Roth FP, Vidal M, Piano F (2005) Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis. Nature 436:861–865
50. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA (2004) Transcriptional regulatory code of a eukaryotic genome. Nature 431:99–104
51. Hardison RC (2003) Comparative genomics. PLoS Biol 1:e58
52. Hensel M, Shea JE, Gleeson C, Jones MD, Dalton E, Holden DW (1995) Simultaneous identification of bacterial virulence genes by negative selection. Science 269:400–403
53. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415:180–183
54. Hogan GJ, Lee CK, Lieb JD (2006) Cell cycle-specified fluctuation of nucleosome occupancy at gene promoters. PLoS Genet 2:e158
55. Hong JW, Quake SR (2003) Integrated nanoliter systems. Nat Biotechnol 21:1179–1183
56. Hongay CF, Grisafi PL, Galitski T, Fink GR (2006) Antisense transcription controls cell fate in Saccharomyces cerevisiae. Cell 127:735–745
57. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH (2000) Functional discovery via a compendium of expression profiles. Cell 102:109–126
58. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK (2003) Global analysis of protein localization in budding yeast. Nature 425:686–691
59. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L (2001) Inte-

grated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292:929–934
60. International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320
61. Ishii N, Nakahigashi K, Baba T, Robert M, Soga T, Kanai A, Hirasawa T, Naba M, Hirai K, Hoque A, Ho PY, Kakazu Y, Sugawara K, Igarashi S, Harada S, Masuda T, Sugiyama N, Togashi T, Hasegawa M, Takai Y, Yugi K, Arakawa K, Iwata N, Toya Y, Nakayama Y, Nishioka T, Shimizu K, Mori H, Tomita M (2007) Multiple high-throughput analyses monitor the response of E. coli to perturbations. Science 316:593–597
62. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 98:4569–4574
63. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 302:449–453
64. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316:1497–1502
65. Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, Welchman DP, Zipperlen P, Ahringer J (2003) Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature 421:231–237
66. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tammana H, Gingeras TR (2007a) RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316:1484–1488
67. Kapranov P, Willingham AT, Gingeras TR (2007b) Genome-wide transcription and the implications for genomic organization. Nat Rev Genet 8:413–423
68. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241–254
69. Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P, Boland F, Brignell SC, Bron S, Bunai K, Chapuis J, Christiansen LC, Danchin A, Debarbouille M, Dervyn E, Deuerling E, Devine K, Devine SK, Dreesen O, Errington J, Fillinger S, Foster SJ, Fujita Y, Galizzi A, Gardan R, Eschevins C, Fukushima T, Haga K, Harwood CR, Hecker M, Hosoya D, Hullo MF, Kakeshita H, Karamata D, Kasahara Y, Kawamura F, Koga K, Koski P, Kuwana R, Imamura D, Ishimaru M, Ishikawa S, Ishio I, Le Coq D, Masson A, Mauel C, Meima R, Mellado RP, Moir A, Moriya S, Nagakawa E, Nanamiya H, Nakai S, Nygaard P, Ogura M, Ohanan T, O'Reilly M, O'Rourke M, Pragai Z, Pooley HM, Rapoport G, Rawlins JP, Rivas LA, Rivolta C, Sadaie A, Sadaie Y, Sarvas M, Sato T, Saxild HH, Scanlan E, Schumann W, Seegers JF, Sekiguchi J, Sekowska A, Seror SJ, Simon M, Stragier P, Studer R, Takamatsu H, Tanaka T, Takeuchi M, Thomaides HB, Vagner V, van Dijl JM, Watabe K, Wipat A, Yamamoto H, Yamamoto M, Yamamoto Y, Yamane K, Yata K, Yoshida K, Yoshikawa H, Zuber U, Ogasawara N (2003) Essential Bacillus subtilis genes. Proc Natl Acad Sci USA 100:4678–4683
70. Korf I, Flicek P, Duan D, Brent MR (2001) Integrating genomic homology into gene structure prediction. Bioinformatics 17 Suppl 1:S140–S148
71. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440:637–643
72. Kumar S, Subramanian S (2002) Mutation rates in mammalian genomes. Proc Natl Acad Sci USA 99:803–808
73. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921
74. Lapidot M, Pilpel Y (2006) Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. EMBO Rep 7:1216–1222
75. Lee I, Date SV, Adai AT, Marcotte EM (2004) A probabilistic functional network of yeast genes. Science 306:1555–1558
76. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 298:799–804
77. Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, Nislow C (2007) A high-resolution atlas of nucleosome occupancy in yeast. Nat Genet 39:1235–1244
78. Lieb JD, Liu X, Botstein D, Brown PO (2001) Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nat Genet 28:327–334
79. Marcotte EM, Xenarios I, Eisenberg D (2001) Mining literature for protein-protein interactions. Bioinformatics 17:359–363
80. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP,

Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

81. Melin J, Quake SR (2007) Microfluidic large-scale integration: the evolution of design rules for biological automation. Annu Rev Biophys Biomol Struct 36:213–231
82. Meyer IM, Durbin R (2004) Gene structure conservation aids similarity based gene prediction. Nucleic Acids Res 32: 776–783
83. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448:553–560
84. Myers CL, Robson D, Wible A, Hibbs MA, Chiriac C, Theesfeld CL, Dolinski K, Troyanskaya OG (2005) Discovery of biological networks from diverse functional genomic data. Genome Biol 6:R114
85. Nadeau JH, Balling R, Barsh G, Beier D, Brown SD, Bucan M, Camper S, Carlson G, Copeland N, Eppig J, Fletcher C, Frankel WN, Ganten D, Goldowitz D, Goodnow C, Guenet JL, Hicks G, Hrabe de Angelis M, Jackson I, Jacob HJ, Jenkins N, Johnson D, Justice M, Kay S, Kingsley D, Lehrach H, Magnuson T, Meisler M, Poustka A, Rinchik EM, Rossant J, Russell LB, Schimenti J, Shiroishi T, Skarnes WC, Soriano P, Stanford W, Takahashi JS, Wurst W, Zimmer A (2001) Sequence interpretation. Functional annotation of mouse genome sequences. Science 291:1251–1255
86. Nobrega MA, Zhu Y, Plajzer-Frick I, Afzal V, Rubin EM (2004) Megabase deletions of gene deserts result in viable mice. Nature 431:988–993
87. Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, Chen F, Platt D, Paabo S, Pritchard JK, Rubin EM (2006) Sequencing and analysis of Neanderthal genomic DNA. Science 314:1113–1118
88. Paddison PJ, Silva JM, Conklin DS, Schlabach M, Li M, Aruleba S, Balija V, O'Shaughnessy A, Gnoj L, Scobie K, Chang K, Westbrook T, Cleary M, Sachidanandam R, McCombie WR, Elledge SJ, Hannon GJ (2004) A resource for large-scale RNA-interference-based screens in mammals. Nature 428:427–431
89. Pan X, Yuan DS, Xiang D, Wang X, Sookhai-Mahadeo S, Bader JS, Hieter P, Spencer F, Boeke JD (2004) A robust toolkit for functional profiling of the yeast genome. Mol Cell 16:487–496
90. Parsons AB, Lopez A, Givoni IE, Williams DE, Gray CA, Porter J, Chua G, Sopko R, Brost RL, Ho CH, Wang J, Ketela T, Brenner C, Brill JA, Fernandez GE, Lorenz TC, Payne GS, Ishihara S, Ohya Y, Andrews B, Hughes TR, Frey BJ, Graham TR, Andersen RJ, Boone C (2006) Exploring the mode-of-action of bioactive compounds by chemical-genetic profiling in yeast. Cell 126:611–625
91. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA 96:4285–4288
92. Pena-Castillo L, Hughes TR (2007) Why are there still over 1000 uncharacterized yeast genes? Genetics 176:7–14
93. Perocchi F, Jensen LJ, Gagneur J, Ahting U, von Mering C, Bork P, Prokisch H, Steinmetz LM (2006) Assessing systems properties of yeast mitochondria through an interaction map of the organelle. PLoS Genet 2:e170
94. Prescott EM, Proudfoot NJ (2002) Transcriptional collision between convergent genes in budding yeast. Proc Natl Acad Sci USA 99:8796–8801
95. Rivas E, Klein RJ, Jones TA, Eddy SR (2001) Computational identification of noncoding RNAs in E. coli by comparative genomics. Curr Biol 11:1369–1373
96. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4:651–657
97. Rockman MV, Kruglyak L (2006) Genetics of global gene expression. Nat Rev Genet 7:862–872
98. Roemer T, Jiang B, Davison J, Ketela T, Veillette K, Breton A, Tandia F, Linteau A, Sillaots S, Marta C, Martel N, Veronneau S, Lemieux S, Kauffman S, Becker J, Storms R, Boone C, Bussey H (2003) Large-scale essential gene identification in Candida albicans and applications to antifungal drug discovery. Mol Microbiol 50:167–181
99. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74: 5463–5467
100. Sassetti CM, Boyd DH, Rubin EJ (2001) Comprehensive identification of conditionally essential genes in mycobacteria. Proc Natl Acad Sci USA 98:12712–12717
101. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, Groop L, Altshuler D, Almgren P, Florez JC, Meyer J, Ardlie K, Bengtsson Bostrom K, Isomaa B, Lettre G, Lindblad U, Lyon HN, Melander O, Newton-Cheh C, Nilsson P, Orho-Melander M, Rastam L, Speliotes EK, Taskinen MR, Tuomi T, Guiducci C, Berglund A, Carlson J, Gianniny L, Hackett R, Hall L, Holmkvist J, Laurila E, Sjogren M, Sterner M, Surti A, Svensson M, Svensson M, Tewhey R, Blumenstiel B, Parkin M, Defelice M, Barry R, Brodeur W, Camarata J, Chia N, Fava M, Gibbons J, Handsaker B, Healy C, Nguyen K, Gates C, Sougnez C, Gage D, Nizzari M, Gabriel SB, Chirn GW, Ma Q, Parikh H, Richardson D, Ricke D, Purcell S (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316:1331–1336
102. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature 422:297–302
103. Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470
104. Scherens B, Goffeau A (2004) The uses of genome-wide yeast mutant collections. Genome Biol 5:229
105. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, Buchanan TA, Watanabe RM, Valle TT, Kinnunen L, Abecasis GR, Pugh EW, Doheny KF, Bergman RN, Tuomilehto J, Collins FS, Boehnke M (2007) A genome-wide

association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316:1341–1345
106. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang JP, Widom J (2006) A genomic code for nucleosome positioning. Nature 442:772–778
107. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728–1732
108. Shutt TE, Shadel GS (2007) Expanding the mitochondrial interactome. Genome Biol 8:203
109. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE (1986) Fluorescence detection in automated DNA sequence analysis. Nature 321:674–679
110. Steinmetz LM, Davis RW (2004) Maximizing the potential of functional genomics. Nat Rev Genet 5:190–201
111. Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, Jones T, Chu AM, Giaever G, Prokisch H, Oefner PJ, Davis RW (2002a) Systematic screen for human disease genes in yeast. Nat Genet 31:400–404
112. Steinmetz LM, Sinha H, Richards DR, Spiegelman JI, Oefner PJ, McCusker JH, Davis RW (2002b) Dissecting the architecture of a quantitative trait locus in yeast. Nature 416:326–330
113. Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science 302:249–255
114. Taft RJ, Pheasant M, Mattick JS (2007) The relationship between non-protein-coding DNA and eukaryotic complexity. Bioessays 29:288–299
115. Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Menard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu AM, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C (2004) Global mapping of the yeast genetic interaction network. Science 303:808–813
116. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proc Natl Acad Sci USA 100:8348–8353
117. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444:1027–1031
118. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37–443
119. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403:623–627
120. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74
121. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res 33:D433–D437
122. Washietl S, Hofacker IL, Stadler PF (2005) Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci USA 102: 2454–2459
123. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562
124. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678
125. Wolfe KH, Sharp PM, Li WH (1989) Mutation rates differ among regions of the mammalian genome. Nature 337: 283–285
126. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature 434:338–345
127. Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HC, Kim C, Nguyen M, Pham P, Cheuk R, Karlin-Newmann G, Liu SX, Lam B, Sakano H, Wu T, Yu G, Miranda M, Quach HL, Tripp M, Chang CH, Lee JM, Toriumi M, Chan MM, Tang CC, Onodera CS, Deng JM, Akiyama K, Ansari Y, Arakawa T, Banh J, Banno F, Bowser L, Brooks S, Carninci P, Chao Q, Choy N, Enju A, Goldsmith AD, Gurjal M, Hansen NF, Hayashizaki Y, Johnson-Hopson C, Hsuan VW, Iida K, Karnes M, Khan S, Koesema E, Ishida J, Jiang PX, Jones T, Kawai J, Kamiya A, Meyers C, Nakajima M, Narusaka M, Seki M, Sakurai T, Satou M, Tamse R, Vaysberg M, Wallender EK, Wong C, Yamamura Y, Yuan S, Shinozaki K, Davis RW, Theologis A, Ecker JR (2003) Empirical analysis of transcriptional activity in the Arabidopsis genome. Science 302:842–846
128. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. Genome Res 14:1107–1118
129. Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ (2005) Genome-scale identification of nucleosome positions in S. cerevisiae. Science 309:626–630

130. Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L (2003) Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nat Genet 35:57–64
131. Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM, Barrett JC, Shields B, Morris AP, Ellard S, Groves CJ, Harries LW, Marchini JL, Owen KR, Knight B, Cardon LR, Walker M, Hitman GA, Morris AD, Doney AS, Burton PR, Clayton DG, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, Ouwehand WH, Samani NJ, Todd JA, Donnelly P, Davison D, Easton D, Evans D, Leung HT, Spencer CC, Tobin MD, Attwood AP, Boorman JP, Cant B, Everson U, Hussey JM, Jolley JD, Knight AS, Koch K, Meech E, Nutland S, Prowse CV, Stevens HE, Taylor NC, Walters GR, Walker NM, Watkins NA, Winzer T, Jones RW, McArdle WL, Ring SM, Strachan DP, Pembrey M, Breen G, St Clair D, Caesar S, Gordon-Smith K, Jones L, Fraser C, Green EK, Grozeva D, Hamshere ML, Holmans PA, Jones IR, Kirov G, Moskvina V, Nikolov I, O'Donovan MC, Owen MJ, Collier DA, Elkin A, Farmer A, Williamson R, McGuffin P, Young AH, Ferrier IN, Ball SG, Balmforth AJ, Barrett JH, Bishop DT, Iles MM, Maqbool A, Yuldasheva N, Hall AS, Braund PS, Dixon RJ, Mangino M, Stevens S, Thompson JR, Bredin F, Tremelling M, et al. (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 316:1336–1341
132. Zhang Z, Hesselberth JR, Fields S (2007) Genome-wide identification of spliced introns using a tiling microarray. Genome Res 17:503–509

Fuzzy Logic

Lotfi A. Zadeh
Department of EECS, University of California, Berkeley, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Conceptual Structure of Fuzzy Logic
The Basics of Fuzzy Set Theory
The Concept of Granulation
The Concepts of Precisiation and Cointensive Precisiation
The Concept of a Generalized Constraint
Principal Contributions of Fuzzy Logic
A Glimpse of What Lies Beyond Fuzzy Logic
Bibliography

Glossary

Cointension A qualitative measure of proximity of meanings/input-output relations.

Extension principle A principle which relates to propagation of generalized constraints.

f-validity fuzzy validity.

Fuzzy if-then rule A rule of the form: if X is A then Y is B. In general, A and B are fuzzy sets.

Fuzzy logic (FL) A precise logic of imprecision, uncertainty and approximate reasoning.

Fuzzy logic gambit Exploitation of tolerance for imprecision through deliberate m-imprecisiation followed by mm-precisiation.

Fuzzy set A class with a fuzzy boundary.

Generalized constraint A constraint of the form X isr R, where X is the constrained variable, R is the constraining relation and r is an indexical variable which defines the modality of the constraint, that is, its semantics. In general, generalized constraints have elasticity.

Generalized constraint language A language generated by combination and propagation of generalized constraints.

Graduation Association of a scale of degrees with a fuzzy set.

Granuland Result of granulation.

Granular variable A variable which takes granules as variables.

Granulation Partitioning of an object/set into granules.

Granule A clump of attribute values drawn together by indistinguishability, equivalence, similarity, proximity or functionality.

Linguistic variable A granular variable with linguistic labels of granular values.

m-precision Precision of meaning.

mh-precisiand m-precisiand which is described in a natural language (human-oriented).

mm-precisiand m-precisiand which is described in a mathematical language (machine-oriented).

p-validity provable validity.

Precisiand Result of precisiation.

Precisiend Object of precisiation.

v-precision Precision of value.

Definition of the Subject

Viewed in a historical perspective, fuzzy logic is closely related to its precursor – fuzzy set theory [70]. Conceptually, fuzzy logic has a much broader scope and a much higher level of generality than traditional logical systems, among them the classical bivalent logic, multivalued logics, model logics, probabilistic logics, etc. The principal objective of fuzzy logic is formalization – and eventual mechanization – of two remarkable human capabilities. First, the capability to converse, communicate, reason and make ra-

tional decisions in an environment of imprecision, uncertainty, incompleteness of information, partiality of truth and partiality of possibility. And second, the capability to perform a wide variety of physical and mental tasks – such as driving a car in city traffic and summarizing a book – without any measurement and any computations.

A concept which has a position of centrality in fuzzy logic is that of a fuzzy set. Informally, a fuzzy set is a class with a fuzzy boundary, implying a gradual transition from membership to nonmembership. A fuzzy set is precisiated through graduation, that is, through association with a scale of grades of membership. Thus, membership in a fuzzy set is a matter of degree. Importantly, in fuzzy logic everything is or is allowed to be graduated, that is, be a matter of degree. Furthermore, in fuzzy logic everything is or is allowed to be granulated, with a granule being a clump of attribute-values drawn together by indistinguishability, equivalence, similarity, proximity or functionality. Graduation and granulation form the core of fuzzy logic. Graduated granulation is the basis for the concept of a linguistic variable – a variable whose values are words rather than numbers [73]. The concept of a linguistic variable is employed in almost all applications of fuzzy logic.

During much of its early history, fuzzy logic was an object of controversy stemming in part from the pejorative connotation of the term "fuzzy". In reality, fuzzy logic is not fuzzy. Basically, fuzzy logic is a precise logic of imprecision and uncertainty.

An important milestone in the evolution of fuzzy logic was the development of the concept of a linguistic variable and the machinery of fuzzy if-then rules [73,90]. Another important milestone was the conception of possibility theory [79]. Possibility theory and probability theory are complimentary. A further important milestone was the development of the formalism of computing with words (CW) [93]. Computing with words opens the door to a wide-ranging enlargement of the role of natural languages in scientific theories.

In the following, fuzzy logic is viewed in a nontraditional perspective. In this perspective, the cornerstones of fuzzy logic are graduation, granulation, precisisation and the concept of a generalized constraint. The concept of a generalized constraint serves to precisiate the concept of granular information. Granular information is the basis for granular computing (GrC) [2,29,37,81,91,92]. In granular computing the objects of computation are granular variables, with a granular value of a granular variable representing an imprecise and/or uncertain information about the value of the variable. In effect, granular computing is the computational facet of fuzzy logic. GrC and CW are closely related. In coming years, GrC and CW are likely to play increasingly important roles in the evolution of fuzzy logic and its applications.

Introduction

Science deals not with reality but with models of reality. In large measure, scientific progress is driven by a quest for better models of reality.

In the real world, imprecision, uncertainty and complexity have a pervasive presence. In this setting, construction of better models of reality requires a better understanding of how to deal effectively with imprecision, uncertainty and complexity. To a significant degree, development of fuzzy logic has been, and continues to be, motivated by this need.

In essence, logic is concerned with formalization of reasoning. Correspondently, fuzzy logic is concerned with formalization of fuzzy reasoning, with the understanding that precise reasoning is a special case of fuzzy reasoning.

Humans have many remarkable capabilities. Among them there are two that stand out in importance. First, the capability to converse, communicate, reason and make rational decisions in an environment of imprecision, uncertainty, incompleteness of information, partiality of truth and partiality of possibility. And second, the capability to perform a wide variety of physical and mental tasks – such as driving a car in heavy city traffic and summarizing a book – without any measurements and any computations. In large measure, fuzzy logic is aimed at formalization, and eventual mechanization, of these capabilities. In this perspective, fuzzy logic plays the role of a bridge from natural to machine intelligence.

There are many misconceptions about fuzzy logic. A common misconception is that fuzzy logic is fuzzy. In reality, fuzzy logic is not fuzzy. Fuzzy logic deals precisely with imprecision and uncertainty. In fuzzy logic, the objects of deduction are, or are allowed to be fuzzy, but the rules governing deduction are precise. In summary, fuzzy logic is a precise system of reasoning, deduction and computation in which the objects of discourse and analysis are associated with information which is, or is allowed to be, imprecise, uncertain, incomplete, unreliable, partially true or partially possible. For illustration, here are a few simple examples of reasoning in which the objects of reasoning are fuzzy.

First, consider the familiar example of deduction in Aristotelian, bivalent logic.

all men are mortal
Socrates is a man

Socrates is mortal

In this example, there is no imprecision and no uncertainty. In an environment of imprecision and uncertainty, an analogous example – an example drawn from fuzzy logic – is

$$\frac{\begin{array}{l}\text{most Swedes are tall}\\ \text{Magnus is a Swede}\end{array}}{\text{it is likely that Magnus is tall}}$$

with the understanding that Magnus is an individual picked at random from a population of Swedes. To deduce the answer from the premises, it is necessary to precisiate the meaning of "most" and "tall," with "likely" interpreted as a fuzzy probability which, as a fuzzy number, is equal to "most". This simple example points to a basic characteristic of fuzzy logic, namely, in fuzzy logic precisiation of meaning is a prerequisite to deduction. In the example under consideration, deduction is contingent on precisiation of "most", "tall" and "likely". The issue of precisiation has a position of centrality in fuzzy logic.

In fuzzy logic, deduction is viewed as an instance of question-answering. Let I be an information set consisting of a system of propositions $p_1, \ldots, p_n$, $I = S(p_1, \ldots, p_n)$. Usually, I is a conjunction of $p_1, \ldots, p_n$. Let q be a question. A question-answering schema may be represented as

$$\frac{\begin{array}{l}I\\ q\end{array}}{\text{ans}(q/I)}$$

where ans(q/I) denotes the answer to q given I. The following examples are instances of the deduction schema

$$\frac{\begin{array}{ll}I: & \text{most Swedes are tall}\\ & \text{Magnus is a Swede}\\ q: & \text{what is the probability that Magnus is tall?}\end{array}}{\text{ans}(q/I)\text{ is likely, likely=most}} \tag{1}$$

$$\frac{\begin{array}{ll}I: & \text{most Swedes are tall}\\ q: & \text{what fraction of Swedes are not tall?}\end{array}}{\text{ans}(q/I)\text{ is }(1-\text{most})} \tag{2}$$

$$\frac{\begin{array}{ll}I: & \text{most Swedes are tall}\\ q: & \text{what fraction of Swedes are short?}\end{array}}{\text{ans}(q/I)} \tag{3}$$

$$\frac{\begin{array}{ll}I: & \text{most Swedes are tall}\\ q: & \text{what is the average height of Swedes?}\end{array}}{\text{ans}(q/I)} \tag{4}$$

$$\frac{\begin{array}{ll}I: & \text{a box contains balls of various sizes}\\ & \text{most are small}\\ & \text{there are many more small balls than large balls}\\ q: & \text{what is the probability that a ball}\\ & \text{drawn at random is neither large nor small?}\end{array}}{\text{ans}(q/I)\ .} \tag{5}$$

In these examples, rules of deduction in fuzzy logic must be employed to compute ans(q/I). For (1) and (2) deduction is simple. For (3)–(5) deduction requires the use of what is referred to as the extension principle [70,75]. This principle is discussed in Sect. "The Concept of a Generalized Constraint".

A less simple example of deduction involves interpolation of an imprecisely specified function. Interpolation of imprecisely specified functions, or interpolative deduction for short, plays a pivotal role in many applications of fuzzy logic, especially in the realm of control.

For simplicity, assume that f is a function from reals to reals, $Y = f(X)$. Assume that what is known about f is a collection of input-output pairs of the form

$$^*f = ((^*a_1, {}^*b_1), \ldots, (^*a_n, {}^*b_n))\ ,$$

where $^*a$ is an abbreviation of "approximately a". Such a collection is referred to as a fuzzy graph of f [74]. A fuzzy graph of f may be interpreted as a summary of f. In many applications, a fuzzy graph is described as a collection of fuzzy if-then rules of the form

$$\text{if}\quad X \text{ is } {}^*a_i \quad \text{then} \quad Y \text{ is } {}^*b_i\ , \qquad i = 1, \ldots, n\ .$$

Let a be a value of X. Viewing $^*f$ as an information set, I, interpolation of f may be expressed as a question-answering schema

$$\frac{\begin{array}{l}I: {}^*f\\ q: {}^*f(^*a)\end{array}}{\text{ans}(q/(^*f, {}^*a))}$$

A very simple example of interpolative deduction is the following. Assume that $^*a_1, {}^*a_2, {}^*a_3$ are labeled small, medium and large, respectively. A fuzzy graph of f may be expressed as a calculus of fuzzy if-then rules.

$$\begin{array}{ll}^*f: & \text{if } X \text{ is small then } Y \text{ is small}\\ & \text{if } X \text{ is medium than } Y \text{ is large}\\ & \text{if } X \text{ is large then } Y \text{ is small.}\end{array}$$

Given a value of X, $X = a$, the question is: What is $^*f(^*a)$? The so-called Mamdani rule [39,74,78] provides an answer to this question.

More generally, in an environment of imprecision and uncertainty, fuzzy if-then rules may be of the form

$$\text{if} \quad X \text{ is } {}^*a_i\,, \quad \text{then usually } (Y \text{ is } {}^*b_i)\,, \quad i = 1, \ldots, n\,.$$

Rules of this form endow fuzzy logic with the capability to model complex causal dependencies, especially in the realms of economics, social systems, forecasting and medicine.

What is not widely recognized is that fuzzy logic is more than an addition to the methods of dealing with imprecision, uncertainty and complexity. In effect, fuzzy logic represents a paradigm shift. More specifically, it is traditional to associate scientific progress with progression from perceptions to numbers. What fuzzy logic adds to this capability are four basic capabilities.

(a) Nontraditional. Progression from perceptions to precisiated words
(b) Nontraditional. Progression from unprecisiated words to precisiated words
(c) Countertraditional. Progression from numbers to precisiated words
(d) Nontraditional. Computing with words (CW)/NL-computation.

These capabilities open the door to a wide-ranging enlargement of the role of natural languages in scientific theories.

Our brief discussion of deduction in the context of fuzzy logic is intended to clarify the nature of problems which fuzzy logic is designed to address. The principal concepts and techniques which form the core of fuzzy logic are discussed in the following. Our discussion draws on the concepts and ideas introduced in [102].

Conceptual Structure of Fuzzy Logic

There are many logical systems, among them the classical, Aristotelian, bivalent logic, multivalued logics, model logics, probabilistic logic, logic, dynamic logic, etc. What differentiates fuzzy logic from such logical systems is that fuzzy logic is much more than a logical system.

The point of departure in fuzzy logic is the concept of a fuzzy set. Informally, a fuzzy set is a class with a fuzzy boundary, implying that, in general, transition from membership to nonmembership in a fuzzy set is gradual rather than abrupt. A set is a class with a crisp boundary (Fig. 1). A set, A, in a space U, $U = \{u\}$, is precisiated through association with a characteristic function which maps U into $\{0, 1\}$. More generally, a fuzzy set, A, is precisiated through graduation, that is, through association with A of a membership function, μ_A – a mapping from U to a grade of

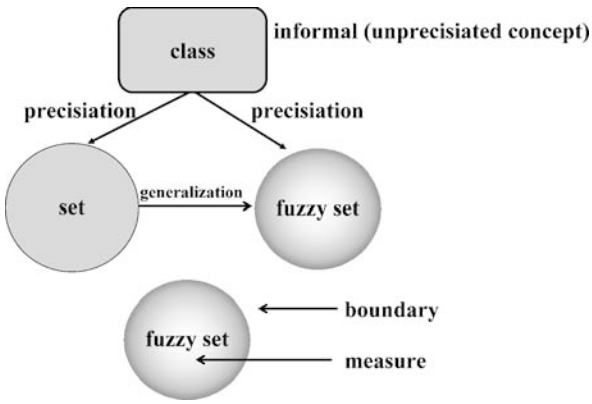

Fuzzy Logic, Figure 1
The concepts of a set and a fuzzy set are derived from the concept of a class through precisiation. A fuzzy set has a fuzzy boundary. A fuzzy set is precisiated through graduation

membership space, G, with $\mu_A(u)$ representing the grade of membership of u in A. In other words, membership in a fuzzy set is a matter of degree. A familiar example of graduation is the association of Richter scale with the class of earthquakes. A fuzzy set is basic if G is the unit interval. More generally, G may be a partially ordered set. L-fuzzy sets [23] fall into this category. A basic fuzzy set is of Type 1. A fuzzy set, A, is of Type 2 if $\mu_A(u)$ is a fuzzy set of Type 1. Recursively, a fuzzy set, A, is of Type n if $\mu_A(u)$ is a fuzzy set of Type $n-1$, $n = 2, 3, \ldots$ [75]. Fuzzy sets of Type 2 have become an object of growing attention in the literature of fuzzy logic [40]. Unless stated to the contrary, a fuzzy set is assumed to be of Type 1 (basic).

Note. A clarification is in order. Consider a concatenation of two words, A and B, with A modifying B, e. g. A is an adjective and B is a noun. Usually, A plays the role of an s-modifier, that is, a modifier which specializes B in the sense that AB is a subset of B, as in convex set. In some instances, however, A plays the role of a g-modifier, that is, a modifier which generalizes B. In this sense, fuzzy in fuzzy set, fuzzy logic and, more generally, in fuzzy B, is a g-modifier. Examples: fuzzy topology, fuzzy measure, fuzzy arithmetic, fuzzy stability, etc. Many misconceptions about fuzzy logic are rooted in incorrect interpretation of fuzzy as an s-modifier.

What is important to note is that (a) $\mu_A(u)$ is name-based (extensional) if u is the name of an object in U, e. g., $\mu_{\text{middle-aged}}(\text{Vera})$; (b) $\mu_A(u)$ is attribute-based (intensional) if the grade of membership is a function of an attribute of u, e. g., age; and (c) $\mu_A(u)$ is perception-based if u is a perception of an object, e. g., Vera's grade of membership in the class of middle-aged women, based on her appearance, is 0.8.

It should be observed that a class of objects, A, has two basic attributes: (a) the boundary of A; and (b) the cardinality (count) or, more generally, the measure of A (Fig. 1). In this perspective, fuzzy set theory is, in the main, boundary-oriented, while probability theory is, in the main, measure-oriented. Fuzzy logic is, in the main, both boundary- and measure-oriented. The concept of a membership function is the centerpiece of fuzzy set theory.

With the concept of a fuzzy set as the point of departure, different directions may be pursued, leading to various facets of fuzzy logic. More specifically, the following are the principal facets of fuzzy logic: the logical facet, FLl; the fuzzy-set-theoretic facet, FLs, the epistemic facet, Fle; and the relational facet, FLr (Fig. 2).

The logical facet of FL, FLl, is fuzzy logic in its narrow sense. FLl may be viewed as a generalization of multivalued logic. The agenda of FLl is similar in spirit to the agenda of classical logic [17,22,25,44,45].

The fuzzy-set-theoretic facet, FLs, is focused on fuzzy sets. The theory of fuzzy sets is central to fuzzy logic. Historically, the theory of fuzzy sets [70] preceded fuzzy logic [77]. The theory of fuzzy sets may be viewed as an entry to generalizations of various branches of mathematics, among them fuzzy topology, fuzzy measure theory, fuzzy graph theory, fuzzy algebra and fuzzy differential equations. Note that fuzzy X is a fuzzy-set-theory-based or, more generally, fuzzy-logic-based generalization of X.

The epistemic facet of FL, FLe, is concerned with knowledge representation, semantics of natural languages and information analysis. In FLe, a natural language is viewed as a system for describing perceptions. An important branch of FLe is possibility theory [15,79,83]. Another important branch of FLe is the computational theory of perceptions [93,94,95].

The relational facet, FLr, is focused on fuzzy relations and, more generally, on fuzzy dependencies. The concept of a linguistic variable – and the associated calculi of fuzzy

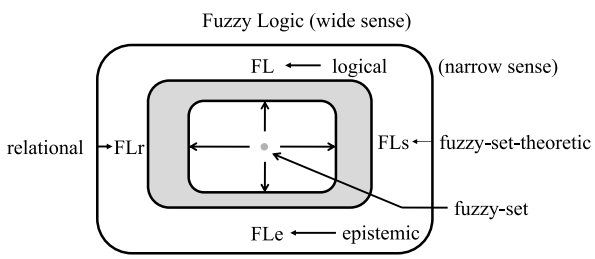

Fuzzy Logic, Figure 2
Principal facets of fuzzy logic (FL). The nucleus of fuzzy logic is the concept of a fuzzy set

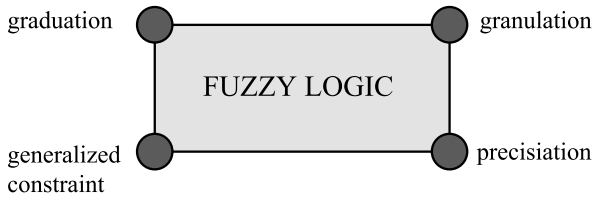

Fuzzy Logic, Figure 3
The cornerstones of a nontraditional view of fuzzy logic

if-then rules – play pivotal roles in almost all applications of fuzzy logic [1,3,6,8,12,13,18,26,28,32,36,40,48,52,65,66,67,69].

The cornerstones of fuzzy logic are the concepts of graduation, granulation, precisiation and generalized constraints (Fig. 3). These concepts are discussed in the following.

The Basics of Fuzzy Set Theory

The grade of membership, $\mu_A(u)$, of u in A may be interpreted in various ways. It is convenient to employ the proposition p: Vera is middle-aged, as an example, with middle-age represented as a fuzzy set shown in Fig. 4. Among the various interpretations of p are the following.

Assume that q is the proposition: Vera is 43 years old; and r is the proposition: the grade of membership of Vera in the fuzzy set of middle-aged women is 0.8.

(a) The truth value of p given r is 0.8.
(b) The possibility of q given p and r is 0.8.
(c) The degree to which the concept of middle-age has to be stretched to apply to Vera is (1–0.8).
(d) 80% of voters in a voting group vote for p given q and r.
(e) The probability of p given q and r is 0.8.

Of these interpretations, (a)–(c) are most compatible with intuition.

If A and B are fuzzy sets in U, then their intersection (conjunction), $A \cap B$, and union (disjunction), $A \cup B$, are defined as

$$\mu_{A\cap B}(u) = \mu_A(u) \wedge \mu_B(u)\,, \quad u \in U$$
$$\mu_{A\cup B}(u) = \mu_A(u) \vee \mu_B(u)\,, \quad u \in U$$

where $\wedge = \min$ and $\vee = \max$. More generally, conjunction and disjunction are defined through the concepts of t-norm and t-conorm, respectively [48].

When U is a finite set, $U = \{u_1, \ldots, u_n\}$, it is convenient to represent A as a union of fuzzy singletons,

$$\mu_A(u_i)/u_i\,, \quad i = 1, \ldots, n\,.$$

Specifically, $\quad A = \mu_A(u_1)/u_1 + \cdots + \mu_A(u_n)/u_n$

- **Imprecision of meaning = elasticity of meaning**
- **Elasticity of meaning = fuzziness of meaning**

Example: middle-age

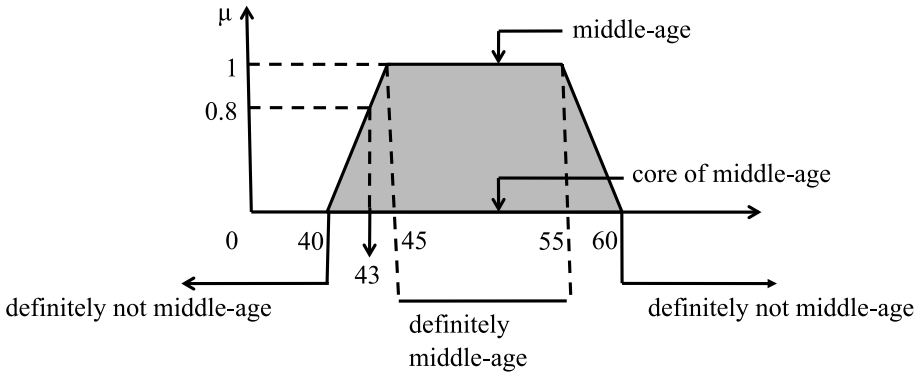

Fuzzy Logic, Figure 4
Precisiation of middle-age through graduation

in which + denotes disjunction. More compactly,

$$A = \Sigma_i \mu_A(u_i)/u_i \,, \quad i = 1, \ldots, n \,.$$

When U is a continuum, A may be expressed as

$$\int_U \mu_A(u)/u \,.$$

A basic concept in fuzzy set theory is that of a level set [70], commonly referred to as an α-cut [48]. Specifically, if A is a fuzzy set in U, then an α-cut, A_α, is defined as (Fig. 5)

$$A_\alpha = \{u | \mu_A(u) \geqslant \alpha\} \,, \quad 0 < \alpha \leqslant 1 \,.$$

The core of A is the α-cut with $\alpha = 1$.

A fuzzy set, A, is a fuzzy interval if all of its α-cuts are intervals. A fuzzy set, A, is convex if all of its α-cuts are convex [70].

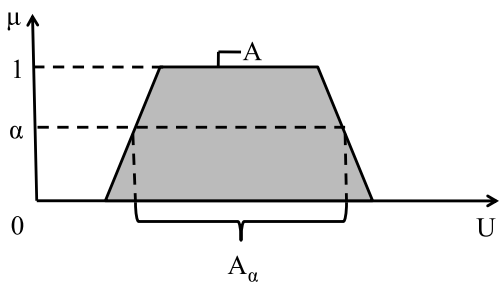

Fuzzy Logic, Figure 5
Definition of α-cut

In most practical applications, a membership function is assumed to have a simple form, e. g. a triangle or a trapezoid (Fig. 6). For convenience, triangular and trapezoidal fuzzy sets are abbreviated as tr-sets and tp-sets, respectively. It should be noted that a singleton is associated with one parameter, an interval with two parameters, a tr-set with three parameters and a tp-set with four parameters. The number of parameters may be interpreted as the number of degrees of freedom. In general, the number of degrees of freedom is covariant with closeness of approximation.

A concept which plays an important role in fuzzy set theory is that of cardinality, that is, the count of elements in a fuzzy set [25,51,64,86]. Basically, there are two ways in which cardinality can be defined: (a) crisp (scaler) cardinality and (b) fuzzy (stratified) cardinality. In the case of (a), the count of elements in a fuzzy set is a crisp number; in the case of (b) it is a fuzzy set.

More specifically, consider a fuzzy set, A, defined in $U = \{u_1, \ldots, u_n\}$ through its membership function,

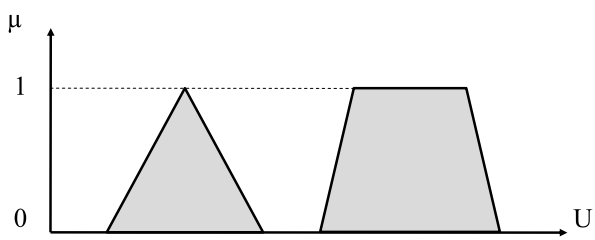

Fuzzy Logic, Figure 6
Triangular and trapezoidal fuzzy sets. tr-sets and tp-sets

$\mu_A(u)$. The sigma-count of A is defined as

$$\Sigma\text{Count}(A) = \sum_{i=1}^{n} \mu_A(u_i) \,.$$

If A and B are fuzzy sets defined in U, then the relative sigma-count, $\Sigma\text{Count}(A/B)$, is defined as

$$\Sigma\text{Count}(A/B) = \frac{\Sigma\text{Count}(A \cap B)}{\Sigma\text{Count}(B)} = \frac{\sum_{i=1}^{n} \mu_A(u_i) \wedge \mu_B(u_i)}{\sum_{i=1}^{n} \mu_B(u_i)} \,,$$

where $\wedge = \min$, and summations are arithmetic.

A stratified count, $\text{S Count}(A)$, is a fuzzy set. As such, it is more informative then $\Sigma\text{Count}(A)$ but has the disadvantage of greater complexity. A stratified count is defined in terms of α-cuts of A [86]. Specifically, let $A_{\alpha 1}, \ldots, A_{\alpha n}$ be α-cuts of A, with $\alpha_1 < \alpha_2 < \cdots < \alpha_n \leq 1$. Then

$$\text{S Count}(A) = \alpha_1/\text{Count}(A_{\alpha 2}) + \cdots + \alpha_n/\text{Count}(A_{\alpha n})$$

or equivalently,

$$\text{S Count}(A) = \{(\alpha_i, A_{\alpha i})\} \,, \quad i = 1, \ldots, n \,.$$

As a simple illustration, consider the fuzzy set

$$A = 0.4/u_1 + 0.8/u_2 + 1/u_3 + 0.9/u_4 + 0.3/u_5 \,.$$

In this case,

$$\begin{aligned}
A_{0.3} &= \{u_1, u_2, u_3, u_4, u_5\} \,, & \text{Count}(A_{0.3}) &= 5 \\
A_{0.4} &= \{u_1, u_2, u_3, u_4\} \,, & \text{Count}(A_{0.4}) &= 4 \\
A_{0.8} &= \{u_2, u_3, u_4\} \,, & \text{Count}(A_{0.8}) &= 3 \\
A_{0.9} &= \{u_3, u_4\} \,, & \text{Count}(A_{0.9}) &= 2 \\
A_1 &= \{u_3\} \,, & \text{Count}(A_1) &= 1
\end{aligned}$$

$$\text{S Count}(A) = 1/1 + 0.9/2 + 0.8/3 + 0.4/4 + 0.3/5$$

$$\Sigma\text{Count}(A) = 3.4 \,.$$

The concept of a membership function has a position of centrality in fuzzy logic. In this context, a natural question is: How can a membership function be constructed? There are three basic approaches: graduation through declaration; graduation through composition/deduction; and graduation through exemplification/ostention/elicitation.

Graduation through declaration is employed in many applications of fuzzy logic, especially in the realm of control systems. In this mode of graduation, a membership function is declared (specified) by the designer of a system. An example is the Honda fuzzy logic transmission (Fig. 7).

Declarative graduation has the potential for important applications in the realm of precisiation/standardization of many concepts and terms in various fields of science, especially in the realms of medicine and economics. A very simple example is standardization of the meaning of normal temperature, mild fever, high fever, etc. It is convenient to standardize the meaning of such terms through

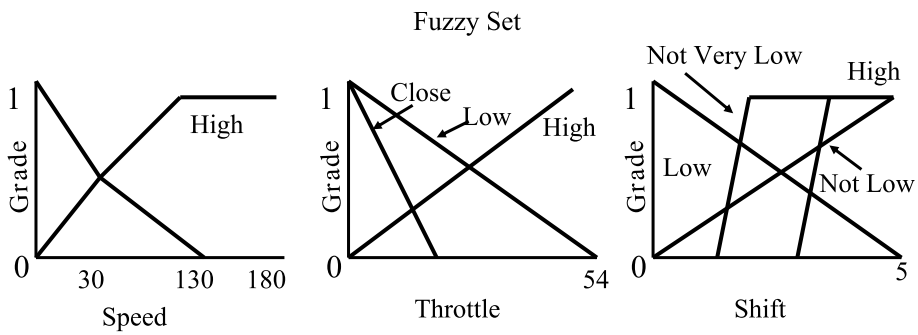

Control fuzzy if-then rules:

1. If (speed is low) and (shift is high) then (-3)
2. If (speed is high) and (shift is low) then (+3)
3. If (throt is low) and (speed is high) then (+3)
4. If (throt is low) and (speed is low) then (+1)
5. If (throt is high) and (speed is high) then (-1)
6. If (throt is high) and (speed is low) then (-3)

Fuzzy Logic, Figure 7
Declared membership functions and associated fuzzy if-then rules in Honda's fuzzy logic transmission

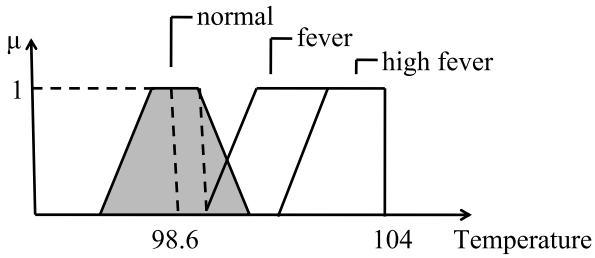

Fuzzy Logic, Figure 8
Declarative graduation/standardization of normal temperature, fever and high fever

the use of trapezoidal membership functions, that is, tp-sets (Fig. 8).

In many cases, a membership function is associated with a number of adjustable parameters, e. g. the four parameters which define a tp-set. The parameters are adjusted through experimentation or self-learning. For this purpose, techniques drawn from neurocomputing and evolutionary computing are commonly employed [54,55].

In construction through composition/deduction, a membership function is composed from other membership functions. Composition/deduction may be simple e. g. a combination of conjunctive or disjunctive operations. More generally, composition/deduction is associated with a chain of deductions which involve generalized constraint propagation. Propagation of generalized constraints is discussed in Sect. "The Concept of a Generalized Constraint".

A special case of composition which plays an important role in fuzzy set theory involves the concept of a convex combination [70]. More specifically, let $A_1, \ldots, A_n$ be a collection of crisp or fuzzy sets in U. Let $a_1, \ldots, a_n$ be numbers in [0,1] which add up to unity. A fuzzy set, A, is a convex combination of the A_i, $i = 1, \ldots, n$, if

$$A = a_1 A_1 + \cdots + a_n A_n \,,$$

with the understanding that

$$\mu_A(u) = a_1 \mu_{A_1}(u) + \cdots + a_n \mu_{A_n}(u) \,.$$

What is important to note is that a convex combination of crisp sets is a fuzzy set. If the coefficients $a_1, \ldots, a_n$ are interpreted as probabilities, then A may be interpreted as the expected value of a random set. This connection between fuzzy sets and random sets has been an object of considerable attention in the literature of fuzzy logic [24,46,47].

Underlying graduation through exemplification is the remarkable human capability to rank-order perceptions. It should be noted that humans learn the meaning of terms and concepts mostly through exemplification.

It is convenient to employ an example to describe graduation through exemplification. Assume that A tells B that Vera is middle-aged. B can elicit A's meaning of middle-aged by asking A to mark on the scale $[0, 1]$ the degree to which a particular age, say 43, fits A's meaning of middle-aged. The process is repeated for various values of age. Eventually, the collected data are employed to approximate to A's meaning of middle-aged by a trapezoidal membership function. In the case of fuzzy sets of Type 2, the mark is a fuzzy point.

A basic feature of fuzzy logic is: In fuzzy logic everything is or is allowed to be graduated.

The Concept of Granulation

The concept of granulation is unique to fuzzy logic and plays a pivotal role in its applications [81,91]. The concept of granulation is inspired by the way in which humans deal with imprecision, uncertainty and complexity.

Granulation is rooted in the concept of a granule. Informally, a granule, G, in a universe of discourse, U, is a clump of elements of U which are drawn together by indistinguishability, equivalence, similarity, proximity or functionality (Fig. 9). For example, an interval is a granule. So is a Gaussian distribution, and so is a fuzzy interval. A granule, G, is precisiated through association of G with a generalized constraint – a concept which will be defined in Sect. "The Concept of a Generalized Constraint". The concept of a generalized constraint is more general than that of a membership function.

A singular variable takes singletons in U as values. A granular variable, X, is a variable which takes granules as values. For example, age is a granular variable if its val-

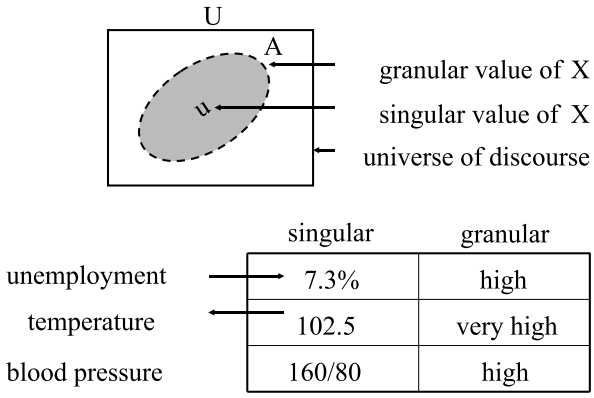

Fuzzy Logic, Figure 9
Singular and granular values

ues are young, middle-aged and old. A linguistic variable is a granular variable whose granular values carry linguistic labels (Fig. 10). The concept of a linguistic variable [73,75] is employed in almost all applications of fuzzy logic.

Granulation is an operation which maps singletons in U into granules. Granulation applied to a singular variable, X, transforms X into a granular variable, $^*X$. For example, if age is a real-valued variable taking values in the interval [0, 120], granulation of X transforms X into a linguistic variable, $^*X$, with values young, middle-aged and old. When convenient, the result of granulation is referred to as the granuland.

Granulation of a set, A, results in a partition of A into granules. For example, granulation applied to the interval [0, l20] results in the granules young, middle-aged and old.

The membership function of a fuzzy set takes values in the interval [0, 1]. Not infrequently, the grade of membership is not known precisely. In this case, it may be expedient to granulate the interval [0, 1], with the granular values of membership being zero, low, medium, high and 1. Such granular values of membership are particularly useful in dealing with fuzzy sets of Type 2.

Granulation may be applied to arbitrarily complex objects. Application of granulation to an expression involves replacement of one or more singular variables in the expression with granular variables. For example, if

$$Z = X + Y$$

is an arithmetic expression, then its granuland may be expressed as

$$^*Z = {}^*X + {}^*Y$$

in which the starred variables are granular variables. In this sense, interval arithmetic may be viewed as the result of granulation of numerical arithmetic. Figure 11 shows an application of granulation to a function, f. The result of

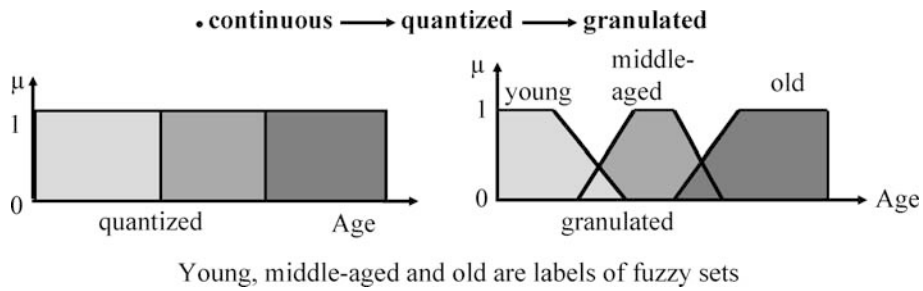

Fuzzy Logic, Figure 10
Granulation of age; young, middle-aged and old are linguistic (granular) values of age

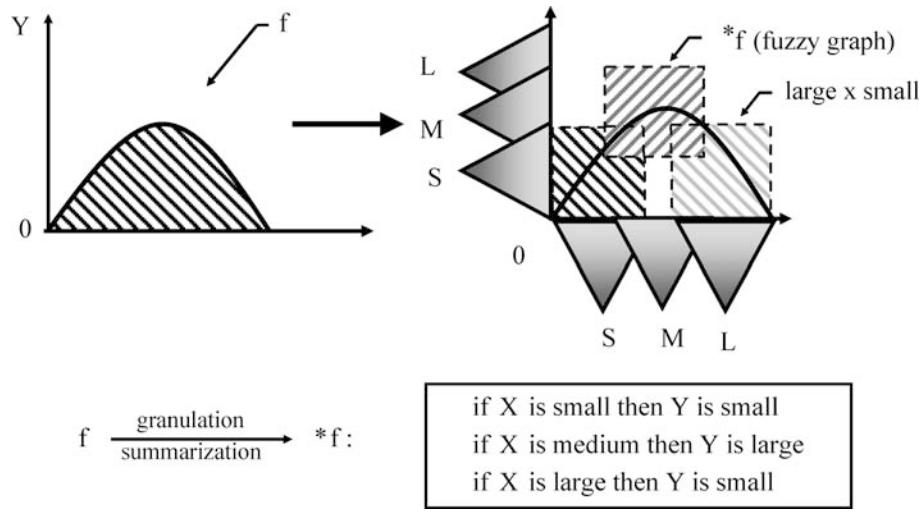

Fuzzy Logic, Figure 11
Granulation of a function. *S* (small), *M* (medium) and *L* (large) are fuzzy sets. The granuland of *f*, $^*f$, may be viewed as a summary of *f*

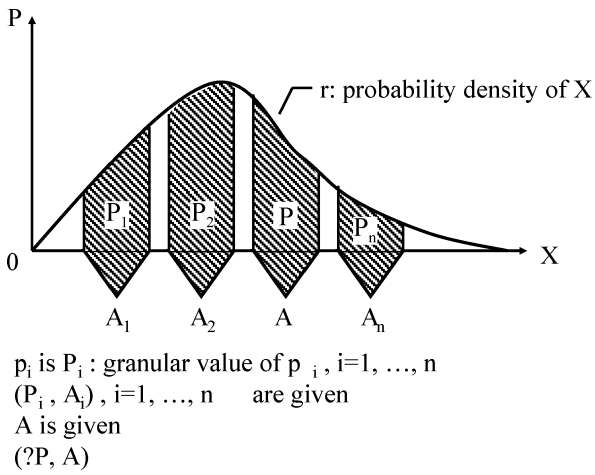

Fuzzy Logic, Figure 12
Granulation and interpolation of a probability distribution

granulation, $^*f$, is the fuzzy graph of f [74,78]. The fuzzy graph of f may be described as a collection of fuzzy if-then rules; $^*f$ may be viewed as a summary of f. Describing a function as a collection of fuzzy if-then rules may be regarded as a form of information compression.

Application of granulation to probability distributions is illustrated in Fig. 12. The granules of X play the role of fuzzy events and the P_i are their granular probabilities [99].

It should be pointed out that in the case of probability distributions it is necessary to differentiate between granular probability distributions and granule-valued probability distributions [96,101] (Fig. 13). It should be noted that there is a connection between the concept of a granular probability distribution and the notion of Perfilieva transform [49].

An instance of a granule-valued distribution is a random set. There is a close connection between granule-valued distributions and the Dempster–Shafer theory of evidence [11,56,57]. More about granular and granule-valued distributions will be said in Sect. "The Concept of a Generalized Constraint".

In addition to the concepts of graduation and granulation there are two basic concepts which play important roles in fuzzy logic. These are the concepts of precisiation [87,97] and generalized constraint [89,93]. These concepts along with the concepts of graduation and granulation form the cornerstones of fuzzy logic. The concepts of precisiation and cointensive precisiation are discussed in the following section. The concept of a generalized constraint is discussed in Sect. "The Concept of a Generalized Constraint".

The Concepts of Precisiation and Cointensive Precisiation

There are not many concepts in science that are as pervasive as the concept of precision. There is an enormous literature. And yet, there are some important facets of the concept of precision which have received little if any attention. In particular, an issue that appears to have been overlooked relates to the need for differentiation between two forms of precision: precision of value, v-precision,

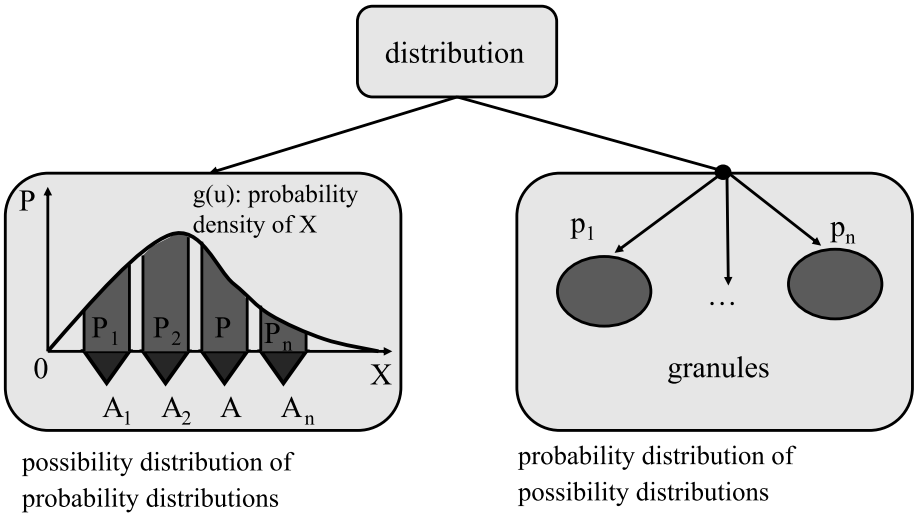

Fuzzy Logic, Figure 13
Granular vs. granule-valued distributions

and precision of meaning, m-precision [101]. Specifically, consider a variable, X, whose value is not known precisely. In this event, the proposition $a \leqslant X \leqslant b$, where a and b are precisely specified numbers, is v-imprecise and m-precise. Similarly, the proposition: X is a normally distributed random variable with mean a and variance b, is v-imprecise and m-precise. On the other hand, the proposition: X is small, is both v-imprecise and m-imprecise if small is not defined precisely. If small is defined precisely as a fuzzy set, then the proposition in question is v-imprecise and m-precise.

Informally, precisiation is an operation which transforms an object, p, into another object, p^*, which is more precisely defined, in some specified sense, than p. The reverse applies to imprecisiation. In the realm of our discourse p is a proposition, predicate, question, command or, more generally, a linguistic expression which has a semantic identity. Unless stated to the contrary, p will be assumed to be a proposition. For convenience, the object and the result of precisiation are referred to as precisiend and precisiand, respectively (Fig. 14).

As in the case of precision/imprecision, there is a need for differentiation between v-precisiation/imprecisiation and m-precisiation/imprecisiation. Example:

$$X = 5 \xrightarrow[\text{m-imprecisiation}]{\text{v-imprecisiation}} X \text{ is small}$$

$$X \text{ is small} \xrightarrow{\text{m-precisiation}} X \text{ is small}$$

(small is defined as a fuzzy set) .

It should be noted that data compression and summarization are forms of v-imprecisiation.

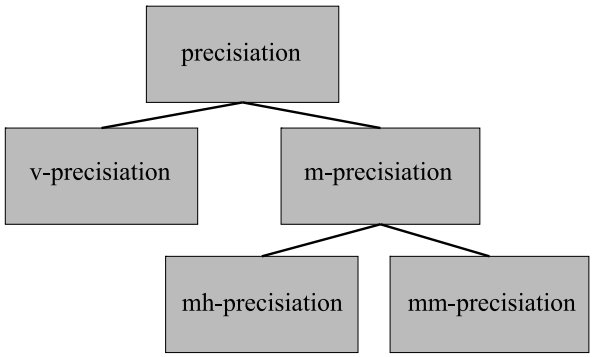

Fuzzy Logic, Figure 15
Modalities of precisiation. mh-precisiation = nonmathematical precisiation of meaning; mm-precisiation = mathematical precisiation of meaning

In the case of m-precisiation, there is a need for additional differentiation between m-precisiation which is human-oriented (non-mathematical), or mh-precisiation for short, and m-precisiation which is machine-oriented (mathematical), or mm-precisiation for short (Fig. 15). In this sense, a dictionary definition is a form of mh-precisiation, with the definiens and the definiendum playing the roles of precisiend and precisiand, respectively. A mathematical definition of a concept, say stability, is an instance of mm-precisiation.

A convenient illustration of mh-precisiation and mm-precisiation is provided by the concept of "bear market".

mh-precisiand declining stock market with expectation of further decline

mm-precisiand 30 percent decline after 50 days, or a 13 percent decline after 145 days (Robert Shuster).

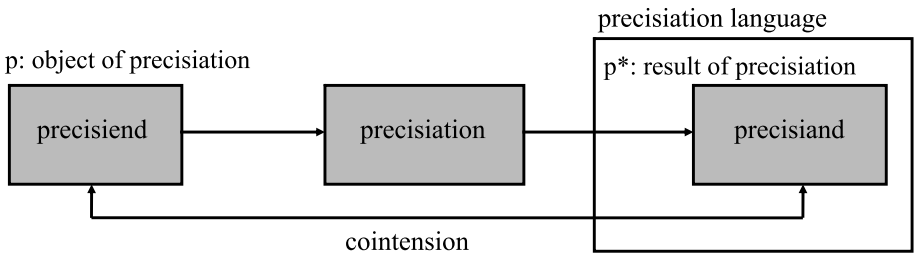

precisiand = model of meaning
cointension = measure of closeness of meanings of p and p*
A precisiend may have many precisiands.

precisiation = translation into precisiation language

Fuzzy Logic, Figure 14
Basic concepts relating to precisiation and cointension

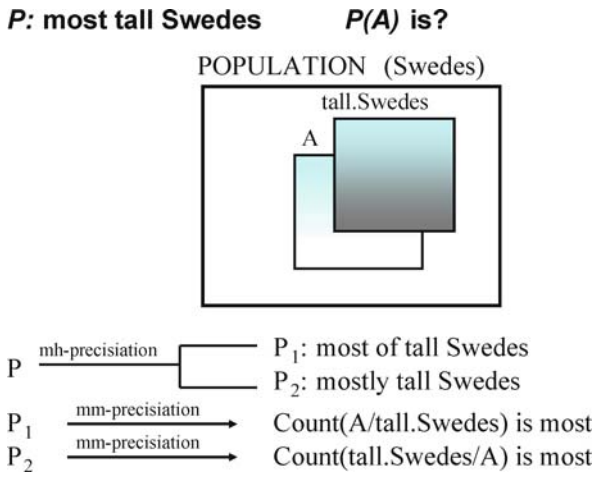

Fuzzy Logic, Figure 16
Disambiguation of *P*: most tall Swedes

It is of interest to note that disambiguation is a form of m-precisiation. As an illustration, alternative interpretations of the predicate, *P*: Most tall Swedes, are shown in Fig. 16.

So far as v-imprecisiation is concerned, there is a need for differentiation between (a) v-imprecisiation which is forced; and (b) v-imprecisiation which is deliberate. To illustrate, if the value of X is described imprecisely because the precise value of X is not known, then what is involved is forced v-imprecisiation. On the other hand, if the value of X is described imprecisely even though the precise value of X is known, then v-imprecisiation is deliberate.

What is the point of deliberate v-imprecisiation? The rationale is that in many cases precision carries a cost. In such cases, deliberate v-imprecisiation serves a useful purpose if it provides a way of reducing cost. Familiar examples of deliberate v-imprecisiation are data compression and summarization. As will be seen at a later point, deliberate v-imprecisiation underlies the fuzzy logic gambit (Sect. "The Concept of a Generalized Constraint"). The fuzzy logic gambit plays an important role in many applications of fuzzy logic, especially in the realm of consumer products – a realm in which cost is an important parameter.

A precisiend may be precisiated in a large, perhaps unbounded, number of ways. As an illustration, Fig. 17a and b shows some of the simpler mm-precisiands of the predicate "approximately *a*," or $^*a$ for short. Note that the simplest mm-precisiand is *a*. This very simple mode of mm-precisiation is widely employed in many fields of science. Probability theory is a case in point. Most real-world probabilities are based on perceptions rather than on measurements. Perceptions are intrinsically imprecise. As a consequence, so are most real-world probabilities. And yet, in most computations involving probabilities, probabilities are treated as exact numbers. For example, $^*0.7$ is treated as 0.7000.

The Concept of Cointension

Let p^* be a precisiand of p. It is expedient to associate with p^* two basic metrics: (a) cointension – a qualitative mea-

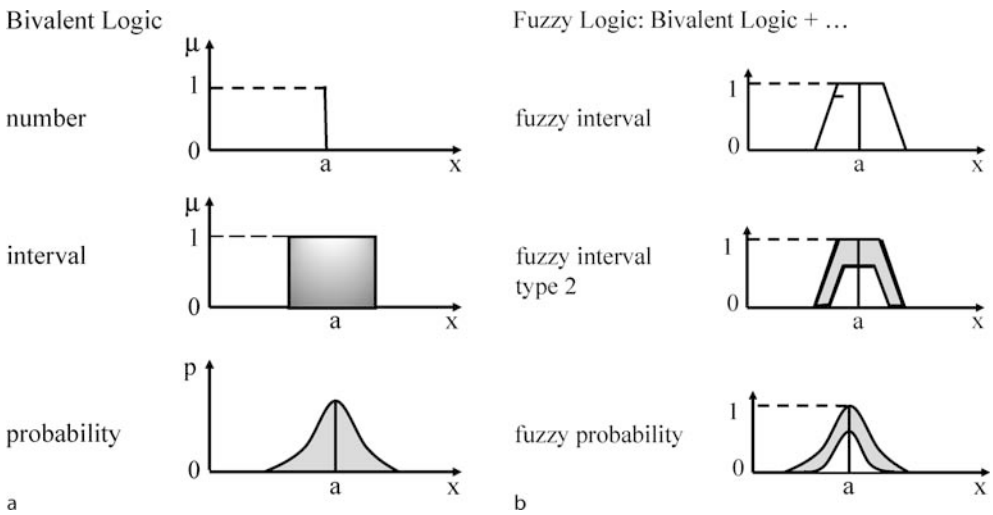

Fuzzy Logic, Figure 17
a Alternative modes of mm-precisiation of "approximately *a*," $^*a$, within the framework of bivalent logic; b Alternative modes of mm-precisiation of "approximately *a*," $^*a$, within the framework of fuzzy logic

sure of the proximity of the meanings of p and p^*; and (b) the computational complexity of p^*. In general, cointension and computational complexity are covariant in the sense that an increase in the cointension of p^* is associated with an increase in the computational complexity of p^*.

Cointension is a new term which is in need of clarification [101]. In logic, intension and extension are defined, respectively, as attribute-based meaning, or i-meaning for short, and name-based meaning, or e-meaning for short [5,10,35]. For our purposes, it is helpful to interpret attribute-based as measurement-based and, name-based as perception-based. What this means is that a precisiend, p, is viewed as a perception of a concept, while its mm-precisiand, p^*, is viewed as a measurement-based definition of p. For example, in the case of the concept of bear market, we have

mm-precisiand 30 percent decline after 50 days, or a 13 percent decline after 145 days (Robert Shuster).

More concretely, if p^* is an mm-precisiand of p, then the cointension of p^* in relation to p, $C(p^*, p)$, is a qualitative measure of the degree of proximity of the i-meanings of p^* and p, or the proximity of the i-meaning of p^* and the e-meaning of p. p^* is cointensive if the degree of proximity is high. In the case of the bear market example, cointension is a measure of the degree to which the mm-precisiand fits our perception of "bear market". Although this degree cannot be assessed precisely, it is evident that the degree is not high.

It should be noted that there is a close analogy between the concept of mm-precisiation and mathematical modeling. More specifically, the analog of mm-precisiation is mathematical modeling; the analog of precisiend is the object of modeling; the analog of precisiand is the model; and the analog of meaning is the input/output relation.

In the context of modeling, cointension is a measure of proximity of the input/output relations of the object of modeling and the model. A model is cointensive if its proximity is high.

The concept of cointension highlights an important issue. Specifically, what should be noted is that, in general, mm-precisiation of p is not the final objective. What matters is the cointension of an mm-precisiand of p. In general, what are sought are mm-precisiands which have high cointension. In other words, the desideratum in not merely mm-precisiation but, more importantly, cointensive mm-precisiation. As will be seen at a later point, a striking feature of fuzzy logic is its high power of cointensive precisiation.

The concept of cointensive mm-precisiation has an important implication for scientific theories. In large measure, scientific theories are based on bivalent logic. As a consequence, definitions of concepts are, generally, bivalent, in the sense that no degrees of truth are allowed. An example is the definition of bear market. The same applies to the definitions of recession, stability, independence, causality, stationarity, etc. The problem is that most of the concepts which are associated with bivalent definitions are in fact fuzzy, that is, are a matter of degree. For instance, the reason why the mm-precisiand of bear market is not cointensive is rooted in the fact that bear market is a fuzzy concept. What can be said in a general way is that bivalent-logic-based definitions of fuzzy concepts cannot be expected to be cointensive, just as linear models of nonlinear systems cannot be expected to be good models. In summary, an important conclusion relating to the concept of cointension may be stated as the Cointension Principle: To Achieve high cointesion it is necessary, in general, to associate a fuzzy precisiend with a fuzzy precisiand.

Note. In the sequel, unless stated to the contrary, precisiation should be understood as cointensive mm-precisiation.

In the foregoing discussion, we talked about mm-precisiation but have not addressed a basic question: How can a proposition, p, be mm-precisiated? In fuzzy logic, a concept which plays a pivotal role in mm-precisiation is that of a generalized constraint. A brief discussion of this concept is presented in the following section.

The Concept of a Generalized Constraint

The concept of a constraint is high on the list of basic concepts in science. There is an extensive literature in the realms of mathematical programming and optimal control. Particularly worthy of note is the rapid growth of interest in constraint programming within computer science and related fields [53].

A basic assumption which is commonly made in the literature is that constraints are hard (inelastic) and are precisely defined. This assumption is not a good fit to reality. In most realistic settings, constraints have some elasticity and are not precisely defined. Familiar examples are:

- Check-out time is l pm. A constraint on check-out time.
- Speed limit is l00 km/h. A constraint on speed.
- Vera has a son in mid-twenties and a daughter in mid-thirties. A constraint on Vera's age.

A first step toward precisiation of such constraints was taken in [4]. A further step was taken in [76]. The concept of a generalized constraint was introduced in [89]. A more detailed description was given in [93] and in [96,98,101].

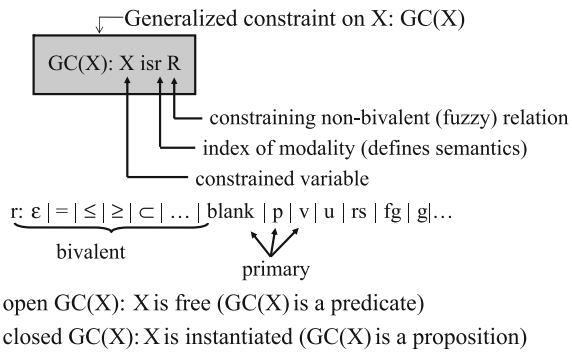

Fuzzy Logic, Figure 18
Principal features of a generalized constraint

The principal features of the concept of generalized constraint are summarized in Fig. 18 and et seq.

Constrained Variable

The constraint variable, X, can assume a variety of forms. Among the principal forms are the following.

X is an n-ary variable, $X = (X_1, \ldots, X_n)$
X is a proposition, e. g., Leslie is tall
X is a relation
X is a function of another variable: $X = f(Y)$
X is conditioned on another variable, X/Y
X is conditioned on another generalized constraint, X isr if Y iss S
X has a structure, e. g., $X =$ Location (Residence(Carol))
X is a generalized constraint, $X : Y$ isr R
X is a group variable. In this case, there is a group, $G : (Name_1, \ldots, Name_n)$, with each member of the group, $Name_i$, $i = 1, \ldots, n$, associated with an attribute-value, h_i, of attribute H. h_i may be vector-valued. Symbolically
$G = (Name_1, \ldots, Name_n)$,
$G[H] = (Name_1/h_1, \ldots, Name_n/h_n)$,
$G[H \text{ is } A] = (\mu_A(h_i)/Name_1, \ldots, \mu_A(h_n)/Name_n)$.
Basically, $G[H]$ is a relation and $G[H \text{ is } A]$ is a fuzzy restriction of $G[H]$ [73,75,76]
The concept of a group variable is closely related to the concept of a fuzzy relation.

Modalities of Generalized Constraints

The indexical variable, r, defines the modality of a generalized constraint, that is, its semantics. The principal modalities are listed below.

$r :=$ equality constraint: $X = R$ is an abbreviation of X is $= R$
$r :\leqslant$ inequality constraint: $X \leqslant R$
$r :\subset$ subsethood constraint: $X \subset R$
$r :$ blank possibilistic constraint; X is R; R is the possibility distribution of X
$r : p$ probabilistic constraint; X isp R; R is the probability distribution of X
$r : v$ veristic constraint; X isv R; R is the verity (truth) distribution of X
Standard constraints bivalent possibilistic, bivalent veristic and probabilistic
$r : rs$ random set constraint; X isrs R; R is the set-valued probability distribution of X
$r : fg$ fuzzy graph constraint; X isfg R; X is a function and R is its fuzzy graph
$r : u$ usuality constraint; X isu R means usually (X is R)
$r : g$ group constraint; X isg R means that R constrains the attribute-values of the group
$r : i$ iterated constraint; X isi R means that X iss S and S ist T.

To define the semantics of various modalities it is convenient to assume that the constrained variable, X, takes values in a finite set $U = (u_1, \ldots, u_n)$. With this assumption, the semantics of various constraints may be defined as follows.

Possibilistic Constraint

Consider the possibilistic constraint

$$X \text{ is } A ,$$

where A is a fuzzy set in U, defined as [73,75].

$$A = \mu_1/u_1 + \cdots + \mu_n/u_n ,$$

with the understanding that $+$ is a separator and μ_i is the grade of membership of u_i in A, $i = 1, \ldots, n$. The meaning of the possibilistic constraint, X is A, is defined as

$$X \text{ is } A \xrightarrow{\text{definition}} \text{Poss}(X = u_i) = \mu_i , \quad i = 1, \ldots, n .$$

Probabilistic Constraint

Let P be a probability distribution defined on U. P may be expressed as [96,101]

$$P = p_1 \backslash u_1 + \cdots + p_n \backslash u_n .$$

In this case, X isp $P \xrightarrow{\text{definition}} \text{Prob}(X = u_i) = p_i$, $i = 1, \ldots, n$.

It should be noted that p_i and u_i are allowed to take granular values, P_i and U_i, respectively, meaning that

$$p_i \text{ is } P_i \text{ and } u_i \text{ is } U_i , \quad i = 1, \ldots, n .$$

In this case, the probability distribution

$$P = P_1 \backslash U_1 + \cdots + P_n \backslash U_n$$

is a granule-valued probability distribution in the sense defined in Sect. "The Concept of Granulation". Alternatively, a granule-valued distribution may be viewed as the result of granulation of the expression

$$P = p_1 \backslash u_1 + \cdots + p_n \backslash u_n \,.$$

A granular probability distribution may be defined as an iterated generalized constraint. More specifically,

$$X \text{ isp } P$$
$$P: \ \text{Prob}(X \text{ is } A_i) \text{ is } P_i \,, \quad i = 1, \ldots, n \,,$$

where the A_i are granules of X and the P_i are granular probabilities (Fig. 5). In this case, as in the case of granule-valued distributions, P is expressed as

$$P = P_1 \backslash U_1 + \cdots + P_n \backslash U_n \,.$$

Note. Two examples will help to clarify the distinction between granular probability distributions and granule-valued probability distributions.

Example (a): Granular probability distribution. X is a real-valued random variable with probability distribution P. What is known about P is: Prob(X is small) is low; Prob(X is medium) is high; Prob(X is large) is low. Example (b): Granule-valued probability distribution. X is a random variable taking the values small, medium and large with respective granular probabilities low, high and low. Question: What are the expected values of these probability distributions?

Note. The concepts of granular and granule-valued probability distributions are closely related to the concepts of granular and granule-valued possibility distributions.

Veristic Constraint [93,98]

In this case, the semantics of X isv A is defined by

$$X \text{ isv } A \xrightarrow{\text{definition}} \text{Ver}(X = u_i) = \mu_i \,, \quad i = 1, \ldots, n \,.$$

where $\text{Ver}(X = u_i)$ is the verity (truth) of the proposition $X = u_i$.

Fuzzy Graph Constraint

In this case, X is a function, f, from U to V. Assume that U and V are granulated, with the granules of U and V being $A_1, \ldots, A_m$ and $B_1, \ldots, B_n$, respectively. The fuzzy graph, R, of f is defined as the disjunction of Cartesian products of the A_i and the $B_{j(i)}$ [73,74,78,90] (Fig. 19).

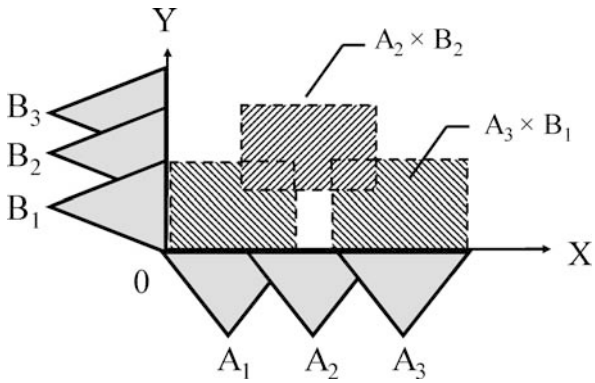

Fuzzy Logic, Figure 19
Fuzzy graph

$$R = A_1 \times B_{j(1)} + \cdots + A_m \times B_{j(m)} \,.$$

The fuzzy graph constraint is defined as the possibilistic constraint

$$f \text{ isfg } R \xrightarrow{\text{definition}} f \text{ is } (A_1 \times B_{j(1)} + \cdots + A_m \times B_{j(m)}) \,.$$

Usuality Constraint [93,96]

The constraint X isu A is defined by

$$X \text{ isu } A \xrightarrow{\text{definition}} \text{Prob}(X \text{ is } A) \text{ is usually} \,,$$

where usually is a fuzzy set in $[0, 1]$ which represents a fuzzy probability.

Primary Constraints

Every conceivable constraint can be viewed as an instance of a generalized constraint. In practice, such generality is rarely needed. What is sufficient for most practical purposes is a subset of generalized constraints which can be generated from so-called primary constraints through combination, projection, qualification, propagation and counterpropagation. The primary constraints are: (a) possibilistic; (b) probabilistic; and (c) veristic. The primary constraints are somewhat analogous to the primary colors: red, green and blue.

Standard Constraints

In large measure, scientific theories are based on what may be called standard constraints – constraints which are

a subset of primary constraints. The standard constraints are: (a) bivalent possibilistic; (b) probabilistic; and (c) bivalent veristic.

What is important to note is that generality of generalized constraints goes far beyond the generality of standard constraints. What this points to is that the concept of a generalized constraint opens the door to a wide-ranging generalization of scientific theories.

Generalized Constraint Language (GCL) [93,98]

The concept of a generalized constraint serves as a basis for construction of what is referred to as the Generalized Constraint Language (GCL). More specifically, GCL is generated by combination, projection, qualification, propagation and counterpropagation of generalized constraints. For example, combination of the probabilistic constraint

$$X \text{ isp } R\,,$$

where X is a variable which takes values in a finite set, and the possibilistic constraint

$$(X, Y) \text{ is } S\,,$$

generates the fuzzy random set constraint [96,98,101]

$$Y \text{ isfrs } T\,.$$

Fuzzy random sets are closely related to fuzzy-set-valued random variables. There is an extensive literature on fuzzy-set-valued random variables [9,24,43,46,47,50,60]. Random sets and set-valued random variables are closely related to the Dempster–Shafer theory of evidence [11,56,57]. An extension of the Dempster–Shafer theory to fuzzy sets and fuzzy probabilities is sketched in [81,83].

GCL is an open language in the sense that generalized constraints may be added to GCL at will. Simple examples of generalized constraints in GCL are:

$$(X \text{ is } R) \text{ and } ((X, Y) \text{ isp } S)$$
$$(X \text{ isu } R) \text{ and } (Y \text{ isu } S)$$
$$(X \text{ is } R) \text{ isp } S$$
$$(Y \text{ isu } B) \text{ if } (X \text{ is } A)\,.$$

In relation to GCL, PCL (Primary Constraint Language) and SCL (Standard Constraint Language) are subsets of GCL which are generated, respectively, by primary and standard constraints (Fig. 20).

Note. GCL is more than a language – it is a language system. A language has descriptive capability. A language system has deductive capability in addition to descriptive capability. GCL has both capabilities.

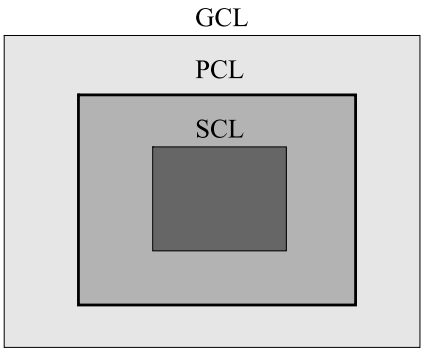

Fuzzy Logic, Figure 20
GCL, PCL and SCL

The concept of a generalized constraint plays a pivotal role in fuzzy logic. In particular, it serves to precisiate the concepts of information and meaning. More specifically, the fundamental thesis of fuzzy logic is that information may be represented as a generalized constraint.

$$\text{information} = \text{generalized constraint}\,,$$

with the understanding that information relates to the value of a variable, X, to which the generalized constraint applies. For example, if the information about X is that X is small, then this information may be represented as a possibilistic constraint

$$X \text{ is small}\,,$$

with small being a granular value of X. It should be noted that the traditional view that information is statistical in nature is a special, albeit important case, of viewing information as a generalized constraint. Another point which should be noted is that in information theory the primary concern is not with the substance of information but with its measure. The fundamental thesis relates to substance rather than measure [16,33].

An important corollary of the fundamental thesis is the meaning postulate of fuzzy logic. More specifically, let p be a proposition. A proposition is a carrier of information. Consequently,

$$\text{meaning of } p = \text{generalized constraint}\,.$$

More concretely,

$$\text{meaning of proposition} = \text{closed generalized constraint}$$
$$\text{meaning of predicate} = \text{open generalized constraint}\,.$$

The meaning postulate leads to an important connection between the concept of a generalized constraint and the concept of mm-precisiation. More specifically, what can be concluded is the equality

$$\text{mm-precisiand} = \text{generalized constraint} .$$

Equivalently,

$$\text{mm-precisiation} = \text{translation into GCL} .$$

This equality may be viewed as a more concrete statement of the meaning postulate. Transparency of translation may be enhanced through annotation. Details may be found in [101].

The meaning postulate points to an important aspect of translation into GCL.

Let SCL denote the subset of GCL which is associated with standard constraints, that is, bivalent possibilistic, probabilistic and bivalent veristic constraints. Let p be a proposition. An mm-precisiand of p, p^*, is an element of GCL.

Let $C(p^*, p)$ be the cointension of p^* in relation to p, and let $\sup_{\mathrm{GCL}} C(p^*, p)$ and $\sup_{\mathrm{SCL}} C(p^*, p)$ be the suprema of $C(p^*, p)$ over GCL and SCL, respectively. Since SCL is a subset of GCL, we have

$$\sup_{\mathrm{GCL}} C(p^*, p) \geqslant \sup_{\mathrm{SCL}} C(p^*, p) .$$

This obvious inequality has an important implication. Specifically, as a meaning representation language, fuzzy logic dominates bivalent logic. As a very simple example consider the proposition p: Speed limit is 65 mph. Realistically, what is the meaning of p? The inequality implies that employment of fuzzy logic for precisiation of p would lead to a precisiand whose cointension is at least as high – and generally significantly higher – than the cointension which is achievable through the use of bivalent logic.

More concretely, assume that A tells B that the speed limit is 65 mph, with the understanding that 65 mph should be interpreted as "approximately 65 mph". B asks A to precisiate what is meant by "approximately 65 mph", and stipulates that no imprecise numbers and no probabilities should be used in precisiation. With this restriction, A is not capable of formulating a realistic meaning of "approximately 65 mph". Next, B allows A to use imprecise numbers but no probabilities. B is still unable to formulate a realistic definition. Next, B allows A to employ imprecise numbers but no imprecise probabilities. Still, A is unable to formulate a realistic definition. Finally, B allows A to use imprecise numbers and imprecise probabilities. This allows A to formulate a realistic definition of "approximately 65 mph". This simple example is intended to demonstrate the need for the vocabulary of fuzzy logic to precisiate the meaning of terms and concepts which involve imprecise probabilities.

In addition to serving as a basis for precisiation of meaning, GCL serves another important function – the function of a deductive question-answering system [100]. In this role, what matters are the rules of deduction. In GCL, the rules of deduction coincide with the rules which govern constraint propagation and counterpropagation. Basically, these are the rules which govern generation of a generalized constraint from other generalized constraints [100,101].

The principal rule of deduction in fuzzy logic is the so-called extension principle [70,75]. The extension principle can assume a variety of forms, depending on the generalized constraints to which it applies. A basic form which involves possibilistic constraints is the following. An analogous principle applies to probabilistic constraints.

Let X be a variable which takes values in U, and let f be a function from U to V. The point of departure is a possibilistic constraint on $f(X)$ expressed as

$$f(X) \text{ is } A ,$$

where A is a fuzzy relation in V which is defined by its membership function $\mu_A(v)$, $v \in V$.

Let g be a function from U to W. The possibilistic constraint on $f(X)$ induces a possibilistic constraint on $g(X)$ which may be expressed as

$$g(X) \text{ is } B ,$$

where B is a fuzzy relation. The question is: What is B?

The extension principle reduces the problem of computation of B to the solution of a variational problem. Specifically,

$$\frac{f(X) \text{ is } A}{g(X) \text{ is } B}$$

where

$$\mu_B(w) = \sup_u \mu_A(f(u))$$

subject to

$$w = g(u) .$$

The structure of the solution is depicted in Fig. 21. Basically, the possibilistic constraint on $f(X)$ counterpropagates to a possibilistic constraint on X. Then, the possibilistic constraint on X propagates to a possibilistic constraint on $g(X)$.

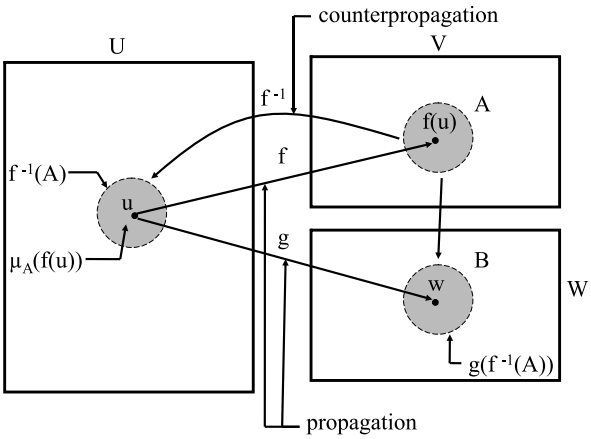

Fuzzy Logic, Figure 21
Structure of the extension principle

There is a version of the extension principle – referred to as the fuzzy-graph extension principle – which plays an important role in control and systems analysis [74,78,90]. More specifically, let f be a function from reals to reals, $Y = f(X)$. Let $^*f$ and $^*X$ be the granulands of f and X, respectively, with $^*f$ having the form of a fuzzy graph (Sect. "The Concept of Granulation").

$$^*f = A_1 \times B_{j(1)} + \cdots + A_m \times B_{j(m)} ,$$

where the A_i, $i = 1, \ldots, m$ and the B_j, $j = 1, \ldots, n$, are granules of X and Y, respectively; $\times$ denotes Cartesian product; and $+$ denotes disjunction (Fig. 22).

In this instance, the extension principle may be expressed as follows.

$$\frac{\begin{array}{l} X \text{ is } A \\ f \text{ is } (A_1 \times B_{j(1)} + \cdots + A_m \times B_{j(m)}) \end{array}}{Y \text{ is } (m_1 \wedge B_{j(1)} + \cdots + m_m \wedge B_{j(m)})}$$

where the m_i are matching coefficients, defined as [78]

$$m_i = \sup(A \cap A_i) , \quad i = 1, \ldots, m$$

and $\wedge$ denotes conjunction (min). In the special case where X is a number, a, the possibilistic constraint on Y may be expressed as

$$Y \text{ is } (\mu_{A_1}(a) \wedge B_{j(1)} + \cdots + \mu_{A_m}(a) \wedge B_{j(m)}) .$$

In this form, the extension principle plays a key role in the Mamdani–Assilian fuzzy logic controller [39].

Deduction

Assume that we are given an information set, I, which consists of a system of propositions $(p_1, \ldots, p_n)$. I will be referred to as the initial information set. The canonical problem of deductive question-answering is that of computing an answer to q, ans$(q|I)$, given I [100,101].

The first step is to ask the question: What information is needed to answer q? Suppose that the needed information consists of the values of the variables $X_1, \ldots, X_n$. Thus,

$$\text{ans}(q|I) = g(X_1, \ldots, X_n) ,$$

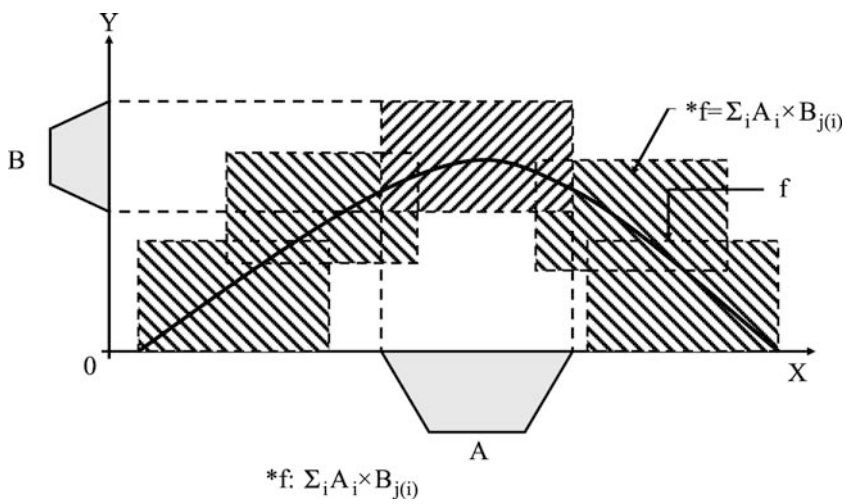

Fuzzy Logic, Figure 22
Fuzzy-graph extension principle. $B = {}^*f(A)$

where g is a known function.

Using GCL as a meaning precisiation language, one can express the initial information set as a generalized constraint on $X_1, \ldots, X_n$. In the special case of possibilistic constraints, the constraint on the X_i may be expressed as

$$f(X_1, \ldots, X_n) \text{ is } A\,,$$

where A is a fuzzy relation.

At this point, what we have is (a) a possibilistic constraint induced by the initial information set, and (b) an answer to q expressed as

$$\text{ans}(q|I) = g(X_1, \ldots, X_n)\,,$$

with the understanding that the possibilistic constraint on f propagates to a possibilistic constraint on g. To compute the induced constraint on g what is needed is the extension principle of fuzzy logic [70,75].

As a simple illustration of deduction, it is convenient to use an example which was considered earlier.

Initial information set,	p : most Swedes are tall
Question,	q : what is the average height of Swedes?

What information is needed to compute the answer to q? Let P be a population of n Swedes, $\text{Swede}_1, \ldots, \text{Swede}_n$. Let h_i be the height of Swede_i, $i = 1, \ldots, n$. Knowing the h_i, one can express the answer to q as

$$\text{av}(h) : \text{ans}(q|p) = \frac{1}{n}(h_1 + \cdots + h_n)\,.$$

Turning to the constraint induced by p, we note that the mm-precisiand of p may be expressed as the possibilistic constraint

$$p \xrightarrow{\text{mm-precisiation}} \frac{1}{n}\sum \text{Count(tall.Swedes) is most}\,,$$

where ΣCount(tall.Swedes) is the number of tall.Swedes in P, with the understanding that tall.Swedes is a fuzzy subset of P. Using this definition of ΣCount [86], one can write the expression for the constraint on the h_i as

$$\frac{1}{n}(\mu_{\text{tall}}(h_1) + \cdots + \mu_{\text{tall}}(h_n)) \text{ is most}\,.$$

At this point, application of the extension principle leads to a solution which may be expressed as

$$\mu_{\text{av}(h)}(v) = \sup_h \left(\frac{1}{n}\mu_{\text{most}}(\mu_{\text{tall}}(h_1) + \cdots + \mu_{\text{tall}}(h_n))\right)\,,$$
$$h = (h_1, \ldots, h_n)$$

subject to

$$v = \frac{1}{n}(h_1 + \cdots + h_n)\,.$$

In summary, the Generalized Constraint Language is, by construction, maximally expressive. Importantly, what this implies is that, in realistic settings, fuzzy logic, viewed as a modeling language, has a significantly higher level of power and generality than modeling languages based on standard constraints or, equivalently, on bivalent logic and bivalent-logic-based probability theory.

Principal Contributions of Fuzzy Logic

As was stated earlier, fuzzy logic is much more than an addition to existing methods for dealing with imprecision, uncertainty and complexity. In effect, fuzzy logic represents a paradigm shift. The structure of the shift is shown in Fig. 23.

Contributions of fuzzy logic range from contributions to basic sciences to applications involving various types of systems and products. The principal contributions are summarized in the following.

Fuzzy Logic as the Basis for Generalization of Scientific Theories

One of the principal contributions of fuzzy logic to basic sciences relates to what is referred to as FL-generalization.

By construction, fuzzy logic has a much more general conceptual structure than bivalent logic. A key element in the transition from bivalent logic to fuzzy logic is the generalization of the concept of a set to a fuzzy set. This generalization is the point of departure for FL-generalization.

More specifically, FL-generalization of any theory, T, involves an addition to T of concepts drawn from fuzzy

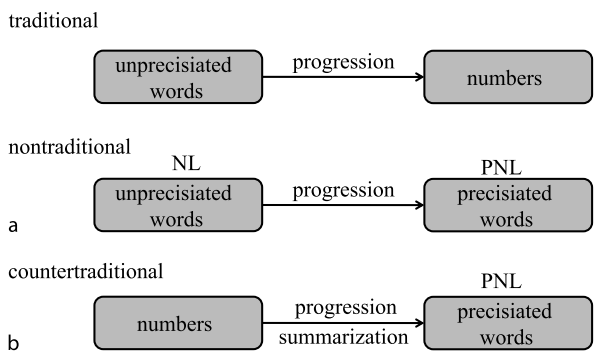

Fuzzy Logic, Figure 23
Fuzzy logic as a paradigm shift

logic. In the limit, as more and more concepts which are drawn from fuzzy logic are added to T, the foundation of T is shifted from bivalent logic to fuzzy logic. By construction, FL-generalization results in an upgraded theory, T^+, which is at least as rich and, in general, significantly richer than T.

As an illustration, consider probability theory, PT – a theory which is bivalent-logic-based. Among the basic concepts drawn from fuzzy logic which may be added to PT are the following [96].

set + fuzzy set
event + fuzzy event
relation + fuzzy relation
probability + fuzzy probability
random set + fuzzy random set
independence + fuzzy independence
stationarity + fuzzy stationarity
random variable + fuzzy random variable
etc.

As a theory, PT^+ is much richer than PT. In particular, it provides a basis for construction of models which are much closer to reality than those that can be constructed through the use of PT. This applies, in particular, to computation with imprecise probabilities.

A number of scientific theories have already been FL-generalized to some degree, and many more are likely to be FL-generalized in coming years. Particularly worthy of note are the following FL-generalizations.

control → fuzzy control
[12,18,20,69,72]
linear programming → fuzzy linear programming
[19,103]
probability theory → fuzzy probability theory
[96,98,101]
measure theory → fuzzy measure theory
[14,62]
topology → fuzzy topology
[38,68]
graph theory → fuzzy graph theory
[34,41]
cluster analysis → fuzzy cluster analysis
[7,27]
Prolog → fuzzy Prolog
[21,42]
etc.

FL-generalization is a basis for an important rationale for the use of fuzzy logic. It is conceivable that eventually the foundations of many scientific theories may be shifted from bivalent logic to fuzzy logic.

Linguistic Variables and Fuzzy If-Then Rules The most visible, the best understood and the most widely used contribution of fuzzy logic is the concept of a linguistic variable and the associated machinery of fuzzy if-then rules [90].

The machinery of linguistic variables and fuzzy if-then rules is unique to fuzzy logic. This machinery has played and is continuing to play a pivotal role in the conception and design of control systems and consumer products. However, its applicability is much broader. A key idea which underlies the machinery of linguistic variables and fuzzy if-then rules is centered on the use of information compression. In fuzzy logic, information compression is achieved through the use of graduated (fuzzy) granulation.

The Concepts of Precisiation and Cointension The concepts of precisiation and cointension play pivotal roles in fuzzy logic [101]. In fuzzy logic, differentiation is made between two concepts of precision: precision of value, v-precision; and precision of meaning, m-precision. Furthermore, differentiation is made between precisiation of meaning which is (a) human-oriented, or mh-precisiation for short; and (b) machine-oriented, or mm-precisiation for short. It is understood that mm-precisiation is mathematically well defined. The object of precisiation, p, and the result of precisiation, p^*, are referred to as precisiend and precisiand, respectively. Informally, cointension is defined as a measure of closeness of the meanings of p and p^*. Precisiation is cointensive if the meaning of p^* is close to the meaning of p. One of the important features of fuzzy logic is its high power of cointensive precisiation. What this implies is that better models of reality can be achieved through the use of fuzzy logic.

Cointensive precisiation has an important implication for science. In large measure, science is bivalent-logic-based. In consequence, in science it is traditional to define concepts in a bivalent framework, with no degrees of truth allowed. The problem is that, in reality, many concepts in science are fuzzy, that is, are a matter of degree. For this reason, bivalent-logic-based definitions of scientific concepts are, in many cases, not cointensive. To formulate cointensive definitions of fuzzy concepts it is necessary to employ fuzzy logic.

As was noted earlier, one of the principal contributions of fuzzy logic is its high power of cointensive precisiation. The significance of this capability of fuzzy logic is underscored by the fact that it has always been, and continues

to be, a basic objective of science to precisiate and clarify what is imprecise and unclear.

Computing with Words (CW), NL-Computation and Precisiated Natural Language (PNL) Much of human knowledge is expressed in natural language. Traditional theories of natural language are based on bivalent logic. The problem is that natural languages are intrinsically imprecise. Imprecision of natural languages is rooted in imprecision of perceptions. A natural language is basically a system for describing perceptions. Perceptions are intrinsically imprecise, reflecting the bounded ability of human sensory organs, and ultimately the brain, to resolve detail and store information. Imprecision of perceptions is passed on to natural languages.

Bivalent logic is intolerant of imprecision, partiality of truth and partiality of possibility. For this reason, bivalent logic is intrinsically unsuited to serve as a foundation for theories of natural language. As the logic of imprecision and approximate reasoning, fuzzy logic is a much better choice [71,80,84,85,86,88].

Computing with words (CW), NL-computation and precisiated natural language (PNL) are closely related formalisms [93,94,95,97]. In conventional modes of computation, the objects of computation are mathematical constructs. By contrast, in computing with words the objects of computation are propositions and predicates drawn from a natural language. A key idea which underlies computing with words involves representing the meaning of propositions and predicates as generalized constraints. Computing with words opens the door to a wide-ranging enlargement of the role of natural languages in scientific theories [30,31,36,58,59,61].

Computational Theory of Perceptions Humans have a remarkable capability to perform a wide variety of physical and mental tasks without any measurements and any computations. In performing such tasks humans employ perceptions. To endow machines with this capability what is needed is a formalism in which perceptions can play the role of objects of computation. The fuzzy-logic-based computational theory of perceptions (CTP) serves this purpose [94,95]. A key idea in this theory is that of computing not with perceptions per se, but with their descriptions in a natural language. Representing perceptions as propositions drawn from a natural language opens the door to application of computing with words to computation with perceptions. Computational theory of perceptions is of direct relevance to achievement of human level machine intelligence.

Possibility Theory Possibility theory is a branch of fuzzy logic [15,79]. Possibility theory and probability theory are distinct theories. Possibility theory may be viewed as a formalization of perception of possibility, whereas probability theory is rooted in perception of likelihood. In large measure, possibility theory and probability theory are complementary rather than competitive. Possibility theory is of direct relevance to, knowledge representation, semantics of natural languages, decision analysis and computation with imprecise probabilities.

Computation with Imprecise Probabilities Most real-world probabilities are perceptions of likelihood. As such, real-world probabilities are intrinsically imprecise [75]. Until recently, the issue of imprecise probabilities has been accorded little attention in the literature of probability theory. More recently, the problem of computation with imprecise probabilities has become an object of rapidly growing interest [63,99].

Typically, imprecise probabilities occur in an environment of imprecisely defined variables, functions, relations, events, etc. Existing approaches to computation with imprecise probabilities do not address this reality. To address this reality what is needed is fuzzy logic and, more particularly, computing with words and the computational theory of perceptions. A step in this direction was taken in the paper "Toward a perception-based theory of probabilistic reasoning with imprecise probabilities"; [96] followed by the 2005 paper "Toward a generalized theory of uncertainty (GTU) – an outline", [98] and the 2006 paper "Generalized theory of uncertainty (GTU) –principal concepts and ideas" [101].

Fuzzy Logic as a Modeling Language Science deals not with reality but with models of reality. More often than not, reality is fuzzy. For this reason, construction of realistic models of reality calls for the use of fuzzy logic rather than bivalent logic.

Fuzzy logic is a logic of imprecision, uncertainty and approximate reasoning [82]. It is natural to employ fuzzy logic as a modeling language when the objects of modeling are not well defined [102]. But what is somewhat paradoxical is that in many of its practical applications fuzzy logic is used as a modeling language for systems which are precisely defined. The explanation is that, in general, precision carries a cost. In those cases in which there is a tolerance for imprecision, reduction in cost may be achieved through imprecisiation, e. g., data compression, information compression and summarization. The result of imprecisiation is an object of modeling which is not precisely defined. A fuzzy modeling language comes into play at this

point. This is the key idea which underlies the fuzzy logic gambit. The fuzzy logic gambit is widely used in the design of consumer products – a realm in which cost is an important consideration.

A Glimpse of What Lies Beyond Fuzzy Logic

Fuzzy logic has a much higher level of generality than traditional logical systems – a generality which has the effect of greatly enhancing the problem-solving capability of fuzzy logic compared with that of bivalent logic.

What lies beyond fuzzy logic? What is of relevance to this question is the so-called incompatibility principle [73]. Informally, this principle asserts that as the complexity of a system increases a point is reached beyond which high precision and high relevance become incompatible. The concepts of mm-precisiation and cointension suggest an improved version of the principle: As the complexity of a system increases a point is reached beyond which high cointension and mm-precision become incompatible. What it means in plain words is that in the realm of complex systems – such as economic systems – it may be impossible to construct models which are both realistic and precise.

As an illustration consider the following problem. Assume that A asks a cab driver to take him to address B. There are two versions of this problem. (a) A asks the driver to take him to B the shortest way; and (b) A asks the driver to take him to B the fastest way. Based on his experience, the driver decides on the route to take to B. In the case of (a), a GPS system can suggest a route that is provably the shortest way, that is, it can come up with a provably valid (p-valid) solution. In the case of (b) the uncertainties involved preclude the possibility of constructing a model of the system which is cointensive and mm-precise, implying that a p-valid solution does not exist. The driver's solution, based on his experience, has what may be called fuzzy validity (f-validity). Thus, in the case of (b) no p-valid solution exists. What exists is an f-valid solution.

In fuzzy logic, mm-precisiation is a prerequisite to computation. A question which arises is: What can be done when cointensive mm-precisiation is infeasible? To deal with such problems what is needed is referred to as extended fuzzy logic (FL+). In this logic, mm-precisiation is optional rather than mandatory.

Very briefly, what is admissible in FL+ is f-validity. Admissibility of f-validity opens the door to construction of concepts prefixed with f, e. g. f-theorem, f-proof, f-principle, f-triangle, f-continuity, f-stability, etc. An example is f-geometry. In f-geometry, figures are drawn by hand with a spray can. An example of f-theorem in f-geometry is the f-version of the theorem: The medians of a triangle are concurrent. The f-version of this theorem reads: The f-medians of an f-triangle are f-concurrent. An f-theorem can be proved in two ways. (a) empirically, that is, by drawing triangles with a spray can and verifying that the medians intersect at an f-point. Alternatively, the theorem may be f-proved by constructing an f-analogue of its traditional proof.

At this stage, the extended fuzzy logic is merely an idea, but it is an idea which has the potential for being a point of departure for construction of theories with important applications to the solution of real-world problems.

Bibliography

Primary Literature

1. Aliev RA, Fazlollahi B, Aliev RR, Guirimov BG (2006) Fuzzy time series prediction method based on fuzzy recurrent neural network. In: Neuronal Informatinformation Processing Book. Lecture notes in computer science, vol 4233. Springer, Berlin, pp 860–869
2. Bargiela A, Pedrycz W (2002) Granular computing: An Introduction. Kluwer Academic Publishers, Boston
3. Bardossy A, Duckstein L (1995) Fuzzy rule-based modelling with application to geophysical, biological and engineering systems. CRC Press, New York
4. Bellman RE, Zadeh LA (1970) Decision-making in a fuzzy environment. Manag Sci B 17:141–164
5. Belohlavek R, Vychodil V (2006) Attribute implications in a fuzzy setting. In: Ganter B, Kwuida L (eds) ICFCA (2006) Lecture notes in artificial intelligence, vol 3874. Springer, Heidelberg, pp 45–60
6. Bezdek J, Pal S (eds) (1992) Fuzzy models for pattern recognition – methods that search for structures in data. IEEE Press, New York
7. Bezdek J, Keller JM, Krishnapuram R, Pal NR (1999) Fuzzy models and algorithms for pattern recognition and image processing. In: Zimmermann H (ed) Kluwer, Dordrecht
8. Bouchon-Meunier B, Yager RR, Zadeh LA (eds) (2000) Uncertainty in intelligent and information systems. In: Advances in fuzzy systems – applications and theory, vol 20. World Scientific, Singapore
9. Colubi A, Santos Domínguez-Menchero J, López-Díaz M, Ralescu DA (2001) On the formalization of fuzzy random variables. Inf Sci 133(1–2):3–6
10. Cresswell MJ (1973) Logic and Languages. Methuen, London
11. Dempster AP (1967) Upper and lower probabilities induced by a multivalued mapping. Ann Math Stat 38:325–329
12. Driankov D, Hellendoorn H, Reinfrank M (1993) An Introduction to Fuzzy Control. Springer, Berlin
13. Dubois D, Prade H (1980) Fuzzy Sets and Systems – Theory and Applications. Academic Press, New York
14. Dubois D, Prade H (1982) A class of fuzzy measures based on triangular norms. Int J General Syst 8:43–61
15. Dubois D, Prade H (1988) Possibility Theory. Plenum Press, New York

16. Dubois D, Prade H (1994) Non-standard theories of uncertainty in knowledge representation and reasoning. Knowl Engineer Rev Camb J Online 9(4):pp 399–416
17. Esteva F, Godo L (2007) Towards the generalization of Mundici's gamma functor to IMTL algebras: the linearly ordered case, Algebraic and proof-theoretic aspects of non-classical logics, pp 127–137
18. Filev D, Yager RR (1994) Essentials of Fuzzy Modeling and Control. Wiley-Interscience, New York
19. Gasimov RN, Yenilmez K (2002) Solving fuzzy linear programming problems with linear membership functions. Turk J Math 26:375–396
20. Gerla G (2001) Fuzzy control as a fuzzy deduction system. Fuzzy Sets Syst 121(3):409–425
21. Gerla G (2005) Fuzzy logic programming and fuzzy control. Studia Logica 79(2):231–254
22. Godo LL, Esteva F, García P, Agustí J (1991) A formal semantical approach to fuzzy logic. In: International Symposium on Multiple Valued Logic, ISMVL'91, pp 72–79
23. Goguen JA (1967) L-fuzzy sets. J Math Anal Appl 18:145–157
24. Goodman IR, Nguyen HT (1985) Uncertainty models for knowledge-based systems. North Holland, Amsterdam
25. Hajek P (1998) Metamathematics of fuzzy logic. Kluwer, Dordrecht
26. Hirota K, Sugeno M (eds) (1995) Industrial applications of fuzzy technology in the world. In: Advances in fuzzy systems – applications and theory, vol 2. World Scientific, Singapore
27. Höppner F, Klawonn F, Kruse R, Runkler T (1999) Fuzzy cluster analysis. Wiley, Chichester
28. Jamshidi M, Titli A, Zadeh LA, Boverie S (eds) (1997) Applications of fuzzy logic – towards high machine intelligence quotient systems. In: Environmental and intelligent manufacturing systems series, vol 9. Prentice Hall, Upper Saddle River
29. Jankowski A, Skowron A (2007) Toward rough-granular computing. In: Proceedings of the 11th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, (RSFDGrC'07), Toronto, Canada, pp 1–12
30. Kacprzyk J, Zadeh LA (eds) (1999) Computing with words in information/intelligent systems part 1. Foundations. Physica, Heidelberg, New York
31. Kacprzyk J, Zadeh LA (eds) (1999) Computing with words in information/intelligent systems part 2. Applications. Physica, Heidelberg, New York
32. Kandel A, Langholz G (eds) (1994) Fuzzy control systems. CRC Press, Boca Raton
33. Klir GJ (2006) Uncertainty and information: Foundations of generalized information theory. Wiley-Interscience, Hoboken
34. Kóczy LT (1992) Fuzzy graphs in the evaluation and optimization of networks. Fuzzy Sets Syst 46(3):307–319
35. Lambert K, Van Fraassen BC (1970) Meaning relations, possible objects and possible worlds. Philosophical problems in logic, pp 1–19
36. Lawry J, Shanahan JG, Ralescu AL (eds) (2003) Modelling with words – learning, fusion, and reasoning within a formal linguistic representation framework. Springer, Heidelberg
37. Lin TY (1997) Granular computing: From rough sets and neighborhood systems to information granulation and computing in words. In: European Congress on Intelligent Techniques and Soft Computing, September 8–12, pp 1602–1606
38. Liu Y, Luo M (1997) Fuzzy topology. In: Advances in fuzzy systems – applications and theory, vol 9. World Scientific, Singapore
39. Mamdani EH, Assilian S (1975) An experiment in linguistic synthesis with a fuzzy logic controller. Int J Man-Machine Stud 7:1–13
40. Mendel J (2001) Uncertain rule-based fuzzy logic systems – Introduction and new directions. Prentice Hall, Upper Saddle River
41. Mordeson JN, Nair PS (2000) Fuzzy graphs and fuzzy hypergraphs. In: Studies in Fuzziness and Soft Computing. Springer, Heidelberg
42. Mukaidono M, Shen Z, Ding L (1989) Fundamentals of fuzzy prolog. Int J Approx Reas 3(2):179–193
43. Nguyen HT (1993) On modeling of linguistic information using random sets. In: Fuzzy sets for intelligent systems. Morgan Kaufmann Publishers, San Mateo, pp 242–246
45. Novak V (2006) Which logic is the real fuzzy logic? Fuzzy Sets Syst 157:635–641
44. Novak V, Perfilieva I, Mockor J (1999) Mathematical principles of fuzzy logic. Kluwer, Boston/Dordrecht
46. Ogura Y, Li S, Kreinovich V (2002) Limit theorems and applications of set-valued and fuzzy set-valued random variables. Springer, Dordrecht
47. Orlov AI (1980) Problems of optimization and fuzzy variables. Znaniye, Moscow
48. Pedrycz W, Gomide F (2007) Fuzzy systems engineering: Toward human-centric computing. Wiley, Hoboken
49. Perfilieva I (2007) Fuzzy transforms: a challenge to conventional transforms. In: Hawkes PW (ed) Advances in images and electron physics, 147. Elsevier Academic Press, San Diego, pp 137–196
50. Puri ML, Ralescu DA (1993) Fuzzy random variables. In: Fuzzy sets for intelligent systems. Morgan Kaufmann Publishers, San Mateo, pp 265–271
51. Ralescu DA (1995) Cardinality, quantifiers and the aggregation of fuzzy criteria. Fuzzy Sets Syst 69:355–365
52. Ross TJ (2004) Fuzzy logic with engineering applications, 2nd edn. Wiley, Chichester
53. Rossi F, Codognet P (2003) Special Issue on Soft Constraints: Constraints 8(1)
54. Rutkowska D (2002) Neuro-fuzzy architectures and hybrid learning. In: Studies in fuzziness and soft computing. Springer
55. Rutkowski L (2008) Computational intelligence. Springer, Polish Scientific Publishers PWN, Warzaw
56. Schum D (1994) Evidential foundations of probabilistic reasoning. Wiley, New York
57. Shafer G (1976) A mathematical theory of evidence. Princeton University Press, Princeton
58. Trillas E (2006) On the use of words and fuzzy sets. Inf Sci 176(11):1463–1487
59. Türksen IB (2007) Meta-linguistic axioms as a foundation for computing with words. Inf Sci 177(2):332–359
60. Wang PZ, Sanchez E (1982) Treating a fuzzy subset as a projectable random set. In: Gupta MM, Sanchez E (eds) Fuzzy information and decision processes. North Holland, Amsterdam, pp 213–220
61. Wang P (2001) Computing with words. Albus J, Meystel A, Zadeh LA (eds) Wiley, New York
62. Wang Z, Klir GJ (1992) Fuzzy measure theory. Springer, New York

63. Walley P (1991) Statistical reasoning with imprecise probabilities. Chapman & Hall, London
64. Wygralak M (2003) Cardinalities of fuzzy sets. In: Studies in fuzziness and soft computing. Springer, Berlin
65. Yager RR, Zadeh LA (eds) (1992) An introduction to fuzzy logic applications in intelligent systems. Kluwer Academic Publishers, Norwell
66. Yen J, Langari R, Zadeh LA (ed) (1995) Industrial applications of fuzzy logic and intelligent systems. IEEE, New York
67. Yen J, Langari R (1998) Fuzzy logic: Intelligence, control and information, 1st edn. Prentice Hall, New York
68. Ying M (1991) A new approach for fuzzy topology (I). Fuzzy Sets Syst 39(3):303–321
69. Ying H (2000) Fuzzy control and modeling – analytical foundations and applications. IEEE Press, New York
70. Zadeh LA (1965) Fuzzy sets. Inf Control 8:338–353
71. Zadeh LA (1972) A fuzzy-set-theoretic interpretation of linguistic hedges. J Cybern 2:4–34
72. Zadeh LA (1972) A rationale for fuzzy control. J Dyn Syst Meas Control G 94:3–4
73. Zadeh LA (1973) Outline of a new approach to the analysis of complex systems and decision processes. IEEE Trans Syst Man Cybern SMC 3:28–44
74. Zadeh LA (1974) On the analysis of large scale systems. In: Gottinger H (ed) Systems approaches and environment problems. Vandenhoeck and Ruprecht, Göttingen, pp 23–37
75. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning Part I. Inf Sci 8:199–249; Part II. Inf Sci 8:301–357; Part III. Inf Sci 9:43–80
76. Zadeh LA (1975) Calculus of fuzzy restrictions. In: Zadeh LA, Fu KS, Tanaka K, Shimura M (eds) Fuzzy sets and their applications to cognitive and decision processes. Academic Press, New York, pp 1–39
77. Zadeh LA (1975) Fuzzy logic and approximate reasoning. Synthese 30:407–428
78. Zadeh LA (1976) A fuzzy-algorithmic approach to the definition of complex or imprecise concepts. Int J Man-Machine Stud 8:249–291
79. Zadeh LA (1978) Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets Syst 1:3–28
80. Zadeh LA (1978) PRUF – a meaning representation language for natural languages. Int J Man-Machine Stud 10:395–460
81. Zadeh LA (1979) Fuzzy sets and information granularity. In: Gupta M, Ragade R, Yager R (eds) Advances in fuzzy set theory and applications. North-Holland Publishing Co., Amsterdam, pp 3–18
82. Zadeh LA (1979) A theory of approximate reasoning. In: Hayes J, Michie D, Mikulich LI (eds) Machine intelligence 9. Halstead Press, New York, pp 149–194
83. Zadeh LA (1981) Possibility theory and soft data analysis. In: Cobb L, Thrall RM (eds) Mathematical frontiers of the social and policy sciences. Westview Press, Boulder, pp 69–129
84. Zadeh LA (1982) Test-score semantics for natural languages and meaning representation via PRUF. In: Rieger B (ed) Empirical semantics. Brockmeyer, Bochum, pp 281–349
85. Zadeh LA (1983) Test-score semantics as a basis for a computational approach to the representation of meaning. Proceedings of the Tenth Annual Conference of the Association for Literary and Linguistic Computing, Oxford University Press
86. Zadeh LA (1983) A computational approach to fuzzy quantifiers in natural languages. Comput Math 9:149–184
87. Zadeh LA (1984) Precisiation of meaning via translation into PRUF. In: Vaina L, Hintikka J (eds) Cognitive constraints on communication. Reidel, Dordrecht, pp 373–402
88. Zadeh LA (1986) Test-score semantics as a basis for a computational approach to the representation of meaning. Lit Linguist Comput 1:24–35
89. Zadeh LA (1986) Outline of a computational approach to meaning and knowledge representation based on the concept of a generalized assignment statement. In: Thoma M, Wyner A (eds) Proceedings of the International Seminar on Artificial Intelligence and Man-Machine Systems. Springer, Heidelberg, pp 198–211
90. Zadeh LA (1996) Fuzzy logic and the calculi of fuzzy rules and fuzzy graphs. Multiple-Valued Logic 1:1–38
91. Zadeh LA (1997) Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets Syst 90:111–127
92. Zadeh LA (1998) Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems. Soft Comput 2:23–25
93. Zadeh LA (1999) From computing with numbers to computing with words – from manipulation of measurements to manipulation of perceptions. IEEE Trans Circuits Syst 45:105–119
94. Zadeh LA (2000) Outline of a computational theory of perceptions based on computing with words. In: Sinha NK, Gupta MM, Zadeh LA (eds) Soft Computing & Intelligent Systems: Theory and Applications. Academic Press, London, pp 3–22
95. Zadeh LA (2001) A new direction in AI – toward a computational theory of perceptions. AI Magazine 22(1):73–84
96. Zadeh LA (2002) Toward a perception-based theory of probabilistic reasoning with imprecise probabilities. J Stat Plan Inference 105:233–264
97. Zadeh LA (2004) Precisiated natural language (PNL). AI Magazine 25(3)74–91
98. Zadeh LA (2005) Toward a generalized theory of uncertainty (GTU) – an outline. Inf Sci 172:1–40
99. Zadeh LA (2005) From imprecise to granular probabilities. Fuzzy Sets Syst 154:370–374
100. Zadeh LA (2006) From search engines to question answering systems – The problems of world knowledge, relevance, deduction and precisiation. In: Sanchez E (ed) Fuzzy logic and the semantic web, Chapt 9. Elsevier, pp 163–210
101. Zadeh LA (2006) Generalized theory of uncertainty (GTU)–principal concepts and ideas. Comput Stat Data Anal 51:15–46
102. Zadeh LA (2008) Is there a need for fuzzy logic? Inf Sci 178:(13)2751–2779
103. Zimmermann HJ (1978) Fuzzy programming and linear programming with several objective functions. Fuzzy Sets Syst 1:45–55

Books and Reviews

Aliev RA, Fazlollahi B, Aliev RR (2004) Soft computing and its applications in business and economics. In: Studies in fuzziness and soft computing. Springer, Berlin

Dubois D, Prade H (eds) (1996) Fuzzy information engineering: A guided tour of applications. Wiley, New York

Gupta MM, Sanchez E (1982) Fuzzy information and decision processes. North-Holland, Amsterdam
Hanss M (2005) Applied fuzzy arithmetic: An introduction with engineering applications. Springer, Berlin
Hirota K, Czogala E (1986) Probabilistic sets: Fuzzy and stochastic approach to decision, control and recognition processes, ISR. Verlag TUV Rheinland, Köln
Jamshidi M, Titli A, Zadeh LA, Boverie S (1997) Applications of fuzzy logic: Towards high machine intelligence quotient systems. In: Environmental and intelligent manufacturing systems series. Prentice Hall, Upper Saddle River
Kacprzyk J, Fedrizzi M (1992) Fuzzy regression analysis. In: Studies in fuzziness. Physica 29
Kosko B (1997) Fuzzy engineering. Prentice Hall, Upper Saddle River
Mastorakis NE (1999) Computational intelligence and applications. World Scientific Engineering Society
Pal SK, Polkowski L, Skowron (2004) A rough-neural computing: Techniques for computing with words. Springer, Berlin
Ralescu AL (1994) Applied research in fuzzy technology, international series in intelligent technologies. Kluwer Academic Publishers, Boston
Reghis M, Roventa E (1998) Classical and fuzzy concepts in mathematical logic and applications. CRC-Press, Boca Raton
Schneider M, Kandel A, Langholz G, Chew G (1996) Fuzzy expert system tools. Wiley, New York
Türksen IB (2005) Ontological and epistemological perspective of fuzzy set theory. Elsevier Science and Technology Books
Zadeh LA, Kacprzyk J (1992) Fuzzy logic for the management of uncertainty. Wiley
Zhong N, Skowron A, Ohsuga S (1999) New directions in rough sets, data mining, and granular-soft computing. In: Lecture Notes in Artificial Intelligence. Springer, New York

Fuzzy Logic, Type-2 and Uncertainty

Robert I. John[1], Jerry M. Mendel[2]
[1] Centre for Computational Intelligence, School of Computing, De Montfort University, Leicester, United Kingdom
[2] Signal and Image Processing Institute, Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Type-2 Fuzzy Systems
Generalized Type-2 Fuzzy Systems
Interval Type-2 Fuzzy Sets and Systems
Future Directions
Bibliography

Glossary

Type-1 fuzzy sets Are the underlying component in fuzzy logic where uncertainty is represented by a number between one and zero.

Type-2 fuzzy sets Are where the uncertainty is represented by a type-1 fuzzy set.

Interval type-2 fuzzy sets Are where the uncertainty is represented by a type-1 fuzzy set where the membership grades are unity.

Definition of the Subject

Type-2 fuzzy logic was first defined in 1975 by Zadeh and is an increasingly popular area for research and applications. The reason for this is because it appears to tackle the fundamental problem with type-1 fuzzy logic in that it is unable to handle the many uncertainties in real systems. Type-2 fuzzy systems are conceptually and mathematically more difficult to understand and implement but the proven applications show that the effort is worth it and type-2 fuzzy systems are at the forefront of fuzzy logic research and applications. These systems rely on the notion of a type-2 fuzzy set where the membership grades are type-1 fuzzy sets.

Introduction

Fuzzy sets [1] have, over the past forty years, laid the basis for a successful method of modeling uncertainty, vagueness and imprecision in a way that no other technique has been able. The use of fuzzy sets in real computer systems is extensive, particularly in consumer products and control applications.

Fuzzy logic (a logic based on fuzzy sets) is now a mainstream technique in everyday use across the world. The number of applications is many, and growing, in a variety of areas, for example, heat exchange, warm water pressure, aircraft flight control, robot control, car speed control, power systems, nuclear reactor control, fuzzy memory devices and the fuzzy computer, control of a cement kiln, focusing of a camcorder, climate control for buildings, shower control and mobile robots. The use of fuzzy logic is not limited to control. Successful applications, for example, have been reported in train scheduling, system modeling, computing, stock tracking on the Nikkei stock exchange and information retrieval.

Type-1 fuzzy sets represent uncertainty using a number in $[0, 1]$ whereas type-2 fuzzy sets represent uncertainty by a function. This is discussed in more detail later in the article. Essentially, the more imprecise or vague the data is, then type-2 fuzzy sets offer a significant improve-

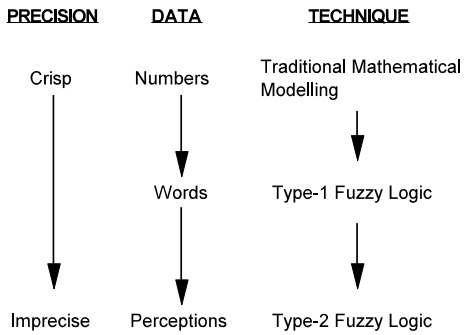

Fuzzy Logic, Type-2 and Uncertainty, Figure 1
Relationships between imprecision, data and fuzzy technique

ment on type-1 fuzzy sets. Figure 1 shows the view taken here of the relationships between levels of imprecision, data and technique.

As the level of imprecision increases then type-2 fuzzy logic provides a powerful paradigm for potentially tackling the problem. Problems that contain crisp, precise data do not, in reality, exist. However some problems can be tackled effectively using mathematical techniques where the assumption is that the data is precise. Other problems (for example, in control) use imprecise terminology that can often be effectively modeled using type-1 fuzzy sets. Perceptions, it is argued here, are at a higher level of imprecision and type-2 fuzzy sets can effectively model this imprecision.

The reason for this lies in some of the problems associated with type-1 fuzzy logic systems. Although successful in the control domain they have not delivered as well in systems that attempt to replicate human decision making. It is our view that this is because a type-1 fuzzy logic system (FLS) has some uncertainties which cannot be modeled properly by type-1 fuzzy logic. The sources of the uncertainties in type-1 FLSs are:

- The meanings of the words that are used in the antecedents and consequents of rules can be uncertain (words mean different things to different people).
- Consequents may have a histogram of values associated with them, especially when knowledge is extracted from a group of experts who do not all agree.
- Measurements that activate a type-1 FLS may be noisy and therefore uncertain.
- The data that are used to tune the parameters of a type-1 FLS may also be noisy.

The uncertainties described all essentially have to do with the uncertainty contained in a type-1 fuzzy set. A type-1 fuzzy set can be defined in the following way:

Let X be a universal set defined in a specific problem, with a generic element denoted by x. A fuzzy set A in X is a set of ordered pairs:

$$A = \{(x, \mu_A(x) \mid x \in X)\} ,$$

where $\mu_A : X \to [0, 1]$ is called the membership function A of and $\mu_A(x)$ represents the degree of membership of the element x in A.

The key points to draw from this definition of a fuzzy set are:

- The members of a fuzzy set are members to some degree, known as a *membership grade* or degree of membership.
- A fuzzy set is fully determined by the membership function.
- The membership grade is the degree of belonging to the fuzzy set. The larger the number (in [0, 1]) the more the degree of belonging.
- The translation from x to $\mu_A(x)$ is known as *fuzzification*.
- A fuzzy set is either continuous or discrete.
- Graphical representation of membership functions is very useful. For example, the fuzzy set 'Tall' might be represented as shown in Fig. 2 where someone who is of height five feet has a membership grade of zero while someone who is of height seven feet is tall to degree one, with heights in between having membership grade between one and zero. The example shown is linear but, of course, it could be any function.

Fuzzy sets offer a practical way of modeling what one might refer to as 'fuzziness'. The real world can be characterized by the fact that much of it is imprecise in one form or other. For a clear exposition (important to the notion of, and argument for, type-2 sets) two ideas of 'fuzziness'

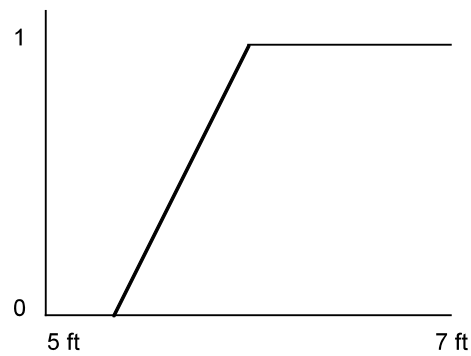

Fuzzy Logic, Type-2 and Uncertainty, Figure 2
The fuzzy set 'Tall'

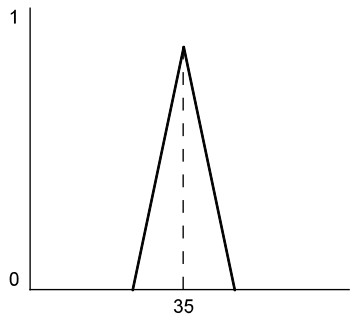

Fuzzy Logic, Type-2 and Uncertainty, Figure 3
The fuzzy number 'About 35'

can be considered important – imprecision and vagueness (linguistic uncertainty).

Imprecision

As has already been discussed, in many physical systems measurements are never precise (a physical property can always be measured more accurately). There is imprecision inherent in measurement. Fuzzy numbers are one way of capturing this imprecision by having a fuzzy set represent a real number where the numbers in an interval near to the number are in the fuzzy set to some degree. So, for example, the fuzzy number 'About 35' might look like the fuzzy set in Fig. 3 where the numbers closer to 35 have membership nearer unity than those that are further away from 35.

Vagueness or Linguistic Uncertainty

Another use of fuzzy sets is where words have been used to capture imprecise notions, loose concepts or perceptions. We use words in our everyday language that we, and the intended audience, know what we want to convey but the words cannot be precisely defined. For example, when a bank is considering a loan application somebody may be assessed as a good risk in terms of being able to repay the loan. Within the particular bank this notion of a good risk is well understood. It is not a black and white decision as to whether someone is a good risk or not – they are a good risk *to some degree*.

Type-2 Fuzzy Systems

A type-1 fuzzy system uses type-1 fuzzy sets in either the antecedent and/or the consequent of type-1 fuzzy if-then rules and a type-2 fuzzy system deploys type-2 fuzzy sets in either the antecedent and/or the consequent of type-2 fuzzy rules.

Fuzzy systems usually have the following features:

- The *fuzzy sets* as defined by their membership functions. These fuzzy sets are the basis of a fuzzy system. They capture the underlying properties or knowledge in the system.
- The *if-then rules* that combine the fuzzy sets – in a rule set or knowledge base.
- The fuzzy *composition* of the rules. Any fuzzy system that has a set of if-then rules has to combine the rules.
- Optionally, the defuzzification of the solution fuzzy set. In many (most) fuzzy systems there is a requirement that the final output be a 'crisp' number. However, for certain fuzzy paradigms the output of the system is a fuzzy set, or its associated word. This solution set is 'defuzzified' to arrive at a number.

Type-1 fuzzy sets are, in fact, crisp and not at all fuzzy, and are two dimensional. A domain value x is simply represented by a number in $[0, 1]$ – the membership grade. The methods for combining type-1 fuzzy sets in rules are also precise in nature.

Type-2 fuzzy sets, in contrast, are three dimensional. The membership grade for a given value in a type-2 fuzzy set is a type-1 fuzzy set. A formal definition of a type-2 fuzzy set is given in the following:

A type-2 fuzzy set, $\tilde{A}$, is characterized by a type-2 membership function $\mu_{\tilde{A}}(x, u)$, where $x \in X$ and $u \in J_x$ subset $JM[0, 1]$, and J_x is called the primary membership, i. e.

$$\tilde{A} = \{((x, u), \mu_{\tilde{A}}(x, u)) \mid \forall x \in X, \forall J_x \text{ subset } JM[0, 1]\}. \tag{1}$$

A useful way of looking at a type-2 fuzzy set is by considering its Footprint of Uncertainty (FOU). This is a two dimensional view of a type-2 fuzzy set. See Fig. 4 for a simple example. The shaded area represents the Union of all the J_x.

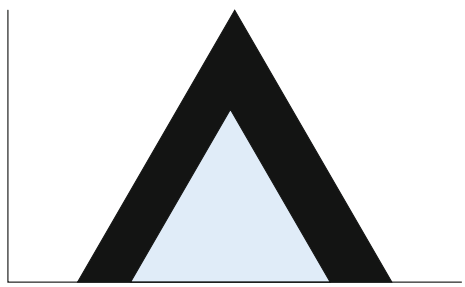

Fuzzy Logic, Type-2 and Uncertainty, Figure 4
A typical FOU of a type-2 set

An effective way to compare type-1 fuzzy sets and type-2 fuzzy sets is by use of a simple example. Suppose, for a particular application, we wish to describe the imprecise concept of 'tallness'. One approach would be to use a type-1 fuzzy set $tall_1$. Now suppose we are only considering three members of this set – Michael Jordan, Danny Devito and Robert John. For the type-1 fuzzy approach one might say that Michael Jordan is $tall_1$ to degree 0.95, Danny Devito to degree 0.4 and Robert John to degree 0.6. This can be written as

$$tall_1 = 0.95/Michael\ Jordan + 0.4/Danny\ Devito + 0.6/Robert\ John\ .$$

A type-2 fuzzy set ($tall_2$) that models the concept of 'tallness' could be

$$tall_2 = High_1/Michael\ Jordan + Low_1/Danny\ Devito + Medium_1/Robert\ John\ ,$$

where $High_1$, Low_1 and $Medium_1$ are type-1 fuzzy sets.

Figure 5 shows what the sets $High_1$, Low_1 and $Medium_1$ might look like if represented graphically. As can be seen the x axis takes values between 0 and 1, as does the y axis (μ).

Type-1 sets have an x axis representing the *domain* – in this case the height of an individual. Type-2 sets employ type-1 sets as the membership grades. Therefore, these fuzzy sets of type-2 allow for the idea that a fuzzy approach does not necessarily have membership grades [0, 1] in but the degree of membership for the member is itself a type-1 fuzzy set. As can be seen, by the simple example, there is an inherent extra fuzziness offered by type-2 fuzzy sets over and above a type-1 approach. So, a type-2 fuzzy set could be called a fuzzy-fuzzy set.

Real situations do not allow for precise numbers in [0, 1]. In a control application, for instance, can we say that a particular temperature, t, belongs to the type-1 fuzzy set hot_1 with a membership grade x precisely? No. Firstly it is highly likely that the membership could just as well be $x - 0.1$ for example. Different experts would attach different membership grades and, indeed, the same expert might well give different values on different days! On top of this uncertainty there is always some uncertainty in the measurement of t. So, we have a situation where an uncertain measurement is matched *precisely* to another uncertain value!! Type-2 fuzzy sets on the other hand, for certain appropriate applications, allow for this uncertainty to be modeled by not using precise membership grades but imprecise type-1 fuzzy sets.

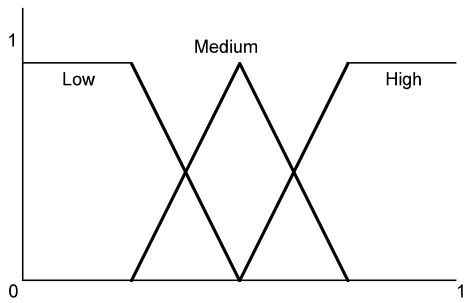

Fuzzy Logic, Type-2 and Uncertainty, Figure 5
The Fuzzy Sets $High_1$, Low_1 and $Medium_1$

So that type-2 sets can be used in a fuzzy system (in a similar manner to type-1 fuzzy sets) a method is required for computing the intersection (AND) and union (OR) of two type-2 sets. Suppose we have two type-2 fuzzy sets, $\tilde{A}$ and $\tilde{B}$ in X and $\mu_{\tilde{A}}(x)$ and $\mu_{\tilde{B}}(x)$ are two secondary membership functions of $\tilde{A}$ and $\tilde{B}$ respectively, represented as:

$$\begin{aligned}\mu_{\tilde{A}}(x) &= f(u_1)/u_1 + f(u_2)/u_2 + \cdots + f(u_n)/u_n \\ &= \sum_i f(u_i)/u_i\ , \\ \mu_{\tilde{B}}(x) &= g(w)/w_1 + f(w_2)/w_2 + \cdots + f(w_m)/w_m \\ &= \sum_j g(w_j)/w_j\ ,\end{aligned}$$

where the functions f and g are membership functions of fuzzy grades and $\{u_i, i = 1, 2, \ldots, n\}$, $\{w_j, j = 1, 2, \ldots, m\}$, are the members of the fuzzy grades.

Union of Type-2 Fuzzy Sets

The union ($\cup$) of two type-2 fuzzy sets ($\tilde{A}$, $\tilde{B}$) corresponding to $\tilde{A}$ OR $\tilde{B}$ is given by:

$$\begin{aligned}\tilde{A} \cup \tilde{B} \Leftrightarrow \mu_{\tilde{A}\cup\tilde{B}}(x) &= \tilde{A}\ \text{join}\ JM\ \tilde{B} \\ &= \sum_{ij} \frac{(f(u_i)JMg(w_j))}{(u_i JMw_j)}\ .\end{aligned}$$

Intersection of Type-2 Fuzzy Sets

The intersection ($\cap$) of two type-2 fuzzy sets ($\tilde{A}$, $\tilde{B}$) corresponding to $\tilde{A}$ AND $\tilde{B}$ is given by:

$$\begin{aligned}\tilde{A} \cap \tilde{B} \Leftrightarrow \mu_{\tilde{A}\cap\tilde{B}}(x) &= \tilde{A}\ \text{meet}\ JM\ \tilde{B} \\ &= \sum_{ij} \frac{(f(u_i)JMg(w_j))}{(u_i JMw_j)}\ ,\end{aligned}$$

where join JM denotes *join* and meet JM denotes *meet*.

Thus the join and meet allow for us to combine type-2 fuzzy sets for the situation where we wish to 'AND' or 'OR' two type-2 fuzzy sets. Join and meet are the building blocks for type-2 fuzzy relations and type-2 fuzzy inferencing with type-2 if-then rules.

Type-2 fuzzy if-then rules (type-2 rules) are similar to type-1 fuzzy if-then rules. An example type-2 if-then rule is given by

$$\text{IF } x \text{ is } \tilde{A} \text{ and } y \text{ is } \tilde{B} \text{ then } z \text{ is } \tilde{C}\,. \tag{2}$$

Obviously the rule could have a more complex antecedent connected by AND. Also the consequent of the rule could be type-1 or, indeed, crisp.

Type-2 output processing can be done in a number of ways through type-reduction (e. g. Centroid, Centre of Sums, Height, Modified Height and Center-of-Sets) that produces a type-1 fuzzy set, followed by defuzzification of that set.

Generalized Type-2 Fuzzy Systems

So far we have been discussing type-2 fuzzy systems where the secondary membership function can take any form – these are known as *generalized* type-2 fuzzy sets. Historically these have been difficult to work with because the complexity of the calculations is too high for real applications. Recent developments [2,3,4] mean that it is now possible to develop type-2 fuzzy systems where, for example, the secondary membership functions are triangular in shape. This is relatively new but offers an exciting opportunity to capture the uncertainty in real applications. We expect the interest in this area to grow considerably but for the purposes of this article we will concentrate on the detail of interval type-2 fuzzy systems where the secondary membership function is always unity.

Interval Type-2 Fuzzy Sets and Systems

As of this date, interval type-2 fuzzy sets (IT2 FSs) and interval type-2 fuzzy logic systems (IT2 FLSs) are most widely used because it is easy to compute using them. IT2 FSs are also known as *interval-valued FSs* for which there is a very extensive literature (e. g., [1], see the many references in this article, [5,6,16,28]). This section focuses on IT2 FSs and IT2 FLSs.

Interval Type-2 Fuzzy Sets

An IT2 FS $\tilde{A}$ is characterized as (much of the background material in this sub-section is taken from [23]; see also [17]):

$$\begin{aligned}\tilde{A} &= \int_{x\in X}\int_{u\in J_x\subseteq[0,1]} 1/(x,u)\\ &= \int_{x\in X}\left[\int_{u\in J_x\subseteq[0,1]} 1/u\right]/x\,,\end{aligned} \tag{3}$$

where x, the *primary variable*, has domain X; $u \in U$, the *secondary variable*, has domain J_x at each $x \in X$; J_x is called the *primary membership* of x and is defined below in (9); and, the *secondary grades* of $\tilde{A}$ all equal 1. Note that for continuous X and U, (3) means: $\tilde{A}\colon X \to \{[a,b] : 0 \leqslant a \leqslant b \leqslant 1\}$.

The bracketed term in (3) is called the *secondary MF*, or *vertical slice*, of $\tilde{A}$, and is denoted $\mu_{\tilde{A}}(x)$, i. e.

$$\mu_{\tilde{A}}(x) = \int_{u\in J_x\subseteq[0,1]} 1/u\,, \tag{4}$$

so that $\tilde{A}$ can be expressed in terms of its vertical slices as:

$$\tilde{A} = \int_{x\in X} \mu_{\tilde{A}}(x)/x\,. \tag{5}$$

Uncertainty about $\tilde{A}$ is conveyed by the union of all the primary memberships, which is called the *footprint of uncertainty* (FOU) of $\tilde{A}$ (see Fig. 6), i. e.

$$\mathrm{FOU}(\tilde{A}) = \bigcup_{\forall x\in X} J_x = \{(x,u)\colon u \in J_x \subseteq [0,1]\}\,. \tag{6}$$

The *upper membership function* (UMF) and *lower membership function* (LMF) of $\tilde{A}$ are two type-1 MFs that bound the FOU (Fig. 6). UMF($\tilde{A}$) is associated with the upper bound of FOU($\tilde{A}$) and is denoted $\bar{\mu}_{\tilde{A}}(x)$, $\forall x \in X$, and LMF($\tilde{A}$) is associated with the lower bound of FOU($\tilde{A}$) and is denoted $\underline{\mu}_{\tilde{A}}(x)$, $\forall x \in X$, i. e.

$$\mathrm{UMF}(\tilde{A}) \equiv \bar{\mu}_{\tilde{A}}(x) = \overline{\mathrm{FOU}(\tilde{A})} \quad \forall x \in X\,, \tag{7}$$

$$\mathrm{LMF}(\tilde{A}) \equiv \underline{\mu}_{\tilde{A}}(x) = \underline{\mathrm{FOU}(\tilde{A})} \quad \forall x \in X\,. \tag{8}$$

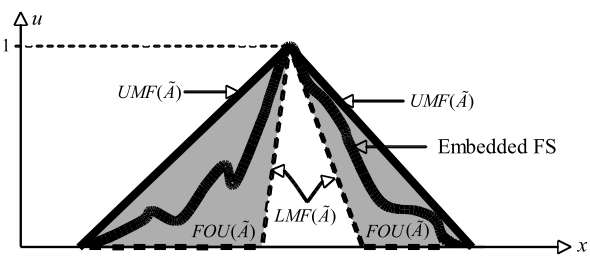

Fuzzy Logic, Type-2 and Uncertainty, Figure 6
FOU (*shaded*), LMF (*dashed*), UMF (*solid*) and an embedded FS (*wavy line*) for IT2 FS $\tilde{A}$

Note that J_x is an *interval set*, i. e.

$$J_x = \left[\underline{\mu}_{\tilde{A}}(x), \bar{\mu}_{\tilde{A}}(x)\right] . \tag{9}$$

This set is discrete when U is discrete and is continuous when U is continuous. Using (9), the $FOU(\tilde{A})$ in (6) can also be expressed as

$$\mathrm{FOU}(\tilde{A}) = \bigcup_{\forall x \in X} \left[\underline{\mu}_{\tilde{A}}(x), \bar{\mu}_{\tilde{A}}(x)\right] . \tag{10}$$

A very compact way to describe an IT2 FS is [22]:

$$\tilde{A} = 1/\mathrm{FOU}(\tilde{A}) , \tag{11}$$

where this notation means that the secondary grade equals 1 for all elements of $\mathrm{FOU}(\tilde{A})$.

For continuous universes of discourse X and U, an *embedded* IT2 FS $\tilde{A}_e$ is

$$\tilde{A}_e = \int_{x \in X} [1/u]/x \quad u \in J_x . \tag{12}$$

Note that (12) means: $\tilde{A}_e \colon X \to \{u \colon 0 \leqslant u \leqslant 1\}$. The set $\tilde{A}_e$ is embedded in $\tilde{A}$ such that at each x it only has one secondary variable (i. e., one primary membership whose secondary grade equals 1). Examples of $\tilde{A}_e$ are $1/\bar{\mu}_{\tilde{A}}(x)$ and $1/\underline{\mu}_{\tilde{A}}(x)$, $\forall x \in X$. In this notation it is understood that the secondary grade equals 1 at all elements in $\underline{\mu}_{\tilde{A}}(x)$ or $\bar{\mu}_{\tilde{A}}(x)$.

For discrete universes of discourse X and U, in which x has been discretized into N values and at each of these values u has been discretized into M_i values, an *embedded* IT2 FS $\tilde{A}_e$ has N elements, where $\tilde{A}_e$ contains exactly one element from $J_{x_1}, J_{x_2}, \ldots, J_{x_N}$, namely $u_1, u_2, \ldots, u_N$, each with a secondary grade equal to 1, i. e.,

$$\tilde{A}_e = \sum_{i=1}^{N} [1/u_i]/x_i , \tag{13}$$

where $u_i \in J_{x_i}$. Set $\tilde{A}_e$ is embedded in $\tilde{A}$, and, there are a total of $n_A = \prod_{i=1}^{N} M_i$ embedded T2 FSs.

Associated with each $\tilde{A}_e$ is an *embedded* T1 FS A_e, where

$$A_e = \int_{x \in X} u/x \quad u \in J_x . \tag{14}$$

The set A_e, which acts as the domain for $\tilde{A}_e$ (i. e., $\tilde{A}_e = 1/A_e$) is the union of all the primary memberships of the set $\tilde{A}_e$ in (12). Examples of A_e are $\bar{\mu}_{\tilde{A}}(x)$ and $\underline{\mu}_{\tilde{A}}(x)$, $\forall x \in X$.

When the universes of discourse X and U are continuous then there is an uncountable number of embedded IT2 and T1 FSs in $\tilde{A}$. Because such sets are only used for theoretical purposes and are not used for computational purposes, this poses no problem.

For discrete universes of discourse X and U, an *embedded* T1 FS A_e has N elements, one each from J_{x_1}, $J_{x_2}, \ldots, J_{x_N}$, namely $u_1, u_2, \ldots, u_N$, i. e.,

$$A_e = \sum_{i=1}^{N} u_i/x_i . \tag{15}$$

Set A_e is the union of all the primary memberships of set A_e and, there are a total of $\prod_{i=1}^{N} M_i$ embedded T1 FSs.

Theorem 1 (Representation Theorem (RT) [19] Specialized to an IT2 FS [22]) *For an IT2 FS, for which X and U are discrete, $\tilde{A}$ is the union of all of its embedded IT2 FSs. Equivalently, the domain of $\tilde{A}$ is equal to the union of all of its embedded T1 FSs, so that $\tilde{A}$ can be expressed as*

$$\begin{aligned}\tilde{A} = \sum_{j=1}^{n_A} \tilde{A}_e^j = 1/\mathrm{FOU}(\tilde{A}) = 1\Big/\sum_{j=1}^{n_A} A_e^j = 1\Big/\sum_{i=1}^{N} u_i^j/x_i \\ = 1\Big/\bigcup_{\forall x \in X} \left\{\underline{\mu}_{\tilde{A}}(x), \ldots, \bar{\mu}_{\tilde{A}}(x)\right\}\end{aligned} \tag{16}$$

and if X is a continuous universe, then the infinite set $\{\underline{\mu}_{\tilde{A}}(x), \ldots, \bar{\mu}_{\tilde{A}}(x)\}$ is replaced by the interval set $[\underline{\mu}_{\tilde{A}}(x), \bar{\mu}_{\tilde{A}}(x)]$.

This RT is arguably the most important result in IT2 FS theory because it can be used as the starting point for solving all problems involving IT2 FSs. It expresses an IT2 FS in terms of T1 FSs, so that all results for problems that use IT2 FSs can be solved using T1 FS mathematics [22]. Its use leads to the structure of the solution to a problem, after which efficient computational methods must be found to implement that structural solution. Table 1 summarizes set theoretic operations and uncertainty measures all of which were computed using the RT. Additional results for similarity measures are in [25].

Interval Type-2 Fuzzy Logic Systems

An interval type-2 fuzzy logic system (IT2 FLS), which is a FLS that uses at least one IT2 FS, contains five components – fuzzifier, rules, inference engine, type-reducer and defuzzifier – that are inter-connected, as shown in Fig. 7 (the background material in this sub-section is taken from [23]). The IT2 FLS can be viewed as a mapping from inputs to outputs (the path in Fig. 7, from "Crisp Inputs" to "Crisp Outputs"), and this mapping can be expressed quantitatively as $y = f(\boldsymbol{x})$, and is also known as *interval*

Fuzzy Logic, Type-2 and Uncertainty, Table 1
Results for IT2 FSs

Set Theoretic Operations [22]	
Union	$\tilde{A} \cup \tilde{B} = 1/\bigcup_{\forall x \in X}\left[\underline{\mu}_{\tilde{A}}(x) \vee \underline{\mu}_{\tilde{B}}(x), \bar{\mu}_{\tilde{A}}(x) \vee \bar{\mu}_{\tilde{B}}(x)\right]$
Intersection	$\tilde{A} \cap \tilde{B} = 1/\bigcup_{\forall x \in X}\left[\underline{\mu}_{\tilde{A}}(x) \wedge \underline{\mu}_{\tilde{B}}(x), \bar{\mu}_{\tilde{A}}(x) \wedge \bar{\mu}_{\tilde{B}}(x)\right]$
Complement	$\bar{\tilde{A}} = 1/\bigcup_{\forall x \in X}\left[1 - \underline{\mu}_{\tilde{A}}(x), 1 - \bar{\mu}_{\tilde{A}}(x)\right]$
Uncertainty Measures [9,21,24]	
Centroid	$C_{\tilde{A}} = [c_l(\tilde{A}), c_r(\tilde{A})] = \left[\frac{\sum_{i=1}^{L} x_i \bar{\mu}_{\tilde{A}}(x_i) + \sum_{i=L+1}^{N} x_i \underline{\mu}_{\tilde{A}}(x_i)}{\sum_{i=1}^{L} \bar{\mu}_{\tilde{A}}(x_i) + \sum_{i=L+1}^{N} \underline{\mu}_{\tilde{A}}(x_i)}, \frac{\sum_{i=1}^{R} x_i \underline{\mu}_{\tilde{A}}(x_i) + \sum_{i=R+1}^{N} x_i \bar{\mu}_{\tilde{A}}(x_i)}{\sum_{i=1}^{R} \underline{\mu}_{\tilde{A}}(x_i) + \sum_{i=R+1}^{N} \bar{\mu}_{\tilde{A}}(x_i)}\right]$ L and R computed using the KM Algorithms in Table 2.
Cardinality	$P_{\tilde{A}} = [p_l(\tilde{A}), p_r(\tilde{A})] = [p(\underline{\mu}_{\tilde{A}}(x)), p(\bar{\mu}_{\tilde{A}}(x))]$, $p(B) = \lvert X \rvert \sum_{i=1}^{N} \mu_B(x_i)/N$
Fuzziness	$F_{\tilde{A}} = [f_1(\tilde{A}), f_2(\tilde{A})] = [f_1(A_{e1}), f_2(A_{e2})]$, $f(A) = h\left(\sum_{i=1}^{N} g(\mu_A(x_i))\right)$ $A_{e1} : \mu_{A_{e1}}(x) = \begin{cases} \bar{\mu}_{\tilde{A}}(x) & \bar{\mu}_{\tilde{A}}(x) \text{ is further away from 0.5 than } \underline{\mu}_{\tilde{A}}(x) \\ \underline{\mu}_{\tilde{A}}(x) & \text{otherwise} \end{cases}$ $A_{e2} : \mu_{A_{e2}}(x) = \begin{cases} \bar{\mu}_{\tilde{A}}(x) & \text{both } \bar{\mu}_{\tilde{A}}(x) \text{ and } \underline{\mu}_{\tilde{A}}(x) \text{ are below 0.5} \\ \underline{\mu}_{\tilde{A}}(x) & \text{both } \bar{\mu}_{\tilde{A}}(x) \text{ and } \underline{\mu}_{\tilde{A}}(x) \text{ are above 0.5} \\ 0.5 & \text{otherwise} \end{cases}$
Variance	$V_{\tilde{A}} = [v_l(\tilde{A}), v_r(\tilde{A})] = \left[\min_{\forall A_e} v_{\tilde{A}}(A_e), \max_{\forall A_e} v_{\tilde{A}}(A_e)\right]$ $v_{\tilde{A}}(A_e) = \sum_{i=1}^{N} [x_i - c(\tilde{A})]^2 \mu_{A_e}(x_i) / \sum_{i=1}^{N} \mu_{A_e}(x_i)$, $c(\tilde{A}) = [c_l(\tilde{A}) + c_r(\tilde{A})]/2$ KM Algorithms are used to compute $v_l(\tilde{A})$ and $v_r(\tilde{A})$
Skew	$S_{\tilde{A}} = [s_l(\tilde{A}), s_r(\tilde{A})] = \left[\min_{\forall A_e} s_{\tilde{A}}(A_e), \max_{\forall A_e} s_{\tilde{A}}(A_e)\right]$ $s_{\tilde{A}}(A_e) = \sum_{i=1}^{N} [x_i - c(\tilde{A})]^3 \mu_{A_e}(x_i) / \sum_{i=1}^{N} \mu_{A_e}(x_i)$, $c(\tilde{A}) = [c_l(\tilde{A}) + c_r(\tilde{A})]/2$ KM Algorithms are used to compute $s_l(\tilde{A})$ and $s_r(\tilde{A})$

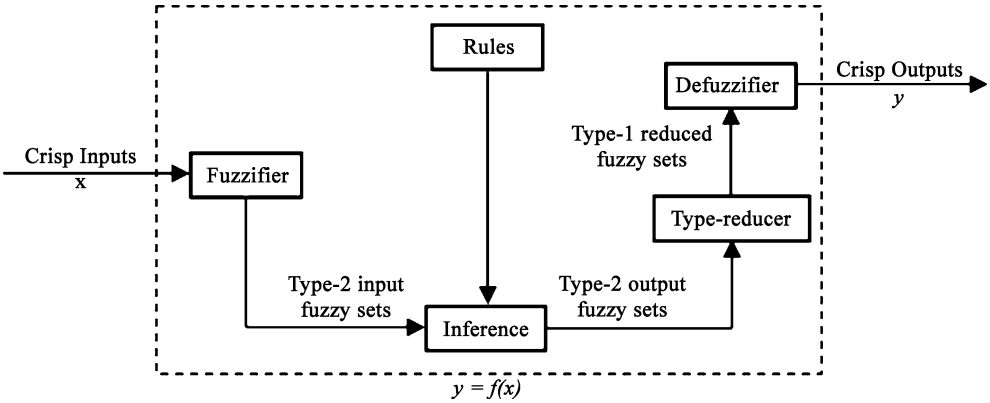

Fuzzy Logic, Type-2 and Uncertainty, Figure 7
Type-2 fuzzy logic system

type-2 fuzzy logic controller (IT2 FLC) [7], *interval type-2 fuzzy expert system, or interval type-2 fuzzy model.*

The inputs to the IT2 FLS prior to fuzzification may be certain (e. g., perfect measurements) or uncertain (e. g., noisy measurements). T1 or IT2 FSs can be used to model the latter measurements.

The IT2 FLS works as follows: the crisp inputs are first fuzzified into either type-0 (known as *singleton fuzzifica-*

tion), type-1 or IT2 FSs, which then activate the inference engine and the rule base to produce output IT2 FSs. These IT2 FSs are then processed by a type-reducer (which combines the output sets and then performs a centroid calculation), leading to an interval T1 FS called the *type-reduced set*. A defuzzifier then defuzzifies the type-reduced set to produce crisp outputs.

Rules are the heart of a FLS, and may be provided by experts or can be extracted from numerical data. In either case, rules can be expressed as a collection of IF-THEN statements. A *multi-input multi-output* (MIMO) rule base can be considered as a group of *multi-input single-output* (MISO) rule bases; hence, it is only necessary to concentrate on a *MISO* rule base. Consider an IT2 FLS having p inputs $x_1 \in X_1, \ldots, x_p \in X_p$ and one output $y \in Y$. We assume there are M rules where the ith rule has the form

$$R^i: \text{ IF } x_1 \text{ is } \tilde{F}_1^i \text{ and } \ldots \text{ and } x_p \text{ is } \tilde{F}_p^i, \quad \text{THEN } y \text{ is } \tilde{G}^i \quad i = 1, \ldots, M\,. \tag{17}$$

This rule represents a T2 relation between the input space, $X_1 \times \cdots \times X_p$, and the output space, Y, of the IT2 FLS. Associated with the p antecedent IT2 FSs, $\tilde{F}_k^i$, are the IT2 MFs $\mu_{\tilde{F}_k^i}(x_k)(k = 1, \ldots, p)$, and associated with the consequent IT2 FS $\tilde{G}^i$ is its IT2 MF $\mu_{\tilde{G}^i}(y)$.

The major result for an interval singleton T2 FLS, i. e. an IT2 FLS in which inputs are modeled as perfect measurements (type-0 FSs, singleton defuzzification) is summarized in the following:

Theorem 2 [11,15] *In an interval singleton T2 FLS using product or minimum t-norm, for input* $\mathbf{x} = \mathbf{x}'$:

(a) *The result of the input and antecedent operations, is an IT1 set called the firing set, i. e.,*

$$F^i(\mathbf{x}') = [\underline{f}^i(\mathbf{x}'), \overline{f}^i(\mathbf{x}')] \equiv [\underline{f}^i, \overline{f}^i] = \left[\underline{\mu}_{\tilde{F}_1^i}(x_1') * \cdots * \underline{\mu}_{\tilde{F}_p^i}(x_p'), \bar{\mu}_{\tilde{F}_1^i}(x_1') * \cdots * \bar{\mu}_{\tilde{F}_p^i}(x_p')\right]. \tag{18}$$

(b) *The rule R^i fired output consequent set, $\mu_{\tilde{B}^i}(y)$, is the IT2 FS*

$$\mu_{\tilde{B}^i}(y) = \int_{b^i \in \left[\underline{f}^i * \underline{\mu}_{\tilde{G}^i}(y), \overline{f}^i * \bar{\mu}_{\tilde{G}^i}(y)\right]} 1/b^i, \quad y \in Y\,, \tag{19}$$

where $\underline{\mu}_{\tilde{G}^i}(y)$ and $\bar{\mu}_{\tilde{G}^i}(y)$ are the lower and upper membership grades of $\mu_{\tilde{G}^i}(y)$.

(c) *Suppose that N of the M rules in the IT2 FLS fire, where $N \leqslant M$, and the combined output fuzzy set, $\mu_{\tilde{B}}(y)$, is obtained by combining the fired output consequent sets by taking the union of the rule R^i fired output consequent sets; then,*

$$\mu_{\tilde{B}}(y) = \int_{b \in \left[\begin{array}{c} \left[\underline{f}^1 * \underline{\mu}_{\tilde{G}^1}(y)\right] \vee \cdots \vee \left[\underline{f}^N * \underline{\mu}_{\tilde{G}^N}(y)\right], \\ \left[\overline{f}^1 * \bar{\mu}_{\tilde{G}^1}(y)\right] \vee \cdots \vee \left[\overline{f}^N * \bar{\mu}_{\tilde{G}^N}(y)\right] \end{array}\right]} 1/b\,, \quad y \in Y\,. \tag{20}$$

We do not necessarily advocate taking the union of these sets. Part (c) of this theorem merely illustrates the calculations if one chooses to do this. Generalizations of this theorem to an input that is a T1 or an IT2 FS are also given in [11,15] and [22].

In Fig. 7, the *type-reduced set* provides an interval of uncertainty for the output of an IT2 FLS, in much the same way that a confidence interval provides an interval of uncertainty for a probabilistic system. The more uncertainties that occur in an IT2 FLS, which translate into more uncertainties about its MFs, the larger will be the type-reduced set, and vice-versa.

Five different type-reduction (TR) methods are described in [10,15]. Each is inspired by what is done in a T1 FLS [when the (combined) output of the inference engine is defuzzified using a variety of defuzzification methods that all do some sort of centroid calculation] and are based on computing the *centroid of an* IT2 FS. Center-of-sets, centroid, center-of-sums, and height type-reduction can all be expressed as

$$\begin{aligned} Y_{\mathrm{TR}}(\mathbf{x}') &= \left[y_l(\mathbf{x}'), y_r(\mathbf{x}')\right] \equiv \left[y_l, y_r\right] \\ &= \int_{y^1 \in [y_l^1, y_r^1]} \cdots \int_{y^M \in [y_l^M, y_r^M]} \int_{f^1 \in \left[\underline{f}^1, \overline{f}^1\right]} \\ &\quad \cdots \int_{f^M \in \left[\underline{f}^M, \overline{f}^M\right]} 1 \Big/ \frac{\sum_{i=1}^M f^i y^i}{\sum_{i=1}^M f^i}\,, \end{aligned} \tag{21}$$

where the multiple integral signs denote the union operation. For a detailed explanation of (21) see [15] and [10]. The most widely used TR is center-of-sets (COS) TR, for which: y_l^i and y_r^i are the left and right end points of the centroid of the consequent of the ith rule [the centroids of all consequent IT2 FSs can be pre-computed using the KM Algorithms (Table 2) and stored for COS TR]; $\underline{f}^i$ and $\overline{f}^i$ are the lower and upper firing degrees of the ith rule, computed using (18); and M is the number of fired rules. For other kinds of TR methods, y_l^i, y_r^i, $\underline{f}^i$, $\overline{f}^i$ and M have different meanings, and are summarized in Table I of [23].

The defuzzified output of the IT2 FLS is simply the average of y_l and y_r, i. e.

$$y(\mathbf{x}') = [y_l(\mathbf{x}') + y_r(\mathbf{x}')]/2\,. \tag{22}$$

Fuzzy Logic, Type-2 and Uncertainty, Table 2
KM Algorithms for computing the centroid end-points of an IT2 FS, $\tilde{A}$, and their properties [9,15,20]. Note that $x_1 \leqslant x_2 \leqslant \cdots \leqslant x_N$

Step	KM Algorithm for c_l $c_l = \min_{\forall \theta_i \in \underline{\mu}_{\tilde{A}}(x_i), \bar{\mu}_{\tilde{A}}(x_i)} \left(\sum_{i=1}^N x_i\theta_i / \sum_{i=1}^N \theta_i\right)$	KM Algorithm for c_r $c_r = \max_{\forall \theta_i \in \underline{\mu}_{\tilde{A}}(x_i), \bar{\mu}_{\tilde{A}}(x_i)} \left(\sum_{i=1}^N x_i\theta_i / \sum_{i=1}^N \theta_i\right)$
1	Initialize θ_i by setting $\theta_i = [\underline{\mu}_{\tilde{A}}(x_i) + \bar{\mu}_{\tilde{A}}(x_i)]/2,\ i = 1, \ldots, N$ (or $\theta_i = \underline{\mu}_{\tilde{A}}(x_i), i \leqslant \lfloor (n+1)/2 \rfloor$ and $\theta_i = \bar{\mu}_{\tilde{A}}(x_i), i > \lfloor (n+1)/2 \rfloor$, where $\lfloor \bullet \rfloor$ denotes the first integer equal to or smaller than $\bullet$), and then compute $c' = c(\theta_1, \ldots, \theta_N) = \sum_{i=1}^N x_i\theta_i / \sum_{i=1}^N \theta_i$	
2	Find $k (1 \leqslant k \leqslant N-1)$ such that $x_k \leqslant c' \leqslant x_{k+1}$	
3	Set $\theta_i = \bar{\mu}_{\tilde{A}}(x_i)$ when $i \leqslant k$, and $\theta_i = \underline{\mu}_{\tilde{A}}(x_i)$ when $i \geqslant k+1$, and then compute $c_l(k) \equiv \frac{\sum_{i=1}^k x_i \bar{\mu}_{\tilde{A}}(x_i) + \sum_{i=k+1}^N x_i \underline{\mu}_{\tilde{A}}(x_i)}{\sum_{i=1}^k \bar{\mu}_{\tilde{A}}(x_i) + \sum_{i=k+1}^N \underline{\mu}_{\tilde{A}}(x_i)}$	Set $\theta_i = \underline{\mu}_{\tilde{A}}(x_i)$ when $i \leqslant k$, and $\theta_i = \bar{\mu}_{\tilde{A}}(x_i)$ when $i \geqslant k+1$, and then compute $c_r(k) = \frac{\sum_{i=1}^k x_i \underline{\mu}_{\tilde{A}}(x_i) + \sum_{i=k+1}^N x_i \bar{\mu}_{\tilde{A}}(x_i)}{\sum_{i=1}^k \underline{\mu}_{\tilde{A}}(x_i) + \sum_{i=k+1}^N \bar{\mu}_{\tilde{A}}(x_i)}$
4	Check if $c_l(k) = c'$. If yes, stop and set $c_l(k) = c_l$ and call k k_L. If no, go to Step 5	Check if $c_r(k) = c'$. If yes, stop and set $c_r(k) = c_r$ and call k k_R. If no, go to Step 5
5	Set $c' = c_l(k)$ and go to Step 2	Set $c' = c_r(k)$ and go to Step 2
Properties of the KM Algorithms [20]		
Convergence is monotonic and super-exponentially fast.		

Because TR is iterative, it may not be possible to use it in a real-time application. Wu and Mendel [26] introduced a method to approximate the TR set by minimax uncertainty bounds. Doing this avoids the computational overheads associated with TR and, as shown in [26] and [12], provides very similar outputs to the IT2 FLSs using TR. These uncertainty bounds are $\underline{y}_l(\mathbf{x}') \leqslant y_l(\mathbf{x}') \leqslant \bar{y}_l(\mathbf{x}')$ and $\underline{y}_r(\mathbf{x}') \leqslant y_r(\mathbf{x}') \leqslant \bar{y}_r(\mathbf{x}')$, where:

$$\bar{y}_l(\mathbf{x}') = \min\left\{\frac{\sum_{i=1}^M \underline{f}^i y_l^i}{\sum_{i=1}^M \underline{f}^i}, \frac{\sum_{i=1}^M \bar{f}^i y_l^i}{\sum_{i=1}^M \bar{f}^i}\right\}, \tag{23}$$

$$\bar{y}_r(\mathbf{x}') = \max\left\{\frac{\sum_{i=1}^M \underline{f}^i y_r^i}{\sum_{i=1}^M \underline{f}^i}, \frac{\sum_{i=1}^M \bar{f}^i y_r^i}{\sum_{i=1}^M \bar{f}^i}\right\}, \tag{24}$$

$$\underline{y}_l(\mathbf{x}') = \bar{y}_l(\mathbf{x}') - \left[\frac{\sum_{i=1}^M (\bar{f}^i - \underline{f}^i)}{\sum_{i=1}^M \bar{f}^i \sum_{i=1}^M \underline{f}^i} \times \frac{\sum_{i=1}^M \underline{f}^i \left(y_l^i - y_l^1\right) \sum_{i=1}^M \bar{f}^i \left(y_l^M - y_l^i\right)}{\sum_{i=1}^M \underline{f}^i \left(y_l^i - y_l^1\right) + \sum_{i=1}^M \bar{f}^i \left(y_l^M - y_l^i\right)}\right], \tag{25}$$

$$\bar{y}_r(\mathbf{x}') = \underline{y}_r(\mathbf{x}') + \left[\frac{\sum_{i=1}^M (\bar{f}^i - \underline{f}^i)}{\sum_{i=1}^M \bar{f}^i \sum_{i=1}^M \underline{f}^i} \times \frac{\sum_{i=1}^M \bar{f}^i \left(y_r^i - y_r^1\right) \sum_{i=1}^M \underline{f}^i \left(y_r^M - y_r^i\right)}{\sum_{i=1}^M \bar{f}^i \left(y_r^i - y_r^1\right) + \sum_{i=1}^M \underline{f}^i \left(y_r^M - y_r^i\right)}\right]. \tag{26}$$

Observe that the four bounds in (23)–(26) can be computed without having to perform TR. Wu and Mendel [26] then approximate the TR set, as $[y_l(\mathbf{x}'), y_r(\mathbf{x}')] \approx [(\underline{y}_l(\mathbf{x}') + \bar{y}_l(\mathbf{x}'))/2, (\underline{y}_r(\mathbf{x}') + \bar{y}_r(\mathbf{x}'))/2]$ and compute the output of the IT2 FLS as

$$y(\mathbf{x}') = \frac{1}{2}\left[\frac{\underline{y}_l(\mathbf{x}') + \bar{y}_l(\mathbf{x}')}{2} + \frac{\underline{y}_r(\mathbf{x}') + \bar{y}_r(\mathbf{x}')}{2}\right] \tag{27}$$

(instead of as in (22)). So, by using the uncertainty bounds, they obtain both an approximate TR set as well as a defuzzified output.

Wu and Mendel [26] still use TR during the design of an IT2 FLS. They define a new objective function that trades off some RMSE with not having to perform TR during the real-time operation of the IT2 FLS. The drawback to this approach is that TR is still performed during the design step. Lynch et al. [12,13] abandon TR completely where they replace all of the IT2 FLS computations with those in (23)–(26) which gave very similar outputs to the IT2 FLSs using TR. Doing this leads to an IT2 FLS real-time architecture.

Applications

Applications for IT2 FLSs or IT2 FSs are very numerous. Those that have appeared in the literature prior to 2001 can be found in (see pp 13–14 in [15]), and those that have appeared between 2001 and 2006 can be found in [16] and (see Table 24.8 in [18]). It is worth mentioning that applications have now appeared for the following general classes of problems: approximation, clustering, communications, control, databases, decision making embedded agents, health care, hidden Markov models, knowledge mining, neural networks, pattern classification, quality control, scheduling, signal and image processing, and

spatial query. Control applications, which were the original bread-and-butter ones for T1 FLSs, are now a major focus of attention for IT2 FLSs, and even general T2 FLSs. [8] provides three important applications that demonstrate that an IT2 FLS can significantly outperform a T1 FLS. Recently, IT2 FLSs have also been implemented in hardware [14]; this should make them more attractive and accessible for industrial applications.

Future Directions

There is much work still to be done in type-2 fuzzy research. But, we see that particular areas for fruitful and interesting work will include:

1. Applications. The two broad areas that fuzzy logic is used we can categorize as control and non-control. As discussed there is already some work in using type-2 for control and we expect this to grow. However, the weakness in type-1 fuzzy logic applications is in non-control where we are trying to emulate human expertise. Type-2 fuzzy sets are well placed to help here.
2. Generalized type-2 fuzzy systems. The relatively new approaches for allowing generalized type-2 fuzzy systems is exciting and we expect many researchers to exploit these approaches in interesting applications and provide new theoretical results.
3. Computing with Words. Because words mean different things to different people, we strongly believe that type-2 fuzzy sets must be used to model words when implementing Zadeh's Computing with Words paradigm.

Bibliography

1. Bustince H (2000) Indicator of inclusion grade for interval-valued fuzzy sets: application to approximate reasoning based on interval-valued fuzzy sets. Int J Approx Reason 23:137–209
2. Coupland S (2007) Type-2 Fuzzy Sets: Geometric Defuzzification and Type-Reduction. In: Proc FOCI, pp 622–629
3. Coupland S, John RI (2005) Towards More Efficient Type-2 Fuzzy Logic Systems. In: Proc Fuzz-IEEE, pp 236–241
4. Coupland S, John RI (2007) Geometric Type-1 and Type-2 Fuzzy Logic Systems. IEEE Trans Fuzzy Syst 15(1):3–15
5. Gorzalczany MB (1987) Decision making in signal transmission problems with interval-valued fuzzy sets. Fuzzy Sets Syst 23:191–203
6. Gorzalczany MB (1988) Interval-valued fuzzy controller based on verbal model of object. Fuzzy Sets Syst 28:45–53
7. Hagras H (2004) A hierarchical type-2 fuzzy logic control architecture for autonomous mobile robots. IEEE Trans Fuzzy Syst 12:524–539
8. Hagras H (2007) Type-2 FLCs: a new generation of fuzzy controllers. IEEE Comput Intel Mag 2:30–43
9. Karnik NN, Mendel JM (2001) Centroid of a type-2 fuzzy set. Inf Sci 132:195–220
10. Karnik NN, Mendel JM, Liang Q (1999) Type-2 fuzzy logic systems. IEEE Trans Fuzzy Syst 7:643–658
11. Liang Q, Mendel JM (2000) Interval type-2 fuzzy logic systems: theory and design. IEEE Trans Fuzzy Syst 8:535–550
12. Lynch C, Hagras H, Callaghan V (2005) Embedded type-2 FLC for real-time speed control of marine and traction diesel engines. In: Proc FUZZ-IEEE, pp 347–353
13. Lynch C, Hagras H, Callaghan V (2006) Using uncertainty bounds in the design of embedded real-time type-2 neuro-fuzzy speed controller for marine diesel engines. In: Proc FUZZ-IEEE, pp 7217–7224
14. Melgarejo M, Pena-Reyes CA (2007) Implementing interval type-2 fuzzy processors. IEEE Comput Intel Mag 2:63–71
15. Mendel JM (2001) Uncertain rule-based fuzzy logic systems: introduction and new directions. Prentice-Hall, Upper Saddle River
16. Mendel JM (2007) Advances in type-2 fuzzy sets and systems. Inf Sci 177:84–110
17. Mendel JM (2007) Type-2 fuzzy sets and systems: an overview. IEEE Comput Intel Mag 2:20–29
18. Mendel JM (2008) On type-2 fuzzy sets as granular models of words. In: Pedrycz W, Skowron A, Kreinovich V (eds) Granular Computing Handbook. Wiley, London
19. Mendel JM, John RI (2002) Type-2 fuzzy sets made simple. IEEE Trans Fuzzy Syst 10:117–127
20. Mendel JM, Liu F (2007) Super-exponential convergence of the Karnik-Mendel algorithms for computing the centroid of an interval type-2 fuzzy set. IEEE Trans on Fuzzy Syst 15:309–320
21. Mendel JM, Wu H (2007) New results about the centroid of an interval type-2 fuzzy set, including the centroid of a fuzzy granule. Inf Sci 177:360–377
22. Mendel JM, John RI, Liu F (2006) Interval type-2 fuzzy logic systems made simple. IEEE Trans Fuzzy Syst 14:808–821
23. Mendel JM, Hagras H, John RI (2006) Standard background material about interval type-2 fuzzy logic systems that can be used by all authors. IEEE Computational Intelligence Society standard. http://ieee.cis.org/standards
24. Wu D, Mendel JM (2007) Uncertainty measures for interval type-2 fuzzy sets. Inf Sci 177:5378–5393
25. Wu D, Mendel JM (2008) A vector similarity measure for linguistic approximation: interval type-2 and type-1 fuzzy sets, vol 177. Inf Sci 178:381–402
26. Wu H, Mendel JM (2002) Uncertainty bounds and their use in the design of interval type-2 fuzzy logic systems. IEEE Trans Fuzzy Syst 10:622–639
27. Zadeh LA (1965) Fuzzy Sets. Inf Control 8:338–353
28. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning–1. Inf Sci 8:199–249

Fuzzy Optimization

Weldon A. Lodwick, Elizabeth A. Untiedt
Department of Mathematical Sciences, University of Colorado Denver, Denver, USA

Article Outline

Glossary

Definition of the Subject
Introduction
Classical Approaches to Fuzzy Optimization
Possibilistic, Interval, Cloud, and Probabilistic Optimization Utilizing IVPM
Future Directions
Bibliography

Glossary

Fuzzy set A set whose membership is characterized by gradualness and uniquely described by a membership function which measures the degree of membership that domain values possess with respect to belonging to the set in question.

Fuzzy set theory The theory of uncertainty associated with sets characterized by gradual membership.

Possibility theory The theory of uncertainty associated with deficiency of information.

Fuzzy number A fuzzy set whose membership function is upper semi-continuous with a unique modal value and whose domain is the set of real numbers.

Possibilistic number A variable described by a possibilistic distribution whose domain is the set of real numbers.

Optimization The mathematical field that studies normative processes.

Definition of the Subject

Fuzzy optimization is normative and as a mathematical model deals with transitional uncertainty and information deficiency uncertainty. Some literature calls these uncertainties *vagueness* and *ambiguity* respectively. Transitional uncertainty is the domain of *fuzzy set theory* while uncertainty resulting from information deficiency is the domain of *possibility theory*. Suppose one is told by one's employer to go to the airport and meet a tall female visitor at the baggage claim of the airport. The notion of "tall" is transitional uncertainty. As the passengers arrive at the baggage claim, one compares each female passenger to the ascribed characteristic (tall woman) as determined by one's evaluation of what the boss' function is for "tall woman" and all the information one might have regarding "tall women". The resulting function is used to obtain a possibility value for each female that appears at the baggage claim. This uncertainty arises from information deficiency.

An example of fuzzy and possibilistic uncertainty in optimization is in the radiation treatment plan of tumors which is a problem that seeks to deposit a tumorcidal dose to cells that are cancerous while sparing all other cells and at the same time minimizing the total amount of radiation used for the treatment. In this context fuzzy uncertainty arises in cell classification because a cell may be precancerous in which case it is both healthy and cancerous at the same time to some degree. Typically, radiation treatment modeling discretizes a CT-scan of a patient into pixels (voxels in three dimensions) where a pixel is a position in space relative to a fixed coordinate system containing the patient and the radiation machine. The nature of a particular pixel in space may be two (or more) things at once. That is, a particular pixel may be precancerous and thus cancerous and healthy at the same time. Classification into discrete states leaves open transitional states. Continuous states are often intractable. That is, given a discrete classification of cells, what type of cell a pixel models is often transitional and thus fuzzy. Optimization models that incorporate uncertainty of these types are confronted with a wider spectrum of uncertainty than merely uncertainty due to probability, that is, frequency.

Possibilistic uncertainty arises when there is the specification, "deposit 60 units of radiation at every tumor cell". The minimal radiation that kills a tumor cell is considered 60 units of radiation by the community of radiation oncologists. The number 60 is derived from mathematical models, research, experience, and expert knowledge. It (60 units) represents the best available information as to a minimum radiation dose that will kill a tumor cell. Of course, if a radiation oncologist were able to attain 59.999 or 60.001 but not 60 units at a tumor cell, this would undoubtedly be satisfactory. The 60 units is "informational" since it is derived from research results and experience whose value as a single mathematical entity, the real number 60, is not precise in fact. It is not just a matter of measurement or probabilistic/statistical uncertainty though both statistics and probability theory may play a part in the determination of the number 60. The number 60 is possibilistic rather than probabilistic (or fuzzy) since the uncertainty associated with the death of a cancer cell given a precise amount of radiation, 60 in this case, as cause/effect (certainly killing a cancerous cell as a result of delivering precisely 60 units of radiation to the cell) is derived from sources beyond frequency analysis. The death of a cancer cell, as a result of radiation, certainly must depend on the type of cancer, its stage of growth, personal genetic characteristics, amount of food in the blood supply at the time of radiation, and so on.

Mathematical models may be considered as being of two types – descriptive (such as simulation) and normative. Optimization models and problems, including fuzzy optimization, are normative in that they impose upon the mathematical system criteria that seek to extract a "best".

This exposition will clearly identify fuzzy optimization as a distinct optimization approach. While it is not exhaustive, this presentation will focus on the salient features of fuzzy optimization most related to optimization under uncertainty. In particular, fuzzy multiobjective programming, fuzzy stochastic programming, and fuzzy dynamic programming are not discussed (see, for example [62,65,86,87,112]). Since intervals (and real numbers) may be considered as fuzzy sets, interval optimization is not covered separately but considered within the family of fuzzy optimization.

Mathematical analyzes that include the normative must embody the idea of order, measure, and distance. The notion of "best" requires an order and measuring with respect to that order. Professor Lothar Collatz in a lecture titled, "Monotonicity and Optimization", given to the Mathematics Department at Oregon State University, August 5, 1974, stated, "The idea of *ordering* is more fundamental than *distance* because from ordering we can get distance but not the other way around". (my notes) The real number system contains within itself the most fundamental mathematical order. Optimal control, stochastic dominance, [97,138], and mean-variance, [89,97,117], are approaches to order functions. Since fuzzy intervals from their definition (see below) relate themselves to *sets* of real numbers, that is, graphs in $\mathbb{R}^2$ (membership functions), the order of fuzzy intervals will need to be derived from that associated with real-valued functions. Moreover, since fuzzy sets model uncertainty on one hand and amplification and flexibility on the other, it is clear that the idea of order and its derived distance (measure) generated will be flexible, that is, require choices, as we shall see. The choice that needs to be made is dependent on the semantics of the problem to a greater extent than in the deterministic setting.

Mathematical modeling involves simplification, the transformation from reality to symbols. Even after a symbolic representation of a system, the process being modeled may, for example, be nonlinear, but is only tractable if linearized. Moreover, a nonlinear process may be the correct model but its linear counterpart may yield acceptable solutions. In the case of radiation therapy models (see [82,83]), scatter, which causes nonlinearities, is typically ignored in developing radiation therapy models that determine the angles and intensities for a given radiation machine and a given patient being treated. Without scatter, the resulting model is a linear model whose results usually produce acceptable treatment plans as long as the breathing of the patient is ignored. On the other hand, obtaining the mathematical model for the location of a cell (in a fixed coordinate system that includes both the patient and the radiation machine) of a patient undergoing radiation therapy, is important to include (while scatter may not be as important in the determination of acceptable angles and intensities). Location of points in the body that are or have been affected by breathing is not a deterministic model. Nor is it a probabilistic model.

The mathematical model development for radiation therapy planning of a particular patient tumor for a particular machine not only is designed to "do the job" (kill all tumor cells while sparing healthy cells) but often adds various "normative" criteria. For example, a radiation oncologist might seek a mathematical model which will, when used, kill the tumor cells, spare the healthy cells, *and* minimize total radiation used to do this. The "minimize total radiation used" is a normative criterion. Of course, there could be many other normative criteria imposed, such as "minimize the probability that healthy cells will become cancerous by the radiation deposited from the radiation treatment".

The thesis of this presentation is two-fold. First, not all uncertainty that occurs and is incorporated in optimization models can be described by the frequency with which its parameters take on various values (probability). Fuzzy and possibility theory are necessary components in some cases (such as in radiation of tumor models). Secondly, optimization under uncertainty models at their most general symbolic level try to capture the true complexity of the systems being modeled so that it is crucial to model transitional processes as fuzzy sets, frequencies as probabilistic distributions, and information deficiencies as possibility distributions. Once this is done, a further simplification may be (usually is) involved, but at the highest symbolic level, it is crucial to be faithful to the nature of the uncertainty, since this will ultimately determine the correct semantics and correct simplifying assumptions. For example, a normative probability model may be translated into a stochastic recourse model which is transformed into a real-valued nonlinear programming problem. Knowing that the underlying process is probabilistic means that the input data must be faithful to the laws of probability even though in the end, the model is a real-valued nonlinear programming model. The solution semantics in this case are those arising from probability and thus frequency based. A fuzzy linear optimization model is often translated into a real-valued linear programming model. Knowing that the underlying processes being modeled are fuzzy means that the nature of both the input data and the semantic interpretation of the output solution are based on the laws of fuzzy set theory and not of probability. Fuzzy optimization models are frequently solved as a real-valued linear or nonlinear programming problems just as in the

case of stochastic optimization. This process, insuring that at the highest symbolic level of mathematical modeling the correct uncertainty is used and only then transforming it into a real-value model, is analogous to first modeling a nonlinear system as a nonlinear system of relations and then linearizing it. It is crucial to create the nonlinear model first (at least in principle) and then to linearize it so that the approximation that is being used is explicit. When one is clear about these steps, one is able to take into account the correct associated approximation errors and underlying assumptions which enable the appropriate interpretation of the solution output.

Likewise, normative models of fuzzy processes require, at the level of mathematical abstraction, fuzzy modeling *before* approximation and transformation to an equivalent real-valued model because the fuzzy model is most faithful to the underlying uncertainty and thus is able to adhere to the basic assumptions associated with its uncertainty type. Moreover, if the starting point is a fuzzy model, when approximations and translations are made, the sources and magnitudes of associated errors are explicit. Lastly, the solution semantics are determined by the context of the input uncertainty. The modeling of normative processes that contain transitional and/or information deficiency uncertainty should utilize fuzzy and/or possibilistic optimization. The symbolic representation faithfully executed according to the associated axioms governing the uncertainty type and its semantics is a necessary first step. That is, mathematics (or any science) tries to bare all of its underlying assumptions. Since mathematical models objectify relationships occurring in reality via symbols, the nature of uncertainty must first be made explicit and then approximated, not the other way around.

Fuzzy optimization, which for this exposition encompasses models affected by both fuzzy and possibilistic uncertainty, is one of the newest optimization fields. Its place is along side stochastic optimization within the field of optimization under uncertainty. Fuzzy optimization began in 1970 with the publication of the seminal Bellman and Zadeh paper [5]. It took three years before the next fuzzy optimization article was published in 1973 by H. Tanaka, T. Okuda, and K. Asai, [1,119], (with the full version [120]). These researchers seem to have been the first to realize the importance of alpha-levels in the mathematical analysis of fuzzy sets in general and fuzzy optimization in particular. The Tanaka, Okuda, Asai article operationalized the theoretical approach developed by Bellman and Zadeh. Independently, in 1974, H.-J. Zimmermann presented a paper at the *ORSA/TIMS* conference in Puerto Rico [147] (with the full version [148]) that not only operationalized the Bellman and Zadeh approach, but greatly simplified and clarified fuzzy optimization, so much so that Zimmermann's approach is a standard to this day. In this same period, the book by C.V. Negoita and D.A. Ralescu [92] contained a description of fuzzy optimization. C.V. Negoita and M. Sularia published in 1976 a set containment approach to fuzzy optimization [93]. From this beginning, fuzzy optimization has become a field of study in its own right with a journal devoted to the subject, *Fuzzy Optimization and Decision Making*, whose first issue came out in February of 2002. Moreover, there have been special sessions devoted solely to fuzzy/possibilistic optimization at the international fuzzy society meetings (*IFSA05*, July 2005, Beijing, China and *IFSA07* June 2007, Cancun, Mexico). There have been two special editions of the journal *Fuzzy Sets and Systems* dealing with fuzzy/possibilistic optimization, the latest being [73]. Two books with edited articles have appeared – [12,54] and there are at least six authored books devoted to fuzzy optimization – [4,60,62,65,103,116].

Thirty-five years of fuzzy optimization research has yielded a wide-ranging set of applications. It is beyond the scope of this exposition to cover applications, but, the interested reader may wish to consult the following set of references: Chap. 6 and 8 in [12], Chap. III in [54], and [4,28,33,36,41,47,53,70,72,73,75,82,83,104,105,107,112,123,125,129,132,133].

The general fuzzy optimization model considered in this presentation is to find a fuzzy optimum of a fuzzy objective function subject to a fuzzy constraint,

$$\widetilde{\underset{x}{\text{opt}}}\, \tilde{z} = f(\tilde{c}, x) \tag{1}$$

$$x \,\tilde{\in}\, \tilde{X} \tag{2}$$

where the tilde, ~, represents fuzzy and/or possibilistic entities or relationships which is made clear by the context and assumptions of the problem. When fuzzy uncertainty needs to be distinguished from possibilistic uncertainty, the tilde, ~, will denote fuzzy uncertainty and the circumflex, $\widehat{\ }$, will denote possibilistic uncertainty. For (1), $\tilde{c}$ is considered to be a known vector of uncertainty parameters characterized by a fuzzy membership function or a possibilistic distribution. The variables (unknown quantities whose values are to be determined by the model) which are denoted here by x, are often called the "decision variables" because in optimization models, it is the value of x that is being computed. For example, x may denote a quantity to be produced by a manufacturing process, or the amount to be transported from a production point to a consuming point, or the intensity of radiation for the angle represented by the variable, and so on.

Introduction

It is assumed that the reader has a basic knowledge of fuzzy set theory at the level of [59] though the basics that are needed for this exposition are set forth. Much of the introductory exposition can also be found in [71].

Basics of Fuzzy Set Theory

Fuzzy set and possibility theory were defined and developed by L. Zadeh beginning with [142] and subsequently [143], and [144]. As is now well-known, the idea was to mathematize and develop analytical tools to solve problems whose uncertainty went beyond probability theory. Classical mathematical sets, for example a set A, have the property that either an element $x \in A$ or $x \notin A$ but not both. There are no other possibilities for classical sets which are also called *crisp* sets. An interval is a classical set. L. Zadeh's idea was to relax this "all or nothing" membership in a set to allow for grades of belonging to a set. When grades of belonging are used, a fuzzy set ensues. To each fuzzy set, $\tilde{A}$, L. Zadeh associated a real-valued *membership function* $\mu_{\tilde{A}}(x)$,which takes x in the domain of interest, the universe Ω, to a value in the interval $[0, 1]$. The membership function $\mu_{\tilde{A}}(x)$ quantifies the degree to which x belongs to $\tilde{A}$ where a value of zero means that x certainly does not belong to $\tilde{A}$ and a value of one means that x certainly belongs to $\tilde{A}$.

$$\mu_{\tilde{A}}(x) \colon \mathbb{R} \to [0, 1] \,.$$

Another way of looking at a fuzzy set is as a set in $\mathbb{R}^2$ as follows.

Definition 1 A **fuzzy set** $\tilde{A}$, as a crisp set in $\mathbb{R}^2$, is the set of ordered pairs

$$\tilde{A} = \{(x, \mu_{\tilde{A}}(x))\} \subseteq \{(-\infty, \infty) \times [0, 1]\} \,. \tag{3}$$

The $\boldsymbol{\alpha - cut}$ of a fuzzy set is the set

$$\tilde{A}_\alpha = \{x \,|\, \mu_{\tilde{A}}(x) \geq \alpha\} \,.$$

Definition 2 A **modal value** of a membership function is a domain value at which the membership function is one. A fuzzy set with at least one modal value is called **normal**. *The **support** of a membership function is the closure of* $\{x \,|\, \mu_{\tilde{A}}(x)) > 0\}$.

Definition 3 (see [29]) A **fuzzy interval**, $\tilde{M}$, defined by its membership function $\mu_{\tilde{M}}(\cdot)$, is a fuzzy continuous subset of the real line such that, if $x, y, z \in \mathbb{R}, z \in [x, y]$, then

$$\mu_{\tilde{M}}(z) \geq \min\{\mu_{\tilde{M}}(x), \mu_{\tilde{M}}(y)\} \,.$$

Like a fuzzy set, a fuzzy interval M is said to be **normal** if $\exists x \in \mathbb{R}$ such that $\mu_{\tilde{M}}(x) = 1$. The set $\{x \,|\, \mu_{\tilde{M}}(x) = 1\}$ is called the **core** (of the fuzzy interval).

Definition 4 A **fuzzy number** is a fuzzy interval with a unique modal value, that is, the core is a singleton.

For all that follows, fuzzy intervals will be assumed to be normal fuzzy intervals with upper semi-continuous membership functions. This means that the $\alpha - cut$ of a fuzzy interval, M_α, is a closed interval. Let $M_1 = \{x \,|\, \mu_{\tilde{M}}(x) = 1\} = [m_1^-, m_1^+]$, be the core of a fuzzy interval $\tilde{M}$ and the open support $M_0 = \{x \,|\, \mu_M(x) > 0\} = (m_0^-, m_0^+)$. For a fuzzy interval M, $\mu_M(x)$, is non-decreasing for $x \in (-\infty, m_1^-]$ and $\mu_M(x)$ is non-increasing for $x \in [m_1^+, \infty)$.

The fact that we have closed intervals at each $\alpha - cut$ means that fuzzy arithmetic can be defined by interval arithmetic (see [90]) on each $\alpha - cut$. Unbounded intervals can be handled by extended interval arithmetic. In fact, when dealing with fuzzy intervals, the operations and analysis can be considered as interval operations and analysis on $\alpha - cuts$ [71]. We define the most common types of fuzzy intervals next.

Definition 5 A **triangular fuzzy number**, $\tilde{A} = (\alpha, \beta, \gamma)$, has a membership function μ_A centered at a value α, with a support (β, γ) such that

$$\mu_A(x) = \begin{cases} 1, & x = \alpha, \\ 0, & x \leq \alpha - \beta, \\ 0, & x \geq \alpha + \gamma, \\ 1 - \frac{x-\alpha}{\beta}, & x \in (\alpha - \beta, \alpha) \\ 1 - \frac{\gamma - x}{\gamma}, & x \in (\alpha, \alpha + \gamma) \,. \end{cases}$$

Definition 6 A **symmetric triangular fuzzy number**, $\tilde{A} = (\alpha, \beta)$, has a membership function μ_A centered at a value α, with a spread β such that

$$\mu_A(x) = \begin{cases} 1, & x = \alpha, \\ 0, & x \leq \alpha - \beta, \\ 0, & x \geq \alpha + \beta, \\ 1 - \frac{|x-\alpha|}{\beta}, & x \in (\alpha - \beta, \alpha + \beta) \,. \end{cases}$$

Definition 7 A **trapezoidal fuzzy interval**, $\tilde{A} = (\alpha, \beta, \gamma, \delta)$, has a membership function μ_A with a core (α, β) and a support (γ, δ) such that

$$\mu_A(x) = \begin{cases} 1, & x \in [\alpha, \beta], \\ 0, & x \notin (\gamma, \delta), \\ 1 - \frac{\alpha - x}{\alpha - \gamma}, & x \in (\gamma, \alpha), \\ 1 - \frac{x - \beta}{\delta - \beta}, & x \in (\beta, \delta) \,. \end{cases}$$

Definition 8 A **L-R fuzzy interval**, $\tilde{A} = (\alpha^R, \alpha^L, \beta^R, \beta^L)_{LR}$ has a membership function μ_A and reference functions L and R. $L\colon [0, \inf) \to [0, 1]$ and $R\colon [0, \inf) \to [0, 1]$ are upper semi-continuous and strictly decreasing in the range (0,1], and μ_A is defined as follows:

$$\mu_A(x) = \begin{cases} 1, & x \in [\alpha^L, \alpha^R], \\ 0, & x \notin (\alpha^L - \beta^L, \alpha^R + \beta^R), \\ L\left(\frac{\alpha^L - x}{\beta^L}\right), & x \in [\alpha^L - \beta^L, \alpha^L], \\ R\left(\frac{x - \alpha^R}{\beta^R}\right), & x \in [\alpha^R, \beta + \alpha^R]\,. \end{cases}$$

The following relations are applicable to fuzzy intervals:

Definition 9 (Equality) Two fuzzy sets, $\tilde{A}$ and $\tilde{B}$ are said to be equal if and only if $\mu_A(x) = \mu_B(x)$ for all $x \in X$.

This definition, given by Bellman and Zadeh [5], is generally accepted. We explore broader interpretations of $\tilde{A} = \tilde{B}$ in Subsect. "Fuzzy Relations".

Definition 10 (Containment) A fuzzy set $\tilde{A}$ is said to be a subset of fuzzy set $\tilde{B}$ if and only if $\mu_A(x) \le \mu_B(x)$ for all $x \in X$.

Definition 11 (Intersection) The *intersection* of $\tilde{A}$ and $\tilde{B}$ is defined as the largest fuzzy set contained in both $\tilde{A}$ and $\tilde{B}$. The membership function of $\tilde{A} \wedge \tilde{B}$ is given by

$$\mu_{A \wedge B}(x) = \min(\mu_A(x), \mu_B(x))\,, \quad x \in X\,.$$

Minimum is only one of a continuum of operators that define intersection, but this discussion is beyond the scope of this paper. The $t-norms$ (see [59]) define a family of intersection operators.

Definition 12 (Relation) A *fuzzy relation* R in the product space $X \times Y$ is a fuzzy set characterized by a membership function μ_R which associates with each ordered pair a grade of membership $\mu_R(x, y)$ in $\mathbb{R}$.

Definition 13 (Non-Interactivity) Consider a fuzzy number $\tilde{C}$ where $C \subseteq \mathbb{R}^2$ which is a direct product of two fuzzy numbers $\tilde{A} \in \mathbb{R}$ and $\tilde{B} \in \mathbb{R}$ such that for all $(x, y) \in \mathbb{R}^2$,

$$\mu_C(x, y) = \mu_A(x) \wedge \mu_B(y)\,. \tag{4}$$

If condition (4) holds for $\tilde{A}$ and $\tilde{B}$, then $\tilde{A}$ and $\tilde{B}$, are said to be non-interactive. Semantically, two numbers are non-interactive if they can be assigned values independently of each other [18]. Every fuzzy and possibilistic model examined in this paper operates on the assumption of non-interactivity of all uncertain entities.

A different and more recent approach to fuzzy entities is possible. Instead of considering a fuzzy interval as a specialized fuzzy set over the set of real numbers, $\mathbb{R}$, Dubois and Prade (see [26]) and Fortin, Dubois, and Fargier (see [29]) introduce the concept of *gradual numbers* to revise the theory of fuzzy intervals so that a (real-valued) fuzzy interval is to a (real-valued) interval what a fuzzy set is to a (classical) set. This restores the algebraic structure of the real numbers to fuzzy arithmetic. (In particular, the usual fuzzy arithmetic which uses interval arithmetic on $\alpha - cuts$ [57] lacks an additive identity, a multiplicative identity, and the distributive law, all properties of algebra of real numbers.) An earlier approach to fuzzy arithmetic which also restores the algebraic structure of real numbers to fuzzy arithmetic is constraint interval arithmetic on $\alpha - cuts$ [69,71].

Gradual numbers were developed by [26,29] not only to restore a richer algebraic structure to fuzzy intervals, but to put fuzzy intervals on a firmer foundation. The theory of gradual numbers is one in which the relationship between real-valued intervals and fuzzy intervals is made quite apparent, clear, and compelling. Special (extreme) gradual numbers serve as endpoints of fuzzy intervals in the same way that real numbers serve as the endpoints of real intervals. The theory of gradual numbers also allows one to deal with the concept of a fuzzy element (of a fuzzy set) in a theoretically sound and meaningful way. The special gradual numbers associated with the endpoints of a fuzzy interval may be thought of as being composed of two parts. The left gradual number endpoint of a fuzzy interval $\tilde{A}$ is the inverse of the membership function $\mu_{\tilde{A}}$ restricted to $(-\infty, m_1^-]$. That is, it is the inverse of

$$\mu_{\tilde{A}}^-(x) = \mu_{\tilde{A}}(x)\,, \quad x \in (-\infty, m_1^-]\,, \tag{5}$$

where $[m_1^-, m_1^+]$ is the core as before, which is non-empty by definition. The second part, the right gradual number endpoint of $\tilde{A}$, is the inverse of the membership function restricted to $[m_1^+, \infty)$. That is, it is the inverse of

$$\mu_{\tilde{A}}^+(x) = \mu_{\tilde{A}}(x)\,, \quad x \in [m_1^+, \infty)\,. \tag{6}$$

Definition 14 (see [29]) A **gradual number** $\tilde{r}$ is defined by an assignment $A_{\tilde{r}}$ from $(0, 1]$ to $\mathbb{R}$.

Note that for a fuzzy interval, $\tilde{A}$, the functions that define the endpoints, $(\mu_{\tilde{A}}^-)^{-1}(\alpha)$ (5), and $(\mu_{\tilde{A}}^+)^{-1}(\alpha)$ (6) are special cases of gradual numbers. Briefly, a gradual number in fuzzy interval $\tilde{A}$ is simply the inverse relation of any unique assignment $r(x)\colon x \in \tilde{A}$ such that $\mu_{\tilde{A}}^-(x) \le r(x) \le \mu_{\tilde{A}}^+(x)$. For the purpose of optimization, continuous strictly-monotonic assignment functions are of interest. Fuzzy intervals may be defined from the point

of view of gradual numbers and and in this context, we have the following.

Definition 15 (see [29]) Using the notion of gradual number, a **fuzzy interval** M is an ordered pair of gradual numbers $(\tilde{m}^-, \tilde{m}^+)$ where $\tilde{m}^-$ is called the fuzzy lower bound or **left profile** and $\tilde{m}^+$ is called the fuzzy upper bound or **right profile**.

To ensure that the left and right profiles adhere to what has been previously defined as a *fuzzy interval*, several properties of $\tilde{m}^-$ and $\tilde{m}^+$ must hold. In particular (see [29]):

1. The domains of the assignment functions, $\tilde{A}_{\tilde{m}^-}$ and $\tilde{A}_{\tilde{m}^+}$, must be in $(0, 1]$.
2. $\tilde{A}_{\tilde{m}^-}$ must be increasing and $\tilde{A}_{\tilde{m}^+}$ must be decreasing.
3. $\tilde{m}^-$ and $\tilde{m}^+$ must be such that $\tilde{A}_{\tilde{m}^-} \leq \tilde{A}_{\tilde{m}^+}$.

Remark 16 Fuzzy intervals with properties 1–3 above possess well-defined inverses which are functions. Note that the endpoints of a crisp interval $[a, b]$ have constant assignments, that is, $\tilde{A}_{\tilde{m}^-}(\alpha) = a$ and $\tilde{A}_{\tilde{m}^+}(\alpha) = b, 0 < \alpha \leq 1$. These are the left and right profiles of the fuzzy interval membership function of a real-valued interval,

$$\mu_{[a,b]}(x) = \begin{cases} 1 & \text{for } x \in [a, b] \\ 0 & \text{otherwise}\,. \end{cases}$$

Since it is constant, it has no fuzziness in it and is simply an interval, not a fuzzy interval. An interval that possesses fuzziness has a non-decreasing, non-constant left profile and a non-increasing, non-constant right profile, but a crisp interval has no fuzziness in the left/right profiles (they are horizontal line segments as gradual numbers). That is, the left and right membership function segments are vertical line segments (no fuzziness), whose inverses are horizontal line segments. Each gradual number that is a horizontal line segment, $y = f(x) = a.0 \leq x \leq 1$, represents the real number (written as a fuzzy set),

$$\mu_a(x) = \begin{cases} 1 & \text{for } x = a \\ 0 & \text{otherwise}\,. \end{cases}$$

The application of gradual numbers to optimization is just beginning. Two of these are [56] and [127].

Extension Principles

Fuzzy extension principles show how to transform real-valued functions into functions of fuzzy sets on one hand and how to compute fuzzy algebraic or fuzzy transcendental expressions on the other. The meaning of fuzzy arithmetic depends directly on the extension principle since arithmetic operations are (continuous) functions over the reals, assuming division by zero is not allowed, and over the extended real numbers, when division by zero is allowed. The fuzzy arithmetic that results from Zadeh's extension principle [142] and its relationship to interval analysis has an extensive recent development (see [70,71,75]). Moreover, there is an intimate interrelationship between the extension principle being used and the analysis that ensues. Since *t-norms* and *t-conorms* capture the way trade-offs among decisions are made in constrained optimization (see [58]), the way one extends union and intersection via *t-norms* and *t-conorms* will determine the constraint set.

The extension principle within the context of fuzzy set theory was first proposed, developed, and defined in [142] and [144].

Definition 17 (Extension Principle of L. Zadeh [142, 144]) Given a real-valued function $f\colon X \to Y$, the function over fuzzy sets $f\colon S(X) \to S(Y)$, where $S(X)$ (respectively $S(Y)$) is the set of all fuzzy sets of X (respectively Y) is given by

$$\mu_{f(\tilde{A})}(y) = \sup\{\mu_{\tilde{A}}(x) \,|\, y = f(x)\} \tag{7}$$

for all fuzzy subsets A of $S(X)$. In particular, if $(X_1, \ldots, X_n)$ is a vector of fuzzy intervals, and $f(x_1, \ldots, x_n)$ a real-valued function, then

$$\begin{aligned} \mu_{f(X_1,\ldots,X_n)} &= \sup_{(x_1,\ldots,x_n)} \min_{1\leq i\leq n} \{\mu_{X_i}(x_i)\} \,|\, z \\ &= f(x_1, \ldots, x_n)\,. \end{aligned} \tag{8}$$

The extension principle is used to define functions of fuzzy sets by taking a function over the real numbers and extending it to a fuzzy-set-valued function with fuzzy set arguments in place of the real numbers. If $X \subset \mathbb{R}$, the set-valued function $F\colon A \subseteq \boldsymbol{P}(X) \to \boldsymbol{P}(Y)$, where $\boldsymbol{P}(X)$ is the power-set of X and $\boldsymbol{P}(Y)$ is the power-set of Y, is an *extension function* of real-valued function $f\colon X \to Y \subset \mathbb{R}$ if:

$$F(A) = \{f(x) \,|\, x \in A\}\,, \tag{9}$$

and

$$\left[\inf_{x\in A} f(x), \sup_{x\in A} f(x)\right] \subseteq F(A)\,. \tag{10}$$

This latter condition (10) needs to be imposed when these set-valued extension functions are approximated on a computer so that even computationally, it is always true that when a set is "retracted" to a point,

$$F(\{x\}) = f(x) \quad \forall x \in X$$

where F is the (set-valued) extension of (the real-valued) f. Now, if we have a set of fuzzy subsets $\tilde{A}$ of X, to ob-

tain a resulting fuzzy set within the set of all fuzzy subsets of Y under the mapping f, the extension function of f over fuzzy sets is denoted $\tilde{F}$ and the membership function (recall that a fuzzy set is uniquely defined by its membership function) $\mu_{\tilde{F}(\tilde{A})}$ is defined by:

$$\mu_{\tilde{F}(\tilde{A})}(y) = \sup\{\mu_{\tilde{A}}(x) \mid y = f(x), x \in X, y \in Y\}$$

The definition (8) of the extension principle has led to fuzzy arithmetic. Moreover, it is one of the main mechanisms used for fuzzy (interval) analysis. Various researchers have dealt with the issue of the extension principle and amplified its applicability. H. Nguyen [98] pointed out, in his 1978 paper, that a fuzzy set needs to be defined to be what Dubois and Prade later called a fuzzy interval (see [26,29]) in order that

$$[f(\tilde{A}, \tilde{B})]_\alpha = f(A_\alpha, B_\alpha)$$

where the function f is assumed to be continuous. In particular A_α and B_α need to be compact (that is closed/bounded intervals) for each α-*cut*. Thus, H. Nguyen defined a fuzzy number as one whose membership function is upper semi-continuous and for which the closure of the support is compact. In this case, the α-*cuts* generated are closed and bounded (compact) sets, that is, real-valued intervals. This is a well-known result in real analysis. That is, when f is continuous, the decomposition by $\alpha - cuts$ can be used to compute $f(\tilde{X}_1, \ldots, \tilde{X}_n)$ via interval analysis by [98] as

$$[f(\tilde{X}_1, \ldots, \tilde{X}_n)]_\alpha = f([X_1]_\alpha, \ldots, [X_n]_\alpha)\,.$$

It should be noted that the gradual number representation of fuzzy intervals allows use of the extension principle without resorting to $\alpha - cuts$.

R. Yager [141] pointed out that by looking at functions as graphs (in the Euclidean plane), the extension principle could be extended to include all graphs thus allowing for analysis of what he calls "non-deterministic" mappings, that is, graphs which are not functions. Now, "non-determinism" as is used by Yager can be considered as point-to-set mappings. Thus, Yager implicitly restores the extension principle to a more general setting of point-to-set mappings.

J. Ramik [100] points out that we can restore L. Zadeh's extension principle to its most general setting of set-to-set mappings explicitly. In fact, a fuzzy mapping is indeed a set-to-set mapping. He defines the image of a fuzzy set-to-set mapping as being the set of α's generated by the function on the $\alpha - cuts$ of the domain.

Lastly, T.Y. Lin's paper [61] is concerned with determining the function space in which the fuzzy set generated by the extension principle "lives". That is, to what space does the range of the fuzzy function belong? The extension principle generates a resultant membership function in the range space. Suppose one is interested in stable controls, then one way to extend is to generate resultant (range-space) membership functions that are continuous. The definition of continuous function states that small perturbations in the input, that is, domain, cause small perturbations in the output, that is, range, which is one way to view the definition of stability. T.Y. Lin develops conditions that are necessary in order that the range membership function has some desired characteristics (such as continuity or smoothness).

The essential purpose of fuzzy extension principles is to define functions over fuzzy sets so that the resulting range maintains various properties of interest specific to both the function and its fuzzy set input. In optimization, this is crucial in computing the output of objective and constraint functions (1) and (2). The theory and computational methods associated with distribution arithmetic including fuzzy and possibilistic arithmetic is discussed in detail in [71].

Basics of Possibility Theory

Possibilistic distributions (of fuzzy intervals or sets) encapsulate the most knowledgeable estimate of the possible values of an entity given the available information. This theory was articulated in [145]. Fuzzy membership function values (of fuzzy intervals or sets) describe the degree to which an entity is that value. Note that if the possibility distribution at x is one, this signifies that the best evidence available indicates x is indeed the entity that the distribution describes. Possibility distributions constructed from first principles require nested sets (see, for example [51]) and normalization. Possibility distributions are normalized since their semantics are tied to existent entities. Since the entity exists, it is always possible for at least one x. For example, if one is hiking from an elevation of 2,000 m to an elevation of 3,000 m, one must traverse the 2,500 m isoline (at least once). That there is a spot at which this occurs is not in question. The location of the spot is and will be dependent on the information at hand (accuracy of the maps, possibly of a global positioning system, compass, altimeter, expert knowledge). Nevertheless, one knows with certainty in this case *that* the 2,500 m isoline is traversed but not *where* it is traversed.

On the other hand, normalization is not required of fuzzy membership functions. Thus, *not all fuzzy sets can give rise to possibility distributions*.

Possibility theory may be derived in at least one of the following ways:

1. Via normalized fuzzy sets (see [145]),
2. Axiomatically from fuzzy measures g that satisfy $g(A \cup B) = \max\{g(A), g(B)\}$ (see [25,59] for example),
3. Via belief functions of Dempster–Shafer theory whose focal elements are normalized and nested (see [59]),
4. By construction (via nested sets with normalization, for example nested $\alpha - level$ sets, see [25,51] and [59]).

The most general derivation of possibility theory (method 2 above) sets up an order among variables with respect to their being an entity. The *magnitudes* associated with this ordering have no significance other than an indication of order. Thus, if $possibility_A(x) = 0.75$ and $possibility_A(y) = 0.25$ all that can be said is that x more likely to be the entity A than y. One *cannot* conclude that x is three times more likely to be A than y is. This means that for optimization problems, if the possibility distributions were constructed using the most general assumptions, then comparisons among several distributions is problematic. In particular, setting the possibility level to be greater than or equal to a certain fixed value, say $0 \leq \alpha \leq 1$, does not have the same meaning as setting a probability level or a membership function to be at least α. In the former case the α has no inherently meaning (other than if one has a $\beta > \alpha$ one prefers the decision that generated β to that which generated α) whereas in the former, the value of α is meaningful. Because of this, optimization methods must assume that the possibility distributions are constructed according to probability based possibility (see [51]) since the possibilities that are so constructed do have meaningful distribution value levels. That is, if the possibility level is α, the the value of α has a quantitative (not just order) meaning in relation to all values in $[0, 1]$.

There is a companion set-valued function to possibility called *necessity*, when the measure of the underlying space is finite. The necessity set-valued function is defined by

$$\text{necessity}(\tilde{A}) = 1 - \text{possibility}(\tilde{A}^C)\,,$$

where $\tilde{A}^C$ denotes the complement of the fuzzy set $\tilde{A}$. Semantically, the necessity of an event measures the impossibility of the opposite event.

Semantics of Fuzzy Sets and Possibility Distributions in Fuzzy Optimization

This study restricts uncertainty to parameters (input data) whose fuzzy membership functions or possibilistic distributions are over intervals of real numbers, that is, fuzzy or possibilistic intervals. That semantics is crucial in the context of fuzzy sets was known early in the development of fuzzy/possibility theory [19] and subsequently elaborated [23,27].

Fuzzy optimization is distinguished from possibilistic optimization by both semantics and optimization procedures. As will be seen below, fuzzy optimization optimizes over sets of real numbers while possibilistic optimization optimizes over sets of (possibility) distributions. In addition, optimization procedures are influenced by the fact that fuzzy and possibilistic distributions have different development when they are derived from first principles.

The semantic distinctions between fuzzy and possibilistic optimization can be found in [37,45] and [47] where it is noted that, for optimization models, *ambiguity* in the coefficients of the model leads to possibility optimization, while *vagueness* in the decision maker's preference is modeled by fuzzy optimization. When this vagueness represents a willingness on the part of the decision maker to relax his or her requirements in order to attain better results, this type of fuzzy optimization is sometimes called *flexible programming*. It has also been said that, in the context of optimization, possibilistic uncertainty is information-based uncertainty and fuzzy uncertainty is preference-based uncertainty [60]. Another point of view on the semantic distinction between fuzzy and possibilistic uncertainty in optimization is evinced in [45].

> The membership grade of a fuzzy goal (fuzzy constraint) represents the *degree of satisfaction*, whereas that of a possibility distribution represents the *degree of occurrence*.

The semantics of an optimization problem are also influenced by where in the optimization problem the uncertainty occurs. Suppose an optimization problem is known to have fuzzy uncertainty in the inequality constraints. In the case of soft constraints, it is the inequality (or equality) itself which is viewed to be fuzzy (i. e. $Ax \tilde{\leq} b$ for a linear programming problem). This is distinct from the case in which the right hand side has vague value, in which case the right hand side is viewed to be fuzzy (i. e. $Ax \leq \tilde{b}$). To illuminate the difference between fuzzy inequalities and fuzzy right hand sides, consider the example of a computer dating service. Suppose woman A specifies that she would like to date a "medium-height" man. Woman A defines "medium" as a fuzzy set characterized by a triangular membership function, centered at $69''$, with a spread of $3''$. This is a hard constraint because the man is *required* to be medium-height, but medium-height is a fuzzy set. This is, therefore, an example of a fuzzy right-hand side (i. e. $Ax = \tilde{b}$). Now consider woman B, who desires to date a man of about $69''$ in height. Unlike woman A, woman B is willing to compromise a little on the height requirement in order to be matched with a date who meets some of her other requirements. Her satisfaction level with a $69''$ man

is 1, with a 68″ or 70″ man is $\frac{2}{3}$, and with a 67″ or 71″ man is $\frac{1}{3}$. This is an example of a soft constraint, represented by a fuzzy equality (i. e. $Ax \doteq b$). Notice that the membership functions in the two fuzzy cases are the same (symmetric triangular centered at 69″ with a spread of 3″), but the semantics are different.

One result of the distinction between fuzziness and possibilistic uncertainty is that they manifest themselves in different regions of the optimization problem. Given the basic linear program,

$$\begin{aligned} &\min z = c^{\mathrm{T}} x \\ &\text{subject to:} \quad Ax \le b\,, \\ &\qquad\qquad x \ge 0\,, \end{aligned} \tag{11}$$

fuzziness can occur in the right-hand-side($\tilde{b}$), and/or in the inequality ($\tilde{\le}$). To date, no model has been proposed which deals with fuzzy objective function coefficients of fuzzy constraint matrix coefficients. However, suppose a constraint, $a_{ij}x \le b_i$, $i \in [1, n]$, is meant to apply only to members of a particular fuzzy set, $\tilde{Y}$. Now suppose that the element represented by row i of the constraint matrix is a member of set Y to a degree defined by $\mu_Y(y)$. Then the constraint i should be multiplied by the membership value of y_i, resulting in fuzzy coefficients. A similar argument can be applied for objective function coefficients. Possibilistic values can occur in the objective function coefficients $\hat{c}$, in the constraint matrix coefficients $\hat{A}$, and/or in the right-hand-side $\hat{b}$. To date, there is neither semantic nor model for a possibilistic inequality.

Fuzzy Relations

Optimization models usually involve constraints consisting of equalities, inequalities, or both. For deterministic optimization, the meaning and computation of the constraint set is clear. In optimization under uncertainty, however, the meaning and computation of "equality" and "inequality" must be determined. To this end, Dubois and Prade, [22,24], give a comprehensive analysis of fuzzy relations with four possible interpretations of fuzzy equalities, called modalities, and four possible interpretations of fuzzy inequalities. Inuiguchi [45] adds two more modalities. However, only the original four modalities are outlined below, using fuzzy intervals $\tilde{M} = [m^-(\alpha), m^+(\alpha)]$, $0 \le \alpha \le 1$, and $\tilde{N} = [n^-(\alpha), n^+(\alpha)]$, $0 \le \alpha \le 1$.

The statement $\tilde{M} \ge \tilde{N}$ can be interpreted in any of the four following ways:

i. $\forall x \in \tilde{M},\ \forall y \in \tilde{N},\ x > y$.
   This is equivalent to $m^-(\alpha) > n^+(\alpha)$.
ii. $\forall x \in \tilde{M},\ \exists y \in \tilde{N},\ x \ge y$.
   This is equivalent to $m^-(\alpha) \ge n^-(\alpha)$.
iii. $\exists x \in \tilde{M},\ \forall y \in \tilde{N},\ x > y$.
   This is equivalent to $m^+(\alpha) > n^+(\alpha)$.
iv. $\exists (x, y) \in \tilde{M} \times \tilde{N},\ x \ge y$.
   This is equivalent to $m^+(\alpha) \ge n^-(\alpha)$.

Inequality relation (i) indicates that x is necessarily greater than y, This is the pessimistic view. The decision maker who requires that $m^- > n+$ in order to satisfy $\tilde{M} > \tilde{N}$ is taking no chances [67]. (iv) indicates that x is possibly greater than y. This is the optimistic view. The decision maker who merely requires that $m^+ > n^-$ in order to satisfy $\tilde{M} > \tilde{N}$ has a hopeful outlook. Inequality relations (ii) and (iii) fall somewhere between the optimistic and pessimistic views.

The statement $\tilde{M} = \tilde{N}$ can be interpreted in any of the following four ways:

i. Zadeh's fuzzy set equality: $\mu_M = \mu_N$
ii. $\forall x \in \tilde{M},\ \exists y \in \tilde{N},\ x = y$
   (which is equivalent to $\tilde{M} \subseteq \tilde{N}$).
iii. $\forall y \in \tilde{N},\ \exists x \in \tilde{M},\ x = y$
   (which is equivalent to $\tilde{N} \subseteq \tilde{M}$).
iv. $\exists (x, y) \in \tilde{M} \times \tilde{N},\ x = y$
   (which is equivalent to $\tilde{N} \cap \tilde{M} \ne \emptyset$).

Equality relation (i) indicates that x is necessarily equal to y (the pessimistic view), (iv) indicates that x is possibly equal to y (the optimistic view), and (ii) and (iii) fall somewhere in between.

Basics of Fuzzy Sets and Possibility Theory in Optimization

The general fuzzy optimization model (1), (2), has two parts – the *objective* (normative criteria), $\widetilde{\text{opt}}\ \tilde{z} = f(\tilde{c}, x)$ (1), and the *constraints*, $x \tilde{\in} \tilde{X}$ (2). This corresponds to what Kacprzyk and Orlovski state (p. 50 in [55]):

> The analysis of real decision making situation is virtually based on two types of information:
>
> - information on feasible alternative decisions (options, choices, alternatives, variants, ...),
> - information making possible the comparison of alternative decisions with each other in terms of "better", "worse", "indifferent", etc.

The set of "feasible alternative decisions" is defined by the constraints while "the comparison of alternative decisions with each other" is accomplished by the objec-

tive. When the objective is a *utility function* that takes a given fuzzy set (or possibility distribution) and maps it to a subset of the real numbers and the constraints are transformed into equations and/or inequalities, the optimization model is a mathematical programming model. "Mathematical programming problems can be considered as decision making problems in which preferences between alternatives are described by means of objective function(s) on a set of alternatives given by constraints in such a way that more preferable alternatives have higher values" [102].

A key component of any mathematical programming problem is the input data. In our radiation therapy planning example, the input data includes the amount of radiation required to kill a cancerous cell, the maximum radiation a healthy cell can tolerate, and how much radiation a beam at some angle will deposit at a particular location in the body. When the input parameters of a mathematical programming model are described by uncertainty distributions (membership functions or possibility distributions), the mathematical programming problem becomes a fuzzy or possibilistic optimization problem. Recall that the decision-maker's flexibility is modeled by fuzzy relations, which results in fuzzy optimization. In addition, we sometimes see fuzzy and possibilistic parameters combined in the same problem, or possibilistic parameters occurring in the statement of fuzzy goals. These situations result in a mixed fuzzy and possibilistic optimization problem. We note that for the models considered here, while parameters, relations, and even the value of the objective function might be fuzzy or possibilistic, all decisions are "crisp". In practice, the solution to a mathematical programming problem is useless if it is not implementable.

There have been many superb surveys of the area of fuzzy optimization. Among all of these, the following are noted: [7,12,38,39,40,41,44,46,55,60,65,85,102,108,109,112,125,130,137], and [150].

It is clear that the fuzzy optimization model (1), (2) is ill-defined. The resolution of this ill-definition depends upon the function space in which the fuzzy optimization problem is solved. To date, two types of function spaces in which fuzzy optimization is "housed" have been used.

1. Fuzzy Banach Spaces (see [16,17,48,113])
2. Real Euclidean Space $\mathbb{R}^n$ (all other approaches)

Each of these two approaches has its own way of mapping the associated transitional and/or information deficiency uncertainty onto an ordered field. This order is necessary for the determination of optimality given the inherent normativeness of optimization.

Key Issues for Fuzzy Optimization Mapped to Fuzzy Banach Spaces Methods for solving optimization problems often involve iteration and/or approximation. For fuzzy optimization problems, a Banach space facilitates the convergence analysis of iterative or approximation algorithms. When each successive iterate or approximation remains a fuzzy entity in a fuzzy Banach space, the convergent entity or approximation is also in the fuzzy Banach space – it is again a fuzzy entity. Once convergence is achieved, a decision can be based on this fuzzy entity. When the problem is solved in a fuzzy Banach space, the fuzzy sets in the optimization problem remain fuzzy sets throughout the solution process and no translation is necessary except that of inclusion into an appropriate Banach space. Diamond and Kloeden [16,17] use a space in which their set of fuzzy sets, which they denoted $\boldsymbol{E}^n$, over $\mathbb{R}^n$ is endowed with a neighborhood system and metric that renders it a Banach space. They then find the Karush–Kuhn–Tucker (KKT) conditions for optimality (see [17]). Saito and Ishii [113] also develop the KKT conditions for optimization over fuzzy numbers. Diamond [15] and Jamison [48,49] consider equivalence classes in developing a Banach space of fuzzy sets. Jamison [48] goes on to use the Banach space developed from the equivalence classes to solve optimization problems. After the fuzzy optimization problem embedded in the fuzzy Banach space is solved, crisp decisions must be made based on fuzzy solution. This mapping from a fuzzy solution to the decision space is often called *defuzzification*. But up to the point of defuzzification all fuzzy entities from the model (1), (2) retain their fuzziness. Further discussion of fuzzy Banach space methods in fuzzy optimization is beyond the scope of this presentation.

Key Issues for Fuzzy Optimization Mapped to Real Euclidean Spaces The ill-definition of (1), (2) may be resolved by mapping the ambiguity and/or vagueness into a complete ordered lattice, namely, the real numbers. This is accomplished by translating the fuzzy objective and fuzzy constraint into real numbers or vectors, and real-valued relations. The various possible mappings distinguish among the types of fuzzy optimization. For a particular optimization problem in the form (1), (2), the answers to the following questions will determine which mapping is used.

1. Is the optimization fuzzy, possibilistic or a mixture?
2. What is meant by fuzzy/possibilistic function $f(\tilde{c}, x)$ and how does one compute $\tilde{z} = f(\tilde{c}, x)$?
3. What is meant by a fuzzy/possibilistic relation $x \tilde{\in} \tilde{X}$? How does one compute the resulting constraint set from fuzzy relations?

4. What is meant by fuzzy/possibilistic optimization of a fuzzy-valued function $\widetilde{\text{opt}}\,\tilde{z}$?

Is the Optimization Fuzzy, Possibilistic, or a Mixture? The first issue is the nature of the uncertainty itself. This depends upon the semantics of the problem, as discussed in Subsect. "Semantics of Fuzzy Sets and Possibility Distributions in Fuzzy Optimization". Recall that fuzzy entities are sets with non-sharp boundaries in which there is a transition between elements that belong and elements that don't belong to the set. Possibilistic entities are known to exist but the evidence associated with whether a particular element belongs to the set or not is incomplete or hard to obtain.

With these definitions in mind, we briefly consider fuzzy, possibilistic, and mixed decision making. Much of what is presented next can be found in [77] and [81].

Fuzzy Decision Making: Given the set of (crisp) decisions, Ω, and fuzzy sets, $\{\tilde{F}_i \mid i = 1 \text{ to } n\}$, find the optimal decision in the set Ω. That is,

$$\sup_{x\in\Omega} h\left(\tilde{F}_1(x), \ldots, \tilde{F}_n(x)\right) , \tag{12}$$

where $h\colon [0,1]^n \to [0,1]$ is an aggregation operator, often taken to be the min function, and $\tilde{F}_i(x) \in [0,1]$ is the fuzzy membership of x in fuzzy set $\tilde{F}_i$. Note that the decision space Ω is a **crisp set** (a set of real numbers) and the optimal decision satisfies a mutual membership condition defined by the aggregation operator h. The methods of Bellman and Zadeh [5], Tanaka, Okuda and Asai [120], and Zimmermann [147,148], who were the first to develop fuzzy mathematical programming fall into this category. While the aggregation operator h historically has been the min operator, it can be, for example, any $t-norm$ that is consistent with the context of the problem and/or decision methods, for example, risk aversion (see [58,106], or [136]). For a discussion of aggregation operators see [59].

Possibilistic Decision Making: Given the set of (crisp) decisions, Ω, and the set of possibility distributions representing the uncertain outcomes from selecting decision $\vec{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$ denoted $\Psi_x = \{\hat{F}_x^i, i = 1, \ldots, n\}$, find the optimal decision that produces the best set of possible outcomes with respect to an ordering U of the outcomes. That is,

$$\sup_{\hat{F}_x^i\in\Psi} U(\hat{F}_x^1, \ldots, \hat{F}_x^n) , \tag{13}$$

where $U(\hat{F}_x^1, \ldots, \hat{F}_x^n)$ represents a "utility" of the set of distributions of possible outcomes $\Psi = \{\Psi_x \mid x \in \Omega\}$. Note that the decision space Ψ is a **set of (possibility) distributions** $\hat{F}_x^i\colon \Omega \to [0,1]$ resulting from taking decision $x \in \Omega$. This semantic is represented by the possibilistic optimization of [37,45,47,50].

Very simply, fuzzy decision making selects from a set of crisp elements ordered by an aggregation operator on corresponding membership functions while possibilistic decision making selects from a set of distributions measured by a utility operator that orders the corresponding distributions. These approaches have two different ordering operators (an aggregation operation like *min* for fuzzy sets and a utility function for possibility) which lead to different optimization methods (see [72]). The underlying sets associated with fuzzy decision making are fuzzy and the decision space consists of crisp elements from operations on these fuzzy sets. The underlying sets associated with possibilistic decision making are crisp and the decision space consists of distributions from operations on crisp sets.

Mixed Fuzzy/Possibilistic Decision Making: The issue of mixed fuzzy and possibility optimization problems has been studied as early as 1989 (see [11,44]). In both these early models the same α -cut defines the level of ambiguity in the possibilistic coefficients and the level at which the decision-maker's requirements are satisfied. We interpret the solution to these models to mean that we have a possibility α of obtaining a solution that satisfies the decision maker to degree α. A recently proposed model [126] allows a trade-off between the fuzzy α-level and the possibilistic α-level. This allows the decision-maker to balance the likelihood of a solution with how satisfactory the solution is.

The use of distinct approaches for mathematical programming problems in which each constraint contains only one single kind of uncertainty is found in [77] and [81]. They aggregate all fuzzy constraint rows according to (12) and find a utility for the possibilistic rows according to (13). If a mixture of uncertainty occurs within one constraint, they apply interval-valued probability measure (IVPM) [78] and [79] to optimization. IVPM applied to optimization is discussed as a separate topic in Sect. "Future Directions".

What is Meant by a Fuzzy-Valued Function and How Does one Compute $\tilde{z} = f(\tilde{c}, x)$? Any function of fuzzy sets (fuzzy-valued function) or possibility distributions must rely on an extension principle, as described in Subsect. "Extension Principles" [71,76], and independently [3]. These authors present practical methods for computing the output of a fuzzy – or possibilistic-valued functions. For this presentation, it is assumed that (8) is used to evaluate fuzzy functions.

What is Meant by a Fuzzy/Possibilistic relation? There are several possible interpretations of fuzzy and possibilistic equalities and inequalities as detailed in Subsect. "Fuzzy Relations". Other comparisons of fuzzy sets may be found in [22,140]. The choice of optimization model for a particular problem depends heavily on which of these relations is used.

What is Meant by Fuzzy Optimization of a Fuzzy-Valued Function? There are three distinct kinds of optimality a fuzzy mathematical program might pursue. *First*, a program might seek to find, among a collection of fuzzy sets, the optimal fuzzy set. This optimality depends on the order dictated by the fuzzy relations described in the previous section.

Secondly, a program may seek to maximize the membership function (or possibility) of the solution chosen for the problem. This typically happens when no fuzzy set satisfies the constraint at a full membership level. Note that $x \tilde{\in} \tilde{X}$ occurs in the transformation of the objective as well as the constraints. This is because some models allow constraint violations at a cost (to the objective function). There are several possible interpretations of fuzzy and possibilistic equalities and inequalities, as detailed in Subsect. "Fuzzy Relations". The choice of optimization model for a particular problem depends heavily on which of these relations is used.

Thirdly, fuzzy optimization problems that use $\mathbb{R}^n$ as the basic space require a mapping of (1) and (2) from fuzzy sets and/or possibility distributions to real numbers and vectors. When the sense of optimization is fuzzy, $\widetilde{\text{opt}}$, optimization is interpreted as a goal, that is, the objective function is considered as a goal. Typically, a target is set and the objective is to come as close as possible to the target. For this presentation, $\widetilde{\text{opt}}$ will be considered as optimization over real numbers. Thus, the general fuzzy optimization problem is:

$$\operatorname*{opt}_x \tilde{z} = f(\tilde{c}, x) \tag{14}$$

$$x \tilde{\in} \tilde{X} . \tag{15}$$

To transform (14) and (15) to $\mathbb{R}^n$, a mapping $T\colon \boldsymbol{E}^n \to \mathbb{R}^{n+1}$ is defined by:

$$\operatorname*{opt}_x T \begin{pmatrix} f(\tilde{c}, x) \\ x \tilde{\in} \tilde{X} \end{pmatrix} = \operatorname*{opt}_x \begin{bmatrix} T_1(f(\tilde{c}, x), x \tilde{\in} \tilde{X}) \\ T_2(x \tilde{\in} \tilde{X}) \end{bmatrix} = \begin{bmatrix} \operatorname*{opt}_x F(c, x) \in \mathbb{R} \\ x \in \Omega \subseteq \mathbb{R}^n \end{bmatrix} . \tag{16}$$

Note that $x \tilde{\in} \tilde{X}$ occurs in the transformation of the objective as well as the constraints. This is because some fuzzy/possibilistic models consider violations of constraints as possible but at a penalty or cost (to the objective function). To make the discussion clearer, the general fuzzy optimization problem (1), (2) is restricted to the fuzzy linear programming model

$$\begin{aligned} \min \tilde{z} &= \tilde{c}^{\mathrm{T}} x \\ \text{subject to: } \tilde{A}x &\tilde{\leq} \tilde{b} \\ x &\geq 0 . \end{aligned} \tag{17}$$

In this context, the *objective* $\widetilde{\text{opt}}\, \tilde{z} = f(\tilde{c}, x)$ becomes

$$\min \tilde{z} = f(\tilde{c}, x) = \tilde{c}^{\mathrm{T}} x . \tag{18}$$

The *constraint* $x \tilde{\in} \tilde{X}$ is

$$\tilde{X} = \{\tilde{A}x \tilde{\leq} \tilde{b}\} \cup \{x \geq 0\} . \tag{19}$$

Interactivity To date, most fuzzy optimization models assume that the fuzzy and possibilistic parameters are non-interactive. Inuiguchi has studied fuzzy optimization with interactivity (dependencies) [39,43]. The issue of interactivity is especially important for mathematical programming models in finance, such as portfolio models where groups of stocks are in fact known to be dependent. However, for this presentation, it is assumed that all uncertainties are non-interactive.

Classical Approaches to Fuzzy Optimization

There have been numerous formulations set forth for modeling imprecise objectives and constraints, each with associated input semantics and solution interpretations. This chapter organizes and reviews a representative selection of fuzzy and possibilistic formulations.

Fuzzy Programming

Vague parameter values leads to fuzzy programming. When the vagueness represents a willingness on the part of the decision-maker to bend the constraints, fuzzy programming can also be called flexible programming. In these cases, the decision-maker is willing to lower the feasibility requirements in order to obtain a more satisfactory objective function value, or, in some cases, simply in order to reach a feasible solution. Such flexible constraints are commonly referred to as *soft constraints*.

Bellman and Zadeh The landmark paper [5] by Bellman and Zadeh is the first to propose an approach to mathematical programming with fuzzy components. Conventional mathematical programming has three principal

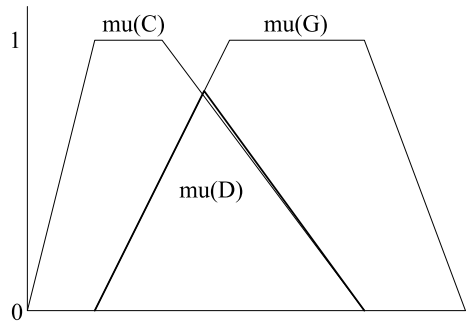

Fuzzy Optimization, Figure 1
Fuzzy Decision Membership Function μ_D

components: A set of alternatives, X, a set of constraints, C, which limit the choice of alternatives, and an objective function, z, which measures the desirability of each alternative. Bellman and Zadeh propose that, in a fuzzy environment, a more natural framework is one in which goal(s) replace the objective function(s), and a symmetry between goals and constraints erases the differences between them. (For this reason, flexible programming is sometimes called symmetric programming.)

Each fuzzy goal, $\tilde{G}_i$, or constraint, $\tilde{C}_j$, is defined as a fuzzy subset of the solution alternatives, X, via a membership function ($\mu_{G_i}(x)$ or $\mu_{C_j}(x)$). Then a fuzzy decision, $\tilde{D}$, is defined as a fuzzy set resulting from the intersection of $\tilde{C}$ and $\tilde{G}$, and is characterized by membership function μ_D as follows:

$$\begin{aligned}\mu_D(x) &= \mu_C(x) \wedge \mu_G(x) \\ &= \min[\mu_C(x), \mu_G(x)]\,. \end{aligned} \tag{20}$$

An optimal decision, in turn, is one which maximizes μ_D. Figure 1 illustrates and provides a visual interpretation of μ_G, μ_C, and μ_D.

Tanaka, Okuda, and Asai [120] Tanaka, Okuda, and Asai suggested an implementation of Bellman and Zadeh's fuzzy decision using *α-cuts* [120]. In the literature, α-cut, α-level, and α-set are used synonymously. If $\tilde{C}$ is a fuzzy constraint in X, then an α-cut of $\tilde{C}$, denoted by C_α, is the following crisp set in X:

$$\begin{aligned} C_\alpha &= \{x \mid \mu_C(x) \geq \alpha\} \quad \text{for } \alpha \in (0,1] \\ C_\alpha &= cls\{x \mid \mu(x) > 0\} \quad \text{for } \alpha = 0 \end{aligned} \tag{21}$$

where $cls\{A\}$ denotes the closure of set $\{A\}$. Tanaka, Okuda and Asai [120] make the following remarkable observation:

$$\sup_x \mu_D(x) = \sup_\alpha [\alpha \wedge \max_{C_\alpha} \mu_G(x)]\,.$$

For illustration, consider two fuzzy sets, $\tilde{C}$ and $\tilde{G}$, depicted in Fig. 2.

In case (i), $\alpha = \alpha_1$, and C_α is the interval between the end-points of the α-cut, $[C^-(\alpha_1), C^+(\alpha_1)]$. The maximum μ_G in this interval is shown in the example. In this case, $\alpha_1 < \max_{C_{\alpha_1}} \mu_G(x)]$, so $[\alpha_1 \wedge \max_{C_{\alpha_1}} \mu_G(x)] = \alpha_1$. In case (ii), $\alpha_2 > \max_{C_{\alpha_2}} \mu_G(x)$, so $[\alpha_2 \wedge \max_{C_{\alpha_2}} \mu_G(x)] = \max_{C_{\alpha_2}} \mu_G(x)$. In case (iii), $\alpha^* = \max_{C_{\alpha^*}} \mu_G(x)$. It should be apparent from Fig. 2 that $\alpha^* = \sup_x \mu_D$. In case (iii), $\alpha = \alpha^*$ is also $\sup_\alpha [\alpha \wedge \max_{C_\alpha} \mu_G(x)]$. For any $\alpha < \alpha_*$, we have case (i), where $[\alpha_1 \wedge \max_{C_{\alpha_1}} \mu_G(x)] = \alpha_1 < \alpha^*$; and for any $\alpha > \alpha_*$, we have case (ii), where $[\alpha_2 \wedge \max_{C_{\alpha_2}} \mu_G(x)] = \max_{C_{\alpha_2}} \mu_G(x) < \alpha^*$. The formal proof, which follows the reasoning illustrated here pictorially, is omitted for brevity's sake. However, the interested reader is referred to Tanaka, Okuda, and Asai [120].

This result allows the following reformulation of the fuzzy mathematical problem:

Determine (α^*, x^*)

$$\alpha^* \wedge f(x^*) = \sup_\alpha [\alpha \wedge \max_{x \in C_\alpha} f(x)]\,. \tag{22}$$

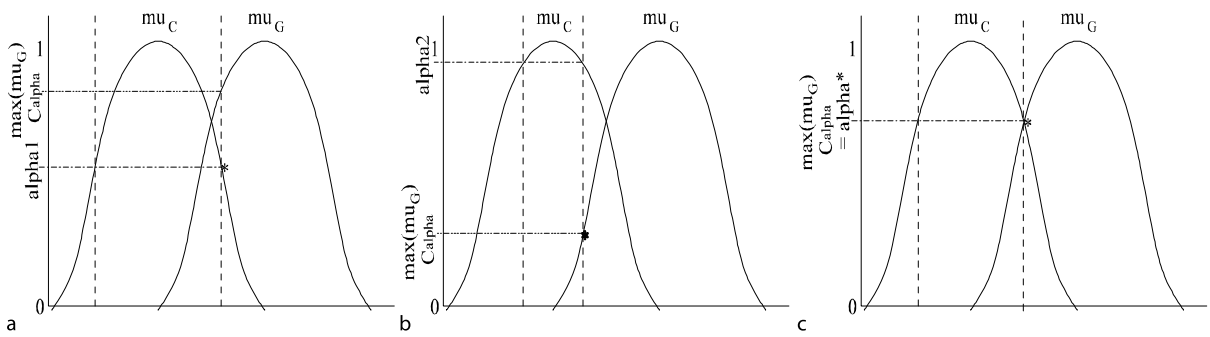

Fuzzy Optimization, Figure 2
Illustration of a Bellman/Zadeh Fuzzy Decision Set

The researchers suggest an iterative algorithm which solves the problem. The algorithm cycles through a series of steps, each of which brings it closer to a solution to the relation $\alpha^* = \max_{C_\alpha^*} f(X)$. When α^* is determined to within a tolerable degree of uncertainty, it is used to find x^* such that $f(x*) = \max_{C_\alpha^*} f(X)$.

This cumbersome solution method is impractical for large-scale problems. It should be noted that Negoita [93] suggested an alternative, but similarly complex, algorithm for the flexible programming problem based on Tanaka's findings.

Zimmermann Just two years after Tanaka, Okuda, and Asai [120] suggested the use of α-cuts to solve the fuzzy mathematical problem, Zimmermann published a linear programming equivalent to the α-cut formulation.

Beginning with the crisp programming problem,

$$\begin{aligned} &\min Z = cx \\ &\text{subject to: } Ax \le b \\ &\qquad x \ge 0\,, \end{aligned} \tag{23}$$

the decision maker introduces flexibility in the constraints, and sets a target value for the objective function, Z,

$$\begin{aligned} cx &\tilde{\le} Z \\ Ax &\tilde{\le} b \\ x &\ge 0\,. \end{aligned} \tag{24}$$

A linear membership function μ_i is defined for each flexible right hand side, including the goal Z as

$$\mu_i\left(\sum_j a_{ij}x_j\right) = \begin{cases} 1 & \sum_j a_{ij}x_j \le b_i, \\ \frac{1-\left(\sum_j a_{ij}x_j - b_i\right)}{d_i} & \sum_j a_{ij}x_j \in (b_i, b_i + d_i), \\ 0 & \sum_j a_{ij}x_j \ge b_i + d_i. \end{cases}$$

where d_i is the decision maker's maximum allowable violation of constraint (or goal) i.

According to the convention set forth by Bellman and Zadeh, the fuzzy decision D is characterized by

$$\mu_D = \min_i \mu_i\left(\sum_j a_{ij}x_j\right), \tag{25}$$

and

$$\max_{x\ge 0} \min_i \mu_i\left(\sum_j a_{ij}x_j\right) \tag{26}$$

is the decision with the highest degree of membership.

The problem of finding the solution is therefore

$$\begin{aligned} &\max \mu_D(x) \\ &\text{subject to: } x \ge 0. \end{aligned} \tag{27}$$

In the membership functions μ_i from (25), Zimmermann substitutes b_i' for b_i/d_i and B_i' for $\sum_j a_{ij}/d_i$. He also drops the 1 (which does not change the solution to the problem) to obtain the simplification $\mu_i = b_i' - B_i'x$. Equation (27) then becomes

$$\begin{aligned} &\min_i (b_i' - B_i'x) \\ &\qquad x \ge 0. \end{aligned} \tag{28}$$

This is equivalent to the following linear program:

$$\begin{aligned} &\max \alpha \\ &\alpha \le b_i' - B_i'x \quad \forall\ i \\ &x \ge 0, \quad 0 \le \alpha \le 1\,. \end{aligned} \tag{29}$$

Equation (29) can easily be solved via the simplex or interior point methods.

Verdegay The solutions examined so far for the fuzzy programming problem have been crisp solutions. Ralescu [99] first suggested that a fuzzy problem should have a fuzzy solution and Verdegay [130] proposes a method for obtaining a fuzzy solution. Verdegay considers a problem with fuzzy constraints,

$$\begin{aligned} &\max z = f(x) \\ &\text{subject to: } x \subseteq \tilde{C}\,, \end{aligned} \tag{30}$$

where the set of constraints have a membership function $\mu_{\tilde{C}}$, with alpha-cuts $\tilde{C}_\alpha$.

Verdegay defines x_α as the set of solutions that satisfy constraints $\tilde{C}_\alpha$. Then a fuzzy solution to the fuzzy linear programming problem is

$$\begin{aligned} &\max_{x\in\tilde{C}_\alpha} z = f(x) \\ &\quad \forall\ \alpha \in [0,1]\,. \end{aligned} \tag{31}$$

Verdegay proposes solving (31) parametrically for $\alpha \in [0,1]$ to obtain a fuzzy solution $\tilde{X}$, with α-cut x_α, which yields fuzzy objective value $\tilde{z}$, with α -cut z_α.

It should be noted that the (crisp) solution obtained via Zimmermann's method corresponds to Verdegay's solution in the following way: If the objective function is transformed into a goal, with membership function μ_G, then Zimmermann's optimal solution, x^*, is equal to Verdegay's optimal value $x(\alpha)$ for the value of α which satisfies

$$\mu_G(z_\alpha^* = cx_\alpha^*) = \alpha\,.$$

In other words, when a particular α-cut of the fuzzy solution, (x_α) yields an objective value $(z_\alpha = c^{\mathrm{T}} x_\alpha)$ whose membership level for goal G, $(\mu_G(z_\alpha))$ is equal to *the same* α, then that solution x_α corresponds to Zimmermann' optimal solution, x^*.

Surprise Recently, Neumaier suggested a way to model fuzzy right-hand side values with crisp inequality constraints $(Ax \leq \tilde{b})$ based on the concept of surprise [83,94]. Neumaier defines a surprise function, $s(x \mid E)$, which corresponds to the amount of surprise a variable x produces, given a statement E. The range of s is $[0, \infty)$, with $s = 0$ corresponding to an entirely true statement E, and $s = \infty$ corresponding to an entirely false statement E.

The most plausible values are those values of x for which $s(x \mid E)$ is minimal, so Neumaier proposes that the best compromise solution for an optimization problem with fuzzy goals and constraints can be found by minimizing the sum of the surprise functions of the goals and constraints. Each fuzzy constraint,

$$(Ax)_i \leq \tilde{b}_i \,,$$

is translated into a fuzzy equality constraint,

$$(Ax)_i = \tilde{\xi}_i \,,$$

where the membership function $\mu_i(\xi)$ of $\tilde{\xi}_i$ is the possibility that $\tilde{b}_i \geq \xi$. These membership functions are subsequently translated into surprise functions by

$$s_i(\xi) = (\mu_i(\xi)^{-1} - 1)^2;$$

and the contribution of all constraints are added to give the total surprise

$$\sum_i s_i(\xi) = \sum_i s_i((Ax)_i) \,.$$

Thus, the fuzzy optimization problem is

$$\min z = \sum_i s_i((Ax)_i) \qquad (32)$$
$$\text{subject to: } x \geq 0 \,.$$

Figure 3 illustrates a surprise function associated with a corresponding fuzzy goal. For triangular and trapezoidal numbers, the surprise function is simple, smooth, and convex, leading to a tractable non-linear programming problem.

The surprise approach is a fuzzy optimization method since the optimization is over sets of crisp values derived from fuzzy sets. In contrast to flexible programming, the constraints are not restricted such that **all** satisfy a minimal level. The salient feature is that surprise uses a dynamic penalty for falling outside distribution/membership values of one. Because the individual penalties are convex functions which become infinite as the values approach the endpoints of the support, this method lends itself to convex programming solution techniques.

It is noted that the surprise approach may be used to handle soft constraints of flexible programming since these soft constraints can be viewed as fuzzy numbers (trapezoidal fuzzy numbers when linear interpolation is used). However, if soft constraints are handled using surprise functions, the **sum** of the failure to meet the constraints is minimized rather than forcing each constraint to meet a minimal feasibility level. One could add a hard constraint to the surprise approach to attain this minimal level of feasibility. The modeler might choose to translate soft constraints to surprise functions (with perhaps a fixed minimal feasibility level) because surprise is usually more computationally efficient than Zimmermann's method of handling soft constraints (see [72]). Therefore, the surprise approach is quite flexible both in terms of semantics as well as in computational robustness.

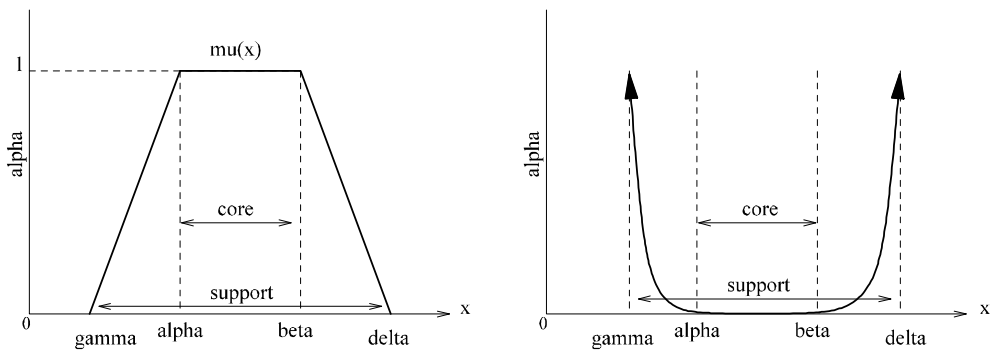

Fuzzy Optimization, Figure 3
Surprise Function

Possibilistic Programming

Recall that fuzzy imprecision arises when elements of a set (for instance, a feasible set) are members of the set to varying degrees, which are defined by the membership function of the set. Possibilistic imprecision arises when elements of a set (say, again, a feasible set) are known to exist as either full members or non-members, but whether they are members or non-members is known with a varying degree of certainty, which is defined by the possibility distribution of the set. This possibilistic uncertainty arises from a lack of information. In this section, we examine possibilistic programming formulations.

Buckley J.J. Buckley has suggested an algorithm for dealing with possibilistic cost coefficients, constraint coefficients, *and* right hand sides. Consider the possibilistic linear program

$$\begin{aligned} &\min Z = \hat{c}x \\ &\text{subject to: } \hat{A}x \geq \hat{b}, x \geq 0\,. \end{aligned} \tag{33}$$

where $\hat{A} = [\tilde{a}_{ij}]$ is an $m \times n$ matrix of trapezoidal possibilistic intervals $\hat{a}_{ij} = (a_{ij\alpha}, a_{ij\beta}, a_{ij\gamma}, a_{ij\delta})$, $\hat{b} = (\tilde{b}_1, \ldots, \tilde{b}_m)^{\mathrm{T}}$ is an $m \times 1$ vector of trapezoidal fuzzy numbers $\hat{b}_i = (b_{i\alpha}, b_{i\beta}, b_{i\gamma}, b_{i\delta})$, and $\hat{c} = (\tilde{c}_1, \ldots, \tilde{c}_n)$ is a $1 \times n$ vector of trapezoidal possibilistic intervals $\tilde{c}_i = (c_{i\alpha}, c_{i\beta}, c_{i\gamma}, c_{i\delta})$. The possibilistic intervals are the possibility distributions associated with the variables, and place a restriction on the possible values the variables may assume [145]. For example, $\mathrm{Poss}[\tilde{a}_{ij} = a] = \mu_a(\tilde{a}_{ij})$ is the possibility that $\tilde{a}_{ij}$ is equal to a. Stated another way, $\mathrm{Poss}[\tilde{a}_{ij} = a] = x$ means that, given the current state of knowledge about the value of a_{ij}, we believe that x is the possibility that variable a_{ij} could take on value a.

Because this is a possibilistic linear programming problem, the objective function will be governed by a possibilistic distribution, $\mathrm{Poss}[Z = z]$. Let us consider this simple example as we follow Buckley's derivation of $\mathrm{Poss}[Z = z]$:

$$\begin{aligned} \min z &= \tilde{d}x_1 + \tilde{e}x_2 \\ \text{subject to } \tilde{f}x_1 + \tilde{g}x_2 &\leq \tilde{h} \\ \tilde{\imath}x_1 + \tilde{\jmath}x_2 &\leq 0 \\ x_1, x_2 &\geq 0 \end{aligned} \tag{34}$$

To derive the possibility function, $\mathrm{Poss}[Z = z]$, Buckley first specifies the possibility that x satisfies the ith constraint. Let

$$\Pi(\hat{a}_i, \hat{b}_i) = \min(\mu_{a_{i1}}(\tilde{a}_{i1}), \ldots, \mu_{a_{in}}(\tilde{a}_{in}), \mu_{b_i}(\tilde{b}_i))\,, \tag{35}$$

which is the simultaneous distribution of $\tilde{a}_{ij}$, $1 \leq j \leq n$, and $\tilde{b}_i$. In our example (34) this corresponds to

$$\Pi(\hat{f}, \hat{g}, \hat{h}) = \min(\mu_F(f), \mu_G(g), \mu_H(h))\,.$$

Then the possibility that $x \geq 0$ is feasible with respect to the ith constraint is:

$$\mathrm{Poss}[x \in \mathcal{F}_i] = \sup_{a_i, b_i}(\Pi(a_i, b_i) \mid a_i x \geq b_i)\,.$$

In our example (34) this corresponds to

$$\begin{aligned} \mathrm{Poss}[x \in \mathcal{F}_1] &= \sup_{f,g,h}(\Pi(f, g, h) \mid fx_1 + gx_2 \leq h) \\ \mathrm{Poss}[x \in \mathcal{F}_2] &= \sup_{i,j,k}(\Pi(i, j, k) \mid ix_1 + jx_2 \leq k)\,. \end{aligned} \tag{36}$$

Now the possibility that $x \geq 0$ is feasible with respect to *all* constraints is

$$[x \in \mathcal{F}] = \min_{1 \leq i \leq m}([x \in \mathcal{F}_i])$$

In our example (34),

$$[x \in \mathcal{F}] = \min([x \in \mathcal{F}_1], [x \in \mathcal{F}_2])$$

Buckley next constructs $\mathrm{Poss}[Z = z \mid x]$, which is the conditional possibility that the objective function Z obtains a particular value z, given values for x. The joint distribution of the possibilistic cost coefficients $\tilde{c}_j$ is

$$\Pi(c) = \min(\mu_{c_1}(\tilde{c}_1), \ldots, \mu_{c_n}(\tilde{c}_n))\,. \tag{37}$$

In our example (34),

$$\Pi(c) = \min(\mu_D(\tilde{d}), \mu_E(\tilde{e}))\,.$$

Therefore

$$\mathrm{Poss}[Z = z \mid x] = \sup_c(\Pi(c) \mid cx = z)\,. \tag{38}$$

In our example (34), to obtain the possibility that the objective function will take on a particular value z, given the solution x, (i. e. $\mathrm{Poss}[Z = z \mid x]$), we consider all the values of d and e for which $dx_1 + ex_2 = z$. Of these, the pair with the highest simultaneous possibility values will yield an objective function value of z with the same possibility level. For example, consider the case illustrated in Fig. 4.

When $\tilde{d} = (2, 2)$ and $\tilde{e} = (3, 2)$ are symmetric triangular fuzzy numbers, $(x_1, x_2) = (1, 1)$ and $z = 7$. We consider values of d and e such that $d \times 1 + e \times 1 = 7$. We could select $d = 2$ and $e = 5$ with $\mu_D(2) = 1$ and

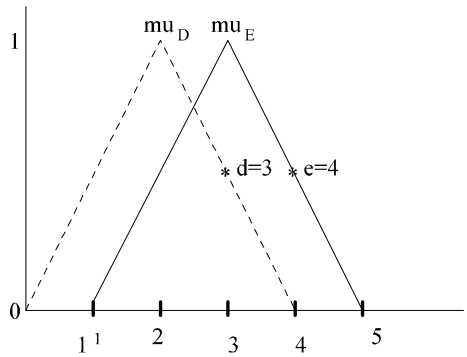

Fuzzy Optimization, Figure 4
Buckley Example

$\mu_E(5) = 0$. The joint possibility of $d = 2$ and $e = 5$ is $\min(\mu_D(2), \mu_E(5)) = 0$. The maximum joint possibility of d and e such that $d \times 1 + e \times 1 = 7$ occurs when $d = 3$ and $e = 4$. The joint possibility that $d = 3$ and $e = 4$, which is $\min(\mu_D(3), \mu_E(4)) = .5$, we set equal to the possibility that $z = 7$ given that $(x_1, x_2) = (1, 1)$. So $\text{Poss}[z = 7 \mid (1, 1)] = .5]$. A combination of (37) and (38) yields the possibility distribution of the objective function:

$$\text{Poss}[Z = z] = \sup_{x \geq 0}[\min(\text{Poss}[Z = z \mid x], \text{Poss}[x \in \mathcal{F}])] . \tag{39}$$

In our example (34), to obtain the possibility that the objective function will take on a particular value z (that is $\text{Poss}[Z = z]$, we consider $\text{Poss}[Z = z \mid x]$, as described above, for all positive x's, and select the x which maximizes $\text{Poss}[Z = z \mid x]$.

This definition of the possibility distribution motivates Buckley's solution method. Recall that because we are dealing with a possibilistic problem, the solution will be governed by a possibilistic distribution. Buckley's method depends upon a static α, chosen a priori. The decision maker defines an acceptable level of uncertainty in the objective outcome, $0 < \alpha \leq 1$. For a given α, we define the left and right end-points of the α-cut of a fuzzy number $\tilde{x}$ as $x^-(\alpha)$ and $x^+(\alpha)$, respectively. Using these, Buckley defines a new objective function:

$$\begin{aligned} &\min Z(\alpha) = c^-(\alpha)x \\ &\text{subject to: } A^+(\alpha)x \geq b^-(\alpha) . \end{aligned} \tag{40}$$

Since both x and α are variables, this is a non-linear programming problem. When α is fixed in advance, however, it becomes linear. We can use either the simplex method or an interior point method to solve for a given a priori chosen value of α. If a maximal value for α is desired, the linear programming method must be applied iteratively.

It should be noted that this linear program is constrained upon the best-case scenario. That is, for a given α-level, each variable is multiplied by the largest possible coefficient $a_{ij}^+(\alpha)$, and is required to be greater than the the smallest possible right hand side $b_i^-(\alpha)$. We should interpret $z(\alpha)$ accordingly. If the solution to the linear program is implemented, the possibility that the objective function will attain the level $z(\alpha)$ is given by α. Stated another way, the best-case scenario is that the objective function attains a value of $z(\alpha)$, and the possibility of the best case scenario occurring is α.

Tanaka, Asai, Ichihashi In the mid 1980s, Tanaka and Asai [118] and Tanaka, Ichahashi, and Asai [121] proposed a technique for dealing with ambiguous coefficients and right hand sides based upon a possibilistic definition of "greater than zero". The reader will note that this approach bears many similarities to the flexible programming proposed by Tanaka, Okuda, and Asai a decade earlier, which was discussed in Subsect. "Tanaka, Okuda, and Asai". Indeed, the 1984 publications refer to *fuzzy* variables. This approach has subsequently been classified [44,60] as possibilistic programming because the imprecision it represents stems from a lack of information about the values of the coefficients.

Consider a programming problem with non-interactive possibilistic $\hat{A}$, $\hat{b}$, and $\hat{c}$, whose possible values are defined by fuzzy matrix $\tilde{A}$, fuzzy vector $\tilde{b}$, and fuzzy vector $\tilde{c}$, respectively:

$$\begin{aligned} &\min \tilde{z} = \tilde{c}x \\ &\text{subject to: } \tilde{A}x \leq \tilde{b} \\ &\qquad x \geq 0 . \end{aligned} \tag{41}$$

Tanaka, Asai, and Ichihashi transform the problem in several steps. First, the objective function is viewed as a goal. As in flexible programming, the goal becomes a constraint, with the aspiration level for the objective function on the right-hand-side of the inequality. Next, a new decision variable x_0 is added. Finally, the b's are moved to the left hand side, so that the possibilistic linear programming problem is

$$\begin{aligned} &\tilde{a}_i' x_i' \geq 0, \quad i = 1, 2, \ldots, m \\ &\quad x' \geq 0 \end{aligned} \tag{42}$$

where $x' = (1, x^{\mathrm{T}})^{\mathrm{T}} = (1, x_1, x_2, \ldots, x_n)^{\mathrm{T}}$, and $\tilde{a}_i = (\tilde{b}_i, \tilde{a}_{i1}, \ldots \tilde{a}_{in})$.

Note that all the parameters, A, b, and c are now grouped together in the new constraint matrix $\tilde{A}$. Because the objective function(s) have become goals, the cost coefficients, c, are part of the constraint coefficient matrix A. Furthermore, because the right-hand side values have been moved to the left-hand side, the b's are also part of the constraint coefficient matrix. The right-hand-sides corresponding the former objective functions are the aspiration levels of the goals.

Each constraint becomes

$$\tilde{Y}_i = \tilde{a}_i x \geq 0\,,$$

where

$$\tilde{a}_i = (\tilde{b}_i, \tilde{a}_{i1}, \ldots, \tilde{a}_{in})\,.$$

"$\tilde{Y}_i$ is almost positive", denoted by $\tilde{Y}_i \tilde{\geq} 0$, is defined by

$$\begin{aligned} &\tilde{Y}_i \tilde{\geq} 0 \\ &\quad\Leftrightarrow \\ &\mu_{Y_i}(0) \leq 1 - h, \\ &\quad x^{\mathrm{T}}\alpha_i \geq 0. \end{aligned} \tag{43}$$

The measure of the non-negativity of $\tilde{Y}_i$ is h: The greater the value of h, the stronger the meaning of "almost positive" (see Fig. 5). Actually, h is $1-\alpha$, where α is the level of membership used by Bellman and Zadeh.

Tanaka and Asai [118] developed this theory for triangular fuzzy numbers, and Tanaka and Asai extended it to trapezoidal fuzzy numbers in [122]. Inuiguchi, Ichihashi, and Tanaka, [44] generalized the approach for L-R fuzzy numbers. For the sake of simplicity, our discussion here will deal with trapezoidal fuzzy numbers, denoted $\tilde{x} = (\gamma, \alpha, \beta, \delta)$, where (γ, δ) is the support of $\tilde{x}$, and (α, β) is the core of $\tilde{x}$.

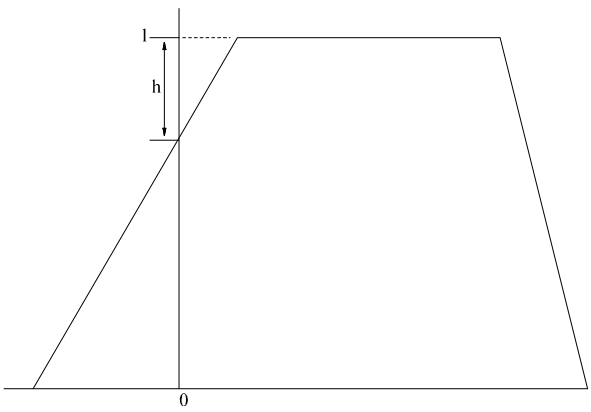

Fuzzy Optimization, Figure 5
h-Level

Using (43), we can rewrite each constraint from (42) as

$$\begin{aligned} \mu_{Y_i}(0) &= 1 - \frac{\alpha_i^{\mathrm{T}} x}{(\alpha_i - \gamma_i)^{\mathrm{T}} x} \\ &\leq 1 - h \\ \alpha_i^{\mathrm{T}} x &\geq 0\,, \end{aligned} \tag{44}$$

where $x > 0$. Then (44) reduces to

$$(\alpha_i - h(\alpha_i - \gamma_i))^{\mathrm{T}} x \geq 0\,.$$

Since we wish to find the largest h that satisfies these conditions, the linear program becomes

$$\begin{aligned} &\max z = h \\ &\text{subject to: } (\alpha_i - h(\alpha_i - \gamma_i))^{\mathrm{T}} x \geq 0\,, \quad \forall i \\ &\qquad h \in [0, 1]\,. \end{aligned} \tag{45}$$

Since both x and h are variables, this is a non-linear programming problem. When h is fixed, it becomes linear. We can use the simplex method or an interior point algorithm to solve for a given value of h. If the decision maker wishes to maximize h, the linear programming method must be applied iteratively.

Fuzzy Max Dubois and Prade [19] suggested that the concept of "fuzzy max" could be applied to constraints with fuzzy parameters. The "fuzzy max" concept was used to solve possibilistic linear programs with triangular possibilistic coefficients by Tanaka, Ichihashi, and Asai [121]. Ramik and Rimanek [101] applied the same technique to L-R fuzzy numbers. For consistency, we will discuss the fuzzy max technique with respect to trapezoidal numbers.

The fuzzy max, illustrated in Fig. 6 where $C = \max[A, B]$, is the extended maximum operator between real numbers, and defined by the extension principle (8) as

$$\mu_C(c) = \max_{\{a,b:\, c = max(a,b)\}} \min[\mu_{\tilde{A}}(a), \mu_{\tilde{B}}(b)]\,.$$

Using fuzzy max, we can define an inequality relation as

$$\begin{aligned} &\tilde{A} \geq \tilde{B} \\ &\quad\Leftrightarrow \\ &\max(\tilde{A}, \tilde{B}) = \tilde{A}\,. \end{aligned} \tag{46}$$

Applying (46) to a fuzzy inequality constraint:

$$\begin{aligned} &f(x, \tilde{a}) \geq g(x, \tilde{b}) \\ &\quad\Leftrightarrow \\ &\max(f(x, \tilde{a}), g(x, \tilde{b})) = f(x, \tilde{a})\,. \end{aligned} \tag{47}$$

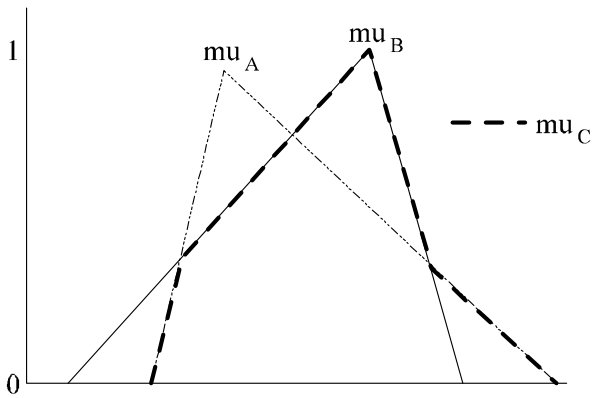

Fuzzy Optimization, Figure 6
Illustration of Fuzzy Max

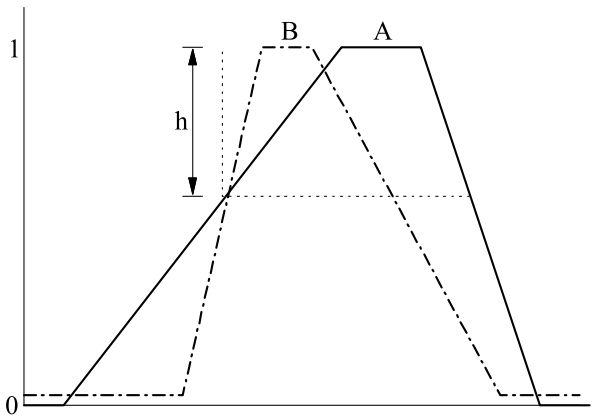

Fuzzy Optimization, Figure 7
Illustration of $\tilde{A} \geq \tilde{B}$

Observe that the inequality relation defined by (46) yields only a partial ordering. That is, it is sometimes the case that neither $\tilde{A} \geq \tilde{B}$ nor $\tilde{B} \geq \tilde{A}$ holds. To improve this, Tanaka, Ichihashi, and Asai, introduce a level h, corresponding to the decision maker's degree of optimism. They define an h-level set of $\tilde{A} \geq \tilde{B}$ as

$$\begin{aligned} &\tilde{A} \geq_h \tilde{B} \\ &\quad\Leftrightarrow \\ &\max[\tilde{A}]_\alpha \geq \max[\tilde{B}]_\alpha \\ &\min[\tilde{A}]_\alpha \geq \min[\tilde{B}]_\alpha \\ &\qquad \alpha \in [1-h, 1]\,. \end{aligned} \tag{48}$$

This definition of $\tilde{A} \geq \tilde{B}$ requires that two of Dubois' inequalities from Subsect. "Fuzzy Relations", (ii) and (iii), hold at the same time, and is illustrated in Fig. 7.

Tanaka, Ichihashi, and Asai, [118] suggest a similar treatment for a fuzzy objective function. A problem with the single objective function, Maximize $z(x, \tilde{c})$, becomes a multi-objective problem with objective functions:

$$\text{maximize} \begin{cases} \inf(z(x,\tilde{c})_\alpha) & \alpha \in [0,1] \\ \sup(z(x,\tilde{c})_\alpha) & \alpha \in [0,1]. \end{cases} \tag{49}$$

Clearly, since α can assume an infinite number of values, (49) has an infinite number of parameters. Since (49) is not tractable, Inuiguchi, Ichihashi, and Tanaka, [44] suggest using the following approximation using a finite set of $\alpha_i \in [0,1]$:

$$\text{maximize} \begin{cases} \min(z(x,\tilde{c}))_{\alpha_i} & i = 1,2,\ldots,p \\ \max(z(x,\tilde{c}))_{\alpha_i} & i = 1,2,\ldots,p \end{cases}. \tag{50}$$

Jamison and Lodwick Jamison and Lodwick [50,74] develop a method for dealing with possibilistic right hand sides that is a possibilistic generalization of the recourse models in stochastic optimization. Violations of constraints are allowable, at a cost determined a priori by the decision maker.

Jamison and Lodwick choose the utility (that is, valuation) of a given interval of possible values to be its expected average (a concept defined by Yager [139].) The expected average (or EA) of a possibilistic distribution of $\tilde{a}$ is defined to be

$$EA(\tilde{a}) = \frac{1}{2}\int_0^1 (\tilde{a}^-(\alpha) + \tilde{a}^+(\alpha))\,\mathrm{d}\alpha. \tag{51}$$

It should be noted that the expected average of a crisp value is the value itself, since $\tilde{a}^-(\alpha) = \tilde{a}^+(\alpha) = a$, $EA(a) = \frac{1}{2}\int_0^1 (a+a)\mathrm{d}x = a\int_0^1 \mathrm{d}x = a$.

Jamison and Lodwick start from the following possibilistic linear program:

$$\begin{aligned} \max\ z &= c^{\mathrm{T}}x \\ Ax &\leq \hat{b}\,, \\ x &\geq 0\,. \end{aligned} \tag{52}$$

By subtracting a penalty term from the objective function, they transform (52) into the following possibilistic nonlinear program:

$$\begin{aligned} \max\ z &= c^{\mathrm{T}}x + p^{\mathrm{T}}\max(0, Ax - \hat{b}) \\ x &\geq 0\,. \end{aligned} \tag{53}$$

The "max" in the penalty is taken component-wise and each $p_i < 0$ is the cost per unit violation of the right hand side of constraint i. The utility, which is the expected average, of the objective function is chosen to be minimized.

The possibilistic programming problem becomes

$$\begin{aligned} &\max z = c^{\mathrm{T}}x + pEA(\max(0, Ax - \hat{b})) \\ &\quad x \in [0, U]. \end{aligned} \tag{54}$$

A closed-form objective function (for the purpose of differentiating when solving) is achieved in [72] by replacing

$$\max(0, Ax - \hat{b})$$

with

$$\frac{\sqrt{(Ax - \hat{b}) + \epsilon^2} + Ax - \hat{b}}{2}.$$

Jamison and Lodwick's method can be extended, [50], to account for possibilistic values for A, b, c, and even the penalty coefficients p with the following formulation:

$$\begin{aligned} EA\tilde{f}(x) = \frac{1}{2}\int_0^1 \Big\{ &\hat{c}^-(\alpha)^{\mathrm{T}}x + \hat{c}^+(\alpha)^{\mathrm{T}}x \\ &- \left[\hat{p}^+(\alpha)\max(0, \hat{A}^-(\alpha)x - \hat{b}^+(\alpha)\right] \\ &- \left[\hat{p}^-(\alpha)\max(0, \hat{A}^+(\alpha)x - \hat{b}^-(\alpha)\right]\Big\}\,\mathrm{d}\alpha\,. \end{aligned} \tag{55}$$

This approach differs significantly from the others we've examined in several regards. First, many of the approaches we've seen have incorporated the objective function(s) as goals into the constraints in the Bellman and Zadeh tradition. Jamison and Lodwick, on the other hand, incorporate the constraints into the objective function. Bellman and Zadeh create a symmetry between constraints and objective, while Jamison and Lodwick temper the objective with the constraints. A second distinction of the expected average approach is the nature of the solution. The other formulations we have examined to this point have produced either (1) a crisp solution for a particular value of α, (namely, the maximal value of α), or, (2) a fuzzy/possibilistic solution which encompasses all possible α values. The Jamison and Lodwick approach provides a crisp solution via the expected average utility which encompasses all alpha values. This may be a desirable quality to the decision maker who wants to account for all possibility levels and still reach a crisp solution.

Luhandjula Luhanjula's [84] formulation of the possibilistic mathematical program depends upon his concept of "more possible" values. He first defines a possibility distribution Π_X with respect to constraint F as

$$\Pi_X = \mu_F(u)\,,$$

where $\mu_F(u)$ is the degree to which the constraint F is satisfied when u is the value assigned to the solution X.

Then the set of more possible values for X, denoted by $V_p(X)$, is given by

$$V_p(X) = \Pi_X^{-1}(\max_u \Pi_X(u))\,.$$

In other words, $V_p(X)$ contains elements of U which are most compatible with the restrictions defined by Π_X. It follows from intuition, and from Luhanjula's formal proof [84], that when Π_X is convex, $V_p(X)$ is a real-valued interval, and when Π_X is strongly convex, $V_p(X)$ is a single real number.

Luhandjula considers the mathematical program

$$\begin{aligned} &\max \tilde{z} = \hat{c}x \\ &\text{subject to: } \hat{A}_i \le \hat{b}_i\,, \\ &\qquad x \ge 0\,. \end{aligned} \tag{56}$$

By replacing the possibilistic numbers $\hat{c}$, $\hat{A}_i$, and $\hat{b}_i$ with their more possible values, $V_p(\hat{c})$, $V_p(\hat{A}_i)$, and $V_p(\hat{b}_i)$, respectively, Luhandjula arrives at a deterministic equivalent to Eq. (56):

$$\begin{aligned} &\max\ z = kx \\ &\text{subject to: } k_i \in V_p(\hat{c}_i) \\ &\sum_i t_i x_i \le s_i \\ &\qquad t_i \in V_p(\hat{a}_{ij}) \\ &\qquad s_i \in V_p(\hat{b}_i) \\ &\qquad x \ge 0\,. \end{aligned} \tag{57}$$

This formulation varies significantly from the other approaches considered thus far. The possibility of each possibilistic component is maximized individually. Other formulations have required that each possibilistic component $\tilde{c}_j$, $\tilde{A}_{ij}$, and $\tilde{b}_i$ achieve the same possibility level defined by α. This formulation also has a distinct disadvantage over the others we've considered, since to date there is no proposed computational method for determining the "more possible" values, V_p, so there is no way to solve the deterministic MP.

Programming with Fuzzy and Possibilistic Components

Sometimes the values of an optimization problem's components are ambiguous *and* the decision-makers are vague (or flexible) regarding feasibility requirements. This sections explores a couple of approaches for dealing with such fuzzy/possibilistic problems.

One type of mixed programming problem that arises (see [20,91]) is a mathematical program with possibilistic constraint coefficients $\hat{a}_{ij}$ whose possible values are defined by fuzzy numbers of the form $\tilde{a}_{ij}$:

$$\begin{aligned} &\max\ cx \\ &\text{subject to: } \hat{a}_i' x' \subseteq \tilde{b}_i \\ &\quad x' = (1, x^t)t \geq 0. \end{aligned} \tag{58}$$

Zadeh [142] defines the set-inclusion relation $\tilde{M} \subseteq \tilde{N}$ as $\mu_{\tilde{M}}(r) \leq \mu_{\tilde{N}}(r) \forall r \in R$. Recall that Dubois [18] interprets the set-inclusive constraint $\tilde{a}_i' x' \subseteq \tilde{b}_i$ as a fuzzy extension of the crisp equality constraint. Mixed programming, however, interprets the set-inclusive constraint to mean that the region in which $\tilde{a}_i' x'$ can possibly occur is restricted to $\tilde{b}_i$, a region which is tolerable to the decision maker. Therefore, the left side of (58) is possibilistic, and the right side is fuzzy.

Negoita [91] defines the fuzzy right hand side as follows:

$$\tilde{b}_i = \{r \mid r \geq b_i\}\,. \tag{59}$$

As a result, we can interpret $\tilde{a}_i' x' \subseteq \tilde{b}_i$ as an extension of an inequality constraint. The set-inclusive constraint (58) is reduced to

$$\begin{aligned} &a_i^+(\alpha)x \leq b_i^+(\alpha) \\ &a_i^-(\alpha)x \geq b_i^-(\alpha) \\ &\text{for all } \alpha \in (0, 1]. \end{aligned} \tag{60}$$

If we abide by Negoita's definition of $\tilde{b}$ (59), $b_i^+ = \infty$ for all values of α, so we can drop the first constraint in (60). Nonetheless, we still have an infinitely (continuum) constrained program, with two constraints for each value of $\alpha \in (0, 1]$. Inuiguchi observes [44] that if the left-hand side of the membership functions for $a_{i0}, a_{i1}, \ldots a_{in}, b_i$ are identical for all i, and the right-hand side of the membership functions for $a_{i0}, a_{i1}, \ldots a_{in}, b_i$ are identical for all i, the constraints are reduced to the finite set,

$$\begin{aligned} &a_{i,\alpha} \geq b_{i,\alpha} \\ &a_{i,\gamma} \geq b_{i,\gamma} \\ &a_{i,\beta} \leq b_{i,\beta} \\ &a_{i,\delta} \leq b_{i,\delta}\,. \end{aligned} \tag{61}$$

As per our usual notation, (γ, δ) is the support of the fuzzy number, and (α, β) is its core. In application, constraint formulation (61) has limited utility because of the narrowly defined sets of memberships functions it admits. For example, if $a_{i0}, a_{i1}, \ldots a_{in}, b_i$ are defined by trapezoidal fuzzy numbers, they must all have the same spread, and therefore the same slope, on the right-hand side; and they must all have the same spread, and therefore the same slope, on the left-hand side if (61) is to be implemented. Recall that in this kind of mixed programming, the a_{ij}'s are possibilistic, reflecting a lack of information about their values, and the b_i are fuzzy, reflecting the decision maker's degree of satisfaction with their possible values. It is possible that n possibilistic components and 1 fuzzy component will share identically-shaped distribution, but it is not something likely to happen with great frequency.

Delgado, Verdegay and Vila Delgado, Verdegay, and Villa [11] propose the following formulation for dealing with ambiguity in the constraint coefficients and right-hand sides, as well as vagueness in the inequality relationship:

$$\begin{aligned} &\max\ cx \\ &\text{subject to: } \hat{A}x \tilde{\leq} \hat{b} \\ &\qquad x \geq 0\,. \end{aligned} \tag{62}$$

In addition to (62), membership functions $\mu_{a_{ij}}$ are defined for the possible values of each possibilistic element of $\hat{A}$, membership functions μ_{b_i} are defined for the possible values of each possibilistic element of $\hat{b}$, and membership function μ_i gives the degree to which the fuzzy constraint i is satisfied. Stated another way, μ_i is the membership function of the fuzzy inequality. The uncertainty in the $\tilde{a}_{ij}$ and the $\tilde{b}_i$ is due to ambiguity concerning the actual value of the parameter, while the uncertainty in the $\tilde{\leq}$ is due to the decision maker's flexibility regarding the necessity of satisfying the constraints in full.

Delgado, Verdegay, and Vila do not define the fuzzy inequality, but leave that specification to decision maker. Any ranking index that preserves ranking of fuzzy numbers when multiplied by a positive scalar is allowed. For instance, one could select any of Dubois' four inequalities from Subsect. "Fuzzy Relations". Once the ranking index is selected, the problem is solved parametrically, as in Verdegay's earlier work [130] (see Subsect. "Verdegay").

To illustrate this approach, let us choose Dubois' pessimistic inequality (i), which interprets $\tilde{A} \geq \tilde{b}$ to mean $\forall x \in A, \forall y \in B, x \geq y$. This is equivalent to $a^+ \geq b^-$. Then (62) becomes

$$\begin{aligned} &\max\ cx \\ &\text{subject to: } a_{ij}^+(\alpha)x \leq b_i^-(\alpha) \\ &\qquad x \geq 0\,. \end{aligned} \tag{63}$$

Fuzzy/Robust Programming The approach covered so far in this section, has the same α cut define the level of am-

biguity in the coefficients *and* the level at which the decision-maker's requirements are satisfied. These α, however, mean very different things. The fuzzy α represents the level at which the decision-maker's requirements are satisfied. The possibilistic α, on the other hand, represents the likelihood that the parameters will take on values which will attain that level. The solution is interpreted to mean that for any value $\alpha \in (0, 1]$ there is a possibility α of obtaining a solution that satisfies the decision maker to degree α. Using the same α value for both the possibilistic and fuzzy components of the problem is convenient, but does not necessarily provide a meaningful model of reality.

A recent model [126] based on Markowitz's mean-variance approach to portfolio optimization (see [89,117]) differentiates between the fuzzy α and the possibilistic α. Markowitz introduced an efficient combination, which has the minimum risk for a return greater than or equal to a given level; or one which has the maximum return for a risk less than or equal to a given level. The decision maker can move among these efficient combinations, or along the efficient frontier, according to her/his degree of risk aversion.

Similarly, in mixed possibilistic and fuzzy programming, one might wish to allow a trade-off between the potential reward of the outcome and the reliability of the outcome, with the weights of the two competing objectives determined by the decision maker's risk aversion. The desire is to obtain an objective function like the following:

$$\max\big[\text{reward} + (\text{risk aversion}) \times (\text{reliability})\big]\,. \tag{64}$$

The reward variable is the α-level associated with the fuzzy constraints and goal(s). It tells the decision maker how satisfactory the solution is. The reliability variable is the α-level associated with the possibilistic parameters. It tells the decision maker how likely it is that the solution will actually be satisfactory. To avoid confusion, let us refer to the fuzzy constraint membership parameter as α and the possibilistic parameter membership level as β.

In addition, let $\mu \in [0, 1]$ be an indicator of the decision maker's valuation of reward and risk-avoidance, with 0 indicating that the decision maker cares exclusively about the reward, and 1 indicating the only risk avoidance is important. Using this notation, the desired objective is

$$\max(1-\mu)\alpha + \mu\beta\,. \tag{65}$$

Suppose we begin with the mixed problem:

$$\begin{aligned} &\max\ \hat{c}^{\mathrm{T}} x \\ &\text{subject to } \hat{A}x \lesssim b \\ &\qquad\qquad x \geq 0. \end{aligned} \tag{66}$$

Incorporating fuzziness from soft constraints in the tradition of Zimmermann and incorporating a pessimistic view of possibility results in the following formulation:

$$\max \quad (1-\mu)\alpha + \mu\beta \tag{67}$$

$$\begin{aligned} \text{subject to} \quad & \alpha \leq -\frac{g}{d_0} + \sum_j \frac{u_j}{d_0} x_j + \sum_j \frac{u_j - w_j}{d_0} x_j \beta \\ & \alpha \leq \frac{b_i}{d_i} - \sum_j \frac{v_{ij}}{d_i} x_j - \sum_j \frac{z_{ij} - v_{ij}}{d_i} x_j \beta \\ & x \geq 0 \\ & \alpha, \beta \in [0, 1]\,. \end{aligned} \tag{68}$$

The last terms in each of the constraints contain βx, so the system is non-linear. It can be fairly easily reformulated as an optimization program with linear objective function and quadratic constraints, but the feasible set is non-convex, so finding to solution to this mathematical programming problem is very difficult.

Possibilistic, Interval, Cloud, and Probabilistic Optimization Utilizing IVPM

This section is taken from [79,80] and begins by defining what is meant by an IVPM. This generalization of a probability measure includes probability measures, possibility/necessity measures, intervals, and clouds (see [95]) which will allow a mixture of uncertainty within one constraint (in)equality. The previous mixed methods was restricted to a single type of uncertainty for any particular (in)equality and are unable to handle cases in which a mixture of fuzzy and possibilistic parameter occurs in the same constraint (in)equality.

The IVPM set function may be thought of as a method for giving a partial representation for an unknown probability measure. Throughout, arithmetic operations involving set functions are in terms of interval arithmetic [90] and the set of all intervals contained in $[0, 1]$ is denoted, $\mathrm{Int}_{[0,1]} \equiv \{[a, b] \mid 0 \leq a \leq b \leq 1\}$. Moreover, S is used to denote the universal set and a set of subsets of the universal set is denoted as $\mathcal{A} \subseteq S$. In particular $\mathcal{A}$ is a set of subset on which a structure has been imposed on it as will be seen and a generic set of the structure $\mathcal{A}$ is denoted by A.

Definition 18 (Weichselberger [135]) Given measurable space $(S, \mathcal{A})$, an interval valued function $i_m : A \subseteq \mathcal{A} \to \mathrm{Int}_{[0,1]}$ is called an **R-probability** if:

(a) $i_m(A) = [i_m^-(A), i_m^+(A)] \subseteq [0, 1]$ with $i_m^-(A) \leq i_m^+(A)$,

(b) $\exists$ a probability measure Pr on $\mathcal{A}$ such that $\forall A \in \mathcal{A}$, $\Pr(A) \in i_m(A)$. By an **R-probability field** we mean the triple $(S, \mathcal{A}, i_m)$.

Definition 19 (Weichselberger [135]) Given an R-probability field $\mathcal{R} = (S, \mathcal{A}, i_m)$ the set

$$\mathcal{M}(\mathcal{R}) = \{\Pr \mid \Pr \text{ is a probability measure on } \mathcal{A} \text{ such that } \forall A \in \mathcal{A}, \Pr(A) \in i_m(A)\}$$

is called the **structure** of $\mathcal{R}$.

Definition 20 (Weichselberger [135]) An R-probability field $\mathcal{R} = (S, \mathcal{A}, i_m)$ is called an **F-probability field** if $\forall A \in \mathcal{A}$:

(a) $i_m^+(A) = \sup\{\Pr(A) \mid \Pr \in \mathcal{M}(\mathcal{R})\}$,
(b) $i_m^-(A) = \inf\{\Pr(A) \mid \Pr \in \mathcal{M}(\mathcal{R})\}$.

It is interesting to note that given a measurable space $(S, \mathcal{A})$ and a set of probability measures P, then defining $i_m^+(A) = \sup\{\Pr(A) \mid \Pr \in P\}$ and $i_m^-(A) = \inf\{\Pr(A) \mid \Pr \in P\}$ gives an F-probability and that P is a subset of the structure.

The following examples show how intervals, possibility distributions, clouds and (of course) probability measures can define R-probability fields on $\mathcal{B}$, the Borel sets on the real line.

Example 21 (An interval defines an F-probability field) Let $I = [a, b]$ be a non-empty interval on the real line. On the Borel sets define

$$i_m^+(A) = \begin{cases} 1 & \text{if } I \cap A \neq \emptyset \\ 0 & \text{otherwise} \end{cases}$$

and

$$i_m^-(A) = \begin{cases} 1 & \text{if } I \subseteq A \\ 0 & \text{otherwise} \end{cases}$$

then

$$i_m(A) = \left[i_m^-(A), i_m^+(A)\right]$$

defines an F-probability field $\mathcal{R} = (R, \mathcal{B}, i_m)$. To see this, simply let P be the set of all probability measures on $\mathcal{B}$ such that $\Pr(I) = 1$.

This example also illustrates that any set A, not just an interval I, can be used to define an F-probability field.

Example 22 (A probability measure is an F-probability field) Let Pr be a probability measure over $(S, \mathcal{A})$. Define

$$i_m(A) = [\Pr(A), \Pr(A)] .$$

This definition of a probability as an IVPM is equivalent to having total knowledge about a probability distribution over S. The concept of a cloud was introduced by Neumaier in [95] as follows:

Definition 23 A **cloud** over set S is a mapping c such that:

1) $\forall s \in S$, $c(s) = \left[\underline{n}(s), \bar{p}(s)\right]$ with $0 \leq \underline{n}(s) \leq \bar{p}(s) \leq 1$
2) $(0, 1) \subseteq \cup_{s \in S} c(s) \subseteq [0, 1]$ In addition, random variable X taking values in S is said to belong to cloud c (written $X \in c$) iff
3) $\forall \alpha \in [0, 1]$, $\Pr(\underline{n}(X) \geq \alpha) \leq 1 - \alpha \leq \Pr\left(\bar{p}(X) > \alpha\right)$

Property 3) above defines when a random variable belongs to a cloud. That any cloud contains a random variable X is proved in section 5 of [96]. This is a significant result as will be seen, since among other things, it means that clouds can be used to define IVPMs. Clouds are closely related to possibility theory.

It is shown in [51] that possibility distributions can be constructed which satisfy the following consistency definition.

Definition 24 ([51]) Let $p: S \to [0, 1]$ be a possibility distribution function with associated possibility measure Pos and necessity measure Nec. Then p is said to be **consistent** with random variable X if $\forall$ measurable sets A, $\text{Nec}(A) \leq \Pr(X \in A) \leq \text{Pos}(A)$.

Possibility distributions constructed in a consistent manner are able to bound (unknown) probabilities of interest. The reason this is significant is twofold. Firstly, possibility and necessity distributions are easier to construct since the axioms they satisfy are more general. Secondly, the algebra on possibility and necessity pairs are much simpler since they are min/max algebras akin to the min/max algebra of interval arithmetic (see, for example, [76]). In particular, they avoid convolutions which are requisite for probabilistic arithmetic. The concept of a cloud can be stated in terms of certain pairs of consistent possibility distributions as shown by the following proposition (which means that clouds may be considered as pairs of consistent possibilities – possibility and necessity pairs, see [51]).

Proposition 25 *Let $\bar{p}, \underline{p}$ be a pair of regular possibility distribution functions over set S such that $\forall s \in S \bar{p}(s) + \underline{p}(s) \geq 1$. Then the mapping $c(s) = \left[\underline{n}(s), \bar{p}(s)\right]$ where $\underline{n}(s) = 1 - \underline{p}(s)$ (i. e. the dual necessity distribution function) is a cloud. In addition, if X is a random variable taking values in S and the possibility measures associated with $\bar{p}, \underline{p}$ are consistent with X then X belongs to cloud c. Conversely, every cloud defines such a pair of possibility distribution functions and their associated possibility measures are consistent with every random variable belonging to c.*

Proof (see [52,78,79]) □

Example 26 (A cloud defines an R-probability field) Let c be a cloud over the real line. Let $\mathrm{Pos}^1, \mathrm{Nec}^1, \mathrm{Pos}^2, \mathrm{Nec}^2$ be the possibility measures and their dual necessity measures relating to $\bar{p}(s)$ and $\underline{p}(s)$ (where $\bar{p}$ and $\underline{p}$ are as in Proposition 18). Define

$$i_m(A) = \left[\max\left\{\mathrm{Nec}^1(A), \mathrm{Nec}^2(A)\right\}, \min\left\{\mathrm{Pos}^1(A), \mathrm{Pos}^2(A)\right\}\right].$$

Neumaier [96] proved that every cloud contains a random variable X. Since consistency requires that $\Pr(X \in A) \in i_m(A)$, the result that every cloud contains a random variable X shows consistency. Thus every cloud defines an R-probability field because the inf and sup of the probabilities are bounded by the lower and upper bounds of $i_m(A)$.

Example 27 (A possibility distribution defines an R-probability field) Let $p: S \to [0,1]$ be a possibility distribution function and let Pos be the associated possibility measure and Nec the dual necessity measure. Define $i_m(A) = [\mathrm{Nec}(A), \mathrm{Pos}(A)]$. Defining a second possibility distribution, $\underline{p}(x) = 1 \forall x$ means that the pair $p, \underline{p}$ define a cloud for which $i_m(A)$ defines the R-probability. Since a cloud defines an R-probability field, this means that this possibility in turn generates a R-probability.

Note that the above example means that every pair of possibility $p(x)$ and necessity $n(x)$ such that $n(x) \le p(x)$ has an associated F-probability field. This is because such a pair defines a cloud and in every cloud there exists a probability distribution. The F-probability field can then be constructed from a *inf*/*sup* over all such enclosed probabilities that are less than or equal to the bounding necessity/possibility distributions.

The application of these concepts to mixed fuzzy and possibilistic optimization is as follows. Suppose the optimization problem is to maximize $f(\vec{x}, \vec{a})$ subject to $g(\vec{x}, \vec{b}) = 0$ (where $\vec{a}$ and $\vec{b}$ are parameters). Assume $\vec{a}$ and $\vec{b}$ are vectors of independent uncertain parameters, each with an associated IVPM. Assume the constraint may be violated at a cost $\vec{p} > 0$ so that the problem becomes one to maximize

$$h(\vec{x}, \vec{a}, \vec{b}) = f(\vec{x}, \vec{a}) - \vec{p}\left|g(\vec{x}, \vec{b})\right|.$$

Given the independence assumption, form an IVPM for the product space $i_{\vec{a}\times\vec{b}}$ for the joint distribution (see the example below and [52,78,79]) and calculate the interval-valued expected value (see [48,140]) with respect to this IVPM. The interval-valued expected value is denoted (there is a lower expected value and an upper expected value)

$$\int_R h(\vec{x}, \vec{a}, \vec{b})\,\mathrm{d}i_{\vec{a}\times\vec{b}}\,. \tag{69}$$

To optimize (69) requires an ordering of intervals, a *valuation function* denoted by $v: \mathrm{Int}_R \to \mathbb{R}^n$. One such ordering is the midpoint of the interval on the principle that in the absence of additional data, the midpoint is the best estimate for the true value so that for this function (midpoint and width), $v: \mathrm{Int}_R \to \mathbb{R}^2$. Thus, for $I = [a, b]$, this particular valuation function is $v(I) = ((a+b)/2, b-a)$. Next a *utility function* of a vector in $\mathbb{R}^n$ is required which is denote by $u: \mathbb{R}^n \to \mathbb{R}$. A utility function operating on the midpoint and width valuation function is $u: \mathbb{R}^2 \to \mathbb{R}$ and particular utility is a weighted sum of the midpoint and width $u(c, d) = \alpha c + \beta d$. Using the valuation and utility functions, the optimization problem is:

$$\max_x u\left(v\left(\int_R h(x, a, b)\,\mathrm{d}i_{a\times b}\right)\right). \tag{70}$$

Thus there are four steps to obtaining a real-valued optimization problem from an IVPM problem. The first step is to obtain interval probability $h(x, a, b)$. The second step is to obtain the IVPM expected value $\int_R h(\vec{x}, \vec{a}, \vec{b})\mathrm{d}i_{\vec{a}\times\vec{b}}$. The third step is to obtain the vector value of $v: \mathrm{Int}_R \to \mathbb{R}^n$. The fourth step is to obtain the utility function value of a vector $u: \mathbb{R}^n \to \mathbb{R}$.

Example 28 Consider the problem

$$\begin{aligned}
&\max f(x, a) = 8x_1 + 7x_2\\
&\quad \text{subject to:}\\
&\qquad g_1(x, b) = 3x_1 + [1, 3]x_2 - 4 = 0\\
&\qquad g_2(x, b) = \tilde{2}x_1 + 5x_2 - 1 = 0\\
&\qquad\qquad \vec{x} \in [0, 2]
\end{aligned}$$

where $\tilde{2} = 1/2/3$, that is, $\tilde{2}$ is a triangular possibilistic number with support [1,3] and modal value at 2. For $\vec{p} = (1, 1)^{\mathrm{T}}$,

$$h(x, a, b) = 5x_1 - \tilde{2}x_1 + [3, 5]x_2 - 6$$

so that

$$\begin{aligned}
&\int_R h(x, a, b)\,\mathrm{d}i_{a\times b}\\
&\qquad = 5x_1 + \left[\int_0^1 (\alpha - 3)\,\mathrm{d}\alpha, \int_0^1 (-1 - \alpha)\,\mathrm{d}\alpha\right] x_1\\
&\qquad\quad + [3, 5]x_2 - 6\\
&\qquad = 5x_1 + \left[-\frac{5}{2}, -\frac{3}{2}\right] x_1 + [3, 5]x_2 - 5\,.
\end{aligned}$$

Since the constant -5 will not affect the optimization, it will be removed (then added at the end), so that

$$v\left(\int_R h(x,a,b)\,\mathrm{d}i_{a\times b}\right) = v\left(\left[\frac{5}{2},\frac{7}{2}\right]x_1 + [3,5]x_2\right)$$
$$= (3,1)x_1 + (4,2)x_2\,.$$

Let $u(\vec{y}) = \sum_{i=1}^{n} y_i$ which for the context of this problem yields

$$\begin{aligned}\max_x z &= u\left(v\left(\int_R h(x,a,b)\,\mathrm{d}i_{a\times b}\right)\right)\\ &= \max_{\vec{x}\in[0,2]}(4x_1+6x_2)-5\\ &= 20-5=15\\ x_1^* &= 2,\ x_2^* = 2\,.\end{aligned}$$

Example 29 (see [124]) Consider

$$\begin{aligned}\max z &= \hat{2}x_1 - \bar{0}x_2 + [3,5]x_3\\ \hat{4}x_1 + [1,5]x_2 - 2x_3 - [0,2] &= 0\\ 6x_1 - \bar{2}x_2 + 9x_3 - 9 &= 0\\ -2x_1 - [1,4]x_2 - \hat{8}x_3 + \bar{5} &= 0\\ 0 \le x_1 \le 3, 1 \le x_2, x_3 &\le 2\,.\end{aligned}$$

Here the $\bar{0}$, $\bar{2}$, and $\bar{5}$ are probability distributions. Note the mixture of three uncertainty types in the third constraint equation. Using the same approach as in the previous example, the optimal values are:

$$\begin{aligned}z^* &= 3.9179\\ x_1^* &= 0\,,\quad x_2^* = 0.4355\,,\quad x_3^* = 1.0121\,.\end{aligned}$$

In [124] it is shown that these mixed problems arising from linear programs remain linear programs. Thus, the complexity of mixed problems is equivalent to that of linear programming.

Future Directions

The applications of fuzzy optimization seems to be headed toward "industrial strength" problems. Increasingly, each year there are a greater number of applications that appear. Given that greater attention is being given to the semantics of fuzzy optimization and as fuzzy optimization becomes increasingly used in applications, associated algorithms that are more sophisticated, robust, and efficient will need to be developed to handle these more complex problems. It would be interesting to develop modeling languages like GAMS [6], MODLER [34], or AMPL [30], that support fuzzy data structures. From the theoretical side, the flexibility that fuzzy optimization has with working with uncertainty data that is fuzzy, flexible, and/or possibilistic (or a mixture of these via IVPM), means that fuzzy optimization is able to provide an ample approach to optimization under uncertainty. Further research into the development of more robust methods that use fuzzy Banach spaces would certainty provide a deeper theoretical foundation to fuzzy optimization. Fuzzy optimization that utilize fuzzy Banach spaces have the advantage that the problem remains fuzzy throughout and only when one needs to make a decision or implement the solution does one map the solution to a real number (defuzzify). The methods that map fuzzy optimization problems to their real number equivalent defuzzify first and then optimize. Fuzzy optimization problems that optimize in fuzzy Banach spaces keep the solution fuzzy and defuzzify as a last step.

Continued development of clear input and output semantics of fuzzy optimization will greatly aid fuzzy optimization's applicability and relevance. When fuzzy optimization is used in, for example, an assembly-line scheduling problem and one's solution is a fuzzy three, how does one convey this solution to the assembly-line manager? Lastly, continued research into handling dependencies in an efficient way would amplify the usefulness and applicability of fuzzy optimization.

Bibliography

1. Asai K, Tanaka H (1973) On the fuzzy – mathematical programming, Identification and System Estimation Proceedings of the 3rd IFAC Symposium, The Hague, 12–15 June 1973, pp 1050–1051
2. Audin J-P, Frankkowska H (1990) Set-Valued Analysis. Birkhäuser, Boston
3. Baudrit C, Dubois D, Fargier H (2005) Propagation of uncertainty involving imprecision and randomness. ISIPTA 2005, Pittsburgh, pp 31–40
4. Bector CR, Chandra S (2005) Fuzzy Mathematical Programming and Fuzzy Matrix Games. Springer, Berlin
5. Bellman RE, Zadeh LA (1970) Decision-Making in a Fuzzy Environment. Manag Sci B 17:141–164
6. Brooke A, Kendrick D, Meeraus A (2002) GAMS: A User's Guide. Scientific Press, San Francisco (the latest updates can be obtained from http://www.gams.com/)
7. Buckley JJ (1988) Possibility and necessity in optimization. Fuzzy Sets Syst 25(1):1–13
8. Buckley JJ (1988) Possibilistic linear programming with triangular fuzzy numbers. Fuzzy Sets Syst 26(1):135–138
9. Buckley JJ (1989) Solving possibilistic linear programming problems. Fuzzy Sets Syst 31(3):329–341
10. Buckley JJ (1989) A generalized extension principle. Fuzzy Sets Syst 33:241–242
11. Delgado M, Verdegay JL, Vila MA (1989) A general model for fuzzy linear programming. Fuzzy Sets Syst 29:21–29
12. Delgado M, Kacprzyk J, Verdegay J-L, Vila MA (eds) (1994) Fuzzy Optimization: Recent Advances. Physica, Heidelberg

13. Demster AP (1967) Upper and lower probabilities induced by multivalued mapping. Ann Math Stat 38:325–339
14. Dempster MAH (1969) Distributions in interval and linear programming. In: Hansen ER (ed) Topics in Interval Analysis. Oxford Press, Oxford, pp 107–127
15. Diamond P (1991) Congruence classes of fuzzy sets form a Banach space. J Math Anal Appl 162:144–151
16. Diamond P, Kloeden P (1994) Metric Spaces of Fuzzy Sets. World Scientific, Singapore
17. Diamond P, Kloeden P (1994) Robust Kuhn–Tucker conditions and optimization under imprecision. In: Delgado M, Kacprzyk J, Verdegay J-L, Vila MA (eds) Fuzzy Optimization: Recent Advances. Physica, Heidelberg, pp 61–66
18. Dubois D (1987) Linear programming with fuzzy data. In: Bezdek JC (ed) Analysis of Fuzzy Information, vol III: Applications in Engineering and Science. CRC Press, Boca Raton, pp 241–263
19. Dubois D, Prade H (1980) Fuzzy Sets and Systems: Theory and Applications. Academic Press, New York
20. Dubois D, Prade H (1980) Systems of linear fuzzy constraints. Fuzzy Sets Syst 3:37–48
21. Dubois D, Prade H (1981) Additions of interactive fuzzy numbers. IEEE Trans Autom Control 26(4):926–936
22. Dubois D, Prade H (1983) Ranking fuzzy numbers in the setting of possibility theory. Inf Sci 30:183–224
23. Dubois D, Prade H (1986) New results about properties and semantics of fuzzy set-theoretical operators. In: Wang PP, Chang SK (eds) Fuzzy Sets. Plenum Press, New York, pp 59–75
24. Dubois D, Prade H (1987) Fuzzy numbers: An overview. Tech. Rep. no 219, (LSI, Univ. Paul Sabatier, Toulouse, France). Mathematics and Logic. In: James Bezdek C (ed) Analysis of Fuzzy Information, vol 1, chap 1. CRC Press, Boca Raton, pp 3–39
25. Dubois D, Prade H (1988) Possibility Theory an Approach to Computerized Processing of Uncertainty. Plenum Press, New York
26. Dubois D, Prade H (2005) Fuzzy elements in a fuzzy set. Proceedings of the 11th International Fuzzy System Association World Congress, IFSA 2005, Beijing, July 2005, pp 55–60
27. Dubois D, Moral S, Prade H (1997) Semantics for possibility theory based on likelihoods. J Math Anal Appl 205:359–380
28. Ertuğrul I, Tuş A (2007) Interactive fuzzy linear programming and an application sample at a textile firm. Fuzzy Optim Decis Making 6:29–49
29. Fortin J, Dubois D, Fargier H (2008) Gradual numbers and their application to fuzzy interval analysis. IEEE Trans Fuzzy Syst 16:2, pp 388–402
30. Fourer R, Gay DM, Kerninghan BW (1993) AMPL: A Modeling Language for Mathematical Programming. Scientific Press, San Francisco, CA (the latest updates can be obtained from http://www.ampl.com/)
31. Fullér R, Keresztfalvi T (1990) On generalization of Nguyen's theorem. Fuzzy Sets Syst 41:371–374
32. Ghassan K (1982) New utilization of fuzzy optimization method. In: Gupta MM, Sanchez E (eds) Fuzzy Information and Decision Processes. North-Holland, Netherlands, pp 239–246
33. Gladish B, Parra M, Terol A, Uría M (2005) Management of surgical waiting lists through a possibilistic linear multiobjective programming problem. Appl Math Comput 167:477–495
34. Greenberg H (1992) MODLER: Modeling by Object-Driven Linear Elemental Relations. Ann Oper Res 38:239–280
35. Hanss M (2005) Applied Fuzzy Arithmetic. Springer, Berlin
36. Hsu H, Wang W (2001) Possibilistic programming in production planning of assemble-to-order environments. Fuzzy Sets Syst 119:59–70
37. Inuiguchi M (1992) Stochastic Programming Problems Versus Fuzzy Mathematical Programming Problems. Jpn J Fuzzy Theory Syst 4(1):97–109
38. Inuiguchi M (1997) Fuzzy linear programming: what, why and how? Tatra Mt Math Publ 13:123–167
39. Inuiguchi M (2007) Necessity measure optimization in linear programming problems with fuzzy polytopes. Fuzzy Sets Syst 158:1882–1891
40. Inuiguchi M (2007) On possibility/fuzzy optimization. In: Melin P, Castillo O, Aguilar LT, Kacprzyk J, Pedrycz W (eds) Foundations of Fuzzy Logic and Soft Computing: 12th International Fuzzy System Association World Congress, IFSA 2007, Cancun, June 2007, Proceedings. Springer, Berlin, pp 351–360
41. Inuiguchi M, Ramik J (2000) Possibilistic linear programming: A brief review of fuzzy mathematical programming and a comparison with stochastic programming in portfolio selection problem. Fuzzy Sets Syst 111:97–110
42. Inuiguchi M, Sakawa M (1997) An achievement rate approach to linear programming problems with an interval objective function. J Oper Res Soc 48:25–33
43. Inuiguchi M, Tanino T (2004) Fuzzy linear programming with interactive uncertain parameters. Reliab Comput 10(5):512–527
44. Inuiguchi M, Ichihashi H, Tanaka H (1990) Fuzzy programming: A survey of recent developments. In: Slowinski R, Teghem J (eds) Stochastic versus Fuzzy Approaches to Multiobjective Mathematical Programming Under Uncertainty. Kluwer, Netherlands, pp 45–68
45. Inuiguchi M, Ichihashi H, Kume Y (1992) Relationships Between Modality Constrained Programming Problems and Various Fuzzy Mathematical Programming Problems. Fuzzy Sets Syst 49:243–259
46. Inuiguchi M, Ichihashi H, Tanaka H (1992) Fuzzy Programming: A Survey of Recent Developments. In: Slowinski R, Teghem J (eds) Stochastic versus Fuzzy Approaches to Multiobjective Mathematical Programming under Uncertainty.Springer, Berlin, pp 45–68
47. Inuiguchi M, Sakawa M, Kume Y (1994) The usefulness of possibilistic programming in production planning problems. Int J Prod Econ 33:49–52
48. Jamison KD (1998) Modeling Uncertainty Using Probabilistic Based Possibility Theory with Applications to Optimization. Ph D Thesis, University of Colorado Denver, Department of Mathematical Sciences. http://www-math.cudenver.edu/graduate/thesis/jamison.pdf
49. Jamison KD (2000) Possibilities as cumulative subjective probabilities and a norm on the space of congruence classes of fuzzy numbers motivated by an expected utility functional. Fuzzy Sets Syst 111:331–339
50. Jamison KD, Lodwick WA (2001) Fuzzy linear programming using penalty method. Fuzzy Sets Syst 119:97–110
51. Jamison KD, Lodwick WA (2002) The construction of consistent possibility and necessity measures. Fuzzy Sets Syst 132(1):1–10

52. Jamison KD, Lodwick WA (2004) Interval-valued probability measures. UCD/CCM Report No. 213, March 2004
53. Joubert JW, Luhandjula MK, Ncube O, le Roux G, de Wet F (2007) An optimization model for the management of South African game ranch. Agric Syst 92:223–239
54. Kacprzyk J, Orlovski SA (eds) (1987) Optimization Models Using Fuzzy Sets and Possibility Theory. D Reidel, Dordrecht
55. Kacprzyk J, Orlovski SA (1987) Fuzzy optimization and mathematical programming: A brief introduction and survey. In: Kacprzyk J, Orlovski SA (eds) Optimization Models Using Fuzzy Sets and Possibility Theory. D Reidel, Dordrecht, pp 50–72
56. Kasperski A, Zeilinski P (2007) Using gradual numbers for solving fuzzy-valued combinatorial optimization problems. In: Melin P, Castillo O, Aguilar LT, Kacprzyk J, Pedrycz W (eds) Foundations of Fuzzy Logic and Soft Computing: 12th International Fuzzy System Association World Congress, IFSA 2007, Cancun, June 2007, Proceedings. Springer, Berlin, pp 656–665
57. Kaufmann A, Gupta MM (1985) Introduction to Fuzzy Arithmetic – Theory and Applications. Van Nostrand Reinhold, New York
58. Kaymak U, Sousa JM (2003) Weighting of Constraints in Fuzzy Optimization. Constraints 8:61–78 (also in the 2001 Proceedings of IEEE Fuzzy Systems Conference)
59. Klir GJ, Yuan B (1995) Fuzzy Sets and Fuzzy Logic. Prentice Hall, Upper Saddle River
60. Lai Y, Hwang C (1992) Fuzzy Mathematical Programming. Springer, Berlin
61. Lin TY (2005) A function theoretic view of fuzzy sets: New extension principle. In: Filev D, Ying H (eds) Proceedings of NAFIPS05
62. Liu B (1999) Uncertainty Programming. Wiley, New York
63. Liu B (2000) Dependent-chance programming in fuzzy environments. Fuzzy Sets Syst 109:97–106
64. Liu B (2001) Fuzzy random chance-constrained programming. IEEE Trans Fuzzy Syst 9:713–720
65. Liu B (2002) Theory and Practice of Uncertainty Programming. Physica, Heidelberg
66. Liu B, Iwamura K (2001) Fuzzy programming with fuzzy decisions and fuzzy simulation-based genetic algorithm. Fuzzy Sets Syst 122:253–262
67. Lodwick WA (1990) Analysis of structure in fuzzy linear programs. Fuzzy Sets Syst 38:15–26
68. Lodwick WA (1990) A generalized convex stochastic dominance algorithm. IMA J Math Appl Bus Ind 2:225–246
69. Lodwick WA (1999) Constrained Interval Arithmetic. CCM Report 138, Feb. 1999. CCM, Denver
70. Lodwick WA (ed) (2004) Special Issue on Linkages Between Interval Analysis and Fuzzy Set Theory. Reliable Comput 10
71. Lodwick WA (2007) Interval and fuzzy analysis: A unified approach. In: Advances in Imaging and Electronic Physics, vol 148. Elsevier, San Diego, pp 75–192
72. Lodwick WA, Bachman KA (2005) Solving large-scale fuzzy and possibilistic optimization problems: Theory, algorithms and applications. Fuzzy Optim Decis Making 4(4):257–278. (also UCD/CCM Report No. 216, June 2004)
73. Lodwick WA, Inuiguchi M (eds) (2007) Special Issue on Optimization Under Fuzzy and Possibilistic Uncertainty. Fuzzy Sets Syst 158:17; 1 Sept 2007
74. Lodwick WA, Jamison KD (1997) A computational method for fuzzy optimization. In: Bilal A, Madan G (eds) Uncertainty Analysis in Engineering and Sciences: Fuzzy Logic, Statistics, and Neural Network Approach, chap 19. Kluwer, Norwell
75. Lodwick WA, Jamison KD (eds) (2003) Special Issue: Interfaces Fuzzy Set Theory Interval Anal 135:1; April 1, 2003
76. Lodwick WA, Jamison KD (2003) Estimating and Validating the Cumulative Distribution of a Function of Random Variables: Toward the Development of Distribution Arithmetic. Reliab Comput 9:127–141
77. Lodwick WA, Jamison KD (2005) Theory and semantics for fuzzy and possibilistic optimization, Proceedings of the 11th International Fuzzy System Association World Congress. IFSA 2005, Beijing, July 2005
78. Lodwick WA, Jamison KD (2006) Interval-valued probability in the analysis of problems that contain a mixture of fuzzy, possibilistic and interval uncertainty. In: Demirli K, Akgunduz A (eds) 2006 Conference of the North American Fuzzy Information Processing Society, 3–6 June 2006, Montréal, Canada, paper 327137
79. Lodwick WA, Jamison KD (2008) Interval-Valued Probability in the Analysis of Problems Containing a Mixture of Fuzzy, Possibilistic, Probabilistic and Interval Uncertainty. Fuzzy Sets and Systems 2008
80. Lodwick WA, Jamison KD (2007) The use of interval-valued probability measures in optimization under uncertainty for problems containing a mixture of fuzzy, possibilistic, and interval uncertainty. In: Melin P, Castillo O, Aguilar LT, Kacprzyk J, Pedrycz W (eds) Foundations of Fuzzy Logic and Soft Computing: 12 th International Fuzzy System Association World Congress, IFSA 2007, Cancun, Mexico, June 2007, Proceedings. Springer, Berlin, pp 361–370
81. Lodwick WA, Jamison KD (2007) Theoretical and semantic distinctions of fuzzy, possibilistic, and mixed fuzzy/possibilistic optimization. Fuzzy Sets Syst 158(17):1861–1872
82. Lodwick WA, McCourt S, Newman F, Humpheries S (1999) Optimization Methods for Radiation Therapy Plans. In: Borgers C, Natterer F (eds) IMA Series in Applied Mathematics – Computational Radiology and Imaging: Therapy and Diagnosis. Springer, New York, pp 229–250
83. Lodwick WA, Neumaier A, Newman F (2001) Optimization under uncertainty: methods and applications in radiation therapy. Proc 10th IEEE Int Conf Fuzzy Syst 2001 3:1219–1222
84. Luhandjula MK (1986) On possibilistic linear programming. Fuzzy Sets Syst 18:15–30
85. Luhandjula MK (1989) Fuzzy optimization: an appraisal. Fuzzy Sets Syst 30:257–282
86. Luhandjula MK (2004) Optimisation under hybrid uncertainty. Fuzzy Sets Syst 146:187–203
87. Luhandjula MK (2006) Fuzzy stochastic linear programming: Survey and future research directions. Eur J Oper Res 174:1353–1367
88. Luhandjula MK, Ichihashi H, Inuiguchi M (1992) Fuzzy and semi-infinite mathematical programming. Inf Sci 61:233–250
89. Markowitz H (1952) Portfolio selection. J Finance 7:77–91
90. Moore RE (1979) Methods and Applications of Interval Analysis. SIAM, Philadelphia
91. Negoita CV (1981) The current interest in fuzzy optimization. Fuzzy Sets Syst 6:261–269
92. Negoita CV, Ralescu DA (1975) Applications of Fuzzy Sets to Systems Analysis. Birkhäuser, Boston
93. Negoita CV, Sularia M (1976) On fuzzy mathematical pro-

gramming and tolerances in planning. Econ Comput Cybern Stud Res 3(31):3–14
94. Neumaier A (2003) Fuzzy modeling in terms of surprise. Fuzzy Sets Syst 135(1):21–38
95. Neumaier A (2004) Clouds, fuzzy sets and probability intervals. Reliab Comput 10:249–272. Springer
96. Neumaier A (2005) Structure of clouds. (submitted – downloadable http://www.mat.univie.ac.at/~neum/papers.html)
97. Ogryczak W, Ruszczynski A (1999) From stochastic dominance to mean-risk models: Semideviations as risk measures. Eur J Oper Res 116:33–50
98. Nguyen HT (1978) A note on the extension principle for fuzzy sets. J Math Anal Appl 64:369–380
99. Ralescu D (1977) Inexact solutions for large-scale control problems. In: Proceedings of the 1st International Congress on Mathematics at the Service of Man, Barcelona
100. Ramik J (1986) Extension principle in fuzzy optimization. Fuzzy Sets Syst 19:29–35
101. Ramik J, Rimanek J (1985) Inequality relation between fuzzy numbers and its use in fuzzy optimization. Fuzzy Sets Syst 16:123–138
102. Ramik J, Vlach M (2002) Fuzzy mathematical programming: A unified approach based on fuzzy relations. Fuzzy Optim Decis Making 1:335–346
103. Ramik J, Vlach M (2002) Generalized Concavity in Fuzzy Optimization and Decision Analysis. Kluwer, Boston
104. Riverol C, Pilipovik MV (2007) Optimization of the pyrolysis of ethane using fuzzy programming. Chem Eng J 133:133–137
105. Riverol C, Pilipovik MV, Carosi C (2007) Assessing the water requirements in refineries using possibilistic programming. Chem Eng Process 45:533–537
106. Rommelfanger HJ (1994) Some problems of fuzzy optimization with T-norm based extended addition. In: Delgado M, Kacprzyk J, Verdegay J-L, Vila MA (eds) Fuzzy Optimization: Recent Advances. Physica, Heidelberg, pp 158–168
107. Rommelfanger HJ (1996) Fuzzy linear programming and applications. Eur J Oper Res 92:512–527
108. Rommelfanger HJ (2004) The advantages of fuzzy optimization models in practical use. Fuzzy Optim Decis Making 3:293–309
109. Rommelfanger HJ, Slowinski R (1998) Fuzzy linear programming with single or multiple objective functions. In: Slowinski R (ed) Fuzzy Sets in Decision Analysis, Operations Research and Statistics. The Handbooks of Fuzzy Sets. Kluwer, Netherlands, pp 179–213
110. Roubens M (1990) Inequality constraints between fuzzy numbers and their use in mathematical programming. In: Slowinski R, Teghem J (eds) Stochastic versus Fuzzy Approaches to Multiobjective Mathematical Programming Under Uncertainty. Kluwer, Netherlands, pp 321–330
111. Russell B (1924) Vagueness. Aust J Philos 1:84–92
112. Sahinidis N (2004) Optimization under uncertainty: State-of-the-art and opportunities. Comput Chem Eng 28:971–983
113. Saito S, Ishii H (1998) Existence criteria for fuzzy optimization problems. In: Takahashi W, Tanaka T (eds) Proceedings of the International Conference on Nonlinear Analysis and Convex Analysis, Niigata, Japan, 28–31 July 1998. World Scientific Press, Singapore, pp 321–325
114. Shafer G (1976) Mathematical A Theory of Evidence. Princeton University Press, Princeton
115. Shafer G (1987) Belief functions and possibility measures. In: James Bezdek C (ed) Analysis of Fuzzy Information. Mathematics and Logic, vol 1. CRC Press, Boca Raton, pp 51–84
116. Sousa JM, Kaymak U (2002) Fuzzy Decision Making in Modeling and Control. World Scientific Press, Singapore
117. Steinbach M (2001) Markowitz revisited: Mean-variance models in financial portfolio analysis. SIAM Rev 43(1):31–85
118. Tanaka H, Asai K (1984) Fuzzy linear programming problems with fuzzy numbers. Fuzzy Sets Syst 13:1–10
119. Tanaka H, Okuda T, Asai K (1973) Fuzzy mathematical programming. Trans Soc Instrum Control Eng 9(5):607–613; (in Japanese)
120. Tanaka H, Okuda T, Asai K (1974) On fuzzy mathematical programming. J Cybern 3(4):37–46
121. Tanaka H, Ichihasi H, Asai K (1984) A formulation of fuzzy linear programming problems based on comparison of fuzzy numbers. Control Cybern 13(3):185–194
122. Tanaka H, Ichihasi H, Asai K (1985) Fuzzy decisions in linear programming with trapezoidal fuzzy parameters. In: Kacprzyk J, Yager R (eds) Management Decision Support Systems Using Fuzzy Sets and Possibility Theory. Springer, Heidelberg, pp 146–159
123. Tang J, Wang D, Fung R (2001) Formulation of general possibilistic linear programming problems for complex industrial systems. Fuzzy Sets Syst 119:41–48
124. Thipwiwatpotjani P (2007) An algorithm for solving optimization problems with interval-valued probability measures. CCM Report, No. 259 (December 2007). CCM, Denver
125. Untiedt E (2006) Fuzzy and possibilistic programming techniques in the radiation therapy problem: An implementation-based analysis. Masters Thesis, University of Colorado Denver, Department of Mathematical Sciences (July 5, 2006)
126. Untiedt E (2007) A robust model for mixed fuzzy and possibilistic programming. Project report for Math 7593: Advanced Linear Programming. Spring, Denver
127. Untiedt E (2007) Using gradual numbers to analyze non-monotonic functions of fuzzy intervals. CCM Report No 258 (December 2007). CCM, Denver
128. Untiedt E, Lodwick WA (2007) On selecting an algorithm for fuzzy optimization. In: Melin P, Castillo O, Aguilar LT, Kacprzyk J, Pedrycz W (eds) Foundations of Fuzzy Logic and Soft Computing: 12th International Fuzzy System Association World Congress, IFSA 2007, Cancun, Mexico, June 2007, Proceedings. Springer, Berlin, pp 371–380
129. Vasant PM, Barsoum NN, Bhattacharya A (2007) Possibilistic optimization in planning decision of construction industry. Int J Prod Econ (to appear)
130. Verdegay JL (1982) Fuzzy mathematical programming. In: Gupta MM, Sanchez E (eds) Fuzzy Information and Decision Processes. North-Holland, Amsterdam, pp 231–237
131. Vila MA, Delgado M, Verdegay JL (1989) A general model for fuzzy linear programming. Fuzzy Sets Syst 30:21–29
132. Wang R, Liang T (2005) Applying possibilistic linear programming to aggregate production planning. Int J Prod Econ 98:328–341
133. Wang S, Zhu S (2002) On fuzzy portfolio selection problems. Fuzzy Optim Decis Making 1:361–377
134. Wang Z, Klir GJ (1992) Fuzzy Measure Theory. Plenum Press, New York
135. Weichselberger K (2000) The theory of interval-probability as a unifying concept for uncertainty. Int J Approx Reason 24:149–170

136. Werners B (1988) Aggregation models in mathematical programming. In: Mitra G (ed) Mathematical Models for Decision Support, NATO ASI Series vol F48. Springer, Berlin
137. Werners B (1995) Fuzzy linear programming – Algorithms and applications. Addendum to the Proceedings of ISUMA-NAPIPS'95, College Park, Maryland, 17–20 Sept 1995, pp A7-A12
138. Whitmore GA, Findlay MC (1978) Stochastic Dominance: An Approach to DecisionDecision-Making Under Risk. Lexington Books, Lexington
139. Yager RR (1980) On choosing between fuzzy subsets. Kybernetes 9:151–154
140. Yager RR (1981) A procedure for ordering fuzzy subsets of the unit interval. Inf Sci 24:143–161
141. Yager RR (1986) A characterization of the extension principle. Fuzzy Sets Syst 18:205–217
142. Zadeh LA (1965) Fuzzy Sets. Inf Control 8:338–353
143. Zadeh LA (1968) Probability measures of fuzzy events. J Math Anal Appl 23:421–427
144. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning. Inf Sci, Part I: 8:199–249; Part II: 8:301–357; Part III: 9:43–80
145. Zadeh LA (1978) Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets Syst 1:3–28
146. Zeleny M (1994) Fuzziness, knowledge and optimization: New optimality concepts. In: Delgado M, Kacprzyk J, Verdegay J-L, M Vila A (eds) Fuzzy Optimization: Recent Advances. Physica, Heidelberg, pp 3–20
147. Zimmermann HJ (1974) Optimization in fuzzy environment. Paper presented at the International XXI TIMS and 46th ORSA Conference, Puerto Rico Meeting, San Juan, Porto Rico, Oct 1974
148. Zimmermann HJ (1976) Description and optimization of fuzzy systems. Int J General Syst 2(4):209–216
149. Zimmermann HJ (1978) Fuzzy programming and linear programming with several objective functions. Fuzzy Sets Syst 1:45–55
150. Zimmermann HJ (1983) Fuzzy mathematical programming. Comput Oper Res 10:291–298

Fuzzy Probability Theory

MICHAEL BEER
National University of Singapore, Kent Ridge, Singapore

Article Outline

Glossary
Definition of the Subject
Introduction
Mathematical Environment
Fuzzy Random Quantities
Fuzzy Probability
Representation of Fuzzy Random Quantities
Future Directions
Bibliography

Glossary

Fuzzy set and fuzzy vector Let $\underline{\mathbf{X}}$ represent a universal set and $\underline{x}$ be the elements of $\underline{\mathbf{X}}$, then

$$\underline{\tilde{A}} = \{(\underline{x}, \mu_A(\underline{x})) \mid \underline{x} \in \underline{\mathbf{X}}\},\ \mu_A(\underline{x}) \geq 0 \quad \forall \underline{x} \in \underline{\mathbf{X}} \quad (1)$$

is referred to as fuzzy set $\underline{\tilde{A}}$ on $\underline{\mathbf{X}}$. $\mu_A(\underline{x})$ is the membership function (characteristic function) of the fuzzy set $\underline{\tilde{A}}$ and represents the degree with which the elements $\underline{x}$ belong to $\underline{\tilde{A}}$. If

$$\sup_{\underline{x} \in \underline{\mathbf{X}}} [\mu_A(\underline{x})] = 1\,, \quad (2)$$

the membership function and the fuzzy set $\underline{\tilde{A}}$ are called normalized; see Fig. 1. In case of a limitation to the Euclidean space $\underline{\mathbf{X}} = \mathbb{R}^n$ and normalized fuzzy sets, the fuzzy set $\underline{\tilde{A}}$ is also referred to as fuzzy vector denoted by $\underline{\tilde{x}}$ with its membership function $\mu(\underline{x})$, or, in the one-dimensional case, as fuzzy variable $\tilde{x}$ with $\mu(x)$.

α-Level set and support The crisp sets

$$\underline{A}_{\alpha_k} = \{\underline{x} \in \underline{\mathbf{X}} \mid \mu_A(\underline{x}) \geq \alpha_k\} \quad (3)$$

extracted from the fuzzy set $\underline{\tilde{A}}$ for real numbers $\alpha_k \in (0, 1]$ are called α-level sets. These comply with the inclusion property

$$\underline{A}_{\alpha_k} \subseteq \underline{A}_{\alpha_i} \quad \forall \alpha_i, \alpha_k \in (0, 1] \text{ with } \alpha_i \leq \alpha_k\,. \quad (4)$$

The largest α-level set $\underline{A}_{\alpha_k \to +0}$ is called support $S(\underline{\tilde{A}})$; see Fig. 1.

σ-Algebra A family $\mathfrak{M}(\underline{\mathbf{X}})$ of sets $\underline{A}_i$ on the universal set $\underline{\mathbf{X}}$ is referred to as σ-algebra $\mathfrak{S}(\underline{\mathbf{X}})$ on $\underline{\mathbf{X}}$, if

$$\underline{\mathbf{X}} \in \mathfrak{S}(\underline{\mathbf{X}})\,, \quad (5)$$

$$\underline{A}_i \in \mathfrak{S}(\underline{\mathbf{X}}) \Rightarrow \underline{A}_i^C \in \mathfrak{S}(\underline{\mathbf{X}})\,, \quad (6)$$

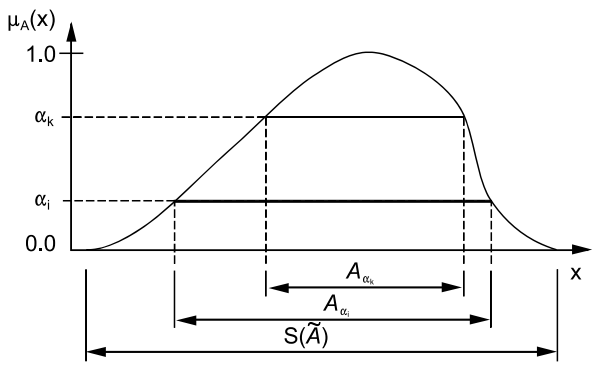

Fuzzy Probability Theory, Figure 1
Normalized fuzzy set with α-level sets and support

and if for every sequence of sets $\underline{A}_i$

$$\underline{A}_i \in \mathfrak{S}(\underline{\mathbf{X}});\ i = 1, 2, \ldots \Rightarrow \bigcup_i^{\infty} \underline{A}_i \in \mathfrak{S}(\underline{\mathbf{X}})\,. \qquad (7)$$

In this definition, $\underline{A}_i^C$ is the complementary set of $\underline{A}_i$ with respect to $\underline{\mathbf{X}}$, a family $\mathfrak{M}(\underline{\mathbf{X}})$ of sets $\underline{A}_i$ refers to subsets and systems of subsets of the power set $\mathfrak{P}(\underline{\mathbf{X}})$ on $\underline{\mathbf{X}}$, and the power set $\mathfrak{P}(\underline{\mathbf{X}})$ is the set of all subsets $\underline{A}_i$ of $\underline{\mathbf{X}}$.

Definition of the Subject

Fuzzy probability theory is an extension of probability theory to dealing with mixed probabilistic/non-probabilistic uncertainty. It provides a theoretical basis to model uncertainty which is only partly characterized by randomness and defies a pure probabilistic modeling with certainty due to a lack of trustworthiness or precision of the data or a lack of pertinent information. The fuzzy probabilistic model is settled between the probabilistic model and non-probabilistic uncertainty models. The significance of fuzzy probability theory lies in the treatment of the elements of a population not as crisp quantities but as set-valued quantities or granules in an uncertain fashion, which largely complies with reality in most everyday situations. Probabilistic and non-probabilistic uncertainty can so be transferred adequately and separately to the results of a subsequent analysis. This enables best case and worst case estimates in terms of probability taking account of variations within the inherent non-probabilistic uncertainty. The development of fuzzy probability theory was initiated by H. Kwakernaak with the introduction of fuzzy random variables in [47] in 1978. Subsequent developments have been reported in different directions and from different perspectives including differences in terminology. The usefulness of the theory has been underlined with various applications beyond mathematics and information science, in particular, in engineering. The application fields are not limited and may be extended increasingly, for example, to medicine, biology, psychology, economy, financial sciences, social sciences, and even to law.

Introduction

The probably most popular example of fuzzy probability is the evaluation of a survey on the subjective assessment of temperature. A group of test persons are asked – under equal conditions – to give a statement on the current temperature as realistically as possible. If the scatter of the statements is considered as random, the mean value of the statements provides a reasonable statistical point estimate for the actual temperature. The statements are, however, naturally given in an uncertain form. The test persons enunciate their perception in a form such as *about 25°C*, *possibly 27°C*, *between 24°C and 26°C*, or they even only come up with linguistic assessments such as *warm*, *very warm*, or *pleasant*. This uncertainty is non-probabilistic but has to be taken into account in the statistical evaluation. It is transferred to the estimated mean value, which is no longer obtained as a crisp number but as a value range or a set of values corresponding to the possibilities within the range of uncertainty of the statements. If the uncertain statements are initially quantified as fuzzy values, the mean value is obtained as a fuzzy value, too; and the probability of certain events is also computed as a fuzzy quantity – referred to as fuzzy probability. This example is a typical materialization of the following general real-world situation.

The numerical representation of a physical quantity with the aid of crisp numbers $x \in \mathbb{R}$ or sets thereof is frequently interfered by uncertainty regarding the trustworthiness of measured, or otherwise specified, values. The perceptions of physical quantities may appear, for example, as imprecise, diffuse, vague, dubious, or ambiguous. Underlying reasons for this phenomenon include the limited precision of any measurement (digital or analog), indirect measurements via auxiliary quantities in conjunction with a – more or less trustworthy – model to eventually determine the value wanted, measurements under weakly specified or arbitrarily changing boundary conditions, and the specification of values by experts in a linguistic manner. The type of the associated uncertainty is non-frequentative and improper for a subjective probabilistic modeling; hence, it is non-probabilistic. This uncertainty is unavoidable and may always be made evident by a respective choice of scale.

If a set of uncertain perceptions of a physical quantity is present in the form of a random sample, then the overall uncertainty possesses a mixed probabilistic/non-probabilistic character. Whilst the scatter of the realizations of the physical quantity possesses a probabilistic character (frequentative or subjective), each particular realization from the population may, additionally, exhibit non-probabilistic uncertainty. Consequently, a realistic modeling in those cases must involve both probabilistic and non-probabilistic uncertainty. This modeling without distorting or ignoring information is the mission of fuzzy probability theory. A pure probabilistic modeling would introduce unwarranted information in the form of a distribution function that cannot be justified and

would thus diminish the trustworthiness of the probabilistic results.

Mathematical Environment

Fuzzy probability is part of the framework of generalized information theory [38] and is covered by the umbrella of granular computing [50,62]. It represents a special case of imprecise probabilities [15,78] with ties to concepts of random sets [52]. This is underlined by Walley's summary of the semantics of imprecise probabilities with the term *indeterminacy*, which arises from ignorance about facts, events, or dependencies. Within the class of mathematical models covered by the term *imprecise probabilities*, see [15, 38,78], fuzzy probability theory has a relationship to concepts known as upper and lower probabilities [28], sets of probability measures [24], distribution envelops [7], interval probabilities [81], and p-box approach [23]. Also, similarities exist with respect to evidence theory (or Dempster–Shafer theory) [20,70] as a theory of infinitely monotone Choquet capacities [39,61]. The relationship to the latter is associated with the interpretation of the measures plausibility and belief, with the special cases of possibility and necessity, as upper and lower probabilities, respectively, [8].

Fuzzy probability shares the common feature of all imprecise probability models: the uncertainty of an event is characterized with a set of possible measure values in terms of probability, or with bounds on probability. Its distinctive feature is that set-valued information, and hence the probability of associated events, is described with the aid of uncertain sets according to fuzzy set theory [83,86]. This represents a marriage between fuzzy methods and probabilistics with fuzziness and randomness as special cases, which justifies the denotation as *fuzzy randomness*. Fuzzy probability theory enables a consideration of a fuzzy set of possible probabilistic models over the range of imprecision of the knowledge about the underlying randomness. The associated fuzzy probabilities provide weighted bounds on probability – the weights of which are obtained as the membership values of the fuzzy sets. Based on α-discretization [86] and the representation of fuzzy sets as sets of α-level sets, the relationship of fuzzy probability theory to various concepts of imprecise probabilities becomes obvious. For each α-level a common interval probability, crisp bounds on probability, or a classical set of probability measures, respectively, are obtained. The α-level sets of fuzzy events can be treated as random sets. Further, a relationship of these random sets to evidence theory can be constructed if a respective basic probability assignment is selected; see [21]. Consistency with evidence theory is obtained if the focal sets are specified as fuzzy elementary events and if the basic probability assignment follows a discrete uniform distribution over the fuzzy elementary events.

The model fuzzy randomness with its two components – fuzzy methods and probabilistics – can utilize a distinction between aleatory and epistemic uncertainty with respect to the sources of uncertainty [29]. This is particularly important in view of practical applications. Irreducible uncertainty as a property of the system associated with fluctuations/variability may be summarized as aleatory uncertainty and described probabilistically, and reducible uncertainty as a property of the analysts, or its perception, associated with a lack of knowledge/precision may be understood as epistemic uncertainty and described with fuzzy sets. The model fuzzy randomness then combines, without mixing, both components in the form of a fuzzy set of possible probabilistic models over some particular range of imprecision. This distinction is retained throughout any subsequent analysis and reflected in the results.

The development of fuzzy probability was initiated with the introduction of fuzzy random variables by Kwakernaak [47,48] in 1978/79. Subsequent milestones were set by Puri and Ralescu [63], Kruse and Meyer [46], Wang and Zhang [80], and Krätschmer [43]. The developments show differences in terminology, concepts, and in the associated consideration of measurability; and the investigations are ongoing [11,12,14,17,35,36,37,45,73]. Krätschmer [43] showed that the different concepts can be unified to a certain extent. Generally, it can be noted that α-discretization is utilized as a helpful instrument. An overview with specific comments on the different developments is provided in [43] and [57]. Investigations were pursued on independent and dependent fuzzy random variables for which parameters were defined with particular focus on variance and covariance [22,32,40,41,60]. Fuzzy random processes were examined to reveal properties of limit theorems and martingales associated with fuzzy randomness [44,65]; see also [71,79] and for a survey [49]. Particular interest was devoted to the strong law of large numbers [10,33,34]. Further, the differentiation and the integration of fuzzy random quantities was investigated in [51,64].

Driven by the motivation for the establishment of fuzzy probability theory considerable effort was made in the modeling and statistical evaluation of imprecise data. Fundamental achievements were reported by Kruse and Meyer [46], Bandemer and Näther [2,5], and by Viertl [75]. Classical statistical methods were extended in order to take account of statistical fluctuations/variability and imprecision simultaneously, and the specific fea-

tures associated with the imprecision of the data were investigated. Ongoing research is reported, for example, in [58,74] in view of evaluating measurements, in [51,66] for decision making, and in [16,42,59] for regression analysis. Methods for evaluating imprecise data with the aid of generalized histograms are discussed in [9,77]. Also, the application of resampling methods is pursued; bootstrap concepts are utilized for statistical estimations [31] and hypothesis testing [26] based on imprecise data. Another method for hypothesis testing is proposed in [27], which employs fuzzy parameters in order to describe a fuzzy transition between rejection and acceptance. Bayesian methods have also been extended by the inclusion of fuzzy variables to take account of imprecise data; see [75] for basic considerations. A contribution to Bayesian statistics with imprecise prior distributions is presented in [76]. This leads to imprecise posterior distributions, imprecise predictive distributions, and may be used to deduce imprecise confidence intervals. A combination of the Bayesian theorem with kriging based on imprecise data is described in [3]. A Bayesian test of fuzzy hypotheses is discussed in [72], while in [67] the application of a fuzzy Bayesian method for decision making is presented.

In view of practical applications, probability distribution functions are defined for fuzzy random quantities [54,75,77,85] – despite some drawback [6]. These distribution functions can easily be formulated and used for further calculations, but they do not uniquely describe a fuzzy random quantity. This theoretical lack is, however, generally without an effect in practical applications so that stochastic simulations may be performed according to the distribution functions. Alternative simulation methods were proposed based on parametric [13] and non-parametric [6,55] descriptions of imprecision. The approach according to [55] enables a direct generation of fuzzy realizations based on a new concept for an incremental representation of fuzzy random quantities. This method is designed to simulate and predict fuzzy time series; it circumvents the problems of artificial uncertainty growth or bias of non-probabilistic uncertainty, which is frequently concerned with numerical simulations.

This variety of theoretical developments provides reasonable margins for the formulation of fuzzy probability theory but does not allow the definition of a unique concept. Choices have to be made within the elaborated margins depending on the envisaged application and environment. For the subsequent sections these choices are made in view of a broad spectrum of possible applications, for example, in civil/mechanical engineering [54]. These choices concern the following three main issues.

First, measurability has to be ensured according to a sound concept. According to [43], the concepts of measurable bounding functions [47,48], of measurable mappings of α-level sets [63], and of measurable fuzzy valued mappings [17,36] are available; or the unifying concept proposed in [43] itself, which utilizes a special topology on the space of fuzzy realizations, may be selected. From a practical point of view the concept of measurable bounding functions is most reasonable due to its analogy to traditional probability theory. On this basis, a fuzzy random quantity can be regarded as a fuzzy set of traditional, crisp, random quantities, each one carrying a certain membership degree. Each of these crisp random quantities is then measurable in the traditional fashion, and their membership degrees can be transferred to the respective measure values. The set of the obtained measure values including their membership degrees then represents a fuzzy probability.

Second, a concept for the integration of a fuzzy-valued function has to be selected from the different available approaches [86]. This is associated with the computation of the probability of a fuzzy event. An evaluation in a mean sense weighted by the membership function of the fuzzy event is suggested in [84], which leads to a scalar value for the probability. This evaluation is associated with the interpretation that an event may occur partially. An approach for calculating the probability of a fuzzy event as a fuzzy set is proposed in [82]; the resulting fuzzy probability then represents a set of measure values with associated membership degrees. This complies with the interpretation that the occurrence of an event is binary but it is not clearly indicated if the event has occurred or not. The imprecision is associated with the observation rather than with the event. The latter approach corresponds with the practical situation in many cases, provides useful information in form of the imprecision reflected in the measure values, and follows the selected concept of measurability. It is thus taken as a basis for further consideration.

Third, the meaning of the distance between fuzzy sets as realizations of a fuzzy random quantity must be defined, which is of particular importance for the definition of the variance and of further parameters of fuzzy random quantities. An approach that follows a set-theoretical point of view and leads to a crisp distance measure is presented in [40]. It is proposed to apply the Hausdorff metric to the α-level sets of a fuzzy quantity and to average the results over the membership scale. Consequently, parameters of a fuzzy random quantity which are associated with a distance between fuzzy realizations reflect the variability within the imprecision merely in an integrated form. For example, the variance of a fuzzy random variable is

obtained as a crisp value. In contrast to this, the application of standard algorithms for operations on fuzzy sets, such as the extension principle, [4,39,86] leads to fuzzy distances between fuzzy sets. Parameters, including variances, of fuzzy random quantities are then obtained as fuzzy sets of possible parameter values. This corresponds to the interpretation of fuzzy random quantities as fuzzy sets of traditional, crisp, random quantities. The latter approach is thus pursued further.

These three selections comply basically with the definitions in [46] and [47]; see also [57]. Among all the possible choices, this set-up ensures the most plausible settlement of fuzzy probability theory within the framework of imprecise probabilities [15,38,78] with ties to evidence theory and random set approaches [21,30]. Fuzzy probability is obtained in the form of weighted plausible bounds on probability. Moreover, in view of practical applications, the treatment of fuzzy random quantities as fuzzy sets of traditional random quantities enables the utilization of established probabilistic methods as kernel solutions in the environment of a fuzzy analysis. For example, sophisticated methods of Monte Carlo simulation [25,68,69] may be combined with a generally applicable fuzzy analysis based on an global optimization approach using α-discretization [53]. If some restricting conditions are complied with, numerically efficient methods from interval mathematics [1] may be employed for the α-level mappings instead of a global optimization approach; see [56]. The selected concept, eventually, enables best-case and worst-case studies within the range of possible probabilistic models.

Fuzzy Random Quantities

With the above selections, the definitions from traditional probability theory can be adopted and extended to dealing with imprecise outcomes from a random experiment. Let Ω be the space of random elementary events $\omega \in \Omega$ and the universe on which the realizations are observed be the n-dimensional Euclidean space $\underline{\mathbf{X}} = \mathbb{R}^n$. Then, a membership scale μ is introduced perpendicular to the hyperplane $\Omega \times \underline{\mathbf{X}}$. This enables the specification of fuzzy sets on $\underline{\mathbf{X}}$ for given elementary events ω from Ω without interaction between Ω and μ. That is, randomness induced by Ω and fuzziness described by μ – only in x-direction – are not mixed with one another. Let $\mathfrak{F}(\underline{\mathbf{X}})$ be the set of all fuzzy quantities on $\underline{\mathbf{X}} = \mathbb{R}^n$; that is, $\mathfrak{F}(\underline{\mathbf{X}})$ denotes the collection of all fuzzy sets $\underline{\tilde{A}}$ on $\underline{\mathbf{X}} = \mathbb{R}^n$, with $\underline{\tilde{A}}$ according to Eq. (1). Then, the imprecise result of the mapping

$$\underline{\tilde{X}}: \Omega \to \mathfrak{F}(\underline{\mathbf{X}}) \tag{8}$$

is referred to as fuzzy random quantity $\underline{\tilde{X}}$. In contrast to real-valued random quantities, a fuzzy realization $\underline{\tilde{x}}(\omega) \in \mathfrak{F}(\underline{\mathbf{X}})$, or $\underline{\tilde{x}}(\omega) \subseteq \underline{\mathbf{X}}$, is now assigned to each elementary event $\omega \in \Omega$; see Fig. 2. These fuzzy realizations may be understood as a numerical representation of granules. Generally, a fuzzy random quantity can be discrete

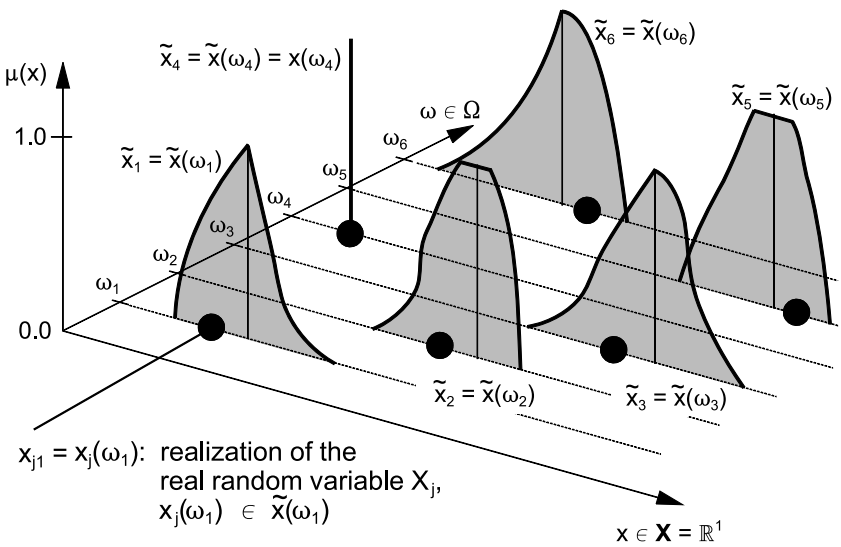

Fuzzy Probability Theory, Figure 2
Fuzzy random variable

or continuous with respect to both fuzziness and randomness. The further consideration refers to the continuous case, from which the discrete case may be derived.

Without a limitation in generality, the fuzzy realizations may be restricted to normalized fuzzy quantities, thus representing fuzzy vectors. Further restrictions can be defined in view of a convenient numerical treatment if the application field allows for. This concerns, for example, a restriction to connected and compact α-level sets $\underline{A}_\alpha = \underline{x}_\alpha$ of the fuzzy realizations, a restriction to convex fuzzy sets as fuzzy realizations (a fuzzy set $\tilde{\underline{A}}$ is convex if all its α-level sets $\underline{A}_\alpha$ are convex sets), or a restriction to fuzzy realizations with only one element $\underline{x}_i$ carrying the membership $\mu(\underline{x}_i) = 1$ as in [47].

For the treatment of the fuzzy realizations, a respective algorithm for operations on fuzzy sets has to be selected. Following the literature and the above selection, the standard basis is employed with the min-operator as a special case of a t-norm and the associated max-operator as a special case of a t-co-norm [18,19,86]. This leads to the min-max operator and the extension principle [4,39,86].

With the interpretation of a fuzzy random quantity as a fuzzy set of real-valued random quantities, according to the above selection, the following relationship to traditional probability theory is obtained. Let $\underline{x}_{ji}$ be a realization of a real-valued random quantity $\underline{\mathrm{X}}_j$ and $\tilde{\underline{x}}_i$ be a fuzzy realization of a fuzzy random quantity $\tilde{\underline{\mathrm{X}}}$ with $\underline{x}_{ji}$ and $\tilde{\underline{x}}_i$ be assigned to the same elementary event ω_i. If $\underline{x}_{ji} \in \tilde{\underline{x}}_i$, then $\underline{x}_{ji}$ is called contained in $\tilde{\underline{x}}_i$. If, for all elementary events $\omega_i \in \Omega,\ i = 1, 2, \ldots$, the $\underline{x}_{ji}$ are contained in the $\tilde{\underline{x}}_i$, the set of the $\underline{x}_{ji},\ i = 1, 2, \ldots$, then constitutes an original $\underline{\mathrm{X}}_j$ of the fuzzy random quantity $\tilde{\underline{\mathrm{X}}}$; see Fig. 2. The original $\underline{\mathrm{X}}_j$ is referred to as completely contained in $\tilde{\underline{\mathrm{X}}}$, $\underline{\mathrm{X}}_j \in \tilde{\underline{\mathrm{X}}}$. Each real-valued random quantity $\underline{\mathrm{X}}$ that is completely contained in $\tilde{\underline{\mathrm{X}}}$ is an original $\underline{\mathrm{X}}_j$ of $\tilde{\underline{\mathrm{X}}}$ and carries the membership degree

$$\mu(\underline{\mathrm{X}}_j) = \max[\alpha | \underline{x}_{ji} \in \underline{x}_{i\alpha} \forall i.] \,. \tag{9}$$

That is, in the Ω-direction, each original $\underline{\mathrm{X}}_j$ must be consistent with the fuzziness of $\tilde{\underline{\mathrm{X}}}$. Consequently, the fuzzy random quantity $\tilde{\underline{\mathrm{X}}}$ can be represented as the fuzzy set of all originals $\underline{\mathrm{X}}_j$ contained in $\tilde{\underline{\mathrm{X}}}$,

$$\tilde{\underline{\mathrm{X}}} = \left\{(\underline{\mathrm{X}}_j, \mu(\underline{\mathrm{X}}_j)) | \underline{x}_{ji} \in \tilde{\underline{x}}_i \forall i.\right\} \,. \tag{10}$$

Each fuzzy random quantity $\tilde{\underline{\mathrm{X}}}$ contains at least one real-valued random quantity $\underline{\mathrm{X}}$ as an original $\underline{\mathrm{X}}_j$ of $\tilde{\underline{\mathrm{X}}}$. Each fuzzy random quantity $\tilde{\underline{\mathrm{X}}}$ that possesses precisely one original is thus a real-valued random quantity $\underline{\mathrm{X}}$. That is, real-valued random quantities are a special case of fuzzy random quantities. This enables a simultaneous treatment of real-valued random quantities and fuzzy random quantities within the same theoretical environment and with the same numerical algorithms. Or, vice versa, it enables the utilization of theoretical results and established numerical algorithms from traditional probability theory within the framework of fuzzy probability theory.

If α-discretization is applied to the fuzzy random quantity $\tilde{\underline{\mathrm{X}}}$, random α-level sets $\underline{\mathrm{X}}_\alpha$ are obtained,

$$\underline{\mathrm{X}}_\alpha = \{\underline{\mathrm{X}} = \underline{\mathrm{X}}_j \,|\, \mu(\underline{\mathrm{X}}_j) \geq \alpha\} \,. \tag{11}$$

Their realizations are α-level sets $\underline{x}_{i\alpha}$ of the respective fuzzy realizations $\tilde{\underline{x}}_i$ of the fuzzy random quantity $\tilde{\underline{\mathrm{X}}}$. A fuzzy random quantity can thus, alternatively, be represented by the set of its α-level sets,

$$\tilde{\underline{\mathrm{X}}} = \left\{(\underline{\mathrm{X}}_\alpha, \mu(\underline{\mathrm{X}}_\alpha)) \mid \mu(\underline{\mathrm{X}}_\alpha) = \alpha \ \forall \alpha \in (0,\ 1]\right\} \,. \tag{12}$$

In the one-dimensional case and with the restriction to connected and compact α-level sets of the realizations, the random α-level sets X_α of the fuzzy random variable $\tilde{\mathrm{X}}$ become closed random intervals $[\mathrm{X}_{\alpha l}, \mathrm{X}_{\alpha r}]$.

Fuzzy Probability

Fuzzy probability is derived as a fuzzy set of probability measures for events the occurrence of which depends on the behavior of a fuzzy random quantity. These events are referred to as fuzzy random events with the following characteristics.

Let $\tilde{\underline{\mathrm{X}}}$ be a fuzzy random quantity according to Eq. (8) with the realizations $\tilde{\underline{x}}$ and $\mathfrak{S}(\underline{\mathbf{X}})$ be a σ-algebra of sets $\underline{A}_i$ defined on $\underline{\mathbf{X}}$. Then, the event

$$\tilde{\mathrm{E}}_i : \tilde{\underline{\mathrm{X}}} \textit{ hits } \underline{A}_i \tag{13}$$

is referred to as fuzzy random event, which occurs if a fuzzy realization $\tilde{\underline{x}}$ of the fuzzy random quantity $\tilde{\underline{\mathrm{X}}}$ hits the set $\underline{A}_i$. The associated probability of occurrence of $\tilde{\mathrm{E}}_i$ is referred to as fuzzy probability $\tilde{P}(\underline{A}_i)$. It is obtained as the fuzzy set of the probabilities of occurrence of the events

$$\mathrm{E}_{ij} : \underline{\mathrm{X}}_j \in \underline{A}_i \tag{14}$$

associated with all originals $\underline{\mathrm{X}}_j$ of the fuzzy random quantity $\tilde{\underline{\mathrm{X}}}$ with their membership values $\mu(\underline{\mathrm{X}}_j)$. Specifically,

$$\begin{aligned}\tilde{P}(\underline{A}_i) = \{&(P(\underline{\mathrm{X}}_j \in \underline{A}_i), \mu(P(\underline{\mathrm{X}}_j \in \underline{A}_i))) \\ &| \underline{\mathrm{X}}_j \in \tilde{\underline{\mathrm{X}}},\ \mu(P(\underline{\mathrm{X}}_j \in \underline{A}_i)) = \mu(\underline{\mathrm{X}}_j)\ \forall j\} \,.\end{aligned} \tag{15}$$

Each of the involved probabilities $P(\underline{\mathrm{X}}_j \in \underline{A}_i)$ is a traditional, real-valued probability associated with the traditional probability space $[\underline{\mathbf{X}}, \mathfrak{S}, P]$ and complying with all

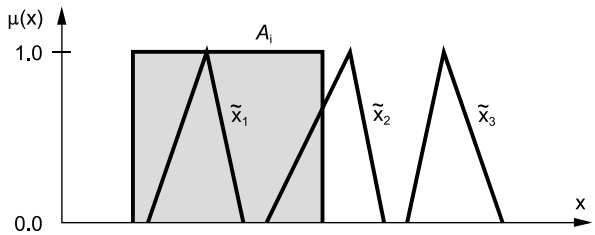

Fuzzy Probability Theory, Figure 3
Fuzzy event $\underline{\tilde{x}}_k$ *hits* $\underline{A}_i$ in the one-dimensional case

established theorems and properties of traditional probability. For a formal closure of fuzzy probability theory, the membership scale μ is incorporated in the probability space to constitute the fuzzy probability space denoted by $[\underline{\mathbf{X}}, \mathfrak{S}, P, \mu]$ or $[\underline{\mathbf{X}}, \mathfrak{S}, \tilde{P}]$.

The evaluation of the fuzzy random event Eq. (13) hinges on the question whether a fuzzy realization $\underline{\tilde{x}}_k$ hits the set $\underline{A}_i$ or not. Due to the fuzziness of the $\underline{\tilde{x}}$, these events appear as fuzzy events $\tilde{\mathrm{E}}_{ik}$: $\underline{\tilde{x}}_k$ *hits* $\underline{A}_i$ with the following three options for occurrence; see Fig. 3:

- The fuzzy realization $\underline{\tilde{x}}_k$ lies completely inside the set $\underline{A}_i$, the event $\tilde{\mathrm{E}}_{ik}$ has occurred.
- The fuzzy realization $\underline{\tilde{x}}_k$ lies only partially inside $\underline{A}_i$, the event $\tilde{\mathrm{E}}_{ik}$ may have occurred or not occurred.
- The fuzzy realization $\underline{\tilde{x}}_k$ lies completely outside the set $\underline{A}_i$, the event $\tilde{\mathrm{E}}_{ik}$ has not occurred.

The fuzzy probability $\tilde{P}(\underline{A}_i)$ takes account of all three options within the range of fuzziness. The fuzzy random quantity $\underline{\tilde{\mathrm{X}}}$ is discretized into a set of random α-level sets $\underline{\mathrm{X}}_\alpha$ according to Eq. (11), and the events $\tilde{\mathrm{E}}_i$ and $\tilde{\mathrm{E}}_{ik}$, respectively, are evaluated α-level by α-level. In this evaluation, the event $\mathrm{E}_{ik\alpha}$: $\underline{x}_{k\alpha}$ *hits* $\underline{A}_i$ admits the following two "extreme" interpretations of occurrence:

- $\mathrm{E}_{ik\alpha l}$: "$\underline{x}_{k\alpha}$ is contained in $\underline{A}_i$: $\underline{x}_{k\alpha} \subseteq \underline{A}_i$", and
- $\mathrm{E}_{ik\alpha r}$: "$\underline{x}_{k\alpha}$ and $\underline{A}_i$ possess at least one element in common: $\underline{x}_{k\alpha} \cap \underline{A}_i \neq \emptyset$".

Consequently, the events $\mathrm{E}_{ik\alpha l}$ are associated with the smallest probability

$$P_{\alpha l}(\underline{A}_i) = P(\underline{\mathrm{X}}_\alpha \subseteq \underline{A}_i)\,, \tag{16}$$

and the events $\mathrm{E}_{ik\alpha r}$ correspond to the largest probability

$$P_{\alpha r}(\underline{A}_i) = P(\underline{\mathrm{X}}_\alpha \cap \underline{A}_i \neq \emptyset)\,. \tag{17}$$

The probabilities $P_{\alpha l}(\underline{A}_i)$ and $P_{\alpha r}(\underline{A}_i)$ are bounds on the probability $\tilde{P}(\underline{A}_i)$ on the respective α-level associated with the random α-level set $\underline{\mathrm{X}}_\alpha$ of the fuzzy random quantity $\underline{\tilde{\mathrm{X}}}$; see Fig. 4. As all elements of $\underline{\mathrm{X}}_\alpha$ are originals $\underline{\mathrm{X}}_j$ of $\underline{\tilde{\mathrm{X}}}$, the probability that an $\underline{\mathrm{X}}_j \in \underline{\mathrm{X}}_\alpha$ hits $\underline{A}_i$ is bounded according to

$$P_{\alpha l}(\underline{A}_i) \leq P(\underline{\mathrm{X}}_j \in \underline{A}_i) \leq P_{\alpha r}(\underline{A}_i) \qquad \forall \underline{\mathrm{X}}_j \in \underline{\mathrm{X}}_\alpha\,. \tag{18}$$

This enables a computation of the probability bounds directly from the real-valued probabilities $P(\underline{\mathrm{X}}_j \in \underline{A}_i)$ associated with the originals $\underline{\mathrm{X}}_j$,

$$P_{\alpha l}(\underline{A}_i) = \min_{\underline{\mathrm{X}}_j \in \underline{\mathrm{X}}_\alpha} P(\underline{\mathrm{X}}_j \in \underline{A}_i)\,, \tag{19}$$

$$P_{\alpha r}(\underline{A}_i) = \max_{\underline{\mathrm{X}}_j \in \underline{\mathrm{X}}_\alpha} P(\underline{\mathrm{X}}_j \in \underline{A}_i)\,. \tag{20}$$

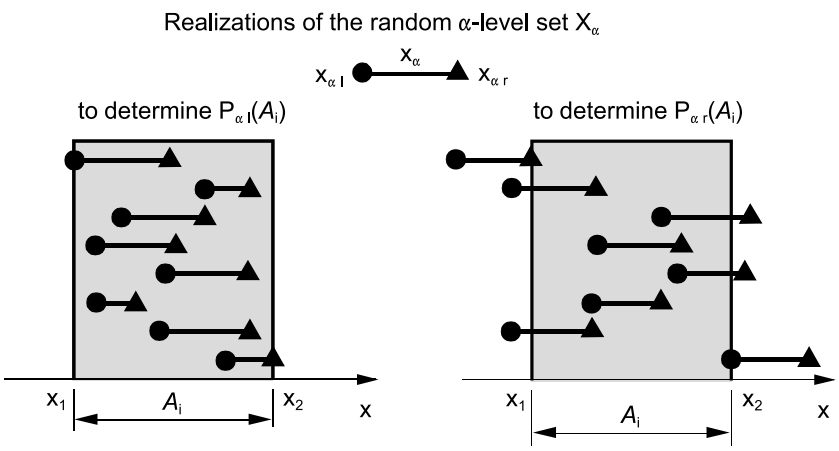

Fuzzy Probability Theory, Figure 4
Events for determining $P_{\alpha l}(\underline{A}_i)$ and $P_{\alpha r}(\underline{A}_i)$ in the one-dimensional case

If the fuzzy random quantity $\underline{\tilde{X}}$ represents a fuzzy set of continuous real-valued random quantities and if the membership functions of all fuzzy realizations $\underline{\tilde{x}}_i$ of $\underline{\tilde{X}}$ are at least segmentally continuous, then the probability bounds $P_{\alpha l}(\underline{A}_i)$ and $P_{\alpha r}(\underline{A}_i)$ determine closed connected intervals $[P_{\alpha l}(\underline{A}_i), P_{\alpha r}(\underline{A}_i)]$. In this case, the fuzzy probability $\tilde{P}(\underline{A}_i)$ is obtained as a continuous and convex fuzzy set, which may be specified uniquely with the aid of α-discretization,

$$\begin{aligned}\tilde{P}(\underline{A}_i) = \big\{&(P_\alpha(\underline{A}_i), \mu(P_\alpha(\underline{A}_i)))|P_\alpha(\underline{A}_i)\\ &= [P_{\alpha l}(\underline{A}_i), P_{\alpha r}(\underline{A}_i)]\,, \mu(P_\alpha(\underline{A}_i)) = \alpha\\ &\qquad \forall\alpha \in (0,1]\big\}\,. \end{aligned} \tag{21}$$

The properties of the fuzzy probability $\tilde{P}(\underline{A}_i)$ result from the properties of the traditional probability measure in conjunction with fuzzy set theory. For example, a complementary relationship may be derived for $\tilde{P}(\underline{A}_i)$ as follows. The equivalence

$$(\underline{X}_\alpha \subseteq \underline{A}_i) \Leftrightarrow \big(\underline{X}_\alpha \cap \underline{A}_i^C = \emptyset\big) \tag{22}$$

with $\underline{A}_i^C$ being the complementary set of $\underline{A}_i$ with respect to the universe $\underline{\mathbf{X}}$, leads to

$$P(\underline{X}_\alpha \subseteq \underline{A}_i) = P\big(\underline{X}_\alpha \cap \underline{A}_i^C = \emptyset\big) \tag{23}$$

for each α-level. If the event $\underline{X}_\alpha \cap \underline{A}_i^C = \emptyset$ is expressed in terms of its complementary event $\underline{X}_\alpha \cap \underline{A}_i^C \neq \emptyset$, Eq. (23) can be rewritten as

$$P(\underline{X}_\alpha \subseteq \underline{A}_i) = 1 - P\big(\underline{X}_\alpha \cap \underline{A}_i^C \neq \emptyset\big)\,. \tag{24}$$

This leads to the relationships

$$P_{\alpha l}(\underline{A}_i) = 1 - P_{\alpha r}\big(\underline{A}_i^C\big)\,, \tag{25}$$

$$P_{\alpha r}(\underline{A}_i) = 1 - P_{\alpha l}\big(\underline{A}_i^C\big)\,, \tag{26}$$

and

$$\tilde{P}(\underline{A}_i) = 1 - \tilde{P}\big(\underline{A}_i^C\big)\,. \tag{27}$$

In the special case that the set $\underline{A}_i$ contains only one element $\underline{A}_i = \underline{x}_i$, the fuzzy probability $\tilde{P}(\underline{A}_i)$ changes to $\tilde{P}(\underline{x}_i)$. The event $\underline{X}_\alpha \cap \underline{A}_i \neq \emptyset$ is then replaced by $\underline{x}_i \in \underline{X}_\alpha$, and $\underline{X}_\alpha \subseteq \underline{A}_i$ becomes $\underline{X}_\alpha = \underline{x}_i$. The probability $P_{\alpha l}(\underline{x}_i) = P(\underline{X}_\alpha = \underline{x}_i)$ may take values greater than zero only if a realization of $\underline{X}_\alpha$ exists that possesses exactly one element $\underline{X}_\alpha = \underline{t}$ with $\underline{t} = \underline{x}_i$ and if this element $\underline{t}$ represents a realization of a discrete original of $\underline{X}_\alpha$. Otherwise, $P_{\alpha l}(\underline{x}_i) = 0$, and the fuzziness of $\tilde{P}(\underline{x}_i)$ is exclusively specified by $P_{\alpha r}(\underline{x}_i)$.

In the one-dimensional case with

$$A_i = \{x \,|\, x \in \mathbf{X};\, x_1 \le x \le x_2\} \tag{28}$$

the fuzzy probability $\tilde{P}(\underline{A}_i)$ can be represented in a simplified manner. If the random α-level sets X_α are restricted to be closed random intervals $[X_{\alpha l}, X_{\alpha r}]$, the associated fuzzy random variable $\tilde{X}$ can be completely described by means of the bounding real-valued random quantities $X_{\alpha l}$ and $X_{\alpha r}$; see Fig. 4. $X_{\alpha l}$ and $X_{\alpha r}$ represent specific originals X_j of $\tilde{X}$. This enables the specification of the probability bounds for each α-level according to

$$\begin{aligned}P_{\alpha l}(A_i) = \max\big[0, &P(X_{\alpha r} = t_r \,|\, x_2, t_r \in \mathbf{X},\ t_r \le x_2)\\ &- P(X_{\alpha l} = t_l \,|\, x_1, t_l \in \mathbf{X},\ t_l < x_1)\big]\,, \end{aligned} \tag{29}$$

and

$$\begin{aligned}P_{\alpha r}(A_i) = &P(X_{\alpha l} = t_l \,|\, x_2, t_l \in \mathbf{X};\ t_l \le x_2)\\ &- P(X_{\alpha r} = t_r \,|\, x_1, t_r \in \mathbf{X};\ t_r < x_1)\,. \end{aligned} \tag{30}$$

From the above framework, the special case of real-valued random quantities $\underline{X}$, may be reobtained as a fuzzy random quantity $\underline{\tilde{X}}$ that contains precisely one original $\underline{X}_j = \underline{X}_1$,

$$\underline{X}_1 = (\underline{X}_{j=1}, \mu(\underline{X}_{j=1}) = 1)\,; \tag{31}$$

see Eq. (10). Then, all $\underline{X}_\alpha$ contain only the sole original $\underline{X}_1$, and both $\underline{X}_\alpha \cap \underline{A}_i \neq \emptyset$ and $\underline{X}_\alpha \subseteq \underline{A}_i$ reduce to $\underline{X}_1 \in \underline{A}_i$. That is, the event $\underline{X}_1$ *hits* $\underline{A}_i$ does no longer provide options for interpretation reflected as fuzziness,

$$P_{\alpha l}(\underline{A}_i) = P_{\alpha r}(\underline{A}_i) = P(\underline{A}_i) = P(\underline{X}_1 \in \underline{A}_i)\,. \tag{32}$$

In the above special one-dimension case, Eqs. (29) and (30), the setting $X = X_{\alpha l} = X_{\alpha r}$ leads to

$$\begin{aligned}P_{\alpha l}(A_i) &= P_{\alpha r}(A_i)\\ &= P(X_1 = t \,|\, x_1, x_2, t \in \mathbf{X};\ x_1 \le t_l \le x_2)\,, \end{aligned} \tag{33}$$

Further properties and computation rules for fuzzy probabilities may be derived from traditional probability theory in conjunction with fuzzy set theory.

Representation of Fuzzy Random Quantities

The fuzzy probability $\tilde{P}(\underline{A}_i)$ may be computed for each arbitrary set $\underline{A}_i \in \mathfrak{S}(\underline{\mathbf{X}})$. If – as a special family of sets $\mathfrak{S}(\underline{\mathbf{X}})$ – the Borel σ-algebra $\mathfrak{S}_0(\mathbb{R}^n)$ of the $\mathbb{R}^n$ is selected, the concept of the probability distribution function may

be applied to fuzzy random quantities. That is, the system $\mathfrak{S}_0(\mathbb{R}^n)$ of the open sets

$$\underline{A}_{i0} = \{\underline{t} = (t_1, \ldots, t_k, \ldots, t_n) | \underline{x} = \underline{x}_i; \underline{x}, \underline{t} \in \mathbb{R}^n; \quad t_k < x_k; k = 1, \ldots, n\} \quad (34)$$

on $\underline{\mathbf{X}} = \mathbb{R}^n$ is considered; $\mathfrak{S}_0(\mathbb{R}^n)$ is a Boolean set algebra.

The concept of fuzzy probability according to Sect. "Fuzzy Probability" applied to the sets from Eq. (34) leads to fuzzy probability distribution functions; see Fig. 5. The fuzzy probability distribution function $\tilde{F}(\underline{x})$ of the fuzzy random quantity $\underline{\tilde{X}}$ on $\underline{\mathbf{X}} = \mathbb{R}^n$ is obtained as the set of the fuzzy probabilities $\tilde{P}(\underline{A}_{i0})$ with $\underline{A}_{i0}$ according to Eq. (34) for all $\underline{x}_i \in \underline{\mathbf{X}}$,

$$\tilde{F}(\underline{x}) = \{\tilde{P}(\underline{A}_{i0}) \, \forall \underline{x}_i \in \underline{\mathbf{X}}\} \,. \quad (35)$$

It is a fuzzy function. Bounds for the functional values $\tilde{F}(\underline{x})$ are specified for each α-level depending on $\underline{x} = \underline{x}_i$ in Eq. (34) and in compliance with Eqs. (19) and (20),

$$F_{\alpha l}(\underline{x} = (x_1, \ldots, x_n)) = 1 - \max_{\underline{X}_j \in \underline{X}_\alpha} P\big(\underline{X}_j = \underline{t} = (t_1, \ldots, t_n) | \underline{x}, \underline{t} \in \underline{\mathbf{X}} = \mathbb{R}^n, \; \exists t_k \geq x_k, \; 1 \leq k \leq n\big) \,, \quad (36)$$

$$F_{\alpha r}(\underline{x} = (x_1, \ldots, x_n)) = \max_{\underline{X}_j \in \underline{X}_\alpha} P\Big(\underline{X}_j = \underline{t} = (t_1, \ldots, t_n) | \underline{x}, \underline{t} \in \underline{\mathbf{X}} = \mathbb{R}^n, \; t_k < x_k, \; k = 1, \ldots, n\Big) \,. \quad (37)$$

For the determination of $F_{\alpha l}(\underline{x})$ the relationship in Eq. (25) is used. If $F_{\alpha l}(\underline{x})$ and $F_{\alpha r}(\underline{x})$ form closed connected intervals $[F_{\alpha l}(\underline{x}), F_{\alpha r}(\underline{x})]$ – see Sect. "Fuzzy Probability" for the conditions – the functional values $\tilde{F}(\underline{x})$ are determined based on Eq. (21),

$$\tilde{F}(\underline{x}) = \{(F_\alpha(\underline{x}), \mu(F_\alpha(\underline{x}))) | F_\alpha(\underline{x}) = [F_{\alpha l}(\underline{x}), F_{\alpha r}(\underline{x})], \; \mu(F_\alpha(\underline{x})) = \alpha \; \forall \alpha \in (0, 1]\} \,, \quad (38)$$

In this case, the functional values of the fuzzy probability distribution function $\tilde{F}(\underline{x})$ are continuous and convex fuzzy sets.

In correspondence with Eq. (15), the fuzzy probability distribution function $\tilde{F}(\underline{x})$ of $\underline{\tilde{X}}$ represents the fuzzy set of the probability distribution functions $F_j(\underline{x})$ of all originals $\underline{X}_j$ of $\underline{\tilde{X}}$ with the membership values $\mu(F_j(\underline{x}))$,

$$\tilde{F}(\underline{x}) = \{(F_j(\underline{x}), \mu(F_j(\underline{x}))) \,|\, \underline{X}_j \in \underline{\tilde{X}}, \; \mu(F_j(\underline{x})) = \mu(\underline{X}_j) \; \forall j\} \,. \quad (39)$$

Each original $\underline{X}_j$ determines precisely one trajectory $F_j(\underline{x})$ within the bunch $\tilde{F}(\underline{x})$ of weighted functions $F_j(\underline{x}) \in \tilde{F}(\underline{x})$.

In the one-dimensional case and with the restriction to closed random intervals $[X_{\alpha l}, X_{\alpha r}]$ for each α-level, the fuzzy probability distribution function $\tilde{F}(x)$ is determined by

$$F_{\alpha l}(x) = P(X_{\alpha r} = t_r \,|\, x, t_r \in \mathbf{X}, \; t_r < x) \,, \quad (40)$$

$$F_{\alpha r}(x) = P(X_{\alpha l} = t_l \,|\, x, t_l \in \mathbf{X}, \; t_l < x) \,. \quad (41)$$

In correspondence with traditional probability theory, fuzzy probability density functions $\tilde{f}(\underline{t})$ or $\tilde{f}(\underline{x})$ are defined in association with the $\tilde{F}(\underline{x})$; see Fig. 5. The $\tilde{f}(\underline{t})$ or $\tilde{f}(\underline{x})$ are fuzzy functions which – in the continuous case with respect to randomness – are integrable for each original $\underline{X}_j$ of $\underline{\tilde{X}}$ and satisfy the relationship

$$F_j(\underline{x}) = \int_{t_1=-\infty}^{t_1=x_1} \cdots \int_{t_k=-\infty}^{t_k=x_k} \cdots \int_{t_n=-\infty}^{t_n=x_n} f_j(\underline{t}) \, \mathrm{d}\underline{t} \,, \quad (42)$$

with $\underline{t} = (t_1, \ldots, t_n) \in \underline{\mathbf{X}}$. For each original $\underline{X}_j$ the integration of the associated trajectory $f_j(\underline{x}) \in \tilde{f}(\underline{x})$ leads to the respective trajectory $F_j(\underline{x}) \in \tilde{F}(\underline{x})$.

For the description of a fuzzy random quantity $\underline{\tilde{X}}$, parameters in the form of fuzzy quantities $\tilde{p}_t(\underline{\tilde{X}})$ may be used. These fuzzy parameters may represent any type of

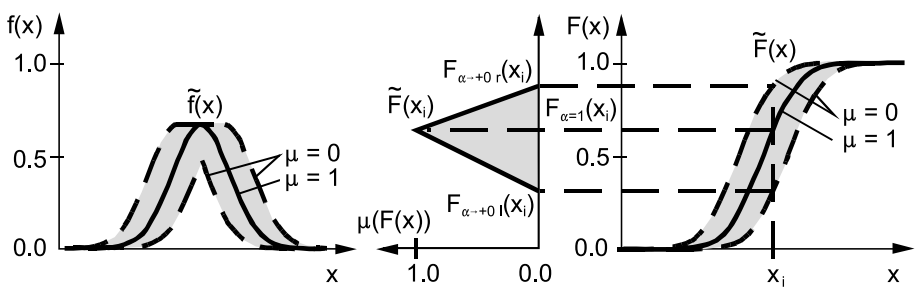

Fuzzy Probability Theory, Figure 5
Fuzzy probability density function $\tilde{f}(\underline{x})$ and fuzzy probability distribution function $\tilde{F}(\underline{x})$ of a continuous fuzzy random variable $\underline{\tilde{X}}$

parameters known from real-valued random quantities, such as moments, weighting factors for different distribution types in a compound distribution, or functional parameters of the distribution functions. The fuzzy parameter $\tilde{p}_t(\underline{\tilde{X}})$ of the fuzzy random quantity $\underline{\tilde{X}}$ is the fuzzy set of the parameter values $p_t(\underline{X}_j)$ of all originals $\underline{X}_j$ with the membership values $\mu(p_t(\underline{X}_j))$,

$$\tilde{p}_t(\underline{\tilde{X}}) = \{(p_t(\underline{X}_j), \mu(p_t(\underline{X}_j))) \,|\, \underline{X}_j \in \underline{\tilde{X}}, \quad \mu(p_t(\underline{X}_j)) = \mu(\underline{X}_j)\ \forall j\} \,. \tag{43}$$

For each α-level, bounds are given for the fuzzy parameter $\tilde{p}_t(\underline{\tilde{X}})$ by

$$p_{t,\alpha l}(\underline{\tilde{X}}) = \min_{\underline{X}_j \in \underline{X}_\alpha} [p_t(\underline{X}_j)] \,, \tag{44}$$

$$p_{t,\alpha r}(\underline{\tilde{X}}) = \max_{\underline{X}_j \in \underline{X}_\alpha} [p_t(\underline{X}_j)] \,. \tag{45}$$

If the fuzzy random quantity $\underline{\tilde{X}}$ represents a fuzzy set of continuous real-valued random quantities, if all fuzzy realizations $\underline{\tilde{x}}_i$ of $\underline{\tilde{X}}$ are connected sets, and if the parameter p_t is defined on a continuous scale, then the fuzzy parameter $\tilde{p}_t(\underline{\tilde{X}})$ is determined by its α-level sets

$$p_{t,\alpha}(\underline{\tilde{X}}) = [p_{t,\alpha l}(\underline{\tilde{X}}), p_{t,\alpha r}(\underline{\tilde{X}})] \,, \tag{46}$$

$$\tilde{p}_t(\underline{\tilde{X}}) = \{(p_{t,\alpha}(\underline{\tilde{X}}), \mu(p_{t,\alpha}(\underline{\tilde{X}}))) \,| \quad \mu(p_{t,\alpha}(\underline{\tilde{X}})) = \alpha\ \forall\, \alpha \in (0,1]\} \,, \tag{47}$$

and represents a continuous and convex fuzzy set.

If a fuzzy random quantity $\underline{\tilde{X}}$ is described by more than one fuzzy parameter $\tilde{p}_t(\underline{\tilde{X}})$, interactive dependencies are generally present between the different fuzzy parameters. If this interaction is neglected, a fuzzy random quantity $\underline{\tilde{X}}_{\text{hull}}$ is obtained, which covers the actual fuzzy random quantity $\underline{\tilde{X}}$ completely. That is, for all realizations of $\underline{\tilde{X}}_{\text{hull}}$ and $\underline{\tilde{X}}$ the following holds,

$$\underline{\tilde{x}}_{i\,\text{hull}} \supseteq \underline{\tilde{x}}_i\ \forall i \,. \tag{48}$$

Fuzzy parameters and fuzzy probability distribution functions do not enable a unique reproduction of fuzzy realizations based on the above description. But they are sufficient to compute fuzzy probabilities correctly for any events defined according to Eq. (13).

The presented concept of fuzzy probability can be extended to fuzzy random functions and processes.

Future Directions

Fuzzy probability theory provides a powerful key to solving a broad variety of practical problems that defy an appropriate treatment with traditional probabilistics due to imprecision of the information for model specification. Fuzzy probabilities reflect aleatory uncertainty and epistemic uncertainty of the underlying problem simultaneously and separately and provide extended information and decision aids. These features can be utilized in all application fields of traditional probability theory and beyond. Respective developments can be observed, primarily, in information science and, increasingly, in engineering. Potential for further extensive fruitful applications exists, for example, in psychology, economy, financial sciences, medicine, biology, social sciences, and even in law. In all cases, fuzzy probability theory is not considered as a replacement for traditional probabilistics but as a beneficial supplement for an appropriate model specification according to the available information in each particular case.

The focus of further developments is seen on both theory and applications. Future theoretical developments may pursue a measure theoretical clarification of the embedding of fuzzy probability theory in the framework of imprecise probabilities under the umbrella of generalized information theory. This is associated with the ambition to unify the variety of available fuzzy probabilistic concepts and to eventually formulate a consistent generalized fuzzy probability theory. Another important issue for future research is the mathematical description and treatment of dependencies within the fuzziness of fuzzy random quantities such as non-probabilistic interaction between fuzzy realizations, between fuzzy parameters, and between fuzzy probabilities of certain events. In parallel to theoretical modeling, further effort is worthwhile towards a consistent concept for the statistical evaluation of imprecise data including the analysis of probabilistic and non-probabilistic dependencies of the data.

In view of applications, the further development of fuzzy probabilistic simulation methods is of central importance. This concerns both theory and numerical algorithms for the direct generation of fuzzy random quantities – in a parametric and in a non-parametric fashion. Representations and computational procedures for fuzzy random quantities must be developed with focus on a high numerical efficiency to enable a solution of real-world problems. For a spread into practice it is further essential to elaborate options and potentials for an interpretation and evaluation of fuzzy probabilistic results such as fuzzy mean values or fuzzy failure probabilities. The most promising potentials for a utilization are seen in worst-case investigations in terms of probability, in a sensitivity analysis with respect to non-probabilistic uncertainty, and in decision-making based on mixed probabilistic/non-probabilistic information.

In summary, fuzzy probability theory and its further developments significantly contribute to an improved uncertainty modeling according to reality.

Bibliography

Primary Literature

1. Alefeld G, Herzberger J (1983) Introduction to interval computations. Academic Press, New York
2. Bandemer H (1992) Modelling uncertain data. Akademie-Verlag, Berlin
3. Bandemer H, Gebhardt A (2000) Bayesian fuzzy kriging. Fuzzy Sets Syst 112:405–418
4. Bandemer H, Gottwald S (1995) Fuzzy sets, fuzzy logic fuzzy methods with applications. Wiley, Chichester
5. Bandemer H, Näther W (1992) Fuzzy data analysis. Kluwer, Dordrecht
6. Beer M (2007) Model-free sampling. Struct Saf 29:49–65
7. Berleant D, Zhang J (2004) Representation and problem solving with distribution envelope determination (denv). Reliab Eng Syst Saf 85(1–3):153–168
8. Bernardini A, Tonon F (2004) Aggregation of evidence from random and fuzzy sets. Special Issue of ZAMM. Z Angew Math Mech 84(10–11):700–709
9. Bodjanova S (2000) A generalized histogram. Fuzzy Sets Syst 116:155–166
10. Colubi A, Domínguez-Menchero JS, López-Díaz M, Gil MA (1999) A generalized strong law of large numbers. Probab Theor Relat Fields 14:401–417
11. Colubi A, Domínguez-Menchero JS, López-Díaz M, Ralescu DA (2001) On the formalization of fuzzy random variables. Inf Sci 133:3–6
12. Colubi A, Domínguez-Menchero JS, López-Díaz M, Ralescu DA (2002) A de[0, 1]-representation of random upper semicontinuous functions. Proc Am Math Soc 130:3237–3242
13. Colubi A, Fernández-García C, Gil MA (2002) Simulation of random fuzzy variables: an empirical approach to statistical/probabilistic studies with fuzzy experimental data. IEEE Trans Fuzzy Syst 10:384–390
14. Couso I, Sanchez L (2008) Higher order models for fuzzy random variables. Fuzzy Sets Syst 159:237–258
15. de Cooman G (2002) The society for imprecise probability: theories and applications. http://www.sipta.org
16. Diamond P (1990) Least squares fitting of compact set-valued data. J Math Anal Appl 147:351–362
17. Diamond P, Kloeden PE (1994) Metric spaces of fuzzy sets: theory and applications. World Scientific, Singapore
18. Dubois D, Prade H (1980) Fuzzy sets and systems theory and applications. Academic Press, New York
19. Dubois D, Prade H (1985) A review of fuzzy set aggregation connectives. Inf Sci 36:85–121
20. Dubois D, Prade H (1986) Possibility theory. Plenum Press, New York
21. Fellin W, Lessmann H, Oberguggenberger M, Vieider R (eds) (2005) Analyzing uncertainty in civil engineering. Springer, Berlin
22. Feng Y, Hu L, Shu H (2001) The variance and covariance of fuzzy random variables and their applications. Fuzzy Sets Syst 120(3):487–497
23. Ferson S, Hajagos JG (2004) Arithmetic with uncertain numbers: rigorous and (often) best possible answers. Reliab Eng Syst Saf 85(1–3):135–152
24. Fetz T, Oberguggenberger M (2004) Propagation of uncertainty through multivariate functions in the framework of sets of probability measures. Reliab Eng Syst Saf 85(1–3):73–87
25. Ghanem RG, Spanos PD (1991) Stochastic finite elements: a spectral approach. Springer, New York; Revised edition 2003, Dover Publications, Mineola
26. González-Rodríguez G, Montenegro M, Colubi A, Ángeles Gil M (2006) Bootstrap techniques and fuzzy random variables: Synergy in hypothesis testing with fuzzy data. Fuzzy Sets Syst 157(19):2608–2613
27. Grzegorzewski P (2000) Testing statistical hypotheses with vague data. Fuzzy Sets Syst 112:501–510
28. Hall JW, Lawry J (2004) Generation, combination and extension of random set approximations to coherent lower and upper probabilities. Reliab Eng Syst Saf 85(1–3):89–101
29. Helton JC, Johnson JD, Oberkampf WL (2004) An exploration of alternative approaches to the representation of uncertainty in model predictions. Reliab Eng Syst Saf 85(1–3):39–71
30. Helton JC, Oberkampf WL (eds) (2004) Special issue on alternative representations of epistemic uncertainty. Reliab Eng Syst Saf 85(1–3):1–369
31. Hung W-L (2001) Bootstrap method for some estimators based on fuzzy data. Fuzzy Sets Syst 119:337–341
32. Hwang C-M, Yao J-S (1996) Independent fuzzy random variables and their application. Fuzzy Sets Syst 82:335–350
33. Jang L-C, Kwon J-S (1998) A uniform strong law of large numbers for partial sum processes of fuzzy random variables indexed by sets. Fuzzy Sets Syst 99:97–103
34. Joo SY, Kim YK (2001) Kolmogorovs strong law of large numbers for fuzzy random variables. Fuzzy Sets Syst 120:499–503
35. Kim YK (2002) Measurability for fuzzy valued functions. Fuzzy Sets Syst 129:105–109
36. Klement EP, Puri ML, Ralescu DA (1986) Limit theorems for fuzzy random variables. Proc Royal Soc A Math Phys Eng Sci 407:171–182
37. Klement EP (1991) Fuzzy random variables. Ann Univ Sci Budapest Sect Comp 12:143–149
38. Klir GJ (2006) Uncertainty and information: foundations of generalized information theory. Wiley-Interscience, Hoboken
39. Klir GJ, Folger TA (1988) Fuzzy sets, uncertainty, and information. Prentice Hall, Englewood Cliffs
40. Körner R (1997) Linear models with random fuzzy variables. Phd thesis, Bergakademie Freiberg, Fakultät für Mathematik und Informatik
41. Körner R (1997) On the variance of fuzzy random variables. Fuzzy Sets Syst 92:83–93
42. Körner R, Näther W (1998) Linear regression with random fuzzy variables: extended classical estimates, best linear estimates, least squares estimates. Inf Sci 109:95–118
43. Krätschmer V (2001) A unified approach to fuzzy random variables. Fuzzy Sets Syst 123:1–9
44. Krätschmer V (2002) Limit theorems for fuzzy-random variables. Fuzzy Sets Syst 126:253–263
45. Krätschmer V (2004) Probability theory in fuzzy sample space. Metrika 60:167–189
46. Kruse R, Meyer KD (1987) Statistics with vague data. Reidel, Dordrecht

47. Kwakernaak H (1978) Fuzzy random variables I. definitions and theorems. Inf Sci 15:1–19
48. Kwakernaak H (1979) Fuzzy random variables II. algorithms and examples for the discrete case. Inf Sci 17:253–278
49. Li S, Ogura Y, Kreinovich V (2002) Limit theorems and applications of set valued and fuzzy valued random variables. Kluwer, Dordrecht
50. Lin TY, Yao YY, Zadeh LA (eds) (2002) Data mining, rough sets and granular computing. Physica, Germany
51. López-Díaz M, Gil MA (1998) Reversing the order of integration in iterated expectations of fuzzy random variables, and statistical applications. J Stat Plan Inference 74:11–29
52. Matheron G (1975) Random sets and integral geometry. Wiley, New York
53. Möller B, Graf W, Beer M (2000) Fuzzy structural analysis using alpha-level optimization. Comput Mech 26:547–565
54. Möller B, Beer M (2004) Fuzzy randomness – uncertainty in civil engineering and computational mechanics. Springer, Berlin
55. Möller B, Reuter U (2007) Uncertainty forecasting in engineering. Springer, Berlin
56. Muhanna RL, Mullen RL, Zhang H (2007) Interval finite element as a basis for generalized models of uncertainty in engineering mechanics. J Reliab Comput 13(2):173–194
57. Gil MA, López-Díaz M, Ralescu DA (2006) Overview on the development of fuzzy random variables. Fuzzy Sets Syst 157(19):2546–2557
58. Näther W, Körner R (2002) Statistical modelling, analysis and management of fuzzy data, chapter on the variance of random fuzzy variables. Physica, Heidelberg, pp 25–42
59. Näther W (2006) Regression with fuzzy random data. Comput Stat Data Analysis 51:235–252
60. Näther W, Wünsche A (2007) On the conditional variance of fuzzy random variables. Metrika 65:109–122
61. Oberkampf WL, Helton JC, Sentz K (2001) Mathematical representation of uncertainty. In: AIAA non-deterministic approaches forum, number AIAA 2001–1645. AIAA, Seattle
62. Pedrycz W, Skowron A, Kreinovich V (eds) (2008) Handbook of granular computing. Wiley, New York
63. Puri ML, Ralescu D (1986) Fuzzy random variables. J Math Anal Appl 114:409–422
64. Puri ML, Ralescu DA (1983) Differentials of fuzzy functions. J Math Anal Appl 91:552–558
65. Puri ML, Ralescu DA (1991) Convergence theorem for fuzzy martingales. J Math Anal Appl 160:107–122
66. Rodríguez-Muñiz L, López-Díaz M, Gil MA (2005) Solving influence diagrams with fuzzy chance and value nodes. Eur J Oper Res 167:444–460
67. Samarasooriya VNS, Varshney PK (2000) A fuzzy modeling approach to decision fusion under uncertainty. Fuzzy Sets Syst 114:59–69
68. Schenk CA, Schuëller GI (2005) Uncertainty assessment of large finite element systems. Springer, Berlin
69. Schuëller GI, Spanos PD (eds) (2001) Proc int conf on monte carlo simulation MCS 2000. Swets and Zeitlinger, Monaco
70. Shafer G (1976) A mathematical theory of evidence. Princeton University Press, Princeton
71. Song Q, Leland RP, Chissom BS (1997) Fuzzy stochastic fuzzy time series and its models. Fuzzy Sets Syst 88:333–341
72. Taheri SM, Behboodian J (2001) A bayesian approach to fuzzy hypotheses testing. Fuzzy Sets Syst 123:39–48
73. Terán P (2006) On borel measurability and large deviations for fuzzy random variables. Fuzzy Sets Syst 157(19):2558–2568
74. Terán P (2007) Probabilistic foundations for measurement modelling with fuzzy random variables. Fuzzy Sets Syst 158(9):973–986
75. Viertl R (1996) Statistical methods for non-precise data. CRC Press, Boca Raton
76. Viertl R, Hareter D (2004) Generalized Bayes' theorem for non-precise a-priori distribution. Metrika 59:263–273
77. Viertl R, Trutschnig W (2006) Fuzzy histograms and fuzzy probability distributions. In: Proceedings of the 11th Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems, Editions EDK, Paris, CD-ROM
78. Walley P (1991) Statistical reasoning with imprecise probabilities. Chapman & Hall, London
79. Wang G, Qiao Z (1994) Convergence of sequences of fuzzy random variables and its application. Fuzzy Sets Syst 63:187–199
80. Wang G, Zhang Y (1992) The theory of fuzzy stochastic processes. Fuzzy Sets Syst 51:161–178
81. Weichselberger K (2000) The theory of interval-probability as a unifying concept for uncertainty. Int J Approx Reason 24(2–3):149–170
82. Yager RR (1984) A representation of the probability of fuzzy subsets. Fuzzy Sets Syst 13:273–283
83. Zadeh LA (1965) Fuzzy sets. Inf Control 8:338–353
84. Zadeh LA (1968) Probability measures of fuzzy events. J Math Anal Appl 23:421–427
85. Zhong Q, Yue Z, Guangyuan W (1994) On fuzzy random linear programming. Fuzzy Sets Syst 65(1):31–49
86. Zimmermann HJ (1992) Fuzzy set theory and its applications. Kluwer, Boston

Books and Reviews

Ayyub BM (1998) Uncertainty modeling and analysis in civil engineering. CRC Press, Boston

Blockley DI (1980) The nature of structural design and safety. Ellis Horwood, Chichester

Buckley JJ (2003) Fuzzy Probabilities. Physica/Springer, Heidelberg

Cai K-Y (1996) Introduction to fuzzy reliability. Kluwer, Boston

Chou KC, Yuan J (1993) Fuzzy-bayesian approach to reliability of existing structures. ASCE J Struct Eng 119(11):3276–3290

Elishakoff I (1999) Whys and hows in uncertainty modelling probability, fuzziness and anti-optimization. Springer, New York

Feng Y (2000) Decomposition theorems for fuzzy supermartingales and submartingales. Fuzzy Sets Syst 116:225–235

Gil MA, López-Díaz M (1996) Fundamentals and bayesian analyses of decision problems with fuzzy-valued utilities. Int J Approx Reason 15:203–224

Gil MA, Montenegro M, González-Rodríguez G, Colubi A, Casals MR (2006) Bootstrap approach to the multi-sample test of means with imprecise data. Comput Stat Data Analysis 51:148–162

Grzegorzewski P (2001) Fuzzy sets b defuzzification and randomization. Fuzzy Sets Syst 118:437–446

Grzegorzewski P (2004) Distances between intuitionistic fuzzy sets and/or interval-valued fuzzy sets based on the Hausdorff metric. Fuzzy Sets Syst 148(2):319–328

Hareter D (2004) Time series analysis with non-precise data. In: Wojtkiewicz S, Red-Horse J, Ghanem R (eds) 9th ASCE specialty conference on probabilistic mechanics and structural reliability. Sandia National Laboratories, Albuquerque

Helton JC, Cooke RM, McKay MD, Saltelli A (eds) (2006) Special issue: The fourth international conference on sensitivity analysis of model output – SAMO 2004. Reliab Eng Syst Saf 91:(10–11):1105–1474

Hirota K (1992) An introduction to fuzzy logic applications in intelligent systems. In: Kluwer International Series in Engineering and Computer Science, Chapter probabilistic sets: probabilistic extensions of fuzzy sets, vol 165. Kluwer, Boston, pp 335–354

Klement EP, Puri ML, Ralescu DA (1984) Cybernetics and Systems Research 2, Chapter law of large numbers and central limit theorems for fuzzy random variables. Elsevier, North-Holland, pp 525–529

Kutterer H (2004) Statistical hypothesis tests in case of imprecise data, V Hotine-Marussi Symposium on Mathematical Geodesy. Springer, Berlin, pp 49–56

Li S, Ogura Y (2003) A convergence theorem of fuzzy-valued martingales in the extended hausdorff metric h(inf). Fuzzy Sets Syst 135:391–399

Li S, Ogura Y, Nguyen HT (2001) Gaussian processes and martingales for fuzzy valued random variables with continuous parameter. Inf Sci 133:7–21

Li S, Ogura Y, Proske FN, Puri ML (2003) Central limit theorems for generalized set-valued random variables. J Math Anal Appl 285:250–263

Liu B (2002) Theory and practice of uncertainty programming. Physica, Heidelberg

Liu Y, Qiao Z, Wang G (1997) Fuzzy random reliability of structures based on fuzzy random variables. Fuzzy Sets Syst 86:345–355

Möller B, Graf W, Beer M (2003) Safety assessment of structures in view of fuzzy randomness. Comput Struct 81:1567–1582

Möller B, Liebscher M, Schweizerhof K, Mattern S, Blankenhorn G (2008) Structural collapse simulation under consideration of uncertainty – improvement of numerical efficiency. Comput Struct 86(19–20):1875–1884

Montenegro M, Casals MR, Lubiano MA, Gil MA (2001) Two-sample hypothesis tests of means of a fuzzy random variable. Inf Sci 133:89–100

Montenegro M, Colubi A, Casals MR, Gil MA (2004) Asymptotic and bootstrap techniques for testing the expected value of a fuzzy random variable. Metrika 59:31–49

Montenegro M, González-Rodríguez G, Gil MA, Colubi A, Casals MR (2004) Soft methodology and random information systems, chapter introduction to ANOVA with fuzzy random variables. Springer, Berlin, pp 487–494

Muhanna RL, Mullen RL (eds) (2004) Proceedings of the NSF workshop on reliable engineering computing. Center for Reliable Engineering Computing. Georgia Tech Savannah, Georgia

Muhanna RL, Mullen RL (eds) (2006) NSF workshop on reliable engineering computing. center for reliable engineering computing. Georgia Tech Savannah, Georgia

Negoita VN, Ralescu DA (1987) Simulation, knowledge-based computing and fuzzy-statistics. Van Nostrand, Reinhold, New York

Oberguggenberger M, Schuëller GI, Marti K (eds) (2004) Special issue on application of fuzzy sets and fuzzy logic to engineering problems. ZAMM: Z Angew Math Mech 84(10–11):661–776

Okuda T, Tanaka H, Asai K (1978) A formulation of fuzzy decision problems with fuzzy information using probability measures of fuzzy events. Inf Control 38:135–147

Proske FN, Puri ML (2002) Strong law of large numbers for banach space valued fuzzy random variables. J Theor Probab 15: 543–551

Reddy RK, Haldar A (1992) Analysis and management of uncertainty: theory and applications, chapter a random-fuzzy reliability analysis of engineering systems. North-Holland, Amsterdam, pp 319–329

Reddy RK, Haldar A (1992) A random-fuzzy analysis of existing structures. Fuzzy Sets Syst 48:201–210

Ross TJ (2004) Fuzzy logic with engineering applications, 2nd edn. Wiley, Philadelphia

Ross TJ, Booker JM, Parkinson WJ (eds) (2002) Fuzzy logic and probability applications – bridging the gap. SIAM & ASA, Philadelphia

Stojakovic M (1994) Fuzzy random variables, expectation, and martingales. J Math Anal Appl 184:594–606

Terán P (2004) Cones and decomposition of sub- and supermartingales. Fuzzy Sets Syst 147:465–474

Tonon F, Bernardini A (1998) A random set approach to the optimization of uncertain structures. Comput Struct 68(6):583–600

Weichselberger K (2000) The theory of interval-probability as a unifying concept for uncertainty. Int J Approx Reason 24(2–3): 149–170

Yukari Y, Masao M (1999) Interval and paired probabilities for treating uncertain events. IEICE Trans Inf Syst E82–D(5):955–961

Zadeh L (1985) Is probability theory sufficient for dealing with uncertainty in ai: A negative view. In: Proceedings of the 1st Annual Conference on Uncertainty in Artificial Intelligence (UAI-85). Elsevier Science, New York, pp 103–116

Zhang Y, Wang G, Su F (1996) The general theory for response analysis of fuzzy stochastic dynamical systems. Fuzzy Sets Syst 83:369–405

Fuzzy Sets Theory, Foundations of

Janusz Kacprzyk
Systems Research Institute,
Polish Academy of Sciences,
Warsaw, Poland

Article Outline

Glossary
Definition of the Subject
Introduction
Fuzzy Sets – Basic Definitions and Properties
Fuzzy Relations
Linguistic Variable, Fuzzy Conditional Statement, and Compositional Rule of Inference
The Extension Principle
Fuzzy Numbers
Fuzzy Events and Their Probabilities
Defuzzification of Fuzzy Sets
Fuzzy Logic – Basic Issues
Bellman and Zadeh's General Approach to Decision Making Under Fuzziness

Concluding Remarks
Bibliography

Glossary

Fuzzy set A mathematical tool that can formally characterize an imprecise concept. Whereas a conventional set to which elements can either belong or not, elements in a fuzzy set can belong to some extent, from zero, which stands for a full nonbelongingness) to one, which stands for a full belongingness, through all intermediate values.

Fuzzy relation A mathematical tool that can formally characterize that which is imprecisely specified, notably by using natural language, relations between variables, for instance, *similar*, *much greater than*, *almost equal*, etc.

Extension principle Makes it possible to extend relations, algorithms, etc. defined for variables that take on nonfuzzy (e. g. real) values to those that take on fuzzy values.

Linguistic variable, fuzzy conditional statement, compositional rule of inference Make it possible to use variables, which take on linguistic (instead of numeric) values to represent relations between such variables, by using fuzzy conditional statements and use them in inference by using the compositional rule of inference.

Fuzzy event and its probability Make it possible to formally define events which are imprecisely specified, like "high temperature" and calculate their probabilities, for instance the probability of a "high temperature tomorrow".

Fuzzy logic Provides formal means for the representation of, and inference based on imprecisely specified premises and rules of inference; can be understood in different ways, basically as fuzzy logic in a narrow sense, being some type of multivalued logic, and fuzzy logic in a broad sense, being a way to formalize inference based on imprecisely specified premises and rules of inference.

Definition of the Subject

We provide a brief exposition of basic elements of Zadeh's [95] fuzzy sets theory. We discuss basic properties, operations on fuzzy sets, fuzzy relations and their compositions, linguistic variables, the extension principle, fuzzy arithmetic, fuzzy events and their probabilities, fuzzy logic, fuzzy dynamic systems, etc. We also outline Bellman and Zadeh's [8] general approach to decision making in a fuzzy environment which is a point of departure for virtually all fuzzy decision making, optimization, control, etc. models.

Introduction

This paper is meant to briefly expose a novice reader to basic elements of theory of fuzzy sets and fuzzy systems viewed for our purposes as an effective and efficient means and calculus to deal with imprecision in the definition of data, information and knowledge, and to provide tools and techniques for dealing with imprecision therein. Our exposition will be as formal as necessary, of more intuitive and constructive a character, so that fuzzy tools and techniques can be useful for the multidisciplinary audience of this encyclopedia. For the readers requiring or interested in a deeper exposition of fuzzy sets and related concepts, we will recommend many relevant references, mainly books. However, as the number of books and volumes on this topic and its applications in a variety of fields is huge, we will recommend some of them only, mostly those better known ones. For the newest literature entries the readers should consult the most recent catalogs of major scientific publishers who have books and edited volumes on fuzzy sets/logic and their applications.

Our discussion will proceed, on the other hand, in the *pure* fuzzy setting, and we will not discuss possibility theory (which is related to fuzzy sets theory). The reader interested in possibility theory is referred to, e. g., Dubois and Prade [29,30] or their article in this encyclopedia.

We will consecutively discuss the idea of a fuzzy set, basic properties of fuzzy sets, operations on fuzzy sets, some extensions of the basic concept of a fuzzy set, fuzzy relations and their compositions, linguistic variables, fuzzy conditional statements, and the compositional rule of inference, the extension principle, fuzzy arithmetic, fuzzy events and their probabilities, fuzzy logic, fuzzy dynamic systems, etc. We also outline Bellman and Zadeh's [8] general approach to decision making in a fuzzy environment which is a point of departure for virtually all fuzzy decision making, optimization, control, etc. models.

Fuzzy Sets – Basic Definitions and Properties

Fuzzy sets theory, introduced by Zadeh in 1965 [95], is a simple yet very powerful, effective and efficient means to represent and handle imprecise information (of vagueness type) exemplified by *tall* buildings, *large* numbers, etc. We will present fuzzy sets theory as some *calculus of imprecision*, not as a new set theory in the mathematical sense.

The Idea of a Fuzzy Set

From our point of view, the main purpose of a (conventional) set in mathematics is to formally characterize some concept (or property). For instance, the concept of "in-

teger numbers which are greater than or equal three and less than or equal ten" may be uniquely represented just by showing all integer numbers that satisfy this condition; that is, given by the following set: $\{x \in I: 3 \leq x \leq 10\} = \{3, 4, 5, 6, 7, 8, 9, 10\}$ where I is the set of integers. Notice that we need to specify first a *universe of discourse* (universe, universal set, referential, reference set, etc.) that contains all those elements which are relevant for the particular concept as, e. g., the set of integers I in our example.

A conventional set, say A, may be equated with its *characteristic function* defined as

$$\varphi_A : X \longrightarrow \{0, 1\} \tag{1}$$

which associates with each element x of a universe of discourse $X = \{x\}$ a number $\varphi(x) \in \{0, 1\}$ such that: $\varphi_A(x) = 0$ means that $x \in X$ does not belong to the set A, and $\varphi_A(x) = 1$ means that x belongs to the set A.

Therefore, for the set verbally defined as integer numbers which are greater than or equal three and less than or equal ten, its equivalent set $A = \{3, 4, 5, 6, 7, 8, 9, 10\}$, listing all the respective integer numbers, may be represented by its characteristic function

$$\varphi_A(x) = \begin{cases} 1 & \text{for } x \in \{3, 4, 5, 6, 7, 8, 9, 10\} \\ 0 & \text{otherwise .} \end{cases}$$

Notice that in a conventional set there is a clear-cut differentiation between elements belonging to the set and not, i. e. the transition from the belongingness to nonbelongingness is clear-cut and abrupt.

However, it is easy to notice that a serious difficulty arises when we try to formalize by means of a set vague concepts which are commonly encountered in everyday discourse and widely used by humans as, e. g., the statement "integer numbers which are *more or less* equal to six." Evidently, the (conventional) set cannot be used to adequately characterize such an imprecise concept because an abrupt and clear-cut differentiation between the elements belonging and not belonging to the set is artificial here.

This has led Zadeh [95] to the idea of a *fuzzy set* which is a class of objects with unsharp boundaries, i. e. in which the transition from the belongingness to nonbelongingness is not abrupt; thus, elements of a fuzzy set may belong to it to *partial degrees*, from the full belongingness to the full nonbelongingness through all intermediate values. Notice that this is presumably the most natural and simple way to formally define the imprecision of meaning.

We should therefore start again with a *universe of discourse* (universe, universal set, referential, reference set, etc.) containing all elements relevant for the (imprecise) concept we wish to formally represent. Then, the characteristic function $\varphi : X \longrightarrow \{0, 1\}$ is replaced by a *membership function* defined as

$$\mu_A : X \longrightarrow [0, 1] \tag{2}$$

such that $\mu_A(x) \in [0, 1]$ is the degree to which an element $x \in X$ belongs to the fuzzy set A: From $\mu_A(x) = 0$ for the full nonbelongingness to $\mu_A(x) = 1$ for the full belongingness, through all intermediate $(0 < \mu_A(x) < 1)$ values.

Now, if we consider as an example the concept of integer numbers which are *more or less* six. Then $x = 6$ certainly belongs to this set so that $\mu_A(6) = 1$, the numbers five and seven belong to this set *almost surely* so that $\mu_A(5)$ and $\mu_A(7)$ are very close to one, and the more a number differs from six, the less its $\mu_A(.)$. Finally, the numbers below one and above ten do not belong to this set, so that their $\mu_A(.) = 0$. This may be sketched as in Fig. 1 though we should bear in mind that although in our example the membership function is evidently defined for the integer numbers (x's) only, it is depicted in a continuous form to be more illustrative.

In practice the membership function is usually assumed to be piecewise linear as shown in Fig. 2 (for the same fuzzy set as in Fig. 1, i. e. the fuzzy set integer numbers which are *more or less* six). To specify the membership function we then need four numbers only: a, b, c, and d as, e. g., $a = 2$, $b = 5$, $c = 7$, and $d = 10$ in Fig. 2.

Notice that the particular form of a membership function is *subjective* as opposed to an *objective* form of a characteristic function. However, this may be viewed as quite natural as the underlying concepts are subjective indeed as, e. g., the set of integer numbers *more or less* six depend on an individual opinion. Unfortunately, this inherent subjectivity of the membership function may lead to some problems in many formal models in which users would rather have a limit to the scope of subjectivity. We will comment on this issue later on.

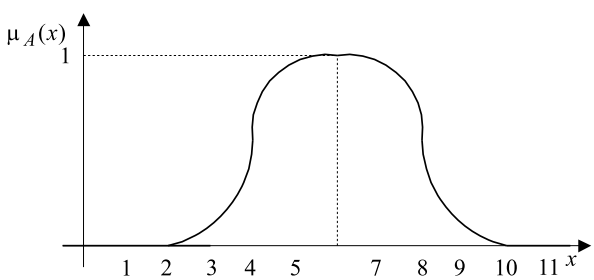

Fuzzy Sets Theory, Foundations of, Figure 1
Membership function of a fuzzy set, integer numbers which are *more or less* six

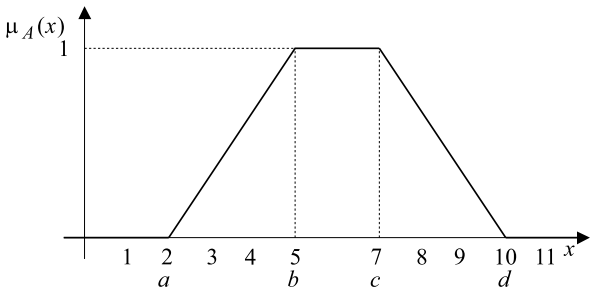

Fuzzy Sets Theory, Foundations of, Figure 2
Membership function of a fuzzy set, integer numbers which are *more or less* six

We will now define formally a fuzzy set in a form that is very often used.

A *fuzzy set* A in a universe of discourse $X = \{x\}$, written A in X, is defined as a set of pairs

$$A = \{(\mu_A(x), x)\} \tag{3}$$

where $\mu_A : X \longrightarrow [0,1]$ is the *membership function* of A and $\mu_A(x) \in [0,1]$ is the *grade of membership* (or a *membership grade*) of an element $x \in X$ in a fuzzy set A.

Needless to say that our definition of a fuzzy set (3) is clearly equivalent to the definition of the membership function (2) because a function may be represented by a set of pairs *argument–value of the function for this argument*. For our purposes, however, the definition (3) is more *set-theoretic-like* which will often be more convenient.

So, in this paper we will practically equate fuzzy sets with their membership functions saying, e. g., a fuzzy set, $\mu_A(x)$, and also very often we will equate fuzzy sets with their labels saying, e. g., a fuzzy set, *large numbers*, with the understanding that the label *large numbers* is equivalent to the fuzzy set mentioned, written $A =$ *large numbers*. However, we will use the notation $\mu_A(x)$ for the membership function of a fuzzy set A in X, and not an abbreviated notation $A(x)$ as in some more technical papers, to be consistent with our *more-set-theoretic-like* convention.

For practical reasons, it is very often assumed (also in this paper) that all the universes of discourse are finite as, e. g., $X = \{x_1, \ldots, x_n\}$. In such a case the pair $\{(\mu_A(x), x)\}$ will be denoted by $\mu_A(x)/x$ which is called a *fuzzy singleton*.

Then, the fuzzy set A in X will be written as

$$\begin{aligned} A &= \{(\mu_A(x), x)\} = \{\mu_A(x)/x\} \\ &= \mu_A(x_1)/x_1 + \ldots + \mu_A(x_n)/x_n = \sum_{i=1}^{n} \mu_A(x_i)/x_i \,, \end{aligned} \tag{4}$$

where $+$ and $\sum$ are meant in the set-theoretic sense. By convention, the pairs $\mu_A(x)/x$ with $\mu_A(x) = 0$ are omitted here.

A conventional (nonfuzzy) set may obviously be written in the fuzzy sets notation introduced above, for instance the (non-fuzzy) set, integer numbers greater than or equal three and less than or equal ten, may be written as

$$A = 1/3 + 1/4 + 1/5 + 1/6 + 1/7 + 1/8 + 1/9 + 1/10 \,.$$

The family of all fuzzy sets defined in X is denoted by $\mathcal{A}$; it includes evidently also the empty fuzzy set to be defined by (9), i. e. $A = \emptyset$ such that $\mu_A(x) = 0$, for each $x \in X$, and the whole universe of discourse X written as $X = 1/x_1 + \ldots + 1/x_n$.

The concept of a fuzzy set as defined above has been the point of departure for the *theory of fuzzy sets* (or *fuzzy sets theory*) which will be briefly sketched below. We will again follow a more intuitive and less formal presentation, which is better suited for this encyclopedia.

Some Extensions of the Concept of Zadeh's Fuzzy Set

The concept of Zadeh's [95] fuzzy set introduced in the previous section is the by far the simplest and most natural way to *fuzzify* the concept of a (conventional) set, and clearly provides what we mean to represent and handle imprecision. However, its underlying elements are the most straightforward possible. This concerns above all the membership function. Therefore, it is quite natural that some extensions have been presented to this basic concept. We will just briefly mention some of them.

First, it is quite easy to notice that though the definition of a fuzzy set by the membership function of the type $\mu_A : X \longrightarrow [0,1]$ is the simplest and most straightforward one, allowing for a gradual transition from the belongingness and nonbelongingness, it can readily be extended. The same role is namely played by a generalized definition by a membership function of the type

$$\mu_A : X \longrightarrow L \,, \tag{5}$$

where L is some (partially) ordered set as, e. g., a lattice.

This obvious, but powerful extension was introduced by Goguen [37] as an *L-fuzzy set*, where *l* stands for a lattice. Notice that by using a lattice as the set of values of the membership function we can accommodate situations *i* which we can encounter elements of the universe of discourse which are not comparable.

Another quite and obvious extension, already mentioned but not developed by Zadeh [95,101] is the concept of a *type 2 fuzzy set*. The rationale behind this concept is obvious. One can easily imagine that the values of

grades of membership of the particular elements of a universe of discourse are fuzzy sets themselves. And further, these fuzzy sets may have grades of membership which are type 2 fuzzy sets, which leads to type 3 fuzzy sets, and one can continue arriving at type n fuzzy sets.

The next, natural extension is that instead of assuming that the degrees of membership are real numbers from the unit interval, one can go a step further and replace these real numbers from $[0, 1]$ by intervals with endpoints belonging to the unit interval. This leads to *interval valued fuzzy sets* which are attributed to Dubois and Gorzałłczany (cf. Klir and Yuan [53]). Notice that by using intervals as values of degrees of membership we significantly increase our ability to represent imprecision.

A more radical extension to the concept of Zade's fuzzy set is the so-called *intuitionistic fuzzy set* introduced by Atanassov [1,2].

An intuitionistic fuzzy set A' in a universe of discourse X is defined as

$$A' = \{\langle x, \mu_{A'}(x), \nu_A(x)\rangle | x \in X\} \tag{6}$$

where:

- The degree of membership is

$$\mu_{A'} : X \to [0, 1] ,$$

- the degree of non-membership is

$$\nu_{A'} : X \to [0, 1] ,$$

- and the condition holds

$$0 \le \mu_{A'}(x) + \nu_{A'}(x) \le 1 ; \quad \text{for each } x \in X .$$

Obviously, each (conventional) fuzzy set A in X corresponds to the following intuitionistic fuzzy set A' in X:

$$A = \{\langle x, \mu_{A'}(x), 1 - \mu_{A'}(x)\rangle | x \in X\} \tag{7}$$

For each intuitionistic fuzzy set A' in X, we call

$$\pi_{A'}(x) = 1 - \mu_{A'}(x) - \nu_{A'}(x) ; \quad \text{for each } x \in X \tag{8}$$

the intuitionistic fuzzy index (or a hesitation margin) of x in A'. The intuitionistic fuzzy index expresses a lack of knowledge of whether an element $x \in X$ belongs to an intuitionistic fuzzy set A' or not.

Notice that the concept of an intuitionistic fuzzy set is a substantial departure from the concept of a (conventional) fuzzy set as it assumes that the degrees of membership and non-membership do not sum up to one, as it is the case in virtually all traditional set theories and their extensions. For more information, we refer the reader to Atanassov's [3] book.

We will not use these extensions in this short introductory article, and the interested readers are referred to the source literature cited.

Basic Definition and Properties Related to Fuzzy Sets

We will now provide a brief account of basic definitions and properties related to fuzzy sets. We illustrate them with simple examples.

A fuzzy set A is said to be *empty*, written $A = \emptyset$, if and only if

$$\mu_A(x) = 0 , \quad \text{for each } x \in X \tag{9}$$

and since we omit the pairs $0/x$, an empty fuzzy set is really void in the notation (4) as there are no singletons in the right-hand side.

Two fuzzy sets A and B defined in the same universe of discourse X are said to be *equal*, written $A = B$, if and only if

$$\mu_A(x) = \mu_B(x) , \quad \text{for each } x \in X \tag{10}$$

Example 1 Suppose that $X = \{1, 2, 3\}$ and

$$A = 0.1/1 + 0.5/2 + 1/3$$
$$B = 0.2/1 + 0.5/2 + 1/3$$
$$C = 0.1/1 + 0.5/2 + 1/3$$

then $A = C$ but $A \neq C$ and $B \neq C$.

It is easy to see that this classic definition of the equality of two fuzzy sets by (10) is rigid and clear-cut, contradicting in a sense our intuitive feeling that the equality of fuzzy sets should be *softer*, and not abrupt, i. e. should rather be to some degree, from zero to one. We will show below one of possible definitions of such an equality to a degree.

Two fuzzy sets A and B defined in X are said to be *equal to a degree* $e(A, B) \in [0, 1]$, written $A =_e B$, and the degree of equality $e(A, B)$ may be defined in many ways exemplified by those given below (cf. Bandler and Kohout [6]).

First, to simplify, we denote:

Case 1: $A = B$ in the sense of (10);

Case 2: $A \neq B$ in the sense of (10) and $T = \{x \in X : \mu_A(x) \neq \mu_B(x)\}$;

Case 3: $A \neq B$ in the sense of (10) and there exists an $x \in X$ such that

$$\begin{aligned} & \mu_A(x) = 0 \quad \text{and} \quad \mu_B(x) \neq 0 \\ \text{or} \quad & \mu_A(x) \neq 0 \quad \text{and} \quad \mu_B(x) = 0 . \end{aligned}$$

Case 4: $A \neq B$ in the sense of (10) and there exists an $x \in X$ such that

$$\begin{aligned} &\mu_A(x) = 0 \quad \text{and} \quad \mu_B(x) = 1 \\ \text{or} \quad &\mu_A(x) = 1 \quad \text{and} \quad \mu_B(x) = 0\,. \end{aligned}$$

Now, the following degrees of equality of two fuzzy sets, A and B, may be defined:

$$e_1(A,B) = \begin{cases} 1 & \text{for case 1} \\ \bigwedge_{x\in T}[\mu_A(x) \wedge \mu_B(x)] & \text{for case 2} \\ 0 & \text{for case 3} \end{cases} \tag{11}$$

$$e_2(A,B) = \begin{cases} 1 & \text{for case 1} \\ \bigwedge_{x\in T}[\mu_A(x)/\mu_B(x) \\ -\mu_B(x)/\mu_A(x)] & \text{for case 2} \\ 0 & \text{for case 3} \end{cases} \tag{12}$$

$$e_3(A,B) = \begin{cases} 1 & \text{for case 1} \\ 1 - \max_{x\in X} \mid \mu_A(x) \\ -\mu_B(x) \mid & \text{for case 2} \\ 0 & \text{for case 4} \end{cases} \tag{13}$$

$$e_4(A,B) = \begin{cases} 1 & \text{for case 1} \\ \max_{x\in X}\{[(1-\mu_A(x)] \\ \wedge[\mu_A(x) \vee (1-\mu_B(x))]\} & \text{for case 2} \\ 0 & \text{for case 4}\,. \end{cases} \tag{14}$$

Now we will proceed to the second basic concept of the containment between two fuzzy sets.

A fuzzy set A defined in X is said to be *contained in* or, alternatively, is said to be a *subset of* a fuzzy set B in X, written $A \subseteq B$, if and only if

$$\mu_A(x) \leq \mu_B(x)\,, \quad \text{for each } x \in X\,. \tag{15}$$

Example 2 Suppose that $X = \{1,2,3\}$ and

$$\begin{aligned} A &= 0.1/1 + 0.5/2 + 1/4 \\ B &= 0.1/1 + 0.4/2 + 0.9/3 \\ C &= 0.1/1 + 0.6/2 + 1/3\,, \end{aligned}$$

then only $B \subseteq A$.

This traditional definition of containment is clearly rigid and clear-cut, and hence there have been proposed many other *softer* definitions in which a *degree of containment*, $c(A,B) \in [0,1]$, has been employed. Once again, Bandler and Kohout's [6] definitions can be mentioned here, and these basically follow the line of reasoning analogous to that behind the degree of equality, i. e. (11)–(14). The above definitions of the degree of equality and containment are popular but not the only possible ones, some remarks can be found in the books by Dubois and Prade [28]!, Klir and Folger [51], Klir and Yuan [53].

Let us proceed now to some further relevant foundational notions.

A fuzzy set A defined in X is said to be *normal* if and only if

$$\max_{x\in X} \mu_A(x) = 1 \tag{16}$$

i. e. when the membership function takes on the value of one for at least one argument. Otherwise, the fuzzy set is said to be *subnormal*.

Example 3 If $X = \{1,2,3\}$, $A = 0.1/1 + 0.5/2 + 1/3$ and $B = 0.1/1 + 0.6/2 + 0.9/3$, then A is normal and B is subnormal.

Normally, it is desirable to work with normal fuzzy sets since they may provide for some sort of *context-free* comparability, or a common ground or denominator. However, in many instances we obtain in the course of algorithms or procedures subnormal fuzzy sets. They are then often normalized although, unfortunately, the normalization is not a straightforward solution and should be applied with care after some consideration.

We have now some important concepts of nonfuzzy sets associated with a fuzzy set.

The *support* of a fuzzy set A in X, written supp$A$, is the following (nonfuzzy) set

$$\text{supp}A = \{x \in X : \mu_A(x) > 0\} \tag{17}$$

and, evidently, $\emptyset \subseteq \text{supp}A \subseteq X$.

Example 4 If $X = \{1,2,\ldots,7\}$ and $A = 0.1/3 + 0.5/4 + 0.8/5 + 1/6$, then supp$A = \{3,4,5,6\} \subset \{1,2,\ldots,7\}$.

The *α-cut*, or *α-level set*, of a fuzzy set A in X, written A_α, is defined as the following (nonfuzzy) set

$$A_\alpha = \{x \in X : \mu_A(x) \geq \alpha\}, \quad \text{for each } \alpha \in (0,1] \tag{18}$$

and if $\geq$ in (18) is replaced by $>$, then we have the *strong α-cut*, or *strong α-level set*, of a fuzzy set A in X. In principle, we will use the α-cuts given by (18) if not otherwise specified.

Example 5 If $X = \{1,2,3,4\}$ and $A = 0.1/1 + 0.5/2 + 0.8/3 + 1/4$, then we obtain the following α-cuts

$$\begin{aligned} &A_{0.1} = \{1,2,3,4\} \quad A_{0.5} = \{2,3,4\} \\ &A_{0.8} = \{3,4\} \quad A_1 = \{4\}\,. \end{aligned}$$

The α-cuts have many interesting and relevant properties, and among them one can mention the following one

$$\alpha_1 \leq \alpha_2 \Longleftrightarrow A_{\alpha_1} \subseteq A_{\alpha_2} \,. \tag{19}$$

The α-cuts play an extremely relevant role in both formal analysis and applications as they make it possible to uniquely replace a fuzzy set by a sequence of nonfuzzy sets. We will widely use them in the sequel, and the interested reader is referred for details and properties to any book on fuzzy sets theory as, e. g., Dubois and Prade [28], Klir and Folger [51] or Klir and Yuan [53].

The following theorem, called the *representation theorem* (cf. Negoita and Ralescu [67]), is very relevant both in theoretical analysis and applications.

Theorem 1 *Each fuzzy set A in X can be represented as*

$$A = \sum_{\alpha \in (0,1]} \alpha A_\alpha \,, \tag{20}$$

where A_α is an α-cut of A defined as (19), Σ is in the set-theoretic sense, and αA_α denotes the fuzzy set whose degrees of membership are

$$\mu_{\alpha A_\alpha}(x) = \begin{cases} \alpha & \text{for} \quad x \in A_\alpha \\ 0 & \text{otherwise} \,. \end{cases} \tag{21}$$

The expression (20) is also called the *resolution identity*.

Example 6 Let $X = \{1, 2, \ldots, 10\}$, and $A = 0.1/2 + 0.3/3 + 0.6/4 + 0.8/5 + 1/6 + 0.7/7 + 0.4/8 + 0.2/9$.

Then:

$$\begin{aligned} A = \sum_{\alpha \in (0,1]} \alpha A_\alpha &= 0.1(1/2 + 1/3 + 1/4 + 1/5 + 1/6 \\ &+ 1/7 + 1/8 + 1/9) + 0.3(1/3 + 1/4 + 1/5 \\ &+ 1/6 + 1/7 + 1/8 + 1/9) + 0.6(1/4 + 1/5 \\ &+ 1/6 + 1/7) + 0.7(1/5 + 1/6 + 1/7) \\ &+ 0.8(1/5 + 1/6) + 1(1/6) = 0.1/2 + 0.3/3 \\ &+ 0.6/4 + 0.8/5 + 1/6 + 0.7/7 + 0.4/8 + 0.2/9 \,. \end{aligned}$$

Notice that the very essence of the presentation theorem is that each fuzzy sets can be uniquely represented by a set of its α-cuts.

An important issue, both in theory and application, is to be able to define the *cardinality* of a fuzzy set, i. e. to define how many elements it contains. Unfortunately, this is a difficult problem, and the definitions proposed have been criticized. We will discuss below two of them which are presumably the most widely used.

A *nonfuzzy cardinality* of a fuzzy set $A = \mu_A(x_1)/x_1 + \ldots + \mu_A(x_n)/x_n$, the so-called *sigma-count*, denoted $\Sigma\text{Count}(A)$, is defined as (cf. Zadeh [100,101])

$$\Sigma\text{Count}(A) = \sum_{i=1}^{n} \mu_A(x_i) \,. \tag{22}$$

Example 7 If $A = 1/x_1 + 0.8/x_2 + 0.6/x_3 + 0.2/x_4 + 0/x_5$, then

$$\Sigma\text{Count}(A) = 1 + 0.8 + 0.6 + 0.2 = 2.6 \,.$$

The ΣCount is very simple, and is hence widely used. However, an immediate objection may be that the set is fuzzy but its cardinality is not. A solution in this respect, a *fuzzy cardinality*, was proposed by Zadeh [101], and it is shown below. Unfortunately, it is more complicated than a nonfuzzy cardinality defined by (22).

Let A be a fuzzy set defined in X, and A_α, for each $\alpha \in (0, 1]$, its α-cuts defined by (19). First, Zadeh [101] introduces the $FG\text{Count}(A)$ as the fuzzy integer defined as

$$FG\text{Count}(A) = \{1/0\} \sum_{\alpha \in (0,1]} \alpha/\text{card}(A_\alpha) \tag{23}$$

where $\sum$ is in the set-theoretic [cf. (4)], $\text{card}(A_\alpha)$ is the usual number of elements in A_α, and 1/0 means the integer number 0.

Equivalently, if A is defined in X such that $\text{card}(X) = n$, then for each non-negative integer $i = 0, 1, \ldots, n$, we denote

$$FG\text{Count}(A)_i = \sum_{\alpha \in (0,1]} \{\alpha \colon \text{card}(A_\alpha) \geq i\} \,. \tag{24}$$

Semantically, $FG\text{Count}(A)_i$ is the truth of the proposition *A contains at least i elements*.

Next, Zadeh [101] introduces the $FL\text{Count}(A)$ which is defined as the truth of the proposition, "A contains at most i elements", i. e.

$$FL\text{Count}(A) = \neg[FG\text{Count}(A)] - 1 \,, \tag{25}$$

where $\neg[.]$ is the complement to be defined by (32), and $-$ is the subtraction in the sense of fuzzy numbers (64).

Similarly as in (24), if A is defined in $X = \{1, 2, \ldots, n\}$, then we can denote

$$FL\text{Count}(A)_i = \sup_{\alpha \in (0,1]} \{\alpha \colon \text{card}(A_\alpha) \geq n - i\} \,. \tag{26}$$

Notice that

$$FG\text{Count}(A)_i = 1 - FL\text{Count}(A)_{i+1} \,, \quad \text{for} \;\; i = 1, 2, \ldots, n \tag{27}$$

Finally, Zadeh [101] introduces the $FE\text{Count}(A)$ as

$$FE\text{Count}(A) = FG\text{Count}(A) \cap FL\text{Count}(A) \tag{28}$$

or, similarly as above,

$$FE\text{Count}(A)_i = FG\text{Count}(A)_i \wedge FL\text{Count}(A)_i \,, \quad i = 1, 2, \ldots, n \tag{29}$$

where $\cap$ and $\wedge$ denote the intersection of two fuzzy sets and the minimum operations, respectively, as in (34).

Example 8 For the same fuzzy set as in Example 7, i. e. $A = 1/x_1 + 0.8/x_2 + 0.6/x_3 + 0.2/x_4 + 0/x_5$, we obtain

$$\begin{aligned} FG\text{Count}(A) &= 1/0 + 1/1 + 0.8/2 + 0.6/3 + 0.2/4 \\ &\quad + 0/5 \\ FL\text{Count}(A) &= 0/0 + 0.2/1 + 0.4/2 + 0.6/3 \\ &\quad + 0.2/4 + 0/5 \\ FE\text{Count}(A) &= 0/0 + 0.2/1 + 0.4/2 + 0.6/3 \\ &\quad + 0.2/4 + 0/5 \,. \end{aligned}$$

The above classical definitions of the cardinality of a fuzzy set are widely employed, in particular the nonfuzzy cardinality ΣCount. However, the problem of how to define the cardinality of a fuzzy set is conceptually difficult, and the best source are here Wygalak's [89,90] books.

An important issue, which is widely used in applications, is a *distance* between two fuzzy sets. In practice, normalized distances are clearly more interesting. In the literature [42], and in this book too, the following two basic definitions are used.

Suppose that we have two fuzzy sets, A and B, both defined in $X = \{x_1, \ldots, x_n\}$. Then, we have the following two basic (normalized) distances:

- The *normalized linear* (Hamming) *distance* between A and B in X defined as

$$l(A, B) = \frac{1}{n} \sum_{i=1}^{n} \mid \mu_A(x_i) - \mu_B(x_i) \mid \,. \tag{30}$$

- The *normalized quadratic* (Euclidean) *distance* between A and B in X defined as

$$q(A, B) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} [\mu_A(x_i) - \mu_B(x_i)]^2} \tag{31}$$

Example 9 If $X = \{1, 2, \ldots, 7\}$, $A = 0.7/1 + 0.2/2 + 0.6/4 + 0.5/5 + 1/6$ and $B = 0.2/1 + 0.6/4 + 0.8/5 + 1/7$, then:

$$l(A, B) = 0.37 \qquad q(A, B) = 0.49 \,.$$

Now we will proceed to the basic set-theoretic and algebraic operations on fuzzy sets. They are clearly crucial for both theoretical analysis and applications.

Basic Operations on Fuzzy Sets

Similarly as in the conventional (nonfuzzy) set theory, the basic operations in fuzzy set theory are also the complement, intersection and union which will be defined below.

The *complement* of a fuzzy set A in X, written $\neg A$, is defined as

$$\mu_{\neg A}(x) = 1 - \mu_A(x) \,, \quad \text{for each } x \in X \tag{32}$$

and the complement corresponds to the negation, *not*.

Example 10 If $X = \{1, 2, 3\}$ and $A = 0.1/1 + 0.7/2 + 1/3$, then $\neg A = 0.9/1 + 0.3/2$.

The idea of the complement can be portrayed as in Fig. 3.

This definition is the most widely used due to its simplicity and some important mathematical properties. Sometimes different definitions can be justified and useful in specific contexts as, for instance when $X = [0, 1]$, we have the following

$$\mu_{\neg A}(x) = \mu_A(1 - x) \,, \quad \text{for each } x \in [0, 1] \,. \tag{33}$$

The *intersection* of two fuzzy sets A and B in X, written $A \cap B$, is defined as

$$\mu_{A \cap B}(x) = \mu_A(x) \wedge \mu_B(x) \,, \quad \text{for each } x \in X \tag{34}$$

where $\wedge$ is the minimum operation, i. e. $a \wedge b = \min(a, b)$; the intersection of two fuzzy sets corresponds to the connective *and*

Example 11 If $X = \{1, 2, 3, 4\}$, and $A = 0.2/1 + 0.5/2 + 0.8/3 + 1/4$ and $B = 1/1 + 0.8/2 + 0.5/3 + 0.2/4$, then we obtain by (34) $A \cap B = 0.2/1 + 0.5/2 + 0.5/3 + 0.2/4$.

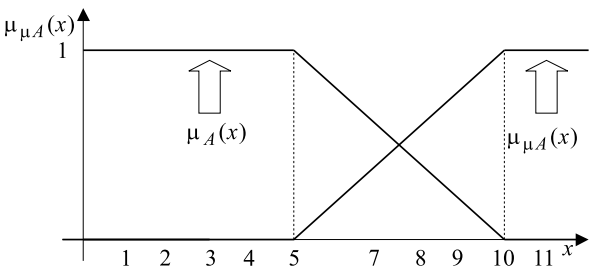

Fuzzy Sets Theory, Foundations of, Figure 3
The complement of a fuzzy set

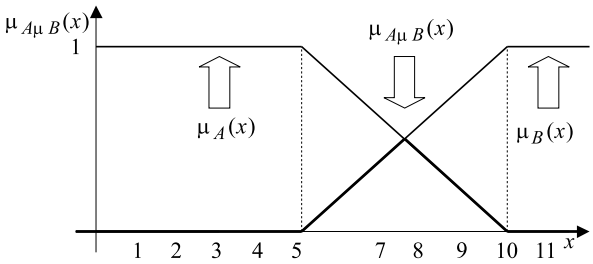

Fuzzy Sets Theory, Foundations of, Figure 4
Membership function of a fuzzy set, *integer numbers which are more or less six*

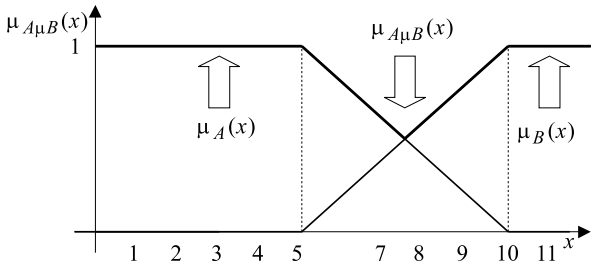

Fuzzy Sets Theory, Foundations of, Figure 5
The union of two fuzzy sets

The intersection can be illustrated as in Fig. 4 where $\mu_{A\cap B}(x)$ is shown in bold line.

The *union* of two fuzzy sets A and B in X, written $A + B$, is defined as

$$\mu_{A+B}(x) = \mu_A(x) \vee \mu_B(x)\,, \quad \text{for each } x \in X \tag{35}$$

where $\vee$ is the maximum operation, i. e. $a \vee b = \max(a, b)$; the union of two fuzzy sets corresponds to the connective "or".

Example 12 If $X = \{1, 2, 3, 4\}$, and $A = 0.2/1 + 0.5/2 + 0.8/3 + 1/4$ and $B = 1/1 + 0.8/2 + 0.5/3 + 0.2/4$, then $A + B = 1/1 + 0.8/2 + 0.8/3 + 1/4$.

The union can be portrayed as in Fig. 5 in which $\mu_{A+B}(x)$ is shown in bold line.

The above definitions of the basic operations are widely used, and justified. However, other definitions are often employed too. In particular, for the intersection and union the t-norms and s-norms are popular.

A *t-norm* is defined as

$$t\colon [0, 1] \times [0, 1] \longrightarrow [0, 1] \tag{36}$$

such that, for each $a, b, c \in [0, 1]$:

1. It has 1 as the unit element, i. e. $t(a, 1) = a$,
2. It is monotone, i. e. $a \le b \Longrightarrow t(a, c) \le t(b, c)$,
3. It is commutative, i. e. $t(a, b) = t(b, a)$, and
4. It is associative, i. e. $t[a, t(b, c)] = t[t(a, b), c]$.

Some more relevant examples of t-norms are:

- The minimum (which is the most widely used)

$$t(a, b) = a \wedge b = \min(a, b)\,, \tag{37}$$

- the algebraic product

$$t(a, b) = a \cdot b\,, \tag{38}$$

- the Łukasiewicz t-norm

$$t(a, b) = \max(0, a + b - 1) \tag{39}$$

and notice that we have written above both $t(a, b)$ and atb.

An *s-norm* (or a *t-conform*) is defined as

$$s\colon [0, 1,] \times [0, 1] \longrightarrow [0, 1] \tag{40}$$

such that, for each $a, b, c \in [0, 1]$:

1. It has 0 as the unit element, i. e. $s(a, 0) = a$,
2. it is monotone, i. e. $a \le b \Longrightarrow s(a, c) \le s(b, c)$,
3. it is commutative, i. e. $s(a, b) = s(b, a)$, and
4. it is associative, i. e. $s[a, s(b, c)] = s[s(a, b), c]$.

Some more relevant examples of s-norms are:

- The maximum (which is the most widely used)

$$s(a, b) = a \vee b = \max(a, b)\,, \tag{41}$$

- the probabilistic product

$$s(a, b) = a + b - ab\,, \tag{42}$$

- the Łukasiewicz s-norm

$$s(a, b) = \min(a + b, 1)\,. \tag{43}$$

Notice that a t-norms is *dual* to an s-norms in that $s(a, b) = 1 - t(1 - a, 1 - b)$.

The t-norms and s-norms are very important for fuzzy sets theory, and among a multitude of works dealing with them, the most complete reference is provided by Klement, Mesiar and Pap's [49] book.

Notice that while defining the basic concepts and operations on fuzzy sets we have always referred to their linguistic sense. This is very important because without such a linguistic connection it would have been very difficult to provide semantics of both the very concept of a fuzzy set and the operations. This is very relevant and has great implications for the development of fuzzy sets related tools and techniques as we will see later.

As to some other operations on fuzzy sets that may be of use in this book, one should also mention the following ones.

The *product of a scalar* $a \in R$ *and a fuzzy set* A in X, written aA, is defined as

$$\mu_{aA}(x) = a\mu_A(x)\,, \quad \text{for each } x \in X \tag{44}$$

where, by necessity, $0 \le a \le 1/\mu_A(x)$, for each $x \in X$.

The *kth power* of a fuzzy set A in X, written A^k, is defined as

$$\mu_{A^k}(x) = [\mu_A(x)]^k\,, \quad \text{for each } x \in X\,, \tag{45}$$

where $k \in R$ and, evidently, $0 \le [\mu_A(x)]^k \le 1$.

An important issue is the *adequacy* of the operations on fuzzy sets, i. e. whether they do reflect the real human perception of their essence, i. e. whether they really reflect the semantics of "not," "and," "or," etc. Diverse approaches have been used to find and justify a particular definition. These approaches may be classified as:

- *Intuitive* as the original Zadeh's [95,97] works in which it is shown by a rational argument that the operations defined are proper,
- *axiomatic* whose line of reasoning is to assume some set of plausible conditions to be fulfilled, and then to show using some analytic tools that definitions assumed are the only possible ones; this may be exemplified by Bellman and Giertz [7],
- *experimental* whose essence is to devise some psychological tests for a group of certain individuals, and then use the responses to find which operation is best justified; this may be exemplified by Zimmermann and Zysno [111].

Now we will present the concept of a fuzzy relation which is, as its nonfuzzy counterpart, crucial for the theory and applications.

Fuzzy Relations

The concept of a relation is crucial for virtually all areas of mathematics and its applications, and the same holds true for fuzzy sets theory and its applications.

A *fuzzy relation* R between two (nonfuzzy) sets $X = \{x\}$ and $Y = \{y\}$ is defined as a fuzzy set in the Cartesian product $X \times Y$, i. e.

$$R = \{(\mu_R(x,y),(x,y))\} = \{\mu_R(x,y)/(x,y)\}\,, \quad \text{for each } (x,y) \in X \times Y \tag{46}$$

where $\mu_R(x,y)\colon X \times Y \longrightarrow [0,1]$ is the membership function of the fuzzy relation R, and $\mu_R(x,y) \in [0,1]$ gives the degree to which the elements $x \in X$ and $y \in Y$ are between each other in relation R.

The above fuzzy relation is defined in the Cartesian product of (nonfuzzy!) two sets, X and Y, and is called a binary fuzzy relation. In general, a fuzzy relation may be defined in the Cartesian product of k sets, $X_1 \times \ldots \times X_k$, and is then called a k-ary fuzzy relation. In this perspective, a fuzzy set is an unary fuzzy relation.

Example 13 If $X = \{\text{horse, donkey}\}$ and $Y = \{\text{mule, cow}\}$, then the fuzzy relation R labeled *similarity* may be exemplified by

$$\begin{aligned} R = \textit{similarity} = {} & 0.8/(\text{horse, mule}) + 0.4/(\text{horse, cow}) \\ & + 0.9/(\text{donkey, mule}) \\ & + 0.2/(\text{donkey, cow}) \end{aligned}$$

to be read as: The horse and the mule are similar (with respect to *our own* subjective aspects!) to degree 0.8, i. e. to a very high extent, the horse and the cow are similar to degree 0.4, i. e. to quite a low extent, etc.

It may easily be seen that the concept of a fuzzy relation makes it possible to express an imprecise, or imprecisely specified, relationship between elements of some sets, as opposed to a precise and abrupt one in the case of a nonfuzzy relation in which any two elements can either be or not in relation.

A fuzzy relation R in $X \times Y$ for X and Y of a sufficiently low dimensionality may be conveniently represented in matrix form exemplified, for the fuzzy relation $R = \textit{similarity}$ in Example 13, by

$R = \textit{similarity} =$

	$y =$ mule	cow
$x =$ horse	0.8	0.4
donkey	0.9	0.2

Since a fuzzy relation is a fuzzy set, all the definitions, properties, operations, etc. on fuzzy sets presented in Sects. "The Idea of a Fuzzy Set" – "Basic Operations on Fuzzy Sets" hold as well, and we will concentrate below on those more specific ones.

The *max-min composition* of two fuzzy relations R in $X \times Y$ and S in $Y \times Z$, written $R \circ_{\max-\min} S$ is defined as a fuzzy relation in $X \times Z$ such that

$$\mu_{R \circ_{\max-\min} S}(x,y) = \max_{y \in Y}[\mu_R(x,y) \wedge \mu_S(y,z)]\,, \quad \text{for each } x \in X\,,\ z \in Z \tag{47}$$

and since this type of composition will be used throughout this paper, if not otherwise specified, then it will be briefly denoted as $R \circ S$.

Example 14 If $X = \{1, 2\}$, $Y = \{1, 2, 3\}$ and $Z = \{1, 2, 3, 4\}$, and the fuzzy relations R and S are as below. Its resulting max-min composition, $R \circ S$, is then:

$R \circ S =$

	$y = 1$	2	3
$x = 1$	0.3	0.8	1
2	0.9	0.7	0.4

$\circ$

	$z = 1$	2	3	4
$y = 1$	0.7	0.6	0.4	0.1
2	0.4	1	0.7	0.2
3	0.5	0.9	0.6	0.8

$=$

	$z = 1$	2	3	4
$x = 1$	0.5	0.9	0.7	0.8
2	0.7	0.7	0.7	0.4

This max-min composition of fuzzy relations is the original Zadeh's definition (cf. Zadeh [97]), and is certainly the most widely used. However, if we notice that the two basic operations involved in the definition of their composition, i. e. *min* ($\wedge$) and *max* ($\vee$) are just specific examples of the t-norm and s-norm (t-conorm) discussed in Sect. "Basic Operations on Fuzzy Sets", then one can well define a much more general type of composition given below.

The $s - t$-*norm composition* of two fuzzy relations R in $X \times Y$ and S in $Y \times Z$, written $R \circ_{s-t} S$, is defined as a fuzzy relation in $X \times Z$ such that

$$\mu_{R \circ_{s-t} S}(x, z) = s_{y \in Y}[\mu_R(x, y) t \mu_S(y, z)] \,, \quad \text{for each } x \in X \,, \; z \in Z \,. \tag{48}$$

Fuzzy relations, similar to their nonfuzzy counterparts, play a crucial role in virtually all aspects of the theory and application of fuzzy sets, notably in tule based fuzzy modeling discussed later in this paper. An important issue is related to so-called fuzzy relational equations. We will not deal with this due to lack of space, and we refer the interested reader to, for instance, the recent book by Peeva and Kyosev [72].

Finally, let us mention two concepts concerning the fuzzy sets that are related to fuzzy relations.

The *Cartesian product* of two fuzzy sets A in X and B in Y, written $A \times B$, is defined as a fuzzy set in $X \times Y$ such that

$$\mu_{A \times B}(x, y) = [\mu_A(x) \wedge \mu_B(y)] \,, \quad \text{for each } x \in X \,, \; y \in Y \,. \tag{49}$$

A fuzzy relation R in $X \times Y \times \ldots \times Z$ is said to be *decomposable* if and only if it can be represented as

$$\mu_R(x, y, \ldots, z) = \mu_{R_x}(x) \wedge \mu_{R_y}(y) \wedge \ldots \wedge \mu_{R_z}(z) \,, \quad \text{for each } x \in X \,, \; y \in Y \,, \ldots \,, \; z \in Z \,, \tag{50}$$

where $\mu_{R_x}(x), \mu_{R_y}(y), \ldots, \mu_{R_z}(z)$ are projections of the fuzzy relation $\mu_R(x, y)$ on $X, Y, \ldots, Z$, respectively, defined as

$$\mu_{R_x}(x) = \sup_{\{y, \ldots, z\} \in Y \times Z} \mu_R(x, y, \ldots, z) \,, \quad \text{for each } x \in X \tag{51}$$

Fuzzy relations play a major role in fuzzy sets theory and its applications, and will also be relevant for some of our next considerations.

Linguistic Variable, Fuzzy Conditional Statement, and Compositional Rule of Inference

We have already mentioned that while defining various concepts and properties of fuzzy sets it is expedient to use natural language because linguistic terms and descriptions best provide semantics. Now we will briefly present the essence of Zadeh's [97] *linguistic approach* in which this fact has been exploited to an even greater extent. This approach has inspired and triggered many new areas of research, notably fuzzy control which has resulted in so many real-world applications in diverse areas and has been a decisive factor in the so-called *fuzzy boom* which was started in the mid-1980s by the launching of a multitude of domestic appliances and professional products exemplified by fuzzy logic-controlled washing machines, cameras, automobile automatic transmissions, cranes, subway trains, etc. These applications have been decisive for a wide acceptance of fuzzy logic as a powerful and potentially useful tool.

Basically, the rationale behind Zadeh's [97] linguistic approach to the analysis of complex systems and decision (and control) processes is that the basic element is a *linguistic variable* exemplified by "temperature" which takes on as their values not conventional numerical values as, e. g., 150°C, but linguistic values as, e. g., "high," "low," etc. that are in turn equated semantically with some fuzzy sets. Notice that such linguistic values are common in human discourse as natural language is the only fully natural human means of communication. Clearly, one can then form more complex linguistic expressions as, e. g., "not very low and not very high," "more or less medium," etc. by using some connectives (e. g., and, or, ...), modifiers (e. g., more or less, very, ...), etc. (see also Sect. "Fuzzy Logic – Basic Issues"), and employing a syntactic analysis.

To represent a relationship between linguistic variables, *fuzzy conditional statements* are employed. For instance, if we have two linguistic variables, a primary one L and a secondary one K, such that the value of L is a fuzzy set A in X, and the value of K is a fuzzy set B in Y, then

a relationship between L and K, in terms of their values A and B, respectively, may be written as

$$\text{IF} \quad L = A \quad \text{THEN} \quad K = B \tag{52}$$

or, shortly

$$\text{IF} \quad A \quad \text{THEN} \quad B\,. \tag{53}$$

This fuzzy conditional statement is now assumed to be equivalent to

$$\text{IF} \quad A \quad \text{THEN} \quad B = A \times B \tag{54}$$

i. e. to the Cartesian product (49) of the two fuzzy sets A and B which is in turn a fuzzy relation in $X \times Y$.

Notice that this is what Mamdani [61] used in his controller, and which is often called Mamdani's implication (though it is not an implication!). This definition is the simplest one, and we can devise more sophisticated ones by using, e. g., various definitions of implication (80)–(85).

It is easy to see that the fuzzy conditional statement (53) may be extended to account for multiple values of A and B obtaining, if we use (54),

$$\begin{aligned} &\text{IF} \quad A_1 \quad \text{THEN} \quad B_1 \quad \text{ELSE} \quad \ldots \quad \text{ELSE} \\ &\text{IF} \quad A_n \quad \text{THEN} \quad B_n = A_1 \times B_1 + \ldots + A_n \times B_n\,, \end{aligned} \tag{55}$$

where A_i's are fuzzy sets in X and B_i's are fuzzy sets in Y, $i = 1, \ldots, n$.

Notice that in (53) and (55) we only specify what happens if the primary variable takes on some value. It is often, however, also relevant to explore what happens if that value is not taken. In such a case (53) becomes

$$\text{IF} \quad A \quad \text{THEN} \quad B \quad \text{ELSE} \quad C \tag{56}$$

and it is represented as

$$\text{IF} \quad A \quad \text{THEN} \quad B \quad \text{ELSE} \quad C = A \times B + \neg A \times C\,. \tag{57}$$

Evidently, one can generalize the above fuzzy conditional statements to involve more than one primary variable. For details, we refer the reader to, e. g., Zadeh [97].

We therefore have some tool to represent a relation between a primary and secondary variable that is represented by some fuzzy relation. An immediate question is then:

> If the primary variable takes on some fuzzy value, say A', and we have a (fuzzy) relation, IF A THEN B, then what will be the implied (inferred) value of the secondary variable B'?

This may be represented by the inference scheme

$$\begin{array}{l} A' \\ \text{IF} \quad A \quad \text{THEN} \quad B \\ \hline \qquad\qquad\qquad B' = \,? \end{array} \tag{58}$$

and, what is the very essence, the fuzzy values A' and A need not be the same (notice that this prohibits the use of conventional logical inference tools).

The answer to the question (58) is provided by the *compositional rule of inference* which states that if R in $X \times Y$ is a fuzzy relation representing a dependence between a primary and secondary variable, represented by a fuzzy conditional statement, and the primary variable takes on a fuzzy value A' in X, then the implied fuzzy value of the secondary variable B' in Y is given by the (max-min) composition (47) of A' and R, i. e.

$$\mu_{B'}(y) = \max_{x \in X}[\mu_A(x) \wedge \mu_R(x, y)], \quad \text{for each} \quad y \in Y \tag{59}$$

and notice that A' is here considered to be a unary fuzzy relation as mentioned in Sect. "Fuzzy Relations".

Example 15 Suppose that $X = \{x\} = \{1, 2, 3\}$ and $Y = \{y\} = \{1, 2, 3, 4\}$, and the fuzzy conditional statement representing the dependence between L and K is

$$\text{IF} \quad L = low \quad \text{THEN} \quad K = high = (low) \circ (high)$$

where

$$\begin{aligned} low &= 1/1 + 0.7/2 + 0.3/3 \\ high &= 0.2/1 + 0.5/2 + 0.8/3 + 1/4 \end{aligned}$$

and is equivalent to the following fuzzy relation

$R = (low) \circ (high) =$

	$y = 1$	2	3	4
$x = 1$	0.2	0.5	0.8	1
2	0.2	0.5	0.7	0.7
3	0.2	0.3	0.3	0.3

If now $L = medium = 0.5/1 + 1/2 + 0.5/3$, then the K induced is given by

$$\begin{aligned} K = (medium) \circ R &= \max_{x \in \{1,2,3\}} [\mu_L(x) \wedge \mu_R(x, y)] \\ &= 0.2/1 + 0.5/2 + 0.7/3 + 0.7/4 \end{aligned}$$

The fuzzy conditional statements may be used to represent simple dependencies and relations between linguistic variables. For more complicated dependencies and relations, fuzzy algorithms may be used (cf. Zadeh [97]) which will not be dealt with here.

A further extension in the spirit of the linguistic approach presented in this section is Zadeh's *computing with words* or, more generally, *computing with words and perceptions*. We will not deal with this and will refer the reader to a comprehensive coverage of main theoretical issues, and many applications, related to this approach which is given in Zadeh and Kacprzyk [105,106].

The Extension Principle

Now we will briefly present the essence of Zadeh's classic *extension principle* (cf. Zadeh [97]) which is one of the most important and powerful tools in fuzzy sets theory. The extension principle addresses the following fundamental issue:

> If there is some relationship (e. g., a function) between *conventional* (nonfuzzy) entities (e. g., variables taking on nonfuzzy values, then what is its equivalent relationship between fuzzy entities (e. g., variables taking on fuzzy values)?

The extension principle makes it therefore possible, for instance, to extend some known conventional models, algorithms, etc. involving nonfuzzy variables to the case of fuzzy variables.

Let $A_1, \ldots, A_n$ be fuzzy sets in $X_1 = \{x_1\}, \ldots, X_n = \{x_n\}$, respectively, and

$$f\colon X_1 \times \ldots \times X_n \longrightarrow Y \tag{60}$$

be some (nonfuzzy) function such that $y = f(x_1, \ldots, x_n)$.

Then, according to the *extension principle*, the fuzzy set B in $Y = \{y\}$ induced by the fuzzy sets $A_1, \ldots, A_n$ via the function f (60) is

$$\mu_B(y) = \max_{(x_1,\ldots,x_n)\in X_1\times\ldots\times X_n\colon\, y=f(x_1,\ldots,x_n)} \bigwedge_{i=1}^{n} \mu_{A_i}(x_i)\,. \tag{61}$$

Example 16 Suppose that: $X_1 = \{1, 2, 3\}$, $X_2 = \{1, 2, 3, 4\}$, f represents the addition, i. e. $y = x_1 + x_2$, $A_1 = 0.1/1 + 0.6/2 + 1/3$ and $A_2 = 0.6/1 + 1/2 + 0.5/3 + 0.1/4$, then

$$\begin{aligned} B &= A_1 + A_2 \\ &= 0.1/2 + 0.6/3 + 0.6/4 + 1/5 + 0.5/6 + 0.1/7 \end{aligned}$$

and notice that $+$ is used here in both the arithmetic (the sum of real and fuzzy numbers – cf. Sect. "Fuzzy Numbers") and set-theoretic sense which should not lead to confusion.

Equivalently, the extension principle (61) may also be written in terms of the α-cuts (18). Namely, suppose for simplicity that $f\colon X \longrightarrow Y$, $X = \{x\}$, $Y = \{y\}$, and A_α, for each $\alpha \in (0, 1]$, are α-cuts of A. Then, the fuzzy set B in Y, induced by A via the extension principle is given as

$$B = f(A) = f\left(\sum_{\alpha\in(0,1]} \alpha \cdot A_\alpha\right) = \sum_{\alpha} \alpha f(A_\alpha) \tag{62}$$

which is clearly implied by the representation theorem (Theorem 1).

Notice that the extension principle plays an extraordinary role in the sense that we have a multitude of nonfuzzy tools like algorithms, procedures, etc. which have been widely used to solve various problems. However, they need precise information, notably real or integer numbers, real intervals, etc. We have neither fuzzy computers nor fuzzy tools of the type mentioned above so that we are not in a position to directly use imprecise information to solve our problems. However, by virtu of the extension principle we can extend our known nonfuzzy tools and techniques (their related algorithms and procedures) to their fuzzy counterparts.

Fuzzy Numbers

The same fundamental role as played by nonfuzzy (real, integer, …) numbers in conventional models, the fuzzy numbers play in fuzzy models.

A *fuzzy number* is defined as a fuzzy set in R, the real line. Usually, but not always, it is assumed to be a normal and convex fuzzy set. For example, the membership function of a fuzzy number that is *more or less six* may be as shown in Fig. 6, i. e. as a bell-shaped function.

For our purposes, operations on fuzzy numbers are the most relevant. It is easy to notice that their definitions may readily be obtained by applying the extension principle (61).

Suppose therefore that A and B are two fuzzy numbers in $R = \{x\}$ characterized by their membership functions

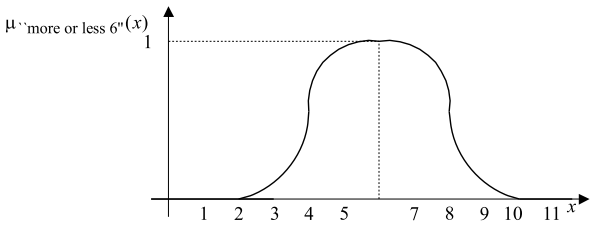

Fuzzy Sets Theory, Foundations of, Figure 6
The membership function of a fuzzy number *more or less six*

$\mu_A(x)$ and $\mu_B(x)$, respectively. Then the extension principle yields the following definitions of the four basic *arithmetic operations* on fuzzy numbers:

- *Addition*

$$\mu_{A+B}(z) = \max_{x+y=z} [\mu_A(x) \wedge \mu_B(y)], \quad \text{for each } z \in R, \tag{63}$$

- *Subtraction*

$$\mu_{A-B}(z) = \max_{x-y=z} [\mu_A(x) \wedge \mu_B(y)], \quad \text{for each } z \in R, \tag{64}$$

- *Multiplication*

$$\mu_{A \cdot B}(z) = \max_{x \cdot y=z} [\mu_A(x) \wedge \mu_B(y)], \quad \text{for each } z \in R, \tag{65}$$

- *Division*

$$\mu_{A/B}(z) = \max_{x/y=z, y \neq 0} [\mu_A(x) \wedge \mu_B(y)], \quad \text{for each } z \in R. \tag{66}$$

In some applications, the following one-argument operations on fuzzy numbers may also be of use:

- The *opposite* of a fuzzy number

$$\mu_{-A}(x) = \mu_A(-x), \quad \text{for each } x \in R, \tag{67}$$

- the *inverse* of a fuzzy number

$$\mu_{A^{-1}}(x) = \mu_A\left(\frac{1}{x}\right), \quad \text{for each } x \in R \setminus \{0\}. \tag{68}$$

In practice, however, such a general definition of fuzzy numbers and operations on them is seldom used. Normally a further simplification is made, namely the fuzzy numbers are assumed to be *triangular* and eventually *trapezoid* fuzzy numbers whose membership functions are sketched in Fig. 7.

For triangular and trapezoid fuzzy numbers, for the four basic operations, i. e. the addition, subtraction, multiplication and division, special formulas can be devised whose calculation is simpler than that of (63)–(66). Moreover, a crucial problem of comparison of two fuzzy numbers may also be simplified. For details, we will refer the reader to a rich literature exemplified by Dubois and Prade [28], Kaufmann and Gupta [48], Klir and Folger [51], Klir and Yuan [53], Maresš [62], Hanss [41], etc.

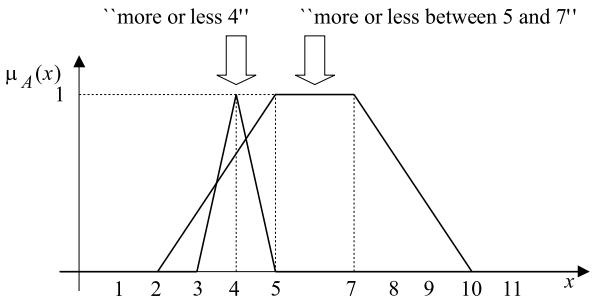

Fuzzy Sets Theory, Foundations of, Figure 7
Triangular (*about* four) and trapezoid (*more or less* between six and seven) fuzzy numbers

Fuzzy Events and Their Probabilities

Fuzziness and randomness are meant, in our perspective, as two aspects of *imperfect information*. Fuzziness is meant to concern entities and relations which are not crisply defined, with gradual transition between the elements belonging and not belonging to a class. Randomness concerns situations in which the event is well defined but it occurrence is uncertain.

However, in practice and in everyday discourse there is an abundance of situations in which there jointly occur fuzziness and randomness, for instance, when we ask about the probability of *cold weather* tomorrow or of a *high inflation* in the next year, we have an imprecise (fuzzy) event – cold weather and high inflation, respectively. To be able to formally deal with such problems, we need a concept of a fuzzy event, and that of a probability of a fuzzy event.

The first approach in this respect is due to Zadeh [96]. Its point of departure is the concept of a *fuzzy event* which is simply a fuzzy set A in $X = \{x\} = \{x_1, \ldots, x_n\}$ whose membership function is Borel measurable. We assume that the probabilities of the (nonfuzzy) elementary events $x_1, \ldots, x_n \in X$ are known and equal to $p(x_1), \ldots, p(x_n) \in [0, 1]$, respectively, with $p(x_1) + \ldots + p(x_n) = 1$.

As to some more important concepts related to fuzzy events, the following may be stated.

Two fuzzy events A and B in X are *independent* if and only if

$$p(AB) = p(A)p(B). \tag{69}$$

The *conditional probability* of a fuzzy event A in X with respect to a fuzzy event B in X is denoted $p(A \mid B)$ and defined as

$$p(A \mid B) = \frac{p(AB)}{p(B)}, \quad p(B) > 0 \tag{70}$$

and if the fuzzy events A and A are independent, then

$$p(A \mid B) = p(A) \tag{71}$$

Notice that both of the above concepts are analogous to their nonfuzzy counterparts.

The (*nonfuzzy*) *probability of a fuzzy event* A in $X = \{x_1, \ldots, x_n\}$ is denoted $p(A)$ and defined by Zadeh [96] as

$$p(A) = \sum_{i=1}^{n} \mu_A(x_i) p(x_i) \tag{72}$$

i. e. as the expected value of the membership function of A, $\mu_A(x)$.

Example 17 Suppose that $X = \{1, 2, \ldots, 5\}$, $p(x_1) = 0.1$, $p(x_2) = 0.1$, $p(x_3) = 0.1$, $p(x_4) = 0.3$, $p(x_5) = 0.4$, and $A = 0.1/2 + 0.5/3 + 0.7/4 + 0.9/5$.

Then

$$p(A) = 0.1 \times 0.1 + 0.1 \times 0.5 + 0.3 \times 0.7 + 0.4 \times 0.9 = 0.73.$$

Notice that Zadeh's [96] (nonfuzzy) probability of a fuzzy event (72) satisfies:

1. $p(\emptyset) = 0$,
2. $p(\neg A) = 1 - p(A)$,
3. $p(A + B) = p(A) + p(B) - p(A \cap B)$,
4. and

$$\begin{aligned} p\left(\sum_{i=1}^{r} A_i\right) = & \sum_{i=1}^{r} p(A_i) - \sum_{j=1}^{r} \sum_{k=1, k<j}^{r} p(A_j \cap A_k) \\ & + \sum_{j=1}^{r} \sum_{k=1, k<j}^{r} \sum_{l=1, l<k}^{r} p(A_j \cap A_k \cap A_l) \\ & + \ldots + (-1)^{r+1} p(A_1 \cap A_2 \cap \ldots \cap A_r) \end{aligned}$$

so it does make sense to term the expression (72) a "probability."

The above Zadeh' [96] classic definition of a (nonfuzzy) probability of a fuzzy event is by far the most popular and most widely used. However, though the event is fuzzy, its probability is nonfuzzy, i. e. is a real number from the unit interval. This may be viewed counter-intuitive but provides simplicity. For some approaches to a fuzzy probability of a fuzzy event, see, e. g., Klir and Folger [51] or Klir and Yuan [53].

Defuzzification of Fuzzy Sets

In many applications we arrive at a fuzzy result. However, in it is a crisp (non-fuzzy) result that should be applied. A notable example is fuzzy control (cf. Driankov, Hellendoorn and Reinfrank [27] or Kacprzyk [44].

Suppose that we have a fuzzy set A defined in $X = \{x_1, x_2, \ldots, x_n\}$, i. e. $A = \mu_A(x_1)/x_1 + \mu_A(x_2)/x_2 + \ldots + \mu_A(x_n)/x_n$. We need to find a crisp number $a \in [x_1, x_n]$ which *best* represents A. Notice that we assume here that A is defined in a finite universe of discourse, but its corresponding defuzzified number a need not be in general any of the finite values of X but should be between the lowest and highest elements of X (evidently, this requires some ordering of x_i's but this is clearly satisfied as x_i's are in virtually all practical cases just real numbers).

The most commonly used defuzzification procedure is certainly the *center-of-area*, also called the *center-of-gravity*, method whose essence is

$$a = \frac{\sum_{i=1}^{n} x_i \mu_A(x_i)}{\sum_{i=1}^{n} \mu_A(x_i)} . \tag{73}$$

The above defuzzification (73) is however often too complex if our analysis involves, e. g., some optimization (cf. Kacprzyk [44]). In such a case one needs to resort to an even simpler defuzzification method which simply assumes that the defuzzified value of a fuzzy value is $x_i \in X = \{x_1, \ldots, x_n\}$ for which $\mu_A(x)$ takes on its maximum values, i. e.

$$\mu_A(a) = \max_{x_i \in X} \mu_A(x) \tag{74}$$

with an obvious extension that if the A determined in (74) is not unique, then we take, say, the mean value of such equivalent a's.

In Sect. "Bellman and Zadeh's General Approach to Decision Making Under Fuzziness" we will provide a justification of the above maximizing-value-type defuzzification procedure in the framework of decision making.

There is a whole array of other defuzzification procedures, and the reader is referred to, e. g., Driankov, Hellendorn and Reinfrank [27], Kacprzyk [44], Klir and Yuan [53] or Yager and Filev [92].

Fuzzy Logic – Basic Issues

The concept of a *fuzzy logic* is not uniquely understood. Basically, it may be meant in (at least) the three following ways:

- As a foundation of reasoning based on ambiguous, vague and imprecise statements (cf. Goguen [38]),
- as a foundation of reasoning based on ambiguous, vague and imprecise statements in which fuzzy set-theoretic tools are used (see, e. g., Zadeh [98]; or Zadeh and Kacprzyk [104]), and

- as a multivalued logic with truth values in the unit interval in which the logical operations of negation, union, intersection, implication, equivalence, etc. are chosen in a special way, and have some fuzzy interpretation (cf. Hájek [40] or Nová, Perfilieva and Močkoř [71]).

It is easy to notice that the meaning of fuzzy logic in the first and second sense is similar, though the generality of the former is clearly higher, while its meaning in the third way is different.

For the purposes of this paper, we will assume the third view on fuzzy logic, and will present a very limited survey of basic issues. The interested reader is referred for more detail on various aspects of fuzzy logic to Zadeh and Kacprzyk's [104] volume which is practically the only up-to-date and exhaustive treatise on fuzzy logic available today.

Notice what some authors mean by fuzzy logic is the whole theory of fuzzy sets and related topics. We will not follow this line of reasoning, and view fuzzy sets theory as more *set-theoretic* while fuzzy logic as more *logical*.

Suppose that we have a statement (predicate) u is P denoted, for brevity, as P and exemplified by temperature (u) is high (P), where u is a variable taking on its values in a universe of discourse $U = \{u\}$, and P is an imprecise term equated with a fuzzy set in U, $P = \{\mu_P(u)/u\}$.

For a specified value $u \in U$ the truth of u is P (or of P) is denoted $\tau(P)$ and meant to be $\tau(u \text{ is } P) = \tau(P) = \mu_P(u)$, for each $u \in U$.

The following general definitions of basic logical operations (in terms of their respective truth values) are usually employed:

- The *negation* of P, i. e. *not* P, denoted by $\neg P$:

$$\tau(\neg P) = 1 - \tau(P) \,, \tag{75}$$

- the *intersection* of P and Q, i. e. P *and* Q, denoted by $P \cap Q$:

$$\tau(P \cap Q) = t[\tau(P), \tau(Q)] \tag{76}$$

where $t\colon [0,1] \times [0,1] \longrightarrow [0,1]$ is a t-norm (36); the original Zadeh's definition is

$$\tau(P \cap Q) = \tau(P) \wedge \tau(Q) = \min[\tau(P), \tau(Q)] \,, \tag{77}$$

- the *union* of P and Q, i. e. P *or* Q, denoted by $P \cup Q$:

$$\tau(P \cup Q) = s[\tau(P), \tau(Q)] \tag{78}$$

where $s\colon [0,1] \times [0,1] \longrightarrow [0,1]$ is an s-norm (40); the original Zadeh's definition is

$$\tau(P \cup Q) = \tau(P) \vee \tau(Q) = \max[\tau(P), \tau(Q)] \,, \tag{79}$$

- the implication, i. e. if P then Q, denoted by $P \Longrightarrow Q$, which may be defined as, e. g.:
  1. The *Łukasiewicz implication*

$$\tau(P \Longrightarrow Q) = \min\{1 - \tau(P) + \tau(Q), 1\} \,, \tag{80}$$

  2. the *Gödel implication*

$$\tau(P \Longrightarrow Q) = \begin{cases} 1 & \text{if } \tau(P) \le \tau(Q) \\ \tau(Q) & \text{otherwise} \,, \end{cases} \tag{81}$$

  3. the *Goguen implication*

$$\tau(P \Longrightarrow Q) = \begin{cases} 1 & \text{if } \tau(P) = 0 \\ \min\left\{1, \frac{\tau(Q)}{\tau(P)}\right\} & \text{otherwise} \,, \end{cases} \tag{82}$$

  4. the *Kleene–Dienes implication*

$$\tau(P \Longrightarrow Q) = \max\{1 - \tau(P), \tau(Q)\} \,, \tag{83}$$

  5. the *Zadeh implication*

$$\tau(P \Longrightarrow Q) = \max\{1 - \tau(P), \min\{\tau(P), \tau(Q)\}\} \,, \tag{84}$$

  6. the *Reichenbach implication*

$$\tau(P \Longrightarrow Q) = 1 - \tau(P) + \tau(P) \cdot \tau(Q) \,, \tag{85}$$

- The *equivalence*, i. e. P is equivalent to Q or if P then Q and if Q then P, denoted by $P \Longleftrightarrow Q$:

$$\tau(P \Longleftrightarrow Q) = \tau[(P \Longrightarrow Q) \cap (Q \Longrightarrow P)] \tag{86}$$

and we can assume an appropriate definition of the intersection $\cap$ (34), and the implication $\Longrightarrow$ (80)–(85).

Among some other important aspects of fuzzy logic, one should mention the use of (fuzzy) linguistic quantifiers, exemplified by *most*, *almost all*, *a few*, etc. which are common in everyday discourse but cannot be handled by conventional logic in which only the two quantifiers are in principle employed, i. e. the universal quantifier *for all* and the existential quantifier *for at least one*.

A *linguistically quantified statement* is exemplified by *most experts are convinced* and may be generally written as

$$Q y's \text{ are } F \,, \tag{87}$$

where Q is a linguistic quantifier (e. g., most), $Y = \{y\}$ is a set of objects (e. g., experts) and F is a property (e. g., convinced).

Importance B may also be added to the linguistically quantified statement (87) yielding

$$QBy'\text{s are } F \tag{88}$$

exemplified by *most (Q) of the important (B) experts* (y's) *are convinced (F)*.

For our purposes, the problem is to find the (degree of) truth of such linguistically quantified statements (87) and (88), denoted $\tau(Qy'\text{s are } F)$ in the former case and $\tau(QBy'\text{s are } F)$ in the latter case, knowing the truth of the statements, y is F, denoted $\tau(y \text{ is } F)$, for all $y \in Y$. Evidently, all these degrees of truth (truths) will be meant as real numbers from the unit interval.

Fortunately enough these truth values may be found by some fuzzy logic calculi [cf. Kacprzyk [44], Yager [91] or Zadeh [101]. For lack of space we are unable to discuss these issues here, and refer the reader to, e. g., the source papers or Kacprzyk's [44] book.

Fuzzy linguistic quantifiers are very relevant both for theory and applications (ranging from decision analysis, social choice, optimization and control to database queries). The fuzzy linguistic quantifiers have a relation to the so-called ordered weighted averaging (OWA) operators introduced by Yager in 1982, and we refer the reader for a comprehensive coverage of their theory and applications to Yager and Kacprzyk's [93] volume.

Bellman and Zadeh's General Approach to Decision Making Under Fuzziness

Fuzzy logic, the essence of which has been presented in previous sections, has found applications in a multitude of areas of science and technology, and a full coverage is beyond the scope of our exposition.

Since *decision making* is by far the most well known, omnipresent problem, we will just sketch the application of fuzzy sets theory to the broadly perceived decision making.

The purpose of this section is to provide the reader with a brief introduction to Bellman and Zadeh's [8] general approach to decision making under fuzziness, originally termed *decision making in a fuzzy environment*, a simple yet extremely powerful framework within which virtually all fuzzy models related to decision making, optimization and control have been dealt with.

In Bellman and Zadeh's [8] setting the imprecision (fuzziness) of the environment within which the decision making (control, ...) process proceeds is modeled by the introduction of the so-called *fuzzy environment* which consists of fuzzy goals, fuzzy constraints, and fuzzy decision.

Suppose that we have some set of possible *options* (or alternatives, variants, choices, decisions, ...), $X = \{x\}$, which contains all the possible (relevant, feasible, ...) values, courses of action, etc.

The *fuzzy goal* is now defined as a fuzzy set in the set G in the set of options X, characterized by its membership function $\mu_G : X \longrightarrow [0,1]$ such that $\mu_G(x) \in [0,1]$ specifies the grade of membership of a particular option $x \in X$ in the fuzzy goal G.

The *fuzzy constraint* is similarly defined as a fuzzy set C in the set of options X, characterized by its membership function $\mu_C : X \longrightarrow [0,1]$ such that $\mu_C(x) \in [0,1]$ specifies the grade of membership of a particular option $x \in X$ in the fuzzy constraint C.

The fuzzy goal and fuzzy constraint are illustrated in Example 18.

As to the interpretation of fuzzy goals and/or constraints (notice that their definitions do indicate an intrinsic analogy!), in some, mostly earlier, works as, e. g., in Bellman and Zadeh [8], the following view on the essence of the fuzzy goal is advocated. Suppose that $f : X \longrightarrow R$ is a conventional performance (objective) function which associates with each option $x \in X$ a real number $f(x) \in R$, and which is bounded, i. e. $f(x) \leq M < \infty$, for each $x \in X$, where $M = \max_{x \in X} f(x)$.

Then the membership function of the fuzzy goal G can be defined as a normalized performance function f, i. e.

$$\mu_G(x) = \frac{f(x)}{M} = \frac{f(x)}{\max_{x \in X} f(x)}, \quad \text{for each } x \in X. \tag{89}$$

A fuzzy goal may however be also viewed from a different perspective that is often convenient, i. e. in terms of Simon's *satisfaction levels*, in particular in view of the representation of the fuzzy goal's membership function in a piecewise linear form as usually assumed, also here. Then the piecewise linear membership function of a fuzzy goal in Fig. 8 should be understood as follows: If the value of x attained is at least $\overline{x}_G (= 8)$, which is the *satisfaction level* of x, i. e. for $x \geq \overline{x}_G$, then $\mu_G(x) = 1$ which means that we are fully satisfied with the x attained. On the other hand, if the x attained does not exceed $\underline{x}_G (= 5)$, which is the lowest possible value of x, then $\mu_G(x) = 0$ which means that we are fully dissatisfied with such a value of x or, in other words, this value is impossible. For the intermediate values, $\underline{x}_G < x < \overline{x}_G$, we have $0 < \mu_G(x) < 1$ which means that our satisfaction as to a particular value of x is intermediate. The interpretation of a fuzzy constraint is analogous.

It is now easy to see that the above interpretation provides a *common denominator* for the fuzzy goal and fuzzy constraint. They may be treated in an analogous

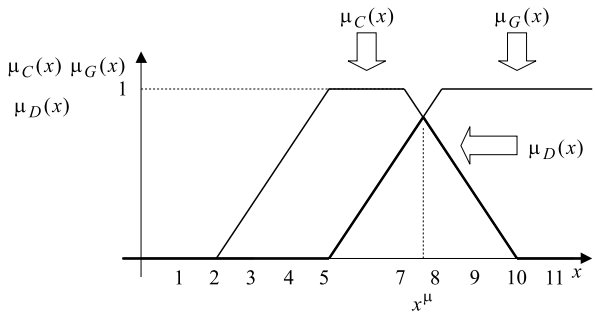

Fuzzy Sets Theory, Foundations of, Figure 8
Fuzzy goal, fuzzy constraint, fuzzy decision, and the optimal (maximizing) decision

way which is one of merits of Bellman and Zadeh's [8] approach.

The above suggests the following general formulation of the decision making problem in a fuzzy environment

$$\text{“Attain } G \text{ \underline{and} satisfy } C\text{”} \tag{90}$$

which should be meant as to determine a decision (an option or a set of options) which simultaneously fulfills the fuzzy goal and fuzzy constraint, and which belongs to the available (or, maybe, relevant, feasible, ...) ones.

The fuzziness of the fuzzy goal and fuzzy constraint implies the above decision, a fuzzy decision, to be a fuzzy set defined in the set of options which results from the intersection (34) of the fuzzy goal and fuzzy constraint. Formally, if G is a fuzzy goal and C is a fuzzy constraint, both defined as fuzzy sets in the set of options $X = \{x\}$, the *fuzzy decision* D is a fuzzy set defined in X given as

$$\mu_D(x) = \mu_G(x) \wedge \mu_C(x)\,, \quad \text{for each } x \in X \tag{91}$$

where $\wedge$ is the minimum operation, i. e. $a \wedge b = \min(a, b)$.

The fuzzy decision (91) is most widely used, but $\wedge$ may clearly be replaced by another operation as, e. g., a t-norm (36).

Example 18 Suppose that G is *x should be much larger than five*, and C is *x should be about 5*, as in Fig. 8.

The membership function of the fuzzy decision is given in heavy line and interpreted as follows. The set of possible options is the interval $[5, 10]$ because $\mu_D(x) > 0$ for $5 \le x \le 10$. The value of $\mu_D(x) \in [0, 1]$ is meant as the degree of satisfaction from the choice of a particular $x \in X$, from zero for full dissatisfaction (impossibility of x) to one for full satisfaction, though all intermediate values, and the higher the value of $\mu_D(x)$, the higher the satisfaction from x.

Notice that in Fig. 8, $\mu_D(x) < 1$ which means that there is no option which fully satisfies both the fuzzy goal and fuzzy constraint. In other words, there is a discrepancy or conflict between the fuzzy goal and constraint.

In practice, however, we need to find a nonfuzzy solution to be implemented. The above interpretation of the fuzzy decision immediately suggests that the best (nonfuzzy) choice in this case would be the one corresponding to the highest value of $\mu_D(x)$.

The *maximizing decision* is therefore defined as an $x^* \in X$ such that

$$\mu_D(x^*) = \max_{x \in X} \mu_D(x) \tag{92}$$

and an example may be found in Fig. 8 where $x^* = 7.5$.

Notice that the above is clearly equivalent to the defuzzification of the fuzzy decision (cf. Sect. "Defuzzification of Fuzzy Sets"), and other defuzzification procedures may also be used in principle. However, (92) is often the only practical choice (cf. Kacprzyk [44]), and will be assumed here.

In non-trivial real problems there are multiple fuzzy goals and fuzzy constraints, and they may be handled within the above framework in quite a straightforward manner.

Suppose a more general situation than the fuzzy constraint C is defined as a fuzzy set in $X = \{x\}$, and the fuzzy goal G is defined as a fuzzy set in $Y = \{y\}$. Moreover, suppose that a function $f\colon x \longrightarrow Y, y = f(x)$, is known. Typically, X and Y may be sets of options and outcomes, notably causes and effects.

Now, the *induced fuzzy goal* G' in X generated by the given fuzzy goal G in Y is defined as

$$\mu_{G'}(x) = \mu_G[f(x)]\,, \quad \text{for each } x \in X\,. \tag{93}$$

Example 19 Let $X = \{1, 2, 3, 4\}$, $Y = \{2, 3, \ldots, 10\}$, and $y = 2x + 1$. If now

$$G = 0.1/2 + 0.2/3 + 0.4/4 + 0.5/5 + 0.6/6 + 0.7/7 + 0.8/8 + 1/9 + 1/10$$

then

$$G' = \mu_G(3)/1 + \mu_G(5)/2 + \mu_G(7)/70.2/1 + 0.5/2 + 0.7/3 + 1/4\,.$$

The *fuzzy decision* is now defined analogously as (91), i. e.

$$\mu_D(x) = \mu_{G'}(x) \wedge \mu_C(x)\,, \quad \text{for each } x \in X\,. \tag{94}$$

Clearly, for $n > 1$ fuzzy goals $G_1, \ldots, G_n$ defined in Y, $m > 1$ fuzzy constraints $C_1, \ldots, C_m$ defined in X, and a function $f: X \longrightarrow Y$, $y = f(x)$, we analogously have

$$\mu_D(x) = \mu_{G_1'}(x) \wedge \ldots \wedge \mu_{G_n'}(x) \wedge \mu_{C_1}(x) \wedge \ldots \wedge \mu_{C_n}(x), \quad \text{for each } x \in X . \tag{95}$$

The *maximizing decision* is defined as (92), i. e.

$$\mu_D(x^*) = \max_{x \in X} \mu_D(x) .$$

The basic conceptual fuzzy decision making model can be used in many specific areas, notably in fuzzy optimization which will be covered elsewhere in this volume.

The models of decision making under fuzziness developed above can also be extended to the case of multiple criteria, multiple decision makers, and multiple stage cases. We will present the last extension, to fuzzy multistage decision making (control) case which makes it possible to account for dynamics.

Multistage Decision Making (Control) Under Fuzziness

In this case it is convenient to use control-related notation and terminology. In particular, decisions will be referred to as *controls*, the discrete time moments at which decisions are to be made – as *control stages*, and the input-output (or cause-effect) relationship – as a *system under control*.

The essence of multistage control in a fuzzy environment may be portrayed as in Fig. 9.

First, suppose that the control space is $U = \{u\} = \{c_1, \ldots, c_m\}$ and the state space is $X = \{x\} = \{s_1, \ldots, s_n\}$. Initially we are in some initial state $x_0 \in X$. We apply a control $u_0 \in U$ subjected to a fuzzy constraint $\mu_{C^0}(u_0)$. We attain a state $x_1 \in X$ via a known cause-effect relationship (i. e. S); a fuzzy goal $\mu_{G^1}(x_1)$ is imposed on x_1. Next, we apply a control u_1 subjected to a fuzzy constraint $\mu_{C^1}(u_1)$, and attain a fuzzy state x_2 on which a fuzzy goal $\mu_{G^2}(x_2)$ is imposed, etc.

Suppose for simplicity that the system under control is deterministic and its temporal evolution is governed by a *state transition equation*

$$f: X \times U \longrightarrow X , \tag{96}$$

such that

$$x_{t+1} = f(x_t, u_t) , \quad t = 0, 1, \ldots \tag{97}$$

where $x_t, x_{t+1} \in X = \{s_1, \ldots, s_n\}$ are the states at control stages t and $t+1$, respectively, and $u_t \in U = \{c_1, \ldots, c_m\}$ is the control at t.

At each t, $u_t \in U$ is subjected to a *fuzzy constraint* $\mu_{C^t}(u_t)$, and on the state attained $x_{t+1} \in X$ a *fuzzy goal* is imposed; $t = 0, 1, \ldots$. The *initial state* is $x_0 \in X$ and is assumed to be known, and given in advance. The *termination time* (planning, or control, horizon), i. e. is the maximum number of control stages, is denoted by $N \in \{1, 2, \ldots\}$, and may be finite or infinite.

The performance (goodness) of the multistage control process under fuzziness is evaluated by the fuzzy decision

$$\mu_D(u_0, \ldots, u_{N-1} \mid x_0) = \mu_{C^0}(u_0) \wedge \mu_{G^1}(x_1) \wedge \ldots \wedge \mu_{C^{N-1}}(u_{N-1}) \wedge \mu_{G^N}(x_N) . \tag{98}$$

In most cases, however, a slightly simplified form of the fuzzy decision (98) is used, namely it is assumed all the subsequent fuzzy controls, $u_0, u_1, \ldots, u_{N-1}$, are subjected to the fuzzy constraints, $\mu_{C^0}(u_0), \mu_{C^1}(u_1), \ldots, \mu_{C^{N-1}}(u_{N-1}$, while the fuzzy goal is just imposed on the final state x_N, $\mu_{G^N}(x_N)$. In such a case the fuzzy decision becomes

$$\mu_D(u_0, \ldots, u_{N-1} \mid x_0) = \mu_{C^0}(u_0) \wedge \ldots \wedge \mu_{C^{N-1}}(u_{N-1}) \wedge \mu_{G^N}(x_N) . \tag{99}$$

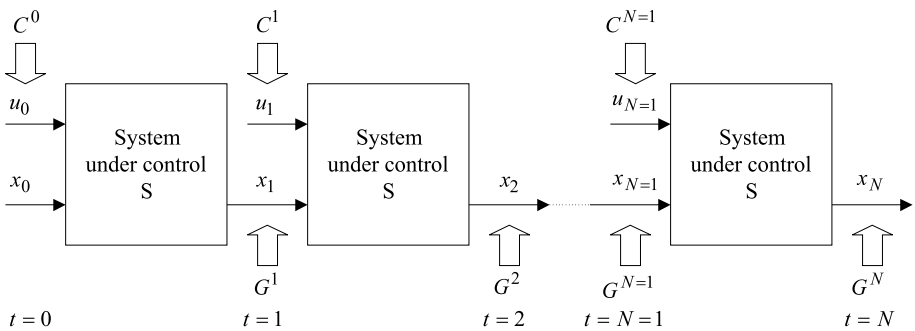

Fuzzy Sets Theory, Foundations of, Figure 9
Essence of the multistage control in a fuzzy environment (under fuzziness)

The multistage control problem in a fuzzy environment is now formulated as to find an optimal sequence of controls $u_0^*, \ldots, u_{N-1}^*$, $u_t^* \in U$, $t = 0, 1, \ldots, N-1$, such that:

$$\mu_D(u_0^*, \ldots, u_{N-1}^* \mid x_0) = \max_{u_0,\ldots,u_{N-1} \in U} \mu_D(u_0, \ldots, u_{N-1} \mid x_0) \,. \quad (100)$$

Usually it is more convenient to express the solution, i. e. the controls to be applied, as a *control policy* $a_t \colon X \longrightarrow U$, such that $u_t = a_t(x_t)$, $t = 0, 1, \ldots$, i. e. the control to be applied at t is expressed as a function of the state at t.

The above basic formulation of multistage control in a fuzzy environment may readily be extended with respect to:

- The type of the termination time (fixed and specified, implicitly specified, fuzzy, and infinite), and
- the type of the system under control (deterministic, stochastic, and fuzzy).

For a detailed analysis of resulting problems and their solutions (by employing dynamic programming, branch-and-bound, genetic algorithms, etc.) we refer the reader to Kacprzyk's [42,44] books.

Concluding Remarks

We provided a brief survey of basic elements of Zadeh's [95] fuzzy sets theory, mainly of basic properties of fuzzy sets, operations on fuzzy sets, fuzzy relations and their compositions, linguistic variables, the extension principle, fuzzy arithmetic, fuzzy events and their probabilities, fuzzy logic, and Bellman and Zadeh's [8] general approach to decision making in a fuzzy environment. Various aspects of fuzzy sets theory will be expanded in other papers in this part of the volume.

Bibliography

1. Atanassov KT (1983) Intuitionistic fuzzy sets. VII ITKR session. Central Sci.-Techn. Library of Bulg. Acad. of Sci., Sofia pp 1697/84. (in Bulgarian)
2. Atanassov KT (1986) Intuitionistic fuzzy sets. Fuzzy Sets Syst 20:87–96
3. Atanassov KT (1999) Intuitionistic fuzzy sets: Theory and applications. Springer, Heidelberg
4. Bandemer H, Gottwald S (1995) Fuzzy sets, fuzzy logic, fuzzy methods, with applications. Wiley, Chichester
5. Bandemer H, Näther W (1992) Fuzzy data analysis. Kluwer, Dordrecht
6. Bandler W, Kohout LJ (1980) Fuzzy power sets for fuzzy implication operators. Fuzzy Sets Syst 4:13–30
7. Bellman RE, Giertz M (1973) On the analytic formalism of the theory of fuzzy sets. Inform Sci 5:149–157
8. Bellman RE, Zadeh LA (1970) Decision making in a fuzzy environment. Manag Sci 17:141–164
9. Belohlávek R, Vychodil V (2005) Fuzzy equational logic. Springer, Heidelberg
10. Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Plenum, New York
11. Black M (1937) Vagueness: An exercise in logical analysis. Philos Sci 4:427–455
12. Black M (1963) Reasoning with loose concepts. Dialogue 2: 1–12
13. Black M (1970) Margins of precision. Cornell University Press, Ithaca
14. Buckley JJ (2004) Fuzzy statistics. Springer, Heidelberg
15. Buckley JJ (2005) Fuzzy probabilities. New approach and applications, 2nd edn. Springer, Heidelberg
16. Buckley JJ (2005) Simulating fuzzy systems. Springer, Heidelberg
17. Buckley JJ (2006) Fuzzy probability and statistics. Springer, Heidelberg
18. Buckley JJ, Eslami E (2002) An introduction to fuzzy logic and fuzzy sets. Springer, Heidelberg
19. Calvo T, Mayor G, Mesiar R (2002) Aggregation operators. New trends and applications. Springer, Heidelerg
20. Carlsson C, Fullér R (2002) Fuzzy reasoning in decision making and optimization. Springer, Heidelberg
21. Castillo O, Melin P (2008) Type-2 fuzzy logic: Theory and applications. Springer, Heidelberg
22. Cox E (1994) The fuzzy system handbook. A practitioner's guide to building, using, and maintaining fuzzy systems. Academic, New York
23. Cross V, Sudkamp T (2002) Similarity and compatibility in fuzzy set theory. Assessment and applications. Springer, Heidelberg
24. Delgado M, Kacprzyk J, Verdegay JL, Vila MA (eds) (1994) Fuzzy optimization: Recent advances. Physica, Heidelberg
25. Dompere KK (2004) Cost-benefit analysis and the theory of fuzzy decisions. Fuzzy value theory. Springer, Heidelberg
26. Dompere KK (2004) Cost-benefit analysis and the theory of fuzzy decisions. Identification and measurement theory. Springer, Heidelberg
27. Driankov D, Hellendorn H, Reinfrank M (1993) An introduction to fuzzy control. Springer, Berlin
28. Dubois D, Prade H (1980) Fuzzy sets and systems: Theory and applications. Academic, New York
29. Dubois D, Prade H (1985) Théorie des possibilités. Applications a la représentation des connaissances en informatique. Masson, Paris
30. Dubois D, Prade H (1988) Possibility theory: An approach to computerized processing of uncertainty. Plenum, New York
31. Dubois D, Prade H (1996) Fuzzy sets and systems (Reedition on CR-ROM of [28]. Academic, New York
32. Fullér R (2000) Introduction to neuro-fuzzy systems. Springer, Heidelberg
33. Gaines BR (1977) Foundations of fuzzy reasoning. Int J Man-Mach Stud 8:623–668
34. Gil Aluja J (2004) Fuzzy sets in the management of uncertainty. Springer, Heidelberg
35. Gil-Lafuente AM (2005) Fuzzy logic in financial analysis. Springer, Heidelberg

36. Glöckner I (2006) Fuzzy quantifiers. A computational theory. Springer, Heidelberg
37. Goguen JA (1967) *L*-fuzzy sets. J Math Anal Appl 18:145–174
38. Goguen JA (1969) The logic of inexact concepts. Synthese 19:325–373
39. Goodman IR, Nguyen HT (1985) Uncertainty models for knowledge-based systems. North-Holland, Amsterdam
40. Hájek P (1998) Metamathematics of fuzzy logic. Kluwer, Dordrecht
41. Hanss M (2005) Applied fuzzy arithmetic. An introduction with engineering applications. Springer, Heidelberg
42. Kacprzyk J (1983) Multistage decision making under fuzziness, Verlag TÜV Rheinland, Cologne
43. Kacprzyk J (1992) Fuzzy sets and fuzzy logic. In: Shapiro SC (ed) Encyclopedia of artificial intelligence, vol 1. Wiley, New York, pp 537–542
44. Kacprzyk J (1996) Multistage fuzzy control. Wiley, Chichester
45. Kacprzyk J, Fedrizzi M (eds) (1988) Combining fuzzy imprecision with probabilistic uncertainty in decision making. Springer, Berlin
46. Kacprzyk J, Orlovski SA (eds) (1987) Optimization models using fuzzy sets and possibility theory. Reidel, Dordrecht
47. Kandel A (1986) Fuzzy mathematical techniques with applications. Addison Wesley, Reading
48. Kaufmann A, Gupta MM (1985) Introduction to fuzzy mathematics – theory and applications. Van Nostrand Reinhold, New York
49. Klement EP, Mesiar R, Pap E (2000) Triangular norms. Springer, Heidelberg
50. Klir GJ (1987) Where do we stand on measures of uncertainty, ambiguity, fuzziness, and the like? Fuzzy Sets Syst 24:141–160
51. Klir GJ, Folger TA (1988) Fuzzy sets, uncertainty and information. Prentice-Hall, Englewood Cliffs
52. Klir GJ, Wierman M (1999) Uncertainty-based information. Elements of generalized information theory, 2nd end. Springer, Heidelberg
53. Klir GJ, Yuan B (1995) Fuzzy sets and fuzzy logic: Theory and application. Prentice-Hall, Englewood Cliffs
54. Kosko B (1992) Neural networks and fuzzy systems. Prentice-Hall, Englewood Cliffs
55. Kruse R, Meyer KD (1987) Statistics with vague data. Reidel, Dordrecht
56. Kruse R, Gebhard J, Klawonn F (1994) Foundations of fuzzy systems. Wiley, Chichester
57. Kuncheva LI (2000) Fuzzy classifier design. Springer, Heidelberg
58. Li Z (2006) Fuzzy chaotic systems modeling, control, and applications. Springer, Heidelberg
59. Liu B (2007) Uncertainty theory. Springer, Heidelberg
60. Ma Z (2006) Fuzzy database modeling of imprecise and uncertain engineering information. Springer, Heidelberg
61. Mamdani EH (1974) Application of fuzzy algorithms for the control of a simple dynamic plant. Proc IEE 121:1585–1588
62. Mareš M (1994) Computation over fuzzy quantities. CRC, Boca Raton
63. Mendel J (2000) Uncertain rule-based fuzzy logic systems: Introduction and new directions. Prentice Hall, New York
64. Mendel JM, John RIB (2002) Type-2 fuzzy sets made simple. IEEE Trans Fuzzy Syst 10(2):117–127
65. Mordeson JN, Nair PS (2001) Fuzzy mathematics. An introduction for engineers and scientists, 2nd edn. Springer, Heidelberg
66. Mukaidono M (2001) Fuzzy logic for beginners. World Scientific, Singapore
67. Negoiţa CV, Ralescu DA (1975) Application of fuzzy sets to system analysis. Birkhäuser/Halstead, Basel/New York
68. Nguyen HT, Waler EA (205) A first course in fuzzy logic, 3rd end. CRC, Boca Raton
69. Nguyen HT, Wu B (2006) Fundamentals of statistics with fuzzy data. Springer, Heidelberg
70. Novák V (1989) Fuzzy sets and their applications. Hilger, Bristol, Boston
71. Novák V, Perfilieva I, Močkoř J (1999) Mathematical principles of fuzzy logic. Kluwer, Boston
72. Peeva K, Kyosev Y (2005) Fuzzy relational calculus. World Scientific, Singapore
73. Pedrycz W (1993) Fuzzy control and fuzzy systems, 2nd edn. Research Studies/Wiley, Taunton/New York
74. Pedrycz W (1995) Fuzzy sets engineering. CRC, Boca Raton
75. Pedrycz W (ed) (1996) Fuzzy modelling: Paradigms and practice. Kluwer, Boston
76. Pedrycz W, Gomide F (1998) An introduction to fuzzy sets: Analysis and design. MIT Press, Cambridge
77. Petry FE (1996) Fuzzy databases. Principles and applications. Kluwer, Boston
78. Piegat A (2001) Fuzzy modeling and control. Springer, Heidelberg
79. Ruspini EH (1991) On the semantics of fuzzy logic. Int J Approx Reasining 5:45–88
80. Rutkowska D (2002) Neuro-fuzzy architectures and hybrid learning. Springer, Heidelberg
81. Rutkowski L (2004) Flexible neuro-fuzzy systems. Structures, learning and performance evaluation. Kluwer, Dordrecht
82. Seising R (2007) The fuzzification of systems. The genesis of fuzzy set theory and its initial applications – developments up to the 1970s. Springer, Heidelberg
83. Smithson M (1989) Ignorance and uncertainty. Springer, Berlin
84. Sousa JMC, Kaymak U (2002) Fuzzy decision making in modelling and control. World Scientific, Singapore
85. Thole U, Zimmermann H-J, Zysno P (1979) On the suitability of minimum and product operator for the intersection of fuzzy sets. Fuzzy Sets Syst 2:167–180
86. Turkşen IB (1991) Measurement of membership functions and their acquisition. Fuzzy Sets Syst 40:5–38
87. Türksen IB (2006) An ontlogical and epistemological perspective of fuzzy set theory. Elsevier, New York
88. Wang Z, Klir GJ (1992) Fuzzy measure theory. Kluwer, Boston
89. Wygralak M (1996) Vaguely defined objects. Representations, fuzzy sets and nonclassical cardinality theory. Kluwer, Dordrecht
90. Wygralak M (2003) Cardinalities of fuzzy sets. Springer, Heidelberg
91. Yager RR (1983) Quantifiers in the formulation of multiple objective decision functions. Inf Sci 31:107–139
92. Yager RR, Filev DP (1994) Essentials of fuzzy modeling and control. Wiley, New York
93. Yager RR, Kacprzyk J (eds) (1996) The ordered weighted averaging operators: Theory, methodology and applications. Kluwer, Boston

94. Yazici A, George R (1999) Fuzzy database modeling. Springer, Heidelberg
95. Zadeh LA (1965) Fuzzy sets. Inf Control 8:338–353
96. Zadeh LA (1968) Probability measures of fuzzy events. J Math Anal Appl 23:421–427
97. Zadeh LA (1973) Outline of a new approach to the analysis of complex systems and decision processes. IEEE Trans Syst, Man Cybern SMC-2:28–44
98. Zadeh LA (1975) Fuzzy logic and approximate reasoning. Synthese 30:407–428
99. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning. Inf Sci (Part I) 8:199–249, (Part II) 8:301–357, (Part III) 9:43–80
100. Zadeh LA (1978) Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets Syst 1:3–28
101. Zadeh LA (1983) A computational approach to fuzzy quantifiers in natural languages. Comput Math Appl 9:149–184
102. Zadeh LA (1985) Syllogistic reasoning in fuzzy logic and its application to usuality and reasoning with dispositions. IEEE Trans Syst Man Cybern SMC-15:754–763
103. Zadeh LA (1986) Fuzzy probabilities. Inf Process Manag 20:363–372
104. Zadeh LA, Kacprzyk J (eds) (1992) Fuzzy logic for the management of uncertainty. Wiley, New York
105. Zadeh LA, Kacprzyk J (eds) (1999) Computing with words in information/intelligent systems. 1 Foundations. Springer, Heidelberg
106. Zadeh LA, Kacprzyk J (eds) (1999) Computing with words in information/intelligent systems. 2 Applications. Springer, Heidelberg
107. Zang H, Liu D (2006) Fuzzy modeling and fuzzy control. Birkhäuser, New York
108. Zimmermann H-J (1976) Description and optimization of fuzzy systems. Int J Gen Syst 2:209–215
109. Zimmermann H-J (1987) Fuzzy sets, decision making, and expert systems. Kluwer, Dordrecht
110. Zimmermann H-J (1996) Fuzzy set theory and its applications, 3rd edn. Kluwer, Boston
111. Zimmermann H-J, Zysno P (1980) Latent connectives in human decision making. Fuzzy Sets Syst 4:37–51

Fuzzy System Models Evolution from Fuzzy Rulebases to Fuzzy Functions

I. Burhan Türkşen
Head Department of Industrial Engineering, TOBB-ETÜ, (Economics and Technology University of the Union of Turkish Chambers and Commodity Exchanges), Ankara, Republic of Turkey

Article Outline

Glossary
Definition of the Subject
Introduction
Type 1 Fuzzy System Models of the Past
Future of Fuzzy System Models
Case Study Applications
Experimental Design
Conclusions and Future Directions
Bibliography

Glossary

Z-FRB Zadeh's linguistic fuzzy rule base.

TS-FR Takagi–Sugeno fuzzy rule base.

c^* the number of rules in the rule base.

nv the number of input variables in the system.

$X = (x_1, x_2, \ldots, x_{nv})$ input vector.

x_j the input (explanatory variable), for $j = 1, \ldots, nv$.

A_{ji} the linguistic label associated with jth input variable of the antecedent in the ith rule.

B_i the consequent linguistic label of the ith rule.

R_i ith rule with membership function $\mu_i(x_j)\colon x_j \to [0, 1]$.

A_i multidimensional type 1 fuzzy set to represent the ith antecedent part of the rules defined by the membership function $\mu_i(x)\colon x \to [0, 1]$.

$a_i = (a_{i,1}, \ldots, a_{i,nv})$ the regression coefficient vector associated with the ith rule.

b_i the scalar associated with the ith rule in the regression equation.

SFF-LSE "Special Fuzzy Functions" that are generated by the Least Squares Estimation.

SFF-SVM "Special Fuzzy Functions" estimated by Support Vector Machines.

The estimate of y_i would be obtained as $Y_i^* = \beta_{i0}^* + \beta_{i1}^* \Gamma_i + \beta_{i2}^* X$ with SFF-LSE

y the dependent variable, assumed to be a linear function.

$\beta_{j'}$ $j = 0, 1, \ldots, nv$, indicate how a change in one of the independent variables affects the dependent variable.

$X = (x_{j,k} \mid j = 1, \ldots, nv,\ k = 1, \ldots, nd)$ the set of observations in a training data set.

m the level of fuzziness, $m = 1.1, \ldots, 2.5$.

c the number of clusters, $c = 1, \ldots, 10$.

J the objective function to be minimized.

$\|\cdot\|_A$ a norm that specifies a distance-based similarity between the data vector x_k and a fuzzy.

$A = I$ the Euclidean norm.

$A = COV^{-1}$ the Mahalonobis norm.

COV the covariance matrix.

m^*, c^* the optimal pair.

$v_{\underset{m^*}{X|Y},i} = (x_{1,i}, x_{2,i}, \ldots, x_{nv,i}, y_i)$ the cluster centers for $m = m^*$ and each cluster $i = 1, \ldots, c^*$.

$v_{X,i} = (x_{1,i}, x_{2,i}, \ldots, x_{nv,i})$ the cluster centers of the "input space" for $m = m^*$ and $c = 1, \ldots, c^*$.

$\gamma_{ik}(x_k)$ the normalized membership values of x data sample in the ith cluster, $i = 1, \ldots, c^*$.

$\Gamma_i = (\gamma_{ik} \mid i = 1, \ldots, c^*; k = 1, \ldots, nd)$ the membership values of x in the ith cluster.

X_i', X_i'', X_i''' potential augmented input matrices in SFF-LSE.

$f(\vec{x}_k) = \hat{y}_k = \langle \vec{w}, \vec{x}_k \rangle + b$ linear Support Vector Regression (SVR) equation.

$l_\varepsilon = |y_k - f(\vec{x}_k)|_\varepsilon = \max\{0, |y - f(x)| - \varepsilon\}$ ε-insensitive loss function.

$\vec{w}, b$ the weight vector and bias term.

$c > 0$ the tradeoff between the empirical error and the complexity term.

$\xi_k \geq 0$ and $\xi_k^* \geq 0$ the slack variables.

α_k and α_k^* Lagrange multipliers.

$K\langle \vec{x}_{k'}, \vec{x}_k \rangle$ the kernel mapping of the input vectors.

$\hat{y}_{ik'}^* = \hat{f}(\vec{x}_{ik'}, \alpha_i, \alpha_i^*) = \sum_{k=1}^{nd} (\alpha_{ik} - \alpha_{ik}^*) K\langle \vec{x}_{ik'}, \vec{x}_{ik} \rangle + b_i$ output value of kth data sample in ith cluster with SFF-SVM.

$\tilde{A} = \{(x, (u, f_x(u))) | x \in X, u \in J_x \subseteq [0, 1]\}$ type 2 fuzzy set, $\tilde{A}$.

$f_x(u): J_x \to [0, 1], \forall u \in J_x \subseteq [0, 1], \forall x \in X$ *secondary membership function.*

$f_x(u) = 1, \forall x \in X, \forall u \in J_x, J_x \subseteq [0, 1]$ *interval value type 2 membership function.*

[27] the domain of the primary membership is discrete and the secondary membership values are fixed to 1. Thus, the proposed method is utilizing discrete interval valued type 2 fuzzy sets in order to represent the linguistic values assigned to each fuzzy variable in each rule. These fuzzy sets can be mathematically defined as follows:

$\tilde{A} = \int_{x \in X} \left[\sum_{u \in J_x} 1/u \right] / x$ Discrete Interval Valued Type 2 Fuzzy Sets (DIVT2FS)

$x \in X \subseteq \Re$, $u \in J_x = \{J_{xr}\}$, $r = 1, \ldots, NM$ a system variable in continuous domain.

$\mathbf{u}$ the primary membership value with discrete domain.

J_{xr} rth primary membership value associated with x.

Definition of the Subject

Fuzzy System Modeling (FSM) is one of the most prominent tools that can be used to identify the behavior of highly nonlinear systems with uncertainty. In the past, FSM techniques utilized type 1 fuzzy sets in order to capture the uncertainty in the system. However, since type 1 fuzzy sets express the belongingness of a crisp value x' of an input variable x in a fuzzy set A by a crisp membership value $\mu_A(x')$, they cannot fully capture the uncertainties associated with higher order imprecision in identifying membership functions. In the future, we are likely to observe higher types of fuzzy sets, such as type 2 fuzzy sets. The use of type 2 fuzzy sets and linguistic logical connectives drew a considerable amount of attention in the realm of fuzzy system modeling in the last two decades. In this paper, we first review type 1 fuzzy system models known as Zadeh, Takagi–Sugeno and Türkşen models; then we review potentially future realizations of type 2 fuzzy systems again under the headings of Zadeh and Takagi–Sugeno and Türkşen fuzzy system models, in contrast to type 1 fuzzy system models. Type 2 fuzzy system models have a higher predictive power. One of the essential problems of type 2 fuzzy system models is computational complexity. In data-driven fuzzy system modeling methods discussed here, Fuzzy C-Means (FCM) clustering algorithm is used in order to identify the system structure.

Introduction

Fuzzy system models are the most successful models to handle uncertainties in decision-making. The major advantages of fuzzy system models are their robustness and transparency. Fuzzy system modeling achieves robustness by using fuzzy sets which incorporates imprecision in system models. In addition, unlike some other system models, such as neural networks, the fuzzy system models are highly descriptive, i. e., transparent.

In the last two decades, researchers proposed several data driven type 1 fuzzy system modeling approaches that can extract the hidden rules of a system behavior automatically by using historical data. The system modeling methods, proposed by Nakanishi et al. [31], Takagi–Sugeno [39], Sugeno and Yasukawa [38], Emami et al. [17], are among the most notable ones. Since these methods utilize only the historical data, i. e., since they do not require expert knowledge, they are strictly data-driven modeling techniques. Thus, in addition to being robust and transparent, these system-modeling techniques can identify system model structure objectively for a given performance measure.

In these traditional fuzzy system models of the past, the structure is characterized by type 1 fuzzy sets. Type 1 fuzzy sets, defined on a universe of discourse, maps an element onto a precise number in the unit interval [0, 1]. This conflicts with the basic philosophy of fuzzy set and logic theory. In the future, it is expected, fuzzy system models will be formed with higher order fuzzy sets, such as type 2 fuzzy sets, which was first proposed by Zadeh [50]. They will be

used more and more in order to capture the uncertainty associated with membership functions. A type 2 fuzzy set can be informally defined as a fuzzy set that is characterized by a fuzzy membership function, i. e., membership values also are in the unit interval [0, 1] in the computational level.

In this paper, we propose to review first the type 1 fuzzy system models as the historically significant but successful modeling activity of the past in the domain of fuzzy control system problems. Next we review the type 2 fuzzy system models as the potential future modeling activity in the domain of fuzzy decision support system problems.

In an historical sense, Zadeh [50], and Takagi–Sugeno [39] versions of type 1 fuzzy system models are typically basic fuzzy rule base models. In contrast, type 1 "special fuzzy function" models, recently proposed by Türkşen [46], are alternate models to fuzzy rule base models. They give better predictions in comparison to type 1 fuzzy rule base models [3,46]. Furthermore, fuzzy rule base models, in general, can not capture the interactive nature of a problem space due to projection deficiency. Whereas fuzzy function models are able to capture the interactions of all variables since they are not subject to projection deficiencies.

For future developments, currently there are mainly two essentially different schools of thought in type 2 fuzzy system modeling research. The first one is based on the properties of the linguistic connectives which causes the generation of interval valued type 2 fuzzy sets. Türkşen [42,43,44,45] mathematically showed that Fuzzy Conjunctive Canonical Forms (FCCF) and Fuzzy Disjunctive Canonical Forms (FDCF) are no longer equivalent to each other for the 16 basic linguistic expressions formed with linguistic connectives "AND", "OR", etc. Furthermore, it is shown that FCCF contains FDCF for certain families of t-norms and t-co-norms. Zimmerman and Zysno [51] empirically showed that linguistic connectives were characterized differently by different people and Türkşen [44] provided the groundwork for this with the Interval Valued type 2 fuzzy sets. These type 2 system models represent uncertainties generated by different linguistic connectives that combine type 1 membership functions. Such combinations generate their FDCF and FCCF expressions and hence identify a particular sort of Interval Valued type 2 representations of systems.

In the second school of thought in type 2 fuzzy system modeling, traditional connectives, i. e., t-norms and co-norms, are utilized directly with the assumption that a t-norm directly corresponds to a linguistic "AND" in its FDCF and a t-co-norm directly corresponds to an "OR" in its FCCF while ignoring FCCF of "AND" and FDCF of "OR". But, each variable is represented by using a type 2 fuzzy set at the beginning of computations. In a series of papers Mendel, Karnik and Lian [23,24,25] extended traditional type 1 system models to type 2 fuzzy system models and proposed inference methods in order to process type 2 fuzzy sets. These studies were explained thoroughly by Mendel in [28]. Several other researchers such as, Starczewski and Rutkowski [37], John and C. Czarnecki [20,21], and Chen and Kawase [10] worked on type 2 fuzzy system models and inference methods.

In general the main inhibitor of the use of the type 2 fuzzy system models is the computational complexity of the inference mechanism that is used to infer a model output by using type 2 fuzzy system models for a given input data vector. Liang and Mendel [26] proposed Interval-Valued (IV) type 2 fuzzy system models in place of full type 2 fuzzy system models together with inference methods to remedy this problem. Thus they use a simplified version of type 2 fuzzy sets, namely, Interval-Valued type 2 fuzzy sets (IVT2FS) rather than full type 2 fuzzy sets (FT2FS).

Recently, Uncu and Türkşen [48] proposed discrete "interval valued type 2 fuzzy sets (DIVT2FS)" which are generated by a variation of the level of fuzziness around fixed cluster centers in applications of FCM. The proposed model representation enables us to identify a computationally efficient inference mechanism.

The rest of this paper is organized as follows: the basic notation, terminology and the three well-known type 1 fuzzy system model structures will be briefly reviewed in Sect. "Introduction" with an emphasis on the more recent "special fuzzy functions". Then the extensions of these well-known fuzzy system modeling structures in type 2 formation will be discussed for potential future studies in Sect. "Future of Fuzzy System Models". Finally, the conclusions will be drawn and the future research directions will be provided in Sect. "Conclusions and Future Directions".

Type 1 Fuzzy System Models of the Past

In general fuzzy system models identify an underlying relationship between input and output variables of a system. In this paper, we deal with Multi-Input Single Output (MISO) version fuzzy system models. Generally fuzzy system models represent relationships between the input and output variables as a collection of: (a) either if-then rules that utilize linguistic labels (i. e., fuzzy sets) or (b) "special fuzzy functions" that takes on membership values and/or their transformations as well as selected input variables as their arguments.

Type 1 Fuzzy Rulebases

In general, the fuzzy rule base structure can be written as follows:

$$R\colon \underset{i=1}{\overset{c^*}{\text{ALSO}}} (\text{ IF antecedent}_i \text{ THEN consequent}_i)\,, \quad (1)$$

where c^* is the number of rules in the rule base. There are several well known fuzzy rule base structures which mainly differ in the representation of their consequents. If the consequent is represented with fuzzy sets then the rule base can be categorized as the Zadeh Fuzzy Rule base, Z-FRB [50] Whereas, if the consequent is represented with a linear equation of input variables, then the rule base structure is known as Takagi–Sugeno Fuzzy Rulebase (TS-FRB) structure [39]. The Z-FRB and TS-FRB structures can be formalized as follows:

Let nv be the number of input variables in the system. Then, the multidimensional antecedent, X, can be defined as $X = (x_1, x_2, \ldots, x_{nv})$, where x_j is the jth input variable of the antecedent in the domain of X. X, can be defined as $X = X_1 \times X_2 \times \ldots \times X_{nv}$, where $X_j \subseteq \Re$ is the domain of variable x_j. Similarly, the domain of the output variable, will be denoted as $Y \subseteq \Re$. Then, the ith rule, r_i, and rulebase, R, in Z-FR structure can be defined as:

$$R_i\colon \text{ IF } \underset{j=1}{\overset{NV}{\text{AND}}} (x_j \in X_j \text{ isr } A_{ji}) \text{ THEN } y \in Y \text{ isr } B_i, \quad \forall i = 1, \ldots, c^* \quad (2)$$

$$R\colon \underset{i=1}{\overset{c^*}{\text{ALSO}}} \left(\text{IF } \underset{j=1}{\overset{NV}{\text{AND}}} (x_j \in X_j \text{ isr } A_{ji}) \text{ THEN } y \in Y \text{ isr } B_i \right), \quad (3)$$

where A_{ji} is the linguistic label associated with jth input variable of the antecedent in the ith rule, R_i, with membership function $\mu_i(x_j)\colon x_j \to [0,1]$ and similarly B_i is the consequent linguistic label of the ith rule with membership function $\mu_i(y)\colon y \to [0,1]$, and c^* is the number of rules in the model. The above structure assumes non-interactivity between input variables because the membership functions of every A_{ji} is obtained by the projection of *nv+1 dimensional internal system representation.* In order to eliminate the non-interactivity assumption, Delgado et al. [12], Babuska et al. [1], and Uncu and Türkşen [48] used multidimensional type 1 fuzzy sets to represent the antecedent part of the rules. Hence, the Z-FRB structure can be expressed as follows:

$$R\colon \underset{i=1}{\overset{c^*}{\text{ALSO}}} (\text{ IF } x \in X \text{ isr } A_i \text{ THEN } y \in Y \text{ isr } B_i)\,, \quad (4)$$

where the multidimensional antecedent fuzzy set of ith rule is defined as $\mu_i(x)\colon x \to [0,1]$. The other well-known fuzzy rulebase structures, namely Takagi–Sugeno (TS-FRB) fuzzy rulebase structure, can be expressed, respectively, as follows:

$$\underset{i=1}{\overset{c^*}{\text{ALSO}}} \left(\text{IF antecedent}_i \text{ THEN } y_i = a_i x^{\mathrm{T}} + b_i \right), \quad (5)$$

where, $a_i = (a_{i,1}, \ldots, a_{i,nv})$ is the regression coefficient vector associated with the ith rule, b_i is the scalar associated with the ith rule.

Type 1 "Special Fuzzy Functions" of the Recent Past

There are at least two ways to form special fuzzy functions: (i) "Special Fuzzy Functions" that are generated by the Least Squares Estimation, SFF-LSE, of Türkşen [46] and (ii) "Special Fuzzy Functions" estimated by Support Vector Machines, SFF-SVM's, of Çelikyilmaz and Türkşen [3]. These "Special Fuzzy Functions" are structurally different from (1) Zadeh's [50] linguistic fuzzy rule bases (Z-FRB), (2) Takagi–Sugeno fuzzy rule base TS-FR [39] (3) "Fuzzy Regression" models of Tanaka et al. [40,41] and its variations, and (5) Hathaway and Bezdek [18] model. Because the proposed "Special Fuzzy Functions" introduces membership values and their transformations as new input variables in addition to the original scalar input variables in function estimations. In particular, these "Special Fuzzy Functions" are structurally different from (1) "Fuzzy Regression" models of Tanaka, et al. [40,41], and its variations, and (2) Hathaway and Bezdek [18] models. This is why we call them "Special Fuzzy Functions".

It ought to be noted that the introduction of membership values and their transformations as new input variables are acceptable since the membership values are obtained from FCM which is a highly nonlinear transformation of the original scalar input variable. Hence there is no co-linearity and thus they can be included in the proposed LSE and SVM structure. For this purpose, first one executes a fuzzy clustering algorithm such as FCM with original selected input variables after an execution of a feature selection algorithm; and then determines (local) optimum number of fuzzy clusters and hence the associated membership values. Then a special fuzzy function to represent each fuzzy cluster, i. e., fuzzy rule, separately can be identified. Thus there are as many fuzzy functions as there are fuzzy clusters similar to Hathaway and Bezdek's [18] Fuzzy C-Regression model (FCRM). But Hathaway and Bezdek use membership values as the weights to be used in the estimation of the functions using weighted least squares algorithm. FCRM updates the membership values as the similarity measure by using estimation error from these functions.

"Special Fuzzy Functions", SFF, are estimated after one generates membership values of each cluster from FCM algorithm. Therefore it is structurally a new and unique approach for the determination of fuzzy system models in place of fuzzy rule bases. This the reason we call them "Special Fuzzy Functions", SFF. These "Special Fuzzy Functions" represent fuzzy rule bases in functional form and structure. When the relationship between input variables and the output variable of the system can be linearly explained in the original dimension space of the data, it is quite reasonable and faster to estimate such "Special Fuzzy Functions" using least squares estimation. When this relationship is more complex and there needs to be a non-linear transformation of the original input variables, it is better to map the input dataset into a higher dimensional space, e. g., a *hyper-space,* where the input dimension is large (maximum n). One of the powerful methods to find these "Special Fuzzy Functions" which define a linear relation between input and output variables in the higher dimension, but a non-linear relationship in the original dimension is the support vector machines which was first proposed by Vapnik [49]. For the regression case, support vector machines for regression algorithm can be applied to find these "Special Fuzzy Functions". Hence, in the next sections, we are going to specify the details of these "Special Fuzzy Functions" estimated using the Least Squares, LSE, and Support Vector machine for Regression, SVR, algorithms.

It is to be noted for the sake of emphasis that the estimated parameters of the inputs are not fuzzy sets in our proposed approach. It should be recalled that membership values and their transformations are augmented into the original selected input set as new and additional variables. In our experience, it is found that this approach is most suitable for those analysts who are familiar with a function estimation technology, e. g., the least squares technology, support vector machines, ridge regression, etc. They only need to develop an understanding of some fuzzy clustering algorithm without studying many aspects of fuzzy theory. All they have to understand is the notion of membership values and how they can be obtained from a fuzzy clustering algorithm such as FCM and/or its variations in addition to their usual background knowledge of a function estimation technique, e. g., LSE, or SVR, etc.

Thus this is a novel approach in order to provide an easy entry into fuzzy system modeling for mathematicians and statisticians who are working in industry and for other novices. For this purpose, we present next our generalization of the LSE algorithm, which includes membership values and their transformations in addition to the original scalar input variables.

Special Fuzzy Functions with LSE (SFF-LSE) Method

In Ordinary LSE (OLSE) method, the dependent variable, y, is assumed to be a linear function of one or more independent, input, variables, x, plus an error component as follows:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{nv} x_{nv} + \varepsilon \,,$$

where y is the dependent output, x_j's are the inputs (explanatory variables), for $j = 1, \ldots, nv$, nv is the number of selected inputs and ε is the independent error term which is typically assumed to be normally distributed. The goal of the least squares method is to obtain estimates of the unknown parameters, β_j's, $j = 0, 1, \ldots, nv$, which indicate how a change in one of the independent variables affects the dependent variable as follows:

$$\beta = (X'X)^{-1} X'Y \,, \tag{6}$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_{nv})$.

The proposed generalization of OLSE as **SFF-LSE**, requires that a fuzzy clustering algorithm, such as FCM [2], be available to determine the interactive (joint) membership values of input-output variables in each of the fuzzy clusters that can be identified for a given training data set.

Let (x_k, y_k), $k = 1, \ldots, nd$, be the set of observations in a training data set, such that

$$X = (x_{j,k} \mid j = 1, \ldots, nv, \;\; k = 1, \ldots, nd) \,. \tag{7}$$

First, one determines the optimal (m^*, c^*) pair for a particular performance measure, i. e., a cluster validity index, with an iterative search and an application of FCM algorithm, where m is the level of fuzziness (in our experiments, we usually take $m = 1.1, \ldots, 2.5$), and c is the number of clusters (in our experiments, we usually take $c = 2, \ldots, 10$). The well known FCM algorithm can be stated as follows:

$$\begin{aligned}
&\min J(U, V) = \sum_{k=1}^{nd} \sum_{i=1}^{c} (u_{ik})^m (\|x_k - v_i\|)_A \\
&\text{s.t.} \quad 0 \le u_{ik} \le 1 \,, \quad \forall i, k \\
&\qquad \sum_{i=1}^{c} u_{ik} = 1 \,, \quad \forall k \\
&\qquad 0 \le \sum_{k=1}^{nd} u_{ik} \le nd \,, \quad \forall i \,,
\end{aligned}$$

where J is objective function to be minimized, $\|.\|_A$ is a norm that specifies a distance-based similarity between the data vector x_k and a fuzzy cluster center v_i. In particular, $A = I$ is the Euclidean norm and $A = COV^{-1}$ is

the Mahalonobis norm, etc., where COV is the covariance matrix.

The optimal pair, (m^*, c^*), can be determined with a user defined cluster validity index, partition entropy or partition coefficient [2]. Another alternative of selecting the optimum pair would be running the overall SFF-LSE model for every (m, c) pair specified by the user and determining the optimal pair from the training RMSE values of each model. The following definitions adopt the idea of using a user defined cluster validity index for the determination of the optimal pair. The experiments in this paper follow the second alternative.

Once the optimal pair (m^*, c^*) is determined with the application of FCM algorithm, one next identifies the cluster centers for $m = m^*$ and each cluster $i = 1, \ldots, c^*$ as:

$$v_{\substack{X|Y,i \\ m^*}} = (x_{1,i}, x_{2,i}, \ldots, x_{nv,i}, y_i) \,. \tag{8}$$

From this, we identify the cluster centers of the "input space" for $m = m^*$ and $c = 1, \ldots, c^*$ as:

$$v_{\substack{X,i \\ m^*}} = (x_{1,i}, x_{2,i}, \ldots, x_{nv,i}) \,. \tag{9}$$

Next, one computes the normalized membership values of each data sample in the training data set with the use of the cluster center values determined in the previous step. There are generally two steps in these calculations:

(a) first we determine the (local) optimum membership values u_{ik}'s and then determine μ_{ik}'s that are above an α-cut in order to eliminate harmonics generated by FCM as:

$$u_{ik} = \left(\sum_{j=1}^{c} \left(\frac{\|x_k - v_{X,i}\|}{\|x_k - v_{X,j}\|} \right)^{\frac{2}{m-1}} \right)^{-1} ,$$

$$\mu_{ik} = \{u_{ik} \geq \alpha\} \,, \tag{10}$$

where μ_{ik} denotes the membership value of the kth vector, $k = 1, \ldots, nd$, in the ith rule, $i = 1, \ldots, c^*$ and x_k denotes the kth vector.

(b) next, we normalize them as:

$$\gamma_{ik}(x_k) = \frac{\mu_{ik}(x_k)}{\sum\limits_{i'=1}^{c} \mu_{i'k}(x_k)} \,, \tag{11}$$

where $\gamma_{ik}(x_k)$'s are the normalized membership values of x data sample in the ith cluster, $i = 1, \ldots, c^*$, which in turn indicate the membership values that will constitute as a new input variable in our proposed scheme of function identification for the representation of ith cluster. Let $\Gamma_i = (\gamma_{ik} | i = 1, \ldots, c^*;\ k = 1, \ldots, nd)$ be the membership values of x, a data sample, in the ith cluster, i. e., ith rule.

Next we determine a new augmented input matrix of x for each of the clusters, which could take on several forms depending on which *transformation* of membership values we want to or need to include in our system structure identification for our intended system analyses. Examples of possible augmented input matrices are:

$$\begin{aligned} X_i' &= [1, \Gamma_i, X] \ , \text{ or} \\ X_i'' &= \left[1, \Gamma_i^2, X\right] \ , \text{ or} \\ X_i''' &= \left[1, \Gamma_i^2, \Gamma_i^m, \exp(\Gamma_i), X\right] \ , \text{ etc.} \,, \end{aligned} \tag{12}$$

where X_i', X_i'', X_i''' are potential augmented input matrices to be used in least squares estimation of a new system structure identification and $\Gamma_i = (\gamma_{ik} | i = 1, \ldots, c^*;\ k = 1, \ldots, nd)$. The choice amongst X_i', X_i'', X_i''' depends on whether we want to or need to include just the membership values or some of their transformations as new input variables in order to obtain the best representation of a system behavior. A new augmented input matrix having a single input variable in the original input space when only membership values themselves are augmented to the dataset, i. e., X_i' may look like this:

$$X_i' = [1, \Gamma_i, X] = \begin{bmatrix} 1 & \gamma_{i,1} & x_{i,1} \\ \vdots & \vdots & \vdots \\ 1 & \gamma_{i,nd} & x_{i,nd} \end{bmatrix}$$

Up to this point, in the proposed system modeling approach, we have defined how the augmented input matrix for each cluster could be formed with the output of an FCM algorithm. Both the proposed SFF-LSE and SFF-SVM approaches implement these steps. From this point forward, the estimation of "Special Fuzzy Functions" takes place for each cluster, where one can implement any function estimation methodology, e. g., LSE or SMV. Different approaches are followed in the estimation of "Special Fuzzy Functions" using a augmented matrix. Here we continue to specify the SFF-LSE models. In the next section the SFF-SVM models will be introduced.

Thus the function of a single input single output model, which includes only the membership values as the additional input variable, $Y_i = \beta_{i_0} + \beta_{i_1}\Gamma_i + \beta_{i2}X$, that represents the ith rule corresponding to the ith interactive (joint) cluster in(Y_i, Γ_i, X) space, would be estimated with SFF-LSE approach as follows:

$$\beta_i^* = \left(X_i'^T X_i'\right)^{-1} \left(X_i'^T Y_i\right) \,, \tag{13}$$

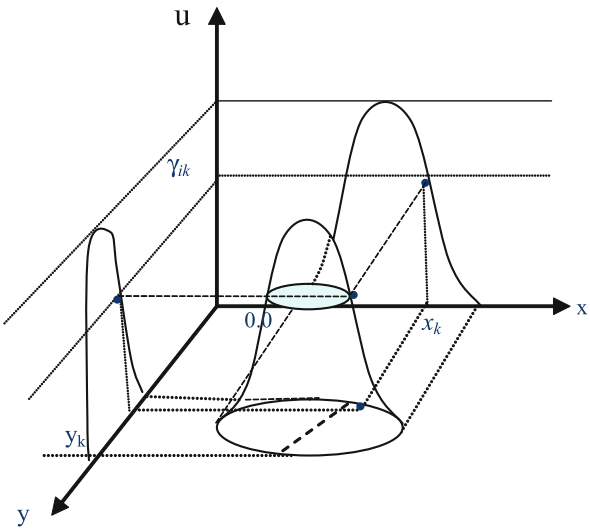

Fuzzy System Models Evolution from Fuzzy Rulebases to Fuzzy Functions, Figure 1
A fuzzy cluster in [$U \times X \times Y$] space

where $\beta_i^* = (\beta_{i0}^*, \beta_{i1}^*, \beta_{i2}^*)$ and $X_i' = [1, \Gamma_i, X]$, provided the inverse of covariance, $(X_i'^T X_i')^{-1}$, exists. The estimate of y_i would be obtained as:

$$Y_i^* = \beta_{i0}^* + \beta_{i1}^* \Gamma_i + \beta_{i2}^* X \,. \tag{14}$$

The single output value is calculated using each output value, one from each cluster, and weighting them with their corresponding membership values as follows:

$$Y_i^* = \frac{\sum_{i=1}^{c^*} \gamma_i Y_i^*}{\sum_{i=1}^{c^*} \gamma_i} \,. \tag{15}$$

Within the proposed framework, the general form of the shape of a cluster for the case of a single input variable X_j and for the ith cluster can be conceptually captured, in a stylistic, imaginary manner, say, by a second order (cone) function when one introduces the square of membership values into the augmented input matrix in the space of $[U \times X \times Y]$ which can be illustrated with a prototype shown in Fig. 1.

In a number of real life case studies, we have in fact found out that generally some second order or exponential function provide a good approximation from amongst some 20 alternatives we have experimented with in the past.

Before we specify the details of the proposed SFF-SVM method, we first review briefly the background of the support vector machines for regression algorithm.

Support Vector Machines for Regression

Support Vector Machine, SVM, is a data-mining tool to build a model of a given system. The foundations of SVM have been developed by Vapnik [49]. SVM is a type of optimization technique in which prediction error and model complexity are simultaneously minimized. Let the training samples be denoted as:

$$X = \{(\vec{x}_k, \vec{y}_k) | k = 1, \ldots, nd\} \,, \tag{16}$$

where X denotes the space of input-output patterns $\vec{x}_k = (x_{jk} | j = 1, \ldots, j_{nv},\ k = 1, \ldots, k_{nd})$ represents each input data vector, and y_k is the output value of the kth, data vector in the dataset. Support vector machines are used to solve classification problems as well as regression models where the output variable is scalar. In linear Support Vector Regression (SVR), the aim is to find a pair$(\vec{w}, b)$, where $\vec{w}$ is the weight vector, and b is the bias term in this regression equation, such that the value of the point, y_k, can be predicted according to a real-valued function:

$$f(\vec{x}_k) = \hat{y}_k = \langle \vec{w}, \vec{x}_k \rangle + b \,, \tag{17}$$

where $\langle . \rangle$ is the dot product representation. The goal is to find a function, that has at most ε deviation from the actually obtained targets, y_k, for all the training data. This concept of ε-insensitive loss function, l_ε, was first introduced by Vapnik [28] as follows:

$$l_\varepsilon = |y_k - f(\vec{x}_k)|_\varepsilon = \max\{0, |y - f(x)| - \varepsilon\} \,. \tag{18}$$

The loss function does not penalize errors below some error, $\varepsilon \geq 0$. Thus the goal of learning is to find a function with a small risk on test samples. This would mean good generalization. This type of SVR is called the ε-insensitive which embodies the Structural Risk Minimization (**SRM**) as displayed as follows:

$$R_{\text{exp}}[f] \leq R_{\text{emp}}[f] + R_{\text{complexity}}[f] \,. \tag{19}$$

In **SRM** of **SVM**, the aim is not only minimize the empirical risk from training samples, R_{emp}, but also find a simple function to minimize the complexity of the model, $R_{\text{complexity}}$. The more flat the functions are, the less complex they would be, in other words they would get simpler, and therefore they would be closer to the linear functions. The more flat function, the smaller the complexity term of the expected risk in SRM, R_{exp}, gets and the smaller would be the weight vector. In support vector regression the complexity term is expressed as weights assigned to all points in the training sample. In order to ensure that the

weight vector is small, Euclidean Norm, i. e., $\|w\|^2$ is used. In mathematical terms the objective function of SVM for regression with two conditions can be stated as follows.

$$\begin{aligned} \min_{\vec{w},b,\xi_k,\xi_k^*} \quad & \varsigma(\vec{w},\xi,\xi^*) = \tfrac{1}{2}\|\vec{w}\|^2 + \tfrac{C}{nd}\sum_{k=1}^{nd}\left(\xi_k + \xi_k^*\right) \\ \text{subject to} \quad & y_k - \langle \vec{w}, \vec{x}_k\rangle - b \le \varepsilon + \xi_k \\ & \langle \vec{w}, \vec{x}_k\rangle + b - y_k \le \varepsilon + \xi_k^* \\ & \xi_k \ge 0 \\ & \xi_k^* \ge 0\,, \end{aligned} \tag{20}$$

where ε represents the ε insensitive value which does not penalize the points whose estimated deviations are lesser/greater than ε, and $\vec{w}$, b unknown values that represent the weight vector and bias term, respectively. The $c > 0$ determines the tradeoff between the empirical error and the complexity term. The slack variables $\xi_k \ge 0$ and $\xi_k^* \ge 0$ are introduced to the model to soften the optimization problem in order to prevent infeasible solutions. The assumption in (20) is that, it is possible to find such a function that approximates all pairs $(\vec{x}_k, y_k)$ with ε precision. The optimization problem is a convex quadratic program, which can be solved by using the well-known Lagrange multiplier method. Therefore by introducing Lagrange multipliers α_i and β_i, one can construct Lagrange function, and the solution to the optimization theorem is given by the saddle point of the Lagrange function using the Karush–Kuhn–Tucker theorem [49] where the primal model is translated into dual quadratic programming problem as follows:

$$\begin{aligned} \max_{\alpha,\alpha^*} = & \frac{1}{2}\sum_{k,k'=1}^{nd}\left(\alpha_k - \alpha_k^*\right)\left(\alpha_{k'} - \alpha_{k'}^*\right)\langle \vec{x}_k, \vec{x}_{k'}\rangle \\ & -\varepsilon\sum_{k=1}^{nd}\left(\alpha_k + \alpha_k^*\right) + \sum_{k=1}^{nd}\left(\alpha_k - \alpha_k^*\right)y_k \\ \text{subject to} \quad & \sum_{k=1}^{nd}\left(\alpha_k - \alpha_k^*\right) = 0 \\ & \alpha_k, \alpha_k^* \in [0, C]\ . \end{aligned} \tag{21}$$

In model (21), we search for the parameters α_k and α_k^*, which are Lagrange multipliers. The weight vector can now be explained using the Lagrange multipliers as:

$$\vec{w} = \sum_{k=1}^{nd}\left(\alpha_k - \alpha_k^*\right)\vec{x}_k\,, \tag{22}$$

where, $\vec{x}_k$is the kth observation and α_k and $\alpha_k^*$are the Lagrange multipliers for the kth observation, and the estimation function of a vector is given as follows:

$$\begin{aligned} f(\vec{x}_{k'}) &= \sum_{k=1}^{nd}\left(\alpha_k - \alpha_k^*\right)\langle \vec{x}_{k'}, \vec{x}_k\rangle + b \\ &= \sum_{k=1}^{nd}\left(\alpha_k - \alpha_k^*\right)\vec{x}_k^{\mathrm{T}}\vec{x}_{k'} + b\,, \end{aligned} \tag{23}$$

where $\vec{x}_{k'}$is the new vector whose output is to be predicted, T represents the transpose operation on vectors and α_k and $\alpha_k^*$are the Lagrange multipliers for the kth observation.

In most cases, there is a non-linear relationship between input and output variables and a non-linear support vector regression algorithm is needed. In order to make the support vector model non-linear, the input vectors are mapped into a higher dimensional feature space using a mapping function, $\phi(x)$. However, in most of the cases, explicit mapping results in infeasible solutions that are computationally hard to obtain. The feasible way to convert a linear SMV's into non-linear SMV is to use kernel mapping which maps the input vectors into a higher dimensional feature space, i. e., $k(x, x') = \langle\phi(x), \phi(x')\rangle$ which changes the SMV algorithm as:

$$\begin{aligned} \max_{\alpha,\alpha^*} = & \frac{1}{2}\sum_{k,k'=1}^{nd}\left(\alpha_k - \alpha_k^*\right)\left(\alpha_{k'} - \alpha_{k'}^*\right)k\langle \vec{x}_k, \vec{x}_{k'}\rangle \\ & -\varepsilon\sum_{k=1}^{nd}\left(\alpha_k + \alpha_k^*\right) + \sum_{k=1}^{nd}\left(\alpha_k - \alpha_k^*\right)y_k \\ \text{subject to} \quad & \sum_{k=1}^{nd}\left(\alpha_k - \alpha_k^*\right) = 0 \\ & \alpha_k, \alpha_k^* \in [0, C]\ . \end{aligned} \tag{24}$$

Note that one may choose various different kernel functions, e. g., Gaussian radial base kernel, polynomial kernel, satisfying the Mercer's condition [49] and the output value of the k'th input vector is calculated using the following function:

$$\hat{y}_{k'} = \hat{f}\left(\vec{x}_{k'}, \alpha, \alpha^*\right) = \sum_{k=1}^{nd}\left(\alpha_k - \alpha_k^*\right)K\langle \vec{x}_{k'}, \vec{x}_k\rangle + b, \tag{25}$$

where the $K\langle \vec{x}_{k'}, \vec{x}_k\rangle$ represents the kernel mapping of the input vectors. From Eq. (25), the dependent variable is estimated using the kernel mapping of the input vectors and their Lagrange multipliers of each vector that are calculated from the optimization algorithm. Note that in Eq. (25), it is not required to calculate the weight vector

explicitly, the input vectors who's Lagrange multipliers are not zero are used in estimating the output and they are called the "**support vectors**". In a sense, the complexity of the functions, i. e., the $\frac{1}{2}\|\vec{w}\|^2$ term, represented by support vectors is independent of the dimensionality of the input space x, and only depends on the number of support vectors.

An additional result of the application of Karush–Kuhn–Tucker (KKT) theorem in SVR is:

$$\begin{aligned}(C-\alpha_k)\xi_k &= 0\\ (C-\alpha_k^*)\xi_k^* &= 0\,.\end{aligned} \tag{26}$$

One of the several conclusions one might make from (26) is that $\alpha_k\alpha_k^* = 0$, i. e., there can never be a set of dual variables for an observation k which are both simultaneously non-zero as this would require non-zero slacks in both directions. Since $c > 0$, then $\xi_k\xi_k^* = 0$ must also be true. In the same sense, there can never be two slack variables $\xi_k, \xi_k^* > 0$ which are both non-zero and equal.

Special Fuzzy Functions with SVM (FF-SVM) Method

As one can build ordinary least squares for the estimation of the Special Fuzzy Functions when the relationship between input variables and the output variable can be linearly defined in the original input space, one may also build support vector regression models to estimate the parameters of the non-linear Special Fuzzy Functions using support vector regression methods. The augmented input matrix is determined from FCM algorithm such that there is one SVR in SFF-SVM for each cluster same as SFF-LSE model. One may choose any membership transformation depending on the input dataset. Then one can apply support vector regression, SVR, algorithm instead of LSE to each augmented matrix, which are comprised of the original selected input variables and the membership values and/or their transformations. Support vector machines' optimization algorithm is applied to each augmented matrix of each cluster (rule) i, $i = 1, \ldots, c^*$, to optimize their Lagrange multipliers, α_{ik} and α_{ik}^*, and find the candidate support vectors, $k = 1, \ldots, nd$. Hence, using SFF-SVM, one finds Lagrange multipliers of each kth train data sample one for each cluster, i. Then the output value of kth data sample in ith cluster is estimated using the Equation (27) as follows:

$$\begin{aligned}\hat{y}_{ik'}^* &= \hat{f}\big(\vec{x}_{ik'}, \alpha_i, \alpha_i^*\big)\\ &= \sum_{k=1}^{nd}\big(\alpha_{ik}-\alpha_{ik}^*\big)K\langle\vec{x}_{ik'}, \vec{x}_{ik}\rangle + b_i\,.\end{aligned} \tag{27}$$

Where the $\hat{y}_{ik'}^*$ is the estimated output of the k'th vector in ith cluster which is estimated using the support vector regression function with the Lagrange multipliers of the ith cluster. The augmented kernel matrix denotes the kernel mapping of the augmented input matrix (as described in SFF-LSE approach) where the membership values and their transformations are used as additional input variables. After the optimization algorithm finds the optimum Lagrange multipliers, one can estimate the output value of each data point in each cluster using Eq. (27).

The inference structure of SFF-SVM is adapted from the Special Fuzzy Functions with least squares where one can estimate a single output of a data point (see Eq. (15) by taking the membership value weighted averages of its output values calculated for each cluster using Eq. (27).

Future of Fuzzy System Models

In the future, fuzzy system models are expected to be structured by type 2 fuzzy sets. For this purpose, we next present basic definitions.

Basic Definitions

Definition 1 A type 2 fuzzy set $\tilde{A}$ on universe of discourse, x, is a fuzzy set which is characterized by a fuzzy membership function, $\tilde{\mu}_A(x)$, where $\tilde{\mu}_A(x)$ is a mapping as shown below:

$$\tilde{\mu}_A(x)\colon X \to [0,1]^{[0,1]}\,. \tag{28}$$

Then type 2 fuzzy set, $\tilde{A}$, can be characterized as follows:

$$\tilde{A} = \{(x,(u, f_x(u)))|x \in X,\ u \in J_x \subseteq [0,1]\}\,. \tag{29}$$

Where u is defined as the primary membership value, $J_x \subseteq [0,1]$ is the domain of u and $f_z(u)$ is the secondary membership function. An alternative definition of type 2 fuzzy set $\tilde{A}$, used by Mendel [28] and inspired by Mizumoto and Tanaka [30], can be given as follows:

Given that x is a continuous universe of discourse, the same type 2 fuzzy set can be defined as:

$$\tilde{A} = \int_{x\in X}\left[\int_{u\in J_x} f_x(u)/u\right]/x\,. \tag{30}$$

And for discrete case, the same type 2 fuzzy set can be defined as:

$$\tilde{A} = \sum\nolimits_{x\in X}\left[\sum\nolimits_{u\in J_x} f_x(u)/u\right]/x\,. \tag{31}$$

The secondary membership function, $f_x(u)$,can be defined as follows:

Definition 2 Secondary membership function, $f_x(u)$, is a function that maps membership values of universe of discourse x onto unit interval $[0,1]$. Thus, $f_x(u)$ can be characterized as follows:

$$f_x(u)\colon J_x \to [0,1], \forall u \in J_x \subseteq [0,1]\,, \quad \forall x \in X\,. \tag{32}$$

With the secondary membership function defined as above, the membership function of type 2 fuzzy set $\tilde{A}$, $\tilde{\mu}_A(x)$, can then be written for continuous and discrete cases, respectively, as follows:

$$\tilde{\mu}_A(x) = \int_{u\in J_x} f_x(u)/u\,, \quad \forall x \in X\,, \tag{33}$$

$$\tilde{\mu}_A(x) = \sum\nolimits_{u\in J_x} f_x(u)/u\,, \quad \forall x \in X\,. \tag{34}$$

An Interval Valued Type 2 Fuzzy Set (IVT2FS), which is a special case of type 2 fuzzy set, can be defined as follows:

Definition 3 Let $\tilde{A}$ be a linguistic label with type-2 membership function on the universe of discourse of base variable x, $\tilde{\mu}_A(x)\colon X \to f_x(u)/u, u \in J_x, J_x \subseteq [0,1]$. The following condition needs to be satisfied in order to consider $\tilde{\mu}_A(x)$ as an interval value type 2 membership functions:

$$f_x(u) = 1, \forall x \in X, \forall u \in J_x, J_x \subseteq [0,1]\,. \tag{35}$$

Thus, the interval valued type 2 membership function is a mapping as shown below:

$$\tilde{\mu}_A(x) : X \to 1/u, u \in J_x, J_x \subseteq [0,1]\,. \tag{36}$$

General Structure of Type 2 Fuzzy System Models

In a series of papers, Mendel, Karnik and Liang [22,23,24,25] extended traditional type 1 inference methods such that these methods can process type 2 fuzzy sets. These studies were explained thoroughly by Mendel in [28]. The classical Zadeh and Takagi–Sugeno type 1 models are modified as type 2 fuzzy rule bases (T2Z-FR and T2TS-FR,), respectively, as follows:

$$\underset{i=1}{\overset{c^*}{\text{ALSO}}}\left[\text{IF } \underset{j=1}{\overset{NV}{\text{AND}}}\left(x_j \in X_j \text{ isr } \tilde{A}_{ji}\right) \text{ THEN } y \in Y \text{ isr } \tilde{B}_i\right], \tag{37}$$

$$\underset{i=1}{\overset{c^*}{\text{ALSO}}}\left[\text{IF } \underset{j=1}{\overset{NV}{\text{AND}}}\left(x_j \in X_j \text{ isr } \tilde{A}_{ji}\right) \text{ THEN } y_i = a_i x^{\mathrm{T}} + b_i\right]. \tag{38}$$

Mendel, Karnik and Liang [22,23,24,25] assumed that the antecedent variables are separable (i. e., non-interactive). After formulating the inference for a full type 2 fuzzy system model, Karnik et al. [22,23,24,25] simplified their proposed methods for the interval values case. In order to identify the structure of the IVT2FS, it was assumed that the membership functions are Gaussian. A clustering method was utilized to identify the mean parameters for the Gaussian functions. However, the clustering method has not been specified. It was assumed that the standard error parameters of the Gaussian membership functions are exactly known. The number of rules was assigned as eight due to the nature of their application. However, the problem of finding the suitable number of rules was not discussed in the paper.

Liang and Mendel [26] proposed another method to identify the structure of IVT2FS. It has been suggested to initialize the inference parameters and to use steepest-descent (or other optimization) method in order to tune these parameters of an IVT2FS. Two different approaches for the initialization phase have been suggested in [26]. Partially dependent approach utilizes a type 1 fuzzy system model to provide a baseline for the type 2 fuzzy system model design. Totally independent approach starts with assigning random values to initialize the inference parameters. Liang and Mendel [26] indicated that the main challenge in their proposed tuning method is to determine the active branches.

Mendel [28] indicated that several structure identification methods, such as one-pass, least-squares, back-propagation (steepest descent), singular value-QR decomposition, and iterative design methods, can be utilized in order to find the most suitable inference parameters of the type 2 fuzzy system models. Mendel [28] provided an excellent summary of the pros and cons of each method.

Several other researchers such as Starczewski and Rutkowski [37], John and Czarnecki [20,21] and Chen and Kawase [10] worked on T2-FSM. Starczewski and Rutkowski [37] proposed a connectionist structure to implement interval valued type 2 fuzzy structure and inference. It was indicated that methods such as, back propagation, recursive least squares or Kalman algorithm-based methods, can be used in order to determine the inference parameters of the structure. John and Czarnecki [20,21] extended the ANFIS structure such that it can process the type 2 fuzzy sets.

All of the above methods assume non-interactivity between antecedent variables. Thus, the general steps of the inference can be listed as: fuzzification, aggregation of the antecedents, implication, and aggregation of the consequents, type reduction and defuzzification.

With the general structure of type 2 fuzzy system models and inference techniques, we next propose discrete interval valued type 2 rule base structures.

Discrete Interval Valued Type 2 Fuzzy Sets (DIVT2FS)

In Discrete Interval Valued Type 2 Fuzzy Sets (DIVT2FS) [48] the domain of the primary membership is discrete and the secondary membership values are fixed to 1. Thus, the proposed method is utilizing discrete interval valued type 2 fuzzy sets in order to represent the linguistic values assigned to each fuzzy variable in each rule. These fuzzy sets can be mathematically defined as follows:

$$\tilde{A} = \int_{x \in X} \left[\sum\nolimits_{u \in J_x} 1/u \right] / x \,, \tag{39}$$

where $x \in X \subseteq \Re, u \in J_x = \{J_{xr}\}$, $r = 1, \ldots, NM$, x is a system variable with continuous domain, u is the primary membership value with discrete domain and J_{xr} is the rth primary membership value associated with x. The discrete interval valued type 2 fuzzy set $\tilde{A}$, can be considered as the union of type 1 fuzzy sets a^r, $r = 1, \ldots, nm$, where nm is the number of embedded type 1 fuzzy sets that form the "discrete interval valued type 2 fuzzy set $\tilde{A}$". (for the purposes of this paper, we assume that nm is known. Currently we are conducting further research to determine nm at the Knowledge-Intelligence Laboratory, University of Toronto). Consequently, the membership function of discrete interval type 2 fuzzy set $\tilde{A}$, $\mu_{\tilde{A}}(x)$, can be represented as a collection of type 1 membership functions as follows:

$$\mu_{\tilde{A}}(x) = \{\mu_A^r(x)\} \,, r = 1, \ldots, NM \,, \tag{40}$$

where nm is the number of discrete membership values assigned to each value of system variable x, $\mu_A^r(x)$ is the membership function associated with rth embedded type 1 fuzzy set, which can be defined as follows:

$$\mu_A^r(x) \colon X \to J_{xr}, x \in X \subseteq \Re \,. \tag{41}$$

One of the goals of this study is to eliminate the non-interactivity assumption used in the existing type 2 fuzzy system models. Hence, we have extended the fuzzy rule base structure proposed by Delgado et al. [12], Babuska et al. [1] and Uncu and Türkşen [48]. Thus, the proposed Discrete Interval Type 2 Zadeh Fuzzy Rule base (DIT2-Z-FR) structure can be written as follows:

$$R\colon \left\{ \underset{i=1}{\overset{c^*}{\text{ALSO}}} \left(\text{IF } x \in X \text{ isr } A_i^r \text{ THEN } y \in Y \text{ isr } B_i^r \right) \right\}, \quad r = 1, \ldots, NM \,, \tag{42}$$

where, A_i^r is the rth embedded type 1 multidimensional fuzzy set associated with the antecedent of the ith rule. A_i^r will be represented with the membership function $\mu_i^r(x)$ in the membership value domain. Similarly, B_i^r is the rth embedded type 1 fuzzy set associated with the consequent of the ith rule. B_i^r will be represented with the membership function $\mu_i^r(y)$ in the membership value domain.

The rule base structure given in (42) can be thought as a collection of embedded type 1 fuzzy rule bases. Hence, the proposed fuzzy rule base structure is named as Discrete Interval Type 2 Fuzzy Rule base (DIT2-Z-FR) structure.

The other well-known fuzzy rule base structures are also extended to type 2 by using the above idea. Hence, Discrete Interval Valued Type 2 Takagi–Sugeno Fuzzy Rule base (DIT2-TS-FR) can be written respectively as follows:

$$R\colon \left\{ \underset{i=1}{\overset{c^*}{\text{ALSO}}} \left(\text{IF } x \in X \text{ isr } A_i^r \text{ THEN } y = a_i^r x^{\mathrm{T}} + b_i^r \right) \right\}, \quad r = 1, \ldots, NM \,, \tag{43}$$

where, $a_i^r = \left(a_{i,1}^r, \ldots, a_{i,NV}^r\right)$ is the regression coefficient vector associated with the ith rule of the rth embedded type 1 fuzzy system model and b_i^r is the scalar associated with the ith rule of the rth embedded type 1 fuzzy system model. As one can observe the Takagi–Sugeno structure is not only extended by representing the antecedent fuzzy sets with discrete type 2 fuzzy sets but also by letting uncertainty in crisp inference parameters in consequents. Thus, the problem of building type 2 fuzzy system models is reduced to finding embedded type 1 fuzzy system models.

Discrete Interval Valued Type 2 "Special Fuzzy Functions", (SFF)

In a similar manner to Discrete Interval Valued Type 2 Fuzzy Rule bases, we could construct Discrete Interval Valued Type 2 "Special Fuzzy Functions" for the cases of SFF-LSE and SFF-SVR. For example, the function of a single input single output model, which includes only the membership values as the additional input variable, $Y_i^r = \beta_{i0}^r + \beta_{i1}^r \Gamma_i^r + \beta_{i2}^r X$, $r = 1, \ldots, NM$ that represents the rth "Special Fuzzy Function" to the ith interactive (joint) type 2 cluster in (Y_i^r, Γ_i^r, X^r), $i = 1, \ldots, c^*$, $r = 1, \ldots, NM$ space, would be estimated with SFF-LSE approach as follows:

$$\beta_i^{r^*} = \left(X_i^{r'T} X_i^{r'} \right)^{-1} \left(X_i^{r\mathrm{T}} Y_i^r \right), \quad r = 1, \ldots, NM \,, \tag{44}$$

where $\beta_i^{r^*} = (\beta_{i0}^{r}{}^*, \beta_{i1}^{r^*}, \beta_{i2}^{r^*})$ and $X_i^{r'} = \left[1, \Gamma_i^r, X\right]$, provided the inverse of covariance, $(X_i^{r'T} X_i^{r'})^{-1}$, exists. The

estimate of y_i would be obtained as:

$$Y_i^{r^*} = \beta_{i0}^{r^*} + \beta_{i1}^{r^*} \Gamma_i^r + \beta_{i2}^{r^*} X \,. \tag{45}$$

The single output value is calculated using each output value, one from each cluster, and weighting them with their corresponding membership values as follows:

$$Y_i^{r^*} = \frac{\sum_{i=1}^{c^*} \gamma_i^r Y_i^{r^*}}{\sum_{i=1}^{c^*} \gamma_i^r}, \quad i = 1, \ldots, c^*, \quad r = 1, \ldots, NM. \tag{46}$$

The development of such Discrete Interval Valued Type 2 "Special Fuzzy Functions" with LSE will be studied in our future investigations. In a similar manner, we propose to develop Discrete Interval Valued Type 2 "Special Fuzzy Functions" with SVM methodology.

Case Study Applications

In order to test the proposed model performances as opposed to fuzzy rule base systems three input-output datasets are considered in this investigation. These are:

(i) Daily price of a stock in stock market.
(ii) Customer Income Prediction model for a major bank.
(iii) The amount of chemicals for a desulphurization process for a steel processing company.

The specifications of each datasets are displayed in the following parts:

Daily Stock Price Dataset

Daily stock price dataset comprises of the daily trend data of stock prices. This dataset was introduced by Sugeno and Yasukawa [38]. The same dataset has been used in various other studies one of which compares six different fuzzy reasoning methods using this dataset.

Out of 100 observations, 50 of them are used for the training purposes and the other 50 was hold-out for testing purposes. There were originally 11 input variables and single output variable in the dataset [48]. Preliminary input selection was applied using Random Forests (RF) method, [48] which estimates variable importance. Based on the results of RF method, only 4 of input variables i. e., x_2, x_4, x_8, x_{10}, were found to have importance on the output variable. The rest of the variables had insignificant effect on the output.

Income Prediction Dataset

The purpose of the Income Prediction Model was to predict the income of the future customers based on the current customer information and 1996 year census data. There were more than 200 variables and hundred of thousands of customers. According to business needs, the dataset was partitioned into 9 different parts based on age, number of different types of investment accounts hold by the customer and different regions of residency. In this study, only one partition was investigated.

We have only selected 10% of the i.i.d. data samples to do our research on using only single partition explained above. The data was cleaned form the outliers using the expert's knowledge and 900 training and 900 testing observations are selected randomly. The dataset was comprised of 11 input variables [48]. Based on the correlation analysis, 3 input variables were discarded from the dataset resulting in 8 input variables. There were no census variables among the selected input variables.

Desulphurization Dataset

A torpedo car desulphurization facility removes sulfur from hot metal leaving the blast furnaces before it is sent to the next process. Generally, desulphurization is carried out by injecting two different powered reagents directly into the hot metal via a lance. The reagents react with the sulfur in the hot metal and residue, which is rich in sulfur, is separated from the iron.

The aim of the data-mining project was to determine the right amounts of the reagents to be added into the hot metal. These regents are expensive materials and precise estimation is required. There are vast quantities of data available on the desulphurization process, which has various characteristics. The original input and output variables are shown in [27]. There are 750 training and 900 verification samples used in the experiments. Based in the variable selection using random forest regression method, only 5 variables are found to be important.

Experimental Design

Using the three different input-output datasets, we build four different fuzzy system model structures, SFF-LSE, SFF-SVM, SY-FRB and TS-FRB. In order to keep the consistency between each model structure, the same training and testing datasets are used for the four fuzzy system models with the same input variables. The categorical variables are transformed into probabilities using logistic regression and are used as additional inputs only in Income Prediction and Desulphurization Datasets in all of the four models.

Sugeno–Yasukawa Models and Takagi–Sugeno Models

The proposed FSM models are compared to two well known Fuzzy Rule Base Models: (i) Sugeno and Yasukawa's fuzzy logic based approach using Partition Type Fuzzy Model, **SY-FRB,** [38] (ii) Takagi and Sugeno's fuzzy system modeling approach, **TS-FRB** [39]. In Sugeno and Yasukawa's FSM approach, they use linguistic variables for both the consequent and the antecedent part of the fuzzy rules and the system learns all inference parameters from the data without the expert intervention. The variable selection method defined in their paper is not applied to these 3 datasets in order to compare the models on the same basis.

In Takagi and Sugeno's FRB (TS-FRB) structure [39], they assume that the antecedent membership functions are to be characterized with triangular membership functions. In their approach, each input variable space is assumed to be partitioned into two clusters and logical connective AND is taken as MIN. Then, the structure identification problem is just to identify the regression equation coefficients for each rule and the antecedent parameters for each input variable in each rule. Researchers proposed several structure identification methods to identify the membership functions from the data, e. g., Delgado et al. [12], Babuska and Verbruggen, etc. In this paper we have used Babuska et al.'s [1] modified Takagi–Sugeno study where the membership functions of the antecedents are identified using fuzzy c-means clustering and projected onto each input vector. The degree of fulfillment of each rule is then calculated using a t-norm operator, i. e. product. Only one aggregate input membership function is identified for each rule. The inference parameters are same as the traditional Takagi–Sugeno inference method [39].

Special Fuzzy Functions Models (SFFM)

In this paper, the modeling performance from the Special Fuzzy Functions with LSE and SVM models, SFF-LSE and SFF-SVM, are compare to the Fuzzy Rule Base structures. Instead of using cluster validity indices to select the optimum model parameters, we measured the optimum model based on the best model performance using RMSE by applying a grid search for each parameter. Note that fuzzy functions with LSE system models have 2 parameters, which are the FCM parameters, i. e., degree of fuzziness and cluster size. Special Fuzzy Functions with SVR models have 4 parameters: FCM parameters and the SVR parameters which are *C-regularization* and *ε-insensitive value (epsilon).*

In both of these special fuzzy function models, membership values and their exponential transformations are used as additional input variables. We applied the LIBSVM program [6] within our special fuzzy function codes for support vector optimization in estimating the "Special Fuzzy Functions". We chose Gaussian RBF as the kernel function, $K(x, x') = exp(-\gamma\|x - x'\|^2)$ in all the experiments and the default values of the kernel parameters in LIBSVM [10] are used. Recall that the SVR regression has two parameters that is set by the user: ε-insensitive zone (*epsilon or* ε) and the regularization parameter, C). The parameters of SVM regression, C and ε, are generated by the grid search, i. e., $C = \{2^{-3}, 2^{-1}, \ldots, 2^3, 2^5\}$, $epsilon(\varepsilon) = \{0.1, \ldots, 0.5\}$, as well as the FCM parameters, i. e., $c = 3, 5, \ldots, 10$, $m = 1.1, \ldots, 2.5$ in all the experiments. Hence, for the fuzzy functions with LSE models 2 parameters are specified for each model, and for the fuzzy functions using SVM models, 4 parameters, m, c, C, and γ are determined using a grid search where the model performance of each model is determined using Root Mean Square Error of the models as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{47}$$

y_i, and $\hat{y}_i$ are the actual and estimated output values of a single observation, N is the total number of observations in the dataset.

Experimental Model Results

The 4 fuzzy system models are applied on the daily stock price of a stock market, income prediction, and reagents estimation in desulphurization process datasets. The results are displayed in Tables 1,2 and 3.

Special Fuzzy Function (SFF) models, when estimated with either *SVM or LSE* algorithms, show better generalization than the fuzzy rule base models. The fuzzy function models can increase the model performance by up to 35% depending on the dataset.

Table 4 displays the optimum parameters of the models from each experiment whose results are displayed in Tables 1–3. In Table 4, *C-reg* indicates the regularization parameter, ε is the ε-insensitive region (epsilon), m refers

Fuzzy System Models Evolution from Fuzzy Rulebases to Fuzzy Functions, Table 1
Daily Stock Price of a Stock in Stock Market

	SFF-SVM	SFF-LSE	TS-FRB	SY-FRB
RMSE(train)	2.43	3.82	2.76	7.16
RMSE(test)	3.64	5.61	5.68	9.93

Fuzzy System Models Evolution from Fuzzy Rulebases to Fuzzy Functions, Table 2
Income Prediction Dataset*

	SFF-SVM	SFF-LSE	TS-FRB	SY-FRB
RMSE(train)	0.40	0.52	0.49	0.58
RMSE(test)	0.64	0.64	0.80	0.70

* The RMSE values are calculated from standardized output values.

Fuzzy System Models Evolution from Fuzzy Rulebases to Fuzzy Functions, Table 3
Reagent Estimation for Desulphurization Process

	SFF-SVM	SFF-LSE	TS-FRB	SY-FRB
Reagent1				
RMSE(train)	30	40	35	69
RMSE(test)	42	45	45	72
Reagent2				
RMSE(train)	4.80	6.49	5.62	10.01
RMSE(test)	6.59	7.19	7.19	10.80

Fuzzy System Models Evolution from Fuzzy Rulebases to Fuzzy Functions, Table 4
Optimum Model Parameters of three datasets

Dataset	Model Type	Optimum model Parameters
Daily Stock Price	SFF-SVM	$C-reg = 32$, $\varepsilon = 0.2$, $c = 8$, $m = 1.9$, $\#sv = 28(\sigma_{sv} = 1.5)$
	SFF-LSE	$c = 8$, $m = 1.6$
Income Prediction	SFF-SVM	$C-reg = 64$, $\varepsilon = 0.2$, $c = 7$, $m = 1.2$, $\#sv = 264(\sigma_{sv} = 5.7)$
	SFF-LSE	$c = 8$, $m = 1.6$
Desulphur-ization	SFF-SVM	Reagent 1: $C-reg = 64$, $\varepsilon = 0.2$, $c = 7$, $m = 1.4$, $\#sv = 121(\sigma_{sv} = 4.8)$ Reagent 2: $C-reg = 64$, $\varepsilon = 0.2$, $c = 7$, $m = 1.5$, $\#sv = 101(\sigma_{sv} = 7.7)$
	SFF-LSE	Reagent 1: $c = 5$, $m = 1.4$ Reagent 2: $c = 6$, $m = 1.5$

to the degree of fuzziness (weighting exponent) of the fuzzy c-mean clustering algorithm, c indicates the number of cluster, and $\#sv$ refers to the average number of support vectors from each support vector regression model build for each cluster of the *SFF-SVM* models. In order to show the dispersion of the number of support vectors among each cluster we also included the standard deviation (σ_{sv}) of the support vectors of the optimum models.

The grid search algorithms applied in this paper try to find the best RMSE value from training data in each experiment and assign these parameters as the optimum model parameters. The algorithm searches for the minimum regression error. Then, verification dataset output is inferred using the optimum parameters. The issue with these grid search algorithms is that, sometimes, the models get stuck in the local minimum which is smaller than the global minimum and this might cause generalization problems. An example to this concept is shown in income prediction dataset (Table 2). The model parameters best fit to the training data when FF-SVM is used but this causes generalization problems. It should also be reminded that, when there is a linear relationship between the inputs and the output, then LSE model performances will be as good as the other model performances. On the other hand, FF-LSE models, in three of the datasets, show more reliable results than the SVM models. One should run both models and determine the optimum model parameters after observing the results from both models.

Conclusions and Future Directions

In this paper, we have outlined basic well known three type 1 system models of the recent past and suggested that type 2 fuzzy system models are likely to be studied more extensively in the future. In particular, we have reviewed: (1) Type 1 Fuzzy Rule bases and (2) Type 1 "Special Fuzzy Functions". Furthermore we discussed the Type 2 Fuzzy Rule bases. But we left the Type 2 "Special Fuzzy Functions" for a future study after providing a structure for their development . As well, we have demonstrated that "Type 1 Special Fuzzy Functions" provide better results then "Type 1 Fuzzy Rule Base" models in three specific case studies.

Bibliography

Primary Literature

1. Babuska R, Verbruggen HB (1997) Constructing fuzzy models by product space clustering. In: Hellendoorn H, Driankov D (eds) Fuzzy model identification: selected approaches. Springer, Berlin, pp 53–90
2. Bezdek JC (1973) Fuzzy mathematics in pattern classification. Ph.D Thesis, Applied Mathematics Center. Cornell University, Ithaca
3. Celikyilmaz A, Türkşen IB (2007) Fuzzy functions with support vector machines. Inf Sci 177:5163–5177
4. Celmins A (1987) Least squares model fitting to fuzzy vector data. Fuzzy Sets Syst 22:245–269
5. Celmins A (1987) Multidimensional least squares model fitting of fuzzy models. Math Model 9:669–690
6. Chang C, Lin C (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

7. Chang PT, Lee ES (1994) Fuzzy linear regression with spreads unrestricted in sign. Compt Math Appl 28:61–71
8. Chang YHO, Ayyub BM (1993) Reliability analysis in fuzzy regression. In: Proc. Annual Conf. of NAFIPS'93. Allentown, IEEE, New York, pp 93–97
9. Chen MS, Wang SW (1999) Fuzzy clustering analysis for optimizing fuzzy membership functions. Fuzzy Sets Syst 103(2):239–254
10. Chen Q, Kawase S (2000) On fuzzy-valued fuzzy reasoning. Fuzzy Sets Syst 113:237–251
11. Chiu S (1994) Fuzzy model identification based on cluster estimation. J Intell Fuzzy Syst 2(3)267–278
12. Delgado M, Gomez–Skermata AF, Martin F (1997) Rapid prototyping of fuzzy models". In: Hellendoorn H, Driankov D (eds) Fuzzy model identification: selected approaches. Springer, Berlin, Germany, pp 53–90
13. Demirci M (1999) Fuzzy functions and their fundamental properties. Fuzzy Sets Syst 106:239–246
14. Demirci M (2003) Foundations of fuzzy functions and vague algebra based on many-valued equivalence relations, Part I: fuzzy functions and their applications. IJ Gen Syst 32:123–155
15. Demirci M, Recasens J (2004) Fuzzy groups, fuzzy functions and fuzzy equivalence relations. Fuzzy Sets Syst 144:441–458
16. Diamond P (1998) Fuzzy least squares. Inf Sci 46:141–157
17. Emami MR, Türkşen IB, Goldenberg AA (1998) Development of a systematic methodology of fuzzy logic modeling. IEEE Tran Fuzzy Syst 63(3):346–361
18. Hathaway RJ, Bezdek JC (1993) Switching regression models and fuzzy clustering. IEEE Trans Fuzzy Syst 1(3):195–203
19. Jang JSR (1993) Anfis: adaptive-network-based fuzzy inference systems. IEEE Trans Syst Man Cybern 23(3):665–685
20. John RI, Czarnecki C (1998) A type 2 adaptive fuzzy inference system. In: Proc. IEEE Conf. Systems, Man and Cybernetics, vol 2. IEEE, New York, pp 2068–2073
21. John RI, Czarnecki C (1999) An adaptive type-2 fuzzy system for learning linguistic membership grades. In: Proc. IEEE International Fuzzy Systems Conference, vol 3. IEEE, New York, pp 1552–1556
22. Karnik NN, Mendel JM (1998) Introduction to type-2 fuzzy logic systems. In: Proc. IEEE Conf. On computational intelligence, vol 2. IEEE, New York, pp 915–920
23. Karnik NN, Mendel JM (1998) Type-2 fuzzy logic systems: type reduction. In: Proc. IEEE Conf. On Systems, Man and Cybernetics, vol 2. IEEE, New York, pp 2046–2051
24. Karnik NN, Mendel JM, Liang Q (1999) Type-2 fuzzy logic systems. IEEE Trans On Fuzzy Syst 7(6):643–658
25. Karnik NN, Mendel JM (2000) Applications of type-2 fuzzy logic systems: handling the uncertainty associated with surveys. In: Proc. IEEE Conf. On Fuzzy Systems, vol 3. pp 1546–1551
26. Liang Q, Mendel JM (2000) Interval type-2 fuzzy logic systems: theory and design. IEEE Trans On Fuzzy Syst 8(5):535–550
27. Mamdani EH, Assilian S (1981) An experiment in linguistic syntesis with a fuzzy logic controller. In: Mamdani EH, Gains BR (eds) Fuzzy Reasoning and Its Applications. Academic Press, New York
28. Mendel JM (2001) Uncertain rule-based fuzzy logic systems: introduction and new directions. Prentice, Upper Saddle River
29. Mizumoto M (1989) Method of fuzzy inference suitable for fuzzy control. J Soc Instrum Control Eng 58:959–963
30. Mizumoto M, Tanaka K (1976) Some properties of fuzzy sets of type 2. Inf Control 31:312–340
31. Nakanishi H, Türkşen IB, Sugeno M (1993) A review and comparison of six reasoning methods. Fuzzy Sets Syst 57: 257–295
32. NR Pal, Bezdek JC (1995) On cluster validity for the fuzzy c-means model. IEEE Trans Fuzzy Syst 3(3)370–379
33. Rutkowska D (2002) Type 2 fuzzy neural networks: an interpretation based on fuzzy inference neural networks with fuzzy parameters. In: Proc. IEEE Conf. On Fuzzy Systems, vol 2. IEEE, New York, pp 1180–1185
34. Savic D, Pedryzc W (1991) Evolution of fuzzy linear regression models. Fuzzy Sets Syst 39:51–63
35. Smola AJ, Scholkopf B (1998) A tutorial on support vector regression. NeuroColT2 Technical Report Series, NC2-Tr-1998–030
36. Sproule BA, Bazoon M, Shulman KI, Turkşen IB, Naranjo CA (2000) Fuzzy logic pharmacokinetic modeling: an application to lithium concentration prediction. Clinical Pharmacology Therapy 62:29–40
37. Starczewski J, Rutkowski L (2002) Connectionist structures of type 2 fuzzy inference systems. In: Wyrzykowski R et al (eds) PPAM 2001, LCNS 2328. Springer, Heidelberg, pp 634–642
38. Sugeno M, Yasukawa T (1993) A fuzzy logic based approach to qualitative modeling. IEEE Trans Fuzzy Syst 1:7–31
39. Takagi T, Sugeno M (1985) Fuzzy identification of systems and its applications to modeling and control. IEEE Trans Syst Man Cybern SMC-15(1):116–132
40. Tanaka H, Ishibuchi H, Yoshikawa S (1995) Exponential possibility regression analysis. Fuzzy Sets Syst 69:305–318
41. Tanaka H, Vegima S, Asai K (1982) Linear regression analysis with fuzzy model. IEEE Trans Syst Man Cybern SMC-2:903–907
42. Türkşen IB (1986) Interval valued fuzzy sets based on normal forms. Fuzzy Sets Syst 20:191–210
43. Türkşen IB (1992) Interval-valued fuzzy sets and 'compensatory AND'. Fuzzy Sets Syst 51:295–307
44. Türkşen IB (1995) Fuzzy normal forms. Fuzzy Sets Syst 69: 319–346
45. Türkşen IB (2002) Type 2 representation and reasoning for cww. Fuzzy Sets Syst 127:17–36
46. Türkşen IB (2008) Fuzzy Functions with LSE. Applied Soft Computing 8(3):1178–1182
47. Uncu Ö, Türkşen IB (2007) A novel feature selection approach: combining feature wrappers and filters. Inf Sci 177:449–466
48. Uncu Ö, Türkşen IB (2007) Discrete Interval Type 2 Fuzzy System Models Using Uncertainty in Learning Parameters. IEEE Fuzzy Syst 15(1):90–106
49. Vapnik NV (1998) Statistical learning theory. Wiley, New York
50. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning. Inf Sci 8:199–249
51. Zimmermann HJ, Zysno P (1980) Latent connectives in human decision-making. Fuzzy Sets Syst 4:37–51

Books and Reviews

Kilic K (2002) A proposed fuzzy system modeling algorithm with an application in pharmacokinetic modeling. Ph.D Thesis, Department of Mechanical and Industrial Engineering. University of Toronto, Toronto

G

Game Theory, Introduction to

Marilda Sotomayor[1,2]
[1] Department of Economics, University of São Paulo, São Paulo, Brasil
[2] Department of Economics, Brown University, Providence, USA

Game theory is the study of decision problems which involves several individuals (the decision-makers or players) interacting rationally. The models of game theory are abstract representations of a number of real-life situations and have applications to economics, political sciences, computer sciences, evolutionary biology, social psychology, law, among others. These applications are also important for the development of the theory, since the questions that emerge may lead to new theoretic results.

This session is an attempt to provide the main features of Game Theory, covering most of the fundamental theoretical aspects under the cooperative, non-cooperative and "general" or "mixed" approaches.

The cooperative approach focuses on the possible outcomes of the decision-maker's interaction by abstracting from the actions or decisions that may lead to these outcomes. Specifically, Cooperative game theory studies the interactions among coalitions of players. Its main question is: Given the sets of feasible payoffs for each coalition, what payoff will be awarded to each player? One can take a positive or normative approach to answering this question, and different solution concepts in the theory lead towards one or the other.

The first cooperative solution concept is the von Neumann–Morgenstern stable sets, treated in ► Cooperative Games (Von Neumann–Morgenstern Stable Sets) (Jun Wako and Shigeo Muto). However, the two best known solution concepts in cooperative game theory are perhaps the core and the Shapley value which are presented and discussed in ► Cooperative Games (Roberto Serrano).

The non-cooperative approach focuses on the actions that the decision-makers can take. Historically, the first contribution to the non-cooperative Game Theory is due to Zermelo (1913), but the idea of a general theory of games was introduced by John von Neumann and Oskar Morgenstern in their famous book of 1944 entitled *Theory of Games and Economic Behavior*. These authors argued that most economic questions should be analyzed as games. They introduced the extensive-form and the strategic-form representations of a game, also known as Dynamic and Static games, respectively.

Dynamic games stress the sequentiality of the various decisions that agents can make. An essential component of a dynamic game is the description of who moves first, who moves second, etc. Static games, on the other hand, abstract from sequentiality of the possible moves and model interactions as simultaneous decisions. All extensive form games can be modeled as static games, and all strategic form games can be modeled as dynamic games. However some situations may be more conveniently modeled as one or the other kind of game. Dynamic games are examined in ► Dynamic Games with an Application to Climate Change Models (Prajit K. Dutta). The structure, as well as its principal results, is discussed in detail. The Chapter ends with an important application, the economics of climate change.

The main ideas and results related to Static games, as well as some interesting relationships that connect equilibrium concepts with the idea of rationality are reviewed in ► Static Games (Oscar Volij). In this chapter it is presented the general Theorem of existence of strategic equilibria, due to Nash (1950). This result extends to more general games the minimax Theorem, which was proved in von Neuman (1928) for two-player zero-sum games. In the literature there are two proofs published by Nash. One of them uses Brower's fixed point theorem. The other one is a simpler proof, attributed to Gale by Nash that uses Kakutani's fixed point theorem. Some version of the proof that uses Brower's fixed point theorem, by Geanakoplos (2003), is presented in this chapter as well as some discussion on Correlated equilibrium and Bayesian games.

The Correlated equilibrium is a Game theoretic solution concept proposed by Aumann (1974, 1987) in order to capture the strategic correlation opportunities that the players face when they take into account the extraneous environment in which they interact. The Chap. ► Correlated Equilibria and Communication in Games (Françoise Forges) focuses on two possible extensions of the correlated equilibrium to Bayesian games: the strategic form correlated equilibrium and the communication equilibrium. The general framework of games with incomplete information is treated in ► Bayesian Games: Games with Incomplete Information (Shmuel Zamir) with special reference to "Bayesian games".

Repeated games deals with situations in which a group of agents engage in a strategic interaction over and over. The Chap. ► Repeated Games with Complete Information (Olivier Gossner and Tristan Tomala) is devoted to Repeated games with complete information. In such games the data of the strategic interaction is fixed over time and is known by all the players. The Chap. ► Repeated Games with Incomplete Information (Jérôme Renault) discusses Repeated games with incomplete information, a situation where several players repeat the same stage game, the players having different knowledge of the stage game which is repeated.

Repeated games have many equilibria, including the repetition of stage game Nash equilibria. At the same time, particularly when monitoring is imperfect, certain plausible outcomes are not consistent with equilibrium. Reputation effects is the term used for the impact upon the set of equilibria (typically of a repeated game) of perturbing the game by introducing incomplete information of a particular kind. This issue is treated in ► Reputation Effects (George Mailath).

Games with two players are of particular significance. The first two-person game studied in the literature was the Zero sum two-person game, first analyzed by von Neumann and Morgenstern (1944). In such a game one player's gain is the other player's loss. Chess, Checkers, Rummy, Two finger Morra, and Tic-Tac-Toe are all examples of zero-sum two-person games. The theory for such games is surveyed in ► Zero-Sum Two Person Games (T.E.S. Raghavan). Recent results on stochastic zero-sum games are presented in ► Stochastic Games (Eilon Solan). Stochastic games are used to model dynamic interactions in which the environment changes in response to the behavior of the players. These games are discussed in Solan's chapter.

Signaling games and Inspection games are also two-player games. Signaling games is the subject of ► Signaling Games (Joel Sobel). They are games of incomplete information in which one player is informed and the other is not. Players can use the actions of their opponents to make inferences about hidden information. The earliest work on this subject is Spence's seminal 1972 work, in which education serves as a signal of ability. Inspection games is covered in ► Inspection Games (Rudolf Avenhaus and Morton J. Canty). These games deal with the problem faced by an *inspector* who is required to control the compliance of an *inspectee* to some legal or otherwise formal undertaking. They started with the analysis of arms control and disarmament problems in the early 1960s and have been applied to auditing, environmental control, material accountancy, etc.

Inspections cause conflict in many real world situations. In economics, there are services of many kinds the fulfillment or payment of which has to be verified. One example is the problem of principal-agent relationships discussed in detail in ► Principal-Agent Models (David Perez-Castrillo and Inez Macho-Stadler). The Principal-agent models provide the theory of contracts under asymmetric information, concerning relationships between owner and manager, insurer and insured, etc. The principal, e. g., an employer, delegates work or responsibility to the agent, the employee, and chooses a payment schedule that best exploits the agent's self-interests. The agent, of course, behaves so as to maximize her own utility given the fee schedule proposed by the principal. The problem faced by the principal is to devise incentives to motivate the agent to act in the principal's interest. This generates some type of transaction cost for the principal, which includes the task of investigating and selecting appropriate agents, gaining information to set performances standards, monitoring agents, bonding payments by the agents, and residual losses.

The Chap. ► Differential Games (Marc Quincampoix) is devoted to Differential games with focus on two-player zero sum and antagonist differential games. These are games in which the state of the players depends on time in a continuous way. The positions of the players are solutions to differential equations. Motivated by military applications in the "Cold War" these games have a wide range of applications from Economics to engineer sciences, and recently to biology and behavioral ecology.

Mechanism designed is the subject of ► Mechanism Design (Ron Lavi). It studies the construction of mechanisms that aim to reach a socially desirable outcome in the presence of rational but selfish players, who care only about their own private utility. More specifically, the question is how to design a mechanism such that the equilibrium behavior of the players in the game induced by the mechanism leads to the socially desired goal.

The theory of mechanism design has been contributed to the development of other research areas as for example auction theory, contract theory and two-sided matching theory. "For having laid the foundations of mechanism design theory" the 2007 Nobel prize in Economics was awarded to Leonid Hurwicz, Eric Maskin and Roger Myerson.

A related theory is the theory of Implementation, the subject of ► Implementation Theory (Luis Corchon). It reverses the usual procedure, namely, fix a mechanism and see what the outcomes are. More precisely, it investigates the correspondence between normative goals and mechanisms designed to achieve those goals.

A class of "mixed" games is that of two-sided matching games, that has been analyzed since Gale and Shapley, 1962, under both cooperative and non-cooperative game theoretic approaches. The two-sided matching theory is surveyed in ► Two-Sided Matching Models (Marilda Sotomayor and Ömer Özak) by focusing on the differences and similarities between some matching models. In their paper, Gale and Shapley formulated and solved the Stable Matching problem for the Marriage and the College Admissions markets. The solution of the College Admissions problem was given by a simple deferred-acceptance algorithm which has been adapted and applied in the reorganization of admission processes of many two-sided matching markets.

Another class of problems that have been discussed from the perspective of cooperative and non-cooperative game theory is the Cost sharing problems, treated in ► Cost Sharing (Maurice Koster). Applications are numerous ranging from environmental issues like pollution, fishing grounds, to sharing multipurpose reservoirs, road systems, communication networks, and the internet. The worth of a "coalition" of such activities is defined as the hypothetical cost of carrying out the activities in that coalition only.

Market games and clubs are treated in ► Market Games and Clubs (Myrna Wooders) with focus on the equivalence between markets – defined as private goods economies where all participants in the economy have utility functions that are linear in the variable money - and games in characteristic function form.

Learning in Games is surveyed in ► Learning in Games (John Nachbar). It covers models in which players are "rational" but not necessarily in equilibrium: players forecast, possibly inaccurately, the future behavior of their opponents and optimize or ε optimize with respect to their forecasts.

Fair Division is reviewed in ► Fair Division (Steven Brams). It provides a rigorous analysis of procedures for allocating goods, or deciding who wins on what issues, in a dispute.

The following two chapters deal with applications to political sciences. The first one, the Chap. ► Voting (Alvaro Sandroni, Antonio Penta, Jonathan Pogach, Deniz Selman and Michela Tincani), presents a game theoretic analysis of voting systems, which are procedures to choose a winner among a set of candidates from the individual preferences of the voters or more ambitiously, allowing to rank all the candidates or a part of them. Such a situation occurs in the field of social choice and welfare, in the field of elections, and in many other fields as games, sports, artificial intelligence, spam detection, web search engines and more generally Internet applications, statistics, and soon. From a practical point of view, it is crucial to be able to announce who is the winner in a "reasonable" time. This raises the question of the complexity of the voting procedures. The second Chap. ► Voting Procedures, Complexity of (Olivier Hudry) details the complexity results about several voting procedures.

The Chap. ► Evolutionary Game Theory (William Sandholm) deals with applications to biology. This field, known as Evolutionary game theory, started in 1972 with the publication of a series of papers by the mathematical biologist John Maynard Smith. Maynard Smith adapted the methods of traditional game theory, which were created to model the behavior of rational economic agents, to the context of biological natural selection.

Network models have a long history in sociology, natural sciences and engineering. However, only recently economists have begun to think of political and economic interactions as network phenomena and to model everything as games of network formation. The Chap. ► Networks and Stability (Frank Page and Myrna Wooders) is devoted to stable Networks and the game theoretic underpinnings of stable networks.

The Chap. ► Game Theory and Strategic Complexity (Kalyan Chatterjee and Hamid Sabourian) deals with some aspect of bounded rationality that has generated important work, namely the presence of constraints on the capacities of players. Various constraints could be considered, for example, limits on the ability to plan ahead in intertemporal decision making or on the ability to compute best responses.

This chapter discusses cognitive costs to players of using strategies that depend on long histories of past play. This is done mainly in the context of bargaining and markets. It is shown that such complexity considerations often enable us to make sharp predictions. It is also considered the issue in the context of repeated games.

Game Theory and Strategic Complexity

Kalyan Chatterjee[1], Hamid Sabourian[2]
[1] Department of Economics, The Pennsylvania State University, University Park, USA
[2] Faculty of Economics, University of Cambridge, Cambridge, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Games, Automata and Equilibrium Concepts
Complexity Considerations in Repeated Games
Complexity and Bargaining
Complexity, Market Games and the Competitive Equilibrium
Discussion and Future Directions
Acknowledgments
Bibliography

Glossary

Game theory A formal model of interaction, usually in human behavior.

Repeated games A series of identical interactions of this kind.

Strategy A complete specification of how a player will play the game.

Strategic complexity A measure of how complex a strategy is to implement.

Equilibrium A solution concept for games in which each player optimizes given his correct prediction of others' behavior.

Equilibrium path The outcome in terms of the play of the game if every player uses his equilibrium strategy.

Continuation game A description of how the play will proceed in a dynamic game once some part of the game has already occurred.

Automata A formal definition of a strategy that captures its complexity.

Definition of the Subject

The subject of this chapter is at the intersection of economics and computer science and deals with the use of measures of complexity obtained from the study of finite automata to help select among multiple equilibria and other outcomes appearing in game-theoretic models of bargaining, markets and repeated interactions. The importance of the topic lies in the ability of concepts that employ bounds on available resources to generate more refined predictions of individual behavior in markets.

Introduction

This chapter is concerned with the concept of strategic complexity and its use in game theory. There are many different meanings associated with the word "complexity", as the variety of topics discussed in this volume makes clear. In this paper, we shall adopt a somewhat narrow view, confining ourselves to notions that measure, in some way, constraints on the ability of economic agents to behave with full rationality in their interactions with other agents in dynamic environments. This will be made more precise a little later. (A more general discussion is available in Rubinstein [45].)

Why is it important to study the effect of such constraints on economic decision-making? The first reason could be to increase the realism of the assumptions of economic models; it is evident from introspection and from observing others that we do not have infinite memory and cannot condition our future actions on the entire corpus of what we once knew; or, for that matter, unlimited computational power. However, only considering the assumptions of a model would not be considered enough if the increased realism were not to expand our ability to explain or to predict. The second reason therefore is that studying the effects of complexity on human decision-making might help us either to make our predictions more precise (by selecting among equilibria) or to generate explanations for behavior that is frequently observed, but incompatible with equilibrium in models that have stronger assumptions about the abilities of agents.

A strategy in a game is an entire plan of how to play the game at every possible history/contingency/eventuality at which the player has to make a move. The particular aspect of complexity that we shall focus on is on the complexity of strategy as a function of the history. One representation of the players' strategies in games is often in terms of (finite) automata. The finiteness need not always be assumed; it can be derived. The ideas of complexity, though often most conveniently represented this way, can also be discussed without referring to finite automata at all but purely in how a strategy depends on the past history of the game.

The number of states in the automaton can be used as a measure of complexity. This may be a natural measure of complexity in a stationary repetitive environment such as repeated games. We shall discuss this measure of complexity as well as other aspects of the complexity of

a strategy that are particularly relevant in non-stationary frameworks.

Note the players are not themselves considered automata in this paper and in the literature it surveys. Also, we do not place restrictions on the ability of players to compute strategies (see [39]), only on the strategies that they can implement. The paper is also not intended as a comprehensive survey of the literature on complexity of implementation in games. The main focus of the paper is inevitably on the works that we have been personally associated with.

The remaining part of this paper is organized as follows: In the next section, we discuss strategies in a game, their representation as finite automata and the basic equilibrium concepts to be used in the paper. Section "Complexity Considerations in Repeated Games" will consider the use of complexity notions in repeated games. Section "Complexity and Bargaining" will focus on extensive form bargaining and the effect of complexity considerations in selecting equilibria. Section "Complexity, Market Games and the Competitive Equilibrium" will extend the analysis of bargaining to markets in which several agents bargain and considers the recent literature that justifies competitive outcomes in market environments by appealing to the aversion of agents to complexity. Section "Discussion and Future Directions" concludes with some thoughts on future research. This paper draws on an earlier survey paper [14] for some of the material in Sects. "Games, Automata and Equilibrium Concepts", "Complexity Considerations in Repeated Games" and "Complexity and Bargaining".

Games, Automata and Equilibrium Concepts

As mentioned in the introduction, this paper will be concerned with *dynamic games*. Though the theory of games has diffused from economics and mathematics to several other fields in the last few decades, we include an introduction to the basic concepts to keep this paper as self-contained as possible. A game is a formal model of interaction between individual agents. The basic components of a game are: (i) Players or agents, whose choices will, in general, have consequences for each other. We assume a finite set of players, denoted by N. We shall also use N sometimes to represent the cardinality of this set. (ii) A specification of the "rules of the game" or the structure of interaction, described by the sequence of possible events in the game, the order in which the players move, what they can choose at each move and what they know about previous moves. This is usually modeled as a tree and is called the "extensive form" of the game (and will not be formalized here, though the formalization is standard and found in all the texts on the subject). (iii) Payoffs for each player associated with every path through the tree from the root. It is easier to describe this as a finite tree and ascribe payoffs to the end nodes z. Let $u_i(z)$ be the real-valued payoff to Player i associated with end node z. The payoffs are usually assumed to satisfy conditions that are sufficient to guarantee that the utility of a probability distribution on a subset of the set of end nodes is the expectation of the utility of the individual end nodes. However, different strands of work on bounded rationality dispense with this assumption. The description above presupposes a tree of finite depth, whilst many of the applications deal with infinite horizon games. However, the definitions are easily modified by associating payoffs with a *play* of the game and defining a node as a set of plays. We shall not pursue this further here.

In the standard model of a game, players are assumed to have all orders of knowledge about the preceding description. Work on bounded rationality also has considered relaxing this assumption.

A strategy is a complete plan of action for playing a game, describing the course of action to be adopted in every possible contingency (or every information set of the player concerned). The plan has to be detailed enough so that it can be played by an agent, even if the principal is not himself or herself in town, and the agent could well be a computer, which is programmed to follow the strategy. Without any loss of generality, a strategy can be represented by an automaton (see below for illustration and Osborne and Rubinstein [38] for a formal treatment in the context of repeated games). Often such a machine description is more convenient in terms of accounting for a complexity of a machine. For example the works that are based on the use of finite automata or Turing machines to represent *strategies* for playing a game impose a natural bound on the set of allowable strategies.

For the types of problem that we shall consider here, it is best to think of a *multistage game with observable actions*, to use the terminology of Fudenberg and Tirole [20]. The game has some temporal structure; let us call each unit of time a period or a stage. In each period, the players choose actions simultaneously and independently. (The actions could include the dummy action.) All the actions taken in a stage are observed and the players then choose actions again. An example is a repeated normal form game, such as the famous Prisoners' Dilemma being repeated infinitely or finitely often. In each stage, players choose whether to *cooperate* or *defect*. The choices are revealed, payoffs received and the choices repeated again and so on. (The reader will recall that in the Prisoners' Dilemma played once, *Defect* is better than *Cooperate* for

each player, no matter what the other player does, but both players choosing *Defect* is strictly worse for each than both choosing *Cooperate*.) What a strategy for a given player would do would be to specify the choice in a given stage as a function of the history of the game up to that stage (for every stage). A finite automaton represents a particular strategy in the following way: It partitions all possible histories in the game (at which the player concerned has to move) using a finite number of elements. Each of these elements is a *state* of the machine. Given a state, the automaton prescribes an action (for example, Cooperate after all histories in which the other party has cooperated). It also specifies how the state of the machine will change as a result of the action taken by the other player. The state-to-action mapping is called the output mapping and the rule that prescribes the state in the next period as a function of today's state and the action of one's opponent in this period is called the transition mapping. The automaton also needs to prescribe what to do in the first stage, when there is no history of past actions to rely on. Thus for example, the famous 'tit-for-tat' strategy in the repeated Prisoners' Dilemma can be represented by the following automaton.

1. Play Cooperate in the first stage. The initial state is denoted as q^1 and in this state the action prescribed is cooperate.
2. As long as the other player cooperates, stay in state q^1.
3. If the other player defects in a state, go to state q^2. The action specified in q^2 is *Defect*.
4. Stay in q^2 as long as the other player defects. If the other player cooperates in a stage, go to q^1.

Denoting the output mapping by $\lambda(\cdot)$, we get $\lambda(q^1) = C$ and $\lambda(q^2) = D$. The transition mapping, $\mu(\cdot,\cdot)$ is as follows: $\mu(q^1, C) = q^1$, $\mu(q^1, D) = q^2$, $\mu(q^2, C) = q^1$, $\mu(q^2, D) = q^2$. Here, of course, C and D denote cooperate and defect respectively. The machine described above has two states and is an instance of a Moore machine in computer science terminology.

The use of a Moore machine to represent a strategy rules out strategies in which histories are arbitrarily finitely partitioned or arbitrarily complex. In fact, the number of states in the machine is a popular measure of the complexity of the machine and the strategy it represents.

Another kind of finite automaton used in the literature is a Mealy machine. The main difference between this and the Moore machine is that now the output is a function both of the state and of an input, unlike the Moore machine where it is only a function of the state. One can always transform a Mealy machine to a Moore machine by making transitions depend on the input and having state transitions after every input. The Mealy machine representation is more convenient for the extensive form game we shall consider in Section "Complexity and Bargaining". We shall briefly address why in that section.

The aim of using the machine framework to describe strategies is to take into account explicitly the cost of complexity of strategies. There is the belief for instance that short-term memory (see [33]) is capable of keeping seven things in mind at any given time and if five of them are occupied by how to play the Prisoners' Dilemma there might be less left over for other important activities.

The standard equilibrium concept in game theory is the concept of Nash equilibrium. This requires each player to choose a best strategy (in terms of payoff) given his or her conjectures about other players' strategies and, of course, in equilibrium the conjectures must be correct. Thus, a Nash equilibrium is a profile of strategies, one for each player, such that every player is choosing a best response strategy given the Nash equilibrium strategies of the other players. In dynamic games Nash equilibrium strategies may not be credible (sequentially rational). In multi-stage games, to ensure credibility, the concept of Nash equilibrium is refined by requiring the strategy of each player to be a best response to the strategies of the others at every well-defined history (subgame) within the game. This notion of equilibrium was introduced by Selten [48] and is called subgame perfect equilibrium. The difference between this concept and that of Nash, which it refines, is that players must specify strategies that are best responses to each other even at nodes in the game tree that would never be reached if the prescribed equilibrium were being played. The Nash concept does not require this. The notion of histories *off the equilibrium path* therefore refers to those that do not occur if every player follows his or her equilibrium strategy. Another useful concept to mention here is that of *payoff in the continuation game*. This refers to the expected payoff from the prescribed strategies in the part of the game remaining to be played after some moves have already taken place. The restriction of the prescribed strategies to the continuation game are referred to here as *continuation strategies*.

Rubinstein [44], Abreu and Rubinstein [1] and others have modified the standard equilibrium concepts to account for complexity costs. This approach is somewhat different from that adopted, for example, by Neyman [35], who restricted strategies to those of bounded complexity. We shall next present the Abreu–Rubinstein definition of *Nash equilibrium with complexity* (often referred to as NEC in the rest of the paper).

The basic idea is a very simple extension of Nash equilibrium. Complexity enters the utility function lexicographically. A player first calculates his or her best re-

sponse to the conjectured strategies of the other players. If there are alternative best responses, the player chooses the less complex one. Thus a Nash equilibrium with complexity has two aspects. First, the strategies chosen by any player must be a best response given his or her conjectures about other players' strategies and, of course, in equilibrium the conjectures must be correct. Second, there must not exist an alternative strategy for a player such that his or her payoff is the same as in the candidate equilibrium strategy, given what other players do, but the alternative strategy is less complex.

In Abreu and Rubinstein [1], the measure of complexity is the number of states in the Moore machine that represents the strategy. The second part of their equilibrium definition restricts the extent to which punishments can be used off the equilibrium path. For example, there is a famous strategy that, if used by all players, gives cooperation in the infinitely repeated Prisoners' Dilemma (for sufficiently high discount factors), namely the "grim" strategy. This strategy can be described by the following machine: Start with *Cooperate*. Play *Cooperate* as long as the other players all cooperate. If in the last period any player has used *Defect*, then switch to playing *Defect* for ever. (That is, never play *Cooperate* again, no matter what the other players do in succeeding periods.) This strategy profile (each player uses the grim strategy) gives an outcome path consisting solely of players cooperating. No one defects because from then until the end of time all the players will be punishing one another.

However, this strategy profile is not a Nash equilibrium with complexity; the grim strategy is a two-state machine in which one state (the one in which a player chooses *Defect*) is never used given that everyone else cooperates on the equilibrium path. Some player can do better, even if lexicographically, by switching to a one-state machine in which he or she cooperates no matter what. Thus even the weak lexicographic requirement has some bite.

Note that the complexity restriction we are considering is on the complexity of *implementation*, not the complexity of *computation*. We know that even a Turing machine, which has potentially infinite memory, might be unable to calculate best responses to all possible strategy profiles of other players in the game (see [2,8]).

To return to the question of defining equilibrium in the machine game, the Abreu–Rubinstein approach is described by them as "buying" states in the machine at the beginning of the game. The complexity cost is therefore a fixed cost per state used. Some recent papers have taken the fixed cost approach further by requiring NEC strategies to be credible. The idea is that players pay an initial fixed cost for the complexity (the notion of complexity in some of these papers differ from counting the states approach) of his/her strategy and then the game is played with strategies being optimal at every contingency as in standard game theory. Chatterjee and Sabourian [15,16] model this by considering Nash equilibrium with complexity costs in (bargaining) games in which machines/strategies can make errors/trembles in output/action. The introduction of errors ensures that the equilibrium strategies are optimal after every history. As the error goes to zero, we are left with subgame perfect equilibria of the underlying game. Chatterjee and Sabourian [15], Sabourian [47], Gale and Sabourian [22] and Lee and Sabourian [31] take a more direct method of introducing credibility into the equilibrium concept with complexity costs by restricting NEC strategies to be subgame perfect equilibrium in the underlying game with no complexity costs. We refer to such an equilibria by perfect equilibrium with complexity costs (PEC).

In contrast to the fixed cost interpretation of complexity cost, Rubinstein in his 1986 paper considers a different approach, namely the choice of "renting" states in the machine for every period the game is played. Formally, the Rubinstein notion of *semi-perfect equilibrium* requires the strategy chosen to have the minimal number of states necessary to play the game at every node on the (candidate) equilibrium outcome path. A state could therefore be dropped if it is not going to be used on the candidate equilibrium path after some period. Thus, to be in the equilibrium machine, it is not sufficient that a state be used on the path, it has to be used in every possible future. Rubinstein called this notion of equilibrium *semi-perfect*, because the complexity of a strategy could be changed in one direction (it could be decreased) after every period. If states could be added as well as deleted every period, we would have yet another definition of equilibrium with complexity, *machine subgame perfect equilibrium*. (See [34].) In contrast, both the NEC and PEC concepts we use here entail a *single choice* of automaton or strategy by players at the beginning of the game.

In all these models, complexity analysis has been facilitated by considering the "machine games". Each player chooses among machines and the complexity of a machine is taken to be the number of states of the machine. In fact, the counting-the-number-of-states measure of complexity has an equivalent measure stated in terms of the underlying strategies that the machine could implement. Kalai and Stanford [29] define complexity of a strategy by the number of *continuation strategies* that the strategy induces at different periods/histories of the game, and establishes that such a measure is equal to the number of the states of the smallest machine that implements the strategy. Thus,

one could equivalently describe any result either in terms of underlying strategies and the cardinality of the set of continuation strategies that they induce or in terms of machines and the number of states in them. The same applies to other measures of complexity discussed in this paper; they can be defined either in terms of the machine specification or in terms of the underlying strategy. In the rest of this paper, to simplify the exposition we shall at times go from one exposition to the other without further explanation.

With this preamble on the concepts of equilibrium used in this literature, we turn to a discussion of a specific game in the next section, the infinitely repeated Prisoners' Dilemma. We will discuss mainly the approach of Abreu and Rubinstein in this section but contrast it with the literature following from Neyman. We also note that the suggestion for using finite automata in games of this kind came originally from Aumann [4].

Complexity Considerations in Repeated Games

Endogeneous Complexity

In this subsection we shall first concentrate on the Prisoners' Dilemma and discuss the work of Abreu and Rubinstein, which was introduced briefly in the last section. For concreteness, consider the following Prisoners' Dilemma payoffs:

	C_2	D_2
C_1	$3, 3$	$-1, 4$
D_1	$4, -1$	$0, 0$

This is the "stage game"; each of the two players chooses an action in each stage, their actions are revealed at the end of the stage and then the next stage begins. The game is repeated infinitely often and future payoffs are discounted with a common discount factor δ.

The solution concept to be used was introduced in the last section; NEC or Nash equilibrium with complexity. Note that here complexity is *endogenous*. A player has a preference for less complex strategies. This preference comes into play lexicographically, that is for any strategies or machines that give the same payoff against the opponent's equilibrium strategy, a player will choose the one with lowest complexity. Thus the cost of complexity is infinitesimal. One could also consider positive but small costs of more complex strategies, but results will then depend on how large the cost of additional complexity is compared to the additional payoff obtained with a more complex strategy.

We saw in the last section that the "grim trigger" strategy, which is a two-state automaton, is not a NEC. The reason is that if Player 2 uses such a strategy, Player 1 can be better off by deviating to a one-state strategy in which she always cooperates.(This will give the same payoff with a less complex strategy.) One-state strategies where both players cooperate clearly do not constitute NEC (deviating and choosing a one-state machine that always plays D is strictly better for a player). However, if both players use a one-state machine that always generates an action of D, this is a NEC.

The question obviously arises if the cooperative outcome in each stage can be sustained as a NEC and the preceding discussion makes clear that the answer is no. Punishments have to be used on the equilibrium path, but we can get arbitrarily close to the cooperative outcome for a high enough discount factor. For example consider the following two-state machine:

$$Q = \{q^1, q^2\}; \lambda(q^1) = D\,, \lambda(q^2) = C\,,$$
$$\mu(q^1, C) = q^1\,, \mu(q^1, D) = q^2\,, \mu(q^2, D) = q^1\,,$$
$$\mu(q^2, C) = q^2\,.$$

Here both players play the same strategy, which starts out playing D. If both players do as they are supposed to, each plays C in the next period and thereafter, so the sequence of actions is (D, D), (C, C), $(C, C)\ldots$ If either player plays C in the first period, the other player keeps playing D in the next period. The transition rule prescribes that if one plays C and one's opponent plays D, one goes back to playing D, so the sequence with the deviation will be (D, C), (D, D), (C, C), $(C, C)\ldots$

Suppose both players use this machine. First, we check it is a Nash equilibrium in payoffs. We only need to check what happens when a player plays C. If Player 2 deviates and plays D, she will get an immediate payoff of 4 followed by payoffs of $0, 3, 3\ldots$ if she thereafter sticks to her strategy for a total payoff of $4 + \delta^2(3/(1-\delta))$ as opposed to $3/(1-\delta)$ if she had not deviated. The net gain from deviation is $1 - 3\delta$, which is negative for $\delta > \frac{1}{3}$. One can check that more complicated deviations are also worse. The second part of the definition needs to be checked as well, so we need to ensure that a player cannot do as well in terms of payoff by moving to a less complex strategy, namely a one-state machine. A one-state machine that always plays C will get the worst possible payoff, since the other machine will keep playing D against it. A one-state machine that plays D will get a payoff of 4 in periods $2,4\ldots$ or a total payoff of $4\delta/(1-\delta^2)$ as against $3\delta/(1-\delta)$. The second is strictly greater for $\delta > \frac{1}{3}$.

This machine gives a payoff close to 3 per stage for δ close to 1. As $\delta \to 1$, the payoff of each player goes to 3, the cooperative outcome.

The paper by Abreu and Rubinstein obtains a basic result on the characterization of payoffs obtained as NEC in the infinitely repeated Prisoners' Dilemma. We recall that the "Folk Theorem" for repeated games tells us that all outcome paths that give a payoff per stage strictly greater for each player than the minmax payoff for that player in the stage game can be sustained by Nash equilibrium strategies. Using endogenous complexity, one can obtain a refinement; now only payoffs on a so-called "cross" are sustainable as NEC. This result is obtained from two observations. First, in any NEC of a two-player game, the number of states in the players' machines must be equal. This follows from the following intuitive reasoning (we refer readers to the original paper for the proofs). Suppose we fix the machine used by one of the players (say Player 1), so that to the other player it becomes part of the "environment". For Player 2 to calculate a best response or an optimal strategy to Player 1's given machine, it is clearly not necessary to partition past histories more finely than the other player has done in obtaining her strategy; therefore the number of states in Player 2's machine need not (and therefore will not, if there are complexity costs) exceed the number in Player 1's machine in equilibrium. The same holds true in the other direction, so the number of states must be equal. (This does not hold for more than two players.) Another way of interpreting this result is that it restates the result from Markov decision processes on the existence of an optimal "stationary" policy (that is depending only on the states of the environment, which are here the same as the states of the other player's machine). See also Piccione [40].

Thus there is a one-to-one correspondence between the states of the two machines. (Since the number of states is finite and the game is infinitely repeated, the machine must visit at least one of the states infinitely often for each player.) One can strengthen this further to establish a one-to-one correspondence between *actions*. Suppose Player 1's machine has $a_t^1 = a_s^1$, where these denote the actions taken at two distinct periods and states by Player 1, with $a_t^2 \neq a_s^2$ for Player 2. Since the states in t and s are distinct for Player 1 and the actions taken are the same, the transitions must be different following the two distinct states. But then Player 1 does not need two distinct states, he can drop one and condition the transition after, say, s on the different action used by Player 2. (Recall the transition is a function of the state and the opponent's action.) But then Player 1 would be able to obtain the same payoff with a less complex machine; so the original one could not have been a NEC machine.

Therefore the actions played must be some combination of (C, C) and (D, D) (the correspondence is between the two Cs and the two Ds) or some combination of (C, D) and (D, C). (By combination, we mean combination over time. For example, $\{C, C)$ is played, say, 10 times for every 3 plays of (D, D)). In the payoff space, sustainable payoffs are either on the line joining (3,3) and (0,0) or on the line joining the payoffs on the other diagonal; hence the evocative name chosen to describe the result-the cross of the two diagonals.

While this is certainly a selection of equilibrium outcomes, it does not go as far as we would wish. We would hope that some equilibrium selection argument might deliver us the co-operative outcome (3,3) uniquely (even in the limit as $\delta \to 1$), instead of the actual result obtained. There is work that does this, but it uses evolutionary arguments for equilibrium selection (see Binmore and Samuelson [9]). An alternative learning argument for equilibrium selection is used by Maenner [32]. In his model, a player tries to infer what machine is being used by his opponent and chooses the simplest automaton that is consistent with the observed pattern of play as his model of his opponent. A player then chooses a best response to this inference. It turns out complexity is not sufficient to pin down an inference and one must use optimistic or pessimistic rules to select among the simplest inferences. One of these gives only (D, D) repeated, whilst the other reproduces the Abreu–Rubinstein NEC results. Piccione and Rubinstein [41] show that the NEC profile of 2-player repeated *extensive form games* is unique if the stage game is one of perfect information. This unique equilibrium involves all players playing their one-shot myopic non-cooperative actions at every stage. This is a strong selection result and involves stage game strategies not being observable (only the path of play is) as well as the result on the equilibrium numbers of states being equal in the two players' machines.

In repeated games with more than two players or with more than two actions at each stage the multiplicity problem may be more acute than just not being able to select uniquely a "cooperative outcome". In some such games complexity by itself may not have any bite and the Folk Theorem may survive even when the players care for the complexity of their strategies. (See Bloise [12] who shows robust examples of two-player repeated games with three actions at each stage such that every individually rational payoff can be sustained as a NEC if players are sufficiently patient.)

Exogenous Complexity

We now consider the different approach taken by Neyman [35,36], Ben Porath [6,7], Zemel [50] and others.

We shall confine ourselves to the papers by Neyman and Zemel on the Prisoners' Dilemma, without discussing the more general results these authors and others have obtained.

Neyman's approach treats complexity as exogenous complexity. Let Player i be restricted to use strategies/automata with the number of states not to exceed m_i. He also considers finitely repeated games, unlike the infinitely repeated games we have discussed up to this point. With the stage game being the Prisoners' Dilemma and the number of repetitions being T (for convenience, this includes the first time the game is played). We can write the game being considered as $G^T(m_1, m_2)$. Note that without the complexity restrictions, the finitely repeated Prisoners' Dilemma has a unique Nash equilibrium outcome path (and a unique subgame perfect equilibrium)–(D, D) in all stages. Thus sustaining cooperation in this setting is obtaining non-equilibrium behavior, though one that is frequently observed in real life. This approach therefore is an example of bounded rationality being used to explain observed behavior that is not predicted in equilibrium.

If the complexity restrictions are severe, it turns out that (C, C) in each period is an equilibrium. For this, we need $2 \leq m_1, m_2 \leq T - 1$. To see this consider the grim trigger strategy mentioned earlier-representable as a two-state automaton- and let $T = 3$. Here $\lambda(q^1) = C$; $\lambda(q^2) = D$; $\mu(q^1, C) = q^1$; $\mu(q^1, D) = q^2$; $\mu(q^2, C \text{ or } D) = q^2$. If each player uses this strategy, (C, C) will be observed. Such a pair of strategies is clearly not a Nash equilibrium-given Player 1's strategy, Player 2 can do better by playing D in stage 3. But if Player 2 defects in the second stage, by choosing a two-state machine where $\mu(q^1, C) = D$, he will gain 1 in the second stage and lose 3 in the third stage as compared to the machine listed above, so he is worse off. But defecting in stage 3 requires an automaton with three states-two states in which C is played and one in which D is played. The transitions in state q^1 will be similar but, if q^2 is the second cooperative state, the transition from q^2 to the defect state will take place no matter whether the other player plays C or D. However, automata with three states violate the constraint that the number of states be no more than 2, so the profitable deviation is out of reach.

Whilst this is easy to see, it is not clear what happens when the complexity is high. Neyman shows the following result: *For any integer k, there exists a T_0, such that for $T \geq T_0$ and $T^{\frac{1}{k}} \leq m_1, m_2 \leq T^k$, there is a mixed strategy equilibrium of $G^T(m_1, m_2)$ in which the expected average payoff to each player is at least $3 - \frac{1}{k}$.*

The basic idea is that rather than playing (C, C) at each stage, players are required to play a complex sequence of C and D and keeping track of this sequence uses up a sufficient number of states in the automaton so that profitable deviations again hit the constraint on the number of states. But since D cannot be avoided on the equilibrium path, only something close to (C, C) each period can be obtained rather than (C, C) all the time.

Zemel's paper adds a clever little twist to this argument by introducing communication. In his game, there are two actions each player chooses at each stage, either C or D as before and a message to be communicated. The message does not directly affect payoffs as the choice of C or D does. The communication requirements are now made sufficiently stringent, and deviation from them is considered a deviation, so that once again the states "left over" to count up to N are inadequate in number and (C, C) can once again be played in each stage/period. This is an interesting explanation of the rigid "scripts" that many have observed to be followed, for example, in negotiations.

Neyman [36] surveys his own work and that of Ben Porath [6,7]. He also generalizes his earlier work on the finitely-repeated Prisoners' Dilemma to show how small the complexity bounds would have to be in order to obtain outcomes outside the set of (unconstrained) equilibrium payoffs in the finitely-repeated, normal-form game (just as (C, C) is not part of an unconstrained equilibrium outcome path in the Prisoners' Dilemma). Essentially, if the complexity permitted grows exponentially or faster with the number of repetitions, the equilibrium payoff sets of the constrained and the unconstrained games will coincide. For sub-exponential growth, a version of the Folk-theorem is proved for two-person games. The first result says:

> For every game G in strategic form and with m_i being the bound on the complexity of $i's$ strategy and T the number of times the game G is played, there exists a constant c such that if $m_i \geq \exp(cT)$, then $E(G^T) = E(G^T(m_1, m_2)$ where $E(.)$ is the set of equilibrium payoffs in the game concerned.

The second result, which generalizes the Prisoners' Dilemma result already stated, considers a sequence of triples $(m_1(n), m_2(n), T(n))$ for a two-player strategic form game, with $m_2 \geq m_1$ and shows that the lim inf of the set of equilibrium payoffs of the automata game as $n \to \infty$ includes essentially the strictly individually rational payoffs of the stage game if $m_1(n) \to \infty$ and $(\log m_1(n))/T(n) \to 0$ as $n \to \infty$. Thus a version of the Folk theorem holds provided the complexity of the players' machines does not grow too fast with the number of repetitions.

Complexity and Bargaining

Complexity and the Unanimity Game

The well-known alternating offers bargaining model of Rubinstein has two players alternating in making proposals and responding to proposals. Each period or unit of time consists of one proposal and one response. If the response is "reject", the player who rejects makes the next proposal but in the following period. Since there is discounting with discount factor δ per period, a rejection has a cost. The unanimity game we consider is a multiperson generalization of this bargaining game, with n players arranged in a fixed order, say $1, 2, 3 \ldots n$. Player 1 makes a proposal on how to divide a pie of size unity among the n people; players $2, 3, \ldots n$ respond sequentially, either accepting or rejecting. If everyone accepts, the game ends. If someone rejects, Player 2 now gets to make a proposal but in the next period. The responses to Player 2's proposal are made sequentially by Players $3, 4, 5 \ldots n, 1$. If Player i gets a share x_i in an eventual agreement at time t, his payoff is $\delta^{t-1} x_i$.

Avner Shaked had shown in 1986 that the unanimity game had the disturbing feature that all individually rational (that is non-negative payoffs for each player) outcomes could be supported as subgame perfect equilibria. Thus the sharp result of Rubinstein [43], who found a unique subgame perfect equilibrium in the two-play stood in complete contrast with the multiplicity of subgame perfect equilibria in the multiplayer game.

Shaked's proof had involved complex changes in expectations of the players if a deviation from the candidate equilibrium were to be observed. For example, in the three-player game with common discount factor δ, the three extreme points $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$ sustain one another in the following way. Suppose Player 1 is to propose $(0, 1, 0)$, which is not a very sensible offer for him or her to propose, since it gives everything to the second player. If Player 1 deviates and proposes, say, $((1-\delta)/2, \delta, (1-\delta)/2)$, then it might be reasoned that Player 2 would have no incentive to reject because in any case he or she can't get more than 1 in the following period and Player 3 would surely prefer a positive payoff to 0. However, there is a counter-argument. In the subgame following Player 1's deviation, Player 3's expectations have been raised so that he (and everyone else, including Player 1) now expect the outcome to be $(0, 0, 1)$, instead of the earlier expected outcome. For sufficiently high discount factor, Player 3 would reject Player 1's insufficiently generous offer. Thus Player 1 would have no incentive to deviate. Player 1 is thus in a bind; if he offers Player 2 less than δ and offers Player 3 more in the deviation, the expectation that the outcome next period will be $(0, 1, 0)$ remains unchanged, so now Player 2 rejects his offer. So no deviation is profitable, because each deviation generates an expectation of future outcomes, an expectation that is confirmed in equilibrium. (This is what equilibrium means.) Summarizing, $(0, 1, 0)$ is sustained as follows: Player 1 offers $(0, 1, 0)$, Player 2 accepts any offer of at least 1 and Player 3 any offer of at least 0. If one of them rejects Player 1's offer, the next player in order offers $(0, 1, 0)$ and the others accept. If any proposer, say Player 1, deviates from the offer $(0, 1, 0)$ to (x_1, x_2, x_3) the player with the lower of $\{x_2, x_3\}$ rejects. Suppose it is Player i who rejects. In the following period, the offer made gives 1 to Player i and 0 to the others, and this is accepted.

Various attempts were made to get around the continuum of equilibria problem in bargaining games with more than two players; most of them involved changing the game. (See [15,16] for a discussion of this literature.) An alternative to changing the game might be to introduce a cost for this additional complexity, in the belief that players who value simplicity will end up choosing simple, that is history independent, strategies. This seems to be a promising approach because it is clear from Shaked's construction that the large number of equilibria results from the players choosing strategies that are history-dependent. In fact, if the strategies are restricted to those that are history-independent (also referred to as *stationary* or *Markov*) then it can be shown (see Herrero [27]) that the subgame perfect equilibrium is unique and induces equal division of the pie as $\delta \to 1$.

The two papers ([15,16]) in fact seek to address the issue of complex strategies with players having a preference for simplicity, just as in Abreu and Rubinstein. However, now we have a game of more than two players, and a single extensive form game rather than a repeated game as in Abreu–Rubinstein. It was natural that the framework had to be broadened somewhat to take this into account.

For each of n players playing the unanimity game, we define a machine or an implementation of the strategy as follows.

A *stage* of the game is defined to be n periods, such that if a stage were to be completed, each player would play each role at most once. A *role* could be as proposer or $n-1$th responder or $n-2$th responder … up to first responder (the last role would occur in the period before the player concerned had to make another proposal). An *outcome* of a stage is defined as a sequence of offers and responses, for example $e = (x, A, A, R; y, R; z, A, R; b, A, A, A)$ in a four-player game where the (x, y, z, b) are proposals made in the four periods and (A, R) refer to accept and reject respectively. From

the point of view of the first player to propose (for convenience, let's call him Player 1), he makes an offer x, which is accepted by Players 2 and 3 but rejected by Player 4. Now it is Player 2's turn to offer, but this offer, y, is rejected by the first responder Player 3. Player 1 gets to play as second responder in the next period, where he rejects Player 3's proposal. In the last period of this stage, a proposal b is made by Player 4 and everyone accepts (including Player 1 as first responder). Any partial history within a stage is denoted by s. For example, when Player 2 makes an offer, he does so after a partial history $s = (x, A, A, R)$. Let the set of possible outcomes of a stage be denoted by E and the set of possible partial histories by S. Let Q_i denote the set of states used in the ith player's machine M_i. The output mapping is given by $\lambda_i : S \times Q_i \to \Lambda$, where Λ is the set of possible actions (that is the set of possible proposals, plus accept or reject). The transition between states now takes place at the end of each stage, so the transition mapping is given as $\mu_i : E \times Q_i \to Q_i$. As before, in the Abreu–Rubinstein setup, there is an initial state $q_{\text{initial},i}$ specified for each player. There is also a termination state F, which is supposed to indicate agreement. Once in the termination state, players will play the null action and make transitions to this state.

Note that our formulation of a strategy naturally uses a Mealy machine. The output mapping $\lambda_i(.,.)$ has two arguments, the state of the machine and the *input* s, which lists the outcomes of previous moves within the stage. The transitions take place at the end of the stage. The benefit of using this formulation is that the continuation game is the same at the beginning of each stage. In Chatterjee and Sabourian [16], we investigate the effects of modifying this formulation, including studying the effects of having a sub-machine to play each role. The different formulations can all implement the same strategies, but the complexities in terms of various measures could differ. We refer the reader to that paper for details, but emphasize that in the general unanimity game, the results from other formulations are similar to the one developed here, though they could differ for special cases, like three-player games.

We now consider a machine game, where players first choose machines and then the machines play the unanimity game in analogy with Abreu–Rubinstein. Using the same lexicographic utility, with complexity coming after bargaining payoffs, what do we find for Nash equilibria of the machine game?

As it turns out, the addition of complexity costs in this setting has some bite but not much. In particular, any division of the pie can be sustained in some Nash equilibrium of the machine game. Perpetual disagreement can, in fact, be sustained by a stationary machine, that is one that makes the same offers and responses each time, irrespective of past history. Nor can we prove, for general *n-player* games that the equilibrium machines will be one-state. (A three-player counter-example exists in [16]; it does not appear to be possible to generate in games that lasted less than thirty periods.) For two-player games, the result that machines must be one-state in equilibrium can be shown neatly ([16]); another illustration that in this particular area, there is a substantial increase of analytical difficulty in going from two to three players.

One reason why complexity does not appear important here is that the definition of complexity used is too restrictive. Counting the number of states is fine, so long as we don't consider how complex a response might be for partial histories *within* a stage. The next attempt at a solution is based on this observation.

We devise the following definition of complexity: Given the machine and the states, if a machine made the same response to different partial stage histories in different states and another machine made different responses, then the second one was more complex (given that the machines were identical in all other respects). We refer to this notion as *response complexity*. (In [15] the concept of response complexity is in fact stated in terms of the underlying strategy rather than in terms of machines.) It captures the intuition that counting states is not enough; two machines could have the same number of states, for example because each generated the same number of distinct offers, but the complexity of responses in one machine could be much lower than that in the other. Note that this notion would only arise in extensive-form games. In normal form games, counting states could be an adequate measure of complexity. Nor is this notion of complexity derivable from notions of transition complexity, due to Banks and Sundaram, for example, which also apply in normal-form games.

The main result of Chatterjee and Sabourian [15] is that this new aspect of complexity enables us to limit the amount of delay that can occur in equilibrium and hence to infer that only one-state machines are equilibrium machines.

The formal proofs using two different approaches are available in Chatterjee and Sabourian [15,16]. We mention the basic intuition behind these results. Suppose, in the three player game, there is an agreement in period 4 (this is in the second stage). Why doesn't this agreement take place in period 1 instead? It must be because if the same offer and responses are seen in period 1 some player will reject the offer. But of course, he or she does not have to do so because the required offer never happens. But a strategy that accepts the offer in period 4 and rejects it

off the equilibrium path in period 1 must be more complex, by our definition, than one that always accepts it whenever it might happen, on or off the expected path. Repeated application of this argument by backwards induction gives the result. (The details are more complicated but are in the papers cited above.) Note that this uses the definition that two machines might have the same number of states and yet one could be simpler than the other. It is interesting, as mentioned earlier, that for two players one can obtain an analogous result without invoking the response simplicity criterion, but from three players on this criterion is essential.

The above result (equilibrium machines have one state each and there are no delays beyond the first stage) is still not enough to refine the set of equilibria to a single allocation. In order to do this, we consider machines that can make errors/trembles in output. As the error goes to zero, we are left with perfect equilibria of our game. With one-state machines, the only subgame perfect equilibria are the ones that give equal division of the pie as $\delta \to 1$. Thus a combination of two techniques, one essentially recognizing that players can make mistakes and the other that players prefer simpler strategies if the payoffs are the same as those given by a more complex strategy, resolves the problem of multiplicity of equilibria in the multiperson bargaining game.

As we mentioned before, the introduction of errors ensures that the equilibrium strategies are credible at every history. We could also take the more direct (and easier) way of obtaining the uniqueness result with complexity costs by considering NEC strategies that are subgame perfect in the underlying game (PEC) (as done in [15]). Then since an history-independent subgame perfect equilibrium of the game is unique and any NEC automaton profile has one state and hence is history-independent, it follows immediately that any PEC is unique and induces equal division as $\delta \to 1$.

Complexity and Repeated Negotiations

In addition to standard repeated games or standard bargaining games, multiplicity of equilibria often appear in dynamic repeated interactions, where a repeated game is superimposed on an alternating offers bargaining game. For instance, consider two firms, in an ongoing vertical relationship, negotiating the terms of a merger. Such situations have been analyzed in several "negotiation models" by Busch and Wen [13], Fernandez and Glazer [18] and Haller and Holden [25]. These models can be interpreted as combining the features of both repeated and alternating-offers bargaining games. In each period, one of the two players first makes an offer on how to divide the total available periodic (flow) surplus; if the offer is accepted, the game ends with the players obtaining the corresponding payoffs in the current and every period thereafter. If the offer is rejected, they play some normal form game to determine their flow payoffs for that period and then the game moves on to the next period in which the same play continues with the players' bargaining roles reversed. One can think of the normal form game played in the event of a rejection as a "threat game" in which a player takes actions that could punish the other player by reducing his total payoffs.

If the bargaining had not existed, the game would be a standard repeated normal form game. Introducing bargaining and the prospect of permanent exit, the negotiation model still admits a large number of equilibria, like standard repeated games. Some of these equilibria involve delay in agreement (even perpetual disagreement) and inefficiency, while some are efficient.

Lee and Sabourian [31] apply complexity considerations to this model. As in Abreu and Rubinstein [1] and others, the players choose among automata and the equilibrium notion is that of NEC and PEC. One important difference however is that in this paper the authors do not assume the automata to be *finite*. Also, the paper introduces a new machine specification that formally distinguishes between the two *roles* – proposer and responder – played by each player in a given period.

Complexity considerations select only *efficient* equilibria in the negotiation model players are sufficiently patient. First, it is shown that if an agreement occurs in some finite period as a NEC outcome then it must occur within the first two periods of the game. This is because if a NEC induces an agreement beyond the first two periods then one of the players must be able to drop the last period's state of his machine without affecting the outcome of the game. Second, given sufficiently patient players, every PEC in the negotiation model that induces perpetual disagreement is at least *long-run* almost efficient; that is, the game must reach a finite date at which the continuation game then on is almost efficient.

Thus, these results take the study of complexity in repeated games a step further from the previous literature in which complexity or bargaining alone has produced only limited selection results. While, as we discussed above, many inefficient equilibria survive complexity refinement, Lee and Sabourian [31] demonstrate that complexity and bargaining in tandem ensure efficiency in repeated interactions. Complexity considerations also allow Lee and Sabourian to highlight the role of transaction costs in the negotiation game. Transaction costs take the form of pay-

ing a cost to enter the bargaining stage of the negotiation game. In contrast to the efficiency result in the negotiation game with complexity costs, Lee and Sabourian also show that introducing transaction costs into the negotiation game dramatically alters the selection result from efficiency to inefficiency. In particular, they show that, for any discount factor and any transaction cost, every PEC in the costly negotiation game induces perpetual disagreement if the stage game normal form (after any disagreement) has a unique Nash equilibrium.

Complexity, Market Games and the Competitive Equilibrium

There has been a long tradition in economics of trying to provide a theory of how a competitive market with many buyers and sellers operates. The concept of competitive (Walrasian) equilibrium (see Debreu [17]) is a simple description of such markets. In such an equilibrium each trader chooses rationally the amount he wants to trade taking the prices as given, and the prices are set (or adjust) to ensure that total demanded is equal to the total supplied. The important feature of the set-up is that agents assume that they cannot influence (set) the prices and this is often justified by appealing to the idea that each individual agent is small relative to the market.

There are conceptual as well as technical problems associated with such a justification. First, if no agent can influence the prices then who sets them? Second, even in a large but finite market a change in the behavior of a single individual agent may affect the decisions of some others, which in turn might influence the behavior of some other agents and so on and so forth; thus the market as a whole may end up being affected by the decision of a single individual.

Game theoretic analysis of markets have tried to address these issues (e. g. see [21,47]). This has turned out to be a difficult task because the strategic analysis of markets, in contrast to the simple and elegant model of competitive equilibrium, tends to be complex and intractable. In particular, dynamic market games have many equilibria, in which a variety of different kinds of behavior are sustained by threats and counter-threats.

More than 60 years ago Hayek [26] noted the competitive markets are simple mechanisms in which economic agents only need to know their own endowments, preferences and technologies and the vector of prices at which trade takes place. In such environments, economic agents maximizing utility subject to constraints make efficient choices in equilibrium. Below we report some recent work, which suggests that the converse might also be true:

> If rational agents have, at least at the margin, an aversion to complex behavior, then their maximizing behavior will result in simple behavioral rules and thereby in a perfectly competitive equilibrium

(Gale and Sabourian [22]).

Homogeneous Markets

In a seminal paper, Rubinstein and Wolinsky [46], henceforth RW, considered a market for a single indivisible good in which a finite number of homogeneous buyers and homogeneous sellers are matched in pairs and bargain over the terms of trade. In their set-up, each seller has one unit of an indivisible good and each buyer wants to buy at most one unit of the good. Each seller's valuation of the good is 0 and each buyer's valuation is 1. Time is divided into discrete periods and at each date, buyers and sellers are matched randomly in pairs and one member of the pair is randomly chosen to be the proposer and the other the responder. In any such match the proposer offers a price $p \in [0,1]$ and the responder accepts or rejects the offer. If the offer is accepted the two agents trade at the agreed price p and the game ends with the seller receiving a payoff p and the buyer in the trade obtaining a payoff $1-p$. If the offer is rejected the pair return to the market and the process continues. RW further assume that there is no discounting to capture the idea that there is no friction (cost to waiting) in the market.

Assuming that the number of buyers and sellers is not the same, RW showed that this dynamic matching and bargaining game has, in addition to a perfectly competitive outcome, a large set of other subgame perfect equilibrium outcomes, a result reminiscent of the Folk Theorem for repeated games. To see the intuition for this, consider the case in which there is one seller s and many buyers. Since there are more buyers than sellers the price of 1, at which the seller receives all the surplus, is the unique competitive equilibrium; furthermore, since there are no frictions $p = 1$ seems to be the most plausible price. RW's precise result, however, establishes that for any price $p^* \in [0,1]$ and any buyer b^* there is a subgame perfect equilibrium that results in s and b^* trading at p^*. The idea behind the result is to construct an equilibrium strategy profile such that buyer b^* is identified as the intended recipient of the good at a price p^*. This means that the strategies are such that (i) when s meets b^*, whichever is chosen as the proposer offers price p^* and the responder accepts, (ii) when s is the proposer in a match with some buyer $b \neq b^*$, s offers the good at a price of $p = 1$ and b rejects and (iii) when a buyer $b \neq b^*$ is the proposer he offers to buy the good at a price of $p = 0$ and s rejects. These strategies pro-

duce the required outcome. Furthermore, the equilibrium strategies make use of of the following punishment strategies to deter deviations. If the seller s deviates by proposing to a buyer b a price $p \neq p^*$, b rejects this offer and the play continues with b becoming the intended recipient of the item at a price of zero. Thus, after rejection by b strategies are the same as those given earlier with the price zero in place of p^* and buyer b in place of buyer b^*. Similarly, if a buyer b deviates by offering a price $p \neq p^*$ then the seller rejects, another buyer $b' \neq b$ is chosen to be the intended recipient and the price at which the unit is traded changes to 1. Further deviations from these punishment strategies can be treated in an exactly similar way.

The strong impression left by RW is that indeterminacy of equilibrium is a robust feature of dynamic market games and, in particular, there is no reason to expect the outcome to be perfectly competitive. However, the strategies required to support the family of equilibria in RW are quite complex. In particular, when a proposer deviates, the strategies are tailor-made so that the responder is rewarded for rejecting the deviating proposal. This requires coordinating on a large amount of information so that at every information set the players know (and agree) what constitutes a deviation.

In fact, RW show that if the amount of information available to the agents is strictly limited so that the agents do not recall the history of past play then the only equilibrium outcome is the competitive one. This suggests that the competitive outcome may result if agents use simple strategies. Furthermore, the equilibrium strategies used described in RW to support non-competitive outcomes are particularly unattractive because they require all players, including those buyers who do not end up trading, to follow complex non-stationary strategies in order to support a non-competitive outcome. But buyers who do not trade and receive zero payoff on the equilibrium path could always obtain at least zero by following a less complex strategy than the ones specified in RW's construction. Thus, RW's construction of non-competitive equilibria is not robust if players prefer, at least at the margin, a simpler strategy to a more complex one.

Following the above observation, Sabourian [47], henceforth S, addresses the role of complexity (simplicity) in sustaining a multiplicity of non-competitive equilibria in RW's model. The concept of complexity in S is similar to that in Chatterjee and Sabourian [15]. It is defined by a partial ordering on the set of individual strategies (or automata) that very informally satisfies the following: If two strategies are otherwise identical except that in some role the second strategy uses more information than that available in the current period of bargaining and the first uses only the information available in the current period, then the second strategy is said to be more complex than the first. S also introduces complexity costs lexicographically into the RW game and shows that any PEC is history-independent and induces the competitive outcome in the sense that all trades take place at the unique competitive price of 1.

Informally, S's conclusions *in the case of a single seller* s and many buyers follows from the following three steps. First, since trading at the competitive price of 1 is the worst outcome for a buyer and the best outcome for the seller, by appealing to complexity type reasoning it can be shown that in any NEC a trader's response to a price offer of 1 is always history-independent and thus he either always rejects 1 or always accepts 1. For example, if in the case of a buyer this were not the case, then since accepting 1 is a worst possible outcome, he could economize on complexity and obtain at least the same payoff by adopting another strategy that is otherwise the same as the equilibrium strategy except that it always rejects 1.

Second, in any non-competitive NEC in which s receives a payoff of less than 1, there cannot be an agreement at a price of 1 between s and a buyer at *any* history. For example, if at some history, a buyer is offered $p = 1$ and he accepts then by the first step the buyer should accept $p = 1$ whenever it is offered; but this is a contradiction because it means that the seller can guarantee himself an equilibrium payoff of one by waiting until he has a chance to make a proposal to this buyer.

Third, in any non-competitive PEC the continuation payoffs of all buyers are positive at every history. This follows immediately from the previous step because if there is no trade at $p = 1$ at any history it follows that each buyer can always obtain a positive payoff by offering the seller more than he can obtain in any subgame.

Finally, because of competition between the buyers (there is one seller and many buyers), in any subgame perfect equilibrium there must be a buyer with a zero continuation payoff after some history. To illustrate the basic intuition for this claim, let m be the worst continuation payoff for s at any history and suppose that there exists a subgame at which s is the proposer in a match with a buyer $\overline{b}$ and the continuation payoff of s at this subgame is m. Then if at this subgame s proposes $m + \epsilon$ ($\epsilon > 0$), $\overline{b}$ must reject (otherwise s can get more than m). Since the total surplus is 1, $\overline{b}$ must obtain at least $1 - m - \epsilon$ in the continuation game in order to reject $s's$ offer and s gets at least m, this implies that the continuation payoff of all $b \neq \overline{b}$ after $\overline{b}'s$ rejection is less than ε. The result follows by making ε arbitrarily small (and by appealing to the finiteness of f).

But the last two claims contradict each other unless the equilibrium is competitive. This establishes the result for the case in which there is one seller and many buyers. The case of a market with more than one seller is established by induction on the number of sellers.

The matching technology in the above model is random. RW also consider another market game with the matching is endogenous: At each date each seller (the short side of the market) chooses his trading partner. Here, they show that non-competitive outcomes and multiplicity of equilibria survive even when the players discount the future. By strengthening the notion of complexity S also shows that in the endogenous matching model of RW the competitive outcome is the only equilibrium if complexity considerations are present.

These results suggest perfectly competitive behavior may result if agents have, at least at the margin, preferences for simple strategies. Unfortunately, both RW and S have too simple a market set-up; for example, it is assumed that the buyers are all identical, similarly for the sellers and each agent trades at most one unit of the good. Do the conclusions extend to richer models of trade?

Heterogeneous Markets

There are good reasons to think that it may be too difficult (or even impossible) to establish a similar set of conclusions as in S in a richer framework. For example, consider a heterogeneous market for a single indivisible good, where buyers (and sellers) have a range of valuations of the good and each buyer wants at most one unit of the good and each seller has one unit of the good for sale. In this case the analysis of S will not suffice. First, in the homogeneous market of RW, except for the special case where the number of buyers is equal to the number of sellers, the competitive equilibrium price is either 0 or 1 and all of the surplus goes to one side of the market. S's selection result crucially uses this property of the competitive equilibrium. By contrast, in a heterogeneous market, in general there will be agents receiving positive payoffs on both sides of the market in a competitive equilibrium. Therefore, one cannot justify the competitive outcome simply by focusing on extreme outcomes in which there is no surplus for one party from trade. Second, in a homogeneous market individually rational trade is by definition efficient. This may not be the case in a heterogeneous market (an inefficient trade between inframarginal and an extramarginal agent can be individually rational). Third, in a homogeneous market, the set of competitive prices remains constant, independently of the set of agents remaining in the market. In the heterogeneous market, this need not be so and in some cases, the new competitive interval may not even intersect the old one. The change in the competitive interval of prices as the result of trade exacerbates the problems associated with using an induction hypothesis because here future prices may be conditioned on past trades even if prices are restricted to be competitive ones.

Despite these difficulties associated with a market with a heterogeneous set of buyers and sellers, Gale and Sabourian [22], henceforth GS, show that the conclusions of S can be extended to the case of a heterogeneous market in which each agent trades at most one unit of the good. GS, however, focus on *deterministic* sequential matching models in which one pair of agents are matched at each date and they leave the market if they reach an agreement. In particular, they start by considering exogenous matching processes in which the identities of the proposer and responder at each date are an exogenous and deterministic function of the set of agents remaining in the market and the date. The main result of the paper is that a PEC is always competitive in in such a heterogeneous market, thus supporting the view that competitive equilibrium may arise in a finite market where complex behavior is costly.

The notion of complexity in GS is similar to that in S [15]. However, in the GS set-up with heterogeneous buyers and sellers the set of remaining agents changes depending who has traded and left the market and who is remaining, and this affects the market conditions. (In the homogeneous case, only the *number* of remaining agents matters.) Therefore, the definition of complexity in GS is with reference to a given set of remaining agents. GS also discuss an alternative notion of complexity that is independent of the set of remaining agents; such a definition may be too strong and may result in an equilibrium set being empty.

To show their result, GS first establish two very useful restrictions on the strategies that form a NEC (similar to the no delay result in Chatterjee and Sabourian [15]). First, they show that if along the equilibrium path a pair of agents k and ℓ trade at a price p with k as the proposer and ℓ as the responder then k and ℓ always trade at p, irrespective of the previous history, whenever the two agents are matched in the same way with the same remaining set of agents. To show this consider first the case of the responder ℓ. Then it must be that at every history with the same remaining set of agents ℓ always accepts p by k. Otherwise, ℓ could economize on complexity by choosing another strategy that is otherwise identical to his equilibrium strategy except that it always accepts p from k without sacrificing any payoff: Such a change of behavior is clearly more simple than sometimes accepting and sometimes re-

jecting the offer and moreover, it results in either agent k proposing p and ℓ accepting, so the payoff to agent ℓ is the same as from the equilibrium strategy, or agent k not offering p, in which case the change in the strategy is not observed and the play of the game is unaffected by the deviation. Furthermore, it must also be that at every history with the same remaining set of agents agent k proposes p in any match with ℓ. Otherwise, k could economize on complexity by choosing another strategy that is otherwise identical to his equilibrium strategy except that it always proposes p to ℓ without sacrificing any payoff on the equilibrium path: Such a change of behavior is clearly more simple and moreover k's payoff is not affected because either agent k and ℓ are matched and k proposes p and ℓ by the previous argument accepts, so the payoff to agent k is the same as from the equilibrium strategy, or agent k and ℓ are not matched with k as the proposer, in which case the change in the strategy is not observed and the play of the game is unaffected by the deviation.

GS's shows a second restriction, again with the same remaining set of agents, namely that in any NEC for any pair of agents k and ℓ, player ℓ's response to k's (on or off-the-equilibrium path) offer is always the same. Otherwise, it follows that ℓ sometimes accepts an offer p by k and sometimes rejects (with the same remaining set of agents). Then by the first restriction it must be that if such an offer is made by k to ℓ on the equilibrium path it is rejected. But then ℓ can could economize on complexity by always rejecting p by k without sacrificing any payoff on the equilibrium path: Such a change of behavior is clearly more simple and furthermore ℓ's payoff is not affected because such a behavior is the same as what the equilibrium strategy prescribes on the equilibrium path.

By appealing to the above two properties of NEC and to the competitive nature of the market GS establish, using a complicated induction argument, that every PEC induces a competitive outcome in which each trade occurs at the same competitive price.

The matching model we have described so far is deterministic and exogenous. The selection result of GS however extends to richer deterministic matching models. In particular, GS also consider a semi-endogenous sequential matching model in which the choice of partners is endogenous but the identity of the proposer at any date is exogenous. Their results extends to this variation, with an endogenous choice of responders. A more radical departure change would be to consider the case where at any date any agent can choose his partner and make a proposal. Such a totally endogenous model of trade generates new conceptual problems. In a recent working paper Gale and Sabourian [24] consider a continuous time version of such a matching model and show that complexity considerations allows one to select a competitive outcome in the case of totally endogenous matching. Since the selection result holds for all the different matching models we can conclude that complexity considerations inducing a competitive outcome seem to be a robust result in deterministic matching and bargaining market games with heterogeneous agents.

Random matching is commonly used in economic models because of its tractability. The basic framework of GS, however, does not extend to such a framework if either the buyers or the sellers are not identical. This is for two different reasons. First, in general in any random framework there is more than one outcome path that can occur in equilibrium with a positive probability; as a result introducing complexity lexicographically may not be enough to induce agents to behave in a simple way (they will have to be complex enough to play optimally along all paths that occur with a positive probability). Second, in Gale and Sabourian [23] it is shown that subgame perfect equilibria in Markov strategies are not necessarily perfectly competitive for the random matching model with heterogeneous agents. Since the definition of complexity in GS is such that Markov strategies are the least complex ones, it follows that with random matching the complexity definition used in GS is not sufficient to select a competitive outcome.

Complexity and Off-The-Equilibrium Path Play

The concept of the PEC (or NEC) used in S, GS and elsewhere was defined to be such that for each player the strategy/automaton has minimal complexity amongst all strategies/automata that are best responses to the equilibrium strategies/automata of others. Although, these concepts are very mild in the treatment of complexity, it should be noted that there are other ways of introducing complexity into the equilibrium concept. One extension of the above set-up is to treat complexity as a (small) positive fixed cost of choosing a more complex strategy and define a Nash (subgame perfect) equilibrium with a fixed positive complexity costs accordingly. All the selection results based on lexicographic complexity in the papers we discuss in this survey also hold for positive small complexity costs. This is not surprising because with positive costs complexity has at least as much bite as in the lexicographic case; there is at least as much refinement of the equilibrium concept with the former as with the latter. In particular, in the case of a NEC (or a PEC), *in considering complexity*, players ignore any consideration of payoffs off the equilibrium path and the trade-off is between the equilibrium payoffs of two strategies and the complexity of the two.

As a result these concepts put more weight on complexity costs than on being "prepared" for off-the-equilibrium-path moves. Therefore, although complexity costs are insignificant, they take priority over optimal behavior after deviations. (See [16] for a discussion.)

A different approach would be to assume that complexity is a less significant criterion than the off-the-equilibrium payoffs. In the extreme case, one would require agents to choose minimally complex strategies among the set of strategies that are best responses on and off the equilibrium path (see Kalai and Neme [28]).

An alternative way of illustrating the differences between the different approaches is by introducing two kinds of vanishingly small perturbations into the underlying game. One perturbation is to impose a small but positive cost of choosing a more complex strategy. Another perturbation is to introduce a small but positive probability of making an error (off-the-equilibrium-path move). Since a PEC requires each agents to choose a minimally complex strategy within the set of best responses, it follows that the limit points of Nash equilibria of the above perturbed game correspond to the concept of PEC if we first let the probability of making an off-the-equilibrium-path move go to zero and then let the cost of choosing a more complex strategy go to zero (this is what Chatterjee and Sabourian [15] do). On the other hand, in terms of the above limiting arguments, if we let the cost of choosing a more complex strategy go to zero and then let the probability of making an off-the-equilibrium-path move go to zero then any limit corresponds to the equilibrium definition in Kalai and Neme [28] where agents choose minimally complex strategies among the set of strategies that are best responses on and off the equilibrium path.

Most of the results reported in this paper on refinement and endogenous complexity (for example Abreu–Rubinstein [1]), Chatterjee and Sabourian [15], Gale and Sabourian [22] and Lee and Sabourian [31] hold only for the concept of NEC and its variations and thus depend crucially on assuming that complexity costs are more important than off-the-equilibrium payoffs. This is because these results always appeal to an argument that involves economizing on complexity if the complexity is not used off the equilibrium path. Therefore, they may be a good predictor of what may happen only if complexity costs are more significant than the perturbations that induce off-the-equilibrium-path behavior. The one exception is the selection result in S [47]. Here, although the result we have reported is stated for NEC and its variations, it turns out that the selection of competitive equilibrium does not in fact depend on the relative importance of complexity costs and off-the-equilibrium path payoffs. It remains true even for the case where the strategies are required to be least complex amongst those that are best responses at every information set. This is because in S's analysis complexity is only used to show that every agent's response to the price offer of 1 is always the same irrespective of the past history of play. This conclusion holds irrespective of the relative importance of complexity costs and off-the-equilibrium payoff because trading at the price of 1 is the best outcome that any seller can achieve at any information set (including those off-the-equilibrium) and a worst outcome for any buyer. Therefore, irrespective of the order, the strategy of sometimes accepting a price of 1 and sometimes rejecting cannot be an equilibrium for a buyer (similar arguments applies for a seller) because the buyer can economize on complexity by always rejecting the offer without sacrificing any payoff off or on-the-equilibrium path (accepting $p = 1$ is a worse possible outcome).

Discussion and Future Directions

The use of finite automata as a model of players in a game has been criticized as being inadequate, especially because as the number of states becomes lower it becomes more and more difficult for the small automaton to do routine calculations, let alone the best response calculations necessary for game-theoretic equilibria. Some of the papers we have explored address other aspects of complexity that arise from the concrete nature of the games under consideration. Alternative models of complexity are also suggested, such as computational complexity and communication complexity.

While our work and the earlier work on which it builds focuses on equilibrium, an alternative approach might seek to see whether simplicity evolves in some reasonable learning model. Maenner [32] has undertaken such an investigation with the infinitely repeated Prisoners' Dilemma (studied in the equilibrium context by Abreu and Rubinstein). Maenner provides an argument for "learning to be simple". On the other hand, there are arguments for increasing complexity in competitive games ([42]). It is an open question, therefore, whether simplicity could arise endogenously through learning, though it seems to be a feature of most human preferences and aesthetics (see [11]).

The broader research program of explicitly considering complexity in economic settings might be a very fruitful one. Auction mechanisms are designed with an eye towards how complex they are – simplicity is a desideratum. The complexity of contracting has given rise to a whole literature on incomplete contracts, where some models postulate a fixed cost per contingency described in the con-

tract. All this is apart from the popular literature on complexity, which seeks to understand complex, adaptive systems from biology. The use of formal complexity measures such as those considered in this survey and the research we describe might throw some light on whether incompleteness of contracts, or simplicity of mechanisms, is an assumption or a result (of explicitly considering choice of level of complexity).

Acknowledgments

We wish to thank an anonymous referee and Jihong Lee for valuable comments that improved the exposition of this chapter. We would also like to thank St. John's College, Cambridge and the Pennsylvania State University for funding Dr Chatterjee's stay in Cambridge at the time this chapter was written.

Bibliography

1. Abreu D, Rubinstein A (1988) The structure of Nash equilibria in repeated games with finite automata. Econometrica 56:1259–1282
2. Anderlini L (1990) Some notes on Church's thesis and the theory of games. Theory Decis 29:19–52
3. Anderlini L, Sabourian H (1995) Cooperation and effective computability. Econometrica 63:1337–1369
4. Aumann RJ (1981) Survey of repeated games. In: Essays in game theory and mathematical economics in honor of Oskar Morgenstern. Bibliographisches Institut, Mannheim/Vienna/Zurich, pp 11–42
5. Banks J, Sundaram R (1990) Repeated games, finite automata and complexity. Games Econ Behav 2:97–117
6. Ben Porath E (1986) Repeated games with bounded complexity. Mimeo, Stanford University
7. Ben Porath E (1993) Repeated games with finite automata. J Econ Theory 59:17–32
8. Binmore KG (1987) Modelling rational players I. Econ Philos 3:179–214
9. Binmore KG, Samuelson L (1992) Evolutionary stability in repeated games played by finite automata. J Econ Theory 57:278–305
10. Binmore KG, Piccione M, Samuelson L (1998) Evolutionary stability in alternating-offers bargaining games. J Econ Theory 80:257–291
11. Birkhoff GD (1933) Aesthetic measure. Harvard University Press, Cambridge
12. Bloise G (1998) Strategic complexity and equilibrium in repeated games. Unpublished doctoral dissertation, University of Cambridge
13. Busch L-A, Wen Q (1995) Perfect equilibria in a negotiation model. Econometrica 63:545–565
14. Chatterjee K (2002) Complexity of strategies and multiplicity of Nash equilibria. Group Decis Negot 11:223–230
15. Chatterjee K, Sabourian H (2000) Multiperson bargaining and strategic complexity. Econometrica 68:1491–1509
16. Chatterjee K, Sabourian H (2000) N-person bargaining and strategic complexity. Mimeo, University of Cambridge and the Pennsylvania State University
17. Debreu G (1959) Theory of value. Yale University Press, New Haven/London
18. Fernandez R, Glazer J (1991) Striking for a bargain between two completely informed agents. Am Econ Rev 81:240–252
19. Fudenberg D, Maskin E (1990) Evolution and repeated games. Mimeo, Harvard/Princeton
20. Fudenberg D, Tirole J (1991) Game theory. MIT Press, Cambridge
21. Gale D (2000) Strategic foundations of general equilibrium: Dynamic matching and bargaining games. Cambridge University Press, Cambridge
22. Gale D, Sabourian H (2005) Complexity and competition. Econometrica 73:739–770
23. Gale D, Sabourian H (2006) Markov equilibria in dynamic matching and bargaining games. Games Econ Behav 54:336–352
24. Gale D, Sabourian H (2008) Complexity and competition II; endogenous matching. Mimeo, New York University/University of Cambridge
25. Haller H, Holden S (1990) A letter to the editor on wage bargaining. J Econ Theory 52:232–236
26. Hayek F (1945) The use of knowledge in society. Am Econ Rev 35:519–530
27. Herrero M (1985) A Strategic theory of market institutions. Unpublished doctoral dissertation, London School of Economics
28. Kalai E, Neme A (1992) The strength of a little perfection. Int J Game Theory 20:335–355
29. Kalai E, Stanford W (1988) Finite rationality and interpersonal complexity in repeated games. Econometrica 56:397–410
30. Klemperer P (ed) (2000) The economic theory of auctions. Elgar, Northampton
31. Lee J, Sabourian H (2007) Coase theorem, complexity and transaction costs. J Econ Theory 135:214–235
32. Maenner E (2008) Adaptation and complexity in repeated games. Games Econ Behav 63:166–187
33. Miller GA (1956) The magical number seven plus or minus two: Some limits on our capacity to process information. Psychol Rev 63:81–97
34. Neme A, Quintas L (1995) Subgame perfect equilibrium of repeated games with implementation cost. J Econ Theory 66:599–608
35. Neyman A (1985) Bounded complexity justifies cooperation in the finitely-repeated Prisoners' Dilemma. Econ Lett 19:227–229
36. Neyman A (1997) Cooperation, repetition and automata in cooperation: Game-theoretic approaches. In: Hart S, Mas-Colell A (eds) NATO ASI Series F, vol 155. Springer, Berlin, pp 233–255
37. Osborne M, Rubinstein A (1990) Bargaining and markets. Academic, New York
38. Osborne M, Rubinstein A (1994) A course in game theory. MIT Press, Cambridge
39. Papadimitriou CH (1992) On games with a bounded number of states. Games Econ Behav 4:122–131
40. Piccione M (1992) Finite automata equilibria with discounting. J Econ Theory 56:180–193
41. Piccione M, Rubinstein A (1993) Finite automata play a repeated extensive game. J Econ Theory 61:160–168
42. Robson A (2003) The evolution of rationality and the Red Queen. J Econ Theory 111:1–22

43. Rubinstein A (1982) Perfect equilibrium in a bargaining model. Econometrica 50:97–109
44. Rubinstein A (1986) Finite automata play the repeated Prisoners' Dilemma. J Econ Theory 39:83–96
45. Rubinstein A (1998) Modeling bounded rationality. MIT Press, Cambridge
46. Rubinstein A, Wolinsky A (1990) Decentralized trading, strategic behaviour and the Walrasian outcome. Rev Econ Stud 57:63–78
47. Sabourian H (2003) Bargaining and markets: Complexity and the competitive outcome. J Econ Theory 116:189–228
48. Selten R (1965) Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit. Z gesamte Staatswiss 12:201–324
49. Shaked A (1986) A three-person unanimity game. In: The Los Angeles national meetings of the Institute of Management Sciences and the Operations Research Society of America, Mimeo, University of Bonn
50. Zemel E (1989) Small talk and cooperation: A note on bounded rationality. J Econ Theory 49:1–9

GARCH Modeling

Christian M. Hafner
Université catholique de Louvain,
Louvain-la-Neuve, Belgium

Article Outline

Glossary
Definition of the Subject
Introduction
Properties of the GARCH(1,1) Model
Estimation and Inference
Testing for ARCH
Asymmetry, Long Memory, GARCH-in-Mean
Non- and Semi-parametric Models
Multivariate GARCH Models
Stochastic Volatility
Aggregation
Future Directions
Bibliography

Glossary

ACF Autocorrelation Function

ARMA Autoregressive Moving Average

BEKK A multivariate GARCH model named after an early unpublished paper by Baba, Engle, Kraft and Kroner.

CCC Constant Conditional Correlation

DCC Dynamic Conditional Correlation

CAPM Capital Asset Pricing Model

GARCH Generalized Autoregressive Conditional Heteroskedasticity

Heteroskedasticity A non-constant variance that depends on the observation or on time.

i.i.d. independent, identically distributed

Kurtosis A standardized fourth moment of a random variable that tells something about the shape of the distribution. A Gaussian distribution has a kurtosis of three. If the kurtosis is larger than three, then typically the distribution will have tails that are thicker than those of the Gaussian distribution.

Lag An operation that shifts the time index of a time series. For example, the first lag of y_t is y_{t-1}.

Long memory Property of covariance stationary processes without absolutely summable ACF, meaning that the ACF decays slowly.

Realized volatility Sum of intra-day squared returns as a measure for daily volatility.

Skewness A standardized third moment of a random variable that tells something about the asymmetry of the distribution. Symmetric distributions have skewness equal to zero.

Volatility Degree of fluctuation of a time series around its mean.

Definition of the Subject

GARCH (Generalized Autoregressive Conditional Heteroskedasticity) is a time series model developed by [44] and [21] to describe the way volatility changes over time. In a GARCH model, the volatility at a given time t, σ_t^2 say, is a function of lagged values of the observed time series y_t. The GARCH model can be written as $y_t = \sigma_t \xi_t$, with ξ_t being an independent, identically distributed (i.i.d.) error term with mean zero and variance one, and where

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i y_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \tag{1}$$

with constant parameters ω, $\alpha_1, \ldots, \alpha_q$ and $\beta_1, \ldots, \beta_p$. Model (1) is also called GARCH(p, q), analogous to ARMA(p, q), as it includes p lagged volatilities and q lagged squared values of y_t. In this model, σ_t^2 is the variance of y_t conditional on the observations until time $t-1$. It is specified as a linear function of lagged squared y_t and lagged conditional variances. Many extensions or modifications of the basic model in (1) have been proposed, the most prominent being the exponential GARCH model of [99] and the threshold GARCH models of [123] and [59]. [71] and [42] provided classes of models that contain a large number of suggested models of the GARCH type.

Introduction

In the late seventies of the last century it became obvious that volatilities of financial assets are indeed not constant, nor deterministic or seasonal, but rather stochastic in nature. There is an unsystematic change between periods of high volatility and periods of low volatility. This 'volatility clustering' had already been remarked in the early works of [93] and [54]. It was one of several stylized facts of financial asset returns, another of which was the observation that the distribution of returns is not Gaussian. Of course, these features were not necessarily treated in an independent way, and in fact it was soon discovered that very likely one of the effects was causing another, such as volatility clustering causing leptokurtosis, or fat tailed distributions. For example, consider the simple model for asset returns y_t,

$$y_t = \sigma_t \xi_t$$

where $\xi_t \sim N(0,1)$ and σ_t is stochastic with $\mathrm{E}[\sigma_t^2] = \sigma^2$, say, and independent of present and future ξ_t. Then it is straightforward to show that the kurtosis of y_t is given by

$$\kappa = \frac{\mathrm{E}[y_t^4]}{\mathrm{E}[y_t^2]^2} = 3 + 3\frac{\mathrm{Var}(\sigma_t^2)}{\sigma^4}\,. \tag{2}$$

Thus, returns in this model are Gaussian distributed if and only if $\mathrm{Var}(\sigma_t^2) = 0$, i. e., volatility is non-stochastic. Moreover, as the second term on the right hand side of (2) is always positive, the kurtosis will be larger than three under stochastic volatility, which often means that its tails are fatter than those of the Gaussian distribution. In other words, extreme events are more likely under stochastic volatility compared with constant volatility.

To illustrate the effects of volatility clustering and fat tails, consider the daily returns on the Dow Jones Industrial Index over the period October 1928 to April 2007. A graph of the (log) index X_t and returns, defined as $y_t = X_t - X_{t-1}$, is given in Fig. 1. Clearly visible is the volatility clustering in the beginning of the sample period and around the year 2000, while the years at the end of the sample showed less volatility. Also visible is the crash of October 1987 where the index dropped by 22 percent.

Figure 2 shows a nonparametric estimator of the logarithmic density of returns, compared with the analogue of a Gaussian distribution. Clearly, the Dow Jones returns distributions has fat tails, i. e., there are more extreme events than one would expect under normality. There are also more returns close to zero than under normality. Concerning volatility clustering, Fig. 3 shows the autocorrelation function of returns and squared returns. While there is very little structure in the ACF of returns, the ACF of squared returns are all positive and highly significant. This positive autocorrelation is explained by the fact that large returns tend to be followed by large returns and small returns tend to be followed by small returns.

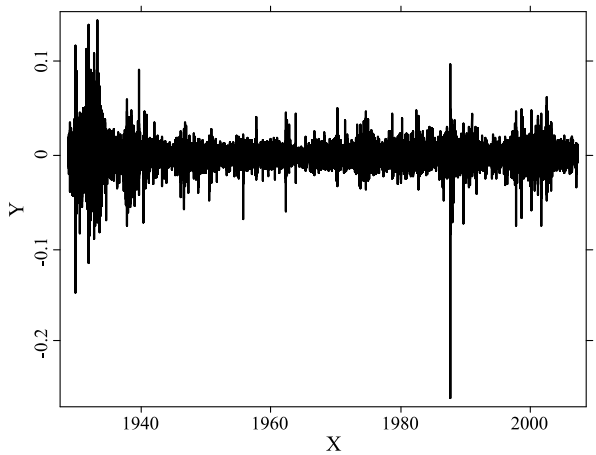

GARCH Modeling, Figure 1
Daily returns of the Dow Jones Index, 1928 to 2007, defined as first difference of the log index

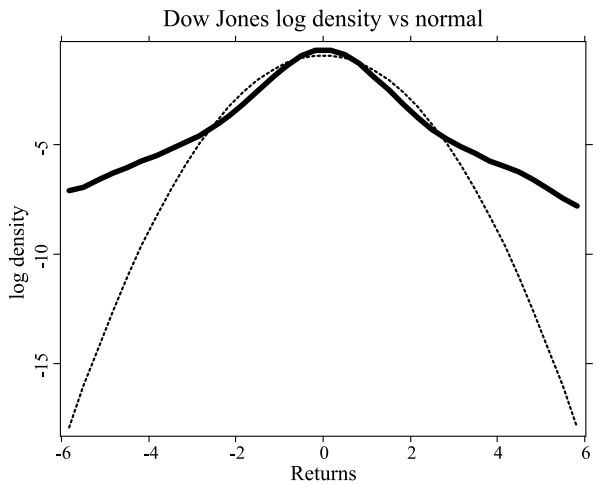

GARCH Modeling, Figure 2
Dow Jones log density versus the Gaussian log density

Yet, realizing that volatility is stochastic does not tell us which model we should use for it. In practice, people are sometimes debating whether they should take historical volatilities over 20 or 100 days, say. They notice that calculating the standard deviation over a shorter period is more accurate when recent upturns or downturns want to be captured, while it is far less efficient than a longer time window when volatility has not changed much. Thus, there is some kind of bias-variance trade-off. The problem is that the optimal window length typically changes

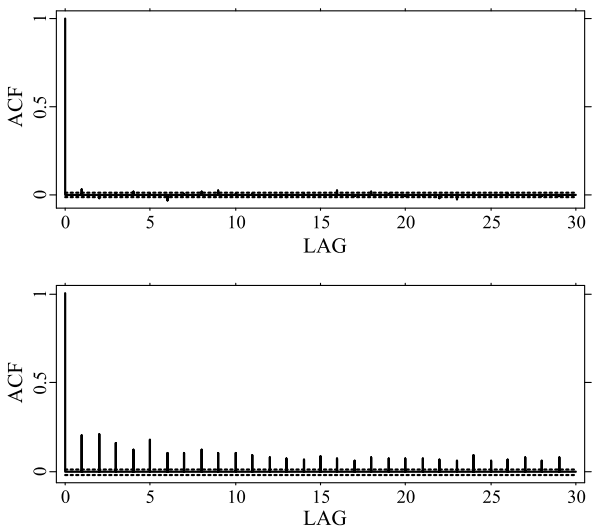

GARCH Modeling, Figure 3
Dow Jones autocorrelation function of returns (*upper panel*) and squared returns (*lower panel*)

over time, and it is virtually impossible to adjust historical volatility windows automatically to market developments. A related problem is that historical volatilities imply a weighting scheme that is highly questionable: Why should k days be incorporated in the calculation with equal weight, but no weights are put to days up to $k+1$ days ago? A smoother weighting scheme seems more natural, and in particular, an exponential scheme seems attractive. Thus, for example, we may specify for σ_t^2

$$\sigma_t^2 = (1-\lambda)\sum_{i=0}^{\infty}\lambda^i y_{t-1-i}^2 \tag{3}$$

with parameter $\lambda \in (0,1)$. Equation (3) can be rewritten as

$$\sigma_t^2 = (1-\lambda)y_{t-1}^2 + \lambda\sigma_{t-1}^2\,, \tag{4}$$

which looks more familiar. It is actually the model used by RiskMetrics of JP Morgan, when the smoothing parameter is fixed to 0.94. RiskMetrics is often used in practice as a means to calculate the Value-at-Risk (VaR) of a portfolio and to assess the market risk of a bank, required by the Basel Committee for Banking Supervision, see e. g., [78] and [95]. The VaR is essentially an extreme quantile of the distribution of portfolio returns. Under Gaussianity, for example, the VaR is a direct function of volatility. The RiskMetrics model is a special case of the integrated GARCH model of [47].

The generalized autoregressive conditional heteroskedasticity – GARCH – model is based on the seminal work of [44] and [21]. The idea is to do exponential smoothing in a more flexible way than RiskMetrics but keeping the model parsimonious. The particular specification reveals many similarities to autoregressive moving average (ARMA) time series models. In its most often used form, the standard GARCH model of order (1,1) reads

$$\sigma_t^2 = \omega + \alpha y_{t-1}^2 + \beta\sigma_{t-1}^2 \tag{5}$$

where ω, α and β are parameters to be estimated from the data. Thus, the conditional variance is a linear function of lagged squared observations y_t and lagged conditional variances. Comparing (5) with the RiskMetrics model (4), it becomes clear that in the GARCH(1,1) model a constant is added, the parameter α takes the role of $1-\lambda$ and β that of λ. But since α and β can be chosen independently, the GARCH model is more flexible than RiskMetrics. In (5), substituting successively for σ_{t-i}^2, one obtains the analogue representation of (3),

$$\sigma_t^2 = \frac{\omega}{1-\beta} + \alpha\sum_{i=0}^{\infty}\beta^i y_{t-1-i}^2\,, \tag{6}$$

which clearly shows the exponential smoothing feature of the GARCH(1,1) model. The basic model can now be extended to allow for more lags. The GARCH(p, q) model is given by

$$\sigma_t^2 = \omega + \sum_{i=1}^{q}\alpha_i y_{t-i}^2 + \sum_{j=1}^{p}\beta_j\sigma_{t-j}^2 \tag{7}$$

extending the number of parameters to $p+q+1$. GARCH models of order higher than (1,1) allow for more complex autocorrelation structures of the squared process. However, in most empirical studies coefficients corresponding to higher lags turned out to be insignificant and thus simple GARCH(1,1) have clearly dominated models of higher order.

Although extremely successful due to its simplicity and yet accurate description of volatility changes, a thorough understanding of its stochastic properties such as stationarity or positivity constraints took many years. For example, [98] shows that, paradoxically at first sight, conditions for strict stationarity are less rigid than those for covariance stationarity if error terms are Gaussian, the reason being that covariance stationarity requires finite variances whereas strict stationarity does not. Moreover, the often given parameter restrictions $\omega > 0$, $\alpha_i, \beta_j \geq 0$ are only sufficient but not necessary for $\sigma_t > 0$ almost surely as demonstrated by [100]. These are just two examples for the subtleties of the theory of univariate GARCH processes.

Nevertheless, the immense success of the simple GARCH(1,1) model to explain many sorts of financial and

macroeconomic time series was irreversible, partly also because it became available in standard statistical programming packages. The theory of estimation and inference developed rapidly, although perhaps still being underway, and estimation time of a GARCH(1,1) model for a thousand or so observations decreased from minutes in the eighties over seconds in the nineties to just fractions of a second nowadays. With these developments it became available to a broad public, and more and more practitioners started using the model, be it for option pricing, portfolio optimization, risk management, or other purposes. Monographs and reviews appeared such as [14,20,60] and [13]. Anniversary issues of renowned journals such as Journal of Applied Econometrics, 2002, were dedicated entirely to new ideas in GARCH modeling. The Nobel price for economics in 2003 was awarded to two time series econometricians, Clive Granger and Robert Engle. The latter has mainly driven the development of a new financial econometrics discipline, based on volatility modeling but spreading also to other areas such as modeling of extreme events and risk management.

The pricing of options and other derivatives is perhaps the most typical example for where models for the volatility of an asset matter. For example, the celebrated option pricing formula of [18] does not depend on the drift of the underlying stock but well on its volatility. In fact, among the ingredients of the Black and Scholes formula, volatility is the most crucial one, the other ones being either fixed such as time of maturity or strike price, or relatively easy to determine such as a riskfree interest rate. Volatility, however, has always been subject to debates about how exactly to find accurate measures for it. The Black and Scholes assumption of constant volatility is actually less crucial to their formula than one often thinks. Actually, if volatility is time-varying but in a deterministic way, then the Black and Scholes formula remains valid. One just has to replace the volatility parameter by the mean of the volatility function from today until the time of maturity of the option contract, see e. g., [90]. If, however, volatility is stochastic, i. e., it has an additional source of randomness, then markets are no longer complete and the Black and Scholes formula breaks down. In that case, assumptions about the volatility risk premium have to be made. In continuous time stochastic volatility models a classical paper is [74], while in a discrete time GARCH framework, [41] derives results for option pricing.

Properties of the GARCH(1,1) Model

For the sake of simplicity let us consider the univariate GARCH(1,1) model given in (5), where we additionally assume that the conditional distribution of y_t is Gaussian. The model can be written as

$$y_t = \sigma_t \xi_t, \quad \xi_t \sim \text{i.i.d. } N(0,1) \tag{8}$$

$$\sigma_t^2 = \omega + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2 \,. \tag{9}$$

In the following we discuss a few properties of model (8). First, the GARCH model specifies the *conditional* variance, where the condition is the information set generated by the process y_t. Formally, it is given by the sigma-algebra $\mathcal{F}_t = \sigma(y_t, y_{t-1}, \ldots)$. With this notation we can write $\sigma_t^2 = \mathrm{Var}(y_t|\mathcal{F}_{t-1})$, since σ_t^2 is $\mathcal{F}_{t-1}$-measurable. As the information set changes over time, the conditional variance also changes. On the other hand, this does not imply that the *unconditional* variance is also time-varying. In fact, for model (8) it is quite straightforward to show that the unconditional variance, if it exists, is constant and given by

$$\mathrm{Var}(y_t) = \frac{\omega}{1-\alpha-\beta}.$$

A necessary and sufficient condition for the existence of the unconditional variance is

$$\alpha + \beta < 1 \,, \tag{10}$$

see [21]. He also shows that condition (10) is necessary and sufficient for the process $\{y_t\}$ to be covariance stationary. In that case, the autocorrelation function of $\{y_t\}$ is given by $\rho_\tau(y_t) = 0, \forall \tau \geq 1$. Moreover, both the conditional and unconditional mean of y_t are zero, so that the process $\{y_t\}$ has the properties of a white noise without being an i.i.d. process. The dependence occurs in higher moments of the process. For example, the autocorrelation function of the squared process, provided that fourth moments exist, is given by

$$\rho_1(y_t^2) = \alpha \frac{1-\alpha\beta-\beta^2}{1-2\alpha\beta-\beta^2} \tag{11}$$

$$\rho_\tau(y_t^2) = (\alpha+\beta)\rho_{\tau-1}(y_t^2), \quad \tau \geq 2 \,. \tag{12}$$

From (11) and (12) it is obvious that in the GARCH(1,1) model all autocorrelations of squared returns are positive with an exponential decay. This decay is slow if $\alpha + \beta$ is close to one, as often found for empirical data. One also characterizes this coefficient as the "persistence" parameter of the GARCH(1,1) model. The closer the persistence parameter is to one, the longer will be the periods of volatility clustering. On the other hand, the larger α relative to β, the higher will be the immediate impact of lagged squared returns on volatility.

The necessary and sufficient condition for finite fourth moments is given by $\beta^2 + 2\alpha\beta + 3\alpha^2 < 1$, see [21]. In that case, the kurtosis of y_t is given by

$$\kappa = 3 + \frac{6\alpha^2}{1 - \beta^2 - 2\alpha\beta - 3\alpha^2} ,$$

which is larger than three since the second term on the right hand side is positive. Hence, the GARCH(1,1) exhibits fat tails compared with a normal distribution.

A link between the GARCH model and an ARMA model is given by considering the squared process, $\{y_t\}$. By simply adding y_t^2 and subtracting σ_t^2 on both sides of Eq. (9), one obtains

$$y_t^2 = \omega + (\alpha + \beta) y_{t-1}^2 - \beta u_{t-1} + u_t \tag{13}$$

where $u_t = y_t^2 - \sigma_t^2$. Equation (13) is an ARMA(1,1) in y_t^2, since u_t is a white noise error term: We have $\mathrm{E}[u_t | \mathcal{F}_{t-1}] = 0$, which implies that all autocorrelations of u_t are zero.

It is possible that the process $\{y_t\}$ is strictly stationary without being covariance stationary, simply because a finite variance is not necessary for strict stationarity. If the process starts in the infinite past, a necessary and sufficient condition for strict stationarity of the GARCH(1,1) process as shown by [98] is given by

$$\mathrm{E}[\log(\alpha \xi_t^2 + \beta)] < 0 , \tag{14}$$

which is indeed weaker than condition (10). This follows directly by noting that (10) is equivalent to $\log(\alpha + \beta) = \log(\mathrm{E}[\alpha \xi_t^2 + \beta]) < 0$. Thus, by Jensen's inequality, $\mathrm{E}[\log(\alpha \xi_t^2 + \beta)] < \log(\mathrm{E}[\alpha \xi_t^2 + \beta]) < 0$. For example, for an ARCH(1) model (i. e., a GARCH(1,1) model with $\beta = 0$), α can be as large as 3.56 and still the process is strictly stationary. Figure 4 shows the different stationarity regions as a function of the two parameters. [25] generalized condition (14) to the GARCH(p, q) case.

Under the sufficient condition (10), the GARCH(1,1) process with Gaussian innovations is also geometrically ergodic and β-mixing with exponential decay as shown by [28].

If condition (14) holds, then the process $\{y_t\}$ has a stationary distribution whose tails are of the Pareto type. That is, for large x and some $a, k > 0$,

$$p(x) = Pr(y_t > x) = kx^{-a} . \tag{15}$$

The coefficient a is known as the tail index. The smaller a, the fatter the tail of the distribution. For all $c, 0 \le c < \alpha$, $\mathrm{E}[|y_t|^c] < \infty$. [43] showed that a stationary ARCH model has Pareto-like tails. Knowledge of the tail index is important for risk management in order to assess the risk of extreme events. The theoretical tail index of a fitted ARCH or GARCH model can be compared with an estimate of the empirical tail index in order to diagnose the goodness-of-fit with respect to the tails. For example, taking logarithms of (15), one obtains $\log p(x) = \log(k) - a \log(x)$ for large x. Replacing x by the largest m order statistics of y_t, and introducing an error term, one obtains the regression

$$\log \frac{i}{n} = \log k - a \log X_{(i)} + \varepsilon_i , \quad i = 1, \ldots, m \tag{16}$$

where $X_{(i)}$ are the largest m order statistics of y_t and ε_i is an error term. One can estimate the coefficients of the linear regression (16) simply by ordinary least squares. More problematic is the choice of m, which involves a bias-variance trade-off. For the Dow Jones returns, Fig. 5 shows the tail index regression using $m = 30$. The OLS estimator of a is 3.12, indicating that fourth moments of returns may not exist. Another simple estimator is the Hill estimator proposed by [72], which is based on a likelihood principle. For the Dow Jones returns, the Hill estimator of a using $m = 30$ is 2.978, which is close to the OLS estimator, suggesting that even third moments may not exist. More elaborate estimators have been proposed and we refer to the detailed discussion in [43].

The presence of autoregressive conditional heteroskedasticity has an effect on the forecast intervals for

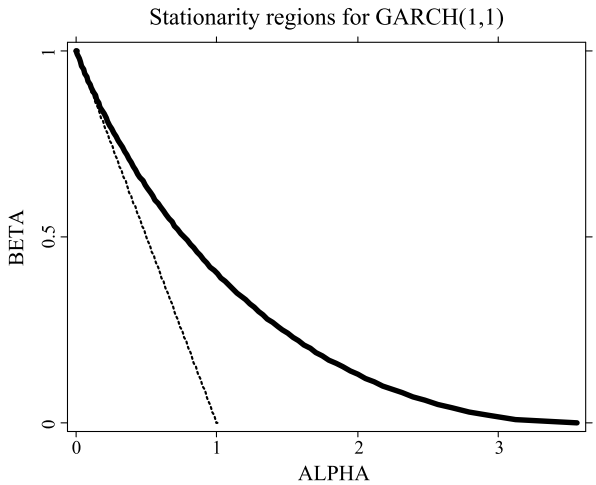

GARCH Modeling, Figure 4
Stationarity regions for a GARCH(1,1) process with Gaussian innovations. To the *left* of the *dashed line* is the region of covariance stationarity, to the *left* of the *thick line* is the region of strict stationarity, and to the *right* of the *thick line* is the region of non-stationarity

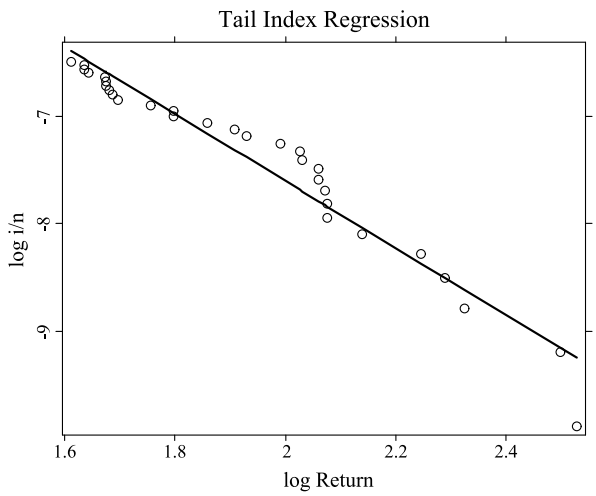

GARCH Modeling, Figure 5
Tail index regression for the Dow Jones returns

predicted y_{t+k} given information at time t. If volatility at time t as measured by the GARCH model is high (low), these will be larger (smaller) than if GARCH effects are ignored. Furthermore, forecasting the volatility itself is easily possible with the standard GARCH model, since analytical expressions can be found for the conditional mean of future volatility as a function of today's information. The conditional mean is the optimal predictor in a mean square prediction error sense. For example, to forecast σ^2_{t+k}, one derives for a forecast horizon of $k \geq 2$,

$$\mathrm{E}[\sigma^2_{t+k}|\mathcal{F}_t] = \omega(1 + (\alpha+\beta) + \cdots + (\alpha+\beta)^{k-2}) + (\alpha+\beta)^{k-1}\sigma^2_{t+1}\,.$$

If the process is covariance stationary, i. e., $\alpha + \beta < 1$, then volatility forecasts converge to the unconditional variance:

$$\lim_{k\to\infty} \mathrm{E}[\sigma^2_{t+k}|\mathcal{F}_t] = \frac{\omega}{1-\alpha-\beta} = \mathrm{Var}(y_t)\,.$$

In the early literature on GARCH models, these were criticized for not providing good forecasts in terms of conventional forecast criteria. For example, when regressing the ex post squared daily returns on the forecasted conditional variance, the obtained R^2 is typically small, of the order of about ten percent. [4] found that the daily squared return is not really the targeted value, but that daily volatility should rather be measured by the sum of *intra-day* squared returns, e. g., on intervals of five minute returns, which they called *realized volatility*. In terms of realized volatility, the forecasting performance of GARCH models improved substantially to levels of about fifty percent R^2. Later, a new branch of volatility modeling opened by noticing that if intra-day data are available, then it is indeed more efficient to measure daily volatility directly by realized volatility and then do forecasting of daily volatility using models fitted to realized volatility, see e. g., [5].

Estimation and Inference

The principal estimation method for GARCH models is maximum likelihood (ML). In most cases one assumes a conditional Gaussian distribution. If the true distribution is Gaussian, then ML estimators are consistent and efficient under quite general conditions. On the other hand, if the true distribution is not Gaussian, then one loses efficiency but again under quite general conditions, consistency is retained if at least the first two conditional moments are correctly specified, see [23]. In the case of misspecification of the conditional distribution one also speaks of *quasi* maximum likelihood (QML), distinguishing it from ML where the true distribution is used, which however in general is unknown.

The log likelihood function, up to an additive constant and conditional on some starting value for the volatility process, reads $L(\theta) = \sum_{t=1}^n l_t(\theta)$, where

$$l_t(\theta) = -\frac{1}{2}\log\sigma^2_t(\theta) - \frac{1}{2}\sum_{t=1}^n \frac{y_t^2}{\sigma^2_t(\theta)}$$

and where $\theta = (\omega, \alpha, \beta)'$ is the parameter vector. The maximum likelihood estimator is then defined as the maximizer of $L(\theta)$ over some compact set Θ,

$$\hat{\theta} = \arg\max_{\theta\in\Theta} L(\theta)\,. \tag{17}$$

Unfortunately, there is no closed form solution to (17) but many numerical optimization procedures exist. For example, a popular algorithm is that of [15].

Figure 6 shows the likelihood function of a GARCH(1,1) process generated using the parameter estimates of the Dow Jones index returns (see Sect. "Asymmetry, Long Memory, GARCH-in-Mean"), Gaussian innovations, and the same sample size of $n = 19727$. The parameter ω has been determined by the variance targeting technique of [50], i. e., $\omega = \sigma^2(1-\alpha-\beta)$, where σ^2 is the sample variance of observed returns. In regions where $\alpha + \beta \geq 1$, ω is set to zero. Note the steep decline of the likelihood for values of α and β that lie beyond the covariance stationarity region ($\alpha + \beta \geq 1$). Figure 7 shows the same function for the observed Dow Jones index returns. No major difference can be detected between both graphs, indicating an appropriate specification of the Gaussian likelihood function.

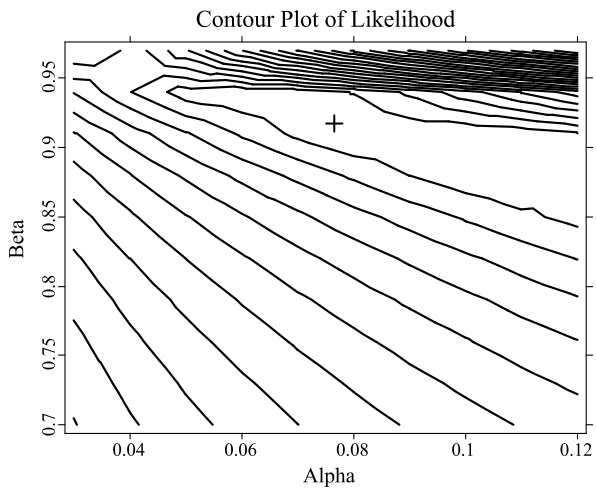

GARCH Modeling, Figure 6
Contour plot of the likelihood function of a generated GARCH(1,1) process using Gaussian innovations and a sample size of $n = 19727$. The abscissa is the parameter α, the ordinate is β. True values, marked by a *cross* in the figure, are $\alpha = 0.0766$ and $\beta = 0.9173$

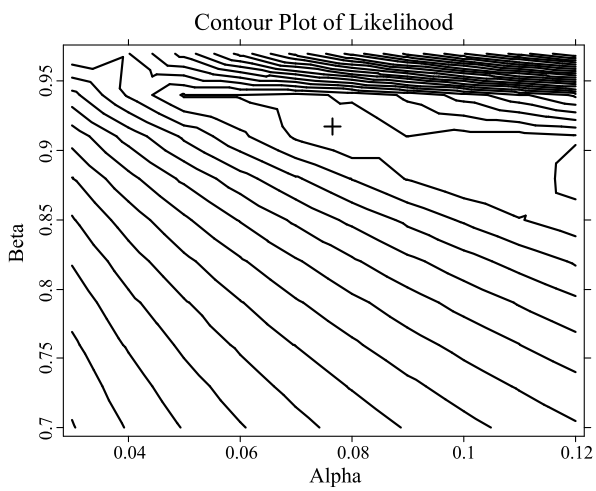

GARCH Modeling, Figure 7
Contour plot of the likelihood function of the GARCH(1,1) model fitted to observed Dow Jones index returns, 1928 to 2007, with sample size $n = 19727$. The *abscissa* is the parameter α, the *ordinate* is β. The maximum, marked by a *cross* in the figure, is obtained for $\alpha = 0.0766$ and $\beta = 0.9173$

If the first two moments of y_t are correctly specified and under further regularity conditions given by [118] and [23], the QML estimator is consistent with asymptotic distribution given by

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} N(0, \mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1}) \tag{18}$$

where

$$\mathcal{I} = \mathrm{E}\left[\frac{\partial l_t}{\partial \theta}\frac{\partial l_t}{\partial \theta'}\right], \quad \mathcal{J} = -\mathrm{E}\left[\frac{\partial^2 l_t}{\partial \theta \partial \theta'}\right],$$

and where the derivatives are evaluated at the true parameter values. In case the conditional distribution is indeed Gaussian, one has the identity $\mathcal{I} = \mathcal{J}$ and the asymptotic covariance matrix reduces to the inverse of the information matrix, $\mathcal{I}^{-1}$. Note that consistency is retained if the conditional distribution is not Gaussian, but efficiency is lost in that case.

It is straightforward to obtain analytical formula for the score vector, the outer product of the score and the Hessian matrix, with which inference on parameter estimates can be done using the result in (18). More primitive conditions than those of [23] have been derived, e. g., by [57,85,91] and [66].

Maximum likelihood estimation using other than Gaussian distributions has been considered, e. g., by [101]. He shows that if the distribution is misspecified, then consistency is no longer guaranteed. In particular, if a symmetric distribution is assumed but the true distribution is asymmetric, then maximum likelihood estimators are inconsistent. In practice, a common distribution used for maximum likelihood estimation is the Student t distribution. Given the results of [101], one should be careful in interpreting parameter estimates if there is evidence for skewness in standardized residuals.

Another estimation strategy based on maximum likelihood is a nonparametric estimation of the error density, which has been advocated by [48]. They suggest to use a first stage estimator of the model parameters, which is consistent but not efficient such as the Gaussian MLE, to construct residuals and then to use nonparametric methods to estimate the error density. Given the estimated error density, one can maximize the likelihood corresponding to this nonparametric density function. These estimators will under regulatory conditions be consistent and more efficient than the Gaussian ML estimator, provided that the true density is different from Gaussian.

A potential practical problem of maximum likelihood estimators is its dependence on numerical optimization routines. Recently, a closed form estimator based on the autocorrelation structure of squared returns has been suggested by [82]. Their estimator is inefficient compared to ML but has the advantage of being uniquely determined by the data. Further Monte Carlo evidence is necessary to see whether it is a serious practical competitor for ML-type estimators. Least squares type estimators of ARCH(q) have been considered by [118] and [103]. Again, these are inefficient compared with maximum likelihood estimators

but simpler to compute. [104] suggest a least absolute deviation estimator for GARCH models that is robust with respect to outliers but does not allow for a closed form. Finally, Bayesian estimation of GARCH-type models has been investigated, e. g., by [12,115] and [32].

Testing for ARCH

In a regression such as

$$y_t = \mu_t + \varepsilon_t \tag{19}$$

where μ_t is measurable w.r.t. $\mathcal{F}_{t-1} = \sigma(y_{t-1}, y_{t-2}, \ldots)$ and ε_t is a white noise sequence, inference on μ_t typically depends on the properties of the error term ε_t. For example, if ε_t is i.i.d. Gaussian and μ_t is linear such as an AR(p) model, then estimation by least squares of the autoregressive coefficients in μ_t is efficient. If, however, ε_t is not i.i.d. and for example conditionally heteroskedastic, then estimation by ordinary least square (OLS) is no longer efficient and some kind of generalized least squares may be employed. Moreover, inference on the parameters in μ_t will be erroneous if homoskedasticity of ε_t is assumed but, in reality, ε_t is conditionally heteroskedastic. In particular, standard errors in that case are typically underestimated. To avoid this, it is essential to test for ARCH type effects in ε_t. The following testing procedure, based on the Lagrange multiplier principle, has been proposed in the original ARCH paper by [44]. The null hypothesis is that ε_t is i.i.d. white noise, the alternative is the presence of ARCH. One first estimated the model (19) by least squares, obtains residuals $\hat{\varepsilon}_t$, and then runs the regression

$$\hat{\varepsilon}_t^2 = \alpha_0 + \alpha_1 \hat{\varepsilon}_{t-1}^2 + \alpha_2 \hat{\varepsilon}_{t-2}^2 + \cdots + \alpha_q \hat{\varepsilon}_{t-q}^2 + \eta_t \tag{20}$$

where η_t is an error term. Under the null hypothesis H_0, $\alpha_1 = \ldots = \alpha_q = 0$. The test statistic is $\lambda = nR^2$, where n is the sample size and R^2 the coefficient of determination of the regression (20). Under H_0, the test statistic follows asymptotically a χ^2 distribution with q degrees of freedom. Hence, it is an elementary exercise to test for ARCH effects in the error term of regression models. Historically, it is remarkable that prior to the introduction of the ARCH model, the above LM test was used by Prof. Clive Granger as an LM test for a bilinear error term, for which it has some power. Then, Prof. Robert Engle discovered that it has more power for another model, which he then introduced as the ARCH model.

An alternative to the LM test of [44] would be a Wald-type test of the hypothesis $H_0\colon \alpha = 0$ in the GARCH(1,1) model (5) using, e. g., the t-ratio as test statistic. However, this test is non-standard since under the null hypothesis the parameter α is on the boundary of the parameter space and the parameter β is not identified. [6] treats this test in a general framework.

Asymmetry, Long Memory, GARCH-in-Mean

In the standard GARCH model in (7), positive and negative values of lagged returns y_{t-i} have the same impact on volatility, since they appear in squares in the equation for σ_t^2. Empirically, it has been frequently noted since [17] that for stock markets, negative returns increase volatility more than positive returns do. Essentially, this so-called leverage effect means that negative news have a stronger impact on volatility than positive ones. To account for this empirical observation, several extensions of the standard GARCH model have been proposed in the literature. The most commonly used are the exponential GARCH model of [99] and the threshold GARCH model of [59] and [123]. The threshold model in its first order variant is given by the process

$$\sigma_t^2 = \omega + \alpha y_{t-1}^2 + \alpha^* y_{t-1}^2 I(y_{t-1} < 0) + \beta \sigma_{t-1}^2$$

where α^* is an additional parameter and $I(\cdot)$ is the indicator function. If $\alpha^* = 0$, then the threshold model reduces to the standard GARCH model. If $\alpha^* > 0$, then negative returns have a stronger impact on volatility than positive ones, which corresponds to the empirical observation for stock markets.

Secondly, the exponential GARCH (EGARCH) model of [99] specifies log-volatility as

$$\log \sigma_t^2 = \omega + \theta \xi_{t-1} + \alpha(|\xi_{t-1}| - \mathrm{E}|\xi_{t-1}|) + \beta \log \sigma_{t-1}^2$$

where $\xi_t = y_t/\sigma_t$ is i.i.d. with a generalized error distribution (GED) which nests the Gaussian and allows for slightly fatter tails. Due to the specification of log-volatility, no parameter restrictions are necessary to keep volatility positive. Moreover, the conditions for weak and strong stationarity coincide. Note that if $\theta \neq 0$, then $\mathrm{Cov}(y_t^2, y_{t-j}) \neq 0$ such that a leverage effect can be captured. A drawback of the EGARCH model is that asymptotic theory for maximum likelihood estimation and inference under primitive conditions are not available yet, but [110] are making much progress in this respect.

Another model allowing for asymmetry is the asymmetric power ARCH (APARCH) model [36]. In its (1,1) order form it specifies volatility as

$$\sigma_t^\delta = \omega + \alpha(|y_{t-1}| - \phi y_{t-1})^\delta + \beta \sigma_{t-1}^\delta$$

where δ is a positive parameter. If $\delta = 2$ and $\phi = 0$, the standard GARCH model is retained. For $\phi \neq 0$, there is

an asymmetric impact of positive and negative lagged returns on volatility. The additional flexibility due to the parameter δ allows to better reproduce the so-called 'Taylor property,' originally noted by [111], which says that the autocorrelations of $|y_t|^d$ are positive even at long lags, and when viewed as a function of d take a maximum for $d \approx 1$ for many financial returns y_t. [70] provide a formal discussion of this issue.

The standard GARCH(p, q) model (7) implies that the decay of the autocorrelation function (ACF) of squared returns is geometrically fast. However, one often finds evidence for a slow hyperbolical decay in financial time series, see for example Fig. 3. The decay pattern of the ACF is related to the structure of coefficients c_j in the ARCH(∞) representation of GARCH models,

$$\sigma_t^2 = c_0 + \sum_{j=1}^{\infty} c_j y_{t-j}^2 \,. \tag{21}$$

For example, in the GARCH(1,1) model, these are given by $c_j = \alpha\beta^{j-1}$. Covariance stationary GARCH models have the property that the autocovariance function of squared returns, $\gamma(\tau) = \mathrm{Cov}(y_t^2, y_{t-\tau}^2)$, is absolutely summable, i. e., $\sum_\tau |\gamma(\tau)| < \infty$. Such a property is commonly called *short memory* as opposed to *long memory* processes for which the ACF is not absolutely summable. Long memory GARCH models have been proposed by [8] and [19], see also the review of long memory processes in econometrics by [7]. An example of a long memory GARCH process would be given by (21) with $c_j = Cj^{-\theta}$ for some constant C and parameter $\theta > 0$. A particular example for such a process is the fractionally integrated GARCH (FIGARCH) model of [8], which can be written as

$$(1-L)^d \sigma_t^2 = \omega + \alpha y_{t-1}^2$$

where L is the lag operator and d a positive parameter. When $d = 1$ one obtains the integrated GARCH (IGARCH) model of [47]. For $d \neq 1$ one can use a binomial extension to obtain after inverting

$$(1-L)^{-d} = \sum_{j=0}^{\infty} \frac{\Gamma(j+d)}{\Gamma(j+1)\Gamma(d)} L^j = \sum_{j=0}^{\infty} c_j L^j \tag{22}$$

where $\Gamma(\cdot)$ is the Gamma function. The coefficient c_j in (22) can be shown to be of the long memory type. A similar long memory EGARCH model has been introduced by [19]. The drawback of these particular specifications is that they share the property with the IGARCH model to have infinite variance. [105] has proposed a long memory GARCH type model that allows for finite variance.

Finally, in the finance literature a link is often made between the expected return and the risk of an asset, since investors are willing to hold risky assets only if their expected returns compensate for the risk. A model that incorporates this link is the GARCH-in-mean or GARCH-M model of [52], given by

$$y_t = \delta g(\sigma_t^2) + \varepsilon_t$$

where ε_t is an ARCH or GARCH error process, δ a parameter, and g a known function such as square root or logarithm. If $\delta > 0$ and g is monotone increasing, then the term $\delta g(\sigma_t^2)$ can be interpreted as a risk premium that increases expected returns $\mathrm{E}[y_t]$ if volatility σ_t^2 is high. It can be shown that such a model, when applied to the market index, is consistent with the capital asset pricing model (CAPM) of [108] and [87], see [24].

As an empirical illustration we estimate alternative models for the Dow Jones index discussed in the introduction. To recall, we have daily returns from October 1928 to April 2007. First order autocorrelation of returns is 0.03, which due to the large number of observations is significant at the level 1%. However, we refrain here from fitting an autoregressive or moving average model to the returns as the results concerning volatility estimation do not change substantially. We only consider a constant conditional mean in the model $y_t = \mu + \varepsilon_t$, where ε_t is one of the discussed GARCH-type models and μ takes into account a non-zero trend in returns. Six alternative GARCH models are considered, all of them being of order (1,1): standard GARCH, TGARCH, EGARCH, GARCH-M, TGARCH-M and EGARCH-M. For the "in-mean" versions, we have chosen the square root specification for the function $g(\cdot)$, which seems to work better than the logarithm or the identity function. Moreover, for all "in-mean" models the constant μ turned out to be insignificant and hence was suppressed from the model. Table 1 summarizes the estimation results.

Note first that all estimated risk premia are positive $\mu > 0$ and $\delta > 0$, as theory would predict. Second, for all models allowing for asymmetry, the leverage effect of negative returns is confirmed, i. e., $\alpha^* > 0$ for the TGARCH models and $\theta < 0$ for the EGARCH models. Third, in all cases persistence of shocks to volatility is very high, measured by $\alpha + \beta$ in the GARCH model, $\alpha + \alpha^*/2 + \beta$ in the TGARCH model (assuming a symmetric innovation distribution), and by β in the EGARCH model. Thus, all models are short memory with exponential decay of the ACF of squared returns, but the models try to adapt to the empirically observed slow decay of the ACF by pushing the persistence parameter close to one. This near-IGARCH

GARCH Modeling, Table 1

Estimation results for the following models: GARCH, EGARCH, TGARCH, GARCH-in-mean, TGARCH-in-mean and EGARCH-in-mean, applied to daily Dow Jones returns from 1928 to 2007. All parameters are significant at the one percent level

	G	TG	EG	GM	TGM	EGM
μ	4.25E-04	2.64E-04	2.36E-04			
δ				0.0591	0.0354	0.0286
ω	8.74E-07	1.03E-06	-0.2184	8.75E-07	1.06E-06	-0.2227
α	0.0766	0.0308	0.1391	0.0766	0.0306	0.1378
α^*		0.0769			0.0761	
θ			-0.0599			-0.0595
β	0.9173	0.9208	0.9879	0.9172	0.9205	0.9874
L	65456.47	65600.53	65589.32	65462.45	65603.79	65590.69

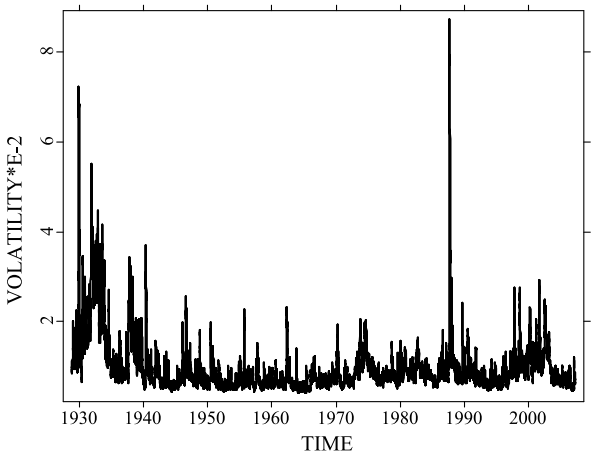

GARCH Modeling, Figure 8

Estimated conditional standard deviation of daily Dow Jones index returns, 1928 to 2007, using the TGARCH-in-mean model

behavior is typical for daily returns. Finally, the goodness-of-fit seems to be best for the TGARCH-in-mean model, taking the log-likelihood as criterion. The estimation results strongly confirm the presence of the leverage effect, high persistence, and positive risk premium in the data. Figure 8 shows the estimated conditional standard deviation of the TGARCH-M model. For the other models, the graph would look quite similar and is therefore not shown here. Notice the very volatile periods at the beginning of the sample in the 1930s, around the year 2000 corresponding to the "new economy" boom and following crash, as well as the spike in 1987 due to the crash of October 17, 1987.

Non- and Semi-parametric Models

Nonparametric methods refrain from associating particular parametric forms to functions or distributions. Instead, only the class of functions is determined, for example the class of squared integrable functions or the degree of smoothness. The price for the flexibility is typically slower convergence rates than parametric models. A combination of the two approaches is often called semiparametric. One such approach has already been mentioned in Sect. "Estimation and Inference" in the context of estimation by maximum likelihood using nonparametric estimates of the error density, as proposed by [48] for GARCH models. [88] shows that this procedure leads to adaptive estimation of the identifiable parameters (α and β in a GARCH(1,1) model) in the sense of [16]. That is, it is possible to achieve the Cramer–Rao lower bound and do as good as if one knew the true error distribution. The scale parameter ω, however, is not adaptively estimable. See also [40] and [37] for related results for univariate GARCH models, and [65] for an extension to semiparametric estimation of multivariate GARCH models.

A different approach is to directly model the volatility process in a nonparametric way. Early models were proposed by [61] and [51]. The qualitative threshold ARCH model of [61] specifies models of the type

$$y_t = \sum_{j=1}^{J} \sigma_j I(y_{t-1} \in A_j)\xi_t \,, \tag{23}$$

where (A_j) is a partition of the real line and σ_j, $j = 1, \ldots, J$, are positive parameters. Thus, volatility is modeled as a piecewise constant function of lagged returns. Note that the threshold ARCH model of [59] and [123] is not a special case of (23) as there the volatility function is piecewise quadratic in lagged returns. Extensions to the ARCH(q) and GARCH(p, q) are straightforward. [51] replaced the piecewise constant functions by piecewise linear functions. In both cases, one may consider their models as nonparametric if the partition becomes finer as the sample size increases.

Consider the model

$$y_t = \sigma(y_{t-1})\xi_t$$

where $\sigma(\cdot)$ is an unknown smooth function, and $\xi_t \sim$ i.i.d. $N(0,1)$. For $\sigma^2(x) = \alpha x^2$ one obtains the parametric ARCH(1) model of [44]. An example of a nonparametric estimator of $\sigma(\cdot)$ is the Nadaraya–Watson estimator given by

$$\hat{\sigma}^2(x) = \frac{\sum_{t=2}^{n} K\{(y_{t-1} - x)/h\} y_t^2}{\sum_{t=2}^{n} K\{(y_{t-1} - x)/h\}}$$

where K is a kernel function satisfying $\int K(x)\mathrm{d}x = 1$ and $\int xK(x)\mathrm{d}x = 0$, and where $h > 0$ is a bandwidth that determines the degree of smoothing. As the Nadaraya–Watson estimator can be interpreted as fitting a constant locally, a generalization consists of fitting a local polynomial instead. This has been derived by [68] for the volatility case.

A general problem of nonparametric methods is the so-called *curse of dimensionality* when smoothing has to operate in high dimensions. Considering a nonparametric ARCH(q) model,

$$y_t = \sigma(y_{t-1}, \ldots, y_{t-q})\xi_t$$

this problem is apparent and in practice very large data sets are required to estimate the function g with appropriate precision. One may be inclined to impose more structure on the g function such as additive or multiplicative separability. Nonparametric multiplicative ARCH models have been proposed by [63] and [120]. Semi-parametric additive ARCH models of the type $\sigma^2(y_{t-1}, \ldots, y_{t-q}) = \sum_{j=1}^{p} \beta^{j-1} g(y_{t-j})$ with some unknown function g and parameter $\beta \in (0,1)$ have been considered by [29].

Extension of nonparametric ARCH(q) models to nonparametric GARCH(1,1) models have also been proposed. However, in its general form $y_t = \sigma_t \xi_t$ with $\sigma_t = g(y_{t-1}, \sigma_{t-1})$, the model is difficult to estimate due to lack of structure. One might consider iterative estimation algorithms, based on some initial estimate of volatility as in [26].

Imposing a similar semi-parametric structure as for the semi-parametric ARCH(q) model of [29], one can write $\sigma_t^2 = g(y_{t-1}) + \beta\sigma_{t-1}^2$, where again $g(\cdot)$ is an unknown smooth function. Note that this model nests many of the proposed parametric models. It has been considered by [119] and [89].

In practice, nonparametric methods may be used whenever it is not a priori clear what functional form fits best the data, either by using them directly, or as a tool to specify a parametric model in a second stage.

Multivariate GARCH Models

In economics and finance, one typically deals with multiple time series that are fluctuating in a non-systematic manner and are considered as realizations of stochastic processes. The interest for applied econometricians is therefore to model their risk, that is, their volatility, but also their inter-dependencies. For example, if one has reasons to assume that the underlying stochastic processes are Gaussian, then the inter-dependencies may be completely described by the correlation structure. In fact, when we say 'multivariate volatility models' we usually mean the modeling of volatilities but also that of correlations. This is also the reason why the extension of univariate volatility to multivariate volatility models is much more complex than that of univariate models for the conditional mean, such as ARMA models, to the multivariate case.

It will be immediately clear that the multivariate case is the one that is by far more relevant in practice when financial markets are under study. The reason is, first, the large number of different assets, or even different types of contracts, assets, exchange rates, interest rates, options, futures, etc. Second, there is usually a strong link between these variables, at least within one group. For example, asset returns in one stock market tend to be quite strongly correlated. One would make big approximation errors when treating the variables as independent by, e. g., using univariate volatility models for the conditional variances and set conditional covariances to zero. Note that, setting conditional covariance to zero is much stronger an assumption than setting the unconditional covariance to zero. Ways must be found to treat the dependence of the series in a flexible yet parsimonious way.

A first step would again be to do exponential smoothing à la RiskMetrics, which can be used not only to obtain the individual variances according to (4), but also to obtain the correlations. To see this, we define just as in (4) an exponential smoother for the covariances as

$$\sigma_{12,t} = (1-\lambda)\varepsilon_{1,t-1}\varepsilon_{2,t-1} + \lambda\sigma_{12,t-1}$$

and then obtain as usual the conditional correlation as

$$\rho_t = \frac{\sigma_{12,t}}{\sigma_{1,t}\sigma_{2,t}},$$

which is guaranteed to be between minus one and one if the same parameter λ is used, typically $\lambda = 0.94$. Figure 9 depicts the RiskMetrics conditional correlation series for the DOW and NASDAQ return series.

Obviously, conditional correlations are not constant, although it is difficult from the graph to verify such a statement statistically. However, one thing to observe is that

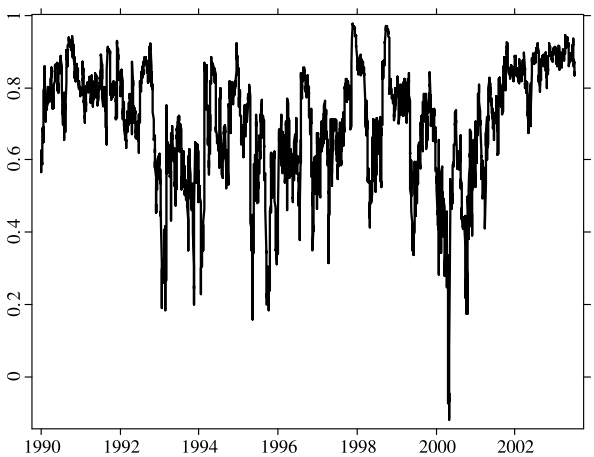

GARCH Modeling, Figure 9
Conditional correlations of the Dow Jones IA and NASDAQ index returns, daily, using the RiskMetrics model

during the New Economy boom in 1999 and 2000, estimated correlations have been substantially lower, sometimes even negative, than at other times. The reason is probably a decoupling due to the higher vulnerability of the NASDAQ with respect to the bubble in high tech and internet stocks. A more thorough analysis of this data set which also compares this model with other, more flexible models is provided by [45]. We see that the RiskMetrics tool, even though very simple, can give some guidelines. One of the objectives of the econometrician is to enhance the model in terms of flexibility (e. g., why should λ be fixed to 0.94?), and to establish a statistical framework in which hypotheses such as constant conditional correlations can be tested.

From an econometrical viewpoint, modeling the volatility of multiple time series is, for several reasons, challenging both theoretically and practically. For the sake of illustration, consider a bivariate GARCH model of the Vec type that was introduced by [24]. Denote by H_t the conditional variance matrix of the asset return vector y_t. Then a bivariate ARCH(1) model reads

$$H_t = \begin{pmatrix} h_{1,t} & h_{12,t} \\ h_{12,t} & h_{2,t} \end{pmatrix}$$

and where

$$\begin{aligned} h_{1t} &= \omega_1 + \alpha_{11}\varepsilon_{1,t-1}^2 + \alpha_{12}\varepsilon_{1,t-1}\varepsilon_{2,t-1} + \alpha_{13}\varepsilon_{2,t-1}^2 \\ h_{12,t} &= \omega_2 + \alpha_{21}\varepsilon_{1,t-1}^2 + \alpha_{22}\varepsilon_{1,t-1}\varepsilon_{2,t-1} + \alpha_{23}\varepsilon_{2,t-1}^2 \\ h_{2t} &= \omega_3 + \alpha_{31}\varepsilon_{1,t-1}^2 + \alpha_{32}\varepsilon_{1,t-1}\varepsilon_{2,t-1} + \alpha_{33}\varepsilon_{2,t-1}^2 \,. \end{aligned}$$

Each conditional variance, h_{1t} and h_{2t}, and conditional covariance, $h_{12,t}$, depends on all lagged squared returns (two in the bivariate case) and all lagged cross-products (one in the bivariate case). The main reason for the rapidly increasing complexity of the model when the dimension is increased lies in the fact that not only all conditional variances with their cross-dependencies have to be modeled, but also all conditional correlations. It is in fact the latter that poses the main problem, as there are a total of $N(N-1)/2$ such correlations when the dimension is N, but only N variances. Thus, modeling variances and correlations simultaneously, a total of $N(N+1)/2$ entries of the conditional covariance matrix need to be modeled. For example, if $N = 10$ (a moderate dimension for many economic or financial problems) this number is 55, if $N = 100$ (such as modeling all stocks of a common stock index), then 5050 series, conditional variances and covariances, are at stake.

It is clear that this is too much to allow for a flexible cross-dependence of the individual series. Without imposing any structure except linearity, the multivariate generalization of the standard GARCH model, the so-called Vec model introduced by [24], is feasible only for low dimensions, two or three say, as otherwise the number of parameters becomes too high relative to the number of observations typically encountered in economic practice. Another problem is that the Vec model does not guarantee a positive definite covariance matrix. Necessary conditions for the latter desirable property are as yet unknown in the general Vec specification.

These are some reasons to look for other models, and in fact, over recent years a broad variety of different approaches to the problem have been suggested in the literature. Roughly speaking, one can divide them into two groups. The first one tries to simplify the problem by imposing more structure on the Vec model. Examples are the BEKK model of [49] and the factor GARCH model by [53]. More recently, the second group tries to separate the problem of modeling the conditional variances and conditional correlations. An early and simple version of this group is to say that conditional variances are just univariate GARCH and conditional correlations are constant over time, as suggested by [22]. In its simplicity, this constant conditional correlation (CCC) model basically does not add any complexity beyond univariate GARCH to the multivariate estimation problem, which renders the model extremely useful in empirical practice. It also introduced the idea of two-step estimation, where in the first step conditional variances are modeled, and in the second step the conditional correlations using the standardized residuals of the first step. However, starting with [45] there have been plenty of arguments in favor of time-varying conditional correlations in financial markets. In par-

ticular, a common finding is that correlations are higher when the market moves up than when it moves down. A test for this correlation asymmetry has been suggested by [73]. Using a dynamic conditional correlation model (DCC), [45] shows that time varying correlations are not uncommon even in normal market situations. In the following we sketch these two branches of the multivariate GARCH literature. It should however be mentioned that there are models that do not fall into these two categories such as a multivariate version of the exponential GARCH model proposed by [79].

Factor GARCH Models

In the following factor GARCH models are discussed as an example of multivariate GARCH models. The main idea of factor models is to reduce the dimension of the system to a tractable two or three factors, which can then be modeled in a standard way. It should be noted that also 'full-factor' models with number of factors equal to the number of variables have been proposed in the literature. For example, [116] propose the model

$$y_t = W f_t$$

where W is a $N \times N$ parameter matrix and f_t is a N-vector with conditional mean zero and diagonal conditional variance matrix, Σ_t say. The individual conditional variances of f_t can be modeled by univariate GARCH(1,1), for example. One can restrict W to be lower triangular, as it is well known that the Choleski decomposition of a positive definite matrix always exists and is unique. Thus, the conditional variance matrix of y_t is given by $H_t = W\Sigma_t W = L_t L_t'$, where $L_t = W\Sigma_t^{1/2}$ is lower triangular. In this model, the parameters in W and those in Σ_t need to be estimated jointly, which may be cumbersome in high dimensions. The empirical performance of such full factor models still remains to be investigated.

It is more common to specify only a few factors and allow for idiosyncratic noise. We will look at such models in the following. Suppose that there are K (observed or unobserved) factors, collected in a K-vector f_t, with $K < N$. Then a simple factor model can be written as

$$y_t = W f_t + v_t \tag{24}$$

where v_t is a white noise vector with $\mathrm{Var}(v_t) = \Omega$ that represents the idiosyncratic noise. Typically, one assumes that Ω is diagonal so that components of the idiosyncratic noise are uncorrelated. In that case, correlation between components of y_t is induced only through the common factors f_t. If y_t represents the error of a time series system, one may constrain f_t to have conditional mean zero. The matrix W is of dimension $N \times K$, of full column rank, and contains the so-called *factor loadings*, the weights of a factor associated with the individual components of y_t.

In finance, model (24) is well known from the arbitrage pricing theory (APT) of [106], where y_t are excess returns of financial assets, f_t are systematic risk factors and v_t is unsystematic risk. It can also be viewed as a generalization of the capital asset pricing model (CAPM) developed by [108] and [87]. For simplicity we assume here that factors are observed. If they are unobserved, identification issues arise that are discussed, e. g., by [107].

For the factors, a low-dimensional GARCH model can be assumed: $\mathrm{Var}(f_t \mid \mathcal{F}_{t-1}) = \Sigma_t$, where Σ_t is a $(K \times K)$ covariance matrix. The conditional covariance matrix of y_t is given by

$$H_t = \mathrm{Var}(y_t \mid \mathcal{F}_{t-1}) = W\Sigma_t W' + \Omega\,. \tag{25}$$

In the case of just one factor, the matrix W reduces to a vector w and the factor volatility, σ_t^2 say, can be modeled by univariate GARCH and the conditional variance of y_t simplifies to

$$H_t = ww'\sigma_t^2 + \Omega\,.$$

If the factors are conditionally uncorrelated, i. e., Σ_t is diagonal with $\Sigma_t = \mathrm{diag}(\sigma_{1t}^2, \ldots, \sigma_{Kt}^2)$, then one can write

$$H_t = \sum_{k=1}^{K} w_k w_k' \sigma_{kt}^2 + \Omega$$

where w_k is the kth column of W. [83] propose methods to test for the number of factors K and derive results for maximum likelihood estimation. For the more general BEKK model class, [31] derived asymptotic theory but assuming moments of order eight of the process, which may exclude many of the typically fat-tailed financial time series.

A popular factor GARCH model is the orthogonal GARCH (OGARCH) model of [3]. In the OGARCH model, factors f_t are the K largest principal components obtained from the (unconditional) sample covariance matrix, and the loading matrix W is the matrix of associated eigenvectors. The loadings represent the sensitivity of an individual series on a specific factor. By construction, the unconditional correlation between the factors is zero, due to the orthogonality of the principal components. However, the *conditional* correlation may be different from zero. Denote the (empirical) covariance matrix of y_t by Σ. The decomposition $\Sigma = \Gamma\Lambda\Gamma'$ gives $\Gamma = (\gamma_1, \ldots, \gamma_N)$ with the eigenvectors γ_i and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$ with corresponding eigenvalues λ_i. We order the columns of Γ

according to the magnitude of the corresponding eigenvalues such that $\lambda_1 > \lambda_2 > \cdots > \lambda_N$. Let us assume here that all eigenvalues are distinct, otherwise Γ may not be identified. For the case of non-distinct eigenvalues, one may use the more general singular value decomposition and go for the GO-GARCH (generalized orthogonal GARCH) model of [114].

The vector of principal components, given by

$$\begin{pmatrix} f_t \\ \varepsilon_t \end{pmatrix} = \Gamma' y_t$$

is partitioned into the first K components f_t, whose volatility will assumed to be stochastic, and the last $N-K$ components ε_t, whose volatility will assumed to be constant. One could speak of K *dynamic* and $N-K$ *static* factors.

Now decompose the matrices as follows:

$$\Gamma_{(N\times N)} = (\Gamma_{1_{(N\times K)}}, \Gamma_{2_{(N\times(N-K))}}) ,$$

and $\Lambda_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$, $\Lambda_2 = \mathrm{diag}(\lambda_{K+1}, \ldots, \lambda_N)$.

The model can then be written as

$$y_t = \Gamma_1 f_t + \Gamma_2 \varepsilon_t \tag{26}$$

where $\mathrm{Var}(f_t \mid \mathcal{F}_{t-1}) = \Sigma_t$ and $\mathrm{Var}(\varepsilon_t \mid \mathcal{F}_{t-1}) = \Lambda_2$. For example, Σ_t may be diagonal or some K-variate GARCH model. Note that this representation is equivalent to that of (24) with $W = \Gamma_1$ and $v_t = \Gamma_2 \varepsilon_t$, except that $\Omega = \Gamma_2 \Lambda_2 \Gamma_2'$ will not be diagonal in general. The conditional variance of y_t is given by

$$\begin{aligned} H_t &= \mathrm{Var}(y_t \mid \mathcal{F}_{t-1}) = \Gamma_1 \Sigma_t \Gamma_1' + \Gamma_2 \Lambda_2 \Gamma_2' \\ &= \Gamma \begin{pmatrix} \Sigma_t & 0 \\ 0 & \Lambda_2 \end{pmatrix} \Gamma' . \end{aligned}$$

If Σ_t follows a K-variate BEKK process, then it can be shown that H_t will follow an N-variate BEKK process with restrictions on the parameter matrices. However, the classical OGARCH assumes that factors are conditionally orthogonal, hence Σ_t is diagonal, additional to the fact that they are unconditionally orthogonal by construction. This assumption is crucial and may not always be justified in practice. It should be emphasized that in the OGARCH model, the factor loadings contained in the matrix Γ and the factor variances contained in Λ are considered as fixed for a given sample covariance matrix. This contrasts the general factor model (24) where factor loadings W are estimated jointly with the parameters describing the factor dynamics.

Instead of using unconditionally orthogonal factors, [55] proposed to use conditionally orthogonal factors by searching numerically for linear combinations of the data such that the conditional correlation between these combinations is minimized under norm constraints. The existence of such linear combinations is tested using bootstrap methods.

Constant and Dynamic Conditional Correlation Models

[22] suggests a multivariate GARCH model with constant conditional correlations. Let H_t be the conditional covariance matrix of a series y_t, and V_t be a diagonal matrix with the conditional standard deviations of y_t on its diagonal. Then the model is simply

$$H_t = V_t R V_t \tag{27}$$

where R is the constant correlation matrix. H_t is positive definite as long as the conditional variances are positive and R is positive definite. For instance, one could specify univariate GARCH models for the individual conditional variances. One the other hand, it is possible to allow for spill-over of volatilities from one series to other series. Note that the CCC model is not nested in the Vec specification. Theory of maximum likelihood estimation for CCC-type models has been established by [77] for consistency and [86] for asymptotic normality.

The assumption of constant correlations simplifies strongly the estimation problem. However, it might sometimes be too restrictive. For example, it is often observed that correlations between financial time series increase in turbulent periods, and are very high in crash situations. A Lagrange Multiplier test against the CCC model has been suggested by [112]. An extension of the CCC model to allow for time-varying correlations is the dynamic conditional correlations (DCC) model introduced by [45]. The DCC model renders the conditional correlation matrix R dependent on time, R_t say. The conditional correlation between the ith and jth component of y_t is modeled as

$$R_{ij,t} = \frac{Q_{ij,t}}{\sqrt{Q_{ii,t} Q_{jj,t}}}$$

where $Q_{ij,t}$ is the ijth element of the matrix Q_t given by

$$Q_t = S(1-\alpha-\beta) + \alpha v_{t-1} v_{t-1}' + \beta Q_{t-1} \tag{28}$$

where α and β are parameters and $v_t = V_t^{-1} y_t$ are the standardized but correlated residuals. That is, the conditional variances of the components of v_t are one, but the conditional correlations are given by R_t. The matrix S is the sample correlation matrix of v_t, so a consistent esti-

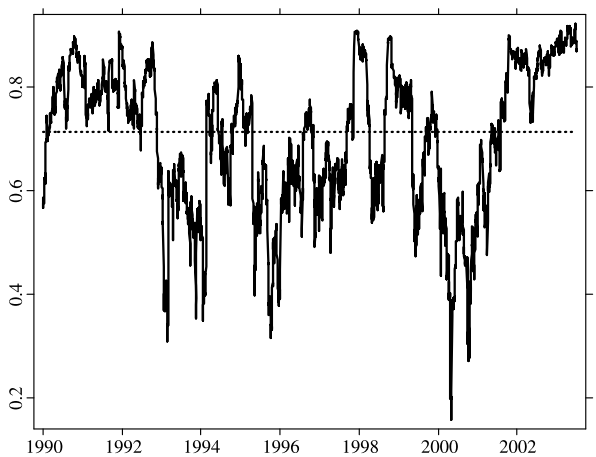

GARCH Modeling, Figure 10
Conditional correlations of the Dow Jones IA and NASDAQ index returns, daily, using the DCC model. *Dashed line*: constant conditional correlation

mate of the unconditional correlation matrix. If α and β are zero, we get the above CCC model. If they are different from zero one gets a kind of ARMA structure for all correlations. Note however that all correlations would follow the same kind of dynamics, since the ARMA parameters are the same for all correlations. The specification of the first term of Q_t ensures that the unconditional mean of Q_t is equal to the sample covariance matrix of v_t, similar to the *variance targeting* technique of [50]. Also it facilitates the estimation, since that can be done in two steps: First, the conditional variances in V_t are estimated using univariate GARCH models, for example, then v_t, the standardized (but correlated) residuals and their covariance matrix S are computed, before in the second step only two remaining parameters, α and β, need to be estimated. A model similar to DCC has been proposed by [113].

Figure 10 depicts the estimated conditional correlations for the DOW Jones and NASDAQ time series, using the DCC and CCC models. Comparing the former with the RiskMetrics estimates of Fig. 9, no substantial difference can be detected visually. However, the parameter estimates of α and β are 0.0322 and 0.9541 with standard errors 0.0064 and 0.0109, respectively, so that the null hypothesis $H_0\colon \alpha = 0.06$ is clearly rejected. Whether or not the difference in estimated conditional correlations matters in empirical applications has been addresses, e. g., by [30], who consider the problem of portfolio selection. [96] compare the performance of CCC, DCC, OGARCH and a model of [84] in forecasting and portfolio selection in high dimensions. They find that the difference is not substantial, but that the CCC model is too restrictive.

To summarize, the whole challenge of multivariate volatility modeling is to balance model complexity and simplicity in such a way that the model is flexible enough to capture all stylized facts in the second moments (and perhaps beyond that) of the series while keeping it simple for estimation and inference.

In the following we sketch some applications of multivariate GARCH models in finance. As an early example, [24] estimate a capital asset pricing model (CAPM) with time-varying betas. The beta is defined as the ratio of the asset return's covariance with the market return, divided by the variance of the market return. Denote by r_{it} the excess return of asset i, and by r_{mt} the excess return of the market. Then the beta-form of the CAPM can be written as

$$r_{it} = \beta_{it} r_{mt} + \varepsilon_i = \frac{\mathrm{Cov}(r_{it}, r_{mt})}{\mathrm{Var}(r_{mt})} + \varepsilon_i$$

where ε_i is idiosyncratic noise whose risk cannot be diversified away and is therefore called unsystematic risk. As we observe time varying second moments, it is clear that betas will also be time varying, not only due to the variance of the market but also due to the covariances of the assets with the market. However, if both returns are covariance stationary, then by definition the unconditional second moments will be constant, and only after conditioning on suitable information sets such as historical returns will second moments become time varying.

Secondly, correlations between exchange rates have been substantially time-varying, as for example in Europe the European exchange rate mechanism enforced increasing correlations. The correlation of the DEM/USD and FRF/USD rates, for instance, increased steadily in the late 1990s until it was virtually one just before the launch of the Euro. See, e. g., [45], who models these data, among others, with alternative correlation models. Thirdly, portfolio selection is another type of application. If, for example, one is interested in the minimum variance portfolio of n assets with covariance matrix Σ, then the well known formula for the optimal weight vector α is given by

$$\alpha = \frac{\Sigma^{-1}\iota}{\iota'\Sigma^{-1}\iota}$$

where ι is an n-vector of ones, see, e. g., [30]. Obviously, if Σ is allowed to be time-varying, then the optimal portfolio weights will in general also depend on time. This has many important practical implications, e. g., for portfolio managers. One of the problems is to determine an optimal reallocation frequency. If variances and covariances change daily and the objective is to minimize the portfo-

lio variance over the next ten days, then one could follow at least two strategies: either calculate the optimal portfolio weights daily and reallocate accordingly. Or, calculate the return distribution over ten days, obtain thus a covariance matrix for ten-day returns, find the optimal weights using this covariance matrix and leave the corresponding portfolio unchanged throughout the ten days. If the objective is to minimize the variance over the ten days, then the first method will usually outperform the second. The intuitive reason is that the second method aggregates data, thus losing valuable information. However, in practice one may still prefer the second method for various reasons, one of which could be the higher transaction costs of the first method.

Stochastic Volatility

GARCH models discussed so far explain the conditional variance at time t as a function of the information set at time $t-1$. In other words, it is measurable with respect to this information set. This is not the case for models of the stochastic volatility (SV) type, which introduce an extra error term in the volatility equation. For example, in the univariate case such a model could take the form

$$\begin{aligned} y_t &= \sigma_t \xi_t \\ \log \sigma_t^2 &= \omega + \beta \log \sigma_{t-1}^2 + \eta_t \end{aligned} \tag{29}$$

where ξ_t and η_t are i.i.d. mean zero random variables with variance equal to one and σ_η^2, respectively. Here, log volatility follows an AR(1) process. Since volatility is unobserved, model (29) is a particular case of a latent variable model. Note that, if the information set at time $t-1$ consists of all lagged values of y_t up to y_{t-1}, then volatility at time t is not measurable with respect to this information set. [27] compare moment properties such as kurtosis and persistence of SV and GARCH models. [58] propose a model that encompasses both GARCH and stochastic volatility and thus allows for testing against each of them.

Estimation is more complicated than for GARCH models because the likelihood is an integral of dimension equal to the sample size, given by

$$L(Y;\theta) = \int p(Y \mid H, \theta) p(H \mid \theta) \mathrm{d}H \tag{30}$$

where $Y = (y_1, \ldots, y_n)$, $H = (\sigma_1^2, \ldots, \sigma_n^2)$, and $\theta = (\omega, \beta, \sigma_\eta^2)$. Maximization of (30) has no closed form and numerical optimization is difficult due to the high dimensional integral. Therefore, other estimation methods have been considered in the literature, for example generalized method of moments (GMM), simulated maximum likelihood with Markov Chain Monte Carlo (MCMC) or Bayesian methods, see e. g., [75,76] and [80]. An application to currency options by [92] compares three alternative estimation algorithms and finds that the estimation error of the volatility series is large for all methods.

In the multivariate case, without imposing structure, estimating a highly dimensional stochastic volatility model seems difficult. One way of imposing structure in multivariate SV models is to assume a factor model as, e. g., in [34,69,81] and [56], or constant correlations. To consider a bivariate extension of stochastic volatility models, one suggestion of [69] is to say that the stochastic variances $\sigma_{1,t}$ and $\sigma_{2,t}$ of the two assets follow univariate stochastic variance processes as in (29), and the stochastic covariance is given by

$$\sigma_{12,t} = \rho \sigma_{1,t} \sigma_{2,t} \ ,$$

where ρ is a constant parameter between -1 and 1. This model, very much in the spirit of the constant conditional correlation GARCH model of [22], is quite parsimonious and can be efficiently estimated using simulated maximum likelihood as demonstrated in [33]. It is straightforward to generalize this specification to higher dimensions. However, estimation may then become trickier. Also the restriction of constant correlation parameters may not be innocuous. More empirical tests are required about goodness of fits, comparing the non-nested GARCH and SV type models of about the same model complexity.

SV models lend themselves naturally to continuous time stochastic volatility models and realized volatility. Indeed, as shown by [9], realized volatility can be used to estimate the volatility of SV models. The monograph of [109] collects influential papers of the stochastic volatility literature.

Aggregation

The frequency at which financial time series are sampled is often not unique. For example, one researcher may be interested in the behavior of returns to the Dow Jones index at a daily frequency, but another one at a weekly or monthly frequency. Considering log-returns, weekly returns can be directly obtained from daily returns by simply summing up intra-week returns. If a model is fitted to daily returns, an important question is what this implies for the weekly returns. In particular, one may ask if the model remains in the same class, which would then be called closed under temporal aggregation. For the univariate GARCH model, [38] have shown that only a weak version of it is closed under temporal aggregation. Instead

of modeling the conditional variance, weak GARCH models the best linear predictor of squared returns in terms of a constant, lagged returns and lagged squared returns. In the weak GARCH(1,1) case, they show how to obtain the parameters of the aggregated process as a function of the parameters of the high frequency process. In particular, denoting the parameters of the aggregated process by $\alpha^{(m)}$ and $\beta^{(m)}$, where m is the aggregation level, then the persistence parameter of the aggregated level is given by $\alpha^{(m)} + \beta^{(m)} = (\alpha + \beta)^m$. Thus, the persistence of the aggregated process declines geometrically fast with the aggregation level. Asymptotically, the process will reduce to white noise. One would therefore expect to see much less conditional heteroskedasticity in monthly returns than in weekly or daily returns. The link between parameters at different frequencies also provides a means for model diagnostics. The results of [38] have been extended to the multivariate case by [64].

Instead of aggregating, one could go the other way and look at "disggregating" the process temporally, i. e., sampling the underlying process at finer intervals. [97] showed that GARCH models can be viewed as approximations of continuous time stochastic volatility models, see also [39]. However, [117] has shown that the GARCH model and its diffusion limit are not equivalent in a statistical experiment sense.

Rather than aggregating temporally, one may alternatively be interested in aggregating contemporaneously in a multivariate context. For example, stock indices are constructed as linear combinations of individual stocks. [102] show that again the aggregated process is only weak GARCH. Rather than aggregating multivariate GARCH models, one can alternatively consider aggregation of univariate heterogenous GARCH processes with random coefficients. In linear ARMA models, this aggregation scheme is known to produce long memory type behavior of the aggregate, see [62]. [35] conjectured that this holds in a similar way for GARCH models. However, [122] shows that although the ACF of the squared aggregate decays hyperbolically, it may be absolutely summable and hence there is no long memory. For the general model class of [94] which includes GARCH, weak GARCH and stochastic volatility as special cases, [121] shows that contemporaneous aggregation leads to long memory properties of the aggregate.

Future Directions

The theory of univariate GARCH models is now well developed and understood. For example, theory of maximum likelihood estimation is available under weak conditions that allow for integrated and even mildly explosive processes. However, theory of multivariate GARCH is still in its infancy and far from closed, due to arising technical difficulties. For general specification such as the BEKK model, no results on asymptotic normality of estimates are available yet that would allow for integrated processes. Most available results on general specifications are high level and only for some special cases, primitive conditions are established. This is certainly one of the main directions for future research.

On the modeling side, there is no clear general answer how to deal with the problem of high dimensions, and in particular how to balance model flexibility with econometric feasibility. More practical experience is necessary to see what type of model performs best for what kind of data. On the application side, a still open issue is how to evaluate the volatility risk for option pricing, and how to efficiently use multivariate GARCH models in portfolio selection or risk management. Other frontiers for GARCH models are discussed by [46].

An interesting new field is the combination of GARCH models with nonparametric distributions to obtain more accurate estimates of the Value-at-Risk, mentioned in Sect. "Introduction". In the univariate case this is quite obvious, but in the multivariate case one has to deal with the "curse of dimensionality", common in the nonparametrics literature. Furthermore, issues such as tail dependence need to be modeled accurately in that case. A joint framework that captures volatilities, correlations, other distributional shape features and tail dependence would be an interesting target for applied research.

Finally, *realized volatilities* (RV) have been mentioned at the end of Sect. "Properties of the GARCH(1,1) Model" as a means to use intra-day data to generate accurate ex post measures of daily volatilities. Using these RV measures, one can build time series models that predict daily volatilities one or more steps ahead, see e. g., [5] for a detailed analysis. It seems that RV provides better forecasts than GARCH, which is not surprising as it uses more information, namely the intra-day returns. The RV literature has evolved as an important second branch of volatility modeling next to the discrete time GARCH or SV models. One direction of research is the treatment of microstructure noise, present in most high frequency data, as e. g., in [1,2] and [67]. Another one is the modeling of jumps using the so-called bi-power variation and the generalization to the multivariate case using realized covariances and bipower co-variation, see e. g., [10] and [11]. Other directions are possible and it seems likely that RV will become the dominant econometric tool to model volatilities provided that high frequency data are available.

Bibliography

Primary Literature

1. Ait-Sahalia Y, Mykland P, Zhang L (2005) A tale of two time scales: Determining integrated volatility with noisy high-frequency data. J Am Stat Assoc 100:1394–1411
2. Ait-Sahalia Y, Mykland P, Zhang L (2005) How often to sample a continuous time process in the presence of market microstructure noise. Rev Financial Stud 18:351–416
3. Alexander C (2001) Orthogonal GARCH. In: Mastering Risk, Financial Times, vol 2. Prentice Hall, London, pp 21–38
4. Andersen TG, Bollerslev T (1998) Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. Int Econ Rev 39:885–905
5. Andersen TG, Bollerslev T, Diebold FX, Labys P (2003) Modeling and forecasting realized volatility. Econometrica 71:579–625
6. Andrews DWK (1999) Estimation when a parameter is on a boundary. Econometrica 67:1341–1383
7. Baillie RT (1996) Long memory processes and fractional integration in econometrics. J Econom 73:5–59
8. Baillie RT, Bollerslev T, Mikkelsen HO (1996) Fractionally integrated generalized autoregressive conditional heteroskedasticity. J Econom 74:3–30
9. Barndorff-Nielsen O, Shephard N (2002) Econometric analysis of realized volatility and its use in estimating stochastic volatility models. J Roy Stat Soc 64:253–280
10. Barndorff-Nielsen O, Shephard N (2004) Econometric analysis of realized covariation: high frequency covariance, regression and correlation in financial economics. Econometrica 72:885–925
11. Barndorff-Nielsen O, Shephard N (2004) Power and bipower variation with stochastic volatility and jumps (with discussion). J Financial Econom 2:1–48
12. Bauwens L, Lubrano M (1998) Bayesian inference on garch models using the gibbs sampler. Econom J 1:C23–C46
13. Bauwens L, Laurent S, Rombouts J (2006) Multivariate garch models: a survey. J Appl Econom 21:79–109
14. Bera A, Higgins M (1993) A survey of ARCH models: properties, estimation and testing. J Econ Surv 7:305–366
15. Berndt EK, Hall BH, Hall RE, Hausman JA (1974) Estimation and inference in nonlinear structural models. Ann Econ Soc Meas 3:653–665
16. Bickel PJ (1982) On adaptive estimation. Ann Stat 10:647–671
17. Black F (1976) Studies in stock price volatility changes. In: Proceedings of the 1976 Meeting of the Business and Economic Statistics Section, American Statistical Association, pp 177–181
18. Black F, Scholes M (1973) The pricing of options and corporate liabilities. J Political Econ 81:637–654
19. Bollerslev T, Mikkelsen HO (1996) Modeling and pricing long-memory in stock market volatility. J Econom 73:151–184
20. Bollerslev T, Engle R, Nelson D (1994) ARCH models. In: Engle R, McFadden D (eds) Handbook of Econometrics. North Holland, Amsterdam, pp 2959–3038
21. Bollerslev TP (1986) Generalized autoregressive conditional heteroscedasticity. J Econom 31:307–327
22. Bollerslev TP (1990) Modelling the coherence in short-run nominal exchange rates: A multivariate generalized arch model. Rev Econ Stat 72:498–505
23. Bollerslev TP, Wooldridge JM (1992) Quasi maximum likelihood estimation of dynamic models with time-varying covariances. Econom Rev 11:143–172
24. Bollerslev TP, Engle RF, Wooldridge JM (1988) A capital asset pricing model with time-varying covariances. J Political Econ 96:116–131
25. Bougerol P, Picard N (1992) Stationarity of garch processes and some nonnegative time series. J Econom 52:115–127
26. Bühlmann P, McNeil AJ (2002) An algorithm for nonparametric GARCH modelling. Comput Stat Data Anal 40:665–683
27. Carnero MA, Pena D, Ruiz E (2004) Persistence and kurtosis in GARCH and stochastic volatility models. J Financial Econom 2:319–342
28. Carrasco M, Chen X (2002) Mixing and moment properties of various garch and stochastic volatility models. Econom Theory 18:17–39
29. Carroll R, Härdle W, Mammen E (2002) Estimation in an additive model when the components are linked parametrically. Econom Theory 18:886–912
30. Chan LKC, Karceski J, Lakonishok J (1999) On portfolio optimization: Forecasting covariances and choosing the risk model. Rev Financial Stud 12:937–674
31. Comte F, Lieberman O (2003) Asymptotic theory for multivariate garch processes. J Multivar Anal 84:61–84
32. Concepcion Ausian M, Galeano P (2007) Bayesian estimation of the gaussian mixture GARCH model. Comput Stat Data Anal 51:2636–2652
33. Danielsson J (1998) Multivariate stochastic volatility models: Estimation and a comparison with vgarch models. J Empir Finance 5:155–174
34. Diebold FX, Nerlove M (1989) The dynamics of exchange rate volatility: A multivariate latent factor arch model. J Appl Econom 4:1–21
35. Ding Z, Granger C (1996) Modelling volatility persistence of speculative returns: a new approach. J Econom 73:185–215
36. Ding Z, Granger CWJ, Engle RF (1993) A long memory property of stock market returns and a new model. J Empir Finance 1:83–106
37. Drost FC, Klaassen CAJ (1997) Efficient estimation in semiparametric garch models. J Econom 81:193–221
38. Drost FC, Nijman T (1993) Temporal aggregation of GARCH processes. Econometrica 61:909–927
39. Drost FC, Werker BJM (1996) Closing the garch gap: Continuous garch modeling. J Econom 74:31–57
40. Drost FC, Klaassen CAJ, Werker BJM (1997) Adaptive estimation in time series models. Ann Stat 25:786–817
41. Duan JC (1995) The GARCH option pricing model. Math Finance 5:13–32
42. Duan JC (1997) Augmented garch(p, q) process and its diffusion limit. J Econom 79:97–127
43. Embrechts P, Klüppelberg C, Mikosch T (1997) Modelling Extremal Events. Springer, Berlin
44. Engle RF (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. Econometrica 50:987–1008
45. Engle RF (2002) Dynamic conditional correlation – a simple class of multivariate garch models. J Bus Econ Stat 20: 339–350
46. Engle RF (2002) New frontiers of ARCH models. J Appl Econom 17:425–446

47. Engle RF, Bollerslev TP (1986) Modelling the persistence of conditional variances. Econom Rev 5:1–50, 81–87
48. Engle RF, Gonzalez-Rivera G (1991) Semiparametric ARCH models. J Bus Econ Stat 9:345–360
49. Engle RF, Kroner KF (1995) Multivariate simultaneous generalized arch. Econom Theory 11:122–150
50. Engle RF, Mezrich J (1996) GARCH for groups. RISK 9:36–40
51. Engle RF, Ng VK (1993) Measuring and testing the impact of news on volatility. J Finance 48:1749–1778
52. Engle RF, Lilien DM, Robins RP (1987) Estimating time varying risk premia in the term structure: The ARCH-M model. Econometrica 55:391–407
53. Engle RF, Ng VK, Rothschild M (1990) Asset pricing with a factor-ARCH covariance structure. J Econom 45:213–237
54. Fama EF (1965) The behavior of stock market prices. J Bus 38:34–105
55. Fan J, Wang M, Yao Q (2008) Modelling multivariate volatilities via conditionally uncorrelated components. J Royal Stat Soc Ser B 70:679–702
56. Fiorentini G, Sentana E, Shephard N (2004) Likelihood-based estimation of latent generalized ARCH structures. Econometrica 72:1481–1517
57. Francq C, Zakoian JM (2004) Maximum likelihood estimation of pure garch and arma-garch processes. Bernoulli 10: 605–637
58. Fridman M, Harris L (1998) A maximum likelihood approach for non-gaussian stochastic volatility models. J Bus Econ Stat 16:284–291
59. Glosten LR, Jagannathan R, Runkle DE (1993) On the relation between the expected value and the volatility of the nominal excess return on stocks. J Finance 48:1779–1801
60. Gouriéroux C (1992) Modèles ARCH et Applications Financières. Economica
61. Gouriéroux C, Monfort A (1992) Qualitative threshold ARCH models. J Econom 52:159–199
62. Granger CWJ (1980) Long memory relationships and the aggregation of dynamic models. J Econom 14:227–238
63. Hafner CM (1998) Estimating high frequency foreign exchange rate volatility with nonparametric ARCH models. J Stat Plan Inference 68:247–269
64. Hafner CM (2008) Temporal aggregation of multivariate GARCH processes. J Econom 142:467–483
65. Hafner CM, Rombouts JVK (2007) Semiparametric multivariate volatility models. Econom Theory 23:251–280
66. Hall P, Yao P (2003) Inference in ARCH and GARCH models with heavy-tailed errors. Econometrica 71:285–317
67. Hansen PR, Lunde A (2006) Realized variance and market microstructure noise, with comments and rejoinder. J Bus Econ Stat 24:127–218
68. Härdle W, Tsybakov A (1997) Local polynomial estimation of the volatility function. J Econom 81:223–242
69. Harvey AC, Ruiz E, Shepard N (1994) Multivariate stochastic variance models. Rev Econ Stud 61:247–264
70. He C, Teräsvirta T (1999) Statistical Properties of the Asymmetric Power ARCH Process. In: Engle RF, White H (eds) Cointegration, Causality, and Forecasting. Festschrift in honour of Clive W.J. Granger, Oxford University Press, pp 462–474
71. Hentschel L (1995) All in the family: Nesting symmetric and asymmetric garch models. J Financial Econ 39:71104
72. Hill B (1975) A simple general approach to inference about the tail of a distribution. Ann Stat 3:1163–1174
73. Hong Y, Tu J, Zhou G (2007) Asymmetries in stock returns: Statistical tests and economic evaluation. Rev Financial Stud 20:1547–1581
74. Hull J, White A (1987) The pricing of options on assets with stochastic volatilities. J Finance 42:281–300
75. Jacquier E, Polson NG, Rossi PE (1994) Bayesian analysis of stochastic volatility models (with discussion). J Bus Econ Stat 12:371–417
76. Jacquier E, Polson NG, Rossi PE (2004) Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. J Econom 122:185–212
77. Jeantheau T (1998) Strong consistency of estimators for multivariate arch models. Econom Theory 14:70–86
78. Jorion P (2000) Value-at-Risk: The New Benchmark for Managing Financial Risk. McGraw-Hill, New York
79. Kawakatsu H (2006) Matrix exponential GARCH. J Econom 134:95–128
80. Kim S, Shephard N, Chib S (1998) Stocahstic volatility: likelihood inference and comparison with ARCH models. Rev Econ Stud 65:361–393
81. King M, Sentana E, Wadhwani S (1994) Volatility and links between national stock markets. Econometrica 62:901–933
82. Kristensen D, Linton O (2006) A closed-form estimator for the garch(1,1) model. Econom Theory 323–337
83. Lanne M, Saikkonen P (2007) A multivariate generalized orthogonal factor GARCH model. J Bus Econ Stat 25:61–75
84. Ledoit O, Santa-Clara P, Wolf M (2003) Flexible multivariate GARCH modeling with an application to international stock markets. Rev Econs Stat 85:735–747
85. Lee SW, Hansen BE (1994) Asymptotic properties of the maximum likelihood estimator and test of the stability of parameters of the GARCH and IGARCH models. Econom Theory 10:29–52
86. Ling S, McAleer M (2003) Asymptotic theory for a vector ARMA-GARCH model. Econom Theory 19:280–310
87. Lintner J (1965) Security prices, risk and maximal gains from diversification. J Finance 20:587–615
88. Linton O (1993) Adaptive estimation in ARCH models. Econom Theory 9:539–569
89. Linton O, Mammen E (2005) Estimating semiparametric ARCH models by kernel smoothing methods. Econometrica 73:771–836
90. Lo A, Wang J (1995) Implementing option pricing models when asset returns are predictable. J Finance 50:87–129
91. Lumsdaine RL (1996) Asymptotic properties of the quasi maximum likelihood estimator in GARCH(1,1) and IGARCH(1,1) models. Econometrica 64:575–596
92. Mahieu R, Schotman P (1998) An empirical application of stochastic volatility models. J Appl Econom 13:333–360
93. Mandelbrot B (1963) The variation of certain speculative prices. J Bus 36:394–419
94. Meddahi N, Renault E (2004) Temporal aggregation of volatility models. J of Econom 119:355–379
95. Morgan JP (1996) Riskmetrics Technical Document, 4th edn. J.P. Morgan, New York
96. Morillo D, Pohlman L (2002) Large scale multivariate GARCH risk modelling for long-horizon international equity portfolios. Proceedings of the 2002 Forecasting Financial Markets conference, London
97. Nelson DB (1990) ARCH models as diffusion approximations. J Econom 45:7–38

98. Nelson DB (1990) Stationarity and persistence in the GARCH(1,1) model. Econom Theory 6:318–334
99. Nelson DB (1991) Conditional heteroskedasticity in asset returns: A new approach. Econometrica 59:347–370
100. Nelson DB, Cao CQ (1992) Inequality constraints in the univariate garch model. J Bus Econ Stat 10:229–235
101. Newey WK, Steigerwald DS (1997) Asymptotic bias for quasi maximum likelihood estimators in conditional heteroskedasticity models. Econometrica 3:587–599
102. Nijman T, Sentana E (1996) Marginalization and contemporaneous aggregation in multivariate GARCH processes. J Econom 71:71–87
103. Pantula SG (1988) Estimation of autoregressive models with ARCH errors. Sankhya Indian J Stat B 50:119–138
104. Peng L, Yao Q (2003) Least absolute deviations estimation for ARCH and GARCH models. Biometrika 90:967–975
105. Robinson PM (1991) Testing for strong serial correlation and dynamic conditional heteroskedasticity in multiple regression. J Econom 47:67–84
106. Ross SA (1976) The arbitrage theory of capital asset pricing. J Econ Theory 13:341–360
107. Sentana E, Fiorentini G (2001) Identification, estimation and testing of conditionally heteroskedastic factor models. J Econom 102:143–164
108. Sharpe WF (1964) Capital asset prices: A theory of market equilibrium under conditions of risk. J Finance 19:425–442
109. Shephard N (2005) Stochastic Volatility: Selected Readings. Oxford University Press, Oxford
110. Straumann D, Mikosch T (2006) Quasi-mle in heteroscedastic times series: a stochastic recurrence equations approach. Ann Stat 34:2449–2495
111. Taylor SJ (1986) Modelling Financial Time Series. Wiley, New York
112. Tse YK (2000) A test for constant correlations in a multivariate GARCH model. J Econom 98:107–127
113. Tse YK, Tsui AKC (2002) A multivariate GARCH model with time-varying correlations. J Bus Econ Stat 20:351–362
114. van der Weide R (2002) Go-garch: A multivariate generalized orthogonal GARCH model. J Appl Econom 17:549–564
115. Vrontos ID, Dellaportas P, Politis DN (2000) Full bayesian inference for GARCH and EGARCH models. J Bus Econ Stat 18:187198
116. Vrontos ID, Dellaportas P, Politis D (2003) A full-factor multivariate garch model. Econom J 6:311–333
117. Wang Y (2002) Asymptotic nonequivalence of GARCH models and diffusions. Ann Stat 30:754–783
118. Weiss AA (1986) Asymptotic theory for ARCH models: Estimation and testing. Econom Theory 2:107–131
119. Yang L (2006) A semiparametric GARCH model for foreign exchange volatility. J Econom 130:365–384
120. Yang L, Härdle W, Nielsen P (1999) Nonparametric autoregression with multiplicative volatility and additive mean. J Time Ser Anal 20:579–604
121. Zaffaroni P (2007) Aggregation and memory of models of changing volatility. J Econom 136:237–249
122. Zaffaroni P (2007) Contemporaneous aggregation of GARCH processes. J Time Series Anal 28:521–544
123. Zakoian JM (1994) Threshold heteroskedastic functions. J Econ Dyn Control 18:931–955

Books and Reviews

Andersen T, Bollerslev T, Diebold F (2004) Parametric and nonparametric measurement of volatility. In: Ait-Sahalia L, Hansen LP (eds) Handbook of Financial Economtrics. Amsterdam (forthcoming)
Bauwens L, Laurent S, Rombouts J (2006) Multivariate GARCH models: A survey. J Appl Econom 21:79–109
Bera A, Higgins M (1993) A survey of ARCH models: properties, estimation and testing. J Econ Surv 7:305–366
Bollerslev T, Chou R, Kroner K (1992) ARCH modelling in finance: a review of the theory and empirical evidence. J Econom 52:5–59
Bollerslev T, Engle R, Nelson D (1994) ARCH models. In: Engle R, McFadden D (eds) Handbook of Econometrics. North Holland Press, Amsterdam, pp 2959–3038
Engle R (1995) ARCH: Selected Readings. Oxford University Press, Oxford
Gouriéroux C (1997) ARCH Models and Financial Applications. Springer, New York
Shephard N (1996) Statistical aspects of ARCH and stochastic volatility. In: Cox DR, Hinkley DV, Barndorff-Nielsen OE (eds) Time Series Models in Econometrics, Finance and Other Fields. Chapman & Hall, London, pp 1–67
Shephard N (2005) Stochastic Volatility: Selected Readings. Oxford University Press, Oxford
Taylor S (1986) Modelling Financial Time Series. Wiley, Chichester

Genetic and Evolutionary Algorithms and Programming: General Introduction and Application to Game Playing

Michael Orlov, Moshe Sipper, Ami Hauptman
Department of Computer Science,
Ben-Gurion University, Beer-Sheva, Israel

Article Outline

Glossary
Definition of the Subject
Introduction
Evolutionary Algorithms
A Touch of Theory
Extensions of the Basic Methodology
Lethal Applications
Evolutionary Games
Future Directions
Bibliography

Glossary

Evolutionary algorithms/evolutionary computation
A family of algorithms inspired by the workings of

evolution by natural selection whose basic structure is to:

1. produce an initial **population** of individuals, these latter being candidate solutions to the problem at hand
2. evaluate the **fitness** of each individual in accordance with the problem whose solution is sought
3. *while* termination condition not met *do*
   (a) **select** fitter individuals for reproduction
   (b) **recombine (crossover)** individuals
   (c) **mutate** individuals
   (d) **evaluate** fitness of modified individuals
   end while

Genome/chromosome An individual's makeup in the population of an evolutionary algorithm is known as a genome, or chromosome. It can take on many forms, including bit strings, real-valued vectors, character-based encodings, and computer programs. The representation issue – namely, defining an individual's genome (well) – is critical to the success of an evolutionary algorithm.

Fitness A measure of the quality of a candidate solution in the population. Also known as *fitness function*. Defining this function well is critical to the success of an evolutionary algorithm.

Selection The operator by which an evolutionary algorithm selects (usually probabilistically) higher-fitness individuals to contribute *genetic* material to the next generation.

Crossover One of the two main genetic operators applied by an evolutionary algorithm, wherein two (or more) candidate solutions (*parents*) are combined in some pre-defined manner to form *offspring*.

Mutation One of the two main genetic operators applied by an evolutionary algorithm, wherein one candidate solution is randomly altered.

Definition of the Subject

Evolutionary algorithms are a family of search algorithms inspired by the process of (Darwinian) evolution in nature. Common to all the different family members is the notion of solving problems by evolving an initially random population of candidate solutions, through the application of operators inspired by natural genetics and natural selection, such that in time *fitter* (i. e., better) solutions emerge. The field, whose origins can be traced back to the 1950s and 1960s, has come into its own over the past two decades, proving successful in solving multitudinous problems from highly diverse domains including (to mention but a few): optimization, automatic programming, electronic-circuit design, telecommunications, networks, finance, economics, image analysis, signal processing, music, and art.

Introduction

The first approach to artificial intelligence, the field which encompasses evolutionary computation, is arguably due to Turing [31]. Turing asked the famous question: "Can machines think?" Evolutionary computation, as a subfield of AI, may be the most straightforward answer to such a question. In principle, it might be possible to evolve an algorithm possessing the functionality of the human brain (this has already happened at least once: in nature).

In a sense, nature is greatly inventive. One often wonders how so many magnificent solutions to the problem of existence came to be. From the intricate mechanisms of cellular biology, to the sandy camouflage of flatfish; from the social behavior of ants to the diving speed of the peregrine falcon – nature created versatile solutions, at varying levels, to the problem of survival. Many ingenious solutions were *invented* (and still are), without any obvious intelligence directly creating them. This is perhaps the main motivation behind evolutionary algorithms: creating the settings for a dynamic environment, in which solutions can be created and improved in the course of time, advancing in new directions, with minimal direct intervention. The gain to problem solving is obvious.

Evolutionary Algorithms

In the 1950s and the 1960s several researchers independently studied evolutionary systems with the idea that evolution could be used as an optimization tool for engineering problems. Central to all the different methodologies is the notion of solving problems by evolving an initially random population of candidate solutions, through the application of operators inspired by natural genetics and natural selection, such that in time *fitter* (i. e., better) solutions emerge [9,16,19,28]. This thriving field goes by the name of *evolutionary algorithms* or *evolutionary computation*, and today it encompasses two main branches – genetic algorithms [9] and genetic programming [19] – in addition to less prominent (though important) offshoots, such as evolutionary programming [10] and evolution strategies [26].

A *genetic algorithm* (GA) is an iterative procedure that consists of a *population* of individuals, each one represented by a finite string of symbols, known as the *genome*, encoding a possible solution in a given problem space.

This space, referred to as the *search space*, comprises all possible solutions to the problem at hand. Generally speaking, the genetic algorithm is applied to spaces which are too large to be exhaustively searched. The symbol alphabet used is often binary, but may also be character-based, real-valued, or any other representation most suitable to the problem at hand.

The standard genetic algorithm proceeds as follows: an initial population of individuals is generated at random or heuristically. Every evolutionary step, known as a *generation*, the individuals in the current population are *decoded* and *evaluated* according to some predefined quality criterion, referred to as the *fitness*, or *fitness function*. To form a new population (the next generation), individuals are *selected* according to their fitness. Many selection procedures are available, one of the simplest being *fitness-proportionate selection*, where individuals are selected with a probability proportional to their relative fitness. This ensures that the expected number of times an individual is chosen is approximately proportional to its relative performance in the population. Thus, high-fitness (*good*) individuals stand a better chance of *reproducing*, while low-fitness ones are more likely to disappear.

Selection alone cannot introduce any new individuals into the population, i. e., it cannot find new points in the search space; these are generated by genetically-inspired operators, of which the most well known are *crossover* and *mutation*. Crossover is performed with probability p_{cross} (the *crossover probability* or *crossover rate*) between two selected individuals, called *parents*, by exchanging parts of their genomes (i. e., encodings) to form one or two new individuals, called *offspring*. In its simplest form, substrings are exchanged after a randomly-selected crossover point. This operator tends to enable the evolutionary process to move toward *promising* regions of the search space. The mutation operator is introduced to prevent premature convergence to local optima by randomly sampling new points in the search space. It is carried out by flipping bits at random, with some (small) probability p_{mut}. Genetic algorithms are stochastic iterative processes that are not guaranteed to converge. The termination condition may be specified as some fixed, maximal number of generations or as the attainment of an acceptable fitness level. Figure 1 presents the standard genetic algorithm in pseudo-code format.

Let us consider the following simple example, demonstrating the GA's workings. The population consists of four individuals, which are binary-encoded strings (genomes) of length 10. The fitness value equals the number of ones in the bit string, with $p_{\text{cross}} = 0.7$ and $p_{\text{mut}} = 0.05$. More typical values of the population size and the genome length are in the range 50–1000. Note that fitness computation in this case is extremely simple, since no complex decoding or evaluation is necessary. The initial (randomly generated) population might look as shown in Table 1.

```
begin GA
    g:=0    { generation counter }
    Initialize population P(g)
    Evaluate population P(g)    { i.e., compute fitness values }
    while not done do
        g:=g+1
        Select P(g) from P(g − 1)
        Crossover P(g)
        Mutate P(g)
        Evaluate P(g)
    end while
end GA
```

Genetic and Evolutionary Algorithms and Programming: General Introduction and Appl. to Game Playing, Figure 1
Pseudo-code of the standard genetic algorithm

Genetic and Evolutionary Algorithms and Programming: General Introduction and Appl. to Game Playing, Table 1
The initial population

Label	Genome	Fitness
p_1	0000011011	4
p_2	1110111101	8
p_3	0010000010	2
p_4	0011010000	3

Using fitness-proportionate selection we must choose four individuals (two sets of parents), with probabilities proportional to their relative fitness values. In our example, suppose that the two parent pairs are $\{p_2, p_4\}$ and $\{p_1, p_2\}$ (note that individual p_3 did not get selected as our procedure is probabilistic). Once a pair of parents is selected, crossover is effected between them with probability p_{cross}, resulting in two offspring. If no crossover is effected (with probability $1 - p_{\text{cross}}$), then the offspring are exact copies of each parent. Suppose, in our example, that crossover takes place between parents p_2 and p_4 at the (randomly chosen) third bit position:

111|0111101

001|1010000

This results in offspring $p'_1 = 1111010000$ and $p'_2 = 0010111101$. Suppose no crossover is performed between parents p_1 and p_2, forming offspring that are exact copies of p_1 and p_2. Our interim population (after crossover) is thus as depicted in Table 2:

Genetic and Evolutionary Algorithms and Programming: General Introduction and Appl. to Game Playing, Table 2
The interim population

Label	Genome	Fitness
p'_1	1111010000	5
p'_2	0010111101	6
p'_3	0000011011	4
p'_4	1110111101	8

Genetic and Evolutionary Algorithms and Programming: General Introduction and Appl. to Game Playing, Table 3
The resulting population

Label	Genome	Fitness
p''_1	1111010000	5
p''_2	0010101101	5
p''_3	0000011011	4
p''_4	1110111111	9

Next, each of these four individuals is subject to mutation with probability p_{mut} per bit. For example, suppose offspring p'_2 is mutated at the sixth position and offspring p'_4 is mutated at the ninth bit position. Table 3 describes the resulting population.

The resulting population is that of the next generation (i. e., p''_i equals p_i of the next generation). As can be seen, the transition from one generation to the next is through application of selection, crossover, and mutation. Moreover, note that the best individual's fitness has gone up from eight to nine, and that the average fitness (computed over all individuals in the population) has gone up from 4.25 to 5.75. Iterating this procedure, the GA will eventually find a perfect string, i. e., with maximal fitness value of ten.

Another prominent branch of the evolutionary computation tree is that of *genetic programming*, introduced by Cramer [7], and transformed into a field in its own right in large part due to the efforts of Koza [19]. Basically, genetic programming (GP) is a GA (genetic algorithm) with individuals in the population being programs instead of bit strings.

In GP we evolve a population of individual LISP expressions[1], each comprising *functions* and *terminals*. The functions are usually arithmetic and logic operators that receive a number of arguments as input and compute a result as output; the terminals are zero-argument functions that serve both as constants and as sensors, the latter being a special type of function that queries the domain environment.

The main mechanism behind GP is precisely that of a GA, namely, the repeated cycling through four operations applied to the entire population: evaluate-select-crossover-mutate. However, the evaluation of a single individual in GP is usually more complex than with a GA since it involves running a program. Moreover, crossover and mutation need to be made to work on trees (rather than simple bit strings), as shown in Fig. 2.

A Touch of Theory

Evolutionary computation is mostly an experimental field. However, over the years there have been some notable theoretical treatments of the field, gaining valuable insights into the properties of evolving populations.

Holland [17] introduced the notion of *schemata*, which are abstract properties of binary-encoded individuals, and analyzed the growth of different schemas when fitness-proportionate selection, point mutation and one-point crossover are employed. Holland's approach has since been enhanced and more rigorous analysis performed; however, there were not many practical consequences on the existing evolutionary techniques, since most of the successful methods are usually much more complex in many aspects. Moreover, the schematic analysis suffers from an important approximation of infinite population size, while in reality schemata can vanish.

Note that the *No Free Lunch* theorem states that "…for any [optimization] algorithm, any elevated performance over one class of problems is exactly paid for in performance over another class" [32].

Extensions of the Basic Methodology

We have reviewed the basic evolutionary computation methods. More advanced techniques are used to tackle complex problems, where an approach of a single population with homogeneous individuals does not suffice. One such advanced approach is coevolution [24].

Coevolution refers to the simultaneous evolution of two or more species with coupled fitness. Such coupled evolution favors the discovery of complex solutions whenever complex solutions are required. Simplistically speaking, one can say that coevolving species can either compete (e. g., to obtain exclusivity on a limited resource) or cooperate (e. g., to gain access to some hard-to-attain resource). In a competitive coevolutionary algorithm the fitness of an individual is based on direct competition with individuals of other species, which in turn evolve separately in their own populations. Increased fitness of one of the species

[1] Languages other than LISP have been used, although LISP is still by far the most popular within the genetic programming domain.

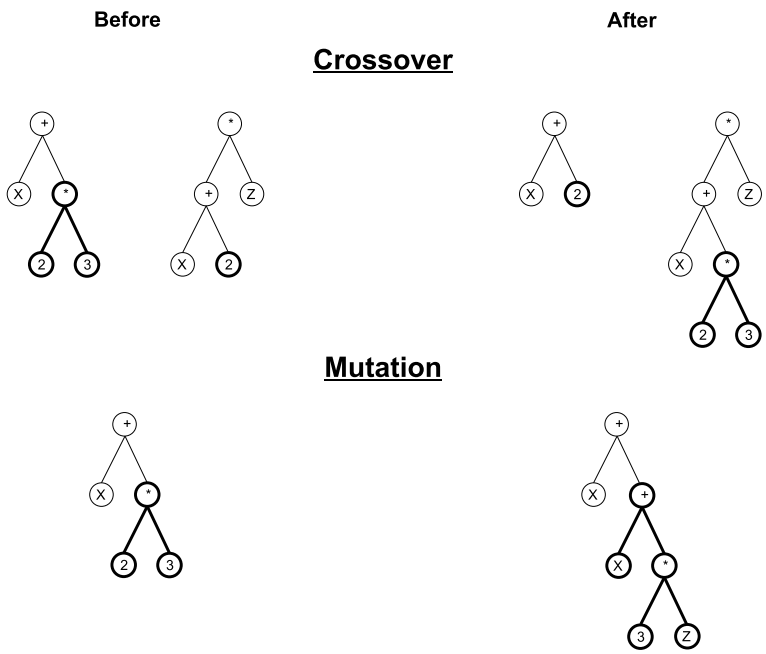

Genetic and Evolutionary Algorithms and Programming: General Introduction and Appl. to Game Playing, Figure 2
Genetic operators in genetic programming. LISP programs are depicted as trees. Crossover (*top*): Two sub-trees (marked in *bold*) are selected from the parents and swapped. Mutation (*bottom*): A sub-tree (marked in *bold*) is selected from the parent individual and removed. A new sub-tree is grown instead

implies a diminution in the fitness of the other species. This evolutionary pressure tends to produce new strategies in the populations involved so as to maintain their chances of survival. This *arms race* ideally increases the capabilities of each species until they reach an optimum.

Cooperative (also called symbiotic) coevolutionary algorithms involve a number of independently evolving species which together form complex structures, well suited to solve a problem. The fitness of an individual depends on its ability to collaborate with individuals from other species. In this way, the evolutionary pressure stemming from the difficulty of the problem favors the development of cooperative strategies and individuals.

Single-population evolutionary algorithms often perform poorly – manifesting stagnation, convergence to local optima, and computational costliness – when confronted with problems presenting one or more of the following features: 1) the sought-after solution is complex, 2) the problem or its solution is clearly decomposable, 3) the genome encodes different types of values, 4) strong interdependencies among the components of the solution, and 5) components-ordering drastically affects fitness [24]. Cooperative coevolution addresses effectively these issues, consequently widening the range of applications of evolutionary computation.

Consider, for instance, the evolution of neural networks [33]. A neural network consists of simple units called neurons, each having several inputs and a single output. The inputs are assigned weights, and a weighted sum of the inputs exceeding a certain threshold causes the neuron to fire an output signal. Neurons are usually connected using a layered topology.

When we approach the task of evolving a neural network possessing some desired property naively, we will probably think of some linearized representation of a neural network, encoding both the neuron locations in the network, and their weights. However, evolving such a network with a simple evolutionary algorithm might prove quite a frustrating task, since much information is encoded in each individual, and it is not homogeneous, which presents us with the difficult target of evolving the individuals as single entities.

On the other hand, this task can be dealt with more sagely via evolving two independently encoded populations of neurons and network topologies. Stanley and Miikkulainen [30] evaluate the fitness of an individual in one of the populations using the individuals of the other. In addition to the simplification of individuals in each population, the fitness is now dynamic, and an improvement in the evolution of topologies triggers a correspond-

ing improvement in the population of neurons, and vice versa.

Lethal Applications

In this section we review a number of applications that – though possibly not killer (death being in the eye of the beholder…) – are most certainly lethal. These come from a sub-domain of evolutionary algorithms, which has been gaining momentum over the past few years: human-competitive machine intelligence. Koza et al. [20] recently affirmed that the field of evolutionary algorithms "now routinely delivers high-return human-competitive machine intelligence", meaning, according to [20]:

- Human-competitive: Getting machines to produce human-like results, e. g., a patentable invention, a result publishable in the scientific literature, or a game strategy that can hold its own against humans.
- High-return: Defined by Koza et al. as a high *artificial-to-intelligence ratio* (**A/I**), namely, the ratio of that which is delivered by the automated operation of the artificial method to the amount of intelligence that is supplied by the human applying the method to a particular system.
- Routine: The successful handling of new problems once the method has been *jump-started*.
- Machine intelligence: To quote Arthur Samuel, getting "machines to exhibit behavior, which if done by humans, would be assumed to involve the use of intelligence."

Indeed, as of 2004 the major annual event in the field of evolutionary algorithms – GECCO (Genetic and Evolutionary Computation Conference; see www.sigevo.org) – boasts a prestigious competition that awards prizes to human-competitive results. As noted at www.human-competitive.org: "Techniques of genetic and evolutionary computation are being increasingly applied to difficult real-world problems – often yielding results that are not merely interesting, but competitive with the work of creative and inventive humans."

We now describe some winners from the HUMIES competition at www.human-competitive.org. Lohn et al. [22] won a Gold Medal in the 2004 competition for an evolved X-band antenna design and flight prototype to be deployed on NASA's Space Technology 5 (ST5) spacecraft:

> The ST5 antenna was evolved to meet a challenging set of mission requirements, most notably the combination of wide beamwidth for a circularly-polarized wave and wide bandwidth. Two evolutionary algorithms were used: one used a genetic algorithm style representation that did not allow branching in the antenna arms; the second used a genetic programming style tree-structured representation that allowed branching in the antenna arms. The highest performance antennas from both algorithms were fabricated and tested, and both yielded very similar performance. Both antennas were comparable in performance to a hand-designed antenna produced by the antenna contractor for the mission, and so we consider them examples of human-competitive performance by evolutionary algorithms [22].

Preble et al. [25] won a gold medal in the 2005 competition for designing photonic crystal structures with large band gaps. Their result is "an improvement of 12.5% over the best human design using the same index contrast platform."

Recently, Kilinç et al. [18] was awarded the Gold Medal in the 2006 competition for designing oscillators using evolutionary algorithms, where the oscillators possess characteristics surpassing the existing human-designed analogs.

Evolutionary Games

Evolutionary games is the application of evolutionary algorithms to the evolution of game-playing strategies for various games, including chess, backgammon, and Robocode.

Motivation and Background

Ever since the dawn of artificial intelligence in the 1950s, games have been part and parcel of this lively field. In 1957, a year after the Dartmouth Conference that marked the official birth of AI, Alex Bernstein designed a program for the IBM 704 that played two amateur games of chess. In 1958, Allen Newell, J. C. Shaw, and Herbert Simon introduced a more sophisticated chess program (beaten in thirty-five moves by a ten-year-old beginner in its last official game played in 1960). Arthur L. Samuel of IBM spent much of the fifties working on game-playing AI programs, and by 1961 he had a checkers program that could play rather decently. In 1961 and 1963 Donald Michie described a simple trial-and-error learning system for learning how to play Tic-Tac-Toe (or Noughts and Crosses) called MENACE (for Matchbox Educable Noughts and Crosses Engine). These are but examples of highly popular games that have been treated by AI researchers since the field's inception.

Why study games? This question was answered by Susan L. Epstein, who wrote:

> There are two principal reasons to continue to do research on games.... First, human fascination with game playing is long-standing and pervasive. Anthropologists have catalogued popular games in almost every culture.... Games intrigue us because they address important cognitive functions.... The second reason to continue game-playing research is that some difficult games remain to be won, games that people play very well but computers do not. These games clarify what our current approach lacks. They set challenges for us to meet, and they promise ample rewards [8].

Studying games may thus advance our knowledge in both cognition and artificial intelligence, and, last but not least, games possess a competitive angle which coincides with our human nature, thus motivating both researcher and student alike.

Even more strongly, Laird and van Lent [21] proclaimed that,

> ...interactive computer games are the killer application for human-level AI. They are the application that will soon need human-level AI, and they can provide the environments for research on the right kinds of problems that lead to the type of the incremental and integrative research needed to achieve human-level AI [21].

Evolving Game-Playing Strategies

Recently, evolutionary algorithms have proven a powerful tool that can automatically *design* successful game-playing strategies for complex games [2,3,13,14,15,27,29].

1. **Chess** (endgames) Evolve a player able to play endgames [13,14,15,29]. While endgames typically contain but a few pieces, the problem of evaluation is still hard, as the pieces are usually free to move all over the board, resulting in complex game trees – both deep and with high branching factors. Indeed, in the chess lore much has been said and written about endgames.
2. **Backgammon** Evolve a full-fledged player for the non-doubling-cube version of the game [2,3,29].
3. **Robocode** A simulation-based game in which robotic tanks fight to destruction in a closed arena (robocode.alphaworks.ibm.com). The programmers implement their robots in the Java programming language, and can test their creations either by using a graphical environment in which battles are held, or by submitting them to a central web site where online tournaments regularly take place. Our goal here has been to evolve Robocode players able to rank high in the international league [27,29].

A strategy for a given player in a game is a way of specifying which choice the player is to make at every point in the game from the set of allowable choices at that point, given all the information that is available to the player at that point [19]. The problem of discovering a strategy for playing a game can be viewed as one of seeking a computer program. Depending on the game, the program might take as input the entire history of past moves or just the current state of the game. The desired program then produces the next move as output. For some games one might evolve a complete strategy that addresses every situation tackled. This proved to work well with Robocode, which is a dynamic game, with relatively few parameters, and little need for past history.

Another approach is to couple a current-state evaluator (e. g., board evaluator) with a next-move generator. One can go on to create a minimax tree, which consists of all possible moves, counter moves, counter counter-moves, and so on; for real-life games, such a tree's size quickly becomes prohibitive. The approach we used with backgammon and chess is to derive a very shallow, single-level tree, and evolve *smart* evaluation functions. Our artificial player is thus had by combining an evolved board evaluator with a simple program that generates all next-move boards (such programs can easily be written for backgammon and chess).

In what follows we describe the definition of six items necessary in order to employ genetic programming: program architecture, set of terminals, set of functions, fitness measure, control parameters, and manner of designating result and terminating run.

Example: Chess

As our purpose is to create a schema-based program that analyzes single nodes thoroughly, in a way reminiscent of human thinking, we did not perform deep lookahead.

We evolved individuals represented as LISP programs. Each such program receives a chess endgame position as input, and, according to its sensors (terminals) and functions, returns an evaluation of the board, in the form of a real number.

Our chess endgame players consist of an evolved LISP program, together with a piece of software that generates all possible (legal) next-moves and feeds them to the program. The next-move with the highest score is selected

(ties are broken stochastically). The player also identifies when the game is over (either by a draw or a win).

Program Architecture As most chess players would agree, playing a winning position (e. g., with material advantage) is very different than playing a losing position, or an even one. For this reason, each individual contains not one but three separate trees: an advantage tree, an even tree, and a disadvantage tree. These trees are used according to the current status of the board. The disadvantage tree is smaller, since achieving a stalemate and avoiding exchanges requires less complicated reasoning. Most terminals and functions were used for all trees.

The structure of three trees per individual was preserved mainly for simplicity reasons. It is actually possible to coevolve three separate populations of trees, without binding them to form a single individual before the end of the experiment. This would require a different experimental setting, and is one of our future-work ideas.

Terminals and Functions While evaluating a position, an expert chess player considers various aspects of the board. Some are simple, while others require a deep understanding of the game. Chase and Simon found that experts recalled meaningful chess formations better than novices [6]. This led them to hypothesize that chess skill depends on a large knowledge base, indexed through thousands of familiar chess patterns.

We assumed that complex aspects of the game board are comprised of simpler units, which require less game knowledge, and are to be combined in some way. Our chess programs use terminals, which represent those relatively simple aspects, and functions, which incorporate no game knowledge, but supply methods of combining those aspects. As we used strongly typed GP [23], all functions and terminals were assigned one or more of two data types: *Float* and *Boolean*. We also included a third data type, named *Query*, which could be used as any of the former two. We also used ephemeral random constants (ERCs).

The Terminal Set We developed most of our terminals by consulting several high-ranking chess players.[2] The terminal set examined various aspects of the chessboard, and may be divided into three groups:

Float values, created using the ERC mechanism. ERCs were chosen at random to be one of the following six values: $\pm 1 \cdot \{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}\} \cdot MAX$ (MAX was empirically set to 1000), and the inverses of these numbers. This guaranteed that when a value was returned after some group of features has been identified, it was distinct enough to engender the outcome.

Simple terminals, which analyzed relatively simple aspects of the board, such as the number of possible moves for each king, and the number of attacked pieces for each player. These terminals were derived by breaking relatively complex aspects of the board into simpler notions. More complex terminals belonged to the next group (see below). For example, a player should capture his opponent's piece if it is not sufficiently protected, meaning that the number of attacking pieces the player controls is greater than the number of pieces protecting the opponent's piece, and the material value of the defending pieces is equal to or greater than the player's. Adjudicating these considerations is not simple, and therefore a terminal that performs this entire computational feat by itself belongs to the next group of complex terminals.

The simple terminals comprising this second group were derived by refining the logical resolution of the previous paragraphs' reasoning: Is an opponent's piece attacked? How many of the player's pieces are attacking that piece? How many pieces are protecting a given opponent's piece? What is the material value of pieces attacking and defending a given opponent's piece? All these questions were embodied as terminals within the second group. The ability to easily embody such reasoning within the GP setup, as functions and terminals, is a major asset of GP.

Other terminals were also derived in a similar manner. See Table 4 for a complete list of simple terminals. Note that some of the terminals are inverted – we would like terminals to always return positive (or true) values, since these values represent a favorable position. This is why we used, for example, a terminal evaluating the player's king's *distance* from the edges of the board (generally a favorable feature for endgames), while using a terminal evaluating the *proximity* of the opponent's king to the edges (again, a positive feature).

Complex terminals: these are terminals that check the same aspects of the board a human player would. Some prominent examples include: the terminal OppPieceCanBeCaptured considering the capture of a piece; checking if the current position is a draw, a mate, or a stalemate (especially important for non-even boards); checking if there is a mate in one or two moves (this is the most complex terminal); the material value of the position; comparing the material value of the position to the original board – this is important since it is easier to consider change than to evaluate the board in an absolute manner. See Table 5 for a full list of complex terminals.

[2] The highest-ranking player we consulted was Boris Gutkin, ELO 2400, International Master, and fully qualified chess teacher.

Genetic and Evolutionary Algorithms and Programming: General Introduction and Appl. to Game Playing, Table 4
Simple terminals for evolving chess endgame players. *Opp:* opponent, *My:* player

Terminal	Description
B=NotMyKingInCheck()	Is the player's king not being checked?
B=IsOppKingInCheck()	Is the opponent's king being checked?
F=MyKingDistEdges()	The player's king's distance form the edges of the board
F=OppKingProximityToEdges()	The opponent's king's proximity to the edges of the board
F=NumMyPiecesNotAttacked()	The number of the player's pieces that are not attacked
F=NumOppPiecesAttacked()	The number of the opponent's attacked pieces
F=ValueMyPiecesAttacking()	The material value of the player's pieces which are attacking
F=ValueOppPiecesAttacking()	The material value of the opponent's pieces which are attacking
B=IsMyQueenNotAttacked()	Is the player's queen not attacked?
B=IsOppQueenAttacked()	Is the opponent's queen attacked?
B=IsMyFork()	Is the player creating a fork?
B=IsOppNotFork()	Is the opponent not creating a fork?
F=NumMovesMyKing()	The number of legal moves for the player's king
F=NumNotMovesOppKing()	The number of illegal moves for the opponent's king
F=MyKingProxRook()	Proximity of my king and rook(s)
F=OppKingDistRook()	Distance between opponent's king and rook(s)
B=MyPiecesSameLine()	Are two or more of the player's pieces protecting each other?
B=OppPiecesNotSameLine()	Are two or more of the opponent's pieces protecting each other?
B=IsOppKingProtectingPiece()	Is the opponent's king protecting one of his pieces?
B=IsMyKingProtectingPiece()	Is the player's king protecting one of his pieces?

Genetic and Evolutionary Algorithms and Programming: General Introduction and Appl. to Game Playing, Table 5
Complex terminals for evolving chess endgame players. *Opp:* opponent, *My:* player. Some of these terminals perform lookahead, while others compare with the original board

Terminal	Description
F=EvaluateMaterial()	The material value of the board
B=IsMaterialIncrease()	Did the player capture a piece?
B=IsMate()	Is this a mate position?
B=IsMateInOne()	Can the opponent mate the player after this move?
B=OppPieceCanBeCaptured()	Is it possible to capture one of the opponent's pieces without retaliation?
B=MyPieceCannotBeCaptured()	Is it not possible to capture one of the player's pieces without retaliation?
B=IsOppKingStuck()	Do all legal moves for the opponent's king advance it closer to the edges?
B=IsMyKingNotStuck()	Is there a legal move for the player's king that advances it away from the edges?
B=IsOppKingBehindPiece()	Is the opponent's king two or more squares behind one of his pieces?
B=IsMyKingNotBehindPiece()	Is the player's king not two or more squares behind one of my pieces?
B=IsOppPiecePinned()	Is one or more of the opponent's pieces pinned?
B=IsMyPieceNotPinned()	Are all the player's pieces not pinned?

Since some of these terminals are hard to compute, and most appear more than once in the individual's trees, we used a memoization scheme to save time [1]: After the first calculation of each terminal, the result is stored, so that further calls to the same terminal (on the same board) do not repeat the calculation. Memoization greatly reduced the evolutionary run-time.

The Function Set The function set used included the If function, and simple Boolean functions. Although our tree returns a real number, we omitted arithmetic functions, for several reasons. First, a large part of contemporary research in the field of machine learning and game theory (in particular for perfect-information games) revolves around inducing logical rules for learning games

Genetic and Evolutionary Algorithms and Programming: General Introduction and Appl. to Game Playing, Table 6
Function set of GP chess player individual. *B:* Boolean, *F:* Float

Function	Description
F=If3(B_1, F_1, F_2)	If B_1 is non-zero, return F_1, else return F_2
B=Or2(B_1, B_2)	Return 1 if at least one of B_1, B_2 is non-zero, 0 otherwise
B=Or3(B_1, B_2, B_3)	Return 1 if at least one of B_1, B_2, B_3 is non-zero, 0 otherwise
B=And2(B_1, B_2)	Return 1 only if B_1 and B_2 are non-zero, 0 otherwise
B=And3(B_1, B_2, B_3)	Return 1 only if B_1, B_2, and B_3 are non-zero, 0 otherwise
B=Smaller(B_1, B_2)	Return 1 if B_1 is smaller than B_2, 0 otherwise
B=Not(B_1)	Return 0 if B_1 is non-zero, 1 otherwise

(for example, see [4,5,11]). Second, according to the players we consulted, while evaluating positions involves considering various aspects of the board, some more important than others, performing logical operations on these aspects seems natural, while mathematical operations does not. Third, we observed that numeric functions sometimes returned extremely large values, which interfered with subtle calculations. Therefore the scheme we used was a (carefully ordered) series of Boolean queries, each returning a fixed value (either an ERC or a numeric terminal, see below). See Table 6 for the complete list of functions.

Fitness Evaluation As we used a competitive evaluation scheme, the fitness of an individual was determined by its success against its peers. We used the random-two-ways method, in which each individual plays against a fixed number of randomly selected peers. Each of these encounters entailed a fixed number of games, each starting from a randomly generated position in which no piece was attacked.

The score for each game was derived from the outcome of the game. Players that managed to mate their opponents received more points than those that achieved only a material advantage. Draws were rewarded by a score of low value and losses entailed no points at all.

The final fitness for each player was the sum of all points earned in the entire tournament for that generation.

Control Parameters and Run Termination We used the standard reproduction, crossover, and mutation operators. The major parameters were: population size – 80, generation count – between 150 and 250, reproduction probability – 0.35, crossover probability – 0.5, and mutation probability – 0.15 (including ERC).

Results We pitted our top evolved chess-endgame players against two very strong external opponents: 1) A program we wrote ('Master') based upon consultation with several high-ranking chess players (the highest being Boris Gutkin, ELO 2400, International Master); 2) CRAFTY – a world-class chess program, which finished second in the 2004 World Computer Speed Chess Championship (www.cs.biu.ac.il/games/). Speed chess (*blitz*) involves a time-limit per move, which we imposed both on CRAFTY and on our players. Not only did we thus seek to evolve good players, but ones that play well *and fast*. Results are shown in Table 7. As can be seen, GP-EndChess manages to hold its own, and even win, against these top players. For more details on GP-EndChess see [13,29].

Genetic and Evolutionary Algorithms and Programming: General Introduction and Appl. to Game Playing, Table 7
Percent of wins, advantages, and draws for best GP-EndChess player in tournament against two top competitors

	%Wins	%Advs	%Draws
Master	6.00	2.00	68.00
CRAFTY	2.00	4.00	72.00

Deeper analysis of the strategies developed [12] revealed several important shortcomings, most of which stemmed from the fact that they used deep knowledge and little search (typically, they developed only *one* level of the search tree). Simply increasing the search depth would not solve the problem, since the evolved programs examine each board very thoroughly, and scanning many boards would increase time requirements prohibitively. And so we turned to evolution to find an optimal way to overcome this problem: How to add more search at the expense of less knowledgeable (and thus less time-consuming) node evaluators, while attaining better performance. In [15] *we evolved the search algorithm itself*, focusing on the *Mate-In-N* problem: find a key move such that even with the best possible counter-plays, the opponent cannot avoid being mated in (or before) move *N*. We showed that our evolved search algorithms successfully solve sev-

eral instances of the Mate-In-*N* problem, for the hardest ones developing 47% less game-tree nodes than CRAFTY. Improvement is thus not over the basic alpha-beta algorithm, but over a world-class program using all standard enhancements [15].

Finally, in [14], we examined a strong evolved chess-endgame player, focusing on the player's emergent capabilities and tactics in the context of a chess match. Using a number of methods we analyzed the evolved player's building blocks and their effect on play level. We concluded that evolution has found combinations of building blocks that are far from trivial and cannot be explained through simple combination – thereby indicating the possible emergence of complex strategies.

Example: Robocode

Program Architecture A Robocode player is written as an event-driven Java program. A main loop controls the tank activities, which can be interrupted on various occasions, called *events*. The program is limited to four lines of code, as we were aiming for the HaikuBot category, one of the divisions of the international league with a four-line code limit. The main loop contains one line of code that directs the robot to start turning the gun (and the mounted radar) to the right. This insures that within the first gun cycle, an enemy tank will be spotted by the radar, triggering a *ScannedRobotEvent*. Within the code for this event, three additional lines of code were added, each controlling a single actuator, and using a single numerical input that was supplied by a genetic programming-evolved sub-program. The first line instructs the tank to move to a distance specified by the first evolved argument. The second line instructs the tank to turn to an azimuth specified by the second evolved argument. The third line instructs the gun (and radar) to turn to an azimuth specified by the third evolved argument (Fig. 3).

Robocode Player

```
while (true)
    TurnGunRight(INFINITY); //main code loop

...

OnScannedRobot() {
    MoveTank(<GP#1>);
    TurnTankRight(<GP#2>);
    TurnGunRight(<GP#3>);
}
```

Genetic and Evolutionary Algorithms and Programming: General Introduction and Appl. to Game Playing, Figure 3
Robocode player's code layout (HaikuBot division)

Genetic and Evolutionary Algorithms and Programming: General Introduction and Appl. to Game Playing, Table 8
Robocode representation. a Terminal set. b Function set (*F:* Float)

Terminal	Description
Energy()	Returns the remaining energy of the player
Heading()	Returns the current heading of the player
X()	Returns the current horizontal position of the player
Y()	Returns the current vertical position of the player
MaxX()	Returns the horizontal battlefield dimension
MaxY()	Returns the vertical battlefield dimension
EnemyBearing()	Returns the current enemy bearing, relative to the current player's heading
EnemyDistance()	Returns the current distance to the enemy
EnemyVelocity()	Returns the current enemy's velocity
EnemyHeading()	Returns the current enemy heading, relative to the current player's heading
EnemyEnergy()	Returns the remaining energy of the enemy
Constant()	An ERC (Ephemeral Random Constant) in the range [−1, 1]
Random()	Returns a random real number in the range [−1, 1]
Zero()	Returns the constant 0

Function	Description
Add(F, F)	Add two real numbers
Sub(F, F)	Subtract two real numbers
Mul(F, F)	Multiply two real numbers
Div(F, F)	Divide first argument by second, if denominator non-zero, otherwise return zero
Abs(F)	Absolute value
Neg(F)	Negative value
Sin(F)	Sine function
Cos(F)	Cosine function
ArcSin(F)	Arcsine function
ArcCos(F)	Arccosine function
IfGreater(F, F, F, F)	If first argument greater than second, return value of third argument, else return value of fourth argument
IfPositive(F, F, F)	If first argument is positive, return value of second argument, else return value of third argument
Fire(F)	If argument is positive, execute fire command with argument as firepower and return 1; otherwise, do nothing and return 0

Terminal and Function Sets We divided the terminals into three groups according to their functionality [27], as shown in Table 8:

1. Game-status indicators: A set of terminals that provide real-time information on the game status, such as

last enemy azimuth, current tank position, and energy levels.
2. Numerical constants: Two terminals, one providing the constant 0, the other being an ERC (ephemeral random constant). This latter terminal is initialized to a random real numerical value in the range [−1, 1], and does not change during evolution.
3. Fire command: This special function is used to curtail one line of code by not implementing the fire actuator in a dedicated line.

Fitness Measure We explored two different modes of learning: using a fixed external opponent as teacher, and coevolution – letting the individuals play against each other; the former proved better. However, not one external opponent was used to measure performance but three, these adversaries downloaded from the HaikuBot league (robocode.yajags.com). The fitness value of an individual equals its average fractional score (over three battles).

Control Parameters and Run Termination The major evolutionary parameters [19] were: population size – 256, generation count – between 100 and 200, selection method – tournament, reproduction probability – 0, crossover probability – 0.95, and mutation probability – 0.05. An evolutionary run terminates when fitness is observed to level off. Since the game is highly nondeterministic a *lucky* individual might attain a higher fitness value than better overall individuals. In order to obtain a more accurate measure for the evolved players we let each of them do battle for 100 rounds against 12 different adversaries (one at a time). The results were used to extract the top player – to be submitted to the international league.

Results We submitted our top player to the HaikuBot division of the international league. At its very first tournament it came in third, later climbing to first place of 28 (robocode.yajags.com/20050625/haiku-1v1.html). All other 27 programs, defeated by our evolved strategy, were written by humans. For more details on GP-Robocode see [27,29].

Backgammon: Major Results

We pitted our top evolved backgammon players against *Pubeval*, a free, public-domain board evaluation function written by Tesauro. The program – which plays well – has become the de facto yardstick used by the growing community of backgammon-playing program developers. Our top evolved player was able to attain a win percentage of 62.4% in a tournament against Pubeval, about 10% higher (!) than the previous top method. Moreover, several evolved strategies were able to surpass the 60% mark, and most of them outdid all previous works. For more details on GP-Gammon see [2,3,29].

Future Directions

Evolutionary computation is a fast growing field. As shown above, difficult, real-world problems are being tackled on a daily basis, both in academia and in industry. In the future we expect major developments in the underlying theory. Partly spurred by this we also expect major new application areas to succumb to evolutionary algorithms, and many more human-competitive results. Expecting such pivotal breakthroughs may seem perhaps a bit of overreaching, but one must always keep in mind Evolutionary Computation's success in Nature.

Bibliography

1. Abelson H, Sussman GJ, Sussman J (1996) Structure and Interpretation of Computer Programs, 2nd edn. MIT Press, Cambridge
2. Azaria Y, Sipper M (2005) GP-Gammon: Genetically programming backgammon players. Genet Program Evolvable Mach 6(3):283–300. doi:10.1007/s10710-005-2990-0
3. Azaria Y, Sipper M (2005) GP-Gammon: Using genetic programming to evolve backgammon players. In: Keijzer M, Tettamanzi A, Collet P, van Hemert J, Tomassini M (eds) Proceedings of 8th European Conference on Genetic Programming (EuroGP2005). Lecture Notes in Computer Science, vol 3447. Springer, Heidelberg, pp 132–142. doi:10.1007/b107383
4. Bain M (1994) Learning logical exceptions in chess. Ph D thesis, University of Strathclyde, Glasgow, Scotland. citeseer.ist.psu.edu/bain94learning.html
5. Bonanno G (1989) The logic of rational play in games of perfect information. Papers 347, California Davis – Institute of Governmental Affairs. http://ideas.repec.org/p/fth/caldav/347.html
6. Charness N (1991) Expertise in chess: The balance between knowledge and search. In: Ericsson KA, Smith J (eds) Toward a general theory of Expertise: Prospects and limits. Cambridge University Press, Cambridge
7. Cramer NL (1985) A representation for the adaptive generation of simple sequential programs. In: Grefenstette JJ (ed) Proceedings of the 1st International Conference on Genetic Algorithms. Lawrence Erlbaum Associates, Mahwah, pp 183–187
8. Epstein SL (1999) Game playing: The next moves. In: Proceedings of the Sixteenth National Conference on Artificial Intelligence. AAAI Press, Menlo Park, pp 987–993
9. Fogel DB (2006) Evolutionary Computation: Toward a New Philosophy of Machine Intelligence, 3rd edn. Wiley-IEEE Press, Hoboken
10. Fogel LJ, Owens AJ, Walsh MJ (1966) Artificial Intelligence Through Simulated Evolution. Wiley, New York
11. Fürnkranz J (1996) Machine learning in computer chess: The next generation. Int Comput Chess Assoc J 19(3):147–161. citeseer.ist.psu.edu/furnkranz96machine.html

12. Hauptman A, Sipper M (2005) Analyzing the intelligence of a genetically programmed chess player. In: Late Breaking Papers at the 2005 Genetic and Evolutionary Computation Conference, distributed on CD-ROM at GECCO-2005, Washington DC
13. Hauptman A, Sipper M (2005) GP-EndChess: Using genetic programming to evolve chess endgame players. In: Keijzer M, Tettamanzi A, Collet P, van Hemert J, Tomassini M (eds) Proceedings of 8th European Conference on Genetic Programming (EuroGP2005). Lecture Notes in Computer Science, vol 3447. Springer, Heidelberg, pp 120–131. doi:10.1007/b107383
14. Hauptman A, Sipper M (2007) Emergence of complex strategies in the evolution of chess endgame players. Adv Complex Syst 10(1):35–59. doi:10.1142/s0219525907001082
15. Hauptman A, Sipper M (2007) Evolution of an efficient search algorithm for the mate-in-*n* problem in chess. In: Ebner M, O'Neill M, Ekárt A, Vanneschi L, Esparcia-Alcázar AI (eds) Proceedings of 10th European Conference on Genetic Programming (EuroGP2007). Lecture Notes in Computer Science vol. 4455. Springer, Heidelberg, pp 78–89. doi:10.1007/978-3-540-71605-1_8
16. Holland JH (1975) Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. University of Michigan Press, Ann Arbor (2nd edn. MIT Press, Cambridge, 1992)
17. Holland JH (1992) Adaptation in Natural and Artificial Systems, 2nd edn. MIT Press, Cambridge
18. Kilinç S, Jain V, Aggarwal V, Cam U (2006) Catalogue of variable frequency and single-resistance-controlled oscillators employing a single differential difference complementary current conveyor. Frequenz: J RF-Eng Telecommun 60(7–8):142–146
19. Koza JR (1992) Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge
20. Koza JR, Keane MA, Streeter MJ, Mydlowec W, Yu J, Lanza G (2003) Genetic Programming IV: Routine Human-Competitive Machine Intelligence. Kluwer, Norwell
21. Laird JE, van Lent M (2000) Human-level AI's killer application: Interactive computer games. In: AAAI-00: Proceedings of the 17th National Conference on Artificial Intelligence. MIT Press, Cambridge, pp 1171–1178
22. Lohn JD, Hornby GS, Linden DS (2005) An evolved antenna for deployment on NASA's Space Technology 5 mission. In: O'Reilly UM, Yu T, Riolo R, Worzel B (eds) Genetic Programming Theory and Practice II, Genetic Programming, vol 8, chap 18. Springer, pp 301–315. doi:10.1007/0-387-23254-0_18
23. Montana DJ (1995) Strongly typed genetic programming. Evol Comput 3(2):199–230. doi:10.1162/evco.1995.3.2.199
24. Peña-Reyes CA, Sipper M (2001) Fuzzy CoCo: A cooperative-coevolutionary approach to fuzzy modeling. IEEE Trans Fuzzy Syst 9(5):727–737. doi:10.1009/91.963759
25. Preble S, Lipson M, Lipson H (2005) Two-dimensional photonic crystals designed by evolutionary algorithms. Appl Phys Lett 86(6):061111. doi:10.1063/1.1862783
26. Schwefel HP (1995) Evolution and Optimum Seeking. Wiley, New York
27. Shichel Y, Ziserman E, Sipper M (2005) GP-Robocode: Using genetic programming to evolve robocode players. In: Keijzer M, Tettamanzi A, Collet P, van Hemert J, Tomassini M (eds) Genetic Programming: 8th European Conference, EuroGP 2005, Lausanne, Switzerland, March 30–April 1, 2005. Lecture Notes in Computer Science, vol 3447. Springer, Berlin, pp 143–154. doi:10.1007/b107383
28. Sipper M (2002) Machine Nature: The Coming Age of Bio-Inspired Computing. McGraw-Hill, New York
29. Sipper M, Azaria Y, Hauptman A, Shichel Y (2007) Designing an evolutionary strategizing machine for game playing and beyond. IEEE Trans Syst, Man, Cybern, Part C: Appl Rev 37(4):583–593
30. Stanley KO, Miikkulainen R (2002) Evolving neural networks through augmenting topologies. Evol Comput 10(2):99–127. doi:10.1162/106365602320169811
31. Turing AM (1950) Computing machinery and intelligence. Mind 59(236):433–460. http://links.jstor.org/sici?sici=0026-4423(195010)2:59:236<433:CMAI>2.0.CO;2-5
32. Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. IEEE Trans Evol Comput 1(1):67–82. doi:10.1109/4235.585893
33. Yao X (1999) Evolving artificial neural networks. Proc IEEE 87(9):1423–1447. doi:10.1009/5.784219

Genetic-Fuzzy Data Mining Techniques

Tzung-Pei Hong[1], Chun-Hao Chen[2], Vincent S. Tseng[2]

[1] Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung, Taiwan

[2] Department of Computer Science and Information Engineering, National Cheng–Kung University, Tainan, Taiwan

Article Outline

Glossary
Definition of the Subject
Introduction
Data Mining
Fuzzy Sets
Fuzzy Data Mining
Genetic Algorithms
Genetic-Fuzzy Data Mining Techniques
Future Directions
Bibliography

Glossary

Data mining Data mining is the process of extracting desirable knowledge or interesting patterns from existing databases for specific purposes. The common techniques include mining association rules, mining sequential patterns, clustering, and classification, among others.

Fuzzy set theory The fuzzy set theory was first proposed by Zadeh in 1965. It is primarily concerned with quantifying and reasoning using natural language in which words can have ambiguous meanings. It is widely used in a variety of fields because of its simplicity and similarity to human reasoning.

Fuzzy data mining The concept of fuzzy sets can be used in data mining to handle quantitative or linguistic data. Basically, fuzzy data mining first uses membership functions to transform each quantitative value into a fuzzy set in linguistic terms and then uses a fuzzy mining process to find fuzzy association rules.

Genetic algorithms Genetic Algorithms (GAs) were first proposed by Holland in 1975. They have become increasingly important for researchers in solving difficult problems since they could provide feasible solutions in a limited amount of time. Each possible solution is encoded as a chromosome (individual) in a population. According to the principle of survival of the fittest, GAs generate the next population by several genetic operations such as crossover, mutation, and reproductions.

Genetic-fuzzy data mining Genetic algorithms have been widely used for solving optimization problems. If the fuzzy mining problem can be converted into an optimization problem, then the GA techniques can easily be adopted to solve it. They are thus called genetic-fuzzy data-mining techniques. They are usually used to automatically mine both appropriate membership functions and fuzzy association rules from a set of transaction data.

Definition of the Subject

Data mining is the process of extracting desirable knowledge or interesting patterns from existing databases for specific purposes. Most conventional data-mining algorithms identify the relationships among transactions using binary values. However, transactions with quantitative values are commonly seen in real-world applications. Fuzzy data-mining algorithms are thus proposed for extracting interesting linguistic knowledge from transactions stored as quantitative values. They usually integrate fuzzy-set concepts and mining algorithms to find interesting fuzzy knowledge from a given transaction data set. Most of them mine fuzzy knowledge under the assumption that a set of membership functions [8,23,24,35,36,50] is known in advance for the problem to be solved. The given membership functions may, however, have a critical influence on the final mining results. Different membership functions may infer different knowledge. Automatically deriving an appropriate set of membership functions for a fuzzy mining problem is thus very important. There are at least two reasons for it. The first one is that a set of appropriate membership functions may not be defined by experts because lots of money and time are needed and experts are not always available. The second one is that data and concepts are always changing along with time. Some mechanisms are thus needed to automatically adapt the membership functions to the changes if needed. The fuzzy mining problem can thus be extended to finding both appropriate membership functions and fuzzy association rules from a set of transaction data.

Recently, genetic algorithms have been widely used for solving optimization problems. If the fuzzy mining problem can be converted into an optimization problem, then the GA techniques can easily be adopted to solve it. They are thus called genetic-fuzzy data-mining techniques. They are usually used to automatically mine both appropriate membership functions and fuzzy association rules from a set of transaction data. Some existing approaches are introduced here. These techniques can dynamically adapt membership functions by genetic algorithms according to some criteria, use them to fuzzify the quantitative transactions, and find fuzzy association rules by fuzzy mining approaches.

Introduction

Most enterprises have databases that contain a wealth of potentially accessible information. The unlimited growth of data, however, inevitably leads to a situation in which accessing desired information from a database becomes difficult. Knowledge discovery in databases (KDD) has thus become a process of considerable interest in recent years, as the amounts of data in many databases have grown tremendously large. KDD means the application of nontrivial procedures for identifying effective, coherent, potentially useful, and previously unknown patterns in large databases [16]. The KDD process [16] is shown in Fig. 1.

In Fig. 1, data are first collected from a single or multiple sources. These data are then preprocessed, including methods such as sampling, feature selection or reduction, data transformation, among others. After that, data-mining techniques are then used to find useful patterns, which are then interpreted and evaluated to form human knowledge.

Especially, data mining plays a critical role to the KDD process. It involves applying specific algorithms for extracting patterns or rules from data sets in a particular representation. Because of its importance, many researchers

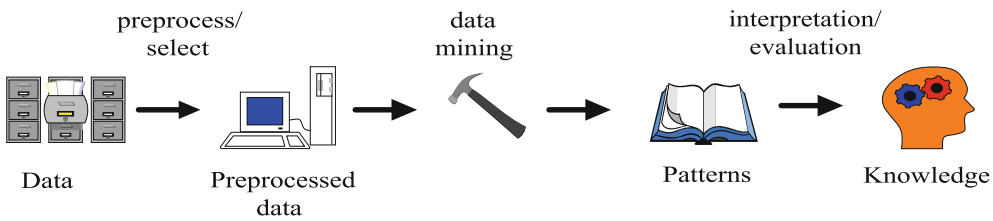

Genetic-Fuzzy Data Mining Techniques, Figure 1
A KDD Process

in the database and machine learning fields are primarily interested in this topic because it offers opportunities to discover useful information and important relevant patterns in large databases, thus helping decision-makers easily analyze the data and make good decisions regarding the domains concerned. For example, there may exist some implicitly useful knowledge in a large database containing millions of records of customers' purchase orders over the recent years. This knowledge can be found by appropriate data-mining approaches. Questions such as "what are the most important trends in customers' purchase behavior?" can thus be easily answered.

Most of the mining approaches were proposed for binary transaction data. However, in real applications, quantitative data exists and should also be considered. Fuzzy set theory is being used more and more frequently in intelligent systems because of its simplicity and similarity to human reasoning [51]. The theory has been applied in fields such as manufacturing, engineering, diagnosis, economics, among others [45,52]. Several fuzzy learning algorithms for inducing rules from given sets of data have been designed and used to good effect within specific domains [7,22]. As to fuzzy data mining, many algorithms are also proposed [8,23,24,35,36,50].

Most of these fuzzy data-mining algorithms assume the membership functions are already known. In fuzzy mining problems, the given membership functions may, however, have a critical influence on the final mining results. Developing effective and efficient approaches to derive both the appropriate membership functions and fuzzy association rules automatically are thus worth being studied. Genetic algorithms are widely used for finding membership functions in different fuzzy applications. In this article, we discuss the genetic-fuzzy data-mining approaches which can mine both appropriate membership functions and fuzzy association rules [9,10,11,12,25,26,27,31,32,33]. The genetic-fuzzy mining problems can be divided into four kinds according to the types of fuzzy mining problems and the ways of processing items. The types of fuzzy mining problems include Single-minimum-Support Fuzzy Mining (SSFM) and Multiple-minimum-Support Fuzzy Mining (MSFM). The ways of processing items include processing all the items together (integrated approach) and processing them individually (divide-and-conquer approach). Each of them will be described in details in the following sections. But first of all, the basic concepts of data mining will be described below.

Data Mining

Data mining techniques have been used in different fields to discover interesting information from databases in recent years. Depending on the type of databases processed, mining approaches may be classified as working on transaction databases, temporal databases, relational databases, multimedia databases, and data streams, among others. On the other hand, depending on the classes of knowledge derived, mining approaches may be classified as finding association rules, classification rules, clustering rules, and sequential patterns, among others.

Finding association rules in transaction databases is most commonly seen in data mining. It is initially applied to market basket analysis for getting relationships of purchased items. An association rule can be expressed as the form $A \rightarrow B$, where A and B are sets of items, such that the presence of A in a transaction will imply the presence of B. Two measures, support and confidence, are evaluated to determine whether a rule should be kept. The support of a rule is the fraction of the transactions that contain all the items in A and B. The confidence of a rule is the conditional probability of the occurrences of items in A and B over the occurrences of items in A. The support and the confidence of an interesting rule must be larger than or equal to a user-specified minimum support and a minimum confidence, respectively.

To achieve this purpose, Agrawal and his coworkers proposed several mining algorithms based on the concept of large item sets to find association rules in transaction data [1,2,3,4]. They divided the mining process into two phases. In the first phase, candidate item sets were gener-

ated and counted by scanning the transaction data. If the number of an itemset appearing in the transactions was larger than a predefined threshold value (called minimum support), the itemset was considered a large itemset. Itemsets containing only one item were processed first. Large itemsets containing only single items were then combined to form candidate itemsets containing two items. This process was repeated until all large itemsets had been found. In the second phase, association rules were induced from the large itemsets found in the first phase. All possible association combinations for each large itemset were formed, and those with calculated confidence values larger than a predefined threshold (called minimum confidence) were output as association rules. In addition to the above approach, there are many other ones proposed for finding association rules.

Most mining approaches focus on binary valued transaction data. Transaction data in real-world applications, however, usually consist of quantitative values. Many sophisticated data-mining approaches have thus been proposed to deal with various types of data [6,46,53]. This also presents a challenge to workers in this research field.

In addition to proposing methods for mining association rules from transactions of binary values, Srikant et al. also proposed a method [46] for mining association rules from those with quantitative attributes. Their method first determines the number of partitions for each quantitative attribute, and then maps all possible values of each attribute into a set of consecutive integers. It then finds large itemsets whose support values are greater than the user-specified minimum-support levels. For example, the following is a quantitative association rule "If *Age* is $[20, \ldots, 29]$, then *Number of Car* is $[0, 1]$" with a support value (60%) and a confidence value (66.6%). This means that if the age of a person is between 20 to 29-years old, then he/she has zero or one car with 66.6%. Of course, different partition approaches for discretizing the quantitative values may influence the final quantitative association rules. Some researches have thus been proposed for discussing and solving this problem [6,53]. Recently, fuzzy sets have also been used in data mining to handle quantitative data due to its ability to deal with the interval boundary problem. The theory of fuzzy sets will be introduced below.

Fuzzy Sets

Fuzzy set theory was first proposed by Zadeh in 1965 [51]. It is primarily concerned with quantifying and reasoning using natural language in which words can have ambiguous meanings. It is widely used in a variety of fields because of its simplicity and similarity to human reasoning [13,45,52]. For example, the theory has been applied in fields such as manufacturing, engineering, diagnosis, economics, among others [19,29,37].

Fuzzy set theory can be thought of as an extension of traditional crisp sets, in which each element must either be in or not in a set. Formally, the process by which individuals from a universal set X are determined to be either members or nonmembers of a crisp set can be defined by a *characteristic or discrimination function* [51]. For a given crisp set A, this function assigns a value $\mu_A(x)$ to every $x \in X$ such that

$$\mu_A(x) = \begin{cases} 1 & \text{if and only if } x \in A \\ 0 & \text{if and only if } x \notin A\,. \end{cases}$$

The function thus maps elements of the universal set to the set containing 0 and 1. This kind of function can be generalized such that the values assigned to the elements of the universal set fall within specified ranges, referred to as the membership grades of these elements in the set. Larger values denote higher degrees of set membership. Such a function is called a membership function, $\mu_A(x)$, by which a fuzzy set A is usually defined. This function is represented by

$$\mu_A : X \to [0, 1]\,,$$

where $[0, 1]$ denotes the interval of real numbers from 0 to 1, inclusive. The function can also be generalized to any real interval instead of $[0, 1]$.

A special notation is often used in the literature to represent fuzzy sets. Assume that x_1 to x_n are the elements in fuzzy set A, and μ_1 to μ_n are, respectively, their grades of membership in A. A is then represented as follows:

$$A = \mu_1/x_1 + \mu_2/x_2 + \cdots + \mu_n/x_n\,.$$

An α-cut of a fuzzy set A is a crisp set A_α that contains all the elements in the universal set X with their membership grades in A greater than or equal to a specified value of α. This definition can be written as

$$A_\alpha = \{x \in X \mid \mu_A(x) \geq \alpha\}\,.$$

The scalar cardinality of a fuzzy set A defined on a finite universal set X is the summation of the membership grades of all the elements of X in A. Thus,

$$|A| = \sum_{x \in X} \mu_A(x)\,.$$

Three basic and commonly used operations on fuzzy sets are complementation, union and intersection, as proposed by Zadeh. They are described as follows.

1. The complementation of a fuzzy set A is denoted by $\neg A$, and the membership function of $\neg A$ is given by

$$\mu_{\neg A}(x) = 1 - \mu_A(x) \quad \forall x \in X .$$

2. The intersection of two fuzzy sets A and B is denoted by $A \cap B$, and the membership function of $A \cap B$ is given by

$$\mu_{A \cap B}(x) = \min\{\mu_A(x), \mu_B(x)\} \quad \forall x \in X .$$

3. The union of two fuzzy sets A and B is denoted by $A \cup B$, and the membership function of $A \cup B$ is given by

$$\mu_{A \cup B}(x) = \max\{\mu_A(x), \mu_B(x)\} \quad \forall x \in X .$$

Note that there are other calculation formula for the complementation, union and intersection, but the above are the most popular.

Fuzzy Data Mining

As mentioned above, the fuzzy set theory is a natural way to process quantitative data. Several fuzzy learning algorithms for inducing rules from given sets of data have thus been designed and used to good effect with specific domains [7,22,41]. Fuzzy data mining approaches have also been developed to find knowledge with linguistic terms from quantitative transaction data. The knowledge obtained is expected to be easy to understand. A fuzzy association rule is shown in Fig. 2.

In Fig. 2, instead of quantitative intervals used in quantitative association rules, linguistic terms are used to represent the knowledge. As we can observe from the rule "If *middle* amount of *bread* is bought, then *high* amount of *milk* is bought", *bread* and *milk* are items, and *middle* and *high* are linguistic terms. The rule means that if the quantity of the purchased item *bread* is *middle*, then there is a high possibility that the associated purchased item is *milk* with *high* quantity.

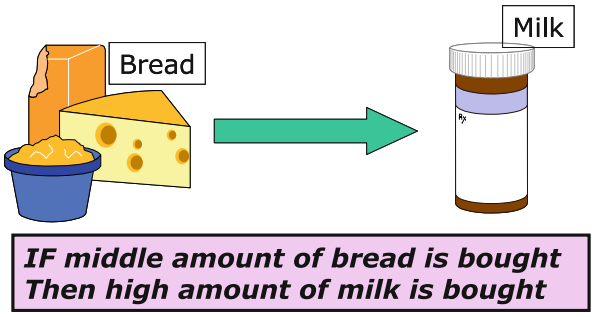

Genetic-Fuzzy Data Mining Techniques, Figure 2
A fuzzy association rule

Many approaches have been proposed for mining fuzzy association rules [8,23,24,35,36,50]. Most of the approaches set a single minimum support threshold for all the items or itemsets and identify the association relationships among transactions. In real applications, different items may have different criteria to judge their importance. Multiple minimum support thresholds are thus proposed for this purpose. We can thus divide the fuzzy data mining approaches into two types, namely Single-minimum-Support Fuzzy Mining (SSFM) [8,23,24,35,50] and Multiple-minimum-Support Fuzzy Mining (MSFM) problems [36].

In the SSFM problem, Chan and Au proposed an F-APACS algorithm to mine fuzzy association rules [8]. They first transformed quantitative attribute values into linguistic terms and then used the adjusted difference analysis to find interesting associations among attributes. Kuok et al. proposed a mining approach for fuzzy association rules. Instead of minimum supports and minimum confidences used in most mining approaches, significance factors and certainty factors were used to derive large itemsets and fuzzy association rules [35]. At nearly the same time, Hong et al. proposed a fuzzy mining algorithm to mine fuzzy rules from quantitative transaction data [23]. Basically, these fuzzy mining algorithms first used membership functions to transform each quantitative value into a fuzzy set in linguistic terms and then used a fuzzy mining process to find fuzzy association rules. Yue et al. then extended the above concept to find fuzzy association rules with weighted items from transaction data [50]. They adopted Kohonen self-organized mapping to derive fuzzy sets for numerical attributes. In general, the basic concept of fuzzy data mining for the SSFM problem is shown in Fig. 3.

In Fig. 3, the process for fuzzy data mining first transforms quantity transactions into a fuzzy representation according to the predefined membership functions. The transformed data are then calculated to generate large itemsets. Finally, the generated large itemsets are used to derive fuzzy association rules.

As to the MSFM problem, Lee et al. proposed a mining algorithm which used multiple minimum supports to mine fuzzy association rules [36]. They assumed that items had different minimum supports and the maximum constraint was used. That is, the minimum support for an itemset was set as the maximum of the minimum supports of the items contained in the itemset. Under the constraint, the characteristic of level-by-level processing was

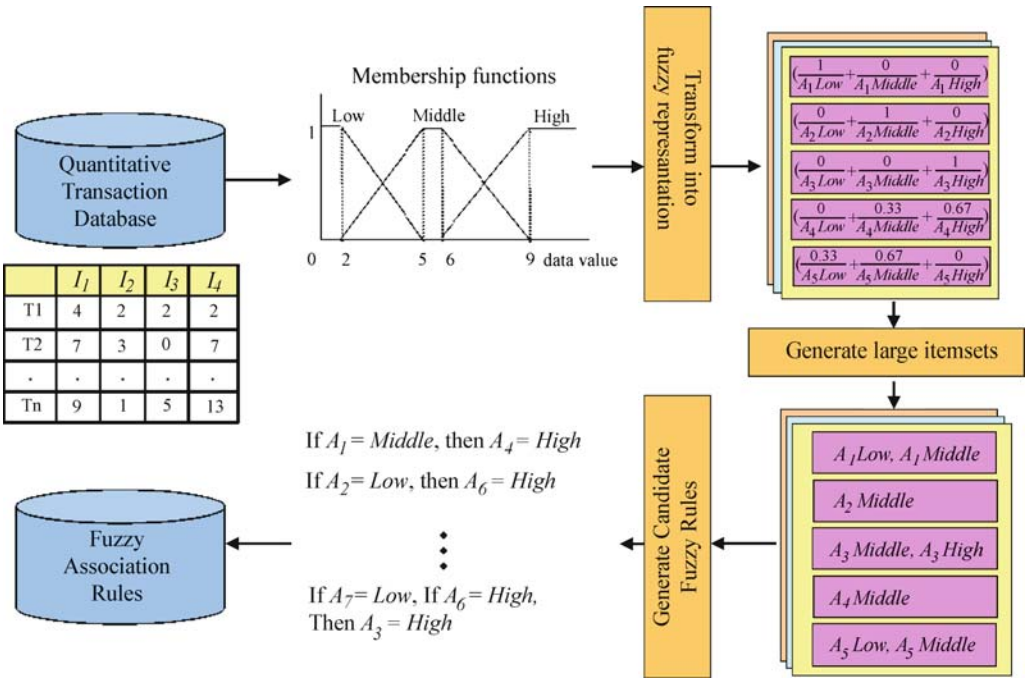

	I_1	I_2	I_3	I_4
T1	4	2	2	2
T2	7	3	0	7
.	.	.	.	.
Tn	9	1	5	13

Genetic-Fuzzy Data Mining Techniques, Figure 3
The concept of fuzzy data mining for the SSFM problem

kept, such that the original a priori algorithm could easily be extended to finding large itemsets. In addition to the maximum constraint, other constraints such as the minimum constraint can also be used with different rationale.

In addition to the above fuzzy mining approaches, fuzzy data mining with taxonomy or fuzzy taxonomy has been developed. Fuzzy web mining is another application of it. Besides, fuzzy data mining is strongly related to fuzzy control, fuzzy clustering, and fuzzy learning.

In most fuzzy mining approaches, the membership functions are usually predefined in advance. Membership functions are, however, very crucial to the final mined results. Below, we will describe how the genetic algorithm can be combined with fuzzy data mining to make the entire process more complete. The concept of the genetic algorithm will first be briefly introduced in the next section.

Genetic Algorithms

Genetic Algorithms (GAs) [17,20] have become increasingly important for researchers in solving difficult problems since they could provide feasible solutions in a limited amount of time [21]. They were first proposed by Holland in 1975 [20] and have been successfully applied to the fields of optimization [17,39,40,43], machine learning [17,39], neural networks [40], fuzzy logic controllers [43], and so on. GAs are developed mainly based on the ideas and techniques from genetic and evolutionary theory [20]. According to the principle of survival of the fittest, they generate the next population by several operations, with each individual in the population representing a possible solution. There are three principal operations in a genetic algorithm.

1. The *crossover* operation: it generates offspring from two chosen individuals in the population by exchanging some bits in the two individuals. The offspring thus inherit some characteristics from each parent.
2. The *mutation* operation: it generates offspring by randomly changing one or several bits in an individual. The offspring may thus possess different characteristics from their parents. Mutation prevents local searches of the search space and increases the probability of finding global optima.
3. The *selection* operation: it chooses some offspring for survival according to predefined rules. This keeps the population size within a fixed constant and puts good offspring into the next generation with a high probability.

On applying genetic algorithms to solving a problem, the first step is to define a representation that describes the

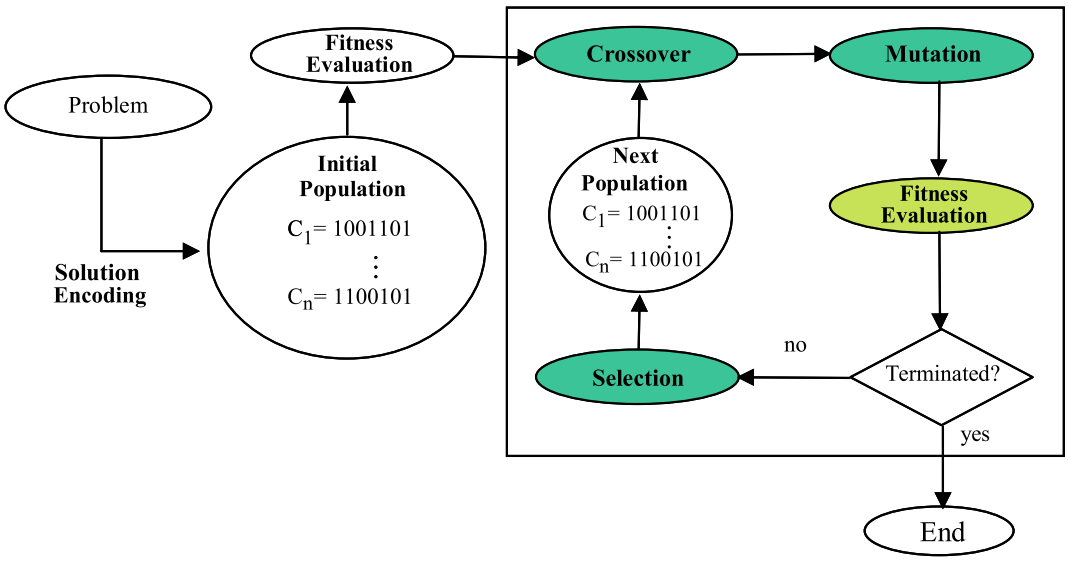

Genetic-Fuzzy Data Mining Techniques, Figure 4
The entire GA process

problem states. The most common way used is the bit string representation. An initial population of *individuals*, called *chromosomes*, is then defined and the three genetic operations (crossover, mutation, and selection) are performed to generate the next generation. Each chromosome in the population is evaluated by a *fitness function* to determine its goodness. This procedure is repeated until a user-specified termination criterion is satisfied. The entire GA process is shown in Fig. 4.

Genetic-Fuzzy Data Mining Techniques

In the previous section for fuzzy data mining, several approaches were introduced, in which the membership functions were assumed to be known in advance. The given membership functions may, however, have a critical influence on the final mining results. Although many approaches for learning membership functions were proposed [14,42,44,47,48], most of them were usually used for classification or control problems. There were several strategies proposed for learning membership functions in classification or control problems by genetic algorithms. Below are some of them:

1. Learning membership functions first, then rules;
2. Learning rules first, then membership functions;
3. Simultaneously learning rules and membership functions;
4. Iteratively learning rules and membership functions.

For fuzzy mining problems, many researches have also been done by combining the genetic algorithm and the fuzzy concepts to discover both suitable membership functions and useful fuzzy association rules from quantitative values. However, most of them adopt the first strategy. That is, the membership functions are first learned and then the fuzzy association rules are derived based on the obtained membership functions. It is done in this way because the number of association rules is often large in mining problems and can not easily be coded in a chromosome.

In this article, we introduce several genetic-fuzzy data mining algorithms that can mine both appropriate membership functions and fuzzy association rules [9,10,11,12,25,26,27,31,32,33]. The genetic-fuzzy mining problems can be divided into four kinds according to the types of fuzzy mining problems and the ways of processing items. The types of fuzzy mining problems include Single-minimum-Support Fuzzy-Mining (SSFM) and Multiple-Minimum-Support Fuzzy-Mining (MSFM) as mentioned above. The ways of processing items include processing all the items together (integrated approach) and processing them individually (divide-and-conquer approach). The integrated genetic-fuzzy approaches encode all membership functions of all items (or attributes) into a chromosome (also called an individual). The genetic algorithms are then used to derive a set of appropriate membership functions according to the designed fitness function. Finally, the best set of membership functions are then used to mine fuzzy

Genetic-Fuzzy Data Mining Techniques, Table 1
The four different genetic-fuzzy data mining problems

	Integrated approach	Divide-and-conquer approach
Single minimum support	IGFSMS Problem	DGFSMS Problem
Multiple minimum supports	IGFMMS Problem	DGFMMS Problem

association rules. On the other hand, the divide-and-conquer genetic-fuzzy approaches encode membership functions of each item into a chromosome. In other words, chromosomes in a population were maintained just for only one item. The membership functions can thus be found for one item after another or at the same time via parallel processing. In general, the chromosomes in the divide-and-conquer genetic-fuzzy approaches are much shorter than those in the integrated approaches since the former only focus on individual items. But there are more application limitations on the former than on the latter. This will be explained later.

The four kinds of problems are thus the Integrated Genetic-Fuzzy problem for items with a Single Minimum Support (IGFSMS) [9,11,26,31,32,33], the Integrated Genetic-Fuzzy problem for items with Multiple Minimum Supports (IGFMMS) [12], the Divide-and-Conquer Genetic-Fuzzy problem for items with a Single Minimum Support (DGFSMS) [10,25,27] and the Divide-and-Conquer Genetic-Fuzzy problem for items with Multiple Minimum Supports (DGFMMS). The classification is shown in Table 1.

Each of the four kinds of genetic-fuzzy data mining problems will be introduced in the following sections.

The Integrated Genetic-Fuzzy Problem for Items with a Single Minimum Support (IGFSMS)

Many approaches have been published for solving the IGFSMS problem [9,11,26,31,32,33]. For example, Hong et al. proposed a genetic-fuzzy data-mining algorithm for extracting both association rules and membership functions from quantitative transactions [26]. They proposed a GA-based framework for searching membership functions suitable for given mining problems and then use the final best set of membership functions to mine fuzzy association rules. The proposed framework is shown in Fig. 5.

The proposed framework consists of two phases, namely mining membership functions and mining fuzzy association rules. In the first phase, the proposed framework maintains a population of sets of membership functions, and uses the genetic algorithm to automatically derive the resulting one. It first transforms each set of membership functions into a fixed-length string. The chromosome is then evaluated by the number of large 1-itemsets and the suitability of membership functions. The fitness value of a chromosome C_q is then defined as

$$f(C_q) = \frac{|L_1|}{\text{suitability}(C_q)},$$

where $|L_1|$ is the number of large 1-itemsets obtained by using the set of membership functions in C_q. Using the number of large 1-itemsets can achieve a trade-off between execution time and rule interestingness. Usually, a larger number of 1-itemsets will result in a larger number of all itemsets with a higher probability, which will thus usually imply more interesting association rules. The evaluation by 1-itemsets is, however, faster than that by all itemsets or interesting association rules. Of course, the number of all itemsets or interesting association rules can also be used in the fitness function. A discussion for different choices of fitness functions can be found in [9].

The suitability measure is used to reduce the occurrence of bad types of membership functions. The two bad types of membership functions are shown in Fig. 6, where the first one is too redundant, and the second one is too separate.

Two factors, called the overlap factor and the coverage factor, are used to avoid the bad shapes. The overlap factor is designed for avoiding the first bad case (too redundant), and the coverage factor is for the second one (too separate). Each factor has its formula for evaluating a value from a chromosome. After fitness evaluation, the approach then chooses appropriate chromosomes for mating, gradually creating good offspring membership function sets. The offspring membership function sets then undergo recursive evolution until a good set of membership functions has been obtained. In the second phase, the final best membership functions are gathered to mine fuzzy association rules. The fuzzy mining algorithm proposed in [24] is adopted to achieve this purpose.

The calculation for large 1-itemsets, however, will still take a lot of time, especially when the database can not totally be fed into the main memory. An enhanced approach, called the cluster-based fuzzy-genetic mining algorithm was thus proposed [11] to speed up the evaluation process and keep nearly the same quality of solutions as that in [26]. That approach also maintains a population of sets of membership functions and uses the genetic algorithm to derive the best one. Before fitness evaluation, the clustering technique is first used to cluster chromosomes. It uses the k-means clustering approach to gather similar

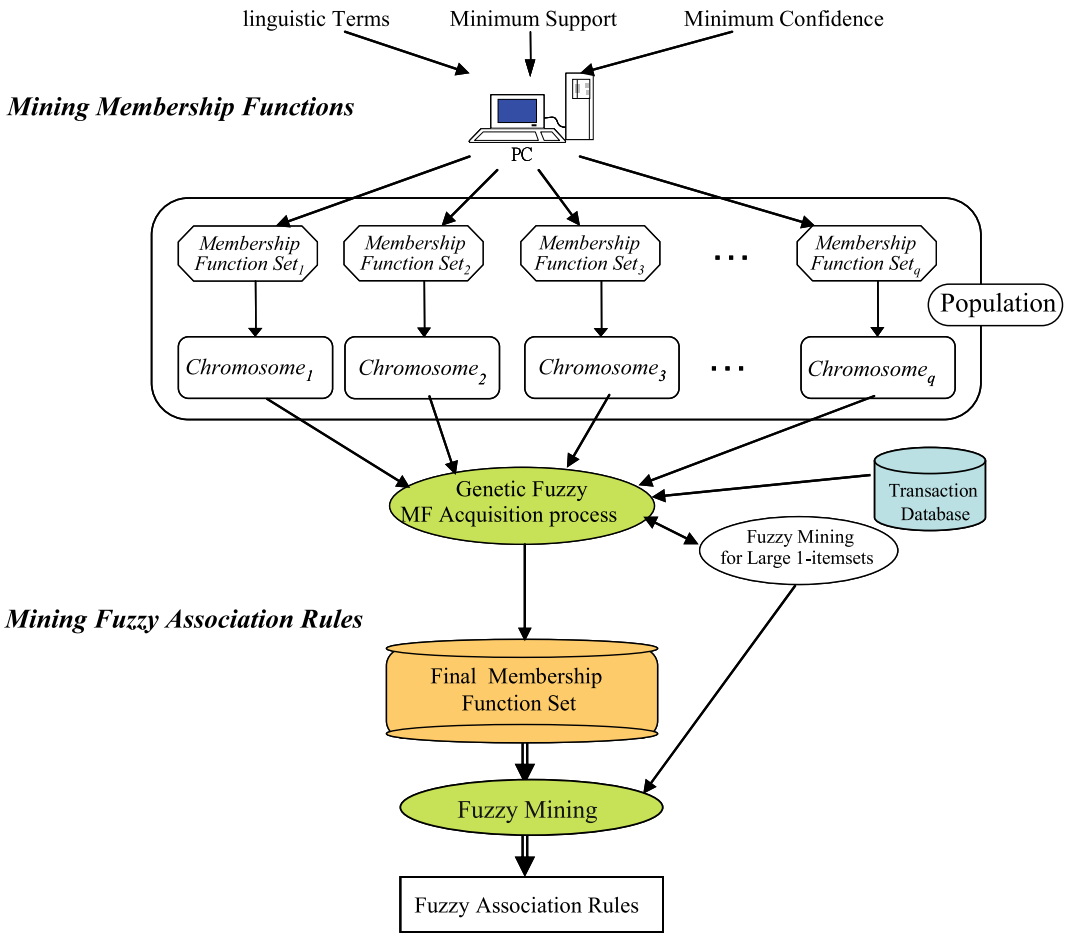

Genetic-Fuzzy Data Mining Techniques, Figure 5
A genetic-fuzzy framework for the IGFSMS problem

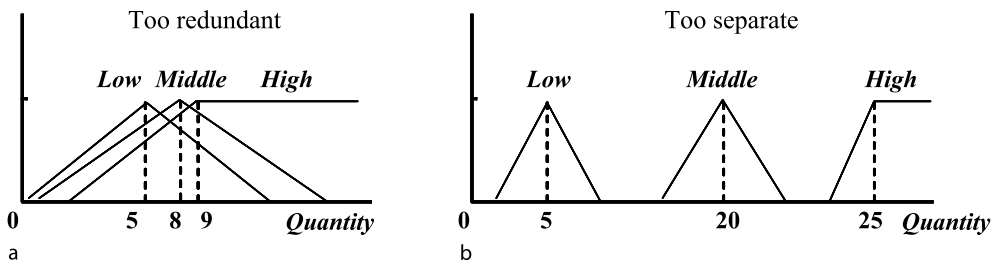

Genetic-Fuzzy Data Mining Techniques, Figure 6
The two bad types of membership functions

chromosomes into groups. The two factors, overlap factor and coverage factor, are used as two attributes for clustering. For example, coverage and overlap factors for ten chromosomes are shown in Table 2, where the column "Suitability" represents the pair (coverage factor, overlap factor).

The k-means clustering approach is then executed to divide the ten chromosomes into k clusters. In this example, assume the parameter k is set at 3. The three clusters found are shown in Table 3. The representative chromosomes in the three clusters are C_5 (4.37, 0.33), C_4 (4.66, 0) and C_9 (4.09, 8.33).

Genetic-Fuzzy Data Mining Techniques, Table 2
The coverage and the overlap factors of ten chromosomes

Chromosome	Suitability	Chromosome	Suitability
C_1	(4, 0)	C_6	(4.5, 0)
C_2	(4.24, 0.5)	C_7	(4.45, 0)
C_3	(4.37, 0)	C_8	(4.37, 0.53)
C_4	(4.66, 0)	C_9	(4.09, 8.33)
C_5	(4.37, 0.33)	C_{10}	(4.87, 0)

Genetic-Fuzzy Data Mining Techniques, Table 3
The three clusters found in the example

Cluster_i	Chromosomes	Representative chromosome
Cluster_1	C_1, C_2, C_5, C_8	C_5
Cluster_2	$C_3, C_4, C_6, C_7, C_{10}$	C_4
Cluster_3	C_9	C_9

All the chromosomes in a cluster use the number of large 1-itemsets derived from the representative chromosome in the cluster and their own suitability of membership functions to calculate their fitness values. Since the number for scanning a database decreases, the evaluation cost can thus be reduced. In this example, the representative chromosomes are chromosomes C_4, C_5, C_9 and it only needs to calculate the number of large 1-itemsets three times. The evaluation results are utilized to choose appropriate chromosomes for mating in the next generation. The offspring membership function sets then undergo recursive evolution until a good set of membership functions has been obtained. Finally, the derived membership functions are used to mine fuzzy association rules.

Kaya and Alhaji also proposed several genetic-fuzzy data mining approaches to derive membership functions and fuzzy association rules [31,32,33]. In [31], the proposed approach tries to derive membership functions, which can get a maximum profit within an interval of user-specified minimum support values. It then uses the derived membership functions to mine fuzzy association rules. The concept of their approaches is shown in Fig. 7.

As shown in Fig. 7a, the approach first derives membership functions from the given quantitative transaction database by genetic algorithms. The final membership functions are then used to mine fuzzy association rules. Figure 7b shows the concept of maximizing the large itemsets of the given minimum support interval. It is used as the fitness function. Kaya and Alhaji also extended the approach to mine fuzzy weighted association rules [32]. Furthermore, fitness functions are not easily defined for GA applications, such that multiobjective genetic algorithms have also been developed [28,30]. In other words, more than one criterion is used in the evaluation. A set of solutions, namely nondominated points (also called Pareto-Optimal Surface), is derived and given to users, instead of only the one best solution obtained by genetic algorithms. Kaya and Alhaji thus proposed an approach based on multiobjective genetic algorithms to learn membership functions, which were then used to generate interesting fuzzy association rules [33]. Three objective functions, namely strongness, interestingness and comprehensibility, were used in their approach to find the Pareto-Optimal Surface. In addition to the above approaches for the IGFSMS problem, some others are still in progress.

The Integrated Genetic-Fuzzy Problem for Items with Multiple Minimum Supports (IGFMMS)

In the above subsection, it can be seen that lots of researches focus on integrated genetic-fuzzy approaches for items with a single minimum support. However, different items may have different criteria to judge their importance. For example, assume among a set of items there are some which are expensive. They are thus seldom bought because of their high cost. Besides, the support values of these items are low. A manager may, however, still be interested in these products due to their high profits. In such cases, the above approaches may not be suitable for this problem. Chen et al. thus proposed another genetic-fuzzy data mining approach [12], which was an extension of the approach proposed in [26], to solve it. The approach combines the clustering, fuzzy and genetic concepts to derive minimum support values and membership functions for items. The final minimum support values and membership functions are then used to mine fuzzy association rules. The genetic-fuzzy mining framework for the IGFMMS problem is shown in Fig. 8.

As shown in Fig. 8, the framework can be divided into two phases. The first phase searches for suitable minimum support values and membership functions of items and the second phase uses the final best set of minimum support values and membership functions to mine fuzzy association rules. The proposed framework maintains a population of sets of minimum support values and membership functions, and uses the genetic algorithm to automatically derive the resulting one.

A genetic algorithm requires a population of feasible solutions to be initialized and updated during the evolution process. As mentioned above, each individual within the population is a set of minimum support values and isosceles-triangular membership functions. Each membership function corresponds to a linguistic term of a certain item. In this approach, the initial set of chromosomes

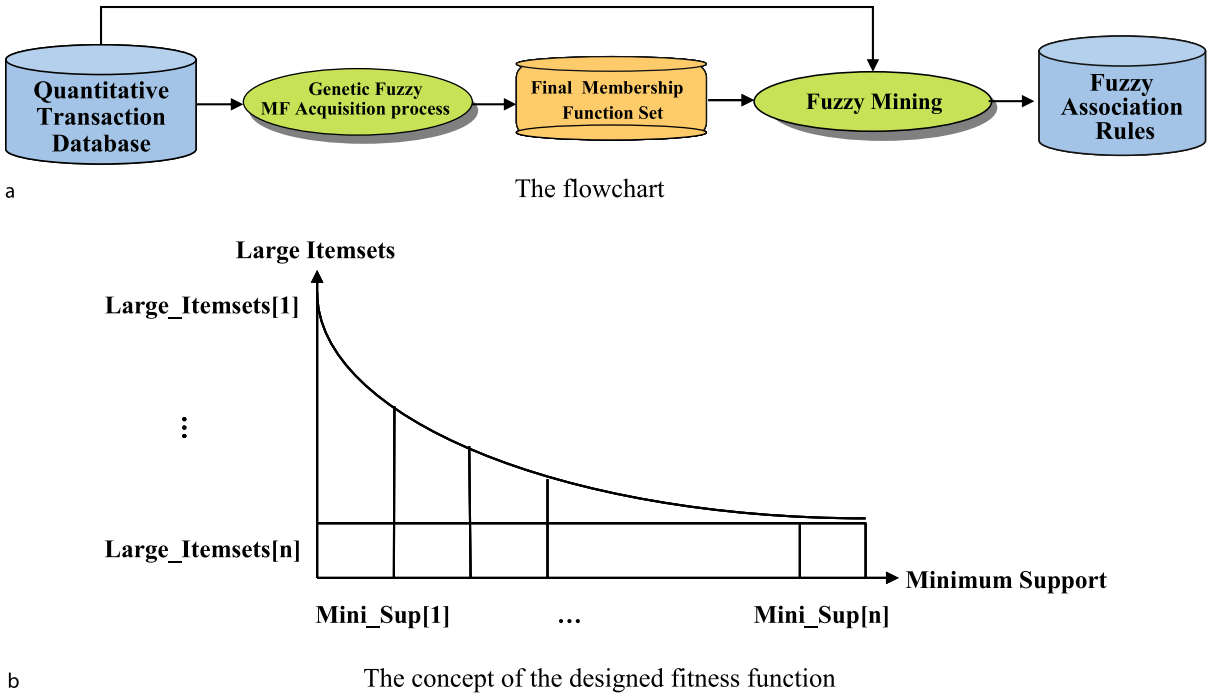

Genetic-Fuzzy Data Mining Techniques, Figure 7
The concept of Kaya and Alhaji's approaches

is generated based on the initialization information derived by the *k*-means clustering approach on the transactions. The frequencies and quantitative values of items in the transactions are the two main factors to gather similar items into groups. The initialization information includes an appropriate number of linguistic terms, the range of possible minimum support values and membership functions of each item. All the items in the same cluster are considered to have similar characteristics and are assigned similar initialization values when a population is initialized. The approach then generates and encodes each set of minimum support values and membership functions into a fixed-length string according to the initialization information.

In this approach, the minimum support values of items may be different. It is hard to assign the values. As an alternative, the values can be determined according to the required number of rules. It is, however, very time-consuming to obtain the rules for each chromosome. As mentioned above, a larger number of 1-itemsets will usually result in a larger number of all itemsets with a higher probability, which will thus usually imply more interesting association rules. The evaluation by 1-itemsets is faster than that by all itemsets or interesting association rules. Using the number of large 1-itemsets can thus achieve a trade-off between execution time and rule interestingness [26].

A criterion should thus be specified to reflect the user preference on the derived knowledge. In the approach, the required number of large 1-itemsets *RNL* is used for this purpose. It is the number of linguistic large 1-itemsets that a user wants to get from an item. It can be defined as the number of linguistic terms of an item multiplied by the predefined percentage which reflects users' preference on the number of large 1-itemsets. It is used to reflect the closeness degree between the number of derived large 1-itemsets and the required number of large 1-itemsets. For example, assume there are three linguistic terms for an item and the predefined percentage p is set at 80%. The *RNL* value is then set as $\lfloor 3 \times 0.8 \rfloor$, which is 2. The fitness function is then composed of the suitability of membership functions and the closeness to the *RNL* value. The minimum support values and membership functions can thus be derived by GA and are then used to mine fuzzy association rules by a fuzzy mining approach for multiple minimum supports such as the one in [36].

The Divide-and-Conquer Genetic-Fuzzy Problem for Items with a Single Minimum Support (DGFSMS)

The advantages of the integrated genetic-fuzzy approaches lie in that they are simple, easy to use, and with few constraints in the fitness functions. In addition to the num-

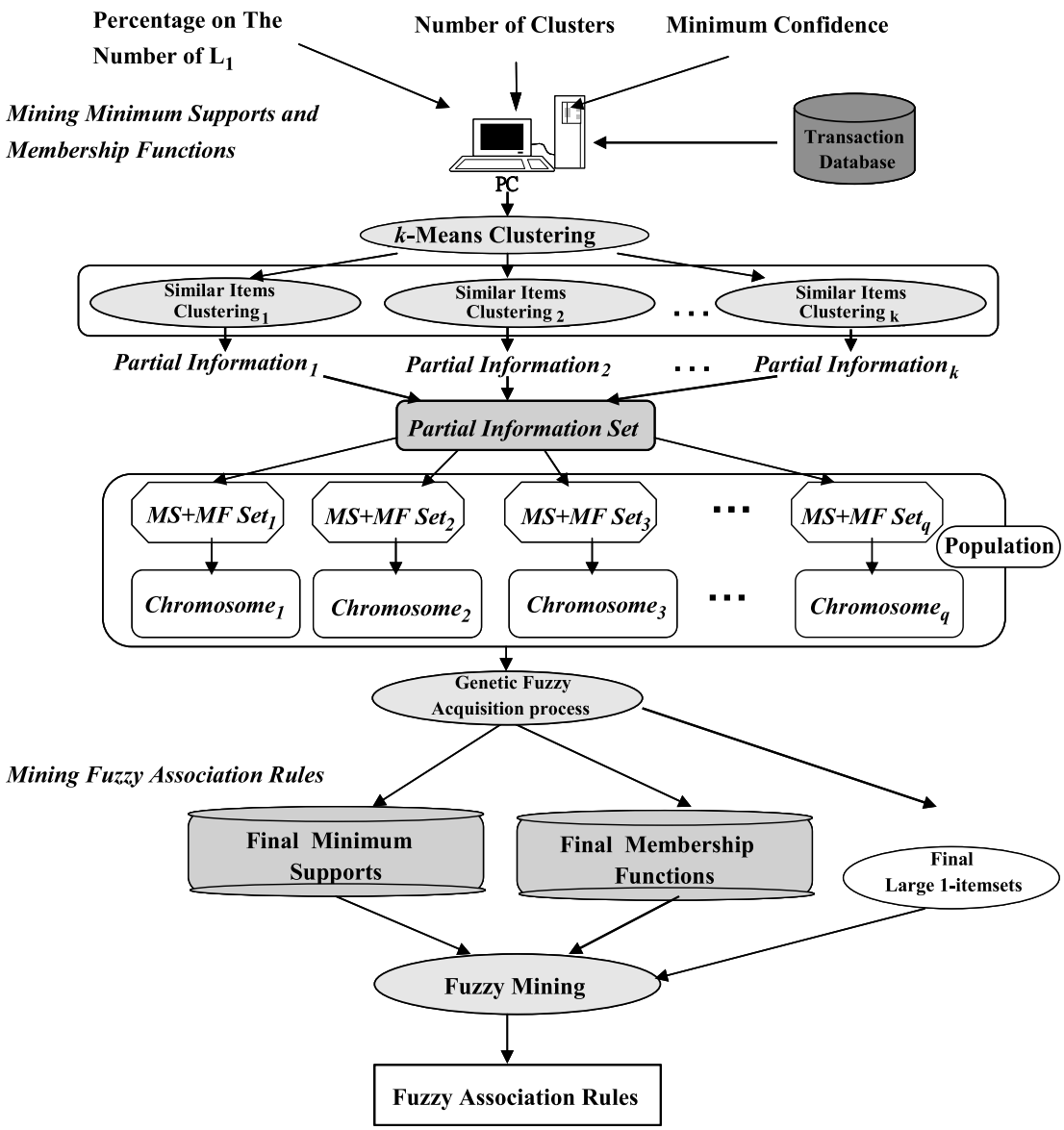

Genetic-Fuzzy Data Mining Techniques, Figure 8
A genetic-fuzzy framework for the IGFMMS problem

ber of large 1-itemsets, the other criteria can also be used. However, if the number of items is large, the integrated genetic-fuzzy approaches may need lots of time to find a near-optimal solution because the length of a chromosome is very long. Recently, the divide-and-conquer strategy has been used in the evolutionary computation community to very good effect. Many algorithms based on it have also been proposed in different applications [5,15, 34,49]. When the number of large 1-itemsets is used in fitness evaluation, the divide-and-conquer strategy becomes a good choice to deal with it since each item can be individually processed in this situation. Hong et al. thus used a GA-based framework with the divide-and-conquer strategy to search for membership functions suitable for the mining problem [27]. The framework is shown in Fig. 9.

The proposed framework in Fig. 9 is divided into two phases: mining membership functions and mining fuzzy association rules. Assume the number of items is m. In the phase of mining membership functions, it maintains m populations of membership functions, with each popula-

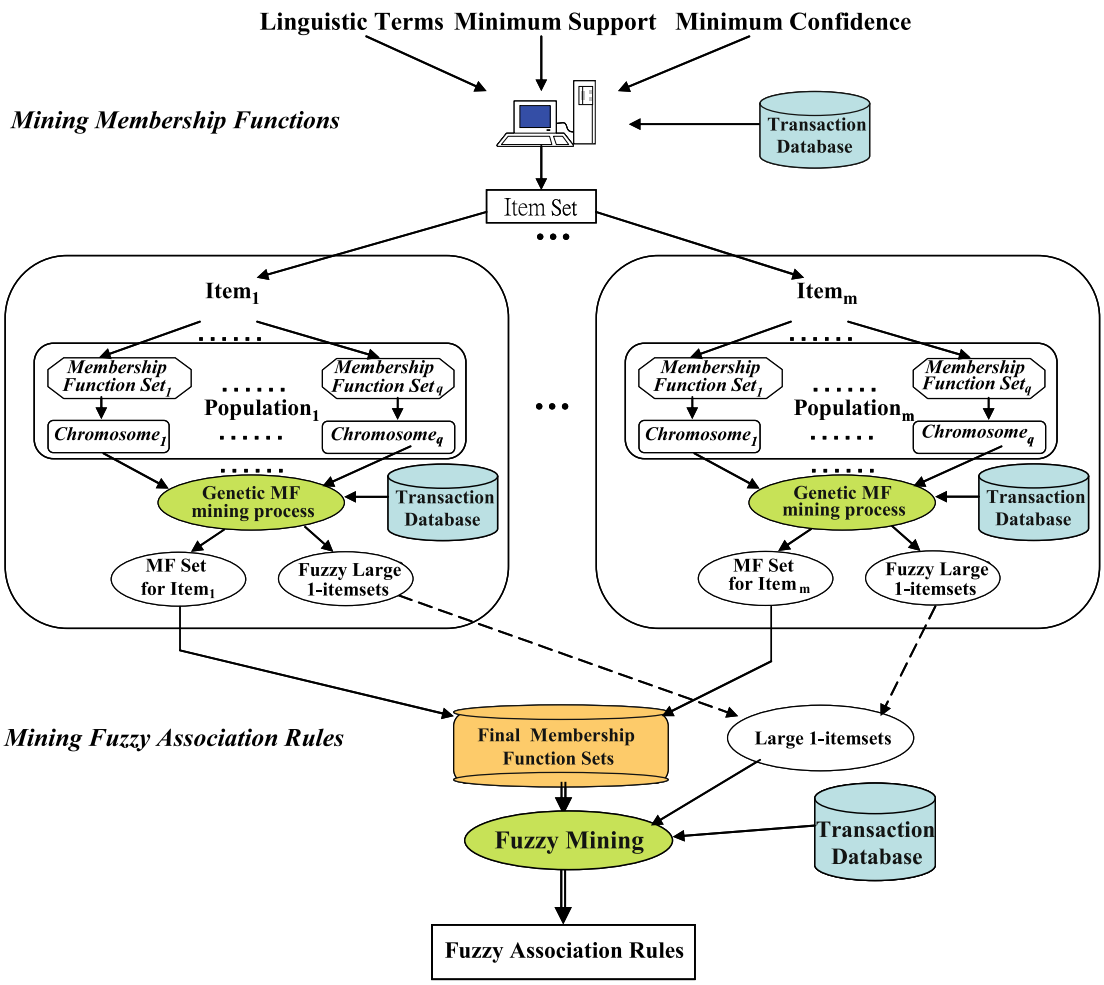

Genetic-Fuzzy Data Mining Techniques, Figure 9
A genetic-fuzzy framework for the DGFSMS problem

tion for an item. Each chromosome in a population represents a possible set of membership functions for that item. The chromosomes in the same population are of the same length. The fitness of each set of membership functions is evaluated by the fuzzy-supports of the linguistic terms in large 1-itemsets and by the suitability of the derived membership functions. The offspring sets of membership functions undergo recursive evolution until a good set of membership functions has been obtained. Next, in the phase of mining fuzzy association rules, the sets of membership function for all the items are gathered together and used to mine the fuzzy association rules from the given quantitative database. An enhanced approach [10], which combines the clustering and the divide-and-conquer techniques, was also proposed to speed up the evaluation process. The clustering idea is similar to that in IGFSMS [11] except that the center value of each membership function is also used as an attribute to cluster chromosomes. The clustering process is thus executed according to the coverage factors, the overlap factors and the center values of chromosomes. For example, assume each item has three membership functions. In total five attributes including one coverage factor, one overlap factor and three center values, are used to form appropriate clusters. Note that the number of linguistic terms for each item is predefined in the mentioned approaches. It may also be automatically and dynamically adjusted [25].

The Divide-and-Conquer Genetic-Fuzzy Problem for Items with Multiple Minimum Supports (DGFMMS)

The problem may be thought of as the combination of the IGFMMS and the *DGFSMS* problems. The framework for the DGFMMS problem can thus be easily designed from

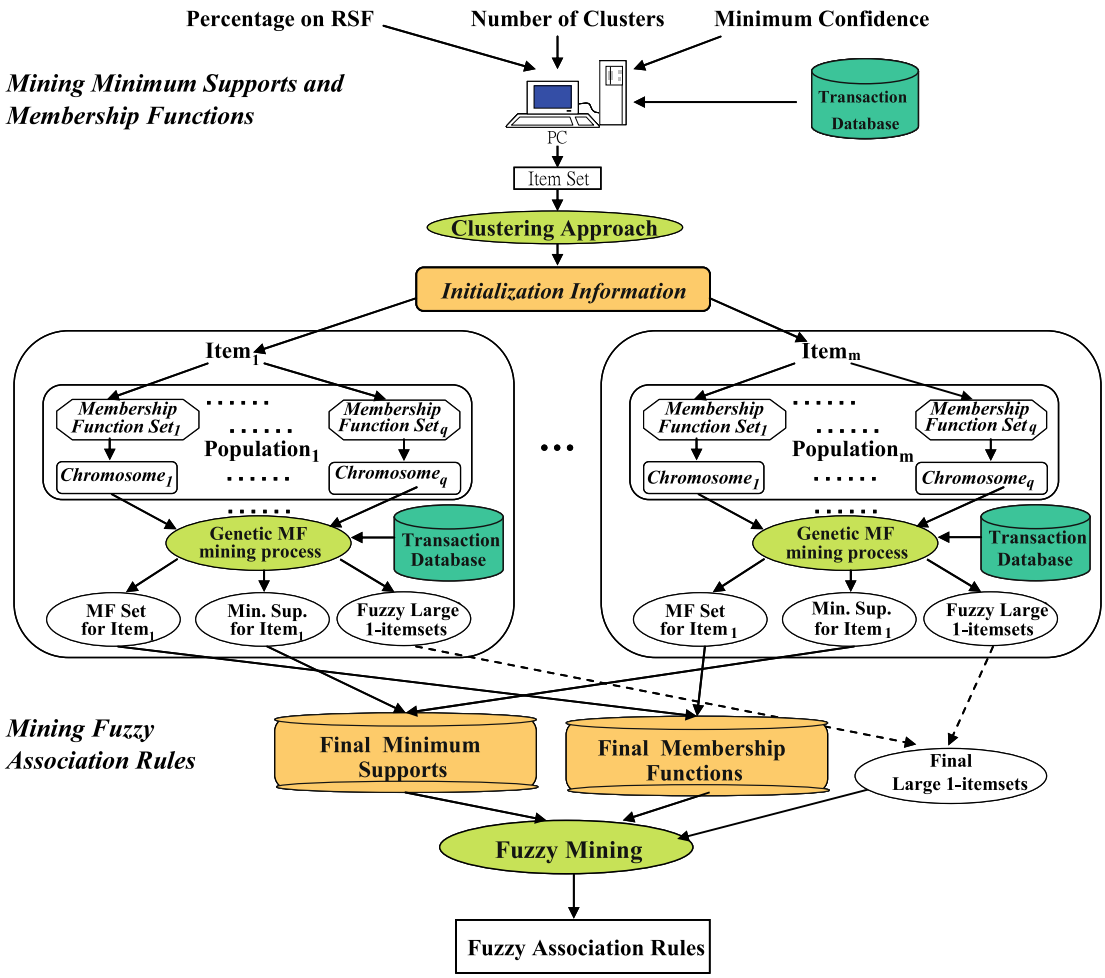

Genetic-Fuzzy Data Mining Techniques, Figure 10
A genetic-fuzzy framework for the DGFMMS problem

the previous frameworks for IGFMMS and DGFSMS. It is shown in Fig. 10.

The proposed framework in Fig. 10 is divided into two phases: mining minimum supports and membership functions, and mining fuzzy association rules. In the first phase, the clustering approach is first used for deriving initialization information which is then used for obtaining better initial populations as used for IGFMMS. It then maintains m populations of minimum supports and membership functions, with each population for an item. Next, in the phase of mining fuzzy association rules, the minimum support values and membership functions for all the items are gathered together and are used to mine fuzzy interesting association rules from the given quantitative database.

Future Directions

In this article, we have introduced some genetic-fuzzy data mining techniques and their classification. The concept of fuzzy sets is used to handle quantitative transactions and the process of genetic calculation is executed to find appropriate membership functions. The genetic-fuzzy mining problems are divided into four kinds according to the types of fuzzy mining problems and the ways of processing items. The types of fuzzy mining problems include Single-minimum-Support Fuzzy-Mining (SSFM) and Multiple-minimum-Support Fuzzy-Mining (MSFM). The methods of processing items include processing all the items together (integrated approach) and processing them individually (divide-and-conquer approach). Each of the

four kinds of problems has been described with some approaches given.

Data mining is very important especially because the data amounts in the information era are extremely large. The topic will continuously grow but with a variety of forms. Some possible research directions in the future about genetic-fuzzy data mining are listed as follows.

- **Applying multiobjective genetic algorithms to the genetic-fuzzy mining problems:** In the article, mined knowledge (number of large itemsets or number of rules) and suitability of membership functions are two important factors used in genetic-fuzzy data mining. Analyzing the relationship between the two factors is thus an interesting and important task. Besides, multiobjective genetic algorithms can also be used to consider the factors at the same time.
- **Analyzing effects of different shapes of membership functions and different genetic operators:** Different shapes of membership functions may have different results on genetic-fuzzy data mining. They may be evaluated in the future. Different genetic operations may also be tried to obtain better results than the ones used in the above approaches.
- **Enhancing performance of the fuzzy-rule mining phase:** The final goal of genetic-fuzzy mining techniques introduced in this article is to mine appropriate fuzzy association rules. However, the phase of mining fuzzy association rules is very time-consuming. How to improve the process of mining interesting fuzzy rules is thus worth studying. Some possible approaches include modifying existing approaches, combining the existing ones with other techniques, or defining new evaluation criteria.
- **Developing visual tools for these genetic-fuzzy mining approaches:** Another interesting aspect of future work is to develop visual tools for demonstrating the genetic-fuzzy mining results. It can help a decision maker easily understand or get useful information quickly. The visual tools may include, for example, how to show the derived membership functions and to illustrate the interesting fuzzy association rules.

Bibliography

1. Agrawal R, Srikant R (1994) Fast algorithm for mining association rules. In: Proceedings of the international conference on very large data bases, pp 487–499
2. Agrawal R, Imielinksi T, Swami A (1993) Database mining: a performance perspective. Trans IEEE Knowl Data Eng 5(6):914–925
3. Agrawal R, Imielinksi T, Swami A (1993) Mining association rules between sets of items in large database. In: Proceedings of the conference ACMSIGMOD, Washington DC, USA
4. Agrawal R, Srikant R, Vu Q (1997) Mining association rules with item constraints. In: Proceedings of the third international conference on knowledge discovery in databases and data mining, Newport Beach, California, August 1997
5. Au WH, Chan KCC, Yao X (2003) A novel evolutionary data mining algorithm with applications to churn prediction. Trans IEEE Evol Comput 7(6):532–545
6. Aumann Y, Lindell Y (1999) A statistical theory for quantitative association rules. In: Proceedings of the ACMSIGKDD international conference on knowledge discovery and data mining, pp 261–270
7. Casillas J, Cordón O, del Jesus MJ, Herrera F (2005) Genetic tuning of fuzzy rule deep structures preserving interpretability and its interaction with fuzzy rule set reduction. Trans IEEE Fuzzy Syst 13(1):13–29
8. Chan CC, Au WH (1997) Mining fuzzy association rules. In: Proceedings of the conference on information and knowledge management, Las Vegas, pp 209–215
9. Chen CH, Hong TP, Vincent Tseng S (2007) A comparison of different fitness functions for extracting membership functions used in fuzzy data mining. In: Proceedings of the symposium IEEE on foundations of computational intelligence, pp 550–555
10. Chen CH, Hong TP, Tseng VS (2007) A modified approach to speed up genetic-fuzzy data mining with divide-and-conquer strategy. In: Proceedings of the congress IEEE on evolutionary computation (CEC), pp 1–6
11. Chen CH, Tseng VS, Hong TP (2008) Cluster-based evaluation in fuzzy-genetic data mining. Trans IEEE Fuzzy Syst 16(1):249–262
12. Chen CH, Hong TP, Tseng VS, Lee CS (2008) A genetic-fuzzy mining approach for items with multiple minimum supports. Accepted and to appear in Soft Computing (SCI)
13. Chen J, Mikulcic A, Kraft DH (2000) An integrated approach to information retrieval with fuzzy clustering and fuzzy inferencing. In: Pons O, Vila MA, Kacprzyk J (eds) Knowledge management in fuzzy databases. Physica, Heidelberg
14. Cordón O, Herrera F, Villar P (2001) Generating the knowledge base of a fuzzy rule-based system by the genetic learning of the data base. Trans IEEE Fuzzy Syst 9(4):667–674
15. Darwen PJ, Yao X (1997) Speciation as automatic categorical modularization. Trans IEEE Evol Comput 1(2):101–108
16. Frawley WJ, Piatetsky-Shapiro G, Matheus CJ (1991) Knowledge discovery in databases: an overview. In: Proceedings of the workshop AAAI on knowledge discovery in databases, pp 1–27
17. Goldberg DE (1989) Genetic algorithms in search, optimization and machine learning. Addison Wesley, Boston
18. Grefenstette JJ (1986) Optimization of control parameters for genetic algorithms. Trans IEEE Syst Man Cybern 16(1):122–128
19. Heng PA, Wong TT, Rong Y, Chui YP, Xie YM, Leung KS, Leung PC (2006) Intelligent inferencing and haptic simulation for Chinese acupuncture learning and training. Trans IEEE Inf Technol Biomed 10(1):28–41
20. Holland JH (1975) Adaptation in natural and artificial systems. University of Michigan Press, Michigan
21. Homaifar A, Guan S, Liepins GE (1993) A new approach on the traveling salesman problem by genetic algorithms. In: Proceedings of the fifth international conference on genetic algorithms

22. Hong TP, Lee YC (2001) Mining coverage-based fuzzy rules by evolutional computation. In: Proceedings of the international IEEE conference on data mining, pp 218–224
23. Hong TP, Kuo CS, Chi SC (1999) Mining association rules from quantitative data. Intell Data Anal 3(5):363–376
24. Hong TP, Kuo CS, Chi SC (2001) Trade-off between time complexity and number of rules for fuzzy mining from quantitative data. Int J Uncertain Fuzziness Knowledge-Based Syst 9(5):587–604
25. Hong TP, Chen CH, Wu YL, Tseng VS (2004) Finding active membership functions in fuzzy data mining. In: Proceedings of the workshop on foundations of data mining in the fourth international IEEE conference on data mining
26. Hong TP, Chen CH, Wu YL, Lee YC (2006) AGA-based fuzzy mining approach to achieve a trade-off between number of rules and suitability of membership functions. Soft Comput 10(11):1091–1101
27. Hong TP, Chen CH, Wu YL, Lee YC (2008) Genetic-fuzzy data mining with divide-and-conquer strategy. Trans IEEE Evol Comput 12(2):252–265
28. Ishibuchi H, Yamamoto T (2004) Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining. Fuzzy Sets Syst 141:59–88
29. Ishibuchi H, Yamamoto T (2005) Rule weight specification in fuzzy rule-based classification systems. Trans IEEE Fuzzy Syst 13(4):428–435
30. Jin Y (2006) Multi-objective machine learning. Springer, Berlin
31. Kaya M, Alhajj R (2003) A clustering algorithm with genetically optimized membership functions for fuzzy association rules mining. In: Proceedings of the international IEEE conference on fuzzy systems, pp 881–886
32. Kaya M, Alhaji R (2004) Genetic algorithms based optimization of membership functions for fuzzy weighted association rules mining. In: Proceedings of the international symposium on computers and communications, vol 1, pp 110–115
33. Kaya M, Alhajj R (2004) Integrating multi-objective genetic algorithms into clustering for fuzzy association rules mining. In: Proceedings of the fourth international IEEE conference on data mining, pp 431–434
34. Khare VR, Yao X, Sendhoff B, Jin Y, Wersing H (2005) Co-evolutionary modular neural networks for automatic problem decomposition. In: Proceedings of the (2005) congress IEEE on evolutionary computation, vol 3, pp 2691–2698
35. Kuok C, Fu A, Wong M (1998) Mining fuzzy association rules in databases, Record SIGMOD 27(1):41–46
36. Lee YC, Hong TP, Lin WY (2004) Mining fuzzy association rules with multiple minimum supports using maximum constraints. In: Lecture notes in computer science, vol 3214. Springer, Heidelberg, pp 1283–1290
37. Liang H, Wu Z, Wu Q (2002) A fuzzy based supply chain management decision support system. In: Proceedings of the world congress on intelligent control and automation, vol 4, pp 2617–2621
38. Mamdani EH (1974) Applications of fuzzy algorithms for control of simple dynamic plants. Proc IEEE 121(12):1585–1588
39. Michalewicz Z (1994) Genetic algorithms + data structures = evolution programs. Springer, New York
40. Mitchell M (1996) An introduction to genetic algorithms. MIT Press, Cambridge MA
41. Rasmani KA, Shen Q (2004) Modifying weighted fuzzy subsethood-based rule models with fuzzy quantifiers. In: Proceedings of the international IEEE conference on fuzzy systems, vol 3, pp 1679–1684
42. Roubos H, Setnes M (2001) Compact and transparent fuzzy models and classifiers through iterative complexity reduction. Trans IEEE Fuzzy Syst 9(4):516–524
43. Sanchez E et al (1997) Genetic algorithms and fuzzy logic systems: soft computing perspectives (advances in fuzzy systems – applications and theory, vol 7). World-Scientific, River Edge
44. Setnes M, Roubos H (2000) GA-fuzzy modeling and classification: complexity and performance. Trans IEEE Fuzzy Syst 8(5):509–522
45. Siler W, James J (2004) Fuzzy expert systems and fuzzy reasoning. Wiley, New York
46. Srikant R, Agrawal R (1996) Mining quantitative association rules in large relational tables. In: Proceedings of the (1996) international ACMSIGMOD conference on management of data, Montreal, Canada, June 1996, pp 1–12
47. Wang CH, Hong TP, Tseng SS (1998) Integrating fuzzy knowledge by genetic algorithms. Trans IEEE Evol Comput 2(4):138–149
48. Wang CH, Hong TP, Tseng SS (2000) Integrating membership functions and fuzzy rule sets from multiple knowledge sources. Fuzzy Sets Syst 112:141–154
49. Yao X (2003) Adaptive divide-and-conquer using populations and ensembles. In: Proceedings of the (2003) international conference on machine learning and application, pp 13–20
50. Yue S, Tsang E, Yeung D, Shi D (2000) Mining fuzzy association rules with weighted items. In: Proceedings of the international IEEE conference on systems, man and cybernetics, pp 1906–1911
51. Zadeh LA (1965) Fuzzy set. Inf Control 8(3):338–353
52. Zhang H, Liu D (2006) Fuzzy modeling and fuzzy control. Springer, New York
53. Zhang Z, Lu Y, Zhang B (1997) An effective partitioning-combining algorithm for discovering quantitative association rules. In: Proceedings of the Pacific-Asia conference on knowledge discovery and data mining, pp 261–270

Genome Organization

Gustavo Glusman, Arian F. A. Smit
Institute for Systems Biology, Seattle, USA

Article Outline

Glossary
Definition
Introduction
Molecular Structure of the Genome
Genome Size vs. Organismal Complexity
The Genome as Habitat for Transposable Elements
Genomes in Constant Flux
Transcriptome Complexity
Transcript Networks
Future Directions
Bibliography

Glossary

Genome The entire genetic complement of a species.

Transcript An RNA molecule derived by copying from a DNA template.

Splicing The process of removing introns from a transcript and thereby joining the exons into a mature RNA molecule.

Alternative splicing The ability to process a transcript by splicing in more than one way, leading to the production of alternative RNA products.

Gene In the simplest cases, a segment of DNA that specifies the structure of a protein or an RNA molecule. The definition becomes less straightforward in complex genomic loci able to produce several overlapping transcripts with potentially disparate functions.

Recombination A process by which two chromosomes exchange segments, typically during the formation of sperm and eggs.

Gene conversion A process by which a segment of one chromosome is copied to another chromosome, replacing the original sequence.

Definition

With rare exceptions, all known living organisms encode their genetic material in the form of double-stranded DNA, in one or more chromosomes, collectively referred to as the "genome". A cell lacking its genome cannot survive for long, since it cannot produce new transcripts in response to environmental challenges. For example, red blood cells lose their DNA and can only function as oxygen shuttles until they break down. The genome includes most of the information needed by the cells to stay alive, to differentiate into new cell types, and to perform their functions in the context of the organism. As such, it is the ultimate resource for identifying the full set of components in the living system. Eukaryotic genomes are much larger than strictly needed to encode the relatively modest set of genes in them, but several mechanisms give rise to a very complex transcriptome.

Introduction

In 1920, the German botanist Hans Winkler wrote: "I propose the expression Genom for the haploid chromosome set, which, together with the pertinent protoplasm, specifies the material foundations of the species …" (Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreiche, Verlag Fischer, Jena). Since then, the word "genome" has evolved in meaning somewhat and gave rise to words like "genomics", defining whole new fields of research.

The word "genome" currently has two main meanings, depending on the context:

1. The genome of a species: The corpus of genetic material that characterizes a species, including its most frequent genetic variants. This meaning of "the genetic endowment of a species" is essentially that intended by Winkler's original definition. For species for which the genome has been sequenced, the "reference genome" denotes the specific sequence that was obtained. Even though the reference genome is frequently a patchwork of sequences derived from several sources, it is taken to be the representative sequence for the species.
2. The genome of an organism or an individual, also known as the "personal genome". This refers to the specific instance of the genetic material present in most cells of a given organism, stressing the individual's variations or polymorphisms, and including both parental copies.

A genome is sometimes taken to represent "the set of genes of an organism". Indeed, some proposed approaches to sequencing personal genomes involve re-sequencing only the exons of known genes. It is known though, that the vast expanses of "junk" DNA within and between genes harbor scattered signals that are important for the regulation of gene expression: A clean-cut subdivision between functional and non-functional DNA is not quite possible. The simplest and most convenient definition of the genome includes in it all the genetic material, regardless of our current state of knowledge about its potential function.

Finally, the word "genome" has also been borrowed to other contexts. For example in Computer Science, when discussing genetic algorithms as a tool for discovering solutions hidden in a complex space that cannot be searched exhaustively, a genome is the set of parameters that define a possible solution to a problem.

The genome has been likened, among other things, to "a blueprint" and "a parts list". Neither of these is entirely appropriate: A blueprint specifies a detailed construction plan, without reference to the origin of the components, while the parts list simply enumerates the components without reference to their interaction. The genome incorporates both aspects.

The genome is an instruction book, providing all the information needed to maintain a cell's structural integrity and to perform its functions, to usher it through the complicated ritual of cell division, and to direct its development into more specialized structures. The genome in a cell includes the instructions for building cellular com-

ponents, many of which are in turn used as tools and as raw materials to make other cellular components, in a complex cascade that requires extensive regulation and control. The information required for that control is also encoded in the genome in many different ways. The construction plan (the "blueprint") is for the most part not explicitly specified in the genome, but rather it emerges from the complex interactions between the cellular components and their spatiotemporal patterns of expression.

Molecular Structure of the Genome

Molecular Topology

Nucleic acids are linear polymers with clearly defined polarity, conventionally denoted as running from 5′ to 3′ end of the molecule. The linearity and polarity severely limit the range of possible molecular topologies: A single-stranded molecule of DNA or RNA can remain linear or form a circle. The specific sequence of nucleotide bases confers variability: Depending on the sequence, local structures can form by base complementarity between different parts of the molecule. A central property of nucleic acids, crucial to the transmission of hereditary information, is the ability to form double-stranded structures, in which two molecules are intertwined and held together by extensive base complementarity. A double-stranded nucleic acid molecule can, again, remain linear or be closed into a circle. Depending on the sequence and the torsional forces induced by various processes, a variety of local structures can emerge and some are used for regulation [69].

Even though a linear molecule would appear to have a simpler topology than a circular structure, the linearity introduces some mechanistic complications. DNA replication requires priming and proceeds linearly along the DNA molecule from 5′ to 3′. While one strand of DNA can be replicated continuously, the complementary strand needs to be synthesized in pieces (called Okazaki fragments), which are then ligated. A consequence of this is that the replication of linear DNA molecules requires special structures (called telomeres) and mechanisms to avoid becoming shorter each generation due to the inability to prime the 3′ ends. Changes in telomere length have been implicated in disease and in senescence [13].

Evolution of Genome Structure

When discussing genome organization, it is important to keep in mind that there are three main lineages of living systems: Eukaryotes, bacteria and archaea, the last two collectively called prokaryotes. There are essentially two main

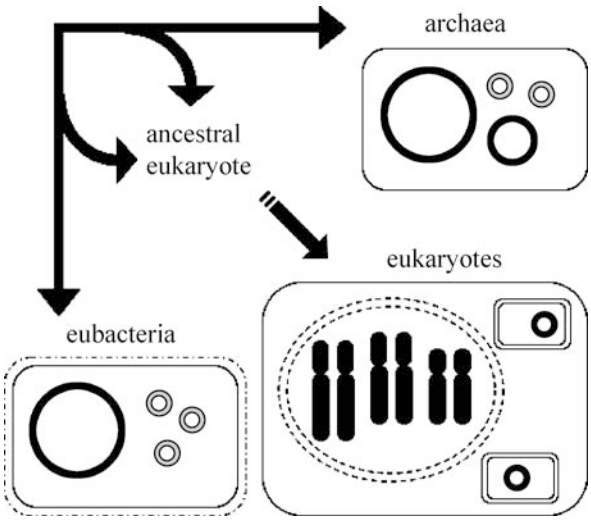

Genome Organization, Figure 1
Evolution of gross genome structure. Two main lineages, leading to modern archaea and modern eubacteria, merged to produce the ancestral eukaryotic genome. This third lineage adopted a large number of innovations including linear chromosomes, homologous chromosome pairs (diploid genome), sequestration in the nucleus (*double dashed lines*), multicellularity, etc. Eukaryotic cells include double-membraned mitochondria, which are the remnants of the endosymbiotic bacteria from the ancestral genomic merger. Gray circles indicate plasmids (non-essential genetic material) while black represents chromosomes (essential genetic material)

types of chromosomal organization, emphasizing circular and linear chromosomes; these are characteristic of prokaryotes and eukaryotes, respectively (Fig. 1). The ancestral eukaryotic genome arose from repeated fusion of archaeal and bacterial genomes, followed by a large number of structural and functional innovations [52,93,120].

The typical prokaryotic genome is organized into a small number of circular DNA molecules. Some of these (and usually the largest) harbor genes that are essential for the survival of the organism, and are considered to be proper chromosomes: These need to be segregated accurately to ensure proper cell division [33]. Prokaryotic genomes typically include also "plasmids", which are much smaller and include only non-essential genes. The prokaryotic genome is located in a region of the cell called the nucleoid, which is not separated from the rest of the cytoplasm by a membrane.

In contrast, the typical eukaryotic genome includes several linear chromosomes, and sometimes a large number of them. The eukaryotic chromosomes are sequestered in the cell's nucleus, a complex structure including a double membrane perforated by pores, which mediate the transport of materials between the nucleus and the cyto-

plasm. During cellular division, the chromosomes adopt a highly condensed "rod" chromatin structure (as depicted in Fig. 1), but during the interphase (between cell divisions) chromosomes expand to fill the entire volume of the nucleus, largely separated into chromosomal territories [15]. The precise organization of the chromosomes within the nucleus has not yet been fully characterized, but there is clear indication of the existence of self-organizing microenvironments within the nucleus [98], which may support the integration of regulatory signals controlling gene expression [132].

In addition to the linear chromosomes in the nucleus, eukaryotic cells contain several copies of very small, circular DNA molecules located in the mitochondria; plants have additional such chromosomes in the chloroplasts. These organellar chromosomes include genes that are essential for the maintenance and function of their respective organelle, and therefore they are essential to the organism. The mitochondrial and chloroplast chromosomes are therefore considered an integral part of the eukaryotic genome, despite their prokaryotic characteristics.

A further, profound difference in genome organization between prokaryotes and eukaryotes relates to the number of copies of the entire set of chromosomes found in the cell. Prokaryotes normally have one copy of the main chromosome per cell, though in fast growing conditions the chromosome may be continually in an intermediate state of replication, yielding a fractional chromosome copy number. In eukaryotes, the entire set of chromosomes (called the haploid set) is frequently doubled (the diploid state) with each chromosome being represented by two homologous copies. Cells in diploid state can divide into two cells retaining the diploid state through the carefully orchestrated process of mitosis, or they can produce haploid cells (which have a single copy of each chromosome) via meiosis. Haploid cells can fuse to generate diploid cells. There is significant variation among species in the life cycle: For example, vertebrate cells spend most of the time in diploid state and generate haploid gametes only for reproduction, while most fungi are normally haploid organisms, and plants switch between both states. Changes in ploidy have profound implications on gene redundancy and on the regulation of gene dosage, and often leads to sterility in animals, while offspring are likely to be isolated from the rest of the population. Polyploidy through whole-genome duplication is therefore a significant disruption requiring many further adaptations, and often leads to speciation with an increase in chromosome number [105]. Polyploidy can arise from mistakes during meiosis, either within one species (autoploidy) or through hybridization of two closely related species (allopolyploidy); the latter may be a rare event, but has the advantage of basically resulting in a diploid with twice the number of chromosomes without the chromosomal segregation issues associated with autoploidy. Relatively recent events are abundant in plants and are easily recognized. Quite a few clear examples exist in animals as well, like the polyploidizations in the ancestors of the salmon family (Salmonidae) and of the *Xenopus* genus of frogs. The only case in mammals is found among the South American octodont rodents, where two of the thirteen species have double the number of chromosomes and double the amount of DNA relative to their sister species.

Genome Size vs. Organismal Complexity

The first fully sequenced genomes were those of viruses, starting with the 3569 nucleotide long bacteriophage MS2, an RNA virus [35] and the 5386 nucleotide long Phage Φ-X174, a DNA virus [102]. While viral genomes are not representative of the genomes of free-living organisms, their organization gave first insights into how compact and "optimized" a genome can be.

The first bacterial genome to be completed was that of *Haemophilus influenzae* [38]. To date, several hundred bacterial genomes have been sequenced, and their full sequences are available through online databases (Table 1). Hundreds of eukaryotic genomes have been sequenced or are currently in draft form, providing an astounding resource for the research community (Table 1). Landmark papers include preliminary analyzes [1,2,19,54,126].

One of the most obvious characteristics of bacterial genomes is the high density of genes, with compact open reading frames that are not interrupted by introns, and that are separated by very little intergenic sequence. The use of DNA by prokaryotes gives the impression of being highly optimized, with 80%–95% of the genome coding for proteins. In extreme cases, the same DNA sequence can be part of more than one ORF, in different frames or strands. Bacterial genes are typically organized into operons, which are sets of consecutive genes on the same strand that are transcribed as a unit.

By extrapolating from the observed characteristics of prokaryotic genomes, it would only be natural to expect the much larger eukaryotic genomes to include many more genes. For a sense of proportion, a prokaryotic genome expanded to the size of a vertebrate genome could encode a few million genes. Such expanded gene numbers would appear to be a simple explanation for the higher organismal complexity observed in eukaryotes. As it turns out, though, vertebrate genomes are much less gene dense than prokaryotic genomes: The coding fraction in eukary-

Genome Organization, Table 1
Selected internet resources

Resource description	URL
Integrated Microbial Genomes system at JGI	http://img.jgi.doe.gov/
TIGR Comprehensive Microbial Resource	http://cmr.tigr.org/tigr-scripts/CMR/shared/Genomes.cgi
NCBI Entrez Genome Database	http://www.ncbi.nlm.nih.gov/sites/entrez?db=Genome
UCSC Genome Bioinformatics (the Genome Browser)	http://genome.ucsc.edu/
miRBase, a database of miRNA sequences and targets	http://microrna.sanger.ac.uk
NHGRI Encyclopedia of DNA Elements (ENCODE)	http://www.genome.gov/10005107
Personal Genome Project	http://arep.med.harvard.edu/PGP/

otes decreases with genome size, down to ~1% in humans. In other words, while the human genome is three orders of magnitude larger than that of *E. coli*, it has only one order of magnitude more genes.

There is no simple correlation between genome size and the observed complexity of the organism, and plenty of outliers exist, e. g. the huge genomes of amoebas. Nevertheless, there is a clear ranking from viruses all the way to multicellular eukaryotes, in terms of genome size, gene number, mobile element number, intron number and size, size of intergenic spacer DNA and complexity of regulatory regions [84].

The reduction in coding fraction results from the presence of introns (discussed below) and huge intergenic distances. Two main hypotheses have been raised to explain the growth in genome sizes: The selfish DNA hypothesis, and the bulk DNA hypothesis [48,122]. The selfish DNA hypothesis postulates that the genome size reflects the ecology of the selfish elements multiplying within it. There is a very large number of elements that can multiply within the genome by various mechanisms, giving rise to extensive families of "interspersed repeats" discussed below, and leading to an accumulation of DNA throughout the genome. In contrast, the bulk DNA hypothesis proposes that genome size is actively selected for, since it dictates many cellular features like cell size [23,24], which in turn affects higher level features of the organism [48].

Selection for fast replication [31] and the metabolic cost of DNA [99] have been invoked as possible alternative explanations for the streamlining of microbial genome sizes. Neither appears to be well supported [82], while a deletional bias [86] may be the major force behind variations in bacterial genome size.

The Genome as Habitat for Transposable Elements

The changes that can befall a genomic sequence are of two kinds. The first can be considered accidental, like mistakes during replication, incomplete repair of environmental damage, or sloppy recombination. The second type is due to the selection-driven propagation of transposable elements (TEs). These mutations generally do not take place in a completely random fashion. Many apparent patterns and correlations between genomic features, "genomic complexity" that appears to hint at an underlying organization, are likely due to mutational skews and not to differential adaptive pressures in different regions of the genome. One example may be the so-called isochore structure in land vertebrate (amniote) genomes: The nucleotide composition of large (100 kb-range) genomic regions can range from 30% to 60% G+C nucleotides. This variation correlates with a number of other characteristics, like gene density and intron size, leading initially to suggestions that this nucleotide heterogeneity is functionally maintained. However, it now seems more likely that the variation arises from regional differences in recombination rates; the associated process of gene conversion, in which a short region of one allele is converted to the other, appears to favor resolution of mismatches to G/C nucleotides [32].

Under relatively stable environmental conditions, a genome's replication, repair and recombination machinery can evolve to reach a balance between stability and flexibility. Given the selective forces on TEs to keep propagating, it is unclear how much control a genome has on the accumulation of mutations of the second type. In fact, the bulk of many eukaryotic genomes is formed by copies of TEs. Being generally without function, these copies have accumulated mutations in a neutral fashion and decayed to various degrees, and are recognized as interspersed repetitive DNA, or "repeats". The most common repeats are derived from retrotransposons, which reproduce via reverse transcription (RNA to DNA) of their transcripts. These include familiar elements like (endogenous) retroviruses, LINE1 and *Alu* in the human genome. The other common elements are DNA transposons, which move by a cut-and-paste mechanism in which the transposase specifically recognizes the element's termini. A major conceptual difference between these two groups is that

the proteins in the first mechanism (reverse transcriptase, integrase, etc.) operate in *cis* and those of the second in *trans*: The cytosol-born transposase does not operate on the transcript close at hand but on any DNA copy in the nucleus. Because of this, retrotransposons are able to survive over long periods of evolutionary time in a genome while DNA transposons generally become extinct through the accumulation of defective copies, and depend on frequent horizontal transfer for evolutionary survival [109].

The isolation of the germline in land vertebrates appears to have had a profound effect on the activity of transposable elements in those organisms, by severely limiting the introduction of new elements through horizontal transfer. This is reflected in the nature of interspersed repeats in amniotes (mammals, birds and reptiles) as compared to those in amphibia, fish, and most other organisms: While DNA transposon copies, which are dependent on horizontal transfer for long-term survival, are common in most organisms, they represent only a very small fraction of interspersed repeats in most amniotic genomes. The exceptional mutagenic potential of DNA transposons through the formation of double-strand breaks after excision, may explain the far less dynamic evolution of amniote genomes [18] compared to that of, for example, insects [29] and plants [9].

This is not to say that retrotransposons necessarily are less deleterious for the host genome. Insertions of both DNA transposons and retrotransposons can incapacitate genes, but retrotransposons tend to carry stronger transcriptional regulatory sites. It is likely that the most common form of gene disruption by TEs is through the introduction of transcriptional termination site within introns [46].

From an information science point-of-view, interspersed repetitive DNA decreases the complexity of the genome, as a large fraction can be expressed as derivatives of a limited number of sequences. From a biological point of view, however, interspersed repeats lead to many complications. For one, they provide homologous seeds for "ectopic" recombination between different sites of the genome, resulting in generally deleterious deletions, insertions, and chromosomal crossovers. A less appreciated effect is the concomitant distribution of binding sites for transcription factors, as most TEs carry highly active promoter sites and other transcriptional regulatory sites. Interspersed repeats that have not decayed too far from their original form may account for much spurious transcription in the genome, and even if local transcription is not disturbed, the distribution of thousands of new binding sites for a cellular transcription factor could influence the balance of the regulatory network.

Genomes may not be able to direct the level of activity of transposable elements, but multiple mechanisms have evolved to suppress it. An often exploited, telltale feature of TE activity of any kind is the presence of transcripts in both orientations, as eventually some copies will be co-transcribed with genes in reverse direction. Some fungal genomes inactivate thus detected TEs by directly altering the genomic copies [42]. In placental mammals a cytosine-amidase has evolved into a family of (APOBEC3) proteins that act as an innate immune system inhibiting the replication of retrotransposons and retroviruses by deaminating cytosines (to uracils) in its nascent DNA strands during reverse transcription; the resulting uracil-containing cDNAs are fodder for degradation by cellular uracyl-DNA-glycolysases [34,53]. Much more widespread mechanisms to suppress TE activity involve transcriptional inactivation by methylation or heterochromatization [16], and the post-transcriptional silencing by RNA interference. The Piwi defense system may explain a characteristic pattern in retrotransposon evolution: A retrotransposon seems to be active in bursts, after which the element either goes extinct, or a modified descendant, perhaps unrecognized by RNA interference, takes over. Considering that TEs must have been an overarching problem since the first genome arose, it is not unlikely that methylation, heterochromatization, and RNA interference in general originated as defense mechanisms against parasitic DNA, and only later were adopted for others purposes like cellular gene regulation.

It is clear that, on top of their mutagenic properties and creation of repetitive DNA, TEs have had an enormous impact on genomic organization. Some of the most obvious features of eukaryotic genomes seem to have found their origin in TEs:

- Most ancient introns probably originated in early eukaryotic evolution from mobile elements, possibly resembling present-day prokaryotic group II introns.
- Telomeres are maintained by the enzyme telomerase, which appears to be a reverse transcriptase derived from a retrotransposon [45].
- The major centromere binding protein CENP-B was derived from related transposases, in a process that happened independently in mammals and in fission yeast, while their genomic binding sites (the centromeres) may be derived from their transposon sequences [22,110].
- The genomic rearrangements that underlie the vertebrate adaptive immune response are performed by a domesticated transposase [60].

The cellular adoption of transposable elements proteins, as in the last three examples, is a form of "exaptation" (the

utilization of a structure for a function other than that for which it developed). There are probably over a hundred exapted TE genes, often recombined with cellular genes, in the human genome alone [54]. Non-coding TE sequences have also been exapted, often with many copies at the time [81,89]. An exciting aspect of the adoption of a group of interspersed sequences is that a regulatory network could arise seemingly from scratch. For example, retroviral long terminal repeats carry a battery of transcriptionally regulatory sites, which at opportune locations may become established and can regulate a set of cellular genes in a coordinated fashion [92,124]. Most of the exapted TE peptides in the human genome are DNA binding domains of transposases, which very specifically recognize binding sites at the termini of the dispersed transposon copies. One reason for the frequency of their adoption could be that, after recombination with a cellular transcription factor, of which multiple examples are known [54,109], such proteins and fortunately located transposon copies could create an entirely new regulatory network.

To summarize, the genome is, in a very real sense, the evolutionary environment of a host of selfish mobile genetic elements. The impact of their propagation within the genome is hard to overstate.

Genomes in Constant Flux

Gene Duplication

The total number of genes is commonly taken as a measure of genomic complexity, though, as we describe below, the number of different gene products and levels of regulation may be a more relevant measure. Whereas prokaryotic gene number can shift drastically by horizontal transfer, almost all increase in eukaryotic gene number is due to gene duplication: Exceptions are genes acquired from organelle genomes, retroviruses and transposable elements.

The gene content of most eukaryotic genomes is highly dynamic. From a variety of approaches it is estimated that 1 out of 50 genes duplicate every million years in vertebrate genomes. However, since the rate at which new genes are inactivated is far higher (10- to 100-fold) than the rate at which they arise, most duplicated genes are short lived. The total gene number may remain stable for long periods of time, but there is likely to be a continuous turnover with respect to which genes have redundant copies. In the human genome, about 6% of the genes have undergone duplication since our separation from most other mammals, while about 12% of the nucleotides have been substituted. The majority of the resulting gene duplicates have been inactivated.

These numbers exclude the increase in gene number through whole genome duplication (polyploidization); such events in the distant past are referred to as paleopolyploidy. The "2R hypothesis" postulates that two such events happened in early vertebrate evolution, initially almost quadrupling the number of genes evolution could tinker with [63]. Paleopolyploidy has probably been a central mechanism for large-scale genome evolution but is typically hard to prove, mostly due to the inevitable fast decay of many of the redundant genes.

Genes can also duplicate individually or as part of a genomic segment (segmental duplication), rather than along with the whole genome. One type of duplication necessarily restricted to a single gene is retroposition (reverse transcription of the gene's transcript and reintegration), which results in an intronless, "processed" gene. This mechanism is common in mammals and over time has given rise to tens of thousands of recognizable copies in the human genome, the great majority of which are inactive processed pseudogenes. This process relies on the LINE1 transposable element, which can retropose any transcript with a poly-A tail. Most other LINE-like transposable elements only interact with transcripts that share the LINE's terminal sequence and are very unlikely to copy cellular genes, as witnessed by the absence of processed pseudogenes in many organisms. For example, the chicken has none, despite the abundant presence of the LINE-like element CR1.

The low "success rate" of producing functional genes through retroposition is due to the frequent 5′ truncation of the transcript during integration and the inherent inability to copy the transcriptionally regulatory sequences along with the gene. Even when a working copy results, there is only a small chance that it will be functionally conserved, as it is initially redundant. Nevertheless, over a thousand retroposed genes appear to be transcriptionally active in humans, and at least 120 of these retrocopies appear to have evolved into *bona fide* human genes [121].

Transposable elements are central to several other mechanisms of gene duplication. There are three common ways in which a gene or a gene fragment can be co-transposed with a transposable element.

1. Transcription of a retrotransposon can proceed through the poly-A signal into neighboring DNA. Such "transduced" regions are often short, but human functional genes duplicated in this fashion have been reported [129]. Given the usual 5′ truncation of retroposed sequences, the transposable element itself may not be visible at the new site, and the frequency of these events may be underestimated.

2. Rolling circle transposons can similarly transduce 3′ flanking sequences, probably by extending the replication process past the original (weakly defined) 3′ signal. Elements up to about 20 kb in length have been observed as very common haplotype differences between strains of maize [71]. The longer elements usually contain fragments of multiple cellular genes, indicating frequent but short read-through with each transposition event. In mammals, rolling circle elements were active in a common ancestor over 150 million years ago, while more recent activity has only been reported in several species of bats [95].
3. Genes sandwiched between two copies of the highly specific recognition site of the transposase of a DNA transposon may be duplicated as a unit. This process can be common in genomes with high DNA transposon activity. For example, there are thousands of "Pack–MULES" in the rice genome that contain fragments of cellular genes [56] although, like retroposed genes, a very small fraction of these seem to have resulted in or contributed to functional genes [59].

Most segmental duplications, especially the larger ones that might contain one or more complete genes, do not arise by transposition. Instead, they are the result of recombination events. Even here transposable elements play an indirect but very significant role. The excision of a DNA transposon, and the integration of a retrotransposon, both create temporary double-strand DNA breaks; such breaks frequently trigger recombination. The great increase in both homologous and non-homologous recombination due to DNA transposon activity has been well documented [9]. Furthermore, the interspersed copies of transposable elements are a common cause of misalignment during meiosis, resulting in a deletion in one daughter chromosome, and a (generally tandem) duplication in the other one. One of the first well-characterized gene duplications, that of the γ-globin gene in the ancestor of hominids and old-world monkeys, was mediated by two LINE1 copies flanking the original gene [37]. Comparison of the human and chimp genome revealed that in the short time (6 million years) of separation of these species, around a thousand homologous recombination events between interspersed repeats has deleted about 2 Mb of DNA in each lineage [28,51]. Indeed, it has been postulated that the unusually high number of interspersed segmental duplications in the primate genomes, as compared to that of other sequenced organisms, was due to the extraordinarily high activity of the *Alu* retrotransposon about 30–40 million years ago, which left half a million copies of almost identical 300 bp sequences [5].

How Do Duplicated Genes Survive?

The previous section described the many physical mechanisms by which genes may duplicate in a genome, but what forces are there to keep an initially redundant copy of a gene in the genome of a species? We address here two processes by which gene function evolves: Neofunctionalization, and subfunctionalization.

The concept of neofunctionalization is intuitive: Once a new gene copy arises, it is "free" to accumulate mutations and evolve a new, useful function. Given the high rate of accumulation of deleterious mutations, though, it is unlikely that one of the copies will evolve a new function quickly enough. This is especially so in organisms with small population sizes, as the relative strength of adaptive selection over genetic drift is directly related to the population size. Moreover, the presence of a second copy of a gene may even be disadvantageous, as the associated higher level of expression may interfere with proper function. This idea seems corroborated by observations in yeast where genes participating in a protein complex are underrepresented among segmental duplications and members of interacting pairs tend to be co-duplicated [91]. In the case of polyploidy (whole genome duplication), most genes are expressed at an increased level; the loss of individual genes could then lead to imbalances. Indeed, duplicates arising from whole genome duplication have a much longer half-life than those arising from segmental duplications. In vertebrates this difference is about 20 to 25-fold (2 million vs. 45 million years). The longer half-life also reflects another advantage of whole genome duplication, i. e. the preservation of all regulatory sequences, even long distance ones. Considering the ease and frequency with which plants become polyploid, the primary mechanism of chromosomal repatterning in plants may be duplication followed by random loss of component genes.

When a polymorphic gene gives a selective advantage to heterozygotes as compared to the homozygotes of either allele, the gene is said to evolve under balancing selection. Such genes would conceivably be more adept at duplicating, as a genome with both alleles at different loci would have the inherent benefit of "heterozygosity". This mechanism may be at the core of the expansion of gene families greatly benefiting from diversity, like those involved in reproduction, immunity, host defense and chemoreception.

The most feasible route through which duplicated genes may be permanently retained is through the alternative process of subfunctionalization (Fig. 2). When a duplicated gene contains coding or regulating regions that are only important for a subset of its activities (be it in time, space, or function), loss of different functionalities

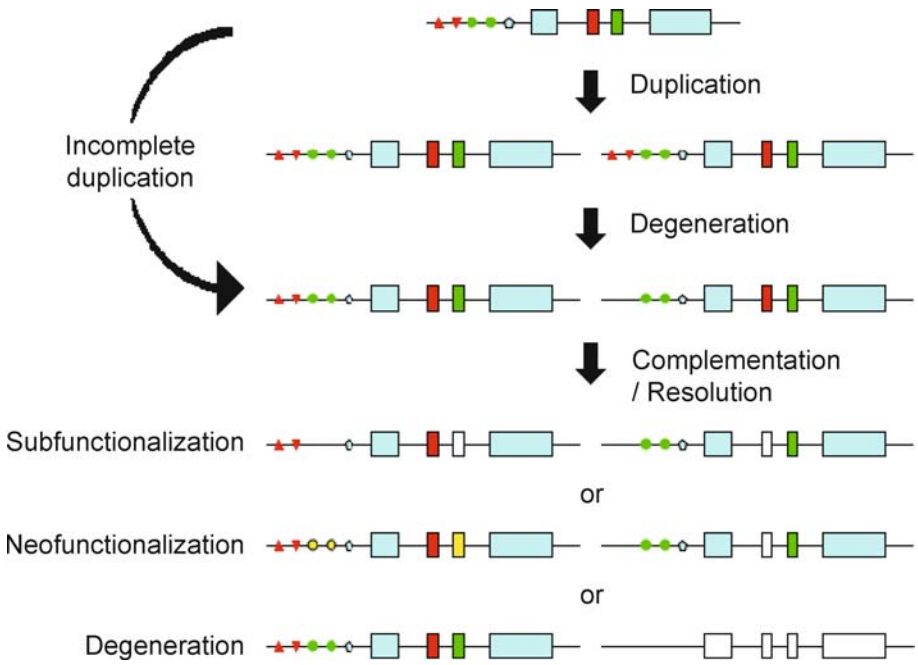

Genome Organization, Figure 2

The Duplication-Degeneration-Complementation (DDC) model for gene duplications. The ancestral gene has two (or more) subfunctions, for each of which independently mutable elements exist. In this particular example, both promoter elements and alternative internal exons exist that are only required for one of two subfunctions. The loss of any one of these regions in one copy of the duplicated gene (which can happen in one or two steps) can be followed by the loss of an independent element in the other copy, after which the copies will be complementary and may both survive. Alternatively though more rarely, the intact copy acquires a new function for the subfunction retained in the damaged copy, after which both copies also may survive. Third, as the damaged copy is still redundant it may simply further decay into a pseudogene. Population genetics predict that the DDC model is both likely, as it depends on common degenerative mutations, and is more apt to occur in small-population species, which tend to have more complex genomes. Figure adapted from Force et al. [39]

in the duplicates can cause the two copies to complement each other, and hence favor the conservation of both in the genome. This Duplication-Degeneration-Complementation (DDC) model [39] is plausible as it invokes common degenerative mutations rather than the exceedingly rare advantageous mutations necessary for neofunctionalization. Subfunctionalization leads to elimination of pleiotropic constraints (i. e. changes cannot interfere with any of the functionalities), so that each duplicate can become optimized to its particular subfunction. This process intuitively adds to increased complexity and flexibility, creating two experts from a single jack-of-all-trades.

An interesting and important prediction is that, through DDC, gene duplications are more likely to become fixated in organisms with small populations sizes (which we observe to have more complex genomes), because each step in the process needs to become fixated in the population before the next step occurs, and the time it takes for neutral mutations to reach fixation is proportional to the population size. In addition, the likelihood of this model does not suffer and may even benefit from incomplete gene duplications, which in segmental duplications may be the rule rather than the exception; absence of, for example, a set of upstream regulatory regions combines the duplication and first degeneration step, creating one copy that already misses a subfunction. For genes that arise via transposition, an alternative way in which the transposed copy may lose some of its functionality upon arrival is through differential epigenetic regulation of the locus it inserts into.

The mechanism by which duplications are fixated will depend on the nature of the gene. While the likelihood of duplication through subfunctionalization increases with greater regulatory and functional complexity of a gene, balancing selection-driven neofunctionalization may be prevalent in the expansion of fast evolving genes like those involved in immune response, reproduction, and xenobiotic recognition (e. g. drug detoxification and odorant

recognition). The olfactory receptor gene family offers a striking example of fast expansions and contractions in vertebrates [44].

Other examples of rapidly expanding gene sets can be found in the nearly 6% of the human euchromatin that is derived from recent (< 40 million years old) segmental duplications. While such duplications are overrepresented in pericentromeric and subtelomeric regions, in hominoid genomes most are actually interspersed through the genome. These interstitial duplicated units (duplicons) are significantly enriched in genes and gene fragments. Moreover, the gene parts of these duplicons are evolving rapidly (most nucleotide substitutions have affected the amino acid sequences) and are more commonly retained in duplicon copies than other regions, suggesting that positive selection has acted not only to diversify sequences but also to fixate gene duplicates in the population [4]. Since the location of interstitial segmental duplications is strongly correlated with unstable sites, Jiang et al. speculated that the association of a gene with such a site was a "predisposing event", opening the way for amplification of this gene in different genomic environments [57].

Transcriptome Complexity

Given the small variability in the number of genes in eukaryotic genomes (and that this number is not that much larger than the respective number in prokaryotes), how does the complex phenotypic diversity of eukaryotes arise? The answer appears to be two-fold: Eukaryotes have the ability of producing many different products from each gene, and have many control mechanisms for regulating and fine-tuning gene expression [83].

The basic mechanism of gene expression involves the recognition, by a set of proteins called the transcriptional complex, of a special DNA sequence called the promoter. The promoter of a gene includes several sequence elements that specify under which conditions the gene should be expressed. The transcriptional complex initiates the copying of the DNA template into an RNA molecule by an enzyme called the RNA polymerase. This process continues until the RNA polymerase encounters a termination signal. The resulting RNA molecule is then further processed, typically leading to the production of a protein molecule.

Prokaryotic genes are easily identified by the presence of an open reading frame (ORF): A contiguous segment of the transcript without any termination codons in (at least) one of the three reading frames in the forward strand. This ORF can be readily translated into a protein sequence, a task carried out in vivo by the translation machinery called the ribosome. Sequences outside the ORF are called untranslated regions (UTRs). A transcript corresponding to a single gene therefore has three well-defined regions: The 5′ UTR, the ORF and the 3′ UTR, delimited by the initiation codon and the termination codon.

In contrast, three decades ago the surprising discovery was made that eukaryotic genes have discontinuous ORFs [11]. For the vast majority of eukaryotic genes, one or more introns may split the ORF and the 5′ UTR into exons (Fig. 3). The boundaries between exons and introns are demarcated by splicing signals: Each intron starts at a "splicing acceptor" and ends at a "splicing donor".

Most of the information required for splicing resides in the introns, though splicing enhancers have been described, that serve for exon definition. Specialized machinery (called the spliceosome) processes the initial (pre-mRNA, or heterogeneous nuclear RNA, hnRNA) transcripts, removing the introns and splicing the exons together into the mature mRNA. Spliceosomes are complex molecular machines involving five small nuclear RNAs (snRNAs) and hundreds of proteins [119]. A large number of genes encode proteins involved in splicing, amounting to approximately 1% of all human genes. Surprisingly, two different kinds of spliceosomes exist: The major and the minor spliceosomes [107,127], which employ different snRNAs (called U2 and U12, respectively) and have different intron specificities. Based on their wide phylogenetic distribution, it is a safe assumption that both types of introns were present in the "stem eukaryote", the ancestor of all extant eukaryotic species (Fig. 1). It is not clear how introns originally arose in the genome, though an origin from mobile elements is probable (see above). As with any other elaboration on the gene structure, their presence increases the risk that the host gene will be rendered defective by subsequent mutations [82]. Nevertheless, once established, the splicing machinery conferred significant advantages to the eukaryotic cell, discussed next.

A crucial aspect of pre-mRNA splicing is that it is not deterministic, and it does not always lead to the same final mRNA form. By a variety of pathways, alternative spliced forms may arise. Alternative transcripts can produce protein products with altered cellular localizations, varying substrate affinities, or different activities [133], in many gene families, e. g. transcription factors [118], kinases and phosphatases [40], kallikreins [70], etc.

Alternative splicing appears to be the norm, particularly among vertebrates [65]: A genome-wide survey using microarrays showed that over 74% of human multi-exon genes are alternatively spliced [58]. In the plant kingdom, alternative splicing is prevalent in flowering plants [97].

There are several different mechanisms that give rise to alternative splice forms from one promoter (Fig. 4), with

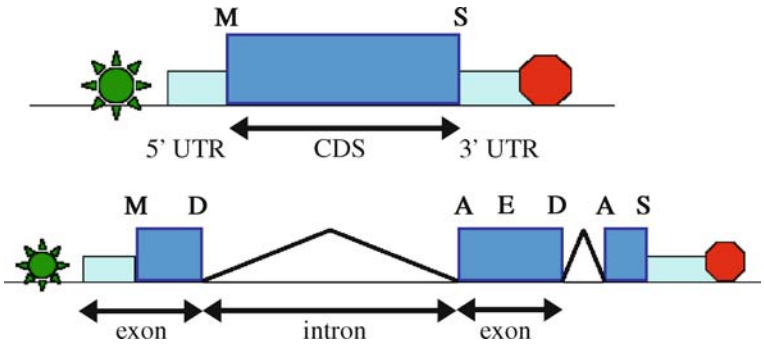

Genome Organization, Figure 3

Schematic of gene structure. *Top*: The basic structure of an intronless gene, with the open reading frame coding region (CDS) flanked by 5′ and 3′ untranslated regions (UTRs). The *star* and the *stop sign* denote the promoter and the poly-adenylation signal, respectively; M and S represent the initiation (methionine) and termination (stop) codons. *Bottom*: The structure of a eukaryotic gene split into exons and introns. D, A and E represent splicing donor, acceptor and enhancer signals, respectively

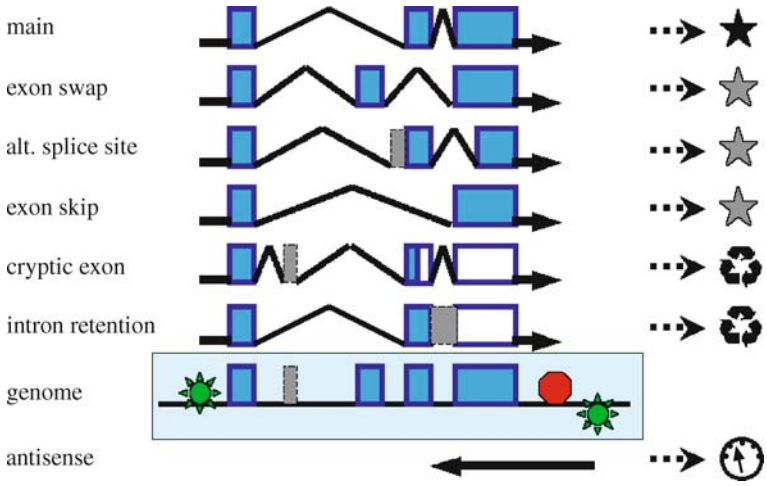

Genome Organization, Figure 4

The different transcripts arising from a genomic locus. The boxed area represents the genomic sequence. From top to bottom: The main transcript, splicing the main exons and giving rise to the main product (*black star*); a secondary transcript in which one exon was swapped with an alternative exon, giving rise to an alternative product (*gray star*); a secondary transcript employing alternative splice sites, be it by extending into an intron or by losing part of an exon; a secondary transcript skipping an exon; a transcript splicing in a cryptic "cassette" exon; a transcript in which an intron is retained instead of being spliced out; an antisense transcript from a downstream promoter, not coding for a protein but functioning in regulation of the main transcripts (*gauge symbol*). Cryptic exons and intron retention typically (but not always) change the reading frame and give rise to an unusable product (*recycling symbol*). Other graphic elements are as in Fig. 3

very different degrees of prevalence [65]. In exon swapping, an exon is replaced by an alternative exon: Such alternative exons typically arise by local duplication. An exon may have alternative splice sites, delimiting a sequence that is sometimes part of the exon, sometimes spliced out with the neighboring intron. Exon skipping refers to the situation in which an exon is spliced out of the transcript with its flanking introns. Conversely, a "cassette" exon may be included in the transcript: These cryptic exons are typically composed of interspersed repeats and other non-coding intergenic sequence, flanked by sequences similar enough to splice sites to be recognized by the spliceosome [114]. Finally, an intron may be retained, bridging its two flanking exons in the final transcript form (Fig. 4).

In multi-exon genes, more than one exon may be alternatively spliced. This can lead to a combinatorial explosion, potentially giving rise to a very large number of

different transcripts from a single locus. Given its potential to produce very different proteins, or even unusable products [116], it is not surprising that alternative splicing is a highly regulated process, involving many proteins and RNA molecules [111]. Central to this process is the family of the serine- and arginine-rich proteins (SR proteins). SR proteins interact with RNA regulatory sequences, and participate in the spliceosome assembly process [17,96]. Taking advantage of the large amounts of sequence information being produced, as well as in vivo splicing reporter systems [125], global analysis approaches are helping elucidate the "cellular code" of splicing [85]. The regulatory sequences include splicing enhancer and silencer elements present in the pre-mRNA, particularly in exons [21,134], but also conserved sequences in introns [113]. There are two main modes of pre-mRNA splicing, involving either the identification of short exons, typically separated by long introns [10,72], or the recognition of short introns [80]. Both modes use the same consensus splice signals [117].

An additional way to achieve a larger variety of alternative transcripts for a gene is by the use of alternative promoters. Many genes are known to have more than one promoter region, each giving rise to its own, unique first exon. These alternative first exons can then be joined with a shared set of exons downstream, adding to the combinatorial capability of transcript production from a single gene locus [79]. Alternatively, the first exon of one gene structure can be an internal exon of another one [26]. Interestingly, transcript elongation and splicing are interrelated in mechanism and regulation, with promoter choice affecting splicing [68]. In turn, the strength of splice sites can affect the choice of transcript termination point, which can lead to the generation of fusion transcripts combining exons from adjacent genes [41].

Several of the mechanisms for alternative splicing produce mature transcripts that include sections of introns, a process called exonization [112]. Such intronic sequences, which were not selected to code for protein sequences, clearly have the potential to interrupt the open reading frame by introducing a premature stop codon. In some cases [77] a proper open reading frame can be restored via a process called RNA editing. The family of adenosine deaminases acting on RNA (ADARs) are able to deaminate adenosines in any long double-stranded RNA, converting them to inosines [64]. Site-specific adenosine deamination can not only correct frame disruptions, but also increase the diversity of transcripts and the resulting proteins [47].

The final quality control of transcripts is effected via a pathway called nonsense-mediated decay (NMD), a surveillance system that targets transcripts with premature termination codons [8,103], as well as transcripts with overly extended 3′ UTRs [87]. This pathway prevents the production of truncated proteins, which may be detrimental to the cell. Interestingly, NMD is also part of regulatory mechanisms of gene expression [30] particularly of splicing regulators like SR proteins, in which alternatively spliced "poison exons" located in ultraconserved elements introduce unavoidable stop codons [73,88].

Are the products of alternative splicing functional, or are they mostly "noise"? Computational analyses indicate that alternative splicing is regulated in a very tissue-specific manner [130] and suggest that this regulation is strongly modified in cancers [108]. As described above, alternative splicing has profound effects on the structure of the resulting proteome [116]. Taken together, these and other lines of evidence suggest that splicing variants are as a rule functional and actively selected for. In contrast, a statistical comparison of conserved vs. lineage-specific alternatively spliced exons indicates that these are two distinct classes, suggesting that many splice forms observed and cataloged in sequence databases may be aberrant and non-functional [115]. Further evidence from analyses of amino acid frequencies suggest relaxed selection on alternative splice forms [25].

Transcript Networks

In the previous section we discussed how a variety of different transcripts can arise from a single "gene" and some aspects of how this process is regulated. The classical view of gene function involves the transmission of information from the genome to the protein via a transcript. During the last few years the perception of RNA as a mere information conduit has changed dramatically, and it is now known that this is not the only way an RNA transcript can be functional. RNA molecules can be structural (e. g. ribosomal RNA, transfer RNA) and catalytic, but most importantly, they provide a complex intermediary network of regulation, between DNA-level transcript production and protein-level interaction networks.

The next level of genomic complexity therefore arises from the interaction between transcripts. In fact, we can conceptually subdivide transcripts into three types (Fig. 5), two of which are heavily invested in regulation at the RNA level. Type I describes the classical gene structure in which the mRNA (after splicing out the introns if any were present) serves as the intermediary form leading to the production of the functional product – the protein. In Type II, the (potentially spliced) mRNA transcript itself represents the functional product, without

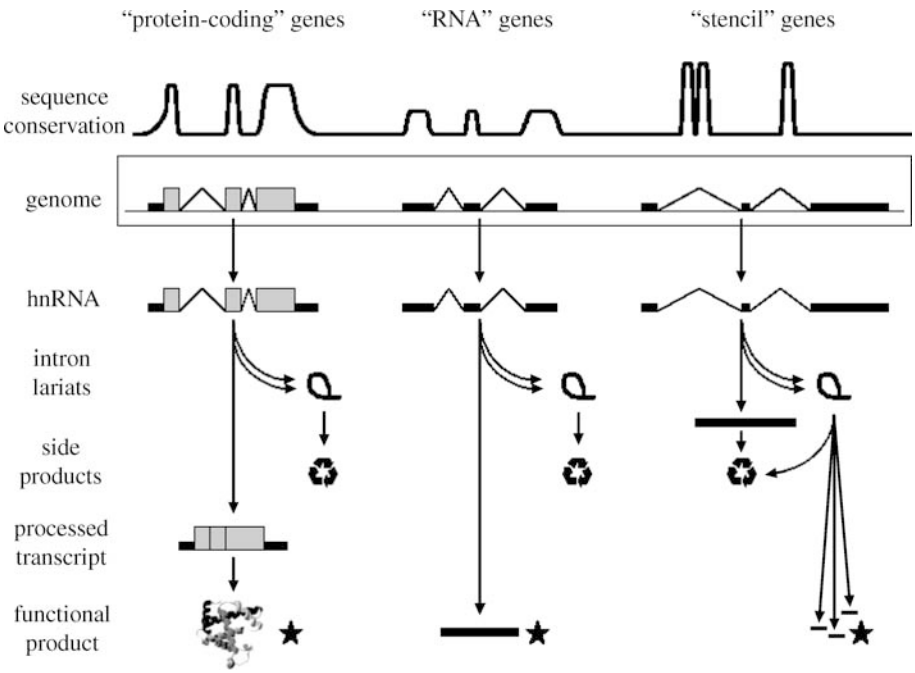

Genome Organization, Figure 5

Schematic representation of the three types of eukaryotic transcripts, as classified by product type: Type I "protein-coding" genes (*left*), Type II "RNA" genes (*center*), and Type III "stencil" genes (*right*). The boxed area represents the genomic sequence, gray boxes indicate coding exons, thick horizontal lines denote UTRs. In protein-coding genes, the introns are spliced out in lariat form and degraded (*recycling symbol*), while the processed transcript gives rise to the functional protein product (*black star*). In RNA genes, the processed transcript is the same as the functional product, without translation to protein. In stencil genes, the processed transcript is meaningless and degraded, while the intron lariats are further processed to produce the functional products – small RNA molecules like miRNAs and snoRNAs. The "sequence conservation" graph reveals the location of the functional elements, extending throughout the exons of protein-coding and RNA genes, but localized in the introns of stencil genes

coding for a protein [50]. Such RNA genes are involved in a large number of cellular processes, including gene regulation [106], and more are being discovered. Type III transcripts differ from the first two types in that, after splicing out of the introns, the resulting spliced mRNA is not the source of the functional product. Instead, specific pathways excise out of the intron lariats short RNA segments including miRNAs (see below) and snoRNAs. We propose to call this third type "stencil genes" because, in similarity to stencils, they are just structural frameworks, while the sections removed from them are those that carry the relevant information. Stencil genes range in size from the very small, e. g. the 5 kb long C6orf48 transcript in the human MHC Class III region [128], to the very large, e. g. the 530 kb long LOC401237 transcript [46]. Interestingly, the three types of transcripts are not mutually exclusive. For example, most snoRNA genes are produced from intron lariats spliced out of transcripts coding for riboproteins [50] and most human miRNA genes are derived from intronic regions [67,101].

The ability of nucleic acids to recognize each other by strand complementarity has afforded several opportunities for the evolution of specific regulatory mechanisms. A large number of genes are now known to be regulated by naturally occurring antisense transcripts [75,131], which can function in *cis* or in *trans* [76]. First identified as specific transcriptional repression by injection of double-stranded RNA in *C. elegans* [36], or RNA interference (RNAi), a more exquisite mechanism for general gene regulation involves the production of short ($\sim$ 22 nt long) RNA molecules called micro RNAs, or miRNAs [90] by cleavage from RNA hairpins. miRNAs form complexes with several proteins including Argonaute gene family members, and the resulting miRNP complexes control gene expression by binding to mRNA, frequently in the 3′ UTR, recognizing sequence motifs by imperfect sequence complementarity [6]. An mRNA thus recognized by a miRNP complex will be either cleaved and degraded, or its translation to protein will be repressed. The identification of miRNAs and their targets is a very active

area of research, with knowledge being warehoused in the miRBase database (Table 1). Related to miRNAs are small interfering RNAs, or siRNAs, which are produced by cleavage of long double-stranded RNAs. Finally, a novel type of small RNAs was recently described, which interact with Piwi proteins, a subfamily of Argonaute proteins. These Piwi-interacting RNAs, or piRNAs, are expressed mainly in the germline and, as described above, are part of an adaptive defense system against transposable elements [3,66].

Computational analyses suggested that a large fraction of the genome is transcribed [46,104], including many regions previously thought to be devoid of genes (and therefore called "gene deserts"). Experimental analyses demonstrated vast networks of overlapping transcripts on both DNA strands, called "gene forests" [12,20,27,61]. Not surprisingly, some transcripts are only observed in specific cell types and/or conditions [100], suggesting that many additional functional transcript forms are yet to be discovered. The observed complexity of transcription and regulation poses a significant challenge to the classical concept of the gene as the discrete element in the genome, and requires updated definitions [43].

If most of both genomic DNA strands are transcribed in complex loci, with many sequences being multifunctional [12,62], what fraction of the genome is actively selected for functional needs? It has been claimed that everything in the genome is functional [123] to different degrees and in various circumstances, or even in aggregate for "genomic bulk" [135]. On the other hand, it has been proposed that a large fraction of the current genomic features has arisen in a "mostly neutral" fashion [84]. In this context, it is important to keep in mind that the typically small population size of eukaryotic species precludes effective selection over most functional sites, particularly those conferring weak advantageous effects.

Future Directions

Genomes have evolved over many hundred million years, by a haphazard accumulation of random mutations ranging from the single nucleotide change to the duplication of the entire genome. In prokaryotes, the large population sizes and fast division times afford for very effective selection against deleterious or wasteful mutations, leading to compact and streamlined genomes. In sharp contrast, given the small population sizes of eukaryotic species, the selective advantage of some variants over others has been tempered by the vagaries of random sampling. This ongoing process has led to the accretion of the highly disorganized genomes observed in eukaryotes. One of the most fundamental and understated results from the analysis of completed genomic sequences is the overwhelming and outstanding lack of evidence for any kind of intelligent design in genome organization (Fig. 6). Even though exceptions exist, in the form of locally coordinated, small clusters of functionally related genes [7], the vast majority of the vertebrate genome is a disordered array of genes in essentially random order and orientation. As genomes evolve, gene order changes following a simple null model of random mutation, with present intergenic distance being the best predictor for future gene linkage [94].

In this context, one major difficulty lies in distinguishing between real biological signals and the background of random variation. We need to understand how the genome is organized and how it evolved, to be able to build suitable null hypotheses for testing whether predictions are likely to be real, and whether specific observations are likely to be biologically meaningful. To surmount this difficulty, the integration of many data types is imperative, as demonstrated by the ENCODE project (Table 1) [12]. Comparative genomics is a powerful tool for distinguishing between evolutionary and functional signal and noise, by identifying what is shared and what is different among related species. In the field of gene prediction, the best results are obtained by integrating several sources of information, reflecting many steps of the process of gene expression, from the genome to the functional product (Fig. 7). Computational and experimental work has shown that many regions of the genome are transcribed in which no classical protein-coding genes can be identified. This suggests that many RNA and stencil genes might remain to be discovered [46].

A deep understanding of genome organization and principles of function will enable the new field of genome engineering. Preliminary work has demonstrated the feasibility of combining two bacterial genomes into one [55] and full genome transplantation [74]. Interestingly, the resulting phenotypes observed either reflected the host genome when combined, or the donor genome when replaced. This suggests that researchers will benefit from a wide range of possible phenotypic effects, to be attained by judicious combinations of genomic sources, both natural and synthetic. At a more modest scale, it will be possible to engineer the minimal genetic modifications required to correct the imbalances and deregulations leading to many diseases, including cancer.

The ongoing development of novel sequencing technologies has reduced the cost of sequencing genomes by orders of magnitude. It is expected that within a few years, obtaining an individual's version of the genome will be extremely affordable, which is one of the essential re-

	non-system	evolved-system	designed-system
purpose	none	survival behavior	purposeful
redundancy	redundant	highly redundant	parsimonious
parts	arbitrary parts	multifunctional parts defunct parts	no extra parts
efficiency	none	wasteful	efficient
response to change	accepts all change	error correction (high turnover) or death	changes outside specification range are catastrophic
examples	mountain river	metazoa bacteria viruses	technology new old
organization	chaotic	messy	organized

Genome Organization, Figure 6

A conceptual comparison between non-systems (random assemblages of objects, e. g. a pile of rocks), living systems (evolved) and technological systems (designed). While designed systems are presumably purposeful and hopefully efficient and parsimonious, non-systems have no purpose – they just are – and as such have no efficiency. The survival behavior of evolved systems gives a functional purpose to their parts, which are typically redundant and multifunctional, or even not used anymore. As such, evolved systems are highly "wasteful". Designed systems typically have specification ranges within which they function normally, and changes outside those predefined ranges are typically catastrophic. Non-systems accept all change and return to passive equilibrium. Evolved systems typically have many methods of error correction that accumulated over many millions of years of selection in changing environments. Overall, non-systems are chaotic and designed systems are well organized, while evolved systems are very messy and disorganized, typically displaying ancient battle scars, relics or defunct parts, and the minimal level of organization required for survival

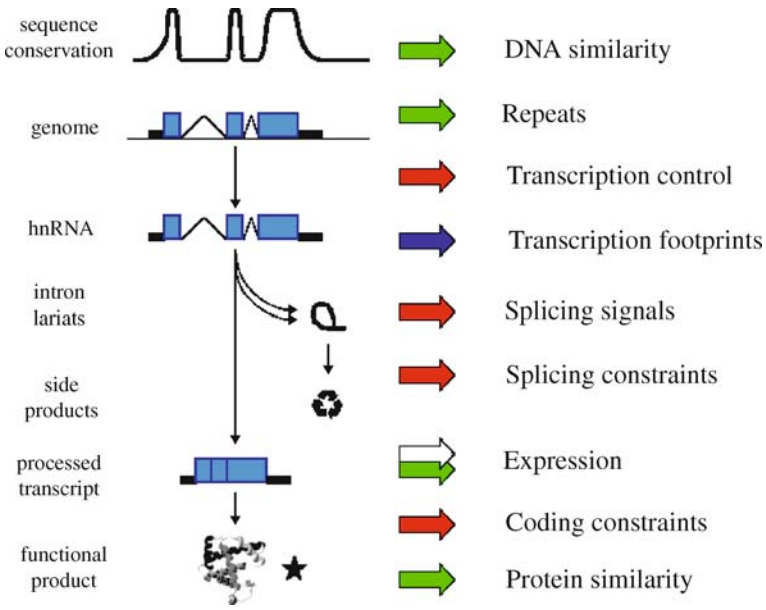

Genome Organization, Figure 7

Sources of information used for predicting and recognizing genes in the genome. *Red arrows*: Ab initio gene prediction methods. *Green arrows*: Similarity-based gene prediction methods. *Blue arrow*: Footprint-based transcript prediction. Observed transcripts are used dually to identify genes (*white arrow*) and to predict similar genes (*green arrow*). Other graphical elements as in Figs. 3 and 4

quirements for personalized medicine [14]. Already, the elucidation of an individual's genome [78] provided insight into the level of polymorphism arising from different types of mutations, and indicated that alterations involving more than a single base pair represent a sizable fraction of the variability among alleles in the human population. As more personal genome sequences become available through the Personal Genome Project (Table 1) and other efforts, it will be possible to obtain a detailed characterization of mutation frequencies, on which to base proper models of neutral evolution. Based on the cumulative results of population sequencing, it will be possible to derive robust "reference" genomes for human populations, leading to a highly detailed characterization of the history of human evolution. The genetic analysis of complex diseases will be strongly impacted by the cumulative genomic knowledge, leading to a comprehensive identification of the genes associated with most major diseases, and a detailed characterization of their interactions.

Bibliography

Primary Literature

1. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD et al (2000) The genome sequence of Drosophila melanogaster. Science 287:2185–2195
2. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM et al (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297:1301–1310
3. Aravin AA, Hannon GJ, Brennecke J (2007) The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science 318:761–764
4. Bailey JA, Eichler EE (2006) Primate segmental duplications: Crucibles of evolution, diversity and disease. Nat Rev Genet 7:552–564
5. Bailey JA, Liu G, Eichler EE (2003) An Alu transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet 73:823–834
6. Bartel DP (2004) MicroRNAs: Genomics, biogenesis, mechanism, and function. Cell 116:281–297
7. Batada NN, Hurst LD (2007) Evolution of chromosome organization driven by selection for reduced gene expression noise. Nat Genet 39:945–949
8. Behm-Ansmant I, Kashima I, Rehwinkel J, Sauliere J, Wittkopp N et al (2007) mRNA quality control: An ancient machinery recognizes and degrades mRNAs with nonsense codons. FEBS Lett 581:2845–2853
9. Bennetzen JL (2007) Patterns in grass genome evolution. Curr Opin Plant Biol 10:176–181
10. Berget SM (1995) Exon recognition in vertebrate splicing. J Biol Chem 270:2411–2414
11. Berget SM, Moore C, Sharp PA (1977) Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci USA 74:3171–3175
12. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR et al (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799–816
13. Blasco MA (2007) Telomere length, stem cells and aging. Nat Chem Biol 3:640–649
14. Blow N (2007) Genomics: The personal side of genomics. Nature 449:627–630
15. Bolzer A, Kreth G, Solovei I, Koehler D, Saracoglu K et al (2005) Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. PLoS Biol 3:e157
16. Bourc'his D, Bestor T (2004) Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature 431:96–99
17. Bourgeois CF, Lejeune F, Stevenin J (2004) Broad specificity of SR (serine/arginine) proteins in the regulation of alternative splicing of pre-messenger RNA. Prog Nucleic Acid Res Mol Biol 78:37–88
18. Bourque G, Zdobnov EM, Bork P, Pevzner PA, Tesler G (2005) Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. Genome Res 15:98–110
19. *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: A platform for investigating biology. Science 282:2012–2018
20. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC et al (2005) The transcriptional landscape of the mammalian genome. Science 309:1559–1563
21. Cartegni L, Chew SL, Krainer AR (2002) Listening to silence and understanding nonsense: Exonic mutations that affect splicing. Nat Rev Genet 3:285–298
22. Casola C, Hucks D, Feschotte C (2008) Convergent domestication of pogo-Like Transposases into centromere-binding proteins in fission yeast and mammals. Mol Biol Evol 25(1):29–41
23. Cavalier-Smith T (1978) Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNAC-value paradox. J Cell Sci 34:247–278
24. Cavalier-Smith T, Beaton MJ (1999) The skeletal function of non-genic nuclear DNA: New evidence from ancient cell chimaeras. Genetica 106:3–13
25. Chen FC, Wang SS, Chen CJ, Li WH, Chuang TJ (2006) Alternatively and constitutively spliced exons are subject to different evolutionary forces. Mol Biol Evol 23:675–682
26. Chen WH, Lv G, Lv C, Zeng C, Hu S (2007) Systematic analysis of alternative first exons in plant genomes. BMC Plant Biol 7:55
27. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S et al (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308:1149–1154
28. Chimpanzee Genome Sequencing Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87
29. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B et al (2007) Evolution of genes and genomes on the drosophila phylogeny. Nature 450:203–218
30. Conti E, Izaurralde E (2005) Nonsense-mediated mRNA decay: Molecular insights and mechanistic variations across species. Curr Opin Cell Biol 17:316–325
31. Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284:601–603

32. Duret L, Eyre-Walker A, Galtier N (2006) A new perspective on isochore evolution. Gene 385:71–74
33. Egan ES, Fogel MA, Waldor MK (2005) Divided genomes: Negotiating the cell cycle in prokaryotes with multiple chromosomes. Mol Microbiol 56:1129–1138
34. Esnault C, Heidmann O, Delebecque F, Dewannieux M, Ribet D et al (2005) APOBEC3G cytidine deaminase inhibits retrotransposition of endogenous retroviruses. Nature 433: 430–433
35. Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D et al (1976) Complete nucleotide sequence of bacteriophage MS2 RNA: Primary and secondary structure of the replicase gene. Nature 260:500–507
36. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE et al (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature 391:806–811
37. Fitch DH, Bailey WJ, Tagle DA, Goodman M, Sieu L et al (1991) Duplication of the gamma-globin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates. Proc Natl Acad Sci USA 88:7396–7400
38. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF et al (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269:496–512
39. Force A, Lynch M, Pickett FB, Amores A, Yan YL et al (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545
40. Forrest AR, Taylor DF, Crowe ML, Chalk AM, Waddell NJ et al (2006) Genome-wide review of transcriptional complexity in mouse protein kinases and phosphatases. Genome Biol 7:R5
41. Frith MC, Carninci P, Kai C, Kawai J, Bailey TL et al (2007) Splicing bypasses 3′ end formation signals to allow complex gene architectures. Gene 403:188–193
42. Galagan J, Selker E (2004) RIP: The evolutionary cost of genome defense. Trends Genet 20:417–423
43. Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J et al (2007) What is a gene, post-ENCODE? History and updated definition. Genome Res 17:669–681
44. Gilad Y, Man O, Glusman G (2005) A comparison of the human and chimpanzee olfactory receptor gene repertoires. Genome Res 15:224–230
45. Gladyshev EA, Arkhipova IR (2007) Telomere-associated endonuclease-deficient Penelope-like retroelements in diverse eukaryotes. Proc Natl Acad Sci USA 104:9352–9357
46. Glusman G, Qin S, El-Gewely MR, Siegel AF, Roach JC et al (2006) A third approach to gene prediction suggests thousands of additional human transcribed regions. PLoS Comput Biol 2:e18
47. Gott JM (2003) Expanding genome capacity via RNA editing. C R Biol 326:901–908
48. Gregory TR (2001) Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol Rev Camb Philos Soc 76:65–101
49. Gregory TR (2002) Genome size and developmental complexity. Genetica 115:131–146
50. Griffiths-Jones S (2007) Annotating noncoding RNA genes. Annu Rev Genomics Hum Genet 8:279–298
51. Han K, Lee J, Meyer TJ, Wang J, Sen SK et al (2007) Alu recombination-mediated structural deletions in the chimpanzee genome. PLoS Genet 3:1939–1949
52. Hedges SB, Chen H, Kumar S, Wang DY, Thompson AS et al (2001) A genomic timescale for the origin of eukaryotes. BMC Evol Biol 1:4
53. Holmes R, Malim M, Bishop K (2007) APOBEC-mediated viral restriction: Not simply editing? Trends Biochem Sci 32:118–128
54. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921
55. Itaya M, Tsuge K, Koizumi M, Fujita K (2005) Combining two genomes in one cell: Stable cloning of the Synechocystis PCC6803 genome in the bacillus subtilis 168 genome. Proc Natl Acad Sci USA 102:15971–15976
56. Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR (2004) Pack-MULE transposable elements mediate gene evolution in plants. Nature 431:569–573
57. Jiang Z, Tang H, Ventura M, Cardone MF, Marques-Bonet T et al (2007) Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. Nat Genet 39:1361–1368
58. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM et al (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 302:2141–2144
59. Juretic N, Hoen DR, Huynh ML, Harrison PM, Bureau TE (2005) The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. Genome Res 15:1292–1297
60. Kapitonov VV, Jurka J (2005) RAG1 core and V(D)J recombination signal sequences were derived from transib transposons. PLoS Biol 3:e181
61. Kapranov P, Drenkow J, Cheng J, Long J, Helt G et al (2005) Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. Genome Res 15:987–997
62. Kapranov P, Willingham AT, Gingeras TR (2007) Genome-wide transcription and the implications for genomic organization. Nat Rev Genet 8:413–423
63. Kasahara M (2007) The 2R hypothesis: An update. Curr Opin Immunol 19:547–552
64. Keegan LP, Leroy A, Sproul D, O'Connell MA (2004) Adenosine deaminases acting on RNA (ADARs): RNA-editing enzymes. Genome Biol 5:209
65. Kim E, Magen A, Ast G (2007) Different levels of alternative splicing among eukaryotes. Nucleic Acids Res 35:125–131
66. Kim VN (2006) Small RNAs just got bigger: Piwi-interacting RNAs (piRNAs) in mammalian testes. Genes Dev 20:1993–197
67. Kim YK, Kim VN (2007) Processing of intronic microRNAs. Embo J 26:775–783
68. Kornblihtt AR (2005) Promoter usage and alternative splicing. Curr Opin Cell Biol 17:262–268
69. Kouzine F, Levens D (2007) Supercoil-driven DNA structures regulate genetic transactions. Front Biosci 12:4409–4423
70. Kurlender L, Borgono C, Michael IP, Obiezu C, Elliott MB et al (2005) A survey of alternative transcripts of human tissue kallikrein genes. Biochim Biophys Acta 1755:1–14
71. Lai J, Li Y, Messing J, Dooner HK (2005) Gene movement by Helitron transposons contributes to the haplotype variability of maize. Proc Natl Acad Sci USA 102:9068–9073
72. Lam BJ, Hertel KJ (2002) A general role for splicing enhancers in exon definition. RNA 8:1233–1241
73. Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE (2007) Unproductive splicing of SR genes associated with

highly conserved and ultraconserved DNA elements. Nature 446:926–929
74. Lartigue C, Glass JI, Alperovich N, Pieper R, Parmar PP et al (2007) Genome transplantation in bacteria: Changing one species to another. Science 317:632–638
75. Lavorgna G, Dahary D, Lehner B, Sorek R, Sanderson CM et al (2004) In search of antisense. Trends Biochem Sci 29:88–94
76. Lehner B, Williams G, Campbell RD, Sanderson CM (2002) Antisense transcripts in the human genome. Trends Genet 18:63–65
77. Lev-Maor G, Sorek R, Levanon EY, Paz N, Eisenberg E et al (2007) RNA-editing-mediated exon evolution. Genome Biol 8:R29
78. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL et al (2007) The diploid genome sequence of an individual human. PLoS Biol 5:e254
79. Li Q, Lee JA, Black DL (2007) Neuronal regulation of alternative pre-mRNA splicing. Nat Rev Neurosci 8:819–831
80. Lim LP, Burge CB (2001) A computational analysis of sequence features involved in recognition of short introns. Proc Natl Acad Sci USA 98:11193–11198
81. Lowe CB, Bejerano G, Haussler D (2007) Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proc Natl Acad Sci USA 104:8005–8010
82. Lynch M (2006) Streamlining and simplification of microbial genome architecture. Annu Rev Microbiol 60:327–349
83. Lynch M (2006) The origins of eukaryotic gene structure. Mol Biol Evol 23:450–468
84. Lynch M, Conery JS (2003) The origins of genome complexity. Science 302:1401–1404
85. Matlin AJ, Clark F, Smith CW (2005) Understanding alternative splicing: Towards a cellular code. Nat Rev Mol Cell Biol 6:386–398
86. Mira A, Ochman H, Moran NA (2001) Deletional bias and the evolution of bacterial genomes. Trends Genet 17:589–596
87. Muhlrad D, Parker R (1999) Aberrant mRNAs with extended 3′ UTRs are substrates for rapid degradation by mRNA surveillance. RNA 5:1299–1307
88. Ni JZ, Grate L, Donohue JP, Preston C, Nobida N et al (2007) Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev 21:708–718
89. Nishihara H, Smit AF, Okada N (2006) Functional noncoding sequences derived from SINEs in the mammalian genome. Genome Res 16:864–874
90. Ouellet DL, Perron MP, Gobeil LA, Plante P, Provost P (2006) MicroRNAs in gene regulation: When the smallest governs it all. J Biomed Biotechnol 2006:69616
91. Papp B, Pál C, Hurst LD (2003) Dosage sensitivity and the evolution of gene families in yeast. Nature 424:194–197
92. Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE et al (2004) Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. Dev Cell 7:597–606
93. Pisani D, Cotton JA, McInerney JO (2007) Supertrees disentangle the chimerical origin of eukaryotic genomes. Mol Biol Evol 24:1752–1760
94. Poyatos JF, Hurst LD (2007) The determinants of gene order conservation in yeasts. Genome Biol 8:R233
95. Pritham EJ, Feschotte C (2007) Massive amplification of rolling-circle transposons in the lineage of the bat Myotis lucifugus. Proc Natl Acad Sci USA 104:1895–1900
96. Reddy AS (2004) Plant serine/arginine-rich proteins and their role in pre-mRNA splicing. Trends Plant Sci 9:541–547
97. Reddy AS (2007) Alternative splicing of pre-messenger RNAs in plants in the genomic era. Annu Rev Plant Biol 58:267–294
98. Rippe K (2007) Dynamic organization of the cell nucleus. Curr Opin Genet Dev 17:373–380
99. Rogozin IB, Makarova KS, Natale DA, Spiridonov AN, Tatusov RL et al (2002) Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. Nucleic Acids Res 30:4264–4271
100. Rozowsky J, Wu J, Lian Z, Nagalakshmi U, Korbel JO et al (2006) Novel transcribed regions in the human genome. Cold Spring Harb Symp Quant Biol 71:111–116
101. Ruby JG, Jan CH, Bartel DP (2007) Intronic microRNA precursors that bypass Drosha processing. Nature 448:83–86
102. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR et al (1977) Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265:687–695
103. Schell T, Kulozik AE, Hentze MW (2002) Integration of splicing, transport and translation to achieve mRNA quality control by the nonsense-mediated decay pathway. Genome Biol 3:REVIEWS1006
104. Sémon M, Duret L (2004) Evidence that functional transcription units cover at least half of the human genome. Trends Genet 20:229–232
105. Sémon M, Wolfe K (2007) Consequences of genome duplication. Curr Opin Genet Dev 17:505–512
106. Shamovsky I, Nudler E (2006) Gene control by large noncoding RNAs. Sci STKE 2006:pe40
107. Sharp PA, Burge CB (1997) Classification of introns: U2-type or U12-type. Cell 91:875–879
108. Skotheim RI, Nees M (2007) Alternative splicing in cancer: Noise, functional, or systematic? Int J Biochem Cell Biol 39:1432–1449
109. Smit AF (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev 9:657–663
110. Smit AF, Riggs AD (1996) Tiggers and DNA transposon fossils in the human genome. Proc Natl Acad Sci USA 93:1443–1448
111. Soller M (2006) Pre-messenger RNA processing and its regulation: A genomic perspective. Cell Mol Life Sci 63:796–819
112. Sorek R (2007) The birth of new exons: Mechanisms and evolutionary consequences. RNA 13:1603–1608
113. Sorek R, Ast G (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. Genome Res 13:1631–1637
114. Sorek R, Lev-Maor G, Reznik M, Dagan T, Belinky F et al (2004) Minimal conditions for exonization of intronic sequences: 5′ splice site formation in alu exons. Mol Cell 14:221–231
115. Sorek R, Shamir R, Ast G (2004) How prevalent is functional alternative splicing in the human genome? Trends Genet 20:68–71
116. Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z et al (2005) Function of alternative splicing. Gene 344:1–20
117. Sterner DA, Carlo T, Berget SM (1996) Architectural limits on split genes. Proc Natl Acad Sci USA 93:15081–15085
118. Taneri B, Snyder B, Novoradovsky A, Gaasterland T (2004) Alternative splicing of mouse transcription factors affects

their DNA-binding domain architecture and is tissue specific. Genome Biol 5:R75
119. Valadkhan S (2005) snRNAs as the catalysts of pre-mRNA splicing. Curr Opin Chem Biol 9:603–608
120. Vellai T, Vida G (1999) The origin of eukaryotes: The difference between prokaryotic and eukaryotic cells. Proc Biol Sci 266:1571–1577
121. Vinckenbosch N, Dupanloup I, Kaessmann H (2006) Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci USA 103:3220–3225
122. Vinogradov AE (2004) Evolution of genome size: Multilevel selection, mutation bias or dynamical chaos? Curr Opin Genet Dev 14:620–626
123. von Sternberg R (2002) On the roles of repetitive DNA elements in the context of a unified genomic-epigenetic system. Ann N Y Acad Sci 981:154–188
124. Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR et al (2007) Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. Proc Natl Acad Sci USA 104:18613–18618
125. Wang Z, Rolish ME, Yeo G, Tung V, Mawson M et al (2004) Systematic identification and analysis of exonic splicing silencers. Cell 119:831–845
126. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562
127. Will CL, Luhrmann R (2005) Splicing of a rare class of introns by the U12-dependent spliceosome. Biol Chem 386:713–724
128. Xie T, Rowen L, Aguado B, Ahearn ME, Madan A et al (2003) Analysis of the gene-dense major histocompatibility complex class III region and its comparison to mouse. Genome Res 13:2621–2636
129. Xing J, Wang H, Belancio VP, Cordaux R, Deininger PL et al (2006) Emergence of primate genes by retrotransposon-mediated sequence transduction. Proc Natl Acad Sci USA 103:17608–17613
130. Xu Q, Modrek B, Lee C (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. Nucleic Acids Res 30:3754–3766
131. Yelin R, Dahary D, Sorek R, Levanon EY, Goldstein O et al (2003) Widespread occurrence of antisense transcription in the human genome. Nat Biotechnol 21:379–386
132. Zaidi SK, Young DW, Javed A, Pratap J, Montecino M et al (2007) Nuclear microenvironments in biological control and cancer. Nat Rev Cancer 7:454–463
133. Zavolan M, Kondo S, Schonbach C, Adachi J, Hume DA et al (2003) Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. Genome Res 13:1290–1300
134. Zheng ZM (2004) Regulation of alternative RNA splicing by exon definition and exon sequences in viral and mammalian gene expression. J Biomed Sci 11:278–294
135. Zuckerkandl E (2002) Why so many noncoding nucleotides? The eukaryote genome as an epigenetic machine. Genetica 115:105–129

Books and Reviews

Brown TA (2002) Genomes. Bios Scientific Publishers, Oxford
Dawkins R (1982) The Extended Phenotype. Oxford University Press, Oxford
Hartwell L et al (2006) Genetics: From Genes to Genomes. McGraw-Hill, Boston
Lynch M (2007) The Origins of Genome Architecture. Sinauer, Sunderland
Margulis L, Sagan D (2003) Acquiring Genomes. Basic Books, New York
Watson JD et al (2003) Molecular Biology of the Gene. Cummings

Geo-complexity and Earthquake Prediction

Vladimir Keilis-Borok[1,2], Andrei Gabrielov[3], Alexandre Soloviev[2,4]
[1] Institute of Geophysics and Planetary Physics and Department of Earth and Space Sciences, University of California, Los Angeles, USA
[2] International Institute of Earthquake Prediction Theory and Mathematical Geophysics, Russian Academy of Sciences, Moscow, Russia
[3] Departments of Mathematics and Earth and Atmospheric Sciences, Purdue University, West Lafayette, USA
[4] The Abdus Salam International Center for Theoretical Physics, Trieste, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Lithosphere as a Hierarchical Complex System
General Scheme of Prediction
Four Paradigms
Earthquake Prediction and Earthquake Preparedness
Further Goals
Acknowledgment
Bibliography

Glossary

Chaos Apparently random or unpredictable behavior in systems governed by deterministic laws. The common element in these systems is a very high sensitivity to initial conditions and to the way in which a system is set in motion (Encyclopedia Britannica).

Complexity An attribute of nonlinear (chaotic) systems. It comprises instability and complex but not random behavior patterns – "order in chaos".

Earthquake An episode of rupture and discontinuous displacement within the solid Earth. Part of the energy accumulated around the rupture is released by inelas-

tic deformation and seismic waves. Both may cause destructive shaking of the ground, if the energy release is sufficiently large.

Earthquake forecasting Probabilistic extrapolation of seismic activity comprising many earthquakes.

Earthquake prediction Prediction of time interval, geographic area, and magnitude range where an individual future strong earthquake will occur. The prediction is meaningful if it includes an estimated rate of false alarms.

Earthquake preparedness A set of actions reducing the damage from the future earthquakes. There are different levels of preparedness.

Extreme events Rare events of low probability but high impact on a system where they occur. In different connotations they are also known as critical transitions, disasters, catastrophes, and crises. Over time they persistently recur in both natural and constructed complex systems. In this article the extreme events are the strong earthquakes. An earthquake might be an extreme event in a certain volume of the lithosphere and part of the background seismicity in a larger volume.

Lithosphere The earthquake-prone outer shell of the solid Earth. In prediction research it is regarded as a hierarchical complex system.

Premonitory seismicity patterns Space-time-magnitude patterns of earthquake occurrences that signal the approach of a strong earthquake.

Definition of the Subject

Definition

The problem of earthquake prediction is to find when and where a strong earthquake will occur. A prediction is formulated as a discrete sequence of alarms (Fig. 1). The accuracy of a prediction method is captured by probabilities of errors (false alarms and failures to predict) and by the total space-time occupied by alarms. (Sect. "Error Diagram").

In terms of prediction studies this is algorithmic prediction of individual extreme events having low probability but large impact. This problem is necessarily intertwined with problems of disaster preparedness, dynamics of solid Earth, and modeling of extreme events in hierarchical complex systems.

Predictability ("order in chaos"). Complex systems, lithosphere included, are not predictable with unlimited precision. However, after a coarse-graining (i. e., in a not-too-detailed scale) certain regular behavior patterns emerge and a system becomes predictable, up to certain limits ([13,20,24,26,36,46,52,83]). Accordingly, earthquake prediction requires a holistic analysis, "from the whole to details". Such analysis makes it possible to overcome the geo-complexity itself and the chronic imperfection of observations as well.

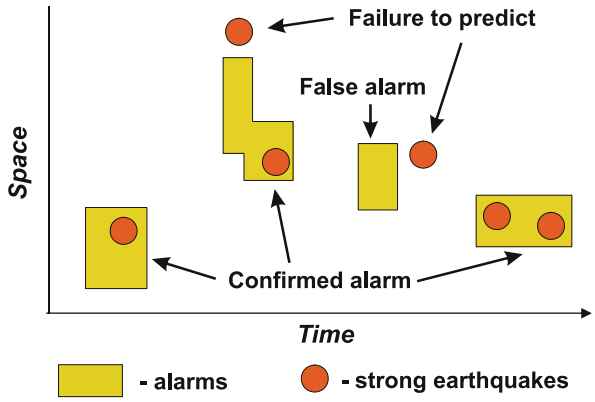

Geo-complexity and Earthquake Prediction, Figure 1
Possible outcomes of prediction

Premonitory patterns. Certain behavior patterns emerge more frequently as a strong earthquake draws near. Called premonitory patterns, they signal destabilization of the earthquake-prone lithosphere and thus an increase in the probability of a strong earthquake. Premonitory patterns do not necessarily contribute to causing a subsequent strong earthquake; both might be parallel manifestations of the same underlying process – the tectonic development of the Earth in multiple time-, space-, and energy- scales. For that reason premonitory patterns might emerge in a broad variety of observable fields reflecting lithosphere dynamics, and in different scales.

The algorithms considered here, based on premonitory seismicity patterns, provide alarms lasting years to months. There is ample evidence that major findings made in developing these algorithms are applicable to premonitory patterns in other fields, to predicting other geological and geotechnical disasters, and probably to determining shorter and longer alarms (Sect. "Further Goals").

Importance

Algorithmic earthquake prediction provides pivotal constraints for fundamental understanding of the dynamics of the lithosphere and other complex systems. It is also critically important for protecting the global population, economy, and environment. Vulnerability of our world to the earthquakes is rapidly growing, due to proliferation of high-risk construction (nuclear power plants, high dams, radioactive waste disposals, lifelines, etc.), deterioration of ground and infrastructure in megacities, destabilization

of environment, population growth, and escalating socio-economic volatility of the global village. Today a single earthquake with its ripple effects may take up to a million lives; destroy a megacity; trigger a global economic depression (e. g. if it occurs in Tokyo); trigger an ecological catastrophe, rendering a large territory uninhabitable; or destabilize military balance in a region. Regions of low seismicity have become highly vulnerable, e. g. European and Indian platforms, and Central and Eastern parts of the U.S. As a result the earthquakes joined the ranks of the major disasters that, in the words of J. Wisner, have become "a threat to civilization survival, as great as was ever posed by Hitler, Stalin or the atom bomb". Earthquake prediction is necessary to reduce the damage by escalating disaster preparedness. Predictions useful for preparedness should have known, but not necessarily high, accuracy. Such is the standard practice in preparedness for all disasters, wars included.

Introduction

Earthquakes occur in some parts of the outer shell of the solid Earth, called the lithosphere; its thickness ranges from a few kilometers near the mid-ocean ridges to a few hundred kilometers in certain continental regions. At many continental margins the lithosphere bends downward penetrating underlying mantle as seismically active subduction zones. In seismically active regions a significant part of tectonic development is realized through the earthquakes.

About a million earthquakes with magnitude 2 (energy about 10^{15} erg) or more are detected each year worldwide by seismological networks. About a hundred of these cause considerable damage and few times in a decade a catastrophic earthquake occurs.

Catalogs of earthquakes provide the data for detecting premonitory seismicity patterns. Typically for complexity studies we do not have a complete set of fundamental equations that govern dynamics of seismicity and unambiguously define earthquake prediction algorithms. This is due to the multitude of mechanisms controlling seismicity – see Sect. "Generalization: Complexity and Extreme Events". In lieu of such equations "…*we have to rely upon the hypotheses obtained by processing of the experimental data*" (A. Kolmogorov on transition to turbulence). Formulating and testing such hypotheses involves exploratory data analysis, numerical and laboratory modeling, and theoretical studies (Sect. "General Scheme of Prediction").

Diversity of methods and urgency of the problem makes learning by doing a major if not *the* major form of knowledge transfer in prediction of extreme events (http://cdsagenda5.ictp.it/full_display.php?da=a06219).

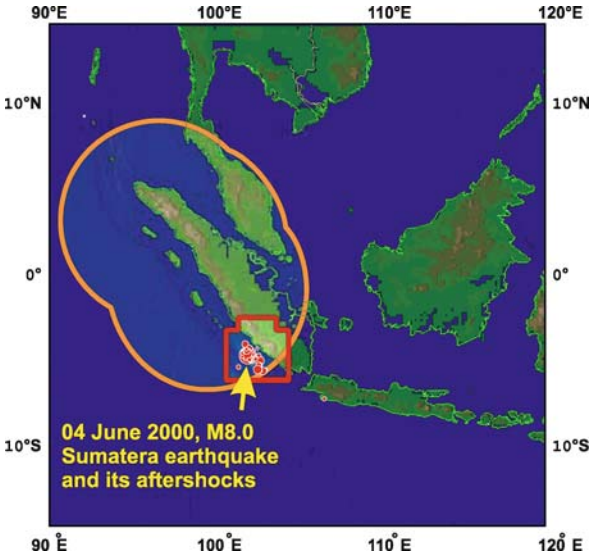

Geo-complexity and Earthquake Prediction, Figure 2
Prediction of the Sumatra earthquake, June 4th, 2000, $M = 8.0$ by algorithms M8 and MSc. The *orange oval curve* bounds the area of alarm determined by algorithm M8, the *red rectangle* is its reducing made by algorithm MSc. *Circles* show epicenters of the Sumatra earthquake and its aftershocks. After [43]

Reliability of the existing algorithms has been tested by continuous prediction of future strong earthquakes in numerous regions worldwide. Each algorithm is self-adapting, i. e. applicable without any changes in the regions with different seismic regimes. Predictions are filed in advance at the websites (http://www.mitp.ru/predictions.html; http://www.phys.ualberta.ca/mirrors/mitp/predictions.html; and http://www.igpp.ucla.edu/prediction/rtp/).

Following is the scoring for four different algorithms.

- *Algorithms* M8 [32] *and* MSc [44] (MSc stands for the Mendocino Scenario). Algorithm M8 gives alarms with characteristic duration years. MSc gives a second approximation to M8, reducing the area of alarm. An example of their application is shown in Fig. 2.

Continually applied since 1992, algorithm M8 has predicted 10 out of 14 large earthquakes (magnitude 8 or more) which have occurred in the major seismic belts. Alarms occupied altogether about 30% of the time-space considered. Both algorithms applied together reduced the time-space alarms to 15%, but three more target earthquakes were missed by prediction.

Geo-complexity and Earthquake Prediction, Figure 3

Prediction of the Northridge, California earthquake, January 28th, 1994, $M = 6.8$ by algorithm SSE. The prediction was made by analysis of aftershocks of the Landers earthquake, June 28th, 1992, $M = 7.6$. An earthquake with $M = 6.6$ or larger was expected during the 18 months after the Landers earthquake within the 169-km distance from its epicenter (shown by a *circle*). The Northridge earthquake occurred on January 28th, 1994, 20 days after the alarm expired. After [43]

- *Algorithm* SSE *or Second Strong Earthquake* [43,91]. Its aim is to predict whether or not a second strong earthquake will follow the one that had just occurred. An alarm lasts 18 months after the first strong earthquake. An example of prediction is shown in Fig. 3. Testing by prediction in advance is set up for California, Pamir and Tien Shan, Caucasus, Iberia and Maghreb, the Dead Sea rift, and Italy. Since 1989 this algorithm made 29 predictions; 24 of which were correct and 5 were wrong.

These scores demonstrate predictability of individual earthquakes. A predictions' accuracy is indeed limited, but sufficient to prevent a considerable part of the damage.

- *Algorithm* RTP *or Reverse Tracing of Precursors* [37,81]. This algorithm gives alarms with a characteristic duration of months. An example of this prediction is shown in Fig. 4. Testing by prediction in advance started only few years ago for California, Japan, the Northern Pacific, Eastern Mediterranean, and Italy with adjacent areas.

Perspective. It is encouraging that only a small part of readily available relevant data, models and theories have been used for prediction so far. This suggests a potential for a substantial increase of prediction accuracy.

Lithosphere as a Hierarchical Complex System

Two major factors turn the lithosphere into a hierarchical dissipative complex system [29,36,87]. The first one is *a hierarchical structure* extending from tectonic plates to grains of rocks. The second factor is *instability* caused by a multitude of nonlinear mechanisms destabilizing the strength and stress fields.

Among extreme events in that system are the strong earthquakes. *An earthquake may be an extreme event in*

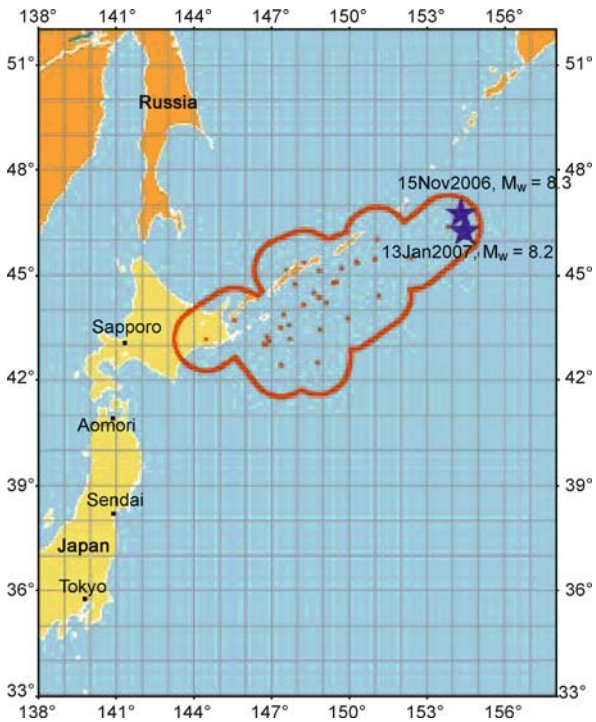

Geo-complexity and Earthquake Prediction, Figure 4
Prediction of Simushir, Kuril Islands earthquakes, November 15th, 2006, *Mw* = 8.3 and January 13th, 2007, *Mw* = 8.2 by Algorithm RTP. An earthquake with magnitude *Mw* ⩾ 7.2 is predicted to occur within the time interval from September 30th, 2006, to June 30th, 2007 in the area bordered by the*red curve*. The *red dots* show epicenters of an earthquake-forming premonitory chain. The *blue stars* show epicenters of the predicted earthquakes

a certain volume of the lithosphere and a part of the background seismicity in a larger volume.

Structure

Blocks The structure of the lithosphere presents a hierarchy of volumes, or blocks, which move relative to each other. The largest blocks are the major tectonic plates, of continental size. They are divided into smaller blocks, such as shields or mountain belts. After 15–20 consecutive divisions we come to about 10^{25} grains of rocks of millimeter size.

Boundary zones Blocks are separated by relatively thin and less rigid boundary zones. They are called fault zones high in the hierarchy, then faults, sliding surfaces, and, finally, interfaces between grains of rock. Except at the bottom of the hierarchy, a boundary zone presents a similar hierarchical structure with more dense division. Some segments of the boundary zones, particularly in tectonically young regions, might be less explicitly expressed, presenting a bundle of small ruptures not yet merged into a fault, of a flexure not yet ruptured, etc.

Nodes *These are even more densely fractured* mosaic structures formed around the intersections and junctions of boundary zones. Their origin is due, roughly saying, to collision of the corners of blocks [16,39,40,55]. The nodes play a singular role in the dynamics of the lithosphere. A special type of instability is concentrated within the nodes and strong earthquakes nucleate in nodes. The epicenters of strong earthquakes worldwide are located only within some specific nodes that can be identified by pattern recognition [19,22].

Nodes are well known in the structural geology and geomorphology and play a prominent textbook role in geological prospecting. However their connection with earthquakes is less widely recognized.

The formalized procedure for dividing a territory into blocks ⇒ faults ⇒ nodes is given in [2].

Fault Network – A Stockpile of Instability

For brevity, the systems of boundary zones and nodes are called here fault networks. They range from the Circum Pacific seismic belt, with the giant triple junctions for the nodes, to interfaces between the grains of rocks, with the corners of grains for the nodes. Their great diversity notwithstanding, fault networks play a similar role in the lithosphere dynamics. Specifically, while tectonic energy is stored in the whole volume of the lithosphere and well beneath, the energy release is to a large extent controlled by the processes in relatively thin fault networks. This contrast is due to the following.

First, the strength of a fault network is smaller than the strength of blocks it separates: fault networks are weakened by denser fragmentation and higher permeability to fluids. For that reason, tectonic deformations are concentrated in fault networks, whereas blocks move essentially as a whole, with a relatively smaller rate of internal deformations. In other words, in the time scale directly relevant to earthquake prediction (hundreds of years or less) the major part of the lithosphere dynamics is realized through deformation of fault networks and relative movement of blocks.

Second, the strength of a fault network is not only smaller, but also highly unstable, sensitive to many processes there. There are two different kinds of such instability. The "physical" one is originated at the micro level by a multitude of physical and chemical mechanisms re-

viewed in the next section. "Geometric" instability is originated at a macro level controlled by the geometry of the fault network (Sect. "Geometric Instability"). These instabilities largely control dynamics of seismicity, the occurrence of strong earthquakes included.

"Physical" Instability [23,29]

As in any solid body, deformations and fracturing in the lithosphere are controlled by the relation of the strength field and stress field. The strength is in turn controlled by a great multitude of interdependent mechanisms concentrated in the fault network. We describe, for illustration, several such mechanisms starting with the impact of fluids.

Rehbinder Effect, or Stress Corrosion [14,85]

Mechanism Many solid substances lose their strength when they come in contact with certain surface-active liquids. The liquid diminishes the surface tension μ and consequently the strength, which is proportional to $\sqrt{\mu}$ by the Griffiths criterion. When the strength drops, cracks may emerge under small stress. Then liquid penetrates the cracks and they grow, with drops of liquid propelling forward, until they dissipate. This greatly reduces the stress required to generate the fracturing. Stress corrosion was first discovered for metals and ceramics. Then such combinations of solid substances and surface-active liquids were recognized among the common ingredients of the lithosphere, e. g. basalt and sulphur solutions. When they meet, the basalt is permeated by a grid of cracks and the efficient strength may instantly drop by a factor of 10 or more due to this mechanism alone.

Geometry of Weakened Areas Orientation of such cracks at each point is normal to the main tensile stress. The stress field in the lithosphere may be very diverse. However, the shape of weakened areas where the cracks concentrate may be of only a few types, determined by the theory of singularities. Some examples are shown in Fig. 5, where thin lines show the trajectories of cracks; each heavy line is a separatrix, dividing the areas with different patterns of trajectories.

If a liquid infiltrates from a place shown in Fig. 5 by arrows, the cracks concentrate in the shaded area, and its strength plummets. A slight displacement of the source across the separatrix may strongly change the geometry of such fatigue; it may be diverted to quite a different place and take quite a different shape, although not an arbitrary

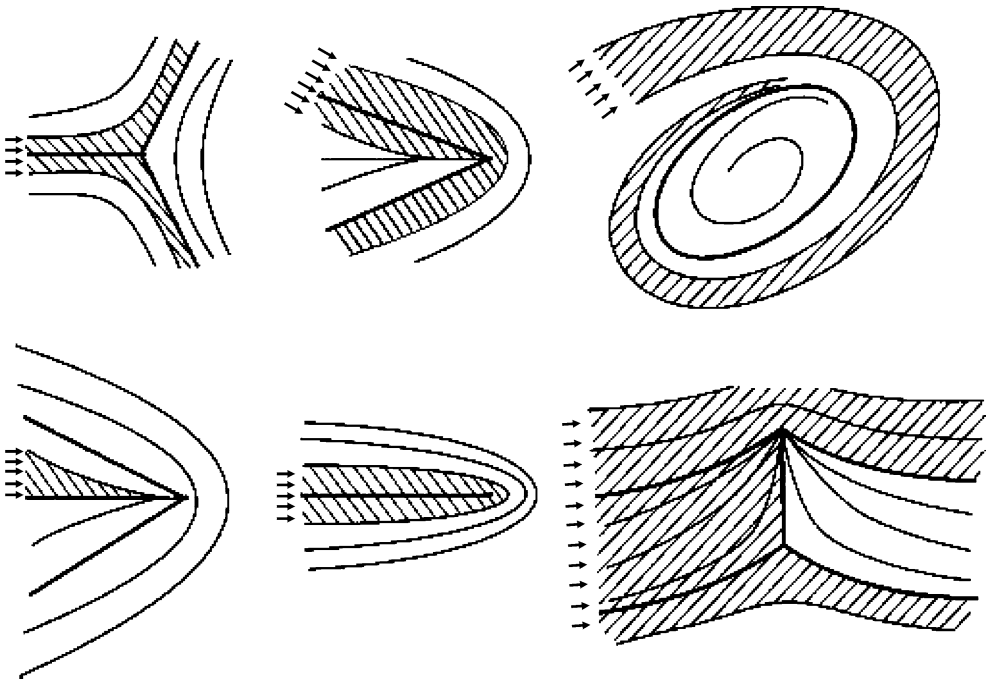

Geo-complexity and Earthquake Prediction, Figure 5
Instability caused by stress corrosion. The geometry of weakened areas depends on the type of singularity and the place where the chemically active fluid comes in. After [14]

one. Furthermore evolution of the stress field may change the type of a singularity, make it disappear or create a new one, and the geometry of fatigue will follow suit.

Stress Corrosion is Highly Sensitive to Geochemistry of Fluids For example, gabbro and dolerite are affected only in the presence of iron oxides; Kamchatka ultrabasic rocks are affected by the andesite lava liquids only in the presence of copper oxide, etc. Migration of fluids would cause observable variations of electromagnetic and geochemical fields.

Summing Up Stress corrosion brings into lithosphere a strong and specific instability, which may explain many observed premonitory seismicity patterns. However the basic configurations of fatigue, as shown in Fig. 5 might be realizable only in not-too-large areas. This limitation stems from the dissipation of fluids and/or from the inhomogeneity of stress field.

Other Mechanisms Boundary zones feature several other mechanisms, potentially as important and certainly as complicated. A few more examples follow.

Mechanical Lubrication by fluids migrating through a boundary zone [7]. The ensuing instability will be enhanced by *fingers of fluids* springing out at the front of migration [6].

Dissolution of Rocks Its impact is magnified by the *Rikke effect* – an increase of solubility of rocks with pressure. This effect leads to a mass transfer. Solid material is dissolved under high stress and carried out in solution along the stress gradient to areas of lower stress, where it precipitates. The Rikke effect might be easily triggered in a crystalline massif at the corners of rock grains, where stress is likely to concentrate.

Petrochemical Transitions Some of them tie up or release fluids, as in the formation or decomposition of serpentines. Other transitions cause a rapid drop of density, such as in the transformation of calcite into aragonite. (This would create a vacuum and unlock the fault; the vacuum will be closed at once by hydrostatic pressure, but a rupture may be triggered.).

Instability is created also by sensitivity of dynamic friction to local physical environment [50], mechanical processes, such as multiple fracturing, buckling, viscous flow, and numerous other mechanisms [49,70].

Most of the above mechanisms are sensitive to variations of pressure and temperature.

Geometric Instability [16]

The geometry of fault networks might be, and often is, incompatible with kinematics of tectonic movements, including earthquakes. This leads to stress accumulation, deformation, fracturing, and the change of fault geometry, jointly destabilizing the fault network. Two integral measures of this instability, both concentrated in the nodes, are *geometric* and *kinematic incompatibility* [16].

Each measure estimates the integrated effect of tectonic movements in a wide range of time scales, from seismicity to geodetic movements to neotectonics.

Geometric Incompatibility The intersection of two strike-slip faults separating moving blocks. Figure 6 is a simple example of geometric incompatibility. If the movements indicated by arrows in Fig. 6a could occur, the corners A and C would penetrate each other and an intersection point would split into a parallelogram (Fig. 6c). In the general case of a finite number of intersecting faults their intersection point would split into a polygon. Such splitting is not possible in reality; the collision at the corners leads to the accumulation of stress and deformations near the intersection followed by fracturing and changes of fault geometry. The divergence of the corners will be realized by normal faulting.

The expansion of that unrealizable polygon with time, $S(t) = Gt^2/2$, measures the intensity of this process. Here, S is the area of the polygon, determined by the slip rates on intersecting faults; t is the elapsed time from the collision, and G is the measure of geometric incompatibility.

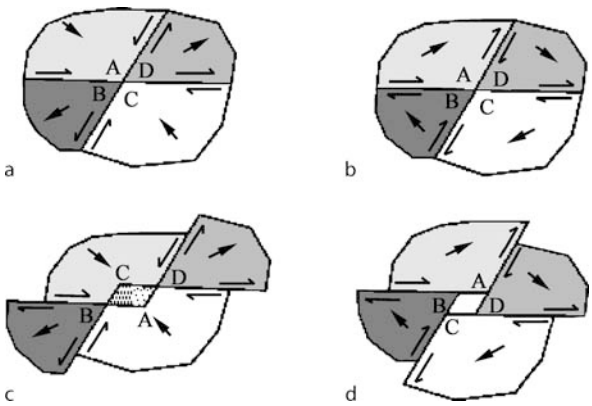

Geo-complexity and Earthquake Prediction, Figure 6
Geometric incompatibility near a single intersection of faults. a, b initial position of the blocks; c, d extrapolation of the blocks' movement; a, c the locked node: movement is physically unrealizable without fracturing or a change in the fault geometry; b, d the unlocked node. After [16]

Such incompatibility of structure and kinematics was first described in [55] for a triple junction. The study established a condition under which a single junction can retain its geometry as the plates move, so that the stress will not accumulate. It was suggested in [39,40] that the general case, when that condition is not satisfied, the ensuing fracturing would not dissolve the stress accumulation, but only redistribute it among newly formed corners. This triggers further similar fracturing with the result that a hierarchy of progressively smaller and smaller faults is formed about an initial intersection. This is a node, recognizable by the dense mosaic structure, with probably self-similar fractal geometry [39].

A real fault network contains many interacting nodes. Incompatibility G is additive, and can be estimated for a network as a whole. An analogue of the Stokes theorem connects the total value of G within a territory with observations on its boundary. This removes the nearly impossible task – to take into account complex internal structure of the nodes. One can instead surround the system of nodes by a contour crossing the less complicated areas. Then the geometric incompatibility can be realistically evaluated from the movements of the fewer faults that cross the contour.

Geometric incompatibility in different nodes is interdependent, because they are connected through the movements of blocks-and-faults system. A strong earthquake in a node would redistribute values G in other nodes thus affecting the occurrence of earthquakes there. Observations indicating the interaction of nodes have been described by [73,74]. These studies demonstrate phenomenon of long-range aftershocks: a rise of seismic activity in the area, where the next strong earthquake is going to occur within about 10 years.

So far, the theory of geometric incompatibility has been developed for the two-dimensional case, with rigid blocks and horizontal movements.

Kinematic Incompatibility Relative movements on the faults would be in equilibrium with the absolute movements of blocks separated by these faults (one could be realized through the other) under the well known Saint-Venant condition of kinematic compatibility [8,56,57]. In the simplest case, shown in Fig. 6, this condition is $K = \sum v_i = 0$, where v_i are slip rates on the faults meeting at the intersection (thin arrows in Fig. 6). The value of K is the measure of the kinematic incompatibility, causing accumulation of stress and deformation in the blocks. A simple illustration of that phenomenon is the movement of a rectangular block between two pairs of parallel faults. The movement of the block as a whole has to be compensated for by relative movements on all the faults surrounding it: if, for example, the movement takes place on only one fault, the stress will accumulate at other faults and within the block itself thus creating kinematic incompatibility.

Like geometric incompatibility the values of K are also additive: one may sum up values at different parts of the network. And an analogue of the Stokes theorem links the value of K for a region with observations on its boundary.

Generalization: Complexity and Extreme Events

Summing up, dynamics of the lithosphere is controlled by a wide variety of mutually dependent mechanisms concentrated predominantly within fault networks and interacting across and along the hierarchy. Each mechanism creates strong instability of the strength-stress field, particularly of the strength. Except for very special circumstances, none of these mechanisms alone prevails in the sense that the others can be neglected.

Even the primary element of the lithosphere, a grain of rock, may act simultaneously as a material point, a viscoelastic body, an aggregate of crystals, a source or absorber of energy, fluids, volume, with its body and surface involved in different processes.

Assembling the set of governing equations is unrealistic and may be misleading as well: A well-known maxim in nonlinear dynamics tells that *one cannot understand chaotic system by breaking it apart* [12]. One may rather hope for a generalized theory (or at least a model), which directly represents the gross integrated behavior of the lithosphere. That brings us to the concept that *the mechanisms destabilizing the strength of fault networks altogether turn the lithosphere into a nonlinear hierarchical dissipative system, with strong earthquakes among the extreme events*. At the emergence of that concept the lithosphere was called a chaotic system [29,66,87]; the more general term is *complex system* [20,24,31,53,78,83].

General Scheme of Prediction

Typically for a complex system, the solid Earth exhibits a permanent background activity, a mixture of interacting processes providing the raw data for earthquake prediction. Predictions considered here are based on detecting premonitory patterns of that activity (Sect. "Definition").

Pattern Recognition Approach

Algorithms described here consider prediction as the pattern recognition problem: *Given* the dynamics of relevant fields in a certain area prior to some time t, *to predict*

whether a strong earthquake will or will not occur within that area during the subsequent time interval $(t, t + \Delta)$. Some algorithms also reduce the area where it will occur.

In terms of pattern recognition, the object of recognition is the time t. The problem is to recognize whether it belongs or not to the time interval Δ preceding a strong earthquake. That interval is often called the *TIP* (an acronym for the *time of increased probability* of a strong earthquake). Such prediction is aimed not at the whole dynamics of seismicity but only at the rare extraordinary phenomena, strong earthquakes.

Pattern recognition of rare events proves to be very efficient in that approach to prediction. This methodology has been developed by the school of I. Gelfand for the study of rare phenomena of complex origin [9,19,34,71].

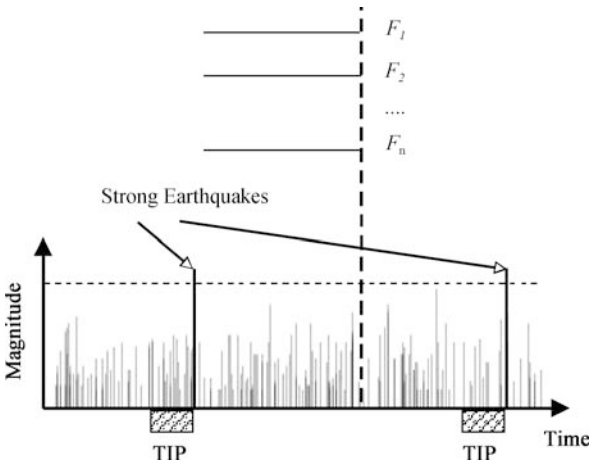

Geo-complexity and Earthquake Prediction, Figure 7
General scheme of prediction. After [29]

Data Analysis

Prediction algorithms are designed by analysis of the learning material – a sample of past critical events and the time series hypothetically containing premonitory patterns. Analysis comprises four following steps:

1. *Detecting premonitory patterns*. Each time series considered is robustly described by the functionals $F_k(t)$, $k = 1, 2, \ldots$, capturing hypothetical patterns (Fig. 7). Hypotheses on what these patterns may be are provided by universal modeling of complex systems (Sect. "Fourth Paradigm: Dual Nature of Premonitory Phenomena"), modeling of Earth-specific processes, exploratory data analysis, and practical experience, even if it is intuitive. Pattern recognition of rare events is an efficient common framework for formulating and testing such hypotheses, their diversity notwithstanding.
   With a few exceptions the functionals are defined in sliding time windows; the value of a functional is attributed to the end of the window. In the algorithms described here the time series were earthquake sequences.
2. *Discretization*. Emergence of a premonitory pattern is defined by the condition $F_k(t) \geqslant C_k$. The threshold C_k is chosen is such a way that a premonitory pattern emerges on one side of the threshold more frequently then on another side. That threshold is usually defined as a certain percentile of the functional F_k. In such robust representation of the data pattern recognition is akin to exploratory data analysis developed in [86].
3. *Formulating an algorithm*. A prediction algorithm will trigger an alarm when a certain combination of premonitory patterns emerges. This combination is determined by further application of pattern recognition procedures [36,71].
4. *Estimating reliability of an algorithm*. This is necessary, since an algorithm inevitably includes many adjustable elements, from selecting the data used for prediction and definition of prediction targets, to the values of numerical parameters. In lieu of the closed theory a priori determining all these elements they have to be adjusted retrospectively, by predicting the past extreme events. That creates the danger of self-deceptive data-fitting: *If you torture the data long enough, it will confess to anything*. Validation of the algorithms requires three consecutive tests.
   - *Sensitivity analysis*: varying adjustable elements of an algorithm.
   - *Out of sample analysis*: applying an algorithm to past data that has not been used in the algorithm's development.
   - *Predicting in advance* – the only decisive test of a prediction algorithm.

Such tests take a lion's share of data analysis [17,19,36,93]. A prediction algorithm makes sense only if its performance is (i) sufficiently better than a random guess, and (ii) not too sensitive to variation of adjustable elements. Error diagrams described in the next section show whether these conditions are satisfied.

Error Diagram

Definition An error diagram shows three major characteristics of a prediction's accuracy. Consider an algorithm applied to a certain territory during the time period T. During the test N strong earthquakes have occurred there and N_m of them have been missed by alarms. Altogether,

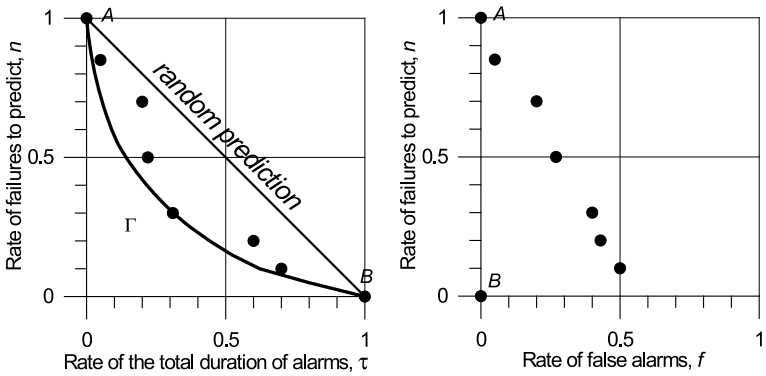

Geo-complexity and Earthquake Prediction, Figure 8

Scheme of an error diagram. Each point shows the performance of a prediction method: the rate of failures to predict, *n*, the relative duration of alarms, τ, and the rate of false alarms, *f*. Different points correspond to different algorithms. The *diagonal* in the *left plot* corresponds to the random guess. Point *A* corresponds to the trivial optimistic strategy, when an alarm is never declared; point *B* marks the trivial pessimistic strategy, when an alarm takes place all the time; other points correspond to non-trivial predictions. Best combinations (n, τ) lie on the envelope of these points Γ. After [63]

A alarms have been declared and A_f of them happened to be false. The total duration of alarms is D.

Performance of an algorithm is characterized by three dimensionless parameters: the relative duration of alarms, $\tau = D/T$; the rate of failures to predict, $n = N_m/N$; and the rate of false alarms, $f = A_f/A$. These three parameters are necessary in any test of a prediction algorithm regardless of a particular methodology. They are juxtaposed on the error diagrams schematically illustrated in Fig. 8. Also called Molchan diagrams, they are used for validation and optimization of prediction algorithms and for joint optimization of prediction and preparedness [59,60,61,62,63]. In many applications parameter f is not yet considered. In early applications they are called ROC diagrams for relative operating characteristics (e. g., [54]).

Four Paradigms

Central for determining premonitory patterns is what we know about them a priori. In other words – what are a priori constraints on the functionals $F_k(t)$ that would capture these patterns (Sect. "Data Analysis"). These constraints are given by the four paradigms described in this section. They have been first found in the quest for premonitory seismicity patterns in the observed and modeled seismicity. There are compelling reasons to apply them also in a wide variety of prediction problems.

Prehistory. New fundamental understanding of the earthquake prediction problem was formed during the last 50 or so years, triggering entirely new lines of research. In hindsight this understanding stems from the following unrelated developments in the early sixties.

- F. Press initiated the installation of the state-of-the-art World-Wide Standardized Seismographic Network (WWSSN) later on succeeded by the Global Seismographic Network (GSN). Thus a uniform data base began to accumulate, augmented by expanding satellite observations.
- E. Lorenz discovered deterministic chaos in an ordinary natural process, thermal convection in the atmosphere [51]. This triggered recognition of deterministic chaos in a multitude of natural and socio-economic processes; however, the turn of seismicity and geodynamics in general came about 30 years later [4,29,66,87]. The phenomenon of deterministic chaos was eventually generalized by less rigorously defined and more widely applicable concept of complexity [20,24,25].
- I. Gelfand and J. Tukey, working independently, created a new culture of exploratory data analysis that allows coping with the complexity of a process (e. g., [19,86]).
- R. Burridge and L. Knopoff [11] demonstrated that a simple system of interacting elements may reproduce a realistically complex seismicity, fitting many basic heuristic constraints. The models of interacting elements developed in statistical physics extended to seismology.
- L. Malinovskaya found a premonitory seismicity pattern reflecting the rise of seismic activity [33]. This is

the first reported earthquake precursor formally defined and featuring long-range correlations and worldwide similarity.

With broader authorship:

- Plate tectonics established the connection between seismicity and large-scale dynamics of the lithosphere [41].
- Research in experimental mineralogy and rocks mechanics revealed a multitude of mechanisms that may destabilize the strength in the fault zones [70].

First Paradigm: Basic Types of Premonitory Patterns

The approach of a strong earthquake is indicated by the following premonitory changes in the basic characteristics of seismicity:

- *Rising:* Seismic activity, earthquakes clustering in space-time, earthquake correlation range, and irregularity of earthquake sequences. Rise of activity sometimes alternates with seismic quiescence.
- *Transforming:* Magnitude distribution (the Gutenberg–Richter relation). Its right end (at larger magnitudes) bends upward, and left end bends downward.
- *Reversing:* territorial distribution of seismicity.
- *Patterns of two more kinds* yet less explored: Rising response to excitation and decreasing dimensionality of the process considered (i. e. rising correlation between its components).

These patterns resemble asymptotic behavior of a thermodynamical system near the critical point in phase transition. Some patterns have been found first in observations and then in models; other patterns have been found in the opposite order. More specifics are given in [15,17,30,31,35,36,67,79,80,83,84,93].

Patterns capturing rise of intensity and clustering, have been validated by statistically significant predictions of real earthquakes [43,65]; other patterns undergo different stages of testing.

Second Paradigm: Long-Range Correlations

The generation of an earthquake is not localized about its future source. A flow of earthquakes is generated by a fault network, rather than each earthquake – by a segment of a single fault. Accordingly, the signals of an approaching earthquake come not from a narrow vicinity of the source but from a much wider area.

What is the size of such areas? Let M and $L(M)$ be the earthquake magnitude and the characteristic length of its source, respectively. In the intermediate-term prediction (on a time scale of years) that size may reach $10L(M)$; it might be reduced down to $3L$ or even to L in a second approximation [43]. On a time scale of about 10 years that size reaches about $100L$. For example, according to [71], the Parkfield (California) earthquake with M about 6 and $L \approx 10$ km "… *is not likely to occur until activity picks up in the Great Basin or the Gulf of California*", about 800 km away.

Historical perspective. An early estimate of the area where premonitory patterns are formed was obtained in [33] for a premonitory rise of seismic activity. C. Richter, who was sceptical about the feasibility of earthquake prediction, made an exception to that pattern, specifically because it was defined in large areas. He wrote [75]: "… *It is important that (the authors) confirm the necessity of considering a very extensive region including the center of the approaching event. It is very rarely true that the major event is preceded by increasing activity in its immediate vicinity.*"

However, such spreading of premonitory patterns has been often regarded as counterintuitive in earthquake prediction research on the grounds that earthquakes can't trigger each other at such distances. The answer is that earthquakes forming such patterns do not trigger each other but reflect an underlying large-scale dynamics of the lithosphere. Among the indisputable manifestations of that correlation are the following phenomena: migration of earthquakes along fault zones [47,52,58,90]; alternate rise of seismicity in distant areas [71] and even in distant tectonic plates [76]. Global correlations have been found also between major earthquakes and other geophysical phenomena, such as Chandler wobble, variations of magnetic field, and the velocity of Earth's rotation [34,72]. These correlations may be explained by several mechanisms not mutually exclusive. Such mechanisms range from micro-fluctuations of large scale tectonic movements to impact of migrating fluids (e. g., [1,5,7,10,69,71,82,84,89]).

Third Paradigm: Similarity

Premonitory phenomena are similar (identical after normalization) in the extremely diverse environments and in a broad energy range (e. g., [1,33,36]). The similarity is not unlimited however and regional variations of premonitory phenomena do emerge.

Normalized prediction algorithms retain their prediction power in active regions and platforms, with the magnitude of target earthquakes ranging from 8.5 to 4.5. Furthermore, similarity extends to induced seismicity, and to multiple fracturing in engineering constructions and lab-

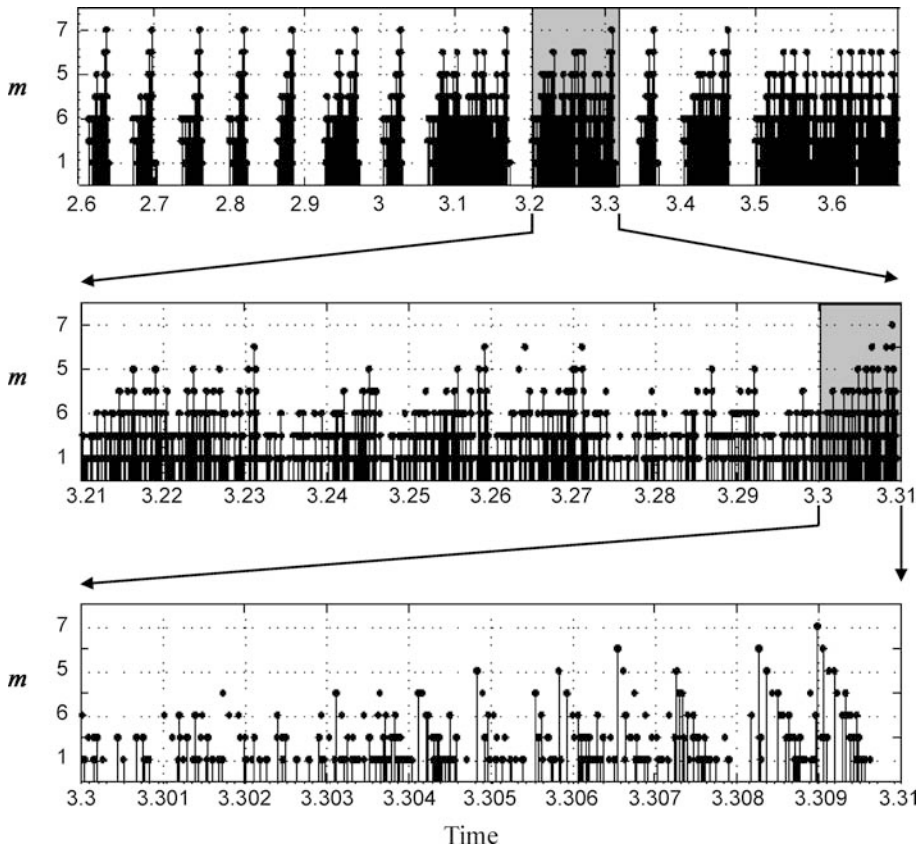

Geo-complexity and Earthquake Prediction, Figure 9
Synthetic earthquake sequence consecutively zoomed. *Shaded areas* mark zoomed intervals. The model shows the rich variety of behavior on different timescales. Note that the ratio of timescales for the *top* and *bottom panels* is 10^2. After [17]

oratory samples (e. g., [3,35,43]). Ultimately, a single but explicit demonstration of similarity was obtained for starquakes – ruptures of the crust of neutron star [45], where the conditions are extremely different than in the Earth. Altogether the corresponding elastic energy release ranges from ergs to 10^{25} ergs (even to 10^{46} ergs if the starquake is counted in).

However, the performance of prediction algorithms does vary from region to region (see [21,35,63]). It is not yet clear whether this is due to imperfect normalization, or to limitations on similarity itself.

Fourth Paradigm: Dual Nature of Premonitory Phenomena

Some premonitory patterns are "universal", common for hierarchical complex systems of different origin; other are specific to geometry of fault networks or to a certain physical mechanism controlling the strength and stress fields in the lithosphere.

Universal patterns. These are most of the patterns so far known. They can be reproduced on models not specific to the Earth only, e. g. models of a statistical physics type (direct or inverse cascade, colliding cascades, percolation, dynamical clustering), models of critical phenomena in fluid dynamics, as well as Earth-specific models themselves.

Complete analytical definition of premonitory patterns was obtained recently on the branching diffusion model [18]. Definition includes only three control parameters, thus strongly reducing uncertainty in data analysis (Sect. "Data Analysis").

Reviews of such models can be found in [15,17,36,66,83,89,93]. Discussion of particular patterns is given also in [25,42,67,68,88,92].

An example of an earthquake sequence generated by a universal model is shown in Fig. 9 [17]. The modeled seismicity exhibits major features of real seismicity: seismic cycle, switching of seismic regime, the Gutenberg–Richter relation, foreshocks and aftershocks, long-range

correlation, and, finally, the premonitory seismicity patterns.

Earth-specific patterns are not yet incorporated in prediction algorithms. We discuss here the patterns reflecting the state of the nodes – structures where the strong earthquakes are nucleated (see Sect. "Structure"). Quantitative characteristics of that state are geometric incompatibility G (Sect. "Geometric Instability"). It shows whether the nodes are locked up or unlocked and quantifies their tendency to fracture and change of the faults geometry. Change of G might create or dissolve such feature as asperities, relaxation barriers, weak links, and replacement of seismicity by creep or "silent" earthquakes [16]. These features would migrate from node to node with velocity typical of seismicity migration: tens to hundreds km/year [90]. All this makes monitoring of G highly relevant to detecting premonitory patterns. A simple pattern of that kind is seismic quiescence around the soon-to-break nodes (e. g., [44,58,77]). A simple highly promising possibility is considering separately premonitory phenomena inside and outside of nodes (e. g., [77]).

Earthquake Prediction and Earthquake Preparedness

Given the limited accuracy of predictions, how do we use them for damage reduction? The key to this is to escalate or de-escalate preparedness depending on the following: content of the current alarm (what and where is predicted), probability of a false alarm, and cost/benefit ratio of disaster preparedness measures. Prediction might be useful if its accuracy is *known*, even if it is not high. Such is the standard practice in preparedness for all disasters, war included.

Diversity of Damage

Earthquakes hurt population, economy, and environment in very different ways: destruction of buildings, lifelines, etc; triggering fires; releasing of toxic, radioactive and genetically active materials; triggering other natural disasters, such as floods, avalanches, landslides, tsunamis, etc.

Equally dangerous are the socio-economic and political consequences of earthquakes: disruption of vital services (supply, medical, financial, law enforcement, etc.), epidemics, drop of production, slowdown of economy, unemployment, disruptive anxiety of population, profiteering and crime. The socio-economic consequences may be inflicted also by the undue release of predictions.

Different kinds of damage are developing at different time and space scales, ranging from immediate damage to chain reaction, lasting tens of years and spreading regionally if not worldwide.

Diversity of Disaster Preparedness Measures Such diversity of damage requires a hierarchy of disaster preparedness measures, from building code and insurance to mobilization of post disaster services to red alert. It takes different times, from decades to seconds to undertake different measures; having different cost they can be maintained for different time periods; and they have to be spread over different territories, from selected sites to large regions. No single stage can replace another one for damage reduction and no single measure is sufficient alone.

On the other hand many important measures are inexpensive and do not require high accuracy of prediction. An example is the Northridge, California, earthquake, 1994, which caused economic damage exceeding $30 billion. Its prediction, published well in advance [48], was not precise – the alarm covered a time period of 18 months and an area 340 km in diameter with dramatically uneven vulnerability. However, low-cost actions, undertaken in response to this prediction (e. g. an out of turn safety inspection) would be well justified if even just a few percent of the damage were prevented.

Joint Optimization of Prediction and Preparedness

The choice of preparedness measures is by no means unique. Different measures may supersede or mutually exclude one another, leaving the decision-maker a certain freedom of choice [38]. The definition of the prediction algorithm is not unique either. The designer of the algorithm has certain freedom to choose the tradeoff between different characteristics of its accuracy (rate of failures to predict, duration of alarms, and rate of failures to predict) by varying adjustable elements of the algorithm (Sect. "General Scheme of Prediction"). That leads to the problem, typical for decision-making with incomplete information: to optimize jointly prediction and preparedness. Figure 10 shows the scheme of such optimization. This figure shows also advantages of a new formulation of prediction: parallel applications of several versions of an algorithm.

Further discussion can be found in [27,28,63,64].

Further Goals

Particularly encouraging for further earthquake prediction research is the wealth of relevant data, models, and theories that are available and yet untapped (the *want amidst plenty* pattern, Conference and School on Predictability of Natural Disasters for our Planet in Danger. A System View: Theory, Models, Data Analysis, 25 June – 6 July

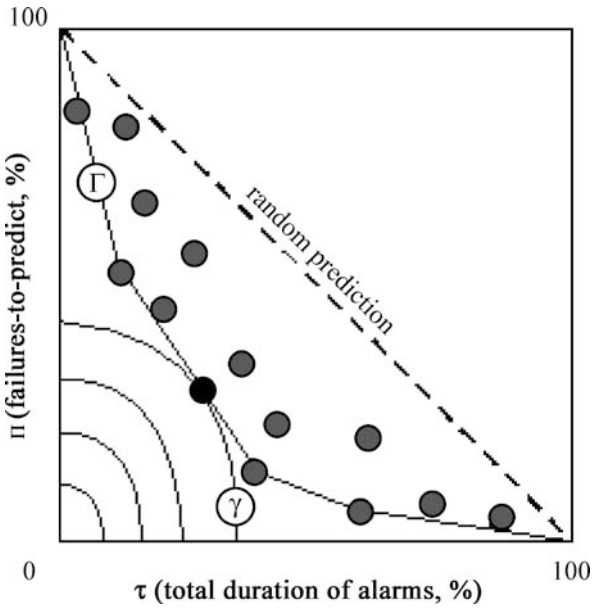

Geo-complexity and Earthquake Prediction, Figure 10
Joint optimization of prediction and preparedness based on the theory of optimal control. Dots show points on the error diagram. Γ is their envelope. Thin contours (γ) show loss curves with constant value of a prevented loss. Optimal strategy is the tangent point of contours Γ and γ. After [63]

2007, Trieste, ICTP, http://cdsagenda5.ictp.it/full_display.php?ida=a06204). Likely within reach is a new generation of prediction algorithms, about five- to ten-fold more accurate than existing ones.

In the general scheme of things, this is a part of wider developments: Emergence of the newly integrated dynamics of the solid Earth, extending from a fundamental concept succeeding plate tectonics to predictive understanding and (with luck) control of geological and geotechnical disasters. And predictive understanding of extreme events (critical phenomena) in the complex systems formed, separately and jointly, by nature and society.

Acknowledgment

The authors are sincerely grateful for insightful comments of Edo Nyland, William Lee, Michele Caputo, and Antoni Correig.

Bibliography

Primary Literature

1. Aki K (1996) Scale dependence in earthquake phenomena and its relevance to earthquake prediction. Proc Natl Acad Sci USA 93:3740–3747
2. Alekseevskaya MA, Gabrielov AM, Gvishiani AD, Gelfand IM, Ranzman EY (1977) Formal morphostructural zoning of mountain territories. J Geophys 43:227–233
3. Allegre CJ, Le Mouël J-L, Provost V (1982) Scaling rules in rock fracture and possible implications for earthquake prediction. Nature 297:47–49
4. Bak P, Chen K, Tang C (1992) A forest-fire model and some thoughts on turbulence. Phys Lett A 147:297–300
5. Barenblatt GI (1993) Micromechanics of fracture. In: Bodner ER, Singer J, Solan A, Hashin Z (eds) Theoretical and Applied Mechanics. Elsevier, Amsterdam, pp 25–52
6. Barenblatt G (1996) Scaling, Self-similarity, and Intermediate Asymptotics. Cambridge University Press, Cambridge
7. Barenblatt GI, Keilis-Borok VI, Monin AS (1983) Filtration model of earthquake sequence. Trans (Doklady) Acad Sci SSSR 269:831–834
8. Bird P (1998) Testing hypotheses on plate-driving mechanisms with global lithosphere models including topography, thermal structure, and faults. J Geophys Res 103(B5):10115–10129
9. Bongard MM, Vaintsveig MI, Guberman SA, Izvekova ML, Smirnov MS (1966) The use of self-learning programs in the detection of oil containing layers. Geol Geofiz 6:96–105 (in Russian)
10. Bowman DD, Ouillon G, Sammis GG, Sornette A, Sornette D (1998) An observational test of the critical earthquake concept. J Geophys Res 103:24359–24372
11. Burridge R, Knopoff L (1967) Model and theoretical seismicity. Bull Seismol Soc Am 57:341–360
12. Crutchfield JP, Farmer JD, Packard NH, Shaw RS (1986) Chaos. Sci Am 255:46–57
13. Farmer JD, Sidorowich J (1987) Predicting chaotic time series. Phys Rev Lett 59:845
14. Gabrielov AM, Keilis-Borok VI (1983) Patterns of stress corrosion: Geometry of the principal stresses. Pure Appl Geophys 121:477–494
15. Gabrielov A, Dmitrieva OE, Keilis-Borok VI, Kossobokov VG, Kuznetsov IV, Levshina TA, Mirzoev KM, Molchan GM, Negmatullaev SK, Pisarenko VF, Prozoroff AG, Rinehart W, Rotwain IM, Shebalin PN, Shnirman MG, Shreider SY (1986) Algorithm of Long-term Earthquakes' Prediction. Centro Regional de Sismologia para America del Sur, Lima
16. Gabrielov AM, Keilis-Borok VI, Jackson DD (1996) Geometric incompatibility in a fault system. Proc Natl Acad Sci USA 93:3838–3842
17. Gabrielov AM, Zaliapin IV, Newman WI, Keilis-Borok VI (2000) Colliding cascade model for earthquake prediction. Geophys J Int 143(2):427–437
18. Gabrielov A, Keilis-Borok V, Zaliapin I (2007) Predictability of extreme events in a branching diffusion model. arXiv: 0708.1542 [nlin.AO]
19. Gelfand IM, Guberman SA, Keilis-Borok VI, Knopoff L, Press F, Ranzman IY, Rotwain IM, Sadovsky AM (1976) Pattern recognition applied to earthquake epicenters in California. Phys Earth Planet Inter 11:227–283
20. Gell-Mann M (1994) The Quark and the Jaguar: Adventures in the Simple and the Complex. Freeman and Company, New York
21. Ghil M (1994) Cryothermodynamics: the chaotic dynamics of paleoclimate. Physica D 77:130–159
22. Gorshkov A, Kossobokov V, Soloviev A (2003) Recognition of

earthquake-prone areas. In: Keilis-Borok VI, Soloviev AA (eds) Nonlinear Dynamics of the Lithosphere and Earthquake Prediction. Springer, Berlin-Heidelberg, pp 239–310
23. Grotzinger J, Jordan TH, Press F, Siever R (2007) Understanding Earth, 5th edn. WH Freeman & Co, New York
24. Holland JH (1995) Hidden Order: How Adaptation Builds Complexity. Addison-Wesley, Reading
25. Huang Y, Saleur H, Sammis C, Sornette D (1998) Precursors, aftershocks, criticality and self-organized criticality. Europhys Lett 41:43–48
26. Kadanoff LP (1976) Scaling, universality and operator algebras. In: Domb C, Green MS (eds) Phase Transitions and Critical Phenomena, vol 5a. Academic Press, London, pp 1–34
27. Kantorovich LV, Keilis-Borok VI (1991) Earthquake prediction and decision-making: social, economic and civil protection aspects. In: International Conference on Earthquake Prediction: State-of-the-Art. Scientific-Technical Contributions, CSEM-EMSC, Strasbourg, pp 586–593
28. Kantorovich LV, Keilis-Borok VI, Molchan GM (1974) Seismic risk and principles of seismic zoning. In: Seismic design decision analysis. Internal Study Report 43, Department of Civil Engineering, MIT, Cambridge (Mass)
29. Keilis-Borok VI (1990) The lithosphere of the Earth as a nonlinear system with implications for earthquake prediction. Rev Geophys 28:19–34
30. Keilis-Borok VI (ed) (1990) Intermediate-Term Earthquake Prediction: Models, Algorithms, Worldwide Tests. Phys Earth Planet Inter, special issue 61(1–2):1–139
31. Keilis-Borok VI (2002) Earthquake prediction: State-of-the-art and emerging possibilities. Annu Rev Earth Planet Sci 30:1–33
32. Keilis-Borok VI, Kossobokov VG (1990) Premonitory activation of earthquake flow: algorithm M8. Phys Earth Planet Inter 61(1–2):73–83
33. Keilis-Borok VI, Malinovskaya LN (1964) One regularity in the occurrence of strong earthquakes. J Geophys Res 69:3019–3024
34. Keilis-Borok VI, Press F (1980) On seismological applications of pattern recognition. In: Allegre CJ (ed) Source Mechanism and Earthquake Prediction Applications. Editions du Centre national de la recherche scientifique, Paris, pp 51–60
35. Keilis-Borok VI, Shebalin PN (eds) (1999) Dynamics of Lithosphere and Earthquake Prediction. Phys Earth Planet Inter, special issue 111(3–4):179–327
36. Keilis-Borok VI, Soloviev AA (eds) (2003) Nonlinear Dynamics of the Lithosphere and Earthquake Prediction. Springer, Berlin-Heidelberg
37. Keilis-Borok V, Shebalin P, Gabrielov A, Turcotte D (2004) Reverse tracing of short-term earthquake precursors. Phys Earth Planet Inter 145(1–4):75–85
38. Keilis-Borok V, Davis C, Molchan G, Shebalin P, Lahr P, Plumb C (2004) Earthquake prediction and disaster preparedness: Interactive algorithms. EOS Trans AGU 85(47), Fall Meet Suppl, Abstract S22B-02a
39. King G (1983) The accommodation of large strains in the upper lithosphere of the earth and other solids by self-similar fault systems: The geometrical origin of b-value. Pure Appl Geophys 121:761–815
40. King G (1986) Speculations on the geometry of the initiation and termination processes of earthquake rupture and its relation to morphology and geological structure. Pure Appl Geophys 124:567–583
41. Knopoff L (1969) The upper mantle of the Earth. Science 163:1277–1287
42. Kossobokov VG, Carlson JM (1995) Active zone size vs. activity: A study of different seismicity patterns in the context of the prediction algorithm M8. J Geophys Res 100:6431–6441
43. Kossobokov V, Shebalin P (2003) Earthquake Prediction. In: Keilis-Borok VI, Soloviev AA (eds) Nonlinear Dynamics of the Lithosphere and Earthquake Prediction. Springer, Berlin-Heidelberg, pp 141–207
44. Kossobokov VG, Keilis-Borok VI, Smith SW (1990) Localization of intermediate-term earthquake prediction. J Geophys Res 95:19763–19772
45. Kossobokov VG, Keilis-Borok VI, Cheng B (2000) Similarities of multiple fracturing on a neutron star and on the Earth. Phys Rev E 61(4):3529–3533
46. Kravtsov YA (ed) (1993) Limits of Predictability. Springer, Berlin-Heidelberg
47. Kuznetsov IV, Keilis-Borok VI (1997) The interrelation of earthquakes of the Pacific seismic belt. Trans (Doklady) Russ Acad Sci, Earth Sci Sect 355A(6):869–873
48. Levshina T, Vorobieva I (1992) Application of algorithm for prediction of a strong repeated earthquake to the Joshua Tree and the Landers earthquakes' aftershock sequence. EOS Trans AGU 73(43), Fall Meet Suppl:382
49. Sir Lighthill J (ed) (1996) A Critical Review of VAN. World Scientific, Singapore-New Jersey-London-Hong Kong
50. Lomnitz-Adler J (1991) Model for steady state friction. J Geophys Res 96:6121–6131
51. Lorenz EN (1963) Deterministic nonperiodic flow. J Atmos Sci 20:130–141
52. Ma Z, Fu Z, Zhang Y, Wang C, Zhang G, Liu D (1990) Earthquake Prediction: Nine Major Earthquakes in China. Springer, New York
53. Ma S-K (1976) Modern Theory of Critical Phenomena. WA Benjamin, Inc, Reading
54. Mason IB (2003) Binary events. In: Jolliffe IT, Stephenson DB (eds) Forecast Verification. A Practioner's Guide in Atmospheric Science. Wiley, Chichester, pp 37–76
55. McKenzie DP, Morgan WJ (1969) The evolution of triple junctions. Nature 224:125–133
56. McKenzie DP, Parker RL (1967) The North Pacific: An example of tectonics on a sphere. Nature 216:1276–1280
57. Minster JB, Jordan TH (1984) In: Crouch JK, Bachman SB (eds) Tectonics and Sedimentation Along the California Margin: Pacific Section, vol 38. Academic, San Diego, pp 1–16
58. Mogi K (1968) Migration of seismic activity. Bull Earth Res Inst Univ Tokyo 46(1):53–74
59. Molchan GM (1990) Strategies in strong earthquake prediction. Phys Earth Planet Inter 61:84–98
60. Molchan GM (1991) Structure of optimal strategies of earthquake prediction. Tectonophys 193:267–276
61. Molchan GM (1994) Models for optimization of earthquake prediction. In: Chowdhury DK (ed) Computational Seismology and Geodynamics, vol 1. Am Geophys Un, Washington, DC, pp 1–10
62. Molchan GM (1997) Earthquake prediction as a decision-making problem. Pure Appl Geophys 149:233–237
63. Molchan GM (2003) Earthquake Prediction Strategies: A Theoretical Analysis. In: Keilis-Borok VI, Soloviev AA (eds) Nonlinear Dynamics of the Lithosphere and Earthquake Prediction. Springer, Berlin-Heidelberg, pp 209–237

64. Molchan G, Keilis-Borok V (2008) Earthquake prediction: Probabilistic aspect. Geophys J Int 173(3):1012–1017
65. Molchan GM, Dmitrieva OE, Rotwain IM, Dewey J (1990) Statistical analysis of the results of earthquake prediction, based on burst of aftershocks. Phys Earth Planet Inter 61:128–139
66. Newman W, Gabrielov A, Turcotte DL (eds) (1994) Nonlinear Dynamics and Predictability of Geophysical Phenomena. Am Geophys Un, Int Un Geodesy Geophys, Washington
67. Newman WI, Turcotte DL, Gabrielov AM (1995) Log-periodic behavior of a hierarchical failure model with application to precursory seismic activation. Phys Rev E 52:4827–4835
68. Pepke GF, Carlson JR, Shaw BE (1994) Prediction of large events on a dynamical model of fault. J Geophys Res 99:6769–6788
69. Pollitz FF, Burgmann R, Romanowicz B (1998) Viscosity of oceanic asthenosphere inferred from remote triggering of earthquakes. Science 280:1245–1249
70. Press F (ed) (1965) Earthquake Prediction: A Proposal for a Ten Year Program of Research. Ad Hoc Panel on Earthquake Prediction. White House Office of Science and Technology, Washington, DC, p 134
71. Press F, Allen C (1995) Patterns of seismic release in the southern California region. J Geophys Res 100(B4):6421–6430
72. Press F, Briggs P (1975) Chandler wobble, earthquakes, rotation and geomagnetic changes. Nature (London) 256:270–273
73. Prozorov AG (1975) Changes of seismic activity connected to large earthquakes. In: Keilis-Borok VI (ed) Interpretation of Data in Seismology and Neotectonics. Comput Seismol vol 8. Nauka, Moscow, pp 71–82 (in Russian)
74. Prozorov AG, Schreider SY (1990) Real time test of the long-range aftershock algorithm as a tool for mid-term earthquake prediction in Southern California. Pure Appl Geophys 133:329–347
75. Richter C (1964) Comment on the paper "One Regularity in the Occurrence of Strong Earthquakes" by Keilis-Borok VI and Malinovskaya LN. J Geophys Res 69:3025
76. Romanowicz B (1993) Spatiotemporal patterns in the energy-release of great earthquakes. Science 260:1923–1926
77. Rundkvist DV, Rotwain IM (1996) Present-day geodynamics and seismicity of Asia minor. In: Chowdhury DK (ed) Computational Seismology and Geodynamics, vol 1. Am Geophys Un, Washington, DC, pp 130–149
78. Rundle JB, Turcotte DL, Klein W (eds) (2000) Geocomplexity and the Physics of Earthquakes. Am Geophys Un, Washington, DC
79. Sammis CG, Sornett D, Saleur H (1996) Complexity and earthquake forecasting. In: Rundle JB, Turcotte DL, Klein W (eds) SFI Studies in the Science of Complexity, vol XXV. Addison-Wesley, Reading
80. Shebalin PN, Keilis-Borok VI (1999) Phenomenon of local "seismic reversal" before strong earthquakes. Phys Earth Planet Inter 111:215–227
81. Shebalin P, Keilis-Borok V, Gabrielov A, Zaliapin I, Turcotte D (2006) Short-term earthquake prediction by reverse analysis of lithosphere dynamics. Tectonophysics 413:63–75
82. Soloviev A, Ismail-Zadeh A (2003) Models of Dynamics of Block-and-Fault Systems. In: Keilis-Borok VI, Soloviev AA (eds) Nonlinear Dynamics of the Lithosphere and Earthquake Prediction. Springer, Berlin-Heidelberg, pp 71–139
83. Sornette D (2004) Critical Phenomena in Natural Sciences: Chaos, Fractals, Selforganization, and Disorder. Concept and Tools, 2nd edn. Springer, Berlin-Heidelberg
84. Sornette D, Sammis CG (1995) Complex critical exponents from renormalization group theory of earthquakes: Implications for earthquake predictions. J Phys I France 5:607–619
85. Traskin VY, Skvortsova ZN (2006) Rehbinder effect in geodynamical processes. In: Kissin IG, Rusinov VL (eds) Fluids and Geodynamics. Nauka, Moscow, pp 147–164 (in Russian)
86. Tukey JW (1977) Exploratory Data Analysis. Addison-Wesley Series in Behavioral Science: Quantitative Methods. Addison-Wesley, Reading
87. Turcotte DL (1997) Fractals and Chaos in Geology and Geophysics, 2nd edn. Cambridge University Press, Cambridge
88. Turcotte DL (1999) Seismicity and self-organized criticality. Phys Earth Planet Inter 111:275–294
89. Turcotte DL, Newman WI, Gabrielov A (2000) A statistical physics approach to earthquakes. In: Geocomplexity and the Physics of Earthquakes. Am Geophys Un, Washington, DC
90. Vil'kovich EV, Shnirman MG (1983) Epicenter migration waves: Examples and models. In: Keilis-Borok VI, Levshin AL (eds) Mathematical models of the structure of the Earth and the earthquake prediction. Comput Sismol, vol 14. Allerton Press, New York, pp 27–36
91. Vorobieva IA, Levshina TA (1994) Prediction of a second large earthquake based on aftershock sequence. In: Chowdhury DK (ed) Computational Seismology and Geodynamics, vol 1. Am Geophys Un, Washington, DC, pp 27–36
92. Yamashita T, Knopoff L (1992) Model for intermediate-term precursory clustering of earthquakes. J Geophys Res 97:19873–19879
93. Zaliapin I, Keilis-Borok V, Ghil M (2003) A Boolean delay model of colliding cascades. II: Prediction of critical transitions. J Stat Phys 111(3–4):839–861

Books and Reviews

Agnew DC, Ellsworth WL (1991) Earthquake prediction and long-term hazard assessment. Rev Geophys Suppl 29:877–889

Aki K (1981) A probabilistic synthesis of precursory phenomena. In: Simpson DV, Richards PG (eds) Earthquake Prediction. An International Review. Manrice Ewing Ser 4. Am Geophys Un, Washington, DC, pp 566–574

Bolt BA (1993) Earthquakes – Newly Revised and Expanded. Freeman, New York

Bongard MM (1967) The Problem of Recognition. Nauka, Moscow (in Russian)

Conference and School on Predictability of Natural Disasters for our Planet in Danger. A System View: Theory, Models, Data Analysis, 25 June–6 July 2007, the Abdus Salam International Centre for Theoretical Physics, Trieste, http://cdsagenda5.ictp.it/full_display.php?ida=a06204

Jaumé SC, Sykes LR (1999) Evolving towards a critical point: A review of accelerating seismic moment/energy release prior to large and great earthquakes. Pure Appl Geophys 155:279–306

Kanamori H (1977) The energy release in great earthquakes. J Geophys Res 82(B20):2981–2988

Mandelbrot B (1983) The Fractal Geometry of Nature. Freeman, New York

Ranzman Ela (1979) Places of Earthquakes and Morphostructures of Mountain Countries. Nauka, Moscow (in Russian)

Varnes DJ (1989) Predicting earthquakes by analyzing accelerating precursory seismic activity. Pure Appl Geophys 130:661–686

Wyss M (ed) (1991) Evaluation of Proposed Earthquake Precursors. Am Geophys Un, Washington DC
Wyss M (1997) Second round of evaluation of proposed earthquake precursors. Pure Appl Geophys 149:3–16
Wyss M, Habermann R (1988) Precursory seismic quiescence. Pure Appl Geophys 126:319–332
Zoller G, Hainzl S, Kurths J (2001) Observation of growing correlation length as an indicator for critical point behaviour prior to large earthquakes. J Geophys Res 106:2167–2176

Geometric Phase and Related Phenomena in Quantum Nanosystems

Shi-Jie Xiong
National Laboratory of Solid State Microstructures
and Department of Physics,
Nanjing University,
Nanjing, China

Article Outline

Glossary
Definition of the Subject
Introduction
Geometric Phase Induced in Different Types of Nanostructures
Berry Phase: Adiabatic Approximation
Theoretical Treatment Beyond Adiabatic Approximation
Effect of Phase Uncertainty
Future Directions
Acknowledgments
Bibliography

Glossary

Target system The system under investigation. Usually the target system is a part of degrees of freedom in a nanostructure whose state is described quantum mechanically.

Environment The system which interacts with the target system and influences its state. The environment may be the parts of degrees of freedom in the nanostructure other than the target, the systems which surround the nanostructure and interact with the target system, or an external electric or magnetic field which is exerted on the target system.

Hamiltonian The quantum operator that describes the energy of the system and acts on the Hilbert space for the quantum states. It is usually denoted by $\hat{H}$ with subscript indicating the described system.

Eigen energy and eigen wave function For time-independent (stationary) Hamiltonian the system can only take specific values of energy, called the eigen energies. Corresponding every eigen energy there is a state of the system described by the eigen wave function. All eigen wave functions are normalized and orthogonal to each other, forming a complete linear space.

Hilbert space A complete linear space spanned by all eigen wave functions of the time-independent Hamiltonian. Any states of the related system can be expressed as a vector in this linear space.

Parameter space The space spanned by parameters which specify the interactions of the environment on the target system. In a dynamical process in which the geometric phase is investigated, the parameters are assumed to be periodically varying in time. The period is denoted as T.

Evolution of wave functions During a period of variation of parameters, the Hamiltonian transverses a circle and returns to its initial situation at the end of the period. In this process the wave function of the target system also undergoes an evolution with the time.

Dynamical phase Even though the Hamiltonian is time independent, the wave function of the target system still has a time dependent phase factor denoted as $e^{i\gamma_d(t)}$. In a period T of a periodic dynamical process the phase γ_d acquired by the target wave function, which is calculated under the assumption that the Hamiltonian is the averaged one and stationary in this period, is called the dynamical phase.

Geometric phase If the Hamiltonian is periodically time-dependent, the phase γ acquired by the target wave function in a period of the evolution is different from the dynamical phase γ_d. The difference $\gamma_g = \gamma - \gamma_d$ is called the geometric phase.

Adiabatic approximation If the evolution of the Hamiltonian is slow enough so that there is no transition between different eigen wave functions of the target system during the evolution, this evolution is regarded as adiabatic. The approximation based on the adiabatic assumption of the process is called adiabatic approximation.

Berry phase The geometric phase calculated in the adiabatic approximation is called the Berry phase.

Quantum interference In the case where the states of a nanostructure are quantum mechanically described by wave functions, the amplitude of a resultant wave function can be constructively enhanced when its components have the same phase, or can be destructively weakened when its components have opposite phases. Generically the resultant wave function de-

pends not only on the amplitudes of its components, but also depends on the phases of the components. This phenomenon is called quantum interference.

Dephasing The processes or mechanisms which cause spatial or temporal uncertainty of phase of the target wave functions.

Definition of the Subject

Owing to the rapid growing of the nanotechnology, now it is possible to design and fabricate different types of nanostructures which have nanometer length scales in one, two, or three dimensions. Examples of nanostructures are: quantum wells [12], quantum wires [33], and quantum dots [15] in which the motion of charge carriers is restricted within nanometer length scale in one, two, and three dimensions, respectively. Usually quantum wells, wires, and dots are made of semiconductor materials and based on the related designing and fabricating technology. Other examples are those of carbon-based nanomaterials, such as graphene [26], carbon nanotubes [17], and fullerenes C_{60} or C_{70} [10], in which the motion of electrons is also restricted within nanometer scale in one, two, and three dimensions, respectively. The electric, optical, and magnetic properties of these nanostructures are mainly depending on the states of electrons. So here we define the electron system in nanostructures as the target system under our investigation. Other degrees of freedom in the nanostructure, such as vibrations of nuclei and other collective excitations, and other materials attaching or near the nanostructure, can be regarded as the environment. In the nanometer scale of some dimensions of the nanostructure the quantum coherence is usually maintained and the effect of quantum interference becomes important. In a dynamical process of the environment, the quantum interference of the target system depends not only on the dynamical phase, but also on the geometric phase. Thus, the investigation of the geometric phase and its effect on nanostructures is of particular interest. Such investigation can reveal the physical origins of peculiar properties of nanostructures observed in experiments and suggest possible methods in designing devices with specific properties.

Introduction

Nanostructures

Recently there has been extensive interest in designing, fabricating, and investigating various types of nanostructures [29]. Owing to the advances of technology, people are now able to arrange atoms into structures that are only a few nanometers in size. A nanometer is about four atoms in diameter. The goal in doing this is in specific self-assembly of nanostructures, which can produce large amounts of artificial materials with new properties. In order to connect nanostructures to the more familiar world of microstructures and microelectronics, one must build up larger assemblies by guided self-assembly.

A common and striking feature of nanostructures is in the length scale of nanometers in one or more dimensions. If an electron is confined within this length scale in one direction, its momentum in this direction will be well quantized due to the boundary condition and, correspondingly, the energy levels characterizing the kinetic motion in this direction are separated so that at low temperatures only one level plays the central role in determining the transport or other properties of the system. This is so-called quantum size effect in nanostructures. An example is the nanowire or carbon nanotube where the motions of electrons in the transverse directions are quantized but the motion in the longitudinal direction is confined and quantized. As a consequence there is only one or few effective energy levels in the transverse directions which correspond one or few traveling channels in the longitudinal direction. If there is only one channel, the system can be viewed as a true one-dimensional one in the longitudinal direction. The quantum size effect is an important manifestation of the quantum effects in nanostructures, implying the central role of the quantum principles.

Besides the quantum size effect, abundant and charming properties of various nanostructures arise from their specific crystal structures, adjusted shapes, and designed arrangement of self-assembly. Such particular structures can provide complicated networks of interfering paths and produce specific band structures which may be useful for applications. A simplest example is the Aharonov–Bohm (AB) rings made of quantum wires which can directly exhibit the quantum interference effect from the AB oscillations of the conductance through the rings as a function of the magnetic flux through the ring [34]. This device makes use of the magnetic flux as a method for controlling the quantum phase of the electron wave functions. Another example is the graphene, a single atomic layer of the graphite where carbon atoms are arranged on a honeycomb lattice. Owing to the specific lattice structure of the graphene, the electrons show massless Dirac relativistic dispersion relation [26]. As we will see below, such a specific band structure gives rise to the Berry or geometric phase and causes exotic transport properties.

The various shapes and structures provide possible methods of adjusting geometric phase j and, in turn, controlling the transport properties in nanostructures. The

physical ingredients taking part in these processes include: (i) the magnetic field or flux through some part of nanostructures [34]; (ii) the spin-orbit coupling which governs the spin precessions during the transport of electrons [28]; (iii) the Jahn–Teller electron-lattice interaction [16]; (iv) the exchange interaction between electrons and magnetic impurities; and (v) the superconducting pairing. Any of these ingredients can take action only in association with special shapes of the nanostructures. For instance, the AB oscillations can occur by varying the magnetic flux only in a ring-shaped device.

Geometric or Berry Phase

We consider that a point (representing a car or a ship) carrying two vectors is traversing along a loop on the surface of a sphere. Both vectors stay tangential to the surface at all times and they remain as parallel as possible to the direction they were pointing before each infinitesimal displacement. In this special example the long vector points "South" and the short one points "East" all the time. After completing the loop, the vectors go back to the original point, but they find themselves rotated with respect to the directions they were pointing at the beginning. Note that the vectors "rotate" despite the fact that we have been careful to keep them parallel during the transport. If the loop is smaller, the rotation angle would have been smaller. For a loop encircling a 1/8, the rotation angle is $\pi/2$. A larger path surrounding for example one quarter of the sphere, would rotate the parallel-transported vectors by π. An even larger loop, a diameter, surrounding (on both sides) half the sphere gives a rotation angle of 2π, i. e. no rotation at all. This rotation originates from the intrinsic curvature of the sphere surface. This phenomenon can not appear for the transport along a flat manifold, such as a plane or a cylinder. So the rotation angle is related to the integral of the curvature over the surface bounded by the loop.

In the quantum version such rotation angles correspond to the Berry phases. In quantum mechanics the state of the target system, i. e., an electron in a nanostructure, is characterized by the wave function ψ which has not only the magnitude $|\psi|$ but also a phase factor $e^{i\gamma}$. Phase γ is equivalent to the angle in the complex plane. Several specific cases had been recognized for many years, in particular the well-studied Aharonov–Bohm effect in the ring of quantum wire threaded with a magnetic flux. In 1984, Berry published an influential formal systematization of the closed-path geometric phase in quantum mechanical problems [8]. In the following we summarize some major points about the Berry phase or geometric phase in nanostructures.

Wave functions of the target system are parameterized vector fields in the Hilbert space. Each wave function has a phase connected to its time evolution. The dynamical phase is not geometrical: it depends on the energy at which the parameterized path is followed, and on the arbitrary choice of a "zero phase" (technically called a gauge choice) at each point along the path. A recipe is needed to get rid of any dynamical phases, and single out the intrinsic, geometrical invariant phases. Let us consider the set of eigen wave functions of a Hamiltonian $\hat{H}$, depending on several classical "external" parameters collectively denoted by q. Changing q from one point to another in the multi-dimensional parameter space, the set of eigen wave functions and corresponding eigen energies of $\hat{H}$ change to totally new and unrelated eigen system. In the adiabatic approximation smooth evolution of $\hat{H}$ is assumed so that one expects that for infinitesimal change Δq in the parameters, the inner product between a given eigen wave function at some point q, and the evolved eigen wave function at a following neighboring point $q + \Delta q$, should be as close to 1 as possible with infinitesimal error of the order Δq^2.

We note that every stationary wave function "rotates" as $e^{-iEt/\hbar}$, where E is the energy of the state, t is time and $\hbar$ is Planck's constant, because of the Schrödinger time evolution. On top of this, since at each point the eigen system is unrelated to that at a different point, nothing forbids to multiply each eigen wave function by an additional arbitrary phase factor $e^{-i\phi(q)}$ (a "gauge choice"). The correct rule for the "parallel transport", as sketched above, permits to calculate the geometrical phase, such that it ignores completely any dynamical and gauge phase. The result depends only on the geometry of the loop. In this sense, Berry phase is said to be "gauge-invariant".

In 1987, Aharonov and Anandan [1] extended this calculation to the non-adiabatic case. By exactly solving the time-dependent Hamiltonian one can obtain the total phase γ acquired by the wave function in the evolution from $t = 0$ to $t = T$. At the same time the acquired dynamical phase in this period can be calculated as

$$\gamma_{\mathrm{d}} = -\frac{1}{\hbar}\int_0^T \langle\psi(t)|\hat{H}(t)|\psi(t)\rangle \mathrm{d}t \,, \tag{1}$$

where $\hat{H}(t)$ is the time dependent Hamiltonian and $|\psi(t)\rangle$ is the investigated wave function. Then the geometric phase acquired in this period is

$$\gamma_{\mathrm{g}} = \gamma - \gamma_{\mathrm{d}} \,. \tag{2}$$

In the following sections we will describe several important phenomena related to the geometric phase in nanostructures.

Geometric Phase Induced in Different Types of Nanostructures

In this section, we outline different groups of phenomena related to the geometric phase in nanostructures. In following sections, detailed mathematical and physical illustrations will be given for particular examples.

As mentioned above, there are three ingredients for the occurrence of the geometric phase in nanostructures: (i) the target system representing the central degrees of freedom under investigation; (ii) the environment which makes the Hamiltonian of the target system evolve periodically in a parameter space; (iii) the coupling or interaction between the target system and the environment. According to these, there are three major groups of the related phenomena: (i) the target is a spin system; (ii) the target is the orbital motion of electrons interacting with a field; (iii) the target is the orbital motion of electrons interacting with nuclei vibrations. We will discuss these phenomena in the following subsections.

Spins as the Target System

A nanostructure consists of electrons and atomic nuclei. For both of them there are two types of degrees of freedom: the orbital (spatial) motions and the spins. Usually the spatial motion of electrons in outer levels of atoms is important in determining the properties of the system. In some cases, however, this spatial motion of electrons is restricted by some factors, such as the potential barriers in quantum dots, so that the spin degrees of freedom become more important. In such cases the spins of the whole or a part of the nanostructure may be regarded as the target system. Under special environment the spin system can exhibit phenomena related to the geometric phase.

Spinning Quantum System in an External Magnetic Field If a spin, say, $s = 1/2$ is aligned to a magnetic field, and the field is made to rotate adiabatically, so that the spinor remains aligned to it, at the end of the loop the spinor acquires a complex phase factor $e^{i\gamma_g}$. For the general component of spin-1/2 spinor along the magnetic field, the Berry phase factor is precisely $e^{-i/2\phi_s}$, where ϕ_s is the solid angle enclosed in the circuit followed by the magnetic field. Thus, the field rotating within a plane (thus enclosing one half of the complete solid angle 4π), the value of γ_g is π. This amounts to a "180 degrees" rotation in the Hilbert space of wave function. Note, however, that in the real three-dimensional space, the spin has actually rotated by 360 degrees. Technically, the factor of two distinguishing integer and half integer s-spinors is related to the 1:1 and 2:1 representations of the three-dimensional rotation group, and equivalently to the 2:1 relation of SU(2) to SO(3) [31].

Spinning Quantum System in the Presence of Spin Orbit Coupling Besides the external magnetic field, the spin degree of freedom as a target system can also be controlled by the spin-orbit coupling of electrons moving in a quantum wire ring. In the presence of spin-orbit interaction there is a momentum-dependent effective magnetic field coupled to the electron spin. If the momentum traverses a circuit in the momentum space which can be viewed as the parameter space for the spin Hamiltonian, and if the spin orientation keeps in the direction of the effective magnetic field, the spin state acquires a Berry phase in its cyclic evolution. This Berry phase is just the spin-orbit Berry phase. During a cyclic motion of the electron along the ring, the spin of electron is precessing due to the spin-orbit coupling. It has been shown that in this case the Berry phase is $-1/2$ of the solid angle subtended by the circuit traced on a sphere by the spin orientation precessing along the ring [4,28]. This investigation has been recently extended to the case including both the external magnetic field and the spin-orbit coupling in a ring [11].

Electron Motion in Magnetic Field

In this subsection we discuss the situation where the target is the spatial motion of electrons and the environment is an external magnetic field. In this case the orbital motion should be important in determining the properties of the system. So there should be some extended dimensions in the nanostructure. The typical nanostructures possessing extended dimensions are the quantum wells having two extended dimensions and quantum wires having one extended dimension. In the following we separately discuss the geometric phase in these two types of nanostructures.

Aharonov–Bohm Effect The Aharonov–Bohm effect occurs in a ring made of quantum wire threaded by a magnetic flux. In this case a geometric phase is acquired by electrons during its cyclic motion around the ring due to differences in local values of the magnetic vector potential. As one must have guessed, however, in most cases the Berry phase vanishes. As one might expect, usually coming back after a loop to the starting point, makes nothing important, and the vector quantities go back to exactly what they used to be before the loop, indicating that the space where those vectors are moved around is flat. So, nonzero phases are clues pointing to nontrivial topological properties underlying the relation between the vector quantities and the adiabatic parameters. In the Aharonov–

Bohm effect the flatness of the space is violated at the point where the flux penetrates. This leads to the geometric phase which depends on the flux. So the topological phase of Aharonov and Bohm can be viewed as a special case of the Berry geometric phase [8].

Quantum Hall Effect If the Aharonov–Bohm ring is widened, the system is changed to a plane. The quantum Hall effect, a striking quantum phenomenon, occurs in two-dimensional electrons in strong magnetic field. Of course the quantum Hall effect is a very complicated phenomenon under extensive investigations. However, owing to the breaking of the flatness of space similar to the Aharonov–Bohm effect, some methods or concept developed in the investigation of the geometric phase can be adopted in the investigation of the quantum Hall effect. Especially, the quantization of the Hall conductance observed in experiments may be explained as a topological invariant represented by the Chern number which is related to the Berry curvature of the space.

Orbital Motion of Electrons Interacting with Nuclei Vibrations

With the Born–Oppenheimer separation of electronic and vibrational motion, the Berry phase is zero on a physical path when the electronic ground state is well separated from the first electronic excitation. An exception is the so-called Jahn–Teller (JT) effect where the ground state is degenerate at some specially symmetric q-point in the parameter space. The distortion of the crystal reduces the symmetry and corresponds to a displacement in the parameter q-space from the symmetric point. To illustrate this idea we consider here the instructive example of the triangular cluster ($E \times e$ JT model). The equilibrium geometry of a triangular cluster is not the equilateral triangle, but a distorted isosceles geometry. In the equilateral configuration, the electronic ground state is twofold degenerate (E). These two electronic wave functions split for any displacement that destroys the equilateral geometry. Due to the symmetry, it can be seen that equivalent isosceles distortions where any two atoms get closer are energetically equivalent. Actually, it may be verified that for linear electron-distortion coupling the equilibrium configurations constitute a flat circular valley when the Born–Oppenheimer potential energy is calculated against the q coordinates. The two Born–Oppenheimer potential surfaces are degenerate at the central $q = 0$ high-symmetry point. So, this "JT valley" encircles a degeneracy point for the electronic Hamiltonian. As a consequence a Berry phase should be present and physically relevant, as this loop is frequently accessed by the low-energy dynamics of the q-distortions. The electronic wave functions undergoes a Berry phase change of π when the distortions loop around the circular valley. In the adiabatic approximation we choose carefully the phase at each point in order to simulate the "parallel transport", i. e., to have positive overlap to the immediately preceding configuration. When a loop along the JT valley is completed, the electronic state changes sign. Another loop is needed to restore the original sign of the electron wave function.

Berry Phase: Adiabatic Approximation

In this section, we present the detailed calculation of the Berry phase of electronic wave functions due to the electron-lattice interaction in graphene as a concrete illustration of the adiabatic approximation.

Due to the advances of material science, the graphene, as a real two-dimensional (2D) system with only one layer of carbon atoms, has been fabricated recently. In spite of the simplest lattice structure, the graphene exhibits striking properties which attracted much attention of both experimentalists and theorists [2,3,6,7,19,21,22,23,24,25,26,27,35,36]. Particularly, the Dirac dispersion relation of electrons in graphene and the Fermi level near the Dirac point lead to specific features different from those in usual metals and semiconductors. Its transport properties are very special. There exists a universal maximal resistivity for all samples with the Fermi level near the Dirac point, independent of their shapes and mobility. This behavior is difficult to be explained from quantum transport theory, as such a minimal conductivity can not be supported in 2D with limited number of tunneling channels. Moreover, it is found that the weak localization is strongly suppressed [22]. This suppression could be attributed to a dephasing effect similar to that of a random magnetic field causing phase uncertainty. Thus, the transport in graphene is more likely to be diffusive or classical-like rather than quantum. There are several theoretical studies using different methods addressing the unusual transport properties in graphene [2,3,19,21,23,24,27,36].

In [37] we discuss one of the possible mechanisms which may cause the phase uncertainty in fermion states of graphene: the geometric phase produced by periodic relative vibrations between two sublattices. By deriving the time-dependent fermion Hamiltonian and exactly solving it with the Floquet scheme, one shows that in an evolution period of the vibration the geometric phases acquired by the fermions can be expressed by the adiabatic theory only in the momentum region where the level splitting of fermions in the evolution path is much larger than $\hbar\omega$ with

ω being the vibration frequency. Outside this region the adiabatic condition is not satisfied. We show that the geometric phase, the fermion average energy, and the energy uncertainty in the non-adiabatic region are essentially different from those in the adiabatic region. Since geometric phases can be easily produced by the fermion-lattice interaction during lattice vibrations and they exhibit random or chaotic behavior in the non-adiabatic region, the resultant phase uncertainty may be a possible dephasing mechanism in such a 2D material.

The band structure in graphene can be described by a tight-binding Hamiltonian with one π orbital per site on the 2D honeycomb lattice [32]:

$$H = \sum_{\langle nn' \rangle} t_{nn'} (a_n^\dagger a_{n'} + a_{n'}^\dagger a_n) , \tag{3}$$

where $a_n^\dagger(a_n)$ creates (annihilates) an electron at site n, $\langle \ldots \rangle$ denotes the nearest-neighbor (NN) sites, and $t_{nn'}$ is the NN hopping. Here the spin indices are not explicitly included. On a honeycomb lattice there are two sublattices, labeled as A and B, and the cells, each of which consists of two atoms, compose a triangle lattice. On a perfect lattice, all the NN hopping integrals are the same, $t_{nn'} = t_0$. When there is a displacement $\mathbf{d}$ between two sublattices, the lengths of three NN bonds connected to a site of sublattice A become

$$l^{(j)} = \sqrt{[d \sin(\theta - \alpha^{(j)})]^2 + [l_0 - d \cos(\theta - \alpha^{(j)})]^2} , \tag{4}$$

where $j = 1, 2, 3$ labels the three NN bonds, θ is the angle of $\mathbf{d}$ related to the x-axis, l_0 is the original bond length, and $\alpha^{(j)} = 0, 2\pi/3, -2\pi/3$ for $j = 1, 2, 3$, respectively, is the original azimuth angle of the jth bond. For a small displacement, we can keep only the first-order terms of d and obtain $l^{(j)} \sim l_0 - d \cos(\theta - \alpha^{(j)})$. Then the hopping integrals as functions of $\mathbf{d}$ can be written as

$$t^{(j)}(\mathbf{d}) = t_0 + \lambda d \cos(\theta - \alpha^{(j)}) , \tag{5}$$

where λ is a coefficient describing the linear dependence of $t^{(j)}(\mathbf{d})$ on $l^{(j)}$.

If one uses the following Bloch transformation for electron operator:

$$a_n = \frac{1}{\sqrt{N}} \sum_{\mathbf{k}} \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}_n} a(b)_{\mathbf{k}} , \quad \text{for} \quad n \in \text{sublattice A(B)} , \tag{6}$$

where $a(b)_{\mathbf{k}}$ is annihilation operator of electron on sublattice A(B) with Bloch wavevector $\mathbf{k}$, $\mathbf{r}_n$ is the position of site n, and N is the total number of cells, the original Hamiltonian becomes

$$H_0 = \sum_{\mathbf{k}} t_0 \left[\mathrm{e}^{\mathrm{i}k_x l_0} + 2\mathrm{e}^{-\mathrm{i}k_x l_0/2} \cos\left(\frac{\sqrt{3}}{2} k_y l_0 \right) \right] a_{\mathbf{k}}^\dagger b_{\mathbf{k}} + \text{H.c.} \tag{7}$$

Expanding it to the first order of $\mathbf{k}$ around two irreducible K points $\left(0, \mp 4\pi/(3\sqrt{3}l_0)\right)$ in the Brillouin zone, one has

$$H_0 = \frac{3}{2} \sum_{\mathbf{k}} \left[t_0 l_0 k_y \hat{\tau}_z \otimes \hat{\sigma}_x - t_0 l_0 k_x \hat{\mathbf{1}} \otimes \hat{\sigma}_y \right] , \tag{8}$$

where $\hat{\mathbf{1}}$ and $\hat{\tau}_z$ are unit and Pauli matrices acting on two valleys, and $\hat{\sigma}_{x,y}$ are Pauli matrices on two sublattices.

At the same time, the interaction between electrons and lattice displacement $\mathbf{d}$ can be described by

$$H_1 = \sum_{\langle nn' \rangle} (t_{nn'}(\mathbf{d}) - t_0)(a_n^\dagger a_{n'} + a_{n'}^\dagger a_n) , \tag{9}$$

where $t_{nn'}(\mathbf{d})$ is the hopping integral between sites n and n' under displacement $\mathbf{d}$. From Eqs. (5) and (6) and keeping only the terms of the first order of $\mathbf{d}$, the interaction Hamiltonian becomes

$$H_1 = \frac{3\lambda}{2} \sum_{\mathbf{k}} (d_x \hat{\mathbf{1}} \otimes \hat{\sigma}_x + d_y \hat{\tau}_z \otimes \hat{\sigma}_y) . \tag{10}$$

In this paper we only consider vibrations at the long wavelength limit, i. e., the relative displacement $\mathbf{d}$ between two sublattices is uniform over the whole system. In this case the vibrations only mix fermion states of particle and hole branches with the same momentum $\mathbf{k}$ and in the same valley. As we will see below, the evolutions of fermion states with different $\mathbf{k}$ during the vibrations are drastically different if the processes are not adiabatic, although the interaction terms in H_1 are independent of $\mathbf{k}$. The vibration modes with finite wavelengths can mix fermion states of two branches with different momenta. This may produce a more complicated level structure of interacting fermions, but, as we will see below, the main features in the non-adiabatic region are chaotic and may not be very sensitive to the details of levels.

We consider the in-plane vibrations of the relative coordinates as $d_x = (\eta_x/\lambda)\cos(\omega t + \alpha_x)$ and $d_y = (\eta_y/\lambda)\cos(\omega t + \alpha_y)$, where $\eta_{x(y)}/\lambda$ and $\alpha_{x(y)}$ are amplitude and initial phase of the vibration in the $x(y)$ direction, respectively. To be specified in this paper, we do not consider other modes of vibrations which may produce similar effects.

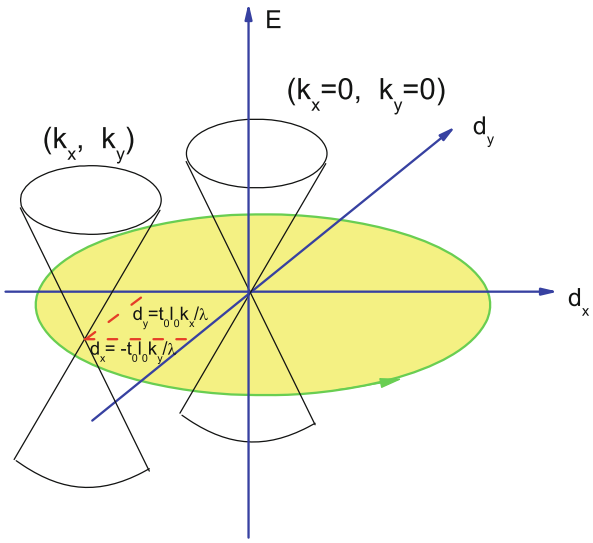

Geometric Phase and Related Phenomena in Quantum Nanosystems, Figure 1
$E_{1,2}^{(+)}$ as functions of d for different values of k. *Green circle* shows the evolution path in the vibrations

The Hamiltonian for electrons becomes time-dependent:

$$H(t) = \frac{3}{2}\sum_{\mathbf{k}}\left[(t_0 l_0 k_y \hat{\tau}_z + \eta_x \cos(\omega t + \alpha_x)\hat{\mathbf{1}}) \otimes \hat{\sigma}_x - (t_0 l_0 k_x \hat{\mathbf{1}} - \eta_y \cos(\omega t + \alpha_y)\hat{\tau}_z) \otimes \hat{\sigma}_y\right] . \quad (11)$$

There is no mixing between two valleys. For a given valley labeled by + or −, the Hamiltonian is

$$H_{\pm}(t) = \frac{3}{2}\sum_{\mathbf{k}}\{t_0 l_0(\pm k_y \hat{\sigma}_x - k_x \hat{\sigma}_y) + \eta_x \cos(\omega t + \alpha_x)\hat{\sigma}_x \pm \eta_y \cos(\omega t + \alpha_y)\hat{\sigma}_y\} . \quad (12)$$

At the adiabatic limit, the instantaneous eigen energies at time t can be obtained by diagonalizing $H_{\pm}(t)$:

$$E_m^{(\pm)}(t) = (-1)^m \tfrac{3}{2}\left|t_0 l_0(\pm k_y + \mathrm{i}k_x) + \eta_x \cos(\omega t + \alpha_x) + \mp \mathrm{i}\eta_y \cos(\omega t + \alpha_y)\right| , \quad (13)$$

where $m = 1$ and $m = 2$ correspond to hole and particle branches, respectively. For given (k_x, k_y), these two instantaneous eigen energies also exhibit conic dependence on parameters d_x and d_y, as shown in Fig. 1. It is interesting to note that the diabolical point, where the poles of two tapers coincide, is determined by $(d_x = \mp t_0 k_y l_0/\lambda,\ d_y = \pm t_0 k_x l_0/\lambda)$, where plus and minus signs refer to two valleys. The vibrations of d_x and d_y result in circular motion with an elliptic track in the $d_x - d_y$ plane. From the Berry theorem, a cyclic motion along a close track in a 2D parameter space can cause a Berry phase of $\pm\pi$ in a wavefunction whose instantaneous eigen energy has a diabolical point enclosed in this circle [8]. Since the position of the diabolic point in the $d_x - d_y$ plane is determined by (k_x, k_y), the Berry phase acquired by the fermions with different momenta are different: It is $\pm\pi$ for the fermions whose diabolic point $(d_x = \mp t_0 k_y l_0/\lambda,\ d_y = \pm t_0 k_x l_0/\lambda)$ is enclosed in the track, while it is zero for the states with the diabolic point in the $d_x - d_y$ parameter space outside the track.

Theoretical Treatment Beyond Adiabatic Approximation

The above conclusion about the Berry phase, however, is valid only in the adiabatic condition, i. e., the energy difference between two blanches in the track is always much larger than $\hbar\omega$. Owing to the existence of energy degeneracy, for any track and any finite frequency there always exist fermion states for which the adiabatic condition can not be satisfied. So we have to investigate the geometric phase from solutions of time-dependent Hamiltonian. Under this Hamiltonian, the wavefunctions of electrons can be expressed in the Floquet form [30]

$$\psi^{\pm}(l, \mathbf{k}, t) = \mathrm{e}^{-\mathrm{i}\frac{\epsilon^{\pm}(l,\mathbf{k})}{\hbar}t} u^{\pm}(l, \mathbf{k}, t) , \quad (14)$$

where l is an index of electron states for given $\mathbf{k}$, and $u^{\pm}(l, \mathbf{k}, t)$ is a periodic function of t with period $T = 2\pi/\omega$. Performing Fourier transformation for $u^{\pm}(l, \mathbf{k}, t)$ with respect to t, one has

$$u^{\pm}(l, \mathbf{k}, t) = \sum_n \begin{pmatrix} A_n^{\pm}(l, \mathbf{k}) \\ B_n^{\pm}(l, \mathbf{k}) \end{pmatrix} \mathrm{e}^{-\mathrm{i}n\omega t} , \quad (15)$$

where $A_n^{\pm}$ and $B_n^{\pm}$ are components on two sublattices. Substituting the Floquet wave function into the time-dependent Schrödinger equation $[H_{\pm}(t) - \mathrm{i}\hbar\partial_t]\psi^{\pm}(l, \mathbf{k}, t) = 0$, we obtain a set of linear homogeneous equations for the components [37]

$$\frac{3t_0 l_0}{2}(\pm k_y + \mathrm{i}k_x)B_n^{\pm} + \frac{3}{4}(\mathrm{e}^{\mathrm{i}\alpha_x}\eta_x \mp \mathrm{i}\mathrm{e}^{\mathrm{i}\alpha_y}\eta_y)B_{n-1}^{\pm} + \frac{3}{4}(\mathrm{e}^{-\mathrm{i}\alpha_x}\eta_x \mp \mathrm{i}\mathrm{e}^{-\mathrm{i}\alpha_y}\eta_y)B_{n+1}^{\pm} = (\epsilon^{\pm} + n\hbar\omega)A_n^{\pm} , \quad (16)$$

$$\frac{3t_0 l_0}{2}(\pm k_y - \mathrm{i}k_x)A_n^{\pm} + \frac{3}{4}(\mathrm{e}^{\mathrm{i}\alpha_x}\eta_x \pm \mathrm{i}\mathrm{e}^{\mathrm{i}\alpha_y}\eta_y)A_{n-1}^{\pm} + \frac{3}{4}(\mathrm{e}^{-\mathrm{i}\alpha_x}\eta_x \pm \mathrm{i}\mathrm{e}^{-\mathrm{i}\alpha_y}\eta_y)A_{n+1}^{\pm} = (\epsilon^{\pm} + n\hbar\omega)B_n^{\pm} . \quad (17)$$

From the requirement of the existence of nonzero solutions, for a given $\mathbf{k}$, one can solve the discrete quasienergies $\epsilon^{\pm}(l,\mathbf{k})$ and the corresponding Floquet states $\psi^{\pm}(l,\mathbf{k},t)$. It is noteworthy that two Floquet states whose quasienergies differ by $n\hbar\omega$ with n being an integer are physically equivalent states [30]. So the quasienergies of physically different Floquet states can be reduced into region $-\hbar\omega/2 \le \epsilon^{\pm} \le \hbar\omega/2$.

In a period of evolution of Hamiltonian $H^{\pm}(t)$, the wavefunction $\psi^{\pm}(l,\mathbf{k},t=T) = \mathrm{e}^{\mathrm{i}\phi^{\pm}(l,\mathbf{k})}\psi^{\pm}(l,\mathbf{k},t=0)$ acquires a phase $\phi^{\pm}(l,\mathbf{k})$. It consists of two parts:

$$\phi^{\pm}(l,\mathbf{k}) = \alpha^{\pm}(l,\mathbf{k}) + \beta^{\pm}(l,\mathbf{k}) , \tag{18}$$

where

$$\alpha^{\pm}(l,\mathbf{k}) = -\frac{\bar{E}^{\pm}(l,\mathbf{k})T}{\hbar} \tag{19}$$

is the dynamical phase with

$$\bar{E}^{\pm}(l,\mathbf{k}) = \frac{1}{T}\int_0^T \mathrm{d}t\langle\psi^{\pm}(l,\mathbf{k},t)|H_{\pm}(t)|\psi^{\pm}(l,\mathbf{k},t)\rangle , \tag{20}$$

and

$$\beta^{\pm}(l,\mathbf{k}) = \phi^{\pm}(l,\mathbf{k}) - \alpha^{\pm}(l,\mathbf{k}) \tag{21}$$

is the geometrical phase [1]. Under the periodic vibrations the Floquet state $\psi^{\pm}(l,\mathbf{k},t)$ is stationary one which returns to the initial state other than phase $\phi^{\pm}(l,\mathbf{k}) = -\epsilon^{\pm}(l,\mathbf{k})$ after a period of evolution. At the same time, the energy is no longer a good quantum number and the average energy of stationary state $\psi^{\pm}(l,\mathbf{k},t)$ can be calculated as

$$\bar{E}^{\pm}(l,\mathbf{k}) = \epsilon^{\pm}(l,\mathbf{k}) + \hbar\sum_n n\omega\left(|A_n^{\pm}(l,\mathbf{k})|^2 + |B_n^{\pm}(l,\mathbf{k})|^2\right) . \tag{22}$$

From this one obtains the geometric phase

$$\beta^{\pm}(l,\mathbf{k}) = 2\pi\sum_n n\left(|A_n^{\pm}(l,\mathbf{k})|^2 + |B_n^{\pm}(l,\mathbf{k})|^2\right) . \tag{23}$$

The energy uncertainty can be specified by the standard variance caused by the vibration

$$\begin{aligned}\Delta^{\pm}(l,\mathbf{k}) &\equiv \langle E^2\rangle - \langle E\rangle^2 \\ &= \hbar^2\omega^2\left\{\sum_n n^2\left(|A_n^{\pm}(l,\mathbf{k})|^2 + |B_n^{\pm}(l,\mathbf{k})|^2\right)\right. \\ &\quad \left. - \left[\sum_n n\left(|A_n^{\pm}(l,\mathbf{k})|^2 + |B_n^{\pm}(l,\mathbf{k})|^2\right)\right]^2\right\} .\end{aligned} \tag{24}$$

From the superposition principle any time-dependent state can be expressed as a linear combination of the Floquet states. But this linear combination is usually not a stationary state, i. e., the state could not return to its initial one with only a phase difference after a period of evolution. For such states the geometric phase cannot be defined. So in this paper we only consider the geometric phase for the Floquet states. Since for given valley and given momentum there are only two unknowns having the same phonon number n in Eqs. (1) and (17), the number of quasienergies within range of $\left[-\hbar\omega/2, \hbar\omega/2\right]$ is 2, corresponding to the particle and hole branches of the Dirac fermions with labels $l = 1$ and $l = 2$, respectively.

Now we begin to investigate the properties of stationary states in the periodic vibrations. We are interested in: (i) the geometric phases acquired by fermion states with various momenta in a period of vibrations; (ii) the deviation of the average energy from the Dirac dispersion relation and the energy uncertainty caused by the vibration. From (i) we can shed some light on the phase variation of the Dirac fermions and the possible dephasing mechanism in graphene. From (ii) we can see the essential effect of vibrations on the basic dispersion relation.

In Fig. 2 we plot the geometric phase acquired by fermion states in a period of vibrations versus the fermion momentum. There is a $\pi/2$ phase difference between the vibrations in the x and y directions and their amplitudes are the same, i. e., $\eta_x = \eta_y = \eta_0$. So this is a circular vibration of the relative coordinates of two sublattices as shown in Fig. 1. According to the Berry theorem, the adiabatic Berry phase is $\pm\pi$ if $|\mathbf{k}| < (\eta_0)/(t_0 l_0)$ and is zero if $|\mathbf{k}| > (\eta_0)/(t_0 l_0)$. We note that from the calculations beyond the adiabatic approximation the obtained geometric phase is still roughly equal to the Berry phase: for $|\mathbf{k}| \ll (\eta_0)/(t_0 l_0)$ it is nearly $\pm\pi$ and for $|\mathbf{k}| \gg (\eta_0)/(t_0 l_0)$ it is almost zero. However, now there appears a "chaotic" region around the border $|\mathbf{k}| = (\eta_0)/(t_0 l_0)$ where the geometric phase randomly oscillates in changing the momentum. This is a direct consequence of the non-adiabaticity, as in this region the adiabatic condition, that the energy difference between two branches is much larger than $\hbar\omega$ in the evolution path, is not satisfied. As can be seen from the inset of Fig. 2c, the width of this chaotic region is increased by increasing the frequency. The chaotic nature of the geometric phase reflects the phase uncertainty of the single-fermion states in this momentum region during vibrations. Such vibrations may exist in the graphene due to thermal excitations at a finite temperature, or due to the zero-point fluctuations even at very low temperatures. So such a fermion-lattice interaction can play roles of a dephasing mechanism for the single-fermion states. We also

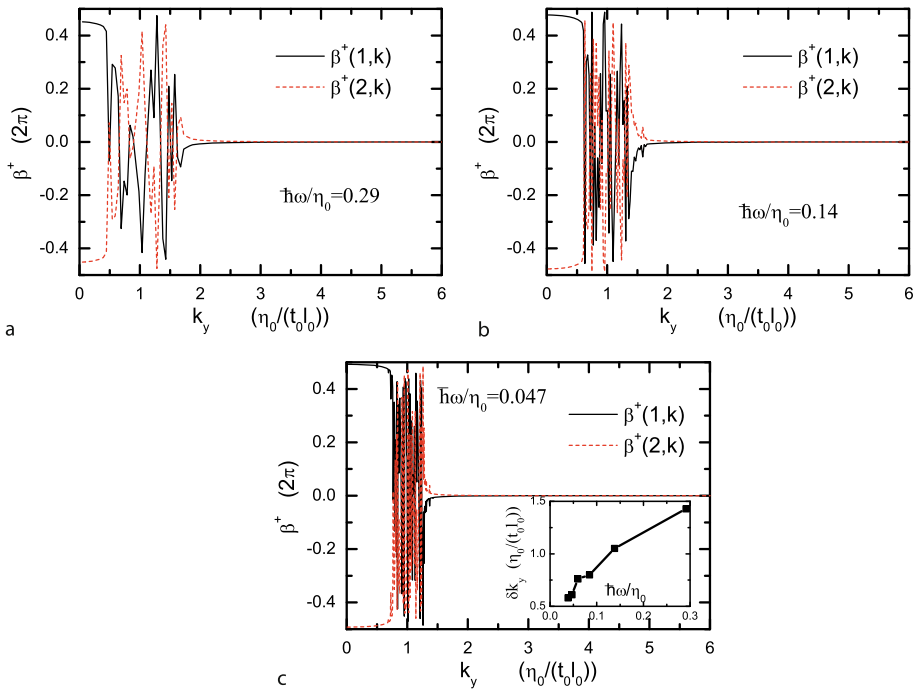

Geometric Phase and Related Phenomena in Quantum Nanosystems, Figure 2

Geometric phase acquired in a period of vibrations by Dirac fermions in particle and hole branches of valley "+" as a function of the fermion momentum. The parameters for vibrations are: $\eta_x = \eta_y = \eta_0, \alpha_x = 0$, and $\alpha_y = -\pi/2$. $k_x = 0$ and units of k_y are set to be $(\eta_0)/(t_0 l_0)$ so that the border between the $\pm\pi$ Berry phase and the zero Berry phase is at $k_y = 1$. The *inset* of c shows the width of the chaotic region δk_y as a function of the frequency

note that in the whole region, even including the chaotic region, the geometric phases acquired by fermions of the same momentum in two branches are opposite to each other. This phase compensation effect of two branches implies that this dephasing mechanism has no effect if the fermions in two branches are paired.

The above result indicates that the exchange of two branches leads to the sign reverse of the geometric phase. From the symmetry shown in Fig. 1 we can see that an equivalent operator which can also cause the sign reverse of the geometric phase is the reverse of the direction of the circular vibration from clockwise to counterclockwise or vice versa.

Another significant effect of vibrations is the uncertainty of fermion energies. As the single-particle energy is not a good quantum number, in Fig. 3 we plot the average energy of single-fermion states as a function of the momentum. From the comparison with the Dirac dispersion relation and with Fig. 2, the energy difference between two branches is enlarged near the Dirac point and for $|\mathbf{k}| \gg (\eta_0)/(t_0 l_0)$, where the geometric phase is near its adiabatic value, $\pm\pi$ or zero, while this energy difference shrinks in the chaotic region where the adiabatic approximation can not be used. Such opposite behaviors in these two regions imply that the failure of the adiabatic theory in the chaotic region has much more profound meaning than reflected from the values of geometric phase. Physically, in the chaotic region the oscillations of electrons between two branches are not able to follow the vibrations, this causes the loss of distinguishability of the two branches and the resultant states trend to take average energies in between, leading to a smaller energy spacing. On the contrary, in a quantum description the coupling of two states always enlarges their energy spacing. So one may expect that the chaotic region corresponds to a classical-like behavior and has a maximum uncertainty of single-fermion energies. To verify this we calculate the standard variance of Eq. (24) and show the results in Fig. 4. As expected, the standard variance exhibits a peak in the chaotic region, and the energy uncertainty increases by increasing the amplitudes of vibrations. Except for the peak in the chaotic region, the standard variance globally increases by increasing the momentum. Especially, the energy uncertainty becomes zero at the Dirac point for all the investigated vibration amplitudes. This originates from the fact that the the energy uncertainty is due to the oscil-

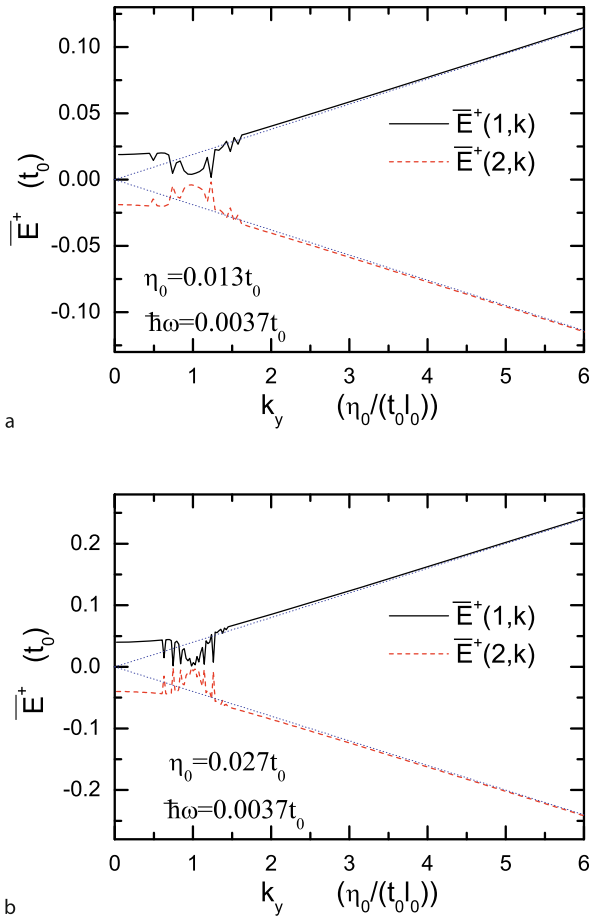

Geometric Phase and Related Phenomena in Quantum Nanosystems, Figure 3

Average energy of fermions in particle and hole branches of valley "+" as a function of the fermion momentum. The parameters for vibrations are $\eta_x = \eta_y = \eta_0$, $\alpha_x = 0$, and $\alpha_y = -\pi/2$. $k_x = 0$. The *blue dotted lines* represent eigenenergies without vibrations

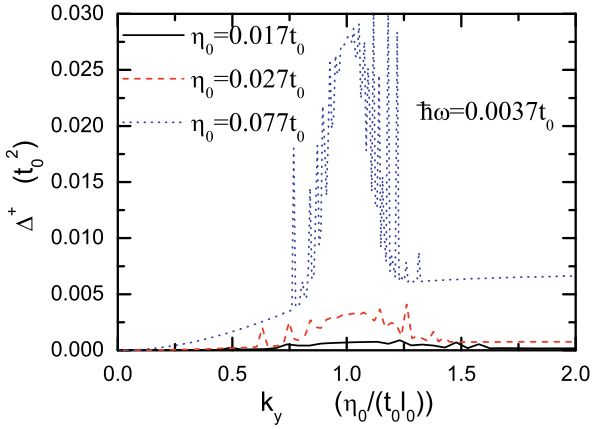

Geometric Phase and Related Phenomena in Quantum Nanosystems, Figure 4

Standard variance of single-fermion energy of valley "+" as a function of the fermion momentum. The parameters for vibrations are $\eta_x = \eta_y = \eta_0$, $\alpha_x = 0$, and $\alpha_y = -\pi/2$. $k_x = 0$. The standard variance is the same for particle and hole branches

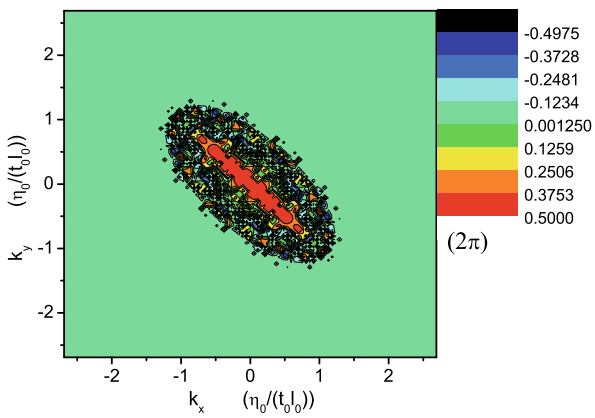

Geometric Phase and Related Phenomena in Quantum Nanosystems, Figure 5

Contour plot of geometric phase acquired by fermions in valley "+" and in the particle branch as a function of k_x and k_y. The parameters for vibrations are $\eta_x = \eta_y = \eta_0 = 0.078t_0$, $\hbar\omega = 0.0037t_0$, $\alpha_x = 0$, and $\alpha_y = 0.6$

lation of the fermion states between two branches, but at the Dirac point the two branches coincide, leading to a zero amplitude of the oscillation. This structure, including the peak and the globally increasing background with the momentum, may catch major features of linewidth distribution obtained from the angle-resolved photoemission spectroscopy in graphene [9].

The evolution path plays a crucial role on the distribution of adiabatic and chaotic regions in the momentum space. To show this in Fig. 5 we plot the geometric phase as a function of k_x and k_y for an elliptic path. The chaotic region forms an orbicular area with a nearly fixed width along a loop defined by $k_x = (\lambda d_y(t))/(t_0 l_0)$ and $k_y = -(\lambda d_x(t))/(t_0 l_0)$. The width depends on ω as shown in the inset of Fig. 2c. As a result, the adiabatic region with a $\pm\pi$ Berry phase is compressed by reducing the short axis of the ellipse, but the chaotic region with random geometric phases can exist even for a linear vibration.

Effect of Phase Uncertainty

The main features of the band structure in the graphene can be well described by a tight-binding Hamiltonian with one π orbital per site on the honeycomb lattice [32]:

$$H = \sum_n \epsilon_n a_n^\dagger a_n \sum_{\langle n,n'\rangle} + t_{nn'}(a_n^\dagger a_{n'} + a_{n'}^\dagger a_n)\,, \qquad (25)$$

where $a_n^\dagger(a_n)$ creates (annihilates) an electron at site n, ϵ_n is the energy level at site n, $\langle\ldots\rangle$ denotes the nearest-neighbor (NN) sites, and $t_{nn'}$ is hopping integral between NN sites n and n'. Here, the spin indices are not explicitly included. In the absence of disorder, $\epsilon_n = 0$ for all n and $t_{nn'} = t_0$ for all $\langle nn'\rangle$, this Hamiltonian leads to the four-component Dirac fermion dispersion relation near the Dirac point $E = 0$. The diagonal and off-diagonal disorder can be introduced by adopting random variables ϵ_n and $t_{nn'}$ satisfying distribution probabilities $P_\mathrm{d}(\epsilon_n)$ and $P_\mathrm{o}(t_{nn'})$, respectively. These probability functions define the strengths and the types of disorder in the system. The diagonal disorder represents the potential fluctuations due to impurities or due to the randomness on the substrate surface. A most commonly used distribution probability for the diagonal disorder is the square function

$$P_\mathrm{d}(\epsilon_n) = \begin{cases} \frac{1}{W}\,, & \text{for} \quad \frac{-W}{2} \le \epsilon_n \le \frac{W}{2}\,, \\ 0\,, & \text{otherwise}\,, \end{cases} \tag{26}$$

where W is the distribution width describing the strength of the disorder. For the OD disorder, there are different distribution functions describing different types of randomness. If we consider slight fluctuations of bond lengths around their average value due to lattice distortions, the following distribution is suitable

$$P_\mathrm{o1}(t_{nn'}) = \begin{cases} \frac{1}{\lambda}\,, & \text{for} \quad t_0 - \frac{\lambda}{2} \le t_{nn'} \le t_0 + \frac{\lambda}{2}\,, \\ 0\,, & \text{otherwise}\,, \end{cases} \tag{27}$$

where $\lambda(< 2|t_0|)$ is a measure of the bond-length fluctuations which make the NN hopping integrals randomly shift from t_0. In $P_\mathrm{o1}(t_{nn'})$ all the hopping integrals have the same sign as t_0, reflecting that only the changes of magnitudes of $t_{nn'}$ are considered. In some cases, however, $t_{nn'}$ not only fluctuate in magnitude, but also change in sign due to specific physical mechanisms. One of these possible mechanisms is the geometric phase which can easily produced for fermions in graphene since the Dirac point could play the roles of a degenerate point of two branches. We can assume that the opposite signs of $t_{nn'}$ are also randomly distributed in the space. Then the distribution of $t_{nn'}$ can be written as

$$P_\mathrm{o}(t_{nn'}) = sP_\mathrm{o1}(t_{nn'}) + (1-s)P_\mathrm{o1}(-t_{nn'})\,, \tag{28}$$

where $s(0 \le s \le 1)$ is a parameter describing how frequently the hopping integrals change sign in the space. When $s = 0$ or $s = 1$, all the hopping integrals have the same sign and $P_\mathrm{o}(t_{nn'})$ reduces to $P_\mathrm{o1}(t_{nn'})$. At $s = 0.5$ the frequency of the sign changes is maximal. In $P_\mathrm{o}(t_{nn'})$ we include both the magnitude disorder and sign disorder of the off-diagonal elements in the Hamiltonian. Below we will set t_0 as the energy units.

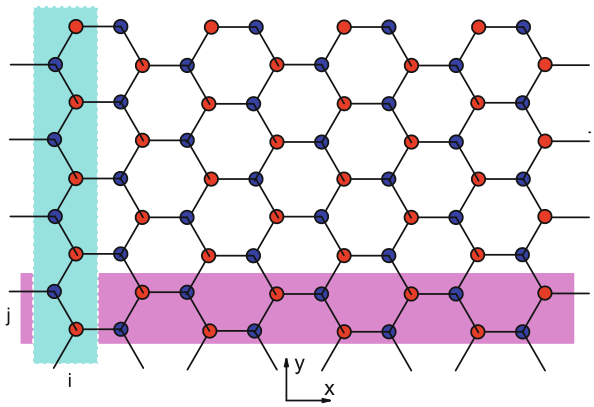

Geometric Phase and Related Phenomena in Quantum Nanosystems, Figure 6
Structure of graphene lattice. *Red* and *blue* sites correspond to sublattices $l = 1$ and $l = -1$, i and j are indices of columns (*light-cyan shaded*) and rows (*light-magenta shaded*), respectively

The wavefunction of a fermion can be written as

$$\psi = \sum_n c_n|n\rangle\,, \tag{29}$$

where c_n is the amplitude on site n. At a given energy E the amplitudes are given be the Schödinger equation $H\psi = E\psi$. The propagation properties of a wavefunction on a finite lattice can be given by a relation between amplitudes on both sides. This is usually formulated by the transfer-matrix method or by the Green function. We note that the graphene lattice can be divided into two sub-lattices as distinguished by red and blue sites in Fig. 6. From the tight-binding nature of the Hamiltonian and the structure shown in Fig. 6, it can be seen that for given E all the amplitudes can be iteratively calculated along the $x(y)$ direction if the amplitudes in the light-cyan shaded column (light-magenta shaded row) are given. From this we can establish a transfer matrix along the x or y direction for the graphene. For this purpose we identify a site n with three integers, $n \equiv (i, j, l)$, where i and j index the column and row, respectively, and $l = \pm 1$ specifies the sublattice which the site n belongs to. In this notation the iterative relation for the amplitudes along the x direction can be derived from the Schödinger equation as [38]

$$c_{i+1,j,-1} = \frac{(E-\epsilon_{i,j,1})c_{i,j,1} - t_{i,j,1;i,j,-1}c_{i,j,-1} - t_{i,j,1;i,j+(-1)^i,-1}c_{i,j+(-1)^i,-1}}{t_{i,j,1;i+1,j,-1}}\,, \tag{30}$$

$$c_{i+1,j,1} = \frac{\begin{array}{c}(E-\epsilon_{i+1,j,-1})c_{i+1,j,-1}\\ -t_{i+1,j,-1;i,j,1}c_{i,j,1}\\ -t_{i+1,j,-1;i+1,j+(-1)^i,1}c_{i+1,j+(-1)^i,1}\end{array}}{t_{i+1,j,-1;i+1,j,1}} . \tag{31}$$

For a system with finite width in the y direction, one can define vectors for the columns by $\mathbf{u}_{i,l} \equiv (c_{i,1,l}, c_{i,2,l}, \ldots, c_{i,M,l})^{\mathrm{T}}$, where M is the number of cells in a column. Then Eqs. (30) and (31) can be written in a matrix form:

$$\mathbf{u}_{i+1,-1} = \hat{T}_{1,i}\mathbf{u}_{i,1} + \hat{T}_{2,i}\mathbf{u}_{i,-1} , \tag{32}$$

$$\mathbf{u}_{i,1} = \hat{T}_{3,i}\mathbf{u}_{i+1,1} + \hat{T}_{4,i}\mathbf{u}_{i+1,-1} , \tag{33}$$

where the elements of the matrices are given by

$$\{\hat{T}_{1,i}\}_{j,j'} = \frac{E-\epsilon_{i,j,1}}{t_{i,j,1;i+1,j,-1}}\delta_{j,j'} , \tag{34}$$

$$\{\hat{T}_{2,i}\}_{j,j'} = -\frac{t_{i,j,1;i,j,-1}}{t_{i,j,1;i+1,j,-1}}\delta_{j,j'} - \frac{t_{i,j,1;i,j+(-1)^i,-1}}{t_{i,j,1;i+1,j,-1}}\delta_{j+(-1)^i,j'} , \tag{35}$$

$$\{\hat{T}_{3,i}\}_{j,j'} = -\frac{t_{i+1,j,-1;i+1,j,1}}{t_{i+1,j,-1;i,j,1}}\delta_{j,j'} - \frac{t_{i+1,j,-1;i+1,j+(-1)^i,1}}{t_{i+1,j,-1;i,j,1}}\delta_{j+(-1)^i,j'} , \tag{36}$$

$$\{\hat{T}_{4,i}\}_{j,j'} = \frac{E-\epsilon_{i+1,j,-1}}{t_{i+1,j,-1;i,j,1}}\delta_{j,j'} . \tag{37}$$

Here, one may use the periodic or open boundary condition at the ends $j = 1$ and $j = M$. We have the iterative relation for the whole column

$$\begin{pmatrix} \mathbf{u}_{i+1,-1} \\ \mathbf{u}_{i+1,1} \end{pmatrix} = \hat{T}_i \begin{pmatrix} \mathbf{u}_{i,-1} \\ \mathbf{u}_{i,1} \end{pmatrix} , \tag{38}$$

where the transfer matrix is

$$\hat{T}_i = \begin{pmatrix} \hat{T}_{2,i} , & \hat{T}_{1,i} \\ -\hat{T}_{3,i}^{-1}\hat{T}_{4,i}\hat{T}_{2,i} , & \hat{T}_{3,i}^{-1} - \hat{T}_{3,i}^{-1}\hat{T}_{4,i}\hat{T}_{1,i} \end{pmatrix} . \tag{39}$$

If there are L columns in the x-direction, the total transfer matrix which gives the relation between amplitudes at two end columns is

$$\hat{T} = \prod_{i=1}^{L} \hat{T}_{L-i+1} . \tag{40}$$

If we want to investigate the propagation properties in the y-direction, we have to establish the transfer matrix for the rows. The relation between two adjacent rows is written as

$$\begin{pmatrix} \mathbf{v}_{j+1,-1} \\ \mathbf{v}_{j+1,1} \end{pmatrix} = \hat{V}_j \begin{pmatrix} \mathbf{v}_{j,-1} \\ \mathbf{v}_{j,1} \end{pmatrix} , \tag{41}$$

where $\mathbf{v}_{j,l}$'s are vectors denoting rows with length M, $\mathbf{v}_{j,l} \equiv (c_{1,j,l}, c_{2,j,l}, \ldots, c_{M,j,l})^{\mathrm{T}}$, and the elements of the transfer matrix are

$$\begin{aligned}\{\hat{V}_j\}_{i,l;i',l'} = \delta_{l,-(-1)^i}\Big\{ & C_{i,j}^{(1)}\delta_{-l,l'}\delta_{i,i'} - C_{i,j}^{(2)}\delta_{l,l'}\delta_{i,i'} \\ & - C_{i,j}^{(3)}\delta_{l,l'}\delta_{i+(-1)^i,i'} + C_{i,j}^{(4)}\Big(C_{i,j}^{(1)}\delta_{l,l'}\delta_{i,i'} - C_{i,j}^{(2)}\delta_{l,-l'}\delta_{i,i'} \\ & - C_{i,j}^{(3)}\delta_{l,-l'}\delta_{i+(-1)^i,i'}\Big) - C^{(5)}\delta_{l,l'}\delta_{l,-(-1)^i}\delta_{i,i'} \\ & - C^{(6)}\Big(C_{i-(-1)^i,j}^{(1)}\delta_{l,-l'}\delta_{i-(-1)^i,i'} \\ & - C_{i-(-1)^i,j}^{(2)}\delta_{l,l'}\delta_{i-(-1)^i,i'} \\ & - C_{i-(-1)^i,j}^{(3)}\delta_{l,l'}\delta_{i-(-1)^i\cdot 2,i'}\Big)\Big\} ,\end{aligned} \tag{42}$$

where

$$\begin{aligned} C_{i,j}^{(1)} &= \frac{E-\epsilon_{i,j,(-1)^i}}{t_{i,j,(-1)^i;i,j+1,-(-1)^i}} , \\ C_{i,j}^{(2)} &= \frac{t_{i,j,(-1)^i;i,j,-(-1)^i}}{t_{i,j,(-1)^i;i,j+1,-(-1)^i}} , \\ C_{i,j}^{(3)} &= \frac{t_{i,j,(-1)^i;i+(-1)^i,j,-(-1)^i}}{t_{i,j,(-1)^i;i,j+1,-(-1)^i}} , \\ C_{i,j}^{(4)} &= \frac{E-\epsilon_{i,j+1,-(-1)^i}}{t_{i,j+1,-(-1)^i;i,j+1,(-1)^i}} , \\ C_{i,j}^{(5)} &= \frac{t_{i,j+1,-(-1)^i;i,j,(-1)^i}}{t_{i,j+1,-(-1)^i;i,j+1,(-1)^i}} , \\ C_{i,j}^{(6)} &= \frac{t_{i,j+1,-(-1)^i;i-(-1)^i,j+1,(-1)^i}}{t_{i,j+1,-(-1)^i;i,j+1,(-1)^i}} . \end{aligned}$$

The total transfer matrix for L rows is

$$\hat{V} = \prod_{i=1}^{L} \hat{V}_{L-i+1} . \tag{43}$$

For a finite $M \times L$ system, there are M propagating channels along the length L. The Hermitian matrix, $(\hat{T}^\dagger\hat{T})^{1/2L}$ or $(\hat{V}^\dagger\hat{V})^{1/2L}$, has $2M$ eigen values. These $2M$ eigen values should be positive and come in inverse pairs due to the unitarity of the matrix [5]. The M positive logarithms of the eigen values, denoted as γ_i in an ascending order with

index i, are Lyapunov exponents (LE) reflecting the exponential decay of the corresponding channels. Obviously, the logarithms of other M eigen values are the negatives of γ_i. The LEs will approach 0 when the disorder vanishes. The localization length is defined as the reciprocal of the smallest LE $\xi \equiv 1/\gamma_1$. The localization length plays a crucial role in the localization theory.

The zero-temperature conductance g of a rectangular $M \times L$ system in units of $2e^2/h$, with two spins being taken into account, can be evaluated using the M-channel Landauer formula [13,14]

$$g = \mathrm{Tr}(\hat{t}^\dagger \hat{t}) \,, \tag{44}$$

where the $M \times M$ transmission matrix $\hat{t}$ describes the transmission of electrons from one lead to the other in the longitudinal direction. Here we assume that the leads are connected to all the M channels and they are perfect metal with band width much larger than that of the graphene. Formula (44) can be expressed with the LEs as [5]

$$g = \sum_{i=1}^{M} \frac{1}{\cosh^2(\gamma_i L)} \,. \tag{45}$$

Numerically, LEs can be calculated by using the standard method of Gram–Schmidt re-orthonormalization after each a few, say, ten steps of multiplication of the transfer matrices [20]. This is equivalent to the diagonalization of $\hat{T}^\dagger \hat{T}$ or $\hat{V}^\dagger \hat{V}$, but avoids terrible overflow and loss of precision on the computer. On the other hand, although the calculations of the transfer matrices along the x and y directions are certainly different, there is no essential difference in the calculated results between these two directions. This is owing to the fact that the dispersion relation is conic around the Dirac point which is isotropic in the x and y directions. Below we will only present the results obtained for the transmission along the y direction.

The quantum transport processes could be drastically changed if the phases of wavefunctions become uncertain. This may happen due to the Berry phase which can be acquired by electrons during a cyclic evolution around a conic point. In the graphene the Berry phase can be easily created owing to the Dirac fermion dispersion relation. The attached Berry phase of π corresponds to sign changes of the wavefunction during the motion of electrons. If the sign changes randomly happen during the spatial motion, they can be modeled by a random distribution of positive and negative signs of hopping integrals. In probability $P_o(t_{nn'})$ of Eq. (28), we include both the sign randomness and the magnitude randomness of the hopping integrals which are specified by parameters s and λ, respectively.

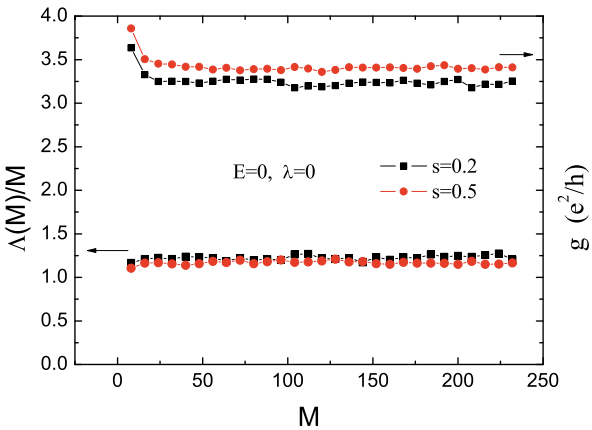

Geometric Phase and Related Phenomena in Quantum Nanosystems, Figure 7

Scaling behavior of the rescaled localization length and the conductance of an $M \times M$ system in the presence of sign randomness of hopping integrals obeying probability $P_o(t_{nn'})$ with different values of s

At first we set $\lambda = 0$ and focus on the effect of sign randomness. In Fig. 7 we plot the scaling behavior of the rescaled localization length and the conductance of an $M \times M$ system for different values of s. Similarly to the case of off-diagonal disorder, the states at the Dirac point in the presence of the sign randomness are also delocalized in the sense of finite-size scaling analysis. Effects of the sign randomness on the rescaled localization length and on the conductance are opposite: $\Lambda(M)/M$ in $s = 0.5$ is smaller than that in $s = 0.2$, but g in $s = 0.5$ is larger than that in $s = 0.2$. This is strange because $\Lambda(M)/M$ corresponds to the channel with the largest localization length which usually provides the leading contribution to the conductance. The only explanation for this is that the other channels also give significant contributions in the case of sign randomness. An expectable consequence of this is a new shape dependence of the conductivity, since the number of the tunneling channels is no longer restricted.

The shape dependence of the conductivity can be investigated by calculating the conductance of $M \times L$ rectangular samples. The obtained conductivity as a function of the ratio M/L for various sizes is shown in Fig. 8. Except for small values of M/L, the conductivity has values near $(4e^2)/h$, independent of M/L and M, in consistence with the experimental findings. Surprisingly, this value is even larger than $(4e^2)/(\pi h)$ obtained in the ballistic graphene [18] in spite of the disorder introduced by the sign randomness of hopping integrals. To get insight into this anomalous feature, we denote the average spacing between successive pairs of Lyapunov exponents as $\delta\gamma$,

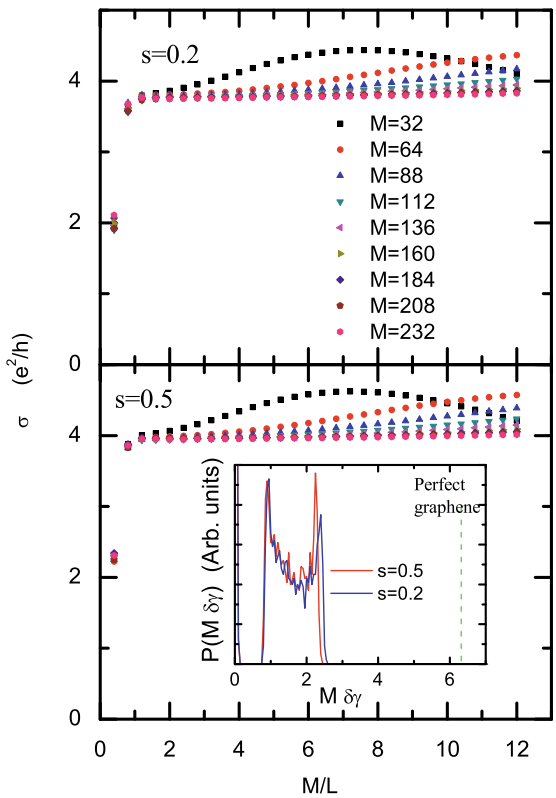

Geometric Phase and Related Phenomena in Quantum Nanosystems, Figure 8

Calculated conductivity as a function of ratio *M/L* for rectangular sheets with different widths *M*. *Inset in the lower panel*: Distribution function *P*(*Mδγ*) of *Mδγ*. The *green dashed line* indicates the *δ*-function distribution at 2*π* in the case of perfect sheet

i. e., $\gamma_{2i+1} - \gamma_{2i-1} \sim \gamma_{2i+2} - \gamma_{2i} \sim \delta\gamma$. Notice that in this case the Lyapunov exponents are grouped into pairs, corresponding to two equivalent valleys. By keeping only the leading exponential term of $\cosh^2(\gamma_i L)$, Eq. (45) can be approximately rewritten as

$$g \approx \sum_{i=1}^{M/2} 8\mathrm{e}^{-2\gamma_{2i-1}L} \approx \frac{8\mathrm{e}^{-2\gamma_1 L}(1-\mathrm{e}^{-(M+2)L\delta\gamma})}{1-\mathrm{e}^{-2L\delta\gamma}}\,, \quad (46)$$

in units of $(2\mathrm{e}^2)/h$. In a perfect system of width M and using the periodic boundary condition, the spacing $\delta\gamma$ for states at $E = 0$ is $\delta\gamma = 2\pi/M$ due to the Dirac fermion dispersion relation [18]. One may reasonably suppose that $\delta\gamma \propto 1/M$ is still held in the case of disorder. Then, for $M \gg L$, the conductivity of an $M \times L$ rectangular sample can be calculated as

$$\sigma = \frac{gL}{M} \approx \frac{4\mathrm{e}^{-2\gamma_1 L}}{\delta\gamma M}\,. \quad (47)$$

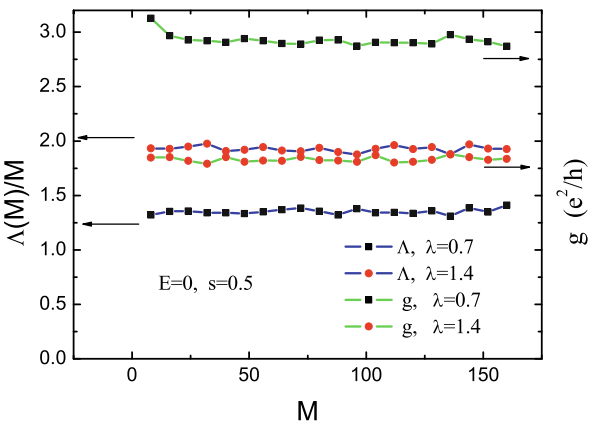

Geometric Phase and Related Phenomena in Quantum Nanosystems, Figure 9

Scaling behavior of the rescaled localization length and the conductance of an *M* × *M* system in the presence of both sign randomness and magnitude randomness of hopping integrals obeying probability $P_0(t_{nn'})$

So for the perfect lattice $\sigma = 2/\pi$ in units of $(2\mathrm{e}^2)/h$ [18]. In the case of sign randomness, the values of $M\delta\gamma$ are in the range [0.9, 2.5] as shown in the inset of Fig. 8. Together with the factor $\mathrm{e}^{-2\gamma_1 L}$, the obtained conductivity may be π times of that of the perfect lattice. Thus, the π factor difference in the conductivity is not trivial and reflects an essential change of the fermion properties due to the phase uncertainty.

It is interesting to investigate how the localization behavior changes in the presence of the magnitude randomness of hopping integrals. The scaling behaviors of the rescaled localization length and the conductance of an $M \times M$ system with both nonzero s and λ are shown in Fig. 9. It can be seen that the states at $E = 0$ are still delocalized and the opposite effects on the rescaled localization length and on the conductance are more evident than those shown in Fig. 7. In Fig. 10 we display the shape dependence of the conductivity of the $M \times L$ rectangular samples for $s = 0.5$ and different nonzero values of λ. Similarly to the case of $\lambda = 0$, the conductivity is still shape independent except for small values of M/L. This means that the sign-randomness induced shape independence of the conductivity is robust against the warping or corrugation disorder which may cause the randomness of magnitudes of hopping integrals. It can also be seen that the value of the conductivity is slightly reduced from $(4\mathrm{e}^2)/h$ by increasing the warping or corrugation disorder. For reasonable strength of the warping disorder ($\lambda < 1$), however, the conductivity is above $(3\mathrm{e}^2)/h$, in the range of measured values in experiments.

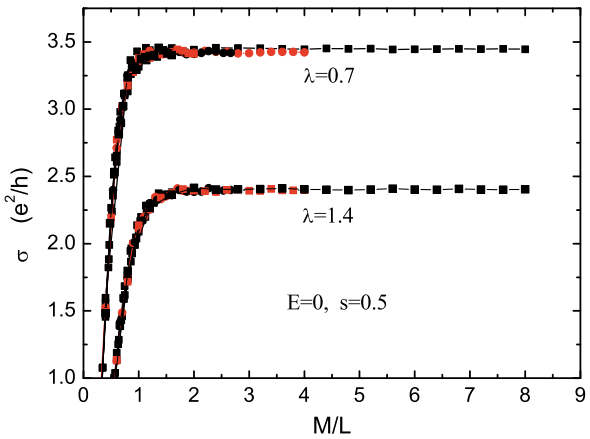

Geometric Phase and Related Phenomena in Quantum Nanosystems, Figure 10

Calculated conductivity as a function of ratio *W*/*L* for rectangular sheets with different widths *M* in the presence of both sign randomness and magnitude randomness of hopping integrals

Future Directions

In physical applications, one often encounters Hamiltonian operators depending on several external parameters, thus potentially affected by Berry phases [38]. The Berry phase is observed by following a nontrivial loop in the parameters space, and performing some kind of interference between the state prior and after the loop. Standard examples are the parallel transport of a vector on a sphere, the Aharonov–Bohm effect, and analogous neutron interference experiments; q indicates some internal degrees of freedom of a system are approximately treated as "slow", classical variables, affecting adiabatically the "fast" quantum dynamics of the other degrees of freedom. Of course, eventually the true quantum nature of the slow variables must be taken into account, as a further crucial step. The standard example for this is the Born–Oppenheimer separation of the "fast" quantum motion of electrons from the "slow" motion of nuclei in molecules and nanocrystals: the nuclear coordinates appear as external parameters defining the potential in which the electrons move. In turn, the electronic ground state energy enters in the potential energy for the adiabatic motion of the nuclei. When the latter is quantized, it gives rise to the vibrational/phononic states observed by spectroscopy. Here we shall be mostly concerned with the latter case. Berry phase for real Hamiltonians A physically relevant and rather simple case is that of the eigensystem of a real Hamiltonian operator H depending on a set of external parameters q. The eigenstates of a real operator may always be chosen to be real, at any q-point. It is straightforward to check that continuous real eigenstates realize parallel transport.

So there are following future directions of investigation on the geometric phase in nanostructures:

(i) The transition from quantum to classical. It is well known that the phases of wave functions are essential for the manifestation of the quantum natures in nanostructures. If the fast degrees of freedom acquire geometric phases from random evolutions of slow degrees of freedom, i. e., the thermally excited vibrations of nuclei, the phase of the fast degrees of freedom becomes uncertain, leading to the transition from the quantum to classical.

(ii) Investigation on the dephasing mechanism in nanostructures. Nanostructures are regarded as potential candidates for the quantum computing and quantum information processing. A prerequisite for such applications is the quantum coherence time as long as possible. Various dephasing mechanisms may prevent these applications as they destroy the quantum coherence of the target system. As one of the dephasing mechanisms the investigation of geometric phase in nanostructures is of particular importance.

(iii) Designing of novel quantum interference devices employing the geometric phase in nanostructures. The geometric phase crucially depends on the geometry of the nanostructure. This provides a new way in designing interference devices by adjusting the shapes of nanostructures.

(iv) Investigation of geometric phase in quantum many-body low-dimensional and nano systems. A many-body system is a complicated mixture of various degrees of freedom. Among them there may be some faster degrees of freedom which are crucially affected by geometric phases created from the slower degrees of freedom. Compared with the one-body nature of the fast degrees of freedom previously investigated, the geometric phases in a many-body fast system are much more complicated and need future investigations.

Acknowledgments

This work was supported by the State Key Programs for Basic Research of China (2005CB623605 and 2006CB921803), and by National Foundation of Natural Science in China Grant Nos. 10474033 and 60676056.

Bibliography

Primary Literature

1. Aharonov Y, Anandan A (1987) Phase change during a cyclic quantum evolution. Phys Rev Lett 58:1593–1596

2. Aleiner IL, Efetov KB (2006) Effect of disorder on transport in graphene. Phys Rev Lett 97:236801
3. Altland A (2006) Low-energy theory of disordered graphene. Phys Rev Lett 97:236802
4. Aronov AG, Lyanda-Geller YB (1993) Spin-orbit Berry phase in conducting rings. Phys Rev Lett 70:343–346
5. Beenakker CWJ (1997) Random-matrix theory of quantum transport. Rev Mod Phys 69:731–808
6. Berger C, Song Z, Li T, Li X, Ogbazghi AY, Feng R, Dai Z, Marchenkov AN, Conrad EH, First PN, de Heer WA (2004) Ultrathin epitaxial graphite: 2D electron gas properties and a route toward graphene-based nanoelectronics. J Phys Chem B 108:19912–19916
7. Berger C, Song Z, Li X, Wu X, Brown N, Naud C, Mayou D, Li T, Hass J, Marchenkov AN, Conrad EH, First PN, de Heer WA (2006) Electronic confinement and coherence in patterned epitaxial graphene. Science 312:1191–1196
8. Berry MV (1984) Quantal phase factors accompanying adiabatic changes. Proc R Soc Lond A 392:45–57
9. Bostwick A, Ohta T, Seyller T, Horn K, Rotenberg E (2007) Quasiparticle dynamics in graphene. Nat Phys 3:36–40
10. Calvert P (1992) Strength in disunity. Nature 357:365–366
11. Capozza R, Giuliano D, Lucignano P, Tagliacozzo A (2005) Quantum interference of electrons in a ring: tuning of the geometrical phase. Phys Rev Lett 95:226803
12. Chang LL, Ploog K (1985) Molecular beam epitaxy and heterostructures. Martinus Nijhoff, Dordrecht
13. Economou EN, Soukoulis CM (1981) Static conductance and scaling theory of localization in one dimension. Phys Rev Lett 46:618–621
14. Fisher DS, Lee PA (1981) Phys Rev B 23:R6851–6854
15. Goldhaber-Gordon D, Shtrikmna H, Mahalu D, Abusch-Magder D, Kastner MA (1998) Kondo effect in a single-electron transistor. Nature (London) 391:156–159, Cronenwett SM, Oosterkamp TH, Kouwenhoven LP (1998) A tunable Kondo effect in quantum dots, Science 281:540–544
16. Ham FS (1987) Berry's geometrical phase and the sequence of states in the Jahn–Teller effect. Phys Rev Lett 58:725–728
17. Iijima S (1991) Helical microtubules of graphitic carbon. Nature 354:56–58
18. Katsnelsona MI (2006) Eur Phys J B 51:157–160
19. Khveshchenko DV (2006) Electron localization properties in graphene. Phys Rev Lett 97:036802
20. MacKinnon A, Kramer B (1981) One-parameter scaling of localization length and conductance in disordered systems. Phys Rev Lett 47:1546–1549
21. McCann E, Kechedzhi K, Fal'ko VI, Suzuura H, Ando T, Altshuler BL (2006) Weak-localization magnetoresistance and valley symmetry in graphene. Phys Rev Lett 97:146805
22. Morozov SV, Novoselov KS, Katsnelson MI, Schedin F, Ponomarenko LA, Jiang D, Geim AK (2006) Strong suppression of weak localization in graphene. Phys Rev Lett 97:016801
23. Morpurgo AF, Guinea F (2006) Intervalley scattering, long-range disorder, and effective time-reversal symmetry breaking in graphene. Phys Rev Lett 97:196804
24. Nomura K, MacDonald AH (2007) Quantum transport of massless Dirac fermions. Phys Rev Lett 98:076602
25. Novoselov KS et al (2005) Two-dimensional gas of massless Dirac fermions in graphene. Nature (London) 438:197–200
26. Novoselov KS, Geim AK, Morozov SV, Jiang D, Zhang Y, Dubonos SV, Grigorieva IV, Firsov AA (2004) Electric field effect in atomically thin carbon films. Science 306:666–669
27. Pereira VM, Guinea F, dos Santos JMBL, Peres NMR, Neto AHC (2006) Disorder induced localized states in graphene. Phys Rev Lett 96:036801
28. Qian TZ, Su ZB (1994) Spin-orbit interaction and Aharonov–Anandan phase in mesoscopic rings. Phys Rev Lett 72:2311–2314
29. Reed MA, Randall JN, Aggarwal RJ, Matyi RJ, Moore TM, Wetsel AE (1988) Observation of discrete electronic states in a zero-dimensional semiconductor nanostructure. Phys Rev Lett 60:535–537
30. Sambe H (1973) Steady states and quasienergies of a quantum-mechanical system in an oscillating field. Phys Rev A 7:2203–2213
31. Simon B (1983) Holonomy, the quantum adiabatic theorem, and Berry's phase. Phys Rev Lett 51:2167–2170
32. Slonczewski JC, Weiss PR (1958) Band structure of graphite. Phys Rev 109:272–279
33. Tersoff J, Tromp RM (1993) Shape transition in growth of strained islands: spontaneous formation of quantum wires. Phys Rev Lett 70:2782–2785
34. Yacoby A, Heiblum M, Mahalu D, Shtrikman H (1995) Coherence and phase sensitive measurements in a quantum dot. Phys Rev Lett 74:4047–4050
35. Zhang Y, Tan YW, Stormer HL, Kim P (2005) Experimental observation of the quantum Hall effect and Berry's phase in graphene. Nature (London) 438:201–204
36. Ziegler K (2006) Robust transport properties in graphene. Phys Rev Lett 97:266802
37. Xiong SJ, Xiong Y (2007) Vibration-induced non-adiabatic geometric phase and energy uncertainty of fermions in graphene. Euro Phys Letters 80(6),60008438
38. Xiong SJ, Xiong Y (2007) Anderson localization of electron states in graphene in different types of disorder. Phys Rev B 76, 214204

Books and Reviews

Bohm A et al (2003) The geometric phase in quantum systems: foundations, mathematical concepts, and applications in molecular and condensed matter physics. Springer, Berlin

Mead CA (1992) The geometric phase in molecular systems. Rev Mod Phys 64:51–85

Yarkony DR (1996) Diabolical conical intersections. Rev Mod Phys 68:985–1013

Glasses and Aging, A Statistical Mechanics Perspective on

Ludovic Berthier[1], Giulio Biroli[2]

[1] Laboratoire des Colloïdes, Verres et Nanomatériaux, Université Montpellier II and CNRS, Montpellier, France

[2] CEA, DSM, Institut de Physique Théorique, IPhT, CNRS, MPPU, URA2306, Saclay, Gif-sur-Yvette, France

Article Outline

Glossary
Definition of the Subject
Phenomenology
Taxonomy of 'Glasses' in Science
Numerical Simulations
Dynamic Heterogeneity
Some Theory and Models
Aging and Off-equilibrium
Future Directions
Acknowledgments
Bibliography

Glossary

In this preliminary section, a few concise definitions of the most important concepts discussed in this article are given.

Glass transition For molecular liquids, the glass transition denotes a crossover from a viscous liquid to an amorphous solid. Experimentally, the crossover takes place at the glass temperature, T_g, conventionally defined as the temperature where the liquid's viscosity reaches the arbitrary value of 10^{12} Pas. The glass transition more generally applies to many different condensed matter systems where a crossover or, less frequently, a true phase transition, takes place between an ergodic phase and a frozen, amorphous glassy phase.

Aging In the glass phase, disordered materials are characterized by relaxation times that exceed common observation timescales, so that a material quenched in its glass phase never reaches equilibrium (neither a metastable equilibrium). It exhibits instead an aging behaviour during which its physical properties keep evolving with time.

Dynamic heterogeneity Relaxation spectra of dynamical observables, e. g. the dynamical structure factor, are very broad in supercooled liquids. This is associated to a spatial distribution of timescales: at any given time, different regions in the liquid relax at different rates. Since the supercooled liquid is ergodic, slow regions eventually become fast, and vice versa. Dynamic heterogeneity refers to the existence of these non-trivial spatio-temporal fluctuations in the local dynamical behaviour, a phenomenon observed in virtually all disordered systems with slow dynamics.

Effective temperature An aging material relaxes very slowly, trying (in vain) to reach its equilibrium state. During this process, the system probes states that do not correspond to thermodynamic equilibrium, so that its thermodynamic properties can not be rigorously defined. Any practical measurement of its temperature becomes a frequency-dependent operation. A 'slow' thermometer tuned to the relaxation timescale of the aging system measures an effective temperature corresponding to the ratio between spontaneous fluctuations (correlation) and linear response (susceptibility). This corresponds to a generalized form of the fluctuation-dissipation theorem for off-equilibrium materials.

Frustration Impossibility of simultaneously minimizing all the interaction terms in the energy function of the system. Frustration might arise from quenched disorder (as in spin glass models), from competing interactions (as in geometrically frustrated magnets), or from competition between a 'locally preferred order', and global, e. g. geometric, constraints (as in hard spheres packing problems).

Definition of the Subject

Glasses belong to a well-known state of matter: we easily design glasses with desired mechanical or optical properties on an industrial scale, they are widely present in our daily life. Yet, a deep microscopic understanding of the glassy state of matter remains a challenge for condensed matter physicists [6,67]. Glasses share similarities with crystalline solids (they are both mechanically rigid), but also with liquids (they both have similar disordered structures at the molecular level). It is mainly this mixed character that makes them fascinating even to non-scientists.

A glass can be obtained by cooling the temperature of a liquid below its glass temperature, T_g. The quench must be fast enough that the more standard first order phase transition towards the crystalline phase is avoided. The glass 'transition' is not a thermodynamic transition at all, since T_g is only empirically defined as the temperature below which the material has become too viscous to flow on a 'reasonable' timescale (and it is hard to define the word 'reasonable' in any reasonable manner). Therefore, T_g does not play a fundamental role, as a phase transition temperature would. It is simply the temperature below which the material looks solid. When quenched in the glass phase below T_g, liquids slowly evolve towards an equilibrium state they cannot reach on experimental timescales. Physical properties are then found to evolve slowly with time in far from equilibriun states, a process known as 'aging' [152].

Describing theoretically and quantifying experimentally the physical mechanisms responsible for the viscosity increase of liquids approaching the glass transition and for aging phenomena below the glass transition certainly stand as central open challenges in condensed mat-

ter physics. Since statistical mechanics aims at understanding the collective behaviour of large assemblies of interacting objects, it comes as no surprise that it is a central tool in that field. We shall therefore summarize the understanding gained from statistical mechanics perspectives into the problem of glasses and aging.

The subject has quite broad implications. A material is said to be 'glassy' when its typical relaxation timescale becomes of the order of, and often much larger than, the typical duration of an experiment or a numerical simulation. With this generic definition, a large number of systems can be considered as glassy materials [173]. One can be interested in the physics of liquids (window glasses are then the archetype), in 'hard' condensed matter (for instance type II superconductors in the presence of disorder such as high-T_c superconducting materials), charge density waves or spin glasses, 'soft' condensed matter with numerous complex fluids such as colloidal assemblies, emulsions, foams, but also granular materials, proteins, etc. All these materials exhibit, in some part of their phase diagrams, some sort of glassy dynamics characterized by a very rich phenomenology with effects such as aging, hysteresis, creep, memory, effective temperatures, rejuvenation, dynamic heterogeneity, non-linear response, etc.

This long list explains why this area of research has received increasing attention from physicists in the last two decades. 'Glassy' topics now go much beyond the physics of simple liquids (glass transition physics) and models and concepts developed for one system often find applications elsewhere in physics, from algorithmics to biophysics [55]. Motivations to study glassy materials are numerous. Glassy materials are everywhere around us and therefore obviously attract interest beyond academic research. At the same time, the glass conundrum provides theoretical physicists with deep fundamental questions since classical tools are sometimes not sufficient to properly account for the glass state. Moreover, simulating in the computer the dynamics of microscopically realistic material on timescales that are experimentally relevant is not an easy task, even with modern computers.

Studies on glassy materials constitute an exciting research area where experiments, simulations and theoretical calculations can meet, where both applied and fundamental problems are considered. How can one observe, understand, and theoretically describe the rich phenomenology of glassy materials? What are the fundamental quantities and concepts that emerge from these descriptions?

The outline of the article is as follows. In Sect. "Phenomenology" the phenomenology of glass-forming liquids is discussed. In Sect. "Taxonomy of 'Glasses' in Science" other type of glasses are described, in particular colloids and granular materials. It is then described how computer simulations have provided deep insights into the glass problem in Sect. "Numerical Simulations". The issue of dynamic heterogeneity is tackled in Sect. "Dynamic Heterogeneity". The main theoretical perspectives currently available in the field are then summarized in Sect. "Some Theory and Models". Aging and off-equilibrium phenomena occupy Sect. "Aging and Off-equilibrium". Finally, issues that seem important for future research are discussed in Sect. "Future Directions".

Phenomenology

Basic Facts

A vast majority of liquids (molecular liquids, polymeric liquids, etc.) form a glass if cooled fast enough in order to avoid the crystallisation transition [6]. Typical values of cooling rate in laboratory experiments are 0.1–100 K/min. The metastable phase reached in this way is called 'supercooled phase'. In this regime the typical timescales increase in a dramatic way and they end up to be many orders of magnitudes larger than microscopic timescales at T_g, the glass transition temperature.

For example, around the melting temperature T_m, the typical timescale τ_α on which density fluctuations relax, is of the order of $\sqrt{ma^2/K_B T}$, which corresponds to few picoseconds (m is the molecular mass, T the temperature, K_B the Boltzmann constant and a a typical distance between molecules). At T_g, which as a rule of thumb is about $2/3T_m$, the typical timescale has become of the order of 100 s, i. e. 14 orders of magnitude larger! This phenomenon is accompanied by a concomitant increase of the shear viscosity η. This can be understood by a simple Maxwell model in which η and τ_α are related by $\eta = G_\infty \tau_\alpha$, where G_∞ is the instantaneous (elastic) shear modulus which does not vary considerably in the supercooled regime. In fact, viscosities at the glass transition temperature are of the order of 10^{12} Pas. In order to grasp how viscous this is, recall that the typical viscosity of water at ambient temperature is of the order of 10^{-2} Pas. How long would one have to wait to drink a glass of water with a viscosity 10^{14} times larger?

As a matter of fact, the temperature at which the liquid does not flow anymore and becomes an amorphous solid, called a 'glass', is protocol dependent. It depends on the cooling rate and on the patience of the people carrying out the experiment: solidity is a timescale dependent notion. Pragmatically, T_g is defined as the temperature at which the shear viscosity is equal to 10^{13} Poise (also 10^{12} Pas).

The increase of the relaxation timescale of supercooled liquids is remarkable not only because of the large number of decades involved but also because of its temperature dependence. This is vividly demonstrated by plotting the logarithm of the viscosity (or the relaxation time) as a function of T_g/T, as in Fig. 1. This is called the 'Angell' plot [6] and is very helpful in classifying supercooled liquids. A liquid is called strong or fragile depending on its position in the Angell plot. Straight lines correspond to 'strong' glass-formers and to an Arrhenius behaviour. In this case, one can extract from the plot an effective activation energy, suggesting quite a simple mechanism for relaxation by 'breaking' locally a chemical bond. The typical relaxation time is then dominated by the energy barrier to activate this process and, hence, has an Arrhenius behaviour. Window glasses fall in this category[1]. If one tries to define an effective activation energy for fragile glass-formers using the slope of the curve in Fig. 1, then one finds that this energy scale increases when the temperature decreases, a 'super-Arrhenius' behaviour. This increase of energy barriers immediately suggests that the glass formation is a collective phenomenon for fragile supercooled liquids. Support for this interpretation is provided by the fact that a good fit of the relaxation time or the viscosity is given by the Vogel–Fulcher–Tamman law (VFT):

$$\tau_\alpha = \tau_0 \exp\left[\frac{DT_0}{(T - T_0)}\right] , \qquad (1)$$

which suggests a divergence of the relaxation time, and therefore a phase transition of some kind, at a finite temperature T_0. A smaller D in the VFT law corresponds to a more fragile glass. Note that there are other comparably good fits of these curves, such as the Bässler law [10],

$$\tau_\alpha = \tau_0 \exp\left(K\left(\frac{T_*}{T}\right)^2\right) ,$$

that only lead to a divergence at zero temperature. Actually, although the relaxation time increases by 14 orders of magnitude, the increase of its logarithm, and therefore of the effective activation energy is very modest, and experimental data do not allow one to unambiguously determine the true underlying functional law without any reasonable doubt. For this and other reasons, physical interpretations in terms of a finite temperature phase transition must always be taken with a grain of salt.

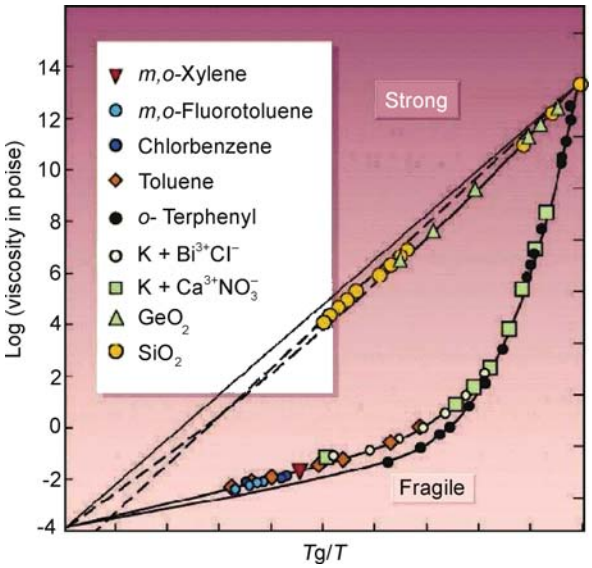

Glasses and Aging, A Statistical Mechanics Perspective on, Figure 1

Arrhenius plot of the viscosity of several glass-forming liquids approaching the glass temperature T_g [67]. For 'strong' glasses, the viscosity increases in an Arrhenius manner as temperature is decreased, $\log \eta \sim E/(K_B T)$, where E is an activation energy and the plot is a straight line, as for silica. For 'fragile' liquids, the plot is bent and the effective activation energy increases when T is decreased towards T_g, as for ortho-terphenyl

However, there are other experimental facts that shed some light and reinforce this interpretation. Among them, is an empirical connection found between kinetic and thermodynamic behaviours. Consider the part of the entropy of the liquids, S_{exc}, which is in excess compared to the entropy of the corresponding crystal. Once this quantity, normalized by its value at the melting temperature, is plotted as a function of T, a remarkable connection with the dynamics emerges. As for the relaxation time one cannot follow this curve below T_g in thermal equilibrium. However, extrapolating the curve below T_g apparently indicates that the excess entropy vanishes at some finite temperature, called T_K, which is very close to zero for strong glasses and, generically, very close to T_0, the temperature at which a VFT fit diverges. This coincidence is quite remarkable: for materials with glass transition temperatures that vary from 50 K to 1000 K the ratio T_K/T_0 remains close to 1, up to a few percents. Examples reported in [145] are provided in Table 1. The chosen subscript for T_K stands for Kauzmann [108] who recognized T_K as a very important temperature in the glass phase di-

[1]The terminology 'strong' and 'fragile' is not related to the mechanical properties of the glass but to the evolution of the short-range order close to T_g. Strong liquids, such as SiO_2, have a locally tetrahedric structure which persists both below and above the glass transition contrary to fragile liquids whose short-range amorphous structure disappears rapidly upon heating above T_g.

Glasses and Aging, A Statistical Mechanics Perspective on, Table 1
Values of glass transition temperature, VFT singularity and Kauzmann temperatures for four supercooled liquids [145]

Substance	o-terphenyl	2-methyltetrahydrofuran	n-propanol	3-bromopentane
T_g	246	91	97	108
T_0	202.4	69.6	70.2	82.9
T_K	204.2	69.3	72.2	82.5
T_K/T_0	1.009	0.996	1.028	0.995

agram. Kauzmann further claimed that some change of behaviour (phase transition, crystal nucleation, etc.) must take place above T_K, because below T_K the entropy of the liquid, a disordered state of matter, becomes less than the entropy of the crystal, an ordered state of matter. This situation that seemed paradoxical at that time is not a serious problem. There is no general principle that would constraint the entropy of the liquid to be larger than that of the crystal. As a matter of fact, the crystallisation transition for hard spheres takes place precisely because the crystal becomes the state with the largest entropy at sufficiently high density [97].

On the other hand, the importance of T_K stands, partially because it is experimentally very close to T_0. Additionally, the quantity S_{exc} which vanishes at T_K, is thought to be a proxy for the so-called configurational entropy, S_c, which quantifies the number of metastable states. A popular physical picture due to Goldstein [91] is that close to T_g the system explores a part of the energy landscape (or configuration space) which is full of minima separated by barriers that increase when temperature decreases. The dynamic evolution in the energy landscape would then consist in a rather short equilibration inside the minima followed by 'jumps' between different minima. At T_g the barriers have become so large that the system remains trapped in one minimum, identified as one of the possible microscopic amorphous configurations of a glass. Following this interpretation, one can split the entropy into two parts. A first contribution is due to the fast relaxation inside one minimum, a second counts the number of metastable states, $S_c = \log N_{\text{metastable}}$, which is called the 'configurational' entropy. Assuming that the contribution to the entropy due to the 'vibrations' around an amorphous glass configuration is not very different from the entropy of the crystal, one finds that $S_{exc} \approx S_c$. In that case, T_K would correspond to a temperature at which the configurational entropy vanishes. This in turn would lead to a discontinuity (a downward jump) of the specific heat and would truly correspond to a thermodynamic phase transition.

Static and Dynamic Correlation Functions

At this point the reader might have reached the conclusion that the glass transition may not be such a difficult problem: there are experimental indications of a diverging timescale and a concomitantly singularity in the thermodynamics. It simply remains to find static correlation functions displaying a diverging correlation length related to the emergence of 'amorphous order', which would indeed classify the glass transition as a standard second order phase transition. Remarkably, this remains an open and debated question despite several decades of research. Simple static correlation function are quite featureless in the supercooled regime, notwithstanding the dramatic changes in the dynamics. A simple static quantity is the structure factor defined by

$$S(q) = \left\langle \frac{1}{N} \delta\rho_{\mathbf{q}} \delta\rho_{-\mathbf{q}} \right\rangle ,$$

where the Fourier component of the density reads

$$\delta\rho_{\mathbf{q}} = \sum_{i=1}^{N} e^{i\mathbf{q}\cdot\mathbf{r}_i} - \frac{N}{V}\delta_{\mathbf{q},0} ,$$

with N is the number of particles, V the volume, and $\mathbf{r}_i$ is the position of particle i. The structure factor measures the spatial correlations of particle positions, but it does not show any diverging peak in contrast to what happens, for example, at the liquid-gas tri-critical point where there is a divergence at small $\mathbf{q}$. More complicated static correlation functions have been studied [66], especially in numerical work, but until now there are no strong indications of a diverging, or at least substantially growing, static lengthscale [133]. A snapshot of a supercooled liquid configuration in fact just looks like a glass configuration, despite their widely different dynamic properties. What happens then at the glass transition? Is it a transition or simply a dynamic crossover? A more refined understanding can be gained studying dynamic correlations or response functions.

A dynamic observable studied in light and neutron scattering experiments is the intermediate scattering function,

$$F(\mathbf{q}, t) = \left\langle \frac{1}{N} \delta\rho_{\mathbf{q}}(t) \delta\rho_{-\mathbf{q}}(0) \right\rangle . \quad (2)$$

Different $F(\mathbf{q}, t)$ measured by neutron scattering in supercooled glycerol [170] are shown for different temperatures in Fig. 2. These curves show a first, rather fast, relaxation to a plateau followed by a second, much slower, relaxation. The plateau is due to the fraction of density

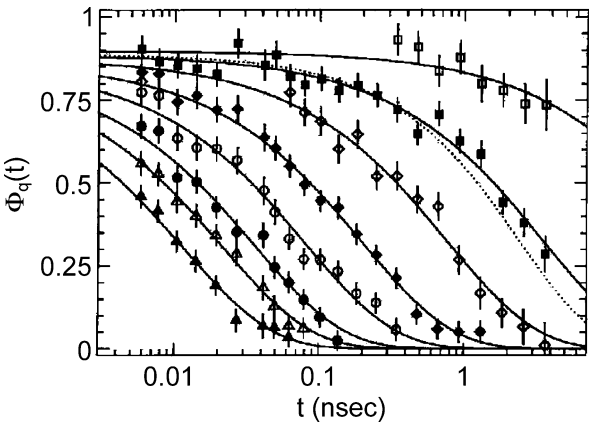

Glasses and Aging, A Statistical Mechanics Perspective on, Figure 2

Temperature evolution of the intermediate scattering function normalized by its value at time equal to zero for supercooled glycerol [170]. Temperatures decrease from 413 K to 270 K from left to right. The solid lines are fit with a stretched exponential with exponent $\beta = 0.7$. The dotted line represents another fit with $\beta = 0.82$

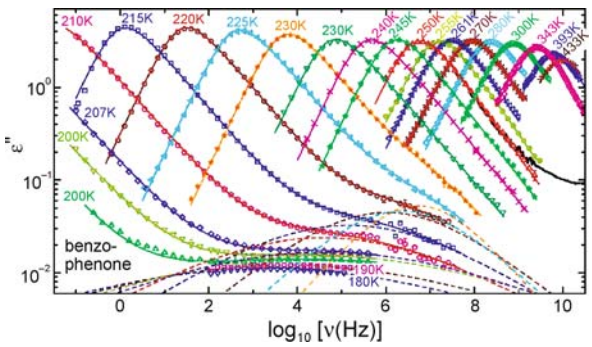

Glasses and Aging, A Statistical Mechanics Perspective on, Figure 3

Temperature evolution of the dielectric susceptibility of the glass-former benzophenone measured over more than 10 decades of relaxation times [142]. Dynamics slows down dramatically as temperature is decreased and relaxation spectra become very broad at low temperature

fluctuations that are frozen on intermediate timescales, but eventually relax during the second relaxation. The latter is called 'alpha-relaxation', and corresponds to the structural relaxation of the liquid. This plateau is akin to the Edwards–Anderson order parameter, q_{EA}, defined for spin glasses which measures the fraction of frozen spin fluctuations [33]. Note that q_{EA} continuously increases from zero below the spin glass transition. Instead, for structural glasses, a finite plateau appears above any transition.

The intermediate scattering function can be probed only on a relatively small regime of temperatures. In order to track the dynamic slowing down from microscopic to macroscopic timescales, other correlators have been studied. A popular one is obtained from the dielectric susceptibility, which is related by the fluctuation-dissipation theorem to the time correlation of polarization fluctuations. It is generally admitted that different dynamic probes reveal similar temperature dependences for the relaxation time. The temperature evolution of the imaginary part of the dielectric susceptibility, $\epsilon''(\omega)$, is shown in Fig. 3 which covers a very wide temperature window [142]. At high temperature, a good representation of the data is given by a Debye law, $\epsilon(\omega) = \epsilon(\infty) + \Delta\epsilon/(1 + i\omega\tau_\alpha)$, which corresponds to an exponential relaxation in the time domain. When temperature is decreased, however, the relaxation spectra become very broad and strongly non-Debye. One particularly well-known feature of the spectra is that they are well fitted, in the time domain, for times corresponding to the alpha-relaxation with a stretched exponential, $\exp(-(t/\tau_\alpha)^\beta)$. In the Fourier domain, forms such as the Havriliak–Negami law are used, $\epsilon(\omega) = \epsilon(\infty) + \Delta\epsilon/(1 + (i\omega\tau_\alpha)^\alpha)^\gamma$, which generalizes the Debye law. The exponents β, α and γ depend in general on temperature and on the particular dynamic probe chosen, but they capture the fact that relaxation is increasingly non-exponential when T decreases towards T_g. A connection was empirically established between fragility and degree of non-exponentiality, more fragile liquids being characterized by broader relaxation spectra [67].

To sum up, there are many remarkable phenomena that take place when a supercooled liquid approaches the glass transition. Striking ones have been presented, but many others have been left out for lack of space [6,33,66,67]. We have discussed physical behaviours, relationships or empirical correlations observed in a broad class of materials. This is quite remarkable and suggests that there is some physics (and not only chemistry) to the problem of the glass transition, which we see as a collective (critical?) phenomenon which is relatively independent of microscopic details. This justifies our statistical mechanics perspective on this problem.

Taxonomy of 'Glasses' in Science

We now introduce some other systems whose phenomenological behaviour is close or, at least, related, to the one of glass-forming liquids, showing that glassiness is truly ubiquitous. It does not only appear in many different physical situations but also in more abstract contexts, such as computer science.

The Jamming Transition of Colloids and Grains

Colloidal suspensions consist of big particles suspended in a solvent [121]. The typical radii of the particles are in the range $R = 1-500$ nm. The solvent, which is at equilibrium at temperature T, renders the short-time dynamics of the particles Brownian. The microscopic timescale for this diffusion is given by $\tau = R^2/D$ where D is the short-time self-diffusion coefficient. Typical values are of the order of $\tau \sim 1$ ms, and thus are much larger than the ones for molecular liquids (in the picosecond regime). The interaction potential between particles depends on the systems, and this large tunability makes colloids very attractive objects for technical applications. A particularly relevant case, on which we will focus in the following, is a purely hard sphere potential, which is zero when particles do not overlap and infinite otherwise. In this case the temperature becomes irrelevant, apart from a trivial rescaling of the microscopic timescale. Colloidal hard spheres systems have been intensively studied [121] in experiments, simulations and theory varying their density ρ, or their volume fraction $\phi = 4/3\pi R^3\rho$. Hard spheres display a fluid phase from 0 to intermediate volume fractions, a freezing-crystallisation transition at $\phi \simeq 0.494$, and a melting transition at $\phi \simeq 0.545$. Above this latter value the system can be compressed until the close packing point $\phi \simeq 0.74$, which corresponds to the FCC crystal. Interestingly for our purposes, a small amount of polydispersity (particles with slightly different sizes) suppresses crystallization. In this case, the system can be more easily 'supercompressed' above the freezing transition without nucleating the crystal, at least on experimental timescales. In this regime the relaxation timescale increases very fast [144]. At a packing fraction $\phi_g \simeq 0.58$ it becomes so large compared to typical experimental timescales that the system does not relax anymore: it is jammed. This 'jamming transition' is obviously reminiscent of the glass transition of molecular systems. In particular, the location ϕ_g of the colloidal glass transition is as ill-defined as the glass temperature T_g.

Actually, the phenomenona that take place increasing the volume fraction are analogous to the ones seen in molecular supercooled liquid: the relaxation timescales increases very fast and can be fitted [52] by a VFT law in density as in Eq. (1), dynamical correlation functions display a broad spectrum of timescales and develop a plateau, no static growing correlation length has been found, etc. Also the phenomenon of dynamic heterogeneity that will be addressed in Sect. "Dynamic Heterogeneity" is present in both cases [109,165]. However, it is important to underline a major difference: because the microscopic timescale for colloids is so large, experiments can only track the first 5 decades of slowing down. A major consequence is that the comparison between the glass and colloidal transitions must be performed by focusing in both cases on the first 5 decades of the slowing down, which corresponds to relatively high temperatures in molecular liquids. Understanding how much and to what extent the glassiness of colloidal suspensions is related to the one of molecular liquids remains an active domain of research.

Another class of systems that have recently been studied from the point of view of their glassiness is driven granular media. Grains are macroscopic objects and, as a consequence, do not have any thermal motion. A granular material is therefore frozen in a given configuration if no energy is injected into the system [104]. However, it can be forced in a steady state by an external drive, such as shearing or tapping. The dynamics in this steady state shows remarkable similarities (and differences) with simple fluids. The physics of granular materials is a very wide subject [104]. In the following we only address briefly what happens to a polydisperse granular fluid at very high packing fractions, close to its random close packed state. As for colloids, the timescales for relaxation or diffusion increase very fast when density is increased, without any noticeable change in structural properties. Again, it is now established [62,110,127] that many phenomonelogical properties of the glass and jamming transitions also occur in granular assemblies. As for colloids, going beyond the mere analogy and understanding how much these different physical systems are related is a very active domain of research.

This very question has been asked in a visual manner by Liu and Nagel [124] who rephrased it in a single picture, reproduced in Fig. 4. By building a common phase diagram for glasses, colloids and grains, they ask whether the glass and jamming transitions of molecular liquids, colloids and granular media are different facets of the same phase. In this unifying 'phase diagram', the 'phase' close to the origin is glassy and can be reached either by lowering the temperature as in molecular liquids, or increasing the packing fraction or decreasing the external drive in colloids and granular media. It remains to provide precise answers to this elegantly formulated, but rather broad, set of questions.

Other 'Glasses' in Physics and Beyond

There are many other physical contexts in which glassiness plays an important role [173]. One of the most famous examples is the field of spin glasses. Real spin glasses are magnetic impurities interacting by quenched random couplings. At low temperatures, their dynamics be-

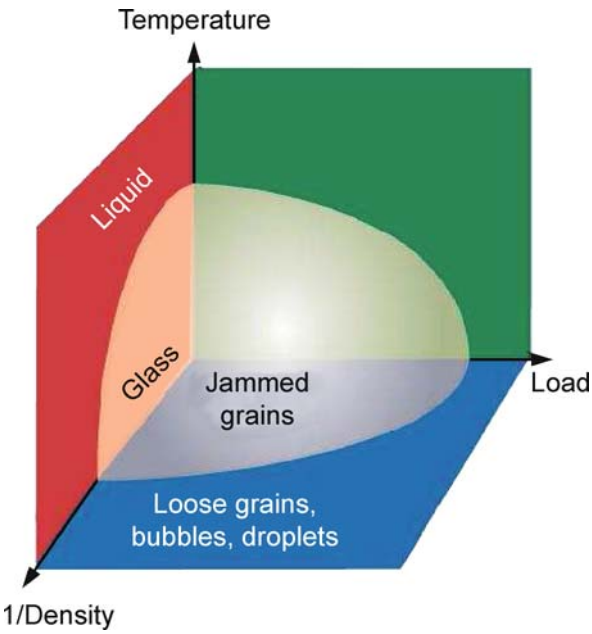

Glasses and Aging, A Statistical Mechanics Perspective on, Figure 4
The 'great unification' phase diagram of jamming and glass transitions [124]. Glassy phases occur at low temperature, low external drive, or high density in different systems

come extremely slow and they freeze in amorphous spin configuration dubbed a 'spin glass' by P. W. Anderson. There are many other physical systems, often characterized by quenched disorder, that show glassy behaviour, like Coulomb glasses, Bose glasses, etc. In many cases, however, one does expect quite a different physics from structural glasses: the similarity between these systems is therefore only qualitative.

Finally, and quite remarkably, glassiness emerges even in other branches of science [55]. In particular, it has been discovered recently that concepts and techniques developed for glassy systems turn out to apply and be very useful tools in the field of computer science. Problems like combinatorial optimization display phenomena completely analogous to phase transitions, actually, to glassy phase transitions. A posteriori, this is quite natural, because a typical optimization problem consists in finding a solution in a presence of a large number of constraints. This can be defined, for instance, as a set of N Boolean variables that satisfies M constraints. For N and M very large at fixed $\alpha = M/N$, this problem very much resembles finding a ground state in a statistical mechanics problem with quenched disorder. Indeed one can define an energy function (a Hamiltonian) as the number of unsatisfied constraints, that has to be minimized, as in a $T = 0$ statmech problem. The connection with glassy systems origins from the fact that in both cases the energy landscape is extremely complicated, full of minima and saddles. The fraction of constraints per degree of freedom, α, plays a role similar to the density in a hard sphere system. A detailed presentation of the relationship between optimization problems and glassy systems is clearly out of the scope of the present review. We simply illustrate it pointing out that a central problem in optimization, random k-satisfiability, has been shown to undergo a glass transition when α increases that is analogous to the one of structural glasses [117].

Numerical Simulations

Studying the glass transition of molecular liquids at a microscopic level is in principle straightforward since one must answer a very simple question: how do particles move in a liquid close to T_g? It is of course a daunting task to attempt answering this question experimentally because one should then resolve the dynamics of single molecules to be able to follow the trajectories of objects that are a few Angstroms large on timescales of tens or hundreds of seconds, which sounds like eternity when compared to typical molecular dynamics usually lying in the picosecond regime. In recent years, such direct experimental investigations have been started using time and space resolved techniques such as atomic force microscopy [161] or single molecule spectroscopy [3], but this remains a very difficult task.

In numerical simulations, by contrast, the trajectory of each particle in the system can, by construction, be followed at all times. This allows one to quantify easily single particle dynamics, as proved in Fig. 5 where the averaged mean-squared displacement $\Delta(t)$ measured in a simple Lennard-Jones glass-former is shown. It is defined by

$$\Delta(t) = \left\langle \frac{1}{N} \sum_{i=1}^{N} |\mathbf{r}_i(t) - \mathbf{r}_i(0)|^2 \right\rangle ,$$

where $\mathbf{r}_i(t)$ represents the position of particle i at time t in a system composed of N particles; the brackets indicate an ensemble average. The particle displacements considerably slow down when T is decreased and the self-diffusion constant decreases by orders of magnitude, mirroring the behaviour of the viscosity shown in Fig. 1 for real systems. Moreover, a rich dynamics is observed, with a plateau regime at intermediate timescales, corresponding to an extended time window during which particles vibrate around their initial positions, exactly as in a crystalline solid. The difference with a crystal is of course that this localization is only transient, and all particles eventu-

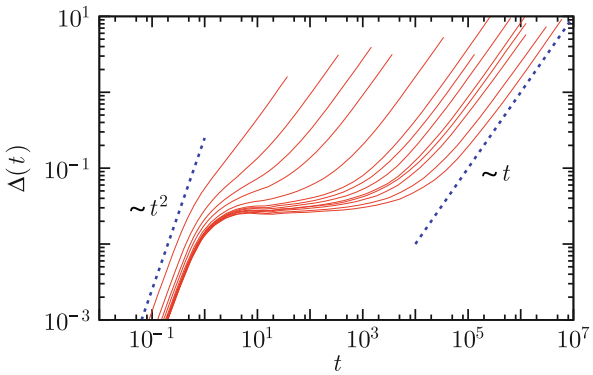

Glasses and Aging, A Statistical Mechanics Perspective on, Figure 5

Mean-squared displacements of individual particles in a simple model of a glass-forming liquid composed of Lennard-Jones particles observed on a wide time window. When temperature decreases (from *left* to *right*), the particle displacements become increasingly slow with several distinct time regimes corresponding, in this order, to ballistic, localized, and diffusive regimes

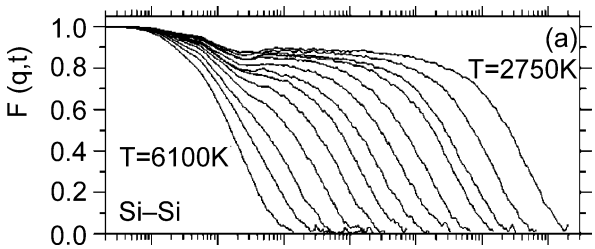

Glasses and Aging, A Statistical Mechanics Perspective on, Figure 6

Intermediate scattering function at wavevector 1.7 Å^{-1} for the Si particles at $T = 2750$ K obtained from molecular dynamics simulations of a model for silica [98]

ally escape and diffuse at long times with a diffusion constant D_s, so that $\Delta(t) \sim 6D_s t$ when $t \to \infty$.

In recent years, computer experiments have played an increasingly important role in glass transition studies. It could almost be said that particle trajectories in numerical work have been studied under so many different angles that probably very little remains to be learnt from such studies in the regime that is presently accessible using present day computers. Unfortunately, this does not imply complete knowledge of the physics of supercooled liquids. As shown in Fig. 5, it is presently possible to follow the dynamics of a simple glass-forming liquid over more than eight decades of time, and over a temperature window in which average relaxation timescales increase by more than five decades. This might sound impressive, but a quick look at Fig. 1 shows, however, that at the lowest temperatures studied in the computer, the relaxation timescales are still orders of magnitude faster than in experiments performed close to the glass transition temperature. They can be directly compared to experiments performed in this high temperature regime, but this also implies that simulations focus on a relaxation regime that is about eight to ten decades of times faster than in experiments performed close to T_g. Whether numerical works are useful to understand the glass transition itself at all is therefore an open, widely debated, question. We believe that it is now possible to numerically access temperatures which are low enough that many features associated to the glass transition physics can be observed: strong decoupling phenomena, clear deviations from fits to the mode-coupling theory (which are experimentally known to hold only at high temperatures), and crossovers towards truly activated dynamics.

Classical computer simulations of supercooled liquids usually proceed by solving a cleverly discretized version of Newton's equations for a given potential interaction between particles [4]. If quantitative agreement with experimental data on an existing specific material is sought, the interaction must be carefully chosen in order to reproduce reality, for instance by combining classical to ab initio simulations. From a more fundamental perspective one rather seeks the simplest model that is still able to reproduce qualitatively the phenomenology of real glass-formers, while being considerably simpler to study. The implicit, but quite strong, hypothesis is that molecular details are not needed to explain the behaviour of supercooled liquids, so that the glass transition is indeed a topic for statistical mechanics, not for chemistry. A considerable amount of work has therefore been dedicated to studying models such as hard spheres, soft spheres, or Lennard-Jones particles. More realistic materials are also studied focusing for instance on the physics of network forming materials, multi-component ones, anisotropic particles, or molecules with internal degrees of freedom. Connections to experimental work can be made by computing quantities that are experimentally accessible such as the intermediate scattering function, static structure factors, $S(\mathbf{q})$, or thermodynamic quantities such specific heat or configurational entropy, which are directly obtained from particle trajectories and can be measured in experiments as well. As an example we show in Fig. 6 the intermediate scattering function $F(\mathbf{q}, t)$ obtained from a molecular dynamics simulation of a classical model for SiO_2 as a function of time for different temperatures [98].

An important role is played by simulations also because a large variety of dynamic and static quantities can

be simultanously measured in a single model system. As we shall discuss below, there exist scores of different theoretical approaches to describe the physics of glass-formers, and they sometimes have their own set of predictions that can be readily tested by numerical work. Indeed, quite a large amount of numerical papers have been dedicated to testing in detail the predictions formulated by the mode-coupling theory of the glass transition, as reviewed recently in [94]. Here, computer simulations are particularly well-suited as the theory specifically addresses the relatively high temperature window that is studied in computer simulations.

While Newtonian dynamics is mainly used in numerical work on supercooled liquids, a most appropriate choice for these materials, it can be interesting to consider alternative dynamics that are not deterministic, or which do not conserve the energy. In colloidal glasses and physical gels, for instance, particles undergo Brownian motion arising from collisions with molecules in the solvent, and a stochastic dynamics is more appropriate. Theoretical considerations might also suggest the study of different sorts of dynamics for a given interaction between particles, for instance, to assess the role of conservation laws and structural information. Of course, if a given dynamics satisfies detailed balance with respect to the Boltzmann distribution, all structural quantities remain unchanged, but the resulting dynamical behaviour might be very different. Several papers [27,88,153] have studied in detail the influence of the chosen microscopic dynamics on the dynamical behaviour in glass-formers using either stochastic dynamics (where a friction term and a random noise are added to Newton's equations, the amplitude of both terms being related by a fluctuation-dissipation theorem), Brownian dynamics (in which there are no momenta, and positions evolve with Langevin dynamics), or Monte-Carlo dynamics (where the potential energy between two configurations is used to accept or reject a trial move). Quite surprisingly, the equivalence between these three types of stochastic dynamics and the originally studied Newtonian dynamics was established at the level of the averaged dynamical behaviour [27,88,153], except at very short times where obvious differences are indeed expected. This strongly suggests that an explanation for the appearance of slow dynamics in these materials originates from their amorphous structure. However, important differences were found when dynamic fluctuations were considered [21,22,27], even in the long-time regime comprising the structural relaxation.

Another crucial advantage of molecular simulations is illustrated in Fig. 7. This figure shows a spatial map of single particle displacements recorded during the simulation

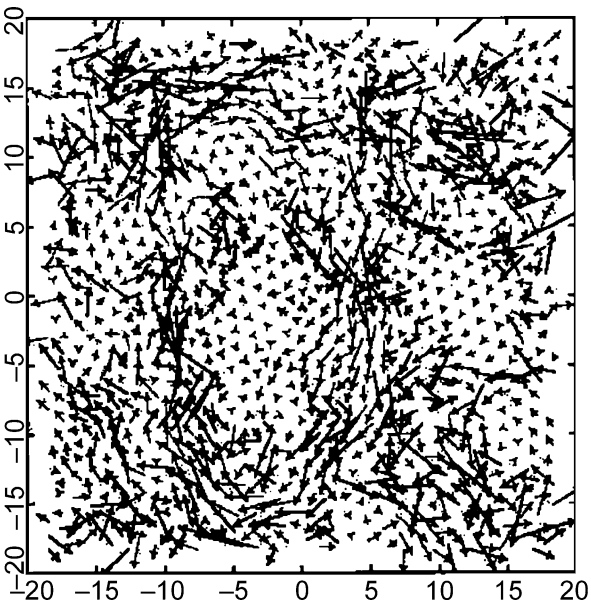

Glasses and Aging, A Statistical Mechanics Perspective on, Figure 7

Spatial map of single particle displacements in the simulation of a binary mixture of soft spheres in two dimensions [99]. Arrows show the displacement of each particle in a trajectory of length about 10 times the structural relaxation time. The map reveals the existence of particles with different mobilities during relaxation, but also the existence of spatial correlations between these dynamic fluctuations

of a binary soft sphere system in two dimensions [99]. This type of measurement, out of reach of most experimental techniques that study the liquid state, reveals that dynamics might be very different from one particle to another. More importantly, Fig. 7 also unambiguously reveals the existence of spatial correlations between these dynamic fluctuations. The presence of non-trivial spatio-temporal fluctuations in supercooled liquids is now called 'dynamic heterogeneity' [72]. This is the phenomenon we discuss in more detail in the next section.

Dynamic Heterogeneity

Existence of Spatio-temporal Dynamic Fluctuations

A new facet of the relaxational behaviour of supercooled liquids has emerged in the last decade thanks to a considerable experimental and theoretical effort. It is called 'dynamic heterogeneity' (DH), and plays now a central role in modern descriptions of glassy liquids [72]. As anticipated in the previous section, the phenomenon of dynamic heterogeneity is related to the spatio-temporal fluctuations of the dynamics. Initial motivations stemmed from the

search for an explanation of the non-exponentiality of relaxation processes in supercooled liquids, related to the existence of a broad relaxation spectrum. Two natural, but fundamentally different, explanations can be put forward. (1) The relaxation is locally exponential, but the typical relaxation timescale varies spatially. Hence, global correlation or response functions become non-exponential upon spatial averaging over this spatial distribution of relaxation times. (2) The relaxation is complicated and inherently non-exponential, even locally. Experimental and theoretical works [72] suggest that both mechanisms are likely at play, but definitely conclude that relaxation is spatially heterogeneous, with regions that are faster and slower than the average. Since supercooled liquids are ergodic materials, a slow region will eventually become fast, and vice versa. A physical characterization of DH entails the determination of the typical lifetime of the heterogeneities, as well as their typical lengthscale.

A clear and more direct confirmation of the heterogenous character of the dynamics also stems from simulation studies. For example, whereas the simulated average mean-squared displacements are smooth functions of time, time signals for individual particles clearly exhibit specific features that are not observed unless dynamics is resolved both in space and time. These features are displayed in Fig. 8. What do we see? We mainly observe that particle trajectories are not smooth but rather composed of a succession of long periods of time where particles simply vibrate around well-defined locations, separated by rapid 'jumps'. Vibrations were previously inferred from the plateau observed at intermediate times in the mean-squared displacements of Fig. 5, but the existence of jumps that are clearly statistically widely distributed in time cannot be guessed from averaged quantities only. The fluctuations in Fig. 8 suggest, and direct measurements confirm, the importance played by fluctuations around the averaged dynamical behaviour.

A simple type of such fluctuations has been studied in much detail. When looking at Fig. 8, it is indeed natural to ask, for any given time, what is the distribution of particle displacements. This is quantified by the self-part of the van-Hove function defined as

$$G_s(\mathbf{r}, t) = \left\langle \frac{1}{N} \sum_{i=1}^{N} \delta(\mathbf{r} - [\mathbf{r}_i(t) - \mathbf{r}_i(0)]) \right\rangle .$$

For an isotropic Gaussian diffusive process, one gets $G_s(\mathbf{r}, t) = \exp(-|\mathbf{r}|^2/(4D_s t))/(4\pi D_s t)^{3/2}$. Simulations reveal instead strong deviations from Gaussian behaviour on the timescales relevant for structural relaxation [116]. In particular they reveal 'fat' tails in the distributions that

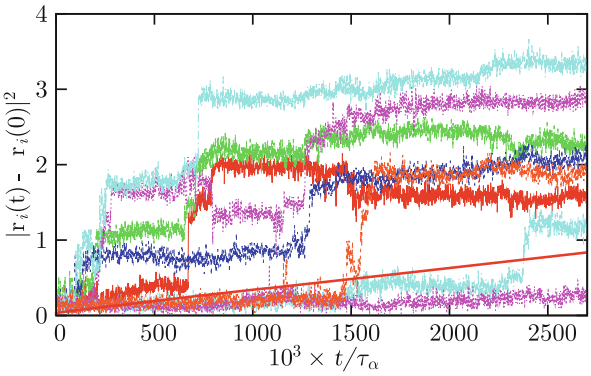

Glasses and Aging, A Statistical Mechanics Perspective on, Figure 8

Time resolved squared displacements of individual particles in a simple model of a glass-forming liquid composed of Lennard-Jones particles. The average is shown as a smooth full line. Trajectories are composed of long periods of time during which particles vibrate around well-defined positions, separated by rapid jumps that are widely distributed in time underlying the importance of dynamic fluctuations

are much wider than expected from the Gaussian approximation. These tails are in fact well described by an exponential, rather than Gaussian, decay in a wide time window comprising the structural relaxation, such that $G_s(\mathbf{r}, t) \sim \exp(-|\mathbf{r}|/\lambda(t))$ [51]. Thus, they reflect the existence of a population of particles that moves distinctively further than the rest and appears therefore to be much more mobile. This observation implies that relaxation in a viscous liquid differs qualitatively from that of a normal liquid where diffusion is close to Gaussian, and that a non-trivial statistics of single particle displacements exists.

A long series of questions immediately follows this seemingly simple observation. Answering them has been the main occupation of many workers in this field over the last decade. What are the particles in the tails effectively doing? Why are they faster than the rest? Are they located randomly in space or do they cluster? What is the geometry, time and temperature evolution of the clusters? Are these spatial fluctuations correlated to geometric or thermodynamic properties of the liquids? Do similar correlations occur in all glassy materials? Can one predict these fluctuations theoretically? Can one understand glassy phenomenology using fluctuation-based arguments? Can these fluctuations be detected experimentally?

Another influential phenomenon that was related early on to the existence of DH is the decoupling of self-diffusion (D_s) and viscosity (η). In the high temperature liquid self-diffusion and viscosity are related by the Stokes–Einstein relation [95], $D_s\eta/T = \text{const}$. For a large

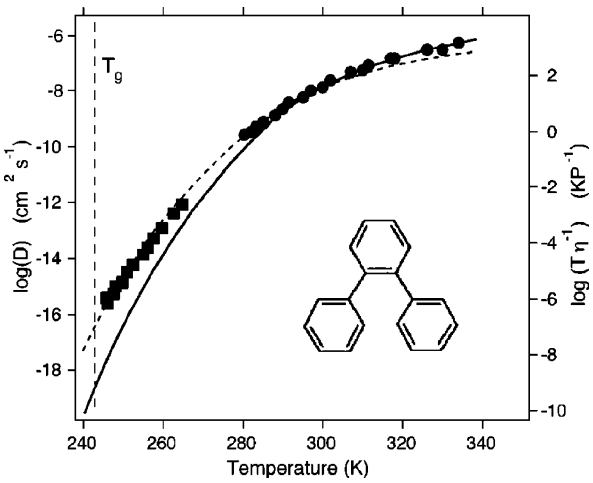

Glasses and Aging, A Statistical Mechanics Perspective on, Figure 9

Decoupling between viscosity (full line) and self-diffusion coefficient (symbols) in supercooled ortho-terphenyl [126] The dashed line shows a fit with a 'fractional' Stokes–Einstein relation, $D_s \sim (T/\eta)^{\zeta}$ with $\zeta \sim 0.82$

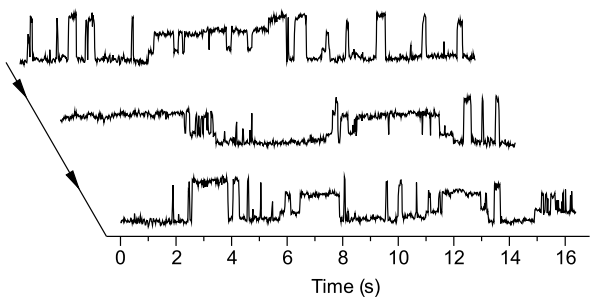

Glasses and Aging, A Statistical Mechanics Perspective on, Figure 10

Time series of polarization in the AFM experiment performed by Vidal Russell and Israeloff [161] on PVAc at $T = 300$ K. The signal intermittently switches between periods with fast or slow dynamics, suggesting that extended regions of space indeed transiently behave as fast and slow regions

particle moving in a fluid the constant is equal to $1/(6\pi R)$ where R is the particle radius. Physically, the Stokes–Einstein relation means that two different measures of the relaxation time R^2/D_s and $\eta R^3/T$ lead to the same timescale up to a constant factor. In supercooled liquids this phenomenological law breaks down, as shown in Fig. 9 for ortho-terphenyl [126]. It is commonly found that D_s^{-1} does not increase as fast as η so that, at T_g, the product $D_s\eta$ has increased by 2–3 orders of magnitude as compared to its Stokes–Einstein value. This phenomenon, although less spectacular than the overall change of viscosity, is a significative indication that different ways to measure relaxation times lead to different answers and, thus, is a strong hint of the existence of a distribution of relaxation timescales.

Indeed, a natural explanation of this effect is that different observables probe differently way the underlying distribution of relaxation times [72]. For example, the self-diffusion coefficient of tracer particles is dominated by the more mobile particles whereas the viscosity or other measures of structural relaxation probe the timescale needed for every particle to move. An unrealistic but instructive example is a model where there is a small, non-percolative subset of particles that are blocked forever, coexisting with a majority of mobile particles. In this case, the structure never relaxes but the self-diffusion coefficient is non-zero because of the mobile particles. Of course, in reality all particles move, eventually, but this shows how different observables are likely to probe different moments of the distribution of timescales, as explicitely shown within several theoretical frameworks [106,154].

The phenomena described above, although certainly an indication of spatio-temporal fluctuations, do not allow one to study how these fluctuations are correlated in space. This is, however, a fundamental issue both from the experimental and theoretical points of view. How large are the regions that are faster or slower than the average? How does their size depend on temperature? Are these regions compact or fractal? These important questions were first addressed in pioneering works using four-dimensional NMR [160], or by directly probing fluctuations at the nanoscopic scale using microscopy techniques. In particular, Vidal Russel and Israeloff using Atomic Force Microscopy techniques [161] measured the polarization fluctuations in a volume of size of few tens of nanometers in a supercooled polymeric liquid (PVAc) close to T_g. In this spatially resolved measurement, the hope is to probe a small enough number of dynamically correlated regions, and detect their dynamics. Indeed, the signal shown in Fig. 10 shows a dynamics which is very intermittent in time, the dynamics switching between moments with intense activity, and moments with no dynamics at all, suggesting that extended regions of space indeed transiently behave as fast and slow regions. A much smoother signal would have been measured if these such dynamically correlated 'domains' were not present. Spatially resolved and NMR experiments are quite difficult. They give undisputed information about the typical lifetime of the DH, but their determination of a dynamic correlation lengthscale is rather indirect and/or performed on a small number of liquids in a small temperature window. Nevertheless, the outcome is that

a non-trivial dynamic correlation length emerges at the glass transition, where it reaches a value of the order of 5–10 molecule diameters [72].

Multi-point Correlation Functions

More recently, substantial progress in characterizing spatio-temporal dynamical fluctuations was obtained from theoretical [21,22,79,159] and numerical results [14,28,76,99,172]. In particular, it is now understood that dynamical fluctuations can be measured and characterized through the use of four-point correlation functions. These multi-point functions can be seen as a generalization of the spin glass susceptibility measuring the extent of amorphous long-range order in spin glasses. In this subsection, we introduce these correlation functions and summarize the main results obtained using them.

Standard experimental probes of the averaged dynamics of liquids give access to the time-dependent auto-correlation function of the spontaneous fluctuations of some observable $O(t)$, $F(t) = \langle \delta O(0) \delta O(t) \rangle$, where $\delta O(t) = O(t) - \langle O \rangle$ represents the instantaneous value of the deviation of $O(t)$ from its ensemble average $\langle O \rangle$ at time t. One can think of $F(t)$ as being the average of a two-point quantity, $C(0,t) = \delta O(0) \delta O(t)$, characterizing the dynamics. A standard example corresponds to O being equal to the Fourier transform of the density field. In this case $F(t)$ is the dynamical structure factor as in Eq. (2). More generally, the correlation functions $F(t)$ measure the global relaxation in the system. Intuitively, in a system with important dynamic correlations, the fluctuations of $C(0,t)$ will be stronger. Quantitative information on the amplitude of those fluctuations is provided by the variance

$$\chi_4(t) = N \left\langle \delta C(0,t)^2 \right\rangle \ , \tag{3}$$

where $\delta C(0,t) = C(0,t) - F(t)$, and N is the total number of particles in the system. The associated spatial correlations show up more clearly when considering a 'local' probe of the dynamics, like for instance an orientational correlation function measured by dielectric or light scattering experiments, which can be expressed as

$$C(0,t) = \frac{1}{V} \int \mathrm{d}^3 r c(\mathbf{r}; 0, t) \ , \tag{4}$$

where V is the volume of the sample and $c(\mathbf{r}; 0, t)$ characterizes the dynamics between times 0 and t around point $\mathbf{r}$. For example, in the above mentioned case of orientational correlations, $c(\mathbf{r}; 0, t) \propto \frac{V}{N} \sum_{i,j=1}^{N} \delta(\mathbf{r} - \mathbf{r}_i) Y(\Omega_i(0)) Y(\Omega_j(t))$, where Ω_i denotes the angles describing the orientation of molecule i, $\mathbf{r}_i(0)$ is the position of that molecule at time 0, and $Y(\Omega)$ is some appropriate rotation matrix element. Here, the 'locality' of the probe comes from the fact that it is dominated by the self-term involving the same molecule at times 0 and t, or by the contribution coming from neighboring molecules. The dynamic susceptibility $\chi_4(t)$ can thus be rewritten as

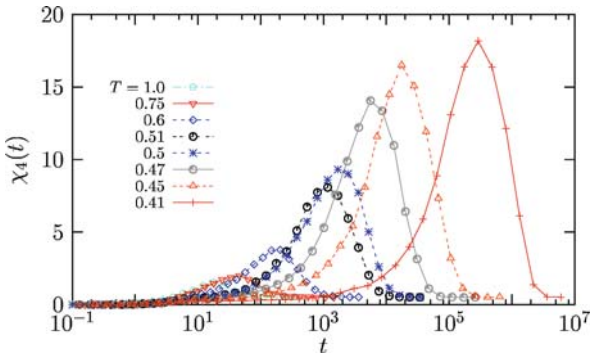

Glasses and Aging, A Statistical Mechanics Perspective on, Figure 11

Time dependence of $\chi_4(t)$ quantifying the spontaneous fluctuations of the intermediate scattering function in a Lennard-Jones supercooled liquid. For each temperature, $\chi_4(t)$ has a maximum, which shifts to larger times and has a larger value when T is decreased, revealing the increasing lengthscale of dynamic heterogeneity in supercooled liquids approaching the glass transition

$$\chi_4(t) = \rho \int \mathrm{d}^3 r G_4(\mathbf{r}; 0, t) \ , \tag{5}$$

where

$$G_4(\mathbf{r}; 0, t) = \langle \delta c(\mathbf{0}; 0, t) \delta c(\mathbf{r}; 0, t) \rangle \ , \tag{6}$$

and translational invariance has been taken into account ($\rho = N/V$ denotes the mean density). The above equations show that $\chi_4(t)$ measures the extent of spatial correlation between dynamical events between times 0 and t at different points of the system, i. e., the spatial extent of dynamically heterogeneous regions over a time span t.

The function $\chi_4(t)$ has been measured by molecular dynamics, Brownian and Monte Carlo simulations in different liquids [14,28,29,76,163]. An example is shown in Fig. 11 for a Lennard-Jones liquid. The qualitative behaviour is similar in all cases [21,79,159]: as a function of time $\chi_4(t)$ first increases, it has a peak on a timescale that tracks the structural relaxation timescale and then it decreases[2]. The peak value measures thus the volume

[2]The decrease at long times constitutes a major difference with spin glasses. In a spin glass, χ_4 would be a monotonically increasing function of time whose long-time limit coincides with the static spin glass susceptibility. Physically, the difference is that spin glasses develop long-range static amorphous order while structual glasses do not or, at least, in a different and more subtle way.

on which the structural relaxation processes are correlated. It is found to increase when the temperature decreases and the dynamics slows down. By measuring directly $G_4(\mathbf{r}; 0, t)$ it has also been checked that the increase of the peak of $\chi_4(t)$ corresponds, as expected, to a growing dynamic lengthscale ξ [14,21,28], although these measurements are much harder in computer simulations, because very large systems need to be simulated to determine ξ unambiguously. Note that if the dynamically correlated regions were compact, the peak of χ_4 would be proportional to ξ^3 in three dimensions, directly relating χ_4 measurements to that of the relevant lengthscale of DH.

These results are also relevant because many theories of the glass transition assume or predict, in a way or another, that the dynamics slows down because there are increasingly large regions on which particles have to relax in a correlated or cooperative way. However, this lengthscale remained elusive for a long time. Measures of the spatial extent of dynamic heterogeneity, in particular $\chi_4(t)$ and $G_4(\mathbf{r}; 0, t)$, seem to provide the long-sought evidence of this phenomenon. This in turn suggests that the glass transition is indeed a critical phenomenon characterized by growing timescales and lengthscales. A clear and conclusive understanding of the relationship between the lengthscale obtained from $G_4(\mathbf{r}; 0, t)$ and the relaxation timescale is still the focus of an intense research activity.

One major issue is that obtaining information on the behaviour of $\chi_4(t)$ and $G_4(\mathbf{r}; 0, t)$ from experiments is difficult. Such measurements are necessary because numerical simulations can only be performed rather far from T_g, see Sect. "Numerical Simulations". Up to now, direct experimental measurements of $\chi_4(t)$ have been restricted to colloidal [166] and granular materials [65,110] close to the jamming transition, because dynamics is more easily spatially resolved in those cases. Unfortunately, similar measurements are currently not available in molecular liquids.

Recently, an approach based on fluctuation-dissipation relations and rigorous inequalities has been developed in order to overcome this difficulty [20,21,22]. The main idea is to obtain a rigorous lower bound on $\chi_4(t)$ using the Cauchy–Schwarz inequality $\langle \delta H(0)\delta C(0,t)\rangle^2 \leq \langle \delta H(0)^2\rangle \langle \delta C(0,t)^2\rangle$, where $H(t)$ denotes the enthalpy at time t. By using fluctuation-dissipation relations the previous inequality can be rewritten as [20]

$$\chi_4(t) \geq \frac{K_B T^2}{c_P} \left[\chi_T(t)\right]^2 , \tag{7}$$

where the multi-point response function $\chi_T(t)$ is defined by

$$\chi_T(t) = \left.\frac{\partial F(t)}{\partial T}\right|_{N,P} = \frac{N}{K_B T^2} \langle \delta H(0)\delta C(0,t)\rangle .$$

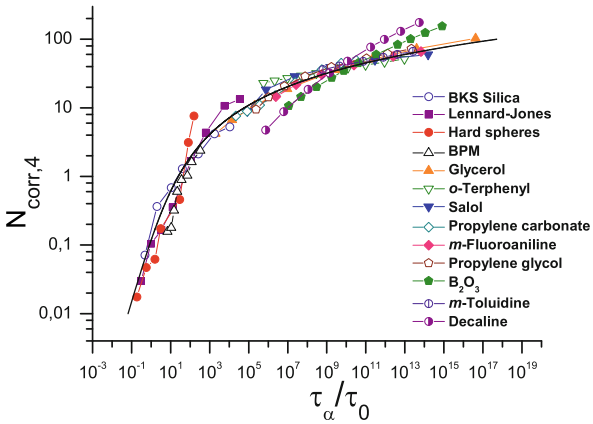

Glasses and Aging, A Statistical Mechanics Perspective on, Figure 12

Universal dynamic scaling relation between number of dynamically correlated particles, $N_{corr,4}$, and relaxation timescale, τ_α, for a number of glass-formers [63], determined using Eq. (7)

In this way, the experimentally accessible response $\chi_T(t)$ which quantifies the sensitivity of average correlation functions $F(t)$ to an infinitesimal temperature change, can be used in Eq. (7) to yield a lower bound on $\chi_4(t)$. Moreover, detailed numerical simulations and theoretical arguments [21,22] strongly suggest that the right hand side of (7) actually provides a good estimation of $\chi_4(t)$, not just a lower bound.

Using this method, Dalle-Ferrier et al. [63] have been able to obtain the evolution of the peak value of χ_4 for many different glass-formers in the entire supercooled regime. In Fig. 12 we show some of these results as a function of the relaxation timescale. The value on the y-axis, the peak of χ_4, is a proxy for the number of molecules, $N_{corr,4}$ that have to evolve in a correlated way in order to relax the structure of the liquid. Note that χ_4 is expected to be equal to $N_{corr,4}$, up to a proportionality constant which is not known from experiments, probably explaining why the high temperature values of $N_{corr,4}$ are smaller than one. Figure 12 also indicates that $N_{corr,4}$ grows faster when τ_α is not very large, close to the onset of slow dynamics, and a power law relationship between $N_{corr,4}$ and τ_α is good in this regime ($\tau_\alpha/\tau_0 < 10^4$). The growth of $N_{corr,4}$ becomes much slower closer to T_g. A change of 6 decades in time corresponds to a mere increase of a factor about 4 of $N_{corr,4}$, suggesting logarithmic rather than power law growth of dynamic correlations. This is in agreement with several theories of the glass transition which are based on activated dynamic scaling [85,155,171].

Understanding quantitatively this relation between timescales and lengthscales is one of the main recent topics addressed in theories of the glass transition, see

Sect. "Some Theory and Models". Furthermore, numerical works are also devoted to characterizing better the geometry of the dynamically heterogeneous regions [7,69].

Some Theory and Models

We now present some theoretical approaches to the glass transition. It is impossible to cover all of them in a brief review, simply because there are way too many of them, perhaps the clearest indication that the glass transition remains an open problem. We choose to present approaches that are keystones and have a solid statistical mechanics basis. Loosely speaking, they have an Hamiltonian, can be simulated numerically, or studied analytically with statistical tools. Of course, the choice of Hamiltonians is crucial and contains very important assumptions about the nature of the glass transition. All these approaches have given rise to unexpected results. One finds more in them than what was supposed at the beginning, which leads to new, testable predictions. Furthermore, with models that are precise enough, one can test (and hopefully falsify!) these approaches by working out all their predictions in great detail, and comparing the outcome to experimental data. This is not possible with 'physical pictures', or simpler approaches of the problem which we have therefore avoided.

Before going into the models, we would like to state the few important questions that face theoreticians.

- Why do the relaxation time and the viscosity increase when T_g is approached? Why is this growth super-Arrhenius?
- Can one understand and describe quantitatively the average dynamical behaviour of supercooled liquids, in particular broad relaxation spectra, non-exponential behaviour, and their evolution with fragility?
- Is there a relation between kinetics and thermodynamics (like $T_0 \simeq T_K$), and why?
- Can one understand and describe quantitatively the spatio-temporal fluctuations of the dynamics? How and why are these fluctuations related to the dynamic slowing down?
- Is the glass transition a collective phenomenon? If yes, of which kind? Is there a finite temperature or zero temperature ideal glass transition? In this case, is the transition of static or purely dynamic origin?
- Is there a geometric, real space explanation for the dynamic slowing down that takes into account molecular degrees of freedom?

The glass transition appears as a kind of 'intermediate coupling' problem, since for instance typical growing length-scales are found to be at most a few tens of particle large close to T_g. It would therefore be difficult to recognize the correct theory even if one bumped into it. To obtain quantitative, testable predictions, one must therefore be able to work out also preasymptotic effects. This is particularly difficult, especially in cases where the asymptotic theory itself has not satisfactorily been worked out. As a consequence, at this time, theories can only be judged by their overall predictive power and their theoretical consistency.

Cooperativity, Chaotic Energy Landscapes and Random First Order Theory

In the last two decades, three *independent* lines of research approaches, Adam–Gibbs theory [2], mode-coupling theory [94] and spin glass theory [137], have merged to produce a theoretical ensemble that now goes under the name of Random First Order Theory (RFOT), a terminology introduced by Kirkpatrick, Thirumalai and Wolynes [111] who also played a major role in its development. Instead of following the rambling development of history, we summarize it in a more modern and unified way.

A key ingredient of RFOT is the existence of a chaotic or complex free energy landscape and its evolution with temperature and/or density. Analysing it in a controlled way for three dimensional interacting particles system is of course an impossible task. This can be achieved, however, in simplified models or using mean-field approximation, that have therefore played a crucial role in the development of RFOT.

A first, concrete example is given by 'lattice glass models' [37]. These are models containing hard particles sitting on the sites of a lattice. The Hamiltonian is infinite if there is more than one particle on a site or if the number of occupied neighbors of an occupied site is larger than a parameter, m, but the Hamiltonian is zero otherwise. Tuning the parameter m, or changing the type of lattice, in particular its connectivity, yields different models. Lattice glasses are constructed as simple statmech models to study the glassiness of hard sphere systems. The constraint on the number of occupied neighbors mimicks the geometric frustration [139] encountered when trying to pack hard spheres in three dimensions. Other models, which have a finite energy and, hence, are closer to molecular glass-formers, can be also constructed [131]. These models can be solved exactly on a Bethe lattice[3], which reveals an astonishing physical behaviour [147]. In particular their free energy landscape can be analyzed in full details and turns

[3] In order to have a well-defined thermodynamics, Bethe lattices are generated as random graphs with fixed connectivity, also called random regular graphs.

out to have the properties that are also found in several 'generalized spin glasses'.

Probably the most studied example of such spin glasses is the p-spin model, defined by the Hamiltonian [93]

$$H = -\sum_{i_1,\ldots,i_p} J_{i_1,\ldots,i_p} S_{i_1} \ldots S_{i_p} \,, \tag{8}$$

where the S_is are Ising or spherical spins, $p > 2$ and $J_{i_1,\ldots,i_p}$ quenched random couplings with zero mean variance $p!/(2N^{p-1})$.

All these models (lattice glasses, their finite energy generalizations and their quenched disorder counterparts) belong the class of one-step replica symmetry breaking systems [137]. This makes reference to the ansatz that is needed [93] when replica techniques are used to compute the thermodynamic behaviour of the model in Eq. (8). This corresponds to the universality class of chaotic (or random) free energy landscapes, as we now explain.

The free energy landscape of these systems is 'rugged', characterized by many minima and saddle points. Actually, the number of stationary points is so large that in order to count them one has to introduce an entropy, called configurational entropy or complexity, $s_c = 1/N \log \mathcal{N}(f)$, where $\mathcal{N}(f)$ is the number of stationary points with a given free energy density f. The density profile corresponding to one given minimum is amorphous and lacks any type of periodic long-range order, and different mimima are very different. Defining a similarity measure between them, an 'overlap', one typically finds that two minima with the same free energy f have zero overlap. The typical shape of the configurational entropy as a function of f is plotted in Fig 13.

At high temperature, there is typically a single minimum, the high temperature liquid state. There is a temperature below which an exponentially large (in the system size) number of minima appears. Within mean-field models, corresponding to Bethe or completely connected lattices, these minima correspond to macroscopic physical states analogous to the periodic minimum corresponding to the crystal[4]. Once the system is in one of these states it remains trapped in it forever, since barriers separating states diverge with the system size. However, when transposed to finite dimensional systems, these states become metastable and have a finite lifetime. As a consequence, in order to compute thermodynamic properties, one has to sum over all of them using the Boltzmann

[4]There is of course no crystal state in disordered systems such as in Eq. (8). In the case of lattice glass models, there is a crystal phase but it can disappear depending whether the Bethe lattice is a Cayley tree or a random regular graph.

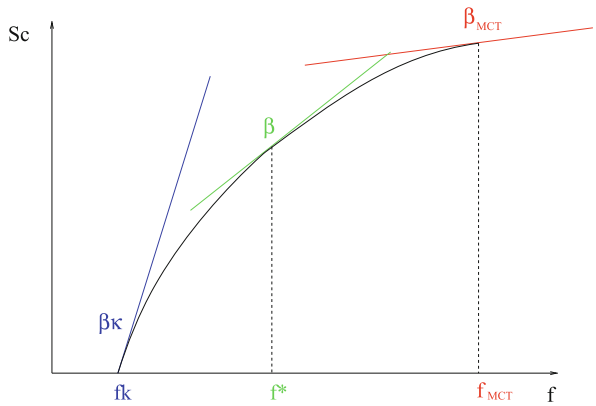

Glasses and Aging, A Statistical Mechanics Perspective on, Figure 13
Typical shape of the configurational entropy, s_c, as a function of free energy density, f in the range $T_k < T < T_{MCT}$ for random first order landscapes. A graphic solution of Eq. (10) is obtained by finding the value of f at which the slope of the curve is β. Note that s_c is also a function of temperature, so this curve in fact changes with T

weight $\exp(-\beta N f_\alpha)$ for each state α [135]:

$$Z = \sum_\alpha \mathrm{e}^{-\beta N f_\alpha} = \int \mathrm{d}f \exp[N s_c(f;T)] \mathrm{e}^{-\beta N f} \,, \tag{9}$$

where $\beta = 1/(K_B T)$. Evaluating this sum by saddle point method yields three regimes. At high temperature, $T > T_{MCT}$, the liquid corresponding to a flat density profile dominates the sum. The landscape is simple and has a single minimum. This is followed by an intermediate temperature regime, $T_K < T < T_{MCT}$, where the sum is dominated by all terms with free energy density satisfying

$$\left.\frac{\partial s_c(f,T)}{\partial f}\right|_{f=f^*} = \beta \,. \tag{10}$$

There are many of them, the logarithm of their number being given by $N s_c(f^*, T)$, see Fig. 13 for a graphical solution of Eq. (10). Upon decreasing the temperature, $s_c(f^*, T)$ decreases until a temperature, T_K, below which the sum in Eq. (9) becomes dominated by only few terms corresponding to states with free energy density f_K given by $s_c(f_K, T) = 0$, see Fig. 13. The entropy in the intermediate temperature range above T_K has two contributions: the one counting the number of minima, given by s_c, and the intra-state entropy, s_{in}, counting the number configurations inside each state. At T_K, the configurational entropy vanishes, $s_c(T_K) = 0$, As a consequence the specific heat undergoes a jump towards a smaller value across T_K,

an exact realization of the 'entropy vanishing' mechanism conjectured by Kauzmann [108].

Let us discuss the dynamical behaviour which results from the above analysis. We have already mentioned that relaxation processes do not occur below T_{MCT} because states have an infinite lifetime. The stability of these states can be analyzed by computing the free energy Hessian in the minima [48]. One finds that states become more fragile when $T \to T_{\mathrm{MCT}}^{-}$, are marginally stable at $T = T_{\mathrm{MCT}}$, unstable for $T > T_{\mathrm{MCT}}$. The dynamics of these models can be analyzed exactly [58]. Coming from high temperature, the dynamics slows down and the relaxation time diverges at T_{MCT} in a power law manner,

$$\tau_\alpha \sim \frac{1}{(T - T_{\mathrm{MCT}})^\gamma} , \tag{11}$$

where γ is a critical exponent. The physical reason is the incipient stable states that appear close to T_{MCT}. The closer the temperature is to T_{MCT}, the longer it takes to find an unstable direction to relax.

Amazingly, the dynamical transition that appears upon approaching T_{MCT} in random first order landscapes is completely analogous to the one predicted to occur in supercooled liquids by the Mode-Coupling Theory of the glass transition, and developed independently by Leuthesser, Bengtzelius, Götze, Sjölander and coworkers [94]. Actually, MCT can be considered as an approximation which becomes controlled and exact for these mean-field models. Originally, MCT was developed using projector operator formalism [13,122] and field-theory methods [64] to yield closed integro-differential equations for the dynamical structure factor in supercooled liquids. These approaches were recently generalized [34,36] to deal with dynamic heterogeneity and make predictions for the multi-point susceptibilities and correlation functions discussed in Sect. "Dynamic Heterogeneity". Within MCT, the relaxation timescale diverges in a power law fashion at T_{MCT}, as in Eq. (11). This divergence is accompanied by critical behaviour that appears both in space (long range spatial dynamic correlations), and in time (power laws in time).

Comparing Eqs. (1) and (11) makes it clear that MCT cannot be used to describe viscosity data close to T_g since it does not predict activated behaviour. It is now recognized that an MCT transition at T_{MCT} does not occur in real materials, so that T_{MCT} is, at best, a dynamical crossover. A central advantage of MCT, compared to many other theories (this includes the $T \approx T_{\mathrm{K}}$ regime of RFOT itself) is that it can yield quantitative predictions from microscopic input obtained for a particular material. As such it has been applied to yield predictions for scores of different systems that can be directly confronted to experimental or numerical measurements. A major drawback is the freedom offered by the 'crossover' nature of the MCT transition, so that 'negative' results can often be attributed to corrections to asymptotic predictions rather than deficiencies of the theory itself. Nevertheless, MCT has proven to be useful and continues to be developed, applied and generalized to study many different physical situations [94], including aging systems and non-linear rheology of glassy materials [18,83,134], see Sect. "Aging and Off-equilibrium".

What happens below T_{MCT} in finite dimensional system if the relaxation time does not diverge as predicted in Eq. (11)? Why is the transition avoided? In fact, the pletora of states that one finds in mean-field are expected to become (at best) metastable in finite dimension, with a finite lifetime, even below T_{MCT}. What is their typical lifetime, and how these metastable states are related to the structural relaxation are issues that still await for a complete microscopic analysis.

There exist, however, phenomenological arguments [38,112,171], backed by microscopic computations [71, 80] that yield a possible solution dubbed 'mosaic state' by Kirkpatrick, Thirumalai and Wolynes [112]. Schematically, the mosaic picture states that, in the regime $T_{\mathrm{K}} < T < T_{\mathrm{MCT}}$, the liquid is composed of domains of linear size ξ. Inside each domain, the system is in one of the mean-field states. The length of the domains is fixed by a competition between energy and configurational entropy. A state in a finite but large region of linear size l can be selected by appropriate boundary conditions that decrease its free energy by an amount which scales as Υl^θ with $\theta \leq 2$. On the other hand, the system can gain entropy, which scales as $s_c l^3$, if it visits the other numerous states. Entropy obviously gains on large lengthscales, the crossover length ξ being obtained by balancing the two terms,

$$\xi = \left(\frac{\Upsilon}{T s_c(T)} \right)^{1/(3-\theta)} . \tag{12}$$

In this scenario, the configurational entropy on scales smaller than ξ is too small to stir the configurations efficiently and win over the dynamically generated pinning field due to the environment, while ergodicity is restored at larger scale. Hence, the relaxation time of the system is the relaxation time, $\tau(\xi)$, of a finite size regions of the system. Barriers are finite, unlike in the mean-field treatment. Smaller length scales are faster but unable to decorrelate, whereas larger scales are orders of magnitude slower. Assuming thermal activation over energy barriers which are supposed to grow with size as ξ^ψ, one predicts finally, us-

ing Eq. (12), that [38]

$$\log\left(\frac{\tau_\alpha}{\tau_0}\right) = c\frac{\Upsilon}{K_B T}\left(\frac{\Upsilon}{T s_c(T)}\right)^{\psi/(3-\theta)}, \tag{13}$$

where c is a constant.

The above argument is rather generic and therefore not very predictive. Recent microscopic computations [71,80] aimed at putting these phenomenological arguments on a firmer basis and computing the exponents θ and ψ. The results are unfortunately not yet conclusive because they involve replica calculations with uncontrolled assumptions, but they do confirm the phenomenological scenario presented above and suggest that $\theta = 2$. Some other phenomenological arguments suggest the value of $\theta = 3/2$ [112]. There are no computation available for ψ, only the suggestion that $\psi = \theta$ [112].

Note that using the value $\theta = 3/2$ with $\theta = \psi$ simplifies Eq. (13) into a form that is well-known experimentally and relates $\log \tau_\alpha$ directly to $1/S_c$, which is the celebrated Adam–Gibbs relation [2] between relaxation time and configurational entropy that is in rather good quantitative agreement with many experimental results [5,96,105]. The Random First Order Theory can be considered, therefore, as a microscopic theory that reformulates and generalizes the Adam–Gibbs mechanism. Furthermore, using the fact that the configurational entropy vanishes linearly at T_K one predicts also a VFT divergence of the relaxation time as in Eq. (1), with the identification that

$$T_0 = T_K\ . \tag{14}$$

The equality (14) between two temperatures that are commonly used in the description of experimental data certainly constitutes a central achievement of RFOT since it accounts for the empirical relation found between the kinetics and the thermodynamics of supercooled liquids. Furthermore RFOT naturally contains MCT, which can be used to describe the first decades of the dynamical slowing down, while the spin glass side of RFOT qualitatively explains the dynamics in terms of the peculiar features of the free energy landscape that have been detailed above. Dynamics first slows down because there appear incipient metastable states, and once this metastable states are formed, the dynamics becomes dominated by the thermally activated barrier crossing from one metastable state to another, which is consistent with the relation between dynamical correlation length and timescale discussed in Sect. "Dynamic Heterogeneity". Quite importantly, microscopic computations of T_{MCT} and T_0 for realistic models of liquids are possible [136]. Remarkably, the jamming transition of hard spheres systems has been also studied with these techniques and a clear connection with the glass transition has emerged within RFOT [143]. This quantitative "side" of RFOT is a most desirable feature, even if the results are not always quantitatively accurate [54,138,156].

Probably the most serious weakness of the RFOT construction is that the theory, although worked out in full details within mean-field models, has remained elusive for finite dimensional systems, for which it has a highly speculative flavour. Worrying is the fact that no simple three-dimensional glassy model, let alone interacting particles in the continuum, has been discovered, for which this theory has been shown to apply, and the entropy driven nucleation theory that leads to the VFT law is not understood completely. Although the ultimate consequences of the theory are sometimes in very good agreement with experiments, as Eq. (14), one should not conclude that RFOT is correct. In fact direct tests of the mosaic state picture are rare, and rather inconclusive [49]. One can hope that in the next few years, joint theoretical and experimental efforts will drive RFOT into a corner, to a point where it can be decided whether it is truly a valid theory for the glass transition.

Free Volume, Defects, and Facilitated Models

In this subsection we motivate and briefly summarize studies of a different family of statistical mechanics models that turns out to yield a rich variety of physical behaviours. Their starting point are physical assumptions that might seem similar to the models described in Sect. "Cooperativity, Chaotic Energy Landscapes and Random First Order Theory", but the outcome yields a different physical explanation of the glass transition. Although the two theoretical approaches cannot be simultaneously correct, they both have been influential and very instructive in order to develop a theoretical understanding of glassy phenomena. Furthermore, despite the 'great unification' phase diagram in Fig. 4, it could be that glass and jamming transitions in colloids, granular media and glass-formers have a different nature, so that different theories could apply to different phenomena.

As in Sect. "Cooperativity, Chaotic Energy Landscapes and Random First Order Theory", we start from the packing considerations that are more appropriate for hard spheres systems. We follow first Kob and Andersen [115] and again use a lattice gas description of the physics and work on a three dimensional cubic lattice. As in a hard sphere system, we assume no interaction between particles beyond the hard-core constraint that the occupation number n_i at site i is at most equal to 1,

$$H\left[\{n_i\}\right] = 0, \quad n_i = 0,\ 1\ . \tag{15}$$

Contrary to the lattice glass model presented above, all configurations respecting the hard-core constraint are allowed and are equally probable. Geometric frustration is instead introduced at the level of the kinetic rules, that are defined as constrained local moves. Namely, a particle can jump to a nearest neighbor site only if that site is empty (to satisfy the hard-core constraint), but, additionnally, only if the sites occupied before and after the move have less than m neighbors, m being an adjustable parameter, which Kob and Andersen choose as $m = 4$ for $d = 3$ ($m = 6$ corresponds to the unconstrained lattice gas). The model captures the idea that if the liquid is locally very dense, no movement is possible while regions with low density move more easily.

Of course, such kinetically constrained lattice gases have been studied in various spatial dimensions, for different values of m, for different constraints, or even different lattice geometries [146]. They can be thought of as models capturing the idea of a 'cage' effect in a strict sense, utilizing the notion that a particle with a dense neighbor shell cannot diffuse. Although the cage seems a purely local concept, it turns out that diffusion in constrained lattice gases arises from cooperative rearrangments, so that slow dynamics can be directly shown to be driven by the growth of dynamic lengthscales for these cooperative moves [77,141,157]. This strongly suggests that such cooperative moves are most probably at work also in real liquids.

In this lattice gas picture, the connection with liquid is not obvious because density ('free volume'), rather than temperature controls the dynamics. Thermal models with similar features can in fact be defined along the following lines. In a liquid, low temperature implies a very small probability to find a location with enough free volume to move. The idea of a small concentration of 'hot spots' is in fact reminiscent of another picture of the glass transition based on the idea of 'defects' which is captured by the defect model proposed by Glarum [87] in the 60's, where relaxation proceeds via the diffusion of a low concentration of independent defects. In the mid-80's, using the conjugated ideas of kinetic constraints and rare defects, Fredrickson and Andersen defined a family of kinetic Ising models for the glass transition [81]. They study an assembly of non-interacting spins,

$$H[\{n_i\}] = \sum_{i=1}^{N} n_i, \quad n_i = 0,\ 1\ , \tag{16}$$

where $n_i = 1$ represent the defects, whose concentration becomes exponentially small at low temperature, $\langle n_i \rangle \approx \exp(-1/T)$. As for the Kob–Andersen lattice gas, the non-trivial ingredient lies in the chosen rates for the transition between states. The kinetic rules stipulate that a transition at site i can happen with a usual Glauber rate, but only if site i is surrounded by at least k defects ($k = 0$ corresponds to the unconstrained limit). Again, one can easily imagine studying such models in different spatial dimensions, on different lattices, and with slightly different kinetic rules, yielding a large number of possible behaviours [125,146]. The similarity between those spin facilitated models and the kinetically constrained lattice gases is striking. Altogether, they form a large family of models generically called kinetically constrained models (KCMs) [146].

The connection between KCMs and the much older concept of free volume is obvious from our presentation. Free volume models are among the most widely used models to analyze experimental data, especially in polymeric systems. They have been thoroughly reviewed before [53,66], and the main prediction is that dynamic slowing down occurs because the free volume available to each particle, v_f, vanishes at some temperature T_0 as $v_f \approx \alpha(T - T_0)$, a relation which connects volume to temperature. Statistical arguments then relate relaxation timescales to free volume assuming that movement is possible if locally there is 'enough' free volume available, more than a typical value v_0. This is clearly reminiscent of the above idea of a kinetic constraint for local moves in lattice gases. An appealing VFT divergence is then predicted:

$$\frac{\tau_\alpha}{\tau_0} \sim \exp\left(\gamma \frac{v_0}{v_f}\right) \sim \exp\left(\frac{\gamma v_0/\alpha}{[T - T_0]^\mu}\right)\ , \tag{17}$$

where γ is a numerical factor and $\mu = 1$. Predictions such as Eq. (17) justify the wide use of free volume approaches, despite the many (justified) criticisms that have been raised.

Initially it was suggested that KCMs would similarly display finite temperature or finite density dynamic transitions similar to the one predicted by the mode-coupling theory of supercooled liquids [81], but it was soon realized [46,82] that most KCMs do not display such singularity, and timescales in fact only diverge in the limit of zero temperature ($T = 0$) or maximal density ($\rho = 1$). Recently, models displaying a $T_0 > 0$ or $\rho_0 < 1$ transition have been introduced and analyzed [158]. They provide a microscopic realization, based on well-defined statistical mechanics models, of the glass transition predicted by free volume arguments. Their relaxation timescale diverges with a VFT-like form but with an exponent $\mu \simeq 0.64$. Understanding their universality classes and how general is the mechanism leading to the transition is still an open problem.

Extensive studies have shown that KCMs have a macroscopic behaviour which resembles the phenomenology of supercooled liquids, displaying in particular Arrhenius or super-Arrhenius increase of relaxation timescales on decreasing the temperature and non-exponential relaxation functions at equilibrium [146]. Early studies also demonstrated that, when suddenly quenched to very low temperatures, the subsequent non-equilibrium aging dynamics of the models compares well with experimental observations on aging liquids [82]. Moreover, the many possibilities to define the models mean that they might exhibit a broad variety of possible behaviours. This is both a positive and a negative aspect: on the one hand one can explore various scenarii to describe glass transition phenomena, but on the other hand, one would like to be able to decide what particular model should be used if one wants to get a predictive quantitative description for a particular liquid. In fact, contrary to MCT, no microscopic calculations have been performed using the framework of KCMs. Rather than predicting the quantitative behaviour of a material in all its microscopic details, it is perhaps more appropriate to use KCMs as theoretical tools to define concepts and obtain new ideas.

It is precisely in this perspective that interest in KCMs continues to increase, in large part since it it was realized that their dynamics is spatially heterogeneous [46,77,84], a central feature of supercooled liquids dynamics. It is only fair to say that in-depth studies of KCMs have greatly contributed to our theoretical understanding of the spatially heterogeneous dynamics in glass and jamming problems. Remarkably, virtually all the aspects related to dynamic heterogeneity mentioned in Sect. "Dynamic Heterogeneity" can be investigated and rationalized, at least qualitatively, in terms of KCMs. The dynamics of these systems can be understood in terms of defects motion [146]. Depending on the particular model, defects can diffuse or have a more complicated motion. Furthermore, they can simply be point-like, or 'cooperative' (formed by point-like defects moving in a cooperative way). A site can relax when it is visited by a defect. As a consequence, the heterogeneous character of the dynamics is entirely encoded in the defect configuration and defect motion [84]. For instance, a snapshot similar to Fig. 7 in a KCM shows clusters which have relaxed within the time interval t [26,167]. These are formed by all sites visited by a defect between 0 and t. The other sites are instead frozen in their initial state. In these models the dynamics slows down because the defect concentration decreases. As a consequence, in the regime of slow dynamics there are few defects and strong dynamic heterogeneity. Detailed numerical and analytical studies have indeed shown that in these systems, non-exponential relaxations patterns do stem from a spatial, heterogeneous distribution of timescales, directly connected to a distribution of dynamic lengthscales [84,103,141,157,158,167]. Decoupling phenomena appear in KCMs and can be shown to be very direct, quantifiable, consequences of the dynamic heterogeneity [106], which also deeply affects the process of self-diffusion in a system close to its glass transition [24]. More fundamentally, multi-point susceptibilities, multi-point spatial correlation functions such as the ones defined in Eqs. (3) and (6) can be studied in much greater detail than in molecular systems, to the point that scaling relations between timescales, lengthscales, and dynamic susceptibilities can be established [22,50,141,159,168]. This type of scaling behaviour has been observed close to $T = 0$ and $\rho = 1$ in spin models and lattice gases without a transition[5]. These particular points of the phase diagram have been shown, by various theoretical means, to correspond to true critical points where timescales and dynamic lengthscales diverge with well-defined critical laws [103,168]. Such 'dynamic criticality' is a useful concept because it implies the possibility that some universal behaviour emerges in the physics of supercooled liquids, precisely of the type observed in Fig. 12.

A central criticism about the free volume approach, that is equally relevant for KCMs concerns the identification, at the molecular level, of the vacancies (in lattice gases), mobility defects (in spin facilitated models), or free volume itself. Attempts to provide reasonable coarse-graining from molecular models with continuous degrees of freedom to lattice models with kinetic rules are so far very limited, and not really convincing [70,163]. On the other hand the proof that kinetic rules can emerge effectively and induce a slow dynamics has been obtained for simple lattice spin models [86], whose dynamics directly maps onto constrained models. Several examples are available but here we only mention the simple case of the bidimensional plaquette model defined by a Hamiltonian of a p-spin type, but in two dimensions on a square lattice of linear size L,

$$H = -J \sum_{i=1}^{L-1} \sum_{j=1}^{L-1} S_{i,j} S_{i+1,j} S_{i,j+1} S_{i+1,j+1} \,, \tag{18}$$

where $S_{i,j} = \pm 1$ is an Ising variable lying at node (i, j) of the lattice. Contrary to KCMs, the Hamiltonian in Eq. (18) contains genuine interactions, which are no less (or no more) physical than p-spin models discussed in

[5] A critical (different) behaviour is expected and predicted for models having a transition [158].

Sect. "Cooperativity, Chaotic Energy Landscapes and Random First Order Theory". Interestingly the dynamics of this system is (trivially) mapped onto that of a KCM by analyzing its behaviour in terms of plaquette variables, $p_{i,j} \equiv S_{i,j}S_{i+1,j}S_{i,j+1}S_{i+1,j+1}$, such that the Hamiltonian becomes a non-interacting one, $H = -J\sum_{i,j} p_{i,j}$, as in Eq. (16). More interestingly, the analogy also applies to the dynamics [86]. The fundamental moves are spin-flips, but when a single spin is flipped the states of the four plaquettes surrounding that spin change. Considering the different types of moves, one quickly realizes that excited plaquettes, $p_{i,j} = +1$, act as sources of mobility, since the energetic barriers to spin flips are smaller in those regions. This observation allows to identify the excited plaquettes as defects, by analogy with KCMs. Spatially heterogenous dynamics, diverging lengthscales accompanying diverging timescales and scaling behaviour sufficiently close to $T = 0$ can be established by further analysis [101], providing a simple, but concrete example, of how an interacting many body system might effectively behave as a model with kinetic constraints[6].

Another essential drawback of facilitated models is that among the microscopic 'details' thrown away to arrive at simple statmech models such as the ones in Eqs. (15) and (16), information on the thermodynamic behaviour of the liquids has totally disappeared. In particular, a possible coincidence between VFT and Kauzmann temperatures, T_0 and T_K is not expected, nor can the dynamics be deeply connected to thermodynamics, as in Adam–Gibbs relations. The thermodynamic behaviour of KCMs appears different from the one of real glass-formers close to T_g [35]. This is probably the point where KCMs and RFOT approaches differ more evidently. Even though the dynamics of KCMs shares similarities with systems characterized with a complex energy landscape [25,169], thermodynamical behaviours are widely different in both cases, as has been recently highlighted in [102] by focusing on the concrete examples of plaquette models such as in Eq. (18).

Finally, when KCMs were first defined, they were argued to display a dynamic transition of a nature very similar to the one predicted by MCT [81]. Although the claim has been proven wrong[7], it bears some truth: both approaches basically focus on the kinetic aspects of the glass transition and they both predict the existence of some dynamic criticality with diverging lengthscales and timescales. This similarity is even deeper, since a mode-coupling singularity is truly present when (some) KCMs are studied on the Bethe lattice [157], but is 'avoided' when more realistic lattice geometries are considered. This underlies the similarity of these two approaches while emphasizing further the mean-field character of the MCT approach.

Geometric Frustration, Avoided Criticality, and Coulomb Frustrated Theories

In all of the above models, 'real space' was present in the sense that special attention was paid to different lengthscales characterizing the physics of the models that were discussed. However, apart from the 'packing models' with hard-core interactions, no or very little attention was paid to the geometric structure of local arrangments in molecular liquids close to a glass transition. This slight 'oversight' is generally justified using concepts such as 'universality' or 'simplicity', meaning that one studies complex phenomena using simple models, a typical statistical mechanics perspective. However, important questions remain: what is the liquid structure within mosaic states? How do different states differ? What is the geometric origin of the defects invoked in KCMs? Are they similar to defects (disclinations, dislocations, vacancies, etc.) found in crystalline materials?

There exists a line of research in this field which attempts to provide answers to these questions. It makes heavy use of the concept of geometric frustration. Broadly speaking, frustration refers to the impossibility of simultaneously minimizing all the interaction terms in the energy function of the system. Frustration might arise from quenched disorder (as in the spin glass models described above), but liquids have no quenched randomness. In that case, frustration has a purely geometric origin. It is attributed to a competition between a short-range tendency for the extension of a 'locally preferred order', and global constraints that prevent the periodic tiling of space with this local structure.

This can be illustrated by considering once more the packing problem of spheres in three dimensions. In that case, the locally preferred cluster of spheres is an icosahedron. However, the 5-fold rotational symmetry characteristic of icosahedral order is not compatible with translational symmetry, and formation of a periodic icosahedral crystal is impossible [75]. The geometric frustration that affects spheres in three dimensional Euclidean space can be relieved in curved space [139]. In Euclidian space, the system possesses topological defects (disclination lines), as the result of forcing the ideal icosahedral ordering into

[6] This type of plaquette models, and other spin models, were introduced originally [123,149] to show how ultra-slow glassy dynamics can emerge because of growing free energy barriers.

[7] Most KCMs do not have a finite temperature dynamical transition and the ones displaying a transition have critical properties different from MCT.

a 'flat' space. Nelson and coworkers have developed a solid theoretical framework based on this picture to suggest that the slowing down of supercooled liquids is due to the slow wandering of these topological defects [139], but their treatment remains so complex that few quantitative, explicit results have been obtained.

This picture of sphere packing disrupted by frustration has been further developed in simple statistical models characterized by geometric frustration, in a pure statistical mechanics approach [155]. To build such a model, one must be able to identify, then capture, the physics of geometric frustration. Considering a locally ordered domain of linear size L, Kivelson et al. [114] suggest that the corresponding free energy scales as

$$F(L,T) = \sigma(T)L^2 - \phi(T)L^3 + s(T)L^5 \ , \tag{19}$$

The first two terms expresses the tendency of growing local preferred order and they represent respectively the energy cost of having an interface between 2 phases and a bulk free energy gain inside the domain. Geometric frustration is encoded in the third term which represents the strain free energy resulting from the frustration. The remarkable feature of Eq. (19) is the super-extensive scaling of the energy cost due to frustration which opposes the growth of local order. The elements in Eq. (19) can then be directly incorporated into ferromagnetic models where 'magnetization' represents the local order, ferromagnetic interactions the tendency to local ordering, and Coulombic antiferromagnetic interactions the opposite effect of the frustration. The following Hamiltonian possesses these minimal ingredients:

$$H = -J\sum_{\langle i,j\rangle} \mathbf{S}_i \cdot \mathbf{S}_j + K\sum_{i\neq j} \frac{\mathbf{S}_i \cdot \mathbf{S}_j}{|\mathbf{x}_i - \mathbf{x}_j|} \ , \tag{20}$$

where the spin $\mathbf{S}_i$ occupies the site i at position $\mathbf{x}_i$. Such Coulomb frustrated models have been studied in great detail, using various approximations to study models for various space and spin dimensions [155].

The general picture is that the ferromagnetic transition occuring at $T = T_c^0$ in the pure model with no frustration, $K = 0$, is either severely depressed to lower temperatures for $K > 0$, sometimes with a genuine discontinuity at $K \to 0$, yielding the concept of 'avoided criticality'. For the simple case of Ising spins in $d = 3$, the situation is different since the second order transition becomes first order between a paramagnetic phase and a spatially modulated phase (stripes). For $K > 0$ and $T < T_c^0$ the system is described as a 'mosaic' of domains corresponding to some local order, whose size increases (but does not diverge!) when T decreases. Tarjus, Kivelson and co-workers clearly demonstrate that such a structuration into mesoscopic domains allows one to understand most of the fundamental phenomena occuring in supercooled liquids [155]. Their picture as a whole is very appealing because it directly addresses the physics in terms of the 'real space', and the presence of domains of course connects to ideas such as cooperativity, dynamic heterogeneity and spatial fluctuations, that directly explains, at least qualitatively, non-exponential relaxation, decoupling phenomena or super-Arrhenius increase of the viscosity. However, as for the RFOT mosaic picture, direct confirmations of this scenario are rare [56]. Icosahedral order has not been clearly linked to the dynamics of hard spheres, while the very notion of local order in more realistic glass-forming liquids (such binary mixtures of spheres, or larger molecules with internal degrees of freedom) is problematic and not easily defined. This makes the basis of the scenario very fragile, and its practical applicability for a particular material difficult.

Aging and Off-equilibrium

Why Aging?

We have dedicated most of the above discussion to properties of materials approaching the glass transition at thermal equilibrium. We discussed a rich phenomenology and serious challenges for both our numerical and analytical capabilities to account for these phenomena. For most people, however, glasses are interesting below the glass transition, so deep in the glass phase that the material seems to be frozen forever in a seemingly arrested amorphous state, endowed with enough mechanical stability for a glass to retain, say, the liquid it contains (preferentially a nice red wine). Does this mean that there is no interesting physics in the glass state?

The answer is clearly 'no'. There is still life (and physics) below the glass transition. We recall that for molecular glasses, T_g is defined as the temperature below which relaxation is too slow to occur within an experimental timescale. Much below T_g, therefore, the equilibrium relaxation timescale is so astronomically large that thermal equilibrium is out of reach. One enters therefore the realm of off-equilibrium dynamics. A full physical understanding of the non-equilibrium glassy state remains a central challenge [9,173].

A first consequence of studying materials in a time window smaller than equilibrium relaxation timescales is that the system can, in principle, remember its complete history, a most unwanted experimental situation since all details of the experimental protocol may then matter. The simplest protocol to study aging phenomena in the glass

phase is quite brutal [152]: take a system equilibrated above the glass transition and suddenly quench it at a low temperature at a 'waiting time' $t_w = 0$ which corresponds to the beginning of the experiment. For $t_w > 0$ the system is left unperturbed at constant temperature where it tries to slowly reach thermal equilibrium, even though it has no hope to ever get there. Aging means that the system never forgets the time t_w spent in the glass phase, its 'age'. The evolution of one time quantities, e. g. the energy, as a function of time are not a good evidence of aging. In order to show that the system never equilibrates two time quantities, such as density-density or spin-spin correlation functions, are much more useful. A typical example is presented in Fig. 14 where the self-part of the intermediate function in Eq. (2) is shown for a Lennard-Jones molecular liquid at low temperature. Immediately after the quench, the system exhibits a relatively fast relaxation: particles still move substantially. However, when the age of the system increases, dynamics slows down and relaxation becomes much slower. When t_w becomes very large, relaxation becomes too slow to be followed in the considered time window and the system seems frozen on that particular timescale: it has become a glass. A striking feature conveyed by these data is that an aging system not only remains out-of-equilibrium for all practical purposes, but its typical relaxation time is in fact set by its age t_w. In simple cases, the effective relaxation time after waiting a time t_w scales at t_w itself, which means that since equilibration timescales have diverged, t_w is the only remaining relevant timescale in the problem.

A popular interpretation of this phenomenon is given by considering trap models [40]. In this picture, reminiscent of the Goldstein view of the glass transition mentioned above [91], the system is described as a single particle evolving in a complex energy landscape with a broad distribution of trap depths—a paradigmatic mean-field approach. Aging in this perspective arises because the system visits traps that are increasingly deep when t_w increases, corresponding to more and more stable states. It takes therefore more and more time for the system to escape, and the dynamics slows down with time, as observed in Fig. 14. This implies that any physical property of the glass becomes an age-dependent quantity in aging protocols, and more generally dependent on how the glass was prepared. One can easily imagine using this property to tune mechanical or optical characteristics of a material by simply changing the way it is prepared, like how fast it is cooled to the glassy state.

A real space alternative picture was promoted in particular in the context of spin glass studies, based on the ideas of scaling and renormalization [42,74]. The physical picture is that of a coarsening process, where the system develops long-range order by growing extended domains of lengthscale $\ell(t_w)$. On lengthscales less than $\ell(t_w)$ the system has ordered since the quench at $t_w = 0$. The domain walls evolve in a random environment. In order to move they have to overcome free energy barriers. It is then assumed an activated dynamic scaling which states that the typical barrier to extend the domain from linear size $\ell(t_w)$ to, say, $2\ell(t_w)$ scales as ℓ^{ψ}, where ψ is some 'barrier' exponent. Using the Arrhenius law to relate dynamics to barriers, one gets that aging corresponds to the logarithmic growth with time of spatially correlated domains, $\ell \sim (T \log t_w)^{1/\psi}$. A domain growth picture of aging in spin glasses can be directly confirmed by numerical simulations [113], only indirectly by experiments.

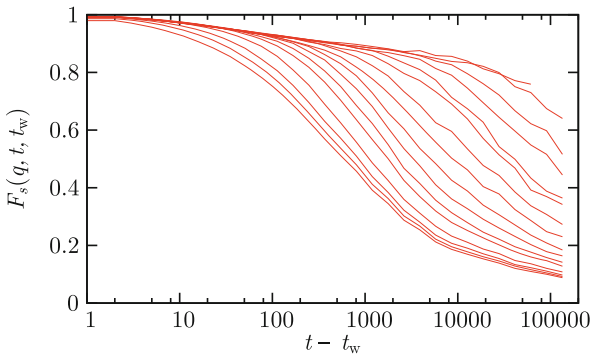

Glasses and Aging, A Statistical Mechanics Perspective on, Figure 14
Aging dynamics in a Lennard-Jones glass-forming liquid at low temperature. The system is quenched at time $t_w = 0$ at low T, where the temperature is kept constant. Two-time self-intermediate scattering functions are then measured for 20 logarithmically spaced waiting times t_w from $t_w = 1$ to $t_w = 10^5$ (from *left* to *right*). The relaxation becomes slower when t_w increases: the system ages

Memory and Rejuvenation Effects

Since the complete history of a sample in the glass phase matters, there is no reason to restrain experimental protocols to the simple aging experiment mentioned above. Indeed, experimentalists have investigated scores of more elaborated protocols that have revealed an incredibly rich, and sometimes quite unexpected, physics [173]. We restrain ourselves here to a short discussion of memory and rejuvenation effects observed during temperature cycling experiments [148] (one can imagine applying a magnetic field or a mechanical constraint, be they constant in time or sinusoidal, etc.). These two effects were first observed in spin glasses, but the protocol was then repeated in

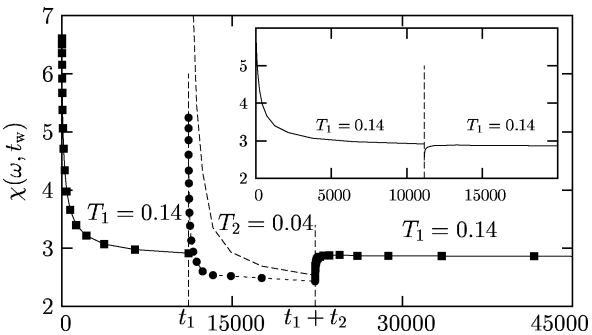

Glasses and Aging, A Statistical Mechanics Perspective on, Figure 15

Memory and rejuvenation effects obtained in the numerical simulation of a three-dimensional Heisenberg spin glass. There is a first aging step, $0 < t_w < t_1$, during which the system slowly tries to reach thermal equilibrium at temperature T_1. The system 'rejuvenates' in the second step at T_2, $t_1 < t < t_1 + t_2$, and it restart aging (rejuvenation). Finally in the third step, temperature is back to T_1, and memory of the first step is kept intact, as shown in the *inset* where relaxation during the second step of the experiment is taken away

many different materials, from polymers and organic liquids to disordered ferroelectrics. After several unsuccessfull attempts, similar effects are now observed in numerical work as well. Results obtained from simulations of a three-dimensional Heisenberg spin glass [32] are presented in Fig. 15.

There are three steps in temperature cycling experiments [148]. The first one is a standard aging experiment, namely a sudden quench from high to low temperature at time $t_w = 0$. The system then ages for a duration t_1 at constant temperature T_1. The system slowly relaxes towards equilibrium and its dynamics slows down, as observed in our spin glass example in Fig. 15 through the measurement of the magnetic susceptibility $\chi(\omega, t_w)$. Temperature is then suddenly shifted to $T_2 < T_1$ at time t_1. There, the material restarts aging (almost) as if the first step had not taken place. This is called 'rejuvenation effect', because the system seems to forget it is already 'old'. At total time $t_1 + t_2$, temperature is then shifted back to its initial value T_1. Then, aging is found to proceed as a quasi-perfect continuation of the first step, as if the second step had not taken place. The system has kept the 'memory' of the first part of the experiment, despite the rejuvenation observed in the intermediate part. The memory effect becomes more spectacular when relaxation during the second step is removed, as in the inset of Fig. 15. The third relaxation appears indeed as a perfect continuation of the first one.

On top of being elegant and quite intriguing, such protocols are relevant because they probe more deeply the dynamics of aging materials, allowing one to ask more precise questions beyond the simplistic observation that 'this material displays aging'. Moreover, the observation of similar effects in many different glassy materials implies that these effects are intrinsic to systems with slow dynamics. Interesting also are the subtle differences observed from one material to the other.

Several experimental, numerical and theoretical papers have been devoted to this type of experiments, and these effects are not 'mysterious' anymore [31]. A clear link between memory effects and typical lengthscales over which the slow dynamics takes place has been established. Because lengthscales depend so sensitively on timescales and on the working temperature, experiments performed at two different temperatures typically probe very different lengthscales, allowing the system to store memory of its state at different temperatures at different lengthscales [23,39]. In return, this link has been elegantly exploited to obtain a rather precise experimental estimate of dynamic lengthscales involved in the aging dynamics of spin glass materials [15], which seems to confirm the slow logarithmic growth law mentioned before.

Discussion of the rejuvenation effect is slightly more subtle. It is indeed not yet obvious that the effect as it is observed in computer simulations and reported, e. g., in Fig. 15 is exactly similar to the one observed in experiments. The difficulty comes from the fact that some seemingly innocuous details of the experimental protocol, such as the necessary use in experiments of finite cooling rates, in fact play a crucial role and influence the physics so that direct comparison between experiments and simulations is difficult. In numerical work, rejuvenation can be attributed to a gradual change with temperature of the nature of spatial correlations between spins that develop with time [23,32]. More drastic changes are predicted to occur in disordered systems as a result of the chaotic evolution with temperature of the metastable states in a spin glass (so-called 'chaos effect' [43]), that could also be responsible for the observed rejuvenation effect [107]. This scenario can be directly discarded in simulations, where spatial correlations can be easily measured and chaos sought (in vain) in a very direct manner. Understanding the very origin of the rejuvenation effect observed in experiments remains, however, a challenge.

Mean-Field Aging and Effective Temperatures

Theoretical studies of mean-field glassy models have provided important insights into the aging dynamics of both structural and spin glasses [60,61]. Although such models are defined in terms of spin degrees of freedom interact-

ing via infinite-ranged interactions, the deep connections between them and the mode-coupling theory of the glass transition make them serious candidates to investigate glassy states in general, not only thermodynamic properties at thermal equilibrium but also non-equilibrium aging dynamics. Despite their often reported 'simplicity', it took several years to derive a proper asymptotic solution of the long-time dynamics for a series of mean-field spin glasses [58]. These results have then triggered an enormous activity [57] encompassing theoretical, numerical and also experimental work trying to understand further these results, and to check in more realistic systems whether they have some reasonable range of applicability beyond mean-field. This large activity, by itself, easily demonstrates the broad interest of these results.

In these mean-field models, thermal equilibrium is never reached, and aging proceeds by downhill motion in an increasingly flat free energy landscape [119], with subtle differences between spin glass and structural glass models. In both cases, however, time translational invariance is broken, and two-time correlation and response functions depend on both their time arguments. In fact, the exact dynamic solution of the equations of motion for time correlators displays behaviours in strikingly good agreement with the numerical results reported in Fig. 14.

In these systems, the equations of motion in the aging regime involve not only time correlations, but also time-dependent response functions. At thermal equilibrium response and correlations are not independent, since the fluctuation-dissipation theorem (FDT) relates both quantities. In aging systems, there is no reason to expect the FDT to hold and both quantities carry, at least in principle, distinct physical information. Again, the asymptotic solution obtained for mean-field models quantitatively establishes that the FDT does not apply in the aging regime. Unexpectedly, the solution also shows that a generalized form of the FDT holds at large waiting times [60]. This is defined in terms of the two-time connected correlation function for some generic observable $A(t)$,

$$C(t, t_{\rm w}) = \langle A(t)A(t_{\rm w})\rangle - \langle A(t)\rangle\langle A(t_{\rm w})\rangle \ , \tag{21}$$

with $t \geq t_{\rm w}$, and the corresponding two-time (impulse) response function

$$R(t, t_{\rm w}) = T\frac{\delta\langle A(t)\rangle}{\delta h(t_{\rm w})}\bigg|_{h=0} \ . \tag{22}$$

Here h denotes the thermodynamically conjugate field to the observable A so that the perturbation to the Hamiltonian (or energy function) is $\delta E = -hA$, and angled brackets indicate an average over initial conditions and any stochasticity in the dynamics. Note that we have absorbed the temperature T in the definition of the response, for convenience. The associated generalized FDT reads then

$$R(t, t_{\rm w}) = X(t, t_{\rm w})\frac{\partial}{\partial t_{\rm w}}C(t, t_{\rm w}) \ , \tag{23}$$

with $X(t, t_{\rm w})$ the so-called fluctuation-dissipation ratio (FDR). At equilibrium, correlation and response functions are time translation invariant, depending only on $\tau = t - t_{\rm w}$, and equilibrium FDT imposes that $X(t, t_{\rm w}) = 1$ at all times. A parametric fluctuation-dissipation (FD) plot of the step response or susceptibility

$$\chi(t, t_{\rm w}) = \int_{t_{\rm w}}^{t} {\rm d}t' \ R(t, t') \ ,$$

against

$$\Delta C(t, t_{\rm w}) = C(t, t) - C(t, t_{\rm w}) \ ,$$

is then a straight line with unit slope. These simplifications do not occur in non-equilibrium systems. But the definition of an FDR through Eq. (23) becomes significant for aging systems [60,61]. In mean-field spin glass models the dependence of the FDR on both time arguments is only through the correlation function,

$$X(t, t_{\rm w}) \sim X(C(t, t_{\rm w})) \ , \tag{24}$$

valid at large wait times, $t_{\rm w} \to \infty$. For mean-field structural glass models, the simplication (24) is even more spectacular since the FDR is shown to be characterized by only two numbers instead of a function, namely $X \sim 1$ at short times (large value of the correlator) corresponding to a quasi-equilibrium regime, with a crossover to a non-trivial number, $X \sim X^{\infty}$ for large times (small value of the correlator). This implies that parametric FD plots are simply made of two straight lines with slope 1 and X^{∞}, instead of the single straight line of slope 1 obtained at equilibrium.

Since any kind of behaviour is in principle allowed in non-equilibrium situations, getting such a simple, equilibrium-like structure for the FD relations is a remarkable result. This immediately led to the idea that aging systems might be characterized by an effective thermodynamic behaviour and the idea of quasi-equilibration at different timescales [59]. In particular, generalized FD relations suggest to define an effective temperature, as

$$T_{\rm eff} = \frac{T}{X(t, t_{\rm w})} \ , \tag{25}$$

such that mean-field glasses are characterized by a unique effective temperature, $T_{\rm eff} = T/X^{\infty}$. It is thought of as the

temperature at which slow modes are quasi-equilibrated. One finds in general that $0 < X^\infty < 1$, such that $T_{\rm eff} > T$, as if the system had kept some memory of its high temperature initial state.

The name 'temperature' for the quantity defined in Eq. (25) is not simply the result of a dimensional analysis but has a deeper, physically appealing meaning that is revealed by asking the following questions. How does one measure temperatures in a many-body system whose relaxation involves well-separated timescales? What is a thermometer (and a temperature) in a far from equilibrium aging material? Answers are provided in Refs. [59, 120] both for mean-field models and for addditional toy models with multiple relaxation timescales. The idea is to couple an additional degree of freedom, such as a harmonic oscillator, $x(t)$, which plays the role of the thermometer operating at frequency ω, to an observable of interest $A(t)$ via a linear coupling, $-\lambda x(t)A(t)$. Simple calculations show then that the thermometer 'reads' the following temperature,

$$\frac{1}{2}K_{\rm B}T^2_{\rm meas} \equiv \frac{1}{2}\omega^2\langle x^2\rangle = \frac{\omega C'(\omega, t_{\rm w})}{2\chi''(\omega, t_{\rm w})}\,, \tag{26}$$

where $C'(\omega, t_{\rm w})$ is the real part of the Fourier transform of Eq. (21), and $\chi(\omega, t_{\rm w})$ the imaginary part of the Fourier transform of Eq. (22), with $h = \lambda x$. The relation (26) indicates that the bath temperature is measured, $T_{\rm meas} = T$, if frequency is high and FDT is satisfied, while $T_{\rm meas} = T_{\rm eff} > T$ if frequency is slow enough to be tuned to that of the slow relaxation in the aging material. The link between the FDR in Eq. (23) and the effective temperature measured in Eq. (26) was numerically confirmed in the computer simulation of a glassy molecular liquid in Ref. [19].

More generally, relaxation in glassy systems occurs in well-separated time sectors [61]; it is then easy to imagine that each sector could be associated with an effective temperature [120]. A thermodynamic interpretation of effective temperatures has also been put forward, relating them to the concept of replica symmetry breaking [78]. Interestingly, the full-step or one-step replica symmetry breaking schemes needed to solve the static problem in these models have a counterpart as the FDR being a function or a number, respectively, in the aging regime. Moreover, we note that these modern concepts are related to, but make much more precise, older ideas of quasi-equilibrium and fictive temperatures in aging glasses [152].

Taken together, these results make the mean-field description of aging very appealing, and they nicely complement the mode-coupling/RFOT description of the equilibrium glass transition described above. Moreover, they have set the agenda for a large body of numerical and experimental work, as reviewed in [57]. In Fig. 16 we present recent numerical data obtained in an aging silica glass [30], presented in the form of a parametric response-correlation plot. The measured correlation functions are the self-part of the intermediate scattering functions defined in Eq. (2), while the conjugated response functions quantify the response of particle displacements to a spatially modulated field conjugated to the density. Plots for silicon and oxygen atoms at different ages of the system are presented. They seem to smoothly converge towards a two-straight line plot, as obtained in mean-field models (note, however, that this could be just a pre-asymptotic, finite "$t_{\rm w}$", effect). Moreover, the second, non-trivial part of the plot is characterized by a slope that appears to be independent of the species, and of the wavevector chosen to quantity the dynamics, in agreement with the idea of a unique asymptotic value of the FDR, possibly related to a well-defined effective temperature.

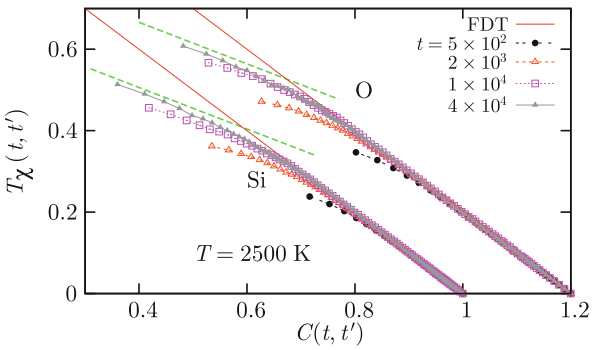

Glasses and Aging, A Statistical Mechanics Perspective on, Figure 16

Parametric correlation-response plots measured in the aging regime of a numerical model for a silica glass, SiO_2 [30]. The plots for both species smoothly converge towards a two-straight line plot of slope 1 at short times (large C values), and of slope $X^\infty \approx 0.51$ at large times (small values of C), yielding an effective temperature of about $T_{\rm eff} = T/X^\infty \approx 4900$ K. Note that the glass transition temperature of SiO_2 is 1446 K

Beyond Mean-Field

Despite successes such as shown in Fig. 16, the broader applicability of the mean-field scenario of aging dynamics remains unclear, however. While some experiments and simulations indeed seem to support the existence of well-behaved effective temperatures [1,92,164], other studies also reveal the limits of the mean-field scenario. Experiments have for instance reported anomalously large FDT violations associated with intermittent

dynamics [11,12,44,45], while theoretical studies of model systems have also found non-monotonic or even negative response functions [68,118,140,162], and ill-defined or observable-dependent FDRs [73]. In principle, these discrepancies with mean-field predictions are to be expected, since there are many systems of physical interest in which the dynamics are not of mean-field type, displaying both activated processes and spatial heterogeneity.

It is thus an important task to understand from the theoretical point of view when the mean-field concept of an FDR-related effective temperature remains viable. However, studying theoretically the interplay between relevant dynamic lengthscales and thermally activated dynamics in the non-equilibrium regime of disordered materials is clearly a challenging task. Nevertheless, this problem has been approached in different ways, as we briefly summarize in this subsection.

A first class of system that displays aging and spatial heterogeneity is given by coarsening systems. The paradigmatic situation is that of an Ising ferromagnetic model (with a transition at T_c) suddenly quenched in the ferromagnetic phase at time $t_w = 0$. For $t_w > 0$ domains of positive and negative magnetizations appear and slowly coarsen with time. The appearance of domains that grow with time proves the presence of both aging and heterogeneity in this situation.

The case where the quench is performed down to $T < T_c$ is well understood. The system becomes scale invariant [41], since the only relevant lengthscale is the growing domain size, $\ell(t_w)$. Correlation functions display aging, and scale invariance implies that $C(t, t_w) \sim f(\ell(t)/\ell(t_w))$. Response functions can be decomposed into two contributions [8,17]: one part stems from the bulk of the domains and behaves as the equilibrium response, and a second one from the domain walls and becomes vanishingly small in the long time limit where $\ell(t_w) \to \infty$ and the density of domain walls vanishes. This implies that for coarsening systems in $d \geq 2$, one has $X^\infty = 0$, or equivalently an infinite effective temperature, $T_{\text{eff}} = \infty$. The case $d = 1$ is special because $T_c = 0$ and the response function remains dominated by the domain walls, which yields the non-trivial value $X^\infty = 1/2$ [89].

Another special case has retained attention. When the quench is performed at $T = T_c$, there is no more distinction between walls and domains and the above argument yielding $X^\infty = 0$ does not hold. Instead one studies the growth with time of critical fluctuations, with $\xi(t_w) \sim t_w^{1/z}$ the correlation length at time t_w, where z is the dynamic exponent. Both correlation and response functions become non-trivial at the critical point [90]. It proves useful in that case to consider the dynamics of the Fourier components of the magnetization fluctuations, $C_q(t, t_w) = \langle m_q(t) m_{-q}(t_w)\rangle$, and the conjugated response $R_q(t, t_w) = (\delta\langle m_q(t)\rangle)(\delta h_{-q}(t_w))$. From Eq. (23) a wavevector dependent FDR follows, $X_q(t, t_w)$, which has interesting properties [128] (see [47] for a review).

In dimension $d = 1$, it is possible to compute $X_q(t, t_w)$ exactly in the aging regime at $T = T_c = 0$. An interesting scaling form is found, and numerical simulations performed for $d > 1$ confirm its validity:

$$X_q(t, t_w) = \mathcal{X}(q^2 t_w) \,, \tag{27}$$

where the scaling function $\mathcal{X}(x)$ is $\mathcal{X}(x \to \infty) \to 1$ at small lengthscale, $q\xi \gg 1$, and $\mathcal{X}(x \to 0) \to 1/2$ (in $d = 1$) at large distance, $q\xi \ll 1$; recall that for $z = 2$ in that case.

Contrary to mean-field systems where geometry played no role, here the presence of a growing correlation lengthscale plays a crucial role in the off-equilibrium regime since $\xi(t_w)$ allows one to discriminate between fluctuations that satisfy the FDT at small lengthscale, $X_q \sim 1$, and those at large lenghtscale which are still far from equilibrium, $0 < X_q \sim X^\infty < 1$. These studies suggest therefore that generalized fluctuation-dissipation relation in fact have a strong lengthscale dependence—a result which is not predicted by mean-field approaches.

Another interesting result is that the FDT violation for global observables (i. e. those at $q = 0$) takes a particularly simple form, since the introduction of a single number is sufficient, the FDR at zero wavevector, $X_{q=0}(t, t_w) \equiv X^\infty = 1/2$ (in $d = 1$). This universal quantity takes non-trivial values in higher dimension, e. g. $X^\infty \approx 0.34$ is measured in $d = 2$ [128]. This shows that the study of global rather than local quantities makes the measurement of X^∞ much easier. Finally, having a non-trivial value of X^∞ for global observables suggests that the possibility to define an effective temperature remains valid, but it has become a more complicated object, related to global fluctuations on large lengthscale.

Kinetically constrained spin models represent a second class of non-mean-field systems whose off-equilibrium has been thoroughly studied recently [125]. This is quite a natural thing to do since these systems have local, finite ranged interactions, and they combine the interesting features of being defined in terms of (effective) microscopic degrees of freedom, having local dynamical rules, and displaying thermally activated and heterogeneous dynamics.

The case of the Fredrickson–Andersen model, described in Sect. "Some Theory and Models", has been studied in great detail [125], and we summarize the main results. Here, the relevant dynamic variables are the Fourier

components of the mobility field, which also correspond in that case to the fluctuations of the energy density. Surprisingly, the structure of the generalized fluctuation-dissipation relation remains once more very simple. In particular, in dimension $d > 2$, one finds a scaling form similar to (27), $X_q(t, t_w) = \mathcal{X}(q^2 t_w)$, with a well-defined limit at large distance $X_{q=0}(t, t_w) \equiv X^\infty$. The deep analogy with critical Ising models stems from the fact that mobility defects in KCMs diffuse in a way similar to domain walls in coarsening Ising models. It is in fact by exploiting this analogy that analytic results are obtained in the aging regime of the Fredrickson-Andersen model [130].

There is however a major qualitative difference between the two families of model. The (big!) surprise lies in the sign of the asymptotic FDR, since calculations show that [129]

$$X^\infty = -3, \quad d > 2 \, .$$

In dimension $d = 1$, one finds $X_{q=0}(t, t_w) = f(t/t_w)$ with $X_{q=0}(t \to \infty, t_w) = (3\pi)(16 - 6\pi) \approx -3.307$. Numerical simulations confirm these calculations. In Fig. 17, we show such a comparison between simulations (symbols) and theory (lines) in the case of the $d = 3$ Fredrickson–Andersen model [129]. Fourier components of the mobility field yield parametric FD plots that follow scaling with the variable $q^2 t_w$, as a direct result of the presence of a growing lengthscale for dynamic heterogeneity, $\xi(t_w) \sim \sqrt{t_w}$. Again, generalized fluctuation-dissipation relations explicitely depend on the spatial lengthscale considered, unlike in mean-field studies. In Fig. 17, the limit $q = 0$ corresponding to global observables is also very interesting since the plot is a pure straight line, as in equilibrium. Unlike equilibrium, however, the slope is not 1 but -3. A negative slope in this plot means a negative FDR, and therefore suggests a negative effective temperature, a very non-intuitive result at first sight.

Negative response functions in fact directly follow from the thermally activated nature of the dynamics of these models [129]. First, one should note that the global observable shown in Fig. 17 corresponds to fluctuations of the energy, $e(t_w)$, whose conjugated field is temperature. In the aging regime the system slowly drifts towards equilibrium. Microscopic moves result from thermally activated processes, corresponding to the local crossing of energy barriers. An infinitesimal change in temperature, $T \to T + \delta T$ with $\delta T > 0$, accelerates these barrier crossings and makes the relaxation dynamics faster. The energy response to a positive temperature pulse is therefore negative, $\delta e < 0$, which directly yields $\delta e / \delta T < 0$, which explains the negative sign of the FDR. This result does not hold in mean-field glasses, where thermal activation plays no role.

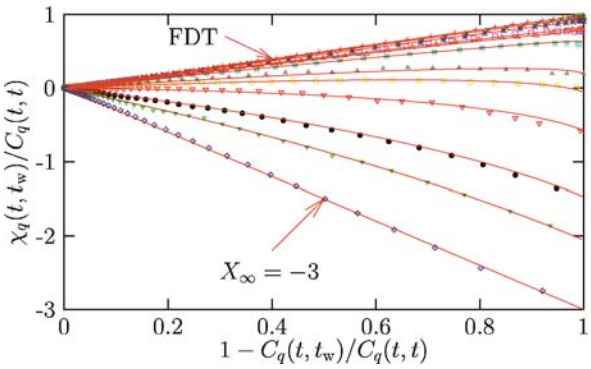

Glasses and Aging, A Statistical Mechanics Perspective on, Figure 17

Parametric response-correlation plots for the Fourier components of the mobility field in the $d = 3$ Fredrickson-Andersen model. Symbols are from simulations, lines from analytic calculations, and wavevectors decrease from top to bottom. The FDT is close to being satisfied at large q corresponding to local equilibrium. At larger distance deviations from the FDT are seen, with an asymptotic FDR which becomes negative. Finally, for energy fluctuations at $q = 0$ (bottom curve), the plot becomes a pure straight line of (negative!) slope −3, as a result of thermally activated dynamics

Finally, another scenario holds for local observables in some KCMs when kinetic constraints are stronger, such as the East model [125] or a bidimensional triangular plaquette model [100]. Here, relaxation is governed by a hierarchy of energy barriers that endow the systems with specific dynamic properties. In the aging regime following a quench, in particular, the hierarchy yields an energy relaxation that arises in discrete steps which take place on very different timescales, reminiscent of the 'time sectors' encountered in mean-field spin glasses. Surprisingly, it is found that to each of these discrete relaxations one can associate a well-defined (positive) value of the fluctuation-dissipation ratio, again reminiscent of the dynamics of mean-field spin glass models. Therefore, even in models that are very far from the mean-field limit the physical picture of a slow relaxation taking place on multiple timescales with each timescale characterized by an effective temperature seems to have some validity.

Driven Glassy Materials

We have introduced aging phenomena with the argument that in a glass phase, the timescale to equilibrate becomes so long that the system always remembers its complete history. This is true in general, but one can wonder whether it

is possible to invent a protocol where the material history could be erased, and the system 'rejuvenated' [132]. This concept has been known for decades in the field of polymer glasses, where complex thermo-mechanical histories are often used.

Let us consider an aging protocol where the system is quenched to low temperature at time $t_w = 0$, but the system is simultaneously forced by an external mechanical constraint. Experimentally one finds that a stationary state can be reached, which explicitely depends on the strength of the forcing: a system which is forced more strongly relaxes faster than a less solicited material, a phenomenon called 'shear-thinning'. The material has therefore entered a driven steady state, where memory of its age is no longer present and dynamics has become stationary: aging is stopped.

Many studies of these driven glassy states have been performed in recent years. In the language of the jamming phase diagram in Fig. 4, these correspond to studies of the (Temperature, Load) plane for molecular liquids, or the (1/Density, Load) plane for colloidal systems. The former studies are relevant for the rheology of supercooled liquids and glasses, and the $T \ll T_g$ limit corresponds to studies of the plasticity of amorphous solids, a broad field in itself. In the colloidal world, such studies are also relevant for the newly-defined field of the rheology of 'soft glassy materials'. These materials are (somewhat tautologically) defined as those for which the non-linear rheological behaviour is believed to result precisely from the competition between intrinsically slow relaxation processses and an external forcing [150]. It is believed that the rheology of dense colloidal suspensions, foams, emulsions, binary mixtures, or even biophysical systems are ruled by such a competition, quite a broad field of application indeed.

From the point of view of statmech modeling, soft glassy rheology can be naturally studied from the very same angles as the glass transition itself. As such trap models [150,151], mean-field spin glasses [18] and the related mode-coupling theory approach [83,134] have been explicitely extended to include an external mechanical forcing. In all these cases, one finds that a driven steady state can be reached and aging is indeed expected to stop at a level that depends on the strength of the forcing. Many of the results obtained in aging systems about the properties of an effective temperature are also shown to apply in the driven case, as shown both theoretically [18] and numerically [16]. A most interesting aspect is that the broad relaxation spectra predicted to occur in glassy materials close to a glass transition directly translate into 'anomalous' laws both for the linear rheological behaviour (seen experimentally in the broad spectrum of elastic, $G'(\omega)$, and loss, $G''(\omega)$, moduli), and the non-linear rheological behaviour (a strong dependence of the viscosity η upon the shear rate $\dot{\gamma}$).

Future Directions

The problem of the glass transition, already very exciting in itself, has ramifications well beyond the physics of supercooled liquids. Glassy systems figures among the even larger class of 'complex systems'. These are formed by a set of interacting degrees of freedom that show an emergent behaviour: as a whole they exhibit properties not obvious from the properties of the individual parts. As a consequence the study of glass-formers as statistical mechanics model characterized by frustrated interactions is a fertile ground to develop new concepts and techniques that will likely be applied to other physical, and more generally, scientific situations.

An example, already cited in this review, are the recent progress obtained in computer science and information theory [55] using techniques originally developed for spin glasses and structural glasses. More progress is certainly expected in the future along these interdisciplinary routes. Concerning physics, glassiness is such an ubiquitous and, yet as we showed, rather poorly understood problem that many developments are very likely to take place in the next decade.

Instead of guessing future developments of the field (and then very likely be proven wrong) we prefer to list a few problems we would like to see solved in the next years.

- Are the jamming transitions of granular media and colloids related to the glass transition of supercooled liquids? If yes, what is the common physical mechanism behind the dramatic slowing down?
- Is the glass transition related to a true phase transition? If yes, a static or a dynamic one? A finite or zero temperature one?
- Do RFOT, defects models, or frustration-based theory form the correct starting points of 'the' theory of the glass transition?
- Is MCT really a useful theory for the first decades of slowing down of the dynamics? Can one find direct evidence that an avoided MCT transition exists and controls the dynamics?
- What is the correct physical picture for the low temperature phase of glass-forming liquids and spin glasses?
- Are there general principles governing off-equilibrium equilibrium dynamics, and in particular aging and sheared materials?

- Do non-disordered, finite-dimensional, finite-range statmech model exist that display a thermodynamically stable amorphous phase at low temperature?

Finally, notice that we did not discuss possible interplays between glassiness and quantum fluctuations. This is a very fascinating topic. Quantum glassiness, and more generally, slow quantum dynamics are research subjects which are still in their infancies but that will likely undergo exciting developments in the near future.

Acknowledgments

We thank C. Marchetti for inviting us to write this review and the collaborators who worked with us on glass physics. We thank J.-P. Bouchaud, A. Lefèvre, T. Sarlat for a careful reading of our manuscript and suggestions. Our work is supported by ANR Grants CHEF, TSANET and DYNHET.

Bibliography

1. Abou B, Gallet F (2004) Phys Rev Lett 93:160603
2. Adam G, Gibbs JH (1958) J Chem Phys 43:139
3. Adhikari AN, Capurso NA, Bingemann D (2007) J Chem Phys 127:114508
4. Allen M, Tildesley D (1987) Computer Simulation of Liquids. Oxford University Press, Oxford
5. Angell CA (1997) J Res NIST 102:171
6. Angell CA (1995) Science 267:1924
7. Appignanesi GA, Rodriguez JA Fris, Montani RA, Kob W (2006) Phys Rev Lett 96:057801
8. Barrat A (1998) Phys Rev E 57:3629
9. Barrat J-L, Dalibard J, Feigelman M, Kurchan J (eds) (2003) Slow relaxations and nonequilibrium dynamics in condensed matter. Springer, Berlin
10. Bassler H (1987) Phys Rev Lett 58:767
11. Bellon L, Ciliberto S, Laroche C (2001) Europhys Lett 53:511
12. Bellon L, Ciliberto S (2002) Physica D 168:325
13. Bengtzelius U, Götze W, Sjölander A (1984) J Phys C 17:5915
14. Bennemann C, Donati C, Baschnagel J, Glotzer SC (1999) Nature 399:246; Lacevic N, Starr FW, Schroder TB, Glotzer SC (2003) J Chem Phys 119:7372
15. Bert F, Dupuis V, Vincent E, Hammann J, Bouchaud J-P (2004) Phys Rev Lett 92:167203
16. Berthier L, Barrat J-L (2002) J Chem Phys 116:6228
17. Berthier L, Barrat J-L, Kurchan J (1999) Eur Phys J B 11:635
18. Berthier L, Barrat J-L, Kurchan J (2000) Phys Rev E 61:5464
19. Berthier L, Barrat J-L (2002) Phys Rev Lett 89:095702
20. Berthier L, Biroli G, Bouchaud J-P, Cipelletti L, D El Masri, L'Hôte D, Ladieu F, Pierno M (2005) Science 310:1797
21. Berthier L, Biroli G, Bouchaud J-P, Kob W, Miyazaki K, Reichman DR (2007) J Chem Phys 126:184503
22. Berthier L, Biroli G, Bouchaud J-P, Kob W, Miyazaki K, Reichman DR (2007) J Chem Phys 126:184504
23. Berthier L, Bouchaud J-P (2002) Phys Rev B 66:054404
24. Berthier L, Chandler D, Garrahan JP (2005) Europhys Lett 69:320
25. Berthier L, Garrahan JP (2003) J Chem Phys 119:4367
26. Berthier L, Garrahan JP (2005) J Phys Chem B 109:3578
27. Berthier L, Kob W (2007) J Phys: Condens Matter 19:205130
28. Berthier L (2004) Phys Rev E 69:020201(R)
29. Berthier L (2007) Phys Rev E 76:011507
30. Berthier L (2007) Phys Rev Lett 98:220601
31. Berthier L, Viasnoff V, White O, Orlyanchik V, Krzakala F, in Reference [9]
32. Berthier L, Young AP (2005) Phys Rev B 71:214429
33. Binder K, Kob W (2005) Glassy materials and disordered solids. World Scientific, Singapore
34. Biroli G, Bouchaud J-P, Miyazaki K, Reichman DR (2006) Phys Rev Lett 97:195701
35. Biroli G, Bouchaud J-P, Tarjus G (2005) J Chem Phys 123:044510
36. Biroli G, Bouchaud JP (2004) Europhys Lett 67:21
37. Biroli G, Mézard M (2001) Phys Rev Lett 88:025501
38. Bouchaud J-P, Biroli G (2004) J Chem Phys 121:7347
39. Bouchaud J-P, Dupuis V, Hammann J, Vincent E (2001) Phys Rev B 65:024439
40. Bouchaud JP (1992) J Phys I France 2:1705
41. Bray AJ (1994) Adv Phys 43:357
42. Bray AJ, Moore MA (1984) J Phys C 17:L463; and in (1987) Heidelberg Colloquim on Glassy Dynamics In: van Hemmen JL, Morgenstern I (eds), Lectures Notes in Physics, vol 275. Springer, Berlin
43. Bray AJ, Moore MA (1987) Phys Rev Lett 58:57
44. Buisson L, Bellon L, Ciliberto S (2003) J Phys Condens Matter 15:S1163
45. Buisson L, Ciliberto S, Garcimartin A (2003) Europhys Lett 63:603
46. Butler S, Harrowell P (1991) J Chem Phys 95:4454 ((1991) J Chem Phys 95:4466)
47. Calabrese P, Gambassi A (2005) J Phys A 38:R133
48. Castellani T, Cavagna A (2005) J Stat Mech P05012
49. Cavagna A, Grigera TS, Verrocchio P (2007) Phys Rev Lett 98:187801
50. Chandler D, Garrahan JP, Jack RL, Maibaum L, Pan AC (2006) Phys Rev E 74:051501
51. Chaudhuri P, Berthier L, Kob W (2007) Phys Rev Lett 99:060604
52. Cheng Z, J, Chaikin PM, Phan S, Russel WB (2002) Phys Rev E 65:041405
53. Cohen MH, Grest GS (1982) Phys Rev B 26:6313
54. Coluzzi B, Verrocchio P (2002) J Chem Phys 116:3789
55. Mézard M, Bouchaud J-P, Dalibard J (eds) (2007) Complex systems. Springer, Berlin
56. Coslovich D, Pastore G (2007) J Chem Phys 127:124504
57. Crisanti A, Ritort F (2003) J Phys A 36:R181
58. Cugliandolo LF in Reference [9]
59. Cugliandolo LF, Kurchan J, Peliti L (1997) Phys Rev E 55:3898
60. Cugliandolo LF, Kurchan J (1993) Phys Rev Lett 71:173
61. Cugliandolo LF, Kurchan J (1994) J Phys A 27:5749
62. D'Anna G, Gremaud G (2001) Nature 413:407
63. Dalle-Ferrier C, Thibierge C, Alba-Simionesco C, Berthier L, Biroli G, Bouchaud J-P, Ladieu F, L'Hôte D, Tarjus G (2007) Phys Rev E 76:041510
64. Das SP, Mazenko GF (1986) Phys Rev A 34:2265
65. Dauchot O, Marty G, Biroli G (2005) Phys Rev Lett 95:265701
66. Debenedetti PG (1996) Metastable liquids. Princeton University Press, Princeton

67. Debenedetti PG, Stillinger FH (2001) Nature 410:259
68. Depken M, Stinchcombe R (2005) Phys Rev E 71:065102
69. Donati C, Douglas J, Kob W, Plimpton SJ, Poole PH, Glotzer SC (1998) Phys Rev Lett 80:2338
70. Downton MT, Kennett MP (2007) Phys Rev E 76:031502
71. Dzero M, Schmalian J, Wolynes PG (2005) Phys Rev B 72:100201
72. Ediger MD (2000) Annu Rev Phys Chem 51:99
73. Fielding S, Sollich P (2002) Phys Rev Lett 88:050603
74. Fisher DS, Huse DA (1986) Phys Rev Lett 56:1601
75. Franck FC (1952) Proc R Soc London 215:43
76. Franz S, Donati C, Parisi G, Glotzer SC (1999) Philos Mag B 79:1827
77. Franz S, Mulet R, Parisi G (2002) Phys Rev E 65:021506
78. Franz S, Mézard M, Parisi G, Peliti L (1998) Phys Rev Lett 81:1758
79. Franz S, Parisi G (2000) J Phys Condens Matter 12:6335
80. Franz S, Stat J (2005) Mech P04001; (2006) Europhys Lett 73:492
81. Fredrickson GH, Andersen HC (1984) Phys Rev Lett 53:1244
82. Fredrickson GH, Brawer SA (1986) J Chem Phys 84:3351
83. Fuchs M, Cates ME (2002) Phys Rev Lett 89:248304
84. Garrahan JP, Chandler D (2002) Phys Rev Lett 89:035704
85. Garrahan JP, Chandler D (2003) Proc Natl Acad Sc USA 100:9710
86. Garrahan JP (2002) J Phys Condens Matter 14:1571
87. Glarum SH (1960) J Chem Phys 33:639
88. Gleim T, Kob W, Binder K (1998) Phys Rev Lett 81:4404
89. Godrèche C, Luck J-M (2000) J Phys A 33:1151; Lippiello E, Zannetti M (2000) Phys Rev E 61:3369
90. Godrèche C, Luck J-M (2000) J Phys A 33:9141
91. Goldstein M (1969) J Chem Phys 51:3728
92. Grigera TS, Israeloff NE (1999) Phys Rev Lett 83:5038
93. Gross J, Mézard M (1984) Nucl Phys B 240:431
94. Götze W (1999) J Phys Condens Matter 11:A1
95. Hansen JP, McDonald IR (1986) Theory of Simple Liquids. Academic, London
96. Hodge I (1997) J Res NIST 102:195
97. Holyst R (2001) Physica A 292:255
98. Horbach J, Kob W (2001) Phys Rev E 64:041503
99. Hurley MM, Harrowell P (1995) Phys Rev E 52:1694
100. Jack RL, Berthier L, Garrahan JP (2006) J Stat Mech P12005
101. Jack RL, Berthier L, Garrahan JP (2005) Phys Rev E 72:016103
102. Jack RL, Garrahan JP (2005) J Chem Phys 123:164508
103. Jack RL, Mayer P, Sollich P (2006) J Stat Mech Theory Exp P03006
104. Jaeger HM, Nagel SR, Behringer RP (1996) Rev Mod Phys 68:1259
105. Johari GP (2000) J Chem Phys 112:8958
106. Jung Y, Garrahan JP, Chandler D (2004) Phys Rev E 69: 061205
107. Jönsson PE, Mathieu R, Nordblad P, Yoshino H, H Aruga Katori, Ito A (2004) Phys Rev B 70:174402
108. Kauzmann AW (1948) Chem Rev 43:219
109. Kegel WK, van Blaaderen A (2000) Science 287:290
110. Keys AS, Abate AR, Glotzer SC, Durian DJ (2007) Nat Phys 3:260
111. Kirkpatrick TR, Thirumalai D (1987) Phys Rev Lett 58:2091; Kirkpatrick TR, Wolynes PG (1987) Phys Rev A 35:3072
112. Kirkpatrick TR, Thirumalai D, Wolynes PG (1989) Phys Rev A 40:1045
113. Kisker J, Santen L, Schreckenberg M, Rieger H (1996) Phys Rev B 53:6418
114. Kivelson D, Kivelson SA, Zhao X-L, Nussinov Z, Tarjus G (1995) Physica A 219:27
115. Kob W, Andersen HC (1993) Phys Rev E 48:4364
116. Kob W, Donati C, Plimpton SJ, Poole PH, Glotzer SC (1997) Phys Rev Lett 79:2827
117. Krzakala F, Montanari A, F Ricci-Tersenghi, Semerjian G, Zdeborová L (2007) Proc Natl Acad Sci 104:10318
118. Krzakala F (2005) Phys Rev Lett 94:077204
119. Kurchan J, Laloux L (1996) J Phys A 29:1929
120. Kurchan J (2005) Nature 433:222
121. Larson RG (1999) The Structure and Rheology of Complex Fluids. Oxford University Press, New York
122. Leutheusser E (1984) Phys Rev A 29:2765
123. Lipowski A, Johnston D, Espriu D (2000) Phys Rev E 62:3404
124. Liu AJ, Nagel SR (1998) Nature 396:21
125. Léonard S, Mayer P, Sollich P, Berthier L, Garrahan JP (2007) J Stat Mech P07017
126. Mapes MK, Swallen SF, Ediger MD (2006) J Chem Phys 124:054710
127. Marty G, Dauchot O (2005) Phys Rev Lett 94:015701
128. Mayer P, Berthier L, Garrahan JP, Sollich P (2003) Phys Rev E 68:016116
129. Mayer P, Léonard S, Berthier L, Garrahan JP, Sollich P (2006) Phys Rev Lett 96:030602
130. Mayer P, Sollich P (2007) J Phys A 40:5823
131. McCullagh GD, Cellai D, Lawlor A, Dawson KA, Phys. Rev. E 71:030102 (2005)
132. McKenna GB, Kovacs AJ (1984) Polym Eng Sci 24:1131
133. Menon N, Nagel SR, Phys. Rev. Lett 74:1230 (1995); Fernandez LA, Martin V-Mayor, Verrocchio P (2006) Phys Rev E 73:020501; and [49]
134. Miyazaki K, Reichman DR (2002) Phys Rev E 66:050501(R)
135. Monasson R (1995) Phys Rev Lett 75:2847
136. Mézard M, Parisi G (1999) Phys Rev Lett 82:747
137. Mézard M, Parisi G, Virasoro M (1988) Spin Glass Theory and Beyond. World Scientific, Singapore
138. Nauroth M, Kob W (1997) Phys Rev E 55:657
139. Nelson DR (2002) Defects and Geometry in Condensed Matter Physics. Cambridge University Press, Cambridge
140. Nicodemi M (1999) Phys Rev Lett 82:3734
141. Pan AC, Garrahan JP, Chandler D (2005) Phys Rev E 72:041106
142. Pardo LC, Lunkenheimer P, Loidl A (2007) Phys Rev E 76:030502(R)
143. Parisi G, Zamponi F (2005) J Chem Phys 123:144501
144. Pusey PN and van Megen W (1986) Nature 320:340
145. Richert R, Angell CA (1998) J Chem Phys 108:9016
146. Ritort F, Sollich P (2003) Adv Phys 52:219
147. Rivoire O, Biroli G, Martin OC, Mézard M (2004) Eur Phys J B 37:55
148. Réfrégier P, Vincent E, Hammann J, Ocio M (1987) J Phys France 48:1533
149. Sethna JP, Shore JD, Huang M (1991) Phys Rev B 44:4943
150. Sollich P, Lequeux F, Hebraud P, Cates ME (1997) Phys Rev Lett 78:2020
151. Sollich P (1998) Phys Rev E 58:738
152. Struik LCE (1978) Physical aging in amorphous polymers and other materials. Elsevier, Amsterdam
153. Szamel G, Flenner E (2004) Europhys Lett 67:779
154. Tarjus G, Kivelson D (1995) J Chem Phys 103:3071

155. Tarjus G, Kivelson SA, Nussinov Z, Viot P (2005) J Phys Condens Matter 17:R1143
156. Thalmann F (2002) J Chem Phys 116:3378
157. Toninelli C, Biroli G, Fisher DS (2004) Phys Rev Lett 92:185504
158. Toninelli C, Biroli G, Fisher DS (2006) Phys Rev Lett 96:035702
159. Toninelli C, Wyart M, Berthier L, Biroli G, Bouchaud J-P (2005) Phys Rev E 71:041505
160. Tracht U, Wilhelm M, Heuer A, Feng H, K Schmidt-Rohr, Spiess HW (1998) Phys Rev Lett 81:2727; Reinsberg SA, Qiu XH, Wilhelm M, Spiess HW, Ediger MD (2001) J Chem Phys 114:7299
161. Vidal Russell E,Israeloff NE (2000) Nature 408:695
162. Viot P, Talbot J, Tarjus G (2003) Fractals 11:185
163. Vogel M, Glotzer SC (2004) Phys Rev E 70:061504
164. Wang P, Song CM, Makse HA (2006) Nat Phys 2:526
165. Weeks E, Crocker JC, Levitt AC, Schofield A, Weitz DA (2000) Science 287:627
166. Weeks ER, Crocker JC, Weitz DA (2007) J Phys Condens Matter 19:205131
167. Whitelam S, Berthier L, Garrahan JP (2005) Phys Rev E 71:026128
168. Whitelam S, Berthier L, Garrahan JP (2004) Phys Rev Lett 92:185705
169. Whitelam S, Garrahan JP (2004) J Phys Chem B 108:6611
170. Wuttke J, Petry W, Pouget S (1996) J Chem Phys 105:5177
171. Xia XY, Wolynes PG (2000) Proc Natl Acad Sci USA 97:2990
172. Yamamoto R, Onuki A (1998) Phys Rev E 58:3515
173. Young AP (ed) (1998) Spin glasses and random fields. World Scientific, Singapore

Gliders in Cellular Automata

Carter Bays
Department of Computer Science and Engineering, University of South Carolina, Columbia, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Other GL Rules in the Square Grid
Why Treat All Neighbors the Same?
Gliders in One Dimension
Two Dimensional Gliders in Non-Square Grids
Three and Four Dimensional Gliders
Future Directions
Bibliography

Glossary

Game of life A particular cellular automaton (CA) discovered by John Conway in 1968.

Neighbor A neighbor of cell x is typically a cell that is in close proximity to (frequently touching) cell x.

Oscillator A periodic shape within a specific CA rule.

Glider A translating oscillator that moves across the grid of a CA.

Generation The discrete time unit which depicts the evolution of a CA.

Rule Determines how each individual cell within a CA evolves.

Definition of the Subject

A cellular automaton is a structure comprising a grid with individual cells that can have two or more states; these cells evolve in discrete time units and according to a rule, which usually involves neighbors of each cell.

Introduction

Although cellular automata has origins dating from the 1950s, interest in that topic was given a boost during the 1980s by the research of Stephan Wolfram, which culminated in 2002 with his publication of the massive tome, "A New Kind of Science" [11]. And widespread popular interest was created when John Conway's "game of life" cellular automaton was initially revealed to the public in a 1970 Scientific American article [8]. The single feature of his game that probably caused this intensive interest was undoubtedly the discovery of "gliders" (translating oscillators). Not surprisingly, gliders are present in many other cellular automata rules; the purpose of this article is to examine some of these rules and their associated gliders.

Cellular automata (CA) can be constructed in one, two, three or more dimensions and can best be explained by giving a two dimensional example. Start with an infinite grid of squares. Each individual square has eight touching neighbors; typically these neighbors are treated the same (a *Moore neighborhood*), whether they touch a candidate square on a side or at a corner. (An exception is one dimensional CA, where position usually plays a role). We now fill in some of the squares; we shall say that these squares are alive. Discrete time units called generations evolve; at each generation we apply a rule to the current configuration in order to arrive at the configuration for the next generation; in our example we shall use the rule below.

(a) If a live cell is touching two or three live cells (called neighbors), then it remains alive next generation, otherwise it dies.
(b) If a non-living cell is touching exactly three live cells, it comes to life next generation.

Figure 1 depicts the evolution of a simple configuration of filled-in (live) cells for the above rule.

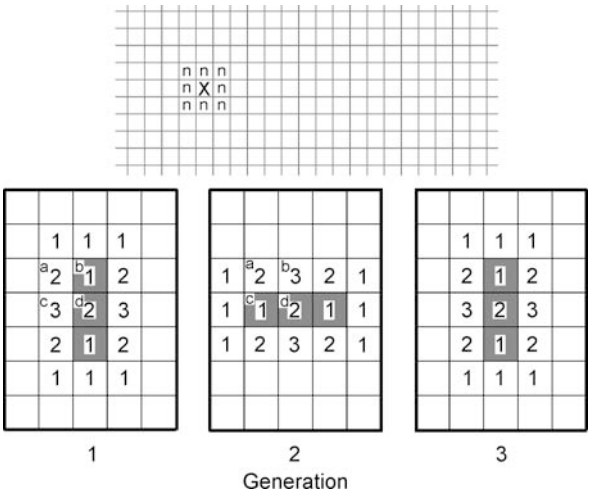

Gliders in Cellular Automata, Figure 1

***Top*: Each cell in a grid has eight neighbors. The cells containing *n* are neighbors of the cell containing the *X*. Any cell in the grid can be either *dead* or *alive*. *Bottom*: Here we have outlined a specific area of what is presumably a much larger grid. At the left we have installed an initial shape. Shaded cells are alive; all others are dead. The number within each cell gives the quantity of live neighbors for that cell. (Cells containing no numbers have zero live neighbors). Depicted are three generations, starting with the configuration at generation one. Generations two then three show the result when we apply the following cellular automata rule: *Live cells with exactly two or three live neighbors remain alive (otherwise they die); dead cells with exactly three live neighbors come to life (otherwise they remain dead)*. Let us now evaluate the transition from generation one to generation two. In our diagram, cell *a* is dead. Since it does not have exactly three live neighbors, it remains dead. Cell *b* is alive, but it needs exactly two or three live neighbors to remain alive; since it only has one, it dies. Cell *c* is dead; since it has exactly three live neighbors, it comes to life. And cell *d* has two live neighbors; hence it will remain alive. And so on. Notice that the form repeats every two generations. Such forms are called oscillators**

There are many notations for describing CA rules; these can differ depending upon the type of CA. For CA of more than one dimension, and in our present discussion, we shall utilize the following notation, which is standard for describing CA in two dimensions with Moore neighborhoods. Later we shall deal with one dimension.

We write a rule as

$$E1\,,E2\,,\ldots/F1\,,F2\,\ldots$$

where the Ei ("environment") specify the number of live neighbors required to keep a living cell alive, and the Fi ("fertility") give the number required to bring a non-living cell to life. The Ei and Fi will be listed in ascending order; hence if $i > j$ then $Ei > Ej$ etc.

Thus the rule for the CA given above is 2, 3/3. This rule, discovered by John Horton Conway, was examined in several articles in Scientific American and elsewhere, beginning with the seminal article in 1970 [8]. It is popularly known as Conway's game of life. Of course it is not really a game in the usual sense, as the outcome is determined as soon as we pick a starting configuration.

Note that the shape in Fig. 1 repeats, with a period of two. A repeating form such as this is called an oscillator. Stationary forms can be considered oscillators with a period of one. In Figs. 2 and 3 we show several oscillators that move across the grid as they change from generation to generation. Such forms are called translating oscillators, or more commonly, gliders. Conway's rule popularized the term; in fact a flurry of activity began during which a great many shapes were discovered and exploited. These shapes were named whimsically – "blinker" (Fig. 1), "boat", "beehive" and an unbelievable myriad of others. Most translating oscillators were given names other than the simple moniker glider – there were "lightweight spaceships", "puffer trains", etc. For this article, we shall call all translating oscillators gliders.

Of course rule 2, 3/3 is not the only CA rule (even though it is the most interesting). Configurations under some rules always die out, and other rules lead to explosive growth. (We say that rules with expansive growth are unstable). We can easily find gliders for many unstable rules; for example Fig. 4 illustrates some simple constructs for rule 2/2. Note that it is practically impossible NOT to create gliders with this rule! Hence we shall only look at gliders for rules that stabilize (i. e. exhibit bounded growth) and eventually yield only zero or more oscillators. We call such rules GL (game of life) rules. Stability can be a rather murky concept, since there may be some carefully constructed forms within a GL rule that grow without bounds. Typically, such forms would never appear in random configurations. Hence, we shall informally define a GL rule as follows:

- All neighbors must be touching the candidate cell and all are treated the same (a Moore neighborhood).
- there must exist at least one translating oscillator (a glider).
- Random configurations must eventually stabilize.

This definition is a bit simplistic; for a more formal definition of a GL rule refer to [5]. Conway's rule 2, 3/3 is the original GL rule and is unquestionably the most famous CA rule known. A challenge put forth by Conway was to create a configuration that would generate an ever increasing quantity of live cells. This challenge was met by William Gosper in 1970 – back when computing time was

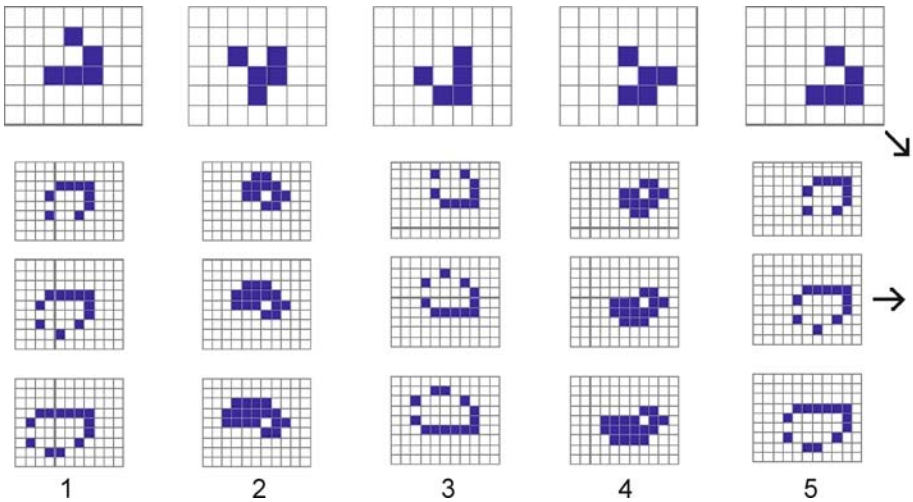

Gliders in Cellular Automata, Figure 2
Here we see a few of the small gliders that exist for 2,3/2. The form at the *top* – the original glider – was discovered by John Conway in 1968. The remaining forms were found shortly thereafter. Soon after Conway discovered rule 2,3/2 he started to give his various shapes rather whimsical names. That practice continues to this day. Hence, the name glider was given only to the simple shape at the top; the other gliders illustrated were called (from *top* to *bottom*) lightweight spaceship, middleweight spaceship and heavyweight spaceship. The numbers give the generation; each of the gliders shown has a period of four. The exact movement of each is depicted by its shifting position in the various small enclosing grids

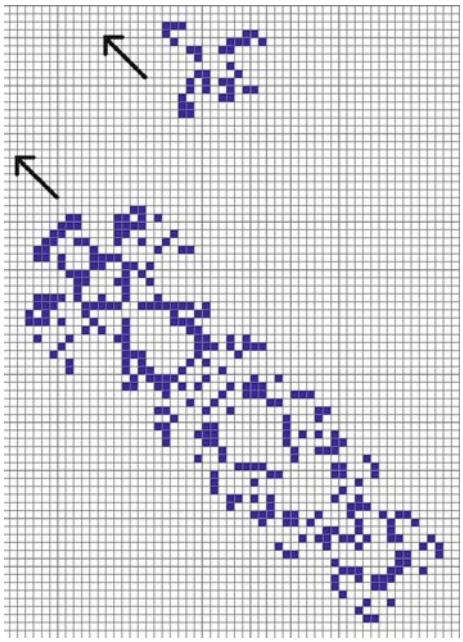

Gliders in Cellular Automata, Figure 3
The rule 2,3/2 is rich with oscillators – both stationary and translating (i. e. gliders). Here are but two of many hundreds of gliders that exist under this rule. The *top form* has a period of five and the *bottom conglomeration*, a period of four

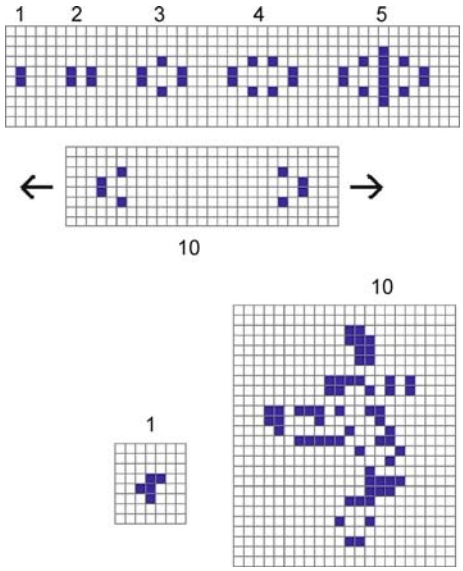

Gliders in Cellular Automata, Figure 4
Gliders exist under a large number of rules, but almost all such rules are unstable. For example the rule 2/2 exhibits rapid unbounded growth, and almost any starting configuration will yield gliders; e. g. just two live cells will produce two gliders going off in opposite directions. But almost any small form will quickly grow without bounds. The form at the *bottom left* expands to the shape at the *right* after only 10 generations. The generation is given with each form

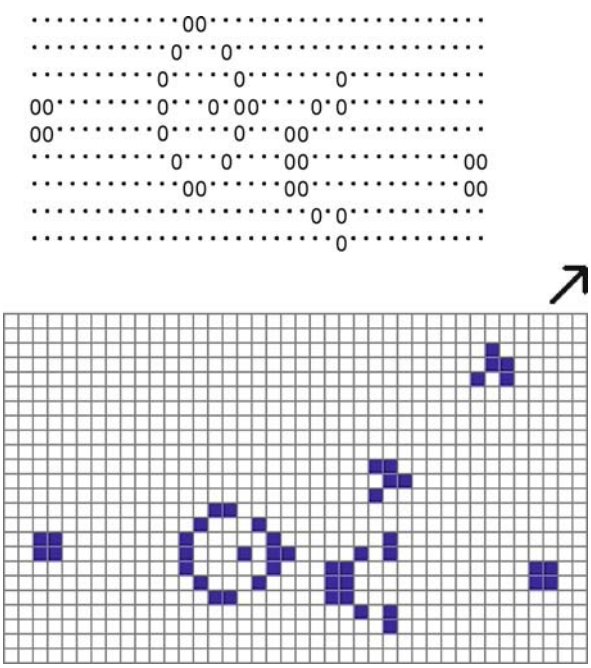

Gliders in Cellular Automata, Figure 5
A fascinating challenge was proposed by Conway in 1970 – he offered $50 to the first person who could devise a form for 2,3/2 that would generate an infinite number of living cells. One such form could be a glider gun – a construct that would create an endless stream of gliders. The challenge was soon met by William Gosper, then a student at MIT. His glider gun is illustrated here. At the *top*, testifying to the primitive computational power of the time, is an early illustration of Gosper's gun. At the *bottom* we see the gun in action, sending out a new glider every thirty generations (here it has sent out two gliders). Since 1970 there have been numerous such guns that generate all kinds of forms – some gliders and some stationary oscillators. Naturally in the latter case the generator must translate across the grid, leaving its intended stationary debris behind

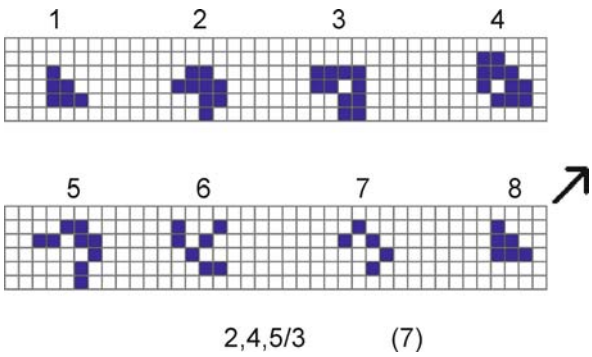

Gliders in Cellular Automata, Figure 6
There are a large number of interesting rules that can be written for the square grid and Rule 2, 3/2 is undoubtedly the most fascinating – but it is not the only GL rule. Here we depict a glider that has been found for the rule 2, 4, 5/3. And since that rule stabilizes, it is a valid GL rule. Unfortunately it is not as interesting as 2, 3/2 because its glider is not as likely to appear in random (and other) configurations – hence limiting the ability of 2, 4, 5/3 to produce interesting moving configurations. Note that the period is seven, indicated in parentheses

expensive and computers were slow by today's standards. He devised a form that spit out a continuous stream of gliders – a "glider gun", so to speak. Interestingly, his gun configuration was displayed not as nice little squares, but as a rather primitive typewritten output (Fig. 5); this emphasizes the limited resources available in 1970 for seeking out such complex structures. Soon a cottage industry developed – all kinds of intricate initial configurations were discovered and exploited; such research continues to this day.

Other GL Rules in the Square Grid

The rule 2, 4, 5/3 is also a GL rule and sports the glider shown in Fig. 6. It has not been seriously investigated and will probably not reveal the vast array of interesting forms that exist under 2, 3/3. Interestingly, 2, 3/3, 8 appears to be a GL rule which not unsurprisingly supports many of the constructs of 2, 3/3. This ability to add terms of high neighbor counts onto known GL rules, obtaining other GL rules, seems to be easy to implement – particularly in higher dimensions or in grids with large neighbor counts such as the triangular grid, which has a neighbor count of 12.

Why Treat All Neighbors the Same?

By allowing only Moore neighborhoods in two (and higher) dimensions we greatly restrict the number of rules that can be written. And certainly we could consider specialized neighborhoods – e. g. treat as neighbors only those cells that touch on sides, or touch only the left two corners and nowhere else, or touch anywhere, but state in our rule that two or more live neighbors of a subject cell must not touch each other, etc. But here we are only exploring gliders. Consider the following rule for finding the next generation.

1) A living cell dies.
2) A dead cell comes to life if and only if its left side touches a live cell.

If we start, say, with a single cell we will obtain a glider of one cell that moves to the right one cell each generation! Such rules are easy to construct, as are more complex

glider-producing positional rules. So we shall not investigate them further. Yet as we shall see, the neighbor position is an important consideration in one dimensional CA.

Gliders in One Dimension

One dimensional cellular automata differ from CA in higher dimensions in that the restrictive grid (essentially a single line of cells) limits the number of rules that can be applied. Hence, many 1D CA involve neighborhoods that extend beyond the immediate two touching neighbors of a cell whose next generation status we wish to evaluate. Or more than the two states (alive, dead) may be utilized. For our discussion about gliders, we shall only look at the simplest rules – those involving just the two adjacent neighbors and two states. Unlike 2D (and higher) dimensions, we usually consider the relative position of the neighbors when giving a rule. Since three cells (center, left, right) are involved in determining the state for the next generation of the central cell, we have $2^3 = 8$ possible initial states, with each state leading to a particular outcome. And since each initial state causes a particular outcome (i. e. the cell in the middle lives or dies next generation) we thus we have 2^8 possible rules. The behavior of these 256 rules has been extensively studied by Wolfram [11] who also introduced a very convenient shorthand that completely describes each rule (Fig. 7).

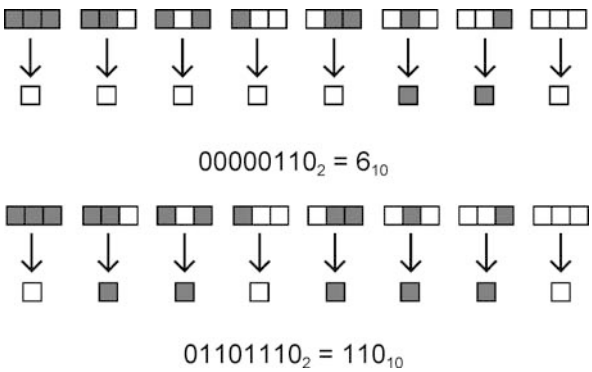

Gliders in Cellular Automata, Figure 7
The one dimensional rules six and 110 are depicted by the diagram shown. There are eight possible states involving a center cell and its two immediate neighbors. The next generation state for the center cell depends upon the current configuration; each possible current state is given. The rule is specified by the binary number depicted by the next generation state of the center cell, This notation is standard for the simplest 1D CA and was introduced by Wolfram (see [11]), who also converts the binary representation to its decimal equivalent. There are 256 possible rules, but most are not as interesting as rule 110. Rule six is one of many that generate nothing but gliders (see Fig. 8)

As we add to the complexity of defining 1D CA we greatly increase the number of possible rules. For example, just by having three states instead of two, we note that now, instead of 2^3 possible initial states, there are 3^3 (Fig. 12). This leads to 27 possible initial states, and we now can create 3^{27} unique rules – more than six trillion! Wolfram observed that even with more complex 1D rules, the fundamental behavior for all rules is typified by the simplest rules [11].

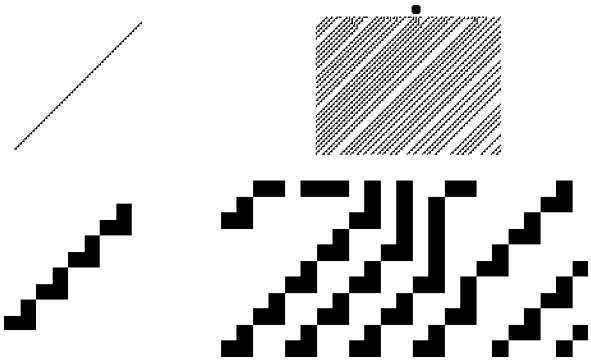

Gliders in Cellular Automata, Figure 8
Rule six (along with many others) creates nothing but gliders. At the *upper left*, we have several generations starting with a single live cell (*top*). (For 1D CA each successive generation moves vertically down one level on the page.) At the *lower left* is an enlargement of the first few generations. By following the diagram for rule six in Fig. 7, the reader can see exactly how this configuration evolves. At the *top right*, we start with a random configuration; at the *lower right* we have enlarged the small area directly under the large dot. Very quickly, all initial random configurations lead solely to gliders heading west

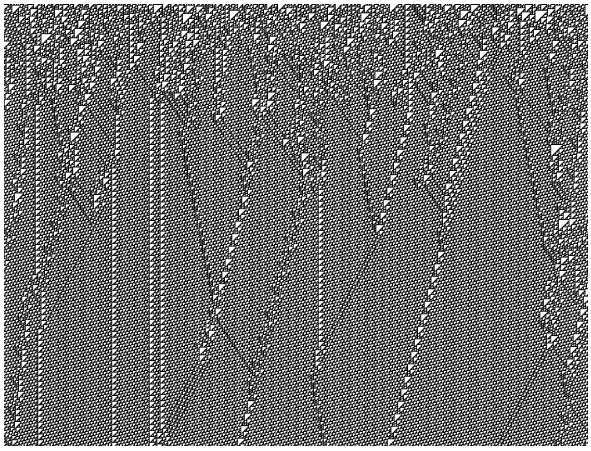

Gliders in Cellular Automata, Figure 9
Evolution of rule 110 for the first 500 generations, given a random starting configuration. With 1D CA, we can depict a great many generations on a 2D display screen

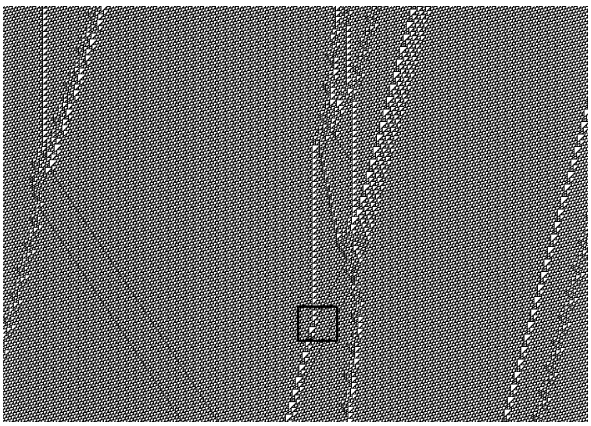

Gliders in Cellular Automata, Figure 10
Rule 110 at generations 2000–2500. The structures that move vertically are stationary oscillators; slanted structures can be considered gliders. Unlike higher dimensions, where gliders move in an unobstructed grid with no other live cells in the immediate vicinity, many 1D gliders reside in an environment of oscillating cells (the background pattern). The black square outlines an area depicted in the next figure

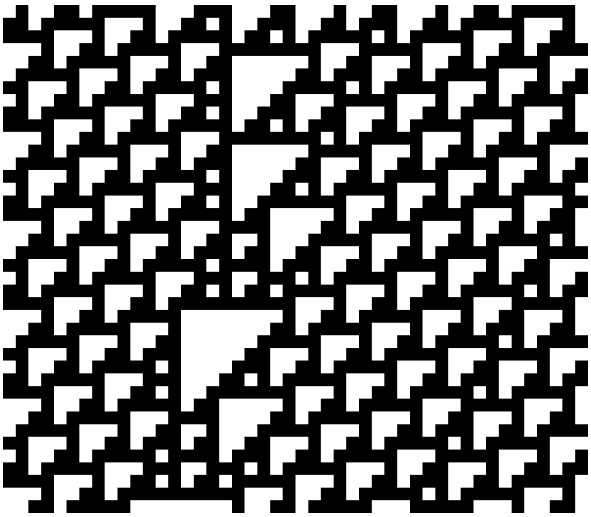

Gliders in Cellular Automata, Figure 11
An area from the previous figure enlarged. One can carefully trace the evolution from one generation to the next. The background pattern repeats every seven generations

Gliders in 1D CA are very common (Figs. 8 and 9) but true GL rules are not, because most gliders for stable rules exist against a uniform patterned background (Figs. 9 through 11) instead of a grid of non-living cells.

Two Dimensional Gliders in Non-Square Grids

Although most 2D CA research involves a square grid, the triangular tessellation has been investigated somewhat.

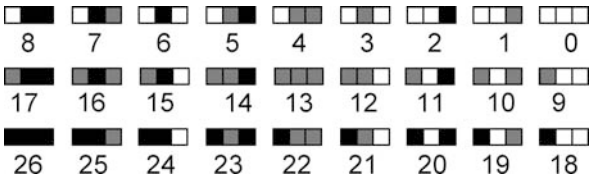

Gliders in Cellular Automata, Figure 12
There are 27 possible configurations when we have three states instead of two. Each configuration would yield some specific outcome as in Fig. 7; thus there would be three possible outcomes for each state, and hence 3^{27} distinct rules

Here we have 12 touching neighbors; as with the square grid, they are all treated equally (Fig. 13). The increased number of neighbors allows for the possibility of more GL rules (and hence several gliders). Figure 14 shows many of these gliders and their various GL rules. The GL rule 2, 7, 8/3 supports two rather unusual gliders (Figs. 15 and 16) and to date is the only known GL rule other than Conway's original 2, 3/3 game of life that exhibits glider guns. Figure 17 shows starting configurations for two of these guns and Fig. 18 exhibits evolution of the two guns after 800 generations. Due to the extremely unusual behavior of the period 80 2, 7, 8/3 glider (Fig. 16), it is highly likely that other guns exist.

The hexagonal grid supports the GL rule 3/2, along with GL rules 3, 5/2, 3, 5, 6/2 and 3, 6/2, which all behave in a manner very similar to 3/2. The glider for these three rules is shown in Fig. 19. It is possible that no other distinct hexagonal GL rules exist, because with only six touching neighbors, the set of interesting rules is quite limited. Moreover the fertility portion of the rule must start with two and rules of the form $^*/2, 3$ are unstable. Thus, any other hexagonal GL rules must be of the form $^*/2, 4$; $^*/2, 4, 5$; etc. (i. e. only seven other fertility combinations).

A valid GL rule has also been found for at least one pentagonal grid (Fig. 19). Since there are several topologically unique pentagonal tessellations (see [10]), probably other pentagonal gliders will be found, especially when all the variants of the pentagonal grid are investigated.

Three and Four Dimensional Gliders

In 1987, the first GL rules in three dimensions were discovered [1,7]. The initially found gliders and their rules are depicted in Fig. 20. It turns out that the 2D rule 2, 3/3 is in many ways contained in the 3D GL rule 5, 6, 7/6. (Note the similarity between the glider at the bottom of Fig. 20 and at the top of Fig. 2). During the ensuing years, several other 3D gliders were found (Figs. 21 and 22). Most of these gliders were unveiled by employing random but

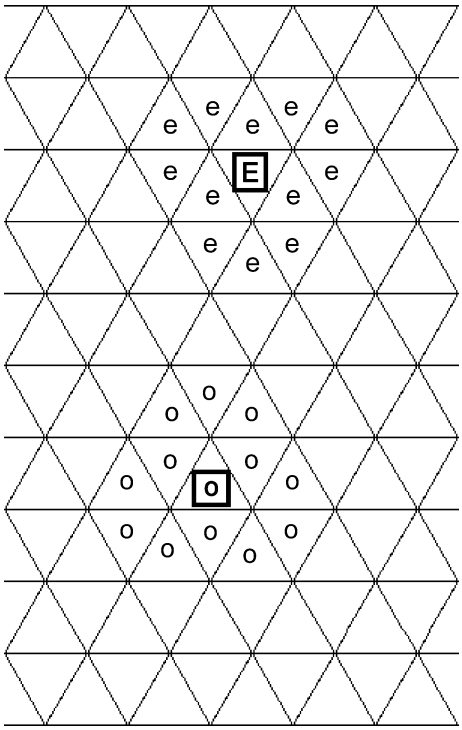

Gliders in Cellular Automata, Figure 13
Each cell in the triangular grid has 12 touching neighbors. The subject central cells can have two orientations, E and O

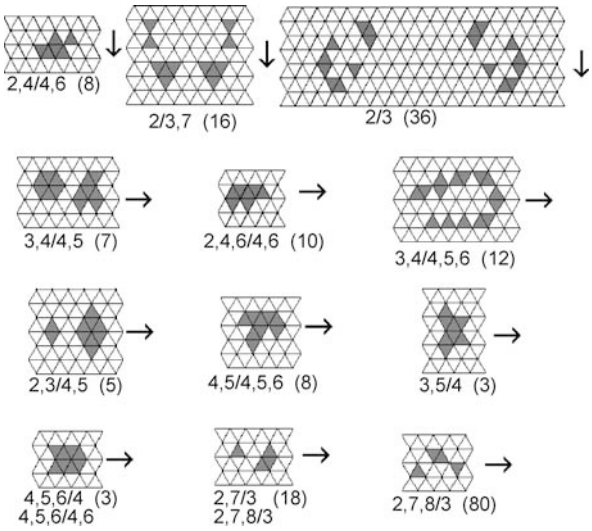

Gliders in Cellular Automata, Figure 14
Most of the known GL rules and their gliders are illustrated. The period for each is given in parentheses

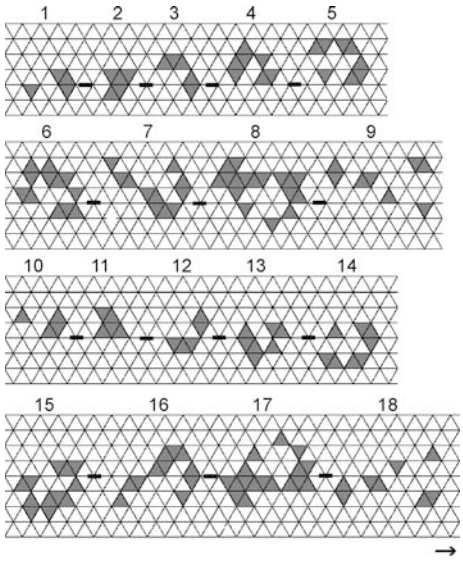

Gliders in Cellular Automata, Figure 15
The small 2, 7, 8/3 glider is shown. This glider also exists for the GL rule 2, 7/3. The small horizontal dash is for positional reference

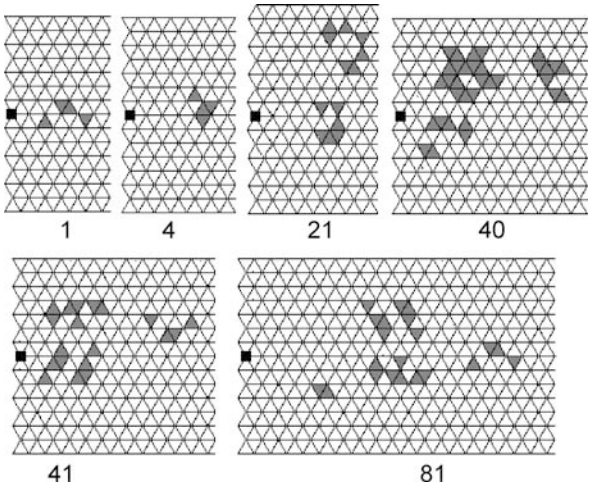

Gliders in Cellular Automata, Figure 16
Here we depict the large 2, 7, 8/3 glider. Perhaps *flamboyant* would be a better description, for this glider spews out much debris as it moves along. It has a period of 80 and its exact motion can be traced by observing its position relative to the black dot. Note that the debris tossed behind does not interfere with the 81st generation, where the entire process repeats 12 cells to the right. By carefully positioning two of these gliders, one can (without too much effort) construct a situation where the debris from both gliders interacts in a manner that produces another glider. This was the method used to discover the two guns illustrated in Figs. 17 and 18

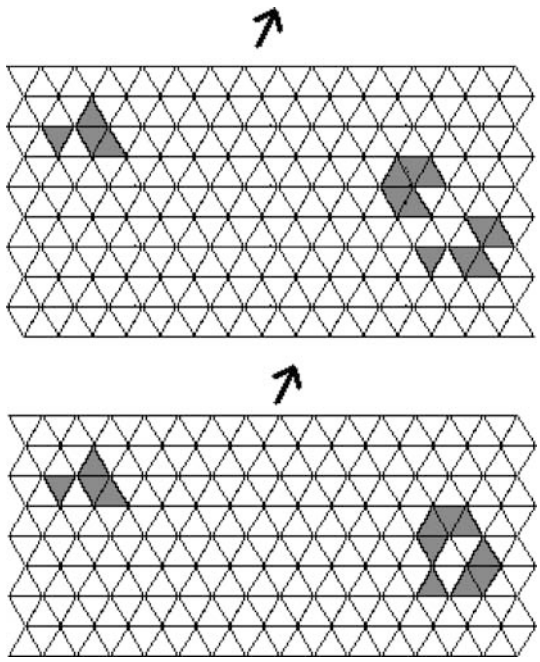

Gliders in Cellular Automata, Figure 17
The GL rule 2, 7, 8/3 is of special interest in that it is the only known GL rule besides Conway's rule that supports glider guns – configurations that spew out an endless stream of gliders. In fact, there are probably several such configurations under that rule. Here we illustrate two guns; the top one generates period 18 (small) gliders and the bottom one creates period 80 (large) gliders. Unlike Gosper's 2, 3/3 gun, these guns translate across the grid in the direction indicated. In keeping with the fanciful jargon for names, translating glider guns are also called "rakes"

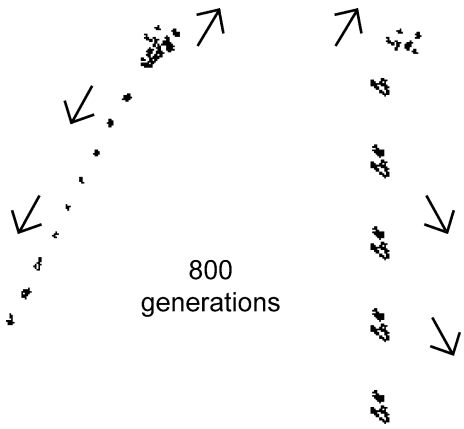

Gliders in Cellular Automata, Figure 18
After 800 generations, the two guns from Fig. 17 will have produced the output shown. Motion is in the direction given by the arrows. The gun at the left yields period 18 gliders, one every 80 generations, and the gun at the right produces a period 80 glider every 160 generations

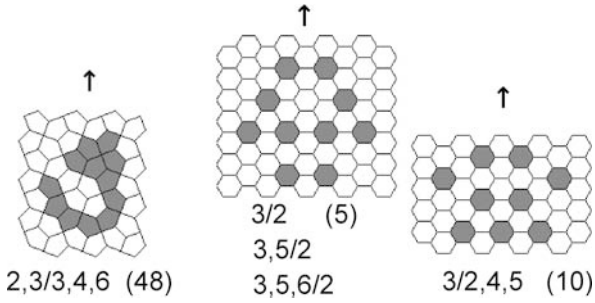

Gliders in Cellular Automata, Figure 19
GL rules are supported in pentagonal and hexagonal grids. The pentagonal grid (*left*) is called the Cairo Tiling, supposedly named after some paving tiles in that city. There are many different topologically distinct pentagonal grids; the Cairo Tiling is but one. At the right are gliders for the hexagonal rules 3/2 and 3/2, 4, 5. The 3/2 glider also works for 3, 5/2; 3, 5, 6/2 and 3, 6/2. All four of these rules are GL rules. The rule 3/2, 4, 5 is unfortunately disqualified (barely) as a GL rule because very large random blobs will grow without bounds. The periods of each glider are given in parentheses

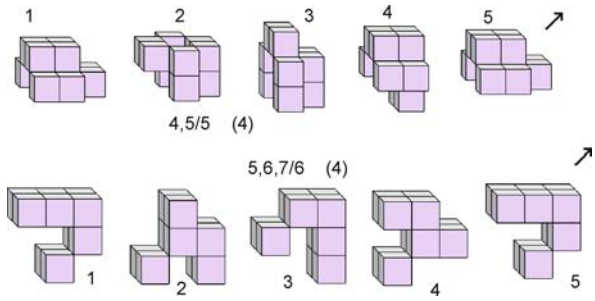

Gliders in Cellular Automata, Figure 20
The first three dimensional GL rules were found in 1987; these are the original gliders that were discovered. The rule 5, 6, 7/6 is analogous to the 2D rule 2, 3/3 (see [1]). Note the similarity between this glider and the one at the top of Fig. 2

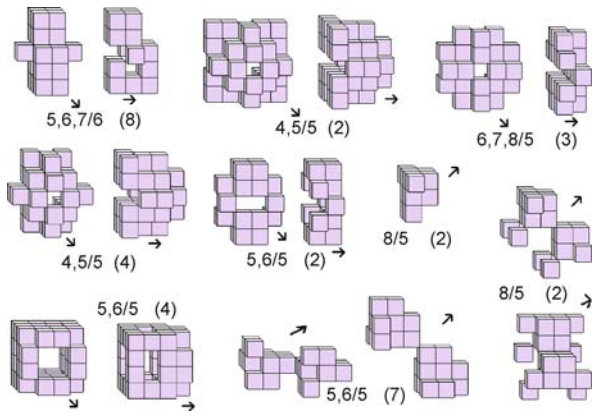

Gliders in Cellular Automata, Figure 21
Several more 3D GL rules were discovered between 1990–1994. They are illustrated here. The 8/5 gliders were originally investigated under the rule 6, 7, 8/5

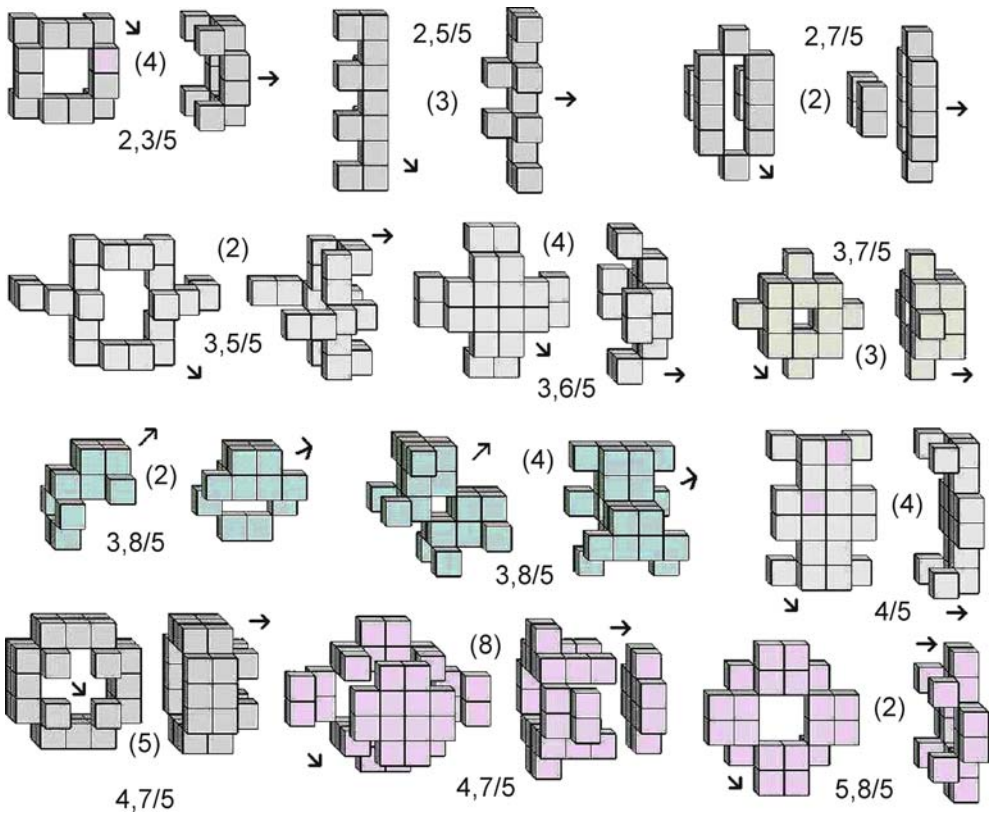

Gliders in Cellular Automata, Figure 22
By 2004, computational speed had greatly increased, so another effort was made to find 3D gliders under GL rules; these latest discoveries are illustrated here

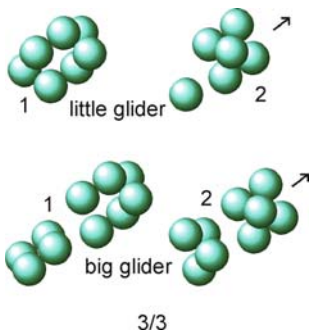

Gliders in Cellular Automata, Figure 23
Some work has been done with the 3D grid of dense packed spheres. Two gliders have been discovered for the rule 3/3, which almost qualifies as a GL rule

symmetric small initial configurations. The large number of live cells in these 3D gliders implies that they are uncommon random occurrences in their respective GL rules; hence it is highly improbable that the plethora of interesting forms (e. g. glider guns) such as those for 2D rule 2, 3/3 exist in three dimensions.

The 3D grid of dense packed spheres has also been investigated somewhat; here each sphere touches exactly 12 neighbors. What is pleasing about this configuration is that each neighbor is identical in the manner that it touches the subject cell, unlike the square and cubic grids, where some neighbors touch on their sides and others at their corners. The gliders for spherical rule 3/3 are shown in Fig. 23. This rule is a borderline GL rule, as random finite configurations appear to stabilize, but infinite ones apparently do not.

Future Directions

Gliders are an important by-product of many cellular automata rules. They have made possible the construction of extremely complicated forms – most notably within the universe of Conway's rule, 2, 3/3. (Figs. 24 and 25 illustrate a remarkable example of this complexity). Needless to say many questions remain unanswered. Can a glider gun be

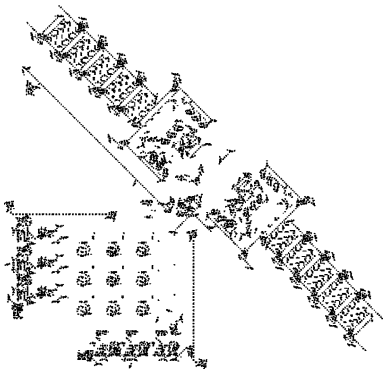

Gliders in Cellular Automata, Figure 24
The discovery of the glider in 2, 3/3, along with the development of several glider guns, has made possible the construction of many extremely complex forms. Here we see a Turing machine, developed in 2001 by Paul Rendell. Figure 25 enlarges a small portion of this structure

Gliders in Cellular Automata, Figure 25
We have enlarged a tiny portion at the upper left of the Turing machine shown in Fig. 24. One can see the complex interplay of gliders, glider guns, and various other stabilizing forms

constructed for some three dimensional rule? This would most likely be rule 5, 6, 7/6, which is the three dimensional analog of 2, 3/3 [7], but so far no example has been found.

The area of cellular automata research is more-or-less in its infancy – especially when we look beyond the square grid. Even higher dimensions have been given a glance; Fig. 26 shows just one of several gliders that are known to exist in four dimensions. Since each cell has 80 touching neighbors, it will come as no surprise that there are a large number of 4D GL rules. But there remains much work do be done in lower dimensions as well. Consider simple one dimensional cellular automata with four possible states. It will be a long time before all 10^{38} possible rules have been investigated!

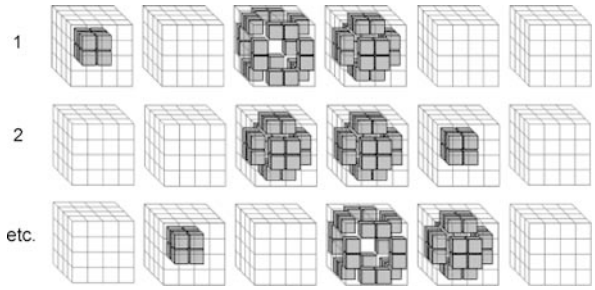

Gliders in Cellular Automata, Figure 26
Some work (not much) has been done in four dimensions. Here is an example of a glider for the GL rule 11, 12/12, 13. Many more 4D gliders exist

Bibliography

1. Bays C (1987) Candidates for the Game of Life in Three Dimensions. Complex Syst 1:373–400
2. Bays C (1987) Patterns for Simple Cellular Automata in a Universe of Dense Packed Spheres. Complex Syst 1:853–875
3. Bays C (1994) Cellular Automata in the Triangular Tessellation. Complex Syst 8:127–150
4. Bays C (1994) Further Notes on the Game of Three Dimensional Life. Complex Syst 8:67–73
5. Bays C (2005) A Note on the Game of Life in Hexagonal and Pentagonal Tessellations. Complex Syst 15:245–252
6. Bays C (2007) The Discovery of Glider Guns in a Game of Life for the Triangular Tessellation. J Cell Autom 2(4):345–350
7. Dewdney AK (1987) The game Life acquires some successors in three dimensions. Sci Am 286:16–22
8. Gardner M (1970) The fantastic combinations of John Conway's new solitaire game 'Life'. Sci Am 223:120–123
9. Preston K Jr, Duff MJB (1984) Modern Cellular Automata. Plenum Press, New York
10. Sugimoto T, Ogawa T (2000) Tiling problem of convex pentagon[s]. Forma 15:75–79
11. Wolfram S (2002) A New Kind of Science. Wolfram Media, Champaign Il

GPS: Applications in Crustal Deformation Monitoring

Jessica Murray-Moraleda
US Geological Survey, Menlo Park, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Global Positioning System Measurements
Applications of GPS Data to the Study of Seismic and Volcanic Hazards
Future Directions

Acknowledgments
Bibliography

Glossary

Confidence ellipse As applied to a vector representing a displacement or velocity estimate, the confidence ellipse defines the region within which the value is estimated at or above a specified confidence level (e. g., 95%). Confidence ellipses are computed by propagation of errors when computing the position. The ellipse is usually plotted at the tip of a GPS vector (e. g., Fig. 9).

GNSS Global Navigation Satellite System. Although this acronym stands for the same phrase as GLONASS, GNSS is a generic term referring to space-based navigation systems like the Global Positioning System (GPS) operated by the U. S., Russia's GLONASS, and the Galileo system under development by the European Union.

Kinematic GPS A method of collecting GPS data in which the receiver is continuously or intermittently in motion. This receiver, called the rover, can receive corrections for ambiguity resolution and common errors from a nearby stationary receiver.

Interferometric synthetic aperture radar (InSAR) A satellite-based imaging technique in which the satellite emits a radar signal and measures the phase of the returning signal after it has been scattered off the surface of the Earth. The difference in phase of the scattered waves measured during two passes of the same satellite can be used to produce a map of deformation, called an interferogram, that occurred during the time between the two satellite passes.

International GNSS service (IGS) An international consortium of agencies worldwide that provide data from permanent GPS and GLONASS sites in order to generate precise orbital and satellite clock parameters.

Ionosphere The electrically charged portion of the atmosphere from ~60 km to ~400 km above sea level. The ionosphere is dispersive, meaning that the degree to which it delays signal propagation depends on the signal's frequency and the electron content of the ionosphere.

Mega-thrust earthquake A type of earthquake which causes rupture of a long portion of the interface between a subducting plate and the over-riding plate. These earthquakes involve slip on a huge surface area, making them among the largest on Earth.

Moment magnitude (M_w) A magnitude scale used to compare the energy released in earthquakes. The moment magnitude is computed from the seismic moment. Therefore, because M_w accounts for the full rupture length of the earthquake, the moment magnitude scale does not saturate for large events in the way that other magnitude scales do.

Reference frame A terrestrial reference frame is defined by a set of points on Earth whose coordinates are precisely determined in a coordinate system with a specified origin and orientation of the axes. In order to compare GPS site positions, displacements, or velocities they must all be transformed into the same reference frame. For GPS, the most commonly used reference frame is the International Terrestrial Reference Frame (ITRF) which is updated periodically.

Rupture The slip that occurs during an earthquake. This term is often used in discussing the way in which the slip progresses with time over the fault surface, as in "the rupture front propagated southeast."

Satellite laser ranging (SLR) A geodetic technique for measuring the position of points on the surface of the Earth. Observation stations emit pulses of light that bounce off retroreflectors on satellites and return to the stations. The stations record the travel time of the light which is used to calculate a range measurement.

Slip The distance that material on one side of a fault moves relative to that on the other side.

Stable North America The stable interior portion of the North American continent that is not affected by plate boundary deformation. Often this term is used in the context of a "stable North American" reference frame, meaning that GPS velocities are transformed so that the velocities at stations considered to be in the stable interior of the continent are essentially zero. Because of factors such as Glacial Isostatic Adjustment (GIA), even some GPS sites in the continental interior have nonzero velocities. These sites are typically omitted when defining a stable North American reference frame.

Strainmeter An instrument that is capable of measuring change in distance over short baselines. These instruments typically come in two forms. The first is installed at the Earth's surface and uses a laser interferometer to measure the changes in distance over baseline lengths of 100s of meters. The second type is installed in a borehole 100s of meters deep to measure subtle changes in the diameter of the borehole. Some borehole strainmeters measure volumetric strain (e. g. the Sacks–Evertson strainmeter) and others measure three independent components of horizontal strain (e. g. the Gladwin tensor strainmeter).

Strong motion seismograph Seismic instrument de-

signed to record high-amplitude shaking near an earthquake rupture. These instruments typically record acceleration, and are sometimes called accelerometers. Data recording is often triggered by the arrival of the first seismic waves, and these instruments can record acceleration several times that of gravity.

Teleseismic Refers to seismic waves recorded at distances greater than 3000 km from the epicenter.

Troposphere The portion of the atmosphere from the Earth's surface to ∼15 km which delays GPS signal propagation. The degree to which the GPS signal is delayed depends on the spatially and temporally varying atmospheric pressure and water vapor content.

Very long baseline interferometry (VLBI) A geodetic positioning technique in which radio signals from distant sources such as quasars received at an array of antennas are used to calculate precise positions.

Definition of the Subject

The Global Positioning System (GPS) is a space-based Global Navigation Satellite System (GNSS). Using signals transmitted by GPS satellites, the positions of ground-based receivers can be calculated to high precision, making it possible to track the movement of points on the Earth's surface over time. Unlike older geodetic surveying methods which involved periodically measuring angles, distances, or elevations between points, GPS can provide three-component (latitude, longitude, and altitude) position information at a range of sampling rates and on a global scale. GPS equipment is easy to use and can be set up to collect data continuously. Since its early geophysical applications in the mid-1980s, this versatile tool, which can be used to track displacements over time periods of seconds to decades, has become indispensable for crustal deformation studies, leading to many important insights and some surprising discoveries.

Introduction

This article focuses on applications of GPS data to the study of tectonic, seismic, and volcanic processes. GPS has become a valuable tool for investigating other types of crustal deformation as well, including landslides (e. g., [25,50,100,128,151]), global sea-level change [150], and the ongoing rebound (termed Glacial Isostatic Adjustment or GIA) of the Earth's crust since the retreat of the ice sheets which covered much of North America and northern Europe during the last ice age (e. g., [19,80,88,95,117,141]), but these topics are beyond the scope of this article. The discussion presented here begins with an overview of how GPS works and how it is used to collect data for geophysical studies. The rest of the paper describes a variety of ways in which GPS data have been used to measure crustal deformation and investigate the underlying processes, as illustrated by examples from the literature. Since GPS is so widely used in geophysical studies, examples of many more applications exist, and the reader is encouraged to explore the literature for more information.

Global Positioning System Measurements

How GPS works

The US Department of Defense developed GPS to provide positioning and timing information, primarily for military purposes, that would be available any time of day, anywhere on Earth, regardless of weather conditions. The first GPS satellites were launched in 1978. Soon afterward the Soviet Union developed a similar system, called GLONASS (which, like the generic acronym GNSS, also stands for Global Navigation Satellite System), and more recently the European Space Agency has designed a satellite navigation system called Galileo which, unlike its predecessors, is dedicated to civilian and commercial, rather than military, use. The rest of this article will focus on GPS. The scope of this article permits only a brief overview. Dzurisin [40] gives a broader discussion with a focus on applications in volcanic investigations. Hofmann-Wellenhof et al. [56] give a thorough treatment of the technical details.

The GPS satellite constellation nominally consists of 24 satellites, as well as several spares. The satellites orbit 20,200 km above the Earth with orbital periods of nearly 12 hours, and each passes over a given point on the Earth's surface once per sidereal day (which is about four minutes shorter than a solar day). From any point on the Earth's surface, at any given time, from four to ten satellites are above the horizon (and thus potentially visible). Each satellite remains visible for approximately five out of every 12 hours [40].

The idea behind satellite positioning is that one can determine the distance between a receiver on the ground and an orbiting satellite from the time it takes a signal to travel from the satellite to the receiver. This calculation, therefore, requires a means for precise time-keeping according to a universally accepted standard. The US Naval Observatory defines "GPS time," and GPS specifications require GPS time to be within one microsecond of Coordinated Universal Time (UTC). The difference between a satellite's or receiver's internal clock and GPS time (due to clock drift) is termed "clock bias" and is accounted for in processing GPS data.

GPS satellites broadcast signals on two carrier frequencies termed L1 and L2 in the microwave band. The "coarse acquisition" (C/A) code is modulated on the L1 carrier, and the precise (P) code is modulated on both L1 and L2. A navigation message, containing information about the satellite orbits, clocks (the time given by the satellite's clock and information about the difference between that satellite's time and GPS time), and state of health, as well as ionospheric conditions, is modulated on both carriers.

The receiver "locks on" to a satellite by generating a replica of one or more of the codes modulated on the satellite signal it receives and continually cross-correlating the internally generated code with that received from the satellite until the two match. Once it has locked on to the satellite it can obtain the navigation message, determine the signal travel time, and measure the carrier signal. The apparent distance, or "pseudorange," between the GPS antenna and the satellite antenna is then calculated by multiplying the time it takes for the signal transmitted by the satellite to travel to the receiver by the speed of light. The term pseudorange emphasizes that the travel time used in this calculation includes the effects of satellite and receiver clock biases as well as a variety of other error sources, some of which may be mitigated during processing, and is therefore not equivalent to the true geometric range between the satellite and receiver. It is possible for the receiver to measure the code to a precision of about 1% of its length (293 meters for C/A-code and 29.3 meters for P-code), which results in 3 meter and 30 cm precision in calculated pseudorange. However, the P-code is generally encrypted by the military, called "anti-spoofing" or A-S, and therefore a civilian receiver cannot use this code for positioning.

The C/A code is modulated on L1, and thus this carrier can be easily measured once the receiver has locked on to the satellite using the C/A code. Because of A-S, in order to obtain the L2 carrier on which the encrypted P-code is modulated, civilian users must have receivers that apply more sophisticated signal processing techniques (see pp. 81–85 in [56]). The L1 carrier has a wavelength of 19 cm and the L2, 24.4 cm. Since the receiver can measure the carrier signal to 1% of a cycle length, much greater resolution can be achieved using the carrier phase measurements, rather than the code information, to calculate the pseudorange.

When the carrier signal is used, the satellite-receiver distance is calculated by multiplying the number of carrier cycles (which is generally not an integer) between the satellite and the receiver by the wavelength of the carrier signal. However, when a receiver locks on to a satellite, it only can measure the initial fraction of a carrier cycle that it receives. Although it measures the number of full cycles thereafter (which changes as the satellite moves overhead), the receiver has no way of knowing how many full cycles in addition to the initial fraction were between it and the satellite to begin with. This unknown number of cycles is often called the integer ambiguity and will be different for each satellite–receiver pair. In order to take advantage of the more precise positioning that can be achieved using the carrier signal, processing techniques have been developed to address the problem of integer ambiguities (see [40] for an overview and [56] for more detail).

The positions of the satellites at any given time are given by their orbital parameters. This information is transmitted as part of the navigation message, however more precise orbital information, available from the International GNSS Service (IGS), is used in scientific applications. With the satellite positions assumed known, once the distance from the receiver to at least four satellites is measured, the position in three coordinate dimensions (e. g., north, east, and vertical) of a GPS antenna on the ground can be found. Four satellites are necessary in order to solve for the three coordinate positions and the satellite and receiver clock bias. However, positioning accuracy is greatly improved with data from additional satellites as it is then possible to estimate some unknown noise sources.

Although GPS receivers are capable of determining a position in real time using internal software, for scientific applications the data are generally downloaded, and one of several processing software packages is used to obtain much more precise positions. This type of software allows the user more control over the way in which the data are processed, for instance by enabling the use of precise orbital parameters, the fixing of ambiguities, and the application of sophisticated models for atmospheric delay and variations in the antenna phase center (the part of the antenna that actually receives the GPS signal).

It is possible to reduce or eliminate certain error sources by differencing data. These errors include atmospheric delays to signal propagation that affect neighboring GPS stations (e. g. within 10 s of km of each other) in a similar way, satellite orbital and clock errors that will be common to data from the same satellite recorded by more than one station at the same time, and receiver clock errors that are common to measurements to multiple satellites made by that receiver at the same time.

Other means of addressing error sources exist as well. For example, the ionosphere is dispersive, meaning the delay in signal propagation that it causes depends on the frequency of the signal. Because GPS uses two frequen-

cies, the effect of this delay can be eliminated from the data. Tropospheric delay is addressed during data processing through a combination of models for the "dry" component (which is dependent on atmospheric pressure, temperature, and elevation), and treatment of the "wet" component (which shows large variability depending on water vapor content) as a stochastic parameter to be estimated. Multipath (when the GPS signal bounces off something before reaching the ground antenna) and set-up error (commonly, operator error in measuring the height of the antenna) are mitigated by antenna design, choice of station location, and surveying technique.

The military has used two means of limiting civilian access to GPS positioning capabilities. The first, called selective availability (SA), involved degrading the accuracy of the satellite orbit information broadcast by the satellites and introducing noise into the satellite clock information. This caused an approximately ten-fold increase in positioning error. Using information from a global network of continuously operating GPS stations it is possible to calculate more accurate satellite orbits for use in post-processing of GPS data. During processing, clock errors can be eliminated by differencing data from multiple stations or estimated along with station positions. Therefore, SA did not pose an insurmountable obstacle to scientific users of GPS. In May 2000 the US government discontinued SA, although it reserves the right to reinstate it if deemed necessary. The other means by which the military can reduce civilian access to GPS data is through A-S. A-S is implemented by encrypting the P-code such that only military users can decipher it, thus preventing outside parties from sending phony "GPS" signals that would prevent accurate positioning. A-S began in 1994 and continues to the present, but its effect on the accuracy of GPS positioning for scientific purposes is generally small.

Methods of GPS Data Collection for Crustal Deformation Studies

For some geological and geophysical applications, typically those which require mapping or recording of sample locations, kinematic GPS methods are used. In these cases, a base station is used to generate corrections that are applied to measurements made by a "roving" antenna which is moved to each site of interest [40]. However, for the crustal deformation applications described in this article, the GPS antenna remains at one location for an extended period of time (usually at least eight hours). Measurements of this type are often classified as "campaign" or "continuous," depending on the frequency of measurement and the way in which the receiver and antenna are installed. Here I give a brief description of each; Blewitt et al. [13] give a detailed discussion as well.

Before the advent of GPS as a tool for monitoring crustal deformation, other types of geodetic observations were made by periodically measuring triangulation, trilateration and leveling networks using traditional land surveying techniques. The objective was to see how the movement of the Earth's surface changed over time, since this reflects a variety of geophysical processes. However, these techniques did not produce 3-component position data, but rather measurements of angles, distances, or elevation changes.

The surveying approach in the early years of GPS monitoring (e. g. mid-1980s to mid-1990s) was similar to that of its predecessors in that instruments were deployed temporarily, for several hours at a time, once a year or so. This approach is called "campaign" or "survey-mode" GPS (SGPS). At that time receivers were prohibitively expensive, limiting their wide-spread use. Also, there were fewer satellites in orbit, and one had to schedule surveying to coincide with the time of day for which there would be satellite coverage in the area of interest.

When collecting SGPS measurements one sets up a tripod which will hold the antenna over a marker, typically a benchmark set in the ground (Fig. 1a,b). The use of such a marker enables one to find the correct spot for future measurements months or years later. The antenna is centered using an optical plummet over a point imprinted on the marker to indicate its center (Fig. 1b). One must then measure the distance between the ground (where the benchmark is) and the antenna (where the GPS signal is actually received), and this distance will be used to convert the positions obtained from data processing to the positions of the benchmarks. As you might guess, there is considerable room for error when having to set up the antenna anew every time observations are made. Height measurement errors are particularly common. The repeated measurement of the same benchmark over time produces a time series like that in Fig. 2a.

As GPS receivers became less expensive, the possibility of permanently installing continuously recording GPS (CGPS) receivers became a real option. Some of the earliest continuous GPS stations for geophysical studies were installed in Japan [144]. The first continuous GPS sites in southern California were installed in 1990 at Pasadena, Pinyon Flat, and La Jolla [77]. Today several countries have extensive CGPS networks for geophysical monitoring purposes, most notably GEONET which consists of over 1000 stations across Japan. Regional CGPS networks have existed in the US for many years covering southern California (Southern California Integrated GPS Network,

a b c

GPS: Applications in Crustal Deformation Monitoring, Figure 1
a Campaign GPS station located in western Nevada. In the foreground the tripod holds the antenna. The receiver will be stored in one of the boxes in the background while data are collected. Rocks are stacked on the feet of the tripod to prevent it from moving during data collection. (USGS photo) b Benchmark that is being observed in a. A benchmark often has a stamping telling which agency installed it and when. The cross in the center is the point over which the tripod is centered every time that benchmark is observed. (USGS photo) c Continuous GPS station located in Fremont, California. This station, P222, is part of the Plate Boundary Observatory. The antenna, protected from the elements by a domed cover, is supported by a monument with legs driven several meters into the ground. The antenna sends data to the receiver, located along with batteries in the box in the background, via a buried cable. Also visible in the background is the solar panel which provides supplemental power to the site. (UNAVCO photo, reprinted with permission)

or SCIGN [60]), the San Francisco Bay Area (Bay Area Regional Deformation, or BARD [72]), the Pacific Northwest (Pacific Northwest Geodetic Array, or PANGA [92]), and the Basin and Range (Basin and Range Geodetic Network, or BARGEN [10]; and Eastern Basin-Range and Yellowstone hotspot, or EBRY, http://www.mines.utah.edu/~rbsmith/RESEARCH/UUGPS.html). A much larger network consisting of ~850 sites is now underway. This new network, called the Plate Boundary Observatory (PBO), will provide CGPS coverage for seismically and volcanically active areas throughout the western continental US and Alaska [147].

In addition to providing daily positions, CGPS sites have the advantage of permanent monumentation, thus eliminating set-up error (Fig. 1c). Having daily positions (e. g., Fig. 2b) enables much better estimates of site velocities, but the more frequent data and added precision means it has become necessary to address additional error sources that were hidden in the noise of the less frequent measurements. These include seasonal signals and time-correlated noise likely due to monument instability, atmospheric effects, reference frame errors, or mismodeled orbits or antenna phase centers [15,33,74,86,166,170]. Despite the proliferation of CGPS stations in recent years, de-

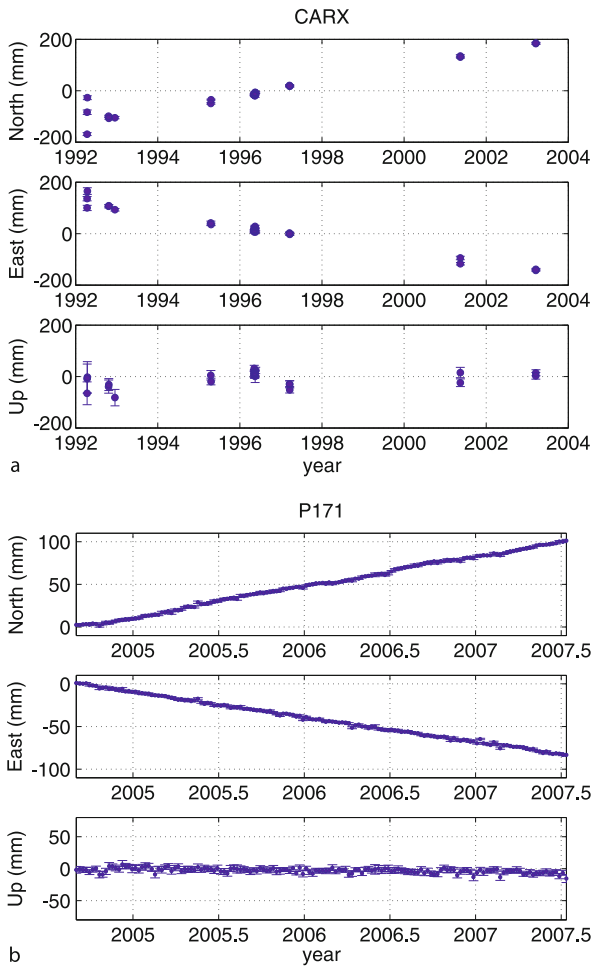

GPS: Applications in Crustal Deformation Monitoring, Figure 2
Time series of changes in station positions. Both these sites are located in central California, west of the San Andreas fault. The position changes are plotted relative to the center of the North American continent. Because of the relative motion between the North American and Pacific tectonic plates, these sites are moving to the northwest over time, relative to the stable interior of the continent which is not affected by strain accumulation due to the interaction of the two plates. a Campaign GPS site CARX near Parkfield, California. b Continuous GPS station P171 of the Plate Boundary Observatory network. This station is located just south of Monterey Bay

pending on factors such as site access and site condition, financial resources, and scientific goals, campaign-style GPS measurements are still frequently collected.

Relative Precision

Daily repeatability (or scatter) in the time series for CGPS sites is typically 0.5 to 1.5 mm in the horizontal and ~3.5 to 6.5 mm in the vertical. If noise that is common to all stations in the network (due, for example, to errors in satellite orbits or atmospheric delay) has been eliminated, short-term repeatabilities of ~0.5 mm in the horizontal and ~2 mm in the vertical can be achieved. With 1.5 to 2.5 years of CGPS data, velocity estimates with ~1 mm/yr uncertainties are possible [13]. In contrast, SGPS position repeatability is ~3 to 5 mm in the horizontal and ~10 mm in the vertical [40], and ~10 years of SGPS data would be required to reach velocity uncertainty of ~1 mm/yr [166].

One way to improve the accuracy of GPS measurements is to extend the length of an observation session, for example from six hours to 24 hours, when estimating daily positions. The additional data mean that multipath noise (which varies throughout the day) averages out better, and it is easier to fix ambiguities and estimate tropospheric delay parameters. The accuracy of GPS measurements has also improved over time. A major reason for this has been the expansion of the global network of tracking stations that are used to calculate precise orbital information for the satellites and to define reference frames for GPS positions. Models for various noise sources have also been refined over time, and some noise sources have been mitigated, for instance by improving the stability of geodetic monuments. It should be noted that the vertical signal is generally much noisier than the horizontal because it is not possible to track satellites below the horizon, and therefore there is no position control from below.

Applications of GPS Data to the Study of Seismic and Volcanic Hazards

GPS data provide important constraints on the underlying processes that lead to observed deformation, especially when used in combination with other data types. One of the major strengths of GPS is its ability to track positions over time spans of seconds to decades; another is that it enables measurement of position changes over continental-scale baselines. This section will discuss several broad areas of study using GPS data and provide specific examples of each. Background on modeling methods will be included as necessary.

Plate Motions

Earthquakes occur in response to stresses in the Earth's crust, and these stresses are largely due to the motion of the Earth's tectonic plates. Knowing the velocity at which the plates move gives us insight into the amount of deformation that must be accommodated on plate boundary faults, and what type of earthquakes might be expected. The Earth's plates are generally thought to be rigid, at

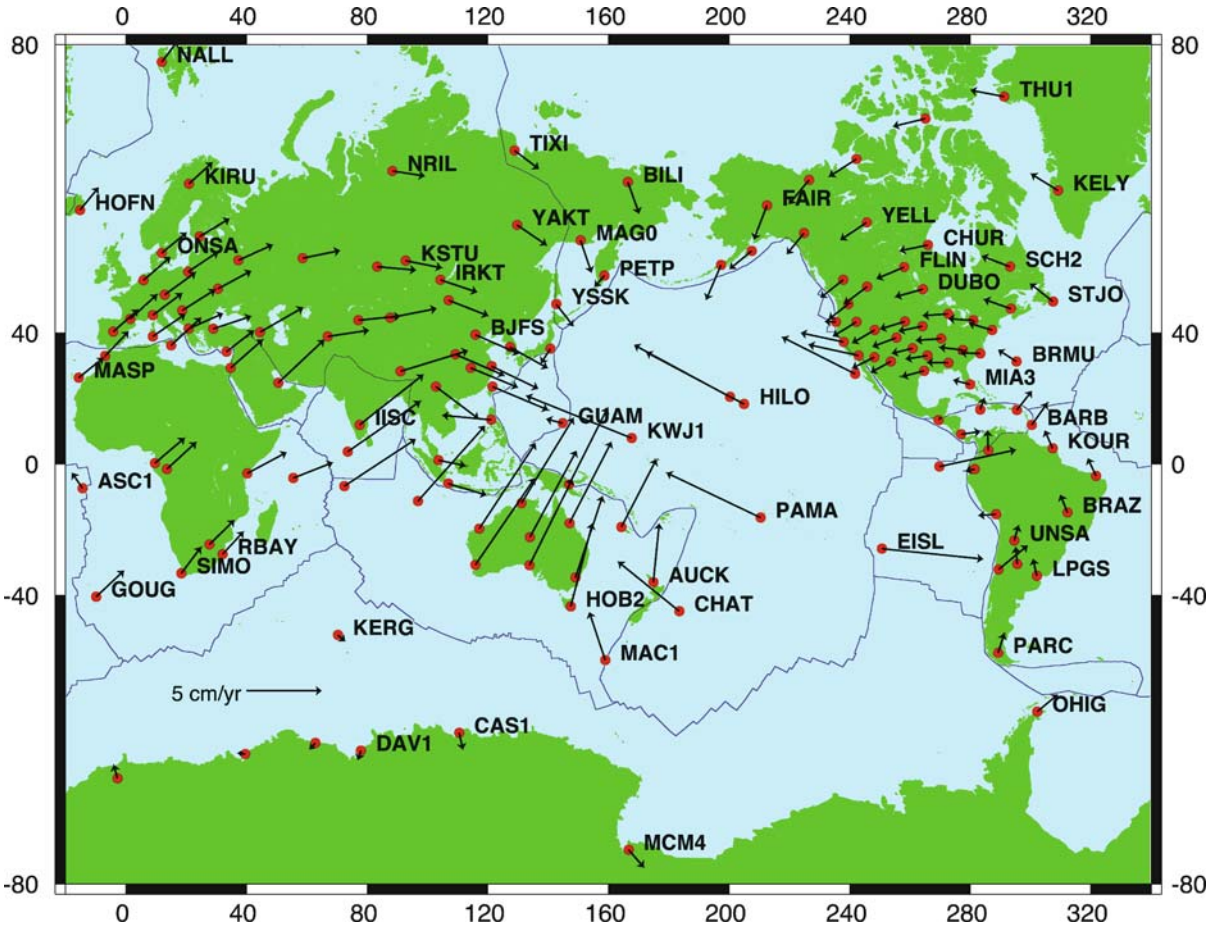

GPS: Applications in Crustal Deformation Monitoring, Figure 3
Velocities for a globally distributed selection of GPS sites. Figure courtesy NASA/JPL-Caltech

least in their interiors (away from plate boundaries where plates interact with one another). One approach to measuring global plate motions combines rate information obtained from the age of magnetic reversals recorded in the basalt of the oceanic crust formed at mid-ocean ridges with information regarding the direction of plate motion gleaned from the orientation of transform faults and direction of slip in earthquakes. Because most such studies use a magnetic anomaly that is ~3 million years old, the rates of plate motion inferred in this manner represent an average over the time since the Pliocene. The most commonly used plate motion model of this type is called NUVEL-1A [29,30]. This model gives relative velocities for pairs of plates. Under the assumption that the lithosphere has no net rotation with respect to the mantle below, Argus and Gordon [2] developed the NNR-NUVEL-1A model of absolute plate motions, where NNR refers to "no net rotation."

In contrast to approaches based on geologic data, geodetic techniques enable the estimation of present-day plate motions using essentially instantaneous measurements. Two such geodetic methods, Satellite Laser Ranging (SLR) and Very Long Baseline Interferometry (VLBI), were used in the 1980s to measure the positions of points on the Earth's surface. Given a good spatial distribution of sites world-wide, it is possible to use such data to estimate plate velocities. However, expense and practical considerations limited the number of sites that could be observed with SLR and VLBI. The advent of GPS provided a cost-effective alternative enabling precise three-component positioning with dense spatial coverage globally (e. g., Fig. 3).

Several studies have used GPS data, either alone or in combination with other observations, to estimate global plate motions (e. g., [4,79,124,140]), and numerous other studies have used such data in analyses focused on subgroups of plates. The plates are modeled as rigid, rotating,

spherical caps. Their motions are expressed as Euler vectors, defined by a location of the Euler pole (point E in Fig. 4a) and the angular velocity of the plate around that pole. These parameters are related to the velocities of the GPS sites by

$$\mathbf{v} = \boldsymbol{\Omega} \times \mathbf{r} \tag{1}$$

where $\mathbf{v}$ is the GPS velocity vector, $\boldsymbol{\Omega}$ is the Euler vector for the plate in question, and $\mathbf{r}$ is the position vector of the GPS site in Cartesian geocentric coordinates (e. g., [79]). The relationship given in (1) may be understood from Fig. 4. The Euler vector and position vector may be expressed as unit vectors ($\boldsymbol{\Omega}_u$ and $\mathbf{r}_u$) multiplied by their magnitudes (ω and R)

$$\boldsymbol{\Omega} = \omega \boldsymbol{\Omega}_u \tag{2}$$

$$\mathbf{r} = R\mathbf{r}_u \tag{3}$$

where ω is the velocity of rotation (generally in degrees or radians per million years), and R is the radius of the Earth (Fig. 4a). As shown in Fig. 4b, the distance, $\mathbf{d}$, that a GPS station travels during a time period t due to rotation about the Euler pole is given by

$$\mathbf{d} = \omega t R \sin\delta \tag{4}$$

where δ is the angular distance between $\boldsymbol{\Omega}$ and $\mathbf{r}$ (Fig. 4a). The quantity $sin\delta$ is simply the cross product of the unit vectors for the Euler vector and the position vector,

$$\sin\delta = \boldsymbol{\Omega}_u \times \mathbf{r}_u \,. \tag{5}$$

Therefore, plugging (2), (3), and (5) into (4), and dividing by time gives (1). The system of equations given in (1) can be solved to estimate the Euler vector that best fits the GPS data for each plate.

The early studies (e. g., [4,79]) were able to estimate velocities for only a handful of plates (e. g., six and eight, respectively). In recent years, however, the global distribution of CGPS sites has grown substantially such that Prawirodirdjo and Bock [124] estimated the velocities of 17 "major and minor tectonic plates" and Sella et al. [140] considered 19 "plates and continental blocks." The increasing precision of GPS measurements has made it possible to more rigorously test plate rigidity and the existence of purported plate boundaries, as well as address the potential systematic velocity error introduced by GIA in North America and Eurasia [124,140].

Velocities of GPS sites used in plate motion studies have repeatedly been found to be consistent with the assumption that plates behave rigidly, as evidenced by the fact that in general a single Euler pole fits the GPS velocities for a given plate well. Moreover, estimates of global plate velocities inferred from these data have agreed with plate motions inferred from geologic and other data (e. g., NUVEL-1A). However, the improvement in size and quality of GPS datasets has highlighted some discrepancies be-

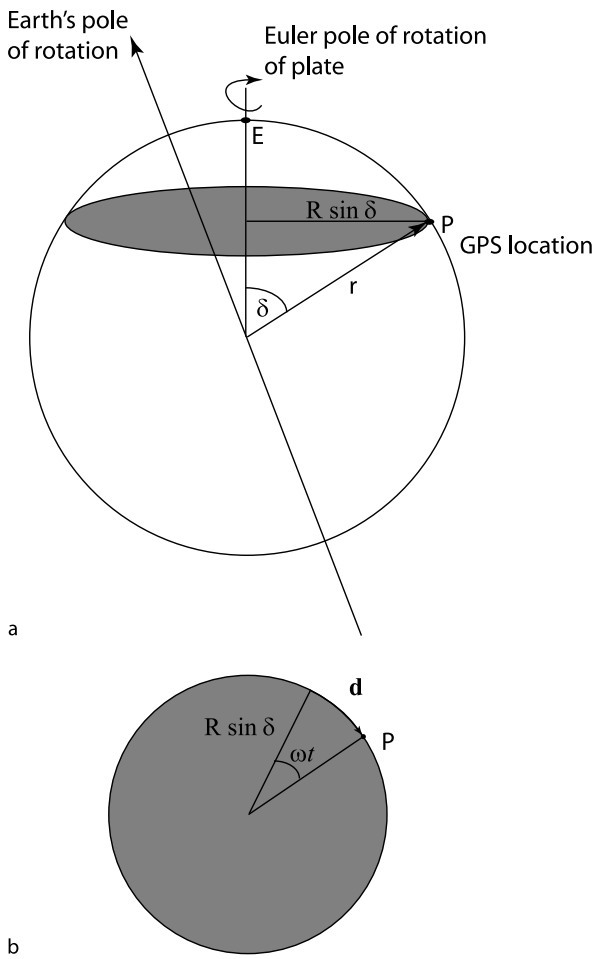

GPS: Applications in Crustal Deformation Monitoring, Figure 4
Relationship between the velocity of a GPS station and an Euler pole of rotation. Point P is the location of a GPS station, and point E is the location of the Euler pole of rotation, $\boldsymbol{\Omega}$, that describes the motion of the plate on which point P sits. The Euler vector points from the center of the Earth to point E, and its magnitude is the rate of rotation, ω. Point P is located at an angular distance of δ from point E. The velocity recorded at point P due to rotation about the Euler pole is given by $\mathbf{v} = \boldsymbol{\Omega} \times r$, where $\mathbf{r}$ is the position vector of the GPS station, and its origin is the center of the earth. Therefore, the length of $\mathbf{r}$ is the radius of the earth, R. The *gray circle* that passes through the point P has radius $R \sin\delta$. b View of the *gray circle* in a looking along the direction of the Euler vector. The distance traveled by point P during a time period t due to the rotation given by $\boldsymbol{\Omega}$ is given by $\mathbf{d}$. See text for details

tween NUVEL-1A and the models arrived at geodetically (e. g., [140]). These differences may be due in some cases to systematic errors in the rates estimated from magnetic anomaly data in tectonically complex areas where relative plate motion is not localized at the spreading center. In other cases the difference may reflect an actual change in the rate of relative plate motion over the past 3 million years. Refining both the geologic and geodetic plate motion models continues to be an area of active research (e. g., [3,28]).

Earthquake and Volcano Source Modeling

Source Potency and Geometry Although the Earth's plates behave rigidly, as evidenced by the velocities of GPS sites in the stable interiors of plates, at their edges neighboring plates interact, colliding, diverging, or sliding past each other. These processes, as well as others such as the movement of magma underground, impart stress to the Earth's crust, distorting the shape of a volume of crustal material. Some of this deformation is permanent, leading for example to mountain building. However, the brittle upper portion of the Earth's crust deforms elastically, meaning that a large portion of crustal deformation is recoverable; once the stress is relieved, the crustal material returns to its pre-stress shape. The discussion presented here will focus primarily on modeling which assumes an elastic crust, however several studies have considered the effects of nonelastic material properties (e. g., [37,65,135]).

Mathematical expressions from continuum mechanics that describe the stress, strain, and displacement in an elastic solid due, for example, to a point source or to movement on a planar dislocation can be applied to the study of crustal deformation sources using surface displacement data like those provided by GPS. Okada [109,110] presented concise analytic expressions that are used in many such studies today. Mogi [99] discussed the special case of surface displacement due to three orthogonal, equal-amplitude point sources of inflation which could represent an inflating or deflating magma body at depth. Dislocation sources are described by their dimensions, orientation, location in the Earth's crust, amount of movement (e. g. slip) which takes place across them, and the direction that the material on one side of the dislocation moves relative to that on the other (sometimes termed the "sense of slip"). Volcanic sources such as dikes and sills can be modeled by dislocations with opening rather than slip. Magma chambers are often represented by inflation sources defined by their locations, amount of inflation, and, in the case of more complicated geometries such as ellipsoidal sources, their shapes. Dzurisin [40] gives a good discussion of approaches for modeling a variety of volcanic deformation sources.

The term "source geometry" refers to all source characteristics except the amount of slip, opening, or inflation. These latter three parameters, which describe the strength of the source, are sometimes referred to as the "source potency." The surface displacement field produced in any deformation event reflects not only the source potency but also the source geometry and characteristics of the crustal material. Figure 5 presents the expected horizontal and vertical displacement due to different modes of shear slip and opening on a planar dislocation and to a point source of inflation. As can be seen, each deformation source produces characteristic surface displacement patterns. As with the estimation of plate motions described earlier, a system of equations can be written that relates a deformation source such as a dislocation in the crust to the displacements measured with GPS at the Earth's surface. In simplified terms this system of equations can be written as

$$\mathbf{d} = \mathbf{G}\mathbf{s} \tag{6}$$

where $\mathbf{d}$ is a vector of station displacements measured by GPS, $\mathbf{s}$ is a vector of source potency (e. g., fault slip), and $\mathbf{G}$ is a matrix which embodies the mathematical expressions relating potency to displacements for an assumed fault geometry and elastic properties. This system of equations can be solved (or "inverted") to estimate the unknown potency that best fits the known displacements. The displacements at the Earth's surface are nonlinearly related to the source geometry, but are linearly related to the source potency. Therefore, when the source geometry is known, inverting for the potency is a linear inverse problem. In the simplest case the potency can be assumed to be uniform for the deformation source. In the case of a fault, this implies that the same amount of slip occurred everywhere on the fault, and the vector $\mathbf{s}$ would have just one element. However, in the presence of multiple sources, the total displacement at the surface is simply the sum of the contributions from all the sources (e. g., Fig. 6). This means that spatially variable fault slip can be estimated by dividing the model fault into a grid of subfaults, each of which contribute to the observed displacements, and estimating the slip on each subfault. In this case the length of the vector $\mathbf{s}$ would be the number of subfaults. The number of subfaults used is generally dictated by how much data are available.

Inversion of geodetic data for characteristics of deformation sources is underdetermined, meaning that a large number of source models can fit the data within errors. Having vertical and horizontal displacement measurements improves the ability to distinguish among different

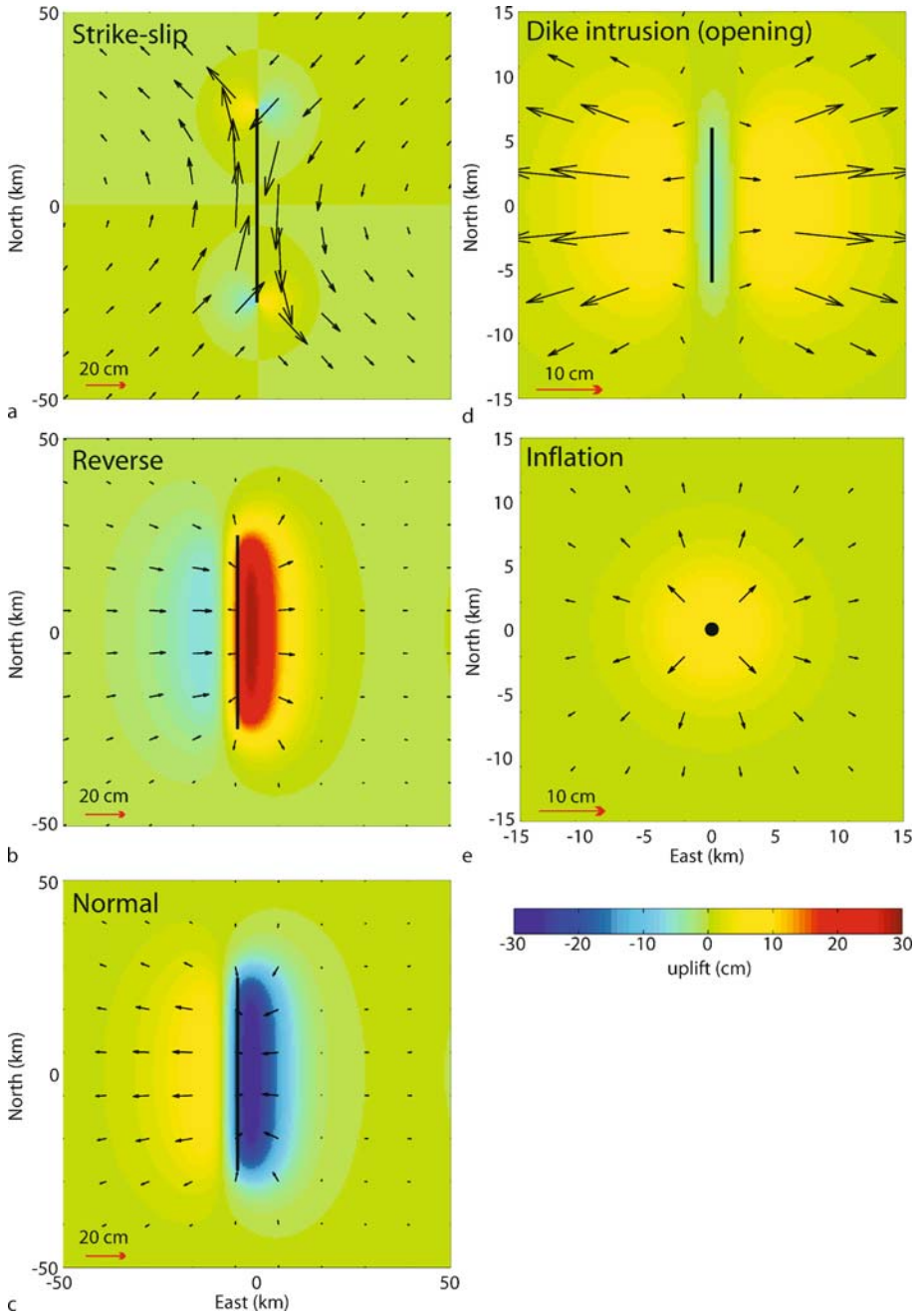

GPS: Applications in Crustal Deformation Monitoring, Figure 5

Predicted displacement due to dislocation and inflation sources assuming a homogeneous elastic half space. *Vectors* show horizontal displacement and *colored background* shows vertical motion. Note change of vector and spatial scales from a – c to d, e. *Heavy black line* in a – d is surface projection of upper edge of source dislocation. *Black circle* in e is surface projection of point source of inflation. See Table 1 for source parameters

GPS: Applications in Crustal Deformation Monitoring, Table 1
Source parameters for deformation sources depicted in Fig. 5

Panel in Figure 5	Description	Length (km)	Width (km)	Depth to top (km)	Dip (degrees)	Slip or opening (m)	Sense of slip	Inflation (10^6 m^3)
a	Dislocation	50	14	0	90	1	Right lateral, strike slip	
b	Dislocation	50	10	6.34	60	1	Reverse	
c	Dislocation	50	10	6.34	60	1	Normal	
d	Dislocation	12	5	3	90	1	Opening	
e	Inflation point source			5				10

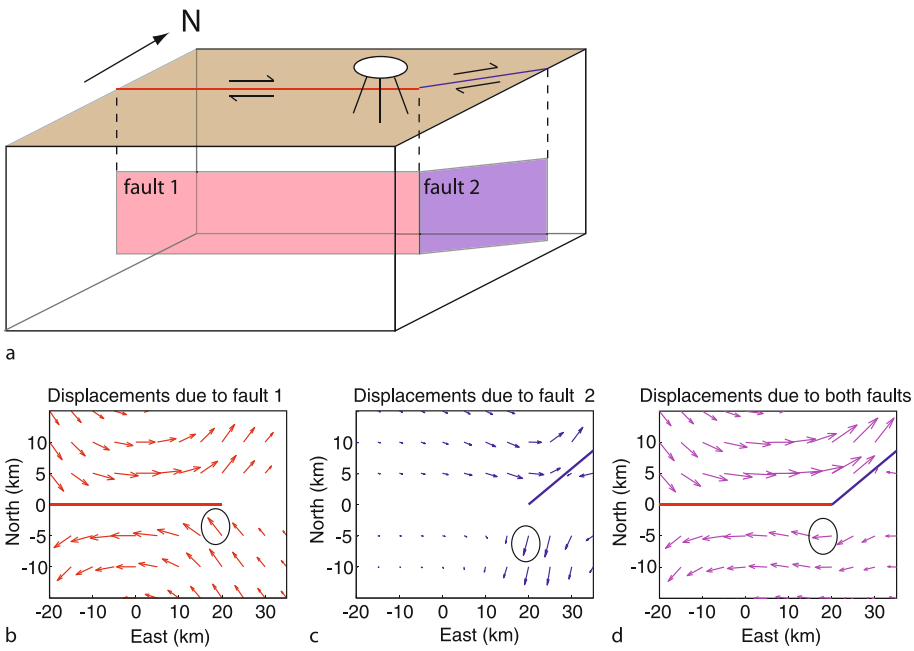

GPS: Applications in Crustal Deformation Monitoring, Figure 6
Predicted displacements due to two sources. a Source geometry consisting of two vertical strike-slip faults and location of a GPS receiver. b Map view of displacement due to slip on the fault shown in *red* in a. c Map view of displacement due to slip on the fault shown in *blue* in a. d Map view of displacement due to slip on both faults. Note how the displacement at the *circled locations* in b and c due to the individual faults is very different from that which would be recorded by GPS in d due to slip on both faults

possible source geometries. For instance, several types of volcanic sources will produce similar patterns of vertical deformation, but with the inclusion of horizontal displacement measurements it is possible to differentiate among them. When using GPS data to estimate the spatial distribution of slip on a fault (by dividing the model fault into subfaults and estimating the slip on each), spatial smoothing is often used to provide added constraints in the inversion. The justification for this is that abrupt changes in the amplitude of slip would result in high stresses on the fault surface, which is physically unlikely. The relative weight given to fitting the data and to spatial smoothness is often determined empirically (e. g., [161]). Non-negativity may also be applied, for instance to include prior knowledge about the sense of slip on a fault as a constraint on fault slip estimates.

Regardless of these means for regularizing inversions, because geodetic measurements are collected at the surface, their sensitivity to the details of a deformation source decreases with depth. GPS measurements will be most sensitive to source processes occurring in the upper few kilometers of the Earth's crust near the GPS receiver's location. Deeper sources will affect GPS sites over a broader region, but the recorded deformation signal will lack detail about the source. This can be best understood if one thinks of an earthquake that causes rupture of the Earth's surface.

A GPS receiver near the fault will record data which primarily reflect the shallow slip and surface rupture close to that receiver. A receiver 15 km away from the fault will not be sensitive to the shallow, near-fault deformation, and the recorded signal will be due to large-scale features of slip on deeper parts of the fault. Likewise, a receiver at the summit of a volcano can record the movement of magma that is collecting near the crater, whereas receivers lower on the flanks of the volcano will likely not record that signal but could be expected to track deformation due to movement of magma at greater depths.

The structure of major faults such as the San Andreas has been studied extensively by mapping, imaging the spatial distribution of background seismicity, and applying geophysical techniques that use, for example, seismic reflection and refraction, gravity, or magnetic data to highlight contrasts in rock properties. Therefore, if an earthquake hypocenter is found to be located on a major fault, the source geometry may be well-known a priori. In the absence of such information, or to refine the source geometry used in inversions, the spatial distribution of aftershocks and the location and extent of any surface rupture are also used.

Traditionally, spatially sparse geodetic measurements were assumed to be insensitive to the details of the model fault geometry used in inversions. With the recent growth of spatially dense GPS networks, however, more physically realistic fault geometries have been required. Maerten et al. [84] showed that using non-planar fault geometries that better represented independent information, for example from surface rupture mapping, led to significant improvement in fits to the GPS displacements for the 1999 M_w 7.1 Hector Mine earthquake. Methods for precisely relocating seismicity (e. g., [163]) have illuminated fine-scale fault structures that were previously obscured in less-precise catalog locations. This has enabled the development of more realistic model fault geometries.

For example, Murray and Langbein [101] used displacements measured with GPS to estimate the slip distribution of the 2004 M6 Parkfield, California earthquake. This event took place on the well-studied San Andreas fault in central California. Earlier work (e. g., [41,149]) had suggested that the fault was essentially vertical, with a strike of 149°. During the 2004 event, the coseismic displacement recorded for one GPS station, CARH, located within about 500 meters of the fault was in the opposite direction to what is predicted for a right lateral strike-slip fault (e. g., Fig. 5a). In order to fit the data for this station, an additional fault structure was needed. A sub-parallel fault called the Southwest Fracture Zone (SWFZ) which had exhibited movement in a previous earthquake at this locale was a likely candidate. Using relocated aftershocks of the 2004 event [159] as a guide, Murray and Langbein [101] developed a non-planar fault geometry that consisted of the primary San Andreas fault and a subsidiary SWFZ. These two model fault surfaces passed through the relocated seismicity and intersected the mapped surface traces of the faults at the Earth's surface. Using this fault geometry, Murray and Langbein [101] inverted the GPS data to image the coseismic and postseismic slip associated with this earthquake.

Unlike in the case of Parkfield described above, often very little is known a priori about the geometry of a deformation source. This is especially true for sources that lie completely underground such as magmatic intrusions and earthquakes that do not cause any surface rupture (e. g., "blind" thrust events). However, as discussed earlier (Figs. 5 and 6), the surface displacement field produced in any deformation event reflects not only the source potency but also the source geometry. Therefore, by observing the spatial pattern of displacement using GPS, it is possible to infer what type of source lies underground by finding the source model that best predicts the observed data, for instance by using the expressions for deformation in an elastic material. Since surface displacements are nonlinearly related to source geometry, parameters describing the geometry cannot be estimated using linear inversion techniques but rather must be found through nonlinear optimization. Cervelli et al. [21] give a good overview of several approaches to this type of problem. When estimating the geometry and potency of an earthquake or opening source, often a two-step approach is employed: first the source geometry is estimated assuming uniform fault slip or dike opening, and then the inferred geometry is held fixed and the spatial distribution of slip or opening is estimated. As described in the following examples, the ability to infer the source geometry can help answer important questions about the underlying processes driving deformation such as whether a fault terminates in a décollement or how magma sources interact.

(a) Deep Structure of the Chelungpu Fault, Taiwan Johnson and Segall [65] used GPS-measured displacements caused by the 1999 M_w 7.6 Chi-Chi Taiwan earthquake to constrain the geometry of its rupture surface. They then went on to answer fundamental questions about the seismotectonics of Taiwan, which has formed due to the collision between the Philippine Sea plate and the Eurasian plate.

The extensive surface rupture that accompanied the Chi-Chi earthquake suggested that oblique reverse / left-lateral slip occurred on the previously known Chelungpu

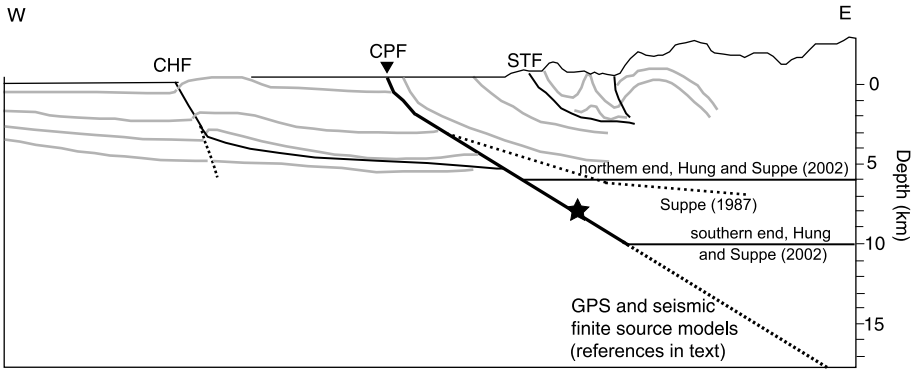

GPS: Applications in Crustal Deformation Monitoring, Figure 7
Cross section of the Chelungpu fault in Taiwan. The *solid, near-horizontal black lines* and the *upper dotted line* represent the ramp and décollement structure of the "thin-skinned" model, while the *steeper dotted line* below 10 km is the deeper extension of the Chelungpu fault envisioned by the "thick-skinned" model. CHF, Changhua fault; CPF, Chelungpu fault; STF, Shuangtung fault; *star*, hypocenter of 1999 Chi-Chi earthquake. Reprinted from [65] with permission from Elsevier

thrust fault which strikes north-south and dips ~30° east. A nearly horizontal décollement structure has been interpreted to exist at depths of 6 to 10 km beneath much of Taiwan based on geologic mapping, seismicity locations, and seismic reflection profiles. In one proposed deformation model, termed the "thin-skinned" model, thrust faults like the Chelungpu fault intersect the décollement at depths of <10 km (Fig. 7). The alternative "thick-skinned" model interprets seismic and gravity data to suggest that, although a décollement may have controlled the long-term tectonic evolution of the island, more recent deformation has taken place on down-dip extensions of thrust faults with the same dip as their shallower portions (Fig. 7). In the case of the Chi-Chi earthquake, aftershocks occurred both at depths which might coincide with a décollement and considerably deeper, potentially on the down-dip extension of the Chelungpu thrust fault.

Johnson and Segall [65] optimized the source geometry of the Chi-Chi earthquake using a half-space with laterally and vertically varying shear modulus. They showed that the GPS data required slip on a thrust plane that transitions into an essentially horizontal décollement structure at ~8 km depth. Moreover, they showed that displacements at the north end of the fault, where the surface rupture changed to a more east-west orientation, could only be fit by an additional thrust fault, which they term a "lateral ramp," extending to the depth of the décollement. The estimated slip distribution of the earthquake using this geometry is shown in Fig. 8. These authors use the results for the Chi-Chi event as the basis for a conceptual model in which deformation follows the thin-skinned model, and lateral ramps form north of the Chelungpu fault due to this fault's orientation oblique to the direction of plate convergence.

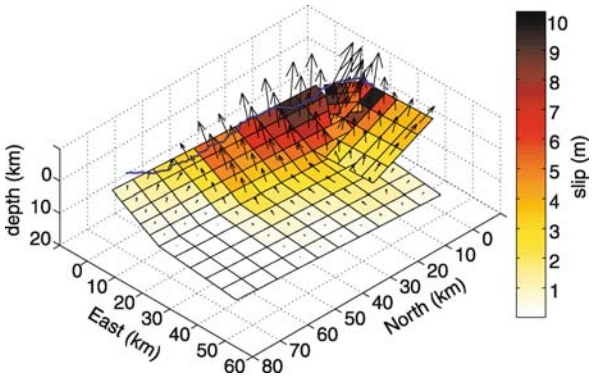

GPS: Applications in Crustal Deformation Monitoring, Figure 8
Slip distribution of the 1999 M_w 7.6 Chi-Chi Taiwan earthquake and optimized fault geometry inferred from GPS data. *Colors* indicate magnitude of slip and *vectors* show the direction that the hanging wall moved relative to the foot wall. The *blue curve* shows the trace of the earthquake rupture at the Earth's surface. Reprinted from [65] with permission from Elsevier

(b) Magma Plumbing System at Kilauea Volcano, Hawaii
Kilauea volcano, on the Big Island of Hawaii, consists of a summit crater as well as two rift zones, themselves consisting of several craters, extending from the summit down the flanks of the volcano (Fig. 9). Since 1983 Pu'u O'o, a collection of volcanic vents in the East Rift Zone, has been the center of eruptive activity, apparently fed by magma flowing through lava tubes from Kilauea summit.

In January 1997 a fissure eruption occurred on the East Rift Zone at Napau crater, ~3 km closer to the summit

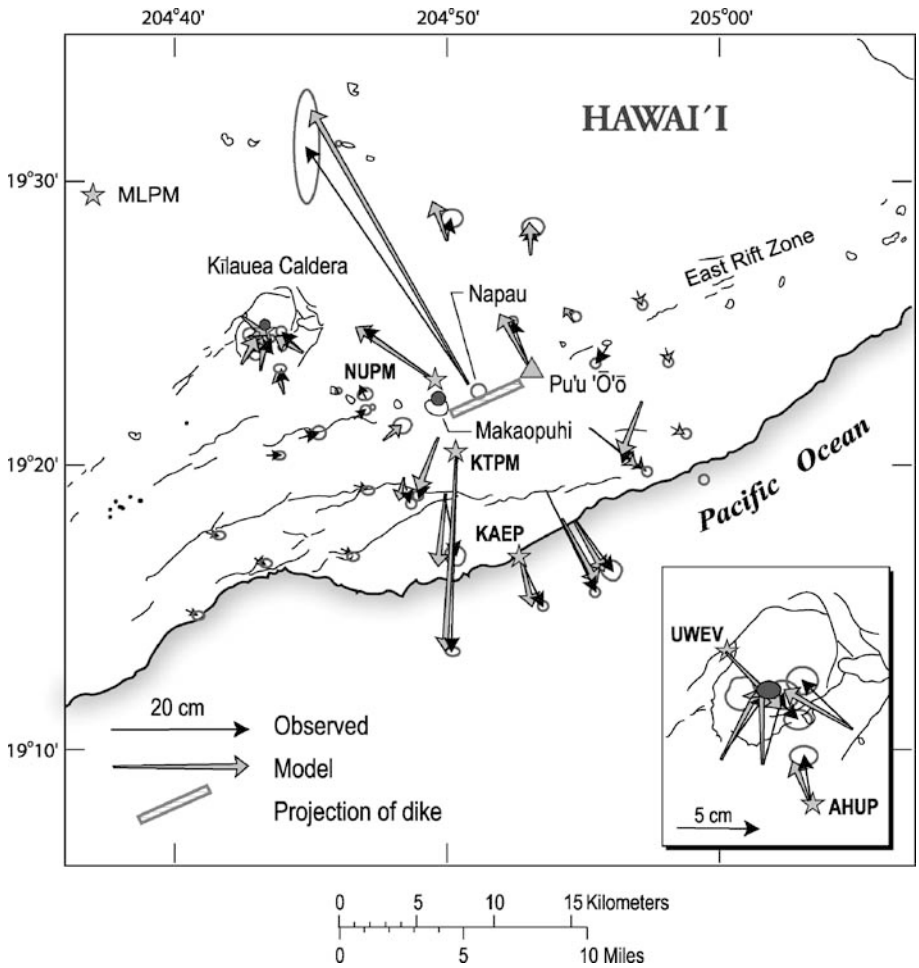

GPS: Applications in Crustal Deformation Monitoring, Figure 9

Displacements measured using GPS during the 1997 fissure eruption at Kilauea volcano. On the basis of the observations (*black vectors* with 95% confidence *ellipses*) Owen et al. [112] inferred the source of the deformation to be a combination of deflation at the summit of Kilauea (inward pointing *arrows*; see *inset*) and the intrusion of a dike along the east rift zone culminating in the eruption. The displacements predicted by this source geometry are shown by the *gray arrows*. CGPS sites are indicated by *stars* with four-character station codes. The deflation sources are shown as *gray circles* and the dike location by the *hachured rectangle*. *Thin black lines* are faults, fractures, and fissures. Adapted with permission from [112] (copyright 2000, American Geophysical Union)

than Pu'u O'o. In addition to uplift, the GPS instruments recorded horizontal displacements around the eruptive fissure at Napau Crater that were directed outward from the rift zone except at the ends of the fissure where they pointed inwards, parallel to the rift (Fig. 9). The magnitude of the displacements died off quickly with distance from the fissure. This displacement pattern is characteristic of a shallow dike intrusion within the rift zone. Prior to the eruption, sites at Kilauea summit showed subsidence and a radially inward pattern due to horizontal shortening across the summit, and this accelerated drastically during the eruption. This pattern suggests deflation, and ultimately emptying, of a magmatic source beneath the summit.

Using nonlinear optimization, Owen et al. [112] found that a source geometry consisting of a steeply dipping dike aligned with the rift and the fissures, combined with deflation both at the summit and at Makaopuhi crater, best fit the GPS observations (Fig. 9). The volume change at the two deflation sources was an order of magnitude less than the volume of the inferred dike intrusion. However, this discrepancy can be remedied if magma previously stored in a lava lake at Pu'u O'o which was seen to drain during this event, along with magma in a conduit thought to con-

nect the summit magma chamber to Pu'u O'o, also contributed to the intrusion. That the period of time leading up to the eruption had been characterized by steady deflation of the summit indicates that dike formation was not in response to magma overpressurization at the summit, but rather some other process. Owen et al. [112] suggest that ongoing southeastward movement of the south flank of the volcano (e. g., [111]) created tensile stresses that encouraged the dike intrusion and fissure eruption at the rift. Following the eruption, inflation resumed at the summit and the lava lake refilled, further evidence of the connection between the two magma reservoirs.

Combined Use of Multiple Data Types Wherever possible, multiple data types are used together to infer crustal deformation source characteristics. For example, GPS data are frequently used in combination with other geodetic measurements such as Interferometric Synthetic Aperture Radar (InSAR) [18] and leveling [38] data. GPS measurements are also often used in combination with seismic records to estimate fault slip.

(a) GPS and InSAR GPS data are commonly used in combination with InSAR data because of the complementary nature of these two data types. GPS observations provide three-component displacements, good horizontal precision, and (in the case of CGPS) good temporal coverage. InSAR, on the other hand, has exceptional spatial coverage, is more sensitive to vertical deformation than GPS, and does not require the deployment of instruments on the ground (thus enabling data collection from otherwise hazardous areas such as volcanoes). InSAR, jointly with GPS where possible, has been widely used to study volcanic deformation (see for example [39,40,119]). Likewise, GPS and InSAR observations, in some cases in combination with seismic data, have been used to infer the slip distribution and rupture history of numerous earthquakes including the 1992 Landers earthquake [54], the 1995 Kobe earthquake [113], the 1999 Hector Mine earthquake (e. g., [70,132,148]), and the 1999 Izmit earthquake [16,27]. Studies have explored the relative constraints on source parameters provided by each data-type, as well as approaches for optimally weighting different data types in inversions, especially when one method, like InSAR, produces many more data points than another, such as GPS (e. g., [71,126,148]). Wright et al. [168] used a combination of InSAR and GPS data to infer the slip distribution of the 2002 Denali earthquake. Because of the remote location of this event, GPS sites were clustered along roads [57,58], thus resulting in poor spatial coverage along some portions of the fault. In this situation InSAR observations helped reduce uncertainty in the slip estimates compared to estimates obtained from GPS data alone.

(b) GPS and Seismic Data It has long been recognized that GPS data are a useful complement to seismic records for estimating fault slip [162]. While seismic data are sensitive to the rupture process of an earthquake (the amount and temporal progression of slip), when using these data to characterize the rupture process, trade-offs exist between the time-history of slip and its spatial distribution. GPS offsets due to an earthquake reflect only the final (or "static") slip distribution, and thus the slip history inferred from the seismic data can be constrained to produce a static slip distribution that fits the geodetic displacements. The combined use of these two data types tends to have the added advantage of improved instrumental coverage over the study area.

Wald and Heaton [162] conducted a comparison of slip distributions for the 1992 M_w 7.2 Landers earthquake inferred from strong motion, teleseismic, and geodetic data (GPS displacements and displacements calculated from trilateration measurements) individually and in a combined inversion. For all inversions they used a consistent fault geometry parametrization comprised of three fault segments based on the aftershock locations and the extensive ground surface rupture that was mapped following the event. They corrected the geodetic data for the effects of the M_w 6.2 Big Bear event and assumed that post-Landers measurements were made soon enough to avoid contamination by postseismic signals.

On inspection of the final slip distributions (Fig. 10) obtained from inversions of each of the three datasets independently, the authors identified several features that were common to all three and thus appeared to be robust regardless of dataset. For example, slip at the hypocenter was moderate and limited to a small depth range. Peak slip at depth occurred along the central portion of the fault, while slip became shallower at the ends (to the northwest and southeast). The greatest near-surface slip was on the Camp Rock / Emerson faults at the northwest end of the rupture. The slip distribution obtained through combined inversion of the three datasets was most similar to that from the geodetic data alone because the timing of slip is an additional degree of freedom in the inversion of teleseismic and strong motion waveform data that is not available in the inversion of static offsets. In addition to the final slip distribution, these authors imaged the temporal progression of slip on the fault surface. The rupture appears to slow as it nears the surface, as well as when it approaches the two step-over regions between fault segments. Furthermore, the authors infer that although the

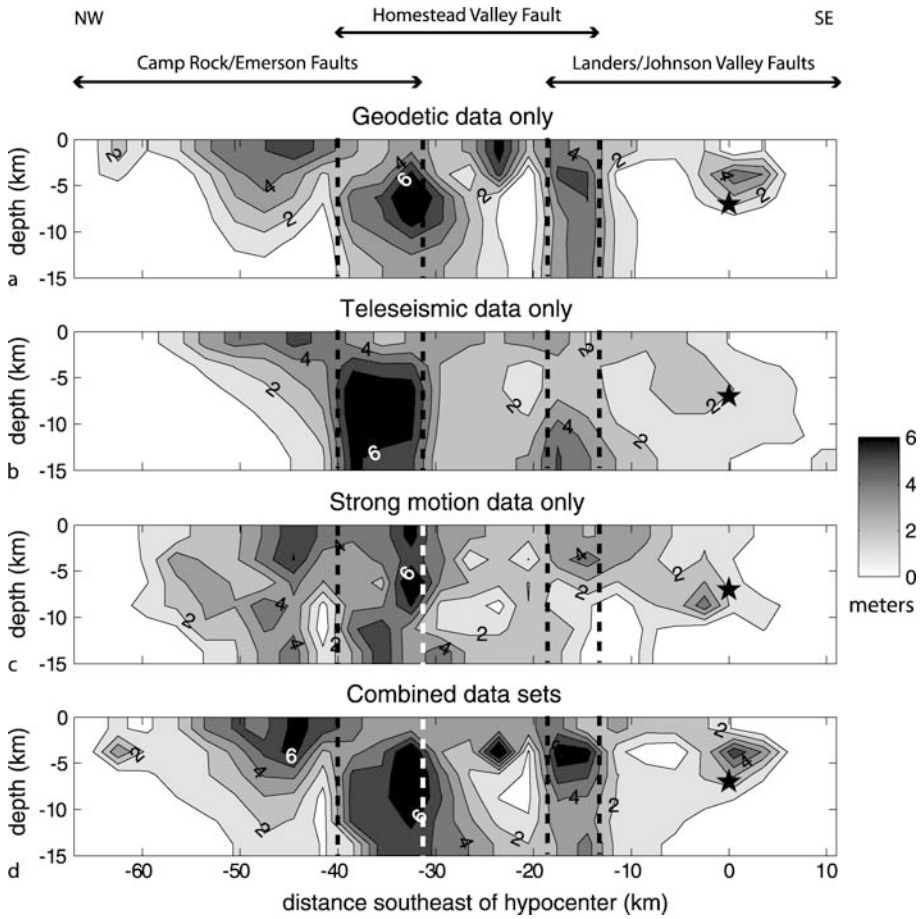

GPS: Applications in Crustal Deformation Monitoring, Figure 10
Slip distributions estimated by Wald and Heaton [162] for the 1992 M_w 7.2 Landers earthquake using different data sets. Contour interval is 1 meter. *Star* marks hypocenter of earthquake. *Dashed lines* indicate along-strike boundaries of faults named at top. a Geodetic data only. b Teleseismic data only. c Strong motion data only. d Combined inversion of all three data sets. Adapted from [162]

rupture generally propagated unilaterally northwest, each time the rupture jumped northwest to a subsequent segment it also propagated backwards down the abandoned portion of this segment southeast of the fault intersection. Thus, the combined use of geodetic and seismic data provided a more complete and robust understanding of the rupture dynamics of this earthquake, which could then be used to investigate, for example, the spatial patterns in the strength of ground shaking due to this event.

Another example of the combined use of geodetic and seismic data comes from the Dec. 26, 2004 Sumatra–Andaman earthquake, a subduction zone megathrust event which ruptured a ~1200 km length of the plate boundary between the Indo-Australian and Eurasian plates [11,81]. This earthquake, which produced peak-to-peak surface wave motions greater than 1 cm worldwide [115] and measurable static offsets at GPS sites at least 4500 km away [6] and probably farther [73], resulted in more than 283,000 deaths, largely due to its triggering of a major tsunami. This is the largest earthquake to have been recorded since the establishment of digital seismic networks and GPS, and both seismic and geodetic data have been critical in describing the rupture process of this earthquake.

The moment magnitude of an earthquake, M_w, is a measure of its size and may be derived from the seismic moment (M_o), which is defined as

$$M_o = \mu s A \tag{7}$$

where μ is the shear modulus of the faulted rock (in units of Pascals), s is the amount of fault slip during the earthquake (meters), A is the area of the surface that slipped in the earthquake (meters^2), and M_o has units of New-

tons × meters. "Tsunami earthquakes" [120] are a subclass of subduction zone earthquakes defined as producing tsunamis larger than would have been predicted from their moment magnitude. Tsunami earthquakes have been observed to have slow rupture velocity with relatively little seismic energy release at high frequencies. Although the moment magnitude should represent the net amount of static slip that occurred in the earthquake, it is typically estimated from seismic data at periods of 100 to 300 seconds [81]. If significant seismic energy is released at longer periods, this method will underestimate moment magnitude. The Harvard CMT (Centroid Moment Tensor) solution for the Sumatra–Andaman event, computed using surface wave data with periods of 300 to 500 seconds, had $M_o = 4.0 \times 10^{22}$ Nm which corresponds to M_w 9.0 [81]. However, researchers quickly began to see evidence that this earthquake was in fact larger.

Ammon et al. [1] analyzed seismic data from a broad frequency range (including data with periods up to 54 minutes) and concluded that these observations could be fit by a model in which the majority of the rupture occurred over a 10 minute time span and slip was concentrated south of 8° N (on the southernmost ∼800 km of the fault surface, Fig. 11). However, they note that the large GPS displacements reported in the Nicobar and Andaman islands would require 2- to 3-times more slip north of 8° N. Studies of the Earth's seismic free oscillations [116,152] used these very low-frequency data (at periods up to one hour) to estimate the moment of the earthquake and found that moment increased with the period of the oscillations. These data suggest that the rupture required ∼10 minutes to travel from south to north, had more slip in the Nicobar and Andaman islands region than originally thought, and led to M_w estimates of 9.13 to 9.3. Park et al. [116] also point out that even slower slip (e. g. over a time span of ∼1 hour) could have occurred and would be difficult to detect in the free oscillation data.

Banerjee et al. [7] compiled GPS data from several studies and used these observations to infer the static slip distribution on the rupture surface. The total moment release associated with their slip estimate is 7.62×10^{22} Nm (corresponding to an M_w 9.22), smaller than that estimated by Stein and Okal [152], but greater than the 6.11×10^{22} Nm (M_w 9.13) inferred from GPS data by Kreemer et al. [73]. Banerjee et al. [7] conclude that from ∼2 to ∼16 meters of slip on portions of the fault north of 8° N is needed to fit the GPS data from the Andaman and Nicobar islands. However, they argue that this slip did not occur slowly over a time span of an hour or more because the continuous GPS site at Phuket, Thailand showed little movement after ∼10 minutes following the earthquake [160], and other sites in Thailand which should be particularly sensitive to slip on the Andaman segment given their location also do not show movement more than 10 to 20 minutes following the earthquake [53]. Although some slip estimates based on geodetic data alone have higher moment than those from seismic data alone, suggesting the occurrence of aseismic slip, Chlieh et al. [24] and Rhie et al. [129] conducted joint inversions of GPS and seismic data and were able to fit both datasets satisfactorily with a single rupture model.

Thus, although seismic data provide information on the details of fault rupture that cannot be obtained from GPS, in the case of the Sumatra–Andaman event the GPS observations provided needed constraints on the extent and duration of fault rupture, both of which had important implications for tsunami generation.

(c) GPS and Gravity Data GPS data have become widely used to study volcanic deformation processes, including the long-term uplift observed at calderas such as Yellowstone (Wyoming), Campi Flegrei (Italy), and Long Valley (California). Geodetic observations can constrain the source geometry and volume change [8]. They cannot, however, discriminate if the deformation is due to an influx of hydrothermal fluids or the intrusion of magma. Battaglia et al. [8,9] address this problem through the combined use of geodetic and gravity data recorded during a period of uplift in Long Valley caldera. Modeling the observed uplift using a point source will produce biased results if the true source does not possess spherical symmetry. Furthermore, the uplift signal of a range of source geometries can be similar, but the horizontal deformation signal can help in distinguishing among different models (e. g., [32]). Battaglia et al. [8] use a combination of vertical and horizontal geodetic measurements (GPS, leveling, and line-length data) to find the best-fitting source geometry, in their case a vertical prolate ellipsoid. Then, with that source geometry uniquely determined, they perform a joint inversion of the uplift and gravity data to infer the volume and mass of the source, from which they obtain a density range of 1180 to 2330 kg/m^3 [9]. Since this density range is too high for hydrothermal fluids to be the sole source of uplift at Long Valley, these authors conclude that a silicic magma body or combination of magma and hydrothermal fluids is required to produce the observed deformation.

Deformation over Time Scales of Decades to Seconds

We have seen that GPS data are useful for constraining models of deformation that happens rapidly, such

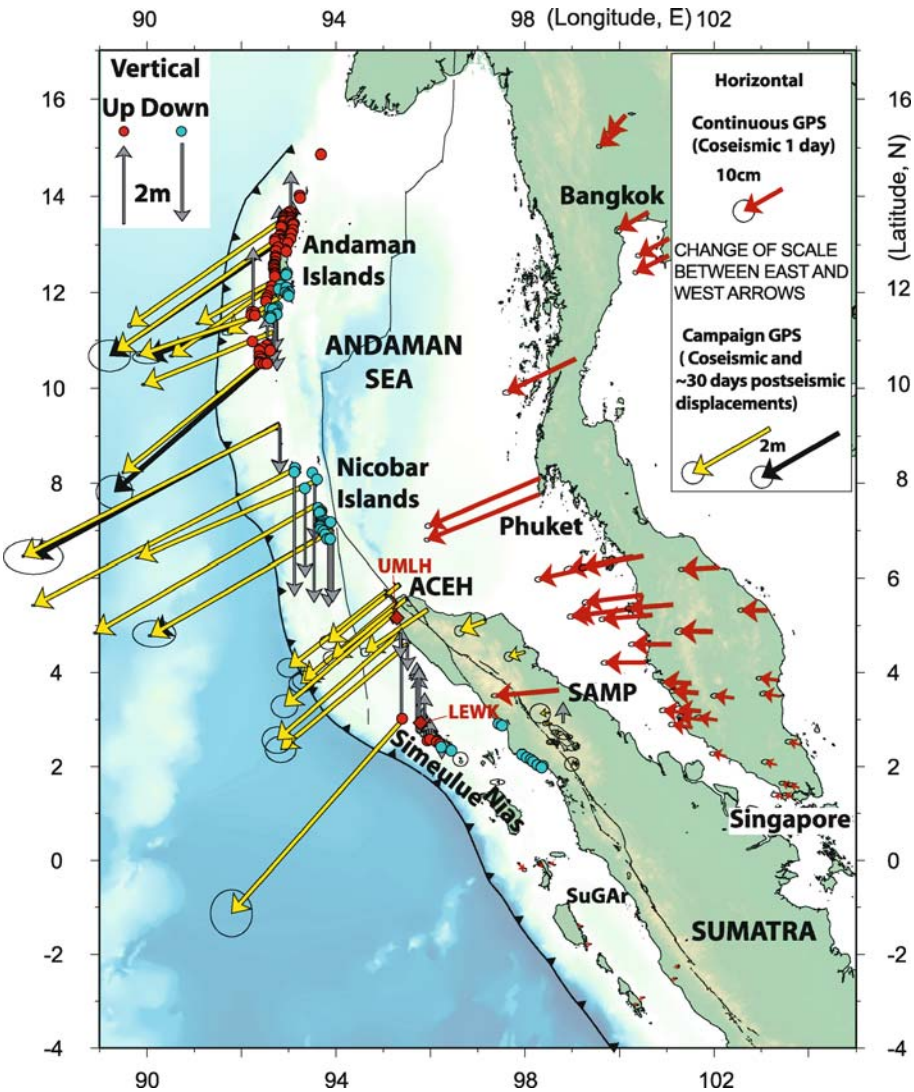

GPS: Applications in Crustal Deformation Monitoring, Figure 11
Summary of geodetic data recorded in the vicinity of the 2004 Sumatra–Andaman earthquake. The campaign GPS (*yellow* and *black*) vectors are compiled from [49,153] and contain approximately one month of postseismic deformation. The continuous GPS data come from Vigny et al. [160]. Note that the near-field vectors (those in the western part of the mapped region) and the far-field vectors use a different scale. *Dots* represent measurements of vertical deformation from satellite imagery [91]. The *gray arrows* indicate uplift and subsidence from GPS data, measurements of the vertical movement of coral heads, and mapping of shoreline changes [12,49,153]. Figure adapted from [24]

as slip in an earthquake. However, perhaps the greatest strength of GPS is its ability to record deformation that occurs over a wide range of time periods. GPS can measure fault slip that happens too slowly to generate seismic waves (termed aseismic slip), volcanic deformation that occurs over several days or several years, and long-term interseismic strain accumulation. GPS can provide much more temporally dense measurements than InSAR and, although less sensitive, is stable to longer time periods than strainmeter data as shown in Fig. 12 [68].

Interseismic Deformation Interseismic deformation refers to the gradual straining of the Earth's crust that occurs during the time between moderate to large earthquakes. This strain can be caused by the build up of stress that will eventually be released during earthquakes, as well

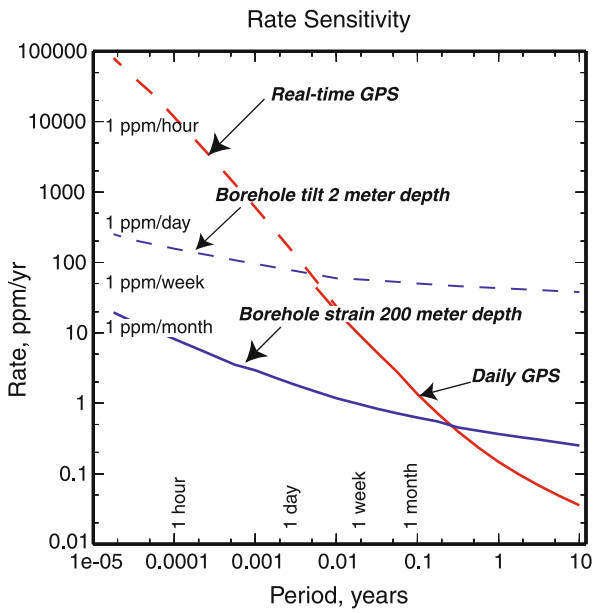

GPS: Applications in Crustal Deformation Monitoring, Figure 12
Comparison of rate sensitivity for GPS, borehole strainmeters at 200 meters depth, and borehole tiltmeters (which measure the gradient in vertical deformation) at 2 meters depth. The *x*-axis indicates the time period that may be spanned by the different data types. For example, daily GPS measurements may span one day or longer. The *y*-axis indicates the strain rate that can be resolved as a function of the period. Strain is the change in length (area or volume) per a unit length (area or volume), and thus is unitless and can be expressed as parts per million (ppm). As can be seen from the plot, the borehole tiltmeters and strainmeters are more sensitive than GPS at shorter periods, but at periods longer than a few days and a few months, respectively, GPS measurements provide better resolution of strain rates. Figure courtesy of John Langbein

as reflect the broader-scale patterns of deformation in response to tectonic plate motion.

(a) Block Versus Continuum Models for Deformation As one moves from the global scale of tectonic plates to the continental scale, a major question is whether continents deform through the movement of many small rigid blocks (like plates on a smaller scale) or through more continuous deformation [157]. GPS data have been used in numerous studies to try to elucidate this hotly debated issue, and one region that has been a focus of study is the Tibetan plateau which accommodates strain due to the collision of India with Eurasia. GPS velocities for a profile of stations spanning central Tibet roughly parallel to the direction of maximum convergence between India and Eurasia show a linear gradient in the component of velocity parallel to the convergence direction. This observation has been cited as evidence for continuously distributed deformation across the region (e. g. [171]) or distributed strain combined with the movement of a small number of crustal blocks [23]. In contrast, Thatcher [158] showed that the velocities predicted by a model in which the Tibetan plateau was divided into a set of rigid, rotating crustal blocks defined by faults and other geologic features could also fit the data since the difference between the observed GPS velocities and those predicted by the block model were relatively small and did not exhibit widespread systematic spatial patterns. Although the rigid block model does not predict a linear velocity gradient across central Tibet, Thatcher [158] notes that both the block and distributed models fit the velocity profile within errors. The spatial sampling provided by GPS remains sparse across large portions of the Tibetan plateau, and it is likely that additional data will be necessary to resolve the outstanding question as to whether continental deformation in this region is primarily block-like or continuous.

(b) Estimating Interseismic Fault Slip Rates Following the 1906 San Francisco earthquake, H. F. Reid [127] used his observations of deformation in that event to formulate a description of the earthquake cycle which he termed "elastic rebound." Reid recognized that strain builds up in the Earth's crust around faults during the interseismic period, and that strain is eventually released in earthquakes. Reid's hypothesis predated the theory of plate tectonics, but we now understand that the source of the ongoing stress affecting the faults is the motion of the Earth's plates. The strain build-up and release causes measurable deformation of the Earth's surface. Areas where the crustal strain measured using GPS is large are likely to have earthquakes to relieve that strain. These events may take place on faults that are not visible on the earth's surface, and in this case the geodetic data can provide an important clue to the existence of seismic hazard. GPS enables measurement of deformation during all phases of the earthquake cycle.

One of the most important pieces of information needed to characterize a region's seismic hazard is an estimate of the slip rates on the major active faults that could affect that region. Most seismically active regions are at plate boundaries. The Earth's rigid plates are always moving at rates that are essentially constant over time periods comparable to earthquake recurrence intervals (e. g., hundreds to thousands of years). At plate boundaries the relative motion between two neighboring plates is accommodated on faults. The portion of the rate of relative motion that is accommodated across a given fault is that fault's slip rate. Slip rates are often estimated from geo-

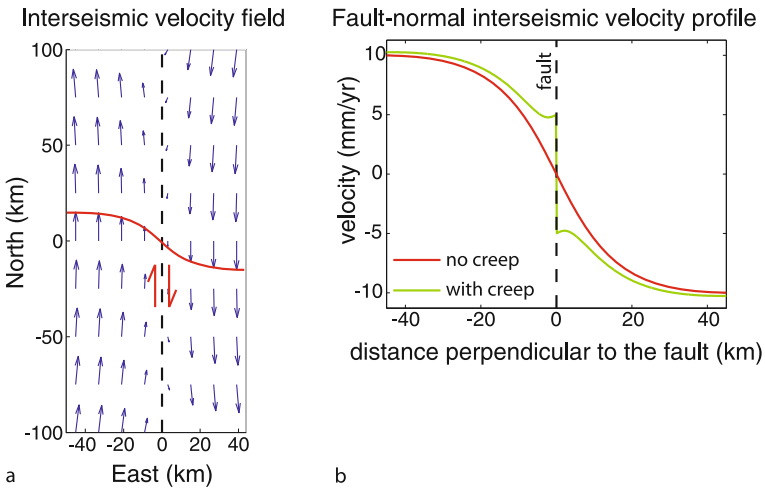

GPS: Applications in Crustal Deformation Monitoring, Figure 13

a Expected velocity field due to interseismic slip below 15 kilometers on a vertical strike slip fault. The *red curve* highlights the difference in velocity near and far from the fault. b Expected velocity profiles for a locked fault and one that exhibits shallow creep. The locked fault is as in a. The creeping fault slips below 15 km depth and creeps in the uppermost 4 km but is locked from 4 to 15 km depth. Note that in the case of a creeping fault, there is an offset in the velocity profile close to the fault. The near-fault inflection in the *green curve* shows the combined effect of the strain due to slip below the locked zone and creep that reaches the Earth's surface

logic data, for example by dating samples collected from a location in which a measurable offset of a stream channel has occurred due to movement on a fault that crosses the stream [146]. However, slip rates can be inferred from geodetic data as well.

Below a certain depth in the Earth's crust (e. g., ~15 km for many strike-slip faults in the San Andreas system) the temperature and pressure are sufficiently high for the crustal rock to behave plastically in response to stress, rather than experiencing brittle failure. Earthquakes occur above this depth, which is termed the "brittle-ductile transition," the "locking depth," or the "transition depth," but not below. In the time between moderate to large earthquakes most faults are largely locked above the transition depth, meaning no movement occurs across them. (It is true that very small earthquakes occur frequently on most faults, but these events affect a relatively small portion of the fault's surface area and release a tiny fraction of the energy, or moment, released in moderate and large events.) The material below the transition depth deforms gradually and continuously in response to plate motion, and a fault that is locked above the transition depth may exist as a zone of distributed shear below that depth. In the vicinity of a locked fault, the constant movement of the material below the transition depth strains the elastic crust above. This is manifest by a characteristic pattern of interseismic velocities for points on the Earth's surface near the fault. For example, the interseismic velocity profile perpendicular to a strike slip fault like the San Andreas will have a characteristic sigmoidal shape as shown in Fig. 13, the details of which reflect the locking depth and slip rate of the fault.

In regions dominated by a small number of major faults of known geometry (e. g., the San Andreas fault system in northern California, or subduction zones of Cascadia and Japan), interseismic slip rates can be inferred in much the same way as was done for earthquake slip (Eq. (6)) using relatively simple dislocation models in which the fault is prescribed to be locked above the transition depth and freely slipping below that (simulating the motion of the tectonic plates). A number of studies (e. g. [47,125]) have taken this approach to model the GPS velocities of sites in northern California as the superposition of interseismic velocity signals due to slip on the multiple sub-parallel faults that make up the San Andreas system. When the slip rate on a fault is estimated from interseismic GPS velocities, the resulting value is generally called the interseismic slip rate to emphasize that it has been estimated from data collected over a time period entirely within the interval between two earthquakes on the fault in question. When slip rates are estimated from geologic data they are often called "long-term average" rates to emphasize that they represent the rate over many earthquake cycles. In the absence of post-seismic effects or other transient deformation (described in more detail in a later section) these two slip rate estimates for a given

fault should be the same. In subduction zones, interseismic velocities are often used to estimate the degree of plate "coupling," which reflects the size of the locked zone that may rupture in a large earthquake [17,44,93,108,154,164].

As discussed earlier in the context of the Tibetan plateau, one interpretation of continental deformation patterns is that they arise from the rotation of fault-bounded blocks, and the GPS data can be used to estimate the Euler poles of rotation for each block. Data from sites near known faults may be discarded from the analysis because these observations will likely reflect the interseismic elastic strain accumulation due to the faults rather than the long-term rigid behavior of the blocks. Slip rates on the faults that bound the blocks can be calculated from the relative rates of block rotation (e. g. [158]). An alternative block modeling approach retains all the data, and the estimated block rotation rates must result in slip rates on block-bounding faults that are compatible with the patterns of strain accumulation recorded in the GPS velocity field. As with the dislocation models, a transition depth is assumed for each fault. The portion of each fault above the transition depth is treated as locked, and slip on the portion below the transition depth drives the observed strain accumulation.

Block modeling of this type lends itself to the many seismically active regions, such as southern California, that are characterized by numerous faults with complex geometries. Unlike models which represent individual faults by separate dislocations, block models are required to be self-consistent in that the rotation rates must be compatible for all blocks, slip rates have to be consistent at fault intersections, and the total slip rate across the region is made to match the relative plate rate. A drawback of this approach, however, is that it is difficult to accommodate dipping faults and to constrain fault-perpendicular motion in a realistic way.

Both block models and dislocation models suffer from trade-offs between slip rate estimates on different faults and sensitivity to poor data coverage. Furthermore, it is difficult to resolve the contribution to the observed GPS velocity from strain accumulation on closely spaced faults (e. g. within two locking depths of each other), and thus slip rate estimates on neighboring faults tend to trade-off with each other and with the assumed locking depth. It will always be difficult to resolve slip rates on faults that are close together given that displacement measurements are confined to the Earth's surface.

Several studies have applied the technique of block modeling with inclusion of elastic strain accumulation to the western United States [26,89,90]. For many faults the slip rates estimated by these studies agree to within errors with those estimated from geologic studies. However, there are some discrepancies. For example, the ~5 mm/yr slip rate Meade and Hager [90] estimated for the San Bernadino segment of the San Andreas fault is considerably lower than the geologic estimate of ~25 mm/yr. Discrepancies between geodetic and geologic slip rates have been found in several locations world wide, but it is not the case that one data type tends to produce consistently higher rate estimates than the other. The differences likely result from a combination of factors, including assumptions made in interpreting the data, the localized nature of geologic estimates, and the different time periods spanned by the two data types. Consistency between geodetic and geologic slip rates is an area of ongoing study.

Time varying deformation is often observed following large earthquakes. One source of this signal in many cases is the viscoelastic response of the material below the transition depth. When a large earthquake occurs it imparts stress to this material, which then deforms slowly, restressing the elastic crust above. The rate of the resulting strain would be expected to be high directly after the earthquake and die off with time. Likewise, the strain accumulation rate on the affected fault would vary throughout the earthquake cycle. These processes have been incorporated into another category of interseismic deformation models called "viscoelastic cycle" models [66,67,133,137,156], which can be used to estimate fault slip rates and earthquake recurrence times. These models account for the strain rate maxima on faults during the interseismic period as well as temporal variations in the months to years following a large earthquake.

(c) Fault Creep An interesting interseismic phenomenon, termed "fault creep," is observed on several faults of the San Andreas system in central California and the San Francisco Bay Area, as well as faults in Taiwan, the Philippines, and Turkey. Faults exhibiting creep slip steadily or episodically at low average rates (e. g. 10 mm/yr). Some creep events are confined to the uppermost ~500 meters of faults, but in other cases creep occurs deeper at what are generally considered to be seismogenic depths (e. g. to a depth of ~15 km on faults of the San Andreas system). However, this fault slip is too slow to generate seismic waves. The ongoing slip of creeping faults is a nuisance, offsetting cultural features like curbs and buildings. However, ongoing creep constantly relieves stress which would otherwise be released in an earthquake and thus reduces the seismic hazard due to that fault. For example, a 2003 study of earthquake probabilities for the San Francisco Bay area found that accounting for creep on the Calaveras fault reduced the predicted rate of $M \geq 6.7$

earthquakes on this fault by a factor of three [167]. Therefore, knowing the extent of creep is valuable in hazard assessments.

Fault creep will have a different interseismic signature than that for a fault that is locked in the seismogenic zone (Fig. 13b). Instead of the smooth transition seen across a locked fault, there will be a step because there is fault offset near the surface. Therefore, GPS data can be used to infer the depth-extent of fault creep. Several studies have used GPS data, in some cases in combination with InSAR, creepmeter, and microseismicity observations, to estimate the spatial distribution of fault creep (e. g. [64,85,103,136]).

Temporally Varying Deformation Continuous GPS measurements are particularly useful for observing transient, or time-varying, deformation. Sources of transient deformation include slow slip events, postseismic response, and volcanic processes.

(a) Slow Slip Events Slow slip events, sometimes also referred to as "slow earthquakes" or "silent earthquakes" (the latter emphasizing the lack of a seismic signature to the event), are a phenomenon in which fault slip occurs at too slow a rate to generate seismic waves. The duration of slow slip events that have been observed geodetically world-wide ranges from days (e. g. [22,106]) to years (e. g. [98,102]). Precursory transient slip with duration of minutes has been observed prior to earthquakes on mid-oceanic ridge transform faults using seismic data (e. g. [62]).

Shallow creep events and a multi-year transient increase in slip rate have been observed along the San Andreas fault where continuous or frequent monitoring using creepmeters, strainmeters, and two-color electronic distance measuring instruments provided temporally dense measurements [51,52,75,82]. In 1996, a couple of years after the establishment of the Japanese CGPS network, evidence for slow slip in the Boso Peninsula region near Tokyo became apparent in the data [131]. As the spatial coverage of CGPS networks worldwide improved it became clear that slow slip events were much more frequent than previously thought, occurring over a variety of spatial and temporal timescales and tectonic settings.

The majority of large slow slip events observed to date have occurred in subduction zones, including those of Japan, the Pacific Northwest of the United States, Mexico, New Zealand, and Alaska. In 2001 a surprising pattern was recognized in time series for several CGPS sites in the Cascadia region in the Pacific Northwest of the United States (Fig. 14a). The subduction interface is thought to be locked near the Earth's surface and freely slipping at greater depths. The ongoing deep slip results in interseismic motion of GPS sites toward the over-riding plate (Fig. 14b). In the case of the Cascadia sites, this means that the long-term average interseismic movement is eastward. However, it was observed that occasionally the sites briefly

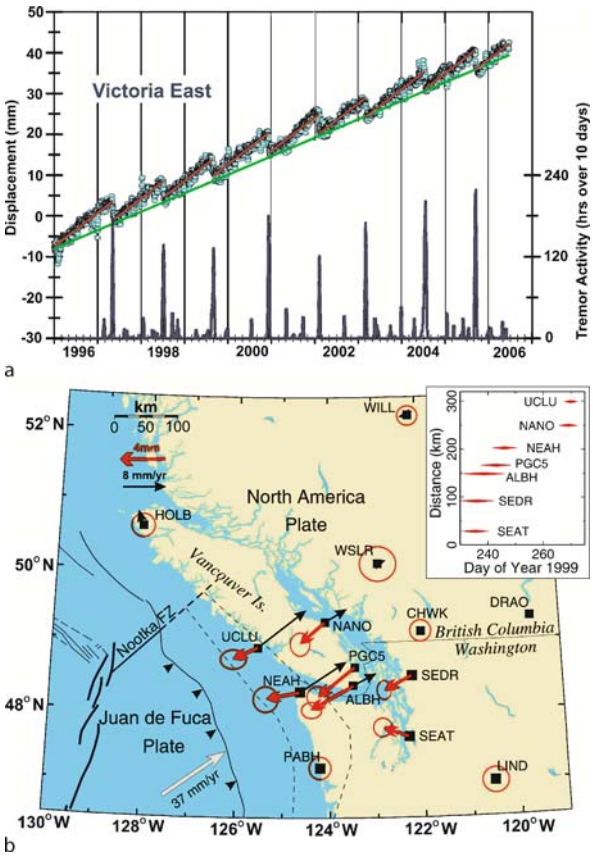

GPS: Applications in Crustal Deformation Monitoring, Figure 14
a Time series for station ALBH located in Victoria British Columbia, Canada relative to stable North America. The *blue dots* are station positions. Although the overall interseismic movement of this site is eastward due to the ongoing strain caused by subduction (*green line*), every ∼14 months this site moves westward (steps in time series). The *red line* represents the average velocity in the time between slow slip events, which is a higher rate than the long-term interseismic movement (*green line*). The *blue curve* represents the time series of nonvolcanic seismic tremor. Periods of increased tremor activity coincide with the times at which the GPS site shows anomalous westward movement [130]. b Portion of the Cascadia subduction zone. *Black vectors* are interseismic velocities of continuous GPS sites relative to stable North America, and *red vectors* are anomalous displacements during the 1999 slow slip event. The *inset* shows relative timing of transient displacements among different sites. From [35]. Copyright, Her Majesty the Queen in right of Canada (2001)

moved in the opposite direction, causing a step-like pattern in the position time series. An intriguing feature of the observed reversals in station velocities is that the pattern was found to repeat approximately every 14 months [94].

That several CGPS sites in the region showed a coherent reversal in the direction of motion at the same time (Fig. 14b) suggested that the source of this signal could be slip on the interface between the down-going slab and over-riding plate of the subduction zone. Modeling of the GPS data [35] indicated that the source region was a portion of the subduction interface between 30 km and 40 km depth that is a transitional zone linking the shallow fully locked and deeper fully slipping parts of the interface.

In 2002 Obara [105] reported a very low frequency seismic signal called tremor, typically observed at active volcanoes, emanating from ∼30 km depth in the subduction zone of southwest Japan. A similar signal was soon discovered in Cascadia, and researchers quickly realized that the tremor occurred simultaneously with slow slip events in those locales (Fig. 14a). Furthermore, they identified the source regions of the tremor and found that it coincides spatially with the inferred source region of the slow slip [106,130]. In volcanic settings tremor is thought to be caused by the movement of fluids through conduits underground. Recent studies have found that slow slip events in subduction zones are accompanied by very low frequency earthquakes and that the tremor in these locales is actually made up of many low amplitude low frequency events [142,143]. One interpretation is that shear slip on the subduction interface, rather than fluid flow, causes nonvolcanic tremor, implying that tremor and slow slip both arise from the same underlying process of shear slip. Although the tremor may not be directly caused by fluid flow, the low frequency earthquakes and slow slip appear to coincide spatially with areas of high fluid pressure in the pore spaces of subduction zone rocks. The high pore pressure, perhaps resulting from metamorphic reactions that release fluid, may encourage shear slip [143].

Like Cascadia, other subduction zones, for example the Guerrero region of Mexico [83] and the Shikoku [106] and Tokai [55] regions of Japan, have also experienced, to varying degrees, quasi-periodic slow slip events. Slow slip, in some cases multiple events, has been observed in other subduction zones such as Alaska [107], New Zealand [34], and the Tokai [98] and Bungo channel [114] regions of Japan, but it remains to be seen if the transient slip is periodic. Even in Cascadia, which shows clear periodicity, there is variation in the periodicity along the strike of the subduction zone. For example, slow slip events have been observed in both northern and southern Vancouver Island with a ∼14 month periodicity, but the events in these two locations are 6 months out of phase with each other. Slow slip events in northern California, also part of the Cascadia subduction zone, have been found to have an ∼11 month recurrence interval [155]. In southwest Japan, the Shikoku region experiences short duration (on the order of a week), small amplitude slow slip events (in fact only detectable in the tiltmeter data) at six-month intervals coincident with tremor [106]. In contrast, the subduction zone beneath the Bungo channel region, which abuts the Shikoku region directly to the southwest, experiences infrequent large slow slip events lasting ∼1.5 years [114]. Similarly, in the Tokai region, short-term and long-term slow slip events seem to occur on nearly overlapping parts of the subduction zone [55,98].

Similar to fault creep, the cumulative effect of several slow slip events may be to relieve stress on the transition zone along the whole length of the subduction interface without a large earthquake. However, slow earthquakes in the subduction zones of Cascadia, Japan, and elsewhere may impart stress to the locked subduction interface up dip and thus increase the likelihood of a large earthquake [35,123]. Understanding both the mechanism of these events and what it means for seismic hazard continues to be a focus of intense study.

Transient slip has also been observed in non-subduction zone settings. GPS data have shown that the south flank of Kilauea volcano in Hawaii moves seaward at a rate of several cm/yr [111], perhaps a manifestation of the instability of the volcanic edifice. However, in 2001 it was observed that this motion sped up for a few days. Cervelli et al. [22] modeled the GPS observables during the period of increased station velocity and concluded that the source was a ∼M 6 slow earthquake lasting 36 hours and generating an average of almost 9 cm of slip on a nearly horizontal thrust fault about 4.5 km underground. This observed transient signal occurred nine days after a storm caused 1 meter of rainfall on this part of the Big Island. Cervelli et al. [22] calculated that given reasonable values for the porosity and permeability of the rocks in the study area, the slow slip could have been triggered by an increase in pore fluid pressure as the rain water penetrated into faults on the volcanic edifice. Subsequently several more very similar transient events have been observed in this locale by GPS [139], not accompanied by anomalously high rainfall. Moreover, it has been recognized that these events are accompanied by increased seismicity in an adjacent area following the onset of the deformation signal. This lends credence to the interpretation that the displacement signal is due to fault slip and suggests that the transient slip triggers the seismicity. In order for slow slip to trigger the observed seismicity, the slip must occur deeper

than originally thought. A model in which slip occurs at the depth of the interface between the volcano and the underlying ocean floor, ~8 km, fits the GPS displacements as well as one in which slip occurs at more shallow depths (~4.5 km) [139].

(b) Postseismic Deformation Often following a moderate or large earthquake, continued aseismic deformation is observed. This may be due to several sources, including continued slip on the fault surface, diffusion of pore fluid, and the viscoelastic response of the lower crust (below the transition depth) and upper mantle. All of these processes are triggered by the stress changes imparted to the surrounding crust and mantle by the earthquake. More than one source can be active simultaneously, and the effects change and decay with time. The spatial and temporal evolution of postseismic signals can provide important insights into the frictional, hydrological, and rheological characteristics of the crust and upper mantle, and enable a better understanding of the way stress is redistributed in the crust, which is an important consideration in assessing seismic hazard.

Afterslip is continued slip on the fault plane after the rapid slip which generates seismic waves has ceased. The effects can start immediately after the event (e. g. [76]) and can last for months or years, typically showing a logarithmic decay with time (e. g. [134]). Afterslip is thought to arise from the stress changes imparted by the earthquake to parts of the fault which have frictional properties that allow slow slip. Typically the total afterslip following an event is a fraction of the amount of slip that occurred coseismically. However, some events such as the Sanriku-haruka-oki [169] and Tokachi-Oki [97] earthquakes in Japan have had afterslip with moment release approaching or, in the case of the 2004 M_w 6 Parkfield earthquake [76], exceeding the coseismic moment. Although the factors controlling the amount of afterslip are not fully understood, the fact that the San Andreas fault near Parkfield, as well as many subduction zone faults, are known to exhibit fault creep may be a factor.

The spatial and temporal evolution of afterslip following the 2003 M_w 8.0 Tokachi-Oki earthquake in Japan was estimated from GPS data using a Kalman-filtering technique [97] (Fig. 15a). The rate of afterslip started quite high and died off gradually. These authors noted that the highest afterslip rates tended to localize around the area which slipped in the earthquake. This is not unexpected as the coseismic slip area would have just experienced a reduction in shear stress which would discourage further slip here. Interestingly, the afterslip seems to avoid areas inferred to have slipped in other earthquakes that occurred in the decades prior to the Tokachi-Oki event as evidenced in the plot of cumulative afterslip in Fig. 15b. This interpretation provides support for the idea that rapid coseismic and gradual aseismic slip may occur on different parts of the fault because of variations in frictional properties. Similar results were found from an analysis of afterslip following the 2005 Nias–Simeulue thrust event [59].

As described earlier in regards to viscoelastic cycle models of interseismic deformation, the stress changes imparted by a moderate or large earthquake to the material below the elastic crust causes time-varying deformation. The geodetically recorded postseismic deformation following several large events has been interpreted to reflect viscoelastic processes (e. g. [156]). The rate at which viscoelastic postseismic deformation decays depends in part on the viscosity of the lower crust and upper mantle. The temporal decay of displacements measured with GPS have been used to infer viscosity values and thus the relative strength of these two layers (e. g. [5,31,45,46,121,122]). Estimates for the viscosity of the lower crust range from 10^{19} to 10^{21} Pa s, and for the upper mantle range from 10^{17} to 10^{19} Pa s.

Stress changes in the crust due to fault slip in an earthquake compress the pore space of rocks in some areas and cause dilation of the pore space elsewhere, depending on the orientation of the fault and sense and distribution of slip. The resulting pore pressure gradients cause fluid to flow from areas of high pressure to those of low pressure. This fluid flow causes further time-dependent strain. Such effects have been observed following several earthquakes including a pair of moderate earthquakes that occurred in 2000 in Iceland [5,69] and the M_w 7.3 Landers event in California [43,118]. If poroelastic effects are confined to the upper few kilometers of the fault zone, the spatial extent of the resulting surface deformation will be localized near the fault. Also, vertical surface displacement is a large component of the poroelastic signal. Because GPS measurements of vertical displacement are noisier than the horizontal data, and because the distribution of GPS stations may be limited near the causative fault, much of the insight into poroelastic deformation following events like the Landers earthquake and the earthquakes in Iceland has come from InSAR data.

It is unlikely that postseismic deformation associated with a given earthquake can be explained by a single process. Given the often limited spatial and temporal data coverage and, in the case of GPS data, the sometimes poor vertical displacement control, it can be difficult to differentiate among different potential sources, e. g., afterslip, viscoelastic, and poroelastic deformation. However, the surface displacements due to each process can exhibit diag-

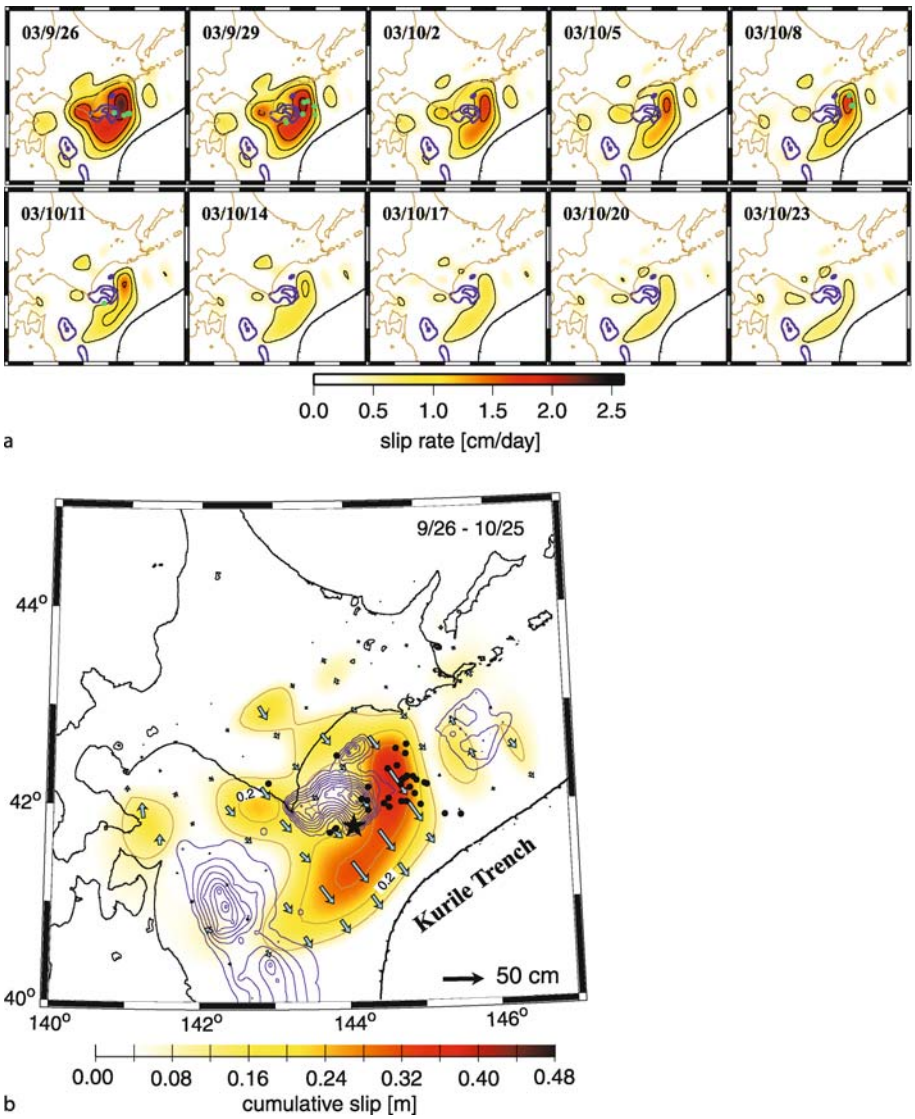

GPS: Applications in Crustal Deformation Monitoring, Figure 15

a Temporal evolution of the rate of afterslip (indicated by *colored shading*) on the subduction interface following the 2003 M_w 8 Tokachi-Oki earthquake. Dates given in upper left of each frame. *Blue contours* are the areas that slipped in previous earthquakes in this region. The *blue contours* in the center of the mapped area are those of the Tokachi-Oki earthquake. The *green dots* are aftershocks. b Cumulative afterslip in the first 30 days following the earthquake estimated from GPS data. Slip magnitude is given by the *colored shading*; the estimated amount and direction that the upper plate (the area northwest of the Kurile trench) moved relative to the lower plate is shown by the *arrows*. Afterslip tends to surround the areas inferred to have slipped in the Tokachi-Oki earthquake (epicenter given by the *black star*) and other events, shown by the *blue contours*. The *black dots* are aftershocks. Adapted with permission from [97] (copyright 2006, American Geophysical Union)

nostic patterns in time and space, which, if observed, make it possible to discern distinct causative processes. For example, due to its deeper source viscoelastic deformation should affect a broader geographic region than poroelastic deformation or shallow afterslip. Because viscoelastic deformation involves the response of material with viscosities on the order of 10^{18} to 10^{19} Pa s, while poroelastic processes involve the flow of aqueous fluid through the ground, the signal due to the former is expected to last considerably longer (e. g., several years) than that of the latter (e. g. several months). Several studies have analyzed postseismic deformation, in some cases observed us-

ing multiple data types, and attributed the observed deformation to a combination of two or more effects [5,43,46].

(c) Volcano Deformation Volcanic deformation is often characterized by transient signals. Magma or hydrothermal fluids migrate beneath the volcanic edifice, causing inflation or deflation and sometimes culminating in an intrusion or eruption. The steep flanks of volcanoes are often unstable, leading to landslides and in some cases collapse of large sections of the edifice. GPS is very well-suited to monitoring these types of deformation signals, and many of the world's volcanoes have CGPS receivers installed for this purpose. Some of the earliest GPS observations of volcanic deformation come from a submarine volcano near the Izu peninsula in Japan. Here GPS data from two receivers recorded deformation several days before volcanic tremor or visible signs of eruption were apparent [48,145].

Loss of instruments in a volcanic eruption is a real and costly risk. InSAR, which does not require any equipment on the ground, is a widely used method for monitoring volcano deformation. However, InSAR does not provide three-component deformation measurements and does not have the temporal resolution that CGPS does. InSAR also suffers from decorrelation when vegetation, snowfall, or lava flows change the land surface during the time between two image acquisitions, however it still generally provides better spatial coverage than GPS. Dzurisin [40] presents a good overview of InSAR as applied to volcano deformation.

During a recent eruption of Augustine, a stratovolcano in Alaska, GPS data proved to be a valuable complement to other monitoring systems [20]. In the summer of 2005 CGPS stations on the volcano began to record an inflation signal, following an increase in microseismicity below the volcano that had begun in May 2005 or perhaps earlier. Beginning in November 2005 the rate of inflation increased rapidly, but then died off somewhat in early January 2006. The inflation signal has been interpreted as evidence for a dike intrusion into the volcanic edifice which nearly reached the surface [20]. On January 11, 2006 a series of explosive eruptions began at Augustine, destroying two of the six CGPS receivers. By January 17th, though, the eruptions had died down, as had seismicity and gas emissions. Only the GPS data from two of the remaining stations showed continued inflation. The quiescence lasted 10 days before another explosive eruption and effusive lava flows took place. This is an example in which GPS data provided early corroboration in the summer of 2005 that increased seismicity was due to magma movement into the volcanic edifice. Moreover, during the ten days of quiescence when other indicators such as seismicity and gas measurements showed little activity, the GPS data showed that further eruptive activity was likely [20].

High-Rate GPS Continuous GPS networks typically record data at 15- or 30-second sampling rates, which is more than adequate for obtaining daily positions. One such CGPS network, the Southern California Integrated GPS Network (SCIGN), recorded data at 30-second intervals for the 1999 M_w 7.1 Hector Mine earthquake. Nikolaidis et al. [104] used these data to obtain positions at every observation epoch for the time spanning the earthquake and demonstrated that these observations, although aliased, agreed with those from nearby strong ground motion instruments. However, GPS receivers are capable of recording data at much higher rates, e. g. 1 Hz or greater. In recent years, for scientific as well as surveying and navigation purposes, an increasing number of CGPS sites have been set to record high-rate data and in some cases transmit them in real-time.

Unlike seismic data, GPS receivers provide a direct measure of displacement, and the instrument stays on scale even during the strong shaking of an earthquake. This makes it possible to obtain high-rate displacement time histories without the error introduced by integrating velocity or acceleration records from seismic instruments and without the data loss that occurs when the shaking exceeds the dynamic range of a seismic instrument (Fig. 16). Although the results of Nikolaidis et al. [104] hinted at a potential application of high-rate data, this kind of observation was not yet available for an earthquake. Such an event occurred in 2002 with the M_w 7.9 Denali earthquake in Alaska. Larson et al. [78], obtained displacement-time histories of seismic waves from 1 Hz GPS data recorded during this event and showed they were in good agreement with those obtained by doubly integrating accelerometer data.

Wang et al. [165] demonstrated that GPS receivers operating at 1 Hz provide a faithful recording of signals with periods of 2 seconds or greater. However, GPS receivers will not replace seismometers for recording earthquakes since they cannot capture the higher-frequency components of the seismic waves. Rather, they provide a valuable complement to seismic data. For example, Emore et al. [42] show that high-rate GPS data can be used as a constraint when integrating accelerograms. Moreover, the combined use of seismic and high-rate GPS data extends the observable frequency and amplitude range of seismic waves considerably [78]. Miyazaki et al. [96] demonstrated that GPS seismograms measured by 1 Hz GPS could be used in a similar way to seismic data to infer the spatial and temporal progression of fault slip dur-

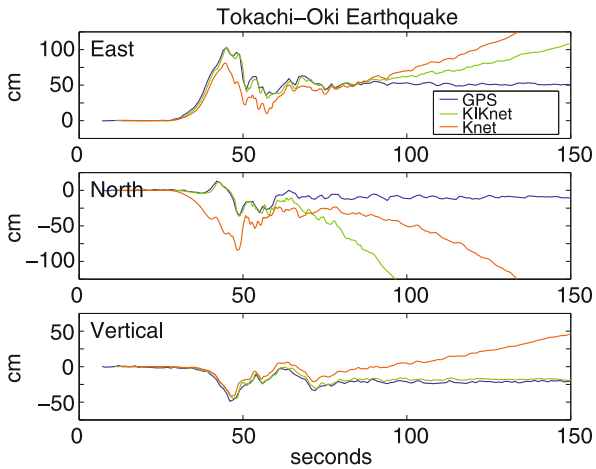

GPS: Applications in Crustal Deformation Monitoring, Figure 16
Comparison of displacement time series from a GPS station and doubly integrated accelerometer data from two nearby stations of the KIKnet and Knet networks in Japan for the 2003 M_w 8 Tokachi-Oki earthquake. The GPS data and that from the KIKnet site (*blue* and *green curves*) agree well for the first ∼50 seconds following the arrival of the seismic signal (at ∼30 seconds). The deviation of the accelerometer data after several seconds is likely due to low frequency noise that is amplified by the double integration required to produce displacements. Additionally, the Knet site (*red curve*) may be affected by tilting during shaking or other problems with sensor orientation, leading to the greater deviation between the *red curve* and those of the other two instruments. Reprinted with permission from [96] (copyright 2006, American Geophysical Union)

ing the earthquake. High-rate GPS observations can be especially useful in this capacity when used in combination with seismic data (e. g. [63]).

High-rate GPS also vastly improves our ability to track early postseismic deformation. Seismic data do not record any processes, like afterslip, that do not generate seismic waves, so postseismic deformation measurements must be made using other instruments including GPS, strainmeters, and InSAR. Until CGPS networks became common, it was generally not possible to deploy instruments for post-earthquake geodetic measurements until field crews could reach the affected area, days or weeks after the event. As a result, any change in measured position from the last pre-earthquake survey until the first post-earthquake survey contained the full coseismic signal and some portion of postseismic displacement. Even the "coseismic" displacements obtained from differencing daily positions from CGPS data collected the day before and the day after the earthquake can be contaminated by postseismic signals that began immediately after the event. High-rate GPS data, on the other hand, provide a means to record the deformation signal continuously, starting in the seconds after an earthquake. An example of this was the 2004 M_w 6 Parkfield earthquake. Langbein et al. [76] showed that rapid postseismic displacement began immediately after the event, and these authors were able to use the high-rate time series to separate the coseismic and postseismic displacements. For other events, including the 1966 M_w ∼6 Parkfield earthquake, the moment release estimated from seismic data has often been substantially lower than that estimated using geodetic data (e. g. [138]). However in the case of the 2004 Parkfield event the slip estimated from the coseismic portion of the GPS displacement signal had moment release in good agreement with that estimated from seismic data [76].

The availability and use of high-rate GPS data in real-time for geophysical applications is not yet widespread, but it is growing as more CGPS sites are installed, telemetry of high bandwidth data becomes more feasible, and approaches to processing the data are refined. These data have great potential for aiding real-time deformation monitoring on volcanoes, for early warning of major earthquakes, and for tsunami warning as well.

For example, Mattia et al. [87] highlight the value of high-rate real-time GPS data in a time-critical situation to quickly differentiate localized volcanic processes (in their case a new volcanic vent opening) from more far-reaching dangers (e. g., a feared flank failure that could have caused a local tsunami).

Real-time high-rate GPS data may soon also contribute to rapid earthquake warning and response. Hudnut et al. [61] envision a system consisting of a pair of GPS receivers straddling a fault such as the San Andreas which record data at high rate and transmit these measurements in real time. In the case of a major earthquake which, within seconds, caused surface offset on the fault of more than a few centimeters, the "GPS slip-sensor" would indicate that the event was large significantly sooner than could be expected from seismic data alone. This rapid information could be used to trigger preventive measures for critical infrastructure such as transportation systems (e. g., slowing trains to help prevent derailment) and factories (e. g. halting processes involving hazardous chemicals).

Blewitt et al. [14], using data from the 2004 Sumatra earthquake, demonstrated that GPS data available in real time, even if recorded at the traditional 30-second sampling rate, can be used to reliably estimate displacements larger than about 10 mm for great earthquakes. Such information would significantly improve the robustness of tsunami warning systems, at least for oceanwide tsunamis, by giving a better indication of the true moment magnitude of the earthquake more quickly (e. g. within 15 minutes) than can be achieved with seismic data.

The patterns of strong shaking that occur during an earthquake provide a good indication as to where damage will be the most severe. The U. S. Geological Survey has implemented a system called ShakeMap which rapidly generates a map of shaking intensity based on instrumental recordings. This is a valuable tool for emergency responders and scientists in the aftermath of an earthquake, and is available to the public as well (http://earthquake.usgs.gov/eqcenter/shakemap/). However, in many areas seismic instrument coverage is sparse, hindering the generation of accurate and detailed maps. Dreger et al. [36] demonstrate that seismic data can be used to rapidly generate a model of fault slip during an earthquake and that such a model can significantly improve the quality of ShakeMap produced for that event. This is especially true when the direction of earthquake rupture away from the hypocenter shows a preferred orientation along strike, since shaking will be stronger at locations that coincide with this directivity. Current work is focused on incorporation of real-time high-rate GPS data into the fault slip analysis with the ultimate goal of further refining ShakeMaps.

Future Directions

As evidenced by the applications described in this article, GPS has become an indispensable tool for monitoring crustal deformation hazards and investigating the underlying processes. It provides an affordable means of obtaining surface displacement measurements with millimeter-level precision at any time of day anywhere on earth in all weather conditions with no line of sight requirement. GPS data complement other observations such as seismic and InSAR data by providing a direct measure of displacement in three dimensions that stays on scale. GPS can record deformation over temporal scales of seconds to decades, thus making it possible to track surface waves generated by an earthquake, postseismic deformation, slow slip events, volcanic unrest, fault creep, interseismic strain accumulation, and plate motions.

For the first decade or so, applications of GPS in crustal deformation tended toward periodic measurements in the style of earlier geodetic surveys. However, the size, power consumption, and cost of receivers have decreased steadily, while the data storage capacity has increased. This has fostered the growth of large CGPS networks, accompanied by more centralized and uniform processing and availability of results. This trend is likely to continue, and many future applications of GPS will focus on further exploiting the ability of this tool to provide temporally dense measurements for tracking deformation.

High-rate data show promise for hazard monitoring and response as well as for providing insight into the physical processes underlying time-varying deformation, especially as more sophisticated methods become available for processing these data in real-time and mitigating error sources such as multipath which are particularly problematic for high rate measurements.

In areas with few CGPS sites or where spatially dense coverage is needed to address specific scientific questions, an alternative to traditional campaign GPS measurements has recently been developed [13]. Termed "semi-permanent" or "semi-continuous" GPS, it involves rotating a pool of GPS receivers through several subsets of GPS sites such that each subset is observed for periods of multiple weeks several times a year. Each site to be observed is outfitted with a specially designed antenna mount set in a rock outcrop so that the antenna is attached in the same location and orientation every time the site is occupied, thus eliminating the set-up error inherent in many SGPS surveys. Blewitt et al. [13] have demonstrated that the scatter in the position time series as well as the uncertainties on velocities estimated using data from a semi-permanent network in Nevada are nearly as low as for CGPS networks over a comparable length of time (e. g., 1.5 years). Although this method requires sites that have rock outcrops and that are secure enough for equipment to be left unattended for extended periods of time, its use is likely to grow as it is arguably more cost-effective than either CGPS or SGPS for many applications.

Finally, a series of changes are being made to modernize the GPS signal. For instance, beginning with a generation of satellites called Block IIR-M, first launched in the fall of 2005, a civilian code is now modulated on the L2 carrier in addition to the C/A code that has always been available on the L1 carrier. This makes it easier to obtain the L2 carrier signal and enables elimination of ionospheric delay when positioning with the code measurements alone. A new series of GPS satellites, the Block IIF, set to launch in 2007 will transmit a third carrier frequency called L5 which will aid in ambiguity resolution. The L5 observable, modulated with a civilian-accessible code, will be broadcast at a higher power than the L1 or L2, making it easier to acquire. Future satellites will transmit a second civilian code on the L1 frequency that will be more robust than the C/A code and will be interoperable with a civilian accessible code to be transmitted by the planned Galileo satellites.

Acknowledgments

John Langbein, Margaret Boettcher, Maurizio Battaglia, Emily Desmarais, Tom Hanks, and Fred Pollitz pro-

vided helpful comments which significantly improved this manuscript.

Bibliography

Primary Literature

1. Ammon C, Ji C, Thio HK, Robinson D, Ni S, Hjorleifsdottir V, Kanamori H, Lay T, Das S, Helmberger D, Ichinose G, Polet J, Wald D (2005) Rupture process of the 2004 Sumatra-Andaman earthquake. Science 308:1133–1139
2. Argus D, Gordon R (1991) No-net-rotation model of current plate velocities incorporating plate motion model NUVEL-1. Geophys Res Lett 18:2039–2042
3. Argus D, Gordon R, Ma C, Eanes R, Heflin M, Owen S, Willis P (2006) GEODVEL: Plate motions from space geodesy. Eos Trans AGU 87(52) Fall Meet Suppl:Abstract G41A-06
4. Argus D, Heflin M (1995) Plate motion and crustal deformation estimated with geodetic data from the Global Positioning System. Geophys Res Lett 22:1973–1976
5. Árnadóttir T, Jónsson S, Pollitz F, Jiang W, Feigl K (2005) Postseismic deformation following the June 2000 earthquake sequence in the south Iceland seismic zone. J Geophys Res 110:B12308; doi:10.1029/2005JB003701
6. Banerjee P, Pollitz F, Bürgmann R (2005) The size and duration of the Sumatra–Andaman earthquake from far-field static offsets. Science 308:1769–1772
7. Banerjee P, Pollitz F, Nagarajan B, Bürgmann R (2007) Coseismic slip distributions of the 26 December 2004 Sumatra–Andaman and 28 March 2005 Nias earthquakes from GPS static offsets. Bull Seismol Soc Amer 97:S86–S102
8. Battaglia M, Segall P, Murray J, Cervelli P, Langbein J (2003) The mechanics of unrest at Long Valley caldera, California: 1. Modeling the geometry of the source using GPS, leveling and two-color EDM data. J Volc Geotherm Res 127:195–217
9. Battaglia M, Segall P, Roberts C (2003) The mechanics of unrest at Long Valley caldera, California: 2. Constraining the nature of the source using geodetic and micro-gravity data. J Volc Geotherm Res 127:219–245
10. Bennett R, Davis J, Wernicke B (1996) First results from the northern Basin and Range continuous GPS network. Eos Trans AGU 77(46) Fall Meet Suppl:150
11. Bilek S, Satake K, Sieh K (2007) Introduction to the special issue on the 2004 Sumatra–Andaman earthquake and the Indian Ocean tsunami. Bull Seis Soc Amer 97:S1–S5
12. Bilham R, Engdahl R, Feldl N, Satyabala S (2005) Partial and complete rupture of the Indo-Andaman plate boundary 1847–2004. Seismol Res Lett 76:299–311
13. Blewitt G, Hammond WC, Kreemer C (2009) Geodetic observation of contemporary strain in the northern Walker Lane: 1, Semi-permanent GPS strategy. In: Oldow JS, Cashman PH (eds) Late Cenozoic Structure and Evolution of the Great Basin – Sierra Nevada Transition. Geol Soc Amer (in press) doi:10.1130/2009.2447(1)
14. Blewitt G, Kreemer C, Hammond W, Plag HP, Stein S, Okal E (2006) Rapid determination of earthquake magnitude using GPS for tsunami warning systems. Geophys Res Lett 33:L11309; doi:10.1029/2006GL026145
15. Bock Y, Nikolaidis R, de Jonge P, Bevis M (2000) Instantaneous geodetic positioning at medium distances with the Global Positioning System. J Geophys Res 105:28223–28253
16. Bos A, Usai S, Spakman W (2004) A joint analysis of GPS motions and InSAR to infer the coseismic surface deformation of the Izmit, Turkey earthquake. Geophys J Int 158:849–863
17. Bürgmann R, Kogan M, Steblov G, Hilley G, Levin V, Apel E (2005) Interseismic coupling and asperity distribution along the Kamchatka subduction zone. J Geophys Res 110:B07405; doi:10.1029/2005JB003648
18. Bürgmann R, Rosen P, Fielding E (2000) Synthetic aperture radar interferometry to measure Earth's surface topography and its deformation. Ann Rev Earth Planet Sci 28:169–209
19. Calais E, Han JY, DeMets C, Nocquet JM (2006) Deformation of the North American plate interior from a decade of continuous GPS measurements. J Geophys Res 111:B06402; doi:10.1029/2005JB004253
20. Cervelli P, Fournier T, Freymueller J, Power J (2006) Ground deformation associated with the precursory unrest and early phases of the January 2006 eruption of Augustine Volcano, Alaska. Geophys Res Lett 33:L18304; doi:10.1029/2006GL027219
21. Cervelli P, Murray M, Segall P, Aoki Y, Kato T (2001) Estimating source parameters from deformation data, with an application to the March 1997 earthquake swarm off the Izu Peninsula, Japan. J Geophys Res 106:11217–11237
22. Cervelli P, Segall P, Johnson K, Lisowski M, Miklius A (2002) Sudden aseismic fault slip on the south flank of Kilauea volcano. Nature 415:1014–1018
23. Chen Q, Freymueller J, Wang Q, Yang Z, Xu C, Liu J (2004) A deforming block model for the present-day tectonics of Tibet. J Geophys Res 109:B01403; doi:10.1029/2002JB002151
24. Chlieh M, Avouac JP, Hjorleifsdottir V, Song TR, Ji C, Sieh K, Sladen A, Hebert H, Prawirodirdjo L, Bock Y, Galetzka J (2007) Coseismic slip and afterslip of the great M_w 9.15 Sumatra–Andaman earthquake of 2004. Bull Seis Soc Amer 97:S152–S173
25. Coe JA, Ellis WL, Godt JW, Savage WZ, Savage JE, Michael JA, Kibler JD, Powers PS, Lidke DJ, Debray S (2003) Seasonal movement of the Slumgullion landslide determined from Global Positioning System surveys and field instrumentation, July 1998 – March 2002. Eng Geol 68:67–101
26. d'Alessio M, Johanson I, Bürgmann R, Schmidt D, Murray M (2005) Slicing up the San Francisco Bay Area: Block kinematics and fault slip rates from GPS-derived surface velocities. J Geophys Res 110:B06403; doi:10.1029/2004JB003496
27. Delouis B, Giardini D, Lundgren P, Salichon J (2002) Joint inversion of InSAR, GPS, teleseismic, and strong-motion data for the spatial and temporal distribution of earthquake slip; application to the 1999 Izmit mainshock. Bull Seis Soc Amer 92:278–299
28. DeMets C, Gordon R, Argus D (2006) Moving beyond NUVEL-1A: The MORVEL estimates of geologically recent global plate motions. Eos Trans AGU 87(52) Fall Meet Suppl:Abstract G41A-05
29. DeMets C, Gordon R, Argus D, Stein S (1990) Current plate motions. Geophys J Int 101:425–478
30. DeMets C, Gordon R, Argus D, Stein S (1994) Effect of recent revisions to the geomagnetic reversal time scale on estimates of current plate motions. Geophys Res Lett 21:2191–2194
31. Deng J, Gurnis M, Kanamori H, Hauksson E (1998) Viscoelastic flow in the lower crust after the 1992 Landers, California, earthquake. Science 282:1689–1692

32. Dieterich J, Decker R (1975) Finite element modeling of surface deformation associated with volcanism. J Geophys Res 80:4094–4102
33. Dong D, Fang P, Bock Y, Cheng MK, Miyazaki S (2002) Anatomy of apparent seasonal variations from GPS-derived site position time series. J Geophys Res 107:2075; doi:10.1029/2001JB000573
34. Douglas A, Beavan J, Wallace L, Townend J (2005) Slow slip on the northern Hikurangi subduction interface, New Zealand. Geophys Res Lett 32:L16305; doi:10.1029/2005GL023607
35. Dragert H, Wang K, James TS (2001) A silent slip event on the deeper Cascadia subduction interface. Science 292:1525–1528
36. Dreger D, Gee L, Lombard P, Murray M, Romanowicz B (2005) Rapid finite-source analysis and near-fault strong ground motions: Application to the 2003 M_w 6.5 San Simeon and 2004 M_w 6.0 Parkfield earthquakes. Seismol Res Lett 76:40–48
37. Du Y, Segall P, Gao H (1997) Quasi-static dislocations in three dimensional inhomogeneous media. Geophys Res Lett 24:2347–2350
38. Dzurisin D (1992) Geodetic leveling as a tool for studying restless volcanoes. In: Ewert J, Swanson D (eds) Monitoring volcanoes: Techniques and strategies used by the staff of the Cascades Volcano Observatory, 1980-1990, USGS Bull. 1966, US Geological Survey, Reston, VA, pp 125–134
39. Dzurisin D (2003) A comprehensive approach to monitoring volcano deformation as a window on the eruption cycle. Rev Geophys 41; doi:10.1029/2001RG000107
40. Dzurisin D (2007) Volcano deformation: Geodetic monitoring techniques. Springer, New York
41. Eberhart-Phillips D, Michael AJ (1993) Three-dimensional velocity structure, seismicity, and fault structure in the Parkfield region, central California. J Geophys Res 98:15737–15758
42. Emore G, Haase J, Choi K, Larson K, Yamagiwa A (2007) Recovering seismic displacements through combined use of 1-Hz GPS and strong-motion accelerometers. Bull Seismol Soc Amer 97:357–378; doi:10.1785/0120060153
43. Fialko Y (2004) Evidence of fluid-filled upper crust from observations of postseismic deformation due to the 1992 M_w 7.3 Landers earthquake. J Geophys Res 109:B08401; doi:10.1029/2004JB002985
44. Fletcher H, Beavan J, Freymueller J, Gilbert L (2001) High interseismic coupling of the Alaska subduction zone SW of Kodiak island inferred from GPS data. Geophys Res Lett 28:443–446
45. Freed A, Bürgmann R (2004) Evidence of power-law flow in the Mojave desert mantle. Nature 430:548–551
46. Freed A, Bürgmann R, Calais E, Freymueller J, Hreinsdóttir S (2006) Implications of deformation following the 2002 Denali, Alaska, earthquake for postseismic relaxation processes and lithospheric rheology. J Geophys Res 111:B01401; doi:10.1029/2005JB003894
47. Freymueller J, Murray M, Segall P, Castillo D (1999) Kinematics of the Pacific-North America plate boundary zone, northern California. J Geophys Res 104:7419–7441
48. Fujinawa Y, Shimada S, Ohmi S, Sekiguchi S, Eguchi T, Okada Y (1991) Fixed point GPS observation of crustal movement associated with the 1989 seismic swarm and submarine volcanic activities of Ito, central Japan. J Phys Earth 39:141–153
49. Gahalaut VK, Nagarajan B, Catherine JK, Kumar S (2006) Constraints on 2004 Sumatra–Andaman earthquake rupture from GPS measurements in Andaman-Nicobar Islands. Earth Planet Sci Lett 242:365–374
50. Gili JA, Corominas J, Rius J (2000) Using Global Positioning System techniques in landslide monitoring. Engineering Geology 55:167–192
51. Gladwin M, Gwyther R, Hart R, Breckenridge K (1994) Measurements of the strain field associated with episodic creep events on the San Andreas fault near San Juan Bautista, California. J Geophys Res 99:4559–4565
52. Gwyther RL, Gladwin MT, Mee M, Hart RHG (1996) Anomalous shear strain at Parkfield during 1993–94. Geophys Res Lett 23:2425–2428
53. Hashimoto M, Hashizume M, Takemoto S, Fukada Y, Fujimori K, Takiguchi H, Satomura M, Otsuka Y, Saito S (2006) Postseismic deformations following the Sumatra–Andaman and Nias earthquakes detected by continuous GPS observation in SE Asia. Seism Res Lett 77:289
54. Hernandez B, Cotton F, Campillo M (1999) Contribution of radar interferometry to a two-step inversion of the kinematic process of the 1992 Landers earthquake. J Geophys Res 104:13083–13099
55. Hirose H, Obara K (2006) Short-term slow slip and correlated tremor episodes in the Tokai region, central Japan. Geophys Res Lett 33:L17311; doi:10.1029/2006GL026579
56. Hofmann-Wellenhof B, Lichtenegger H, Collins J (2001) Global Positioning System theory and practice, 5th edn. Springer, New York
57. Hreinsdóttir S, Freymueller JT, Bürgmann R, Mitchell J (2006) Coseismic deformation of the 2002 Denali Fault earthquake: Insights from GPS measurements. J Geophys Res 111:B03308; doi:10.1029/2005JB003676
58. Hreinsdóttir S, Freymueller JT, Fletcher HJ, Larsen CF, Bürgmann R (2003) Coseismic slip distribution of the 2002 M_w 7.9 Denali Fault earthquake, Alaska, determined from GPS measurements. Geophys Res Lett 30:1670; doi:10.1029/2003GL017447
59. Hsu Y, Simons M, Avouac J-P, Galetzka J, Sieh K, Chlieh M, Natawidjaja D, Prawirodirdjo L, Bock Y (2006) Frictional afterslip following the 2005 Nias-Simeulue earthquake, Sumatra. Science 312:1921–1926
60. Hudnut K (1997) The Southern California Integrated GPS Network (SCIGN). Open-File Report, US Geological Survey, Report: OF 97-0467:10-13
61. Hudnut K, Anderson G, Aspiotes A, King N, Moffitt R, Stark K (2002) GPS fault slip sensors. APEC Symposium on Confronting Urban Earthquakes/Seismic Early Warning. Academia Sinica, Taipei, pp 93–96
62. Ihmlé PF, Jordan TH (1994) Teleseismic search for slow precursors to large earthquakes. Science 266:1547–1551
63. Ji C, Larson K, Tan Y, Hudnut K, Choi K (2004) Slip history of the 2003 San Simeon earthquake constrained by combining 1-Hz GPS, strong motion, and teleseismic data. Geophys Res Lett 31:L17608; doi:10.1029/2004GL020448
64. Johanson I, Bürgmann R (2005) Creep and quakes on the northern transition zone of the San Andreas fault from GPS and InSAR data. J Geophys Res 32:L14306; doi:10.1029/2005GL023150
65. Johnson K, Segall P (2004) Imaging the ramp-décollement geometry of the Chelungpu fault using coseismic GPS displacements from the 1999 Chi-Chi, Taiwan earthquake. Tectonophysics 378:123–139

66. Johnson K, Segall P (2004) Viscoelastic cycle models of deep stress driven creep along the San Andreas Fault. J Geophys Res 109; doi:10.1029/2004JB003096
67. Johnson K, Segall P (2005) A viscoelastic earthquake cycle model for Taiwan. J Geophys Res 110:B10404; doi:10.1029/2004JB003516
68. Johnston M, Linde A (2002) Implications of crustal strain during convetional, slow, and silent earthquakes. Handbook of Earthquake and Engineering Seismology 81A:589–605
69. Jónsson S, Segall P, Pedersen R, Björnsson G (2003) Post-earthquake ground movements correlated to pore-pressure transients. Nature 424:179–183
70. Jónsson S, Zebker H, Segall P, Amelung F (2002) Fault Slip Distribution of the 1999 M_w 7.1 Hector Mine, California, Earthquake, estimated from Satellite Radar and GPS Measurements. Bull Seis Soc Amer 92:1377–1389
71. Kaverina A, Dreger D, Price E (2002) The combined inversion of seismic and geodetic data for the source process of the 16 October 1999 M_w 7.1 Hector Mine, California, earthquake. Bull Seis Soc Amer 92:1266–1280
72. King N, Murray M, Prescott W, Clymer R, Romanowicz B (1994) The Bay Area Regional Deformation (BARD) permanent GPS array. Eos Trans AGU 75(44) Fall Meet Suppl:470
73. Kreemer C, Blewitt G, Hammond W, Plag HP (2006) Global deformation from the great 2004 Sumatra–Andaman earthquake observed by GPS: Implications for rupture process and global reference frame. Earth Planets Space 58:141–148
74. Langbein J (2004) Noise in two-color electronic distance meter measurements revisited. J Geophys Res 109:B04406; doi:10.1029/2003JB002819
75. Langbein J, Gwyther RL, Hart RHG, Gladwin MT (1999) Slip-rate increase at Parkfield in 1993 detected by high-precision EDM and borehole tensor strainmeters. Geophys Res Lett 26:2529–2532
76. Langbein J, Murray J, Snyder HA (2006) Coseismic and initial postseismic deformation from the 2004 Parkfield, California, earthquake, observed by Global Positioning System, electronic distance meter, creepmeters, and borehole strainmeters. Bull Seismol Soc Amer 96:S304–S320; doi:10.1785/0120050823
77. Larson K (1995) Crustal deformation. Rev Geophys 33:371–378; doi:10.1029/95RG00439
78. Larson K, Bodin P, Gomberg J (2003) Using 1-Hz GPS data to measure deformations caused by the Denali fault earthquake. Science 300:1421–1424; doi:10.1126/science.1084531
79. Larson K, Freymueller J, Philipsen S (1997) Global plate velocities from the Global Positioning System. J Geophys Res 102:9961–9981
80. Larson K, van Dam T (2000) Measuring postglacial rebound with GPS and absolute gravity. Geophys Res Lett 27:3925–3928
81. Lay T, Kanamori H, Ammon C, Nettles M, Ward S, Aster R, Beck S, Bilek S, Brudzinski M, Butler R, DeShon H, Ekstrom G, Satake K, Sipkin S (2005) The great Sumatra–Andaman earthquake of 26 December 2004. Science 308:1127–1133
82. Linde A, Gladwin M, Johnston M, Gwyther R, Bilham R (1996) A slow earthquake sequence on the San Andreas fault. Nature 383:65–68
83. Lowry A (2006) Resonant slow fault slip in subduction zones forced by climatic load stress. Nature, 442:802–805
84. Maerten F, Resor P, Pollard D, Maerten L (2005) Inverting for slip on three-dimensional fault surfaces using angular dislocations. Bull Seismol Soc Amer 95:1654–1665
85. Manaker D, Bürgmann R, Prescott W, Langbein J (2003) Distribution of interseismic slip rates and the potential for significant earthquakes on the Calaveras fault, central California. J Geophys Res 108:B62287; doi:10.1029/2002JB001749
86. Mao A, Harrison C, Dixon T (1999) Noise in GPS coordinate time series. J Geophys Res 104:2797–2816
87. Mattia M, Rossi M, Guglielmino F, Aloisi M, Bock Y (2004) The shallow plumbing system of Stromboli Island as imaged from 1 Hz instantaneous GPS positions. Geophys Res Lett 31:L24610; doi:10.1029/2004GL021281
88. Mazzotti S, James TS, Henton J, Adams J (2005) GPS crustal strain, postglacial rebound, and seismic hazard in eastern North America: The Saint Lawrence valley example. J Geophys Res 110:B11301; doi:10.1029/2004JB003590
89. McCaffrey R (2005) Block kinematics of the Pacific-North America plate boundary in southwestern United States from inversion of GPS, seismological, and geologic data. J Geophys Res 110:B07401; doi:10.1029/2004JB003307
90. Meade B, Hager B (2005) Block models of crustal motion in southern California constrained by GPS measurements. J Geophys Res 110:B03403; doi:10.1029/2004JB003209
91. Meltzner A, Sieh K, Abrams M, Agnew D, Hudnut K, Avouac JP, Natawidjaja DH (2006) Uplift and subsidence associated with the great Aceh-Andaman earthquake of 2004. J Geophys Res 111:B02407; doi:10.1029/2005JB003891
92. Miller M, Johnson D, Rubin C, Dragert H, Endo E, Humphreys E, Nabelek J, Qamar A (1997) GPS Monitoring of the Cascadia Margin: The Pacific Northwest Geodetic Array (PANGA). Eos Trans AGU 78(46) Fall Meet Suppl:167
93. Miller M, Johnson D, Rubin C, Dragert H, Wang K, Qamar A, Goldfinger C (2001) GPS-determination of along-strike variation in Cascadia margin kinematics: Implications for relative plate motion, subduction zone coupling, and permanent deformation. Tectonics 20:161–176
94. Miller M, Melbourne T, Johnson D, Sumner W (2002) Periodic slow earthquakes from the Cascadia subduction zone. Science 295:2423
95. Milne GA, Davis JL, Mitrovica JX, Scherneck HG, Johansson JM, Vermeer M, Koivula H (2001) Space-geodetic constraints on glacial isostatic adjustment in Fennoscandia. Science 291:2381–2385
96. Miyazaki S, Larson K, Choi K, Hikima K, Koketsu K, Bodin P, Haase J, Emore G, Yamagiwa A (2004) Modeling the rupture process of the 2003 September 25 Tokachi-Oki (Hokkaido) earthquake using 1-Hz GPS data. Geophys Res Lett 31:L21603; doi:10.1029/2004GL021457
97. Miyazaki S, Segall P, Fukuda J, Kato T (2004) Space time distribution of afterslip following the 2003 Tokachi-oki earthquake: Implications for variations in fault zone frictional properties. Geophys Res Lett 31:L06623; doi:10.1029/2003GL019410
98. Miyazaki S, Segall P, McGuire J, Kato T, Hatanaka Y (2006) Spatial and temporal evolution of stress and slip rate during the 2000 Tokai slow earthquake. J Geophys Res 111:B03409; doi:10.1029/2004JB003426
99. Mogi K (1958) Relations between the eruptions of various volcanoes and the deformations of the ground surfaces around them. Bull Seismol Soc Amer 36:111–123
100. Mora P, Baldi P, Casula G, Fabris M, Ghirotti M, Mazzini E, Pesci

A (2003) Global Positioning Systems and digital photogrammetry for the monitoring of mass movements: Application to the Ca' di Malta landslide (northern Appenines, Italy). Eng Geol 68:103–121
101. Murray J, Langbein J (2006) Slip on the San Andreas fault at Parkfield, California, over two earthquake cycles, and the implications for seismic hazard. Bull Seismol Soc Amer 96:S283–S303
102. Murray J, Segall P (2005) Spatiotemporal evolution of a slip-rate increase on the San Andreas fault near Parkfield, CA. J Geophys Res 110:B09407; doi:10.1029/2005JB003651
103. Murray J, Segall P, Cervelli P, Prescott W, Svarc J (2001) Inversion of GPS data for spatially variable slip-rate on the San Andreas Fault near Parkfield, CA. Geophys Res Lett 28:359–362
104. Nikolaidis R, Bock Y, de Jonge P, Shearer P, Agnew D, Domselaar M (2001) Seismic wave observations with the Global Positioning System. J Geophys Res 106:21897–21916
105. Obara K (2002) Nonvolcanic deep tremor associated with subduction in southwest Japan. Science 296:1679–1681
106. Obara K, Hirose H, Yamamizu F, Kasahara K (2004) Episodic slow slip events accompanied by non-volcanic tremors in southwest Japan subduction zone. Geophys Res Lett 31; doi:10.1029/2004GL020848
107. Ohta Y, Freymueller J, Hreinsdóttir S, Suito H (2006) A large slow slip event and the depth of the seismogenic zone in the south central Alaska subduction zone. Earth Plan Sci Lett 247:108–116
108. Ohta Y, Kimata F, Sagiya T (2004) Reexamination of the interplate coupling in the Tokai region, central Japan, based on the GPS data in 1997-2002. Geophys Res Lett 31:L24604; doi:10.1029/2004GL021404
109. Okada Y (1985) Surface deformation due to shear and tensile faults in a half-space. Bull Seismol Soc Amer 75:1135–1154
110. Okada Y (1992) Internal deformation due to shear and tensile faults in a half-space. Bull Seismol Soc Amer 82:1018–1040
111. Owen S, Segall P, Lisowski M, Miklius A, Denlinger R, Sako M (2000) Rapid deformation of Kilauea volcano: GPS measurements between 1990 and 1996. J Geophys Res 105:18983–18998
112. Owen S, Segall P, Lisowski M, Murray M, Bevis M, Foster J (2000) The January 30, 1997 eruptive event on Kilauea Volcano, Hawaii, as monitored by continuous GPS. Geophys Res Lett 27:2757–2760
113. Ozawa S, Murakami M, Fujiwara S, Tobita M (1997) Synthetic aperture radar interferogram of the 1995 Kobe earthquake and its geodetic inversion. Geophys Res Lett 24:2327–2330
114. Ozawa S, Suito H, Imakiire T, Murakmi M (2007) Spatiotemporal evolution of aseismic interplate slip between 1996 and 1998 and between 2002 and 2004, in Bungo channel, southwest Japan. J Geophys Res 112:B05409; doi:10.1029/2006JB004643
115. Park J, Anderson K, Aster R, Butler R, Lay T, Simpson D (2005) Global seismographic network records the great Sumatra–Andaman earthquake. Eos Trans AGU 86:57, 60–61
116. Park J, Song TR, Tromp J, Okal E, Stein S, Roult G, Clevede E, Laske G, Kanamori H, Davis P, Berger J, Braitenberg C, Van Camp M, Lei X, Sun H, Xu H, Rosat S (2005) Earth's free oscillations excited by the 26 December 2004 Sumatra–Andaman earthquake. Science 308:1139–1144
117. Park K, Nerem RS, Davis JL, Schenewerk MS, Milne GA, Mitrovica JX (2002) Investigation of glacial isostatic adjustment in the northeast US using GPS measurements. Geophys Res Lett 29:1509; doi:10.1029/2001GL013782
118. Peltzer G, Rosen P, Rogez F, Hudnut K (1998) Poroelastic rebound along the Landers 1992 earthquake surface rupture. J Geophys Res 103:30131–30145
119. Poland M, Hamburger M, Newman A (2006) The changing shapes of active volcanoes: History, evolution, and future challenges for volcano geodesy. J Volc Geotherm Res 150:1–13
120. Polet J, Kanamori H (2000) Shallow subduction zone earthquakes and their tsunamigenic potential. Geophys J Int 142:684–702
121. Pollitz F (2005) Transient rheology of the upper mantle beneath central Alaska inferred from the crustal velocity field following the 2002 Denali earthquake. J Geophys Res 110:B08407; doi:10.1029/2005JB003672
122. Pollitz F, Wicks C, Thatcher W (2001) Mantle flow beneath a continental strike-slip fault: postseismic deformation after the 1999 Hector Mine earthquake. Science 293:1814–1818
123. Pratt T (2006) Do Episodic Tremor and Slip (ETS) Events Affect Seismicity in the Northern Cascadia Subduction Zone? Eos Trans AGU 87(52), Fall Meet Suppl, Abstract T54A-04
124. Prawirodirdjo L, Bock Y (2004) Instantaneous global plate motion model from 12 years of continuous GPS observations. J Geophys Res 109:B08405; doi:10.1029/2003JB002944
125. Prescott W, Savage J, Svarc J, Manaker D (2001) Deformation across the Pacific-North America plate boundary near San Francisco, California. J Geophys Res 106:6673–6682
126. Pritchard M, Norabuena E, Ji C, Boroschek R, Comte D, Simons M, Dixon T, Rosen P (2007) Geodetic, teleseismic, and strong motion constraints on slip from recent southern Peru subduction zone earthquakes. J Geophys Res 112:B03307; doi:10.1029/2006JB004294
127. Reid HF (1910) The California Earthquake of April 18, 1906. In: Report of the state earthquake investigation commission, vol 2, Carnegie Institute, Washington DC
128. Reid M, LaHusen R, Schmidt K (2004) Capturing 3-D displacements in active landslides using GPS. Abstracts with Programs. Geol Soc Amer 36:331
129. Rhie J, Dreger D, Bürgmann R, Romanowicz B (2007) Slip of the 2004 Sumatra–Andaman earthquake from joint inversion of long-period global seismic waveforms and GPS static offsets. Bull Seismol Soc Amer 97:S115–S127
130. Rogers G, Dragert H (2003) Episodic tremor and slip on the Cascadia subduction zone: the chatter of silent slip. Science 300:1942–1943
131. Sagiya T (2004) Interplate coupling in the Kanto district, central Japan, and the Boso Peninsula silent earthquake in May 1996. Pure Appl Geophys 161:2327–2342; doi:10.1007/s00024-004-2566-6
132. Salichon J, Lundgren P, Delouis B, Giardini D (2004) Slip history of the 16 October 1999 M_w 7.1 Hector Mine earthquake (California) from the inversion of InSAR, GPS, and teleseismic data. Bull Seismol Soc Amer 94:2015–2027
133. Savage J, Prescott W (1978) Asthenosphere readjustment and the earthquake cycle. J Geophys Res 83:3369–3376
134. Savage J, Svarc J, Yu SB (2005) Postseismic relaxation and transient creep. J Geophys Res 110:B11402; doi:10.1029/2005JB003687
135. Schmalzle G, Dixon T, Malservisi R, Govers R (2006) Strain accumulation across the Carrizo segment of the San Andreas

fault, California: Impact of laterally varying crustal properties. J Geophys Res 111:B05403; doi:10.1029/2005JB003843
136. Schmidt D, Bürgmann R, Nadeau R, d'Alessio M (2005) Distribution of aseismic slip rate on the Hayward fault inferred from seismic and geodetic data. J Geophys Res 110:B08406; doi:10.1029/2004JB003397
137. Segall P (2002) Integrating geologic and geodetic estimates of slip rate on the San Andreas Fault system. Int Geol Rev 44:62–82
138. Segall P, Harris R (1986) Slip deficit on the San Andreas fault at Parkfield, California, as revealed by inversion of geodetic data. Science 233:1409–1413
139. Segall P, Desmarais E, Shelly D, Miklius A, Cervelli P (2006) Earthquakes triggered by silent slip events on Kilauea volcano, Hawaii. Nature 442:71–74; doi:10.1038/nature04938
140. Sella G, Dixon T, Mao A (2002) REVEL: A model for recent plate velocities from space geodesy. J Geophys Res 107:B42081; doi:10.1029/2000JB000033
141. Sella G, Stein S, Dixon T, Craymer M, James T, Mazzotti S, Dokka R (2007) Observation of glacial isostatic adjustment in "stable" North America with GPS. Geophys Res Lett 34: L02306; doi:10.1029/2006GL027081
142. Shelly D, Beroza G, Ide S (2007) Non-volcanic tremor and low-frequency earthquake swarms. Nature 446:305–307
143. Shelly D, Beroza G, Ide S, Nakamula S (2006) Low-frequency earthquakes in Shikoku, Japan, and their relationship to episodic tremor and slip. Nature 442:188–191
144. Shimada S, Bock Y (1992) Crustal deformation measurements in central Japan determined by a Global Positioning System fixed point network. J Geophys Res 97:12437–12455
145. Shimada S, Fujinawa Y, Sekiguchi S, Ohmi S, Eguchi T, Okada Y (1990) Detection of a volcanic fracture in Japan using Global Positioning System measurements. Nature 343:631–633
146. Sieh K, Jahns RH (1984) Holocene activity of the San Andreas fault at Wallace Creek, California. Geol Soc Amer Bull 95:883–896
147. Silver P, Bock Y, Agnew D, Henyey T, Linde A, McEvilly T, Minster JB, Romanowicz B, Sachs I, Smith R, Solomon S, Stein S (1999) A Plate Boundary Observatory. IRIS Newsletter XVI:3–9
148. Simons M, Fialko Y, Rivera L (2002) Coseismic deformation from the 1999 M_w 7.1 Hector Mine, California, earthquake as inferred from InSAR and GPS observations. Bull Seismol Soc Amer 92:1390–1402
149. Sims JD (1990) Geologic map of the San Andreas fault in the Parkfield 7.5-minute quadrangle, Monterey and Fresno counties, California. US Geol Surv Misc Field Studies Map MF-2115
150. Snay R, Cline M, Dillinger W, Foote R, Hilla S, Kass W, Ray J, Rohde J, Sella G, Soler T (2007) Using global positioning system-derived crustal velocities to estimate rates of absolute sea level change from North American tide gauge records. J Geophys Res 112:B04409; doi:10.1029/2006JB004606
151. Squarzoni C, Delacourt C, Allemand P (2005) Differential single-frequency GPS monitoring of the La Valette landslide (French Alps). Eng Geol 79:215–229
152. Stein S, Okal E (2005) Speed and size of the Sumatra earthquake. Nature 434:581–582
153. Subarya C, Chlieh M, Prawirodirdjo L, Avouac JP, Bock Y, Sieh K, Meltzner AJ, Natawidjaja DH, McCaffrey R (2006) Plate-boundary deformation associated with the great Sumatra–Andaman earthquake. Nature 440:46–51
154. Suwa Y, Miura S, Hasegawa A, Sato T, Tachibana K (2006) Interplate coupling beneath NE Japan inferred from three-dimensional displacement field. J Geophys Res 111:B04402; doi:10.1029/2004JB003203
155. Szeliga W, Melbourne T, Miller M, Santillan V (2004) Southern Cascadia episodic slow earthquakes. Geophys Res Lett 31:L16602; doi:10.1029/2004GL020824
156. Thatcher W (1983) Nonlinear strain buildup and the earthquake cycle on the San Andreas fault. J Geophys Res 88:5893–5902
157. Thatcher W (1995) Microplate versus continuum descriptions of active tectonic deformation. J Geophys Res 100:3885–3894
158. Thatcher W (2007) Microplate model for the present-day deformation of Tibet. J Geophys Res 112:B01401; doi:10.1029/2005JB004244
159. Thurber C, Zhang H, Waldhauser F, Hardebeck J, Michael A, Eberhart-Phillips D (2006) Three-dimensional compressional wavespeed model, earthquake relocations, and focal mechanisms for the Parkfield, California, region. Bull Seism Soc Amer 96:S38–S49
160. Vigny C, Simons WJF, Abu S, Bamphenyu R, Satirapod C, Choosakul N, Subarya C, Socquet A, Omar K, Abidin HZ, Ambrosius BAC (2005) Insight into the 2004 Sumatra–Andaman earthquake from GPS measurements in southeast Asia. Nature 436:201–206
161. Wahba G (1990) Spline Models for Observational Data. SIAM, Philadelphia PA
162. Wald D, Heaton T (1994) Spatial and temporal distribution of slip for the 1992 Landers, California, earthquake. Bull Seismol Soc Amer 84:668–691
163. Waldhauser F, Ellsworth W (2000) A double-difference earthquake location algorithm: Method and application to the northern Hayward fault, California. Bull Seismol Soc Amer 90:1353–1368
164. Wallace L, Beavan J, McCaffrey R, Darby D (2004) Subduction zone coupling and tectonic block rotations in the North Island, New Zealand. J Geophys Res 109:B12406; doi:10.1029/2004JB003241
165. Wang GQ, Boore D, Tang G, Zhou X (2007) Comparisons of ground motions from collocated and closely spaced one-sample-per-second Global Positioning System and accelerograph recordings of the 2003 M 6.5 San Simeon, California, earthquake in the Parkfield region. Bull Seismol Soc Amer 97:76–90
166. Williams S, Bock Y, Fang P, Jamason P, Nikolaidis R, Prawirodirdjo L, Miller M, Johnson D (2004) Error analysis of continuous GPS position time series. J Geophys Res 109. doi:10.1029/2003JB002741
167. Working Group on California Earthquake Probabilities (2003) Earthquake probabilities in the San Francisco Bay Region: 2002-2031. US Geol Surv Open File Report 03-214
168. Wright T, Lu Z, Wicks C (2004) Constraining the slip distribution and fault geometry of the M_w 7.9, 3 November 2002, Denali Fault earthquake with interferometric synthetic aperture radar and Global Positioning System data. Bull Seismol Soc Amer 94:175–189
169. Yagi Y, Kikuchi M, Nishimura T (2003) Co-seismic slip, post-seismic slip, and largest aftershock associated with the 1994 Sanriku-haruka-oki, Japan, earthquake. Geophys Res Lett 30:2177; doi:10.1029/2003GL018189
170. Zhang J, Bock Y, Johnson H, Fang P, Williams S, Genrich J, Wdowinski S, Behr J (1997) Southern California permanent

GPS geodetic array: Error analysis of daily position estimates and site velocities. J Geophys Res 102:18035–18055
171. Zhang PZ, Shen Z, Wang M, Gan W, Bürgmann R, Molnar P, Wang Q, Niu Z, Sun J, Wu J, Hanrong S, Xinzhao Y (2004) Continuous deformation of the Tibetan Plateau from global positioning system data. Geol 32:809–812

Books and Reviews

Bolt B (1999) Earthquakes, 4th edn. W H Freeman and Company, New York
Menke W (1989) Geophysical data analysis: Discrete inverse theory, rev edn. In: Dmowska R, Holton J (eds) International geophysics series 45. Academic Press, San Diego
Misra P, Enge P (2001) Global Positioning System: Signals, measurements, and performance. Ganga-Jamuna Press, Lincoln, MA
Schwartz S, Rokosky J (2007) Slow slip events and seismic tremor at circum-pacific subduction zones. Rev Geophys 45:RG3004
Shearer P (1999) Introduction to seismology. Cambridge University Press, New York
Strang G, Borre K (1997) Linear algebra, geodesy, and GPS. Wellesley-Cambridge Press, Wellesley, MA

Granular Computing and Data Mining for Ordered Data: The Dominance-Based Rough Set Approach

Salvatore Greco[1], Benedetto Matarazzo[1], Roman Słowiński[2,3]
[1] Faculty of Economics, University of Catania, Catania, Italy
[2] Poznań University of Technology, Institute of Computing Science, Poznan, Poland
[3] Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Article Outline

Glossary
Definition of the Subject
Introduction: Granular Computing and Ordered Data
Philosophical Basis of DRSA Granular Computing
Dominance-Based Rough Set Approach
Fuzzy Set Extensions of the Dominance-Based Rough Set Approach
Variable-Consistency Dominance-Based Rough Set Approach (VC-DRSA)
Dominance-Based Rough Approximation of a Fuzzy Set
Monotonic Rough Approximation of a Fuzzy Set Versus Classical Rough Set
Dominance-Based Rough Set Approach to Case-Based Reasoning
An Algebraic Structure for Dominance-Based Rough Set Approach
Conclusions
Future Directions
Bibliography

Glossary

Case-based reasoning Case-based reasoning is a paradigm in machine learning whose idea is that a new problem can be solved by noticing its similarity to a set of problems previously solved. Case-based reasoning regards the inference of some proper conclusions related to a new situation by the analysis of similar cases from a memory of previous cases. Very often similarity between two objects is expressed on a graded scale and this justifies application of fuzzy sets in this context. Fuzzy case-based reasoning is a popular approach in this domain.

Decision rule Decision rule is a logical statement of the type "*if…, then…*", where the premise (condition part) specifies values assumed by one or more condition attributes and the conclusion (decision part) specifies an overall judgment.

Dominance-based rough set approach (DRSA) DRSA permits approximation of a set in universe U based on available ordinal information about objects of U. Also the decision rules induced within DRSA are based on ordinal properties of the elementary conditions in the premise and in the conclusion, such as "*if property f_{i1} is present in degree at least α_{i1} and … property f_{ip} is present in degree at least α_{ip}, then property f_{iq} is present in degree at least α_{iq}*".

Fuzzy sets Differently from ordinary sets in which an object belongs or does not belong to a given set, in a fuzzy set an object belongs to a set in some degree. Formally, in universe U a fuzzy set X is characterized by its membership function $\mu_X : U \to [0, 1]$, such that for any $y \in U$, y certainly does not belong to set X if $\mu_X(y) = 0$, y certainly belongs to X if $\mu_X(y) = 1$, and y belongs to X with a given degree of certainty represented by the value of $\mu_X(y)$ in all other cases.

Granular computing Granular computing is a general computation theory for using granules such as subsets, classes, objects, clusters, and elements of a universe to build an efficient computational model for complex applications with huge amounts of data, information, and knowledge. Granulation of an object a leads to a collection of granules, with a granule being a clump of points (objects) drawn together by indiscernibility,

similarity, proximity, or functionality. In human reasoning and concept formulation, the granules and the values of their attributes are fuzzy rather than crisp. In this perspective, fuzzy information granulation may be viewed as a mode of generalization, which can be applied to any concept, method, or theory.

Ordinal properties and monotonicity Ordinal properties in description of objects are related to graduality of the presence or absence of a property. In this context, it is meaningful to say that a property is more present in one object than in another object. It is important that the ordinal descriptions are handled properly, which means, without introducing any operation, such as sum, averages, or fuzzy operators, like t-norm or t-conorm of Łukasiewicz, taking into account cardinal properties of data not present in the considered descriptions, which would, therefore, give not meaningful results. Monotonicity is strongly related to ordinal properties. It regards relationships between degrees of presence or absence of properties in the objects, like "the more present is property f_i, the more present is property f_j", or "the more present is property f_i, the more absent is property f_j". The graded presence or absence of a property can be meaningfully represented using fuzzy sets. More precisely, the degree of presence of property f_i in object $y \in U$ is the value given to y by the membership function of the set of objects having property f_i.

Rough set A rough set in universe U is an approximation of a set based on available information about objects of U. The rough approximation is composed of two ordinary sets called *lower and upper approximation*. Lower approximation is a maximal subset of objects which, according to the available information, certainly belong to the approximated set, and upper approximation is a minimal subset of objects which, according to the available information, possibly belong to the approximated set. The difference between upper and lower approximation is called *boundary*.

Definition of the Subject

This article describes the dominance-based rough set approach (DRSA) to granular computing and data mining. DRSA was first introduced as a generalization of the rough set approach for dealing with multicriteria decision analysis, where preference order is important. The ordering is also important, however, in many other problems of data analysis. Even when the ordering seems absent, the presence or the absence of a property can be represented in ordinal terms, because if two properties are related, the presence, rather than the absence, of one property should make more (or less) probable the presence of the other property. This is even more apparent when the presence or the absence of a property is graded or fuzzy, because in this case, the more credible the presence of a property, the more (or less) probable the presence of the other property. Since the presence of properties, possibly fuzzy, is the basis of any granulation, DRSA can be seen as a general basis for granular computing.

After presenting the main ideas of DRSA for granular computing and its philosophical basis, the article introduces the basic concepts of DRSA, followed by its extensions in a fuzzy context and in probabilistic terms. This prepares the ground for treating the rough approximation of a fuzzy set, which is the core of the subject. It is also explained why the classical rough set approach is a specific case of DRSA. The article continues with presentation of DRSA for case-based reasoning, where the main ideas of DRSA for granular computing are fruitfully applied. Finally, some basic formal properties of the whole approach are presented in terms of an algebra modeling the logic of DRSA.

Introduction: Granular Computing and Ordered Data

Granular computing originated in the research of Lin [37,38,39,40,41,42] and Zadeh [57,58,59,60] and gained considerable interest in the last decade. The basic components of granular computing are granules, such as subsets, classes, objects, clusters, and elements of a universe. Granulation of an object a leads to a collection of granules, with a granule being a clump of points (objects) drawn together by indiscernibility, similarity, proximity, or functionality. In human reasoning and concept formulation, the granules and the values of their attributes are fuzzy rather than crisp. In this perspective, fuzzy information granulation may be viewed as a mode of generalization, which can be applied to any concept, method, or theory. Moreover, the theory of fuzzy granulation provides a basis for computing with words, due to the observation that in a natural language, words play the role of labels of fuzzy granules.

Since fuzzy granulation plays a central role in fuzzy logic and in its applications, and rough set theory can be considered as a crisp granulation of set theory, it is interesting to study the relationship between fuzzy sets and rough sets from this point of view. Moreover, noticing that fuzzy granulation that leads to fuzzy logic underlies all applications of granulation, hybridization of fuzzy sets and rough sets can lead to a more general theory of granulation with a potential of application to any domain of human in-

vestigation. This explains the interest in putting together rough sets and fuzzy sets.

Recently, it has been shown that a proper way of handling graduality in rough set theory is to use the DRSA [29]. This implies that DRSA is also a proper way of handling granulation within rough set theory. Let us explain this point in detail.

The rough set approach has been proposed to approximate some relationships existing between concepts. For example, in medical diagnosis the concept of "disease Y" can be represented in terms of such concepts as "low blood pressure" and "high temperature", or "muscle pain" and "headache". The classical rough approximation is based on a very coarse representation, that is, for each aspect characterizing a concept ("low blood pressure", "high temperature", "muscle pain", etc.), only its presence or its absence is considered relevant. In this case, the rough approximation involves a very primitive idea of monotonicity related to a scale with only two values: "presence" and "absence".

Monotonicity gains importance when a finer representation of the concepts is considered. A representation is finer when, for each aspect characterizing a concept, not only its presence or its absence is taken into account, but also the *degree* of its presence or absence is considered relevant. Graduality is typical for fuzzy set philosophy [56] and, therefore, a joint consideration of rough sets and fuzzy sets is worthwhile. In fact, rough sets and fuzzy sets capture the two basic complementary aspects of monotonicity: rough sets deal with relationships between different concepts and fuzzy sets deal with expression of different dimensions in which the concepts are considered. For this reason, many approaches have been proposed to combine fuzzy sets with rough sets (see e. g. [3,5,45,49,51]). Our combination of rough sets and fuzzy sets presents some important advantages with respect to the other approaches, which are discussed below.

The main preoccupation in almost all the studies combining rough sets with fuzzy sets was related to a fuzzy extension of Pawlak's definition of lower and upper approximations using fuzzy connectives [10,34]. In fact, there is no rule for the choice of the "right" connective, so this choice is always arbitrary to some extent. Another drawback of fuzzy extensions of rough sets involving fuzzy connectives is that they are based on cardinal properties of membership degrees. In consequence, the result of these extensions is sensitive to order preserving transformation of membership degrees. For example, consider the t-conorm of Łukasiewicz as a fuzzy connective; it may be used in the definition of both fuzzy lower approximation (to build fuzzy implication) and fuzzy upper approximation (as a fuzzy counterpart of a union). The t-conorm of Łukasiewicz is defined as

$$T^*(\alpha,\beta) = \min(\alpha+\beta,1)\,, \quad \alpha,\beta \in [0,1]\,.$$

$T^*(\alpha,\beta)$ can be interpreted as follows. Given two fuzzy propositions p and q, putting $v(p)=\alpha$ and $v(q)=\beta$, $T^*(\alpha,\beta)$ can be interpreted as $v(p \vee q)$, the truth value of the proposition $p \vee q$. Let us consider the following values of arguments:

$$\alpha = 0.5\,, \quad \beta = 0.3\,, \quad \gamma = 0.2\,, \quad \delta = 0.1\,,$$

and their order preserving transformation:

$$\alpha' = 0.4\,, \quad \beta' = 0.3\,, \quad \gamma' = 0.2\,, \quad \delta' = 0.05\,.$$

The values of the t-conorm are in the two cases as follows:

$$T^*(\alpha,\delta) = 0.6\,, \qquad T^*(\beta,\gamma) = 0.5\,,$$
$$T^*(\alpha',\delta') = 0.45\,, \qquad T^*(\beta',\gamma') = 0.5\,.$$

One can see that the order of the results has changed after the order preserving transformation of the arguments. This means that the Łukasiewicz t-conorm takes into account not only the ordinal properties of the truth values, but also their cardinal properties. A natural question arises: is it reasonable to expect from truth values a cardinal content instead of ordinal only? Or, in other words, is it realistic to claim that a human is able to say in a meaningful way not only that

(a) "proposition p is more credible than proposition q" but even something like
(b) "proposition p is two times more credible than proposition q"?

It is much safer to consider information of type (a), because information of type (b) is rather meaningless for a human.

Since fuzzy generalization of rough set theory using DRSA takes into account only *ordinal properties* of fuzzy membership degrees, it is the proper way of fuzzy generalization of rough set theory.

Moreover, the classical rough set approach [46,47] can be seen as a specific case of our general model. This is important for several reasons. In particular, this interpretation of DRSA gives an insight into fundamental properties of the classical rough set approach and permits its further generalization.

Rough set theory [46,47] relies on the idea that some knowledge (data, information) is available about objects of a universe of discourse U. Thus, a subset of U is defined using the available knowledge about the objects and not on the base of information about membership or non-mem-

bership of the objects to the subset. For example, knowledge about patients suffering from a certain disease may contain information about body temperature, blood pressure, etc. All patients described by the same information are *indiscernible* in view of the available knowledge, and form groups of similar objects. These groups are called *elementary sets*, and can be considered as basic granules of the available knowledge about patients. Elementary sets can be combined into compound concepts. For example, elementary sets of patients can be used to represent a set of patients suffering from a certain disease. Any union of elementary sets is called a crisp set, while other sets are referred to as *rough sets*. Each rough set has boundary line objects, i. e. objects which, in view of the available knowledge, cannot be classified with certainty as members of the set or of its complement. Therefore, in the rough set approach, any set is associated with a pair of crisp sets, called the lower and the upper approximation. Intuitively, in view of the available information, the *lower approximation* consists of all objects which certainly belong to the set and the *upper approximation* contains all objects which possibly belong to the set. The difference between the upper and the lower approximation constitutes the *boundary* region of the rough set. Analogously, for a partition of universe U into classes, one may consider rough approximation of the partition. It appeared to be particularly useful for analysis of classification problems, being the most common decision problems.

The rough set approach operates on an *information table* composed of a set U of objects described by a set Q of attributes. If in the set Q disjoint sets (C and D) of *condition* and *decision* attributes are distinguished, then the information table is called a *decision table*. It is often assumed, without loss of generality, that set D is a singleton $\{d\}$, and thus decision attribute d makes a partition of set U into decision classes corresponding to its values. Data collected in such a decision table correspond to a multiple attribute classification problem. The classical Indiscernibility-Based Rough Set Approach (IRSA) is naturally adapted to analysis of this type of decision problems, because the set of objects can be identified with examples of classification and it is possible to extract all the essential knowledge contained in the decision table using indiscernibility or similarity relations. However, as pointed out by the authors (see e. g. [17,20,26,54]), IRSA cannot extract all the essential knowledge contained in the decision table if a background knowledge about *monotonic relationships* between evaluation of objects on condition attributes and their assignment to decision classes has to be taken into account. Such a background knowledge is typical for data describing various phenomena, as well as for data describing multiple criteria decision problems (see e. g. [9]), e. g., "the larger the mass and the smaller the distance, the larger the gravity", "the more a tomato is red, the more it is ripe" or "the better the school marks of a pupil, the better his overall classification". The monotonic relationships, typical for multiple criteria decision problems, follow from preferential ordering of value sets of attributes (scales of criteria), as well as preferential ordering of decision classes.

In order to take into account the *ordinal* properties of the considered attributes and the monotonic relationships between condition and decision attributes, a number of methodological changes to the original rough set theory were necessary. The main change was the replacement of the indiscernibility relation with a *dominance relation*, which permits approximation of ordered sets. The dominance relation is a very natural and rational concept within multiple criteria decision analysis. The dominance-based rough set approach (DRSA) has been proposed and characterized by the authors (see e. g. [17,20,24,25,26,54]).

Let us mention that ordered value sets of attributes and a kind of order dependency among attributes has also been considered by specialists of relational databases (see, e. g., [13]). There is, however, a striking difference between consideration of orders in database queries and consideration of orders in knowledge discovery. More precisely, assuming ordered domains of attributes, knowledge discovery tends to discover monotonic relationships between ordered attributes, e. g., if a student is at least medium in networks, and at least good in databases, then his overall evaluation is at least medium. On the other hand, assuming an order of attribute value sets and order dependency in databases, one can exploit this given information for a more efficient answer to a query, e. g., when the dates of bank checks and their numbers are ordered, there is an order dependency between these two attributes, because in day x a check cannot hold a number smaller than checks from day $x - 1$, which permits us to prune the search tree and make the search more efficient.

Looking at DRSA from a granular computing perspective, one can observe that DRSA permits us to deal with *ordered data* by considering a specific type of information granules defined by means of *dominance-based constraints* having a syntax of the type: "x is at least R" or "x is at most R", where R is a qualifier from a properly ordered scale. In evaluation space, such granules are *dominance cones*. In this sense, the contribution of DRSA consists of:

- Extending the paradigm of granular computing to problems involving ordered data,
- Specifying a proper syntax and modality of information granules (the dominance-based constraints which

should be adjoined to other modalities of information constraints, such as possibilistic, veristic, and probabilistic [60]),
- Defining a methodology dealing properly with this type of information granules, and resulting in a theory of computing with words and reasoning about data in the case of ordered data.

Let us observe that other modalities of information constraints, such as veristic, possibilistic, and probabilistic, have also to deal with ordered values (with qualifiers relative to grades of truth, possibility, and probability). Therefore, granular computing with ordered data and DRSA as a proper way of reasoning about ordered data, are very important in the future development of the whole domain of granular computing.

Indeed the DRSA approach proposed in [15,16] avoids arbitrary choice of fuzzy connectives and not meaningful operations on membership degrees. It exploits only ordinal character of the membership degrees and proposes a methodology of fuzzy rough approximation that infers the most cautious conclusion from available imprecise information. In particular, any approximation of knowledge about concept Y using knowledge about concept X is based on positive or negative relationships between premises and conclusions, i. e.:

(i) "The more x is X, the more it is Y" (positive relationship),
(ii) "The more x is X, the less it is Y" (negative relationship).

The following simple relationships illustrate i) and ii):

- "The larger the market share of a company, the greater its profit" (positive relationship), and
- "The greater the debt of a company, the smaller its profit" (negative relationship).

These relationships have the form of *gradual decision rules* [4]. Examples of these decision rules are: "if a car is speedy with credibility at least 0.8 and it has high fuel consumption with credibility at most 0.7, then it is a good car with a credibility at least 0.9", and "if a car is speedy with credibility at most 0.5 and it has high fuel consumption with credibility at least 0.8, then it is a good car with a credibility at most 0.6". It is worth noting that the syntax of gradual decision rules is based on monotonic relationships between degrees of credibility, that can also be found in dominance-based decision rules induced from preference-ordered data. This explains why one can build a fuzzy rough approximation using DRSA.

Finally, the fuzzy rough approximation taking into account monotonic relationships can be applied to case-based reasoning [28]. In this perspective, it is interesting to consider monotonicity of the type "the more similar is y to x, the more credible is that y belongs to the same class as x". Application of DRSA in this context leads to decision rules similar to the gradual decision rules:

> "the more object z is similar to a referent object x w.r.t. condition attribute s, the more z is similar to a referent object x w.r.t. decision attribute t",

or, equivalently, but more technically,

$$s(z,x) \geq \alpha \quad \Rightarrow \quad t(z,x) \geq \alpha \,,$$

where functions s and t measure the credibility of similarity with respect to condition attribute and decision attribute, respectively. When there are multiple condition and decision attributes, functions s and t aggregate similarity with respect to these attributes.

The decision rules induced on the basis of DRSA do not need the aggregation of the similarity with respect to different attributes into one comprehensive similarity. This is important because it permits us to avoid using aggregation operators (weighted average, min, etc.) which are always arbitrary to some extent. Moreover, the DRSA decision rules permit us to consider different thresholds for degrees of credibility in the premise and in the conclusion.

The article is organized as follows. Next Section presents the philosophical basis of DRSA granular computing. Section "Dominance-Based Rough Set Approach" recalls the main steps of DRSA for multiple criteria classification, or, in general, for ordinal classification. Section "Fuzzy Set Extensions of the Dominance-Based Rough Set Approach" presents, a review of fuzzy set extensions of DRSA based on fuzzy connectives. In Sect. "Variable-Consistency Dominance-Based Rough Set Approach (VC–DRSA)", a "probabilistic" version of DRSA, the variable-consistency dominance-based rough set approach is presented. Dominance-based rough approximation of a fuzzy set is presented in Sect "Dominance-Based Rough Approximation of a Fuzzy Set". The explanation that IRSA is a particular case of DRSA is presented in Sect. "Monotonic Rough Approximation of a Fuzzy Set versus Classical Rough Set". Section "Dominance-Based Rough Set Approach to Case-Based Reasoning" is devoted to DRSA for case-based reasoning. In Sect. "An Algebraic Structure for Dominance-Based Rough Set Approach" an algebra modeling logic of DRSA is presented. Section "Conclusions" contains conclusions. In Sect. "Future Directions" some issues for further developments are presented.

Philosophical Basis of DRSA Granular Computing

It is interesting to analyze the relationships between DRSA and granular computing from the point of view of the philosophical basis of rough set theory proposed by Pawlak. Since according to Pawlak [48], rough set theory refers to some ideas of Gottlob Frege (vague concepts), Gottfried Leibniz (indiscernibility), George Boole (reasoning methods), Jan Łukasiewicz (multi-valued logic), and Thomas Bayes (inductive reasoning), it is meaningful to give an account for DRSA generalization of rough sets, justifying it in reference to some of these main ideas recalled by Pawlak.

The *identity of indiscernibles* is a principle of analytic ontology first explicitly formulated by Gottfried Leibniz in his Discourse on Metaphysics, Sect. 9 [43]. Two objects x and y are defined indiscernible, if x and y have the same properties. The principle of identity of indiscernibles states that

if x and y are indiscernible, then $x = y$. (II1)

This can be expressed also as *if* $x \neq y$, *then x and y are discernible*, i. e. there is at least one property that x has and y does not, or vice versa. The converse of the principle of identity of indiscernibles is called *indiscernibility of identicals* and states that *if* $x = y$, *then x and y are indiscernible*, i. e. they have the same properties. This is equivalent to saying that if there is at least one property that x has and y does not, or vice versa, then $x \neq y$. The conjunction of both principles is often referred to as "Leibniz's law".

Rough set theory is based on a weaker interpretation of Leibniz's law, having as objective the ability to classify objects falling under the same concept. This reinterpretation of Leibniz's law is based on a reformulation of the principle of identity of indiscernibles as follows:

if x and y are indiscernible,
then x and y belong to the same class . (II2)

Let us observe that the word "class" in the previous sentence can be considered as synonymous with "granule". Thus, from the point of view of granular computing, (II2) can be rewritten as

if x and y are indiscernible,
then x and y belong to the same granule of classification . (II2')

Notice also that the principle of indiscernibility of identicals cannot be reformulated in analogous terms. In fact, such an analogous reformulation would amount to stating that *if x and y belong to the same class, then x and y are indiscernible*. This principle is too strict, however, because there can be two discernible objects x and y belonging to the same class. Thus, within rough set theory, the principle of indiscernibility of identicals should continue to hold in its original formulation (i. e. *if* $x = y$, *then x and y are indiscernible*). It is worthwhile to observe that the relaxation in the consequence of the implication from (II1) to (II2), implies an implicit relaxation also in the antecedent. In fact, one could say that two objects are identicals if they have the same properties, if one would be able to take into account all conceivable properties. For human limitations this is not the case, therefore, one can imagine that (II2) can be properly reformulated as

if x and y are indiscernible taking into account
a given set of properties, (II2")
then x and y belong to the same class .

This weakening in the antecedent of the implication means also that the objects indiscernible with respect to a given set of properties can be seen as a granule, such that, finally, the (II2) could be rewritten in terms of granulation as

if x and y belong to the same granule with respect
to a given set of properties,
then x and y belong to the same classification granule . (II2"')

For this reason, rough set theory needs a still weaker form of the principle of identity of indiscernibles. Such a principle can be formulated using the idea of vagueness due to Gottlob Frege. According to Frege "the concept must have a sharp boundary – to the concept without a sharp boundary there would correspond an area that had not a sharp boundary-line all around". Therefore, following this intuition, the principle of identity of indiscernibles can be further reformulated as

if x and y are indiscernible,
then x and y should belong to the same class . (II3)

In terms of granular computing, (II3) can be rewritten as

if x and y belong to the same granule with respect
to a given set of properties,
then x and y should belong to the same classification
granule . (II3')

This reformulation of the principle of identity of indiscernibles implies that there is an inconsistency in the statement that x and y are indiscernible, and x and y be-

long to different classes. Thus, Leibniz's principle of identity of indiscernibles and Frege's intuition about vagueness found the basic idea of the rough set concept proposed by Pawlak.

The above reconstruction of the basic idea of Pawlak's rough set should be completed, however, by referring to another basic idea. This is the idea of Georg Boole that concerns a property which is satisfied or not satisfied. It is quite natural to weaken this principle admitting that a property can be satisfied to some degree. This idea of graduality can be attributed to Jan Łukasiewicz and his proposal of many-valued logic where, in addition to well-known truth values "true" and "false", other truth values representing partial degrees of truth were present. Łukasiewicz's idea of graduality has been reconsidered, generalized and fully exploited by Zadeh [56] within fuzzy set theory, where graduality concerns membership to a set.

In this sense, any proposal of putting rough sets and fuzzy sets together can be seen as a reconstruction of the rough set concept, where Boole's idea of binary logic is abandoned in favor of Łukasiewicz's idea of many-valued logic, such that Leibniz's principle of identity of indiscernibles and Frege's intuition about vagueness are combined with the idea that a property is satisfied to some degree.

Putting aside, for the moment, Frege's intuition about vagueness, but taking into account the concept of graduality, the principle of identity of indiscernibles can reformulated as follows:

if the grade of each property for x is greater than or equal to the grade for y,
then x belongs to the considered class in a grade at least as high as y . (II4)

Taking into account the paradigm of granular computing, (II4) can be rewritten as

if x belongs to the granules defined by considered properties more than y, because the grade of each property for x is greater than or equal to the grade for y,
then x belongs to the considered classification granule in a grade at least as high as y . (II4')

Considering the concept of graduality together with Frege's intuition about vagueness, one can reformulate the principle of identity of indiscernibles as follows:

if the grade of each property for x is greater than or equal to the grade for y,
then x should belong to the considered class in a grade at least as high as y . (II5)

In terms of granular computing, (II5) can be rewritten as

if x belongs to the granules defined by considered properties more than y, because the grade of each property for x is greater than or equal to the grade for y,
then x should belong to the considered classification granule in a grade at least as high as y . (II5')

The formulation (II5') of the principle of identity of indiscernibles is perfectly concordant with the rough set concept defined within the dominance-based rough set approach [20].

DRSA has been proposed by the authors to deal with ordinal properties of data related to preferences in decision problems [26,54]. The fundamental feature of DRSA is that it handles monotonicity of comprehensive evaluation of objects with respect to preferences relative to evaluation of these objects on particular attributes. For example, the more preferred is a car with respect to such attributes as maximum speed, acceleration, fuel consumption, and price, the better is its comprehensive evaluation. The type of monotonicity considered within DRSA is also meaningful for problems where relationships between different aspects of a phenomenon described by data are to be taken into account, even if preferences are not considered. Indeed, monotonicity concerns, in general, mutual trends existing between different variables, like distance and gravity in physics, or inflation rate and interest rate in economics. Whenever a relationship between different aspects of a phenomenon is discovered, this relationship can be represented by a monotonicity with respect to some specific measures of the considered aspects. Formulation (II5) of the principle of identity of indiscernibles refers to this type of monotonic relationships. So, in general, the monotonicity permits us to translate into a formal language a primitive intuition of relationships between different concepts of our knowledge corresponding to the principle of identity of indiscernibles formulated as (II5').

Dominance-Based Rough Set Approach

This section presents the main concepts of the dominance-based rough set approach (for a more complete presentation see, for example, [17,20,26,54]).

Information about objects is represented in the form of an information table. The rows of the table are labeled by objects, whereas columns are labeled by attributes and entries of the table are attribute-values. Formally, an information system (table) is the 4-tuple $\boldsymbol{S} = \langle U, Q, V, \phi \rangle$, where U is a finite set of objects, Q is a finite set of attributes, $V = \bigcup_{q \in Q} V_q$ and V_q is the set of values of the

attribute q, and $\phi\colon U \times Q \to V_q$ is a total function such that $\phi(x,q) \in V_q$ for every $q \in Q$, $x \in U$, called an information function [47]. The set Q is, in general, divided into set C of condition attributes and set D of decision attributes.

Condition attributes with value sets ordered according to decreasing or increasing preference are called *criteria*. For criterion $q \in Q$, $\succeq_q$ is a *weak preference* relation on U such that $x \succeq_q y$ means "x is at least as good as y with respect to criterion q". It is supposed that $\succeq_q$ is a complete preorder, i. e. a strongly complete and transitive binary relation, defined on U on the basis of evaluations $\phi(\cdot, q)$. Without loss of generality, the preference is supposed to increase with the value of $\phi(\cdot, q)$ for every criterion $q \in C$, such that for all $x, y \in U$, $x \succeq_q y$ if and only if $\phi(x,q) \geq \phi(y,q)$.

Furthermore, it is supposed that the set of decision attributes D is a singleton d. Values of decision attribute d makes a partition of U into a finite number of decision classes, $\boldsymbol{Cl} = \{Cl_t, t = 1, \ldots, n\}$, such that each $x \in U$ belongs to one and only one class $Cl_t \in Cl$. It is supposed that the classes are preference-ordered, i. e. for all $r, s \in \{1, \ldots, n\}$, such that $r > s$, the objects from Cl_r are preferred to the objects from Cl_s. More formally, if $\succeq$ is a *comprehensive weak preference relation* on U, i. e. if for all $x, y \in U$, $x \succeq y$ means "x is at least as good as y", it is supposed: $[x \in Cl_r, y \in Cl_s, r{>}s] \Rightarrow [x \succeq y$ and *not* $y \succeq x]$. The above assumptions are typical for consideration of *ordinal classification problems* (also called *multiple criteria sorting problems*).

The sets to be approximated are called *upward union* and *downward union* of classes, respectively:

$$Cl_t^{\geq} = \bigcup_{s \geq t} Cl_s\,, \quad Cl_t^{\leq} = \bigcup_{s \leq t} Cl_s\,, \quad t = 1, \ldots, n\,.$$

The statement $x \in Cl_t^{\geq}$ means "x belongs to at least class Cl_t", while $x \in Cl_t^{\leq}$ means "x belongs to at most class Cl_t". Let us remark that $Cl_1^{\geq} = Cl_n^{\leq} = U$, $Cl_n^{\geq} = Cl_n$ and $Cl_1^{\leq} = Cl_1$. Furthermore, for $t = 2, \ldots, n$,

$$Cl_{t-1}^{\leq} = U - Cl_t^{\geq} \quad \text{and} \quad Cl_t^{\geq} = U - Cl_{t-1}^{\leq}\,.$$

The key idea of the rough set approach is representation (approximation) of knowledge generated by decision attributes, by "*granules of knowledge*" generated by condition attributes.

In DRSA, where condition attributes are criteria and decision classes are preference ordered, the represented knowledge is a collection of *upward* and *downward unions of classes* and the "granules of knowledge" are sets of objects defined using a dominance relation.

x dominates y with respect to $P \subseteq C$ (shortly, *x P-dominates y*), denoted by xD_Py, if for every criterion $q \in P$, $\phi(x,q) \geq \phi(y,q)$. The relation of P-dominance is reflexive and transitive, that is it is a partial preorder.

Given a set of criteria $P \subseteq C$ and $x \in U$, the "granules of knowledge" used for approximation in DRSA are:

- A set of objects dominating x, called *P-dominating set*, $D_P^+(x) = \{y \in U : yD_Px\}$,
- A set of objects dominated by x, called *P-dominated set*, $D_P^-(x) = \{y \in U : xD_Py\}$.

Note that the "granules of knowledge" defined above have the form of upward (positive) and downward (negative) *dominance cones* in the evaluation space.

Let us recall that the *dominance principle* (or Pareto principle) requires that an object x dominating object y on all considered criteria (i. e. x having evaluations at least as good as y on all considered criteria) should also dominate y on the decision (i. e. x should be assigned to at least as good a decision class as y). This principle is the only objective principle that is widely agreed upon in the multiple criteria comparisons of objects.

Given $P \subseteq C$, the inclusion of an object $x \in U$ to the upward union of classes $Cl_t^{\geq}$, $t = 2, \ldots, n$, is *inconsistent with the dominance principle* if one of the following conditions holds:

- x belongs to class Cl_t or better but it is P-dominated by an object y belonging to a class worse than Cl_t, i. e. $x \in Cl_t^{\geq}$ but $D_P^+(x) \cap Cl_{t-1}^{\leq} \neq \emptyset$,
- x belongs to a worse class than Cl_t but it P-dominates an object y belonging to class Cl_t or better, i. e. $x \notin Cl_t^{\geq}$ but $D_P^-(x) \cap Cl_t^{\geq} \neq \emptyset$.

If, given a set of criteria $P \subseteq C$, the inclusion of $x \in U$ to $Cl_t^{\geq}$, where $t = 2, \ldots, n$, is inconsistent with the dominance principle, then x belongs to $Cl_t^{\geq}$ *with some ambiguity*. Thus, x belongs to $Cl_t^{\geq}$ *without any ambiguity* with respect to $P \subseteq C$, if $x \in Cl_t^{\geq}$ and there is no inconsistency with the dominance principle. This means that all objects P-dominating x belong to $Cl_t^{\geq}$, i. e. $D_P^+(x) \subseteq Cl_t^{\geq}$.

Furthermore, *x possibly belongs to $Cl_t^{\geq}$* with respect to $P \subseteq C$ if one of the following conditions holds:

- According to decision attribute d, x belongs to $Cl_t^{\geq}$,
- According to decision attribute d, x does not belong to $Cl_t^{\geq}$, but it is inconsistent in the sense of the dominance principle with an object y belonging to $Cl_t^{\geq}$.

In terms of ambiguity, x possibly belongs to $Cl_t^{\geq}$ with respect to $P \subseteq C$, if x belongs to $Cl_t^{\geq}$ with or without any ambiguity. Because of the reflexivity of the dominance relation D_P, the above conditions can be summarized as

follows: *x possibly belongs* to class Cl_t or better, with respect to $P \subseteq C$, if among the objects P-dominated by x there is an object y belonging to class Cl_t or better, i. e. $D_P^-(x) \cap Cl_t^{\geq} \neq \emptyset$.

The *P-lower approximation* of $Cl_t^{\geq}$, denoted by $\underline{P}Cl_t^{\geq}$, and the *P-upper approximation* of $Cl_t^{\geq}$, denoted by $\overline{P}\left(Cl_t^{\geq}\right)$, are defined as follows ($t = 1, \ldots, n$):

$$\underline{P}\left(Cl_t^{\geq}\right) = \left\{x \in U : D_P^+(x) \subseteq Cl_t^{\geq}\right\} ,$$
$$\overline{P}\left(Cl_t^{\geq}\right) = \left\{x \in U : D_P^-(x) \cap Cl_t^{\geq} \neq \emptyset\right\} .$$

Analogously, one can define the *P-lower approximation* and the *P-upper approximation* of $Cl_t^{\leq}$ as follows ($t = 1, \ldots, n$):

$$\underline{P}\left(Cl_t^{\leq}\right) = \left\{x \in U\colon D_P^-(x) \subseteq Cl_t^{\leq}\right\} ,$$
$$\overline{P}\left(Cl_t^{\leq}\right) = \left\{x \in U\colon\ D_P^+(x) \cap Cl_t^{\leq} \neq \emptyset\right\} .$$

The P-lower and P-upper approximations so defined satisfy the following *inclusion properties* for each $t \in \{1, \ldots, n\}$ and for all $P \subseteq C$:

$$\underline{P}\left(Cl_t^{\geq}\right) \subseteq Cl_t^{\geq} \subseteq \overline{P}\left(Cl_t^{\geq}\right) ,$$
$$\underline{P}\left(Cl_t^{\leq}\right) \subseteq Cl_t^{\leq} \subseteq \overline{P}\left(Cl_t^{\leq}\right) .$$

The P-lower and P-upper approximations of $Cl_t^{\geq}$ and $Cl_t^{\leq}$ have an important *complementarity property*, according to which,

$$\underline{P}\left(Cl_t^{\geq}\right) = U - \overline{P}\left(Cl_{t-1}^{\leq}\right) \quad \text{and}$$
$$\overline{P}\left(Cl_t^{\geq}\right) = U - \underline{P}\left(Cl_{t-1}^{\leq}\right) , \quad t = 2, \ldots, n ,$$

$$\underline{P}\left(Cl_t^{\leq}\right) = U - \overline{P}\left(Cl_{t+1}^{\geq}\right) \quad \text{and}$$
$$\overline{P}\left(Cl_t^{\leq}\right) = U - \underline{P}\left(Cl_{t+1}^{\geq}\right) , \quad t = 1, \ldots, n-1 .$$

The *P-boundary* of $Cl_t^{\geq}$ and $Cl_t^{\leq}$, denoted by $Bn_P\left(Cl_t^{\geq}\right)$ and $Bn_P\left(Cl_t^{\leq}\right)$ respectively, are defined as follows ($t = 1, \ldots, n$):

$$Bn_P\left(Cl_t^{\geq}\right) = \overline{P}\left(Cl_t^{\geq}\right) - \underline{P}\left(Cl_t^{\geq}\right) ,$$
$$Bn_P\left(Cl_t^{\leq}\right) = \overline{P}\left(Cl_t^{\leq}\right) - \underline{P}\left(Cl_t^{\leq}\right) .$$

Because of the complementarity property, $Bn_P(Cl_t^{\geq}) = Bn_P(Cl_{t-1}^{\leq})$, for $t = 2, \ldots, n$.

The dominance-based rough approximations of upward and downward unions of classes can serve to induce "*if…, then…*" decision rules. It is meaningful to consider the following five types of decision rules:

1) Certain $D_{\geq}$-decision rules: if $x_{q1} \succeq_{q1} r_{q1}$ and $x_{q2} \succeq_{q2} r_{q2}$ and … $x_{qp} \succeq_{qp} r_{qp}$, then *certainly x belongs to* $Cl_t^{\geq}$, where, for each $w_q, z_q \in X_q$, "$w_q \succeq_q z_q$" means "w_q is *at least* as good as z_q".
2) Possible $D_{\geq}$-decision rules: if $x_{q1} \succeq_{q1} r_{q1}$ and $x_{q2} \succeq_{q2} r_{q2}$ and … $x_{qp} \succeq_{qp} r_{qp}$, then *x possibly belongs to* $Cl_t^{\geq}$.
3) Certain $D_{\leq}$-decision rules: if $x_{q1} \preceq_{q1} r_{q1}$ and $x_{q2} \preceq_{q2} r_{q2}$ and … $x_{qp} \preceq_{qp} r_{qp}$, then *certainly x belongs to* $Cl_t^{\leq}$, where, for each $w_q, z_q \in X_q$, "$w_q \preceq_q z_q$" means "w_q is *at most* as good as z_q".
4) Possible $D_{\leq}$-decision rules: if $x_{q1} \preceq_{q1} r_{q1}$ and $x_{q2} \preceq_{q2} r_{q2}$ and … $x_{qp} \preceq_{qp} r_{qp}$, then *x possibly belongs to* $Cl_t^{\leq}$.
5) Approximate $D_{\geq\leq}$-decision rules: if $x_{q1} \succeq_{q1} r_{q1}$ and … $x_{qk} \succeq_{qk} r_{qk}$ and $x_{q(k+1)} \preceq_{q(k+1)} r_{q(k+1)}$ and … $x_{qp} \preceq_{qp} r_{qp}$, then $x \in Cl_s^{\geq} \cap Cl_t^{\leq}$, where $s < t$.

The rules of type 1) and 3) represent certain knowledge extracted from the decision table, while the rules of type 2) and 4) represent possible knowledge. Rules of type 5) represent doubtful knowledge.

Fuzzy Set Extensions of the Dominance-Based Rough Set Approach

The concept of dominance can be refined by introducing gradedness through the use of fuzzy sets. Here are basic definitions of fuzzy connectives [10,34]. For each proposition p, one can consider its truth value $v(p)$ ranging from $v(p) = 0$ (p is definitely false) to $v(p) = 1$ (p is definitely true); and for all intermediate values, the greater $v(p)$, the more credible is the truth of p. A negation is a non-increasing function $N\colon [0,1] \to [0,1]$ such that $N(0) = 1$ and $N(1) = 0$. Given proposition p, $N(v(p))$ states the credibility of the negation of p. A t-norm T and a t-conorm T^* are two functions $T\colon [0,1] \times [0,1] \to [0,1]$ and $T^*\colon [0,1] \times [0,1] \to [0,1]$, such that given two propositions, p and q, $T(v(p), v(q))$ represents the credibility of the conjunction of p and q, and $T^*(v(p), v(q))$ represents the credibility of the disjunction of p and q. t-norm T and t-conorm T^* must satisfy the following properties:

$$T(\alpha, \beta) = T(\beta, \alpha) \quad \text{and} \quad T^*(\alpha, \beta) = T^*(\beta, \alpha) ,$$
$$\text{for all } \alpha, \beta \in [0,1] ,$$

$$T(\alpha, \beta) \leq T(\gamma, \delta) \quad \text{and} \quad T^*(\alpha, \beta) \leq T^*(\gamma, \delta) ,$$
$$\text{for all } \alpha, \beta, \gamma, \delta \in [0,1]$$
$$\text{such that } \alpha \leq \gamma \quad \text{and} \quad \beta \leq \delta ,$$

$$T(\alpha, T(\beta, \gamma)) = T(T(\alpha, \beta), \gamma) \quad \text{and}$$
$$T^*(\alpha, T^*(\beta, \gamma)) = T^*(T^*(\alpha, \beta), \gamma) ,$$
$$\text{for all } \alpha, \beta, \gamma \in [0,1] ,$$
$$T(1, \alpha) = \alpha \quad \text{and} \quad T^*(0, \alpha) = \alpha ,$$
$$\text{for all } \alpha \in [0,1] .$$

A negation is strict iff it is strictly decreasing and continuous. A negation N is involutive iff, for all $\alpha \in [0,1]$, $N(N(\alpha)) = \alpha$. A strong negation is an involutive strict negation. If N is a strong negation, then (T, T^*, N) is a de Morgan triplet iff $N(T^*(\alpha,\beta)) = T(N(\alpha), N(\beta))$. A fuzzy implication is a function $I\colon [0,1] \times [0,1] \to [0,1]$ such that, given two propositions p and q, $I(v(p), v(q))$ represents the credibility of the implication of q by p. A fuzzy implication must satisfy the following properties (see [10]):

$$
\begin{aligned}
&I(\alpha,\beta) \geq I(\gamma,\beta) \quad \text{for all } \alpha,\beta,\gamma \in [0,1]\,,\\
&\text{such that } \alpha \leq \gamma\,,\\
&I(\alpha,\beta) \geq I(\alpha,\gamma) \quad \text{for all } \alpha,\beta,\gamma \in [0,1]\,,\\
&\text{such that } \beta \geq \gamma\,,\\
&I(0,\alpha) = 1,\ I(\alpha,1) = 1 \quad \text{for all } \alpha \in [0,1]\,,\\
&I(1,0) = 0\,.
\end{aligned}
$$

An implication $I_{N,T^*}^{\rightarrow}$ is a T^*-implication if there is a t-conorm T^* and a strong negation N such that $I_{N,T^*}^{\rightarrow}(\alpha,\beta) = T^*(N(\alpha),\beta)$. A fuzzy similarity relation on the universe U is a fuzzy binary relation (i. e. function $R\colon U \times U \to [0,1]$) reflexive ($R(x,x) = 1$ for all $x \in U$), symmetric ($R(x,y) = R(y,x)$ for all $x,y \in U$) and transitive (given t-norm T, $T(R(x,y), R(y,z)) \leq R(x,z)$ for all $x,y,z \in U$).

Let $\succeq_q$ be a fuzzy weak preference relation on U with respect to criterion $q \in C$, i. e. $\succeq_q\colon U \times U \to [0,1]$, such that, for all $x,y \in U$, $\succeq_q (x,y)$ represents the credibility of the proposition "x is at least as good as y with respect to criterion q". Suppose that $\succeq_q$ is a fuzzy partial T-preorder, i. e. that it is reflexive ($\succeq_q (x,x) = 1$ for each $x \in U$) and T-transitive ($T(\succeq_q (x,y), \succeq_q (y,z)) \leq \succeq_q (x,z)$, for each $x,y,z \in U$) (see [10]). Using the fuzzy weak preference relations $\succeq_q$, $q \in C$, a fuzzy dominance relation on U (denotation $D_P(x,y)$) can be defined, for all $P \subseteq C$, as follows:

$$D_P(x,y) = T_{q\in P}(\succeq_q (x,y))\,.$$

Given $(x,y) \in U \times U$, $D_P(x,y)$ represents the credibility of the proposition "x is at least as good as y with respect to each criterion q from P". Since the fuzzy weak preference relations $\succeq_q$ are supposed to be partial T-preorders, then also the fuzzy dominance relation D_P is a partial T-preorder. Furthermore, let $\boldsymbol{Cl} = \{Cl_t, t = 1,\ldots,n\}$ be a set of fuzzy classes in U such that, for each $x \in U$, $Cl_t(x)$ represents the membership function of x to Cl_t. It is supposed, as before, that the classes of $\boldsymbol{Cl}$ are increasingly ordered, i. e. that for all $r,s \in \{1,\ldots,n\}$ such that $r > s$, the objects from Cl_r have a better comprehensive evaluation than the objects from Cl_s. On the basis of the membership functions of the fuzzy class Cl_t, fuzzy membership functions of two other sets can be defined as follows:

1) The upward union fuzzy set $Cl_t^{\geq}$, whose membership function $Cl_t^{\geq}(x)$ represents the credibility of the proposition "x is at least as good as the objects in Cl_t":

$$Cl_t^{\geq}(x) = \left\{ \begin{array}{ll} 1 & \text{if } \exists s \in \{1,\ldots,n\}\colon Cl_s(x) > 0 \\ & \text{and } s > t \\ Cl_t(x) & \text{otherwise} \end{array} \right\};$$

2) the downward union fuzzy set $Cl_t^{\leq}$, whose membership function $Cl_t^{\leq}(x)$ represents the credibility of the proposition "x is at most as good as the objects in Cl_t":

$$Cl_t^{\leq}(x) = \left\{ \begin{array}{ll} 1 & \text{if } \exists s \in \{1,\ldots,n\}\colon Cl_s(x) > 0 \\ & \text{and } s < t \\ Cl_t(x) & \text{otherwise} \end{array} \right\}.$$

The P-lower and the P-upper approximations of $Cl_t^{\geq}$ with respect to $P \subseteq C$ are fuzzy sets in U, whose membership functions, denoted by $\underline{P}[Cl_t^{\geq}(x)]$ and $\overline{P}[Cl_t^{\geq}(x)]$ respectively, are defined as:

$$
\begin{aligned}
\underline{P}\left[Cl_t^{\geq}(x)\right] &= T_{y\in U}\left(T^*\left(N(D_P(y,x)), Cl_t^{\geq}(y)\right)\right)\,,\\
\overline{P}\left[Cl_t^{\geq}(x)\right] &= T^*_{y\in U}\left(T\left(D_P(x,y), Cl_t^{\geq}(y)\right)\right)\,.
\end{aligned}
$$

$\underline{P}[Cl_t^{\geq}(x)]$ represents the credibility of the proposition "for all $y \in U$, y does not dominate x with respect to criteria from P or y belongs to $Cl_t^{\geq}$", while $\overline{P}[Cl_t^{\geq}(x)]$ represents the credibility of the proposition "there is at least one $y \in U$ dominated by x with respect to criteria from P which belongs to $Cl_t^{\geq}$".

The P-lower and P-upper approximations of $Cl_t^{\leq}$ with respect to $P \subseteq C$, denoted by $\underline{P}[Cl_t^{\leq}(x)]$ and $\overline{P}[Cl_t^{\leq}(x)]$ respectively, can be defined, analogously, as:

$$
\begin{aligned}
\underline{P}\left[Cl_t^{\leq}(x)\right] &= T_{y\in U}\left(T^*\left(N(D_P(x,y)), Cl_t^{\leq}(y)\right)\right)\,,\\
\overline{P}\left[Cl_t^{\leq}(x)\right] &= T^*_{y\in U}\left(T\left(D_P(y,x), Cl_t^{\leq}(y)\right)\right)\,.
\end{aligned}
$$

$\underline{P}\left[Cl_t^{\leq}(x)\right]$ represents the credibility of the proposition "for all $y \in U$, x does not dominate y with respect to criteria from P or y belongs to $Cl_t^{\leq}$", while $\overline{P}[Cl_t^{\leq}(x)]$ represents the credibility of the proposition "there is at least one $y \in U$ dominating x with respect to criteria from P which belongs to $Cl_t^{\leq}$".

Let us remark that, using the definition of the T^*-implication, it is possible to rewrite the definitions of $\underline{P}[Cl_t^{\geq}(x)]$, $\overline{P}[Cl_t^{\geq}(x)]$, $\underline{P}[Cl_t^{\leq}(x)]$ and $\overline{P}[Cl_t^{\leq}(x)]$, in the

following way:

$$\underline{P}\left[Cl_t^{\geq}(x)\right] = T_{y\in U}\left(I_{T^*,N}^{\rightarrow}\left(D_P(y,x), Cl_t^{\geq}(y)\right)\right) ,$$
$$\overline{P}\left[Cl_t^{\geq}(x)\right] = T_{y\in U}^{*}\left(N\left(I_{T^*,N}^{\rightarrow}\left(D_P(x,y), N\left(Cl_t^{\geq}(y)\right)\right)\right)\right) ,$$
$$\underline{P}\left[Cl_t^{\leq}(x)\right] = T_{y\in U}\left(I_{T^*,N}^{\rightarrow}\left(D_P(y,x), Cl_t^{\leq}(y)\right)\right) ,$$
$$\overline{P}\left[Cl_t^{\leq}(x)\right] = T_{y\in U}^{*}\left(N\left(I_{T^*,N}^{\rightarrow}\left(D_P(y,x), N\left(Cl_t^{\leq}(y)\right)\right)\right)\right) .$$

The following results can be proved:

1) for each $x \in U$ and for each $t \in \{1,\ldots,n\}$,

$$\underline{P}\left[Cl_t^{\geq}(x)\right] \leq Cl_t^{\geq}(x) \leq \overline{P}\left[Cl_t^{\geq}(x)\right] ,$$
$$\underline{P}\left[Cl_t^{\leq}(x)\right] \leq Cl_t^{\leq}(x) \leq \overline{P}\left[Cl_t^{\leq}(x)\right] ;$$

2) if (T, T^*, N) constitute a de Morgan triplet and if $N[Cl_t^{\geq}(x)] = Cl_{t-1}^{\leq}(x)$ for each $x \in U$ and $t = 2,\ldots,n$, then

$$\underline{P}\left[Cl_t^{\geq}(x)\right] = N\left(\overline{P}\left[Cl_{t-1}^{\leq}(x)\right]\right) ,$$
$$\overline{P}\left[Cl_t^{\geq}(x)\right] = N\left(\underline{P}\left[Cl_{t-1}^{\leq}(x)\right]\right) , \quad t = 2,\ldots,n ,$$
$$\underline{P}\left[Cl_t^{\leq}(x)\right] = N\left(\overline{P}\left[Cl_{t+1}^{\geq}(x)\right]\right) ,$$
$$\overline{P}\left[Cl_t^{\leq}(x)\right] = N\left(\underline{P}\left[Cl_{t+1}^{\geq}(x)\right]\right) , \quad t = 1,\ldots,n-1 ;$$

3) for all $P \subseteq R \subseteq C$, for all $x \in U$ and for each $t \in \{1,\ldots,n\}$,

$$\underline{P}\left[Cl_t^{\geq}(x)\right] \leq \underline{R}\left[Cl_t^{\geq}(x)\right] , \quad \overline{P}\left[Cl_t^{\geq}(x)\right] \geq \overline{R}\left[Cl_t^{\geq}(x)\right] ,$$
$$\underline{P}\left[Cl_t^{\leq}(x)\right] \leq \underline{R}\left[Cl_t^{\leq}(x)\right] , \quad \overline{P}\left[Cl_t^{\leq}(x)\right] \geq \overline{R}\left[Cl_t^{\leq}(x)\right] .$$

Results 1) to 3) can be read as fuzzy counterparts of the following results well-known within the classical rough set approach:

1) (*inclusion property*) says that $Cl_t^{\geq}$ and $Cl_t^{\leq}$ include their P-lower approximations and are included in their P-upper approximations;
2) (*complementarity property*) says that the P-lower (P-upper) approximation of $Cl_t^{\geq}$ is the complement of the P-upper (P-lower) approximation of its complementary set $Cl_{t-1}^{\leq}$, (analogous property holds for $Cl_t^{\leq}$ and $Cl_{t+1}^{\geq}$);
3) (*monotonicity with respect to sets of attributes*) says that enlarging the set of criteria, the membership to the lower approximation does not decrease and the membership to the upper approximation does not increase.

Greco, Inuiguchi, and Słowiński [14] proposed, moreover, the following fuzzy rough approximations based on dominance, which go in line with the fuzzy rough approximation by Dubois and Prade [3,5], concerning classical rough sets:

$$\underline{P}\left[Cl_t^{\geq}(x)\right] = \inf_{y\in U}\left(I\left(D_P(y,x), Cl_t^{\geq}(y)\right)\right) ,$$
$$\overline{P}\left[Cl_t^{\geq}(x)\right] = \sup_{y\in U}\left(T\left(D_P(x,y), Cl_t^{\geq}(y)\right)\right) ,$$
$$\underline{P}\left[Cl_t^{\leq}(x)\right] = \inf_{y\in U}\left(I\left(D_P(x,y), Cl_t^{\leq}(y)\right)\right) ,$$
$$\underline{P}\left[Cl_t^{\leq}(x)\right] = \sup_{y\in U}\left(T\left(D_P(y,x), Cl_t^{\leq}(y)\right)\right) .$$

Using fuzzy rough approximations based on DRSA, one can induce decision rules having the same syntax as the decision rules obtained from crisp DRSA. In this case, however, each decision rule has a fuzzy credibility.

Variable-Consistency Dominance-Based Rough Set Approach (VC-DRSA)

The definitions of rough approximations introduced in Sect. "Dominance-Based Rough Set Approach" are based on a strict application of the dominance principle. However, when defining non-ambiguous objects, it is reasonable to accept a limited proportion of negative examples, particularly for large data tables. Such an extended version of DRSA is called the variable-consistency DRSA model (VC-DRSA) [31].

For any $P \subseteq C$, $x \in U$ belongs to $Cl_t^{\geq}$ *without any ambiguity* at consistency level $l \in (0,1]$, if $x \in Cl_t^{\geq}$ and at least $l * 100\%$ of all objects $y \in U$ dominating x with respect to P also belong to $Cl_t^{\geq}$, i. e., for $t = 2,\ldots,n$,

$$\frac{\left|D_P^+(x) \cap Cl_t^{\geq}\right|}{\left|D_P^+(x)\right|} \geq l .$$

The level l is called *consistency level* because it controls the degree of consistency with respect to objects qualified as belonging to $Cl_t^{\geq}$ without any ambiguity. In other words, if $l < 1$, then at most $(1-l) * 100\%$ of all objects $y \in U$ dominating x with respect to P do not belong to $Cl_t^{\geq}$ and thus contradict the inclusion of x in $Cl_t^{\geq}$.

Analogously, for any $P \subseteq C$, $x \in U$ belongs to $Cl_t^{\leq}$ *without any ambiguity* at consistency level $l \in (0,1]$, if $x \in Cl_t^{\leq}$ and at least $l * 100\%$ of all the objects $y \in U$ dominated by x with respect to P also belong to $Cl_t^{\leq}$, i. e., for $t = 1,\ldots,n-1$,

$$\frac{\left|D_P^-(x) \cap Cl_t^{\leq}\right|}{\left|D_P^-(x)\right|} \geq l .$$

The concept of non-ambiguous objects at some consistency level l leads naturally to the corresponding definition of P-lower approximations of the unions of classes

$Cl_t^{\geq}$ and $Cl_t^{\leq}$, respectively:

$$\underline{P}^l\left(Cl_t^{\geq}\right) = \left\{x \in Cl_t^{\geq} : \frac{\left|D_P^+(x) \cap Cl_t^{\geq}\right|}{\left|D_P^+(x)\right|} \geq l\right\} ,$$
$$t = 2, \ldots, n ,$$
$$\underline{P}^l\left(Cl_t^{\leq}\right) = \left\{x \in Cl_t^{\leq} : \frac{\left|D_P^-(x) \cap Cl_t^{\leq}\right|}{\left|D_P^-(x)\right|} \geq l\right\} ,$$
$$t = 1, \ldots, n-1 .$$

Given $P \subseteq C$ and consistency level l, the corresponding *P-upper approximations* of $Cl_t^{\geq}$ and $Cl_t^{\leq}$, denoted by $\overline{P}^l(Cl_t^{\geq})$ and $\overline{P}^l(Cl_t^{\leq})$, respectively, can be defined as a complement of $\underline{P}^l(Cl_{t-1}^{\leq})$ and $\underline{P}^l(Cl_{t+1}^{\geq})$ with respect to U:

$$\overline{P}^l\left(Cl_t^{\geq}\right) = U - \underline{P}^l\left(Cl_{t-1}^{\leq}\right) , \quad t = 2, \ldots, n ,$$
$$\overline{P}^l\left(Cl_t^{\leq}\right) = U - \underline{P}^l\left(Cl_{t+1}^{\geq}\right) , \quad t = 1, \ldots, n-1 .$$

$\overline{P}^l(Cl_t^{\geq})$ can be interpreted as a set of all the objects belonging to $Cl_t^{\geq}$, *possibly ambiguous* at consistency level l. Analogously, $\overline{P}^l(Cl_t^{\leq})$ can be interpreted as a set of all the objects belonging to $Cl_t^{\leq}$, *possibly ambiguous* at consistency level l. The *P-boundaries* (P-doubtful regions) of $Cl_t^{\geq}$ and $Cl_t^{\leq}$ at consistency level l are defined as:

$$Bn_P^l\left(Cl_t^{\geq}\right) = \overline{P}^l\left(Cl_t^{\geq}\right) - \underline{P}^l\left(Cl_t^{\geq}\right) , \quad t = 2, \ldots, n$$
$$Bn_P^l\left(Cl_t^{\leq}\right) = \overline{P}^l\left(Cl_t^{\leq}\right) - \underline{P}^l\left(Cl_t^{\leq}\right) , \quad t = 1, \ldots, n-1 .$$

The *variable consistency* model of the dominance-based rough set approach provides some degree of flexibility in assigning objects to lower and upper approximations of the unions of decision classes. The following properties can be easily proved: for $0 < l' < l \leq 1$,

$$\underline{P}^l\left(Cl_t^{\geq}\right) \subseteq \underline{P}^{l'}\left(Cl_t^{\geq}\right) \quad \text{and} \quad \overline{P}^l\left(Cl_t^{\geq}\right) \supseteq \overline{P}^{l'}\left(Cl_t^{\geq}\right) ,$$
$$t = 2, \ldots, n ,$$
$$\underline{P}^l\left(Cl_t^{\leq}\right) \subseteq \underline{P}^{l'}\left(Cl_t^{\leq}\right) \quad \text{and} \quad \overline{P}^l\left(Cl_t^{\leq}\right) \supseteq \overline{P}^{l'}\left(Cl_t^{\leq}\right) ,$$
$$t = 1, \ldots, n-1 .$$

The following two basic types of variable-consistency decision rules can be considered:

1. $D_{\geq}$-decision rules with the following syntax: "if $\phi(x, q1) \geq r_{q1}$ and $\phi(x, q2) \geq r_{q2}$ and … $\phi(x, qp) \geq r_{qp}$, then $x \in Cl_t^{\geq}$" with confidence α (i. e. in fraction α of considered cases), where $P = \{q1, \ldots, qp\} \subseteq C$, $(r_{q1}, \ldots, r_{qp}) \in V_{q1} \times V_{q2} \times \ldots \times V_{qp}$ and $t = 2, \ldots, n$;
2. $D_{\leq}$-decision rules with the following syntax: "if $\phi(x, q1) \leq r_{q1}$ and $\phi(x, q2) \leq r_{q2}$ and … $\phi(x, qp) \leq r_{qp}$, then $x \in Cl_t^{\leq}$" with confidence α, where $P = \{q1, \ldots, qp\} \subseteq C$, $(r_{q1}, \ldots, r_{qp}) \in V_{q1} \times V_{q2} \times \ldots \times V_{qp}$ and $t = 1, \ldots, n-1$.

The *variable consistency* model is inspired by the *variable precision* model proposed by Ziarko [61,62] within the classical indiscernibility-based rough set approach.

Dominance-Based Rough Approximation of a Fuzzy Set

This section shows how the dominance-based rough set approach can be used for rough approximation of fuzzy sets.

A *fuzzy information base* is the 3-tuple $\boldsymbol{B} = \langle U, F, \varphi \rangle$, where U is a finite set of *objects* (universe), $F = \{f_1, f_2, \ldots, f_m\}$ is a finite set of *properties*, and $\varphi\colon U \times F \to [0, 1]$ is a function such that $\varphi(x, f_h) \in [0, 1]$ expresses the credibility that object x has property f_h. Each object x from U is described by a vector

$$\mathrm{Des}_F(x) = [\varphi(x, f_1), \ldots, \varphi(x, f_m)] ,$$

called *description* of x in terms of the degrees to which it has properties from F; it represents the available information about x. Obviously, $x \in U$ can be described in terms of any non-empty subset $E \subseteq F$ and in this case

$$\mathrm{Des}_E(x) = [\varphi(x, f_h), f_h \in E] .$$

For any $E \subseteq F$, the dominance relation D_E can be defined as follows: for all $x, y \in U$, x dominates y with respect to E (denotation xD_Ey) if, for any $f_h \in E$,

$$\varphi(x, f_h) \geq \varphi(y, f_h) .$$

Given $E \subseteq F$ and $x \in U$, let

$$D_E^+(x) = \{y \in U\colon yD_Ex\}, \quad D_E^-(x) = \{y \in U\colon xD_Ey\}.$$

Let us consider a fuzzy set X in U, with its membership function $\mu_X\colon U \to [0, 1]$. For each cutting level $\alpha \in [0, 1]$ and for $* \in \{\geq, >\}$, the E-lower and the E-upper approximation of $X^{*\alpha} = \{y \in U\colon \mu_X(y) * \alpha\}$ with respect to $E \subseteq F$ (denotation $\underline{E}(X^{*\alpha})$ and $\overline{E}(X^{*\alpha})$, respectively), can be defined as:

$$\underline{E}(X^{*\alpha}) = \left\{x \in U : D_E^+(x) \subseteq X^{*\alpha}\right\} ,$$
$$\overline{E}(X^{*\alpha}) = \left\{x \in U : D_E^-(x) \cap X^{*\alpha} \neq \emptyset\right\} .$$

Rough approximations $\underline{E}(X^{*\alpha})$ and $\overline{E}(X^{*\alpha})$ can be expressed in terms of unions of granules $D_E^+(x)$ as follows:

$$\underline{E}(X^{*\alpha}) = \bigcup_{x\in U} \left\{D_E^+(x) : D_E^+(x) \subseteq X^{*\alpha}\right\} ,$$
$$\overline{E}(X^{*\alpha}) = \bigcup_{x\in U} \left\{D_E^+(x) : D_E^-(x) \cap X^{*\alpha} \neq \emptyset\right\} .$$

Analogously, for each cutting level $\alpha \in [0,1]$ and for $\diamond \in \{\leq, <\}$, the E-lower and the E-upper approximation of $X^{\diamond\alpha} = \{y \in U : \mu_X(y) \diamond \alpha\}$, with respect to $E \subseteq F$ (denotation $\underline{E}(X^{\diamond\alpha})$ and $\overline{E}(X^{\diamond\alpha})$, respectively), can be defined as:

$$\underline{E}(X^{\diamond\alpha}) = \left\{x \in U : D_E^-(x) \subseteq X^{\diamond\alpha}\right\} ,$$
$$\overline{E}(X^{\diamond\alpha}) = \left\{x \in U : D_E^+(x) \cap X^{\diamond\alpha} \neq \emptyset\right\} .$$

The rough approximations $\underline{E}(X^{\diamond\alpha})$ and $\overline{E}(X^{\diamond\alpha})$ can be expressed in terms of unions of granules $D_E^-(x)$ as follows:

$$\underline{E}(X^{\diamond\alpha}) = \bigcup_{x\in U} \left\{D_E^-(x) : D_E^-(x) \subseteq X^{\diamond\alpha}\right\} ,$$
$$\overline{E}(X^{\diamond\alpha}) = \bigcup_{x\in U} \left\{D_E^-(x) : D_E^+(x) \cap X^{\diamond\alpha} \neq \emptyset\right\} .$$

Let us remark that the rough approximations $\underline{E}(X^{\geq\alpha})$, $\overline{E}(X^{\geq\alpha})$, $\underline{E}(X^{\leq\alpha})$ and $\overline{E}(X^{\leq\alpha})$ can be rewritten as follows:

$$\underline{E}(X^{\geq\alpha}) = \left\{x \in U : \forall w \in U, wD_Ex \Rightarrow w \in X^{\geq\alpha}\right\} ,$$
$$\overline{E}(X^{\geq\alpha}) = \left\{x \in U : \exists w \in U \text{ such that } xD_Ew \text{ and } w \in X^{\geq\alpha}\right\} ,$$
$$\underline{E}(X^{\leq\alpha}) = \left\{x \in U : \forall w \in U, xD_Ew \Rightarrow w \in X^{\leq\alpha}\right\} ,$$
$$\overline{E}(X^{\leq\alpha}) = \left\{x \in U : \exists w \in U \text{ such that } wD_Ex \text{ and } w \in X^{\leq\alpha}\right\} .$$

Rough approximations $\underline{E}(X^{>\alpha})$, $\overline{E}(X^{>\alpha})$, $\underline{E}(X^{<\alpha})$ and $\overline{E}(X^{<\alpha})$ can be rewritten analogously by a simple replacement of "$\geq$" with ">", and "$\leq$" with "<".

This reformulation of the rough approximations is concordant with the syntax of decision rules obtained in DRSA. For example, $\underline{E}(X^{\geq\alpha})$ is concordant with decision rules of the type

> "*if object y has property f_{i1} to degree at least h_{i1}, and has property f_{i2} to degree at least h_{i2}, …, and has property f_{ip} to degree at least h_{ip}, then object y belongs to set X to degree at least α*",

where $\{i1, \ldots, ip\} = E$ and $h_{i1} = \varphi(x, f_{i1}), \ldots, h_{ip} = \varphi(x, f_{ip})$.

Let us remark that in the above approximations, even if $X^{\geq\alpha} = Y^{\leq\alpha}$, their approximations are, in general, different due to the different directions of cutting the membership functions of X and Y. Of course, a similar remark holds also for $X^{>\alpha}$ and $Y^{<\alpha}$. Considerations of the directions in the cuts $X^{\geq\alpha}$, $X^{>\alpha}$ and $X^{\leq\alpha}$, $X^{<\alpha}$ are important in the definition of the rough approximations of unions and intersections of cuts.

The rough approximations $\underline{E}(X^{\geq\alpha})$, $\overline{E}(X^{\geq\alpha})$, $\underline{E}(X^{\leq\alpha})$, $\overline{E}(X^{\leq\alpha})$ and $\underline{E}(X^{>\alpha})$, $\overline{E}(X^{>\alpha})$, $\underline{E}(X^{<\alpha})$, $\overline{E}(X^{<\alpha})$ satisfy the following inclusion properties: for any $0 \leq \alpha \leq 1$,

$$\underline{E}(X^{\geq\alpha}) \subseteq X^{\geq\alpha} \subseteq \overline{E}(X^{\geq\alpha}) ,$$
$$\underline{E}(X^{\leq\alpha}) \subseteq X^{\leq\alpha} \subseteq \overline{E}(X^{\leq\alpha}) ,$$
$$\underline{E}(X^{>\alpha}) \subseteq X^{>\alpha} \subseteq \overline{E}(X^{>\alpha}) ,$$
$$\underline{E}(X^{<\alpha}) \subseteq X^{<\alpha} \subseteq \overline{E}(X^{<\alpha}) .$$

Furthermore, the following complementary properties hold: for any $0 \leq \alpha \leq 1$,

$$\underline{E}(X^{\geq\alpha}) = U - \overline{E}(X^{<\alpha}) , \quad \underline{E}(X^{\leq\alpha}) = U - \overline{E}(X^{>\alpha}) ,$$
$$\underline{E}(X^{>\alpha}) = U - \overline{E}(X^{\leq\alpha}) , \quad \underline{E}(X^{<\alpha}) = U - \overline{E}(X^{\geq\alpha}) .$$

The following properties of monotonicity with respect to sets of properties also hold: for any $E_1 \subseteq E_2 \subseteq F$ and for any $0 \leq \alpha \leq 1$,

$$\underline{E}_1(X^{\geq\alpha}) \subseteq \underline{E}_2(X^{\geq\alpha}) , \quad \underline{E}_1(X^{>\alpha}) \subseteq \underline{E}_2(X^{>\alpha}) ,$$
$$\underline{E}_1(X^{\leq\alpha}) \subseteq \underline{E}_2(X^{\leq\alpha}) , \quad \underline{E}_1(X^{<\alpha}) \subseteq \underline{E}_2(X^{<\alpha}) ,$$
$$\overline{E}_1(X^{\geq\alpha}) \supseteq \overline{E}_2(X^{\geq\alpha}) , \quad \overline{E}_1(X^{>\alpha}) \supseteq \overline{E}_2(X^{>\alpha}) ,$$
$$\overline{E}_1(X^{\leq\alpha}) \supseteq \overline{E}_2(X^{\leq\alpha}) , \quad \overline{E}_1(X^{<\alpha}) \supseteq \overline{E}_2(X^{<\alpha}) .$$

One can consider also fuzzy rough approximations $\underline{X}_E^{\uparrow}, \underline{X}_E^{\downarrow}, \overline{X}_E^{\uparrow}, \overline{X}_E^{\downarrow}$, which are fuzzy sets with membership functions defined, respectively, as follows: for any $y \in U$,

$$\mu_{\underline{X}_E^{\uparrow}}(y) = \max\left\{\alpha \in [0,1] : y \in \underline{E}(X^{\geq\alpha})\right\} ,$$
$$\mu_{\underline{X}_E^{\downarrow}}(y) = \min\left\{\alpha \in [0,1] : y \in \underline{E}(X^{\leq\alpha})\right\} ,$$
$$\mu_{\overline{X}_E^{\uparrow}}(y) = \max\left\{\alpha \in [0,1] : y \in \overline{E}(X^{\geq\alpha})\right\} ,$$
$$\mu_{\overline{X}_E^{\downarrow}}(y) = \min\left\{\alpha \in [0,1] : y \in \overline{E}(X^{\leq\alpha})\right\} .$$

The membership function $\mu_{\underline{X}_E^{\uparrow}}(y)$ is defined as the upward lower fuzzy rough approximation of X with respect to E and can be interpreted in the following way. For any $\alpha, \beta \in [0,1]$, $\alpha < \beta$ implies $X^{\geq\alpha} \supseteq X^{\geq\beta}$. Therefore, the greater the cutting level α, the smaller $X^{\geq\alpha}$ and, consequently, the smaller also its lower approximation $\underline{E}(X^{\geq\alpha})$. Thus, for each $y \in U$ and for each fuzzy set X, there is a threshold $k(y)$, $0 \leq k(y) \leq \mu_X(y)$, such that $y \in \underline{E}(X^{\geq\alpha})$ if $\alpha \leq k(y)$, and $y \notin \underline{E}(X^{\geq\alpha})$ if $\alpha > k(y)$.

Since $k(y) = \mu_{\underline{X}_E^{\uparrow}}(y)$, this explains the interest of $\mu_{\underline{X}_E^{\uparrow}}(y)$. Analogous interpretation holds for $\mu_{\overline{X}_E^{\uparrow}}(y)$, defined as the membership function of the upward upper fuzzy rough approximation of X with respect to E.

The membership function $\mu_{\underline{X}_E^{\downarrow}}(y)$ is defined as the downward lower fuzzy rough approximation of X with respect to E and can be interpreted as follows. For any $\alpha, \beta \in [0,1]$, $\alpha < \beta$ implies $X^{\leq\alpha} \subseteq X^{\leq\beta}$. Therefore, the greater the cutting level α, the greater $X^{\leq\alpha}$ and, consequently, its lower approximation $\underline{E}(X^{\geq\alpha})$. Thus, for each $y \in U$ and for each fuzzy set X, there is a threshold $h(y)$, $\mu_X(y) \leq h(y) \leq 1$, such that $y \in \underline{E}(X^{\leq\alpha})$ if $\alpha \geq h(y)$, and $y \notin \underline{E}(X^{\leq\alpha})$ if $\alpha < h(y)$. Observe that $h(y) = \mu_{\underline{X}_E^{\downarrow}}(y)$. Analogous interpretation holds for $\mu_{\overline{X}_E^{\downarrow}}(y)$, defined as the membership function of the downward upper fuzzy rough approximation of X with respect to E.

The membership functions of the upward and downward lower and upper fuzzy rough approximations can also be rewritten in the following equivalent formulations, which has been proposed and investigated by Greco, Inuiguchi, and Słowiński [16]:

$$\mu_{\underline{X}_E^{\uparrow}}(y) = \min\left\{\mu_X(z) : z \in D_E^{+}(y)\right\} ,$$
$$\mu_{\overline{X}_E^{\uparrow}}(y) = \max\left\{\mu_X(z) : z \in D_E^{-}(y)\right\} ,$$
$$\mu_{\underline{X}_E^{\downarrow}}(y) = \max\left\{\mu_X(z) : z \in D_E^{-}(y)\right\} ,$$
$$\mu_{\overline{X}_E^{\downarrow}}(y) = \min\left\{\mu_X(z) : z \in D_E^{+}(y)\right\} .$$

The membership functions of the fuzzy rough approximations $\mu_{\underline{X}_E^{\uparrow}}(y)$, $\mu_{\overline{X}_E^{\uparrow}}(y)$, $\mu_{\underline{X}_E^{\downarrow}}(y)$ and $\mu_{\overline{X}_E^{\downarrow}}(y)$, satisfy the following inclusion properties: for any $y \in U$,

$$\mu_{\underline{X}_E^{\uparrow}}(y) \leq \mu_X(y) \leq \mu_{\overline{X}_E^{\uparrow}}(y) ,$$
$$\mu_{\underline{X}_E^{\downarrow}}(y) \leq \mu_X(y) \leq \mu_{\overline{X}_E^{\downarrow}}(y) .$$

Furthermore, the following complementary property holds: for any $y \in U$,

$$\mu_{\underline{X}_E^{\uparrow}}(y) = \mu_{\overline{X}_E^{\downarrow}}(y) , \quad \mu_{\underline{X}_E^{\downarrow}}(y) = \mu_{\overline{X}_E^{\uparrow}}(y) .$$

The following properties of monotonicity with respect to sets of properties also hold: for any $E_1 \subseteq E_2 \subseteq F$,

$$\mu_{\underline{X}_{E_1}^{\uparrow}}(y) \leq \mu_{\underline{X}_{E_2}^{\uparrow}}(y) , \quad \mu_{\underline{X}_{E_1}^{\downarrow}}(y) \geq \mu_{\underline{X}_{E_2}^{\downarrow}}(y) ,$$
$$\mu_{\overline{X}_{E_1}^{\uparrow}}(y) \geq \mu_{\overline{X}_{E_2}^{\uparrow}}(y) , \quad \mu_{\overline{X}_{E_1}^{\downarrow}}(y) \leq \mu_{\overline{X}_{E_2}^{\downarrow}}(y) .$$

Monotonic Rough Approximation of a Fuzzy Set Versus Classical Rough Set

What is the relationship between the classical rough set and the DRSA approximation of a fuzzy set? Greco, Matarazzo, and Słowiński [27,29] proved that the former is a particular case of the latter, as shown below.

Let us remember that in the classical rough set approach [46,47], the original information is expressed by means of an *information system*, that is the 4-tuple $\boldsymbol{S} = \langle U, Q, V, \phi \rangle$, where U is a finite set of *objects* (universe), $Q = \{q_1, q_2, \ldots, q_m\}$ is a finite set of *attributes*, V_q is the set of values of the attribute q, $V = \bigcup_{q \in Q} V_q$ and $\phi\colon U \times Q \to V$ is a total function such that $\phi(x, q) \in V_q$ for each $q \in Q$, $x \in U$, called *information function*.

Therefore, each object x from U is described by a vector

$$\mathrm{Des}_Q(x) = [\phi(x, q_1), \phi(x, q_2), \ldots, \phi(x, q_m)] ,$$

called *description* of x in terms of the evaluations of the attributes from Q; it represents the available information about x. Obviously, $x \in U$ can be described in terms of any non-empty subset $P \subseteq Q$.

With every (non-empty) subset of attributes P there is associated an *indiscernibility relation* on U, denoted by I_P:

$$I_P = \{(x, y) \in U \times U\colon \phi(x, q) = \phi(y, q), \forall q \in P\} .$$

If $(x, y) \in I_P$, it is said that the objects x and y are P-indiscernible. Clearly, the indiscernibility relation thus defined is an equivalence relation (reflexive, symmetric, and transitive). The family of all the equivalence classes of the relation I_P is denoted by $U|I_P$, and the equivalence class containing an element $x \in U$ by $I_P(x)$, i. e.

$$I_P(x) = \{y \in U\colon \phi(y, q) = \phi(x, q), \forall q \in P\} .$$

The equivalence classes of the relation I_P are called *P-elementary sets*.

Let $\boldsymbol{S}$ be an information system, X a non-empty subset of U and $\emptyset \neq P \subseteq Q$. The *P-lower approximation* and the *P-upper approximation* of X in $\boldsymbol{S}$ are defined, respectively, as:

$$\underline{P}(X) = \{x \in U\colon I_P(x) \subseteq X\} ,$$
$$\overline{P}(X) = \{x \in U\colon I_P(x) \cap X \neq \emptyset\} .$$

The elements of $\underline{P}(X)$ are all and only those objects $x \in U$ which belong to the equivalence classes generated by the indiscernibility relation I_P, *contained* in X; the elements of $\overline{P}(X)$ are all and only those objects $x \in U$ which belong to the equivalence classes generated by the indiscernibility relation I_P, *containing at least one* object x belonging to X. In other words, $\underline{P}(X)$ is the largest union of the P-elementary sets included in X, while $\overline{P}(X)$ is the smallest union of the P-elementary sets containing X.

Granular Computing and Data Mining for Ordered Data: The Dominance-Based Rough Set Approach, Table 1
Information base *B*

	f_1	f_2
x_1	0	1
x_2	1	0
x_3	1	1

Any information system can be expressed in terms of a specific type of an information base. An *information base* is called *Boolean* if $\varphi\colon U \times F \to \{0,1\}$. A partition $\boldsymbol{F} = \{F_1, \ldots, F_r\}$ of the set of properties F, with *card* $(F_k) \geq 2$ for all $k = 1, \ldots, r$, is called *canonical* if, for each $x \in U$ and for each $F_k \in \boldsymbol{F}$, $k = 1, \ldots, r$, there exists only one $f_j \in F_k$ for which $\varphi(x, f_j) = 1$ (and, therefore, for each $f_i \in F_k - \{f_j\}$, $\varphi(x, f_i) = 0$). The condition *card* $(F_k) \geq 2$ for all $k = 1, \ldots, r$, is necessary because, otherwise, there would be at least one class $F_k = \{f'\}$ of the partition such that $\varphi(x, f') = 1$ for all $x \in U$, and this would mean that property f' gives no information and can be removed. Observe now that any *information system* $\boldsymbol{S} = \langle U, Q, V, \phi \rangle$ can be transformed to a Boolean information base $\boldsymbol{B} = \langle U, F, \varphi \rangle$ assigning to each $v \in V_q$, $q \in Q$, one property $f_{qv} \in F$ such that $\varphi(x, f_{qv}) = 1$ if $\phi(x, q) = v$, and $\varphi(x, f_{qv}) = 0$ otherwise. Let us remark that $\boldsymbol{F} = \{F_1, \ldots, F_r\}$, with $F_q = \{f_{qv}, v \in V_q\}$, $q \in Q$, is a canonical partition of F. The opposite transformation, from a Boolean information base to an information system, is not always possible, i. e. there may exist Boolean information bases which cannot be transformed into information systems, because their sets of properties do not admit any canonical partition, as shown by the following example.

Example 1 Let us consider a Boolean information base $\boldsymbol{B}$, such that $U = \{x_1, x_2, x_3\}$, $F = \{f_1, f_2\}$ and function φ is defined by Table 1. One can see that $\boldsymbol{F} = \{\{f_1, f_2\}\}$ is not a canonical partition because $\varphi(x_3, f_1) = \varphi(x_3, f_2) = 1$, while the definition of canonical partition $\boldsymbol{F}$ does not allow that for an object $x \in U$, $\varphi(x, f_1) = \varphi(x, f_2) = 1$. Therefore, this Boolean information base has no equivalent information system. Let us remark that also the Boolean information base $\boldsymbol{B}'$ presented in Table 2, where $U = \{x_1, x_2, x_4\}$ and $F = \{f_1, f_2\}$, cannot be transformed to an information system, because partition $\boldsymbol{F} = \{\{f_1, f_2\}\}$ is not canonical. Indeed, $\varphi(x_4, f_1) = \varphi(x_4, f_2) = 0$, while definition of canonical partition $\boldsymbol{F}$ does not allow that for an object $x \in U$, $\varphi(x, f_1) = \varphi(x, f_2) = 0$.

The above says that consideration of rough approximation in the context of a Boolean information base is more general than the same consideration in the context of an information system. This means, of course, that the rough approximation considered in the context of a fuzzy information base is yet more general.

Granular Computing and Data Mining for Ordered Data: The Dominance-Based Rough Set Approach, Table 2
Information base B'

	f_1	f_2
x_1	0	1
x_2	1	0
x_4	0	0

Granular Computing and Data Mining for Ordered Data: The Dominance-Based Rough Set Approach, Table 3
Information system *S*

	q_1	q_2
x_1	0	1
x_2	1	0
x_3	1	1

It is worth stressing that the Boolean information bases $\boldsymbol{B}$ and $\boldsymbol{B}'$ are not Boolean information systems. In fact, on one hand, a Boolean information base provides information about absence ($\varphi(x, f) = 0$) or presence ($\varphi(x, f) = 1$) of properties $f \in F$ in objects $x \in U$. On the other hand, a Boolean information system provides information about values assigned by attributes $q \in Q$, whose sets of values are $V_q = \{0, 1\}$, to objects $x \in U$, such that $\phi(x, q) = 1$ or $\phi(x, q) = 0$ for all $x \in U$ and $q \in Q$. Observe, therefore, that to transform a Boolean information system $\boldsymbol{S}$ into a Boolean information base $\boldsymbol{B}$, each attribute q of $\boldsymbol{S}$ corresponds to two properties f_{q0} and f_{q1} of $\boldsymbol{B}$, such that for all $x \in U$

- $\varphi(x, f_{q0}) = 1$ and $\varphi(x, f_{q1}) = 0$ if $\phi(x, q) = 0$,
- $\varphi(x, f_{q0}) = 0$ and $\varphi(x, f_{q1}) = 1$ if $\phi(x, q) = 1$.

Thus, the Boolean information base $\boldsymbol{B}$ in Table 1 and the Boolean information system $\boldsymbol{S}$ in Table 3 are different, though they could seem identical. In fact, the Boolean information system $\boldsymbol{S}$ in Table 3 can be transformed into the Boolean information base $\boldsymbol{B}''$ in Table 4, which is clearly different from $\boldsymbol{B}$.

The equivalence between rough approximations in the context of a fuzzy information base and the classical definition of rough approximations in the context of an information system can be stated as follows [27,29]. Let us consider an information system and the corresponding Boolean information base; for each $P \subseteq Q$, let E^P be the

Granular Computing and Data Mining for Ordered Data: The Dominance-Based Rough Set Approach, Table 4
Information base B''

	$f_{q_1 0}$	$f_{q_1 1}$	$f_{q_2 0}$	$f_{q_2 1}$
x_1	1	0	0	1
x_2	0	1	1	0
x_3	0	1	0	1

set of all the properties corresponding to values v of attributes in P. Let X be a non-fuzzy (classical) set in U (i. e. $\mu_X : U \to \{0, 1\}$ and, therefore, for any $y \in U$, $\mu_X(y) = 1$ or $\mu_X(y) = 0$); then it is:

$$\underline{E^P}(X^{\geq 1}) = \underline{P}(X^{\geq 1}) , \quad \overline{E^P}(X^{\geq 1}) = \overline{P}(X^{\geq 1}) ,$$
$$\underline{E^P}(X^{\leq 0}) = \underline{P}(U - X^{\geq 1}) , \quad \overline{E^P}(X^{\leq 0}) = \overline{P}(U - X^{\geq 1}) .$$

This result proves that the rough approximation of a non-fuzzy set X in a Boolean information base admitting a canonical partition is equivalent to the classical rough approximation of set X in the corresponding information system. Therefore, the classical rough approximation is a particular case of the dominance-based rough approximation in a fuzzy information base.

Dominance-Based Rough Set Approach to Case-Based Reasoning

This section presents rough approximation of a fuzzy set using a similarity relation in the context of case-based reasoning [28].

Case-based reasoning (for a general introduction to case-based reasoning see e. g. [35]; for a fuzzy set approach to case-based reasoning see [6]) is a paradigm in machine learning whose idea is that a new problem can be solved by noticing its similarity to a set of problems previously solved. Case-based reasoning regards the inference of some proper conclusions related to a new situation by the analysis of similar cases from a memory of previous cases. It is based on two principles [36]:

a) Similar problems have similar solutions;
b) Types of encountered problems tend to recur.

Gilboa and Schmeidler [12] observed that the basic idea of case-based reasoning can be found in the following sentence of Hume [33]: "From causes which appear *similar* we expect similar effects. This is the sum of all our experimental conclusions." Rephrasing Hume, one can say that "the more similar are the causes, the more similar one expects the effects." Therefore, measuring similarity is the essential point of all case-based reasoning and, particularly, of the fuzzy set approach to case-based reasoning [6]. This explains the many problems that measuring similarity generates within case-based reasoning. Problems of modeling similarity are relative to two levels:

- At the level of similarity with respect to single features: how to define a meaningful similarity measure with respect to a single feature?
- At the level of similarity with respect to all features: how to properly aggregate the similarity measures with respect to single features in order to obtain a comprehensive similarity measure?

For the above reasons, [28] proposes a DRSA approach to case-based reasoning, which tries to be possibly "neutral" and "objective" with respect to similarity relation. At the level of similarity concerning single features, the DRSA approach to case-based reasoning considers only ordinal properties of similarity, and at the level of aggregation, it does not impose any particular functional aggregation based on some very specific axioms (see, for example, [12]), but it considers a set of decision rules based on the general monotonicity property of comprehensive similarity with respect to similarity of single features. Therefore, the DRSA approach to case-based reasoning is only very little "invasive", compared to the many other existing approaches.

Let us consider a *pairwise fuzzy information base* being the 3-tuple

$$\boldsymbol{B} = \langle U, F, \sigma \rangle ,$$

where U is a finite set of *objects* (universe), $F = \{f_1, f_2, \ldots, f_m\}$ is a finite set of *features*, and $\sigma : U \times U \times F \to [0, 1]$ is a function such that $\sigma(x, y, f_h) \in [0, 1]$ expresses the credibility that object x is similar to object y w.r.t. feature f_h. The minimal requirement function σ must satisfy is that, for all $x \in U$ and for all $f_h \in F$, $\sigma(x, x, f_h) = 1$. Therefore, each pair of objects $(x, y) \in U \times U$ is described by a vector

$$\mathrm{Des}_F(x, y) = [\sigma(x, y, f_1), \ldots, \sigma(x, y, f_m)] ,$$

called *description* of (x, y) in terms of the credibilities of similarity with respect to features from F; it represents the available information about similarity between x and y. Obviously, similarity between x and y, $x, y \in U$, can be described in terms of any non-empty subset $E \subseteq F$ as follows:

$$\mathrm{Des}_E(x, y) = [\sigma(x, y, f_h), f_h \in E] .$$

With respect to any $E \subseteq F$, the dominance relation D_E can be defined on $U \times U$ as follows: for any $x, y, w, z \in U$,

(x, y) dominates (w, z) with respect to E (denotation $(x, y)D_E(w, z)$) if, for any $f_h \in E$,

$$\sigma(x, y, f_h) \geq \sigma(w, z, f_h) .$$

Given $E \subseteq F$ and $x, y \in U$, let

$$D_E^+(y, x) = \{w \in U\colon (w, x)D_E(y, x)\} ,$$
$$D_E^-(y, x) = \{w \in U\colon (y, x)D_E(w, x)\} .$$

In the pair (y, x), x is considered to be a *reference object*, while y can be called a *limit object*, because it is conditioning the membership of w in $D_E^+(y, x)$ and in $D_E^-(y, x)$.

For each $x \in U$ and $\alpha \in [0, 1]$ and $* \in \{\geq, >\}$, the lower approximation of $X^{*\alpha}$, $\underline{E}_\sigma(X^{*\alpha})$, and the upper approximation of $X^{*\alpha}$, $\overline{E}_\sigma(X^{*\alpha})$, based on similarity σ with respect to $E \subseteq F$ and x, respectively, can be defined as:

$$\underline{E}(x)_\sigma(X^{*\alpha}) = \left\{y \in U : D_E^+(y, x) \subseteq X^{*\alpha}\right\} ,$$
$$\overline{E}(x)_\sigma(X^{*\alpha}) = \left\{y \in U : D_E^-(y, x) \cap X^{*\alpha} \neq \emptyset\right\} .$$

For the sake of simplicity, in the following only $\underline{E}(x)_\sigma(X^{\geq\alpha})$ and $\overline{E}(x)_\sigma(X^{\geq\alpha})$ with $x \in X^{\geq\alpha}$ are considered. Of course, analogous considerations hold for $\underline{E}(x)_\sigma(X^{>\alpha})$ and $\overline{E}(x)_\sigma(X^{>\alpha})$. Observe that the lower approximation of $X^{\geq\alpha}$ with respect to x contains all the objects $y \in U$ such that any object w, being similar to x at least as much as y is similar to x w.r.t. all the considered features $E \subseteq F$, also belongs to $X^{\geq\alpha}$. Thus, the data from the fuzzy pairwise information base $\boldsymbol{B}$ confirm that if w is similar to x not less than $y \in \underline{E}(x)_\sigma(X^{\geq\alpha})$ is similar to x w.r.t. all the considered features $E \subseteq F$, then w belongs to $X^{\geq\alpha}$. In other words, x is a reference object and $y \in \underline{E}(x)_\sigma(X^{\geq\alpha})$ is a limit object which belongs "certainly" to set X with credibility at least α; the limit is understood such that all objects w that are similar to x w.r.t. considered features at least as much as y is similar to x, also belong to X with credibility at least α.

Analogously, the upper approximation of $X^{\geq\alpha}$ with respect to x contains all objects $y \in U$ such that there is at least one object w, being similar to x at most as much as y is similar to x w.r.t. all the considered features $E \subseteq F$, which belongs to $X^{\geq\alpha}$. Thus, the data from the fuzzy pairwise information base $\boldsymbol{B}$ confirm that if w is similar to x not less than $y \in \overline{E}(x)_\sigma(X^{\geq\alpha})$ is similar to x w.r.t. all the considered features $E \subseteq F$, then it is possible that w belongs to $X^{\geq\alpha}$. In other words, x is a reference object and $y \in \overline{E}(x)_\sigma(X^{\geq\alpha})$ is a limit object which belongs "possibly" to set X with credibility at least α; the limit is understood such that all objects $z \in U$ similar to x not less than y w.r.t. considered features, possibly belong to $X^{\geq\alpha}$.

For each $x \in U$ and $\alpha \in [0, 1]$ and $\diamond \in \{\leq, <\}$, the lower approximation of $X^{\diamond\alpha}$, $\underline{E}(x)_\sigma(X^{\diamond\alpha})$, and the upper approximation of $X^{\diamond\alpha}$, $\overline{E}(x)_\sigma(X^{\diamond\alpha})$, based on similarity σ with respect to $E \subseteq F$ and x, respectively, can be defined as:

$$\underline{E}(x)_\sigma(X^{\diamond\alpha}) = \left\{y \in U : D_E^-(y, x) \subseteq X^{\diamond\alpha}\right\} ,$$
$$\overline{E}(x)_\sigma(X^{\diamond\alpha}) = \left\{y \in U : D_E^+(y, x) \cap X^{\diamond\alpha} \neq \emptyset\right\} .$$

For the sake of simplicity, in the following only $\underline{E}(x)_\sigma(X^{\leq\alpha})$ and $\overline{E}(x)_\sigma(X^{\leq\alpha})$ with $x \in X^{\leq\alpha}$ are considered. Of course, analogous considerations hold for $\underline{E}(x)_\sigma(X^{<\alpha})$ and $\overline{E}(x)_\sigma(X^{<\alpha})$. Observe that the lower approximation of $X^{\leq\alpha}$ with respect to x contains all the objects $y \in U$ such that any object w, being similar to x at most as much as y is similar to x w.r.t. all the considered features $E \subseteq F$, also belongs to $X^{\leq\alpha}$. Thus, the data from the fuzzy pairwise information base $\boldsymbol{B}$ confirm that if w is similar to x not more than $y \in \underline{E}(x)_\sigma(X^{\leq\alpha})$ is similar to x w.r.t. all the considered features $E \subseteq F$, then w belongs to $X^{\leq\alpha}$. In other words, x is a reference object and $y \in \underline{E}(x)_\sigma(X^{\leq\alpha})$ is a limit object which belongs "certainly" to set X with credibility at most α; the limit is understood such that all objects w that are similar to x w.r.t. considered features at most as much as y is similar to x, also belong to X with credibility at most α.

Analogously, the upper approximation of $X^{\leq\alpha}$ with respect to x contains all the objects $y \in U$ such that there is at least one object w, being similar to x at least as much as y is similar to x with respect to all the considered features $E \subseteq F$, which belongs to $X^{\leq\alpha}$. Thus, the data from the fuzzy pairwise information base B confirm that if w is similar to x not more than $y \in \overline{E}(x)_\sigma(X^{\leq\alpha})$ is similar to x w.r.t. all the considered features $E \subseteq F$, then it is possible that w belongs to $X^{\leq\alpha}$. In other words, x is a reference object and $y \in \overline{E}_\sigma(X^{\leq\alpha})$ is a limit object which belongs "possibly" to set X with credibility at most α; the limit is understood such that all objects $z \in U$ similar to x not more than y w.r.t. considered features, possibly belong to $X^{\leq\alpha}$.

Observe that the rough approximations $\underline{E}(x)_\sigma(X^{\geq\alpha})$, $\overline{E}(x)_\sigma(X^{\geq\alpha})$, $\underline{E}(x)_\sigma(X^{\leq\alpha})$ and $\overline{E}(x)_\sigma(X^{\leq\alpha})$ can be rewritten as

$$\underline{E}(x)_\sigma(X^{\geq\alpha}) = \left\{y \in U : \forall w \in U, (w, x)D_E(y, x) \Rightarrow w \in X^{\geq\alpha}\right\} ,$$
$$\overline{E}(x)_\sigma(X^{\geq\alpha}) = \left\{y \in U : \exists w \in U \text{ such that } (y, x)D_E(w, x) \text{ and } w \in X^{\geq\alpha}\right\} ,$$
$$\underline{E}(x)_\sigma(X^{\leq\alpha}) = \left\{y \in U : \forall w \in U, (y, x)D_E(w, x) \Rightarrow w \in X^{\leq\alpha}\right\} ,$$
$$\overline{E}(x)_\sigma(X^{\leq\alpha}) = \left\{y \in U : \exists w \in U \text{ such that } (w, x)D_E(y, x) \text{ and } w \in X^{\leq\alpha}\right\} .$$

This formulation of the rough approximation is concordant with the syntax of the decision rules induced by means of DRSA from a fuzzy pairwise information base. More precisely,

- $\underline{E}(x)_\sigma(X^{\geq\alpha})$ is concordant with decision rules of the type: "if object w is similar to object x w.r.t. feature f_{i1} to degree at least h_{i1} and w.r.t. feature f_{i2} to degree at least h_{i2} and … and w.r.t. feature f_{ip} to degree at least h_{ip}, then object w belongs to set X to degree at least α",
- $\overline{E}(x)_\sigma(X^{\geq\alpha})$ is concordant with decision rules of the type: "if object w is similar to object x w.r.t. feature f_{i1} to degree at least h_{i1} and w.r.t. feature f_{i2} to degree at least h_{i2} and … and w.r.t. feature f_{ip} to degree at least h_{ip}, then object w could belong to set X to degree at least α",
- $\underline{E}(x)_\sigma(X^{\leq\alpha})$ is concordant with decision rules of the type: "if object w is similar to object x w.r.t. feature f_{i1} to degree at most h_{i1} and w.r.t. feature f_{i2} to degree at most h_{i2} and … and w.r.t. feature f_{ip} to degree at most h_{ip}, then object w belongs to set X to degree at most α",
- $\overline{E}(x)_\sigma(X^{\leq\alpha})$ is concordant with decision rules of the type: "if object w is similar to object x w.r.t. feature f_{i1} to degree at most h_{i1} and w.r.t. feature f_{i2} to degree at most h_{i2} and … and w.r.t. feature f_{ip} to degree at most h_{ip}, then object w could belong to set X to degree at most α", where $\{i1, \ldots, ip\} = E$ and $h_{i1}, \ldots, h_{ip} \in [0, 1]$.

The above definitions of rough approximations and the syntax of decision rules are based on ordinal properties of similarity relations only. In fact, no algebraic operation, such as sum or product, involving cardinal properties of function σ measuring credibility of similarity relations is considered. This is an important characteristic of our approach in comparison with alternative approaches to case-based reasoning.

Let us remark that, similarly to DRSA approximation in an information base, in the case of DRSA approximation in a fuzzy pairwise information base, even if for two fuzzy sets X and Y it is $X^{\geq\alpha} = Y^{\leq\alpha}$, their approximations may be different due to the different directions of cutting the membership functions of sets X and Y.

Rough approximations in a fuzzy pairwise information base satisfy the following interesting properties.

Theorem 1 *Given a fuzzy pairwise information base $\boldsymbol{B} = \langle U, F, \sigma\rangle$ and a fuzzy set X in U with membership function $\mu_X(\cdot)$, the following properties hold for any $E \subseteq F$:*

1. For any $0 \leq \alpha \leq 1$,

$$\underline{E}(x)_\sigma(X^{\leq\alpha}) \subseteq X^{\leq\alpha} \subseteq \overline{E}(x)_\sigma(X^{\leq\alpha}),$$
$$\underline{E}(x)_\sigma(X^{\geq\alpha}) \subseteq X^{\geq\alpha} \subseteq \overline{E}(x)_\sigma(X^{\geq\alpha}),$$
$$\underline{E}(x)_\sigma(X^{<\alpha}) \subseteq X^{<\alpha} \subseteq \overline{E}(x)_\sigma(X^{<\alpha}),$$
$$\underline{E}(x)_\sigma(X^{>\alpha}) \subseteq X^{>\alpha} \subseteq \overline{E}(x)_\sigma(X^{>\alpha}).$$

2. For any $0 \leq \alpha \leq 1$,

$$\underline{E}(x)_\sigma(X^{\leq\alpha}) = U - \overline{E}(x)_\sigma(X^{>\alpha}),$$
$$\underline{E}(x)_\sigma(X^{\geq\alpha}) = U - \overline{E}(x)_\sigma(X^{<\alpha}).$$

3. For any $0 \leq \alpha \leq \beta \leq 1$,

$$\underline{E}(x)_\sigma(X^{\leq\alpha}) \subseteq \underline{E}(x)_\sigma(X^{\leq\beta}),$$
$$\underline{E}(x)_\sigma(X^{<\alpha}) \subseteq \underline{E}(x)_\sigma(X^{<\beta}),$$
$$\underline{E}(x)_\sigma(X^{\geq\alpha}) \supseteq \underline{E}(x)_\sigma(X^{\geq\beta}),$$
$$\underline{E}(x)_\sigma(X^{>\alpha}) \supseteq \underline{E}(x)_\sigma(X^{>\beta}),$$
$$\overline{E}(x)_\sigma(X^{\leq\alpha}) \subseteq \overline{E}(x)_\sigma(X^{\leq\beta}),$$
$$\overline{E}(x)_\sigma(X^{<\alpha}) \subseteq \overline{E}(x)_\sigma(X^{<\beta}),$$
$$\overline{E}(x)_\sigma(X^{\geq\alpha}) \supseteq \overline{E}(x)_\sigma(X^{\geq\beta}),$$
$$\overline{E}(x)_\sigma(X^{>\alpha}) \supseteq \overline{E}(x)_\sigma(X^{>\beta}).$$

4. For any $x, y, w, z \in U$ and for any $0 \leq \alpha \leq 1$,

$$\left[(y,x)D_E(w,x) \text{ and } w \in \underline{E}(x)_\sigma(X^{\geq\alpha})\right] \Rightarrow y \in \underline{E}(x)_\sigma(X^{\geq\alpha}),$$
$$\left[(y,x)D_E(w,x) \text{ and } w \in \underline{E}(x)_\sigma(X^{>\alpha})\right] \Rightarrow y \in \underline{E}(x)_\sigma(X^{>\alpha}),$$
$$\left[(y,x)D_E(w,x) \text{ and } w \in \overline{E}(x)_\sigma(X^{\geq\alpha})\right] \Rightarrow y \in \overline{E}(x)_\sigma(X^{\geq\alpha}),$$
$$\left[(y,x)D_E(w,x) \text{ and } w \in \overline{E}(x)_\sigma(X^{>\alpha})\right] \Rightarrow y \in \overline{E}(x)_\sigma(X^{>\alpha}),$$
$$\left[(w,x)D_E(y,x) \text{ and } w \in \underline{E}(x)_\sigma(X^{\leq\alpha})\right] \Rightarrow y \in \underline{E}(x)_\sigma(X^{\leq\alpha}),$$
$$\left[(w,x)D_E(y,x) \text{ and } w \in \underline{E}(x)_\sigma(X^{<\alpha})\right] \Rightarrow y \in \underline{E}(x)_\sigma(X^{<\alpha}),$$
$$\left[(w,x)D_E(y,x) \text{ and } w \in \overline{E}(x)_\sigma(X^{\leq\alpha})\right] \Rightarrow y \in \overline{E}(x)_\sigma(X^{\leq\alpha}),$$
$$\left[(w,x)D_E(y,x) \text{ and } w \in \overline{E}(x)_\sigma(X^{<\alpha})\right] \Rightarrow y \in \overline{E}(x)_\sigma(X^{<\alpha}).$$

5. For any $E_1 \subseteq E_2 \subseteq F$ and for any $0 \le \alpha \le 1$,

$$\underline{E_1}(x)_\sigma(X^{\le\alpha}) \subseteq \underline{E_2}(x)_\sigma(X^{\le\alpha}),$$
$$\underline{E_1}(x)_\sigma X^{<\alpha}) \subseteq \underline{E_2}(x)_\sigma(X^{<\alpha}),$$
$$\underline{E_1}(x)_\sigma(X^{\ge\alpha}) \subseteq \underline{E_2}(x)_\sigma(X^{\ge\alpha}),$$
$$\underline{E_1}(x)_\sigma(X^{>\alpha}) \subseteq \underline{E_2}(x)_\sigma(X^{>\alpha}),$$
$$\overline{E_1}(x)_\sigma(X^{\le\alpha}) \supseteq \overline{E_2}(x)_\sigma(X^{\le\alpha}),$$
$$\overline{E_1}(x)_\sigma(X^{<\alpha}) \supseteq \overline{E_2}(x)_\sigma(X^{<\alpha}),$$
$$\overline{E_1}(x)_\sigma(X^{\ge\alpha}) \supseteq \overline{E_2}(x)_\sigma(X^{\ge\alpha}),$$
$$\overline{E_1}(x)_\sigma(X^{>\alpha}) \supseteq \overline{E_2}(x)_\sigma(X^{>\alpha}).$$

An Algebraic Structure for Dominance-Based Rough Set Approach

This section presents an algebraic characterization of DRSA in terms of the bipolar complemented de Morgan Brouwer–Zadeh distributive lattice [30], being a generalization of the de Morgan Brouwer–Zadeh distributive lattice [2], already proposed to characterize the classical rough set approach [1].

A system $\langle \Sigma, \Sigma^+, \Sigma^-, \wedge, \vee, '^+, '^-, \sim^+, \sim^-, 0, 1\rangle$ is a *bipolar complemented quasi-Brouwer–Zadeh distributive lattice* if the following properties (1b)–(4b) hold:

(1b) Σ is a distributive lattice with respect to the join and the meet operations $\vee$ and $\wedge$

(1b') $\Sigma^+, \Sigma^- \subseteq \Sigma$ are distributive lattices with respect to the join and the meet operations $\vee$ and $\wedge$. Σ is bounded by the least element 0 and the greatest element 1, which implies that also Σ^+ and Σ^- are bounded.

(2b) The unary operations $'^+ \colon \Sigma^+ \to \Sigma^-$ and $'^- \colon \Sigma^- \to \Sigma^+$ are Kleene (also Zadeh or fuzzy) bipolar complementations, that is for arbitrary $a, b \in \Sigma^+$ and $c, d \in \Sigma^-$

(K1b) $a'^{+'^-} = a, \quad c'^{-'^+} = c,$
(K2b) $(a \vee b)'^+ = a'^+ \wedge b'^+,$
$(c \vee d)'^- = c'^- \wedge d'^-,$
(K3b) $a \wedge a'^+ \le b \vee b'^+, \quad c \wedge c'^- \le d \vee d'^-.$

(3b) The unary operations $\sim^+ \colon \Sigma^+ \to \Sigma^-$ and $\sim^- \colon \Sigma^- \to \Sigma^+$ are Brouwer (or intuitionistic) bipolar complementations, that is for arbitrary $a, b \in \Sigma^+$ and $c, d \in \Sigma^-$

(B1b) $a \wedge a^{\sim^+\sim^-} = a, \quad c \wedge c^{\sim^-\sim^+} = c$
(B2b) $(a \vee b)^{\sim^+} = a^{\sim^+} \wedge b^{\sim^+},$
$(c \vee d)^{\sim^-} = c^{\sim^-} \wedge d^{\sim^-},$
(B3b) $a \wedge a^{\sim^+} = 0, \quad c \wedge c^{\sim^-} = 0.$

(4b) Complementation $'^+$ and complementation $\sim^+$ in one hand and complementation $'^-$ and complementation $\sim^-$ in the other hand are linked by the interconnection rule, that is, for arbitrary $a \in \Sigma^+$ and arbitrary $b \in \Sigma^-$: *(in-b)* $a^{\sim^+} \le a'^+$, $b^{\sim^-} \le b'^-$.

A structure $\langle \Sigma, \Sigma^+, \Sigma^-, \wedge, \vee, '^+, '^-, \sim^+, \sim^-, 0, 1\rangle$ is a *bipolar complemented Brouwer–Zadeh distributive lattice* if it is a quasi-Brouwer–Zadeh distributive lattice satisfying the stronger interconnection rule, that is, for arbitrary $a \in \Sigma^+$ and arbitrary $b \in \Sigma^-$: *(s-in-b)* $a^{\sim^+\sim^-} = a^{\sim^+'^-}$, $b^{\sim^-\sim^+} = b^{\sim^-'^+}$.

A bipolar complemented Brouwer–Zadeh distributive lattice is a *bipolar complemented de Morgan Brouwer–Zadeh distributive lattice*, if it satisfies also the $\vee$ de Morgan property that is, for arbitrary $a, b \in \Sigma^+$ and $c, d \in \Sigma^-$:

(B2a-b) $(a \wedge b)^{\sim^+} = a^{\sim^+} \vee b^{\sim^+}$
$(c \wedge d)^{\sim^-} = c^{\sim^-} \vee d^{\sim^-}.$

The bipolar complemented de Morgan Brouwer–Zadeh distributive lattice is an algebraic structure which can be given to the collection of all rough approximations within the dominance-based rough set approach as follows. Let us consider a fuzzy information base $\boldsymbol{B} = \langle U, F, \varphi\rangle$. Given $G \subseteq F$, for each $X \subseteq U$, its *upward lower approximation* $\underline{G}^{(>)}(X)$ and its *upward upper approximation* $\overline{G}^{(>)}(X)$ can be defined as:

$$\underline{G}^{(>)}(X) = \{x \in U \colon D_G^+(x) \subseteq X\},$$
$$\overline{G}^{(>)}(X) = \{x \in U \colon D_G^-(x) \cap X \ne \emptyset\}.$$

Analogously, given $G \subseteq F$, for each $X \subseteq U$, its *downward lower approximation* $\underline{G}^{(<)}(X)$ and its *downward upper approximation* $\overline{G}^{(<)}(X)$ can be defined as:

$$\underline{G}^{(<)}(X) = \{x \in U \colon D_G^-(x) \subseteq X\},$$
$$\overline{G}^{(<)}(X) = \{x \in U \colon D_G^+(x) \cap X \ne \emptyset\}.$$

Fixed $G \subseteq F$, for any $X \subseteq U$ let us consider the pairs $\langle \underline{G}^{(<)}(X), U - \overline{G}^{(<)}(X)\rangle$ and $\langle \underline{G}^{(>)}(X), U - \overline{G}^{(>)}(X)\rangle$ and the sets

$$B = \{(I, E) \colon I, E \subseteq U \text{ such that } I \cap E = \emptyset\},$$
$$B^- = \{(I, E) \colon \exists X \subseteq U \text{ for which } I = \underline{G}^{(<)}(X) \text{ and } E = U - \overline{G}^{(<)}(X)\},$$

$$B^+ = \left\{(I,E)\colon \exists X \subseteq U \text{ for which } I = \underline{G}^{(>)}(X) \text{ and } E = U - \overline{G}^{(>)}(X)\right\}.$$

The following result [30] holds.

Theorem 2 *The structure* $\langle B, B^+, B^-, \sqcap, \sqcup, ^{--}, ^{-+}, ^{\approx -}, ^{\approx +}, \langle \emptyset, U\rangle, \langle U, \emptyset\rangle\rangle$ *where, for any* $\langle I_1, E_1\rangle, \langle I_2, E_2\rangle \in B$, $\langle I_3, E_3\rangle \in B^-$, $\langle I_4, E_4\rangle \in B^+$,

$$\langle I_1, E_1\rangle \sqcap \langle I_2, E_2\rangle = \langle I_1 \cap I_2, E_1 \cup E_2\rangle$$

$$\langle I_1, E_1\rangle \sqcup \langle I_2, E_2\rangle = \langle I_1 \cup I_2, E_1 \cap E_2\rangle$$

$$\langle I_3, E_3\rangle^{--} = \langle E_3, I_3\rangle,\ \langle I_4, E_4\rangle^{-+} = \langle E_4, I_4\rangle$$

$$\langle I_3, E_3\rangle^{\approx -} = \langle E_3, U - E_3\rangle,\ \langle I_4, E_4\rangle^{\approx +} = \langle E_4, U - E_4\rangle$$

is a bipolar complemented de Morgan Brouwer–Zadeh distributive lattice.

Let us observe that within the bipolar complemented de Morgan Brouwer–Zadeh distributive lattice, Σ^+ has to be interpreted as the set of "positive" concepts, while Σ^- as the set of "negative" concepts. For example, in a problem of evaluation of students, the concept of "good students" is "positive", while the concept of "bad students" is "negative". Within DRSA, each concept is represented by the pair (I, E), where I (the interior) is the lower approximation of a set $X \subseteq U$ and E (the exterior) is the complement in U of the upper approximation of X. Intuitively, each concept is represented by I, being the set of objects that surely belong to the concept, and E being the set of objects that surely does not belong to the concept. Therefore, the positive concept "good students" is represented by the pair (I_G, E_G), where I_G represents the set of students "surely good" and E_G represents the set of students "surely not good". Analogously, the negative concept "bad students" is represented by the pair (I_B, E_B), where I_B represents the set of students "surely bad" and E_B represents the set of students "surely not bad". To illustrate the concept of join $\sqcup$ and the meet $\sqcap$, let us consider the concepts of "good students in Mathematics", represented by the pair (I_{G_M}, E_{G_M}), and "good students in Literature", represented by the pair (I_{G_L}, E_{G_L}). In this context, the concept of "good students in Mathematics *or* Literature" is represented by

$$(I_{G_M}, E_{G_M}) \sqcup (I_{G_L}, E_{G_L}) = (I_{G_M} \cup I_{G_L}, E_{G_M} \cap E_{G_L}).$$

This means that to the concept of "good students in Mathematics or Literature" surely belongs the set of students "surely good in Mathematics" or "surely good in Literature", i. e. $I_{G_M} \cup I_{G_L}$, while surely does not belong the set of students "surely not good in Mathematics" and "surely not good in Literature", i. e. $E_{G_M} \cap E_{G_L}$. Analogously, the concept of "good students in Mathematics *and* Literature" is represented by

$$(I_{G_M}, E_{G_M}) \sqcap (I_{G_L}, E_{G_L}) = (I_{G_M} \cap I_{G_L}, E_{G_M} \cup E_{G_L}).$$

This means that to the concept of "good students in Mathematics and Literature" surely belongs the set of students "surely good in Mathematics" and "surely good in Literature", i. e. $I_{G_M} \cap I_{G_L}$, while surely does not belong the set of students "surely not good in Mathematics" or "surely not good in Literature", i. e. $E_{G_M} \cup E_{G_L}$.

To the concept of "good students" (I_G, E_G), the Kleene bipolar complementation $^{-+}$ associates the set (E_G, I_G), that is to the negation of the concept of "good students" surely belongs the set of student "surely not good" E_G, and surely does not belong the set of all the students "surely good" I_G. Considering the Brouwer bipolar complementation $^{\approx +}$, the negation of the concept of "good students" (I_G, E_G) is given by $(E_G, U - E_G)$, that is to the negation of the concept of "good students" surely belongs the set of students "surely not good" E_G, as in the case of the Kleene bipolar complementation, and surely does not belong the set of all the other students $U - E_G$, differently from the case of the Kleene bipolar complementation. Similar considerations hold for the Kleene bipolar complementation $^{--}$ and the Brouwer bipolar complementation $^{\approx -}$ of negative concepts.

Conclusions

This article provides arguments for the claim that the dominance-based rough set approach is a proper way of handling monotonically ordered data in granular computing. Referring to some ideas of Leibniz, Frege, Boole, and Łukasiewicz, DRSA represents fundamental concepts of rough set theory in terms of a generalization that takes into account ordinal properties of data, permitting us to deal with the graduality of fuzzy sets. DRSA in the context of ordinal classification, its fuzzy extension, and a rough probabilistic model of DRSA (variable consistency-dominance-based rough set approach) have been presented. Moreover, dominance-based rough approximation of a fuzzy set has been discussed, which infers the most cautious conclusions from available imprecise information; different to almost all known fuzzy rough set approaches, the dominance-based rough approximation of a fuzzy set does not require any fuzzy connective which is always arbitrary to some extent.

Knowledge induced from dominance-based rough approximations of fuzzy sets, or more generally, from ordi-

nal data, is represented in terms of gradual decision rules. The dominance-based rough approximations of fuzzy sets generalize the classical rough approximations of crisp sets, as proved by showing that the classical rough set approach is one of its particular cases. Because of considering only ordinal character of the graduality of fuzzy sets, and due to eliminating all fuzzy connectives, the dominance-based rough approximations of fuzzy sets give a new insight into both rough sets and fuzzy sets, and enable further generalizations of both of them. The recently proposed DRSA for fuzzy case-based reasoning is an example of this capacity. This article also exhibits some important merits of DRSA within granular computing, which can be summarized as follows:

- DRSA extends the paradigm of granular computing to problems involving ordered data,
- It specifies a syntax and modality of information granules, defined by means of dominance-based constraints, which are appropriate for dealing with ordered data,
- It provides a methodology for dealing with this type of information granules, which results in a theory of computing with words and reasoning about ordered data,
- It is supported by a robust and meaningful algebraic model, the bipolar complemented de Morgan Brouwer–Zadeh distributive lattice, and this ensures the solidity of the obtained results.

Future Directions

Granular computing with ordered data is a very general problem, because also other modalities of information constraints, such as veristic, possibilistic, and probabilistic modalities, have to deal with ordered value sets (with qualifiers relative to grades of truth, possibility, and probability). For this reason, granular computing with ordered data based on DRSA is a very promising research field. There is a great potential in both theoretical investigations, such as extension of DRSA to other algebraic models than the Brouwer–Zadeh lattice (for a general survey of algebraic structures for rough set theory see Chap. 12 in [50]), and practical applications, such as customization of DRSA to case-based reasoning for specific domains, like medical diagnosis or computer security.

Bibliography

1. Cattaneo G, Ciucci D (2004) Algebraic structures for rough sets. In: Transaction on rough sets II. LNCS, vol 3135. Springer, Berlin, pp 208–252
2. Cattaneo G, Giuntini R, Pilla R (1999) BZMV^{dM} algebras and stonian MV-algebras, (applications to fuzzy sets and rough approximations). Fuzzy Sets Syst 108:201–222
3. Dubois D, Prade H (1990) Rough fuzzy sets and fuzzy rough sets. Int J General Syst 17:191–209
4. Dubois D, Prade H (1992) Gradual inference rules in approximate reasoning. Inf Sci 61:103–122
5. Dubois D, Prade H (1992) Putting rough sets and fuzzy sets together. In: Słowiński R (ed) Intelligent decision support – Handbook of applications and advances of the rough sets theory. Kluwer, Dordrecht pp 203–232
6. Dubois D, Prade H, Esteva F, Garcia P, Godo L, Lopez de Mantara R (1998) Fuzzy set modelling in case-based reasoning. Int J Intell Syst 13:345–373
7. Dubois D, Grzymala-Busse J, Inuiguchi M, Polkowski L (eds) (2004) Transations on rough sets II: Rough sets and fuzzy sets. LNCS, vol 3135. Springer, Berlin
8. Dyer J (2005) MAUT – Multiattribute utility theory, In: Figueira J, Greco S, Ehrgott M (eds) Multiple criteria decision analysis: State of the art surveys. Springer, Berlin, pp 266–294
9. Figueira J, Greco S, Ehrgott M (eds) (2005) Multiple criteria decision analysis: State of the art surveys. Springer, Berlin
10. Fodor J, Roubens M (1994) Fuzzy preference modelling and multicriteria decision support. Kluwer, Dordrecht
11. Fortemps P, Greco S, Słowiński R (2008) Multicriteria decision support using rules that represent rough-graded preference relations. Eur J Operational Res 188:206–223
12. Gilboa I, Schmeidler D (2001) A theory of case-based decisions. Cambridge University Press, Cabmridge
13. Ginsburg S, Hull R (1983) Order dependency in the relational model. Theor Comput Sci 26:149–195
14. Greco S, Inuiguchi M, Słowiński R (2002) Dominance-based rough set approach using possibility and necessity measures. In: Alpigini JJ, Peters JF, Skowron A, Zhong N (eds) Rough sets and current trends in computing. LNAI, vol 2475. Springer, Berlin, pp 85–92
15. Greco S, Inuiguchi M, Słowiński R (2004) A new proposal for rough fuzzy approximations and decision rule representation. In: Dubois D, Grzymala-Busse J, Inuiguchi M, Polkowski L (eds) Transations on rough sets II: Rough sets and fuzzy sets. LNCS, vol 3135. Springer, Berlin, pp 156–164
16. Greco S, Inuiguchi M, Słowiński R (2006) Fuzzy rough sets and multiple-premise gradual decision rules. Int J Approx Reason 41:179–211
17. Greco S, Matarazzo B, Słowiński R (1999) The use of rough sets and fuzzy sets in MCDM. In: Gal T, Stewart T, Hanne T (eds) Advances in multiple criteria decision making. Kluwer, Boston, pp 14.1–14.59
18. Greco S, Matarazzo B, Słowiński R (2000) Rough set processing of vague information using fuzzy similarity relations. In: Calude C, Paun G (eds) From finite to infinite. Springer, Berlin, pp 149–173
19. Greco S, Matarazzo B, Słowiński R (2000) A fuzzy extension of the rough set approach to multicriteria and multiattribute sorting. In: Fodor J, De Baets B, Perny P (eds) Preferences and decisions under incomplete information. Physica, Heidelberg, pp 131–154
20. Greco S, Matarazzo B, Słowiński R (2001) Rough sets theory for multicriteria decision analysis. Eur J Operational Res 129:1–47

21. Greco S, Matarazzo B, Słowiński R (2001) Rough set approach to decisions under risk. In: Ziarko W, Yao Y (eds) Rough sets and current trends in computing. LNAI, vol 2005. Springer, Berlin, pp 160–169
22. Greco S, Matarazzo B, Słowiński R (2002) Preference representation by means of conjoint measurement and decision rule model. In: Bouyssou D, Jacquet-Lagréze E, Perny P, Słowiński R, Vanderpooten D, Vincke P (eds) Aiding decisions with multiple criteria – Essays in Honor of Bernard Roy. Kluwer, Dordrecht, pp 263–313
23. Greco S, Matarazzo B, Słowiński R (2004) Axiomatic characterization of a general utility function and its particular cases in terms of conjoint measurement and rough-set decision rules. Eur J Operational Res 158:271–292
24. Greco S, Matarazzo B, Słowiński R (2004) Dominance-based rough set approach to knowledge discovery (I) – General perspective. In: Zhong N, Liu J (eds) Intelligent technologies for information analysis. Springer, Berlin, pp 513–552
25. Greco S, Matarazzo B, Słowiński R (2004) Dominance-based rough set approach to knowledge discovery (II) – Extensions and applications. In: Zhong N, Liu J (eds) Intelligent technologies for information analysis. Springer, Berlin, pp 553–612
26. Greco S, Matarazzo B, Słowiński R (2005) Decision rule approach. In: Figueira J, Greco S, Ehrgott M (eds) Multiple criteria decision analysis: State of the art surveys. Springer, Berlin, pp 507–563
27. Greco S, Matarazzo B, Słowiński R (2005) Generalizing rough set theory through dominance-based rough set approach. In: Slezak D, Yao J, Peters J, Ziarko W, Hu X (eds) Rough sets, fuzzy sets, data mining, and granular computing. LNAI, vol 3642. Springer, Berlin, pp 1–11
28. Greco S, Matarazzo B, Słowiński R (2006) Dominance-based rough set approach to case-based reasoning. In: Torra V, Narukawa Y, Valls A, Domingo-Ferrer J (eds) Modelling decisions for artificial intelligence. LNAI, vol 3885. Springer, Berlin, pp 7–18
29. Greco S, Matarazzo B, Słowiński R (2007) Dominance-based rough set approach as a proper way of handling graduality in rough set theory. In: Transactions on rough sets VII. LNAI, vol 4400. Springer, Berlin, pp 36–52
30. Greco S, Matarazzo B, Słowiński R (2008) An algebraic structure for dominance-based rough set approach. In: Proc. 3rd Int Conference on rough sets and knowledge technology (RSKT 2008), LNAI. Springer, Berlin, pp 252–259
31. Greco S, Matarazzo B, Słowiński R, Stefanowski J (2001) Variable consistency model of dominance-based rough set approach, In: Ziarko W, Yao Y (eds) Rough sets and current trends in computing. LNAI, vol 2005. Springer, Berlin, pp 170–181
32. Greco S, Predki B, Słowiński R (2002) Searching for an equivalence between decision rules and concordance-discordance preference model in multicriteria choice problems. Control Cybern 31:921–935
33. Hume D (1748) An enquiry concerning human understanding. Oxford, Clarendon Press
34. Klement EP, Mesiar R, Pap E (2000) Triangular norms. Kluwer, Dordrecht
35. Kolodner J (1993) Case-based reasoning. Morgan Kaufmann, San Mateo
36. Leake DB (1996) CBR in context: the present and future. In: Leake D (ed) Case-based reasoning: Experiences, lessons, and future directions. AAAI Press/MIT Press, Menlo Park, pp 1–30
37. Lin TY (1988) Neighborhood systems and relational databases. In: Proceedings of the ACM Conference on Computer Science, Atlanta, p 725
38. Lin TY (1989) Neighborhood systems and approximation in database and knowledge base systems. In: Proceedings of the Fourth International Symposium on Methodologies of Intelligent Systems, Poster Session, October 12–15, pp 75–86
39. Lin TY (1992) Topological and fuzzy rough sets. In: Slowinski R (ed) Intelligent decision support – Handbook of application and advances of the rough sets theory. Kluwer, Dordrecht, pp 287–304
40. Lin TY (1997) Granular computing. Announcement of the BISC Special Interest Group on Granular Computing
41. Lin TY (1998) Granular computing on binary relations I: Data mining and neighborhood systems. In: Skowron A, Polkowski L (eds) Rough sets in knowledge discovery. Physica, Heidelberg, pp 107–121
42. Lin TY (1998) Granular computing on binary relations II: Rough set representations and belief functions. In: Skowron A, Polkowski L (eds) Rough sets in knowledge discovery. Physica, Heidelberg, pp 121–140
43. Loemker L (ed and trans), Leibniz GW (1969) Philosophical papers and letters, 2nd edn. Reidel, Dordrecht
44. Nakamura A, Gao JM (1991) A logic for fuzzy data analysis. Fuzzy Sets Syst 39:127–132
45. Pal SK, Skowron A (eds) (1999) Rough-fuzzy hybridization: A new trends in decision making. Springer, Singapore
46. Pawlak Z (1982) Rough sets. Int J Comput Inf Sci 11:341–356
47. Pawlak Z (1991) Rough sets. Kluwer, Dordrecht
48. Pawlak Z (2001) Rough set theory. Künstliche Intelligenz 3: 38–39
49. Peters JF, Skowron A, Dubois D, Grzymala-Busse J, Inuiguchi M, Polkowski L (eds) (2005) Rough sets and fuzzy sets, transaction on rough sets II. Springer, Berlin
50. Polkowski L (2002) Rough set: mathematical foundations. Physica, Heidelberg
51. Radzikowska AM, Kerre EE (2002) A comparative study of fuzzy rough sets. Fuzzy Sets Syst 126:137–155
52. Słowiński R, Greco S, Matarazzo B (2002) Axiomatization of utility, outranking and decision-rule preference models for multiple-criteria classification problems under partial inconsistency with the dominance principle. Control Cybern 31:1005–1035
53. Słowiński R, Greco S, Matarazzo B (2002) Mining decision-rule preference model from rough approximation of preference relation. In: Proc. 26th IEEE Annual Int. Conference on Computer Software & Applications (COMPSAC 2002), Oxford, pp 1129–1134
54. Słowiński R, Greco S, Matarazzo B (2005) Rough set based decision support. In: Burke EK, Kendall G (eds) Search methodologies: Introductory tutorials in optimization and decision support techniques. Springer, New York, pp 475–527
55. Stewart T (2005) Dealing with uncertainties in MCDA. In: Figueira J, Greco S, Ehrgott M (eds) Multiple criteria decision analysis: State of the art surveys. Springer, Berlin, pp 445–470
56. Zadeh LA (1965) Fuzzy sets. Inf Control 8:338–353
57. Zadeh LA (1979) Fuzzy sets and information granularity. In: Gupta M, Ragade RK, Yager RR (eds) Advances in fuzzy set theory and applications. North-Holland, Amsterdam, pp 3–18

58. Zadeh LA (1996) Key roles of information granulation and fuzzy logic in human reasoning, concept formulation and computing with words. In: Proceedings of the 5th IEEE International Conference on Fuzzy Systems, New Orleans, p 1
59. Zadeh LA (1997) Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets Syst 90:111–127
60. Zadeh LA (1999) From computing with numbers to computing with words – from manipulation of measurements to manipulation of perception. IEEE Trans Circuits Syst – I: Fundament Theor Appl 45:105–119
61. Ziarko W (1993) Variable precision rough sets model. J Comput Syst Sci 46:39–59
62. Ziarko W (1998) Rough sets as a methodology for data mining. In: Polkowski L, Skowron A (eds) Rough Sets in Knowledge Discovery, vol 1. Physica, Heidelberg, pp 554–576

Granular Computing, Information Models for

Steven A. Demurjian
Department of Computer Science & Engineering, The University of Connecticut, Storrs, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Candidate Information Models
Suitability of Information Models for Granular Computing
Future Directions
Bibliography

Glossary

Attributed-based data model (ABDM) ABDM is a data (information) model that organizes data (information) in attribute-value pairs, providing a means for both definition and access.

Data (information) model A data (information) model is used in database design to capture the structure or schema of the data. All data (information) that is entered into a database must conform to the definitions of the data (information) model.

Data (information) table A set of similar granules used in granular computing are collected into a data (information) table in order to allow them to be conceptualized and reasoned on in a formal manner.

Extensible markup language (XML) XML is a data (information) model that is a standard for exchanging and sharing data (information) across the internet, among databases, and so on.

Functional data model (FDM) FDM is a semantic data (information) model to represent data (information) as it appears in the "real-world" with a functional/logic basis.

Granular computing A discipline of information theory/computer science that uses a formal theory to reason about and analyze data (information) in granules.

Granule A piece of data (information) of varied size and complexity that is used to represent data (information) as it occurs in some "real-world" context.

Relational data model (RDM) RDM is a dominant data (information) model in commercial and open source database management systems with a basis in set theory.

Definition of the Subject

Granular computing (GrC) is an emerging discipline of information theory that strives to allow reasoning and analysis based on varying levels of information granularity (from fine to coarse). In GrC, the entire information universe can be organized based on many different criteria, allowing the information to be abstracted, aggregated, classified, generalized, and so on, based on various characteristics (e. g., data similarity, operational usage, etc.). As a result, GrC relies on information models to describe the universe, the elements of the universe, and the composition of each element. In this chapter, alternative candidate information models for GrC are explored, including: the attribute-based data model, the Relational data model, the functional data model, and the extensible markup-language (XML). This includes both a description of these models and an analysis of the suitability in support of GrC.

Introduction

In information theory, one approach to support the reasoning and analysis of information based on varied levels of conceptualization is the emerging disciple of *Granular computing (GrC)* [10,13,14,26], a term jointly coined by Lin and Zadeh and having a basis in research conducted by Lin on neighborhoods (e. g., granules) in databases [7] and computer security [8]. Information granulation is an approach that partitions the entire information universe into *granules* [11,12,24,25]. This partitioning can be based on many different criteria: aggregating objects that demonstrate similarity of features or usage by classification (e. g., defining a relation such as a Course table in a university application); abstracting away details via generalization (e. g., creating an object-oriented class Person that would

be the root of a hierarchy); distinguishing among different characteristics of objects by specialization (e. g., Students or Faculty sub-classes of Person); summarizing the characteristics of similar objects into a coarser conceptualization (e. g., collecting multiple Computer Science Faculty instances into a single object); and so on. In all of these situations, the information universe is described in its entirety, and from that initial characterization, it is possible to define elements and components of elements.

In support of GrC, one of the key initial considerations is to define the information universe via an appropriate data representation. For GrC, this data representation has most frequently been based on data tables as formally defined for rough sets [15], where a *data table* has many different equivalent nomenclatures, including: a knowledge representation system, an attribute-value system, and an information table. For a data table, columns are labeled as attributes (of the objects) and rows are labeled as instances (of the objects). Formally, the data table is characterized as a pair that contains a finite non-empty universe U and a finite non-empty set of primitive attributed A. Additionally, for each a in A, there is a non-empty set of values, and a function that maps the universe U to these values. Collectively, these data tables based on rough sets have been used as the basis to define foundational models for granular computing [13,14,18] and to achieve privacy protection for medical data by partitioning attributes into identifying, easily known, and sensitive categories [20].

The main objective of this chapter is to explore alternative candidate information models for GrC to support the data table (and dependencies that may exist both within a table and across multiple data tables). Specifically, a variety of models that are old and new are explored, including: the attribute-based data model [6], the relational data model [1], the functional data model [17], and the extensible markup-language [22]. The *attribute-based data model (ABDM)*, proposed in the early 1970, was touted for its ability to formally define "real-world" records of information, to support their retrieval, and has been shown capable of capturing data from relational, hierarchical, network, and functional data models [3,4]. The *relational data model (RDM)* proposed in mid-1970 along with the current SQL 1992 (SQL2) has evolved into the dominant database model for commercial and open source data base systems, with extensions to support object-oriented and other features underway (SQL3). The *functional data model (FDM)* and its associated Daplex programming language, proposed in 1981, offers unique programming-like features and its functional formal underpinnings can provide a means to reason about data and constraints, and moreover, to capture semantics. Finally, the *extensible markup-language (XML)* has emerged as an information-representation standard format, allowing information to be modeled and more easily exchanged among programs, web-sites, databases, etc. All of these models offer different capabilities in support of data tables and GrC.

The remainder of this chapter contains two sections and a conclusion. In Sect. "Candidate Information Models", the four information models, ADBM, RDM, FDM, and XML, are examined. Using this as a basis, in Sect. "Suitability of Information Models for Granular Computing", the suitability of the four information models in support of GrC is investigated by considering them against three different criteria. Finally, Sect. "Future Directions" offers concluding remarks.

Candidate Information Models

In this section, background on the four information models to be considered for their suitability for GrC are reviewed, namely: the attribute-based data model (ABDM) [6], the relational data model (RDM) [1], the functional data model (FDM) [17], and the extensible markup-language (XML) [22]. To serve as a common context for the discussion, a university database is utilized as a example. In the university database, faculty are tracked by name, identifier, office phone, and department (e. g., computer science, mathematics, English, etc.), and students are tracked by name, identifier, grade point average, and campus address (e. g., dormitory). For each course in the catalog, there is a unique course number (e. g., a combination of department and number such as CS123, MATH233, etc.), the course title and description, and the list of prerequisites for the course. The courses offered are also tracked, by course number/section number pairs (e. g., courses may be offered in multiple sections), and the FacultyID teaching each offering (by faculty identifier), and the term (e. g., semester, quarter, etc.) of the offering (e. g., Fall2007, Spring2008, Summer2008, etc.). Likewise, the students enrolled by the course number/section number pair for each term are tracked.

The Relational Data Model (RDM)

Relational data is organized into tuples of *relations*. A database is a collection of relations. The attributes of a relation are distinct. The tuples of a relation have the property that no two tuples are identical. In addition, one or more attributes of the relation may be defined as the *primary key* of the relation; a relation can have multiple candidate keys, one of which is chosen as the primary key. The primary key of the relation is used to uniquely identify

the tuples of the relation. To establish dependencies across two or more relations, a *foreign key* can be defined consisting of one or more attributes of one relation that reference the primary key of another relation. Foreign keys are employed to establish referential integrity constraints across two relations where the foreign key of one relation references the primary key of another relation (with self- reference allowed). Non-primary key attributes are allowed to have null values; this includes a foreign key. A definition of the University database in tabular (or relational) form is as below:

```
Student(Name, StudentID*, GPA, CampusAddress)
Faculty(Name, FacultyID*, Ophone, Department)
Courses(Course#*, Title, Description,
                                  PCourse#*)
OfferedCourses(Course#*, Section#*,
                          FacultyID*, Term*)
EnrolledStudents(Course#*, Section#*,
                          StudentID*, Term*)
```

Briefly, let us describe the data relationships of this database. First, each student of the Student relation is uniquely identified by a StudentID (primary key), has a Name, a grade point average (GPA), and a CampusAddress. Next, each faculty member of the Faculty relation also has a unique FacultyID (primary key), a Name, an office phone number, and a Department affiliation. Third, each course of the Courses relation (catalog) is uniquely identified by a Course# (primary key), has a Title, a Description and may have zero or more prerequisite courses (PCourse#-foreign key). Fourth, the OfferedCourses relation tracks the courses offered by sections taught by faculty each term, with these four attributes forming a compound primary key. Fifth, the EnrolledStudents relation tracks the students enrolled in sections of courses for each term, again by a compound primary key. In terms of referential integrity, the foreign key PCourse# self references the relation Course; note that a course without a prerequisite will have a null foreign key.

The Attribute-Based Data Model (ABDM)

In ABDM, the initial step in defining a database is the specification of a collection of the attributes that are in the database. The *database records* in ADBM consist of sets of attribute-value pairs. An *attribute-value pair* is a member of the Cartesian product of the attribute name and the value domain of the attribute. As an example, <Department, Computer Science> is an attribute-value pair having Computer Science as the value for the Department attribute. Database records have the property that for a given set of attribute-value pairs, no two pairs in the set have the same attribute name. In ADBM, data is considered in the following constructs: database, file, record, attribute-value pair (keyword) attribute-value range, record body, directory, directory keyword and non-directory keyword. Informally, a *database* consists of a collection of files. Each *file* contains a group of records which are characterized by a unique set of keywords. A *record* is composed of two parts. The first part is a collection of *attribute-value pairs (keywords)*. The second part of the record, which is optional, is for unformatted textual information, and is referred to as the *record body*. A record contains at most one attribute-value pair for each attribute defined in the database. An example of a record equivalent to the Faculty relation is:

```
(<File, Faculty>, <Record Number, 123>,
<Name, John Smith>, <FacultyID, 12121212>,
<Ophone, 5551111>, <Department, Computer
                                  Science>,
{This is an assistant professor up for
tenure in 2012})
```

The angle brackets, <, >, enclose an attribute-value pair, i. e., keyword. The curly brackets, {, }, include the record body. The record is enclosed in parentheses. The first attribute-value pair of all records of a file, by convention, is the same. In particular, the attribute is File and the value is the file name (akin to a relation name in RDM). Similar records for Student, Courses, CourseOfferings, and EnrolledStudents can also be defined; referential integrity is achieved by using the unique record numbers that are given to every instance.

Finally, in ABDM, the indexing criteria for a given database is also a critical part of the definition of the database. In particular, for indexing purposes, there are two different types of attribute-value pairs in a record (or a file), with all of the indexing data maintained in the *directory* of the database. Certain attribute-value pairs of a record (or a file) are called the *directory keywords* of the record (file), because either the attribute-value pairs or their attribute-value ranges are kept in a directory for identifying the records (files). Those attribute-value pairs which are not kept in the directory are called *non-directory keywords*. In the example record above, the File, Record Number, and FacultyID attributes would be directory keywords, while the other attributes would be non-directory keywords. The identification of database records are by either directory or non-directory keywords. When

directory keywords are used, the search space is clearly defined using the indexing criteria. When non-directory keywords are used, the entire file of records must be searched. Note that null values are not allowed for directory attributes.

The Functional Data Model (FDM)

An *entity* in a functional database is always associated with a collection of distinct functions that can be applied to the entity to return either individual data values or one or more objects. The term *object* is used to refer to the actual data values (values for the functions) for an entity. Thus, an object can be considered as an instantiation (occurrence, in the earlier terminology) of the entity. An object is analogous to a tuple in the RDM or a record in ABDM. The *functions* of an entity are applied to the entity to return a particular value that is associated with that entity, i. e., to return a portion of the object. There are two types of functions, scalar-valued functions and entity-valued functions. Each type of function may be either single-valued (returning one value) or set-valued (returning zero or more values). *Scalar-valued functions* return one or more typical database values (i. e., string, integer, and float). *Entity-valued functions* return one or more entity objects as their values.

Additionally, in the FDM, relationships between the entities may be defined that lead to one or more generalization hierarchies for the database. Generalization [19], is an abstraction technique that is used to organize commonalities from multiple entities into a single entity (e. g., the name and identifier for students and faculty can be organized in a person entity). Because of the presence of generalization hierarchies, there are two different categories of entities, entity subtypes and entity supertypes. An *entity subtype* exists at the inner and leaf nodes of a generalization hierarchy, and inherits all of the characteristics of its ancestors, i. e., inherits all of the functions of its ancestors. An *entity supertype*, or, more simply, an *entity type*, is at the root of a generalization hierarchy and has the property that no two instantiations (occurrences) of the entity type return the same values for all of the entity's functions. Additionally, one or more single-valued, scalar-valued functions in an entity type may be defined to be the key of the entity. Functional data is organized into *generalization hierarchies* of *entities*. A database is a collection of entities organized into generalization hierarchies. References among entities are accomplished when a function is entity valued, with self-referencing possible. In the following, the definition of the FDM version of the University database using the Daplex data definition language is presented.

```
DATABASE University IS

TYPE        Person;
SUBTYPE     Student;
SUBTYPE     Faculty;
TYPE        Courses;
TYPE        CoursesOffered;

TYPE Person IS
     Name       : STRING(1..30);
     Identifier : STRING(1..9);
END ENTITY;

TYPE Courses IS
     Course#    : STRING(1..8);
     Title      : STRING(1..20);
     Descrip    : STRING(1..100);
     PCourse#   : SET OF Courses;
END ENTITY;

TYPE CoursesOffered IS
     Course#    : STRING(1..8);
     Section#   : INTEGER;
     TERM       : STRING(1..10);
END ENTITY;

SUBTYPE Student IS Person
     StTakes : SET OF CoursesOffered;
     GPA     : FLOAT;
     CampusAddress : STRING(1..100);
END ENTITY;

SUBTYPE Faculty IS Person
     Teaches : SET OF CoursesOffered;
     Ophone  : STRING(1..7);
     Department : STRING(1..30);
END ENTITY;

UNIQUE Course# WITHIN Courses;
UNIQUE Section# WITHIN Formats;
UNIQUE Identifier WITHIN Person;
UNIQUE Course#, Section#, Term WITHIN
                            CoursesOffered;

END UNIVERSITY;
```

The FDM version of the University database has subtle differences from its RDM counterpart. Note that Student and Faculty in the RDM version have been reorganized into a generalization hierarchy with Person (root entity type with commonalities) and Student and Faculty (children or entity subtypes). Courses is similar, but the foreign key PCourse# has been replaced by an entity-valued function. The other two differences are: CoursesOffered no longer includes the faculty identifier and has been replaced by a Teaches entity-valued function in Faculty; and, EnrolledStudents has been replaced by a Takes entity-valued func-

tion in Student. The end of the definition of the database includes the various uniqueness constraints (akin to primary keys).

The Extensible Markup Language (XML)

The eXtensible Markup Language (XML) [22] has emerged as a standard for information modeling and exchange for web-based applications, database interoperability, common software tool formats, patient record data, etc. XML allows information content to be hierarchically organized and tagged to highlight important/relevant content. The *tags* capture not only the content, but can be leveraged to represent the meaning of the information (semantics). In a web-based setting, XML is the successor to HTML to allow information content to be hierarchically organized and tagged to highlight important and relevant content. The tags can be exploited to capture both information content and the meaning of the information (semantics). In addition, XML allows the definition of templates to capture the known structure information for an application by the creating of XML schema files called Document Type Definitions (DTDs). The resulting XML document instances that are created contain both information content (data) as well as semantic notations (tags) that indicate the meaning of the information.

The Extensible Markup Language, XML [22] provides a flexible means to store and transmit data between different information systems and platforms, and has emerged as a dominant means for interoperability on the World Wide Web (as a successor to HTML) and as a standard for information format and exchange (e. g., the OpenDocument project for office applications). Both HTML and XML use tags to identify data/elements. However, while the HTML tags are predefined and specify the way the data within the tags is to be displayed, the XML tags are user-defined and can be employed to identify the data structure in a hierarchical fashion. For instance in HTML, < b > or < / b > are all predefined tags to display the data/characters between them in a bold font. In XML, you will not have those kinds of predefined tags. All of the tags are defined by the users to indicate the structure of the data elements. An XML document for faculty data would be as simple as:

```
<?xml version='1.0' encoding='ISO-8859-1' ?>
 <Faculty>
  <Name> John Smith </Name>
  <FacultyID> 12121212 </FacultyID>
  <Ophone> 5551111 </Ophone>
  <Department> Computer Science </Department>
 </Faculty>
```

In the XML fragment, the tags <Faculty>, <Name>, <FacultyID>, etc., are defined, based on the application and/or modeling requirements. These tags capture data content (with attribute names) as well as data dependencies; they do not provide any clues on formatting or displaying the data elements between them. In addition to these structural capabilities of an XML file, each of the elements can be further quantified with tags to track semantic aspects of the data. For example, the Name could be augmented with a tag that indicates whether it is a married or maiden name, and the Department could be tagged with a title (e. g., Professor, Associate Professor, Head, etc.). If the XML structure contained numerical data (e. g., GPA), then the additional tags may denote the scale of the data (e. g., out of 4.0, 100, etc.).

The overall appearance and structure of an XML file can also be defined and validated by an XML schema file. An XML schema file defines the tags or attributes that can be used and where they appear in an XML file. There are two types of XML schema files: the Document Type Definition (DTD) and the XML Schema (a more sophisticated construct). A sample XML Schema for the previous XML fragment can be defined as:

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<xs:schema xmlns:xs=
         "http://www.w3.org/2001/XMLSchema">

<xs:element name="Faculty">
   <xs:complexType>
      <xs:sequence>
         <xs:element name="Name"
            type="xs:string"/>
         <xs:element unique name="FacultyID"
            type="xs:string"/>
         <xs:element name="Ophone"
            type="xs:string"/>
         <xs:element name="Department"
            type="xs:string"/>
      </xs:sequence>
    </xs:complexType>
</xs:element>
```

This sample XML Schema file defines that in the corresponding XML file, the Faculty element is a complex type (akin to a relation in RDM, file in ABDM, and entity type in FDM) must have one Name, one FacultyID, one Ophone, and one Department element. Uniqueness is achieved via the "unique" keyword attached to the FacultyID element. With a XML Schema, an XML parser can be used to validate whether the elements within a XML file are valid or not with respect to the XML Schema. Concep-

tually, a XML Schema is analogous to type declaration for a RDM relation, a ABDM file, or a FDM entity type (subtype).

Suitability of Information Modelsbreak for Granular Computing

This section, compares the four information models as given in Sect. "Candidate Information Models" (ABDM, RDM, FDM, and XML) for their suitability in support of granular computing. To facilitate the comparison, criteria are utilized that are relevant for comparing data models [3,4], in conjunction with various concepts proposed in [25]. The criteria chosen are:

- Formalism and Theory: This criterion is used to assess the degree to which a formal model exists for the data model that is capable of supporting GrC.
- Expressive Power: This criterion is used to evaluate the ability of each data model to support the definition, usage, and analysis of granules,
- Extensibility and Utility: This criterion is used to determine the usability of the data model, from many different perspectives.

In the remainder of this section, these criteria are explored against the four models, as a means to assess their suitability for GrC.

Formalism and Theory

All of the data models given in Sect. "Candidate Information Models", have a certain degree of formalism/theory upon which they are based. GrC relies on an information model that has a formal basis, in order to capture the required information within a data table [15] and to employ a theory of rough sets and fuzzy logic for granular analysis. As a result, it is vital that an underlying data model is rich enough to support the formalisms necessary for GrC. To begin, the attributed-based data model [6] was proposed by David K. Hsiao (founder of ACM Transactions on Database Systems) and Frank Harary (noted graph theorist) and has a formal basis that has been explored by other researchers in the 1970s [16], 1980s [4], and 1990s [9], with applicability in many different contexts including paging environments for OS, data-model transformations for federated databases, and polyinstantiation for security, respectively. Likewise, the relational data model [1] has a rich history and formal basis using set theoretic concepts, and is the dominant approach to date for GrC researchers [2,13,20].

The functional data model [17] was proposed in an era in the early-to-mid 1980s when semantic data models (successors to the entity-relationship data model) were advocated to break away from relatively flat approaches (e. g., relational, hierarchical, and network data models and databases systems) to an approach that was semantically rich and able to model the data as it occurs in the "real world". In fact, in a workshop in 1993 later published as special journal issue [5], the overriding theme emphasized the functional model and its potential for a unifying paradigm, providing formalisms based on functional and logic programming that could be used to capture both relational data and object-oriented data in a formal way. Lastly, XML [22] has emerged as a de facto standard across many different disciplines, being used in a wide variety of contexts (web semantics, patient health records/standards, database interoperability, information exchange format, etc.). XML is based on detailed specifications [23] and has a grammar in Extended Backus–Naur Form. As a result, there are XML parsers that have been built for many purposes to parse XML files and XML Schemas (see Sect. "The Extensible Markup Language (XML)" again) in different ways. Such a formal-grammar basis means that XML can be utilized in a very rigorous manner to reason using context-free grammar concepts from automata theory. In summary, it is clear that from a formal basis, each of these four data models can realize the information/data table needs of GrC, offering different strengths in their support.

Expressive Power

The expressive power needed to support information granules can be impacted by many factors; we choose three for our comparison:

- *Grouping capability* which involves the reason that objects are grouped into granules (equivalence classes) and can be based on object relationships, similarity, proximity, semantics, etc. [25].
- *Linguistic representation of granules* which assigns a linguistic value to each granule that can be accomplished using a naming convention or by providing sample representative objects [25].
- *Constraint and type checking* which involves the ability to enforce intra-granular and inter-granular dependencies (constraint checking) while simultaneously insuring that the granule always adheres to the defined structure (type checking).

From an information model perspective, these three granulation factors are directly related to the features and characteristics of the data models as presented in Sect. "Candidate Information Models", and includes: aggregation

(records in ABDM, relations in RDM, types and subtypes in FDM, and XML Schemas in XML), generalization (inheritance in FDM), identity and uniqueness (directory attributes in ABDM, primary keys in RDM, uniqueness in FDM, and unique in XML), and relationships (record references in ABDM, foreign keys in RDM, entity-valued functions in FDM, and hierarchical structure in XML).

Specifically, for the grouping capability factor, the assembly of objects into granules (equivalence classes) is aggregation, which is provided by all four data models. Aggregation is simply a result of the creation of the equivalence class, and the aggregation can be based on similarity, proximity, etc. The process of grouping objects into granules is part of GrC, but once the granulation has been completed, then any of these four models have sufficient aggregation to represent a granule. If granules are related to one another hierarchically (or the grouping occurs based on a hierarchical relationship), then either FDM or XML is appropriate for this resultant aggregation. For the linguistic representation of granules factor, again, the same aggregation capability applies; all four models provide a naming convention, and the examples as given in Sect. "Candidate Information Models" for ABDM and XML illustrated data instances (sample representative objects).

Finally, for the constraints and type checking factor, each model offers different capabilities. For ABDM, uniqueness of records, record references across objects (instances), support for attribute–value pairs (used by many granulation researchers), and directory attributes (akin to keys), all provide the basis for constraint checking; the definition of the attributes (name and type) in each record provides type checking. For RDM, no duplicate tuples per relation, primary keys, foreign keys, and referential integrity are relevant for constraint checking; a relational table definition with its attributes and data types provides type checking. For FDM, the types and subtypes provide explicit support for type checking (i. e., these types and subtypes are akin to programming languages types and subtypes in Ada), while uniqueness constraints preserve properties (constraint checking). For XML, the XML Schema is a template that an XML instance must follow (type checking), and in its definition, provides the ability for dependencies (constraint checking); as such, for different granules, there would be different XML Schemas, and XML provides the ability to parse and map from one XML Schema to another via XSL Transformations (http://www.w3.org/TR/xslt) and XQuery (http://www.w3.org/TR/xquery). Thus, XML is much more powerful than the other models, particularly in type checking; the use of XML can facilitate the ability to transform one granule to another while still preserving content and semantics.

Extensibility and Utility

The final criterion considered is extensibility and utility, which is focused on the usability of the various models, both today and in the future. This can be examined from many different perspectives:

- Database Platform Support: While Sect. "Formalism and Theory" considered the theoretical basis for GrC, when one moves from theory to practice in GrC, a seamless transition to the corresponding database management system for a chosen data model would be very useful. In that regard, RDM has the definite advantage, with a wide variety of commercial (Oracle, SQL Server, Informix, etc.) and Open Source (MySQL, PostgreSQL, Ingres, etc.) products available. XML is also a strong player in this regard, since it is a dominant technique for information exchange (data interoperability) among databases, particularly across networked and distributed database solutions.
- Future Potential: XML has the greatest potential for future usage in GrC as an information model, due to its widespread and growing usage. In fact, there are many open source XML databases that are starting to be released, including: Apache XIndice, Senda XML DBMS, X-Hive/DB, etc. Many of these are in their earliest release stages. For RDM, SQL3, which will include object-oriented extensions, is still being considered from a standards perspective. These object-oriented capabilities are critical to allow granules to be formed in a representation similar to the way that they occur in the "real world", rather than having the granule be flattened into a relational table. Once approved, the relational database vendors will have to implement SQL3 and release new versions. One can easily hypothesize that XML database systems may make SQL3 obsolete before it even is released.
- Data Model Translations: The ABDM as given in Sect. "The Attribute-Based Data Model (ABDM)", has been shown to be capable of subsuming the features of RDM and FDM [4] at both a model and system level. This means that a single model can be used in place of other models without a loss of information.

Overall, the clear leader from a model perspective is ABDM and its ability to subsume RDM and FDM model capabilities. From a practical perspective, while RDM may have the edge in commercial and open source database platforms, the future seems pointed towards XML for sharing and database usage.

Future Directions

In this chapter, candidate information models and their suitability for granular computing (GrC) have been explored. This chapter considered four different models, presented in Sect. "Candidate Information Models": the attribute-based data model (ABDM) [6], the relational data model (RDM) [1], the functional data model (FDM) [17], and the extensible markup-language (XML) [22]. Two of these models (RDM and XML) are dominant from a database system perspective today, but all four models have different capabilities to offer when modeling, reasoning, and analyzing for GrC. To place the four models in their proper perspective, in Sect. "Suitability of Information Models for Granular Computing", the models were explored using three different criteria: Formalism and Theory, Expressive Power, and Extensibility and Utility: For the first criterion (Sect. "Formalism and Theory"), each of the four models have a strong formal basis and long history of usage; to varying degrees all are suitable for a formal basis for GrC. For the second criterion (Sect. "Expressive Power"), we compared based on three factors: grouping capability, linguistic representation of granules, and constraint and type checking. In the analysis, ABDM, RDM, and FDM are very comparable in what they offer, but XML has the potential to transcend all three, with both its current capabilities and its emerging characteristics and features. Finally, for the third criterion (Sect. "Extensibility and Utility"), there were different strengths for different models: for ADBM, its strength was its ability to subsume RDM and FDM at the data model level [4]; for RDM, its strength was in the deployed database platforms (commercial and open source) and the future potential of SQL3 with object-oriented capabilities (which will improve granule representation); and, for XML, the usage in relational databases for information exchange, coupled with the arrival of XML database systems, may have the greatest potential for the future. This leaves us with two recommendations: from a theory perspective, while all four data models are appropriate to some degree for GrC, ABDM and RDM have to be strongly considered due to their long history of usage and formal basis, with FDM an alternative when there is a desire to have a functional/logic based basis to GrC; and, from a practical perspective, if one is to transition from GrC theory to actual usage, RDM is the choice today, but XML has an excellent chance to be the choice of the future.

Bibliography

1. Codd E (1970) A relational model of data for large shared data banks. Commun ACM 13(6):377–387
2. Demchenko Y et al (2005) Security architecture for open collaborative environment. In: Sloot P et al (eds) Advances in grid computing – EGC2005, vol 3470. LNCS. Springer, Heidelberg
3. Demurjian S, Hsiao D (1987) The multi-lingual database system. In: Proceedings of the 3rd international conference on data engineering. IEEE Computer Society, Washington DC
4. Demurjian S, Hsiao D (1988) Towards a better understanding of data models through the multilingual database system. IEEE Trans Softw Eng 14(7):946–958
5. Gray P, King P, Kerschberg L (1999) Functional approach to intelligent information systems. Special Issue J Intell Inf Syst (JIIS) 12(2–3):107–111
6. Hsiao D, Harary F (1970) A formal system for information retrieval from files. Commun ACM 13(2), Corrigenda 13(4): 67–73
7. Lin T (1988) Neighborhood systems and approximation in relational database and knowledge bases. In: Proceedings of the fourth international symposium on methodologies of intelligent systems, poster session. Oct. 1989. IEEE Computer Society, Washington DC
8. Lin T (1989) Chinese wall security policy – an aggressive model. In: Proceedings of the fifth aerospace computer security application conference. Dec. 1989. IEEE Computer Society, Washington DC
9. Lin T (1992) Attribute based data model and polyinstantiation. In: Aiken R (ed) Education and society – information processing '92, vol 2, proceedings of the IFIP 12th world computer congress. North-Holland, 1992
10. Lin T (1997) Granular computing. In: Announcement of the BISC special interest group on granular computing
11. Lin T (1998) Granular computing on binary relations I: Data mining and neighborhood systems. In: Skowron A, Polkowski L (eds) Rough sets in knowledge discovery. Physica, Heidelberg, pp 107–121
12. Lin T (1998) Granular computing on binary relations II: Rough set representations and belief functions. In: Skowron A, Polkowski L (eds) Rough sets in knowledge discovery. Physica, Heidelberg, pp 121–140
13. Lin T (2005) Granular computing: A problem solving paradigm. In: Proceedings of 14th international conference on fuzzy systems, 2005. IEEE Computer Society, Washington DC
14. Lin T (2006) A roadmap from rough set theory to granular computing. In: Wang G et al (eds) Proceedings of first international conference on rough sets and knowledge technology – RSKT 2006, vol 4062. LNCS. Springer, Heidelberg pp 33–41
15. Pawlak Z (1991) Rough sets – theoretical aspects of reasoning about data. Kluwer, Heidelberg
16. Rothnie Jr JB (1974) Attribute based file organization in a paged memory environment. Commun ACM 17(2):63–69
17. Shipman D (1981) The functional data model and the data language DAPLEX. ACM Trans Database Syst 6(1):140–173
18. Skowron A, Stepaniuk J (2001) Information granules: Towards foundations of granular computing. Int J Intell Syst 16(1): 57–86
19. Smith J, Smith D (1977) Database abstractions: Aggregation and generalization. ACM Trans Database Syst 2(2):405–413
20. Wang D et al (2004) Medical privacy protection based on granular computing. Artif Intell Medicine 32(2):137–149
21. Wong E, Chiang T (1971) Canonical structure in attribute based file organization. Commun ACM 14(9):593–597

22. World Wide Web Consortium. http://www.w3c.org/XML/. Accessed 10 July 2008
23. XML 1.0 Fourth Edition Specification. http://www.w3.org/TR/2006/REC-xml-20060816/. Accessed 10 July 2008
24. Zadeh L (1996) Fuzzy logic = Computing with words. IEEE Trans Fuzzy Syst 4(2):103–111
25. Zadeh L (1997) Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets Syst 90(2):111–127
26. Zadeh L (1998) Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems. Soft Comput 2(1):23–25

Granular Computing, Introduction to

Tsau Young Lin
Department of Computer Science, San Jose State University, San Jose, USA

Article Outline

Granular Computing from Rough Set Theory
Granular Computing in Database Theory
Granular Computing in Social Networks
Granular Computing and Fuzzy Set Theory
Grid/Cloud Computing – A New Addition to GrC
General Issues in Granular Computing
Conclusions
Bibliography

What is granular computing (GrC)? It is a shifting paradigm. Let us start with a few words about how the term was coined. In the academic year 1996–97, when Lin (this section editor) took his sabbatical leave at UC-Berkeley, Zadeh suggested granular mathematics (GrM) as his research area. To limit the scope, Lin proposed the term *granular computing* [14]. What was GrC at that time? Zadeh had outlined it in his 1997 seminal paper [15]. Lin took an incremental approach: he mapped his neighborhood system [5] to Zadeh's intuitive definition [12] and used it as his First GrC model [8,9,10]. It may be important to point out that the concept of neighborhood systems, which was motivated from approximate retrieval in databases [7], is a generalization of the topological neighborhood system that formalizes the ancient intuition of infinitesimal granules.

Much progress has been achieved since then. This section has been organized to represent this progress and to reflect the current state of GrC. We believe in the incremental approach, namely, that each new step is based on solid results and moves forward. So many special theories and applications are gathered here. Jointly, they reflect the current state of GrC and may also implicitly hint at the ultimate goals of GrC.

To grasp the main idea from such a diverse collection of papers, a roadmap will be helpful: We suggest the reader start with the first four sections of T.Y. Lin's paper ► Granular Computing: Practices, Theories, and Future Directions. There, the reader may want to pay special attention to the first three examples:

E1) Human body is granulated into head, neck, etc.
So far, there is no flawless formal model. Obvious models do not work satisfactorily; we need a much more subtle theory.

E2) Space-time is granulated intuitively into infinitesimal granules.

E3) The Heisenberg uncertainty principle.

The last two examples have been fully digested by mathematicians and scientists. There are two solutions to the first example, namely, topology and non-standard analysis, so the readers may want to skim through Lin's Section "Second GrC Models and Modern Examples" (about neighborhood system/pre-topology) in ► Granular Computing: Practices, Theories, and Future Directions and the two articles, ► Non-standard Analysis, An Invitation to by Wei-Zhe Yang, and ► Granular Computing and Modeling of the Uncertainty in Quantum Mechanics by Kow-Lung Chang. These are classical subjects; the authors are a mathematician and a physicist, respectively.

Having digested these readings, the reader may proceed to Zadeh's ► Fuzzy Logic to gain some feeling about the modern view, and may use it to examine the very first example, E1. The article is full of revolutionary ideas; it is, we believe, one of his best papers in presenting an overview of his idea.

Here we paraphrase or quote some of his assertions from his article: "There are many misconceptions about fuzzy logic. A common misconception is that fuzzy logic is fuzzy. In reality, fuzzy logic is not fuzzy. Fuzzy logic deals precisely with imprecision and uncertainty. In fuzzy logic, the objects of deduction are, or are allowed to be, fuzzy, but the rules governing deduction are precise."

Fuzzy logic is much more than a logical system. More specifically, fuzzy logic has many facets. "The principal facets are: the logical facet, FLl; the fuzzy-set-theoretic facet, FLs, the epistemic facet, Fle; and the relational facet, FLr."

To gain a glimpse into the nature of fuzzy logic, we examine some examples taken from his article. The left-hand column is "the familiar example of deduction in Aristotelian, bivalent logic. In this example, there is no impre-

cision and no uncertainty." On the other hand, the right-hand column is an example of fuzzy logic which is "in an environment of imprecision and uncertainty."

all men are mortal	most Swedes are tall
Socrates is a man	Magnus is a Swede
Socrates is mortal	it is likely that Magnus is tall

To deduce the answer from the premises (for the right-hand example), "it is necessary to precisiate the meaning of "most" and "tall," with "likely" interpreted as a fuzzy probability which, as a fuzzy number, is equal to "most." This simple example points to a basic characteristic of fuzzy logic, namely, in fuzzy logic precisiation of meaning is a prerequisite to deduction." In this example, "deduction is contingent on precisiation of "most," "tall," and "likely." The issue of precisiation has a position of centrality in fuzzy logic." "In fuzzy logic, the deduction is viewed as an instance of question-answering".
[Digression] To this point, we may want to observe that there is some similarity to Lin's context based reasoning method, which is a derivative of (ϵ, δ)-definition; see Lin's article "Second GrC Models and Modern Examples" in ► Granular Computing: Practices, Theories, and Future Directions on the Meaning of "Near".

To avoid mis-representation, we will not offer a summary here, but urge the reader to read Zadeh's article.

There are five areas that have great interactions with GrC, namely, (in alphabetical order) databases (especially data mining), fuzzy theory (especially fuzzy control), grid/cloud computing, rough set theory (implicitly extended to topology) and Social Networks. We will group these articles (plus general issues) accordingly. We shall start from rough set theory.

Granular Computing from Rough Set Theory

Briefly, rough set theory (RST) is a theory of equivalence relations (which are equivalent to partitions), so the RST community views GrC as a generalized partition theory. Recall that an equivalence relation is a reflexive, symmetric and transitive binary relation. Thus, the "next" generalizations are the tolerance relation (dropping transitive), the partial ordering relation (dropping symmetric), and the fuzzified equivalence relation.

We have collected the corresponding articles: ► Granulation of Knowledge: Similarity Based Approach in Information and Decision Systems by Lech Polkowski, ► Granular Computing and Data Mining for Ordered Data: The Dominance-Based Rough Set Approach by Salvatore Greco, Benedetto Matarazzo1, and Roman Slowinski, ► Rough and Rough-Fuzzy Sets in Design of Information Systems by Theresa Beaubouef and Frederick Petry and ► Multi-Granular Computing and Quotient Structure by Ling Zhang and Bo Zhang.

Thanks to Professor Polkowski's diligent search, we are aware that the term "tolerance relation" first occurred much earlier than one might expect. However, in the rough set community, Nieminen (1988) and Lin [7] were probably the first to use the terms "tolerance equality" and "neighborhood of tolerance," respectively. In fact, progress on the tolerance relation is quite far reaching; it is a complete Pawlak Theory (CPT); it is implicitly in [11].

Recall that a RST is called a CPT, if it includes analogous theories covered in Pawlak's book (except logic), namely, a theory of approximations and a complete knowledge representation system; the latter notion is defined below (Definition 1).

There are more results in this directions: Let (U, β) be a Global GrC Model with β being a full covering. Then the model (U, β) is a CPT, if β is a semi-group under intersection. However, there are counter examples, namely, some GrC on partial ordering cannot be a CPT.

The quotient structures, addressed by Zhang–Zhang, have been neglected by the rough set community, except some brief results by Lin. Quotient structures are essential for information hiding and knowledge representations.

The quotient structure of a partition is simply a classical set. Zhang–Zhang actually has considered topological cases, and Lin has considered pre-topological cases (neighborhood systems). In general, it can be much more complicated. We shall conclude this paragraph with an example from algebraic geometry: Let the universe U be the polynomial ring over the complex numbers, and β be the collection of all prime ideals. Then the quotient structure is the complex line plus a generic point; see "Granular and Related Structures" of Lin's article ► Granular Computing: Practices, Theories, and Future Directions for more details.

Granular Computing in Database Theory

Relational databases can be related to GrC in two views, Fifth or Fourth GrC models (Relational or Multi-Binary GrC Models).

A Fourth GrC Model $GDM_1 = (U, \beta)$ is called a Granular Data Model (GDM), if the granular structure $\beta = \{R_1, R_2, \ldots\}$ is a collection of equivalence relations. By giving each equivalence relation and its equivalence classes meaningful names, GDM becomes a relation instance, called an information table (IT) in RST. This naming process is called a knowledge representation. Let

$IT_1 = (U, A_1, A_2, \ldots)$ be a knowledge representation of GDM_1, where A_i, called an attribute, denotes the meaningful names of R_i. Pawlak observed a beautiful phenomena:

Theorem 1 $(U, R_1, R_2, \ldots) \Longleftrightarrow (U, A_1, A_2, \ldots)$, *up to isomorphism.*

From this theorem, we define:

Definition 1 A knowledge representation of the Fourth GrC model is said to be a complete knowledge representation if the representation (the process) induces an equivalence. In other words, the knowledge representation can re-capture the model.

Base on this theorem, RST can transform data mining on IT to granule-processing, namely, granular computing, on GDM. This view clarifies the nature of automated data mining and simplifies its computations. For example, the association rule mining is reduced to counting the cardinality of set theoretical expressions of granules. However, we should also observe that in such computations the semantic part of attribute values are ignored. In general, the semantics of attribute values are not utilized in the RST/GrC approach. So some data mining techniques, such as clustering, cannot be approached by RST, since clustering uses the metric of the ambient space of attribute values. In this respect, GrC may use techniques in computing with words to deal with semantic issues. However, we should also note that computing with words is still in its inception stage.

In this group, we have collected the following articles, ► Granular Computing, Information Models for by Steven A. Demurjian, ► Rule Induction, Missing Attribute Values and Discretization by Grzymala-Busse, ► Cooperative Multi-hierarchical Query Answering Systems by Zbigniew W. Ras, Agnieszka Dardzinska, ► Dependency and Granularity in Data-Mining by Tsumoto–Hirano, ► Rough Set Data Analysis by Shusaku Tsumoto, ► Granular Model for Data Mining by Anita Wasilewska, Ernestina Menasalvas, Here, we would like to note that Steven A. Demurjian's article views GrC from the database, not the RST, perspective.

Granular Computing in Social Networks

By interpreting a granule as an "ordered set" or a tuple (in some relations), Lin formalizes Social Networks into a Fifth GrC model, called the Relational GrC Model. Articles ► Granular Computing System Vulnerabilities: Exploring the Dark Side of Social Networking Communities by Steve Webb, James Caverlee, Calton Pu, and ► Social Networks and Granular Computing by Churn-Jung Liau investigate two issues in social networks: security and positional equivalence.

This area actually started much earlier. In the mid-1970s, Atkin started to investigate mathematical structure in human affairs [1]. In fact, he reached a simplicial complex in [2], which is symmetric Fifth GrC model or equivalently a Second GrC model(Global GrC model).

Granular Computing and Fuzzy Set Theory

Basically, any assertion about classical sets can be fuzzified, so the intention here is to address GrC applications that extend beyond set theory.

A fuzzy set is defined by a membership function, which is a bounded non-negative real-valued function. So it is rather natural to investigate the generalization of fuzzy set theory to general function theory; even to more general cases, such as random variables (measurable functions) and to generalized functions (such as Dirac Functions). So, we have proposed a Sixth GrC model, which is a function-based GrC model.

In most known applications, we often need some extra properties on the granular structure, such as the universal approximation property. Using Banach space's language, it implies that the granular structure is a Schauder base. The collection of membership functions used in fuzzy control, and the activation functions in neural networks (e. g., Radial–Basis–Functions), all have such a property.

For this group, we have selected four papers: ► Genetic-Fuzzy Data Mining Techniques by Tzung-Pei Hong, Chun-Hao Chen and Vincent S. Tseng, ► Fuzzy System Models Evolution from Fuzzy Rulebases to Fuzzy Functions by I.B. Turksen, ► Granular Neural Network by Yan-Qing Zhang, and ► Fuzzy Probability Theory by Michael Beer.

We should note that many soft computing papers can be classified in this category; however, it has its own section, so we recommend that the readers visit that section. We specifically recommend reading some papers on Type II fuzzy sets, whose membership functions can be viewed as granule-valued (fuzzy-number-valued) membership functions.

We also recommend that readers compare Michael Beer's ► Fuzzy Probability Theory and non-standard probability theory (in Wei-Zhe Yang's article). From them, the reader might draw his own conclusions as to what a granular probability (granule-valued probability) theory might be. One could ask even further questions, such as, what is a granular function (granule valued function)?

Grid/Cloud Computing – A New Addition to GrC

In the end, we would like to note that a new technology, called cloud computing (a subset of grid computing) is intrinsically a technology of granular computing. This is a most natural example of a Seveth GrC models (a Turing machine-based GrC Model), even though the name GrC is foreign to the experts in this field. We are expecting, in the near future, that this area will be the top group that influences GrC development.

General Issues in Granular Computing

Two articles, ► Granular Computing, Principles and Perspectives of by Jianchao Han and Nick Cercone, and ► Granular Computing, Philosophical Foundation for by Zhengxin Chen, are about general views on GrC. However, they are basically addressing First, Second and Third models. Professor Chen's three-dimensional views, philosophical, technical, and social/application dimensions. are quite intriguing and may trigger more general investigations into all of GrC. In this group, we will add Kow-Lung Chang, Tsau Young Lin, Wei-Zhe Yang, Lotfi A. Zadeh Zadeh's articles.

Conclusions

As GrC is an emerging field, many directions are possible (Zadeh [13,14,15]; Lin [6] pre-CrC, [7,8,9]; Bargiela and Pedrycz [3]; Jankoswki and Skowron [4]). Some important works, due to lack of time and some other factors (including ignorance of this section editor), may not be included here. We apologize for the omissions, with the hope that many of them are referenced by some of the collected articles.

In editing this section, we have adopted an incremental approach: all papers are scientific papers, the only visionary paper is Lotfi Zadeh's "Fuzzy Logic"; and even in that paper, many of the visions are already accomplished.

Looking at the collection, we believe GrC is a promising field; here are some indications:

1. The recent emergence of cloud computing to GrC (our view) may give GrC a momentum to the real world applications.
2. From the abstract point of view (category theory based models) GrC and databases have the same abstract structures. This implies that GrC as an abstract concept can be technologically realized.
3. GrC has many links to various branch of mathematics, namely, algebraic topology (simplical complex) homological algebra (extension functor), algebraic geometry (Spec(R)). Topological spaces, probability and belief functions. These facts indicate, that mathematically, GrC contains some important common structures that are touched by many distinct mathematical fields.

Finally, we would like to extend our thanks to all of the authors for their support of this project, to Professor Zadeh for his generous guidance and advice, to Dr. Robert Meyers for the opportunity to edit this section.

Bibliography

1. Atkin RH (1974) Mathematical strucutres in human affairs. In: Heinemann educaton book. Pitman Press, Bath. ISBN 0 435 820257
2. Atkin RH (1977) Combinatorial connectivites in socail systems. Birkhäuser, Basel
3. Bargiela A, Pedrycz W (2002) Granular computing. Kluwer, Boston
4. Jankowski A, Skowron A (2007) Toward rough-granular computing. In: Proceedings of the 11th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, (RSFDGrC'07). Toronto, Canada. Lecture Notes in AI. Springer, pp 1–12
5. Lin TY (1988) Neighborhood systems and relational database. In: Proceedings of CSC'88. ACM Press, New York, pp 725
6. Lin TY (1989) Chinese wall security policy – an Aggressive Model. In: Proceedings of the Fifth Aerospace Computer Security Application Conference, December 4–8, 1989. Kluwer 2004, pp 286–293
7. Lin TY (1989) Neighborhood systems and approximation in database and knowledge base systems. In: Proceedings of the Fourth International Symposium on Methodologies of Intelligent Systems (Poster Session), pp 75–86
8. Lin TY (1998) Granular computing on binary relations I: data mining and neighborhood systems. In: Skoworn A, Polkowski L (eds) Rough Sets In Knowledge Discovery. Physica-Verlag, Heidelberg, pp 107–121
9. Lin TY (1998) Granular computing on binary relations II: rough set representations and belief functions. In: Skoworn A, Polkowski L (eds) Rough sets in knowledge discovery. Physica-Verlag, Heidelberg, pp 121–140
10. Lin TY (1999) Granular computing: fuzzy logic and rough sets. In: Zadeh L, Kacprzyk J (eds) Computing with words in information/intelligent systems. Springer Lecture Notes in AI. Physica-Verlag, Heidelberg, pp 183–200
11. Lin TY (2006) A roadmap from rough set theory to granular computing. RSKT 33–41
12. Zadeh LA (1996) The key roles of information granulation and fuzzy logic in human reasoning. In: IEEE International Conference on Fuzzy Systems, September 8–11, 1
13. Zadeh LA(1979) Fuzzy sets and information granularity. In: Gupta M, Ragade R, Yager R (eds) Advances in fuzzy set theory and applications. North-Holland, Amsterdam, pp 3–18
14. Zadeh LA (1998) Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems. Soft Comput 2:23–25
15. Zadeh LA (1998) Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets Syst 90:111–127

Granular Computing and Modeling of the Uncertainty in Quantum Mechanics

KOW-LUNG CHANG
Physics Department, National Taiwan University, Taipeh, Taiwan

Article Outline

Glossary
Definition of the Subject
Introduction
Quantum Postulates and Associated Propositions
Dirac's Bra and Ket Notations and the Realization of Wave Function
Realization of q-Representation in Quantum Mechanics
Fourier Transformation and p-Representation
Quantum Uncertainty and Non-Compatibility of Observables P and X
Conclusion
Future Directions
Bibliography

Glossary

Newtonian mechanics The classical mechanics based on Newton's law of motion.

Uncertainty Also commonly referred to as error; deviation from the average value.

Extrinsic uncertainty The uncertainty due to the systematic error and random error.

Intrinsic uncertainty The uncertainty due to the nature of particle-wave duality in the quantum system.

$\boldsymbol{\psi(r)}$ q-representation of the quantum state, called state function or wave function in coordinate space.

$\boldsymbol{\varphi(p)}$ p-representation of the quantum state, called the state function or the wave function in momentum space.

$\boldsymbol{\rho(r)}$ Probability density in coordinate space, defined as $\rho(\boldsymbol{r}) = \psi^*(\boldsymbol{r})\psi(\boldsymbol{r}) = |\psi(\boldsymbol{r})|^2$.

$\boldsymbol{\tilde{\rho}(p)}$ Probability density in momentum space, defined as $\tilde{\rho}(\boldsymbol{p}) = \varphi^*(\boldsymbol{p})\varphi(\boldsymbol{p}) = |\varphi(\boldsymbol{p})|^2$.

$\boldsymbol{|\psi(r)|}$ Probability amplitude of state $\psi(\boldsymbol{r})$.

Particle-wave duality A quantum object exhibits both the nature of a particle and a wave.

Wave vector $\boldsymbol{k} = \boldsymbol{p}/\hbar$.

Wave number One-dimensional wave vector, $k = p/\hbar$.

Quantum state The physical state in a subatomic quantum system.

0th Postulate of quantum mechanics Regarding quantum states as elements of Hilbert space $\mathcal{H}$.

1st Postulate of quantum mechanics Assigning each dynamical variable in a quantum system a unique linear Hermitean operator in Hilbert space $\mathcal{H}$.

2nd Postulate of quantum mechanics The set of eigenvectors of a given observable forms the bases of a Hilbert space.

3rd Postulate of quantum mechanics Poisson brackets in classical mechanics are replaced by commutators of the corresponding observables according to the relations in Eq. (5)

$\boldsymbol{[R, S]}$ Commutator of operator R and operator S defined as $RS - SR$.

$\boldsymbol{\{R, S\}}$ Anti-commutator of operator R and operator S defined as $RS + SR$.

Eigenvalue and eigenvector For operator A acting upon a particular state ψ_a, such that $A\psi_a = a\psi_a$, then the scalar number a and the state ψ_a are called respectively the eigenvalue and eigenvector of operator A.

Inner product A numerically valued function of the ordered pair of vectors ψ and φ, denoted by (ψ, φ) such that $(\psi, \varphi) = (\varphi, \psi)^*$. In Dirac's notation, it takes the forms $\langle\psi|\varphi\rangle = \langle\varphi|\psi\rangle^*$.

Hilbert space A complete vector space with norm defined as the inner product.

Dual space The space formed by the set of all functionals satisfying the linearity conditions.

Hermitean operator An operator which is self-adjoint; that is, an operator A equals its Adjoint conjugate, $A = A^+$.

Adjoint conjugate operator For an operator A and a pair of vectors ψ and φ, the Adjoint operator to A, denoted by A^+ if it satisfies the relation

$$(\psi, A\varphi) = (\varphi, A^+\psi)^* = (A^+\psi, \varphi)\,.$$

$\boldsymbol{|\psi\rangle}$ A ket vector in Hilbert space.

$\boldsymbol{\langle\psi|}$ A bra vector in dual space.

Compatible observables Physical observables that commute to each other.

Projection operator $|a_i\rangle\langle a_i|$ a Hermitean operator that projects any vector in $\mathcal{H}$ onto a subspace.

Closure relation Direct sum of all project operators equals identity operator.

$\boldsymbol{\delta}$-function A distribution function denoted by $\delta(x - a)$ such that

$$\int_D f(x)\delta(x-a)\mathrm{d}x = f(a) \quad \text{and}$$
$$\int_D f(x)\delta'(x-a)\mathrm{d}x = -f'(a) \quad \text{if } a \in D\,,$$
$$\int_D f(x)\delta(x-a)\mathrm{d}x = 0 \quad \text{if } a \notin D\,.$$

Cauchy–Schwarz inequality Given vectors $|\alpha\rangle$ and $|\beta\rangle$ such that

$$\langle\alpha|\alpha\rangle\langle\beta|\beta\rangle \geq \langle\alpha|\beta\rangle\langle\beta|\alpha\rangle \,.$$

$\boldsymbol{G_{\mathrm{d}}}$ Deviation operator with respect to observable G defined as $G_{\mathrm{d}} = G - \langle G\rangle I$

Definition of the Subject

For data collection or data acquisition in granular computing or modeling, particular attention should be paid to data that is involved with the position measurement and/or the momentum measurement of a particle in a physical system. For a system that is governed by Newton's law of motion, measurements of the positions and the momenta of a classical mass point can be made as precise as one wishes if the best measurement conditions are available. When we perform physical measurements in microscopic systems at subatomic level, the laws of Newtonian mechanics cease to apply and instead a new theory, quantum theory, is introduced to account for the quantum phenomena of the subatomic system.

Uncertainty in the measurement of a quantum system comes from two origins. The first is called the extrinsic uncertainty. Quantum measurements, just as in any scientific measurement, are never exact. There exists a systematic error or uncertainty as well as random ones during the processes of measurement for various reasons such as inaccurate calibration of instruments, limitation in resolution of apparatus or meters, variations in temperature or humidity or even the position adopted by the individual observer in reading data etc. The second is called the intrinsic uncertainty which arises only in the case of a quantum mechanical system.

This particular quantum uncertainty is inherent in the very nature of particle-wave duality [1] for a subatomic quantum system. The intrinsic uncertainty exists always even if the extrinsic uncertainty is eliminated or minimized to the level of negligibility.

In order to describe the particle-wave duality for a subatomic quantum system, one introduces the wave function, or state function, $\psi(\boldsymbol{r})$ to represent the quantum state. $\psi(\boldsymbol{r})$, which provides all the dynamic quantities of the system, and is related to the probability density by $\rho(\boldsymbol{r}) = \psi^*(\boldsymbol{r})\psi(\boldsymbol{r}) = |\psi(\boldsymbol{r})|^2$. Therefore, the average position of the quantum object $\langle\boldsymbol{r}\rangle$ can be obtained through the following integration

$$\langle\boldsymbol{r}\rangle = \int \rho(\boldsymbol{r})\boldsymbol{r}\mathrm{d}^3\boldsymbol{r} = \int \psi^*(\boldsymbol{r})\boldsymbol{r}\psi(\boldsymbol{r})\mathrm{d}^3\boldsymbol{r} \,. \tag{1}$$

The uncertainty in position, dented by $\Delta\boldsymbol{r}$, can then be expressed by

$$\begin{aligned}(\Delta\boldsymbol{r})^2 &= \int \rho(\boldsymbol{r})(\boldsymbol{r} - \langle\boldsymbol{r}\rangle)^2\mathrm{d}^3\boldsymbol{r} \\ &= \int \psi^*(\boldsymbol{r})(\boldsymbol{r} - \langle\boldsymbol{r}\rangle)^2\psi(\boldsymbol{r})\mathrm{d}^3\boldsymbol{r} \,.\end{aligned} \tag{2}$$

Similarly, the uncertainty in momentum, $\Delta\boldsymbol{p}$ can be written as

$$\begin{aligned}(\Delta\boldsymbol{p})^2 &= \int \tilde{\rho}(\boldsymbol{p})(\boldsymbol{p} - \langle\boldsymbol{p}\rangle)^2\mathrm{d}^3\boldsymbol{p} \\ &= \int \varphi^*(\boldsymbol{p})(\boldsymbol{p} - \langle\boldsymbol{p}\rangle)^2\varphi(\boldsymbol{p})\mathrm{d}^3\boldsymbol{p} \,,\end{aligned} \tag{3}$$

where $\varphi(\boldsymbol{p})$ is the state function of the quantum system in momentum space.

The quantum uncertainties of the system obey the following relation [2]:

$$\begin{aligned}\Delta x\Delta p_x &\geq \hbar/2 \\ \Delta y\Delta p_y &\geq \hbar/2 \\ \Delta z\Delta p_z &\geq \hbar/2 \,,\end{aligned} \tag{4}$$

where $\hbar = h/2\pi$, and h is called the Planck's constant.

Introduction

Some basic mathematical tools are necessary in order to understand the uncertainty relations of quantum mechanics in a more rigorous sense. As discussed in the last section, it is only a vague answer to the statement that a quantum object possesses both the nature of a point particle and the nature of a wave at the same time, so that one visualizes the quantum object as a matter wave with some sort of distribution in coordinate space as well as in wave vector space. Therefore, neither can one find a localized point quantum object, nor a strictly monochromatic wave motion in the quantum system.

Next we shall discuss the relations of the quantum state and the element in Hilbert space as well as the physical observables and the linear Hermitean operator. The connection between the eigenvalue problems and the preparation of a quantum state for a particular measurement of physical observables are also explored in depth. Dirac's notations of the bra and ket vector in Hilbert space are also reviewed and their connection with the q-representation and p-representation of wave functions are derived. The uncertainty relations can therefore formally be proved as the consequence of non-compatibility of two physical observables of position and momentum.

Quantum Postulates and Associated Propositions

0th Postulate of Quantum Mechanics

For every quantum system, there exists an abstract entity, called the state (or state function or wave function in q-representation) which provides the information of all the dynamical quantities of the system; such as positions, momenta, energy, angular momentum, spin, charge, ... etc. All the possible states $\psi, \varphi, \chi, \ldots$ etc. of a given quantum system are elements of Hilbert space $\mathcal{H}$.

1st Postulate of Quantum Mechanics

For the measurement of each dynamical variable in the quantum system, such as the total energy of the system, or 3rd component of the orbital angular momentum ... etc., there associates a unique linear Hermitean operator in Hilbert space $\mathcal{H}$ corresponding to each dynamical variable. The set of linear Hermitean operators in $\mathcal{H}$ are called physical observables, or simply observables.

Dynamical variable		Physical observable
E (energy)	$\rightarrow$	H (Hamiltonian operator)
ℓ_3 (3rd component of $\boldsymbol{\ell}$)	$\rightarrow$	L_3 (3rd component of angular momentum operator $\boldsymbol{L}$)

The physical quantity a corresponding to the observable of the quantum state ψ is obtained from the inner product of the order pair ψ and $A\psi$, where A is the corresponding linear Hermitean operator associated with dynamical variable a. The inner product $(\psi, A\psi)$ is called the expectation value, or the average value associated with the dynamical operator A for state ψ of the quantum system.

$$a = (\psi, A\psi) .$$

In general, the operator A acting on ψ will change ψ into another element φ in Hilbert space $\mathcal{H}$, which implies that the action of the measurement usually would disturb the quantum system and brings the original quantum state ψ into a new state φ by the external perturbation during the process of measurement, i. e.

$$A\psi = \varphi .$$

In a particular case, if an operator A, acts on ψ_a such that

$$A\psi_a = a\psi_a ,$$

i. e. when A acts upon a particular quantum state ψ_a, the resultant state is the same as the one before except multiplying by a scalar number a, then it is said that the quantum system is prepared for the measurement of the dynamical variable associated with the physical observable A. This particularly prepared quantum state ψ_a is called the eigenstate of the operator A. The scalar numerical value a, called the eigenvalue of operator A, is the result of the measurement for the dynamical observable.

Since the value of the physical measurement is always a real quantity, we therefore conclude the first Proposition:

Proposition 1 *The eigenvalues for a Hermitean operator are real.*

For $A = A^+$, and letting ψ_a be the eigenstate of A, then

$$\begin{aligned} a(\psi_a, \psi_a) &= (\psi_a, A\psi_a) = (A^+\psi_a, \psi_a) \\ &= (A\psi_a, \psi_a) = (\psi_a, A\psi_a)^* = a^*(\psi_a, \psi_a) \end{aligned}$$

Since $(\psi_a, \psi_a) \neq 0$, we have $a = a^*$.

It is also ready to show Proposition 2.

Proposition 2 *Two eigenvectors of the same Hermitean operator are orthogonal if the corresponding eigenvalues are unequal.*

Let

$$A\psi_1 = a_1\psi_1 , \quad A\psi_2 = a_2\psi_2 .$$

And if $a_1 \neq a_2$, then

$$\begin{aligned} (\psi_1, A\psi_2) &= a_2(\psi_1, \psi_2) = (A^+\psi_1, \psi_2) \\ &= (A\psi_1, \psi_2) = a_1(\psi_1, \psi_2) . \end{aligned}$$

Therefore, $(a_1 - a_2)(\psi_1, \psi_2) = 0$, that concludes that ψ_1 is orthogonal to ψ_2 if $a_1 \neq a_2$. Namely $(\psi_1, \psi_2) = 0$.

With these propositions, we can formulate the 2nd postulate of quantum mechanics.

2nd Postulate of Quantum Mechanics

The set of eigenvectors ψ_a for all possible value a of a given Hermitean operator corresponding to a physical observable form the bases of a Hilbert space $\mathcal{H}$.

If there exists a complete set of linearly independent vectors ψ_a which are eigenvectors of both operators R and S, then the two corresponding physical observables, R and S are said to be compatible, we then have:

Proposition 3 *If two observables are compatible, their corresponding operators commute, i. e. if*

$$\begin{aligned} R\psi_a &= r_a\psi_a \\ S\psi_a &= s_a\psi_a \end{aligned}$$

Then $(RS - SR)\psi_a = [R, S]\psi_a = 0$.

Any vector ψ in $\mathcal{H}$ can be expressed in terms of linear combination of ψ_a, i. e.

$$\psi = \sum_a \alpha_a\psi_a ,$$

one can easily verify that it also satisfies $[R, S]\psi = 0$ *for any state* ψ *in* $\mathcal{H}$.

Therefore, the observable R and the observable S commute.

From the point of view of physical measurement, the compatibility of two observables implies that one is able to prepare a single quantum system for the precise measurements of two different dynamical quantities corresponding to R and S respectively at the same time.

Unfortunately, not all physical observables are compatible. The position and the momentum are just an example, and the non-compatibility of position operator X and momentum operator P is in fact the foundation of quantum theory that forms the 3rd Postulate of Quantum Mechanics.

3rd Postulate of Quantum Mechanics

Every Poisson bracket in classical mechanics for canonical variables (p_i, q_i) is replaced by the commutator of the corresponding operators in the quantum system with the following relations

$$\begin{array}{lcl} \text{Poisson bracket} & & \text{Quantum commutator operator} \\ [q_i, q_j] = 0 & \rightarrow & [X_i, X_j] = 0 \\ [p_i, p_j] = 0 & \rightarrow & [P_i, P_j] = 0 \\ [p_i, x_i] = \delta_{ij} & \rightarrow & [P_i, X_j] = \hbar / i\delta_{ij} \,. \end{array} \tag{5}$$

The non-commuting of the dynamical observables P_i and X_i will in fact not only lead to the fundamental quantization of the microscopic physical system, it also leads to the deviations in the position measurement and the momentum measurement. The non-commutativity of P and X in the 3rd Quantum Postulate signifies that one is not able to prepare a quantum state for the simultaneous measurement of momentum and position with absolute precision. Therefore, one shall regard the uncertainty relations of Eq. (4) as the direct consequence of this Quantum Postulate.

Dirac's Bra and Ket Notations and the Realization of Wave Function

To understand that quantum uncertainty is the direct consequence of two non-commuting physical observables, it is convenient to introduce Dirac's notations of bra vectors and ket vectors. A ket vector, written as $|a\rangle$, represents the abstract quantum state ψ_a, with a inside the notation $|a\rangle$ stands for the eigenvalue of the observable A, and the eigenvalue equation becomes

$$A|a\rangle = a|a\rangle \,.$$

Similarly we denote the states $\psi, \varphi, \chi, \ldots$ etc. respectively by the ket vector notations by $|\psi\rangle, |\varphi\rangle, |\chi\rangle \ldots$ etc.

The bra vectors are the elements in another vector space that are dual to ket vector space.

The bra vector takes the form $\langle\psi|$. The inner product in Dirac's notation is expressed by taking a pair of dual vectors, i. e. a bra vector $\langle\psi|$ and a ket vector $|\varphi\rangle$, and putting them side by side as $\langle\psi| \bullet |\varphi\rangle$ of which one always simplifies the notation as $\langle\psi|\varphi\rangle$. The inner product is a liner functional that defined a complex number for every ket vector in Hilbert space $\mathcal{H}$, such that

$$\langle\psi| \bullet (\alpha|\varphi\rangle + \beta|\chi\rangle) = \alpha\langle\psi|\varphi\rangle + \beta\langle\psi|\chi\rangle \,.$$

The set of bra vectors that define all linear functionals associated with the ket vector in $\mathcal{H}$ forms a dual Hilbert space $\widetilde{\mathcal{H}}$. The inner product is a complex number, i. e.

$$\langle\psi|\varphi\rangle = \langle\varphi|\psi\rangle^* \,,$$

because bra vectors and ket vectors are elements of two different Hilbert spaces, and they are dual to each other. The inner product of ψ and φ denoted previously by (ψ, φ) is then replaced by $\langle\psi|\varphi\rangle$. The expectation value of the physical observable A becomes

$$\langle\psi|A|\psi\rangle = (\psi, A\psi) \,.$$

The advantage of using Dirac's notations is in the construction of the Projection Operator, of which a vector in the vector space is projected onto a subspace. Let us denote the projector operator $\wp_a$ by putting the bra vector followed by a ket vector immediately as

$$\wp_a = |a\rangle\langle a| \,,$$

therefore the vector $|\psi_a\rangle$ is automatically constructed when the Projection operator $\wp_a$ applies upon any vector $|\psi\rangle$, that is

$$|\psi_a\rangle = \wp_a|\psi\rangle = |a\rangle\langle a|\psi\rangle \,.$$

Let $|i\rangle$ be the set of the eigenkets of observable A, with the discrete eigenvalue a_i, then

$$A|i\rangle = a_i|i\rangle \,.$$

It is often normalized to the eigenvector such that

$$\langle i|j\rangle = \delta_{ij} \,.$$

Any vector $|\psi\rangle$ in $\mathcal{H}$ can be expressed in terms of the base vectors with the coefficients $\langle i|\psi\rangle$ in the following linear combination,

$$|\psi\rangle = \sum_i (\langle i|\psi\rangle)|i\rangle = \sum_i |i\rangle\langle i|\psi\rangle \,.$$

The summation $\sum_i |i\rangle\langle i|$ in fact is just the direct sum of projection operator $\wp_i = |i\rangle\langle i|$, which obeys the following closure relation

$$\mathrm{I} = \sum_i \wp_i = \sum_i |i\rangle\langle i| \,.$$

In many cases, the eigenvalue of observables becomes continuous. Taking observable X, the position operator for instance in the quantum system of interest, and for the sake of simplicity, we shall consider only a one-dimensional case, the eigenvalue equation of this quantum system reads

$$\mathrm{X}|x\rangle = x|x\rangle \,,$$

where $|x\rangle$ is the eigenket of observable X with eigenvalue x that is a continuous parameter representing the position of the quantum object.

The normalization of an eigenvector with continuous eigenvalue is taken as a δ-function

$$\langle x|x'\rangle = \delta(x - x') \,,$$

and the corresponding closure relation becomes

$$\mathrm{I} = \int |x\rangle \mathrm{d}x \langle x| \,.$$

Realization of *q*-Representation in Quantum Mechanics

Let us introduce a unitary operator

$$U(P;\xi) = \mathrm{e}^{-\frac{i}{\hbar}\xi_P}$$

where ξ is a real continuous parameter and P is the momentum operator. It is unitary because

$$U^+(P;\xi) = \mathrm{e}^{\frac{i}{\hbar}\xi_P}$$

and

$$U(P;\xi)U^+(P;\xi) = U^+(P;\xi)U(P;\xi) = \mathrm{I} \,.$$

If one expands $U(P;\xi)$ in terms of a power series in P, i. e.

$$U(P;\xi) = \mathrm{I} + \left(-\frac{i}{\hbar}\xi P\right) + \frac{1}{2!}\left(-\frac{i}{\hbar}P\xi\right)^2 + \ldots \,,$$

and calculates the commutator $[X, P^n]$ team by team, and makes use of

$$[X, P^n] = i\hbar n P^{n-1} \,,$$

which is based on the 3rd postulate of Quantum Mechanics, then one obtains

$$[X, U(P;\xi)] = \xi U(P;\xi)$$

or

$$XU(P;\xi) = U(P;\xi)(X + \xi\mathrm{I}) \,.$$

It can be easily verified that $U(P;\xi)|x\rangle$ is also an eigenket of X with eigenvalue $(x + \xi)$, i. e.

$$\begin{aligned} XU(P;\xi)|x\rangle &= U(P;\xi)(X + \xi I)|x\rangle \\ &= (x + \xi)U(P;\xi)|x\rangle \quad \text{and} \\ U(P;\xi)|x\rangle &= |x + \xi\rangle \,, \end{aligned}$$

where we have chosen the phase factor of the new eigenket $|x + \xi\rangle$ to be 1.

Consider the matrix element of the operator, defined as $\langle x|U(P;\xi)|x'\rangle$ for infinitesimal value of ξ, then by keeping up to first order in ξ, one has

$$\langle X|U(P;\xi)|x'\rangle = \langle x|x' + \xi\rangle \doteq \left\langle x \left| \left(\mathrm{I} - \frac{i}{\hbar}\xi P\right) \right| x' \right\rangle \,.$$

With the properties of the δ-function, one concludes

$$\begin{aligned} \langle x|P|x'\rangle &= \frac{\hbar}{i}\lim_{\xi\to 0}\frac{1}{\xi}\{\delta(x - x') - \delta(x - x' - \xi)\} \\ &= \frac{\hbar}{i}\frac{\mathrm{d}}{\mathrm{d}x}\delta(x - x') = -\frac{\hbar}{i}\frac{\mathrm{d}}{\mathrm{d}x'}\delta(x - x') \,. \end{aligned}$$

Therefore one is ready to establish the connection between the axiomatic quantum postulates and the q-representation in quantum mechanics by defining the wave function $\psi(x)$ to be the inner product of $|\psi\rangle$ and $|x\rangle$, i. e.

$$\psi(x) = \langle x|\psi\rangle \,,$$

and q-representation of observable P, denoted by $\wp$ leads to

$$\begin{aligned} \wp\psi(x) = \langle x|P|\psi\rangle &= \int \langle x|P|x'\rangle \mathrm{d}x' \langle x'|\psi\rangle \\ &= -\frac{\hbar}{i}\int \frac{\mathrm{d}}{\mathrm{d}x'}\delta(x - x')\mathrm{d}x'\psi(x') = \frac{\hbar}{i}\frac{\mathrm{d}}{\mathrm{d}x}\psi(x) \,, \end{aligned}$$

namely, the q-representation of the momentum operator becomes the differential operator with respect to coordinate x. The quantum theory formulated in the previous

sections can be translated into the q-representation as follows

Abstract formulation	q-representation
State ψ	Wave function $\psi(x)$
Observable X	Coordinate x
Observable P	Differential operator $\wp = \frac{\hbar}{i}\frac{\mathrm{d}}{\mathrm{d}x}$
Any observable $F(X,P)$	$F(x,\wp) = F\left(x, \frac{\hbar}{i}\frac{\mathrm{d}}{\mathrm{d}x}\right)$

Fourier Transformation and *p*-Representation

As we have mentioned a quantum system can be described either in the coordinate space of x, or in the momentum space (or wave number space) of p. The p-representation of the quantum state $\varphi(p)$ is similarly defined as the inner product of the order pair vectors $|p\rangle$ and $|\psi\rangle$, i. e.

$$\varphi(p) = \langle p|\psi\rangle \,.$$

It is of interest to relate $\varphi(p)$ with $\psi(x)$ by inserting an Identity Projection operator before the state $|\psi\rangle$, i. e.

$$\varphi(p) = \int \langle p|x'\rangle \mathrm{d}x' \langle x'|\psi\rangle = \int \langle p|x'\rangle \mathrm{d}x' \psi(x') \,. \quad (6)$$

Since

$$p\langle p|x\rangle = \langle p|P|x\rangle = \int \langle p|x'\rangle \mathrm{d}x' \langle x'|P|x\rangle$$
$$= \frac{\hbar}{i}\int \langle p|x'\rangle \frac{\mathrm{d}}{\mathrm{d}x'}\delta(x-x') = -\frac{\hbar}{i}\frac{\mathrm{d}}{\mathrm{d}x}\langle p|x\rangle \,,$$

one finds, with proper normalization, that

$$\langle p|x\rangle = \frac{1}{\sqrt{2\pi\hbar}} \mathrm{e}^{-\frac{i}{\hbar}px} \,,$$

and Eq. (6) can be expressed as

$$\varphi(p) = \frac{1}{\sqrt{2\pi\hbar}} \int \mathrm{e}^{-\frac{i}{\hbar}px} \psi(x) \mathrm{d}x \,.$$

Conversely, $\psi(x)$ is the inverse Fourier Transform of $\varphi(x)$, i. e.

$$\psi(x) = \frac{1}{\sqrt{2\pi\hbar}} \int \mathrm{e}^{\frac{i}{\hbar}px} \varphi(p) \mathrm{d}p \,.$$

Quantum Uncertainty and Non-Compatibility of Observables *P* and *X*

According to the 3rd Quantum Postulate that $[P,X] = \frac{\hbar}{i}I$, where I is an Identity operator, one can show that the deviation operators, or the uncertainty operators of P and X, defined by

$$P_\mathrm{d} = P - \langle P\rangle I \,,$$
$$X_\mathrm{d} = X - \langle X\rangle I \,,$$
$$\text{where } \langle P\rangle = \langle\psi|P|\psi\rangle \equiv \bar{p} \,,$$
$$\text{and } \langle X\rangle = \langle\psi|X|\psi\rangle \equiv \bar{x}$$

are the average value of position and momentum respectively and the commutator of these deviation operators P_d and X_d will also follow the same quantization rule, i. e.

$$[P_\mathrm{d}, X_\mathrm{d}] = \frac{\hbar}{i} I \,.$$

Let us introduce the anti-commutator operator, defined by

$$P_\mathrm{d}X_\mathrm{d} + X_\mathrm{d}P_\mathrm{d} = \{P_\mathrm{d}, X_\mathrm{d}\} \,, \quad \text{and denote it by } \Sigma \,,$$

which is a Hermitean operator. Therefore, one can express the product operator of X_d and P_d as

$$P_\mathrm{d}X_\mathrm{d} = \frac{1}{2}[P_\mathrm{d}, X_\mathrm{d}] + \frac{1}{2}\sum \,,$$
$$X_\mathrm{d}P_\mathrm{d} = -\frac{1}{2}[P_\mathrm{d}, X_\mathrm{d}] + \frac{1}{2}\sum \,.$$

If we define the ket vectors $|\alpha\rangle$ and $|\beta\rangle$ as

$$X_\mathrm{d}|\psi\rangle = |\alpha\rangle \,,$$
$$P_\mathrm{d}|\psi\rangle = |\beta\rangle \,,$$

then the square of position uncertainty $(\Delta x)^2$ as well as the square of the momentum uncertainty $(\Delta p)^2$ can be expressed respectively as

$$(\Delta x)^2 = \langle\alpha|\alpha\rangle = \left\langle\psi\middle|X_\mathrm{d}^2\middle|\psi\right\rangle \,,$$
$$(\Delta p) = \langle\beta|\beta\rangle = \left\langle\psi\middle|P_\mathrm{d}^2\middle|\psi\right\rangle \,.$$

The quantum uncertainty relation takes the expression

$$(\Delta x)^2(\Delta p)^2 = \langle\alpha|\alpha\rangle\langle\beta|\beta\rangle$$
$$\geq \langle\alpha|\beta\rangle\langle\beta|\alpha\rangle = \left\langle\psi\middle|\left(-\frac{\hbar}{2i}I + \frac{1}{2}\sum\right)\psi\right\rangle$$
$$\times \left\langle\psi\middle|\left(\frac{\hbar}{2i} + \frac{1}{2}\sum\right)\middle|\psi\right\rangle \,,$$
$$(\Delta x)^2(\Delta p)^2 \geq \frac{\hbar^2}{4} + \frac{1}{4}\left|\left\langle\sum\right\rangle\right|^2 \,,$$

where Cauchy–Schwarz inequality was applied to obtain the right hand side of the first inequality sign. Since $|\langle\sum\rangle|^2$

is also positive definite because of the Hermiticity property of the operator Σ, we conclude that

$$\Delta x \Delta p \geq \hbar/2 \,,$$

which is the quantum uncertainty in a one-dimensional case. The equality sign in the last equation holds if the conditions $|\beta\rangle = \lambda|a\rangle$ and $\langle\psi|\sum|\psi\rangle = 0$ are met. By making use of

$$\langle\psi_{\min}|[P,X]|\psi_{\min}\rangle = \frac{\hbar}{i} \quad \text{and} \quad \left\langle\psi_{\min}\middle|\sum\middle|\psi_{\min}\right\rangle = 0 \,,$$

one obtains $\lambda = \frac{i\hbar}{2(\Delta x)^2}$, and the condition for the state function of the minimal uncertainty in a one-dimensional quantum system satisfies the following equations

$$\left(\frac{\hbar}{i}\frac{\mathrm{d}}{\mathrm{d}x} - \bar{p}\right)\psi_{\min}(x) = \frac{i\hbar}{2(\Delta x)^2}(x-\bar{x})\psi_{\min}(x) \,,$$

which can be solved with the solution behaving as a traveling Gaussian wave packet as follows,

$$\psi_{\min}(x) = \frac{1}{\sqrt[4]{2\pi(\Delta x)^2}} \mathrm{e}^{-\frac{(x-\bar{x})^2}{2(\Delta x)^2} + \frac{i}{\hbar}\bar{p}x} \,.$$

Conclusion

The uncertainties in quantum mechanics exist not only in the simultaneous measurement of the positions and the momenta as we have treated previously, uncertainties also exist in the simultaneous measurement of any pair of the dynamical variables that are not compatible, or their corresponding physical observables do not commute. For example, one is not able to obtain the data with absolute precision in the simultaneous measurements of any two components of the angular momentum L_i and L_j, because of the non-vanishing commutator $[L_i, L_j] = i\hbar \in_{ijk} L_k$. If we denote the deviation operator for the observable corresponding to angular momentum by

$$\boldsymbol{L}_{\mathrm{d}} = \boldsymbol{L} - \langle\boldsymbol{L}\rangle I \,,$$

one can easily show that

$$[L_{\mathrm{d}i}, L_{\mathrm{d}j}] = i\hbar \in_{ijk} L_k \,,$$

and the square of uncertainty in the ith and jth components of angular momentum, denoted by $(\Delta\ell_i)^2$ and $(\Delta\ell_j)^2$ respectively can be expressed as

$$(\Delta\ell_i)^2(\Delta\ell_j)^2 = \langle\psi|L_{\mathrm{d}i}^2|\psi\rangle\langle\psi|L_{\mathrm{d}j}^2|\psi\rangle \,.$$

Because of the Hermiticity of $\boldsymbol{L}$, and making use of the Cauchy–Schwarz inequality, one can derive that

$$(\Delta\ell_i)^2(\Delta\ell_j)^2 \geq \frac{\hbar^2}{4} \in_{ijk} \bar{L}_k^2 + \frac{1}{4}\bar{K}_{ij}^2 \,,$$

where $\bar{L}_k = \langle\psi|L_k|\psi\rangle$ and $\bar{K}_{ij} = \langle\psi|\{L_i, L_j\}|\psi\rangle$.

The uncertainty in simultaneous measurement of L_i and L_j depends upon the average value of the physical observable of the third component of angular momentum, as well as a positive definite number $\bar{K}_{ij}^2$. Therefore, the minimum of the uncertainty

$$(\Delta\ell_i)(\Delta\ell_j) \quad \text{becomes} \quad (\Delta\ell_i)(\Delta\ell_j)\Big|_{\min} = \frac{\hbar}{2}|\in_{ijk}\bar{L}_k| \,,$$

Therefore,

$$\Delta\ell_i\Delta\ell_j \geq \frac{\hbar}{2}|\in_{ijk}\bar{L}_k| \,,$$

which indicates that the uncertainty relation for the product of two components of a given angular momentum in a quantum system can not be independent of the average value of the third component. Many peculiar quantum phenomena that would not be happening in classical physics do exist in the subatomic quantum system. Quantum uncertainties corresponding to a pair of non-compatible observables can only be obtained through detailed investigation based on quantum dynamical theory.

Future Directions

The limitations in pursuit of accurate scientific measurement of some data in quantum mechanical systems are the direct consequence of the quantum dynamics in the commutation relations among various physical observables. One is able to minimize but not to eliminate the uncertainty in the simultaneous measurement in any pair of dynamical variables of which the commutator of the corresponding quantum operators does not vanish. For instance, a plane wave of one-dimension can be prepared in a quantum system to achieve the utmost monochromatic frequency only at the expense of the total accuracy in locality. As we have treated in the previous section Heisenberg's Uncertainty Relations in position measurement and momentum measurement can only be minimized by preparing the quantum state of a traveling Gaussian wave packet. In fact we can easily show that the state function in momentum space is also of a Gaussian traveling wave packet due to the property that the Fourier Transformation of a Gaussian distribution is again a Gaussian distribution. Therefore, it is imperative that one should execute uncertainty control or error management before performing the measurements for physical observables of non-compati-

bility in the quantum mechanical system. As we have already demonstrated in the simultaneous measurement of any two components of an angular momentum, the uncertainty can be minimized by preparing a quantum state with the least mean value of its third component.

Bibliography

1. de Broglie L (1923) Nature 112:540. (1924) Thesis, Paris. (1925) Ann. Phys 3:22
2. Merzbacher E (1970) Quantum Mechanics, 2nd edn. Wiley, New York
3. Jordan TF (1982) Linear operators for quantum mechanics, 2nd edn. Wiley, New York
4. Dirac PAM (1958) The principles of quantum mechanics, 4th edn. Clarendon Press, Oxford

Granular Computing, Philosophical Foundation for

Zhengxin Chen
Department of Computer Science, University of Nebraska at Omaha, Omaha, USA

Article Outline

Glossary
Definition of the Subject
Introduction
The Road to Granular Computing
The Nature of Granule and Granule Computing
Granule Measurement
Granular Structure
Granulation Provides a Unified View for Intelligent Problem Solving
Relationship with Soft Computing and Natural Computing
Relationship with Fundamental Issues of Computing and Complex Systems Problem Solving
Summary
Future Directions
Acknowledgments
Bibliography

Glossary

Fuzzy set and fuzzy logic Unlike a conventional set, in a fuzzy set, a fuzzy membership function is used to define the degree of an element belonging to the set. Fuzzy logic is a superset of conventional (Boolean) logic that has been extended to handle the concept of partial truth as defined by membership functions. Fuzzy logic contributes to the machinery of granular computing.

Granular computing (GrC) In a broad sense, granular computing is the general term referring to any computing theory/technology that involves elements and granules, with granule, granulated view, granularity, and hierarchy as its key concepts.

Granular structure Granular structure is a collection of granules in which the internal structure of each granule is visible.

Granularity The granularity of a level refers to the collective properties of granules in a level with respect to their sizes.

Granulation Granulation refers to the process of forming granules.

Granule As the fundamental concept in granular computing, a granule is a clump of elements drawn together by various criteria such as indistinguishability, equivalence, similarity, proximity or functionality.

Hierarchy In granular computing, hierarchy captures the ordering of levels.

Neighborhood system A neighborhood system of a point (an element) in the universe is the nonempty family of subsets (referred to as the neighborhood of that point) associated to it.

Rough set Rough set is a formal approximation of a conventional set, using a pair of sets as the lower and the upper approximations of the original set. Rough sets provide a single-layered granulation structure of the universe.

Definition of the Subject

In a little more than a decade, granular computing (GrC) has emerged as a major research field for further abstraction, generalization and unification of granule-based key concepts traditionally scattered throughout a wide range of scientific disciplines, offering new opportunities for systematic studies of challenging computational issues cross multiple disciplines. An examination of foundations of granular computing, particularly on its philosophical dimension, is extremely important because it should reveal the hidden nature of granular computing, shed useful light for future directions and provide guidelines for researchers working in this area. In order to make this article useful to the research community in granular computing, instead of using philosophical/epistemological jargon, we will stay with basic terminology used in granular computing.

An important advantage of studying the philosophical foundation of granular computing is to identify what

is still missing. In particular, we point out that the future of granular computing is inevitably tied to important features of complex adaptive systems such as self-organization, learning, evolution and adaptation, as well as key issues in contemporary artificial intelligence (AI) such as embodiment, emergence and emotion.

The objective of this article is to provide a fair review of certain important aspects related to philosophical foundation of granular computing. We summarize and analyze existing literature to find out what has led to GrC, which ideas have been explicitly stated, which are implied, and which are still lacking (or overlooked). Due to the huge amount of literature in this area and diverse viewpoints of researchers, comprehensive coverage of all perspectives and approaches is impossible. Only selected materials of influential work are included. Materials presented in this article (including the author's own opinions, some of which have never been published before) is not necessarily an indication that they have been accepted by the GrC community in general or approved by any specific individual.

Introduction

In order to answer the question of what is meant by "foundations of granular computing", we need to answer the question: What constitutes the foundations of granular computing? Just like in the examination of foundations of data mining (where granular computing can play an important role in discovery of hidden knowledge buried in huge amounts of data) [8], foundations of granular computing can be examined from philosophical, technical, and social/application dimensions.

In general, philosophy of science refers to the study of philosophical assumptions, foundations, and implications of science. Topics to be studied include the character and development of concepts and terms, propositions and hypotheses, arguments and conclusions as they function in science; the manner and types of reasoning, the formulation, scope and limits of scientific methods, the implication of scientific methods and models, etc. Philosophy of science is closely related to epistemology, as a field focusing on the theory of knowledge it studies the nature, methods, limitations, and validity of knowledge and belief [6].

In order to study the philosophy of science, either as a whole or as it applies to particular fields, we have to analyze the contents of theories as well as practices of scientists working in these fields, to identify the important philosophical thoughts are implied. Therefore, study of philosophy of science is not confined to "professional" philosophers; researchers in various science fields can (and should) take the philosophical aspects of their research discipline in their own hands. This is our attitude toward the study of philosophical foundation of granular computing.

All the issues related to the philosophical foundation of granular computing boil down to one question: *What is granular computing?* (Or: What is the *nature* of granular computing?) A simple question, with complex, possibly conflicting, yet incomplete answers! The philosophical foundation of granular computing requires a holistic examination on the foundations of granular computing as a whole.

Before we attempt to answer the magic question of what granular computing is, we have to address the issue of *how to* study it. Note that this is a multi-faceted question. First, the general methodology: Just like the study of philosophy of science in general, we have to take a retrospective/reflective approach to study what researchers have done in published literature related to granular computing: What is explicitly stated (i. e., what researchers in granular computing states what it is), what is implied (i. e., what researchers assume or think what granular computing might be), as well as what is yet to be addressed.

The next aspect of dealing with the "how to" question is to identify specific issues that need to be investigated. The answers to the question of "what granular computing is" can be studied from several aspects, in light of interests from the perspective of philosophy of science but stated in terms of granular computing:

- Character and development of concepts and terms, propositions and hypotheses:
  - Where is granular computing from? How does this new paradigm emerge (or converge from multiple predecessors)? In order to understand what granular computing is, we need a historical examination of the roots of granular computing.
  - Now that granular computing has established itself as a research discipline, we need to study its basic terms such as granule, granularity, granulation, etc. What does each of them mean? Without an understanding of these terms and what lies behind them, we will not be able to understand the nature of granular computing.
- Manner and types of reasoning, the formulation, scope and limits of scientific methods:
  - What are the major research issues of concern to the granular computing community? These issues form a set of descriptive features which collectively characterize what granular computing is.

  - What kinds of problems can granular computing solve, and where are concepts such as granularity and granulation are useful? The universality and diversity of these problems not only demonstrates the power of granular computing, but also reveal the nature of granular computing from yet another perspective.
- Implication of scientific methods and models:
  - In order to understand the nature of granular computing, it is equally important to examine how it distinguishes itself from other related disciplines. This leads us to examine its relationship with research fields such as soft computing and cognitive informatics.
  - Finally, it is time to examine a more fundamental issue: How is granular computing related to basic issues of computational modeling and complex problem solving? These more advanced issues help identify where granular computing stands in the larger picture and helps our understanding of granular computing.

Answers to the questions listed above form the bulk of the rest of this article.

Note that the study of the philosophical foundation of granular computing is inevitably intertwined with other dimensions of the foundations of granular computing. For example, one topic of study in the philosophy of science involves the implications of scientific methods and models, along with the technology that arises from scientific knowledge, for society at large [5]. This justifies the connection of the philosophical foundation with the other two dimensions of foundations. In the case of granular computing, the study of philosophical foundations requires a close examination of the models and algorithms developed (the task of the technical dimension) as well as the implication of various applications (the task of the social/application dimension).

The Road to Granular Computing

Granule, granularity and granulation are basic concepts involved in all aspects of human intelligence. First, it is in the language of our daily life, such as "extra fine granulated sugar". More profoundly, these concepts also form the foundation of science and technology, including biology, mathematics, etc. For example, mathematics, as the study of the measurement, properties, and relationships of quantities using numbers and symbols, has been dealing with computation problems on such granules for thousands of years.

For more recent history, we may consider computerized information systems, such as database management systems (DBMS) or information retrieval systems (IR). For example, in a relational database, tuples form the basic granularity level for information storage and retrieval, yet from the transaction processing perspective, it is the entire relation, rather than its constituent tuples, which serves as the basic granule for a commit. In fact, the theory and practice of transaction processing is largely built around concepts of granularity and granulation: In traditional DBMSs, transactions are the granules to commit and abort, and in order to assure this, ACID (atomicity, consistence, isolation and duration) properties are required. In more advanced DBMS systems such as object-oriented DBMSs, due to the significantly increased complexity of transactions, sub-transactions may be used as granules for commit, and accordingly, some of the rigid ACID properties may be relaxed somewhat. As a main issue of transaction processing, concurrency control is another showcase of granularity-driven activities: Although many concurrency control protocols use individual data items as the granules to perform synchronization, multiple granularity has also been used, where different locking modes are imposed on data at different granularity levels of hierarchy, such as the entire database, a partition of the database, a file, a record, and so on. As for the physical storage implementation, B-tree (and its variations) is a popular indexing structure for storing individual key values for records (here the individual data points serve as granules), while R-tree (and its variations) is B-tree's counterpart dealing with spatial data with two-dimensional minimum bounding boxes (mbbs) as granules for operation, the concept shared by many other proposed data structures for handling spatial data. Similarly, in information retrieval (IR), granules include document, chapters, sections, paragraphs, sentences, phrases, words, etc. (For basics of DBMS and a very brief overview on IR, see [36]).

Even today, an examination of relational databases from an information theoretic perspective [24] has been considered as an important piece of work related to the pre-history of granular computing. Yet, what is really interesting in today's granular computing community is what is now usually referred to as the *information granule* [38], a concept emphasizing its role in *intelligent* information problem solving. Another important aspect distinguishing granular computing from other areas in information technology is that it goes beyond concepts related to granularity and directly addresses *computational aspects* related to granulations.

As noted in [1,2], human-centered information processing has been pioneered by Zadeh through his intro-

duction of the concept of fuzzy sets in the mid-60s. GrC arose as a synthesis of insights into human-centered information processing by Zadeh in the later 1990s and the term was coined by T. Y. Lin in 1997. Below we take a closer look at contributions of several research fields which led to granular computing.

According to Zadeh [43], granulation refers to partitioning (crisp or fuzzy) of an object into a collection of granules, with a granule being a clump of elements drawn together by various criteria such as indistinguishability, equivalence, similarity, proximity or functionality. For example, in rough set theory partitioning is based on the criterion of equivalence, while in granular computing partitioning is based on similarity, proximity or functionality (note that a rough set is a formal approximation of a conventional set, using a pair of sets as the lower and the upper approximations of the original set).

A granular variable is a variable which takes granular values. Granulation of a variable (or granular variable), such as age, goes from continuous, to quantized, and finally, becomes granulated. Several principal types of granules can be distinguished. It has been noted that in fuzzy logic, granulation is achieved through graduation – "In fuzzy logic everything is or is allowed to be graduated, that is, be a matter of degree or, equivalently, fuzzy" [44]. In addition, granulation of a function is done by summarization, as shown in rules such as "If X is large then Y is small" [45]. On the other hand, developments in rough set theory provide another example of the universality of the basic concepts of granule and granulation.

As pointed out in historic remarks [30], while fuzzy set theory and Dempster–Shafer theory are considered as examples of non-partition theories (where the universe is not partitioned), rough set theory and algebraic theory of relational databases in the 1980s are examples of partition theory (where the universe is partitioned into subspaces). Rough set theory is about equivalence relations which is a subset of binary relations studied earlier. Rough set theory is about partition which is a subset of covering. Later, the concept of neighborhood system (NS) was proposed, which is the generalization of the rough set concept in that it incorporates various concepts of topological space, covering, binary relation, as well as α-cuts from fuzzy sets. Furthermore, each neighborhood can be regarded as a granule, and because of this, early granular computing research has the neighborhood system as its mathematical model. Therefore, the key in a road map from rough set to granular computing lies in the mathematic concepts of coverings, binary relations, and neighborhood systems, where a covering is a collection of granules (subsets) whose union is the whole space. Neighborhood systems have been used for the study of relational databases, and the granulation structures induced by neighborhood systems have been studied. A binary relation is equivalent to a special type of neighborhood system.

The Nature of Granule and Granule Computing

The Nature of Granule

The concept of granule has been studied in the AI community for decades [20]. According to [44,45], a granule is a clump of elements drawn together by various criteria such as indistinguishability, equivalence, similarity, proximity or functionality. In addition, attributes of a granule include probability measure, possibility measure, verity measure, length, volume, etc. There are numerous assumptions implied by various researchers regarding basic notions in granular computing, and the concept of granule in particular. In order to have a better understanding about the nature of granules, below we examine some of the hidden assumptions and provide a brief comparison.

- *Granule vs. words*: Traditionally, artificial intelligence (AI) has been considered as a research field of reasoning with individual symbols (a symbol is a token of meaning). These days, many researchers in computational intelligence have put much emphasis on computing with words (CW). As Zadeh noted, granular computing serves as a basis for the methodology of computing with words (CW) [44]. Although efforts have been made in this aspect, few researchers in granular computing have focused on issues related to linguistics. In current study of granular computing, semantics is demonstrated not through individual granules, but rather on the grouping of granules. Reference [1] proposed an information processing "pyramid" consisting of three layers organized by the size of information granules, with the lowest one concerned with numeric processing, the intermediate one concerning large information granules, and highest one devoted to symbol-based processing. But this is just one perspective.
- *Granule vs. attribute or entity (in database modeling)*: In the entity-relationship approach, although we talk about attributes of information granules, such as size, capacity and dimension, in general, granules are not distinguished from each other mainly based on descriptive features (or attributes). Granules have low individuality: A granule is a primitive concept while entity is a first order citizen in data modeling. Although entities of the same type can form an entity set, entity sets are not mainly used to form hierarchical levels (rather,

they are the "end" products of conceptual modeling), but granulation is.

- *Positive granule vs. object in object-oriented (OO) paradigms*: In OO, objects are related through superclass/class/subclass relationships and properties can be inherited or overridden. The OO paradigm has an anthropomorphic nature. On the other hand, granule is a general concept directly tied to computational aspects; although granules can also be related together through hierarchies, the relationship between various levels in the hierarchy is much more complex (see later discussion). In addition, although hierarchies play a key role in granular computing, other forms of relationships exist between granules (such as the volume of a cylinder and the surface of that cylinder).
- *Positive granule vs. cluster (in data analysis/data mining)*: Similarity in cluster analysis is usually determined by topology which is largely absent in granular computing; i. e., distance is usually not used as the basic factor to relate individual granules.
- *Granule vs. system vs. graph*: Granularity is a measure of the size of the components, or descriptions of components, that make up a system. A system, either natural or man-made, is a set of interacting or interdependent components forming an integrated whole. A system can also consist of a number of subsystems. Ideas from systems theory have grown with diversified areas, exemplified by the work of B. Banathy, as well as others. Systems described in terms of large components (i. e., larger granularity) are referred to as coarse-grained, otherwise they are termed as fine-grained. Since graph data structure is a useful tool to describe systems, it can also be a useful modeling tool for the study of granules (see ► Granular Computing: Practices, Theories, and Future Directions for the modeling examples). However, since system theory is devoted to the study of inter-relationship of its components to optimize system performance, it does not address the majority of issues of concern in granular computing.

The Nature of Granular Computing

According to [44,45], the rationale of granulation can be either imperative (forced) or intentional (deliberate), when precision is costly and there is a tolerance for imprecision. We can distinguish different forms of granulation, such as spatial (including interval), temporal granulation, etc. Granular Computing is an emerging conceptual and computing paradigm of information processing. It has been motivated by the urgent need for intelligent processing of empirical data, that is now commonly available in vast quantities, into humanly manageable abstract knowledge. In this sense, granular computing offers a landmark change from the current machine-centric approach to information and knowledge to a human-centric approach [1,2,3]. According to [31] by T. Y. Lin who first coined the term granular computing, any computing theory/technology that involves elements and granules (subsets or generalized subsets) may be called granular computing. Intuitively, elements are the data, and granules are the basic knowledge. This simple definition has the advantage of unifying research activities of researchers from diverse backgrounds. In addition, this simple definition allows us to exploit the interesting idea of *granulate and conquer*, a softer version of classical divide and conquer; fuzzy granulate and conquer is the cornerstone of the success of fuzzy controls. A very common technique used in the classical "non-partitioning" recursive call is dynamic programming [17].

According to [1], the fundamental features of GrC consist of the following: allow for multiple abstraction levels (granularity levels), allow for several methods of traversing various levels of hierarchy (referred to as encoding-decoding mechanisms), and allow for non-homogeneous methods. In addition, Zadeh [45] noted that granulation is a core concept surrounded by natural language computation, rough set theory, computational theory of perceptions, as well as granular computing. Therefore, there is a simple formula of "granular computing = ballpark computing".

As an example of recent development in regard to basic concepts in GrC, the study in [15] considered two granular worlds which are defined on the same universe of discourse but employ different granulations of this universe. In order to translate information from one granular world to the other, there is a need to regranulate the information so that it matches the information granularity of the target world.

Granule Measurement

As indicated earlier, according to [44,45], a granule is a clump of elements drawn together by various criteria such as indistinguishability, equivalence, similarity, proximity or functionality. In addition, attributes of a granule include probability measure, possibility measure, verity measure, length, volume, etc. Therefore, granules can be considered as knowledge, compared with constituting elements which represent data [27]. In addition, granules may be interpreted from three semantic views: a granule is a unit of basic knowledge; a granule is a unit lacking precise knowledge; and a granule is a sub-problem of com-

puting (see ► Granular Computing: Practices, Theories, and Future Directions). We will get back to the last perspective in a later section. Here we examine the first two perspectives: Granules are knowledge units, yet consist of uncertainty. Together, these two perspectives imply an intrinsic property associated with granules: Measurement, which refers to the estimation of the magnitude of some attribute of an object, such as its length or weight, relative to a unit of measurement. Consequently, granularity is a measure of the size of the components. In the simplest form, measures can be made through the size of a granule, which indicates the degree of abstraction, concreteness, or detail, of the granule. In the set-theoretic setting, the size of a granule can be the cardinality of the granule. Various uncertainty theories related to granular computing, such as fuzzy set theory or rough set theory, all have specific ways to deal with measurement.

The need for measurement in studying granules implies the importance of dealing with topology, a key concept underlying many subfields in mathematics, such as mathematical analysis. Take a look, for example, at the well-known *Heine–Borel theorem*: Each open covering of a closed and bounded set of real numbers has a *finite* subcovering. The granules involved here include the closed and bounded set, the elements in the original open covering, the elements in the resulting subcovering, etc. Therefore, from a granular computing perspective, we may say that the Heine–Borel theorem is concerned with *reduction* from infinite to finite, or *mapping* from an infinite number of granules to a finite number of granules.

There are various kinds of granules, and they can be related through hierarchy or other relationships or through aspects such as measurement. For example, a granule could be 3-dimensional (3D) object, or could be the surface of this object. Yet the calculation of the volume of a 3D object can be converted to the calculation of the area of the surface through various theorems developed in mathematical analysis. Here the computation problem (i. e., the volume) on one granule (i. e., the 3D object) is *converted* to another computation problem (i. e., the area) on another granule (i. e., the surface of the 3D object). This is exemplified in *Green's theorem*, a special case of the more general Stokes' theorem.

As a more recent example concerning the importance of measures for granules, let us consider the case of relational database design theory as well as its contemporary extension to XML file design. Relations (or flat tables) are the most visible granules in relational database design. But what should comprise a relation? This issue has been addressed by normalization theory, where important normal forms such as Third Normal Form (3NF) or Boyce–Codd Normal Form has been developed, based on the notion of functional dependency and related concepts. But how could this kind of result be generalized to the case of semistructured XML files? A recent study by Libkin and his research group [25] addressed this issue by proposing to use entropy-based measures to determine the quality of the granulation for relational database design, and to further apply this principle to XML file design. First the concept of relative information content (RIC) was proposed, subsequently other related measures were proposed, such as the Price. Furthermore, Price (3NF) = 1/2. Therefore, the conformation of a "good" granule (such as a relational table) is justified by a quantified approach. Although these studies are conducted by researchers outside of GrC community, these results can shed new insight for research in GrC.

Granular Structure

Quotient View – Information Hiding

As noted in ► Granular Computing: Practices, Theories, and Future Directions, granular structure is the collection of granules in which the internal structure of each granule is visible. In this sense, granular structure is a collection of white box granules. In contrast, a quotient structure, as derived from the mathematical concept of quotient space, refers to the collection of granules in which each granule is regarded as an element (or point) of a set, and the "intersections" among granules are abstract to interactions among points (or elements). Therefore, a quotient structure is a collection of black box granules. This perspective provides a quotient view on a hierarchical structure consisting of granules at different levels of detail. With details encapsulated at lower levels while only abstract information is shown on higher levels, quotient view-based granular structure supports information hiding.

Earlier, in the Introduction section, we mentioned that DBMS is a major source behind GrC. In fact, database design is a perfect example of information hiding: details of the instances are hidden at the more abstract levels of a schema. The success of the well-known entity-relationship approach lies on the information hiding by focusing on schemas (while not completely throwing away the instances, such as the notion of mapping cardinality involved in relationship sets). More recent developments in management of semistructured data necessitates the need for XML schema discovery (see, e. g., [12,19]), which in essence is to identify the quotient structure and support information hiding.

The theory of hierarchy provides a multi-layered framework based on levels. Mathematically, a hierarchy

may be viewed as a partially ordered set. Two types of partial ordering can be distinguished, i. e., strong vs. weak dependencies: A neighborhood system X is weakly dependent on Y if every neighborhood of X is a subset of Y; or strongly dependent if every neighborhood of Y is a union of the neighborhood of X [26,27,33]. Note that strong = weak in the case of partitions. (More discussion on hierarchy will be given later in the section on emergence.)

A granule in a lower level schema may be a more detailed description of a granule in a higher level with added information. In the other direction, a granule in a higher level is a coarse-grained description of a granule in a lower level omitting irrelevant details. Note that although we are talking about granules of various granularities connected together in a hierarchical manner, granules are related to each other in a complex network. In this sense, what we call hierarchy in GrC is actually a kind of "information hiding" with detailed network information filtered out.

A taxonomy of types of granularity has been discussed [23]. First, one can identify the main differences in types of granularity based on:

(1) Whether scale matters: Arbitrary scale vs. non-scale-dependent granularity;
(2) How levels (and their contents) in a perspective relate to each other;
(3) How perception and (mathematical) representation are related, such as based on set theory or mereology.

The philosophical implication of this taxonomy has also been examined; for example, it is noted that the difference between scale and non-scale dependency and their formal representations roughly fits with Sowa's epistemic and intentional granularity [39] based on Peirce's three categories.

Granular Computing and Ontologies

The concept of hierarchy plays a key role in many other contexts studied in last decades, including frame systems in artificial intelligence and object-oriented approaches, just name a few. Here we take a look at the recent surge of research related to ontologies, an issue originated in philosophy [6] that has received ever-increasing attention in recent years from a much wider community.

Even though hierarchies play a significant role in both granular computing and ontologies, the relationship between these two research fields is not well-studied, and this is likely due to the fact that although granular computing sets place emphasis on both computation and semantics, the study of ontologies is largely "semantics without computation". Nevertheless, ontological aspects have been occasionally studied by granular computer researchers, including the work on taxonomy of types of granularity summarized in an earlier section [23]. In addition, granular computing can be used to aid the study of ontologies. One example is depicted in [35] where the concepts of domain granule and domain granule lattice were introduced, and an algorithm for generating the lattice was proposed to capture the ontology. The relationship between granular computing and ontology can take the opposite direction as well; for example, [48] described a granular computing model based on ontology.

In addition, [22] presented a rough-granular approach which incorporates domain knowledge. This additional knowledge, represented by ontology of concepts, is used to make it feasible to search for features (condition attributes) relevant for the approximation of concepts on different levels of the concept hierarchy defined by a given ontology. Yet, the question of how ontologies of concepts can be discovered from sensory data remains as one of the greatest challenges for many interdisciplinary projects on learning of concepts.

Information Hiding and Emergence

A basic task of granular computing, granularity conversion, is to change views with respect to different levels of granularity (with changing details). However, it would be naive to believe that relationships between different levels of granulations can be described by mappings.

Information hiding as demonstrated in a granulation hierarchy may shed important light on revealing the nature of emergence [21], where complex phenomena are produced by interaction of simple, tiny things. Think about the operations of zoom in and zoom out: Just like using zoom lenses on a camera, we can show a smaller area of an image at a higher resolution or magnification or a larger area at a lower resolution or magnification. Note that there are apparently two kinds of "zoom in": For example, when we are examining an online interactive map, we may click a particular country on the initial world map, then click a particular region or state, then click a particular city, and so on, to take a close look at particular streets or buildings in which we are interested. On the other hand, sometimes the so-called "zoom in" only gives us larger prints of he original map without any additional information added: Although this latter type of "zoom in" is useful to senior citizens with vision problems, it is useless for most people. Similarly, we may have two types of zoom out: We can get the entire picture as a whole rather than focusing on a particular part of the picture (with all the details kept),

or we summarize or abstract information not available at the more detailed level.

Apparently, zoom in and zoom out are directly concerned with getting information at different granulation levels. When we "zoom out" from the lower level, we may not necessarily just lose the details at the lower level, but we may also gain insight not demonstrated at the lower level. As a typical example, think about the painting "Gala Contemplating the Mediterranean Sea which at Twenty Meters becomes a Portrait of Abraham Lincoln" by Spanish surrealist Salvador Dali. As the title suggests, at the detailed level is Gala's picture and the surroundings, while at a coarse granulation level, Lincoln's portrait emerges. This is where the excitement of emergence comes from, and granular computing can play a critical role in finding hidden rules which produce the emergent phenomena.

Due to the complexity of the networks in which granules are interrelated to each other, it would be questionable to assume that an algorithmic study of procedures for constructing granules and related computation is a universal and realistic approach in GrC. Just like the more general study of complex adaptive systems, the tasks faced in the GrC community are closely related to the task of understanding the nature of emergence. As the movement from low-level simple rules among granules to higher level sophistication, emergence is a ubiquitous feature of the world around us. The hallmark of emergence is the sense of much coming from little. Complex systems present behaviors by drawing upon populations of relatively "unintelligent" individual agents, rather than a single, "intelligent" agent. Yet, as John Holland noted, regardless of how mysterious it might appear, emergence can be studied as a scientific discipline because low-level rules and laws can be discovered, and the movement from low-level rules to higher level sophistication is what we call emergence. This view suggests a hierarchy of levels from which emergence takes place, and the phenomena of emergence are typically bottom-up in nature. Holland underlines a number of features of emergent (or complex) systems and suggests that emergence is a product of coupled, context-dependent interactions. These interactions, along with the resulting system, are nonlinear, where the overall behavior of the system cannot be obtained by summing up the behavior of its constituent parts. Persistent patterns at one level of observation can become building blocks for persistent patterns at still more complex levels. At each level of observation the persistent combinations of the previous level constrain what emerges at the next level. As Holland put it, "This kind of interlocking hierarchy is one of the central features of the scientific endeavor. It will lead us into, and out of, the thorny thicket known as reduction – roughly, the idea that we can reduce explanations to the interactions of simple parts". Yet, there are no systematic ways described there on how to discover much needed laws or rules so that emergence can take place.

Think about one of the most mysterious things we can ever have, namely, life. Regardless whether we are talking about how the first life appeared on the Earth millions or billions of years ago, or how a particular biological species (such as rhino or elephant) first evolved from other species, or how a particular baby named John J. Johnson was conceived by his parents in the last year, a life always emerges from a hierarchy of extremely complex interactions, starting from the lowest level (where the granules are the chemical elements) all the way up along the hierarchy.

An important lesson learned from making a life is: Emergence is a bottom-up process, and top-down is not symmetrical to bottom-up. A top-down analysis of a new born baby (using a partitioning approach) is the task of anatomy, but this will not reveal how he or she is made. In other words, although top-down lets us examine the product, bottom-up reveals the process. This justifies the need for both top-down and bottom-up directions of research in GrC.

Just like the study of emergence, the study of granular computing is intended to provide a unified approach of studying complex phenomena across various domains. In addition, just like the case of emergence, granular computing models are made concrete in specific application domains. Although emergence is a complex issue and has not caught much attention by researchers in the GrC community, certain aspects have been addressed under the notion of information integration ► Granular Computing, Introduction to. More systematic research along this direction is needed. Below we discuss several notable points.

- *Interaction among granules:* The interactions of granules were the beginning of GrC. The issues in knowledge representations, additive measures, belief functions (non-additive measures) and approximations were discussed in [26,27]. In addition, although the Choquet integral [37] usually is not considered within GrC realm, it is also based on belief functions. A Choquet integral is a way of measuring the expected utility of an event where the event is uncertain, and is applied specifically to capacities. The Choquet integral focuses on interactions among attributes, resulting in non-additive measures. Therefore, the contribution of two attributes to a classification variable is more than the sum of the contributions made by these two attributes alone.
- *Relationships of levels within a hierarchy:* In granular computing, granules in a higher level may have greater

integrity and higher bond strength than those in a lower level. The structures need to be fully explored to establish a basis of granular computing. However, in the study of emergence, levels in the hierarchy interact in a much more complex way; for example, a completely new form may take place, dramatically different from what can be found at a lower level (such as life coming from non-living chemical materials). In this respect, the relationship should be examined between granular computing and mereology, which is a collection of axiomatic formal systems dealing with parts and their respective wholes and the parts of the parts in a whole.

- *Other important features:* Other important features demonstrated in complex systems, such as self-organization, evolution, adaptation, etc., are not well-supported by current studies in granular computing. However, this situation seems to be changing. For example, [16] described a self-learning algorithm for uncertain information processing.
- *Building blocks:* Both granular computing and the study of emergence have emphasized the importance of building blocks for model construction but for different purposes.

Granulation Provides a Unified View for Intelligent Problem Solving

Granular computing and related concepts play an important role in a wide range of intelligent problem solving tasks, including data mining, which is intended to discover implicit, interesting knowledge patterns from massive data [18]. We now take a look at this issue, which is addressed through three tiers, from the most basic form to the most advanced form, and accordingly this section is divided into three subsections. We first present how to *interpret* problem domains in terms of concepts in granular computing, and then discuss how to *cast* issues of intelligent problem solving in terms of granular computing. Neither of these two tiers involves any specific algorithms or methods developed in granular computing. Finally, we take on the issue of granular computing algorithms using granular computing for data mining as an example, which is further illustrated by an implemented case study. Our discussion through these tiers demonstrates that granular computing and its related concepts serve as a unified view for intelligent problem solving in various degrees.

Re-Examination of Existing Studies from a GrC Perspective

Although granular computing should not be simple restatement of existing results, re-examining or re-interpreting existing results from the standpoint of granularity can help reshape the problem definition and establish a new start of granular computing-oriented solutions – so long as our investigation does not stop at the re-interpretation.

Here we use the case of bioinformatics as an example (note this is not intended as a general discussion on the relationship between granular computing and bioinformatics). Various subjects studied in bioinformatics, such as sequence, block, pattern, motif, and protein, can all be viewed as granules at different levels. Below we describe some sample scenarios to illustrate the applicability of GrC in bioinformatics. The key idea behind our discussion is that the relationship between GrC is twofold: GrC can be applied to shed light on bioinformatics problem solving, and bioinformatics also poses interesting challenges for GrC. Materials presented here are mainly taken from resources cited in [11].

- *Granulation construction under constraints*: The complexity involved in structural bioinformatics problem solving can take advantage of techniques developed in granular computing, and demand new solutions yet to be developed in granular computing. As discussed before, one of the basic concepts in granular computing is granulation. An essential question related to granulation is how to construct higher-level granulation from lower-level granules. As an example, here we examine the fragment assembly problem, which arises when a DNA sequence is broken into small fragments. These fragments must then be assembled to reconstitute the original molecule. (In a sense, the fragments form a covering in the context of GrC). The importance of this problem lies in the fact that with current technology it is impossible to sequence directly contiguous stretches of more than a few hundred bases. On the other hand, there is a technology to cut random pieces of a long DNA molecule and to produce enough copies of the pieces to sequence. Therefore, a typical approach to sequencing long DNA molecules is to sample and then sequence fragments from them. Solving this complex problem consists of assembling a collection of fragments coming from a long, unknown DNA sequence into a correct order and orientation. This can be cast as a problem in GrC, which is concerned with construction of larger granulation from smaller granules *under certain constraints* (such as imposing or preserving a certain order and orientation).
- *Characterization of granulation*: Once granulations have been constructed, the next question is to uncover the basic features of these granulations (namely, find

an *interpretation* of the constructed granulations). As an example, consider the following. Reference [37] described a data-mining algorithm for discovering regions of locus control, i. e., these regions that are instrumental for activating genes. One type of such elements of locus control is the matrix attachment regions (MARs). The discovery of MARs is based on the observation that a group of patterns are bonded together by the virtue of their similar function. After such a grouping, a search for the patterns in a given group can be performed to identify three regions in the query DNA sequence. If a large subset of members of a functionally related group of patterns is found in a given region of the DNA sequence, one can justifiably classify it as an MAR. The detection problem is to identify observation within a given region. Therefore, discovery of MARs is actually done by grouping for granulation (here in the form of patterns) and then finding the statistical characteristics of those granulations – somewhat similar to solving a mathematical equation. A more general indication of this approach is the task of studying *characterizing the granulations*, namely, identifying the significant features of the constructed granulations. Although characterization rule mining has been studied by many researchers in the data mining community [16], the task described in the above example goes beyond the reach of many existing approaches in characterization of rule mining, because the granulations considered here may be constructed in a dynamic manner. Therefore, a study in characterization of granulation will enrich both GrC and basic theory for characterization rule mining.

Forming Intelligent Problem Solving Tasks in Terms of Concepts in Granular Computing

We now examine several issues of intelligent problem solving.

Intelligent Query Answering Intelligent query processing has very rich contents. GrC (integrated with appropriate AI and data mining techniques) can contribute to intelligent query answering involving aggregate data, mainly due to its power of grouping data in a dynamic and flexible manner. The concepts of constructing equivalent classes, partitioning and covering, as widely used in GrC literature, can be incorporated into the discovery of query-relevant rules from the database. In addition, GrC can be more directly absorbed into presentation techniques applied on original extensional answers to produce final answers with an intensional flavor. Here we take a look at the issue of answering queries at the right granulation levels as originally presented in [10].

Although discovery of query-relevant rules is an interesting direction of study, in many other scenarios we may want to generate aggregate answers from existing extensional answers. Comparing with the previous topic, this direction can be considered as a "low key" solution; nevertheless, it deserves equal attention due to its practicality. As the online availability of data is exploding, business firms have begun to use such data to help to *make better decisions* for their business. For example, an auto manufacturer may want to know which type of vehicles made the highest profit in the previous year, or which generation of people could be the ideal target population for auto purchase and what their "tastes" may be. For such inquiries, conventional answers, usually extensional answers consisting of all the involved individual tuples, are not the most appropriate answers. Frequently, people feel the need for better decision support based on data analysis and knowledge discovery from the primitive data. As a more concrete example, suppose the Board of Directors of a retail firm wants to know something about the salary levels of its employees. A question could be: "*In our company, what kind of people are making more than $ 100 000 a year?*" We may expect to receive several types of answer to this query. An answer could be "All directors and most managers". It gives some qualitative description about the common characteristics of those qualified individuals. An alternative way is using an aggregate expression. An *aggregate expression* is a sequence of terms that are in the format of "$r/t\ C$", where C represents a concept with a total number of t individuals, while r is the number of these individuals who belong to the answer. As to the above query, an aggregate answer could be "8/8 directors + 27/32 managers". From this example it can be seen that an aggregate answer is superior to the qualitative answer "all directors and most managers", in that the aggregate answer not only covers all the information released in the latter, but also provides more quantitative information towards the overall picture of the database. In general, there exists more than one answer to a query. The question is how to obtain the best (optimal) one from those candidate answers.

An approach of using *query entropy* (or Q-entropy for short) for intelligent query answering concerning aggregate answers (with a prototype experiment developed) has been developed for answering aggregate queries at the appropriate data granulation levels [10]. The particular aspect we are looking for is that, given a conceptual hierarchy C and a query Q against a database D, when multiple answers exist at different granulation levels, how do we evaluate them and select the best answer? This ap-

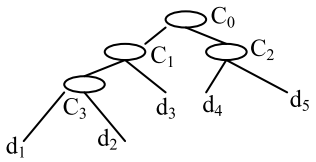

Granular Computing, Philosophical Foundation for, Figure 1
A bi-level taxonomy To of concepts

proach can be examined from a granular computing perspective.

An expression is a collection of terms. A term is represented by the ratio of qualified individuals and total individuals in that concept. The length of an expression measures conciseness, while the entropy of the expression measures preciseness. An expression is said to be the optimal answer to a given query if it has the lowest entropy value. For a real world problem, to obtain the optimal answer to a query, a hierarchy has to be built first from the primitive information stored in a given database.

Shannon entropy and associated conditional entropy are widely used today in information theory. Although the intuition to apply the entropy concept in information theory to intelligent database query answering is well-known, practical application of entropy theory is hindered by the inherent computing burden of Shannon and conditional entropy. One has to be very clear with taxonomy in order to obtain the probability of the event that an individual in a concept belongs to the extensional answer and the probability of the event that a randomly chosen individual belongs to a concept.

To take advantage of the approximation of Shannon entropy and the simplification of conditional entropy, the concept of Q-entropy was introduced. Let R be a finite space consisting of mutually disjointed subsets $R_1, R_2, \ldots, R_k$ with cardinalities $|C_1|, |C_2|, \ldots, |C_k|$ respectively. The probability of the event that an individual in set R_i is a qualified individual is denoted as r_i. Then the Q-entropy is defined as:

$$Q = \sum_{i=1}^{k} |C_i| * G_i = \sum_{i=1}^{k} -|C_i| * \left(r_i - \frac{1}{2}\right)^2 .$$

In addition to concepts C_i, we also use d_i to denote extensional answers (which are actual tuples). As an example, consider a taxonomy T of four concepts with $C_0 = \{d_1, d_2, d_3, d_4, d_5\}$ as the root concept, and let $C_1 = \{C_3, d_3\}$, $C_2 = \{d_4, d_5\}$ and $C_3 = \{d_1, d_2\}$, as shown in Fig. 1. (As a more concrete example, C_0 would be employee, C_1 could be worker, C_2 could be manager, etc.).

There are three candidate expressions:

I) "$1/2\, C_3 + 2/2\, C_2$";
II) "$1/3\, C_1 + 2/2\, C_2$";
III) "$3/5\, C_0$".

Calculating Q-entropy using the formula introduced above, we have $Q_{\mathrm{I}} < Q_{\mathrm{II}} < Q_{\mathrm{III}}$. Therefore the best answer, (I), is selected.

The Q-entropy approach for query answering was developed prior to our study of GrC. Nevertheless, we can reexamine this approach from a granular computing (GrC) perspective.

We first note that partition can be viewed as a special kind of granulation. However, in classification rule mining, partition is conducted on training samples. In contrast, in Q-entropy approach partition is conducted on extensional answers for queries. We also note that in the Q-entropy approach for any two aggregate answers of the same query, one answer may not necessarily be a refinement or coarsening of the other. For example, suppose in Fig. 1 C_2 also has a child concept C_4. Then the answer "1/2 of C_3 and 1/2 of C_4" is not a refinement or coarsening of the answer "1/4 of C_2 and 1/4 of C_2".

The approach of using Q-entropy for selecting appropriate levels of query answering also differs from classification rule mining in several other aspects, including size of granules, the construction of the tree, the way in which entropy is used, and so on. Because of these features, a GrC perspective can be incorporated into presentation techniques applied on original extensional answers and can produce final answers in an intensional flavor.

Granular Computing and Data Mining/OLAP

Relationship Between These Two Fields The interesting relationship between granular computing and data mining can be further examined in many different aspects; here we take a look at just one example of granular operators for property preservation. Granulation allows different representations of the same problem in different levels of detail. It is naturally expected that the same problem must be consistently represented. Granulation and its related computing methods are meaningful only if they preserve certain desired properties. For example, [46] studied the "false-preserving" property, which states that if a coarse-grained space has no solution for a problem then the original fine-grained space has no solution. Such a property can be explored to improve the efficiency of problem solving by eliminating a more detailed study in a coarse-grained space. In the context of hierarchical planning, one may impose similar properties, such as the upward solution property, the downward solution property, monotonicity, etc. This property resembles a priori principle as discussed in the context of asso-

ciation rule mining [18], but significantly differs from the view of new properties appearing at higher levels of hierarchies which are not found in lower levels in the study of emergence [20].

Granular Computing as a Basis for Data Mining and OLAP Data mining from databases [18], a key step in the knowledge discovery from databases (KDD), is intended to discover hidden knowledge buried under the ocean of data. As indicated in [33], "In essence data mining may be viewed as a form of summarization of very large datasets, while granular computing may be viewed as operations on summaries of small datasets. The common rule of summarization in data mining and granular computing is the principle reason why granular computing is of high relevance to data mining". For example, in [29], Lin showed that generalized association rules can be expressed as union of basic granules. More recently, [32] further explored the issue on deductive data mining using granular Computing.

In a sense, data mining is finding hidden equivalence relations. For example, cluster analysis aims to find clusters, which are a sort of equivalence relation, so at the cluster level only clusters (rather than original data points) are seen – a typical case of information hiding. Similarly, characterization or discrimination analysis [18] aim to identify dominant features to characterize target class or to discriminate target class versus comparison class (with details of attribute values removed) – again a typical example of information hiding. An example of the impact of granulation size to data mining is multi-level association rule mining [18], which results in revised algorithms and reduced support/confidence level at coarser granulation levels.

Related to data mining is On-Line Analytical Process (OLAP) technology. Unlike data mining, however, OLAP does not resort to a reasoning process; rather, it is more closed tied to database and data warehouse technology for analyzing summary and historical data. Since information granules are formed from data granules, perspectives from GrC also contribute directly to OLAP. In fact, OLAP operations such as rollup and drill down are directly concerned with granulation sizes. In addition, OLAP techniques have been coupled with data mining, giving GrC more room to maneuver. For example, as discussed in [18], complex aggregations using multi-featured data cubes can facilitate data mining type queries to allow computation of aggregates at different granularity levels. In addition, a unified On-Line Analytical Mining (OLAM) has been proposed [18]. Additional discussion about this integrated data mining/OLAP and the role of GrC can be found in [7,9,47].

Relationship with Soft Computing and Natural Computing

Relationship with soft computing: Probably, the most ancient granule is the intuitive infinitesimals (which refers to the idea of objects so small that there is no way to see them or to measure them). It was the inspiring source of calculus and Zeno's stationary and moving paradox. A granule is an atom of uncertainty (see ► Granular Computing: Practices, Theories, and Future Directions and [31]). Yet research related to uncertainty in computational research has focused on the technical side rather than on the philosophical side; therefore, the bulk of research on uncertainty has been conducted in the realm of soft computing rather than granular computing. Soft computing has been defined as an association of computing methodologies which includes its unique, complementary and symbiotic constituent members fuzzy logic, neurocomputing, evolutionary computing and probabilistic computing. Soft computing differs from conventional (hard) computing in that, unlike hard computing, it is tolerant of imprecision, uncertainty, partial truth, and approximation. In particular, the primary contribution of fuzzy logic is the machinery of granular computing, which serves as a basis for the methodology of computing with words [43]. Suffice it to say that even granular computing and soft computing have significant overlap in contents and in research individuals, and are both driven by human-centered intelligent problem solving. They also differ in the perspectives used in problem solving: Unlike granular computing which places emphasis on granulation-based reasoning aspects of human information processing, soft computing sets emphasis on the "soft" side of human information processing which tolerates imprecision and uncertainty.

Relationship with natural computing: Since granular computing is closely related to soft computing and since soft computing is also closely related to natural computing, the relationship between granular computing and natural computing should not be overlooked. Natural computing is the term used to encompass all AI approaches based on some inspiration from nature, soft computing, man-made complex systems, biologically inspired computing, and novel computing paradigms rooted in nature in a higher or lesser degree. However, as noted by Zadeh [43], in effect, the role model for soft computing is the human mind. There is a significant overlap between natural computing and soft computing in content; yet they differ in *perspectives* and *emphases*: Natural computing is more concerned with how to *model nature* while soft computing focuses more on how to *achieve "softness"*. A brief examination of the contents of natural computing can re-

veal the relationship between granular computing and natural computing more directly: Recall that natural computing refers to three types of approaches [14]: *Computing inspired by nature* (such as generic algorithms) which makes use of nature as inspiration for the development of problem solving techniques; *simulation and emulation of natural phenomena in computers*, which is basically a synthetic process aimed at creating patterns, forms, behaviors, and organisms that (do not necessarily) resemble "life-as-we-know-it"; and *computing with natural material*, such as quantum computing. From a GrC perspective, a quantum can be viewed as a granule, and granule is organized in such a way that it could grow exponentially, so granulation may provide a new framework of attacking classical theory of computing [17]. In addition, genetic algorithms can be interpreted as producing new granules from existing granules through specific operations (such as crossover or mutation).

Taking a step further, we note that natural computing is an important field to demonstrate the pervasive concepts of emergence [20], embodiment [34], as well as other key concepts such as interactivity, adaptation, learning, evolution, and self organization. Granular computing can play an important role in the study of natural computing as in the case of learning theory. As noted in [17], learning is interpolation of data based on background knowledge (granules). Neural network, support vector machine, and Smale's mathematical foundation of learning [13] can all be so formulated.

Relationship with Fundamental Issues of Computing and Complex Systems Problem Solving

We now examine a number of more advanced perspectives relating granular computing to basics of computing and complex systems of problem solving.

Granular Computing as Infrastructure for AI-Engineering

As endorsed by Zadeh [42] at the very beginning, granular computing should serve as foundation of human problem solving. As such, GrC serves as the infrastructure for AI-engineering: uncertainty management, data mining, knowledge engineering, and learning [30]. Structures, representations, and applications of granular computing have been discussed in this context [28]. Note that unlike the human problem solving perspective, this loose definition is computation-driven – even the key notions of granular computing were originally motivated by human-centered information processing.

It has been noted that a simpler and more effectively computable view of knowledge is needed. Rough set theory takes a courageous step and assumes that partitions (classifications) are the essence of human knowledge. Granular computing takes a softer view: generalized subsets are basic knowledge [17].

This perspective seems to serve as the de facto standard of "what granular computing is" for the majority of researchers who are active in this area. For example, [40] apparently echoes this perspective, where several learning paradigms for granular computing are discussed, although from a fuzzy set perspective.

Note also according to this definition, there is no urgency calling for a new computer model or AI architecture; we will stay with the von Neumann architecture, as traditional AI has done. Nevertheless, proponents of this perspective do relate granular computing to new computational models, such as quantum computing (as already mentioned earlier).

Granular Computing and New Computational Models

References [1,2] proposed to interpret granulation in the context of axiomatic set theory. They define information granulation as a *semantically* meaningful grouping of elements based on their indistinguishability, similarity, proximity or functionality. The semantics of granules are derived from the domain that has, in general, higher cardinality than the cardinality of the granulated sets. The next question is then how the meaning (semantics) are instilled into real-life information granules. It has been argued that Turing's computer on its own is unable to respond to external, physical stimuli. To overcome the problem, we may have to advocate the idea of embodiment in AI [34], and embrace Bain's approach [4] of using inherent computational ability of physical phenomena in conjunction with the numerical information processing ability of a Universal Turing Machine (UTM). It has been argued that in a broad context a UTM implementing the virtual computational intelligence function can be referred to as computing with perceptions or computing with words.

Summary

As a new research discipline, granular computing (GrC) is a subject of study for further abstraction, generalization and unification of granule-based key concepts. However, as the discussion of this articles shows, due to the complexity of this subject researchers differ in the scope, methodology, and perspective on the study of GrC. Although there has been a widespread enthusiasm for GrC, a unified theory of GrC is still missing. Nevertheless, there are a num-

ber of basic elements serving as the backbone of GrC such as granular structure, quotient structure and information hiding. We have also addressed certain issues not being emphasized by some researchers in GrC, such as granule measurement, as well as the relationship between hierarchy and emergence.

In the main text, while recognizing differences among different researchers, we have tried to present the philosophical foundation of granular computing as a coherent whole. As we are closing this article, we want to point out a few major differences in regard to philosophical foundations observed from various researchers.

As it currently stands, granular computing seems to be a loosely structured, somewhat interdisciplinary field: A big umbrella which accommodates researchers in rough set theory, fuzzy set theory, as well as soft computing and data mining – *so long as we all respect granulation as the core of human-centered information processing*. As for the theoretical core, much of today's study in granular computing is still largely an extension of rough set theory. Should this situation continue? Some may argue that this multidisciplinary nature is not all bad – after all, researchers with different backgrounds and different interests can now gather together, exchange their thoughts and foster new ideas – for example, the increasing publication of hybridization of fuzzy/rough set approaches seems to be a solid proof. Ironically, such successes may also have hampered the development of a unified theory of granular computing because there seems to be no urgency of bothering with it at all.

For those who have a long-term vision of granular computing, viewpoints vary significantly. Although seemingly everybody agrees granular computing is about human-centered information processing, what is its ultimate objective? As shown in the previous sections, answers could be:

- Assisting human problem solving in new ways of thinking,
- Providing infrastructure to AI-engineering, and
- Serving as the starting point toward new computational models.

Our examination has also addressed important issues related to embodiment and emergence. In addition, to cite Zadeh: "In coming years, granular computing is likely to play an increasingly important role in scientific theories-especially in human-centric theories in which human judgment, perception and *emotions* are of pivotal importance" (from [3], our italics). Therefore, as with the study of human-centered information processing, granular computing should also incorporate the role of emotions into its research agenda.

Future Directions

Born in 1997, granular computing is now a happy teenager, enjoying a wide range of research fruits and publicity. But will its multiple personality survive adolescence to its adulthood? The answer is stated (or implied) in other articles on granular computing in this encyclopedia, and is thus not in the scope of this current article.

As for the research on philosophical foundations of granular computing, what we want to emphasize is that the future of granular computing lies in the development of algorithms. Yet a continued examination of philosophical foundations will benefit this direction of research. In addition, as noted in [41], in the past 10 years, granular computing publications have experienced a linear growth rate; however, the granular computing community has less interaction with other research communities other than fuzzy sets and rough sets. Thr granular computing community should be aware of this and take steps to promote interdisciplinary research activities.

The research agenda of GrC includes a series of methodological and algorithmic issues [1], including construction of information granules, characterization of dimension (granularity) of information granules (to provide better insight as to the essence of the granulation process and its implications), development of the encoding and decoding mechanisms between levels of hierarchy; research on interoperability which is crucial to the design of systems operating within the realm of various formalisms of information granularity, as well as others.

There is a need for focusing on key ideas behind granular computing that have been overlooked (or at least not emphasized enough), such as computing with words. In addition, as indicated earlier, important features needed to support complex systems such as self-organization, adaptation, evolution, etc. are currently not well-supported by granular computing research, and this situation should be changed.

Acknowledgments

The author thanks Dr. T. Y. Lin's useful comments for the improvement of the paper.

Bibliography

Primary Literature

1. Bargiela A, Pedrycz W (2002) Granular computing as an emerging paradigm of information processing. In: Bargiela A, Perycz W (eds) Granular computing: An introduction. Kluwer, Boston

2. Bargiela A, Pedryez W (2006) The roots of granular computing. Proc IEEE Conf Granul Comput, pp 806–809
3. Bargiela A, Pedryez W (2008) Toward a theory of granular computing for human-centered information processing. IEEE Trans Fuzzy Syst 16(2):320–330
4. Bains S (2004) Physical computation and the design of anticipatory systems. Proc IEEE SMC 2004
5. Boyd R, Gasper P, Trout JD (1991) The philosophy of science. Blackwell Publishers, Cambridge
6. Burkhardt H, Dufour CA (1991) Part/Whole I: History. In: Burkhardt H, Smith B (eds) Handbook of metaphysics and ontology. Philosophia, München
7. Chen Z (1999) An integrated architecture for OLAP and data mining. In: Bramer M (ed) Knowledge discovery and data mining: Theory and practice. IEE, London, Chap 6, pp 114–136
8. Chen Z (2002) The three dimensions of data mining foundation. In: Proc ICDM 2002 workshop on the foundation of data mining and discovery, Dec 2002, pp 119–124
9. Chen Z (2003) On Granular computing for aggregate data mining. In: 7th international conference on computer science and informatics, Proc JCIS, pp 431–434
10. Chen Z (2003) Discovering rules and answering queries at the right granulation levels. In: 7th international conference on computer science and informatics, Proc JCIS, pp 435–438
11. Chen Z (2003) Bioinformatics as a study of structured granules. In: International conference on computational intelligence and natural computing, Proc JCIS, pp 1629–1632
12. Chidlovskii B (2001) Schema extraction from XML data: A grammatical inference approach. In: KRDB'01 Workshop (knowledge representation and databases) Rome, Italy, 2001. http://citeseer.ist.psu.edu/chidlovskii01schema.html
13. Cucker F, Smale S (2002) On the mathematical foundations of learning. Bull Amer Math Soc 39:1–49
14. De Castro LN (2006) Fundamentals of natural computing: Basic concepts, algorithms, and applications. Chapman and Hall/CRC, Boca Raton
15. Dick S, Schenker A, Pedrycz W, Kandel A (2007) Regranulation: A granular algorithm enabling communication between granular worlds. Inf Sci 177(2):408–435
16. Gan Q, Wang G, Hu J (2006) A self-learning model based on granular computing. In: Proc 2006 Int Conf Granul Comput (IEEE GrC 2006), pp 530–533
17. Granular computing information center. http://www.cs.sjsu.edu/~grc/grcinfo_center/grcinfo_index.php
18. Han J, Kamper M (2006) Data mining: Concepts and techniques, 2nd edn. Morgan Kaufmann, San Francisco
19. He J (2002) Schema discovery of the semi-structured and hierarchical data. In: Proc IDEAL (intelligent data engineering and automated learning) 2002. LNCS, vol 2412. Springer, Berlin, pp 129–134
20. Hobbs JR (1985) Granularity. In: Proceedings of the 9th international joint conference on artificial intelligence, pp 432–435
21. Holland JH (1999) Emergence: From chaos to order. Addison-Wesley, Reading
22. Jankowski A, Skowron A (2007) Toward rough-granular computing. In: Rough sets, fuzzy sets, data mining and granular computing (Proc RSFDGrC 2007). LNCS, vol 4482. Springer, Berlin, pp 1–12
23. Keet CM (2006) A taxonomy of types of granularity. In: Proc 2006 Int Conf Granul Comput (IEEE GrC 2006), pp 106–111
24. Lee TT (1987) An information-theoretic analysis of relational databases – Part I: Data dependencies and information metric. IEEE Tran Softw Eng 13(10):1049–1061
25. Libkin L (2007) Normalization theory for XML. In: Proc XSym 2007, pp 1–13
26. Lin TY (1998) Granular computing on binary relations I: Data mining and neighborhood systems. In: Skowron A, Polkowski L (eds) Rough sets in knowledge discovery. Physica, Heidelberg, pp 107–121
27. Lin TY (1998) Granular computing on binary relations II: Rough set representations and belief functions. In: Skowron A, Polkowski L (eds) Rough sets in knowledge discovery. Physica, Heidelberg, pp 121–140
28. Lin TY (2003) Granular computing: Structures, representations, and applications, rough sets, fuzzy sets, data mining, and granular computing. In: Proc RSFDGrC 2003. LNCS, vol 2639. Springer, Berlin, pp :16–24
29. Lin TY (2004) Mining associations by linear inequalities. ICDM 2004:154–161
30. Lin TY (2006) A roadmap from rough set theory to granular computing. LNAI, vol 4062, pp 33–41. Updated version: Granular computing on coverings: A roadmap from rough set theory to granular computing
31. Lin TY (2006) Granular computing II: Infrastructure for AI-engineering. In: Proc 2006 Int Conf Granul Comput (IEEE GrC 2006), pp 2–7
32. Lin TY (2003) Deductive data mining: Mathematical foundation of database mining. LNCS, vol 2369, pp 403–404
33. Lin TY, Zadeh LA (2004) Special issue on granular computing and data mining. Int J Intell Syst 19(7):565–566
34. Pfeifer R, Bongard J (2007) How the body shapes the way we think: A new view of intelligence. MIT Press, Cambridge
35. Qiu T, Chen X, Huang H, Liu Q (2006) Ontology capture based on granular computing. In: Proc 6th Int Conf Intell Syst Design Appl (ISDA'06), pp 770–774
36. Silberschatz A, Korth H, Sudhashan S (2006) Database system concepts, 5th edn. McGraw-Hill, Boston
37. Singh GB (2002) Discovering matrix attachment regions (MARs) in genomic databases. SIGKDD Explor 1(2):39–45
38. Skowron A, Stepaniuk J (2001) Information granules: Towards foundations of granular computing. Int J Intell Syst 16:57–85
39. Sowa JF (1999) Knowledge representation: Logical, philosophical, and computational foundations. Brooks/Cole Publishing, Pacific Grove
40. Yager RR (2006) Some learning paradigms for granular computing. In: Proc 2006 Int Conf Granul Comput (IEEE GrC 2006), pp 25–29
41. Yao JT (2007) A ten-year review of granular computing. GrC 2007:734–739
42. Zadeh LA (2006) Granular computing – The concept of generalized constraint-based computation. RSCTC 2006 (LNCS, vol 4259), 12–14
43. Zadeh LA (1998) Some reflections on soft computing, granular computing and their roles in the conception, design and untilization of information/intelligent systems. Soft Comput 2:23–25
44. Zadeh LA (2007) Granular computing – Computing with uncertain, imprecise and partially true data. In: Proc 5th Int Sym Spat Data Qual (ISSDQ 2007). http://www.itc.nl/ISSDQ2007/Documents/keynote_Zadeh.pdf
45. Zadeh LA (2007) Granular computing and rough set theory. In:

Kryszkiewicz M, Peters JF, Rybinski H, Skowron A (eds) Rough sets and intelligent systems paradigms. Proc RSEISP 07, pp 1–4
46. Zhang L, Zhang B (2003) The quotient space theory of problem solving. LNCS, vol 2639. Springer, Berlin, pp 11–15
47. Zhang X, Chen Z (2002) Algorithms for finding influential association rules. In: Proc 1st IEEE international conference on fuzzy systems and knowledge discovery, pp 499–503
48. Zhou G, Liang J (2006) Granular computing model based on ontology. IEEE Int Conf Granul Comput 321–324

Books and Reviews

Johnson S (2001) Emergence: The connected lives of ants, brains, cities, and software. Scribner, New York
Lin TY (1997) Granular computing. In: Announcement of the BISC special interest group on granular computing.
Minsky M (2006) The emotion machine: Commonsense thinking, artificial intelligence, and the future of human mind. Simon and Schuster. Online draft at Minsky's website http://web.media.mit.edu/~minsky/
Pawlak Z (1991) Rough sets: Theoretical aspects of reasoning about data. Kluwer Academic Publishers, Dordrecht
Pawlak Z (1998) Granularity of knowledge, indiscernibility and rough sets. In: Proceedings of 1998 IEEE International Conference on Fuzzy Systems, pp 106–110
Wang Z, Leung K-S, Klir GJ (2008) Integration on finite sets. Int J Intell Syst 21:1073–1092
Zadeh LA (1997) Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Set Syst 19:111–127

Granular Computing: Practices, Theories, and Future Directions

Tsau Young Lin
Department of Computer Science,
San Jose State University, San Jose, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Classical Examples of Granulation
Formal Models of Granulation
Granular and Related Structures
Integration – A Dual of Granulation
Semantic Views
Future Directions
Bibliography

Glossary

All terms are explained in classical sets, but implicitly, we are assuming all terms and assertions do include fuzzified versions (if fuzzifiable).

Granulation Granulation is an operation or a process of forming granules, with a granule being a collection of objects (points) that are drawn together by some constraints, such as indistinguishability, similarity or functionality.

Granular structure Granular structure is the collection of granules, in which the internal structure of each granule is visible as a sub-structure. Informally speaking, granular structure is a collection of white box granules.

Quotient structure A quotient structure is the mathematical structure of the collection of granules, in which each granule is regarded as an element (point) of a set, but the interactions among granules are preserved. Informally speaking, a quotient structure is a collection of black box granules.

The collection of $\{\ldots,-2,0,2,\ldots\}$ and $\{\ldots,-3,1,3,\ldots\}$ is a granular structure. Let E be the first subset (even integers) and O be the second subset (the odd integers). We write the two subsets by $[E]$ and $[O]$, when we think of them as points. Then the collection of $[E]$ and $[O]$ (as points) is the quotient structure.

Neighborhood system (local granular model abb. local GrC model) A domain of interests (a classical set) U is called the universe. To each point p in the universe, a family of subsets is assigned. Such a family (could be empty) and each such subset is called a neighborhood system $\mathrm{NS}(p)$ at p and a neighborhood at p, respectively. The collection β of such a family at every point of the universe is called a neighborhood system $\mathrm{NS}(U)$ of the universe. Neighborhood and neighborhood system are pre-GrC language; in granular computing, they are called granule and the granular structure, respectively. The pair (U,β) is called a local granular model, since each granule is associated with some points.

Topological neighborhood system A neighborhood system is called a topological neighborhood system, if it satisfies the axioms of topology.

Binary neighborhood system (binary granular model; binary GrC model) A binary neighborhood system is a neighborhood system defined by a binary relation R. A (right) neighborhood is defined as follows: $B(p) = \{x \mid (p,x) \in R\}$. The collection B of $B(p)$ at each p is the (right) binary neighborhood system. Similarly, we can define a left version: A left neighborhood system L is defined by the $L(p) = \{x \mid (x,p) \in R\}$ at every point p. Note that the right and left neighborhood systems determine each other. The pair (U,β) is called a binary granular model, where β is right (left) neighborhood system or R.

pre-Topology Pre-topology is a general term referring to

the neighborhood system, which includes the topological neighborhood system and binary neighborhood system as special cases. Technically, it is equivalent to the neighborhood system.

Bag A bag is similar to a set, but allows an element to appear more than once. For example $\{1, 2, 1, 2, 1\}$ is a bag, but not a set. If a bag contains n elements, we may say it is an n-bag. For example, the previous bag is a 5-bag.

Relational structure (relational granular model; relational GrC model) A family of classical sets is called a universe and denoted by $\mathcal{U}$. A Cartesian product of an n-bag of $\mathcal{U}$ is called an n-product set. A n-ary relation is a subset of an n-product set. A collection β of n-ary relations (n could vary) is called a relational structure. The pair (U, β) is called a Relational GrC Model. Note that this relational structure, except the size, is similar to the relational structure (without functions) of the First Order Logic; the Relational GrC Model permits n to be any cardinal number.

Partial covering (global granular model; global GrC model) Let U be a classical set, called the universe. Let $\beta = \{F^1, F^2, \ldots\}$ be a family of subsets. Such a β is a partial covering, and (full)covering, if the union of β is the whole universe. The pair (U, β) is called a Global Granular Model (Global GrC Model).

Equivalence Relation A binary relation $\mathcal{R}$ is called an equivalence relation, if it has the following properties: Let u, v, and w be elements of U.
reflexive: $u\mathcal{R}u$
symmetric: $u\mathcal{R}v$ implies $v\mathcal{R}u$
transitive: $u\mathcal{R}v$ and $u\mathcal{R}w$ implies $u\mathcal{R}w$

Partition A partition $\mathcal{P}$ of a classical set U is a collection of subsets that are mutually disjoint and their union is U. Each subset is called an equivalence class. This name is derived from the fact that partition is equivalent to the following equivalence relation: We define $u\mathcal{R}v$, if and only if u and v belong to the same equivalence class. Such $\mathcal{R}$ is the equivalence relation corresponding to the partition $\mathcal{P}$. Note that a partition is a special type of a granular structure, so an equivalence class is a special granule.

Definition of the Subject

Granular Computing (GrC) is still in its inception stage, we use motivation as its statements of importance.

How Important is Granular Computing?

Granulation seems to be a natural methodology deeply rooted in human thinking. Many daily "things" are routinely granulated into sub"things"; for example, the human body has been granulated into the head, neck, and so forth. The notion is intrinsically fuzzy, vague, and imprecise. Formalization is difficult, mathematicians idealized/simplified it into the notion of partitions (= equivalence relations), and have developed it into a fundamental part of mathematics, for example, congruence in Euclidean geometry, quotient structures (groups, rings, etc) in algebra, the concept of "a. e." (almost every where) in analysis. Nevertheless, the notion of partitions (see glossary), which absolutely does not permit any overlapping among its granules, seems to be too restrictive for real world problems. Even in natural science, classification *does* permit a small degree of overlapping; there are beings that are both appropriate subjects of zoology and botany. So a more general theory, namely, Granular Computing is needed.

What is Granular Computing?

It has been a changing paradigm. We believe in incremental developments. Its development may be similar, though not as glorious, to that of classical geometry. Specific geometries, such as, Euclidean, hyperbolic, elliptic geometries, had appeared before the the unified theory was proposed in Klein's Erlangen program. Except some details, we are on the final line; in GrC2008 keynote, we (this section editor) verbally had proposed to regard the category theory-based model (Eighth GrC Model) as the final GrC model; see Sect. "Formal Models of Granulation."

Granular Computing is a recent label coined by Lin and Zadeh (see Sect. "Introduction") to denote a set of common, even ancient, concepts and practices. The subject will be presented from various angles: What are the target concept (defined by examples), key constituents, and the current interpretations or semantic views? How far has it been formalized? What are the important applications?

1) The "Definition" of Target Concept: The key concept in Granular Computing is the concept of granulation, which has been "defined" implicitly by a set of intuitive examples (see Sect. "Classical Examples of Granulation" for more), including Zadeh's intuitive view. Here is a list of representative examples. In the following, we will call the domain of interests the universe and denote it by U.
   1. Many daily things have been routinely granulated into "sub"things; for example, the human body is granulated into head, neck, etc. The notion is intrinsically fuzzy, vague and imprecise. A formal model just had been proposed by this writer in the keynote

of GrC2008. Its final form should appear soon in the International Journal of Granular Computing, Rough Sets, and Intelligent Systems.

2. Space and time has been granulated intuitively into infinitesimal granules; a circle was viewed as a polygon with infinitesimal sides. The idea was known to Zeno (490 BC, implicitly in his paradox), Archimedes (287–212 BC) etc. It led to the invention of calculus, topology, and nonstandard analysis; we have modeled it in the First GrC Model.
3. The simplest kind of granulation is partition (see glossary). Its algebraic correspondence, equivalence relation, has played an important role in Euclidean Geometry. (300 BC)
4. The Heisenberg uncertainty principle states that, in general, neither the momentum nor the position of a particle can be determined simultaneously with arbitrary great precision. In other words, a precise measurement of the momentum can only determine a "neighborhood"(granule) of positions and vice versa. The idea is abstract into the Third GrC Model.
5. A collection of fuzzy sets (in fuzzy control) or functions (e. g. Radial-Basis-Functions) which has the universal approximation property is useful granular structure in a function space or a set of fuzzy sets. The idea is modeled in the Sixth GrC Model.
6. A committee in a human society is a granule in a social network. Observe that each member may play different roles. By viewing the collection of roles as a relational schema, a committee is a tuple, not necessary a subset. The idea is modeled in the Fifth GrC Model and the Second GrC Model.
7. A mathematical proof or computer program often contains some lemmas or subprograms. These lemmas or subprograms are granules. These are conceptual examples. They are modeled in the Seventh GrC Model for computable domain, and in the Ninth GrC Model for general cases.
8. Computers or clusters of computers in Grid/Cloud computing are granules. These are hardware examples. This also belongs to the Seventh GrC Model.
9. Zadeh's informal definition [38]: "information granulation involves partitioning a class of objects (points) into granules, with a granule being a clump of objects (points) which are drawn together by indistinguishability, similarity or functionality."

2) A Category Theory based Formal model (Eighth GrC Models) is proposed to be *the* Formal Model for GrC. It realizes all classical examples given above; see Sect. "Formal Models of Granulation." By specifying the abstract category to various ones, eight common models have been explained in this article. Two models are illustrated at the end of this section;

3) The Key Constituents: two operators, three semantic views, and four structures.

3.1) Two Operators: information granulation and integration. In granulating a problem, a dual action, namely, integrating the solutions of sub-problems is triggered. Here, we highlight some unusual points; see Sect. "Integration – A Dual of Granulation."

1. Recursively granulating a problem may take Np hard time to terminate. Dynamic programming technique has been used in the classical cases.
2. Two distinct problems may have the same granulation. Divide/granulate and conquer may have deeper meanings. The EXTENSION Functor of Homological Algebra of Commutative group is illustrated. This is an unexplored area.

3.2) Three Semantic Views: Granules may be interpreted from:

1. Uncertainty Theory: A granule is a unit of lacking precise knowledge.
2. Knowledge Engineering: A granule is a unit of Basic Knowledge (Information).
3. How-to-Solve/Compute-it: A granule is a subproblem or software unit. It is a special type of basic knowledge.

Each view may have its own GrC theory; for example, Concept Approximations are useful in the Second View, while Information Hiding is in Third View; see Section Semantic Views.

3.3) Four Structures: Granular, Quotient, Knowledge and Linguistic Structures:

1. Granular Structure (GrS): It is the collection of all granules. In the case of partition, GrS is the collection of the equivalence classes.
2. Quotient Structure (QS): If each granule is abstracted into a point and the intersections of granules are abstract to the interactions of points, then such a collection of points is called a quotient structure. In the case of partition, the quotient structure is a classical set, called quotient set. Informally, GrS is a collection of white boxes (the content of granules are visible), while the quotient set is a collection of black boxes (the contents of granules are hided). The process of abstracting a granular structure into a quotient structure is called information hiding; see Examples in Sect. "Granular and Related Structures."
3. Knowledge structure: By giving each granule (point) in the quotient structure a meaning-

ful symbol then the named quotient structure is called a knowledge structure. The knowledge structure provides an intuitive view of the quotient structure; the symbols and interaction among symbols are in sync with the granules (points) and interactions among granules (points).
In the case of n partitions (equivalence relations), the knowledge structure can be arranged into an n-column relational table. In the case of n binary relations, the table has been called binary information table, granular table or topological table [15,22].

4. Linguistic structure: By giving each granule in the granular structure a word that reflects its meaning. The interactions among these words are reflected implicitly in precisiated natural language (in a knowledge structure, the interactions among symbols are explicitly reflected from the quotient structure). The linguistic structure is the domain of computing with words

4) Applications to Computer Security, Web Technology, and Complex Data.

1. Discretionary Access Control Model (DAC) [8,17]. This structure has been captured in the Third GrC Model. On the basis of this, information flows on DAC can be analyzed; this has been considered a very "difficult" area. As a consequence the Aggressive Chinese Wall Security Policy can be enforced, namely, the system can guarantee that a company's data will never flow into "enemy" hands, where "enemy" is a granule of companies that are in conflict.
2. Documents can be clustered into a simplicial complex (a common structure in Combinatorial Topology [34]) of keywords and co-occurring keyword sets. The set of keywords can be regarded as a set of vertices, and the collection of co-occurring keywords (within a small neighborhood) is a set of simplexes. Together, they form a simplicial complex [21,24]. Simplexes are granules; a simplicial complex is a Second GrC model.
3. Granular computing has been used to solve the modeling problem of complex architectures [25]. It uses the Fifth GrC model.

Illustration of Working Formal Models

The most general model is expressed in the category theory. However, for easiness, we explain the Second GrC model first.

Granular Computing: Practices, Theories, and Future Directions, Table 1
Generalize 2nd to 5th GrC Models

2nd GrC	Generalized to	5th GrC
U	$\rightarrow$	$\mathcal{U} = \{U_j^h, h, j, = 1, 2, \ldots\}$
$F^1 \subseteq U$	$\rightarrow$	$R^1 \subseteq U_1^1 \times U_2^1 \times \cdots$
…	…	…
$F^j \subseteq U$	$\rightarrow$	$R^j \subseteq U_1^j \times U_2^j \times \cdots$
…	…	…

1) Second GrC Model Let U be a classical set, called the universe. Let $\beta = \{F^1, F^2, \ldots\}$ be a family of subsets. Then the pair (U, β), called the Global GrC Model or the Second GrC Model. The β, some time, is called Partial Covering (PCov).

A *granule* can be intuitively defined as a clump of objects that are drawn together by the constraints X, where X can be indistinguishability, similarity or functionality and etc [14,38]. In the Second GrC model, a granule is a set, namely, we have implicitly assumed that the constraints are uniform. In general, each object may receive distinct constraints, so in the Fifth GrC model, a granule is a tuple, not necessarily a set: One can regard the constraints as a schema (of a relational database), and a granule is a tuple under such a schema. In this case the collection of granules are tuples from various relations.

To understand the next model, Table 1, that explains the generalization process, may be helpful.

2) Fifth GrC Model

1. Let $\mathcal{U} = \{U_j^h, h, j, = 1, 2, \ldots\}$ be a given family of classical sets, called the universe. Note that distinct indices do not imply that the sets are distinct.
2. Let $U_1^j \times U_2^j \times \cdots; j = 1, 2, \ldots$ be a family of Cartesian products of various lengths.
3. Recall that an n-ary relation is a subset $R^j \subseteq U_1^j \times U_2^j \times \cdots U_n^j$.
4. Let $\beta = \{R^1, R^2, \ldots\}$ be a given family of n-ary relations for various n.

Then the pair $(\mathcal{U}, \beta)$, called the Relational GrC Model, is the formal definition of the Fifth GrC Model.

Introduction

What is Granular Computing (GrC)? The best approach is to trace how the intuitions have been evolved. For this purpose, let us recall the event. In the academic year 1996–97, when Lin had his sabbatical leave at Berkeley, Zadeh suggested granular mathematics (GrM) to be his research area. To limit the scope, Lin proposed the term granular

computing [40]. So, at the beginning, roughly GrC is the computable part of granular mathematics.

What is Granular Mathematics (GrM)? Zadeh in his 1979 paper [37] had implicitly explained his view. Here, we take a simpler view: It is a new mathematics, in which "points" are replaced by or associated with "granules." We call this process granulation, and the collection of granules the granular structure.

What is a Granule? This is the main topic. There is an obvious candidate, namely, an equivalence class of a *partition* that is an ancient notion in mathematics. In general, a granule can be a crisp/fuzzy subset, a function, an algorithm, a random variable (measurable function), a generalized functions etc.

Traditionally, "how to solve it" [31] cannot be any part of formal mathematics, however, "how to compute it" is an integral part of computing. So GrC has included the mathematical/computational problem solving practices.

Classical Examples of Granulation

Ancient Examples

E1 Granulation of the Human Body: The Granules Are Head, Neck, Body, Hand etc. Many daily things are routinely granulated into "sub"things, probably since ancient time. Human body is granulated into head, neck, etc. Currently, there are no flawless formal models to capture such intuitive concept yet. The obvious one does not work well: One can easily write down a fuzzy membership function to represent a head, a neck or a body. However, the likelihood of any two persons to write down the same membership function for the same granule (e. g., head) is extremely low. Hence we need a much more subtle theory. A new proposal on qualitative fuzzy set theory is in preparation by this author, this example can be realized.

E2 Granulation of Space and Time The space and time has been granulated intuitively into infinitesimal granules by early scientists; this notion was known to Zeno, Archimedes, etc. This intuitive notion led to the invention of calculus by Newton and Leibniz. However, its formalizations, theory of limit (18th century), topology (early 20th century [28]) and nonstandard analysis (mid 20th century [32]) are relatively recent. This ancient example, inspired two models, Local GrC model (First GrC Model) and Global GrC Model (Second GrC Model).

E3 Partition The simplest kind of granulation is partition (see glossary). Its algebraic correspondence, equivalence relation, has played an important role in the Euclidean Geometry (300 BC).

Modern Examples

E4 Local Granules of Uncertainty – from Quantum Mechanics Heisenberg's uncertainty principle states that, in general, neither the momentum nor the position of a particle can be determined simultaneously with arbitrary great precision. In other words, a point of the momentum (a precise measurement) can determine only a "neighborhood" of positions and vice versa. The idea is simplified into the Third GrC Model (Binary GrC Model).

E5 Local Granules of Basic Knowledge – Discretionary Access Control Models in Computer Security This example is a common data structure in many computer systems. For each user, there is a set of users (friends) who can access his files, or a set of users (foe), who cannot access his files; this set is called explicitly denied access list.

Examples E4 and E5 are "serious" examples. The idea is modeled in the third GrC Model. Mathematically, it is equivalent to a binary relation. Geometrically, a binary relation is a graph, or network.

E6 Global Granules of Knowledge – Simplicial Complexes A simplicial complex consists of two objects: One is a finite set of vertices, another one is a family of subsets, called simplexes, of vertices that satisfy the closed condition, namely a subset of a simplex is a simplex. It is a common mathematical structure in combinatorial topology. Currently, it is finding its way to web technology [21].

Further Examples

Next, we give examples of non-commutative granules, which is a generalization of a binary relation (Binary GrC Model; Third GrC Model).

E7 A committee in a Human Society Is a Granule in a Social Network A committee in a human society (a set of human beings) is a granule. Observe that each member may play different roles, so the committee may not consist of homogenous members; so the members cannot exchange their roles. By viewing the collection of roles as a relational schema, a committee is a tuple. This idea is modeled in the Fifth GrC Model. Observe that different types of committees have different schema. The set of committees under the same schema forms a relation. A collection of n-nary relations for various n in a society can be viewed as a granulation of the society.

In these examples, granules can be fuzzy sets. Since a fuzzy set is characterized by a membership function, so we have generalized the idea further.

E8 A Granule Can Be a Function, a Random Variable (A Measurable Function) or Even a Generalized Function (Such as the Dirac Delta Function)
This class of examples led to the Sixth GrC model.

Traditionally, "how to solve it" [31] has not been any part of formal mathematics, however, "how to compute it" is an integral part of computing. So GrC includes some mathematical/computational practices.

E9 A granule can be a lemma in a mathematical proof, a subprogram in a program, or a machine/cluster of machines in a grid/clouding computing environment. Formally, within the computable domains, a granule is a sub-Turing machine. This is modeled in the Seventh GrC Model. For general cases, they are included in the Ninth GrC Model

E10 Computers or clusters of computers in Grid/Cloud computing are granules. These are hardware examples. This also belongs to the Seventh GrC Model

Formal Models of Granulation

On the basis of these examples, nine models are discussed. The category theory-based model (Eighth GrC Model) has been proposed by this writer in the keynote of GrC2008 as *the formal model of GrC*. The rest of the eight models are basically "convenient models" in the sense that they can be derived from the most general model, but for convenience, they are modeled independently.

First GrC Models and Ancient Examples

This model is derived from the Ancient Example E2, which probably is the most lively example in the early notion of granules. It led to the invention of calculus by Newton and Leibniz. Actually the idea was much more ancient; it was in the mind of Archimedes, Zeno, etc. Yet the solutions were in modern time. Two formalizations had emerged. One was the concept of limit (19th century) that led to the notion of topology (early 20th century). The other one was the nonstandard analysis (mid 20th century), which formally realized the original intuition.

These developments conclude that the following concepts are "equivalent" (not in the formal mathematical sense, as the first one is merely an intuitive notion):

1. The ancient intuitive notion of infinitesimal granule.
2. The formal infinitesimal granule in the non standard analysis.
3. Topological Neighborhood System (TNS) in the standard world.

In other words, the ancient intuition of infinitesimal granules (with the required properties) is realized, not by a set, but by a family of subsets, that satisfies the axioms of topology. Nevertheless, in this paper a (modern) granule will refer to a neighborhood, but, not to the whole neighborhood system.

The notion of topology can be defined in two ways:

1. A topology τ is a family of subsets, called open sets, that satisfies the (global version) axioms of topology.
2. A topology, called topological neighborhood system (TNS), is an assignment that associates to each point p a family of subsets, TNS(p), that satisfies the (local version) axioms of topology.

These two definitions lead us to First and Second GrC Models (Local and Global GrC Models). In this subsection, we will focus on the First GrC Model: Let U and V be two classical sets. Let NS be a mapping, called neighborhood system (NS),

$$\mathrm{NS}\colon V \to 2^{(P(U))},$$

where $P(X)$ is the family of all crisp/fuzzy subsets of X. 2^Y is the family of all crisp subsets of Y, where $Y = P(U)$. In other words, NS associates each point p in V, a family NS(p) of crisp/fuzzy subsets of U. Such a subset is called a neighborhood (granule) at p, and NS(p) is called a neighborhood system at p.

Definition 1 (First GrC Model) The 3-tuple (V, U, β) is called the **Local GrC Model**, where β is a neighborhood system (NS). If $V = U$, the 3-tuple is reduced to a pair (U, β). In addition, if we require NS to satisfy the topological axioms, then it becomes a TNS.

Some Intuition Behind the NS The following arguments are adopted from my pre-GrC paper [13]

(1) The Meaning of "Near"

The notion of near is rather difficult to formalize. Let us examine the following two examples.

1. Is Santa Monica "near" Los Angeles? Answers could vary. For local residents, who have cars, answers are often "yes." For visitors, who have no cars, answers may be "no."
2. Is 1.73 "near" $\sqrt{3}$? Again answers vary; it depends on what should be the appropriate tolerance radius.

Intrinsically "near" is a subjective judgment. One might wonder whether there is a scientific theory for such subjective judgments? Mathematicians have offered a nice solution. They simply include all contexts into its formalism.

Here is the formalism of the second question: Given the radius of an acceptable error, say, radius of errors 1/100 (a given context)

Is 1.73 "near" $\sqrt{3}$?

With the agreement 1/100 is acceptable, then 1.73 is near $\sqrt{3}$! In this case, all possible contexts are ε that represents all positive real numbers. Similarly, if a neighborhood system has been assigned to each city in the Los-Angeles-area. For example, based on car driving, public transportation, walking etc., we assign a neighborhood to *each* city for *each* context. Under such a concept of a neighborhood system, we could have a definite answer for Example 1. So a proper formulation for such a question is:

Assuming that we are taking public transportation (a given context), Is p near q?

Therefore, a neighborhood system is a good infrastructure for addressing the concept of "near"! These analysis leads to the following conclusions.

1. In Modeling, a neighborhood system is a good infrastructure for providing all possible contexts.
2. Under this model, in an application, selecting a context means selecting *a fixed* neighborhood as a unit of tolerance (uncertainty).

Now, under this concept, we will re-examine previous examples.

Example 1 If we have chosen "driving half an hour" as acceptable distance, then Santa Monica is "near" Los Angels.

Example 2 Let the collection of ε-neighborhoods be the neighborhood system for the real numbers R; then (R, ϵ-neighborhoods) is a First GrC Model, where ε could take any real value. Now, we re-state the previous example using this First GrC Model

1. Assuming we have agreed $\epsilon = 1/100$ is acceptable, then 1.73 is "near" $\sqrt{3}$.
2. But, if we have only agreed $\epsilon = 1/1000$ then 1.73 is not "near" $\sqrt{3}$.
3. Next let us consider a deeper question:

   Is the sequence $1, 1/2, 1/3, \ldots, 1/n, \ldots$ "near" zero?

   Then, it is possible, we can have a "yes" answer for all contexts: For any given context, namely, $\epsilon > 0$, there is a number $N = [1/\epsilon] + 1$, such that, for all $n > N$, $1/n$ is "near" zero, where $[1/\epsilon]$ denotes the biggest integer $\leq 1/\epsilon$.

Readers who are familiar with the standard (ϵ, δ)-definition of limit can spot the origin of neighborhood systems. Such a context-free (all possible contexts) answer is precisely the classical notion of limits, $\lim_{n\to\infty} 1/n = 0$. Using our language, we may say that limit is the context free answers of "near".

Perhaps we should also point out here that there are no context free answers for the question whether two points are "near."

Brief pre-GrC Historical Notes

1. In 1988–1989, Lin generalized TNS to the Neighborhood Systems (NS) by simply dropping the (local version) axioms of topology [7,9] and apply it to approximate retrievals. Each neighborhood was treated as a unit of uncertainty.
2. In the same year (1989), Lin also examined a non-reflexive and symmetric binary relation (conflict of interests) for computer security from the view of NS [8].
3. Abstractly, Lin imposed the NS structure on the attribute domains for approximate retrieval. Taking this view, we should mention that earlier D. Hsiao imposed equivalence relations on the domain for access precision in early 1970 [6,36]. In 1980, S. Ginsburg and R. Hull had imposed partial ordering on attribute domains [4,5].
4. In much earlier, NS was studied in [33] as a generalization of topology. Note that however, there are fundamental differences, for example, the concept of closures are different. The term pre-topology also has been used for referring NS and TNS.
5. In the early GrC period, Lin, by mapping the NS onto Zadeh's intuitive definition, used NS as his first mathematical GrC model [14,15,16].

Second GrC Models and Modern Examples

As in the previous case, by dropping the global axioms of topology, we have the Second GrC model.

Definition 2 (Second GrC Model) The Pair (U, β), where β is a family of subsets of U, is called **Global GrC Model**. The β, some time, is referred to as a partial covering (PCov).

Note that the Second GrC model is a special case of the First GrC model: If we regard the sub-collection of all members of the partial covering β, that contains p, as a neighborhood system at p, then this Second GrC model is an example of the First GrC model.

The modern example, simplicial complexes, is an important example of such a model: A simplicial complex

consists of a set of vertices and a family of subsets, called simplexes, that satisfies the closed condition [34].
[Digression] Perhaps, it is worthwhile to note that

- The closed condition of the simplicial complex is the a priori principle in association (rules) mining.

This observation plays an important role in document clustering [24].

Third and Fourth GrC Models and Modern Examples

In this section, we will build a new model that realizes modern example E4 and E5. Recall that E4 concludes that a precise measure of the momentum can only determine a (probabilistic) "neighborhood" of positions; and E5 concludes that in computer security, the Discretionary Access Control Model (DAC) assigns to each user p a family of users, Y_i, $i = 1, \ldots$, who can access p's data. In other words, each p is assigned a granule of friends.

To formalize these examples, let U and V be two classical sets. Each $p \in V$ is assigned a subset, $B(p)$, of "basic knowledge" (a set of friends or a "neighborhood" of positions).

$$p \to B(p) = \{Y_i, i = 1, \ldots\} \subseteq U\,.$$

Such a set $B(p)$ is called a (right) binary neighborhood and the collection $\{B(p) \mid \forall p \in V\}$ is called the binary neighborhood system (BNS).

Definition 3 (Third GrC Model) The 3-tuple (U, V, β), where β is a BNS, is called a **Binary GrC Model**. If $U = V$, then the 3-tuple is reduced to a pair (U, β).

Observe that BNS is equivalent to a binary relation (BR):

$$\mathrm{BR} = \{(p, Y) \mid Y \in B(p) \text{ and } p \in V\}\,.$$

Conversely, a binary relation defines a (right) BNS as follows:

$$p \to B(p) = \{Y \mid (p, Y) \in \mathrm{BR}\}\,.$$

So, both modern examples give rise to BNS which was called a binary granular structure in [14]. We would like to note that based on this (right) BNS, the (left) BNS can also be defined:

$$D(p) = \{Y \mid p \in B(Y)\} \text{ for all } p \in V\}\,.$$

Note that BNS is a special case of NS, namely, it is the case when the collection $\mathrm{NS}(p)$ is a singleton $B(p)$. So the Third GrC Model is a special case of the First GrC Model.

The algebraic notion, binary relations, in computer science, is often represented geometrically as graphs, networks, forests etc. So the Third GrC Model has captured most of the mathematical structure in computer science.

Next, instead of a single binary relation, we consider the case where β is a set of binary relations. It was called a [binary] knowledge base [14]. Such a collection naturally defines an NS.

Definition 4 (Fourth GrC Model) The Pair (U, β), where β is a set of binary relations, is called Multi-Binary GrC Model. This model is most useful in data bases; hence it has been called Binary Granular Data Model (BGDM), in the case of equivalence relations, it is called Granular Data Model (GDM)

Observe that a Fourth GrC Model can be converted, say by a mapping $\mathcal{G}$, to a First Model. Conversely, a First GrC Model induces, say by $\mathcal{F}$, to a Fourth Model. So First and Fourth models are equivalent, but not naturally, namely, $\mathcal{G}$ and $\mathcal{F}$ are not inverse to each other.

Models for Further Examples

We have observed in Sect "Definition of the Subject" that the collection of n objects that are "drawn together" is, not necessary a subset, but is a tuple in an n-ary relation. For example, if the universe is a human society then a group of people may be drawn into a committee with distinct roles, such as the chair, vice chair, secretary, treasurer, etc. As every member has a different role, they can not be swapped around. So the committee is not a set; it is a tuple under the schema that consists of distinct roles.

Definition 5 (Fifth GrC Model)

1. Let $\mathcal{U} = \{U_j^h, h, j, = 1, 2, \ldots\}$ be a given family of classical sets, called the universe. Note that distinct indices do not imply that the sets are distinct.
2. Let $U_1^j \times U_2^j \times \cdots$ be a family of Cartesian products of various length.
3. Recall that an n-ary relation is a subset $R^j \subseteq U_1^j \times U_2^j \times \cdots U_n^j$.
4. Let $\beta = \{R^1, R^2, \ldots\}$ be a given family of n-ary relations for various n.

The pair $(\mathcal{U}, \beta)$, called the Relational GrC Model, is a formal definition of the Fifth GrC Model.

Note that this granular structure is the relational structure (without functions) in the First Order Logic, if n only varies through finite cardinal number.

For the next two models, we will use the language of category theory in next sub-section. We may note that we have not committed ourselves to every specific details yet.

Definition 6 The Sixth GrC Model is in the categories of functions, random variables, and even generalized functions.

Fuzzy sets are described by membership functions, so granules can be regarded as membership functions; note that the First to Fifth GrC Models include fuzzy sets. Hence, we consider further generalizations: granules are functions, random variables (measurable functions), generalized functions (e. g. Dirac delta functions).

In the case, a granule is a function, we may require that the granular structure (the collection of granules) has the universal approximation property, namely, any function in the universe can be approximated by the functions in the collections. The membership functions selected in fuzzy controls do have such properties. In neural networks, the functions generated by the activation functions also have such a property [29].

In the case of probability/measure theory, quantum mechanics may be a good guiding example.

Definition 7 The Seventh GrC Model is in the category of Turing machines.

For example, a collection of lemmas in a mathematical proof (mechanizable), a set of subprograms in a computer program, or a computer or cluster of computers in grid/cloud computing are granules in the model.

Definition 8 The Ninth GrC Model is in the category of qualitative fuzzy sets.

This model was proposed after the Eighth GrC model. It has not been published in printing form yet. The idea is similar to the model that we have called it sofset (this is not a typo) [12]. It associates to each "real world" fuzzy set, a collection of membership functions; please watch for new development.

Category Theory Based Models

Now we generalize the category of sets to general categories. It is somewhat a surprise that this is basically the same as the category of relational databases [10]. In other words, the abstract structures of data and knowledge are similar. After analysis, it seems reasonable; because in GrC approach, the basic unit of knowledge is a granule of data.

Let us set up some language for the Category Theory. A category consists of

1. A class of objects, and
2. A set $\mathrm{Mor}(X, Y)$ of morphisms for every ordered pair of objects X and Y, which satisfies certain properties. For this paper, the formal details are not important; we only need the language loosely.

Here are some examples.

1. The Category of Sets: The objects are classical sets. The morphisms are the maps.
2. The Category of Sets with binary relations as morphisms: The objects are classical sets. The morphisms are binary relations. This is the Category of Entity Relationship Models.
3. The Category of Power Sets: The object U_X is the power set $P(X)$ of a classical set X. Let U_Y be another object, where Y is another classical set. The morphisms are the maps, $P(f)\colon U_X \to U_Y$ that are induced by maps $f\colon X \to Y$.

Let CAT be a given category.

Definition 9 (Category Theory Based GrC Model)

1. $C = \{C_j^h, h, j, = 1, 2, \ldots\}$ is a family of objects in the category CAT.
2. There are families (which are bags; see the glossary) of Cartesian products, $C_1^j \times C_2^j \times \cdots$ of objects, $j = 1, 2, \ldots$ of various lengths. They are called product objects.
3. An n-ary relation object R^j is a sub-object of the product object $C_1^j \times C_2^j \times \cdots C_n^j$.
4. $\beta = \{R^1, R^2, \ldots\}$ be a family of n-ary relations (n could vary).

The pair(C, β), called the Categorical GrC Model (Eighth GrC model), is the formal model of granulation.

By specifying the general category to various special cases, we have all models: By specifying the category to be the category of sets, we have the Fifth GrC model. Further by limiting n to 2, we have the First GrC Model and the Fourth GrC Model. By assuming the symmetry for all n-ary relations, we have the Second GrC Model. By restricting the number of relations to be one and $n = 2$, we have the Third GrC Model. For the Sixth and the Seventh, further researches are needed.

Overview of Early GrC Models

Schematically we summarize the relationships based on the Granular Structures of early GrC Models as follows: (the diagram will be different, if it is based on their approximation spaces). "$\Rightarrow\Leftarrow$" is a two-way generalization but they are not inverse to each other. "$\Rightarrow$, $\Uparrow$, and $\Downarrow$" are one

way generalizations. "GM" means GrC Models and RST means Rough Set Model.

$$\text{RST} \Longrightarrow \left[\begin{array}{lcl} \text{Global GM} & \Rightarrow & \text{Local GM} \\ \Downarrow & & \\ \text{Multi Binary GM (BGDM)} & \begin{array}{c}\Leftarrow \\ \Rightarrow\end{array} & \text{Local GM} \\ \Uparrow & & \\ \text{Binary GM} & \Rightarrow & \text{Relation GM} \end{array}\right]$$

$$\Rightarrow \text{Category GM}\,.$$

Granular and Related Structures

Four structures are related to each other, they are all briefly explained in Sect. "Definition of the Subject." We will focus on a quotient structure here.

A granule can be examined from three states: We will use the following example to illustrate the idea. Let U be the set $(Z, +)$ of the integers as an additive group, let E and O be the set of even and odd integers respectively.

1. Isolated State (Internal State): Let us consider the case E is in the isolated state, that is, independent from the universe U. In this case, only the internal structure of E is available to us: we can only know that the number 2 is playing the role of the identity in an additive group, but is *not* aware of the existence of 1 in the outside of E. So in isolated state, E is the additive group $(Z, +)$. Some time, this state may also be called internal state as only the internal structure is involved.
2. Embedded State (Conceptual State): In this state, E is a subgroup of U. We not only know that the number 2 is playing the role of the identity in E and we also *do know* that 2 is an even integer in U.
   [Digress] In category theory, a subobject (in this case a subgroup) is often represented by a pair (a group, the mapping of the group into the subgroup),
   $$(Z, +) \rightarrow_{\text{homomorphism}} E \subseteq (Z, +)\,.$$
   The pair (the first $(Z, +)$, $\rightarrow_{\text{homomorphism}}$) is the subobject (subgroup). The first component $(Z, +)$ represents the internal structure of (or isolated state of) even integers, the middle E is the embedded state of even integers. Since this state involves with every aspect of the concept of a granule, one may call it conceptual state.
3. Quotient State (External State): Note that the quotient structure Q consists of two elements, namely, the two subset of even and odd integers (E and O, respectively). But the two subsets appear as two elements, $[E]$, and $[O]$ in Q. Note the bracketed symbols emphatically denote their roles as points (elements) of a set; no contents of subsets are visible. The interactions between the two subsets induce interactions between these two elements. So Q is the integer mod 2 (as additive group). In notations,
   $$Q = (Z_2, +) \equiv \{[0]_2, [1]_2\}$$
   as an additive group. So the quotient state of E is the element $[0]_2$ in Q. Since the quotient state only involves the external relationships with other granules, this state may also be called external state.

With these preparations, we can define the concepts of granular structure (GrS), quotient structure (QS) and knowledge structures (KS): GrS is the collection of granules in embedded states. The quotient structure is the structure of granules in quotient states, namely, each granule is abstracted to an element (point) and the interactions among granules are abstracted to the interactions among elements (points).

Among these four structures only the quotient structure is a difficult concept. Here, we will illustrate only the quotient structure. We will start from the simplest case, namely, the GrS is a partition. By definition, a partition is a collection of equivalence classes that are mutually disjoint and their union is the whole universe. Hence, by abstracting each equivalence class to a point, we have a set of points that have no interactions. This implies:

Proposition 1 *The quotient structure of a partition is a classical set.*

Next, we need a lemma.

Lemma 1 *The collection of complete inverse image of a given mapping $f: U \rightarrow V$ defines a partition. Namely, the collection $\{f^{-1}(f(p)) \mid p \in U\}$ forms a partition.*

So NS and BNS induce partitions on U. These partitions have been called the derived partition (or derived equivalence relation); see [14].

Let $B: U \rightarrow P(U)$ and $\text{NS}: U \rightarrow 2^{P(U)}$ be a binary neighborhood system (BNS) and neighborhood system (NS), respectively. Then (U, NS) and (U, B) are First and Third GrC Models, respectively. By the previous Lemma, the complete inverse images of the two mappings, B and NS, are partitions (equivalence relations). The two partitions, denoted by E_B and E_{NS}, induce two quotient sets U/E_B, U/E_{NS}, respectively. Moreover, the pre-topologies, BNS and NS, of U naturally induce the respective pre-topologies on the quotient sets U/E_B and U/E_{NS}, respectively [9,11].

Proposition 2 *Such pre-topological spaces U/E_B (BNS-space) and U/E_{NS} (NS-space) are the quotient structures of*

Binary and Local GrC Models (Third and First GrC Models) respectively.

This example indicates that there are additional pre-topological structures on the information tables [2,22].

Let (U, β) be a Global GrC Model, where $\beta = \{F^1, F^2, \ldots\}$ be a partial covering, We will use $[F^{j_1}], [F^{j_2}], \ldots$ to denote a set of points (when we think of $F^1, F^2, \ldots$ as points).

Proposition 3 *The granular structure, β, generates a semi-group $S(\beta)$ under set theoretical intersection. Then the quotient structure of $(U, S(\beta))$ is a semi-group generated by $[F^{j_1}], [F^{j_2}], \ldots$ under "intersection" $[F^{j_1}] \circ [F^{j_2}] = [F^{j_1} \cap F^{j_2}]$.*

This example indicates that there are additional algebraic structures on the information tables [20].

We observe that each quotient structure is not easy to determine. It depends on what are the mathematical structures under consideration. Let us recall some works from pure mathematics. Let the universe U be the ring Z of integers. In ring theory, the collection β of all prime ideals is the center of attentions. Let p be a prime number, then the prime ideal consists of following set

$$\{\ldots, -2p, -p, 0, p, 2p, \ldots\},$$

together with some algebraic structure.

By regarding each prime ideal as a point, the collection of prime ideals is a set, often denoted by Spec(Z) in algebraic geometry. The structure of prime ideals turns Spec(Z) into a topological space under Zariski topology.

Example 3 (Two granular and quotient structures)

1. The quotient structure of (U, β), where β is the collection of the prime ideals, is a topological space Spec(Z) [27].
2. The quotient structure of $(U, S(\beta))$, where $S(\beta)$ is the semi-group generated by the intersection, is isomorphic to the semigroup of positive integers.

In RST, the approximation spaces of the first and second examples are different. However in GrC, the approximations (under the knowledge engineering view) of the two examples are the same. In GrC, granules represent known basic units of knowledge (known concepts), hence the intersections of known concepts are known concepts. So in GrC we take all possible intersections of granules to approximate unknown concepts. This example clearly indicates the effectiveness of this view.

Integration – A Dual of Granulation

Classical Divide (partitioning) and Conquer is extended to Granulate and Conquer. In other words, we are considering the cases, in which the sub-problems may not be logically independent from each other. So the standard thinking may not work, for example

- It may take Np-hard time to granulate a problem, from top to bottom.

This issue is not new. Some cases have been addressed in the topic of "dynamic programming" in data structure courses. We have used "topological divide" to solve the cases in the Third GrC model [18].

In this section, we will explore the following problem: For convenience, we will use the term, sub-structure, to denote (1) the collection of the internal structures of each granule and (2) the quotient structure. Now, we raise the following questions:

- Can we find more than one universes (granular models) that have the same sub-structure?
- Equivalently: Could two distinct problems, after granulation, have the same sub-structure?

In homological algebra, this is called the extension problem; in computer science, we feel that integration is a better name.

We will illustrate the concept in the following two cases

1) Simple Integration: no information structure
2) Integration with information structure.

Simple Examples of Integration

This is a truly unexplored area; we will explain the case when granulations are partitions.

A Simple Partition We will illustrate the key concepts by simple examples. Let the universe U be the *set* of all integers, $Z = \{\ldots, 1, 0, 1, \ldots\}$ with the following granular structure, namely, a given collection of two granules

$$\{\{\ldots, -2, 0, 2, \ldots\}, \{\ldots, -3, 1, 3, \ldots\}\}.$$

We name the granules E and O (the even and odd integers). Using the languages of Sect. "Granular and Related Structures," we say E and O are in the embedded state. We will use $[E]$ and $[O]$ to denote the quotient state. So the set $Q = \{[E], [O]\}$ is the quotient set. Intuitively, Q is a set of black boxes; the internals of black boxes are invisible.

Now we will summarize the important structures of U so that integration can be formulated: Let Int(X) denote

the internal structure of a granule X, that is, the structure of X in isolation (and not a subset of U).

1. $A = \text{Int}(\{\ldots, -2, 0, 2, \ldots\})$ is the set of integers Z.
2. $B = \text{Int}(\{\ldots, -3, 1, 3, \ldots\})$ is the set of integers Z.
3. Two copies of integers, A and B, are mapped to even integers and odd integers respectively.
4. Q is the quotient set that consists of $[E]$ and $[O]$ as two elements (points).

Schematically, we give the following situation:

$$Z = \left\{ \begin{array}{l} A = \text{Int}(E) \to E \\ B = \text{Int}(O) \to O \end{array} \right\} (\subseteq) \, U \to Q = \{[E], [O]\} \, .$$

Integrations on Partitions From the point of view of problem solving,

- The quotient structure represents the recipe (a set of higher level instructions) of integrating the sub-solutions. It is the "Main Program" that integrates the returns of sub-program calls.

Here a sub-solution means the solution of a sub-problem that has been solved in isolation. Now the integration problem can be formulated as follows:

1. We only know that two copies of integers, A and B, are mapped to the two subsets of unknown universe.
2. The images of the mappings form a partition in the unknown universe U
3. But we do know its quotient structure $Q = \{[E], [O]\}$, namely a set of two elements, where $[E]$ represents the point that is abstracted from the image E of A in U; similarly $[O]$ is that of B.

Schematically, we can summarize it as follows: The unknown U has two unknown subsets, image A and image B, that form a partition on unknown U, but with a known quotient set which is equivalent to the integers mod 2.

$$Z = \left\{ \begin{array}{l} A\sim \to \text{Image A} \\ B \to \text{Image B} \end{array} \right\} (\subseteq) \ \text{Unknown} \ \ U \to Q = Z_2$$

Can we construct the unknown universe? And is it unique? The answer is "yes" and it is unique. It is the Cartesian product $Q \times Z$. Observe that from a set-theoretical point of view, $Q \times Z$ is equivalent to Z, so the constructed one is the old friend.

Integration on Partition with Information Structure

Let us consider a second view on the set of integers. But this time, the universe carries additional information, namely, the additive structure of integers, $(Z, +)$. This universe is denoted by $(U, +)$. Then

1. $\text{Int}(E)$ is the *additive group* $(Z, +)$ of integers, and $\text{Int}(O)$ is a *set* Z of integers.
2. The quotient structure $(Q, +) = \{[E], [O]; +\}$ is an additive group: $[E]+[E] = [E]$, $[E]+[O] = [O]+[E] = [O]$, $[O] + [O] = [E]$. This $(Q, +)$ is often called the integer mod 2, and denoted by $(Z_2, +)$.

Again, we give a similar situation

$$\left\{ \begin{array}{lll} (Z, +) = \text{Int}(E) & \to_{(\text{homomorphism})} & E \\ Z = \text{Int}(O) & \to_{(\text{map})} & O \end{array} \right\} \subseteq (U, +) \to (Q, +) \, .$$

The upper arrow

$$(Z, +) = \text{Int}(E) \to_{(\text{homomorphism})} (U, +) \to (Q, +)$$

is very similar to a short exact sequence (SES) in homological algebra.

Now, we will extend the SES to granules:

$$\text{Int(GrS)} \to_{\text{map}} U \to QS \, ,$$

where (1) Int(GrS) is the collection of the internal structures of all granules, (2) its image in U (under the map $\to_{\text{map}}$) is GrS, the granular structure in U, and (3) QS denote the quotient structure of GrS. The map, $\to_{\text{map}}$, is on-to-one within each granule, but two granules may have overlapping images in U. We will call this a *granular exact sequence*.

Can $(U, +)$ be reconstructed back?

The answer is more than yes; there are *two* solutions! The two solutions are: $(Z_2 \times Z, +)$ and $(Z, +)$. They are not equivalent as additive groups. This fact can be expressed by the extension functor, namely, $EXT(Z, Z_2) \neq 0$.

The important question is

Could such a functor be extended to formal granular models?

The answer may be "yes." The granular exact sequences shall play a similar role as short exact sequences. There are some preliminary results in [19]; for example, the fact that the knowledge representation of a symmetric binary relation is complete, implies that the integration is unique.

Integrations on Non-Partition Case

We will consider a simple example using the concept of granular exact sequence. Let $U = \{a, b, c\}$, let β be the set

of all subsets of U. So the internal structure of the granules are: (we will use n-granule to denote a granule of n elements) One copy of 3-granule; three copies of 2-granule, and three copies of 1-granule. This is a Global GrC Model. The quotient structure is a collection of 7 elements (to find the quotient structure is not easy, but easy to verify). These elements are related by partial ordering, called "a face of." We will use the geometry to represent this quotient structure; it is a triangle spanned by $\mathbf{i}(1,0,0)$, $\mathbf{j}=(0,1,0)$, $\mathbf{k}=(0,0,1)$. The triangle can be viewed geometrically (a simplicial complex of closed triangle) as ONE open triangle, THREE open segments, and THREE vertices. Its a partial ordered set of 7 elements.

Now the granular exact sequence (three steps) can be described as follows:

1. 3-granule $\{g_1^3, g_2^3, g_3^3\} \to$ a subset of unknown universe $\to$ open triangle $\Delta(\mathbf{ij}k)$
2. first 2-granule $\{g_1^2, g_2^2\} \to$ a subset of unknown universe $\to$ first open segment $\Delta(\mathbf{ij})$
3. second 2-granule $\{g_3^2, g_4^2\} \to$ a subset of unknown universe $\to$ second open segment$\Delta(\mathbf{ik})$
4. third 2-granule $\{g_5^2, g_6^2\} \to$ a subset of unknown universe $\to$ third open segment$\Delta(\mathbf{jk})$
5. first 1-granule $\{g_1^1\} \to$ a subset of unknown universe $\to$ first vertex $\Delta(\mathbf{i})$
6. second 1-granule $\{g_2^1\} \to$ a subset of unknown universe $\to$ second vertex $\Delta(\mathbf{j})$
7. third 1-granule $\{g_3^1\} \to$ a subset of unknown universe $\to$ third vertex $\Delta(\mathbf{k})$

We will use $I(g_j^i)$ to denote the image in the unknown universe. We do know many distinct g_j^i mapped to the same point in the quotient structure. Here are the mappings

1. 3-granule $\{I(g_1^3), I(g_2^3), I(g_3^3)\} \to \Delta(\mathbf{ijk})$
2. first 2-granule $\{I(g_1^2), I(g_2^2)\} \to \Delta(\mathbf{ij})$
3. second 2-granule $\{I(g_3^2), I(g_4^2)\} \to \Delta(\mathbf{ik})$
4. third 2-granule $\{I(g_5^2), I(g_6^2)\} \to \Delta(\mathbf{jk})$
5. first 1-granule $\{I(g_1^1)\} \to \Delta(\mathbf{i})$
6. second 1-granule $\{I(g_2^1)\} \to \Delta(\mathbf{j})$
7. third 1-granule $\{I(g_3^1)\} \to \Delta(\mathbf{k})$

From the maps above, we may make some identifications: Observe that from the maps, 1st, 2nd, 5th, we can identify, without losing generality, that $I(g_1^3) = I(g_1^2) = I(g_1^1)$. By similar arguments, we have

1. $I(g_1^3) = I(g_1^2) = I(g_3^2) = I(g_1^1) \to \Delta(\mathbf{i})$
2. $I(g_2^3) = I(g_2^2) = I(g_5^2) = I(g_2^1) \to \Delta(\mathbf{j})$
3. $I(g_3^3) = I(g_4^2) = I(g_6^2) = I(g_3^1) \to \Delta(\mathbf{k})$

So the 12 points are actually three points; they will be denoted by $\{I(g_1^3), I(g_2^3), I(g_3^3)\}$. Thus, the granular structure of the unknown universe is: one 3-granule $\{I(g_1^3), I(g_2^3), I(g_3^3)\}$, three of its 2-subgranules $\{I(g_1^3), I(g_2^3)\}$, $\{I(g_1^3), I(g_3^3)\}$, $\{I(g_2^3), I(g_3^3)\}$ and three of 1-subgranules $\{I(g_1^3)\}$, $\{I(g_2^3)\}$, $\{I(g_3^3)\}$. So we have re-captured the unknown U and β. – This is the integration. The key question is: Is this U unique? The answer is "no"; we will skip it here.

Semantic Views

Granules may be interpreted from three views.

1. Uncertainty Theory: A granule is a unit of lacking precise knowledge. Both L. A. Zadeh and T. Y. Lin, who coined the label, started from the uncertainty theory. Lin took his neighborhood system as a system of uncertainty. Zadeh has a grand project [40].
2. Knowledge Engineering: A granule is a unit of basic knowledge (Information). In their book, D. Stanat and D. McAllister state "Knowledge varies in sophistication from simple classification to ..." [35]. A classification is a partition, so an equivalence class is a (unit of) basic knowledge; this view is also promoted by Pawlak [30]. So we believe: a granule is a (unit of) basic knowledge.
3. How-to-Solve/Compute-it: A granule is a sub-problem or a software unit. It is a special type of basic knowledge.

Each view may have its own GrC theory: Some fundamental operators are:

1) *Information Hiding*: It is a transformation of granular structures into quotient structures (see glossary).

A quotient structure is the mathematical structure of the collection of granules, in which each granule is regarded as an element (point), and the interactions among granules are transformed into the interactions among elements. For example, in group theory a quotient group is a collection of cosets, in which each coset is regarded as an element and the multiplication of cosets is abstracted into the multiplication of elements in the quotient group. Using the software engineering language, a granule in a quotient structure is a black box, while in a granular structure, it is a white box. These are easy cases, for granulations that have nonempty overlappings, the quotient structure may not be easy to determine; see Section Granular and Related Structures.

2) *Information Integration*: see Sect. "Integration – A Dual of Granulation."

3) *Knowledge Representation*: This amounts to give each point in the quotient structure (Sect. "Granular and Related Structures") a meaningful name; we will call it naming map. So knowledge representation is a composition of information hiding (a map from a granular structure to a quotient structure) and the naming map.

Knowledge representation is another way to discover new knowledge by organizing the knowledge structure in an appropriate fashion, such as, an information table (= a relation in database theory). In GrC, the table often has additional algebraic and/or topological structures [18,22].

4) *Concept Approximation*: Among three semantic views, the knowledge engineering view is the most suitable view for this operation. Under this view, concept approximation is to express approximately the unknown concepts (arbitrary subsets of the universe) in terms of known basic knowledges (granules).

It is reasonable to regard that "and" ($\cap$) or "or" ($\cup$) operations of two basic units of known knowledge are also a known knowledge. So, we take finite intersection and any number of union as acceptable knowledge operations; the last one is derived from the topological spaces. However, we believe a negation of a known knowledge is not necessary a piece of known knowledge; so negation is not an acceptable operation.

Note that this approximation is different from *Rough Set Approximations*, which, including its generalizations, are based on the sole operation "or" ($\cup$). In other words, RST does not regard the "and" of two known concepts as also a known concept. So strictly speaking, rough set approximation is not a concept approximation.

GrC has many models; each model has a slightly different approximation theory. First, we will explain
The *Second GrC Model (Global GrC Model)*: Let C_1 be a given collection. Let $\mathcal{G}_1$ be the collection of all possible finite intersections of C_1. Then, by definition, the pair (U, β) is a Second GrC Model (Global GrC Model), where $\beta = C_1$. Let G be a variable that varies through the collection $\mathcal{G}_1$, then we define

Definition 10 Three approximations

1. Upper approximation:
   $C[X] = \overline{\beta}[X] = \{p\colon \forall G$, such that, $p \in G \ \ \& \ \ G \cap X \neq \emptyset\}$.
2. Lower approximation:
   $I[X] = \underline{\beta}[X] = \{p\colon \exists \ a \ \ G$, such that, $p \in G \ \ \& \ \ G \subseteq X\}$.
3. Closed set-based upper approximation: [33] used *closed* closure operator. It applies closure operator repeatedly (transfinitely many steps) until the results stop growing. The space is called Frechet(V)-space or (V)-space. $Cl[X] = X \cup C[X] \cup C[C[X]] \cup C[C[C[X]]] \ldots$ (transfinite). For such a closure, it is a closed set.

Theorem 1 *The concept approximation space of the Global GrC Model (Second GrC Model) is a topological space.*

We should also note that under the rough set approximation, this model is not a topological space.

Next, let us consider the approximation theory of
The *First GrC Model (Local GrC Model)*: Before, we proceed, let us examine the classical case, the topological space (U, τ). A subset $N(p) \subseteq U$ is a neighborhood of p, if $N(p)$ contains an open set that contains p. The union of all such open sets is the interior points of $N(p)$, which is the largest open set in $N(p)$; let us denote it by $O(p)$. Observe that $O(p)$ consists of every point that regards $N(p)$ as its neighborhood.

Now, we will generalize this idea to the First GrC Model. Let NS(p) be the neighborhood system at p. Let $G(p)$ be the collection of all finite intersections of all neighborhoods in NS(p). Let G be a variable that varies through $G(p)$.

Definition 11 With such a G, the previous equations given above define the appropriate notions of $C[X]$, $I[X]$, $Cl[X]$ for the First, Third and Fourth GrC Models.

We should caution here that in the Third GrC Model, there is at most one neighborhood at each point, so there is no "true" intersection.

Let $N(p)$ represent an arbitrary neighborhood of NS(p). Let $C_N(p)$, called the center set of $N(p)$, consist of all those points that have $N(p)$ as its neighborhood. (Note that $C_N(p)$ is the generalization of $O(p)$ in TNS).

Now we will observe some harder question: Do the intersections of neighborhoods at distinct points belong to some $G(p)$?

Proposition 4 (Theorem of Intersections in NS)

1. *$N(p) \cap N(q)$ is in $G(p) = G(q)$, iff $C_N(p) \cap C_N(q) \neq \emptyset$.*
2. *$N(p) \cap N(q)$ is not in any $G(p) \forall p$, iff $C_N(p) \cap C_N(q) = \emptyset$.*

If we regard $N(p)$ as a known basic knowledge, then we should define the knowledge operations: Let $^\circ$ be the "and" operation of the basic knowledge (a neighborhood). For technical reasons, the $\emptyset$ is regarded as a piece of the given basic knowledge.

Definition 12 (New operations in NS)

1. $N(p) \circ N(q) = N(p) \cap N(q)$, iff $C_N(p) \cap C_N(q) \neq \emptyset$.
2. $N(p) \circ N(q) = \emptyset$, iff $C_N(p) \cap C_N(q) = \emptyset$.

Observe that BNS is a special cases of NS. So we have:

Definition 13 Let B be a BNS, then

1. $B(p) \circ B(q) = B(p) = B(q)$, iff $C_B(p) = C_B(q)$.
2. $B(p) \circ B(q) = \emptyset$, iff $C_B(p) \cap C_B(q) = \emptyset$. Note that $B(p) \cap B(q)$ may not be empty, but it is not a neighborhood of any point.

Observe that in Binary GrC Model, two basic knowledges are either the same or the set theoretical intersection does not represent any basic known knowledge.

5) Higher Order Concept Approximations

In the Fifth GrC model, we consider the relations (subsets of product space) as basic knowledge. Any subset in a product space is a new unknown concept. We will illustrate the idea in the following case: U^j is either a copy of V or U. Moreover, in each product space, there is at most one copy of V, but no restrictions on the number of copies of U. If a Cartesian product has no V component, it is called a U-product space. If there is one and only one copy of V, it is called a product space with unique V.

1. u and u^1 are said to be directly related, if u and u^1 are in the same tuple (of a relation in β), where $u \in U$ and u^1 could be an element of U or V.
2. u and u^2 is said to be indirectly related, if there is a finite sequence $u_i, i = 1, 2, \ldots, t$ such that (1) u_i and u_{i+1} are directly related for every i, and (2) $u = u_1$ and $u^2 = u_t$.
3. An element $u \in U$ is said to be v-related ($v \in V$), if u and v are directly or indirectly related.
4. v-neighborhood, U_v, consists of all the $u \in U$ that are v-related.

In such a relational GrC model (U^j with unique V) induces a map:

$$B\colon V \to 2^U; v \to U_v\ .$$

Such a map defines a binary neighborhood system (BNS), where U_v is a v-neighborhood in U, and hence induces a *binary GrC model* (U, V, B). Next, we will consider the case $U = V$ and define

Definition 14 The high order approximations of the Fifth GrC model are the approximations based on the v-neighborhood system.

6) Concept Approximations in other Categories

In a category of functions, we will be interested in those granular structures that have the universal approximation property. For a category of Turing machines (algorithms), it is still unclear how to define the concept approximations.

Future Directions

Granular Computing is still in its inception stage; possible directions are wide open. Here we will focus only on those issues that are touched in this article.

1) Developments of Categories
In this paper, a category based model is proposed as *the* Formal Model for GrC. It can be specialized into various models to realize all the classical examples, including the first example, the granulation of the human body. We should note that the claim on the realization of the first example is not in printing form yet. However, the author feels that it is important to inform the readers that is occurring.
The key to realize the first example is based on the category of qualitative fuzzy sets or sofsets (this is not a typo); please watch for new development. The categories of functions, random variables (measurable functions) and Turing machines must be developed, too.
2) Developments of Granular Structures
Given a granular structure, we associate it with four structures (including itself). Among them, quotient and knowledge structures are mathematical consequences of a granular structure (if it is given mathematically). However, the linguistic structure is not a mathematical formalism but is a natural language formulation. In this paper, there is no report on this direction. We urge the readers to read Zadeh's article.
3) Imported Concepts
For information integration (this may correspond to Zadeh's term, "organization"), we have illustrated the idea imported from homological algebra. It is unclear if we have imported the correct thinking; but it does point out essential problems in granulate and conquer.
4) GrC and RST
RST has been served as the "model" of GrC developments. So, even there are a lot of similarities here, we would like to caution the readers that there are fundamental differences. For example, the fundamental views of uncertainty are quite different; Pawlak used "unable to specify" as the base of uncertainty, while GrC regards a granule as a unit of uncertainty (such as

uncertainty in quantum mechanics) Also the approximation theories are different. Of course, there are other differences; we skip.

5) GrC, Databases and Data Mining
As we have pointed out that the categorical structures of databases and GrC are similar; at the same time, we need to point out the differences in semantics. Nevertheless, we are looking forward to the transfer of database technology to GrC. For data mining, please see the database section on the articles by this author on deductive data mining using GrC, and mining decision rules using RST.
6) GrC and Fuzzy Logic
Most of the expositions have been based on classical sets (and fuzzifiable concepts). For more intrinsic fuzzy view, we strongly recommend the readers to read Zadeh's article.
7) GrC and Clouding Computing
Theoretically, cloud computing can be related to the GrC on the category of Turing machines. We expect some strong interactions in near future.

As we have observed that GrC is deeply rooted in human thinking, we expect GrC will have many interactions with wide variety of areas.

Bibliography

1. Baliga P, Lin TY (2005) Kolmogorov complexity based automata modeling for intrusion detection. In: 2005 IEEE international conference on granular computing, Beijing, 25–27 July 2005. pp 387–392
2. Chiang IJ, Lin TY, Liu Y (2005) Table representations of granulations revisited. In: Proc 10th international conference, RSFDGrC 2005, Regina, Canada, 31 Aug–3 Sept 2005, part I. Lecture notes in computer science, vol 3641. Springer, Berlin, pp 728–737
3. Dubois D, Prade H (1992) Putting rough sets and fuzzy sets together. In: Slowinski R (ed) Intelligent decision support: Handbook of applications and advances of the rough sets theory. Kluwer, Dordrecht, pp 203–232
4. Ginsburg S, Hull R (1981) Ordered attribute domains in the relational model. In: XP2 workshop on relational database theory, June 22–24, Pennsylvania State University, USA
5. Ginsburg S, Hull R (1983) Order dependency in the relational model. Theor Comput Sci 26:149–195
6. Hsiao D, Harary F (1970) A formal system for information retrieval from files. Commun ACM 13(2):67–73 (Corrigenda 13(4))
7. Lin TY (1988) Neighborhood systems and relational database. In: Proc of the 16th ACM annual conference on computer science, Atlanta, 23–25 Feb 1988. p 725
8. Lin TY (1989) Chinese wall security policy – An aggressive model. In: Proc of the 5th aerospace computer security application conference, Tuscon, 4–8 December, pp 286–293
9. Lin TY (1989) Neighborhood systems and approximation in database and knowledge base systems. In: Proc of the 4th international symposium on methodologies of intelligent systems (poster session), Charlotte, 12 Oct 1989. pp 75–86. (Available to the public from the National Technical Information Services, U. S. Departmetnof Commerce, 5285 Port Royal Rd., Springfield, VA 22161. NTIS price codes rited copy: A11 Micorfiche A01. Available to DOE and DOE contractors from the office od scientific and technical information P. O. Box 62, Oak Ridge, TN 37831; prices available from (615) 576-8401)
10. Lin TY (1990) Relational data models and category theory (abstract). In: CSC'90, Proc of the ACM 18th annual computer science conference on cooperation, Sheraton Washington Hotel, Washington DC, 20–22 Feb 1990. p 424
11. Lin TY (1992) Topological and fuzzy rough sets. In: Slowinski R (ed) Decision support by experience – Application of the rough sets theory. Kluwer, Dordrecht, pp 287–304
12. Lin TY (1996) A set theory for soft computing. In: Proceedings of 1996 IEEE international conference on fuzzy systems, New Orleans, 8–11 Sept 1996, pp 1140–1146
13. Lin TY (1997) Neighborhood systems – A qualitative theory for fuzzy and rough sets. In: Wang P (ed) Advances in machine intelligence and soft computing, vol IV. Duke University, Durham, North Carolina, USA, pp 132–155
14. Lin TY (1998) Granular computing on binary relations I: Data mining and neighborhood systems. In: Skoworn A Polkowski L (eds) Rough sets in knowledge discovery. Physica, Heidelberg, pp 107–121
15. Lin TY (1998) Granular computing on binary relations II: Rough set representations and belief functions. In: Skoworn A, Polkowski L (eds) Rough sets in knowledge discovery. Physica, Heidelberg, pp 121–140
16. Lin TY (1999) Granular computing: Fuzzy logic and rough sets. In: Zadeh L, Kacprzyk J (eds) Computing with words in information/intelligent systems. Physica, Heidelberg, pp 183–200
17. Lin TY (2003) Chinese wall security policy models: Information flows and confining trojan horses. In: De Capitani di Vimercati S, Ray I, Ray I (eds) Data and applications security XVII: Status and prospects, IFIP TC-11 WG 11.3 17th annual working conference on data and application security, Estes Park, 4–6 Aug 2003. Kluwer, Boston, pp 275–287
18. Lin TY (2005) Divide and conquer in granular computing: Topological partition. In: Proc of the 2005 North American fuzzy information processing society annual conference: Computing for real world applications, Ann Arbor, 22–25 June 2005
19. Lin TY (2006) A roadmap from rough set theory to granular computing. In: Proc of the 1st international conference rough sets and knowledge technology, RSKT 2006, Chongquing, 24–26 July 2006. Lecture notes in computer science, vol 4062. Springer, Berlin, pp 33–41
20. Lin TY (2006) Granular computing on partitions, coverings, and neighborhood systems. J Nanchang Inst Technol 25(2):1–7 (Special issue as the Proc of the international forum on theory of GrC from rough set prospective.)
21. Lin TY, Chiang IJ (2005) A simplicial complex, a hypergraph, structure in the latent semantic space of document clustering. Int J Approx Reason 40(1–2):55–80
22. Lin TY, Liau CJ (2005) Granular Computing and Rough Sets. In: Maimon O, Rokach L (eds) The data mining and knowledge discovery handbook. Springer, New York, pp 535–561
23. Lin TY, Huang KJ, Liu Q, Chen W (1990) Rough sets, neighborhood systems and approximation. In: Proc of the 5th international symposium on methodologies of intelligent systems, se-

lected papers, Knoxville, 25–27 Oct 1990, pp 130–141. Library of Congress Catalog Card Number:90-84690
24. Lin TY, Sutojo A, Hsu JD (2006) Concept analysis and web clustering using combinatorial topology. In: Workshops Proc of the 6th IEEE international conference on data mining (ICDM 2006), Hong Kong, 18–22 Dec 2006. pp 412–416
25. Liu Y, Lin TY, Xu C, Zhang Q, Huang L, He P (2008) Modeling complex architectures based on granular computing on ontology. Submitted to IEEE Trans Fuzzy Syst
26. Maclane S (1995) Homology. Classics in mathematics, reprint of the 1975 edn. Springer, Berlin, x+422 pp
27. Mumford D (1988) Introduction ot algebriac geometry. In: The red book of varieties and schemes, mimeographed notes from the Harvard mathematics department 1967. Lecture Notes in Mathematics, vol 1348. Springer, Berlin
28. Munkres J (2000) Topology, 2nd edn. Prentice-Hall
29. Park JW, Sandberg IW (1991) Universal approximation using radial-basis-function networks. Neural Comput 3:246–257
30. Pawlak Z (1991) Rough sets. Theoretical Aspects Of Reasoning About Data. Kluwer, Dordrecht
31. Polya G (1957) How to solve it, 2nd edn. Princeton University Press, Princeton
32. Robinson A (1966) Non-standard analysis. North-Holland, Amsterdam
33. Sierpenski W (trans by Krieger C) (1952) General Topology. Mathematical exposition, no 7. University of Toronto Press, Toronto
34. Spanier EH (1966) Algebraic topology. Springer, New York
35. Stanat D, McAllister D (1977) Discrete mathematics in computer science. Prentice-Hall, Englewood Cliffs
36. Wong E, Chiang TC (1971) Canonical structure in attribute based file organization. Commun ACM 14(9):593–597
37. Zadeh LA (1979) Fuzzy sets and information granularity. In: Gupta M, Ragade R, Yager R (eds) Advances in fuzzy set theory and applications. North-Holland, Amsterdam, pp 3–18
38. Zadeh LA (1996) The key roles of information granulation and fuzzy logic in human reasoning. In: 1996 IEEE international conference on fuzzy systems, New Orleans, 8–11 Sept, p 1
39. Zadeh LA (1997) Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets Syst 90:111–127
40. Zadeh LA (1998) Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems. Soft Comput 2:23–25
41. Zimmerman H (1991) Fuzzy set theory – and its applications. Kluwer, Dordrecht

Granular Computing, Principles and Perspectives of

Jianchao Han[1], Nick Cercone[2]
[1] Department of Computer Science, California State University, Dominguez Hills, Carson, USA
[2] Faculty of Science and Engineering, York University, Toronto, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
Neighborhood System: A Mathematical Structure of Granular Computing
Rough Set System: An Equivalence Neighborhood System
Fuzzy Set System: A Fuzzy Neighborhood System
Granular Computing in Data Mining
Granular Computing in Software Engineering
Future Directions
Bibliography

Glossary

Granular computing Granular computing is a strategy of problem solving. The basic idea of granular computing comes from the strategy of divide-and-conquer. It is a twofold process: First it granulates a complex program into granules and then it computes these granules and integrates results to form a solution to the complex problem. Granulation of problems into granules is of different forms such as chucking, clustering, partitioning, division, or decomposition, while granules are clumps of objects or points. Computations with granules are either within granules or granule with environment.

Neighborhood system Neighborhood system is a mathematical structure of granular computing to model granules, and can be used to compute structure of granules and/or between granules and ambient spaces. A neighborhood system at a point is a framework to capture the concept of "near" objects, and any subset of objects can be approximated by a set of neighborhoods. A neighborhood system defines a set of binary relations, and a set of binary relationships can be used to define a neighborhood system.

Fuzzy set theory Unlike the classic set theory where a set is represented as an indicator function to specify if an object belongs or not to it, a fuzzy set is an extension of a classic set where a subset is represented as a membership function to characterize the degree that an object belongs to it. The indicator function of a classic set takes value of 1 or 0, whereas the membership function of a fuzzy set takes value between 1 and 0.

Rough set theory The rough set theory deals with inexact information systems. In an information system, a decision table consists of a set of objects which are characterized by a set of condition attributes and decision at-

tributes. Objects in the decision table can be classified into equivalence classes using an *indiscernibility* relation, and equivalence classes are explored to approximate crisp sets of objects.

Data mining Data mining is a very important step of knowledge discovery in databases to extract nontrivial, previously unknown, and potentially useful patterns that are hidden from large data sets. Data mining tasks include classification and clustering analysis of objects into categories, discovery of associations and correlations among data items, characterization and summarization of subsets of objects, finding sequential patterns and similarities in ordered data, etc.

Software engineering Software engineering is an engineering discipline that applies a systematic approach to produce reliable and efficient software. Software development process consists of different phases, including requirements, analysis, design, specification, implementation, testing, deployment, and maintenance. Object-oriented methodology to software engineering is based on several techniques and principles such as inheritance, modularity, polymorphism and encapsulation.

Definition of the Subject

Granular computing (abbreviated, GrC) is a general computing paradigm that effectively deals with elements and granules that are generated from elements. Granules are vaguely viewed as generalized subsets of the universe of discourse and may be in different forms such as classes, clusters, groups, and intervals. The essence of GrC is to build an efficient computational model for complex applications with huge amounts of data, information and knowledge. Though the term has been proposed recently, the basic ideas and principles of granular computing have been explored in a variety of fields under the different names such as divide-and-conquer, and have been extensively applied in many problem solving disciplines such as politics, sociology, economics, computational intelligence with different forms either consciously or unconsciously by human beings.

Research on GrC will not only reveal the computational structures of complex problems and systems, build the mathematical foundations of complex problem solving, and formalize the common problem solving strategy, but also help practitioners explore novel specific problem solving strategies in special application domains. Additionally, by consciously using GrC strategies and heuristics, human being has a better chance to find a good solution to a complex problem.

Introduction

The terminology of granular computing (GrC) was first proposed by Professor T. Y. Lin in 1996 as a label of family of theories, methodologies, and techniques in computational intelligence that make use of granules, although this strategy can be dated back to the ancient time. Its basic ideas and principles have been studied in various application domains. Especially in the form of partitions, the theory has been accumulated for hundred of years in mathematics.

Before the term, granular computing, was invented, some results and thoughts on granular computing had been achieved. The explicit study of granular computing is originated from Zadeh. In 1979, Zadeh [25] introduced the notion of information granulation and suggested that fuzzy set theory might find potential applications in this regard.

Though the granular computing is focusing on the non-partition theories, nevertheless, the partition case was the main source of inspiration, especially the partition theory in computer sciences. In 1982, Pawlak [21] proposed Rough Set Theory (RST) to deal with inexact information. It is an uncertainty theory using a very special form of granules, called equivalence classes. It is primarily due to the rough set theory (partition theory) that causes researchers to realize the importance of generalized granular computing notion.

In 1985, Hobbes [7] presented a theory of granularity as the base of knowledge representation, abstraction, heuristic search, and reasoning. In his theory the problem world is represented as various grains and only interesting ones are abstracted to learn concepts. The conceptualization of the world can be performed at different granularities and switched between granularities. Even though his discussions mainly concentrated on the partition cases, his model is more general than rough sets, and includes reflexive and symmetric binary relations.

In 1988, from the approximation retrieval, Lin introduced the notion of the neighborhood systems (NS) as a model of uncertainty in database systems [10,11], where a neighborhood is a unit of uncertainty. Lin studied the mathematics foundation of the neighborhood system that originates from the perspective of topology. A topological neighborhood system attaches to every point a collection of subsets that satisfy a set of axioms, called the axioms of topology. Lin removed the axioms and extended the theory to a quite general notion [12,13,14], where a neighborhood system corresponds to a set of binary relationships.

In 1992, Giunchigalia and Walsh [4] presented a theory of abstraction to improve the conceptualization of

granularities, where granules are abstracted in hierarchical levels so that relevant elements/attributes can be extracted but irrelevant details are ignored.

In 1997, Zadeh [26,27] refined his information granulation, presented granular mathematics, and applied to word computation. About the same time, Lin used the term "granular computing" to label this growing research field (1996). Since then, granular computing has received more and more attentions and much research has been conducted in various aspects of this area, and has begun to play important roles in various fields, including machine learning and data mining, bioinformatics, e-Business, network security, high-performance computing and wireless mobile computing, information hiding, and social works [3,17,18,22,23,27]. International conferences on Granular Computing sponsored by IEEE Computational Intelligent Society have been annually held since 2005 [8,28]. The essence of these models and applications has been addressed by researchers to build efficient computational algorithms for handling huge amounts of data, information and knowledge. The objectives of these computation models are computer-centered and mainly concern the efficiency, effectiveness, and robustness of using granules such as classes, clusters, subsets, groups and intervals in problem solving. In recent years, some researchers have investigated the granular computing paradigm from perspectives of philosophy, cognitive science, and human thinking as well as the general strategies of interactions between granules and operations on granule coverings.

Generally speaking, granular computing is a twofold process: Granulation and computation, where the former transforms the problem domain to one with granules, whereas the latter computes these granules to solve the problem. Granulation is a very natural concept and appears almost everywhere in different names, such as chucking, clustering, partitioning, division, decomposition, just to name a few. According to Zadeh [25] and Lin [13],

> "information granulation is a collection of granules, with a granule being a clump of objects (points) which are drawn towards an object. In other words, each object is associated a family of clumps."

Up to now, modeling and applying general GrC is not as successful as desired, although some special models have been innovated with great success. So it is immature at this point to speculate what the general definition, principle and methodology of granular computing should be. Therefore, we will look at some concrete theories that may be informally called granular computing. The discussion will start from the neighborhood system and extend to the fuzzy set system, the rough set theory, and the quotient space. Then, the perspectives of granular computing in data mining and software engineering will be introduced, including granularity and granulation, granular relationships, and computation with granules.

Neighborhood System: A Mathematical Structure of Granular Computing

According to Lin [12], granular computing can be mathematically structured as a neighborhood system, which is abstracted from the geometric notion of "near" or "negligible distance". Roughly, a neighborhood system assigns each object in the universe a (possibly empty, finite, or infinite) family of non-empty subsets, called neighborhoods, to represent the semantics of "near". Neighborhoods play the most fundamental role in granular computing and can be considered as "granules". A neighborhood is precisely a topological term of "granules", and a neighborhood system is vaguely a topological space.

Neighborhood System

Fundamental notions on neighborhood systems were presented by Lin [12,13,14]. Assume U is the universe of discourse and p is an object in U.

1. An object x is a neighbor of p if x is "near" to p. The notion "near" is a subjective judgment and can be semantically interpreted in different ways, for example, if they satisfy some relationship or condition. For a specific case, let's consider a binary relationship R in U. x is "near" to p, if $x\mathrm{R}p$.
2. A neighborhood of p, denoted by $\mathrm{N}(p)$, in U is a non-empty subset of U, which may or may not contain p. A neighborhood system of p, denoted as $\mathrm{NS}(p)$, in U is a family of neighborhoods of p. If p has no neighborhoods, then $NS(p)$ is an empty family.
3. A neighborhood system of U, denoted by $NS(U)$, is the collection of $NS(p)$ for $\forall p \in U$. Formally, $NS(U) = \{NS(p) | \forall p \in U\}$.
4. Consider two neighborhood systems of U, say $NS_1(U)$ and $NS_2(U)$. $NS_1(U)$ is said a refinement of $NS_2(U)$, if for any neighborhood N_1 in $NS_1(U)$, there exists a neighborhood N_2 in $NS_2(U)$ such that $N_1 \subseteq N_2$. Similarly, $NS_2(U)$ is said a weak coarsening of $NS_1(U)$. (see [13] for strong refinement and coarsening).

From above, one can see, a neighborhood system is induced from the fundamental neighbors of objects, which are imprecisely defined as "near" to. If the "near" is explicitly defined according to a single binary relation, the result

neighborhood system is called a binary neighborhood system.

A binary neighborhood system BNS is defined as $\text{BNS} = \langle U, \text{B}, V \rangle$, where U and V are two universes, and B is a mapping function from U to V.

$$B\colon U \to V\ .$$

For $\forall v \in V$, let B_v denote the subset of U, that are source of v under the map B, that is,

$$B_v = \{u | u \in U \ \text{ and } \ B(u) = v\}\ .$$

Each B_v is called a basic neighborhood or a binary neighborhood.

One can infer that each neighborhood system $\text{BNS} = \langle U, \text{B}, V \rangle$ defines a binary relation $R \subseteq U \times V$: $\forall v \in V$ and $u \in U$, uRv iff $u \in B_v \subseteq U$. Thus the binary relation R is fully determined by B. Conversely B is fully determined by a binary relation R.

If B is a binary relation $R \subseteq U \times V$, then for each object $v \in V$, the binary neighborhood in terms of v will be

$$N_v = \{u | u \in U \ \text{ and } \ uRv\}\ .$$

This is a subset of U, whose all elements are related to v by R.

A binary relation, in the usual sense, is a very special form of a neighborhood system, where each object has either zero or one neighborhood.

If $U = V$ and the mapping function B is considered as a binary relation, then we have a simple binary neighborhood system, denoted as $\text{BNS} = \langle U, R \rangle$.

For simplicity, we call this binary neighborhood system with a binary relation as binary relation neighborhood system (BRNS).

Let's consider two binary relation neighborhood systems BRNS1 and BRNS2 built on the same universes, where $\text{BRNS1} = \langle U, R1, V \rangle$ and $\text{BRNS2} = \langle U, R2, V \rangle$. BRNS1 is said a weak refinement of BRNS2, if $R2 \subseteq R1$, and a strong refinement if every neighborhood of BNS1 is a union of neighborhoods of BNS2. On the other hand, BRNS1 is said a weak/strong coarsening of BRNS2, if $R1 \subseteq R2$.

Approximation

Assume $\text{BNS} = \langle U, \text{B}, V \rangle$ is a binary neighborhood system and $X \subseteq U$. X can be approximated by neighborhoods of BNS.

Following topological spaces, the interior approximation, interior for short, of X is defined as either

$$\text{Interior}(X) = \cup_{x \in X} \{N(x) | N(x) \subseteq X\}\ ,$$

or

$$\text{Interior}(X) = \{x | x \in U \ \text{ and } \ \exists N(x) \subseteq X\}\ .$$

Similarly, the exterior approximation, exterior for short, of X can be defined as either

$$\text{Exterior}(X) = \cup_{x \in X} \{N(x) | N(x) \cap X \neq \emptyset\}\ ,$$

or

$$\text{Exterior}(X) = \{x | x \in U \ \text{ and } \ \forall N(x) \cap X \neq \emptyset\}\ .$$

The interior of a subset X of the universe can be approximated by the union of all neighborhoods of the universe that are completely contained in the subset, or by a subset of the universe that consists of all objects that have at least one neighborhood contained in X; while the exterior of a subset X can be approximated by the union of all neighborhoods that have non-empty intersection with X, or by a subset of the universe that consists of all objects whose all neighborhoods have non-empty intersection with X.

A subset X of U is open if $\forall p \in X$, there is a neighborhood $N(p) \subseteq X$. X is closed, if its complement is open. A neighborhood system of p, $NS(p)$, is open, if all neighborhoods of p are open. A neighborhood system of U, $\text{NS}(U)$, is open, if $\forall p \in U$, $\text{NS}(p)$ is open.

If the universe U is a topological space and a neighborhood system of U, $\text{NS}(U)$, is open, then $NS(U)$ also defines the same topological space on U. In such a case, both $NS(U)$ and the collection of open sets are called topology.

An object p is a limit point of a set E, if every neighborhood of p contains a point of E other than p. The set of all limit points of E is called derived set. E together with its derived set is a closed set.

One can easily verify that, $NS(U)$ is discrete if every $NS(p)$ is singleton and $NS(U)$ is indiscrete if $NS(U)$ is a singleton $\{U\}$.

A covering of U is an open neighborhood system $NS(U)$.

A partition of U is a covering of U, where $\forall p, q \in U$, and $p \neq q$, either $NS(p) = NS(q)$, or $NS(p) \cap NS(q) = \emptyset$.

Rough Set System: An Equivalence Neighborhood System

Let's consider a special case of the binary neighborhood system $\text{BNS} = \langle U, R \rangle$, where R is an equivalence binary relation on $U \times U$. For an element $u \in U$, by definition, the binary neighborhood of u is defined as

$$N_u = \{x | x \in U, \ \text{ and } \ xRu\}\ .$$

Since R is an equivalence relation, N_u is the equivalence class $[u]$ induced by u with respect to R. Thus, the neigh-

borhood system (the collection of all neighborhoods) will be the family of equivalence classes induced by R.

Partition

Rough set theory was proposed by Pawlak in 1982 to deal with inexact information by using rough sets to approximate a crisp set [21]. By investigating the granularity of knowledge from the point of view of the rough set theory, one can view the rough set theory as a granular computing model based on partitions [22]. Basically, suppose U is a finite and nonempty universe. Let $R \subseteq U \times U$ be an equivalence relation on U. The pair $\langle U, R\rangle$ is called an approximation space. The relation R is a special BNS; it can be conveniently represented by a mapping B_R from U to the power set of U, and defined as

$$B_R : U \to 2^U ,$$

$$B_R(x) = [x]_R = \{y \in U | xRy\} , \quad \forall x \in U .$$

The subset $[x]_R$ is the equivalence class containing x, and is called an elementary subset. The family of all equivalent classes, denoted by $U/R = \{[x]_R | \ x \in U\}$, defines a partition of the universe U, namely, a family of pairwise disjoint subsets whose union covers the whole universe.

From the perspective of the rough set theory, each equivalence class is considered as a whole granule instead of many individuals. With these elementary subsets, any subset of U can be approximated.

Approximation

Let X be a subset of U and R be an equivalence relation on U. The lower approximation of X based on R, denoted by $\mathrm{Low}_R(X)$, is defined as

$$\mathrm{Low}_R(X) = \cup\{Y \in U/R | Y \subseteq X\} ,$$

which contains all elementary subsets of U that are completely included in X.

The upper approximation of X based on R, denoted by $\mathrm{Upp}_R(X)$, is defined as

$$\mathrm{Upp}_R(X) = \cup\{Y \in U/R | Y \cap X \neq \emptyset\} ,$$

which contains all elementary subsets of U that have nonempty intersection with X.

In the partition model, these forms of approximations are equivalent to the interior and exterior defined in the BNS, respectively. However, they are not appropriate to be generalized to the *NS*. For example, in the topological space, $\mathrm{Upp}_R(X)$ is always the whole space, no matter what you choose.

Information Systems

The rough set theory has found its applications in data mining, especially in classification and decision-making problem domains, called information systems.

An information system IS is defined as: $\mathrm{IS} = \langle U, C, D, \{V_a\}_{a \in C \cup D}\}, f\rangle$, where $U = \{u_1, u_2, \ldots, u_n\}$ is a non-empty set of objects (tuples), called data set or decision table, C is a non-empty set of condition attributes, and D is a non-empty set of decision attributes and $C \cap D = \emptyset$. Va is the domain of attribute a with at least two elements. f is a function: $U \times (C \cup D) \to V = \cup_{a \in C \cup D} V_a$, which maps each pair of object and attribute to an attribute value.

Let $A \subseteq C \cup D$, and $t, s \in U$. A binary relation R_A, called an *indiscernibility* relation, is defined as follows: $R_A = \{\langle t, s\rangle \in U \times U | \ \forall a \in A, t[a] = s[a]\}$, where $t[a]$ indicates the value of attribute $a \in A$ of object t. The *indiscernibility* relation, denoted by $\mathrm{IND}(A)$, is an equivalence relation on U. With this equivalence relation R_A, one can construct a binary relation neighborhood system $\langle U, \mathrm{IND}(A)\rangle$.

Let X be a subset of U in an information system and represent a concept. X can be lower-approximated and upper-approximated by finding its lower approximation and upper approximation using elementary subsets of $U/\mathrm{IND}(A)$. The lower approximation of X contains all objects in U which are *definitely* included in X, while the upper approximation of X contains all objects in U which are *potentially* included in X. On the other hand, one can see that the complement of the upper approximation of X contains all objects in U which are definitely excluded by X. As a concept, X has its lower approximation as the positive region, whereas the complement of its upper approximation as its negative region. From the perspective of machine learning or classification, the positive region contains the positive examples of X, while the negative region contains the negative examples of X. Between the positive region and the negative region is called the boundary region of X.

One main application of concept approximation using granules in the rough set theory is to reduce the size of a large decision table by finding an attribute reduct. A condition attribute $a \in C$ is a core attribute of C in U with respect to D if

$$\forall X \in U/\mathrm{IND}(D), \ \ \mathrm{Low}_{\mathrm{IND}(C)}(X) \neq \mathrm{Low}_{\mathrm{IND}(C-\{a\})}(X).$$

A reduct of the condition attribute set is a minimum subset of it that has the same classification capability as itself. Thus a reduct of the condition attribute set can be used to represent the entire condition attribute set.

Formally, a subset A of C, $A \subseteq C$, is defined as a reduct of C in U with respect to D if

$$\forall X \in U/\mathrm{IND}(D)\,,\ \ \mathrm{Low}_{\mathrm{IND}(A)}(X) = \mathrm{Low}_{\mathrm{IND}(C)}(X)\,,$$
$$\text{and}\ \ \forall B \subset R\,,\ \ \mathrm{Low}_{\mathrm{IND}(B)}(X) \neq \mathrm{Low}_{\mathrm{IND}(C)}(X)\,.$$

A condition attribute $a \in C$ is said to be a reduct attribute if $\exists B \subseteq C$, B is a reduct of C and $a \in B$. A reduct R of C is called a minimum reduct of C if $\forall Q \subset R$, Q is not a reduct of C.

Assume $P \subseteq C \cup D$ and $Q \subseteq C \cup D$, the positive region of Q with respect to P, denoted $\mathrm{POS}_P(Q)$, is defined as

$$\mathrm{POS}_P(Q) = \cup_{X \in U/\mathrm{IND}(Q)}\ \mathrm{Low}_{\mathrm{IND}(P)}(X)\,,$$

which contains all objects in U that can be classified using the information contained in P. With this definition, the degree of dependency of Q from P, denoted $\gamma_P(Q)$, is defined as $\gamma_P(Q) = \mathrm{POS}_P(Q)/|U|$, where $|X|$ denotes the cardinality of the set X.

The degree of attribute dependency provides a measure how an attribute subset is dependent on another attribute subset. $\gamma_P(Q) = 1$ means that Q totally depends on P, $\gamma_P(Q) = 0$ indicates that Q is totally independent from P, while $0 < \gamma_P(Q) < 1$ denotes a partial dependency of Q from P. Particularly, assume $P \subset C$ and $Q = D$, then $\gamma_P(D)$ can be used to measure the dependency of the decision attributes on a subset of conditional attributes. The task of the rough set attribute reduction is to find a subset of the conditional attributes that functions as the original conditional attribute set without loss of classification capability. This subset of the conditional attribute set is called *reduct*. One can prove that $R \subseteq C$ is a *reduct* of C, if and only if $\mathrm{POS}_R(D) = \mathrm{POS}_C(D)$, or equivalently, $\gamma_R(D) = \gamma_C(D)$.

Thus, an attribute reduct is such a subset of condition attributes that the decision attribute has the same dependency degree on it as that on the entire set of condition attributes, and no attribute can be eliminated from it without affecting the dependency degree. Given an information system, however, for a given decision table, there may exist more than one reduct. Each reduct can be used as an alternative group of attributes to represent the original information system. It has been proved that every reduct must contain all the core attributes. However, finding all reducts of the condition attribute set is unfortunately *NP*-hard.

Quotient Space

There multiple reducts of condition attributes in an information system may exist. Each reduct can be used to define an equivalence relation. Thus it can be applied to construct a quotient space. From the perspectives of mathematics and topology, a quotient space is a space where equivalent points are glued together, and therefore a new space can be constructed.

Assume the universe of discourse U is a topological space, R is an equivalence relation on U. U/R, the set of all equivalence classes of R, is called a quotient space, where the topology on U/R is defined as below:

> A subset $X \subseteq U/R$ is open if and only if their union is open .

Topological properties of a quotient space as well as relationships between quotient spaces have been studied to identify and select attributes to form high-quality reducts. The quotient space theory has also been explored to solve information hiding, network security, and other problems [29].

Fuzzy Set System: A Fuzzy Neighborhood System

In this section, another special case of binary neighborhood system is considered. Assume a binary neighborhood system $\mathrm{BNS} = \langle U, \mathrm{B}, V\rangle$, where U is a universe of discourse (the data space), V is a concept space, and B is a mapping function from $U \times V$ to the unit interval. Formally,

$$B\colon U \times V \to [0, 1]\,.$$

Since each point in the concept space is actually a concept, and can be assigned a name, Lin [14] suggested that this neighborhood system be defined as a 4-tuple system $\langle U, \mathrm{B}, V, C\rangle$, where U, B, and V are the same as before, while C is the concept name space. This system is called a fuzzy granular structure, and each concept in the concept space is a fundamental fuzzy clump or fuzzy neighborhood. In the simple case, B is a fuzzy binary relation and thus the 4-tuple system above is a single level granulation, called a fuzzy binary granular structure.

Information Granularity

Information granularity with Fuzzy set was proposed by L. A. Zadeh in 1979 [25] to provide a basis for the construction of more general theories in which the evidence is allowed to be fuzzy in nature. In the information granularity, the data granules are viewed as a proposition in the general form of

$$g\colon \text{“}X\ \text{is}\ G\text{”}\ \text{is}\ \lambda\,,$$

where X is a variable taking values in a universe of discourse U, G is a fuzzy set of the universe and is characterized by its membership function μ_G, and λ is a fuzzy probability characterized by a possibility distribution over an unit interval.

A typical example of such a proposition is given in the universe U of real numbers, as below:

$$g\colon \text{“}X \text{ is small”} \text{ is likely},$$

where the fuzzy subset G is “small” and the fuzzy probability λ is a fuzzy subset of the unit interval.

Each fuzzy subset G of the universe can be generally understood as a concept, for example, “small”, “very large”, “old”, “young”, etc. All concepts are put together to form a concept space, denoted by C,

$$C = \{G_1, G_2, \ldots, G_N\}.$$

Each proposition g as above is considered as an evidence, and all evidences compromise a collection of propositions,

$$V = \{g_1, g_2, \ldots, g_N\}, \quad \text{where } g_i\colon \text{“}X \text{ is } G_i\text{”} \text{ is } \lambda_i, \quad i = 1, 2, \ldots, N.$$

The proceeding information granularity can be characterized as a neighborhood system. Consider a mapping function B from $U \times V$ to [0,1], defined as:

$$B\colon U \times V \to [0, 1],$$
$$B(x, g) = \mu_\lambda(\mu_{G(g)}(x)), \quad \forall x \in U \text{ and } \forall v \in V.$$

where $\mu_{G(g)}$ is the membership function of the concept G in the proposition g. Thus the 4-tuple $\langle U, \mathrm{B}, V, C\rangle$ forms a neighborhood system in the form of fuzzy binary granular structure.

A special case of fuzzy granules is called non-probability-qualified granule and in the form of “X is G”. In this case, the corresponding fuzzy neighborhood system can be re-written as a 4-tuple system $\mathrm{FNS} = \langle U, \mathrm{B}, V, C\rangle$, where U is the universe of discourse (data space), V is a family of fuzzy subsets of U, C is the name space whose names correspond to the fuzzy subsets in V, and B is a mapping function:

$$B\colon U \times V \to [0, 1],$$
$$B(x, v) = \mu_v(x), \quad \forall x \in U \text{ and } \forall v \in V.$$

Consider the following example: $U = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, $V = \{V_1, V_2, V_3, V_4\}$, where (using discrete fuzzy subset notation)

$$V_1 = 1.0/0 + 0.67/1 + 0.33/2 + 0/3 + 0/4 + 0/5 + 0/6 + 0/7 + 0/8 + 0/9,$$
$$V_2 = 0/0 + 0.33/1 + 0.67/2 + 1.0/3 + 0.67/4 + 0.33/5 + 0/6 + 0/7 + 0/8 + 0/9,$$
$$V_3 = 0/0 + 0/1 + 0/2 + 0/3 + 0.33/4 + 0.67/5 + 1.0/6 + 0.67/7 + 0.33/8 + 0/9,$$
$$V_4 = 0/0 + 0/1 + 0/2 + 0/3 + 0/4 + 0/5 + 0/6 + 0.33/7 + 0.67/8 + 1.0/9,$$

and the collection of words $C = \{\text{Low}, \text{Median-Low}, \text{Median-High}, \text{High}\}$ to name fuzzy subsets V_1, V_2, V_3, and V_4.

Conditional granules can be represented as fuzzy rules such as “If $X = u$ Then Y is G”. These kinds of granules can also be represented as fuzzy set membership functions or expressions of these functions [14].

Approximation

In a fuzzy neighborhood system $\mathrm{FBS} = \langle U, \mathrm{B}, V, C\rangle$, the concept space V (a family of fuzzy subsets of U) is also called a fuzzy covering of U, and each v in V is called a cover of U. For each $u \in U$, $v \in V$ is called a fundamental fuzzy neighborhood of u, if $B(u, v) = \mu_v(u) > 0$. Since each object u in the data space U may have multiple fundamental fuzzy neighborhoods, u can be represented by a vector of these fundamental fuzzy neighborhoods, called multi-valued representation for multi-level granulation. All fundamental fuzzy neighborhoods of u, put together, is denoted by $\mathrm{FNS}(u)$.

Given a fuzzy neighborhood system $\mathrm{FBS} = \langle U, \mathrm{B}, V, C\rangle$, a subset X of U can be approximated by a membership function. The minimum approximation $\min(X)$ of X in FBS can be defined as the following membership function:

$$\mu_X(x) = \min\{\mu_v(x) | v \in \mathrm{FNS}(x)\} \text{ if } x \in X, \text{ and 0 otherwise}.$$

Similarly, the maximum approximation $\max(X)$ of X in FBS can be defined by the following membership function:

$$\mu_X(x) = \max\{\mu_v(x) | v \in \mathrm{FNS}(x)\} \text{ if } x \in X, \text{ and 0 otherwise}.$$

One can see from above that the minimum (maximum) approximation of a subset of U is a fuzzy subset of U that minimizes (maximizes) the membership degrees of its elements to their fuzzy fundamental neighborhoods.

Computing with words is to make inferences from words to words and is a very important topic of information granularity [27]. Fundamental neighborhoods can also be linearly combined to approximate formal words (concept names) [14]. Basically, let $C = \{C_1, C_2, \ldots, C_m\}$ be the concept space of a fuzzy neighborhood system, and $V = \{V_1, V_2, \ldots, V_m\}$ be the corresponding fuzzy subsets of the universe U. Assume $r_1, r_2, \ldots, r_m$ are real numbers. Consider the following linear expression:

$$r_1 C_1 + r_2 C_2 + \ldots + r_m C_m = \sum_{i=1,\ldots,m} r_i C_i \,.$$

Mathematically, the collection of all such expressions forms an abstract vector space, and each vector in the space is called a formal word, corresponding to a new concept. This new concept can be approximated by a fuzzy subset of U, of which the membership function is defined in terms of the membership functions $\mu_{Vi}, i = 1, 2, \ldots, m$, for example, the minimum approximation or the maximum approximation discussed previously. The membership function of the new concept can also be defined as a linear combination of membership functions of its all fundamental fuzzy neighborhoods, i. e.

$$\mu_{\text{New Concept}}(x) = \sum_{i=1,\ldots,m} r_i \mu_{Vi}(x) \,.$$

In above example, a concept Median can be defined in the middle between the concepts Median-Low and Median-High, that is,

$$\begin{aligned}\text{Median} \;&= 0 * \text{Low} + 1/2 * \text{Median-Low} \\ &\quad + 1/2 * \text{Median-High} + 0 * \text{High} \,.\end{aligned}$$

This concept (word) can be approximated by the fuzzy subset defined below:

$$\mu_{\text{Median}} = 1/2 * \mu_{V2} + 1/2 * \mu_{V3} \,,$$

and the fuzzy subset is:

$$\begin{aligned}&0/0 + 0.16/1 + 0.33/2 + 0.5/3 + 0.5/4 + 0.5/5 \\ &\quad + 0.5/6 + 0.33/7 + 0.16/8 + 0/9.\end{aligned}$$

Granular Computing in Data Mining

Granular computing, as a strategy of problem solving and human thinking, has been extensively applied in knowledge discovery and data mining. For example, machine learning is to interpolate data and integrate concepts based on background knowledge that is viewed as granules. In neural network computing, activation functions can also be regarded as granules, and thus the propagation and processing of activation functions can be viewed as a form of computations with granules. A gene is a granule, genetic algorithms process and transform genes which are actually granules, and therefore genetic algorithms can also be considered as one form of granular computing. In this section classification and clustering analysis and especially association mining are reviewed from the perspective of granular computing.

Granular Computing in Classification and Clustering

Classification problem is, given a training set of examples with class labels, to construct a classifier that is able to assign a class label to a new example without class label that is not in the training set [6]. All examples are described with a given set of attributes. On the other hand, clustering is partitioning a data set by the similarity into clusters. Classes in classification and clusters in clustering are granules, subsets of examples, and the granulation is to partition the set of examples into granules. Computation with classes is to induce classification rules to characterize each class and predict the classes of future examples [9]. Similarly, computation with clusters is to summarize or characterize each cluster and discriminate between clusters.

The classification can be based on the refinement and coarsening relationships between granules [24]. A granule g_1 is defined as a refinement of another granule g_2, or equivalently, g_2 is a coarsening of g_1, if every sub-granule or object of g_1 is contained in some sub-granules of g_2. Partitions and coverings are two simple and commonly used granulations of a universe. A partitioning of a universe is a collection of its non-empty and pair-wise disjoint subsets, whose union is the universe. On the other hand, a covering of a universe is a collection of its non-empty subsets that cover the whole universe. A partitioning is a special case of covering.

On the other hand, the clustering can be built upon the similarity relationship [24]. Similarity between granules is a basic and fundamental granular relationship and a key to forming an intrarelationship of a granule. Furthermore, it can be used to measure closeness or nearness amongst granules. Various measures can be exploited to calculate the similarity between two granules, which can be defined as the average distance between individuals in the two granules, and the distance between individuals can be measured as, for example, the Euclidean distance between points, the Levenshtein edit distance between strings or texts, the amino acid distance in biology, and the Mahalanobis squared distance between statistics data items.

Granular Computing in Association Mining

Discovering association rules are another important area of data mining. The basic task of mining association rules is to find frequent itemsets which are granules, and then extract associations or correlations between items in the frequent itemsets, which is the computation with granules.

An association rule is an implication of the form $A \rightarrow B$, where A and B are subsets of regarded attributes, and $A \cap B = \phi$. For different types of attribute values the implication of the association form may have different meanings. The problem of mining association rules can be interpreted from the perspective of granular computing.

Mining association rules in large databases was first presented by Agrawal, Imielinski, and Swami in 1993 [2] to solve the basket data problem, in which, given a large database of customer transactions (basket data) that consist of items purchased by customers, significant associations between items are pursued. For example, "most often transactions that purchase bread and butter also purchase milk", and "a customer purchasing tea is likely to also purchase coffee". These significant associations are described as a set of association rules and quantitatively measured with support and confidence.

Formally, let $I = \{I_1, I_2, \ldots, I_m\}$ be a set of items, each I_k being an item, where m is the number of items. Let D be a database of transactions, where each transaction corresponds to a tuple in D, denoted by T. A transaction T is a set of items represented as a binary vector, with $T[k] = 1$ if T contains the item I_k, and $T[k] = 0$ otherwise, for $1 \leq k \leq m$. Assume $X \subseteq I$ is a subset of items in I. A transaction T satisfies X if for all items $I_k \in X$, $T[k] = 1$.

In the basket data problem, an association rule is of the form $X \rightarrow Y$, where $X, Y \subset I$, and $X \cap Y = \phi$. The rule $X \rightarrow Y$ holds in D with confidence c if at least c% transactions in D that contain X also contain Y, and support s if at least s% transactions in D contain $X \cup Y$.

The task of mining association rules in D is to generate above implications with support and confidence greater than or equal to the user-specified support threshold δ and confidence threshold σ respectively. Confidence is a measure of the rule strength, while support corresponds to statistic significance.

The problem of mining association rules can be decomposed into two subproblems:

1. Discovering all frequent itemsets: A subset X of I is a frequent itemset if its support is at least δ;
2. Generating association rules: For each frequent itemset X, construct implications of the form $X - Y \rightarrow Y$, for all $Y \subset X$. If the confidence of such a form is at least σ, it is considered as an association rule.

Currently many approaches to solving the basket data problem have been developed and efficient algorithms have been implemented. The classical solution exploits level-wise search to generate all candidate frequent itemsets and then prunes them [1]. The level-wise methods are based on the observation that if an itemset is frequent then all of its subsets are frequent. Thus, the basic idea is to evaluate 1-itemsets first, find frequent singleton itemsets, and then evaluate their supersets 2-itemsets, and so forth. Such evaluations are repeated until no frequent itemset is found or the size of the itemsets approaches some threshold.

Louie and Lin reformulate the association relations into bit data model [19] and Lin also reformulates these relations into granular model [15] which has been extended to a theory of attributes (features) [16].

From the perspective of granular computing, association rule mining can be formalized as a binary neighborhood system BNS $= \langle I, B, D \rangle$, where the universe of discourse is the set I of all items, transaction database D (the collection D of tuples) is a subset of the power set of I, that is, $D \subset 2^I$, and the binary relation B can be understood as: For all items i and j, iBj, if and only if there exists an tuple in D that contains both i and j. D forms a covering of I, but not a partition of I.

An itemset $X \subset I$ can be approximated by a subset $C(X)$ of D, where C is defined as

$$C(X) = \{Y | Y \in D\,, \;\; X \subseteq Y\}\,, \;\; \text{for all} \;\; X \subseteq I\,.$$

X is a frequent itemset, if the cardinality of $C(X)$, denoted by $|C(X)|$, is greater than or equal to the support threshold δ. A frequent itemset is called granule. The level-wise search process can be applied to find all granules.

The computation with granules in generating association rules is mainly within granules. For each subset Y of such a granule (a frequent itemset) X, one can find the approximation $C(X)$ and $C(Y)$ of X and Y respectively. The confidence of implication $X - Y \rightarrow Y$ can be calculated as $|C(X)|/|C(Y)|$.

Granular Computing in Software Engineering

In software engineering, the strategies of granular computing have been broadly used in all phases. The Granulate-and-conquer is a softer version of classical divide-and-conquer strategy that employs recursive algorithms to solve problems. A very common technique used in the classical "non-partitioning" recursive call is dynamic programming. Functional decomposition is to partition user requirement into granules (functions). Structured programming is to organize the computer programs as a collection of modules (procedures, functions, routines). The

intrarelationships among these granules are based on their ingredients such as input parameters, output results, executable statements, whereas the interrelationships involve the module interfaces and procedure calls. In object-oriented programming, granules are classes, intrarelationships are the interactions between components of classes such as instance fields, methods and constructors in the Java language; and the interrelationships are class inheritance, aggregation, association, delegation, dependency, etc.

In this section, the different phases of modern software engineering process are investigated from the granular computing perspective, based on object-oriented methodology [20], including requirement analysis, system analysis, design, and implementation.

User Requirements

In the object-oriented methodology, user requirements analysis involves two main aspects: Identifying business actors and use cases. Both aspects can be conducted with granular computing paradigm. However, only identifying use cases will be investigated in this subsection.

Basically, each use case is a snippet of the business and may involve two-way communication between actors and the system. If the business requirement is considered as the problem domain, then the use cases will be the granules. Although there's no a set of rule for deciding how to granulate the business into use cases and most analysts partition the business process into use cases based on common sense, business logic, and their experience. The basic ingredients of the business requirement analysis from the granular computing perspective can be summarized as follows.

Granules: Use cases.

Granulation method: Top–down decomposition for complete covering and bottom–up combination to form a hierarchy of use cases.

Granular relationships: Some relationships between use cases can be recognized. For example, inheritance is a specialization and generalization relationship in which a large case is decomposed into small cases or a set of small cases are combined to constitute a large case. Inheritance is often referred as an is-a relationship: A granule g_1 is a special case of another granule g_2. The is-a relationship plays a very important role in the object-oriented methodology and concept analysis, since "g_1 is-a g_2" reveals that g_1 inherits all features and functions of g_2 [24]. Other relationships between use cases include: Inclusion where the source case has some of its steps provided by the target case, and extension where the source case adds steps to the target case [20].

Granular computation: Identifying the process of use cases, including input data, output data and business logic, as well as the communication with actors.

System Analysis

The basic goal of system analysis is to find candidate classes that describe the objects that might be relevant to the system, relationships between the classes, as well as attributes for the classes. With granular computing terminology, classes are granules that all together should cover the system requirement and satisfy the needs of use cases that have been identified before. The ingredients of system analysis from the granular computing perspective are summarized as follows:

Granules: Classes.

Granulation method is to identify classes. Candidate classes are often indicated by nouns in the use cases except those that represent the system, actors, boundaries, and trivial types. Two basic rules should be followed to identify classes: High cohesion inside classes and low coupling between classes. Class identification might be top–down and/or bottom–up, but the general complete coverage is necessary.

Granular relationships correspond to class relationships, which can include the following types [20]: Inheritance (is-a relationship) where a subclass inherits all of the attributes and behavior of its superclass(es); association, where objects of one class are associated with objects of another class; aggregation (strong association), where an instance of one class is made up of instances of another class; composition (strong aggregation) where the composed object can't be shared by other objects and dies with its composer; and others as well.

Computation with granules is two-fold. First, determine the internal structure and behaviors of objects of classes, including the instance fields and method/constructor prototypes; and second, interface the connections between classes. Associating classes often affects the design of internal structures of classes.

System Design

Basically the software system design is to decompose a system into physical and logical components, and determine the technologies to be used to implement the system. Traditional system design focuses on the system functional partitioning, while modern software system design is based on object-oriented technology. In this subsection,

object-oriented system architecture design and technology are discussed from the perspective of granular computing.

System design involves multiple steps, including system topology, software partitioning, concurrency and security policies as well as communications between components. The following discussion will concentrate in the topology design and system partitioning.

The first task of current object-oriented system design is normally to determine the topology of a networked system. One popular system topology is the three-tier or multi-tier architecture to separate user interfaces, program logic and data in the system. In the three-tier architecture, the client tier presents the user interface to the user so that he/she can enter data and view results; the middle tier – also known as the business logic tier or server tier – runs multi-thread program code using large processors and lots of memory; and the data tier stores the data and provides safe concurrent access to it, typically with the help of a database management system. Thus, the system is granulated into three components. Besides the design of these three tiers, the protocols between tiers are also important.

To granulate the software, thus, object-oriented system design partitions the system components into layers: *User Interface*, *Control*, *Network*, *Server*, *Business*, *Persistence*, and *Database*.

System Implementation

Software system implementation can also be viewed from the perspective of granular computing. The typical program structure in object-oriented software system is actually a granular structure. Consider a Java-based software system, which can be characterized as follows:

Software system
- *Packages*
  - *Classes*
    - *Instance fields*
    - *Constructors*
    - *Methods*
  - *Interfaces*
    - *Method prototypes*

In this structure, a software system is designed as a set of packages, which may be installed in different physical devices. Each package contains a set of interfaces whose subgranules are method prototypes and a set of classes which is partitioned into three types of components: instance fields, constructors, and methods.

During the design of classes, algorithms are one of designer's main concerns. Granular computing is also an important strategy for this purpose. Let's consider the divide-and-conquer technique that is frequently used in algorithm design. The general paradigm of divide-and-conquer [5] is

Divide: *Divide the problem* **S** *in two or more disjoint subsets* $\boldsymbol{S}1, \boldsymbol{S}2, \ldots$

Recur: *Solve the subproblems recursively.*

Conquer: *Combine the solutions to* $\boldsymbol{S}1, \boldsymbol{S}2, \ldots$, into a solution to $\boldsymbol{S}$.

The base cases for the recursion are subproblems of constant size.

In the above paradigm, each subproblem is a granule, which can be granulated further into smaller subproblems. Granulation process usually follows the top-down method and the obtained granules should completely cover the parent problem. The granularity constitutes a hierarchy, where the top is the problem originally given, while at the bottom are all base cases. The computing with granules consists of not only dividing a non-base problem or solving a base problem but also combining solutions to subproblems to form a solution to the parent problem.

Another important general algorithm design paradigm is dynamic programming, which is actually a special case of divide-and-conquer, and thus a granular computing strategy. The main difference between them is: divide-and-conquer solves each subproblems individually, while dynamic programming stores solutions to all subproblems so that when a subproblem is reencountered, the solution can be obtained directly without resolving it.

Future Directions

In this paper, basic principles and models of granular computing were briefly summarized. The ingredients of granular computing were investigated in terms of granulation methods and criteria, granular relationships, and computations with granules as well. Data mining technologies, especially association and correlation mining, and object-oriented technologies were reviewed from the perspective of granular computing. Granular computing methodology has attracted researchers and practitioners from various fields, but the unified architecture and generally accepted framework of granular computing have not been established. It is still an open question what granular computing is and how granular computing is performed. To answer such kind of questions, the future research on granular computing will be focusing on the following directions.

Foundations of granular computing: The mathematical foundation of granular computing will be one of important research topics, which will develop the gran-

ular structures of granules, granular relationships between granules, and computation with granules from the point of view of mathematics. Neighborhood systems, fuzzy set theory, rough set theory, and quotient space theory have initialized this research direction, but from different concerns and computing tasks. Granulation is based on partitioning and covering in terms of the nature of applications. Although it is difficult to propose a general process for granulation and in most applications granulation is performed by human experts, the basic principles and rules will be built for practitioners to follow. For example, granulation should meet the universal approximation properties. The infinite input space is granulated into finite granules, but finite granules should be able to approximate the universe of discourse in some ways.

Framework of granular computing: The hierarchical structure of granular computing will be established to represent the granules, relationships between granules, and computation with granules from the point of view of machine-centered computations. This research is originated from human thinking and problem solving and will be developed from philosophy, psychology, and behavior sciences. The general framework of granular computing will concentrate in the granulation and granules transformation.

Information integration: Basically, granular computing granulates complex problems into subproblems, conquer subproblems, and integrate solutions to subproblems to obtain the solution to the original complex problem. Granulate-and-conquer is the natural extension of the divide-and-conquer strategy, and addresses the integration of the subsolutions on granules into the total solutions on the original problem. For general granular computing, it is possible that two distinct problems be granulated into the same collection of granules. Different integrations of the same subsolutions may lead to the different results.

Applications of granular computing: Granular computing is strongly domain dependent. For example, in fuzzy logic control, the unknown control functions can be expressed by the membership functions of fuzzy granules; in structural programming, the programmers partition the functions according to their experience and background knowledge; while in rough set theory, the indiscernibility equivalence relation is provided in priori by experts and unknown concepts are approximated by equivalence classes. These examples demonstrate that granulation is not algorithmic and strongly domain dependent. Similarly, computation with granules is dependent on the domain knowledge and the computation task. Extracting the common properties and features of granular computing in various application domains will be emphasized to enhance the general framework of granular computing and contribute to the foundation of granular computing.

Bibliography

1. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile, pp 487–499
2. Agrawal R, Imielinski T, Swami A (1993) Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD Conference, Washington DC, 1993
3. Bargiela A, Pedrycz W (2002) Granular computing: An introduction. Kluwer Academic, Boston
4. Giunchglia F, Walsh T (1992) A theory of abstraction. Artif Intell 56:323–390
5. Goodrich M, Tamassia R (2004) Algorithm design: Foundations, analysis, and internet examples, 2nd edn. Wiley, New Jersey
6. Gordon AD (1981) Classification. Chapman Hall, London
7. Hobbs J (1985) Granularity. In: Proc. of International Joint Conference on Artificial Intelligence, pp 432–435
8. Hu X, Liu Q, Skowron A, Lin TY, Yager RR, Zhang B (2005) Proc. of IEEE International Conference on Granular Computing. IEEE Press, Los Alamitos
9. James M (1985) Classification algorithms. William Collins, Glasgow
10. Lin TY (1988) Neighborhood systems and relational database. Abstract. In: Proceedings of CSC '88, pp 725
11. Lin TY (1989) Neighborhood systems and approximation in database and knowledge base systems. In: Proceedings of the Fourth International Symposium on Methodologies of Intelligent Systems, pp 75–86
12. Lin TY (1997) Neighborhood systems – a qualitative theory for fuzzy and rough sets. In: Wang P (ed) Advances in machine intelligence and soft computing, vol IV. Duke University, North Carolina, pp 132–155
13. Lin TY (1998) Granular computing on binary relations I: Data mining and neighborhood systems. In: Skowron A, Polkowski L (eds) Rough sets in knowledge discovery. Physica, New York, pp 107–121
14. Lin TY (1999) Granular computing: Fuzzy logic and rough sets. In: Zadeh LA, Kacprzyk J (eds) Computing with words in information/intelligent systems. Physica, New York, pp 183–200
15. Lin TY (2000) Data mining and machine oriented modeling: A granular computing approach. Appl Intell 13(2):113–124
16. Lin TY (2002) Attribute (feature) completion – the theory of attributes from data mining prospect. In: Proceedings of International Conference on Data Mining, pp 282–289
17. Lin TY (2003) Granular computing. In: Lecture notes in computer science, vol 2639. Springer, Berlin, pp 16–24
18. Lin TY, Yao YY, Zadeh LA (eds) (2002) Data mining, rough sets and granular computing. Physica, Heidelberg
19. Louie E, Lin TY (2000) Finding association rules using fast bit computation: Machine-oriented modeling. In: Proceedings of International Symposium on Methodologies for Information Systems, pp 486–494

20. O'Docherty M (2005) Object-oriented analysis and design. Wiley, New Jersey
21. Pawlak Z (1982) Rough sets. Int J Comput Inf Sci 11:341–356
22. Pawlak Z (1998) Granularity of knowledge, indiscernibility and rough sets. In: Proceedings of IEEE International Conference on Fuzzy Systems, pp 106–110
23. Skowron A, Stepaniuk J (2001) Information granules: Towards foundations of granular computing. Int J Intell Syst 16:57–85
24. Yao JT (2006) Information granulation and granular relationships. In: Proc. of IEEE GrC, pp 326–329
25. Zadeh LA (1979) Fuzzy sets and information granularity. In: Gupta M, Ragade R, Yager R (eds) Advances in fuzzy set theory amd applications. North-Holland, Amsterdam, pp 3–18
26. Zadeh LA (1997) Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets Syst 19:111–127
27. Zadeh LA (1998) Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems. Soft Comput 2:23–25
28. Zhang Y, Lin TY (2006) Proc. of IEEE International Conference on Granular Computing. IEEE Press, Los Alamitos
29. Zhang L, Zhang B (2003) The quotient space theory of problem solving. In: Lecture Notes in Computer Science, vol 2639. Springer, Berlin, pp 11–15

Granular Computing System Vulnerabilities: Exploring the Dark Side of Social Networking Communities

Steve Webb[1], James Caverlee[2], Calton Pu[1]
[1] College of Computing, Georgia Institute of Technology, Atlanta, USA
[2] Department of Computer Science, Texas A&M University, College Station, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Background on Social Networking Communities
Traditional Attacks Targeting Social Networking Communities
New Attacks Against Social Networking Communities
Future Directions
Bibliography

Glossary

Social networking community A community of individuals (users) that are connected based on mutually shared activities, beliefs, goals, or relationships.

Profile A user-controlled web page that contains a picture of a community user along with various pieces of personal information for that user. This is often the online representation of an individual in a social networking community.

Friend request A communication mechanism that allows community users to initiate connections between their profiles and the profiles of their friends.

Definition of the Subject

Online social networking communities are connecting hundreds of millions of individuals across the globe and facilitating new modes of interaction. Due to their immense popularity, an important question is whether these communities are safe for their users. In this paper, we address this safety question and show that social networking communities are susceptible to numerous attacks. Specifically, we identify two attack classes: traditional attacks that have been adapted to these communities (e. g., malware propagation, spam, and phishing) and new attacks that have emerged through malicious social networking profiles (e. g., rogue advertising profiles and impersonating profiles). Concretely, we describe examples of these attack types that are observable in MySpace, which is currently the most popular social networking community.

Introduction

Over the past few years, social networking communities have experienced unprecedented growth. Communities such as MySpace and Facebook are connecting people in a variety of new and exciting ways, and as a result, individuals are attaching an increasing amount of value to their online personas. Unfortunately, the rising importance and prominence of these communities have also made them prime targets for attack by malicious entities.

We observe two distinct attack classes that threaten social networking communities and the privacy of their users. First, traditional attacks that have plagued Internet users for many years (e. g., malware propagation, spam, and phishing) have been adapted to take advantage of the unique properties of online communities. These traditional attacks are thriving in their new environment, and the severity of these attacks promises to grow as attackers become more sophisticated. The second attack class consists of new attacks that have emerged from the very fabric of these communities. One prominent example is the use of deceptive profiles (e. g., rogue advertising profiles and impersonating profiles) that are becoming more widespread, difficult to detect, and extremely costly to legitimate community members.

In this paper, we provide detailed descriptions for each of these attack classes, and we show that the continued success of social networking communities is contingent upon their ability to mitigate the risks associated with these attacks. For concreteness, we describe attacks that are observable in MySpace, which is the most popular social networking community in terms of unique visitors (more than 65 million in February 2008) [3], total traffic (4.29% of all US Internet visits in February 2008) [6], and user base (more than 110 million active accounts as of February 2008) [14]. Our observations show the practical importance of these attacks and their impact on millions of users. Additionally, since other social networking communities are both functionally and structurally similar to MySpace, the attacks we describe can be easily adapted to most (if not all) social networking communities.

The rest of the paper is organized as follows. Sect. "Background on Social Networking Communities" provides background information about social networking communities. In Sect. "Traditional Attacks Targeting Social Networking Communities", we describe traditional attacks that have adapted to target social networking environments, including malware propagation, spam, and phishing. In Sect. "New Attacks Against Social Networking Communities", we present new attacks that specifically target social networking communities by utilizing deceptive profiles such as rogue advertising profiles and impersonating profiles. Sect. "Future Directions" concludes the paper.

Background on Social Networking Communities

Social networking communities provide an online platform for people to manage existing relationships, form new ones, and engage in a variety of social interactions. Typically, a user's online presence in these communities is represented by profile, which is a user-controlled web page that contains a picture of the user along with various pieces of personal information. Profiles connect to other profiles through explicitly declared friend relationships and numerous messaging mechanisms.

An example of a MySpace profile is shown in Fig. 1a. Some of a profile's personal information is mandatory (e. g., the user's name, age, gender, location, etc.), and some of it is optional (e. g., the user's interests, relationship status, occupation, etc.). To facilitate self-expression, each profile also has free text "About me" and "Who I'd like to meet" sections, and users are allowed to embed various objects in their profiles such as pictures, audio clips, and videos. In addition to personal information and embedded

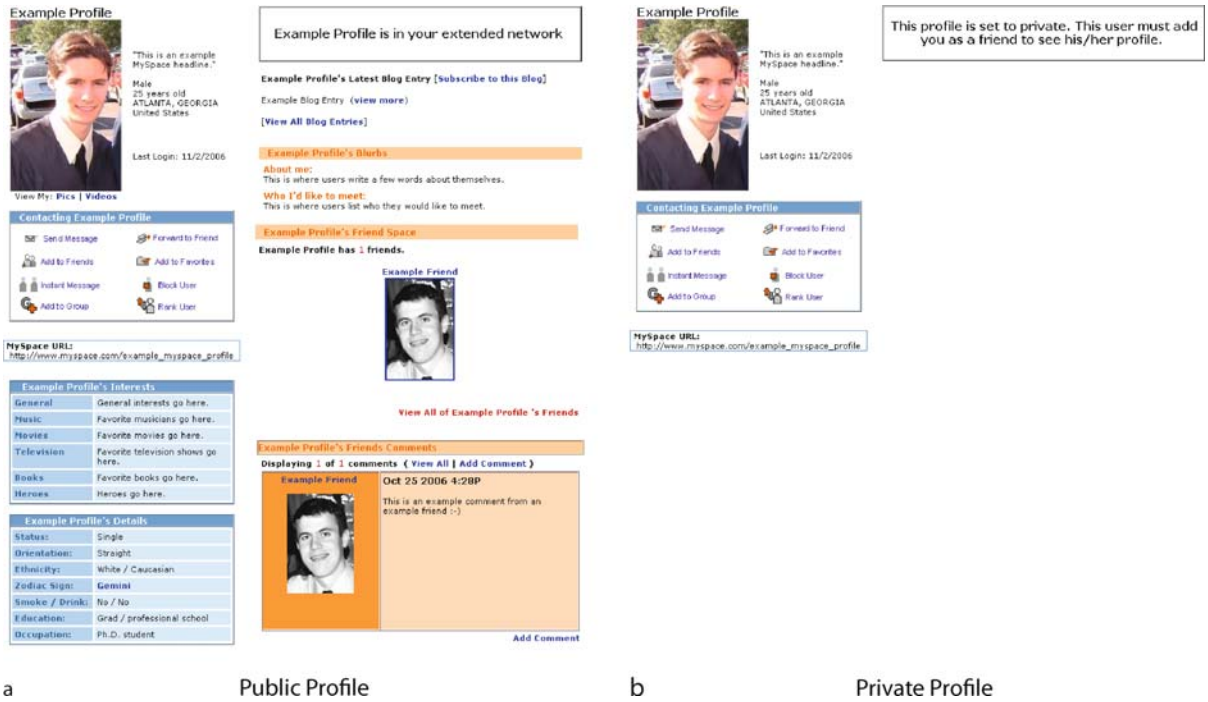

Granular Computing System Vulnerabilities: Exploring the Dark Side of Social Networking Communities, Figure 1
Example MySpace profiles. In a, the profile is publicly accessible. In b, the profile is private

content, a user's profile also contains a list of links to the profiles of that user's friends. These friend links are bidirectional because a link is established only after both parties acknowledge the friendship. To initiate a friendship, a user sends a friend request to another user. If the other user accepts this request, the friendship is established, and a friend link is added to both users' profiles.

Aside from friend requests, MySpace provides a number of other communication facilities that enable users to communicate with each other within the community. These facilities include messaging, bulletin, commenting, blogging, and instant messaging (IM) systems. The messaging system allows users to exchange intra-community email messages with any other user (i. e., both friends and strangers). The bulletin system is essentially an exclusive bulletin board that only a user's friends can view, enabling users to communicate with all of their friends at once. Users can post comments on their friends' profiles using the commenting system, and the blogging system allows users to maintain blogs on their profiles, which other users can read and comment on. Finally, the IM system provides users with a mechanism to send intra-community instant messages to any other user.

Due to the wealth of private information that is accessible on user profiles and the various means of communication that are available, MySpace provides mechanisms to protect the privacy of its users. First and foremost, users have the ability to choose between making their profiles publicly viewable (the default option) or private. If a user's profile is designated as private, only the user's friends are allowed to view the profile's detailed personal information (e. g., the user's interests, blog entries, comments, etc.). However, as Fig. 1b shows, a private profile still reveals the user's name, picture, headline, gender, age, location, and last login date. MySpace also provides a few finer-grained privacy mechanisms. Users can control who is allowed to IM them (everyone, only friends, or no one) and who is allowed to leave blog comments (everyone or only friends). Users can also maintain a block list to prevent specific users from contacting them at all. Unfortunately, as we will see in the following sections, these privacy mechanisms have been unable to prevent a number of attacks.

Traditional Attacks Targeting Social Networking Communities

Since social networking communities include communication facilities that are fundamentally similar to traditional means (i. e., email, blogs, instant messaging, etc.), many of the attacks that are effective against those traditional communication media have been adapted to exploit social network communications. Due to the massive size of many social networking communities, their tightly connected nature, and their relatively naïve user bases, these communities are target rich environments for attackers. In this section, we describe three of the adapted attacks that have been observed in MySpace: malware propagation, spam, and phishing.

Malware Propagation

Malware creators aim to spread their malicious content to as many victims as possible. Since MySpace is the most popular community on the web, it has become a prime target for malware propagation. In fact, over the past couple of years, MySpace was attacked by at least one instance of each of the following malware categories: worms, spyware, and adware. For the remainder of this section, we will detail the most interesting occurrences of these attacks, and we will explain the threats they pose to social networking communities.

Worms The most successful example of rapid worm propagation in a social networking community occurred in MySpace on October 4, 2005. The worm was called the "Samy worm," and it generated more than a million friend requests for its creator (Samy) over the course of a single day. The basic operation of this worm was quite simple but extremely clever [11]. First, Samy wrote the worm using Javascript, and then, he embedded it in his MySpace profile. MySpace disallows users from adding scripts to their profiles by filtering scripting tags and removing specific strings (e. g., "javascript"). To evade these filters, Samy exploited the behavior of popular web browsers such as Internet Explorer. Specifically, he hid the code inside a Cascading Style Sheet (CSS) tag and obfuscated the strings that MySpace would filter (e. g., "javascript" became "java\nscript"). Thus, even though MySpace had security mechanisms in place, the lax security of certain web browsers allowed the obfuscated code to execute successfully.

When a MySpace user accessed Samy's profile, the embedded Javascript code sent Samy a friend request on that user's behalf. The code also embedded itself in the user's profile (in the same manner that it was embedded in Samy's profile). Consequently, when other MySpace users visited the newly infected profile, the code would send Samy friend requests from those users and propagate itself to their profiles. As Samy described it, "If 5 people viewed my profile, that's 5 new friends. If 5 people viewed each of their profiles, that's 25 more new friends."

Fortunately, this worm was relatively harmless for its infected users because it only affected their MySpace profiles and not their actual machines. However, the same cannot be said for MySpace. Due to the worm's viral growth pattern, the MySpace administrators were forced to temporarily shut down the site to stop the worm's propagation and remove the worm's code from the infected users' profiles. The service outage was relatively brief, but this incident clearly illustrates the potential damage that social networking worms are capable of inflicting. Specifically, this worm teaches two very important lessons. First, the success of the worm's obfuscated code clearly highlights the importance of web site security as well as web browser security. A delicate balance exists between functionality and security, and this balance must be respected when designing and developing online communities. Second, the worm's propagation speed showcases how quickly the entire community could become infected. In this case, more than a million users were affected in less than 24 hours.

A much more dangerous social networking worm appeared on MySpace in the middle of July 2006. Similar to the Samy worm, this worm propagated itself through the profiles of unsuspecting MySpace users. However, unlike the Samy worm, this worm's code was hidden inside a malformed Shockwave Flash (.swf) file that directly exploited a vulnerability on users' machines. Additionally, instead of generating innocuous friend requests, this worm actually redirected users' web browsers to a politically charged blog posting.

The manner in which this worm spread is as follows. First, the worm's creator generated a malformed .swf file and embedded it in a MySpace profile. This .swf file exploited a critical vulnerability in Macromedia Flash Player v8.0.24.0 and earlier versions, which allows an attacker to execute code on an affected machine. When a MySpace user with a vulnerable Macromedia Flash Player accessed the profile, the code in the .swf file automatically redirected the user's browser to a blog posting that contained political propaganda. Finally, the code propagated itself by embedding a copy of the .swf file in the infected user's profile.

This worm was also relatively harmless; however, it was far more troubling than the Samy worm because it actually exploited a vulnerability on users' computers, which allowed it to execute code. Fortunately, the worm's creator was only interested in spreading political ideals because nothing prevented the worm from installing any number of malicious utilities on its victims' machines. For example, the worm could have installed spyware or adware, and it could have even zombified the infected machine (i. e., compromise the computer and enlist it in a botnet [4]). Combine these frightening, yet completely realistic, scenarios with the viral propagation patterns of these worms, and it becomes immediately obvious how devastating worms could be in a social networking environment and why they must be prevented.

Spyware Although spyware has yet to be distributed within the payload of a social networking worm, it has already made a few appearances in social networking communities. The most prominent example of spyware appearing in MySpace occurred towards the beginning of July 2006. At that time, an advertisement for deckoutyourdeck.com, which contained a malformed Windows Metafile (.wmf) image, was inserted into one of the ad networks that MySpace uses. This malformed .wmf image exploited a critical vulnerability in the Graphics Rendering Engine of Windows, which allows remote code execution.

MySpace displays an ad banner, which is randomly selected from MySpace's supplying ad networks, at the top of every profile. When a user accessed a profile that displayed the deckoutyourdeck.com ad, the user was prompted to download the embedded .wmf image. Figure 2 shows a screenshot of this prompted download. If the user was running an unpatched version of Windows, the .wmf file installed an assortment of programs, including known spyware utilities such as PurityScan [9]. These spyware utilities pose a serious threat to the user because they track the user's web browsing behaviors, and they install various other third-party applications without the user's consent.

Unfortunately, despite the fact that Microsoft released a patch for this vulnerability more than six months before the ad appeared on MySpace, over a million users were affected. Thus, this incident reveals two very important points. First, many users have a false sense of security in these communities. One of the most basic secure browsing principles is never to download questionable content, yet more than a million users gladly accepted this suspicious (and completely unsolicited) download request. Second, an alarming number of users are not vigilant about protecting themselves against security threats. This incident proves that at least a million users neglected to install security patches for more than six months. Therefore, social networking communities must assume that their users are completely vulnerable and take every precaution necessary to protect them from malicious content.

Adware An interesting instance of adware appearing in MySpace occurred around the same time as the previous spyware example. A security researcher was browsing MySpace profiles and found two that were named after a known adware company called "Zango" (formerly

Granular Computing System Vulnerabilities: Exploring the Dark Side of Social Networking Communities, Figure 2
Embedded .wmf image within a deckoutyourdeck.com advertisement

known as "180solutions"). Both profiles were created to deceive users into downloading and installing the Zango Search Assistant and Toolbar (two known adware programs) [2]. The first profile claimed that the programs could "protect kids from predators," and the second profile was even more deceptive.

When a user accessed the second Zango profile, a popup with a license agreement immediately launched, asking the user to accept the license in order to play a video file. This license popup is shown in Fig. 3. If the user accepted the license by clicking the "Play Now" button, a video file began to play, but secretly, the Zango Search Assistant and Toolbar were also installed on the user's system. As Fig. 3 illustrates, this popup contained a number of deceptive elements. First, the popup did not explicitly indicate that an installation was taking place. The text in the top-left corner of the popup mentions the Zango Search Assistant and Toolbar, but it does not tell the user that they will be installed. Additionally, none of the popup's buttons mention an installation. Instead, they are entitled "Play Now" and "Play," which implies that the only action will be the play-back of a video file. Finally, the most prominent features of the popup are the video's preview window in the top-right corner and the "Play Now" button. The actual license agreement and its preselected checkbox are positioned at the bottom, where they are likely to be overlooked. Consequently, many users played the video before they noticed the license agreement, and the adware was successfully installed on their machines.

According to Zango, both profiles were created by a Zango developer that was acting against the company's policy of not targeting MySpace. As a result, the deceptive video clip was removed, and the company released a public apology. However, this occurrence clearly illustrates the potential for abuse by adware companies in social networking communities. Additionally, the success of this deception further illustrates the naïvety of users and the need for these communities to provide effective security mechanisms that protect against malicious content.

Spam

In addition to malware propagation, another type of attack on traditional communication media (e. g., email, blogs, instant messaging, etc.) is spam. Since MySpace provides similar communication facilities, they are also susceptible to spamming abuse. In fact, almost all spamming activities that occur outside the community can be recreated inside the community. To aggravate the problem, spammers can also use the profile information posted by users to target them outside the social networking community. Although MySpace's Terms of Use Agreement prohibits users from providing contact information such as email addresses and URLs on their profiles, many users still do so at their own peril. Consequently, spammers can use that information to spam those users using traditional techniques.

Spamming relies on the open nature of communications; thus, the vulnerabilities of MySpace's communication facilities are proportionate to the openness of their access. On the conservative side, the commenting and bulletin systems are the most spam resistant because they can only be used by a user's friends. The blogging, IM, and

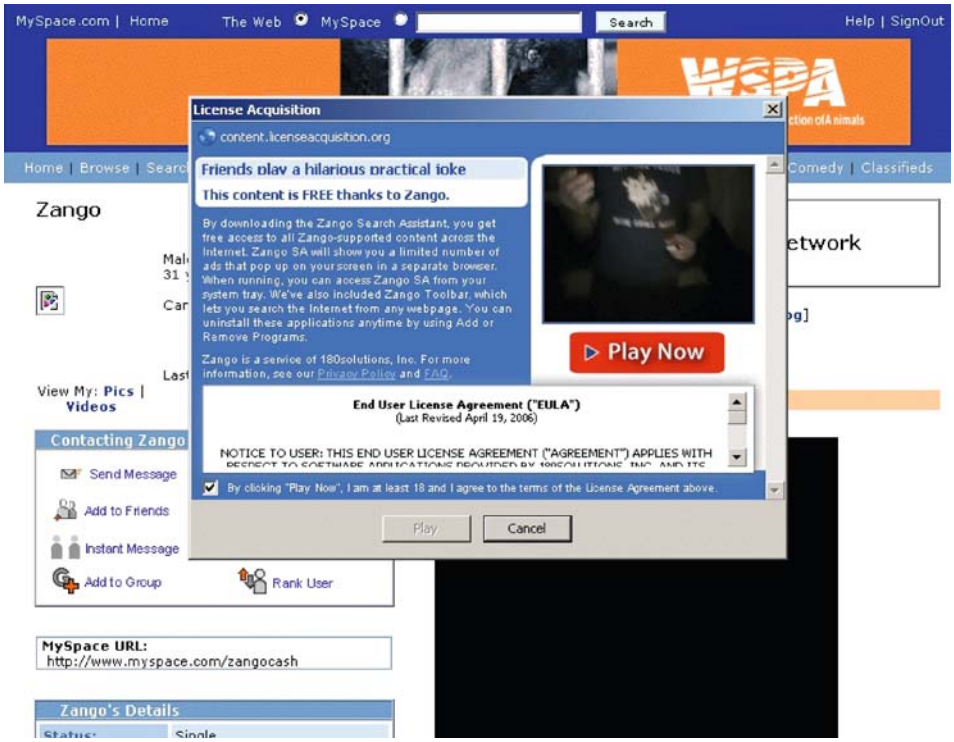

Granular Computing System Vulnerabilities: Exploring the Dark Side of Social Networking Communities, Figure 3
Deceptive Zango popup license agreement

friend request systems also provide the option of disallowing non-friends from using them to contact a user, but this protection is not enabled by default. Thus, these three systems are more susceptible to spamming because any user can use them to contact other users that have not enabled this protection. The messaging system is the most vulnerable to abuse because it does not provide an effective spam prevention mechanism. It allows users to report a message as being spam, but those users have no assurances that immediate action will be taken. Therefore, users may receive a number of spam messages before MySpace administrators eliminate the offending spammer from the system. MySpace also provides a general block listing mechanism that allows users to completely prevent all communication from specific users. However, this feature is ineffective due to the ease with which users can create new MySpace profiles (i. e., if spammers have been blocked, they can create new profiles and continue their spamming activities undeterred).

Spammers abuse these communication systems for the same reason they abuse traditional communication facilities: promotion. Spammers want to expose their products, web sites, and viewpoints to as many individuals as possible, and spamming provides them with a technique to accomplish that goal. For social networking communities, this spamming activity represents a huge problem for a couple of reasons. First, spam content wastes a significant amount of resources, including storage space, bandwidth, and users' time. This last wasted resource leads us to the second major consequence of social network spam. When users waste time dealing with spam, it has a negative effect on their overall experience with these communities because it prevents them from participating in their desired activities. For example, when users are forced to sift through an inbox full of annoying spam messages, they are unable to send and receive messages. Similarly, when users are burdened with removing an endless stream of spam comments on their blogs, they are unable to post new content. Consequently, if users become overwhelmed with spam, they will have no incentive to continue interacting with these communities, and the communities will be forced to shut down.

In addition to spamming the social networking community, spammers are also able to use the information on user profiles to more effectively spam users outside the community. By mining email addresses, IM screen names,

and other contact information from these profiles, spammers can spam users using traditional techniques (e. g., email spam, spim, blog comment spam, etc.). Additionally, spammers can use the personal information on these profiles to construct user-specific spam content that is more relevant to the user, making it more likely to be read. For example, if a user's profile contains a great deal of sports-related information, a spammer can leverage this information to create a sports-oriented spam message for that user. Since the message's content matches the user's interests, the user is much more likely to read it, and as a result, the spam's sales pitch is more likely to be successful. To make their messages even more deceptive, spammers can also masquerade as users' friends. A user's profile contains a list of that user's friends; thus, spammers can use this information to make their messages appear as if they were sent by those friends. Since a user is much more likely to read and trust a message from a friend, the spammer has a higher probability of success with these disguised messages [7].

Phishing

In a traditional phishing attack, a phisher seeks to obtain a targeted user's sensitive information through means of deception. Historically, phishers have been interested in credit card information, banking information, and login information for various web sites (e. g., eBay, PayPal, etc.). As social networking communities have become more popular, complex networks of friends have been established within them. Consequently, social networking communities have become a prime target for phishers because they are portals to a large number of potential victims.

A MySpace phishing attack utilizes the same deceptive techniques that are employed in traditional phishing attacks. First, a user is presented with a seemingly legitimate URL (called a *phishing URL*) that appears to be affiliated with MySpace. This phishing URL is typically propagated in two distinct manners: the communication systems provided by MySpace and traditional communication channels (e. g., email, blogs, instant messaging, etc.). For example, in May 2006, a phisher used the MySpace spamming techniques described above to send a phishing message to various MySpace users. This message had "CHECK OUT these old school pictures ..." as its subject and a phishing URL in its body [15]. Upon accessing one of these phishing URLs, the user is directed to a fraudulent web page that appears identical to the authentic MySpace login page. When the user enters the necessary MySpace login information, the fraudulent page stores that information and uses it to redirect the user to the authentic MySpace community. Thus, the user is completely unaware of the attack, while the phisher successfully obtains the user's sensitive login information.

Phishing attacks represent a serious threat to MySpace users for three reasons. First, victimized users often lose control of their profiles. In many cases, phishers will immediately change the login information of profiles they have compromised, locking victims out of their own profiles. Since users spend a great deal of time and energy building new friendships, writing blogs, and customizing their profiles, it is somewhat traumatic when they lose control of their creations. Even in cases where a compromised profile's login information remains the same, the phisher's activities severely damage that profile's credibility. For example, if a phisher compromises a user's profile and begins spamming that user's friends, those friends will eventually distrust the compromised profile and remove it from their lists of friends. The second reason these attacks are dangerous is due to the viral propagation patterns that are possible in social networking communities. As shown with the malware examples above, one compromised user can quickly escalate into a network full of compromised users. Specifically, once a single user's login information is compromised, the phisher can use that user's profile to propagate the attack to the user's friends. In our spam discussion, we mentioned a few MySpace communication systems that are somewhat spam resistant. However, the spam resistance of those systems assumes that a user's friends can be trusted (i. e., they are not spammers, phishers, etc.). Thus, if a user's profile becomes compromised, that user's friends are immediately at risk because the phisher can contact them under the guise of a profile they trust. As a result, a compromised user's friends are highly likely to become phishing victims (i. e., access the phishing URL), and by the time they realize they should distrust the compromised profile, it will be too late. The final threat posed by MySpace phishing attacks relies on the knowledge that many computer users reuse the same login information at many different sites [10]. Thus, if a user's MySpace login information is compromised, that user's login information at those other sites is automatically compromised as well.

In addition to launching phishing attacks that specifically target social networking communities, phishers can also use the private information found in these communities to make their traditional phishing attacks more effective. As previously mentioned, many MySpace users include various pieces of personal information on their profiles (e. g., location, interests, occupation, etc.). Phishers can easily use this personal information to construct user-specific messages that a potential victim would be much

more likely to read and trust. For example, a phisher could send the potential victim a fraudulent eBay email that contains auction information for products the victim would be interested in buying. Each of these products could be selected using the user's interests and associated with one or more phishing URLs. Since this email message contains user-specific content, the victim is more likely to read it and access one of its phishing URLs. As a result, this user-specific attack has a higher probability of success than a generic phishing attack.

New Attacks Against Social Networking Communities

Initially, social networking communities were composed of ordinary people – typically kids and young adults that created profiles to interact with their friends. However, as these communities began to expand, new parties became interested because the communities evolved into portals for connecting and interacting with these ordinary people. The first wave of new participants were entertainers (e. g., musicians, comedians, and actors) that created profiles to communicate with their current fans and reach out to potential new ones. Not long after that, companies began partnering with communities to create advertising profiles (i. e., profiles that advertise the company and its products). Eventually, other public figures such as politicians became involved, creating profiles to promote their campaigns, solicit volunteers, and request donations. Unfortunately, the explosive growth of these communities and the conflicting interests of their various participants have generated a range of new attacks. In this section, we describe attacks that utilize two types of malicious social networking profiles: rogue advertising profiles and impersonating profiles.

Rogue Advertising Profiles

According to analysts at eMarketer, companies currently spend around $280 million on social network advertising in the US, and by 2010, that figure is expected to grow to $1.9 billion [8]. Thus, it is not surprising that MySpace has partnerships with numerous companies, which allow those companies to advertise legitimately in the social networking community. For example, both Burger King and Wendy's have embraced the potential of social network advertising by creating advertising profiles on MySpace. Burger King uses the King, a marketing character that appears in the company's commercials, to promote its MySpace profile at http://www.myspace.com/burgerking. Similarly, Wendy's uses a small hamburger patty named Smart to promote its MySpace profile at http://www.myspace.com/wendysquare. However, this new advertising craze is not exclusive to fast food restaurants. MySpace profiles also exist for television shows (e. g., "It's Always Sunny in Philadelphia"), commercial products (e. g., Herbal Essences shampoo), and movies (e. g., "Talladega Nights").

Although MySpace's Terms of Use Agreement strictly prohibits commercial use of the community without prior approval, a number of companies violate this policy by creating advertising profiles without MySpace's consent. We refer to these profiles as *rogue advertising profiles*, and they appear in many forms with varying degrees of deception. The most innocuous group of rogue advertising profiles is created by small web site operators that are merely trying to generate web traffic for their sites. These profiles are relatively harmless because they simply include URLs for these small sites, and they are openly affiliated with a company (i. e., they do not attempt to deceive users into believing the profiles belong to an ordinary individual). However, these profiles are still unacceptable because they pollute the community with unwanted and unauthorized advertising.

A much more deceptive group of rogue advertising profiles is generated by nefarious companies such as gambling and pornographic web sites. Unlike legitimate advertising profiles, which are clearly identifiable as a marketing device, these deceptive rogue advertising profiles appear as though they are maintained by an individual (and not a company). As a result, naïve users can easily mistake these rogue advertising profiles as ordinary MySpace profiles that were created by ordinary MySpace users. The deceptive profile construction process proceeds as follows. First, a fraudulent MySpace profile is created using a picture of an attractive female. This profile also contains a provocative description in its "About me" section that visitors assume was written by the pictured female. Conveniently, this description usually includes at least one reference to the URL of a nefarious web site. If the profile is particularly deceptive, the description includes an instant messenger screen name instead of a URL. The inclusion of a screen name is especially manipulative because most users assume that a conversation over IM can only be accomplished by a real person. However, the screen names found on these rogue advertising profiles are almost always attached to an IM bot (i. e., a computer program that emulates a real conversation) that attempts to direct its victims to the URL of a nefarious web site. Once the profile is completed, its creator sends friend requests to males near a specific geographic location, which the profile is also associated with (i. e., the pictured female claims to live there). When the males receive these requests, a num-

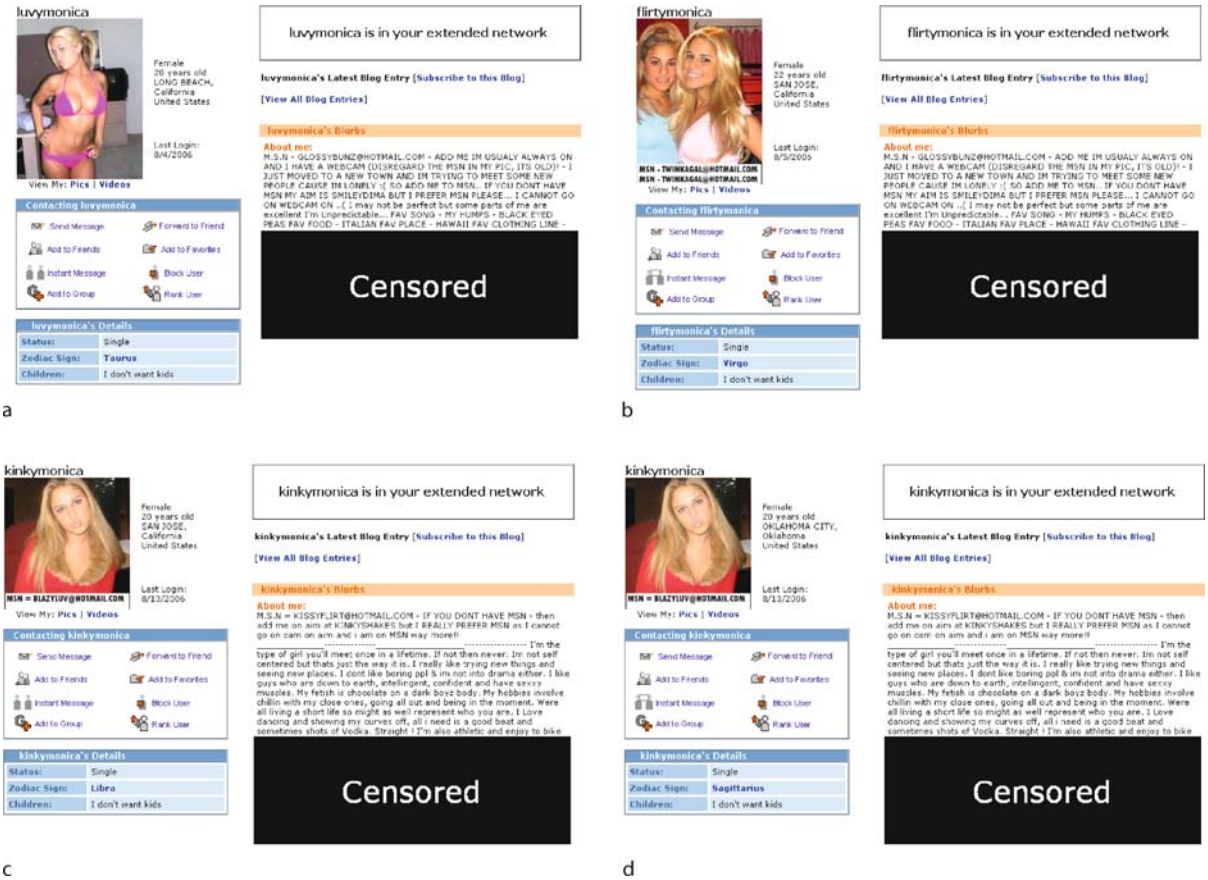

Granular Computing System Vulnerabilities: Exploring the Dark Side of Social Networking Communities, Figure 4
Pornographic rogue advertising profiles

ber of them accept and visit the profile, making the deception a success.

Figure 4 shows four examples of deceptive rogue advertising profiles, which we slightly modified to hide objectionable content. By analyzing these examples, a number of interesting observations emerge. First, the profiles are extremely similar. Figures 4a and 4b share the exact same "About me" text, and Figs. 4c and 4d share the same "About me" text and profile picture. Thus, these two pairs of profiles were probably created by the same individuals. The next observation addresses the content of each profile's "About me" text. All four of these profiles are particularly deceptive because they all provide instant messenger screen names to induce users to contact IM bots. Finally, all of the profiles claim to be located in different cities. Despite their uncanny similarities, the profiles in Figs. 4a and 4b are supposedly located in different parts of California, and the profiles in Figs. 4c and 4d are in completely different parts of the country. These location differences exist because each profile is meant to target males in a different geographic area. This also explains why the profiles are allowed to be so similar. As long as the profiles in a given location are unique, users are less likely to become suspicious that the profiles are fraudulent.

Deceptive rogue advertising profiles represent a serious security threat because they directly manipulate the behavior of users. Since the profiles appear to be maintained by ordinary (albeit very attractive) people, other users mistakenly trust their content. Additionally, users often believe they have made a connection with the people pictured on these profiles because the users received friend requests from these profiles. Consequently, users access the URLs that are on the profiles (or in IM conversations) and become victims to the content found on the corresponding web pages. In most cases, these pages contain pornographic or gambling-related material, but nothing prevents malicious individuals from embedding malware on the pages. Thus, to protect its users, social networking communities must develop techniques to identify and eliminate these profiles.

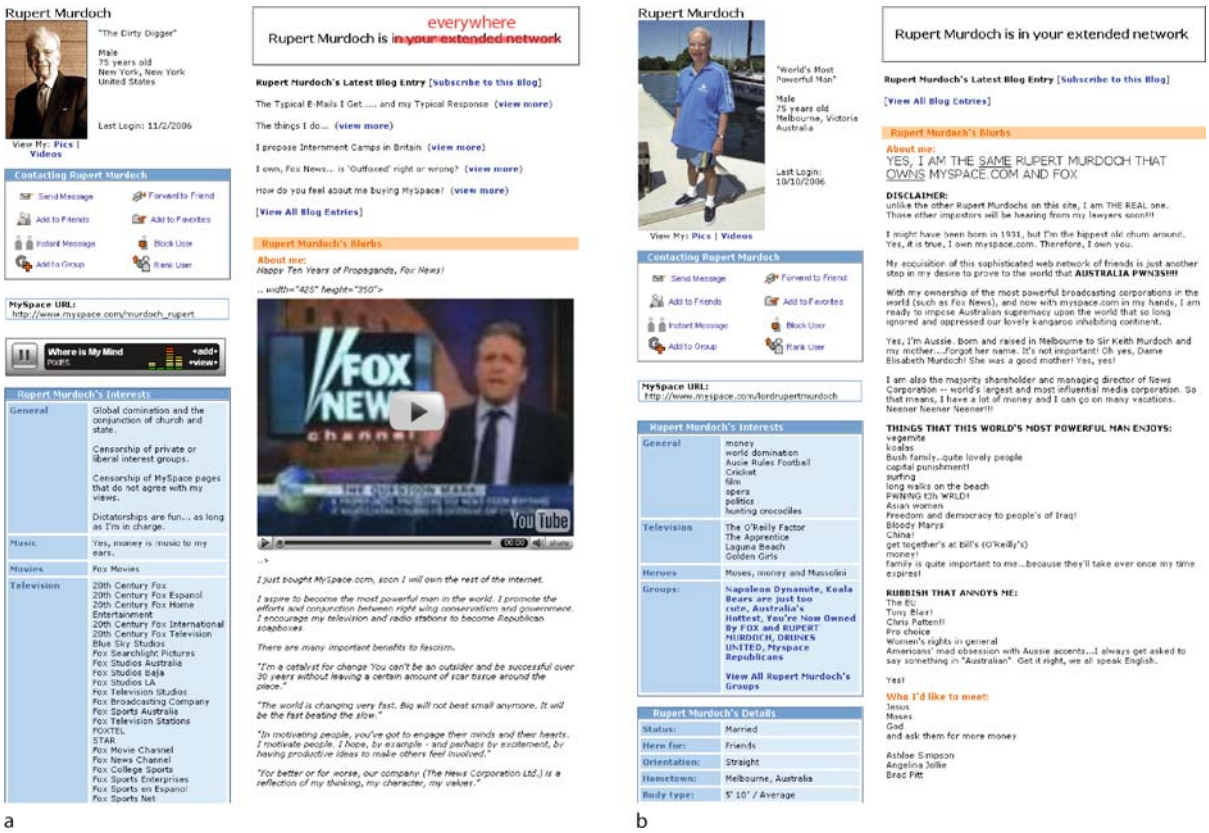

Granular Computing System Vulnerabilities: Exploring the Dark Side of Social Networking Communities, Figure 5
Example impersonating profiles for Rupert Murdoch

Impersonating Profiles

Due to the enormous popularity of social networking communities, they have emerged as forums for individuals to voice their opinions about various topics as well as other people. In MySpace, some users are more tactful and express their views in blog postings on their profiles or in comments on their friends' profiles. However, other users have taken their crusade to an entirely new level, creating new MySpace profiles that directly target specific ideals, companies, and even people. We refer to these profiles as *impersonating profiles* because they impersonate their targets to convey their message.

MySpace contains profiles that impersonate actors (e. g., Tom Hanks), athletes (e. g., Michael Jordan), technologists (e. g., Bill Gates), and politicians (e. g., George W. Bush). Even ancient philosophers, such as Socrates and Aristotle, have profiles dedicated to them. In fact, most of these individuals are impersonated by multiple distinct profiles. In some cases, these profiles are meant to be an homage to the individuals in question. However, more often than not, the impersonating profiles are meant to be slanderous to the targets of the impersonation. For example, after News Corporation purchased MySpace in July 2005, impersonating profiles for Rupert Murdoch (News Corporation's CEO) began to appear, making claims such as, "I just bought MySpace.com, soon I will own the rest of the internet" and "Dictatorships are fun … as long as I'm in charge." Two examples of these profiles are shown in Figs. 5a and 5b. To make matters worse, impersonators are able to use MySpace's communication facilities to contact other users under the guise of these profiles. Thus, these counterfeit Rupert Murdoch profiles can express any number of scandalous opinions, and the negative backlash will be directed at the real Rupert Murdoch.

From the impersonators' standpoint, these profiles appear to be amusing satire. However, from the victims' standpoint, these profiles represent a form of identity theft as well as a public relations nightmare. The general public is unable to verify the authenticity of these profiles; thus, any slanderous comments found on the profiles (or received under the guise of the profiles) will be incorrectly

associated with the victims of the impersonation. Using the example above, nothing prevents users from believing that Rupert Murdoch actually created one of those profiles or that he actually shares the viewpoints it contains. Admittedly, these viewpoints are somewhat absurd, but it is easy to envision subtler comments that can severely damage an individual's reputation. Since the success of most public figures is completely contingent upon the strength of their fan bases, an impersonating profile represents a direct threat to its victim's credibility and livelihood.

Sadly, this growing epidemic is not limited to high profile individuals. Private citizens such as teachers, principals, and police officers have also been victimized by these impersonating profiles. In fact, over the past couple of years, the news has been filled with stories about these types of attacks. For example, in December 2005, a 16-year-old boy posted an impersonating profile of a local police officer, which contained various derogatory statements about the police officer's appearance, intelligence, and sexual orientation [13]. Then, in April 2006, an eighth grader created impersonating profiles of his English teacher that contained racist remarks and falsely represented the teacher as a pornographer and child molester [12]. In September 2006, a high school assistant principal sued two students for creating an impersonating profile that falsely identified her as a lesbian [1]. These examples clearly illustrate the magnitude of this problem, and unfortunately, they represent a small sample of a growing list of incidents.

In addition to the risks already mentioned, impersonating profiles can also be used to amplify the severity of the other attacks we have already discussed. Specifically, a spammer could create an impersonating profile for a MySpace user and use the profile to spam that user's friends. Malware creators and phishers could also use this approach to spread their malicious content to the user's friends. Since these friends would be under the impression that they were communicating with the authentic profile, they would be much more likely to trust the communication, and as a result, the attacks would be far more successful.

Future Directions

In only a few years, social networking communities have made staggering strides in popularity (MySpace welcomed more than 65 million unique visitors in February 2008 [3]) and importance (YouTube was acquired by Google for $1.65 billion in October 2006 [5]). Aside from their social and economic impact, one of the most important questions is whether social networking environments are safe for their users. In this paper, we have analyzed this safety question and described several security and privacy threats that have translated into real attacks in MySpace. Some of these attacks have been adapted from other Internet environments, including variants of malware propagation, spam, and phishing. Other attacks are new and unique to social networking environments, including the creation of rogue advertising profiles and impersonating profiles. From our real world observations, it is clear that additional efforts should be made to protect social networking users. We hope that researchers will identify and counteract emerging variations of these known attacks and provide measures for safeguarding the future of these valuable communities.

Bibliography

1. Associated Press (2006) Official Sues Students Over MySpace Page. http://www.sfgate.com/cgi-bin/article.cgi?file=/news/archive/2006/09/22/national/a092749D95.DTL
2. Boyd C (2006) Teenagers used to push Zango on Myspace. http://www.vitalsecurity.org/2006/07/teenagers-used-to-push-zango-on.html
3. Freiert M (2008) February Top Social Networks - Make way for the new guys. http://blog.compete.com/2008/03/07/top-social-networks-traffic-feb-2008
4. Geer D (2005) Malicious bots threaten network security. IEEE Computer 38(1):18–20
5. Google Press Center (2006) Google To Acquire YouTube for $1.65 Billion in Stock. http://www.google.com/press/pressrel/google_youtube.html
6. Hitwise (2008) Hitwise US - Top 20 Websites - February, 2008. http://www.hitwise.com/datacenter/rankings.php
7. Jagatic T et al (2007) Social Phishing. Communications of the ACM 50(10):94–100
8. King R (2006) Marketing to Kids Where They Live. http://www.businessweek.com/technology/content/sep2006/tc20060908_974400.htm?campaign_id=bier_tcs.g3a.091106a
9. Krebs B (2006) Hacked Ad Seen on MySpace Served Spyware to a Million. http://blog.washingtonpost.com/securityfix/2006/07/myspace_ad_served_adware_to_mo.html
10. Ross B et al (2005) Stronger password authentication using browser extensions. In: Proceedings of the 14th Usenix Security Symposium, Baltimore, 31 Jul – 5 Aug 2005, pp 17–32
11. Samy (2005) Technical explanation of the MySpace worm. http://namb.la/popular/tech.html
12. Sanchez M (2006) Pranksters posting fake profiles on MySpace. http://www.dfw.com/mld/dfw/news/local/15255785.htm?template=contentModules/printstory.jsp
13. Seibel J (2006) Boy charged in creating fake Myspace profile. http://www.jsonline.com/story/index.aspx?id=413620
14. Swartz J (2008) Social-networking sites going global. http://www.usatoday.com/money/industries/technology/2008-02-10-social-networking-global_N.htm
15. Technocrat (2006) Myspace Phishing Attacks on the Rise. http://djtechnocrat.blogspot.com/2006/05/myspace-phishing-attacks-on-rise.html

Granular Flows

James W. Dufty
Department of Physics, University of Florida, Gainesville, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Granular Fluid and Its Statistical Mechanics
Macroscopic Balance Equations
"Normal" States and Hydrodynamics
Navier–Stokes Approximation
Future Directions
Acknowledgments
Appendix
Bibliography

Glossary

Granular matter a system comprised of a large number of grains, or particles, of macroscopic size. Examples include powders, sand, seeds, and the surface of Mars.

Granular fluid an activated (driven) state of granular matter such that the grains move (flow) and collide frequently.

Statistical mechanics a field of physics that addresses systems with many degrees of freedom based on the fundamental microscopic laws to describe derived macroscopic properties.

Macrostate a statistical description of a system with many degrees of freedom in terms of limited information about that system.

Hydrodynamic fields the local densities of mass, energy, and momentum defined at each point in the system of interest.

Normal state a macrostate whose time evolution is described entirely through that of the average hydrodynamic fields.

Balance equations exact equations for the time derivatives of the hydrodynamic fields in terms of associated fluxes and sources.

Constitutive equations expressions for the fluxes and sources of the balance equations as functionals of the hydrodynamic fields.

Hydrodynamics a macroscopic description of the system in terms of a closed, deterministic set of equations for the average hydrodynamic fields, resulting from the exact balance equations with approximate constitutive equations.

Navier–Stokes hydrodynamics local, first order in time, partial differential equations for states with small spatial gradients in the hydrodynamic fields (constitutive equations calculated to first order in the spatial gradients).

Definition of the Subject

The terminology granular matter refers to systems with a large number of hard objects (grains) of mesoscopic size ranging from millimeters to meters. Geological examples include desert sand and the rocks of a landslide. But the scope of such systems is much broader, including powders and snow, edible products such a seeds and salt, medical products like pills, and extraterrestrial systems such as the surface regolith of Mars and the rings of Saturn. The importance of a fundamental understanding for granular matter properties can hardly be overestimated. Practical issues of current concern range from disaster mitigation of avalanches and explosions of grain silos to immense economic consequences within the pharmaceutical industry. In addition, they are of academic and conceptual importance as well as examples of systems far from equilibrium.

Under many conditions of interest, granular matter flows like a normal fluid [1]. In the latter case such flows are accurately described by the equations of hydrodynamics. Attention is focused here on the possibility for a corresponding hydrodynamic description of granular flows. The tools of nonequilibrium statistical mechanics [3], developed over the past fifty years for fluids composed of atoms and molecules [4,5], are applied here to a system of grains for a fundamental approach to both qualitative questions and practical quantitative predictions. Applications of basic atomic physics principles to granular fluids have accelerated during the past decade, starting with an emphasis on molecular dynamics (MD) simulations [6] and kinetic theory [7,8], and more recently with the theoretical methods of the type described here [9,10,11,12,13,14].

Introduction

To start with the familiar, consider a jar of vitamin pills, mustard seeds, or peanuts. Remove the lid and pour them into a bowl, observing that the "flow", or their collective motion, has some similarity to that of a normal fluid such as water. The collective motion in both cases is the consequence of collisions among their constituents, grains or atoms, and their large number. It is tempting to make the correspondence of grains to atoms in considering the similarities of flows in these two types of fluids. The objective here is to explore in formal detail the extent to

which that correspondence is conceptually and quantitatively justified. An important prerequisite is the integrity of the grains during their motion. Each grain is comprised of a large number of atoms or molecules. Integrity refers to their retention of mass and shape following interactions with other grains or with their environment. As such, the grains behave as "particles" whose detailed internal structure is not essential to their description, which is captured instead by a few parameters describing their shape, mass, and collisional properties with other grains. However, an important consequence of this underlying molecular structure is a redistribution of translational kinetic energy of the grains and internal energy of the constituent molecules. At the mesoscopic level this appears as an energy loss on collisions between pairs of grains. This is a central feature of granular fluids differentiating them from atomic fluids: the inelasticity of granular pair collisions.

Granular matter occurs in two classes of states, compact and activated [15]. In the first case, the grains form a static packed configuration within the container due to the effects of gravity on their relatively large mass and their inelastic collisions. Any initial motion is quickly dissipated and their kinetic energy becomes negligible relative to the gravitational potential energy. Important questions arise about the possible and probable packing configurations that determine the stresses within the system and the distribution of forces on the container. For example, chains of particles in contact can occur as arches to support matter above them while reducing their force on the matter below. There is an intense interest in the study of such states, generically referred to as contact mechanics.

Activated states refer to continuously driven systems, or gravity free conditions. For example, a container of grains in a compact configuration can be shaken to impose kinetic energy and motion among the grains. Similarly, unrestrained systems in a gravitational field will flow towards lower potential energy (e. g., hopper flow or flow down an incline). Initial activation in space laboratory experiments provides another example (self-sustained fluidization). For the flows considered here as candidates for a hydrodynamic description continual collisions are essential. This means that within each small cell, still containing many particles, the particles are moving randomly relative to the collective motion of that cell. Thus, ballistic motion or beams with all particles moving independently in the same direction are excluded.

Both compact and activated grains may occur immersed in a continuum such as water or air that may have a strong or weak effect on their collective properties. For compact systems water may provide a lubrication effect that affects the dominant class of configurations. For activated systems it can provide an additional dissipative drag between collisions among the grains. When the medium plays an important role the systems is said to be wet. In the opposite limit it is said to be dry. Finally, it is possible for compact and activated components of a system to coexist as heterogeneous states. Here, only the simplest case of dry systems in fully activated flows are considered. These are referred to in the following as granular fluids.

Advances in the study of granular fluids have arisen from many communities, including chemical engineering, materials sciences, and physics. The additional academic and conceptual importance of granular matter as practical systems for exploring the relevance of many-body fluid methods is primarily for the physics community. Granular matter, viewed as a system of particles with inelastic interactions, provides new opportunities to test the qualitative and quantitative limits of many-body methods developed over the past century for atomic and molecular systems. This is the field of non-equilibrium statistical mechanics [3,4,5]. Granular matter provides a new testing ground for a reconsideration of the most fundamental concepts and tools [1], with the potential for enhanced understanding of their place in atomic and molecular systems as well.

Statistical mechanics addresses the difficult many-body problem of extracting macroscopic properties of experimental interest from the very large number of constituent particles. The results express these properties in terms of the fundamental "microscopic" features of these particles, such as mass, shape, degree of inelasticity, and collisional properties. In this way the properties of the vitamin pills, mustard seeds, and peanuts are distinguished at a fundamental level. Also, conceptual issues such as the limitations of a macroscopic description are exposed through the insistence on their logical evolution from the fundamental microdynamics. In the next section, the granular fluid is described as a system of particles interacting via pairwise additive, nonconservative forces. The microscopic dynamics of these particles leads to balance equations for the mass density, energy density, and momentum density. Their averages define the "hydrodynamic fields" which are candidates for a macroscopic, continuum mechanics description. These exact equations are described in Sect. "Macroscopic Balance Equations" and the need for "constitutive equations" to provide a closure is described. The origin of constitutive equations, and consequently the origin of hydrodynamics, is associated with the concept of "normal states" [16] in Sect. "Normal States and Hydrodynamics". The normal state for the case of small spatial deviations from homogeneity is constructed

formally in Sect. "Navier–Stokes Approximation", resulting in the constitutive equations for Navier–Stokes hydrodynamics [17]. This derivation also provides insight into the context in which such a description should hold, and differences from the corresponding results for a normal fluid are noted. Empirical evidence [18,19], simulations [20,21,22], and corresponding results from kinetic theory [23,24,25,26] support the applicability of this hydrodynamic description under appropriate conditions. Finally, the results are summarized in Sect. "Future Directions" and some comments on the outlook for future developments are offered.

The presentation here is focused on recent work of the author and his collaborators for application of statistical mechanics to explore hydrodynamics for a granular gas. Consequently, the references quoted are heavily weighted toward those developmental studies. Apologies are offered for the exclusion of the vast and important complementary literature on simulations, kinetic theory, and experiments also bearing on this topic. Many of these can be found in the list of Books and Reviews given here.

Granular Fluid and Its Statistical Mechanics

Nonequilibrium Statistical Mechanics

Consider a system of $N \gg 1$ identical grains (hereafter referred to as particles) in a volume V, whose initial positions $\{\mathbf{q}_i\}$ and velocities $\{\mathbf{v}_i\}, 1 \le i \le N$, are specified. The positions and velocities define a point in a $6N$ dimensional space denoted by $\Gamma \equiv \{\mathbf{q}_i, \mathbf{v}_i\}$, defining the microstate of the system. A macrostate is defined by a probability density $\rho(\Gamma)$ in this space, representing statistical rather than precise knowledge of the system. The field of statistical mechanics addresses properties of macrostates, based on the recognition that for very large N the details of microstates are neither experimentally accessible nor practically calculable. Properties of interest are represented by functions $A(\Gamma)$, and their values for a macrostate $\rho(\Gamma)$ are determined from the expectations

$$\langle A; \rho \rangle \equiv \int \mathrm{d}\Gamma \rho(\Gamma) A(\Gamma) \ . \tag{1}$$

In this section, a brief overview of the essential ingredients of nonequilibrium statistical mechanics is given, broadened from its usual form [3] to include granular matter.

The dynamics of macrostates is determined from the underlying dynamics of the microstates. The initial point Γ changes in time since the particles have velocities and move to new positions. They move in straight lines until one or more come within the force field of other particles, at which point their velocities change as well as their positions. The forces are taken to be pairwise additive, such that the total force on particle i is $\mathbf{F}_i = \sum_j \mathbf{F}_{ij}$, where $\mathbf{F}_{ij}$ is the force on particle i due to particle j. This does not mean that the interactions are pairwise sequential; three or more particles can interact simultaneously. The pair forces are restricted by Newton's third law, $\mathbf{F}_{ij} = -\mathbf{F}_{ji}$, with conservation of momentum. Otherwise quite general forces can be considered to represent the shape of the particles and their degree of inelasticity. It is assumed here that the force range vanishes outside a distance $\sigma/2$ from the center of each particle so that σ characterizes the size of the particles. Furthermore, the particles are taken to be strongly repulsive so that their mean maximum overlap d on collision is small compared to their size, $d/\sigma < 1$. However, their size can be large or small compared to the mean distance between particles $(V/N)^{1/3}$, depending on whether the density of the system is small or large, respectively. Most importantly for the purposes here, these forces do not conserve energy. This property captures the feature of real grains that center of mass kinetic energy is lost as they distort during pair collisions. Further details of the force law are not required at this point.

The dynamics consists of straight line motion along the direction of the velocity at time t (free streaming), until the force range of any pair of particles, say i, j, overlaps. The relative velocity $\mathbf{g}_{ij} = \mathbf{v}_i - \mathbf{v}_j$ of that pair changes according to Newton's second law for the chosen force law $\mathbf{F}_{ij}$. Subsequently, all particles continue to stream freely until another pair has a force range of overlap, and the collisional change is repeated for that pair. In this way a trajectory $\Gamma_t \equiv \{\mathbf{q}_1(t), \ldots, \mathbf{q}_N(t), \mathbf{v}_1(t), \ldots, \mathbf{v}_N(t)\}$ is generated for $t > 0$. This trajectory is unique and invertible. The statistical mechanics for a fluid of inelastic particles [9,10,11,12] is comprised of the dynamics just described, a macrostate specified in terms of a probability density $\rho(\Gamma)$, and a set of observables generically denoted by $A(\Gamma)$. The expectation value for an observable at time $t > 0$ for a state $\rho(\Gamma)$ given at $t = 0$ is defined by

$$\langle A(t); 0 \rangle \equiv \int \mathrm{d}\Gamma \rho(\Gamma) A(\Gamma_t) \equiv \int \mathrm{d}\Gamma \ \rho(\Gamma) \mathrm{e}^{tL} A(\Gamma) \tag{2}$$

where $A(t) = A(\Gamma_t)$, and $\Gamma_t \equiv \{\mathbf{q}_1(t), \ldots, \mathbf{q}_N(t), \mathbf{v}_1(t), \ldots, \mathbf{v}_N(t)\}$ is the phase point evolved to time t from $\Gamma = \Gamma_{t=0}$. The dynamics can be represented in terms of a generator L defined by the second equality of (2). There are two components to the generator, corresponding to the two steps of free streaming and velocity changes due to interactions

$$L = \sum_{i=1}^{N} \mathbf{v}_i \cdot \nabla_i + \frac{1}{2m} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \mathbf{F}_{ij} \cdot \left(\nabla_{\mathbf{v}_i} - \nabla_{\mathbf{v}_j} \right) \ . \tag{3}$$

An alternative equivalent representation of the dynamics is obtained by transferring the dynamics from the observable $A(\Gamma)$ to the state $\rho(\Gamma)$ by the definition

$$\int \mathrm{d}\Gamma\, \rho(\Gamma)\mathrm{e}^{tL}A(\Gamma) \equiv \int \mathrm{d}\Gamma\, \left(\mathrm{e}^{-t\overline{L}}\rho(\Gamma)\right) A(\Gamma) \equiv \int \mathrm{d}\Gamma\, \rho(\Gamma, t)A(\Gamma)\,. \tag{4}$$

The representation in terms of a dynamical state $\rho(\Gamma, t)$ is referred to as Liouville dynamics. Its generator $\overline{L}$ is the formal adjoint of L which is found to be

$$\overline{L} = L + \frac{1}{2m}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\left(\nabla_{\mathbf{v}_i} - \nabla_{\mathbf{v}_j}\right)\cdot \mathbf{F}_{ij}\,. \tag{5}$$

The difference between L and $\overline{L}$ arises because the forces are non-conservative and therefore depend on the relative velocities of each pair as well as their positions. Time correlation functions for two observables A and B are defined in a similar way

$$\langle A(t)B;0\rangle \equiv \int \mathrm{d}\Gamma\, \left(\mathrm{e}^{tL}A(\Gamma)\right)\rho(\Gamma)B(\Gamma) = \int \mathrm{d}\Gamma\, A(\Gamma)\left(\mathrm{e}^{-t\overline{L}}\rho(\Gamma)\right)\left(\mathrm{e}^{-tL}B(\Gamma)\right)\,. \tag{6}$$

or

$$\langle A(t)B;0\rangle \equiv \langle AB(-t);t\rangle\,. \tag{7}$$

In summary, averages like $\langle A(t);0\rangle$ and correlation functions $\langle A(t)B;0\rangle$ are the central properties of interest for a macroscopic description of physical systems. The microscopic dynamics can be represented in terms of the observables $A(\Gamma, t)$ or the states $\rho(\Gamma, t)$ which are determined from specified initial values and the equations

$$(\partial_t - L)\,A(\Gamma, t) = 0\,, \quad \left(\partial_t + \overline{L}\right)\rho(\Gamma, t) = 0\,. \tag{8}$$

In the following most of the analysis is done in terms of the states, and the associated equation of motion is known as the Liouville equation.

Liouville Equation and Cooling

For an isolated system, the total energy decreases monotonically due to the loss of energy on each pair collision. This is reflected in a decrease of the average kinetic energy of the particles between collisions and hence is referred to as collisional "cooling". The energy per particle at time t and its loss are

$$\epsilon\,(t) \equiv N^{-1}\langle E;t\rangle\,, \quad \omega(t) \equiv -\partial_t\epsilon\,(t) = N^{-1}\langle LE;t\rangle\,. \tag{9}$$

This cooling effect is common to all solutions to the Liouville equation and it is useful to separate the dynamics into that due to this cooling and the residual time dependence

$$\rho(\Gamma, t) \equiv \rho(\Gamma, \epsilon(t), t)\,. \tag{10}$$

The Liouville equation then can be written

$$\partial_t\rho(\Gamma, \epsilon, t)\,|_\epsilon + \left(-\omega\,(\epsilon, t)\,\partial_\epsilon + \overline{L}\right)\rho(\Gamma, \epsilon, t) = 0\,. \tag{11}$$

The time derivative is now taken at constant ϵ. The notation $\omega\,(\epsilon, t)$ reflects the fact that it is a linear functional of $\rho(\Gamma, \epsilon, t)$, from its definition (9). This is a useful form that isolates a primary effect of the nonconservative forces (cooling) from the residual dynamics that will be associated with relaxation of the spatial inhomogeneities of interest below. For notational simplicity (11) is written

$$\left(\partial_t + \overline{\mathcal{L}}\right)\rho(\Gamma, \epsilon, t) = 0\,, \quad \overline{\mathcal{L}} \equiv -\omega\,(\epsilon, t)\,\partial_\epsilon + \overline{L}\,. \tag{12}$$

The corresponding equation for observables is

$$(\partial_t - \mathcal{L})\,A(\Gamma, \epsilon, t) = 0\,, \quad \mathcal{L} \equiv -\partial_\epsilon\omega\,(\epsilon, t) + L\,. \tag{13}$$

where it is understood that ∂_ϵ operates on everything to its right.

Stationary Homogeneous State

An isolated normal fluid supports an equilibrium state. This is a stationary solution to the Liouville equation with translational invariance, the Gibbs states. From the discussion above it is clear that isolated granular fluids have no truly stationary state due to cooling. However, there is a "universal" homogeneous state similar to the Gibbs state in the sense that a wide class of homogeneous initial states rapidly approach this state, on the time scale of a few collisions per particle. It is simple in the sense that all of its time dependence is that associated with cooling

$$\rho_0(\Gamma, t) = \rho_0(\{\mathbf{q}_{ij}, \mathbf{v}_i\}, \epsilon(t))\,. \tag{14}$$

Here, $\mathbf{q}_{ij} = \mathbf{q}_i - \mathbf{q}_j$ so the solution also has translational invariance. In the representation (12) it is seen to be a stationary solution to the Liouville equation

$$\overline{\mathcal{L}}\rho_0 = 0\,. \tag{15}$$

There is no longer any explicit time dependence since $\omega\,(\epsilon, t) = \omega\,(\epsilon)$ and $\overline{\mathcal{L}} = -\omega\,(\epsilon)\,\partial_\epsilon + \overline{L}$ for this state. This solution is referred to as the homogeneous cooling state (HCS). Clearly, it is the close analogue of the Gibbs state for a normal fluid. It is an example of a "normal" state in the sense that all of its time dependence occurs through one of the hydrodynamic fields (the energy). This concept is sharpened below.

Macroscopic Balance Equations

The origins of a macroscopic description for a fluid are the balance equations for the average mass density $\langle m(\mathbf{r});t\rangle$, energy density $\langle e(\mathbf{r});t\rangle$, and momentum density $\langle \mathbf{g}(\mathbf{r});t\rangle$, where $\mathbf{r}$ denotes an arbitrary field point within the system [16]. These will be referred to as the hydrodynamic fields since they are the ones that are expected to obey the hydrodynamic equations under appropriate conditions. The phase functions $m(\Gamma,\mathbf{r})$, $e(\Gamma,\mathbf{r})$, and $\mathbf{g}(\Gamma,\mathbf{r})$ are well known and their explicit forms will not be needed here. They will be denoted collectively by $a_\alpha(\mathbf{r})$

$$a_\alpha(\mathbf{r},t) \leftrightarrow \{m(\mathbf{r}), e(\mathbf{r}), \mathbf{g}(\mathbf{r})\} \ . \tag{16}$$

It follows from (8) that they obey the microscopic balance equations [14]

$$\partial_t a_\alpha(\mathbf{r},t) = La_\alpha(\mathbf{r},t) = -\nabla\cdot\mathbf{b}_\alpha(\mathbf{r},t) - \delta_{\alpha 2} w(\mathbf{r},t) . \tag{17}$$

To obtain this result, it has been recognized that the quantity $-La_\alpha(\mathbf{r},t)$ can be written as the sum of a divergence $\nabla\cdot\mathbf{b}_\alpha(\mathbf{r},t)$ plus a remainder $w(\mathbf{r},t)$ that cannot be so represented. For a normal fluid $w(\mathbf{r},t)$ vanishes and (17) become the local conservation laws for mass, energy, and momentum. The $\mathbf{b}_\alpha(\mathbf{r},t)$ are the corresponding fluxes. This clarifies why $a_\alpha(\mathbf{r},t)$ are selected for a macroscopic description. Their time dependence is determined by the scale of the spatial gradients of the fluxes, and averages of the latter become small as the system approaches homogeneity. Consequently, on long time scales the $\langle a_\alpha(\mathbf{r});t\rangle$ are the only surviving dynamical variables, and it is under these conditions that these fields obey hydrodynamic equations. The mass and momentum are conserved for a granular fluid as well, but there is a loss of energy $w(\mathbf{r},t)$ due to the non-conservative forces. It is no longer obvious that the energy is still one of the slow variables since its time scale is coupled to $w(\mathbf{r},t)$ which does not become small for nearly homogeneous states. Thus, an additional requirement for the existence of a macroscopic description in terms of $\langle a_\alpha(\mathbf{r});t\rangle$ is that the time scale of $\langle e(\mathbf{r});t\rangle / \langle w(\mathbf{r});t\rangle$ must be larger than that for non-hydrodynamic properties. This issue is discussed further below.

The macroscopic balance equations follow from the averages of (17)

$$\partial_t \langle a_\alpha(\mathbf{r});t\rangle + \nabla\cdot\langle \mathbf{b}_\alpha(\mathbf{r});t\rangle = -\delta_{\alpha 2}\langle w(\mathbf{r});t\rangle \ . \tag{18}$$

These equations are formally exact, but of little practical use as they do not form a closed (self-deterministic) set of equations for $\langle a_\alpha(\mathbf{r});t\rangle$. Closure requires expressing the average flux $\langle \mathbf{b}_\alpha(\mathbf{r});t\rangle$ and energy loss $\langle w(\mathbf{r});t\rangle$ as functionals of the fields $\langle a_\alpha(\mathbf{r});t\rangle$. Such relationships are called "constitutive equations". The combination of the exact balance equations with some form of constitutive equations provides the most general definition of hydrodynamics.

Construction of the constitutive equations is simplified by extracting the effects of convection. The velocity $\mathbf{U}(\mathbf{r},t)$ of a cell at point $\mathbf{r}$ is defined in terms of the average momentum

$$\langle \mathbf{g}(\mathbf{r});t\rangle \equiv \langle m(\mathbf{r});t\rangle\,\mathbf{U}(\mathbf{r},t) . \tag{19}$$

The fluxes are functions of the positions and velocities $\mathbf{b}_\alpha(\mathbf{r}) = \mathbf{b}_\alpha(\mathbf{r};\{\mathbf{q}_i,\mathbf{v}_i\}) = \mathbf{b}_\alpha(\mathbf{r};\{\mathbf{q}_i,\mathbf{V}_i+\mathbf{U}(\mathbf{r})\})$, where the velocity in the local rest frame has been introduced, $\mathbf{V}_i = \mathbf{v}_i - \mathbf{U}(\mathbf{r},t)$. Then defining the microscopic flux in the rest frame by $\mathbf{b}'_\alpha(\mathbf{r}) = \mathbf{b}_\alpha(\mathbf{r};\{\mathbf{q}_i,\mathbf{V}_i\})$ it follows that the average flux has the form [3]

$$\begin{aligned}\langle \mathbf{b}_\alpha(\mathbf{r});t\rangle = {} & \langle \mathbf{b}'_\alpha(\mathbf{r});t\rangle + \mathbf{c}_{\alpha\eta}\mathbf{U}(\mathbf{r},t)\cdot\langle \mathbf{b}'_\eta(\mathbf{r});t\rangle \\ & + \mathbf{U}(\mathbf{r},t) d_\alpha(\{\langle a_\nu(\mathbf{r});t\rangle\}) \ .\end{aligned} \tag{20}$$

The first term is the flux of mass, energy, and momentum in a fluid element at rest, and represents the dissipative processes. The second and third terms are proportional to the flow velocity $\mathbf{U}(\mathbf{r},t)$ and are associated with convection. The coefficients of these terms are explicit functions of the fields $\langle a_\alpha(\mathbf{r});t\rangle$ (as is $\mathbf{U}(\mathbf{r},t)$). For a normal fluid, neglect of the rest frame fluxes leads to the perfect fluid Euler hydrodynamic equations. Hence, determination of the constitutive equations is reduced to expressing the rest frame fluxes and energy loss as functionals of the fields.

"Normal" States and Hydrodynamics

A hydrodynamic description is a closed set of equations for the hydrodynamic fields, $\langle a_\alpha(\mathbf{r});t\rangle$. This follows from the exact macroscopic balance equations if the energy loss and fluxes can be represented as functionals of these fields

$$\begin{aligned}\langle w(\mathbf{r});t\rangle &\to \omega(\mathbf{r} \mid \langle a_\alpha;t\rangle) , \\ \langle \mathbf{b}_\alpha(\mathbf{r});t\rangle &\to \boldsymbol{\beta}_\alpha(\mathbf{r} \mid \langle a_\alpha;t\rangle) .\end{aligned} \tag{21}$$

The arrow is used to indicate that such a functional representation need not be valid on all length and time scales, and any such restrictions constitute the domain of validity for hydrodynamics. The notation here and below is such that $f(\mathbf{r},t,\{\langle a_\alpha(\mathbf{r});t\rangle\})$ denotes a *function* of $\mathbf{r}$, t and of the fields $\langle a_\alpha(\mathbf{r});t\rangle$ at the point $\mathbf{r}$, while $f(\mathbf{r},t \mid \langle a_\alpha;t\rangle)$ denotes a function of $\mathbf{r}$, t and a *functional* of the $\langle a_\alpha;t\rangle$ at all space points. With such constitutive re-

lations the macroscopic balance Eqs. (18) become hydrodynamic equations

$$\partial_t \langle a_\alpha(\mathbf{r}); t\rangle + \nabla \cdot \boldsymbol{\beta}(\mathbf{r} \mid \langle a_\alpha; t\rangle) = -\delta_{\alpha 2}\omega(\mathbf{r} \mid \langle a_\alpha; t\rangle) . \tag{22}$$

The average energy loss and fluxes are averages of specific functions of the particle positions and velocities, and hence are linear functionals of the solution to the Liouville equation. The existence of constitutive equations is therefore related to a special property of the solution which will be called "normal" (this terminology originates in a related context for derivation of hydrodynamics from the Boltzmann kinetic equation [3]). The class of "normal" distributions is defined by the functional forms

$$\rho_{\mathrm{n}}(\Gamma, t) = \rho_{\mathrm{n}}\left(\{\mathbf{q}_{ij}, \mathbf{v}_i\} \mid \langle a_\alpha; t\rangle\right) . \tag{23}$$

All time dependence and all the breaking of translational invariance for normal states occurs only through the hydrodynamic fields. A familiar example of a normal distribution for real fluids is the *local* Gibbs distribution

$$\rho_{e\ell}(\Gamma \mid \langle a_\alpha; t\rangle) = \exp\left\{q - \int \mathrm{d}\mathbf{r} y_\alpha(\mathbf{r}, t)\, a_\alpha(\mathbf{r})\right\} . \tag{24}$$

Here q is a normalization constant, and $y_\alpha(\mathbf{r}, t)$ are conjugate fields determined by the requirement that the averages of $a_\alpha(\mathbf{r})$ give the specified values $\langle a_\alpha(\mathbf{r}); t\rangle$. In this way $y_\alpha(\mathbf{r}, t)$ are functionals of the hydrodynamic fields and $\rho_{e\ell}(\Gamma \mid \langle a_\alpha; t\rangle)$ is normal. The importance of normal solutions is that they yield directly the desired functionals of (21)

$$\omega(\mathbf{r} \mid \langle a_\alpha; t\rangle) = \int \mathrm{d}\Gamma \rho_{\mathrm{n}}(\Gamma \mid \langle a_\alpha; t\rangle)\, w(\mathbf{r}) \tag{25}$$

$$\boldsymbol{\beta}_\alpha(\mathbf{r} \mid \langle a_\alpha; t\rangle) = \int \mathrm{d}\Gamma \rho_{\mathrm{n}}(\Gamma \mid \langle a_\alpha; t\rangle)\, \mathbf{b}_\alpha(\mathbf{r}) . \tag{26}$$

The normal state in (25) and (26) must be a solution to the Liouville equation. In general, the time derivative in the Liouville equation can be separated into that which occurs through $\langle a_\alpha; t\rangle$ plus the residual time dependence, generalizing (10)

$$\rho(\Gamma, t) = \rho(\Gamma, t \mid \langle a_\alpha; t\rangle) . \tag{27}$$

The Liouville equation then becomes

$$\partial_t \rho \mid_{\langle a_\alpha; t\rangle} - \int \mathrm{d}\mathbf{r} \frac{\delta\rho}{\delta\langle a_\alpha(\mathbf{r}); t\rangle} \cdot \{\nabla \cdot \langle \mathbf{b}_\alpha(\mathbf{r}); t\rangle + \delta_{\alpha 2}\langle w(\mathbf{r}); t\rangle\} + \overline{L}\rho = 0 . \tag{28}$$

A normal solution results when $\partial_t \rho_{\mathrm{n}} \mid_{\langle a_\alpha; t\rangle} \to 0$. For specified fields, (28) becomes an equation for the Γ dependence of the normal phase space density as a functional of the fields. This dependence then allows determination of the normal forms in (25) and (26). Finally, with the form of the hydrodynamic equation determined at that point, their solution with suitable initial and boundary conditions provides the explicit forms for the fields, and completes the normal solution. The existence and determination of this solution is the central problem for establishing a hydrodynamic description for both normal and granular fluids.

The concept of a normal solution and its use in the macroscopic balance equations makes no special reference to whether the fluid is atomic or granular, and is not restricted to states near homogeneity. In this general context, hydrodynamics is not a simple set of local partial differential equations such as the familiar Navier–Stokes equations. The latter are a special case of this more general idea, and their inadequacy for some conditions should not be interpreted as the absence of a more complex hydrodynamic description.

In closing this Section a qualitative explanation of why a normal solution can be expected is given, by analogy with the similar expectation for atomic fluids. For a wide class of initial states there is a first stage of rapid velocity relaxation in each small region toward the universal homogeneous state (HCS or Gibbs, respectively). However, the hydrodynamic parameters of that universal state are specific to each region so it is only locally homogenous, as in (24) for the atomic fluid. Subsequently, these differences in the parameters of neighboring cells are decreased by the fluxes of mass, energy, and momentum across their boundaries. It is this second stage where a normal description in terms of the hydrodynamic fields can be expected, indicating also that the space and time scales for a hydrodynamic description should be large compared to those of the first stage. This basic conceptual picture is essentially the same for both atomic and granular fluids, and the rapid approach of the first stage is indeed observed in molecular dynamics simulation studies of both equilibrium and the HCS.

Navier–Stokes Approximation

Equation (28) presents a formidable problem and further progress requires specialization to specific cases of interest. Perhaps the simplest of these are weakly inhomogeneous states. These are states for which all spatial gradients of first order are small and all higher order derivatives are negligible. Small gradients means that the relative change in the hydrodynamic fields over the largest micro-

scopic length scale ℓ_0 is small: $\ell_0 \partial_r \ln \langle a_\alpha; t\rangle \ll 1$. There are two characteristic length scales, the mean free path and the grain diameter. For a dilute gas the mean free path is largest, while for a dense fluid the grain size is largest. Under these conditions a solution to the Liouville equation can be sought as an expansion to leading order in these small gradients. This will be referred to as the Navier–Stokes approximation.

According to the discussion at the end of the last section, a normal solution is expected after the system has relaxed to its local HCS form, denoted by $\rho_{0\ell}\left(\Gamma \mid \langle a_\alpha; t\rangle\right)$, representing the fluid as having each cell in its own HCS. Define the deviations of the hydrodynamic fields from some common reference value by

$$\delta\left\langle a_\alpha; t\right\rangle = \left\langle a_\alpha; t\right\rangle - a_{0\alpha}\,, \tag{29}$$

where $a_{0\alpha}$ is the same for all cells. Then, the local HCS must satisfy the conditions

$$\rho_{0\ell}\left(\Gamma \mid a_{0\alpha} + \delta\left\langle a_\alpha; t\right\rangle\right)|_{\delta\langle a_\alpha;t\rangle=0} = \rho_0\left(\Gamma; a_{0\alpha}\right)\,, \tag{30}$$

$$\frac{\partial \rho_0}{\partial a_{0\alpha}} = \int \mathrm{d}\mathbf{r} \frac{\delta \rho_{0\ell}\left(\Gamma \mid a_{0\alpha} + \delta\left\langle a_\alpha; t\right\rangle\right)}{\delta\left\langle a_\alpha\left(\mathbf{r}\right); t\right\rangle}|_{\delta\langle a_\alpha;t\rangle=0}, \cdots \tag{31}$$

i. e., the local HCS and all of its functional derivatives must agree with those of the HCS in the homogenous limit. Also, as a normal distribution its time dependence is through the exact hydrodynamic fields for the fluid state considered. This means the averages of the corresponding microscopic fields $a_\alpha\left(\mathbf{r}\right)$ for the local HCS and for the solution to the Liouville equation must be the same

$$\int \mathrm{d}\Gamma\left(\rho - \rho_{0\ell}\right) a_\alpha\left(\mathbf{r}\right) = 0\,. \tag{32}$$

A more complete discussion of the construction of $\rho_{0\ell}$ from knowledge of ρ_0 is given elsewhere [14]. For the purposes here properties (30), (31), and (32) are sufficient.

The local HCS distribution, $\rho_{0\ell}$, is not a solution to the Liouville equation except in limit that all hydrodynamic fields become the same for each cell. Instead, it is a reference state approximating the actual solution after its first stage of velocity relaxation. To construct a solution ρ define its deviation from $\rho_{0\ell}$ by

$$\rho\left(\Gamma, t \mid \langle a_\alpha; t\rangle\right) = \rho_{0\ell}\left(\Gamma, t \mid \langle a_\alpha; t\rangle\right) + \Delta\left(\Gamma, t \mid \langle a_\alpha; t\rangle\right)\,. \tag{33}$$

The Liouville Equation (28) gives

$$\begin{aligned}\partial_t \Delta &- \int \mathrm{d}\mathbf{r}' \frac{\delta\Delta}{\delta\left\langle a_\alpha\left(\mathbf{r}'\right); t\right\rangle} \\ &\cdot \left\{\nabla\cdot\left\langle \mathbf{b}_\alpha\left(\mathbf{r}\right); t\right\rangle + \delta_{\alpha 2}\left\langle w\left(\mathbf{r}\right); t\right\rangle\right\} + \overline{L}\Delta \\ &= \int \mathrm{d}\mathbf{r}' \frac{\delta\rho_{0\ell}}{\delta\left\langle a_\alpha\left(\mathbf{r}'\right); t\right\rangle} \\ &\cdot \left\{\nabla\cdot\left\langle \mathbf{b}_\alpha\left(\mathbf{r}\right); t\right\rangle + \delta_{\alpha 2}\left\langle w\left(\mathbf{r}\right); t\right\rangle\right\} - \overline{L}\rho_{0\ell}\,.\end{aligned} \tag{34}$$

This equation is still exact, but if only small gradient states are considered it simplifies by retaining terms only of first order in the gradients. To be precise, the ultimate use of this solution is to calculate local properties of the form

$$\begin{aligned}&A(\mathbf{r}, t \mid \{y_\alpha\left(t\right)\}) \\ &= \int \mathrm{d}\Gamma\, a(\Gamma, \mathbf{r})\rho\left(\Gamma, t \mid \langle a_\alpha(\mathbf{r}); t\rangle + \delta\left\langle a_\alpha; t\right\rangle\right)\,.\end{aligned} \tag{35}$$

Therefore, in the following analysis the gradient expansions are referred to the field point $\mathbf{r}$ of interest, $\langle a_\alpha; t\rangle = \langle a_\alpha(\mathbf{r}); t\rangle + \delta\left\langle a_\alpha; t\right\rangle$, i. e. the common reference values in (29) are the exact field values at the chosen point, $a_{0\alpha} = \langle a_\alpha(\mathbf{r}); t\rangle$. The gradient expansion is carried out relative to these values. Of course the results will be general and applicable to any choice for $\mathbf{r}$.

The details of the gradient expansion are given in the Appendix. The solution to the Liouville equation to first order in the gradients is

$$\begin{aligned}\rho\left(\Gamma, t \mid \langle a_\alpha; t\rangle\right) &= \rho_0\left(\Gamma, \langle a_\alpha(\mathbf{r}); t\rangle\right) + (1-\mathcal{P}) \\ &\cdot \left(\mathbf{M}_\beta\left(\Gamma, \langle a_\alpha(\mathbf{r}); t\rangle\right) + \int_0^t \mathrm{d}t' \left(\mathrm{e}^{-\left(I\overline{\mathcal{L}}+K^{\mathrm{T}}\right)t'}\right)_{\beta\nu} \right. \\ &\left. \cdot (1-\mathcal{P})\,\boldsymbol{\Upsilon}_\nu\left(\Gamma, \langle a_\alpha(\mathbf{r}); t\rangle\right)\right) \cdot \nabla\left\langle a_\beta\left(\mathbf{r}\right); t\right\rangle\,.\end{aligned} \tag{36}$$

with the definitions

$$\mathbf{M}_\beta = \int \mathrm{d}\mathbf{r}' \left(\frac{\delta\rho_{0\ell}}{\delta\left\langle a_\alpha\left(\mathbf{r}'\right); t\right\rangle}\right)_{\delta\langle a_\alpha;t\rangle=0} \mathbf{r}'\,, \tag{37}$$

$$\boldsymbol{\Upsilon}_\alpha = -\left(I\overline{\mathcal{L}} + K^{\mathrm{T}}\right)_{\alpha\beta} \mathbf{M}_\beta\,. \tag{38}$$

The generator for the dynamics $I\overline{\mathcal{L}} + K^{\mathrm{T}}$ has a contribution from $\overline{\mathcal{L}}$ which is the same as in (12), with ω evaluated for the HCS as a function of the exact hydrodynamic fields at the point $\mathbf{r}$ and time t

$$\overline{\mathcal{L}} = -\omega_0(\langle a_\alpha\left(\mathbf{r}\right); t\rangle)\partial_{\langle e(\mathbf{r}); t\rangle} + \overline{L}\,. \tag{39}$$

The second contribution to the generator of the dynamics is the transpose of the matrix $K_{\alpha\beta}$

$$K_{\alpha\beta} = \delta_{\alpha 2} \frac{\partial \omega_0(\langle a_\alpha(\mathbf{r}); t\rangle)}{\partial \langle a_\beta(\mathbf{r}); t\rangle} . \tag{40}$$

Finally, $\mathcal{P}$ is a projection operator

$$\begin{aligned} \mathcal{P}X &= \Psi_\beta \int \mathrm{d}\Gamma A_\beta X , \\ A_\beta &= V^{-1} \int \mathrm{d}\mathbf{r} a_\alpha(\mathbf{r}) , . \\ \Psi_\beta &\equiv \frac{\partial \rho_0}{\partial \langle a_\beta(\mathbf{r}); t\rangle} \end{aligned} \tag{41}$$

The phase functions A_β and Ψ_β form a biorthogonal set in the sense

$$\int \mathrm{d}\Gamma A_\alpha \Psi_\beta = \delta_{\alpha\beta} . \tag{42}$$

The A_α are the usual global invariants of the Liouville operator $\overline{L}$ for a normal fluid; it is shown in the Appendix that the Ψ_β are the invariants of the new generator for dynamics in a granular fluid

$$\left(I\overline{\mathcal{L}}_T + K^{\mathrm{T}}\right)_{\nu\beta} \Psi_\beta = 0 . \tag{43}$$

Equation (36) is not quite the normal solution desired. All terms depend on time through $\langle a_\beta(\mathbf{r}); t\rangle$ as required, except for the last term which has an additional explicit time dependence through the upper limit of the time integral. This time dependence becomes negligible if the integrand is effectively non-zero after some short time scale τ. Then for $t \gg \tau$ the time integral becomes independent of t and can be taken formally to infinity. Thus, a normal solution is attained for this time scale

$$\begin{aligned} \rho_{\mathrm{n}}(\Gamma, \langle a_\alpha(\mathbf{r}); t\rangle) &= \rho_0(\Gamma, \langle a_\alpha(\mathbf{r}); t\rangle) \\ &\quad + (1-\mathcal{P}) \Big(\mathbf{M}_\beta(\Gamma, \langle a_\alpha(\mathbf{r}); t\rangle) \\ &\quad + \lim_{t_0\to\infty} \int_0^{t_0} \mathrm{d}t' \left(\mathrm{e}^{-(I\overline{\mathcal{L}}+K^{\mathrm{T}})t'}\right)_{\beta\nu} \\ &\quad \cdot (1-\mathcal{P}) \boldsymbol{\Upsilon}_\nu(\Gamma, \langle a_\alpha(\mathbf{r}); t\rangle) \Big) \\ &\quad \cdot \nabla \langle a_\beta(\mathbf{r}); t\rangle . \end{aligned} \tag{44}$$

It is expected that the integrand should have this property of a short time scale since the domain of operation for the generator of time dependence is functions with translational invariance (as a consequence of the gradient expansion). Hence there are no explicit slow hydrodynamic modes of finite wavelength. Also, there is no contribution from the homogeneous hydrodynamics (that for the invariants) due to the orthogonal projection $(1-\mathcal{P})$. The appearance of this projection is an essential self-consistency of the analysis, and occurs as well for normal fluids. The expression (44) is only formal and the actual limit should be taken in the weak sense only after (36) has been used to define average properties. A technical complication is the occurrence of periodic time dependence, the Poincare recurrence time. This can be removed by considering the thermodynamic limit of $V \to \infty, N \to \infty$ at constant N/V. Therefore, averages using the normal solution to the Liouville equation are understood as having the thermodynamic limit followed by the long time limit at constant $\langle a_\alpha(\mathbf{r}); t\rangle$.

An alternative equivalent form results from performing the integral in (44) using the explicit form (38) and the property (99) of the Appendix

$$\begin{aligned} \rho_{\mathrm{n}}(\Gamma, \langle a_\alpha(\mathbf{r}); t\rangle) &= \rho_0(\Gamma, \langle a_\alpha(\mathbf{r}); t\rangle) \\ &\quad + \lim_{t_0\to\infty} (1-\mathcal{P}) \left(\mathrm{e}^{-(I\overline{\mathcal{L}}+K^{\mathrm{T}})t_0}\right)_{\beta\nu} \\ &\quad \cdot \mathbf{M}_\nu(\Gamma, \langle a_\alpha(\mathbf{r}); t\rangle) \cdot \nabla \langle a_\beta(\mathbf{r}); t\rangle . \end{aligned} \tag{45}$$

The decay time for the integrand of (44) now becomes the time after which (45) reaches its normal form.

Constitutive Equations

The exact macroscopic balance equations are given by (22), and the necessary constitutive equations are given by (26) and (27) as averages over the normal solution. These can be made more explicit now using the small gradient result (44). Since the latter is a local function of the fields, the constitutive equations also will be local. Furthermore, since all components of the gradients in (45) depend on the common value $\langle \mathbf{g}(\mathbf{r}); t\rangle$, this can be eliminated through a Galilean transformation so that all properties refer to a fluid element at rest. Of course, the gradients of $\langle \mathbf{g}(\mathbf{r}); t\rangle$ in that fluid element are nonzero.

Consider first the energy loss function ω

$$\begin{aligned} \omega(\langle a_\alpha(\mathbf{r}); t\rangle) &= \int \mathrm{d}\Gamma \rho_{\mathrm{n}}(\Gamma, \langle a_\alpha(\mathbf{r}); t\rangle) w(\mathbf{r}) \\ &= \int \mathrm{d}\Gamma \rho_{\mathrm{n}}(\Gamma, \langle a_\alpha(\mathbf{r}); t\rangle) \overline{w} . \end{aligned} \tag{46}$$

The coefficients of the gradient in the normal solution have translational invariance and the average is independent of $\mathbf{r}$, except through its parametrization by $\langle a_\alpha(\mathbf{r}); t\rangle$.

The second equality takes this into account by replacing $w(\mathbf{r})$ by its average $\overline{w}$

$$\overline{w} = V^{-1} \int \mathrm{d}\mathbf{r} w(\mathbf{r}) \,. \tag{47}$$

Since $\omega(\langle a_\alpha(\mathbf{r}); t\rangle)$ is a scalar, fluid symmetry restricts the contributions to first order in the gradients to

$$\begin{aligned}&\omega(\langle a_\alpha(\mathbf{r}); t\rangle)\\ &= \omega_0(\langle a_\alpha(\mathbf{r}); t\rangle) + \omega_1(\langle a_\alpha(\mathbf{r}); t\rangle)\nabla\cdot\mathbf{U}(\mathbf{r}, t) \,.\end{aligned} \tag{48}$$

Here, the flow velocity $\mathbf{U}(\mathbf{r}, t)$ of (19) has been used in place of the momentum density. The first term is the contribution from the HCS distribution

$$\omega_0(\langle a_\alpha(\mathbf{r}); t\rangle) = \int \mathrm{d}\Gamma \rho_0(\Gamma, \langle a_\alpha(\mathbf{r}); t\rangle)\,\overline{w} \,. \tag{49}$$

The coefficient of $\nabla\cdot\mathbf{U}(\mathbf{r}, t)$ is

$$\begin{aligned}\omega_1(\langle a_\alpha(\mathbf{r}); t\rangle) &= \lim_{t_0\to\infty} C_\omega(t_0, \langle a_\alpha(\mathbf{r}); t\rangle)\\ &= C_\omega(0, \langle a_\alpha(\mathbf{r}); t\rangle)\\ &\quad + \lim_{t_0\to\infty}\int_0^{t_0} \partial_{t'} C_\omega(t', \langle a_\alpha(\mathbf{r}); t\rangle)\mathrm{d}t'\end{aligned} \tag{50}$$

with the correlation function defined by

$$C_\omega(t) = \int \mathrm{d}\Gamma\,\overline{w}\,(1-\mathcal{P})\,\mathrm{e}^{-\overline{\mathcal{L}}t} M_U, \tag{51}$$

$$M_U \equiv \frac{1}{3}\int \mathrm{d}\mathbf{r}'\mathbf{r}'\cdot\left(\frac{\delta\rho_{0\ell}}{\delta\mathbf{U}(\mathbf{r}', t)}\right)_{\delta\langle a_\alpha; t\rangle=0} \,. \tag{52}$$

The coefficient ω_0 defines an "equation of state" for the granular hydrodynamics, and gives the first non-trivial result of this analysis. It is similar to the pressure (given below) and is an inherent property of the local state of each cell, independent of the gradients between cells. In contrast, ω_1 is a true transport coefficient characterizing communication between cells. The first equality of (50) provides the Helfand form for this coefficient, while the second equality gives the equivalent Green–Kubo form. Each has its practical utility, depending on the method used for its approximate evaluation. Both forms have proven useful for normal fluids, and further discussion is provided below. Both ω_0 and the transport coefficient ω_1 vanish for normal fluids since they characterize collisional energy loss.

The fluxes $\boldsymbol{\beta}_\alpha$ of (26) can be determined in a similar way. As indicated in (20), only the rest frame flux $\boldsymbol{\beta}'_\alpha$ is required. Furthermore, since all components of the gradients in (45) depend on the common value $\langle\mathbf{g}(\mathbf{r}); t\rangle$, this can be eliminated through a Galilean transformation so that all properties refer to a fluid element at rest. Of course, the gradients of $\langle\mathbf{g}(\mathbf{r}); t\rangle$ in that fluid element are nonzero. The component $\boldsymbol{\beta}'_1$ is the rest frame mass flux which is expected to vanish in order to give the continuity equation. This follows from the fact that $\mathbf{b}'_1(\mathbf{r})$ is the microscopic momentum density

$$\begin{aligned}\boldsymbol{\beta}'_1 &= \int \mathrm{d}\Gamma\,\mathbf{g}(\mathbf{r})\,\rho_0(\Gamma, \langle a_\alpha(\mathbf{r}); t\rangle)\\ &\quad + \lim_{t_0\to\infty}\int \mathrm{d}\Gamma\,\mathbf{g}(\mathbf{r})\,(1-\mathcal{P})(\cdots) = 0\end{aligned} \tag{53}$$

The first term vanishes since $\langle\mathbf{g}(\mathbf{r}); t\rangle = 0$ in the rest frame, and the second term vanishes since $(1-\mathcal{P})$ projects orthogonal to the mass, energy, and momentum. Thus, the expected continuity equation is verified.

The fluxes $\boldsymbol{\beta}'_2$ and $\boldsymbol{\beta}'_\alpha$ for $\alpha = 3, 4, 5$ are the rest frame energy and momentum fluxes. The energy flux transforms like a vector and therefore fluid symmetry (translational and rotational invariance) requires that it can depend only on gradients of scalars

$$\begin{aligned}\boldsymbol{\beta}'_2 &= -\lambda(\langle a_\alpha(\mathbf{r}); t\rangle)\,\nabla T(\mathbf{r}, t)\\ &\quad - \mu(\langle a_\alpha(\mathbf{r}); t\rangle)\,\nabla\langle m(\mathbf{r}); t\rangle \,.\end{aligned} \tag{54}$$

To make the connection with Fourier's law for an atomic fluid, a temperature $T(\mathbf{r}, t)$ has been introduced through the definition

$$\langle e(\mathbf{r}); t\rangle \equiv e_0(\langle m(\mathbf{r}); t\rangle, T(\mathbf{r}, t)). \tag{55}$$

For an atomic fluid the function $e_0(\langle m(\mathbf{r}); t\rangle, T(\mathbf{r}, t))$ is chosen to be the thermodynamic internal energy density. As there is no thermodynamics for a granular fluid this function is arbitrary and simply constitutes a change of variables from $\langle e(\mathbf{r}); t\rangle, \langle m(\mathbf{r}); t\rangle$ to $T(\mathbf{r}, t)), \langle m(\mathbf{r}); t\rangle$. In this form (54) is a generalization of Fourier's law where λ is the thermal conductivity [27]. The contribution from the gradient of the mass density is new to granular fluids ($\mu = 0$ for atomic fluids). These coefficients are given by

$$\begin{aligned}\lambda(\langle a_\alpha(\mathbf{r}); t\rangle) &= \lim_{t_0\to\infty} C_\lambda(t_0, \langle a_\alpha(\mathbf{r}); t\rangle)\\ &= C_\lambda(0, \langle a_\alpha(\mathbf{r}); t\rangle)\\ &\quad + \lim_{t_0\to\infty}\int_0^{t_0} \partial_{t'} C_\lambda(t', \langle a_\alpha(\mathbf{r}); t\rangle)\mathrm{d}t'\end{aligned} \tag{56}$$

$$\begin{aligned}\mu(\langle a_\alpha(\mathbf{r}); t\rangle) &= \lim_{t_0\to\infty} C_\mu(t_0, \langle a_\alpha(\mathbf{r}); t\rangle)\\ &= C_\mu(0, \langle a_\alpha(\mathbf{r}); t\rangle)\\ &\quad + \lim_{t_0\to\infty}\int_0^{t_0} \partial_{t'} C_\mu(t', \langle a_\alpha(\mathbf{r}); t\rangle)\mathrm{d}t'\end{aligned} \tag{57}$$

with the correlation functions

$$C_\lambda(t) = \frac{1}{3}\int \mathrm{d}\Gamma \boldsymbol{\beta}'_2 \cdot (1-\mathcal{P})\,\mathrm{e}^{-(\overline{\mathcal{L}}+K_{22})t}\mathbf{M}_T\,, \tag{58}$$

$$C_\mu(t) = \frac{1}{3}\int \mathrm{d}\Gamma \boldsymbol{\beta}'_2 \cdot (1-\mathcal{P})\left(\mathrm{e}^{-\overline{\mathcal{L}}t}\left(\mathbf{M}_m + \frac{K_{21}}{K_{22}}\mathbf{M}_T\right) + \mathrm{e}^{-(\overline{\mathcal{L}}+K_{22})t} \cdot \left(\frac{\partial e_0}{\partial\langle m(\mathbf{r});t\rangle} - \frac{K_{21}}{K_{22}}\right)\frac{\partial T}{\partial e_0}\,|_{\langle m(\mathbf{r});t\rangle}\,\mathbf{M}_T\right)\,, \tag{59}$$

$$\mathbf{M}_T \equiv \int \mathrm{d}\mathbf{r}'\mathbf{r}'\left(\frac{\delta\rho_{0\ell}}{\delta T(\mathbf{r}',t)}\right)_{\delta\langle a_\alpha;t\rangle=0}\,. \tag{60}$$

$$\mathbf{M}_m \equiv \int \mathrm{d}\mathbf{r}'\mathbf{r}'\left(\frac{\delta\rho_{0\ell}}{\delta\langle m(\mathbf{r}');t\rangle}\right)_{\delta\langle a_\alpha;t\rangle=0}\,. \tag{61}$$

Finally, the set of vectors $\boldsymbol{\beta}'_\alpha$ for $\alpha = 3,4,5$ define the pressure tensor $\boldsymbol{\beta}'_\alpha \Leftrightarrow P_{ij}$. Fluid symmetry then determines that it can couple only to the momentum gradients, or equivalently the flow velocity gradients, in the form

$$P_{ij} = p(\langle a_\alpha(\mathbf{r});t\rangle)\,\delta_{ij} - \eta(\langle a_\alpha(\mathbf{r});t\rangle) \cdot \left(\partial_i U_j(\mathbf{r},t) + \partial_j U_i(\mathbf{r},t) - \frac{2}{3}\delta_{ij}\nabla\cdot\boldsymbol{U}(\mathbf{r},t)\right) - \kappa(\langle a_\alpha(\mathbf{r});t\rangle)\,\delta_{ij}\nabla\cdot\boldsymbol{U}(\mathbf{r},t)\,. \tag{62}$$

The scalar function $p(\langle a_\alpha(\mathbf{r});t\rangle)$ is the pressure, now identified as

$$p(\langle a_\alpha(\mathbf{r});t\rangle) = \int \mathrm{d}\Gamma\rho_0(\Gamma,\langle a_\alpha(\mathbf{r});t\rangle)\,\beta'_{3x}\,. \tag{63}$$

The transport coefficients in (62) are the shear viscosity $\eta(\langle a_\alpha(\mathbf{r});t\rangle)$ and the bulk viscosity $\kappa(\langle a_\alpha(\mathbf{r});t\rangle)$ given by

$$\eta(\langle a_\alpha(\mathbf{r});t\rangle) = \lim_{t_0\to\infty} C_\eta(t_0,\langle a_\alpha(\mathbf{r});t\rangle) = C_\eta(0,\langle a_\alpha(\mathbf{r});t\rangle) + \lim_{t_0\to\infty}\int_0^{t_0}\partial_{t'}C_\eta(t',\langle a_\alpha(\mathbf{r});t\rangle)\mathrm{d}t' \tag{64}$$

$$\kappa(\langle a_\alpha(\mathbf{r});t\rangle) = \lim_{t_0\to\infty} C_\kappa(t_0,\langle a_\alpha(\mathbf{r});t\rangle) = C_\kappa(0,\langle a_\alpha(\mathbf{r});t\rangle) + \lim_{t_0\to\infty}\int_0^{t_0}\partial_{t'}C_\kappa(t',\langle a_\alpha(\mathbf{r});t\rangle)\mathrm{d}t' \tag{65}$$

with the correlation functions

$$C_\eta(t) = \int \mathrm{d}\Gamma\beta'_{3y}\cdot(1-\mathcal{P})\,\mathrm{e}^{-\overline{\mathcal{L}}t}M_\eta\,, \tag{66}$$

$$C_\kappa(t) = \int \mathrm{d}\Gamma\beta'_{3x}\cdot(1-\mathcal{P})\,\mathrm{e}^{-\overline{\mathcal{L}}t}M_\kappa\,, \tag{67}$$

$$M_\eta \equiv \int \mathrm{d}\mathbf{r}'x'\left(\frac{\delta\rho_{0\ell}}{\delta U_y(\mathbf{r}',t)}\right)_{\delta\langle a_\alpha;t\rangle=0}\,. \tag{68}$$

$$M_\kappa \equiv \int \mathrm{d}\mathbf{r}'y'\left(\frac{\delta\rho_{0\ell}}{\delta U_y(\mathbf{r}',t)}\right)_{\delta\langle a_\alpha;t\rangle=0}\,. \tag{69}$$

This completes the formal derivation of the constitutive equations leading to the nonlinear Navier–Stokes equations, including expressions for the cooling rate, energy flux, and pressure tensor including contributions up through first order in the gradients of the hydrodynamic fields. These expressions are functions of the hydrodynamic fields to be determined by their detailed many-body analysis of the correlation functions.

Green–Kubo Expressions

To contrast the results here with those for an atomic fluid, it is instructive to focus on the Green–Kubo forms for the transport coefficients [3]. These are given by the second equalities of (50), (56), (57), (64), and (65); the first equalities are the corresponding Helfand forms [28]. For atomic fluids there is no counter part to ω_1 and μ. However, there are Green–Kubo expressions for the thermal conductivity and the two viscosities. For the discussion here only the thermal conductivity is considered, whose Green–Kubo expression is

$$\lambda(\langle a_\alpha(\mathbf{r});t\rangle) = C_\lambda(0,\langle a_\alpha(\mathbf{r});t\rangle) + \lim_{t_0\to\infty}\int_0^{t_0}\partial_{t'}C_\lambda(t',\langle a_\alpha(\mathbf{r});t\rangle)\mathrm{d}t' \tag{70}$$

$$\partial_t C_\lambda(t) = \frac{1}{3}\int \mathrm{d}\Gamma\boldsymbol{\beta}'_2\cdot(1-\mathcal{P})\,\mathrm{e}^{-(\overline{\mathcal{L}}+K_{22})t}\boldsymbol{\Upsilon}_\lambda\,, \tag{71}$$

$$\boldsymbol{\Upsilon}_e = -\left(\overline{\mathcal{L}}+K_{22}\right)\mathbf{M}_e\,. \tag{72}$$

In contrast the thermal conductivity for an atomic fluid is

$$\lambda(\langle a_\alpha(\mathbf{r});t\rangle) \to \lim_{t_0\to\infty}\int_0^{t_0}\partial_{t'}C_\lambda(t',\langle a_\alpha(\mathbf{r});t\rangle)\mathrm{d}t' \tag{73}$$

$$\partial_t C_\lambda(t) \to \frac{1}{3T^2}\int \mathrm{d}\Gamma\beta'_2\cdot(1-\mathcal{P})\,\mathrm{e}^{-Lt}\boldsymbol{\beta}'_2\rho_e\,, \tag{74}$$

In this last expression ρ_e is the equilibrium Gibbs ensemble, and it is understood that $\boldsymbol{\beta}'_2$ is the microscopic expression for the energy flux for a dynamics with conservative forces, and L is the Liouville generator for the corresponding dynamics.

There are several similarities and differences between the granular and atomic fluid expressions [14]. The latter is the time integral of a energy flux – energy flux equilibrium time correlation function. The granular fluid is similar, with one of the fluxes the same but the other flux is generated from the local HCS state. Also, the generator for the dynamics in the granular case has two additional effects, L replaced by $\overline{\mathcal{L}} + K_{22}$, to represents homogeneous cooling of the reference state and its homogeneous response to perturbations. The projection orthogonal to the invariants of each dynamics $(1 - \mathcal{P})$ occurs in both cases as a necessary condition for the long time limit of the time integral, and the corresponding existence of the normal state. Finally, the contribution from $C_\lambda(0, \langle a_\alpha(\mathbf{r}); t\rangle)$ vanishes for a normal fluid (except for singular forces) but is non-zero for the granular fluid due to the non-conservative forces.

Navier–Stokes Hydrodynamic Equations

In closing this section it is appropriate to record the results of substituting the Navier–Stokes constitutive equations, valid to first order in the gradients, into the exact macroscopic balance equations. This defines the Navier–Stokes hydrodynamic equations for a granular fluid

$$D_t m + m\nabla_{\mathbf{r}} \cdot \mathbf{U} = 0 \tag{75}$$

$$D_t e_0 + \omega_0 + \left(p + \omega_1 + \left(\frac{2}{3}\eta - \kappa\right)\nabla \cdot \mathbf{U}\right)\nabla \cdot \mathbf{U} - \eta\left(\partial_\alpha U_\beta + \partial_\beta U_\alpha\right)\partial_\alpha U_\beta - \boldsymbol{\nabla} \cdot (\lambda \boldsymbol{\nabla} T + \mu \boldsymbol{\nabla} m) = 0\,, \tag{76}$$

$$D_t U_\alpha + m^{-1}\partial_\alpha\left(p - \left(\frac{2}{3}\eta + \kappa\right)\nabla \cdot \mathbf{U}\right) - m^{-1}\partial_\beta \eta\left(\partial_\alpha U_\beta + \partial_\beta U_\alpha\right) = 0\,. \tag{77}$$

For simplicity of notation, $m \equiv \langle m(\mathbf{r}); t\rangle$ in these equations. They are a set of five nonlinear partial differential equations for the variables m, e_0, and $\mathbf{U}$. They are a closed set of equations since ω_0, p, and the transport coefficients ω_1, λ, μ, η, and κ are defined as functions of these variables. These definitions for the constitutive equations are the primary accomplishment of the statistical mechanical basis for the hydrodynamic equations. The form of (64)–(65) could have been guessed from the outset based on the macroscopic balance equations and fluid symmetry. The underlying basis in the microdynamics of the particles provides the necessary details for how the parameters of these equations must depend on the fields. Here only the formal definitions have been identified. It is only the first half of the problem of completing these equations, as the evaluation of these definitions poses a difficult many body problem. Still, without this first half, the starting point for that detailed analysis would not be possible. This the case for atomic fluids as well.

Future Directions

The objective here has been to formulate the basis for a macroscopic description of granular fluids using the fundamental principles of nonequilibrium statistical mechanics. The analysis presented follows that for an atomic fluid. First, the exact macroscopic balance equations are identified. Next, their closure is linked to the concept of a normal state and corresponding normal solution to the Liouville equation. This defines the domain of hydrodynamics in its most general sense, both for atomic and granular fluids. The construction of a normal solution is quite difficult in general, but can be accomplished for states with small gradients relative to locally homogeneous conditions. This gives the Navier–Stokes approximation described here.

Navier–Stokes hydrodynamics is applicable for most common states of atomic fluids, while deviations occur primarily for more complex polymeric molecular fluids. The latter have rheological properties corresponding to larger gradients relative to additional microscopic length and time scales. The construction of normal states in these cases is more difficult and is still at the semi-phenomenological stage [29]. Granular fluids provide a new motivation for renewed efforts to describe these more complex normal states. The reason is that even structurally simple granular fluids composed of spherically symmetric particles can exhibit rheology and other phenomena beyond the Navier–Stokes domain of validity [30,31] . This is due to the cooling rate in the energy balance equation which provides a new internal time scale, that can set the size of hydrodynamic gradients beyond any control through boundary conditions. For example, new steady states are possible for granular fluids due to the balance of this internal cooling with external forcing. In many cases this implies that the hydrodynamic description required is beyond the Navier–Stokes domain. The understanding of constitutive equations in these cases is poor at this point. It is hoped that the formal structure described here will provide the appropriate basis for studies of these problems.

The context of hydrodynamics depends on the formation of a normal state from more complex conditions. Above this has been described qualitatively as a two stage process of rapid velocity relaxation in each cell to a state near the local HCS, followed by hydrodynamic relaxation through exchange of mass, energy, and momentum between the cells on a longer time scale. This separation of microscopic and hydrodynamic time scales is essential to the dominance of the hydrodynamic excitations over all others at large space and time scales. It is justified for atomic fluids since the hydrodynamic times are determined by the wavelength of the phenomena studied. As the system approaches homogeneity, these time scales become much larger than the microscopic excitations and hydrodynamics prevails at large times. However, there is an additional hydrodynamic time scale for granular fluids, the cooling rate, which is not set by the wavelength alone. It would seem that this additional time scale must be large as well, implying a weak cooling rate. This condition is too strong. What matters is the rate of the approach to the homogeneous state, not any dynamics of that final state. In the above derivation of hydrodynamics the final form for the solution to the Liouville equation, Eq. (44) or (45), has a dynamics generated by $I\overline{\mathcal{L}} + K^{\mathrm{T}}$ rather than simply that for the trajectories $\overline{L}$. This is significant since the former has the additional compensation for the cooling and for the homogeneous perturbations of that cooling. Hence the approach to the time dependent normal state is determined only by the remaining non-hydrodynamic relaxation. The time scale for relaxation to the normal state is independent of the hydrodynamic time scales of that normal state. Quantitative verification of these concepts is another important future direction for research on a hydrodynamic description for granular fluids.

Acknowledgments

The author is indebted to Professor J. Javier Brey and Dr. Aparna Baskaran of Syracuse University for their collaboration on closely related linear response methods for granular fluids.

Appendix

Gradient expansion In this Appendix the Liouville equation in the form (34) is written to first order in the gradients and solved. Also the invariants of the associated dynamics are identified.

Consider first the right side of (34) which can be written equivalently as

$$\begin{aligned}
&\int \mathrm{d}\mathbf{r}' \frac{\delta\rho_{0\ell}}{\delta\langle a_\alpha(\mathbf{r}');t\rangle}\{\nabla\cdot\langle \mathbf{b}_\alpha(\mathbf{r});t\rangle + \delta_{\alpha 2}\langle w(\mathbf{r});t\rangle\} - \overline{L}\rho_{0\ell} \\
&= -\int \mathrm{d}\mathbf{r}' \frac{\delta\rho_{0\ell}}{\delta\langle a_\alpha(\mathbf{r}');t\rangle}\left\langle L a_\alpha(\mathbf{r}');t\right\rangle - \overline{L}\rho_{0\ell} \\
&= \int \mathrm{d}\mathbf{r}' \frac{\delta\rho_{0\ell}}{\delta\langle a_\alpha(\mathbf{r}');t\rangle}\int \mathrm{d}\Gamma\, a_\alpha(\mathbf{r})\,\overline{L}\,(\rho_{0\ell}+\Delta) - \overline{L}\rho_{0\ell}\,.
\end{aligned} \tag{78}$$

The first equality follows from (17) and (21). The first two terms are determined by the local HCS which can be expanded to first order in the gradients

$$\begin{aligned}
\rho_{0\ell} &= \rho_0(\langle a_\alpha(\mathbf{r});t\rangle) \\
&\quad + \int \mathrm{d}\mathbf{r}' \left(\frac{\delta\rho_{0\ell}}{\delta\langle a_\alpha(\mathbf{r}');t\rangle}\right)_{\delta\langle a_\alpha;t\rangle=0} \\
&\quad \cdot \left(\left\langle a_\alpha(\mathbf{r}');t\right\rangle - \langle a_\alpha(\mathbf{r});t\rangle\right) + \ldots \\
&= \rho_0(\langle a_\alpha(\mathbf{r});t\rangle) \\
&\quad + \mathbf{m}_\beta(\mathbf{r},\langle a_\alpha(\mathbf{r});t\rangle)\cdot\nabla\left\langle a_\beta(\mathbf{r});t\right\rangle + \ldots
\end{aligned} \tag{79}$$

The functional derivatives are

$$\begin{aligned}
\left(\frac{\delta\rho_{0\ell}}{\delta\langle a_\alpha(\mathbf{r}');t\rangle}\right)_{\delta\langle a_\alpha;t\rangle=0} &= \delta(\mathbf{r}'-\mathbf{r})\left(\frac{\partial\rho_0(\langle a_\alpha(\mathbf{r});t\rangle)}{\partial\langle a_\alpha(\mathbf{r});t\rangle}\right. \\
&\quad \left. + \frac{\partial\mathbf{m}_\beta(\mathbf{r},\langle a_\alpha(\mathbf{r});t\rangle)}{\partial\langle a_\alpha(\mathbf{r});t\rangle}\cdot\nabla\left\langle a_\beta(\mathbf{r});t\right\rangle\right) \\
&\quad + \mathbf{m}_\alpha(\mathbf{r},\langle a_\alpha(\mathbf{r});t\rangle)\cdot\nabla\delta(\mathbf{r}'-\mathbf{r}) + \ldots
\end{aligned} \tag{80}$$

Here,

$$\begin{aligned}
&\mathbf{m}_\beta(\mathbf{r},\langle a_\alpha(\mathbf{r});t\rangle) \\
&\quad \equiv \int \mathrm{d}\mathbf{r}' \left(\frac{\delta\rho_{0\ell}}{\delta\left\langle a_\beta(\mathbf{r}');t\right\rangle}\right)_{\delta\langle a_\alpha;t\rangle=0}(\mathbf{r}'-\mathbf{r})\,,
\end{aligned} \tag{81}$$

and $\rho_0(\langle a_\alpha(\mathbf{r});t\rangle)$ is the actual HCS with its global density, energy, and momentum evaluated at the common values $\langle a_\alpha(\mathbf{r});t\rangle$. It follows from (32) that the averages of $a_\alpha(\mathbf{r})$ for ρ, $\rho_{0\ell}$, and ρ_0 are all the same. This in turn gives

$$\int \mathrm{d}\Gamma\, a_\alpha(\mathbf{r})\,\mathbf{m}_\beta = 0 = \int \mathrm{d}\Gamma\, a_\alpha(\mathbf{r})\,\Delta\,. \tag{82}$$

With these results and the fact that Δ is of first order in the gradients, (78) to first order in the gradients becomes

$$\begin{aligned}
&\int \mathrm{d}\mathbf{r}' \frac{\delta\rho_{0\ell}}{\delta\langle a_\alpha(\mathbf{r}');t\rangle}\{\nabla\cdot\langle \mathbf{b}_\alpha(\mathbf{r});t\rangle + \delta_{\alpha 2}\langle w(\mathbf{r});t\rangle\} - \overline{L}\rho_{0\ell} \\
&\to \overline{\mathcal{L}}\rho_0 - (1-\mathcal{P})\left(I\overline{\mathcal{L}} + K^{\mathrm{T}}\right)_{\alpha\beta}\mathbf{m}_\beta\cdot\nabla\langle a_\alpha;t\rangle + \mathcal{P}\overline{\mathcal{L}}\Delta\,.
\end{aligned} \tag{83}$$

The matrix K^{T} is the transpose of K

$$K_{\alpha\beta} = \delta_{\alpha 2} \frac{\partial \omega(\langle a_\alpha(\mathbf{r}); t\rangle)}{\partial \langle a_\beta(\mathbf{r}); t\rangle} , \tag{84}$$

and I is the unit matrix. The generator $\overline{\mathcal{L}}$ is the same as that of (12) with $\omega \to \omega_0(\langle a_\alpha(\mathbf{r}); t\rangle)$ for the HCS evaluated at the common values $\langle a_\alpha(\mathbf{r}); t\rangle$

$$\overline{\mathcal{L}} = -\omega_0(\langle a_\alpha(\mathbf{r}); t\rangle)\partial_{\langle e(\mathbf{r}); t\rangle} + \overline{L} . \tag{85}$$

Finally, $\mathcal{P}$ is the projection operator

$$\mathcal{P}X = \frac{\partial \rho_0}{\partial \langle a_\alpha(\mathbf{r}); t\rangle} \int \mathrm{d}\Gamma\, a_\alpha(\mathbf{r})\, X . \tag{86}$$

The first term of (83) vanishes by definition of the HCS, ρ_0, confirming that the right side of the Liouville Eq. (34) is of first order in the gradients.

At this point, the Liouville Eq. (34) becomes

$$\partial_t \Delta - \int \mathrm{d}\mathbf{r}' \frac{\delta \Delta}{\delta \langle a_2(\mathbf{r}'); t\rangle} \omega_0(\langle a_\alpha(\mathbf{r}'); t\rangle) + \mathcal{P}\overline{L}\Delta = (1-\mathcal{P})\, \boldsymbol{\Upsilon}'_\alpha \cdot \nabla \langle a_\alpha; t\rangle , \tag{87}$$

$$\boldsymbol{\Upsilon}'_\alpha \equiv -\left(I\overline{\mathcal{L}} + K^{\mathrm{T}}\right)_{\alpha\beta} \mathbf{m}_\beta . \tag{88}$$

This equation is still exact up through contributions of first order in the gradients. It has solutions of the form

$$\Delta(\Gamma, t \mid \langle a_\alpha(\mathbf{r}); t\rangle) = \mathbf{G}_\nu(\Gamma, t, \langle a_\alpha(\mathbf{r}); t\rangle) \cdot \nabla \langle a_\nu(\mathbf{r}); t\rangle , \tag{89}$$

Substitution into (87) gives the corresponding equation for $\mathbf{G}_\nu$

$$\partial_t \mathbf{G}_\nu + (1-\mathcal{P})\left(I\overline{\mathcal{L}} + K^{\mathrm{T}}\right)_{\nu\beta} \mathbf{G}_\beta = (1-\mathcal{P})\, \boldsymbol{\Upsilon}'_\nu , \tag{90}$$

with the solution

$$\mathbf{G}_\nu(\Gamma, t, \langle a_\alpha(\mathbf{r}); t\rangle) = \int_0^t \mathrm{d}t' \left(\mathrm{e}^{-(1-\mathcal{P})(I\overline{\mathcal{L}}+K^{\mathrm{T}})t'}\right)_{\nu\beta} (1-\mathcal{P})\, \boldsymbol{\Upsilon}'_\beta . \tag{91}$$

It is possible to add to (87) an arbitrary solution to the homogeneous equation corresponding to (90). As described in the text, this represents the dynamics of the first stage of rapid velocity relaxation to the local HCS. The interest here is in the second stage where possible formation of a normal solution occurs. Hence, it is simpler to choose that stage for initial conditions (initial local HCS).

Define the derivatives of the HCS by

$$\Psi_\beta(\Gamma, \langle a_\alpha(\mathbf{r}); t\rangle) \equiv \frac{\partial \rho_0}{\partial \langle a_\beta(\mathbf{r}); t\rangle} . \tag{92}$$

Then differentiate the equation for ρ_0

$$\frac{\partial}{\partial \langle a_\beta(\mathbf{r}); t\rangle} \overline{\mathcal{L}} \rho_0 = 0 , \tag{93}$$

to get

$$\left(I\overline{\mathcal{L}}_T + K^{\mathrm{T}}\right)_{\nu\beta} \Psi_\beta = 0 . \tag{94}$$

Since $\left(I\overline{\mathcal{L}}_T + K^{\mathrm{T}}\right)$ is the generator for the dynamics in (91) this shows that Ψ_β are the invariants of that dynamics.

The projection operator $\mathcal{P}$ in (95) acts only on phase functions with translational invariance. In that case (86) simplifies to

$$\mathcal{P}X = \Psi_\beta \int \mathrm{d}\Gamma\, A_\beta X , \quad A_\beta = V^{-1} \int \mathrm{d}\mathbf{r}\, a_\alpha(\mathbf{r}) . \tag{95}$$

The first equality of (82) becomes $\mathcal{P}\mathbf{m}_\beta = 0$. This in turn gives

$$\begin{aligned} \mathbf{m}_\beta &= (1-\mathcal{P})\, \mathbf{m}_\beta = (1-\mathcal{P})\, \mathbf{M}_\beta , \\ \mathbf{M}_\beta &\equiv \int \mathrm{d}\mathbf{r}' \left(\frac{\delta \rho_{0\ell}}{\delta \langle a_\alpha(\mathbf{r}'); t\rangle}\right)_{\delta\langle a_\alpha; t\rangle = 0} \mathbf{r}' . \end{aligned} \tag{96}$$

Then $(1-\mathcal{P})\, \boldsymbol{\Upsilon}'_\alpha$ simplifies to

$$\begin{aligned} (1-\mathcal{P})\, \boldsymbol{\Upsilon}'_\alpha &= -(1-\mathcal{P})\left(I\overline{\mathcal{L}} + K^{\mathrm{T}}\right)_{\alpha\beta} (1-\mathcal{P})\, \mathbf{M}_\beta \\ &\equiv (1-\mathcal{P})\, \boldsymbol{\Upsilon}_\alpha \end{aligned} \tag{97}$$

with

$$\boldsymbol{\Upsilon}_\alpha = -\left(I\overline{\mathcal{L}} + K^{\mathrm{T}}\right)_{\alpha\beta} \mathbf{M}_\beta . \tag{98}$$

Use has been made of the identity

$$(1-\mathcal{P})\left(I\overline{\mathcal{L}} + K^{\mathrm{T}}\right)\mathcal{P} = 0 . \tag{99}$$

This same identity leads to a simplification of the dynamics in (91)

$$\mathrm{e}^{-(1-\mathcal{P})(I\overline{\mathcal{L}}+K^{\mathrm{T}})t'} (1-\mathcal{P}) = (1-\mathcal{P})\, \mathrm{e}^{-(I\overline{\mathcal{L}}+K^{\mathrm{T}})t'} (1-\mathcal{P}) . \tag{100}$$

In summary, the solution to the Liouville equation to first order in the gradients is

$$\begin{aligned} \rho(\Gamma, t \mid \langle a_\alpha; t\rangle) = {} & \rho_0(\Gamma, \langle a_\alpha(\mathbf{r}); t\rangle) \\ & + (1-\mathcal{P})\Big(\mathbf{M}_\beta(\Gamma, \langle a_\alpha(\mathbf{r}); t\rangle) \\ & + \int_0^t \mathrm{d}t' \left(\mathrm{e}^{-(I\overline{\mathcal{L}}+K^{\mathrm{T}})t'}\right)_{\nu\beta} \\ & \cdot (1-\mathcal{P})\, \boldsymbol{\Upsilon}_\beta(\Gamma, \langle a_\alpha(\mathbf{r}); t\rangle)\Big) \\ & \cdot \nabla \langle a_\nu(\mathbf{r}); t\rangle . \end{aligned} \tag{101}$$

Bibliography

Primary Literature

1. Kadanoff LP (1999) Built upon sand: Theoretical ideas inspired by granular flows. Rev Mod Phys 71:435–444
2. Haff PK (1983) Grain flow as a fluid mechanical phenomenon. J Fluid Mech 134:401–430
3. McLennan JA (1989) Introduction to Nonequilibrium Statistical Mechanics. Prentice-Hall, New Jersey
4. Hansen J-P, McDonald I (1986) Theory of Simple Liquids. Elsevier Press, London
5. Resibois P, De Leener M (1977) Classical Kinetic Theory of Fluids. Wiley, New York
6. See, for instance, Goldhirsch I, Tan ML, Zanetti G (1993) A molecular dynamical study of granular fluids: the unforced granular gas. J Sci Comput 8:1–40; McNamara S, Young WR (1996) Dynamics of a freely evolving, two-dimensional granular medium. Phys Rev E 53:5089–5100; Deltour P, Barrat JL (1997) Quantitative study of a freely cooling granular medium. J Phys I 7:137–151
7. Brilliantov N, Pöschel T (2004) Kinetic Theory of Granular Gases. Oxford, New York
8. Dufty JW (2001) Kinetic theory and hydrodynamics for a low density gas. Adv Complex Syst 4:397–407. cond-mat/0109215.201
9. Brey JJ, Dufty JW, Santos A (1997) Dissipative dynamics for hard spheres. J Stat Phys 87:1051–1066
10. Van Noije TPC, Ernst MH (2001) Kinetic theory of granular gases. In: Pöschel T, Luding S (eds) Granular Gases. Springer, New York
11. Dufty JW, Brey JJ, Lutsko J (2002) Diffusion in a granular fluid. I. Theory Phys Rev E 65:051303; Lutsko J, Dufty JW, Brey JJ (2002) Diffusion in a granular fluid. II. Simulation. Phys Rev E 65:051305; Dufty JW, Garzó V (2001) Mobility and diffusion in granular fluids. J Stat Phys 105:723–744
12. Dufty JW (2000) Statistical mechanics, kinetic theory, and hydrodynamics for rapid granular flow. J Phys: Condens Matter 12:A47–A56
13. Dufty JW, Baskaran A, Brey JJ (2006) Linear response for a granular fluid. JSTAT L08002:1–8
14. Dufty JW, Baskaran A, Brey JJ (2007) Linear response and hydrodynamics for granular fluids. Phys Rev E77, 031310; Baskaran A, Dufty JW, Brey JJ (2007) Transport coefficients for the hard sphere granular fluid. Phys Rev E77, 031311
15. See for example articles (2002) In: Halsey T, Metha A (eds) Challenges in Granular Physics. World Scientific, Singapore
16. Dufty J, Brey JJ (2005) Origins of hydrodynamics for a granular gas. In: Pareschi L, Russo G, Toscani G, (eds) Modelling and Numerics of Kinetic Dissipative Systems Nova Science, New York, pp 17–30, cond-mat/0410133
17. Here the Navier–Stokes approximation is defined by calculating the cooling rate, energy flux, and momentum flux to first order in the gradients. However, the fluxes occur under a gradient in the macroscopic balance equations while the cooling rate does not. Hence the equations themselves do not have all terms to second order in the gradients (i. e., the additional terms of second order contributing to the cooling rate).
18. Huan C, Yang X, Candela D, Mair RW, and Walsworth RL (2004) NMR experiments on a three-dimensional vibrofluidized granular medium. Phys Rev E 69:041302
19. Bizon C, Shattuck MD, Swift JB, Swinney HL (1999) Transport coefficients for granular media from molecular dynamics simulations. Phys Rev E 60:4340–4351; Rericha EC, Bizon C, Shattuck MD, Swinney HL (2001) Shocks in supersonic sand. Phys Rev Lett 88:014302
20. See, for instance, Brey JJ, Ruiz-Montero MJ, Cubero D (1999) On the validity of linear hydrodynamics for low-density granular flows described by the Boltzmann equation. Europhys Lett 48:359–364; Brey JJ, Ruiz-Montero MJ, Cubero D, García-Rojo R (2000) Self-diffusion in freely evolving granular gases. Phys Fluids 12:876–883; Garzó V, Montanero JM (2002) Transport coefficients of a heated granular gas. Physica A 313:336–356; Montanero JM, Santos A, Garzó V (2005) DSMC evaluation of the Navier–Stokes shear viscosity of a granular fluid. In: Capitelli M (ed) Rarefied Gas Dynamics 24 (AIP Conf Proc, vol 72), pp 797–802
21. Brey JJ, Ruiz-Montero MJ, Moreno F, Garcia-Rojo R (2002) Transversal inhomogeneities in dilute vibrofluidized granular fluids. Phys Rev E 65:061302; Brey JJ, Ruiz-Montero MJ, Moreno F (2001) Hydrodynamics of an open vibrated granular system. Phys Rev E 63:061305
22. Brey JJ, Ruiz-Montero MJ, Maynar P, Garzia de Soria MI (2005) Hydrodynamic modes, Green–Kubo relations, and velocity correlations in dilute granular gases. J Phys Cond Mat 17:S2489–S2502
23. Brey JJ, Dufty JW, Kim CS, Santos A (1998) Hydrodynamics for granular flow at low density. Phys Rev E 58:4638–4653; Sela N, Goldhirsch I (1998) Hydrodynamic equations for rapid flows of smooth inelastic spheres, to Burnett order. J Fluid Mech 361:41–74
24. Dufty JW, Brey JJ (2002) Green–Kubo expressions for a low density granular gas. J Stat Phys 109:433–448. cond-mat 0201361
25. Dufty JW (2005) Some aspects of the Boltzmann equation for granular gase. In: Capitelli M (ed) Rarefied Gas Dynamics (AIP Conf Proc 762, New York), pp 789–796
26. Dufty JW, Brey JJ (2005) Hydrodynamic modes for granular gases. Phys Rev E 68:030302; Brey JJ, Dufty JW (2005) Hydrodynamic modes for a granular gas from kinetic theory. Phys Rev E 72:011303
27. Dufty JW (2007) Fourier's law for a granular fluid. J Phys Chem B 111:15605–15612
28. Helfand E (1960) Transport coefficients from dissipation in a canonical ensemble. Phys Rev 119:1–9
29. Bird R, Armstrong R, Hassager O (1977) Dynamics of Polymeric Liquids. Wiley, New York
30. Santos A, Garzo V, Dufty JW (2004) Inherent rheology of a granular fluid in uniform shear flow. Phys Rev E 69:061303. cond-mat 0309320
31. Hrenya C (2007) (private communication, and to be published)

Books and Reviews

Campbell CS (1990) Rapid granular flows. Ann Rev Fluid Mech 22:57–92

Mehta A (ed) (1993) Granular Matter, An Interdisciplinary Approach. Springer, New York

Jaeger HM, Nagel SR, Behringer RP (1996) Granular solids, liquids, and gases. Rev Mod Phys 68:1259–1273

Duran J (2000) Sands, powders, and grains: an introduction to the physics of granular materials. Springer, New York

Pöschel T, Luding S (eds) (2001) Granular Gases. Springer, New York
Halsey T, Metha A (eds) (2002) Challenges in Granular Physics. World Scientific, Singapore
Campbell CS (2002) Granular shear flows in the elastic limit. J Fluid Mech 465:261–291
Pöschel T, Brilliantov N (eds) (2003) Granular Gases Dynamics. Springer, New York
Goldhirsch I (2003) Rapid granular flows. Annu Rev Fluid Mech 35:267–293
Brilliantov N, Pöschel T (2004) Kinetic Theory of Granular Gases. Oxford, New York
Hinrichsen H, Wolf D (eds) (2004) The Physics of Granular Media. Wiley-VCH, Berlin
Coniglio A, Fierro A, Herrmann H, Nicodemi M (eds) (2004) Unifying Concepts in Granular Media and Glasses. Elsevier, Amsterdam
Pöschel T, Schwager T (2005) Computational Granular Dynamics: Models and Algorithms. Springer, New York
Dufty JW (2007) Nonequilibrium statistical mechanics and hydrodynamics for a granular fluid. Six lectures at the Second Warsaw School on Statistical Physics. Kazimierz, Poland. arXiv:0707.3714

Granular Model for Data Mining

Anita Wasilewska[1], Ernestina Menasalvas[2]
[1] Computer Science Department, Stony Brook University, Stony Brook, USA
[2] Departamento de Lenguajes y Sistemas Informaticos, Facultad de Informatica, Madrid, Spain

Article Outline

Glossary
Definition of the Subject
Introduction
Granular Model: Syntax and Semantics for Data Mining
Semantic Model
Descriptive Model
Granular Model Revisited: Satisfaction and Truth
Future Directions
Bibliography

Glossary

Data mining Data mining is a process that includes the following phases: creating the target data, data preprocessing, data mining proper, pattern evaluation, and knowledge presentation.

Syntax Syntax, or syntactical concepts, refer to simple relations among symbols and expressions of formal, symbolic languages. A symbolic language is a pair $\mathcal{L} = (\mathcal{A}, \mathcal{E})$, where $\mathcal{A}$ is an alphabet and $\mathcal{E}$ is the set of expressions of $\mathcal{L}$. The expressions are often the formulae of the language, but can also represent sets of sequences of formulae, sets of clauses, or other sets built out of symbols of the alphabet, or their subsets. The expressions of formal languages, even if created with a specific meaning in mind, do not carry themselves any meaning, they are just finite sequences of certain symbols. The meaning is being assigned to them by establishing a proper semantics.

Semantics Semantics for a given symbolic language $\mathcal{L}$ assigns a specific interpretation in some domain to all symbols and expressions of the language. It also involves related ideas such as truth and model. They are called semantical concepts to distinguish them from the syntactical ones.

Model The word *model* is used in many situations and has many meanings but they all reflect some parts, if not all, of its following formal meaning. A structure M, called also an interpretation, is a model for a set $\mathcal{E}_0 \subseteq \mathcal{E}$ of expressions of $\mathcal{L}$ if and only if every expression $E \in \mathcal{E}_0$ it true in M.

Definition of the Subject

We present here three abstract models: descriptive, semantic, and granular. All of them are abstract structures that allow us to formalize some general properties of the data mining process and address the semantics-syntax duality inherent to any data mining process. A descriptive model formalizes the syntactical concepts and properties of the process. A semantic model formalizes its semantical properties. Finally, they form components of the granular model in which we establish a relationship between the descriptive and semantic models and provide a formal semantics for syntactical expressions of the language $\mathcal{L}$ of the descriptive model, thus justifying a use of the word model.

Moreover, we provide, within the granular model, a formal definition of data mining as the process of information generalization. The notion of generalization is defined in terms of granularity of steps of the process. In the model the data is represented in a form of knowledge systems. Each knowledge system has a granularity associated with it and the process changes, or not, its granularity. The notion of granularity is crucial for defining some notions and components of the model, hence the granular model name.

Introduction

Data mining, as defined in 1996 by Piatetsky-Shapiro [13] is a step (crucial, but a step nevertheless) in a KDD (knowledge discovery in databases) process. The Piatetsky-Shapiro's definition states that the KDD process consists of the following steps: developing an understanding of

the application domain; creating a target data set; choosing the data mining task i. e. deciding whether the goal of the KDD process is classification, regression, clustering, etc ...; choosing data preprocessing algorithms; choosing data mining algorithm(s); interpreting mined patterns; deciding if a re-iteration is needed; consolidating the discovered knowledge.

Since then the data mining (DM) term has evolved to become a name for all of the KDD process, or some parts of it, or even to be used as a name of an application of a data mining or learning algorithm.

In 1997 the Cross–Industry Standard Process for Data Mining (CRISP-DM) was proposed [15] to establish a standard for what they called, and others adopted, a data mining process. CRISP-DM standard was developed for business purposes and it included all of KDD process steps plus some extra steps such as a business understanding, business goal understanding followed by the KDD standard steps. Hence the KDD process became often a data mining process for industrial applications and was and is more and more often called just by the name of data mining.

To clarify these naming confusions we follow the standard terminology developed by data mining research in which we understand by data mining a KDD process in which its original data mining phase is now called *data mining proper phase*. For short we say that

Data mining (DM) is a process that includes between the others the following phases: creating the target data, data preprocessing, data mining proper, pattern evaluation, and knowledge presentation.

We formalize, within our granular model, the intuitive notions of the data mining preprocessing and data mining proper processes (Sect. "Data Mining Process").

One of the main goals of data mining is to *provide comprehensible descriptions* of information extracted from the data bases. The descriptions come in different forms. In the case of classification problems it might be a set of characteristic or discriminant rules, it might be a decision tree or a neural network with fixed set of weights. In the case of association analysis it is a set of associations (frequent itemsets), or association rules with accuracy parameters. In the case of cluster analysis it is a set of clusters, each of which has its own description and a cluster name. In the case of approximate classification by the rough set analysis it is usually a set of discriminant or characteristic rules (with or without accuracy parameters) or a set of decision tables.

If the number of descriptions is relatively small and all descriptions are of a reasonable length say that we obtained a generalized knowledge from the initial target database; that we mined more comprehensive, more general information. If an algorithm would reduce, for example, 100,000 records of the database of size 500 (number of attributes of the initial database) to 20 descriptions of size 10, we surely would say that we well generalized our information. Once it is done, a natural question of a *measure of quality* of such syntactic generalization and hence the validity of our method arises and this is being handled by our semantic model (Sect. "Semantic Model"). The format of the descriptions is defined in our descriptive model (Sect. "Descriptive Model"). The correctness and quality of syntactical descriptions is defined in our granular model by a satisfaction relation. The satisfaction relation (Sects. "K- Satisfaction and K- Truth", "Granular Model Revisited: Satisfaction and Truth") establishes a connection between syntactical and semantical approaches, and hence describe formally the syntax-semantic duality of the data mining process.

Granular Model: Syntax and Semantics for Data Mining

We usually view data mining results and present them to the user in their descriptive, i. e. syntactic form as it is the most natural form of communication. But the data mining process is deeply semantical in its nature. We hence build our granular model on two levels: syntactic and semantic. The syntactic level is represented by a descriptive model, and semantic level is described by the semantic model. The semantics-syntax duality of data mining process is expressed in our granular model by the satisfiability relation.

Moreover, we use a granular model to provide a formal definition of data mining as the process of information generalization (Sect. "Data Mining as Generalization"). In the model the data preprocessing and data mining algorithms are defined as certain operators that act on data represented in a form of knowledge systems (Definition 9). Each knowledge system has a granularity associated with it (Definition 10) and the operators change, or not, its granularity. The notion of granularity is crucial for defining all notions and components of the model, hence the granular model name.

Granular Model Definition

For a given data mining application one defines in detail all the components of the model (see Examples 7, 14). We provide here, as a part of the granular model definition, a general form of such definitions.

Definition 1 *A* **granular model** *is a system* $\mathbf{GM} = (\mathcal{SM}, \mathcal{DM}, \models)$ *where*:

- $\mathcal{S}$**M** *is a* **semantic model**;
- $\mathcal{D}$**M** *is a* **descriptive model**;
- $\models \subseteq \mathcal{P}(U) \times \mathcal{E}$ *is called* **a satisfaction relation**, *where* U *is the universe of* $\mathcal{S}$**M** *and* $\mathcal{E}$ *is the set of descriptions defined by the* $\mathcal{D}$**M**.

The satisfaction relation $\models$ (Definition 31) establishes the relationship between expressions of the semantic and descriptive models. It hence established formally the syntax-semantics duality of the data mining process. The models are $\mathcal{S}$**M** and $\mathcal{D}$**M** are defined in Definitions 13 and 21, respectively.

Motivation and Examples

The semantic model is the most important component of our granular model. The intuitions behind the definition of the semantic model are as follows. When we perform the data mining procedures the first step in any of them is to drop the key attribute. This step allows us to introduce similarities in the database as records do not have anymore their unique identification. The input into the data mining process is hence always a data table obtained from the target data by removal of the key attribute. We call it *a target data table*.

As the next step we represent, following the rough set model, our target data table as Pawlak's information system [11] with the universe U by adding a new, non attribute column for the record names, i. e. objects of U. We take this set U as the universe of our model of $\mathcal{S}$**M**.

Data mining, as it is commonly said, is a process of generalization. In order to model this process we have first to define what does it mean from semantical point of view that one stage of the process is more general then the other. The idea behind is very simple. It is the same as saying that $(a+b)^2 = a^2 + 2ab + b^2$ is a more general formula then the formula $(2+3)^2 = 2^2 + 2 \cdot 2 \cdot 3 + 3^2$. This means that one description (formula) is more general then the other if it describes more objects. From a semantical point of view it means that data mining process consists of putting objects (records) in sets of objects. From a syntactical point of view the data mining process consists of building descriptions (in terms of attribute, values of attributes pairs) of these sets of objects, with some extra parameters, if needed.

To model a situation that allows us to talk about descriptions of sets of records we extend the notion of Pawlak's model of an information system to our notion of a *knowledge system* (Definition 9). The universe the knowledge system consists of some subsets of U, i. e. elements of $\mathcal{P}(U)$. For example a target data table (after preprocessing) and the corresponding representation by Pawlak's information system, and a knowledge system with universe U of *granularity* one are as follows.

Example 2
Target Data Table T_0

a_1	a_2	a_3
small	small	medium
medium	small	medium
small	small	medium
big	small	small
medium	medium	big
small	small	medium
big	small	small
medium	medium	big
small	small	medium
big	small	medium
medium	medium	small
small	small	medium
big	small	big
medium	medium	small

Target Information System I_0

U	a_1	a_2	a_3
x_1	small	small	medium
x_2	medium	small	medium
x_3	small	small	medium
x_4	big	small	small
x_5	medium	medium	big
x_6	small	small	medium
x_7	big	small	small
x_8	medium	medium	big
x_9	small	small	medium
x_{10}	big	small	medium
x_{11}	medium	medium	small
x_{12}	small	small	medium
x_{13}	big	small	big
x_{14}	medium	medium	small

A *knowledge system* corresponding to target table T_0 is a system that all its objects are one element sets $\{x\}$ for corresponding $x \in U$ of T_0. We call such knowledge systems (Definition 11) a target knowledge system based on the target data table T_0. The knowledge systems with the property that all its objects are one element sets are called systems of *granularity one*.

Example 3
Target Knowledge System K_0

$\mathcal{P}^1(U)$	a_1	a_2	a_3
$\{x_1\}$	small	small	medium
$\{x_2\}$	medium	small	medium
$\{x_3\}$	small	small	medium
$\{x_4\}$	big	small	small
$\{x_5\}$	medium	medium	big
$\{x_6\}$	small	small	medium
$\{x_7\}$	big	small	small
$\{x_8\}$	medium	medium	big
$\{x_9\}$	small	small	medium
$\{x_{10}\}$	big	small	medium
$\{x_{11}\}$	medium	medium	small
$\{x_{12}\}$	small	small	medium
$\{x_{13}\}$	big	small	big
$\{x_{14}\}$	medium	medium	small

Assume now that we have applied some algorithm ALG_1 and it has returned a following set $\mathcal{D} = \{D_1, D_2, \ldots, D_7\}$ of descriptions (in the granular model, $D_i \in \mathcal{E}$ and $\mathcal{E}$ is defined in the descriptive model. We write s for the attribute value *small*, m for *medium* and b for *big*.

$$
\begin{aligned}
D_1 &: (a_1 = s) \cap (a_2 = s) \cap (a_3 = m)\,,\\
D_2 &: (a_1 = m) \cap (a_2 = s) \cap (a_3 = m)\,,\\
D_3 &: (a_1 = m) \cap (a_2 = m) \cap (a_3 = b)\,,\\
D_4 &: (a_1 = m) \cap (a_2 = m) \cap (a_3 = s)\,,\\
D_5 &: (a_1 = b) \cap (a_2 = s) \cap (a_3 = s)\,,\\
D_6 &: (a_1 = b) \cap (a_2 = s) \cap (a_3 = m)\,,\\
D_7 &: (a_1 = b) \cap (a_2 = s) \cap (a_3 = b)\,.
\end{aligned}
$$

Now, a natural question arises: how well this set of descriptions describes our original data i. e. how accurate is the algorithm ALG_1 we have used to find them, how accurate is the *knowledge* we have thus obtained out of our data. To answer this, we use the target information system with the universe U, and for any $D \in \mathcal{D}$ we examine a set $S(D) = x \in U : D$ called a truth set for D. We define it formally in the Definition 27. Intuitively, the sets $S(D) = \{x \in U : D\}$ contain all records (i. e. their identifiers) with the same description given in terms of attribute, values of attribute pairs. The descriptions do not need to utilize all attributes of the target data, as it is often the case, and one of ultimate goals of data mining is to find descriptions with as few attributes as possible.

In association analysis the descriptions can represent the frequent itemsets. For example, for a frequent three itemset $D = i_1 i_2 i_3$, the truth set $S(D)$ represents all transactions that contain items i_1, i_2, i_3.

Descriptions come in different forms, depending on the data mining goal and application. We define formally a general form of descriptions as a part of the descriptive model (Definition 21).

For the data presented in the Examples 2, 3 and their descriptions $D_i \in \mathcal{D}$ the sets $S(D_i)$ are as follows.

$$
\begin{aligned}
S_1 &= S(D_1) = \{x \in U : D_1\} = \{x_1, x_3, x_6, x_9, x_{12}\}\,,\\
S_2 &= S(D_2) = \{x \in U : D_2\} = \{x_2\}\,,\\
S_3 &= S(D_3) = \{x \in U : D_3\} = \{x_5, x_8\}\,,\\
S_4 &= S(D_4) = \{x \in U : D_4\} = \{x_{11}, x_{14}\}\,,\\
S_5 &= S(D_5) = \{x \in U : D_5\} = \{x_4, x_7\}\,,\\
S_6 &= S(D_6) = \{x \in U : D_6\} = \{x_{10}\}\,,\\
S_7 &= S(D_7) = \{x \in U : D_7\} = \{x_{13}\}\,.
\end{aligned}
$$

We represent our results in a form of a *knowledge system* as follows. We write, as before, s for the attribute value *small*, m for *medium* and b for *big* and we put the names of proper subsets of U in the second representation of K_1.

Example 4
Resulting Knowledge System K_1

$K_1(U)$	a_1	a_2	a_3
$\{x_1, x_3, x_6, x_9, x_{12}\}$	s	s	m
$\{x_2\}$	m	s	m
$\{x_5, x_8\}$	m	m	b
$\{x_{11}, x_{14}\}$	m	m	s
$\{x_4, x_7\}$	b	s	s
$\{x_{10}\}$	b	s	s
$\{x_{13}\}$	b	s	b

Resulting Knowledge System K_1

$K_1(U)$	a_1	a_2	a_3a
S_1	s	s	m
S_2	m	s	m
S_3	m	m	b
S_4	m	m	s
S_5	b	s	s
S_6	b	s	s
S_7	b	s	b

The representation of data mining results in a form of a knowledge system allows us to define how good is the *knowledge* obtained by a given algorithm. In our case the knowledge obtained describes 100% of our target data as

$$S_1 \cup S_2 \cup S_3 \cap \cdots \cup S_7 = \{x_1, x_2, \ldots, x_{14}\} = U\,.$$

Observe that the sets $S_1, \ldots, S_7$ are also disjoint and non-empty, i. e. they form a partition of the universe U). In such a case the knowledge system (and the algorithm) will be called *exact* (see Definition 12).

Moreover, we can see that the resulting system K_1 is *more general* then the input data, as represented by the target system K_0 because its *granularity* (i. e. maximum of cardinality of its granules, i. e. elements of its universe) is higher the granularity of K_0. This observation motivates the formal Definition 15. The granularity of our K_0 is one, as is the granularity of all target knowledge systems.

The *granularity* of K_1 is $\max\{|S_1|, \ldots, |S_7|\} = \max\{5, 1, 2, \} = 5$.

Let now assume that we have applied to out target data T (represented by K_0) another algorithm ALG_2 and it returned two descriptions D_1, D_2 under a condition that we need only descriptions of length 2, such that the corresponding attribute- value pairs appear in the target data with frequency $\geq 30\%$.

Example 5 Consider the following two descriptions D_1, D_2.

- $D_1 : (a_1 = s) \cap (a_2 = s)$, $D_2 : (a_2 = s) \cap (a_3 = m)$.
- Now we evaluate: $S_1 = S(D_1) = \{x_1, x_3, x_6, x_9, x_{12}\}$,
- $S_2 = S(D_2) = \{x_1, x_2, x_3, x_6, x_9, x_{10}, x_{12}\}$.

The description D_1 fulfills the algorithm condition because it's of length 2 and $\frac{|S_1|}{|U|} = \frac{5}{14} \sim 0.3508$. The frequency% is greater then 30%.

The description D_2 fulfills the algorithm condition because it's of length 2 and $\frac{|S_1|}{|U|} = \frac{7}{14} = 0.5$. The frequency% is hence greater then 30%.

Incorporating the parameters of # of attributes in the description and their frequency imposed by the ALG_2 into our knowledge system we obtain the following table. The table is incomplete, as there are more descriptions fulfilling the algorithm conditions.

Example 6
Knowledge System K_2

$K_2(U)$	a_1	a_2	a_3	#of attr	frequency%
S_1	s	s	–	2	35
S_2	–	s	m	2	50

The sets S_1, S_2 do not form a partition of the universe U as $S_1 \cap S_2 \neq \emptyset$ and moreover, $S_1 \cup S_2 \neq U$. The algorithm ALG_2 is hence *not exact*. It describes only 57% of the target data and what is described is described following the frequency conditions. Of course K_2 is more general then K_0 and the algorithm ALG_2 generalized the target data, even if in an incomplete way. Now we form a new set of descriptions D_3, D_4 associated with the resulting knowledge system K_2.

Example 7

- $D_3 : (a_1 = s) \cap (a_2 = s) \cap (\text{\#of attr} = 2) \cap (\text{frequency\%} = 36)$,
- $D_4 : (a_2 = s) \cap (a_3 = m) \cap (\text{\#of attr} = 2) \cap (\text{frequency\%} = 50)$.

We re-write D_3, D_4 as follows. $D_3 : D_1 \cap (\text{\#of attr} = 2) \cap (\text{frequency\%} = 36)$, $D_4 : D_2 \cap (\text{\#of attr} = 2) \cap (\text{frequency\%} = 50)$, where D_1, D_2 are descriptions considered in Example 5.

We say that the set $S_1 = S(D_1) = \{x \in U : D_1\}$ is *a truth set* for the description D_3 under restrictions defined by the descriptions $(\text{\#of attr} = 2) \cap (\text{frequency\%} = 36)$. The set $S_2 = S(D_2)$ is the truth set for D_4 under restrictions $(\text{\#of attr} = 2) \cap (\text{frequency\%} = 50)$.

Semantic Model

The definition of a semantic model presented here is a simpler and more comprehensible version of [17,18,21,22]. The initial investigations of the subject appeared also in [3,8,9,10]. The notion of a knowledge system is central to the definition and development of the semantic model, as we have seen in Sect. "Motivation and Examples". We define it formally in the next section. We next use the knowledge system to formalize a notion of a granule and granularity (Definition 10) that is central to our granular model.

Knowledge Systems and Granularity

Observe that the table depicting the knowledge system K_2 (Example 6) is incomplete, i. e. not all attributes have assigned values. This fact in the formal definition will be represented by a *partial function* g. Also K_2 contains some new *attributes describing properties of its granules*. In the formal definition we will call these new attributes the *knowledge attributes*.

The formal definitions of information system, knowledge and target knowledge systems, and their granularity and exactness are as follows.

Definition 8 *A Pawlak's* **information system** *is a system* $I = (U, A, V_A, f)$, *where* $U \neq \emptyset$ *is called a set of* **objects**; $A \neq \emptyset$, $V_A \neq \emptyset$ *are called the set of* **attributes** *and* **values of attributes**, *respectively*; $f : U \times A \longrightarrow V_A$ *is called an* **information** *function*.

Definition 9 *A* **knowledge system** *based on* $I = (U, A, V_A, f)$ *is a system*

$$K = (K(U), A, E, V_A, V_E, g)$$

where:

- *The* **universe** $K(U)$ *of* K *is a subset of the set* $\mathcal{P}(U)$ *of all subsets of the universe* U *of* I, *i. e.* $K(U) \subseteq \mathcal{P}(U)$;
- E *is a finite set of* **knowledge attributes** *(k-attributes) such that* $A \cap E = \emptyset$;
- V_E *is a finite set of* **values of k- attributes**;
- g *is a partial function called* **a knowledge function** *(k-function)*;
- $g : \mathcal{P} \times (A \cup E) \longrightarrow (V_A \cup V_E)$ *is such that*:
  (i) $g \mid (\bigcup_{x \in U}\{x\} \times A) = f$;
  (ii) $\forall S \in \mathcal{P}, \forall a \in A((S, a) \in \mathrm{dom}(g) \Rightarrow g(S, a) \in V_A)$;
  (iii) $\forall S \in \mathcal{P}, \forall e \in E((S, e) \in \mathrm{dom}(g) \Rightarrow g(S, e) \in V_E)$;

We use the above notion of a knowledge system to define the granules of the universe and the granularity of the system, an hence later, the granularity of the data mining process.

Definition 10 *Any set* $S \in \mathcal{P}(U)$ *i. e.* $S \subseteq U$ *is called a* **granule** *of* U. *The cardinality* $|S|$ *of* S *is called a* **granularity of** S. *The set*

$$Gr_K = \{S \in \mathcal{P} : \exists b \in (E \cup A)((S, b) \in \mathrm{dom}(g))\}$$

is called a **granule universe** *of* K. *A number* $gr_K = \max\{|S| : S \in Gr_K\}$ *is called a* **granularity** *of* K.

Observe that by conditions (ii), (iii) of Definition 9, $Gr_K = K(U)$ and it justifies its name. The condition (i) of Definition 9 says that when $E = \emptyset$, the k-function g is total on the set $\{\{x\} : x \in U\} \times A$ and $\forall x \in U \forall a \in A\,(g(\{x\}, a) = f(x, a))$. We denote $\mathcal{P}^1(U) = \{\{x\} : x \in U\}$.

Definition 11 *Let* $I = (U, A, V_A, f)$. *Any system* $K^1 = K^1 = (\mathcal{P}^1(U), A, \emptyset, V_A, \emptyset, g) = (\mathcal{P}^1(U), A, V_A, g) = (\mathcal{P}^1(U), A, V_A, g)$ *is called a* **target** *knowledge system based on* I.

Observe that any target knowledge system has granularity one. Finally, we define the *exactness of a knowledge system* as follows.

Definition 12 *A knowledge system* $K = (K(U), A, E, V_A, V_E, g)$ *is called* **exact** *if all its granules* Gr_K *form a partition of the universe* U.

The system K_2 from Example 4 is exact, the system K_3 from Example 6 is not exact.

In our model we view data mining algorithms as certain *operators*. For example, our ALG_1 is represented in the semantic model by an operator p_1 acting on some subset of a set $\mathcal{K}$ of knowledge systems, such that $p_1(K_0) = K_1$. ALG_2 is represented in the model by an operator p_2 also acting on some (may be different) subset of the set $\mathcal{K}$ of knowledge systems, such that $p_2(K_0) = K_2$. We put all the above observations into a formal notion of a semantic model.

Definition 13 *A* **semantic model** *is a system* $\mathbf{SM} = (\mathcal{P}(U), \mathcal{K}, \mathcal{G})$, *where*:

- $U \neq \emptyset$ *is the* **universe**;
- $\mathcal{K} \neq \emptyset$ *is a set of knowledge systems, called also data mining process states*;
- $\mathcal{G} \neq \emptyset$ *is the set of* **operators**;
- *Each operator* $p \in \mathcal{G}$ *is a partial function on the set of all data mining process states, i. e.* $p : \mathcal{K} \longrightarrow \mathcal{K}$.

The semantic model is always being built for a given application. The target data is represented first in a form the target information system with the universe U, as in Example 2 and then in the form of target knowledge system as in the Example 3.

The semantic model based on our Examples 2, 3, and 4 are as follows.

Example 14 $\mathbf{SM}_1 = (\mathcal{P}(U), \mathcal{K}, \mathcal{G})$, where: $U = \{x_1, x_2, \ldots, x_{14}\}$; $\mathcal{K} = \{K_0, K_1, K_2\}$; $\mathcal{G} = \{p_1, p_2\}$; Each $p_i \in \mathcal{G}$, $(i = 1, 2)$ is a partial function $p_i : \mathcal{K}_1 \longrightarrow \mathcal{K}_1$, such that $p_1(K_0) = K_1, p_2(K_0) = K_2$.

Data Mining as Generalization

In order to model within our semantic model data mining as a process of generalization we first introduce the following definition of *generalization relation* based on a notion of granularity.

Definition 15 *A relation* $\preceq \subseteq \mathcal{K} \times \mathcal{K}$ *is called a* **generalization relation** *if the following condition holds for any* $K, K' \in \mathcal{K}$.

$$K \preceq K' \text{ if and only if } \mathrm{gr}_K \leq gr_{K'} ,$$

where gr_K *denotes the granularity of* K *(Definition 10).*

Observe that for K_0, K_1, K_2 from Example 14, $gr_{K_0} = 1 \leq 5 = gr_{K_1} \leq 7 = gr_{K_2}$, and the system K_2 is the most general. But at the same time K_1 is exact and K_2 is not exact, so we have a trade off between exactness and generality.

As the next step we use the generalization relation to define the notion of a generalization operator as follows.

Definition 16 *An operator* $g \in \mathcal{G}$ *is called a* **generalization operator** *if for any* $K, K' \in \mathcal{K}$ *such that* $g(K) = K'$, *we have that* $K \preceq K'$.

Observe that both operators p_1, p_2 in Example 14 are generalization operators.

Data Mining Operators

In the data mining process the preprocessing and data mining are disjoint, inclusive/exlusive categories. The preprocessing is an integral and very important stage of the data mining process and needs as careful analysis as the data mining itself. Our framework allows us distinguish two disjoint classes of operators: the preprocessing operators $\mathcal{G}_{\text{prep}}$ and data mining proper operators $\mathcal{G}_{\text{dm}}$ and we put

$$\mathcal{G} = \mathcal{G}_{\text{prep}} \cup \mathcal{G}_{\text{dm}} .$$

The paper [20] contains detailed formal definitions, their motivation, and discussion of these two classes. The model presented in Example 14 didn't include the preprocessing stage; it used the data mining proper operators only.

The main idea behind the concept of the operator is to capture not only the fact that data mining techniques generalize the data, but also to categorize existing methods. We want to do it in as exclusive/inclusive sense as possible. It means that we want to make sure that our categorization will be to distinguish as it should, for example clustering from classification, making at the same time all classification methods fall into one category, called the classification operator while all clustering methods would fall into the category called the clustering operator. The third category is the association analysis described in our framework by the association operator. We don't include in our analysis purely statistical methods like regression, etc … This gives us only three data mining classes of operators to consider: classification $\mathcal{G}_{\text{class}}$, clustering $\mathcal{G}_{\text{clust}}$, and association $\mathcal{G}_{\text{assoc}}$. The motivation, discussion and formal definition of these classes is included in the paper [22]

Theorem 17 *Let $\mathcal{G}_{\text{class}}, \mathcal{G}_{\text{clust}}$ and $\mathcal{G}_{\text{assoc}}$ be the sets of all classification, clustering, and association operators, respectively. The following conditions hold.*

(1) $\mathcal{G}_{\text{class}} \neq \mathcal{G}_{\text{clust}} \neq \mathcal{G}_{\text{assoc}}$
(2) $\mathcal{G}_{\text{assoc}} \cap \mathcal{G}_{\text{class}} = \emptyset$,
(3) $\mathcal{G}_{\text{assoc}} \cap \mathcal{G}_{\text{clust}} = \emptyset$.

Data Mining Process

We model here two stages of the data mining process: preprocessing and data mining proper, as we were able to distinguish the data preprocessing and data mining proper operators as disjoint categories. We adopt the following definitions.

Definition 18 *Any sequence $K_1, K_2, \ldots, K_n$ $(n \geq 1)$ of data mining states is called a* **data preprocessing process**, *if there is a preprocessing operator $G \in \mathcal{G}_{\text{prep}}$, such that $G(K_i) = K_{i+1}$, $i = 1, 2, \ldots, n-1$.*

Definition 19 *Any sequence $K_1, K_2, \ldots, K_n$ $(n \geq 1)$ of data mining states is called a* **data mining proper process**, *if there is a data mining proper operator $G \in \mathcal{G}_{\text{dm}}$, such that $G(K_i) = K_{i+1}$, $i = 1, 2, \ldots, n-1$.*

The data mining process consists of the preprocessing process (that might be empty) and the data mining proper process. We know that the sets $\mathcal{G}_{\text{prep}}$ and $\mathcal{G}_{\text{dm}}$ are disjoint. This justifies the following definition.

Definition 20 *A* **data mining process** *is any sequence $K_1, K_2, \ldots, K_n$ $(n \geq 1)$ of data mining states, such that $K_1, \ldots, K_i (0 \leq i \leq n)$ is a preprocessing process and $K_{i+1}, \ldots, K_n$ is a data mining proper process.*

Observe that in the semantic model $S\mathbf{M_1}$ (Example 14) we have only two processes: K_0, K_1 and K_0, K_2. Both of them are data mining proper processes.

Descriptive Model

Given a semantic model $S\mathbf{M} = (\mathcal{P}(U), \mathcal{K}, \mathcal{G})$ We associate with it its descriptive counterpart defined below.

Definition 21 *A* **descriptive model** *is a system $\mathcal{D}\mathbf{M} = (\mathcal{L}, \mathcal{E}, \mathcal{D}K)$ where:*

- *$\mathcal{L} = (\mathcal{A}, \mathcal{E})$ is called a* **descriptive language**;
- *$\mathcal{A}$ is a countably infinite set called the* **alphabet**;
- *$\mathcal{E} \neq \emptyset$ and $\mathcal{E} \subseteq \mathcal{A}^*$ is the set of* **descriptive expressions** *of $\mathcal{L}$;*
- *$\mathcal{D}K \neq \emptyset$ and $\mathcal{D}K \subseteq \mathcal{P}(\mathcal{E})$ is a set of* **descriptions of knowledge states**.

As in the case of a semantic model, we build the descriptive model for a given application. We define here only a general form of the model. We assume however, that whatever the application is, the descriptions are always built in terms of attributes and values of the attributes, some logical connectives, some predicates and some extra parameters, if needed. For example, a neural network with its nodes and weights can be seen as a formal description (in an appropriate descriptive language), and the knowledge states would represent changes in parameters during the neural network training process, or a final, converged network.

The model we build here is a model for what we call a **descriptive data mining**, i. e. the data mining for which the goal of the data mining process is to produce a set

of descriptions in a language easily comprehensible to the user. For that reason we identify, for example, a decision tree constructed by the classification by the decision tree algorithm with the set of discriminant rules obtained from the tree.

In our Examples 2–14 we have used descriptions in the form $(a = v)$ to denote that the attribute a has a value v, but one might also use, like it is often done a predicate form $a(v)$ or $a(x, v)$ instead.

We define the components of $\mathcal{D}$**M** in the following stages.

Stage 1 For each $K \in \mathcal{K}$, we define (Sect. "K-Descriptive Language") its own descriptive language $\mathcal{L}_K = (\mathcal{A}_K, \mathcal{E}_K)$.

Stage 2 For each $K \in \mathcal{K}$, and descriptive expression $F \in \mathcal{E}_K$, we define what does it mean that D satisfied in K; i. e. we define (Definition 25) a satisfaction relation $\models_K$.

Stage 3 For each $K \in \mathcal{K}$, and descriptive expression $F \in \mathcal{E}_K$, we define what does it mean that D is true K, i. e. $\models_K D$ (Definition 28).

Stage 4 For each $K \in \mathcal{K}$, we define its set $\mathcal{D}K \subseteq \mathcal{P}(\mathcal{E}_K)$ of descriptions of its own knowledge (Definition 26).

Stage 5 We use the languages $\mathcal{L}_K$ to define the descriptive language $\mathcal{L}$ (Definition 29).

Stage 6 We use the descriptive expressions $\mathcal{E}_K$ of $\mathcal{L}_K$ to define the set $\mathcal{E}$ of descriptive expressions of $\mathcal{L}$ (Definition 30).

Later Stage We use the satisfaction relations $\models_K$ to define the satisfaction relation $\models$ of our granular model **GM** (Definitions 1 and 31, respectively).

K-Descriptive Language

As Stage 1 of our construction of the descriptive model, we define for each $K \in \mathcal{K}$, its own description language $\mathcal{L}_K$. The language depends on the semantic model and the goal of the data mining process. As we have said, the descriptions produced by data mining algorithms come in different forms. We build here a model for a **descriptive data mining**, i. e. we assume that the descriptions are built from attributes and values of attributes and two logical connectives: conjunction (see descriptions associated with Examples 2–14) and implication. The implication connective is needed to model the different kind of rules that are being mined by data mining algorithms: discriminant and characteristic rules in classification analysis; association rules by association analysis; or other rules obtained by hybrid systems.

Definition 22 *For any $K \in \mathcal{K}$, of S**M**, $K = (\mathcal{P}(U), A, E, V_A, V_E, g)$, we define the* **descriptive language** *of K, $\mathcal{L}_K$ as*

$$\mathcal{L}_K = (\mathcal{A}_K, \mathcal{E}_K)\,,$$

where $\mathcal{A}_K$ is called an **alphabet**, *$\mathcal{E}_K$ the set of* **descriptive expressions** *such that*

$$\mathcal{E}_K = \mathcal{D}_K \cup \mathcal{F}_K\,,$$

for $\mathcal{D}_K$ the set of **descriptive formulae** *and $\mathcal{F}_K$ the set of* **formulae** *of $\mathcal{L}_K$ (Definition 23).*

The alphabet $\mathcal{A}_K = VAR_K \cup \{\cap, \Rightarrow\} \cup \{(,)\}$, where $\{\cap, \Rightarrow\}$ is the set of logical connectives of $\mathcal{L}_K$. The set VAR_K of variables of K is also called its set of **atomic descriptions**. We put $VAR_K = \mathcal{D}_A \cup \mathcal{D}_E$, where $\mathcal{D}_A = \{(a = v) : a \in A, v \in V_A\}, \mathcal{D}_E = \{(a = v) : a \in E, v \in V_E\}$.

Atomic descriptions, i. e. the elements of $VAR_K = \mathcal{D}_A \cup \mathcal{D}_E$ represent minimal blocks of semantical description. Elements of $\mathcal{D}_A$ are atomic descriptions of minimal blocks built with use of the attributes of the initial target database. Elements of $\mathcal{D}_E$ are atomic descriptions of minimal blocks built with use of knowledge attributes used (if any) during the process of data mining. We use them to define the sets of descriptions and formulae as follows.

Definition 23 *The set $\mathcal{D}_K$ of all* **descriptive formulae** *of $\mathcal{L}_K$ is the set*

$$\mathcal{D}_K = A\mathcal{D}_K \cup E\mathcal{D}_K \cup \mathcal{F}_K\,,$$

where $A\mathcal{D}_K, E\mathcal{D}_K, \mathcal{F}_K$ are defined as below.

The set $A\mathcal{D}_K \subseteq \mathcal{A}_K{}^*$ is called the set of **data attribute descriptions** and is the smallest set such that the following conditions hold.

(1) $\mathcal{D}_A \subseteq A\mathcal{D}_K$ (data attribute **atomic** description);
(2) If $D_1, D_2 \in A\mathcal{D}_K$, then $D_1 \cap D_2 \in A\mathcal{D}_K$.

The set $E\mathcal{D}_K \subseteq \mathcal{A}_K{}^*$ is called the set of **knowledge attribute descriptions** and is the smallest set such that the following conditions hold.

(1) $\mathcal{D}_E \subseteq E\mathcal{D}_K$ (knowledge attribute **atomic** description);
(2) If $D_1, D_2 \in E\mathcal{D}_K$, then $D_1 \cap D_2 \in E\mathcal{D}_K$.

We distinguish two categories of formulae: one $\mathcal{F}_A$ describes a certain knowledge, the other, $\mathcal{F}_E$, the uncertain, or approximate, knowledge. The certain knowledge is expressed in our model in terms of attributes and values of attributes of the initial, target data only. The description

of the approximate knowledge includes the extra parameters, describing properties of granules as defined by the knowledge attributes of K (Definition 9).

Definition 24 *The set $\mathcal{F}_K$ of* **formulae** *of $\mathcal{L}_K$ is a union of two sets of formulae; $\mathcal{F}_A, \mathcal{F}_E$, i. e.*

$$\mathcal{F}_K = \mathcal{F}_A \cup \mathcal{F}_E \,,$$

where $\mathcal{F}_A = A\mathcal{D}_K \cup A\mathcal{R}_K$ and $\mathcal{F}_E = AE\mathcal{D}_K \cup E\mathcal{R}_K$ are defined below.

The set $AE\mathcal{D}_K = \{D_1 \cap D_2 : D_1 \in A\mathcal{D}_K, D_2 \in E\mathcal{D}_K\}$ is the set of **knowledge description formulae**, with knowledge attribute descriptions depicting uncertainty measures. The set $A\mathcal{R}_K = \{(D_1 \Rightarrow D_2) : D_1, D_2 \in \mathcal{D}_A\}$ is the set of **attribute rule-formulae** depicting certain rules obtained during the data mining process. The set $E\mathcal{R}_K = \{(D_1 \Rightarrow D_2) \cap D_3 : D_1, D_2 \in \mathcal{D}_A, D_3 \in \mathcal{D}_E\}$ is the set of **knowledge rule-formulae** depicting rules with uncertainty measures described obtained by descriptions from $\mathcal{D}_E$.

The formulae of the language $\mathcal{L}_K$ **are not** yet the data mining **rules**. They only describe a syntactical form of the rules appropriate for a given application. Formulae from $\mathcal{F}_K$ become **rules determined by** K only when they do relate semantically to K, i. e. reflect the properties of our initial target database. In this case we say that they are **true**, or **true under the measures** described by descriptions from $\mathcal{D}_E$. We define, in the next section the notion of truthfulness, as we always do, via notion of satisfiability.

K- Satisfaction and K- Truth

In this section we define (Definition 25) a satisfaction relation $\models_K \subseteq \mathcal{P}(U) \times \mathcal{E}_K$ that establishes the relationship between descriptive expressions of the $\mathcal{L}_K$ and what they semantically represent in K. For any $(S, F) \in \models_K$ we say that S **satisfies** F **in** K and write it symbolically as

$$S \models_K F \,.$$

We call our satisfaction relation a k-satisfaction, as it represents a satisfaction relation relative to the system K. As the next step we define the notion of K-truth (Definition 28) i. e. we define what does it mean that a descriptive expression F **is true in** K, symbolically expressed as

$$\models_K F \,.$$

Let $K = (K(U), A, E, V_A, V_E, g)$ and let $\mathcal{L}_K = (\mathcal{A}_K, \mathcal{E}_K)$ be the description language defined by K (Definition 22), where $\mathcal{E}_K = \mathcal{D}_K \cup \mathcal{F}_K$.

Definition 25 *A* **k-satisfaction relation** $\models_K \subseteq \mathcal{P}(U) \times \mathcal{E}_K$ *is defined by induction over the level of complexity of any descriptive expression $F \in \mathcal{E}_K$ as follows.*

1. $F \in \mathcal{D}_K$.
   (i) Let $(a = v) \in \mathcal{D}_A$ be an **atomic** attribute description. We define, for any $S \in \mathcal{P}(U)$, for any $a \in A, v \in V_A$, $S \models_K (a = v)$ if and only if $(S, a) \in \mathrm{dom}(g)$ and $g(S, a) = v$;
   (ii) Let $(e = v) \in \mathcal{D}_E$ is an **atomic** knowledge attribute description. We define, for any $S \in \mathcal{P}(U)$, for any $e \in E, v \in V_E$, $S \models_K (a = v)$ if and only if $(S, e) \in \mathrm{dom}(g)$ and $g(S, e) = v$;
2. $F \in \mathcal{F}_K$.
   We extend $\models_K$ to the set $\mathcal{F}_A = A\mathcal{D}_K \cup A\mathcal{R}_K$
   (iii) Let $D \in A\mathcal{D}_K$ and $D = D_1 \cap \cdots \cap D_n$. We define, for any $S \in \mathcal{P}(U)$, $S \models_K D$ if and only if $\forall (1 \leq i \leq n)(S \models_K D_i)$;
   (iv) Let $(D_1 \Rightarrow D_2) \in A\mathcal{R}_K$, i. e. $D_1, D_2 \in A\mathcal{D}_K$. We define, for any $S \in \mathcal{P}(U)$, $S \models_K (D_1 \Rightarrow D_2)$ if and only if $S \models_K D_1$ and $S \models_K D_2$;
   (v) Let $(D_1 \Rightarrow D_2) \cap D_3 \in E\mathcal{R}_K$, i. e. $(D_1 \Rightarrow D_2) \in A\mathcal{R}_K$ and $D_3 \in \mathcal{D}_E$. We define, for any $S \in \mathcal{P}(U)$, $S \models_K (D_1 \Rightarrow D_2) \cap D_3$ if and only if $S \models_K (D_1 \Rightarrow D_2)$ and $S \models_K D_3$.

This ends Stage 2 of the definition of the descriptive model and we are ready for Stage 3, i. e. to define the set of all descriptions of knowledge states.

Definition 26 *Let $K \in \mathcal{K}$, the set $\mathcal{D}K \subseteq \mathcal{P}(\mathcal{E}_K)$ of* **descriptions of knowledge states** *of K is defined as follows*

$$\mathcal{D}K = \{F \in \mathcal{E}_K : \exists S \in Gr_K(S \models_K F)\} \,.$$

Now we are ready to define a notion of F *true in* K, for any *formula* $F \in \mathcal{F}_K$, symbolically expressed by $\models_K F$. This notion relates the satisfaction relation $\models_K$, i. e. satisfiability of formulae in the system K with the initial target data as out ultimate point of reference. This connection is being established by the notion of the *truth set* defined below.

Definition 27 *For any attribute data description $D \in A\mathcal{D}_K \cup E\mathcal{D}_K$ the set $S(D) = \{x \in U : D\}$ is called a* **truth set** *for D.*

We define $\models_K F$ by induction over the level of complexity of any descriptive expression $F \in \mathcal{F}_K$ as follows.

Definition 28

1. $F \in A\mathcal{D}_K$.
   (i) *For any attribute data description $D \in A\mathcal{D}_K$, we define: $\models_K D$ if and only if $S(D) \models_K D$.*

(ii) *For any attribute description formula* $D \in A\mathcal{D}_K$, *we put* $\models_K D$ *if and only if* $S(D) \models_K D$.

2. $F \in \mathcal{F}_A$.

(iii) *For an attribute rule-formula* $(D_1 \Rightarrow D_2) \in A\mathcal{R}_K$, *we define* $\models_K (D_1 \Rightarrow D_2)$ *if and only if* $S(D_1) \models_K D_1, S(D_1) \models_K D_2$, *and* $S(D_1) \subseteq S(D_2)$ *or* $(D_1) \cap S(D_2) \neq \emptyset$.

3. $F \in AE\mathcal{D}_K$.

(iv) *For any knowledge attribute description* $D_1 \cap D_2 \in AE\mathcal{D}_K, D_1 \in A\mathcal{D}_K, D_2 \in E\mathcal{D}_K$ *we define*: $\models_K D_1 \cap D_2$ *if and only if* $S(D_1) \cap S(D_2) \models_K D_1$.

4. $F \in \mathcal{F}_E$.

(v) *For an attribute rule-formula* $(D_1 \Rightarrow D_2) \cap D_3 \in AE\mathcal{R}_K$, *we define* $\models_K (D_1 \Rightarrow D_2) \cap D_3$ *if and only if* $S(D_1) \cap S(D_3) \models_K D_1, S(D_1) \cap S(D_3) \models_K D_2$, *and* $S(D_1) \cap S(D_3) \subseteq S(D_2) \cap S(D_3)$ *or* $(D_1) \cap S(D_2) \cap S(D_3) \neq \emptyset$.

In the case when $F \in \mathcal{F}_E$ and $\models_K F$ we say that F **is true in** K **under the measures** D, for a certain $D \in AE\mathcal{D}_K$.

If $\models_K (D_1 \Rightarrow D_2)$ and the condition $S(D_1) \subseteq S(D_2)$ holds, then we call the formula $(D_1 \Rightarrow D_2)$ **C- true in** K, or **C- true under the measures** D, if $\models_K (D_1 \Rightarrow D_2) \cap D$.

If $\models_K (D_1 \Rightarrow D_2)$ and the condition $(D_1) \cap S(D_2) \neq \emptyset$ holds, then we call the formula $(D_1 \Rightarrow D_2)$ **D- true in** K, or **D - true under the measures** D, if $\models_K (D_1 \Rightarrow D_2) \cap D$.

The notion of D-truth reflects the semantics needed to define the discriminant rules for the classification analysis, and C-truth is needed for the characteristic rules, hence the names. As we have said before, the knowledge attributes from the set E describe uncertainty measures for the granules $S \in K(U)$. The formulae that incorporate the knowledge attribute descriptions $D \in E\mathcal{D}_K$ can be only *true* in K under some uncertainty measures (Example 5).

Descriptive Language and Descriptive Model

We have already constructed (Sects. "Descriptive Language and Descriptive Model") all sub-components of the Definition 21 of the descriptive model $\mathcal{D}\mathbf{M}$ and now we proceed to complete it as follows.

Definition 29 We define the language of $\mathcal{D}\mathbf{M}$ as $\mathcal{L} = (\mathcal{A}, \mathcal{E})$, where $\mathcal{A} = \bigcup\{\mathcal{A}_K : K \in \mathcal{K}\}$, and $\mathcal{E} = \bigcup\{\mathcal{E}_K : K \in \mathcal{K}\}$.

The set $\mathcal{D}K$ of all descriptions of knowledge states of the semantic model $S\mathbf{M} = (\mathcal{P}(U), \mathcal{K}, \mathcal{G})$ is the following.

Definition 30 Let $\mathcal{D}K$ be the set of descriptions of knowledge states of K (Definition 26). The set $\mathcal{D}K = \bigcup\{\mathcal{D}_K : K \in \mathcal{K}\}$ is a set of **descriptions of knowledge states** of $S\mathbf{M}$.

This completes the definition of the descriptive model as a system $\mathcal{D}\mathbf{M} = (\mathcal{L}, \mathcal{E}, \mathcal{D}K)$ where: $\mathcal{L} = (\mathcal{A}, \mathcal{E})$ is a descriptive language; $\mathcal{A}$ an alphabet and $\mathcal{E}$ is the set of descriptive expressions (Definition 29); $\mathcal{D}K$ is a set of descriptions of knowledge states (Definition 30).

Granular Model Revisited: Satisfaction and Truth

The granular model is defined (Sect. "Granular Model: Syntax and Semantics for Data Mining") as a system $\mathbf{GM} = (S\mathbf{M}, \mathcal{D}\mathbf{M}, \models)$ where: $S\mathbf{M}$ is a semantic model as defined in the Definition 13; $\mathcal{D}\mathbf{M}$ is a descriptive model (Definition 21); and $\models \subseteq \mathcal{P}(U) \times \mathcal{E}$ is a satisfaction relation. All the components of **GM** except the satisfaction relation have already been defined. As the last step we define the satisfaction relation and the notion of truth in **GM** as follows.

Definition 31 *For any* $S \in \mathcal{P}(U)$ *and for any* $F \in \mathcal{E}$, $S \models F$ *if and only if* $\exists K \in \mathcal{K}(S \models_K F)$

Definition 32 *We say that* $F \in \mathcal{E}$ *is* **true** *in* **GM** (*symbolically* $\models F$) *if and only if* $\exists K \in \mathcal{K}(\models_K F)$.

Future Directions

The models presented here set a general framework for future foundational investigations. They can be carried on three levels. One, the most general, would deal with further developments within the granular model. The second, a more specific one, would deal with applications of the methods and language developed within the granular model to specific domains. On this level one would build semantic, descriptive, and granular models for different data mining domains; i. e., build and examine specific models for classification, clustering, and for association analysis. Finally, the most specific third level would deal with building models for basic descriptive data mining algorithms: decision trees and rough sets classification, association and classification by association, and clustering. The general methodology of analysis of the data mining process developed for the granular model can hence serve as unifying language in which different data mining domains and algorithms can be studied, discussed, and compared.

Bibliography

1. Chapman P (2000) CRISP-DM 1.0 – Step-by-step Data Mining CRISP-DM Consortium: 1.0
2. Greco S, Matarazzo B, Slowinski R, Stefanowski J (2002) Importance and interaction of conditions in decision rules. In: Proceedings of Third International Conference RSCTC'02, Malvern,

USA. Lecture Notes in Artificial Intelligence, vol 2475. Springer, Berlin, pp 255–262
3. Hadjimichael M, Wasilewska A (1998) A hierarchical model for information generalization. In: Proceedings of the 4th Joint Conference on Information Sciences, Rough Sets, Data Mining and Granual Computing (RSDMGrC'98) North Carolina, vol II, pp 306–309
4. Han J, Kamber M (2000) Data mining: Concepts and techniques. Morgan Kauffman
5. Inuiguchi M, Tanino T (2003) Classification versus approximation oriented generalization of rough sets. Bull Int Rough Set Soc 7(1/2):19–25
6. Komorowski J (2002) Modelling biological phenomena with rough sets. In: Proceedings of Third International Conference RSCTC'02, Malvern, USA. Lecture Notes in Artificial Intelligence, vol 2575. Springer, Berlin, p 13
7. Lin TY (2002) Database mining on derived attributes. In: Proceedings of Third International Conference RSCTC'02, Malvern, USA. Lecture Notes in Artificial Intelligence, vol 2575. Springer, Berlin, pp 14–32
8. Martinez FJ, Menasalvas E, Wasilewska A, Fernández C, Hadjimichael M (2002) Extension of relational management system with data mining capabilities. In: Proceedings of Third International Conference RSCTC'02, Malvern, USA. Lecture Notes in Artificial Intelligence, vol 2475. Springer, Berlin, pp 421–424
9. Menasalvas E, Wasilewska E, Fernández C (2001) The lattice structure of the KDD process: Mathematical expression of the model and its operators. Fundamenta Informaticae: Special Issue Int J Info Syst 48–62
10. Menasalvas E, Wasilewska A, Fernández C, Martinez FJ (2002) Data mining – a semantical model. In: Proceedings of 2002 World Congres on Computational Intelligence, Honolulu, pp 435–441
11. Pawlak Z (1981) Information systems. Theor Found Inf Syst 6:205–218
12. Pawlak Z (1991) Rough sets – theoretical aspects reasoning about data. Kluwer, Norwell
13. Fayyad, Piatetsky-Shapiro, Smyth (1996) From data mining to knowledge discovery: An overview. In: Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy (eds) Advances in Knowledge Discovery and Data Mining. AAAI Press, Menlo Park, pp 1–34
14. Polkowski Lech (2002) Rough sets – mathematical foundations. Physica, Heidelberg
15. Shearer C (2000) The CRISP-DM model: The new blueprint for data mining. J Data Warehous 5(4):13–22
16. Skowron A (1993) Data filtration: a rough set approach. In: Proceedings of rough sets, fuzzy sets and knowledge discovery. RSKD, pp 108–118
17. Wasilewska A, Ruiz EM, Fernández-Baizan MC (1998) A model for RSMD implementation. 1st International Conference on Rough Sets and Current Trends in Computing (RSCTC'98), 22–26 June, Warsaw, pp 186–193
18. Wasilewska A, Menasalvas E (2004) Data preprocessing and data mining as generalization process. In: Proceedings of ICDM'04, The Fourth IEEE International Conference on Data Mining, Brighton, 1–4 Nov, pp 133–137
19. Wasilewska A, Menasalvas E (2004) Data mining operators. In: Proceedings of ICDM'04, The Fourth IEEE International Conference on Data Mining, Brighton, pp 209–214
20. Wasilewska A, Menasalvas E, Scharff C (2005) Uniform Model for Data Mining. In: Proceedings of FDM05 (Foundations of Data Mining). In: ICDM2005, Fifth IEEE International Conference on Data Mining. Austin, Texas, pp 19–27
21. Wasilewska A, Ruiz EM (2006) Data mining as generalization: A formal model. In: Lin TY, Ohsuga S, Liau CJ, Hu X (eds) Foundations and Novel Approaches in Data Mining. Studies in Computational Intelligence, vol 9. Springer, Berlin, pp 99–126
22. Wasilewska A, Ruiz EM (2008) Data preprocessing and data mining as generalization. In: Lin TY, Xie Y, Wasilewska A, Liau C-J (eds) Data Mining: Foundations and Practice. Studies in Computational Intelligence. Springer, pp 453–468
23. Wasilewska A, Ruiz EM (2005) A classification model: Syntax and semantics for classification. In: Proceedings of 10th International RSFDGrC2005 Conference, Regina, Canada, August/September 2005. Lecture Notes in Artificial Intelligence, vol 2. Springer, Berlin, pp 59–68
24. Ziarko W, Fei X (2002) VPRSM approach to WEB searching. In: Proceedings of Third International RSCTC'02 Conference, Malvern, USA, October 2002. Lecture Notes in Artificial Intelligence, vol 2475. Springer, Berlin, pp 514–522
25. Ziarko W (1993) Variable precision rough set model. J Comput Syst Sci 46(1):39–59
26. Yao JT, Yao YY (2002) Induction of classification rules by granular computing. In: Proceedings of Third International RSCTC'02 Conference, Malvern, USA, October 2002. Lecture Notes in Artificial Intelligence, vol 2475. Springer, Berlin, pp 331–338

Granular Neural Network

Yan-Qing Zhang
Department of Computer Science,
Georgia State University, Atlanta, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Basic Architecture
Granular Learning Algorithms
Applications in Bioinformatics
Applications in Computational Web Intelligence
Applications in Brain Informatics
Conclusions
Future Directions
Bibliography

Glossary

Granular data Granular data include various data granules such as classes, clusters, subsets, groups, linguistic values and intervals.

Granular neuron An artificial neuron maps granular data inputs to granular data outputs.

Granular link A granular link connects two different granular neurons.

Granular weights A granular weight represents connection strength between two granular neurons by using a granular value that is not limited to a traditional numerical value.

Granular neural network An intelligent neural network consists of granular neurons and granular links that connect relevant granular neurons.

Definition of the Subject

Granular computing is a general computation theory for effectively using granules such as classes, clusters, subsets, groups and intervals to build an efficient computational model for complex applications with huge amounts of data, information and knowledge. Relevant existing techniques are divide and conquer, cluster analysis, fuzzy sets, rough sets, Yin–Yang computing, neutrosophic computing, interval computing, quotient space theory, belief functions, machine learning, databases, and so on.

A biological neural network in the human brain consists of a huge number of biological neurons that are networked to process various data sets (numerical and nonnumerical). Different complex data sets such as images, linguistic terms, numbers, patterns and sounds can be recognized by biological neural networks. In other words, biological neural networks can process granules such as data clusters, linguistic words, information classes, image groups from human sensor networks (i. e, eyes, ears, nose, etc.).

Based on granular computing and biological neural networks, a granular neural network is designed to deal with numerical-linguistic data fusion and granular knowledge discovery in numerical-linguistic databases. From a data granulation point of view, the granular neural network can process granular data in a database. From a data fusion point of view, the granular neural network makes decisions based on different kinds of granular data. From a knowledge discovery point of view, the granular neural network is able to learn internal granular relations between numerical-linguistic inputs and outputs, and predict new relations in a database. The granular neural network is also capable of greatly compressing low-level granular data to high-level granular knowledge with some compression error and a data compression rate.

Introduction

Similar to a biological neural network in the human brain, a traditional artificial neural network consists of lots of artificial neurons that are connected by links. Each artificial neuron is simply a nonlinear mathematical mapping function, and each link has a numerical weight value showing connection strength between two neurons. The learning algorithm uses training data sets to optimize these weights of an artificial neural network, and the trained neural network can be used in prediction and decision-making applications. Traditional artificial neural networks based on the basic principle of biological neural networks in the human brain have been widely used in a lot of applications. However, traditional artificial neural networks use numerical values like 3.14 and −23.5. Internal weights are also presented by numerical values that are not meaningful for users (i. e., the black box problem). Generally, the traditional artificial neural network has two long-term problems: (1) the black box problem: Learned knowledge in a neural network is represented by non-meaningful numerical weights (i. e., people cannot understand these numerical values because they are not linguistic values, therefore people cannot understand how the neural network makes a decision), and (2) the curse of dimensionality: When a number of inputs and outputs is increased, the number of weights is increased dramatically (even exponentially). So the traditional artificial neural network cannot be used for large-scale applications with a large number of inputs and outputs.

Granular neural networks are used to process linguistic data like fuzzy terms to do granular knowledge discovery [20]. Granular neural networks with different structures can process information granules such as fuzzy sets and rough sets by using training algorithms [10]. The granular neural network uses linguistic weights and the rules using linguistic arithmetic to speed up learning process [1]. Granular neural networks are used to real applications such as land use classification [14]. In summary, the granular neural learning algorithm can discover meaningful granular rules that can be understood by ordinary people. So the granular neural network is useful to tackle the black box problem.

A set of training data is given to a computational model, and then the problem is how the network gleans useful knowledge from these training data sets in order to use that discovered knowledge to make right decisions for real applications. The two challenging problems for both computer science and cognitive science are the black box problem (i. e., the uninterpretability of a network of numerical weights or connection strengths) and the curse of dimensionality (i. e., the intractability complex cognitive problems for linear neural networks). This proposal plans to design a new architecture of a granular neural network with a new learning algorithm, and then use data from both cognitive science and computer science to test and to improve the new learning algorithm in terms of the two challenging problems.

In recent years, hybrid neural networks such as fuzzy neural network, granular neural networks and genetic neural networks have been investigated to solve complex application problems with high dimensionality. However, a single-stage hybrid neural network suffers from curse of dimensionality. A single-stage hybrid neural network suffers from curse of dimensionality because the number of parameters of the single-stage hybrid neural network is increasing exponentially with the increasing of the number of inputs. The hierarchical network structures based on multistage fuzzy reasoning have been used to solve the dimensionality problem. The ANFIS (Adaptive-Network-Based Inference System), based on TSK fuzzy reasoning, is useful in many applications. Because it still cannot extract the commonly used fuzzy rules with both fuzzy IF part and fuzzy THEN part, it makes some difficulties in acquiring fuzzy knowledge from experts in a natural manner and learning commonly used fuzzy rules from data.

Therefore, how to design a powerful granular neural network with high learning speed, low training error and low prediction error is still a challenging problem of artificial intelligence, computational intelligence and cognitive science.

Basic Architecture

There are different types of granular neural networks. Now two granular neural networks are introduced in the two following sections, respectively.

First Architecture of Granular Neural Network

A database may contain numerical values, linguistic words, images, sounds, music pieces, and texts. The lowest data granulation technique deals with raw multimedia data collected directly from a real world environment, whereas a higher data granulation technique classifies raw multimedia data into higher-level granules (i. e., classes, clusters, categories, groups, sets, etc.) to simplify data processing and data mining. For high-level data granulation, basic processing elements are these granules (not low-level raw data) [20]. In this sense, data granulation is related to data mining, data fusion, information fusion and knowledge discovery. Interestingly, the human brain has a very strong ability to process multimedia granules, compute with linguistic words, and discover useful information from multimedia data. There should be links between raw multimedia data in an outside real world and biological neural networks inside a human brain. Similarly, links should be established between multimedia granules in databases and inputs-outputs of artificial neural networks. Since an artificial neural network cannot directly process these multimedia granules in most cases (because it is designed for directly processing numerical data), a conversion or interpretation of a granule, a link between a multimedia database and a neural network, becomes crucial to designing a granular neural network. In other words, how to convert multimedia granules into corresponding numerical features is very important for granular neural networks. For example, the linguistic data feature extraction system can convert fuzzy linguistic data such as very high, too small and almost 100 into typical numerical features [20].

Generally speaking, a granular neural network is capable of processing various granular data (granules). Granules could be a class of numbers, a cluster of images, a set of concepts, a group of objects, a category of data, etc. These granules are inputs and outputs as multimedia data are inputs and outputs of biological neural networks in the human brain. Therefore, a granular-data-based granular neural network is more useful and more effective to process multimedia granules than a conventional numerical-data-based neural network. Here, the granular neural network is made by Fuzzy Neural Networks with Knowledge Discovery [18].

Generally speaking, a granular neural network is capable of processing various granular data (granules) [20]. Granules could be a class of numbers, a cluster of images, a set of concepts, a group of objects, a category of data, etc. These granules are inputs and outputs of granular neural networks as multimedia data are inputs and outputs of biological neural networks in the human brain.

Therefore, granular-data-based granular neural networks are more useful and more effective to process multimedia granules than conventional numerical-data-based neural networks. Here, the FNN is a powerful Fuzzy Neural Network with Knowledge Discovery (FNNKD). The functions of different layers are described layer by layer as follows [20,21]:

Layer 1: Linguistic Feature Extraction Layer In this layer, numerical-linguistic X and Y are transformed to corresponding fuzzy feature vectors $(a1, b1, c1, d1)$ and $(a2, b2, c2, d2)$, respectively.

Layer 2: Multi-FNNKD Layer This layer consists of 4 dedicated FNNKDs (i. e., FNNKD1, FNNKD2, FNNKD3, and FNNKD4). FNNKD1 is a $2 \times k_1 \times 1$ fuzzy neural network which generates a crisp center a of an output fuzzy set by using k_1 fuzzy rules. FNNKD2 is a $2 \times k_2 \times 1$ fuzzy neural network which generates a crisp width b of an output fuzzy set by using k_2 fuzzy rules. FNNKD3 is a $2 \times k_3 \times 1$ fuzzy neural network which generates c of

an output fuzzy set by using k_3 fuzzy rules. FNNKD4 is a $2 \times k_4 \times 1$ fuzzy neural network which generates d of an output fuzzy set by using k_4 fuzzy rules.

An n-input-1-output normal fuzzy system has m fuzzy IF-THEN rules which are described by:

IF x_1 is A_1^k and … and x_n is A_n^k THEN y is B^k, where x_i and y are input and output fuzzy linguistic variables, respectively.

Layer 3: Output Layer

Case 1: A fuzzy linguistic output is Z represented by a fuzzy feature vector (a, b, c, d).

Case 2: A crisp numerical output is a.

For simplicity, the detailed learning algorithm is given in [19]. Once the learning procedure has been completed, all parameters for a FNNKD have been adjusted and optimized. As a result, all m fuzzy rules have been discovered from training data. Finally, the trained FNNKD can generate new values for new given input data.

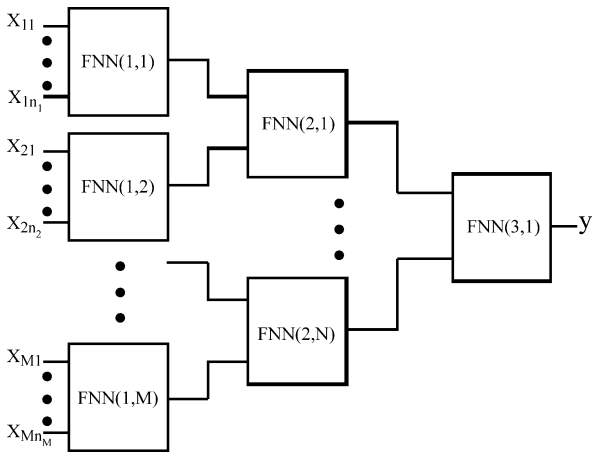

Granular Neural Network, Figure 1
Architecture of a 3-stage granular neural network [17]

Second Architecture of Granular Neural Network

Here, we plan to use traditional artificial intelligence, new computational intelligence including fuzzy logic, neural networks, evolutionary computation, granular computing and cognitive science to investigate the new evolutionary granular cognitive neural network that can discover hidden decision-making knowledge from data (i. e., know how the human brain learns new knowledge and makes a decision). The traditional neural-learning algorithms can generate a lot of numerical weights that are not meaningful (i. e., the black box problem). Now we plan to design basic neuron granules that can contain meaningful knowledge like rules, and make new granular learning methods based on cognitive science and computational intelligence [5].

A new evolutionary granular cognitive neural network is proposed based on the normal fuzzy reasoning. The hybrid evolutionary granular cognitive neural learning algorithm using the divide-and-conquer strategy is developed to enhance learning quality in term of discovered knowledge, training error and prediction error. Simulations have shown that the evolutionary granular cognitive neural network is an effective data mining and knowledge discovery system which discovers meaningful fuzzy rules, has low training error, and generates low prediction error. There are two main problems of curse of dimensionality for a soft computing system: (1) structural dimensionality (i. e., the total number of soft rules increases exponentially with the number of input variables in single-stage reasoning process, and (2) parametric dimensionality (i. e., the total number of adjustable system parameters increases exponentially with the number of input variables). For example, if a fuzzy system has n linguistic variables, each variable has m fuzzy linguistic sets, and each fuzzy linguistic set has k parameters, then the fuzzy system will have m^n fuzzy rules and $(km)^n$ parameters. In this case, the structural complexity and the parametric complexity of the fuzzy system are $O(m^n)$ and $O[(km)^n]$, respectively. To solve the problems of curse of dimensionality, different hybrid soft computing architectures such as the multistage fuzzy neural network and the hierarchical system have been proposed to reduce system complexity in terms of dimensionality, time and space. In general, the key technology for solving the curse of dimensionality of a hybrid soft computing system (particularly Fuzzy Neural Network (FNN)) is the divide-and-conquer method. In other words, an original single-stage FNN with high system complexity can be divided into small multi-stage soft computing building blocks with low system complexity so as to reduce both structural dimensionality and parametric dimensionality significantly.

In general, the kth-stage FNN uses outputs generated from relevant $k-1$th-stage FNNs as inputs, and then generates an output which will become an input of relevant $k+1$th-stage FNNs. A final stage FNN will generate a final output. For example, Fig. 1 shows the 3-stage GCFNN. The FNN(i, j) is the ith-stage-$j$thFNN. So FNN$(1, j)$ uses x_{1nj}as inputs for $j = 1, 2, \ldots, M$, similarly FNN$(2, l)$ for $l = 1, 2, \ldots, N$ uses outputs generated from relevant 1st-stage FNNs as inputs, and finally FNN(3,1) generates the final output y. Therefore, the final output y is a function of all original inputs $xj1, xj2, \ldots, x_{jnj}$ for $j = 1, 2, \ldots, M$.

Recently, hybrid neural networks such as fuzzy neural networks and granular neural networks have been inves-

tigated to solve complex application problems with high dimensionality. Such hybrid neural networks provide remarkable abilities to derive knowledge from complicated or imprecise datasets. Therefore, they have already been applied to extract patterns and detect trends that are too complex to be noticed by either human beings or other computer techniques. In this case, the trained neural networks can be thought of as an "expert" in the category of information it has been given to analyze.

In general, the key technology for solving the curse of dimensionality (including structural dimensionality and parametric dimensionality) of a hybrid soft computing system, such as a Fuzzy Neural Network (FNN), is the divide-and-conquer method. In other words, an original single-stage FNN with high system complexity can be divided into small multi-stage soft computing building blocks with low system complexity so as to reduce both structural dimensionality and parametric dimensionality significantly.

In addition, GGCFNN provides parameters for deciding the number of iterations and termination criteria in order to control its learning performance. The hybrid genetic forward-wave-backward-wave learning algorithm is developed to enhance the system's learning quality. Our simulation results generated by genetic granular cognitive fuzzy neural network are compared to human persons and are analyzed in terms of computer science and cognitive science.

Granular Learning Algorithms

The traditional neural learning algorithms can generate a lot of numerical weights that are not meaningful (i. e., the black box problem). We plan to use traditional artificial intelligence, cognitive science and new computational intelligence including fuzzy logic, neural networks, evolutionary computation and granular computing to investigate the new evolutionary neural cognitive learning algorithms that can discover hidden decision-making knowledge from data (i. e., know how the human brain extracts rule-like relational knowledge from the inputs of experience and uses this knowledge in problem-solving and decision-making).

We plan to use traditional artificial intelligence, cognitive science and computational intelligence to investigate the new granular cognitive neural network architecture with cognitive neural granules with a small number of meaningful linguistic parameters. Each cognitive neural granule consists of a number of artificial cognitive neurons with nonlinear numerical/linguistic mapping functions. So cognitive neural granules are basic building blocks for the new granular cognitive neural network to effectively solve the problem of curse of dimensionality. The effective topological structure of the granular cognitive neural network will be investigated.

Suppose that the GCFNN has K inputs and one output. Given input data vectors X^p (i. e., $X^p = (X_1^p, X_2^p, \ldots, X_k^p)$) and an output data vector Y^p for $p = 1, 2, \ldots, L$. The global energy function is defined by

$$GE^p = \frac{1}{2}\left[f\left(X_1^p, \ldots, X_K^p\right) - Y^p\right]^2 . \tag{1}$$

The basic learning algorithm for the GCFNN was described in [19]. Here, for clarity and convenience, the relevant algorithms are introduced below. The new hybrid genetic-algorithms-based learning algorithm will be proposed in Subsect. "Forward-Wave-Backward-Wave Learning".

Local Forward-Wave Learning

In general, suppose that a FNN has n inputs and one output. Given input data vectors x^p (i. e., $x^p = (x_1^p, x_2^p, \ldots, x_n^p)$) and one-dimensional output data vector y^p for $p = 1, 2, \ldots, N$. The energy function is defined by

$$E^p = \frac{1}{2}\left[f(X_1^p, \ldots X_n^p) - y^p\right]^2 \tag{2}$$

For simplicity, let E and f^p denote E^p and $f(x_1^p, x_2^p, \ldots, x_n^p)$, respectively.

Based on the learning algorithm in [17,18], basic steps of the local forward-wave learning algorithm for a FNN are given below:

Procedure of Local-Forward-Wave-Learning

Step 1: Begin.

Step 2: Heuristic Initialization of Parameters.

Step 3: Gradient Descending Learning From Data.

Then we can get the following learning algorithms for $i = 1, 2, \ldots, n$, $k = 1, 2, \ldots, m$, $p = 1, 2, \ldots, N$, $t = 0, 1, 2, \ldots$ and a learning rate $\lambda > 0$.

Step 3.1: Train b^k

$$b^k(t+1) = b^k(t) - \lambda \frac{\partial E}{\partial b^k}|_t , \tag{3}$$

Step 3.2: Train η^k

$$\eta^k(t+1) = \eta^k(t) - \lambda \frac{\partial E}{\partial \eta^k}|_t , \tag{4}$$

Step 3.3: Train a_i^k

$$a_i^k(t+1) = a_i^k(t) - \lambda \frac{\partial E}{\partial a_i^k}|_t , \tag{5}$$

Step 3.4: Train σ_i^k

$$\sigma_i^k(t+1) = \sigma_i^k(t) - \lambda \frac{\partial E}{\partial \sigma_i^k}|_t \,, \tag{6}$$

Step 3.5: Train $w\text{left}_i^k$ or $w\text{right}_i^k$

$$\text{IF} \quad x_i^k \le a_i^k$$

$$\text{THEN} \quad w\text{left}_i^k(t+1) = w\text{left}_i^k(t) - \lambda \frac{\partial E}{\partial w\text{left}_i^k}|_t \,, \tag{7}$$

$$\text{ELSE} \quad w\text{right}_i^k(t+1) = w\text{right}_i^k(t) - \lambda \frac{\partial E}{\partial w\text{right}_i^k}|_t \,. \tag{8}$$

Step 4: End.

Global Backward-Wave Learning

Based on the traditional back-propagation learning method, the global backward-wave learning algorithm is proposed to train all local FNNs in the back-propagation manner. The following procedure is a general algorithm for updating parameters of any FNN in GCFNN.

Procedure of Global-Backward-Wave-Learning

Step 1: Begin.

Step 2: Calculate a back-propagation error δ^1 to the relevant $(l-1)$th stage FNNs based on an error $\delta^{(l+1)}$ from the $(l+1)$th stage FNN.

$$\delta^l = \frac{\partial E}{\partial y^{l-1}} = \delta^{(l+1)} \frac{\partial y^l}{\partial y^{l-1}} \,, \tag{9}$$

where y^1 and $y^{(l-1)}$ are outputs of the lth stage FNN and the $(l-1)$th stage FNN, respectively.

Step 3: Update parameters using the back-propagation error δ^1.

Step 3.5: Train $wleft_i^k$ or $wright_i^k$

Step 4: Discovering Fuzzy Knowledge.

Step 5: End.

Forward-Wave-Backward-Wave Learning

The new hybrid learning algorithm for training is called *forward-wave-backward-wave learning algorithm*, which combines the techniques of genetic algorithms, local forward-wave learning and global backward-wave learning.

Procedure Genetic-Forward-Wave-Backward-Wave-Learning

Step 1: Begin.

Step 2: Partition the original training data sets input data vectors X^p and an output data vector Y^p for $p = 1, 2, \ldots, L$ into M local training data sets where M is the number of 1st-stage granular neural networks.

Step 3: Local Forward-Wave Learning: *Local-forward-wave-learning()*.

Step 4: Global Backward-Wave Learning: *Global-backward-wave-learning()*.

Step 5: End.

Applications in Bioinformatics

A new granular neural network is used to form the new tertiary architecture [12]. The neurons in the new granular neural network are SVM machines that classify the input data into two classes of protein structures. The two classes are the binary classes that the SVM machines are actually trained for. The tertiary classifier's architecture is explained in subsequent sections.

The new tertiary classifier makes use of both one-versus-one as well as one-versus-rest binary classifiers. The novel architecture makes use of all the six binary classifiers in neural net architecture. The architecture is shown in the Fig. 2 [12].

There are two hidden layers; The output of the first one is the same as the output of the individual SVM.

Outputs of first hidden layer

$$O_1 = \text{SVM}(H/\sim H)$$
$$O_2 = \text{SVM}(E/\sim E)$$
$$O_3 = \text{SVM}(C/\sim C)\,.$$

The output of the second hidden layer considers the output of the first layer as well as the SVM machine stored

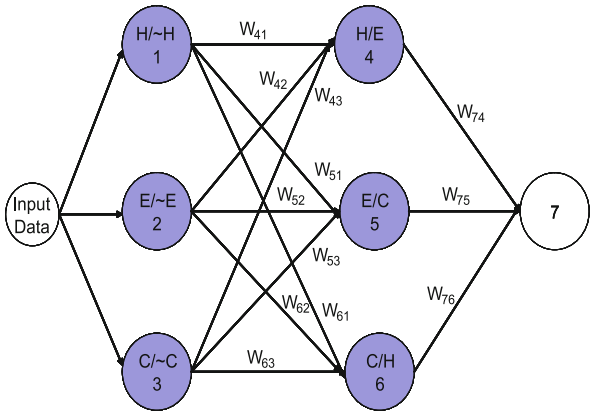

Granular Neural Network, Figure 2
Genetic Neural Support Vector Machines

in that layer. For example the output of the neuron 4 has an SVM binary classifier that positively classifies H and negatively classifies E, the result of this SVM is combined with that of the first layer outputs. This method uses the outputs of the three one-versus-rest classifier in a single neuron.

In the formulations the output of the second hidden layer is formed by adding the output of the SVM sitting inside the neuron with the product of the weight and output of the corresponding neuron (i. e. the neuron which positively classifies the same class as the current neuron) and by subtracting the products of the other two neurons in the first layer.

The output of the second layer are calculated as

$$O_4 = SVM(H/E) + W_{41}O_1 - W_{42}O_2 - W_{43}O_3 \,.$$

In the above formula we add the values of the SVM that positively classify the same class (H) and subtract those that positively classify other classes. Here W_{41} means weight between neuron 1 and neuron 4. Similarly other weight corresponds to output, input naming pattern.

Similarly outputs of other two neurons in the second hidden layer are calculated as

$$O_5 = SVM(E/C) + W_{52}O_2 - W_{51}O_1 - W_{53}O_3$$
$$O_6 = SVM(H/E) + W_{63}O_3 - W_{62}O_2 - W_{61}O_1 \,.$$

The final output layer does not have any SVM embedded in it. It calculates its results based on maximum of the three outputs of second hidden layer. There is only one neuron in this layer. The final output is one among the three classes (H, E or C), which ever neuron produces the maximum output after multiplying it with appropriate weight with the second hidden layer output is considered as final output.

So the output of the third layer is as follows

If (Max $(W_{74}O_4\,,\ \ W_{75}O_5\,,\ \ W_{76}O_6) = W_{74}O_4)$
Then
$O_7 = H$
Else If (Max $(W_{74}O_4\,,\ \ W_{75}O_5\,,\ \ W_{76}O_6 = W_{75}O_5))$
Then
$O_7 = E$
Else
$O_7 = C$

For optimizing the weights, genetic algorithms are used. The weight range for the Genetic Granular Neural Network (GGNN) is selected to be between 0 and 1 so that the architecture performs to its full potential.

The same tertiary classifiers were demonstrated with binary classifiers using multiple windows scheme. This resulted in increase in total accuracy level, which is expected as the binary classifiers formed using multiple window scheme are better when compared to single window encoding scheme. The binary classifier used is constructed using three consecutive windows each of size 5 with gaps between the first and second window as well as between second window and third window. The results of these simulations are shown in Table 1.

Granular Neural Network, Table 1
Accuracy Comparisons

Tertiary Classifier	Accuracy (%)
SVM_VOTE	62.6
SVM_REPRESNT.	64.8
GGNN	68.0

Granular Neural Network, Table 2
Meaning of data parameters

dd/mm/yyyy	Date	Date of the day
Double	Open	Initial price
Double	Low	Lowest traded price
Double	High	Highest traded price
Double	Close	Price of the last trade
Double	Volume	Number of traded stocks

Applications in Computational Web Intelligence

Concepts of Input Data

Database tables have been generated using stock-historical data from www.yahoo.com site. Each line of the data set must contain 5 values date, open, high, low, close values of the stock. Each value is separated by space (see Table 2) [21].

Overview of Implementation

A full run of the program implementation will be described, going through all the main features of the program [21]:

- Download historical data from the Internet.
- Copy the whole data into a text and run the program, which inserts the data into the database.
- A program is written through which users can buy and sell the stock shares, and the corresponding data is stored in the database.
- A program is written through which each user can see his/her own transactions.
- Algorithm, which trains the granular neural networks using the mean square error as stop criterion for learning, while never exceeding the maximum number of

cycles which can take testing data from the initial date to the user entered date, and predicts the future stock closing values.
- A program is written, which compares the predicted values with real values.

One of the most important factors here is to construct a neural network deciding on what the network will learn. A neural network must be trained on some input data. The two major problems in implementing the training are

- Defining the set of input to be used (the learning environment)
- Deciding on an algorithm.

Performance

The performance of the granular neural network algorithm is compared with the performance of the BP algorithm by training the same set of data and predicting the future stock values. If the training error was set at 0.03 and the neural network was trained for dow stock data using both the algorithms. The granular neural network took 2 min 58 s to train the neural network where as BP took 2 h and 55 min. The granular neural network's average error was 1.39 where as BP gave 3.38.

The average error for granular neural network is less compared to the average error for BP algorithm. From the average error and the graph it is conclusive that, granular neural network produced closer future stock values with the real stock values compared to the BP algorithm using less training error. If the training error was set at 0.07 and the neural network was trained for csco stock data using both the algorithms. The granular neural network took 2 min to train the neural network where as BP took 1 h and 48 min. The granular neural network's average error was 6.09 where as BP gave 7.16.

The average error for the granular neural network is less compared to the average error for BP algorithm. From the average error and the graph it is conclusive that, the granular neural network produced closer future stock values with the real stock values compared to the BP algorithm using less training error. Based on the above two simulations, the overall performance with the granular neural network technique is better than BP technique.

Applications in Brain Informatics

Recent advances in granular computing, soft computing and cognitive science have allowed an increase understanding of normal and abnormal brain functions, especially in the research of human's pattern recognition by means of computational intelligence.

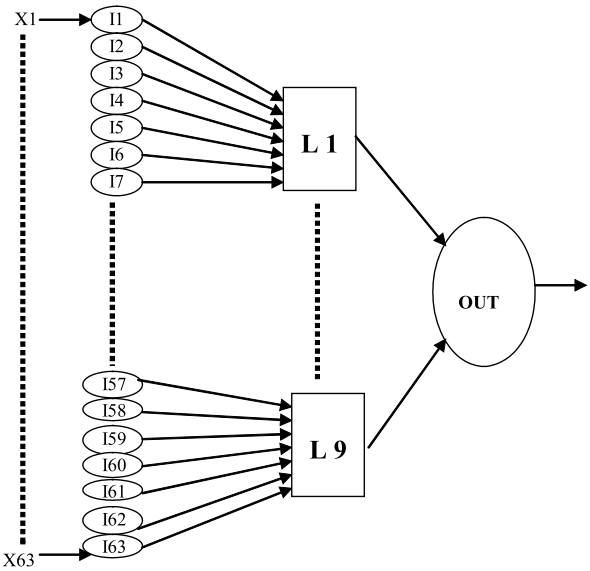

Granular Neural Network, Figure 3
The genetic granular neural network structure for pattern recognition [5]

The simulations of the genetic granular neural network for pattern recognition have two ways: (1) testing whether a pair of noised pattern images can still be identified by the genetic granular neural network as similar as its original version after horizontal or vertical transformation, and (2) testing whether the genetic granular neural network can recognize vertically or horizontally reversed pattern pairs with noise disturbance [5].

During each test, specific pattern samples are used for training and other groups of samples are served for testing. Each pattern sample is depicted by 7×9 pixel matrix, so totally 63 inputs are required in order to present one pattern, as shown in Fig. 3. To accommodate the input requirement, the genetic granular neural network's granular layer has 9 local genetic granular neural networks with 7 inputs, so that 63 pixel inputs are satisfied. In its second hidden layer, each the genetic granular neural network has one output; the output layer has 9 inputs and 1 output. Therefore, the total number of inputs of the system is 126 [5].

Finally, the system will make a final decision whether or not two images in each testing sample are similar or symmetric based on the maximum output of the three genetic granular neural networks. Human beings are chosen to make their subjective judgment on pattern samples, by answering what degree the two patterns are similar or symmetric to each other [5].

Group A contains 50 pairs of arrow pattern samples within the following five noise levels, which are 0–10%,

Granular Neural Network, Figure 4
a In Group A on the *left*, the second arrow pattern is transformed from the first after horizontal reversal; b In Group B on the *right*, the second cup-like pattern is transformed from the first after vertical reversal [5]

Granular Neural Network, Table 3
Simulation results

Group name	Different answers	Total samples	Agreement percentage
Group A	5	50	90%
Group B	22	50	56%
Total	27	100	73%

Granular Neural Network, Table 4
Different answers to vertical reversal within 5 noise levels

Noise level	Different answers	Total samples	Disagreement percentage
40–50% noise	0	10	0%
30–40% noise	0	10	10%
20–30% noise	1	10	10%
10–20% noise	1	10	10%
0–10% noise	3	10	30%

Granular Neural Network, Table 5
Different answers to horizontal reversal within 5 noise levels

Noise level	Different answers	Total samples	Disagreement percentage
40–50% noise	3	10	30%
30–40% noise	4	10	40%
20–30% noise	4	10	40%
10–20% noise	4	10	40%
0–10% noise	7	10	70%

10–20%, 20–30%, 30–40%, 40–50% respectively. In each pair of arrow pattern sample, as the Fig. 4a, the second pattern is transformed from the first pattern after horizontal reversal at the first, and then added into the same level of noises.

As shown in Fig. 4b, Group B includes 50 pairs of cup-like pattern samples within the same five noise levels as Group A. In each pair of cup-like-pattern sample, the second pattern is transformed from the first pattern after vertical reversal and then added into the same level of noises.

Table 3 shows the comparisons of decisions made by human beings and the genetic granular neural network for scenario1. As can be seen, the genetic granular neural network and the human brain make 73% identical choices with total 100 testing data. In Group A, Only 10% of 50 testing samples are different between the genetic granular neural network and human; while, in Group B, almost half of the testing samples, which are more than four times higher than Group A, introduce contradictive decisions. Thus, human and computer system in Group B make much more different answers than in Group A. It may indicate that the horizontal reversal causes more confusions than vertical reversal for human or the genetic granular neural network [5].

According to Table 4 and Table 5, we discovered an interesting phenomenon: As noise level increases, the disagreement on pattern's similarity between human and GGCFNN tends to decrease. More specifically, as shown in Table 4, disagreement percentage is exactly 0% in (40%, 50%) while it increases to 16.67% in (0%, 30%). In Table 5, disagreement percentage is increased from 36% within the noise level (20%, 50%) to 55% in (0%, 20%). Additionally, the disagreement percentages experience prominent increases or decreases within low noise levels.

The genetic granular neural network has exhibited good learning performance when stimulating human's pattern recognition in term of symmetry and similarity. It achieved higher agreement ratios when recognizing symmetrical patterns than similar patterns, and vertical transformation than horizontal transformation, even within high noise levels. The preliminary experiments indicated that the genetic granular neural network is able to simulate the human's recognition ability and adapt its learning experience to new patterns and then achieve quite similar results as human beings.

Conclusions

The human brain has high intelligence to recognize different geometrical patterns in terms of the similarity and symmetry, but a systematic framework of biological neural network has not been established. The granular neural network aims at simulating functionality of the human brain in terms of processing granular data such as linguistic terms, clusters, sets and classes.

Since the human brain consists of biological neural networks that are the major components performing intelligent tasks, it is very important to develop advanced algorithms and mathematical models for the granular neural networks.

Future Directions

In the future, the challenging problem is how to design more powerful granular neural networks for more complex applications effectively and efficiently. A lot of advanced intelligent techniques such as type-1/type-2 fuzzy sets [4,15], rough sets [6,7,8], granular computing [8,9], granular systems [16], granular support vector machines [13], granular kernel machines [3] and neurofuzzy systems [2,17,19] can be merged to design a hybrid granular neural networks with various functions and outstanding performance.

Bibliography

1. Dick S, Kandel A (2001) Granular weights in a neural network. In: IFSA World Congress and 20th NAFIPS International Conference, vol 3, Vancouver, July 2001, pp 1708–1713
2. Jang J-SR (1993) ANFIS: Adaptive-network-based fuzzy inference system. IEEE Trans Syst Man Cybern 23(3):665–685
3. Jin B, Zhang Y-Q, Wang BH (2007) Granular kernel trees with parallel genetic algorithms for drug activity comparisons. Int J Data Mining Bioinform 1(3):270–285
4. Karnik NN, Mendel JM, Liang Q (1999) Type-2 fuzzy logic systems. IEEE Trans Fuzzy Syst 7:643–658
5. Lin C, Li J, Barrett N, Zhang Y-Q, Washburn DA (2006) Genetic granular cognitive fuzzy neural networks and human brains for pattern recognition. In: The WICI International Workshop on Web Intelligence (WI) meets Brain Informatics (BI) (WImBI 2006), Beijing, 15–16 December 2006
6. Lin TY (1997) Granular computing: From rough sets and neighborhood systems to information granulation and computing in words. In: European Congress on Intelligent Techniques and Soft Computing, 8–12 September 1997, pp 1602–1606
7. Lin TY (1999) Granular computing: Fuzzy logic and rough sets. In: Zadeh L, Kacprzyk J (eds) Computing with Words in Information/Intelligent Systems. Physica, Heidelberg, pp 184–200
8. Lin TY, Yao YY, Zadeh L (2002) Data mining, rough sets and granular computing. Physica, Heidelberg
9. Pedrycz W (2001) Granular computing: An emerging paradigm. Physica, Heidelberg
10. Pedrycz W, Vukovich G (2001) Granular neural networks. Neurocomputing 36:205–224
11. Pedrycz W, Kandel A, Zhang Y-Q (1997) Neurofuzzy Systems. In: Dubois D, Prade H (eds) Fuzzy systems: Modeling and control. Kluwer, Norwell, pp 311–380
12. Reyaz-Ahmed A, Zhang Y-Q (2007) Protein secondary structure prediction using genetic neural support vector machines. In: Proc of IEEE 7th International Conference on BioInformatics and BioEngineering, Boston, 14–17 October 2007, pp 1355–1359
13. Tang YC, Jin B, Zhang Y-Q (2005) Granular support vector machines with association rules mining for protein homology prediction. Artif Intell Med (Special Issue on Computational Intelligence Techniques in Bioinformatics) 35(1–2):121–134
14. Vasilakos A, Stathakis D (2005) Granular neural networks for land use classification. Soft Comput 9:332–340
15. Zadeh LA (1979) Fuzzy sets and information granulation. In: Gupta N, Ragade R, Yager R (eds) Advances in fuzzy set theory and applications. Elsevier, St. Louis, pp 3–18
16. Zhang Y-Q (2005) Constructive granular systems with universal approximation and fast knowledge discovery. IEEE Trans Fuzzy Syst 13(1):48–57
17. Zhang Y-Q, Chung F (2002) Fuzzy neural network tree with heuristic back-propagation learning. In: Proc of IJCNN of World Congress on Computational Intelligence 2002, Honolulu, May 2002, pp 553–558
18. Zhang Y-Q, Kandel A (1998) Compensatory genetic fuzzy neural networks and their applications. In: Series in machine perception artificial intelligence, vol 30. World Scientific, Singapore
19. Zhang Y-Q, Kandel A (1998) Compensatory neurofuzzy systems with fast learning algorithms. IEEE Trans Neural Netw 9(1):83–105
20. Zhang Y-Q, Fraser MD, Gagliano RA, Kandel A (2000) Granular neural networks for numerical-linguistic data fusion and knowledge discovery. IEEE Trans Neural Netw 11(3):658–667
21. Zhang Y-Q, Akkaladevi S, Vachtsevanos G, Lin TY (2001) Fuzzy neural web agents for stock prediction. In: Proc of FLINT2001, UC Berkeley, 14–18 August 2001, pp 101–105

Granulation of Knowledge: Similarity Based Approach in Information and Decision Systems

Lech Polkowski
Polish-Japanese Institute of Information Technology, Warsaw, Poland

Article Outline

Glossary
Definition of the Subject
Introduction
Similarity
Information and Decision Systems
Granulation of Knowledge: General Ideas and Technique of Rough Inclusions
Calculus on Granules
Some Principal Fields of Applications
Complexity Issues
Future Directions
Bibliography

Glossary

Knowledge The notion of knowledge can be described by positing the structure of the perceivable world as a system of states of things. These states are related among themselves by relations which in turn form a network of interdependent combinations. States and relations,

as well as dependencies among them, are reflected in knowledge: things are reflected in objects or notions, relations in notions or concepts, states of things in sentences. Sentences form knowledge [6]; for instance, in Artificial Intelligence, a judiciously chosen set of sentences forms a knowledge base for a logical agent (see [52]).

Knowledge representation A chosen symbolic system (language) by means of which notions are encoded and reasoning is formalized.

Description logic A language of concepts commonly used for knowledge representation in Artificial Intelligence logical systems (see [2]). The variant which is interesting to us is constructed in the attribute-value context in which objects are described in terms of chosen attributes (features) and their values. Primitive formulas of *description logic* are *descriptors* of the form $(a = v)$ where a is an attribute and v is one of its values. Descriptors are interpreted in the universe U of objects: the meaning $[(a = v)]$ of the descriptor $(a = v)$ is defined as the set $\{u \in U\colon a(u) = v\}$. Descriptors are extended to formulas of description logic with the help of connectives of sentential calculus: when α, β are formulas then $\alpha \vee \beta$, $\alpha \wedge \beta$, $\neg\alpha$, $\alpha \Rightarrow \beta$ are formulas as well. The meaning of formulas is defined by recursion: $[\alpha \vee \beta] = [\alpha] \cup [\beta]$, $[\alpha \wedge \beta] = [\alpha] \cap [\beta]$, $[\neg\alpha] = U \setminus [\alpha]$. The meaning of implication $\alpha \Rightarrow \beta$ depends on the chosen interpretation of implication; when it is interpreted as in sentential calculus, that is, as $\neg\alpha \vee \beta$, the meaning is already defined by the given formulas for $\vee, \neg$.

Rough sets: knowledge as classification Rough set theory proposed by Pawlak [30,32], understands knowledge as classification: a set of equivalence relations, each of which is understood as a certain classification of objects into its equivalence classes. Together, these relations constitute the knowledge base $\mathcal{R}$. Objects that belong in one class $[u]_R$ of a relation R in the knowledge base are *R-indiscernible*. Objects that belong in the intersection $\bigcap_{R\in\mathcal{R}}[u]_R$ are *$\mathcal{R}$-indiscernible*: they cannot be discerned by means of the knowledge available in $\mathcal{R}$.

Indiscernibility relations are formally defined for sets of relations: given $\mathcal{T} \subseteq \mathcal{R}$, the $\mathcal{T}$-indiscernibility relation $\mathrm{ind}(\mathcal{T})$ is defined as the intersection $\bigcap_{R\in\mathcal{T}} R$. Concepts are divided into two classes: *$\mathcal{T}$-exact concepts* are sets of objects which can be expressed as unions of $\mathcal{T}$-indiscernibility classes; other concepts are *$\mathcal{T}$-rough*. For each concept X, there exist the greatest exact concept $\underline{\mathcal{T}}$ contained in X (the *$\mathcal{T}$-lower approximation to X*) and the smallest exact concept $\underline{\mathcal{T}}$ which contains X (the *$\mathcal{T}$-upper approximation to X*). Rough concepts are perceived as uncertain sets sandwiched between a lower approximation and an upper approximation.

Information system A system commonly used for representing information about a given world fragment. An information system can be represented as a collection of pairs of the form (U, A) where U is a set of objects – representing things – and A is a set of attributes; each attribute a is modeled as a mapping $a\colon U \to V_a$ from the set of objects into the value set V_a. For an attribute a, the equivalence relation R_a is the set $\{(u, v) \in U \times U\colon a(u) = a(v)\}$ and the collection $\mathcal{R} = \{R_a\colon a \in A\}$ is the knowledge base. For each set $B \subseteq A$ of attributes, the *B-indiscernibility relation* $\mathrm{ind}(B)$ is the set $\{(u, v) \in U \times U\colon a(u) = a(v)$ for each $a \in B\}$. Each object $u \in U$ can be described in two ways: first, by means of its indiscernibility class $[u]_A = \{v \in U\colon (u, v) \in \mathrm{ind}(A)\}$; next, by means of its descriptor logic formula $\phi^A(u)\colon \bigwedge_{a\in A}(a = a(u))$; as $[\phi^A(u)] = [u]_A$, both descriptions are equivalent and u is described up to its equivalence class.

Decision system A particular form of an information system, a triple (U, A, d) in which d is the *decision*, the attribute not in A, that expresses the evaluation of objects by an external oracle, an expert. Attributes in A are called *conditional* which expresses the fact that they are set by us as conditions describing objects in our language. The principal problem related to a decision system is to find a description of dependence between conditional knowledge given by (U, A) and the expert knowledge $(U, \{d\})$.

Decision rule A *decision rule* is a formula in descriptor language that expresses a particular relation among conditional attributes in the attribute set A and the decision d, of the form: $\bigwedge_{a\in A}(a = v_a) \Rightarrow (d = v)$ with the semantics defined in (Glossary: Description logic). The formula is *true*, or *certain* in case $[\bigwedge_{a\in A}(a = v_a)] = \bigcap_{a\in A}[(a = v_a)] \subseteq [(d = v)]$. Otherwise, the formula is *partially true*, or *possible*. An object o which satisfies the rule $a(o) = v_a$ for $a \in A$ can be assigned to the class $[(d = v)]$; often a partial match based on a chosen similarity measure must be performed.

Similarity Similarity relations are weaker forms of indiscernibility: indiscernible objects are similar but not necessarily vice versa an example attributable to Henri Poincaré [35] explains this best: assume that points x, y in the real line are similar whenever their distance is less than a fixed threshold δ: $|x - y| < \delta$. Then certainly each x is similar to x, and if x is similar to y then

y is similar to x, but from x similar to y and y similar to z it need not follow that x is similar to z: $|x - z|$ can be close to 2δ. Relations of this kind are reflexive and symmetric: they were called *tolerance relations* in Zeeman [67]. Tolerance relations in the framework of rough set theory information systems were considered in Nieminen [26] and tolerance reducts and related notions in information systems were discussed in Polkowski, Skowron and Zytkow [50]. Some authors relax tolerance relations to similarity relations that need not be symmetric: for instance, rough inclusions need not be symmetric [37].

In many applications, metrics are taken as similarity measures: a metric on a set U is a function $d\colon U \times U \to R^+$ into non-negative reals such that 1. $d(x, y) = 0$ if and only if $x = y$. 2. $d(x, y) = d(y, x)$. 3. $d(x, y) \leq d(x, z) + d(z, y)$ for each x, y, z in X (the *triangle inequality*); similarity induced by a metric d is usually of the form $d(x, y) < \delta$ for a certain $\delta > 0$, Poincaré-style.

Rough inclusion A ternary relation μ on a set $U \times U \times [0, 1]$ which satisfies conditions: 1. $\mu(x, x, 1)$. 2. $\mu(x, y, 1)$ is a binary partial order relation on the set X. 3. $\mu(x, y, 1)$ implies that for each object z in U: if $\mu(z, x, r)$ then $\mu(z, y, r)$. 4. $\mu(x, y, r)$ and $s < r$ imply that $\mu(x, y, s)$. The formula $\mu(x, y, r)$ is read as "the object x is a part in object y to a degree at least r". The partial containment idea encompasses the idea of an exact part, that is, mereological theory of concepts Leśniewski [15]. Similarity induced by a rough inclusion can be introduced as an asymmetric relation: $\mu(x, y, r)$, or a symmetric one: $\mu(x, y, r) \wedge \mu(y, x, r)$.

Granulation of knowledge The concept of granulation was proposed by Zadeh [67]. In fuzzy calculus, granules are naturally induced as the inverse images of fuzzy membership functions and fuzzy computing is a fortiori computing with granules. A similar case happens with rough sets, as primitive objects in this paradigm are indiscernibility classes that are elementary granules, whereas their unions are granules of knowledge; reasoning in rough sets means, by definition, reasoning with granules. Each similarity relation induces granules understood as its classes; in comparison to indiscernibility relations, granules induced by similarity relations are more intricate due to lack of transitivity: relations among granules are more difficult to ascertain. In defining granules, attention is focused on the property that singles out objects belonging in the granule; usually, this property is a relation to an object chosen as the granule center: this means that distinct objects in the granule need not be in the relation. This feature distinguishes granules from *clusters*: aggregates in which any two objects have the defining property, for example, distance less than the chosen threshold. Also, relations between pairs of distinct granules are difficult to ascertain, contrary to the case of (for example) clusters, in which case the distance between distinct clusters is kept above a fixed threshold.

Classification of objects Classification entails assigning to each element in a set of objects (test sample) a class (a decision) to which the given element should belong; it is accomplished on the basis of knowledge induced from the given collection of examples (the training sample). To perform this task, objects are usually mapped onto vectors in a multi-dimensional real vector space (feature space). Paradigms invented to perform classification are many, from heuristics based on cognitive networks such as neural networks, through probabilistic tools such as Bayesian networks to tools based on distance such as k-nn methods and prototype methods (see [7,12]).

Fusion of knowledge A combination of knowledge coming from few distinct sources. For example, in a mobile robot, fusion of knowledge by means of a Kalman filter (see [52]), consists in producing an output – a robot's predicted position at the current moment – on the basis of inputs: a robot's position estimated as the last moment preceding the current moment, the current controls, and sensor readings in the current position.

Reasoning Processes of reasoning include an effort by means of which sentences are created; various forms of reasoning depend on the chosen system of notions, symbolic representation of notions, forms of manipulating symbols (see [6]).

Mereology Mereology, see Leśniewski [15], is a theory of concepts based on the notion of a part instead – as with naive set theory – on the notion of an element. The relation of being a part is irreflexive and transitive. A secondary notion of an ingredient means either being a part or being the whole object and it sets a partial order on objects. By means of the notion of an ingredient the notion of a class operator is introduced in Leśniewski [15] which converts ontological notions, that is, collections of objects into an individual object; it is used by us in granule formation.

Definition of the Subject

The creator of Fuzzy Set Theory, Lotfi A. Zadeh, proposed computing with granules in Zadeh [66]. The idea was natural, as fuzzy reasonings are carried out in terms of fuzzy

membership functions. A fuzzy membership function μ_X maps a universe U of objects into the interval $[0, 1]$ and it represents the membership in a set X as a membership to a degree. The value $\mu_X(x) = r$ is interpreted as the statement that the object x is an element of the set X to the degree of r. The mapping $U \to U/\mu_X$ which sends each object x to its fibre $\mu_X^{-1}(\mu_X(x))$ includes into the granule $g_\mu(r)$ all objects that belong in X to the degree r. All fuzzy constructs are then expressed in terms of those granules. In this sense, fuzzy reasoning is, in a natural way, reasoning with granules. Granules in this reasoning are constructed in a uniform way, that is, all objects in a granule share the same property of external character: they belong to an "oracle" X to the same degree; changing X produces a variety of granules to reason with. The relation forming any granule is an equivalence R_X: the universe is decomposed into fibres of a fuzzy membership function μ_X, and a knowledge base is composed of all relations R_X for subsets $X \subseteq U$ of the universe of objects.

The same conclusion concerns rough set reasoning: elementary objects in reasoning are indiscernibility classes – elementary granules which are elements of a partition of the universe of objects by an indiscernibility relation ind(B) (see Glossary: Information systems). These granules are the smallest objects which can be described in terms of attributes and their values, that is, by description logic, and they are used in forming description of objects, in building decision rules and then classifiers as well as control algorithms. Lin [16,17] was the first to recognize the topological character of granules and to form the basic notion of a neighborhood system as a collection of granules on the universe of objects (see also [18,19,20,21,22]). Lin [17] recognized the import of tolerance relations by discussing tolerance induced neighborhoods.

In all hybrid approaches involving fuzzy or rough sets along with neural networks, genetic algorithms etc., one is therefore bound to compute with granules, testimony to the importance of granular structures.

In search of adequate similarity relations, various forms of granules have been proposed and considered as well as experimentally verified as to their effectiveness. In information systems, *templates* were proposed as granules, that is, generalized descriptors of the form $(a \in W_a)$ where $W_a \subseteq V_a$ with the meaning $[(a \in W_a)] = \{u \in U\colon a(u) \in W_a\}$ [27]. Clearly, templates are aggregates in the ontological sense of descriptors, that is, they form "big" granules. Their usage is motivated by their greater descriptive force *vis-a-vis* descriptors; a judicious choice of sets W_a should allow construction of a similarity relation that accurately reflects the decision's dependence on conditional attributes.

Rough inclusions have been considered as a means for granule construction; a rough inclusion is (see Glossary: Rough inclusions) a generic term for a ternary predicate that formalizes the phrase: "the object x is a part of the object y to a degree r", introduced in [47,48]. Granulation by means of rough inclusions has been studied, for example, in [37,38,39,40,41,42,43,44]. The idea of granule formation comes form mereology and it rests on using the class operator (see Glossary: Mereology).

Granules formed by rough inclusions are used in fusion of knowledge models, in rough-neural computing and in building many-valued logics reflecting the rough set ideology in reasoning.

Introduction

Granulation of knowledge can be considered from a few angles:

1. General purpose of granulation.
2. Granules from binary relations.
3. Granules in information systems from indiscernibility.
4. Granules from generalized descriptors.
5. Granules from rough inclusions – mereological approach.

General Purpose of Granulation

Granulation of knowledge comes into existence for a few reasons: the principal one is founded on the underlying assumption of basically all paradigms, *viz.*, that reality exhibits a fundamental continuity, that is, objects with identical descriptions in a given paradigm should exhibit the same properties with respect to classification or decision making. For instance, fuzzy set theory assumes that objects with identical membership descriptions should behave identically and rough set theory assumes that objects indiscernible with respect to a group of attributes should behave identically; in particular, they should fall into the same decision class.

Thus, granulation is implied by assumptions of the paradigm and is unavoidable once the paradigm is accepted and applied. Granules induced in the given paradigm by necessity form the first level of granulation.

In the search for the most appropriate similarities for use in applications such as classification or decision making, more complex granules are constructed, for example, as unions of granules of the first level, or more complex functions of them, resulting, for example, in fusion of various granules from distinct sources.

Among granules of the first two levels, some kinds can be exhibited by means of various operators, for example, the class operator associated with a rough inclusion.

Granules from Binary Relations

Granulation on the basis of binary general relations as well as, for example, tolerance relations, has been studied by T. Y. Lin [16,17,18,19,20,21,22,23] in particular as an important notion of neighborhood systems; see also Yi Yu Yao [64,65]. This extends the approach based on indiscernibility as a special case. A general form of this approach, according to Yao, exploits the classical notion of Galois connection: two mappings $f: X \to Y, g: Y \to X$ form a Galois connection between ordered sets $(X, <)$ and $(Y, \prec)$ if and only if the equivalence $x < g(y) \Leftrightarrow f(x) \prec y$ holds. In an information system (U, A), for a binary relation R on U, one considers the sets $xR = \{y \in U: xRy\}$ and $Rx = \{y \in U: yRx\}$, called, respectively, the *successor neighborhood* and the *predecessor neighborhood* of x. These sets are considered as granules formed by a specific relation of being affine to x in the sense of the relation R. Other forms of granulation can be obtained by comparing objects with identical neighborhoods, for example, $x \equiv y \Leftrightarrow xR = yR$ etc. Saturation of sets of objects X, Y with respect to the relation R leads to sets X^*, Y^* such that $X * R = Y* \Leftrightarrow RY^* = X^*$, forming a Galois connection. This approach is closely related to the Formal Concept Analysis of Wille [61].

Granules in Information Systems from Indiscernibility

Granules based on indiscernibility in information/decision systems are constructed as indiscernibility classes: given an information system (U, A) (see Glossary: Information systems; Decision systems), and the collection $\mathrm{IND} = \{\mathrm{ind}(B): B \subseteq A\}$ of indiscernibility relations, each elementary granule is of the form $[u]_B = \{v \in U: (u, v) \in \mathrm{ind}(B)\}$ for some B. Among those granules, there are minimal ones: granules of the form $[u]_A$ induced from the set A of all attributes. Granules $[u]_A$ form the finest partition of the universe U; given the class $[u]_B$ and the class $[u]_{A\setminus B}$, we have $[u]_A = [u]_B \cap [u]_{A\setminus B}$ hence $[u]_B = \bigcup_{v\in[u]_B\cap \mathrm{DIS}_{A\setminus B}(u)}[v]_A$, where $v \in \mathrm{DIS}_{A\setminus B}(u)$ if and only if there exists an attribute $a \in A \setminus B$ such that $a(u) \neq a(v)$. It is manifest that granules $[u]_B$ can be arranged into a tree with the root $[u]_\emptyset = U$ and leaves of the form $[u]_A$.

Granules based on indiscernibility form a complete Boolean algebra generated by atoms of the form $[u]_A$: unions of these atomic granules are closed on intersections and complements and these operations induce into granules the structure of a field of sets.

Atomic granules are aggregated into some important unions by approximation operators: given a concept $X \subseteq U$, the lower approximation $\underline{B}X$ to X over the set B of attributes is defined as the union $\bigcup\{[u]_B: [u]_B \subseteq X\}$; the operator L_B: Concepts $\to$ Granules sending X to $\underline{B}X$ is monotone increasing and idempotent: $X \subseteq Y$ implies $LX \subseteq LY$ and $L \circ L = L$. Similarly, the upper approximation $\overline{B}X = \bigcup\{[u]_B: [u]_B \cap X \neq \emptyset\}$ to X over B, brings some elementary granules into the union; the operator U^B sending concepts into upper approximations is also monotone increasing and idempotent.

Granules from Generalized Descriptors

Some authors have made use of generalized descriptors called *templates* (see Nguyen S H [27]). A template is a formula $T: (a \in W_a)$, where $W_a \subseteq V_a$ (see Glossary: Information systems) with the meaning $g(T): \{u \in U: a(u) \in W_a\}$. The granule $g(T)$ can be represented as the union $\bigcup_{u\in W_a}[u]_a$; granules of the form $[u]_a$ are also called *blocks*, see Grzymala-Busse [10,11].

Granules from Rough Inclusions – Mereological Approach

Rough inclusions (see Glossary: Rough inclusions) are ternary predicates of the form $\mu(x, y, r)$ which read: "x is a part of y to degree at least r"; intuitively, one can expect of μ the fulfillment of conditions:

a. $\mu(x, x, 1)$;
b. $\mu(x, y, 1)$ if and only if $x \,\mathrm{ing}_\pi\, y$, where π is a part relation (see Glossary: Mereology) and ing_π is the associated ingredient relation: $x \,\mathrm{ing}_\pi\, y$ if and only if $x \,\pi\, y$ or $x = y$;
c. If $\mu(x, y, 1)$ then for each z: if $\mu(z, x, r)$ then $\mu(z, y, r)$ which is a condition of monotonicity of μ;
d. If $r > s$ and $\mu(x, y, r)$ then $\mu(x, y, s)$.

Granule $g_r(u)$ of the radius r about the center u is defined as the class of property $\Phi(v)$: $\mu(v, u, r)$. The class operator

$$\mathrm{Cls}: \text{Ontological entities (collections of objects)} \to \text{Mereological entities (individual objects)}$$

acts on collections, or properties of objects and yields individual objects; an example of such a class operator in set theory with the relation $\subset$ of part and the associated relation $\subseteq$ of ingredient is the union of a family of sets operator $\bigcup$: it converts a family (a property) of sets $\mathcal{F}$ into its union $\bigcup \mathcal{F}$ which is a single set.

The class operator Cls applied to a non-vacuous property Φ of objects, yields the object $\mathrm{Cls}\,\Phi$ which is defined

as the unique object with the properties:

1. If $\Phi(u)$ then $\mathrm{ing}_\pi \mathrm{Cls}\,\Phi$;
2. If $u\,\mathrm{ing}_\pi \mathrm{Cls}\,\Phi$ then there exist objects v, w such that $v\,\mathrm{ing}_\pi\, u$, $v\,\mathrm{ing}_\pi\, w$, $\Phi(w)$.

In other words, $\mathrm{Cls}\,\Phi$ absorbs all objects whose each part has a part in common with an object which satisfies Φ.

Hence: $g_r(u)$ is $\mathrm{Cls}\{v\colon \mu(v,u,r)\}$. Properties of granules defined in this way depend on properties of rough inclusion μ.

There are many rough inclusions to select from, and most are defined from one of three sources (see Polkowski [37,38,39,40,41,42,43,44])

A From archimedean t-norms: an archimedean t-norm t see Polkowski [36], admits a representation $t(x,y) = g_t(f_t(x) + f_t(y))$ with a continuous decreasing $f_t\colon [0,1] \to [0,1]$ and its pseudo-inverse g_t, see Ling [24] or Polkowski [36].
We define the rough inclusion μ_t as follows:

$$\mu_t(u,v,r) \quad \text{if and only if} \quad g_t\left(\frac{|\,\mathrm{DIS}_A(u,v)|}{|A|}\right) \geq r\,,$$

where $\mathrm{DIS}_A(u,v) = \{a \in A\colon a(u) \neq a(v)\}$.
The most important example of rough inclusions obtained in this way is the rough inclusion μ_L induced from the Łukasiewicz t-norm $t_L(x,y) = \max\{0, x + y - 1\}$ for which $f(x) = 1 - x$ and $g(y) = 1 - y$, see Ling [24] or Polkowski [36]. Introducing the complement $\mathrm{IND}_A(u,v) = U \times U \setminus \mathrm{DIS}_A(u,v)$, one can write the rough inclusion μ_L as,

$$\mu_L(u,v,r) \quad \text{if and only if} \quad \frac{|\,\mathrm{IND}_A(u,v)|}{|A|} \geq r$$

Thus, objects u, v are similar at least to degree r if and only if the probability of randomly choosing an attribute which does not discern between u and v is at least r. This rough inclusion is also induced by the relative Hamming distance on information sets of objects in the set U of the information system (see [43]).

B From continuous t-norms (see [43]): a continuous t-norm t induces the *residual implication* $\Rightarrow_t$ by the formula,

$$x \Rightarrow_t y \leq z \quad \text{if and only if} \quad t(x,z) \leq y\,,$$

and it is well-known that $x \Rightarrow_t y = 1$ if and only if $x \leq y$. Thus, residual implications are proper candidates for rough inclusions, once proper set functions $\Phi(u,v), \Psi(u,v)$ are found such that $\mu_t(u,v,r)$ holds if and only if $\Phi(u,v) \Rightarrow_t \Psi(u,v) \geq r$ holds. Examples are given in Polkowski [43].

C It is evident that for a metric d on the universe U, the predicate $\mu_d(u,v,r)$ is satisfied if and only if $d(u,v) \leq 1 - r$ is a rough inclusion. In particular, for the Hamming metric relative to the set A of attributes, $h(u,v) = \frac{|\{a \in A\colon a(u) \neq a(v)\}|}{|a|}$, the rough inclusion μ_h coincides with μ_L.
A rough inclusion μ_t is *transitive* in cases where following the property holds:

$$\text{If } \mu_t(u,v,r) \text{ and } \mu(v,w,s) \text{ then } \mu(u,w,t(r,s))\,.$$

It has been proved that rough inclusions obtained by methods 1 and 2 are transitive [43]. One can also verify that rough inclusions obtained from metrics by method 3 are transitive with the Łukasiewicz rough inclusion as the transitivity measure: for a metric d and $\mu_d(u,v,r), \mu_d(v,w,s)$, it follows that $d(u,v) \leq 1 - r$ and $d(v,w,s) \leq 1 - s$ and the triangle inequality implies that $d(u,w) \leq (1-r) + (1-s)$ hence $\mu_d(u,w,r+s-1)$, that is, $\mu_d(u,w,t_L(r,s))$.
Reasoning about granules defined from rough inclusions is carried out on lines of the mereological deduction rule (Leśniewski [15]):
(D(eduction) R(ule) in M(ereology)) For objects x, y: if for each object z, from $z\,\mathrm{ing}_\pi\, x$ it follows that there exists an object w such that $w\,\mathrm{ing}_\pi\, z$ and $z\,\mathrm{ing}_\pi\, y$ then $x\,\mathrm{ing}_\pi\, y$.

Similarity

Similarity relations came to the attention of eminent theorists in the beginning of the twentieth century due to their interest in problems regarding mechanisms of thinking and perception. Henri Poincaré [35] considered relations of discernibility, for example, two points in space are *indiscernible* if and only if their distance is less than a fixed threshold value δ. This relation is certainly reflexive (a point cannot be discerned from itself) and symmetric, but may lack the transitivity property; again, in connection with perception mechanisms, these relations were called *tolerance relations* in [67].

Formally, a *tolerance relation* τ is a relation which is reflexive: $\tau(x,x)$ for each x, and symmetric: if $\tau(x,y)$ then $\tau(y,x)$ for each pair x, y.

Tolerance relations are weaker than equivalence relations of, for example, indiscernibility (see Glossary: Information systems); lack of transitivity makes them more difficult to analyze. In analogy to equivalence relations, one

introduces *tolerance classes*: a tolerance class c_τ is defined as a maximal set with the property: for each pair $x, y \in c_\tau$ the relation $\tau(x, y)$ holds.

Contrary to equivalence classes, distinct tolerance classes can intersect: the collection of all pairwise distinct tolerance classes is a *covering* of the universe U of objects.

From the maximality principle in set theory (Teichmuller's, Vaught's or Zorn's, see [14]) it follows that for each pair x, y with $\tau(x, y)$ there exists a tolerance class $c_{\tau(x,y)}$ which contains both x, y.

We denote by the symbol $C_\tau(x)$ the collection of all tolerance classes which contain x. Then the principal property of tolerance relations holds: For each pair x, y of objects in the universe U and a tolerance relation τ on this universe: $\tau(x, y)$ if and only if $C_\tau(x) \cap C_\tau(y) \neq \emptyset$.

This fact shows that a canonical tolerance relation is defined as the non-empty intersection of sets, and each tolerance relation can be reduced to it.

Tolerance relations in information systems should factor through indiscernibility: if $\tau(u, v)$ and $u \operatorname{ind}(A) u'$, $v \operatorname{ind}(A) v'$ then $\tau(u', v')$. Thus, tolerance relations in information systems should be defined in terms of attributes and their values.

Examples are: metrics-induced tolerance relations, for example, $\tau(u, v)$ if and only if $\max_a |a(u) - a(v)| < \delta$ for a fixed value of δ (tolerance based on the Manhattan metric) or $\tau(u, v)$ if and only if $\sum_a (|a(u) - a(v)|^2)^{\frac{1}{2}} < \delta$ (tolerance based on the Euclidean metric).

Rough inclusions also supply examples of tolerance relations: given a rough inclusion μ, and a real number $r \in [0, 1]$, the tolerance relation $\tau_{\mu,r}$ can be defined as follows: $\tau_{\mu,r}(u, v)$ if and only if $\mu(u, v, r)$ and $\mu(v, u, r)$.

Weaker yet is the class of similarity relations which are merely reflexive: this kind of similarity relations encompasses relations of rough inclusions induced by asymmetric part relations as well as relations discussed in Słowiński et al. [8].

A reflexive similarity relation τ can be described in terms of classes $C_\tau(x) = \{y \in U : \tau(x, y)\}$ for $x \in U$. Then: $\tau(x, y)$ if and only if $y \in C_\tau(x)$. Thus, reflexive similarity relations can be canonically described as relations of membership between objects and their collections.

Information and Decision Systems

The scope of our discussion is limited to knowledge representation in the form of information systems (see Glossary: Information systems, Decision systems, Description logic, Decision rules).

Information systems are means for formal rendering of data usually given in the form of data tables: rows in the table correspond to objects being described, columns correspond to features used in object descriptions.

Each data table involves the set U of objects, called often the *universe of objects* and the set A of *attributes* formally representing features used in object descriptions. The pair (U, A) is called the *information system*.

For an object $u \in U$ and a set of attributes B, the *information set* of u is the set $\inf_B(u) = \{(a = a(u)) : a \in B\}$.

Thus, the universe U can be represented as the set $\operatorname{Inf}_B = \{\inf_B(u) : u \in U\}$.

A principal assumption of the information system (U, A) is that all information about objects is supplied, and constructs derived from it should extend to the world of possible objects.

Therefore: objects u, v which satisfy the condition $\inf_B(u) = \inf_B(v)$ are regarded as *indiscernible* over B and perceived through the filter of attributes in B should be treated as identical with regard to description over B.

Formal rendering of this postulate is effected by means of *B-indiscernibility relation* $\operatorname{ind}(B)$ (see Glossary: Rough sets: knowledge as classification): $\operatorname{ind}(B) = \{(u, v) \in U \times U : \inf_B(u) = \inf_B(v)\}$. This is an equivalence relation.

Attribute sets produce a variety of indiscernibility relations with obvious properties such as: $\operatorname{ind}(C) \subseteq \operatorname{ind}(D)$ if and only if $D \subseteq C$; $\operatorname{ind}(C \cap D) = \operatorname{ind}(C) \cap \operatorname{ind}(D)$; $\operatorname{ind}(A) \subseteq \operatorname{ind}(C) \subseteq \operatorname{ind}(\emptyset) = U \times U$ for each $C \subseteq A$.

Classes $[u]_B = \{v \in U : (u, v) \in \operatorname{ind}(B)\}$ are minimal sets which can be described exactly over the set B: the class $[u]_B$ is the meaning of the descriptor formula $\bigwedge_{a \in B}(a = a(u))$.

Indiscernibility relations form the knowledge base from which inferences are derived about relations among objects, attributes and attribute sets.

An important example of a property of attribute sets is the reduct property [31,54]: a set B of attributes is a *reduct* if and only if B is a minimal (with respect to set inclusion) set of attributes such that $\operatorname{ind}(B) = \operatorname{ind}(A)$. Thus, each reduct fully preserves both classification and knowledge.

The meaning of a reduct is also shown by means of dependencies among sets of attributes. The *functional dependency* between sets C and D of attributes, see [28,29,31], means that there exists a mapping $f_{C,D} : \operatorname{Inf}_C \to \operatorname{Inf}_D$ with the property that $f(\inf_C(u)) = \inf_D(u)$ for each $u \in U$. In this case, D *depends functionally* on C, in symbols, $C \mapsto D$.

Functional dependence $C \mapsto D$ means equivalently that $\operatorname{ind}_C \subseteq \operatorname{ind}_D$; in particular, as $\operatorname{ind}_B \subseteq \operatorname{ind}_C$ for each reduct B, the functional dependence $B \mapsto C$ takes place for each set C of attributes: each reduct functionally determines all values of attributes.

Finding reducts is computationally difficult, see [54]. In theory, reducts can be found as prime implicants of the Boolean discernibility function. Decision systems are (see Glossary: Decision systems) information systems of the form $(U, A \cup \{d\})$ where $d \notin A$ is a *decision*.

The most important problem concerning decision systems consists in finding a set of *decision rules* (see Glossary: Decision rules) which adequately describes the decision attribute in terms of attributes in A (*conditional attributes*).

Classification methods can be divided, according to the adopted methodology, into classifiers based on reducts and decision rules, classifiers based on templates and similarity, classifiers based on descriptor search, classifiers based on granular descriptors, and hybrid classifiers.

For a decision system (U, A, d), classifiers are sets of decision rules. Induction of rules has been a subject of research in rough set theory since its beginning. In most general terms, building a classifier consists of searching the pool of descriptors for conjuncts that describe decision classes sufficiently well. As distinguished in Stefanowski [58], there are three main kinds of classifiers to search for: *minimal*, that is, consisting of minimum possible number of rules describing decision classes in the universe, *exhaustive*, that is, consisting of all possible rules, and *satisfactory*, that is, containing rules tailored to a specific use. Classifiers are evaluated globally with respect to their ability to properly classify objects, usually by *error* which is the ratio of the number of correctly classified objects to the number of test objects, *total accuracy* being the ratio of the number of correctly classified cases to the number of recognized cases, and *total coverage*, the ratio of the number of recognized test cases to the number of test cases.

Minimum size algorithms include the LEM2 algorithm of Grzymala-Busse [10,11] and the covering algorithm in the RSES package [51]; exhaustive algorithms include, for example, the LERS system of Grzymala-Busse [9], systems based on discernibility matrices and Boolean reasoning by Skowron [53], Bazan [3], Bazan et al. [4], implemented in the RSES package [51].

Minimal consistent sets of rules were introduced in Skowron and Rauszer [54]; they were shown to coincide with rules induced on the basis of local reducts in Wróblewski [63]. Further developments include dynamic rules, approximate rules, and relevant rules as described in [3,4] as well as local rules [3,4], effective in implementations of algorithms based on minimal consistent sets of rules. Rough set based classification algorithms, especially those implemented in the RSES system [51], were discussed extensively in [3]; Skowron and Stepaniuk [55] and Skowron and Swiniarski [56] discussed rough set classifiers along with some attempt at analysis of granulation in the process of knowledge discovery.

In [3], a number of techniques were verified in experiments with real data, based on various strategies:

`discretization` of attributes (codes: N-no discretization, S-standard discretization, D-cut selection by dynamic reducts, G-cut selection by generalized dynamic reducts);
`dynamic selection of attributes` (codes: N-no selection, D-selection by dynamic reducts, G-selection based on generalized dynamic reducts);
`decision rule choice` (codes: A-optimal decision rules, G-decision rules on basis of approximate reducts computed by Johnson's algorithm, simulated annealing and Boltzmann machines etc., N-without computing of decision rules);
`approximation of decision rules` (codes: N-consistent decision rules, P-approximate rules obtained by descriptor dropping);
`negotiations among rules` (codes:S-based on strength, M-based on maximal strength,R-based on global strength, D-based on stability).

Any choice of a strategy in particular areas yields a compound strategy denoted with the alias being concatenation of symbols of strategies chosen in consecutive areas, for example, NNAND etc.

We record here in Table 1 an excerpt from the comparison (Table 8, 9, 10 in [3]) of the best of these strategies with results based on other paradigms in classification for two sets of data: Diabetes and Australian credit from the UCI Repository [55].

An adaptive method of classifier construction was proposed in Wróblewski [63]; reducts are determined by means of a genetic algorithm and in turn reducts induce sub-tables of data regarded as classifying agents; choice of optimal ensembles of agents is done by a genetic algorithm.

Classifiers constructed by means of similarity relations are based on templates matching a given object, or closest to it, with respect to a certain distance function, or on coverings of the universe of objects by tolerance classes and assigning the decision value on basis of some of them [27]. We include in Table 2 excerpts from classification results in [27].

A combination of rough set methods with the k-nearest neighbor idea is a further refinement of classification based on similarity or analogy in Wojna [62]. In this approach, training set objects are endowed with a metric, and the test objects are classified by a vote of k nearest training objects for some k that is subject to optimization.

Granulation of Knowledge: Similarity Based Approach in Information and Decision Systems, Table 1
A comparison of errors in classification by rough set and other paradigms

Paradigm	System/method	Diabetes	Austr. credit
Stat. Methods	Logdisc	0.223	0.141
Stat. Methods	SMART	0.232	0.158
Neural Nets	Backpropagation2	0.248	0.154
Neural Networks	RBF	0.243	0.145
Decision Trees	CART	0.255	0.145
Decision Trees	C4.5	0.270	0.155
Decision Trees	ITrule	0.245	0.137
Decision Rules	CN2	0.289	0.204
Rough Sets	NNANR	0.335	0.140
Rough Sets	DNANR	0.280	0.165
Rough Sets	best result	0.255(DNAPM)	0.130 (SNAPM)

Granulation of Knowledge: Similarity Based Approach in Information and Decision Systems, Table 2
Accuracy of classification by template and similarity methods

Paradigm	System/method	Diabetes	Austr. credit
Rough Sets	Simple.templ./Hamming	0.6156	0.8217
Rough Sets	Gen.templ./Hamming	0.742	0.855
Rough Sets	Simple.templ./Euclidean	0.6312	0.8753
Rough Sets	Gen.templ./Euclidean	0.7006	0.8753
Rough Sets	Match.tolerance	0.757	0.8747
Rough Sets	Clos.tolerance	0.743	0.8246

Granulation of Knowledge: General Ideas and Technique of Rough Inclusions

Granulation of knowledge means that, given a representation of knowledge, for example, an information system, a fortiori, the set of information vectors of objects in its universe, these information vectors will undergo the process of aggregation, that is, of granule formation (granules being collections of information vectors). The process of granulation should be driven by a similarity relation on information vectors, and as the result of this process, similar to a satisfactory degree, information vectors/objects will find themselves in one single granule. Parameters of this process are an object u around which the granule is formed (the granule center) and the degree $r \in [0, 1]$ to which objects in the granule are similar to u.

A General Form of Similarity

For our purposes of granulation of knowledge, we introduce a general form of graded similarity in the form a relation $\tau \subseteq U \times U \times [0, 1]$ on triples of the form (u, v, r) with $u, v \in U$, the universe of an information system (U, A), and $r \in [0, 1]$. We require of the similarity τ the following:

1. $\tau(u, u, 1)$;
2. $r < s$ implies if $\tau(u, v, s)$ then $\tau(u, v, r)$;
3. $\tau(u, v, 1)$ is a partial ordering of the universe U.

Granulation: Tools

Given a similarity τ, we form a basic granule $g_\tau(u, r)$ of the radius r about an object u, by collecting together all objects v with the property that $\tau(v, u, r)$. Then, by definition of τ, one has the following properties of granules:

a. $u \in g_\tau(u, r)$ for each $r \in [0, 1]$;
b. $v \in g_\tau(u, r)$ implies $v \in g_\tau(u, s)$ for each $s < r$;

Granulation thus requires:

- a similarity measure,
- an aggregation mechanism.

Granulation by Means of Rough Inclusions

In this section, we apply the general ideas derived in Sect. "Similarity" and Sect. "Information and Decision Systems" in a specific context of rough inclusions (see Glossary: Rough Inclusions).

Mereological Theory of Concepts As a first step to granulation by means of rough inclusions, we introduce the reader to the mereological theory of concepts which is the basis of our approach to granulation.

Mereology, as conceived by Stanisław Leśniewski [15], is based on the primitive binary relation of *part*, π, on the universe U of objects, which is subject to the following requirements:

(P1) $u \pi v$ implies $u \neq v$;
(P2) $u \pi v$ and $v \pi w$ imply $u \pi w$.

Thus, the part relation π is non-reflexive and transitive.

Given a part relation, which by (P1) means a proper part relation, one forms an improper part counterpart, called the ingredient, ing, defined from the part relation as follows:

(I) v ing u if and only if $v \pi u$ or $v = u$.

Thus, the ingredient relation is a partial ordering on the universe U:

(U1) u ing u for each u;
(U2) u ing v and v ing u imply $v = u$;
(U3) u ing v and v ing w imply u ing w.

Mereology provides an operator, the *class operator*, whose purpose is to represent collections of objects as an individual object; clearly, this will be also our principal granulation tool when formal properties of granules are involved. The class operator Cls is applied to any non-empty collection (a property) Φ of individual objects in U, making this collection into an individual Cls Φ; its definition is as follows Leśniewski [15]:

(Cl1) If $v \in \Phi$ then v ing Cls Φ;
(Cl2) If v ing Cls Φ then for some $w, z \in U$ one has that w ing v, w ing z, $z \in \Phi$.

The object Cls Φ is unique for a given Φ.

As an example, the reader will verify that the relation $\subset$ of the proper set/concept containment is a part relation, the relation $\subseteq$ of the improper containment is the adjoint relation of an ingredient, and given a non-empty family of sets Φ, the union $\bigcup \Phi$ is the class Cls Φ with respect to the ingredient relation $\subseteq$. Thus, the mereological context generalizes the standard set-theoretical context adopted prevalently by computer science.

Rough Inclusions In the process of development of rough set theory, it has turned out that indiscernibility should be relaxed to similarity: in Polkowski, Skowron and Zytkow [50], attention was focused on tolerance relations, that is, relations which are reflexive and symmetric but need not be transitive. An example of such a relation was given in Poincaré [35]: given a metric ρ and a fixed small positive δ, one declares points x, y in the relation sim(δ) if and only if $\rho(x, y) < \delta$. The relation sim(δ) is a tolerance relation but it is equivalence for non-archimedean ρ's only, that is, when $\rho(x, y) \leq \max\{\rho(x, z), \rho(z, y)\}$.

We continue this example by introducing a graded version of sim(δ), viz., for a real number $r \in [0, 1]$, we define the relation sim(δ, r) by letting,

$$\mathrm{sim}(\delta, r)(x, y) \text{ iff } \rho(x, y) \leq 1 - r\,. \tag{1}$$

The collection sim(δ, r) of relations have the following properties evident by the properties of the metric ρ:

(SIM1) sim(δ, 1)(x, y) iff $x = y$;
(SIM2) sim(δ, 1)(x, y) and sim(δ, r)(z, x) imply sim(δ, r)(z, y);
(SIM3) sim(δ, r)(x, y) and $s < r$ imply sim(δ, s)(x, y).

Properties (SIM1)–(SIM3) induced by the metric ρ refer to the ingredient relation $=$ whose corresponding relation of part is empty; a generalization can thus be obtained by replacing the identity with an ingredient relation ing in a mereological universe (U, π).

Consequently, a relation $\mu(u, v, r)$ is defined that satisfies the following conditions:

(RM1) $\mu(u, v, 1)$ iff u ing v;
(RM2) $\mu(u, v, 1)$ and $\mu(w, u, r)$ imply $\mu(w, v, r)$;
(RM3) $\mu(u, v, r)$ and $s < r$ imply $\mu(u, v, s)$.

Any relation μ which satisfies the conditions (RM1)–(RM3) is called a *rough inclusion*. This relation is a similarity relation which is not necessarily symmetric, but is reflexive. It is read as "the relation of a part to a degree".

Rough Inclusions: Case of Information Systems The problem of methods by which rough inclusions could be introduced in information/decision systems has been studied in Polkowski [37,38,39,40,41,42,43,44]. Here we recapitulate the results and add new ones. We recall that an *information system* is a method of representing knowledge about a certain phenomenon in the form of a table of data; formally, it is a pair (U, A) where U is a set of *objects* and A is a set of *conditional attributes*; any object $u \in U$ is described by means of its *information set* $\mathrm{Inf}(u) = \{(a, a(u)) : a \in A\}$. The indiscernibility relation Ind, definable sets and non-definable sets are defined from Inf as indicated in (Glossary: Information Systems).

Rough Inclusions from Metrics As observed in Sect. "Introduction", any metric ρ defines a rough inclusion μ_ρ by means of the equivalence $\mu_\rho(u, v, r) \Leftrightarrow \rho(u, v) \leq 1 - r$.

A very important example of a rough inclusion obtained on these lines is the rough inclusion μ_h with $h(u, v)$ being the reduced Hamming distance on information vectors of u and v, that is, $h(u, v) = \frac{|\{a \in A : (a, a(u)) \neq (a, a(v))\}|}{|A|}$, $|X|$ denoting the cardinality of the set X.

Thus, $\mu_h(u,v,r)$ if and only if $h(u,v) \leq 1-r$; introducing sets $\mathrm{DIS}(u,v) = \{a \in A\colon (a,a(u)) \neq (a,a(v))\}$ and $\mathrm{IND}(u,v) = A \setminus \mathrm{DIS}(u,v) = \{a \in A\colon a(u) = a(v)\}$, one can write down the formula for μ_h either as,

$$\mu_h(u,v,r) \Leftrightarrow \frac{|\,\mathrm{DIS}(u,v)|}{|A|} \leq 1-r\,, \tag{2}$$

or,

$$\mu_h(u,v,r) \Leftrightarrow \frac{|\,\mathrm{IND}(u,v)|}{|A|} \geq r\,. \tag{3}$$

The formula (3) witnesses that the rough inclusion μ_h is an extension of the indiscernibility relation Ind to a graded indiscernibility.

In a similar manner one should be able to compute rough inclusions induced by other metrics standardly used on information sets such as Euclidean, Manhattan etc.

Rough inclusions induced by metrics possess an important property of *functional transitivity* expressed in general form by the rule,

$$\frac{\mu_\rho(u,v,r), \mu_\rho(v,w,s)}{\mu_\rho(u,w,L(r,s))}\,, \tag{4}$$

where $L(r,s) = \max\{0, r+s-1\}$ is the Łukasiewicz t-norm, see Polkowski [66]. We offer a short proof of this fact: assuming that $\mu_\rho(u,v,r), \mu_\rho(v,w,s)$ which means in terms of the metric ρ that $\rho(u,v) \leq 1-r, \rho(v,w) \leq 1-s$; by the triangle inequality, $\rho(u,w) \leq (1-r)+(1-s)$, that is, $\mu_\rho(u,w,r+s-1)$.

Rough Inclusions from Functors of Many-Valued Logics A functor (t-norm) $t\colon [0,1]\times[0,1] \to [0,1]$ is *non-archimedean* only when the equality $t(x,x) = x$ holds for $x = 0,1$; it is known (see e.g. [36]) that the only such t-norms are the Łukasiewicz L and the product t-norm $P(x,y) = x \cdot y$.

Each of these t-norms admits a functional representation: $t(x,y) = g(f(x)+f(y))$ (see, for example, Polkowski [36]).

One defines a rough inclusion μ_t by letting (see Polkowski [35,37,38,39,40,41,42,43]),

$$\mu_t(u,v,r) \Leftrightarrow g\left(\frac{|\,\mathrm{DIS}(u,v)|}{|A|}\right) \geq r\,. \tag{5}$$

In particular, in case of the t-norm L, one has $g(x) = 1-x$ (see, for example, [36]), and thus the rough inclusion μ_L is expressed by means of the formula (3).

Another systematic method for defining rough inclusions is by means of residual implications of continuous t-norms(see Polkowski [38,39,40,41,42,43,44]).

For a continuous t-norm t, the *residual implication* $x \Rightarrow_t y$ is a mapping from the square $[0,1]^2$ into $[0,1]$ defined as follows(see, for example, Polkowski [36]),

$$x \Rightarrow_t y \geq z \;\text{ iff }\; t(x,z) \leq y\,; \tag{6}$$

thus, $x \Rightarrow_t y = \max\{z\colon t(x,z) \leq y\}$.

Proposition 1 *The residual implication $x \Rightarrow_t y$ does induce a rough inclusion $\mu_t^{\Rightarrow}$ by means of the formula: $\mu_t^{\Rightarrow}(x,y,r)$ if and only if $x \Rightarrow_t y \geq r$ for every continuous t-norm t.*

We include a short argument for the sake of completeness; clearly, $\mu_t^{\Rightarrow}(x,x,1)$ holds as $x \Rightarrow_t y \geq 1$ is equivalent to $x \leq y$. Assuming $\mu_t^{\Rightarrow}(x,y,1)$, that is, $x \leq y$, and $\mu_t^{\Rightarrow}(z,x,r)$, that is, $z \Rightarrow x \geq r$ hence $t(z,r) \leq x$ we have $t(z,r) \leq y$, that is, $z \Rightarrow y \geq r$ so finally $\mu_t^{\Rightarrow}(x,y,r)$. Clearly, by definition, from $\mu_t^{\Rightarrow}(x,y,r)$ and $s < r$ it does follow that $\mu_t^{\Rightarrow}(x,y,s)$.

We recall here the basic cases of rough inclusions obtained from the most frequently applied t-norms. In all cases, $\mu_t^{\Rightarrow}(x,y,1)$ if $x \leq y$, so the associated ing relation is $\leq$ and the underlying part relation is $<$. For $r < 1$, that is, $x > y$, one has

Case 1 $t = L$; in this case $x \Rightarrow_L y = \min\{1, 1-x+y\}$, hence $\mu_L^{\Rightarrow}(x,y,r)$ if and only if $1-x+y \geq r$.

Case 2 $t = P$ where $P(x,y) = x \cdot y$; in this case, $x \Rightarrow_P y = \frac{y}{x}$ when $x \neq 0$ and 1 when $x = 0$ hence $\mu_P^{\Rightarrow}(x,y,r)$ if and only if $y \geq x \cdot r$.

Case 3 $t = \min(x,y)$; in this case $x \Rightarrow_{\min} y$ is y hence $\mu(x,y,r)$ if and only if $y \geq r$.

Transitivity of Rough Inclusions It has been proved [37,38,39,40,41,42,43,44] that all rough inclusions induced from either non-archimedean or continuous t-norms t in the manner as above are transitive in the sense of the formula:

$$\frac{\mu(u,v,r), \mu(v,w,s)}{\mu(u,w,t(r,s))}\,.$$

Applications to Granulation of Knowledge Formal theory of rough inclusions allows for a formal mechanism of granulation of knowledge; we assume an information system (U,A) is given. Granulation of knowledge, proposed as a paradigm by L.A. Zadeh [66], means grouping objects into collections called granules, objects within a granule being similar with respect to a chosen measure; granular computing means computing with granules in place of objects.

The mechanism of granule formation based on rough inclusions has been presented by the author in a few works (see [37,38,39,40,41,42,43,44]) and we recall it here. We assume an information system (U, A) is given. The basic tool in establishing properties of granules is the class operator of mereology, along with the Leśniewski deduction rule (DRM) of Sect. "Introduction".

Given a rough inclusion μ on the universe U, for each object u and each $r \in [0, 1]$, the granule $g_\mu(u, r)$ of the radius r about u relative to μ is defined as the class of the property $\Phi(u, r, \mu) = \{v\colon \mu(v, u, r)\}$:

$$g_\mu(u, r) \text{ is } \mathrm{Cls}\,\Phi(u, r, \mu)\,. \tag{7}$$

In case of t-norm-induced rough inclusions, by their transitivity, the following important property holds (see Polkowski [43]):

$$v \text{ ing } g_\mu(u, r) \;\text{ iff }\; \mu(v, u, r)\,, \tag{8}$$

That is, the granule $g_\mu(u, r)$ is $\{v\colon \Phi(u, r, \mu)(v)\}$ with no synergy effect.

Rough Inclusions on Granules because mereology operates (due to the class operator) only on the level of individuals, one can extend rough inclusions from objects to granules. The formula for extending a rough inclusion μ to a rough inclusion $\overline{\mu}$ on granules is a modification of the mereological deduction rule of Sect. "Introduction": (DRMG): $\overline{\mu}(g, h, r)$ if and only if for each object z with z ing g there exists an object w with w ing h such that $\mu(z, w, r)$.

The fact that $\overline{\mu}$ is a rough inclusion can be established by means of the rule (DRMG).

Modifications and Variants of Rough Inclusions In applications to be presented in this special section, some modified rough inclusions or weaker similarity measures will be instrumental. We include a discussion of them here.

Modification by Means of Metrics on Attribute Values For the rough inclusion μ_L, the formula $\mu_L(v, u, r)$ means that $\frac{|\mathrm{IND}(v,u)|}{|A|} \geq r$, that is, at least $r \cdot 100$ percent of attributes agree on u and v; an extension of this rough inclusion depends on a chosen metric ρ bounded by 1 in the attribute value space V (we assume a simple case that ρ works for each attribute).

Then, given an $\varepsilon \in [0, 1]$, we let $\mu^\varepsilon(v, u, r)$ if and only if $|\{a \in A\colon \rho(a(v), a(u)) < \varepsilon\}| \geq r \cdot |A|$; it is manifest that μ^ε is a rough inclusion if ρ is a non-archimedean metric, that is, $\rho(u, w) \leq \max\{\rho(u, v), \rho(v, w)\}$; otherwise the monotonicity condition rm2 of Sect. "Calculus on Granules" need not be satisfied and this takes place with most popular metrics (Euclidean, Manhattan etc.).

In this case, a remedy is to define a rough inclusion μ^* as follows: $\mu^*(v, u, r)$ if and only if there exists an ε such that $\mu^\varepsilon(v, u, r)$. It is easy, then, to verify that μ^* is a rough inclusion.

Modification by transfer Assume that $\Rightarrow_t$ is chosen, and for an information system (U, A), with an ingredient relation ing on U, a mapping $\phi\colon U \to [0, 1]$ is given such that $\phi(u) \leq \phi(v)$ if and only if u ing v. Then, the relation,

$$\mu_\phi(v, u, r) \;\text{ iff }\; \phi(u) \Rightarrow_t \phi(v) \geq r\,, \tag{9}$$

is a rough inclusion on U. We include a short proof of this fact: that $\mu_\phi(u, v, 1)$ is equivalent to $\phi(u) \leq \phi(v)$ hence to u ing v was observed in Subsect. "Granules from rough inclusions – mereological approach" b. Assuming that $\mu_\phi(u, v, 1)$ and $\mu_\phi(w, u, r)$, the proof that (RM2) (see Sect. "Granulation of Knowledge: General Ideas and Technique of Rough Inclusions") holds, that is, $\mu_\phi(w, v, r)$ proceeds like the proof of (RM2) in Proposition 1 of Sect. "Granulation of Knowledge: General Ideas and Technique of Rough Inclusions". Finally, the property (RM3) is evident.

As ϕ depends on one argument, let us select an object $s \in U$, and consider:

1. $\phi_1 = \mathrm{dis}(u) = \frac{|\{a \in A\colon a(s) \neq a(u)\}|}{|A|}$;
2. $\phi_2 = \mathrm{ind}(u) = \frac{|\{a \in A\colon a(s) = a(u)\}|}{|A|}$;
3. $\phi_3 = \mathrm{dis}_\varepsilon(u) = \frac{|\{a \in A\colon \rho(a(s), a(u)) \geq \varepsilon\}|}{|A|}$;
4. $\phi_4 = \mathrm{ind}_\varepsilon(u) = \frac{|\{a \in A\colon \rho(a(s), a(u)) \leq \varepsilon\}|}{|A|}$, where ρ is a chosen metric on the set of attribute values V, and ε is a chosen threshold in $[0, 1]$.

Proposition 2 *In all cases $i = 1, 2, 3, 4$, the relation $\mu_{\phi_i}(v, u, r)$ defined with ϕ_i as ϕ is a rough inclusion.*

The reference object s can be chosen as an "ideal object," for example, whose conditional class is contained in its decision class, etc. The formula (9) can be generalized by considering a set $S = \{s_1, \ldots, s_k\}$ of reference objects.

Comparison of objects u, v on the lines of this section, do not necessarily lead to rough inclusions due to a possible violation of property (RM2); yet, such variants are of importance as they allow for a direct comparison among objects, rules and granules. We introduce a generalization of rough inclusions:

For given objects u, v, and $\varepsilon \in [0,1]$, we introduce the factors:

$$\mathrm{dis}_\varepsilon(u,v) = \frac{\left|\{a \in A\colon \rho(a(u), a(v)) \geq \varepsilon\}\right|}{|A|} ,$$

and

$$\mathrm{ind}_\varepsilon(u,v) = \frac{\left|\{a \in A\colon \rho(a(u), a(v)) < \varepsilon\}\right|}{|A|} ,$$

where ρ is a metric on attribute value sets.

Then, we modify the formula (9) to the form,

$$\nu(u,v,r) \ \text{ iff } \ \mathrm{dis}_\varepsilon(u,v) \rightarrow_t \mathrm{ind}_\varepsilon(u,v) \geq r . \tag{10}$$

Clearly, ν has properties: 1. $\nu(u,u,1)$; 2. $\nu(u,v,r)$ and $s < r$ imply $\nu(u,v,s)$ but monotonicity property (RM2) need not hold. Rough inclusions and their weak variants will be essentially exploited in data mining tasks presented in the sequel.

Calculus on Granules

We recall that we define a granule $g_\mu(u,r)$ about $u \in U$ of the radius r, relative to the rough inclusion μ, as follows:

$$g_\mu(u,r) \text{ is } \mathrm{Cls}_\mu\, \Pi^\mu(u,r) , \tag{11}$$

where $\Pi^\mu(u,r)$ is the property defined by means of,

$$\Pi^\mu(u,r)(v) \ \text{ iff } \ \mu(v,u,r) . \tag{12}$$

General properties of granules are collected below,

$$\begin{aligned}
&1.\ \text{if } y \text{ ing } x \ \text{ then } \ y \text{ ing } g_r x;\\
&2.\ \text{if } y \text{ ing } g_r x \text{ and } z \text{ ing } y \ \text{ then } \ z \text{ ing } g_r x;\\
&3.\ \text{if } \mu(y,x,r) \ \text{ then } \ y \text{ ing } g_r x;\\
&4.\ \text{if } s < r \ \text{ then } \ g_r x \text{ ing } g_s x ,
\end{aligned} \tag{13}$$

which follow from properties (RM1)–(RM3) of rough inclusions (see sect. V) and the fact that ing is a partial order, and, in particular, is transitive.

More generally,

$$\text{if } y \text{ ing } g_r x \ \text{ then } \ g_s y \text{ ing } g_{T(r,s)} x . \tag{14}$$

The last statement follows directly from the transitivity property of rough inclusions and from the deduction rule (DRM) of Sect. "Introduction".

It is natural to regard a granule system $\{g_r^{\mu_t}(x)\colon x \in U; r \in (0,1)\}$ as a neighborhood system for a topology on U that may be called the *granular topology*; we refer here to the idea of a neighborhood system introduced by Lin [16,17].

In order to make this idea explicit, we define classes of the form $N^T(x,r) = \mathrm{Cls}(\psi_{r,x}^{\mu_T})$, where

$$\psi_{r,x}^{\mu_T}(y) \Leftrightarrow \exists s > r . \mu_T(y,x,s) . \tag{15}$$

We declare the system $\{N^T(x,r)\colon x \in U; r \in (0,1)\}$ to be a neighborhood basis for a topology θ_μ. This is justified by the following

Theorem 1 *Here are properties of the system* $\{N^T(x,r)\colon x \in U; r \in (0,1)\}$*:*

$$\begin{aligned}
&1.\ y \text{ ingr } N^t(x,r) \ \Rightarrow \ \exists \delta > 0 . N^t(y,\delta) \text{ ingr } N(x,r);\\
&2.\ s > r \ \Rightarrow \ N^t(x,s) \text{ ingr } N^t(x,r);\\
&3.\ z \text{ ingr } N^t(x,r) \wedge z \text{ ingr } N^t(y,s) \ \Rightarrow \ \exists \delta > 0.\\
&\quad N^t(z,\delta) \text{ ingr } N^t(x,r) \wedge N^t(z,\delta) \text{ ingr } N^t(y,s) .
\end{aligned} \tag{16}$$

An argument for (16) is as follows:

For Property 1. y ingr $N^t(x,r)$ implies that there exists an $s > r$ such that $\mu_t(y,x,s)$. Let $\delta < 1$ be such that $t(u,s) > r$ whenever $u > \delta$; δ exists by continuity of t and the identity $t(1,s) = s$. Thus, if z ingr $N^t(y,\delta)$, then $\mu_t(z,y,\eta)$ with $\eta > \delta$ and $\mu_t(z,x,t(\eta,s))$ hence z ingr $N^t(x,r)$.

Property 2. follows by (RM3) and Property 3 is a corollary to properties 1 and 2. This concludes the argument for (16).

Granule systems as defined above form a basis for applications where approximate reasoning is a crucial ingredient.

We begin with a basic application in which approximate reasoning itself is codified as a many-world (intensional) logic where granules serve as possible worlds.

Some Principal Fields of Applications

Applications of the proposed mechanism of granulation have been indicated in many areas of reasoning under uncertainty, in particular in:

- Reasoning about uncertainty in information/decision systems by means of rough mereological granular logics;
- Reasoning in distributed systems (fusion of knowledge);
- Reasoning in cognitive schemes (rough neural hybrid systems);
- Granular classifiers.

We devote some space to these applications.

Reasoning About Uncertainty

The idea of a granular rough mereological logic (see [38, 48]) consists of measuring the meaning of a unary predicate in the model, which is a universe of an information system, against a granule defined to a certain degree by means of a rough inclusion. The result can be regarded as the degree of truth (the logical value) of the predicate with respect to the given granule. The logics thus obtained are intensional as they can be regarded as mappings from the set of granules (possible worlds) to the set of logical values in the interval $[0, 1]$, the value at a given granule regarded as the extension at that granule of the generally defined intension (see [60] for a general introduction to intensional logics).

For our purpose it is essential to extend rough inclusions to sets; we use the t-norm L along with the representation $L(r, s) = g(f(r) + f(s))$ already mentioned. We denote these kind of inclusions with the generic symbol ν.

For finite sets X, Y, we let,

$$\nu_L(X, Y, r) \text{ iff } g\left(\frac{|X \setminus Y|}{|X|}\right) \geq r\,; \tag{17}$$

as $g(x) = 1 - x$, see Polkowski [7], we have that $\nu_L(X, Y, r)$ holds if and only if $\frac{|X \cap Y|}{|X|} \geq r$. Let us observe that ν_L is *regular*, that is, $\nu_L(X, Y, 1)$ if and only if $X \subseteq Y$ and $\nu_L(X, Y, r)$ only with $r = 0$ if and only if $X \cap Y = \emptyset$.

Thus, the ingredient relation associated with a regular rough inclusion is the improper containment $\subseteq$, whereas the underlying part relation is the strict containment $\subset$.

Other rough inclusion on sets we we exploit is the 3-valued rough inclusion ν_3 defined via the formula, see Polkowski [38],

$$\nu_3(X, Y, r) \text{ iff } \begin{cases} X \subseteq Y \text{ and } r = 1 \\ X \cap Y = \emptyset \text{ and } r = 0 \\ r = \frac{1}{2} \text{ otherwise}\,. \end{cases} \tag{18}$$

We assume that an information/decision system (U, A, d) is given, along with a rough inclusion ν on the subsets of the universe U; for a collection of predicates (unary) Pr, interpreted in the universe U (meaning that for each predicate $\phi \in Pr$ the meaning $[\phi]$ is a subset of U), we define the intensional logic grm_ν on Pr by assigning to each predicate ϕ in Pr its intension $I_\nu(\phi)$ defined by its extension $I_\nu^\vee(g)$ at particular granules g, as,

$$I_\nu^\vee(g)(\phi) \geq r \text{ iff } \nu(g, [\phi], r)\,. \tag{19}$$

With respect to the rough inclusion ν_L, the formula (19) becomes,

$$I_{\nu_L}^\vee(g)(\phi) \geq r \text{ iff } \frac{|g \cap [\phi]|}{|g|} \geq r\,. \tag{20}$$

The counterpart for ν_3 is specified by definition (18).

We say that a formula ϕ interpreted in the universe U of an information system (U, A) is *true* at a granule g with respect to a rough inclusion ν if and only if $I_\nu^\vee(g)(\phi) = 1$.

Thus, for every regular rough inclusion ν, a formula ϕ interpreted in the universe U, with meaning $[\phi]$, is true at a granule g with respect to ν if and only if $g \subseteq [\phi]$. In particular, for a decision rule $r\colon p \Rightarrow q$ in the descriptor logic, the rule r is true at a granule g with respect to a regular rough inclusion ν if and only if $g \cap [p] \subseteq [q]$.

The formula $\nu(g, [\phi], r) = 1$ stating the truth of ϕ at g, ν with ν regular can be regarded as a condition of orthogonality type, with the usual consequences.

1. If ϕ is true at granules g, h then it is true at $g \cup h$.
2. If ϕ is true at granules g, h then it is true at $g \cap h$.
3. If ϕ, ψ are true at a granule g then $\phi \vee \psi$ is true at g.
4. If ϕ, ψ are true at a granule g then $\phi \wedge \psi$ is true at g.
5. If ψ is true at a granule g then $\phi \Rightarrow \psi$ is true at g for every formula ϕ.
6. If ϕ is true at a granule g then $\phi \Rightarrow \psi$ is true at g if and only if ψ is true at g.

The graded relaxation of truth is given obviously by the condition, a formula ϕ is *true to a degree at least r at g*, ν if and only if $I_\nu^\vee(g)(\phi) \geq r$, that is, $\nu(g, [\phi], r)$ holds. In particular, ϕ is *false* at g, ν if and only if $I_\nu^\vee(g)(\phi) \geq r$ implies $r = 0$, i. e. $\nu(g, [\phi], r)$ implies $r = 0$. Properties 1–6 above suggest that the notion of truth in rough mereological logics has similar properties with respect to connectives of logical calculi to those of classical sentential calculus. Therefore, we introduce the following semantics of sentential connectives $\neg, \vee, \wedge, \Rightarrow$.

1. $[\neg\alpha] = U \setminus [\alpha]$.
2. $[\alpha \vee \beta] = [\alpha] \cup [\beta]$.
3. $[\alpha \wedge \beta] = [\alpha] \cap [\beta]$.
4. $[\alpha \Rightarrow \beta] = (U \setminus [\alpha]) \cup [\beta]$.

With respect to these semantics, the following properties hold.

1. For each regular ν, a formula α is true at g, ν if and only if $\neg\alpha$ is false at g, ν.
2. For $\nu = \nu_L, \nu_3$, $I_\nu^\vee(g)(\neg\alpha) \geq r$ if and only if $I_\nu^\vee(g)(\alpha) \geq s$ implies $s \leq 1 - r$.
3. For $\nu = \nu_L, \nu_3$, the implication $\alpha \Rightarrow \beta$ is true at g if and only if $g \cap [\alpha] \subseteq [\beta]$ and $\alpha \Rightarrow \beta$ is false at g if and only if $g \subseteq [\alpha] \setminus [\beta]$.
4. For $\nu = \nu_L$, if $I_\nu^\vee(g)(\alpha \Rightarrow \beta) \geq r$ then $\Rightarrow_L (t, s) \geq r$ where $I_\nu^\vee(g)(\alpha) \geq t$ and $I_\nu^\vee(g)(\beta) \geq s$.

The functor $\Rightarrow_L$ in 4. is the Łukasiewicz implication of many-valued logic: $\Rightarrow_L (t, s) = \min\{1, 1 - t + s\}$.

Further analysis should be split into the case of ν_L and the case of ν_3 as the two differ essentially with respect to the form of reasoning they imply.

Reasoning with ν_L The last property 4. shows in principle that the value of $I_\nu^\vee(g)(\alpha \Rightarrow \beta)$ is bounded from above by the value of $\Rightarrow_L (I_\nu^\vee(g)(\alpha), I_\nu^\vee(g)(\beta))$.

This suggests that the idea of collapse due to S. Lesniewski can be applied to formulas of rough mereological logic in the following form: for a formula $q(x)$ we denote by the symbol q^* the formula q regarded as a sentential formula (that is, with variable symbols removed) subject to relations $\neg q(x)^*$ is $\neg(q(x)^*)$ and $p(x) \Rightarrow q(x)^*$ is $p(x)^* \Rightarrow q(x)^*$. As the value $[q^*]_g$ of the formula $q(x)^*$ we admit the value of $\frac{|g \cap [q(x)]|}{|g|}$. Thus, the item 4 can be rewritten in the form.

$$I_\nu^\vee(g)(\alpha \Rightarrow \beta) \leq \Rightarrow_L ([\alpha^*]_g, [\beta^*]_g) . \tag{21}$$

The following statement is then obvious: if $\alpha \Rightarrow \beta$ is true at g then the collapsed formula has a truth value of 1 at the granule g. This gives a necessity condition for verification of implications of rough mereological logic: if $\Rightarrow_L ([\alpha^*]_g, [\beta^*]_g) < 1$ then the implication $\alpha \Rightarrow \beta$ is not true at g. This concerns decision rules in particular: for a decision rule $p(v) \Rightarrow q(v)$, it follows that the decision is true on a granule g if and only if $[p^*]_g \leq [q^*]_g$.

Possibility and Necessity Possibility and necessity are introduced in rough set theory by means of upper and the lower approximations, respectively. A logical rendering of these modalities in rough mereological logic exploits the approximations. Denoting the lower approximation as $\underline{X} = \{u \in U : [u]_A \subseteq X\}$ and the upper approximation as $\overline{X} = \{u \in U : [u]_A \cap X \neq \emptyset\}$, we define two modal operators: M (possibility) and L (necessity) by means of their semantics.

To this end, we let

$$\begin{aligned} I_\nu^\vee(g)(M\alpha) \geq r \ \text{ iff } \ \nu_L(g, \overline{[\alpha]}, r) \\ I_\nu^\vee(g)(L\alpha) \geq r \ \text{ iff } \ \nu_L(g, \underline{[\alpha]}, r) . \end{aligned} \tag{22}$$

Then we have the following criteria for necessarily or possibly true formulas.

A formula α is *necessarily true at a granule* g if and only if $g \subseteq \underline{[\alpha]}$; α is *possibly true at* g if and only if $g \subseteq \overline{[\alpha]}$.

This semantics of modal operators M, L can be applied to show that rough set structures carry the semantics of S5 modal logic, that is, the following relations hold at each granule g.

Granulation of Knowledge: Similarity Based Approach in Information and Decision Systems, Table 3
Truth values for implication in L_3

$\Rightarrow$	0	1	$\frac{1}{2}$
0	1	1	1
1	0	1	$\frac{1}{2}$
$\frac{1}{2}$	$\frac{1}{2}$	1	1

Granulation of Knowledge: Similarity Based Approach in Information and Decision Systems, Table 4
Truth values for implication $\alpha \Rightarrow \beta$ in logic based on ν_3

$\Rightarrow$	$I_\nu^\vee(g)(\beta) = 0$	$I_\nu^\vee(g)(\beta) = 1$	$I_\nu^\vee(g)(\beta) = \frac{1}{2}$
$I_\nu^\vee(g)(\alpha) = 0$	1	1	1
$I_\nu^\vee(g)(\alpha) = 1$	0	1	$\frac{1}{2}$
$I_\nu^\vee(g)(\alpha) = \frac{1}{2}$	$\frac{1}{2}$	1	1 when $g \cap [\alpha] \subseteq [\beta]$; $\frac{1}{2}$ otherwise

1. $L(\alpha \Rightarrow \beta) \Rightarrow [(L\alpha) \Rightarrow L(\beta)]$.
2. $L\alpha \Rightarrow \alpha$.
3. $L\alpha \Rightarrow LL\alpha$.
4. $M\alpha \Rightarrow LM\alpha$.

Proofs can be found in Polkowski [38].

Reasoning with ν_3 In case of $\nu = \nu_3$, one can check on the basis of definitions that $I_\nu^\vee(g)(\neg\alpha) \geq r$ if and only if $I_\nu^\vee(g)(\alpha) \leq 1-r$; thus the negation functor in rough mereological logic based on ν_3 is the same as the negation functor in the 3-valued Łukasiewicz logic. For implication, the relations between rough mereological and 3-valued logics L_3 follow from tables of values of truth.

Table 3 shows truth values of implications for the 3-valued logic L_3 and Table 4 shows truth values of implications for rough mereological logic based on ν_3.

To express the relation between the two implications, we introduce a new notion: we say that a formula ϕ is *acceptable* in either logic (L_3 or grm_{ν_3}) at a granule g if and only if $[\phi^*]_g \geq \frac{1}{2}$, respectively, $I_\nu^\vee(g)(\phi) \geq \frac{1}{2}$.

From Tables 3, 4 one infers that $I_\nu^\vee(g)(\phi) \geq [\phi^*]_g$. This crucial relationship implies that: if ϕ^* is acceptable at g then ϕ is acceptable at g; if ϕ^* is true at g then ϕ is true at g. Also, if ϕ at false at g then ϕ^* is false at g.

This fact allows for a sufficiency criterion: if ϕ^* is a theorem of the logic L_3 than ϕ is a true formula of the logic grm_{ν_3}. This fact supplies a characterization of decision rules as well.

Reasoning in Distributed Systems

We now begin with the first of the principal applications of the apparatus developed so far (cf., [49]). Rough inclu-

sions and granular intensional logics based on them can be applied to describe workings of a collection of intelligent agents, which we will call granular agents, as they are endowed with granular logic.

A granular agent ag in its simplest form is a tuple

$$ag^* = (U_{ag}, A_{ag}, \mu_{ag}, \mathrm{Pred}_{ag}, \mathrm{UncProp}_{ag}, \mathrm{GSynt}_{ag}, \mathrm{LSynt}_{ag}) ,$$

where,

1. $(U_{ag}, A_{ag}) = I_{ag}$ is an information system of the agent ag;
2. μ_{ag} is a rough inclusion induced from I_{ag};
3. Pred_{ag} is a set of first-order predicates interpreted in U_{ag};
4. $\mathrm{UncProp}_{ag}$ is the function that describes how uncertainty measured by rough inclusions at agents connected to ag propagates to ag;
5. The operator GSynt_{ag}, the granular synthesizer at ag, takes granules sent to the agent from agents connected to it, and makes those granules into a granule at ag;
6. LSynt_{ag}, the logic synthesizer at ag, takes formulas sent to the agent ag by its connecting neighbors and makes them into a formula describing objects at ag.

A network of granular agents is a directed acyclic graph $N = (Ag, C)$, where Ag is its set of vertices, that is, granular agents, and C is the set of edges, that is, connections among agents, along with disjoint subsets $In, Out \subset Ag$ of, respectively, input and output agents.

We assume for simplicity that N consists of three agents connected into a tree, and we show a simple analysis of the direct fusion of knowledge; clearly, more complex schemes will require a deeper and more complex analysis, but will follow the lines indicated in the example that follows.

Fusion of Knowledge: An Example We consider an agent $ag \in Ag$ and – for simplicity reasons – we assume that ag has two incoming connections from agents ag_1, ag_2; the number of outgoing connections is of no importance as ag sends the same information along each of them.

We assume that each agent i applies the rough inclusion μ_L induced by the Łukasiewicz t-norm L (see Subsect. "Granulation: Tools" A). Each agent also applies rough inclusion on sets of the form (17) in evaluations related to extensions of formulae intensions.

Clearly, there exists a fusion operator o_{ag} that assembles from objects $x \in U_{ag_1}, y \in U_{ag_2}$ the object $o(x, y) \in U_{ag}$; we assume that $o_{ag} = id_{ag_1} \times id_{ag_2}$, that is, $o_{ag}(x, y) = (x, y)$.

Similarly, we assume that the set of attributes at ag, equals: $A_{ag} = A_{ag_1} \times A_{ag_2}$, that is, attributes in A_{ag} are pairs (a_1, a_2) with $a_i \in A_{ag_i}$ $(i = 1, 2)$ and that the value of this attribute is defined as:

$$(a_1, a_2)(x, y) = (a_1(x), a_2(y)) .$$

It follows that the condition holds:

$$o_{ag}(x, y)\, \mathrm{IND}_{ag}\, o_{ag}(x', y') \quad \text{iff } x\, \mathrm{IND}_{ag_1}\, x' \text{ and } y\, \mathrm{IND}_{ag_2}\, y' .$$

Concerning the function $\mathrm{UncProp}_{ag}$, we consider objects x, x', y, y'; clearly,

$$\begin{aligned} &\mathrm{DIS}_{ag}(o_{ag}(x, y), o_{ag}(x', y')) \\ &\quad \subseteq \mathrm{DIS}_{ag_1}(x, x') \times A_{ag_2} \cup \\ &\quad\quad A_{ag_1} \times \mathrm{DIS}_{ag_2}(y, y') , \end{aligned} \tag{23}$$

and hence,

$$\begin{aligned} &|\,\mathrm{DIS}_{ag}(o_{ag}(x, y), o_{ag}(x', y'))| \\ &\quad \leq |\,\mathrm{DIS}_{ag_1}(x, x')| \cdot |A_{ag_2}| \\ &\quad + |A_{ag_1}| \cdot |\,\mathrm{DIS}_{ag_2}(y, y')| . \end{aligned} \tag{24}$$

By (24),

$$\begin{aligned} &\mu_{ag}(o_{ag}(x, y), o_{ag}(x', y'), t) \\ &= 1 - \frac{|\,\mathrm{DIS}_{ag}(o_{ag}(x, y), o_{ag}(x', y'))|}{|A_{ag_1}| \cdot |A_{ag_2}|} \\ &\geq 1 - \frac{|\,\mathrm{DIS}_{ag_1}(x, x')| \cdot |A_{ag_2}| + |A_{ag_1}| \cdot |\,\mathrm{DIS}_{ag_2}(y, y')|}{|A_{ag_1}| \cdot |A_{ag_2}|} \\ &= 1 - \frac{|\,\mathrm{DIS}_{ag_1}(x, x')|}{|A_{ag_1}|} + 1 - \frac{|\,\mathrm{DIS}_{ag_2}(y, y')|}{|A_{ag_2}|} - 1 . \end{aligned} \tag{25}$$

It follows that,

$$\begin{aligned} &\text{if } \mu_{ag_1}(x, x', r), \mu_{ag_2}(y, y', s) \\ &\text{then } \mu_{ag}(o_{ag}(x, y), o_{ag}(x', y'), L(r, s)). \end{aligned} \tag{26}$$

Hence, $\mathrm{UncProp}(r, s) = L(r, s)$, the value of the Łukasiewicz t-norm L on the pair (r, s).

In consequence, the granule synthesizer GSynt_{ag} can be defined in our example as,

$$\mathrm{GSynt}_{ag}(g_{ag_1}(x, r), g_{ag_2}(y, s)) = (g_{ag}(o_{ag}(x, y), L(r, s)). \tag{27}$$

The definition of the logic synthesizer LSynt_{ag} follows directly from our assumptions,

$$\mathrm{LSynt}_{ag}(\phi_1, \phi_2) = \phi_1 \wedge \phi_2 . \tag{28}$$

Finally, we consider extensions of the operators of our intensional logic.

We have for the extension $I(\mu_{ag})^{\vee}_{\mathrm{GSynt}_{ag}(g_1,g_2)}(\mathrm{LSynt}_{ag}(\phi_1,\phi_2))$:

$$I(\mu_{ag})^{\vee}_{\mathrm{GSynt}_{ag}(g_1,g_2)}(\mathrm{LSynt}_{ag}(\phi_1,\phi_2)) = I(\mu_{ag_1})^{\vee}_{g_1}(\phi_1)\cdot I(\mu_{ag_2})^{\vee}_{g_2}(\phi_2)\,, \quad (29)$$

which follows directly from (27), (28).

Let us note that

$$I(\mu_{ag_1})^{\vee}_{g_1}(\phi_1)\cdot I(\mu_{ag_2})^{\vee}_{g_2}(\phi_2) = P(I(\mu_{ag_1})^{\vee}_{g_1}(\phi_1), I(\mu_{ag_2})^{\vee}_{g_2}(\phi_2))\,,$$

where P is the Product Archimedean t-norm.

Thus, in the case of parallel fusion, where each agent works according to the Łukasiewicz t-norm, uncertainty propagation and granule synthesis are described by the Łukasiewicz t-norm L and extensions of logical intensions propagate according to the Product t-norm P.

Reasoning in Cognitive Schemes: Rough Neural Reasoning

In neural models of computation (see Bishop [5]), an essential feature of neurons is differentiability of transfer functions; hence, we introduce a special type of rough inclusions, called *gaussian* in Polkowski [40] because of their form, by letting,

$$\mu_G(x,y,r) \text{ iff } \mathrm{e}^{-\left|\sum_{a\in \mathrm{DIS}(x,y)} w_a\right|^2} \geq r\,, \quad (30)$$

where $w_a \in (0,+\infty)$ is a weight associated with the attribute a for each attribute $a \in A$; clearly, we retain the notation of previous sections. One may notice that the gaussian rough inclusion is a modification of rough inclusion μ_P obtained from the Product t-norm.

Let us observe in passing that μ_G can be factored through the indiscernibility relation $\mathrm{IND}(A)$, and thus its arguments can be objects as well as indiscernibility classes; we will freely use this fact.

Properties of gaussian rough inclusions are the following (cf., [41]):

- $x \operatorname{ing} y$ iff $\mathrm{DIS}(x,y) = \emptyset$;
- There exists a function $\eta(r,s)$ such that $\mu_G(x,y,r), \mu_G(y,z,s)$ imply $\mu_G(x,z,\eta(r,s))$;
- If $x \operatorname{ing} g^{\mu_G}_r y, x \operatorname{ing} g^{\mu_G}_s z$, then $g^{\mu_G}_t x \operatorname{ing} g^{\mu_G}_r y$, $g^{\mu_G}_t x \operatorname{ing} g^{\mu_G}_s z$ for $t \geq \max\{r^4, s^4\}$.

Property 1 follows by definition, and Property 2 may be verified with $\eta(r,s) = r\cdot s\cdot \mathrm{e}^{2\cdot(\log r\cdot\log s)^{1/2}}$ (op.cit.). Property 3 can be verified by observing that t should satisfy conditions $\eta(r,t)\geq r, \eta(s,t)\geq s$ because of Property 2, see Polkowski [41].

Rough Mereological Perceptron The rough mereological perceptron is modeled on the perceptron (see, for example, [5]). It consists of an intelligent agent ag, endowed with a gaussian rough inclusion μ_{ag} on the information system $I_{ag} = (U_{ag}, A_{ag})$ of the agent ag.

The input to ag is in the form of a finite tuple $\overline{x} = (x_1,\ldots,x_k)$ of objects, and the input $\overline{x}$ is converted at ag into an object $x = O_{ag}(\overline{x}) \in U_{ag}$ by means of an operator O_{ag}.

The rough mereological perceptron is endowed with a set of *target concepts* $T_{ag} \subseteq U_{ag}/\mathrm{IND}(A_{ag})$, each target concept a class of the indiscernibility IND_{ag}.

Formally, a rough mereological perceptron is thus a tuple

$$\mathrm{RMP} = (ag, I_{ag}, \mu_{ag}, O_{ag}, T_{ag})\,.$$

The output $res_{ag}(\overline{x})$ to RMP, at the input $\overline{x}$, is a granule of knowledge $g_{r(res)}x$ with

$$r(res) = \max\{r : \text{there is } y \in T_{ag}\mu_{ag}(x,y,r)\}\,. \quad (31)$$

Formula (31) shows that the input $\overline{x}$ is classified at ag as the collection of indiscernibility classes that are as close to x as the closest target class (closeness with respect to μ_{ag}).

Applications in Data Mining: Granular Classifiers

Granulation of knowledge by means of μ_h consists of forming, for each $r \in [0,1]$ and each $u \in U$, a granule $g_h(u,r) = \{v \in U\colon \mu_h(v,u,r)\}$. As, clearly, μ_h is symmetric, from $v \in g_h(u,r)$ it follows that $u \in g_h(v,r)$.

Granular data sets were proposed in Polkowski [40, 42,43] by means of the following constructions: Given $r \in [0,1]$, the set of all granules $G^h_r = \{g_h(u,r)\colon u \in U\}$ is defined. From this set, a covering $\mathrm{Cov}^h_r(\mathcal{G})$ is chosen according to a strategy $\mathcal{G}$. Granules in $\mathrm{Cov}^h_r(\mathcal{G})$ form a new universe of objects. For each $g \in \mathrm{Cov}^h_r(\mathcal{G})$, and each attribute $a \in A$, a factored attribute a_h is defined as $a_h(g) = S(\{a(u)\colon u \in g\})$.

The new information system $I^h_r = (\mathrm{Cov}^h_r(\mathcal{G}), \{a_h\colon a \in A\})$ is a *granular reflection* of the original information system I. The same procedure is applied to a decision system $D = (U, A, d)$ to form the reflection $D^h_r = (\mathrm{Cov}^h_r(\mathcal{G}), \{a_h\colon a \in A\}, d_h)$. The object $o(g)$ defined for

Granulation of Knowledge: Similarity Based Approach in Information and Decision Systems, Table 5
Best results for Australian credit by some rough set based algorithms

Source	Method	Accuracy	Coverage
Bazan [3]	SNAPM(0.9)	error = 0.130	–
Nguyen SH [22]	simple.templates	0.929	0.623
Nguyen SH [22]	general.templates	0.886	0.905
Nguyen SH [22]	tolerance.gen.templ.	0.875	1.0
Wroblewski [57]	adaptive.classifier	0.863	–

Granulation of Knowledge: Similarity Based Approach in Information and Decision Systems, Table 6
Best results for Australian credit by granular approach

Source	Method	Accuracy	Coverage
[45]	granular radius 0.642857	0.867	1.0
[45]	granular radius 0.714826	0.875	1.0
[45]	granular.concept dependent radius 0.785	0.9970	0.9995

a granule g by means of $\inf(o(g)) = \{(a_h, a_h(g)) \colon a \in A\}$ according to a strategy S is called an *S-reflection of the granule* g; clearly, $o(g)$ need not be a real object in the training or test sets.

The idea of data classification by means of granular reflections consists in splitting a given data set into training and test parts, forming a granular reflection of the training set, inducing classification rules from this new data set, and applying the induced rules in classifying data in the original test set.

We include some examples showing the high efficiency of this approach (see [1,43,45]).

In tests with μ_h, the exhaustive classifier (for its public version, see [51]) was applied (cf., [1,43]). The strategy $\mathcal{G}$ was a random choice of an irreducible covering from the set of all granules of a given radius and the strategy S was selected as majority voting with random tie resolution.

We show results of tests with the Australian credit data set (see UCI Repository [59]), well studied in rough set literature. For comparison, we include some best results, obtained by means of rough set – based methods, in Table 5. Classification quality is expressed by means of two factors: accuracy and coverage. Accuracy (also called *total accuracy*) is the ratio of the number of correctly classified objects to the number of recognized test objects, and *total coverage*, $\frac{\text{rec}}{\text{test}}$, where rec is the number of recognized test cases and test is the number of test cases.

Tests on these data with the granular approach described above were carried out by splitting the Australian credit data set into training and test sets in a 1:1 ratio. The training sample was granulated and a granular reflection was formed from which, by means of the RSES exhaustive algorithm, a classifier was produced which was applied to the test part of the data to find the quality of classification.

Granules were calculated in a twofold way: first as indicated above and, second, by a modified procedure of *concept dependent granulation* (see [1]). In the latter procedure, the granule $g_h^c(u, r) = g_h(u, r) \cap [u]_d$ was computed relative to the *concept*, that is, the decision class, to which u belonged. The results of the tests are given in Table 6 in which the best results obtained with various granulation radii are shown, taken from [45].

Results in Table 6 attest that the granular approach gives results fully comparable with other results for satisfactorily large radii of granulation, and that concept-dependent granulation gives better results than any other existing approach.

In order to test the impact of the choice of granular covering on classification, we have carried out 10 experiments with the RSES exhaustive algorithm, with random coverings on the Heart dataset (Cleveland) (see UCI Repository [59]). Results are given in Table 7. Total accuracy was found to be 0.807, and total coverage 1.0 with an exhaustive algorithm on the full data. These values are achieved here with radius of at least 0.538462, and beginning with the radius of 0.384615, the error in total accuracy is at most 0.07, and the error in total coverage is at most 0.007.

This attests to a very high stability of the granular approach showing the essential independence of results of a choice of a granular covering for inducing a granular reflection of data.

Granules from Parameterized Variants of Rough Inclusions μ_h in Classification of Data Given an $\varepsilon \in [0, 1]$, we let $\mu_h^\varepsilon(v, u, r)$ if and only if $|\{a \in A \colon \rho(a(v), a(u)) < \varepsilon\}| \geq r \cdot |A|$. It is manifest that μ^ε is a rough inclusion if ρ is a non-archimedean metric, that is, $\rho(u, w) \leq$

Granulation of Knowledge: Similarity Based Approach in Information and Decision Systems, Table 7
Effect of a choice of a granular covering on classification

Radius	Total accuracy	Total coverage
0.0	0.0	0.0
0.0769231	0.0	0.0
0.153846	0.0	0.0
0.230769	0.0–0.789	0.0–1.0
0.307692	0.01.0	0.0–1.0
0.384615	0.737–0.799	0.993–1.00
0.461538	0.778–0.822	0.996–1.0
0.538462	0.881–0.911	1.0
0.615385	0.874–0.907	1.0
0.692308	0.963–0.974	1.0
0.769231	1.0	1.0
0.846154	1.0	1.0
0.923077	1.0	1.0
1.0	1.0	1.0

$\max\{\rho(u,v), \rho(v,w)\}$; otherwise the monotonicity condition (RM2)of Sect. "Introduction" need not be satisfied and this takes place with most popular metrics (Euclidean, Manhattan, etc.).

In this case, a rough inclusion μ^* is defined as follows: $\mu^*_h(v,u,r)$ if and only if there exists an ε such that $\mu^\varepsilon_h(v,u,r)$. It is easy, then, to check that μ^* is a rough inclusion. The parameter r is called the catch radius.

Granules induced by the rough inclusion μ^*_h with $r = 1$ have a simple structure: a granule $g^\varepsilon_h(u,1)$ consists of all $v \in U$ such that $\rho(a(u), a(v)) \leq \varepsilon$.

The idea poses itself to use granules defined in this way to assign a decision class to an object u in the test set. First on the training set, rules are induced by an exhaustive algorithm. Then, given a set *Rul* of these rules, and an object u in the test set, a granule $g^\varepsilon_h(u,1)$ is formed in the set *Rul*. In this manner, the duality between objects and rules is exploited, as rules and objects can be written down in the same format, that of information sets. This also allows the use of training objects instead of rules in forming granules and permits majority voting for a decision.

Thus, $g^\varepsilon_h(u,1) = \{r \in Rul : \rho(a(u), a(r)) \leq \varepsilon$ for each attribute $a \in A(r)\}$ where $a(r)$ is the value of the attribute a in the premise of the rule and $A(r)$ is the set of attributes in the premise of r.

Rules in the granule $g^\varepsilon_h(u,1)$ take part in a voting process: for each value c of a decision class, the following factor is computed,

$$\text{param}(c) = \frac{\text{sum of supports of rules pointing to } c}{\text{cardinality of } c \text{ in the training set}}, \quad (32)$$

cf., Bazan [3], Michalski et al. [25] for a discussion of various strategies of voting for decision values.

The class c_u assigned to u is decided by

$$\text{param}(c_u) = \max_c \text{param}(c)\,, \quad (33)$$

with random resolution of ties.

In computing granules, the parameter ε is normalized to the interval $[0,1]$ as follows: first, for each attribute $a \in A$, the value $\text{train}(a) = \max_{\text{training set}} a - \min_{\text{training set}} a$ is computed and the real line $(-\infty, +\infty)$ is contracted to the interval $[\min_{\text{training set}} a, \max_{\text{training set}} a]$ by the mapping f_a,

$$f_a(x) = \begin{cases} \min_{\text{training set}} a \text{ in case } x \leq \min_{\text{training set}} \\ x \text{ in case } x \in [\min_{\text{training set}} a, \max_{\text{training set}} a] \\ \max_{\text{training set}} a \text{ in case } x \geq \max_{\text{training set}} a\,. \end{cases} \quad (34)$$

When the value $a(u)$ for a test object u is off the range $[\min_{\text{training set}} a, \max_{\text{training set}} a]$, it is replaced with the value $f_a(a(u))$ in the range. For an object v, or a rule r with the value $a(v)$, respectively, $a(r)$ of a denoted $a(v,r)$, the parameter ε is computed as $\frac{|a(v,r) - f_a(a(u))|}{\text{train}(a)}$. The metric ρ was chosen as the metric $|x-y|$ in the real line. We show results of experiments with rough inclusions discussed in this work. Our data set was a subset of Australian credit data in which the training set had 100 objects from class 1 and 150 objects from class 0 (which approximately yields the distribution of classes in the whole data set). The test set had 100 objects, 50 from each class. The RSES exhaustive classifier (see Skowron et al. [51]) applied to this data set gave accuracy of 0.79 and coverage of 1.0.

In Fig. 1 results of classification are given in function of ε for accuracy as well as for coverage.

We return to the rough inclusion $\mu^*_h(v,u,r)$ with general radius r. The procedure applied in case of $\mu^\varepsilon_h(v,u,1)$ can be repeated in the general setting. The resulting classifier is a function of two parameters ε, r. In Table 8 results are included where against values of the catch radius r the best value for ε's marked by the optimal value *optimal eps* is given for accuracy and coverage.

As shown in Sect. "Granulation of Knowledge: General Ideas and Technique of Rough Inclusions", residual implications of continuous t-norms can supply rough inclusions according to a general formula, see (9),

$$\mu_\phi(v,u,r) \text{ iff } \phi(u) \Rightarrow_t \phi(v) \geq r\,, \quad (35)$$

where ϕ maps the set U of objects into $[0,1]$ and $\phi(u) \leq \phi(v)$ if and only if u ing v (ing is an ingredient relation of the underlying mereology; $\Rightarrow_t$ is the residual implication induced by the t-norm.

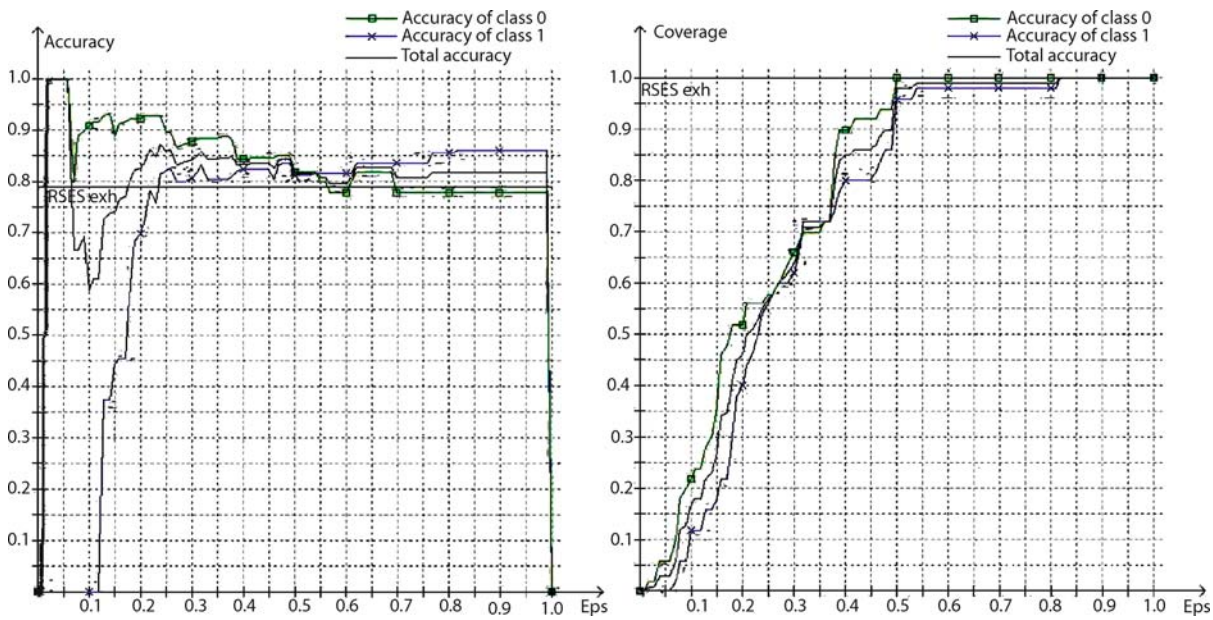

Granulation of Knowledge: Similarity Based Approach in Information and Decision Systems, Figure 1
Results for algorithm 1_v1, best result for ε = 0.62: accuracy = 0.828283, coverage = 0.99

Granulation of Knowledge: Similarity Based Approach in Information and Decision Systems, Table 8
(40%–60%)(1–0); Australian credit; Algorithm 1_v2. r_catch= catch radius, optimal_eps=Best ε, acc= accuracy, cov= coverage

r_catch	optimal eps	acc	cov
nil	nil	0.79	1.0
0.071428	0	0.06	1.0
0.142857	0	0.66	1.0
0.214286	0.01	0.74	1.0
0.285714	0.02	0.83	1.0
0.357143	0.07	0.82	1.0
0.428571	0.05	0.82	1.0
0.500000	0	0.82	1.0
0.571429	0.08	0.84	1.0
0.642857	0.09	0.84	1.0
0.714286	0.16	0.85	1.0
0.785714	0.22	0.86	1.0
0.857143	0.39	0.84	1.0
0.928571	0.41	0.828283	0.99
1.000000	0.62	0.828283	0.99

A weak interesting variant of this class of rough inclusions is indicated. This variant uses sets

$$\mathrm{dis}_\varepsilon(u,v) = \frac{\left|\{a \in A\colon \rho(a(u), a(v)) \geq \varepsilon\}\right|}{|A|},$$

and

$$\mathrm{ind}_\varepsilon(u,v) = \frac{\left|\{a \in A\colon \rho(a(u), a(v)) < \varepsilon\}\right|}{|A|},$$

for $u, v \in U$, $\varepsilon \in [0,1]$, where ρ is a metric $|x-y|$ on attribute value sets.

The resulting weak variant of the rough inclusion μ_ϕ is,

$$\mu_t(u,v,r) \text{ iff } \mathrm{dis}_\varepsilon(u,v) \rightarrow_t \mathrm{ind}_\varepsilon(u,v) \geq r\,. \tag{36}$$

Basic variants for three principal t-norms: the Łukasiewicz t-norm $L = \max\{0, x+y-1\}$, the product t-norm $P(x,y) = x \cdot y$, and $\min\{x,y\}$ are: (the value in all variants is 1 if and only if $x \leq y$ so we give values only in the contrary case)

$$\mu_t(u,v,r) \text{ iff } \begin{cases} 1 - \mathrm{dis}_\varepsilon(u,v) + \mathrm{ind}_\varepsilon(u,v) \geq r & \text{for L} \\ \frac{\mathrm{ind}_\varepsilon(u,v)}{\mathrm{dis}_\varepsilon(u,v)} \geq r & \text{for P} \\ \mathrm{ind}_\varepsilon(u,v) \geq r & \text{for min}\,. \end{cases} \tag{37}$$

Objects in the class c in the training set vote for decision at the test object u according to the formula:

$$p(c) = \frac{\sum_{v \in c} w(v,t)}{|c| \text{ in the training set}}$$

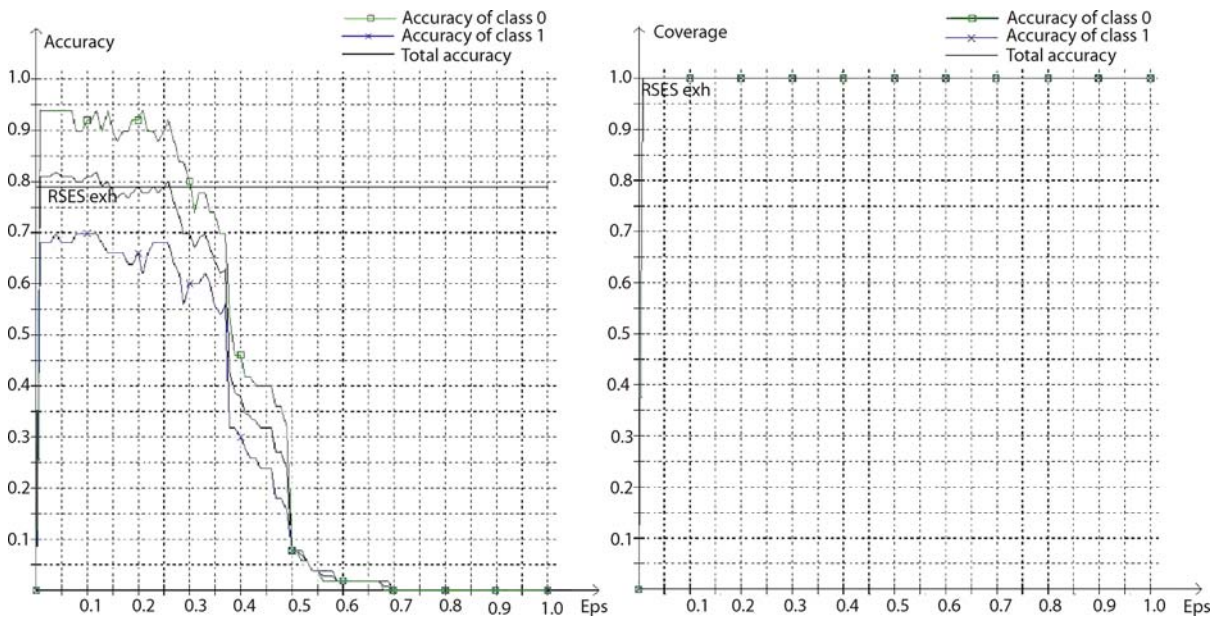

Granulation of Knowledge: Similarity Based Approach in Information and Decision Systems, Figure 2
Results for algorithm 5_v1, best result for $\varepsilon = 0.04$, accuracy = 0.82, coverage = 1

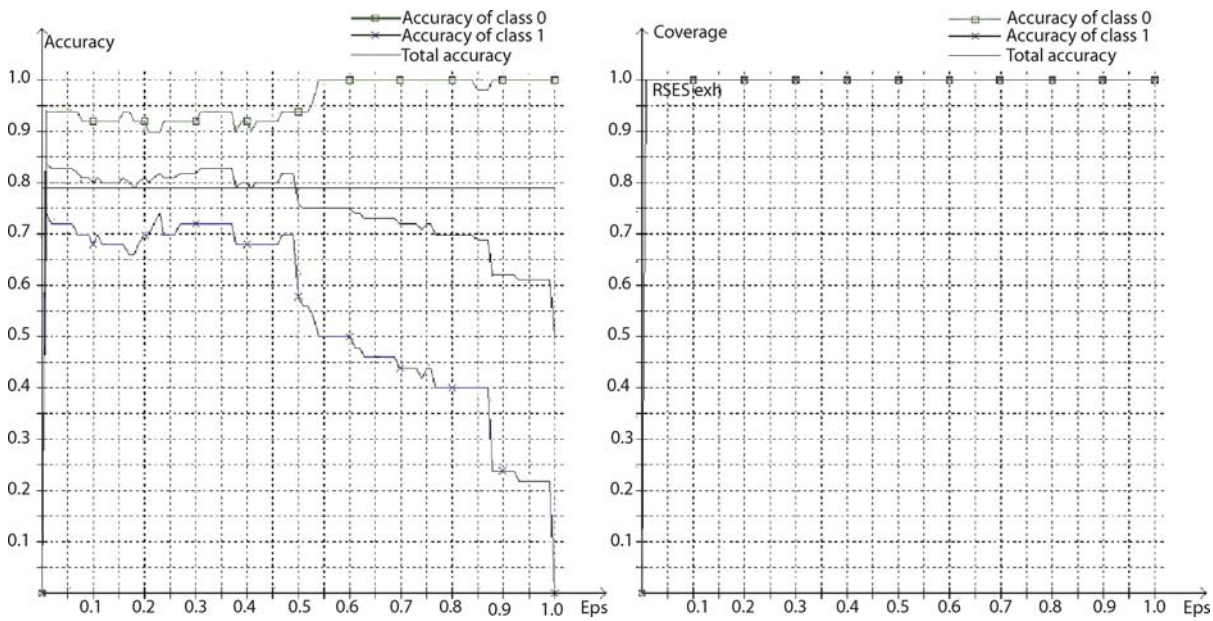

Granulation of Knowledge: Similarity Based Approach in Information and Decision Systems, Figure 3
Results for algorithm 5_v2, best result for $\varepsilon = 0.01$, accuracy = 0.84, coverage = 1

where weight $w(v, t)$ is $\mathrm{dis}_\varepsilon(u, v) \rightarrow_t \mathrm{ind}_\varepsilon(u, v)$; rules induced from the training set pointing to the class c vote according to the formula:

$$p(c) = \frac{\sum_r w(r, t) \cdot \mathrm{support}(r)}{|c| \text{ in the training set}} .$$

In either case, the class c^* with $p(c*) = \max p(c)$ is chosen. We include here results of tests with training objects and $t = \min$ (Fig. 2) and rules and $t = \min$ (Fig. 3).

Similarly, we include in Figs. 4, 5 results of tests with granules of training objects and rules for $t = P$, the product t-norm.

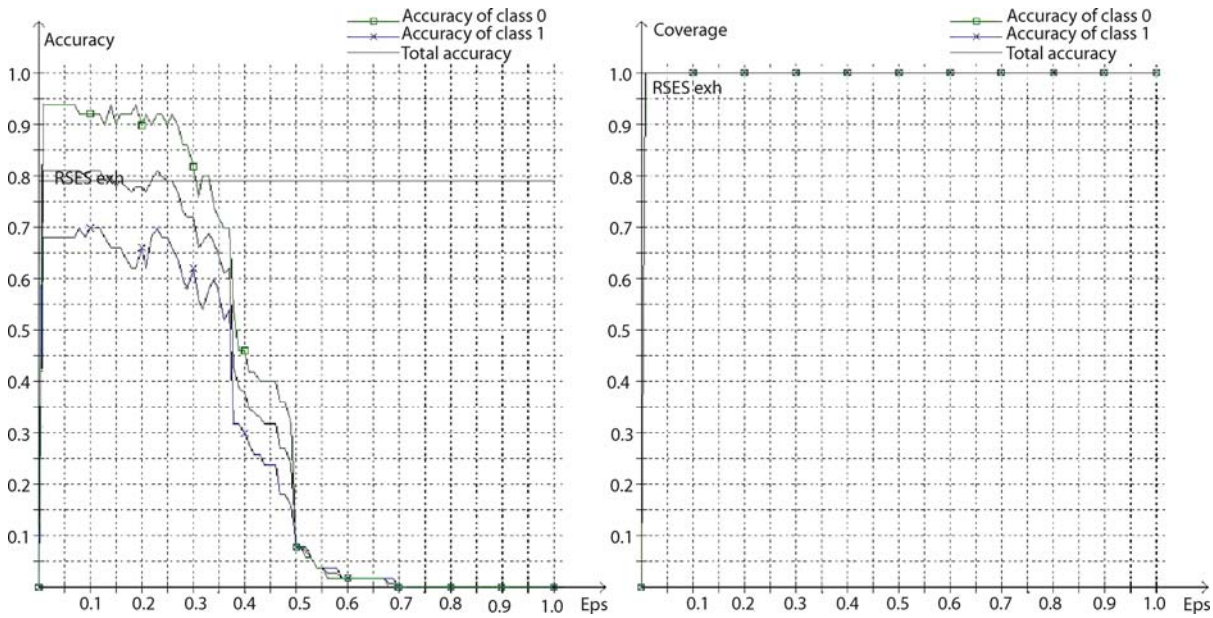

Granulation of Knowledge: Similarity Based Approach in Information and Decision Systems, Figure 4
Results for algorithm 6_v1, best result for $\varepsilon = 0.01$, accuracy = 0.81, coverage = 1

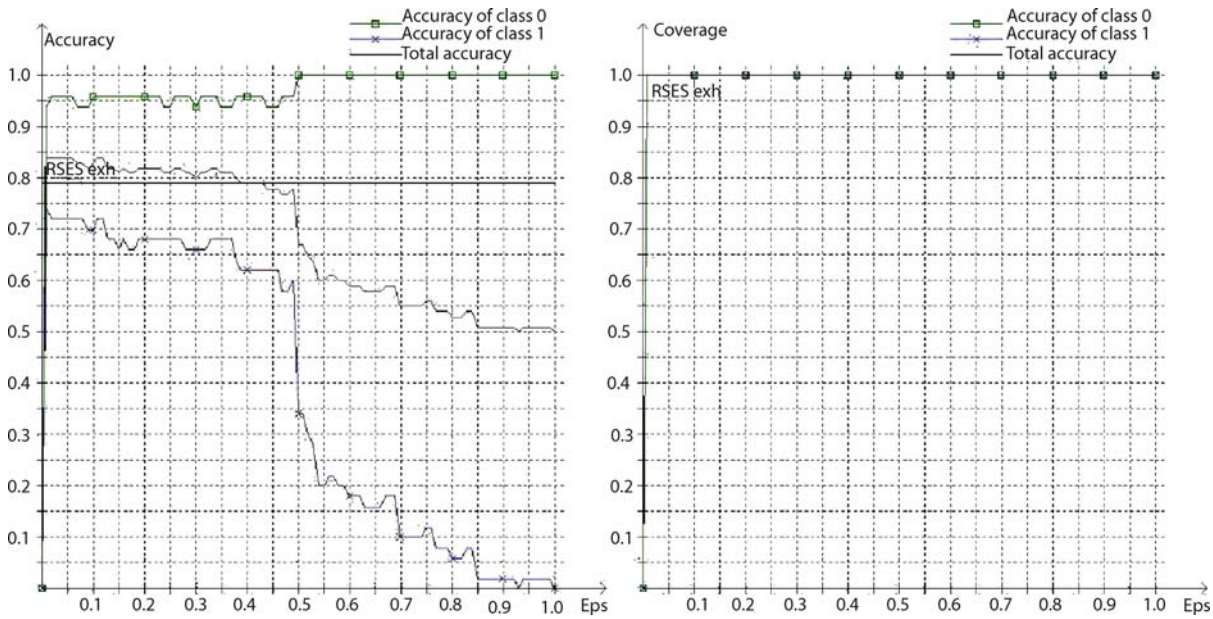

Granulation of Knowledge: Similarity Based Approach in Information and Decision Systems, Figure 5
Results for algorithm 6_v2, Best result for *varepsilon* = 0.01, accuracy = 0.84, coverage = 1

The results of tests in best cases for optimal values of ε exceed results obtained with the standard exhaustive algorithm (the RSES exhaustive algorithm in this data set gives accuracy of 0.79 and coverage 1.0).

Complexity Issues

Concerning granulation in information systems, it can be proved that finding a granule about an object u is at most

$O(|U| \cdot |A|)$. Therefore, finding all granules of the given radius is $O(|U|^2 \cdot |A|)$.

Selecting a coverage by granules of the given radius is therefore $O(|U|^2 \cdot |A|)$.

Future Directions

As results of the last section attest, granulation is a very important ingredient in such diverse areas as reasoning about uncertainty, reasoning in distributed and cognitive schemes, and data mining. Especially in this last area, it is possible to objectively judge the contribution of granulation by comparing the quality of classifiers using granulation with those which dispense with granulation. These results show clearly that granulation produces results better, as a rule, than the standard approach of rule induction by exhaustive algorithm, and the best results are as good as any obtained by methods based on rule building by descriptors with optimization.

Further research should be aimed at better understanding the role of granulation, of the trade-off between the size of granules and quality of reasoning/classification, of the dynamics of granule formation and at finding better similarity measures adapted to particular data sets.

Bibliography

1. Artiemjew P (2007) Classifiers from granulated data sets: Concept dependent and layered granulation. In: Proceedings RSKD'07, Workshop at ECML/PKDD'07. Warsaw Univ Press, Warsaw, pp 1–9
2. Baader F, Calvanese D, McGuinness D, Nardi D, Patel-Schneider P (eds) (2004) The description logic handbook: theory, implementation and applications. Cambridge U Press, Cambridge
3. Bazan JG (1998) A comparison of dynamic and non-dynamic rough set methods for extracting laws from decision tables. In: Polkowski L, Skowron A (eds) Rough sets in knowledge discovery 1. Physica, Heidelberg, pp 321–365
4. Bazan JG, Nguyen HS, Nguyen SH, Synak P, Wróblewski J (2000) Rough set algorithms in classification problems. In: Polkowski L, Tsumoto S, Lin TY (eds) Rough set methods and applications. New developments in knowledge discovery in information systems. Physica, Heidelberg, pp 49–88
5. Bishop CM (1997) Neural networks for pattern recognition, 2nd edn. Clarendon Press, Oxford
6. Bocheński JM (1954) Die zeitgenössischen Denkmethoden. A Francke AG, Bern
7. Duda RO, Hart PE, Stork DG (2001) Pattern classification. Wiley, New York
8. Greco S, Matarazzo B, Słowiński R (1999) On joint use of indiscernibility, similarity and dominance in rough approximation of decision classes. In: Proceedings of the 5th International Conference of the Decision Sciences Institute, Athens, Greece. pp 1380–82
9. Grzymala-Busse JW (1992) LERS – a system for learning from examples based on rough sets. In: Słowiński R (ed) Intelligent decision support: Handbook of advances and applications of the rough sets theory. Kluwer, Dordrecht, pp 3–18
10. Grzymala-Busse JW (2004) Data with missing attribute values: Generalization of indiscernibility relation and rule induction. In: Transactions on rough sets, vol I. Lecture Notes in Computer Science, vol 3100. Springer, Berlin, pp 78–95
11. Grzymala-Busse JW, Hu M (2000) A comparison of several approaches to missing attribute values in data mining. In: Rough sets and current trends in computing. Lecture Notes in CS, vol 2005. Springer, Berlin, pp 378–385
12. Klösgen W, Żytkow J (eds) (2002) Handbook of data mining and knowledge discovery. Oxford University Press, Oxford
13. Krawiec et al (1998) Learning decision rules from similarity based rough approximations. In: Polkowski L, Skowron A (eds) Rough sets in knowledge discovery 2. Physica, Heidelberg, pp 37–54
14. Kuratowski K, Mostowski A (1966) Set theory. Polish Scientific Publishers, Warsaw
15. Leśniewski S (1916) Podstawy ogólnej teoryi mnogosci (On the foundations of set theory), in Polish. The Polish Scientific Circle, Moscow. see also a later digest: Topoi, vol 2, 1982, pp 7–52
16. Lin TY (1988) Neighborhood systems and relational database. Abstract. In: Proceedings of CSC'88, 1988. p 725
17. Lin TY (1989) Neighborhood systems and approximation in database and knowledge based systems. In: Proceedings of the 4th International Symposium on Methodologies of Intelligent Systems, Poster Session, 1989. pp 75–86
18. Lin TY (1992) Topological and fuzzy rough sets. In: Słowiński R (ed) Intelligent decision support-handbook of applications and advances of the rough sets theory. Kluwer, Dordrecht, pp 287–304
19. Lin TY (1997) From rough sets and neighborhood systems to information granulation and computing with words. In: Proceedings of the European Congress on Intelligent Techniques and Soft Computing. pp 1602–1606
20. Lin TY (1999) Granular computing: fuzzy logic and rough sets. In: Zadeh LA, Kacprzyk J (eds) Computing with words in information/intelligent systems 1. Physica, Heidelberg, pp 183–200
21. Lin TY (2003) Granular computing. In: Lecture Notes in Computer Science, vol 2639. Springer, Berlin, pp 16–24
22. Lin TY (2005) Granular computing: Examples, intuitions, and modeling. In: Proceedings of IEEE 2005 Conference on Granular Computing GrC05, Beijing, China, 2005. IEEE Press, New York, pp 40–44
23. Lin TY (2006) A roadmap from rough set theory to granular computing. In: Proceedings RSKT 2006, 1st International Conference on Rough Sets and Knowledge Technology, Chongqing, China, 2006. Lecture Notes in Artificial Intelligence, vol 4062. Springer, Berlin, pp 33–41
24. Ling C-H (1965) Representation of associative functions. Publ Math Debrecen 12:189–212
25. Michalski RS et al (1986) The multi-purpose incremental learning system AQ15 and its testing to three medical domains. In: Proceedings of AAAI-86. Morgan Kaufmann, San Mateo CA, pp 1041–1045
26. Nieminen J (1988) Rough tolerance equality and tolerance black boxes. Fundamenta Informaticae 11:289–296
27. Nguyen SH (2000) Regularity analysis and its applications in Data Mining. In: Polkowski L, Tsumoto S, Lin TY (eds) Rough set methods and applications. New developments in knowl-

edge discovery in information systems. Physica, Heidelberg, pp 289–378
28. Novotny M, Pawlak Z (1988) Partial dependency of attributes. Bulletin Polish Academy. Ser Sci Math 36:453–458
29. Novotny M, Pawlak Z (1992) On a problem concerning dependence spaces. Fundamenta Informaticae 16:275–287
30. Pawlak Z (1982) Rough sets. Intern J Comput Inf Sci 11:341–356
31. Pawlak Z (1985) On rough dependency of attributes in information systems. Bulletin Polish Academy. Ser Sci Tech 33: 551–559
32. Pawlak Z (1991) Rough sets: theoretical aspects of reasoning about data. Kluwer, Dordrecht (Holland)
33. Pawlak Z, Skowron A (1993) A rough set approach for decision rules generation. In: Proceedings of IJCAI'93 Workshop W12. The Management of Uncertainty in AI, France, 1993. also ICS Research Report 23/93, Warsaw University of Technology, Institute of Computer Science
34. Pawlak Z, Skowron A (1994) Rough membership functions. In: Yaeger RR, Fedrizzi M, Kasprzyk J (eds) Advances in the Dempster–Schafer Theory of evidence. Wiley, New York, pp 251–271
35. Poincaré H (1905) Science and hypothesis. Walter Scott Publishing, London
36. Polkowski L (2002) Rough sets. mathematical foundations. Physica, Heidelberg
37. Polkowski L (2003) A rough set paradigm for unifying rough set theory and fuzzy set theory. In: Proceedings RSFDGrC03, Chongqing, China, 2003. Lecture Notes in AI, vol 2639. Springer, Berlin, pp 70–78. also in: Fundamenta Informaticae 54:67–88
38. Polkowski L (2004) A note on 3-valued rough logic accepting decision rules. Fundamenta Informaticae 61:37–45
39. Polkowski L (2004) Toward rough set foundations. Mereological approach. In: Proceedings RSCTC04, Uppsala, Sweden. Lecture Notes in AI, vol 3066. Springer, Berlin, pp 8–25
40. Polkowski L (2005) Formal granular calculi based on rough inclusions. In: Proceedings of IEEE 2005 Conference on Granular Computing GrC05, Beijing, China. IEEE Press, New York, pp 57–62
41. Polkowski L (2005) Rough-fuzzy-neurocomputing based on rough mereological calculus of granules. Intern J Hybrid Intell Syst 2:91–108
42. Polkowski L (2006) A model of granular computing with applications. In: Proceedings of IEEE 2006 Conference on Granular Computing GrC06, Atlanta, USA. IEEE Press, New York, pp 9–16
43. Polkowski L (2007) Granulation of knowledge in decision systems: The approach based on rough inclusions. The method and its applications. In: Proceedings RSEiSP'07 (in memory of Z. Pawlak). Lecture Notes in AI, vol 4585. pp 69–79
44. Polkowski L (2007) The paradigm of granular rough computing. In: Proceedings ICCI'07. 6th IEEE Intern. Conf. on Cognitive Informatics. IEEE Computer Society, Los Alamitos, pp 145–163
45. Polkowski L, Artiemjew P (2007) On granular rough computing: Factoring classifiers through granular structures. In: Proceedings RSEISP'07, Warsaw. Lecture Notes in AI, vol 4585. Springer, Berlin, pp 280–289
46. Polkowski L, Semeniuk-Polkowska M (2005) On rough set logics based on similarity relations. Fundamenta Informaticae 64:379–390
47. Polkowski L, Skowron A (1994) Rough mereology. In: Proceedings of ISMIS'94. Lecture notes in AI, vol 869. Springer, Berlin, pp 85–94
48. Polkowski L, Skowron A (1997) Rough mereology: a new paradigm for approximate reasoning. Intern J Approx Reason 15(4):333–365
49. Polkowski L, Skowron A (1999) Towards an adaptive calculus of granules. In: Zadeh LA, Kacprzyk J (eds) Computing with words in information/intelligent systems,1. Physica, Heidelberg, pp 201–228
50. Polkowski L, Skowron A, Żytkow J (1994) Tolerance based rough sets. In: Lin TY, Wildberger M (eds) Soft computing: rough sets, fuzzy logic, neural networks, uncertainty management, knowledge discovery. Simulation Councils Inc, San Diego CA, pp 55–58
51. RSES: Skowron A et al (1994) A system for data analysis. http://logic.mimuw.edu.pl/~rses/. Accessed 19 Sept 2008
52. Russell SJ, Norvig P (2003) Artificial intelligence. A modern approach, 2nd edn. Prentice Hall (Pearson Education Inc), Upper Saddle River NJ
53. Skowron A (1993) Boolean reasoning for decision rules generation. In: Komorowski J, Ras Z (eds) Proceedings of ISMIS'93. Lecture Notes in AI, vol 689. Springer, Berlin, pp 295–305
54. Skowron A, Rauszer C (1992) The discernibility matrices and functions in decision systems. In: Słowiński R (ed) Intelligent decision support. handbook of applications and advances of the rough sets theory. Kluwer, Dordrecht, pp 311–362
55. Skowron A, Stepaniuk J (2001) Information granules: towards foundations of granular computing. Intern J Intell Syst 16: 57–85
56. Skowron A, Swiniarski RW (2004) Information granulation and pattern recognition. In: Pal SK, Polkowski L, Skowron A (eds) Rough – neural computing. Techniques for computing with words. Springer, Berlin, pp 599–636
57. Stefanowski J (1998) On rough set based approaches to induction of decision rules. In: Polkowski L, Skowron A (eds) Rough sets in knowledge discovery 1. Physica, Heidelberg, pp 500–529
58. Stefanowski J (2007) On combined classifiers, rule induction and rough sets. In: Transactions on rough sets, vol VI. Lecture Notes in Computer Science, vol 4374. Springer, Berlin, pp 329–350
59. UCI Repository: available at http://archive.ics.uci.edu/ml/. Accessed 19 Sept 2008
60. van Benthem J (1988) A manual of intensional logic. CSLI Stanford University, Stanford CA (US)
61. Wille R (1982) Restructuring lattice theory: An approach based on hierarchies of concepts. In: Rival I (ed) Ordered sets. Reidel, Dordrecht, pp 445–470
62. Wojna A (2005) Analogy-based reasoning in classifier construction. In: Transactions on Rough Sets vol IV. Lecture Notes in Computer Science, vol 3700. Springer, Berlin, pp 277–374
63. Wróblewski J (2004) Adaptive aspects of combining approximation spaces. In: Pal SK, Polkowski L, Skowron A (eds) Rough – neural computing. Techniques for computing with words. Springer, Berlin, pp 139–156
64. Yao YY (2000) Granular computing: Basic issues and possible solutions. In: Proceedings of the 5th Joint Conference on Information Sciences I. Assoc Intell Machinery, Atlantic NJ, pp 186–189
65. Yao YY (2005) Perspectives of granular computing. In: Proceedings of IEEE 2005 Conference on Granular Computing GrC05, Beijing, China. IEEE Press, New York, pp 85–90
66. Zadeh LA (1979) Fuzzy sets and information granularity. In:

Gupta M, Ragade R, Yaeger RR (eds) Advances in fuzzy set theory and applications. North-Holland, Amsterdam, pp 3–18
67. Zeeman EC (1965) The topology of the brain and the visual perception. In: Fort MK (ed) Topology of 3-manifolds and selected topics. Prentice Hall, Englewood Cliffs NJ, pp 240–256

Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures

P. Martin Mai
Swiss Seismological Service, Institute of Geophysics, ETH, Zürich, Switzerland

Article Outline

Glossary
Definition of the Subject
Introduction
Characterizing Earthquake Source Complexity
Wave Propagation in Complex Media: Path and Site Effects
Ground-Motion Scaling Relations
Future Directions
Acknowledgments
Bibliography

Glossary

Attenuation relation (ground-motion prediction equation) The term "attenuation relation" is a former shorthand notation in earthquake engineering for "empirical ground-motion attenuation relationship", now referred to as "ground-motion prediction equation" (GMPE). Attenuation relations represent empirical scaling equations that relate observed ***ground-motion intensity measures*** to parameters of the earthquake source, the wave propagation from the source to the observer and the site response at the observer location.

Dynamic rupture model Dynamic rupture models build a physical understanding of the earthquake rupture based on the material properties around the source volume, and the initial and boundary conditions for the forces/stresses acting on the fault plane. The distribution of on-fault ***slip-rate*** vectors and the temporal rupture evolution is obtained by solving the elastodynamic equations of motion under an assumed constitutive law (friction model), considering the energy balance at the crack tip (Chap. 11 in [6]). See also ***kinematic rupture model***.

Ground motion intensity measures Earthquake shaking due to seismic waves, observed at recording sites or experienced by people and structures, is commonly reported in terms of various scalar intensity measures that capture parts of the transient wave-field. Seismogram-based ground-motion intensity measures are, for instance, peak ground acceleration (*PGA*) and peak ground velocity (*PGV*), while the modified Mercalli intensity (*MMI*) is a damage-related measure. In earthquake engineering, ground-motion intensities are often reported as *response spectra*: the response of an idealized building (modeled as a single-degree-of-freedom oscillator) of given eigenperiod T and damping ζ (usually 5%) to a given ground-motion time series. Spectral acceleration (S_A), spectral velocity (S_V) and spectral displacement (S_D) are analyzed considering the period of the structure.

Ground motion uncertainty In ground-motion prediction for engineering purposes, random (aleatory) variability and scientific (epistemic) uncertainty are distinguished. The latter is due to incomplete knowledge and/or limited data, and is captured by alternative empirical attenuation relations or different ground-motion simulation strategies. Aleatory variability is quantified in terms of a standard deviation of an attenuation relation or by a large number of model realizations within a particular simulation method. The distinction between aleatory variability and epistemic uncertainty is particularly useful in probabilistic seismic hazard analysis (PSHA).

Kinematic rupture model A kinematic rupture model characterizes the time-dependent displacement field on the rupture plane without considering the forces or stresses acting on the fault and causing its motion. The rupture process is completely specified by the spatio-temporal distribution of the slip vector, the local ***slip-velocity*** function on the fault, and the ***rupture velocity*** with which the rupture propagates over the fault plane. See also ***dynamic rupture model***.

Path effects Seismic waves propagating through the Earth are sensitive to the detailed geologic structure along the wave path, generating pronounced waveform complexities. Considering crust and upper-mantle structure (relevant for near-field ground motions) three major elements to path effects are distinguished in practice: (a) waves in a flat-layered attenuating Earth; (b) basins and other deterministic deviations from a flat-layered model; (c) random heterogeneities in the three-dimensional velocity-density structure. The distinction between path effects and ***site effects*** is often ambiguous.

Rise time The rise time (or slip duration) τ_r measures how long each point on the fault moves during the rupture process, and must not be confused with the ***rupture duration***. The rise time is related to ***the slip-velocity function***, and is usually measured as the time it takes to attain 5–95% of the final slip at each point. For simple parametric slip-functions (e. g. boxcar, isosceles triangle or combinations thereof), the rise time is generally given by the width of this function.

Rupture duration The rupture duration characterizes the total time for the earthquake rupture process to complete, starting at the nucleation point (hypocenter) and lasting until the last point on the rupture plane stops sliding. Rupture duration therefore depends on ***rupture velocity*** and scales with the size (source dimension) of the earthquake.

Rupture velocity Earthquake ruptures, either modeled as propagating cracks or slip-pulses, expand over the fault plane at rupture speeds (v_r) close to the local shear-wave velocity (v_S), typically in the range $0.5 \cdot v_s \leq v_r \leq 0.9 \cdot v_S$, or about 1.0–3.5 km/sec for crustal earthquakes. However, the crack tip, the transition region from unbroken, intact rock to the currently slipping zone, does not necessarily travel at constant rupture speed. Rupture velocity may locally slow down or accelerate, even to super-shear velocities (in which case the crack front travels at speeds faster than the local shear-wave velocity), depending on the initial and boundary conditions that govern the dynamic rupture process.

Site effects Site effects refer to wave-propagation effects in the immediate proximity to the observation point; they are distinguished from ***path effects*** which comprise the complete path from the source to the receiver (although the boundary between these two is often ill defined). The local sedimentary cover, topography, strong geologic contrasts or water-table variations may contribute to site effects that modify the incoming "bedrock" seismic motions.

Slip distribution The slip distribution represents the cumulative slip on each point on the fault acquired during the co-seismic rupture process (i. e. small contributions from post-seismic slip episodes are ignored). A slip distribution for an earthquake is computed from the space-time integration of ***slip-velocity functions*** on the rupture plane.

Slip-velocity function (Slip-rate function) Each point participating in the rupture process experiences a time-dependent slip history during which the two sides of the fault go through a stage of acceleration, stable sliding, deceleration, and final stopping. This local displacement trajectory is often represented in terms of a slip-velocity function (or slip function) whose details depend on the dynamic rupture process and the constitutive behavior of the host rock. Slip-rate functions are often modeled using simple parametric functions.

Source effects Amplitudes and waveform character of seismic waves are strongly affected by source effects, i. e. by the details of the earthquake rupture process. Far-field signals carry the signature of the overall "point-source" earthquake source mechanism; near-field recordings are very sensitive to the spatio-temporal details of the rupture process, characterized in a finite-fault source model either as ***kinematic*** or ***dynamic rupture model***.

Static stress drop The static stress drop $\Delta\sigma$ represents the difference between the initial and final stress across the fault before and after an earthquake, and is related to slip on the fault. It is defined, based on a shear crack with uniform stress drop, as $\Delta\sigma = C \cdot \mu \cdot D/L_c$, where μ is the shear-modulus, D the mean slip over the fault, L_c a characteristic length scale (usually the smallest dimension of the rupturing fault), and C a constant of order unity which depends on the source geometry. Using the fault width W as characteristic length, static stress drop is related to the seismic moment, $M_o = \mu \cdot L \cdot W \cdot D = C \cdot \Delta\sigma \cdot A^{3/2}$, where A is fault area, and L is fault length. Inferred values of static stress drop are in the range of 0.1–10 MPa, independent of seismic moment, leading to the generally assumed self-similar constant stress-drop scaling (see ► Earthquake Scaling Laws). The static stress drop must not be confused with the dynamic stress drop [164] which captures the time-dependent stress change on a point of the fault during the dynamic faulting event, and may be significantly higher or lower than the static stress drop.

Definition of the Subject

The accurate prediction of the level and variability of (potentially damaging) near-source strong-ground motions in future earthquakes is one of the key challenges for seismologists and earthquake engineers. The increasing number of near-source recordings collected by dense strong-motion networks exemplifies the inherent complexity of near-field ground shaking, governed by a number of (partially interacting) physical processes. Characterizing, quantifying, and modeling (either by means of empirical scaling relations or by numerical simulations) ground-motion complexity requires the joint investigation

of three dominant ingredients: (I) the physics of earthquake rupture; (II) the details of wave-propagation in heterogeneous media; (III) the effects of local site conditions.

This article discusses briefly the beginnings of strong-motion seismology and the recognition of ground-motion complexity. Using two well recorded recent earthquakes, I introduce observational aspects of near-field ground shaking and the basic mathematical description for computing ground motion. The article proceeds by describing each of the three "ground-motion ingredients" in some detail, but does not attempt to provide an in-depth review of all the scientific advancements in these fields. Rather, I explain the key elements for characterizing and modeling ground-motion complexity, supplemented with a concise overview of the underlying physical processes. Current research increasingly incorporates advanced physical concepts into standard practice, therefore leading to improved strong-motion simulation approaches to accurately predict intensity and variability of near-source shaking.

Introduction

As early as 1910 Reid [148] identified strong spatial variations of seismic shaking due to the 1906 San Francisco earthquake, which he correctly attributed to the specific geologic conditions at the locations from which shaking intensities were reported. The space-time complexity of the seismic wavefield in the vicinity of the causative fault became more widely recognized since the beginnings of strong-motion seismology in the 1930s and the establishment of strong-motion networks in the 1960s [9]. Observed peak accelerations frequently reach 1 g, occasionally even exceed 2 g, where nearby stations often show not only much lower peak amplitudes but also very different waveform character. Since the advent of modern digital instruments in the 1990s and corresponding online near-real-time databases, seismologists and earthquake engineers have access to high-quality recordings which irrefutably show the complexity of earthquake shaking. Given a particular location, near-source ground motions vary strongly for different earthquakes, and so do the motions for a single event when recorded at different sites.

While hundreds of earthquakes happen annually in the magnitude range $6.0 \leq \mathbf{M} \leq 6.9$, only about 15 events occur in the range $7.0 \leq \mathbf{M} \leq 7.9$, most of which happening in remote areas and do not cause large damage. Those events are typically recorded at rather large distances (so called teleseismic events) and have weak (low amplitude) motions. Their ground-motion variability is due to differences in the large-scale wave paths through the Earth and due to the properties of the point-source representation for far-distant earthquakes. Weak-motion recordings are also obtained from small nearby earthquakes. In this article, however, we will be concerned with strong-motion observation in the near-field of (large) earthquakes. Such observations are much harder to obtain than teleseismic recordings because we cannot anticipate in detail where future earthquakes may happen (in order to optimize the instrumentation) and also because large crustal earthquakes are rather infrequent.

Investigating the characteristics of near-source ground-motions is of great importance for earthquake engineers who are concerned with seismically safe design, and for seismologists who study the physical processes leading to ground-motion complexity. To illustrate the large ground-motion variability in a single earthquake, I plot horizontal peak ground accelerations and peak ground velocities recorded at 47 sites in the Sept 28, 2004, **M** 6.0 Parkfield earthquake (Fig. 1), contrasted with ground-motion intensities (*PGA*, *PGV*, *PGD*) for the Sept 20, 1999, **M** 7.6 Taiwan (Chi-Chi) earthquake, shown at 441 locations (Figs. 2–4) (waveforms from the COSMOS database, http://db.cosmos-eq.org). First-order observations from these data are: (i) the maximum *PGA* is larger for the smaller earthquake while the highest *PGV* values are roughly identical; (ii) ground-motions tend to be larger close to the fault trace and decay in amplitude with increasing distance from the fault. However, closer inspection reveals strong variability in the near-field motions for each earthquake. The Parkfield data exhibit ground-motion differences between the two horizontal components of motion (fault-parallel and fault-normal, both for *PGA* and *PGV*) and large variability between neighboring sites. This site dependence is even more pronounced for the Chi-Chi event. *PGA*-values are highest for sites very close to the fault, on the hanging-wall (east of the fault trace) and at a few locations in the northward and southward extension of the fault-trace. This pattern changes, however, for the recorded *PGV* and *PGD*-values which are largest on the foot-wall (west side of the fault trace) and towards the northern end of the rupture.

In addition to the scalar measures of shaking amplitude (intensity), Figs. 1–4 display recorded waveforms for selected sites which further illustrate the complexity and variability of near-field motions. Acceleration time series accentuate the high-frequency content of ground-motion and display an almost "shaped random noise" character, with amplitudes that increase rapidly in the beginning and then taper off gradually in the seismic coda (the incoherent wavefield after the arrival of prominent seismic phases). In contrast, velocity waveforms usually show very distinct arrivals and energetic pulses (directivity

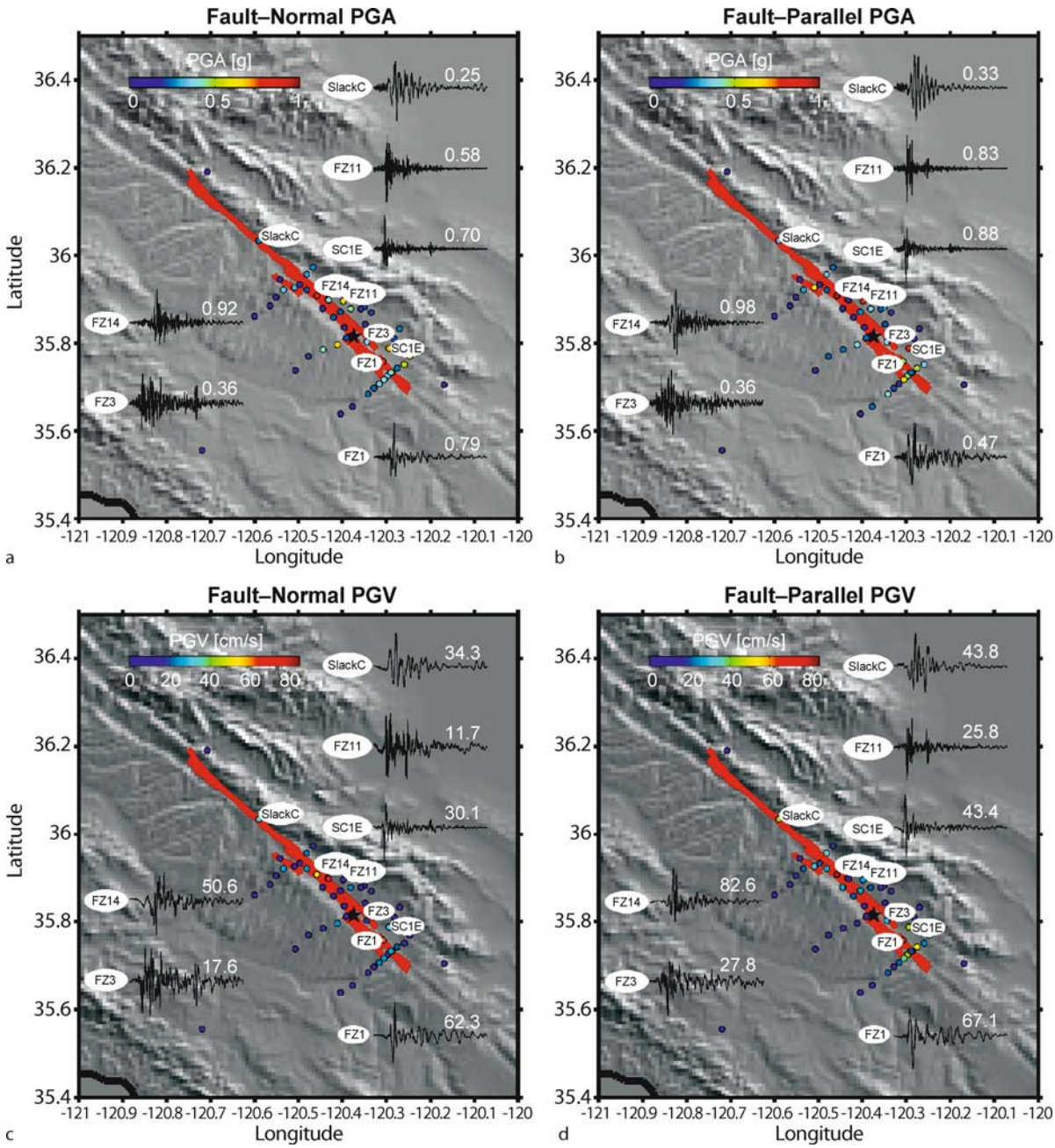

Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 1

Ground-motion intensities (*colored circles*) for the Sept 28, 2004 Parkfield (M 6.0) earthquake, recorded at near-field stations, and selected waveforms (data from COSMOS database). The *black star* denotes the epicenter, (Lon = −120.374, Lat = 35.815) *red-lines* show the mapped fault trace of the San Andreas Fault in the Parkfield area, the *gray background* displays a shaded relief map. a Peak ground accelerations (PGA) of fault-normal component; b PGA of fault-parallel component; c Peak ground velocity (PGV) of fault-normal component; d PGV of fault-parallel component. Seismic traces are shown for 30s, the small number indicates the corresponding PGA or PGV value of that record

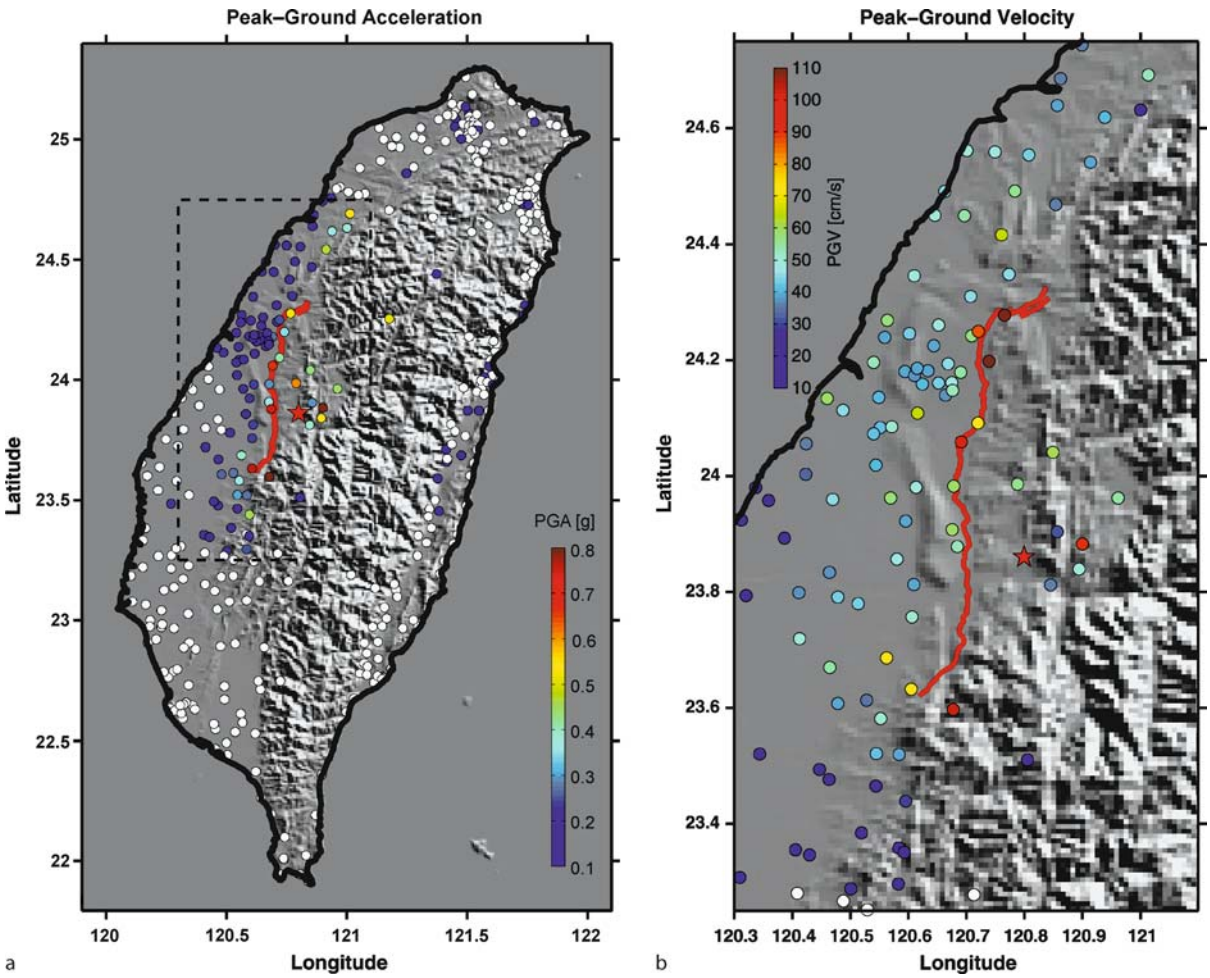

Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 2
Ground-motion intensities (*colored circles*) for the Sept 20, 1999 Taiwan (M 7.6) earthquake, recorded at 441 locations (data from COSMOS database). The *red star* denotes the epicenter (Lon = 120.7995, Lat = 23.860), *red-lines* show the mapped fault trace of the Chelungpu fault, *heavy black lines* trace the coast line, the *gray background* displays a shaded relief map. a Large-scale view of PGA, white dotes are sites with PGA < 0.1 g. b Zoomed view for PGV (white dotes mark PGV < 10 cm/s). c Zoomed view for PGD (white dotes mark PGD < 10 cm)

pulses) whose frequency-dependent amplitude and waveform character change with source-site geometry and rupture propagation direction. Their waveform character also varies strongly between closely spaced stations for a given event, illustrating the importance of localized source properties and site effects. Note also the large differences in ground-motion amplitudes and waveform shapes between the two events for stations at similar distance to the fault, indicating the dependence on earthquake source properties of near-field ground motions. Section "Future Directions" compares the near-field recordings for these two earthquakes against empirical predictions for several published attenuation relations to further illustrate ground-motion complexity.

The variability in shaking intensity and the complexity in near-fault seismograms results from three physical processes: (I) the complex dynamics of earthquake rupture and the associated radiation of seismic waves; (II) the propagation of these seismic waves through the heterogeneous Earth; (III) the interaction of the seismic wavefield with local geology/morphology, referred to as the site conditions at each observation points. Mathematically, the time-dependent ground displacement $u_k(t)$ at a particular location k is described as

$$u_k(t) = s(t) * g_k(t) * l_k(t) \tag{1}$$

where $*$ denotes the convolution operator, $s(t)$ represents the **source effects** due to the earthquake rupture pro-

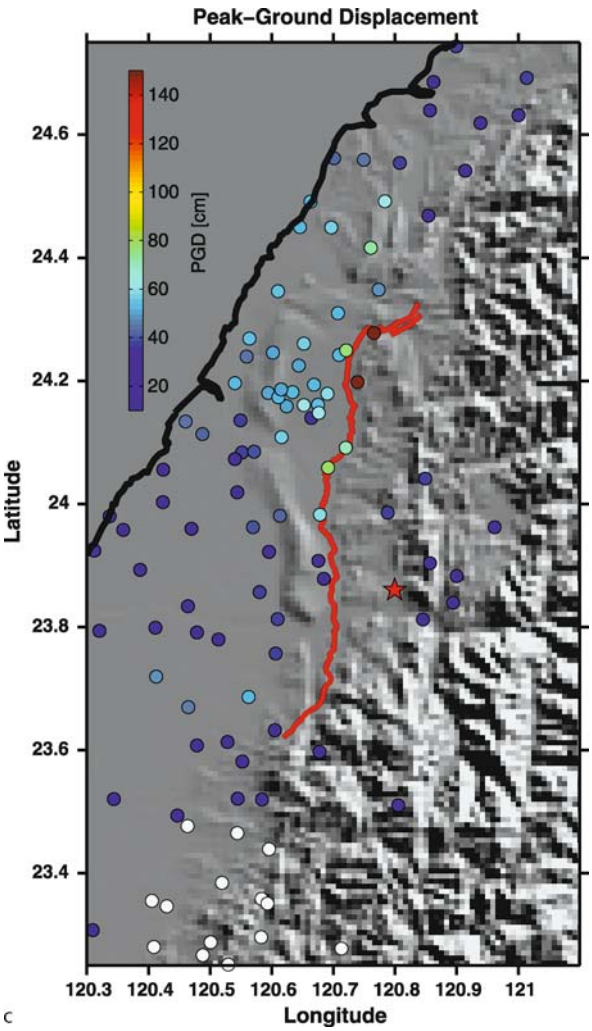

Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 2
(continued)

cess, $g_k(t)$ describes the **path effects** due to wave propagation from the source to site k, and comprises $l_k(t)$ the **local site effects** due to the small-scale geological conditions at the kth observation point. Equation (1) quantifies ground-motion generation, omitting for simplicity an additional instrument response $i_k(t)$ that modulates the seismic recording. The earthquake-source contribution $s(t)$ can be further subdivided into effects originating from the local time-dependent particle motion on the fault, $p_{ij}(t)$, and from rupture finiteness, $f(t)$. The term $g_k(t)$ represents the Earth transfer function and contains contributions from layered Earth structure and seismic-wave attenuation, but may also comprise effects due to random heterogeneities in the Earth and/or basin and topographic structures. The factor $l_k(t)$ describes the effects of soil structure and may also account for non-linear soil behavior.

In the following, I will use Eq. (1) as a "roadmap" for this article to illustrate the various factors of ground-motion complexity. In Sect. "Characterizing Earthquake Source Complexity", I present methods and relevant parameters to quantify the earthquake rupture process. Wave propagation in complex media, both deterministic and stochastic, are described in Sect. "Wave Propagation in Complex Media: Path and Site Effects". A large body of literature exists for each of these topics, both from an observational/experimental view and from theoretical work; it is beyond the scope of this article to provide an in-depth review of all relevant material. Instead, I will focus on some of the key aspects of earthquake ruptures and waves in inhomogeneous media that are most relevant for understanding the complexity of near-field ground motions. Note also that a number of specialized articles in this encyclopedia provide more detailed information on earthquake source physics (see ► Earthquake Nucleation Process, ► Earthquake Scaling Laws) and wave-propagation phenomena in complex media (see ► Seismic Wave Propagation in Media with Complex Geometries, Simulation of, ► Seismic Waves in Heterogeneous Earth, Scattering of). Section "Wave Propagation in Complex Media: Path and Site Effects" also includes various aspects of local site conditions that lead to pronounced site effects, in particular non-linear site phenomena. Section "Ground-Motion Scaling Relations" focuses on empirical ground-motion prediction equations (GMPE's). Since many studies have been published on non-linear site effects and ground-motion prediction equations, both from the seismology and the earthquake-engineering communities, an extensive review of the developments in these fields is not attempted. Instead, I concentrate on non-linear soil behavior directly beneath a site of interest that substantially affects the ground-shaking levels, and summarize some of the latest findings related to ground-motion attenuation relations. The article concludes with an outlook onto future tasks and challenges for characterizing, quantifying and predicting ground-motion complexity.

Characterizing Earthquake Source Complexity

Geologic faults are generally geometrically complex multi-scale structures, characterized by one or more main fault strands, with associated subsidiary branches (fault segments) that form three-dimensional (3D) fault networks. However, in many cases and for almost all practical purposes, faults are approximated as planar surfaces. Using this plane-fault approximation, Fig. 5 illustrates the three

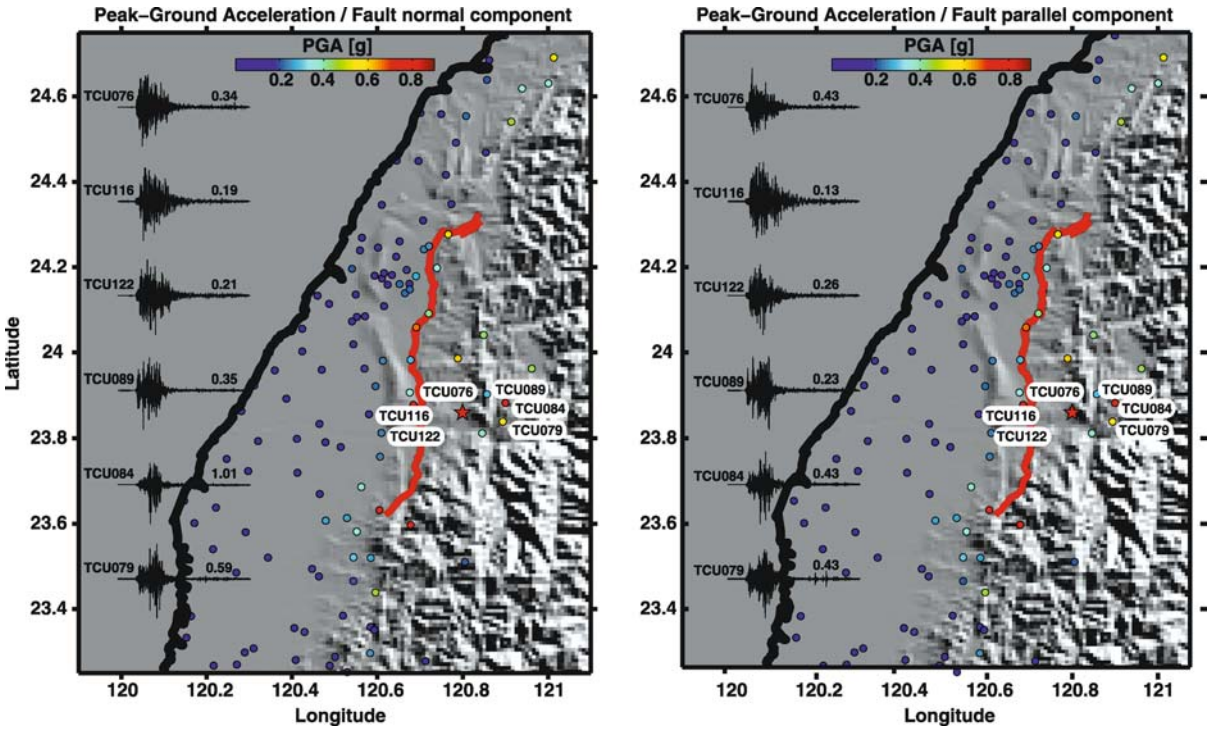

Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 3

Near-field ground accelerations for the fault-normal (*left*) and fault-parallel (*right*) motion during the 1999 Chi-Chi earthquake; *waveforms* are plotted for selected stations in the vicinity of the epicenter (*red star*). Sites located on the footwall, i. e. west of the fault-trace, generally show lower ground-motions than sites located on the hanging-wall (east of the fault trace) despite similar epicentral distance

main factors affecting near-source ground-motion complexity – source, path and site effects. This Section focuses on the earthquake rupture process, i. e. the source, and its properties important for ground-motion generation. A fault (shown as a planar surface with color-coded slip distribution) is embedded in a rock volume, and has a specific orientation in space described by the strike-angle Φ (representing the azimuth of the fault's projection onto the surface, measured clockwise from North) and the dip-angle δ (measured downward from the surface to the fault in the vertical plane perpendicular to the strike). The strike direction is defined such that, using the right-hand rule, the dip-angle is smaller than 90° (for a vertically dipping fault, $\delta = 90°$, the strike direction is arbitrarily either direction).

Given the overall source geometry, the slip-vector on the fault plane defines the relative motion between the two blocks. The angle of slip, or rake angle λ, measured in the fault plane from the strike direction, shows the movement of the hanging wall relative to the foot wall (see inset in Fig. 5). The following definitions apply: $\lambda = 0°$ – *left-lateral strike-slip*, i. e. the hanging wall (or near-side of a vertical fault) moves horizontally to the right, so the opposite side moves to the left; $\lambda = 180°$ – *right-lateral strike-slip*, i. e. the hanging wall (or near-side of a vertical fault) moves horizontally to the left, so the opposite side moves to the right. For $\lambda = 90°$, the hanging wall moves upward (*thrust faulting*), for $\lambda = 270°$ the hanging wall moves downward (*normal faulting*). Figure 5 also displays the rupture model for a right-lateral strike-slip earthquake with $\lambda = 180°$ (indicated by the black arrows) on a fault-plane that dips 80°. The strike in this hypothetical case is undefined. The amount of displacement (slip) on each point of the fault is color-coded, white contours (at $\Delta t = 1$ s spacing) show the expanding rupture front. Examples of past earthquakes show that the rake angle may also vary over the fault plane (Fig. 6c, d). Geometrical complexity, manifested in several fault segments, is often present (Fig. 9); in these cases the faulting-style is characterized by the predominant fault direction and slip angle.

The strength of seismic radiation is typically quantified by an earthquake magnitude, whereby a variety of magnitude scales exist. The seismic moment, defined as $M_0 = \mu \cdot A \cdot D$, ($\mu$: shear modulus in the source region;

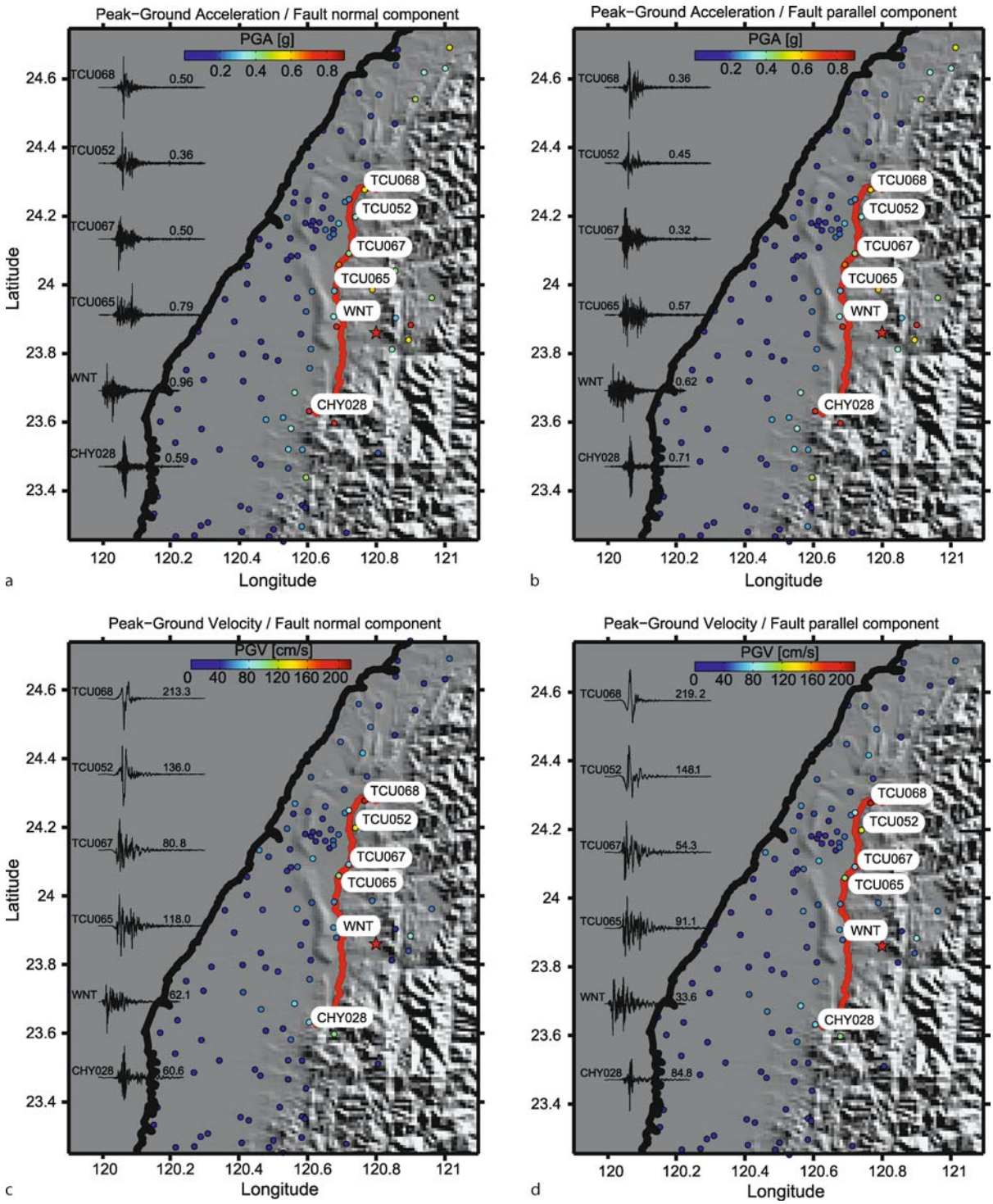

Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 4

Near-field ground accelerations (*top*) for the fault-normal a and fault-parallel b motions during the 1999 Chi-Chi earthquake; *waveforms* are plotted for selected stations along the fault trace (*red line*). c and d show the corresponding ground-velocities. Sites located at either end of the fault trace recorded waveforms with shorter duration of the dominant wave-energy. Velocity records in the North show strong rupture-directivity effects as high-amplitude short-duration velocity pulses

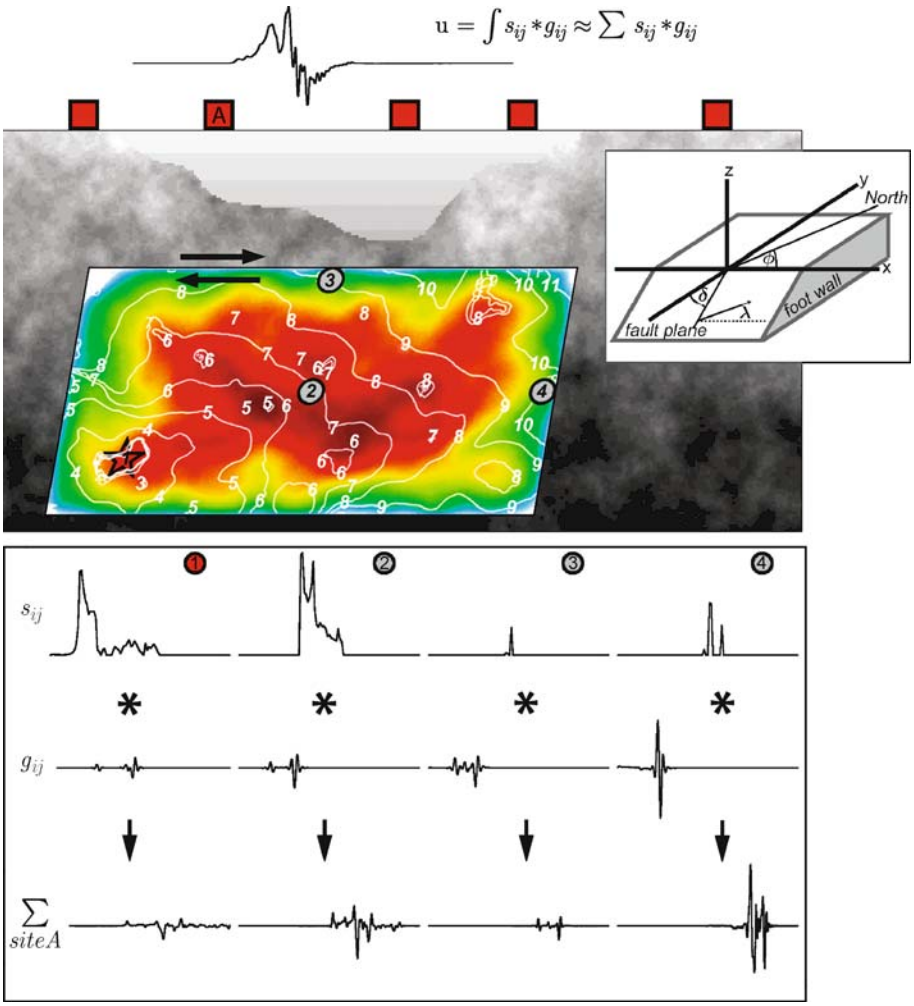

Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 5
Sketch of the physical components of near-source ground motions (*top-right inset*: illustrates the source-geometry definition). A simulated earthquake rupture occurs on an assumed planar fault, with color-coded slip amplitudes (0–2.5 m) and white contours of the propagating rupture front (emanating from the hypocenter shown by the *red star*). The rupture is embedded in a gray-scale model of the Earth' crust. Darker gray-tones denote higher v_P, v_S, ρ than lighter gray, reflecting the overall trend of higher velocities at greater depth, locally disturbed by random heterogeneities. A basin-structure with less compliant layered sedimentary rocks exhibits complex sub-surface topography. A near-field seismogram (for an arbitrary horizontal component) is shown at a representative site A. Ground-motions are composed as the summation of the slip- functions s_{ij} at grid-points i, j on the fault, convolved with the corresponding Green's functions g_{ij} for this observer and grid points (the site term l_k in Eq. (1) is neglected for simplicity)

$A = L \cdot W$: fault area, given by fault-length L and fault-width W; D: average displacement on the fault) is considered the best scalar quantity characterizing earthquake size. Ben-Zion [20,23] proposes to use the more basic concept of seismic potency, $P_0 = A{\cdot}D$, to avoid a source quantification that includes ambiguously defined or poorly known material properties; however, seismic potency is rarely used in the earthquake engineering. The moment magnitude follows from $M_w = 2/3{\cdot}\log M_0 - 6.07$ (with M_0 given in Nm). The static stress drop $\Delta\sigma$ is related to the ratio of average slip to a characteristic fault length, therefore seismic moment is proportional to static stress drop: $M_0 \propto \Delta\sigma \cdot A^{3/2}$. Observational evidences [106,186,187] confirm this scaling for moderate to large earthquakes, suggesting – constant (i. e. scale-invariant) static stress drop and self-similar earthquake source scaling: large earthquakes are only scaled-up versions of small earthquakes. Other scaling relations between observable fault parameters (fault area, length, and width; average slip) and magnitude (seismic moment) [79,122,169,187] partially confirm the self-

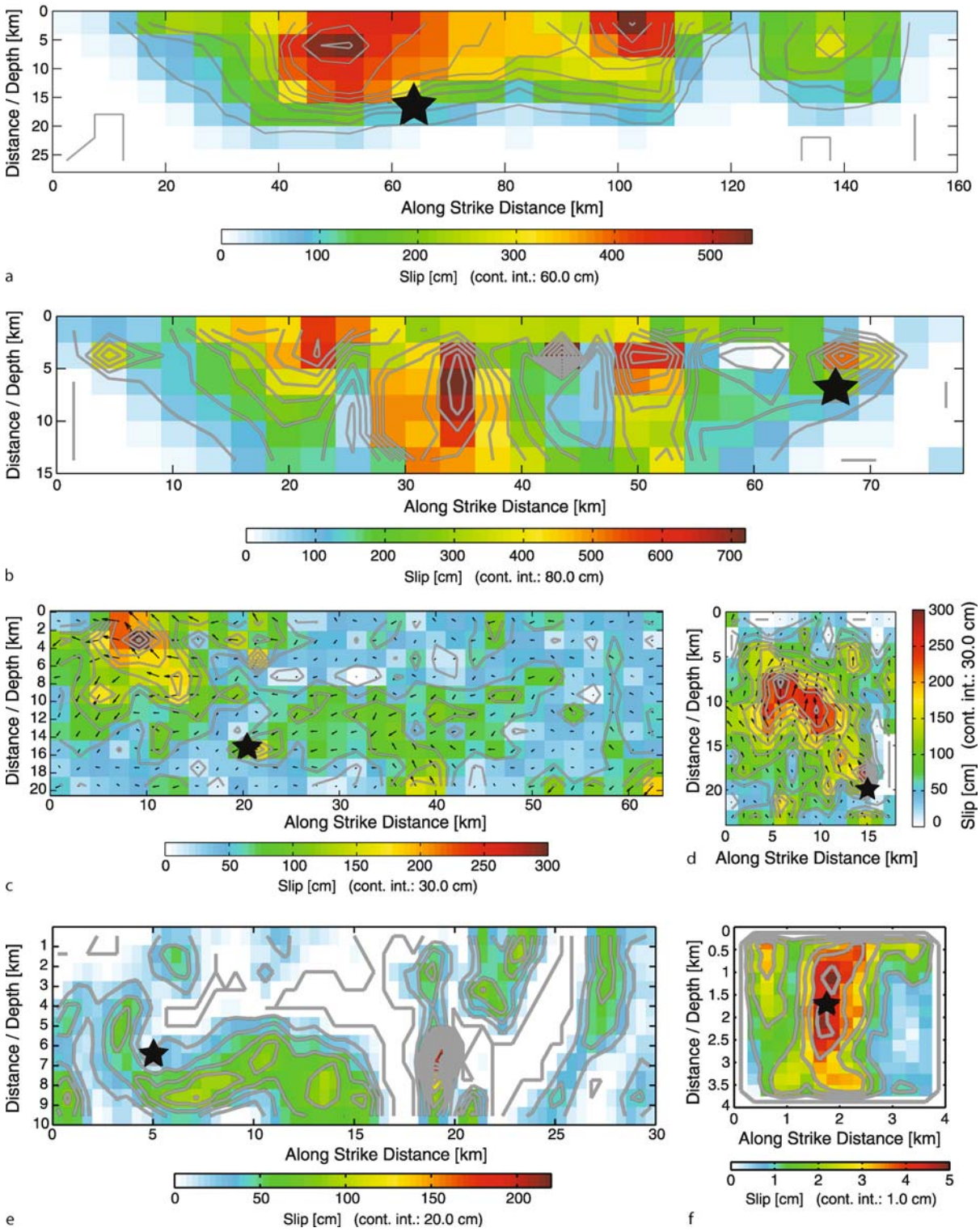
Distance / Depth [km]
Along Strike Distance [km]
Slip [cm] (cont. int.: 60.0 cm)
a
Slip [cm] (cont. int.: 80.0 cm)
b
Slip [cm] (cont. int.: 30.0 cm)
c
d
Slip [cm] (cont. int.: 20.0 cm)
e
Slip [cm] (cont. int.: 1.0 cm)
f

◀ **Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 6**
Selection of finite-source rupture models, obtained from inversion of seismic recordings and geodetic data, for six earthquakes in the magnitude range $4.5 \leq M_w \leq 7.5$, illustrating the heterogeneous distribution of earthquake slip (color-coded using different color scale for each model). The *black star* denotes the hypocenter, generally located in the vicinity of a high-slip patch (also called "asperity"). a 1999 Izmit earthquake (*M*7.5) [51]; b 1992 Landers earthquake (*M*7.3) [184]; c 1995 Kobe event (*M*7.0) [165]; d 1994 Northridge earthquake (*M*6.7) [185]; e 1984 Morgan Hill event (*M*6.1) [29]; f Hida Mountains event #5 (*M*4.5) [96]. The rupture length of b is about half that of a; models b, c and d are drawn to scale with respect to each other. Model e has fault length $L = 30$ km, a factor of two shorter than c, while f has source dimensions 4×4 km^2

similarity of earthquake rupture, but also provide evidence for the break-down of self-similarity for very small earthquakes [23] and very long strike-slip earthquakes [154]. The exact physics leading to these source-scaling properties is a current topic of active research [90,91].

The overall faulting-style, defined by the dip and rake angle, affects the radiation pattern of P- and S-waves (see Chap. 4 in [6] for more details), and thus the resulting ground-motion characteristics. Together with the earthquake magnitude, the style-of-faulting is used as a predictor variable in empirical ground-motion prediction equations (see Sect. "Future Directions"). However, ground-motion complexity arises largely from the details of the rupture process (Fig. 5). The earthquake nucleates at the hypocenter (indicated by the red star), and propagates over the fault plane (white contour lines) with a rupture velocity that may exhibit local variations due to initial stress conditions and frictional properties on the fault. Stress state and friction also determine the slip-time history s_{ij} for all points ij on the fault (which start sliding once reached by the propagating rupture front). To first order, the shape of the local slip-velocity function depends on the overall fault dimensions (i. e. fault aspect ratio), the relative position of each point with respect to the rupture nucleation point and its distance to the closest fault edge [52,58,59,78] (variations in s_{ij} at four different locations on the fault plane of lower panel). The duration of slip (rise time) is related to the length of these local slip-functions. Once sliding has stopped at all points on the fault, the rupture has attained its final slip distribution (color-coded in Fig. 5) within a characteristic time (called rupture duration) which depends on fault dimensions and rupture velocity (and marginally on rise time). As depicted in Fig. 5, the slip distribution, the rupture velocity and the rise times are highly heterogeneous over the fault plane, illustrating earthquake source complexity in terms of kinematic source parameters.

The governing equation that relates ground displacement to the motion on the fault is given by a representation theorem (Chap. 3 in [6]), which we use in the following notation:

$$\mathbf{u}(\mathbf{x}, t) = \int \mathrm{d}\tau \int_S \Delta u(\xi, \tau) c_{ijkl} \nu_j G_{nk,l}(\xi, \tau; \mathbf{x}, t)\, \mathrm{d}S\,. \quad (2)$$

Equation (2) states that the time-dependent ground displacement at observer location $\mathbf{x}$ and time t depends on the space-time integral over the space-time-dependent slip function $\Delta u(\xi, \tau)$ on the fault plane ($\boldsymbol{\xi}$ defines the position on the fault, τ is time), the elasticity tensor c_{ijkl}, the fault-normal vector ν_j, and the Green's tensor $G_{nk,l}$ (subscripted comma indicate the derivative with respect to the subsequent variable). Equation (2) is related to Eq. (1), but contains the functional dependency on the source explicitly; in this context the Green's function typically neglects the site effects $l_k(t)$. The term $\int_S \Delta u(\xi, \tau)\, \mathrm{d}S$ contains the time-dependent local particle motion on the fault, $p_{ij}(t)$, and the effects of fault finiteness, $f(t)$.

Equation (2) can be applied to retrieve the spatio-temporal slip distribution on the fault plane from seismic recordings (and other data) by either forward-modeling or a formal inversion procedure. Earthquake-source inversions have been carried out since the early 1980ies, manifesting the complexity of the rupture process as depicted in Figs. 6 and 9 (see also database of finite-source rupture models [121]). These earthquake source images represent kinematic rupture models that quantify $\Delta u(\xi, \tau)$ based on observations, but do not explicitly derive $\Delta u(\xi, \tau)$ from physical principles as attempted in dynamic rupture models e. g. [14,15,28,78].

For examining earthquake source complexity in more detail, it is mandantory we need to distinguish kinematic from dynamic rupture models, and we need to quantify slip heterogeneity (and its associated stress-change distribution) on the fault plane. It is also important to note that the position of the rupture nucleation point is not arbitrary on the fault plane for a given slip (stress) distribution but adheres to fundamental concepts of energy balance during the rupture process. This Section characterizes source complexity in space and time before introducing to *isochrone theory*, a powerful tool to visualize how the details of the rupture process determines near-fault ground-motions.

Kinematic Rupture Models

Kinematic rupture models characterize the space-time evolution of earthquake rupture in terms of a time-dependent displacement field (distribution of slip vectors)

on a predefined fault plane without considering the forces and stresses that cause these motions on the fault. The local slip-rate (or slip-velocity) function is specified along with the rupture propagation properties. Ground motions can then be computed using Eq. (2) with any kinematic rupture model $\Delta u(\xi, \tau)$. Current ground-motion simulation approaches are largely based on kinematic source models since they can be efficiently generated; recent advancements try to capture at least the basic principles of source dynamics by proposing *pseudo-dynamic* source models [77,78,120]. Earthquake source inversions parameterized in terms of Eq. (2) retrieve only a kinematic rupture model which, in principle, does not need to obey any physical laws. The database of finite-source rupture models [121] provides a compilation of kinematic source models for past earthquakes, obtained by applying Eq. (2) to a variety of data and using different methods to solve the inverse problem.

Dynamic Rupture Models

Dynamic rupture models build a physical understanding of the earthquake rupture process based on the material properties around the source volume, and the initial and boundary conditions for the forces/stresses acting on the fault plane [80]. The distribution of slip vectors $\Delta u(\xi, \tau)$ on the fault plane is obtained by solving the elasto-dynamic equations of motion under the assumption of a constitutive law (i. e. a friction model), considering essentially the energy balance at the crack tip during rupture growth (for details see Chap. 11 in [6]). Dynamic rupture models have been developed for (i) canonical model to study general feature of dynamic rupture e. g. [10,58,59] (ii) for existing kinematic source models to infer their specific dynamic rupture process e. g. [14,15,126,181] (iii) for heterogeneous initial conditions in stress and/or friction and material distribution to investigate rupture details for classes of events e. g. [11,13,63,68,78,83,133,151]. Due to the high computational demands, dynamic rupture models are not (yet) developed routinely for ground-motion simulations, but rather to investigate source physics for general cases, earthquakes of special interest, and to study rupture behavior for certain classes of initial conditions.

Quantifying Slip Heterogeneity

Recent approaches to characterize and quantify slip heterogeneity demonstrate that slip distributions exhibit statistical properties and empirical laws grounded in physical principles. First-order observations for these slip distributions (obtained in finite-source inversions based on Eq. (2)), can be made in Fig. 6, a compilation of slip models for six earthquakes in the magnitude range $4.5 \leq M_w \leq 7.5$. The source dimensions increase from $4 \times 4\,\text{km}^2$ to $25 \times 160\,\text{km}^2$ while the corresponding maximum fault displacements grow by two orders of magnitude (from ~5 cm to over 500 cm). Mai and Beroza [122] have shown that the overall source-scaling relations of such rupture models is roughly consistent with global earthquake scaling laws (i. e. $M_o \propto A^{3/2}$), but that there is evidence that the commonly assumed self-similar constant stress-drop scaling may not hold, because slip on the fault does not saturate but keeps increasing for growing fault dimensions (albeit at a progressively lower rate). The topic of general earthquake scaling laws is still hotly debated e. g. (see ► Earthquake Scaling Laws) [90,91], and directly affects ground-motion prediction [188,189] (see Sect. "Future Directions").

More fundamentally, earthquake slip is heterogeneous on the rupture plane (Fig. 6), i. e. regions of little displacement are separated from areas of high slip (often called "asperities"). The distribution and properties of these high-slip patches strongly influences seismic radiation and hence near-fault ground motions. Characterizing and quantifying slip heterogeneity is thus important for accurately predicting and simulating ground-motions for future earthquakes, but also to better understand the physics of earthquake rupture. Two basic avenues have been pursued in the recent past to quantify slip complexity: (i) a deterministic approach that counts the number of high-slip patches and extracts their properties in terms of size, displacement and other quantities [127,169]; (ii) a stochastic approach that characterizes slip heterogeneity in terms of a random-field model [113,114,123]. Results of both methods can be used for simulating stochastic slip distributions for ground-motion calculation to investigate how source complexity affects near-field ground shaking.

Before interpreting and utilizing inferred source-rupture models for future science or engineering applications, the reliability and resolution of these source-inversion images needs to be addressed. Examining the intra-event variability (source models for the same earthquake obtained by different research groups) often reveals large differences in slip distributions and temporal rupture parameters for the set of source models (online-database [121] for examples and [25] for a critical review). The degree of (dis-)similarity between these slip histories is controlled by differences in the Green's functions (due to computational methods and the choice of the Earth model), variations in the fault parameterization, different inversion schemes and their control parameters (tuning parameters in non-linear inversion; damping/smoothing constraints in linearized multi-time-window inversion). Source-rup-

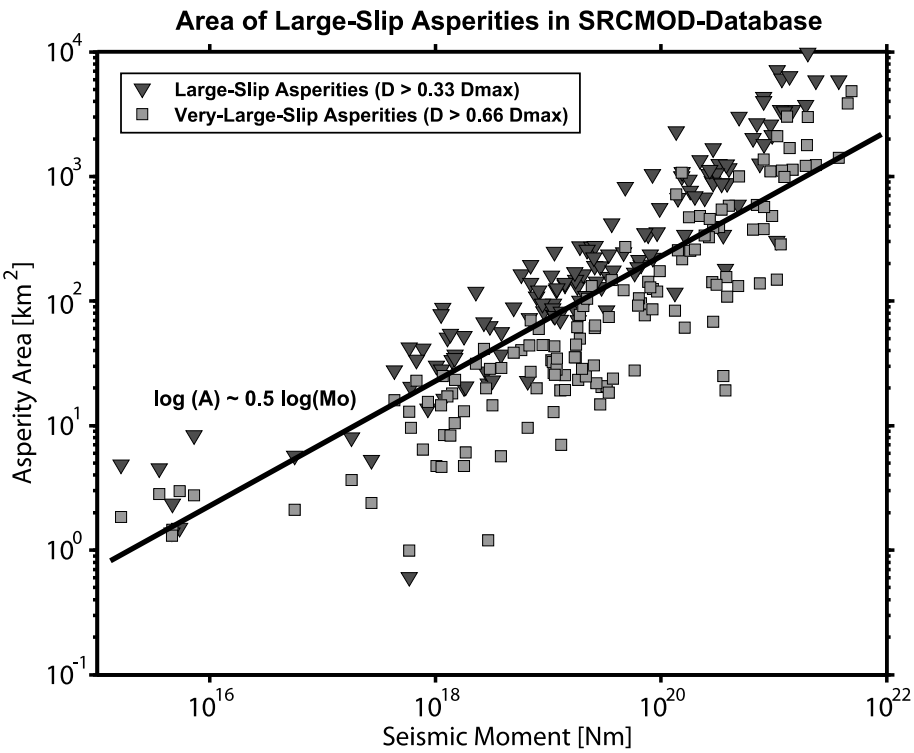

Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 7
Asperity area, measured for ∼90 rupture models (http://www.seismo.ethz.ch/srcmod), plotted against seismic moment. A scaling of $\log_{10}(A) \propto 0.5 \cdot \log_{10}(M_o)$ is shown for reference. Asperity size is measured based on slip values being a certain fraction of the maximum slip on the fault [127]: $1/3 \cdot D_{\max} \leq D < 2/3 \cdot D_{\max}$ for "large-slip asperities", $D \geq 2/3 \cdot D_{\max}$ for "very-large-slip" asperities

ture models are also affected by the type of data (seismic, geodetic, or both; additional geologic data), their selection and processing, and their weighting in the inversion.

Consequently, variations in inverted source models are expected, but the actual uncertainties are rarely quantified. Only recently, more efforts are devoted to perform rigorous uncertainty estimation by testing different inversion algorithms, model parameterizations, and data selection criteria for the same earthquake [56,86,116,117]. The work by Monelli and Mai [128] even estimates posteriori probability density functions for the model parameters of interest, using a non-linear inversion strategy coupled to a Bayesian inference technique. Despite the variability in imaged slip distributions for a given earthquake, several source quantities are stably estimated: rupture dimension, seismic moment and average displacement (both within a factor of 2 generally), and also the slip near the hypocenter. Estimates for the average rupture velocity may vary strongly (up to ∼ 30%) between models for the same event because rupture speed estimates trade off with the slip-rate function. The locations of high-slip patches on the fault ("asperities") are relatively well located if sufficient data are used, but exceptions exist (i. e. the 1999 Izmit earthquake). The intra-event variability of maximum fault slip is generally quite high, and depends on the particular inversion strategy. However, estimates of correlation lengths or dominant wave-lengths of heterogeneous slip maps are robust for the different models of a given earthquake [123], indicating that the statistical properties of earthquake rupture are well imaged. Also the hypocenter location with respect to the regions of high-slip is estimated reliably in finite-source inversions [127].

Measuring slip heterogeneity deterministically by defining large-slip ($\frac{1}{3} \cdot D_{\max} \leq D < \frac{2}{3} \cdot D_{\max}$) and very-large-slip ($D \geq \frac{2}{3} \cdot D_{\max}$) asperities ($D = D(x, z)$ is the local slip on the rupture plane, $D_{\max}$ the corresponding maximum slip), [127] examine the scaling of total asperity area (A_{ta}) with respect to seismic moment for 90 finite-source rupture models (Fig. 7). Without a formal regression, the measurements suggest the scaling $\log_{10}(A_{\text{ta}}) \propto 0.5 \cdot \log_{10}(M_o)$, meaning that the area occupied by high-slip patches grows slower with increasing magnitude than the total fault size (which scales as $\log_{10}(A) \propto \frac{2}{3} \cdot \log_{10}(M_o)$). This scaling requires relatively larger maximum displace-

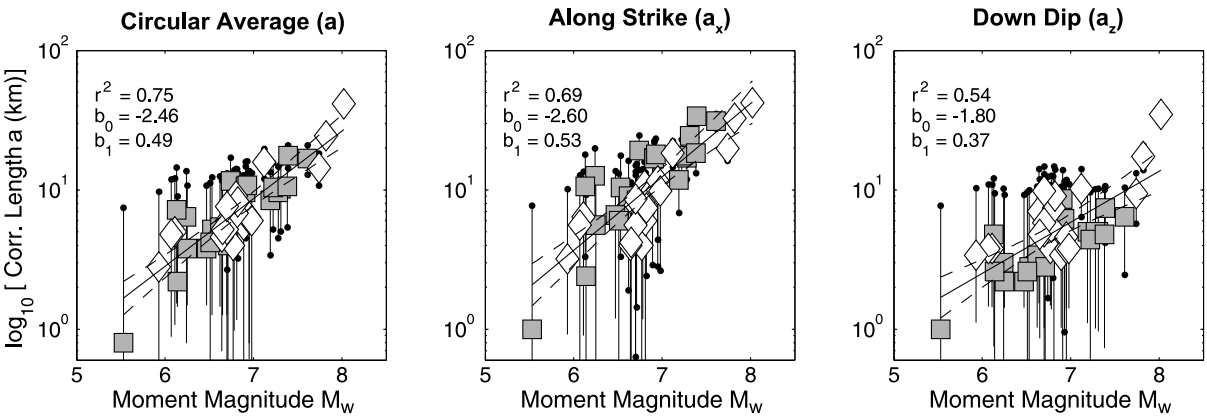

Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 8
Correlation lengths, a, a_x, a_z, versus moment magnitude M_w for 44 slip models (modified after [123]). *Filled squares* denote strike-slip earthquakes, *open diamonds* represent dip-slip events, *vertical lines* and *black dots* mark the 1σ-error-estimates. The slope of $b_1 \approx 0.5$ of the regression curves (*solid lines*) indicates that the correlation lengths scale self-similarly with moment magnitude

ments on the high-slip patches for large and very large earthquakes, which in turn affects seismic radiation and the scaling properties of near-fault ground motions. Large fault slip, concentrated on relatively small portions of the rupture plane, lead to large stress heterogeneity (with high static stress drop locally), thus implying large ground-motion variability for moderate to large earthquakes.

Alternatively, slip distributions (Fig. 6) can be treated as spatial random fields [113,123] to estimate the fractal dimension or the correlation lengths for an auto-correlation function (ACF). Such measurements are typically based on the two-dimensional power spectrum $P(\mathbf{k})$ of the slip map. Assuming a simple fractal model for a random field, its power spectral density is given as

$$P(\mathbf{k}) \propto \mathbf{k}^{-\nu-1} \tag{3}$$

with wavenumber vector $\mathbf{k}$ and scaling exponent ν. [113] estimate one-dimensional scaling exponents for a small number of rupture models, using only horizontal slices of slip distributions, and find $0.8 \leq \nu \leq 1.5$, while [123] find for the two-dimensional exponents $(\nu + 1)$ values of about 1.7, implying a fractal dimension $D = 2.3$. The differences in these estimated scaling exponents can be reconciled when accounting for slightly different initial processing of the slip distributions and fitting strategies [114]. Testing a variety of auto-correlation functions to fit the spectral properties of a large number of slip distributions, [123] observe that a von Karman auto-correlation function, with its power-spectral density given by

$$P(\mathbf{k}) = \frac{4\pi H}{K_H(0)} \cdot \frac{a_x \cdot a_z}{(1 + \mathbf{k}^2)^{H+1}} , \tag{4}$$

with magnitude-dependent correlation lengths a_x, a_z, and a scale-invariant Hurst number $H \approx 0.7$ best matches the slip heterogeneity spectra (K_H: modified Bessel function order H; $\mathbf{k}$ wavenumber). Figure 8 displays their estimated correlation lengths for 44 slip models [123], along with a least-squares regression that exhibits scaling relations of the form

$$\begin{aligned} \log_{10}(a_x) &\propto \frac{1}{2} M_w \\ \log_{10}(a_z) &\propto \frac{1}{3} M_w \, . \end{aligned} \tag{5}$$

Interestingly, the scaling law in Eq. (5) for correlation lengths of heterogeneous slip maps is similar to the relationship inferred for the deterministic measure of asperity size. This corroborates the previous argument that correlation lengths (asperity size) increases with increasing magnitude, though at a lower rate than the overall fault dimensions. Thus, to accommodate the corresponding seismic moment, the displacements (and associated stress drops) on these high-slip patches need to grow faster than a self-similar scaling would predict. This conclusion agrees with the conjecture of Heaton [88] that large earthquakes require a "very strong kick" (due to an area of large stress drop) in order to be able to grow into a very large rupture.

Treating earthquake slip as being distributed on a *planar* fault surface strongly simplifies geologic observations of geometrical fault complexity, where the degree of complexity depends on fault-zone maturity [20,90]. Earthquakes that break several fault segments can still be imaged with Eq. (2) using an appropriate parameterization, but planar-fault ruptures exhibit very different rupture dy-

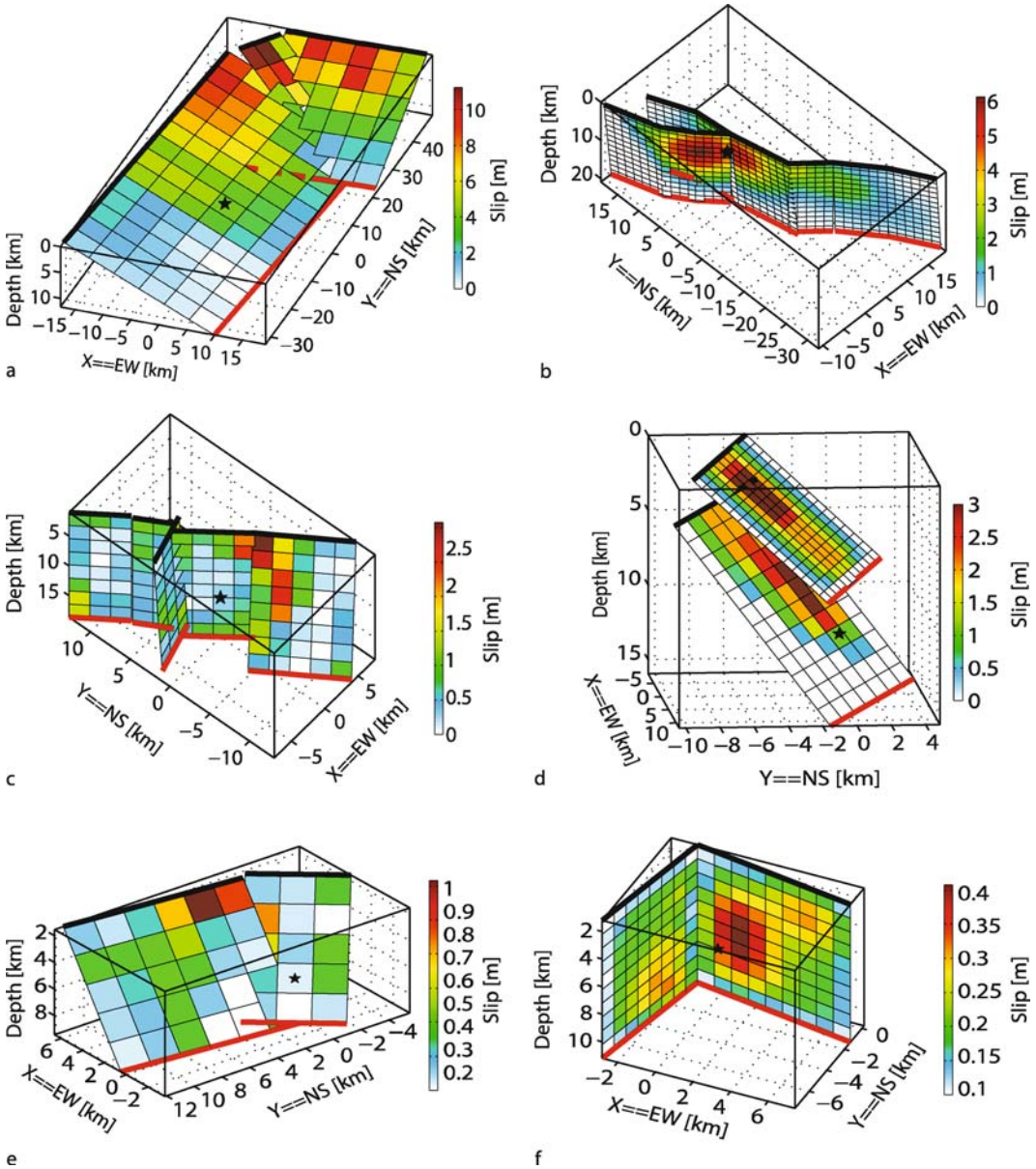

Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 9

Three-dimensional views of heterogeneous slip on geometrically complex faults, imaged using seismic and/or geodetic data. The *black star* shows the hypocenter, *thick black lines* are the top of the fault segments, *thick red lines* the bottom. a The 1999 Chi-Chi earthquake (M_w7.6) [101]; b the 1999 Hector Mine earthquake (M_w7.2) [102]; c the 2000 Tottori earthquake (M_w6.6) [100]; d the 1971 San Fernando event (M_w6.6) [87]; e the 2003 Miyagi-hokubo event (M_w6.1) [89]; f the 1997 Kagoshima earthquake (M_w6.0) [94]

namics than earthquakes that have to overcome geometrical obstacles [82,84,131,134]. Figure 9 shows a collection of such geometrically complex fault models, involving two or more segments that form either a system of sub-parallel fault planes (Fig. 9d), or a branching fault (Fig. 9b), or fault planes oriented at arbitrary angles to each other (Fig. 9a, c, e, f). The seismic radiation from earthquakes on geometrically complex faults is more complicated than from a single-plane rupture, a topic that is investigated in the earthquake source dynamics community e. g. [14,15,43,81,82,84,131,134]. Unfortunately, little attention has been devoted in the past to rigorously exam-

ine and quantify the degree of ground-motion complexity generated by such fault systems. Thus, future research, using advanced numerical techniques on high-performance computing architecture, needs to properly represent fault systems at sufficient spatial resolution to capture the intricacies of rupture dynamics and associated seismic radiation on geometrically complex faults.

Rupture Nucleation and Directivity

Besides the slip heterogeneity and the temporal rupture evolution (discussed later), the location of the hypocenter (point of rupture nucleation) is a critical factor affecting near-source ground motions. Some metric of hypocentral distance enters empirical attenuation relations (Sect. "Ground-Motion Scaling Relations"), while the on-fault hypocenter location determines the directivity effect [170,171]. This global directivity effect due to fault-hypocenter-site geometry results in large velocity-pulses, in particular at sites close to the fault, because of the constructive interference of S-wave energy which is continuously radiated from the propagating crack front and arrives within a short time window in the forward-direction of rupture propagation (see velocity records in Fig. 1 at station SlackC, FZ1, SC1E for the Parkfield event; Fig. 4 at station TCU052, TCU068 for the Chi-Chi event). In contrast, for sites in the backward direction of rupture propagation the arriving seismic energy is spread over a longer time interval, generating lower-frequency motions with smaller amplitudes. The directivity effect is therefore not a purely source-related phenomenon, but also depends on the observer location.

Additionally, there is an "on-fault" directivity effect which operates on smaller scales and is most prominent for large earthquakes and very near-fault sites. The on-fault directivity effect can be efficiently quantified using isochrone theory (see Sect. "Isochrone Theory"). For an explanatory description, consider the slip distribution and hypocenter location in Fig. 6a, for which the large-scale "global" directivity effect was observed towards the right (i. e. East in this case of the 1999 Izmit earthquake). However, any site located at about $30 \leq X \leq 50$ km in along-strike distance (above the left-most high-slip region) experiences strong "local" directivity effects as the rupture propagates from the hypocenter towards the site and across the high-slip patch. In this case, the integrated slip along this rupture trajectory generates a high-amplitude short-duration pulse. Considering a more westerly hypocenter (e. g. at $X = 30$ km instead of $X = 65$ km) the large-scale directivity effect remains essentially unchanged while the integrated slip along the rupture trajectory for stations at $30 \leq X \leq 50$ km is much smaller, thus diminishing on-fault directivity. The spatial relation between the hypocenter and large-slip regions plays an important role for ground-motion complexity, but is also fundamentally related to the dynamics of the rupture process.

Examining the hypocenter positions in Fig. 6 and Fig. 9, two observations are evident: (a) ruptures generally do not nucleate at the fault boundaries, but in the interior of the fault plane, but rarely exactly in the center; (b) the nucleation point is generally not located in regions of low slip (light colors, $D \leq 0.1 \cdot D_{\max}$) but in areas where $0.2 \cdot D_{\max} \leq D \leq 0.4 \cdot D_{\max}$ and the distance to a nearby large-slip zone (asperity) is small compared to the overall source dimensions [127]. This study investigated these first-order observations statistically using a database of ∼80 finite-source rupture models, and concluded that hypocenters are not randomly located on a fault but are located either within or close to regions of large slip (Fig. 10). More specifically, ruptures nucleate within or very close to large-slip asperities ($\frac{1}{3} \cdot D_{\max} \leq D < \frac{2}{3} \cdot D_{\max}$), but rarely start on very-large-slip asperities ($D \geq \frac{2}{3} \cdot D_{\max}$; often located far away from the hypocenter).

These observational constraints of slip heterogeneity and rupture nucleation are rooted in the energy balance of earthquake source physics, and have been confirmed by dynamic rupture simulations [78,133,151]. The essence of the physical mechanism is that a sustained large earthquake can only be generated if a sufficiently large amount of energy is furnished to the propagating crack tip to facilitate rupture growth. In a simplistic view, an earthquake can only grow in size if the energy absorbed to create new crack surface (fracture energy) balances the available elasto-static energy and the energy radiated by seismic waves (Chap. 11 in [6]). If the hypocenter is located in regions of low slip (low stress drop) and far away from any point of significant stress drop, the required fracture energy will become too large to allow further crack propagation, and the rupture will stop prematurely. The spatial correlation between hypocenter and asperity is thus consistent with this simplified energy budget of dynamic rupture. Moreover this hypocenter-asperity relation manifests the "on-fault" directivity and strongly effects seismic radiation and thus ground-motion complexity.

Temporal Rupture Evolution

The characteristics of the slip distribution and the hypocenter position alone are insufficient to parameterize $\Delta u(\xi, \tau)$ in Eq. (2) because the temporal rupture evolution is not yet specified. Rupture velocity and the local slip-velocity function with its associated rise time (slip duration)

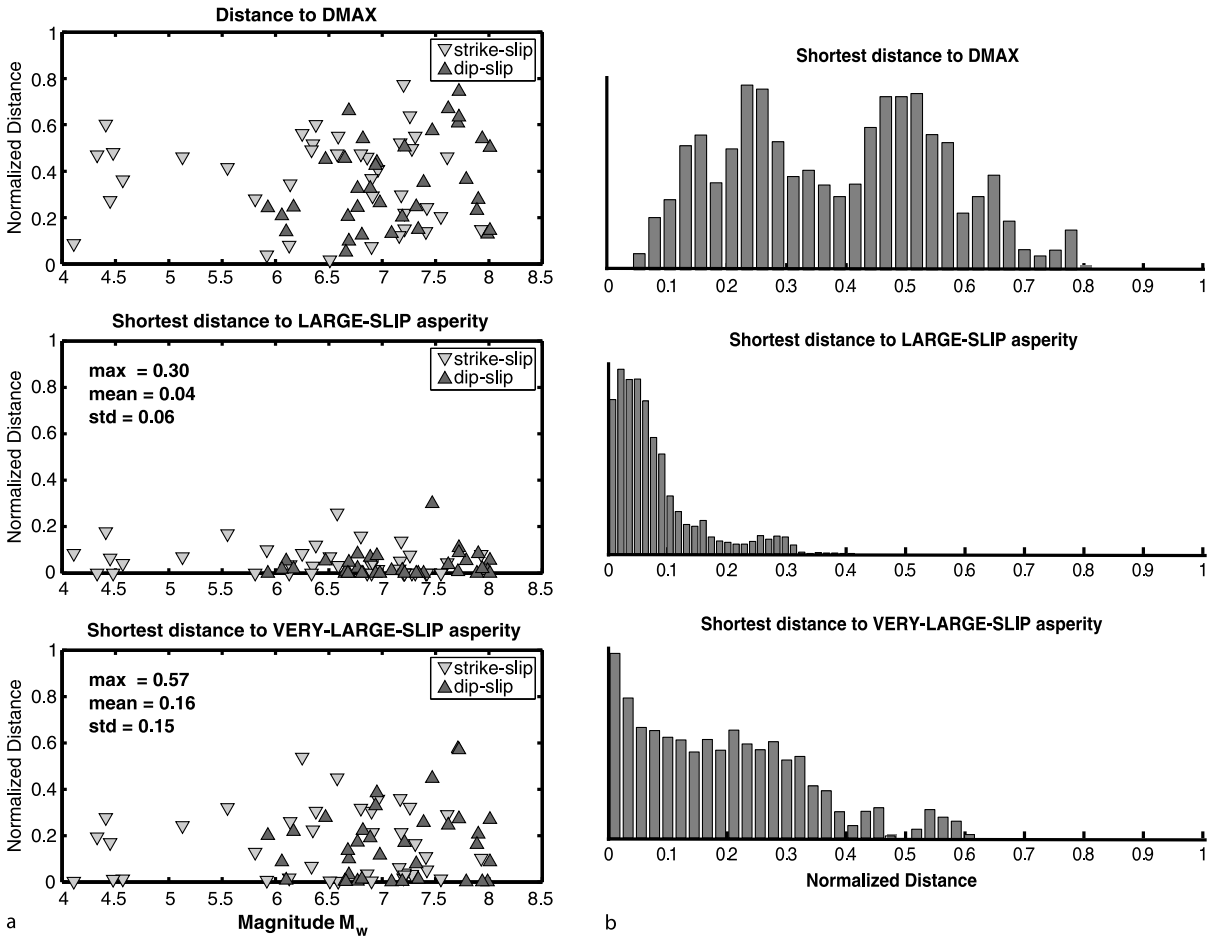

Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 10
Hypocenter distance to "large-slip" and "very-large-slip" regions in finite-source rupture models. a Raw measurements of shortest distances between hypocenter and point of maximum displacement (*top*), hypocenter and closest large-slip asperity (*middle*), and hypocenter and closest very-large-slip asperity (*bottom*), separated into strike-slip and dip-slip earthquakes. b Distributions of shortest distances between the hypocenter and the point of peak-slip, the closest large-slip, and the closest very-large-slip asperity, generated using Monte Carlo simulations that include hypocenter uncertainties. The distributions indicate that 16% of all hypocenters occur on a very-large-slip asperity, 35% within a large-slip asperity, and 48% right outside an asperity (compiled from [127])

need to be given, either pre-defined by means of a kinematic rupture modeling approach or solved for from physical principles in a dynamic rupture model.

Rupture Velocity After earthquake nucleation the rupture propagates over the fault plane with a rupture velocity v_r which generally is close to the local shear-wave velocity (typically in the range $0.6 \cdot v_s \leq v_r \leq 0.9 \cdot v_S$). Seismic radiation and thus near-field motions strongly depend on rupture velocity: slow earthquakes radiate little seismic energy while fast ruptures generate higher ground-motion amplitudes. Earthquake source inversions mostly assume a constant rupture velocity over the fault plane, although dynamic rupture modeling and evidence from particularly well recorded earthquakes indicates that rupture speed may vary significantly on the fault plane [16,44]. Dynamic modeling shows that the initial rupture speed (during and right after the nucleation phase) may be small (e. g. $v_r^{\text{init}} \leq 0.5 \cdot v_S$), but then rapidly increases during the rupture's dynamic breakout and propagation phase (Chap. 11 in [6]). Depending on the stress conditions and the frictional parameters on the fault, and the geometrical properties of the rupturing fault, the rupture speed may actually exhibit very strong small-scale variations (see Fig. 5 for example). Variations in rupture velocity are a source of high-frequency seismic radiation [119,173], and thus strongly contribute to near-field ground motion complexity.

While the rupture speed for most earthquakes is lower than the local S-wave velocity (i. e. sub-shear rupture propagation), there is evidence from laboratory measurements [155] and from seismic data and source modeling e. g. [16,44,134] that the crack front may travell at super-shear speeds at least over parts of the fault plane. Such super-shear ruptures generate rather distinct ground-motion characteristics at the few sites where they were recorded, but it is not yet clear what the "generic" ground-motion signature of super-sonic rupture speed would be. [1] address some aspects of this question by kinematic rupture modeling with various rupture speeds and computing the resulting near-field motions, indicating very peculiar "mode-switching" of the seismic-energy distribution on the two horizontal components as the rupture speed becomes very large. Furthermore dynamic rupture simulations [10,58,59,63,68,134] provide physical models for the occurrence of super-shear propagation, which depends on local stress and frictional conditions on the fault plane. The potential occurrence of super-shear rupture burst [62,63] and rupture speed variations in general are critical for understanding ground-motion complexity. To efficiently include such effects into ground-motion simulations, [78] developed an initial physics-based representation of rupture-velocity heterogeneity for kinematic source modeling.

Slip-Velocity Function and Rise Time As the propagating rupture sweeps over the fault plane points on the fault plane are "activated" and start slipping. Each point traces an individual slip-time history (slip-velocity function) whose shape and duration depends on the stress conditions and frictional properties on the fault, but also on overall fault size and the position of each point with respect to rupture nucleation [59]. Figure 5 illustrates the variability of local slip-velocity functions s_{ij} on the fault. The slip duration (rise time τ_r) is usually defined by integrating the slip-velocity function and measuring the time it takes to complete 5–95% of the total displacement at each point.

Several approximations of these complicated slip-velocity functions are in use: a boxcar or isosceles-triangle function with a rise time equal to the width of the function, or symmetric and asymmetric cosine-functions. The classical approximation uses the solution of a quasi-dynamic crack model [110], showing an $1/\sqrt{t}$-decay after a rapid onset. Recent modifications to this Kostrov-type slip function [129,180] are compatible with earthquake dynamics, while ground-motion simulations and source inversion often assume simple parameterization of overlapping triangles [78,92].

As Fig. 5 indicates, rise time varies over the fault plane. However many kinematic source inversions and simple source-model simulations assume constant rise time. A self-similarly expanding crack model predicts longer rise times in the center of the fault e. g. [58,110] where the rupture nucleates, but short rise times at the crack periphery due to the earlier arrival of the healing front from the crack rim. Analysis of source-inversion results [88] and dynamic rupture modeling [28] however suggest that rise times are in general short, i. e. the rupture propagates as a self-healing pulse over the fault plane with rise times determined by local healing of the rupture front.

For fixed rupture velocity and slip distribution, ground-motions are very sensitive to rise time variations. Shorter rise times lead to larger ground-motions as the seismic energy is released in a shorter time interval. The detailed shape of the slip-velocity function is less important for the seismic waveforms, because they are determined by the summation of slip-functions, convolved with the appropriate Green's function, over the entire fault plane (effectively filtering out small-scale features of the slip-velocity parameterizations). However, the peak slip-velocity value that the slip-rate function may attain is crucial for the final ground-motion amplitudes.

Isochrone Theory

The effects of various kinematic source parameters (slip and slip-velocity distribution, rupture propagation) on near-field ground-motions can be efficiently visualized in the framework of the isochrone theory, a high-frequency (ray-theory) approximation to calculate seismic radiation from earthquake ruptures [27,173]. By approximating the elastic wave Green's functions with far-field body waves, surface-waves and low-frequency near-field terms are not considered in the isochrone method. It is correct, however, at higher frequencies, where far-field terms dominate, and in the distance range to the fault where surface waves have not yet developed but near-field terms are unimportant. The main aspect of this approach is that the space-time integration in Eq. (2) can be replaced by a series of line integrals over the fault for simple slip-function parameterizations. For each time t_i in the observer seismogram a line integral is computed for which the integration path comprises only those points on the fault which radiate seismic waves that arrive at the observer location at exactly the time t_i. Defining an arrival-time function, composed of the rupture-front arrival time at each point on the fault and the corresponding seismic-wave travel time to the observer, the integration path represents an isochrone of this arrival-time function. The corresponding isochrone veloc-

ity, **c** (the spatial derivative of the arrival-time function) is related to rupture velocity v_r, and thus resembles the directivity function. Since ground velocity is proportional to isochrone velocity, the characteristics of high-frequency body-waves in near-field seismograms can easily be examined with respect to arbitrarily heterogeneous slip and rupture velocity distributions.

The omission of near-field terms and surface-waves restricts the isochrone method to cases in which seismic radiation from compact slip zones or sudden changes in rupture velocity dominate the ground-motions. Such source behavior has been reported for many earthquakes, consistent with [119] who showed that the high-frequency far-field radiation is emitted at the propagating crack tip, with additional radiation coming from stopping phases as the rupture heals. The isochrone theory has also been successfully applied to earthquake source inversions [29].

The detailed theoretical development of the isochrone theory is beyond the scope of this article, but a brief description will help to illustrate the concept. Spudich and Frazer [173] link the representation theorem, Eq. (2), to geometrical ray theory (without Fraunhofer approximation) for the far-field displacements of P- and S-waves. According to [176] a simple, yet versatile, parameterization of the slip function $s(\mathbf{y}, t)$, is given by

$$\mathbf{s}(\mathbf{y}, t) = \mathbf{s_r}(\mathbf{y}) f_r[t - t_r(\mathbf{y})] \,, \tag{6}$$

with rupture time t_r at positions **y** on the fault plane, position-independent shape-function f_r for the slip-velocity function, and position-dependent amplitude **s**. Any heterogeneous rupture model can be approximated by Eq. (6). Inserting this expression into Eq. (2), shown here for S-waves only, one obtains

$$\mathbf{u}^S(\mathbf{x}, t) = f_r(t) \cdot \int_S \mathbf{s_r} \cdot G_a^S \delta\left(t - t_a^S\right) \mathrm{d}S \tag{7}$$

where $t_a^S(\mathbf{y}, \mathbf{x}) = t_r(\mathbf{y}) + t^S(\mathbf{y}, \mathbf{x})$ is the arrival time function for an observer at location **x** due to an S-wave radiated at point **y** on the fault G_a^S is the corresponding Green's function. The surface integral in Eq. (7) is non-zero only if the argument of the δ-function is zero. Curves $\mathbf{y}(t_a^S, \mathbf{x})$ define the contours of equal arrival time (isochrones) at observer **x**; the surface-integral in Eq. (7) thus reduces to a line integral

$$\mathbf{u}^S(\mathbf{x}, t) = f_r(t) \cdot \int_{y(t,\mathbf{x})} \left(\mathbf{s_r} \cdot G_a^S\right) \cdot c(\mathbf{y}, \mathbf{x}) \, \mathrm{d}l \tag{8}$$

in which $c(\mathbf{y}, \mathbf{x}) = |\nabla t^S(\mathbf{y}, \mathbf{x})|^{-1}$ represents the "isochrone velocity" of these curves along the fault surface (∇ denotes the gradient operator). Note that the isochrone velocity is related to the directivity function; isochrone theory therefore helps to visualize on-fault directivity effects [172].

As an application of the isochrone method, we compute near-field seismograms for a hypothetical rupture (M_w 6.7) with constant rupture velocity (radially spreading rupture front) at three sites, each at fault-perpendicular position $y = 0.2$ km, but various positions along-strike (Fig. 11). A simple layered medium is assumed, and anelastic attenuation is excluded. For each observer, the S-wave travel time function (travel times from the site to each point on the fault) is shown (top panels), the color-coded integrand (Eq. (8)) with isochrone contours (spacing $\mathrm{d}t = 0.5$ sec) (2nd row of panels) and the isochrone velocity with isochrone contours (3rd row of panels). S-wave pulses in the fault-normal velocity seismograms (bottom) can then be attributed to isochrone properties. Consider for example observer D. The first isochrone appears at ~7 sec, consistent with the S-wave arrival time in the seismogram. Isochrones up to 8 sec are entirely in areas where the integrand is negative (yellow to red colors), hence the down-ward initial motion. The peak velocity (53.2 cm/sec) occurs at $t \approx 9$ sec when the corresponding isochrones hit the large-amplitude regions of the integrand (green areas). Observer D exhibits stronger directivity than observers A or B due to its position with respect to the hypocenter and the dominant asperity, generating large amplitudes due to high isochrone velocity over regions of large slip. At site D, the seismic energy radiated from the high-slip patch at $x \approx -15$ km, $y \approx 10$ km arrives within a short time window, whereas at site A, the contributions from this slip patch are spread over a longer time window. For site A, the isochrone velocity is smaller than for D, hence directivity is less pronounced. As a results, site A exhibits lower peak ground motions but a longer shaking duration.

The current implementation of the isochrone method [176] could be extended to include near-field terms [104] and to be applicable in three-dimensional velocity structures by adding 3D ray-tracing capabilities. Spatially variable slip-velocity functions and anelastic attenuation are further possible extensions. Finally, effects of wave-scattering in random media could be implemented using scattering operators [125,192,193,194] convolved with the local slip-rate function $s(\mathbf{y}, t)$.

Wave Propagation in Complex Media: Path and Site Effects

Returning to Eq. (1), this Section discusses path effects, $g_k(t)$, due to source-to-site wave propagation, and local site effects, $l_k(t)$, due to the small-scale geological con-

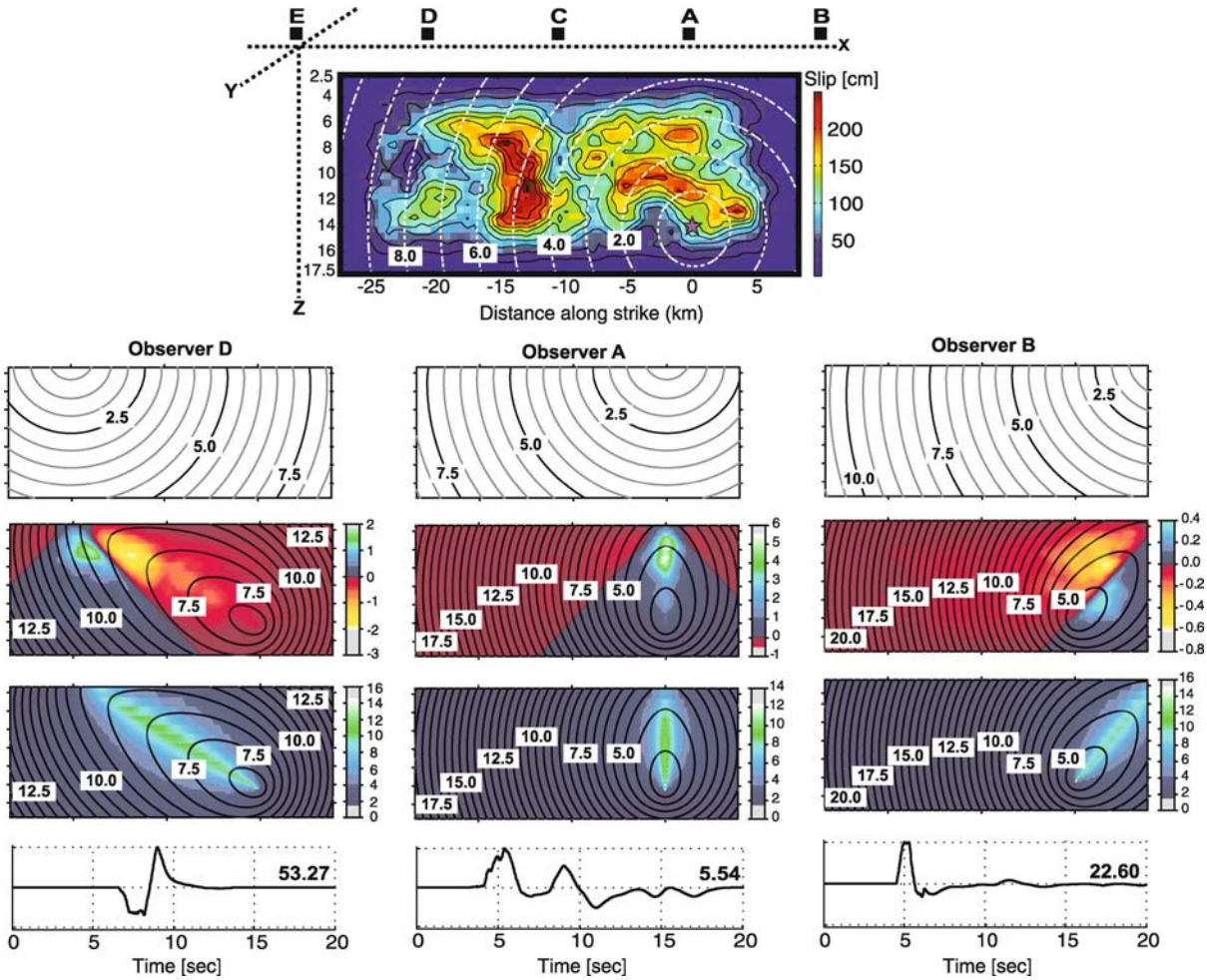

Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 11

Ground-motion simulation using isochrone theory for a hypothetical M_w 6.7 earthquake, buried at 2.5 km depth and embedded in a layered medium. *Top graph*: Rupture spreads from the hypocenter (red star) at constant rupture speed ($v_r = 2.7$ km/s, white contours) over the fault with heterogeneous slip (color-coded). *Bottom graphs*: Isochrone quantities at three sites. *Top-most panels*: S-wave arrival time function; 2nd *row of panels:* isochrones overlain over the real part of the integrand in Eq. (8) (in units of $s_r \cdot G_a^S \cdot c(y, x)$); 3rd *row of panels*: isochrones overlain over isochrone velocity (in km/s). *Lower-most panels*: Resulting fault-normal velocity seismograms dominated by large S-wave pulses (in cm/s; peak velocity indicated at the end of each trace)

ditions at each observation point. The wave propagation from the earthquake to the observation point depends on the complexity of Earth structure, which can be separated into two parts: (i) deterministic wave-propagation within a layered medium, three-dimensional basin effects, topographic features and large-scale geological structures; (ii) wave-propagation in a stochastic random medium with small-scale heterogeneity leading to incoherently scattered wave energy. Local site effects in the shallow near-surface structure (usually the top-most 100m) beneath the observation site further complicate near-fault ground-motions, leading to either increased or decreased motions (compared to bed-rock level), increased shaking duration, shifts in the dominant frequency of ground-motion, and perhaps nonlinearity effect. These wave-propagation effects have to be incorporated into the characterization of ground-motion complexity, and should be applied in strong-motion simulations for predicting the intensity and variability of seismic shaking. However, the success in incorporating these effects hinges on the available knowledge of seismic properties of the Earth, which in turn depends on the dominant wavelengths (frequen-

cies) at which the Earth is sampled by observational data. Similarly, inversion for Earth structure and forward simulation of seismic waves require an appropriate spatial discretization of the modeling region for the numerical calculations, which essentially provides an upper limit on either the maximum resolvable frequency or the size of the computational domain for ground-motion simulations.

In this Section, I first I review some aspects of "deterministic" wave propagation in a flat-layered attenuating Earth, and then qualitatively describe effects of basin structures, Earth topography and local features (e. g. fault zones, narrow belts of low shear-wave velocity). I then examine aspects of wave propagation in random media and corresponding simulation methods. Site effects and nonlinearity conclude this section. Due to the large body of literature on these topics, a detailed quantitative review is not attempted in this Section. Instead I merely select a few key elements of particular interest for characterizing ground-motion complexity.

Wave-Propagation in a Flat-Layered, Attenuating Earth

The first-order approximation of Earth' internal structure is a radially layered medium, which for short source-site distances (< 600 km) can be treated as a flat-layered structure. Many research branches in seismology successfully apply this approximation of a one-dimensional depth-dependent velocity-density model: earthquake location, arrival-time measurements of various seismic wave types, focal-mechanism determination, moment-tensor inversion, finite-source inversion. A number of well tested methods exist to compute Greens functions for a flat-layered Earth model, although the exact frequency range in which this model is accurate is difficult to determine and depends on the particular application and region. At long periods ($T > 20$ sec and corresponding long wavelengths ($\lambda \approx 120$ km for P-waves traveling at speed of $v_P \approx 6$ km/sec), used for instance in moment-tensor inversions), basin structures, mountain ranges and other geologic features with minor changes in physical properties (density, wave-speed) can be ignored at spatial scales on the order of tens of kilometers, thus allowing accurate modeling of the complete low-frequency seismogram. For small areas with relatively simple Earth structure, body-wave Greens functions can be accurately synthesized up to frequencies of $f \approx 1$ Hz, a property that is often used in finite-source inversions. In these cases, ground-motion complexity is assumed to be largely due to the heterogeneity in the rupture process, since the local, observer-specific Greens functions (containing the impulse response in a layered medium due to point-source excitation) are simple (examples are shown in Fig. 5).

In many cases, however, sites of interest may be located on top of basin structures, in zones that are geologically different on a small-scale, or even in a narrow fault-zone belt which may have distinctly different seismic properties. By examining observed ground-motions at many sites, site-specific one-dimensional Earth models can be used to compute corresponding Greens functions, a cumbersome and error prone approach since source and receiver are located at points with different seismic properties. In such cases, 2D or 3D-models are preferable for the Greens function calculation, as they allow to appropriately include basin effects, local geology and potentially also topography, usually at the expense of reduced frequency resolution and/or a smaller computational domain.

Effects of Sedimentary Basins, Fault-Zones, Topography

Geological basins, containing layers of compliant sedimentary rock units and covered by potentially poorly consolidated, low shear-wave-velocity sediments, have a variety of effects on incoming seismic waves. Characterizing and quantifying the basin-related ground-motion complexity is particularly important for seismic hazard studies since many major urban areas are built in geological basins that are located within or close to a seismically active region (e. g. Mexico City, Tokyo, Los Angeles).

For instance, the 1985 M_w 8.0 Michoacan earthquake generated major damage in Mexico City, several hundreds of kilometers away from the epicenter. While the near-source strong-motion stations did not show unusual ground motions, the shaking level in Mexico City was unexpectedly large in those parts of the town constructed on lake bed sediments [9]. Ground-motions and associated damage exhibited high spatial variability; long-period surface waves, generated at the edge of the basin, had particularly severe consequences for tall high-rise building with eigenperiods of several seconds. Many of them suffered complete collapse. Strong basin effects have been simulated for earthquakes in the greater Los Angeles area, indicating localized basin-response effects caused by surface waves generated at the edge of the basin [74,75,135,136,138] and the focusing of seismic energy due to basin geometry [76,108,138]. A particular interesting effect has been found for the 1994 M_w 6.7 Northridge earthquake, attributed to an isolated high-damage area in Santa Monica to focusing effects caused by a small-scale localized 3D lens-like high-velocity structures [57].

Other prominent examples of strong basin response are given for the Wellington Basin (New Zealand) [24], for the Tokyo plain (Japan) [109,159,160,162], and the Kobe area (Japan), which was strongly damaged during the 1995 Kobe earthquake (M_w 6.9). In the latter case, the largest damage did not occur in direct vicinity of the rupturing fault, but was concentrated within an elongated band offset to the southeast of the fault. Modeling studies showed that this "damage belt" was caused by constructive interference of seismic waves taking different paths: (i) directly through the low-velocity basin structure and (ii) through high-velocity rocks outside the basin and then reflected back into the basin [107,143,144]. Additionally, topographic effects may have contributed to the complexity of near-fault motions for the Kobe earthquake [145]. Another example for strong topographic effects on near-source motion is the Tarzana (California) site, located atop a small hill, which recorded a value of $PGA = 1.78$ g on the EW-component of motion during the 1994 M_w 6.7 Northridge earthquake [174]. However, it is difficult to separate topographic effects from subsurface-structure properties beneath the topographic features, and firm conclusions on the contribution of topographic effects are not well defined [108].

These basin-induced or topographic site effects strongly depend on the direction from which the incoming wave field arrives, as shown by strong-motion simulations for the Basel (Switzerland) area [139] or the greater Los Angeles region [118,138]. Strong variations in the basin response, and hence ground-motions at individual sites, depend on whether the incoming waves are predominantly polarized in the direction parallel or perpendicular to the main geologic structures (i. e. the large-scale basin shape and major faults inside and bounding the basin).

Low-frequency ground-motions, modulated by basin effects, topography, or narrow fault-zone-related regions of low shear-wave velocity, exhibit also longer shaking duration (due to surface-waves arriving after the dominant S-arrivals), and potentially larger amplitudes in case of constructive interference of wave packets (often occurring as trapped waves) [141]. Due to the dependency of the basin response on the direction of the incoming wavefield and the detailed small-scale structure of the basin, individual sites within a particular basin will experience very different ground-shaking. The ground-motion complexity is thus greatly increased by these effects, and general scaling relations for ground-motion (de-)amplification and prolongation due to basin and topographic structures are difficult to derive. Only numerical simulations for particular earthquakes or specific scenario events can help to understand the corresponding ground-motion variability which can then be related to standard ground-motion attenuation relationships (see Sect. "Ground-Motion Scaling Relations").

While there has been considerable progress in interpreting and modeling strong-motion waveforms for frequencies $f < 1$ Hz, one of the major challenges is to calculate reliable broadband near-source seismograms for the frequency range of engineering interest which extends to f $\sim$ 10Hz. The works described above examine and model the low-frequency wavefield contribution, using 1D-, 2D- or 3D-finite-element, finite-difference or spectral-element techniques (see ► Seismic Wave Propagation in Media with Complex Geometries, Simulation of). Seismic source properties, Earth structure, and site effects strongly affect high-frequency motions, but computational limitations still prohibit purely deterministic ground-motion simulations for frequencies above $f \approx 1$ Hz. Instead, at least some part of the simulation procedure needs to involve a stochastic component since Earth structure, and to some extent the earthquake source, are essentially unknown at short spatial scales required for accurate high-frequency simulations. The difficulty is thus to capture and correctly quantify the scattering properties of the Earth at the scale-lengths and frequencies that are of interest for seismic hazard purposes.

Scattering in Inhomogeneous Media

The seismic coda, the energy in the seismogram after the prominent direct P- and S-wave arrivals, consist of P-, S- and surface-wave energy scattered in the inhomogeneous rock volume between the source and the recording site. Figures 1, 3, 4, display near-field waveforms for which the coda-characteristics show large variability. In some cases, the coda waves decay fast, in other cases they persist for a long time; the frequency content of coda waves also appears to be site dependent. These observations manifest that, scattering and attenuation of seismic waves in inhomogeneous media strongly contribute to ground-motion complexity.

The basic method of estimating attenuation properties of seismic waves uses the Fourier amplitude spectrum of observed ground motion $\mathbf{u}(r; f)$, for a spherical S-wave of frequency f given as

$$\mathbf{u}^S(r; f) \propto \frac{1}{r} \cdot \exp\left[-\frac{\pi \cdot f \cdot r}{Q_S \cdot \beta}\right] \tag{9}$$

where r is distance, β the S-wave velocity, and Q_S is the attenuation coefficient for S-waves, containing both intrinsic anelastic attenuation and scattering attenuation. For S-waves, Q-values in the lithosphere range from

$20 \leq Q_S \leq 10\,000$, for P-waves they are about a factor of two larger [158]. However, the frequency dependence of $Q_{S,P}$ above about $f \approx 1$ Hz indicates that attenuation increases for higher frequencies as $Q_{S,P}^{-1} \propto f^{-n}$ (with $0.5 \leq n \leq 1.0$) [161]. While the effects of seismic wave attenuation due to geometrical spreading are well understood and can be readily measured and incorporated into ground-motion simulation methods. The scattering losses of seismic waves are more difficult to estimate.

Based on seismic array measurements, recorded coda envelopes or sonic logs from borehole measurements, several studies estimated characteristic scale lengths of seismic scattering in the Earth [71,93,152,190]. Alternatively, numerical simulations for assumed random-media realizations have been used to assess scattering properties of the Earth e. g. [65,66]. These works show that the heterogeneity spectrum of velocity fluctuations in the Earth (i. e. the random variations of P- and S-wave velocities around a depth-dependent velocity gradient) can be adequately modeled as a fractal medium (e. g. Eq. (3)) or with a correlation function that does not decay too rapidly in wavenumber domain. Smoothly varying Gaussian-type correlation functions are thus inappropriate whereas an exponential or a more general van Karman correlation function Eq. (4) well models seismic scattering in the Earth crust [93,159].

Due the tectonic history of the Earth and the multiscale nature of geologic structures, scattering parameters depend on region and depth. Correlation lengths of $a \approx 10$ km are inferred for wave-scattering in the lower crust e. g. [66], whereas [86] assumed $a \approx 5$ km for upper-crustal wave-field simulations. Upper-mantle heterogeneities may be modeled with correlation length $a \approx 20$ km. The actual velocity fluctuations are on the order of 2–10%. A concise review of seismic scattering and estimated scattering parameters is given in (see ► Seismic Waves in Heterogeneous Earth, Scattering of) [161], a thorough introduction is published by Sato and Fehler [158].

Ground-motion simulations for seismic hazard assessment or studies on the nature of ground-motion complexity should incorporate seismic scattering, i. e. or at least to some degree the stochastic nature of high-frequency seismograms. In earthquake engineering random-vibration theory has been used for this purpose e. g. [35,103], which oversimplifies the earthquake rupture process but appropriately accounts for the apparent randomness of high-frequency ground-motions. More advanced techniques combine deterministic low-frequency motions with stochastic high-frequency signals [99,146,147], in which the stochastic part reflects the short-scale variability in source properties and seismic-wave scattering. The stochastic signal used in these methods is generated as random white noise and does not contain any physical scattering mechanism.

More physical approaches calculate the 3D-seismic wavefield in a 3D-heterogeneous medium, a computationally expensive task which cannot be carried out for large-scale simulations or many scenario events for which high-frequency ground-motions are needed. To devise a simplified method, [86] have combined realistic small-scale heterogeneity in the source properties with scattering operators for a von-Karman random medium to compute broadband time histories. Their study indicates that the effects of scattering are masked by the heterogeneity in the kinematic source characterization. Note also that, near-field seismic wave scattering in a heterogeneous medium may also replicate apparent nonlinear site effects [130] (see Section "Site Effects and Nonlinearity").

A computationally less demanding approach to include seismic scattering into ground-motion simulation uses radiative transfer theory to model the space-time distribution of the seismic-energy envelope due to scattered waves [192,194]. In this formulation, a time-dependent multiple S-to-S scattering process occurs due to a shear-dislocation point source embedded in a 3D-medium with background velocity v_0 in which point-like isotropic scatterers of cross section σ_0 are randomly distributed with density N. The total scattering coefficient is thus $g_0 = N \cdot \sigma_0$. The detailed theoretical developments are beyond the scope of this article, but it is worth noting that the calculation of the energy-density envelopes is computationally efficient and can be used for high-frequency ground-motion simulations [193].

The multiple S-to-S scattering theory has recently been applied to compute hybrid broadband near-field seismograms [125]. Their technique joins low-frequency 3D-finite-difference (FD) synthetics (which may contain the effects of a sedimentary basin) with site-dependent high-frequency scattering operators to synthesize broadband ground motions. The high-frequency scattering operators are convolved with an appropriate source-time function, and then these site-dependent "scatterograms" are combined in the Fourier domain with the corresponding low-frequency synthetics using a phase-matching optimization technique [124] (conceptually depicted in Fig. 12). Comparing data and simulations for a site that recorded the 1994 Northridge earthquake indicates a good agreement in terms of waveforms, amplitude spectra and spectral acceleration (Fig. 13). The model bias (logarithm of the ratio between the observed and simulated quantity) for spectral acceleration at 30 sites that recorded the 1994 Northridge event (Fig. 14a) shows only small deviations from zero, ex-

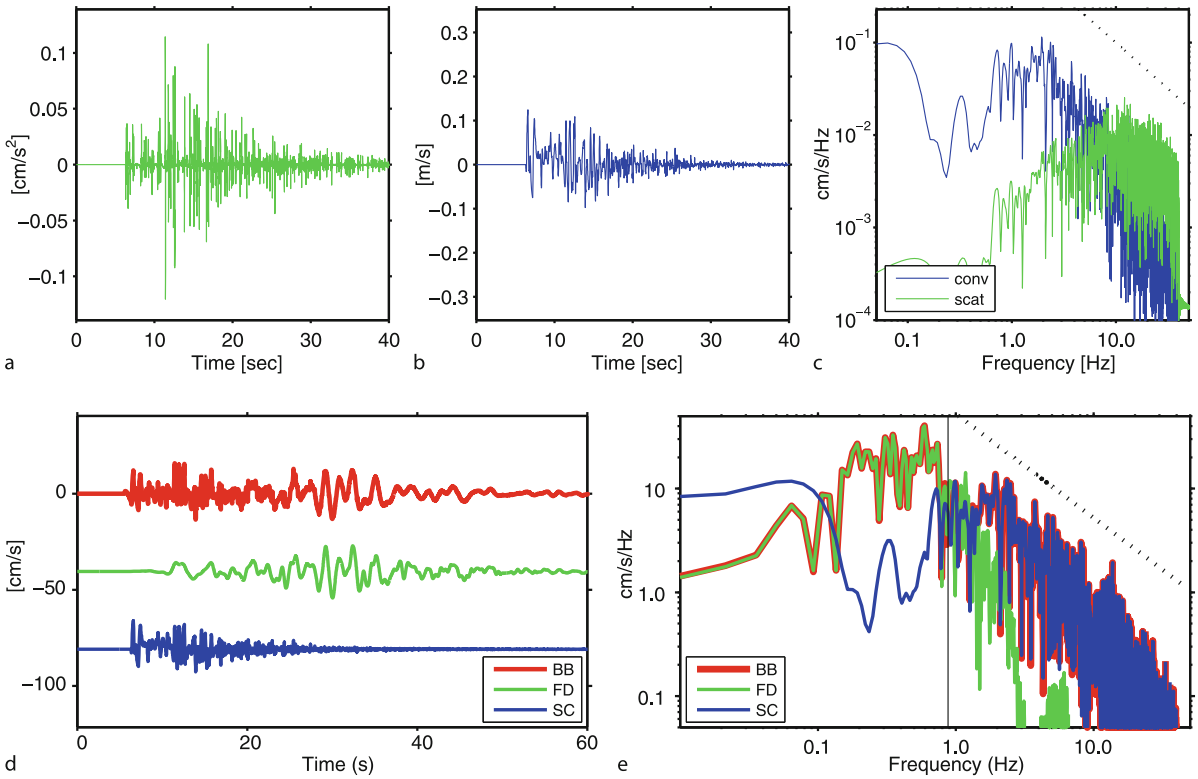

Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 12

Conceptual diagram for computing hybrid broadband seismograms. a site-specific scattering Greens function for a point-source at the hypocenter. b time series representing the local high-frequency scattering contribution (obtained for instance by convolving the scattering Greens function in a with an appropriate slip-rate function). c Fourier amplitude spectra for the time series in a and b; the velocity-scatterogram decays as $1/\omega$ (*dotted line*) beyond the corner frequency. d broadband seismogram (*top*) computed by combining the LF-seismogram (*center*) with the site-specific HF-scatterogram (*bottom*). e amplitude spectra for the time series in d; the spectra of the broad-band synthetics exactly represent the LF-motions at low-frequencies and the HF-scattering contribution at high frequencies

hibiting also narrow 90%-confidence limits. Considering also the general agreement between observed and simulated PGV and PGA values (Fig. 14b) indicates that such hybrid broadband wavefield simulations can reproduce ground-motion complexity, where differences at individual site remain due to unmodeled site and path effects.

Adding seismic scattering into ground-motion characterization and simulation helps to capture the large degree of complexity in the near-source seismic wavefield, and thus presents an important future research topic. Here I have only briefly discussed a few key elements of seismic scattering and its application to a specific case study. However, to fully explore the range of realistic scattering parameters, combined with the inherent complexity of the source-rupture process, requires extensive numerical simulations that need to be calibrated and validated against observational data.

Site Effects and Nonlinearity

Site effects play a major role in characterizing and quantifying strong ground motion as they may amplify or deamplify the incoming "bedrock" motions in the uppermost velocity layers beneath the observer. Since site-amplification factors may reach two orders of magnitude [108], these effects cannot be neglected in earthquake engineering practice. A comprehensive review of various approaches to estimate and model site effects is given by Kawase [108], so I restrict the following discussion to a few key aspects.

There is no strict boundary between site effects and path effects, but site-effects usually refer to wave-field modifications in the immediate vicinity of the observer location (top 30 m of the local sedimentary cover). Usually, the time-averaged shear-wave velocity in the top 30–100 m

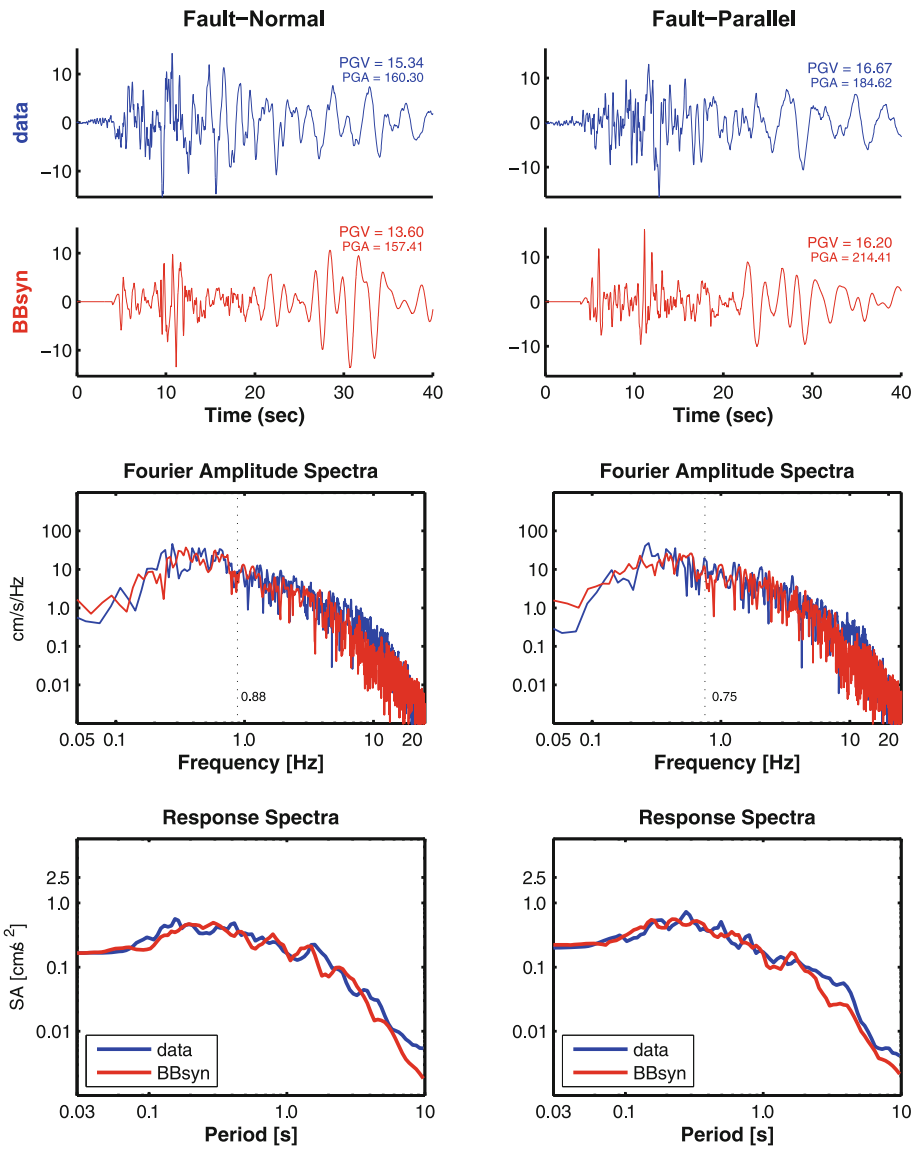

Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 13
Comparison of data (*blue*) and broadband simulations (*red*) at a site that recorded the 1994 M_w 6.7 Northridge earthquake (*left*: fault-normal; *right*: fault parallel motions). Note the consistency between the Fourier amplitude spectra (vertical line denotes the matching frequency) and between the response spectra ($\zeta = 5\%$ damping). PGA- and PGV-values are also similar

(V_{S30}) is used to define the station-specific soil classification. Site effects may be affected by the water table, and even soil-building interaction can be considered a site effect.

The classical method to estimate site effects from seismic observations is based on Eq. (1), and attempts to separate source and path from site effects. Expressing Eq. (1) in the Fourier domain and considering the ground motions due to earthquake j observed at site k, we obtain source-site specific Fourier amplitude spectra as

$$A_{jk}(f) = S_j(f) \cdot G_{jk}(f) \cdot L_k(f) \,. \tag{10}$$

The Green's function term, $G_{jk}(f)$ can be expressed in terms of geometrical spreading factors with respect to the source-site distances, r_{jk}; combining intrinsic and scattering attenuation into a common $Q(f)$ one obtains (for

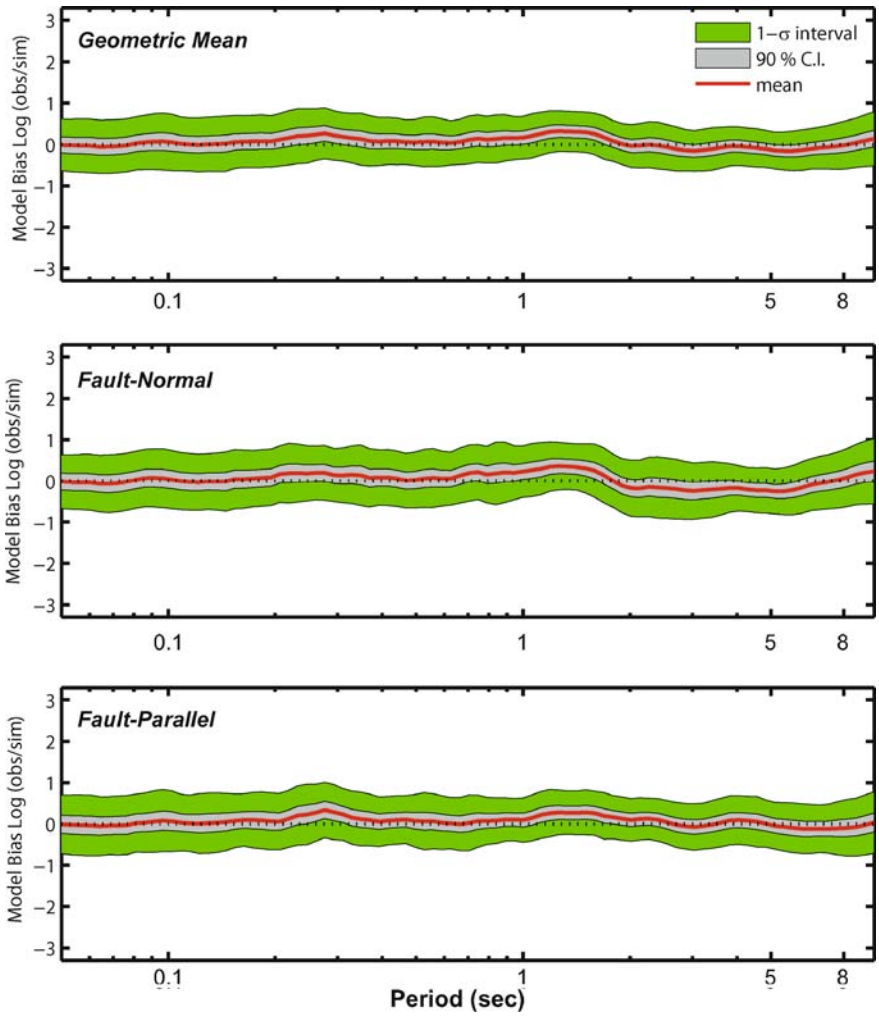

Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 14a
Hybrid broadband ground-motion calculations using low-frequency finite-difference synthetics and high-frequency scattering operators. The *three panel* show, for the geometric mean (*top*), the fault-normal (*center*), and fault-parallel (*bottom*) component, the model bias, computed as log ($\frac{\text{obs}}{\text{sim}}$), for spectral acceleration at 30 sites that recorded the 1994 Northridge earthquake

S-waves)

$$G_{jk}(f) = \frac{1}{r_{jk}} \cdot \exp\left[-\frac{\pi \cdot r_{jk}}{\beta \cdot Q(f)}\right] \tag{11}$$

where β is the representative shear-wave velocity over the entire path. Eq. (10) can then be solved by adding at least one independent constraint on either source, site or path. A common approach uses the average site factor as a reference, meaning that the logarithm of the sum of all site factors is equal to zero. Other methods select the site with the smallest site factor as reference, or make an a priori reference-site selection based on geologic or other information.

A recent method for site-effect estimation uses microtremor, ambient noise measurements, or aftershock recordings. Taking the spectral ratio of the horizontal component of motion with respect to the vertical component (so called H/V-ratios), these microtremor ratios often show similar site characteristic as inferred by independent methods [108]. However, the exact physical meaning of this H/V-ratio is not well understood, and since H/V-measurements do not directly represent the true soil amplification, those need to be supplemented and calibrated by numerical modeling [69,178]. Nonetheless, H/V-measurements are useful for microzonation studies which the detailed properties of the uppermost soil layers for a small

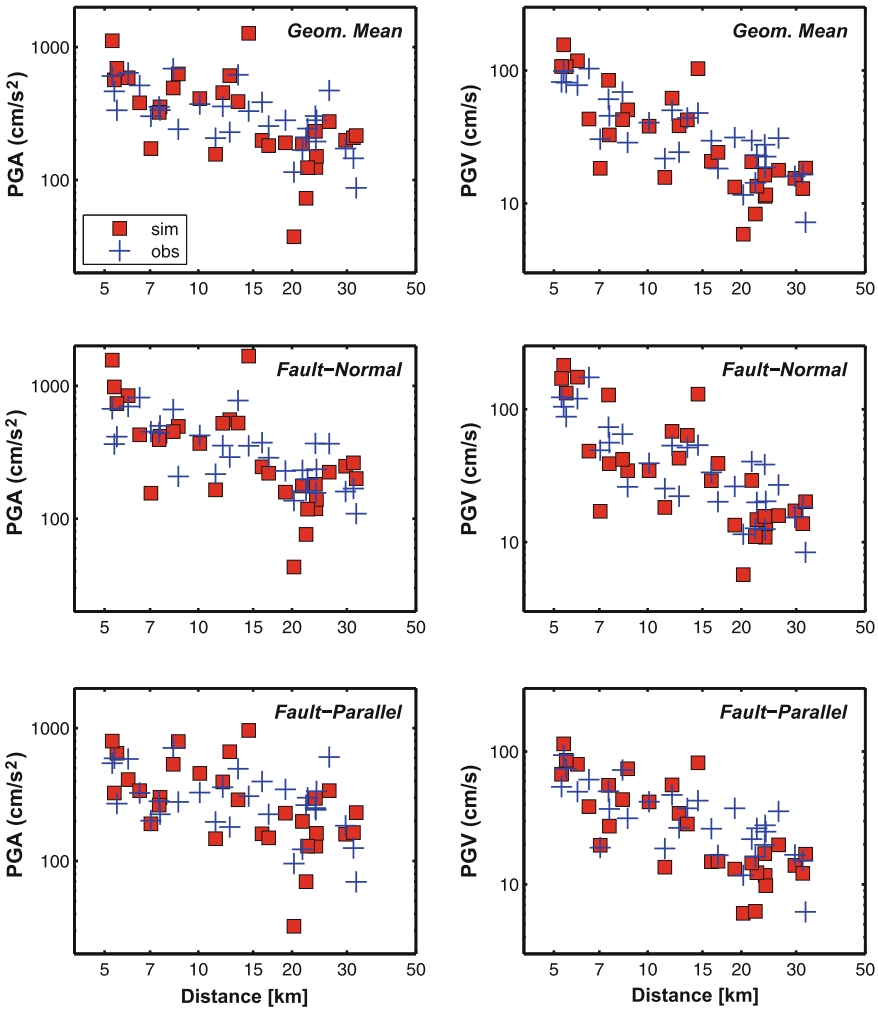

Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 14b
Comparison of observed (*blue crosses*) and simulated ((*red squares*) peak-ground acceleration (*left*) and peak-ground velocity(*right*) for 30 strong-motion sites that recorded the 1994 M_w 6.7 Northridge earthquake. Note the overall agreement between simulated and recorded motions. Discrepancies at individual stations are due to unmodeled ground-motion complexity

area, thus characterizing the susceptibility of particular locations to strong site effects [70].

Site-correction factors have been derived also from the seismic coda. Because coda waves essentially consist of scattered S-waves, local site amplification of the coda itself should be an average of the S-wave amplification factor for S-waves that arrive from different azimuths under different incidence angles. However, the assumption that the seismic coda consists only of scattered S-waves is too simplistic, because surface waves and strong P-coda waves may contaminate the S-wave coda. Site-effect estimation based only on S-coda measurements should be therefore interpreted with care [108].

A common assumption in seismic wave propagation is that strong and weak motions are affected in an identical manner. Analysis of near-field ground-motions provides evidence for nonlinear effects (i. e. a non-linear stress-strain relationship) [26,64,156], due to lower shear-wave velocity and increased damping within the sedimentary cover [54]. Local (or temporal) modification of the seismic properties of the rock material underneath a site or along the wave path may shift resonant modes to lower frequencies and generate reduced amplitudes. Nonlinearity occurs once a certain ground-motion intensity threshold is exceeded. Previous studies suggest $PGA > 0.3$ g, or $PGV > 20$ cm/sec, or peak-strains in excess of 0.06% [9],

roughly consistent with values inferred for strong-motion recordings of the 1989 M_w 6.9 Loma Prieta earthquake for which the threshold for large S-wave travel-time delays in repeating earthquakes occurs at $PGA > 0.4\,\mathrm{g}$ or $PGV > 40\,\mathrm{cm/sec}$ [156].

Effects of nonlinearity greatly increases the complexity of ground-motions, but only on a very localized scale. Recently, detailed spectrogram analyses of seismic recordings of the 2003 M_w 8:3 Tokachi-Oki (Japan) earthquake demonstrate liquefaction, quantified by a dramatically reduced high-frequency contents of the waveforms [55]. However, it is very difficult to establish general scaling laws that account for potential nonlinearity effects. In case of water-saturated loose sands, strong shaking may result in nonlinear soil liquefaction due to a rapid, temporary increase of pore water pressure, resulting in a dramatic loss of soil stiffness. Liquefaction effects are extremely important for earthquake engineering, but have received little attention from seismology in the past.

Ground-Motion Scaling Relations

Despite the fact that observed near-source ground-motions show complicated time histories and large variability in intensity measures, there is the need in earthquake engineering and seismic-hazard analysis to devise simple approaches for estimating expected ground-motions in future earthquakes. For this reason, ground-motion prediction equations (GMPE's) are developed, providing mathematical expressions that relate shaking intensities to seismological quantities, source-site geometry, and potentially site-specific parameters. Thus, the complexity of the earthquake rupture process, the wave-propagation effects from the source to the site, and the detailed site conditions are condensed into relative simple functional forms, comprising a few parameters that constitute empirical ground-motion scaling relations.

Development of Empirical Scaling Relations Based on Recordings of Past Events

Using strong-motion observations, ground-motion prediction equations are developed using a number of different parameterizations. The most fundamental form is given as [52]

$$Y = a_1 \cdot \mathrm{e}^{a_2 M} \cdot R^{a_3} \cdot \mathrm{e}^{a_4 r} \cdot \mathrm{e}^{a_5 F} \cdot \mathrm{e}^{a_6 S} \cdot \mathrm{e}^{\sigma} \tag{12}$$

where Y is the ground-motion intensity measure of interest (PGA, PGV, S_A,...). M is the earthquake's magnitude, r the source-to-site distance; F is a parameter that characterizes the type of faulting, and S captures the local site conditions. R is an additional magnitude-dependent distance function which can take on alternative forms, for example [52]

$$R = \begin{cases} r + c_7 \cdot \mathrm{e}^{c_8 M} \\ \sqrt{r^2 + \left[c_7 + \mathrm{e}^{c_8 M}\right]^2}\,. \end{cases} \tag{13}$$

This model contains some of the basic physics of earthquakes (terms with M and F), the attenuation due to geometrical spreading of seismic waves (terms related to distance r and R) and the site conditions (S). Due to the approximately log-normal distribution of ground-motion intensities, attenuation relations are parameterized using the natural logarithm of Y. Correspondingly, the standard deviation of the zero-mean random error term σ is estimated as the standard error of $\ln Y (\sigma_{lnY})$.

Over the years, different models have been proposed to capture various aspects of the inherent complexity of near-fault shaking. One class of models assumes a functional form whose shape is magnitude-independent for all distances. A simple parametric form is given by [38]

$$\begin{aligned} &\ln(Y)_{M,R,F} \\ &\quad = a_1 + a_2 \cdot M + a_3 \cdot M^2 + a_4 \cdot \ln(r + a_5) + a_6 \cdot F\,. \end{aligned} \tag{14}$$

However, many observations show that ground-motion intensities saturate at close distances to the fault, i. e. moderate-magnitude earthquakes ($5.0 \le M \le 6.5$) may generate about the same level of high-frequency shaking as large magnitude events ($M > 6.5$). The second class of ground-motion scaling relations captures these observations. For instance, starting from Eq. (14) Abrahamson and Silva [5] propose the relation

$$\begin{aligned} \ln(Y)_{M,R,F} = {} & a_1 + a_2 \cdot M + a_3 \cdot M^2 \\ & + (a_4 + f_1(M)) \cdot \ln(r + a_5) + a_6 \cdot F \end{aligned} \tag{15}$$

with $f_1(M) = a_7 M$, while Sadigh et al. [157] developed an attenuation model given by

$$\begin{aligned} &\ln(Y)_{M,R,F} \\ &\quad = a_1 + a_2 \cdot M + a_3 \cdot M^2 + a_4 \cdot \ln(r + f_2(M)) + a_6 \cdot F \end{aligned} \tag{16}$$

where $f_2(M) = c_7 \cdot \mathrm{e}^{c_8 M}$. The difference between these two scaling relations becomes evident at larger distances ($r > 50\,\mathrm{km}$), while for short distances they lead to about the same ground-motion intensities (Fig. 15).

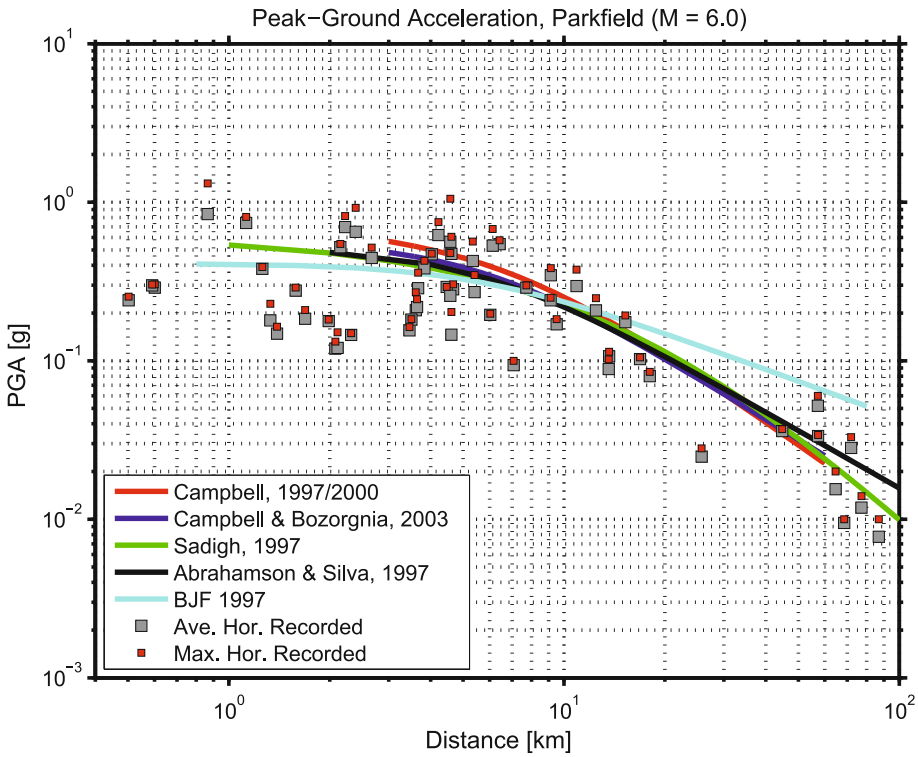

Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 15
Observed peak ground accelerations for the Sept 28, 2004 M_w 6.0 Parkfield earthquake, compared against five different empirical attenuation relationships. Data are plotted using the geometrical mean of the two horizontal components (fault-normal and fault-parallel) ($PGA = \sqrt{PGA_{FN} \cdot PGA_{FP}}$), and for the maximum of the two components. The median of the observations is generally in good agreement with the empirical predictions, but the data exhibit large variability with much higher (and lower) motions than empirically predicted, independent of the closest distance to the fault. Note that the recording at station "Fault Zone 16" is not shown as it was not available at the COSMOS database. The PGA-value for site "Fault Zone 16", located at ∼ 0.6 km distance from the fault, is estimated to be in the range 2.0–2.5 g [167]

Empirical ground-motion prediction like those presented in Eqs. (14), (15), (16) have been widely applied to global and regional ground-motion data sets, but a number of observations call for modification to these "generic" models. First of all, attenuation models for different tectonic provinces have been published accounting for variations in the regional geology [7,8,17,18,31,67,175]. While there is the need to distinguish between stable-cratonic areas (like Eastern North America), extensional regimes (like the Basin and Range province in the western US), zones of complex active tectonics (like California), or subduction zones (like in Mexico, Chile, Japan), recent work [7] suggests that data from different tectonically active areas with crustal seismicity can still be jointly analyzed. In certain regions the geologic conditions cause the presence of several of such tectonic regimes over a rather small area (e. g. Italy, Switzerland, Japan) which further complicates the development of adequate ground-motion prediction equations.

Besides the local geology, the source-site geometry greatly influences seismic shaking. In the near-fault regime, ground-motions are strongly affected by geometrical source-site effects, leading to large long-period pulses of motion termed "directivity pulses" (see Sect. "Isochrone Theory"). Recent work includes such directivity pulses in the attenuation equation [170,171]. Large permanent offset of the ground due to the rupture process will also generate long-period large motions; this so-called "fling-step" is distinguished from the directivity pulse [32]. For thrust-faulting earthquakes, the "hanging-wall" effect has been shown to generate very different ground motions [3] for sites located on the hanging-wall than on the footwall (Fig. 3). This effect is explicitly included in several recent attenuation models [3,53].

An intriguing observation has been made following the large surface-breaking earthquakes in 1999 in Turkey (the Aug. 17 M_w 7.5 Izmit and the Nov. 12 M_w 7.1 Duzce events) and Taiwan (the Sept. 20 M_w 7.6 Chi-Chi event):

near-field ground-motions were significantly lower than empirically predicted [105]. This counter-intuitive result can partly be explained by the general differences in stress-release patterns between buried-faulting and surface-rupturing earthquakes. The primary cause for the ground-motion variation between these two classes of earthquakes lies in general differences in the dynamic rupture process and the associated energy balance of the system (see ► Earthquake Scaling Laws). Ultimately, these effects could be related to depth-dependent fault-zone properties which exhibits highly damaged rocks with intense microcracking and lower seismic velocity close to the surface, but more compliant and less damaged material at depths [22]. Fault-zone structure affect the dynamics of the rupture process [13,21,22,83] and the resulting seismic radiation [141], and therefore strongly influence the resulting ground-motion intensities.

Besides the functional form and the given geological, geometrical, and physical parameters a number of technical issues affect ground-motion scaling equations (a comprehensive review is given in [60]): (i) the choice of the ground-motion intensity measure Y determines the values for the coefficients a_i (Eqs. (12)–(16)) derived by regression analysis; (ii) the specific approach for the regression (data fitting) matters (common techniques are weighted non-linear least-squares regression, two-step regression, or random-effects regression); (iii) the data selection criteria for assembling the individual strong-motion database; (iv) data corrections may be applied to the recorded strong-motions (e. g. uniform instrument response for all records, baseline correction, frequency filtering) and also to the meta-data (magnitudes, distances, site information etc.) which represent the independent parameters.

In developing ground-motion prediction equations, a uniform and well-calibrated magnitude definition is particularly important, but in practice often difficult to achieve when merging datasets from various institutions whose magnitude values may not be compatible. The style-of-faulting factor F in Eqs. (14), (15), (16) is generally well defined, but the site-effect factor S is often poorly known. One of the most critical parameters in ground-motion scaling is the distance r between the site and the source. In particular for extended rupture planes, the applied distance metric for the source-to-site geometry becomes crucial; different source-to-site distance definitions are in use [4] but the particular choice for r in turn will affect the seismic-hazard calculations [163].

In an effort to harmonize and calibrate ground-motion data, related meta-data, and the development of empirical ground-motion prediction equations, the Pacific Earthquake Engineering Research Center (PEER) carried out the New Generation Attenuation of Ground Motions (NGA) Project (completed in Jan. 2008). In this context, [37] for instance refine Eq. (14) to accommodate the effects of anelastic attenuation when modeling far-distance recordings ($R > 80$ km) and to include an "effective" magnitude-dependent geometrical spreading (allowing to predict ground-motion amplitudes out to distances $R = 400$ km). Their data-driven equation includes only terms that are truly needed to adequately fit the data, involving (i) a complicated magnitude-scaling function, $f_M(M)$; (ii) a versatile distance function, $f_D(R_{JB}, M)$; (iii) a site-effect function that includes potential nonlinearity effects, $f_S(V_{S30}, R_{JB}, M)$:

$$\ln(Y)_{M,R,F} = f_M(M) + f_D(R_{JB}, M) + f_S(V_{S30}, R_{JB}, M) + \epsilon \cdot \sigma_T \,. \tag{17}$$

All terms in Eq. (17) are period-dependent; M is moment magnitude, R_{JB} is the Joyner–Boore distance (the closest distance to the surface projection of the fault) ϵ is the fractional number of standard deviations of a single predicted value of $ln(Y)$, and $\sigma_T = \sqrt{\sigma^2_{\text{intra}} + \sigma^2_{\text{inter}}}$ describes the uncertainty by combining the intra-event and inter-event aleatory uncertainty. In this context, the exact definition for the ground-motion parameter Y, usually taken as the geometric mean of the two horizontal components, has received increased attention. Because of misaligned recording instruments or complicated (multi-)pathing of radiated waves, simple measures of "mean ground-motion" appeared to be incorrect when taking the sensor orientation as installed in the field. Works by [30,39,150] discuss the usefulness of orientation-independent ground-motion measures based on data and numerical simulations.

Relating Earthquake Source-Scaling to Ground-Motion Prediction

When using empirical ground-motion attenuation relations for predicting the expected shaking level at a given site the analyst needs quantitative information regarding the type of faults and their properties, the fault locations with respect to the site, and the site conditions. Probabilistic seismic hazard analysis (PSHA) requires additional information about the recurrence rate of earthquakes on the chosen faults [111]. Therefore, earthquake statistics (see ► Geo-complexity and Earthquake Prediction) [111] and earthquake physics (see ► Earthquake Scaling Laws) [6] are major components for accurate seismic hazard calculations. However, ground-motion prediction Eqs. (14),

(15), (16) only contain a faulting-style factor (F) and the magnitude dependence as earthquake-source related parameters. A correct estimation of the potential magnitudes for earthquakes occurring on the selected faults is the most crucial step in any seismic hazard study. Earthquake physics (see ► Earthquake Scaling Laws) [6] provides the theoretical foundation for defining these magnitudes and to explain observational data and empirical relations e. g. [106,122,187].

In modern seismic hazard analysis, the magnitude M relates to moment magnitude, M_w, derived from the seismic moment as $M_w = 2/3 \cdot \log M_0 - 6.07$. Since $M_0 = \mu \cdot L \cdot W \cdot D$, source-scaling relations between observable fault dimensions (length L, width W, area $A = L \cdot W$, average displacement D) and magnitude are used in empirical ground-motion prediction for obtaining self-consistent input parameters.

Based on geologic observations it is often possible to estimate the length of a fault, considering also cases with multiple segments that may rupture in individual events or jointly (these cases are then treated in PSHA using logic trees) [188]. Estimating fault width is more difficult; it may be inferred from the location of background seismicity in a crustal tectonic setting, or from modeling interseismic deformation which constrains the fault's locking depth, but often a generic fault width of $W = [10 - 20]$ km is assumed. Several studies have published earthquake scaling relations between various fault parameters [79,120,122,169,186,187]; these source-scaling relations are not only useful for seismic hazard analysis but also provide important insight into earthquake mechanics (see ► Earthquake Scaling Laws) [164]. A generic relation between magnitude M_w and fault area A (in km^2) is given by:

$$M_w = p + q \cdot \log_{10} A \tag{18}$$

For $q \equiv 1$, Eq. (18) is consistent with self-similar constant average-stress-drop scaling (see ► Earthquake Scaling Laws). Rule-of-thumb values for the coefficients in Eq. (18) are $p = 4$ and $q = 1$ (adapted from [187] who find $p = 3.98, q = 1.02$ for strike-slip earthquakes, and $p = 4.07, q = 0.98$ for all faulting styles), leading to a magnitude M_w 7 earthquake for a fault area $A = 1000\,\text{km}^2$. However, a number of studies have found considerable deviations from self-similar earthquake scaling [79,122,169], and reported values in the range $3.97 \le p \le 4.39$ and $0.97 \le q \le 1.33$ [120]; these differences strongly affect empirically predicted ground-motion intensities and seismic hazard calculations [188,189]. For instance, [188] proposes $p = 4.2$, $q = 1$, resulting in an M_w 7.2 earthquake for $A = 1000\,\text{km}^2$, with twice the seismic moment and twice the displacement as an M_w 7.0 rupture with identical source area. Correspondingly, Eqs. (14)–(17) lead to higher ground-motion estimates for the larger event, consequently also different seismic hazard, due to small changes in the source-scaling relation.

Quantifying Uncertainty in Ground-Motion Prediction Equations

Analyzing, quantifying and modeling strong-motion uncertainties is a mandatory part of any reliable seismic-hazard study [168]. The variability in source-scaling relations reflects in part the complexity of the earthquake rupture process, treated as an epistemic uncertainty (i. e. as more data become available and seismologists better understand earthquake dynamics, this variability will eventually decrease). The standard deviation of the residuals, obtained when deriving ground-motion prediction equations by regression analysis measures the aleatory variability (randomness) of ground-motion parameters. This standard deviation, σ_{lnY}, is then partitioned into two error terms for the intra-event and inter-event variability,

$$\sigma_{lnY} = \sqrt{\sigma_{\text{inter}}^2 + \sigma_{\text{intra}}^2}\,.$$

If the regression uses the geometric mean of the two horizontal components and the component-to-component variability is needed, a third term σ_{comp} has to be included. The intra-event variability can be further separated into a site-to-site component σ_{comp}, and the remaining variability σ_0 (after accounting for source and site effects): $\sigma_{\text{intra}} = \sqrt{\sigma_s^2 + \sigma_0^2}$ [52].

The standard deviation σ_{lnY} has previously been found to be a function of magnitude (e. g. [5,157]) with decreasing variability for increasing magnitudes. However, this counter-intuitive result could be affected by a sampling bias due to fewer near-source strong-motion recordings for moderate-to-large earthquakes. More recent work suggests that σ_{lnY} is independent of magnitude [2]. The exact value of σ_{lnY} varies between different studies and ground-motion parameters, but is generally in the range $0.4 \le \sigma_{lnY} \le 0.8$. Including this standard deviation in the ground-motion estimation is indispensable to assess the variability in shaking intensity – common practice is to use at least one or two standard deviations. However, for critical infrastructures (e. g. nuclear power plants, nuclear waste repositories), it is not clear where to truncate the ground-motion distribution (i. e. how many $\sigma's$ to include, [177]) since very long return periods need to be

considered and conclusive physical arguments for upper limits on near-field ground motions have not yet been made. Bommer et al. [33] demonstrate that the choice of the truncation level will significantly affect seismic hazard estimates at very low probabilities (annual frequencies of exceedance of 10^{-6} and lower), while the effect on hazard at probabilities traditionally used in PSHA (annual frequencies of exceedance of 10^{-4}) is small. The reason is that low-probability seismic hazard may contain contributions from rare but extreme ground-motion values.

The preceding discussion raises the issue of defining upper bounds of near-source ground motion [33]. For instance, the maximum values for *PGA* in the NGA-database is $PGA = 1.56\,\mathrm{g}$ (for the Tarzana record of the 1994 M_w 6.7 Northridge earthquake), but values of over 2 g have been reported for the 2004 Parkfield earthquake [167,168] or the 2003 Miyagi (Japan) earthquake. Note that sites and events showing the largest *PGA*-values generally do not coincide with those showing the highest *PGV*-values. For instance, the largest *PGV*-value reported in the NGA-database is 205 cm/s (observed at station TCU068 for the M_w 7.6 Chi-Chi earthquake, Fig. 4), but the *PGA* at this site was only 0.53 g. *PGV*-values in the NGA-database frequently exceed 75 cm/s, and values of 100 cm/s are not uncommon. Closer inspection of the rapidly growing online databank of near-source recordings, obtained with modern dense observational networks (e. g. the K-Net and KiK-net stations in Japan) may return even higher maximum *PGA*- and *PGV*-values than listed above.

Current research addresses the physical limits to maximum ground motions, imposed by the complexity of the source-rupture process, the wave propagation, the site conditions and the strength of the rock. Investigating the combination of these complex physical processes with numerical methods and large-scale simulations helps to better understand the overall distribution of ground-motion parameters and their complexity.

An innovative observational method to examine maximum ground-motion has been developed by Brune and colleagues [19,46,47,48,49,50], based on precariously balanced rocks. These are free-standing (large) boulders, created by erosional processes, which appear as if small ground-accelerations could overturn (topple) them; this toppling-acceleration can be measured. By also dating the age of these rocks it is possible to determine which ground-motion levels have not been exceeded in the corresponding time interval in this region. These "natural seismoscopes" therefore provide important constraints for maximum shaking levels and probabilistic seismic hazard analysis.

Comparing Observations with Empirical Predictions

The introduction of this article presents near-field ground-motion observations for two recent well recorded earthquakes, the M_w 6.0 strike-slip earthquake in Parkfield (09/28/2004) and the M_w 7.6 thrust-faulting earthquake in Taiwan (09/20/1999) (Figs. 1–4). I now compare these records against empirical predictions for a number of widely used ground-motion attenuation relations. Figures 15 and 16 display recorded *PGA* for the Parkfield event and spectral acceleration ($S_A^{T=2\,\mathrm{sec}}$) for the Chi-Chi earthquake, respectively. In both cases, the empirical predictions are generally consistent with the median of the observations, but the large scatter in the data also leads to observations occur far outside the standard 1σ-bounds of the predictions.

For the Parkfield data, observed *PGA*-values are both higher and lower than empirically predicted (for $r < 100\,\mathrm{km}$), and no obvious pattern is visible that may explain the large ground-motion variability (Fig. 15). Peak ground acceleration close to and above 1 g were observed at a number of stations. At one site (station "Fault Zone 16"), the *PGA*-value is estimated to be on the order of 2.0–2.5 g [167]. The plotted *PGA*-values are not corrected or grouped according to the soil classification for each site, i. e. parts of the observed variability can be attributed to site effects. However, much of the variability originates from the particular position of each station with respect to the spatial complexity of the earthquake rupture process, and, to a lesser extent, from localized wave-propagation phenomena.

Examining recorded motions and empirical predictions for the Chi-Chi earthquake reveals an overall consistency between the observed median spectral acceleration ($S_A^{T=2\,\mathrm{sec}}$) and ground-motion attenuation relations, despite significant scatter in the data(Fig. 16). At large distances ($R_{JB} \geq 50\,\mathrm{km}$) many of the recorded motions are significantly larger than any of the empirical prediction, while at distances close to the fault ($R_{JB} \leq 10\,\mathrm{km}$) most of the S_A-values are significantly lower than predicted. This unexpected behavior of lower than-predicted ground-motions for large surface-breaking earthquakes has been recently observed for a number of earthquakes [105], and is attributed to effects of earthquake source dynamics occurring in the uppermost shallow and more compliant crustal layers [126].

These two examples illustrate the large variability of near-source motions recorded at a number of stations for the same earthquake, thus characterizing the intra-event variability described above. Site, path, and source effects

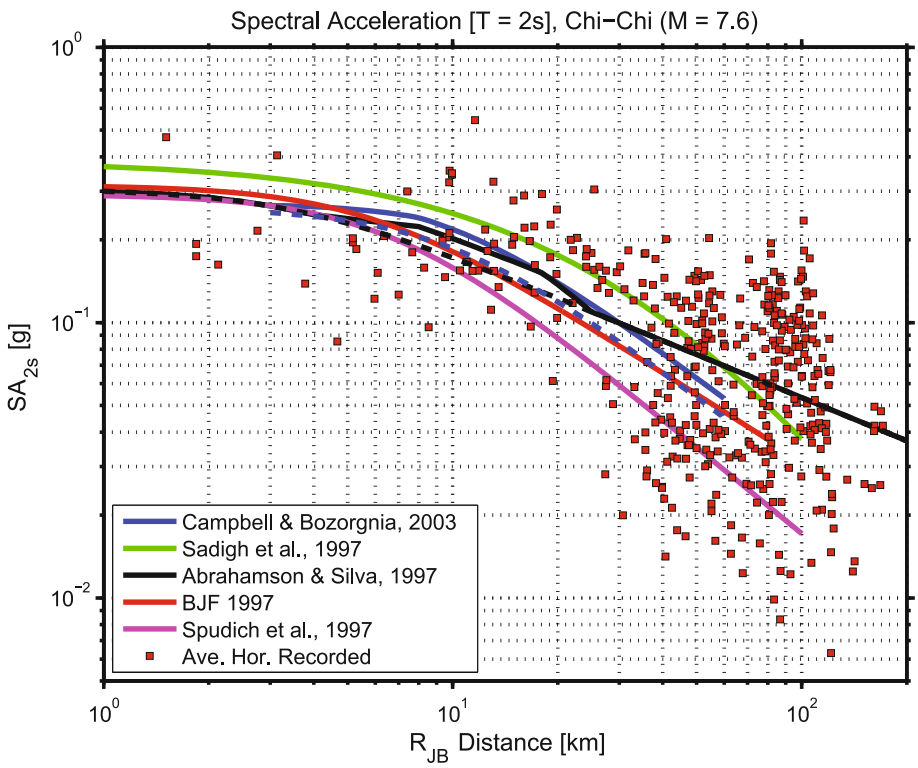

Ground Motion: Complexity and Scaling in the Near Field of Earthquake Ruptures, Figure 16
Spectral accelerations (*SA*) at $T = 2$ s for ground-motion recordings of the Sept 20, 2004 M_w 7.6 Chi-Chi (Taiwan) earthquake, compared against five different empirical attenuation relationships. Data are plotted using the geometrical mean of the two horizontal components (fault-normal and fault-parallel) ($SA = \sqrt{SA_{FN} \cdot SA_{FP}}$). The Campbell & Bozorgnia [53] and Abrahamson & Silva [5] relations contain a hanging-wall factor, indicated by the *solid blue* and *black lines*, respectively; the corresponding regular relations are plotted with *dashed lines*. The median of the observations is consistent the empirical predictions, but the data exhibit large variability, with much higher (and lower) motions than empirically predicted. Note that at large distances ($R_{JB} \geq 50$ km) many sites reveal higher-than-predicted spectral accelerations while at very short distances ($R_{JB} \leq 10$ km) many sites exhibit significantly lower ground-motions

are responsible for this large degree of ground-motion complexity, but separating the contributions of each of these effects is difficult. Deciphering the detailed physical processes and corresponding parameters that lead to a specific near-field ground-motion is an active area of research. Ultimately, this will allow us to better model ground-motions and their variability for future earthquakes, a key ingredient for improved seismic hazard analysis.

Future Directions

This article reviews the complexity of near-field ground motions, generated by the space-time-dependent heterogeneous earthquake rupture process, transformed by wave-propagation through complex geologic structure and inhomogeneous media, and finally subjected to localized site conditions. As a consequence, the waveform complexity recorded at dense strong-motion arrays provides both a challenge and an opportunity for future research.

A largely overlooked aspect of ground-motion complexity pertains to the rotational motions (also called vorticity) of the displacement field. Translational ground displacements (velocity, acceleration) are recorded as seismograms used for monitoring seismic activity and ground motions; strain measurements capture the deformation of the Earth, but the theoretically predicted [6] ground rotation (a vectorial quantity) has rarely been reliably measured in the past. Observations of rotational motions are challenging, due to their small amplitudes [42] and inadequate instrument sensitivity, but recent work provides evidence that rotational motions are significant [95,97,98,179]. Moreover, they are particularly important for engineering applications [182,183] (see ► Earthquake Source: Asymmetry and Rotation Effects for

a comprehensive review on rotational motions). As improved sensors are developed to measure rotational motions, they may become a new observable for (engineering) seismology, potentially adding complementary information on earthquake source processes, Earth structure, and ground shaking [98]. For deciphering ground-motion complexity, rotational seismology constitutes a new emerging research field; consequently, the International Working Group on Rotational Seismology (IWGoRS) was formed in 2006 to foster the exchange of ideas, data, and software on rotational seismology.

As strong-motion arrays become more abundant and better equipped, the recorded near-source seismograms provide a wealth of information on the rupture process, the wave-propagation phenomena and the site structure. Harvesting these data to learn more about these complex physical processes will be a research focus for years to come. For example, current earthquake source inversions generally use low-pass filtered seismograms ($f \leq 1\,\mathrm{Hz}$) because of incomplete knowledge of Earth structure at shorter wavelengths and the increasing ill-conditioning of the inverse problem as higher frequencies are included. This limitation does not allow proper imaging of high-frequency radiation on the fault, excited at small-scale geometrical complexities or by sudden changes in the dynamic rupture process. Those higher frequencies, however, are particularly damaging to much of the built environment and critical for reliable seismic hazard studies.

Currently, seismologists either estimate earthquake source properties and assume an Earth model, or simplify the source to a known mechanism to understand Earth structure. Future work will increasingly consider earthquake source modeling and imaging of Earth structure as a coupled (joint) inverse problem, potentially including shorter wavelength and/or deriving appropriate stochastic media characterizations. With the advent of high-performance computing facilities and innovative numerical methods (see ► Seismic Wave Propagation in Media with Complex Geometries, Simulation of) [61,112,149], multi-level optimization strategies will be developed to solve the highly non-linear inverse problem of inferring earthquake rupture dynamics from radiated seismic waves that propagate through heterogeneous media.

Empirical ground-motion equations remain important in engineering practice, but accurately quantifying ground-shaking variability for GMPE's continues to be difficult, despite an increasing number of near-source recordings. It is likely that each future large earthquake, recorded by a dense network, may generate unusual ground-motions outside the commonly assumed standard-deviation of current empirical predictions. Precise and reliable ground-motion estimation is the key challenge for seismologists and earthquake engineers for proper seismic hazard assessment and earthquake loss mitigation in future events. This task requires the ability to model not only median values of ground-motion intensity, but also to accurately capture their variability. Moreover, it will become increasingly important to not only model scalar ground-motion intensities, but to compute large suites of realistic synthetic near-field seismograms that reproduce the observed ground-motion complexity. Thus, innovative source-modeling approaches are required to capture the complexity of the dynamic rupture process (e. g. [77,120]). Multi-scale dynamic simulations with initial conditions based on stochastic parameter distributions [133,151], need to be coupled to large-scale broadband wave-propagation computations (e. g. [138,139]) for many realizations of heterogeneous Earth models (perhaps with stochastic properties at short wave-lengths). Such simulation-based ground-motion prediction is needed to advance the current practice in seismic hazard assessment.

Acknowledgments

I am indebted to J. Bühler for generating several figures for this article. J. Ripperger provided simulation data for Fig. 5. Thanks to P. Spudich for computing isochrone quantities and synthetics shown in Fig. 11. Strong-motion data were taken from the COSMOS strong-motion database (http://db.cosmos-eq.org). Critical comments and helpful suggestions by J. Clinton, G. Cua, and S. Jonsson greatly improved the article. I am also grateful for constructive reviews by R. Harris and Y. Ben-Zion. Parts of this work was supported by the Southern California Earthquake Center; SCEC is funded by NSF Cooperative Agreement EAR-0106924 and USGS Cooperative Agreement 02HQAG0008. This is SCEC contribution number 1154.

Bibliography

Primary Literature

1. Aagaard B, Heaton TH (2004) Near-source ground motions from simulations of sustained intersonic and supersonic fault ruptures. Bull Seis Soc Am 94(6):2064–2078
2. Abrahamson NA, Silva WJ (2005) Preliminary results of the A&S 2005 attenuation relation, presented at the USGS workshop on Attenuation Relations to be used in the National USGS Seismic Hazard Maps, Menlo Park CA, October 24
3. Abrahamson NA, Somerville PG (1996) Effects of the hanging-wall and foot-wall on ground motions recorded during the Northridge earthquake. Bull Seis Soc Am 86:93–99

4. Abrahmson NA, Shedlock KM (1997) Overview. Seis Res Lett 68(1):9–23
5. Abrahmson NA, Silva W (1997) Empirical response spectral attenuation relations for shallow crustal earthquakes. Seis Res Lett 68(1):94–127
6. Aki K, Richards PG (2002) Quantitative Seismology. University Science Books, Sausalito
7. Akkar S, Bommer JJ (2007) Empirical prediction equations for peak ground velocity derived from strong-motion records from Europe and the Middle East. Bull Seis Soc Am 97(2):511–530
8. Ambraseys NN (1995) The prediction of earthquake peak ground acceleration in Europe. Earthq Eng Struct Dyn 24(4):467–490
9. Anderson JG (2002) Strong-motion seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part B. Academic Press, San Diego, 937–965
10. Andrews DJ (1976) Rupture propagation with finite stress in antiplane strain. J Geophys Res 81(20):3575–3582
11. Andrews DJ (1980) A stochastic fault model; 1. Static case. J Geophys Res 85(B7):3867–3877
12. Andrews DJ (1981) A stochastic fault model; 2. Time-dependent case. J Geophys Res 86(11):10821–10834
13. Andrews DJ, Ben-Zion Y (1997) Wrinkle-like slip pulse on a fault between different materials. J Geophys Res 102(1): 552–571
14. Aochi H, Fukuyama E (2002) Three-dimensional non-planar simulation of the 1992 Landers earthquake. J Geophys Res 107(B2):art 2035. doi:10.1029/2000JB000061
15. Aochi H, Madariaga R (2003) The 1999 Izmit, Turkey, earthquake: Non-planar fault structure, dynamic rupture process, and strong ground motion. Bull Seis Soc Am 93(3):1249–1266
16. Archuleta RJ (1984) A faulting model for the 1979 Imperial Valley earthquake. J Geophys Res 89(6):4559–4585
17. Atkinson G, Boore DM (1995) New ground motion relations for eastern North America. Bull Seis Soc Am 85:17–30
18. Bay F, Fäh D, Malagnini L, Giardini D (2003) Spectral shear-wave ground-motion scaling in Switzerland. Bull Seis Soc Am 93(1):414–429
19. Bell JW, Brune JN, Zeng Y (2004) Methodology for obtaining constraints of ground motion from precariously balanced rocks. Bull Seis Soc Am 94:285–303
20. Ben-Zion Y (2003) Appendix 2, Key Formulas in Earthquake Seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part B. Academic Press, San Diego, pp 1857–1875
21. Ben-Zion Y, Andrews DJ (1998) Properties and implications of dynamic rupture along a material interface. Bull Seis Soc Am 88(4):1085–1094
22. Ben-Zion Y, Sammis CG (2003) Characterization of Fault Zones. Pure Appl Geophys 160:677–715
23. Ben-Zion Y, Zhu L (2002) Potency-magnitude scaling relations for southern California earthquakes with $1.0 < M_L < 7.0$. J Geophys Int 148:F1–F5
24. Benites R, Olsen KB (2005) Modeling strong ground motion in the Wellington Metropolitan Area, New Zealand. Bull Seis Soc Am 95(6):2180–2196
25. Beresnev IA (2003) Uncertainties in finite-fault slip inversions: To what extent to believe? (A critical review). Bull Seis Soc Am 93(6):2445–2458
26. Beresnev IA, Wen KL (1996) Nonlinear soil response – A reality. Bull Seis Soc Am 86:1964–1978
27. Bernard P, Madariaga R (1984) A new asymptotic method for the modeling of near field accelerograms. Bull Seis Soc Am 74:539–558
28. Beroza GC, Mikumo T (1996) Short slip duration in dynamic rupture in the presence of heterogeneous fault properties. J Geophys Res 101(10):22,449–22,460
29. Beroza GC, Spudich P (1988) Linearized Inversion for Fault Rupture Behavior. Application- to the 1984 Morgan-Hill, California, Earthquake. J Geophys Res 93(B6):6275–6296
30. Beyer K, Bommer JJ (2006) Relationships between median values and between aleatory variabilities for different definitions of the horizontal component of motion. Bull Seis Soc Am 96(4):1512–1522
31. Bindi D, Parolai S, Grosser H, Milkereit C, Durukal E (2007) Empirical ground-motion prediction equations for northwestern Turkey using the aftershocks of the 1999 Kocaeli earthquake. Geophys Res Lett 34:L08305. doi:10.1029/2007GL029222
32. Bolt BA, Abrahamson NA (2002) Estimation of strong seismic ground motions. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part B. Academic Press, San Diego, 983–1001
33. Bommer JJ, Abrahamson NA, Strasser FO, Pecker A, Bard P-Y, Bungum H, Cotton F, Fäh D, Sabetta F, Scherbaum F, Studer J (2004) The challenge of defining upper bounds on earthquake ground motions. Seis Res Lett 75(1):82–95
34. Bommer JJ, Martinez-Pereira A (1998) The effective duration of earthquake strong motion. J Earthq Eng 3(2):127–172
35. Boore DM (1983) Stochastic simulation of high frequency ground motions based on seismological models of the radiation spectra. Bull Seismol Soc Am 73:1865–1894
36. Boore DM (2001) Comparisons of ground motions from the 1999 Chi-Chi earthquake with empirical predictions largely based on data from California. Bull Seis Soc Am 91(5):1212–1217
37. Boore DM, Atkinson GM (2008) Updated Reports of Next Generation Attenuation (NGA) Models, PEER Lifelines Program, report and auxiliary material available at http://peer.berkeley.edu/products/Boore-Atkinson-NGA.html
38. Boore DM, Joyner WB et al (1997) Equations for estimating horizontal response spectra and peak acceleration from western North American earthquakes; a summary of recent work. Seis Res Lett 68 (1):128–153
39. Boore DM, Lamprey JW, Abrahamson NA (2006) Orientation-independent measures of ground motion. Bull Seism Soc Am 96(4A):1502–1511. doi:10.1785/0120050209
40. Borcherdt RD (1994) Estimates of site-dependent response spectra for design (methodology and justification). Earthquake Spectra 10(4):617–653
41. Borcherdt RD (2002) Empirical evidence for acceleration-dependent amplification factors. Bull Seis Soc Am 92(2):761–782
42. Bouchon M, Aki K (1982) Strain, tilt, and rotation associated with strong ground motion in the vicinity of earthquake faults. Bull Seismol Soc Am 72:1717–1738
43. Bouchon M, Streiff D (1997) Propagation of a shear crack on a nonplanar fault: A method of calculation. Bull Seis Soc Am 87(1):61–66

44. Bouchon M, Vallee M (2003) Observation of long supershear rupture during the magnitude 8.1 Kunlunshan earthquake. Science 301(5634):824–826
45. Brune JN (1970) Tectonic Stress and Spectra of Seismic Shear Waves from Earthquakes. J Geophys Res 75(26):4997–5009
46. Brune JN (1996) Precariously balanced rocks and ground-motion maps for southern California. Bull Seis Soc Am 86(1): 43–54
47. Brune JN (1999) Precarious rocks along the Mojave section of the San Andreas Fault, California: Constraints on ground motion from great earthquakes. Seis Res Lett 70:29–33
48. Brune JN (2003) Precarious rock evidence for low near-source accelerations for trans-tensional strike-slip earthquakes. Physics of the Earth and Planetary Interiors 137(1–4):229–239
49. Brune JN, Anooshehpoor A, Purvance MD (2006) Band of precariously balanced bocks between the Elsinore and San Jacinto, California, fault zones: Constraints on ground motion for large earthquakes. Geology 34:137–140
50. Brune JN, Whitney JW (1992) Precariously balanced rocks with rock varnish–Paleoindicators of maximum ground acceleration? Seis Res Lett 63:21
51. Cakir Z, de Chabalier JB, Armijo R, Meyer B, Barka A, Peltzer G (2003) Coseismic and early post-seismic slip associated with the 1999 Izmit earthquake (Turkey), from SAR interferometry and tectonic field observations. Geophys J Int 155 (1):93–110
52. Campbell KW (2002) Strong-motion attenuation relations. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part B. Academic Press, San Diego, pp 1003–1012
53. Campbell KW, Bozorgnia Y (2003) Updated near-source ground-motion (attenuation) relations for the horizontal and vertical components of peak ground acceleration and acceleration response spectra. Bull Seis Soc Am 93(3):1413–1413
54. Choi Y, Stewart JP (2005) Nonlinear site amplification as function of 30 m shear-wave velocity. Earthq Spect 21(1):1–30
55. Clinton JF (2004) Modern digital seismology-instrumentation, and small amplitude studies in the engineering world. http://resolver.caltech.edu/CaltechETD:etd-05202004-225044
56. Custodio S, Lui P-C, Archuleta RJ (2005) The 2004 M_w 6.0 Parkfield, California, earthquake: inversion of near-source ground motion using multiple data sets. Geophys Res Lett 32:L23312
57. Davis P, Rubinstein JL, Liu K, Gao S, Knopoff L (2000) Northridge earthquake damage caused by geologic focusing of seismic waves. Science 289(5485):1746–1750
58. Day SM (1982a) Three-dimensional finite difference simulation of fault dynamics; rectangular faults with fixed rupture velocity. Bull Seis Soc Am 72(3):705–727
59. Day SM (1982b) 3-Dimensional Simulation of Spontaneous Rupture – the Effect of Non-uniform Prestress. Bull Seis Soc Am 72(6):1881–1902
60. Douglas J (2003) Earthquake ground motion estimation using strong-motion records: a review of equations for the estimation of peak ground acceleration and response spectral ordinates. Earth-Science Reviews 61:43–104
61. Dumbser M, Käser M (2006) An arbitrary high-order discontinuous Galerkin method for elastic waves on unstructured meshes II– The three dimensional case. Geophys J Int 167:319–336
62. Dunham E, Archuleta RJ (2004) Evidence for a super-shear transient during the 2002 Denali earthquake. Bull Seis Soc Am 94:S256–S268
63. Dunham E, Favreau P, Carlson J (2003) A supershear transition mechanism for cracks. Science 299:1557–1559
64. Field EH, Zeng YH, Johnson PA, Beresnev I (1998) Nonlinear sediment response during the 1994 Northridge earthquake: Observations and finite source simulations. J Geophys Res 103(B11):26869–26883
65. Frankel A, Clayton RW (1984) A finite-difference simulation of wave propagation in two-dimensional random media. Bull Seis Soc Am 74(6):2167–2186
66. Frankel A, Clayton RW (1986) Finite-Difference Simulations of Seismic Scattering Implications– for the Propagation of Short-Period Seismic-Waves in the Crust and Models of Crustal Heterogeneity. J Geophys Res 91(B6):6465–6489
67. Fukushima Y, Tanaka T (1990) A new attenuation relation for peak horizontal acceleration of strong earthquake ground motion in Japan. Bull Seis Soc Am 80(4):757–783
68. Fukuyama E, Olsen KB (2002) A condition for super-shear rupture propagation in a heterogeneous stress field. Pure Appl Geophys 159(9):2047–2056
69. Fäh D, Kind F, Giardini D (2001) A theoretical investigation of average H/V ratios. Geophys J Int 145(2):535–549
70. Fäh D, Rüttener E, Noack T, Kruspan P (1997) Microzonation of the city of Basel. J of Seism 1:87–102
71. Goff JA, Holliger K (1999) Nature and origin of upper crustal seismic velocity fluctuations and associated scaling properties: Combined stochastic analyses of KTB velocity and lithology logs. J Geophys Res 104(B6):13169–13182
72. Graves RW (1996) Simulating seismic wave propagation in 3D elastic media using staggered-grid finite differences. Bull Seis Soc Am 86(4):1091–1106
73. Graves RW (1998) Three-dimensional finite-difference modeling of the San Andreas fault: source parameterization and ground-motion levels. Bull Seis Soc Am 88(4):881–897
74. Graves RW, Clayton RW (1992) Modeling path effects in 3-dimensional basin structures. Bull Seis Soc Am 82(1):81–103
75. Graves RW, Pitarka A, Somerville P (1998) Ground-motion amplification in the Santa Monica area: Effects of shallow basin-edge structure. Bull Seis Soc Am 88(5):1224–1242
76. Graves RW, Wald DJ (2004) Observed and simulated ground motions in the San Bernardino basin region for the Hector Mine, California, earthquake. Bull Seis Soc Am 94(1):131–146
77. Guatteri M, Mai PM, Beroza GC (2004) A pseudo-dynamic approximation to dynamic rupture models for strong ground motion prediction. Bull Seis Soc Am 94(6):2051–2063
78. Guatteri M, Mai PM, Beroza GC, Boatwright J (2003) Strong-ground motion prediction from stochastic-dynamic source models. Bull Seis Soc Am 93(1):301–313
79. Hanks TC, Bakun WH (2002) A bilinear source-scaling model for M-log A observations of continental earthquakes. Bull Seis Soc Am 92(5):1841–1846
80. Harris RA (2004) Numerical simulations of large earthquakes: dynamic rupture propagation on heterogeneous faults. Pure Appl Geophys 161:2171–2181
81. Harris RA, Archuleta RJ et al (1991) Fault steps and the dynamic rupture process; 2-D numerical simulations of a spontaneously propagating shear fracture. Geophys Res Lett 18(5):893–896
82. Harris RA, Day SM (1993) Dynamics of fault interaction; parallel strike-slip faults. J Geophys Res 98(3):4461–4472

83. Harris RA, Day SM (1997) Effects of a low-velocity zone on a dynamic rupture. Bull Seis Soc Am 87(5):1267–1280
84. Harris RA, Day SM (1999) Dynamic 3D simulations of earthquakes on en echelon faults. Geophys Res Lett 26(14):2089–2092
85. Hartzell S, Guatteri M, Mai PM, Liu P-C, Fiske M (2005) Calculation of broadband time histories of ground motion: Part II, Kinematic and dynamic modeling with theoretical Green's functions. Bull Seis Soc Am 95(2):614–645
86. Hartzell S, Liu P-C, Mendoza C, Ji C, Larson K (2007) Stability and uncertainty of finite-fault slip inversions: Application to the 2004 Parkfield, California, earthquake. Bull Seis Soc Am 97:1911–1934
87. Heaton TH (1982) The 1971 San-Fernando Earthquake – a Double Event? Bull Seis Soc Am 72(6):2037–2062
88. Heaton TH (1990) Evidence for and implications of self-healing pulses of slip in earthquake rupture. Phys Earth Planet Int 64:1–20
89. Hikima K, Koketsu K (2004) Source processes of the foreshock, mainshock and largest aftershock in the 2003 Miyagi-ken Hokubu, Japan, earthquake sequence. Earth Planets Space 56:87–93
90. Hillers G, Mai PM, Ampuero J-P, Ben-Zion Y (2007) Statistical properties of seismicity of fault zones at different evolutionary stages. Geophys J Int 169:515–533. doi:10.1111/j.1365-246X2006.03275.x
91. Hillers G, Wesnousky S (2008) Scaling relation of strike-slip earthquakes with different rate-state-dependendent properties at depth. Bull Seis Soc Am 98(3):1085–1101
92. Hisada Y (2001) A theoretical omega-square model considering spatial variation in slip and rupture velocity. Part 2: Case for a two-dimensional source model. Bull Seis Soc Am 91(4):651–666
93. Holliger K (1997) Seismic scattering in the upper crystalline crust based on evidence from sonic logs. Geophys J Int 128(1):65–72
94. Horikawa H (2001) Earthquake doublet in Kagoshima, Japan: Rupture of asperities in a stress shadow. Bull Seis Soc Am 91(1):112–127
95. Huang BS (2003) Ground rotational motions of the 1991 Chi-Chi, Taiwan earthquake as inferred from dense array observations. Geophys Res Lett 30(6):1307–1310
96. Ide S (2001) Complex source processes and the interaction of moderate earthquakes during the earthquake swarm in the Hida-Mountains, Japan, 1998. Tectonophysics 334(1):35–54
97. Igel H, Cochard A, Wassermann J, Schreiber U, Velikoseltsev A, Dinh NP (2007) Broadband observations of rotational ground motions. Geophys J Int 168(1):182–197
98. Igel H, Schreiber U, Flaws A, Schuberth B, Velikoseltsev A, Cochard A (2005) Rotational motions induced by the M8.1 Tokachi-oki earthquake, September 25, 2003. Geophys Res Lett 32:L08309. doi:10.1029/2004GL022336
99. Irikura K, Kamae K (1994) Estimation of strong ground motion in broad-frequency band based on a seismic source scaling model and an empirical Green's function technique. Annali Di Geofisica 37:1721–1743
100. Iwata T, Sekiguchi H (2002) Source process of the 2000 western Tottori Prefecture earthquake and near-source strong ground motion. Proceedings of the Japan. Earthquake Eng. Symposium, vol 11. pp 125–128
101. Johnson KM, Hsu YJ, Segall P, Yu SB (2001) Fault geometry and slip distribution of the 1999 Chi-Chi, Taiwan earthquake imaged from inversion of GPS data. Geophys Res Lett 28(11):2285–2288
102. Jonsson S, Zebker H, Segall P, Amelung F (2002) Fault slip distribution of the 1999 M-w 7.1 Hector Mine, California, earthquake, estimated from satellite radar and GPS measurements. Bull Seis Soc Am 92(4):1377-1389
103. Joyner WB, Boore DM (1988) Measurement, characterization, and prediction of strong ground motion. Conference on Earthquake Engineering and Soil Dynamics II: Recent Advances in Ground Motion Evaluation, Park ASCE City UT, USA, p 43–102
104. Joyner WB, Spudich P (1994) Including near-field terms in the isochrones integration method for application to finite-fault or Kirchhoff boundary integral problems. Bull Seis Soc Am 84:1260–1265
105. Kagawa T, Irikura K, Somerville PG (2004) Differences in ground motion and fault rupture process between surface and buried rupture earthquakes. Earth Planets Space 56(1):3–14
106. Kanamori H, Anderson DL (1975) Theoretical basis of some empirical relations in seismology. Bull Seis Soc Am 65(5):1073–1095
107. Kawase H (1996) The cause of the damage belt in Kobe: The basin edge effect – constructive interference of the direct S-wave with the basin-induced diffracted Rayleigh waves. Seis Res Lett 67(5):25–34
108. Kawase H (2003) Site effects on strong ground motions In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part B. Academic Press, San Diego, pp 1013–1030
109. Koketsu K, Kikuchi M (2000) Propagation of seismic ground motion in the Kanto basin, Japan. Science 288:1237–1239
110. Kostrov BV (1964) Self-similar problems of propagating shear cracks. App J Math Mech 28:1077–1078
111. Kramer SL (1996) Geotechnical Earthquake Engineering. Prentice Hall, Upper Saddle River, New Jersey
112. Käser M, Mai PM, Dumbser M (2007) Geometrically complex finite source rupture models and their accurate treatment using tetrahedral meshes. Bull Seis Soc Am 97(5):1570–1586
113. Lavallee D, Archuleta R (2003) Stochastic modeling of slip spatial complexities for the 1979 Imperial Valley, California, earthquake. Geophys Res Lett 30:1245. doi:10.1029/2002GL015839
114. Lavallee D, Liu PC, Archuleta R (2006) Stochastic model of heterogeneity in earthquake slip spatial distributions. Geophys J Int 165(2):622–640
115. Lee WHK, Shin T-C, Kuo KW, Chen KC, Wu CF (2001) CWB free-field strong-motion data from the 21 September Chi-Chi, Taiwan, earthquake. Bull Seis Soc Am 91(5):1370–1376
116. Liu P, Archuleta RJ (2004) A new nonlinear finite fault inversion with three-dimensional Green's functions: Application to the 1989 Loma Prieta, California, earthquake. J Geophys Res 109:B02318
117. Liu P-C, Custodio S, Archuleta RJ (2006) Kinematic inversion of the 2004 M 6.0 Parkfield earthquake including an approximation to site effects. Bull Seis Soc Am 96:143–158
118. Ma S, Archuleta RJ, Page MT (2007) Effects of large-scale surface topography on ground motions, as demonstrated by a study of the San Gabriel Mountains, Los Angeles, California. Bull Seis Soc Am 97(6):2066–2079

119. Madariaga R (1983) High frequency radiation from dynamic earthquake fault models. Ann Geophys 1(1):17–23
120. Mai PM (2001) Characterizing earthquake source complexity for improved strong motion prediction. Ph D thesis, Department of Geophysics, Stanford University, California
121. Mai PM (2004) SRCMOD: an online database of finite-source rupture models. http://www.seismo.ethz.ch/srcmod, updated July 2007; last accessed Aug 2007
122. Mai PM, Beroza GC (2000) Source-scaling properties from finite-fault rupture models. Bull Seis Soc Am 90(3):604–615
123. Mai PM, Beroza GC (2002) A spatial random-field model to characterize complexity in earthquake slip. J Geophys Res 107(B11):2308. doi:10.1029/2001JB000588
124. Mai PM, Beroza GC (2003) A hybrid method for calculating near-source broadband seismograms: application to strong motion prediction. Phys Earth Planet Int 137:183–199
125. Mai PM, Olsen KB (2005) Broadband ground motion simulations using finite-difference synthetics with local scattering operators. 2005 Annual SCEC Meeting, Palm Springs CA, Southern California Earthquake Center
126. Mai PM, Somerville P, Pitarka A, Dalguer L, Miyake H, Beroza G, Song S-G, Irikura K (2006) Fracture-energy scaling in dynamic rupture models of past earthquakes. Earthquakes: Radiated Energy and the Physics of Faulting Geophysical Monograph Series, vol. 170. American Geophysical Union, 10.1029/170GM28:283–294
127. Mai PM, Spudich P, Boatwright J (2005) Hypocenter locations in finite-source rupture models. Bull Seis Soc Am 95(3):965–980
128. Monelli D, Mai PM (2008) Bayesian inference of kinematic earthquake rupture parameters through fitting of strong motion data. Geophys J Int 173:220–232. doi:10.1111/j.1365-246X.2008.03733.x
129. Nakamura H, Miyatake T (2000) An approximate expression of slip velocity time functions for simulation of near-field strong ground motion. Zishin (J Seis Soc Jpn) 53:1–9
130. O'Connell DRH (1999) Replication of apparent nonlinear seismic response with linear wave propagation models. Science 283(5410):2045–2050
131. Oglesby DD, Archuleta RJ (2003) The three-dimensional dynamics of a nonplanar thrustfault. Bull Seis Soc Am 93(5):2222–2235
132. Oglesby DD, Day SM (2001) Fault geometry and the dynamics of the 1999 Chi-Chi (Taiwan) earthquake. Bull Seis Soc Am 91(5):1099–1111
133. Oglesby DD, Day SM (2002) Stochastic fault stress: Implications for fault dynamics and ground motion. Bull Seis Soc Am 92(8):3006–3021
134. Oglesby DD, Dreger DS, Harris RA, Ratchkovski N, Hansen R (2004) Inverse kinematic and forward dynamic models of the 2002 Denali fault earthquake, Alaska. Bull Seis Soc Am 94(6):S214–S233
135. Olsen KB (2000) Site amplification in the Los Angeles basin from three-dimensional modeling of ground motion. Bull Seis Soc Am 90(6):77–94
136. Olsen KB, Archuleta RJ (1996) Three-dimensional simulation of earthquakes on the Los Angeles fault system. Bull Seis Soc Am 86(3):575–596
137. Olsen KB, Archuleta RJ, Matarese JR (1995) 3-Dimensional simulation of a magnitude-7.75 Earthquake on the San-Andreas fault. Science 270(5242):1628–1632
138. Olsen KB, Day SM, Minster JB, Cui Y, Chourasia A, Faerman M, Moore R, Maechling P, Jordan T (2006) Strong shaking in Los Angeles expected from southern San Andreas earthquake. Geophys Res Lett 33:L07305. doi:10.1029/2005GL025472
139. Oprsal I, Fäh D, Mai PM, Giardini D (2005) Deterministic earthquake scenario for the Basel area Simulating – strong motion and site effects for Basel (Switzerland). J Geophys Res 110:B04305. doi:10.1029/2004JB003188
140. PEER (2008) Next Generation Attenuation of Ground Motions (NGA) Project. http://peer.berkeley.edu/products/nga_project.htmt
141. Peng ZG, Ben-Zion Y (2006) Temporal changes of shallow seismic velocity around the Karadere–Duzce branch of the North Anatolian Fault and strong ground motion. Pure Appl Geophys 163:567–600
142. Piatanesi A, Cirella A, Spudich P, Cocco M (2007) A global search inversion for earthquake kinematic rupture history: application to the 2000 western Tottori, Japan earthquake. J Geophys Res B07314. doi:10.1029/2006JB004821
143. Pitarka A, Irikura K (1996) Modeling 3D surface topography by finite-difference method; Kobe-JMA station site, Japan, case study. Geophys Res Lett 23(20):2729–2732
144. Pitarka A, Irikura K, Iwata T, Kagawa T (1996) Basin structure effects in the Kobe area inferred from the modeling of ground motions from two aftershocks of the January 17,1995, Hyogo-ken Nanbu earthquake. J Phys Earth 44(5):563–576
145. Pitarka A, Irikura K, Iwata T, Sekiguchi H (1998) Three-dimensional simulation of the near-fault ground motion for the 1995 Hyogo-ken Nanbu (Kobe), Japan, Earthquake. Bull Seis Soc Am 88(2):428–440
146. Pitarka A, Somerville P, Fukushima Y, Uetake T, Irikura K (2000) Simulation of near-fault strong-ground motion using hybrid Green's function. Bull Seis Soc Am 90(3):566–586
147. Pulido N, Kubo T (2004) Near-fault strong motion complexity of the 2000 Tottori earthquake (Japan) from a broadband source asperity model. Tectonophysics 390:177–192
148. Reid HF (1910) The Mechanics of the Earthquake: The California Earthquake of April 18, 1906, Report of the State Investigation Commission, Vol. 2. Carnegie Institution of Washington, Washington DC
149. Ripperger J, Mai PM (2004) Fast computation of static stress changes on 2D faults from final slip distributions. Geophys Res Lett 31(18):L18610. doi:10.1029/2004GL020594
150. Ripperger J, Mai PM, Ampuero J-P (2008) Near-Field Ground Motion from Dynamic Earthquake Rupture Simulations. Bull Seis Soc Am 98(3):1207–1228
151. Ripperger J, Mai PM, Ampuero J-P, Giardini D (2007) Earthquake source characteristics from dynamic rupture with constrained stochastic fault stress. J Geophys Res 112:B04311. doi:10.1029/2006JB004515
152. Ritter JRR, Mai PM, Stoll G, Fuchs K (1997) Scattering of teleseismic waves in the lower crust Observations – in Massif Central, France. Phys Earth Planet Int 104:127–146
153. Robinson DP, Brough C et al (2006) The M-w 7.8:2001 Kunlunshan earthquake: Extreme rupture speed variability and effect of fault geometry. J Geophys Res B08303. doi:10.1029/2005JB004137
154. Romanowicz B (1992) Strike-slip earthquakes on quasi-vertical transcurrent faults; inferences for general scaling relations. Geophys Res Lett 19(5):481–484

155. Rosakis AJ, Samudrala O, Coker D (1999) Cracks faster than the shear wave speed. Science 284(5418):1337–1340
156. Rubinstein JL, Beroza GC (2004) Evidence for widespread nonlinear strong ground motion in the M-W 6.9 Loma Prieta Earthquake. Bull Seis Soc Am 94(5):1595–1608
157. Sadigh K, Chang CY, Egan JA, Makdisi F, Youngs RR (1997) Attenuation relationships for shallow crustal earthquakes based on California strong motion data. Seis Res Lett 68(1):180–189
158. Sato H, Fehler M (1998) Seismic wave propagation and scattering in the heterogeneous Earth, Press AIP/Springer, New York
159. Sato H, Fehler M, Wu R-S (2003) Scattering and attenuation of seismic waves in the lithosphere. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part A. Academic Press, San Diego, p 195–208
160. Sato T, Graves RW, Somerville PG (1999) Three-dimensional finite-difference simulations of long-period strong motions in the Tokyo metropolitan area during the 1990 Odawara earthquake (MJ 5.1) and the great 1923 Kanto earthquake (MS 8.2) in Japan. Bull Seis Soc Am 89:579–607
161. Sato T, Graves RW, Somerville PG, Kataoka S (1998) Estimates of regional and local strong motions during the great 1923 Kanto, Japan, earthquake (Ms 8.2). Part 2: Forward simulation of seismograms using variable-slip rupture models and estimation of near-fault long-period ground motions. Bull Seis Soc Am 88(1):206–227
162. Sato T, Helmberger DV, Somerville PG, Graves RW, Saikia CK (1998) Estimates of regional and local strong motions during the great 1923 Kanto, Japan, earthquake (Ms 8.2). Part 1: Source estimation of a calibration event and modeling of wave propagation paths. Bull Seis Soc Am 88(1):183–205
163. Scherbaum F, Schmedes J, Cotton F (2004) On the conversion of source-to-site distance measures for extended earthquake Source models. Bull Seis Soc Am 94(3):1053–1069
164. Scholz C (2002) The mechanics of earthquakes and faulting, 2nd edn. Cambridge University Press, Cambridge
165. Sekiguchi H, Irikura K, Iwata T (2002) Source inversion for estimating the continuous slip distribution on a fault-introduction of Green's functions convolved with a correction function to give moving dislocation effects in subfaults. Geophys J Int 150(2):377–391
166. Shakal A, Haddadi H, Graizer V, Lin K, Huang M (2006) Some key features of the strong-motion data from the M 6.0 Parkfield, California, earthquake of 28 September 2004. Bull Seis Soc Am 96(4B):S90–S118
167. Shakal AF, Haddadi HR, Huang MJ (2006) Note on the Very-High-Acceleration Fault Zone 16 Record from the 2004 Parkfield Earthquake. Bull Seis Soc Am 96(3):S119–S128
168. Sigbjörnsson R, Ambraseys NN (2003) Uncertainty analysis of strong-motion and seismic hazard. Bull Earthq Eng 1:321–347
169. Somerville P, Irikura K, Graves R, Sawada S, Wald DJ, Abrahmason N, Iwasaki Y, Kagawa T, Smith N, Kowada A (1999) Characterizing crustal earthquake slip models for the prediction of strong ground motion. Seis Res Lett 70(1):59–80
170. Somerville PG (2003) Magnitude scaling of the near fault rupture directivity pulse. Phys Earth Planet Int 37:201–212
171. Somerville PG, Smith NF, Graves RW, Abrahamson NA (1997) Modification of empirical strong ground motion attenuation relations to include the amplitude and duration effects of rupture directivity. Seis Res Lett 68(1):199–222
172. Spudich P, Chiou BSJ, Graves R, Collins N, Somerville P (2004) A formulation of directivity for earthquake sources using isochrone theory. United States Geological Survey, Open-File Report 2004–1268
173. Spudich P, Frazer LN (1984) Use of ray theory to calculate high-frequency radiation from earthquake sources having spatially variable rupture velocity and stress drop. Bull Seis Soc Am 74(6):2061–2082
174. Spudich P, Hellweg M, Lee WHK (1996) Directional topographic site response at Tarzana observed in aftershocks of the 1994 Northridge, California, earthquake; implications for mainshock motions. Bull Seis Soc Am 86(1, Part Suppl B):193–208
175. Spudich P, Joyner WB, Lindh AG, Boore DM, Margaris BM, Fletcher JB (1999) SEA99: a revised ground motion prediction relation for use in extensional tectonic regimes. Bull Seis Soc Am 89(1):1156–1170
176. Spudich P, Xu L (2003) Software for calculating earthquake ground motions from finite faults in vertically varying media In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part B. Academic Press, San Diego, pp 1857–1875
177. Strasser FO, Bommer JJ, Abrahamson NA (2008) Truncation of the distribution of ground-motion residuals. J Seis 12:79–105
178. Sørensen MB, Oprsal I, Bonnefoy SC, Atakan K, Mai PM, Pulido N, Yalciner C (2006) Local site effects in Ataköy, Istanbul, Turkey, due to a future large earthquake in the Marmara Sea. Geophys J Int 167(3):1413–1424
179. Takeo M (1998) Ground rotational motions recorded in near-source region of earthquakes. Geophys Res Lett 25(6):789–792
180. Tinti E, Fukuyama E, Piatenesi A, Cocco M (2005) A kinematic source-time function compatible with earthquake dynamics. Bull Seis Soc Am 95(4):1211–1223
181. Tinti E, Spudich P, Cocco M (2005) Earthquake fracture energy inferred from kinematic rupture models on extended faults. J Geophys Res 110:B12303. doi:10.1029/2005JB003644
182. Trifunac MD (1982) A note on rotational components of earthquake motions on ground surface for incident body waves. Soil Dyn Earthq Eng 1:11–19
183. Trifunac MD (2006) Effects of torsional and rocking excitations on the response of structures. In: Teisseyre R et al (eds) Earthquake Source Asymmetry, Structural Media and Rotation Effects. Springer, Heidelberg, pp 569–582
184. Wald DJ, Heaton TH (1994) Spatial and temporal distribution of slip for the 1992 Landers, California, earthquake. Bull Seis Soc Am 84(3):668–691
185. Wald DJ, Heaton TH, Hudnut KW (1996) The slip history of the 1994 Northridge, California, earthquake determined from strong-motion, teleseismic, GPS, and leveling data. Bull Seis Soc Am 86(1):S49–S70
186. Wang JH, Ou SS (1998) On scaling of earthquake faults. Bull Seis Soc Am 88(2):758–766
187. Wells DL, Coppersmith KJ (1994) New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement. Bull Seis Soc Am 84(4):974–1002
188. Working Group on California Earthquake Probabilities (1999) Earthquake probabilities in the San Francisco Bay region: 2000 to 2030–A summary of findings. Open USGS-File Report 99–517

189. Working Group on California Earthquake Probabilities (2003) Earthquake probabilities in the San Francisco Bay region: 2002–2031. Open USGS-File Report 03–214
190. Wu RS, Aki K (1985) The fractal nature of the inhomogeneities in the lithosphere evidenced from seismic-wave scattering. Pure Appl Geophy 123(6):805–818
191. Wu Y-M, Shin T-C, Chang C-H (2001) Near real-time mapping of peak ground acceleration and peak ground velocity following a strong earthquake. Bull Seis Soc Am 91(5):1218–1228
192. Zeng YH (1993) Theory of scattered P-wave and S-wave energy in a random isotropic scattering medium. Bull Seis Soc Am 83(4):1264–1276
193. Zeng YH, Anderson JG, Su F (1995) Subevent rake and random scattering effects in realistic strong ground motion simulation. Geophys Res Lett 22(1):17–20
194. Zeng YH, Su F, Aki K (1991) Scattering wave energy propagation in a random Isotropic scattering medium. 1. Theory J Geophys Res 96(B1):607–619

Books and Reviews

Abercrombie R, McGarr A, Kanamori H, Di Toro G (eds) (2006) Earthquakes: Radiated Energy and the Physics of Faultingq. Geophysical Monograph Series, 170. American Geophysical Union, 10.1029/170GM28

Lay T, Wallace TC (1995) Modern Global Seismology. Academic Press, San Diego

Lee WHK, Kanamori H, Jennings PC, Kisslinger C (2002) International Handbook of Earthquake an Engineering Seismology, Part A and B. Academic Press, San Diego

Web Links

Database of finite-source rupture models http://www.seismo.ethz.ch/srcmod

COSMOS strong-motion database http://db.cosmos-eq.org

Kyoshin Network (Japan) http://www.k-net.bosai.go.jp

PEER (Pacific Earthquake Engineering Research Center) http://peer.berkeley.edu

NGA Project page at PEER http://peer.berkeley.edu/products/nga_project.html

International Working Group on Rotational Seismology (IWGoRS) http://www.rotationalseismology.org

Group Model Building

Etiënne A. J. A. Rouwette, Jac A. M. Vennix
Institute for Management Research, Radboud University, Nijmegen, The Netherlands

Article Outline

Glossary
Definition of the Subject
Introduction
Group Model Building: Basic Ideas and Concepts
Designing Group Model Building Projects
Conducting Group Model Building Sessions
Researching Group Model Building Effectiveness
Future Directions
Bibliography

Glossary

Facilitator Person who guides the group process in group model building.

Gatekeeper Person who forms the linking pin between modeling team and management team.

Knowledge elicitation Process of capturing the knowledge contained in the mental models of team members of the management team.

Modeler Person who constructs the quantified model during group model building.

Recorder Person who takes notes during group model building sessions and constructs workbooks.

Reference mode Graph(s) showing the behavior of the problem over time.

Workbook Booklet which contains summary of previous group model building sessions and prepares for subsequent sessions.

Client Person (or agency) who buys a model.

Definition of the Subject

Computer (simulation) models have been used to support policy and decision making in the decades after World War II. Over the years modelers learned that the application of these models to policy problems was not as straightforward as had been thought initially. As of the beginning of the 1970s studies started to appear that questioned the use of large-scale computer models to support policy and decision making (cf. [24,31]). Lee's article bears the significant title: "Requiem for large scale models", a statement that leaves little room for ambiguity. Other authors who have studied the impact records of computer models also seem rather sceptical (e. g. [9,22,25,73]). It is interesting to note that Greenberger et al., after interviewing both modelers and policy makers (for whom the models were constructed) found that modelers generally pointed to the fact that they learned a lot from modeling a particular policy issue. Policy makers on the other hand indicated that they did not really understand the models nor had much confidence in them. The results of these studies pointed in the direction of learning from computer models, i. e. conceptual or enlightenment use rather than instrumental use, where policy recommendations could straightforwardly be deduced from the model analysis and outcomes. In other words it is in the process of modeling a policy problem where the learning takes place which

is required to (re)solve a problem. And it is also in this process that one needs to anticipate the implementation of policy changes. By the end of the 1970s system dynamics modelers pointed out that implementation of model outcomes was a neglected area (e. g. [50,74]) and that modelers sometimes naively assumed that implementation was straightforward, thereby neglecting organizational decision making as a political arena.

In other words client participation in the process of model construction and analysis is required for successful modeling and implementation of insights from the model into policy making. Or as Meadows and Robinson put it:

> Experienced consultants state that the most important guarantee of modelling success is the interested participation of the client in the modelling process (p. 408 in [34]).

Over the years this has given rise to all kinds of experiments to involve clients in the process of model construction. In the 1990s the term Group Model Building was introduced to refer to more or less structured approaches for client involvement in system dynamics model construction and analysis.

Introduction

From the early days of the field, the topic of client involvement in the process of model construction has raised attention in the system dynamics literature. Jay Forrester, the founder of the field of system dynamics, has repeatedly indicated that most of the knowledge needed to construct a system dynamics models can be found in the mental database of the participants of the system to be modeled [20,21]. Over the years several system dynamics modelers have experimented with approaches to involve client (groups) in model construction and analysis. This development in the system dynamics community parallels a movement in the operational research and systems fields towards more attention for stakeholders' opinions. A number of authors (e. g. [1]) criticized traditional OR and systems approaches as unsuitable for ill-structured problems that arise from differences between stakeholders' views on the problem. For ill-structured problems a range of new methods was developed [35].

The developments in the system dynamics, operational research and systems communities have given rise to a set of distinct methods and approaches. However, practitioners work on problems that have clear similarities to those encountered in other disciplines and frequently borrow techniques from one another. The boundaries between methods are therefore difficult to draw and there is a degree of overlap between approaches in and between fields. Below we first describe the distinguishing characteristics of system dynamics, as this separates group model building most clearly from other approaches fostering client involvement. We then describe a number of distinct group model building approaches.

System Dynamics

System dynamics is most easily characterized by its emphasis on two ideas: (a) the importance of closed loops of information and action for social systems, i. e. social systems as information feedback systems and (b) the need to use formal models to study these loops. System dynamicists assume that the dynamic behavior of a social system is the result of its underlying feedback structure. Actors use the information about the structure as input to their decisions, and by implementing their decision influence system behavior. This creates an interlocked chain of action and information which is also known as a feedback loop. Richardson (see p. 1 in [44]) describes a feedback loop as follows:

> The essence of the concept … is a circle of interactions, a closed loop of action and information. The patterns of behavior of any two variables in such a closed loop are linked, each influencing, and in turn responding to the behavior of the other.

As an illustration of the use of information on the system state in decisions, imagine a simple example on customer behavior. Let us assume that if customers perceive that a product's functionality increases, more products will be bought. This will increase profits and thereby the design budget. An increased design budget can be used to improve the product's design, which will lead more customers to buy the product, and so on. Thus, decisions of actors within the system have an important influence on the system's behavior. If we continue to add other factors and relations to our example and capture these in a model, the diagram in Fig. 1 may result.

As Fig. 1 shows, a causal loop diagram consists of variables, relationships, and feedback loops. Relations can be of two types: positive and negative. A positive relation indicates that both variables change in the same direction. In the model above, an increase in retail price will lead to an increase in profits, indicating a positive relationship. Variables in a negative relationship change in opposite directions. An increase in costs will decrease profits, indicating a negative relationship. The snowball rolling down the slope in the right hand side of Fig. 1 indicates a positive feedback loop. We assumed that an increase in profit re-

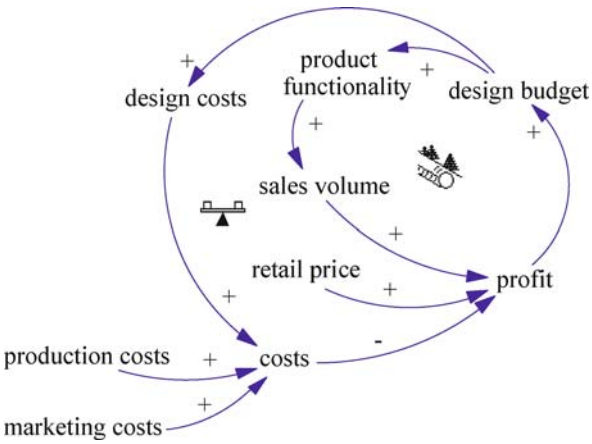

Group Model Building, Figure 1
Example of a causal loop diagram

sults in a direct increase in the design budget. A higher budget allows for increased product functionality, which increases sales volume and finally profit. Starting from an increase in profit, the result is a further increase in profit. This is a so-called positive or self-reinforcing loop. However, if we assume that the design department uses its complete budget each year, an increased budget will contribute to design costs and lower profits. This is a negative or balancing loop, indicated by the balance symbol.

The second important idea in system dynamics is that formal models are necessary to understand the consequences of system structure. Since system dynamics models contain many (often non-linear) relations and feedback loops, it becomes very difficult to predict their behavior without mathematical simulation. Systems are assumed to consist of interacting feedback loops, which may change in dominance over time. Diagrams such as the one depicted above are frequently used in the interaction with clients. Before the dynamic consequences of the structure captured in Fig. 1 can be studied, it is necessary to further specify both the variables and relations used in the model. Two categories of variables are distinguished: stocks and flows. Stocks are entities existing at a certain time period, for example supplies, personnel, or water in a reservoir. Flows are entities measured over a time period, such as deliveries, recruitment, or inflow of water. Relationships are separated into physical flows and information flows. If we capture differences between stocks and flows and information and physical flows in a diagram, a stock and flows diagram results.

As can be seen in Fig. 2, information links are depicted with a single arrow and physical flows with a double arrow. The physical human resources flow is separated in three stocks: number of rookies, number of junior researchers, and number of senior researchers. Recruitment will lead to an increase in the number of rookies. Two other flows influence the number of people in the stocks: rookies may be promoted to junior researchers and junior researchers may be promoted to senior researchers. The human resources flow is related to the project flow with information links, for instance indicating that acquisition of research projects is determined by the number of senior researchers.

Group Model Building Approaches

As pointed out before, client involvement has been important to system dynamics from the start of the field. System dynamics emphasizes feedback loops and the use of formal models. In this section we describe how models based on these ideas are built in interaction with actors and stakeholders in the problem at hand. A number of different participative model building formats can be identified, i. e. the reference group approach [43,61]; the stepwise approach [76]; the Strategic Forum [49]; modeling as learning [29]; strategy dynamics [71,72] and Hines' "standard method" [39]. Below we describe each approach briefly.

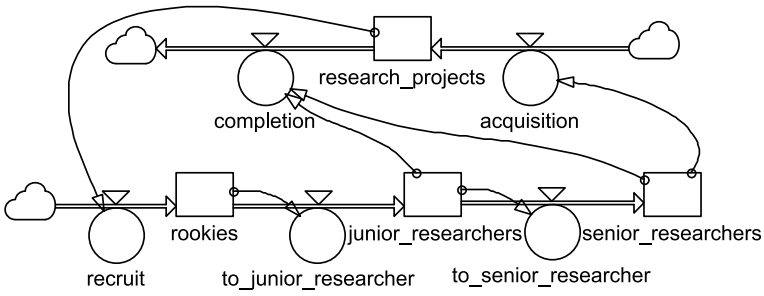

Group Model Building, Figure 2
Example of a stock and flows diagram

In the Reference Group approach [43,61] participation takes the form of frequent interaction between the modeling team and a group of eight to ten clients. The approach starts with the identification of interest groups, of which representatives are invited to contribute to the modeling effort. The representatives are referred to as referents. In a series of interviews and meetings, the problem to be addressed is defined more specifically. On the basis of this definition and the information gathered in the interviews and meetings, the modeling team develops a preliminary model. In the remainder of the project the modelers are responsible for model improvements while the referents function as critics. This model is elaborated in a series of meetings and is at the same time used as a tool for structuring the discussion. In later sessions, model output is used for developing scenarios. In a scenario discussion the model is run and results are described and analyzed by the modelers. The reference group is then asked to determine to what extent the model's behavior corresponds to their expectations about reality, and if it does not, to suggest changes. These suggestions can trigger changes in the model structure, initiating a new round in the discussions.

The stepwise approach [76] is founded on the idea that full quantification of models is not always possible or desirable. The approach starts with a definition of the problematic behavior. If possible, this definition is given in the form of a behavior over time of the problem of interest. Modeling starts by roughly sketching the feedback loops responsible for this behavior. The key variables related to the cause for concern are identified, followed by the system resources connected to these key variables and their initial states. The resources are used to derive the central stocks in the system. From the resources, the resource flows can then be sketched with the associated rates of conversion. Delays are added to these flows if they are significant. Next, organizational boundaries, flows of information and strategies through which the stocks influence the flows, are added. Again, if there are significant delays, these are added to the information linkages. In the final step, information flows and strategies linking different resource flows are added. The steps are repeated until the relevant feedback loops have all been included. Wolstenholme indicates that these steps often provide the insights necessary to infer system behavior from the structure, which reduces the need for quantification. Models can also be analyzed in a qualitative manner.

The steps that make up the Strategic Forum [49] provide a detailed insight of how clients are encouraged to participate in modeling. The strategic forum consists of eight steps, of which the first two are conducted before the actual meeting (also called the forum) with the client group. The process begins with interviews prepared by a small questionnaire, in which three issues are addressed: ideas on the current situation, a statement of the vision for the future, and agreement on a preliminary map of the problem. On the basis of the interviews, the modeler constructs an integrated map and accompanying computer model. In the second step the project team designs a number of small group exercises that will be used during the forum. The exercises are aimed at discovering important structural and behavioral elements and are similar to the scenario discussions in the reference group approach. The most important difference is that before simulation results are shown, participants have to 'put a stake in the ground', i. e. they have to make a prediction of model behavior on the basis of a change in a policy variable and values for connected parameters. The model is then simulated and results are compared with participants' expectations. Discrepancies between predictions and simulations are identified, and might point to inconsistencies in participants' ideas or lead to model improvements. In the following steps the participants meet in a series of workshops. Each workshop opens with an introduction and a big picture discussion. The heart of the session consists of exercises aimed at internal consistency checks, addressing the consistency between the group's mental model and the computer model. As in the other approaches, model structure will be changed if inconsistencies with the participants' ideas on the problem are revealed. In the final phase of policy design, potential consequences of strategic policies are addressed and the existing capability of realizing the strategic objectives. A wrap-up discussion and identification of follow-up activities concludes the Strategic Forum.

Richmond (see p. 146 in [49]) emphasizes that the main purpose of the Strategic Forum is to check the consistency of strategy. The insights gained by the client therefore frequently lead to changes in strategy or operating policies, but less frequently to changes in objectives or the mission statement. One important element of ensuring an impact on participants' ideas is the (dis)confirmation of expectations on simulation outcomes.

Lane [29] describes a modeling approach developed at Shell International Petroleum, known as 'modeling as learning'. Lane explicitly sets this approach apart from the widely used expert consultancy methodology (e. g. [58]). His approach also puts strong emphasis on involving decision makers in the modeling process. By showing decision makers the benefits of participation early on in the process, an attempt is made to persuade them to spend time in direct interaction with the model. The approach centers on capturing and expressing the client's ideas, initiating a dis-

cussion on the issue with 'no a-priori certainty regarding quantification, or even cause and effect' (see p. 70 in [29]). The modelers also strive to include both hard as well as 'soft' aspects of the problematic situation. In doing this, it is hoped that the clients' ideas are included in the model and that ownership is created. This is encouraged by making models and model output transparent to participants, helping the client 'to learn whichever techniques are used in a project' (see p. 71 in [29]). Lane states that the focus throughout the approach is on a process of learning, using such elements as experimentation with the model, testing of assumptions and representing and structuring ideas in a logical way.

Hines' approach [39] starts off by diagnosing the problem. This step comes down to gathering and clustering problem variables. Problem behavior is visualized by sketching the graph over time of the problematic behavior. In the second step the structure underlying the problematic behavior is captured in a causal diagram. This so-called dynamic hypothesis incorporates many of the problem variables identified earlier. The diagram helps to clarify the boundary of the problem that will be addressed and thus limits the project scope. The next step is to identify accumulations in the system, which will form the stocks in the system dynamics model. In the construction of the computer model most work is done by the modelers, with client participation limited to providing data such as numerical values and details of the work processes relevant to the problem at hand. Model structure and behavior is then explained to the client. Discussions with the client then lead to a series of model iterations, increasing confidence of the client in model calibration and validity. Similar to other participative approaches, policy runs are used to test proposed interventions in the problem.

Warren [71,72] describes an approach to participative modeling that strongly focuses on identifying accumulations (stocks) in the system. In order to identify central accumulations, clients are asked to identify the strategic resources in the problem at hand. Increases and decreases in resources then lead to the identification of flows. Warren's approach differs from the ones described above in the sense that stocks and flows are differentiated from the outset. This means that causal loop diagrams are not used. In addition, graphs over time are recorded next to each variable in the model. By gradually adding elements to the model while visually relating structure and behavior, the clients' understanding of the problem is gradually increased.

As mentioned before, the boundaries around approaches are not easy to draw and one method may 'borrow' techniques of another. Insights and practices from the operational research and system fields have been merged with those in system dynamics to develop combined methods. For example, modeling as learning is one of the approaches incorporating elements of soft operational research methodologies. Lane and Oliva [30] describe the theoretical basis for integrating system dynamics and soft systems methodology. The cognitive mapping approach (e. g. [17]) also offers tools and techniques that are used in system dynamics studies.

In addition to combining different methods, approaches are sometimes also tailored to use in specific content areas. An example is van den Belt's [63] mediated modeling, which combines insights from participative system dynamics modeling and consensus building on environmental issues.

Group Model Building: Basic Ideas and Concepts

The separate approaches described in the last section continue to be developed and used in practical problems. Although we are not sure that all proponents of these approaches would characterize themselves as using "group model building", this term has been used more and more in the last decades to refer to system dynamics approaches with client involvement in a general sense. The two approaches that coined the term group model building evolved more or less simultaneously, with considerable cross-fertilization of ideas, at SUNY at Albany and Radboud University Nijmegen in the Netherlands (see [36,69]). In an early application at Radboud University, participants were involved in a Delphi study consisting of mailed questionnaires and workbooks, followed by workshops [66]. In the dissertations by Verburgh [70] and Akkermans [4] a similar approach is used under the name of participative policy modeling and participative business modeling, respectively. In its latest version group model building is a very open approach, which allows for the use of preliminary models or a start from scratch, uses individual interviews, documents and group sessions, qualitative or quantitative modeling and small as well as large models. Vennix [64,65] provides a set of guidelines for choosing among these different approaches, building on and adding to the studies mentioned above. Andersen and Richardson [5] provide a large number of "scripts" that can help in setting up modeling projects. The procedures described are a long way from the earlier descriptions of a set of steps that seem to prescribe standard approaches applicable to most modeling projects. Instead, the guidelines offered have more the appearance of tool boxes, from which the appropriate technique can be selected on the basis of problem characteristics and the clients involved.

Group model building is generally conducted with a group of at least six and up to 15 people. The group is guided by at least two persons: a facilitator and a modeler/recorder. The group is seated in a semi circle in front of a whiteboard and/or projection screen, which serves as a so-called group memory. A projection screen is typically used when a model is constructed with the aid of system dynamics modeling software with a graphic interface (e. g. Vensim, Powersim, Ithink). This group memory documents the model under construction and is used as a parking lot for all kinds of unresolved issues which surface during the deliberations of the group.

In Fig. 3, the small circles indicate the persons present in the session. Apart from the participants, there is a facilitator and a recorder. The facilitator has the most important role in the session as he or she guides the group process. His/her task, as a neutral outsider, is to (a) elicit relevant knowledge from the group members, (b) to (help) translate elicited knowledge into system dynamics modeling terms, and (c) make sure that there is an open communication climate so that in the end consensus and commitment will result. The recorder keeps track of the elements of the model. In Fig. 3 (s)he is seated behind a computer and the model is projected on the screen in front of the group. A separate whiteboard (upper right hand corner) is used to depict the reference mode of behavior and record comments or preliminary model structure. As the model is visible to all participants, it serves as a group memory that at each moment reflects the content of the discussion up to that point. A group model building session is generally conducted in the so-called chauffeured style, where only the facilitator uses electronic support and projection equipment, while participants do not have access to electronic communication media [38]. The central screen or whiteboard will be used to depict the model, as shown in Fig. 3.

The role of liaison between the organization and the modeling team is performed by the gatekeeper, who is generally also a member of the participant group. The gatekeeper is the contact between both parties, and has an important role in the decision which participants to involve in the sessions. Apart from the gatekeeper, the facilitator and the recorder, two other roles may be important in a modeling session [46], i. e. a process and a modeling coach. The process coach functions as an observer and primarily pays attention to the group process. The model coach needs to be experienced in system dynamics modeling but might also be an expert in the content area as well. As Richardson and Andersen [46] point out, all roles are important in group model building but not all of them have to be taken up by a single person. One person might

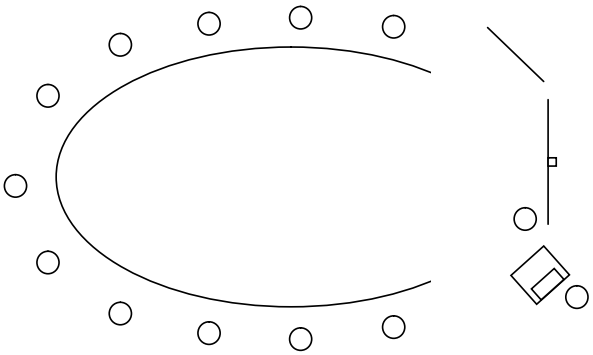

Group Model Building, Figure 3
Typical room layout for group model building with participants seated in a semi-circle, white board and facilitator in front, and computer and overhead projector (adapted from [5])

for instance combine the roles of facilitator and process coach. Taken together, these different roles constitute the facilitation or modeling team.

In principle the group follows the normal steps in the construction of a system dynamics model. This means that the first step is the identification of the strategic issue to be discussed, preferably in the form of a so-called reference mode of behavior, i. e. a time series derived form the system to be modeled which indicates a (historical) undesirable development over time. As an example let us take the sales of a software product. An initial problem statement might be falling profit. Typically the problematic behavior will be depicted in a graph over time as in Fig. 4.

In the graph above, a projection of the expected behavior is included for the years after 2008.

The next step is to elicit relevant variables with which the model construction process can be started. Depending on the type of problem this will take the form of either a causal loop diagram or a stocks and flow diagram and is generally referred to as the conceptualization stage. The following step is to write mathematical equations (model formulation) and to quantify the model parameters. As described in the introductory section, most of the model formulation work is done backstage as it is quite time consuming and members of a management team generally are not very much interested in this stage of model construction. In this stage, the group is only consulted for crucial model formulations and parameter estimations. Experienced group model builders will start to construct a simple running model as soon as possible and complicate it from there on if required. In the end the model should of course be able to replicate the reference mode of behavior (as one of the many validity tests) before it can be sensibly be used as a means to simulate the potential effects of strategies and scenarios.

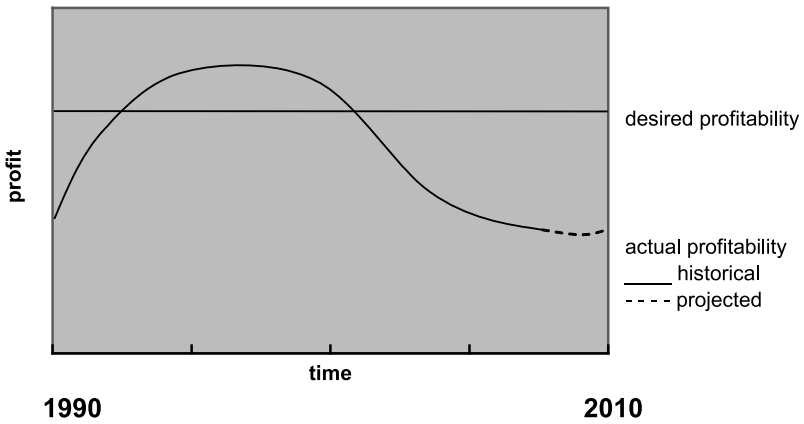

Group Model Building, Figure 4
Problem description in graphical form: reference mode of behavior

Objectives of Group Model Building

As mentioned in the introduction, the founder of system dynamics has repeatedly pointed out that much of the knowledge and information which is needed to construct a model can be found in the mental models of system participants. At first sight it may seem that the most important objective of building a system dynamics model is to find a robust strategy to solve the problem of the organization. In the end that is why one builds these models. From this perspective the most important issue in group model building is how to elicit the relevant knowledge from the group. However, as stated before, decision making in organizations has its own logic, and in many cases there is quite some disagreement about the problem and how it should be tackled. No wonder that implementation of model outcomes is difficult if the model building process is not well integrated with decision making processes in organizations, when it comes to creating agreement and commitment with a decision. From that perspective knowledge elicitation is only one element in the process of model construction. It is not so much the model but to a greater extent the process of model construction which becomes important. Somewhat simplified one could say that in the "standard" approach when an organization is confronted with a strategic problem it hires a modeler (or group of modelers) to construct a model and come up with recommendations to "solve" the problem. However, in most cases these recommendations become part of the discussion in the management team and get misunderstood, or adapted or frequently just disappear as a potential solution from the discussion. Hence Watt's title of his paper: "Why won't anyone believe us?" becomes very much understandable from the point of view of the modeler. So rather than creating a situation where modelers "take away" the problem from the organization and (after considerable time) return with their recommendations, the model building process is now used to structure the problem, guide communication about it and test the robustness of strategies taking into account other criteria and information which is not included in the model, but does play a role for the organization when making the decision. Stated differently, the model building process now becomes intertwined with the process of decision making in an organization. And this in turn means that other objectives than knowledge elicitation become important.

Simultaneously with the attempts to involve clients in the process of constructing system dynamics models the objectives of group model building have been defined at several levels, i. e. the individual, the group and the organizational level (cf. [7,65]). The main goal at the individual level is change of mental models and learning. The idea is that participants should better understand the relationship between structure and dynamics and how their interventions may create counterintuitive results. Unfortunately research has revealed that this is hardly the case. Even after extensive training people have difficulty to understand the relationship between structure and dynamics (for a review see [8,53]). A second goal at the individual level is behavioral change. Frequently the conclusions of a modeling intervention point in the direction of behavioral change, for example implementing a new job rotation scheme, or a change in purchasing policy. The question can then be asked how insights from the modeling intervention are translated to changes in behavior. Rouwette (e. g. [52,53,57]) uses a framework from social psychology to understand the impact of modeling on behavior. The theory of Ajzen [2,3] explains behavior on the basis of

Group Model Building, Table 1
Objectives of group model building (cf. [53])

Individual	Positive reaction
	Mental model refinement
	Commitment
	Behavioral change
Group	Increased quality of communication
	Creation of a shared language
	Consensus and alignment
Organization	System changes
	System improvement or results
Method	Further use
	Efficiency

(a) attitude, (b) perceptions of norms and (c) perceptions of control. It seems likely that each of these concepts is influenced in modeling sessions. When for example model simulations reveal unexpected levers for improving system behavior, we can expect that perceived control will increase. Another example: let's imagine that a manager is participating in a modeling session, where another participant reveals positive outcomes of a certain policy option. If these positive outcomes were previously not known to the manager, hearing them might make his/her attitude towards that option more positive (cf. [42]).

At the group level objectives refer to mental model alignment [28] and fostering consensus [51,67,75]). Creating consensus should not be confused with premature consensus, i. e. not discussing conflicting viewpoint. Here it concerns creation of consensus after critical debate and discussion of opinions has taken place. This type of discussion which needs to take place in a cooperative communication climate is helpful to also create commitment with the resulting decision.

At the organizational level goals have been discussed as system process change (are things done differently) and system outcome change (are customers impacted differently) [11]. Although it has to be pointed out that in many cases system changes are the result of changes in attitude and behavior of participants in the system. An overview of group model building objectives is given in Table 1. In this table finds a number of additional objectives such as positive reaction and creation of a shared language, that are more fully reviewed by Huz et al. [28], Rouwette et al. [55] and Rouwette and Vennix [53].

Designing Group Model Building Projects

When designing group model building projects there are a number of questions that need to be addressed. The first concerns the suitability of system dynamics for the problem at hand. System dynamicists generally say that a problem needs to be dynamically complex in order to be suitable to model it through system dynamics. This means that one should at least hypothesize that there are positive and negative feedback process underlying the problem. From a more practical point of view one could say that one should be able to represent the problem in the form of a reference mode of behavior. If the latter is not possible one should seriously question the use of system dynamics for the problem.

A second issue which needs to be given some thought is the question whether to use qualitative or quantitative modeling. Within the system dynamics community there is still a debate about the question whether qualitative modeling (or: mapping) can be considered system dynamics (see [13,14,23]). In short those who disagree point out that without quantification and simulation one cannot reliably develop a robust policy simply because the human mind is not capable of predicting the dynamic effects of (interventions in) a dynamically complex structure. Those who do use mapping on the other hand point out that mapping in itself can have the beneficial effect to structure the problem and at least will make managers aware of potential underlying feedback loops and their potential counterintuitive effects when intervening in a dynamically complex system. Basically the issue to quantify or not depends on the goals of the group model building intervention. If the ultimate goal is to find robust policies then quantification is required. However, if the aim is to structure a problem and to create consensus on a strategic issue then qualitative modeling may be all that is needed. This links up with Zagonel's [77] distinction between the use of models as micro worlds or as boundary objects. When used as a boundary object the emphasis is on supporting negotiation and exchange of viewpoints in a group. This is clearly the case when problems are messy, i. e. connected to other problems and when there is much divergence of opinion on what the problem is and sometimes even whether there is a problem at all.

A third issue is the question who to involve in the sessions. There are a number of criteria which are generally employed. First it is important to involve people who have the power to make decisions and changes. A second criterion is to involve people who are knowledgeable about the problem at hand. A third criterion is to involve a wide variety of viewpoints, in order to make sure that all relevant knowledge about the problem is included. Of course these guidelines may create dilemmas. For example, involving more people in the process will make the group communication process more difficult. This may in turn endanger the creation of consensus and commitment.

Group Model Building, Table 2
Group model building scripts (cf. [5])

Phase in modeling	Script
Defining a problem	Presenting reference modes
	Eliciting reference modes
	Audience, purpose, and policy options
Conceptualizing model structure	Sectors, a top down approach
	Maintain sector overview while working within a sector
	Stocks and flows, by sector
	Name that variable or sector
Eliciting feedback structure	Direct feedback loop elicitation
	Capacity utilization script
	System archetype templates
	"Black box" means-ends script
Equation writing and parametrization	Data estimation script
	Model refinement script
	"Parking lot" for unclear terms
Policy development	Eliciting mental model-based policy stories
	Create a matrix that links policy levers to key system flows
	"Complete the graph" policy script
	Modeler/reflector feedback about policy implications
	Formal policy evaluation using multiattribute utility models
	Scripts for "ending with a bang"

Another issue is whether to use a preliminary model or to start from scratch (see [45]). Although using a preliminary model may speed up the process the inherent danger is that it will be difficult to build group ownership over the model. Group ownership is clearly required to create consensus and commitment.

Finally, a range of methods and techniques is available to elicit relevant knowledge both from individuals and from groups. When it comes to individuals, well known methods are interviews, questionnaires and so-called workbooks. The latter are a kind of modified questionnaires, which are used in between sessions to report back to the group and ask new question in preparation of the next session. Interviews are being used routinely as a preparation for group model building sessions.

If a decision is made on the issues discussed above, the next important question is how to plan and execute the modeling sessions. This question is a central topic in the group model building literature and its success heavily depends on the correct choice of available techniques and the quality of the facilitator.

Conducting Group Model Building Sessions

Although careful preparation of group model building sessions is a necessity, the most important part of the whole project is what happens in the group model building sessions themselves. During the sessions not only the analysis of the problem takes place (and the model is constructed), but also the interaction process between members of the management team unfolds. It is this interaction process which needs to be guided in such a way that consensus and commitment will emerge and implementation of results will follow. As pointed out the process is guided by the group facilitator, generally someone who is not only specialized in facilitation of group processes but also in system dynamics model construction. The facilitator is supported by a recorder or modeler who helps constructing the system dynamics model while the facilitator interacts with the management team.

The facilitator may choose from a wide variety of techniques in setting up and conducting a session. As a foundation for choosing techniques, Andersen and Richardson [5] develop a set of guiding principles and so-called scripts for group model building sessions. Guiding principles capture basic ideas in the interaction with clients, such as break task/group structure several times each day, clarify group products, maintain visual consistency and avoid talking heads. Scripts are more concrete instances of these principles and refer to small elements of the interaction process [5,32]. The Table 2 shows scripts described in Andersen and Richardson's [5] original paper.

In choosing a script it is first important to be aware of the phase that is relevant in the project at that time. A common starting point, as we saw in the description of group model building approaches, is to define the central problem of interest. The reference mode of behavior can function as a guideline for involving clients in this phase. Once the central problem is clear, a logical next step is to move towards model conceptualization. In this step again a number of options are available. Andersen and Richardson [5] describe a script for identifying sectors that are important in the problem. An alternative is to start with more concrete variables in the problem, using a Nominal Group Technique [15].

Whatever scripts and techniques a facilitator employs it is important that (s)he displays the right attitude and uses the correct skills. Several different aspects of attitude are important. First of all the facilitator is not the person who thinks (s)he knows the best solution, but needs to be helpful in guiding the group to find a solution to the strategic problem the organization is faced with. Second, a facilitator should be neutral with respect to the problem that is being discussed. Being too knowledgeable about a particular problem area (e. g. strategic alliances) may thus be dangerous, because it creates the tendency to participate in the discussions. Rather than being an expert, having an inquiry attitude (i. e. asking questions rather than providing answers) is more helpful to the group. Finally, integrity and being authentic is important. Relying on tricks to guide the process will prove counterproductive, because people will look through them.

When it comes to skills, a thorough knowledge and experience in constructing system dynamics models is of course indispensable. Second, a facilitator needs to be knowledgeable about group process and have the skills to structure both the strategic problem as well as the group interaction process. For the latter, special group process techniques (e. g. brainstorming, Nominal Group Technique, Delphi) may be used, and knowledge about and skills in applying these techniques is of course a prerequisite for a successful group model building intervention. Finally, communication skills are important. Reflective listening is a skill which will help to prevent misunderstanding in communication, both between participants and the facilitator and between group members. For a more thorough discussion of these attitudes and skill in the context of group model building we refer to Vennix [64].

Researching Group Model Building Effectiveness

In the previous sections we described goals of group model building projects and principles and scripts for guiding the modeling process. In this section we consider the empirical evidence for a relation between modeling interventions and these intended outcomes. Empirical evidence can be gathered using a variety of research strategies, such as (field) experiments, surveys or (in-depth) case studies. According to the review of modeling studies by Rouwette et al. [55], the case study is the most frequently used design to study group model building interventions. We first report on the results found by these authors and then turn to other designs.

In the meta-analysis of Rouwette et al. [55], the majority of group model building studies uses a case study design and assesses outcomes in a qualitative manner. Data are collected using observation, and a minority of studies employs individual group interviews. Case reports may be biased towards successful projects and are frequently not complete. The outcomes of the modeling projects were scored along the dimensions depicted in Table 1 in the section on modeling goals. The findings show positive outcomes in almost all dimensions of outcomes. Learning about the problem seems to be a robust outcome of group model building, for example:

- Of 101 studies that report on learning effects, 96 indicate a positive effect;
- Of 84 studies focusing on implementation of results, 42 report a positive effect.

Another set of studies, using quantitative assessment of results is described by Rouwette and Vennix [53]. Although the research surveyed so far indicates positive effects of modeling on outcomes such as mental model refinement, consensus and implementation of results, important challenges remain. Research so far has paid little attention to the complexity of the intervention as described in the previous section. Pawson and Tilley [40] urge us not to assume that interventions are similar and lead to similar effects, since this would confuse meaningful differences between studies. Rouwette and Vennix [53] describe two ways to learn more about the process through which outcomes of modeling are created: base research more on theory and/or to conduct research in more controlled settings. At present only few studies address elements of group model building in a controlled setting. Shields [59,60] investigates the effect of type of modeling and facilitation on a group task. Most research on the use of system dynamics models concerns so-called management flight simulators. These studies aim to mimic the important characteristics of decision making in complex, dynamic problems, and test the effectiveness of various decision aids. Results are reviewed by Sterman [62], Hsiao and Richardson [26] and Rouwette, Größler and Vennix [56].

Increased attention to theories may shed more light on the way in which modeling effects group decisions. Theories can help in specifying relations which can then be tested. Explanatory research is needed to connect the components and outcomes of group model building interventions (see p. 194 in [7]). In the field of system dynamics modeling, two attempts at formulating theories on modeling components and outcomes are the work of Richardson et al. [47] and Rouwette [52]. The framework formulated by Rouwette [57] builds on the theories of Ajzen [2,3] and Petty and Cacioppa [42] described earlier. Richardson et al. [47] separate mental models into means, ends and means-ends models. The ends model contains goals, while the means model consists of strategies, tactics, and policy levers. The means-ends model contains the connection between the two former types of models and may contain either detailed "design" logic or more simple "operator" logic. On the basis of research on participants in a management flight simulator [6], the authors conclude that operator logic, or high level heuristics, is a necessary condition for improving system performance. Therefore, providing managers with operator knowledge is the key to implementation of system changes.

Future Directions

The success of group model building and problem structuring methodologies in general depends on a structured interaction between theory, methodology refinement and application in practical project accompanied by systematic empirical evaluation.

Rouwette and Vennix [53] indicate three areas for further development of theories:

- Review related methodologies used in complex organizational problems, to determine which theories are used to explain effects. Examples that come to mind are theories used in the operational research and systems fields [37].
- Forge a closer connection to research on electronic meeting systems. In this field, studies are usually conducted in controlled settings [16,18,41] and theory development seems to be at a more advanced stage. Research on electronic meeting systems is interesting both because of the empirical results and explanatory theories used and because of insights on the intervention process. A recent development in the field is research on ThinkLets [10]. A ThinkLet is defined as a named, packaged facilitation intervention and thus seems very similar to the concept of a group model building script.
- A third source of theories is formed by research in psychology and group decision making. Theories from these fields inform the definition of central concepts in group model building (see Table 1) and theories on modeling effectiveness. Rouwette and Vennix [54] review literature on group information processing and relate this to elements of group model building interventions.

From theories and evaluation research will come insights to further develop the methodology along the lines of (a) determining what kind of problem structuring methodology is best suited in what kind of situation, (b) refinement of procedures, (c) better understanding the nature of the intervention, and (d) better guidelines for facilitators how to work in different kinds of groups and situations.

Bibliography

Primary Literature

1. Ackoff RA (1979) The future of operational research is past. J Oper Res Soc 30(2):93–104
2. Ajzen I (1991) The theory of planned behavior. Organ Behav Human Decis Process 50:179–211
3. Ajzen I (2001) Nature and operation of attitudes. Annu Rev Psychol 52:27–58
4. Akkermans HA (1995) Modelling with managers: Participative business modelling for effective strategic decision-making. Ph D Dissertation, Technical University Eindhoven
5. Andersen DF, Richardson GP (1997) Scripts for group model-building. Syst Dyn Rev 13(2):107–129
6. Andersen DF, Maxwell TA, Richardson GP, Stewart TR (1994) Mental models and dynamic decision making in a simulation of welfare reform. Proceedings of the 1994 International System Dynamics Conference: Social and Public Policy. Sterling, Scotland, pp 11–18
7. Andersen DF, Richardson GP, Vennix JAM (1997) Group model-building: Adding more science to the craft. Syst Dyn Rev 13(2):187–201
8. Andersen DF, Vennix JAM, Richardson GP, Rouwette EAJA (2007) Group model building: Problem structuring, policy simulation and decision support. J Oper Res Soc 58(5):691–694
9. Brewer GD (1973) Politicians, bureaucrats and the consultant: A critique of urban problem solving. Basic Books, New York
10. Briggs RO, de Vreede GJ, Nunamaker JFJ (2003) Collaboration engineering with ThinkLets to pursue sustained success with group support systems. J Manag Inf Syst 19:31–63
11. Cavaleri S, Sterman JD (1997) Towards evaluation of systems thinking interventions: A case study. Syst Dyn Rev 13(2): 171–186
12. Checkland P (1981) Systems thinking, systems practice. Wiley, Chichester/New York
13. Coyle G (2000) Qualitative and quantitative modelling in system dynamics: Some research questions. Syst Dyn Rev 16(3):225–244
14. Coyle G (2001) Maps and models in system dynamics: Rejoinder to Homer and Oliva. Syst Dyn Rev 17(4):357–363
15. Delbecq AL, van de Ven AH, Gustafson DH (1975) Group techniques for program planning: A guide to nominal group and delphi processes. Scott, Foresman and Co, Glenview

16. Dennis AR, Wixom BH, van den Berg RJ (2001) Understanding fit and appropriation effects in group support systems via meta-analysis. Manag Inf Syst Q 25:167–183
17. Eden C, Ackermann F (2001) SODA – The principles. In: Rosenhead J, Mingers J (eds) Rational analysis for a problematic world revisited. Problem structuring methods for complexity, uncertainty and conflict. Wiley, Chichester, pp 21–42
18. Fjermestad J, Hiltz SR (1999) An assessment of group support systems experimental research: Methodology and results. J Manag Inf Syst 15:7–149
19. Fjermestad J, Hiltz SR (2001) A descriptive evaluation of group support systems case and field studies. J Manag Inf Syst 17:115–160
20. Forrester JW (1961) Industrial Dynamics. MIT Press, Cambridge
21. Forrester JW (1987) Lessons from system dynamics modelling. Syst Dyn Rev 3(2):136–149
22. Greenberger M, Crenson MA, Crissey BL (1976) Models in the policy process: Public decision making in the computer era. Russel Sage Foundation, New York
23. Homer J, Oliva R (2001) Maps and models in system dynamics: A response to Coyle. Syst Dyn Rev 17(4):347–355
24. Hoos IR (1972) Systems analysis in public policy: A critique. University of California Press, Berkeley/Los Angeles/London
25. House PW (1982) The art of public policy analysis: The arena of regulations and resources, 2nd printing. Sage, Thousand Oaks
26. Hsiao N, Richardson GP (1999) In search of theories of dynamic decision making: A literature review. In: Cavana RY et al (eds) Systems thinking for the next millennium – Proceedings of the 17th international conference of the system dynamics society. Wellington, New Zealand
27. Huz S (1999) Alignment from group model building for systems thinking: Measurement and evaluation from a public policy setting. Unpublished doctoral dissertation, SUNY, Albany
28. Huz S, Andersen DF, Richardson GP, Boothroyd R (1997) A framework for evaluating systems thinking interventions: an experimental approach to mental health system change. Syst Dyn Rev 13(2):149–169
29. Lane DC (1992) Modelling as learning: A consultancy methodology for enhancing learning in management teams. Eur J Oper Res 59(1):64–84. Special issue of Eur J Oper Res: Morecroft JDW, Sterman JD (eds) Modelling for learning
30. Lane DC, Oliva R (1998) The greater whole: Towards a synthesis of system dynamics and soft systems methodology. Eur J Oper Res 107(1):214–235
31. Lee DB (1973) Requiem for large-scale models. J Am Inst Plan 39(1):163–178
32. Luna-Reyes L, Martinez-Moyano I, Pardo T, Cresswell A, Andersen D, Richardson G (2006) Anatomy of a group model-building intervention: Building dynamic theory from case study research. Syst Dyn Rev 22(4):291–320
33. Maxwell T, Andersen DF, Richardson GP, Stewart TR (1994) Mental models and dynamic decision making in a simulation of welfare reform. In: Proceedings of the 1994 international system dynamics conference, Stirling, Scotland. Social and Public Policy. System Dynamics Society, Lincoln, pp 11–28
34. Meadows DH, Robinson JM (1985) The electronic oracle: Computer models and social decisions. Wiley, Chichester/New York
35. Mingers J, Rosenhead J (2004) Problem structuring methods in action. Eur J Oper Res 152:530–554
36. Morecroft JDW, Sterman JD (1992) Modelling for learning. Special issue Eur J Oper Res 59(1):28–41
37. Morton A, Ackermann F, Belton V (2003) Technology-driven and model-driven approaches to group decision support: focus, research philosophy, and key concepts. Eur J Inf Syst 12:110–126
38. Nunamaker JF, Dennis AR, Valacich JS, Vogel DR, George JF (1991) Electronic meetings to support group work. Commun ACM 34(7):40–61
39. Otto PA, J Struben (2004) Gloucester fishery: Insights from a group modeling intervention. Syst Dyn Rev 20(4):287–312
40. Pawson R, Tilley N (1997) Realistic evaluation. Sage, London
41. Pervan G, Lewis LF, Bajwa DS (2004) Adoption and use of electronic meeting systems in large Australian and New Zealand organizations. Group Decis Negot 13(5):403–414
42. Petty RE, Cacioppo JT (1986) The elaboration likelihood model of persuasion. Adv Exp Soc Psychol 19:123–205
43. Randers J (1977) The potential in simulation of macro-social processes, or how to be a useful builder of simulation models. Gruppen for Ressursstudier, Oslo
44. Richardson GP (1991) Feedback thought in social science and systems theory. University of Pennsylvania Press, Philadelphia
45. Richardson G (2006) Concept models. Größler A, Rouwette EAJA, Langer RS, Rowe JI, Yanni JM (eds) Proceedings of the 24th International Conference of the System Dynamics Society, Nijmegen
46. Richardson GP, Andersen DF (1995) Teamwork in group model-building. Syst Dyn Rev 11(2):113–137
47. Richardson GP, Andersen DF, Maxwell TA, Stewart TR (1994) Foundations of mental model research. Proceedings of the 1994 international system dynamics conference, Stirling, Scotland. pp 181–192. System Dynamics Society, Lincoln
48. Richmond B (1987) The strategic forum: From vision to strategy to operating policies and back again. High Performance Systems Inc, Highway, Lyme
49. Richmond B (1997) The strategic forum: Aligning objectives, strategy and process. Syst Dyn Rev 13(2):131–148
50. Roberts EB (1978) Strategies for effective implementation of complex corporate models. In: Roberts EB (ed) Managerial applications of system dynamics. Productivity Press, Cambridge, pp 77–85
51. Rohrbaugh JW (1992) Collective challenges and collective accomplishments. In: Bostron RP, Watson RT, Kinney ST (eds) Computer augmented team work: A guided tour. Van Nostrand Reinhold, New York, pp 299–324
52. Rouwette EAJA (2003) Group model building as mutual persuasion. Ph D Dissertation, Radboud University Nijmegen
53. Rouwette EAJA, Vennix JAM (2006) System dynamics and organizational interventions. Syst Res Behav Sci 23(4):451–466
54. Rouwette EAJA, Vennix JAM (2007) Team learning on messy problems. In: London M, Sessa VI (eds) Work group learning: Understanding, improving & assessing how groups learn in organizations. Lawrence Erlbaum Associates, Mahwah, pp 243–284
55. Rouwette EAJA, Vennix JAM, van Mullekom T (2002) Group model building effectiveness: A review of assessment studies. Syst Dyn Rev 18(1):5–45
56. Rouwette EAJA, Größler A, Vennix JAM (2004) Exploring influencing factors on rationality: A literature review of dynamic decision-making studies in system dynamics. Syst Res Behav Sci 21(4):351–370

57. Rouwette EAJA, Vennix JAM, Felling AJA (2008) On evaluating the performance of problem structuring methods: An attempt at formulating a conceptual framework. Paper accepted for Group Decision and Negotiation
58. Schein EH (1987) Process consultation, vol II. Addison-Wesley, Reading
59. Shields M (2001) An experimental investigation comparing the effects of case study, management flight simulator and facilitation of these methods on mental model development in a group setting. In: Hines JH, Diker VG, Langer RS, Rowe JI (eds) Proceedings of the 20th international conference of the system dynamics society, Atlanta
60. Shields M (2002) The role of group dynamics in mental model development. In: Davidsen PI, Mollona E, Diker VG, Langer RS, Rowe JI (eds) Proceedings of the 20th international conference of the system dynamics society, Palermo
61. Stenberg L (1980) A modeling procedure for public policy. In: Randers J (ed) Elements of the system dynamics method. Productivity Press, Cambridge, pp 292–312
62. Sterman JD (1994) Learning in and about complex systems. Syst Dyn Rev 10(2–3):291–330
63. van den Belt M (2004) Mediated modeling. A system dynamics approach to environmental consensus building. Island Press, Washington
64. Vennix JAM (1996) Group model-building: Facilitating team learning using system dynamics. Wiley, Chichester
65. Vennix JAM (1999) Group model building: Tackling messy problems. Syst Dyn Rev 15(4):379–401
66. Vennix JAM, Gubbels JW, Post D, Poppen HJ (1990) A structured approach to knowledge elicitation in conceptual model-building. Syst Dyn Rev 6(2):31–45
67. Vennix JAM, Scheper W, Willems R (1993) Group model-building: What does the client think of it? In: Zepeda E, Machuca J (eds) The role of strategic modelling in international competitiveness. Proceedings of the 1993 international system dynamics conference Mexico, Cancun, pp 534–543
68. Vennix JAM, Akkermans HA, Rouwette EAJA (1996) Group model-building to facilitate organizational change: An exploratory study. Syst Dyn Rev 12(1):39–58
69. Vennix JAM, Andersen DF, Richardson GP (eds) (1997) Special issue on group model building. Syst Dyn Rev 13(2):187–201
70. Verburgh LD (1994) Participative policy modelling: Applied to the health care industry. Ph D Dissertation, Radboud University Nijmegen
71. Warren K (2000) Competitive strategy dynamics. Wiley, Chichester
72. Warren K (2005) Improving strategic management with the fundamental principles of system dynamics. Syst Dyn Rev 21(4):329–350
73. Watt CH (1977) Why won't anyone believe us? Simulation 28:1–3
74. Weil HB (1980) The evolution of an approach for achieving implemented results from system dynamic projects. In: Randers J (ed) Elements of the system dynamics method. MIT, Cambridge
75. Winch GW (1993) Consensus building in the planning process: Benefits from a "hard" modelling approach. Syst Dyn Rev 9(3):287–300
76. Wolstenholme EF (1992) The definition and application of a stepwise approach to model conceptualisation and analysis. Eur J Oper Res 59:123–136
77. Zagonel AA (2004) Reflecting on group model building used to support welfare reform in New York state. Unpublished doctoral dissertation, SUNY, Albany

Books and Reviews

Meadows DH, Richardson J, Bruckmann G (1982) Groping in the dark: The first decade of global modelling. Wiley, Chicester
Schwartz RM (1994) The skilled facilitator: Practical wisdom for developing effective groups. Jossey-Bass, San Francisco

Growth Models for Networks

Sergey N. Dorogovtsev[1,2]
[1] Departamento de Física, Universidade de Aveiro, Aveiro, Portugal
[2] A. F. Ioffe Physico-Technical Institute, St. Petersburg, Russia

Article Outline

Glossary
Definition of the Subject
Introduction
Equilibrium and Nonequilibrium Random Networks
Random Recursive Graphs (Trees)
Recursive Trees Versus Equilibrium Connected Trees
The Barabási–Albert Model
General Preferential Attachment
Finite Size Effects
Hidden Variables
Condensation of Edges
Weighted Growing Networks
Connected Components in Growing Networks
Significance of Loops
Accelerated Growth of Networks
Critique of the Preferential Attachment Mechanism
Optimization-Based Models
Deterministic Graphs
Future Directions
Acknowledgments
Bibliography

Glossary

Random network A random network is a statistical ensemble: a given set of graphs and their statistical weights – realization probabilities. Formulating a network model, one defines the full set of members of the ensemble and indicates corresponding statistical weights or rules generating these weights.

Nonequilibrium random network Nonequilibrium random networks are nonequilibrium statistical ensembles of graphs. A growing network is a very particular case of a nonequilibrium one, with a growing number of edges and vertices.

Recursive network A recursive network is a growing network, where new edges connect only new vertices to already existing ones. In recursive networks, no new edges emerge between already existing vertices.

Tree A tree is a graph without loops (cycles). The absence of loops crucially simplifies the description of these networks.

Random recursive graph (tree) In graph theory, the random recursive graph is a tree growing in the following way: at each time step, a new vertex is attached to one of the existing vertices chosen with equal probability.

Degree Degree – a total number of connections of a vertex – is the simplest local property of a network. A vertex degree distribution is the probability that a vertex in a network has a given degree.

Heavy-tailed degree distributions These are degree distributions with divergent higher moments in the infinite network limit. In particular, power-law distributions are heavy tailed.

Scale-free networks In many natural and artificial networks, a sufficiently wide region of a degree distribution may be approximately fitted by a power law. These networks are called scale-free.

Preferential attachment In the preferential attachment process, vertices for connection/attachment are chosen preferentially – usually, with probability proportional to a given function of their degree. The preferential attachment mechanism effectively generates networks with complex architectures including networks with heavy-tailed degree distribution.

Definition of the Subject

For several decades mathematicians, biologists, and physicists have studied growing networks and their models. In physics and physical chemistry, models of this kind were first proposed in the theory of polymers [1,2]. In graph theory and its applications, e. g., to biology and to computer science, very simple growing networks – random recursive trees – were an issue of wide interest for many years, see for example early articles [3,4]. Growth models for networks have become, maybe, the hottest topic in statistical mechanics, graph theory, and multidisciplinary research after the work [5], where these models were used to explain universal complex structures of the Internet, the World Wide Web, and other real networks.

Introduction

Numerous natural and artificial networks have essentially more complex architectures than classical random graphs in graph theory. The classical random graphs have a rapidly decreasing Poisson distribution of the number of connections of a vertex – degree distributions. In contrast, real-world networks have degree distributions decaying in a much slower fashion. Usually, networks with a more complex structure than the classical random graphs are called complex networks. Complex networks include practically all important real-world nets: various cellular networks, the Internet, the World Wide Web, social networks, and many others.

Why do real-world networks have complex architectures? This fundamental problem essentially determines the evolution of a multidisciplinary research field, which is often called the science of complex networks. A few concepts explaining complex network structures were developed. Two of them – self-organization and optimization – were extensively studied during the last years. Modeling the creation of complex network structures is mostly based on these concepts. A significant part of real networks are evolving, usually growing networks, which strongly differ from equilibrium nets. Furthermore, technically, it is easier to arrive at heavy-tailed distributions of connections in evolving networks than in equilibrium ones. These are the main reasons for the intense interest in growth models for networks.

In this very brief survey we only discuss basic demonstrative models highlighting the set of key ideas of this field. A huge number of growth models, including more realistic ones, remain out of the scope of this article. For the sake of demonstration, we discuss only networks which are undirected (networks with undirected connections) and one-partite (there is only one sort of vertices). Many important real networks are directed, the World Wide Web, for example. Chemical reaction networks and many others are bipartite. Fortunately, generalizations to directed and multi-partite networks are rather straightforward.

Equilibrium and Nonequilibrium Random Networks

In statistical mechanics and graph theory, a random network is not a single graph but a statistical ensemble. A given set of graphs are members of this ensemble. Each of these graphs has its probability of realization – statistical weight. These weights may be stationary or may be evolving, that is depending on time t. One of the possible ways to define a statistical ensemble is by a master equation for statistical weights. In a wide range of situations, these

equations are linear. Let $\mathcal{P}(g,t)$ be the statistical weight of a graph g belonging to the set of graphs of a statistical ensemble G, $g \in G$, and $W(g,g')$ be transition rates from graph g' to graph g. Then the master equation is of the form

$$\frac{\partial \mathcal{P}(g,t)}{\partial t} = \sum_{g' \in G} [W(g,g')\mathcal{P}(g',t) - W(g',g)\mathcal{P}(g,t)]. \quad (1)$$

In equilibrium random networks, the left-hand side of the master equation is zero, and the right-hand side provides one with a set of detailed balance conditions for stationary statistical weights $\mathcal{P}(g)$. In nonequilibrium networks, $\partial\mathcal{P}(g,t)/\partial t \neq 0$, and a detailed balance is not reached. Growing networks represent a particular case of nonequilibrium ones, and in many network growth models, equations describing the evolution of statistical properties of a network are simple consequences of Eq. (1). Some characteristics of a growing network may become stationary as $t \to \infty$. This does not mean that the network approaches an equilibrium state in this limit. For example, in numerous network growth models, a degree distribution approaches a stationary limit, which is very convenient for their analysis.

Random Recursive Graphs (Trees)

In recursive networks new edges connect only new vertices to already existing vertices, new connections between already existing vertices are not possible. Most simple growth models for networks are recursive. The simplest recursive network is *the random recursive graph* from graph theory [6]. In the random recursive graph, the growth starts with a single vertex. At each time step,

(i) add a new vertex and
(ii) attach it to one of the already existing vertices chosen with equal probability.

The random recursive graph has no loops – it is a tree by definition. This model is one of two fundamental random graphs in graph theory: the Erdős–Rényi graph [7] is a basic equilibrium network, and the random recursive graph is a basic growing network.

The degree distribution $P(q)$ of the random recursive graph (the average fraction of vertices of degree q) decays exponentially. In the infinite network limit,

$$P(q) = 2^{-q-1}, \quad (2)$$

which is valid for all degrees. This distribution decreases much slower than the Poisson degree distribution of the Erdős–Rényi model, which asymptotically decays as $1/q!$. On the other hand, all moments of this distribution are finite.

The absence of loops allows one to easily describe the global organization of the network. The first characteristic of the global organization is *a mean intervertex distance* $\overline{\ell}$ – the average length of the shortest path between two vertices in the network. The dependence of $\overline{\ell}$ on the total number of vertices in the network, N, characterizes the network compactness. In finite-dimensional networks – "large worlds" – $\overline{\ell} \sim N^{1/d}$, where d may be regarded as the dimensionality of a network (more rigorously, its Hausdorff dimension). In networks with the small-world property, which are infinite-dimensional dimensional objects – "small-worlds" – $\overline{\ell}$ increases with N slower than any positive power of N. In the random recursive tree,

$$\overline{\ell}(N) = \frac{2}{N-1}[(N+1)H_N - 2N] \overset{N\gg 1}{\cong} 2\ln N - (4-2\gamma_e) \quad (3)$$

(see [8] and references therein) so that this network is a small world. Here $H_N = \sum_{i=1}^{N} 1/i$ is the Nth harmonic number and $\gamma_e = 0.5772\ldots$ is Euler's number. The average distance of vertices from the root, approaching $(1-\gamma_e)\ln N$ as $N \to \infty$, and the diameter (the average maximum distance between two vertices in a graph of this statistical ensemble) behave similarly.

Remarkably, the relative width of the intervertex distance distribution, $\mathcal{R}(\ell, N)$, of the random recursive tree approaches zero in the infinite network limit. The width of this distribution grows as $\sqrt{\overline{\ell}} \sim \sqrt{\ln N} \ll \overline{\ell}$, and $\mathcal{R}(\ell, N) \to \delta(\ell - \overline{\ell}(N))$ as $N \to \infty$. Here $\delta(x)$ is the δ-function. This property is rather typical for small worlds.

Recursive Trees Versus Equilibrium Connected Trees

A difference between growing and equilibrium networks may be striking. For the sake of demonstration, let us compare two random networks – the random recursive trees and the equilibrium (or homogeneous) connected trees, [9]. The equilibrium statistical ensemble of connected trees is defined in the following way. Each of its members is a labeled connected tree. (The term "a labeled network" means the following. All vertices in these networks have individual labels. Labeled networks which differ only by permutations of the labels are counted as distinct. We discuss only labeled networks as is usual in statistical mechanics and graph theory.) The ensemble includes all possible labeled connected trees of a given size N. All the members have equal probabilities of realization. It turns out that the mean intervertex distance of this

random network scales with N as $\bar{\ell}(N) \sim N^{1/2}$, i. e., this kind of network may be regarded as a two-dimensional object. This is in a sharp contrast to the random recursive trees, which are small worlds, as we explained above. In the recursive tree ensemble, only those connected labeled trees are present that are allowed by the causality of the construction. So the great majority of the members – labeled trees – of the equilibrium tree ensemble are absent in the random recursive tree. This results in a quite different global organization of this equilibrium network. As for local properties, the degree distribution of equilibrium connected trees is $P(q) = \mathrm{e}^{-1}/(q-1)!$, which is a much more rapidly decaying function than the exponential degree distribution (2) of the random recursive trees.

The Barabási–Albert Model

The preferential attachment process is a very straightforward way to generate networks with numerous hubs. In this process, highly connected vertices receive new connections with higher probability ("popularity is attractive"); as a result, they become even more attractive for connection, get new edges, and so on – a network organizes itself. Actually, this very general idea is not new. In a non-network aspect, Yule in 1925, [10], and later Simon, [11], used this idea to explain skewed distributions in nature. Price applied this self-organization mechanism to scientific citations [12]. Nonetheless, it was the very timely work of Barabási and Albert [5] (see also [13]) that triggered an avalanche of studies using the preferential attachment to explain complex networks.

The Barabási–Albert model is a recursive graph, growing according to the following rules. The growth starts from some particular initial configuration of vertices and edges, which is not very important. At each time step,

(i) add a new vertex and
(ii) attach it to $m \geq 1$ preferentially selected already existing vertices. Each of these vertices is chosen with a probability proportional to its degree – proportional preference.

That is, in the Barabási–Albert model, the probability of attachment to vertex i of degree q_i is $\Pi_i = q_i / \sum_{j=1}^{N} q_j$. In particular, if the initial configuration is a single vertex, and $m = 1$, i. e. each new vertex has only one connection, the Barabási–Albert network is a nonuniform recursive tree [14]. However, it turns out that the resulting degree distribution does not depend on m (we ignore a normalization factor), and in this respect loops generated for $m > 1$ are not important. The degree distribution approaches a stationary form in the large network limit,

$$P(q) = \frac{2m(m+1)}{q(q+1)(q+2)} \overset{q\to\infty}{\sim} q^{-3} \,. \tag{4}$$

This degree distribution is heavy-tailed, its second and higher moments diverge.

One of the simple ways to obtain the proportional preference is as follows. Choose at random an edge and attach a new vertex to one or to both ends of this edge [15]. One can see that the end vertices will indeed be selected with probability proportional to their degrees. Furthermore, the attachment of a new vertex to all nearest neighbors of a randomly chosen vertex leads to the same effect – proportional preference. In a more specific situation, in uncorrelated networks, one can arrive at the proportional preference also in the following way: choose a vertex at random and then choose at random one of its nearest neighbors for attachment. The last algorithm is sometimes applicable, since the proportional preferential attachment generates recursive networks with very weak correlations (see below).

In respect of degree–degree correlations, the Barabási–Albert model is very special. The simplest degree correlations in a network are correlations between degrees of the nearest neighbor vertices. These correlations are present if the joint distribution of the degrees of the nearest neighbors does not factor in the following way: $P(q, q') \neq qP(q)q'P(q')/\langle q\rangle^2$. If there is equality, the network is uncorrelated. Empirically, it is convenient to use a less informative object for these correlations. This is the dependence of the mean degree of the nearest neighbors of a vertex versus degree of this vertex, $\langle q\rangle_{\mathrm{nn}}(q)$ [16,17]. In uncorrelated networks, $\langle q\rangle_{\mathrm{nn}}$ does not depend on q. These correlations were observed in practically all real-world networks. We stress that only equilibrium networks may be uncorrelated. Growing (and, more general, nonequilibrium) networks, including the Barabási–Albert model and random recursive graphs, have these correlations. However, in the Barabási–Albert model and other recursive models based on the proportional preferential attachment, the degree–degree correlations are anomalously weak, and so the dependence $\langle q\rangle_{\mathrm{nn}}(q)$ is nearly flat. This is a quite unusual situation for growing networks.

When $m > 1$, this network has loops (or cycles in terms of graph theory). A loop is a closed path visiting each vertex only once. Loops of length 3 (triangles) are directly related to *clustering*. The clustering coefficient C_i of vertex i is the ratio of the number n_i of triangles attached to this vertex and $q_i(q_i-1)/2$ which is the maximum possible number of these triangles. Here q_i is the degree of vertex i. *The mean clustering* is the average clustering co-

efficient of a vertex:

$$\langle C\rangle = \sum_q P(q)C(q)\,. \tag{5}$$

The clustering coefficient of a network is

$$\begin{aligned} C &= \frac{\sum_q P(q)C(q)q(q-1)}{\sum_q P(q)q(q-1)} \\ &= \frac{3\times(\text{total number of triangles})}{\text{number of connected triples of vertices}}\,. \end{aligned} \tag{6}$$

In uncorrelated networks, $C = \langle C\rangle$. In growing networks, this is not the case. The mean number of triangles in the Barabási–Albert model is $(m-1)m(m+1)(\ln N)^3/48$, while the mean number of pairs of adjacent edges is $m(m+1)(N\ln N)/2$ [18]. So the clustering coefficient of the Barabási–Albert model equals

$$C = \frac{m-1}{8}\,\frac{(\ln N)^2}{N}\,. \tag{7}$$

The mean clustering differs from expression (7) only by a factor, $\langle C\rangle \sim (\ln N)^2/N$ [19]. Thus, a nonzero clustering coefficient in these networks is only a finite size effect: C and $\langle C\rangle$ approach 0 as $N\to\infty$. More generally, the number of finite loops of length L in the Barabási–Albert model is also relatively small, $\sim(\ln N)^L$ as $N\to\infty$ [20]. This indicates that this network has a locally tree-like structure.

The Barabási–Albert model is a small world. However, the size dependence of the mean intervertex distance (and of diameter) noticeably differs for different m. As $N\to\infty$,

$$\bar{\ell}(N, m=1)\sim \ln N, \quad \bar{\ell}(N, m>1)\sim \frac{\ln N}{\ln\ln N} \tag{8}$$

[21], which demonstrates the importance of loops in the network.

General Preferential Attachment

In general, the attachment probability may be some given function $f(q)$ of the degree of a vertex. For linear preference, $f(q) = q + A$, where the constant $A \equiv am > -m$ is an additional attractiveness, the resulting network is scale-free. The asymptotics of the degree distribution is power-law, $P(q)\sim q^{-\gamma}$ in the infinite network limit, where exponent γ is

$$\gamma = 3 + \frac{A}{m}\,. \tag{9}$$

So γ can be in the range from 2 to ∞. For an arbitrary q and $N\to\infty$,

$$\begin{aligned} P(q) &= (2+a)\frac{\Gamma(m(1+a)+a+2)}{\Gamma(m(1+a))} \\ &\quad\cdot\frac{\Gamma(k+ma)}{\Gamma(k+ma+3+a)} \overset{q\to\infty}{=\!\sim} q^{-\gamma} \end{aligned} \tag{10}$$

[22], where $\Gamma(x)$ is the gamma-function. At $a = A/m = 0$, we arrive at the degree distribution (4) of the Barabási–Albert model.

A similar effect – changing exponent γ – can be obtained by combining random attachment and preferential one, with proportional preference. Say, for each new connection, with a probability p, choose a vertex at random and, with the complementary probability $1-p$, choose a vertex with probability proportional to its degree. In this case, $\gamma = 1 + 2/(1-p)$, for details and more realistic examples see book [23].

A more detailed characteristic in these networks is a degree distribution of a vertex born at time s, which we denote by $p(q, s, t)$. The age of the network asymptotically approaches its size, $t \cong N$, if the growth starts from a finite cluster. The following scaling form is valid for growing scale-free networks:

$$p(q, s, t) = (s/t)^\beta f(q(s/t)^\beta)\,, \tag{11}$$

where $f(x)$ a scaling function, and exponent β is related to the γ exponent of the degree distribution [22]:

$$\beta = \frac{1}{\gamma-1}\,. \tag{12}$$

Usually, even in growing networks with wide degree distributions, the function $f(x)$ has a narrow peak form. The mean degree of a vertex born at time s measured in a scale-free network of large age t is asymptotically expressed as

$$\langle q\rangle(s, t) \propto \left(\frac{s}{t}\right)^{-\beta}\,. \tag{13}$$

We have mentioned that these networks have strong correlations between degrees of the nearest neighbors [24]. In particular, in the scale-free recursive networks,

$$\langle q\rangle_{\rm nn}(q) \propto q^{\gamma-3}\,. \tag{14}$$

This demonstrates that

(i) if $\gamma > 3$, the correlations are *assortative* on terms of [25], that is the nearest neighbors of a strongly connected vertex, as a rule, also have many connections and vice versa;
(ii) if $\gamma < 3$, the correlations are *disassortative*, that is the nearest neighbors of a strongly connected vertex, as a rule, have small degrees and vice versa; and
(iii) in the unique case of $\gamma = 3$, as in the Barabási–Albert model, these correlations are very weak.

More general forms of the preference function were studied, but it turned out that only the linear preference pro-

vides power-law degree distributions. In particular, networks with a power-law preference $f(q) = q^y$ were investigated, [26]. When $0 \le y < 1$, the resulting degree has a stretched exponential form:

$$P(q) \propto q^{-y} \exp\left[-g(y)q^{1-y}\right] , \tag{15}$$

where $g(y)$ depends only on exponent y. On the other hand, if $y > 1$, a small number of the oldest vertices attract all connections. In particular, if $y > 2$, there is even a finite probability that the first vertex gets all links, and the network is a star.

In networks, which we discussed, once added, vertices never disappear. The network evolution, however, may combine two parallel processes – both the addition and removal of vertices. Remarkably, the deletion of vertices with all their connections results in the preferential removal of edges from vertices with many connections. Similarly to Sect. "The Barabási–Albert Model", the probability of this removal is proportional to degree – proportional preference. In a network model studied in work [27], (i) new vertices are attached to preferentially chosen existing vertices and, simultaneously, (ii) randomly selected vertices with all their connections are removed from the network. The form of the resulting degree distribution depends on the relation between the rates of these processes and also on the mean degree of the network. The exponent of the degree distribution grows as the removal rate increases. Moreover, at sufficiently high removal rates, the degree distribution may become exponentially decaying. Networks, in which connections of a removed vertex are not deleted but are redistributed in the neighborhood of this vertex, were described in [28]. The degree distributions of these networks also vary from a power law to an exponential one.

Finite Size Effects

Even the largest artificial network, the World Wide Web, is actually small – between 10^{10} and 10^{11} vertices (Web pages). So finite size effects in networks are of primary importance. Here we only touch upon the size dependence of a degree distribution in scale-free recursive networks. It readily follows from Eq. (11) that the degree distribution is of the form:

$$P(q, N) = q^{-\gamma} F(q/N^{1/(\gamma-1)}) , \tag{16}$$

where $F(x)$ is an exponentially rapidly decreasing scaling function whose precise form, in principle, depends on initial conditions. A typical scaling function $F(x)$ for exponent $\gamma = 3$ is shown in Fig. 1 from [15]. The scaling functions for arbitrary γ were computed in Refs. [29,30].

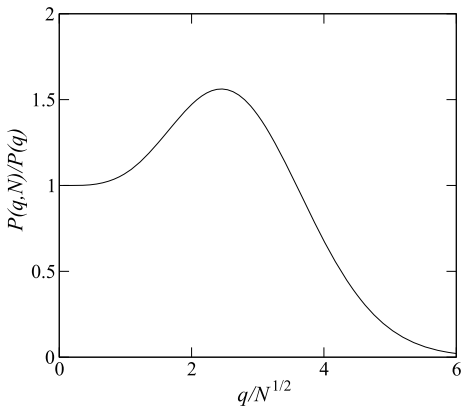

Growth Models for Networks, Figure 1
Ratio $P(q, N)/P(q, N\to\infty)$ versus the scaling variable $q/N^{1/2}$ for a scale-free recursive network with degree distribution exponent 3 [15]. The growth starts with a pair of vertices with a connecting edge. With other initial configurations, the form of the hump will be different

Network finiteness leads to a cutoff of the degree distribution, $q_{\text{cut}}(N)$. In the scale-free recursive networks, $q_{\text{cut}}(N) \sim N^{1/(\gamma-1)}$ for arbitrary γ, as Eq. (16) shows. This dependence coincides with a so-called "natural cutoff", whose meaning is clear from the following estimate: $N \int_{q_{\text{cut}}(N)}^{\infty} \mathrm{d}q P(q) \sim 1$, where $P(q) = P(q, N\to\infty)$. That is, in a single realization of a network, only one vertex of degree above the natural cutoff value may be found, and so such vertices are practically unobservable. Note, however, that the cutoff is essentially model dependent, and in some models its position is far below the natural cutoff value.

Hidden Variables

Equation (13) shows that it is the oldest vertices that become hubs in the Barabási–Albert model and other scale-free recursive networks, where a preference function is $f(q) = Bq + A$ with constant A and B. This is certainly not the case in real-world networks. Google, for example, has received a far greater number of hyperlinks than a huge majority of earlier created Web sites. There is a simple way to weaken "the older the richer effect" in growing networks, which was proposed in [31]. Suppose that B and A be random variables with values varying from vertex to vertex, $A(s)$ and $B(s)$, where $s = 0, 1, \ldots, t$ is a label of a vertex. These $A(s)$ and $B(s)$ are usually called *hidden variables*, the term *fitness* is also used. It turns out that under some conditions, even with random $A(s)$ and $B(s)$, a degree distribution remains heavy tailed, see [31] and, for more detail, [32]. On the other hand, the randomness of the hidden variables leads to strong fluctuations of degrees

of vertices with a given birth time. Even a young vertex s with large $A(s)$ or/and $B(s)$ may become rapidly a strongly connected hub.

More generally, a random preference function for a recursive graph may be a function of three variables: (i) the degree q of a vertex, which is selected for attachment, (ii) its birth time s, and (iii) the age t of the network. That is, in general, one should consider $f(q, s, t)$. In principle, it is possible to arrive at heavy-tailed degree distributions even when the preference function is independent of q. For example, a power-law $f(s, t) \propto (s + 1)^{-\alpha}$ results in scale-free networks. Other forms of $f(q, s, t)$ allow one, in particular, to take into account aging of vertices, where $f(q, s, t) = g(q, t - s)$ [33].

Condensation of Edges

If the coefficient $B(s)$ of the preference function strongly fluctuates, a new phenomenon may occur [34]. The large fluctuations of $B(s)$ may lead to condensation of edges on a few vertices in the growing network. Namely, a finite fraction of connections in a large network turns out to be attracted by a few vertices (or even by a single one).

This effect is observable even when only a single vertex in a recursive network, say, vertex u, has a fitness essentially greater than others: $B(s) = 1 + g\delta_{su}$. Here δ_{su} is a Kronecker symbol. If g is greater than some critical value $g_c = \gamma_0 - 1$, where γ_0 is the degree distribution exponent in the "pure" network with $g = 0$, then a finite fraction of edges will be attracted by vertex u. This fraction is proportional to the difference $g - g_c$. Other vertices have a power-law degree distribution with exponent $\gamma = 1 + g > \gamma_0$.

Weighted Growing Networks

The edges of weighted networks have their own *weights* – positive real numbers. The introduction of the weights allows one to take into account the diversity of edges in many real networks [35]. For example, transport networks are weighted. Let w_{ij} be the weight of the edge connecting vertices i and j. Then *the strength* of vertex i is defined as the sum of the weights of its edges: $s_i = \sum_j w_{ij}$. Empirical data show that in many real-world weighted networks both the strength distribution and the degree distribution are power-law [35,36]. In some of these networks, the strength of a vertex, in average, is proportional to its degree, $s_i \propto k_i$, and in other ones, $s_i \propto k_i^\theta$ with exponent $\theta > 1$.

A simple growth model for scale-free weighted networks is defined as follows [35,36]. At each time step,

(i) add a new vertex and attach it to one or several preferentially chosen existing vertices by edges with some initial weight, say 1; the probability to select vertex i is taken to be proportional to its degree s_i;
(ii) in addition, increase the weights of the edges of each of these selected vertices: for vertex i, the edge weights are increased as $w_{ij} \to w_{ij}(1 + \delta/s_i)$, where δ is a given positive parameter.

Thus, the attachment to a vertex increases its degree by 1 and its strength by $1 + \delta$. In the resulting network, the strengths of vertices are proportional to their degrees, $s_i \propto k_i$, and the strength and degree distributions are power laws with the same exponent

$$\gamma = 2 + \frac{1}{1 + 2\delta} \,. \tag{17}$$

One can arrive at the same results in another way [37]: choose an edge with a probability proportional to its weight, increase this weight by δ, and simultaneously attach a new vertex by a unit weight link to one of the ends of this selected edge.

Bianconi showed how to construct a growing network with a power-law relation between vertex strength and degree [38]. Simply combine two kinds of the preferential attachment. Namely, new vertices should be attached at random either to vertices selected with a probability proportional to a vertex strength or to vertices selected with a probability proportional to a vertex degree.

Connected Components in Growing Networks

So far we have discussed networks consisting of a single connected component, which is a rather special situation. All vertices in these networks are mutually approachable. In general, a network consists of a set of connected components of various sizes. Typically, in a random network, either (i) all connected components are relatively small (much smaller than N) or (ii) one of the components is "giant" and the rest are relatively small. The emergence of *the giant connected component* (containing, by definition, a finite fraction of all vertices in an infinite net) crucially changes the architecture of a network. Furthermore, the statistics of connected components essentially characterizes the network organization. These statistics for growing random networks may remarkably differ from those for equilibrium networks.

A specific transition of the birth of the giant connected component in a growing network was discovered in work [39], where a very simple network was studied. The growth starts from some initial configuration, which

is not important over long periods. The network grows due to two parallel processes:

(i) With the unit rate, bare vertices are added to the network.
(ii) With a rate b, new edges interconnect randomly chosen vertices.

The rate b plays the role of a control parameter. When b is greater than some critical value b_c, a giant connected component is present. The degree distribution of this random graph has an exponential decay. In contrast to equilibrium networks, the birth of the giant component in this model is an infinite order phase transition. At the critical point, the size of the giant connected component has a specific singularity

$$S \propto \exp(-\text{const}/\sqrt{b - b_c}) , \tag{18}$$

coinciding with that in the famous Berezinskii–Kosterlitz–Thouless phase transition. It was found that the result (18) is also valid for growing scale-free networks [40], where the value of the constant in the exponential is determined by exponent γ. This singularity was later observed in many other growing networks [41,42,43,44].

The size distribution of finite components in the networks with this transition has an exponential decay in the presence of the giant connected component. Without a giant component, the component size distribution is a power-law in scale-free and non-scale-free growing networks. For comparison, if an uncorrelated network has a rapidly decreasing degree distribution, then the component size distribution slowly decreases only at the birth point of a giant component.

Significance of Loops

We have emphasized that trees are simple for analysis. However, it is loops that make networks really interesting. In numerous studied random networks, in the infinite network limit, any finite environment of a vertex has no loops, i. e., it is tree-like, and the so-called *tree ansatz* is valid for many problems. It turns out, however, that it is infinite loops that are important for the global organization of networks. In sparse random recursive networks, there are usually few finite loops and many loops of length greater than a network diameter. These large loops may make a loopy network much more compact than a tree.

Let us compare a mean intervertex distance $\overline{\ell}(N)$ in the two kinds of scale-free recursive networks: trees ($m = 1$) and sparse networks with large loops ($m > 1$), see Sects. "The Barabási–Albert Model" and "General Preferential Attachment".

(i) *Recursive trees*. For any $\gamma > 2$, asymptotically $\overline{\ell}(N) \sim \ln N$ [45,46].
(ii) *Sparse recursive networks with loops*.
  (a) When $\gamma > 3$, $\overline{\ell}(N) \sim \ln N$.
  (b) When $\gamma = 3$, $\overline{\ell}(N) \sim \ln N/\ln\ln N$ [21].
  (c) When $2 < \gamma < 3$, we expect $\overline{\ell}(N) \sim \ln\ln N$ or, maybe, another slow dependence of this sort (in this range, we know only estimates for uncorrelated networks [47]).

Thus, the difference in $\overline{\ell}(N)$ in these networks is strong if $\gamma \leq 3$.

Accelerated Growth of Networks

In networks, which we discussed above, the number of edges L grows with time proportionally to (or linearly in) the total number of vertices N. Asymptotically, $L/N \cong$ const in large networks – the growth is linear. A nonlinear growth, where the dependence $L(N)$ is nonlinear at large N, is poorly studied. A growing network with a power-law $L \propto N^{1+a}$, where $a < 1$, was considered in [48]. Since the number a in real networks is usually positive, this type of growth is called *accelerated*. The diverse architectures of these networks are strongly model dependent. Their degree distributions may be nonstationary even in the limit $N \to \infty$. Some of these networks are scale-free, with exponent γ in a wide range from 1 to ∞. Cutoffs in these degree distributions strongly differ from those for linearly growing networks.

The accelerated growth essentially changes the form of the size dependence of the mean intervertex distance. Since these networks become more dense with N, the dependence $\overline{\ell}(N)$ may be stationary or even decreasing. Numerous real-world examples of this effect are described in [49].

Critique of the Preferential Attachment Mechanism

The preferential attachment mechanism and, more generally, self-organization models for networks have been criticized in [50]. The authors of that work stressed that these models are "only evocative" and "not explanatory". An evocative model only "can reproduce the phenomenon of interest but does not necessarily capture and incorporate the true underlying cause". In contrast, an explanatory model "also captures the causal mechanisms (why and how, in addition to what)".

This critique was based on two arguments. First, it was demonstrated in [50] that the Barabási–Albert model fails to reproduce many empirical data for real networks. We believe that this part of the critique is not very strong.

Indeed, the Barabási–Albert model is only the simplest model in a wide class of models. There exist much more realistic models with preferential attachment. The second argument is essentially more serious. That is: the preferential attachment concept itself does not provide us with a reason for the preferential attachment. In this respect, the preferential attachment model falls short of explaining the genuine nature of complex network architectures. The authors of [50] suggested that optimization-based models explain complex networks in a more natural and convincing way. Unfortunately, the optimization concept of complex networks is much less developed than the preferential attachment. The difficulty is that typically, numerical optimization of networks is a heavy computational task. On the other hand, few optimization models for complex networks were studied analytically.

Optimization-Based Models

In the optimization-based models, new connections optimize (e. g., minimize) some functional of network characteristics. Optimization may be global or local. In the global optimization, each new connection is made in such a way that it minimizes a function or functional of global network characteristics. For example, this *cost function* may be a linear function of the mean intervertex distance and other global characteristics of a network [51]. In this case, optimization leads to more compact architectures. The global optimization of a network structure is a time consuming computational problem. Only small networks were generated using these algorithms. Their sizes did not allow one to arrive at solid conclusions about their structures.

In the local optimization, for a new connection, we select "the best vertex" in a network. The variables of a cost function in the local optimization are characteristics of a vertex, e. g., a shortest path distance from some other vertex or vertices. For each selected vertex, the cost function is minimum among all vertices of the network.

A local optimization model proposed in [52] allows a strict analytical treatment, since it is a tree. The model is formulated in the following way. The growth starts from a single vertex (root with label 0) at some point of a restricted two-dimensional (which is not that important) area. At each time step,

(i) place a new vertex t at a random point of this area and
(ii) attach it to that existing vertex i where the cost function $\ell_{0,i} + \alpha(t) d_{i,t}$ has the minimum value.

Here $\ell_{0,i}$ is the shortest-path network distance between vertex i and the root – "the depth" of vertex i; $d_{i,t}$ is the Euclidean distance between vertices t and i in this plane; the coefficient $\alpha(t)$ is some given function of the network size $t = N + 1$. One can see that this model is based on "optimized trade-offs" between two conflicting objectives: a network objective and a geographic one – connect to the root (the center of a network) versus connect to the geographically closest vertex. It turns out that when $\alpha(t)$ is a power law, the resulting degree distribution has a power-law region. Unfortunately, depending on a form of the function $\alpha(t)$, either this power-law region is quite narrow or the great majority of vertices are leaves [53]. So this model does not produce a really scale-free network in the limit $N \to \infty$.

Another local optimization model was described in [54]. This is also a tree. The growth starts from the root 0 at some point of a one-dimensional (which is important) interval. At each time step,

(i) Place a new vertex t at a random point of this interval, and
(ii) attach it to that existing vertex i where the cost function $\ell_{0,i} + \alpha n_{i,t}$ has the minimum value.

In this cost function, $\ell_{0,i}$ is the shortest-path network distance between vertex i and the root; $n_{i,t}$ is the number of already existing vertices between vertices t and i in this interval; and the coefficient α is a constant. One can show that this very special optimization process actually generates preferential attachment. The resulting degree distribution has a power-law form with a cutoff depending only on the value of α: $q_{\text{cut}} \sim 1/\alpha$.

Deterministic Graphs

Deterministic growing graphs provide a wide range of complex architectures. These networks are quite simple: for each size, a deterministic network "ensemble" consists of a single member – a single graph. Nonetheless, one can build deterministic graphs with properties surprisingly similar to those of random networks. One can construct deterministic graphs with a small-world feature, deterministic scale-free graphs, where a discrete spectrum of degrees has a power-law envelope, and so on. Figure 2 shows two simple constructions proposed in Refs. [55,56] which are scale-free deterministic small worlds. For other deterministic graphs with a small-world property, see, e. g., Refs. [57,58].

Future Directions

In this brief survey, we discussed only a few basic growth models. For example, we did not mention a mechanism

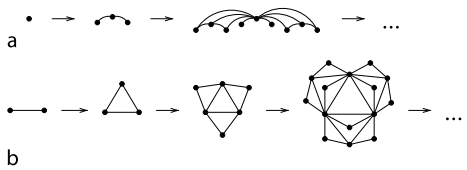

Growth Models for Networks, Figure 2
Two examples of scale-free deterministic graphs. Graphs a and b were introduced in [55] and [56], respectively

based on the merging of vertices [59]. This merging results in strongly connected hubs and power laws in networks. Moreover, we did not touch upon some important network characteristics. Among these missed characteristics, there are for example, the betweenness centrality and its distribution [60,61]. We considered only exponential and power-law degree distributions and omitted more general multifractal distributions. We did not discuss numerous applications of these models and empirical data. The reader should refer to books and reviews in the list of references for more detail.

Here we mention a few works where the growth models for networks were used for analysis and interpretation of empirical data. A link copying model of the World Wide Web was described in [62]. In this model links are directed. Each new vertex connects to one of the nearest neighbors of a randomly chosen vertex, i. e. "copies" the corresponding directed link. This process effectively generates preferential attachments and so leads to heavy-tailed degree distributions. The authors of work [16] applied the preferential attachment concept to the Internet both on the Autonomous System level and for the network of routers. Duplication of proteins also leads to preferential attachment and explains specific architectures of protein interaction networks [42], see also [63]. Several networks of scientific collaborations were empirically studied in the work [64]. With these data, the parameters of the preferential attachment process were found. One should note that collaboration networks may be treated as bipartite graphs where the vertices of the first type are collaborators and the vertices of the second type are acts of collaboration (scientific papers for example). The preferential attachment mechanism has been generalized to networks of this kind in [65].

Finally, we present a short list of research directions which, we believe, give a good perspective.

(i) More realistic models for specific real-world networks. These models should not only reproduce basic network characteristics but also uncover the underlying mechanism.
(ii) Optimization-based models. The goal is to find efficient optimization algorithms generating large networks with heavy-tailed degree distributions.
(iii) Accounting for loops in growing networks. The aim is to find a way to compute the contribution of loops to the properties and characteristics of networks.
(iv) Cooperative models on growing networks, flows, and processes taking place in networks. The problem is how the architectures of growing networks influence phenomena in interacting systems placed on these networks.
(v) Co-evolving systems – networks and interacting agents on them. Interacting agents placed on the top of a network may also influence the structure of the network itself. How do these complex systems evolve? How are they organized?

Acknowledgments

The author thanks A. V. Goltsev, J. F. F. Mendes, and A. N. Samukhin for numerous discussions. This work was partially supported by projects POCI: FAT/46241, MAT/46176, FIS/61665, and BIA-BCM/62662, and DYSONET.

Bibliography

Primary Literature

1. Flory PJ (1941) Molecular size distribution in three-dimensional polymers: I, II, III. J Am Chem Soc 63:83–100
2. Stockmayer WH (1943/1944) Theory of molecular size distribution and gel formation in branched chain polymers. J Chem Phys 11:45–55; 12:125–134
3. Na HS, Rapoport A (1970) Distribution of nodes of a tree by degree. Math Biosci 6:313–329
4. Moon JW (1974) The distance between nodes in recursive trees. Lond Math Soc Lect Notes Ser 13:125–132
5. Barabási A-L, Albert R (1999) Emergence of scaling in complex networks. Science 286:509–512
6. Mahmoud H (1991) Limiting distributions for path lengths in recursive trees. Probab Eng Info Sci 5:53–59
7. Erdős P, Rényi A (1959) On random graphs. Publ Math Debr 6:290–297
8. Dobrow RP (1996) On the distribution of distances in recursive trees. J Appl Prob 33:749–757
9. Bialas P, Burda Z, Jurkiewicz J, Krzywicki A (2003) Tree networks with causal structure. Phys Rev E 67:066106
10. Yule GU (1925) A mathematical theory of evolution based on the conclusions of Dr. JC Willis. Phil Trans Royal Soc Lond B 213:21–87
11. Simon HA (1955) On a class of skew distribution functions. Biometrica 42:425–440
12. Price DJ de S (1976) A general theory of bibliometric and other cumulative advantage processes. J Amer Soc Inform Sci 27:292–306

13. Albert R, Barabási A-L, Jeong H (1999) Mean-field theory for scale-free random networks. Phys A 272:173–187
14. Szymański J (1987) On a nonuniform random recursive trees. Ann Discret Math 33:297–306
15. Dorogovtsev SN, Mendes JFF, Samukhin AN (2001) Size-dependent degree distribution of a scale-free growing network. Phys Rev E 63:062101
16. Pastor-Satorras R, Vázquez A, Vespignani A (2001) Dynamical and correlation properties of the Internet. Phys Rev Lett 87:258701
17. Vázquez A, Pastor-Satorras R, Vespignani A (2002) Large-scale topological and dynamical properties of the Internet. Phys Rev E 65:066130
18. Bollobás B, Riordan O (2003) Mathematical results on scale-free graphs. In: Bornholdt S, Schuster HG (eds) Handbook of Graphs and Networks. Wiley, Weinheim, pp 1–34
19. Klemm K, Eguíluz VM (2002) Growing networks with small-world behavior. Phys Rev E 65:057102
20. Bianconi G, Capocci A (2003) Number of loops of size *h* in growing scale-free networks. Phys Rev Lett 90:078701
21. Bollobás B, Riordan OM (2004) The diameter of a scale-free random graph. Combinatorica 24:5–34
22. Dorogovtsev SN, Mendes JFF, Samukhin AN (2000) Exact solution of the Barabási-Albert model. Phys Rev Lett 85:4633–4636
23. Dorogovtsev SN, Mendes JFF (2003) Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford University Press, Oxford
24. Krapivsky PL, Redner S (2001) Organization of growing random networks. Phys Rev E 63:066123
25. Newman MEJ (2002) Assortative mixing in networks. Phys Rev Lett 89:208701
26. Krapivsky PL, Redner S, Leyvraz F (2000) Connectivity of growing random networks. Phys Rev Lett 85:4629–4632
27. Bauke H, Sherrington D (2007) Topological phase transition in complex networks. arXiv:0710.0831
28. Ben-Naim E, Krapivsky PL (2007) Addition-deletion networks. J Phys A 40:8607–8619
29. Krapivsky PL, Redner S (2002) Finiteness and fluctuations in growing networks. J Phys A 35:9517–9534
30. Waclaw B, Sokolov IM (2007) Finite size effects in Barabási–Albert growing networks. Phys Rev E 75:056114
31. Bianconi G, Barabási A-L (2001) Competition and multiscaling in evolving networks. Europhys Lett 54:436–442
32. Ergun G, Rodgers GJ (2002) Growing random networks with fitness. Phys A 303:261–272
33. Dorogovtsev S, Mendes JFF (2000) Evolution of networks with aging of sites. Phys Rev E 62:1842–1845
34. Bianconi G, Barabási A-L (2001) Bose–Einstein condensation in complex networks. Phys Rev Lett 86:5632–5635
35. Barrat A, Barthélemy M, Pastor-Satorras R, Vespignani A (2004) The architecture of complex weighted networks. Proc Natl Acad Sci 101:3747–3752
36. Barrat A, Barthélemy M, Vespignani A (2004) Weighted evolving networks: Coupling topology and weights dynamics. Phys Rev Lett 92:228701
37. Dorogovtsev S, Mendes JFF (2005) Evolving weighted scale-free networks. AIP Conf Proc 776:29–36
38. Bianconi G (2005) Emergence of weight-topology correlations in complex scale-free networks. Europhys Lett 71:1029–1035
39. Callaway DS, Hopcroft JE, Kleinberg JM, Newman MEJ, Strogatz SH (2001) Are randomly grown graphs really random? Phys Rev E 64:041902
40. Dorogovtsev SN, Mendes JFF, Samukhin AN (2001) Anomalous percolation properties of growing networks. Phys Rev E 64:066110
41. Lancaster D (2002) Cluster growth in two growing network models. J Phys A 35:1179–1194
42. Kim J, Krapivsky PL, Kahng B, Redner S (2002) Infinite-order percolation and giant fluctuations in a protein interaction network. Phys Rev E 66:055101
43. Coulomb S, Bauer M (2003) Asymmetric evolving random networks. Eur Phys J B 35:377–389
44. Krapivsky PL, Derrida B (2004) Universal properties of growing networks. Phys A 340:714–724
45. Szabó G, Alava M, Kertész J (2002) Shortest paths and load scaling in scale-free trees. Phys Rev E 66:026101
46. Bollobás B, Riordan OM (2004) Shortest paths and load scaling in scale-free trees. Phys Rev E 69:036114
47. Cohen R, Havlin S (2003) Scale-free networks are ultra-small. Phys Rev Lett 90:058701
48. Dorogovtsev SN, Mendes JFF (2001) Effect of the accelerating growth of communications networks on their structure. Phys Rev E 63:025101(R)
49. Leskovec J, Kleinberg J, Faloutsos C (2007) Laws of graph evolution: Densification and shrinking diameters. ACM TKDD (1)2 physics/0603229
50. Willinger W, Govindan R, Jamin S, Paxson V, Shenker S (2002) Scaling phenomena in the Internet: Critically examining criticality. Proc Natl Acad Sci 99:2573–2580
51. Ferrer I, Cancho R, Sole RV (2003) Optimization in complex networks. In: Pastor-Satorras R, Rubi M, Diaz-Guilera A (eds) Statistical Mechanics of Complex Networks. Springer, Berlin, pp 114–125; cond-mat/0111222
52. Fabrikant A, Koutsoupias E, Papadimitriou CH (2002) Heuristically optimized trade-offs: A new paradigm for power laws in the Internet. Lecture Notes in Computer Science, vol 2380. Springer, Berlin, pp 110–122
53. Berger N, Bollobás B, Borgs C, Chayes J, Riordan O (2003) Degree distribution of the FKP network model. Lecture Notes in Computer Science, vol 2719. Springer, Berlin, pp 725–738
54. D'Souza RM, Borgs C, Chayes JT, Berger N, Kleinberg RD (2007) Emergence of tempered preferential attachment from optimization. Proc Natl Acad Sci 104:6112–6117
55. Barabási A-L, Ravasz E, Vicsek T (2001) Deterministic scale-free networks. Phys A 299:559–564
56. Dorogovtsev SN, Goltsev AV, Mendes JFF (2002a) Pseudofractal scale-free web. Phys Rev E 65:066122
57. Jung S, Kim S, Kahng B (2002) Geometric fractal growth model for scale-free networks. Phys Rev E 65:056101
58. Andrade JS Jr., Herrmann HJ, Andrade RFS, da Silva LR (2005) Apollonian networks. Phys Rev Lett 94:018702
59. Kim BJ, Trusina A, Minnhagen P, Sneppen K (2005) Self organized scale-free networks from merging and regeneration. Eur Phys J B 43:369–372
60. Goh KI, Kahng B, Kim D (2001) Universal behavior of load distribution in scale-free networks. Phys Rev Lett 87:278701–278704
61. Goh KI, Oh E, Jeong H, Kahng B, Kim D (2002) Classification of scale-free networks. Proc Natl Acad Sci 99:12583–12588
62. Kleinberg JM, Kumar R, and Raghavan P, Rajagopalan S, Tomkins AS (1999) The Web as a graph: Measurements, mod-

els and methods. Lecture Notes in Computer Science, vol 1627. Springer, Berlin, pp 1–17
63. Colizza V, Flammini A, Maritan A, Vespignani A (2005) Characterization and modeling of protein-protein interaction networks. Phys A 352:1–27
64. Barabási A-L, Jeong H, Neda Z, Ravasz E, Schubert A, Vicsek T (2002) Evolution of the social network of scientific collaborations. Phys A 311:590–614
65. Ramasco JJ, Dorogovtsev SN, Pastor-Satorras R (2004) Self-organization of collaboration networks. Phys Rev E 70:036106

Books and Reviews

Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D-U (2006) Complex networks: Structure and dynamics. Phys Rep 424:175–308
Caldarelli G (2007) Scale-Free Networks: Complex Webs in Nature and Technology. Oxford Finance Series. Oxford University Press, Oxford
Cohen R, Havlin S, Ben-Avraham D (2003) Structural properties of scale free networks. In: Bornholdt S, Schuster HG (eds) Handbook of Graphs and Networks. Wiley, Weinheim, pp 85–110
Dorogovtsev SN, Mendes JFF (2002) Evolution of networks. Adv Phys 51:1079–1187
Dorogovtsev SN, Goltsev AV, Mendes JFF (2008) Critical phenomena in complex networks. Rev Mod Phys 80(3) arXiv:0705.00100
Durrett R (2006) Random Graph Dynamics. Cambridge University Press, Cambridge
Kim J, Krapivsky PL, Kahng B, Redner S (2002) Infinite-order percolation and giant fluctuations in a protein interaction network. Phys Rev E 66:055101
Newman MEJ (2003) The structure and function of complex networks. SIAM Review 45:167–256
Pastor-Satorras R, Vespignani A (2004) Evolution and Structure of the Internet: A Statistical Physics Approach. Cambridge University Press, Cambridge
Pastor-Satorras R, Vazquez A, Vespignani A (2001) Dynamical and correlation properties of the Internet. Phys Rev Lett 87:258701–258704
Smythe RT, Mahmoud HM (1995) A survey of recursive trees. Theor Prob Math Statist 51:1–27

Growth Phenomena in Cellular Automata

Janko Gravner
Mathematics Department, University of California, Davis, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Final Set
Asymptotic Shapes
Nucleation
Future Directions
Bibliography

Glossary

Asymptotic density The proportion of sites in a lattice occupied by a specified subset is called asymptotic density, or, in short, density.

Asymptotic shape The shape of a growing set, viewed from a sufficient distance so that the boundary fluctuations, holes, and other lower order details disappear, is called the asymptotic shape.

Cellular automaton A cellular automaton is a sequence of configurations on a lattice which proceeds by iterative applications of a homogeneous local update rule. A configuration attaches a state to every member (also termed a site or a cell) of the lattice. Only configurations with two states, coded 0 and 1, are considered here. Any such configuration is identified with its set of 1's.

Final set A site whose state changes only finitely many times is said to fixate, or attain a final state. If this happens for every site, then the sites whose final states are 1 comprise the final set.

Initial set A starting set for a cellular automaton evolution is called initial set, and may be deterministic or random.

Metastability Metastability refers to a long, but finite, time period in an evolution of a cellular automaton rule, during which the behavior of the iterates has identifiable characteristics.

Monotone cellular automaton A cellular automaton is monotone if addition of 1's to the initial configuration always results in more 1's in any subsequent configuration.

Nucleation Nucleation refers to (usually small) pockets of activity, often termed nuclei, with long range consequences.

Solidification A cellular automaton solidifies if any site which achieves state 1 remains forever in this state.

Definition of the Subject

In essence, analysis of growth models is an attempt to study properties of physical systems far from equilibrium (e. g., [52] and its more than 1300 references). Cellular automata (CA) growth models, by virtue of their simplicity and amenability to computer experimentation [59,66], have become particularly popular in the last

30 years in many fields, such as physics [15,59,60], biology [18], chemistry [15,50], social sciences [12], and artificial life [51]. In contrast to voluminous empirical literature on CA in general and their growth properties in particular, precise mathematical results are rather scarce. A general CA theory is out of the question, since a Turing machine can be embedded in a CA, so that examples as "simple" as elementary one-dimensional CA [17] and Conway's Game of Life [7] are capable of universal computation. Even the most basic parameterized families of CA systems exhibit a bewildering variety of phenomena: self-organization, metastability, turbulence, self-similarity, and so forth [1,22,24,43]. From a mathematical point of view, CA can be rightly viewed as discrete counterparts to partial differential equations, and so they are able to emulate many aspects of the physical world, while at the same time they are easy to experiment using widely available platforms (from many available simulation programs we mention just [66]).

Despite their resistance to traditional methods of deductive analysis, CA have been of interest to mathematicians from their inception and we will focus on the rigorous mathematical results about their growth properties. The scope will be limited to CA with deterministic update rule – random rules are more widely used in applications [12,18,50], but fit more properly within probability theory (however, see e. g. [37] for a connection between deterministic and random aspects of CA growth). Adding randomness to the rule in fact often makes them more tractable, as ergodic properties of random systems are much better understood than those of deterministic ones ([10] provides good examples).

Even though mathematical arguments are the ultimate objective, computer simulations are an indispensable tool for providing the all important initial clues for the subsequent analysis. However, as we explain in a few examples in the sequel, caution needs to be exercised while making predictions based on simulations. Despite the increased memory and speed of commercial computers, for some CA rules highly relevant events occur on spatial and temporal scales far beyond the present-day hardware. Simply put, mathematics and computers are both important, and one ignores each ingredient at one's own peril.

By nature, a subject in the middle of active research contains many exciting unresolved conjectures, vague ideas that need to be made precise, and intriguing examples in search of proper mathematical techniques. It is clear from what has already been accomplished that, often in sharp contrast with the simplicity of the initially posed problem, such techniques may be surprising, sophisticated, and drawn from such diverse areas as combinatorics, geometry, probability, number theory, and PDE. This is a field to which mathematicians of all stripes should feel invited.

Introduction

Let us begin with the general set-up. We will consider exclusively binary CA. Accordingly, a *configuration* will be a member of $\{0,1\}^{\mathbb{Z}^d}$, that is, an assignment of 0 or 1 to every site in the d-dimensional lattice $\mathbb{Z}^d$. This divides the lattice into two sets, those that are assigned state 1, called the *occupied* sites, and those in state 0, which are the *empty* sites. A configuration is thus represented by its occupied set A. This set will change in discrete time, its evolution given by $A_0, A_1, A_2, \cdots \subset \mathbb{Z}^d$.

The configuration changes subject to a CA rule, which is, in general, specified by the following two ingredients. The first is a finite *neighborhood* $\mathcal{N} \subset \mathbb{Z}^d$ of the origin, its translate $x + \mathcal{N}$ then being the neighborhood of point x. By convention, we assume that $\mathcal{N}$ contains the origin. Typically, $\mathcal{N} = B_\nu(0,\rho) \cap \mathbb{Z}^d$, where $B_\nu(0,\rho) = \{x \in \mathbb{R}^d : ||x||_\nu \le \rho\}$ is the ball in the ℓ^ν-norm $||\cdot||_\nu$ and ρ is the *range*. When $\nu = 1$ the resulting $\mathcal{N}$ is called the Diamond neighborhood, while if $\nu = \infty$ it is referred to as the Box neighborhood. (In particular, when $d = 2$, range 1 Diamond and Box neighborhoods are also known as von Neumann and Moore neighborhoods, respectively.) The second ingredient is a map $\pi : 2^{\mathcal{N}} \to \{0,1\}$, which flags the *sufficient configurations* for occupancy. More precisely, for a set $A \subset \mathbb{Z}^d$, we let $\mathcal{T}(A) \subset \mathbb{Z}^d$ consist of every $x \in \mathbb{Z}^d$ for which $\pi((A_t - x) \cap \mathcal{N}) = 1$. Then, for a given *initial subset* $A_0 \subset \mathbb{Z}^d$ of occupied points, we define $A_1, A_2, \ldots$ recursively by $A_{t+1} = \mathcal{T}(A_t)$.

To explain this notation on arguably the most famous CA of all time, the Game of Life [7,26] has $d = 2$, Moore neighborhood $\mathcal{N}$ consisting of the origin and nearest eight sites, so that the neighbor of x is

$$x + \mathcal{N} = \begin{matrix} \bullet & \bullet & \bullet \\ \bullet & x & \bullet \\ \bullet & \bullet & \bullet \end{matrix},$$

and $\pi(S) = 1$ precisely when either $0 \in S$ and $|S| \in \{3,4\}$, or $0 \notin S$ and $|S| = 3$. Here, $|S|$ is the size (cardinality) of $S \subset \mathcal{N}$ and note that the center x of the neighborhood itself is counted in the occupation number.

Usually, our starting set A_0 will consist of a possibly large, but finite set of 1's surrounded by 0's. However, other initial states are worthy of consideration, for example, half-spaces, wedges, and sets with finite complements, called *holes*. Finally, for understanding self-organizational abilities and statistical tendencies of the CA rule, the most

natural starting set is the random "soup" $\Pi(p)$ to which every site is adjoined independently with probability p.

As already mentioned, we need to consider special classes if we hope to formulate a general theorem. Mathematically, the most significant restriction is to the class of *monotone* (or *attractive*) CA rules, for which $S_1 \subset S_2$ implies $\pi(S_1) \leq \pi(S_2)$. To avoid the trivial case we will also assume that monotone CA have $\pi(\mathcal{N}) = 1$.

Another important notion is that of solidification: we say that the CA *solidifies* if $\pi(S) = 1$ whenever $0 \in S$. In words, this means that once a site becomes occupied, it cannot be removed. To *every CA* on $\mathbb{Z}^d$ given by the rule $(\mathcal{N}, \pi)$ one can associate "space-time" solidification CA on $\mathbb{Z}^d \times \mathbb{Z}$, with unique solidification rule given by the neighborhood $\mathcal{N}' = (\mathcal{N} \times \{-1\}) \cup \{0^{d+1}\}$, and π' such that $\pi'(S \times \{-1\}) = \pi(S)$ for $S \subset \mathcal{N}$. This construction is useful particularly for one-dimensional CA whose space-time version interprets their evolution as a two-dimensional object [64], but we prefer to focus on the growth phenomena in the rule's "native" space.

A more restrictive, but still quite rich, class of rules is the *Threshold Growth (TG)* CA, which is a general *totalistic* monotone solidification CA rule. For such rules, $\pi(S)$ depends only on the cardinality $|S|$ of S whenever $0 \notin S$; therefore, for such S there exists a *threshold* $\theta \geq 0$ such that $\pi(S) = 0$ whenever $|S| < \theta$ and $\pi(S) = 1$ whenever $|S| \geq \theta$.

We will universally assume that a 1 cannot spontaneously appear in a sea of 0's, that is, that 1's only grow *by contact*: $\pi(\emptyset) = 0$. We also find it convenient to assume that π is *symmetric*: $-\mathcal{N} = \mathcal{N}$ and $\pi(-S) = \pi(S)$. This is not a necessary assumption in many contexts, but its absence makes many statements unnecessarily awkward.

Next is a very brief historical lesson. The first paper in CA modeling is surely [62], a precursor to the research into nucleation and self-organization in CA. The follow-up to this pioneering work had to wait until the 70's, when the influential work [42] appeared. The earliest work on CA growth is that of S. Willson [63,64,65], which still stands today as one of the notable achievements of mathematical theory. The importance of growth properties of CA, from theoretical and modeling perspectives, was more widely recognized in the mid-80's [53,54,59]. At about the same time, statistical physicists recognized the value of mathematical arguments in studying nucleation and metastability and hence the need to build tractable models [60,61]. Bootstrap percolation ([2,4,67], and references therein), one of the most studied CA, which we discuss in some detail in Sect. "Nucleation", originates from that period. Since the beginning of the 90's there has been a great expansion in the popularity of CA modeling [18,50], while mathematical theory, which we review in the next three sections, proceeds at a much more measured pace.

The rest of the article is organized as follows. In Sect. "Final Set" we consider properties of the set which the CA rule generates "at the end of time." In particular, we discuss when the CA eventually occupies the entire available space and, when it fails to do so, what proportion of space it does fill. Section "Asymptotic Shapes" then focuses on the occupation mechanism, in particular on shapes attained from finite seeds. The main theme of Sect. "Nucleation" are sparse randomly populated initializations. We conclude with Sect. "Future Directions", a summary of issues in need of further research.

Final Set

Perhaps the most basic question that one may ask is: what proportion of space does a CA rule ultimately fill? Clearly we need to specify more precisely what is meant by this, but it should be immediately suspected that the answer in general depends on the initial state, even if we only restrict to finite ones. Indeed, consider the TG CA with Moore neighborhood and $\theta = 3$. It is easy to construct an initial set which stops growing, say, one containing fewer than 3 sites. It is not much harder to convince oneself that there exist finite sets (even some with only 3 sites) which eventually make every site occupied. It is a combinatorial exercise to show that these two are the only possibilities in this example. Is this dichotomy valid in any generality? This is one of the questions we address in this section.

Assume a fixed CA rule and the associated transformation $\mathcal{T}$, and fix an initial state A_0. If every $x \in \mathbb{Z}^d$ *fixates*, that is, changes state only finitely many times, then the *final set* $A_\infty = \mathcal{T}^\infty(A_0)$ exists. Notice that this is automatically true for every solidification rule, in which no site can change state more than once.

We say that A_0 *fills space* if $\mathcal{T}^\infty(A_0) = \mathbb{Z}^d$. One cannot imagine a greater ability of a CA rule to "conquer" the environment than if a finite set is able to fill space. Thus it is natural to ask whether there exist general conditions that assure this property, and indeed they do for monotone CA.

Induced by $\mathcal{T}$ is a growth transformation $\tilde{\mathcal{T}}$ on closed subsets of $\mathbb{R}^d$, given by

$$\tilde{\mathcal{T}}(B) = \{x \in \mathbb{R}^d : 0 \in \mathcal{T}((B - x) \cap \mathbb{Z}^d)\} .$$

In words, one translates the lattice so that $x \in \mathbb{R}^d$ is at the origin, and applies $\mathcal{T}$ to the intersection of Euclidean set B with the translated lattice. It is easy to verify that the two transformations are *conjugate*,

$$\mathcal{T}(B \cap \mathbb{Z}^d) = \tilde{\mathcal{T}}(B) \cap \mathbb{Z}^d .$$

It will become immediately apparent why $\tilde{\mathcal{T}}$ is convenient. Let S^{d-1} be the set of unit vectors $\mathbb{R}^d$ in and let

$$H_u^- = \{x \in \mathbb{R}^d : \langle x, u\rangle \le 0\}$$

be the closed half-space with unit outward normal $u \in S^{d-1}$. Then, *provided that the CA rule is monotone*, there exists a $w(u) \in \mathbb{R}$ so that

$$\tilde{\mathcal{T}}(H_u^-) = H_u^- + w(u) \cdot u$$

and consequently

$$\mathcal{T}^t(H_u^- \cap \mathbb{Z}^d) = (H_u^- + tw(u) \cdot u) \cap \mathbb{Z}^d \,.$$

Monotone CA with $w(u) > 0$ for every u are called *supercritical*. A supercritical CA hence enlarges every half-space.

Theorem 1 *Assume a monotone CA rule. A finite set A_0 which fills space exists if and only if $w(u) > 0$ for every direction $u \in S^{d-1}$.*

See [30,63] for a proof. Before we proceed, a few remarks are in order. First, we should note that one direction of the above theorem has a one-line proof: if $w(u) \le 0$ for some u, then monotonicity prevents the CA from ever occupying a point outside a suitable translate of H_u^-. The other direction is proved by constructing a sufficiently "smooth" initial set. Moreover, supercriticality can be checked on a finite number of directions, in particular one can prove that a two-dimensional TG CA is supercritical if and only if $\theta \le \frac{1}{2}(|\mathcal{N}| - \max\{|\mathcal{N} \cap \ell| : \ell$ a line through $0\})$ [30]. Thus, among the TG CA with Moore neighborhood, exactly those with $\theta \le 3$ are supercritical, while this is true for range 2 Box neighborhood when $\theta \le 10$.

A finite set A_0 for which $\cup_t A_t$ is infinite is said to *generate persistent growth*. Further, a CA for which any set that generates persistent growth has $A_\infty = \mathbb{Z}^d$ is called *omnivorous* [30]. For an omnivorous rule a finite seed has either a bounded effect or it fills space.

Is every supercritical TG CA omnivorous? The answer is no, and a counterexample in $d = 2$ is obtained by taking the neighborhood to be the cross of radius 2: $\mathcal{N} = \{(0,0), (0,\pm1), (0,\pm2), (\pm1,0), (\pm2,0)\}$, and $\theta = 2$. It is easy to check that for $A_0 = \{(0,0), (1,0)\}$ the final set A_∞ consists of the x-axis, while initialization with a 2×2 box results in $A_\infty = \mathbb{Z}^2$. On the other hand, the following theorem holds.

Theorem 2 *The two-dimensional TG CA is omnivorous provided either of the two conditions are satisfied:*

(1) $\mathcal{N}$ is box neighborhood of arbitrary range.

(2) $\mathcal{N} = \tilde{\mathcal{N}} \cap \mathbb{Z}^2$, where $\tilde{\mathcal{N}}$ is a convex set with the same symmetries as $\mathbb{Z}^2$, and $\theta \le \sigma^2/2$, where σ is the range of the largest box neighborhood contained in $\mathcal{N}$.

The theorem is proved in [8] and [9] by rather delicate combinatorial arguments involving analysis of invariant, or nearly invariant, quantities. The lack of robust methods makes conditions in the theorem far from necessary. In particular, proving a general analogue of Theorem 2 without solidification (while keeping monotonicity) is an intriguing open problem.

For non-monotone solidification rules, any general theory appears impossible, but one can analyze specific examples, and we list some recent results below. All are two-dimensional, therefore we assume $d = 2$ for the rest of this section.

In many interesting cases, it is immediately clear from computer simulations that $A_\infty \ne \mathbb{Z}^d$, but at least A_∞ is spread out fairly evenly. This motivates the following definition. Pick a set $A \subset \mathbb{Z}^2$. Let μ_ϵ be ϵ^2 times the counting measure on $\epsilon \cdot A$. We say that A has *asymptotic density* ρ if μ_ϵ converges to $\rho \cdot \lambda$ as $\epsilon \to 0$. Here λ is Lebesgue measure on $\mathbb{R}^2$ and the convergence holds in the usual sense:

$$\int f \, \mathrm{d}\mu_\epsilon \to \rho \cdot \int f \, \mathrm{d}\lambda \tag{1}$$

for any $f \in C_c(\mathbb{R}^2)$. Equivalently, for any square $R \subset \mathbb{R}^2$, the quantity $\epsilon^2 \cdot |R \cap (\epsilon \cdot A)|$ converges to the area of R as $\epsilon \to 0$.

For totalistic solidification CA, the rule is determined by the neighborhood and a *solidification list* of neighborhood counts which result in occupation at the next time. Three neighborhoods have been studied so far: *Diamond* rules with von Neumann neighborhood, *Box* rules with Moore neighborhood and *Hex* rules with the neighborhood $\mathcal{N}$ consisting of $(0,0)$ and the six sites $(\pm1,0)$, $(0,\pm1)$, and $\pm(1,1)$. (We note that this last neighborhood is a convenient way to represent the triangular lattice [59].) These rules are often referred to as *Packard snowflakes* [11,36,53]. As an example, in *Hex 135* rule, a 0 turns into a 1 exactly when it "sees" an odd number of already occupied neighbors in its hexagonal neighborhood.

We will assume that 1 is on the solidification list, for otherwise the analysis quickly becomes too difficult (see however [33] and [45] for some results on *Box 2* and *Box 3*) rules. Further, for *Hex* and *Diamond* cases, we will assume 2 is not on this list (or else the dynamics is too similar to a TG CA). We now summarize the main results of [36] and [11].

Theorem 3 *To each of the four Diamond and 16 Hex Packard snowflakes there corresponds a number $\rho \in (0,1]$,*

the asymptotic density of A_∞, which is independent of the finite seed A_0. The densities in Diamond cases are

$$\rho_1 = 2/3,\ \rho_{13} = 2/3,\ \rho_{14} = 1,\ \rho_{134} = 29/36\,.$$

The Hex densities are exactly computable in 8 cases:

$$\rho_{13} = \rho_{135} = 5/6,\ \rho_{134} = \rho_{1345} = 21/22,\ \rho_{136} = \rho_{1356} = \rho_{1346} = \rho_{13456} = 1\,.$$

In six other Hex cases, one can estimate, within ± 0.0008,

$$\rho_1 \approx 0.6353,\ \rho_{14}, \rho_{145} \approx 0.9689,\ \rho_{15} \approx 0.8026,$$
$$\rho_{16} \approx 0.7396,\ \rho_{156} \approx 0.9378\,.$$

The final two Hex rules have densities very close to 1:

$$\rho_{146} \in (0.995, 1),\ \ \rho_{1456} \in (0.9999994, 1)\,.$$

The indices in the densities of course refer to the respective rule.

Note that, in each of the two cases, $\rho_{14} > \rho_{134}$, testimony to the fundamentally nonmonotone nature of these rules. It is also shown in [36] that observing *Hex 1456* from $A_0 = \{0\}$ on even the world's most extensive graphics array, with millions of sites on a side, one would still be led to the conclusion that $A_\infty = \mathbb{Z}^2$. In fact, the site in A^c_∞ closest to the origin is still at distance of the order 10^9. Nevertheless, A^c_∞ has a positive density and contains arbitrarily large islands of 0's. This is one illustration of limitations in making empirical conclusions based on simulations.

The fundamental tool to prove the above theorem is the fact that these dynamics have an additive component which creates an impenetrable *web* of occupied sites [36]. This web consists of sites at the edge of the light cone, or, to be more precise, the sites which are occupied at the same time at which TG CA with the same neighborhood and $\theta = 1$ would occupy them.

The web makes at least an approximate recursion possible, and the basic renewal theory applies. The delicacy of such results is conveyed effectively by comparison to *Box* solidification. There are 128 such rules with 1 on the solidification list. Although snowflake-like recursive carpets emerge in a great many cases, and exact computations are sometimes feasible, there is no hope of a complete analysis as in the *Hex* and *Diamond* settings, and many fascinating problems remain. For instance, the density of *Box 1*, provided it exists at all, can depend on the initial seed. Namely, it is shown in [33] that *Box 1* solidification yields density 4/9 starting from a singleton. Later, D. Hickerson (private communication) engineered finite initial seeds with asymptotic densities 29/64 and 61/128.

Growth Phenomena in Cellular Automata, Table 1
Densities of A_∞ from $A_0 = \{0\}$ for some *Box* rules

Rule	Density
12	2/3
13	28/45
15	43/72
16	385/576
17	35/72
18	4/9

The latter is achieved by an ingenious arrangement of 180 carefully placed occupied cells around the boundary of an 83×83 grid. The highest density with which Box 1 solidification can fill the plane is not known, and neither is whether any seed fills with density less than 4/9. Most initial seeds generate what seems to be a chaotic growth with density about 0.51.

Many other *Box* rules have known asymptotic densities started from a singleton. Table 1 is a sample (D. Griffeath, private communication).

All exact density computations presented in this section are based on explicit recursions, made possible by an additive web. These recursions are in some cases far from simple, for example, D. Griffeath has shown that in the *Box 12* case, the following formula holds for $a_n = |A_\infty \cap B_\infty(0, 2^n - 1)|$, $n \geq 12$:

$$a_n = \frac{8}{3}\cdot 4^n + r_1\gamma_3^n - \frac{8}{3}\cdot 3^n - \frac{16}{15}\cdot 2^n + \frac{2}{51}\cdot(-2)^n + 4n - 3 + \frac{8}{3}\cdot(-1)^n + r_2\gamma_1^n + r_3\cdot\gamma_2^n\,,$$

where

$$\gamma_1 \approx -.675\,,\ \gamma_2 \approx .461\,,\ \gamma_3 \approx 3.214$$

are the three real roots of the equation $\gamma^3 - 3\gamma^2 - \gamma + 1 = 0$, while

$$r_1 \approx -6.614\,,\ r_2 \approx -2.126\,,\ r_3 \approx 2.434$$

solve $3145r^3 + 19832r^2 - 22688r - 107648 = 0$.

Apparently very similar rules to those in the above table seem unsolvable, such as *Box 14*, and the "odd" rule *Box 1357* which does have an additive component, but the resulting web from $A_0 = \{0\}$ "leaks" and growth is apparently chaotic. The same problem plagues almost all *Box* rules started from general initial set. The sole exception seems to be the *12* rule, the best candidate for a general general theorem among the 128 rules, due to its quasiadditive web [49]. We should also mention that embedded additive dynamics have been used to study other models [23].

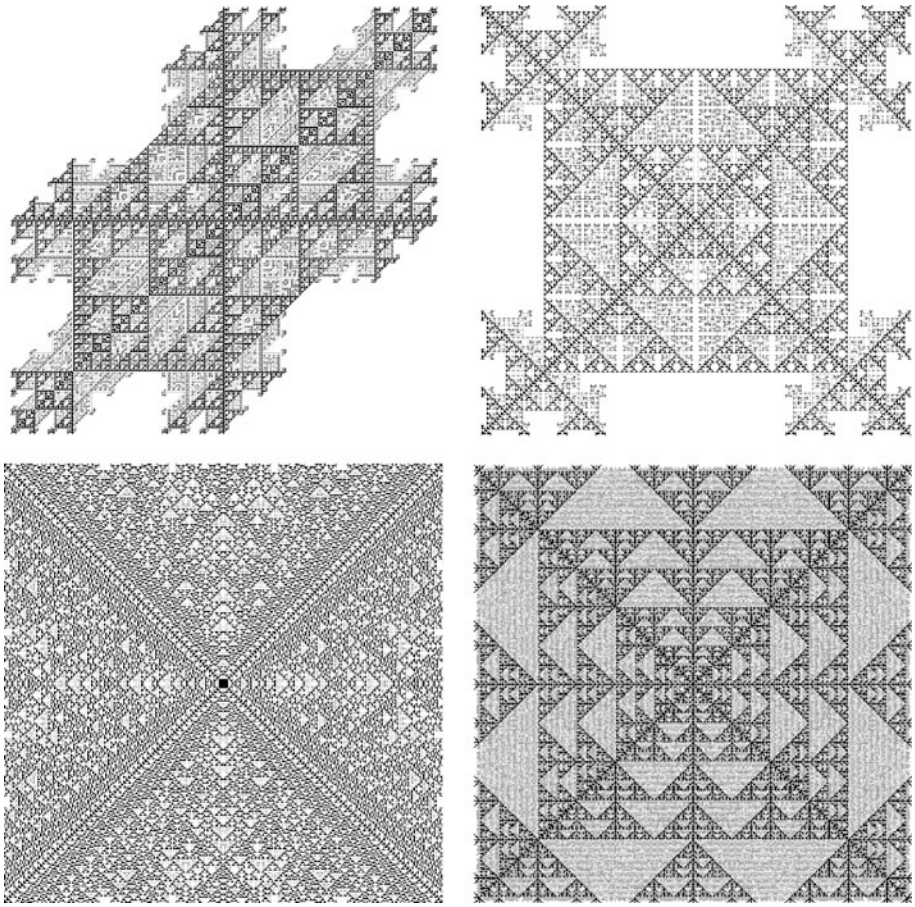

Growth Phenomena in Cellular Automata, Figure 1
Some Packard snowflakes. Clockwise from top *left*: *Hex 1*; *Box 1*; *Box 1357*; and again *Box 1*. First three are started from {0}, and the last from an 8 × 8 box. The web is black, otherwise the updates are periodically shaded. Note that the chaotic growth can result from a chaotic web (*bottom left*) or from a leaky web (*bottom right*)

In all considered cases, the web consist of several copies of the final set generated by the space-time solidification associated to a one-dimensional CA. When this CA is linear, the web's fractal dimension can be computed using the method from [65]. For example, the properly scaled webs in the top two frames on Fig. 2 approach a set with Hausdorff dimension $\log 3/\log 2$, while for the bottom right web this dimension is $\log(1+\sqrt{5})/\log 2$.

Given that all exactly given densities so far are rational, a natural question is whether there is an example of A_∞ with irrational density. Such example was given by Griffeath and Hickerson in [44], where an initial state for the *Game of Life* is provided for which the set $t^{-1}A_t$ converges to an asymptotic density $(3-\sqrt{5})/90$ on an appropriate finite set L. This formulation masks the fact that every site x eventually periodically changes its state, so A_∞ does not exist. However, a closer look at the construction shows that the final periods are uniformly bounded. Therefore, if p is the lowest common multiple of all final periods, the p'th iterate of the *Game of Life* rule will generate A_∞ from the same A_0 and with the same density.

This is the only known example of a computable irrational density, and there is a good reason, which we now explain, why such examples are difficult to come by.

By analogy with statistical physics, we would call a set $A \subset \mathbb{Z}^2$ *exactly solvable*, if there exists a formula which decides whether a given x is an element of A. More formally, we require that there exists a finite automaton which, upon encountering x as input, decides whether $x \in A$. Representation of x as input is given as $(\pm i_1^1, \pm i_1^2, i_2^1, i_2^2, \ldots)$, where i_1^1, i_1^2 are the most significant binary digits of the first and second coordinate of x; i_2^1, i_2^2 the next most significant, etc. (Some initial i_k^1's or i_k^2's may be 0, and the representation is finite but of arbitrary length.) This means that A is *auto-*

matic [5], or equivalently a *uniform tag system* [16]. With a slight abuse of terminology we call a solidification CA *exactly solvable* (from A_0) if A_∞ is exactly solvable.

To our knowledge, the simplest nontrivial example of an exactly solvable CA is *Diamond 1* solidification, for which it can be shown by induction that $x \notin A_\infty$ iff $\max\{k\colon i_k^1 = 1\} = \max\{k\colon i_k^2 = 1\}$. It is easy to construct a (two-state) finite automaton that checks this condition, and the density ρ of A_∞ evidently must satisfy the equation $\rho = 1/2 + \rho/4$, so that $\rho = 2/3$ as stated in Theorem 3. In fact, all of the CA in Theorem 3 with exactly given densities are exactly solvable, and then, by [16], Theorem 6, these densities must be rational. Therefore, the Griffeath–Hickerson example given above is *not* exactly solvable, and the mechanism that forms A_∞ must be more complex in this precise sense. We note that none of the other examples from Theorem 3 are exactly solvable either, but for a different reason [36].

This section's final example, like many other fascinating CA rules, is due to D. Hickerson (private communication). His *Diamoeba* is a rule with the Moore neighborhood and $\pi(S) = 1$ whenever one of the following two conditions is satisfied:

$$0 \notin S, \text{ and } |S| \in \{3, 5, 6, 7, 8\}, \text{ or}$$
$$0 \in S, \text{ and } |S| \in \{6, 7, 8, 9\}\,.$$

This would be an easily analyzed monotone rule if the 3 were replaced by a 9, with $A_\infty = \emptyset$ for every finite A_0. At first, it seems that the Diamoeba shares this fate. In fact, D. Hickerson has demonstrated that, starting from $A_0 = B_\infty(0, r) \cap \mathbb{Z}^2$, $A_t = \emptyset$ at the smallest t given by

$$12r - 8 - 4r_1 + r_{11} + (r \mod 2)\,,$$

where r_1 and r_{11} are, respectively, the number of 1's and the number of 11's in the binary representation of r. This interesting formula only gives a small taste of things to come (see [33] for a detailed discussion). One of the most intriguing examples is when A_0 is a 2×59 rectangle with a corner cell removed. This grows to a fairly large set in about a million updates, then apparently stops for several million more, after which another growth spurt is possible. The question whether $A_\infty = \mathbb{Z}^2$ for this A_0 is tantalizingly left open. However, there does exist an A_0 for which $A_\infty = \mathbb{Z}^2$. This initialization was discovered by D. Bell, and is an adaptation of a spaceship found by a search algorithm designed by D. Eppstein [21]. This startling object attests to the remarkable design expertise that *Game of Life* researchers have developed over the years.

Asymptotic Shapes

After addressing a rule's *ability* to grow in the previous section, we now turn to the *geometry* of growth: is it possible to predict the shape that the set of 1's attains as it spreads? It turns out that the complete answer is known in the monotone case.

Naturally, we need a notion of convergence of sets, and the most natural definition is due to Hausdorff (see [29,31] for an introduction to such issues). We say that a sequence of compact sets $K_n \subset \mathbb{R}^d$ converges to a compact set $K \subset \mathbb{R}^d$ (in short, $K_n \to K$) if, for every $\epsilon > 0$, $K_n \subset K + B_2(0, \epsilon)$ and $K \subset K_n + B_2(0, \epsilon)$, for n large enough. Then we say that a CA has a *linear asymptotic shape* L from a finite initial seed A_0 if

$$\frac{1}{t} A_t \to L$$

as $t \to \infty$.

Turning to monotone CA, we recall the definition of half-space velocities w, and set

$$K_{1/w} = \cup\{[0, 1/w(u)] \cdot u\colon u \in S^{d-1}\}$$

and let L be the polar transform of $K_{1/w}$, that is,

$$L = K_{1/w}^* = \{x \in \mathbb{R}^d\colon \langle x, u\rangle \le w(u), \quad \text{for every } u \in S^{d-1}\}\,.$$

Growth Phenomena in Cellular Automata, Figure 2
The Bell–Eppstein initial set (*left*) that results in $A_\infty = \mathbb{Z}^2$ for the *Diamoeba* rule. The set A_t, whose linear asymptotic shape is a rhombus with vertices $(\pm 1/7, 0)$ and $(0, \pm 1/8)$, is shown at $t = 500$

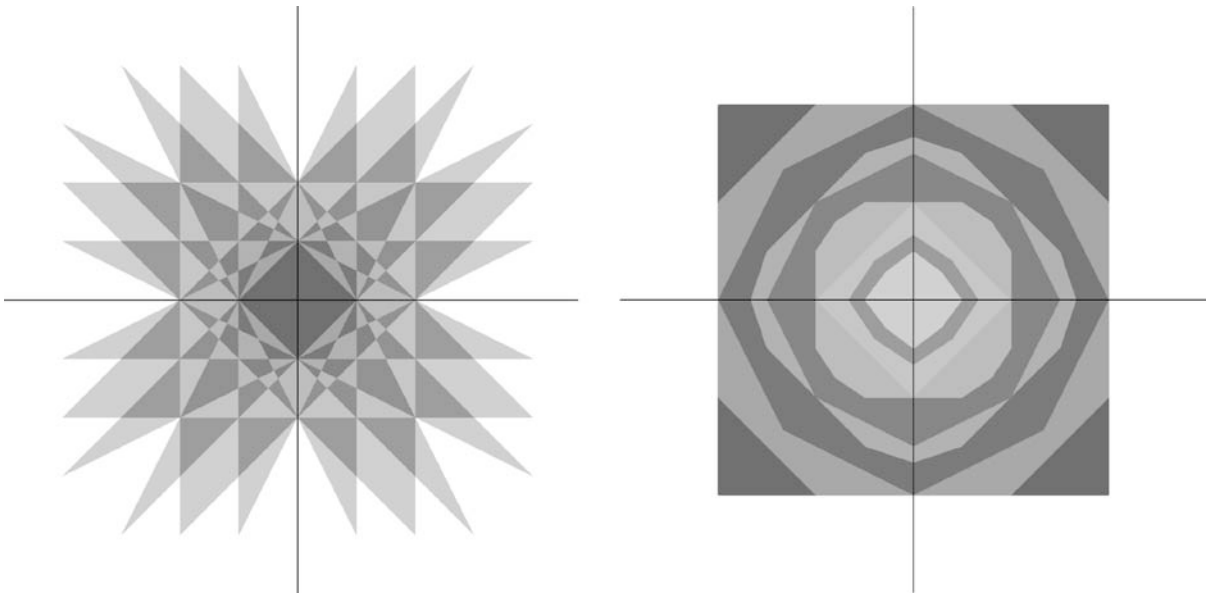

Growth Phenomena in Cellular Automata, Figure 3
The sets $K_{1/w}$ (*left*) and the asymptotic shapes for all 10 supercritical range 2 TG CA. Note that there are only 9 shapes, as those with $\theta = 7$ and $\theta = 8$ coincide

In general, the polar of a set $K \subset \mathbb{R}^d$ is given by $K^* = \{y \in \mathbb{R}^d : \langle x, y\rangle \leq 1 \text{ for every } x \in K\}$. The set L is known as a *Wulff shape*, and is a very important notion in crystallography and statistical physics [55]. The next theorem was proved in the classic paper [64]. The core methods in its proof, as well as proofs of similar results [29], are those of convex and discrete geometry.

Theorem 4 *Assume a monotone CA rule with all $w(u) \geq 0$. Then there exists a large enough r so that for every finite initial set A_0, which contains $B_2(0, r) \cap \mathbb{Z}^d$, the linear asymptotic shape from A_0 equals the Wulff shape L. Even more, the difference between A_t and tL is bounded: there exists a constant C, which depends on the rule and on A_0, so that $A_t \subset tL + B_2(0, C)$ and $tL \subset A_t + B_2(0, C)$ for every $t \geq 0$.*

Note that supercriticality is not assumed here. If $w(u) = 0$ for some u, then $K_{1/w(u)}$ is an infinite object and L has dimension less than d. (The one trivial case is when $w \equiv 0$ and $L = \{0^d\}$.) Finally, note that if there exists a u so that $w(u) < 0$ (and hence $w(-u) < 0$, by symmetry), then A_t is sandwiched between two hyperplanes which approach each other and so eventually $A_t = \emptyset$.

It is also important to point out that $K_{1/w}$ is always a polytope, L is always a convex polytope and both are, for small neighborhoods, readily computable by hand or by computer [30,31,32]. For example, the Moore neighborhood TG CA with $\theta = 3$ has $K_{1/w}$ with 16 vertices, of which three successive ones are $(0, 1)$, $(1, 2)$, $(1, 1)$, and the remaining 13 are then continued by symmetry. It then follows that the limiting shape L is the convex hull of $(\pm 1/2, 0), (0, \pm 1/2), (\pm 1/3, \pm 1/3)$.

Matters become much murkier when the monotonicity assumption is dropped. We discuss a few interesting two-dimensional solidification examples next. They all hinge on *recursive specification* of iterates A_t for every t (see [33] for a definition). This is far from a general approach (and appears to fail even for simple monotone cases), but is the primary technique available.

We begin with the *Box 25* solidification, starting from $A_0 = B_2(0, r + 1/2) \cap \mathbb{Z}^2$. As was observed in [33], and can be quickly checked by computer, the linear asymptotic shape exists for $r = 2$, $r = 9$ and $r = 13$, but is in each case different, in fact it is convex in the first case, and nonconvex in the other two cases. This demonstrates that such shapes may depend on the initial seed.

A very interesting example was discovered by D. Hickerson (private communication). Consider *Box 37* solidification, with $A_0 = B_2(0, 7/2) \cap \mathbb{Z}^2$. Then $t^{-1/2}A_t$ converges to $B_\infty(0, 2\sqrt{2/3})$ as $t \to \infty$. This demonstrates the possibility of nontrivial *sublinear* asymptotic shapes.

We turn next to the *Hex* rules [36]. These exhibit subsequential limiting shapes, which are not always polygons, as we explain next.

Theorem 5 *Take any of the 16 Hex rules as in Theorem 3, and fix a finite A_0. There exists a one-parameter family of sets S_a, $a \in [0,1]$, so that the following holds: for $t_n = a \cdot 2^n$,*

$$2^{-n} A_{t_n} \to S_a \ ,$$

as $n \to \infty$.

Furthermore, when 3 and 4 are not both on the solidification list, the family S_a is called simple and is independent of the initial set. In the opposite, diverse case, initial sets are divided into two classes, distinguished by two different families S_a.

For rational a, it can be shown that the Hausdorff dimension of ∂S_a always exists, and is in principle computable. For example, for the simple S_a this dimension equals 5/4 for $a = 14/15$, evidently producing a non-polygonal subsequential shape.

This discussion brings forth the following question, which is probably the most interesting open problem on CA growth. For a *prescribed* set L, can we find a CA with linear asymptotic shape L, attained from a "generic" collection of initial sets? In particular, can L be a circle, thereby giving rise to asymptotic isotropy?

We note that the isotropic construction is possible for probabilistic CA [39], so it seems likely that the answer is yes for a properly constructed chaotic growth. However, techniques for such an approach are completely lacking at present. We should also remark that computational universality should allow for a construction of a CA and a *carefully engineered* initial state with circular (or any other) shape – although this has never been explicitly done. This would, however, violate the requirement of generic initialization.

We conclude this section by a short review of *reverse* shapes [34]. The question here is, if the initial set A_0 is a large hole, and evolves until shortly before the entire lattice is occupied, what is the resulting shape? The initial state has a large and persistent effect on the dynamics and thus the reverse shape geometry will depend on it. The detailed analysis depends on technical convexity arguments, but the cleanest instance is given by the following result.

Theorem 6 *Assume a monotone CA, with $w \geq 0$ but not identically 0 on S^{d-1}. Assume also that its rule $\mathcal{T}$ preserves all symmetries of the lattice $\mathbb{Z}^d$. Pick a closed convex set $H \subset \mathbb{R}^d$, which has all symmetries of $\mathbb{Z}^d$, and let $A_0 = (mH)^c \cap \mathbb{Z}^d$ for some large m. Moreover, let*

$$T = \inf\{t\colon 0 \in A_t\} \ .$$

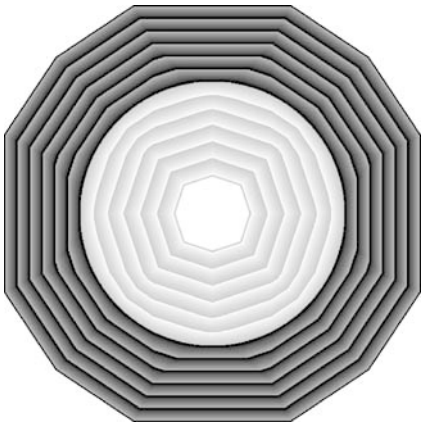

Growth Phenomena in Cellular Automata, Figure 4
Superimposed convergence to the linear asymptotic shape and to the reverse shape, from, respectively, the interior and the exterior, of a large lattice circle. The rule is TG CA with range 2 and $\theta = 6$. Iterates are periodically shaded

There is a nonempty bounded convex subset $\mathcal{R}(H) \subset \mathbb{R}^d$ such that

$$\lim_{M\to\infty} \lim_{m\to\infty} \frac{1}{M} \cdot A_{T-M} = \mathcal{R}(H)^c \ ,$$

in the Hausdorff sense. Moreover, if

$$h_0 = \max\{h > 0\colon h \cdot H^* \subset K_{1/w}\} \ ,$$

then

$$\mathcal{R}(H) = (h_0 \cdot H^* \cap \partial K_{1/w})^* \ .$$

In words, one scales the polar H^* so that it touches the boundary of $K^{1/w}$; at this point, the intersection determines the reverse shape. (The shape does not change if H is multiplied by a constant, so h_0 determines its natural scale.) The paper [34] has many more details and examples.

Nucleation

In this section we assume that the initial state A_0 is the product measure $\Pi(p)$, with density $p > 0$ that is typically very small. Initially, then, there will be no significant activity on most of the space. Certainly this is no surprise as most of the space is empty, but isolated 1's or small islands of them are often not able to accomplish much either. Most of the lattice is thus in a *metastable* state. However, at certain rare locations there may, by chance, occur local configurations, which are able to spread their influence over large distances until they statistically dominate

the lattice. These locations are called *nuclei*, and their frequency and mechanism of growth are the main self-organizational aspects of the CA rule. The majority of results are confined to two-dimensions, so we will assume $d = 2$ for the rest of this section and relegate higher dimensions to remarks.

We start with a simple example, for which we give a few details to introduce the basic ideas and demonstrate that a CA can go through more than one metastable state. For this example we do not specify the map π, but instead give a more informal description. In a configuration A, we call an *insurance* five sites in a cross formation in the state 1, or, more formally, a translate of von Neumann neighborhood which is inside A. The map $\mathcal{T}$ changes any 0 with a 1 in its von Neumann neighborhood to 1. Moreover, it automatically changes any 1 to 0, except that any 1 whose von Neumann neighborhood intersects with an insurance remains 1. Then, for every $\epsilon > 0$, as $p \to 0$,

$$P(0 \in A_t^c \text{ for all } t \le p^{-1/2+\epsilon}) \to 1\,,$$
$$P(0 \in (A_t \text{ xor } A_{t+1}) \text{ for all } p^{-1/2-\epsilon} \le t \le p^{-5/2+\epsilon}) \to 1\,,$$
$$P(0 \in A_t \text{ for all } p^{-5/2-\epsilon} \le t) \to 1\,.$$

(Here, xor is the exclusive union.) Roughly, most sites are 0 up to time $p^{-1/2}$, then periodic with period 2 up to time $p^{-5/2}$, and 1 afterwards. (In fact, stronger statements, along the lines of Theorem 7 below, are possible.)

The proof has two phases: the first deterministic and the second probabilistic. For the deterministic one, let $d_1(x)$ be the ℓ^1 distance from x to A_0, and assume that A_0 contains no insurance. Then one can prove by induction that, first, none of the A_t contain an insurance, and second, that for every x and $t \ge d_1(x)$, $x \in A_t$ precisely when $(t - d_1(x)) \mod 2 = 0$. On the other hand, an insurance in A_0 centered at the origin will result in $x \in A_t$ for every $t \ge d_1(x) - 1$. The probabilistic part consists of noting that, with overwhelming probability when p is small, $B_1(0, p^{-1/2+\epsilon})$ (resp. $B_1(0, p^{-5/2+\epsilon})$) contains no 1 (resp. insurance) in $A_0 = \Pi(p)$, while $B_1(0, p^{-1/2-\epsilon})$ (resp. $B_1(0, p^{-5/2-\epsilon})$) does.

The bulk of the mathematical theory of nucleation and metastability addresses monotone CA, although some work has been done on the *Game of Life* [28] and its generalizations [1,22], excitable media dynamics [19,24,25,42], and artificial life models [51].

Our first general class are supercritical monotone solidification CA. (In fact, the solidification assumption is not necessary, but reduces technical details so much that it is assumed in most published works.) Such rules have two *nucleation parameters*. Let γ be the smallest i for which there exists an A_0 with $|A_0| = i$ that generates persistent growth. Moreover, let ν be the number of sets A_0 of size γ that generate persistent growth and have the leftmost among their lowest sites at the origin. (The last requirement ensures that ν counts the number of distinct smallest "shapes" that grow.) We call the rule *voracious* if, started from any of the ν initial sets A_0 described above, $A_\infty = \mathbb{Z}^2$. Voracity is a weak condition, which assures a minimal regularity of growth and can, for any fixed rule, be checked on finitely many cases (which is not true for the more restrictive omnivorous property).

Growth Phenomena in Cellular Automata, Table 2
Nucleation parameter ν for small Box neighborhood TG CA

	$\theta = 2$	$\theta = 3$	$\theta = 4$	$\theta = 5$	$\theta = 6$	$\theta = 7$
$\rho = 1$	12	42				
$\rho = 2$	40	578	4683	24938	94050	259308

For illustration, we briefly discuss these for range ρ Box neighborhood TG CA. For relatively small θ, $\gamma = \theta$; for example, when $\rho = 1$, $\gamma = \theta$ for all three supercritical rules, while when $\rho = 2$, γ exceeds θ only for $\theta = 10$, when it equals 11. For large ρ, and $\theta \sim \alpha\rho^2$, γ is asymptotically the smallest possible (that is, $\gamma \sim \alpha\rho^2$) when $\alpha < \alpha_c$ for some $\alpha_c \in (1.61, 1.66)$ [32]. One can also compute some ν, before they become too large.

Returning to $A_0 = \Pi(p)$, the most natural statistics to study is

$$T = \inf\{t\colon 0 \in A_t\}\,,$$

the first time the CA occupies the origin.

Theorem 7 *Assume a monotone, supercritical, and voracious CA, with nucleation parameters γ and ν. Then, as $p \to 0$,*

$$\sqrt{\nu p^\gamma} \cdot T$$

converges in distribution to a nontrivial random variable τ, which is a functional of a Poisson point location $\mathcal{P}$ with unit intensity.

That $T \approx p^{-\gamma/2}$ can be easily guessed (and proved), but the more precise asymptotics described above require a considerable argument [30], as interaction between growing droplets is nontrivial. In particular, the higher dimensional version has not been proved, and the description of the limiting "movie" from $\mathcal{P}$ probably cannot avoid viscosity methods from PDE [58].

The most exciting nucleation results have been proved about *critical* models, for which $w(u)$ vanishes for some direction u but is positive for others. Although a general framework is presented in [35], we will instead focus on

the most studied examples. Of these the most popular has been the *bootstrap percolation (BP)*, which is TG CA with von Neumann neighborhood and $\theta = 2$ [2,3,4,67]. Its *modified* version *(MBP)* has the same neighborhood, still solidifies, but when $0 \notin S$, $\pi(S) = 1$ precisely when $\{\pm e_1\} \cap S \neq \emptyset$ and $\{\pm e_2\} \cap S \neq \emptyset$. (Here e_1 and e_2 are the basis vectors.)

Now $w(\pm e_1) = w(\pm e_2) = 0$, so no finite set can generate persistent growth, and it is not immediately clear that $P(T < \infty) = 1$. This is true [67], as very large sets are able to use sparse but helpful smattering of 1's around them and so are unlikely to be stopped. To determine the size of T, one needs more information about the necessary size of these nuclei, and the likelihood of their formation. This was started in [4] and culminated by the following theorem by A. Holroyd [46], which is arguably the crowning achievement of CA nucleation theory to date.

Theorem 8 *For BP let $\lambda = \pi^2/18$, and for MBP let $\lambda = \pi^2/6$. Then, for every $\epsilon > 0$,*

$$P(p \log T \in [\lambda - \epsilon, \lambda + \epsilon]) \to 1$$

as $p \to 0$.

To summarize, $T \approx \exp(\lambda/p)$, which is for small p a long time indeed and amply justifies the description of the almost empty lattice as metastable.

The most common formulation of the theorem above involves finite $L \times L$ squares with periodic boundary instead of infinite lattices. Then

$$I(L, p) = P(\text{the entire square is eventually occupied})$$

and, as $p \to 0$,

$$I(L, p) \to 1 \quad \text{if} \quad p \log L \geq \lambda + \epsilon\,,$$
$$I(L, p) \to 0 \quad \text{if} \quad p \log L \leq \lambda - \epsilon\,.$$

Here L is of course assumed to increase with p. Before the value of λ was known, this second formulation was used to estimate it by simulation. For example, [3] used L close to 30,000 and obtained $\lambda \approx 0.245$ for BP, about a factor of two smaller than the true value $0.548\ldots$. Other simulations of BP, MBP, and related models all exhibit a similar discrepancy. The reason apparently is that nuclei are, for realistic values of p, quite a bit more frequent than the asymptotics would suggest. Indeed, the following result from [38] confirms this.

Theorem 9 *For BP and MBP,*

$$I(L, p) \to 1 \quad \textit{if} \quad p \log L \geq \lambda - c(\log L)^{-1/2}\,,$$

for an appropriate constant c.

This alone indicates that to halve the error in approximating λ on an $L \times L$ system it is necessary to replace L by L^4. In addition, [38] shows that for the more tractable MBP one can do explicit calculations to conclude that to get an estimate of λ within 2%, one would need L at least 10^{500}, a non-achievable size.

For BP, the quantity $p \log L$ is the "order parameter," the quantity that, when varied, causes a phase transition (which, in addition, is sharp by Theorem 8). We will list now some other models with known order parameters – we also indicate the status of phase transition, when known:

- CA with von Neumann neighborhood and $\pi(S) = 1$ when $|S \setminus \{0\}| \geq 2$: $p^2 \log L$ [57];
- TG CA with range ρ Box neighborhood, $\theta \in [2\rho^2 + \rho + 1, 2\rho^2 + 2\rho]$: $p^{\theta - 2\rho^2 - \rho} \log L$ [30];
- TG CA with $\mathcal{N} = \{(0,0), (0, \pm 1), (\pm 1, 0), (\pm 2, 0)\}$, $\theta = 2$: $p^{3/2} L$, not sharp [30];
- TG CA with $\mathcal{N} = \{(0,0), (0, \pm 1), (\pm 1, 0), (\pm 2, 0)\}$, $\theta = 3$: $(-\log p)^{-2} p \log L$ [30,68];
- TG CA with range ρ cross neighborhood $\mathcal{N} = \{(x, y) \colon |x| \leq \rho, |y| \leq \rho, xy = 0\}$ and $\theta = \rho + 1$: $p \log L$, sharp at $\lambda = \pi^2/(3(\rho + 1)(\rho + 2))$ [48];
- TG CA on $\mathbb{Z}^d$ with $\mathcal{N} = B_1(0, 1) \cap \mathbb{Z}^d$ and $\theta \in [3, d]$, $p^{1/(d-\theta+1)} \log_{\theta-1} L$ (where $\log_k$ is the k'th iterate of log) [14], sharp at $\lambda = \pi^2/6$ for the modified version when $\theta = d$ [47].

Note that when $\theta = d = 3$ the last example gives the metastable scale $\exp(\exp(\lambda/p))$ [13,56], making an even modest experimental approximation of λ impossible.

There are other interesting issues about critical growth models, which do not have to do with nucleation. One is decay rate for the first passage time T [6], which is connected to the properties of the very last holes to be filled in. Another is its ability to overtake random obstacles [40].

Apart from sending $p \to 0$, one could vary other parameters to get the metastability phenomena, and one natural example is the range. We explain this scenario on a non-monotone CA known as the *Threshold Voter Automaton (TVA)* [20,31]. For simplicity, assume $\mathcal{N}$ is range ρ Box neighborhood, and fix a threshold θ. This rule makes a site change its "opinion" by contact with at least θ of the opposite opinions:

$$\pi(S) = 1 \quad \text{iff} \quad (0 \in S \text{ and } |S^c| < \theta) \text{ or } (0 \notin S \text{ and } |S| \geq \theta)\,.$$

As the two opinions are symmetric, the most natural initial state of TVA is $\Pi(1/2)$. We also assume that ρ is large and the scaling $\theta = a|\mathcal{N}|$, for some $a \in (0, 1)$.

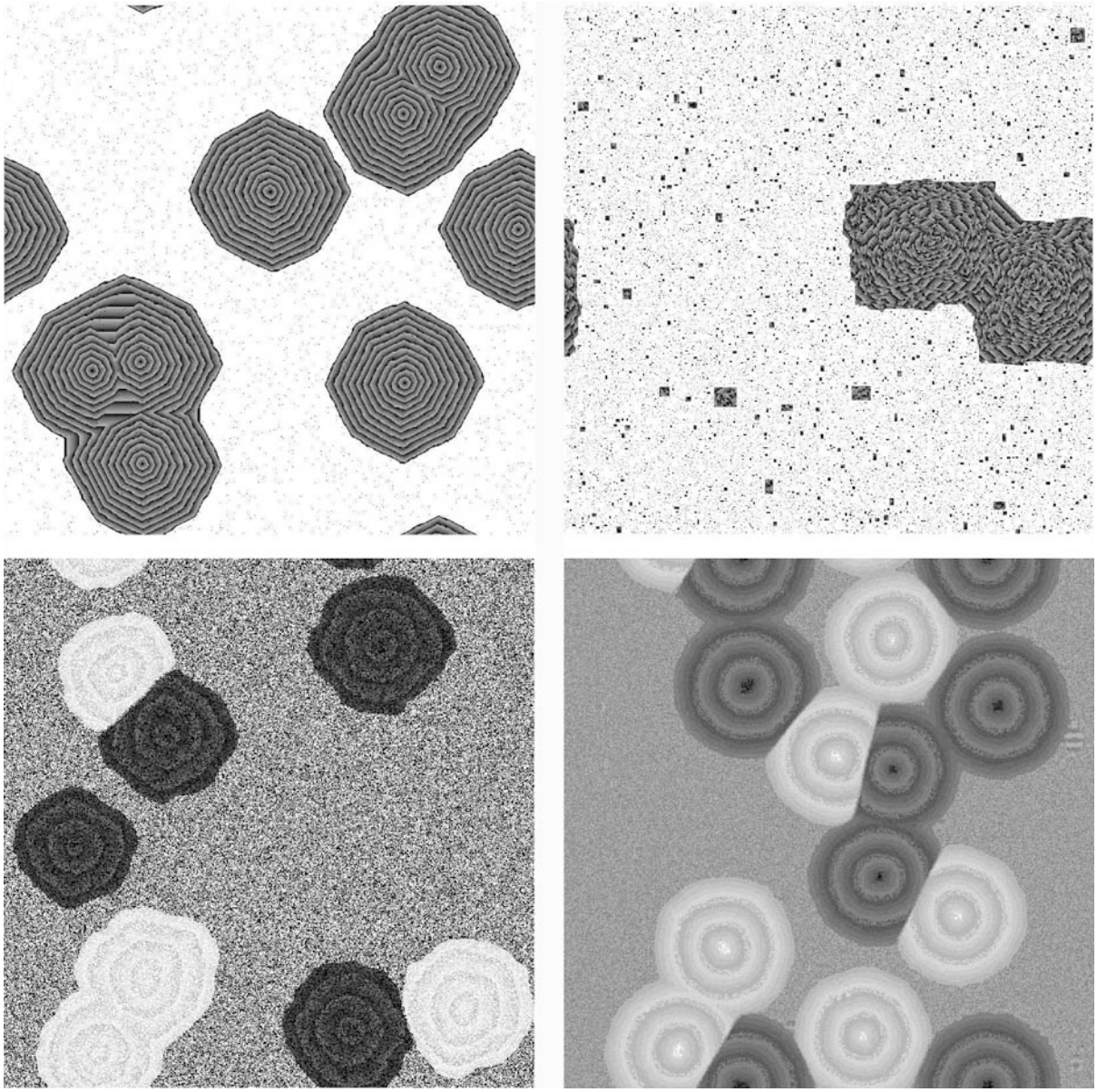

Growth Phenomena in Cellular Automata, Figure 5
Four nucleation examples, each on an 800 × 800 array with periodic boundary. Clockwise from *top left*: TGM CA with Moore neighborhood, $\theta = 3$, and $p = 0.006$; bootstrap percolation with $p = 0.041$; TVA with $\rho = 10$ and $\theta = 194$; TVA with $\rho = 10$ and $\theta = 260$. The iterates are periodically colored to indicate growth and, in the TVA frames, the lighter shades indicate 0's

It is proved in [20] that when $a > 3/4$, any fixed $x \in \mathbb{Z}^2$ changes its opinion only finitely many times with probability approaching 1 as $\rho \to \infty$ — and the rigorous results end there. The most interesting rare nucleation questions arise when $a \in (1/4, 3/4) \setminus \{1/2\}$. According to simulations, under this assumption the nuclei are rare and eventually tessellate the lattice into regions of consensus with either stable or periodic boundaries [31]. However, the definition of a nucleus is unclear and consequently their density cannot be estimated. Two torus simulations are given in Fig. 5; it is important to point out that, for such *finite* systems, Lyapunov methods of [27] imply that every site eventually fixates or becomes periodic with period 2.

The *majority* TVA, when $a = 1/2$, is perhaps the most appealing of all [43]. The nucleation is now not rare; instead, this CA quickly self-organizes into visually attrac-

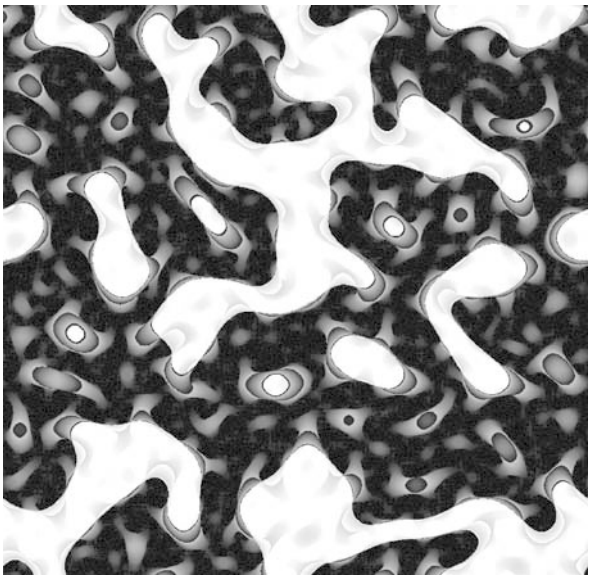

Growth Phenomena in Cellular Automata, Figure 6
Majority vote: TGM with $\rho = 10$, $\theta = 221$ on a 1000×1000 array with periodic boundary. Again, iterates are periodically colored with the lighter shades reserved for 0's

tive curvature driven dynamics. (Note that flat interfaces between the two opinions are now stable, so one opinion can advance only when it forms a concavity.) For any fixed ρ this must eventually stop, as finite islands with uniformly small enough curvature of either opinion are stable. However, when ρ is increased this effect is with large probability not felt on any fixed finite portion of space. (A similar effect is achieved by the Vichniac "twist" [61].) Many fascinating questions remain about this case, especially on the initial nucleating phase, whose analysis depends on delicate properties of random fields and remains an open problem.

Future Directions

We will identify seven themes, which connect to open problems discussed in previous sections. Progress on each is bound to be a challenge, but also a significant advance in understanding CA growth.

Regularity of Growth

It is often important, and of independent interest, to be able to conclude that a cellular automaton rule generates growth without arbitrarily large tentacles, holes, or other undesirable features. An omnivorous CA, for example, has this property. The natural goal would be to develop techniques to establish such regularity for much more general monotone and non-monotone CA, and for arbitrary dimension. Many rules give the impression that regular growth is a generic trait, i. e., holds for a majority of initial sets.

Oscillatory Growth

Does there exist a class of CA with growing sets that oscillate on different scales? Hickerson's *Diamoeba* might be able to accomplish this from some initial sets, but perhaps there are other, more tractable, examples with identifiable mechanisms.

Analysis of Chaotic Growth

One look at the growth of *Box 1* solidification from a 8×8 initial box (bottom left frame in Fig. 1) would convince most observers that it has a square asymptotic shape. However, there are no tools to prove, or disprove, this statement. A fully rigorous theory of chaotic CA, tailored to address such asymptotic issues, is almost nonexistent and constitutes perhaps the most important challenge for mathematicians in this area.

Nucleation Theory for Non-monotone Models

Once nucleation centers are established, growth most often proceeds in a random environment, which consists of debris left over from the nucleation phase. This may help the analysis, as it adds a random perturbation to what may otherwise be intractable dynamics, but on the other hand random environment processes are notoriously tricky to analyze [41].

Robust Exact Constants and Sharp Transitions

The nucleation phase transition has been proved sharp for a few critical models, by rather delicate arguments. A more robust approach would extend them, and would perhaps provide further insights into the error terms, for which only one sided estimates are now known. The apparent *crossover* [3] phenomenon would also be interesting to understand rigorously.

Three-Dimensional Nucleation and Growth

With advances in computer power, extensive three-dimensional CA simulations have become viable on commercial hardware. Therefore, it may be possible to investigate nucleation, droplet interaction, clustering mechanisms, and other staples of two-dimensional CA research,

at least experimentally. Proper visualization tools of complex three-dimensional phenomena may well require some novel ideas in computer graphics.

Generic Properties of CA with Large Range

A TG CA with range ρ, say, has on the order of ρ^2 possible thresholds θ. When can it be established that some property holds for the majority of relevant choices? One such property (sensitivity of shapes to random perturbations in the rule) was analyzed from this perspective in [37], but it would be interesting to provide further examples for TG or other classes of CA.

Bibliography

Primary Literature

1. Adamatzky A, Martínez GJ, Mora JCST (2006) Phenomenology of reaction-diffusion binary-state cellular automata. Int J Bifurc Chaos Appl Sci Eng 16:2985–3005
2. Adler J (1991) Bootstrap percolation. Physica A 171:453–4170
3. Adler J, Staufer D, Aharony A (1989) Comparison of bootstrap percolation models. J Phys A: Math Gen 22:L279–L301
4. Aizenman M, Lebowitz J (1988) Metastability effects in bootstrap percolation. J Phys A: Math Gen 21:3801–3813
5. Allouche J-P, Shallit J (2003) Automatic sequences: theory, applications, generalizations. Cambridge University Press, Cambridge
6. Andjel E, Mountford TS, Schonmann RH (1995) Equivalence of decay rates for bootstrap percolation like cellular automata. Ann Inst H Poincaré 31:13–25
7. Berlekamp ER, Conway JH, Guy RK (2004) Winning ways for your mathematical plays, vol 4, 2nd edn. Peters, Natick
8. Bohman T (1999) Discrete threshold growth dynamics are omnivorous for box neighborhoods. Trans Am Math Soc 351:947–983
9. Bohman T, Gravner J (1999) Random threshold growth dynamics. Random Struct Algorithms 15:93–111
10. Bramson M, Neuhauser C (1994) Survival of one-dimensional cellular automata under random perturbations. Ann Probab 22:244–263
11. Brummitt CD, Delventhal H, Retzlaff M (2008) Packard snowflakes on the von Neumann neighborhood. J Cell Autom 3:57–80
12. Bäck T, Dörnemann H, Hammel U, Frankhauser P (1996) Modeling urban growth by cellular automata. In: Lecture Notes in Computer Science. Proceedings of the 4th International Conference on Parallel Problem Solving from Nature, vol. 1141. Springer, Berlin, pp 636–645
13. Cerf R, Cirillo ENM (1999) Finite size scaling in three-dimensional bootstrap percolation. Ann Probab 27:1837–1850
14. Cerf R, Manzo F (2002) The threshold regime of finite volume bootstrap percolation. Stoch Process Appl 101:69–82
15. Chopard B, Droz M (1998) Cellular automata modeling of physical systems. Cambridge University Press, Cambridge
16. Cobham A (1972) Uniform tag sequences. Math Syst Theory 6:164–192
17. Cook M (2005) Universality in elementary cellular automata. Complex Syst 15:1–40
18. Deutsch A, Dormann S (2005) Cellular automata modeling of biological pattern formation. Birkhäuser, Boston
19. Durrett R, Steif JE (1991) Some rigorous results for the Greenberg–Hastings model. J Theor Probab 4:669–690
20. Durrett R, Steif JE (1993) Fixation results for threshold voter systems. Ann Probab 21:232–247
21. Eppstein D (2002) Searching for spaceships. In: More Games of no Chance (Berkeley, CA, 2000). Cambridge University Press, Cambridge, pp 351–360
22. Evans KM (2001) Larger than Life: digital creatures in a family of two-dimensional cellular automata. In: Cori R, Mazoyer J, Morvan M, Mosseri R (eds) Discrete Mathematics and Theoretical Computer Science, vol AA. pp 177–192
23. Evans KM (2003) Replicators and larger than life examples. In: Griffeath D, Moore C (eds) New constructions in cellular automata. Oxford University Press, New York, pp 119–159
24. Fisch R, Gravner J, Griffeath D (1991) Threshold-range scaling for the excitable cellular automata. Stat Comput 1:23–39
25. Fisch R, Gravner J, Griffeath D (1993) Metastability in the Greenberg–Hastings model. Ann Appl Probab 3:935–967
26. Gardner M (1976) Mathematical games. Sci Am 133:124–128
27. Goles E, Martinez S (1990) Neural and Automata Networks. Kluwer, Dordrecht
28. Gotts NM (2003) Self-organized construction in sparse random arrays of Conway's game of life. In: Griffeath D and Moore C (eds) New Constructions in Cellular Automata. Oxford University Press, New York, pp 1–53
29. Gravner J, Griffeath D (1993) Threshold growth dynamics. Trans Am Math Soc 340:837–870
30. Gravner J, Griffeath D (1996) First passage times for the threshold growth dynamics on $\mathbb{Z}^2$. Ann Probab 24:1752–1778
31. Gravner J, Griffeath D (1997) Multitype threshold voter model and convergence to Poisson-Voronoi tessellation. Ann Appl Probab 7:615–647
32. Gravner J, Griffeath D (1997) Nucleation parameters in discrete threshold growth dynamics. Exp Math 6:207–220
33. Gravner J, Griffeath D (1998) Cellular automaton growth on $\mathbb{Z}^2$: theorems, examples and problems. Adv Appl Math 21:241–304
34. Gravner J, Griffeath D (1999) Reverse shapes in first–passage percolation and related growth models. In: Bramson M, Durrett R (eds) Perplexing Problems in Probability. Festschrift in Honor of Harry Kesten, Birkhäuser, Bostons, pp 121–142
35. Gravner J, Griffeath D (1999) Scaling laws for a class of critical cellular automaton growth rules. In: Révész P, Tóth B (eds) Random Walks. János Bolyai Mathematical Society, Budapest, pp 167–186
36. Gravner J, Griffeath D (2006) Modeling snow crystal growth. I. Rigorous results for Packard's digit snowflakes. Exp Math 15:421–444
37. Gravner J, Griffeath D (2006) Random growth models with polygonal shapes. Ann Probab 34:181–218
38. Gravner J, Holroyd AE (2008) Slow convergence in bootstrap percolation. Ann Appl Probab 18:909–928
39. Gravner J, Mastronarde N Shapes in deterministic and random growth models (in preparation)
40. Gravner J, McDonald E (1997) Bootstrap percolation in a polluted environment. J Stat Phys 87:915–927
41. Gravner J, Tracy C, Widom H (2002) A growth model in a random environment. Ann Probab 30:1340–1368

42. Greenberg J, Hastings S (1978) Spatial patterns for discrete models of diffusion in excitable media. SIAM J Appl Math 4:515–523
43. Griffeath D (1994) Self-organization of random cellular automata: four snapshots. In: Grimmett G (ed) Probability and Phase Transition. Kluwer, Dordrecht, pp 49–67
44. Griffeath D, Hickerson D (2003) A two-dimensional cellular automaton with irrational density. In: Griffeath D, Moore C (eds) New constructions in cellular automata. Oxford University Press, Oxford, pp 119–159
45. Griffeath D, Moore C (1996) Life without death is **P**-complete. Complex Syst 10:437–447
46. Holroyd AE (2003) Sharp metastability threshold for two-dimensional bootstrap percolation. Probab Theory Related Fields 125:195–224
47. Holroyd AE (2006) The metastability threshold for modified bootstrap percolation in *d* dimensions. Electron J Probab 11:418–433
48. Holroyd AE, Liggett TM, Romik D (2004) Integrals, partitions, and cellular automata. Trans Am Math Soc 356:3349–3368
49. Jen E (1991) Exact solvability and quasiperiodicity of one-dimensional cellular automata. Nonlinearity 4:251–276
50. Kier LB, Seybold PG, Cheng C-K (2005) Cellular Automata Modeling of Chemical Systems. Springer, Dordrecht
51. Lindgren K, Nordahl MG (1994) Evolutionary dynamics of spatial games. Physica D 75:292–309
52. Meakin P (1998) Fractals, scaling and growth far from equilibrium. Cambridge University Press, Cambridge
53. Packard NH (1984) Lattice models for solidification and aggregation. Institute for Advanced Study preprint. Reprinted in: Wolfram S (ed) (1986) Theory and application of cellular automata. World Scientific, Singapore, pp 305–310
54. Packard NH, Wolfram S (1985) Two-dimensional cellular automata. J Stat Phys 38:901–946
55. Pimpinelli A, Villain J (1999) Physics of Crystal Growth. Cambridge University Press, Cambridge
56. Schonmann RH (1992) On the behavior of some cellular automata related to bootstrap percolation. Ann Probab 20:174–193
57. Schonmann RH (1990) Finite size scaling behavior of a biased majority rule cellular automaton. Physica A 167:619–627
58. Song M (2005) Geometric evolutions driven by threshold dynamics. Interfaces Free Bound 7:303–318
59. Toffoli T, Margolus N (1997) Cellular Automata Machines. MIT Press, Cambridge
60. Vichniac GY (1984) Simulating physics with cellular automata. Physica D 10:96–116
61. Vichniac GY (1986) Cellular automata models of disorder and organization. In: Bienenstock E, Fogelman-Soulie F, Weisbuch G (eds) Disordered Systems and Biological Organization. Springer, Berlin, pp 1–20
62. Wiener N, Rosenblueth A (1946) The math foundation of the problem of conduction of impulses in a network of connected excitable elements, specifically in cardiac muscle. Arch Inst Cardiol Mex 16:205–265
63. Willson SJ (1978) On convergence of configurations. Discret Math 23:279–300
64. Willson SJ (1984) Cellular automata can generate fractals. Discret Appl Math 8:91–99
65. Willson SJ (1987) Computing fractal dimensions for additive cellular automata. Physica D 24:190–206
66. Wójtowicz M (2001) Mirek's Cellebration: a 1D and 2D Cellular Automata Explorer, Version 4.20. http://www.mirwoj.opus.chelm.pl/ca/
67. van Enter ACD (1987) Proof of Straley's argument for bootstrap percolation. J Stat Phys 48:943–945
68. van Enter ACD, Hulshof T (2007) Finite-size effects for anisotropic bootstrap percolation: logarithmic corrections. J Stat Phys 128:1383–1389

Books and Reviews

Adamatzky A (1995) Identification of cellular automata. Taylor & Francis, London

Allouche J-P, Courbage M, Kung J, Skordev G (2001) Cellular automata. In: Encyclopedia of physical science and technology, vol 2, 3rd edn. Academic Press, San Diego, pp 555–567

Allouche J-P, Courbage M, Skordev G (2001) Notes on cellular automata. Cubo, Matemática Educational 3:213–244

Durrett R (1988) Lecture Notes on particle Systems and Percolation. Wadsworth & Brooks/Cole, Pacific Grove

Durrett R (1999) Stochastic spatial models. SIAM Rev 41:677–718

Gravner J (2003) Growth phenomena in cellular automata. In: Griffeath D, Moore C (eds) New Constructions in Cellular Automata. Oxford University Press, New York, pp 161–181

Holroyd AE (2007) Astonishing cellular automata. Bull Centre Rech Math 10:10–13

Ilachinsky A (2001) Cellular Automata: A Discrete Universe. World Scientific, Singapore

Liggett TM (1985) Interacting Particle Systems. Springer, New York

Liggett TM (1999) Stochastic Interacting Systems: contact, voter and exclusion processes. Springer, New York

Rothman DH, Zaleski S (1997) Lattice-Gas Cellular Automata. Cambridge University Press, Cambridge

Toom A (1995) Cellular automata with errors: problems for students of probability. In: Snell JL (ed) Topics in contemporary probability and its applications. CRC Press, Boca Raton, pp 117–157

H

Hamiltonian Perturbation Theory (and Transition to Chaos)

Henk W. Broer[1], Heinz Hanssmann[2]
[1] Instituut voor Wiskunde en Informatica, Rijksuniversiteit Groningen, Groningen, The Netherlands
[2] Mathematisch Instituut, Universiteit Utrecht, Utrecht, The Netherlands

Article Outline

Glossary
Definition of the Subject
Introduction
One Degree of Freedom
Perturbations of Periodic Orbits
Invariant Curves of Planar Diffeomorphisms
KAM Theory: An Overview
Splitting of Separatrices
Transition to Chaos and Turbulence
Future Directions
Bibliography

Glossary

Bifurcation In parametrized dynamical systems a bifurcation occurs when a qualitative change is invoked by a change of parameters. In models such a qualitative change corresponds to transition between dynamical regimes. In the generic theory a finite list of cases is obtained, containing elements like 'saddle-node', 'period doubling', 'Hopf bifurcation' and many others.

Cantor set, Cantor dust, Cantor family, Cantor stratification Cantor dust is a separable locally compact space that is perfect, i. e. every point is in the closure of its complement, and totally disconnected. This determines Cantor dust up to homeomorphisms. The term Cantor set (originally reserved for the specific form of Cantor dust obtained by repeatedly deleting the middle third from a closed interval) designates topological spaces that locally have the structure $\mathbb{R}^n \times$ *Cantor dust* for some $n \in \mathbb{N}$. Cantor families are parametrized by such Cantor sets. On the real line $\mathbb{R}$ one can define Cantor dust of positive measure by excluding around each rational number p/q an interval of size

$$\frac{2\gamma}{q^\tau}\,, \quad \gamma > 0\,, \quad \tau > 2\,.$$

Similar Diophantine conditions define Cantor sets in $\mathbb{R}^n$. Since these Cantor sets have positive measure their Hausdorff dimension is n. Where the unperturbed system is stratified according to the co-dimension of occurring (bifurcating) tori, this leads to a Cantor stratification.

Chaos An evolution of a dynamical system is chaotic if its future is badly predictable from its past. Examples of non-chaotic evolutions are periodic or multi-periodic. A system is called chaotic when many of its evolutions are. One criterion for chaoticity is the fact that one of the Lyapunov exponents is positive.

Diophantine condition, Diophantine frequency vector A frequency vector $\omega \in \mathbb{R}^n$ is called Diophantine if there are constants $\gamma > 0$ and $\tau > n-1$ with

$$|\langle k, \omega\rangle| \geq \frac{\gamma}{|k|^\tau} \quad \text{for all} \quad k \in \mathbb{Z}^n \backslash \{0\}\,.$$

The Diophantine frequency vectors satisfying this condition for fixed γ and τ form a Cantor set of half lines. As the Diophantine parameter γ tends to zero (while τ remains fixed), these half lines extend to the origin. The complement in any compact set of frequency vectors satisfying a Diophantine condition with fixed τ has a measure of order $O(\gamma)$ as $\gamma \downarrow 0$.

Integrable system A Hamiltonian system with n degrees of freedom is (Liouville)-integrable if it has n functionally independent commuting integrals of motion. Locally this implies the existence of a torus action, a feature that can be generalized to dissipative sys-

tems. In particular a mapping is integrable if it can be interpolated to become the stroboscopic mapping of a flow.

KAM theory Kolmogorov–Arnold–Moser theory is the perturbation theory of (Diophantine) quasi-periodic tori for nearly integrable Hamiltonian systems. In the format of quasi-periodic stability, the unperturbed and perturbed system, restricted to a Diophantine Cantor set, are smoothly conjugated in the sense of Whitney. This theory extends to the world of reversible, volume-preserving or general dissipative systems. In the latter KAM theory gives rise to families of quasi-periodic attractors. KAM theory also applies to torus bundles, in which case a global Whitney smooth conjugation can be proven to exist, that keeps track of the geometry. In an appropriate sense invariants like monodromy and Chern classes thus also can be defined in the nearly integrable case. Also compare with ► Kolmogorov–Arnold–Moser (KAM) Theory.

Nearly integrable system In the setting of perturbation theory, a nearly integrable system is a perturbation of an integrable one. The latter then is an integrable approximation of the former. See an above item.

Normal form truncation Consider a dynamical system in the neighborhood of an equilibrium point, a fixed or periodic point, or a quasi-periodic torus, reducible to Floquet form. Then Taylor expansions (and their analogues) can be changed gradually into normal forms, that usually reflect the dynamics better. Often these display a (formal) torus symmetry, such that the normal form truncation becomes an integrable approximation, thus yielding a perturbation theory setting. See above items. Also compare with ► Normal Forms in Perturbation Theory.

Persistent property In the setting of perturbation theory, a property is persistent whenever it is inherited from the unperturbed to the perturbed system. Often the perturbation is taken in an appropriate topology on the space of systems, like the Whitney C^k-topology [72].

Perturbation problem In perturbation theory the unperturbed systems usually are transparent regarding their dynamics. Examples are integrable systems or normal form truncations. In a perturbation problem things are arranged in such a way that the original system is well-approximated by such an unperturbed one. This arrangement usually involves both changes of variables and scalings.

Resonance If the frequencies of an invariant torus with multi- or conditionally periodic flow are rationally dependent, this torus divides into invariant sub-tori. Such resonances $\langle h, \omega \rangle = 0$, $h \in \mathbb{Z}^k$, define hyperplanes in ω-space and, by means of the frequency mapping, also in phase space. The smallest number $|h| = |h_1| + \cdots + |h_k|$ is the order of the resonance. Diophantine conditions describe a measure-theoretically large complement of a neighborhood of the (dense!) set of all resonances.

Separatrices Consider a hyperbolic equilibrium, fixed or periodic point or invariant torus. If the stable and unstable manifolds of such hyperbolic elements are codimension one immersed manifolds, then they are called separatrices, since they separate domains of phase space, for instance, basins of attraction.

Singularity theory A function $H\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ has a critical point $z \in \mathbb{R}^n$ where $\mathrm{D}H(z)$ vanishes. In local coordinates we may arrange $z = 0$ (and similarly that it is mapped to zero as well). Two germs $K\colon (\mathbb{R}^n, 0) \longrightarrow (\mathbb{R}, 0)$ and $N\colon (\mathbb{R}^n, 0) \longrightarrow (\mathbb{R}, 0)$ represent the same function H locally around z if and only if there is a diffeomorphism η on $\mathbb{R}^n$ satisfying

$$N = K \circ \eta\,.$$

The corresponding equivalence class is called a singularity.

Structurally stable A system is structurally stable if it is topologically equivalent to all nearby systems, where 'nearby' is measured in an appropriate topology on the space of systems, like the Whitney C^k-topology [72]. A family is structurally stable if for every nearby family there is a re-parametrization such that all corresponding systems are topologically equivalent.

Definition of the Subject

The fundamental problem of mechanics is to study Hamiltonian systems that are small perturbations of integrable systems. Also, perturbations that destroy the Hamiltonian character are important, be it to study the effect of a small amount of friction, or to further the theory of dissipative systems themselves which surprisingly often revolves around certain well-chosen Hamiltonian systems. Furthermore there are approaches like KAM theory that historically were first applied to Hamiltonian systems. Typically perturbation theory explains only part of the dynamics, and in the resulting gaps the orderly unperturbed motion is replaced by random or chaotic motion.

Introduction

We outline perturbation theory from a general point of view, illustrated by a few examples.

The Perturbation Problem

The aim of perturbation theory is to approximate a given dynamical system by a more familiar one, regarding the former as a perturbation of the latter. The problem then is to deduce certain dynamical properties from the unperturbed to the perturbed case.

What is familiar may or may not be a matter of taste, at least it depends a lot on the dynamical properties of one's interest. Still the most frequently used unperturbed systems are:

- Linear systems
- Integrable Hamiltonian systems, compare with ▶ Dynamics of Hamiltonian Systems and references therein
- Normal form truncations, compare with ▶ Normal Forms in Perturbation Theory and references therein
- Etc.

To some extent the second category can be seen as a special case of the third. To avoid technicalities in this section we assume all systems to be sufficiently smooth, say of class C^∞ or real analytic. Moreover in our considerations ε will be a real parameter. The unperturbed case always corresponds to $\varepsilon = 0$ and the perturbed one to $\varepsilon \neq 0$ or $\varepsilon > 0$.

Examples of Perturbation Problems To begin with consider the autonomous differential equation

$$\ddot{x} + \varepsilon\dot{x} + \frac{\mathrm{d}V}{\mathrm{d}x}(x) = 0\,,$$

modeling an oscillator with small damping. Rewriting this equation of motion as a planar vector field

$$\begin{aligned}\dot{x} &= \quad y\\ \dot{y} &= -\varepsilon y - \frac{\mathrm{d}V}{\mathrm{d}x}(x)\,,\end{aligned}$$

we consider the energy $H(x, y) = \frac{1}{2}y^2 + V(x)$. For $\varepsilon = 0$ the system is Hamiltonian with Hamiltonian function H. Indeed, generally we have $\dot{H}(x, y) = -\varepsilon y^2$, implying that for $\varepsilon > 0$ there is dissipation of energy. Evidently for $\varepsilon \neq 0$ the system is no longer Hamiltonian.

The reader is invited to compare the phase portraits of the cases $\varepsilon = 0$ and $\varepsilon > 0$ for $V(x) = -\cos x$ (the pendulum) or $V(x) = \frac{1}{2}\lambda x^2 + \frac{1}{24}bx^4$ (Duffing).

Another type of example is provided by the non-autonomous equation

$$\ddot{x} + \frac{\mathrm{d}V}{\mathrm{d}x}(x) = \varepsilon f(x, \dot{x}, t)\,,$$

which can be regarded as the equation of motion of an oscillator with small external forcing. Again rewriting as a vector field, we obtain

$$\begin{aligned}\dot{t} &= \quad 1\\ \dot{x} &= \quad y\\ \dot{y} &= -\frac{\mathrm{d}V}{\mathrm{d}x}(x) + \varepsilon f(x, y, t)\,,\end{aligned}$$

now on the generalized phase space $\mathbb{R}^3 = \{t, x, y\}$. In the case where the t-dependence is periodic, we can take $\mathbb{S}^1 \times \mathbb{R}^2$ for (generalized) phase space.

Remark

- A small variation of the above driven system concerns a parametrically forced oscillator like

  $$\ddot{x} + (\omega^2 + \varepsilon\cos t)\sin x = 0\,,$$

  which happens to be entirely in the world of Hamiltonian systems.
- It may be useful to study the Poincaré or period mapping of such time periodic systems, which happens to be a mapping of the plane. We recall that in the Hamiltonian cases this mapping preserves area. For general reference in this direction see, e. g., [6,7,27,66].

There are lots of variations and generalizations. One example is the solar system, where the unperturbed case consists of a number of uncoupled two-body problems concerning the Sun and each of the planets, and where the interaction between the planets is considered as small [6,9,107,108].

Remark

- One variation is a restriction to fewer bodies, for example only three. Examples of this are systems like Sun–Jupiter–Saturn, Earth–Moon–Sun or Earth–Moon–Satellite.
- Often Sun, Moon and planets are considered as point masses, in which case the dynamics usually are modeled as a Hamiltonian system. It is also possible to extend this approach taking tidal effects into account, which have a non-conservative nature.
- The Solar System is close to resonance, which makes application of KAM theory problematic. There exist, however, other integrable approximations that take resonance into account [3,63].

Quite another perturbation setting is local, e. g., near an equilibrium point. To fix thoughts consider

$$\dot{x} = Ax + f(x)\,, \quad x \in \mathbb{R}^n$$

with $A \in \mathrm{gl}(n, \mathbb{R})$, $f(0) = 0$ and $\mathrm{D}_x f(0) = 0$. By the scaling $x = \varepsilon \bar{x}$ we rewrite the system to

$$\dot{\bar{x}} = A\bar{x} + \varepsilon g(\bar{x}) .$$

So, here we take the linear part as an unperturbed system. Observe that for small ε the perturbation is small on a compact neighborhood of $\bar{x} = 0$.

This setting also has many variations. In fact, any normal form approximation may be treated in this way ► Normal Forms in Perturbation Theory. Then the normalized truncation forms the unperturbed part and the higher order terms the perturbation.

Remark In the above we took the classical viewpoint which involves a perturbation parameter controlling the size of the perturbation. Often one can generalize this by considering a suitable topology (like the Whitney topologies) on the corresponding class of systems [72]. Also compare with ► Normal Forms in Perturbation Theory, ► Kolmogorov–Arnold–Moser (KAM) Theory and ► Dynamics of Hamiltonian Systems.

Questions of Persistence

What are the kind of questions perturbation theory asks? A large class of questions concerns the *persistence* of certain dynamical properties as known for the unperturbed case. To fix thoughts we give a few examples.

To begin with consider equilibria and periodic orbits. So we put

$$\dot{x} = f(x, \varepsilon) , \quad x \in \mathbb{R}^n , \quad \varepsilon \in \mathbb{R} , \tag{1}$$

for a map $f\colon \mathbb{R}^{n+1} \to \mathbb{R}^n$. Recall that equilibria are given by the equation $f(x, \varepsilon) = 0$. The following theorem that continues equilibria in the unperturbed system for $\varepsilon \neq 0$, is a direct consequence of the implicit function theorem.

Theorem 1 (Persistence of equilibria) *Suppose that $f(x_0, 0) = 0$ and that*

$$\mathrm{D}_x f(x_0, 0) \ \textit{has maximal rank} .$$

Then there exists a local arc $\varepsilon \mapsto x(\varepsilon)$ with $x(0) = x_0$ such that

$$f(x(\varepsilon), \varepsilon) \equiv 0 .$$

Periodic orbits can be approximated in a similar way. Indeed, let the system (1) for $\epsilon = 0$ have a periodic orbit γ_0. Let Σ be a local transversal section of γ_0 and $P_0\colon \Sigma \to \Sigma$ the corresponding Poincaré map. Then P_0 has a fixed point $x_0 \in \Sigma \cap \gamma_0$. By transversality, for $|\varepsilon|$ small, a local Poincaré map $P_\varepsilon\colon \Sigma \to \Sigma$ is well-defined for (1). Observe that fixed points x_ε of P_ε correspond to periodic orbits γ_ε of (1). We now have, again as another direct consequence of the implicit function theorem.

Theorem 2 (Persistence of periodic orbits) *In the above assume that*

$$P_0(x_0) = x_0 \ \textit{and} \ \mathrm{D}_x P_0(x_0) \ \textit{has no eigenvalue 1} .$$

Then there exists a local arc $\varepsilon \mapsto x(\varepsilon)$ with $x(0) = x_0$ such that

$$P_\varepsilon(x(\varepsilon)) \equiv x_\varepsilon .$$

Remark

- Often the conditions of Theorem 2 are not easy to verify. Sometimes it is useful here to use Floquet Theory, see [97]. In fact, if T_0 is the period of γ_0 and Ω_0 its Floquet matrix, then $\mathrm{D}_x P_0(x_0) = \exp(T_0 \Omega_0)$.
- The format of the Theorems 1 and 2 with the perturbation parameter ε directly allows for algorithmic approaches. One way to proceed is by perturbation series, leading to asymptotic formulae that in the real analytic setting have positive radius of convergence. In the latter case the names of Poincaré and Lindstedt are associated with the method, cf. [10].
  Also numerical continuation programmes exist based on the Newton method.
- The Theorems 1 and 2 can be seen as special cases of a a general theorem for *normally hyperbolic invariant manifolds* [73], Theorem 4.1. In all cases a contraction principle on a suitable Banach space of graphs leads to persistence of the invariant dynamical object.

  This method in particular yields existence and persistence of stable and unstable manifolds [53,54].

Another type of dynamics subject to perturbation theory is quasi-periodic. We emphasize that persistence of (Diophantine) quasi-periodic invariant tori occurs both in the conservative setting and in many others, like in the reversible and the general (dissipative) setting. In the latter case this leads to persistent occurrence of families of quasi-periodic attractors [125]. These results are in the domain of Kolmogorov–Arnold–Moser (KAM) theory. For details we refer to Sect. "KAM Theory: An Overview" below or to [24], ► Kolmogorov–Arnold–Moser (KAM) Theory, the former reference containing more than 400 references in this area.

Remark

- Concerning the Solar System, KAM theory always has aimed at proving that it contains many quasi-periodic motions, in the sense of positive Liouville measure. This would imply that there is positive probability that a given initial condition lies on such a stable quasi-periodic motion [3,63], however, also see [85].
- Another type of result in this direction compares the distance of certain individual solutions of the perturbed and the unperturbed system, with coinciding initial conditions over time scales that are long in terms of ε. Compare with [24].

Apart from persistence properties related to invariant manifolds or individual solutions, the aim can also be to obtain a more global persistence result. As an example of this we mention the Hartman–Grobman Theorem, e. g., [7,116,123]. Here the setting once more is

$$\dot{x} = Ax + f(x)\,, \quad x \in \mathbb{R}^n\,,$$

with $A \in \mathrm{gl}(n, \mathbb{R})$, $f(0) = 0$ and $\mathrm{D}_x f(0) = 0$. Now we assume A to be hyperbolic (i. e., with no purely imaginary eigenvalues). In that case the full system, near the origin, is topologically conjugated to the linear system $\dot{x} = Ax$. Therefore all global, *qualitative* properties of the unperturbed (linear) system are persistent under perturbation to the full system. For details on these notions see the above references, also compare with, e. g., [30].

It is said that the hyperbolic linear system $\dot{x} = Ax$ is (locally) *structurally stable*. This kind of thinking was introduced to the dynamical systems area by Thom [133], with a first, successful application to catastrophe theory. For further details, see [7,30,69,116].

General Dynamics

We give a few remarks on the general dynamics in a neighborhood of Hamiltonian KAM tori. In particular this concerns so-called superexponential stickiness of the KAM tori and adiabatic stability of the action variables, involving the so-called Nekhoroshev estimate.

To begin with, emphasize the following difference between the cases $n = 2$ and $n \geq 3$ in the classical KAM theorem of Subsect. "Classical KAM Theory". For $n = 2$ the level surfaces of the Hamiltonian are three-dimensional, while the Lagrangian tori have dimension two and hence codimension one in the energy hypersurfaces. This means that for open sets of initial conditions, the evolution curves are forever trapped in between KAM tori, as these tori foliate over nowhere dense sets of positive measure. This implies perpetual adiabatic stability of the action variables. In contrast, for $n \geq 3$ the Lagrangian tori have codimension $n - 1 > 1$ in the energy hypersurfaces and evolution curves may escape.

This actually occurs in the case of so-called *Arnold diffusion*. The literature on this subject is immense, and we here just quote [5,9,93,109], for many more references see [24].

Next we consider the motion in a neighborhood of the KAM tori, in the case where the systems are real analytic or at least Gevrey smooth. For a definition of Gevrey regularity see [136]. First we mention that, measured in terms of the distance to the KAM torus, nearby evolution curves generically stay nearby over a *superexponentially* long time [102,103]. This property often is referred to as superexponential stickiness of the KAM tori, see [24] for more references.

Second, nearly integrable Hamiltonian systems, in terms of the perturbation size, generically exhibit *exponentially* long adiabatic stability of the action variables, see e. g. [15,88,89,90,93,103,109,110,113,120], ▶ Nekhoroshev Theory and many others, for more references see [24]. This property is referred to as the *Nekhoroshev estimate* or the *Nekhoroshev theorem*. For related work on perturbations of so-called superintegrable systems, also see [24] and references therein.

Chaos

In the previous subsection we discussed persistent and some non-persistent features of dynamical systems under small perturbations. Here we discuss properties related to splitting of separatrices, caused by generic perturbations.

A first example was met earlier, when comparing the pendulum with and without (small) damping. The unperturbed system is the undamped one and this is a Hamiltonian system. The perturbation however no longer is Hamiltonian. We see that the equilibria are persistent, as they should be according to Theorem 1, but that none of the periodic orbits survives the perturbation. Such qualitative changes go with perturbing away from the Hamiltonian setting.

Similar examples concern the breaking of a certain symmetry by the perturbation. The latter often occurs in the case of normal form approximations. Then the normalized truncation is viewed as the unperturbed system, which is perturbed by the higher order terms. The truncation often displays a reasonable amount of symmetry (e. g., toroidal symmetry), which *generically* is forbidden for the class of systems under consideration, e. g. see [25].

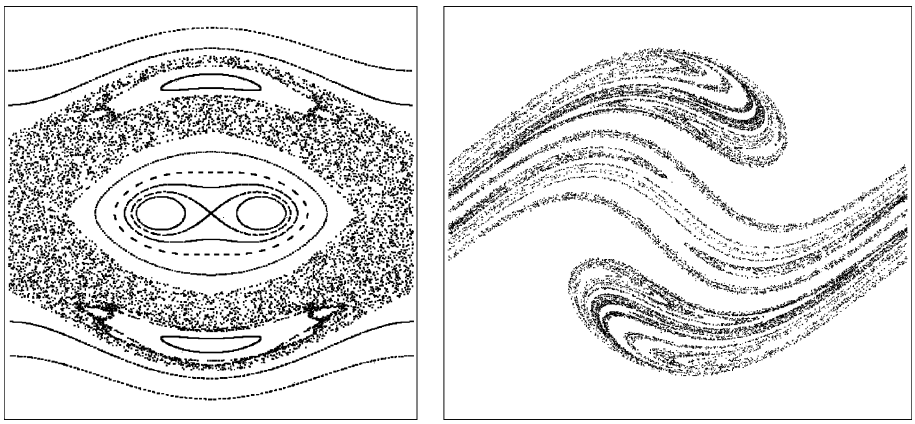

Hamiltonian Perturbation Theory (and Transition to Chaos), Figure 1
Chaos in the parametrically forced pendulum. *Left:* Poincaré map $P_{\omega,\varepsilon}$ near the 1 : 2 resonance $\omega = \frac{1}{2}$ and for $\varepsilon > 0$ not too small. *Right:* A dissipative analogue

To fix thoughts we reconsider the conservative example

$$\ddot{x} + (\omega^2 + \varepsilon \cos t) \sin x = 0$$

of the previous section. The corresponding (time dependent, Hamiltonian [6]) vector field reads

$$\begin{aligned} \dot{t} &= 1 \\ \dot{x} &= y \\ \dot{y} &= -(\omega^2 + \varepsilon \cos t) \sin x \,. \end{aligned}$$

Let $P_{\omega,\varepsilon} \colon \mathbb{R}^2 \to \mathbb{R}^2$ be the corresponding (area-preserving) Poincaré map. Let us consider the unperturbed map $P_{\omega,0}$ which is just the flow over time 2π of the free pendulum $\ddot{x} + \omega^2 \sin x = 0$. Such a map is called *integrable*, since it is the stroboscopic map of a two-dimensional vector field, hence displaying the $\mathbb{R}$-symmetry of a flow. When perturbed to the *nearly integrable* case $\varepsilon \neq 0$, this symmetry generically is broken. We list a few of the generic properties for such maps [123]:

- The homoclinic and heteroclinic points occur at transversal intersections of the corresponding stable and unstable manifolds.
- The periodic points of period less than a given bound are isolated.

This means generically that the separatrices split and that the resonant invariant circles filled with periodic points with the same (rational) rotation number fall apart. In any concrete example the issue remains whether or not it satisfies appropriate genericity conditions. One method to check this is due to Melnikov, compare [66,137], for more sophisticated tools see [65]. Often this leads to elliptic (Abelian) integrals.

In nearly integrable systems chaos can occur. This fact is at the heart of the celebrated non-integrability of the three-body problem as addressed by Poincaré [12,59,107,108,118]. A long standing open conjecture is that the clouds of points as visible in Fig. 1, left, densely fill sets of positive area, thereby leading to ergodicity [9].

In the case of dissipation, see Fig. 1, right, we conjecture the occurrence of a Hénon-like strange attractor [14,22,126].

Remark

- The persistent occurrence of periodic points of a given rotation number follows from the Poincaré–Birkhoff fixed point theorem [74,96,107], i. e., on topological grounds.
- The above arguments are not restricted to the conservative setting, although quite a number of unperturbed systems come from this world. Again see Fig. 1.

One Degree of Freedom

Planar Hamiltonian systems are always integrable and the orbits are given by the level sets of the Hamiltonian function. This still leaves room for a perturbation theory. The recurrent dynamics consists of periodic orbits, equilibria and asymptotic trajectories forming the (un)stable manifolds of unstable equilibria. The equilibria organize the phase portrait, and generically all equilibria are elliptic (purely imaginary eigenvalues) or hyperbolic (real eigenvalues), i. e. there is no equilibrium with a vanishing eigenvalue. If the system depends on a parameter such van-

ishing eigenvalues may be unavoidable and it becomes possible that the corresponding dynamics persist under perturbations.

Perturbations may also destroy the Hamiltonian character of the flow. This happens especially where the starting point is a dissipative planar system and e. g. a scaling leads for $\varepsilon = 0$ to a limiting Hamiltonian flow. The perturbation problem then becomes twofold. Equilibria still persist by Theorem 1 and hyperbolic equilibria moreover persist as such, with the sum of eigenvalues of order $\mathcal{O}(\varepsilon)$. Also for elliptic eigenvalues the sum of eigenvalues is of order $\mathcal{O}(\varepsilon)$ after the perturbation, but here this number measures the dissipation whence the equilibrium becomes (weakly) attractive for negative values and (weakly) unstable for positive values. The one-parameter families of periodic orbits of a Hamiltonian system do not persist under dissipative perturbations, the very fact that they form families imposes the corresponding fixed point of the Poincaré mapping to have an eigenvalue one and Theorem 2 does not apply. Typically only finitely many periodic orbits survive a dissipative perturbation and it is already a difficult task to determine their number.

Hamiltonian Perturbations

The Duffing oscillator has the Hamiltonian function

$$H(x, y) = \frac{1}{2}y^2 + \frac{1}{24}bx^4 + \frac{1}{2}\lambda x^2 \tag{2}$$

where b is a constant distinguishing the two cases $b = \pm 1$ and λ is a parameter. Under variation of the parameter the equations of motion

$$\begin{aligned} \dot{x} &= y \\ \dot{y} &= -\frac{1}{6}bx^3 - \lambda x \end{aligned}$$

display a Hamiltonian pitchfork bifurcation, supercritical for positive b and subcritical in case b is negative. Correspondingly, the linearization at the equilibrium $x = 0$ of the anharmonic oscillator $\lambda = 0$ is given by the matrix

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

whence this equilibrium is parabolic.

The typical way in which a parabolic equilibrium bifurcates is the center-saddle bifurcation. Here the Hamiltonian reads

$$H(x, y) = \frac{1}{2}ay^2 + \frac{1}{6}bx^3 + c\lambda x \tag{3}$$

where $a, b, c \in \mathbb{R}$ are nonzero constants, for instance $a = b = c = 1$. Note that this is a completely different unfolding of the parabolic equilibrium at the origin. A closer look at the phase portraits and in particular at the Hamiltonian function of the Hamiltonian pitchfork bifurcation reveals the symmetry $x \mapsto -x$ of the Duffing oscillator. This suggests the addition of the non-symmetric term μx. The resulting two-parameter family

$$H_{\lambda,\mu}(x, y) = \frac{1}{2}y^2 + \frac{1}{24}bx^4 + \frac{1}{2}\lambda x^2 + \mu x$$

of Hamiltonian systems is indeed structurally stable. This implies not only that all equilibria of a Hamiltonian perturbation of the Duffing oscillator have a local flow equivalent to the local flow near a suitable equilibrium in this two-parameter family, but that every one-parameter family of $\mathbb{Z}_2$-symmetric Hamiltonian systems that is a perturbation of (2) has equivalent dynamics. For more details see [36] and references therein.

This approach applies mutatis mutandis to every non-degenerate planar singularity, cf. [69,130]. At an equilibrium all partial derivatives of the Hamiltonian vanish and the resulting singularity is called non-degenerate if it has finite multiplicity, which implies that it admits a versal unfolding H_λ with finitely many parameters. The family of Hamiltonian systems defined by this versal unfolding contains all possible (local) dynamics that the initial equilibrium may be perturbed to. Imposing additional discrete symmetries is immediate, the necessary symmetric versal unfolding is obtained by averaging

$$H_\lambda^G = \frac{1}{|G|}\sum_{g\in G} H_\lambda \circ g$$

along the orbits of the symmetry group G.

Dissipative Perturbations

In a generic dissipative system all equilibria are hyperbolic. Qualitatively, i. e. up to topological equivalence, the local dynamics is completely determined by the number of eigenvalues with positive real part. Those hyperbolic equilibria that can appear in Hamiltonian systems (the eigenvalues forming pairs $\pm\nu$) do not play an important role. Rather, planar Hamiltonian systems become important as a tool to understand certain bifurcations triggered off by non-hyperbolic equilibria. Again this requires the system to depend on external parameters.

The simplest example is the Hopf bifurcation, a co-dimension one bifurcation where an equilibrium loses stability as the pair of eigenvalues crosses the imaginary axis,

say at $\pm$i. At the bifurcation the linearization is a Hamiltonian system with an elliptic equilibrium (the co-dimension one bifurcations where a single eigenvalue crosses the imaginary axis through 0 do not have a Hamiltonian linearization). This limiting Hamiltonian system has a one-parameter family of periodic orbits around the equilibrium, and the non-linear terms determine the fate of these periodic orbits. The normal form of order three reads

$$\begin{aligned}\dot{x} &= -y\left(1 + b(x^2 + y^2)\right) + x\left(\lambda + a(x^2 + y^2)\right)\\ \dot{x} &= x\left(1 + b(x^2 + y^2)\right) + y\left(\lambda + a(x^2 + y^2)\right)\end{aligned}$$

and is Hamiltonian if and only if $(\lambda, a) = (0, 0)$. The sign of the coefficient distinguishes between the supercritical case $a > 0$, in which there are no periodic orbits coexisting with the attractive equilibria (i. e. when $\lambda < 0$) and one attracting periodic orbit for each $\lambda > 0$ (coexisting with the unstable equilibrium), and the subcritical case $a < 0$, in which the family of periodic orbits is unstable and coexists with the attractive equilibria (with no periodic orbits for parameters $\lambda > 0$). As $\lambda \to 0$ the family of periodic orbits shrinks down to the origin, so also this Hamiltonian feature is preserved.

Equilibria with a double eigenvalue 0 need two parameters to persistently occur in families of dissipative systems. The generic case is the Takens–Bogdanov bifurcation. Here the linear part is too degenerate to be helpful, but the nonlinear Hamiltonian system defined by (3) with $a = 1 = c\lambda$ and $b = -3$ provides the periodic and heteroclinic orbit(s) that constitute the nontrivial part of the bifurcation diagram. Where discrete symmetries are present, e. g. for equilibria in dissipative systems originating from other generic bifurcations, the limiting Hamiltonian system exhibits that same discrete symmetry. For more details see [54,66,82] and references therein.

The continuation of certain periodic orbits from an unperturbed Hamiltonian system under dissipative perturbation can be based on Melnikov-like methods, again see [66,137]. As above, this often leads to Abelian integrals, for instance to count the number of periodic orbits that branch off.

Reversible Perturbations

A dynamical system that admits a reflection symmetry R mapping trajectories $\varphi(t, z_0)$ to trajectories $\varphi(-t, R(z_0))$ is called reversible. In the planar case we may restrict to the reversing reflection

$$\begin{array}{rccc} R: & \mathbb{R}^2 & \longrightarrow & \mathbb{R}^2 \\ & (x, y) & \mapsto & (x, -y)\,. \end{array} \tag{4}$$

All Hamiltonian functions $H = \frac{1}{2}y^2 + V(x)$ which have an interpretation "kinetic + potential energy" are reversible, and in general the class of reversible systems is positioned between the class of Hamiltonian systems and the class of dissipative systems. A guiding example is the perturbed Duffing oscillator (with the roles of x and y exchanged so that (4) remains the reversing symmetry)

$$\begin{aligned}\dot{x} &= -\frac{1}{6}y^3 - y + \varepsilon xy\\ \dot{y} &= x\end{aligned}$$

that combines the Hamiltonian character of the equilibrium at the origin with the dissipative character of the two other equilibria. Note that all orbits outside the homoclinic loop are periodic.

There are two ways in which the reversing symmetry (4) imposes a Hamiltonian character on the dynamics. An equilibrium that lies on the symmetry line $\{y = 0\}$ has a linearization that is itself a reversible system and consequently the eigenvalues are subject to the same constraints as in the Hamiltonian case. (For equilibria z_0 that do not lie on the symmetry line the reflection $R(z_0)$ is also an equilibrium, and it is to the union of their eigenvalues that these constraints still apply.) Furthermore, every orbit that crosses $\{y = 0\}$ more than once is automatically periodic, and these periodic orbits form one-parameter families. In particular, elliptic equilibria are still surrounded by periodic orbits.

The dissipative character of a reversible system is most obvious for orbits that do not cross the symmetry line. Here R merely maps the orbit to a reflected counterpart. The above perturbed Duffing oscillator exemplifies that the character of an orbit crossing $\{y = 0\}$ exactly once is undetermined. While the homoclinic orbit of the saddle at the origin has a Hamiltonian character, the heteroclinic orbits between the other two equilibria behave like in a dissipative system.

Perturbations of Periodic Orbits

The perturbation of a one-degree-of-freedom system by a periodic forcing is a perturbation that changes the phase space. Treating the time variable t as a phase space variable leads to the extended phase space $\mathbb{S}^1 \times \mathbb{R}^2$ and equilibria of the unperturbed system become periodic orbits, inheriting the normal behavior. Furthermore introducing an action conjugate to the "angle" t yields a Hamiltonian system in two degrees of freedom.

While the one-parameter families of periodic orbits merely provide the typical recurrent motion in one degree of freedom, they form special solutions in two or more de-

grees of freedom. Arcs of elliptic periodic orbits are particularly instructive. Note that these occur generically in both the Hamiltonian and the reversible context.

Conservative Perturbations

Along the family of elliptic periodic orbits a pair $e^{\pm i\Omega}$ of Floquet multipliers passes regularly through roots of unity. Generically this happens on a dense set of parameter values, but for fixed denominator q in $e^{\pm i\Omega} = e^{\pm 2\pi i p/q}$ the corresponding energy values are isolated. The most important of such resonances are those with small denominators q.

For $q = 1$ generically a periodic center-saddle bifurcation takes place where an elliptic and a hyperbolic periodic orbit meet at a parabolic periodic orbit. No periodic orbit remains under further variation of a suitable parameter.

The generic bifurcation for $q = 2$ is the period-doubling bifurcation where an elliptic periodic orbit turns hyperbolic (or vice versa) when passing through a parabolic periodic orbit with Floquet multipliers -1. Furthermore, a family of periodic orbits with twice the period emerges from the parabolic periodic orbit, inheriting the normal linear behavior from the initial periodic orbit.

In case $q = 3$, and possibly also for $q = 4$, generically two arcs of hyperbolic periodic orbits emerge, both with three (resp. four) times the period. One of these extends for lower and the other for higher parameter values. The initial elliptic periodic orbit momentarily loses its stability due to these approaching unstable orbits.

Denominators $q \geq 5$ (and also the second possibility for $q = 4$) lead to a pair of subharmonic periodic orbits of q times the period emerging either for lower or for higher parameter values. This is (especially for large q) comparable to the behavior at Diophantine $e^{\pm i\Omega}$ where a family of invariant tori emerges, cf. Sect. "Invariant Curves of Planar Diffeomorphisms" below.

For a single pair $e^{\pm i\Omega}$ of Floquet multipliers this behavior is traditionally studied for the (iso-energetic) Poincaré-mapping, cf. [92] and references therein. However, the above description remains true in higher dimensions, where additionally multiple pairs of Floquet multipliers may interact. An instructive example is the Lagrange top, the sleeping motion of which is gyroscopically stabilized after a periodic Hamiltonian Hopf bifurcation; see [56] for more details.

Dissipative Perturbations

There exists a large class of local bifurcations in the dissipative setting, that can be arranged in a perturbation theory setting, where the unperturbed system is Hamiltonian. The arrangement consists of changes of variables and rescaling. An early example of this is the Bogdanov–Takens bifurcation [131,132]. For other examples regarding nilpotent singularities, see [23,40] and references therein.

To fix thoughts, consider families of planar maps and let the unperturbed Hamiltonian part contain a center (possibly surrounded by a homoclinic loop). The question then is which of these persist when adding the dissipative perturbation.

Usually only a definite finite number persists. As in Subsect. "Chaos", a Melnikov function can be invoked here, possibly again leading to elliptic (Abelian) integrals, Picard Fuchs equations, etc. For details see [61,124] and references therein.

Invariant Curves of Planar Diffeomorphisms

This section starts with general considerations on circle diffeomorphisms, in particular focusing on persistence properties of quasi-periodic dynamics. Our main references are [2,24,29,31,70,71,139,140]. For a definition of rotation number, see [58]. After this we turn to area preserving maps of an annulus where we discuss Moser's twist map theorem [104], also see [24,29,31]. The section is concluded by a description of the holomorphic linearization of a fixed point in a planar map [7,101,141,142].

Our main perspective will be perturbative, where we consider circle maps near a rigid rotation. It turns out that generally parameters are needed for persistence of quasi-periodicity under perturbations. In the area preserving setting we consider perturbations of a pure twist map.

Circle Maps

We start with the following general problem. Given a two-parameter family

$$P_{\alpha,\varepsilon}\colon \mathbb{T}^1 \to \mathbb{T}^1, \quad x \mapsto x + 2\pi\alpha + \varepsilon a(x,\alpha,\varepsilon)$$

of circle maps of class C^∞. It turns out to be convenient to view this two-parameter family as a one-parameter family of maps

$$P_\varepsilon\colon \mathbb{T}^1 \times [0,1] \to \mathbb{T}^1 \times [0,1],$$
$$(x,\alpha) \mapsto (x + 2\pi\alpha + \varepsilon a(x,\alpha,\varepsilon), \alpha)$$

of the cylinder. Note that the unperturbed system P_0 is a family of rigid circle rotations, viewed as a cylinder map, where the individual map $P_{\alpha,0}$ has rotation number α. The question now is what will be the fate of this rigid dynamics for $0 \neq |\varepsilon| \ll 1$.

The classical way to address this question is to look for a conjugation Φ_ε, that makes the following diagram commute

$$\begin{array}{ccc} \mathbb{T}^1 \times [0,1] & \xrightarrow{P_\varepsilon} & \mathbb{T}^1 \times [0,1] \\ \uparrow \Phi_\varepsilon & & \uparrow \Phi_\varepsilon \\ \mathbb{T}^1 \times [0,1] & \xrightarrow{P_0} & \mathbb{T}^1 \times [0,1] \,, \end{array}$$

i. e., such that

$$P_\varepsilon \circ \Phi_\varepsilon = \Phi_\varepsilon \circ P_0 \,.$$

Due to the format of P_ε we take Φ_ε as a skew map

$$\Phi_\varepsilon(x,\alpha) = (x + \varepsilon U(x,\alpha,\varepsilon), \alpha + \varepsilon\sigma(\alpha,\varepsilon)) \,,$$

which leads to the *nonlinear* equation

$$\begin{aligned} &U(x + 2\pi\alpha, \alpha, \varepsilon) - U(x,\alpha,\varepsilon) \\ &= 2\pi\sigma(\alpha,\varepsilon) + a\,(x + \varepsilon U(x,\alpha,\varepsilon), \alpha + \varepsilon\sigma(\alpha,\varepsilon), \varepsilon) \end{aligned}$$

in the unknown maps U and σ. Expanding in powers of ε and comparing at lowest order yields the linear equation

$$U_0(x + 2\pi\alpha, \alpha) - U_0(x,\alpha) = 2\pi\sigma_0(\alpha) + a_0(x,\alpha)$$

which can be directly solved by Fourier series. Indeed, writing

$$\begin{aligned} a_0(x,\alpha) &= \sum_{k\in\mathbf{Z}} a_{0k}(\alpha) \mathrm{e}^{\mathrm{i}kx} \,, \\ U_0(x,\alpha) &= \sum_{k\in\mathbf{Z}} U_{0k}(\alpha) \mathrm{e}^{\mathrm{i}kx} \end{aligned}$$

we find $\sigma_0 = -1/(2\pi) a_{00}$ and

$$U_{0k}(\alpha) = \frac{a_{0k}(\alpha)}{\mathrm{e}^{2\pi\mathrm{i}k\alpha} - 1} \,.$$

It follows that in general a formal solution exists if and only if $\alpha \in \mathbb{R} \setminus \mathbb{Q}$. Still, the accumulation of $\mathrm{e}^{2\pi\mathrm{i}k\alpha} - 1$ on 0 leads to the celebrated *small divisors* [9,108], also see [24,29,31,55].

The classical solution considers the following Diophantine non-resonance conditions. Fixing $\tau > 2$ and $\gamma > 0$ consider $\alpha \in [0,1]$ such that for all rationals p/q

$$\left|\alpha - \frac{p}{q}\right| \geq \gamma q^{-\tau} \,. \tag{5}$$

This subset of such αs is denoted by $[0,1]_{\tau,\gamma}$ and is well-known to be nowhere dense but of large measure as $\gamma > 0$ gets small [115]. Note that Diophantine numbers are irrational.

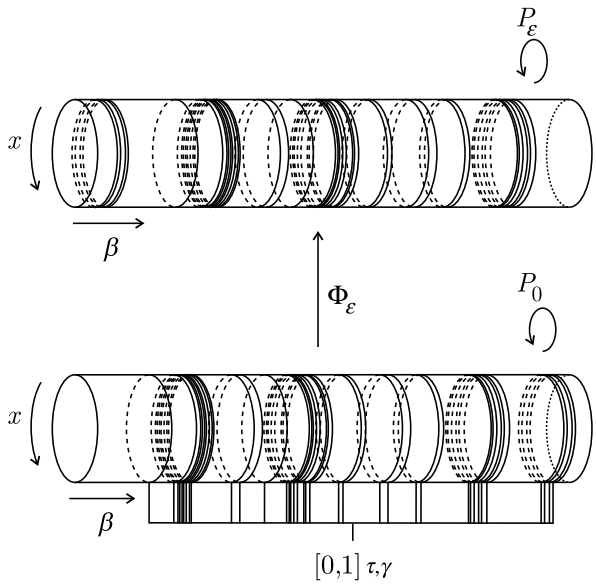

Hamiltonian Perturbation Theory (and Transition to Chaos), Figure 2
Skew cylinder map, conjugating (Diophantine) quasi-periodic invariant circles of P_0 and P_ε

Theorem 3 (Circle Map Theorem) *For γ sufficiently small and for the perturbation εa sufficiently small in the C^∞-topology, there exists a C^∞ transformation $\Phi_\varepsilon \colon \mathbb{T}^1 \times [0,1] \to \mathbb{T}^1 \times [0,1]$, conjugating the restriction $P_0|_{[0,1]_{\tau,\gamma}}$ to a subsystem of P_ε.*

Theorem 3 in the present structural stability formulation (compare with Fig. 2) is a special case of the results in [29,31]. We here speak of *quasi-periodic stability*. For earlier versions see [2,9].

Remark

- Rotation numbers are preserved by the map Φ_ε and irrational rotation numbers correspond to quasi-periodicity. Theorem 3 thus ensures that *typically* quasi-periodicity occurs with *positive* measure in the parameter space. Note that since Cantor sets are perfect, quasi-periodicity typically has a non-isolated occurrence.
- The map Φ_ε has no dynamical meaning inside the gaps. The gap dynamics in the case of circle maps can be illustrated by the Arnold family of circle maps [2,7,58], given by

  $$P_{\alpha,\varepsilon}(x) = x + 2\pi\alpha + \varepsilon \sin x$$

  which exhibits a countable union of open resonance tongues where the dynamics is periodic, see Fig. 3. Note that this map only is a diffeomorphism for $|\varepsilon| < 1$.

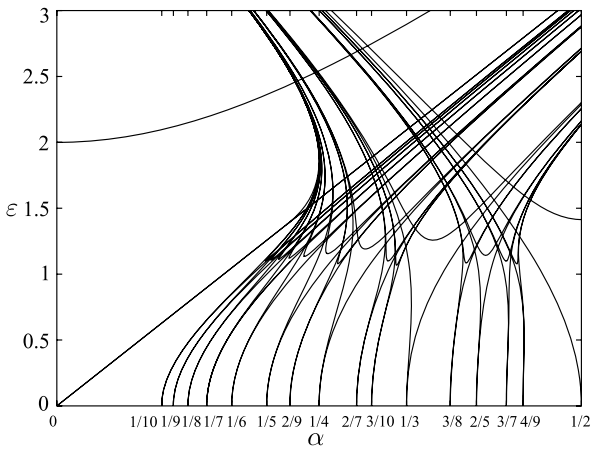

Hamiltonian Perturbation Theory (and Transition to Chaos), Figure 3
Arnold resonance tongues; for $\varepsilon \geq 1$ the maps are endomorphic

- We like to mention that non-perturbative versions of Theorem 3 have been proven in [70,71,139].
- For simplicity we formulated Theorem 3 under C^∞-regularity, noting that there exist many ways to generalize this. On the one hand there exist C^k-versions for finite k and on the other hand there exist fine tunings in terms of real-analytic and Gevrey regularity. For details we refer to [24,31] and references therein. This same remark applies to other results in this section and in Sect. "KAM Theory: An Overview" on KAM theory.

A possible application of Theorem 3 runs as follows. Consider a system of weakly coupled Van der Pol oscillators

$$\ddot{y}_1 + c_1\dot{y}_1 + a_1 y_1 + f_1(y_1, \dot{y}_1) = \varepsilon g_1(y_1, y_2, \dot{y}_1, \dot{y}_2)$$
$$\ddot{y}_2 + c_2\dot{y}_2 + a_2 y_2 + f_2(y_2, \dot{y}_2) = \varepsilon g_2(y_1, y_2, \dot{y}_1, \dot{y}_2)\,.$$

Writing $\dot{y}_j = z_j$, $j = 1, 2$, one obtains a vector field in the four-dimensional phase space $\mathbb{R}^2 \times \mathbb{R}^2 = \{(y_1, z_1), (y_2, z_2)\}$. For $\varepsilon = 0$ this vector field has an invariant two-torus, which is the product of the periodic motions of the individual Van der Pol oscillations. This two-torus is normally hyperbolic and therefore persistent for $|\varepsilon| \ll 1$ [73]. In fact the torus is an attractor and we can define a Poincaré return map within this torus attractor. If we include some of the coefficients of the equations as parameters, Theorem 3 is directly applicable. The above statements on quasi-periodic circle maps then directly translate to the case of quasi-periodic invariant two-tori. Concerning the resonant cases, generically a tongue structure like in Fig. 3 occurs; for the dynamics corresponding to parameter values inside such a tongue one speaks of *phase lock*.

Remark

- The celebrated synchronization of Huygens' clocks [77] is related to a $1:1$ resonance, meaning that the corresponding Poincaré map would have its parameters in the main tongue with rotation number 0. Compare with Fig. 3.
- There exist direct generalizations to cases with n-oscillators ($n \in \mathbb{N}$), leading to families of invariant n-tori carrying quasi-periodic flow, forming a nowhere dense set of positive measure. An alteration with resonance occurs as roughly sketched in Fig. 3. In higher dimension the gap dynamics, apart from periodicity, also can contain strange attractors [112,126]. We shall come back to this subject in a later section.

Area-Preserving Maps

The above setting historically was preceded by an area preserving analogue [104] that has its origin in the Hamiltonian dynamics of frictionless mechanics.

Let $\Delta \subseteq \mathbb{R}^2 \setminus \{(0,0)\}$ be an annulus, with sympectic polar coordinates $(\varphi, I) \in \mathbb{T}^1 \times \mathbf{K}$, where $\mathbf{K}$ is an interval. Moreover, let $\sigma = \mathrm{d}\varphi \wedge \mathrm{d}I$ be the area form on Δ.

We consider a σ-preserving smooth map $P_\varepsilon : \Delta \to \Delta$ of the form

$$P_\varepsilon(\varphi, I) = (\varphi + 2\pi\alpha(I), I) + O(\varepsilon)\,,$$

where we assume that the map $I \mapsto \alpha(I)$ is a (local) diffeomorphism. This assumption is known as the *twist condition* and P_ε is called a *twist map*. For the unperturbed case $\varepsilon = 0$ we are dealing with a pure twist map and its dynamics are comparable to the unperturbed family of cylinder maps as met in Subsect. "Circle Maps". Indeed it is again a family of rigid rotations, parametrized by I and where $P_0(., I)$ has rotation number $\alpha(I)$. In this case the question is what will be the fate of this family of invariant circles, as well as with the corresponding rigidly rotational dynamics.

Regarding the rotation number we again introduce Diophantine conditions. Indeed, for $\tau > 2$ and $\gamma > 0$ the subset $[0,1]_{\tau,\gamma}$ is defined as in (5), i. e., it contains all $\alpha \in [0,1]$, such that for all rationals p/q

$$\left|\alpha - \frac{p}{q}\right| \geq \gamma q^{-\tau}\,.$$

Pulling back $[0,1]_{\tau,\gamma}$ along the map α we obtain a subset $\Delta_{\tau,\gamma} \subseteq \Delta$.

Theorem 4 (Twist Map Theorem [104]) *For γ sufficiently small, and for the perturbation $O(\varepsilon)$ sufficiently*

small in C^∞-topology, there exists a C^∞ transformation $\Phi_\varepsilon\colon \Delta \to \Delta$, conjugating the restriction $P_0|_{\Delta_{\tau,\gamma}}$ to a subsystem of P_ε.

As in the case of Theorem 3 again we chose the formulation of [29,31]. Largely the remarks following Theorem 3 also apply here.

Remark

- Compare the format of the Theorems 3 and 4 and observe that in the latter case the role of the parameter α has been taken by the action variable I. Theorem 4 implies that *typically* quasi-periodicity occurs with positive measure in phase space.
- In the gaps typically we have coexistence of periodicity, quasi-periodicity and chaos [6,9,35,107,108,123,137]. The latter follows from transversality of homo- and heteroclinic connections that give rise to positive topological entropy. Open problems are whether the corresponding Lyapunov exponents also are positive, compare with the discussion at the end of the introduction.

Similar to the applications of Theorem 3 given at the end of Subsect. "Circle Maps", here direct applications are possible in the conservative setting. Indeed, consider a system of weakly coupled pendula

$$\ddot{y}_1 + \alpha_1^2 \sin y_1 = \varepsilon \frac{\partial U}{\partial y_1}(y_1, y_2)$$
$$\ddot{y}_2 + \alpha_2^2 \sin y_2 = \varepsilon \frac{\partial U}{\partial y_2}(y_1, y_2)\,.$$

Writing $\dot{y}_j = z_j$, $j = 1, 2$ as before, we again get a vector field in the four-dimensional phase space $\mathbb{R}^2 \times \mathbb{R}^2 = \{(y_1, y_2), (z_1, z_2)\}$. In this case the energy

$$H_\varepsilon(y_1, y_2, z_1, z_2) = \frac{1}{2}z_1^2 + \frac{1}{2}z_2^2 - \alpha_1^2 \cos y_1 - \alpha_2^2 \cos y_2 + \varepsilon U(y_1, y_2)$$

is a constant of motion. Restricting to a three-dimensional energy surface $H_\varepsilon^{-1} = \text{const.}$, the iso-energetic Poincaré map P_ε is a twist map and application of Theorem 4 yields the conclusion of quasi-periodicity (on invariant two-tori) occurring with positive measure in the energy surfaces of H_ε.

Remark As in the dissipative case this example directly generalizes to cases with n oscillators ($n \in \mathbb{N}$), again leading to invariant n-tori with quasi-periodic flow. We shall return to this subject in a later section.

Linearization of Complex Maps

The Subsects. "Circle Maps" and "Area-Preserving Maps" both deal with smooth circle maps that are conjugated to rigid rotations. Presently the concern is with planar holomorphic maps that are conjugated to a rigid rotation on an open subset of the plane. Historically this is the first time that a small divisor problem was solved [7,101,141,142] and ► Perturbative Expansions, Convergence of.

Complex Linearization Given is a holomorphic germ $F\colon (\mathbb{C}, 0) \to (\mathbb{C}, 0)$ of the form $F(z) = \lambda z + f(z)$, with $f(0) = f'(0) = 0$. The problem is to find a biholomorphic germ $\Phi\colon (\mathbb{C}, 0) \to (\mathbb{C}, 0)$ such that

$$\Phi \circ F = \lambda \cdot \Phi\,.$$

Such a diffeomorphism Φ is called a *linearization* of F near 0.

We begin with the formal approach. Given the series $f(z) = \sum_{j\geq 2} f_j z^j$, we look for $\Phi(z) = z + \sum_{j\geq 2} \phi_j z^j$. It turns out that a solution always exists whenever $\lambda \neq 0$ is not a root of unity. Indeed, direct computation reveals the following set of equations that can be solved recursively:
For j = 2 get the equation $\lambda(1-\lambda)\phi_2 = f_2$
For j = 3 get the equation $\lambda(1-\lambda^2)\phi_3 = f_3 + 2\lambda f_2 \phi_2$
For j = n get the equation $\lambda(1-\lambda^{n-1})\phi_n = f_n + \text{known}$.
The question now reduces to whether this formal solution has a positive radius of convergence.

The hyperbolic case $0 < |\lambda| \neq 1$ was already solved by Poincaré, for a description see [7]. The elliptic case $|\lambda| = 1$ again has small divisors and was solved by Siegel when for some $\gamma > 0$ and $\tau > 2$ we have the Diophantine non-resonance condition

$$|\lambda - \mathrm{e}^{2\pi \mathrm{i} \frac{p}{q}}| \geq \gamma |q|^{-\tau}\,.$$

The corresponding set of λ constitutes a set of full measure in $\mathbb{T}^1 = \{\lambda\}$.

Yoccoz [141] completely solved the elliptic case using the Bruno-condition. If

$$\lambda = \mathrm{e}^{2\pi i \alpha} \quad \text{and when} \quad \frac{p_n}{q_n}$$

is the nth convergent in the continued fraction expansion of α then the Bruno-condition reads

$$\sum_n \frac{\log(q_{n+1})}{q_n} < \infty\,.$$

This condition turns out to be necessary and sufficient for Φ having positive radius of convergence [141,142].

Cremer's Example in Herman's Version As an example consider the map

$$F(z) = \lambda z + z^2 ,$$

where $\lambda \in \mathbb{T}^1$ is not a root of unity.

Observe that a point $z \in \mathbb{C}$ is a periodic point of F with period q if and only if $F^q(z) = z$, where obviously

$$F^q(z) = \lambda^q z + \cdots + z^{2^q} .$$

Writing

$$F^q(z) - z = z\left(\lambda^q - 1 + \cdots + z^{2^q-1}\right) ,$$

the period q periodic points exactly are the roots of the right hand side polynomial. Abbreviating $N = 2^q - 1$, it directly follows that, if $z_1, z_2, \ldots, z_N$ are the nontrivial roots, then for their product we have

$$z_1 \cdot z_2 \cdot \ldots \cdot z_N = \lambda^q - 1 .$$

It follows that there exists a nontrivial root within radius

$$|\lambda^q - 1|^{1/N}$$

of $z = 0$.

Now consider the set of $\Lambda \subset \mathbb{T}^1$ defined as follows: $\lambda \in \Lambda$ whenever

$$\liminf_{q\to\infty} |\lambda^q - 1|^{1/N} = 0 .$$

It can be directly shown that Λ is *residual*, again compare with [115]. It also follows that for $\lambda \in \Lambda$ linearization is impossible. Indeed, since the rotation is irrational, the existence of periodic points in any neighborhood of $z = 0$ implies zero radius of convergence.

Remark

- Notice that the residual set Λ is in the complement of the full measure set of all Diophantine numbers, again see [115].
- Considering $\lambda \in \mathbb{T}^1$ as a parameter, we see a certain analogy of these results on complex linearization with the Theorems 3 and 4. Indeed, in this case for a full measure set of λs on a neighborhood of $z = 0$ the map $F = F_\lambda$ is conjugated to a rigid irrational rotation. Such a domain in the z-plane often is referred to as a Siegel disc. For a more general discussion of these and of Herman rings, see [101].

KAM Theory: An Overview

In Sect. "Invariant Curves of Planar Diffeomorphisms" we described the persistent occurrence of quasi-periodicity in the setting of diffeomorphisms of the circle or the plane. The general perturbation theory of quasi-periodic motions is known under the name Kolmogorov–Arnold–Moser (or KAM) theory and discussed extensively elsewhere in this encyclopedia ▶ Kolmogorov–Arnold–Moser (KAM) Theory. Presently we briefly summarize parts of this KAM theory in broad terms, as this fits in our considerations, thereby largely referring to [4,80,81,119,121,143,144], also see [20,24,55].

In general quasi-periodicity is defined by a smooth conjugation. First on the n-torus $\mathbb{T}^n = \mathbb{R}^n/(2\pi Z)^n$ consider the vector field

$$\mathbb{X}_\omega = \sum_{j=1}^{n} \omega_j \frac{\partial}{\partial \varphi_j} ,$$

where $\omega_1, \omega_2, \ldots, \omega_n$ are called frequencies [43,106]. Now, given a smooth (say, of class C^∞) vector field X on a manifold M, with $T \subseteq M$ an invariant n-torus, we say that the restriction $X|_T$ is *parallel* if there exists $\omega \in \mathbb{R}^n$ and a smooth diffeomorphism $\Phi\colon T \to \mathbb{T}^n$, such that $\Phi_*(X|_T) = \mathbb{X}_\omega$. We say that $X|_T$ is *quasi-periodic* if the frequencies $\omega_1, \omega_2, \ldots, \omega_n$ are independent over $\mathbb{Q}$.

A quasi-periodic vector field $X|_T$ leads to an integer affine structure on the torus T. In fact, since each orbit is dense, it follows that the self conjugations of $\mathbb{X}_\omega$ exactly are the translations of $\mathbb{T}^n$, which completely determine the affine structure of $\mathbb{T}^n$. Then, given $\Phi\colon T \to \mathbb{T}^n$ with $\Phi_*(X|_T) = \mathbb{X}_\omega$, it follows that the self conjugations of $X|_T$ determines a natural affine structure on the torus T. Note that the conjugation Φ is unique modulo translations in T and $\mathbb{T}^n$.

Note that the composition of Φ by a translation of $\mathbb{T}^n$ does not change the frequency vector ω. However, the composition by a linear invertible map $S \in \mathrm{GL}(n, \mathbb{Z})$ yields $S_*\mathbb{X}_\omega = \mathbb{X}_{S\omega}$. We here speak of an *integer affine* structure [43].

Remark

- The transition maps of an integer affine structure are translations and elements of $\mathrm{GL}(n, \mathbb{Z})$.
- The current construction is compatible with the integrable affine structure on the Liouville tori of an integrable Hamiltonian system [6]. Note that in that case the structure extends to all parallel tori.

Classical KAM Theory

The classical KAM theory deals with smooth, nearly integrable Hamiltonian systems of the form

$$\begin{aligned} \dot{\varphi} &= \omega(I) + \varepsilon f(I, \varphi, \varepsilon) \\ \dot{I} &= \varepsilon g(I, \varphi, \varepsilon)\,, \end{aligned} \tag{6}$$

where I varies over an open subset of $\mathbb{R}^n$ and φ over the standard torus $\mathbb{T}^n$. Note that for $\varepsilon = 0$ the phase space as an open subset of $\mathbb{R}^n \times \mathbb{T}^n$ is foliated by invariant tori, parametrized by I. Each of the tori is parametrized by φ and the corresponding motion is parallel (or multi- periodic or conditionally periodic) with frequency vector $\omega(I)$.

Perturbation theory asks for persistence of the invariant n-tori and the parallelity of their motion for small values of $|\varepsilon|$. The answer that KAM theory gives needs two essential ingredients. The first ingredient is that of *Kolmogorov non-degeneracy* which states that the map $I \in \mathbb{R}^n \mapsto \omega(I) \in \mathbb{R}^n$ is a (local) diffeomorphism. Compare with the twist condition of Sect. "Invariant Curves of Planar Diffeomorphisms". The second ingredient generalizes the Diophantine conditions (5) of that section as follows: for $\tau > n-1$ and $\gamma > 0$ consider the set

$$\mathbb{R}^n_{\tau,\gamma} = \{\omega \in \mathbb{R}^n \mid |\langle \omega, k\rangle| \geq \gamma |k|^{-\tau}, k \in \mathbb{Z}^n \setminus \{0\}\}. \tag{7}$$

The following properties are more or less direct. First $\mathbb{R}^n_{\tau,\gamma}$ has a closed half line geometry in the sense that if $\omega \in \mathbb{R}^n_{\tau,\gamma}$ and $s \geq 1$ then also $s\omega \in \mathbb{R}^n_{\tau,\gamma}$. Moreover, the intersection $\mathbb{S}^{n-1} \cap \mathbb{R}^n_{\tau,\gamma}$ is a Cantor set of measure $\mathbb{S}^{n-1} \setminus \mathbb{R}^n_{\tau,\gamma} = O(\gamma)$ as $\gamma \downarrow 0$, see Fig. 4.

Completely in the spirit of Theorem 4, the classical KAM theorem roughly states that a Kolmogorov non-degenerate nearly integrable system $(6)_\varepsilon$, for $|\varepsilon| \ll 1$ is smoothly conjugated to the unperturbed version $(6)_0$, provided that the frequency map ω is co-restricted to the Diophantine set $\mathbb{R}^n_{\tau,\gamma}$. In this formulation smoothness has to be taken in the sense of Whitney [119,136], also compare with [20,24,29,31,55,121].

As a consequence we may say that in Hamiltonian systems of n degrees of freedom typically quasi-periodic invariant (Lagrangian) n-tori occur with positive measure in phase space. It should be said that also an iso-energetic version of this classical result exists, implying a similar conclusion restricted to energy hypersurfaces [6,9,21,24]. The Twist Map Theorem 4 is closely related to the iso-energetic KAM Theorem.

Remark

- We chose the quasi-periodic stability format as in Sect. "Invariant Curves of Planar Diffeomorphisms". For regularity issues compare with a remark following Theorem 3.
- For applications we largely refer to the introduction and to [24,31] and references therein.
- Continuing the discussion on affine structures at the beginning of this section, we mention that by means of the symplectic form, the domain of the I-variables in $\mathbb{R}^n$ inherits an affine structure [60], also see [91] and references therein.

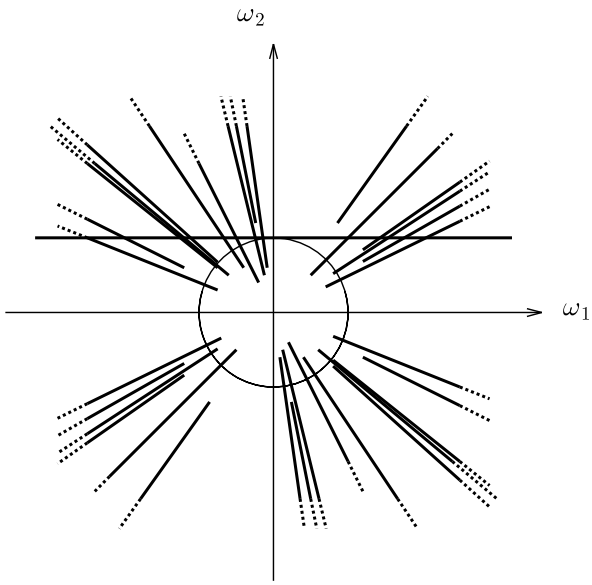

Hamiltonian Perturbation Theory (and Transition to Chaos), Figure 4
The Diophantine set $\mathbb{R}^n_{\tau,\gamma}$ has the closed half line geometry and the intersection $\mathbb{S}^{n-1} \cap \mathbb{R}^n_{\tau,\gamma}$ is a Cantor set of measure $\mathbb{S}^{n-1} \setminus \mathbb{R}^n_{\tau,\gamma} = O(\gamma)$ as $\gamma \downarrow 0$

Statistical Mechanics deals with particle systems that are large, often infinitely large. The *Ergodic Hypothesis* roughly says that in a bounded energy hypersurface, the dynamics are ergodic, meaning that any evolution in the energy level set comes near every point of this set.

The taking of limits as the number of particles tends to infinity is a notoriously difficult subject. Here we discuss a few direct consequences of classical KAM theory for many degrees of freedom. This discussion starts with Kolmogorov's papers [80,81], which we now present in a slightly rephrased form. First, we recall that for Hamiltonian systems (say, with n degrees of freedom), typically the union of Diophantine quasi-periodic Lagrangian invariant n-tori fills up positive measure in the phase space and also in the energy hypersurfaces. Second, such a collection of KAM tori immediately gives rise to non-ergodic-

ity, since it clearly implies the existence of distinct invariant sets of positive measure. For background on Ergodic Theory, see e. g. [9,27] and [24] for more references. Apparently the KAM tori form an obstruction to ergodicity, and a question is how bad this obstruction is as $n \to \infty$. Results in [5,78] indicate that this KAM theory obstruction is not too bad as the size of the system tends to infinity. In general the role of the Ergodic Hypothesis in Statistical Mechanics has turned out to be much more subtle than was expected, see e. g. [18,64].

Dissipative KAM Theory

As already noted by Moser [105,106], KAM theory extends outside the world of Hamiltonian systems, like to volume preserving systems, or to equivariant or reversible systems. This also holds for the class of general smooth systems, often called dissipative. In fact, the KAM theorem allows for a Lie algebra proof, that can be used to cover all these special cases [24,29,31,45]. It turns out that in many cases parameters are needed for persistent occurrence of (Diophantine) quasi-periodic tori.

As an example we now consider the dissipative setting, where we discuss a parametrized system with normally hyperbolic invariant n-tori carrying quasi-periodic motion. From [73] it follows that this is a persistent situation and that, up to a smooth (in this case of class C^k for large k) diffeomorphism, we can restrict to the case where $\mathbb{T}^n$ is the phase space. To fix thoughts we consider the smooth system

$$\begin{aligned} \dot{\varphi} &= \omega(\mu) + \varepsilon f(\varphi, \mu, \varepsilon) \\ \dot{\mu} &= 0 \,, \end{aligned} \tag{8}$$

where $\mu \in \mathbb{R}^n$ is a multi-parameter. The results of the classical KAM theorem regarding $(6)_\varepsilon$ largely carry over to $(8)_{\mu,\varepsilon}$.

Now, for $\varepsilon = 0$ the product of phase space and parameter space as an open subset of $\mathbb{T}^n \times \mathbb{R}^n$ is completely foliated by invariant n-tori and since the perturbation does not concern the $\dot{\mu}$-equation, this foliation is persistent. The interest is with the dynamics on the resulting invariant tori that remains parallel after the perturbation; compare with the setting of Theorem 3. As just stated, KAM theory here gives a solution similar to the Hamiltonian case. The analogue of the Kolmogorov non-degeneracy condition here is that the frequency map $\mu \mapsto \omega(\mu)$ is a (local) diffeomorphism. Then, in the spirit of Theorem 3, we state that the system $(8)_{\mu,\varepsilon}$ is smoothly conjugated to $(8)_{\mu,0}$, as before, provided that the map ω is co-restricted to the Diophantine set $\mathbb{R}^n_{\tau,\gamma}$. Again the smoothness has to be taken in the sense of Whitney [29,119,136,143,144], also see [20,24,31,55].

It follows that the occurrence of normally hyperbolic invariant tori carrying (Diophantine) quasi-periodic flow is typical for families of systems with sufficiently many parameters, where this occurrence has positive measure in parameter space. In fact, if the number of parameters equals the dimension of the tori, the geometry as sketched in Fig. 4 carries over in a diffeomorphic way.

Remark

- Many remarks following Subsect. "Classical KAM Theory" and Theorem 3 also hold here.
- In cases where the system is degenerate, for instance because there is a lack of parameters, a path formalism can be invoked, where the parameter path is required to be a generic subfamily of the Diophantine set $\mathbb{R}^n_{\tau,\gamma}$, see Fig. 4. This amounts to the Rüssmann non-degeneracy, that still gives positive measure of quasi-periodicity in the parameter space, compare with [24,31] and references therein.
- In the dissipative case the KAM theorem gives rise to families of quasi-periodic attractors in a typical way. This is of importance in center manifold reductions of infinite dimensional dynamics as, e. g., in fluid mechanics [125,126]. In Sect. "Transition to Chaos and Turbulence" we shall return to this subject.

Lower Dimensional Tori

We extend the above approach to the case of lower dimensional tori, i. e., where the dynamics transversal to the tori is also taken into account. We largely follow the setup of [29,45] that follows Moser [106]. Also see [24,31] and references therein. Changing notation a little, we now consider the phase space $\mathbb{T}^n \times \mathbb{R}^m = \{x(\mathrm{mod}\, 2\pi), y\}$, as well a parameter space $\{\mu\} = P \subset \mathbb{R}^s$. We consider a C^∞-family of vector fields $X(x, y, \mu)$ as before, having $\mathbb{T}^n \times \{0\} \subset \mathbb{T}^n \times \mathbb{R}^m$ as an invariant n-torus for $\mu = \mu_0 \in P$.

$$\begin{aligned} \dot{x} &= \omega(\mu) + f(y, \mu) \\ \dot{y} &= \Omega(\mu)\, y + g(y, \mu) \\ \dot{\mu} &= 0 \,, \end{aligned} \tag{9}$$

with $f(y, \mu_0) = O(|y|)$ and $g(y, \mu_0) = O(|y|^2)$, so we assume the invariant torus to be of Floquet type.

The system $X = X(x, y, \mu)$ is integrable in the sense that it is $\mathbb{T}^n$-symmetric, i. e., x-independent [29]. The interest is with the fate of the invariant torus $\mathbb{T}^n \times \{0\}$ and

its parallel dynamics under small perturbation to a system $\tilde{X} = \tilde{X}(x, y, \mu)$ that no longer needs to be integrable.

Consider the smooth mappings $\omega: P \to \mathbb{R}^n$ and $\Omega: P \to \mathrm{gl}(m, \mathbb{R})$. To begin with we restrict to the case where all eigenvalues of $\Omega(\mu_0)$ are simple and nonzero. In general for such a matrix $\Omega \in \mathrm{gl}(m, \mathbb{R})$, let the eigenvalues be given by $\alpha_1 \pm \mathrm{i}\beta_1, \ldots, \alpha_{N_1} \pm \mathrm{i}\beta_{N_1}$ and $\delta_1, \ldots, \delta_{N_2}$, where all α_j, β_j and δ_j are real and hence $m = 2N_1 + N_2$. Also consider the map spec: $\mathrm{gl}(m, \mathbb{R}) \to \mathbb{R}^{2N_1+N_2}$, given by $\Omega \mapsto (\alpha, \beta, \delta)$. Next to the internal frequency vector $\omega \in \mathbb{R}^n$, we also have the vector $\beta \in \mathbb{R}^{N_1}$ of normal frequencies.

The present analogue of Kolmogorov non-degeneracy is the Broer–Huitema–Takens (BHT) non-degeneracy condition [29,127], which requires that the product map $\omega \times (\mathrm{spec}) \circ \Omega: P \to \mathbb{R}^n \times \mathrm{gl}(m, \mathbb{R})$ at $\mu = \mu_0$ has a surjective derivative and hence is a local submersion [72].

Furthermore, we need Diophantine conditions on both the internal and the normal frequencies, generalizing (7). Given $\tau > n - 1$ and $\gamma > 0$, it is required for all $k \in \mathbb{Z}^n \setminus \{0\}$ and all $\ell \in \mathbb{Z}^{N_1}$ with $|\ell| \leq 2$ that

$$|\langle k, \omega\rangle + \langle \ell, \beta\rangle| \geq \gamma |k|^{-\tau} . \tag{10}$$

Inside $\mathbb{R}^n \times \mathbb{R}^{N_1} = \{\omega, \beta\}$ this yields a Cantor set as before (compare Fig. 4). This set has to be pulled back along the submersion $\omega \times (\mathrm{spec}) \circ \Omega$, for examples see Subsects. "$(n-1)$-Tori" and "Quasi-periodic Bifurcations" below.

The KAM theorem for this setting is quasi-periodic stability of the n-tori under consideration, as in Subsect. "Dissipative KAM Theory", yielding typical examples where quasi-periodicity has positive measure in parameter space. In fact, we get a little more here, since the normal linear behavior of the n-tori is preserved by the Whitney smooth conjugations. This is expressed as normal linear stability, which is of importance for quasi-periodic bifurcations, see Subsect. "Quasi-periodic Bifurcations" below.

Remark

- A more general set-up of the normal stability theory [45] adapts the above to the case of non-simple (multiple) eigenvalues. Here the BHT non-degeneracy condition is formulated in terms of versal unfolding of the matrix $\Omega(\mu_0)$ [7]. For possible conditions under which vanishing eigenvalues are admissible see [29,42,69] and references therein.
- This general set-up allows for a structure preserving formulation as mentioned earlier, thereby including the Hamiltonian and volume preserving case, as well as equivariant and reversible cases. This allows us, for example, to deal with quasi-periodic versions of the Hamiltonian and the reversible Hopf bifurcation [38,41,42,44].
- The Parameterized KAM Theory discussed here a priori needs many parameters. In many cases the parameters are distinguished in the sense that they are given by action variables, etc. For an example see Subsect. "$(n-1)$-Tori" on Hamiltonian $(n-1)$-tori Also see [127] and [24,31] where the case of Rüssmann non-degeneracy is included. This generalizes a remark at the end of Subsect. "Dissipative KAM Theory".

Global KAM Theory

We stay in the Hamiltonian setting, considering Lagrangian invariant n-tori as these occur in a Liouville integrable system with n degrees of freedom. The union of these tori forms a smooth $\mathbb{T}^n$-bundle $f: M \to B$ (where we leave out all singular fibers). It is known that this bundle can be non-trivial [56,60] as can be measured by monodromy and Chern class. In this case global action angle variables are not defined. This non-triviality, among other things, is of importance for semi-classical versions of the classical system at hand, in particular for certain spectrum defects [57,62,134,135], for more references also see [24].

Restricting to the classical case, the problem is what happens to the (non-trivial) $\mathbb{T}^n$-bundle f under small, non-integrable perturbation. From the classical KAM theory, see Subsect. "Classical KAM Theory" we already know that on trivializing charts of f Diophantine quasi-periodic n-tori persist. In fact, at this level, a Whitney smooth conjugation exists between the integrable system and its perturbation, which is even Gevrey regular [136]. It turns out that these local KAM conjugations can be glued together so to obtain a global conjugation at the level of quasi-periodic tori, thereby implying global quasi-periodic stability [43]. Here we need unicity of KAM tori, i. e., independence of the action-angle chart used in the classical KAM theorem [26]. The proof uses the integer affine structure on the quasi-periodic tori, which enables taking convex combinations of the local conjugations subjected to a suitable partition of unity [72,129]. In this way the geometry of the integrable bundle can be carried over to the nearly-integrable one.

The classical example of a Liouville integrable system with non-trivial monodromy [56,60] is the spherical pendulum, which we now briefly revisit. The configuration space is $\mathbb{S}^2 = \{q \in \mathbb{R}^3 \mid \langle q, q\rangle = 1\}$ and the phase space $T^*\mathbb{S}^2 \cong \{(q, p) \in \mathbb{R}^6 \mid \langle q, q\rangle = 1$ and $\langle q, p\rangle = 0\}$. The two integrals $I = q_1 p_2 - q_2 p_1$ (angular momentum) and $E = \frac{1}{2}\langle p, p\rangle + q_3$ (energy) lead to

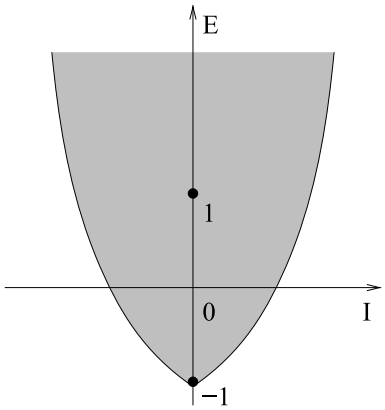

Hamiltonian Perturbation Theory (and Transition to Chaos), Figure 5
Range of the energy-momentum map of the spherical pendulum

an energy momentum map $\mathcal{EM}\colon T^*\mathbb{S}^2 \to \mathbb{R}^2$, given by $(q,p) \mapsto (I,E) = \left(q_1p_2 - q_2p_1, \frac{1}{2}\langle p,p\rangle + q_3\right)$. In Fig. 5 we show the image of the map $\mathcal{EM}$. The shaded area B consists of regular values, the fiber above which is a Lagrangian two-torus; the union of these gives rise to a bundle $f\colon M \to B$ as described before, where $f = \mathcal{EM}|_M$. The motion in the two-tori is a superposition of Huygens' rotations and pendulum-like swinging, and the non-existence of global action angle variables reflects that the three interpretations of 'rotating oscillation', 'oscillating rotation' and 'rotating rotation' cannot be reconciled in a consistent way. The singularities of the fibration include the equilibria $(q,p) = ((0,0,\pm 1),(0,0,0)) \mapsto (I,E) = (0,\pm 1)$. The boundary of this image also consists of singular points, where the fiber is a circle that corresponds to Huygens' horizontal rotations of the pendulum. The fiber above the upper equilibrium point $(I,E) = (0,1)$ is a pinched torus [56], leading to non-trivial monodromy, in a suitable bases of the period lattices, given by

$$\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \in \mathrm{GL}(2,\mathbb{R})\,.$$

The question here is what remains of the bundle f when the system is perturbed. Here we observe that locally Kolmogorov non-degeneracy is implied by the non-trivial monodromy [114,122]. From [43,122] it follows that the non-trivial monodromy can be extended in the perturbed case.

Remark

- The case where this perturbation remains integrable is covered in [95], but presently the interest is with the nearly integrable case, so where the axial symmetry is broken. Also compare [24] and many of its references.
- The global conjugations of [43] are Whitney smooth (even Gevrey regular [136]) and near the identity map in the C^∞-topology [72]. Geometrically speaking these diffeomorphisms also are $\mathbb{T}^n$-bundle isomorphisms between the unperturbed and the perturbed bundle, the basis of which is a Cantor set of positive measure.

Splitting of Separatrices

KAM theory does not predict the fate of close-to-resonant tori under perturbations. For fully resonant tori the phenomenon of frequency locking leads to the destruction of the torus under (sufficiently rich) perturbations, and other resonant tori disintegrate as well. In the case of a single resonance between otherwise Diophantine frequencies the perturbation leads to quasi-periodic bifurcations, cf. Sect. "Transition to Chaos and Turbulence".

While KAM theory concerns the fate of most trajectories and for all times, a complementary theorem has been obtained in [93,109,110,113]. It concerns all trajectories and states that they stay close to the unperturbed tori for *long* times that are exponential in the inverse of the perturbation strength. For trajectories starting close to surviving tori the diffusion is even superexponentially slow, cf. [102,103]. Here a form of smoothness exceeding the mere existence of infinitely many derivatives of the Hamiltonian is a necessary ingredient, for finitely differentiable Hamiltonians one only obtains polynomial times.

Solenoids, which cannot be present in integrable systems, are constructed for generic Hamiltonian systems in [16,94,98], yielding the simultaneous existence of representatives of all homeomorphy-classes of solenoids. Hyperbolic tori form the core of a construction proposed in [5] of trajectories that venture off to distant points of the phase space. In the unperturbed system the union of a family of hyperbolic tori, parametrized by the actions conjugate to the toral angles, form a normally hyperbolic manifold. The latter is persistent under perturbations, cf. [73,100], and carries a Hamiltonian flow with fewer degrees of freedom. The main difference between integrable and non-integrable systems already occurs for periodic orbits.

Periodic Orbits

A sharp difference to dissipative systems is that it is generic for hyperbolic periodic orbits on compact energy shells in Hamiltonian systems to have homoclinic orbits, cf. [1] and references therein. For integrable systems these form to-

gether a pinched torus, but under generic perturbations the stable and unstable manifold of a hyperbolic periodic orbit intersect transversely. It is a nontrivial task to actually check this genericity condition for a given non-integrable perturbation, a first-order condition going back to Poincaré requires the computation of the so-called Mel'nikov integral, see [66,137] for more details. In two degrees of freedom normalization leads to approximations that are integrable to all orders, which implies that the Melnikov integral is a flat function. In the real analytic case the Melnikov criterion is still decisive in many examples [65].

Genericity conditions are traditionally formulated in the universe of smooth vector fields, and this makes the whole class of analytic vector fields appear to be non-generic. This is an overly pessimistic view as the conditions defining a certain class of generic vector fields may certainly be satisfied by a given analytic system. In this respect it is interesting that the generic properties may also be formulated in the universe of analytic vector fields, see [28] for more details.

$(n-1)$-Tori

The $(n-1)$-parameter families of invariant $(n-1)$-tori organize the dynamics of an integrable Hamiltonian system in n degrees of freedom, and under small perturbations the parameter space of persisting analytic tori is Cantorized. This still allows for a global understanding of a substantial part of the dynamics, but also leads to additional questions.

A hyperbolic invariant torus $\mathbb{T}^{n-1}$ has its Floquet exponents off the imaginary axis. Note that $\mathbb{T}^{n-1}$ is not a normally hyperbolic manifold. Indeed, the normal linear behavior involves the $n-1$ zero eigenvalues in the direction of the parametrizing actions as well; similar to (9) the format

$$\begin{aligned}\dot{x} &= \omega(y) + O(y) + O(z^2)\\ \dot{y} &= O(y) + O(z^3)\\ \dot{z} &= \Omega(y)z + O(z^2)\end{aligned}$$

in Floquet coordinates yields an x-independent matrix Ω that describes the symplectic normal linear behavior, cf. [29]. The union $\{z=0\}$ over the family of $(n-1)$-tori is a normally hyperbolic manifold and constitutes the center manifold of $\mathbb{T}^{n-1}$. Separatrices splitting yields the dividing surfaces in the sense of Wiggins et al. [138].

The persistence of elliptic tori under perturbation from an integrable system involves not only the internal frequencies of $\mathbb{T}^{n-1}$, but also the normal frequencies. Next to the internal resonances the necessary Diophantine conditions (10) exclude the normal-internal resonances

$$\langle k, \omega\rangle = \alpha_j \tag{11}$$

$$\langle k, \omega\rangle = 2\alpha_j \tag{12}$$

$$\langle k, \omega\rangle = \alpha_i + \alpha_j \tag{13}$$

$$\langle k, \omega\rangle = \alpha_i - \alpha_j\,. \tag{14}$$

The first three resonances lead to the quasi-periodic center-saddle bifurcation studied in Sect. "Transition to Chaos and Turbulence", the frequency-halving (or quasi-periodic period doubling) bifurcation and the quasi-periodic Hamiltonian Hopf bifurcation, respectively. The resonance (14) generalizes an equilibrium in 1 : 1 resonance whence $\mathbb{T}^{n-1}$ persists and remains elliptic, cf. [78]. When passing through resonances (12) and (13) the lower-dimensional tori lose ellipticity and acquire hyperbolic Floquet exponents. Elliptic $(n-1)$-tori have a single normal frequency whence (11) and (12) are the only normal-internal resonances. See [35] for a thorough treatment of the ensuing possibilities.

The restriction to a single normal-internal resonance is dictated by our present possibilities. Indeed, already the bifurcation of equilibria with a fourfold zero eigenvalue leads to unfoldings that simultaneously contain all possible normal resonances. Thus, a satisfactory study of such tori which already may form one-parameter families in integrable Hamiltonian systems with five degrees of freedom has to await further progress in local bifurcation theory.

Transition to Chaos and Turbulence

One of the main interests over the second half of the twentieth century has been the transition between orderly and complicated forms of dynamics upon variation of either initial states or of system parameters. By 'orderly' we here mean equilibrium and periodic dynamics and by complicated quasi-periodic and chaotic dynamics, although we note that only chaotic dynamics is associated to unpredictability, e. g. see [27]. As already discussed in the introduction systems like a forced nonlinear oscillator or the planar three-body problem exhibit coexistence of periodic, quasi-periodic and chaotic dynamics, also compare with Fig. 1.

Similar remarks go for the onset of turbulence in fluid dynamics. Around 1950 this led to the scenario of Hopf–Landau–Lifschitz [75,76,83,84], which roughly amounts to the following. Stationary fluid motion corresponds to an equilibrium point in an ∞-dimensional state space

of velocity fields. The first transition is a Hopf bifurcation [66,75,82], where a periodic solution branches off. In a second transition of similar nature a quasi-periodic two-torus branches off, then a quasi-periodic three-torus, etc. The idea is that the motion picks up more and more frequencies and thus obtains an increasingly complicated power spectrum. In the early 1970s this idea was modified in the Ruelle–Takens route to turbulence, based on the observation that, for flows, a three-torus can carry chaotic (or strange) attractors [112,126], giving rise to a broad band power spectrum. By the quasi-periodic bifurcation theory [24,29,31] as sketched below these two approaches are unified in a generic way, keeping track of measure theoretic aspects. For general background in dynamical systems theory we refer to [27,79].

Another transition to chaos was detected in the quadratic family of interval maps

$$f_\mu(x) = \mu x(1-x)\,,$$

see [58,99,101], also for a holomorphic version. This transition consists of an infinite sequence of period doubling bifurcations ending up in chaos; it has several universal aspects and occurs persistently in families of dynamical systems. In many of these cases also homoclinic bifurcations show up, where sometimes the transition to chaos is immediate when parameters cross a certain boundary, for general theory see [13,14,30,117]. There exist quite a number of case studies where all three of the above scenarios play a role, e. g., see [32,33,46] and many of their references.

Quasi-periodic Bifurcations

For the classical bifurcations of equilibria and periodic orbits, the bifurcation sets and diagrams are generally determined by a classical geometry in the product of phase space and parameter space as already established by, e. g., [8,133], often using singularity theory. Quasi-periodic bifurcation theory concerns the extension of these bifurcations to invariant tori in nearly-integrable systems, e. g., when the tori lose their normal hyperbolicity or when certain (strong) resonances occur. In that case the dense set of resonances, also responsible for the small divisors, leads to a Cantorization of the classical geometries obtained from Singularity Theory [29,35,37,38,39,41,44,45,48,49,67,68,69], also see [24,31,52,55]. Broadly speaking, one could say that in these cases the Preparation Theorem [133] is partly replaced by KAM theory. Since the KAM theory has been developed in several settings with or without preservation of structure, see Sect. "KAM Theory: An Overview", for the ensuing quasi-periodic bifurcation theory the same holds.

Hamiltonian Cases To fix thoughts we start with an example in the Hamiltonian setting, where a robust model for the quasi-periodic center-saddle bifurcation is given by

$$\begin{aligned} &H_{\omega_1,\omega_2,\mu,\varepsilon}(I,\varphi,p,q) \\ &\quad = \omega_1 I_1 + \omega_2 I_2 + \frac{1}{2}p^2 + V_\mu(q) + \varepsilon f(I,\varphi,p,q) \end{aligned} \tag{15}$$

with $V_\mu(q) = \frac{1}{3}q^3 - \mu q$, compare with [67,69]. The unperturbed (or integrable) case $\varepsilon = 0$, by factoring out the $\mathbb{T}^2$-symmetry, boils down to a standard center-saddle bifurcation, involving the fold catastrophe [133] in the potential function $V = V_\mu(q)$. This results in the existence of two invariant two-tori, one elliptic and the other hyperbolic. For $0 \neq |\varepsilon| \ll 1$ the dense set of resonances complicates this scenario, as sketched in Fig. 6, determined by the Diophantine conditions

$$\begin{aligned} |\langle k,\omega\rangle| &\geq \gamma |k|^{-\tau}\,, \quad \text{for } q<0\,, \\ |\langle k,\omega\rangle + \ell\beta(q)| &\geq \gamma |k|^{-\tau}\,, \quad \text{for } q>0 \end{aligned} \tag{16}$$

for all $k \in \mathbb{Z}^n \setminus \{0\}$ and for all $\ell \in \mathbb{Z}$ with $|\ell| \leq 2$. Here $\beta(q) = \sqrt{2q}$ is the normal frequency of the elliptic torus given by $q = \sqrt{\mu}$ for $\mu > 0$. As before, (cf. Sects. "Invariant Curves of Planar Diffeomorphisms", "KAM Theory: An Overview"), this gives a Cantor set of positive measure [24,29,31,45,69,105,106].

For $0 < |\varepsilon| \ll 1$ Fig. 6 will be distorted by a near-identity diffeomorphism; compare with the formulations of the Theorems 3 and 4. On the Diophantine Cantor set the dynamics is quasi-periodic, while in the gaps generically there is coexistence of periodicity and chaos, roughly comparable with Fig. 1, at left. The gaps at the border furthermore lead to the phenomenon of parabolic resonance, cf. [86].

Similar programs exist for all cuspoid and umbilic catastrophes [37,39,68] as well as for the Hamiltonian Hopf bifurcation [38,44]. For applications of this approach see [35]. For a reversible analogue see [41]. As so often within the gaps generically there is an infinite regress of smaller gaps [11,35]. For theoretical background we refer to [29,45,106], for more references also see [24].

Dissipative Cases In the general dissipative case we basically follow the same strategy. Given the standard bifurcations of equilibria and periodic orbits, we get more complex situations when invariant tori are involved as well.

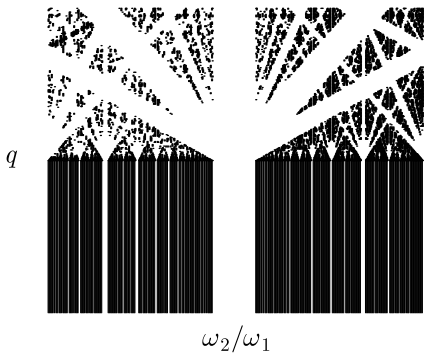

Hamiltonian Perturbation Theory (and Transition to Chaos), Figure 6
Sketch of the Cantorized Fold, as the bifurcation set of the quasi-periodic center-saddle bifurcation for $n = 2$ [67], where the horizontal axis indicates the frequency ratio $\omega_2 : \omega_1$, cf. (15). The lower part of the figure corresponds to hyperbolic tori and the upper part to elliptic ones. See the text for further interpretations

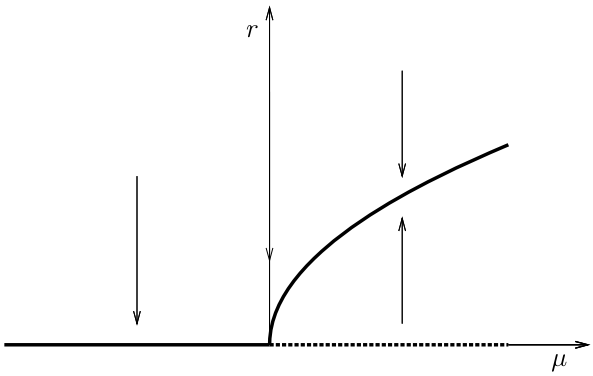

Hamiltonian Perturbation Theory (and Transition to Chaos), Figure 7
Bifurcation diagram of the Hopf bifurcation

The simplest examples are the quasi-periodic saddle-node and quasi-periodic period doubling [29] also see [24,31].

To illustrate the whole approach let us start from the Hopf bifurcation of an equilibrium point of a vector field [66,75,82,116] where a hyperbolic point attractor loses stability and branches off a periodic solution, cf. Subsect. "Dissipative Perturbations". A topological normal form is given by

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \end{pmatrix} = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} - (y_1^2 + y_2^2) \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \tag{17}$$

where $y = (y_1, y_2) \in \mathbb{R}^2$, ranging near $(0, 0)$. In this representation usually one fixes $\beta = 1$ and lets $\alpha = \mu$ (near 0) serve as a (bifurcation) parameter, classifying modulo topological equivalence. In polar coordinates (17) so gets the form

$$\dot{\varphi} = 1 \,,$$
$$\dot{r} = \mu r - r^3 \,.$$

Figure 7 shows an amplitude response diagram (often called the bifurcation diagram). Observe the occurrence of the attracting periodic solution for $\mu > 0$ of amplitude $\sqrt{\mu}$.

Let us briefly consider the Hopf bifurcation for fixed points of diffeomorphisms. A simple example has the form

$$P(y) = \mathrm{e}^{2\pi(\alpha + i\beta)} y + O(|y|^2) \,, \tag{18}$$

$y \in \mathbb{C} \cong \mathbb{R}^2$, near 0. To start with β is considered a constant, such that β is not rational with denominator less than five, see [7,132], and where $O(|y|^2)$ should contain generic third order terms. As before, we let $\alpha = \mu$ serve as a bifurcation parameter, varying near 0. On one side of the bifurcation value $\mu = 0$, this system has by normal hyperbolicity and [73], an invariant circle. Here, due to the invariance of the rotation numbers of the invariant circles, no topological stability can be obtained [111]. Still this bifurcation can be characterized by many persistent properties. Indeed, in a generic two-parameter family (18), say with both α and β as parameters, the periodicity in the parameter plane is organized in resonance tongues [7,34,82]. (The tongue structure is hardly visible when only one parameter, like α, is used.) If the diffeomorphism is the return map of a periodic orbit for flows, this bifurcation produces an invariant two-torus. Usually this counterpart for flows is called Neĭmark–Sacker bifurcation. The periodicity as it occurs in the resonance tongues, for the vector field is related to phase lock. The tongues are contained in gaps of a Cantor set of quasi-periodic tori with Diophantine frequencies. Compare the discussion in Subsect. "Circle Maps", in particular also regarding the Arnold family and Fig. 3. Also see Sect. "KAM Theory: An Overview" and again compare with [115].

Quasi-periodic versions exist for the saddle-node, the period doubling and the Hopf bifurcation. Returning to the setting with $\mathbb{T}^n \times \mathbb{R}^m$ as the phase space, we remark that the quasi-periodic saddle-node and period doubling already occur for $m = 1$, or in an analogous center manifold. The quasi-periodic Hopf bifurcation needs $m \geq 2$. We shall illustrate our results on the latter of these cases, compare with [19,31]. For earlier results in this direction see [52]. Our phase space is $\mathbb{T}^n \times \mathbb{R}^2 = \{x (\mathrm{mod}\, 2\pi), y\}$, where we are dealing with the parallel invariant torus $\mathbb{T}^n \times \{0\}$. In the integrable case, by $\mathbb{T}^n$-symmetry we can

reduce to $\mathbb{R}^2 = \{y\}$ and consider the bifurcations of relative equilibria. The present interest is with small non-integrable perturbations of such integrable models.

We now discuss the *quasi-periodic Hopf bifurcation* [17,29], largely following [55]. The unperturbed, integrable family $X = X_\mu(x, y)$ on $\mathbb{T}^n \times \mathbb{R}^2$ has the form

$$X_\mu(x, y) = [\omega(\mu) + f(y, \mu)]\partial_x + [\Omega(\mu)y + g(y, \mu)]\partial_y \,, \tag{19}$$

were $f = O(|y|)$ and $g = O(|y|^2)$ as before. Moreover $\mu \in P$ is a multi-parameter and $\omega\colon P \to \mathbb{R}^n$ and $\Omega\colon P \to \mathrm{gl}(2, \mathbb{R})$ are smooth maps. Here we take

$$\Omega(\mu) = \begin{pmatrix} \alpha(\mu) & -\beta(\mu) \\ \beta(\mu) & \alpha(\mu) \end{pmatrix} ,$$

which makes the ∂_y component of (19) compatible with the planar Hopf family (17). The present form of Kolmogorov non-degeneracy is Broer–Huitema–Takens stability [29,42,45], requiring that there is a subset $\Gamma \subseteq P$ on which the map

$$\mu \in P \mapsto (\omega(\mu), \Omega(\mu)) \in \mathbb{R}^n \times \mathrm{gl}(2, \mathbb{R})$$

is a submersion. For simplicity we even assume that μ is replaced by

$$(\omega, (\alpha, \beta)) \in \mathbb{R}^n \times \mathbb{R}^2 \,.$$

Observe that if the non-linearity g satisfies the well-known Hopf non-degeneracy conditions, e. g., compare [66,82], then the relative equilibrium $y = 0$ undergoes a standard planar Hopf bifurcation as described before. Here α again plays the role of bifurcation parameter and a closed orbit branches off at $\alpha = 0$. To fix thoughts we assume that $y = 0$ is attracting for $\alpha < 0$. and that the closed orbit occurs for $\alpha > 0$, and is attracting as well. For the integrable family X, qualitatively we have to multiply this planar scenario with $\mathbb{T}^n$, by which all equilibria turn into invariant attracting or repelling n-tori and the periodic attractor into an attracting invariant $(n + 1)$-torus. Presently the question is what happens to both the n- and the $(n + 1)$-tori, when we apply a small near-integrable perturbation.

The story runs much like before. Apart from the BHT non-degeneracy condition we require Diophantine conditions (10), defining the Cantor set

$$\Gamma^{(2)}_{\tau,\gamma} = \{(\omega, (\alpha, \beta)) \in \Gamma \mid |\langle k, \omega\rangle + \ell\beta| \geq \gamma|k|^{-\tau} \,, \quad \forall k \in \mathbb{Z}^n \setminus \{0\} \,, \forall \ell \in \mathbb{Z} \text{ with } |\ell| \leq 2\} \,, \tag{20}$$

In Fig. 8 we sketch the intersection of $\Gamma^{(2)}_{\tau,\gamma} \subset \mathbb{R}^n \times \mathbb{R}^2$

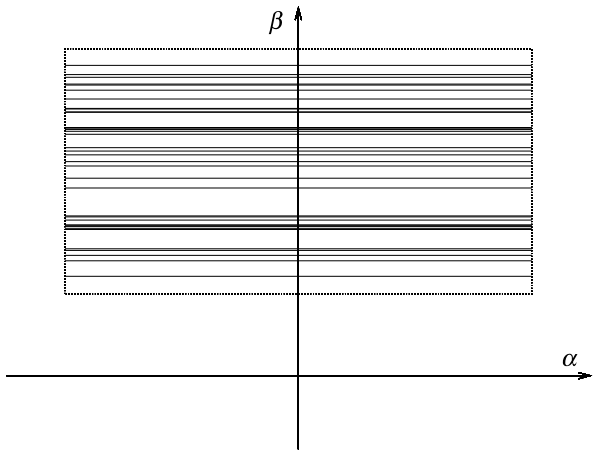

Hamiltonian Perturbation Theory (and Transition to Chaos), Figure 8
Planar section of the Cantor set $\Gamma^{(2)}_{\tau,\gamma}$

with a plane $\{\omega\} \times \mathbb{R}^2$ for a Diophantine (internal) frequency vector ω, cf. (7).

From [17,29] it now follows that for any family $\tilde{X}$ on $\mathbb{T}^n \times \mathbb{R}^2 \times P$, sufficiently near X in the C^∞-topology a near-identity C^∞-diffeomorphism $\Phi\colon \mathbb{T}^n \times \mathbb{R}^2 \times \Gamma \to \mathbb{T}^n \times \mathbb{R}^2 \times \Gamma$ exists, defined near $\mathbb{T}^n \times \{0\} \times \Gamma$, that conjugates X to $\tilde{X}$ when further restricting to $\mathbb{T}^n \times \{0\} \times \Gamma^{(2)}_{\tau,\gamma}$. So this means that the Diophantine quasi-periodic invariant n-tori are persistent on a diffeomorphic image of the Cantor set $\Gamma^{(2)}_{\tau,\gamma}$, compare with the formulations of the Theorems 3 and 4.

Similarly we can find invariant $(n + 1)$-tori. We first have to develop a $\mathbb{T}^{n+1}$ symmetric normal form approximation [17,29] and ► Normal Forms in Perturbation Theory. For this purpose we extend the Diophantine conditions (20) by requiring that the inequality holds for all $|\ell| \leq N$ for $N = 7$. We thus find another large Cantor set, again see Fig. 8, where Diophantine quasi-periodic invariant $(n + 1)$-tori are persistent. Here we have to restrict to $\alpha > 0$ for our choice of the sign of the normal form coefficient, compare with Fig. 7.

In both the cases of n-tori and of $(n + 1)$-tori, the nowhere dense subset of the parameter space containing the tori can be fattened by normal hyperbolicity to open subsets. Indeed, the quasi-periodic n- and $(n + 1)$-tori are infinitely normally hyperbolic [73]. Exploiting the normal form theory [17,29] and ► Normal Forms in Perturbation Theory to the utmost and using a more or less standard contraction argument [17,53], a fattening of the parameter domain with invariant tori can be obtained that leaves out only small 'bubbles' around the resonances, as sketched

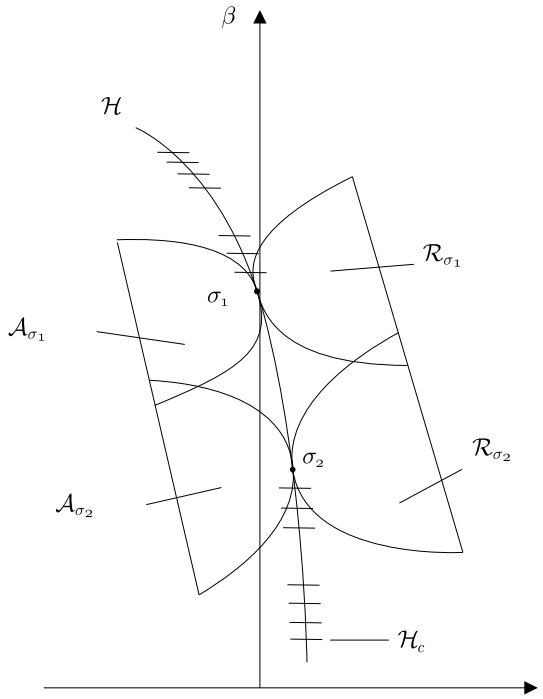

Hamiltonian Perturbation Theory (and Transition to Chaos), Figure 9

Fattening by normal hyperbolicity of a nowhere dense parameter set with invariant *n*-tori in the perturbed system. The curve $\mathcal{H}$ is the Whitney smooth (even Gevrey regular [136]) image of the β-axis in Fig. 8. $\mathcal{H}$ interpolates the Cantor set $\mathcal{H}_c$ that contains the non-hyperbolic Diophantine quasi-periodic invariant *n*-tori, corresponding to $\Gamma^{(2)}_{\tau,\gamma}$, see (20). To the points $\sigma_{1,2} \in \mathcal{H}_c$ discs $\mathcal{A}_{\sigma_{1,2}}$ are attached where we find attracting normally hyperbolic *n*-tori and similarly in the discs $\mathcal{R}_{\sigma_{1,2}}$ repelling ones. The contact between the disc boundaries and $\mathcal{H}$ is infinitely flat [17,29]

and explained in Fig. 9 for the n-tori. For earlier results in the same spirit in a case study of the quasi-periodic saddle-node bifurcation see [49,50,51], also compare with [11].

A Scenario for the Onset of Turbulence

Generally speaking, in many settings quasi-periodicity constitutes the order in between chaos [31]. In the Hopf–Landau–Lifschitz–Ruelle–Takens scenario [76,83,84,126] we may consider a sequence of typical transitions as given by quasi-periodic Hopf bifurcations, starting with the standard Hopf or Hopf–Neĭmark–Sacker bifurcation as described before. In the gaps of the Diophantine Cantor sets generically there will be coexistence of periodicity, quasi-periodicity and chaos in infinite regress. As said earlier, period doubling sequences and homoclinic bifurcations may accompany this.

As an example consider a family of maps that undergoes a generic quasi-periodic Hopf bifurcation from circle to two-torus. It turns out that here the Cantorized fold of Fig. 6 is relevant, where now the vertical coordinate is a bifurcation parameter. Moreover compare with Fig. 3, where also variation of ε is taken into account. The Cantor set contains the quasi-periodic dynamics, while in the gaps we can have chaos, e. g., in the form of Hénon like strange attractors [46,112]. A fattening process as explained above, also can be carried out here.

Future Directions

One important general issue is the mathematical characterization of chaos and ergodicity in dynamical systems, in conservative, dissipative and in other settings. This is a tough problem as can already be seen when considering two-dimensional diffeomorphisms. In particular we refer to the still unproven ergodicity conjecture of [9] and to the conjectures around Hénon like attractors and the principle 'Hénon everywhere', compare with [22,32]. For a discussion see Subsect. "A Scenario for the Onset of Turbulence". In higher dimension this problem is even harder to handle, e. g., compare with [46,47] and references therein. In the conservative case a related problem concerns a better understanding of Arnold diffusion.

Somewhat related to this is the analysis of dynamical systems without an explicit perturbation setting. Here numerical and symbolic tools are expected to become useful to develop computer assisted proofs in extended perturbation settings, diagrams of Lyapunov exponents, symbolic dynamics, etc. Compare with [128]. Also see [46,47] for applications and further reference. This part of the theory is important for understanding concrete models, that often are not given in perturbation format.

Regarding nearly-integrable Hamiltonian systems, several problems have to be considered. Continuing the above line of thought, one interest is the development of Hamiltonian bifurcation theory without integrable normal form and, likewise, of KAM theory without action angle coordinates [87]. One big related issue also is to develop KAM theory outside the perturbation format.

The previous section addressed persistence of Diophantine tori involved in a bifurcation. Similar to Cremer's example in Subsect. "Cremer's Example in Herman's Version" the dynamics in the gaps between persistent tori displays new phenomena. A first step has been made in [86] where internally resonant parabolic tori involved in a quasi-periodic Hamiltonian pitchfork bifurcation are considered. The resulting large dynamical instabilities may be further amplified for tangent (or flat)

parabolic resonances, which fail to satisfy the iso-energetic non-degeneracy condition.

The construction of solenoids in [16,94] uses elliptic periodic orbits as starting points, the simplest example being the result of a period-doubling sequence. This construction should carry over to elliptic tori, where normal-internal resonances lead to encircling tori of the same dimension, while internal resonances lead to elliptic tori of smaller dimension and excitation of normal modes increases the torus dimension. In this way one might be able to construct solenoid-type invariant sets that are limits of tori with varying dimension.

Concerning the global theory of nearly-integrable torus bundles [43], it is of interest to understand the effects of quasi-periodic bifurcations on the geometry and its invariants. Also it is of interest to extend the results of [134] when passing to semi-classical approximations. In that case two small parameters play a role, namely Planck's constant as well as the distance away from integrability.

Bibliography

1. Abraham R, Marsden JE (1978) Foundations of Mechanics, 2nd edn. Benjamin
2. Arnold VI (1961) Small divisors I: On mappings of the circle onto itself. Izv Akad Nauk SSSR Ser Mat 25:21–86 (in Russian); English translation: Am Math Soc Transl Ser 2(46):213–284 (1965); Erratum: Izv Akad Nauk SSSR Ser Mat 28:479–480 (1964, in Russian)
3. Arnold VI (1962) On the classical perturbation theory and the stability problem of the planetary system. Dokl Akad Nauk SSSR 145:487–490
4. Arnold VI (1963) Proof of a theorem by A.N. Kolmogorov on the persistence of conditionally periodic motions under a small change of the Hamilton function. Russ Math Surv 18(5):9–36 (English; Russian original)
5. Arnold VI (1964) Instability of dynamical systems with several degrees of freedom. Sov Math Dokl 5:581–585
6. Arnold VI (1978) Mathematical Methods of Classical Mechanics, GTM 60. Springer, New York
7. Arnold VI (1983) Geometrical Methods in the Theory of Ordinary Differential Equations. Springer
8. Arnold VI (ed) (1994) Dynamical Systems V: Bifurcation Theory and Catastrophe Theory. Encyclopedia of Mathematical Sciences, vol 5. Springer
9. Arnold VI, Avez A (1967) Problèmes Ergodiques de la Mécanique classique, Gauthier-Villars; English edition: Arnold VI, Avez A (1968) Ergodic problems of classical mechanics. Benjamin
10. Arnol'd VI, Kozlov VV, Neishtadt AI (1988) Mathematical Aspects of Classical and Celestial Mechanics. In: Arnold VI (ed) Dynamical Systems, vol III. Springer
11. Baesens C, Guckenheimer J, Kim S, MacKay RS (1991) Three coupled oscillators: Mode-locking, global bifurcation and toroidal chaos. Phys D 49(3):387–475
12. Barrow-Green J (1997) Poincaré and the Three Body Problem. In: History of Mathematics, vol 11. Am Math Soc, Providence; London Math Soc, London
13. Benedicks M, Carleson L (1985) On iterations of $1 - ax^2$ on $(-1, 1)$. Ann Math 122:1–25
14. Benedicks M, Carleson L (1991) The dynamics of the Hénon map. Ann Math 133:73–169
15. Benettin G (2005) Physical applications of Nekhoroshev theorem and exponential estimates. In: Giorgilli A (ed) Hamiltonian dynamics theory and applications, Cetraro 1999, Lecture Notes in Mathematics, vol 1861. Springer, pp 1–76
16. Birkhoff BD (1935) Nouvelles recherches sur les systemes dynamiques. Mem Pont Acad Sci Novi Lyncaei 1(3):85–216
17. Braaksma BLJ, Broer HW (1987) On a quasi-periodic Hopf bifurcation. Ann Inst Henri Poincaré, Anal non linéaire 4(2): 115–168
18. Bricmont J (1996) Science of chaos or chaos in science? In: Gross PR, Levitt N, Lewis MW (eds) The Flight from Science and Reason (New York, 1995), Ann New York Academy of Sciences, vol 775. New York Academy of Sciences, New York, pp 131–175; Also appeared in: Phys Mag 17:159–208 (1995)
19. Broer HW (2003) Coupled Hopf-bifurcations: Persistent examples of n-quasiperiodicity determined by families of 3-jets. Astérisque 286:223–229
20. Broer HW (2004) KAM theory: the legacy of Kolmogorov's 1954 paper. Bull Am Math Soc (New Series) 41(4):507–521
21. Broer HW, Huitema GB (1991) A proof of the isoenergetic KAM-theorem from the "ordinary" one. J Differ Equ 90:52–60
22. Broer HW, Krauskopf B (2000) Chaos in periodically driven systems. In Krauskopf B, Lenstra D (eds) Fundamental Issues of Nonlinear Laser Dynamics. American Institute of Physics Conference Proceedings 548:31–53
23. Broer HW, Roussarie R (2001) Exponential confinement of chaos in the bifurcation set of real analytic diffeomorphisms. In: Broer HW, Krauskopf B, Vegter G (eds) Global Analysis of Dynamical Systems, Festschrift dedicated to Floris Takens for his 60th birthday. Bristol and Philadelphia IOP, pp 167–210
24. Broer HW, Sevryuk MB (2007) KAM Theory: quasi-periodicity in dynamical systems. In: Broer HW, Hasselblatt B, Takens F (eds) Handbook of Dynamical Systems, vol 3. North-Holland (to appear)
25. Broer HW, Takens F (1989) Formally symmetric normal forms and genericity. Dyn Rep 2:36–60
26. Broer HW, Takens F (2007) Unicity of KAM tori. Ergod Theory Dyn Syst 27:713–724
27. Broer HW, Takens F (2008) Dynamical Systems and Chaos. To be published by Epsilon Uitgaven
28. Broer HW, Tangerman FM (1986) From a differentiable to a real analytic perturbation theory, applications to the Kupka Smale theorems. Ergod Theory Dyn Syst 6:345–362
29. Broer HW, Huitema GB, Takens F, Braaksma BLJ (1990) Unfoldings and bifurcations of quasi-periodic tori. In: Memoir AMS, vol 421. Amer Math Soc, Providence
30. Broer HW, Dumortier F, van Strien SJ, Takens F (1991) Structures in dynamics, finite dimensional deterministic studies. In: de Jager EM, van Groesen EWC (eds) Studies in Mathematical Physics, vol II. North-Holland
31. Broer HW, Huitema GB, Sevryuk MB (1996) Quasi-Periodic Motions in Families of Dynamical Systems: Order amidst Chaos. In: Lecture Notes in Mathematics, vol 1645. Springer
32. Broer HW, Simó C, Tatjer JC (1998) Towards global models

near homoclinic tangencies of dissipative diffeomorphisms. Nonlinearity 11(3):667–770
33. Broer HW, Simó C, Vitolo R (2002) Bifurcations and strange attractors in the Lorenz-84 climate model with seasonal forcing. Nonlinearity 15(4):1205–1267
34. Broer HW, Golubitsky M, Vegter G (2003) The geometry of resonance tongues: a singularity theory approach. Nonlinearity 16:1511–1538
35. Broer HW, Hanßmann H, Jorba À, Villanueva J, Wagener FOO (2003) Normal-internal resonances in quasi-periodically forced oscillators: a conservative approach. Nonlinearity 16:1751–1791
36. Broer HW, Hoveijn I, Lunter G, Vegter G (2003) Bifurcations in Hamiltonian systems: Computing Singularities by Gröbner Bases. In: Lecture Notes in Mathematics, vol 1806. Springer
37. Broer HW, Hanßmann H, You J (2005) Bifurcations of normally parabolic tori in Hamiltonian systems. Nonlinearity 18:1735–1769
38. Broer HW, Hanßmann H, Hoo J, Naudot V (2006) Nearly-integrable perturbations of the Lagrange top: applications of KAM theory. In: Denteneer D, den Hollander F, Verbitskiy E (eds) Dynamics & Stochastics: Festschrift in Honor of MS Keane Lecture Notes, vol 48. Inst. of Math. Statistics, pp 286–303
39. Broer HW, Hanßmann H, You J (2006) Umbilical torus bifurcations in Hamiltonian systems. J Differ Equ 222:233–262
40. Broer HW, Naudot V, Roussarie R (2006) Catastrophe theory in Dulac unfoldings. Ergod Theory Dyn Syst 26:1–35
41. Broer HW, Ciocci MC, Hanßmann H (2007) The quasi-periodic reversible Hopf bifurcation. In: Doedel E, Krauskopf B, Sanders J (eds) Recent Advances in Nonlinear Dynamics: Theme section dedicated to André Vanderbauwhede. Intern J Bifurc Chaos 17:2605–2623
42. Broer HW, Ciocci MC, Hanßmann H, Vanderbauwhede A (2009) Quasi-periodic stability of normally resonant tori. Phys D 238:309–318
43. Broer HW, Cushman RH, Fassò F, Takens F (2007) Geometry of KAM tori for nearly integrable Hamiltonian systems. Ergod Theory Dyn Syst 27(3):725–741
44. Broer HW, Hanßmann H, Hoo J (2007) The quasi-periodic Hamiltonian Hopf bifurcation. Nonlinearity 20:417–460
45. Broer HW, Hoo J, Naudot V (2007) Normal linear stability of quasi-periodic tori. J Differ Equ 232:355–418
46. Broer HW, Simó C, Vitolo R (2008) The Hopf–Saddle-Node bifurcation for fixed points of 3D-diffeomorphisms, the Arnol'd resonance web. Bull Belg Math Soc Simon Stevin 15:769–787
47. Broer HW, Simó C, Vitolo R (2008) The Hopf–Saddle-Node bifurcation for fixed points of 3D-diffeomorphisms, analysis of a resonance 'bubble'. Phys D Nonlinear Phenom (to appear)
48. Broer HW, Hanßmann H, You J (in preparation) On the destruction of resonant Lagrangean tori in Hamiltonian systems
49. Chenciner A (1985) Bifurcations de points fixes elliptiques I, courbes invariantes. Publ Math IHÉS 61:67–127
50. Chenciner A (1985) Bifurcations de points fixes elliptiques II, orbites périodiques et ensembles de Cantor invariants. Invent Math 80:81–106
51. Chenciner A (1988) Bifurcations de points fixes elliptiques III, orbites périodiques de "petites" périodes et élimination résonnante des couples de courbes invariantes. Publ Math IHÉS 66:5–91
52. Chenciner A, Iooss G (1979) Bifurcations de tores invariants. Arch Ration Mech Anal 69(2):109–198; 71(4):301–306
53. Chow S-N, Hale JK (1982) Methods of Bifurcation Theory. Springer
54. Chow S-N, Li C, Wang D (1994) Normal Forms and Bifurcation of Planar Vector Fields. Cambridge University Press, Cambridge
55. Ciocci MC, Litvak-Hinenzon A, Broer HW (2005) Survey on dissipative KAM theory including quasi-periodic bifurcation theory based on lectures by Henk Broer. In: Montaldi J, Ratiu T (eds) Geometric Mechanics and Symmetry: the Peyresq Lectures, LMS Lecture Notes Series, vol 306. Cambridge University Press, Cambridge, pp 303–355
56. Cushman RH, Bates LM (1997) Global Aspects of Classical Integrable Systems. Birkhäuser, Basel
57. Cushman RH, Dullin HR, Giacobbe A, Holm DD, Joyeux M, Lynch P, Sadovskií DA and Zhilinskií BI (2004) CO_2 molecule as a quantum realization of the 1 : 1 : 2 resonant swing-spring with monodromy. Phys Rev Lett 93:024302
58. Devaney RL (1989) An Introduction to Chaotic Dynamical Systems, 2nd edn. Addison-Wesley, Redwood City
59. Diacu F, Holmes P (1996) Celestial Encounters. The Origins of Chaos and Stability. Princeton University Press, Princeton
60. Duistermaat JJ (1980) On global action-angle coordinates. Commun Pure Appl Math 33:687–706
61. Dumortier F, Roussarie R, Sotomayor J (1991) Generic 3-parameter families of vector fields, unfoldings of saddle, focus and elliptic singularities with nilpotent linear parts. In: Dumortier F, Roussarie R. Sotomayor J, Zoladek H (eds) Bifurcations of Planar Vector Fields: Nilpotent Singularities and Abelian Integrals. LNM 1480, pp 1–164
62. Efstafhiou K (2005) Metamorphoses of Hamiltonian systems with symmetries. LNM, vol 1864. Springer, Heidelberg
63. Féjoz J (2004) Démonstration du "théorème d'Arnold" sur la stabilité du système planétaire (d'après Herman). Ergod Theory Dyn Syst 24:1–62
64. Gallavotti G, Bonetto F, Gentile G (2004) Aspects of Ergodic, Qualitative and Statistical Theory of Motion. Springer
65. Gelfreich VG, Lazutkin VF (2001) Splitting of Separatrices: perturbation theory and exponential smallness. Russ Math Surv 56:499–558
66. Guckenheimer J, Holmes P (1983) Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields. Springer
67. Hanßmann H (1988) The quasi-periodic centre-saddle bifurcation. J Differ Equ 142:305–370
68. Hanßmann H (2004) Hamiltonian Torus Bifurcations Related to Simple Singularities. In: Ladde GS, Medhin NG, Sambandham M (eds) Dynamic Systems and Applications, Atlanta 2003. Dynamic Publishers, pp 679–685
69. Hanßmann H (2007) Local and Semi-Local Bifurcations in Hamiltonian Dynamical Systems – Results and Examples. In: Lecture Notes in Mathematics, vol 1893. Springer, Berlin
70. Herman M (1977) Mesure de Lebesgue et nombre de rotation. In: Palis J, do Carmo M (eds) Geometry and Topology. In: Lecture Notes in Mathematics, vol 597. Springer, pp 271–293
71. Herman MR (1979) Sur la conjugaison différentiable des difféomorphismes du cercle à des rotations. Publ Math IHÉS 49:5–233
72. Hirsch MW (1976) Differential Topology. Springer
73. Hirsch MW, Pugh CC, Shub M (1977) Invariant Manifolds. In: Lecture Notes in Mathematics, vol 583. Springer
74. Hofer H, Zehnder E (1994) Symplectic invariants and Hamiltonian dynamics. Birkhäuser

75. Hopf E (1942) Abzweigung einer periodischen Lösung von einer stationären Lösung eines Differentialsystems. Ber Math-Phys Kl Sächs Akad Wiss Leipzig 94:1–22
76. Hopf E (1948) A mathematical example displaying features of turbulence. Commun Appl Math 1:303–322
77. Huygens C Œvres complètes de Christiaan Huygens, (1888–1950), vol 5, pp 241–263 and vol 17, pp 156–189. Martinus Nijhoff, The Hague
78. de Jong HH (1999) Quasiperiodic breathers in systems of weakly coupled pendulums: Applications of KAM theory to classical and statistical mechanics. Ph D Thesis, Univ. Groningen
79. Katok A, Hasselblatt B (1995) Introduction to the Modern Theory of Dynamical Systems. Cambridge University Press, Cambridge
80. Kolmogorov AN (1954) On the persistence of conditionally periodic motions under a small change of the Hamilton function. Dokl Akad Nauk SSSR 98:527–530 (in Russian); English translation: Stochastic Behavior in Classical and Quantum Hamiltonian Systems, Volta Memorial Conference (Como, 1977). In: Casati G, Ford J (eds) Lecture Notes in Physics, vol 93. Springer, Berlin pp 51–56 (1979); Reprinted in: Bai Lin Hao (ed) Chaos. World Scientific, Singapore, pp 81–86 (1984)
81. Kolmogorov AN (1957) The general theory of dynamical systems and classical mechanics. In: Gerretsen JCH, de Groot J (eds) Proceedings of the International Congress of Mathematicians, vol 1 (Amsterdam, 1954), North-Holland, Amsterdam, pp 315–333 (in Russian); Reprinted in: International Mathematical Congress in Amsterdam, (1954) (Plenary Lectures). Fizmatgiz, Moscow, pp 187–208 (1961); English translation as Appendix D in: Abraham RH (1967) Foundations of Mechanics. Benjamin, New York, pp 263–279; Reprinted as Appendix in [1], pp 741–757
82. Kuznetsov YA (2004) Elements of Applied Bifurcation Theory, 3rd edn. In: Applied Mathematical Sciences, vol 112. Springer, New York
83. Landau LD (1944) On the problem of turbulence. Akad Nauk 44:339
84. Landau LD, Lifschitz EM (1959) Fluid Mechanics. Pergamon, Oxford
85. Laskar J (1995) Large scale chaos and marginal stability in the Solar System, XIth International Congress of Mathematical Physics (Paris, 1994). In: Iagolnitzer D (ed) Internat Press, Cambridge, pp 75–120
86. Litvak-Hinenzon A, Rom-Kedar V (2002) Parabolic resonances in 3 degree of freedom near-integrable Hamiltonian systems. Phys D 164:213–250
87. de la Llave R, González A, Jorba À, Villanueva J (2005) KAM theory without action-angle variables. Nonlinearity 18:855–895
88. Lochak P (1999) Arnold diffusion; a compendium of remarks and questions. In: Simó C (ed) Hamiltonian systems with three or more degrees of freedom (S'Agaró, 1995), NATO ASI Series C: Math Phys Sci, vol 533. Kluwer, Dordrecht, pp 168–183
89. Lochak P, Marco J-P (2005) Diffusion times and stability exponents for nearly integrable analytic systems, Central Eur J Math 3:342–397
90. Lochak P, Neĭshtadt AI (1992) Estimates of stability time for nearly integrable systems with a quasiconvex Hamiltonian. Chaos 2:495–499
91. Lukina O (2008) Geometry of torus bundles in Hamiltonian systems, Ph D Thesis, Univ. Groningen
92. MacKay RS (1993) Renormalisation in area-preserving maps. World Scientific
93. Marco J-P, Sauzin D (2003) Stability and instability for Gevrey quasi-convex near-integrable Hamiltonian systems. Publ Math Inst Hautes Etud Sci 96:199–275
94. Markus L, Meyer KR (1980) Periodic orbits and solenoids in generic Hamiltonian dynamical systems. Am J Math 102: 25–92
95. Matveev VS (1996) Integrable Hamiltonian systems with two degrees of freedom. Topological structure of saturated neighborhoods of points of focus-focus and saddle-saddle types. Sb Math 187:495–524
96. McDuff D, Salamon D (1995) Introduction to Symplectic Geometry. Clarendon/Oxford University Press
97. Meyer KR, Hall GR (1992) Introduction to Hamiltonian Dynamical Systems and the *N*-Body Problem. In: Applied Mathematical Sciences, vol 90. Springer
98. Meiss JD (1986) Class renormalization: Islands around islands. Phys Rev A 34:2375–2383
99. de Melo W, van Strien SJ (1991) One-Dimensional Dynamics. Springer
100. Mielke A (1991) Hamiltonian and Lagrangian Flows on Center Manifolds – with Applications to Elliptic Variational Problems. In: Lecture Notes in Mathematics, vol 1489. Springer
101. Milnor JW (2006) Dynamics in One Complex Variable, 3rd edn. In: Ann. Math. Studies, vol 160. Princeton University Press, Princeton
102. Morbidelli A, Giorgilli A (1995) Superexponential Stability of KAM Tori. J Stat Phys 78:1607–1617
103. Morbidelli A, Giorgilli A (1995) On a connection between KAM and Nekhoroshev's theorems. Physica D 86:514–516
104. Moser JK (1962) On invariant curves of area-preserving mappings of an annulus. Nachr Akad Wiss Göttingen II, Math-Phys Kl 1:1–20
105. Moser JK (1966) On the theory of quasiperiodic motions. SIAM Rev 8(2):145–172
106. Moser JK (1967) Convergent series expansions for quasi-periodic motions. Math Ann 169:136–176
107. Moser JK (1968) Lectures on Hamiltonian systems. Mem Am Math Soc 81:1–60
108. Moser JK (1973) Stable and random motions in dynamical systems, with special emphasis to celestial mechanics. In: Ann. Math. Studies, vol 77. Princeton University Press, Princeton
109. Nekhoroshev NN (1977) An exponential estimate of the time of stability of nearly-integrable Hamiltonian systems. Russ Math Surv 32:1–65
110. Nekhoroshev NN (1985) An exponential estimate of the time of stability of nearly integrable Hamiltonian systems II. In: Oleinik OA (ed) Topics in Modern Mathematics, Petrovskii Seminar No.5. Consultants Bureau, pp 1–58
111. Newhouse SE, Palis J, Takens F (1983) Bifurcations and stability of families of diffeomorphisms. Publ Math IHÉS 57:5–71
112. Newhouse SE, Ruelle D, Takens F (1978) Occurrence of strange Axiom A attractors near quasi-periodic flows on $\mathbb{T}^m$, $m \geq 3$. Commun Math Phys 64:35–40
113. Niederman L (2004) Prevalence of exponential stability among nearly-integrable Hamiltonian systems. Ergod Theory Dyn Syst 24(2):593–608
114. Nguyen Tien Zung (1996) Kolmogorov condition for integrable systems with focus-focus singularities. Phys Lett A 215(1/2):40–44

115. Oxtoby J (1971) Measure and Category. Springer
116. Palis J, de Melo M (1982) Geometric Theory of Dynamical Systems. Springer
117. Palis J, Takens F (1993) Hyperbolicity & Sensitive Chaotic Dynamics at Homoclinic Bifurcations. Cambridge University Press, Cambridge
118. Poincaré H (1980) Sur le problème des trois corps et les équations de la dynamique. Acta Math 13:1–270
119. Pöschel J (1982) Integrability of Hamiltonian systems on Cantor sets. Commun Pure Appl Math 35(5):653–696
120. Pöschel J (1993) Nekhoroshev estimates for quasi-convex Hamiltonian systems. Math Z 213:187–216
121. Pöschel J (2001) A lecture on the classical KAM Theorem. In: Proc Symp Pure Math 69:707–732
122. Rink BW (2004) A Cantor set of tori with monodromy near a focus-focus singularity. Nonlinearity 17:347–356
123. Robinson C (1995) Dynamical Systems. CRC Press
124. Roussarie R (1997) Smoothness properties of bifurcation diagrams. Publ Mat 41:243–268
125. Ruelle D (1989) Elements of Differentiable Dynamics and Bifurcation Theory. Academic Press
126. Ruelle D, Takens F (1971) On the nature of turbulence. Commun Math Phys 20:167–192; 23:343–344
127. Sevryuk MB (2007) Invariant tori in quasi-periodic non-autonomous dynamical systems via Herman's method. DCDS-A 18(2/3):569–595
128. Simó C (2001) Global dynamics and fast indicators. In: Broer HW, Krauskopf B, Vegter G (eds) Global Analysis of Dynamical Systems, Festschrift dedicated to Floris Takens for his 60th birthday. IOP, Bristol and Philadelphia, pp 373–390
129. Spivak M (1970) Differential Geometry, vol I. Publish or Perish
130. Takens F (1973) Introduction to Global Analysis. Comm. 2 of the Math. Inst. Rijksuniversiteit Utrecht
131. Takens F (1974) Singularities of vector fields. Publ Math IHÉS 43:47–100
132. Takens F (1974) Forced oscillations and bifurcations. In: Applications of Global Analysis I, Comm 3 of the Math Inst Rijksuniversiteit Utrecht (1974); In: Broer HW, Krauskopf B, Vegter G (eds) Global Analysis of Dynamical Systems, Festschrift dedicated to Floris Takens for his 60th birthday. IOP, Bristol and Philadelphia, pp 1–62
133. Thom R (1989) Structural Stability and Morphogenesis. An Outline of a General Theory of Models, 2nd edn. Addison-Wesley, Redwood City (English; French original)
134. Vũ Ngọc San (1999) Quantum monodromy in integrable systems. Commun Math Phys 203:465–479
135. Waalkens H, Junge A, Dullin HR (2003) Quantum monodromy in the two-centre problem. J Phys A Math Gen 36:L307-L314
136. Wagener FOO (2003) A note on Gevrey regular KAM theory and the inverse approximation lemma. Dyn Syst 18:159–163
137. Wiggins S (1990) Introduction to Applied Nonlinear Dynamical Systems and Chaos. Springer
138. Wiggins S, Wiesenfeld L, Jaffe C, Uzer T (2001) Impenetrable barriers in phase-space. Phys Rev Lett 86(24):5478–5481
139. Yoccoz J-C (1983) C^1-conjugaisons des difféomorphismes du cercle. In: Palis J (ed) Geometric Dynamics, Proceedings, Rio de Janeiro (1981) Lecture Notes in Mathematics, vol 1007, pp 814–827
140. Yoccoz J-C (1992) Travaux de Herman sur les tores invariants. In: Séminaire Bourbaki, vol 754, 1991–1992. Astérisque 206:311–344
141. Yoccoz J-C (1995) Théorème de Siegel, nombres de Bruno et polynômes quadratiques. Astérisque 231:3–88
142. Yoccoz J-C (2002) Analytic linearization of circle diffeomorphisms. In: Marmi S, Yoccoz J-C (eds) Dynamical Systems and Small Divisors, Lecture Notes in Mathematics, vol 1784. Springer, pp 125–174
143. Zehnder E (1974) An implicit function theorem for small divisor problems. Bull Am Math Soc 80(1):174–179
144. Zehnder E (1975) Generalized implicit function theorems with applications to some small divisor problems, I and II. Commun Pure Appl Math 28(1):91–140; (1976) 29(1):49–111

Hamilton–Jacobi Equations and Weak KAM Theory

Antonio Siconolfi
Dip. di Matematica, "La Sapienza" Università di Roma, Roma, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Subsolutions
Solutions
First Regularity Results for Subsolutions
Critical Equation and Aubry Set
An Intrinsic Metric
Dynamical Properties of the Aubry Set
Long-Time Behavior of Solutions to the Time-Dependent Equation
Main Regularity Result
Future Directions
Bibliography

Glossary

Hamilton–Jacobi equations This class of first-order partial differential equations has a central relevance in several branches of mathematics, both from a theoretical and an application point of view. It is of primary importance in classical mechanics, Hamiltonian dynamics, Riemannian and Finsler geometry, and optimal control theory, as well. It furthermore appears in the classical limit of the Schrödinger equation. A connection with Hamilton's equations, in the case where the Hamiltonian has sufficient regularity, is provided by the classical Hamilton–Jacobi method which shows that the graph of the differential of any regular, say C^1,

global solution to the equation is an invariant subset for the corresponding Hamiltonian flow. The drawback of this approach is that such regular solutions do not exist in general, even for very regular Hamiltonians. See the next paragraph for more comments on this issue.

Viscosity solutions As already pointed out, Hamilton–Jacobi equations do not have in general *global* classical solutions, i. e. everywhere differentiable functions satisfying the equation pointwise. The method of characteristics just yields *local* classical solutions. This explains the need of introducing weak solutions. The idea for defining those of viscosity type is to consider C^1 functions whose graph, up to an additive constant, touches that of the candidate solution at a point and then stay locally above (resp. below) it. These are the viscosity test functions, and it is required that the Hamiltonian satisfies suitable inequalities when its first-order argument is set equal to the differential of them at the first coordinate of the point of contact. Similarly it is defined the notion of viscosity sub, supersolution. Clearly a viscosity solution satisfies pointwise the equation at any differentiability points. A peculiarity of the definition is that a viscosity solution can admit no test function at some point, while the nonemptiness of both classes of test functions is equivalent to the solution being differentiable at the point. Nevertheless powerful existence, uniqueness and stability results hold in the framework of viscosity solution theory. The notion of viscosity solutions was introduced by Crandall and Lions at the beginning of the 1980s. We refer to Bardi and Capuzzo Dolcetta [2], Barles [3], Koike [24] for a comprehensive treatment of this topic.

Semiconcave and semiconvex functions These are the appropriate regularity notions when working with viscosity solution techniques. The definition is given by requiring some inequalities, involving convex combinations of points, to hold. These functions possess viscosity test functions of one of the two types at any point. When the Hamiltonian enjoys coercivity properties ensuring that any viscosity solution is locally Lipschitz-continuous then a semiconcave or semiconcave function is the solution if and only if it is classical solution almost everywhere, i. e. up to a set of zero Lebesgue measure.

Metric approach This method applies to stationary Hamilton–Jacobi equations with the Hamiltonian only depending on the state and momentum variable. This consists of defining a length functional, on the set of Lipschitz-continuous curves, related to the corresponding sublevels of the Hamiltonian. The associated length distance, obtained by performing the infimum of the intrinsic length of curves joining two given points, plays a crucial role in the analysis of the equation and, in particular, enters in representation formulae for any viscosity solution. One important consequence is that only the sublevels of the Hamiltonian matter for determining such solutions. Accordingly the convexity condition on the Hamiltonian can be relaxed, just requiring quasiconvexity, i. e. convexity of sublevels. Note that in this case the metric is of Finsler type and the sublevels are the unit cotangent balls of it.

Critical equations To any Hamiltonian is associated a one-parameter family of Hamilton–Jacobi equations obtained by fixing a constant level of Hamiltonian. When studying such a family, one comes across a threshold value under which no subsolutions may exist. This is called the critical value and the same name is conferred to the corresponding equation. If the ground space is compact then the critical equation is unique among those of the family for which viscosity solutions do exist. When, in particular, the underlying space is a torus or, in other terms, the Hamiltonian is $\mathbb{Z}^N$-periodic then such functions play the role of correctors in related homogenization problems.

Aubry set The analysis of the critical equation shows that the obstruction for getting subsolutions at subcritical levels is concentrated on a special set of the ground space, in the sense that no critical subsolution can be strict around it. This is precisely the Aubry set. This is somehow compensated by the fact that critical subsolutions enjoy extra regularity properties on the Aubry set.

Definition of the Subject

The article aims to illustrate some applications of weak KAM theory to the analysis of Hamilton–Jacobi equations. The presentation focuses on two specific problems, namely the existence of C^1 classical subsolutions for a class of stationary (i. e. independent of the time) Hamilton–Jacobi equations, and the long-time behavior of viscosity solutions of an evolutive version of it.

The Hamiltonian is assumed to satisfy mild regularity conditions, under which the corresponding Hamilton equations cannot be written. Consequently PDE techniques will be solely employed in the analysis, since the powerful tools of the Hamiltonian dynamics are not available.

Introduction

Given a continuous or more regular Hamiltonian $H(x, p)$ defined on the cotangent bundle of a boundaryless manifold M, where x and p are the state and the momentum variable, respectively, and satisfying suitable convexity and coercivity assumptions, is considered the family of Hamilton–Jacobi equations

$$H(x, Du) = a \quad x \in M\,, \tag{1}$$

with a a real parameter, as well as the time-dependent version

$$w_t + H(x, Dw) = 0 \quad x \in M\,,\ t \in (0, +\infty)\,, \tag{2}$$

where Du and u_t stand for the derivative with respect to state and time variable, respectively. As a matter of fact, it will be set, for the sake of simplicity, to either $M = \mathbb{R}^N$ (noncompact case) or $M = \mathbb{T}^N$ (compact case), where $\mathbb{T}^N$ indicates the flat torus endowed with the Euclidean metric and with the cotangent bundle identified to $\mathbb{T}^N \times \mathbb{R}^N$.

The main scope of this article is to study the existence of the C^1 classical subsolution to (1), and the long-time behavior of viscosity solutions to (2) by essentially employing tools issued from weak KAM theory.

Some of the results that will be outlined are valid with the additional assumption of compactness for the underlying manifold, in particular those concerning the asymptotics of solutions to (2).

For both issues it is crucial to perform a qualitative analysis of (1) for a known value of the parameter a, qualified as *critical*; accordingly Eq. (1) is called critical when a is equal to the critical value. This analysis leads to detection of a special closed subset of the ground space, named after Aubry, where any locally Lipschitz-continuous subsolution to (1) enjoys some additional regularity properties, and behaves in a peculiar way. This will have a central role in the presentation.

The requirements on H will be strengthened to obtain some theorems, but we remain in a setting where the corresponding Hamilton equations cannot be written. Consequently PDE techniques will be solely employed in the analysis, since the powerful tools of Hamiltonian dynamics are not available.

Actually the notion of critical value was independently introduced by Ricardo Mañé at the beginning of the 1980s, in connection with the analysis of integral curves of the Euler–Lagrange flow enjoying some global minimizing properties, and by P.L. Lions, S.R.S. Varadhan and G. Papanicolaou [25] in 1987 in the framework of viscosity solutions theory, for studying the periodic homogenization of Hamilton–Jacobi equations.

The Aubry set was determined and analyzed by Serge Aubry, in a pure dynamical way, as the union of the supports of integral curves of Euler–Lagrange flow possessing suitable minimality properties. John Mather defined (1986), in a more general framework, a set, contained in the Aubry set, starting from special probability measures invariant with respect to the flow. See Contreras and Iturriaga, [9], for an account on this theory.

The first author to point out the link between Aubry–Mather theory and weak solutions to the critical Hamilton–Jacobi equation was Albert Fathi, see [16,17], with the so-called weak KAM theory (1996); he thoroughly investigated the PDE counterpart of the dynamical phenomena occurring at the critical level. However, his investigation is still within the framework of the dynamical systems theory, requires the Hamiltonian to be at least C^2, and requires the existence of associated Hamiltonian flow as well.

The work of Fathi and Siconolfi (2005) [19,20], completely bypassed such assumptions and provided a geometrical analysis of the critical equation independent of the flow, which made it possible to deal with nonregular Hamiltonians. The new idea being the introduction of a length functional, intrinsically related to H, for *any* curve, and of the related distance, as well. The notion of the Aubry set was suitably generalized to this broader setting.

Other important contributions in bridging the gap between PDE and the dynamical viewpoint have been made by Evans and Gomes [14,15].

The material herein is organized as follows: the Sects. "Subsolutions", "Solutions" are of introductory nature and illustrate the notions of viscosity (sub)solution with their basic properties. Some fundamental techniques used in this framework are introduced as well. Section "First Regularity Results for Subsolutions" deals with issues concerning regularity of subsolutions to (1). The key notion of the Aubry set is introduced in Sect. "Critical Equation and Aubry Set" in connection with the investigation of the critical equation, and a qualitative analysis of it, specially devoted to looking into metric and dynamical properties, is performed in Sects. "An Intrinsic Metric", "Dynamical Properties of the Aubry Set". Sections "Long-Time Behavior of Solutions to the Time-Dependent Equation", "Main Regularity Result" present the main results relative to the long-time behavior of solutions to (2) and the existence of C^1 subsolutions to (1). Finally, Sect. "Future Directions" gives some ideas of possible developments in the topic.

Subsolutions

First I will detail the basic conditions postulated throughout the paper for H. Additional properties required for obtaining some particular result, will be introduced when needed. The Hamiltonian is assumed to be

$$\text{continuous in both variables}\,, \tag{3}$$

to satisfy the coercivity assumption

$$\{(x,p)\colon H(y,p)\le a\} \quad \text{is compact for any } a \tag{4}$$

and the following quasiconvexity conditions for any $x \in M$, $a \in \mathbb{R}$

$$\{p\colon H(x,p)\le a\} \text{ is convex} \tag{5}$$

$$\partial\{p\colon H(x,p)\le a\} = \{p\colon H(x,p)=a\} \tag{6}$$

where ∂, in the above formula, indicates the boundary. The a-sublevel of the Hamiltonian appearing in (5) will be denoted by $Z_a(x)$. It is a consequence of the coerciveness and convexity assumptions on H that the set-valued map $x \mapsto Z_a(x)$ possesses convex compact values and, in force of (3), is upper semicontinuous; it is in addition continuous at any point x where $\operatorname{int} Z_a(x) \neq \emptyset$. Here (semi)continuous must be understood with respect to the Hausdorff metric.

Next will be given four different definitions of weak subsolutions to Eq. (1) and their equivalence will be proved. From this it can be seen that the family of functions so detected is intrinsically related to the equation. As a matter of fact, it will be proved, under more stringent assumptions, that this family is the closure of the classical (i. e. C^1) subsolutions in the locally uniform topology.

Some notations and definitions must be preliminarily introduced. Given two continuous functions u and v, it is said that v is a *(strict) supertangent* to u at some point x_0 if such point is a (strict) local maximizer of $u - v$. The notion of *subtangent* is obtained by replacing a maximizer with a minimizer. Since (sub, super)tangents are involved in the definition of viscosity solution, they will be called in the sequel viscosity test functions. It is necessary to check:

Proposition 1 *Let u be a continuous function possessing both C^1 supertangents and subtangents at a point x_0, then u is differentiable at x_0.*

Recall that by Rademacher Theorem a locally Lipschitz function is differentiable almost everywhere (for short a.e.), with respect to the Lebesgue measure. For such a function w the (Clarke) generalized gradient at any point x is defined by

$$\partial w(x) = \operatorname{co}\{p = \lim_i D\,w(x_i) : x_i \text{ differentiability point of } u\,,\ \lim_i x_i = x\}\,,$$

where co denotes the convex hull.

Remark 2 Record for later use that this set of weak derivatives can be retrieved even if the differentiability points are taken not in the whole ground space, but just outside a set of vanishing Lebesgue measure.

The generalized gradient is nonempty at any point; if it reduces to a singleton at some x, then the function w is *strictly differentiable* at x, i. e. it is differentiable and Du is continuous at x. The set-valued function $x \mapsto \partial w(x)$ possesses convex compact values and is upper semicontinuous. The following variational property holds:

$$0 \in \partial w(x) \quad \text{at any local minimizer or maximizer of } w\,, \tag{7}$$

furthermore, if ψ is C^1 then

$$\partial(w-\psi)(x) = \partial w(x) - D\psi(x)\,. \tag{8}$$

First definition of weak subsolution A function u is said to be an *a.e.* subsolution to (1) if it is locally Lipschitz-continuous and satisfies

$$H(x, Du(x)) \le a \quad \text{for } x \text{ in a subset of } M \text{ with full measure}\,.$$

Second definition of weak subsolution A function u is said to be a *viscosity* subsolution *of first type* to (1) if it is continuous and

$$H(x_0, D\psi(x_0)) \le a\,,$$

or equivalently

$$D\psi(x_0) \in Z_a(x_0)\,,$$

for any $x_0 \in M$, any C^1 supertangent ψ to u at x_0.

The previous definition can be equivalently rephrased by taking test functions of class C^k, $1 < k \le +\infty$, or simply differentiable, instead of C^1.

Third definition of weak subsolution A function u is a *viscosity* subsolution *of second type* if it satisfies the previous definition with subtangent in place of supertangent.

Fourth definition of weak subsolution A function u is a subsolution *in the sense of Clarke* if

$$\partial u(x) \subset Z_a(x) \quad \text{for all } x \in M\,.$$

Note that this last definition is unique based on a condition holding at *any* point of M, and not just at the differentiability points of u or at points where some test function exists. This fact will be exploited in the forthcoming definition of *strict* subsolution.

Proposition 3 *The previous four definitions are equivalent.*

We first show that the viscosity subsolutions of both types are locally Lipschitz-continuous. It is exploited that Z_a, being upper semicontinuous, is also locally bounded, namely for any bounded subset B of M there is a positive r with

$$Z_a(x) \subset B(0,r) \quad \text{for } x \in B\,, \tag{9}$$

where $B(0,r)$ is the Euclidean ball centered at 0 with radius r. The minimum r for which (9) holds true will be indicated by $|Z_a|_{\infty,B}$. The argument is given for a viscosity subsolution of first type, say u; the proof for the others is similar.

Assume by contradiction that there is an open bounded domain B_1 where u is not Lipschitz-continuous, then consider an open bounded domain B_2 containing B_1 such that

$$\alpha := \inf\{|x-y|\,,\ x \in B_1\,,\ y \in \partial B_2\} > 0\,,$$

and choose an l_0 such that

$$|Z_a|_{\infty,B_2} < l_0 \tag{10}$$

$$\sup_{B_2} u - \inf_{B_1} u - l_0\alpha < 0 \tag{11}$$

Since u is not Lipschitz-continuous on B_1, a pair of points x_0, x_1 in B_1 can be found satisfying

$$u(x_1) - u(x_0) > l_0|x_1 - x_0|\,, \tag{12}$$

which shows that the function $x \mapsto u(x) - u(x_0) - l_0|x - x_0|$), has a positive supremum in B_2. On the other hand such a function is negative on ∂B_2 by (11), and so it attains its maximum in B_2 at a point $\overline{x} \in B_2$, or, in other terms, $x \mapsto u(x_0) + l_0|x - x_0|)$ is supertangent to u at $\overline{x} \neq x_0$. Consequently

$$l_0 \frac{x_1 - x_0}{|x_1 - x_0|} \in Z_a(x_1)\,,$$

in contradiction with (10).

Since at every differentiability point u can be taken as a test function of itself, then any viscosity subsolution of both types is also an a.e. subsolution. The a.e. subsolutions, in turn, satisfy the fourth definition above, thanks to the definition of generalized gradient, Remark 2, and the fact that H is convex in p and continuous in both arguments.

Finally, exploit (7) and (8) to see that the differential of any viscosity test function to u at some point x_0 is contained in $\partial u(x_0)$. This shows that any subsolution in the Clarke sense is also a viscosity subsolution of the first and second type.

In view of Proposition 3, from now on any element of this class of functions will be called a subsolution of (1) without any further specification, similarly the notion of a subsolution in an open subset of M can be given. Note that for any subsolution u, any bounded open domain B, the quantity $|Z_a|_{\infty,B}$ is a Lipschitz constant in B for every subsolution, consequently the family of all subsolutions to (1) is locally equiLipschitz-continuous.

A *conjugate* Hamiltonian $\check{H}$ can be associated to H, it is defined by

$$\check{H}(x,p) = H(x,-p) \quad \text{for any } x, p\,. \tag{13}$$

Note that $\check{H}$ satisfies, as H does, assumptions (3)–(5). The two corresponding conjugate Hamilton–Jacobi equations have the same family of subsolutions, up to a change of sign, as is apparent looking at the first definition of subsolution.

Next we will have a closer look at the family of subsolutions to (1), denoted from now on by S_a; the properties deduced will be exploited in the next sections. Advantage is taken of this to illustrate a couple of basic arguments coming from viscosity solutions theory.

Proposition 4 *The family S_a is stable with respect to the local uniform convergence.*

The key point in the proof of this result is to use the *same* C^1 function, at different points, for testing the limit as well as the approximating functions. This is indeed the primary trick to obtaining stability properties in the framework of viscosity solutions theory.

Let u_n be a sequence in S_a and $u_n \to u$ locally uniformly in M. Let ψ be a supertangent to u at some point x_0, it can be assumed, without loss of generality, that ψ is a strict subtangent, by adding a suitable quadratic term. Therefore, there is a compact neighborhood U of x_0 where x_0 itself is the *unique* maximizer of $u - \psi$. Any sequence x_n of maximizers of $u_n - \psi$ in U converges to a maximizer of $u - \psi$, and so $x_n \to x_0$, and consequently lies in the interior of U for n large enough. In other terms ψ is supertangent to u_n at x_n, when n is sufficiently large.

Consequently $H(x_n, D\psi(x_n) \leq a$, which implies, exploiting the continuity of the Hamiltonian and passing at the limit, $H(x_0, D\psi(x_0)) \leq a$, as desired.

Take into account the equiLipschitz character of subsolutions to (1), and the fact that the subsolution property is not affected by addition of a constant, to obtain by slightly adjusting the previous argument, and using the Ascoli Theorem:

Proposition 5 *Let $u_n \in S_{a_n}$, with a_n converging to some a. Then the sequence u_n converges to $u \in S_a$, up to addition of constants and extraction of a subsequence.*

Before ending the section, a notion which will have some relevance in what follows will be introduced.

A subsolution u is said to be *strict* in some open subset $\Omega \subset M$ if

$$\partial u(x) \subset \operatorname{int} Z_a(x) \quad \text{for any } x \in \Omega\,,$$

where int stands for interior. Since the multivalued map $x \mapsto \partial u(x)$ is upper semicontinuous, this is equivalent to

$$\operatorname{ess\,sup}_{\Omega'} H(x, Du(x)) < a \qquad \text{for any } \Omega' \text{ compactly contained in } \Omega\,,$$

where the expression *compactly contained* means that the closure of Ω' is compact and is contained in Ω. Accordingly, the maximal (possibly empty) open subset W_u where u is strict is given by the formula

$$W_u := \{x\colon \partial u(x) \subset \operatorname{int} Z_a(x)\}\,. \tag{14}$$

Solutions

Unfortunately the relevant stability properties pointed out in the previous section for the family of subsolutions, do not hold for the a.e. *solutions*, namely the locally Lipschitz-continuous functions satisfying the equation up to a subset of M with vanishing measure. Take, for instance, the sequence u_n in $\mathbb{T}^1$, obtained by linear interpolation of

$$u_n\left(\frac{k}{2n}\right) = 0 \quad \text{for } k \text{ even}, 0 \leq k \leq 2n$$

$$u_n\left(\frac{k}{2n}\right) = \frac{1}{2n} \quad \text{for } k \text{ odd}, 0 \leq k \leq 2n\,,$$

then it comprises a.e. solutions of (1), with $H(x, p) = |p|$ and $a = 1$. But its uniform limit is the null function, which is an a.e. subsolution of the same equation, according to Proposition 4, but fails to be an a.e. solution. This lack of stability motivates the search for a stronger notion of weak solution. The idea is to look at the properties of S_a with respect to the operations of sup and inf.

Proposition 6 *Let $\tilde{S} \subset S_a$ be a family of locally equibounded functions, then the function defined as the pointwise supremum, or infimum, of the elements of $\tilde{S}$ is a subsolution to (1).*

Set $u(x) = \inf\{v(x)\colon v \in \tilde{S}\}$. Let ψ, u_n be a C^1 subtangent to u at a point x_0, and a sequence of functions in $\tilde{S}$ with $u_n(x_0) \to u(x_0)$, respectively.

Since the sequence u_n is made up of locally equibounded and locally equiLipschitz-continuous functions, it locally uniformly converges, up to a subsequence, by Ascoli Theorem, to a function w which belongs, in force of Proposition 4, to S_a. In addition $w(x_0) = u(x_0)$, and w is supertangent to u at x_0 by the very definition of u. Therefore, ψ is also subtangent to v at x_0 and so $H(x_0, D\psi(x_0)) \leq a$, which shows the assertion. The same proof, with obvious adaptations, allows us to handle the case of the pointwise supremum.

Encouraged by the previous result, consider the subsolutions of (1) enjoying some extremality properties. A definition is preliminary. A family $\tilde{S}$ of locally equibounded subsolutions to (1) is said to be *complete* at some point x_0 if there exists ε_{x_0} such that if two subsolutions u_1, u_2 agree outside some neighborhood of x_0 with radius less than ε_{x_0} and $u_1 \in \tilde{S}$ then $u_2 \in \tilde{S}$.

The interesting point in the next proposition is that the subsolutions which are extremal with respect to a complete family, possess an additional property involving the viscosity test functions.

Proposition 7 *Let u be the pointwise supremum (infimum) of a locally equibounded family $\tilde{S} \subset S_a$ complete at a point x_0, and let ψ be a C^1 subtangent (supertangent) to u at x_0. Then $H(x_0, D\psi(x_0)) = a$.*

Only the case where u is a pointwise supremum will be discussed. The proof is based on a *push up* method that will be again used in the sequel.

If, in fact, the assertion were not true, there should be a C^1 strict subtangent ψ at x_0, with $\psi(x_0) = u(x_0)$, such that $H(x_0, D\psi(x_0)) < a$. The function ψ, being C^1, is a (classical) subsolution of (1) in a neighborhood U of x_0. Push up a bit the test function to define

$$v = \begin{cases} \max\{\psi + \varepsilon\,,\, u\} & \text{in } B(x_0, \varepsilon) \\ u & \text{otherwise} \end{cases} \tag{15}$$

with the positive constant ε chosen so that $B(x_0, \varepsilon) \subset U$ and $\varepsilon < \varepsilon_{x_0}$, where ε_{x_0} is the quantity appearing in the definition of a complete family of subsolutions at x_0. By the

Proposition 6, the function v belongs to S_a, and is equal to $u \in \tilde{S}$ outside $B(x_0, \varepsilon)$. Therefore, $v \in \tilde{S}$, which is in contrast with the maximality of u because $v(x_0) > u(x_0)$.

Proposition 7 suggests the following two definitions of weak solution in the viscosity sense, or viscosity solutions for Eq. (1).

The function u is a *viscosity solution of the first type* if it is a subsolution and for any x_0, any C^1 subtangent to u at x_0 one has $H(x_0, D\psi(x_0)) = a$. The *viscosity solutions of the second type* are definite by replacing the subtangent with supertangent. Such functions are clearly a.e. solutions.

The same proof of Proposition 4, applied to C^1 subtangents as well as supertangents, gives:

Proposition 8 *The family of viscosity solutions (of both types) to (1) is stable with respect the local uniform convergence.*

Moreover, the argument of Proposition 6, with obvious adaptations, shows:

Proposition 9 *The pointwise infimum (supremum) of a family of locally equibounded viscosity solutions of the first (second) type is a viscosity solution of the first (second) type to (1).*

Morally, it can be said that the solutions of the first type enjoy some maximality properties, and some minimality properties hold for the others. Using the notion of the strict subsolution, introduced in the previous section, the following partial converse of Proposition 7 can be obtained:

Proposition 10 *Let Ω, u, φ be a bounded open subset of M, a viscosity solution to (1) of the first (second) type, and a strict subsolution in Ω coincident with u on $\partial\Omega$, respectively. Then $u \geq \varphi$ ($u \leq \varphi$) in Ω.*

The proof rests on a regularization procedure of φ by mollification. Assume that u is a viscosity solution of the first type, the other case can be treated similarly. The argument is by contradiction, then admit that the minimizers of $u - \varphi$ in $\overline{\Omega}$ (the closure of Ω) are in an open subset Ω' compactly contained in Ω. Define for $x \in \Omega'$, $\delta > 0$,

$$\varphi_\delta(x) = \int \zeta_\delta(y - x)\varphi(y)\,\mathrm{d}y\,,$$

where ζ_δ is a standard C^∞ mollifier supported in $B(0, \delta)$. By using the convex character of the Hamiltonian and Jensen Lemma, it can be found

$$H(x, D\varphi_\delta(x)) \leq \int \zeta_\delta(y - x)H(x, D\varphi(y))\,\mathrm{d}y.$$

Therefore, taking into account the stability of the set of minimizers under the uniform convergence, and that φ is a strict subsolution, δ can be chosen so small that φ_δ is a C^∞ strict subsolution of (1) in Ω' and, in addition, is subtangent to u at some point of Ω'. This is in contrast with the very definition of viscosity solution of the first type. The above argument will be used again, and explained with some more detail, in the next section.

The family of viscosity solutions of first and second type coincide for conjugate Hamilton–Jacobi equations, with Hamiltonian H and $\check{H}$, up to a change of sign. More precisely:

Proposition 11 *A function u is a viscosity solution of the first (second) type to*

$$\check{H}(x, Du) = a \quad \text{in } M \tag{16}$$

if and only if $-u$ is a viscosity solution of the second (first) type to (1).

In fact, if u, ψ are a viscosity solution of the first type to (16) and a C^1 supertangent to $-u$ at a point x_0, respectively, then $-u$ is a subsolution to (1), and $-\psi$ is supertangent to u at x_0 so that

$$a = \check{H}(x_0, -D\psi(x_0)) = H(x_0, -D\psi(x_0))\,,$$

which shows that $-u$ is indeed a viscosity solution of the second type to (1). The other implications can be derived analogously.

The choice between the two types of viscosity solutions is just a matter of taste, since they give rise to two completely equivalent theories. In this article those of the first type are selected, and they are referred to from now on as (viscosity) solutions, without any further specification.

Next a notion of regularity is introduced, called semiconcavity (semiconvexity), which fits the viscosity solutions framework, and that will be used in a crucial way in Sect. "Main Regularity Result". The starting remark is that even if the notion of viscosity solution of the first (second) type is more stringent than that of the a.e. solution, as proved above, the two notions are nevertheless equivalent for concave (convex) functions. In fact a function of this type, say u, which is locally Lipschitz-continuous, satisfies the inequality

$$u(y) \leq (\geq)\, u(x_0) + p\,(y - x)\,,$$

for any x_0, y, $p \in \partial u(x_0)$. It is therefore apparent that it admits (global) linear supertangents (subtangents) at any point x_0. If there were also a C^1 subtangent (supertangent), say ψ, at x_0, then u should be differentiable at x_0 by Proposition 1, and $Du(x_0) = D\psi(x_0)$, so that if u were an a.e. solution then $H(x_0, D\psi(x_0)) = a$, as announced.

In the above argument the concave (convex) character of u was not exploited to its full extent, but it was just used to show the existence of C^1 super(sub)tangents at any point x and that the differentials of such test functions make up ∂u. Clearly such a property is still valid for a larger class of functions. This is the case for the family of so-called *strongly semiconcave (semiconvex)* functions, which are concave (convex) up to the subtraction (addition) of a quadratic term. This point will now be outlined for strongly semiconcave functions, a parallel analysis could be performed in the strongly semiconvex case.

A function u is said to be strongly semiconcave if $u(x) - k|x - x_0|^2$ is concave for some positive constant k, some $x_0 \in M$. Starting from the inequality

$$u(\lambda x_1 + (1-\lambda)x_2) - k|\lambda x_1 + (1-\lambda)x_2 - x_0|^2 \geq \lambda(u(x_1) - k|x_1 - x_0|^2) + (1-\lambda)(u(x_2) - k|x_2 - x_0|^2)$$

which holds for any $x_1, x_2, \lambda \in [0, 1]$, it is derived through straightforward calculations

$$u(\lambda x_1 + (1-\lambda)x_2) - \lambda u(x_1) - (1-\lambda)u(x_2) \geq -k\lambda(1-\lambda)|x_1 - x_2|^2 , \quad (17)$$

which is actually a property equivalent to strong semiconcavity. This shows that for such functions the subtraction of $k|x - x_0|^2$, for *any* $x_0 \in M$, yields the concavity property. Therefore, given any x_0, and taking into account (8), one has

$$u(x) - k|x - x_0|^2 \leq u(x_0) + p(x - x_0) ,$$

for any x, any $p \in \partial u(x_0)$ which proves that the generalized gradient of u at x_0 is made up by the differentials of the C^1 supertangents to u at x_0. The outcome of the previous discussion is summarized in the following statement.

Proposition 12 *Let u be a strongly semiconcave (semiconvex) function. For any x, $p \in \partial u(x)$ if and only if it is the differential of a C^1 supertangent (subtangent) to u at x. Consequently, the notions of a.e. solution to (1) and viscosity solution of the first (second) type coincide for this class of functions.*

In Sect. "Main Regularity Result" a weaker notion of semiconcavity will be introduced, obtained by requiring a milder version of (17), which will be crucial to proving the existence of C^1 subsolutions to (1).

Even if the previous analysis, and in particular the part about the equivalent notions of subsolutions, is only valid for (1), one can define a notion of viscosity solution for a wider class of Hamilton–Jacobi equations than (1), and even for some second-order equations. To be more precise, given a Hamilton–Jacobi equation $G(x, u, , Du) = 0$, with G nondecreasing with respect to the second argument, a continuous function u is called the viscosity solution of it, if for any x_0, any C^1 supertangent (subtangent) ψ to u at x_0 the inequality

$$G(x_0, u(x_0), D\psi(x_0)) \leq (\geq) 0$$

holds true.

Loosely speaking the existence of comparison principles in this context is related to the strict monotonicity properties of the Hamiltonian with respect to u or the presence in the equation of the time derivative of the unknown. For instance such principles hold for (2), see Sect. "Long-Time Behavior of Solutions to the Time-Dependent Equation".

Obtaining uniqueness properties for viscosity solutions to (1) is a more delicate matter. Such properties are actually related to the existence of strict subsolutions, since this, in turn, allows one to slightly perturb any solution obtaining a strict subsolution. To exemplify this issue, Proposition 10 is exploited to show:

Proposition 13 *Let Ω, g be an open-bounded subset of M and a continuous function defined on $\partial\Omega$, respectively. Assume that there is a strict subsolution φ to (1) in Ω. Then there is at most one viscosity solution to (1) in Ω taking the datum g on the boundary.*

Assume by contradiction the existence of two viscosity solutions u and v with $v > u + \varepsilon$ at some point of Ω, where ε is a positive constant. The function $v_\lambda := \lambda\varphi + (1-\lambda)v$ is a strict subsolution to (1) for any $\lambda \in]0, 1]$, by the convexity assumption on H. Further, λ can be taken so small that the points of Ω for which $v_\lambda > u + \frac{\varepsilon}{2}$ make up a nonempty set, say Ω', are compactly contained in Ω. This goes against Proposition 10, because u and $v_\lambda + \frac{\varepsilon}{2}$ agree on $\partial\Omega'$ and the strict subsolution $v_\lambda + \frac{\varepsilon}{2}$ exceeds u in Ω'.

First Regularity Results for Subsolutions

A natural question is when does a classical subsolution to (1) exist? The surprising answer is that it happens whenever there is a (locally Lipschitz) subsolution, provided the assumptions on H introduced in the previous section are strengthened a little. Furthermore, any subsolution can be approximated by regular subsolutions in the topology of locally uniform convergence.

This theorem is postponed to Sect. "Main Regularity Result". Some preliminary results of regularity for subsolution, holding under the assumptions (3)–(6), are presented below.

Firstly the discussion concerns the existence of subsolutions to (1) that are regular, say C^1, at least on some distinguished subset of M. More precisely an attempt is made to determine when such functions can be obtained by mollification of subsolutions. The essential output is that this smoothing technique works if the subsolution one starts from is strict, so that, loosely speaking, some room is left to perturb it locally, still obtaining a subsolution. A similar argument has already been used in the proof of Proposition 10.

In this analysis it is relevant the critical level of the Hamiltonian which is defined as the one for which the corresponding Hamilton–Jacobi equation possesses a subsolution, but none of them are strict on the whole ground space. It is also important subset of M, named after Aubry and indicated by $\mathcal{A}$, made up of points around which no critical subsolution (i. e. subsolution to (1) with $a = c$) is strict.

According to what was previously outlined, to smooth up a critical subsolution around the Aubry set, seems particulary hard, if not hopeless. This difficulty will be overcome by performing a detailed qualitative study of the behavior of critical subsolutions on $\mathcal{A}$.

The simple setting to be first examined is when there is a strict subsolution, say u, to (1) satisfying

$$H(x, Du(x)) \leq a - \varepsilon \quad \text{for a.e. } x \in M, \text{ and some } \varepsilon > 0\,, \tag{18}$$

and, in addition, H is uniformly continuous on $M \times B \subset T^*M$, whenever B is a bounded subset of $\mathbb{R}^N$. In this case the mollification procedure plainly works to supply a regular subsolution. The argument of Proposition 10 can be adapted to show this. Define, for any x, any $\delta > 0$

$$u_\delta(x) = \int \zeta_\delta(y - x)u(y)\,\mathrm{d}y\,,$$

where ζ_δ is a standard C^∞ mollifier supported in $B(0, \delta)$, and by using the convex character of the Hamiltonian and Jensen Lemma, get

$$H(x, Du_\delta(x)) \leq \int \zeta_\delta(y - x)H(x, Du(y))\,\mathrm{d}y\,,$$

so that if $o(\cdot)$ is a continuity modulus of H in $M \times B(0, r)$, with r denoting a Lipschitz constant for u, a δ can be selected in such a way that $o(\delta) \leq \frac{\varepsilon}{2}$, and consequently u_δ is the desired smooth subsolution, and is, in addition, still strict on M.

Even if condition (18) does not hold on the whole underlying space, the previous argument can be applied locally, to provide a smoothing of any subsolution u, at least in the open subset W_u where it is strict (see (14) for the definition of this set), by introducing countable open coverings and associated C^∞ partition of the unity. The uniform continuity assumption on H as well as the global Lipschitz-continuity of u can be bypassed as well. It can be proved:

Proposition 14 *Given $u \in S_a$ with W_u nonempty, there exists $v \in S_a$, which is strict and of class C^∞ on W_u.*

Note that the function v appearing in the statement is required to be a subsolution on the whole M. In the proof of Proposition 14 an extension principle for subsolutions will be used that will be explained later.

Extension principle *Let v and C be a subsolution to (1) and a closed subset of M, respectively. Any continuous extension of $v|_C$ which is a subsolution on $M \setminus C$ is also a subsolution in the whole M.*

The argument for showing Proposition 14 will also provide the proof, with some adjustments, of the main regularity result, i. e. Theorem 35 in Sect. "Main Regularity Result".

By the very definition of W_u an open neighborhood U'_x, compactly contained in W_u, can be found, for all $x \in W_u$, in such a way that

$$H(y, Du(y)) < a - \varepsilon_x \quad \text{for a.e. } y \in U'_x \text{ and some } \varepsilon_x > 0. \tag{19}$$

Through regularization of u by means of a C^∞ mollifier ζ_δ supported in $B(0, \delta)$, for $\delta > 0$ suitably small, a smooth function can be then constructed still satisfying (19) in a neighborhood of x slightly smaller than U'_x, say U_x. The next step is to extract from $\{U_x\}$, $x \in W_u$, a countable locally finite cover of W_u, say $\{U_{x_i}\}$, $i \in \mathbb{N}$. In the sequel the notations U_i, ε_i are adopted in place of U_{x_i}, ε_{x_i}, respectively. The regularized function is denoted by u_i.

Note that such functions are not, in general, subsolutions to (1) on M, since their behavior outside U_i cannot be controlled. To overcome this difficulty a C^∞ partition of the unity β_i subordinated to U_i is introduced.

The crucial point here is that the mollification parameters, denoted by δ_i, can be adjusted in such a way that the uniform distance $|u - u_i|_{\infty, U_i}$ is as small as desired. This quantity, more precisely, is required to be small with respect to $\frac{1}{|D\beta_i|_\infty}$, $\frac{1}{2^i}$ and the ε_j corresponding to indices j such that $U_j \cap U_i \neq \emptyset$. In place of $\frac{1}{2^i}$ one could take the terms of any positive convergent series with sum 1. Define v via the formula

$$v = \begin{cases} \sum \beta_i u_i & \text{in } W_u \\ u & \text{otherwise}\,. \end{cases} \tag{20}$$

Note that a finite number of terms are involved in the sum defining v in W_u since the cover $\{U_i\}$ is locally finite. It can be surprising at first sight that the quantity $\sum \beta_i u_i$, with u_i subsolution in U_i and β_i supported in U_i, represents a subsolution to (1), since, by differentiating, one gets

$$D\left(\sum \beta_i u_i\right) = \sum \beta_i Du_i + \sum D\beta_i u_i \,,$$

and the latter term does not seem easy to handle. The trick is to express it through the formula

$$\sum D\beta_i u_i = \sum D\beta_i (u_i - u) \,, \tag{21}$$

which holds true because $\sum \beta_i \equiv 1$, by the very definition of partition of unity, and so $\sum D\beta_i \equiv 0$. From (21) deduce

$$\left|\sum D\beta_i u_i\right| \le \sum |D\beta_i|_\infty \, |u - u_i|_{\infty, U_i} \,,$$

and consequently, recalling that $|u - u_i|_{\infty,U_i}$ is small with respect to $\frac{1}{|D\beta_i|_\infty}$

$$D\left(\sum \beta_i u_i\right) \sim \sum \beta_i Du_i \,.$$

Since the Hamiltonian is convex in p, β_i is supported in U_i and u_i is a strict subsolution to (1) in U_i, we finally discover that v, defined by (20), is a strict subsolution in W_u.

Taking into account the extension principle for subsolutions, it is left to show, for proving Proposition 14, that v is continuous. For this, first observe that for any $n \in \mathbb{N}$ the set

$$\overline{\cup_{i \le n} U_i}$$

is compact and disjoint from ∂W_u, and consequently

$$\min\{i : x \in U_i\} \to +\infty \qquad \text{when } x \in W_u \text{ approaches } \partial W_u \,.$$

This, in turn, implies, since $|u - u_i|_{\infty,U_i}$ is small compared to $\frac{1}{2^i}$,

$$\begin{aligned}\left|\sum \beta_i(x) u_i(x) - u(x)\right| &\le \sum_{\{i:\, x \in U_i\}} \beta_i(x) |u - u_i|_{\infty,U_i} \\ &\le \sum_{\{i:\, x \in U_i\}} \beta_i(x) \frac{1}{2^i} \to 0 \,,\end{aligned}$$

whenever $x \in W_u$ approaches ∂W_u. This shows the assertion.

The next step is to look for subsolutions that are strict in a subset of M as large as possible. In particular a strict subsolution to (1) on the whole M does apparently exist at any level a of the Hamiltonian with

$$a > \inf\{b : H(x, Du) = b \text{ has a subsolution}\} \,. \tag{22}$$

The infimum on the right-hand side of the previous formula is the critical value of H; it will be denoted from now on by c. Accordingly the values $a > c$ (resp. $a < c$) will be qualified as *supercritical* (resp. *subcritical*). The inf in (22) is actually a minimum in view of Proposition 5. By the coercivity properties of H, the quantity $\min_p H(x, p)$ is finite for any x, and clearly

$$c \ge \sup_x \min_p H(x, p) \,,$$

which shows that $c > -\infty$, but it can be equal to $+\infty$ if M is noncompact. In this case no subsolutions to (1) should exist for any a. In what follows it is assumed that c is finite. Note that the critical value for the conjugate Hamiltonian $\check{H}$ does not change, since, as already noticed in Sect. "Subsolutions", the family of subsolutions of the two corresponding Hamilton–Jacobi equations are equal up to a change of sign.

From Proposition 14 can be derived:

Theorem 15 *There exists a smooth strict subsolution to (1) for any supercritical value a.*

Critical Equation and Aubry Set

Here the attention is focused on the critical equation

$$H(x, Du) = c \,. \tag{23}$$

A significant progress in the analysis is achieved by showing that there is a critical subsolution v with W_v, see (14) for the definition, enjoying a maximality property. More precisely the following statement holds:

Proposition 16 *There exists $v \in S_c$ with*

$$W_v = W_0 := \bigcup \{W_u : \ u \text{ is a critical subsolution}\} \,.$$

This result, combined with Proposition 14, gives the

Proposition 17 *There exists a subsolution to (23) that is strict and of class C^∞ on W_0.*

To construct v appearing in the statement of Proposition 16 a covering technique to W_0, as in Proposition 14, is applied and then the convex character of H is exploited. Since no regularity issues are involved, there is no need to introduce smoothing procedures and partitions of unity, so the argument is altogether quite simple.

Any point $y \in W_0$ possesses a neighborhood U_y where some critical subsolution v_y satisfies

$$H(x, Dv_y(x) \le c - \varepsilon_y \quad \text{for a.e. } x \in U_y, \text{ some positive } \varepsilon_y .$$

A locally finite countable subcover $\{U_{y_i}\}$, $i \in \mathbb{N}$, can be extracted, the notations U_i, v_i, ε_i are used in place of U_{y_i}, v_{y_i}, ε_{y_i}. The function v is defined as an infinite convex combination of u_i, more precisely $v = \sum \frac{1}{2^i} v_i$.

To show that v has the properties asserted in the statement, note that the functions v_i are locally equiLipschitz-continuous, being critical subsolutions, and can be taken, in addition, locally equibounded, up to addition of a constant. The series $\sum \lambda_i u_i$, $\sum \lambda_i Du_i$ are therefore locally uniformly convergent by the Weierstrass M-test. This shows that the function v is well defined and is Lipschitz-continuous, in addition

$$Dv(x) = \sum \lambda_i Du_i(x) \quad \text{for a.e. } x .$$

If x belongs to the full measure set where v and all the v_i are differentiable, one finds, by exploiting the convex character of H

$$H\left(x, \sum_{i \le n} \lambda_i Du_i(x)\right) \le \sum_{i \le n} \lambda_i H(x, Du_i(x)) + \left(1 - \sum_{i \le n} \lambda_i\right) H(x, 0) ,$$

for any fixed n. This implies, passing to the limit for $n \to +\infty$

$$H(x, Dv(x)) = H\left(x, \sum_{i=1}^{\infty} \lambda_i Du_i(x)\right) \le \sum_{i=1}^{\infty} \lambda_i H(x, Du_i(x)) .$$

The function v is thus a critical subsolution, and, in addition, one has

$$H(x, Dv(x)) \le \sum_{i \ne j} H(x, Dv_i(x)) + \lambda_j H(x, Dv_j(x)) \le c - \lambda_j \varepsilon_j , \tag{24}$$

for any j and a.e. $x \in U_j$. This yields Proposition 16 since $\{U_j\}$, $j \in \mathbb{N}$, is a locally finite open cover of W_0, and so it comes from (24) that the essential sup of v, on any open set compactly contained in W_0, is strictly less that c.

The Aubry set $\mathcal{A}$ is defined as $M \setminus W_0$. According to Propositions 14, 16 it is made up by the *bad* points around which no function of S_c can be regularized through mollification still remaining a critical subsolution. The points of $\mathcal{A}$ are actually characterized by the fact that no critical subsolution is strict around them. Note that a local as well as a global aspect is involved in such a property, for the subsolutions under investigation must be subsolutions on the whole space. Note further that $\mathcal{A}$ is also the Aubry set for the conjugate critical equation with Hamiltonian $\check{H}$.

A qualitative analysis of $\mathcal{A}$ is the main subject of what follows. Notice that the Aubry set must be nonempty if M is compact, since, otherwise, one could repeat the argument used for the proof of Proposition 16 to get a *finite* open cover $\{U_i\}$ of M and a *finite* family u_i of critical subsolutions satisfying

$$H(x, Du_i(x) \le c - \varepsilon_i \quad \text{for a.e. } x \in U_i \text{ and some } \varepsilon_i > 0,$$

and to have for a *finite* convex combination $u = \sum_i \lambda_i u_i$

$$H(x, Du(x)) \le c - \min_i \{\lambda_i \varepsilon_i\} ,$$

in contrast with the very definition of critical value. If, on the contrary, M is noncompact Hamiltonian such that the corresponding Aubry set is empty, then it can be easily exhibited.

One example is given by $H(x, p) = |p| - f(x)$, in the case where the potential f has no minimizers. It is easily seen that the critical level is given by $-\inf_M f$, since, for a less than this value, the sublevels $Z_a(\cdot)$ are empty at some $x \in M$ and consequently the corresponding Hamilton–Jacobi equation does not have any subsolution; on the other side

$$H(x, 0) = -f(x) < -\inf_M f ,$$

which shows that any constant function is a strict critical subsolution on M. This, in turn, implies the emptiness of $\mathcal{A}$.

In view of Proposition 14, one has:

Proposition 18 *Assume that M is noncompact and the Aubry set is empty, then there exists a smooth strict critical subsolution.*

The points y of $\mathcal{A}$ are divided into two categories according to whether the sublevel $Z_c(y)$ has an empty or nonempty interior. It is clear that a point y with $\operatorname{int} Z_c(y) = \emptyset$ must belong to $\mathcal{A}$ because for such a point

$$H(y, p) = c \quad \text{for all } p \in Z_c(y) , \tag{25}$$

and, since any critical subsolution u must satisfy $\partial u(y) \subset Z_c(y)$, it cannot be strict around y. These points

are called equilibria, and $\mathcal{E}$ indicates the set of all equilibria. The reason for this terminology is that if the regularity assumptions on H are enough to write the Hamilton's equations on T^*M, then (y, p_0) is an equilibrium of the related flow with $H(y, p_0) = c$ if and only if $y \in \mathcal{E}$ and $Z_c(y) = \{p_0\}$. This point of view will not be developed further herein. From now on the subscript c will be omitted to ease notations.

Next the behavior of viscosity test functions of any critical subsolution at points belonging to the Aubry set is investigated. The following assertion holds true:

Proposition 19 *Let u, y, ψ be a critical subsolution, a point of the Aubry set and a viscosity test function to u at y, respectively. Then $H(y, D\psi(y)) = c$.*

Note that the content of the proposition is an immediate consequence of (25) if, in addition, y is an equilibrium. In the general case it is not restrictive to prove the statement when ψ is a strict subtangent. Actually, if the inequality $H(y, D\psi(y)) < c$ takes place, a contradiction is reached by constructing a subsolution v strict around y by means of the push-up argument introduced in Sect. "Subsolutions" for proving Proposition 7.

By using the previous proposition, the issue of the existence of (viscosity) solutions to (23) or, more generally, to (1) can be tackled. The starting idea is to fix a point y in M, to consider the family

$$\tilde{S}_a^y = \{u \in S_a : u(y) = 0\}\,, \tag{26}$$

and to define

$$w_a^y(x) = \sup_{\tilde{S}_a^y} u(x)\,, \tag{27}$$

Since $\tilde{S}_a^y$ is complete (this terminology was introduced in Sect. "Solutions") at any $x \neq y$, the function w_a^y is a subsolution to (1) on M, and a viscosity solution to $M \setminus \{y\}$, by Propositions 6, 7.

If $a = c$, and the point y belongs to $\mathcal{A}$ then, in view of Proposition 19, w_c^y is a critical solution on the whole M. On the contrary, the fact that $y \notin \mathcal{A}$ i. e. $y \in W_0$, prevents this function from being a global solution. In fact in this case, according to Propositions 14, 16, there is a critical subsolution φ, which is smooth and strict around y, and it can be also assumed, without any loss of generality, to vanish at y. Therefore, φ is subtangent to w_c^y at y for the maximality property of w_c^y and $H(y, D\varphi(y)) < c$. A characterization of the Aubry set then follows:

First characterization of $\mathcal{A}$ *A point y belongs to the Aubry set if and only if the function w_c^y, defined in (27) with $a = c$, is a critical solution on the whole M.*

If $\mathcal{A} \neq \emptyset$, which is true when M is compact, then the existence of a critical solution can be derived. Actually in the compact case the critical level is the unique one for which a viscosity solution to (1) does exist. If, in fact $a > c$, then, by Theorem 15, the equation possesses a smooth strict critical subsolution, say φ, which is subtangent to any other function f defined on M at the minimizers of $f - u$, which do exist since M is assumed to be compact. This rules out the possibility of having a solution of (1) since $H(x, D\varphi(x)) < a$ for any x.

Next is discussed the issue that in the noncompact case a solution does exist at the critical as well as at any supercritical level.

Let a be supercritical. The idea is to exploit the noncompact setting, and to throw away the points where the property of being a solution fails, by letting them go to infinity.

Let $w_n := w_a^{y_n}$ be a sequence of subsolutions given by (27), with $|y_n| \to +\infty$. The w_n are equiLipschitz-continuous, being subsolutions to (1), and locally equibounded, up to the addition of a constant. One then gets, using Ascoli Theorem and arguing along subsequences, a limit function w. Since the w_n are solutions around any fixed point, for n suitably large, then, in view of the stability properties of viscosity solutions, see Proposition 8, w is a solution to (1) around any point of M, which means that w is a viscosity solution on the whole M, as announced. The above outlined properties are summarized in the next statement.

Proposition 20

(i) If M is compact then a solution to (1) does exist if and only if $a = c$.
(ii) If M is noncompact then (1) can be solved in the viscosity sense if and only if $a \geq c$.

An Intrinsic Metric

Formula (27) gives rise to a nonsymmetric semidistance $S_a(\cdot, \cdot)$ by simply putting

$$S_a(y, x) = w_a^y(x)\,.$$

This metric viewpoint will allow us to attain a deeper insight into the structure of the subsolutions to (1) as well as of the geometric properties of the Aubry set.

It is clear that S_a satisfies the triangle inequality and $S_a(y, y) = 0$ for any y. It fails, in general to be symmetric and non-negative. It will be nevertheless called, from now on, distance to ease terminology.

An important point to be discussed is that S_a is a *length distance*, in the sense that a suitable length functional ℓ_a

can be introduced in the class of Lipschitz-continuous curves of M in such a way that, for any pairs of points x and y, $S_a(y,x)$ is the infimum of the lengths of curves joining them. Such a length, will be qualified from now on as *intrinsic* to distinguish it from the natural length on the ground space, denoted by ℓ. It only depends on the corresponding sublevels of the Hamiltonian. More precisely one defines for a (Lipschitz-continuous) curve ξ parametrized in an interval I

$$\ell_a(\xi) = \int_I \sigma_a(\xi, \dot{\xi})\, \mathrm{d}t\,, \tag{28}$$

where σ_a stands for the support function of the a-sublevel of H. More precisely, the function σ_a is defined, for any $(x,q) \in TM$ as

$$\sigma_a(x,q) = \max\{p\,q \,:\, p \in Z_a(x)\}\,,$$

it is accordingly convex and positively homogeneous in p, upper semicontinuous in x, and, in addition, continuous at any point possessing a sublevel with a nonempty interior. The positive homogeneity property implies that the line integral in (28) is invariant under change of parameter preserving the orientation. The intrinsic length ℓ_a is moreover lower semicontinuous for the uniform convergence of a equiLipschitz-continuous sequence of curves, by standard variational argument, see [7]. Let $\overline{S}_a$ denote the length distance associated to ℓ_a, namely

$$\overline{S}_a(y,x) = \inf\{\ell_a(\xi)\colon \quad \xi \text{ connects } y \text{ to } x\}\,.$$

The following result holds true:

Proposition 21 *$\overline{S}_a$ and S_a coincide.*

Note that by the coercivity of the Hamiltonian $\ell_a(\xi) \le r\ell(\xi)$ for some positive r. Taking into account that the Euclidean segment is an admissible junction between any pair of points, deduce the inequality

$$|\overline{S}_a(y,x)| \le r|y-x| \quad \text{for any } y, x\,,$$

which, combined with the triangle inequality, implies that the function $x \mapsto \overline{S}_a(y,x)$ is locally Lipschitz-continuous, for any fixed y. Let y_0, x_0 be a pair of points in M. Since $u := S_a(y_0,\cdot)$ is locally Lipschitz-continuous, one has

$$S_a(y_0,x_0) = u(x_0) - u(y_0) = \int_I \frac{\mathrm{d}}{\mathrm{d}t} u(\xi(t)\,\mathrm{d}t$$

for any curve ξ connecting y_0 to x_0, defined in some interval I. It is well known from [8] that

$$\frac{\mathrm{d}}{\mathrm{d}t} u(\xi(t)) = p(t)\,\dot{\xi}(t) \qquad \text{for a.e. } t \in I, \text{ some } p(t) \in \partial u(\xi(t))\,,$$

and, since $\partial u(x) \subset Z_a(x)$ for any x, derive

$$S_a(y_0,x_0) \le \ell_a(\xi) \quad \text{for all every curve } \xi \text{ joining } y_0 \text{ to } x_0,$$

which, in turn, yields the inequality $S_a(y_0,x_0) \le \overline{S}_a(x_0, y_0)$. The converse inequality is obtained by showing that the function $w := \overline{S}_a(y_0,\cdot)$ is a subsolution to (1), see [20,29].

From now on the subscript from Z_a, S_a and σ_a will be omitted in the case where $a = c$.

It is clear that, in general, the intrinsic length of curves can have any sign. However, if the curve is a cycle such a length must be non-negative, according to Proposition 21, otherwise going several times through the same loop the identity $S_a \equiv -\infty$ would be obtained. This remark will have some relevance in what follows.

Proposition 21 allows us to determine the intrinsic metric related to the a-sublevel of the conjugate Hamiltonian $\check{H}$, denoted by $\check{Z}_a(\cdot)$. Since $\check{Z}_a(x) = -Z_a(x)$, for any x, because of the very definition of $\check{H}$, the corresponding support function $\check{\sigma}$ satisfies

$$\check{\sigma}_a(x,q) = \sigma_a(x,-q) \quad \text{for any } x, q\,.$$

Therefore, the intrinsic lengths ℓ_a and $\check{\ell}_a$ coincide up to a change of orientation. In fact, given ξ, defined in $[0,1]$, and denoted by $\gamma(s) = \xi(1-s)$ the curve with opposite orientation, one has

$$\check{\ell}_a(\xi) = \int_0^1 \sigma_a(\xi, -\dot{\xi})\,\mathrm{d}s\,,$$

and using $r = 1-t$ as a new integration variable one obtains

$$\int_0^1 \sigma_a(\xi,-\dot{\xi})\,\mathrm{d}s = \int_0^1 \sigma_a(\gamma,\dot{\gamma})\,\mathrm{d}r = \ell_a(\gamma).$$

This yields

$$\check{S}_a(x,y) = S_a(y,x) \quad \text{for any } x, y\,,$$

where $\check{S}_a$ stands for the conjugate distance. The function $S_a(\cdot,y)$ is thus the pointwise supremum of the family

$$\{v\colon v \text{ is a subsolution to (16) and } v(y) = 0\}\,,$$

and accordingly $-S_a(\cdot,y)$ the pointwise infimum of $\{u \in S_a\colon u(y) = 0\}$. Summing up:

Proposition 22 *Given $u \in S_a$, $y \in M$, the functions $S_a(y,\cdot)$, $-S_a(\cdot,y)$ are supertangent and subtangent, respectively, to u at y.*

The Extension Principle for subsolutions can now be proved. Preliminarily the fifth characterization of the family of subsolutions is given.

Proposition 23 *A continuous function u is a subsolution to (1) if and only if*

$$u(x) - u(y) \leq S_a(y, x) \quad \textit{for any } x, y\,. \tag{29}$$

It is an immediate consequence that any subsolution satisfies the inequality in the statement. Conversely, let ψ be a C^1 subtangent to u at some point y. By the inequality (29) the subsolution $x \mapsto S_a(y, x)$ is supertangent to u at y, for any y. Therefore, ψ is subtangent to $x \mapsto S_a(y, x)$ at the same point, and so one has $H(y, D\psi(y)) \leq a$, which shows the assertion taking into account the third definition of subsolution given in Sect. "Subsolutions".

To prove the Extension Lemma, one has to show that a function w coincident with some subsolution to (1) on a closed set C, and being a subsolution on $M \setminus C$, is a subsolution on the whole M. The intrinsic length will play a main role here. Two facts are exploited:

(i) if a curve connects two points belonging to C then the corresponding variation of u is estimated from above by the its intrinsic length because of Proposition 23, and since u coincides with a subsolution to (1) on C,
(ii) the same estimate holds true for any pair of points if the curve joining them lies outside C in force of the property that u is a subsolution in $M \setminus C$.

Let ε be a positive constant, x, y a pair of points and ξ a curve joining them, whose intrinsic length approximates $S_a(y, x)$ up to ε. The interval of definition of ξ can be partitioned in such a way that the portion of the curve corresponding to each subinterval satisfies the setting of one of the previous items (i) and (ii). By exploiting the additivity of the intrinsic length one finds

$$u(x) - u(y) \leq \ell_a(\xi) \leq S_a(y, x) + \varepsilon\,,$$

and the conclusion is reached taking into account the characterization of a subsolution given by Proposition 23, and the fact that ε is arbitrary.

To carry on the analysis, it is in order to discuss an apparent contradiction regarding the Aubry set. Let $y_0 \in \mathcal{A} \setminus \mathcal{E}$ and $p_0 \in \operatorname{int} Z(y)$, then p_0 is also in the interior of the c-sublevels at points suitably close to y, say belonging to a neighborhood U of y, since Z is continuous at y. This implies

$$p\,(x - y_0) \leq \ell_c(\xi)\,, \tag{30}$$

for any $x \in U$, any curve ξ joining y to X and *lying in* U. However, this inequality does not imply by any means that the function $x \mapsto p(x - y)$ is subtangent to $x \mapsto S(y, x)$ at y. This, in fact, should go against Proposition 19, since $H(y, p_0) < c$.

The unique way to overcome the contradiction is to admit that, even for points very close to y, the critical distance from y is realized by the intrinsic lengths of curves going out of U. In this way one could not deduce from the inequality (30) the previously indicated subtangency property. This means that S is not localizable with respect to the natural distance, and the behavior of the Hamiltonian in points far from y in the Euclidean sense can affect it.

There thus exist a sequence of points x_n converging to y and a sequence of curves joining y to x_n with intrinsic length approximating $S(y, x_n)$ up to $\frac{1}{n}$ and going out U. By juxtaposition of ξ_n and the Euclidean segment from x_n to y, a sequence of cycles γ_n can be constructed based on y (i. e. passing through y) satisfying

$$\ell_c(\gamma_n) \to 0\,, \quad \inf_n \ell(\gamma_n) > 0\,. \tag{31}$$

This is a threshold situation, since the critical length of any cycle must be non-negative. Next it is shown that (31) is indeed a metric characterization of the Aubry set.

Metric characterization of the Aubry set *A point y belongs to $\mathcal{A}$ if and only if there is a sequence γ_n of cycles based on y and satisfying* (31).

What remains is to prove that the condition (31) holds at any equilibrium point and, conversely, that if it is true at some y, then such a point belongs to $\mathcal{A}$.

If $y \in \mathcal{E}$ then this can be directly proved exploiting that $\operatorname{int} Z(y)$ is empty and, consequently the sublevel, being convex, is contained in the orthogonal of some element, see [20].

Conversely, let $y \in M \setminus \mathcal{A}$, according to Proposition 17, there is a critical subsolution u which is of class C^∞ and strict in a neighborhood U of y. One can therefore find a positive constant δ such that

$$Du(x)\,q \leq \sigma(x, q) - \delta \quad \text{for any } x \in U, \text{ any unit vector } q. \tag{32}$$

Let now ξ be a cycle based on y and parametrized by the Euclidean arc-length in $[0, \ell(\xi)]$, then $\xi(t) \in U$, for t belonging to an interval that can be assumed without loss of generality of the form $[0, t_1]$ for some $t_1 \geq \operatorname{dist}(y, \partial U)$ (where dist indicates the Euclidean distance of a point from a set). This implies, taking into account (32) and that ξ is a cycle

$$\begin{aligned}\ell_c(\xi) &= \ell_c\left(\xi\big|_{[0,t_1]}\right) + \ell_c\left(\xi\big|_{[t_1,T]}\right)\\ &\geq (u(\xi(t_1) - u(\xi(0)) + \delta t_1 + (u(\xi(T) - u(\xi(t_1))\\ &\geq \delta \operatorname{dist}(y, \partial U)))\,.\end{aligned}$$

This shows that the condition (31) cannot hold for sequences of cycles passing through y. By slightly adapting the previous argument, a further property of the intrinsic critical length, to be used later in Sect. "Long-Time Behavior of Solutions to the Time-Dependent Equation", can be deduced.

Proposition 24 *Let M be compact. Given $\delta > 0$, there are two positive constants α, β such that any curve ξ lying at a distance greater than δ from $\mathcal{A}$ satisfies*

$$\ell_c(\xi) \geq -\alpha + \beta\ell(\xi)\,.$$

An important property of the Aubry set is that it is a *uniqueness set* for the critical equation, at least when the ground space is compact. This means that two critical solutions coinciding on $\mathcal{A}$ must coincide on M. More precisely it holds:

Proposition 25 *Let M be compact. Given an admissible trace g on $\mathcal{A}$, i. e. satisfying the compatibility condition*

$$g(y_2) - g(y_1) \leq S(y_1, y_2)\,,$$

the unique viscosity solution taking the value g on $\mathcal{A}$ is given by

$$\min\{g(y) + S(y,\cdot)\colon y \in \mathcal{A}\}\,.$$

The representation formula yields indeed a critical solution thanks to the first characterization of the Aubry set and Proposition 9. The uniqueness can be obtained taking into account that there is a critical subsolution which is strict and C^∞ in the complement of $\mathcal{A}$ (see Proposition 17), and arguing as in Proposition 13.

Some information on the Aubry set in the one-dimensional case can be deduced from both the characterizations of $\mathcal{A}$

Proposition 26 *Assume M to have dimension 1, then*

(i) if M is compact then either $\mathcal{A} = \mathcal{E}$ or $\mathcal{A} = M$,
(ii) if M is noncompact then $\mathcal{A} = \mathcal{E}$, and, in particular, $\mathcal{A} = \emptyset$ if $\mathcal{E} = \emptyset$.

In the one-dimensional case the c-sublevels are compact intervals. Set $Z(x) = [\alpha(x), \beta(x)]$ with α, β continuous, and consider the Hamiltonian

$$\overline{H}(x,p) = H(x, p - \alpha(x))\,.$$

It is apparent that u is a critical (sub)solution for H if and only if $u + F$, where F is any antiderivative of α, is a (sub)solution to $\overline{H} = c$, and in addition, u is strict as a subsolution in some $\Omega \subset M$ if and only if $u + F$ is strict in the same subset. This proves that c is also the critical value for $\overline{H}$.

Further, $u \in \tilde{S}_c^y$ for some y, with $\tilde{S}_c^y$ defined as in (26), if and only if $u + F_0$, where F_0 is the antiderivative of α vanishing at y, is in the corresponding family of subsolutions to $\overline{H} = c$. Bearing in mind the first characterization of $\mathcal{A}$, it comes that the Aubry sets of the two Hamiltonians H and $\overline{H}$ coincide.

The advantage of using $\overline{H}$ is that the corresponding critical sublevels $\overline{Z}(x)$ equal $[0, \beta(x) - \alpha(x)]$, for any x, and accordingly the support function, denoted by $\overline{\sigma}$, satisfies

$$\overline{\sigma}(x,q) = \begin{cases} q\,(\beta(x) - \alpha(x)) & \text{if } q > 0 \\ 0 & \text{if } q \leq 0 \end{cases}$$

for any x, q. This implies that the intrinsic critical length related to $\overline{H}$, say $\overline{\ell}_c$, is non-negative for all curves. Now, assume M to be noncompact and take $y \notin \mathcal{E}$, the claim is that $y \notin \mathcal{A}$. In fact, let $\varepsilon > 0$ be such that

$$m := \inf\{\beta(x) - \alpha(x)\colon x \in I_\varepsilon :=]y - \varepsilon, y + \varepsilon[\,\} > 0\,,$$

given a cycle ξ based on y, there are two possibilities: either ξ intersects ∂I_ε or is entirely contained in I_ε. In the first case

$$\overline{\ell}_c(\xi) \geq m\,\varepsilon\,, \tag{33}$$

in the second case ξ can be assumed, without loosing generality, to be parametrized by the Euclidean arc-length; since it is a cycle one has

$$\int_0^{\ell(\xi)} \dot{\xi}\,\mathrm{d}s = 0\,,$$

so that $\dot{\xi}(t) = 1$ for t belonging to a set of one-dimensional measure $\frac{\ell(\xi)}{2}$. One therefore has

$$\overline{\ell}_c(\xi) \geq m\,\frac{\ell(\xi)}{2}\,. \tag{34}$$

Inequalities (33), (34) show that $\overline{\ell}_c(\xi)$ cannot be infinitesimal unless $\ell(\xi)$ is infinitesimal. Hence item (ii) of Proposition 26 is proved. The rest of the assertion is obtained by suitably adapting the previous argument.

Dynamical Properties of the Aubry Set

In this section the convexity and the coercivity assumptions on H are strengthened and it is required, in addition

to (3),

$$H \text{ is convex in } p \tag{35}$$

$$\lim_{|p|\to+\infty} \frac{H(x,p)}{|p|} = +\infty \quad \text{uniformly in } x. \tag{36}$$

The *Lagrangian* L can be therefore defined through the formula

$$L(x,q) = \max\{p\,q - H(x,p)\colon p \in \mathbb{R}^N\}\,.$$

A curve γ, defined in some interval I, is said to be *critical* provided that

$$S(\gamma(t_1),\gamma(t_2)) = \int_{t_1}^{t_2} L(\gamma,\dot\gamma) + c\,\mathrm{d}s = -S(\gamma(t_2),\gamma(t_1))\,, \tag{37}$$

for any t_1, t_2 in I. It comes from the metric characterization of $\mathcal{A}$, given in the previous section, that any critical curve is contained in $\mathcal{A}$. In fact if such a curve is supported on a point, say x_0, then $L(x_0,0) = -c$ and, consequently, the critical value is the minimum of $p \mapsto H(x,p)$. This, in turn, implies, in view of (35), that the sublevel $Z(x_0)$ has an empty interior so that $x_0 \in \mathcal{E} \subset \mathcal{A}$.

If, on the contrary, a critical curve is nonconstant and x_1, x_2 are a pair of different points lying in its support, then $S(x_1,x_2) + S(x_2,x_1) = 0$ and there exist two sequences of curves, ξ_n and η_n whose intrinsic length approximates the $S(x_1,x_2)$ and $S(x_2,x_1)$, respectively. Hence the trajectories obtained through juxtaposition of ξ_n and η_n are cycles with critical length infinitesimal and natural length estimated from below by a positive constant, since they contain x_1 and x_2, with $x_1 \neq x_2$. This at last implies that such points, and so the whole support of the critical curve, are contained in the Aubry set.

Next the feature of the parametrization of a critical curve γ is investigated, since it apparently matters for a curve to be critical. For this purpose let x_0, $q_0 \neq 0$, and $p_0 \in Z(x_0)$ with $\sigma(x_0,q_0) = p_0\,q_0$, it comes from the definition of the Lagrangian

$$\sigma(x_0,q_0) - c = p_0\,q_0 - H(x_0,p_0) \le L(x_0,q_0)\,,$$

by combining this formula with (37), and recalling the relationship between intrinsic length and distance, one gets

$$L(\gamma,\dot\gamma) + c = \sigma(\gamma,\dot\gamma) \quad \text{for a.e. } t\,. \tag{38}$$

A parametrization is called *Lagrangian* if it satisfies the above equality. As a matter of fact it is possible to prove that any curve η, which stays far from $\mathcal{E}$, can be endowed with such a parametrization, see [10].

A relevant result to be discussed next is that the Aubry set is fully covered by critical curves. This property allows us to obtain precious information on the behavior of critical subsolution on $\mathcal{A}$, and will be exploited in the next sections in the study of long-time behavior of solutions to (2). More precisely the following result can be shown:

Theorem 27 *Given $y_0 \in \mathcal{A}$, there is a critical curve, defined in $\mathbb{R}$, taking the value y_0 at $t = 0$.*

If $y_0 \in \mathcal{E}$ then the constant curve $\xi(t) \equiv y_0$ is critical, as pointed out above. It can therefore be assumed $y_0 \notin \mathcal{E}$. It is first shown that a critical curve taking the value y_0 at 0, and defined in a bounded interval can be constructed.

For this purpose start from a sequence of cycles ξ_n, based on y_0, satisfying the properties involved in the metric characterization of $\mathcal{A}$, and parametrized by (Euclidean) arc length in $[0, T_n]$, with $T_n = \ell(\xi_n)$. By exploiting Ascoli Theorem, and arguing along subsequences, one obtains a uniform limit curve ξ of the ξ_n, with $\xi(0) = y_0$, in an interval $[0, T]$, where T is strictly less than $\inf_n T_n$. It is moreover possible to show that ξ is nonconstant.

A new sequence of cycles γ_n can be defined through juxtaposition of ξ, the Euclidean segment joining $\xi(T)$ to $\xi_n(T)$ and $\xi_n\big|_{[T,T_n]}$. By the lower semicontinuity of the intrinsic length, the fact that $\xi_n(T)$ converges to $\xi(T)$, and consequently the segment between them has infinitesimal critical length, one gets

$$\lim_n \ell_c(\gamma_n) = 0\,. \tag{39}$$

The important thing is that all the γ_n coincide with ξ in $[0, T]$, so that if $t_1 < t_2 < T$, $S(\xi(t_1),\xi(t_2))$ is estimated from above by

$$\ell_c\left(\xi\big|_{[t_1,t_2]}\right) = \ell_c\left(\gamma_n|_{[t_1,t_2]}\right)\,,$$

and $S(\xi(t_2),\xi(t_1))$ by the intrinsic length of the portion of γ_n joining $\xi(t_2)$ to $\xi(t_1)$. Taking into account (39) one gets

$$0 = \lim_n \ell_c(\gamma_n) \ge S(\xi(t_1),\xi(t_2)) + S(\xi(t_2),\xi(t_1))\,,$$

which yields the crucial identity

$$S(\xi(t_2),\xi(t_1)) = -S(\xi(t_1),\xi(t_2))\,. \tag{40}$$

In addition the previous two formulae imply that $\xi\big|_{[t_1,t_2]}$ is a minimal geodesic whose intrinsic length realizes $S(\xi(t_1,\xi(t_2))$, so that (40) can be completed as follows:

$$S(\xi(t_2),\xi(t_1)) = \int_{t_1}^{t_2} \sigma(\xi,\dot\xi)\,\mathrm{d}s = -S(\xi(t_1),\xi(t_2))\,. \tag{41}$$

Finally, ξ has a Lagrangian parametrization, up to a change of parameter, so that it is obtained in the end

$$\int_{t_1}^{t_2} \sigma(\xi, \dot{\xi})\, \mathrm{d}s = \int_{t_1}^{t_2} (L(\xi, \dot{\xi}) + c)\, \mathrm{d}s\,.$$

This shows that ξ is a critical curve. By applying Zorn lemma ξ can be extended to a critical curve defined in $\mathbb{R}$, which concludes the proof of Theorem 27.

As a consequence a first, perhaps surprising, result on the behavior of a critical subsolution on the Aubry set is obtained.

Proposition 28 *All critical subsolutions coincide on any critical curve, up to an additive constant.*

If u is such a subsolution and ξ any curve, one has

$$S(\xi(t_1), \xi(t_2)) \geq u(\xi(t_2)) - u(\xi(t_1)) \geq -S(\xi(t_2), \xi(t_1))\,,$$

by Proposition 22. If in addition ξ is critical then the previous formula holds with equality, which proves Proposition 28.

Next is presented a further result on the behavior of critical subsolutions on $\mathcal{A}$. From this it appears that, even in the broad setting presently under investigation (the Hamiltonian is supposed to be just continuous), such subsolutions enjoy some extra regularity properties on $\mathcal{A}$.

Proposition 29 *Let ξ be a critical curve, there is a negligible set E in $\mathbb{R}$ such that, for any critical subsolution u the function $u \circ \xi$ is differentiable whenever in $\mathbb{R} \setminus E$ and*

$$\frac{\mathrm{d}}{dt} u(\xi(t)) = \sigma(\xi(t), \dot{\xi}(t))\,.$$

More precisely E is the complement in $\mathbb{R}$ of the set of Lebesgue points of $\sigma(\xi, \dot{\xi})$ where, in addition, ξ is differentiable. E has a vanishing measure thanks to Rademacher and Lebesgue differentiability theorem. See [10] for a complete proof of the proposition.

The section ends with the statement of a result, that will be used for proving the forthcoming Theorem 34. The proof can be obtained by performing Lagrangian reparametrizations.

Proposition 30 *Let ξ be a curve defined in $[0, 1]$. Denote by Ξ the set of curves obtained through reparametrization of ξ in intervals with right endpoint 0, and for $\gamma \in \Xi$ indicate by $[0, T(\gamma)]$ its interval of definition. One has*

$$\ell_c(\xi) = \inf\left\{ \int_0^{T(\gamma)} (L(\gamma, \dot{\gamma}) + c)\; ds\colon \gamma \in \Xi \right\}\,.$$

Long-Time Behavior of Solutions to the Time-Dependent Equation

In this section it is assumed, in addition to (3), (36),

$$M \text{ is compact} \tag{42}$$

$$H \text{ is strictly convex in } p\,. \tag{43}$$

A solution of the time-dependent Eq. (2) is said to be *stationary* if it has the variable-separated form

$$u_0(x) - at\,, \tag{44}$$

for some constant a. Note that if ψ is a supertangent (subtangent) to u_0 at some point x_0, then $\psi - at$ is supertangent (subtangent) to $u_0 - at$ at (x_0, t) for any t, so that the inequality

$$-a + H(x_0, D\psi(x_0)) \leq (\geq)\ 0$$

holds true. Therefore, u_0 is a solution to (1) in M. Since such a solution does exist only when $a = c$, see Proposition 20, it is the case that in (44) u_0 is a critical solution and a is equal to c. The scope of this section is to show that any solution to the time-dependent equation uniformly converge to a stationary solution, as t goes to $+\infty$.

In our setting there is a comparison principle for (2), stating that two solutions v, w, issued from initial data v_0, w_0, with $v_0 \geq w_0$, satisfies $v \geq w$, from any $x \in M$, $t > 0$. In addition there exists a viscosity solution v, for any continuous initial datum v_0, which is, accordingly, unique, and is given by the Lax–Oleinik representation formula:

$$v(x, t) = \inf\left\{ v_0(\xi(0)) + \int_0^t L(\xi, \dot{\xi})\, \mathrm{d}s\colon \xi \text{ is a curve with } \xi(t) = x \right\}\,. \tag{45}$$

This shows that Eq. (2) enjoys the semigroup property, namely if w and v are two solutions with $w(\cdot, 0) = v(\cdot, t_0)$, for some $t_0 > 0$, then

$$w(x, t) = v(x, t_0 + t)\,.$$

It is clear that the solution of (2), taking a critical solution u_0 as initial datum, is stationary and is given by (44). For any continuous initial datum v_0 it can be found, since the underlying manifold is compact, a critical solution u_0 and a pair of constants $\alpha > \beta$ such that

$$u_0 + \alpha > v_0 > u_0 + \beta\,,$$

and consequently by the comparison principle for (2)

$$u_0 + \alpha > v(\cdot, t) + ct > u_0 + \beta \quad \text{for any } t\,.$$

This shows that the family of functions $x \mapsto v(x,t) + ct$, for $t \geq 0$, is equibounded. It can also be proved, see [10], that it is also equicontinuous, so that, by Ascoli Theorem, every sequence $v(\cdot, t_n) + ct_n$, for $t_n \to +\infty$, is uniformly convergent in M, up to extraction of a subsequence. The limits obtained in this way will be called *ω-limit* of $v + ct$. The first step of the analysis is to show:

Proposition 31 *Let v be a solution to (2). The pointwise supremum and infimum of the ω-limit of $v + ct$ are critical subsolutions.*

The trick is to introduce a parameter ε small and consider the functions

$$v^\varepsilon(x,t) = v(x, t/\varepsilon) \quad \text{for } \varepsilon > 0\,.$$

Arguing as above, it can be seen that the family $v^\varepsilon + ct$ is equibounded, moreover $v^\varepsilon + ct$ is apparently the solution to

$$\varepsilon w_t + H(x, Dw) = c\,,$$

for any ε. Hence, we exploit the stability properties that have been illustrated in Sects "Subsolutions", "Solutions", to prove the claim.

The following inequality will be used

$$L(x,q) + c \geq \sigma(x,q) \quad \text{for any } x, q,$$

which yields, by performing a line integration

$$\int_0^t (L(\gamma, \dot\gamma) + c)\, \mathrm{d}s \geq \ell_c(\gamma)\,, \tag{46}$$

for any $t > 0$, any curve γ defined in $[0, t]$, moreover, taking into account Lax–Oleinik formula and that M is assumed in this section to be compact

$$v(x,t) + c\,t \geq v_0(y) + S(y,x) \quad \text{for some } y \text{ depending on } x\,, \tag{47}$$

for a solution v of (2) taking a function v_0 as initial datum. If, in addition, v_0 is a critical subsolution, invoke Proposition 23 to derive from (47)

$$v(x,t) \geq v_0(x) - c\,t \quad \text{for any } x, t\,. \tag{48}$$

A crucial point to be exploited, is that, for such a v_0, the evolution induced by (2) on the Aubry set takes place on the critical curves. Given $t > 0$ and $x \in \mathcal{A}$, pick a critical curve ξ with $\xi(t) = x_0$, whose existence is guaranteed by Theorem 27, and then employ Proposition 29, about the behavior of the subsolution to (23) on critical curves, to obtain

$$v_0(x) - c\,t = v_0(\xi(0)) + \int_0^t L(\xi, \dot\xi)\, \mathrm{d}s \geq v(x,t)\,. \tag{49}$$

By combining (49) and (48), one finally has

$$v(x,t) = v_0(\xi(0)) + \int_0^t L(\xi, \dot\xi)\, \mathrm{d}s = v_0(x) - c\,t\,,$$

which actually shows the announced optimal character of critical curves with respect to the Lax–Oleinik formula, and, at the same time, the following

Proposition 32 *Let v be a solution to (2) taking a critical subsolution v_0 as initial datum at $t = 0$. Then*

$$v(x,t) = v_0(x) - c\,t \quad \textit{for any } x \in \mathcal{A}\,.$$

Summing up: stationary solutions are derived by taking as initial datum solutions to (23); more generally, solutions issued from a critical subsolution are stationary at least on the Aubry set. The next step is to examine the long-time behavior of such solutions on the whole M.

Proposition 33 *Let v be a solution to (2) taking a critical subsolution v_0 as initial datum at $t = 0$. One has*

$$\lim_{t \to +\infty} v(x,t) + c\,t = u_0(x) \quad \textit{uniformly in } x,$$

where u_0 is the critical solution with trace v_0 on $\mathcal{A}$.

The starting remark for getting the assertion is that, for any given x_0, an ε-optimal curve for $v(x_0, t_0)$, say ξ, must be close to $\mathcal{A}$ for some $t \in [0, t_0]$, provided ε is sufficiently small and t_0 large enough.

If in fact ξ stayed far from $\mathcal{A}$ for any $t \in [0, t_0]$ then $L(\xi(s), 0) + c$ could be estimated from below by a positive constant, since $\mathcal{E} \subset \mathcal{A}$, and the same should hold true, by continuity, for $L(\xi(s), q) + c$ if $|q|$ is small. One should then deduce that $\ell(\xi)$, and consequently (in view of Proposition 24) $\ell_c(\xi)$ were large. On the other side

$$u_0(x_0) \geq v(x_0, t_0) + c\,t_0 \geq v_0(\xi(0)) + \ell_c(\xi) - \varepsilon\,, \tag{50}$$

by (46) and the comparison principle for (2), which shows that the critical length of ξ is bounded from above, yielding a contradiction.

It can be therefore assumed that, up to a slight modification, the curve ξ intersects $\mathcal{A}$ at a time $s_0 \in [0, t_0]$ and satisfies

$$\begin{aligned} v(x_0, t_0) &\geq v_0(\xi(0)) + \int_0^{t_0} L(\xi, \dot\xi)\, \mathrm{d}s - \varepsilon \\ &\geq v(\xi(s_0), s_0)) + \int_{s_0}^{t_0} L(\xi, \dot\xi)\, \mathrm{d}s - \varepsilon\,. \end{aligned}$$

It is known from Proposition 32 that $v(\xi(s_0), s_0)) = v_0(\xi(s_0)) - c\, s_0$, so we have from the previous inequality, in view of (46)

$$\begin{aligned} v(x_0, t_0) &\geq v_0(\xi(s_0)) - c\, t_0 + \ell_c\left(\xi\big|_{[s_0,t_0]}\right) - \varepsilon \\ &\geq v_0(\xi(s_0)) - c\, t_0 + S(\xi(s_0), x_0) - \varepsilon\,. \end{aligned}$$

Bearing in mind the representation formula for u_0 given in Proposition 25, we obtain in the end

$$v(x_0, t_0) \geq u_0(x_0) - c\, t_0 - \varepsilon\,,$$

and conclude exploiting $u_0 \geq v_0$ and the comparison principle for (2).

The previous statement can be suitably generalized by removing the requirement of v_0 being a critical subsolution. One more precisely has:

Theorem 34 *Let v be a viscosity solution to (2) taking a continuous function v_0 as initial datum for $t = 0$, then*

$$\lim_{t\to+\infty} v(x,t) + c\, t = u_0(x) \quad \textit{uniformly in } x,$$

where u_0 is the critical solution given by the formula

$$u_0(x) = \inf_{y\in\mathcal{A}} \inf_{z\in M} \big(v_0(z) + S(z,y) + S(y,x)\big)\,. \tag{51}$$

The claim is that u_0, as defined in (51), is the critical solution with trace

$$w_0 := \inf_{z\in M} v_0(z) + S(z,\cdot) \tag{52}$$

on the Aubry set. This can indeed be deduced from the representation formula given in Proposition 25, once it is proved that w_0 is a critical subsolution. This property, in turn, comes from the characterization of critical subsolutions in terms of critical distance, presented in Proposition 23, the triangle inequality for S, and the inequalities

$$\begin{aligned} &w_0(x_1) - w_0(x_2) \\ &\quad \leq v_0(z_2) + S(z_2, x_1) - v_0(z_2) - S(z_2, x_2) \leq S(x_2, x_1)\,, \end{aligned}$$

which hold true if z_2 is a point realizing the infimum for $w_0(x_2)$. If, in particular, v_0 itself is a critical subsolution, then it coincides with w_0, so that, as announced, Theorem 34 includes Proposition 33.

In the general case, it is apparent that $w_0 \leq v_0$, moreover if $z \in M$ and $\overline{w}_0$ is a critical subsolution with $\overline{w}_0 \leq v_0$, one deduces from Proposition 23

$$\overline{w}_0(x) \leq v_0(z) + S(z,x) \quad \text{for any } z,$$

which tells that $\overline{w}_0 \leq w_0$, therefore w_0 is the maximal critical subsolution not exceeding v_0.

A complete proof of Theorem 34 is beyond the scope of this presentation. To give an idea, we consider the simplified case where the equilibria set $\mathcal{E}$ is a uniqueness set for the critical equation.

Given $x_0 \in \mathcal{E}$, $\varepsilon > 0$, take a z_0 realizing the infimum for $w_0(x_0)$, and a curve η, connecting z_0 to x_0, whose intrinsic length approximates $S(z_0, x_0)$ up to ε. By invoking Proposition 30 one deduces that, up to a change of the parameter, such a curve, defined in $[0, T]$, for some $T > 0$, satisfies

$$\int_0^T (L(\eta, \dot\eta) + c)\, \mathrm{d}s < \ell_c(\eta) + \varepsilon\,.$$

Therefore, taking into account the Lax–Oleinik formula, one discovers

$$\begin{aligned} w_0(x_0) &\geq v_0(z_0) + \int_0^T (L(\eta,\dot\eta) + c)\, \mathrm{d}s - 2\varepsilon \\ &\geq v(x_0, T) + cT - 2\varepsilon\,. \end{aligned} \tag{53}$$

Since $L(x_0, 0) + c = 0$, by the very definition of equilibrium, it can be further derived from Lax–Oleinik formula that $t \mapsto v(x_0, t)$ is nonincreasing, so that the inequality (53) still holds if T is replaced by every $t > T$. This, together with the fact that ε in (53) is taken arbitrarily, shows in the end, that any ω-limit ψ of $v + ct$ satisfies

$$w_0(x_0) \geq \psi(x_0) \quad \text{for any } x_0 \in \mathcal{E}\,. \tag{54}$$

On the other side, the initial datum v_0 is greater than or equal to w_0, and the solution to (2) with initial datum w_0, say w, has as unique ω-limit the critical solution u_0 with trace w_0 on $\mathcal{E}$ [recall that $\mathcal{E}$ is assumed to be a uniqueness set for (23)]. Consequently, by the comparison principle for (2), one obtains

$$u_0 \leq \psi \quad \text{in } M\,, \tag{55}$$

which, combined with (54), implies that ψ and w_0 coincide on $\mathcal{E}$. Further, by Proposition 31, the maximal and minimal ψ are critical subsolutions, and u_0 is the maximal critical subsolution taking the value w_0 on $\mathcal{E}$. This finally yields $\psi = u_0$ for any ψ, and proves the assertion of Theorem 34.

In the general case the property of the set of ω-limits of critical curves (i. e. limit points for $t \to +\infty$) of being a uniqueness set for the critical equation must be exploited. In this setting the strict convexity of H is essential, in [4,10] there are examples showing that Theorem 34 does not hold for H just convex in p.

Main Regularity Result

This section is devoted to the discussion of

Theorem 35 *If the Eq. (1) has a subsolution then it also admits a C^1 subsolution. Moreover, the C^1 subsolutions are dense in S_a with respect to the local uniform convergence.*

Just to sum up: it is known that a C^1 subsolution does exist when a is supercritical, see Theorem 15, and if the Aubry set is empty, see Proposition 18. So the case where $a = c$ and $\mathcal{A} \neq \emptyset$ is left. The starting point is the investigation of the regularity properties of *any* critical subsolution on $\mathcal{A}$.

For this, and for proving Theorem 35 conditions (43), (35) on H are assumed, and (3) is strengthened by requiring

$$H \text{ is locally Lipschitz-continuous in both arguments.} \tag{56}$$

This regularity condition seems unavoidable to show that the functions $S_a(y,\cdot)$ and $\check{S}_a(y,\cdot) = S_a(\cdot, y)$ enjoy a weak form of semiconcavity in $M \setminus \{y\}$, for all y, namely, if v is any function of this family, x_1, x_2 are different from y and $\lambda \in [0,1]$, then

$$v(\lambda x_1 + (1-\lambda)x_2) - \lambda v(x_1) - (1-\lambda)v(x_2)$$

can be estimated from below by a quantity of the same type as that appearing on the right-hand side of (17), with $|x_1 - x_2|^2$ replaced by a more general term which is still infinitesimal for $|x_1 - x_2| \to 0$. Of course the fundamental property of possessing C^1 supertangents at any point different from y and that the set made up by their differentials coincides with the generalized gradient is maintained in this setting.

To show the validity of this semiconcavity property, say for $S_a(y,\cdot)$, at some point x it is crucial that for any neighborhood U of x suitably small there are curves joining y to x and approximating $S_a(y,x)$ which stays in U for a (natural) length greater than a fixed constant depending on U. This is clearly true if $x \neq y$ and explains the reason why the initial point y has been excluded. However, if $a = c$ and $y \in \mathcal{A}$, exploiting the metric characterization of the Aubry set, it appears that this restriction on y can be removed, so that the following holds:

Proposition 36 *Let $y \in \mathcal{A}$, then the functions $S(y,\cdot)$ and $S(\cdot, y)$ are semiconcave (in the sense roughly explained above) on the whole M. This, in particular, implies that both functions possess C^1 supertangents at y and their differentials comprise the generalized gradient.*

From this the main regularity property of critical subsolutions on $\mathcal{A}$ are deduced. This reinforces the results given in Propositions 28, 29 under less stringent assumptions.

Theorem 37 *Every critical subsolution is differentiable on $\mathcal{A}$. All have the same differential, denoted by $p(y)$, at any point $y \in \mathcal{A}$, and*

$$H(y, p(y)) = c\,.$$

Furthermore, the function $y \mapsto p(y)$ is continuous on $\mathcal{A}$.

It is known from Proposition 22 that $S(y,\cdot)$, $-S(\cdot, y)$ are supertangent and subtangent, respectively, to every critical subsolution u at any y. If, in particular, $y \in \mathcal{A}$ then $S(y,\cdot)$ and $-S(\cdot, y)$ admit C^1 supertangents and subtangents, respectively, thanks to Proposition 36. This, in turn, implies that u is differentiable at y by Proposition 1 and, in addition, that all the differentials of supertangents to $S(y,\cdot)$ coincide with $Du(y)$, which shows that its generalized gradient reduces to a singleton and so $S(y,\cdot)$ is strictly differentiable at y by Proposition 36. If $p(y)$ denotes the differential of $S(y,\cdot)$ at y, then $Du(y) = p(y)$ for any critical subsolution and, in addition, $H(y, p(y)) = c$ by Proposition 19. Finally, the strict differentiability of $S(y,\cdot)$ at y gives that $p(\cdot)$ is continuous on $\mathcal{A}$.

The first application of Theorem 37 is relative to critical curves, and confirm their nature of generalized characteristics. If ξ is such a curve (contained in $\mathcal{A}$ by the results of Sect. "Dynamical Properties of the Aubry Set") and u a critical subsolution, it is known from Proposition 29 that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}u(\xi(t) &= p(\xi(t))\,\dot{\xi}(t) = \sigma(\xi(t), \dot{\xi}(t)) \\ &= L(\xi(t), \dot{\xi}(t)) + c\,,\end{aligned}$$

for a.e. $t \in \mathbb{R}$. Bearing in mind the definition of L, it is deduced that $p(\xi(t))$ is a maximizer of $p \mapsto p\,\dot{\xi}(t) - H(\xi(t), p)$, then by invoking (7) one obtains:

Proposition 38 *Any critical curve ξ satisfies the differential inclusion*

$$\dot{\xi} \in \partial_p H(\xi, p(\xi)) \quad \textit{for a.e. } t \in \mathbb{R}\,.$$

In the statement ∂_p denotes the generalized gradient with respect to the variable p.

The proof of Theorem 35 is now attacked. Combining Proposition 17 and Theorem 37, it is shown that there exists a critical subsolution, say w, differentiable at any point of M, strict and of class C^∞ outside the Aubry set, and with $Dw|_{\mathcal{A}} = p(\cdot)$ continuous. The problem is therefore to adjust the proof of Proposition 14 in order to have continuity of the differential on the whole M.

The first step is to show a stronger version of the Proposition 16 asserting that it is possible to find a critical subsolution u, which is not only strict on $W_0 = M \setminus \mathcal{A}$,

but also strictly differentiable on $\mathcal{A}$. Recall that this means that if $y \in \mathcal{A}$ and x_n are differentiability points of u with $x_n \to y$, then $Du(x_n) \to Du(y) = p(y)$. This implies that if $\overline{p}(\cdot)$ is a continuous extension of $p(\cdot)$ in M, then $\overline{p}(x)$ and $Du(x)$ are close at every differentiability point of u close to $\mathcal{A}$.

Starting from a subsolution u enjoying the previous property, the idea is then to use for defining the sought C^1 subsolution, say v, the same formula (20) given in Proposition 14, i. e.

$$v = \begin{cases} \sum \beta_i u_i & \text{in } W_0 \\ u & \text{in } \mathcal{A} \end{cases}$$

where β_i is a C^∞ partition of unity subordinated to a countable locally finite open covering U_i of W_0, and the u_i are obtained from u through suitable regularization in U_i by mollification. Look at the sketch of the proof of Proposition 16 for the precise properties of these objects. It must be shown

$$D\left(\sum \beta_i u_i(x_n)\right) \to Du(y) = p(y)\,,$$

for any sequence x_n of elements of W_0 converging to $y \in \mathcal{A}$, or equivalently

$$\left|\overline{p}(x_n) - D\left(\sum \beta_i u_i(x_n)\right)\right| \to 0\,.$$

One has

$$\left|\overline{p}(x_n) - D\left(\sum \beta_i u_i(x_n)\right)\right| \le \sum_i \beta_i(x_n)|Du_i(x_n) - \overline{p}(x_n)|D\beta_i(x_n)|u_i(x_n) - u(x_n)|\,.$$

The estimates given in the proof of Proposition 14 show that the second term of the right-hand side of the formula is small. To estimate the first term calculate

$$|Du_i(x_n) - \overline{p}(x_n)| \le \int \zeta_{\delta_i}(z - x_n)\big(|Du(z) - \overline{p}(z)| + |\overline{p}(z) - \overline{p}(x_n)|\big)\,\mathrm{d}z\,,$$

and observe first that $|Du(z) - \overline{p}(z)|$ is small, as previously explained, since $x_n \to y \in \mathcal{A}$ and z is close to x_n, second, that the mollification parameter δ_i can be chosen in such a way that $|\overline{p}(z) - \overline{p}(x_n)|$ is also small.

What is left is to discuss the density issue of the C^1 subsolutions. This is done still assuming $a = c$ and $\mathcal{A}$ nonempty. The proof in the other cases is simpler and goes along the same lines.

It is clear from what was previously outlined that the initial subsolution u and v obtained as a result of the regularization procedure are close in the local uniform topology. It is then enough to show that any critical subsolution w can be approximated in the same topology by a subsolution enjoying the same property of u, namely being strict in W_0 and strictly differentiable on the Aubry set.

The first property is easy to obtain by simply performing a convex combination of w with a C^1 subsolution, strict in W_0, whose existence has been proved above. It can be, in turn, suitably modified in a neighborhood of $\mathcal{A}$ in order to obtain the strict differentiability property, see [20].

Future Directions

This line of research seems still capable of relevant developments. In particular to provide exact and approximate correctors for the homogenization of Hamilton–Jacobi equations in a stationary ergodic environment, see Lions and Souganidis [26] and Davini and Siconolfi [11,12], or in the direction of extending the results about long-time behavior of solutions of time-dependent problems to the noncompact setting, see Ishii [22,23]. Another promising field of utilization is in mass transportation theory, see Bernard [5], Bernard and Buffoni [6], and Villani [30]. The generalization of the model in the case where the Hamiltonian presents singularities should also make it possible to tackle through these techniques the N-body problem. With regard to applications, the theory outlined in the paper could be useful for dealing with topics such as the analysis of dielectric breakdown as well as other models in fracture mechanics.

Bibliography

1. Arnold WI, Kozlov WW, Neishtadt AI (1988) Mathematical aspects of classical and celestial mechanics. In: Encyclopedia of Mathematical Sciences: Dynamical Systems III. Springer, New York
2. Bardi M, Capuzzo Dolcetta I (1997) Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman equations. Birkhäuser, Boston
3. Barles G (1994) Solutions de viscosité des équations de Hamilton–Jacobi. Springer, Paris
4. Barles G, Souganidis PE (2000) On the large time behavior of solutions of Hamilton–Jacobi equations. SIAM J Math Anal 31:925–939
5. Bernard P (2007) Smooth critical subsolutions of the Hamilton–Jacobi equation. Math Res Lett 14:503–511
6. Bernard P, Buffoni B (2007) Optimal mass transportation and Mather theory. J Eur Math Soc 9:85–121
7. Buttazzo G, Giaquinta M, Hildebrandt S (1998) One-dimensional Variational Problems. In: Oxford Lecture Series in Mathematics and its Applications, 15. Clarendon Press, Oxford

8. Clarke F (1983) Optimization and nonsmooth analysis. Wiley, New York
9. Contreras G, Iturriaga R (1999) Global Minimizers of Autonomous Lagrangians. In: 22nd Brazilian Mathematics Colloquium, IMPA, Rio de Janeiro
10. Davini A, Siconolfi A (2006) A generalized dynamical approach to the large time behavior of solutions of Hamilton–Jacobi equations. SIAM J Math Anal 38:478–502
11. Davini A, Siconolfi A (2007) Exact and approximate correctors for stochastic Hamiltonians: the 1-dimensional case. to appear in Mathematische Annalen
12. Davini A, Siconolfi A (2007) Hamilton–Jacobi equations in the stationary ergodic setting: existence of correctors and Aubry set. (preprint)
13. Evans LC (2004) A survey of partial differential methods in weak KAM theory. Commun Pure Appl Math 57
14. Evans LC, Gomes D (2001) Effective Hamiltonians and averaging for Hamilton dynamics I. Arch Ration Mech Anal 157: 1–33
15. Evans LC, Gomes D (2002) Effective Hamiltonians and averaging for Hamilton dynamics II. Arch Rattion Mech Anal 161: 271–305
16. Fathi A (1997) Solutions KAM faibles et barrières de Peierls. C R Acad Sci Paris 325:649–652
17. Fathi A (1998) Sur la convergence du semi–groupe de Lax–Oleinik. C R Acad Sci Paris 327:267–270
18. Fathi A () Weak Kam Theorem in Lagrangian Dynamics. Cambridge University Press (to appear)
19. Fathi A, Siconolfi A (2004) Existence of C^1 critical subsolutions of the Hamilton–Jacobi equations. Invent Math 155:363–388
20. Fathi A, Siconolfi A (2005) PDE aspects of Aubry–Mather theory for quasiconvex Hamiltonians. Calc Var 22:185–228
21. Forni G, Mather J (1994) Action minimizing orbits in Hamiltonian systems. In: Graffi S (ed) Transition to Chaos in Classical and Quantum Mechanics. Lecture Notes in Mathematics, vol 1589. Springer, Berlin
22. Ishii H (2006) Asymptotic solutions for large time Hamilton–Jacobi equations. In: International Congress of Mathematicians, vol III. Eur Math Soc, Zürich, pp 213–227
23. Ishii H () Asymptotic solutions of Hamilton–Jacobi equations in Euclidean n space. Anal Non Linéaire, Ann Inst H Poincaré (to appear)
24. Koike S (2004) A beginner's guide to the theory of viscosity solutions. In: MSJ Memoirs, vol 13. Tokyo
25. Lions PL (1987) Papanicolaou G, Varadhan SRS, Homogenization of Hamilton–Jacobi equations. Unpublished preprint
26. Lions PL, Souganidis T (2003) Correctors for the homogenization of Hamilton-Jacobi equations in the stationary ergodic setting. Commun Pure Appl Math 56:1501–1524
27. Roquejoffre JM (2001) Convergence to Steady States of Periodic Solutions in a Class of Hamilton–Jacobi Equations. J Math Pures Appl 80:85–104
28. Roquejoffre JM (2006) Propriétés qualitatives des solutions des équations de Hamilton–Jacobi et applications. Séminaire Bourbaki 975, 59ème anné. Société Mathématique de France, Paris
29. Siconolfi A (2006) Variational aspects of Hamilton–Jacobi equations and dynamical systems. In: Encyclopedia of Mathematical Physics. Academic Press, New York
30. Villani C () Optimal transport, old and new. http://www.umpa.ens-lyon.fr/cvillani/. Accessed 28 Aug 2008
31. Weinan E (1999) Aubry–Mather theory and periodic solutions of the forced Burgers equation. Commun Pure and Appl Math 52:811–828

Health Care in the United Kingdom and Europe, System Dynamics Applications to

Eric Wolstenholme[1,2]
[1] South Bank University, London, UK
[2] Symmetric SD, Brighton, UK

Article Outline

Glossary
Definition of the Subject
Introduction
The History of System Dynamics
The Need for System Dynamics
The Components of System Dynamics
An Overview of Health and Social Care in the UK and Europe
A Case Study: Using System Dynamics to Influence Health and Social Care Policy Nationally in the UK – Delayed Hospital Discharges
Review of System Dynamics Studies in Epidemiology in Europe
Review of System Dynamics Studies in Health and Social Care Management in Europe
System Dynamics Workforce Planning Models to Support Health Management
Future Directions
Acknowledgments
Bibliography

This paper describes the application of system dynamics to health and social care in Europe.

Systems thinking and the simulation tool set of system dynamics are introduced together with an overview of current strategic health issues and responses in the UK and Europe. A case study is then presented to demonstrate how effective and apposite system dynamics studies can be. This is followed by a pan-European review of applications of system dynamics in epidemiology and in health treatment and diagnosis in different sectors of health and social care, based on an extensive bibliography. Reference is also made to health workforce planning studies. Lastly, a review of future directions is described.

The knowledge base of this paper is located in published work by internal and external consultants and Universities, but it should also be said that there is far more

work in system dynamics in health than is referred to in these sources. Many internal and external consultancies undertake studies which remain unpublished.

The description of the subject and the applications described are comprehensive, but the review is a personal interpretation of the current state of a fast-moving field by the author and apologies are made in advance for any unintended omissions.

The case study in Sect. "A Case Study: Using System Dynamics to Influence Health and Social Care Policy Nationally in the UK – Delayed Hospital Discharges" is extracted from material published by Springer-Verlag, US and published with their permission.

Glossary

System dynamics

System A collection of elements brought together for a purpose and whose sum is greater than the parts.

Systems thinking The process of interpreting the world as a complex, self regulating and adaptive system.

System dynamics A method based on quantitative computer simulation to enhance learning and policy design in complex systems.

Qualitative system dynamics The application of systems thinking and system dynamics principles, without formal simulation.

Dynamic complexity The number of interacting elements contained in a system and the consequences of their interactions over time.

Human activity system Any system created and regulated by human intervention.

Reductionism The opposite of systemic – seeing the world only in its constituent parts.

Feedback Feedback refers to the interaction of the elements of the system where a system element, X, affects another system element, Y, and Y in turn affects X perhaps through a chain of causes and effects. Feedback thus controls the performance of the system. Feedback can be either natural or behavioral (created by human intervention) (System Dynamics Society).

Unintended consequences Undesirable consequences arising well intended action – or vice versa.

Continuous simulation The aggregate method of computer simulation used in system dynamics based on a continuous time analogy with fluid dynamics and used to test out patterns of behavior over time.

System structure The term used in system dynamics to refer to the total structure of a system (composing processes, organization boundaries, information feedback, policy and delays).

System behavior The term used in system dynamics to refer to the behavior over time of a particular structure.

Reference mode of behavior An observed past trend and future projected trends used to assist defining model scope and time frame.

Discrete entity simulation A method of simulation based on the movement of individual entities through systems over time either as processes or as interactions between entities.

Health and Social Care

Epidemiology The study of factors affecting the health and the incidence and prevalence of illness of populations.

Health treatment The application of drugs, therapies, and medical/surgical interventions to treat illness.

National health service (NHS) The organization in the UK responsible for the delivery of health care.

Primary care trusts (PCTs) The local operating agencies of the NHS, which both commission (buy) and deliver health services.

General practitioners (GPs) Locally-based general clinicians who deliver primary care services and control access to specialist health services.

Social services In England, care services which provide non-health related care, mainly for children and older people, located within local government in the UK.

Nursing/residential home care In England, private and public residential establishments for the care of older people.

Domiciliary care In England, care for older people in their own homes.

Acute hospitals Hospital dealing with short term conditions requiring mainly one-off treatment.

Outliers Patients located in hospital in wards not related to their condition, due to bed capacity issues.

Intermediate care Short term care to expedite the treatment of non-complex conditions.

Definition of the Subject

All too often complexity issues are ignored in decision making simply because they are just too difficult to represent. Managers feel that to expand the boundaries of the decision domain to include intricate, cross-boundary interconnections and feedback will detract from the clarity of the issue at stake. This is particularly true when the interconnections are behavioral and hard to quantify. Hence, the focus of decision making is either very subjective or based on simple, linear, easy to quantify components. However, such a reductionist stance, which ig-

nores information feedback (for example, the effects of health supply on health demand management) and multiple-ownership of issues can result in unsustainable, short term benefits with major unintended consequences.

System dynamics is a particular way of thinking and analyzing situations, which makes visible the dynamic complexity of human activity systems for decision support.

It is particularly important in the health and social care field where there are major issues of complexity associated with the incidence and prevalence of disease, an aging population, a profusion of new technologies and multiple agencies responsible for the prevention and treatment of illness along very long patient pathways. Health is also linked at every stage to all facets of life and health policy has a strong political dimension in most countries.

Introduction

This paper describes and reviews work in applying system dynamics to issues of health and social care in the UK and Europe. Although the fundamental issues in health and social care and many of the strategies adopted are similar the world over, there are differences in culture, operational policies and funding even over short geographical distances. Additionally, the health field can be dissected in many different ways both internally and between countries.

There is, moreover, a fundamental dilemma at the center of health that determines both its structure and emphasis. Although the real long term and systemic solution to better health lies in the prevention of illness, the health field focuses on the study of the incidence and prevalence of disease (Epidemiology) and on the 'health service' issues of how to manage ill health (Health Diagnosis and Treatment).

There are many reasons for this, not the least being that illness prevention is in fact the province of a field much bigger than health, which includes economics, social deprivation, drugs, poverty, power and politics.

The field of system dynamics in health reflects this dilemma. Whilst all studies would conclude that prevention is better than the cure, the majority of applications focus on illness. Whilst more studies are required on the truly systemic goal of moving attention away from the status quo, for example, modeling the German system of health care and drug addicts [52], the major focus and impact of system dynamics in Europe in recent years has been in terms of Epidemiology and Health Treatment. Hence, it is these categories that will be the focus of this paper. However, work often transcends the two and models often include both disease and treatment states. For example, work on AIDS covers both prevalence and drug treatment and work on long term conditions, particularly mental health conditions, covers condition progression as well as alternative therapies.

It is important to emphasize what this paper does not cover. By definition system dynamics is a strategic approach aimed at assisting with the understanding of high level feedback effects at work in organizations. It is therefore separate from the many applications of spreadsheets and discrete entity simulation methods applied to answer short term operational level issues in health [9,21,29].

It is also important to note where the knowledge base of this paper is located. System dynamics applications in health in Europe began in the 1980s and are expanding rapidly. However, as will be seen from the bibliography to this paper, much of the work is applied by internal and external consultants and Universities for health care managers and reported in management, operational research and system dynamics journals. Little of the work so far has been addressed directly at clinicians or published in the health literature. It should also be said that there is far more work in system dynamics in health than is referred to in this publication. Many internal and external consultancies undertake studies which remain unpublished.

Initially the fundamentals of system dynamics will be described followed by an overview of current health issues and responses in the UK and Europe. This is followed by a case study to demonstrate how effective and apposite system dynamics studies can be. There then follows a review of applications in epidemiology and in both physical and mental health diagnosis and treatment. Mention is also made of health workforce planning studies. Lastly, a review of future directions is described.

The History of System Dynamics

System dynamics was conceived at MIT, Boston in the late 60s and has now grown into a major discipline [25,47] which was formally celebrated and reviewed in 2008 [48]. It is widely used in the private business sector in production, marketing, oil, asset management, financial services, pharmaceuticals and consultancy. It is also used in the public sector in defense, health and criminal justice.

System dynamics has a long history in the UK and Europe. The first formal university group was established at the University of Bradford in England 1970. Today there are at least a dozen university departments and business schools offering courses in system dynamics and numerous consultancies of all types using the method in one form or another. Thousands of people have attended pri-

vate and university courses in system dynamics and, additionally, there are almost one hundred UK members of the System Dynamics Society, which is the largest national grouping outside the US.

The Need for System Dynamics

Most private and public organizations are large and complex. They exhibit both 'detailed' complexity (the number of elements they contain), but more importantly 'dynamic' complexity (the number of interconnections and interactions they embrace). They have long processes which transcend many sectors, each with their own accounting and performance measures. In the case of health and social care organizations this translates into long patient pathways across many agencies. Complexity and decision making in the public sector is also compounded by a multitude of planning time horizons and the political dimension.

Long processes mean that there are many opportunities for intervention, but that the best levers for overall improvement are often well away from symptoms of problems. Such interventions may benefit sectors other than those making the investments and require an open approach to improving patient outcomes, rather than single agency advantage.

The management of complex organizations is complicated by the fact that human beings have limited cognitive ability to understand interconnections and consequently have limited mental models about the structure and dynamics of organizations.

A characteristic of complex organizations is a tendency for management to be risk averse, policy resistant and quick to blame. This usually means they prefer to stick to traditional solutions and reactive, short term gains. In doing this managers ignore the response of other sectors and levels of the organization. In particular, they underestimate the role and effect of behavioral feedback.

Such oversight can result in unintended consequences in the medium term that undermine well-intended actions. Self organizing and adaptive responses in organizations can lead to many types of informal coping actions, which in turn, inhibit the realization of improvement attempts and distort data. A good example of these phenomena, arising from studies described here, is the use of 'length of stay' in health and social care services as a policy lever to compensate for capacity shortages.

Planning within complex organization reflects the above characteristics. The core of current planning tends to be static in nature, sector-based and reliant on data and financial spreadsheets with limited transparency of assumptions. For example the planning of new acute hospitals can quickly progress to detailed levels without assessment of trends in primary and post acute care; that is, where hospital patients come from and go to.

In contrast, sustainable solutions to problems in complex organizations often require novel and balanced interventions over whole processes, which seem to defy logic and may even be counterintuitive.

However, in order to realize such solutions requires a leap beyond both the thinking and planning tools commonly used today. In order to make significant changes in complex organizations it is necessary to think differently and test ideas before use. System dynamics provides such a method.

The Components of System Dynamics

System dynamics is based on the idea of resisting the temptation to be over reactive to events, learning instead to view patterns of behavior in organizations and ground these in the structure (operational processes and policies) of organizations. It uses purpose-built software to map processes and policies at a strategic level, to populate these maps with data and to simulate the evolution of the processes under transparent assumptions, polices and scenarios.

System dynamics is founded upon:

- Non linear dynamics and feedback control developed in mathematics, physics and engineering,
- Human, group and organizational behavior developed in cognitive and social psychology and economics,
- Problem solving and facilitation developed in operational research and statistics.

System dynamics provides a set of *thinking* skills and a set of *modeling* tools which underpin the current trend of 'whole systems thinking' in health and social care.

System Dynamics Thinking Skills for the Management of Complex Organizations

In order to understand and operate in complex organizations it is necessary to develop a wide range of thinking skills [45]. The following are summarized after Richmond [42].

- **Dynamic thinking** – The ability to conceptualize how organizations behave over time and how we would like them to behave.
- **System-as-cause thinking** – The ability to determine plausible explanations for the behavior of the organization over time in terms of past actions.
- **Forest thinking** – The ability to see the "big picture" (transcending organizational boundaries).

- **Operational thinking** – The ability to analyze the contribution made to the overall behavior by the interaction of processes, information feedback, delays and organizational boundaries.
- **Closed-loop thinking** – The ability to analyze feedback loops, including the way that results can feedback to influence causes.
- **Quantitative thinking** – The ability to determine the mathematical relationships needed to model cause and effect.
- **Scientific thinking** – The ability to construct and test hypotheses through modeling.

System Dynamics Modeling Tools for Planning in Complex Organizations

A useful way to appreciate the tool set of system dynamics is by a brief comparison with other computer based management tools for decision support.

System dynamics is, by definition, a strategic rather than operational tool. It can be used in a detailed operational role, but is first and foremost a *strategic* tool aimed at integrating policies across organizations, where behavioral feedback is important. It is unique in its ability to address the strategic domain and this places it apart from more operational toolsets such as process mapping, spreadsheets, data analysis, discrete entity simulation and agent-based simulation.

System dynamics is based on representing process flows by 'stock' and 'rate' variables. Stocks are important measurable accumulations of physical (and non-physical) resources in the world. They are built and depleted over time as input and output rates to them change under the influence of feedback from the stocks and outside factors. Recognizing the difference between stocks and rates is fundamental to understanding the world as a system. The superimposition of organizational sectors and boundaries on the processes is also fundamental to understanding the impact of culture and power on the flows. System dynamics also makes extensive use of causal maps to both help conceptualize models and to highlight feedback processes within models.

Applying System Dynamics with Management Teams

However, the success of system dynamics lies as much in its process of application as in the tool set and hence demands greater skill in conceptualization and use than spreadsheets.

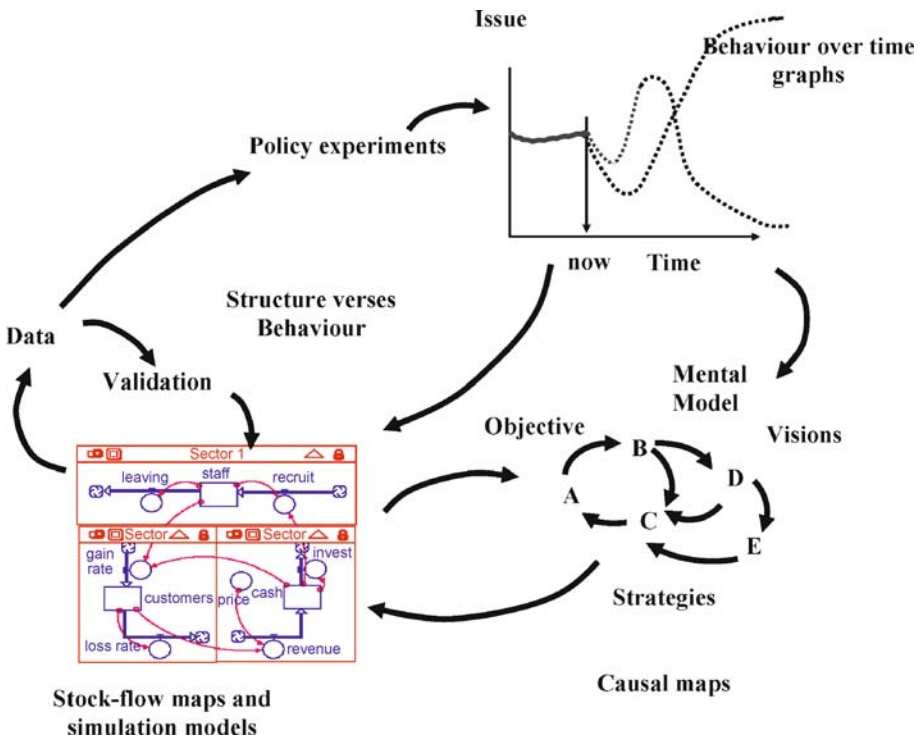

Health Care in the United Kingdom and Europe, System Dynamics Applications to, Figure 1
The systems thinking/system dynamics method

Figure 1 shows the overall process of applying system dynamics. A key starting point is the definition of an initial significant issue of managerial concern and the establishment of a set of committed and consistent management teams from all agencies involved in the issue. Another requirement is a set of facilitators experienced in both conceptualizing and formulating system dynamics models. The models created must be shared extensions of the mental models of the management teams, not the facilitators and, importantly owned by the team.

The next step is the analysis of existing trends in major performance measures of the organizations and of their future trajectories, desired and undesired. This is referred to as the reference model of behavior of the issue and helps with the establishment of the time scale of the analysis. The key contribution of system dynamics is then to formulate a high level process map, at an appropriate level of aggregation, linking operations across organizations and to populate this with the best data available. Once validated against past data, the mental models of the management team and shown capable of reproducing the reference mode of behavior of the issue ('what is'), the model is used to design policies to realize desired futures ('what might be'). Maps and models are constructed in relatively inexpensive purpose-built software (for example *ithink, Vensim* and *Powersim*) with very transparent graphical interfaces.

The key is to produce the simplest model possible consistent with maintaining its transparency and having confidence in its ability to cast new light on the issue of concern. This means keeping the resolution of the model at the highest possible level and this distinguishes it from most spreadsheets and process maps.

An Overview of Health and Social Care in the UK and Europe

Ensuring that all residents have access to health and social care services is an important goal in all EU countries and all have universal or almost universal health care coverage (European Observatory 'Healthcare in Transition' profiles and OECD Health Data 2004). Even in the Netherlands, where only 65% of the population are covered by a compulsory scheme, with voluntary private insurance available to the remainder, only 1.6% of the population are without health insurance.

At the present time, most care in the EU is publicly financed, with taxation and social insurance provide the main sources of funding. Taxation is collected at either the national level or local level, or both and social insurance contributions are generally made by both employees and employers. The role of private insurance varies between countries and generally private insurance is as a supplement to, rather than as a substitute for, the main care system. The exceptions to this are Germany and the Netherlands. Further, people are increasingly required to pay part of the cost of medical care.

The delivery of health and social care is a mixture of public and private with only 10 countries not having any private delivery sector at all.

This paper is primarily concerned with health and social care supply issues. Although the structure and terminology associated with supply varies across the EU the underlying issues tend to be similar between countries. Hence the major issues will be described for England.

Health in England is primarily managed and delivered by the National Health Service (NHS) and is at the center of a modernization agenda, whereby the government sets out a program of change and targets against which the public may judge improved services.

A major mechanism for reform tends to be via frequent changes to organizational structure. The current structure consist of large primary care trusts (PCTs), which both deliver services such as General Practitioner Services (GPs), but also purchase (commission) more specialist services from other agencies, both public and private. A key driver of structural change is to enhance primary care and to take the pressure off acute hospitals (acute is a word used to differentiate short term hospitals from long stay ones). Initiatives here center on providing new services, such as diagnostic and treatment centers and shorter term 'intermediate' care. Emphasis is on bringing the services to the users, patient choice, payment by results (rather than through block contracts) and service efficiency, the latter being driven by target setting and achievement. The government has made reform of public services a key plank in its legislative program and pressure to achieve a broad range of often conflicting targets is therefore immense. However, despite continual increases in funding new initiatives are slow to take effect and the performance and viability of the service is problematic with money often being used to clear deficits rather than generate new solutions.

Social care in England is delivered both by a public sector located with Local Government Social Services Directorates and a private sector. It consists of numerous services to support children and older people. The latter consisting of care homes, nursing homes and domiciliary (at home) care.

Many patient processes, particularly for older people, transcend health and social care boundaries and hence create a serious conflict of process structure and organi-

zational structure, where the relative power of the different agencies is a major determinant of resource allocation [64]. Consequently, emphasis in this paper will be on joint health and social care work.

A Case Study: Using System Dynamics to Influence Health and Social Care Policy Nationally in the UK – Delayed Hospital Discharges

In order to give a flavor of the relevance and impact of applying system dynamics to health and social care issues a concise case study will be presented [65,67,70].

Issue

Delayed hospital discharge was an issue which first came onto the UK legislative agenda in late 2001. The 'reference mode' of behavior over time for this situation was that of increasing numbers of patients occupying hospital beds, although they had been declared "medically fit". In March 2002, 4,258 people were "stuck" in hospital and some were staying a long time, pushing up the number of bed days and constituting significant lost capacity.

The government's approach to this issue was to find out who was supposed to "get the patients out" of acute hospitals and threaten them with 'fines' if they did not improve performance. This organization proved to be social services for older people, who are located within the local government sector and who are responsible for a small, but significant, number of older people needing ex-hospital ('post-acute') care packages. Such patients are assessed and packages organized by hospital social workers. There was also pressure on the government from hospitals claiming that some of the problem was due to lack of hospital capacity.

The idea of fines was challenged by the Local Government Association (LGA), which represents the interests of all local government agencies at the national level) who suggested that a 'system' approach should be undertaken to look at the complex interaction of factors affecting delayed hospital discharges. This organization, together with the NHS Confederation (the partner organization representing the interests of the National Health Service organizations at a national level) then commissioned a system dynamics study to support their stance.

The remit was for consultants working with the representatives of the two organizations to create a system dynamics model of the 'whole patient pathway' extending upstream and downstream from the stock of people delayed in hospital, to identify and test other interventions affecting the issue.

Model

A system dynamics model was developed interactively with managers from the LGA and NHS, using national data to simulate pressures in a sample health economy covering primary, acute and post acute care over a 3 year period. The model was driven by variable demand including three winter pressure "peaks" when capacity in each sector was stretched to the limit. Figure 2 shows an overview of the sectors of the model.

The patient flows through the model were broken down into medical flows and surgical with access to the medical and surgical stocks of beds being constrained by bed capacity. The medical flows were mainly emergencies patients and the surgical flows mainly non-emergency 'elective' patients, who came via referral processes and wait lists.

Further, medical patients were broken down into 'fast' and 'slow' streams. The former were the normal patients who had a short stay in hospital and needed few post acute services and the latter the more complex cases (mainly older people), who require a longer stay and hospital and complex onward care packages from social services. This split was because although the slow patients were few in number they constituted most of the people who caused delayed discharges.

The post hospital health and social care services of intermediate care, nursing/residential home care, and domiciliary care were included in the model and were also capacity constrained in terms of the number of care packages they could provide.

The model incorporated a number of mechanisms by which hospitals coped during periods of high demand, for example, moving medical patients to surgical beds (outliers) and early discharges with allowance for readmissions.

Configuration of the Model

The model was set up to simulate a typical sample health economy over a 3 year period when driven by a variable demand (including three winter "peaks"). The capacity constrained sectors of the model were given barely sufficient capacity to cope. This situation was designed to create shocks against which to test alternative policies for performance improvement. Major performance measures in use in the various agencies were incorporated. These included:

1. Cumulative episodes of elective surgery.
2. Elective wait list size and wait time.

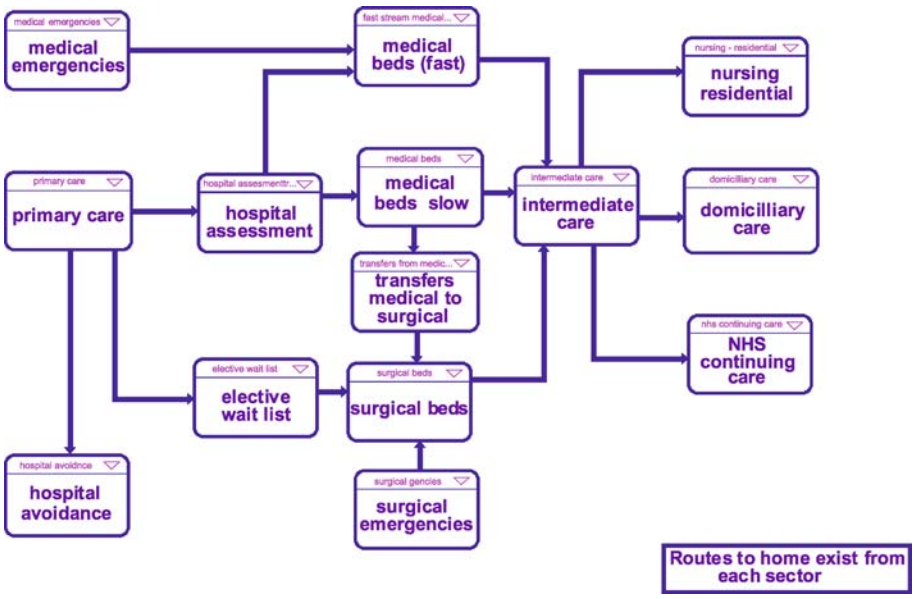

Health Care in the United Kingdom and Europe, System Dynamics Applications to, Figure 2
An overview of the sectors of the delayed discharge model

3. Numbers of patients in hospital having completed treatment and assessment, but not yet discharged (delayed discharges).
4. Number of 'outliers'.

The model was initially set up with a number of fixed experiments, to introduce people to the range of experiments that yielded useful insights into the behavior of the whole system. From there, they were encouraged to devise their own experiments and develop their own theories of useful interventions and commissioning strategies.

The three main polices tested in the fixed runs were:

1. Adding additional acute hospital bed capacity. This is the classic response used over many years by governments throughout the world to solve any patient pathway capacity problems and was a favorite 'solution' here.
2. Adding additional post acute capacity, both nursing and residential home beds but also more domiciliary capacity.
3. Diverting more people away from hospital admission by use of pre-hospital intermediate capacity and also expansion of treatment in primary care GP surgeries.

Example Results from the Delayed Hospital Discharge Model

Figures 3, 4 and 5 show some typical outputs for the delayed hospital discharge model. Figure 3 captures the way capacity utilization was displayed (actual beds occupied *v* total available for both medical and surgical sectors of the hospital) and shows the occurrence of 'outliers' (transfers of patients from medical to surgical beds) whenever medical capacity was reached.

Figures 4 and 5 show comparative graphs of 3 policy runs for 2 major performance measures for 2 sectors of the patient pathway – delayed discharges for post acute social services and cumulative elective procedures for acute hospitals. In each case the base run is line 1. Line 2 shows the effect of increasing hospital beds by 10% and line 3 shows the effect of increasing post acute capacity by 10%.

The interesting feature of this example output is that the cheaper option of increasing post acute capacity gives lower delayed discharges and higher elective operations whereas the more expensive option of increasing acute hospital beds benefits the hospital but makes delayed discharges worse. The key to this counter intuitive effect is that increasing post acute capacity results in higher hospital discharges which in turn reduces the need for the 'outlier' coping policy in the hospital, hence freeing up surgical capacity for elective operations.

Outcomes

Common Sense Solutions Can Be Misleading The obvious unilateral solution of adding more acute capacity was shown to exacerbate the delayed discharge situation. Increasing hospital capacity means facilitating more hos-

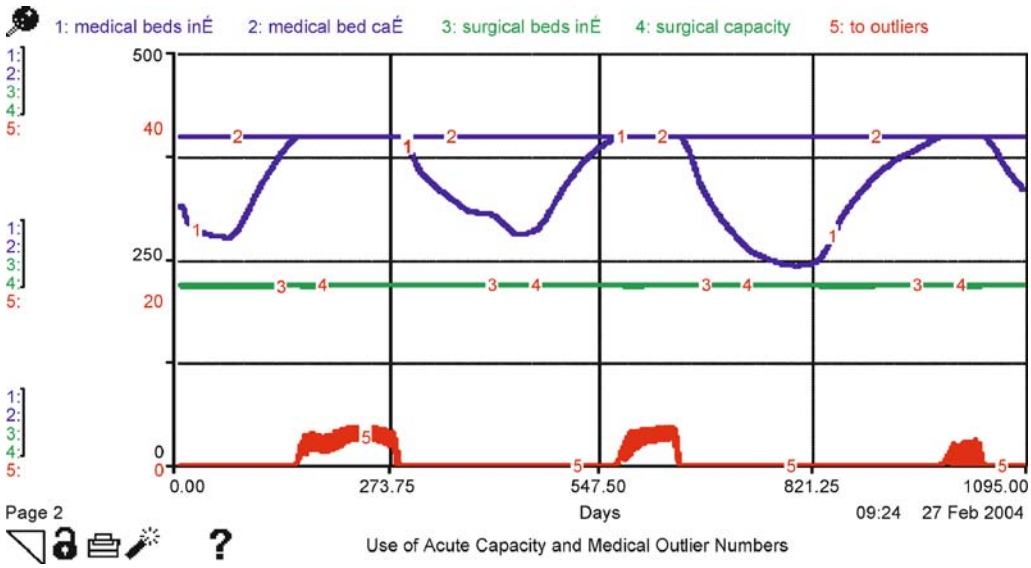

Health Care in the United Kingdom and Europe, System Dynamics Applications to, Figure 3
Medical and surgical bed utilization's in hospital and 'outliers'

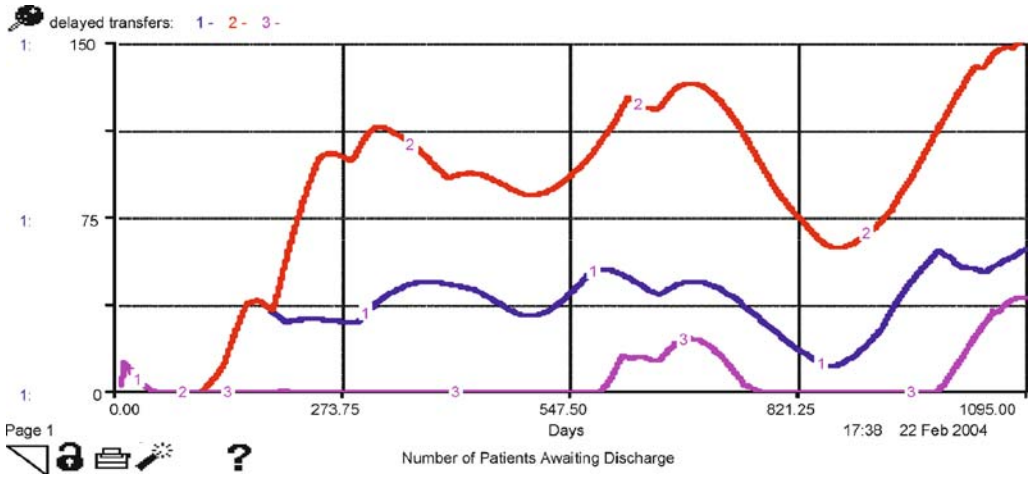

Health Care in the United Kingdom and Europe, System Dynamics Applications to, Figure 4
Delayed hospital discharges for 3 policy runs of the model

pital admissions, but with no corresponding increase in hospital discharges. Hence, the new capacity will simply fill up and then more early discharges and outliers will be needed.

Fines May Have Unintended Consequences This solution was shown to depend on where the money raised by fines was spent. If the money levied from social services was given to the acute sector to finance additional capacity it was clearly demonstrated that this would make delayed discharges worse. It would be worse still if it causes the post-acute sector to cut services. The effects of service cuts may also then spill over into other areas of local government including housing and education.

It was demonstrated that there were some interventions that could help:

1. Increasing post acute capacity gives a win-win solution to both health and social care because it increases all acute and post acute sector performance measures. Such action allows hospital discharges to directly increase, and eliminates the need for the hospitals to apply coping policies, which in turn increases elective operations and reduces elective wait times. Further,

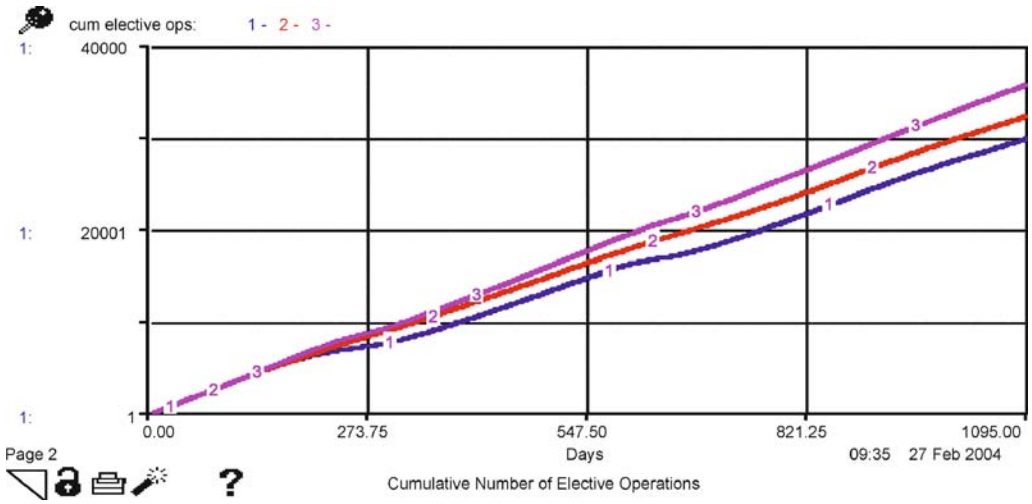

Health Care in the United Kingdom and Europe, System Dynamics Applications to, Figure 5
Cumulative elective operations for 3 policy runs of the model

counter intuitively, increasing medical capacity in hospital is more effective than increasing surgical capacity for reducing elective wait times.

2. Reducing assessment times and lengths of stay in all sectors is beneficial to all performance measures, as is reducing variation in flows, particularly reinforcing feedback loops like re-admission rates.
3. Increasing diversion from hospitals into pre-admission intermediate care was almost as beneficial as increasing post acute capacity.
4. If fines are levied they need to be re-invested from a whole systems perspective. This means re-balancing resources across all the sectors (NOT just adding to hospital capacity).
5. In general the model showed that keeping people out of hospital is more effective than trying to get them out faster. This is compounded by the fact that in-patients are more prone to infections so the longer patients are in hospital, the longer they will be in hospital.
6. Improving the quality of data was shown to be paramount to realizing the benefits of all policies. This is an interesting conclusion associated with many system dynamics studies, where explicit representation of the structure of the organization can lead to a review and redesign of the information needed systems to really manage the organization.

An interesting generalization of the findings was that increasing stock variables where demand is rising (such as adding capacity) is an expensive and unsustainable solution. Whereas increasing rate variables, by reducing delays and lengths of stay, is cheaper and sustainable.

Impact

This model was shown at the Political Conferences of 2002 and generated considerable interest. It was instrumental in causing re-thinking of the intended legislation, so that social services was provided with investment funding to address capacity issues, and the implementation of fines was delayed for a year. Reference to the model was made in the House of Lords.

> Moving the main amendment, Liberal Democrat health spokesperson Lord Clement-Jones asked the House to agree that the Bill failed to tackle the causes of delayed discharges and would create perverse incentives which would undermine joint working between local authorities and the NHS and distort priorities for care of elderly people by placing the requirement to meet discharge targets ahead of measures to avoid hospital admission … **He referred to "ithink", the whole systems approach being put forward by the Local Government Association, health service managers and social services directors involving joint local protocols and local action plans prepared in co-operation.**

Postscript

This case study demonstrates the ability of system dynamics to be applied quickly and purposefully to shed rigor and insight on an important issue. The study enabled the development of a very articulate and compelling case for the government to move from a reactive position of blaming social services to one of understanding and acting on

a systemic basis. The whole project including modeling and communication of the outcomes was completed in 6 weeks.

Review of System Dynamics Studies in Epidemiology in Europe

The potential for system dynamics in population health and disease control began in the UK in the late eighties/early nineties with the extensive studies carried out on AIDS modeling. The majority of these studies were by Prof. Brian Dangerfield and Carole Roberts and were ongoing until 2000 [14,15,16,17,18,19,20].

The earlier studies [16] used a transition model to portray the nature of the disease and to better specify the types of data collection required for further developments of the model. The model was then developed further over the years [18,19] and was fed with time-series data of actual cases. This enabled projections of future occurrence to be forecast. The latter models were more concerned with examining the resource and cost implications of treatments given to HIV positive individuals and at their varying stages up until the ensuing onset of AIDS.

A recent study by Dangerfield et al. [20] saw further development of the original model with parameter optimization and recent data on the spread of AIDS in the UK was also integrated. The rationale for the update of the model was to investigate the recent dramatic decrease in diagnosed Aids cases in the West. The model assesses the effects of relatively new emergent triple antiretroviral therapy given to HIV patients causing this reduction and examines the possibility of continuity of the effectiveness of this therapy.

Dangerfield explains some of the reasons [13] why system dynamics acts as an excellent tool for epidemiological modeling. The positive and negative feed-back loops help imitate the natural disposition of the spread and containment of diseases amongst the general population. Further, system dynamics allows delays associated with the incubation predisposition of infectious diseases to be accurately and easily modeled without the need for complicated mathematical representation.

The work in the UK was complemented by work in Holland on simulation as a tool in the decision-making process to prevent HIV incidence among homosexual men [23] and on models for analysis and evaluation of strategies for preventing AIDS [32]. Further epidemiological studies in system dynamics in the UK related to the outbreak out of BSE and the subsequent infection of humans with its human form nvCJD [12].

These models are all characterized by modeling the flow of people through different stocks over time representing the different stages of the disease progression. The purpose of the model is then to test the effects of interventions aimed at slowing down the rate of progression of the condition or indeed moving people 'upstream' to less severe states of the condition.

Review of System Dynamics Studies in Health and Social Care Management in Europe

By far the greatest number of studies and publications in the use of system dynamics in health and social care is associated with patient flow modeling for health care planning. That is, the flow of patients through multiple service delivery channels. Patient pathway definition has been an area of health modernization and these pathways lend themselves to representation as stock/flow resource flows in system dynamics. The purpose of this type of modeling is to identify bottlenecks, plan capacity, reduce wait lists, improve the efficiency of patient assessments and times and the design of alternative pathways with shorter treatment times, (for example, intermediate care facilities both pre and post hospital treatment).

A characteristic of patient flows is that they are long and pass through multiple agencies and hence confront the major health issues of working across boundaries and designing integrated policies. Studies in this area have examined the flow of many different populations of patients and often resulted in arrayed models to represent the flow of different 'populations' or 'needs groups' through several parallel service channels.

The studies have covered both physical and mental conditions and have sometimes combined both the dynamic progression of people through undiagnosed and untreated disease states and the dynamic progression of diagnosed people through treatment pathways.

The Modeling of the Diagnosis and Treatment of Physical Conditions

Here the most common set of models are associated with the flow of patients from primary care, through acute hospitals and onwards into post acute care such as social services provisions for home care, nursing care and residential care. The populations have often been split between the simple everyday cases and the complex cases associated with older people needing greater degrees of care. They have also involves medical and surgical splits. There are a number of review papers which supplement the work described below [1,18,19].

In addition to work in the 1990s on the interface between health and social care [59,60,61] and the national level UK work on older people flows through hospitals [65,67,71,72], Wolstenholme has reported that system dynamics applications are currently underway by the authors in 10 local health communities around the UK with the objectives of modeling patient flows across agency boundaries to provide a visual and quantitative stimulus to strategic multi-agency planning [65].

Lane has reported work in Accident and Emergency Departments [33] and in mapping acute patient flows [34] whilst Royston worked with the NHS to help develop and implement policies and programs in health care in England [43]. Taylor has undertaken award winning modeling of the feedback effects of reconfiguring health services [49,50,51], whilst Lacey has reported numerous UK studies to support the strategic and performance management roles of health service management, including provision of intermediate care and reduction of delayed hospital discharges [31]. Other intermediate care and social care delivery studies are described by Bayer [6,7] and further hospital capacity studies by Coyle [11]. Elsewhere, there have been specific studies on bed-blocking [24] and screening [37].

In Norway system dynamics-based studies have focused on mapping the flows of patients in elderly non-acute care settings [10]. The purpose of this study according to Chen is to differentiate between acute and non-acute settings and thereby increase understanding of the complexity and dynamics caused by influencing elements in the system. Also it is to provide a tool for local communities in Norway for their long term budget planning in the non-acute health sector for the elderly.

Work on reducing waiting lists has been reported in Holland [30,53,54,57]). Also in Holland Vennix has reported comprehensive work on modeling a regional Dutch health care system [56].

Work has been undertaken to balance capacities in individual hospitals in Italy [44] and in Norway [38,41]. Whist normally the realm of more operational types of simulation system dynamics has proved very effective here. There has also been work to assess the impact on health and social care of technological innovation, particularly telecare [5,8]. Additionally, system thinking has been undertaken by doctors to examine the European time directive [40].

Given the similar nature of a lot of these studies further detail here will focus on the work of Vennix in participative model building and Wolstenholme in extracting insights from numerous studies.

Participative Model Building

A characteristic of all Vennix's work has been group model building [55]. The main objectives of this [27] are communication and learning and integration of multiple perspectives where the process of model building is frequently more important than the resulting model itself [56]. Vennix brought together strategic managers and important stakeholders to participate in the process of building a system dynamics model of the Dutch healthcare system. The policy problem which is modeled in Vennix's 1992 study is related to the gradual, but persistent, rise in health care costs in the Netherlands. Vennix [56] attempts to find the underlying causes of those increases that emanate from within the health care system itself rather than focusing on exogenous factors. By doing so Vennix stands to identify potential levers within the health care system that can be practically and appropriately be adjusted to reduce cost increases.

Vennix attempts to extract important assumptions from the key players by posing three straight forward questions;

a) What factors have been responsible for the increase in health care costs?
b) How will health care costs develop in the future?
c) What are the potential effects of several policy options to reduce these costs?

Participants are asked if they agreed or disagreed with the statements and why they thought the statements were true or not. The most frequently given reasons for the verbal statements were then incorporated in to the statements to create causal arguments from the participant's mental models.

Similar methods were adopted to identify policies which represent the aggregate of many individual actions. For example, why a GP may decide on such matters as frequency of patients appointments, drugs choice, referral to other medical specialist or a combination of all these. Vennix's model was subsequently formalized and quantified and converted into a computer-based learning environment for use by a wider range of health personnel.

The idea of using system dynamics as a means of participative modeling for learning is also inherent in other work [35].

Offering Insights into Managing the Demand for Health Care

Wolstenholme reports the insights from many applications of his own and other work. He suggests a hypothesis that the 'normal' mode of operation for many health and

social care organizations today is often well beyond their safe design capacity. This situation arises from having to cope with whatever demand arrives at their door irrespective of their supply capability. Risk levels can be high in these organizations and the consequences could be catastrophic for patients [71,72].

Evidence for the hypothesis has emerged at many points along patient pathways in health and social care from a number of studies carried out using system dynamics simulation to identify and promote systemic practice in local health communities. The rigor involved in knowledge-capture and quantitative simulation model construction and running has identified mismatches between how managers claim their organizations work and the observed data and behavior. The discrepancies can only be explained by surfacing informal coping policies. For example, transferring medical emergency patients to surgical wards, resulting in canceled elective procedures, also reported by Lane [35]. Indeed, the data itself becomes questionable as it reflects more the actions of managers than the true characteristics of patients.

The result of capacity pressure can mean that managers are unable, physically and financially, to break out from a fire-fighting mode to implement better resource investment and development policies for systemic and sustainable improvement. The insights reported are important for Health and Social Care management, the meaning of data and for modeling. The key message here is that much-needed systemic solutions and whole system thinking can never be successfully implemented until organizations are allowed to articulate and dismantle their worst coping strategies and return to working within best practice capacities.

The Modeling of the Treatment of Mental Health Diagnosis and Treatments in the UK

Modeling to assist mental health reform has recently developed as a separate strand of health work in the UK [46,69,72].

Mental health services in the UK over the past 50 years have undergone numerous major reforms. The National Institute for Clinical Excellence [36] has recently published extensive research-based guidelines on the way stepped care might be best achieved. These involved moves towards a balanced, mixed community/institutional provision of services set within a range of significant reforms to the National Health Service. The latest and perhaps most significant reform is that associated with the introduction of 'stepped care'. Stepped care is aimed at bringing help to more patients more cheaply by developing intermediate staff, services and treatments between GPs and the specialist health hospitals.

Having decided on the new treatments at each step and having designed the basic patient pathways, modeling has been used in the North West of England to help with communication of the benefits and to overcome anticipated problems with resource reallocation issues [69]. Further work in Lincolnshire UK [58] reports the increasing use of 'matrix' modeling in mental health to capture the dynamics of both patient needs and treatments. This work also demonstrates the dangers of over-investment in situations where much demand is in accrued backlogs and incidence is reducing due to better and more successful interventions.

The depression work has also led to work at the Department of Health in the UK to help analyze the national impact of stepped services for mental health on the totality of the labor market and unemployment [72]. This work is an example of the value that system dynamics can add to conventional cost benefit analysis. A static cost benefit analysis was developed into a system dynamics model. By developing a bigger picture of the issue, both upstream to where patients go after treatment and downstream from where patients originate in the labor market, and by simulation of the enhanced vision, the dynamic cost benefit analysis is shown to advance understanding of the issue and plans.

The work questions the magnitude of the potential benefits, introduces phasing issues, surfaces structural insights, takes account of the dynamics of the lab-our market and forces linkages between the plan and other initiatives to get people back to work. The paper suggests that cost benefit analysis and system dynamics are very complementary and should be used together in strategic planning.

Other mental health capacity planning studies have been carried out for individual mental health hospitals and trusts. One such study [71] describes the application of system dynamics to assist decision making in the reallocation of resources within a specialist mental health trust in south London. Mental health service providers in the UK are under increasing pressure to both reduce their own costs and to move resources upstream in mental health patient pathways to facilitate treating more people, whilst not compromising service quality.

The investigation here focused on the consequences of converting an existing specialist service ward in a mental health hospital into a 'triage' ward, where patients are assessed and prioritized during a short stay for either discharge or onward admission to a normal ward. Various policies for the transition were studied together with the

implications for those patients needing post hospital services and relocation within the community. The model suggested that the introduction of a triage ward could meet the strategic requirement of a 10% shift away from institutional care and into community services. The paper includes a number of statements from the management team involved on the benefits of system dynamics and the impact of its application on their thinking.

System Dynamics Workforce Planning Models to Support Health Management

It is also important to mention that work has been carried out in a number of countries in the field of workforce planning related to health. In the UK the NHS has deployed sophisticated workforce planning models to determine the training and staffing needs associated with numerous alternative service configurations. In the Spanish Health system modeling has been used to determine the number of doctors required for a number of specialists services and to attempt to explore solutions for the current imbalance among supply and demand of physicians [2,4,5]. Elsewhere the factors affecting staff retention has been studied [28] and in the Netherlands, an advisory body of the Dutch government was given the responsibility of implying a new standard for the number of rheumatologists [39]. One of the main factors that were studied in the scenario analysis stage was the influences of changing demographics on the demand of manpower in the health system. Other studies have covered time reduction legislation on doctor training [22].

Future Directions

System dynamics has already made a significant impact on health and social care thinking across the EU. Many policy insights have been generated and the organizations are increasingly being recognized as complex adaptive systems. However, true understanding and implementation of the messages requires much more work and too many organizations are still locked into a pattern of short-termism which leads them to focus on the things they feel able to control – usually variables within their own individual spheres of control. There are also some aspects of system reform in some countries that are producing perverse incentives which encourage organizations to apply short-term policies.

Wider communication of existing studies and further studies are necessary to demonstrate the advantages of sustainable, systemic solutions. The key challenge lies in demonstrating to a wider audience of managers and clinicians that they can add value to the whole whilst remaining autonomous. An important element is to train more people capable of modeling and facilitating studies and to simplify the process and software of system dynamics.

Acknowledgments

The author would like to acknowledge the many practitioners of system dynamics in health and social care throughout Europe for their dedication to using these methods to improve health and wellbeing. Particular thanks are due to my colleagues in the International Systems Dynamics community and its Society and those in the UK Chapter of the Society, without whom I would not have the knowledge base to undertake this review. Special thanks are also due to my colleagues in Symmetric SD without whom much of the personal work reported here would not have been carried out at all.

Bibliography

Primary Literature

1. Abdul-Salam O (2006) An overview of system dynamics applications. In: A dissertation submitted to the University of Salford Centre for Operational Research and Applied Statistics for the degree of MSc Centre for Operational Research and Applied Statistics, University of Salford
2. Alonso Magdaleno MI (2002) Administrative policies and MIR vacancies: Impact on the Spanish Health System. In: Proceedings of the 20th International Conference of the System Dynamics Society, Palermo, 2002
3. Alonso Magdaleno MI (2002) Dynamic analysis of some proposals for the management of the number of physicians in Spain. In: Proceedings of the 20th International Conference of the System Dynamics Society, Palermo, 2002
4. Alonso Magdaleno MI (2002) Elaboration of a model for the management of the number of specialized doctors in the spanish health system. In: Proceedings of the 20th International Conference of the System Dynamics Society, System Dynamics Society, Palermo, 2002
5. Bayer S (2001) Planning the implementation of telecare services. In: The 19th International Conference of the System Dynamics Society, System Dynamics Society, Atlanta, 2001
6. Bayer S (2002) Post-hospital intermediate care: Examining assumptions and systemic consequences of a health policy prescription. In: Proceedings of the 20th International Conference of the System Dynamics Society, System Dynamics Society, Palermo, 2002
7. Bayer S, Barlow J (2003) Simulating health and social care delivery. In: Proceedings of the 21st International Conference of the System Dynamics Society, System Dynamics Society, New York, 2003
8. Bayer S, Barlow J (2004) Assessing the impact of a care innovation: Telecare. In: 22nd International Conference of the System Dynamics Society, System Dynamics Society, Oxford, 2004
9. Brailsford SC, Lattimer VA (2004) Emergency and on-demand health care: Modelling a large complex system. J Operat Res Soc 55:34–42

10. Chen Y (2003) A system dynamics-based study on elderly non-acute service in norway. In: Proceedings of the 21st International Conference of the System Dynamics Society, System Dynamics Society, New York, 2003
11. Coyle RG (1996) A systems approach to the management of a hospital for short-term patients. Socio Econ Plan Sci 18(4):219–226
12. Curram S, Coyle JM (2003) Are you vMad to go for surgery? Risk assessment for transmission of vCJD via surgical instruments: The contribution of system dynamics. In: Proceedings of the 21st International Conference of the System Dynamics Society, New York, 2003
13. Dangerfield BC (1999) System dynamics applications to european health care issues. System dynamics for policy, strategy and management education. J Operat Res Soc 50(4):345–353
14. Dangerfield BC, Roberts CA (1989) A role for system dynamics in modelling the spread of AIDS. Trans Inst Meas Control 11(4):187–195
15. Dangerfield BC, Roberts CA (1989) Understanding the epidemiology of HIV infection and AIDS: Experiences with a system dynamics. In: Murray-Smith D, Stephenson J, Zobel RN (eds) Proceedings of the 3rd European Simulation Congress. Simulation Councils Inc, San Diego, pp 241–247
16. Dangerfield BC, Roberts CA (1990) Modelling the epidemiological consequences of HIV infection and AIDS: A contribution from operational research. J Operat Res Soc 41(4):273–289
17. Dangerfield BC, Roberts CA (1992) Estimating the parameters of an AIDS spread model using optimisation software: Results for two countries compared. In: Vennix JAM, Faber J, Scheper WJ, Takkenberg CA (eds) System Dynamics. System Dynamics Society, Cambridge, pp 605–617
18. Dangerfield BC, Roberts CA (1994) Fitting a model of the spread of AIDS to data from five european countries. In: Dangerfield BC, Roberts CA (eds) O.R. Work in HIV/AIDS 2nd edn. Operational Research Society, Birmingham, pp 7–13
19. Dangerfield BC, Roberts CA (1996) Relating a transmission model of AIDS spread to data: Some international comparisons. In: Isham V, Medley G (eds) Models for infectious human diseases: Their structure and relation to data. Cambridge University Press, Cambridge, pp 473–476
20. Dangerfield BC, Roberts CA, Fang Y (2001) Model-based scenarios for the epidemiology of HIV/AIDS: The consequences of highly active antiretroviral therapy. Syst Dyn Rev 17(2):119–150
21. Davies R (1985) An assessment of models in a health system. J Operat Res Soc 36:679–687
22. Derrick S, Winch GW, Badger B, Chandler J, Lovett J, Nokes T (2005) Evaluating the impacts of time-reduction legislation on junior doctor training and service. In: Proceedings of the 23rd International Conference of the System Dynamics Society, Boston, 2005
23. Dijkgraaf MGW, van Greenstein GJP, Gourds JLA (1998) Interactive simulation as a tool in the decision-making process to prevent HIV incidence among homosexual men in the Netherlands: A proposal. In: Jager JC, Rotenberg EJ (eds) Statistical Analysis and Mathematical Modelling of AIDS. OUP, Oxford, pp 112–122
24. El-Darzi E, Vasilakis C (1998) A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in hospitals. Health Care Manag Sci 1(2):143–149
25. Forrester JW (1961) Industrial Dynamics. MIT Press
26. Gonzalez B, Garcia R (1999) Waiting lists in spanish public hospitals: A system dynamics approach. Syst Dyn Rev 15(3):201–224
27. Heyne G, Geurts JL (1994) DIAGNOST: A microworld in the healthcare for elderly people. In: 1994 International System Dynamics Conference. System Dynamics Society, Sterling
28. Holmstroem P, Elf M (2004) Staff retention and job satisfaction at a hospital clinic: A case study. In: 22nd International Conference of the System Dynamics Society, Oxford, 2004
29. Jun JB, Jacobson SH, Swisher JR (1999) Application of discrete-event simulation in health care clinics: A survey. J Operat Res Soc 50(2):109–123
30. Kim DH, Gogi J (2003) System dynamics modeling for long term care policy. Proceedings of the 21st International Conference of the System Dynamics Society. System Dynamics Society, New York
31. Lacey P (2005) Futures through the eyes of a health system simulator, Paper presented to the System Dynamics Conference, Boston 2005
32. Lagergren M (1992) A family of models for analysis and evaluation of strategies for preventing AIDS. In: Jager JC (eds) Scenario Analysis. Elsvier, Amsterdam, pp 117–145
33. Lane DC (2000) Looking in the wrong place for healthcare improvements: A system dynamics study of an accident and emergency department. J Operat Res Soc 51(5):518
34. Lane DC, Husemann E (2008) System dynamics mapping of acute patient flows. J Oper Res Soc 59:213–224
35. Lane DC, Monefeldt C et al (2003) Client involvement in simulation model building: Hints and insights from a case study in a London hospital. Health Care Manag Sci 6(2):105–116
36. National Institute for Clinical Excellence (2004) Depression: Management of depression in primary and secondary care – NICE guidance. National Clinical Practice Guideline 23
37. Osipenko L (2006) System dynamics model of a new prenatal screening technology (Screening pregnant women). In: 24th International Conference of the System Dynamics Society, Nijmegen, The Netherlands, 23–27 July 2006
38. Petersen LO (2000) How should the capacity for treating heart decease be expanded? In: 18th International Conference of the System Dynamics Society, Bergen, 2000
39. Posmta TJBM, Smits MT (1992) Personnel planning in health care: An example in the field of rheumatology. In: Proceedings of the 1992 International System Dynamics Conference of the System Dynamics Society. System Dynamics Society, Utrecht
40. Ratnarajah M (2005) European union working time directive. In: 7th Annual Gathering. System Dyanamics Society, Harrogate, February 2005
41. Ravn H, Petersen LO (2007) Balancing the surgical capacity in a hospital. Int J Healthcare Technol Manag 14:4023–4089
42. Richmond B (1994) ISEE Systems Inc, Hanover
43. Royston G, Dost A (1999) Using system dynamics to help develop and implement policies and programmes in health care in England. Syst Dyn Rev 15(3):293–315
44. Sedehi H (2001) HDS: Health department simulator. In: The 19th International Conference of the System Dynamics Society. System Dynamics Society, Atlanta, 2001
45. Senge P (1990) The fifth discipline doubleday. New York
46. Smith G, Wolstenholme EF (2004) Using system dynamics in modeling health issues in the UK. In: 22nd International Conference of the System Dynamics Society. The System Dynamics Society, Oxford, 2004

47. Sterman J (2000) Business dynamics: System thinking and modelling for a complex world. McGraw-Hill, Boston
48. Sterman T (ed) (2008) Exploring the next frontier: System Dynamics at 50. Syst Dyn Rev 23:89–93
49. Taylor KS (2002) A system dynamics model for planning and evaluating shifts in health services: The case of cardiac catheterisation procedures in the NHS London, London School of Economics and Political Science. J Opr Res Soc 56:659–1229
50. Taylor KS, Dangerfield BC (2004) Modelling the feedback effects of reconfiguring health services. J Operat Res Soc 56:659–675 (Published on-line Sept 2004)
51. Taylor KS, Dangerfield BC, LeGrand J (2005) Simulation analysis of the consequences of shifting the balance of health care: A system dynamics approach. J Health Services Res Policy 10(4):196–202
52. Tretter F (2002) Modeling the system of health care and drug addicts. In: Proceedings of the 20th International Conference of the System Dynamics Society, Palermo, 2002
53. Van Ackere A, PC Smith (1999) Towards a macro model of national health service waiting lists. Syst Dyn Rev 15(3):225
54. Van Dijkum C, Kuijk E (1998) Experiments with a non-linear model of health-related actions. In: 16th International Conference of the System Dynamics Society. System Dynamics Society, Quebec, 1998
55. Vennix AM (1996) Group Model Building. Wiley, Chichester
56. Vennix JAM, JW Gubbels (1992) Knowledge elicitation in conceptual model building: A case study in modeling a regional dutch health care system. Europ J Operat Res 59(1):85–101
57. Verburgh LD, Gubbels JW (1990) Model-based analyses of the dutch health care system. In: System Dynamics '90: Proceedings of the 1990 International Systems Dynamics Conference, International System Dynamics Society, Chestnut Hill, 1990
58. Wolstenholme E, McKelvie D, Monk D, Todd D, Brad C (2008) Emerging opportunities for system dynamics in UK health and social care – the market-pull for systemic thinking. Paper submitted to the 2008 System Dynamics Conference, Athens 2008
59. Wolstenholme EF (1993) A case study in community care using systems thinking. J Operat Res Soc 44(9):925–934
60. Wolstenholme EF (1996) A management flight simulator for community care. In: Cropper S (ed) Enhancing decision making in the NHS. Open University Press, Milton Keynes
61. Wolstenholme EF (1999) A patient flow perspective of UK health services: Exploring the case for new intermediate care initiatives. Syst Dyn Rev 15(3):253–273
62. Wolstenholme EF (2004) Using generic system archetypes to support thinking and learning. Syst Dyn Rev 20(2):341–356
63. Wolstenholme EF, McKelvie D (2004) Using system dynamics in modeling health and social care commissioning in the UK. In: 22nd International Conference of the System Dynamics Society, Oxford, 2004
64. Wolstenholme EF, Monk D (2004) Using system dynamics to influence and interpret health and social care policy in the UK. In: 22nd International Conference of the System Dynamics Society, Oxford, 2004
65. Wolstenholme EF, Monk D, Smith G, McKelvie D (2004) Using system dynamics in modelling health and social care commissioning in the UK. In: Proceedings of the 2004 System Dynamics Conference, Oxford, 2004
66. Wolstenholme EF, Monk D, Smith G, McKelvie D (2004) Using system dynamics in modelling mental health issues in the UK. In: Proceedings of the 2004 System Dynamics Conference, Oxford, 2004
67. Wolstenholme EF, Monk D, Smith G, McKelvie D (2004) Using system dynamics to influence and interpret health and social care policy in the UK. In: Proceedings of the 2004 System Dynamics Conference, Oxford, 2004
68. Wolstenholme EF, Arnold S, Monk D, Todd D, McKelvie D (2005) Coping but not coping in health and social care – masking the reality of running organisations well beyond safe design capacity. Syst Dyn Rev 23(4):371–389
69. Wolstenholme EF, Arnold S, Monk D, Todd D, McKelvie D (2006) Reforming mental health services in the UK – using system dynamics to support the design and implementation of a stepped care approach to depression in north west england. In: Proceedings of the 2006 System Dynamics Conference, Nijmegen, 2006
70. Wolstenholme EF, Monk D, McKelvie D (2007) Influencing and interpreting health and social care policy in the UK In: Qudrat-Ullah H, Spector MJ, Davidsen PI (eds) Complex decision making: Theory and practice. Springer, Berlin
71. Wolstenholme EF, Monk D, McKelvie D, Gillespie P, O'Rourke D, Todd D (2007) Reallocating mental health resources in the borough of Lambeth, London, UK. In: Proceedings of the 2007 System Dynamics Conference, Boston, 2007
72. Wolstenholme EF, Monk D, McKelvie D, Todd D (2007) The contribution of system dynamics to cost benefit analysis – a case study in planning new mental health services in the UK. In: Proceedings of the 2007 System Dynamics Conference, Boston

Books and Reviews

Dangerfield BC (1999) System dynamics applications to european health care issues. J Operat Res Soc 50(4):345–353
Dangerfield BC, Roberts CA (eds) (1994) Health and health care dynamics. Special issue of Syst Dyn Rev 15(3)
Wolstenholme EF (2003) Towards the definition and use of a core set of archetypal structures in system dynamics. Syst Dyn Rev 19(1):7–26

Health Care in the United States, System Dynamics Applications to

Gary Hirsch[1], Jack Homer[2]
[1] Independent Consultant, Wayland, USA
[2] Independent Consultant, Voorhees, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Four Applications
Future Directions
Bibliography

Glossary

Chronic illness A disease or adverse health state that persists over time and cannot in general be cured, although its symptoms may be treatable.

Stock An accumulation or state variable, such as the size of a population.

Flow A rate-of-change variable affecting a stock, such as births flowing into a population or deaths flowing out.

Feedback loop A closed loop of causality that acts to counterbalance or reinforce prior change in a system state.

Definition of the Subject

Health care involves a complex system of interactions among patients, providers, payers, and other stakeholders. This system is difficult to manage in the United States because of its free market approach and relative lack of regulation. System Dynamics simulation modeling is an effective method for understanding and explaining causes of dysfunction in U.S. health care and for suggesting approaches to improving health outcomes and slowing rising costs. Applications since the 1970s have covered diverse areas in health care including the epidemiology of diseases and substance abuse, as well as the dynamics of health care capacity and delivery and their impacts on health. Many of these applications have dealt with the mounting burden of chronic illnesses, such as diabetes. In this article four such applications are described.

Introduction

Despite remarkable successes in some areas, the health enterprise in the United States faces difficult challenges in meeting its primary goal of reducing the burden of disease and injury. These challenges include the growth of the underinsured population, epidemics of obesity and asthma, the rise of drug-resistant infectious diseases, ineffective management of chronic illness [33], long-standing racial and ethnic health disparities [32], and an overall decline in the health-related quality of life [64]. Many of these complex problems have persisted for decades, often proving resistant to attempts to solve them [36].

It has been argued that these interventions fail because they are made in piecemeal fashion, rather than comprehensively and from a whole-system perspective [15]. This compartmentalized approach is engrained in the financial structures, intervention designs, and evaluation methods of most health agencies. Conventional analytic methods are generally unable to satisfactorily address situations in which population needs change over time (often in response to the interventions themselves), and in which risk factors, diseases, and health resources are in a continuous state of interaction and flux [52].

The term *dynamic complexity* has been used to describe such evolving situations [56]. Dynamically complex problems are often characterized by long delays between causes and effects, and by multiple goals and interests that may in some ways conflict with one another. In such situations, it is difficult to know how, where, and when to intervene, because most interventions will have unintended consequences and will tend to be resisted or undermined by opposing interests or as a result of limited resources or capacities.

The systems modeling methodology of System Dynamics (SD) is well suited to addressing the challenges of dynamic complexity in public health. The methodology involves the development of causal diagrams and policy-oriented computer simulation models that are unique to each problem setting. The approach was developed by computer pioneer Jay W. Forrester in the mid-1950s and first described at length in his book *Industrial Dynamics* [11] with some additional principles presented in later works [8,9,10,12]. The International System Dynamics Society was established in 1983, and within the Society a special interest group on health issues was organized in 2003.

SD modeling has been applied to health and health care issues in the U.S. since the 1970s. Topic areas have included:

- Disease epidemiology including work in heart disease [24,40], diabetes [24,34,43], obesity [25], HIV/AIDS [29], polio [57] and drug-resistant pneumococcal infections [28];
- Substance abuse epidemiology covering heroin addiction [37], cocaine prevalence [30], and tobacco reduction policy [50,58];
- Health care capacity and delivery in such areas as population-based HMO planning [21], dental care [20,38], and mental health [38], and as affected by natural disasters or terrorist acts [16,22,41]; and
- Interactions between health care or public health capacity and disease epidemiology [17,18,19,23,27].

Most of these modeling efforts have been done with the close involvement of clinicians and policymakers who have a direct stake in the problem being modeled. Established SD techniques for group model building [60] can help to harness the insights and involvement of those who deal with public health problems on a day-to-day basis.

It is useful to consider how SD models compare with those of other simulation methods that have been applied to public health issues, particularly in epidemio-

logical modeling. One may characterize any population health model in terms of its degree of aggregation, that is, the extent to which individuals in the population are combined together in categories of disease, risk, or age and other demographic attributes. At the most aggregate end of the scale are lumped contagion models [3,35]; more disaggregated are Markov models [13,31,44]; and the most disaggregated are microsimulations at the level of individuals [14,51,63].

The great majority of SD population health models are high or moderately high in aggregation. This is related to the fact that most SD models have a broad model boundary sufficient to include a variety of realistic causal factors, policy levers, and feedback loops. Although it is possible to build models that are both broad in scope and highly disaggregated, experience suggests that such very large models nearly always suffer in terms of their ability to be easily and fully tested, understood, and maintained. In choosing between broader scope and finer disaggregation, SD modelers tend to opt for the former, because a broad scope is generally needed for diagnosing and finding effective solutions to dynamically complex problems [55,56].

The remainder of this article describes four of the System Dynamics modeling applications cited above, with a focus on issues related to chronic illnesses and their care and prevention. The U.S. Centers for Disease Control and Prevention (CDC) estimates that chronic illness is responsible for 70% of all deaths and 75% of all health care costs in the U.S. [5]. The applications discussed below address:

- Diabetes and heart failure management at the community level;
- Diabetes prevention and management from an epidemiological perspective;
- General chronic illness care and prevention at a community level; and
- General chronic illness care and prevention at the national level.

The article concludes with a discussion of promising areas for future work.

Four Applications

Diabetes and Heart Failure Management at the Community Level

Two hours north of Seattle in the state of Washington lies Whatcom County, with a population of about 170 thousand. The county embarked on a major effort to address chronic illness care and was selected by the Robert Wood Johnson Foundation as one of seven sites in a larger chronic care program called Pursuing Perfection [24]. The program initially concentrated on two chronic illnesses as prototypes for improved care: diabetes and congestive heart failure. Both of these illnesses affect millions of people in the U.S. and other countries and exact a heavy toll in terms of direct medical expenditures as well as indirect costs due to disability and premature mortality [2,45,47]. The prevalence of both diseases is growing rapidly as the numbers of people above age 65 increase, and also due to the epidemic rise in obesity, which is a risk factor for both diabetes and heart disease [7,46].

Leaders of the Whatcom County program had two critical needs for making decisions about potential interventions for improving the care of chronic illnesses such as diabetes and heart failure. First, they wanted to get a sense of the overall impact of these interventions on incidence and prevalence of diabetes and heart failure, health care utilization and cost, and mortality and disability rates in the community. Second, they wanted to understand the impact of the various interventions on individual health care providers in the community and on those who pay for care—insurers, employers, and patients themselves. There was a concern that the costs and benefits of the program be shared equitably and that providers who helped produce savings should not suffer a resulting loss of revenue to their businesses.

These analytic needs could not be met with spreadsheet and other models that project impacts in a simple, linear fashion. Interventions in chronic illness do not have simple direct impacts. The aging of the population, incidence of new cases, progression of disease, deaths, and the interventions themselves all create a constantly changing situation. Interventions ideally reduce mortality rates, but this leaves more people with the disease alive and requiring care for years to come.

Figure 1 presents a simplified view of the stock-and-flow structure used in modeling non-insulin-dependent (Type 2) diabetes. The actual model has two separate structures like those shown in Fig. 1, one for the 18-to-64 age group and one for the 65-and-older age group, which are linked by flows of patients turning 65. The model also calculates an inflow of population turning 18, death outflows from each stock based on patient age and stage of illness, and flows of migration into and out of the county. The rectangular boxes in Fig. 1 represent sub-populations with particular characteristics. The arrows signify flows of people from one population group to another (e. g., from uncontrolled to controlled diabetes at a particular stage). Lines from ovals (programmatic interventions such as disease management) to population flows indicate control of or influence on those flows.

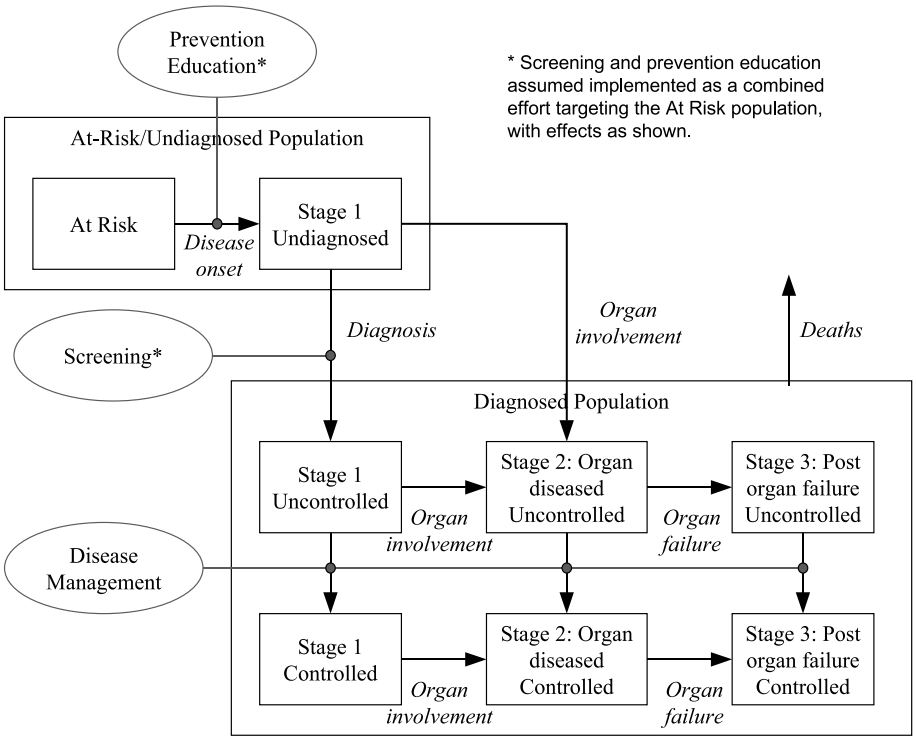

Health Care in the United States, System Dynamics Applications to, Figure 1
Disease stages and intervention points in the Whatcom County Diabetes Model

The three stages of diabetes portrayed in this figure were identified through discussions with clinicians in Whatcom County. The population At Risk includes those with family history, the obese, and, most directly, those with a condition of moderate blood sugar known as prediabetes. Further increases in blood sugar lead to Stage 1 diabetes, in which blood vessels suffer degradation, but there is not yet any damage to organs of the body, nor typically any symptoms of the encroaching disease. More than half of Stage 1 diabetics are undiagnosed. If Stage 1 diabetics go untreated, most will eventually progress to Stage 2, marked by organ disease. In Stage 2 diabetes, blood flow disturbances impair the functioning of organ systems and potentially lead to irreversible damage. A patient who has suffered irreversible organ damage, or organ failure, is said to be in Stage 3; this would include diabetics who suffer heart attacks, strokes, blindness, amputations, or endstage renal disease. These patients are at the greatest risk of further complications leading to death.

Several studies have demonstrated that the incidence, progression, complications, and costs of diabetes can be reduced significantly through concerted intervention [1,4,6,59,61]. Such intervention may include primary prevention or disease management. As indicated in Fig. 1, primary prevention would consist of efforts to screen the at-risk population and educate them about the diet and activity changes they need to prevent progression to diabetes. Disease management, on the other hand, addresses existing diabetics. A comprehensive disease management approach, such as that employed by the Whatcom County program, can increase the fraction of patients who are able to keep their blood sugar under effective control from the 40% or less typically seen without a program up to perhaps 80% or more.

The SD model of diabetes in Whatcom County was first used to produce a 20-year status quo or baseline projection, which assumes that no intervention program is implemented. In this projection, the prevalence of diabetes among all adults gradually increases from 6.5% to 7.5%, because of a growing elderly population; the prevalence of diabetes among the elderly is 17%, compared with 5% among the non-elderly. Total costs of diabetes, including direct costs for health care and pharmaceuticals and indirect economic losses due to disability, grow substantially in this baseline projection.

The next step was to use the model to examine the impact of various program options. These included: (1) a partial approach enhancing disease management but not pri-

mary prevention, (2) a full implementation approach combining enhancement of both disease management and primary prevention, and (3) an approach that goes beyond full implementation by also providing greater financial assistance to the elderly for purchasing drugs needed for the control of diabetes.

Simulations of these options projected results in terms of various outcome variables, including deaths from complications of diabetes and total costs of diabetes. Figure 2 shows typical simulation results obtained by projecting these options, in this case, the numbers of deaths over time that might be expected due to complications of diabetes. "Full VCTIS" refers to the complete program of primary prevention and disease management. Under the status quo projection, the number of diabetes-related deaths grows continuously along with the size of the diabetic population. The partial (disease management only) approach is effective at reducing deaths early on, but becomes increasingly less effective over time. The full program approach (including primary prevention) overcomes this shortcoming and by the end of the 20 year simulation reduces diabetes-related deaths by 40% relative to the status quo. Addition of a drug purchase plan for the elderly does even better, facilitating greater disease control and thereby reducing diabetes related deaths by 54% relative to the status quo.

With regard to total costs of diabetes, the simulations indicate that the full program approach can achieve net savings only two years after the program is launched. Four years after program launch, a drug plan for the elderly generates further reductions in disability costs beyond those provided by the program absent such a plan. The partial program approach, in contrast, achieves rapid net savings initially, but gives back most of these savings over time as diabetes prevalence grows. By the end of 20 years, the full program approach results in a net savings amounting to 7% of the status quo costs, two-thirds of that savings coming from reduction in disability-related costs. The model suggests that these anticipated net savings are the result of keeping people in the less severe stages of the diseases for a longer period of time and reducing the number of diabetes-related hospitalizations.

The simulations provided important information and ideas to the Whatcom County program planners, as well as supporting detailed discussions of how various costs and benefits could be equitably distributed among the participants. This helped to reassure participants that none of them would be unfairly affected by the proposed chronic illness program. Perhaps the most important contribution of modeling to the program planning process was its ability to demonstrate that the program, if implemented in its full form, would likely reduce total costs, even though it would extend the longevity of many diabetics requiring costly care. Given the sensitivity of payers who were already bearing high costs, this finding helped to motivate their continued participation in the program.

Diabetes Prevention and Management from an Epidemiological Perspective

Another SD model of diabetes in the population was developed for the CDC's Division of Diabetes Translation [34]. This model, a structural overview of which is presented in Fig. 3, builds upon the Whatcom County work but looks more closely at the drivers of diabetes onset, including the roles of prediabetes and obesity. The core of the CDC model is a chain of population stocks and flows portraying the movement of people among the stages of normal blood glucose, prediabetes, uncomplicated diabetes, and complicated diabetes. The prediabetes and diabetes stages are further divided among stocks of people whose conditions are diagnosed or undiagnosed. Also shown in Fig. 3 are the potentially modifiable influences in the model that affect the rates of population flow. These flow-rate drivers include obesity and the detection and management of prediabetes and of diabetes.

The model's parameters were calibrated based on historical data available for the U.S. adult population, as well as estimates from the scientific literature. The model is able to reproduce historical time series, some going as far back as 1980, on diagnosed diabetes prevalence, the diagnosed fraction of diabetes, prediabetes prevalence, the obese fractions of people with prediabetes and diabetes, and the health burden (specifically, the mortality, morbidity, and costs) attributable to diabetes. The model suggests that two forces worked in opposition to affect the diabetes health burden from 1980 to 2004. The first force is a rise in the prevalence of obesity, which led to a greater incidence and prevalence of prediabetes and diabetes through the chain of causation seen in Fig. 3. The second and opposing force is a significant improvement in the control of diabetes, achieved through greater efforts to detect and manage the disease. The second force managed to hold the health burden of diabetes more or less flat during 1980 to 2004.

Looking toward the future, a baseline scenario assumes that no further changes occur in obesity prevalence after 2006, and that inputs affecting the detection and management of prediabetes and diabetes remain fixed at their 2004 values through 2050. This fixed-inputs assumption for the baseline scenario is not meant to represent a forecast of what is most likely to in the future but does pro-

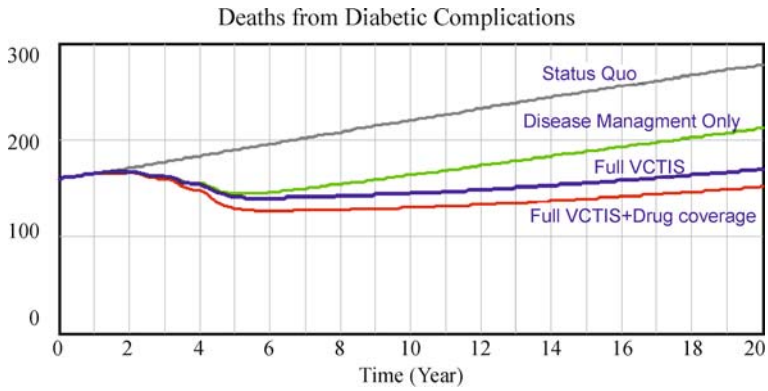

Health Care in the United States, System Dynamics Applications to, Figure 2
Typical results from policy simulations with Whatcom County Diabetes Model

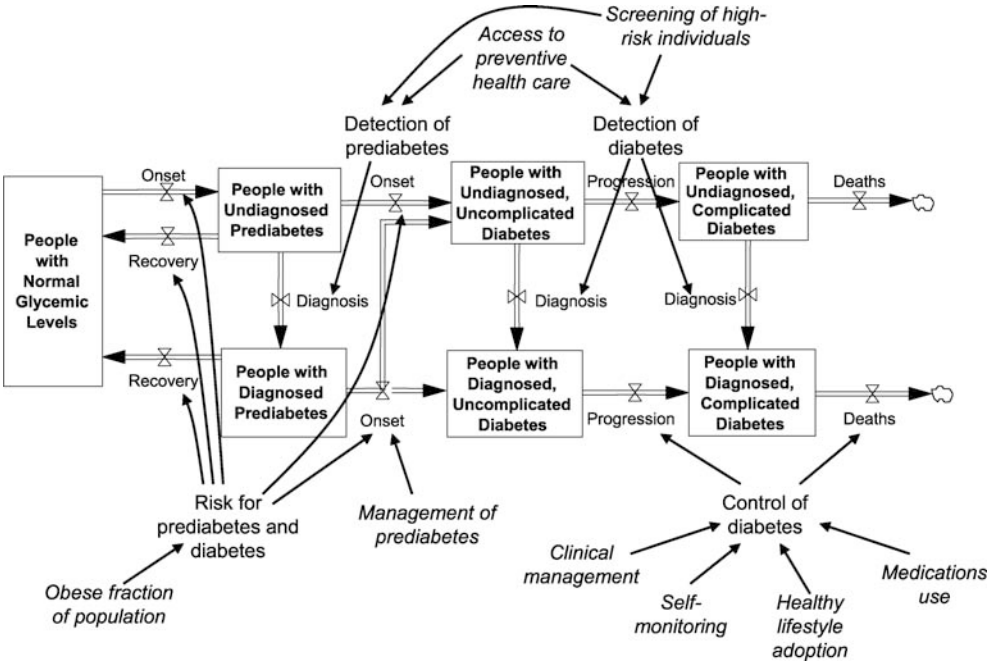

Health Care in the United States, System Dynamics Applications to, Figure 3
Structure of the CDC Diabetes Model

vide a useful and easily-understood starting point for policy analysis.

The baseline simulation indicates a future for diabetes burden outcomes for the period 2004–2050 quite different from the past. With obesity prevalence fixed, by assumption, at a high point of 37% from 2006 onward, the diabetes onset rate remains at a high point as well, and diabetes prevalence consequently continues to grow through 2050, becoming more level (after about 2025) only when the outflow of deaths starts to catch up with the inflow of onset.

The CDC model has been used to examine a variety of future scenarios involving policy interventions (singly or in combination) intended to limit growth in the burden of diabetes. These include scenarios improving the management of diabetes, increasing the management of prediabetes, or reducing the prevalence of general population obesity over time. Enhanced diabetes management can significantly reduce the burden of diabetes in the short term, but does not prevent the growth of greater burden in the longer term due to the growth of diabetes prevalence. Indeed, the effect of enhanced dia-

betes management on diabetes prevalence is not to decrease it at all, but rather to increase it somewhat by increasing the longevity of people with diabetes. Increased management of prediabetes does, in contrast, reduce diabetes onset and the growth of diabetes prevalence. However, it does not have as much impact as one might expect; this is because many people with prediabetes are not diagnosed, and also because the policy does nothing to reduce the growth of prediabetes prevalence due to obesity in the general population. A reduction in prediabetes can be achieved only by reducing population obesity. Significant obesity reduction may take 20 years or more to accomplish fully, but the model suggests that such a policy can be quite a powerful one in halting the growth of diabetes prevalence and burden even before those 20 years are through.

Overall, the CDC model suggests that no single type of intervention is sufficient to limit the growth of the diabetes burden in both the short term and the long term. Rather, what is needed is a combination of disease management for the short term and primary prevention for the longer term. The model also suggests that effective primary prevention may require obesity reduction in the general population a focus on managing diagnosed prediabetes.

At the state and regional level, the CDC model has become the basis for a model-based workshop called the "Diabetes Action Lab". Participants have included state and local public health officials along with non-governmental stakeholders including health care professionals, leaders of not-for-profit agencies, and advocates for people living with diabetes. The workshops have helped the participants improve their intervention strategies and goals and become more hopeful and determined about seeing their actions yield positive results in the future.

The CDC diabetes model has led to other SD modeling efforts at the CDC emphasizing disease prevention, including studies of obesity [25] and cardiovascular risk. The obesity study involved the careful analysis of population survey data to identify patterns of weight gain over the entire course of life from childhood to old age. It explored likely impacts decades into the future of interventions to reduce or prevent obesity that may be targeted at specific age categories. Tentative findings included (1) that obesity in the U.S. should be expected to grow at a much slower pace in the future than it did in the 1980s and 1990s; (2) that the average amount of caloric reduction necessary to reverse the growth of obesity in the population is less than 100 calories per day; (3) that the current trend of focusing intervention efforts on school-age children will likely have only a small impact on future obesity in the adult population; and (4) that it may take decades to see the full impacts of interventions to reduce obesity in the overall population.

General Health Care and Illness Prevention at a Community Level

Hirsch and Immediato [19] describe a comprehensive view of health at the level of a community. Their "Health Care Microworld", depicted in highly simplified form in Fig. 4, simulates the health status and health care delivery for people in the community. The Microworld was created for a consortium of health care providers who were facing a wide range of changes in the mid-1990s and needed a means for their staffs to understand the implications of those changes for how they managed. The underlying SD model consists of many hundreds of equations and was designed to reflect with realistic detail a typical American community and its providers, with data taken from public sources as well as proprietary surveys. Users of the Microworld have a wide array of options for expanding the capacity and performance of the community's health care delivery system such as adding personnel and facilities, investing in clinical information systems, and process redesign. They have a similar range of alternatives for improving health status and changing the demand for care including screening for and enhanced maintenance care of people with chronic illnesses, programs to reduce behavioral risks such as smoking and alcohol abuse, environmental protection, and longer-term risk reduction strategies such as providing social services, remedial education, and job training.

The Microworld's comprehensive view of health status and health care delivery can provide insights not available from approaches that focus on one component of the system at a time. For example, users can play roles of different providers in the community and get a better understanding of why many attempts at creating integrated delivery systems have failed because participating providers care more about their own bottom lines and prerogatives than about creating a viable system. When examining strategies for improving health status, users can get a better sense of how a focus on enhanced care of people with chronic illnesses provides short-term benefits in terms of reduced deaths, hospital admissions, and costs, but how better long-term results can be obtained by also investing in programs that reduce social and behavioral health risks.

General Health Care and Illness Prevention at the National Level

Despite rapid growth in health care spending in the U.S. in recent decades, the health of Americans has not noticeably

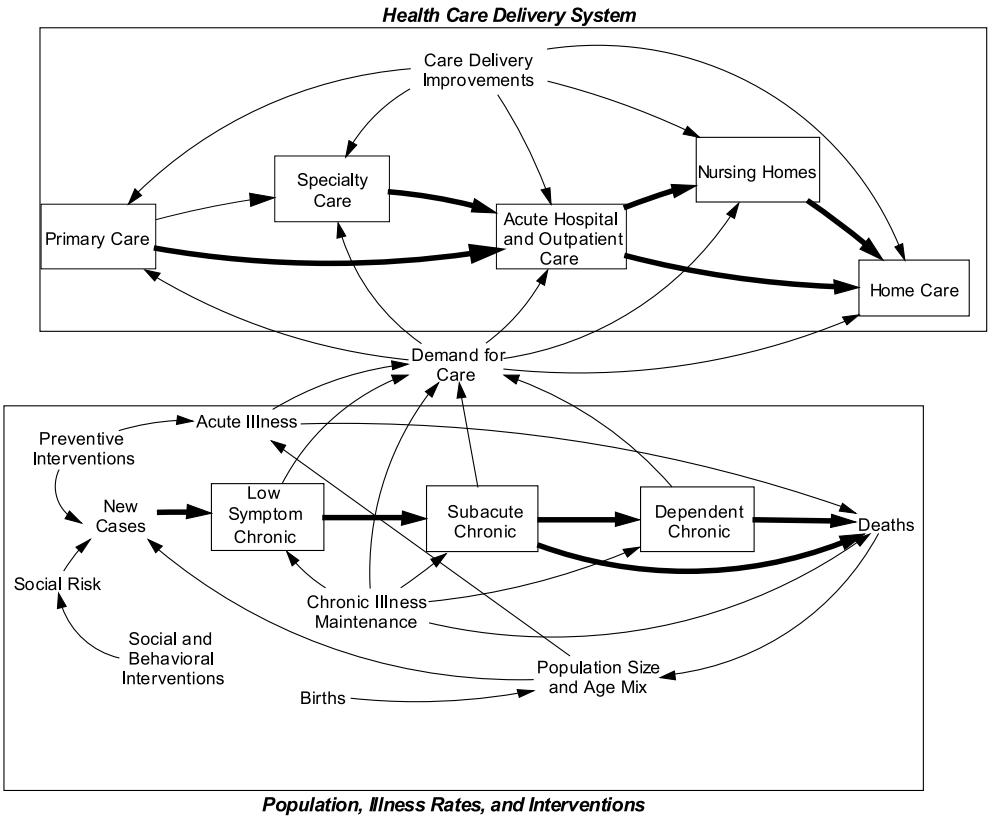

Health Care in the United States, System Dynamics Applications to, Figure 4
Overview of the Health Care Microworld

improved. A recent SD model [23] addresses the question of why the U.S. has not been more successful in preventing and controlling chronic illness. This model can faithfully reproduce patterns of change in disease prevalence and mortality in the U.S., but its structure is a generic one and should be applicable to other countries. The model examines the growing prevalence of disease and responses to it, responses which include the treatment of complications as well as disease management activities designed to slow the progression of illness and reduce the occurrence of future complications. The model shows how progress in complications treatment and disease management has slowed since 1980 in the U.S., largely due to a behavioral tug-of-war between health care payers and providers that has resulted in price inflation and an unstable climate for health care investments. The model is also used to demonstrate the impact of moving "upstream" by managing known risk factors to prevent illness onset, and moving even further upstream by addressing adverse behaviors and living conditions linked to the development of these risk factors in the first place.

An overview of the model's causal structure is presented in Fig. 5. The population stock of disease prevalence is increased by disease incidence and decreased by deaths. The death rate can be reduced by a greater extent of disease care, including urgent care and disease management. Disease incidence draws from a stock of risk prevalence, where risk refers to physical or psychological conditions or individual behaviors that may lead to disease. Effective risk management can reduce the flow of people from risk to disease, and may also in some cases allow people to return to a condition of being no longer at risk. Such management may include changes in nutrition or physical activity, stress management, or the use of medications. The risk prevalence stock is increased by adverse behaviors and living conditions. Adverse behaviors may include poor diet, lack of physical activity, or substance abuse. Adverse living conditions can encompass many factors, including crime, lack of access to healthy foods, inadequate regulation of smoking, weak social networks, substandard housing, poverty, or poor educational opportunities.

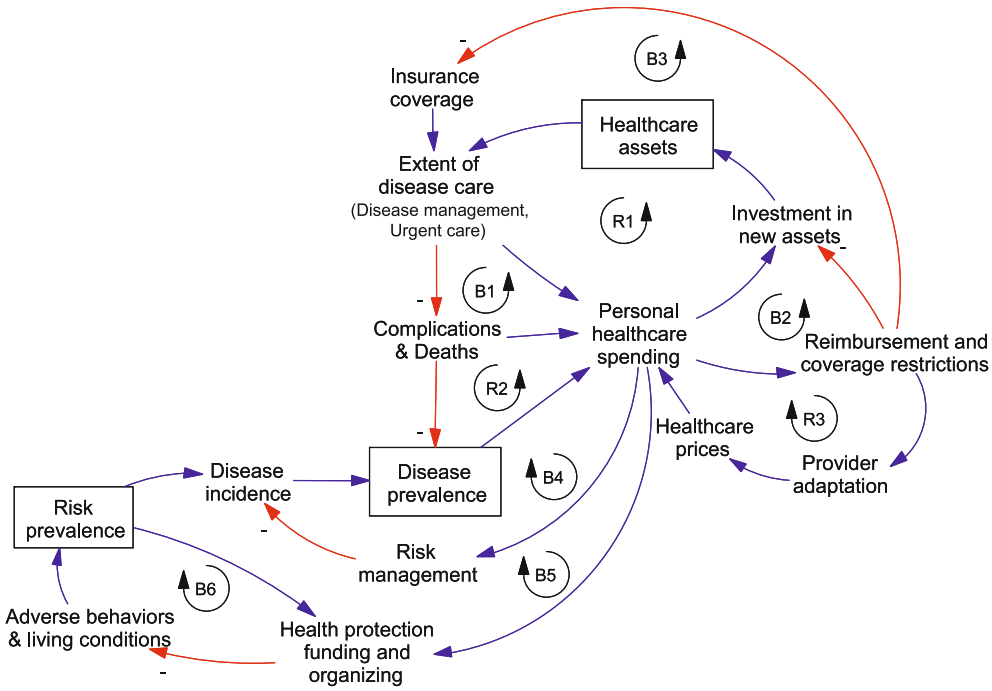

Health Care in the United States, System Dynamics Applications to, Figure 5
Overview of a National-Level Model of Health Care and Illness Prevention. Key to feedback loops ("R" denotes self-reinforcing, "B" denotes counterbalancing):
R1 Health care revenues are reinvested for further growth
B1 Disease management reduces need for urgent care
R2 Disease care prolongs life and further increases need for care
B2 Reimbursement restriction limits spending growth
B3 Insurance denial limits spending growth
R3 Providers circumvent reimbursement restrictions, leading to a tug-of-war with payers
B4 Risk management proportional to downstream spending can help limit it
B5 Health protection proportional to downstream spending can help limit it
B6 Health protection (via sin taxes) proportional to risk prevalence can help limit it

The extent of care is explained in the model by two key factors: the abundance of health care assets, and insurance coverage. Health care assets are the structures and fixed equipment used directly for health care or for the production of health care products, as well as the human capital of personnel involved. Insurance coverage refers to the fraction of the population with some form of health care insurance, either with a private insurer or through a government plan. The uninsured are less likely than the insured to receive health care services, especially disease management services, something which most of the uninsured cannot afford whereas in most cases they can get urgent care at a hospital emergency department.

The stock of assets is increased by investments, which may be viewed as the reinvestment of some fraction of health care revenues. Such reinvestment drives further growth of care and revenue, and the resulting exponential growth process is identified as loop R1 in Fig. 5. The data indicate, however, that the reinvestment process has slowed significantly since 1980. It is hypothesized that this decline in the reinvestment rate has been the response by potential investors to various forms of cost control, including the restriction of insurance reimbursements, which affect the providers of health care goods and services. With increasing controls and restrictions, these potential investors face greater risk and uncertainty about the future return on their investments, and the result is a greater reluctance to build a new hospital wing, or to purchase an expensive new piece of equipment, or even, at an individual level, to devote a decade or more of one's life to the hardship of medical education and training. Health care costs and cost controls have also led to elimination of private health insurance coverage by some employers, although some of the lost coverage has been replaced by publicly-funded insurance.

One additional part of the downstream health care story portrayed in Fig. 5 is the growth of health care prices. Health care prices are measured in terms of a medical care consumer price index (CPI), which since 1980 has grown much more rapidly than the general CPI for the overall economy. For the period 1980–2004, inflation in medical care prices averaged 6.1% versus general inflation of 3.5%. Why has health care inflation exceeded that of the general economy? Several different phenomena have contributed to health care inflation, but not all have contributed with sufficient magnitude or with the timing necessary to explain the historical pattern. One phenomenon that does appear to have such explanatory power is shown in Fig. 5 as "provider adaptation". This is the idea that, in response to cost containment efforts, providers may "increase fees, prescribe more services, prescribe more complex services (or simply bill for them), order more follow-up visits, or do a combination of these…" [49] Many tests and procedures are performed that contribute little or no diagnostic or therapeutic value, thereby inflating the cost per quality of care delivered. By one estimate, unnecessary and inflationary expense may have represented 29% of all personal health care spending in the year 1989 [23].

The dynamics involving the extent of disease care are portrayed in Fig. 5 in the feedback loops labeled R1, B1, R2, B2, B3, and R3. Taken together, one may view these loops—with the exception of Loop R3—as the story of a "rational" downstream health care system that favors growth and investment until the resulting costs get to a point where further increases are perceived to be no longer worth the expected incremental improvements in health and productivity. Loop R3, however, introduces dysfunction into this otherwise rational system. The loop describes a tug-of-war between payers restricting reimbursement in response to high health care costs, and providers adapting to these restrictions by effectively raising health care prices in an attempt to circumvent the restrictions and maintain their incomes. Because this loop persistently drives up health care costs, it ends up hurting health care investments and insurance coverage (through Loops B2 and B3, respectively), thus dampening growth in the extent of care.

Simulations of the model suggest that there are no easy downstream fixes to the problem of an underperforming and expensive health care system in the U.S. mold. The simulations seem to suggest—perhaps counterintuitively—that health insurance should be stable and nonrestrictive in its reimbursements, so as to avoid behavioral backlashes that can trigger health care inflation and underinvestment. Although a broad mandate of this sort would likely be politically infeasible in the U.S., movement in this direction could perhaps start with the government's own Medicare and Medicaid insurance programs, and then diffuse naturally to private insurers over time. It is interesting to consider whether a more generous and stable approach to reimbursement could not only combat illness better than the current restrictive approach, but do it more efficiently and perhaps even at lower cost.

The model also includes structure for evaluating the upstream prevention of disease incidence. There are two broad categories of such efforts described in the literature: Risk management for people already at risk, and health protection for the population at large to change adverse behaviors and mitigate unhealthy living conditions. While spending on population-based health protection and risk management programs has grown somewhat, it still represents a small fraction of total U.S. health care spending, on the order of 5% in 2004 [23].

Figure 5 includes three balancing loops to indicate how, in general terms, efforts in risk management and health protection might be funded or resourced more systematically and in proportion to indicators of capability or relative need. Loop B4 suggests that funding for programs promoting risk management could be made proportional to spending on downstream care, so that when downstream care grows funding for risk management would grow as well. Loop B5 suggests something similar for health protection, supposing that government budgets and philanthropic investments for health protection could be set in proportion to recent health care spending. Loop B6 takes a different approach to the funding of health protection, linking it not to health care spending but to risk prevalence, the stock which health protection most directly seeks to reduce. The linkage to risk prevalence can be made fiscally through "sin taxes" on unhealthy items, such as cigarettes (already taxed throughout the U.S. to varying extents [39]) and fatty foods [42]. In theory, the optimal magnitude of such taxes may be rather large in some cases, as the taxes can be used both to discourage unhealthy activities and promote healthier ones [48].

Simulations of the model suggest that whether the approach to upstream action is risk management or health protection, such actions can reduce illness prevalence and ultimately save money. However, the payback time, in terms of reduced downstream health care costs, may be a relatively long one, perhaps on the order of 20 years. It should be noted, however, that the model does not include losses in productivity to employers and society at large. The Whatcom County models described above suggest that when these losses are taken into account, the payback on upstream action may shrink to a much shorter time period that may be acceptable to the public as well

as to those decision makers in a position to put upstream efforts into effect [24].

Future Directions

As long as there are dynamically complex health issues in search of answers, the SD approach will have a place in the analytic armamentarium. There is still much to be learned about the population dynamics of individual chronic conditions like hypertension and risk factors like obesity. SD models could also address multiple interacting diseases and risks, giving a more realistic picture of their overall epidemiology and policy implications, particularly where the diseases and risks are mutually reinforcing. For example, it has been found that substance abuse, violence, and AIDS often cluster in the same urban subpopulations, and that such "syndemics" are resistant to narrow policy interventions [53,54,62]. This idea could also be extended to the case of mental depression, which is often exacerbated by other chronic illnesses, and may, in turn, interfere with the proper management of those illnesses. An exploratory simulation model has indicated that SD can usefully address the concept of syndemics [26].

There is also more to be learned about health care delivery systems and capacities, with the inclusion of characteristics specific to selected real-world cases. Models combining delivery systems and risk and disease epidemiology could help policymakers and health care providers understand the nature of coordination required to put ambitious public health and risk reduction programs in place without overwhelming delivery capacities. Such models could reach beyond the health care delivery system per se to examine the potential roles of other delivery systems, such as schools and social service agencies, in health risk reduction.

The more complete view of population health dynamics advocated here may also be extended to address persistent challenges in the U.S. that will likely require policy changes at a national and state level, and not only at the level of local communities. Examples include the large underinsured population, persistent racial and ethnic health disparities, and the persistent shortage of nurses. SD modeling can help to identify the feedback structures responsible for these problems, and point the way to policies that can make a lasting difference.

Bibliography

1. American Diabetes Association and National Institute of Diabetes and Digestive and Kidney Diseases (2002) The prevention or delay of Type 2 diabetes. Diabetes Care 25:742–749
2. American Heart Association (2000) 2001 Heart and Stroke Statistical Update. AHA, Dallas
3. Anderson R (1994) Populations, infectious disease and immunity: A very nonlinear world. Phil Trans R Soc Lond B346:457–505
4. Bowman BA, Gregg EW, Williams DE, Engelgau MM, Jack Jr L (2003) Translating the science of primary, secondary, and tertiary prevention to inform the public health response to diabetes. J Public Health Mgmt Pract November (Suppl):S8–S14
5. Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion (2007) Chronic Disease Overview. Available at http://www.cdc.gov/nccdphp/overview.htm
6. Diabetes Prevention Program Research Group (2002) Reduction in the incidence of Type 2 diabetes with lifestyle intervention or metformin. New Engl J Med 346:393–403
7. Flegal KM, Carroll MD, Ogden CL, Johnson CL (2002) Prevalence and trends in obesity among US adults, 1999-2000. JAMA 288:1723–1727
8. Forrester JW, Senge PM (1980) Tests for building confidence in system dynamics models. In: System Dynamics, TIMS Studies in the Management Sciences. North-Holland, New York, pp 209–228
9. Forrester JW (1971) Counterintuitive behavior of social systems. Technol Rev 73:53–68
10. Forrester JW (1980) Information sources for modeling the national economy. J Amer Stat Assoc 75:555–574
11. Forrester JW (1961) Industrial Dynamics. MIT Press, Cambridge
12. Forrester JW (1969) Urban Dynamics. MIT Press, Cambridge
13. Gunning-Schepers LJ (1989) The health benefits of prevention: A simulation approach. Health Policy Spec Issue 12:1–255
14. Halloran EM, Longini IM, Nizam A, Yang Y (2002) Containing bioterrorist smallpox. Science 298:1428–1432
15. Heirich M (1999) Rethinking Health Care: Innovation and Change in America. Westview Press, Boulder
16. Hirsch GB (2004) Modeling the consequences of major incidents for health care systems. In: 22nd International System Dynamics Conference. System Dynamics Society, Oxford. Available from: http://www.systemdynamics.org/publications.htm
17. Hirsch G, Homer J (2004) Integrating chronic illness management, improved access to care, and idealized clinical practice design in health care organizations: A systems thinking approach. In: International Conference on Systems Thinking in Management. AFEI/University of Pennsylvania, Philadelphia. Available from: http://www.afei.org/documents/CDRomOrderForm.pdf
18. Hirsch G, Homer J (2004) Modeling the dynamics of health care services for improved chronic illness management. In: 22nd International System Dynamics Conference. System Dynamics Society, Oxford. Available from: http://www.systemdynamics.org/publications.htm
19. Hirsch GB, Immediato CS (1999) Microworlds and generic structures as resources for integrating care and improving health. Syst Dyn Rev 15:315–330
20. Hirsch GB, Killingsworth WR (1975) A new framework for projecting dental manpower requirements. Inquiry 12:126–142
21. Hirsch G, Miller S (1974) Evaluating HMO policies with a computer simulation model. Med Care 12:668–681
22. Hoard M, Homer J, Manley W, Furbee P, Haque A, Helmkamp J (2005) Systems modeling in support of evidence-based

disaster planning for rural areas. Int J Hyg Environ Health 208:117–125
23. Homer J, Hirsch G, Milstein B (2007) Chronic illness in a complex health economy: The perils and promises of downstream and upstream reforms. Syst Dyn Rev 23(2–3):313–334
24. Homer J, Hirsch G, Minniti M, Pierson M (2004) Models for collaboration: How system dynamics helped a community organize cost-effective care for chronic illness. Syst Dyn Rev 20:199–222
25. Homer J, Milstein B, Dietz W, Buchner D, Majestic E (2006) Obesity population dynamics: Exploring historical growth and plausible futures in the U.S. In: 24th International System Dynamics Conference. System Dynamics Society, Nijmegen. Available from: http://www.systemdynamics.org/publications.htm
26. Homer J, Milstein B (2002) Communities with multiple afflictions: A system dynamics approach to the study and prevention of syndemics. In: 20th International System Dynamics Conference. System Dynamics Society, Palermo. Available from: http://www.systemdynamics.org/publications.htm
27. Homer J, Milstein B (2004) Optimal decision making in a dynamic model of community health. In: 37th Hawaii International Conference on System Sciences. IEEE, Waikoloa. Available from: http://csdl.computer.org/comp/proceedings/hicss/2004/2056/03/2056toc.htm
28. Homer J, Ritchie-Dunham J, Rabbino H, Puente LM, Jorgensen J, Hendricks K (2000) Toward a dynamic theory of antibiotic resistance. Syst Dyn Rev 16:287–319
29. Homer JB, St. Clair CL (1991) A model of HIV transmission through needle sharing. Interfaces 21:26–49
30. Homer JB (1993) A system dynamics model of national cocaine prevalence. Syst Dyn Rev 9:49–78
31. Honeycutt AA, Boyle JP, Broglio KR et al (2003) A dynamic Markov model for forecasting diabetes prevalence in the United States through 2050. Health Care Mgmt Sci 6:155–164
32. Institute of Medicine (Board on Health Sciences Policy) (2003) Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. National Academies Press, Washington, DC
33. Institute of Medicine (Committee on Quality of Health Care in America) (2001) Crossing the Quality Chasm: A New Health System for the 21st Century. National Academies Press, Washington, DC
34. Jones AP, Homer JB, Murphy DL, Essien JDK, Milstein B, Seville DA (2006) Understanding diabetes population dynamics through simulation modeling and experimentation. Am J Public Health 96(3):488–494
35. Kaplan EH, Craft DL, Wein LM (2002) Emergency response to a smallpox attack: The case for mass vaccination. In: Proceedings of the Natl Acad of Sciences 99:10935–10940
36. Lee P, Paxman D (1997) Reinventing public health. Annu Rev Public Health 18:1–35
37. Levin G, Roberts EB, Hirsch GB (1975) The Persistent Poppy. Ballinger, Cambridge
38. Levin G, Roberts EB, Hirsch GB, Kligler DS, Wilder JF, Roberts N (1976) The Dynamics of Human Service Delivery. Ballinger, Cambridge
39. Lindblom E (2006) State cigarette excise tax rates and rankings. Campaign for Tobacco-Free Kids, Washington, DC. Available from: http://www.tobaccofreekids.org/research/factsheets/pdf/0097.pdf
40. Luginbuhl W, Forsyth B, Hirsch G, Goodman M (1981) Prevention and rehabilitation as a means of cost-containment: The example of myocardial infarction. J Public Health Policy 2:1103–1115
41. Manley W, Homer J et al (2005) A dynamic model to support surge capacity planning in a rural hospital. In: 23rd International System Dynamics Conference, Boston. Available from: http://www.systemdynamics.org/publications.htm
42. Marshall T (2000) Exploring a fiscal food policy: The case of diet and ischaemic heart disease. BMJ 320:301–305
43. Milstein B, Jones A, Homer J, Murphy D, Essien J, Seville D (2007) Charting Plausible Futures for Diabetes Prevalence in the United States: A Role for System Dynamics Simulation Modeling. Preventing Chronic Disease, 4(3), July 2007. Available at: http://www\ignorespaces.cdc.gov/pcd/issues/2007/jul/06_0070.htm
44. Naidoo B, Thorogood M, McPherson K, Gunning-Schepers LJ (1997) Modeling the effects of increased physical activity on coronary heart disease in England and Wales. J Epidemiol Community Health 51:144–150
45. National Institute of Diabetes and Digestive and Kidney Diseases (2004) National Diabetes Statistics. Available from: http://diabetes.niddk.nih.gov/dm/pubs/statistics/index.htm
46. National Institute of Diabetes and Digestive and Kidney Diseases (2004) Statistics Related to Overweight and Obesity. Available from: http://www.niddk.nih.gov/health/nutrit/pubs/statobes.htm#preval
47. O'Connell JB, Bristow MR (1993) Economic impact of heart failure in the United States: Time for a different approach. J Heart Lung Transplant 13(suppl):S107-S112
48. O'Donoghue T, Rabin M (2006) Optimal sin taxes. J Public Econ 90:1825–1849
49. Ratanawijitrasin S (1993) The dynamics of health care finance: A feedback view of system behavior. Ph.D. Dissertation, SUNY Albany
50. Roberts EB, Homer J, Kasabian A, Varrell M (1982) A systems view of the smoking problem: Perspective and limitations of the role of science in decision-making. Int J Biomed Comput 13:69–86
51. Schlessinger L, Eddy DM (2002) Archimedes: A new model for simulating health care systems—the mathematical formulation. J Biomed Inform 35:37–50
52. Schorr LB (1997) Common Purpose: Strengthening Families and Neighborhoods to Rebuild America. Doubleday/Anchor Books, New York
53. Singer M, Clair S (2003) Syndemics and public health: Reconceptualizing disease in bio-social context. Med Anthropol Q 17:423–441
54. Singer M (1996) A dose of drugs, a touch of violence, a case of AIDS: Conceptualizing the SAVA syndemic. Free Inquiry 24:99–110
55. Sterman JD (1988) A skeptic's guide to computer models. In: Grant L (ed) Foresight and National Decisions. University Press of America, Lanham, pp 133–169
56. Sterman JD (2000) Business Dynamics: Systems Thinking and Modeling for a Complex World. Irwin/McGraw-Hill, Boston
57. Tompson KM, Tebbens RJD (2007) Eradication versus control for poliomyelitis: An economic analysis. Lancet 369:1363–1371. doi:10.1016/S0140-6736(07)60532-7
58. Tengs TO, Osgood ND, Chen LL (2001) The cost-effectiveness of intensive national school-based anti-tobacco education: Results from the tobacco policy model. Prev Med 33:558–70

59. U.K. Prospective Diabetes Study Group (1998) Tight blood pressure control and risk of macrovascular and microvascular complications in Type 2 diabetes. The Lancet 352:703–713
60. Vennix JAM (1996) Group Model-building: Facilitating Team Learning Using System Dynamics. Wiley, Chichester
61. Wagner EH, Sandhu N, Newton KM et al (2001) Effect of improved glycemic control on health care costs and utilization. JAMA 285(2):182–189
62. Wallace R (1988) A synergism of plagues. Environ Res 47:1–33
63. Wolfson MC (1994) POHEM: A framework for understanding and modeling the health of human populations. World Health Stat Q 47:157–176
64. Zack MM, Moriarty DG, Stroup DF, Ford ES, Mokdad AH (2004) Worsening trends in adult health-related quality of life and self-rated health–United States, 1993-2001. Public Health Rep 119:493–505

Hierarchical Dynamics

Martin Nilsson Jacobi
Complex Systems Group, Department of Energy and Environment, Chalmers University of Technology, Gothenburg, Sweden

Article Outline

Glossary
Definition
Introduction
Overview
Temporal Hierarchies: Separation of Time Scales
Structural Hierarchies: Foliations
Conclusion
Future Directions
Acknowledgment
Bibliography

Glossary

In this section some definitions and results that are important as background for the later exposition are summarized. Out of necessity the definitions are very short and the reader might want to consult standard textbooks on the subjects. This is especially true in the section on group theory, which is dense and only meant to recapitulate main results to a reader already familiar with the basics of the subject. It would have been appropriate to also include a brief background on differential geometry. Unfortunately this subject is so large that even a minimal introduction must span several pages. The reader is advised to consult a standard book on differential geometry and Lie group theory, e. g. [4], to find the necessary definitions and results.

Liouville's Theorem, Conservative and Dissipative Systems

Consider a dynamical system $\dot{x} = f(x)$ with a phase space $x \in \mathbb{R}^n$. Let $\rho(t,x)$ be a probability density on the phase space, defined so that $\rho(t,x)d^n x$ is the probability of finding the system in the phase space volume $d^n x$ at time t. Given an initial value, the differential equation has a unique solution. This observation results in the following continuity equation

$$\frac{\partial \rho}{\partial t} + \sum_i \frac{\partial (f_i \rho)}{\partial x_i} = 0 \,, \tag{1}$$

for the density under the flow f. In general, we may write the time evolution of a probability density as

$$\frac{\partial \rho}{\partial t} = -\sum_i \frac{\partial (f_i \rho)}{\partial x_i} = -\mathcal{L}\rho \,, \tag{2}$$

where $\mathcal{L}$ is called the Liouville operator. In quantum mechanics the convention $\frac{\partial \rho}{\partial t} = -i\mathcal{L}\rho$ is often used to ensure that $\mathcal{L}$ is a Hermitian operator. We do not use this convention here. The evolution of the probability density along a trajectory is given by the total time derivative:

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = \frac{\partial \rho}{\partial t} + \sum_i f_i \frac{\partial \rho}{\partial x_i} = -(\nabla \cdot f)\rho \,, \tag{3}$$

where Eq. (1) is used in the last step. The factor $\nabla \cdot f$ measures the phase space contraction under the flow of the dynamical system. If the system is Hamiltonian, then the degrees of freedom are given by x_i and p_i and there exist a function $H(x,p,t)$ (the Hamiltonian) such that the dynamics can be written on the form:

$$\dot{x}_i = f_i(x,p) = -\frac{\partial H}{\partial p_i} \,, \quad \dot{p}_i = f_{i+d}(x,p) = \frac{\partial H}{\partial x_i} \,,$$

$i = 1, \ldots, d$. It then follows that

$$\begin{aligned} \nabla \cdot f &= \sum_{i=1}^{d} \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^{d} \frac{\partial f_{i+d}}{\partial p_i} \\ &= -\sum_{i=1}^{d} \frac{\partial^2 H}{\partial x_i \partial p_i} + \sum_{i=1}^{d} \frac{\partial^2 H}{\partial p_i \partial x_i} = 0 \,. \end{aligned}$$

We conclude that $\frac{\mathrm{d}\rho}{\mathrm{d}t} = 0$ for a Hamiltonian system, i. e. the phase space volume is conserved. This result is called Liouville's theorem and it is valid for closed Hamiltonian systems. We also note that, with the standard scalar product $(f,g) = \int \mathrm{d}x \overline{f(x)} g(x)$, the Liouville operator is antisymmetric ($\mathcal{L}^\dagger = -\mathcal{L}$) for Hamiltonian systems.

If $\nabla \cdot f < 0$ in some region of the phase space the phase space volume is contracting, we say that the system is dissipative. A consequence of a contracting phase space is that there exists an attractor. An attractor is a subset to which trajectories from some regions of initial values evolve asymptotically. For chaotic dissipative systems, the attractor is usually not smooth and can have non-integer Hausdorff dimension. By contrast, a conservative system do not have any attractors.

The Langevin and Fokker–Planck Equations

One of the simplest stochastic differential equations is the Langevin equation

$$\dot{v} = -\gamma v + \zeta(t) , \tag{4}$$

where $\zeta(t)$ represents white noise with zero mean $\langle \zeta(t) \rangle_t = 0$ and variance defined as $\langle \zeta(t)\zeta(t + \mathrm{d}t) \rangle_t = 2B\delta(\mathrm{d}t)$ (δ denotes a Dirac delta function). The first moments of v can be shown to be [55]:

$$\langle v(t) \rangle = 0 \quad \langle v(t)^2 \rangle = B/\gamma .$$

If v is interpreted as velocity, then by the definition of temperature $\langle v(t)^2 \rangle = k_B T$ (assuming unit mass) and the fluctuation-dissipation theorem follows from the variance:

$$B = \gamma k_\mathrm{B} T . \tag{5}$$

For a deterministic dynamical system, the time evolution of a probability density on the phase space is given by Liouville's equation (2). For a stochastic differential equation, such as the Langevin equation, the time evolution of a probability density is given by a Fokker–Planck equation:

$$\frac{\partial \rho}{\partial t} = -\gamma \nabla \cdot \rho + B \Delta \rho .$$

Note that the stochastic part of the dynamics shows up as a diffusion term. In general, for a stochastic differential equation on the form

$$\dot{x}_i = f_i(x) + \zeta_i(t) , \tag{6}$$

with $\langle \zeta_i(t) \rangle = 0$ and

$$\langle \zeta_i(t_1)\zeta_j(t_2) \rangle = 2B_{ij}(x)\delta(t_1 - t_2) ,$$

the corresponding Fokker–Planck equation reads

$$\frac{\partial \rho}{\partial t} = -\sum_i \frac{\partial}{\partial x_i}(f_i \rho) + \sum_{ij} \left(\frac{\partial}{\partial x_i} B_{ij} \frac{\partial}{\partial x_j} \right) \rho . \tag{7}$$

It is important to be aware of subtleties with the interpretation of stochastic differential equations. The fundamental problem lies in the ambiguous representation of the noise $\zeta(t)$, which is not a regular function (it is not even continuous). There are essentially two different interpretations of (6), called Itō and Stratonovich. They lead to different forms of the diffusion term in Eq. (7). This issue are not of central concern to the current presentation and we do not discuss it further. The interested reader is recommended to read two standard references on the Fokker–Planck equation: Risken & Frank and Gardiner [13,41].

Ergodicity and Mixing

Let (X, Σ, μ) be a probability space, where X is the space (or set), Σ is a σ-algebra, and μ is a probability measure on Σ. Consider a map T that is measure preserving, $\mu(T^{-1}(E)) = \mu(E)$ for any $E \in \Sigma$. The map T is called ergodic if, for almost all $x \in X$ and any Lebegue measurable function f,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i x) = \int f \, \mathrm{d}\mu ,$$

where μ is an invariant measure. Alternatively one may define ergodic as follows: whenever $T(E) = E$, i. e. E is an invariant measure, then it follows that either $\mu(E) = 0$ or $\mu(E) = 1$. Intuitively ergodicity means that we can study the map's properties by looking at a single trajectory. The definition of ergodicity for continuous dynamical system is analogous.

A measure preserving map T is called (strongly) mixing if

$$\lim_{n\to\infty} \mu\left(T^n(A) \cap B\right) = \mu(A)\mu(B),$$

where A, B are measurable sets. Intuitively, mixing means that different sets of initial distributions become intertwined with each other as the system evolves. The generic metaphor is a drop of dye in a glass of water. Mixing systems are always ergodic but the converse is not necessarily true. For example, an irrational flow on a torus is ergodic but not mixing. The mixing property is strongly connected with chaotic motion. The exponential stretching and folding of the phase space associated with nonzero Lyaponov exponents results in a mixing behavior for chaotic dynamics.

Some Concepts from Group Theory

Group A group is a set G endowed with a binary operation $*\colon G \times G \to G$. We often denote the group

by $(G, *)$, or if there is no risk for confusion just G. The operator $*$ fulfills three group axioms: Associativity $a * (b * c) = (a * b) * c$; Identity element, there exist a unique element $e \in G$ such that $\forall a \in G \quad a * e = e * a = a$; Inverse, for each element $a \in G$ there exist a unique element $a^{-1} \in G$ such that $a * a^{-1} = a^{-1} * a = e$.

Subgroup A subset H of G is called a subgroup if it is closed under the group operation ($\forall a, b \in H$ it follows that $a * b \in H$), and if H is a group in itself ($e \in H$ and $\forall a \in H$ it follows that $a^{-1} \in H$).

Classes We say that a and a' are said to be conjugates if $\exists b \in G$ such that $a' = b^{-1} * a * b$. Conjugacy is an equivalence relation and each element $a \in G$ belong to exactly one class. We call the resulting group a direct product of G_1 and G_2, denoted by $G_1 \otimes G_2$. Note that G_1 and G_2 are both normal subgroups of $G_1 \otimes G_2$.

Normal subgroup If N is a subgroup of G and furthermore $\forall b \in N$ and $\forall a \in G$ it is true that $a * b = b' * a$ for some $b' \in N$, then N is called a normal subgroup of G. A normal subgroup can be identified with a union of classes.

Direct product Let $(G_1, *)$ and $(G_2, \cdot)$ be two groups. We can form ordered pairs (a_1, a_2) where $a_1 \in G_1$ and $a_2 \in G_2$. We can then define a group product on the ordered pair as $(a_1, a_2) \circ (a_1', a_2') = (a_1 * a_1', a_2 \cdot a_2')$.

Coset and quotient group If H is a subgroup of G, a left (right) coset of an element $a \in G$ is defined as $aH = \{a * b \colon b \in H\}(Ha = \{b * a \colon b \in H\})$. The collection of $aH(Ha)\forall a \in G$ are called the left (right) cosets of H in G. The left and right cosets of H in G coincide ($aH = Ha$) if and only if H is a normal subgroup of G. In fact the importance of normal subgroups comes from a possibility to define a quotient group (or factor group) G/N. The elements in the quotient group G/N can be identified with the cosets of N in G. The group operation $\cdot$ on G/N is defined as $(aN) \cdot (bN) \doteq (a * b)N$ (note that this definition makes sense since $Nb = bN$).

Homomorphism, isomorphism, and automorphism Let $(G, *)$ and $(F, \cdot)$ be two groups. A map $\phi \colon G \to F$ is called a homomorphism if $\phi(a * b) = \phi(a) \cdot \phi(b)$ for all $a, b \in G$. It follows directly that $\phi(e_G) = e_F$ and $\phi(a^{-1}) = \phi(a)^{-1}$. If the map ϕ is also bijektive, it is called an isomorphism and in this case G and F are considered "essentially the same", denoted $G \simeq F$. If $\Psi \colon G \to G$ is an isomorphism from G to itself, then Ψ it is called an automorphism. The kernel of an homomorphism $\phi \colon G \to F$, denoted $\ker(\phi)$ is defined as the pre-image of the identity element in F, i. e. $\ker(\phi = \{a \in G \colon \phi(a) = e_F\}$. In general, the pre-image of any subgroup in F is a subgroup in G. The pre-image of the trivial subgroup e_F, i. e. $\ker(\phi)$, is also a normal subgroup in G (this follows from $a \in \ker(\phi)$ gives $\phi(b * a * b^{-1}) = \phi(b) \cdot \phi(a) \cdot \phi(b^{-1}) = \phi(b) \cdot e_F \cdot \phi(b^{-1}) = \phi(e_G) = e_F$). With these definitions we can state the fundamental theorem of isomorphisms: $\phi(G) \simeq G/\ker(\phi)$.

Semi-direct product Let $(N, *)$ and $(H, \cdot)$ be groups and $\Phi \colon H \to \mathrm{Aut}(N)$ map from H to the set of automorphisms on N, we use the notation $\Psi(h)(\cdot) \doteq \Psi_h(\cdot)$. A semi-direct product of N and H with respect to Ψ, denoted $N \rtimes_\Psi H$, is defined by the set of pairs (n, h) where $n \in N$ and $h \in H$, and the group operator defined by $(n_1, h_1) \circ (n_2, h_2) = (n_1 * \Psi_{h_1}(n_2), h_1 \cdot h_2)$. N is a normal subgroup of $N \rtimes_\Psi H$ (if H is also a normal subgroup then $N \rtimes_\Psi H \simeq N \otimes H$). Conversely, if N is a normal subgroup and H a subgroup of G such that for each $g \in G$, $g = n \circ h$ for some $n \in N$ and $h \in H$ and $N \cap H = e_G$, then $G \simeq N \rtimes_\Psi H$ with $\Psi_h n = h \circ n \circ h^{-1}$. The normal subgroup can be eliminated by the quotient group construction $(N \rtimes_\Psi H)/N \simeq H$.

Semigroup A semigroup is a set S with an associative binary operation $\circ$. A semigroup can trivially always be extended with an identity element e by considering $e \cup S$, and let $e \circ a = a \circ e = a \forall a \in S$. It is the absence of inverse elements that makes semigroups fundamentally different from groups. A dynamical system can be viewed as a transformation semigroup acting on a state space (see below). semigroups are also important in the study of finite automata (e. g. Krohn–Rhodes decomposability theorem [24]) and theoretical computer science in general.

Lie group and Lie algebra A Lie group is a d dimensional differentiable manifold with a group structure. Both the group operation $* \colon G \times G \to G$ and the inversion map $i \colon G \to G$, $i(a) = a^{-1}$ are smooth maps between manifolds. The vector fields spanning the tangent space at the identity element in a Lie group spans an algebra, called a Lie algebra $\mathfrak{g} \simeq TG|_e$. The vector fields in the Lie algebra are characterized by their invariance under group multiplication (from the right or left). The Lie algebra is then a vector space together with a bilinear operator $\mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$, called the Lie bracket. The Lie bracket is defined as the commutator of two vector fields $[\mathbf{v}, \mathbf{w}] = \mathbf{v}(\mathbf{w}) - \mathbf{w}(\mathbf{v})$. It is clear that the Lie algebra is closed under the Lie bracket since the vector fields span a tangent space (they must form an involution). Any group element $g \in G$, connected to the identity element, can be expressed in terms of a finite succession of

exponentials of the vector fields in the Lie algebra $g = e^{\epsilon_1 \mathbf{v}_{i1}} e^{\epsilon_2 \mathbf{v}_{i2}} \cdots e^{\epsilon_d \mathbf{v}_{id}}$, for some ϵ_j. We therefore view the Lie algebra as a generator of the Lie group. Most properties of the Lie group can be analyzed in terms of the Lie algebra. These concepts are similar in spirit to how all global information about an analytic function is contained in a local Taylor expansion.

Transformation group A group G with an action ψ such that G can act on a set M, $\psi : G \times M \to M$ is called a transformation group. Examples of transformation groups are rotations acting on $\mathbb{R}^n$. Let $a, b \in G$ and $x \in M$. A transformation group action then fulfills the following properties: $\psi(a, \psi(b, x)) = \psi(a * b, x)$ and $\psi(e, x) = x$. An orbit of a point $x \in M$ is a set in M that is invariant under the action ψ. Let $\mathcal{O}$ denote an orbit then if $x \in \mathcal{O}$ it follows that $\psi(a, x) \in \mathcal{O} \forall a \in G$. It is clear that an element $x \in M$ belongs to exactly one orbit, i. e. the orbits form a partition of M through their associated equivalence classes. A group action is called transitive if there is only one orbit covering the entire set M. The set of all orbits is defined as a quotient set (or quotient manifold if M is a manifold), denoted by M/G. In the case when G is a Lie group and M is a smooth manifold, the regularity of the quotient manifold M/G is not guaranteed unless the following extra conditions are assumed for the group action: ψ is regular. Semi-regular means that all orbits have the same dimension as submanifolds in M. Regular means that the action is semi-regular, and in addition for each $x \in M$ there exists a neighborhood U of x such that each orbit intersecting U is a pathwise connected subset.

Finite state automata A finite state automaton is defined as a triplet $\mathcal{A} = (A, X, \delta)$ where A is a finite (nonempty) state set, X is an input alphabet, and $\delta : X \times A \to A$ is a transition function. There is an optional output alphabet but this could be identified with A. The transition function can be extended to act on words: $\delta(uv, a) = \delta(v, \delta(u, a))$ etc. The set of all words formed out of the alphabet X is denoted by X^*, and the set of nonempty finite words are denoted by X^+. There is a natural equivalence defined on $X^* : U \sim V$ if $\delta(U, a) = \delta(V, a)$ for all $a \in A$ (where $U, V \in X^*$). The equivalence classes $X^+/\sim$ together with the concatenation operation form the characteristic semigroup of the automaton $S(\mathcal{A})$. The characteristic semigroup $S(\mathcal{A})$ acting on A with the action δ defines a transformation semigroup. In this way any finite automaton can be analyzed as a transformation semigroup.

Definition

The concept of hierarchies is often used in our descriptions of the world. Sometimes the hierarchical structure is primarily a construct of our mind, and can therefore be considered subjective. Examples are the Linnean taxonomy, Chomsky's linguistic hierarchies, hierarchies in object oriented programming, and hierarchical structures within human organizations. In other situations, one may argue that the physical world, external to the human mind (as far as something of that nature actually exists), is objectively organized in a hierarchical way. Examples of the latter situation can be physical hierarchies of particles and length/time scales, evolutionary taxonomies in biology, hierarchical organization within organisms (cells versus organs), ecosystems (food chains), and computational complexity classes in theoretical computer science. The distinction between subjective and objective is provocative. Not only is it unclear if the distinction actually makes sense in all relevant cases, it is in fact problematic to argue sharply in any of the examples given. The Linnean taxonomy can, for example, be argued for as objective since it is based on important morphological differences and similarities between the organisms. The characteristics used are also central for the function, which is often reproduction. Is it then fair to to say that the result is just a reflection of the human visual cortex combined with higher cognitive functions in the human brain? On the other side, is not our description of the universe in terms of physical law, expressed as mathematical relations, subjective *in absurdum*? However, the important difference between the subjective and objective hierarchies listed above lies in whether or not the hierarchical structure is pre-assumed a priori to the model building, and included as a central part of the description of the system; or, as in the hierarchies listed as objective, a posteriori result derived from a model of reality that does not pre-assume the existence of a hierarchical organization (we may say that the discovery of a hierarchy includes some element of surprise).

While the idea of using hierarchies to describe the world goes back to Aristotle, Simon was one of the first to discuss hierarchical organization as a key ingredient in evolvable and/or controllable complex systems [51]. The important insight was to acknowledge how hard it is to make a complex system robust and adaptable if it is not organized hierarchically. For a system to remain robust as the overall complexity increases, the internal complexity must be confined by recursively "hiding" internal degrees of freedom into modules. The modules then communicate with each other only through narrow channels. In this way, a combinatorial explosion of complexity stemming from

"everyone talking to everyone about everything" can be avoided. In conclusion, hierarchies through modularity is a central theme when trying to combine complexity with robustness, controllability, and evolvability.

Understanding hierarchal organization in dynamical systems may also lead to insights into mechanisms behind emergence in systems that are, at least in principle, reductionistic. Emergence is one of the central themes in complex systems science. Methods and ideas used for analyzing hierarchical dynamics are likely to play a central role in any theoretical framework addressing emergence. Finally, many of the techniques that we discuss in the context of hierarchical dynamics originated in the related fields of model reduction and multi-scale simulation. In model reduction one typically seeks systematic methods for reducing the complexity of a specific model, or a class of models, thereby making them more manageable in terms of simulation or analytic studies. As we will see, this is a special case of hierarchical dynamics (a case of great practical and conceptual importance).

Introduction

Technically, a definition of a hierarchy must contain some measure that defines the levels, i. e. the objects in the hierarchy must be ordered according to their complexity, or some other measure that defines the hierarchy. In this presentation we focus on hierarchies in dynamical systems. The most natural measure to form a hierarchy around in this case is the dimensionality of the phase space, or in the discrete case the cardinality of the state space. Furthermore, it is natural to assume an upward causality on the hierarchy based on projections from a higher dimensional phase space (or if the system is discrete state space of higher cardinality) to a lower dimensional phase space (state space of lower cardinality). An interesting exception is the renormalizing projections used to analyze critical phenomena. It builds a self-similar hierarchy without reducing the cardinality of the state space. This is possible in an infinite system (the thermodynamic limit) since the cardinality of e. g. the natural numbers does not change when half of the states are eliminated. Returning to the discussion on projections that do reduce the dimensionality of the phase space. To avoid arbitrary projections from defining new hierarchical levels, we must also require the resulting dynamics to be closed. Closure means that the dynamics on the higher level (lower dimensionality) is self-contained, i. e. the evolution of the system can be effectively described using only its internal degrees of freedom, not information from the lower levels in the hierarchy. However, this definition is still not strict enough. We must also require that the resulting system is a first order ordinary differential equation (it it is a continuous time system) and Markovian, which means that the future evolution of the system depends only on the current state, not on the history of the system. To demonstrate the importance of the Markov requirement, we consider a dynamical system on the form

$$\dot{x} = f(x, y)$$
$$\dot{y} = g(x)y\,.$$

Formally we can solve the second equation for y and put the result into the first equation. The result reads

$$\frac{\partial x}{\partial t} = f\left(x, y_0 e^{\int_0^t \mathrm{d}s\; g(x(s))}\right)\,.$$

We end up with a one-dimensional equation but with an infinite memory term. This reduction corresponds to using the projection $\pi(x, y) = x$ (detailed explanation of the meaning of this projection is given in Sect. "Structural Hierarchies: Foliations"). The exact form of the right hand side of $\dot{y}$ was chosen for simplicity and clarity. In principle the result generalizes, at least locally, to generic functional forms $\dot{y} = g(x, y)$. The projection in this example was arbitrary and did not reflect any structure in the dynamical system. It is clear that such projections are not interesting when defining a hierarchy. By requiring the projected system to be Markovian we restrict attention to projections that do reflect hierarchical organization of the dynamics.

The fact that no information is lost by a generic projection of a dynamical system is useful in practice. Delay-time embedding for attractor reconstruction is for example based on this observation [38,45,53]. In our context we are interested in projections that actually do hide some information on the lower levels in the hierarchy. Takens' embedding theorem states that such projections, if they exist, constitute singular (with regards to a some natural measure) points in the space of all possible projections. The purpose here is to identify the constraints that have these singular projections as solutions. We conclude the introduction by repeating the central theme that will be our guide throughout this presentation:

Definition 1 Each level in a hierarchy should be a self-contained Markovian dynamical system.

For a more extensive discussion on objectivity in hierarchical dynamics and the Markov property see [49].

Overview

There are mainly two types of hierarchies in dynamical systems: structural and temporal hierarchies. Temporal hi-

erarchies are defined by separation of time scales between the different levels. The local nature of physical interactions connects time scales to length scales. As a consequence, the levels in temporal hierarchies are also often associated with separation in length scales. One may turn the argument around and claim that our choice of metric is a reflection of how the interactions between objects in the universe behave. In any case the result is the same, there is a tight coupling between time and length scales. Structural hierarchies are derived from geometric properties of the dynamics. These geometric properties stems from decomposability, or skew-product decomposability (to be defined later), of the vector field defining the flow. Simple examples of structural hierarchies are non-interacting subsystems or systems with constants of motion.

Temporal hierarchies are discussed extensively in the literature. The mechanisms behind structural hierarchies are also very carefully analyzed in the context of classical mechanics and quantum mechanics. In classical mechanics, the connection between symmetries of the Hamiltonian and invariants of the motion was clarified by Noether's theorem. In quantum mechanics one is often interested in the result of composing multiple particles with certain symmetries into a single object with a larger symmetry group. The general mathematical setting for modern gauge theory is fiber bundles. As we discuss in Sect. "Structural Hierarchies: Foliations" this is also the natural framework for working with structural hierarchies. In this presentation however, our aim is not to clarify the connection between modern theoretical physics and hierarchical dynamics, but rather to show the connection between temporal and structural hierarchies in dynamical systems. This connection is usually not emphasized in the literature.

Temporal Hierarchies: Separation of Time Scales

Dynamical systems with high dimensionality often display dynamics on vastly different time scales. As a consequence, when the system is analyzed at a specific time scale, some degrees of freedom evolve so slowly that they can effectively be treated as (adiabatic) constants, while other degrees of freedom evolve very fast compared to the time scale of interest. We wish to systematically eliminate both the very fast and the very slow degrees of freedom. The fast dynamics must be treated differently depending on its characteristics: it can be represented by its average influence; it can be treated as white or colored noise; or its dynamics is dissipative and the dynamics quickly relaxes to an adiabatic fixed point. For a review on model reduction in dynamical systems with time scale separation, see Givon et al. [14].

Elimination of Slaved Degrees of Freedom

Self-organization can be defined as the tendency for a system to increase internal order without influence from the outside. From a more technical perspective, self-organization is a result of a collapse of the phase space volume. As a result the effective dimensionality of the system's phase space is reduced. As was discussed in Sect. "Liouville's Theorem, Conservative and Dissipative Systems", a shrinking volume of phase space elements in the Liouville equation is a result of energy dissipation, or "friction". It is clear that a closed physical system cannot be self-organizing since this would break the first and second law of thermodynamics, as well as Liouville' theorem that ensures conserved phase space volume for Hamiltonian systems. Self-organizing systems are open, and often kept in a out-of-equilibrium, but stationary, state by external energetic driving. A simple example of a driven dissipative systems is a forced damped pendulum:

$$\ddot{\theta} + \gamma\dot{\theta} + \sin(\theta) = A\sin(\omega t + \phi)\,. \tag{8}$$

If we re-write this system as a first order differential equation we can show that $\nabla \cdot f = -\gamma$. Equation (3) then implies that the phase space volume shrinks exponentially. The attractor for Eq. (8) is a limit cycle. Note that the shrinking of the phase space volume only depend on the dissipation, not on the driving on the right hand side. Naively one could have expected that the driving would tend to expand the phase space volume, but this is not the case. Liouville's theorem holds true also for mechanical systems with time dependent Hamiltonians.

A generic feature of driven dissipative systems is that fast degrees of freedom, due to large negative exponent associated with dissipation, often relaxes to an adiabatic fixed point, i. e. a point in the phase space that appears effectively fixed on the time scale of the fast dynamics but that changes on the time scale set by the slow degrees of freedom. The overall dynamics is therefore slaved to a slow positively invariant, or inertial, manifold (or more correctly attractor) and the resulting dimensionality is reduced. It should be noted that the geometry of the reduced system is often very complicated, taking e. g. the form of a strange attractor [42]. This picture of self-organization has been advanced by Haken in his work on synergetics [18]. Lately the same idea has also been revitalized in the turbulence community, primarily by a proof of existence of inertial manifolds in a class of hyperbolic dynamical systems [11]. Positive invariant manifolds are also used

in model reduction schemes in chemical kinetics [17]. In general it is hard to strictly prove the existence of global inertial manifolds even though it is often suspected that they exist and lead to spontaneous dimensional reduction in driven dissipative systems.

We now present the generic setup for inertial manifolds. Consider a dissipative dynamical system for a function $u(t,x)$ on the form

$$\dot{u} = Au + F(u)\,, \tag{9}$$

where A is a symmetric linear operator with compact resolvent, defined on a Hilbert space H and F is a nonlinear function. The operator A contains a spatial coupling through space derivatives. The operator F is nonlinear but contains no derivatives (sometimes F is also defined to include derivatives of lower order than A, but here we assume that this is not the case). For systems of the type (9) the main result is [29]:

Theorem 2 *Assume that F is Lipschitz continuous with constant c. If there is a gap in the spectrum of A such that $|\lambda_n - \lambda_{n+1}| > c$, then Eq. (9) then has an inertial manifold of dimension n.*

Consider the special case when the operator A is the Laplacian ($A = \Delta$). Common examples of this situation are reaction-diffusion equations [28]. If $A = \Delta$ in d-dimensions, the spectrum scales as $\lambda_k \sim -k^{2/d}$, and it follows that

$$\begin{aligned} \lambda_k - \lambda_{k-1} &\sim k \qquad d = 1 \\ \lambda_k - \lambda_{k-1} &\sim 1 \qquad d = 2\,. \end{aligned}$$

We conclude that the spectral gap becomes arbitrarily large for $d = 1$, which shows that inertial manifolds do exist. For $d = 2$ the spectral gap is constant so the existence of an inertial manifold is not clear from the general scaling argument. However, in $d = 2$ on finite domains there are more advanced arguments showing that under certain general conditions the spectral gap can still become large [28]. Reaction-diffusion equations in two dimensions do possess inertial manifolds.

Now let us assume the existence of an inertial manifold of dimension n. We proceed by defining a spectral projection operator $\mathcal{P}$ onto the first n eigenfunctions of A

$$\mathcal{P}u = \sum_{i=1}^{n} (u, v_i) v_i\,,$$

where v_i are normalized eigenfunctions of A and $(\cdot,\cdot)$ denotes a scalar product on H. The operator $\mathcal{P}$ is idempotent, i. e. $\mathcal{P}^2 = \mathcal{P}$, and the complement $\mathcal{Q} = \mathcal{I} - \mathcal{P}$ is also a projection operator and $\mathcal{PQ} = 0$. Furthermore, $\mathcal{P}$ commutes with A, $\mathcal{P}A = A\mathcal{P}$. We split the Eq. (9) according to

$$\frac{\partial \mathcal{P}u}{\partial t} = \mathcal{P}A\mathcal{P}u + \mathcal{P}F(\mathcal{P}u + \mathcal{Q}u) \tag{10}$$

$$\frac{\partial \mathcal{Q}u}{\partial t} = \mathcal{Q}A\mathcal{Q}u + \mathcal{Q}F(\mathcal{P}u + \mathcal{Q}u)\,. \tag{11}$$

The inertial manifold can be expressed implicitly as a graph $\Phi\colon \mathcal{P}H \to \mathcal{Q}H$. Inserting $\mathcal{Q}u = \Phi(\mathcal{P}u)$ into Eq. (10) gives

$$\frac{\partial \mathcal{P}u}{\partial t} = \mathcal{P}A\mathcal{P}u + \mathcal{P}F(\mathcal{P}u + \Phi(\mathcal{P}u))\,,$$

which is a closed evolution equation for the slow dynamics $\mathcal{P}u$. The problem is to calculate Φ. The crudest approximation is to set $\Phi = 0$. This is referred to as the linear Galerkin method. In contrast, there are many numerical schemes for nonlinear Galerkin methods that provides more nontrivial approximations of Φ, see e. g. [10,39]. In general one assumes that the fast dynamics relaxes on a time scale τ and use some method for solving Eq. (11) under this condition. Implicit Euler gives $\Phi(\mathcal{P}u)$ as a solution to the fixed point map

$$\mathcal{Q}u \to -\tau(I + \tau A\mathcal{Q})^{-1}\mathcal{Q}F(\mathcal{P}u + \mathcal{Q}u)\,. \tag{12}$$

This map is a contraction for small enough τ. Due to the time scale separation we can chose $\tau \sim \lambda_{n+1}^{-1}$. In [10] the fixed point is approximated by one application of (12) on the initial linear Galerkin guess $\mathcal{Q}u = 0$.

There exist many methods for calculating invariant manifolds in dissipative systems on more general forms than Eq. (9). Chemical kinetics and transport theory have been especially active areas [17].

Example 1 As a simple explicit example of slaving we look at the system (from [14])

$$\begin{aligned} \dot{x}_1 &= -x_2 - x_3 \\ \dot{x}_2 &= x_1 + x_2/5 \\ \dot{x}_3 &= 1/5 - 5x_3 + y \\ \dot{y} &= \epsilon^{-1}(x_1 x_3 - y)\,, \end{aligned}$$

where $\epsilon \ll 1$ defines the spectral gap ($|\lambda_x - \lambda_y| \sim \epsilon^{-1}$). We assume that the y variable has time to relax to its adiabatic fixed point $\dot{y} = \mathcal{O}(\epsilon)$. The graph that defines the inertial manifold is in this case given by $\Phi(y) = x_1 x_3 + \mathcal{O}(\epsilon)$. The resulting reduced equations read

$$\begin{aligned} \dot{X}_1 &= -X_2 - X_3 \\ \dot{X}_2 &= X_1 + X_2/5 \\ \dot{X}_3 &= 1/5 - 5X_3 + X_1 X_3\,, \end{aligned}$$

which is recognized as the Rössler system.

Averaging

Averaging is a phenomenon whose effect is reminiscent of the slaving described in the last section. However, in an averaging system the fast degrees of freedom typically do not evolve to an adiabatic fixed point. They stay active and continue to evolve on the fast time scales. Their effect on the slow degrees of freedom can be described by their average influence. As a result it is possible to derive closed equations for the slow degrees of freedom, while only taking the average affect of the fast dynamics into account. As an example, consider a system on the form

$$\begin{aligned} \dot{x} &= f(x, y) \\ \dot{y} &= \epsilon^{-1} g(x, y) \,, \end{aligned} \tag{13}$$

where $\epsilon \ll 1$ and f and g are assumed to be of order unity. If, for a fixed x the y-dynamics is ergodic, Anasov's theorem states that the slow dynamics will on some finite time interval converge uniformly to an average equation ([44]):

$$\dot{X} = F(X) \,,$$

where

$$\begin{aligned} F(\xi) &= \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathrm{d}\tau \; f(\xi, \eta[\tau, y; x]) \\ &= \int f(\xi, y) \mu_\xi(\mathrm{d}y) \,, \end{aligned}$$

where $\eta[\tau/\epsilon, y; x]$ is the solution to the differential equation (fast dynamics with x fixed and initial value y, which will not matter asymptotically due to ergodicity):

$$\frac{\partial \eta[\tau, y; x]}{\partial \tau} = \epsilon^{-1} g(x, \eta[\tau, y; x]) \quad \eta[\tau = 0, y; x] = y \,, \tag{14}$$

and μ_ξ is the invariant ergodic measure of y for a fixed $x = \xi$. The main result is that $|x(t) - X(t)|$ is of order $\mathcal{O}(\epsilon)$ for $\epsilon \ll 1$. In practice, the time average is taken over an interval proportional to the separation in time scale, i. e. (using a slight abuse of notation) $\epsilon \ll T \ll 1$. There are also many results on averaging in situations when the ergodicity assumption is not valid, see [14] and references therein. For extensive expositions on averaging and systematic perturbation methods see [25,44]. At the end of the next section we return to the issue of averaging as a limiting case of a noise approximation of the fast degrees of freedom.

White Noise Reduction: Mori–Zwanzig Projections

In non-equilibrium statistical physics dimensional reduction often means going from a deterministic high dimensional model to a reduced Langevin-type model that includes noise. The randomness stems from fast motion that, on the time scale of the relevant (slow) degrees of freedom, can be approximated as white noise. The result is a Markovian dynamics for the slow degrees of freedom. This idea was first formalized by Zwanzig [54] and has later matured into the general framework described e. g. in [13,41]. The classic example of bulk degrees of freedom behaving like noise is a heath bath in contact with a heavy particle, see e. g. [55] for details. More modern studies are often focused on low dimensional chaotic fast subsystems as noise generators [3,8,43]. In the current presentation we review some of the later findings. Many of the results in this section are based on a series of papers by Just et al., e. g. [21,22,40].

Assume we have a dynamical system whose corresponding Liouville equation reads

$$\frac{\partial \rho_t}{\partial t} = -\mathcal{L}\rho_t \,. \tag{15}$$

We like to project away some part of the system (the fast degrees of freedom). To this end we define a projection operator $\mathcal{P}$, which splits the phase space density according to $\rho_t = \mathcal{P}\rho_t + \mathcal{Q}\rho_t$, where $\mathcal{Q} = 1 - \mathcal{P}$ and $\mathcal{P}\rho$ is the subsystem we are interested in (the slow dynamics). The projection operator is idempotent, i. e. $\mathcal{P}^2 = \mathcal{P}$, which also implies that $\mathcal{P}\mathcal{Q} = 0$. Equation (15) splits into

$$\frac{\partial \mathcal{P}\rho_t}{\partial t} = -\mathcal{P}\mathcal{L}(\mathcal{P}\rho_t + \mathcal{Q}\rho_t) \tag{16}$$

$$\frac{\partial \mathcal{Q}\rho_t}{\partial t} = -\mathcal{Q}\mathcal{L}(\mathcal{P}\rho_t + \mathcal{Q}\rho_t) \,. \tag{17}$$

(Compare these equations to Eqs. (10)–(11).) We want to find a closed equation for the time evolution of $\mathcal{P}\rho_t$. In general, if $\mathcal{P}\mathcal{L}\mathcal{Q}\rho_t = 0$, then Eq. (16) is closed. This is fulfilled, for example, if the projection operator commutes with the Liouville operator, $\mathcal{P}\mathcal{L} = \mathcal{L}\mathcal{P}$, then the closure follows immediately from $\mathcal{P}\mathcal{Q} = 0$. This is an interesting special case, connected to symmetries of the dynamical system, that will be discussed extensively in Sect. "Structural Hierarchies: Foliations". To achieve this one may naively consider projections on the form [26] (based on Cauchy's formula):

$$\mathcal{P} = \frac{1}{2\pi i} \int_\Gamma \mathrm{d}\zeta (1 - \zeta\mathcal{L})^{-1} \,, \tag{18}$$

where Γ is a closed curve in the complex plane. This is a spectral projection, i. e. a generalization of the projections used in Eq. (10). The operator (18) projects onto the space spanned by the eigenfunctions corresponding to the

eigenvalues contained within the closed curve Γ. Projections defined as in (18) commutes with $\mathcal{L}$ by definition. However, projections on the density space do not in general correspond to dimensional reduction on the phase space. Therefore we cannot simply follow the same line of manipulations as in Eq. (10). We will discuss explicit forms of projections that do correspond to dimensional reductions shortly, but for now it is enough to accept that we often want to consider projections where $\mathcal{PL} \neq \mathcal{LP}$. In this case we proceed be formally solving Eq. (17) for $\mathcal{Q}\rho_t$ as

$$\mathcal{Q}\rho_t = \mathcal{Q}\rho_0 - \int_0^t \mathrm{d}\tau \mathrm{e}^{\mathcal{QL}(\tau-t)}\mathcal{QLP}\rho_\tau \,.$$

(Note that this approach does not work for Eq. (10) since F is a nonlinear function.) We are interested in the asymptotic solutions, and we can therefore ignore the term that comes from the initial distribution. Inserting the solution for $\mathcal{Q}\rho_t$ into Eq. (16) gives a closed equation for $\mathcal{P}\rho_t$:

$$\frac{\partial \mathcal{P}\rho_t}{\partial t} = -\mathcal{PLP}\rho_t + \mathcal{PL}\int_0^t \mathrm{d}\tau \mathrm{e}^{-\tau\mathcal{QL}}\mathcal{QLP}\rho_{t-\tau} \,. \quad (19)$$

The structure of Eq. (19) is interesting. As discussed in the Introduction, the dimensional reduction comes at a cost: the system has (infinite) memory, i. e. it is not Markovian. It should be noted that Eq. (19) is not simpler than Eq. (15). So far we have achieved nothing. The central idea is that, if there is a clear separation of time scales between $\mathcal{P}\rho_t$ and $\mathcal{Q}\rho_t$, then Eq. (19) can be simplified by using a Markovian approximation. To proceed we use a dynamical system where the fast and the slow degrees of freedom have been separated explicitly. Let x and y denote the slow and the fast degrees of freedom respectively. The dynamical system is given by

$$\begin{aligned} \dot{x} &= f(x,y) \\ \dot{y} &= \epsilon^{-1} g(x,y) \,, \end{aligned} \quad (20)$$

with $0 < \epsilon \ll 1$ measuring the scale separation, f, and g are on the order of unity. There is a hidden subtlety in Eq. (20). As it stands, it is identical to the situation that was analyzed in connection to averaging, Eq. (13). In fact, in the $\epsilon \to 0^+$ limit, the y-dependent part of the back-coupling in $f(x,y)$ must scale as $\mathcal{O}(\epsilon^{-1/2})$. The intuitive rationale for this scaling is that the y-term in Eq. (20) should behave as noise.

The corresponding Liouville operator splits into a fast and a slow part:

$$-\mathcal{L} = \underbrace{\sum_i \frac{\partial}{\partial x_i} f_i(x,y)}_{-\mathcal{L}_s} + \frac{1}{\epsilon}\underbrace{\sum_i \frac{\partial}{\partial y_i} g_i(x,y)}_{-\mathcal{L}_f} \,. \quad (21)$$

The idea is now to find a closed evolution for the slow degrees of freedom x. We proceed by assuming that the fast variables have time to relax to a stationary state before$x$ changes. We define the projection operator as

$$\mathcal{P}\rho_t(x,y) = \rho_{\mathrm{ad}}(y|x)\int \mathrm{d}y' \rho_t(x,y') \,, \quad (22)$$

where $\rho_{\mathrm{ad}}(y|x)$ denotes the adiabatic equilibrium of the fast degrees of freedom, y, for a given value of x, i. e.

$$\rho_{\mathrm{ad}}(y|x) = \lim_{\tau\to\infty} \mathrm{e}^{\tau\mathcal{L}_f}\rho_t(x,y) \quad (23)$$

$$\mathcal{L}_f \rho_{\mathrm{ad}}(y|x) = 0 \,. \quad (24)$$

Equation (24) means that $\rho_{\mathrm{ad}}(y|x)$ is a zero eigenfunction to $\mathcal{L}_f$ and Eq. (23) is based on the assumption that the zero eigenfunction is unique for all fixed x, which is often true for a mixing system.

At this point it is interesting to make a connection to a result called Trotter's theorem. Assume that we have an operator that can be decomposed as $\mathcal{L} = \mathcal{L}_s + \epsilon^{-1}\mathcal{L}_f$. Then the following limit is well defined and converges [46]:

$$\mathrm{e}^{t(\mathcal{L}_s+\epsilon^{-1}\mathcal{L}_f)} = \lim_{N\to\infty}\left(\mathrm{e}^{t\mathcal{L}_f/2\epsilon N}\mathrm{e}^{t\mathcal{L}_s/N}\mathrm{e}^{t\mathcal{L}_f/2\epsilon N}\right)^N \,, \quad (25)$$

where $\mathcal{L}_s$ and $\mathcal{L}_f$ do not commute in general, $\mathcal{L}_f\mathcal{L}_s \neq \mathcal{L}_s\mathcal{L}_f$ (if the operators do commute, the relation is trivial). The system can be accurately integrated from $\tau = 0$ to $\tau = t$ using N steps, where $1 \ll N/t \ll 1/\epsilon$. The intuitive rational for the step size, and behind defining the projection as in Eq. (22), is the assumption that the fast degrees of freedom has time to effectively equilibrate before the slow degrees of freedom change significantly. In Eq. (25) the scheme is explicit: $\mathrm{e}^{t\mathcal{L}_f/\epsilon N}$ relaxes the fast dynamics, whereas $\mathrm{e}^{t\mathcal{L}_s/N}$ evolves the slow dynamics. Technically this means that the limit in Eq. (23) actually converge as $\tau \sim \mathcal{O}(\epsilon) \ll 1$. The (unique) adiabatic equilibrium distribution of the fast variables is used to generate noise, which drives the slow degrees of freedom.

To proceed with this scheme we need to re-write Eq. (19) as an expansion in ϵ. From (22), (23) and (24) follows the relation:

$$\mathcal{PL}_f = \mathcal{L}_f\mathcal{P} = 0 \,. \quad (26)$$

Using this fundamental relation we can formally close Eq. (16). Under the assumption that the fast variables relaxes to a stationary state, we set

$$\frac{\partial \mathcal{Q}\rho_t}{\partial t} = 0 \,,$$

and use this in Eq. (17) to solve for $\mathcal{L}\mathcal{Q}\rho_t$. Then we use this result to close Eq. (16) (formally identifying the pseudo-inverse relation $\mathcal{Q}^{-1}\mathcal{Q} = \mathcal{Q}$, which follows from $\mathcal{Q}\mathcal{Q} = \mathcal{Q}$):

$$\frac{\partial \mathcal{P}\rho_t}{\partial t} = -\mathcal{P}\mathcal{L}_s\mathcal{P}\rho_t + \mathcal{P}\mathcal{L}_s\left(\mathcal{L}_s + \frac{1}{\epsilon}\mathcal{L}_f\right)^{-1}\mathcal{Q}\mathcal{L}_s\mathcal{P}\rho_t\,,$$

where we have used (26) to set $\mathcal{P}\mathcal{L} = \mathcal{P}\mathcal{L}_s$ and $\mathcal{L}\mathcal{P} = \mathcal{L}_s\mathcal{P}$. Expanding to first order in ϵ gives (under the assumption that $\alpha(\epsilon) \ll 1/\epsilon$):

$$\frac{\partial \mathcal{P}\rho_t}{\partial t} = -\mathcal{P}\mathcal{L}_s\mathcal{P}\rho_t + \epsilon\mathcal{P}\mathcal{L}_s\mathcal{L}_f^{-1}\mathcal{Q}\mathcal{L}_s\mathcal{P}\rho_t\,, \tag{27}$$

where $\mathcal{L}_f^{-1}$ is well defined since its null-space is projected out by $\mathcal{Q}$. Note that for the second term in Eq. (27) to be non-zero in the $\epsilon \to 0$ limit, $\mathcal{L}_s$ must be of order $\mathcal{O}(1/\sqrt{\epsilon})$. As discussed earlier, this scaling must come from the y-dependent part of the coupling $f(x, y)$. We will see this effect explicitly in an example below. To use Eq. (27) in practice the eigenfunctions of $\mathcal{L}_f$ must be found. Solutions of (27) can then be expressed in terms of series expansions in these eigenfunction basis. In most examples where this approach is successful the fast variables have a stochastic dynamics, i. e. $\mathcal{L}_f$ is a Fokker–Planck operator, see [13,41] for derivation of e. g. the Smoluchowski equation using a Hermitian polynomial basis. However, Eq. (27) is not the most convenient form for deriving a closed dynamics for the slow variables if the eigenfunctions of $\mathcal{L}_f$ are not easy to find (as is usually the case when the fast variables have deterministic dynamics). One can often introduce a trick by adding a small amount of noise to the fast degrees of freedom to make the stationary distribution smooth and well defined, i. e. to "fatten" the usually fractal attractor. Here we discuss an alternative, more direct, approach. We start by using Dyson's operator identity,

$$\begin{aligned} \mathrm{e}^{-\tau\mathcal{Q}\mathcal{L}} &= \mathrm{e}^{-\tau\mathcal{Q}\mathcal{L}_f} + \int_0^\tau \mathrm{d}\sigma \frac{\partial}{\partial\sigma}\left(\mathrm{e}^{-\sigma\mathcal{Q}\mathcal{L}}\mathrm{e}^{-(\tau-\sigma)\mathcal{Q}\mathcal{L}_f}\right) \\ &= \mathrm{e}^{-\tau\mathcal{Q}\mathcal{L}_f} - \int_0^\tau \mathrm{d}\sigma \mathrm{e}^{-\sigma\mathcal{Q}\mathcal{L}}\mathcal{Q}\mathcal{L}_s\mathrm{e}^{(\sigma-\tau)\mathcal{Q}\mathcal{L}_f}\,, \end{aligned} \tag{28}$$

to write a perturbation expansion of Eq. (19) in terms of ϵ (using Eq. (26)):

$$\begin{aligned} \frac{\partial \mathcal{P}\rho_t}{\partial t} = &-\mathcal{P}\mathcal{L}_s\mathcal{P}\rho_t + \mathcal{P}\mathcal{L}_s\int_0^t \mathrm{d}\tau\, \mathrm{e}^{\mathcal{Q}\mathcal{L}_f(\tau-t)/\epsilon}\mathcal{Q}\mathcal{L}_s\mathcal{P}\rho_\tau \\ &-\epsilon^2\mathcal{P}\mathcal{L}_s\int_0^{t/\epsilon}\mathrm{d}\tau\int_0^{\tau/\epsilon}\mathrm{d}\sigma\, \mathrm{e}^{\mathcal{Q}\mathcal{L}_f(\tau-\sigma)}\mathcal{Q}\mathcal{L}_s \\ &\cdot \mathrm{e}^{-\mathcal{Q}(\mathcal{L}_f+\epsilon\mathcal{L}_s)\tau}\mathcal{Q}\mathcal{L}_s\mathcal{P}\rho_{t-\epsilon\sigma}\,. \end{aligned}$$

The last term is of order ϵ^2 will from now on be dropped. We now define adiabatic averages as (the last equality assumes ergodicity in the fast degrees of freedom):

$$\begin{aligned} \langle h\rangle_{\mathrm{ad}}(x) &\doteq \int \mathrm{d}y\; h(x, y)\rho_{\mathrm{ad}}(y|x) \\ &= \lim_{T\to\infty}\frac{1}{T}\int_0^T \mathrm{d}\tau\; h(x, \eta[\tau/\epsilon, y; x])\,, \end{aligned} \tag{29}$$

where $\eta[\tau/\epsilon, y; x]$ is defined as in Eq. (14). Using this formalism we can write Eq. (28) as:

$$\frac{\partial\bar{\rho}_t}{\partial t} = -\langle\mathcal{L}_s\rangle_{\mathrm{ad}}\,\bar{\rho}_t + \int_0^t \mathrm{d}\tau\,\left\langle\mathcal{L}_s\mathrm{e}^{-\tau\mathcal{L}_f/\epsilon}\mathcal{Q}\mathcal{L}_s\right\rangle_{\mathrm{ad}}\,\bar{\rho}_{t-\tau}\,, \tag{30}$$

where $\bar{\rho}_t(x) = \int \mathrm{d}y\rho_t(x, y)$ is the density for the slow degrees of freedom only. It remains to analyze the kernel in the last term. For the dynamics to become approximately Markovian we need to assume that the fast degrees of freedom are exponentially mixing. Under this assumption, the correlation functions captured by the kernel decay rapidly on the time scale ϵ, and we write:

$$\begin{aligned} &\int_0^t \mathrm{d}\tau\,\left\langle\mathcal{L}_s\mathrm{e}^{-\tau\mathcal{L}_f/\epsilon}\mathcal{Q}\mathcal{L}_s\right\rangle_{\mathrm{ad}}\,\bar{\rho}_{t-\tau} \\ &= \int_0^\infty \mathrm{d}\tau\,\left\langle\mathcal{L}_s\mathrm{e}^{-\tau\mathcal{L}_f/\epsilon}\mathcal{Q}\mathcal{L}_s\right\rangle_{\mathrm{ad}}\,\bar{\rho}_t\,. \end{aligned}$$

After some further algebraic manipulations we arrive at a Fokker–Planck equation for the slow degrees of freedom:

$$\begin{aligned} \frac{\partial\bar{\rho}_t}{\partial t} = &-\sum_i \frac{\partial}{\partial x_i}D_i^{(1)}(x)\bar{\rho}_t(x) \\ &+\sum_{ij}\frac{\partial^2}{\partial x_i\partial x_j}D_{ij}^{(2)}(x)\bar{\rho}_t(x)\,. \end{aligned} \tag{31}$$

The drift term is defined as:

$$\begin{aligned} D_i^{(1)}(x) = \langle f_i\rangle_{\mathrm{ad}}(x) + \sum_j\int_0^\infty \mathrm{d}\tau\,\Big\langle f_j(x, y) \\ \frac{\partial}{\partial x_j}\delta f_i(x, \eta[\tau/\epsilon, y; x]\Big\rangle_{\mathrm{ad}}\,, \end{aligned} \tag{32}$$

where the we use the notation

$$\delta f_i(x, y) = f_i(x, y) - \langle f_i\rangle_{\mathrm{ad}}(x)$$

as an abbreviation for the fluctuations around the adiabatic equilibrium. The diffusion term is defined as:

$$D_{ij}^{(2)}(x) = \int_0^\infty \mathrm{d}\tau\,\left\langle\delta f_i(x, \eta[\tau/\epsilon, y; x])\,\delta f_j(x, y)\right\rangle_{\mathrm{ad}}\,. \tag{33}$$

Example 2 As a simple example of the method we consider a circle map coupled to a fast evolving Lorenz system in the chaotic regime.

$$\begin{aligned}
\dot{x}_1 &= x_2 + x_1\left(\frac{\gamma + \alpha y_1}{\sqrt{x_1^2 + x_2^2}} - 1\right) \\
\dot{x}_2 &= -x_1 + x_2\left(\frac{\gamma + \alpha y_1}{\sqrt{x_1^2 + x_2^2}} - 1\right) \\
\dot{y}_1 &= -\frac{3(y_1 - y_2)}{\epsilon} \\
\dot{y}_2 &= -\frac{y_1 y_3 + 26.5 y_1 - y_2}{\epsilon} \\
\dot{y}_3 &= \frac{y_1 y_2 - y_3}{\epsilon}\,.
\end{aligned} \tag{34}$$

Changing to cylindrical variables for the slow degrees of freedom, their dynamics simplifies as

$$\begin{aligned}
\dot{r} &= -\gamma(r-1) + \alpha y_1 \\
\dot{\theta} &= 1\,.
\end{aligned} \tag{35}$$

The derivation of the terms in (31) is now quite straightforward:

$$\begin{aligned}
\langle f_r\rangle_{\mathrm{ad}}(r,\theta) &= -\gamma(1-r) \\
D^{(1)}_{rr}(r,\theta) &= D^{(1)}_{\theta\theta}(r,\theta) = D^{(1)}_{r\theta}(r,\theta) = 0 \\
\langle f_\theta\rangle_{\mathrm{ad}}(r,\theta) &= 1 \\
D^{(2)}_{rr}(r,\theta) &= \alpha^2 \int_0^\infty \mathrm{d}\tau \langle y_1(\tau)\big|_{y_1(0)=y_1} y_1\rangle_{\mathrm{ad}} \\
\delta f_r(r,\theta,y_1) &= y_1 \\
D^{(2)}_{\theta\theta}(r,\theta) &= D^{(2)}_{r\theta}(r,\theta) = 0 \\
\delta f_\theta(r,\theta,y_1) &= 0
\end{aligned}$$

and the resulting approximative Langevin equation (4) for r is

$$\dot{r} = -\gamma(r-1) + \zeta(t)\,, \tag{36}$$

with ζ representing white noise: $\langle\zeta(t)\rangle = 0$ and $\langle\zeta(t_1)\zeta(t_2)\rangle = 2D^{(2)}_{rr}\delta(t_1 - t_2)$ (see e. g. [13,41,55] for details on how to relate a Langevin dynamics to a Fokker–Planck equation). The mixing time for y_1 is $\mathcal{O}(\epsilon)$ and this limits the effective support in the integral defining $D^{(2)}_{rr}$ so that $D^{(2)}_{rr} = \mathcal{O}(\alpha^2\epsilon)$. From this we conclude that, in the limit $\epsilon \to 0$, $D^{(2)}_{rr} \to 0$ unless $\alpha \sim 1/\sqrt{\epsilon}$. This is in agreement with earlier remarks on the size of the back-coupling from the fast degrees of freedom to the slow dynamics.

The resulting dynamics is shown in Fig. 1. A numerical simulation with parameters $\gamma = 5, \alpha = 1$, and $\epsilon = 0.005$ is shown. From the trajectory of the fast dynamics we measure $D^{(2)}_{rr} = 0.082$. The Langevin equation (36) with white noise predicts

$$\langle (r(t)-1)^2\rangle = \frac{D^{(2)}_{rr}}{\gamma}\,,$$

from the fluctuation-dissipation theorem. We measure $\langle (r(t)-1)^2\rangle = 0.076/\gamma$ from the simulation, which can be considered in good agreement with expectations with an error of order $\mathcal{O}(\epsilon)$.

Separation of Time-Scales in Discrete Markov Chains

We now look at the discrete equivalent of separation of time scales discussed in Sect. "Temporal Hierarchies: Separation of Time Scales". The model problem is a Markov process with a transition matrix that is approximately block-diagonal:

$$T = \epsilon \begin{pmatrix} \frac{1}{\epsilon^{(1)}}T^{(1)} & Q^{(12)} & \cdots & Q^{(1n)} \\ Q^{(21)} & \frac{1}{\epsilon^{(2)}}T^{(2)} & \cdots & Q^{(2n)} \\ \vdots & \vdots & \ddots & \vdots \\ Q^{(n1)} & Q^{(n2)} & \cdots & \frac{1}{\epsilon^{(n)}}T^{(n)} \end{pmatrix}, \tag{37}$$

where the elements in the Q-matrices are of order unity or smaller and the ϵs represent small numbers. Intuitively this system consists of n subsystems with internal dynamics $T^{(i)}$. On the time scale of $\mathcal{O}(1)$ the dynamics typically remains within one of these subsystems, but on a time scale of the order $\mathcal{O}(1/\epsilon)$ the systems switch from being in one subsystem to being in another. The slow dynamics therefore consists of n states whereas the fast dynamics have a varying number of states depending on the dimensionality of $T^{(i)}$. Note that the setup is analogous to the continuous case with $\mathcal{L} = \mathcal{L}_s + \epsilon^{-1}\mathcal{L}_f$, where the Liouville operator is viewed as a linear transition operator.

Normally, the transition matrix is not given on the near block diagonal form as in (37), but has the columns and rows mixed in random order just like in Sect. "Hierarchies in Markov Chains Through Aggregation of States". Finding a permutation matrix that transforms the transition matrix to the near block diagonal form is however a much simpler problem than finding a hidden tensor decomposition. Effective algorithms for finding optimal permutations are based on the observation that a block diagonal matrix has a degeneracy of the stationary distribution corresponding to the number to blocks, i. e. there is a n-fold degeneracy of the eigenvalue 1 (the Perron roots).

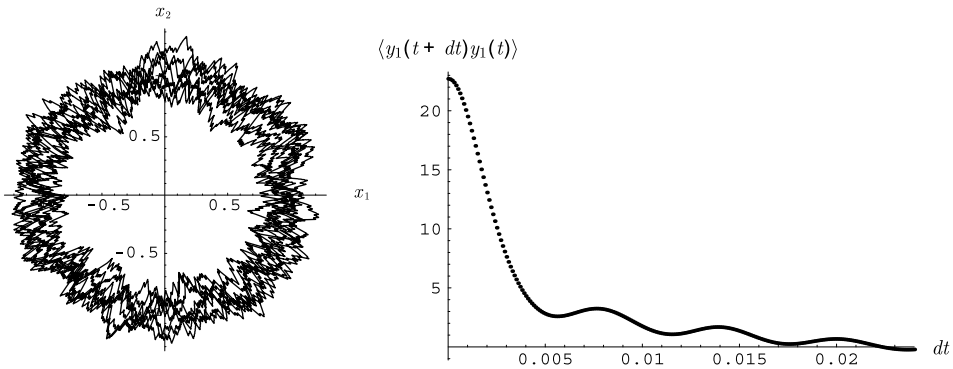

Hierarchical Dynamics, Figure 1
To the *left*: The trajectory in the $x_1 - x_2$-plane of the dynamical system defined in (34). The parameters used: $\alpha = 1, \epsilon = 0.005$ and $\gamma = 5$. To the *right*: the auto-correlation function $\langle y_1(t + dt)y_1(t)\rangle$ used to calculate $D_{rr}^{(2)}$

This also reflects the broken ergodicity for a Markov process with a block diagonal transition matrix. For a system where the transition matrix can be transformed into near block diagonal form, there is a set of eigenvalues close to 1, separated from the rest of the eigenvalues by a spectral gap. The aggregates of states building up the slow dynamics are identified by the approximately identical sign structure in the corresponding (right) eigenvectors. See [7] for details.

It is worth mentioning that Markov processes of the type given in (37) are often used as approximations of stochastic differential equations of Langevin type (4):

$$\ddot{x} = -\nabla U(x) - \gamma \dot{x} + \zeta(t) .$$

If the free energy potential U has multiple minima, then the jumping between the basins around these minima can sometimes be described as a Markov process of the type in (37). The spectral gap $\sim 1/\epsilon$ is then depending on the height and width of the potential barriers between the local minima, as well as the temperature. In the crudest approximation one only considers the second derivatives at the minima to estimate the transition times. For more details see e. g. [5,13,19].

Computational Mechanics

We end the presentation with a pointer to a framework that is useful for complementing the hierarchical decomposition of discrete systems; computational mechanics. Computational mechanics [6,47,48] is a technique for deriving optimal predictors for stochastic processes. The predictors, called ϵ-machines, are automata whose nodes are equivalence classes, causal states, of observed histories of states. All the states in a causal state must have the same probability distribution of future observed states. An ϵ-machine is the minimal and maximally efficient model of the observed process [48]. In practice an ϵ-machine can be acquired approximately from generated time series or other statistics [50]. In the context of hierarchical dynamics, the ϵ-machines can be used to find an optimal Markovian dynamics for a discrete system (with finite memory), on which we can apply the methods presented below to infer the hierarchical structure.

Invariants of the Motion

Invariants of the motion play a central role in the analysis of mechanical systems. The oldest roots in this tradition can be traced to the systematic study of continuous symmetries, advanced by Lagrange, Poisson, Jacobi, Lie, and Noether. The resulting reduction schemes eliminate inactive, i. e. constant, degrees of freedom. The most elegant product of this line of thought is Noethers' theorem, which directly connects continuous symmetries of the Hamiltonian to conserved quantities such as momentum, energy, angular momentum (spin) etc. Reduction of mechanical systems with symmetries is still a very active field, see for example the recent article [15] on reduction by stages in the Kepler problem. The literature in this field is voluminous. For a brief introduction, look in Arnold [1] (and also in Smale's article on topology in classical mechanics [52]). For moderns reviews see Marsden et al. [31,32] and Marmo et al. [30]. In the current context it is interesting to note that invariants of the motion can be classified as both temporal and structural: a constant has an infinitely slow dynamics, and at the same time the dynamics is trivially self-contained (see Sect. "Structural Hierarchies: Foliations" for definition of structural hierarchies). Alternatively, one may argue that projections that eliminates, or projects onto, invariants of the motion do not qualify to define hierarchies since they merely reflect that the origi-

nal description of the system is "over-determined". Which way we choose to interpret inactive degrees of freedom is not critical for the definition of hierarchies, so we leave it for the reader to decide for her or himself.

Structural Hierarchies: Foliations

Hierarchies in dynamical systems is a more general concept than dimensional reduction. To define a hierarchy in a more general setting, we focus on projections of the dynamics that constitutes a new "self-contained" (Markovian) dynamical system. The levels in the hierarchy do not necessarily evolve on different time scales. The general idea of autonomy naturally leads to the concept of preserved fibrations and foliations, as we define shortly.

Let a dynamical system be defined on a manifold M through a flow ϕ_t generated by a vector field $\mathbf{v}$. We define a new hierarchical level through a map π from M onto a lower dimensional manifold N. For N to define a new level of description we require π to induce a well defined flow on N, i. e., the differential $\pi_*(\mathbf{v})$ should be a well defined vector field. The hierarchical organization of dynamical systems is most naturally expressed in the language of fiber bundles. Roughly, a fiber bundle consist of a map $\pi : M \to N$ of a total space M onto a base space N such that all pre-image spaces $\pi^{-1}(x)$, where $x \in N$, are considered equivalent. The pre-image spaces $\pi^{-1}(x)$ are called fibers over a base point x. We say that a fibration π is preserved by a flow ϕ_t if fibers are carried over to fibers. In other words, if x and y belongs to the same fiber, i. e. $\pi(x) = \pi(y)$, then $\phi_t(x)$ and $\phi_t(y)$ also belong to the same fiber: $\pi(\phi_t(x)) = \pi(\phi_t(y))$. It should be clear that a preserved fibration also defines a new level in the dynamical hierarchy as defined in Fig. 2.

Foliation are geometric constructs, closely related to fibration. A foliation of an open subset of $\mathbb{R}^n$ is a union of disjoint subsets called leaves. We say that a foliation is preserved by a flow ϕ_t if leaves are carried over to leaves, i. e., if x and y belong to the same leaf L_1 then $\phi_t(x)$ and $\phi_t(y)$ also belong to the same leaf L_2. Note that we do not require $L_1 = L_2$. If $L_1 = L_2$, then we say that the foliation is invariant under the flow. Further, if L_1 has the structure of a manifold, then it is an invariant manifold of the flow. If the set of leaves is taken as the base space then a foliation becomes a fibration, with the projection defined by collapsing all points on a leaf to the same point on N. Hierarchical dynamics can be expressed either in terms of preserved fibration or preserved foliations.

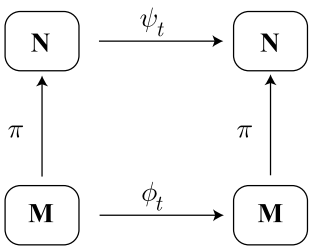

Hierarchical Dynamics, Figure 2
The manifold M is the original phase space, N is a lower dimensional phase space, and ϕ_t and ψ_t denotes flows on M respective N. The projective map π describes a new level of description if the diagram commutes

If we let $\mathbf{v}$ denote the vector field that generates the flow ϕ_t, the criterion for π to be a preserved fibration under the flow generated by the vector field $\mathbf{v}$ can be expressed as:

$$\pi(x) = \pi(y) \Rightarrow \pi_*(\mathbf{v}|_x) = \pi_*(\mathbf{v}|_y) \,, \tag{38}$$

where $\pi_* : TM \to TN$ denotes the differential of the map $\pi : M \to N$. The foliations we are considering in this paper are such that the leaves L are immersed integral submanifolds of M. It then follows from Frobenius' theorem that the vector fields spanning the tangent space of the leaves TL form an involution [4], i. e., if $\mathbf{w}_k \in TL$ then

$$[\mathbf{w}_k, \mathbf{w}_l] = \sum_m g_{kl}^m(x)\mathbf{w}_m \,, \tag{39}$$

for some smooth functions g_{kl}^m. Expressed as a fibration, π is invariant under the translations along the vector fields in TL. The invariance of the fibration π is then expressed infinitesimally as

$$\mathbf{w}_k(\pi) = 0 \,, \tag{40}$$

for all $\mathbf{w}_k \in TL$. The space of leaves, or the base space, N ($\pi : M \to N$) carries the structure of a quotient manifold, $N = M/L$. We have the following central result:

Theorem 3 *Let M be a manifold of dimension m. Consider a (singular) foliation F of M through a partition of M into connected immersed integral submanifolds (leaves). Let the tangent space of a leaf at a point x, $TL|_x$, be spanned by an involution of vector fields $\mathbf{w}_k$ at x. The foliation is preserved under the flow generated by the vector field $\mathbf{v}$ if and only if*

$$[\mathbf{v}, \mathbf{w}_k] = \sum_l f_k^l(x)\mathbf{w}_l \,. \tag{41}$$

Proof The proof of this theorem is given e. g. in [35], see also [33,34] for a discussion on how to use this result for constructing geometric integrators. We sketch the proof idea as follows. For π to be a preserved fibration, we require

$$\pi(x) = \pi(y) \Rightarrow \pi(\exp(t\mathbf{v})x) = \pi(\exp(t\mathbf{v})y) \,.$$

Now, if $\pi(x) = \pi(y)$ then $x = \exp(\mathbf{w})y$ for some $\mathbf{w} \in TL$. So

$$\begin{aligned}\pi(\exp(t\mathbf{v})x) &= \pi(\exp(t\mathbf{v})\exp(\mathbf{w})y)\\ &= \pi(\exp(\widetilde{\mathbf{w}})\exp(t\mathbf{v})y)\\ &= \pi(\exp(t\mathbf{v})y)\,,\end{aligned}$$

for some vector field $\widetilde{\mathbf{w}} \in TL$, given explicitly by the Baker–Campbell–Hausdorff formula. □

An important special case of Eq. (41) is $f_k^l(x) = 0$, i. e., when

$$[\mathbf{v}, \mathbf{w}_k] = 0 \quad \forall k\,. \tag{42}$$

The vector field $\mathbf{w}_k$ is then a symmetry of the dynamics generated by $\mathbf{v}$ (and vice versa), see [37] for details. In practice, it is much easier to search for solutions to Eq. (42) than Eq. (41). Equation (42) is a closed partial differential equation whereas Eq. (41) contains the unknown functions $f_k^l(x)$ which can be difficult to handle.

Quotient Manifold Projection

Given a set of vector fields that generates a preserved foliation, we want to construct the corresponding reduced dynamical system. The most straightforward approach is to find the projection π by solving Eq. (40). In local coordinates, if $\mathbf{w}_k = \sum_i \eta_k^i(x)\frac{\partial}{\partial x^i}$, Eq. (40) can be written as a set of quasi-linear first order partial differential equations:

$$\mathbf{w}_k(\pi) = 0\,, \tag{43}$$

or in local coordinates

$$\sum_j \eta_k^j(x)\frac{\partial \pi^i}{\partial x^j} = 0\,,$$

for all k and j. To find an explicit expression for π we need to recursively solve this system, using e. g. the method of characteristics.

There are technical complications with the quotient manifold construction presented above. Regularity is not guaranteed. The resulting quotient manifold may not even be Hausdorff. However, if we assume that the involution $\mathbf{w}_i$ form regular submanifolds, it follows that the vector fields can be defined so that the functions $f_k^l(x)$ in (41) and g_{kl}^m in (39) are independent of x and $\mathbf{w}_k$ form a Lie algebra [35]. The corresponding Lie group has a regular action on M. In this case the quotient manifold M/G is smooth and well defined.

Example 3 (Linear dynamics projected onto the real projective plane.) In general, any linear system $\dot{x} = Ax$ has two trivial symmetries: $\mathbf{w}_1 = \sum_{ij} A_{ij}x_i\frac{\partial}{\partial x_j}$ and $\mathbf{w}_2 = \sum_i x_i\frac{\partial}{\partial x_i}$. The first symmetry is just the dynamics itself and the corresponding projection maps onto an invariant of the motion. The latter symmetry comes from the trivial observation that the identity matrix commutes with A, but it actually gives a non-trivial foliation. According to Eq. (43), the projection must fulfill

$$\sum_j x_j\frac{\partial \pi_i(x)}{\partial x_j} = 0\,,$$

for all components i. The general solution reads

$$\pi_i(x) = F_i\left(\frac{x_i}{x_j}\right)\,,$$

for some arbitrary coordinate x_j and general functions F_i. Note that $\pi_j(x)$ is a constant. This reflects the reduction of dimensionality by the projective map. Just as in the case of the circle there is no single projective map valid over the entire phase space. For different choices of coordinate x_β, π provides "coordinate charts" valid in regions where $x_j \neq 0$. The resulting manifold can be identified as the real projective plane, $P\mathbb{R}^{d-1}$, if the original dynamics was in $\mathbb{R}^d$.

The circle is a special case, $A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ where it is natural to choose

$$\pi(x, y) = \arctan(x/y)\,,$$

in the coordinate chart where $y \neq 0$, and

$$\pi(x, y) = \pi/2 - \arctan(y/x)\,,$$

in the coordinate chart where $x \neq 0$. The projection is onto the angular coordinate in the cylindric coordinate system. Note that $P\mathbb{R}^1 \simeq S^1$.

The other projections of a linear system follows directly from the Jordan form of A that fully resolves all invariant subspaces of the dynamics. Each invariant subspace can then be projected out to form a new level in the hierarchy. A set of non-redundant eigenvalues λ_i can be projected immediately using

$$\pi(x) = \left(\prod_i (A - \lambda_i \mathbf{I})\right)x\,, \tag{44}$$

or the more abstract formula in Eq. (18). Note that Eq. (44) is not strictly a projection since $P^2 \neq P$ while Eq. (18) is a projection. If it is important, this can be fixed by using a pseudo-inverse, see Sect. "Hierarchies in Markov Chains

Through Aggregation of States". Equation (44) is perhaps the most natural projection of a linear system. The special case in the example above is focused on projecting onto one degree of freedom corresponding to one half of a complex conjugate pair. This is the reason why the topology of the resulting reduced manifold becomes non-trivial.

Example 4 (Skew-product systems) Consider a dynamical system with the following nonlinear skew-product structure:

$$\begin{aligned} \dot{x} &= f(x) \\ \dot{y} &= g(x, y)\,, \end{aligned} \tag{45}$$

where $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^{m-n}$. The family of vector fields $\mathbf{w}_1 = \frac{\partial}{\partial y_1}, \ldots, \mathbf{w}_{m-n} = \frac{\partial}{\partial y_{m-n}}$ forms an involution. The corresponding foliation $\pi(x, y) = F(x)$ for any function F is in accordance with $\dot{x} = f(x)$ being a self-contained subsystem. Note that this result is also consistent with the discussion in Sect. "White Noise Reduction: Mori–Zwanzig Projections" since if $f(x, y)$ in Eq. (20) is independent of y it follows that $\delta f(x, y) = 0$ and therefore $D^{(1)} = D^{(2)} = 0$.

Non-autonomous Dynamical Systems

It is straight forward to generalize the foliation framework to non-autonomous dynamical systems:

$$\dot{x} = f(t, x)\,. \tag{46}$$

To analyze Eq. (46) we need to find a involution of vector fields $\mathbf{v}_k$ that can generate the dynamics, i. e.

$$\begin{aligned} f(t, x) &= \mathrm{span}\{\mathbf{v}_i\} \quad \forall t \\ [\mathbf{v}_k, \mathbf{v}_l] &= \sum_m g_{kl}^m(x)\mathbf{v}_m\,. \end{aligned}$$

Not that if f is free to take any form, then $\mathbf{v}_i$ must span the entire tangent space. The condition (41) is generalized to:

$$[\mathbf{v}_k, \mathbf{v}_l] = \sum_m f_{kl}^m(x)\mathbf{w}_m \quad \forall k\,. \tag{47}$$

Hierarchies in Discrete Dynamical System Through Normal Subgroup Extensions

In this section we present the analogy of non-autonomous foliation (Sect. "Non-autonomous Dynamical Systems") for discrete systems. Let H be a semigroup with a finite number of generators acting on a finite state space Σ. The semigroup H together with the state space Σ define a dynamical system. We define a new finite group N, also acting on Σ. Further, we define the joint group $F = \mathrm{span}(H, N)$ generated by the elements in both H and N: $f = \prod_i f_i$ where $f_i \in H \cup N$. The action ψ of F on Σ is well defined through the action of H and N individually. Two elements f_1 and f_2 are identical if and only if $\psi(f_1, \sigma) = \psi(f_2, \sigma), \forall \sigma \in \Sigma$. Note that both H and N are subgroups of F.

Assume that N has been carefully chosen so that, for each $h \in H$, $\sigma \in \Sigma$, and $n \in N$, there exist a $n' \in N$ such that

$$\psi(h * n, \sigma) = \psi(n' * h, \sigma)\,. \tag{48}$$

Note that the element n' may depend on the state σ, i. e. $\psi(h * n, \sigma_1) = \psi(n' * h, \sigma_1)$ and $\psi(h * n, \sigma_2) = \psi(n'' * h, \sigma_2)$ do not generally imply $n' = n''$. Relation (48) ensures that N is a normal subgroup of F (for each fixed $\sigma \in \Sigma$). Furthermore, the commutation relation in Eq. (48) ensures that, for each $\sigma \in \Sigma$, the elements in the product $f = \prod_i f_i$ can be re-arranged so that $f = \prod_i n_i \prod_j h_j$, i. e. for each $f \in F$ there exist $n \in N$ and $h \in H$ such that $f^{(\sigma)} = n^{(\sigma)} * h^{(\sigma)}$ (the superscript indicates that the decomposition may be different for different σ). It then follows that (see Sect. "Some Concepts from Group Theory")

$$F^{(\sigma)} \simeq N \rtimes_{\Psi^{(\sigma)}} H\,, \tag{49}$$

with the automorphism mapping $\Psi_h^{(\sigma)}(n) = n'$, where $\psi(h * n, \sigma) = \psi(n' * h, \sigma)$.

Equations (48) and (49) are the equivalent of the ideal relation for vector fields used in Eq. (41) or more generally Eq. (47). It is therefore the central relation in the reduction of discrete dynamical systems. Note that the definition of the automorphism map $\Psi^{(\sigma)}$ plays the same role as the commutator of the generating vector fields $\exp(t\mathbf{v})\exp(\mathbf{w}) = \exp(\widetilde{\mathbf{w}})\exp(t\mathbf{v})$ in the proof of reduction in continuous systems using preserved foliations. In the continuous case, the solution to the equation $\psi(h*n, \sigma) = \psi(n' * h, \sigma)$ is given explicitly by the Baker–Campbell–Hausdorff formula. In the discrete case there is no such explicit approach. However, as we shall see there is no need for deriving the explicit expression for $\Psi^{(\sigma)}$.

To form the quotient projection in the discrete case, we form equivalence classes on Σ: $\sigma_1 \sim \sigma_2$ if there exist a $n \in N$ such that $\sigma_1 = \psi(n, \sigma_2)$, i. e. the orbits of N form the equivalence classes. The equivalence classes are preserved under the action of H. To see this, let $\sigma_1 = \psi(n, \sigma_2)$. Then when H acts on σ_1 and σ_2 we have

$$\psi(h, \sigma_1) = \psi(h * n, \sigma_2) = \psi(n' * h, \sigma_2) \sim \psi(h, \sigma_2)$$

for some $n' \in N$. Thus $\psi(h, \sigma_1) \sim \psi(h, \sigma_2)$, so the equivalence classes are preserved. This implies that the action

of H is well defined on the set of equivalence classes, i. e., the quotient set Σ/N. We say that the resulting dynamical system, i. e. H acting on Σ/N, is reduced.

The algebraic structure presented in this section makes a clear connection to reduction through foliations for continuous dynamical systems. Moreover, since any automaton can be represented as a semigroup (see Sect. "Some Concepts from Group Theory"), it also shows how to make hierarchical decompositions of automata. The algebraic structure is also connected to Wreath products and Krohn–Rhodes theory for finite automata [9,24]. Furthermore, automata theory can also be applied to Markov processes. The Markov process must then be decomposed into a Bernoulli shift, combined with a set of deterministic transition matrices [27]. Intuitively this means that at each time step the system is updated using a randomly chosen deterministic transition matrix. The semigroup H is composed of all the deterministic transition matrices that generate the Markov process. However, for a Markov process the approach introduced in this section is unnecessarily complicated. In the next section we discuss a more useful technique.

Hierarchies in Markov Chains Through Aggregation of States

Consider a stochastic Markov process with transition matrix T and state space consisting of symbols in an alphabet Σ. If T can be decomposed as a tensor, or Kronecker, product

$$T = T^{(1)} \otimes T^{(2)} \otimes \cdots \otimes T^{(N)} , \tag{50}$$

then $T^{(i)}$ are transition matrices for independent processes. Hence, a projection onto a state space representing one, or many, subsystems results in a new Markov process over a state space with reduced cardinality. Let $T^{(2:N)} = T^{(2)} \otimes \cdots \otimes T^{(N)}$ and $T^{(1)}_{i,j}$ denote the elements $T^{(1)}$. From the definition of a tensor product it then follows that

$$T = \begin{pmatrix} T^{(1)}_{1,1} T^{(2:N)} & T^{(1)}_{1,2} T^{(2:N)} & \cdots & T^{(1)}_{1,K} T^{(2:N)} \\ T^{(1)}_{2,1} T^{(2:N)} & T^{(1)}_{2,2} T^{(2:N)} & \cdots & T^{(1)}_{2,K} T^{(2:N)} \\ \vdots & \vdots & \ddots & \vdots \\ T^{(1)}_{K,1} T^{(2:N)} & T^{(1)}_{K,2} T^{(2:N)} & \cdots & T^{(1)}_{K,K} T^{(2:N)} \end{pmatrix} ,$$

where $T^{(1)}_{ij} T^{(2:N)}$ represents sub-matrices in T. Since $T^{(1)}$ is a transition matrix, its column sum is normalized. It follows that

$$T^{(2:N)} = \sum_{j=1}^{K} T^{(1)}_{i,j} T^{(2:N)} \quad i = 1, \ldots, K ,$$

or explicitly

$$T^{(2:N)}_{k,l} = \sum_{j=0}^{K-1} T_{k+iK,l+jK} \quad i = 1, \ldots, K . \tag{51}$$

On the one hand, under the assumption that T can be decomposed as in the ansatz (50), then Eq. (51) gives a Markov process on the reduced state space. On the other hand, Eq. (51) also tests the ansatz since, for each element $T^{(2:N)}_{k,l}$ the right hand side should evaluate identically for all choices of the index i. The last observation can be used to identify hierarchies in Markov processes.

When a subsystem $T^{(1)}$ is projected out according to Eq. (51), the corresponding reduction on the state space can, as usual, be defined in terms of equivalence classes. Assume that the state vector is ordered in the default manner, so that the element in position i represents the probability that the system is in a state denoted by $\sigma_i \in \Sigma$. The reduction described by (51) is then equivalent to forming equivalence classes according to $\sigma_i \sim \sigma_j$ iff $(i - j) \mod K = 0$. It is clear that the composition of the equivalence classes depends on the ordering in the state vector. As a result, to find a tensor decomposition of the transition matrix presented with the rows and columns in random ordering, the validity of Eq. (51) should be tested for all permutations (excluding those not affecting the equivalence classes). Formally we express this as:

Theorem 4 *An $N \times N$ transition matrix T can be decomposed as a tensor product of an $K \times K$ matrix and an $N/K \times N/K$ matrix if and only if there exists a $P \in \mathrm{Sym}(\Sigma)$ such that*

$$\sum_{j=0}^{K-1} (P^T T P)_{k+iK,l+jK} \tag{52}$$

gives the same result independent of the index i.

A major problem with this result is that the cardinality of the symmetric group increase extremely fast with the number of states in the state space (actually $|\mathrm{Sym}(\Sigma)| = |\Sigma|!$). The practical usefulness of the summation condition is of limited. Below we present a criterion that is more useful for identify projections onto a Markov process with reduced state space. Before this however, we generalize from direct (Kronecker) product to semi-direct products.

The decomposition in (51) is not the most general form permitting projections onto well-defined Markov subsystems. In (51) the subsystems are completely decoupled. In a more general case it may happen that a subsystem can influence the dynamics of another but not vice

versa. In this situation it is still possible to project onto the fist subsystem, but not onto the second. This situation is equivalent with the one presented in Eq. (45) for the continuous case. In the discrete setting a decomposition of this type uses a semi-direct product of submatrices. It is straightforward to define consistency equations on the form in Eq. (51) in this more general case. The size of the different equivalence classes in the partitioning of the state space may vary, and this needs to be taken into account in the summation. We start by introducing some notation.

We consider a Markov chain X_t, $t = 0, 1, \ldots$, with a finite state space $\Sigma = \{1, \ldots, N\}$ and transition probability matrix $P = [p_{ij}]$. The transition matrix operates from the right $x_{t+1} = x_t P$. A reduction of the state space is a lumping, or state aggregation, by a partition of the state space $\widetilde{\Sigma} = \{L_1, \ldots, L_M\}$ where L_k is a nonempty subset of Σ, $L_k \cap L_l = \emptyset$ if $k \neq l$ and $\bigcup_k L_k = \Sigma$. Clearly $M \leq \mathrm{N}$. The reduction can be defined by an $N \times M$ matrix $\Pi = [\pi_{ik}]$ where $\pi_{ik} = 1$ if $i \in L_k$ and $\pi_{ik} = 0$ if $i \notin L_k$. We call the reduction, or the aggregation, a Lumping $\widetilde{\Sigma}$. A lumping induces a quotient process $\tilde{x}_t = x_t \Pi$. If the process $\tilde{x}_t$ is a Markov process (which is, as always, not usually the case), then we say that the Markov chain is *strongly lumpable* with respect to Π (or $\widetilde{\Sigma}$). The following criterion is a generalization of the summation condition above and is necessary and sufficient for a Markov chain with transition matrix P to be lumpable with respect to Π or $\widetilde{\Sigma}$ [23,26]:

1. $\Pi \widetilde{P} = P \Pi$, where $\widetilde{P} = \Pi^+ P \Pi$ is the transition matrix on the reduced phase space ($\Pi^+ \doteq (\Pi^T \Pi)^{-1} \Pi^T$ is the left psudo-inverse of Π, $\Pi^+ \Pi = I$).
2. $\ker(\Pi)$ is P-invariant, i. e. $y\Pi = 0 \Rightarrow yP\Pi = 0$.
3. For any $L_k, L_l \in \widetilde{\Sigma}$, the total probability of going from any state $i \in L_k$ to L_l, i. e. $\sum_{j \in L_l} p_{ij}$, is independent of i.

Note that condition 2 is equivalent to Eq. (43) and condition 1 is related to Eq. (18). If the Markov chain is lumpable, then the probability distribution over the reduced state space is updated according to $\tilde{x}_{t+1} = \tilde{x}_t \widetilde{P}$, where $\widetilde{P}$ is defined in Criterion 1. Furthermore, the transition matrix for the reduced Markov chain is given by

$$\tilde{p}_{kl} = \sum_{j \in L_l} p_{ij} \quad i \in L_k \,, \tag{53}$$

where we note that $\tilde{P} = [\tilde{p}_{kl}]$ is well defined since the sum is independent of which representative $i \in L_k$ we chose according to criterion 3.

As mentioned in connection with the summation criterion in Eq. (52), criteria 1–3 not immediately useful for identifying lumpings of a Markov chain. Barr and Thomas [2] presented a necessary lumpability criterion involving the left eigenvectors of the transition matrix. It was first noted that the spectrum of $\tilde{P}$ must be a subset of the spectrum of P (this is also discussed in detail in a more general setting in [26]). It also follows that if $vP = \lambda v$ then $\tilde{v}\tilde{P} = \lambda \tilde{v}$, with $\tilde{v} = v\Pi$. It follows that if λ is an eigenvalue of both P and $\tilde{P}$, then $\tilde{v}$ is an eigenvector of $\tilde{P}$, but if λ is not an eigenvalue of $\tilde{P}$ then $\tilde{v} = v\Pi = 0$. Intuitively this observation can be understood as Π eliminating a subset of the eigenvectors of P. This is also clear from criterion 1 and 2, as well as from our previous discussion on linear systems in general. Equation (44).

Barr and Thomas' result suggests a search for lumpings defined by Π such that $v^\alpha \Pi = 0$ for some subset of the left eigenvectors of P, $\{v^\alpha\}_{\alpha \in J}$, $J \subseteq \Sigma$. Since Π should be a matrix with zeros and ones, $v^\alpha \Pi = 0$ essentially means searching for eigenvectors with subsets of elements that sums to zero. For lumpings only involving agglomeration of two states this is straightforward since the eigenvector(s) must have pairs of elements $v_i^\alpha = -v_j^\alpha$. However, agglomeration of k states means searching for partial sums evaluating to zero and involving k elements. This leads back to the combinatorial explosion of possibilities discussed before. As Barr and Thomas point out, there is no obvious algorithm to generate Π based on their result.

In a recent study a more useful method for identifying possible partitions of the state space leading to a reduced Markov process is presented [36]. The key is an observation that the dual of the probability space can be used to identify lumpings. The criterion $v^\alpha \Pi = 0$ is viewed as an orthogonality condition between the left eigenvectors $\{v^\alpha\}_{\alpha \in J}$ and the column space of Π. The orthogonal complement of a set of left eigenvectors is spanned by complementary right eigenvectors (defined naturally in the dual vector space). These complementary eigenvectors span the column space of Π. Requiring that Π consists of zeros and ones does corresponds to a criterion of repeated elements within each complementary right eigenvector. Clearly, identifying repeated elements in the right eigenvectors is algorithmically straight forward. The precise result reads as follows (proof is to be published [36]):

Theorem 5 *Assume that P is a diagonalizable transition matrix with full rank describing a Markov process $x_{t+1} = x_t P$. Consider a set of linearly independent right eigenvectors of P, $Pu^\alpha = \lambda^\alpha u^\alpha$. Let $I \subseteq \Sigma$ be the set of indices for the eigenvectors. Form state equivalence classes defined by states with identical elements in all eigenvectors u^α, i. e. $i \sim j$ iff $u_i^\alpha = u_j^\alpha \, \forall \alpha \in I$. The equivalence classes define a partitioning $\widetilde{\Sigma}$ of the state space. This partitioning*

is a lumping of the Markov chain if the number of partition elements equals the number of eigenvectors, i. e. $|\widetilde{\Sigma}| = |I|$.

Conversely, if $\widetilde{\Sigma}$ *is a lumping then there exist* $|\widetilde{\Sigma}|$ *linearly independent right eigenvectors that are invariant under permutations within the lumps.*

There are two important points to make in connection to this result. First, the result can be viewed as an extension of the result in Sect. "Separation of Time-Scales in Discrete Markov Chains". A block diagonal matrix can be constructed as a semi-direct product of the identity matrix and a set of matrices appearing in the blocks. The second point is that the criterion in general provides a relatively efficient method for generating projections of a given Markov chain. The most time consuming part of the algorithm is to diagonalize the transition matrix. This problem scales approximately cubic in the number of states, i. e. not exponential like the naive approaches. However, there are several subtleties to take into consideration. Degenerate eigenvalues can cause difficulties since the corresponding eigenvectors are in this situation no longer uniquely defined. The criterion is still valid but one must chose which linear combination of eigenvectors to use. The partitions suggested by the eigenvectors can also be nested to produce an exponential number of possible reductions. This problem is intrinsic, no algorithm can find all possible reductions in linear time since there might be an exponential number of possible projections. The identity matrix is a good example where any projection is clearly an acceptable lumping.

Example 5 (Kronecker product system) As a simple example we use the method on a process with transition matrix $T_{\mathcal{A}}$ described in Fig. 3. If we sort the state vector according to the default ordering: $\{\sigma_1, \sigma_2.\sigma_3, \sigma_4\}$, then the corresponding transition matrix reads

$$T_{\mathcal{A}} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1-p & 0 & p & 0 \\ 0 & p & 0 & 1-p \\ 0 & 0 & 1 & 0 \end{pmatrix} .$$

As it stands, $T_{\mathcal{A}}$ is not directly decomposable but if we change the ordering of the states to $\{\sigma_1, \sigma_4, \sigma_3, \sigma_2\}$, we end up with a decomposable transition matrix:

$$P^T T_{\mathcal{A}} P = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1-p & 0 & p \\ 1-p & 0 & p & 0 \end{pmatrix}$$

$$= \underbrace{\begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}}_{T_{\mathcal{C}}} \otimes \underbrace{\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}}_{T_{\mathcal{B}}} ,$$

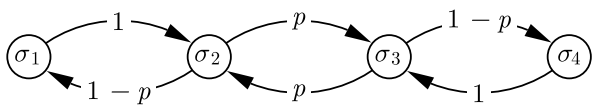

Hierarchical Dynamics, Figure 3
Example process $T_{\mathcal{A}}$ over the state space $\Sigma = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$. The edges are labeled with transition probabilities

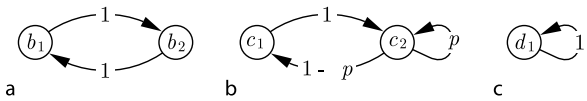

Hierarchical Dynamics, Figure 4
Dynamics (a) $T_{\mathcal{B}}$, (b) $T_{\mathcal{C}}$ and (c) $T_{\mathcal{D}}$ (a trivial process with one state) resulting from projections of the example process $T_{\mathcal{A}}$ in Fig. 3. The states are defined in terms of equivalence classes as follows: $b_1 = \{\sigma_1, \sigma_3\}$, $b_2 = \{\sigma_2, \sigma_4\}$, $c_1 = \{\sigma_1, \sigma_4\}$, $c_2 = \{\sigma_2, \sigma_3\}$, $d_1 = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$

where P is the permutation matrix providing the re-ordering: $\{\sigma_1, \sigma_2.\sigma_3, \sigma_4\} \to \{\sigma_1, \sigma_4.\sigma_3, \sigma_2\}$. From this it follows that the process $T_{\mathcal{A}}$ can be projected onto $T_{\mathcal{B}}$, $T_{\mathcal{C}}$, or the trivial process with one state $T_{\mathcal{D}}$, all described in Fig. 4.

The eigenvalues of $P_{\mathcal{A}}$ are $\lambda_1 = 1$, $\lambda_2 = -1$, $\lambda_3 = 1-p$ and $\lambda_4 = p-1$. The corresponding right eigenvectors are $u_1 = (1,1,1,1)^{\mathrm{T}}$, $u_2 = (-1,1,-1,1)^{\mathrm{T}}$, $u_3 = (-1, p-1, 1-p, 1)^{\mathrm{T}}$, and $u_4 = (1, p-1, p-1, 1)^{\mathrm{T}}$. By the condition on repeated elements, the two pairs of eigenvectors $\{u_1, u_2\}$ and $\{u_1, u_4\}$ implies the two possible lumpings found above: $b_1 = \{\sigma_1, \sigma_3\}$, $b_2 = \{\sigma_2, \sigma_4\}$, $c_1 = \{\sigma_1, \sigma_4\}$, $c_2 = \{\sigma_2, \sigma_3\}$, $d_1 = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$.

Example 6 (Semi-direct product system) Consider the transition matrix

$$P = \begin{pmatrix} a+b+(c-1)/2 & 1-a-b & (1-c)/2 \\ -a+(c+1)/2 & a & (1-c)/2 \\ 1-b-c & b & c \end{pmatrix} ,$$

with $0 \le a, b, c \le 1$. P has the eigenvalues $\lambda^1 = 1$, $\lambda^2 = 2a+b-1$ and $\lambda^3 = (3c-1)/2$, and the eigenvectors

$$u^1 = (1,1,1)^{\mathrm{T}} ,$$
$$u^2 = (1+c-2a-2b, 2(a-c), 2b+c-1)^{\mathrm{T}} \text{and}$$
$$u^3 = (-1,-1,2)^{\mathrm{T}} .$$

There are three possible lumpings of P:

$$\widetilde{\Sigma}_1 = \{\{1,2,3\}\} \qquad \text{from } \{u^1\} ,$$
$$\widetilde{\Sigma}_2 = \{\{1,2\},\{3\}\} \qquad \text{from } \{u^1, u^3\} \text{ and}$$
$$\widetilde{\Sigma}_3 = \{\{1\},\{2\},\{3\}\} \qquad \text{from } \{u^1, u^2, u^3\} ,$$

where $\widetilde{\Sigma}_2$ is valid if $2a+b-1 \neq 0$ and $3c-1 \neq 0$.

Conclusion

Analysis of hierarchical dynamics contains two separate steps. Initially the states, or degrees of freedom, that should be eliminated in each hierarchical transition need to be identified. After the identification, the reduced dynamics can be derived. The first step is by far the most computationally expensive. This is especially transparent in the analysis of decomposition of transition matrices for Markov processes (see Sect. "Hierarchies in Markov Chains Through Aggregation of States"), but the conclusion holds in general. In the discussion on temporal hierarchies (Sect. "Temporal Hierarchies: Separation of Time Scales"), the problem of separating the slow and the fast degrees of freedom was tactically assumed to be given a priori (with the exception of inertial manifolds where numerical algorithms do exist [10,17,39]), and the nearly decomposable Markov chains discussed in Sect. "Separation of Time-Scales in Discrete Markov Chains", see also [7]. In the section on structural hierarchies (Sect. "Structural Hierarchies: Foliations"), the conditions on the projections are clearly stated. However the practical applicability of the methods presented depends critically on efficient algorithms for finding the projections. Naive approaches typically lead to algorithms that are exponentially slower than solving the original system. There are in fact reasons to believe that the Markov chains are typical. As we have seen, naive approaches to finding reductions lead to exponentially slow algorithms. There are more efficient methods that work most of the time, like to eigenvector criterion presented in Sect. "Hierarchies in Markov Chains Through Aggregation of States", but in worst case scenarios these methods also become exponentially slow. In conclusion, methods for deriving appropriate projections for reduction is an area that is in need of more extensive study.

Future Directions

In the physics community, one of the main areas where hierarchical dynamics is discussed is renormalization of systems with critical behavior. The most prominent example is the Ising model in two or three dimensions [16]. The main idea in renormalization theory is that the state of the system, drawn from an ensemble of possible states with probability defined by the Hamiltonian, is structurally self-similar under the projection operator. Clearly, since the projection in general reduces the dimensionality of the phase space, self-similarity can only be well defined if the dimensionality of the system is infinite. For certain parameter values (often the temperature is the parameter), the dynamics is not only structurally self-similar but the interaction strength between the agglomerated states does not decay, i. e. the projection operator has a non-trivial fixed point on the space of Hamiltonians (or rather a parametrization of the Hamiltonians) . In this situation the system is said to be in a critical state, characterized by self-similarity and non-vanishing fluctuations on all length scales. It seems obvious that renormalization theory and hierarchical dynamics are closely connected. In fact it is relatively straight forward to formulate the renormalization group in the framework of structural hierarchies as presented in Sect. "Structural Hierarchies: Foliations". A detailed investigation of the relation remains to be done. Furthermore, the more interesting aspect of this connection would be to explore possible generalizations where the constraint of self-similarity is relaxed and we focus on non-vanishing interactions with different structure on different levels in the hierarchy. The result of such an effort should, if successful, be of central interest in complex systems, especially for studying emergence (the author would like to acknowledge private discussions with Nils Baas on this subject).

The remarkable ingenuity in the renormalization group is the step of abstraction when the projection is viewed as generator of a discrete dynamical system on the space of Hamiltonians. This leap enables us to understand the origin of universality, as the details of the physical system become unimportant for the analysis of the behavior (the fixed points) of the dynamical system generated by the projection. It is exciting to speculate about the possibility of formulating similar "meta-models" for the dynamics generated by the projections used to define general hierarchical dynamics.

Finally, one of the central problems with defining emergence in complex dynamical systems is the lack of a framework for describing creation and destruction of objects. On the basic level, the number of degrees of freedom in a dynamical system does not change during the time evolution. From this perspective one may argue that dynamical systems are not suited for describing systems consisting of entities that can be created, destroyed and change internal properties [12]. These features are central in many complex systems, especially in biological and social systems. Methods used to analyze hierarchical dynamics can possibly be used to define the emergence of objects and organizations in dynamical systems. Near decomposability into independent subsystems (as defined in Sect. "Hierarchies in Markov Chains Through Aggregation of States") can for example be used to define objects, see [20] for an alternative definition. Emergence, creation, destruction and change can be understood as different regions of the phase space permitting different

hierarchical structures. A "meta-model", as described in the previous paragraph, could possibly be used to understand emergent higher-order organization in dynamical systems.

Acknowledgment

The author would like to thank Olof Görnerup and Kolbjørn Tuntrøm for discussions and comments on the manuscript. The author would also like to acknowledge support from the EU integrated project FP6-IST-FET PACE, by EMBIO, a European Project in the EU FP6 NEST initiative, and by MORPHEX, a European Project in the EU FP6 NEST initiative.

Bibliography

Primary Literature

1. Arnold VI (1989) Mathematical Methods of Classical Mechanics, Graduate Texts in Mathematics, 2nd edn. Springer, New York
2. Barr DR, Thomas MU (1977) An eigenvector condition for Markov chain lumpability. Oper Res 25(6):1028–1031
3. Beck C, Schlögl F (1993) Thermodynamics of chaotic systems, Cambridge Nonlinear Science Series, vol 4. Cambridge University Press, Cambridge
4. Boothby WM (2002) An introduction to differentiable manifolds and Riemannian geometry, Pure and Applied Mathematics, vol 120, 2nd edn. Academic Press
5. Caroli B, Caroli C, Roulet B (1979) Diffusion in a bistable potential: a systematic wkb treatment. J Stat Phys 21:415–536
6. Crutchfield J, Young K (1989) Inferring statistical complexity. Phys Rev Lett 63:105
7. Deuflhard P, Huisinga W, Fischer A, Schütte C (2000) Identification of almost invariant aggregates in reversible nearly uncoupled markov chains. Linear Algebra Appl 315:39–59
8. Dorfman JR (1998) An Introduction to Chaos in Nonequilibrium Statistical Mechanics, Cambridge Lecture Notes in Physics, vol 14. Cambridge University Press, Cambridge
9. Egri-Nagy A (2005) Algebraic hierarchical decompositions of finite state automata a computational approach. Ph D thesis, University of Hertfordshire
10. Foias C, Jolly MS, Kevrekidis IG, Sell GR, Titi ES (1988) On the computation of inertial manifolds. Phys Lett A 131(7–8): 433–436
11. Foias C, Sell GR, Temam R (1988) Inertial manifolds for non-linear evolutionary equations. J Differ Equ 73:309–353
12. Fontana W, Buss LW (1996) The barrier of objects: From dynamical systems to bounded organizations. In: Casti J, Karlqvist A (eds) Barriers and Boundaries. Addison-Wesley, Reading, pp 56–116
13. Gardiner C (2004) Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences. Springer Series in Synergetics, vol 13, 3rd edn. Springer, Berlin
14. Givon D, Kupferman R, Stuart A (2004) Extracting macroscopic dynamics: model problems and algorithms. Nonlinearity 17:55–127
15. Godfreyt SE, Princet GE (1991) A canonical reduction of order for the kepler problem. J Phys A: Math Theor 24: 5465–5475
16. Goldenfeld N (1992) Lectures on Phase Transitions and the Renormalization Group. Perseus Books, Oxford
17. Gorban A, Karlin I (2005) Invariant Manifolds for Physical and Chemical Kinetics. Lecture Notes in Physics. Springer, Berlin
18. Haken H (1983) Synergetics, an Introduction: Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry, and Biology, 3rd edn. Springer, New York
19. Hänggi P, Talkner P, Borkovec M (1990) Reaction-rate theory: fifty years after kramers. Rev Mod Phys 62(2):251–341
20. Jost J, Bertschingera N, Olbricha E, Aya N, Frankela S (2007) An information theoretic approach to system differentiation on the basis of statistical dependencies between subsystems. Physica A 378(1):1–10
21. Just W, Kantz H, Rödenbeck C, Helm M (2001) Stochastic modelling: Replacing fast degrees of freedom by noise. J Phys A: Math Theor 34:3199–3213
22. Just W, Gelfert K, Baba N, Riegert A, Kantz H (2003) Elimination of fast chaotic degrees of freedom: On the accuracy of the born approximation. J Stat Phys 112:277–292
23. Kemeny JG, Snell JL (1976) Finite Markov Chains, 2nd edn. Springer, New York
24. Krohn K, Rhodes J (1965) Algebraic theory of machines. i. prime decomposition theorem for finite semigroups and machines. Trans Am Math Soc 116:450–464
25. Lichtenberg AJ, Lieberman MA (1983) Regular and Stochastic Motion, Applied Mathematical Sciences, vol 38. Springer, New York
26. Lorch E (1962) Spectral theory. Oxford University Press, New York
27. Maler O (1995) A decomposition theorem for probabilistic transition systems. Theor Comput Sci 145(1–2):391–396
28. Mallet-Paret J, Sell GS (1988) Inertial manifolds for reaction diffusion equations in higher space dimensions. J Am Math Soc 1(4):805–866
29. Mane R (1977) Reduction of semilinear parabolic equations of finite dimensional c^1 flows. In: Geometry and Topology, no. 597 in Lecture Notes in Mathematics. Springer, New York, pp 361–378
30. Marmo G, Saletan EJ, Simoni A, Vitale B (1985) Dynamical Systems: a differential geometric approach to symmetry and reduction. Wiley, New York
31. Marsden JE, Ratiu TS (2002) Introduction to Mechanics and Symmetry: A Basic Exposition of Classical Mechanical Systems, Texts in Applied Mathematics, 2nd edn. Springer, New York
32. Marsden J, Misiolek G, Ortega JP, Perlmutter M, Ratiu T (2007) Hamiltonian Reduction by Stages. Lecture Notes in Mathematics. Springer, New York
33. McLachlan RI, Perlmutter M, Quispel GRW (1998) Numerical integrators that preserve symmetries and reversing symmetries. SIAM J Numer Anal 35(2):586–599
34. McLachlan RI, Perlmutter M, Quispel GRW (2003) Lie group foliations: dynamical systems and integrators. Futur Gener Comput Syst 19(7):1207–1219
35. Molino P (1988) Riemannian Foliations. Birkhäuser, Boston
36. Nilsson Jacobi M, Görnerup O A dual eigenvector condition on lumpability in markov chains. To be published
37. Olver P (2000) Applications of Lie Groups to Differential Equa-

tions, Graduate Texts in Mathematics, 2nd edn. Springer, New York
38. Packard N, Crutchfield J, Farmer D, Shaw R (1980) Geometry from a time series. Phys Rev Lett 45:712–716
39. Rega G, Troger H (2005) Dimension reduction of dynamical systems: Methods, models. Nonlinear Dyn 41:1–15
40. Riegert A, Baba N, Gelfert K, Just W, Kantz H (2005) Hamiltonian chaos acts like a finite energy reservoir: Accuracy of the Fokker–Planck approximation. Phys Rev Lett 94:54–103
41. Risken H, Frank T (1996) The Fokker–Planck Equation: Methods of Solutions and Applications, 2nd edn. Springer Series in Synergetics. Springer, Berlin
42. Ruelle D (1989) Chaotic evolution and strange attractors. Cambridge University Press, Cambridge
43. Ruelle D (1999) Smooth dynamics and new theoretical ideas in nonequilibrium statistical mechanics. J Stat Phys 95(1–2):393–468
44. Sanders JA, Verhulst F (1985) Averaging Methods in Nonlinear Dynamical Systems, Applied Mathematical Sciences, vol 59. Springer, New York
45. Sauer T, Yorke JA, Casdagli M (1991) Embedolgy. J Stat Phys 65:579–616
46. Schulman LS (1996) Techniques and applications of path integration. John Wiley, New York
47. Shalizi C (2001) Causal architecture, complexity and self-organization in time series and cellular automata. Ph D thesis, University of Wisconsin
48. Shalizi C, Crutchfield J (2001) Computational mechanics: Pattern and prediction, structure and simplicity. J Stat Phys 104:816
49. Shalizi C, Moore C (2003) What is a macrostate? Subjective observations and objective dynamics. http://arxiv.org/abs/cond-mat/0303625
50. Shalizi C, Shalizi K (2004) Blind construction of optimal nonlinear recursive predictors for discrete sequences. In: AUAI '04: Proceedings of the 20th conference on Uncertainty in artificial intelligence, AUAI Press, Arlington, pp 504–511
51. Simon H (1962) The architecture of complexity. Proc Am Philos Soc 106(6):467–482
52. Smale S (1970) Topology and mechanics. Invent Math 10(4):305–331
53. Takens F (1981) Detecting strange attractors in turbulence. In: Rand D, Young L (eds) Dynamical Systems and Turbulence, Warwick 1980. Springer, Berlin, p 366
54. Zwanzig R (1960) Ensemble methods in the theory of irreversibility. J Chem Phys 33:1338
55. Zwanzig R (2001) Nonequilibrium statistical mechanics. Oxford University Press, New York

Books and Reviews

Anderson PW (1972) More is Different. Science 177(4047):393–396
Badii R, Politi A (1997) Complexity: Hierarchical structures and scaling in physics. Cambridge Nonlinear Science Series, vol 6. Cambridge University Press, Cambridge
van Kampen NG (1985) Elimination of fast variables. Phys Rep 124(2):69–160
Laughlin RB, Pines D (2000) The theory of everything Proc Natl Acad Sci 97(1):28–31
Laughlin RB, Pines D, Schmalian J, Stojkovic BP, Wolynes P (2000) The middle way. Proc Natl Acad Sci 97(1):32–37

Human Behavior, Dynamics of

Douglas R. White[1,2]
[1] University of California, Irvine, USA
[2] Santa Fe Institute, Santa Fe, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Networks and Cohesion in HB Dynamics
Cooperation, Connectivity-*k* and "Critical Mass" in Collective Action
Transition Models with Thresholds
Aggregate ("Sufficient Unit") Equation-Based Modeling
Institutions, Network, Economic Models and Experiments: Testing Causality
Future Directions
Acknowledgment
Bibliography

Glossary

Connectivity-*k* (*k*-connected, *k*-cohesive, structural cohesion, cohesive.blocks) refer to the Menger (1927) theorem for structure/traversal isomorphism in graph theory, as explained in the text, where k-components are the largest possible expansion (maximal group) that preserve structural k-cohesion. Computation is provided by cohesive.blocks in the igraph R package.

Scale-free network where the probability that a node i in the network connects with k other nodes is inversely proportional to the number of k's links (see: power law), more generally, $p_i(k) \sim k^{-\gamma}$, with $\lambda = 1$ for scale-free.

Nonindependence is characteristic of complex phenomena with built-in interdependencies, where distributions of attributes or relations should not be not directly subject to statistical inference using the null hypothesis of independence, as in structural measures sampled from networks, and autocorrelated time series or autocorrelated spatial distributions.

Sufficient statistic a *sufficient statistic* for a statistical model is one that captures the information relevant to statistical inference within the context of the model, including the size and composition of the units of study. Let $X_1, \ldots, X_M$ be a random sample, governed by the density or probability mass function $f(x|\theta)$. The statistic $T(x)$ is sufficient for θ if the conditional distribu-

tion of x, given $T(x) = t$, is independent of θ. Equivalently, the functional form of $f_{\theta|x}(x)$ does not involve θ, and the Fisher–Neyman factorization theorem may be used to help spot sufficient statistics. The likelihood ratio test can often be reduced to a sufficient statistic of the data for hypothesis testing. The minimum variance unbiased estimator of a parameter θ can be characterized in parameter estimation by sufficient statistics and the Rao–Blackwell Theorem. See Scharf *Statistical Signal Processing* [107]. A sufficient unit is one for which a random sample of aggregate statistics are sufficient.

Aggregate ("sufficient unit") equation modeling assumes that causality can be found with quantitative equation models that use **sufficient statistics**, which implies that the aggregate units studied have cohesive mass or entitivity for causal interactions to act on their aggregate characteristics. See Sect. "Aggregate ("Sufficient Unit") Equation-based Modeling".

NP-complete (NPC) algorithms require an order of nondeterministic polynomial time (NP) but are exceptionally difficult: if a deterministic polynomial time solution can be found for any of them, it would provide a solution to every other problem in NP and empty out the class of NPC.

Dictator game where the first player proposes a split of some endowment and the second, entirely passive, receives the remainder. Not formally a game at all (as the term is used in game theory, where every player's outcome must depend on the actions of others), it is used in decision theory to test the *homo economicus* model of individual behavior, where selfishness would dictate allocation entirely to oneself. Henrich et al. [50] discovered in a 15-society cross cultural study that people do allocate a share of the endowment to others. Skyrms [113] gives the dynamics of an evolutionary game theory variant.

Concentration indices such as the Laakso–Taagepera Index $1/\Sigma_i p_i$, where p_i is an effective proportion weighting for each unit, are used for problems such as "what are the effective numbers of political parties self-weighted by their membership (for polities: by their population or area)", e. g., US party proportions {.49, .49, .02} would have an effective number of 2.08 while France with 101 parties (each weighted by its number of members) might have effective party number of 22.1.

Power law is a Pareto distribution where probability $p(x) \sim x^{-\alpha}$, as for example: "multifractals have tails that follow a power law" (p. 209 in [75]) in how the frequency of similar units at different scales varies with the scale; see multifractal. Power laws tend to become ubiquitous when what is studied involves dimensional constraints. Power-law *growth* is expressed as $N = K/(t_0 - t)^k$ where K is an initial constant, t is calendrical time, and t_0 is the calendrical singularity date at which $K/(t_0 - t) = K/0$, where division by zero produces dynamical instability as $K/(t_0 - t) \to \infty$.

Fractal is a pattern or object (e. g. geometrical) whose parts echo the whole, only scaled down, i. e., scale invariant; invariant at any scale of magnification or reduction. Fractal prices occur when positive and negative changes in prices (daily, weekly, monthly, yearly) follow a power law. "To improve almost any fractal model it is a good idea to replace it with a multifractal one" (p. 209 in [75]). A multifractal (with root and generator) is a composite pattern that begins with an initial root (e. g., a straight line) that is successively replaced with a generator (e. g., a zagged line) that replaces every instance of the initial element. See power law.

Causality is a relation holding between two variables such that manipulation of one of the variables (the potential cause) is reliably associated with variation in the other (the response), for some configuration of the values of other potential causes of the response. Estimation includes classical structural equations approaches [74], the treatment effects framework [102,103], the directed acyclic graph (DAG) probabilistic approach [95], and the settable system probabilistic approach that unifies all three [141]. Another aspect of causation is probabilistic evaluation and decision theory, in which case the effect of evidence in revising beliefs about causation can be studied in a Bayesian framework [28,112]. Probability of causation is not causation of probability, although there are probabilistic causative models.

Definition of the Subject

Dynamics of human behavior (abbreviations DHB, HB, HD) deals with the effects of multiple causal forces in human behavior, including network interactions, groups, social movements, and historical transitions, among many other concerns. Description of movement and change distinguishes kinematics from statics, while dynamics considers causes of movement and change. Pearl [95] summarizes issues of causality with two fundamental questions: (1) What empirical evidence is required for legitimate inference of cause-effect relationships? (2) Given that we are willing to accept causal information about a phenomenon, what inferences can we draw from such information, and how? Policy issues entail beliefs about causation and open

a second framework for evaluating beliefs about causality [28,112]. HB dynamics is a field replete with new discoveries—and applications of methods derived from problems and principles that apply across disciplines. Insights transfer across disciplinary boundaries. This is because research strategies for studying causality in a hierarchy of sciences are typically not a reductionism of one level to another but involve recognition of emergence at different levels. Common principles that apply are often shared but with different detailed applications more finely tuned to irreducible aspects of concurrent phenomena. Mathematics and physical principles apply at various levels in the scientific disciplines, but principles discovered in the human and evolutionary sciences are increasingly found to apply and generalize as well.

DHB takes into account the distinctive behaviors of humans and the range of their sociopsychocultural variations. Focusing on causes, HB dynamics may refer, for different levels of social entities, to spatial and temporal, local and long-distance interactions, growth and decline, oscillations, changes in distributional properties, and synchronous or time-lagged causality in dynamical evolution. Examples of precursors in DHB include Ibn-Khaldun's (c.1379) dynamical characterizations of the oscillations of Muslim and Berber political dynasties and charismatic tribal initiatives [60]. Ibn-Khaldun's work was an extraordinary early precursor of the empirical study of oscillatory sociopolitical dynamics (as contrasted with beliefs in cycles of renewal, for example, derived from experience with cycles in nature) and is incorporated into contemporary DHB modeling. Similarly, Richardson's *Statistics of deadly quarrels* (1960) searched for causality of war and posed behavioral dynamic equation-based decision models with basins of attraction for stability, disarmament, or the arms race [108,109]. Schelling's "focal point" solution in the study of strategic behavior and bargaining ("each person's expectation of what the other expects him to expect to be expected to do") advanced the game theoretic policy sciences while his *Micromotives and Macrobehavior* [101] was seminal for modeling complex causal feedbacks. Interest in lower-level processes and how they link to higher levels motivates much of HB dynamical modeling. This is the case as well in biological modeling, as in SFI researcher David Krakauer's statement of research on "the evolutionary history of information processing mechanisms in biology, with an emphasis on robust information transmission, signaling dynamics and their role in constructing novel, higher level features. The research spans several levels of organization finding analogous processes in genetics, cell biology, microbiology and in organismal behavior" [68].

Introduction

HB dynamics is grounded within an evolutionary framework and interacts well with research in biology and primate and human ethology. Fundamental problems in new and old approaches to HB dynamics include general approaches to identify and model (1) units of analysis, (2) interaction equations and structures, and (3) levels of analysis, with (4) sufficient statistics. Many problems concerned with the "units" of investigation, organized into systems, are multifractal, and are explored through detailed study of social organization, biological reproduction, evolutionary phylogeny, and developmental ontology. A focus on networks recognizes the fluidity of dynamical interactions in living systems (i. e., recognizing the limits of hard-unit and hard-wired modeling). Network analysis also links to hydrodynamics, nonlinear synchronization, percolation, and other physical processes as well as models derived from the study of graphs and lattices. Generalizations of entropy measures may also provide approaches for testing general principles in physics that are more useful than mechanics, solid state physics, or conventional models of entropy. While many principles of complexity sciences will apply across many disciplines, how they apply varies with subject matter.

Formal approaches to HB dynamics—where *formal* means theories have been stated in a formalized language, usually mathematical, that does not allow for variable readings [70,120,121]—require construction on the basis of careful descriptive, qualitative, and quantitative research about human behavior and institutions such as are independently carried out in the disciplines (history, sociology, economics, psychology, cognitive science, political science, linguistics, and anthropology, including ethnography, archeology and other domains) as well as in cross-disciplinary fields including those of complexity sciences.

The modeling of human behavior is still in its infancy, and there are likely to be widespread advances in many different areas in coming years. The examples here show a range of concepts and practices but are not intended to cover all of the definitive techniques for modeling human behavior. Among the formal and complexity science approaches in HB dynamics, some of the examples include network modeling, aggregate equation-based modeling, and simulation modeling (equation or agent-based, or both), and how these deal with problems of non-independence. Network modeling depends on finding means of bounding and measuring fields of interaction where particular kinds of units and their causal interrelations can be specified. "Sufficient unit" modeling looks for aggregates at particular scales that represent relative closures of

systems in which causality from internal dynamics can be studied for certain types of relatively well-bounded units that occur within limited ranges of scale. Briefly exemplified are institutional studies of the evolution of market systems extended by experiments in network economics. Not covered are generalized "open system" entropy maximization [119], fields such as fractal dynamics that have challenged fundamental economic axioms and start with the notion that "units" of behavior operate with memory compressed through repetitions of structure that are not dependent on scale. The topics and examples presented form an overall outline about structural k-cohesion and resistance as measurable social forces in human behavior; what enhances or limits scalability in cohesion; what produces and inhibits resistance; and the multiple ways that these two social forces, very different from physical forces, interact dynamically.

Networks and Cohesion in HB Dynamics

The interconnected theme of these illustrative examples will vary from basic measurement to exploratory models to findings built on the mathematics of universality classes, focusing on two features of human ethology that make for unusual dynamics. One is an open-field bonding ability, like chimpanzees, gorillas, and orangutans, which involves recognition of community organized by weak rather than strong ties [39,78]. Humans are additionally equipped with a huge range of social and cultural abilities that derive from our use of symbols, which can widen community and cohesion and enable scalable networks of trust through strong ties as well [135]. These emergents can alter the scale and especially the dynamics of human social organizations. One foundational base for a theory of such emergents are the scalable cohesive groups whose boundaries are identified with the concept of *structural k-cohesion* in sociology [83,134], with new parallels recently discovered in the signaling properties of human and biological networks [104]. Another is the role of *k-cohesive resistance* in human ethology. Taken together, the scalability of cohesive human groups, which allow the scale-up of group sizes that contribute greatly to political expansion and warfare, and the role of decentralized cohesive resistance in pushing back political aggression, exhibit some of the properties of laws of momentum and of proportional reaction, not atypical of complex systems with complex interiors.

To understand the potential for such regularities in phenomena as irregular as human sociopolitical histories (ones that were not lost on the pre-Einsteinian Henry Adams [53]), the concepts underlying indefinite extensibility of scalably emergent cohesive human groups need to be carefully drawn. Rather than harking back to Ibn Khaldun, they draw on Menger's 1927 theorem [80] for graphs or networks, which is now in use in sociology [83,97,134,138] and anthropology [17] even if rarely used in physics or chemistry, although applications are beginning in graph-theoretic formalization of biological signaling network models [104]. In a network of connected elements, a maximal group (one that cannot be expanded further without losing the property) with structural cohesion k is one that (a) cannot be disconnected without removal of at least k elements, and which, as proved by Menger [80], is equivalent to its having (b) at least k element-disjoint paths between every pair of elements. Property (a) provides *external resistance* to complete disruption (i. e., removing fewer than k elements leaves the structurally k-cohesive group connected), and property (b) proves the existence of a measure k of *internal cohesive traversal* through concomitant existence of at least k redundant paths of transmission or potential communication between every pair of elements. Neither the internal nor the external *cohesive* properties can be surpassed by extending its boundary to include others, whereby each structural cohesion k-group has a unique social boundary. Perfect *scalability* occurs for the numeric size of the *intensive* variable k by any scale-up *extensive* multiplier m because while a structurally k-cohesive group of size n requires only $k < n$ links per element, the same is true at size nm. Note that while dying or migrating might be due to external forces or attractions that remove people from groups, sometimes group members themselves decide to leave, or are expelled. This raises the point that cohesion models and measures are appropriate where the inter-element or interpersonal ties are positive, not antagonistic or negative, by restriction on what should be included in such a model.

Broad problem areas of HB dynamics can be understood from the pairing of (1) the indefinite extensibility of scalably emergent structurally cohesive groups (which have an indefinite supportive potential for scale-up in size of cooperative groups and community) with (2) the contending abilities to form both emergent centralized social structures and (3) cohesive resistance to invasion or centralized authority. HB dynamical processes that can be phrased in terms of symbolic and social interactions of types (1)–(3) are discussed in Sect. "Cooperation, Connectivity-k and "Critical Mass" in Collective Action". Central to these issues, John Turner's (2002) *Face to Face: Toward a Sociological Theory of Interpersonal Behavior* [128] presents evidence for the deeply rooted ethological two-sidedness of humans as a species pitting *cohesion* against

resistance. A reviewer's summary is worth quoting:

> Turner forcefully argues that we are not the solidarity-seeking emotional animals that theorists like Durkheim, Goffman, and Mead would have us to be. Nor are we normally the tortured beings of the Freudian perspective. Reflecting our origins among the great apes, we are a deeply ambivalent species of two minds, craving strong emotional attachments and at the same time, bridling against the constraints in closed social circles of even strong interpersonal ties. Turner argues that this two-sidedness is rooted deeply in our biology, and is not simply the product of historically specific ideologies and social structures. Clearly this viewpoint has enormous implications for the study of face-to-face interactions, as well as many other aspects of sociology. However, in his deep respect for the traditional perspectives in this field, these implications are often obscured and hidden in Turner's exegesis of the general problems and principles in this area of study. None of the other theorists analyzed here have created a better model of ambivalence. Capturing the two-sided nature of social linkages was not a key part of theorists such as Mead, Goffman, and Schutz. Freud made ambivalence central to his model, but locked it into a narrow sexual model. As Neil Smelser has argued, the future of sociological theory will depend in large part on its ability to deal with ambivalence, and Turner's model goes a long way in this regard [45].

Issues of two-sidedness, through a number of steps in logic and measurement, are not unrelated to those of scalability in structural cohesion. To clarify the first three steps in this logic, we can refer to the number of elements in a maximally-sized *k-cohesive* group as its *k-cohsize* or extension and so state, for clarity, that *k-cohesion* and *k-cohsize* $= n$ ($> k$ by definition) of such a group can vary independently for a given level of *k-cohesion* that defines the boundaries of a particular subgroup in a network:

Step 1. *Intensive versus extensive aspects of structural k-cohesion are independent*. Evidence of the causal effect of *k-cohesion* is found in empirical studies and is unrelated to *k-cohsize*. There are three major tests of this to date, one where the major variance in student attachment to high school [83] (as measured by a half-dozen validated questions) was consistently predicted, in multiple tests (ten American high schools randomly selected from the 100-school sample of US Adolescent Health network surveys [10]), by level of *k*-cohesion in which each student was embedded in the school's network of friendships. With complete data on students and networks in each school, replication of this result was achieved in logistic regressions where all other attribute and pertinent network measures competed in accounting for variance. The influence coefficients for *k*-cohesion replicated in each of the ten independent populations [83]; and the *k-cohsize* of the friendship groups for individual students did not account for school attachment. Since these groups varied in size for each level of *k*-cohesion, this is evidence that the causal effect of *k*-cohesion is not diluted by size, that is, it is an intensive predictive property independent of its scalability in size.

In a second major study (Powell et al. 2006 [97]), Attraction to *k*-cohesion along with recruitment of diversity were the major predictors in a 12-year time-series analysis of variables accounting for tie-formation probabilities proportional to *k* in the biotech industry. Because of the recruitment of new entrants, with fewer ties the overall industry levels of cohesion varied relatively little and oscillated in alternation with 3-year to 4-year waves of variation in attracting new recruits. While *k* did vary slightly for the maximally cohesive core of the industry, it grew neither uniformly nor uniformly decreased over time. There is a consistency here with the finding that *greater cohesion* was the attractor in tie formation and not greater network centrality as hypothesized in the Barabási scale-free network model [8]. The tie-preference attractor was a *sufficient level of k-cohesion that is scalable by addition of members to the structurally cohesive group*, as is shown to occur over time in the biotech industry study.

A third study, of cohesive decay (White and Harary 2001 [134]), tested predictions of how a single 4-cohesive group disintegrated into two competing and eventually disconnected groups. With leaders in opposing groups, order of dissolution of ties followed the pattern predicted, as individuals dissolved their ties successively on the side of the leader with whom they had less *k*-cohesion, and if cohesion was equal, dissolved these ties to the opposing side that had the longer path lengths. While the larger 4-cohesive group dissolved, ties redistributed to the two smaller 4-cohesive groups that formed around the disputant leaders.

Step 2. *Further evidence for scalability* is that *k-cohesion* and *k-cohsize* measures find cohesive groups on much larger scales than do density-based measures called *community detection* [84,87] that split networks into mutually exclusive groups such that higher densities are within rather than between them. Calling these density groups "communities" ignores the fact that such groups overlap and form *k-cohesive* groups on much larger scales. Community detection lacks the scalability of structural *k*-co-

hesion. White et al. [138] demonstrate how the boundaries of these much larger cohesive groups can be approximated in extremely large networks, which is needed because cohesive.blocks computation is NP-complete. Further, they show how density-based and row-column correlation-based algorithms fail in 30 out of 31 methodological studies in a meta-analysis of a classical small-network dataset. They also show analytically, as do Harary and White [134], how *k*-cohesive components of networks stack hierarchically for successive values of *k*, providing core-group centralization in addition to horizontal cohesion within a group. The analytical properties of multiconnectivity (*aka* structural cohesion and also *k*-cohesion) as a precisely measurable and scalable concept (connectivity-*k* in graph theory) for hierarchical and overlapping group-cohesion boundaries make it an ideal construct for studying the relation between micro (small group and local network properties) and macro properties of social networks, those of political and other social units, and the social construction of roles and institutions [59]. A large number of studies show cohesive scalability in the ways in which symbols and attachments are deployed in human groups and networks, as will be discussed in Sect. "Cooperation, Connectivity-*k* and "Critical Mass" in Collective Action", although some effects are preserved only up to certain scale-up thresholds in group size.

Step 3. *Evidence discussed in Sect.* "Cooperation, Connectivity-*k* and "Critical Mass" in Collective Action" *supports the hypothesis that k-cohesive components of human social networks amplify transmission quality and the utility of information that can be cross-checked from multiple independent paths.* For example, distinguishing carefully between dominance (force or force threat) and prestige (freely conferred deference), generalized prestige rankings are scalable along with the transmission quality of multiple channels in *k*-cohesive groups, while dyadic dominance hierarchies are not. Henrich and Gil-White [47] tested and found support from data across the social sciences for the predictions of a prestige model of social learning as opposed to dominance imprinting. This supports their argument that "natural selection favored social learners who could evaluate potential models and copy the most successful among them," and that prestige rankings were an emergent product of psychological adaptations that evolved to improve the quality of information acquired via cultural transmission. Finally, studies of networks where utility is gained from long-range interactions [21,56,117] show a variety of network topologies that may combine the benefits of centralized hubs (which are often thought in network economics to maximize efficiency by minimizing redundancy) with those of redundancy in *k*-cohesive components.

The approach to cohesion taken here—also contrasting to methods for the partitioning of roles [100]—is not that of trying to specify analytical boundaries using matrix-based methods (Newman [85,86]), which are insufficient as tools to capture the precise boundaries and overlaps in the concept of *k*-connectivity. The analogy between physical forces and social cohesion or repulsion breaks down because the latter do not involve the kinds of hard-body (Hamiltonian) equations used to describe simple systems such as a bouncing ball, billiard balls, a pendulum, or an oscillating spring. The algorithmic complexity identifying *k*-cohesive units given an arbitrary graph is NP-complete and not susceptible to matrix-analytic detection, although humans are often better at perceiving accessible and simple but algorithmically complex patterns than are computers.

Because a great many fundamental issues in HB dynamics can be framed in the context of the pairing of cohesion and resistance—similar to but much more complex than the concept of attractive and repulsive force—this pairing is used to organize many of the research questions and findings presented here, not the least of which is related to the problem of the units of analysis needed to for tests of HB dynamics, and how these units interact or embed in one another.

How, for example, do scalability and resistance play out in terms of HB dynamics on the larger historical scale? Peter Turchin's [121] examination of 50 cases (Fig. 1) in the historical military expansion of agrarian states in European history over the last two millennia is illustrative as a test of historical DHB theories that engage concepts of social cohesion. What happens when agrarian states or empires invade a sizeable group that differs in major metaethnic markers (multiple cultural differentiations in *religion, language, and ethnicity*) that are internally cohesive for the group invaded? The framing of this problem is given initially by comparison of dynamical equation-based models for ordinary differential equations of zero-order (unbounded growth or decline), first-order (bounded growth or decline), and second-order (oscillatory growth and decline) [120,121]. Empires show growth and collapse that fail to conform to the first two types of dynamical equations, but could be governed by a second-order dynamics in which there are time lags and negative feedback. The next steps in this study engage the ethological issues that will also be examined here. For example: *What accounts for the resistive capabilities of human social groups, e.g., against outside invasion?* (It is useful to recall that this research was finished before 11 September

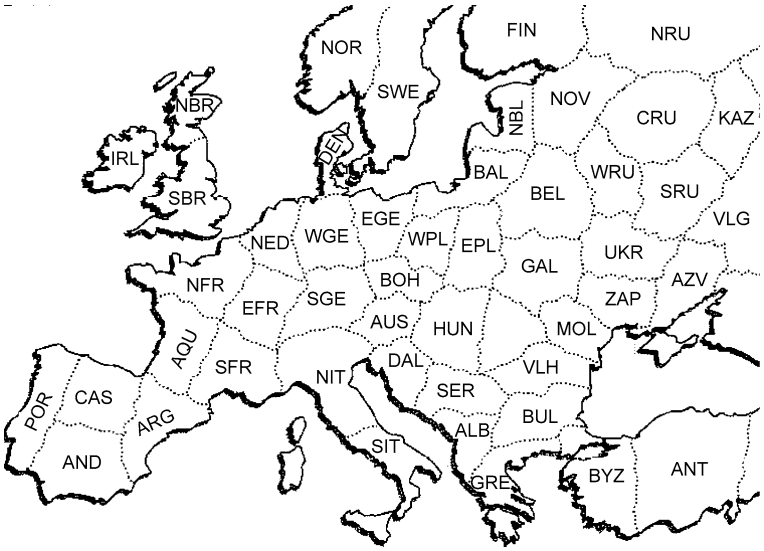

Human Behavior, Dynamics of, Figure 1
Turchin's [121] 50 cultural regions used as geographical units in the statistical analysis of the relationship between metaethnic frontiers and polity size (courtesy of the author)

2001). The study draws parallels with the dynamical theory of Ibn Khaldun, who used the term *asabiya* for collective solidarity:

> Ibn Khaldun was clearly aware of the nested nature of ethnic groups, and that each level has its own *asabiya* associated with it.... [T]he leading or ruling element within a group must be vested in a family or lineage that has the strongest and most natural claim to the control of the available *asabiya* (Ibn Khaldun [60]). Only the leader who controls an *asabiya* of sufficient strength may succeed in founding a dynasty (pp. 38–39 in [47]).

Ibn Khaldun is widely credited as a thoroughly modern sociological scientist of culture, knowledge, politics, and urban life and in his theory of oscillations of Arab and Berber polities. His theory and historical analysis is framed in terms of second-order dynamics:

> It [the theory] is held together by his central concept of "*asabiyyah*", or "social cohesion." It is this cohesion, which arises spontaneously in tribes and other small kinship groups, but which can be intensified and enlarged by a religious ideology, that provides the motive force that carries ruling groups to power. Its inevitable weakening, due to a complex combination of psychological, sociological, economic, and political factors, which Ibn Khaldun analyzes with consummate skill, heralds the decline of a dynasty or empire and prepares the way for a new one, based on a group bound by a stronger cohesive force [29].

The thesis of the 50-case study of European military expansion is that "areas where imperial and metaethnic frontiers coincide act as *asabiya* incubators" (p. 56 [121]), areas where new ethnies (i. e., ethnicities, nationalities) are born in the growth of collective resistance. These solidary groups with high *asabiya* have the attributes of *k*-connectivity: "An important element of the theory is the ability of ethnic groups to scale up without splintering into subgroups" (p. 57 [121]). Examples of integrative mechanisms in this particular context of differing ethnies are *religion*, *society-wide mechanisms of male socialization*, and *rulership with primogeniture*.

External conflict has long been seen to stimulate cohesion on both sides of the conflict boundaries [24,111], as exemplified in the fault line frontiers in history [79,127] and in the marcher state [9] conflicts along these frontiers. A remarkable display of the dynamics of history for the 50-case study is provided by the maps constructed to show, for the regions included in Fig. 1 and for each of the last 20 centuries, the invasions by European empires across metaethnic frontiers and the resultant appearance of new nationalities as resistive movements and states [122].

This mathematical model for empire expansion lacks "a well-developed theory that would connect micro-level individual actions"—like those deriving from structural cohesion—"to macro-level dynamics of *asabiya*" [121] al-

though the altruism of *asabiya* is seen to follow a conditional altruist model (like that of kin-selection [118]) that depends on cohesion with other altruists—discussed in Sect. "Cooperation, Connectivity-*k* and "Critical Mass" in Collective Action" below. A provisional model (later improved) is given, in its simplest form (pp. 64–66 in [121]), for a polity with a spatial scale h of power projection (imperial "reach") over an area $A > 0$ and the resistant cohesion $0 < S < 1$ of *asabiya* with an everpresent/constant minimum geopolitical pressure a from the hinterland across a metaethnic frontier of size b. This is given as two dynamical equations with negative feedback that give an unstable equilibrium with a single boom/bust cycle (c_0 and r_0 in these equations are constants):

$$\dot{A} = c_0 AS\left(1 - \frac{A}{h}\right) - a$$

$$\dot{S} = r_0\left(1 - \frac{A}{2b}\right) S(1 - S)\,.$$

Here, change in area is a function of the polity area and of cohesion, limited by overextension, while the function for change in cohesion has an in-built oscillatory dynamic affected by the size of the metaethnic frontier. These dynamics, although intended only to characterize the problem, are informative as to how its parameters play into dynamical complexity. If "reach" h is not much greater than frontier width b, the empire can reach a stable equilibrium, while if $b < h/4$ the boom/bust cycle, but only one can occur, not more. Only when the model incorporates the discounting of expansionary power with distance in a spatial simulation is a second-order type of effect is obtained, that of oscillatory growth and decline.

Rather than having this model serve to study attacks and resistance, and the influence of *relative* cohesiveness in outcomes of politicomilitary contests (which is difficult to measure), Turchin's frontier theory is tested instead with the time-lagged prediction for each of two millennia that *when the metaethnic frontier is intense in the first half of the millennium*, for one of the cultural areas in Fig. 1, then *large territorial polities (empires) will originate in the second* [121]. The evaluation is whether the expansive tendency *originated* in a contest of respectively cohesive entities rather than trying to predict the outcome of the battle, the more relevant outcome being that—having developed its cohesion through external conflict—the unit that is initially attacked may eventually enlarge to become an empire. This holds for 11 out of 15 cultural regions that were on the metaethnic frontier, while out of 34 regions that were not on the frontier, only 1 developed an empire in the first millennium AD; and it holds for 22 of 28 Fig. 1 frontiers in 500 AD–1500 and empires in 1000 AD–2000. The four exceptions in the first case and the six in the second were regions incorporated into an empire centered in a neighboring region.

Human Behavior, Dynamics of, Table 1
Cross tabulations for politics that start on frontiers and end as empire a millenium later

0–1000CE	Starts as frontier	No frontier
Becomes empire	11	1
No Empire	4	34

1000–1900CE	Starts as frontier	No frontier
Becomes empire	22	3
No Empire	6	19

50 regions, $p < .0000004$

How many empires *were observed that lacked the temporal precondition of a metaethnic frontier* (with subsequent growth of resistant cohesion)? The exceptions are 1 and 3 for the two periods, respectively. The first, and two of the latter cases, occur where the existence of the frontier was of short duration. One polity (Savoy-Sardinia, founding Italy) remained as a true exception, in a population formed by Celts and Romans, but with no clear causal path from metaethnic frontiers to polity expansion. But the major result is that the empires of the later periods *did* (and not just may) result in almost all cases from cohesive resistance to the attacks of the previous period, at long time scales.

Thus, these results are fully consistent with the scalability of structural cohesion as a basis of sociopolitical support for military expansion of polities (potentially into empires) but more importantly are supportive of the theory that *k*-cohesive structural resistance, which grows slowly on the metaethnic frontiers of expanding empires in such a way as to facilitate the growth of resistive "nationalistic" ethnic solidarity and eventually of consolidation of resistive metaethnic frontier groups themselves into expanding polities and empires, with long time-lags in their development.

Lim, Metzler and Bar-Yam [72] analyze local conflicts between distinct ethnic or cultural groups within multiethnic states (India and former Yugoslavia), matching actual conflicts to spatial population structure in a simulation model of type separation, where cohesion emerges through movement to more homogeneous regions and through avoidance of conflict. Conflicts are predicted due to the structure of boundaries rather than between the groups themselves, consistent with Turchin's [121] findings. The local ethnic patch serves as an "order parameter" to which aspects of behavior are coupled in the dynam-

ics of a universality class of collective behavior. Similarly, the multilevel evolutionary model of Garcia and van den Bergh [36] shows how parochialism, as altruistic behavior specifically targeted towards in-group members, can result from group selection operating on direct conflict between groups.

Cooperation, Connectivity-*k* and "Critical Mass" in Collective Action

If structural cohesion is scalable, what are the factors, aside from external conflict, that would prevent or facilitate the scale-up of cohesion? Or of group size generally, assuming some modicum of cooperation [73]? The major problem in explaining why cooperation should occur at all in human groups, in the absence of external conflict, is that of the benefits of selfishness to free-riders when others bear the cost of altruism. One component of "The Tragedy of the Commons" [46] is that collective goods [92] are nonexcludable: Once achieved (like peace, clean water or air, public transport, or wage contracts) they are available to everyone. Many if not most such goods have jointness of supply (available to all), i. e., their cost does not increase proportionally to group size. The initial problem is that if it takes only some initial investment and costs by those who bring such goods into existence, why should anyone else bear these costs when they can have them for free? This creates "the dilemma of cooperation" [92] and of collective action. And the larger the group the easier it is to ride free. Evolutionary game theorym [89], with a replicator dynamic that favors those with lower cost for the same benefits, predicts that without some compensation for altruism, even starting from a small number of free-riders in a population, selfishness becomes the norm. The secondary problem of collective goods is who will bear the costs to maintain them?

Reputation may attach positively to altruism and negatively to free-riding. In this respect, two recent experimental papers are strongly supportive (although unaware) of connectivity-*k* in helping to explain cooperation in human groups [18]. In one study the judged veracity of gossip is shown to increase considerably if it came from more different sources [54], not if one source kept repeating the same gossip, while another relates gossip to reputation and cooperation in general [114]. James West in *Plainsville* (1945) [130] was the first to connect gossip and the maintenance of the unity of groups. According to Gluckman (p. 308 in [37]), however, West misinterpreted the extent to which "gossip does not have isolated roles in community life, but is part of the very blood tissue of that life." Gluckman refers to Colson's *Makah Indians* [23] ethnography to illustrate the importance of gossip to the unity of groups.

While diffusion of reputation along the node-disjoint paths of *k*-components can provide benefits to altruism, its influence diminishes as groups grow larger and average network distance grows large, reducing the scalability of *k*-components with high levels of cooperation. Further, if a group has *too much* *k*-connectivity (as in completely connected cliques), the benefits of reputation diminish because of the "echoing" effects of conformity and diminution of independent sources of information [19,55,145]. In cliques or overly-connected groups, single dominant individuals have the potential to influence everyone and thus to distort the robust veracity of information. Further, studies of human friendships and other long-term relationships show that the success of reciprocal strategies (such as tit-for-tat) relies on a combination of medium-term accounting, forgiveness, and propensity to defect with strangers if they already have an adequate number of partners [55].

Benefits of punishment within a group have a similar profile of optimality to reputation. Like gossip and reputation, punishment can be effectively delivered through *k*-cohesive independent paths and thus diffuse coherently respecting the boundaries of *k*-cohesion [18] (although for a given k-cohesive group, the paths used to diffuse reputation need not be the same as those used to deliver punishments). This works best for groups with moderate average distances in the network and with (or defined by) moderate connectivity-*k*. Similar to reputation, "cohesion extends punishment even beyond the community network and protects insiders against trouble-making outsiders, [especially] when community members come to defend fellow community members against norm-violating outsiders" [18], while incidental defectors at the margins of the cohesive group may have little impact on behavior within the group. Henrich et al. [50,51] ethnographic-psychological study of 15 societies from five continents, representing the breadth of human production systems, found that willingness to use punishment in the dictator game covaries with altruism across populations in a manner consistent with coevolutionary theories. But while appropriate punishments diminish the relative rewards of free-riding, they also incur costs to the enforcers.

Punishment as third-party intervention tends to rely more on dominance and perceptions of the use of force, entailing higher risks in some cases, than on reputation in the modeling of advantages of cooperative behavior. After observing a group of macaques in captivity in which a small number of individuals fulfilled policing tasks, making interventions into dyadic conflicts, temporary ex-

perimental "network knockout" removals of the policing monkeys showed it was their presence that prevented the group from falling apart into small clusters [34,35]. Here, Jessica Flack and coworkers note [33], "the degree to which one individual perceives another as capable of using force is communicated using a special dominance signal. Group consensus about an individual's capacity to use force arises from the network of signaling interactions." Consistent with studies of k-cohesion, this research found that "coarse-grained information stored at the group level—behavioral macrostates – "was more useful than detailed information at the individual level". Because "successful intervention relies on consensus among combatants about the intervener's capacity to use force," use of " a formalism to quantify consensus in the network," and with consensus as a measure of power, showed that "the power distribution is fat tailed and power [here: consensus] is a strong predictor of social variables including request for support, intervention cost, and intensity." This modeling of power distributions shows how dominance *signaling* strategies "promote robust power distributions despite individual signaling errors" [34].

Third-party interventions in conflicts resemble recognition of community membership in that such interventions rarely occur with respect to outsiders. Recognized community boundaries (as distinct from k-cohesion, which may extend beyond these boundaries) provide the most probable context for the dyadic construction of cooperativity through reciprocity [118], dominance in third-party intervention, and dyadic game theoretic strategies that achieve cooperativity (such as tit-for-tat or lose-shift in Prisoner's dilemma) [113]. These, together with generalized reciprocity [71], i. e., altruism in the expectation of indirect return, are also among the most potent constructors of community-building strong ties in social networks [135], especially if they are navigable [1], as in many elite groups and non-Western [136] societies. Bowles and Gintis [15] summarize the game-theoretic work on cooperation showing that the critical condition for cooperative outcomes, which otherwise deteriorate with increases in group size, is the presence of *strong reciprocators*, who cooperate with one another and punish defectors, even if they sustain net costs, provided that they are more likely to interact with one another than at random. Thus, network structure and preferences (positive assortment) prove to be central to an evolutionary path to large-scale cooperativity. Pepper and Smuts [107] show how positive assortment through environmental feedback can play the same role. There are, then, evolutionary paths to the scale-up of k-cohesion for indefinitely large groups.

Putting together these principles of primate (reciprocity, policing) and human social networking, we can also see compatibility of k-connectivity with the theory of "critical mass" in collective action [77,90,91]. Group size *does* increase the probability of a critical mass of people who develop common goods through collective or cooperative action. This relates directly to the scalability of k-connectivity, wherein as the size $n$of such a structurally cohesive group expands, it is still only k links per person that are needed for k-connectivity. But there is always an expected excess of ties, upwards of k, for some members of such a group, and an increase in n increases the probability of formation for a group with a critical mass of connectivity $k + 1$ or higher ($k + l > 1$). This relates to the "paradox of group size" for collective action groups: "When groups are heterogeneous and a good has high jointness of supply [i. e., with cost that does not increase proportionally to size], a larger interest group [size n] can have a [relatively] smaller critical mass," which could also be a critical mass with connectivity $k + l$. The problem of mobilizing collective action is whether there is a mechanism that connects enough people with appropriate interests and resources so that they can act to construct a collective good [77]. Structural cohesion provides just such a mechanism [17,83,97].

An extended feature of this model of critical mass, which has been investigated through simulation [77], is an *accelerative function* for what has been called network externality [4], where every new participant in creating a collective good makes it more attractive for the next participant to join. Different forms of collective action have some mix of this source of nonindependent decisions and/or a *decelerative function* wherein free-riding is more likely on the belief that others will do the job. Since success in collective action is partly a problem of coordination, there is some advantage to members of a critical mass in collective action having greater centrality. But again, if the collective action group at large has connectivity k and the leadership critical mass has connectivity $k + l$, the latter is achieved by the hierarchical embedding of higher orders of connectivity and not necessarily by greater centrality of a single leader.

So there are two aspects to consider for the dynamics of growth and decline in size of cooperative human groups: (1) reinforcement mechanisms of community, which tend to be self-limiting with respect of structural cohesion, and (2) critical mass in collective action or positive assortative reciprocators [15], which both tend to be self-enabling. While collective action to produce a collective good also requires a model of group process that cannot be deduced from simple models of individual behavior [77], the prob-

lem for understanding how large-scale societies and polities can achieve sustainability may be solved by assortative strong reciprocators [15]. The former problem—of sustainability of cooperation in a community—is different. All of the mechanisms there—reciprocity, third-party intervention, reputation, and punishment—depend on relative stability of community membership. Prior to the electronic age (which poses somewhat different problems of stability in virtual communities but makes for more independence of *k*-connectivity from local density and geographic distance), stable communities carried designations such as "settlement" or "nomadic group" as nominal indicators of relative stability in proximal spatial interactions among their members.

There is a herd cohesion solution to the stability problem (follow the *surest* neighbor!) [25,26], but also an advanced social cohesion solution (the formation of *k*-connectivity groups with a stable core) [18]. Empirical studies of neighborhoods [105,106] show that a certain threshold of residential stability is a crucial factor for the efficacy of mechanisms for community-level enforcement of the cooperative norms of third-party intervention, reputation, and punishment (all but strictly dyadic reciprocity unless it is strong and assortative). "For cooperation to be maintained at the community level, the network as a whole must be relatively more stable than patterns of individual actions" [18]. Combining community mechanisms and critical mass in collective action, we have the foundations for an evolutionary theory of cooperativity and cohesion in human groups. Many of these features (but not structural cohesion) have been brought under the Darwinian umbrella in a way that shows how the co-evolution of culture and genes jointly influence cultural transmission (dual inheritance theory) through the vehicles of human behavior and psychology [48,52]. This framework allows the integration of work on kinship, friendship, reciprocity, reputation, social norms, and ethnicity into a generally applicable mathematical characterization that may contribute to solving the problem of cooperation and extending on to the evolution of evolution and of economic systems [49,50].

Beyond adding *k*-connectivity into dual inheritance theory, there are also newer models of achieving minimum punishment and maximum crime reduction through policing concentration on an arbitrary push-down set of offenders [64]. The theory here, validated in simulation and case study, is that a fair and effective law enforcement strategy can only succeed if it approximates one with a stable target set of offenders at whom punishment is directed until recidivism ceases, individual by individual, replacing each nonrecidivist on the pushdown list by another known offender chosen with a probability *proportional* to rate of current offenses but otherwise *arbitrarily*, i. e., *fairly*. Policing an arbitrarily stabilized set of offenders mirrors the requirement for stability in cooperative neighborhoods. *Stability seems to be a key ingredient for cooperativity.*

What are the implications of these findings for considerations of the scalability of human communities and of human polities? Although *k*-connected groups are scalable, the properties of third-party intervention, reputation, and punishment to maintain cooperativity are not scalable, nor is dyadic reciprocity except under very special conditions [135,136]. Scalability through conflict—resistance to threat—and through collective action to produce collective goods, organized by a "critical mass" or through assortative strong reciprocity [15] is, however, scalable.

So why do human groups not simply grow larger at all scales [73], as challenged by competing groups, or by possibilities such as establishing collective goods capable of sustaining growth? Prestate societies only rarely sustain continued growth in size, but rather split, and then remix through intermarriage and mating (fission and remixing), with transition to a higher-order political form occurring extremely sporadically. It might be thought that if politically independent groups fission but still remain linked, through intermarriage or *k*-connectivity, then fusion into larger political groups would be easy. Many anthropological theories assume a stage-wise progression such as band to tribe or tribe to chiefdom [110]. Comparative ethnographic, historical, and archaeological studies, such as those of Wright [143,144], however, make it clear that passages from band to tribe (concepts with serious conceptual problems) to chiefdom to state are extremely difficult and unlikely transitions. And as we have seen [121], growth in state societies and empires is followed by collapse and the rise of other polities instead. Models of political fission might provide necessary conditions for transitions in successions of forms of leadership as polities develop with different sets of roles. New role set configurations might also create founder effects in the emergence of economic or political forms.

Transition Models with Thresholds

Transitions such as chiefdom to state can be modeled in an evolutionary dynamics of human behavior framework that includes the interaction of ethological characteristics—general human behavioral tendencies—and forms of sociopolitical organization. Social anthropologist Christopher Boehm [13,14], whose field studies range from Montenegro [12] to wild chimpanzees at Gombe, called at-

tention to the *human tendency to resist domination* (consistent with Turchin's findings [121] in Sect. "Networks, and Cohesion in HB Dynamics"), which is not shared with other great apes (consistent with Henrich and Gil-White [47]). In a substantial cross-cultural survey of societies in a wide variety of social and ecological settings, Boehm selected those with egalitarian behavior, and found that their behavior was not shaped by these settings but rather was deliberately shaped by their members, guided by a nearly universal ethos in these societies "that disapproves of hierarchical behavior in general and of bossiness in leaders in particular." His survey reveals the wide variety of means by which "the political rank and file" evict leaders who evince excessive authoritarian tendencies. This "creates a reverse dominance hierarchy, a social arrangement that has important implications for cross-phylogenetic comparisons and for the theory of state formation" [13] that might be called a "law of human ethological resistance", consistent with [121]. One of these mechanisms of resistance is fission, the break-away from a group that is growing large or with too many settlements under a single leader.

Surveys of archaeological, historical, and ethnographic cases not only show transitions from chiefdoms to states to be very rare but also show that states are based on a radically different principle of a hierarchy of roles to which decision-makers are recruited. Henry Wright [143,144] shows that primary and secondary states have three or more levels of mobilization of resources upwards and passing of information both upwards and downwards through a hierarchy of divided offices and a division of political labor [115]. Chiefdoms, unlike states, are characterized by paramount leaders who delegate as little authority as possible, in contrast to states with their delegated division of labor for authority [144]. Paramount chiefs may govern subdivided territories with village chiefs and ritual specialists, but there are nearly always no more than *two* levels of chiefly resource mobilization conducting directly to the chief and *all political decisions are integrated into the chiefly persona*.

To assume a simple quantitative increase in network size and complexity as chiefdoms develop into states is therefore inappropriate. Chiefdoms are also characterized by a reverse ranking hierarchy [13,14], not an actual reversal of dominance but one of prestige ranking [47] in which leaders are expected to exhibit altruism to followers through redistribution of goods or forms of reciprocity and bestowal of favors or gifts to counterbalance the processes whereby resources were concentrated through interpersonal network ties, although the reciprocity is rarely balanced in any material sense [27,96,98]. In their dynamics of growth, chiefdoms—with their structural cohesion and cohesive hierarchies based on intermarriage, exchange, and cross-cutting ties—tend to increase in size through internal growth or annexation of settlements, then to give way to fission at times of crisis following growth, especially if these crises coincide with issues of political succession. There is no tendency in these dynamics for gradual cumulative evolution in complexity toward state organization. The mosaic of sub-chief territories mapped into the chiefly ranking are segments that recurrently separate and then re-form in successive periods of political change

Griffin and Stanish (p. 2,24 in [43]) provide evidence and a model for a tipping-point synchronicity threshold in the transition from chiefdoms to emergent pristine states in the Lake Titicaca case of Tiwanaku, c. 500 AD (outgrowing the territorially larger political formation at Pucara). The transition occurred archeologically and in a detailed simulation model after a long period of cycling in which multiple chiefdoms climb the population size gradient only to be fragmented by fission. There is strong empirical evidence for cycling in growth and fission. During the period of cycling, primary centers, population concentrations, and increase in both the overall productivity and population of the region occur sporadically without synchronization. Then, in one rapid burst, archeologically and in repeated probabilistic simulations, these previously unsynchronized features emerge synchronously, pushing past a probability threshold for fission. Figure 2, reflecting results from the simulation model, shows the variables affecting the fission of chiefdoms plotted against time for growth; then, as cycling occurs, setting the cycling time back to that of an earlier equal scale in size. The simulation data were also reaggregated for X as number of settlements under the chief, and could be estimated analytically for other variables such as X for communication time from center to furthest outlier. The first variable, Y_1, is one of exponential decay in the ratio of resistance strength to the leader's strength, which can be expressed as an exponential probability density function supported on the interval $[0, \infty)$, where $\lambda > 0$ is the rate parameter of the distribution. Since this is a discrete exponential distribution with $X \geq 1$:

$$p(X; \lambda) \sim \lambda e^{-\lambda X} .$$

The second variable, Y_2, increases with X on the assumption that the likelihood of resistance to the chief increases with time, or variables that cycle with time, such as number of settlements. The third variable, $Y_3 = Y_1 \bullet Y_2$, is the probability of fission due to resistance, which is humped because it is a product of distributions that have higher

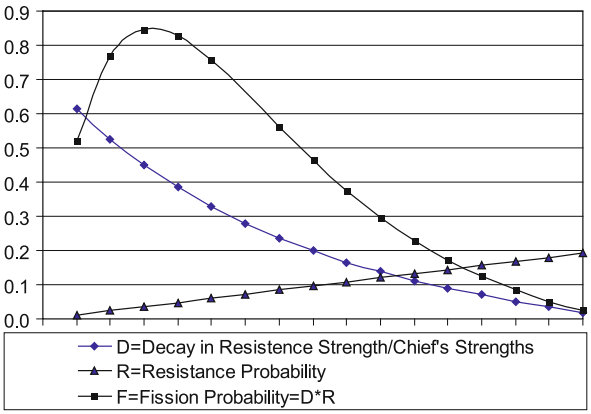

Human Behavior, Dynamics of, Figure 2
The transition threshold from Chiefdom to State: $Y_1 \sim$ Exponential Decay in the ratio of Resistance strength/Leader's strength; $Y_2 \sim$ Increasing] probability of Resistance as number of settlements increases; $Y_3 = Y_1 \bullet Y_2 =$ the probability of fission

resistance probabilities at opposite magnitudes of X. The shape of the Y_3 distribution emerges as an average over many simulation runs, and opens further questions for investigation. What emerged in the Lake Titicaca region, consistent with the simulation, were two dominant polities, separated in time, one an incipient state, the other smaller in population but larger in territory, along with extensive trade networks including the smaller centers. This is a typical multisite trading configuration of early states [3].

One can speculate as to whether resistive transition thresholds might occur also from band to tribe or big man (having occasional fission), to chiefdom and from chiefdoms (having occasional fission) to minimal states. Fissioning is by no means universal, and one study of the Titicaca region itself shows that village but not chiefly fissioning had ceased long before state formation with emergence of a regional religious tradition [7]. Are there resistive transition thresholds from minimal states (having occasional fission) to urbanized states, from urbanized states (having occasional fission through colonization) to dynastic states, or from dynastic states (with occasional fission with the death of a ruler and partition of domains under obligatory personal inheritance) to territorial agrarian states? Or do nonterritorial state expansions collapse, replaced by others? The territorial state, given institutional sovereignty over territory, is less likely to fission at a size threshold and its growth dynamics are shown in Sect. "Aggregate ("Sufficient Unit") Equation-based Modeling" to involve shrinkage following times of scarcity in population/resource ratios. This creates amplifications of inequality and internal conflict [126]. Modern mega-corporation growth is often arrested by national and international legal regulations mandating breakup of monopolies but there are no early barriers against corporate growth in size, although some corporations do fission for reasons other than size constraints.

The temporal scaling of long-term transitions in populations and sizes of the largest polities does show clear transitions, over 5000 years of world history, as shown on Fig. 3 [116]. The lower line in the figure is an exponential fitting of the effective number of polities (Laakso–Taagepera concentration index $1/\Sigma_i p_i$, where p_i is the effective proportion-weighting for each unit) weighted by their populations, and the upper line by their geographical areas. More even proportions for p_i, such as {.4 .3 .2 .1}, compared to higher concentrations like {.7 .1 .1 .1}, will have a higher effective numbers, 3.33 versus 1.92, while extremely concentrated proportions, e. g., {.97 .01 .01 .01} with effective number 1.06, approach unity. The declining slopes in Fig. 3 show a *decrease* in effective polity numbers $1/\Sigma_i p_i$ with greater concentration of population than of area (slopes differ by 2, the fitted exponential population roughly the square root of area). Over these five millennia the fitted effective number of political entities weighted by area decreased from circa one million to circa 64, and from circa one thousand to circa 8 weighted by population. For Fig. 3:

> Three sudden increases in polity sizes occur: [fewer large polity concentrations] around 3000 BC [urban revolution in Mesopotamia], 600 BC, and AD 1600 [the seafaring trade revolution]. This study tests the exponential model against area and population data [for polities] over five millennia. It also gives tables and graphs of area versus time for all major polities since AD 600. The median duration of large polities at more than half the peak size has been 130 years, and it has not changed over 5000 years. (p. 475 in [116])

Two of the three solid lines superimposed on the original figure show how two of the three elbows of change in the lower of the empirical data lines (circa 2600 BC, 1200 BC, and 200 AD), from polity population concentrations to dispersals, are followed with short time-lags by two similar elbows of change in the upper line from polity area concentrations to dispersals over the next hundred years. Whether these transitions represent eras of crises in urban empires is unclear, as are most extrapolations from so few data points. The first case might reflect the short-lived breakup of early Bronze-age polities (Mesopotamian and Indus) and the second the breakup

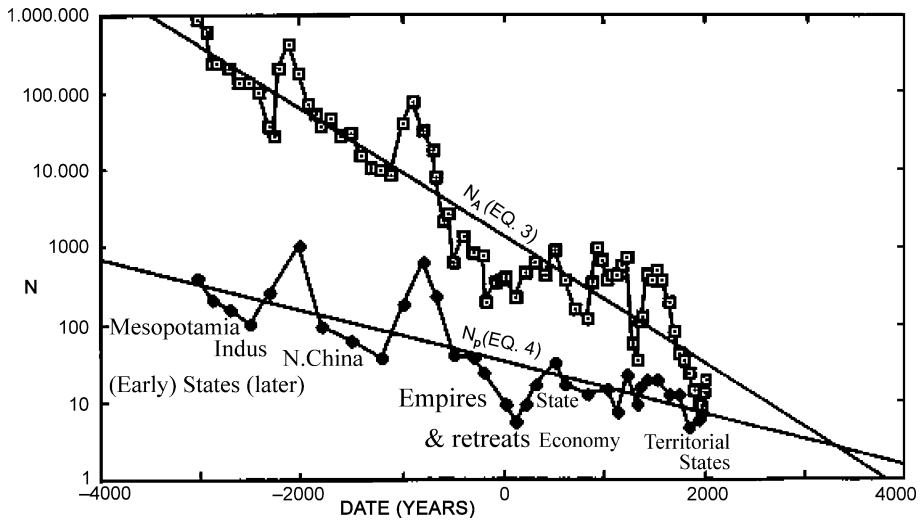

Human Behavior, Dynamics of, Figure 3

Transition thresholds for States and Empires (effective number of polities), based on area and on population (Taagepera 1997 Fig. 5, courtesy of the author, who notes that individual polities that expand slower tend to last slightly longer (p. 475 in [116]); *arrows* mark his dates for large polity concentrations; others are marked by *lighter lines*)

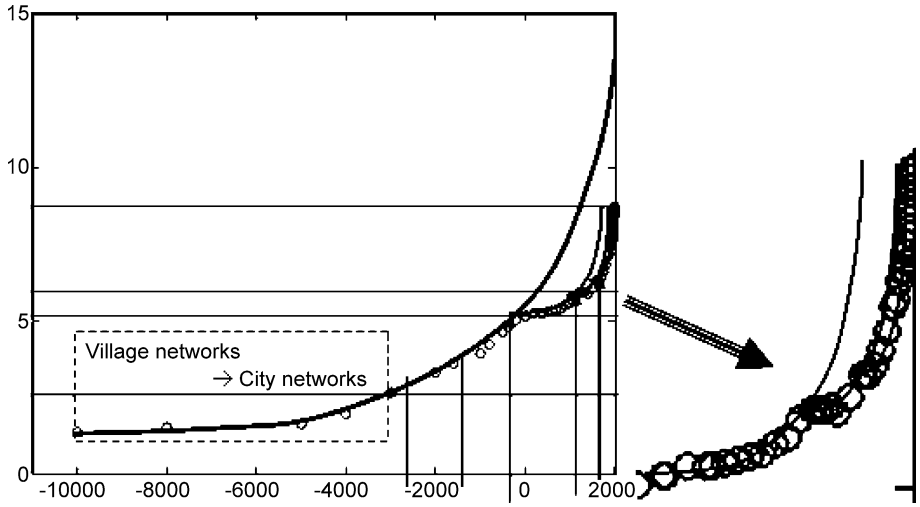

Human Behavior, Dynamics of, Figure 4

World population power-law growth spurts and flattening as shown in a semilog plot of Kremer's (1993) [69] data with successive power-law fits

of North China states. At 50 AD the third elbow of transition from population concentration to dispersion (e. g., for classical empires such as those of the Romans, Satavaphana and Han, which actively discouraged market developments) occurs in the dispersion phase commensurate with that of polity area. Population and area reconcentration in the next phase (ca. 500–850 AD) are also roughly commensurate. A downward spike of population concentration recurs circa 1050 AD when again it has a lagged effect on polity area concentration. It might be surmised that changes in population-area interactions in the era of power-law city growth are increasingly subject to market-driven trade routes (e. g., Silk Roads in Eurasia from 100 BC–1300 AD). This is a context, from 900 AD forward, of new national market economies that diffuse from Sung China to the west, where the Abbasid, Carolingian and related polities encouraged widely articulated market systems. Such changes are studied and modeled by

Modelski and Thompson [82]. Variability in the ways that markets change these political oscillations is particularly evident from 1800 and the industrial revolution, as nationalism and markets consolidate the effective number of polities geometrically weighted by size, up until 1990 with the breakups of the Soviet Union. There has been little dynamical modeling of the multiple causality in these coupled/decoupled oscillations. Taagapera (p. 488 in [116]) notes that while population concentration can continue to increase in the present era area concentration must stabilize because jumps to higher concentrations in earlier eras occurred with acquisition of control over large and sparsely populated areas (desert, steppe, deserts, tundra, respectively for Sargon, Mongol, British, Russian empires), and such areas are much less available now.

To express some of the consequences of the transition to networks of cities connected by trade routes, and eventually to market-driven trade, Fig. 4 sketches the suggestion that world population begins to grow not exponentially but in power-law growth spurts, correlated initially with the transitions noted in Fig. 3 [133].

Cities act as attractors for skilled, unskilled and intellectual labor as well as entrepreneurs and merchants, with a concomitant drain on settlements of smaller size. This enables power-law growth, at growth rates proportional to city size, i. e., cities as "attractors" as in the scale-free network model of Barabási 2000 [8] (but see [139]), while rural areas and smaller settlements do not diminish their population but with elevated birth rates can replenish their losses from outmigrants. This pattern allows world population to grow in power-law spurts, but power-law growth is self-limited by population crashes as it would otherwise grow to infinity in a finite time [129]. The places where the polity transition crises occur in Fig. 3, e. g., 2600–2400 BC, 1200–1100 BC, and 200–100 BC, and 1300 AD correspond to those crises in the larger states and cities where power-law growth in their (and world) population hits some sort of limit, growth flattens, and resets the starting parameters for a new upswing of power-law growth (a pattern first noticed but not explained in Korotayev et al. 2006 [66]). The largest world empires, of the Golden Horde Mongols and the British, appear as the result of two of these more recent upswings (a topic currently under investigation by Christopher Chase-Dunn in one of the National Science Foundation's Human Social Dynamics research awards).

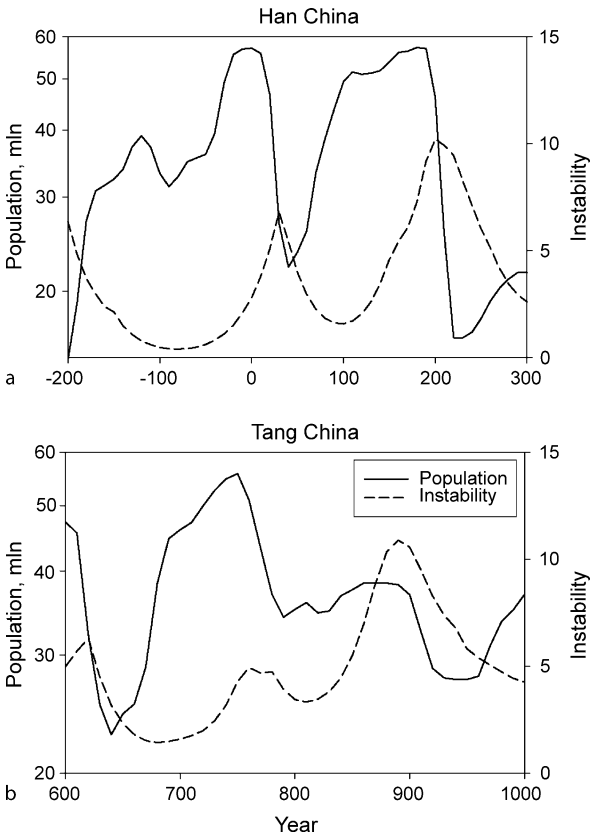

Human Behavior, Dynamics of, Figure 5
Population and Sociopolitical Instability for Han and Tang China (Turchin 2005a [124], courtesy of the author)

Aggregate ("Sufficient Unit") Equation-Based Modeling

This approach aggregates to the unit size and boundaries at which to define causal variables and interactions and to attempt to explain behavioral dynamics of these units by appropriate equations. This requires the "sufficient unit" condition that the aggregate units of study have the kinds of cohesive mass or entitivity for causal interactions to act on their aggregate characteristics. Time-series will have periods in which this condition is satisfied because of relative "endogeneity" of interactions where there are few external disturbances or exogenous shocks to the unit.

Using this approach, Peter Turchin [123,124] extended a realistic and empirical approach to historical processes—not caricatures of imperial collapse—for basic Malthusian models of population pressure on resources and time-lagged negative feedback effects with internal conflict (see also [125]). He simply uses "standard quantitative methods of natural sciences, such as time-series analysis, regression, and cross-validation. The statistical analysis reveals strong and repeatable patterns in the data on population numbers and the intensity of internal war. And history of science suggests that strong empirical regularities are usually associated with the action of funda-

mental laws" [22], some yet to be discovered for complex systems science. Examples from the Han and Tang China data [123,124] (Chap. 8 in [121]) are shown in Fig. 5 for population and sociopolitical instability (internecine wars), which are related by time-lagged feedback effects.

For the stationary X (population) and Y (internecine wars) variables in these figures, standard time-lagged regression is used to estimate regression constants $a_i = \{a_0, a_1, a_2\}$ where τ is the time lag of 30 years (approximating a human generation), t is time, and ε_t is an error term assumed to be normally distributed [16]:

$$X(t) = a_0 + a_1 X(t-\tau) + a_2 Y(t-\tau) + \varepsilon_t\,, \quad \text{(Model(1))}$$

(and an analogous model for $Y(t)$, reversing the definitions of X and Y). Further:

> One possible objection to the procedure outlined above is that there is some positive autocorrelation between $X(t)$ and $X(t-\tau)$ due to the time-series nature of the data, and it is conceivable that the excellent correlations between the observed $X(t)$ and predicted $X(t)^*$ are entirely due to this "inertial" effect. To eliminate this possibility [the analyzes were redone] with a different dependent variable, $\Delta X(t) = X(t) - X(t-\tau)$. $\Delta X(t)$ is a measure of the rate of change, and by using it we break the autocorrelation arising from the time-series nature of the data. In fact, $\Delta X(t)$ is none other than the realized per capita rate of population change, which is the standard dependent variable in the analyzes of population data.... There can still be some predictive relationship between $\Delta X(t)$ and $X(t)$, so we need to compare two alternative models:
>
> $$\Delta X(t) = a_0 + a_1 X(t-\tau) + \varepsilon_t\,, \quad \text{(Model(2))}$$
>
> ... the *inertial* model (with an analogous (2) for $Y(t)$), and
>
> $$\Delta X(t) = a_0 + a_1 X(t-\tau) + a_2 Y(t-\tau) + \varepsilon_t\,, \quad \text{(Model(3))}$$
>
> ... the *interactive* model (with an analogous (3) for $Y(t)$). The interactive model has an extra parameter, but in a cross-validation setting this does not matter (if the extra independent variable does not have a systematic influence on the dependent variable, then adding it to the model actually *decreases* to the ability of the model to predict out-of-sample data) [124].

The comparisons of the inertial and interactive predictions in Table 2 show consistent effects of dynamical time-lagged interactions between population and sociopolitical instability (civil conflict) that cannot be attributed simply to the inertial dynamics of each of these variables separately. The interactive effects, documented in detailed case studies [126], are those of oscillations: rising population creating resource scarcity, which amplifies inequality, making the value of property rise while that of labor falls, which, if lasting longer than a generation, causes civil unrest and conflict, causing population in turn to decline, with a lag until the cycle recurs as civil conflict ceases, allowing population to rise again (see [38]). Replications of similar findings are obtained by Turchin for the English Tudor cycle (1485–1730) [124], the Medieval English Plantagenet Cycle (1150–1485), French Capetian cycle (1150–1450) and Valois (1450–1660) cycles, Roman republican (350–30 BCE) and principate (30 BCE–285 CE) cycles, Russian Muscovite and Romanov cycle [126], and the Pueblo cycle, where Kohler et al. [65] examine the Turchin model with data from Southwest Colorado between AD 600 and 1300. They find that "it fits well during those periods when this area is

Human Behavior, Dynamics of, Table 2

Comparing Out-of-Sample Predictions of the Inertial and Interactive Models (Turchin 2005a [124], courtesy of the author)

Source of data	Dependent variable	Correlation between predicted and observed			
		1st half → 2nd half		2nd half → 1st half	
		Inertial	Interactive	Inertial	Interactive
England	Population	−0.57	0.94	−0.07	0.44
England	Instability	−0.13	0.80	−0.53	0.89
Han China	Population	0.45	0.57	0.73	0.48
Han China	Instability	0.39	0.87	0.37	0.68
Tang China	Population	0.56	0.80	0.61	0.90
Tang China	Instability	0.57	0.78	0.66	0.92

a more or less closed system. It fits poorly during the time from about AD 1000–1200 when this area is heavily influenced first by the spread of the Chacoan system, and then, by its collapse and the local political reorganization that follows. The model is helpful in isolating periods in which the relationship between violence and population size is not as expected."

Institutions, Network, Economic Models and Experiments: Testing Causality

Studies of historical HB dynamics often lead to different conclusions. In many cases these differences result from the aspects of social process that are focused upon. Contending views may have more general points of consensus when we look at these processes more abstractly. The concepts of structural (k-)cohesion and resistance may help to provide more points of consensus.

There are many views of the formative processes of a market economy based on impersonal exchange and its prior institutional bases. Conceptualized as a network, a market economy requires k-cohesiveness simply to attain $k > 3$ alternatives for buyers and sellers, the minimum "many" players in the market without which the advantages of competitive pricing cannot be obtained. Competition itself, however, is simultaneously a *resistive* as well as a cohesive process, a differentiation of the interests and identities of the competitors. The goods exchanged, for competitive markets, must be alienable, which entails a change of hands in property rights. Players at one time and place may be groups or corporations as property owners party to exchange. At other times they are individuals; or, parties to exchange may be a heterogeneous mix of individuals and groups. For parties to exchange they must have rights: rights to hold property and to alienate property, rights that can be agreed upon by contract, rights and *institutions* that can enforce the contract. Effective "coercion-constraining" institutions that prevent the abuse of others' property rights "influence whether individuals will bring their goods to the market in the first place" (p. 727 in [41]). These give rise to agency, as the capacity for human beings to make choices within a social world and to enforce the rights that those choices impose on the world, whether agency is for the selfsame agent or on behalf of another. The social world is complexly layered at the level of rights, obligations, agents, agency—and institutions as cohesive and resistant social constructions exist for the enforcement of norms. Competing views and agendas are entailed.

These kinds of interlocking components of social worlds do not fall into place quickly, but are built up incrementally over time, just as social networks are built up incrementally and their structural configurations may change slowly even while specific individuals come and go. Market institutions, for example, "co-evolve through a dynamic inter-play between contract-enforcement and coercion constraining institutions" (p. 727 in [40]) along with resistive social movements, movements to create collective goods against the resistance of free-riders, and more episodic events.

The institution-building perspective is one that has received very detailed effort in modeling actual social networks and institutional change in their historical context, abstracting the ways that social players and agents have come to effectively optimize their interactions from their multiple interests and perspectives. One of the most formidable projects of this sort over the last decade, building on the earlier work of North [88], has been to trace social foundations and historical development of institutions in pre-modern Eurasia that facilitate impersonal exchange and lead to paths toward competitive markets, while other developmental paths lead in a variety of other directions [42]. In the words of one reviewer, this work of Avner Greif:

> strips economic transactions down to their elements [and] focuses on the core question: who (or what) were the watchdogs that allowed the merchants to trust one another and to bear with the princes who could confiscate the fruits of all their efforts? And who (or what) were the watchdogs' watchdogs? [The work] repeatedly and carefully relates these questions to economic theory [and] illustrates them with real transactions of medieval merchants. He takes the right approach to economic development, and thereby achieves an original and important new perspective on its causes [2].

In each of Greif's case studies, dynamical game theory is used to test the fit between the observed data and the known historical development of institutions as well as the cultures and behaviors of the players and actors. One of the shortcomings of Greif's work is that the early modern merchants did not face the same problems as those developing markets de novo in early Mesopotamia, India, China, and Mesoamerica. But further evaluation of the replicability of Greif's model is carried on by network economists using experimental real-world simulations that engage participants in the knowledge, payoffs, and choices of the context that is modeled, testing the experimental models against the observed or recorded historical processes and outcomes [62]. To quote:

> North (2005) argues that belief systems and the stock of local knowledge, the internal representations of the human experience, are intimately intertwined with the external institutions that humans build. We investigate this relationship by varying the degree to which property rights are enforced in yesterday's institutions before the opportunities for long-distance trade present themselves with perfectly enforced property rights. Specifically, in the new experiment we report here, three-fourths of the subjects in an economy are drawn from two different treatment histories in *Build8* sessions, one in which property rights in personal goods are perfectly enforced for all of the participants, *though they must rely on trust and repeat interactions to enforce exchange agreements*, and another in which no property rights of any kind are enforced. Hence, in both sets of history-inducing sessions, there is no external enforcement of exchange contracts and, as found [in an earlier experiment], no need for such [62].

The findings of the experimental study are "that a history of un-enforced property rights hinders our subjects' ability to develop the requisite *personal* social arrangements necessary to support specialization and effectively exploit *impersonal* long-distance trade." Thus we might understand through network economic experiments some replicable elements of the origin of impersonal market system. These, like cooperativity, require but go beyond structural cohesion to the social constructions of institutions that secure trust and the benefits of interpersonal trade, i. e., network elements that reinforce the scalability and benefits of structural *k*-cohesion as discussed above.

In Greif's analysis, while the institutional supports for impersonal long-distance trade only developed slowly in medieval Europe (and elsewhere) the full protections of "coercion-constraining" institutions "that prevent the abuse of others' property rights" and "influence whether individuals will bring their goods to the market in the first place" were still not in place even in England after the "Glorious Revolution of 1688," which did not secure such rights beyond "the landed, commercial, and financial elite" (p. 786 [41]). Rarely is linear progression of rights entailed in the ups and downs of the precursor elements of fully competitive markets that vary from one country to another. In England, after 1688, for example, although

> parliament gained supremacy, it was not in the business of protecting property rights per se. Its policy reflected the interests of those who controlled it.... The subsequent history is thus marked by gross abuses of property rights.... Yet, a state controlled by its landed, commercial, and financially elite and later empowered by the Industrial Revolution was a boon for the extensions of markets. The evolution of the modern markets reached its zenith.... Europeans shared a common heritage of individualism, self-governance, a broad distribution of coercive powers, and man-made laws. Reversing their institutional developments and enabling market extensions was relatively easy (pp. 775–776 in [41]).

Eurocentrism is not intended in the use of this example, as this project entailed equally detailed historical and modeling analyzes of China and the Muslim world.

Greif's analysis of land-based institutions and exchange example provides a contrastive comparison against Erikson and Bearman's 2006 [30]) network study of English maritime trade between 1600 and 1831. Here an entirely different account is given of the emergence of the competitive market system. The shared elements are the *k*-cohesive extensions of trade routes, extensive by sea as by land, and the institutional development of English rights in property, commercial exchange, protection, and agency. Here, however, the resistive element is paramount, and the "new economy" arises through malfeasance of the sea captains of the English East India Company. Their work is carried on preemptively, out of self-interest, exploiting the opportunity of delay. Instead of bringing English goods to the orient and returning with oriental goods in one single return cycle, in order to stay beyond the time when the ships could return by the monsoon winds, they traded from port to port on their own behalf with their own goods and retained the profits. Over time, the density of this network became so great that the sheer volume of overlapping circular routes, crisscrossing the net of visited ports, pushed the *k*-connectivity of the market exchanges beyond 0, 1, or 2 for different subregions often up to 7–8, a veritable revolution in creating new market opportunities and competitive market pricing through shear volume of malfeasance behavior: malfeasance because this was all conducted against the policy of the home company, which was powerless to prevent it.

Future Directions

The topics covered here, of cohesion and resistance as measurable social forces in human behavior, and the multiple ways that these two social forces dynamically interact—and what enhances or limits scale-up and scale-down of both cohesion and conflict or resistance – leaves open many researchable questions. Lim, Metzler and Bar-Yam's (2007) study supports a more advanced even if

partial view of the dynamics of cohesion and resistance, group separations, and segregative conflicts along insufficiently demarcated boundaries. Other parts of the human cohesion/resistance dynamics covered in this review show some of the other ways in which cohesion and resistance interact. Human capacities for structural cohesion, for example, support cultural differentiation of groups. Transition thresholds characterize evolutionary bouts of scale-up in group size through central authority, oscillating against resistance from egalitarian preferences for autonomy. With scale-up in size, expansions of political units encounter boundaries of cultural and ethnic differentiation where resistance scales up as oppositional cohesion in positive feedback cycles, creating further expansion of polities that began only as resistive groups. These support growth of population sizes, which lead in turn to scarcity relative to resources within regions. With generational time lags there develop both greater differential inequality and conflictual resistances to inequality. Large polities develop institutional and economic frameworks that can provide benefits to internal differentiation, while the enhanced potential for cohesiveness and economic growth can find ways, as in the biotechnology industry illustration, to utilize the recruitment of diversity to create innovation [93] while stabilizing the costs of cohesive integration.

The problems of modern states and institutions may be seen to devolve on how to minimize the costs of the conflicts that are generated by the oscillations between oppositional cohesion and integrative cohesion. For HB dynamics more generally, solid causal analysis using the most advanced techniques is only possible with current and future data collected systematically on historically documented entities compared over different time scales, up to millennial time series. These data can be analyzed with processual models, network analytic models, institutional, cultural, and evolutionary game-theoretic and economic analyzes. Many of the algorithms needed at this level of complexity have developed in computer science, e. g., by Pearl [5,95] and, by including the crucial element of agency in a new econometrics framework [140,141,142] economics—the otherwise dismal science—can be investigated by causal modeling algorithms. In the modeling of causality that is relevant here to HB dynamics, the analytical power recently gained in econometric models may be neatly illustrated by a comparison of statistical results and conclusions reached by fractal economics, survivorship analysis of successful mutual fund managers, bootstrap models of the same problem, and market simulations of intelligent agents that place orders to trade at random. The first case involves the discovery of fractal pricing in cotton markets [75] and the Dow-Jones [76], contradicting the standard assumption in economics hypothesized by Bachelier [6] that Brownian movement (Gaussian price deviations) is descriptive of market price dynamics. If volatility is predictable in markets, but not price and direction, the implication is that value might not be useful as a concept in economics [75] (consider market collapse when no trader wants to trade in an uncertain market (pp. 3–4 in [58]). Parallel evidence from experimental studies rejected the reference-independent framing of judgments for the "value" of expected utility theory as originally framed by Bernoulli [11], and questioning the assumption that utilities are stable [61]. Similarly, survivorship analysis of successful mutual fund managers showed no evidence that the top ranked funds were any better than random as they lacked measurable persistence [20]. Finely tuned bootstrap estimation models that are oriented toward testing causal models in econometrics, however, showed that while income fund managers did no better than random, growth fund managers showed persistence in their ability to pick stocks (Kosowski et al. [67]). And finally, a baseline market simulation model of for intelligent agents that place orders to trade at random, with only one free parameter, accounted for 96% of the best buying and selling prices (the spread), and 76% of the variance of the price diffusion rate [32], which "demonstrates the existence of simple laws relating prices to order flows, and in a broader context, because it suggests that there are circumstances where the strategic behavior of agents may be dominated by other considerations." "One of the virtues of this model is that it provides a benchmark to separate properties that are driven by the statistical mechanics of the market institution from those that are driven by the strategic behavior of agents. It suggests that institutions strongly shape our behavior, so that some of the properties of markets may depend more on the structure of institutions than on the rationality of individuals." These examples are all indicators of complex dynamics.

The challenge of HB dynamics is to assemble better data related to aspects of the problems modeled, including those of competing hypotheses: more complete data, data better grounded in diverse historical circumstances, and more contextual detail. A second challenge is to have better statistical estimators, identification and correction for sources of bias, attention to nonindependence, careful modeling of richly grounded historical data, attention to causal modeling, and multiple-level models. These efforts are facilitated by sharing of data, collaborative analysis of potential biases in data, sharing of documentation of software and source code, verification of source code,

and replication of results. Extensive effort has gone into the archaeological and geocoded data related to the evaluation of the model in Fig. 2 of transition probabilities. Ten years of effort went in locating and coding the data in Fig. 3, for example. Good data on population numbers at all levels, such as aggregated in Figs. 3 and 4, is extremely important for modeling, hard to come by, and demands careful analysis for bias detection, bias correction and data reconstruction. The data on internecine warfare in Fig. 5 were patiently transcribed episode by episode in two compendia of scholarly work over millennia. These are but a few of many thousands of databases, many of which have not been made sharable or are not conserved or not well documented. Also needed are data analytic routines able to make accurate estimates using probabilistic bootstrap methods with small samples and for different kinds of data, e. g., continuous or discrete, nominal or ordinal. A great deal of documented open source code is now available.

Causal modeling is the core of dynamical analysis, and future directions will include modeling of the types illustrated here, and many more, but with integrated datasets for different foci and levels of analysis. A host of intersecting and mutually enriching integratable time-coded longitudinal datasets—like Turchin's data for the 50-region data in Fig. 1 [121], or the sufficient size data for Fig. 2 [124]—are needed from comparable local contexts and processes up to the global, e. g., google-earth-like sharable data structures, equipped with analytic routines for time-series causal modeling and testing. There are many separate projects on shared issues, but overall integration is needed. Geographic Information Systems (GIS), for example, need to be reintegrated around open source code (e. g., GRASS, written in open source R) that includes temporal and network modeling. Autocorrelation and other techniques and models for dealing with nonindependence of cases will figure heavily in causal modeling.

Among new network analytic methods that are becoming standard in many disciplines are the censuses of different types of cycles ("motifs") in large networks that make up cohesive k-components, and accessible software to compute k-connectivity. Analysis of these sorts of data allow testing of where and how the internal micro and middle-range structures come from in k-components [44,81,132]. Which structures come from preferences and which from the marginals or limits on how data were collected or spatially distributed? This kind of work is now being done in biology but also in anthropological network studies, where new software packages have been developed that are specifically designed to deal with certain problems, such as kinship networks or the kinds of generative kinship computations that people actually use in their social cognition [70,99].

Entropy maximization "open system" models conditioned on biased random processes will increasingly become integrated with HB dynamics as we come to understand how to connect them to foundational problems in the human sciences, some of which are discussed in the ENTROPY entry in these volumes. Simple entropy models, for example, are currently being used to fill in missing data from what is known from an archaeological site [31]. Tsallis entropy [119], in contrast, would provide a one-parameter modification for least energy maximization channeled through networks, with diffusion gradients that have multiplicative effects. Generative network models for cohesive cycles, as studied to date [137] show consistent distributions of numbers of links that are all in the Tsallis entropy family, so there are promising avenues in this research area.

As seen in examples here, more integration of theory and data is needed, from macrohistorical models where large-unit aggregation relates to sufficient statistics for causality, through cascades of spatio-temporal processes to the micro level of interactions between individuals [94]. Kirman [63], for example, argues that "The emergence and evolution of the networks that govern the interaction in the economy plays a crucial role and it may well be the case that the standard notion of equilibrium is irrelevant in such as context." Multilevel network analyzes will be aggregated structurally in new ways for which new modeling techniques are needed to analyze the composition of units and of processes [131].

The construction of theory and hypotheses in this article illustrate only a few potential causal links among major topics on evolutionary and historical dynamics, institutional and economic models, game theory and social networks, organized around a few core dimensions of ethological importance (structural cohesion and resistance). The point of these illustrations has been that there are truly major forces in history—k-cohesion and cohesive resistance for example—but these have very different properties than forces in physics, or the dynamics of chemistry and biology (although one LANL biologist, asked to pinpoint the major threats to survival, responded with "human behavior" [57]), and they require very different measurements and theory. But there is no closure on the topics of HB dynamics: rather, there is an abundance of theory and results in social, historical, and simulation modeling that lend themselves to the evaluation of causality. Taking causality and dynamics seriously rather than dismissively leads to very different theoretical and analytical perspec-

tives. Longitudinal data on human social, historical, and network phenomena are sufficient to support high-level theoretical and integrative research that can further benefit from the most advanced of methods in the complexity sciences. But there is a pressing work to be done in analytic methods and in constructing valid and reliable datasets and variables on appropriate and comparable units and processes under analysis.

Acknowledgment

As a Santa Fe Institute external faculty member, the author greatly appreciates SFI support and the benefits of interactions with scores of SFI researchers, visitors, and staff. Special thanks to Jeroen Bruggeman for sharing his prepublication book manuscript and the cross-fertilization of ideas presented in Sect. "Cooperation, Connectivity-*k* and "Critical Mass" in Collective Action", to Henry Wright for detailed commentary and suggestions, Eric Smith for comments on the coalescence of the argument, Peter Turchin for sharing historical data, and to Bruggeman, Turchin, Art Griffin, Charles Stanish, and Emily Erikson for suggestions on the discussions of their findings, to Lilyan Brudner-White for many editorial suggestions, and to D. Eric Smith and many other colleagues at SFI for discussions.

Bibliography

1. Adamic L, Lukose RM, Huberman BA (2002) Local search in unstructured networks. In: Bornholdt S, Schuster HG (eds) Handbook of graphs and networks: From the genome to the internet. Wiley-VCH, Berlin
2. Akerlof GA (2007) Book commentary on Avner Greif (2007a) Institutions and the Path to the Modern Economy: Lessons from Medieval Trade. Amazon.com editorial review
3. Algaze G (2005) The Sumerian takeoff. Struct Dyn 1(1):2
4. Arthur B (1994) Increasing returns and path dependence in the economy. University of Michigan Press, Ann Arbor
5. Avin C, Shpitser I, Pearl J, Identifiability of path-specific effects. UCLA Cognitive Systems Laboratory, Technical Report (R-321), June 2005. In: Proceedings of International Joint Conference on Artificial Intelligence, Edinburgh, August 2005
6. Bachelier L (1900) Théorie de la Spéculation, 1900, Annales de l'Ecole normale superiure (trans. Random Character of Stock Market Prices)
7. Bandy MS (2004) Fissioning, scalar stress, and social evolution in early village societies. Am Anthropol 106(2):322–333
8. Barabási A-L (2002) Linked: The New Science of networks. Perseus Publishing, Cambridge. http://www.nd.edu/~networks/linked
9. Barfield TJ (1989) The perilous frontier: Nomadic empires and China. Blackwell, Oxford
10. Bearman PS, Burns L (1998) Adolescents, health and school: Early findings from the national longitudinal study of adolescent health. NASSP Bull 82:601–23
11. Bernoulli D (1738) Exposition of a new theory on the Measurement of Risk (trans. 1954). Econometrica 22:123–136
12. Boehm C (1983) Montenegrin social organization and values: Political ethnography of a refuge area tribal adaption. AMS Press, New York
13. Boehm C (1993) Egalitarian behavior and reverse dominance hierarchy. Curr Anthropol 34:227–254
14. Boehm C (1999) Hierarchy in the forest: Egalitarian society and the evolution of democratic politics. Harvard University Press, Cambridge
15. Bowles S, Gintis H (2008) Cooperation. In: Blume LE, Durlauf SN (eds) New palgrave encyclopedia of economics, 2nd edn. Palgrave Macmillan, Basingstoke, http://www.dictionaryofeconomics.com/article?id=pde2008_C00059789=GintisAopicid=result_number=2
16. Box GEP, Jenkins G (1976) Time series analysis: Forecasting and control. Holden-Day, San Francisco
17. Brudner LA, White DR (1997) Class, property and structural endogamy: Visualizing networked histories. Theory Soc 26:161–208
18. Bruggeman J (2008) Social networks: An introduction. Routledge, New York
19. Burt RS (2001) Bandwidth and echo: Trust, information, and gossip in social networks, In: Casella A, Rauch JE (eds) Networks and markets. Russell Sage Foundation, New York
20. Carhart MM, Carpenter JN, Lynch AW, Musto DK (2002) Mutual fund survivorship. Rev Financ Stud 15:1439–1463
21. Carvalho R, Iori G (2007) Socioeconomic networks with long-range interactions. ECCS 2007, European Conference on Complex Systems 2007—Dresden. Paper #334. http://arxiv.org/abs/0706.0024
22. Cliodynamics (harvested 2006), Web site of Peter Turchin, http://www.eeb.uconn.edu/people/turchin/Clio.htm
23. Colson E (1953) The Makah indians. Manchester University Press, Manchester, University of Minnesota Press, Minneapolis
24. Coser L (1967) Continuities in the study of social conflict. The Free Press, New York
25. Couzin I (2007) Collective minds. Nature 445:715
26. Couzin I et al (2005) Effective leadership and decision making in animal groups on the move. Nature 433:513–516
27. Earle TK (1977) A reappraisal of redistribution: Complex Hawaiian chiefdoms. In: Earle TK, Ericson J (eds) Exchange Systems in Prehistory. Academic Press, New York
28. Eells E, Skyrms B (eds) (1994) Probability and conditionals, Belief revision and rational decision. In: Cambridge studies in probability, induction, and decision theory. Cambridge University Press, Cambridge
29. Encyclopedia Brittannica (2007) Ibn-Khaldun. Harvested December 8. http://www.britannica.com/eb/article-225307/Ibn-Khaldun
30. Erikson E, Bearman PS (2006) Routes into networks: The structure of English East Indian trade, 1600–1831. Am J Sociol 112(1):195–230
31. Evans T, Knappett C, Rivers R (2008) Using statistical physics to understand relational space: A case study from mediterranean prehistory. In: Lane D, Pumain D, van der Leeuw S, West G (eds) Complexity perspectives on innovation and social change. Springer Methodos series, in press, http://www3.imperial.ac.uk/pls/portallive/docs/1/7292491.PDF

32. Farmer JD, Patelli P, Zovko II (2005) The predictive power of zero intelligence in financial markets. PNASUSA 102(11):2254–2259, http://www.santafe.edu/~jdf/papers/zero.pdf
33. Flack JC, Krakauer DC (2006) Encoding power in communication networks. Am Nat 168:97–102
34. Flack JC, de Waal FBM, Krakauer DC (2005) Social structure, robustness, and policing cost in a cognitively sophisticated species. Am Nat 165:E126–E139
35. Flack JC et al (2006) Policing stabilizes construction of social niches in primates. Nature 439:426–429
36. Garcia J, van den Bergh J (2007) Evolution of parochialism requires group selection. ECCS 2007, European Conference on Complex Systems 2006, Oxford. Paper #120
37. Gluckman M (1963) Gossip and scandal. Curr Anthropol 4:307–315
38. Goldstone JA (1991) Revolution and rebellion in the early modern world. University of California Press, Berkeley
39. Granovetter M (1973) The strength of weak ties. Am J Sociol 78:1360–1380
40. Greif A (1997) On the social foundations and historical development of institutions that facilitate impersonal exchange: From the community responsibility system to individual legal responsibility in pre-modern europe, Working Papers 97016, Stanford University, Department of Economics. http://www-econ.stanford.edu/faculty/workp/swp97016.pdf
41. Greif A (2007) Commitment, coercion and markets: The nature and dynamics of institutions supporting exchange. In: Ménard C, Shirley MM (eds) Handbook of new institutional economics. Springer, New York
42. Greif A (2007) Institutions and the path to the modern economy: Lessons from medieval trade. In: Political economy of institutions and decisions. Stanford University Press, Palo Alto
43. Griffin AF, Stanish C (2007) An agent-based model of prehistoric settlement patterns and political consolidation in the lake Titicaca basin of Peru and Bolivia. Struct Dyn 1(1):2
44. Hamberger K, Houseman M, Daillant I, White DR, Barry L (2004) Matrimonial ring structures. Mathématiques et sciences humaines 43(168):83–121
45. Hammond M (2002) Review of [15]. Canad J Sociol Online Nov–Dec 2002
46. Hardin G (1968) The tragedy of the commons. Science 162(3859):1243–1248
47. Henrich J, Gil-White F (2001) The Evolution of prestige: freely conferred status as a mechanism for enhancing the benefits of cultural transmission. Evol Hum Behav 22:1–32
48. Henrich J, Henrich N (2006) Culture, evolution and the puzzle of human cooperation. Cogn Syst Res 7:221–245
49. Henrich J, McElreath R (2003) The evolution of cultural evolution. Evol Anthropol 12:123–135
50. Henrich J et al (2005) Economic man in cross-cultural perspective: Ethnography and experiments from 15 small-scale societies. Behav Brain Sci 28:795–855
51. Henrich J et al (2006) Costly punishment across human societies. Science 312:1767–1770.http://www.sciencemag.org/cgi/content/full/312/5781/1767
52. Henrich N, Henrich J (2007) Why humans cooperate. Oxford University Press, New York
53. Henry A (1906, 1918) The education of Henry Adams. Houghton Mifflin, Boston
54. Hess NH, Hagen EH (2006) Psychological adaptations for assessing gossip veracity. Hum Nat 17:337–354
55. Hruschka D, Henrich J (2006) Friendship, cliquishness, and the emergence of cooperation. J Theor Biol 239:1–15
56. Jackson MO, Wolinsky A (1996) A strategic model of social and economic networks. J Econ Theory 71:44–74
57. Johnson N, LANL scientist, Personal communication
58. Jorion PJ (2007) Reasons vs. causes: Emergence as experienced by the human agent. Struct Dyn 2(1):1–6
59. Jost J (2005) Formal aspects of the emergence of institutions. Struct Dyn 1(2):2
60. Khaldun I (1958) [c.1379] The Muqaddimah: An introduction to history, (trans. from the Arabic by Rosenthal F). Pantheon books, New York
61. Kahneman D, Tversky A (2003) Maps of bounded rationality: A perspective on intuitive judgment and choice. In: Frangsmyr T (ed) Les Prix Nobel 2002. Almquist & Wiksell International, Stockholm
62. Kimbrough E, Smith VL, Wilson B (2006) Historical property rights, sociality, and the emergence of impersonal exchange in long-distance trade, Am Econ Rev (forthcoming) Interdisciplinary Center for Economic Science working paper at http://www.ices-gmu.net/article.php/433.html
63. Kirman A (1999) Aggregate activity and economic organisation. Revue européenne des sciences sociales 37(113):189–230
64. Kleiman MAR (2008) When brute force fails: Strategy for crime control (in progress)
65. Kohler TA, Cole S, Ciupe S (2008) Population and warfare: A test of the Turchin model in Pueblo societies. In: Shennan S (ed) Pattern and process in cultural evolution. University of California Press, Santa Fe, in press Institute Working Paper 06-06-018): http://www.santafe.edu/research/publications/wpabstract/200606018
66. Korotayev A, Malkov A, Khaltourina D (2006) Introduction to social macrodynamics: Secular cycles and millennial trends. URSS, Moscow
67. Kosowski R, Timmermann A, White H, Wermers R (2006) Can mutual fund "stars" really pick stocks? New evidence from a bootstrap analysis. J Financ 61(6):2551–2595. http://weber.ucsd.edu/~mbacci/white/pub_files/hwcv-101.pdf
68. Krakauer D (2007) Self-description of work in progress, SFI web site, Harvested 12/11
69. Kremer M (1993) Population growth and technological change: One million BC to (1990). Q J Econ 108(3):681–716
70. Leaf M (2007) Empirical formalism. Structure and Dynamics 2(1):2
71. Lévi-Strauss C (1949, 1969) The elementary structures of kinship. Beacon, Boston
72. Lim M, Metzler R, Bar-Yam Y (2007) Global pattern formation and ethnic/cultural violence. Science 317:1540–1544
73. Machalek R (1992) The evolution of macrosociety: Why are large societies rare? In: Freese L (ed) Advances in human ecology, vol 1. JAI Press, Greenwich, pp 33–64
74. Malinvaud E (1985) Econometric methodology at the Cowles commission: Rise and maturity. Abstracted from the Cowles fiftieth anniversary volume. http://cowles.econ.yale.edu/archive/reprints/50th-malinvaud.htm
75. Mandelbrot B (1963) The variation of certain speculative prices. J Bus 36:394–419

76. Mandelbrot B, Hunter RL (2004) The (mis)behavior of markets: A fractal view of risk, ruin, and reward. Basic Books, Hudson
77. Marwell G, Oliver P (1993) The critical mass in collective action: A micro-social theory. Cambridge University Press, Cambridge
78. Maryanski AR (1987) African ape social structure: Is there strength in weak ties? Soc Netw 9:191–215
79. McNeill WH (1963) The Rise of the west. New American Library, New York
80. Menger K (1927) Zur allgemeinen Kurventheorie. Fundamenta Mathematicae 10:96–115
81. Milo R et al (2002) Network motifs: Simple building blocks of complex networks. Science 298:824–827
82. Modelski G, Thompson WR (1996) Leading sectors and world power: The coevolution of global politics and economics. University of South Carolina, Columbia
83. Moody J, White DR (2003) Structural cohesion and embeddedness. Am Sociol Rev 69:103–127
84. Newman MEJ (2004) Detecting community structure in networks. Eur Phys J B (38):321–330
85. Newman MEJ (2006) Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103:8577–8582
86. Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74:036104
87. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69:026113
88. North DC (1990) Institutions, institutional change and economic performance. Cambridge University Press, Cambridge
89. Nowak MA (2006) Five rules for the evolution of cooperation. Science 314:1560–1563
90. Oliver P, Marwell G, Teixeira R (1985) A theory of the critical mass, I. interdependence, group heterogeneity, and the production of collective goods. Am J Sociol 91:522–556
91. Oliver PE, Marwell G (1988) The paradox of group size in collective action: A theory of the critical mass. II. Am Sociol Rev 53(February):1–8
92. Olson M (1965) The logic of collective action. Harvard University Press, Cambridge
93. Page SE (2007) The difference: How the power of diversity creates better groups, firms, schools, and societies. Princeton University Press, Princeton
94. Pande, Rohini and Christopher Udry. Institutions and development: A view from below. Economic Growth Center, Discussion Paper No. 928, Yale University. http://www.econ.yal.edu/growth_pdf/cdp928.pdf
95. Pearl J (2000) Causality: Models, reasoning and inference. Press MIT, Cambridge
96. Peebles C, Kus S (1977) Some archaeological correlates of ranked societies. Am Antiq 42:421–448
97. Powell WW, White DR, Koput KW, Owen-Smith J (2005) Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. Am J Sociol 110(4):1132–1205
98. Pryor LP, Graburn NH (1980) The myth of reciprocity. In: Gergen KJ, Greenberg MS, Willis RH (eds) Social exchange: Advances in theory and research. Plenum Press, New York
99. Read D, Behrens C (1990) KAES: An expert system for the algebraic analysis of kinship terminologies. J Quant Anthropol 2:353–393
100. Reichardt J, White DR (2007) Role models for complex networks. Eur Phys J B 60:217–224
101. Richardson LF (1960) Statistics of deadly quarrels. Boxwood Press, Pacific Grove. http://shakti.trincoll.edu/~pbrown/armsrace.html
102. Rosenbaum P, Rubin D (1983) The central role of the propensity score in observational studies for causal effects, Biometrika 70:41–55
103. Rubin D (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educal Psychol 66:688–701
104. Ruths DA, Nakhleh L, Iyengar SM, Reddy SAG, Ram PT (2006) Hypothesis generation in signaling networks. J Comput Biol 13(9):1546–1557
105. Sampson RJ, Raudenbush SW, Earls F (1997) Neighborhoods and violent crime: A multilevel study of collective efficacy. Science 277:918–924
106. Sampson RJ, Morenoff JD, Earls F (1999) Beyond social capital: Spatial dynamics of collective efficacy for children. Am Sociol Rev 64:633–660
107. Scharf L (1991) Statistical signal processing: Detection, estimation, and time series analysis. Addison-Wesley, New York
108. Schelling T (1960) The strategy of conflict. http://en.wikipedia.org/wiki/Schelling_point
109. Schelling T (1978) Micromotives and macrobehavior. W.W. Norton, New York
110. Service ER (1962) Primitive social organization: An evolutionary perspective. Random House, New York
111. Simmel G (1890) The sociology of Georg Simmel (trans: 1950). Free Press, New York
112. Skyrms B (1988) Probability and causation. J Economet 39(1-2):53–68
113. Skyrms B (1996) Evolution of the social contract. Cambridge University Press, Cambridge
114. Sommerfeld RD et al (2007) Gossip as an alternative for direct observation in games of indirect reciprocity. Proceedings of the National Academy of Sciences 104:17435–17440
115. Spencer C (2004) Primary state formation in Mesoamerica. Annu Rev Anthropol 33:173–199
116. Taagepera R (1997) Expansion and contraction patterns of large polities: Context for Russia? Int Stud Q 41:482–504
117. Tambayong L (2007) Dynamics of network formation processes in the co-author model. J Artif Soc Soc Simul 10(3):2
118. Trivers RL (1971) The evolution of reciprocal altruism. Q Rev Biol 46:35–57
119. Tsallis C (2008) Entropy. Encyclopedia of complexity and systems science. Springer, Berlin
120. Turchin P (2003) Complex population dynamics. Princeton University Press, Princeton
121. Turchin P (2003) Historical dynamics. Princeton University Press, Princeton
122. Turchin P (2004) Dynamic maps of evolution of the state system and metaethnic frontiers in Europe during the two millennia. http://www.eeb.uconn.edu/people/turchin/Clio.htm
123. Turchin P (2005) A primer on statistical analysis of dynamical systems in historical social sciences (with a particular emphasis on secular cycles). Struct Dyn 1(1):4
124. Turchin P (2005) Dynamical feedbacks between population growth and sociopolitical instability in agrarian states. Struct Dyn 1(1):3

125. Turchin P, Korotayev A (2003) Population dynamics and internal warfare: a reconsideration. Soc Evolut Hist 5(2):112–147
126. Turchin P, Nefedov S (2008) Secular cycles. Princeton University Press, Princeton
127. Turner JH (1921) The frontier in american history. H. Holt and Company, New York
128. Turner JH (2002) Face to face: toward a sociological theory of interpersonal behavior. Stanford University Press, Palo Alto
129. von Foerster H, Mora PM, Amiot LW (1960) Doomsday: Friday, 13 November, A.D., (2026). Science 132:1291–1295
130. West J (1945) Plainsville. Columbia University Press, New York
131. White DR (1974) Mathematical anthropology. In: Honigmann JJ (ed) Handbook of social and cultural anthropology. Rand McNally, Chicago, pp 369–446. http://eclectic.ss.uci.edu/~drwhite/pub/MathAnth74-1.pdf
132. White DR (2004) Ring cohesion theory in marriage and social networks. Mathématiques et sciences humaines 43(168): 5–28
133. White DR (2008) Innovation in the context of networks, hierarchies, and cohesion. In: Lane D, Pumain D, van der Leeuw S, West G (eds) Complexity perspectives on innovation and social change. Springer Methodos series, in press. http://eclectic.ss.uci.edu/~drwhite/pub/ch5revMay-20.pdf
134. White DR, Harary F (2001) The cohesiveness of blocks in social networks: Node connectivity and conditional density. Sociolog Methodol 31:305–359
135. White DR, Houseman M (2002) The navigability of strong ties: small worlds, tie strength and network topology. Complexity 8(1):72–81. http://eclectic.ss.uci.edu/~drwhite/Complexity/K&C-a.pdf
136. White DR, Johansen U (2005) Network analysis and ethnographic problems: Process models of a Turkish nomad clan. Lexington, Oxford. A summary is at http://en.wikipedia.org/wiki/Network_Analysis_and_Ethnographic_Problems
137. White DR, Kejžar N, Tsallis C, Farmer D, White S (2006) A generative model for feedback networks. Phys Rev E 11:016119
138. White DR, Owen-Smith J, Moody J, Powell WW (2004) Comput Math Organ Theory 10(1):95–117
139. White DR, Tambayong L, Kejžar N (2008) Oscillatory dynamics of city-size distributions in world historical systems. In: Modelski G, Devezas T, Thompson W (eds) Globalization as evolutionary process: Modeling, simulating, and forecasting global change. Routledge, London, pp 190–225
140. White H (2009) Settable systems: An extension of Pearl's causal model with optimization, equilibrium, and learning. Submitted to J Machine Learn Res. http://eclectic.ss.uci.edu/~drwhite/center/ppt_pdf/A_Comparison_of_Pearl_s_Causal_Models_and_Settable_Systems.pdf
141. White H, Chalak K (2006). A unified framework for defining and identifying causal effects. Department UCSD of Economics Discussion Paper http://www.economics.ucr.edu/seminars/spring06/econometrics/HalWhite_Lect_6-2-06.pdf
142. White H, Chalak K (2007) Independence and conditional independence in causal systems. http://eclectic.ss.uci.edu/~drwhite/center/ppt_pdf/Independence_and_Conditional_Independence_in_Causal_Systems.pdf
143. Wright HT (2000) Instability, collapse, and the transformation to state organization. System shocks-system resilience. Workshop Paper, Abisko, Sweden, May 20–23:2000
144. Wright HT (2006) Atlas of Chiefdoms and Early States. Structure and Dynamics: 1(4)1
145. Zuckerman EW (2003) On Networks and Markets by Rauch and Casella (eds). J Econ Lit 41:545–565

Human–Environment Interactions, Complex Systems Approaches for Dynamic Sustainable Development

Lenore Lauri Newman
Royal Roads University, Victoria, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
Placing Dynamic Sustainable Development in a Historical Context
Feedback Loops and Reactive, Proactive, and Adaptive Management
Path Dependence and Lock-in
Future Directions
Bibliography

Glossary

Emergence The appearance of complex structures or behaviors within a complex adaptive system that is unpredictable from the starting conditions or materials.

Lock-in The inability to change non-optimal behavior due to the engrained social, financial, or technical cost of changing the behavior.

Negative feedback Outcomes of an action within a complex adaptive system that tend to decrease the magnitude of the originating action.

Panarchy A model of behavior of complex adaptive systems proposing that such systems progress through cycles of growth, collapse, and renewal.

Path dependence The dependence on system behavior upon prior system behavior.

Positive feedback Outcomes of an action within a complex adaptive system that tend to increase the magnitude and impact of the originating action.

Resilience The ability of a complex adaptive system to maintain its form in the face of disruption.

Transformability The ability of a complex adaptive system to transform into a different state better suited to existing conditions.

Definition of the Subject

Dynamic sustainable development is a process-based environmental management theory based upon the recognition of human systems as complex adaptive systems that are, in turn, subsets of ecosystems that are also complex adaptive systems. Human systems and ecosystems exist in an evolving relationship, and as human systems grow in size adaptive management techniques must be employed to ensure that the containing ecosystems are not overwhelmed by human resource demands or by the waste products of human activity. Dynamic sustainable development is the latest step in an ongoing process of evolution within the larger sustainable development discourse away from goal-based, utopian models of sustainability to a process-based, iterative, adaptive approach.

Introduction

The concept of dynamic sustainable development arose as a means of reconciling the desire to create less environmentally damaging human enterprises with the complexity inherent in the field of resource management. Dynamic sustainable development suggests that there are two central pillars to the successful adaptive management of human–environment interactions; the need for resilience and the need for transformability. Resilience, defined as the magnitude of disturbance that can be absorbed before a structural change occurs within a system [30], has been identified as a desirable goal of adaptive management. However while a system should be able to maintain a degree of stability in the face of surrounding change, sometimes a system change is required for long term sustainability; the system must possess transformability, the ability to totally alter itself if needed [71]. This combination of stability and change represents a break with the goal-based models of steady-state human–environmental interactions more prevalent in earlier environmental discourses.

Some of the general principles of dynamic sustainable development are set out in [46]; these general principles draw on the continuing evolution of the general sustainable development concept. As such dynamic sustainable development is part of a larger trend to incorporate complex adaptive systems theory into environmental management. The first general principle is that *dynamic sustainable development is a process, not a goal*. This principle reflects a broader shift within the sustainable development discourse. In recent literature, sustainable development is described as a continuous process of change [36], and is described as a process that must be treated as an evolution of ideas [53]. This shift is well founded; complexity theory has shown us that change is the norm in social and ecological systems. Any static or "climax community" will eventually fall prey to the "inevitable accident" [38]. Change is the process that allows evolving systems to thrive. Rammel notes, " … there cannot be any best state, or stable equilibrium, or optimal path of development" [51].

One of the key figures of the shift from goal to process within the environmental management and planning field is C.S. Holling, whose contributions include panarchy theory and many of the formative writings on adaptive management. The panarchy theory was developed to help clarify the role of change within adaptive systems. The term was chosen as a reference to the Greek God of nature, Pan, and as a reference to the combination of change and persistence found within complex systems. Panarchy theory focuses on the interplay between different levels of activity within a system and outlines an evolutionary cycle of growth, collapse, and rebirth that systems undergo. Different time scales within systems are also highlighted: fast scales of events experiment and test, while slower levels stabilize and conserve. The interplay of these scales and levels create dynamic structures. Holling argues that seeing sustainable development in the light of complex adaptive systems resolves the critique that sustainable development is an oxymoron:

> Sustainability is the capacity to create, test, and maintain adaptive capability. Development is the process of creating, testing, and maintaining opportunity. The phrase that combines the two, 'sustainable development' thus refers to the goal of fostering adaptive capabilities and creating opportunities. It is, therefore, not an oxymoron but a term that describes a logical partnership [29].

Holling's vision of sustainable development is an iterative one; it can emerge organically from unsustainable behavior in manageable steps. Norms cannot be imposed in advance [56], but will emerge as part of an adaptation process. Instead of being a final objective, sustainable development has to be understood as a continuous process of change [36], and a fruitful approach to this process is to treat it as an evolution [53]. Treating sustainable development as a process creates the need for an indefinite program of monitoring and adjustment. Every successful adaptation is only a temporary "solution" to changing selective conditions [53]. This principle of sustainable development means that it is always a moving target [60].

The second principle of dynamic sustainable development is that *dynamic sustainable development must cope with the inherent unpredictability of the systems it addresses*.

Uncertainty is inherent to complex adaptive systems; in particular the behavior of ecosystems and social systems are notoriously difficult to predict. These systems manifest emergent properties; human society in particular is highly heterogeneous, far from equilibrium, and complexity is our society's defining feature. Human society is also highly non-ergodic. Ergodicity is the tendency of a system to move towards equilibrium, maximizing entropy. Human societies do not settle down into stable patterns for long. They constantly innovate, grow and change, posing a challenge for those trying to adjust our interactions with the biosphere.

Though we might wish to design a perfect and stable society, history suggests such experiments end in failure. Sustainable development models must therefore be robust enough to mitigate the ecological effects of a non-ergotic society. Positive feedback loops, which will be discussed in detail later, allow accidents of history to get magnified in outcome [70]. This leads to many results of small actions being unintended [35] and unpredictable from the initial conditions. Our predictions of the future are at best temporary guides, leaving us in the need to iteratively monitor feedback loops and continually adjust our models and our actions accordingly. This inherent unpredictability represents a "strong uncertainty"; not only are we unable to predict the consequences of events we are unable to determine which events are the ones that will lead to future change [65]. To use the language of complex adaptive systems theory, when human societies interact with natural systems they show sensitive dependence on initial conditions. The changes that can arise in a complex system involving society and the environment can be particularly perplexing, as they can involve changes in human knowledge and awareness, changes in technology, and also changes in public perception [24]. These three aspects can all be present at the same time and interact.

The third guiding principle of dynamic sustainability is that *innovation processes greatly affect sustainable development initiatives*. Early models of environmental societies were almost always static, steady state societies that changed very little over time. Goal orientated utopian models of environmental action range from Skinner's "Walden Two" [65] to the steady state economics proposed by Herman Daly [14] to the models presented in the landmark document "Limits to Growth" [43]. These models, however, ignore innovation as a fundamental component of human society. We use technical ingenuity to create new technology, but social ingenuity reforms old institutions and social arrangements into new ones [31]. Managing our interaction with the Earth's ecosystems would be much easier without the complicating factor of innovation constantly changing the nature of this interaction, but this process is constantly occurring on a number of scales. At the smaller scale we see incremental innovations, which are small refinements that occur relatively continuously. At a larger scale, there are radical innovations representing large shifts in technologies. These are not predictable, and may happen at any time. There are systematic innovations that can create entirely new fields [48]. Such sudden shifts can provide new technologies to protect ecosystems, can shift our resource use from one resource base to another, and can also increase our impact on ecosystems in new and unexpected ways.

The process of innovation should be rather familiar to those who work with complex systems; Complexity has been called the "science of surprise" precisely because unpredictable behaviors and structures are emergent within complex systems. Emergence is the creation of new behavior and properties that cannot be understood from an investigation of the system's parts [20]; emergent behavior underlies the need for transformability within systems as the need for sudden changes to preserve the integrity of ecological systems and social systems can arise at any time. Lewin calls emergence "the central feature of the new science of complexity", p. 175 in [40]. It is certainly one of the more surprising features of complex adaptive systems. Emergence can be observed in the biosphere by even a casual investigator. As ecologist Chaia Heller says, a seed doesn't grow; it becomes something new p. 106 in [27]. An emergence process drives the agricultural processes that allow us to harness energy and survive. C.D Broad, a philosopher, coined the term emergence in the 1920s, to describe those things that appear even though there is no hint of them at lower levels, p. 28 in [9].

The problem is that while it is easy to understand effects that can intuited from the behavior of the system, predicting emergent behavior is a different matter. This unpredictability deeply affects the quest for sustainable development. The existence of emergent behavior requires our definitions of sustainability to be dynamic emergent behavior can provide sudden unexpected innovation, and it can also spawn bursts of ecological destruction. For this reason, the existence of emergent structure raises the need for safety factors and resiliency.

The prominence of emergence and innovation do not sit well with many members of the environmental movement. There is a historical uneasiness with innovation that informs the development of sustainable development dialogs. Though technology can be seen as an "adaptive answer" to problems [53], there is a fundamental disconnect between the world of the information society and the groundings of sustainable development due to differing

values held by the actors involved [36]. This uneasiness is made more acute by the inherent uncertainty in the process of innovation [7].

Placing Dynamic Sustainable Development in a Historical Context

The application of complex systems theory to questions of resource management and environmental preservation was first attempted relatively early in the twentieth century, long before the concept of sustainable development itself was developed. What was then called general systems theory addressed environmental issues under the assumption that human organizations could not be successfully analyzed by only considering their parts. In effect, systems theory reverse engineered the study of complexity from the study of specific complex adaptive systems associated with human society.

General systems theory arose from the necessity of managing complex technologies in difficult situations. General systems theory sought common principles in the structure and the operation of systems of all shapes and sizes [63], a quest that continues in the field of adaptive resource management. In 1950, Ludwig von Bertalanffy, one of the founders of general systems theory, brought together several of his ideas to argue that systems have properties independent of discipline. Von Bertalanffy felt ever-increasing specialization was not useful to the study of such systems [69]. The study of cybernetics had already proven useful during the Second World War, and there was a growing movement among ecologists to consider systems in a transdisciplinary way.

In 1954, Anotol Rapoport, Ludwig von Bertalanffy, Kenneth Boulding, Ralph Gerrard, and James Miller founded the Society for General Systems Research [25]. Von Bertalanffy had already introduced the concept of general systems theory in 1937. He understood that social systems are far from equilibrium and coined the term "open system".

General systems theory is devoted to understanding function of systems, including organizations [69]. It was meant to be a very practical study, and has been applied widely to the understanding of business organizations. Boulding, who coined the term "spaceship earth", divided study into special systems theory, which focused on modeling, and general systems theory, which was a broader philosophical inquiry into the overall dynamics of social systems [25].

Several of the elements of complex adaptive systems studied by physicists have also been studied by systems theorists. Erich Jantsch felt dissipative structures were of particular importance to the understanding of human systems [34], as they are far from equilibrium but resilient. He felt dissipative structures grow to limits governed by internal, not external constraints, a belief that reflects a general interest among systems theorists in autopoiesis, the process by which dissipative structures continually renew and regulate themselves in order to maintain themselves. The Society for General Systems Research contributed to the *Limits to Growth* Report, which is discussed below as a foundational document of sustainable development.

An unlikely contributor to the incorporation of complex systems theory within environmental discourse emerged from the mathematical study of discontinuous phase transitions. Mathematician Rene Thom developed a method for examining such discontinuous changes in mathematical and physical systems, a method he came to call catastrophe theory. In Thom's theory, a catastrophe is the discontinuous transition between stable but divergent paths [73]. Thom based the theory on rules of topology, and the assumption of underlying structural stability. This assumption relies on the presence of unseen order; Thom assumed there was a hidden order beneath the discontinuity that he could not quantify. He also understood the uniqueness of each region of discontinuity. Even though it went against the grain of traditional model of mathematics, Thom insisted catastrophes were local, and that no global models of such systems exist, p. 7 in [66].

Catastrophe theory provided a new way of thinking about change, and gave a metaphor for the abrupt changes seen in natural and human systems, unlike the continuous, linear models used at the time. The theory thus represented a rather significant mathematical breakthrough, and enjoyed a spectacular rise in general popularity, perhaps due to its very striking name. Catastrophe theory was quickly trumpeted as describing the actual workings of natural and social systems. Controversy began to surround the theory as it was used as a blanket metaphor for a wide range of systems. The application of catastrophe theory became a craze that spread far beyond the mathematical sciences. Unfortunately, the application of catastrophe theory to social and environmental issues could not actually live up to the hype surrounding it. This was partly due to the difficulty of the mathematics underlying the theory, and partly due to the fact that many of the social systems it was applied to did not lead at the time to quantifiable models [20]. The bubble of popularity Thom and his theory enjoyed collapsed. However the brief flirtation with this theory planted a seed of interest in complex adaptive systems within practitioners of ecological management.

The evolution of the sustainable development concept itself really began in earnest with the publication of *Limits to Growth* by the Club of Rome, group which was founded by British scientist Alex King and Italian industrialist Aurelio Peccei. The first meeting of roughly forty international thinkers occurred on the seventh of August, 1967 in Rome, Italy. The *Limits to Growth* report features computer models of resource use created by Donella and Dennis Meadows. Their most advanced model, World Three, shows three possible futures: overshoot and collapse, overshoot and oscillation, and sigmoid growth, p. 123 in [44]. The futures predicted in World Three were highly pessimistic of future progress. Overshoot refers to a society that is using resources and creating wastes at a rate faster than can be supplied or absorbed by the biosphere. This crude beginning founded an ongoing movement to build complex models of human–ecological interactions.

Limits to Growth focused on what the authors called the "world problematique". This group of linked problems included poverty amid plenty, environmental destruction, urban sprawl, and economic problems p. 10 in [43]. In the opinion of Meadows and her co-authors, we are far past the point at which overshoot has occurred. Her models commonly predict overshoot and collapse. In short, not only must our societies not grow any more, they must contract.

Response was immediate and was highly polarized. It was pointed out that World Three underestimates the ability of technology to postpone catastrophe [12]. The models rely on tables composed of extrapolated data generated through an iteration process, causing runaway errors, p. 32 in [12]. The *Limits to Growth* report was heavily critiqued for ignoring innovation [10]. In many ways what was missing from the models were non-linear effects such as feedback and emergence.

When the Club of Rome published *Limits to Growth* in 1970, it sparked intense international debate over its basic premise, which challenges the basic tenets of traditional economics. The prevalent economic vision sees the economy as an isolated system: a circular flow of exchange value between firms and households. The economy is the system of interest and natural systems are simply sources of resources and sinks for wastes. Nature may be finite, but these natural sources and sinks can be infinitely substituted for by human capital, without limiting overall growth in any important way.

The problem with exponential growth in a finite system is simple enough to describe; the economy strains nature with respect to sources and waste sinks [32]. In effect, the finite expanse of the natural world has a carrying capacity that can be expressed in different ways; economic carrying capacity is the maximum global economic welfare derivable from the sustainable throughput flows of the ecosphere [72].

Following *Limits to Growth* it became popular to speak of a finite and firm upper limit to the scope of human activity within the biosphere, and to imagine static societies that would exist at an equilibrium point with the biosphere that respected clear limits. The first Earth Day was held in 1970, attracting 20 million people to peaceful demonstrations across the US. The deteriorating state of the Great Lakes prompted a number of clean-up efforts. Many governments began to allocate resources to establish ministries of the environment. Of particular importance was the founding of Greenpeace in 1971. Greenpeace mustered grass-roots support for campaigns against clear-cut ecological abuse. Spurred by public interest in the environment, several major commissions and reports established by various levels of government began to highlight the environmental damage caused by economic growth. The Stockholm Conference of 1972 focused attention on the growth of industrial pollution. The *Global 2000* report to President Carter of the United States warned that "if present trends continue, the world in 2000 will be more crowded, more polluted, less stable ecologically … despite greater material output the world's people will be poorer in many ways than they are today", p. 1 in [5].

The conservation movement helped to fuel the debate over humanity's appropriation of natural resources and the physical limitations scarcity posed to economic growth. However, the early movement was often fixated with scaling back human activity to remove the complexity from our interactions with the environment, rather than addressing issues through a complex adaptive lens. In 1983, the United Nations assembled the World Commission on Environment and Development, and charged them with the task of gathering opinion on the state of the environment and its potential effect on development. In 1987, the group produced *Our Common Future*, which introduced the concept of sustainable development into common use. Known as the Brundtland Report in reference to Chairperson Gro Harlem Brundtland, this document drew on scientists, government and communities to pinpoint the resource issues most pressing to global development. The Brundtland report set out as an objective the achievement of a sustainable society by the year 2000, and quickly became a milestone in the discussion of Sustainable Development. *Our Common Future* also called for a conference to be called in five years time to set out an action plan for the achievement of these goals.

The lasting effects of the Brundtland report were mixed, and reflect the tensions between those still work-

ing within a goal-based mindset and those arguing for process-based management. *Our Common Future* created consensus between many different stakeholders, and formalized many of the assumptions now taken for granted in the discussion of sustainability. Two of these are particularly important; the Brundtland report made clear the link between economic and ecological health, condemning the treatment of the natural sphere and the human sphere as separate entities. The report defines development as what we do in an attempt to improve our lot within these conjoined spheres. *Our Common Future* begins by giving a very vague definition of "sustainable development", saying simply that it should "meet the needs of the present without compromising the ability of future generations to meet their own needs" [8]. The Brundtland report puts forward many initiatives designed to reach a state of sustainability, which was still framed in a goal-oriented way. They call for action on all levels, including monitoring at the global level. They ask that "the ability to anticipate and prevent environmental damage requires that the ecological dimensions of policy be considered at the same time as the economic, trade, energy, agricultural and other dimensions", p. 10 in [8]. In short, they wish for less reactive repair of environmental problems and more proactive monitoring to prevent environmental damage.

The lasting importance of the Brundtland report is due to several factors. The commission's use of roundtables and participatory democracy allowed the widest possible input of voices to be heard, leading to a document that spoke to a variety of stakeholders. The Brundtland report made it clear that sustainability is a needed goal, but left the definitions vague enough to allow further discussion. The report also called for a follow up meeting to be held five years later at which a roadmap to sustainability would be set out.

The years following the publication of *Our Common Future* saw a peak in environmental concern, cumulating in the publication of *Agenda 21*, the follow-up report to *Our Common Future* [58]. Agenda 21 is a broad action plan adopted at the 1992 Rio Conference to promote environmentally sound and sustainable development in all countries of the world. Agenda 21 was signed on 13 June 1992 by over one hundred heads of state representing ninety eight percent of the world's population. Agenda 21 is not legally binding; it is a flexible guide for achieving a sustainable world.

Agenda 21 is divided into six themes composed of subareas with specified action plans. The first theme, quality of life, addresses areas such as limiting poverty, changing consumption patterns, controlling population growth, and ensuring the availability of adequate health care. The second theme, efficient use of resources, focuses on land use planning, water conservation and management, energy resources, food production, forest management, and the protection of biodiversity. The third theme, protection of the global commons, discusses management of the atmosphere and the oceans. The fourth theme, management of human settlements, considers urban issues and the provision of adequate shelter. The fifth theme, waste management, focuses on the classification and disposal of chemical, solid, and radioactive wastes. The final theme, sustainable economic growth, discusses trade, development, and technology transfer.

Agenda 21 has been criticized for not including strong positions on transport, energy issues, and tourism. The action plan has also been criticized for being too focused on increasing trade. Agenda 21 stresses that removing distortions in international trade is essential and environmental concerns should not restrain trade, positions that are critiqued by antiglobalisation groups.

The success of Agenda 21 has been mixed. The action plan has been successful at linking environment and poverty, and many local working groups have been formed that include the multiple stakeholders such as youth, indigenous people, scientists, and farmers called for in the plan. The Commission on Sustainable Development, which is charged with monitoring the progress of Agenda 21's implementation, has reported positive developments in the areas of controlling population growth, increasing food production, and improving local environments. However they also report an increase in inequality, increasing water scarcity, and extensive loss of agricultural land. Implementation in the European countries has been more successful than in other regions.

The action plan's mixed success can be attributed in part to a lack of commitment of funding for the initiatives in the plan. Every action in Agenda 21 included a projected cost; but no source of funding was secured at the time of signing. The Commission on Sustainable Development continues to monitor the implementation of *Agenda 21*, and follow-up meetings known as the Earth Summits and nicknamed Rio +5 and Rio +10, were held in 1997 in New York and in 2002 in Johannesburg. In 2002 the United Nations General Assembly called the progress of *Agenda 21's* implementation extremely disappointing, and at Johannesburg a plan was developed to speed the implementation of *Agenda 21*.

Agenda 21 and *Our Common Future* set out as a goal the achievement of a sustainable world by the year 2000. This goal was not achieved, though much good did come out of these documents. This failed legacy will also shape any attempt to form a dynamically sustainable society. We

can learn a lot by considering this large attempt at building a sustainable world. The importance of securely funding initiatives is made clear, for example. Also, the importance of entrenching a culture of sustainability at the individual level is made clear as well. *Our Common Future* and *Agenda 21* were largely off the radar of popular culture, and so did not get the attention they needed to move forward.

Though *Our Common Future* propelled the concept of sustainable development into the public eye, the vague definition given within the report of sustainable development led to a long and ongoing debate over what sustainable development actually means. Since the introduction of sustainable development into common parlance, numerous variations have emerged, such as sustainability, sustainable growth, sustainable economic growth, and sustainable environmental or ecological development. Although disagreement exists among different sectors and communities about the usefulness of the concept of sustainable development, it is recognized internationally and it does avoid most of the traditional left-right polarization and discourse about growth versus no-growth, by bringing together the terms sustainable and development in a constructive ambiguity that has stimulated greater dialog between sectors. Despite its ambiguity, it has succeeded in uniting widely divergent theoretical and ideological perspectives into a single conceptual framework [19]. More fundamentally, it has brought a wide diversity of industrialists, environmentalists, public policy practitioners and politicians to round tables, in their attempts to define, deal with and actualize the concept. In order to provide some contextual appreciation, a brief examination of some of the earlier definitions of sustainable development follows. Human societies everywhere will place a different emphasis on the former and on the latter, according to their ecological, social and economic conditions.

Though the term was popularized by the Brundtland commission, it predates Our Common Future. In 1980, the World Conservation Strategy, IUCN, UNEP, WWF, and others argued that integration of conservation and development is particularly important, because unless patterns of development that also conserve living resources are widely adopted, it will become impossible to meet the needs of today without foreclosing the achievement of tomorrow's needs. Meadows et al. [44] defined a sustainable society as one that had in place informational, social and institutional mechanisms to keep in check the positive feedback loops that cause exponential population and capital growth. In other words, means that birth rates roughly equal death rates, and investment rates roughly equal depreciation rates, unless and until technical changes and social decisions justify a considered and controlled change in the levels of population or capital. In order to be socially sustainable the combination of population, capital and technology in the society would have to be configured so that the material living standard is adequate and secure for everyone. In order to be physically sustainable the society's material and energy throughput has to meet economist Herman Daly's [15] three conditions: its rates of use of renewable resources do not exceed their rates of regeneration; its rates of use of nonrenewable resources do not exceed the rate at which sustainable renewable substitutes are developed, and its rates of pollution emission do not exceed the assimilative capacity of the environment. Some scholars differentiate between sustainability and sustainable development, while others use them interchangeably. Sustainability derives from the Latin root *sus-tinere*, which literally means to under-hold or hold up from underneath. Sustainability describes a characteristic of relations (states or processes) that can be maintained for a very long time or indefinitely [37].

The outcome of this dialog produced a generic, rather than a specific definition for sustainable development. It is a process of reconciliation of three imperatives. These are the ecological imperative to live within global biophysical carrying capacity and to maintain biodiversity, the social imperative to ensure the development of democratic systems of governance that can effectively propagate and sustain the values that people wish to live by, and the economic imperative to ensure that basic needs are met worldwide [13,57]. This definition, however, remains general enough to allow for sustainable development to be interpreted differently in specific socio-geographic situations and to be sufficiently responsive in the face of unpredictable change and uncertainty. It also responds to the dynamic interplay between the imperatives, reflecting the complexity of modern human society.

The move from the above "three pillar model" to a process oriented version of the same model is largely a product of the last ten years. Various organizations have fed this shift by studying ecological systems with complex adaptive techniques. Perhaps the most well-know of these organizations is the Resilience Alliance, a multidisciplinary group based in the United States that a focuses on the exploration of the dynamics of social-ecological systems. Their mission is to assemble knowledge that embraces resilience, adaptability and transformability and that influences sustainable development policy and practice. The Resilience Alliance publishes the journal "Ecology and Society", which was started by C.S Holling under the name "Conservation Ecology". The Society for Human Ecology has a similar focus as the Resilience Alliance,

and holds one of the central annual conferences in the field. The journal is one of the central ones in the areas of sustainable development as process, along with the journals *Sustainable Development* and *International Journal of Sustainable Development*. On the modeling side of the field, the Santa Fe Institute is a definite leader. The Santa Fe Institute was founded as an independent multidisciplinary research institute to study complex adaptive systems theory and how it can be applied to address environmental, technological, biological, economic, and political challenges. These centers of knowledge creation continue to fuel the robust development of dynamic sustainable development.

Feedback Loops and Reactive, Proactive, and Adaptive Management

The next few sections look at the complex adaptive features that shape dynamic sustainable development; the first of these is the existence of feedback process within social-environmental systems. Complex adaptive systems are far more than a collection of elements; they are bound together by the flow of energy, matter and information. This flow is often two-way, forming feedback loops within the complex system. Feedback loops are a central feature of all complex adaptive systems, and an understanding of feedback loops is critical to the understanding of human complex adaptive systems. Achieving sustainability is fundamentally a question of observing and responding to feedback. Feedback allows control within complex adaptive systems, and also allows growth. Feedback loops form the nervous systems of complex adaptive systems, allowing the flow of information between elements and between the system and the environment.

Feedback is a process in which a change in an element alters other elements, which in turn affect the original element [35]. Feedback is an iterative process, and is a fundamental part of what makes a system both complex and adaptive.

There are two main types of feedback within complex systems: positive and negative feedback. Negative feedback loops moderate a system, but this process does not always lead to stability. Too much negative feedback can cause a system to become stagnant and unable to adapt to suddenly changing situations. A system composed only of negative feedbacks will become out of step with its surrounding environment and perish.

In Waldrop's opinion, it is the mixture of positive and negative feedback that creates complex systems [70]. Negative feedback provides stability [35] that holds systems away from run-away growth and collapse. Negative feedback is often associated with the concept of equilibrium. As an example, a thermostat uses negative feedback to maintain homeostatic equilibrium, keeping the surrounding area at a stable temperature. Most systems, thus, have built in processes that can bypass negative feedback in emergency situations. Writer and urban planner Jane Jacobs provides as an example of such an "override" mechanism: our reflexive ability to stop breathing upon being immersed suddenly in water, p. 115 in [33].

Positive feedback is the driving force behind sensitive dependence on initial conditions, which is also called the "butterfly effect" after an example given by Lorentz in a 1972 talk entitled, "Predictability: Does the Flap of a Butterfly's Wings in Brazil Set Off a Tornado in Texas?" [41]. The concept of positive feedback can be difficult to grasp, as we tend to believe in a "conservation of complexity" [11] in which simple causes lead to simple effects, and complex causes lead to complex effects. This is a restatement of Newton's second law: *for every action there is an equal and opposite reaction*. This correlation does not hold in complex adaptive systems. Positive feedback can reinforce a small event again and again until it becomes a system-wide phenomenon.

Positive feedback allows growth, and fuels expansion and diversity [33,35]. As Waldrop comments, positive feedback loops allow accidents of history to get magnified in outcome [70]. If negative feedback loops hold a system stable, positive feedback loops allow systems to explore their environment and follow new paths of development. As they magnify random small variations, positive feedback loops add an element of surprise to the system's behavior. This leads to many results of small actions being unintended and unpredictable from the initial conditions. A forest free from disturbances will evolve towards an equilibrium state that was once called "climax" ecology. Little changes in such an ecosystem until a positive feedback process occurs; such as a tiny blaze that builds upon itself to raze the forest and make room for new growth. By its nature, positive feedback has the ability to lead complex systems into precarious territory. Positive feedback loops are either dampened by negative feedback loops or they crash [33]. The ability to dampen positive feedback loops is necessary to the survival of complex adaptive systems.

Both positive and negative feedback take a non-zero amount of time to propagate through a complex system. If this time is overly large the feedback loop may be ineffective at regulating the system or might have an unintended effect on the system. Delayed feedback is quite common within natural ecosystems' responses to human stresses, as the two complex systems function on such different time scales. Delayed feedback poses a large challenge to the

achievement of dynamic sustainability. As it might take us some time for the biosphere to respond to our actions, we can suffer creeping environmental decay. Damage occurs slowly and by the time we notice, it has already become severe.

The existence of positive feedback and sensitive dependence on initial conditions within society has profound consequences for sustainable development. As we can never trust our predictions of the future entirely, there can be no perfect model of a sustainable society that will hold up for all time. Instead, we must monitor feedback loops carefully and continually adjust our models and our actions accordingly. In the larger context of dynamic sustainable development, negative feedback processes provide resilience and positive feedback processes provide transformability.

Within the environmental movement there is a great desire to move away from the reactive approach of dealing with environmental problems after they develop to a proactive approach in which feedback processes are anticipated before they begin.

One proposed method used to mitigate uncertainty is to use what is called a precautionary principle. The concept of a precautionary principle is often credited to the German principle of Vorsorgeprinzip, or foresight planning, which began to receive attention in the 1970s [45]. The concept has evolved over time; what began as a "measure" shifted to an "approach" and finally to a "principle" [1].

In the United Nation's Rio Declaration, the use of a precautionary principle is urged. Principle 15 of the Rio Declaration states that where there are threats of serious or irreversible damage, lack of full scientific certainty shouldn't be used as a reason for postponing cost-effective measures to prevent environmental degradation [45]. A stronger definition known as the "Wingspread definition" emerged from a conference of environmental thinkers in 1998. The Wingspread definition of the precautionary principle states that when an activity raises threats of harm to human health or the environment, precautionary measures should be taken even if some cause-and-effect relationships are not fully established scientifically [49].

Intuitively, the precautionary principle is straightforward, but it is notoriously difficult to apply [1,49]. The general idea of the precautionary principle is to avoid serious and irreversible damage [64]. As Raffensperger states, the precautionary principle can be used to prevent, not just redress, harm [52]. What is simple to describe, however, is not necessarily simple to put into use. There must be some evidence a hazard exists if the precautionary principle isn't to lead to ruling out any action [61]. If the precautionary principle is not to stifle progress, it should be coherent, utilize known information and theories, have explanatory power, and possess simplicity [55]. Low complexity solutions should be preferred to high complexity solutions if the precautionary principle is to avoid simply creating further problems [64]. And given that not all results of an action within a complex system can be predicted before they occur, no purely proactive management regime will be sufficient.

A solution to this problem that is gaining in popularity is a combination of proactive and reactive management known as adaptive management. Adaptive co-management involves the combination of proactive and reactive management techniques through an iterative process of feedback monitoring. Adaptive management keeps careful track of uncertainties within the system, partially through local control of day to day ecosystem management, as it is assumed that undesirable effects will appear locally first. Adaptive co-management seeks to include past, present and future stakeholders. Adaptive co-management is an emergent and self-organizing.

Adaptive co-management attempts to correct for runaway feedback through a place-based approach. Local voices are incorporated into Adaptive co-management of ecosystems in which both local actors and larger level effects are considered is critical to the creation of resilience and transformability. Use of local ecological knowledge builds resilience [6]; local solutions tailored to local conditions are necessary for healthy ecosystem interactions, and if many variables are to be monitored changes in those variables will first be observed "on the ground" locally. Without local involvement, management tends to shift towards exploitation; as local resources are depleted new resources are substituted in other locations.

The need for local knowledge and observation is not, however, an argument in favor of moving to locally isolated small-scale enterprise. As even local action can have global consequences, resilience emerges from both cross-scale and within scale interactions [47]. Transformability in the face of external changes requires an outward focus to the larger scale. Connectivity allows resilience and movement, but the existence of local network structure buffers against cascades of disaster from the larger world [2]. That said, for adaptive co-management to work it must be collaborative; without a shared sense of purpose stakeholders at different scales are likely to have very different interests [26]. In some cases global stakeholders are likely to value short-term financial gain over local ecosystem integrity, but in other cases local actors might value short-term employment prospects over larger ecological

needs. Actors on all scales must work in concert, not at cross-purposes.

Complex systems are filled with uncertainty, and no amount of precaution will eliminate all risks. Management must, thus, build system resilience. There are several factors that influence the resilience of a system. These include the system's latitude, which is the maximum amount of stress that a system can absorb without changing to a new state, the system's ability to resist change, and its precariousness, which refers to how fragile it is [71]. The more resilient an ecosystem or society is, the better it will be at responding and adapting to unpredictable changes.

There are several ways system resilience can be increased. Primarily, we can increase resilience by ensuring that as we undertake a course of action we leave room for alternatives. Preventative measures should allow for more flexibility in the future [22]. Especially in cases of irreversibility, options should be kept open [4]. We can increase resilience through increasing the buffering capacity of the system, managing for processes at multiple scales, and nurturing sources of renewal [23].

Path Dependence and Lock-in

A critical feature of complex adaptive systems that has direct application on human–environment interactions is the occurrence of branching points during the evolution of complex systems. These points occur when the system can evolve in multiple and exclusive ways. Such systems are thus path dependent; the state of the system depends on its earlier history. Path dependence is important to the study of complex adaptive systems such as human societies and ecosystems because it limits the ability to act reactively when attempting to intervene in a complex system. The field of restoration ecology was one of the first to encounter this aspect of complex systems. To restore a damaged ecosystem, one cannot simply plant the species that were found in the undamaged ecosystem; one must plant the pioneer species for that ecosystem and then allow the ecosystem to follow a succession cycle to the desired state. To do this, one must know the ecosystem's *trajectory*; the history of its succession over time. Determining this can be very difficult as the pioneer species are often not found in the desired mature ecosystem. In short, the status of a system does not only depend on the state of a set of variables, but how they were reached [35]. Path dependence makes the job of understanding complex adaptive systems much more difficult.

Path dependence is central to the development of ecosystems and human development. Historical records are, thus, critically important. Knowledge gathered over time locally can provide historical knowledge that is not accessible to more abstract methods of data collection [21]. Local knowledge allows for adaptive management of local resources [6] precisely because it captures the on-going path dependence of complex systems.

Local complex problems require general scientific knowledge, but local knowledge plays a complimentary role. That said, the needed historical data are not often easily found; most local knowledge is passed on orally and is based upon generations of observation; this knowledge as younger generations move to urban areas and older generations die. Practitioners have uncovered a wealth of unlikely environmental information locally, including records of planting times, bird logs, temperature records, and species lists.

Local knowledge can add another layer of uncertainty even as it provides crucial data. There is an assumption that locals have intimate knowledge of the local environment, but not all locals have the same knowledge [16]. Experts can usually be identified by asking locals who holds the greatest knowledge locally. Using local knowledge can be frustrating for all involved [21], and thus must be used appropriately. Local knowledge can determine changes in local conditions, outline ecological trajectories, and identify path dependent processes.

Path dependence can be described as "reactive sequences" in which each event precipitated by previous self-reinforcing sequences [42]. In short, history matters, and a random event can ensure a suboptimal technology or process becomes the norm. Rammel points out sometimes rather mediocre solutions dominate a natural selection process in the short term (2003); systems, particularly ones of great complexity, can prove very inflexible. Arthur calls this re-enforcement of certain historical paths nonergodic behavior; in his words, path dependence matters [3] deeply. Though this problem could be minimized by careful use of precautionary principles at the beginning of the development of a technological path, in many cases negatives of new technologies tend to appear after implementation [39]. The ozone-depleting properties of CFCs are a good example of a technology that had to be completely replaced due to an unforeseen danger that became apparent after lock-in had occurred. Identifying a problem and choosing a solution is difficult, but it is often, as Homer-Dixon notes, implementation that is the true problem, p. 23 in [31].

From a practical standpoint, the largest impact of path dependence is that it actively constrains the actions one can take to implement sustainable development programs. Even if one is carefully monitoring feedback, and a signal arrives within a system suggesting change is needed,

the needed change might not be possible. The problem of technologies' ideas and behavior patterns becoming entrenched is known in the literature as lock-in. Lock in occurs because within complex societies innovations do not stand alone, they co-develop as entire networks of supporting technologies, much as keystone species co-develop in nature. If we suddenly need to change one of these keystone technologies we can find there is significant social and economic reluctance to do so. Diamond calls this the "sunk cost effect" as we are reluctant to abandon what we have even if it doesn't work, p. 432 in [18]. Lock-in arises naturally out of two properties of complex systems; path dependence and increasing returns. One cannot simply change them without setting off cascading changes throughout the system.

Recognition of this problem is not new. In his work on the need to shift away from fossil fuels, Unruh calls lock-in a technological "cul de sac" that at its worst cumulates in an embedded techno-institutional complex [67]. There is a need to tackle path dependence and lock in [53], but as Unruh notes "the question of how to overcome lock-in has been little explored" [69]. Lock-in has been related to the issue of diversity discussed as condition three: the encouragement of niche markets is a possible way of breaking lock in [68]. However research has shown that in extremely locked in systems, research into options falls to near zero [54]. As Arthur notes, there is a minimum cost for a transition and changing by fiat sometimes necessary [3].

Future Directions

The role of complex systems theory in sustainable development continues to develop. At the qualitative level the greatest challenge is an educational one; factors such as feedback loops, sensitive dependence on initial conditions and path dependence are not widely understood by decision makers. One of the largest policy challenges will be understanding and controlling lock-in within complex markets. It is likely that the leaders in the field will continue to be found in the areas of ecosystem management led by members of groups such as the Resilience Alliance. Given the difficulty of the underlying mathematics and physics, knowledge of complex systems theory is likely to remain outside of the general knowledge employed by sustainable development practitioners aside from qualitative models and metaphorical applications of general concepts such as self organization and feedback. Fostering this understanding of general concepts should be encouraged.

If sustainable development initiatives are to create the kind of changes needed to ensure the survival of our societies and the ecosystems they reside in, the move to adaptive and process-based environmental management will need to proceed at a faster pace. It has been rightly remarked that the transition to sustainable development is alarmingly slow [52], partly due to the magnitude of the changes required.

The greatest factor in the continuing development of quantitative complex models in this field is the ongoing increase in available computational power. New computational techniques such as evolutionary computing [17] will also increase the range of techniques for modeling the sustainability of complex systems. These computational techniques could add needed rigor to popular sustainable development tools such as the Natural Step and the Ecological Footprint, both of which are widely used and are based on iterative processes that would benefit from complex systems approaches. It is likely the Santa Fee institute will continue to play a leading role in approaches to complex modeling. Areas that will benefit from these technological advances could include the management of electricity grids that include a large number of small intermittent renewable energy sources, traffic management to reduce emissions, and climate modeling.

The lasting influence of sustainable development lies in its ability to evolve as a concept. Given the advances in the understanding of complex adaptive systems and the application of this understanding to ecological and social systems, there are likely many fruitful avenues that combine sustainable development and complex systems theory. In the long run, the most successful sustainable development initiatives will likely look rather a lot like ecosystems; diverse, complex, and evolving.

Bibliography

Primary Literature

1. Adams M (2002) The precautionary principle and the rhetoric behind it. J Risk Res 5:301–316
2. Anderson E (2006) Urban landscapes and sustainable cities. Ecol Soc 11(1):34. http://www.ecologyandsociety.org/vol11/iss1/art34/
3. Arthur B (1994) Increasing returns and path dependence in the economy. University of Michigan Press, Ann Arbor
4. Arrow K, Fisher A (1974) Environmental preservation, uncertainty, and irreversibility. Q J Econ 88:312–319
5. Barney G (1980) The global 2000 report to the president. US Government Printing Office, Washington
6. Berkes F, Colding J, Folkes C (2000) Rediscovery of traditional ecological knowledge as adaptive management. Ecol Appl 10:1251–1262
7. Buenstorf G (2000) Self-organization and sustainability: Energetics of evolution and implications for ecological economics. Ecol Econ 33:119–134

8. Brundtland G (1987) Our common future: World commission on environment and development. Oxford University Press, New York
9. Capra F (1996) The web of life. Anchor Books, New York
10. Cleaveland C, Ruth M (1997) When, where, and by how much do biophysical limits constrain the economic process? Ecol Econ 22:203–223
11. Cohen J, Steward I (1994) The collapse of chaos: Discovering simplicity in a complex world. Penguin Books, New York
12. Cole H (1973) Introduction. In: Cole H, Freeman C, Jahoda M, Pavittetal K (eds) Models of doom: A critique of the Limits to Growth. Universe Books, New York
13. Dale A (2001) At the edge: sustainable development in the 21st Century. UBC Press, Vancouver
14. Daly H (1991) Steady state economics, 2nd edn. Island Press, Washington
15. Daly H, Cobb J (1994) For the common good, 2nd edn. Beacon Press, Boston
16. Davis A, Wagner J (2003) Who knows? On the importance of identifying experts when researching local ecological knowledge. Human Ecol 31:463–489
17. De Jong K (2006) Evolutionary computation: a unified approach. MIT Press, Cambridge
18. Diamond J (2005) Collapse: How societies choose to fail or succeed. Viking Press, New York
19. Estes R (1993) Toward sustainable development: from theory to praxis. Soc Dev Issues 15:1–29
20. Eve R, Horsefall S, Lee M (1997) Chaos, complexity and sociology. Sage Publications, London
21. Fischer F (2000) Citizens, experts, and the environment: the politics of local knowledge. Duke University Press, Durham
22. Gollier C, Jullien B, Treich N (2000) Scientific progress and irreversability: An economic interpretation of the precautionary principle. J Pub Econ 75:229–253
23. Gunderson L (2000) Ecological resilience: In theory and application. Annu Rev Ecol Syst 31:425–439
24. Hadfield L (1999) A Co-evolutionary model of change in environmental management. Futures 31:577–592
25. Hammond D (2003) The science of systems: Exploring the social implications of general systems theory. University Press of Colorado, Boulder
26. Hein L, Van Kloppen K, De Groot R, Van Lerland E (2005) Spatial scales, stakeholders, and the valuation of Ecosystem services. Ecol Econ 57:209–228
27. Heller C (1999) Ecology of everyday life. Black Rose Books, Montreal
28. Holling C (1976) Resilience and stability of ecosystems. In: Jantsch E, Waddington C (eds) Evolution and consciousness: human systems in transition. Addison Wesley, Reading, pp 73–92
29. Holling C (2001) Understanding the complexity of economic, ecological, and social systems. Ecosystems 4:390–405
30. Holling C, Gunderson L (2002) Resilience and adaptive cycles. In: Gunderson L, Holling C (eds) Panarchy: Understanding transformations in Human and Natural Systems. Island Press, Washington
31. Homer-Dixon T (2000) The ingenuity gap. Alfred A Knopf, New York
32. Hudson R (2005) Towards sustainable economic practices, flows, and spaces: Or is the necessary impossible and the impossible necessary? Sustain Dev 13:239–252
33. Jacobs J (2001) The nature of economies. Vintage Canada, Toronto
34. Jantsch E (1975) Design for evolution: Self-organization and planning in the life of human systems. George Braziller, New York
35. Jervis R (1997) System effects: Complexity in political and social life. Princeton University Press, Princeton
36. Jokinen P, Malaska P, Kaivo-Oja J (1998) The environment in an information society: A transition stage towards more sustainable development. Futures 306:485–498
37. Judes U (2000) Towards a culture of sustainability. In: Filho W (ed) Communicating sustainability. Peter Land, Frankfurt
38. Kay J, Regier H, Boyle M, Francis G (1999) An ecosystem approach for sustainability: addressing the challenge of complexity. Futures 1999:721–742
39. Konnola T, Unruh G, Carrillo-Hermosilla J (2006) Prospective voluntary agreements for escaping techno-institutional lock-in. Ecol Econ 57:239–252
40. Lewin R (1992) Complexity: life at the edge of chaos. Machmillan Publishing Company, New York
41. Lorentz E (1993) The essence of chaos. University of Washington Press, Seattle
42. Mahoney J (2000) Path dependence in historical sociology. Theor Soc 29:507–548
43. Meadows D et al (1972) The limits to growth: A report for the Club of Rome's project on the predicament of mankind. Universe Books, New York
44. Meadows D et al (1992) Beyond the limits: confronting global collapse, envisioning a sustainable future. McClelland and Stewart, Toronto
45. Morris J (2000) Defining the precautionary principle. In: Morris J (ed) Rethinking risk and the precautionary principle. Butterworth, Oxford, pp 1–14
46. Newman L (2005) Uncertainty, innovation, and dynamic sustainable development. Sustain: Sci Pract Policy 1(2)25–31. http://ejournal.nbii.org/archives/vol1iss2/TOC.html
47. Peterson G (2000) Political ecology and ecological resilience: An integration of human and ecological dynamics. Ecol Econ 35:323–336
48. Pereira P (1994) New technologies: Opportunities and threats. In: Salomon J, Sagasti J, Sachs-Jeantet C (eds) The uncertain quest: science, technology, and development. United Nations University Press, Tokyo, pp 448–462
49. Raffensperger C (2002) Precaution and security: The labyrinthine challenge. Whole Earth 113. http://findarticles.com/p/articles/mi_m0GER/is_2002_Fall/ai_93135763
50. Raffensperger C (2003) Constitutional experiments: protecting the environment and future generations. Conserv Biol 17:1587–1488
51. Rammel C (2003) Sustainable development and innovations: Lessons from the red queen. Int J Sustain Dev 6:395–416
52. Rammel C, Kastenhofer K (2006) Obstacles to and potentials of the societal implementation of sustainable development. Sustain: Sci Pract Policy 1(2)5–13. http://ejournal.nbii.org/archives/vol1iss2/TOC.html
53. Rammel C, Van Den Berg J (2003) Evolutionary policies for sustainable development. Ecol Econ 47:121–133
54. Redding S (2002) Path dependence, endogenous innovation, and growth. Int Econ Rev 43:1215–1248
55. Resnik D (2003) Is the precautionary principle unscientific? Stud Hist Philos Biol Biomed Sci 34:329–344

56. Robinson J (2003) Future subjunctive: Backcasting as social learning. Futures 35:839–856
57. Robinson J, Tinker J (1997) Reconciling ecological, economic and social imperatives: a new conceptual framework. In: Schrecker T (ed) Surviving globalism Social and environmental dimensions. Macmillan, London
58. Robinson N (1993) Agenda 21: Earth's action plan. Oceana Publications, New York
59. Saltelli A, Funtowicz S (2005) The precautionary principle: Implications for risk management strategies. Human Ecol Risk Manag 11:69–83
60. Salwasser H (1993) Sustainability needs more than better science. Ecol App 3:587–589
61. Sandin P, Peterson M, Hansson S, Ruden C, Juthe A (2002) Five charges against the precautionary principle. J Risk Res 5:287–299
62. Skinner B (1976) Waldon two. Person Education Canada, Toronto
63. Skyttner L (2001) General systems theory: Ideas and applications. World Scientific, Singapore
64. Som C, Hilty L, Ruddy T (2004) The precautionary principle in the information society. Human Ecol Risk Assess 10: 787–799
65. Spash C (2002) Greenhouse economics: Values and ethics. Routledge Press, London
66. Thom R (1975) Structural stability and mophogenesis. W. Benjamin, Reading
67. Unruh G (2000) Understanding carbon lock-in. Energy Policy 28:817–830
68. Unruh G (2002) Escaping carbon lock-in. Energy Policy 30:317–325
69. Von Bertalanffy L (1968) General systems theory: Foundations, development, applications. George Braziller Press, New York
70. Waldrop M (1992) Complexity: The emerging science at the edge of order and chaos. Simon and Schuster, New York
71. Walker B, Holling C, Carpenter S, Kinzig A (2004) Resilience, adaptability, and transformability in Social-Ecological systems. Ecol Soc 9(2):5. http://www.ecologyandsociety.org/vol9/iss2/art5
72. Wetzel K, Wetzel J (1995) Sizing the earth: Recognition of economic carrying capacity. Ecol Econ 12:13–21
73. Woodcock A, Davis M (1978) Catastrophe theory. EP Dutton, New York

Books and Reviews

Bak P (1996) How nature works: The science of self-organized criticality. Copernicus Press, New York
Fisher M, Frohlich F (2001) Knowledge, complexity and innovation systems. Springer, Berlin
Gladwell M (2000) The tipping point: How little things can make a big difference. Little, Brown and Company, Boston
Holland J (1995) Hidden order: How adaptation builds complexity. Helix Books, Reading
Kauffman S (1996) At home in the universe: The search for laws of self-organization and complexity. Oxford University Press, New York
King I (2000) Social science and complexity: The scientific foundations. Nova Science Publishing, Huntington
Ormerod P (1998) Butterfly economics: A new general theory of social and economic behavior. Faber and Faber, New York

Human Robot Interaction

David Feil-Seifer[1], Maja J. Matarić[2]
[1] Computer Science Department, University of Southern California, Los Angeles, USA
[2] USC Center for Robotics and Embedded Systems, University of Southern California, Los Angeles, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Major HRI Influences in Popular Culture
Prominent Research Challenges
Benchmarks and Ethical Issues for HRI
Notable Conferences
Bibliography

Glossary

Anthropomorphic Resembling, or having the attributes, human form. Human qualities have been ascribed to inanimate objects, computer-animated characters, and mechanical objects, among others. When referring to a robot, anthropomorphism refers to how human-like that robot is.

AR Assistive Robotics refers to robot systems that are designed to give aid or support to a human user. Traditionally, the assistance has been physical. However, AR has more recently expanded to encompass other types of assistance, including social, motivational, and cognitive. AR thus includes socially assistive robotics (defined below), rehabilitation robots, wheelchair robots and other mobility aids, companion robots, and educational robots.

Autonomy The ability to exert independent control, to self-direct. When referring to robotics, autonomy is used to indicate how much control of a robot results from the robot itself (based on its sensory inputs and internal computation), and how much is exerted by a human operator through tele-operation (defined below).

Benchmarks A standard used to measure performance. In the robotics context, benchmarks can be practical, relating to the safety and task performance of a system, or more abstract, relating to the ethical and other aspects of the system.

Embodied Having physical form, a body. Robots are inherently embodied, having physical form and existing in the physical world. In contrast, computer characters

may or may not be embodied. Some animated characters have three-dimensional bodies with simulated physics, thereby satisfying the condition of being embodied, but not being in the physical world.

GSR Galvanic Skin Response. Measure of electrodermal activity (EDA), or skin conductance, a function of the eccrine gland. GSR has been shown to be related to emotional stimuli, making it a potential sensor for determining user state.

HCI Human–Computer Interaction. HCI is the study of interaction between humans and computers. HCI includes interface design, issues with usability, ethics, interaction, and hardware and software design.

HRI Human–Robot Interaction. HRI is the study of interaction dynamics between humans and robots. In contrast to HCI, which addresses human-computer interaction, HRI addresses the dynamics of interaction between humans and physical, embodied robots.

SAR Socially Assistive Robotics is the study of robots capable of providing assistance through social rather than physical interaction. SAR is the intersection of SIR (defined below) and AR (defined above). SAR work is focused on addressing societal needs, such as eldercare, education, and cognitive, physical, and social therapy.

SIR Socially Interactive Robotics describes robots that interact with humans through social interaction rather than through tele-operation. SIR is a subset of HRI that addresses the challenges of social interaction between humans and robots. SIR can also be referred to as social robotics.

Robot A mechanical system that takes inputs from sensors, processes them, and acts on its environment to perform tasks.

Tele-Operation The act of controlling a device (such as a robot) remotely. The use of tele-operation for a robot decreases the autonomy of that robot.

Definition of the Subject

Human–robot interaction (HRI) is the interdisciplinary study of interaction dynamics between humans and robots. Researchers and practitioners specializing in HRI come from a variety of fields, including engineering (electrical, mechanical, industrial, and design), computer science (human–computer interaction, artificial intelligence, robotics, natural language understanding, and computer vision), social sciences (psychology, cognitive science, communications, anthropology, and human factors), and humanities (ethics and philosophy).

Introduction

Robots are poised to fill a growing number of roles in today's society, from factory automation to service applications to medical care and entertainment. While robots were initially used in repetitive tasks where all human direction is given a priori, they are becoming involved in increasingly more complex and less structured tasks and activities, including interaction with people required to complete those tasks. This complexity has prompted the entirely new endeavor of Human–Robot Interaction (HRI), the study of how humans interact with robots, and how best to design and implement robot systems capable of accomplishing interactive tasks in human environments. The fundamental goal of HRI is to develop the principles and algorithms for robot systems that make them capable of direct, safe and effective interaction with humans. Many facets of HRI research relate to and draw from insights and principles from psychology, communication, anthropology, philosophy, and ethics, making HRI an inherently interdisciplinary endeavor.

Major HRI Influences in Popular Culture

Robots got their name in Čapek's play R.U.R. (Rossum's Universal Robots, 1921) [18]. In R.U.R., robots were man-made beings created to work for people and, as in many fictional stories thereafter, they went on to rebel and destroy the human race. In the 1950s, Isaac Asimov coined the term "robotics" and first examined the fundamental concepts of HRI, most prominently in his book *I, Robot* [3].

HRI has continued to be a topic of academic and popular culture interest. In fact, real-world robots have come into existence long after plays, novels, and movies developed them as notions and began to ask questions regarding how humans and robots would interact, and what their respective roles in society could be. While not every one of those popular culture works has affected the field of robotics research, there have been instances where ideas in the research world had their genesis in popular culture. In this section, significant popular culture products relating to HRI are overviewed, and their impact discussed.

The original benchmarks for HRI were proposed by Isaac Asimov in his now famous three laws of robotics:

1. *A robot may not injure a human being or, through inaction, allow a human being to come to harm.*
2. *A robot must obey orders given it by human beings except where such orders would conflict with the First Law.*
3. *A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.*

In *I, Robot* [3], the three laws were examined relative to commands that humans give robots, methods for humans to diagnose malfunctions, and ways in which robots can participate in society. The theoretical implications of how the three laws are designed to work has impacted the way that robot and agent systems operate today [140], even though the type of autonomous reasoning needed for implementing a system that obeys the three laws does not exist yet.

Philip K. Dick's novel *Do Androids Dream of Electric Sheep* [23] is set in a future world (originally in the late '90s) where robots (called replicants) mingle with humans. The replicants are humanoid robots that look and act like humans, and special tests are devised to determine if an individual is a human or a replicant. The test is related to the Turing Test [132], in that both involve asking probing questions that require human experiences and capacities in order to answer correctly. As is typical, the story also featured a battle between humans and replicants.

George Lucas' *Star Wars* movies (starting in 1977) feature two robot characters (C3P0 and R2D2) as key characters, which are active, intuitive, even heroic. One of the most interesting features from a robot design point of view is that, while one of the robots is humanoid in form (C3PO) and the other (R2D2) is not, both interact effectively with humans through social, assistive, and service interactions. C3P0 speaks, gestures, and acts as a less-than-courageous human. R2D2, on the other hand, interacts socially only through beeps and movement, but is understood and often preferred by the audience for its decisiveness and courage.

In the television show *Star Trek: The Next Generation* (1987–1994), an android named Data is a key team member with super-human intelligence but no emotions. Data's main dream was to become more human, finally mastering emotion. Data progressed to becoming an actor, a poet, a friend, and often a hero, presenting robots in a number of potentially positive roles.

The short story and movie *The Bicentennial Man* [4], features a robot who exhibits human-like creativity, carving sculptures from wood. Eventually, he strikes out on his own, on a quest to find like-minded robots. His quest turns to a desire to be recognized as a human. Through cooperation with a scientist, he develops artificial organs in order for him to bridge the divide between himself and other humans, benefiting both himself and humanity. Eventually, he is recognized as a human when he creates his own mortality.

These examples, among many others, serve to frame to scope of HRI research and exploration. They also provide some of the critical questions regarding robots and society that have become benchmarks for real-world robot systems.

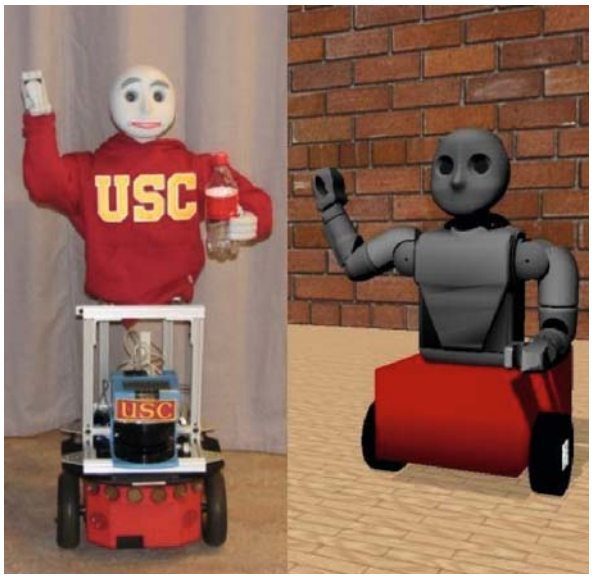

Human Robot Interaction, Figure 1
An example of an HRI testbed: a humanoid torso on a mobile platform, and a simulation of the same system

Prominent Research Challenges

The study of HRI contains a wide variety of challenges, some of them of basic research nature, exploring concepts general to HRI, and others of domain-specific nature, dealing with direct uses of robot systems that interact with humans in particular contexts. In this paper, we overview the following major research challenges within HRI: multi-modal sensing and perception; design and human factors; developmental and epigenetic robotics; social, service and assistive robotics; and robotics for education. Each is discussed in turn.

Multi-modal Perception

Real-time perception and dealing with uncertainty in sensing are some of the most enduring challenges of robotics. For HRI, the perceptual challenges are particularly complex, because of the need to perceive, understand, and react to human activity in real-time.

The range of sensor inputs for human interaction is far larger than for most other robotic domains in use today. HRI inputs include vision and speech, both major open challenges for real-time data processing. Computer vision methods that can process human-oriented data such as facial expression [10] and gestures [25] must be capa-

ble of handling a vast range of possible inputs and situations. Similarly, language understanding and dialog systems between human users and robots remain an open research challenge [49,141]. Tougher still is to obtain understanding of the connection between visual and linguistic data [106] and combining them toward improved sensing [110] and expression [14].

Even in the cases where the range of input for HRI-specific sensors is tractable, there is the added challenge of developing systems that can accomplish the sensory processing needed in a low-latency timeframe that is suitable for human interaction. For example, Kismet [13], an animated robotic head designed for infant-like interactions with a human, using object tracking for active vision, speech and prosody detection and imitation, and an actuated face for facial expressions, required several computers running in tandem to produce engaging if non-sensical facial and speech behavior. The humanoid ASIMO has been adapted to use a combination visual-auditory system for operation in indoor environments [109]. ASIMO's subsystems were used for perception, planning, and action with the goal of enabling human–robot interaction. Adding meaning to the facial and physical expressions and speech, and combining all of those capabilities in real time on a mobile, self-contained robot platform, is still an open research problem in HRI.

Even though most implemented HRI systems are necessarily domain-specific, as all physical systems, they still require the additional step of generalization to make them work beyond the research lab context. Computer vision solutions often depend on specific lighting conditions [51], ambient colors [116], and objects in the scene [15]. Beyond the lab, either the environment must be constrained to match the acceptable conditions for system operation [130], or the system capabilities must be extended in order to meet the range of conditions in the specific destination environment [83].

In addition to robot sensors that mimic the functionality of human perception (speech recognition, computer vision, etc.), sensors are being developed that cater to alternative sensing opportunities presented by an autonomous system such as a robot. These sensors enable a machine to observe people and the environment in ways that may be beyond human capability. Physiological signals, such as heart rate, blood pressure, galvanic skin response (GSR, the measure of skin conductance using a galvanometer), provide information about the user's emotional state [62,86,114] that may not otherwise be observable. Work by Mower et al. [90] used GSR as part of an HRI system to model and predict when a user is about to quit a rehabilitation-type task.

Body pose and movement are important sources of information for social interaction [106]. For example, social and expressive gestures are crucial components of human–human and human–robot interaction [121]. Computer vision can provide such information in limited contexts. In others, wearable sensors may be an effective means of obtaining human activity data in real time with high accuracy [85]. Such wearable systems have been used in HRI tasks applied to physical rehabilitation post-stroke [32], and for social interaction [124].

In addition to developing new and improving existing sensors toward particular needs of HRI, researchers are also developing algorithms for integrating multi-sensor multi-modal data inherent to HRI domains [29,44,91,93,109]. For example, Kapoor and Picard [57] implemented an affect recognition system that applies Gaussian models to fuse multiple sensors. Multi-modal sensing has also been used for a robot to detect the attention of human users in order to determine if a user is addressing the robot [71], integrating person tracking, face recognition [12], sound source localization [135], and leg detection [84].

Design and Human Factors

The design of the robot, particularly the human factor concerns, are a key aspect of HRI. Research in these areas draws from similar research in human–computer interaction (HCI) but features a number of significant differences related to the robot's physical real-world embodiment. The robot's physical embodiment, form and level of anthropomorphism, and simplicity or complexity of design, are some of the key research areas being explored.

Embodiment The most obvious and unique attribute of a robot is its physical embodiment. By studying the impact of physical embodiment on social interaction, HRI researchers hope to find measurable distinctions and trade-offs between robots and non-embodied systems (e. g., virtual companion agents, personal digital assistants, intelligent environments, etc.).

Little empirical work to date has compared robots to other social agents. Work by Bartneck et al. [9] claimed that robotic embodiment has no more effect on people's emotions than a virtual agent. Compelling recent work [60] used three characters, a human, a robot, and an animated character, to verbally instruct participants in a block stacking exercise. The study reported differences between the embodied and non-embodied agents: the robot was more engaging to the user than a simulated agent. Woods et al. [146] studied perception dif-

ferences between live and video recorded robot performances. They proposed using video recordings during system development as a complementary research tool for HRI.

Recent findings [138,139] suggest that there are several key differences between a robot and virtual agent in the context of human–machine interaction. The three conditions explored in that work (a physical robot body, a physical robot located elsewhere through a video link, and a simulation of a robot) were an attempt to control variables in order to isolate the effects of embodiment from realism. The researchers surveyed the participants regarding various properties related to the interaction. The results showed that the embodied robot was viewed by participants as more watchful, helpful, and appealing than either the realistic or non-realistic simulation.

Much work remains to be done in order to address the complex issues of physical embodiment in human–machine interaction. One confounding factor of this study involves the robot's form, discussed next.

Anthropomorphism The availability and sophistication of humanoid robots has recently soared. The humanoid form allows for exploring the use of robots for a vast variety of general tasks in human environments. This propels forward the various questions involved in studying the role of anthropomorphism in HRI. Evidence from communications research shows that people anthropomorphize computers and other objects, and that that anthropomorphism affects the nature of participant behavior during experiments [104].

HRI studies have verified that there are differences in interaction between anthropomorphic and non-anthropomorphic robots. For example, children with autism are known to respond to simple mobile car-like robots as well as to humanoid machines. However, pilot experiments have suggested that humanoid robots may be overwhelming and intimidating, while others have shown therapeutic benefit [107,111]. Biomimetic, and more specifically, anthropomorphic forms all for human-like gestures and direct imitation movements, while non-biomimetic form preserves the appeal of computers and mechanical objects.

Several examinations have been performed of the effects of anthropomorphic form on HRI [28]. These include studies of how people perceive humanoid robots compared to people and non-humanoid robots [99], possible benchmarks for evaluating the role of humanoid robots and their performance [54], and how the design of humanoid robots can be altered to affect user interacts with robots [24].

Simplicity/Complexity of Robot Design The simplicity/complexity of the robot's expressive behavior is related to the biomimetic/anthropomorphic property. Researchers are working to identify the effect that simple v. complex robot behavior has on people interacting with robots. For example, Parise et al. [100] examined the effects of life-like agents on task-oriented behavior. Powers and Kiesler [103] examined how two forms of agent embodiment and realism affect HRI for answering medical questions. Wainer et al. [138,139] used a similar experimental design to explore the effects of realism on task performance. In those studies, the more realistic or complex a robot was, the more watchful it seemed. However, it was also found that participants were less likely to share personal information with a realistic or complex robot.

Other Attributes In Reeves and Nass [104], several human factors concepts are explored in relation to human–computer interaction (HCI). As researchers work to better understand human–robot interaction, human factors insights from HCI can be valuable, but may not always be relevant. Lee and Nass [75] examined the relationship between a a virtual agent's voice and its personality. The authors found that users experienced a stronger sense of social presence from the agent when the voice type and personality matched, than when they did not. In an HRI study, Tapus and Matarić [126] showed that when a robot's expressive personality matched the user's personality, task performance was better than when the personalities were mismatched. Robles et al. [108] used agents that gave feedback for a speed-dating application to examine users' feelings regarding monitoring (public and private), conformity, and self-consciousness. This study correlated users' actions with surveyed perceptions regarding feedback to determine how feedback can be most effectively given, and how it can be given in as effective a context as possible. Kidd and Breazeal [60] used a similar design to evaluate how a robot (compared to a computer agent or to a human) can give feedback for making decisions.

Ongoing research is also exploring how cultural norms and customs can affect the use of computer agent and robot systems. For example, Takeuchi et al. [125] designed an experiment to test the differences in behavior reciprocity between users of a virtual agent in the US and users in Japan. They discovered that users from both countries expressed attitudes consistent with behavior reciprocity, but only US users exhibited reciprocal behavior. However, they discovered that when recognizable brands from popular culture were used, then reciprocal behavior was exhibited in Japanese users as well.

Developmental/Epigenetic Robotics

Developmental robotics, sometimes referred to as epigenetic robotics, studies robot cognitive development. Developmental roboticists are focused on creating intelligent machines by endowing them with the ability to autonomously acquire skills and information [142]. Research into developmental/epigenetic robotics spans a broad range of approaches. One effort has studied teaching task behavior using shaping and joint attention [15], a primary means used by children in observing the behavior of others in learning tasks [92,94]. Developmental work includes the design of primitives for humanoid movements [26], gestures [69], and dialog [115].

While developmental/epigenetic robotics is not a direct subset of HRI research, there is significant overlap in the goals of the two areas. Developmental techniques for information acquisition share much in common with multi-modal perception. Epigenetic research into pronoun learning has overlap with social robotics [41]. Finally, techniques for automated teaching and learning of skills has direct applications for algorithm development for education robotics [66,95]. This work involves estimating behavior from human actions [67]. In the broader field of robot learning, a variety of methods are being developed for robot instruction from human demonstration [44,68,96,102], from reinforcement learning [134], and from genetic programming [97], among others.

Social, Service, and Assistive Robotics

Service and assistive robotics [31] include a very broad spectrum of application domains, such as office assistants [5,43], autonomous rehabilitation aids [79], and educational robots [128]. This broad area integrates basic HRI research with real-world domains that required some service or assistive function. The study of social robots (or socially interactive robots) focuses on social interaction [35], and so is a proper subset of problems studied under HRI.

Assistive robotics itself has not been formally defined or surveyed. An assistive robot is broadly defined as one that gives aid or support to a human user. Research into assistive robotics includes rehabilitation robots [16,27,47,53,78], wheelchair robots and other mobility aides [2,40,118,147], companion robots [8,101,136], manipulator arms for the physically disabled [39,42,59], and educational robots [55]. These robots are intended for use in a range of environments including schools, hospitals, and homes. In the past, assistive robotics (AR) has largely referred to robots developed to assist people through physical interaction. This definition has been significantly broadened in the last several years, in response to the growing field of AR in which assistive robots provide help through non-contact, social interaction, defining the new field of socially assistive robotics (SAR).

Socially assistive robotics (SAR) is a growing area of research with potential benefits for elder care, education, people with social and cognitive disorders, and rehabilitation, among others [33]. SAR is the intersection of assistive robotics, which focuses on robots whose primary goal is assistance, and socially interactive robotics [35], which addresses robots whose primary feature is social interaction. SAR arose out of the large and growing body of problem domains suitable for robot assistance that involves social rather than physical interaction [77,129,144].

In rehabilitation robotics, an area that has typically developed physically-assistive robots, non-contact assistive robots are now being developed and evaluated. These robots fulfill a combined role of coach, nurse, and companion in order to motivate and monitor the user during the process of rehabilitation therapy. Observing the user's progress, the robots provide personalized encouragement and guidance. Applications for post-operative cardiac surgery recovery [56] and post-stroke rehabilitation [79] have been studied. Other rehabilitation projects have explored using a robot as a means of motivating rehabilitation through mutual storytelling [72,101]. In these experiments, a robot and a user constructs a story, which, when acted out, require the user to perform physical therapy exercises.

A variety of assistive robotics systems have been studied for use by the elderly. Such robots are meant to be used in the home, in assisted living facilities, and in hospital settings. They work to automate some physical tasks that an elderly person may not be able to do, including feeding [59], brushing teeth [131], getting in and out of bed, getting into and out of a wheelchair, and adjusting a bed for maximum comfort [52]. In some cases, the robots are envisioned as part of a ubiquitous computing system [52], which combines cameras and other sensors in the environment and computer controlled appliances (such as light switches, doors, and televisions) [8]. In others, the robots serve SAR roles such as promoting physical and cognitive exercise [127].

HRI systems have been used as companion robots in the public areas of nursing homes, aimed at increasing resident socialization. These robots are designed not to provide a specific therapeutic function, but to be a focus of resident attention. One such example is the huggable, a robot outfitted with several sensors to detect different types of touch [123]. Another such example is NurseBot, a robot used to guide users around a nursing home [87]. Paro [136,137], an actuated stuffed seal, behaves in re-

Human Robot Interaction, Figure 2
Examples of SAR research. *Left*: post-cardiac surgery convalescence. *Middle:* post-stroke rehabilitation. *Right*: cognitive and physical exercises

sponse to touch and sound. Its goal is to provide the benefits of pet-assisted therapy, which can affect resident quality of life [30], in nursing homes that cannot support pets. Initial studies have shown lowered stress levels in residents interacting with this robot, as well as an overall increase in the amount of socialization among residents in the common areas of the same facility.

Finally, HRI is being studied as a tool for diagnosis [111,112] and socialization [22,70,83,143] of children with autism spectrum disorders (ASD). When used for diagnosis, robots can observe children in ways that humans cannot. In particular, eye-tracking studies have shown remarkable promise when evaluating children for the purposes of diagnosing ASD. In terms of socialization, robots are a more comfortable social partner for children with ASD than people. These robots encourage social behavior, such as dancing, singing, and playing, with the robot and with other children or parents in the hope of making such behavior more natural.

Educational Robotics

Robotics has been shown to be a powerful tool for learning, not only as a topic of study, but also for other more general aspects of science, technology, engineering, and math (STEM) education. A central aspect of STEM education is problem-solving, and robots serve as excellent means for teaching problem-solving skills in group settings. Based on the mounting success of robotics courses world-wide, there is now is an active movement to develop robot hardware and software in service of education, starting from the youngest elementary school ages and up [50,80,81]. Robotics is becoming an important tool for teaching computer science and introductory college engineering [81].

Robot competition leagues such as Botball [122], RoboCup [120] and FIRST [98] have become vastly popular. The endeavors encourage focused hands-on problem solving, team work, and innovation, and range from middle- and high-school-age children up to university teams. Educators are also using robots as tools for service learning, where projects are designed for assistive domains. Innovative teaching methods include competitions to develop robot toys for children with ASD [82] and other assistive environments [48].

In some specific domains, robots have been shown to be better for instruction than people [73]. While some automated systems are used for regular academic instruction [45], others are used for social skill instruction. In particular, robots can be used to teach social skills such as imitation [107], and self-initiation of behavior [65], in addition, they are being explored as potentially powerful tools for special education [58].

Benchmarks and Ethical Issues for HRI

As HRI systems are being developed, their impact on users and society at large are increasingly being considered [34]. Currently, it is difficult to compare robotic systems designed for different problem domains, yet it is important to do so in order to establish benchmarks for effective and ethical HRI design. Kahn et al. [54] argued for comparative methods and proposed benchmarks for HRI, with a particular focus on gaining a better understanding humanoid robots designed for HRI.

One of the most challenging aspects of establishing such benchmarks is that many aspects of HRI are difficult to measure. Establishing whether or not a robot can make eye contact with a person is comparatively simple (if not always easy to implement), but evaluating how the person reacts to and is affected by the robot's gaze and behavior is much more difficult. Does the user get bored or frustrated? Does the user consider the robot helpful and effective? Is the robot perceived as competent? Is it trusted to perform its intended tasks?

These and related questions lead to ethical considerations and legal guidelines that need to be addressed when developing HRI systems. Not only do roboticists

need to act ethically, the robots themselves must do so as well. Challenges to be considered include unintended uses of the robot, allowable tasks, and unintended situations that might be encountered. For example, if the user needs emergency attention, what is the robot's responsibility? Furthermore, the issue of control has important implications. While it is assumed the user is in control, in a variety of situations (dispensing medicine, dealing with cognitively incapacitated users) the control responsibility must rest with the machine. The issue of control and authority thus extends to all involved with the machine, including caretakers, and even designers and programmers. Well-studied ethical challenges are gradually making their way into HRI as the systems are growing in complexity and usefulness, and as their likelihood of entering human daily life increases.

General Benchmarks and Ethical Theory

While no specific ethical guidelines have yet been established, active discussions and task forces have taken up this challenging problem. Turkle [133] addressed the attachment that occurs between humans and robots when residents of a nursing home are asked to "care for" a baby-like robot. The users in the experiment ascribed human-like qualities to the robot, resulting in side-effects with ethical ramifications. What happens when the robot breaks down? What if the robot is taken away? Some benchmarks address the disparity between machines that exist only to serve a human "master" and those that exist in cooperation with their users and act with autonomy [54]. Is it acceptable for people to treat a social being like a slave?

The nature of morality for androids and other artificially intelligent entities has also been explored [140] and the difference between top-down and bottom-up morality defined. A top-down approach to morality is any approach that takes an ethical theory and guides the design and implementation of algorithms and subsystems capable of implementing that ethical theory. A bottom-up approach involves treating values as implicit to the design of the robot. In that work, morality (either implied or explicitly programmed) helps guide the behavior of robots to effectively work with humans in social situations.

Yanco [147] described the evaluation of an assistive robot, stating that such evaluation can be done through user tests and comparison to a human in the same assistive role. Long-term studies were recommended in order to evaluate effectiveness in real-world settings. Others advocated a human-centered approach to design, suggesting ecological studies of the use of the robots in the intended environment rather than long-term user studies [37].

Robot Evaluation

Any robot is a physical and technological platform that must be properly evaluated. In this section, two evaluation benchmarks of particular concern to HRI, safety and scalability, are discussed.

Safety Safety is an important benchmark for HRI: *How safe is the robot itself, and how safe can the robot make life for its user?*

A robot's safety in its given domain is the primary concern when evaluating an HRI system. If a robot is not designed with safety in mind, it could harm the very users it is designed to interact with. A key advantage of HRI over physically assistive robots is the minimization of the inherent safety risk associated with physical contact. When discussing safety pertaining to a mobile platform, we refer to the ability to maneuver about a scene without unwanted contact or collisions. Safety also refers to protection (as much as it is possible) of a robot's user and of the robot itself. This concept, as a benchmark, refers to safety in a bottom-up fashion, rather than Asimov's laws which refer to the concept in a top-down fashion [140].

Safety for assistive robots has been studied in depth in the contexts of obstacle avoidance for guide-canes and wheelchairs [7,105,147]. Robots have also been designed to help users navigate through a nursing home [40,89]. The need for safety assessment for HRI systems designed for vulnerable user populations is a topic of growing importance as HRI systems are increasingly being developed toward users from such populations.

Scalability The majority of current HRI work occurs in research laboratories, where systems are engineered for one environment and a pre-determined prototype user population. As HRI becomes more widespread in homes, schools, hospitals, and other daily environments, the question of scalability and adaptability arises: *How well will such HRI systems perform outside of the lab?* and *How well does a robot perform with users from the general population?*

The scalability benchmark does not imply that roboticists should design each robot for a large a variety of situations where assistance is required. Rather, it is important to stress that, even within a group that needs assistance, there is a great difference between a "prototypical" user or environment and the range of real-world users and environments.

Another key question to address is: *How many people can be helped by such a robot?* Consider, for example, a robot that uses speech recognition for understand-

ing a user's intentions. How does speech recognition perform when the speaker has recently suffered a stroke? Can the robot interact with someone who cannot speak? If the robot is meant to be a companion for a user, can the robot adapt its behavior to different users? How difficult is it for the robot to be modified for different needs?

In addition to user population scalability, the range of usable environments is an important benchmark. Most systems to date have been tested in research labs or controlled hospital and managed care settings. In the future, however, HRI systems will be used in homes and other more unpredictable environments. In such domains, the following benchmark becomes relevant: *Can the robot operate in the most relevant environments for the user?*

Social Interaction Evaluation

A critical benchmark of HRI is the evaluation of the robot as a social platform. Social interaction and engagement are both the primary means of interaction and the driving force behind the design. When assessing a robot in terms of social performance, we must also consider the larger goal of the robot in its application context.

Previously proposed benchmarks for humanoid robots [54] are directly relevant to HRI as well. In many respects, the same comparisons and evaluations that hold for humanoid robotics also hold for HRI. However, the goal of HRI is not to make as interesting or realistic a robot as possible, but to make a robot that can best carry out its task. It is important, therefore, to evaluate HRI not only from a perspective of modeling human characteristics, but also from a user-oriented perspective. The following sections describe how some of the previously identified humanoid benchmarks that relate to HRI.

Autonomy Autonomy is a complex property in the HRI context. It is favorable, when constructing a system that is designed to stand in for a human in a given situation, to have a degree of autonomy which allows it to perform well in its desired tasks. Autonomy can speed up applications for HRI by not requiring human input, and by providing rich and stimulating interactions. For example, HRI systems for proactive social interaction with children with ASD [22] and motivational robot tools [79,126,138] require such autonomy. However, autonomy can also lead to undesirable behavior. In situations such as medication dispension and therapy monitoring [38], for example, autonomy is not desirable.

In general, HRI contexts require engaging and believable social interaction, but the user must clearly retain authority. For example, rehabilitation should terminate if the user is in pain. Social interaction should only occur when it is tolerable for the user. Partial or adjustable autonomy on the part of the HRI system allows for an appropriate adjustment of both authority and autonomy.

Imitation Alan Turing proposed a test of artificial intelligence (AI), whereby a system is evaluated by whether it could fool a human user communicating with it through teletype [132]. This test was later elaborated to the Total Turing Test [46], where a system communicating in human-like ways (text, speech, facial expressions) tries to fool a human user into believing it is human. Since that time, one of the benchmarks for success in AI and HRI has been how well the system can imitate human behavior. However, when dealing with goal-oriented systems not primarily relating to human behavior but rather to assistance and treatment, imitating human behavior is necessarily desirable.

It has been shown that a robot's personality can effect a user's compliance with that robot [61]. When exhibiting a serious personality, the robot could provoke a greater degree of compliance than displaying a playful personality. It has also been shown that when the robot's extroversion/introversion personality traits matched the user's, task performance was improved [126]. Thus, the imitation benchmark proposed by Kahn could be revised for HRI: *How do imitation and reciprocity affect task performance?*

While no definitive evidence yet exists, there is a good deal of theory regarding a negative correlation between the robot's physical realism and its effectiveness in human–robot interaction. Realistic robotics introduces new complications to social robot design [28] and it has been implied that anthropomorphism has a negative influence on social interaction when the robot's behavior does not meet a user's expectations [117]. The Uncanny Valley theory suggests that as a robot becomes very similar in appearance to a human, that robot appears less, rather than more, familiar [88]. Physical similarity that attempts in imitation of human-like appearance and behavior could cause discord. This leads to two possible benchmark for imitation: *Does the interaction between the human and the robot reflect an accurate and effective impression of the robot's capabilities?* and *Does the interaction between the human and the robot allow for the expression of the human's capabilities?*

Privacy The presence of a robot inherently affects a user's sense of privacy [54]. In contrast to ubiquitous systems [11,63,76] where a user has no idea of when the system may be watching, robots are tangible and their perception limited and observable. A robot can be told to

leave when privacy is desired, and the user can observe when privacy is achieved. Because of its synthetic nature, a robot is perceived as less of a privacy invasion than a person, especially in potentially embarrassing situations. Since privacy is such a concern for designers of assistive systems [6]. Therefore, a possible benchmark from an HRI perspective asks: *Does the user's perceived sense of privacy relate to better robot performance as an assistive presence?*

Task-Oriented Benchmarks

The interactive, task-oriented nature of HRI suggests some additional benchmarks. Task performance is described as the ability of the robot to assist a user in a given task. The benchmarks then pertain to how the social aspects of the robot affect the overall task performance of the robot and its user. As with the other benchmarks, discussed above, these could apply to all social robots, but when put into an assistive context, the task-related effects highlight these features.

Social Success *Does the robot successfully achieve the desired social identity?* This is perhaps the most amorphous of benchmarks, but its evaluation is simple. When the robot is intended to be playful, do users find the robot playful? If the robot is supposed to be a social peer, do users act as if it were a social peer? How does the intended social identity compare to what occurs in practice? This benchmark is not meant to judge the ability of the robot system designer to generate a suitable robot personality. The social success of the robot is a fundamental component of HRI applications. As discussed above, the social identity of the robot (both the personality and the role of the robot) has an effect on the user's task performance.

Understanding of Domain Understanding of social dynamics is a critical component of HRI. Roboticists employ user and activity modeling as means of achieving such understanding. Efforts to understand a user of an HRI system include emotion recognition [19,20], and integration of vocalizations, speech, language, motor acts, and gestures [17,74] for effectively modeling user state.

Sensing social understanding and engagement can be be assessed through a variety of means. Roboticists have also used radio frequency identification (RFID) tags and position tracking to observe children in school hallways to detect when users were in social range, and who they were interacting with over time [55], to help the robot determine appropriate social responses. Thus, social understanding in HRI can come from both human-oriented social perception (such as the interpretation of gestures, speech, and facial expressions), and from an evaluation of user physiologic state (such as GSR, heart rate, temperature, etc.). How such data are used leads to the following benchmark: *Does a robot's social understanding of human behavior help task performance?*

Evaluation as an Assistive Tool

For the domains of HRI, impact on user's care, impact on caregivers, impact on the user's life, and the role of the robot are the key benchmarks for an assistive platform. An important way to view how an assistive robot performs when caring for people is by first observing how people care for other people in similar situations. The role of an assistive robot may be that of a stand-in for a human caregiver, a complement for a human caregiver, or an assistant to a human caregiver. Naturally, the benchmarks have different application in various scenarios. As with the other benchmarks, discussed above, this is not meant to be a comprehensive list, but a consideration of some of the most relevant benchmarks.

Success Relative to Human Caregiver A good place to start when evaluating the effect a robot has on a user's care is to compare the results of care with a robot caregiver to that of care with a human caregiver: *How does the robot perform relative to a human performing the same task?* When such evaluation is possible, existing metrics can be applied. For example, in rehabilitation tasks, functional improvement can be a metric [79]. For learning tasks, overall learning measures such as grades, tests, or evaluations can be used. In a spirometry task where a robot instructed a cardiac surgery patient to do breathing exercises [56], compliance with the robot compared to compliance with a human was a suitable metric. For companion robots, evaluating user satisfaction is most relevant.

A key role of assistive HRI is to provide care where human care is not available. In many cases, the type of interaction that is established in HRI is not directly comparable to human care, and in some instances, human care is not available for comparison. In all cases, the user satisfaction and motivation to engage in the relevant activities is a key metric of system effectiveness, on par with functional measures of task performance.

Cost/Benefit Analysis The robot can perform in several different capacities for any given task. For example, in a rehabilitation setting a robot could serve as therapist, giving advice on specific movements, a motivational coach, giving general encouragement and monitoring progress, a cognitive orthotic, reminding the users of important

items, a companion, a learning aid, or as a demonstration, showing a user how to do specific exercises. The role of the robot for a given task can inform the complexity and sophistication of the robot and its social and assistive capacities.

HRI is intended as a tool for creating robotic systems capable of providing cost-effective solutions to a variety of applications. Cost/benefit analysis can thus a benchmark for success for such systems. In domains where no alternatives exist, and where HRI systems provide a novel and only solution, have the potential of creating major societal impact. Health care is one such domain. This suggests two benchmarks for HRI: *Does the use of the robot (a) change the cost/benefit ratio of providing such care or (b) make such care available where it was not previously possible?*

Impact on Caregivers In some cases, the goal of automation is not to increase the efficiency, productivity, or standard of care, but to make the user's or caregivers' job easier and more manageable. For example, the goal of the robot described above in Kang et al. [56] was to reduce the overall workload for cardiac nurses, given the overall nurse shortage in the US and world-wide. The robot visited cardiac patients post-surgery, approached each patient's bed, encouraged the patient to perform the breathing exercise, monitored the number and depth of the breaths taken, and collected performance data. By automating the prompting and monitoring of spirometry, which must be performed ten times per hour for the critical post-surgery period, the robot made it possible for caregivers to attend to other tasks and provide more individualized services. However, in this case, the robot did not provide any care not already provided by a human caregiver.

Caregiver impact is thus a useful benchmark: *Does the job condition of the caregiver improve as a result of the robot?* Additionally, it is important to observe cooperation: *How well does the caregiver work with the robot?* This arises out of a concern that trained and experienced caregivers are not used to working with robots, and may need to adjust their work habits [113].

Satisfaction with Care User satisfaction is an important aspect of assistive therapy success. Users' impression of a nurse robot's personality affects compliance with that robot, both positively and negatively [61]. Satisfaction, therefore can be a useful benchmark for success. Questionnaires are being explored [138,139] to measure satisfaction, although little work to date has directly related satisfaction with a robot system to task performance or user compliance. An important question when designing an assistive system is raised: *Does user satisfaction with a system affect the assistive task performance and/or user compliance?*

Existing Quality of Life Measures Evaluating the effects of a particular therapy regimen must be done relative to the overall quality of life (QoL) of the user [145]. Some recommend using repeated measures with the same survey to capture changes over time. The SF-36 survey is designed for patient rating of health-related quality of life [1]. This survey assesses the comprehensive quality of life from the patient's point of view. The 15-D survey produces quality of life numbers along several dimensions [119]. In addition to such quantifiable measures, experiential measures, such as the Dementia Care Mapping (DCM), are also used broadly [64]. Such measures bring to the forefront the users of a particular type of service [148], as well as the notion that socially-sensitive care (involving eye-contact, favorable attention, etc.) is important to the overall outcome. This leads to a suitable HRI benchmark: *Does the robot result in a general increase in the quality of life as perceived by the user?*

Impact on the Role in Community/Society The introduction of automation and HRI-capable systems has an affect the user community. When fish tanks were introduced into a nursing home environment to test the effects on residents, observers found an overall increase in nutrition on the part of the participating residents [30]. A side-effect of the installation of the fish tanks was that residents gathered around those situated in common areas and engaged in more conversation than was previously observed. The introduction of new objects of social interest into an environment can thus change the dynamics of the community.

When roboticists introduced the robot seal Paro into the common areas of a nursing home [136,137], they found a reduction of stress proteins in the urine of the participants. Another positive effect of the experiment was that residents were in the common areas longer and socialized more. The Robovie project was able to use a robot to stimulate social interaction among a group of elementary school students [55]. By telling "secrets" about itself, the robot was able to elevate a student's status in the group by giving him/her special information [21].

An ethnographic study used readily-available low-cost robot vacuum cleaners to determine the role that the robots played in household [36]. The study used home tours and semi-structured interviews to create an ecological model of the home. The data provided insights into how a service robot might be treated, and how close the real users came to the design intention of the robot. Some treated the robot as if it were a member of the house-

hold, with status roughly between the vacuum cleaner and a pet. Others treated it strictly as a device with a purpose. An interesting observation is that men got more involved in cleaning tasks associated with the Roomba (pre-cleaning, activation, and emptying the unit when the task was completed).

A potential critique of assistive robotics is that social robots capable of HRI could reduce the amount of human contact for their users. Thus, when assessing a particular robot-assisted therapy, it is important to note not only the immediate effects on a single user, but also the effects that the robot has on the community as a whole: *Does the robot increase or decrease the amount of socialization in its user community?* and: *Are changes in community due to a robot positive or negative?*

Notable Conferences

HRI is an active and growing area of research. Progress in the field is discussed and showcased at a number of conferences, symposia, and workshops. Research results are published both in new and growing HRI conferences and journals, and the more established venues of the parent fields of HRI, namely robotics and AI.

Human–Robot Interaction-Specific Conferences

- **Conference on Human Robot Interaction (HRI)** This conference, created in 2006, is focused specifically on HRI research. Attendees and submissions to this conference are mostly from engineering (electrical engineering and computer science) with contributions from allied fields, such as psychology, anthropology, and ethics.
- **International Workshop on Robot and Human Interactive Communication (RO-MAN)** RO-MAN provides a forum for an interdisciplinary exchange for researchers dedicated to advancing knowledge in the field of human–robot interaction and communication. Importantly, RO-MAN has traditionally adopted a broad perspective encompassing research issues of human–machine interaction and communication in networked media as well as virtual and augmented tele-presence environments. RO-MAN is somewhat longer-standing than HRI.
- **International Conference on Development and Learning (ICDL)** This conference brings together the research community at the convergence of artificial intelligence, developmental psychology, cognitive science, neuroscience, and robotics, aimed at identifying common computational principles of development and learning in artificial and natural systems. The goal of the conference is to present state-of-the-art research on autonomous development in humans, animals and robots, and to continue to identify new interdisciplinary research directions for the future of the field.
- **Computer/Human Interaction (CHI) Conference** CHI is an established conference in Human–Computer Interaction (HCI). Every year, it is a venue for 2000 HCI professionals, academics, and students to discuss HCI issues and research and make lasting connections in the HCI community. HRI representation in this meeting is small, but the two fields (HRI and HCI) have much to learn and gain from each other.

General Robotics and AI Conferences

- **Association for the Advancement of Artificial Intelligence (AAAI)** AAAI's annual conference affords participants a setting where they can share ideas and learn from each other's artificial intelligence (AI) research. Topics for the symposia change each year, and the limited seating capacity and relaxed atmosphere allow for workshoplike interaction.
- **AAAI Spring and Fall Symposia** These annual symposia cover a broad range of focused topics. With the rapid growth of HRI, the topic and related areas (e. g., service robotics, socially assistive robotics, etc.) symposia are held in each session.
- **Epigenetic Robotics (EpiRob)** The Epigenetic Robotics annual workshop has established itself as an opportunity for original research combining developmental sciences, neuroscience, biology, and cognitive robotics and artificial intelligence is being presented.
- **International Conference on Robotics and Automation (ICRA)** This is one of two most major robotics conferences, covering all areas of robotics and automation. In recent years, the themes of the conference have included many areas of HRI research, such as "Humanitarian Robotics," "Ubiquitous Robotics," and "Human-Centered Robotics", reflecting the rapid growth in the field.
- **International Conference on Intelligent Robots and Systems (IROS)** This is the other major international robotics conference, featuring a very large number of papers, with a growing representation of HRI. Tutorials and workshops, as well as organized/special sessions in HRI are featured regularly.
- **International Symposium on Experimental Robotics (ISER)** ISER is a single-track symposium featuring around 50 presentations on experimental research in robotics. The goal of these symposia is to provide a forum dedicated to experimental robotics research with

principled foundations. HRI topics have become a regular part of this venue.

Bibliography

1. Aaronson NK, Acquadro C, Alonso J, Apolone G, Bucquet D, Bullinger M, Bungay K, Fukuhara S, Gandek B, Keller S, Razavi D, Sanson-Fisher R, Sullivan M, Wood-Dauphinee S, Wagner A Jr JEW (2004) International quality of life assessment (iqola) project. Qual Life Res 1(5):349–351
2. Aigner P, McCarragher B (1999) Shared control framework applied to a robotic aid for the blind. Control Syst Mag IEEE 19(2):40–46
3. Asimov I (1950) I, robot. Doubleday, New York
4. Asimov I (1976) Bicentennial man. Ballantine Books, New York
5. Asoh H, Hayamizu S, Hara I, Motomura Y, Akaho S, Matsui T (1997) Socially embedded learning of the office-conversant mobile robot jijo-2. In: International joint conference on artificial intelligence (IJCAI), Nagoya, Japan
6. Baillie L, Pucher M, Képesi M (2004) A supportive multimodal mobile robot for the home. In: Stary C, Stephanidis C (eds) User-centered interaction paradigms for universal access in the information society, Lecture notes in computer science, vol 3196/2004. Springer, Berlin, pp 375–383
7. Baker M, Yanco H (2005) Automated street crossing for assistive robots. In: Proceedings of the international conference on rehabilitation robotics, Chicago, Il, pp 187– 192
8. Baltus G, Fox D, Gemperle F, Goetz J, Hirsh T, Magaritis D, Montemerlo M, Pineau J, Roy N, Schulte J, Thrun S (2000) Towards personal service robots for the elderly. In: Proceedings of the workshop on interactive robots and entertainment, Pittsburgh, PA
9. Bartneck C, Reichenbach J, v Breemen A (2004) In your face, robot! the influence of a character's embodiment on how users perceive its emotional expressions. In: Proceedings of the design and emotion 2004 conference, Ankara, Turkey
10. Betke M, Mullally W, Magee J (2000) Active detection of eye scleras in real time. In: Proceedings of the IEEE workshop on human modeling, analysis and synthesis, Hilton Head, South Carolina
11. Bien Z, Park K, Bang W, Stefanov D (2002) LARES: An Intelligent Sweet Home for Assisting the Elderly and the Handicapped. In: Proceedings of the 1st Cambridge workshop on universal access and assistive technology (CWUAAT), Cambridge, UK, pp 43–46
12. Bradski G et al (1998) Computer vision faxce tracking for use in a perceptual user interface. Intel Technol J 2(2):12–21
13. Breazeal C (2000) Infant-like social interactions between a robot and a human caretaker. Adapt Behav 8(1):49–74
14. Breazeal C, Edsinger A, Fitzpatrick P, Scassellati B (2001) Active vision for sociable robots. IEEE Transactions on Man, Cybern Syst 31(5)
15. Breazeal C, Hoffman G, Lockerd A (2004) Teaching and working with robots as a collaboration. In: Proceedings of the international joint conference on autonomous agents and multiagent systems, vol 3. New York, pp 1030–1037
16. Burgar C, Lum P, Shor P, van der Loos H (2002) Development of robots for rehabilitation therapy: The palo alto va/standford experience. J Rehabil Res Dev 37(6):663–673
17. Busso C, Deng Z, Yildirim S, Bulut M, Lee C, Kazemzadeh A, Lee S, Neumann U, Narayanan S (2004) Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of the international conference on multimodal interfaces, State Park, PA, pp 205–211
18. Capek K (2001) Rossum's universal robots. Dover Publications, New York
19. Cassell J, Sullivan J, Prevost S, Churchill E (2000) Embodied conversational agents. MIT Press, Cambridge
20. Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor JG (2001) Emotion recognition in human–computer interaction. IEEE Signal Process Mag 18(1): 32–80
21. Cowley S, Kanda H (2005) Friendly machines: Interaction-oriented robots today and tomorrow. Altern 12(1a):79–106
22. Dautenhahn K, Werry I (2002) A quantitative technique for analysing robot-human interactions. In: Proceedings of the IEEE/RSJ, international conference on intelligent robots and systems, Lausanne, Switzerland, pp 1132–1138
23. Dick P (1968) Do androids dream of electric sheep. Doubleday, New York
24. DiSalvo C, Gemperle F, Forlizzi J, Kiesler S (2002) All robots are not created equal: Design and the perception of humanoid robot heads. In: Proceedings of the conference on designing interactive systems: processes, practices, methods, and techniques, London, England, pp 321–326
25. Drumwright E, Jenkins OC, Matarić MJ (2004) Exemplar-based primitives for humanoid movement classification and control. In: IEEE International conference on robotics and automation, pp 140–145
26. Drumwright E, Ng-Thow-Hing V, Matarić MJ (2006) Toward a vocabulary of primitive task programs for humanoid robots. In: International conference on development and learning, Bloomington, IN
27. Dubowsky S, Genot F, Godding S, Kozono H, Skwersky A, Yu H, Yu LS (2000) PAMM – a robotic aid to the elderly for mobility assistance and monitoring. In: IEEE International conference on robotics and automation, vol 1. San Francisco, CA, pp 570–576
28. Duffy B (2003) Anthropomorphism and the social robot. Robot Autonom Syst 42(3):177–190
29. Duquette A, Mercier H, Michaud F (2006) Investigating the use of a mobile robotic toy as an imitation agent for children with autism. In: Proceedingsof the international conference on epigenetic robotics: modeling cognitive development in robotic systems, Paris, France
30. Edwards N, Beck A (2002) Animal-assisted therapy and nutrition in Alzheimer's disease. West J Nurs Res 24(6):697–712
31. Engelberger JF (1989) Robotics in service. MIT Press, Cambridge
32. Eriksson J, Matarić MJ, Winstein C (2005) Hands-off assistive robotics for post-stroke arm rehabilitation. In: Proceedings of the international conference on rehabilitation robotics, Chicago, Il, pp 21– 24
33. Feil-Seifer D, Matarić MJ (2005) Defining socially assistive robotics. In: Proceedings of the international conference on rehabilitation robotics, Chicago, Il, pp 465–468
34. Feil-Seifer DJ, Skinner KM, Matarić MJ (2007) Benchmarks for evaluating socially assistive robotics. Interact Stud: Psychol Benchmarks Human–Robot Interact 8(3):423–439

35. Fong T, Nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. Robot Autonom Syst 42(3-4): 143–166
36. Forlizzi J, DiSalvo C (2006) Service robots in the domestic environment: A study of the Roomba vacuum in the home. In: Proceeding of the 1st ACM SIGCHI/SIGART conference on human–robot interaction. Academic, New York, pp 258–265
37. Forlizzi J, DiSalvo C, Gemperle F (2004) Assistive robotics and an ecology of elders living independently in their homes. Human–Comp Interact 19(1,2):25–59
38. Fortescue E, Kaushal R, Landrigan C, McKenna K, Clapp M, Federico F, Goldmann D, Bates D (2003) Prioritizing strategies for preventing medication errors and adverse drug events in pediatric inpatients. Pediatr 111(4):722–729
39. Giminez A, Balaguer C, Sabatini SM, Genovese V (2003) The MATS robotic system to assist disabled people in their home environments. In: Proceedings of the international conference on intelligent robots and systems, vol 3. Las Vegas, Nevada, pp 2612–2617
40. Glover J, Holstius D, Manojlovich M, Montgomery K, Powers A, Wu J, Kiesler S, Matthews J, Thrun S (2003) A robotically-augmented walker for older adults. Tech. Rep. CMU-CS-03-170, Carnegie Mellon University, Computer Science Department, Pittsburgh, PA
41. Gold K, Scassellati B (2005) Learning about the self and others through contingency. In: AAAI spring symposium on developmental robotics, Stanford, CA
42. Graf B, Hans M, Kubacki J, Schraft R (2002) Robotic home assistant care-o-bot II. In: Proceedings of the joint EMBS/BMES conference, vol 3. Houston, TX, pp 2343–2344
43. Green A, Huttenrauch H, Norman M, Oestreicher L, Eklundh K (2000) User centered design for intelligent service robots. In: Proceedings of the international workshop on robot and human interactive communication, Osaka, Japan, pp 161–166
44. Grollman D, Jenkins O (2007) Learning elements of robot soccer from demonstration. In: Proceedings of the international conference on development and learning (ICDL), London, England
45. Grynszpan O, Martin J, Nadel J (2007) Exploring the influence of task assignment and output modalities on computerized training for autism. Interact Stud 8(2):241–266
46. Harnad S (1989) Minds, machines and searle. J Exp Theor Artif Intell 1:5–25
47. Harwin W, Ginige A, Jackson R (1988) A robot workstation for use in education of the physically handicapped. IEEE Trans Biomed Eng 35(2):127–131
48. Hobson RS (2000) The changing face of classroom instructional methods: servicelearning and design in a robotics course. In: Frontiers in education conference, vol 2. Kansas City, MO, pp F3C 20–25
49. Horvitz E, Paek T (2001) Harnessing models of users' goals to mediate clarification dialog in spoken language systems. In: Proceedings of the eighth international conference on user modeling, pp 3–13
50. Hsiu T, Richards S, Bhave A, Perez-Bergquist A, Nourbakhsh I (2003) Designing a low-cost, expressive educational robot. In: Proceedings of the conference on intelligent robots and systems, vol 3, pp 2404–2409
51. Hunke M, Waibel A (1994) Face locating and tracking for human–computer interaction. In: Conference record of the conference on signals, systems and computers, Pacific Grove, CA, vol 2, pp 1277–1281
52. Jung J, Do J, Kim Y, Suh K, Kim D, Bien Z (2005) Advanced robotic residence for the elderly/the handicapped : Realization and user evaluation. In: Proceedings of the international conference on rehabilitation robotics, Chicago, Il, pp 492–495
53. Kahn L, Verbuch M, Rymer Z, Reinkensmeyer D (2001) Comparison of robot-assisted reaching to free reaching in promoting recovery from chronic stroke. In: Proceedings of the international conference on rehabilitation robotics. IOS Press, Evry, France, pp 39–44
54. Kahn PH, Ishiguro H, Friedman B, Kanda T (2006) What is a human? – Toward psychological benchmarks in the field of human–robot interaction. In: IEEE Proceedings of the international workshop on robot and human interactive communication (RO-MAN), Hatfield, UK
55. Kanda T, Hirano T, Eaton D, Ishiguro H (2003) Person identification and interaction of social robots by using wireless tags. In: IEEE/RSJ International conference on intelligent robots and systems (IROS2003), Las Vegas, NV, pp 1657–1664
56. Kang K, Freedman S, Matarić MJ, Cunningham M, Lopez B (2005) Hands-off physical therapy assistance robot for cardiac patients. In: Proceedings of the international conference on rehabilitation robotics, Chicago, Il, pp 337–340
57. Kapoor A, Picard RW (2005) Multimodal affect recognition in learning environments. In: Proceedings of the 13th annual ACM international conference on Multimedia, Singapore, pp 677–682
58. Karna-Lin E, Pihlainen-Bednarik K, Sutinen E, Virnes M (2006) Can robots teach? preliminary results on educational robotics in special education. In: Proceedings of the sixth IEEE international conference on advanced learning technologies (ICALT), pp 319–321
59. Kawamura K, Bagchi S, Iskarous M, Bishay M (1995) Intelligent robotic systems in service of the disabled. Proc IEEE Trans Rehabil Eng 3(1):14–21
60. Kidd CD, Breazeal C (2004) Effect of a robot on user perceptions. In: IEEE/RSJ International conference on intelligent robots and systems, Sendai, Japan, pp 3559–3564
61. Kiesler S, Goetz J (2002) Mental models and cooperation with robotic assistants. In: Proceedings of the conference on human factors in computing systems. ACM Press, Minneapolis, Minnesota, USA, pp 576–577
62. Kim K, Bang S, Kim S (2004) Emotion recognition system using short-term monitoring of physiological signals. Med Biol Eng Comput 42(3):419–427
63. Kim Y, Park K, Seo K, Kim C, Lee W, Song W, Do J, Lee J, Kim B, Kim J et al (2003) A report on questionnaire for developing Intelligent Sweet Home for the disabled and the elderly in Korean living conditions. In: Proceedings of the ICORR (the eighth international conference on rehabilitation robotics)
64. Kitwood T, Bredin K (1992) A new approach to the evaluation of dementia care. J Adv Health Nurs Care 1(5):41–60
65. Koegel L, Carter C, Koegel R (2003) Teaching children with autism self-initiations as a pivotal response. Top Lang Disord 23:134–145
66. Koenig N, Matarić MJ (2006) Behavior-based segmentation of demonstrated task. In: International conference on development and learning, Bloomington, IN

67. Koenig N, Matarić MJ (2006) Behavior-based segmentation of demonstrated task. In: International conference on development and learning, Bloomington, IN
68. Koenig N, Matarić MJ (2006) Demonstration-based behavior and task learning. In: Working notes, AAAI spring symposium to boldy go where no human–robot team has gone before, Stanford, California
69. Kopp S, Wachsmuth I (2002) Model-based animation of coverbal gesture. In: Proceedings of the computer animation, IEEE computer society, Washington, DC, USA
70. Kozima H, Nakagawa C, Yasuda Y (2005) Interactive robots for communication-care: a case-study in autism therapy. In: IEEE International workshop on robot and human interactive communication (ROMAN), Nashville, TN, pp 341–346
71. Lang S, Kleinehagenbrock M, Hohenner S, Fritsch J, Fink GA, Sagerer G (2003) Providing the basis for human–robot-interaction: A multi-modal attention system for a mobile robot. In: Proceedings of the international conference on multimodal interfaces. ACM, Vancouver, Canada, pp 28–35
72. Lathan C, Vice J, Tracey M, Plaisant C, Druin A, Edward K, Montemayor J (2001) Therapeutic play with a storytelling robot. In: Conference on human factors in computing systems. ACM Press, New York, NY, USA, pp 27–28
73. Lathan C, Boser K, Safos C, Frentz C, Powers K (2007) Using cosmo's learning system (CLS) with children with autism. In: Proceedings of the international conference on technology-based learning with disabilities, Dayton, OH, pp 37–47
74. Lee C, Narayanan S (2005) Towards detecting emotions in spoken dialogs. IEEE Trans Speech Audio Process 13(2): 293–302
75. Lee KM, Nass C (2003) Designing social presence of social actors in human computer interaction. In: Proceedings of the SIGCHI conference on human factors in computing systems, Ft. Lauderdale, Fl, vol 5, pp 289–296, http://portal.acm.org/citation.cfm?id=642662
76. Lee N, Keating D (1994) Controllers for use by disabled people. Comput Control Eng J 5(3):121–124
77. Lord C, McGee J (eds) (2001) Educating children with autism. National Academy Press, Washington
78. Mahoney R, van der Loos H, Lum P, Burgar C (2003) Robotic stroke therapy assistant. Robotica 21:33–44
79. Matarić MJ, Eriksson J, Feil-Seifer D, Winstein C (2007) Socially assistive robotics for post-stroke rehabilitation. J NeuroEng Rehabil 4(5)
80. Matarić MJ, Koenig N, Feil-Seifer DJ (2007) Materials for enabling hands-on robotics and stem education. In: AAAI Spring symposium on robots and robot venues: resources for AI education, Stanford, CA
81. Matarić MJ, Fasola J, Feil-Seifer DJ (2008) Robotics as a tool for immersive, hands-on freshmen engineering instruction. In: American society for engineering education, Proceedings of the ASEE annual conference & exposition, Pittsburgh, PA
82. Michaud F, Clavet A (2001) Robotoy contest – designing mobile robotic toys for autistic children. In: Proceedings of the american society for engineering education (ASEE), Alberqueque, New Mexico, http://citeseer.nj.nec.com/michaud01robotoy.html
83. Michaud F, Laplante JF, Larouche H, Duquette A, Caron S, Letourneau D, Masson P (2005) Autonomous spherical mobile robot for child-development studies. IEEE Trans Syst Man Cybern 35(4):471–480
84. Mikolajczyk K, Schmid C, Zisserman A (2004) Human Detection Based on a Probabilistic Assembly of Robust Part Detectors. Computer Vision, ECCV 2004: Proceedings of the 8th European conference on computer vision, Prague, Czech Republic, 11-14 May 2004
85. Miller N, Jenkins O, Kallman M, Matarić MJ (2004) Motion capture from inertial sensing for untethered humanoid teleoperation. In: Proceedings, IEEE-RAS International conference on humanoid robotics (Humanoids-2004), Santa Monica, CA
86. Mohan A, Picard R (2004) Health0: a new health and lifestyle management paradigm. Stud Health Technol Inform 108: 43–8
87. Montemerlo M, Prieau J, Thrun S, Varma V (2002) Experiences with a mobile robotics guide for the elderly. In: Procedings of the AAAI national conference on artificial intelligence. Edmunton, Alberta, pp 587–592
88. Mori M (1970) Bukimi no tani (The uncanny valley). Energy 7:33–35
89. Morris A, Donamukkala R, Kapuria A, Steinfeld A, Matthews J, Dunbar-Jacob J, Thrun S (2003) A robotic walker that provides guidance. In: Proceedings of the 2003 IEEE international conference on robotics and automation. ICRA, Taipei, Taiwan, pp 25–30
90. Mower E, Feil-Seifer D, Matarić MJ, Narayanan S (2007) Investigating implicit cues for user state estimation in human robot interaction. In: Proceedings of the international conference on human–robot interaction (HRI)
91. MP Michalowski HK S Sabanovic (2007) A dancing robot for rhythmic social interaction. In: Proceedings of the conference on human–robot interaction (HRI), Washington, DC
92. Mundy P, Card J, Fox N (2000) Fourteen-month cortical activity and different infant joint attention skills. Dev Psychobiol 36:325–338
93. Mutlu B, Krause A, Forlizzi J, Guestrin C, Hodgins J (2007) Robust, low-cost, non-intrusive recognition of seated postures. In: Proceedings of 20th ACM symposium on user interface software and technology, Newport, RI
94. Nagai Y, Hosoda K, Asada M (2003) How does an infant acquire the ability of joint attention?: A constructive approach. In: Proceedings of the third international workshop on epigenetic robotics: modeling cognitive development in robotic systems, Boston, MA, pp 91–98
95. Nicolescu M, Matarić MJ (2003) Linking perception and action in a control architecture for human–robot interaction. In: Hawaii international conference on system sciences, (HICSS-36), Hawaii, USA
96. Nicolescu M, Matarić MJ (2005) Task learning through imitation and human–robot interaction. In: Dautenhahn K, Nehaniv C (eds) Models and mechanisms of imitation and social learning in robots, humans and animals: behavioural, social and communicative dimensions. Cambridge University Press, New York
97. Nordin P (1997) An on-line method to evolve behavior and to control a miniature robot in real time with genetic programming. Adapt Behav 5(2):107–140
98. Oppliger D (2001) University-pre college interaction through FIRST robotics competition. Oslo, Norway, pp 11–16
99. Oztop E, Franklin DW, Chaminade T, Cheng G (2005) Human–humanoid interaction: Is a humanoid robot perceived as a human? Int J Human Robot 2(4):537–559

100. Parise S, Kiesler S, Sproull L, Waters K (1999) Cooperating with life-like interface agents. Comp Human Behav 15(2):123–142
101. Plaisant C, Druin A, Lathan C, Dakhane K, Edwards K, Vice J, Montemayor J (2000) A storytelling robot for pediatric rehabilitation. In: Proceedings of the fourth international ACM conference on assistive technologies, Arlington, VA, pp 50–55
102. Pomerleau D (1993) Knowledge-based training of artificial neural networks for autonomous robot driving. In: Connell JH, Mahadevan S (eds) Robot Learning. Kluwer, Boston, pp 19–43
103. Powers A, Kiesler S (2006) The advisor robot: Tracing people's mental model from a robot's physical attributes. In: Proceedings of the 2006 ACM conference on human–robot interaction. ACM Press, Salt Lake City, UT, pp 218–225
104. Reeves B, Nass C (1996) The media equation: how people treat computers, television, and new media like real people and places. Cambridge University Press, New York
105. Rentschler A, Cooper R, Blasch B, Boninger M (2003) Intelligent walkers for the elderly: Performance and safety testing of VA-PAMAID robotic walker. J Rehabil Res Dev 40(5): 423–431
106. Rizzolatti G, Arbib M (1998) Language within our grasp. Trends Neurosci 21(5):188–194
107. Robins B, Dautenhahn K, Boekhorst R, Billard A (2005) Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills? Univers Access Inf Soc 4(2):105–120
108. Robles EA, Sukumaran A, Rickertsen K, Nass C (2006) Being watched or being special: how i learned to stop worrying and love being monitored, surveilled, and assessed. In: Proceedings of the ACM SIGCHI conference on human factors in computing systems, Montreal, Quebec, Canada, pp 831–839
109. Sakagami Y, Watanabe R, Aoyama C, Matsunaga S, Higaki N, Fujimura K, Ltd H, Saitama J (2002) The intelligent ASIMO: system overview and integration. In: International conference on intelligent robots and system, 2002, EPFL, Switzerland, vol 3, pp 2478–2483
110. Scassellati B (2003) Investigating models of social development using a humanoid robot. Proc Int Joint Conf Neural Networks 4:2704– 2709
111. Scassellati B (2005) Quantitative metrics of social response for autism diagnosis. In: IEEE International workshop on robots and human interactive communication (ROMAN), Nashville, TN, pp 585–590
112. Scassellati B (2005) Using social robots to study abnormal social development. In: Proceedings of the fifth international workshop on epigenetic robotics: modeling cognitive development in robotic systems, Nara, Japan, pp 11–14
113. Scholtz J (2002) Evaluation methods for human–system performance of intelligent systems. In: Proceedings of the 2002 performance metrics for intelligent systems (PerMIS) workshop, Gaithersburg, MD
114. Shalom DB, Mostofsky SH, Hazlett RL, Goldberg MC, Landa RJ, Faran Y, McLeod DR, Hoehn-Saric R (2006) Normal physiological emotions but differences in expression of conscious feelings in children with high-functioning autism. J Autism Dev Disord 36(3):295–400
115. Shin J, Narayanan S, Gerber L, Kazemzadeh A, Byrd D (2002) Analysis of user behavior under error conditions in spoken dialogs. In: Proceedings of ICSLP, Denver, CO
116. Shin MC, Chang KI, Tsap LV (2002) Does colorspace transformation make any difference on skin detection? In: Proceedings of sixth IEEE workshop on applications of computer vision, 2002 (WACV 2002), pp 275–279, http://citeseer.nj.nec.com/542214.html
117. Shneiderman B (1989) A nonanthropomorphic style guide: Overcoming the humpty-dumpty syndrome. Comput Teacher 16(7):5
118. Simpson R, Levine S (1997) Development and evaulation of voice control for a smart wheelchair. In: Proceedings of the rehabilitation engineering society of North America annual conference, Pittsburgh, PA, pp 417–419
119. Sintonen H (1994) The 15-d measure of health related quality of life: reliability, validity and sensitivity of its health state descriptive system. Working Paper 41, Center for Health Program Evaluation, West Heidelberg, Victoria, Australia
120. Sklar E, Eguchi A, Johnson J (2003) RoboCupJunior: learning with educational robotics. Robocup 2002: Robot Soccer World Cup VI
121. Sowa T, Kopp S (2003) A cognitive model for the representation and processing of shape-related gestures. In: Schmalhofer F, Young R, Katz G (eds) Proceedings of the European cognitive science conference (EuroCogSci03), Lawrence Erlbaum Assoc, New Jersey, p 441
122. Stein C (2002) Botball: Autonomous students engineering autonomous robots. In: Proceedings of the ASEE conference, Montreal, Quebec, Canada
123. Stiehl WD, Lieberman J, Breazeal C, Basel L, Lalla L, Wolf M (2006) The design of the huggable: A therapeutic robotic companion for relational, affective touch. In: Proceedings of the AAAI fall symposium on caring machines: AI in eldercare, Washington, DC
124. Stone M, DeCarlo D (2003) Crafting the illusion of meaning: Template-based specification of embodied conversational behavior. In: Proceedings of the international conference on computer animation and social agents, pp 11–16
125. Takeuchi Y, Katagriri Y, Nass CI, Fogg BJ (2000) Social response and cultural dependency in human–computer interaction. In: Proceedings of the CHI 2000 conference, Amsterdam, The Netherlands
126. Tapus A, Matarić MJ (2006) User personality matching with hands-off robot for post-stroke rehabilitation therapy. In: Proceedings of the international symposium on experimental robotics (ISER), Rio de Janeiro, Brazil
127. Tapus A, Fasola J, Matarić MJ (2008) Socially assistive robots for individuals suffering from dementia. In: ACM/IEEE 3rd human–robot interaction international conference, workshop on robotic helpers: user interaction, interfaces and companions in assistive and therapy robotics, Amsterdam, The Netherlands
128. Tartaro A, Cassell J (2006) Authorable virtual peers for autism spectrum disorders. In: Combined workshop on language enabled educational technology and development and evaluation of robust dialog system, ECAI
129. Taub E, Uswatte G, King D, Morris D, Crago J, Chatterjee A (2006) A placebo-controlled trial of constraint-induced movement therapy for upper extremity after stroke. Stroke 37(4):1045–9
130. Thrun S, Bennewitz M, Burgard W, Cremers A, Dellaert F, Fox D, Hahnel D, Rosenberg C, Roy N, Schulte J, Schulz D (1999) MINERVA: A second-generation museum tour-guide

robot. In: Proceedings of the IEEE international conference on robotics and automation (ICRA '99), Detroit, Michigan
131. Topping M, Smith J (1999) The development of handy, a robotic system to assist the severly disabled. In: Proceedings of the international conference on rehabilitation robotics, Stanford, CA, http://rose.iinf.polsl.gliwice.pl/~kwadrat/www.csun.edu/cod/conf2001/proceedings/0211topping.html
132. Turing A (1950) Computing machinery and intelligence. Mind 49:433–460
133. Turkle S (2005) Relational artifacts/children/elders: The complexities of cybercompanions. In: Toward social mechanisms of android science: A CogSci 2005 workshop, Stresa, Italy, p 62–33
134. Uchibe E, Asada M, Hosoda K (1998) Cooperative behavior acquisition in multi-mobile robots environment by reinforcement learning based on state vector estimation. In: Proceedings of the international conference on robotics and automation, Leuven, Belgium, pp 1558–1563
135. Valin J, Michaud F, Rouat J, Letourneau D (2003) Robust sound source localization using a microphone array on a mobile robot. Proceedings of the 2003 IEEE/RSJ international conference on intelligent robots and systems, 2003(IROS 2003), vol 2, pp 1228–1233
136. Wada K, Shibata T, Saito T, Tanie K (2002) Analysis of factors that bring mental effects to elderly people in robot assisted activity. In: Proceedings of the international conference on intelligent robots and systems, Lausanne, Switzerland, vol 2, pp 1152–1157
137. Wada K, Shibata T, Saito T, Sakamoto K, Tanie K (2005) Psychological and social effects of one year robot assisted activity on elderly people at a health service facility for the aged. In: Proceedings of the IEEE international conference on robotics and automation (ICRA), pp 2785–2790
138. Wainer J, Feil-Seifer D, Shell D, Matarić MJ (2006) The role of physical embodiment in human–robot interaction. In: IEEE Proceedings of the international workshop on robot and human interactive communication, Hatfield, United Kingdom, pp 117–122
139. Wainer J, Feil-Seifer D, Shell D, Matarić MJ (2007) Embodiment and human–robot interaction: A task-based perspective. In: Proceedings of the international conference on human–robot interaction
140. Wallach W, Allen C (2005) Android ethics: Bottom-up and top-down approaches for modeling human moral faculties. In: Proceedings of the 2005 CogSci workshop: toward social mechanisms of android science, Stresa, Italy, pp 149–159
141. Wang D, Narayanan S (2007) An acoustic measure for word prominence in spontaneous speech. IEEE Trans Speech Audio Lang Process 15(2):690–701
142. Weng J, McClelland J, Pentland A, Sporns O, Stockman I, Sur M, Thelen E (2001) Autonomous mental development by robots and animals. Science 291(5504):599–600
143. Werry I, Dautenhahn K, Ogden B, Harwin W (2001) Can social interaction skills be taught by a social agent? The role of a robotic mediator in autism therapy. Lecture notes in computer science, vol 2117. Springer, Heidelberg, pp 57–74
144. Wolf S, Thompson P, Morris D, Rose D, Winstein C, Taub E, Giuliani C, Pearson S (2005) The EXCITE trial: Attributes of the wolf motor function test in patients with subacute stroke. Neurorehabil Neural Repair 19:194–205
145. Wood-Dauphinee S (1999) Assessing quality of life in clinical research: From where have we come and where are we going? J Clin Epidemiol 52(4):355–363
146. Woods S, Walters M, Koay KL, Dautenhahn K (2006) Comparing human robot interaction scenarios using live and video based methods: towards a novel methodological approach. In: Proceedings the 9th international workshop on advanced motion control, Istanbul
147. Yanco H (2002) Evaluating the performance of assistive robotic systems. In: Proceedings of the workshop on performance metrics for intelligent systems, Gaithersburg, MD
148. Younger D, Martin G (2000) Dementia care mapping: an approach to quality audit of services for people with dementia in two health districts. J Adv Nurs 32(5):1206–1212

Human Sexual Networks

Fredrik Liljeros
Stockholm University, Stockholm, Sweden

Article Outline

Glossary
Definition of the Subject
Introduction
Non-Complex Models of Contagious Diseases
The Core Group Theory
Lessons Learned from the Early AIDS Epidemic
Clustering
The Effect of Geographical Space
The Long Tail
The Importance of Concurrent Relationship
Assortative Interaction
Data Sources
Future Directions
Bibliography

Glossary

Basic reproduction rate The most common way to calculate the epidemic threshold is to calculate the basic reproduction rate, R_0, which is usually defined as the average number of secondary infections caused by one infectious individual that enters into a totally susceptible population. The basic reproduction rate may underestimate the risk of epidemic outbreaks if the variation in number of contacts is large, as is usually the case with sexual contacts.

Core group A subgroup of individuals in a population characterized by a high partner turnover rate and a high tendency for having sexual contacts within the group. The existence of a core group may push the population above the epidemic threshold.

Epidemic threshold The probability that an epidemic will occur is determined by the contagiousness of the disease, the duration of infectiousness, and the interaction structure in the population. Contagious diseases are nonlinear phenomena in the sense that small changes in any of these parameters may push the population from a state in which a large epidemic is not possible to a state in which an epidemic may easily occur if infection is introduced into the population. The specific point at which an epidemic is possible is referred to as the epidemic threshold.

Random homogeneous mixing When modeling outbreaks of contagious diseases in a population, the individuals are often assumed to have the same probability of interacting with everyone else in the population. This assumption has been shown to be less valid for sexually transmitted infections because they are characterized by a large variation in number of contacts.

Sexually transmitted infection Many contagious infections can be spread through sexual contact. Sexually transmitted infections are, however, generally defined as being spread through vaginal intercourse, anal intercourse, and oral sex. They include *Chlamydia trachomatis*, gonorrhea, and HIV. The reason why the expression "sexually transmitted infection" is used instead of "sexually transmitted disease" is that a state of infection and infectiousness do not necessarily result in disease.

Definition of the Subject

Human sexual networks are the network structures that emerge when individuals have sexual contact with each other. In general, use of the term "sexual contact" is restricted in this article to mean vaginal or anal intercourse or oral sex – contacts by which sexually transmitted infections (STIs) can be transmitted. Sexual networks are important because an understanding of their structure and how they facilitate the spread of infection can help us understand how the spread of this type of infection can best be prevented.

Introduction

Although the type of contact that spreads STIs occurs less frequently than is the case for most other types of contact that spread disease, the spread of STIs has turned out to be surprisingly hard to limit. The difficulties in getting STIs under control have led to an interest in sexual contact patterns [24,28,29]. In the present chapter we discuss a variety of explanations related to the structural properties of sexual networks that have been advanced for why STIs are so widespread, and why such diseases are so hard to eradicate. We begin this chapter by presenting a family of models in which no explicit assumptions are made about the interaction structure other than that all individuals are assumed to have the same probability of interacting with everyone else in the population. These models are then used as a baseline when discussing more realistic assumptions about the sexual contact structure. We then move on to more realistic assumptions about the contact structure by introducing one of the first theories about sexual contact structure, the core-group theory.

The emerging AIDS epidemic in the early 1980s raised new questions about how sexually transmitted diseases are spread in human populations [19]. We discuss different structural explanations for why the number of HIV-infected not did grow as fast as models based on random homogeneous mixing predicted. We then look at two other structural properties of importance for the understanding of the spread of STIs, assortative interaction and concurrent relationship. Sexual networks are usually very difficult to study empirically for several different reasons. These difficulties and ways of handling them are introduced in the succeeding section. Some remarks on the future challenges for research on sexual networks conclude the chapter.

Non-Complex Models of Contagious Diseases

To help to understand the complexity of human sexual networks and how their structures may facilitate the spread of STIs, we will first introduce a simple family of models based on the assumption of random homogeneous mixing. Random homogeneous mixing means that every person in the population has the same probability of interacting with every other person in the population; hence no assumption is made about any structural properties of the contact network. Models based on this assumption often model outbreaks of highly contagious diseases surprisingly well, while they are usually less good for modeling less contagious diseases. These simple models and their characteristic behaviors will then be used as a baseline when we introduce and discuss different structural properties of human sexual networks that deviate from random homogeneous mixing.

In this type of standard model, individuals in a population are assumed to be in one of three states: susceptible (S), infected (I), or removed (R). The latter can, depending on the disease under study, result from immunity or death. It is conventional to distinguish between SI, SIS, and SIR models. Children's diseases are best modeled by an SIR model because infection confers lifelong immunity, that is,

removal from the pool of those susceptible. For most sexually transmitted diseases, the *SIS* model makes most sense since few sexually transmitted diseases confer any immunity after infection, and people remain susceptible. An important exception is HIV, which is still appropriately described, at least in the Western world, using the *SI* model.

The random homogenous mixing assumption is explained by Eq. (1), which represents the *SI* model as a continuous model in its simplest form, consisting of a system of two differential equations:

$$\frac{dS}{dt} = \frac{-c\beta S(t)I(t)}{N}, \quad \frac{dI}{dt} = \frac{c\beta S(t)I(t)}{N}. \tag{1}$$

There are two dependent variables in Eq. (1): The number of susceptible persons, S, and the number of infected persons, I. The number of susceptible and infectious persons are constrained by the size of the population N so that $S(t) + I(t) = N$. Note that this constraint means that the two variables, $S(t)$ and $I(t)$, are linearly dependent, so that the two differential equations actually are redundant. As is evident, this model is homogenous across the population as each person is assumed to have the same number of contacts per time unit, c and the same probability of infection per contact, B. This is what we refer to as the assumption of *random homogeneous mixing*.

For many diseases, such as measles or flu, that are spread by aerosolized droplets by all infected persons, random interaction is a reasonable assumption and probably a good approximation. It describes an abundance of everyday situations in which a person is exposed to such infections, for instance on public transportation, in the workplace, and in shops. A significant advantage of the random interaction assumption is that it can easily be modeled with differential equations, and these models can be studied analytically [1,7]. The equilibriums for the system, for example, can easily be found by first setting the left side of Eq. (1) to 0, and then solving the expression as a system of equations.

The solution to Eq. (1) yields an S-shaped trajectory as shown in Fig. 1. One important property of such models is that they predict that the number of infected persons will grow exponentially during the early stages of an outbreak. This growth cannot, however, continue to accelerate at this rate for long because the number of susceptible persons is low during later stages of the process. In later stages, most infected persons will interact only with other infectious persons. Consequently, the growth in the number of infected persons is largest at about the midpoint of the process.

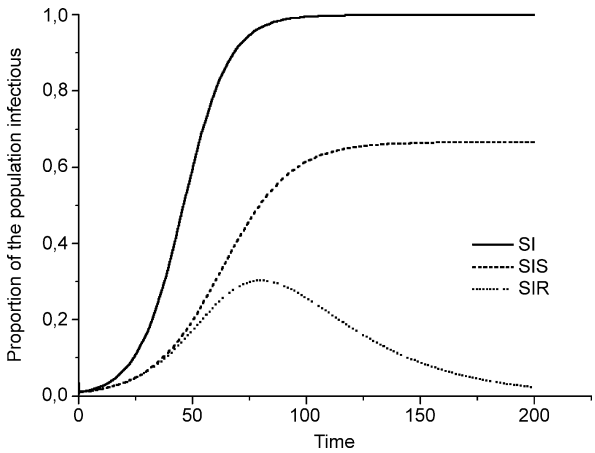

Human Sexual Networks, Figure 1
The typical growth in the number of infectious persons during an outbreak for an *SI*, an *SIS* and an *SIR* model, when random homogeneous mixing is assumed for the population

The *SIS* model can be written as a system of two differential equations as follows:

$$\frac{dS}{dt} = \frac{-c\beta S(t)I(t)}{N} + \frac{I(t)}{D}, \quad \frac{dI}{dt} = \frac{c\beta S(t)I(t)}{N} - \frac{I(t)}{D}. \tag{2}$$

The equation for the *SIS* model differs from that of the *SI* model in the sense that the term $\frac{I(t)}{D}$ that describes the rate of individuals who recover from the disease and become susceptible is added to both equations. The solution to the *SIS* equations also shows that we should expect an S-shaped trajectory with an exponential growth in the number of infected persons. The *SIS* trajectory, however, differs from the *SI* trajectories in the sense that the number of infected persons never reaches that of the entire population. The process equilibrates at a point where exactly as many infectious individuals become susceptible as susceptible ones become infected.

The last model we are going to discuss here is the *SIR* model, which can be formulated as a set of differential equations in its simplest form as:

$$\frac{dS}{dt} = \frac{-c\beta S(t)I(t)}{N}, \quad \frac{dI}{dt} = \frac{c\beta S(t)I(t)}{N} - \frac{I(t)}{D}, \quad \frac{dR}{dt} = \frac{I(t)}{D}. \tag{3}$$

The *SIR* model is portrayed here in the form of a system of three differential equations that describe the change in

the three different states (of which two are independent, since we now have $S(t) + I(t) + R(t) = N$). The solution to the *SIR* model also shows an S-shaped form during the early stages of the epidemic. The *SIR* model differs from the *SI* and the *SIS* models by its tendency to result in zero infectious individuals in the long run.

These three basic models can be adapted to the characteristics of specific diseases by, for example, letting persons be immune only for a certain time interval, making it possible for new individuals to enter into the population through birth and immigration, and allowing persons leave the population because of emigration or death. This makes it possible to generate more complex types of trajectories such as cyclic behavior [1].

A critical notion in disease epidemiology is the basic reproduction number, R_0. In the homogenous deterministic *SIS*-model and *SIR*-model this number tells us how many uninfected persons an infectious individual will, on average, infect in a totally susceptible population [1]:

$$R_0 = c\beta D\,. \tag{4}$$

R_0 has received special attention because in the homogenous model it is quite an intuitive measure of the epidemic threshold. If R_0 is less than 1, then the disease will become extinct. With R_0 equaling exactly 1, we have an unstable equilibrium with no change in the number of infected or susceptible persons, that is, the disease is endemic. A value greater than 1 upsets the replacement conditions, which means that if $R_0 > 1$ the outcome is an epidemic. (Note that R_0 will always be larger than one in the *SI* model since an infectious individual is assumed to be infectious for an infinitely long time.)

The Core Group Theory

Core group theory is an early, and probably also the best known, explanation for why STIs can be endemic despite the fact that the average number of sexual contacts in most national populations is such that we should expect R_0 to be lower than 1, that is, below the epidemic threshold if the contacts were evenly distributed in the population [13]. According to core group theory, the reproduction of STIs can be explained by the existence of several distinct subgroups in the general population that are all characterized by high-risk sexual behavior (high partner turnover rate and unprotected sex) and extensive intergroup interaction. The existence of core groups, according to the theory, makes it possible for STIs to reproduce within the core group because R_0 is greater than 1 in these groups. The core groups thus constitute a reservoir allowing the sexually transmitted infection to remain endemic in a general population in which the R_0 is lower than the critical value 1. The core groups also make it possible for the rest of the population to be infected through contacts with these groups.

Lessons Learned from the Early AIDS Epidemic

The discovery of a progressive outbreak of AIDS in the early 1980s gave rise to intensive research efforts to understand the path of contagion and the disease's dynamic course [8]. One of the most important findings was made by Anderson and May [1] who showed that the expression for R_0 in Eq. (4) is not suitable if the variance in number of potential infectious contacts is high, as is the case with STIs. They showed instead that R_0 would be more accurately estimated by using the following equation:

$$R_0 = \rho_0 \left(1 + \frac{\sigma^2}{\mu^2}\right)\,, \tag{5}$$

where ρ_0 is the average number of infections produced by an infected person in an uninfected population, σ^2 is the variance in the number of contacts, and μ is the mean number of contacts in the population. From this equation, it is clear that the larger the variance in number of partners in the population for a given μ, the less infectious an infection needs to be to continue to reproduce itself, that is, to generate an epidemic.

It became clear relatively early that the AIDS epidemic did not behave the way traditional infection epidemiological models expected. The assumption of random homogeneous mixing makes these models predict an exponential increase in the number of contagious persons at the beginning of an epidemic (see Fig. 1). The AIDS epidemic did not seem to follow this pattern, however. Instead of increasing exponentially, it seemed to increase more slowly [5]. This was seen to occur in spite of the fact that the epidemic was very far from global saturation, so that the slowing down could not be due to a global saturation effect. Exponential growth is characterized by a constant time required for the number of infected persons to double. However, in the case of AIDS, that time became increasingly longer as the epidemic spread. At the end of the 1980s, the AIDS epidemic was demonstrated to exhibit a polynomial pattern of spread, and it was shown that epidemics have had this functional form before any changes in behavior patterns could have had an effect. Then Colgate et al. [5] argued that a polynomial pattern of spread could not be explained as a direct effect of interventions or by changed behavior patterns caused by a general understanding of how HIV was spread. They also showed that polynomial spread could be observed in groups of indi-

viduals of different ethnic backgrounds. There are at least three structural properties of contact structures that may slow down spread in the observed way: clustering, embedding in a low-dimensional space, and a large variation in partner turnover rate.

Clustering

One structural property of contact networks that has been shown to slow an epidemic's rate of spread is *clustering* (or *transitivity* as it is also called). A large amount of clustering is typical of many social networks. For example, if Charles is a good friend of Paul and Ben's, it is quite probable that Paul and Ben know each other. In an outbreak of a highly contagious disease, the contacts of an infected individual in a clustered network will often already have been infected by common contact or by a contact only a few steps away in the network. A common way of estimating clustering in a network is to estimate its relative number of triangles, or more exactly, to calculate the fraction C of all paths of length three in the network that form a triangle:

$$C = \frac{3n_{\text{triangle}}}{n_{\text{triple}}}, \tag{6}$$

where n_{triangle} is the number of triangles and n_{triple} is the number of triples of vertices connected by two or three contacts. The factor three is needed to normalize C to the interval $[0, 1]$. One useful property of C is that $1 - C$ gives the average proportion of outgoing links from all contacts directly connected to an individual that potentially can transmit the disease further to the rest of the network.

Clustering, that is, two individuals with a common sexual contact who also have sexual relations with each other by definition cannot exist in a heterosexual network unless bisexual relationships are allowed. It is possible, however, for two individuals of the same sex to have had two or more sexual partners in common. This phenomenon is called *mutuality*, M. For computational reasons, M is often defined in a way similar to $1 - C$ as

$$M = \frac{\text{mean number of nodes two steps away from a node}}{\text{mean number of paths of length two between those nodes}}, \tag{7}$$

where a node here would be a person, and a path would be a route along any number of links from one node to another. Mutuality can thus be thought of as the opposite of clustering (transitivity). For example, M for a given node A reaches its maximum value Eq. (1) when A reaches as many two-hop neighbors as possible with the number of two-hop paths coming from A. Thus, high mutuality leads to faster epidemic spreading, while high transitivity tends to confine epidemic spreading.

Recently, Balázs Szendroi and Gábor Csányi [45] have proposed that the polynomial pattern of spreading could be explained by the fact that sexual networks should show a high degree of transitivity. This may hold true for a population of gay men; it is not clear, however, that this thesis would be valid for heterosexual sexual contacts. In heterosexual networks, there are no tricycles per definition, so in these parts of the sexual network, an epidemic course must be slowed down by four cycles. One study of a romantic and sexual network presents a result that indicates the existence of a norm against changing each other's partner, which may decrease the effect of local clustering even more [3].

The Effect of Geographical Space

Contagious diseases that are spread by a contact network that is embedded in a two-dimensional space where only local interaction takes place will also have a polynomial pattern of spread. This property can easily be understood by the following thought experiment. Assume that we have a large population distributed on a square lattice so that there is one individual on each vertex of the lattice. If we then infect an individual in the center of the square lattice with a chronic disease that is spread to all neighbors every 24 h, it is easy to understand that the cumulative number of infected persons after t days and nights will be $(2 \cdot t + 1)^2$. That is to say, it will have a polynomial pattern of spread. It is not likely, however, that sexual contacts are sufficiently local for a polynomial pattern of spread to be observed. This is because strong evidence indicates that enough contacts extend over social and geographical distances, usually referred to as *social and spatial bridges* [47], for the sexual contact network to exhibit a so called *small world* quality [46]. That is to say, the average distance increases logarithmically with the network's size in the same way a random network does. Figure 2, for example, shows how different parts of the region Värmland in Sweden are sexually connected to each other and to the rest of Sweden by sexual contacts of individuals that tested positive for chlamydia in Värmland [36].

The Long Tail

According to Colgate et al. [5], the polynomial pattern of spread could be explained by a great variation in behavior that involves the risk of being infected by HIV.

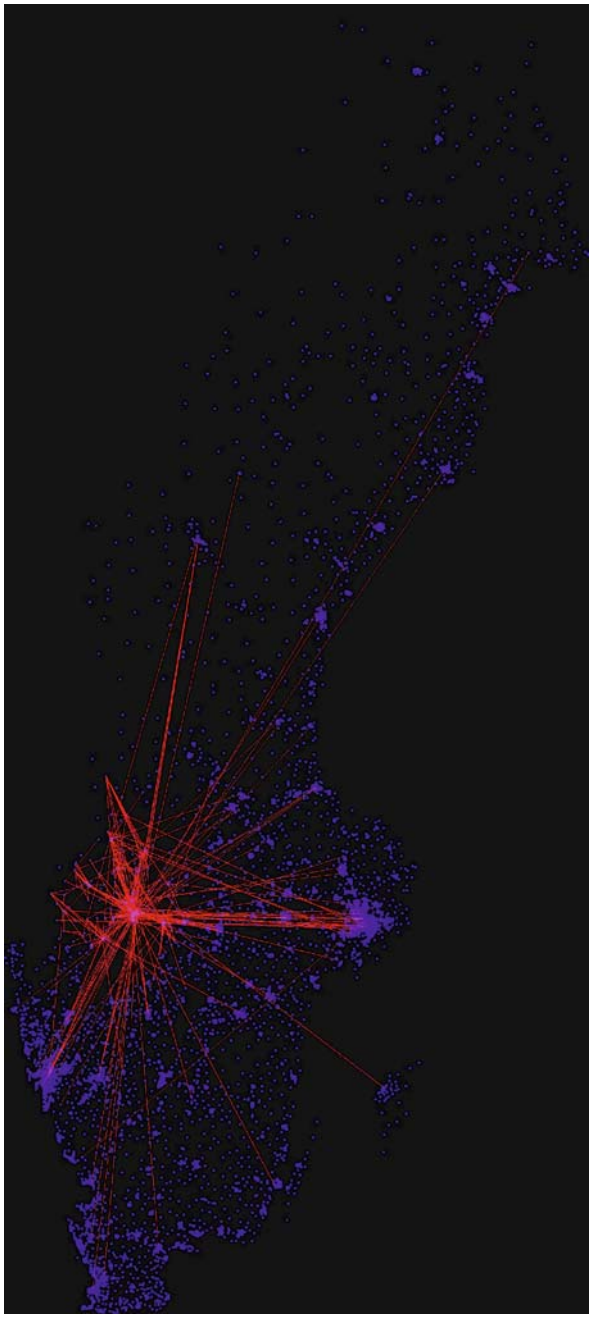

Human Sexual Networks, Figure 2
The spatial distribution of the sexual contacts outside of Värmland County (contacts outside Sweden not shown) [36]

They demonstrated that if a behavior involving risk follows a power-law function, that is to say, the largest group indulges in a relatively safe kind of behavior, the next largest is a little riskier, etc., this would be enough, together with a tendency to have sex with individuals exhibiting similar risk behavior, to generate a polynomial pattern of spread. First, the small high-risk group will be quickly infected to the point of local saturation. The disease will then spread gradually to groups with lower and lower risk levels. In these groups, the infection will spread increasingly slowly as the epidemic continues. What is interesting is that Colgate et al. presented empirical evidence that at an STD clinic the distribution of the number of sexual contacts in the risk category of homosexual men followed a power-law with exponent −3.

There are many examples that show the error of trying to generalize a certain behavior pattern from a group of individuals in a community to the whole community. To do so, it is necessary for these individuals to have been chosen at random, independently of each other. The first study indicating that the frequency distribution for the number of sexual contacts of an entire nation is very close to a power-law tail was reported by Liljeros et al. [23], using the results of a study of the sexual habits of Swedish citizens, "Sex in Sweden" [22]. It was shown that the number of sexual contacts for men and women did fit a power law for the upper tail of reported contacts during the past twelve months and reported contacts over the lifetime. The data for the Sex in Sweden study were gathered in 1996.

The method used to estimate tail inclination was criticized after the results of the Swedish study were published [11,17,18,24]. The study was conducted on a rather small dataset but had the additional strength of showing a power-law tail for both partners during the past 12 months and for the number of partners over a lifetime. Recently, however, Schneeberger et al. [43] reported a similar power-law-looking distributions on a dataset that can be described as the "Rolls Royce" of national sexual studies that has been carried out so far, the NATSAL 2000 [16]. The NATSAL 2000 was a study of 10 000 individuals in Great Britain who answered a survey at their home on a laptop that was brought to them by a research assistant. The research assistant did not sit in the same room while the respondent answered the questions, and the respondent was told that those answers were encrypted so that the research assistant would be unable to read them. Schneeberger et al. [43] reported the slope of the power law at 2.5 for heterosexual men, 3.1 for heterosexual women, 3.3 for homosexual women, and 1.6 for homosexual men. There will probably always be questions about human sexual behavior that cannot be answered due to lack of empirical data, such as the location of the upper cutoff of the distribution that by necessity must exist due to space and time constraints. Interestingly, a recent study of flirting on an Internet dating community reported a similar power-law-

like tail in the number of contacts that people had with each other on the site [15]. This data source is, however associated with large uncertainty when it comes to mapping the actual sexual network, because no information exists about actual sexual contact.

A power-law distribution in the number of sexual contacts has, under some specific circumstances, recently been shown to have serious consequences for the potential for eradicating sexually transmitted infections [37,38]. In an infinite population with homogeneous mixing, a slope smaller than three makes the second moment of the distribution infinitely large, and therefore also the variance of the distribution [25]. It can easily be seen from Eq. (5) that this will result in an infinitely large R_0 for the population. This has the bizarre consequence that all individuals must be tested and treated at the same time, otherwise it will be impossible to eradicate the disease. There are, however, several reasons to believe that this is not the case for sexual networks. We know, for example, that every human population must by definition have a finite size. Even though it has been shown that a power-law distribution in number of contacts will increase R_0 significantly in a finite population (as compared to a population with the same mean number of contacts but a low variance in number of contacts – see again Eq. (5), and [39], this does not imply that an epidemic outbreak of an STI cannot be stopped, or at least curtailed. Another factor that probably mitigates the effect of the skewed distribution is that there must also be an upper limit for how many sexual contacts a single individual can have per unit of time. A study of prostitutes in the United States shows, however, that this limit can be very large. The median value for number of partners was found to be as large as 103 during the previous six months [4]. One advantage of the skewed distribution, from an STI prevention perspective, is that it is predicted that R_0 can be drastically reduced and the epidemic eventually stopped if the individuals who change partners frequently can be tested and convinced to practice safe sex [6]. It may at first glance seem difficult to identify, for a specific targeted intervention, individuals who change partners frequently, except perhaps for specific groups such as prostitutes, or gay men who visit video clubs where anonymous sex takes place. Contact tracing, however (which is discussed in more detail in Sect. "Data Sources"), has the positive side effect that individuals who have many contacts have a higher probability of showing up – because they have a larger group of sources of infection – than do individuals with fewer contacts.

There are several ways to generate networks with a distribution of contacts similar to the one observed in sexual networks. The model so far given the most attention is

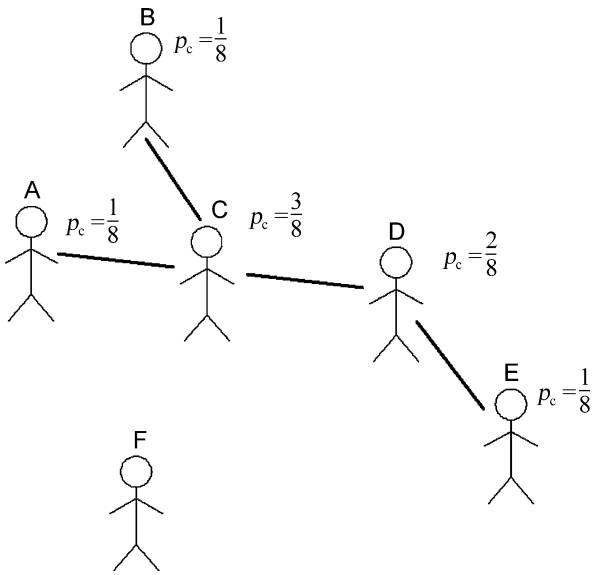

Human Sexual Networks, Figure 3
A snapshot at an early stage of a BA model for generating a network with a degree distribution power-law tail. The probability p_c that an already connected individual (i. e., one of A–E) will connect to the new individual F is proportional to the connected individual's number of contacts

one proposed by Albert-László Barabási and Réka Albert (the BA model) [2]. This model is based on an idea that can be traced back to the work of Herbert Simon [44] and J.D. Price [41] and works in the following way: start with a small number of vertices, and continuously add new vertices. Let the new vertices connect to one or several of the already existing vertices. Do this with a probability that is linear in proportion with the number of contacts these already existing vertices have.

Figure 3 shows a snapshot of the BA model, where the new vertex F is about to connect to the network of already connected vertices. The probability of F connecting to vertex D at this stage is twice the probability of F connecting to vertex E; and F's probability of connecting to C is three times the probability of F connecting to vertex E. The BA model generates a network, whose degree distribution could be described by a Yule distribution where $\alpha = 3$.

$$p(k|\alpha) = \frac{(\alpha - 1)\Gamma(k)\Gamma(\alpha)}{\Gamma(k + \alpha)} . \tag{8}$$

The BA model was originally proposed as a model for how the World Wide Web grows over time, that is to say, how a new homepage links to already existing homepages. As other types of networks have started to be analyzed, several modified BA models have been proposed that aim at finding special properties in these networks. The original

BA model assumes that new connections are made only between old and new vertices, which is not correct for sexual networks. In other words, it does not permit the updating of new links to already existing pages. Réka Albert and Albert-László Barabási [2] have shown that it is possible to generate power-law distributions leaning more steeply than 3 if the formation and resolving of links between already existing nodes is also permitted.

Surveys of sexual behavior are usually both too small and contain too little information about how the individual's number of partners grows over time to use standard methods to measure preferential attachment. One exception is an MLE-based expectation–maximization fitting technique developed especially for estimating preferential attachment in sexual survey data [10]. This method has demonstrated a significant effect of sublinear preferential attachment in partner growth over time on Norwegian survey data.

Theoretically, it is possible to argue for at least three different mechanisms that may cause preferential attachment. Like most other social behavior, it ought to be possible to get better at flirting and picking up through practice. An individual may also, under some circumstances, be considered to be more attractive the more partners he/she has had. Finally, we know that getting a new partner, like any other initially pleasurable behavior, can be psychologically addictive. To date we have no empirical results verifying the extent to which each of these different preferential attachment mechanisms are operative.

The Importance of Concurrent Relationship

The number of unprotected sexual contacts is clearly associated with an individual's risk for both being infected by a sexually transmitted infection, and for passing on the infection once infected. It is, however, possible that the number of sexual contacts is not the most important risk factor for getting an STI per se. Morris and Kretzschmar suggested in a series of articles [20,30,31] that it is the frequency of concurrent relations (partnerings that overlap over short time periods) in a population that is the most important factor in the transmission of sexually transmitted infections. Concurrent relations are important because potential contacts for the transmission of an STI come much closer in time if individuals have concurrent relations than if they practice serial monogamy.

This approach gives rise to a special type of graph called a line graph [12] in which the contacts between the persons are seen as nodes in a network. When we let a contact between two persons define a node on the graph, an edge is present whenever a person has more than one contact. Two graphical examples of sexual contacts and their corresponding line graphs are shown in Fig. 4. Both contact networks displayed in Fig. 4a and b have the same average degree. Despite this, it is much easier for an STI to propagate in the left network than in the right one.

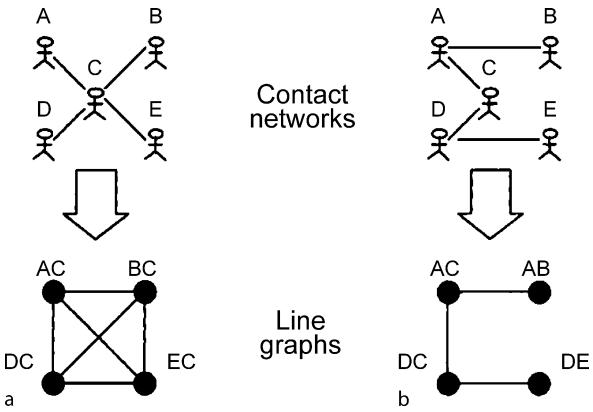

Human Sexual Networks, Figure 4

Two contact networks and corresponding line graphs following Morris and Kretzschmar [30]. In the line graph, every edge in the contact network is translated into a node; for example, edge A–C in a becomes the node AC. Nodes with a degree > 1 in the contact network will contribute to new edges in the line graph. For example, in b there is an edge between D and C and D and E in the contact network (C has a degree of 3), thus there will be an edge between DC and DE in the line graph

This is due to the fact that the level of concurrent sexual relations is much higher in network Fig. 4a than in network Fig. 4b, as can be seen by considering the corresponding line graphs. A measure for assessing the level of concurrency in a line graph has been suggested by Morris and Kretzschmar [20]. The concurrency, κ_2, is given by the following equation:

$$\kappa_2 = L_2 \left(\frac{N_2(N_2-1)}{2} \right)^{-1} = \begin{cases} 1 & \text{all pairs adjacent} \\ 0 < \kappa_2 < 1 & \text{some pairs adjacent} \\ 0 & \text{no pairs adjacent (monogamy)} \end{cases} . \tag{9}$$

Here L_2 is the number of links and N_2 is the number of nodes in the line graph (i. e., κ_2 is the density of the line graph). By further calculating the mean number of concurrent relationships per relationship, the index of concurrency κ_3, it has been demonstrated that concurrency is a function of the mean and the standard deviation of the degree distribution [20]. Since these properties can be calculated solely on the basis of ego-network data (i. e., lo-

cal information), it is possible to estimate concurrency by using random samples of the population [20,30], which is very difficult for several other network measures such as density and component size [9].

The idea of using line graphs [20,30] has been developed even further by taking into consideration that an STI can propagate between two non-concurrent sexual relationships that occur relatively closely in time [42]. To handle this possibility in defining the line graph, it has been suggested that a sexual relation should be viewed as active for some period after the sexual partnership ended, depending on the type of disease.

Assortative Interaction

A tendency for individuals to prefer sexual contact with persons similar to themselves is usually referred to as *assortative interaction* in the STI literature. A tendency toward assortative interaction has been reported for social factors such as social class and ethnicity [24]. An important property of most sexually networks is that they are assortative by number of contacts [32,33,34]. This means that individuals who have many contacts tend to have contact with other individuals who also have many contacts. High assortativity decreases the epidemic threshold because a large interconnected component will emerge at a lower average density. The standard measure of assortativity is the assortative mixing coefficient r; that is, Pearson's correlation coefficient between the individuals' degrees on each side of the edges [32,33]. As the edges are undirected, we need a coefficient that is invariant to edge reversal. This can be obtained for an edge (i, j) by including both (k_i, k_j) and (k_j, k_i) in the correlation coefficient, which can be expressed mathematically as:

$$r = \frac{4\langle k_1 k_2\rangle - \langle k_1 + k_2\rangle^2}{2\langle k_1^2 + k_2^2\rangle - \langle k_1 + k_2\rangle^2}\,, \tag{10}$$

where k_1 (k_2) is the degree of the first (and second) argument as it appears in the edge list.

Data Sources

The study of sexual networks can to some extent be compared to the study of planets in other solar systems in the sense that that they can only be studied indirectly. Our knowledge comes from at least four different sources, each of which has its specific advantages and disadvantages. The first consists of national surveys of sexual behavior [16,21,22]. The advantage of such studies is that, given that they are based on random population samples, they are the only kind of study that can theoretically yield knowledge about sexual behavior for the whole population. Unfortunately, such studies have several drawbacks. For one thing, it is very hard to control for the accuracy of respondents' answers. Also, the response rate is also usually too low to guarantee reliable precision in the estimates.

One striking thing about national surveys, for example, is that men on average report a significantly higher number of sexual partners than do women. This discrepancy has been explained by a tendency for men as a group to over-report the number of sexual partners [27]. A recent study has also shown that this difference may be explained by the fact that prostitutes are usually not included in the samples [4]. National surveys have another disadvantage that is probably more important than the validity problem mentioned above, namely that they are only able to give us information about the behavior of the respondents, and not information about the behavior of their sexual partners (and *their* partners). National surveys cannot therefore give us information about the global properties of a network, such as level of clustering, size of the largest interconnected component, or average distance between the individuals.

A second source of information about sexual networks is the network data generated by *contact tracing* [47], which is the process whereby the contacts of an individual who has tested positive for an STI are traced and tested. If this procedure is also continued for the contacts that tested positive, and for their contacts in turn, it is eventually possible to generate a subgraph of a sexual network. This subnetwork can then be used to analyze global structural properties of the network that cannot be studied with national survey data. This kind of sampling is also associated with severe biases. It is not always possible for an individual, even if s/he is cooperative, to give enough information about any given sexual partner to be able to identify him or her. A more serious bias is that, by definition, contact tracing has a tendency to identify the subnetworks of the general sexual networks in which it is easiest for the disease to spread. The latter may not, however, always be a problem. If, for example, the purpose of a study is to find ways to mitigate the epidemic in the parts of the network in which the STIs are spread, this type of biased data really can be very useful.

A third type of data source is the mapping of the sexual network. This has so far only been carried out at the local level, probably for practical reasons, for example with a high risk group for HIV of drug users and prostitutes in a town in the United States (see Fig. 5) [40].

Another example is the study of a network of romantic relationships in a high school in the United States [3].

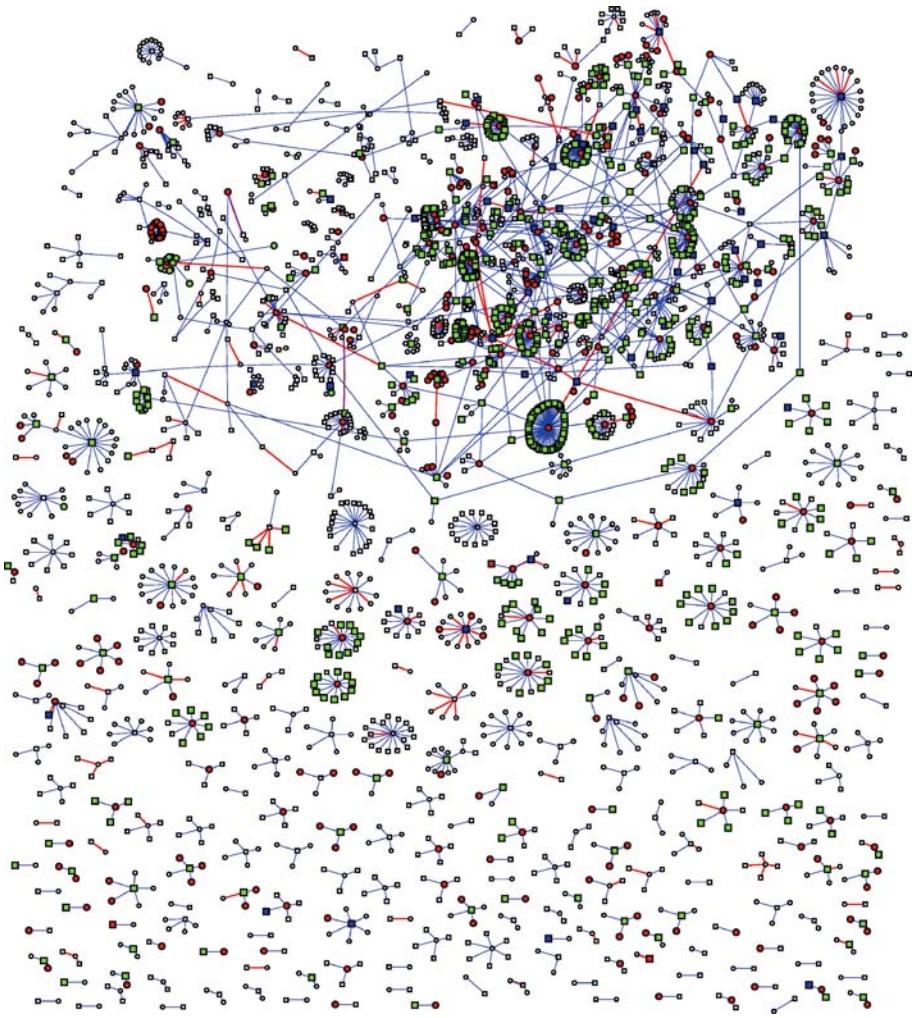

Human Sexual Networks, Figure 5
A sexual network of a high risk group for HIV infection in a town in the US. (Line colorings are *blue* for vaginal sexual contact, *red* for anal sexual contact. Node colorings are *red* for prostitutes, *blue* for pimps, and *green* for sex-buyers) [40]

These studies have the same problem as contact tracing studies when it comes to identifying and finding partners. It is also important to remember that these two studies cannot straightforwardly be generalized to the rest of the population, as studies based on a random sample can.

The fourth and last data source is data generated from Internet dating communities [15]. Because much of the activity that takes place in such communities is logged, it is possible to extract a network about which members interact. The advantage of this data source is that all contacts in the community can be traced, and that networks of a significant size can be generated. This data source is, however, probably the least valid network discussed here when it comes to mapping the actual sexual network, because no information exists about actual sexual contact.

The classification of data sources into the four groups cited here should be seen as a preliminary one. There are also combinations of such as surveys based on non-random samples, for example, individuals who are seen at STI clinics [5] or convenient samples of university students. This short survey of data sources for sexual networks shows that all data sources are associated with different kinds of validity problems. This is quite a common situation in the social sciences. One way to mitigate, if not to solve, the problem is to try to confirm an empirical result with data from different sources, a procedure usually referred to as triangulation.

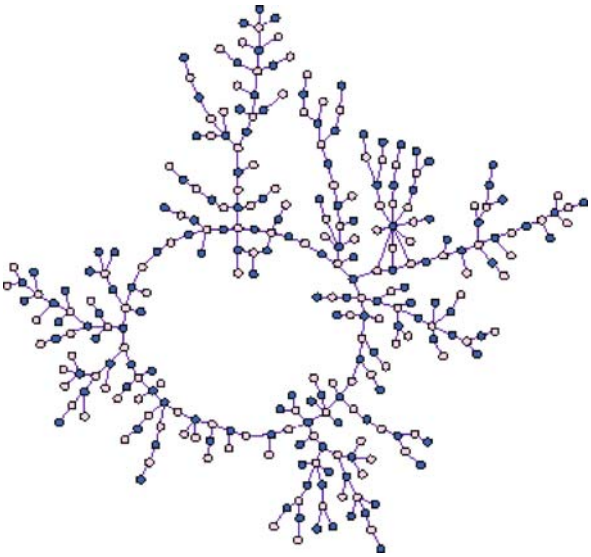

Human Sexual Networks, Figure 6
The largest component in network of romantic relationships in a high school in the United States (romantic relationships with non-students not included) [3]

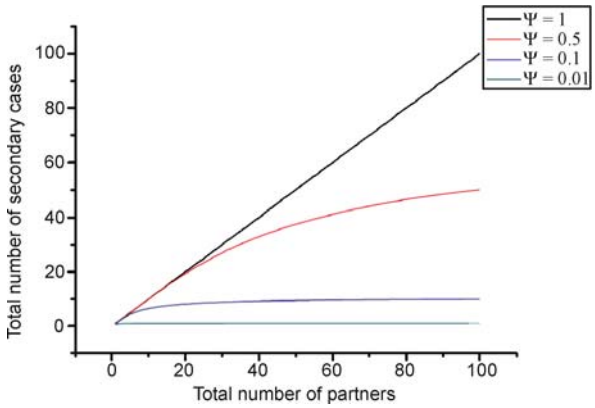

Human Sexual Networks, Figure 7
Total number of secondary cases per total number of partners, when each individual has 100 sexual intercourses evenly distributed among his or her partners (Ψ = probability of transmission per act of sexual intercourse) [35]

Future Directions

To date, the risk of becoming infected with an STI and the risk of passing on the STI have been assumed to grow with partner turnover rate. A recent analysis of a small Swedish survey which contained unusually detailed information on sexual behavior has, to some extent, put this assumption into question [35]. The study suggested that the increased risk of high partner turnover rate may be compensated for by the fact that individuals with a high partner turnover rate seem to have a tendency to have fewer sexual intercourses with each partner, and therefore a lower risk for STI transmission per partner. The effect of this tendency seems to be especially important for STIs with a low risk of transmission per intercourse as shown in Fig. 7.

The high incidence of the most common STI, *Chlamydia trachomatis*, that has been reported in most Western societies may look like an anomaly in light of these results. Recent work based on a simulation model suggests, however, that a small set of individuals who report a large number of partners in sexual surveys are not necessarily as indispensable for transmitting STIs as generally thought [29]. The authors argue instead that multiconnected components, that is, network components such as network cycles in which each individual reaches each other individual in mutually exclusive ways, may be of greater importance for the spread of STIs. Interestingly, their preliminary simulation results show that large bicomponents, that is, components in which each individual can reach each other individual in two different, mutually exclusive ways, may emerge in networks that have a low variance and a relatively low mean number of contacts.

There is a tendency in most network research to simplify dynamic networks to their static equivalents, where the links that exist during a more or less long period are assumed to be concurrent. This also holds true of research about sexual networks even though the formation of and resolution of sexual relationship is to some extent taken into consideration in the notion of concurrent relations. Concurrency cannot, however, capture the fact that A can infect C, but not vice versa, in Fig. 8. One exception to this is Moody's pioneering "The importance of relationship timing for diffusion" [26] in which he presents a formalized framework for describing dynamic networks.

The results presented here indicate that many different structural properties must be taken into consideration to understand the spread of STIs. It is therefore not likely that a single solution to the problem will ever be found. The solution probably lies in a combination of broad and targeted interventions. It is also important to remember that efforts to control STIs in many Western societies have already drastically decreased their incidence, especially in the case of gonorrhea (even though the disease has not been totally eradicated). The fact that an increase in the incidence of STIs has been reported in some Western countries is a warning that must be taken seriously [14]. It is also important to remember that STIs are now a global problem, and as long as travel patterns between countries

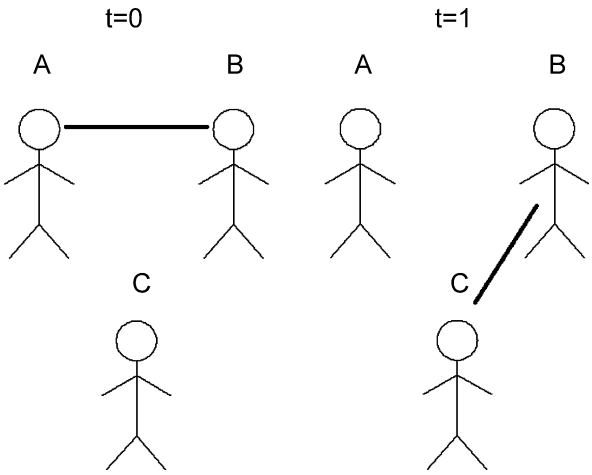

Human Sexual Networks, Figure 8
An example of how the order of relationship may be of importance for the transmission of a sexually transmitted disease. A may indirectly infect C but not vice versa

and continents persist, STIs will never be eradicated in one part of the world as long as STIs are still endemic in the rest of the world.

Bibliography

1. Anderson R, May RM (1991) Infectious diseases of humans. Oxford University Press, Oxford
2. Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512
3. Bearman PS, Moody J et al (2004) Chains of affection: The structure of adolescent romantic and sexual networks. Am J Sociol 110(1):44–91
4. Brewer DD, Potterat JJ, Garrett SB, Muth SQ, Roberts JM, Kazprzyk D, Montano DE, Darrow WW (2000) Prostitution and the sex discrepancy in reported number of sexual partners. Proc Natl Acad Sci USA 97:12385–12388
5. Colgate SA, Stanley EA et al (1989) Risk behavior-based model of the cubic growth of acquired immunodeficiency syndrome in the United States. Proc Nat Acad Sci US 86(12):4793–4797
6. Dezso Z, Barabasi AL (2002) Halting viruses in scale-free networks. Phys Rev E Stat Nonlin Soft Matter Phys 65(5 Pt 2): 055103
7. Diekman O, Heesterbeek JAP (2000) Mathematical epidemiology of infectious disease. John Wiley and Son, Chichester
8. Foulkes MA (1998) Advances in HIV/AIDS statistical methodology over the past decade. Stat Med 17(1):1–25
9. Frank O (1971) Statistical inference in graphs. FOA, Stockholm
10. Freiesleben de Blasio B, Svensson B et al (2007) Preferential attachement in sexual networks. PNAS 104(26):10762–10767
11. Handcock MS, Jones JH (2004) Likelihood-based inference for stochastic models of sexual network formation. Theor Popul Biol 65(4):413–22
12. Harary F (1969) Graph theory. Addison-Wesley, Reading
13. Hethcote H, Yorke JA (1984) Gonorrhea transmission dynamics and control. Springer, New York
14. Hiltunen-Back E, Haikala O et al (2003) Nationwide increase of Chlamydia trachomatis infection in Finland – Highest rise among adolescent women and men. Sex Transm Dis 30(10): 737–741
15. Holme P, Edling CR et al (2004) Structure and time evolution of an Internet dating community. Social Netw 26(2):155–174
16. Johnson AM, Mercer CH et al (2001) Sexual behaviour in Britain: partnerships, practices, and HIV risk behaviours. Lancet 358(9296):1835–1842
17. Jones JH, Handcock MS (2003) An assessment of preferential attachment as a mechanism for human sexual network formation. Proc Biol Sci 270(1520):1123–1128
18. Jones JH, Handcock MS (2003) Social networks: Sexual contacts and epidemic thresholds. Nature 423(6940):605–606; discussion 606
19. Klovdahl AS (1985) Social networks and the spread of infectious diseases: the AIDS example. Soc Sci Med 21(11):1203–1216
20. Kretzschmar M, Morris M (1996) Measures of concurrency in networks and the spread of infectious disease. Math Biosci 133(2):165–195
21. Laumann EO, Gagnon JH et al (1994) The social organization of sexuality. University of Chicago Press, Chicago
22. Lewin B (ed) (2000) Sex in Sweden. The Swedish National Institute of Public Health, Stockholm
23. Liljeros F, Edling CR et al (2001) The web of human sexual contacts. Nature 411(6840):907–908
24. Liljeros F, Edling CR et al (2003) Sexual networks: implications for the transmission of sexually transmitted infections. Microbes Infect 5(2):189–196
25. Lloyd AL, May RM (2001) Epidemiology. How viruses spread among computers and people. Science 292(5520): 1316–1317
26. Moody J (2002) The importance of relationship timing for diffusion. Social Forces 81(1):25–56
27. Morris M (1993) Telling tails explain the discrepancy in sexual partner reports. Nature 365(6445):437–440
28. Morris M (ed) (2004) Network epidemiology: A handbook for survey design and data collection. Oxford University Press Inc, New York
29. Morris M, Goodreau S et al (2007) Sexual networks, concurrency, and STD/HIV. In: Holmes KK, Sparling PF, Stamm WE (eds) Sexually transmitted diseases. McGraw-Hill, New York
30. Morris M, Kretzschmar M (1995) Concurrent partnerships and transmission dynamics in networks. Social Netw 17(3–4): 299–318
31. Morris M, Kretzschmar M (1997) Concurrent partnerships and the spread of HIV. Aids 11(5):641–648
32. Newman MEJ (2002) Assortative mixing in networks. Phys Rev Lett 89(20):1–4
33. Newman MEJ (2003) Mixing patterns in networks. Phys Rev E 67(2):1–13
34. Newman MEJ (2003) Properties of highly clustered networks. Phys Rev E 68(2):026121
35. Nordvik MK, Liljeros F (2006) Number of sexual encounters involving intercourse and the transmission of sexually transmitted infections. Sex Transm Dis 33(6):342–349
36. Nordvik MK, Liljeros F et al (2007) Spatial bridges and the spread of Chlamydia: the case of a county in Sweden. Sex Transm Dis 34(1):47–53
37. Pastor-Satorras R, Vespignani A (2001) Epidemic dynamics and

endemic states in complex networks. Phys Rev E 63(066117): 1–8
38. Pastor-Satorras R, Vespignani A (2001) Epidemic spreading in scale-free networks. Phys Rev Lett 86:3200–3203
39. Pastor-Satorras R, Vespignani A (2002) Epidemic dynamics in finite size scale-free networks. Phys Rev E Stat Nonlin Soft Matter Phys 65(3 Pt 2A): 035108
40. Potterat JJ, Woodhouse DE et al (2004) Network dynamism: history and lessons of the Colorado Springs study. In: Morris M (ed) Network epidemiology: A Handbook for survey d esign and data collection. Oxford University Press Inc, New York, pp 87–114
41. Price DJ (1976) A general theory of bibliometric and other cumulative advantage processes. J Am. Soc. Inform. Sci 27: 292–306
42. Riolo CS, Koopman JS et al (2001) Methods and measures for the description of epidemiologic contact networks. J Urban Health 78(3):446–457
43. Schneeberger A, Mercer CH et al (2004) Scale-free networks and sexually transmitted diseases: a description of observed patterns of sexual contacts in Britain and Zimbabwe. Sex Transm Dis 31(6):380–387
44. Simon HA (1955) On a class of skew distribution functions. Biometrika 42:425–440
45. Szendroi B, Csányi G (2004) Polynomial epidemics and clustering in contact networks. Proc Biol Sci Aug 7:271 Suppl 5:S364-6
46. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440–442
47. Wylie JL, Jolly A (2001) Patterns of chlamydia and gonorrhea infection in sexual networks in Manitoba, Canada. Sex Transm Dis 28(1):14–24

Hybrid Control Systems

Andrew R. Teel[1], Ricardo G. Sanfelice[2], Rafal Goebel[3]
[1] Electrical and Computer Engineering Department, University of California, Santa Barbara, USA
[2] Department of Aerospace and Mechanical Engineering, University of Arizona, Tucson, USA
[3] Department of Mathematics and Statistics, Loyola University, Chicago, USA

Article Outline

Glossary
Notation
Definition of the Subject
Introduction
Well-posed Hybrid Dynamical Systems
Modeling Hybrid Control Systems
Stability Theory
Design Tools
Applications
Discussion and Final Remarks
Future Directions
Bibliography

Glossary

Global asymptotic stability The typical closed-loop objective of a hybrid controller. Often, the hybrid controller achieves global asymptotic stability of a compact set rather than of a point. This is the property that solutions starting near the set remain near the set for all time and all solutions tend toward the set asymptotically. This property is robust, in a practical sense, for well-posed hybrid dynamical systems.

(Well-posed) Hybrid dynamical system System that combines behaviors typical of continuous-time and discrete-time dynamical systems, that is, combines both flows and jumps. The system is said to be well-posed if the data used to describe the evolution (consisting of a flow map, flow set, jump map, and jump set) satisfy mild regularity conditions; see conditions (C1)–(C3) in Subsect. "Conditions for Well-posedness".

Hybrid controller Algorithm that takes, as inputs, measurements from a system to be controlled (called the plant) and combines behaviors of continuous-time and discrete-time controllers (i. e. flows and jumps) to produce, as outputs, signals that are to control the plant.

Hybrid closed-loop system The hybrid system resulting from the interconnection of a plant and a controller, at least one of which is a hybrid dynamical system.

Invariance principle A tool for studying asymptotic properties of bounded solutions to (hybrid) dynamical systems, applicable when asymptotic stability is absent. It characterizes the sets to which such solutions must converge, by relying in part on invariance properties of such sets.

Lyapunov stability theory A tool for establishing global asymptotic stability of a compact set without solving for the solutions to the hybrid dynamical system. A Lyapunov function is one that takes its minimum, which is zero, on the compact set, that grows unbounded as its argument grows unbounded, and that decreases in the direction of the flow map on the flow set and via the jump map on the jump set.

Supervisor of hybrid controllers A hybrid controller that coordinates the actions of a family of hybrid controllers in order to achieve a certain stabilization objective. Patchy control Lyapunov functions provide a means of constructing supervisors.

Temporal regularization A modification to a hybrid controller to enforce a positive lower bound on the amount of time between jumps triggered by the hybrid control algorithm.

Zeno (and discrete) solutions A solution (to a hybrid dynamical system) that has an infinite number of jumps in a finite amount of time. It is discrete if, moreover, the solution never flows, i. e., never changes continuously.

Notation

- $\mathbb{R}^n$ denotes n-dimensional Euclidean space. $\mathbb{R}$ denotes the real numbers. $\mathbb{R}_{\geq}$ denotes the nonnegative real numbers, i. e., $\mathbb{R}_{\geq} = [0, \infty)$. $\mathbb{Z}$ denotes the integers. $\mathbb{N}$ denotes the natural numbers including 0, i. e., $\mathbb{N} = \{0, 1, \ldots\}$.
- Given a set S, $\overline{S}$ denotes its closure.
- Given a vector $x \in \mathbb{R}^n$, $|x|$ denotes its Euclidean vector norm.
- $\mathbb{B}$ is the closed unit ball in the norm $|\cdot|$.
- Given a set $S \subset \mathbb{R}^n$ and a point $x \in \mathbb{R}^n$, $|x|_S := \inf_{y \in S} |x - y|$.
- Given sets S_1, S_2 subsets of $\mathbb{R}^n$, $S_1 + S_2 := \{x_1 + x_2 | \, x_1 \in S_1, x_2 \in S_2\}$.
- A function is said to be *positive definite with respect to a given compact set* in its domain if it is zero on that compact set and positive elsewhere. When the compact set is the origin, the function will be called *positive definite*.
- Given a function $h\colon \mathbb{R}^n \to \mathbb{R}$, $h^{-1}(c)$ denotes its c-level set, i. e. $h^{-1}(c) := \{z \in \mathbb{R}^n \,|\, h(z) = c\}$.
- The double-arrow notation e. g., $g\colon D \rightrightarrows \mathbb{R}^n$, indicates a set-valued mapping, in contrast to a single arrow used for functions.

Definition of the Subject

Control systems are ubiquitous in nature and engineering. They regulate physical systems to desirable conditions. The mathematical theory behind engineering control systems developed over the last century. It started with the elegant stability theory of linear dynamical systems and continued with the more formidable theory of nonlinear dynamical systems, rooted in knowledge of stability theory for attractors in nonlinear differential or difference equations. Most recently, researchers have recognized the limited capabilities of control systems modeled only by differential or difference equations. Thus, they have started to explore the capabilities of hybrid control systems. Hybrid control systems contain dynamical states that sometimes change continuously and other times change discontinuously. These states, which can flow and jump, together with the output of the system being regulated, are used to produce a (hybrid) feedback control signal. Hybrid control systems can be applied to classical systems, where their added flexibility permits solving certain challenging control problems that are not solvable with other methods. Moreover, a firm understanding of hybrid dynamical systems allows applying hybrid control theory to systems that are, themselves, hybrid in nature, that is, having states that can change continuously and also change discontinuously. The development of hybrid control theory is in its infancy, with progress being marked by a transition from ad-hoc methods to systematic design tools.

Introduction

This article will present a general framework for modeling hybrid control systems and analyzing their dynamical properties. It will put forth basic tools for studying asymptotic stability properties of hybrid systems. Then, particular aspects of hybrid control will be described, as well as approaches to successfully achieving control objectives via hybrid control even if they are not solvable with classical methods. First, some examples of hybrid control systems are given.

Hybrid dynamical systems combine behaviors typical of continuous-time dynamical systems (i. e., flows) and behaviors typical of discrete-time dynamical systems (i. e., jumps). Hybrid control systems exploit state variables that may flow as well as jump to achieve control objectives that are difficult or impossible to achieve with controllers that are not hybrid. Perhaps the simplest example of a hybrid control system is one that uses a relay-type hysteresis element to avoid cycling the system's actuators between "on" and "off" too frequently.

Consider controlling the temperature of a room by turning a heater on and off. As a good approximation, the room's temperature T is governed by the differential equation

$$\dot{T} = -T + T_0 + T_\Delta u \,, \tag{1}$$

where T_0 represents the natural temperature of the room, T_Δ represents the capacity of the heater to raise the temperature in the room by being always on, and the variable u represents the state of the heater, which can be either 1 ("on") or 0 ("off").

A typical temperature control task is to keep the temperature between two specified values $T_{\min}$ and $T_{\max}$ where

$$T_0 < T_{\min} < T_{\max} < T_0 + T_\Delta \,.$$

For purposes of illustration, consider the case when $T_{\min} = 70°\text{F}$, $T_{\max} = 80°\text{F}$. An algorithm that accomplishes this control task is

```
input T, u
if u=1 and T >= 80 then
  u=0
elseif u = 0 and T <= 70 then
  u=1
end
```

In words, if the heater is "on" and the temperature is larger than 80°F, then the heater is turned off, while if the heater is "off " and the temperature is smaller than 70°F, then the heater is turned on. By programming the controller with the algorithm above and closing the loop, the temperature of the system will evolve as shown in Fig. 1 (the values $T_0 = 60°\text{F}$ and $T_\Delta = 30°\text{F}$ were used in these simulations).

Following the logic in the algorithm above, the controller can be expressed by the following conditional difference equation

$$\begin{aligned} u^+ &= 0 \quad \text{if} \quad u = 1, T \geq 80 \\ u^+ &= 1 \quad \text{if} \quad u = 0, T \leq 70\,, \end{aligned}$$

where u^+ is the value of u after a jump. Combining this difference equation with the differential equation (1) leads to the closed-loop system

$$\left.\begin{aligned} \dot{T} &= -T + T_0 + T_\Delta u \\ \dot{u} &= 0 \end{aligned}\right\} \quad \begin{aligned} u = 0, T \geq 70 \quad &\text{or} \\ u = 1, T \leq 80 \quad & \end{aligned} \tag{2}$$

$$\left.\begin{aligned} T^+ &= T \\ u^+ &= 1 - q \end{aligned}\right\} \quad u = 1, T \geq 80 \quad \text{or} \quad u = 0, T \leq 70\,. \tag{3}$$

This closed-loop system is a hybrid dynamical system. The state variables are T and u; the continuous dynamics or *flows* as well as the constraints on the continuous evolution are given by (2); the discrete dynamics or *jumps* as well as the constraints on the discrete evolution are given by (3).

As in the temperature control problem above, a hybrid control system can arise from controlling a classical system with a hybrid controller. In a more general setting, hybrid control systems emerge when controlling hybrid systems with nonlinear and/or hybrid controllers. Consider controlling the vertical motion of a ball through collisions with an actuated robot. Figure 2 depicts such a scenario. The impacts between the ball and the robot generate a jump in their velocity. When the control task is to stabilize the ball to a periodic motion, like the height pattern in Fig. 2b, this system is referred to as the *one degree-of-freedom juggler*.

The dynamics of the ball in between the collisions are given by classical physics laws and can be written in terms of the ball's height and vertical velocity, denoted by x_{11} and x_{12}, respectively, as follows:

$$\begin{aligned} \dot{x}_{11} &= x_{12} \\ \dot{x}_{12} &= -\gamma\,, \end{aligned} \tag{4}$$

where γ is the gravity constant. We denote the mass of the ball by m_1.

We consider a robot with a vertical velocity control input u and denote the robot's height and velocity by x_{21} and x_{22}, respectively. Then, its dynamics are given by

$$\begin{aligned} \dot{x}_{21} &= x_{21} \\ \dot{x}_{22} &= u\,. \end{aligned} \tag{5}$$

The mass of the actuated robot is denoted by m_2.

Following [10,58], impacts between the ball and the robot are assumed to conserve momentum, i. e.,

$$m_1 x_{12}^+ + m_2 x_{22}^+ = m_1 x_{12} + m_2 x_{22}\,, \tag{6}$$

and, for some restitution coefficient $e \in (0, 1)$, satisfy

$$x_{12}^+ - x_{22}^+ = -e(x_{12} - x_{22})\,. \tag{7}$$

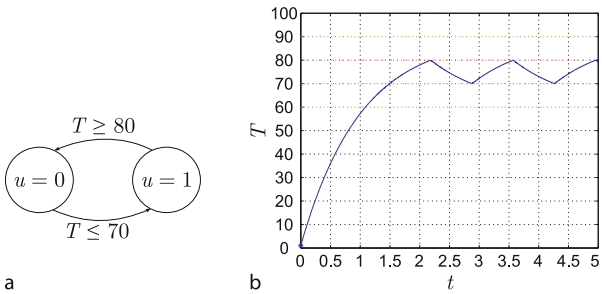

Hybrid Control Systems, Figure 1
Temperature control. a Flow diagram of control algorithm. b Evolution of temperature with control algorithm

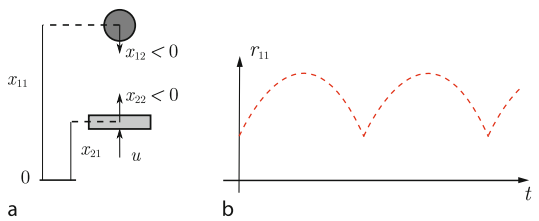

Hybrid Control Systems, Figure 2
One degree-of-freedom juggler and a juggling task. Positions are denoted by x_{11}, x_{21} and velocities by x_{12}, x_{22}, respectively. The control input is denoted by u. A desired height pattern is denoted by r_{11}. a Ball and actuated robot. b Desired ball's height pattern

Defining $\lambda := \frac{m_1}{m_1+m_2}$, Eqs. (6) and (7) can be combined to obtain the update law for velocities

$$\begin{bmatrix} x_{12}^+ \\ x_{12}^+ \end{bmatrix} = \begin{bmatrix} -e+\lambda(1+e) & (1-\lambda)(1+e) \\ \lambda(1+e) & 1-\lambda(1+e) \end{bmatrix} \begin{bmatrix} x_{12} \\ x_{22} \end{bmatrix} =: \Gamma(\lambda, e) \begin{bmatrix} x_{12} \\ x_{22} \end{bmatrix} .$$

The update law for positions is given by

$$x_{11}^+ = x_{11} , \quad x_{21}^+ = x_{21} .$$

Impacts between the ball and the actuated robot occur when their heights are the same, that is, $x_{11} = x_{21}$, and when their velocities indicate that they are not moving away from each other, that is, $x_{12} \le x_{22}$.

The derivation above leads to the following model for the one degree-of-freedom juggler system in Fig. 2:

Flows:

$$\left.\begin{aligned} \dot{x}_{11} &= x_{12} , \quad \dot{x}_{12} - \gamma \\ \dot{x}_{21} &= x_{22} , \quad \dot{x}_{22} = u \end{aligned}\right\} \quad x_{11} - x_{21} \ge 0 .$$

Jumps:

$$\left.\begin{aligned} x_{11}^+ &= x_{11} \\ x_{12}^+ &= \begin{bmatrix} 1 & 0 \end{bmatrix} \Gamma(\lambda, e) \begin{bmatrix} x_{12} \\ x_{22} \end{bmatrix} \\ x_{21}^+ &= x_{21} \\ x_{22}^+ &= \begin{bmatrix} 1 & 0 \end{bmatrix} \Gamma(\lambda, e) \begin{bmatrix} x_{12} \\ x_{22} \end{bmatrix} \end{aligned}\right\} \text{ and } \quad \begin{aligned} x_{11} - x_{21} &= 0 \\ x_{12} - x_{22} &\le 0 . \end{aligned}$$

A controller designed to accomplish stabilization of the ball to a periodic pattern will only be able to measure the ball's state at impacts. During flows, it will be able to control the robot's velocity through u. Regardless of the nature of the controller, the closed-loop system will be a hybrid system by virtue of the dynamics of the one degree-of-freedom system. Figure 3 shows a trajectory to the closed-loop system with a controller that stabilizes the ball state to the periodic pattern in Fig. 2b (note the discontinuity in the velocity of the ball at impacts); see the control strategy in [55].

Following the modeling techniques illustrated by the examples above, the next section introduces a general modeling framework for hybrid dynamical systems. The framework makes possible the development of a robust stability theory for hybrid dynamical systems and prepares the way for insights into the design of robust hybrid control systems.

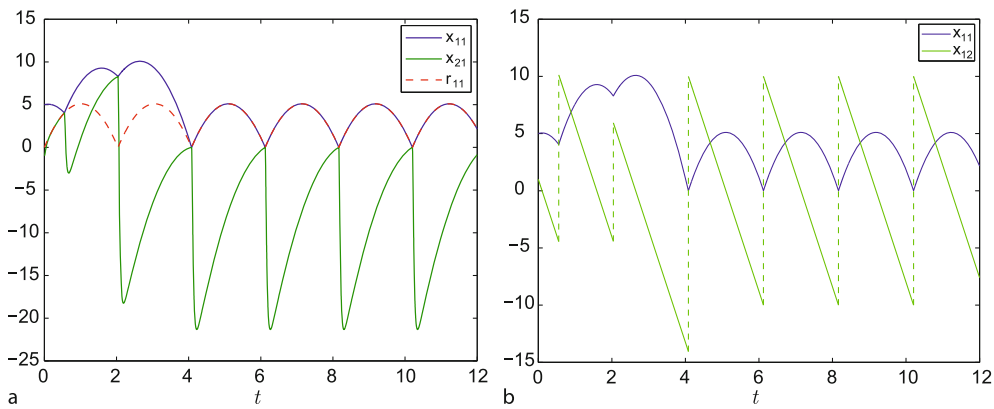

Hybrid Control Systems, Figure 3
Trajectories to the one degree-of-freedom juggler. Their positions are denoted by x_{11}, x_{21} and their velocities by x_{12}, x_{22}, respectively. The desired height pattern is denoted by r_{11}. a Juggling on ball. b Ball's position and velocity

Well-posed Hybrid Dynamical Systems

For numerous mathematical problems, well-posedness refers to the uniqueness of a solution and its continuous dependence on parameters, for example on initial conditions. Here, well-posedness will refer to some mild regularity properties of the data of a hybrid system that enable the development of a robust stability theory.

Hybrid Behavior and Model

Hybrid dynamical systems combine continuous and discrete dynamics. Such a combination may emerge when controlling a continuous-time system with a control algorithm that incorporates discrete dynamics, like in the temperature control problem in Sect. "Introduction", when controlling a system that features hybrid phenomena, like in the juggling problem in Sect. "Introduction", or as a modeling abstraction of complex dynamical systems.

Solutions to hybrid systems (sometimes referred to as *trajectories*, *executions*, *runs*, or *motions*) can evolve both continuously, i. e. *flow*, and discontinuously, i. e. *jump*. Figure 4 depicts a representative behavior of a solution to a hybrid system.

For a purely continuous-time system, flows are usually modeled by differential equations, and sometimes by differential inclusions. For a purely discrete-time system, jumps are usually modeled by difference equations, and sometimes by difference inclusions. Set-valued dynamics naturally arise as regularizations of discontinuous difference and differential equations and represent the effect of state perturbations on such equations, in particular, the effect of state measurement errors when the equations represent a system in closed loop with a (discontinuous) feedback controller. For the continuous-time case, see the work by Filippov [22] and Krasovskii [34], as well as [26,27,]; for the discrete-time case, see [33].

When working with hybrid control systems, it is appropriate to allow for set-valued discrete dynamics in order to capture decision making capabilities, which are typical in hybrid feedback. Difference inclusions, rather than equations, also arise naturally in modeling of hybrid automata, for example when the discrete dynamics are generated by multiple so-called "guards" and "reset maps"; see, e. g. [8,11,40] or [56] for details on modeling guards and resets in the current framework.

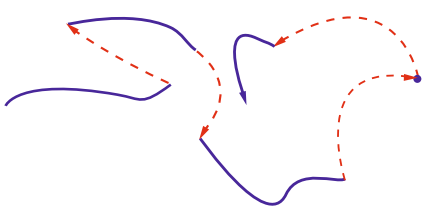

Hybrid Control Systems, Figure 4
Evolution of a hybrid system: continuous motion during flows (*solid*), discontinuous motion at jumps (*dashed*)

Naturally, differential equations and difference inclusions will be featured in the model of a hybrid system. In most hybrid systems, or even in some purely continuous-time systems, the flow modeled by a differential equation is allowed to occur only on a certain subset of the state space $\mathbb{R}^n$. Similarly, the jumps modeled by a difference equation may only be allowed from a certain subset of the state space. Hence, the model of the hybrid system stated above will also feature a *flow set*, restricting the flows, and a *jump set*, restricting the jumps.

More formally, a hybrid system will be modeled with the following data:

- The flow set $C \subset \mathbb{R}^n$;
- The flow map $f\colon C \to \mathbb{R}^n$;
- The jump set $D \subset \mathbb{R}^n$;
- The (set-valued) jump map $G\colon D \rightrightarrows \mathbb{R}^n$.

A shorthand notation for a hybrid system with this data will be $\mathcal{H} = (f, C, G, D)$. Such systems can be written in the suggestive form

$$\mathcal{H}: \quad x \in \mathbb{R}^n \quad \begin{cases} \dot{x} = f(x)\,, & x \in C \\ x^+ \in G(x)\,, & x \in D\,, \end{cases} \tag{8}$$

where $x \in \mathbb{R}^n$ denotes the state of the system, $\dot{x}$ denotes its derivative with respect to time, and x^+ denotes its value after jumps. In several control applications, the state x of the hybrid system can contain logic states that take value in discrete sets (representing, for example, "on" or "off" states, like in the temperature control problem in Sect. "Introduction").

Two parameters will be used to specify "time" in solutions to hybrid systems: t, taking values in $\mathbb{R}_{\geq 0}$, and representing the elapsed "real" time; and j, taking values in $\mathbb{N}$, and representing the number of jumps that have occurred. For each solution, the combined parameters (t, j) will be restricted to belong to a hybrid time domain, a particular subset of $\mathbb{R}_{\geq 0} \times \mathbb{N}$. Hybrid time domains corresponding to different solutions may differ.

Note that with such a parametrization, both purely continuous-time and purely discrete-time dynamical systems can be captured. Furthermore, for truly hybrid solutions, both flows and jumps are parametrized "symmetrically" (cf. [8,40]).

A subset E of $\mathbb{R}_{\geq 0} \times \mathbb{N}$ is a *hybrid time domain* if it is the union of infinitely many intervals of the

form $[t_j, t_{j+1}] \times \{j\}$, where $0 = t_0 \leq t_1 \leq t_2 \leq \ldots$, or of finitely many such intervals, with the last one possibly of the form $[t_j, t_{j+1}] \times \{j\}$, $[t_j, t_{j+1}) \times \{j\}$, or $[t_j, \infty) \times \{j\}$. On each hybrid time domain there is a natural ordering of points: we write $(t, j) \preceq (t' j')$ for $(t, j), (t', j') \in E$ if $t \leq t'$ and $j \leq j'$.

Solutions to hybrid systems are given by functions, which are called *hybrid arcs*, defined on hybrid time domains and satisfying the dynamics and the constraints given by the data of the hybrid system. A hybrid arc is a function $x\colon \operatorname{dom} x \to \mathbb{R}^n$, where $\operatorname{dom} x$ is a hybrid time domain and $t \mapsto x(t, j)$ is a locally absolutely continuous function for each fixed j. A hybrid arc x is a solution to $\mathcal{H} = (f, C, G, D)$ if $x(0, 0) \in C \cup D$ and it satisfies

Flow condition:

$$\dot{x}(t, j) = f(x(t, j)) \quad \text{and} \quad x(t, j) \in C \tag{9}$$

for all $j \in \mathbb{N}$ and almost all t such that $(t, j) \in \operatorname{dom} x$;

Jump condition:

$$x(t, j+1) \in G(x(t, j)) \quad \text{and} \quad x(t, j) \in D \tag{10}$$

for all $(t, j) \in \operatorname{dom} x$ such that $(t, j+1) \in \operatorname{dom} x$.

Figure 5 shows a solution to a hybrid system $\mathcal{H} = (f, C, G, D)$ flowing (as solutions to continuous-time systems do) while in the flow set C and jumping (as solutions to discrete-time systems do) from points in the jump set D.

A hybrid arc x is said to be *nontrivial* if $\operatorname{dom} x$ contains at least one point different from $(0, 0)$ and *complete* if $\operatorname{dom} x$ is unbounded (in either the t or j direction, or both). It is said to be *Zeno* if it has an infinite number of jumps in a finite amount of time, and *discrete* if it has an infinite number of jumps and never flows. A solution x to a hybrid system is *maximal* if it cannot be extended, i. e., there is no solution x' such that $\operatorname{dom} x$ is a proper subset of $\operatorname{dom} x'$ and x agrees with x' on $\operatorname{dom} x$. Obviously, complete solutions are maximal.

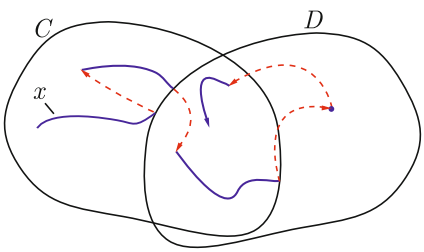

Hybrid Control Systems, Figure 5
Evolution of a solution to a hybrid system. Flows and jumps of the solution *x* are allowed only on the flow set *C* and on the jump set *D*, respectively

Conditions for Well-Posedness

Many desired results in stability theory for dynamical systems, like invariance principles, converse Lyapunov theorems, or statements about generic robustness of stability, hinge upon some fundamental properties of the space of solutions to the system. These properties may involve continuous dependence of solutions on initial conditions, completeness and sequential compactness of the space of solutions, etc.

To begin addressing these or similar properties for hybrid systems, one should establish a concept of distance between solutions. In contrast to purely continuous-time systems or purely discrete-time systems, the uniform metric is not a suitable indicator of distance: two solutions experiencing jumps at close but not the same times will not be close in the uniform metric, even if (intuitively) they represent very similar behaviors. For example, Fig. 6 shows two solutions to the juggling problem in Sect. "Introduction" starting at nearby initial conditions for which the velocities are not close in the uniform metric.

A more appropriate distance notion should take possibly different jump times into account. We use the following notion: given $T, J, \varepsilon > 0$, two hybrid arcs $x\colon \operatorname{dom} x \to \mathbb{R}^n$ and $y\colon \operatorname{dom} y \to \mathbb{R}^n$ are said to be (T, J, ε)-*close* if:

(a) for all $(t, j) \in \operatorname{dom} x$ with $t \leq T$, $j \leq J$ there exists s such that $(s, j) \in \operatorname{dom} y$, $|t - s| < \varepsilon$, and

$$|x(t, j) - y(s, j)| < \varepsilon\,,$$

(b) for all $(t, j) \in \operatorname{dom} y$ with $t \leq T$, $j \leq J$ there exists s such that $(s, j) \in \operatorname{dom} x$, $|t - s| < \varepsilon$, and

$$|y(t, j) - x(s, j)| < \varepsilon\,.$$

An appealing geometric interpretation of (T, J, ε)-closeness of x and y can be given. The graph of a hybrid arc $x\colon \operatorname{dom} x \to \mathbb{R}^n$ is the subset of $\mathbb{R}^{n+2}$ given by

$$\operatorname{gph} x := \{(t, j, z) \,|\, (t, j) \in \operatorname{dom} x,\ z = x(t, j)\}\,.$$

Hybrid arcs x and y are (T, J, ε)-close if the restriction of the graph of x to $t \leq T$, $j \leq J$, i. e., the set $\{(t, j, z) | (t, j) \in \operatorname{dom} x, t \leq T, j \leq J, z = x(t, j)\}$, is in the ε-neighborhood of $\operatorname{gph} y$, and vice versa: the restriction of the graph of y is in the ε-neighborhood of $\operatorname{gph} x$. (The neighborhoods of $\operatorname{gph} x$ and $\operatorname{gph} y$ should be understood in the norm for which the unit ball is $[-1, 1] \times [-1, 1] \times \mathbb{B}$.)

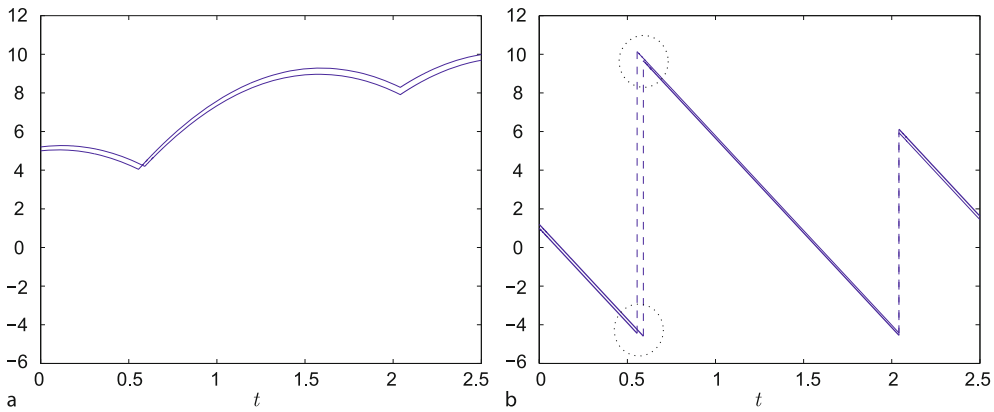

Hybrid Control Systems, Figure 6
Two solutions to the juggling system in Sect. "Introduction" starting at nearby initial conditions. Velocities are not close in the uniform metric near the jump times. a Ball's heights. b Ball's velocities

The (T, J, ε)-closeness can be used to quantify the concept of graphical convergence of a sequence of hybrid arcs; for details, see [23]. Here, it is only noted that graphical convergence of a sequence of mappings is understood as convergence of the sequence of graphs of these mappings. Such a convergence concept does have solid intuitive motivation when the mappings considered are associated with solutions to hybrid systems.

It turns out that when (T, J, ε)-closeness and graphical convergence are used to study the properties of the space of solutions to a hybrid system $\mathcal{H} = (f, C, G, D)$, only mild and easy to verify conditions on the data of $\mathcal{H}$ are needed to ensure that $\mathcal{H}$ is "well-posed". These conditions are:

(C1) the flow set C and jump set D are closed;
(C2) the flow map $f\colon C \to \mathbb{R}^n$ is continuous;
(C3) the jump map $G\colon D \rightrightarrows \mathbb{R}^n$ is outer semicontinuous and locally bounded.

Only (C3) requires further comment: $G\colon D \rightrightarrows \mathbb{R}^n$ is outer semicontinuous if for every convergent sequence $x_i \in D$ with $x_i \to x$ and every convergent sequence $y_i \in G(x_i)$ with $y_i \to y$, one has $y \in G(x)$; G is locally bounded if for each compact set $K \subset \mathbb{R}^n$ there exists a compact set $K' \subset \mathbb{R}^n$ such that $G(x) \subset K'$ for all $x \in K$. Any system $\mathcal{H} = (f, C, G, D)$ meeting (C1), (C2), and (C3) will be referred to as well-posed.

One important consequence of a hybrid system $\mathcal{H}$ being well-posed is the following:

($\star$) Every sequence of solutions to $\mathcal{H}$ has a subsequence that graphically converges to a solution to $\mathcal{H}$, which holds under very mild boundedness assumptions about the sequence of solutions in question. The assumptions hold, for example, if the sequence is uniformly bounded, i. e., there exists a compact set $K \subset \mathbb{R}^n$ such that, for each i, $x_i(t, j) \in K$ for all $(t, j) \in \operatorname{dom} x_i$, where $\{x_i\}_{i=1}^{\infty}$ is a sequence of solutions.

Another important consequence of well-posedness is the following *outer-semicontinuity* (or *upper-semicontinuity*) property:

($\star\star$) For every $x^0 \in \mathbb{R}^n$, every desired level of closeness of solutions $\varepsilon > 0$, and every (T, J), there exists a level of closeness for initial conditions $\delta > 0$ so that for every solution x_δ to $\mathcal{H}$ with $|x_\delta(0, 0) - x^0| < \delta$, there exist a solution x to $\mathcal{H}$ with $x(0, 0) = x^0$ such that x_δ and x are (T, J, ε)-close.

This property holds at each $x^0 \in \mathbb{R}^n$ from which all maximal solutions to $\mathcal{H}$ are either complete or bounded.

Properties ($\star$) and ($\star\star$), while being far weaker than any kind of continuous dependence of solutions on initial conditions, are sufficient to develop basic stability characterizations. Continuous dependence of solutions on initial conditions is rare in hybrid systems, as in its more classical meaning it entails uniqueness of solutions from each initial point. If understood in a set-valued sense, in order for inner-semicontinuity (or lower-semicontinuity) to be present, it still requires many further assumptions on the data.

Property ($\star$) is essentially all that is needed to establish invariance principles, which will be presented in Subsect. "Invariance Principles". Property ($\star\star$) is useful in describing, for example, uniformity of convergence and

of overshoots in an asymptotically stable hybrid system. For the analysis of robustness properties of well-posed hybrid systems, strengthened versions of ($\star$) and ($\star\star$) are available. They take into account the effect of small perturbations; for example a stronger version of ($\star$) makes the same conclusion not about a sequence of solutions to $\mathcal{H}$, but about a sequence of solutions to $\mathcal{H}$ generated with vanishing perturbations. (More information can be found in [23].) It is the stronger versions of the two properties that make converse Lyapunov results possible; see Subsect. "Converse Lyapunov Theorems and Robustness", where more precise meaning to perturbations is also given.

In the rest of this article, the analysis results will assume that (C1)–(C3) hold, i. e., the hybrid system under analysis is well-posed. The control algorithms will be constructed so that the corresponding closed-loop systems are well-posed. Such algorithms will be called well-posed controllers.

Modeling Hybrid Control Systems

Hybrid Controllers for Classical Systems

Given a nonlinear control system of the form

$$\mathcal{P}: \quad \begin{cases} \dot{x} = f(x,u)\,, & x \in C_P \\ y = h(x)\,, \end{cases} \tag{11}$$

where C_P is a subset of the state space where the system is allowed to evolve, a general output-feedback hybrid controller $\mathcal{K} = (\kappa, \phi, C_K, \psi, D_K)$ takes the form

$$\mathcal{K}: \quad \begin{cases} u = \kappa(y,\eta) \\ \dot{\eta} = \phi(y,\eta)\,, & (y,\eta) \in C_K \\ \eta^+ \in \psi(y,\eta)\,, & (y,\eta) \in D_K\,, \end{cases}$$

where the output of the plant $y \in \mathbb{R}^p$ is the input to the controller, the input to the plant $u \in \mathbb{R}^m$ is the output of the controller, and $\eta \in \mathbb{R}^k$ is the controller state. When system (11) is controlled by $\mathcal{K}$, their interconnection results in a hybrid closed-loop system given by

$$\begin{aligned} \left.\begin{aligned} \dot{x} &= f(x,\kappa(h(x),\eta)) \\ \dot{\eta} &= \phi(h(x),\eta) \end{aligned}\right\} & \quad (x,\eta) \in C \\ \left.\begin{aligned} x^+ &= x \\ \eta^+ &\in \psi(h(x),\eta) \end{aligned}\right\} & \quad (x,\eta) \in D\,, \end{aligned} \tag{12}$$

where

$$\begin{aligned} C &:= \{(x,\eta)\,|\,x \in C_P, (h(x),\eta) \in C_K\} \\ D &:= \{(x,\eta)\,|\,(h(x),\eta) \in D_K\}\,. \end{aligned}$$

We now cast some specific situations into this framework.

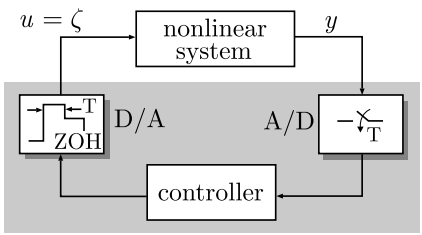

Hybrid Control Systems, Figure 7
Sample-and-hold control of a nonlinear system

Sample-and-hold Control

Perhaps the simplest example of a hybrid system that arises in control system design is when a continuous-time plant is controlled via a digital computer connected to the plant through a sample-and-hold device. This situation is ubiquitous in feedback control applications. A hybrid system emerges by considering the nonlinear control system (11), with $C_P = \mathbb{R}^n$, where the measurements y are sampled every $T > 0$ seconds, producing a sampled signal y_s that is processed through a discrete-time algorithm

$$\begin{bmatrix} z^+ \\ u_s \end{bmatrix} = \chi(z, y_s)$$

to generate a sequence of input values u_s each of which is held for T seconds to generate the input signal u.

When combined with the continuous-time dynamics, this algorithm will do the following:

- At the beginning of the sample period, it will update the value of u and the value of the controller's internal state z, based on the values of z and y (denoted y_s) at the beginning of the sampling period.
- During the rest of the sampling period, it will hold the values of u and z constant.

A complete model of this behavior is captured by defining the controller state to be

$$\eta := \begin{bmatrix} z \\ \zeta \\ \tau \end{bmatrix},$$

where ζ keeps track of the input value to hold during a sampling period and τ is a timer state that determines when the state variables z and ζ should be updated. The hybrid controller is specified in the form of the previous

section as

$$\left.\begin{aligned} u &= \zeta \\ \dot{z} &= 0 \\ \dot{\zeta} &= 0 \\ \dot{\tau} &= 1 \end{aligned}\right\} \quad (y, z, \zeta, \tau) \in C_K$$

$$\left.\begin{aligned} \begin{bmatrix} z \\ \zeta \end{bmatrix}^+ &= \chi(z, y) \\ \tau^+ &= 0 \end{aligned}\right\} \quad (y, z, \zeta, \tau) \in D_K \,,$$

where

$$C_K := \{(y, z, \zeta, \tau) \,|\, \tau \in [0, T]\} \,,$$
$$D_K := \{(y, z, \zeta, \tau) \,|\, \tau = T\} \,.$$

The overall closed-loop hybrid system is given by

$$\left.\begin{aligned} \dot{x} &= f(x, \zeta) \\ \dot{z} &= 0 \\ \dot{\zeta} &= 0 \\ \dot{\tau} &= 1 \end{aligned}\right\} \quad (x, z, \zeta, \tau) \in C$$

$$\left.\begin{aligned} x^+ &= x \\ \begin{bmatrix} z \\ \zeta \end{bmatrix}^+ &= \chi(z, h(x)) \\ \tau^+ &= 0 \end{aligned}\right\} \quad (x, z, \zeta, \tau) \in D \,,$$

where

$$C := \{(x, z, \zeta, \tau) \,|\, \tau \in [0, T]\} \,,$$
$$D := \{(x, z, \zeta, \tau) \,|\, \tau = T\} \,.$$

Notice that if the function χ is discontinuous then this may fail to be a well-posed hybrid system. In such a case, it becomes well-posed by replacing the function χ by its set-valued regularization

$$\bar{\chi}(z, y) := \bigcap_{\delta > 0} \overline{\chi((z, y) + \delta \mathbb{B})} \,.$$

This corresponds to allowing, at points of discontinuity, values that can be obtained with arbitrarily small perturbations of z and y. Allowing these values is reasonable in light of inevitable control systems perturbations, like measurement noise and computer round-off error.

Networked Control Systems

Certain classes of networked control systems can be viewed as generalizations of systems with a sample-and-hold device. The networked control systems generalization allows for multiple sample-and-hold devices operating simultaneously and asynchronously, and with a variable sampling period. Compared to the sample-and-hold closed-loop model in Subsect. "Sample-and-hold Control", one can think of u as a large vector of inputs to a collection of i plants, collectively modeled by $\dot{x} = f(x, u)$. The update rule for u may only update a certain part of u at a given jump time. This update rule may depend not only on z and y but perhaps also on u and a logic variable, which we denote by ℓ, that may be cycling through the list of i indices corresponding to connections to different plants. Several common update protocols use algorithms that are discontinuous functions, so this will be modeled explicitly by allowing a set-valued update rule. Finally, due to time variability in the behavior of the network connecting the plants, the updates may occur at any time in an interval $[T_{\min}, T_{\max}]$ where $T_{\min} > 0$ represents the minimum amount of time between transmissions in the network and $T_{\max} > T_{\min}$ represents the maximum amount of time between transmissions. The overall control system for the network of plants is given by

$$\left.\begin{aligned} u &= \zeta \\ \dot{z} &= 0 \\ \dot{\zeta} &= 0 \\ \dot{\ell} &= 0 \\ \dot{\tau} &= 1 \end{aligned}\right\} \quad (y, z, \zeta, \ell, \tau) \in C_K$$

$$\left.\begin{aligned} \begin{bmatrix} z \\ \zeta \end{bmatrix}^+ &\in \chi(z, y, \zeta, \ell) \\ \ell^+ &= (\ell \bmod i) + 1 \\ \tau^+ &= 0 \end{aligned}\right\} \quad (y, z, \zeta, \ell, \tau) \in D_K \,,$$

where

$$C_K := \{(y, z, \zeta, \ell, \tau) \,|\, \tau \in [0, T_{\max}]\} \,,$$
$$D_K := \{(y, z, \zeta, \ell, \tau) \,|\, \tau \in [T_{\min}, T_{\max}]\} \,.$$

The closed-loop networked control system has the form

$$\left.\begin{aligned} \dot{x} &= f(x, \zeta) \\ \dot{z} &= 0 \\ \dot{\zeta} &= 0 \\ \dot{\ell} &= 0 \\ \dot{\tau} &= 1 \end{aligned}\right\} \quad (x, z, \zeta, \ell, \tau) \in C$$

$$\left.\begin{aligned} x^+ &= x \\ \begin{bmatrix} z \\ \zeta \end{bmatrix}^+ &\in \chi(z, h(x), \zeta, \ell) \\ \ell^+ &= (\ell \bmod i) + 1 \\ \tau^+ &= 0 \end{aligned}\right\} \quad (x, z, \zeta, \ell, \tau) \in D \,,$$

where

$$C := \{(x, z, \zeta, \ell, \tau) \,|\, \tau \in [0, T_{\max}]\} \ ,$$
$$D := \{(x, z, \zeta, \ell, \tau) \,|\, \tau \in [T_{\min}, T_{\max}]\} \ .$$

Reset Control Systems

The first documented reset controller was created by Clegg [19]. Consisting of an operational amplifier, resistors, and diodes, Clegg's controller produced an output that was the integral of its input subject to the constraint that the sign of the output and input agreed. This was achieved by forcing the state of the circuit to jump to zero, a good approximation of the behavior induced by the diodes, when the circuit's input changed sign with respect to its output. Consider such a circuit in a negative feedback loop with a linear control system

$$\dot{x} = Ax + Bu \ , \quad x \in \mathbb{R}^n \ , \ u \in \mathbb{R}$$
$$y = Cx \ , \quad y \in \mathbb{R} \ .$$

Use η to denote the state of the integrator. Then, the hybrid model of the Clegg controller is given by

$$\begin{aligned} u &= \eta \\ \dot{\eta} &= -y \ , \quad (y, \eta) \in C_K \\ \eta^+ &= 0 \ , \quad (y, \eta) \in D_K \ , \end{aligned}$$

where

$$C_K := \{(y, \eta) \,|\, \eta y \leq 0\} \ ,$$
$$D_K := \{(y, \eta) \,|\, \eta y \geq 0\} \ .$$

One problem with this model is that it exhibits discrete solutions, as defined in Subsect. "Hybrid Behavior and Model". Indeed, notice that the jump map takes points with $\eta = 0$, which are in the jump set D_K, back to points with $\eta = 0$. Thus, there are complete solutions that start with $\eta = 0$ and never flow, which corresponds to the definition of a discrete solution.

There are several ways to address this issue. When a reset controller like the Clegg integrator is implemented through software, a temporal regularization, as discussed next in Subsect. "Zeno Solutions and Temporal Regularization", can be used to force a small amount of flow time between jumps. Alternatively, and also for the case of an analog implementation, one may consider a more detailed model of the reset mechanism, as the model proposed above is not very accurate for the case where η and y are small. This modeling issue is analogous to hybrid modeling issues for a ball bouncing on a floor where the simplest model is not very accurate for small velocities.

Zeno Solutions and Temporal Regularization

Like for reset control systems, the closed-loop system (12) may exhibit Zeno solutions, i. e., solutions with an infinite number of jumps in a finite amount of time. These are relatively easy to detect in systems with bounded solutions, as they exist if and only if there exist discrete solutions, i. e., solutions with an infinite number of jumps and no flowing time. Discrete solutions in a hybrid control system are problematic, especially from an implementation point of view, but they can be removed by means of temporal regularization. The temporal regularization of a hybrid controller $\mathcal{K} = (\kappa, \phi, C_K, \psi, D_K)$ is generated by introducing a timer variable τ that resets to zero at jumps and that must pass a threshold defined by a parameter $\delta \in (0, 1)$ before another jump is allowed. The regularization produces a well-posed hybrid controller $\mathcal{K}_\delta := (\tilde{\kappa}, \tilde{\phi}, C_{K,\delta}, \tilde{\psi}, D_{K,\delta})$ with state $\tilde{\eta} := (\eta, \tau) \in \mathbb{R}^{k+1}$, where

$$\begin{aligned} \tilde{\kappa}(\tilde{\eta}) &:= \kappa(\eta) \\ \tilde{\phi}(\tilde{\eta}) &:= \phi(\eta) \times \{1 - \tau\} \ , \\ \tilde{\psi}(\tilde{\eta}) &:= \psi(\eta) \times \{0\} \ , \\ C_{K,\delta} &:= \big(C_K \times \mathbb{R}_{\geq 0}\big) \cup (\mathbb{R}^k \times [0, \delta]) \ , \\ D_{K,\delta} &:= D_K \times [\delta, 1] \ . \end{aligned}$$

This regularization is related to one type of temporal regularization introduced in [32]. The variable τ is initialized in the interval $[0, 1]$ and remains there for all time. When $\delta = 0$, the controller accepts flowing only if $(y, \eta) \in C_K$, since $\dot{\tau} = 1 - \tau$ and the flow condition for τ when $(y, \eta) \notin C_K$ is $\tau = 0$. Moreover, jumping is possible for $\delta = 0$ if and only if $(y, \eta) \in D_K$. Thus, the controller with $\delta = 0$ has the same effect on the closed loop as the original controller $\mathcal{K}$. When $\delta > 0$, the controller forces at least δ seconds between jumps since $\dot{\tau} \leq 1$ for all $\tau \in [0, \delta]$. In particular, Zeno solutions, if there were any, are eliminated. Based on the remarks at the end of Subsect. "Conditions for Well-Posedness", we expect the effect of the controller for small $\delta > 0$ to be close to the effect of the controller with $\delta = 0$. In particular, we expect that this temporal regularization for small $\delta > 0$ will not destroy the stability properties of the closed-loop hybrid system, at least in a practical sense. This aspect is discussed in more detail in Subsect. "Zeno Solutions, Temporal Regularization, and Robustness".

Hybrid Controllers for Hybrid Systems

Another interesting scenario in hybrid control is when a hybrid controller

$$\mathcal{K}\begin{cases} u = \kappa(y,\eta) & \\ \dot{\eta} = \phi(y,\eta)\,, & (y,\eta) \in C_K \\ \eta^+ \in \psi(y,\eta)\,, & (y,\eta) \in D_K \end{cases}$$

is used to control a plant that is also hybrid, perhaps modeled as

$$\begin{aligned} \dot{x} &= f(x,u)\,, \quad x \in C_P \\ x^+ &= g(x)\,, \quad x \in D_P \\ y &= h(x)\,. \end{aligned}$$

This is the situation for the juggling example presented in Sect. "Introduction". In this scenario, the hybrid closed-loop system is modeled as

$$\begin{aligned} &\left.\begin{aligned} \dot{x} &= f(x,\kappa(h(x),\eta)) \\ \dot{\eta} &= \phi(h(x),\eta) \end{aligned}\right\} && (x,\eta) \in C \\ &\begin{bmatrix} x \\ \eta \end{bmatrix}^+ \in G(x,\eta)\,, && (x,\eta) \in D \end{aligned}$$

where

$$\begin{aligned} C &:= \{(x,\eta)\,|\,x \in C_P, (h(x),\eta) \in C_K\} \\ D &:= \{(x,\eta)\,|\,x \in D_P \text{ or } (h(x),\eta) \in D_K\} \end{aligned}$$

and

$$\begin{aligned} G_P(x,\eta) &:= \begin{bmatrix} g(x) \\ \eta \end{bmatrix} \\ G_K(x,\eta) &:= \begin{bmatrix} \{x\} \\ \psi(h(x),\eta) \end{bmatrix} \\ G(x,\eta) &:= \begin{cases} G_P(x,\eta)\,, & x \in D_P\,,\ (h(x),\eta) \notin D_K \\ G_K(x,\eta)\,, & x \notin D_P\,,\ (h(x),\eta) \in D_K \\ G_P(x,\eta) \cup G_K(x,\eta)\,, & \\ & x \in D_P\,,\ (h(x),\eta) \in D_K\,. \end{cases} \end{aligned}$$

As long as the data of the controller and plant are well-posed, the closed-loop system is a well-posed hybrid system. Also, it can be verified that the only way this model can exhibit discrete solutions is if either the plant exhibits discrete solutions or the controller, with constant y, exhibits discrete solutions. Indeed, if the plant does not exhibit discrete solutions then a discrete solution for the closed-loop system would eventually have to have x, and thus y, constant. Then, if there are no discrete solutions to the controller with constant y, there can be no discrete solutions to the closed-loop system.

Stability Theory

Lyapunov stability theory for dynamical systems typically states that asymptotically stable behaviors can be characterized by the existence of energy-like functions, which are called Lyapunov functions. This theory has served as a powerful tool for stability analysis of nonlinear dynamical systems and has enabled systematic design of robust control systems. In this section, we review some recent advances on Lyapunov-based stability analysis tools for hybrid dynamical systems.

Global (Pre-)Asymptotic Stability

In a classical setting, say of differential equations with Lipschitz continuous right-hand sides, existence of solutions and completeness of maximal ones can be taken for granted. These properties, together with a Lyapunov inequality ensuring that the Lyapunov function decreases along each solution, lead to a classical concept of asymptotic stability. On its own, the Lyapunov inequality does not say anything about the existence of solutions. It is hence natural to talk about a concept of asymptotic stability that is related only to the Lyapunov inequality. This appears particularly natural for the case of hybrid systems, where existence and completeness of solutions can be problematic.

The compact set $\mathcal{A} \subset \mathbb{R}^n$ is *stable* for $\mathcal{H}$ if for each $\varepsilon > 0$ there exists $\delta > 0$ such that any solution x to $\mathcal{H}$ with $|x(0,0)|_{\mathcal{A}} \le \delta$ satisfies $|x(t,j)|_{\mathcal{A}} \le \varepsilon$ for all $(t,j) \in \operatorname{dom} x$; it is globally *pre-attractive* for $\mathcal{H}$ if any solution x to $\mathcal{H}$ is bounded and if it is complete then $x(t,j) \to \mathcal{A}$ as $t + j \to \infty$; it is globally *pre-asymptotically stable* if it is both stable and globally pre-attractive. When every maximal solution to $\mathcal{H}$ is complete, the prefix "pre" can be dropped and the "classical" notions of stability and asymptotic stability are recovered.

Globally Pre-Asymptotically Stable Ω-Limit Sets Suppose all solutions to the hybrid system are bounded, there exists a compact set S such that all solutions eventually reach and remain in S, and there exists a neighborhood of S from which this convergence to S is uniform. (We are not assuming that the set S is forward invariant and thus it may not be stable.) Moreover, assume there is at least one complete solution starting in S. In this case, the hybrid system admits a nonempty compact set $\mathcal{A} \subset S$ that is globally pre-asymptotically stable. Indeed, one such set is

the so-called Ω-limit set of S, defined as

$$\Omega_{\mathcal{H}}(S) := \left\{ y \in \mathbb{R}^n \;\middle|\; \begin{array}{l} y = \lim_{i\to\infty} x_i(t_i, j_i)\,, \\ t_i + j_i \to \infty\,, (t_i, j_i) \in \operatorname{dom} x_i \\ x_i \text{is a solution to } \mathcal{H} \\ \text{with } x_i(0,0) \in S \end{array} \right\}.$$

In fact, this Ω-limit set is the smallest compact set in S that is globally pre-asymptotically stable.

To illustrate this concept, consider the temperature control system in Sect. "Introduction" with additional heater dynamics given by

$$\dot{h} = -3h + (2h_\Delta + h)u\,,$$

where h is the heater temperature and h_Δ is a constant that determines how hot the heater can get due to being on. That is, when the heater is "on" ($u = 1$), its temperature rises asymptotically towards h_Δ. There is a maximum temperature $h_{\bar{\sigma}} < h_\Delta$ for which the heater can operate safely, and another temperature $h_{\underline{\sigma}} < h_{\bar{\sigma}}$ corresponding to a temperature far enough below $h_{\bar{\sigma}}$ that it is considered safe to turn the heater back on. For the desired range of temperatures for T given by $T_{\min} = 70°\text{F}$, $T_{\max} = 80°\text{F}$ and $T_\Delta = 30°\text{F}$, let the overheating constant be $h_\Delta = 200°\text{F}$, the maximum safe temperature be $h_{\bar{\sigma}} = 150°\text{F}$, and the lower temperature $h_{\underline{\sigma}} = 50°\text{F}$. Then, to keep the temperature T in the desired range and prevent overheating, the following algorithm is used:

- When the heater is "on" ($u = 1$) and either $T \geq 80$ or $h \geq 150$, then turn the heater off ($u^+ = 0$).
- When the heater is "off " ($u = 0$) and $T \leq 70$ and $h \leq 50$, then turn the heater on ($u^+ = 1$).

These rules define the jump map for u, which is given by $u^+ = 1 - u$, and the jump set of the hybrid control system. The resulting hybrid closed-loop system, denoted by $\mathcal{H}_T$, is given by

$$\left.\begin{array}{l} \dot{T} = -T + T_0 + T_\Delta u \\ \dot{h} = -3h + (2h_\Delta + h)u \\ \dot{u} = 0 \end{array}\right\} \quad \left\{\begin{array}{l} u = 1, \\ (T \leq 80 \text{ and } h \leq 150) \\ \qquad \text{or} \\ u = 0, \\ (T \geq 70 \text{ or } h \geq 50) \end{array}\right.$$

$$\left.\begin{array}{l} T^+ = T \\ h^+ = h \\ u^+ = 1 - u \end{array}\right\} \quad \left\{\begin{array}{l} u = 1, (T \geq 80 \text{ or } h \geq 150) \\ \qquad \text{or} \\ u = 0, (T \leq 70 \text{ and } h \leq 50)\,. \end{array}\right.$$

For this system, the set

$$S := [70, 80] \times [0, 150] \times \{0, 1\} \tag{13}$$

is not forward invariant. Indeed, consider the initial condition $(T, h, u) = (70, 150, 0)$ which is not in the jump set, so there will be some time where the heater remains off, cooling the room to a value below 70°F. Nevertheless, all trajectories converge to the set S and a neighborhood of initial conditions around S produce solutions that reach the set S in a uniform amount of time. Thus, the set $\Omega_{\mathcal{H}_T}(S) \subset S$ is a compact globally pre-asymptotically stable set for the system $\mathcal{H}_T$.

Converse Lyapunov Theorems and Robustness

For purely continuous-time and discrete-time systems satisfying some regularity conditions, global asymptotic stability of a compact set implies the existence of a smooth Lyapunov function. Such results, known as converse Lyapunov theorems, establish a necessary condition for global asymptotic stability. For hybrid systems $\mathcal{H} = (f, C, G, D)$, the conditions for well-posedness also lead to a converse Lyapunov theorem.

If a compact set $\mathcal{A} \subset \mathbb{R}^n$ is globally pre-asymptotically stable for the hybrid system $\mathcal{H} = (f, C, G, D)$, then there exists a smooth Lyapunov function; that is, there exists a smooth function $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ that is positive definite with respect to $\mathcal{A}$, radially unbounded, and satisfies

$$\begin{aligned} \langle \nabla V(x), f(x) \rangle &\leq -V(x) \quad \forall x \in C\,, \\ \max_{g \in G(x)} V(g) &\leq \frac{V(x)}{2} \quad \forall x \in D\,. \end{aligned}$$

Converse Lyapunov theorems are not of mere theoretical interest as they can be used to characterize robustness of asymptotic stability. Suppose that $\mathcal{H} = (f, C, G, D)$ is well-posed and that $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is a smooth Lyapunov function for some compact set $\mathcal{A}$. The smoothness of V and the regularity of the data of $\mathcal{H}$ imply that V decreases along solutions when the data is perturbed. More precisely:

Global pre-asymptotic stability of the compact set $\mathcal{A} \subset \mathbb{R}^n$ for $\mathcal{H}$ is equivalent to semiglobal practical pre-asymptotic stability of $\mathcal{A}$ in the size of perturbations to $\mathcal{H}$, i.e., to the following: there exists a continuous, nondecreasing in the first argument, nonincreasing in second argument function $\beta : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with the property that $\lim_{s \searrow 0} \beta(s, t) = \lim_{t\to\infty} \beta(s, t) = 0$ and, for each $\varepsilon > 0$ and each compact set $K \subset \mathbb{R}^n$, there exists $\rho^* > 0$, such that for each perturbation level

$\rho \in (0, \rho^*]$ each solution x, $x(0,0) \in K$, to the ρ-perturbed hybrid system

$$\mathcal{H}_\rho: \quad x \in \mathbb{R}^n \quad \begin{cases} \dot{x} \in F_\rho(x)\,, & x \in C_\rho \\ x^+ \in G_\rho(x)\,, & x \in D_\rho\,, \end{cases}$$

where, for each $x \in \mathbb{R}^n$,

$$F_\rho(x) := \overline{\text{co}} f((x + \rho\mathbb{B}) \cap C) + \rho\mathbb{B}\,,$$
$$G_\rho(x) := \{v \in \mathbb{R}^n \,|\, v \in z + \rho\mathbb{B},\, z \in G((x + \rho\mathbb{B}) \cap D)\}\,,$$

and

$$C_\rho := \{z \in \mathbb{R}^n \,|\, (z + \rho\mathbb{B}) \cap C \neq \emptyset\}\,,$$
$$D_\rho := \{z \in \mathbb{R}^n \,|\, (z + \rho\mathbb{B}) \cap D \neq \emptyset\}\,,$$

satisfies

$$|x(t,j)|_{\mathcal{A}} \leq \max\{\beta(|x(0,0)|_{\mathcal{A}}, t+j), \varepsilon\}\ \forall (t,j) \in \operatorname{dom} x.$$

The above result can be readily used to derive robustness of (pre-)asymptotic stability to various types of perturbations, such as slowly-varying and weakly-jumping parameters, "average dwell-time" perturbations (see [16] for details), and temporal regularizations, as introduced in Subsect. "Zeno Solutions and Temporal Regularization". We clarify the latter robustness now.

Zeno Solutions, Temporal Regularization, and Robustness Some of the control systems we will design later will have discrete solutions that evolve in the set we are trying to asymptotically stabilize. So, these solutions do not affect asymptotic stability adversely, but they are somewhat problematic from an implementation point of view. We indicated in Subsect. "Zeno Solutions and Temporal Regularization" how these solutions arising in hybrid control systems can be removed via temporal regularization. Here we indicate how doing so does not destroy the asymptotic stability achieved, at least in a semi-global practical sense.

The assumption is that stabilization via hybrid control is achieved as a preliminary step. In particular, assume that there is a well-posed closed-loop hybrid system $\mathcal{H} := (f, G, C, D)$ with state $\xi \in \mathbb{R}^n$, and suppose that the compact set $\mathcal{A} \subset \mathbb{R}^n$ is globally pre-asymptotically stable. Following the prescription for a temporal regularization of a hybrid controller in Subsect. "Zeno Solutions and Temporal Regularization", we consider the hybrid system $\mathcal{H}_\delta := (\tilde{f}, C_\delta, \tilde{G}, D_\delta)$ with the state $x := (\xi, \tau) \in \mathbb{R}^{n+1}$, where $\delta \in (0,1)$ and

$$\tilde{f}(x) := f(\xi) \times \{1 - \tau\}\,,$$
$$\tilde{G}(x) := G(\xi) \times \{0\}\,,$$
$$C_\delta := \left(C \times \mathbb{R}_{\geq 0}\right) \cup (\mathbb{R}^n \times [0, \delta])\,,$$
$$D_\delta := D \times [\delta, 1]\,.$$

As observed before, the system $\mathcal{H}_0 = (\tilde{f}, C_0, \tilde{G}, D_0)$ has the compact set $\tilde{\mathcal{A}} := \mathcal{A} \times [0,1]$ globally pre-asymptotically stable. When $\delta > 0$, in each hybrid time domain of each solution, each time interval is at least δ seconds long, since $\dot{\tau} \leq 1$ for all $\tau \in [0, \delta]$. In particular, Zeno solutions, if there were any, have been eliminated. Regarding pre-asymptotic stability, note that

$$C_\delta \subset \left\{z \in \mathbb{R}^{n+1} \,\middle|\, (z + \delta\mathbb{B}) \cap C_0 \neq \emptyset\right\}$$

(with $\mathbb{B} \subset \mathbb{R}^{n+1}$), while $D_\delta \subset D_0$. Hence, following the discussion above, one can conclude that for $\mathcal{H}_\delta$, the set $\tilde{\mathcal{A}}$ is semi-globally practically asymptotically stable in the size of the temporal regularization parameter δ. Broadly speaking, temporal regularization does not destroy (practical) pre-asymptotic stability of $\mathcal{A}$.

Lyapunov Stability Theorem

A Lyapunov function is not only necessary for asymptotic stability but also sufficient. It is a convenient tool for establishing asymptotic stability because it eliminates the need to solve explicitly for solutions to the system. In its sufficiency form, the requirements of a Lyapunov function can be relaxed somewhat compared to the conditions of the previous subsection.

For a hybrid system $\mathcal{H} = (f, C, G, D)$ the compact set $\mathcal{A} \subset \mathbb{R}^n$ is globally pre-asymptotically stable if there exists a continuously differentiable function $V: \mathbb{R}^n \to \mathbb{R}_{\geq}$ that is positive definite with respect to $\mathcal{A}$, radially unbounded, and, with the definitions,

$$u_c(x) := \begin{cases} \langle \nabla V(x), f(x) \rangle & x \in C \\ -\infty & \text{otherwise} \end{cases} \tag{14}$$

$$u_d(x) := \begin{cases} \max_{g \in G(x)} V(g) - V(x) & x \in D \\ -\infty & \text{otherwise}\,, \end{cases} \tag{15}$$

satisfies

$$u_d(x) \leq 0 \quad \forall x \in \mathbb{R}^n \tag{16}$$

$$u_c(x) < 0\,, \quad u_d(x) < 0 \quad \forall x \in \mathbb{R}^n \setminus \mathcal{A}\,. \tag{17}$$

In light of the converse theorem of the previous subsection, this sufficient condition for global asymptotic stability is reasonable. Nevertheless, finding a Lyapunov function is often difficult to do. Thus, there is motivation for stability analysis tools that relax the Lyapunov conditions. There are several directions in which to go. One is in the direction of invariance principles, which are presented next.

Invariance Principles

An important tool to study the convergence of solutions to dynamical systems is LaSalle's invariance principle. LaSalle's invariance principle [35,36] states that bounded and complete solutions converge to the largest invariant subset of the set where the derivative or the difference (depending whether the system is continuous-time or discrete-time, respectively) of a suitable energy function is zero. In situations where the condition (17) holds with nonstrict inequalities, the invariance principle provides a tool to extract information about convergence of solutions.

By relying on the sequential compactness property of solutions in Subsect. "Conditions for Well-Posedness", several versions of LaSalle-like invariance principles can be stated for hybrid systems. Like for continuous-time and discrete-time systems, to make statements about the convergence of a solution one typically assumes that the solution is bounded, that its hybrid time domain is unbounded (i. e., the solution is complete), and that a Lyapunov function does not increase along it. To obtain information about the set to which the solution converges, an invariant set is to be computed. For hybrid systems, since solutions may not be unique, the standard concept of invariance needs to be adjusted appropriately.

Following [35], but in the setting of hybrid systems, we will insist that a (weakly) invariant set be both *weakly forward invariant* and *weakly backward invariant*. The word "weakly" indicates that only one solution, rather than all, needs to meet some invariance conditions. By requiring both forward and backward invariance we refine the sets to which solutions converge. For a given set $\mathcal{M}$ and a hybrid system $\mathcal{H}$, these notions were defined, in [54], as follows:

- *Forward invariance: if for each point $x^0 \in \mathcal{M}$ there exists at least one complete solution x to $\mathcal{H}$ that starts at x^0 and stays in the set $\mathcal{M}$ for all $(t, j) \in \operatorname{dom} x$.*
- *Backward invariance: if for each point $q \in \mathcal{M}$ and every positive number N there exists a point x^0 from which there exists at least one solution x to $\mathcal{H}$ and $(t^*, j^*) \in \operatorname{dom} x$ such that $x(t^*, j^*) = q$ and $x(t, j) \in \mathcal{M}$ for all $(t, j) \in \operatorname{dom} x$, $(t, j) \preceq (t^*, j^*)$.*

Then, the following invariance principle can be stated:

Let $V : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and suppose that $\mathcal{U} \subset \mathbb{R}^n$ is nonempty. Let x be a bounded and complete solution to a hybrid system $\mathcal{H} := (f, C, G, D)$. If x satisfies $x(t, j) \in \mathcal{U}$ for each $(t, j) \in \operatorname{dom} x$ and

$$u_c(z) \le 0, \quad u_d(z) \le 0 \quad \text{for all } z \in \mathcal{U},$$

then, for some constant $r \in V(\mathcal{U})$, the solution x approaches the largest weakly invariant set contained in

$$\left[u_c^{-1}(0) \cup \left(u_d^{-1}(0) \cap G\left(u_d^{-1}(0)\right)\right)\right] \cap V^{-1}(r) \cap \mathcal{U}. \tag{18}$$

Note that the statement and the conclusion of this invariance principle resemble the ones by LaSalle for differential/difference equations. In particular, the definition of the set (18) involves both the zero-level set of u_c and u_d as the continuous and discrete-time counterparts of the principle. For more details and other invariance principles for hybrid systems, see [54].

The invariance principle above leads to the following corollary on global pre-asymptotic stability:

For a hybrid system $\mathcal{H} = (f, C, G, D)$ the compact set $\mathcal{A} \subset \mathbb{R}^n$ is globally pre-asymptotically stable if there exists a continuously differentiable function $V : \mathbb{R}^n \to \mathbb{R}_{\ge}$ that is positive definite with respect to $\mathcal{A}$, radially unbounded, such that, with the definitions (14)–(15),

$$u_c(x) \le 0, \quad u_d(x) \le 0 \quad \forall x \in \mathbb{R}^n,$$

and, for every $r > 0$, the largest weakly invariant subset in (18) is empty.

This corollary will be used to establish global asymptotic stability in the control application in Subsect. "Source Localization".

Design Tools

Supervisors of Hybrid Controllers

In this section, we discuss how to construct a single, globally asymptotically stabilizing, well-posed hybrid controller from several individual well-posed hybrid controllers that behave well on particular regions of the state-space but that are not defined globally.

Suppose we are trying to control a nonlinear system

$$\dot{x} = f(x, u), \quad x \in C_0 \tag{19}$$

and suppose that we have constructed a finite family of well-posed hybrid controllers $\mathcal{K}_q$, that work well individually, on a particular region of the state space. We will make this more precise below.

For simplicity, the controllers will share the same state. This can be accomplished by embedding the states of the individual controllers into a common state space. Each

controller $\mathcal{K}_q$, with $q \in \mathcal{Q}$ and $\mathcal{Q}$ being a finite index set not containing 0, is given by

$$\mathcal{K}_q \begin{cases} u = \kappa_q(x,\eta) \\ \dot{\eta} = \phi_q(x,\eta)\,, & (x,\eta) \in C_q \\ \eta^+ \in \psi_q(x,\eta)\,, & (x,\eta) \in D_q. \end{cases} \tag{20}$$

The sets C_q and D_q are such that $(x,\eta) \in C_q \cup D_q$ implies $x \in C_0$.

The union of the regions over which these controllers operate is the region over which we want to obtain robust, asymptotic stability. We define this set as $\Theta := \cup_{q\in\mathcal{Q}} \left(C_q \cup D_q\right)$. To achieve our goal, we will construct a hybrid supervisor that makes decisions about which of the hybrid controllers should be used based on the state's location relative to a collection of closed sets $\Psi_q \subset C_q \cup D_q$ that cover Θ. The hybrid supervisor will have its own state $q \in \mathcal{Q}$. The composite controller, denoted $\mathcal{K}$ with flow set C and jump set D, should be such that 1) $C \cup D = \Theta \times \mathcal{Q}$, 2) all maximal solutions of the interconnection of $\mathcal{K}$ with (19), denoted $\mathcal{H}$, starting in $\Theta \times \mathcal{Q}$ are complete, 3) and the compact set $\mathcal{A} \times \mathcal{Q}$, a subset of $\Theta \times \mathcal{Q}$, is globally asymptotically stable for the system $\mathcal{H}$.

We now clarify what we mean by the family of hybrid controllers working well individually. Let $\mathcal{H}_q$ denote the closed-loop interconnection of the system (19) with the hybrid controller (20). For each $q \in \mathcal{Q}$, the solutions to the system $\mathcal{H}_q$ satisfy:

I. The set $\mathcal{A}$ is globally pre-asymptotically stable.
II. Each maximal solution is either complete or ends in

$$\left(\bigcup_{i\in\mathcal{Q},i>q} \Psi_i\right) \cup \overline{\Theta \setminus \left(C_q \cup D_q\right)}\,.$$

III. No maximal solution starting in Ψ_q reaches

$$\overline{\Theta \setminus \left[C_q \cup D_q \cup \left(\bigcup_{i\in\mathcal{Q},i>q} \Psi_i\right)\right]} \setminus \mathcal{A}\,.$$

Item III holds for free for the minimum index $q_{\min}$ since $\Psi_{q_{\min}} \subset C_{q_{\min}} \cup D_{q_{\min}}$ and $\cup_{i\in\mathcal{Q}} \Psi_i = \Theta$. The combination of the three items for the maximum index $q_{\max}$ implies that the solutions to $\mathcal{H}_{q_{\max}}$ that start in $\Psi_{q_{\max}}$ converge to $\mathcal{A}$.

Intuitively, the hybrid supervisor will attempt to reach its goal by guaranteeing completeness of solutions and making the evolution of q eventually monotonic while (x,η) does not belong to the set $\mathcal{A}$. In this way, the (x,η) component of the solutions eventually converges to $\mathcal{A}$ since q is eventually constant and because of the first assumption above. The hybrid supervisor can be content with sticking with controller q as long as $(x,\eta) \in C_q \cup D_q$, it can increment q if $(x,\eta) \in \cup_{i\in\mathcal{Q},i>q}\Psi_i$, and it can do anything it wants if $(x,\eta) \in \mathcal{A}$. These are the only three situations that should come up when starting from Ψ_q. Otherwise, the hybrid controller would be forced to decrease the value of q, taking away any guarantee of convergence. This provides the motivation for item III above. Due to a disturbance or unfortunate initialization of q, it may be that the state reaches a point that would not otherwise be reached from Ψ_q. From such conditions, the important thing is that the solution is either complete (and thus converges to $\mathcal{A}$) or else reaches a point where either q can be incremented or where q is allowed to be decremented. This is the motivation for item II above.

The individual hybrid controllers are combined into a single, well-posed hybrid controller $\mathcal{K}$ as follows: Define $\Phi_q := \cup_{i\in\mathcal{Q},i>q}\Psi_i$ and then

$$\mathcal{K} \begin{cases} u = \kappa_q(x,\eta) \\ \dot{\eta} = \phi_q(x,\eta)\,, & (x,\eta) \in \tilde{C}_q \\ \begin{bmatrix}\eta \\ q\end{bmatrix}^+ \in G_q(x,\eta)\,, & (x,\eta) \in \tilde{D}_q\,, \end{cases} \tag{21}$$

where

$$\tilde{D}_q := D_q \cup \Phi_q \cup \overline{\Theta \setminus (C_q \cup D_q)}\,,$$

$\tilde{C}_q$ is closed and satisfies

$$\overline{C_q \setminus \Phi_q} \subset \tilde{C}_q \subset C_q\,,$$

and the set-valued mapping G_q is constructed via the following definitions:

$$\begin{aligned} D_{q,a} &:= \Phi_q \\ \overline{D_q \setminus \Phi_q} &\subset D_{q,b} \subset D_q \\ D_{q,c} &:= \overline{\Theta \setminus \left(C_q \cup D_q \cup \Phi_q\right)} \end{aligned}$$

and

$$\begin{aligned} G_{q,a}(x,\eta) &:= \begin{bmatrix} \{\eta\} \\ \{i \in \mathcal{Q} \,|\, i > q\,,\ (x,\eta) \in \Psi_i\} \end{bmatrix} \\ G_{q,b}(x,\eta) &:= \begin{bmatrix} \psi_q(x,\eta) \\ \{q\} \end{bmatrix} \\ G_{q,c}(x,\eta) &:= \begin{bmatrix} \{\eta\} \\ \{i \in \mathcal{Q} \,|\, (x,\eta) \in \Psi_i\} \end{bmatrix} \\ G_q(x,\eta) &:= \bigcup_{\{j\in\{a,b,c\}\,,\,(x,\eta)\in D_{q,j}\}} G_{q,j}(x,\eta)\,. \end{aligned} \tag{22}$$

This hybrid controller is well-posed and induces complete solutions from $\Theta \times \mathcal{Q}$ and global asymptotic stability of the compact set $\mathcal{A} \times \mathcal{Q}$.

Uniting Local and Global Controllers As a simple illustration, consider a nonlinear control system $\dot{x} = f(x, u)$, $x \in \mathbb{R}^n$, and the task of globally asymptotically stabilizing the origin using state feedback while insisting on using a particular state feedback κ_2 in a neighborhood of the origin. In order to solve this problem, one can find a state feedback κ_1 that globally asymptotically stabilizes the origin and then combine it with κ_2 using a hybrid supervisor. Suppose that the feedback κ_2 is defined on a closed neighborhood of the origin, denoted C_2, and that if the state x starts in the closed neighborhood $\Psi_2 \subset C_2$ of the origin then the closed-loop solutions when using κ_2 do not reach the boundary of C_2. Then, using the notation of this section, we can take $\Psi_1 = C_1 = \mathbb{R}^n$ and $D_1 = D_2 = \emptyset$. With these definitions, the assumptions above are satisfied and a hybrid controller can be constructed to solve the posed problem. The controller need not use the additional variable η. Its data is defined as $G_q(x) := 3 - q$, $\tilde{D}_1 := \Psi_2 = \Phi_1$, $\tilde{D}_2 := \overline{\mathbb{R}^n \backslash C_2}$, $\tilde{C}_1 := \overline{C_1 \backslash \Psi_2}$ and $\tilde{C}_2 := C_2$.

Additional examples of supervisors will appear in the applications section later.

Patchy Control Lyapunov Functions

A key feature of (smooth) control Lyapunov functions (CLFs) is that their decrease along solutions to a given control system can be guaranteed by an appropriate choice of the control value, for each state value. It is known that, under mild assumptions on the control system, the existence of a CLF yields the existence of a robust (non hybrid) stabilizing feedback. It is also known that many nonlinear control systems do not admit a CLF. This can be illustrated by considering the question of robust stabilization of a single point on a circle, which faces a similar obstacle as the question of robust stabilization of the set $\mathcal{A} = \{0, 1\}$ for the control system on $\mathbb{R}$ given by $\dot{x} = f(x, u) := u$. Any differentiable function on $\mathbb{R}$ that is positive definite with respect to $\mathcal{A}$ must have a maximum in the interval $(0, 1)$. At such a maximum, say $\bar{x}$, one has $\nabla V(\bar{x}) = 0$ and no choice of u can lead to $\langle \nabla V(\bar{x}), f(\bar{x}, u) \rangle < 0$.

(Smooth) patchy control Lyapunov functions (PCLFs) are, broadly speaking, objects consisting of several local CLFs the domains of which cover $\mathbb{R}^n$ and have certain weak invariance properties. PCLFs turn out to exist for far broader classes of nonlinear systems than CLFs, especially if an infinite number of patches (i. e., of local CLFs) is allowed. They also lead to robust hybrid stabilizing feedbacks. This will be outlined below. A brief illustration of the concept, for the control system on $\mathbb{R}$ mentioned above, would be to consider functions $V_1(x) = x^2$ on $(-\infty, 2/3)$ and $V_2(x) = (x-1)^2$ on $(1/3, \infty)$. These functions are local CLFs for the points, respectively, 0 and 1; their domains cover $\mathbb{R}$; and for each function, an appropriate choice of control will not only lead to the function's decrease, but will also ensure that solutions starting in the function's domain will remain there.

While the example just mentioned outlines a general idea of a PCLF, the definition is slightly more technical. For the purposes of this article, a *smooth patchy control Lyapunov function* for a nonlinear system

$$\dot{x} = f(x, u) \quad x \in \mathbb{R}^n \,,\; u \in U \subset \mathbb{R}^m \tag{23}$$

with respect to the compact set $\mathcal{A}$ consists of a finite set $Q \subset \mathbb{Z}$ and a collection of functions V_q and sets Ω_q, Ω'_q for each $q \in Q$, such that:

(i) $\{\Omega_q\}_{q \in Q}$ and $\{\Omega'_q\}_{q \in Q}$ are families of nonempty open subsets of $\mathbb{R}^n$ such that

$$\mathbb{R}^n = \bigcup_{q \in Q} \Omega_q = \bigcup_{q \in Q} \Omega'_q \,,$$

and for all $q \in Q$, the unit (outward) normal vector to Ω_q is continuous on $\partial\Omega_q \setminus \bigcup_{i>q} \Omega'_i$, and

$$\overline{\Omega'_q} \subset \Omega_q \,;$$

(ii) for each q, V_q is a smooth function defined on a neighborhood of $\overline{\Omega_q \setminus \bigcup_{i>q} \Omega'_i}$;

and the following conditions are met: There exist a continuous, positive definite function $\alpha \colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, and positive definite, radially unbounded functions $\underline{\gamma}$, $\bar{\gamma}$ such that

(iii) for all $q \in Q$, all $x \in \Omega_q \setminus \bigcup_{i>q} \Omega'_i$,

$$\underline{\gamma}(|x|_{\mathcal{A}}) \leq V_q(x) \leq \bar{\gamma}(|x|_{\mathcal{A}}) \,;$$

(iv) for all $q \in Q$, all $x \in \Omega_q \setminus \bigcup_{i>q} \Omega'_i$, there exists $u_{q,x} \in U$ such that

$$\langle \nabla V_q(x), f(x, u_{q,x}) \rangle \leq -\alpha(|x|_{\mathcal{A}}) \,;$$

(v) for all $q \in Q$, all $x \in \partial\Omega_q \setminus \bigcup_{i>q} \Omega'_i$, the $u_{q,x}$ of (iii) can be chosen such that

$$\langle n_q(x), f(x, u_{q,x}) \rangle \leq -\alpha(|x|_{\mathcal{A}}),$$

where $n_q(x)$ is the unit (outward) normal vector to Ω_q at x.

Suppose that, for each $x, v \in \mathbb{R}^n$ and $c \in \mathbb{R}$, the set $\{u \in U \mid \langle v, f(x,u)\rangle \le c\}$ is convex, as always holds if $f(x,u)$ is affine in u and U is convex. For each $q \in Q$ let

$$C_q = \overline{\Omega_q \setminus \bigcup_{i>q} \Omega'_i}$$

and

$$\Psi_q = \overline{\Omega'_q \setminus \bigcup_{i\in Q, i>q} \Omega'_i}\,.$$

It can be shown, in part via arguments similar to those one would use when constructing a feedback from a CLF, that for each $q \in Q$ there exists a continuous mapping

$$k_q \colon C_q \to U$$

such that, for all $x \in C_q$,

$$\langle \nabla V_q(x), f(x, k_q(x))\rangle \le -\frac{\alpha(|x|_{\mathcal{A}})}{2}\,;$$

all maximal solutions to

$$\dot{x} = f(x, k_q(x))$$

are either complete or end in

$$\left(\bigcup_{i>q, i\in Q} \Psi_i\right) \cup \overline{\mathbb{R}^n \setminus C_q}\,;$$

and no maximal solution starting in Ψ_q reaches

$$\overline{\mathbb{R}^n \setminus \left(C_q \cup \bigcup_{i\in Q, i>q} \Psi_i\right)}\,.$$

The feedbacks k_q can now be combined in a hybrid feedback, by taking $D_q = \emptyset$ for each $q \in Q$, and following the construction of Subsect. "Supervisors of Hybrid Controllers". Indeed, the properties of maximal solutions just mentioned ensure conditions I, II and III of that section; the choice of C_q and Ψ_q also ensures that $\Psi_q \subset C_q$ and the union of Ψ_q's covers $\mathbb{R}^n$.

Among other things, this construction illustrates that the idea of hybrid supervision of hybrid controllers applies also to combining standard, non hybrid controllers.

Applications

In this section, we make use of the following:

- For vectors in $\mathbb{R}^2$, we will use the following multiplication rule, conjugate rule, and identity element:

$$z \otimes x := \begin{bmatrix} z_1x_1 - z_2x_2 \\ z_2x_1 + z_1x_2 \end{bmatrix}, \quad x^c := \begin{bmatrix} x_1 \\ -x_2 \end{bmatrix}, \quad \mathbf{1} := \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

  The multiplication rule is commutative, associative, and distributive. Note that $x = \mathbf{1} \otimes x = x \otimes \mathbf{1}$ and note that $x^c \otimes x = x \otimes x^c = |x|^2\mathbf{1}$. Also, $(z \otimes x)^c = x^c \otimes z^c$.
- For vectors in $\mathbb{R}^4$, we will use the following multiplication rule, conjugate rule, and identity element (vectors are partitioned as $x = [x_1\, x_2^{\mathrm{T}}]^T$ where $x_2 \in \mathbb{R}^3$):

$$z \otimes x = \begin{bmatrix} z_1x_1 - z_2^{\mathrm{T}}x_2 \\ z_2x_1 + z_1x_2 + z_2 \times x_2 \end{bmatrix},$$

$$x^c = \begin{bmatrix} x_1 \\ -x_2 \end{bmatrix},$$

$$\mathbf{1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

  The multiplication rule is associative and distributive but not necessarily commutative. Note that $x = \mathbf{1} \otimes x = x \otimes \mathbf{1}$ and $x^c \otimes x = x \otimes x^c = |x|^2\mathbf{1}$. Also, $(z \otimes x)^c = x^c \otimes z^c$.

Overcoming Stabilization Obstructions

Global Stabilization and Tracking on the Unit Circle In this section, we consider stabilization and tracking control of the constrained system

$$\dot{\xi} = \xi \otimes v(\omega)\,, \quad v(\omega) := \begin{bmatrix} 0 \\ \omega \end{bmatrix} \quad \xi \in S^1\,, \tag{24}$$

where S^1 denotes the unit circle and $\omega \in \mathbb{R}$ is the control variable. Notice that S^1 is invariant regardless of the choice of ω since $\langle \xi, \xi \otimes v(\omega)\rangle = 0$ for all $\xi \in S^1$ and all $\omega \in \mathbb{R}$. This model describes the evolution of orientation angle of a rigid body in the plane as a function of the angular velocity ω, which is the control variable. We discuss robust, global asymptotic stabilization and tracking problems which cannot be solved with classical feedback control, even when discontinuous feedback laws are allowed, but can be solved with hybrid feedback control.

Stabilization First, we consider the problem of stabilizing the point $\xi = \mathbf{1}$. We note that the (classical) feedback control $\omega = -\xi_2$ would almost solve this problem. We would have $\dot{\xi}_1 = \xi_2^2 = 1 - \xi_1^2$ and the derivative of the energy function $V(\xi) := 1 - \xi_1$ would satisfy

$$\langle \nabla V(\xi), \xi \otimes v(\omega)\rangle = -\left(1 - \xi_1^2\right)\,.$$

We note that the energy will remain constant if ξ starts at $\pm\mathbf{1}$. Thus, since the goal was (robust) global asymptotic stability, this feedback does not achieve the desired goal. One could also consider the discontinuous feedback $\omega = -\mathrm{sgn}(\xi_2)$ where the function "sgn" is defined arbitrarily in the set $\{-1, 1\}$ when its argument is zero. This feedback is not robust to arbitrarily small measurement noise which can keep the trajectories of the system arbitrarily close to the point $\xi = -\mathbf{1}$ for all time. To visualize this, note that from points on the circle with $\xi_2 < 0$ and close to $\xi = -\mathbf{1}$, this control law steers the trajectories towards $\xi = \mathbf{1}$ counterclockwise, while from points on the circle with $\xi_2 > 0$ and close to $\xi = -\mathbf{1}$, it steers the trajectories towards $\mathbf{1}$ clockwise. Then, from points on the circle arbitrarily close to $\xi = -\mathbf{1}$, one can generate an arbitrarily small measurement noise signal e that changes sign appropriately so that $-\mathrm{sgn}(\xi_2 + e)$ is always pushing trajectories towards $-\mathbf{1}$.

In order to achieve a robust, global asymptotic stability result, we consider a hybrid controller that uses the controller $\omega = -\xi_2$ when the state is not near $-\mathbf{1}$ and uses a controller that drives the system away from $-\mathbf{1}$ when it is near that point. One way to accomplish the second task is to build an almost global asymptotic stabilizer for a point different from $-\mathbf{1}$ such that the basin of attraction contains all points in a neighborhood of $-\mathbf{1}$. For example, consider the feedback controller $\omega = -\xi_1 =: \kappa_1(\xi)$ which would almost globally asymptotically stabilize the point $\zeta := (0, -1)$ with the only point not in the basin of attraction being the point ζ^c.

Strangely enough, each of the two controllers can be thought of as globally asymptotically stabilizing the point $\mathbf{1}$ if their domains are limited. In particular, let the domain of applicability for the controller $\omega = -\xi_1$ be

$$C_1 := S^1 \cap \{\xi \,|\, \xi_1 \le -1/3\} \ ,$$

and let the domain of applicability for the controller $\omega = -\xi_2$ be

$$C_2 := S^1 \cap \{\xi \,|\, \xi_1 \ge -2/3\} \ .$$

Notice that $C_1 \cup C_2 = S^1 =: \Theta$. Thus, we are in a situation where a hybrid supervisor, as discussed in Subsect. "Supervisors of Hybrid Controllers", may be able to give us a hybrid, global asymptotic stabilizer. (There is no state η in the controllers we are working with here.) Let us take

$$\Psi_1 := C_1 \ , \quad \Psi_2 := \overline{S^1 \backslash C_1} \ .$$

We have $\Psi_1 \cup \Psi_2 = S^1$. Next, we check the assumptions of Subsect. "Supervisors of Hybrid Controllers". For each $q \in \{1, 2\}$, the solutions of $\mathcal{H}_q$ (the system we get by using $\omega = \kappa_q(\xi)$ and restricting the flow to C_q), are such that the point $\mathbf{1}$ is globally pre-asymptotically stable. For $q = 1$, this is because there are no complete solutions and $\mathbf{1}$ does not belong to C_1. For $q = 2$, this is because C_2 is a subset of the basin of attraction for $\mathbf{1}$. We note that every maximal solution to $\mathcal{H}_1$ ends in Ψ_2. Every maximal solution to $\mathcal{H}_2$ is complete and every maximal solution to $\mathcal{H}_2$ starting in Ψ_2 does not reach $\overline{S^1 \backslash C_2}$. Thus, the assumptions for a hybrid supervisor are in place.

We follow the construction in Subsect. "Supervisors of Hybrid Controllers" to define the hybrid supervisor that combines the feedback laws κ_1 and κ_2. We take

$$\begin{aligned}
\omega &:= \kappa_q(\xi) \\
\tilde{C}_q &:= C_q \\
\tilde{D}_1 &:= \Psi_2 \cup \left(\overline{S^1 \backslash C_1}\right) = \overline{S^1 \backslash C_1} \\
\tilde{D}_2 &:= \overline{S^1 \backslash C_2} \ .
\end{aligned}$$

For this particular problem, jumps toggle the mode q in the set $\{1, 2\}$. Thus, the jump map G_q can be simplified to $G_q := 3 - q$.

Tracking Let $\zeta : \mathbb{R}_{\ge 0} \to S^1$ be continuously differentiable. Suppose we want to find a hybrid feedback controller so that the state of (24) tracks the signal ζ. This problem can be reduced to the stabilization problem of the previous section. Indeed, first note that $\zeta^c \otimes \zeta = \mathbf{1}$ and thus the following properties hold:

$$\begin{aligned}
\dot{\zeta}^c \otimes \zeta &= -\zeta^c \otimes \dot{\zeta} \\
\zeta^c \otimes \dot{\zeta} &= \begin{bmatrix} 0 \\ \zeta_1 \dot{\zeta}_2 - \zeta_2 \dot{\zeta}_1 \end{bmatrix} \ .
\end{aligned} \tag{25}$$

Then, with the coordinate transformation $\xi = z \otimes \zeta$, we have:

I. By multiplying the coordinate transformation on the right by ζ^c we get $z = \xi \otimes \zeta^c$ and $z \otimes z^c = \xi \otimes \xi^c = \mathbf{1}$, so that $z \in S^1$.
II. $\xi = \zeta \iff z = \mathbf{1}$.
III. The derivative of z satisfies

$$\begin{aligned}
\dot{z} &= \dot{\xi} \otimes \zeta^c + \xi \otimes \dot{\zeta}^c \\
&= \xi \otimes \nu(\omega) \otimes \zeta^c + \xi \otimes \dot{\zeta}^c \otimes \zeta \otimes \zeta^c \\
&= \xi \otimes \left[\nu(\omega) + \dot{\zeta}^c \otimes \zeta\right] \otimes \zeta^c \\
&= z \otimes \zeta \otimes \left[\nu(\omega) - \zeta^c \otimes \dot{\zeta}\right] \otimes \zeta^c \ .
\end{aligned}$$

Our desire is to pick ω so that we have

$$\dot{z} = z \otimes \nu(\Omega)$$

and then to choose Ω to globally asymptotically stabilize the point $z = \mathbf{1}$. Due to (25) and the properties of multiplication, the vectors $\zeta^c \otimes \dot{\zeta}$ and $\zeta^c \otimes v(\Omega) \otimes \zeta$ are in the range of $v(\omega)$. So, we can pick $v(\omega) = \zeta^c \otimes \dot{\zeta} + \zeta^c \otimes v(\Omega) \otimes \zeta$ to achieve the robust, global tracking goal. In fact, since the multiplication operation is commutative in $\mathbb{R}^2$, this is equivalent to the feedback $v(\omega) = \zeta^c \otimes \dot{\zeta} + v(\Omega)$.

Global Stabilization and Tracking for Unit Quaternions In this section, we consider stabilization and tracking control of the constrained system

$$\dot{\xi} = \xi \otimes v(\omega)\,, \quad v(\omega) := \begin{bmatrix} 0 \\ \omega \end{bmatrix} \quad \xi \in S^3\,, \tag{26}$$

where S^3 denotes the hypersphere in $\mathbb{R}^4$ and $\omega \in \mathbb{R}^3$ is the control variable. Notice that S^3 is invariant regardless of the choice of ω since $\langle \xi, \xi \otimes v(\omega) \rangle = 0$ for all $\xi \in S^3$ and all $\omega \in \mathbb{R}^3$. This model describes the evolution of orientation angle of a rigid body in space as a function of angular velocities ω, which are the control variables. The state ξ corresponds to a unit quaternion that can be used to characterize orientation. We discuss robust, global asymptotic stabilization and tracking problems which cannot be solved with classical feedback control, even when discontinuous feedback laws are allowed, but can be solved with hybrid feedback control.

Stabilization First, we consider the problem of stabilizing the point $\xi = \mathbf{1}$. We note that the (classical) feedback control

$$\omega := \begin{bmatrix} 0 & I \end{bmatrix} v(-\xi_2) =: \kappa_2(\xi)$$

(i. e., $\omega = -\xi_2$ where ξ_2 refers to the last three components of the vector ξ) would almost solve this problem. We would have $\dot{\xi}_1 = \xi_2^{\mathrm{T}} \xi_2 = 1 - \xi_1^2$ and the derivative of the energy function $V(\xi) := 1 - \xi_1$ would satisfy

$$\langle \nabla V(\xi), \xi \otimes v(\omega) \rangle = -\left(1 - \xi_1^2\right)\,.$$

We note that the energy will remain constant if ξ starts at $\pm\mathbf{1}$. Thus, since the goal is (robust) global asymptotic stabilization, this feedback does not achieve the desired goal. One could also consider the discontinuous feedback $\omega = -\mathrm{sgn}(\xi_2)$ where the function "sgn" is the componentwise sign and each component is defined arbitrarily in the set $\{-1, 1\}$ when its argument is zero. This feedback is not robust to arbitrarily small measurement noise which can keep the trajectories of the system arbitrarily close to the point $-\mathbf{1}$.

In order to achieve a robust, global asymptotic stability result, we consider a hybrid controller that uses the controller above when the state is not near $-\mathbf{1}$ and uses a controller that drives the system away from $-\mathbf{1}$ when it is near that point. One way to accomplish the second task is to build an almost global asymptotic stabilizer for a point different from $-\mathbf{1}$ and so that the basin of attraction contains all points in a neighborhood of $-\mathbf{1}$. For example, consider stabilizing the point $\zeta := (0, -1, 0, 0)^T$ using the feedback controller (for more details see the next subsection)

$$z = \xi \otimes \zeta^c\,, \quad \omega = \begin{bmatrix} 0 & I \end{bmatrix} \left(\zeta^c \otimes v(-z_2) \otimes \zeta\right) =: \kappa_1(\xi)$$

(i. e. $\omega = (-\xi_1, \xi_4, -\xi_3)^T$ where now the subscripts refer to the individual components of ξ). This feedback would almost globally asymptotically stabilize the point ζ with the only point not in the basin of attraction being the point ζ^c.

The two feedback laws, $\omega = \kappa_2(\xi)$ and $\omega = \kappa_1(\xi)$, are combined into a single hybrid feedback law via the hybrid supervisor approach given in Subsect. "Supervisors of Hybrid Controllers". In fact, the construction is just like the construction for the case of stabilization on a circle with the only difference being that S^1 is replaced everywhere by S^3.

Tracking Let $\zeta : \mathbb{R}_{\geq 0} \to S^3$ be continuously differentiable. Suppose we want to find a hybrid feedback controller so that the state of (26) tracks the signal ζ. This problem can be reduced to the stabilization problem of the previous section. Indeed, first note that $\zeta^c \otimes \zeta = \mathbf{1}$ and thus the following properties hold:

$$\begin{aligned} \dot{\zeta}^c \otimes \zeta &= -\zeta^c \otimes \dot{\zeta} \\ \zeta^c \otimes \dot{\zeta} &= \begin{bmatrix} 0 \\ \zeta_1 \dot{\zeta}_2 - \dot{\zeta}_1 \zeta_2 - \zeta_2 \times \dot{\zeta}_2 \end{bmatrix}\,. \end{aligned} \tag{27}$$

Then, with the coordinate transformation $\xi = z \otimes \zeta$, we have:

I. By multiplying the coordinate transformation on the right by ζ^c we get $z = \xi \otimes \zeta^c$ and $z^c \otimes z = \xi^c \otimes \xi = \mathbf{1}$ so that $z \in S^3$.
II. $\xi = \zeta \iff z = \mathbf{1}$.
III. The derivative of z satisfies

$$\begin{aligned} \dot{z} &= \dot{\xi} \otimes \zeta^c + \xi \otimes \dot{\zeta}^c \\ &= \xi \otimes \widehat{\omega} \otimes \zeta^c + \xi \otimes \dot{\zeta}^c \otimes \zeta \otimes \zeta^c \\ &= \xi \otimes \left[\widehat{\omega} + \dot{\zeta}^c \otimes \zeta\right] \otimes \zeta^c \\ &= z \otimes \zeta \otimes \left[\widehat{\omega} - \zeta^c \otimes \dot{\zeta}\right] \otimes \zeta^c\,. \end{aligned}$$

Our desire is to pick ω so that we have

$$\dot{z} = z \otimes \nu(\Omega)$$

and then to choose Ω to globally asymptotically stabilize the point $z = \mathbf{1}$. Due to (27) and the properties of multiplication, the vectors $\zeta^c \otimes \dot{\zeta}$ and $\zeta^c \otimes \nu(\Omega) \otimes \zeta$ are in the range of $\nu(\omega)$. So, we can pick $\nu(\omega) = \zeta^c \otimes \dot{\zeta} + \zeta^c \otimes \nu(\Omega) \otimes \zeta$ to achieve the robust, global tracking goal.

Stabilization of a Mobile Robot

Consider the global stabilization problem for a model of a unicycle or mobile robot, given as

$$\begin{aligned} \dot{x} &= \xi \vartheta \\ \dot{\xi} &= \xi \otimes \nu(\omega) \end{aligned} \tag{28}$$

where $x \in \mathbb{R}^2$ denotes planar position from a reference point (in meters), $\xi \in S^1$ denotes orientation, $\vartheta \in \mathcal{V} := [-3, 30]$ denotes velocity (in meters per second), and $\omega \in [-4, 4]$ denotes angular velocity (in radians per second). Both ϑ and ω are control inputs. Due to the specification of the set $\mathcal{V}$, the vehicle is able to move more rapidly in the forward direction than in the backward direction. We define $\mathcal{A}_0$ to be the point $(x, \xi) = (0, \mathbf{1})$. The controllers below will all use a discrete state $p \in \mathcal{P} := \{-1, 1\}$. We take $\mathcal{A} := \mathcal{A}_0 \times \mathcal{P}$ and $\Theta := \mathbb{R}^2 \times S^1 \times \mathcal{P}$.

This system also can be modeled as

$$\begin{aligned} \dot{x} &= \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} \vartheta \\ \dot{\theta} &= \omega \end{aligned}$$

where $\theta = 0$ corresponds to $\xi = \mathbf{1}$ and $\theta > 0$ is in the counterclockwise direction. The set $\mathcal{A}_0$ in these coordinates is given by the set $\{0\} \times \{\theta | \theta = 2k\pi \ , \ k \in \mathbb{Z}\}$. Even for the point $(x, \theta) = (0, 0)$, this control system fails Brockett's well-known condition for robust local asymptotic stabilization by classical (even discontinuous) time-invariant feedback [9,26,52]. Nevertheless, for the control system (28), the point $(0, \mathbf{1})$ can be robustly, globally asymptotically stabilized by hybrid feedback. This is done by building three separate hybrid controllers and combining them with a supervisor. The three controllers are the following:

- The first hybrid controller, $\mathcal{K}_1$, uses $\vartheta = \mathrm{Proj}_{\mathcal{V}}(k_1 \xi^T x)$, where $k_1 < 0$ and $\mathrm{Proj}_{\mathcal{V}}$ denotes the projection onto $\mathcal{V}$, while the feedback for ω is given by the hybrid controller in Subsect. "Global Stabilization and Tracking on the Unit Circle" for tracking on the unit circle with reference signal for ξ given by $-x/|x|$. The two different values for q in that controller should be associated with the two values in $\mathcal{P}$. The particular association does not matter. Note that the action of the tracking controller causes the vehicle eventually to use positive velocity to move x toward zero. The controller's flow and jump sets are such that

$$C_1 \cup D_1 = \left\{ x \in \mathbb{R}^2 \, ||x| \geq \varepsilon_{11} \right\} \times S^1 \times \mathcal{P}$$

  where $\varepsilon_{11} > 0$, and C_1, D_1 are constructed from the hybrid controller in Subsect. "Global Stabilization and Tracking on the Unit Circle" for tracking on the unit circle.
- The second hybrid controller, $\mathcal{K}_2$, uses $\vartheta = \mathrm{Proj}_{\mathcal{V}}(k_2 \xi^T x)$, $k_2 \leq 0$, while the feedback for ω is given as in Subsect. "Global Stabilization and Tracking on the Unit Circle" for stabilization of the point $\mathbf{1}$ on the unit circle. Again, the q values of that controller should be associated with the values in $\mathcal{P}$ and the particular association does not matter. The controller's flow and jump sets are such that

$$\begin{aligned} C_2 \cup D_2 = \big(\big(&\left\{ x \in \mathbb{R}^2 \, ||x| \leq \varepsilon_{21} \right\} \times S_1 \big) \\ &\cap \left\{ (x, \xi) \, \middle| 1 - \xi_1 \geq \varepsilon_{22} |x|^2 \right\} \big) \times \mathcal{P} \, , \end{aligned}$$

  where $\varepsilon_{21} > \varepsilon_{11}$, $\varepsilon_{22} > 0$, and C_2, D_2 are constructed from the hybrid controller in Subsect. "Global Stabilization and Tracking on the Unit Circle" for stabilization of the point $\mathbf{1}$ on the unit circle.
- The third hybrid controller, $\mathcal{K}_3$, uses $\vartheta = \mathrm{Proj}_{\mathcal{V}}(k_3 \xi^T x)$, $k_3 < 0$, while the feedback for ω is hybrid as defined below. The controller's flow and jump sets are designed so that

$$\begin{aligned} C_3 \cup D_3 = \big(&\{ x \, ||x| \leq \varepsilon_{31} \} \times S^1 \big) \\ &\cap \left\{ (x, \xi) \, \middle| 1 - \xi_1 \leq \varepsilon_{32} |x|^2 \right\} \times \mathcal{P} =: \Lambda_3 \, , \end{aligned}$$

  where $\varepsilon_{31} > \varepsilon_{21}$ and $\varepsilon_{32} > \varepsilon_{22}$. The control law for ω is given by $\omega = pk$, where $k > 0$ and the discrete state p has dynamics given by

$$\dot{p} = 0, \quad p^+ = -p \, .$$

  The flow and jump sets are given by

$$\begin{aligned} C_3 := &\Lambda_3 \cap \big(\{ \sigma(p) \xi_2 \leq 0 \} \\ &\cup \left\{ \sigma(p) \xi_2 \geq 0, 1 - \xi_1 \leq \varepsilon_{22} |x|^2 \right\} \big) \\ D_3 := &\overline{\Lambda_3 \backslash C_3} \, . \end{aligned}$$

This design accomplishes the following: controller $\mathcal{K}_1$ makes ξ track $-x/|x|$ as long as $|x|$ is not too small, and

thus the vehicle is driven towards $x = 0$ eventually using only positive velocity; controller $\mathcal{K}_2$ drives ξ towards $\mathbf{1}$ to get the orientation of the vehicle correct; and controller $\mathcal{K}_3$ stabilizes ξ to $\mathbf{1}$ in a persistently exciting manner so that ϑ can be used to drive the vehicle to the origin.

This control strategy is coordinated through a supervisor by defining

$$\begin{aligned}\Psi_1 &:= C_1 \cup D_1\\ \Psi_2 &:= \left(\overline{\Theta \backslash \Psi_1}\right) \cap (C_2 \cup D_2)\\ \Psi_3 &:= \left(\left(\{x \,|\, |x| \le \varepsilon_{21}\} \times S^1\right)\right.\\ &\qquad \left. \cap \left\{(x,\xi) \,\middle|\, 1 - \xi_1 \le \varepsilon_{22}|x|^2\right\}\right) \times \mathcal{P}\,.\end{aligned}$$

It can be verified that $\cup_{q\in\mathcal{Q}} \Psi_q = \Theta$ and that the conditions in Subsect. "Supervisors of Hybrid Controllers" for a successful supervisor are satisfied.

Figure 8 depicts simulation results of the mobile robot with the hybrid controller proposed above for global asymptotic stabilization of $\mathcal{A} \times \mathcal{Q}$. From the initial condition $x(0,0) = (10,10)$ (in meters), $\xi(0,0)$ corresponding to an angle of $\frac{\pi}{4}$ radians, the mobile robot backs up using controller $\mathcal{K}_1$ until its orientation ξ corresponds to about $\frac{3\pi}{4}$ radians, at which x is approximately $(10, 9.5)$. The green $\star$ denotes a jump of the hybrid controller $\mathcal{K}_1$. From this configuration, the mobile robot is steered towards a neighborhood of the origin with orientation given by $-x/|x|$. About a fifth of a meter away from it, a jump of the hybrid supervisor connects controller $\mathcal{K}_3$ to the vehicle input (the location at which the jump occurs is denoted by the red $\star$). Note that the trajectory is such that controller $\mathcal{K}_2$ is bypassed since at the jump of the hybrid supervisor, while $|x|$ is small the orientation ξ is such that the system state does not belong to Ψ_2 but to Ψ_3; that is, the orientation is already close enough to $\mathbf{1}$ at that jump. Figure 8a shows a zoomed version of the x trajectory in Fig. 8b. During this phase, controller $\mathcal{K}_3$ is in closed-loop with the mobile robot. The vehicle is steered to the origin with orientation close to $\mathbf{1}$ by a sequence of "parking" maneuvers. Note that after about seven of those maneuvers, the vehicle position is close to $(0, 0.01)$ with almost the desired orientation.

Source Localization

Core of the Algorithm Consider the problem of programming an autonomous vehicle to find the location of a maximum for a continuously differentiable function by noting how the function values change as the vehicle moves. Like before, the vehicle dynamics are given by

$$\begin{aligned}\dot{x} &= \xi\vartheta\\ \dot{\xi} &= \xi \otimes \nu(\omega)\,, \quad \xi \in S^1\,.\end{aligned}$$

The vehicle is to search for the maximum of the function $\varphi\colon \mathbb{R}^2 \to \mathbb{R}$. The function is assumed to be such that its maximum is unique, denoted x^*, that $\nabla\varphi(x) = 0$ if and only if $x = x^*$, and the union of its level sets over any interval of the form $[c, \infty)$, $c \in \mathbb{R}$, yields a compact set. For simplicity, we will assume that the sign of the derivative of the function φ in the direction $\dot{x} = \xi\vartheta$ is available as a measurement. We will discuss later how to approximate this quantity by considering the changes in the value of the function φ along solutions. We also assume that the vehi-

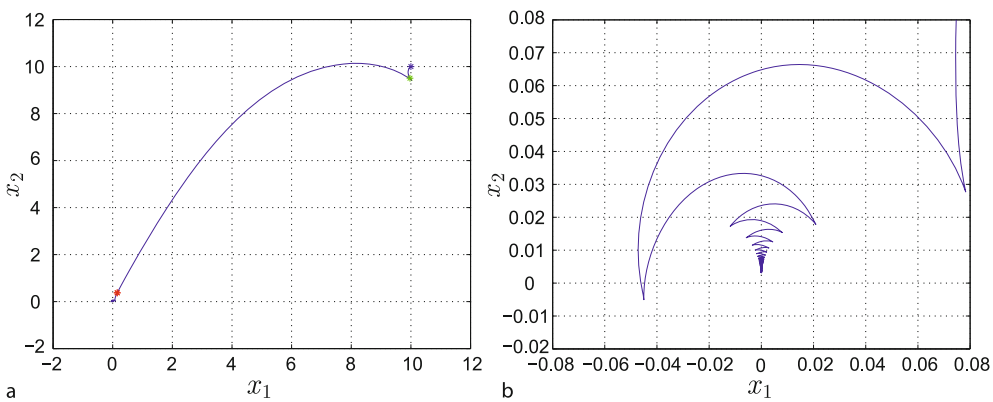

Hybrid Control Systems, Figure 8

Global stabilization of a mobile robot to the origin with orientation (1, 0). Vehicle starts at $x(0,0) = (10,10)$ (in meters) and $\xi(0,0)$ corresponding to an angle of $\frac{\pi}{4}$ radians. a The vehicle is initially steered to a neighborhood of the origin with orientation $-x/|x|$. At about 1/5 meters away from it, controller $\mathcal{K}_3$ is enabled to accomplish the stabilisation task. b Zoomed version of trajectory in a around the origin. Controller $\mathcal{K}_3$ steers the vehicle to $x = (0,0)$ and $\xi = 1$ by a sequence of "parking" maneuvers

cle's angle, ξ, can make jumps, according to

$$\xi^+ = \xi \otimes \zeta\,, \quad (\xi,\zeta) \in S^1 \times S^1\,.$$

We will discuss later how to solve the problem when the angle ξ cannot change discontinuously.

We propose the dynamic controller

$$\left.\begin{aligned} \vartheta &= \bar{\vartheta} \\ \begin{bmatrix} \dot{z} \\ \omega \end{bmatrix} &= \phi(z) \end{aligned}\right\} \quad (x,\xi,z) \in C$$

$$\left.\begin{bmatrix} z^+ \\ \zeta \end{bmatrix} \in \psi(z) \right\} \quad (x,\xi,z) \in D$$

where $\bar{\vartheta}$ is a positive constant,

$$C := \left\{(x,\xi,z) \,\middle|\, \langle \nabla\varphi(x), \xi\bar{\vartheta}\rangle \geq 0\,,\ \xi \in S^1\,,\ z \in \Upsilon\right\}$$
$$D := \left\{(x,\xi,z) \,\middle|\, \langle \nabla\varphi(x), \xi\bar{\vartheta}\rangle \leq 0\,,\ \xi \in S^1\,,\ z \in \Upsilon\right\}\,,$$

(note that C and D use information about the sign of the derivative of φ in the direction of the flow of x) and the required properties for the set Υ, the function ϕ, and the set-valued mapping ϕ are as follows:

I. (a) The set Υ is compact.
   (b) The maximal solutions of the continuous-time system

$$\begin{bmatrix} \dot{z} \\ \omega \end{bmatrix} = \phi(z) \quad z \in \Upsilon$$

   and the maximal solutions to the discrete-time system

$$\begin{bmatrix} z^+ \\ \zeta \end{bmatrix} \in \psi(z) \quad z \in \Upsilon$$

   are complete.

II. There are no non-trivial solutions to the system

$$\left.\begin{aligned} \dot{x} &= \xi\bar{\vartheta} \\ \dot{\xi} &= \xi \otimes \nu(\omega) \\ \begin{bmatrix} \dot{z} \\ \omega \end{bmatrix} &= \phi(z) \end{aligned}\right\} \quad (x,\xi,z) \in C_0$$

   where

$$C_0 := \left\{(x,\xi,z) \,\middle|\, \langle \nabla\varphi(x), \xi\bar{\vartheta}\rangle = 0\,,\ \xi \in S^1\,,\ z \in \Upsilon\right\}\,.$$

III. The only complete solutions to the system

$$\left.\begin{aligned} x^+ &= x \\ \xi^+ &= \xi \otimes \zeta \\ \begin{bmatrix} z^+ \\ \zeta \end{bmatrix} &\in \psi(z) \end{aligned}\right\} \quad (x,\xi,z) \in D$$

   start from $x_\circ = x^*$.

The first assumption on (Υ,ϕ,ψ) above guarantees that the control algorithm can generate commands by either flowing exclusively or jumping exclusively. This permits arbitrary combinations of flows and jumps. Thus, since $C \cup D = \mathbb{R}^2 \times S^1 \times \Upsilon$, all solutions to the closed-loop system are complete. Moreover, the assumption that Υ is compact guarantees that the only way solutions can grow unbounded is if x grows unbounded.

The second assumption on (Υ,ϕ,ψ) guarantees that closed-loop flows lead to an increase in the function φ. One situation where the assumption is easy to check is when $\omega = 0$ for all $z \in \Upsilon$ and the maxima of φ along each search direction are isolated. In other words, the maxima of the function $\varphi_{\zeta,x} : \mathbb{R} \to \mathbb{R}$ given by $\lambda \mapsto \varphi(x + \lambda\zeta)$ are isolated for each $(\zeta, x) \in S^1 \times \mathbb{R}^2$. In this case it is not possible to flow while keeping φ constant.

The last assumption on (Υ,ϕ,ψ) guarantees that the discrete update algorithm is rich enough to be able to find eventually a direction of decrease for φ for every point x. (Clearly the only way this assumption can be satisfied is if $\nabla\varphi(x) = 0$ only if $x = x^*$, which is what we are assuming.) The assumption prevents the existence of discrete solutions at points where $x \neq x^*$.

One example of data (Υ,ϕ,ψ) satisfying the three conditions above is

$$\Upsilon = \{0\}\,, \quad \phi(z) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}\,, \quad \psi(z) = \begin{bmatrix} z \\ \begin{bmatrix} 0 \\ 1 \end{bmatrix} \end{bmatrix}\,.$$

For this system, the state z does not change, the generated angular velocity ω is always zero, and the commanded rotation at each jump is $\pi/2$ radians. Other more complicated orientation-generating algorithms that make use of the dynamic state z are also possible. For example, the algorithm in [42] uses the state variable z to generate conjugate directions at the update times.

With the assumptions in place, the invariance principle of Subsect. "Invariance Principles" can be applied with the function $-\varphi$ to conclude that the closed-loop system has the compact set $\mathcal{A} := \{x^*\} \times S^1 \times \Upsilon$ globally asymptotically stable. Moreover, because of the robustness of global asymptotic stability to small perturbations, the results that we obtain are robust, in a practical sense, to slow variations in the characterization of the function φ, including the point where it obtains its maximum.

Practical Modifications The assumptions of the previous section that we would like to relax are that ξ can change discontinuously and that the derivative of φ is available as a measurement.

The first issue can be addressed by inserting a mode after every jump where the forward velocity is set to zero and a constant angular velocity is applied for the correct amount of time to drive ξ to the value $\xi \otimes \zeta$. If it is not possible to set the velocity to zero then some other open-loop maneuver can be executed so that, after some time, the orientation has changed by the correct amount while the position has not changed.

The second issue can be addressed by making sure that, after the direction is updated, values of the function φ along the solution are stored and compared to the current value of φ to determine the sign of the derivative. The comparison should not take place until after a sufficient amount of time has elapsed, to enhance robustness to measurement noise. The robustness of the nominal algorithm to temporal regularization and other perturbations permits such a practical implementation.

Discussion and Final Remarks

We have presented one viewpoint on hybrid control systems, but the field is still developing and other authors will give a different emphasis.

For starters, we have stressed a dynamical systems view of hybrid systems, but authors with a computer science background typically will emphasize a hybrid automaton point of view that separates discrete-valued variables from continuous-valued variables. This decomposition can be found in the early work [61] and [59], and in the more recent work [2,8,58]. An introduction to this modeling approach is given in [7]. Impulsive systems, as described in [4], are closely related to hybrid systems but rely on ordinary time domains and usually do not consider the case where the flow set and jump set overlap. The work in [18] on "left-continuous systems" is closely linked to such systems. Passing to hybrid time domains (under different names) can be seen in [20,24,39,40]; other generalized concepts of time domains can be found in [43] and in the literature on dynamical systems on time scales, see [41] for an introduction.

For simplicity, we have taken the flow map to be a function rather than a set-valued mapping. A motivation for set-valued mappings satisfying basic conditions is given in [56] where the notion of generalized solutions is developed and shown to be equivalent to solutions in the presence of vanishing perturbations or noise. Set-valued dynamics, with some consideration of the regularity of the mappings defining them, can be found in [1,2]. Implications of data regularity on the basic structural properties of the set of solutions to a system were outlined concurrently in [20,24]. The work in [24] emphasized the implications for robustness of stability theory. The latter work preceded the rigorous derivations in [23] where the proofs of statements in the Subsect. "Conditions for Well-Posedness" can be found. To derive stronger results on continuous dependence of solutions to initial conditions, extra assumptions must be added. An early result in this direction appeared in [59]; see also [11,20,40]. The work in [11] exhibited a continuous selection of solutions, continuous dependence in the set-valued sense was addressed in [17].

Sufficient Lyapunov stability conditions, for hybrid or switching systems, and relying on various concepts of a solution, appeared in [6,21,40,62]. The results in Subsect. "Lyapunov Stability Theorem" are contained in [54] which develops a general invariance principle for hybrid systems. A part of the latter result is quoted in Subsect. "Invariance Principles". Other invariance results for hybrid or switching systems have appeared in [3,18,28,30,40]. Some early converse results for hybrid systems, relying on nonsmooth and possibly discontinuous Lyapunov functions, were given in [62]. The results quoted in Subsect. "Converse Lyapunov Theorems and Robustness" come from [15].

The development of hybrid control theory is still in its formative stages. This article has focused on the development of supervisors of hybrid controllers and related topics, as well as to applications where hybrid control gives solutions that overcome obstacles faced by classical control. For hybrid supervisors used in the context of adaptive control, see [63] and the references therein. Other results related to supervisors include [47], which considers the problem discussed at the end of Subsect. "Supervisors of Hybrid Controllers", and [57]. The field of hybrid control systems is moving in the direction of systematic design tools, but the capabilities of hybrid control have been recognized for some time. Many of the early observations were in the context of nonholonomic systems, like the result for mobile robots we have presented as an application of supervisors. For example, see [29,31,37,44,49]. More recently, in [48] and [50] it has been established that every asymptotically controllable nonlinear system can be robustly asymptotically stabilized using logic-based hybrid feedback. Other recent results include the work in [25] and its predecessor [12], and the related work on linear reset control systems as considered in [5,45] and the references therein.

This article did not have much to do with the control of hybrid systems, other than the discussion of the juggling problem in the introduction and the structure of hybrid controllers for hybrid systems in Subsect. "Hybrid Controllers for Hybrid Systems". A significant amount of work on the control of hybrid systems has been done, although

typically not in the framework proposed here. Notable references include [10,46,51] and the references therein.

Other interesting topics and open questions in the area of hybrid control systems are developed in [38].

Future Directions

What does the future hold for hybrid control systems? With a framework in place that mimics the framework of ordinary differential and difference equations, it appears that many new results will become available in directions that parallel results available for nonlinear control systems. These include certain types of separation principles, control algorithms based on zero dynamics (results in this direction can be found in [60] and [13]), and results based on interconnections and time-scale separation. Surely there will be unexpected results that are enabled by the fundamentally different nature of hybrid control as well.

The theory behind the control of systems with impacts will continue to develop and lead to interesting applications. It is also reasonable to anticipate further developments related to the construction of robust, embedded hybrid control systems and robust networked control systems. Likely the research community will also be inspired by hybrid control systems discovered in nature.

The future is bright for hybrid control systems design and it will be exciting to see the progress that is made over the next decade and beyond.

Bibliography

Primary Literature

1. Aubin JP, Haddad G (2001) Cadenced runs of impulse and hybrid control systems. Internat J Robust Nonlinear Control 11(5):401–415
2. Aubin JP, Lygeros J, Quincampoix M, Sastry SS, Seube N (2002) Impulse differential inclusions: a viability approach to hybrid systems. IEEE Trans Automat Control 47(1):2–20
3. Bacciotti A, Mazzi L (2005) An invariance principle for nonlinear switched systems. Syst Control Lett 54:1109–1119
4. Bainov DD, Simeonov P (1989) Systems with impulse effect: stability, theory, and applications. Ellis Horwood, Chichester; Halsted Press, New York
5. Beker O, Hollot C, Chait Y, Han H (2004) Fundamental properties of reset control systems. Automatica 40(6):905–915
6. Branicky M (1998) Multiple Lyapunov functions and other analysis tools for switched hybrid systems. IEEE Trans Automat Control 43(4):475–482
7. Branicky M (2005) Introduction to hybrid systems. In: Levine WS, Hristu-Varsakelis D (eds) Handbook of networked and embedded control systems. Birkhäuser, Boston, pp 91–116
8. Branicky M, Borkar VS, Mitter SK (1998) A unified framework for hybrid control: Model and optimal control theory. IEEE Trans Automat Control 43(1):31–45
9. Brockett RW (1983.) Asymptotic stability and feedback stabilization. In: Brockett RW, Millman RS, Sussmann HJ (eds) Differential Geometric Control Theory. Birkhauser, Boston, MA, pp 181–191
10. Brogliato B (1996) Nonsmooth mechanics models, dynamics and control. Springer, London
11. Broucke M, Arapostathis A (2002) Continuous selections of trajectories of hybrid systems. Syst Control Lett 47:149–157
12. Bupp RT, Bernstein DS, Chellaboina VS, Haddad WM (2000) Resetting virtual absorbers for vibration control. J Vib Control 6:61
13. Cai C, Goebel R, Sanfelice R, Teel AR (2008) Hybrid systems: limit sets and zero dynamics with a view toward output regulation. In: Astolfi A, Marconi L (eds) Analysis and design of nonlinear control systems – In Honor of Alberto Isidori. Springer, pp 241–261 http://www.springer.com/west/home/generic/search/results?SGWID=4-40109-22-173754110-0
14. Cai C, Teel AR, Goebel R (2007) Results on existence of smooth Lyapunov functions for asymptotically stable hybrid systems with nonopen basin of attraction. In: Proc. 26th American Control Conference, pp 3456–3461, http://www.ccec.ece.ucsb.edu/~cai/
15. Cai C, Teel AR, Goebel R (2007) Smooth Lyapunov functions for hybrid systems - Part I: Existence is equivalent to robustness. IEEE Trans Automat Control 52(7):1264–1277
16. Cai C, Teel AR, Goebel R (2008) Smooth Lyapunov functions for hybrid systems - Part II: (Pre-)asymptotically stable compact sets. IEEE Trans Automat Control 53(3):734–748. See also [14]
17. Cai C, Goebel R, Teel A (2008) Relaxation results for hybrid inclusions. Set-Valued Analysis (To appear)
18. Chellaboina V, Bhat S, Haddad W (2003) An invariance principle for nonlinear hybrid and impulsive dynamical systems. Nonlin Anal 53:527–550
19. Clegg JC (1958) A nonlinear integrator for servomechanisms. Transactions AIEE 77(Part II):41–42
20. Collins P (2004) A trajectory-space approach to hybrid systems. In: 16th International Symposium on Mathematical Theory of Networks and Systems, CD-ROM
21. DeCarlo R, Branicky M, Pettersson S, Lennartson B (2000) Perspectives and results on the stability and stabilizability of hybrid systems. Proc of IEEE 88(7):1069–1082
22. Filippov A (1988) Differential equations with discontinuous right-hand sides. Kluwer, Dordrecht
23. Goebel R, Teel A (2006) Solutions to hybrid inclusions via set and graphical convergence with stability theory applications. Automatica 42(4):573–587
24. Goebel R, Hespanha J, Teel A, Cai C, Sanfelice R (2004) Hybrid systems: Generalized solutions and robust stability. In: Proc. 6th IFAC Symposium in Nonlinear Control Systems, pp 1–12, http://www-ccec.ece.ucsb.edu/7Ersanfelice/Preprints/final_nolcos.pdf
25. Haddad WM, Chellaboina V, Hui Q, Nersesov SG (2007) Energy- and entropy-based stabilization for lossless dynamical systems via hybrid controllers. IEEE Trans Automat Control 52(9):1604–1614, http://ieeexplore.ieee.org/iel5/9/4303218/04303228.pdf
26. Hájek O (1979) Discontinuous differential equations, I. J Diff Eq 32:149–170
27. Hermes H (1967) Discontinuous vector fields and feedback control. In: Differential Equations and Dynamical Systems, Academic Press, New York, pp 155–165

28. Hespanha J (2004) Uniform stability of switched linear systems: Extensions of LaSalle's invariance principle. IEEE Trans Automat Control 49(4):470–482
29. Hespanha J, Morse A (1999) Stabilization of nonholonomic integrators via logic-based switching. Automatica 35(3): 385–393
30. Hespanha J, Liberzon D, Angeli D, Sontag E (2005) Nonlinear norm-observability notions and stability of switched systems. IEEE Trans Automat Control 50(2):154–168
31. Hespanha JP, Liberzon D, Morse AS (1999) Logic-based switching control of a nonholonomic system with parametric modeling uncertainty. Syst Control Lett 38:167–177
32. Johansson K, Egerstedt M, Lygeros J, Sastry S (1999) On the regularization of zeno hybrid automata. Syst Control Lett 38(3):141–150
33. Kellet CM, Teel AR (2004) Smooth Lyapunov functions and robustness of stability for differential inclusions. Syst Control Lett 52:395–405
34. Krasovskii N (1970) Game-Theoretic Problems of capture. Nauka, Moscow
35. LaSalle JP (1967) An invariance principle in the theory of stability. In: Hale JK, LaSalle JP (eds) Differential equations and dynamical systems. Academic Press, New York
36. LaSalle J (1976) The stability of dynamical systems. SIAM's Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia
37. Lucibello P, Oriolo G (1995) Stabilization via iterative state steering with application to chained-form systems. In: Proc. 35th IEEE Conference on Decision and Control, pp 2614–2619
38. Lygeros J (2005) An overview of hybrid systems control. In: Levine WS, Hristu-Varsakelis D (eds) Handbook of Networked and Embedded Control Systems. Birkhäuser, Boston, pp 519–538
39. Lygeros J, Johansson K, Sastry S, Egerstedt M (1999) On the existence of executions of hybrid automata. In: Proc. 38th IEEE Conference on Decision and Control, pp 2249–2254
40. Lygeros J, Johansson K, Simić S, Zhang J, Sastry SS (2003) Dynamical properties of hybrid automata. IEEE Trans Automat Control 48(1):2–17
41. M Bohner M, Peterson A (2001) Dynamic equations on time scales. An introduction with applications. Birkhäuser, Boston
42. Mayhew CG, Sanfelice RG, Teel AR (2007) Robust source seeking hybrid controllers for autonomous vehicles. In: Proc. 26th American Control Conference, pp 1185–1190
43. Michel A (1999) Recent trends in the stability analysis of hybrid dynamical systems. IEEE Trans Circuits Syst – I Fund Theory Appl 45(1):120–134
44. Morin P, Samson C (2000) Robust stabilization of driftless systems with hybrid open-loop/feedback control. In: Proc. 19th American Control Conference, pp 3929–3933
45. Nesic D, Zaccarian L, Teel A (2008) Stability properties of reset systems. Automatica 44(8):2019–2026
46. Plestan F, Grizzle J, Westervelt E, Abba G (2003) Stable walking of a 7-dof biped robot. IEEE Trans Robot Automat 19(4): 653–668
47. Prieur C (2001) Uniting local and global controllers with robustness to vanishing noise. Math Contr, Sig Syst 14(2): 143–172
48. Prieur C (2005) Asymptotic controllability and robust asymptotic stabilizability. SIAM J Control Opt 43:1888–1912
49. Prieur C, Astolfi A (2003) Robust stabilization of chained systems via hybrid control. IEEE Trans Automat Control 48(10):1768–1772
50. Prieur C, Goebel R, Teel A (2007) Hybrid feedback control and robust stabilization of nonlinear systems. IEEE Trans Automat Control 52(11):2103–2117
51. Ronsse R, Lefèvre P, Sepulchre R (2007) Rhythmic feedback control of a blind planar juggler. IEEE Transactions on Robotics 23(4):790–802, http://www.montefiore.ulg.ac.be/services/stochastic/pubs/2007/RLS07
52. Ryan E (1994) On Brockett's condition for smooth stabilizability and its necessity in a context of nonsmooth feedback. SIAM J Control Optim 32(6):1597–1604
53. Sanfelice R, Goebel R, Teel A (2006) A feedback control motivation for generalized solutions to hybrid systems. In: Hespanha JP, Tiwari A (eds) Hybrid Systems: Computation and Control: 9th International Workshop, vol LNCS, vol 3927. Springer, Berlin, pp 522–536
54. Sanfelice R, Goebel R, Teel A (2007) Invariance principles for hybrid systems with connections to detectability and asymptotic stability. IEEE Trans Automat Control 52(12):2282–2297
55. Sanfelice R, Teel AR, Sepulchre R (2007) A hybrid systems approach to trajectory tracking control for juggling systems. In: Proc. 46th IEEE Conference on Decision and Control, pp 5282–5287
56. Sanfelice R, Goebel R, Teel A (2008) Generalized solutions to hybrid dynamical systems. ESAIM: Control, Optimisation and Calculus of Variations 14(4):699–724
57. Sanfelice RG, Teel AR (2007) A "throw-and-catch" hybrid control strategy for robust global stabilization of nonlinear systems. In: Proc. 26th American Control Conference, pp 3470–3475
58. van der Schaft A, Schumacher H (2000) An introduction to hybrid dynamical systems. Lecture notes in control and information sciences. Springer, London
59. Tavernini L (1987) Differential automata and their discrete simulators. Nonlinear Anal 11(6):665–683
60. Westervelt E, Grizzle J, Koditschek D (2003) Hybrid zero dynamics of planar biped walkers. IEEE Trans Automat Control 48(1):42–56
61. Witsenhausen HS (1966) A class of hybridstate continuous-time dynamic systems. IEEE Trans Automat Control 11(2):161–167
62. Ye H, Mitchel A, Hou L (1998) Stability theory for hybrid dynamical systems. IEEE Trans Automat Control 43(4):461–474
63. Yoon TW, Kim JS, Morse A (2007) Supervisory control using a new control-relevant switching. Automatica 43(10): 1791–1798

Books and Reviews

Aubin JP, Cellina A (1984) Differential inclusions. Springer, Berlin
Haddad WM, Chellaboina V, Nersesov SG (2006) Impulsive and hybrid dynamical systems: stability, dissipativity, and control. Princeton University Press, Princeton
Levine WS, Hristu-Varsakelis D (2005) Handbook of networked and embedded control systems. Birkhäuser, Boston
Liberzon D (2003) Switching in systems and control. Systems and control: Foundations and applications. Birkhäuser, Boston
Matveev AS, Savkin AV (2000) Qualitative theory of hybrid dynamical systems. Birkhäuser, Boston

Michel AN, Wang L, Hu B (2001) Qualitative theory of dynamical systems. Dekker
Rockafellar RT, Wets RJ-B (1998) Variational analysis. Springer, Berlin

Hybrid Soft Computing Models for Systems Modeling and Control

Oscar Castillo, Patricia Melin
Division of Graduate Studies and Research,
Tijuana Institute of Technology, Tijuana, Mexico

Article Outline

Glossary
Definition of the Subject
Introduction
Genetic Algorithm for Optimization
Evolution of Fuzzy Systems
Application to Anesthesia Control
Application to the Control of the Bar and Ball System
Hierarchical Genetic Algorithms for Neural Networks
Experimental Results for Time Series Prediction
Conclusions
Future Directions
Bibliography

Glossary

Hybrid intelligent systems Intelligent Systems that are build using a combination of soft computing techniques. In particular, Soft Computing includes fuzzy logic, neural networks, genetic algorithms or hybrid approaches.

Intelligent control The application of intelligent techniques for achieving the control of non-linear plants. In particular, the use of fuzzy logic, neural networks, genetic algorithms or hybrid approaches for designing intelligent controllers.

Soft computing Soft Computing is a new area of Computer Science that deals with new intelligent methodologies that combine symbolic and numerical calculations. In particular, Soft Computing includes, at the moment, methodologies like fuzzy logic, neural networks, genetic algorithms or hybrid approaches.

Fuzzy systems Intelligent systems that are developed based on the theory of fuzzy logic, fuzzy inference and membership functions, and fuzzy rules. Fuzzy systems are able to manage the uncertainty of the decision process of humans, and for this reason are able to mimic the expert decision process in automation applications.

Genetic algorithms Genetic algorithms are search optimization techniques that mimic natural evolution for finding solution to complex problems. In particular, genetic algorithms use operators to generate new candidate solutions based on the selection of previous good solutions.

Evolution of fuzzy systems Application of evolutionary algorithms to the optimization of number of fuzzy rules and membership functions, as well as the parameter values of the fuzzy system.

Definition of the Subject

The evolutionary design of hybrid intelligent systems using hierarchical genetic algorithms will be described in this paper. The evolutionary approach can be used for fuzzy system optimization in intelligent control. In particular, we consider the problem of optimizing the number of rules and membership functions using an evolutionary approach. The hierarchical genetic algorithm enables the optimization of the fuzzy system design for a particular application. We illustrate the approach with two cases of intelligent control. Simulation results for both applications show that we are able to find an optimal set of rules and membership functions for the fuzzy control system. We also describe the application of the evolutionary approach for the problem of designing hybrid intelligent systems in time series prediction. In this case, the goal is to design the best predictor for complex time series. Simulation results show that the evolutionary approach optimizes the hybrid intelligent systems in time series prediction.

Introduction

The application of a Hierarchical Genetic Algorithm (HGA) for fuzzy system optimization [14] will be described in this paper. In particular, we consider the problem of finding the optimal set of rules and membership functions for a specific application [24]. The HGA is used to search for this optimal set of rules and membership functions, according to the data about the problem. We consider, as an illustration, the case of a fuzzy system for intelligent control [4].

Fuzzy systems are capable of handling complex, nonlinear and sometimes mathematically intangible dynamic systems using simple solutions [10]. Very often, fuzzy systems may provide a better performance than conventional non-fuzzy approaches with less development cost [17]. However, to obtain an optimal set of fuzzy membership functions and rules is not an easy task [12]. It requires

time, experience and skills of the designer for the tedious fuzzy tuning exercise [22]. In principle, there is no general rule or method for the fuzzy logic set-up, although a heuristic and iterative procedure for modifying the membership functions to improve performance has been proposed [19]. Recently, many researchers have considered a number of intelligent schemes for the task of tuning the fuzzy system [1]. The noticeable Neural Network (NN) approach [9] and the Genetic Algorithm (GA) approach [8] to optimize either the membership functions or rules, have become a trend for fuzzy logic system development.

The HGA approach differs from the other techniques [5] in that it has the ability to reach an optimal set of membership functions and rules without a known fuzzy system topology [20]. During the optimization phase, the membership functions need not be fixed. Throughout the genetic operations [7], a reduced fuzzy system including the number of membership functions and fuzzy rules will be generated [25]. The HGA approach has a number of advantages:

1. An optimal and the least number of membership functions and rules are obtained,
2. no pre-fixed fuzzy structure is necessary, and
3. simpler implementing procedures and less cost are involved.

We consider in this paper the case of automatic anesthesia control in human patients for testing the optimized fuzzy controller. We did have, as a reference, the best fuzzy controller that was developed for the automatic anesthesia control [13], and we consider the optimization of this controller using the HGA approach. After applying the genetic algorithm the number of fuzzy rules was reduced from 12 to 9 with a similar performance of the fuzzy controller. Of course, the parameters of the membership functions were also tuned by the genetic algorithm. We did compare the simulation results of the optimized fuzzy controllers obtained with the HGA against the best fuzzy controller that was obtained previously with expert knowledge, and control is achieved in a similar fashion. Since simulation results are similar, and the number of fuzzy rules was reduced, we can conclude that the HGA approach is a good alternative for designing fuzzy systems.

We also describe the application of the evolutionary approach for the problem of designing hybrid intelligent systems in time series prediction. In this case, the goal is to design the best predictor for complex time series. Simulation results show that the evolutionary approach optimizes the hybrid intelligent systems in time series prediction.

Genetic Algorithm for Optimization

In this paper, we used a floating-point genetic algorithm [3] to adjust the parameter vector θ, specifically we used the Breeder Genetic Algorithm (BGA). The genetic algorithm is used to optimize the fuzzy system for control that will be described later. A BGA can be described by the following equation:

$$\mathrm{BGA} = (P_g^0, N, T, \Gamma, \Delta, \mathrm{HC}, F, \mathrm{term}) \tag{1}$$

where: P_g^0 = initial population, N = the size of the population, T = the truncation threshold, Γ = the recombination operator, Δ = the mutation operator, HC = the hill climbing method, F = the fitness function, term = the termination criterion.

The BGA uses a selection scheme called truncation selection. The $\%T$ best individuals are selected and mated randomly until the number of offspring is equal the size of the population. The offspring generation is equal to the size of the population. The offspring generation replaces the parent population. The best individual found so far will remain in the population. Self-mating is prohibited. As a recombination operator we used "extended intermediate recombination", defined as: If $x = (x_i, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ are the parents, then the successor $z = (z_1, \ldots, z_n)$ is calculated by:

$$z_i = x_i + \alpha_i(y_i - x_i) \quad i = 1, \ldots n\,. \tag{2}$$

The mutation operator is defined as follows: A variable x_i is selected with probability p_m for mutation. The BGA normally uses $p_m = 1/n$. At least one variable will be mutated. A value out of the interval $[-\mathrm{range}_i, \mathrm{range}_i]$ is added to the variable. range_i defines the mutation range. It is normally set to $(0.1 \times \mathrm{searchinterval}_i)$. $\mathrm{searchinterval}_i$ is the domain of definition for variable x_i. The new value z_i is computed according to

$$z_i = x_i \pm \mathrm{range}_i \cdot \delta\,. \tag{3}$$

The + or − sign is chosen with probability 0.5.δ is computed from a distribution which prefers small values. This is realized as follows

$$\delta = \sum_{i=0}^{15} \alpha_i 2^i \quad \alpha_i \in 0, 1\,. \tag{4}$$

Before mutation we set $\alpha_i = 0$. Then each α_i is mutated to 1 with probability $p_\delta = 1/16$. Only $\alpha_i = 1$ contributes to the sum. On the average there will be just one α_i with value 1, say α_j. Then δ is given by

$$\delta = 2^{-j}\,. \tag{5}$$

The standard BGA mutation operator is able to generate any point in the hypercube with center x defined by $x_i \pm \text{range}_i$. But it generates values much more often in the neighborhood of x. In the above standard setting, the mutation operator is able to locate the optimal x_i up to a precision of $\text{range}_i \cdot 2^{-150}$.

We also solved the problem with a LMS algorithm, with the purpose of have a good reference mark. To monitor the convergence rate of the LMS algorithm, we computed a short term average of the squared error $e^2(n)$ using

$$\text{ASE}(m) = \frac{1}{K} \sum_{k=n+1}^{n+K} e^2(k) \tag{6}$$

where $m = n/K = 1, 2, \ldots$. The averaging interval K may be selected to be (approximately) $K = 10N$. The effect of the choice of the step size parameter Δ on the convergence rate of LMS algorithm may be observed by monitoring the ASE(m).

Genetic Algorithm for Optimization

The proposed genetic algorithm is as follows:

1. We use real numbers as a genetic representation of the problem.
2. We initialize variable i with zero ($i = 0$).
3. We create an initial random population P_i, in this case (P_0). Each individual of the population has n dimensions and, each coefficient of the fuzzy system corresponds to one dimension. Since we are using a fuzzy system of 25 coefficients, then our search space is of $n = 25$. The generated individuals have their coefficients in $[-3, 3]$.
4. We calculate the normalized fitness of each individual of the population using linear scaling with displacement, in the following form:

$$f_i' = f_i + \frac{1}{N} \sum |f_i| + \left| \min_i(f_i) \right| \quad \forall i .$$

5. We normalize the fitness of each individual using:

$$F_i = \frac{f_i'}{\sum_{i=1}^{N} f_i'} \quad \forall i .$$

6. We sort the individuals from greater to lower fitness.
7. We use the truncated selection method, selecting the %T best individuals, for example if there are 500 individuals and, then we select 0.30*500 = 150 individuals.
8. We apply random crossover, to the individuals in the population (the 150 best ones) with the goal of creating a new population (of 500 individuals). Crossover with it self is not allowed, and all the individuals have to participate. To perform this operation we apply the genetic operator of extended intermediate recombination as follows:
   If $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ are the parents, then the successors $z = (z_1, \ldots, z_n)$ are calculated by, $z_i = x_i + \alpha_i(y_i - x_i)$ for $i = 1, \ldots, n$ where α is a scaling factor selected randomly in the interval $[-d, 1 + d]$. In intermediate recombination $d = 0$, and for extended $d > 0$, a good choice is $d = 0.25$, which is the one that we used.
9. We apply the mutation genetic operator of BGA. In this case, we select an individual with probability $p_m = 1/n$(where n represents the working dimension, in this case $n = 25$, which is the number of coefficients in the membership functions). The mutation operator calculates the new individuals z_i of the population in the following form: $z_i = x_i \pm \text{range}_i \delta$ we can note from this equation that we are actually adding to the original individual a value in the interval: $[-\text{range}_i, \text{range}_i]$ the range is defined as the search interval, which in this case is the domain of variable x_i, the sign $\pm$ is selected randomly with probability of 0.5, and is calculated using the following formula,

$$\delta = \sum_{i=0}^{m-1} \alpha_i 2^{-i} \quad \alpha_i \in 0, 1 .$$

   Common used values in this equation are $m = 16$ and $m = 20$. Before mutation we initiate with $\alpha_i = 0$, then for each α_i we mutate to 1 with probability $p_\delta = 1/m$.
10. Let $i = i + 1$, and continue with step 4.

Evolution of Fuzzy Systems

Ever since the very first introduction of the fundamental concept of fuzzy logic [26], its use in engineering disciplines has been widely studied. Its main attraction undoubtedly lies in the unique characteristics that fuzzy logic systems possess [27]. They are capable of handling complex, non-linear dynamic systems using simple solutions. Very often, fuzzy systems provide a better performance than conventional non-fuzzy approaches with less development cost [24].

However, to obtain an optimal set of fuzzy membership functions and rules is not an easy task [11]. It requires time, experience, and skills of the operator for the tedious

fuzzy tuning exercise [2]. In principle, there is no general rule or method for the fuzzy logic set-up [15]. Recently, many researchers have considered a number of intelligent techniques for the task of tuning the fuzzy set.

Here, another innovative scheme is described [23]. This approach has the ability to reach an optimal set of membership functions and rules without a known overall fuzzy set topology. The conceptual idea of this approach is to have an automatic and intelligent scheme to tune the membership functions and rules, in which the conventional closed loop fuzzy control strategy remains unchanged, as indicated in Fig. 1.

In this case, the chromosome [6] of a particular system is shown in Fig. 2. The chromosome consists of two types of genes, the control genes and parameter genes. The control genes, in the form of bits, determine the membership function activation, whereas the parameter genes are in the form of real numbers to represent the membership functions.

To obtain a complete design for the fuzzy control system, an appropriate set of fuzzy rules is required to ensure system performance [25]. At this point it should be stressed that the introduction of the control genes is done to govern the number of fuzzy subsets in the system.

Once the formulation of the chromosome has been set for the fuzzy membership functions and rules, the genetic operation cycle can be performed. This cycle of operation for the fuzzy control system optimization using a genetic algorithm is illustrated in Fig. 3.

There are two population pools, one for storing the membership chromosomes and the other for storing the fuzzy rule chromosomes. We can see this in Fig. 3 as the membership population and fuzzy rule population, respectively. Considering that there are various types of gene structure, a number of different genetic operations can be used. For the crossover operation, a one-point crossover is applied separately for both the control and parameter genes of the membership chromosomes within certain operation rates. There is no crossover operation for fuzzy rule chromosomes since only one suitable rule set can be assisted.

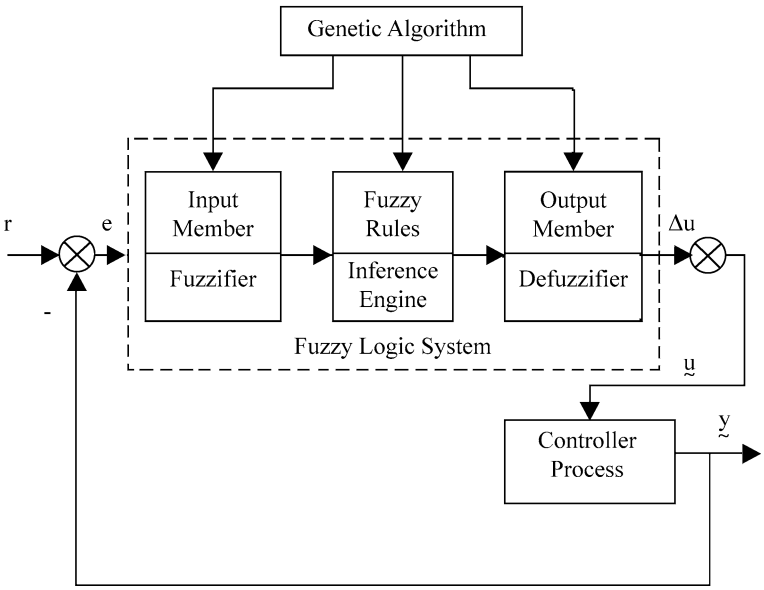

Hybrid Soft Computing Models for Systems Modeling and Control, Figure 1
Genetic algorithm for a fuzzy control system

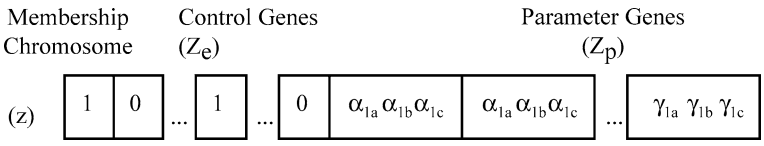

Hybrid Soft Computing Models for Systems Modeling and Control, Figure 2
Chromosome structure for the fuzzy system

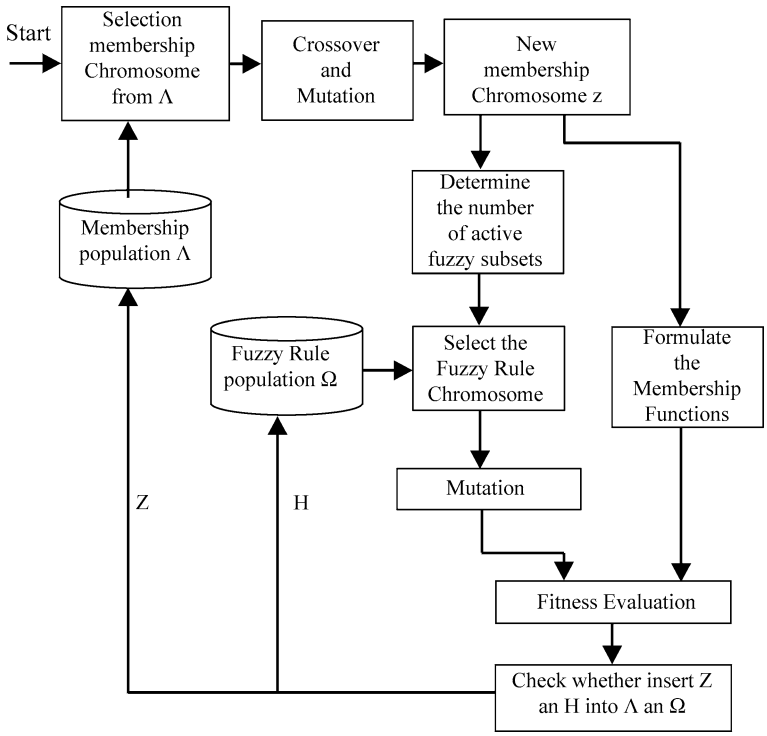

Hybrid Soft Computing Models for Systems Modeling and Control, Figure 3
Genetic cycle for fuzzy system optimization

Bit mutation is applied for the control genes of the membership chromosome. Each bit of the control gene is flipped if a probability test is satisfied (a randomly generated number is smaller than a predefined rate). As for the parameter genes, which are real number represented, random mutation is applied.

The complete genetic cycle continues until some termination criteria, for example, meeting the design specification or number of generation reaching a predefined value are fulfilled.

The fitness function can be defined in this case as follows:

$$f_i = \sum |y(k) - r(k)| \tag{7}$$

where $\sum$ indicates the sum for all the data points in the training set, and $y(k)$ represents the real output of the fuzzy system and $r(k)$ is the reference output. This fitness value measures how well the fuzzy system is approximating the real data of the problem.

Application to Anesthesia Control

We consider the case of controlling the anesthesia given to a patient as the problem for finding the optimal fuzzy system for control [13]. The complete implementation was done in the MATLAB programming language. The fuzzy systems were build automatically by using the Fuzzy Logic Toolbox, and the genetic algorithm was coded directly in the MATLAB language. The fuzzy systems for control are the individuals used in the genetic algorithm, and these are evaluated by comparing them to the ideal control given by the experts. In other words, we compare the performance of the fuzzy systems that are generated by the genetic algorithm, against the ideal control system given by the experts in this application. We give more details below.

Anesthesia Control Using Fuzzy Logic

The main task of the anesthesist, during and operation, is to control anesthesia concentration. In any case, anesthesia concentration can't be measured directly. For this reason, the anesthesist uses indirect information, like the heartbeat, pressure, and motor activity. The anesthesia concentration is controlled using a medicine, which can be given by a shot or by a mix of gases. We consider here the use of isoflurance, which is usually given in a concentration of 0 to 2% with oxygen. In Fig. 4 we show a block diagram of the controller.

The air that is exhaled by the patient contains a specific concentration of isoflurance, and it is recirculated to the

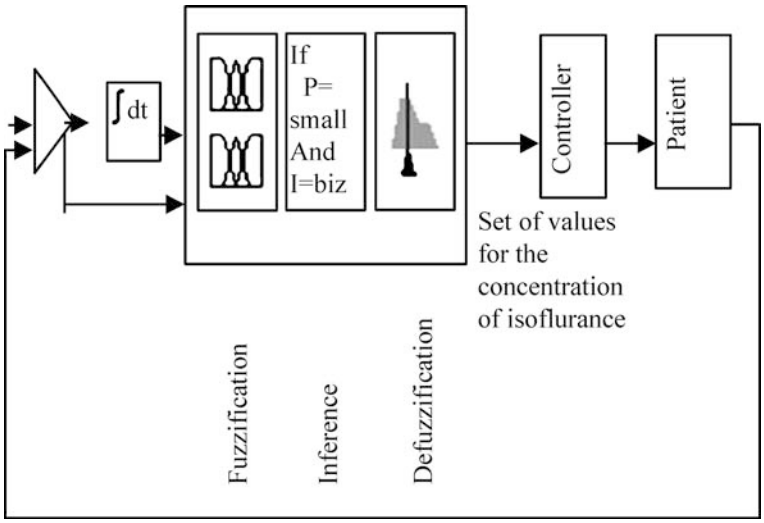

Hybrid Soft Computing Models for Systems Modeling and Control, Figure 4
Architecture of the fuzzy control system

patient. As consequence, we can measure isoflurance concentration on the inhaled and exhaled air by the patient, to estimate isoflurance concentration on the patient's blood. From the control engineering point of view, the task by the anesthesist is to maintain anesthesia concentration between the high level W (threshold to wake up) and the low level E (threshold to success). These levels are difficult to be determine in a changing environment and also are dependent on the patient's condition. For this reason, it is important to automate this anesthesia control, to perform this task more efficiently and accurately, and also to free the anesthesist from this time consuming job. The anesthesist can then concentrate in doing other task during operation of a patient.

The first automated system for anesthesia control was developed using a PID controller in the 60's. However, this system was not very successful due to the non-linear nature of the problem of anesthesia control. After this first attempt, adaptive control was proposed to automate anesthesia control, but robustness was the problem in this case. For these reasons, fuzzy logic was proposed for solving this problem. An additional advantage of fuzzy control is that we can use in the rules the same vocabulary as the medical doctors use. The fuzzy control system can also be easily interpreted by the anesthesists.

Characteristics of the Fuzzy Controller

In this section we describe the main characteristics of the fuzzy controller for anesthesia control. We will define input and output variable of the fuzzy system. Also, the fuzzy rules of fuzzy controller previously designed will be described.

The fuzzy system is defined as follows:

1. Input variables: Blood pressure and Error.
2. Output variable: Isoflurance concentration.
3. Nine fuzzy if-then rules of the optimized system, which is the base for comparison.
4. 12 fuzzy if-then rules of an initial system to begin the optimization cycle of the genetic algorithm.

The linguistic values used in the fuzzy rules are the following:

- PB = Positive Big
- PS = Positive Small
- ZERO = zero
- NB = Negative Big
- NS = Negative Small

We show below a sample set of fuzzy rules that are used in the fuzzy inference system that is represented in the genetic algorithm for optimization:

- if Blood pressure is NB and error is NB then conc_isoflurance is PS
- if Blood pressures is PS then conc_isoflurance is NS
- if Blood pressure is NB then conc_isoflurance is PB
- if Blood pressure is PB then conc_isoflurance is NB
- if Blood pressure is ZERO and error is ZERO then conc_isoflurance is ZERO
- if Blood pressure is ZERO and error is PS then conc_isoflurance is NS

- if Blood pressure is ZERO and error is NS then conc_isoflurance is PS
- if error is NB then conc_isoflurance is PB
- if error is PB then conc_isoflurance is NB
- if error is PS then conc_isoflurance is NS
- if Blood pressure is NS and error is ZERO then conc_isoflurance is NB
- if Blood pressure is PS and error is ZERO then conc_isoflurance is PS.

Genetic Algorithm Specification

The general characteristics of the genetic algorithm that was used are the following:

- **NIND** = 40; % Number of individuals in each subpopulation.
- **MAXGEN** = 100; % Maximum number of generations allowed.
- **GGAP** = .6; % "Generational gap", which is the percentage from the complete population of new individuals generated in each generation.
- **PRECI** = 120; % Precision of binary representations.
- **SelCh** = select('rws', Chrom, FitnV, GGAP); % Roulette wheel method for selecting the individuals participating in the genetic operations.
- **SelCh** = recombin('xovmp',SelCh,0.7); % Multi-point crossover as recombination method for the selected individuals.
- **ObjV** = FuncionObjDifuso120_555(Chrom, sdifuso); Objective function is given by the error between the performance of the ideal control system given by the experts and the fuzzy control system given by the genetic algorithm.

Representation of the Chromosome

In Table 1 we show the chromosome representation, which has 120 binary positions. These positions are divided in two parts, the first one indicates the number of rules of the fuzzy inference system, and the second one is divided again into fuzzy rules to indicate which membership functions are active or inactive for the corresponding rule.

Hybrid Soft Computing Models for Systems Modeling and Control, Table 1
Binary chromosome representation

Bit assigned	Representation
1 to 12	Which rule is active or inactive
13 to 21	Membership functions active or inactive of rule 1
22 to 30	Membership functions active or inactive of rule 2
…	Membership functions active or inactive of rule …
112 to 120	Membership functions active or inactive of rule 12

Simulation Results for the Case of Anesthesia Control

We describe in this section the simulation results that were achieved using the hierarchical genetic algorithm for the optimization of the fuzzy control system, for the case of anesthesia control. The genetic algorithm is able to evolve the topology of the fuzzy system for the particular application. We used 50 generations of 40 individuals each to achieve the minimum error. We show in Fig. 5 the final results of the genetic algorithm, where the error has been minimized. This is the case in which only nine fuzzy rules are needed for the fuzzy controller. The value of the minimum error achieved with this particular fuzzy logic controller was of 0.0064064, which is considered a small number in this application.

In Fig. 6 we show the simulation results of the fuzzy logic controller produced by the genetic algorithm after evolution. We used a sinusoidal input signal with unit amplitude and a frequency of 2 radians/second, with a transfer function of $[1/(0.5\,s + 1)]$. In this figure we can appreciate the comparison of the outputs of both the ideal controller (1) and the fuzzy controller optimized by the genetic algorithm (2). From this figure it is clear that both controllers are very similar and as a consequence we can conclude that the genetic algorithm was able to optimize

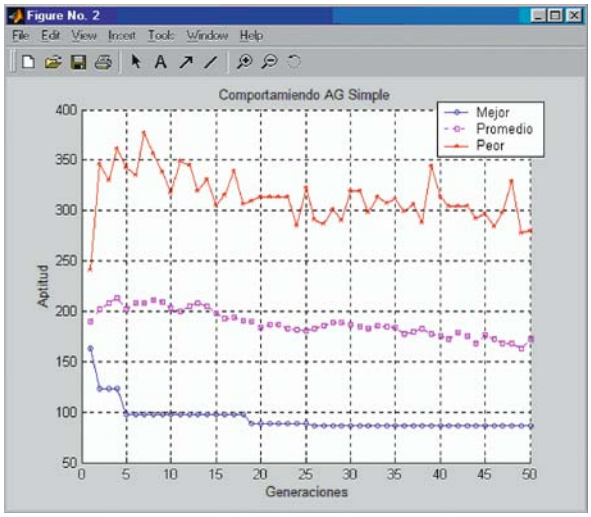

Hybrid Soft Computing Models for Systems Modeling and Control, Figure 5
Plot of the error after 50 generations of the HGA

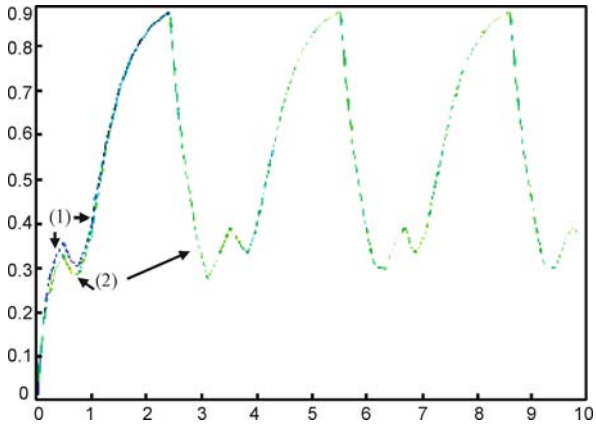

Hybrid Soft Computing Models for Systems Modeling and Control, Figure 6
Comparison between outputs of the ideal controller (1) and the fuzzy controller with the HGA (2)

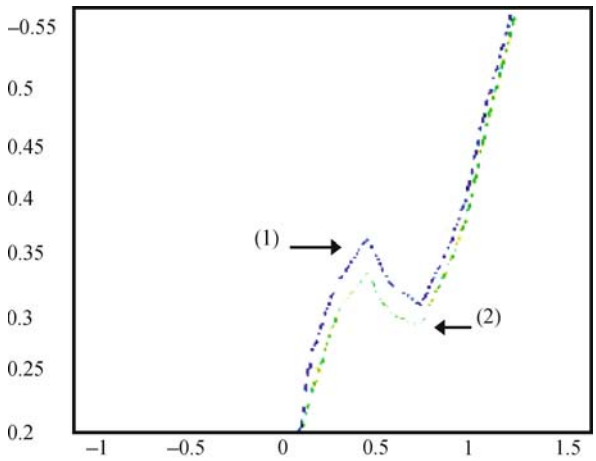

Hybrid Soft Computing Models for Systems Modeling and Control, Figure 7
Zoom in of Fig. 6 to view in more detail the difference between the controllers

the performance of the fuzzy logic controller. We can also appreciate this fact more clearly in Fig. 7, where we have amplified the simulation results from Fig. 6 for a better view.

Application to the Control of the Bar and Ball System

In this section, we describe the ball and beam experiment [23], which was also used as a basis for testing the genetic approach of fuzzy controller optimization.

A ball is placed on a beam, see Fig. 8, where it is allowed to roll with one degree of freedom along the length of the beam. A lever arm is attached to the beam at one

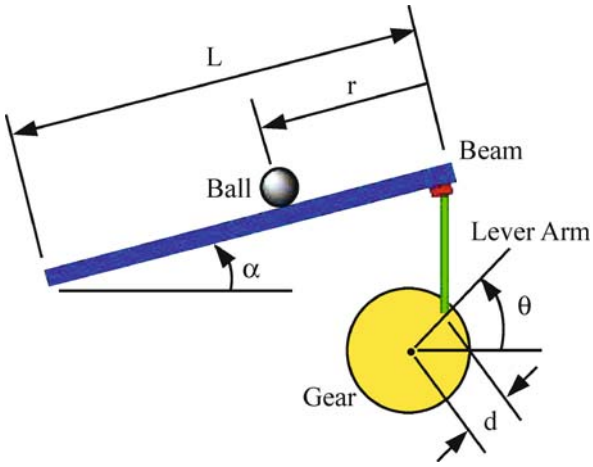

Hybrid Soft Computing Models for Systems Modeling and Control, Figure 8
Diagram of the ball and beam system

end and a servo gear at the other. As the servo gear turns by an angle theta, the lever changes the angle of the beam by a magnitude "alpha". When the angle is changed from the horizontal position, gravity causes the ball to roll along the beam. A controller will be designed for this system so that the ball's position can be manipulated.

For this problem, we will assume that the ball rolls without slipping and friction between the beam and ball is negligible. The constants for this example are defined as follows:

M mass of the ball 0.11 kg
R radius of the ball 0.015 m
d lever arm offset 0.03 m
g gravitational acceleration 9.8 m/s^2
L length of the beam 1.0 m
J moment of inertia 9.99×10^{-6} kg m^2
r ball position coordinate
α beam angle coordinate
θ servo gear angle

The Lagrangian equation of motion for the ball is then given by the following equation:

$$(J/R^2 + m)r'' + mg\sin\alpha - mr(\alpha')^2 = 0\,. \tag{8}$$

Simulation Results with the Complete HGA

- NIND = 50; % Number of individuals in the population
- MAXGEN = 80; % Maximum number of generations
- GGAP = 0.8; % Generation gap
- PRECI = 120; % Length of the Chromosome

In this case, we use the complete HGA with the following parameters:

Hybrid Soft Computing Models for Systems Modeling and Control, Table 2
Fuzzy rules in indexed and linguistic form

Indexed Rules	Linguistic Rules
1 1, 1 (1): 1	If error is N and derror is N then Angle is NG
1 2, 2 (1): 1	If error is N and derror is Z then Angle is N
2 1, 2 (1): 1	If error is Z and derror is N then Angle is N
2 2, 3 (1): 1	If error is Z and derror is Z then Angle is Z
2 3, 4 (1): 1	If error is Z and derror is P then Angle is P
3 2, 4 (1): 1	If error is P and derror is Z then Angle is P
3 3, 5 (1): 1	If error is P and derror is P then Angle is PG

Hybrid Soft Computing Models for Systems Modeling and Control, Table 3
Fuzzy rules in indexed and linguistic form

Indexed Rules	Linguistic Rules
1 1, 1 (1): 1	If error is N and derror is N then Angle is NG
1 2, 2 (1): 1	If error is N and derror is Z then Angle is N
2 1, 2 (1): 1	If error is Z and derror is N then Angle is N
2 2, 3 (1): 1	If error is Z and derror is Z then Angle is Z
2 3, 4 (1): 1	If error is Z and derror is P then Angle is P
3 2, 4 (1): 1	If error is P and derror is Z then Angle is P

At the end of the genetic evolution seven fuzzy rules were obtained, which are shown in Table 2.

Simulation Results with a Method Dased Only on Mutation

In this section, we show results for the HGA method based only on mutation (no crossover was used). The parameters of the genetic algorithm are:

- NIND = 50; % Number of individuals in the population
- MAXGEN = 250; % Maximum number of generations
- GGAP = 0.8; % Generation gap
- PRECI = 120; % Length of the chromosome

Table 3 shows the fuzzy rules obtained at the end with this type of genetic algorithm.

Simulation Results with a Method Based Only on Crossover

In this section, we show results for the HGA method based only on crossover (no mutation was used). The parameters of the genetic algorithm are the same. The fuzzy rules obtained at the end of the genetic evolution are shown in Table 4.

Hybrid Soft Computing Models for Systems Modeling and Control, Table 4
Fuzzy rules in indexed and linguistic form

Indexed rules	Linguistic rules
1 1, 1 (1): 1	If error is N and derror is N then Angle is NG
1 3, 3 (1): 1	If error is N and derror is P then Angle is Z
2 1, 2 (1): 1	If error is Z and derror is N then Angle is N
2 3, 4 (1): 1	If error is Z and derror is P then Angle is P
3 2, 4 (1): 1	If error is P and derror is Z then Angle is P
3 3, 5 (1): 1	If error is P and derror is P then Angle is PG

Comparison of the Simulation Results

In this section we show the comparison of the fuzzy controllers that were obtained with the different types of genetic algorithms that were considered. We show in Figs. 9 and 10 the comparison of the responses of the different fuzzy controllers. From these figures we can appreciate that the different types of HGA work. However, a pseudobacterial genetic algorithm (PBGA) does not give a valid fuzzy controller. The PBGA is not described in this paper, but can be seen with detail in [16] and [18].

Table 5 shows the comparison between the different methods used. This table summarizes response times of the controllers and the number of fuzzy rules of the corresponding controllers. This table also contains the information of a PID controller. We also show in Fig. 11 the behavior of the PID controller to have a basis for comparison with the other controllers.

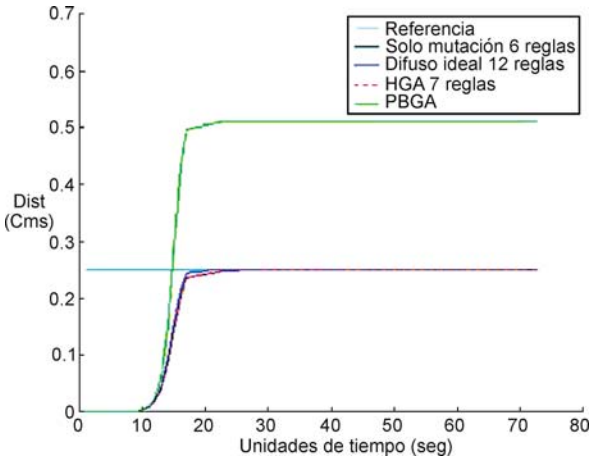

Hybrid Soft Computing Models for Systems Modeling and Control, Figure 9
Responses of the fuzzy controllers obtained (*green* = PBGA, *red* = HGA complete, *blue* = ideal fuzzy, *black* = HGA with only mutation)

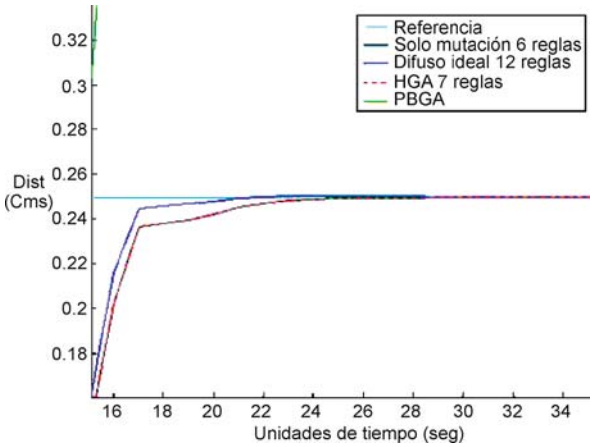

Hybrid Soft Computing Models for Systems Modeling and Control, Figure 10
Magnification of Fig. 9

Hybrid Soft Computing Models for Systems Modeling and Control, Table 5
Comparison of the different methods

Methods	Time	Reference	Number of rules
Only mutation	12.000	0.25	6
Ideal Fuzzy System (Not optimized)	13.8000	0.25	12
Complete HGA	12.000	0.25	7
PBGA	Did not control	0.25	9
Only crossover	Did not control	0.25	7
Traditional (PID)	400	0.25	Transfer Function

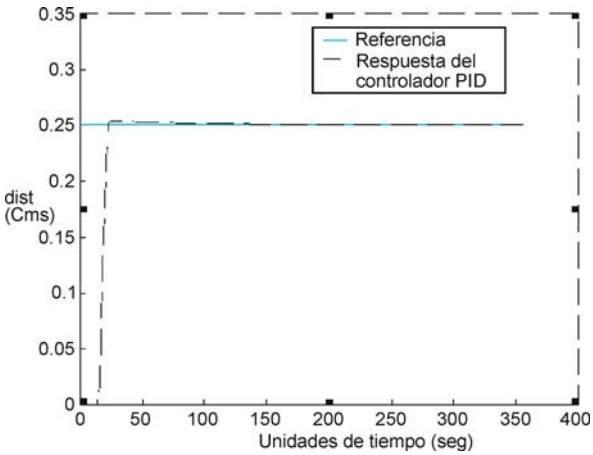

Hybrid Soft Computing Models for Systems Modeling and Control, Figure 11
Response of the PID controller (*blue* = reference, *black* = response of the controller)

We can appreciate from Table 5 that the best result corresponds to the method using only mutation, because it achieved a lower number of fuzzy rules without losing the efficiency of the control goal. The complete HGA method also achieves control but with one more fuzzy rule. The ideal fuzzy controller (from the experts) also achieves control but with 12 fuzzy rules. The other methods do not achieve the control goal.

Hierarchical Genetic Algorithms for Neural Networks

The bottleneck problem for NN application lies within the optimization procedures that are used to obtain an optimal NN topology. Hence, the formulation of the Hierarchical Genetic Algorithm (HGA) is applied for this purpose. The HGA differs from the standard GA with a hierarchy structure in that each chromosome consists of multilevel genes. Each chromosome consists of two types of genes, i. e. control genes and connection genes. The control genes in the form of bits, are the genes for layers and neurons for activation. The connection genes, a real value representation, are the genes for connection weightings and neuron bias.

With such a specific treatment, a structural chromosome incorporates both active and inactive genes. It should be noted that the inactive genes remain in the chromosome structure and can be carried forward for further generations. Such an inherent genetic variation in the chromosome avoids any trapping at local optima, which has the potential to cause premature convergence. Thus it maintains a balance between exploiting its accumulated knowledge and exploring the new areas of the search space. This structure also allows larger genetic variations in chromosome while maintaining high viability by permitting multiple simultaneous genetic changes. As a result, a single change in high level genes will cause multiple changes (activation or deactivation in the whole level) in lower level genes. In the case of the traditional GA, this is only possible when a sequence of many random changes takes place. Hence the computational power is greatly improved.

The fitness function used in this work combines the information the error objective and also the information about the number of nodes as a second objective. This is shown in the following equation.

$$f(z) = \left(\frac{1}{\alpha * \text{Ranking}(\text{ObjV1}) + \beta * \text{ObjV2}} \right) * 10 \,. \quad (9)$$

The first objective is basically the average sum of squared of errors as calculated by the predicted outputs of the MNN compared with real values of the function. This is

given by the following equation.

$$f_1 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - y_i)\,. \tag{10}$$

The parameters of the genetic algorithm for this case are as follows:

- Type of crossover operator: Two-point crossover
- Crossover rate: 0.8
- Type of mutation operator: Binary mutation
- Mutation rate: 0.05
- Population size per generation: 10
- Total number of generations: 100

The evolution of neural networks is used in the following section to optimize hybrid intelligent systems for time series prediction. In particular, in the neuro-genetic approach the evolution is used to optimize the neural network for prediction, and in the neuro-fuzzy-genetic approach the evolution is used to optimize the neuro-fuzzy system.

Experimental Results for Time Series Prediction

In this section, we illustrate the application of interval type-2 fuzzy logic to the problem of time series prediction. The Mackey-Glass time series is used to compare the results of interval type-2 fuzzy logic with the results of other intelligent methods. In particular, a comparison is made with type-1 fuzzy systems, neural networks, neuro-fuzzy systems and neuro-genetic and fuzzy-genetic approaches.

Prediction with a Neural Network

In this case, we did find a neural network model for this time series using 5 time delays and 6 periods. The architecture of the neural network used was 4–12–1 with 150 epochs, and 500 data points for training/500 data points for validation. The mean square error for prediction is 0.0043 (see Fig. 12).

Prediction with an Adaptive Neuro–Fuzzy Inference System (ANFIS)

To find the ANFIS model an analysis was made of the data, and the decision was to use a 5 time delay with 6 periods. The time series was divided into 500 data points for training and 500 data points for validation. The ANFIS model was designed with 4 inputs (2 membership functions each) and 1 output with 16 linear functions, giving a total of 16 fuzzy rules. The training was of 50 epochs and the mean squared error of prediction is of 0.0016 (see Fig 13).

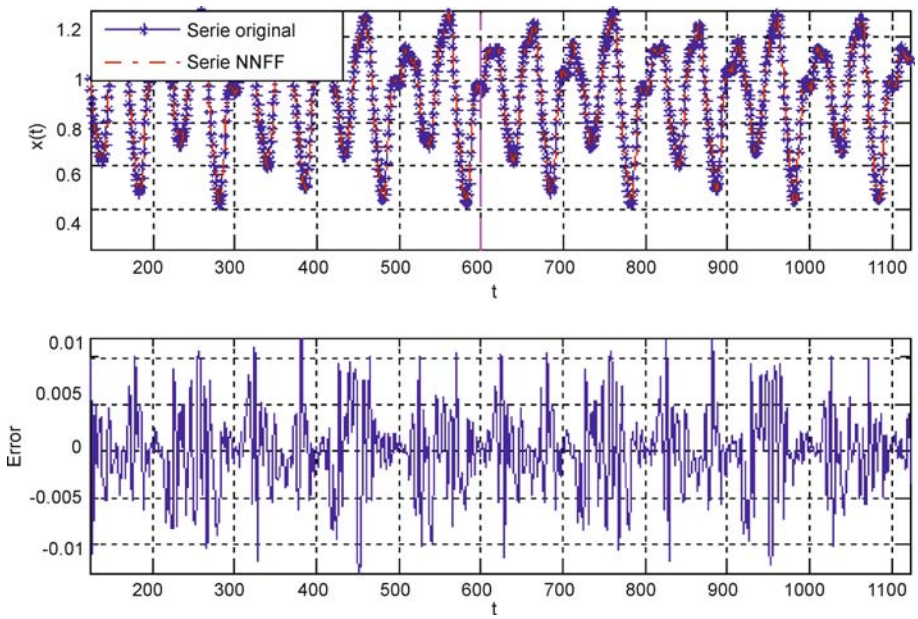

Hybrid Soft Computing Models for Systems Modeling and Control, Figure 12
Forecasting the Mackey-Glass time series with a neural network. *On top* it is shown the comparison of the prediction with the neural network (*red*) and the original time series (*blue*). *Below* it is shown the forecasting error

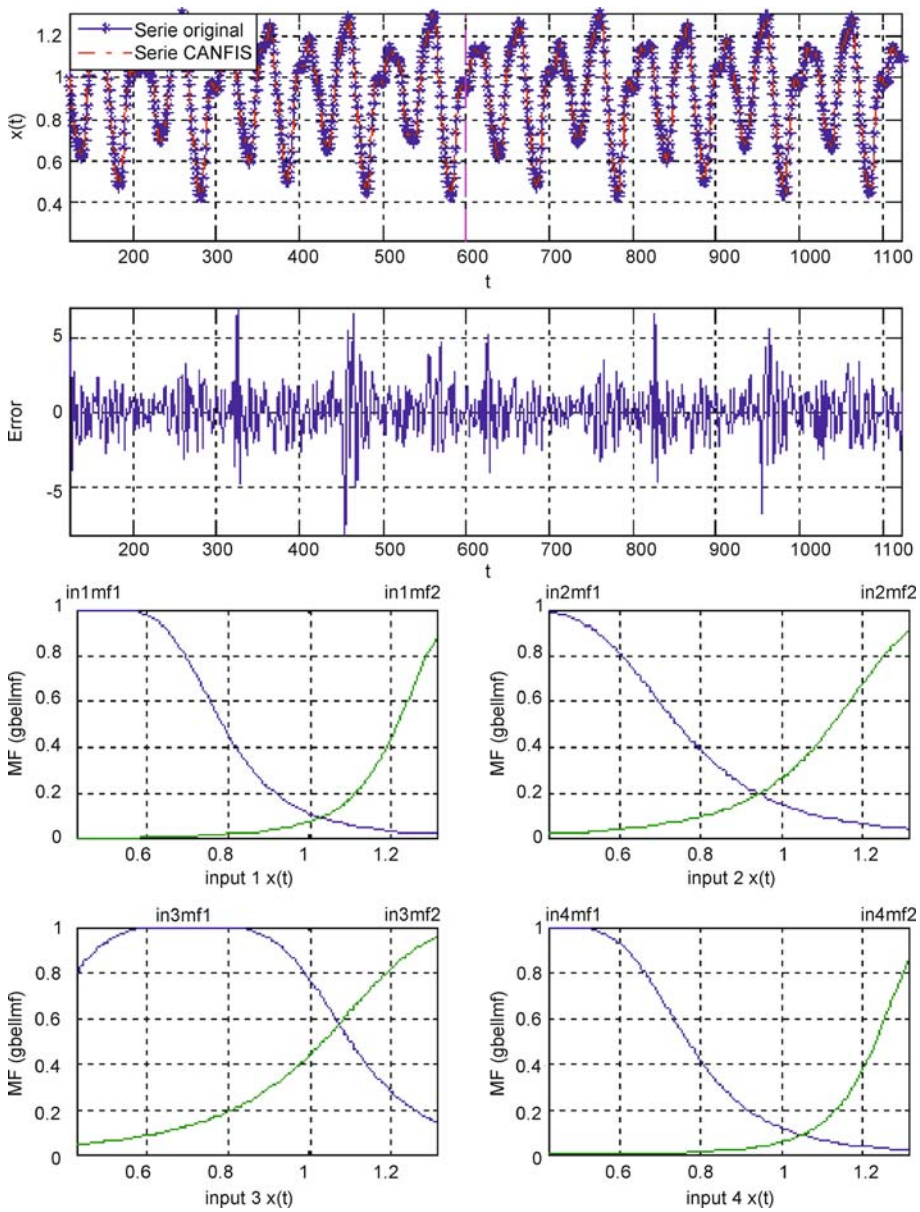

Hybrid Soft Computing Models for Systems Modeling and Control, Figure 13
Forecasting the Mackey-Glass time series with ANFIS. The figure on *top* shows the comparison of the predicted values and the original time series. The *following* figures show the error plot and the membership functions obtained with ANFIS

Prediction with an Interval Type-2 TSK Fuzzy Model

To find the type-2 fuzzy model an analysis was made of the data, and the decision was to use a 5 time delay with 6 periods. The time series was divided into 500 data points for training and 500 data points for validation. The model was designed with 4 inputs (with two interval type-2 (igbellmtype2) membership functions each) and one output with 16 interval linear functions, giving a total of 16 fuzzy rules. The training was of 50 epochs and the mean squared error of prediction is of 0.00023 (see Fig. 14).

Prediction with a Neuro-Genetic Model

In this case, we use a genetic algorithm for training a neural network model for this time series using 5 time de-

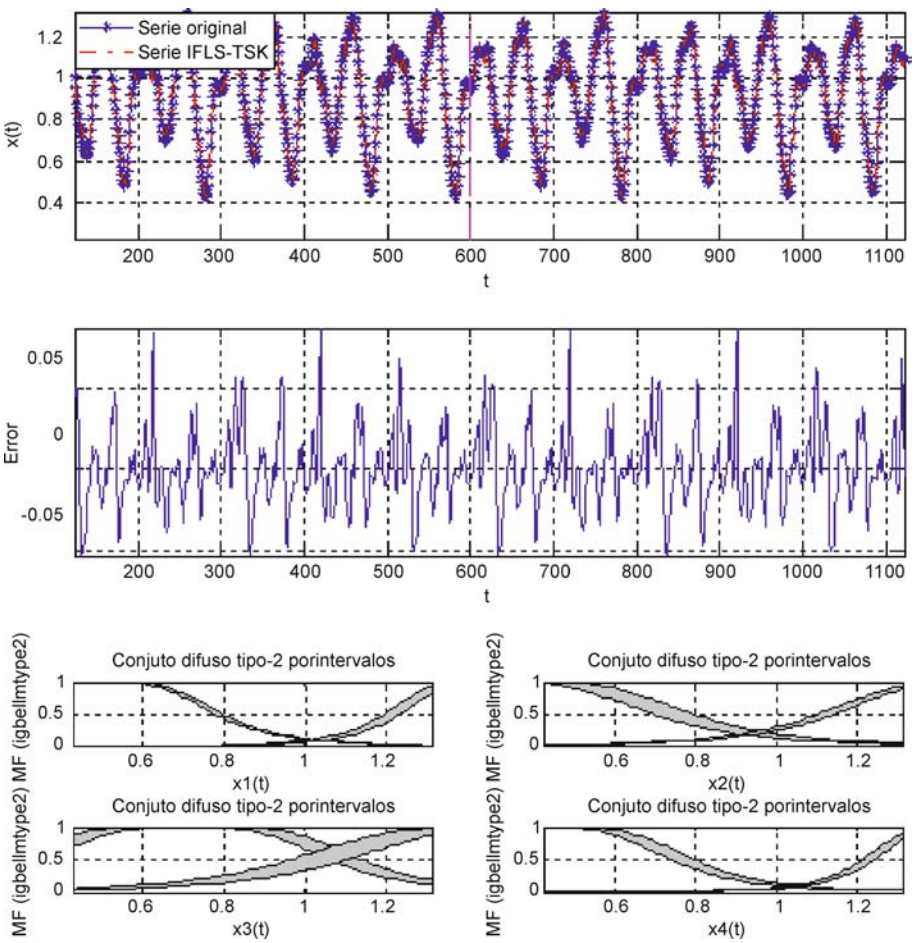

Hybrid Soft Computing Models for Systems Modeling and Control, Figure 14
Forecasting the Mackey-Glass time series with an interval type-2 TSK fuzzy model. The figure on *top* shows the comparison of the predicted values and the original time series. The *following* figures show the error plot and the interval type-2 membership functions obtained

lays and 6 periods. The architecture of the neural network used was 4–12–1 with 200 generations, and 500 data points for training/500 data points for validation. The mean square error for prediction is 0.00064 (see Fig. 15). The genetic parameters are: population size = 30, crossover ratio = 0.75, and mutation ratio = 0.01.

Prediction with Type-1 Fuzzy Models Optimized with Genetic Algorithms

In this section, we show prediction results of type-1 fuzzy models that were optimized using genetic algorithms. In all cases, we have 4 inputs (with 2 membership functions) and 1 output with 16 functions. The parameters of the genetic algorithms are the same as in the previous section. We show in Fig. 16 the results of a TSK model (mean squared error of 0.00647) and in Fig. 17 the results of a Mamdani model (mean squared error of 0.00692).

The Mackey-Glass time series shows chaotic behavior an for this reason has been chosen many times as a benchmark problem for prediction methods. We show in Table 6 a summary of the results using the methods mentioned previously. Based on mean squared error (RMSE) of forecasting, we can conclude that the interval type-2 fuzzy model (IT2 FLS on Table 6) is the best one to predict future values of this time series. Also, based on the training required, we can conclude that the interval type-2 fuzzy model is the best one because it only requires one epoch.

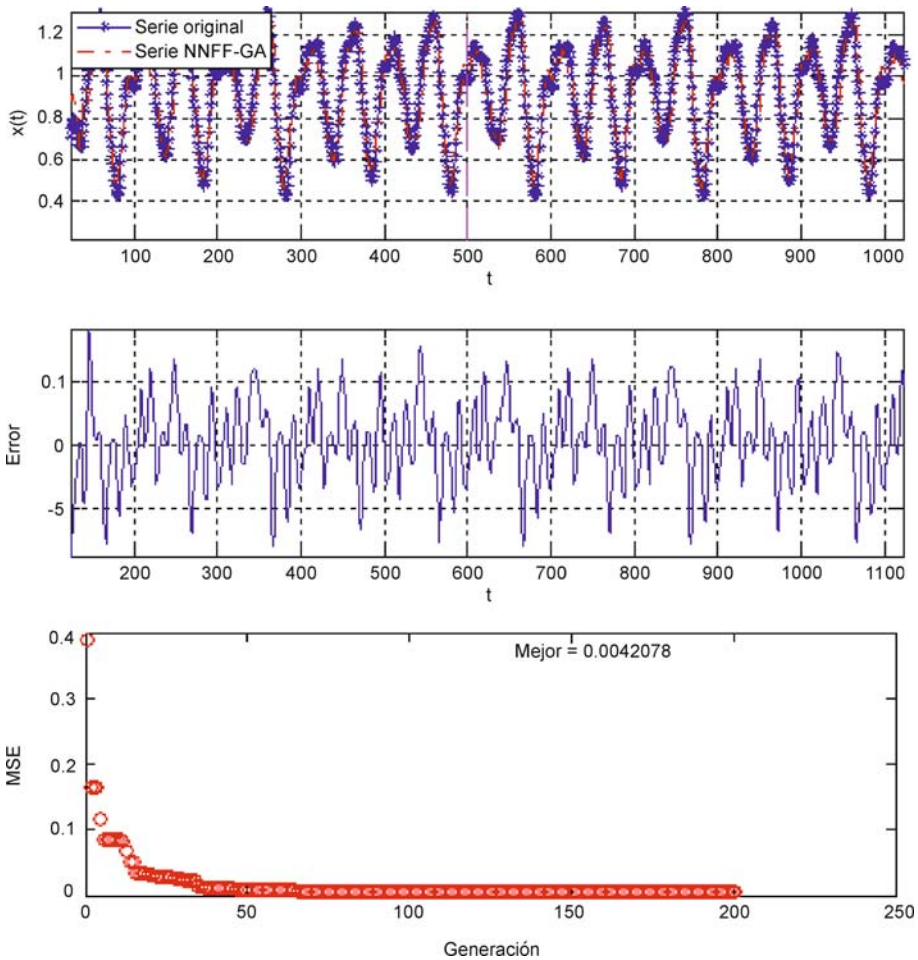

Hybrid Soft Computing Models for Systems Modeling and Control, Figure 15
Forecasting the Mackey-Glass time series with a neuro-genetic model. The figure on *top* shows the comparison of the predicted values and the original time series. The figure on the *middle* shows the error plot. The figure on the *bottom* shows the evolution of the mean squared error as the generations increase up to 250

Hybrid Soft Computing Models for Systems Modeling and Control, Table 6
Summary of results for the mackey-glass time series

Method	RMSE	Data Training/ Checking	Epochs or Generations
NNFF (Fig. 12)	0.00430	500/500	150
CANFIS (Fig. 13)	0.00160	500/500	50
IT2FLS(TSK) (Fig. 14)	0.00023	500/500	1
NNFF-GA (Fig. 15)	0.00064	500/500	150
FLS(TSK)-GA (Fig. 16)	0.00647	500/500	200
FLS(MAM)-GA (Fig. 17)	0.00693	500/500	200

Finally, from Table 6 we can say that the use of evolution helps in the design of hybrid intelligent systems. The results of the hybrid intelligent systems shown in the last three rows (of Table 6) are very good, and this is a consequence of using genetic algorithm for optimizing the neural networks or the fuzzy systems.

Conclusions

We consider in this paper the case of automatic anesthesia control in human patients for testing the optimized fuzzy controller. We did have, as a reference, the best fuzzy controller that was developed for the automatic anesthesia control, and we consider the optimization of this controller using the HGA approach. After applying the genetic algorithm the number of fuzzy rules was reduced from 12 to 9 with a similar performance of the fuzzy controller.

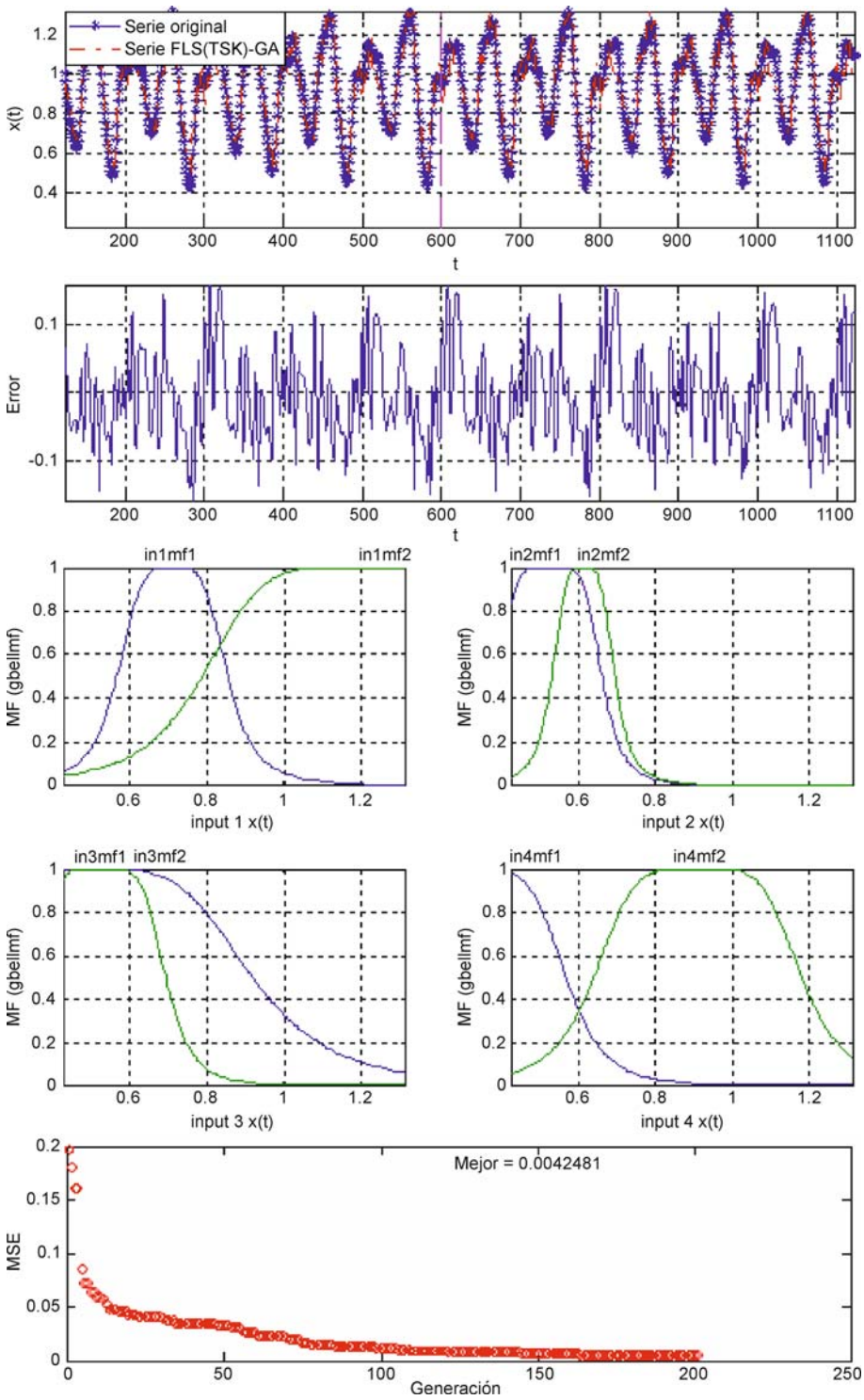

Hybrid Soft Computing Models for Systems Modeling and Control, Figure 16
Forecasting the Mackey-Glass time series with a TSK fuzzy model optimized using a genetic algorithm

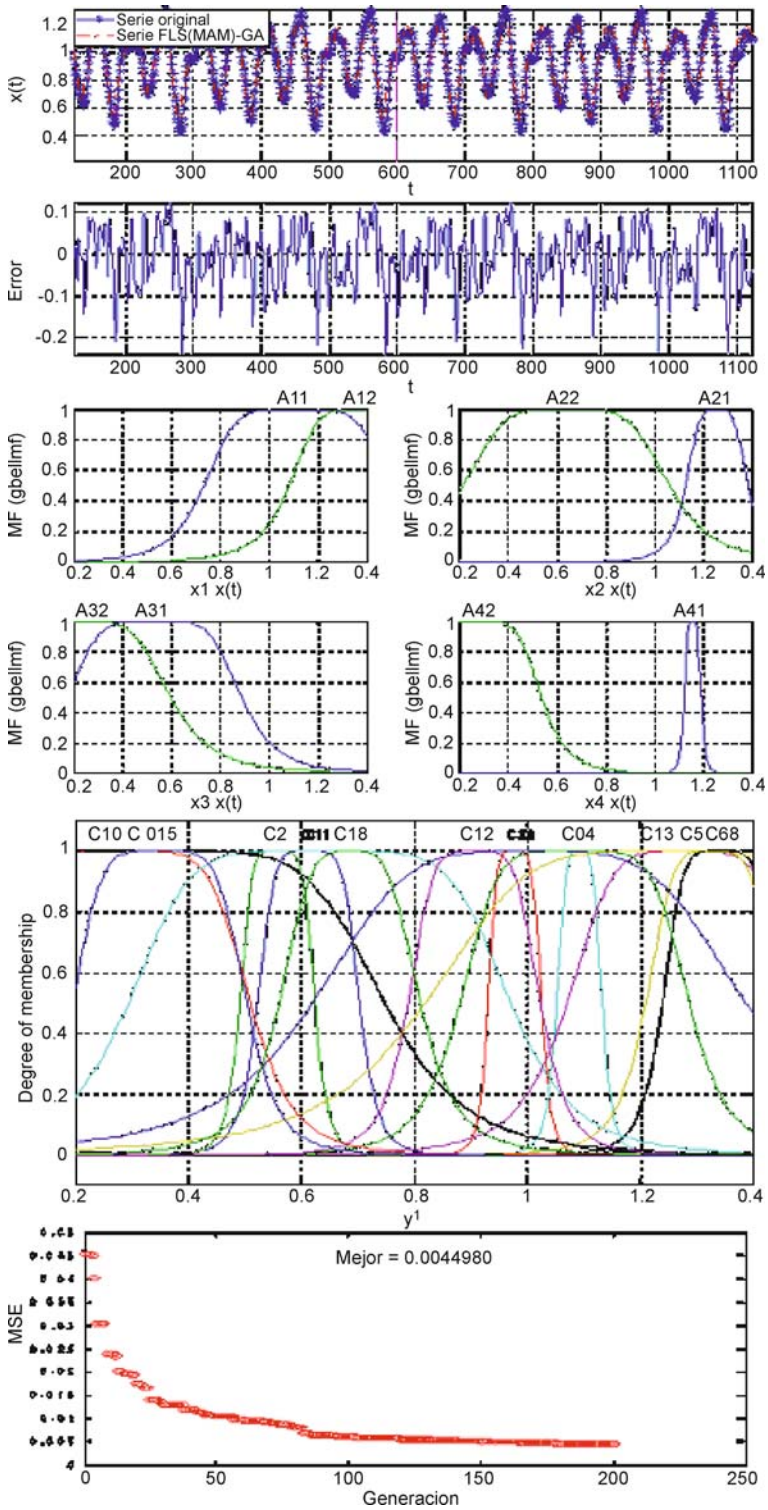

Hybrid Soft Computing Models for Systems Modeling and Control, Figure 17
Forecasting the Mackey-Glass time series with a Mamdani fuzzy model optimized using a genetic algorithm

Of course, the parameters of the membership functions were also tuned by the genetic algorithm. We did compare the simulation results of the optimized fuzzy controllers obtained with the HGA against the best fuzzy controller that was obtained previously with expert knowledge, and control is achieved in a similar fashion. Since simulation results are similar, and the number of fuzzy rules was reduced, we can conclude that the HGA approach is a good alternative for designing fuzzy systems. We also consider the case of controlling the bar and ball system, and the genetic approach was able to optimize the number of rules from 12 to 6. In conclusion, the HGA approach is a good alternative in optimizing fuzzy controllers. Future work will include testing the proposed approach with the optimization of other fuzzy controllers. We also described the application of the evolutionary approach for the problem of designing hybrid intelligent systems in time series prediction. In this case, the goal is to design the best predictor for complex time series. Simulation results show that the evolutionary approach optimizes the hybrid intelligent systems in time series prediction.

Future Directions

The evolutionary approach is a good alternative in optimizing fuzzy controllers. Future work will include testing the proposed approach with the optimization of other fuzzy controllers and comparison with results of existing approaches. We also described the application of the evolutionary approach for the problem of designing hybrid intelligent systems in time series prediction. In this case, the goal is to design the best predictor for complex time series. Simulation results show that the evolutionary approach optimizes the hybrid intelligent systems in time series prediction. In this case, future work will include applying the hybrid approach to other problems of time series prediction, and compare the results with existing approaches.

Bibliography

1. Baruch S, Garrido R (2005) A direct adaptive neural control scheme with integral terms. Int J Intell Syst 20(2):213224
2. Castillo O, Melin P (2001) Soft computing for control of nonlinear dynamical systems. Springer, Heidelberg
3. Castillo O, Melin P (2003) Soft computing and fractal theory for intelligent manufacturing. Springer, Heidelberg
4. Castillo O, Huesca G, Valdez F (2004) Evolutionary computing for fuzzy system optimization in intelligent control. Proceedings of IC-AI'04, Las Vegas, vol 1. CSREA Press, Las Vegas, pp 98–104
5. Davis L (1991) Handbook of genetic algorithms. Van Nostrand Reinhold, New York
6. Goldberg D (ed) (1989) Genetic algorithms in search, optimization and machine learning. Addison Wesley, Reading
7. Holland J (1975) Adaptation in natural and artificial systems. University of Michigan Press, Ann Arbor
8. Homaifar, McCormick E (1995) Simultaneous design of membership functions and rule sets for fuzzy controllers using genetic algorithms. IEEE Trans Fuzzy Syst 3:129–139
9. Jang JSR, Sun CT (1995) Neurofuzzy fuzzy modeling and control. Proc IEEE 83:378–406
10. Jang JSR, Sun CT, Mizutani E (1997) Neuro-fuzzy and soft computing, a computational approach to learning and machine intelligence. Prentice Hall, Upper Saddle River
11. Karr CL, Gentry EJ (1993) Fuzzy control of pH using genetic algorithms. IEEE Trans Fuzzy Systems 1:46–53
12. Langari R (1990) A framework for analysis and synthesis of fuzzy linguistic control systems. Ph D thesis, University of California, Berkeley
13. Lozano (2004) Optimización de un sistema de control difuso por medio de algoritmos genéticos jerarquicos. Thesis, Dept of Computer Science, Tijuana Institute of Technology
14. Man KF, Tang KS, Kwong S (1999) Genetic algorithms: Concepts and designs. Springer, London
15. Melin P, Castillo O (2002) Modelling, simulation and control of non-linear dynamical systems. Taylor and Francis, London
16. Nawa NE, Furuhashi T (1999) Fuzzy system parameters discovery by bacterial evolutionary algorithm. IEEE Trans Fuzzy Syst 7:608–616
17. Procyk TJ, Mamdani EM (1979) A linguistic self-organizing process controller. Automatica 15(1):15–30
18. Salmeri M, Re M, Petrongari E, Cardarilli GC (1999) A novel bacterial algorithm to extract the rule base from a training set. Technical Report, Dept. of Electronic Engineering, University of Rome
19. Sepulveda R, Castillo O, Melin P, Montiel O, Rodriguez-Diaz A (2005) Handling uncertainty in controllers using type-2 fuzzy logic. J Intell Syst 14:237–262
20. Tang KS, Man KF, Liu ZF, Kwong S (1998) Minimal fuzzy memberships and rules using hierarchical genetic algorithms. IEEE Trans Ind Electron 45(1):142–150
21. Vachtsevanos G, Farinwata S (1996) Fuzzy logic control. In: Patyra MJ, Mlynek DM (eds) A systematic design and performance assessment methodology. Wiley, New York
22. Valdes M, Gomez-Skarmeta AF, Botia JA (2005) Toward a framework for the specification of hybrid fuzzy modeling. Int J Intell Syst 20(2):225–252
23. Valdez F, Castillo O (2004) Comparative study of evolutionary computing methods for fuzzy system optimization in intelligent control. Proc IS-IC'04. Tijuana, México, pp 1–5
24. Yen J, Langari R(1999) Fuzzy logic: intelligence, control and information. Prentice Hall, Upper Saddle River
25. Yoshikawa T, Furuhashi T, Uchikawa Y (1996) Emergence of effective fuzzy rules for controlling mobile robots using DNA coding method. Proc ICEC'96. Nagoya, Japan, pp 581–586
26. Zadeh L (1965) Fuzzy sets. J Inf Control 8:338–353
27. Zadeh L (1987) Fuzzy sets and applications. In: Yager RR, Ovchinnikov S, Tong RM, Nguyen HT (eds) Selected papers. Wiley, New York

Hyperbolic Conservation Laws

Alberto Bressan
Department of Mathematics, Penn State University, University Park, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Examples of Conservation Laws
Shocks and Weak Solutions
Hyperbolic Systems in One Space Dimension
Entropy Admissibility Conditions
The Riemann Problem
Global Solutions
Hyperbolic Systems in Several Space Dimensions
Numerical Methods
Future Directions
Bibliography

Glossary

Conservation law Several physical laws state that certain basic quantities such as mass, energy, or electric charge, are globally conserved. A conservation law is a mathematical equation describing how the density of a conserved quantity varies in time. It is formulated as a partial differential equation having divergence form.

Flux function The flux of a conserved quantity is a vector field, describing how much of the given quantity moves across any surface, at a given time.

Shock Solutions to conservation laws often develop shocks, i. e. surfaces across which the basic physical fields are discontinuous. Knowing the two limiting values of a field on opposite sides of a shock, one can determine the speed of propagation of a shock in terms of the Rankine–Hugoniot equations.

Entropy An entropy is an additional quantity which is globally conserved for every smooth solution to a system of conservation laws. In general, however, entropies are not conserved by solutions containing shocks. Imposing that certain entropies increase (or decrease) in correspondence to a shock, one can determine a unique physically admissible solution to the mathematical equations.

Definition of the Subject

According to some fundamental laws of continuum physics, certain basic quantities such as mass, momentum, energy, electric charge..., are globally conserved. As time progresses, the evolution of these quantities can be described by a particular type of mathematical equations, called conservation laws.

Gas dynamics, magneto-hydrodynamics, electromagnetism, motion of elastic materials, car traffic on a highway, flow in oil reservoirs, can all be modeled in terms of conservation laws. Understanding, predicting and controlling these various phenomena is the eventual goal of the mathematical theory of hyperbolic conservation laws.

Introduction

Let $u = u(x,t)$ denote the density of a physical quantity, say, the density of mass. Here t denotes time, while $x = (x_1, x_2, x_3) \in \mathbb{R}^3$ is a three-dimensional space variable. A *conservation law* is a partial differential equation of the form

$$\frac{\partial}{\partial t} u + \operatorname{div} \mathbf{f} = 0 \,. \tag{1}$$

which describes how the density u changes in time. The vector field $\mathbf{f} = (f_1, f_2, f_3)$ is called the *flux* of the conserved quantity. We recall that the divergence of $\mathbf{f}$ is

$$\operatorname{div} \mathbf{f} = \frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2} + \frac{\partial f_3}{\partial x_3} \,.$$

To appreciate the meaning of the above Eq. (1), consider a fixed region $\Omega \subset \mathbb{R}^3$ of the space. The total amount of mass contained inside Ω at time t is computed as

$$\int_\Omega u(x,t)\mathrm{d}x \,.$$

This integral may well change in time. Using the conservation law (1) and then the divergence theorem, one obtains

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega u(x,t)\mathrm{d}x = \int_\Omega \frac{\partial}{\partial t} u(x,t)\mathrm{d}x &= -\int_\Omega \operatorname{div}\mathbf{f}\mathrm{d}x \\ &= -\int_\Sigma \mathbf{f}\cdot\mathbf{n}\mathrm{d}\Sigma \,.\end{aligned}$$

Here Σ denotes the boundary of Ω, while the integrand $\mathbf{f}\cdot\mathbf{n}$ denotes the inner product of the vector $\mathbf{f}$ with the unit outer normal $\mathbf{n}$ to the surface Σ. According to the above identities, no mass is created or destroyed. The total amount of mass contained inside the region Ω changes in time only because some of the mass flows in or out across the boundary Σ.

Assuming that the flux $\mathbf{f}$ can be expressed as a function of the density u alone, one obtains a closed equation. If the initial density $\bar{u}$ at time $t = 0$ is known, then the values

of the function $u = u(x, t)$ at all future times $t > 0$ can be found by solving the initial-value problem

$$u_t + \operatorname{div} \mathbf{f}(u) = 0 \qquad u(0, x) = \bar{u}(x) .$$

More generally, a *system of balance laws* is a set of partial differential equations of the form

$$\begin{cases} \frac{\partial}{\partial t} u_1 + \operatorname{div}\ \mathbf{f}_1(u_1, \ldots, u_n) = \phi_1 , \\ \qquad\qquad \ldots \\ \frac{\partial}{\partial t} u_n + \operatorname{div}\ \mathbf{f}_n(u_1, \ldots, u_n) = \phi_n . \end{cases} \tag{2}$$

Here $u_1, \ldots, u_n$ are the conserved quantities, $\mathbf{f}_1, \ldots, \mathbf{f}_n$ are the corresponding fluxes, while the functions $\phi_i = \phi_i(t, x, u_1, \ldots, u_n)$ represent the source terms. In the case where all ϕ_i vanish identically, we refer to (2) as a *system of conservation laws*.

Systems of this type express the fundamental balance equations of continuum physics, when small dissipation effects are neglected. A basic example is provided by the equations of non-viscous gases, accounting for the conservation of mass, momentum and energy. This subject is thus very classical, having a long tradition which can be traced back to Euler [21] and includes contributions by Stokes, Riemann, Weyl and von Neumann, among several others.

In spite of continuing efforts, the mathematical theory of conservation laws is still largely incomplete. Most of the literature has been concerned with two main cases: (i) a single conservation law in several space dimensions, and (ii) systems of conservation laws in one space dimension. For systems of conservation laws in several space dimensions, not even the global-in-time existence of solutions is presently known, in any significant degree of generality. Several mathematical studies are focused on particular solutions, such as traveling waves, multi-dimensional Riemann problems, shock reflection past a wedge, etc...

Toward a rigorous mathematical analysis of solutions, the main difficulty that one encounters is the lack of regularity. Due to the strong nonlinearity of the equations and the absence of dissipation terms with regularizing effect, solutions which are initially smooth may become discontinuous within finite time. In the presence of discontinuities, most of the classical tools of differential calculus do not apply. Moreover, the Eqs. (2) must be suitably reinterpreted, since a discontinuous function does not admit derivatives in a classical sense.

Topics which have been more extensively investigated in the mathematical literature are the following:

- Existence and uniqueness of solutions to the initial-value problem. Continuous dependence of the solutions on the initial data [8,10,24,29,36,49,57].
- Admissibility conditions for solutions with shocks, characterizing the physically relevant ones [23,38,39,46].
- Stability of special solutions, such as traveling waves, w.r.t. small perturbations [34,47,49,50,62].
- Relations between the solutions of a hyperbolic system of conservation laws and the solutions of various approximating systems, modeling more complex physical phenomena. In particular: vanishing viscosity approximations [6,20,28], relaxations [5,33,48], kinetic models [44,53].
- Numerical algorithms for the efficient computation of solutions [32,41,41,56,58].

Some of these aspects of conservation laws will be outlined in the following sections.

Examples of Conservation Laws

We review here some of the most common examples of conservation laws. Throughout the sequel, subscripts such as u_t, f_x will denote partial derivatives.

Example 1 (Traffic flow) Let $u(x, t)$ be the density of cars on a highway, at the point x at time t. This can be measured as the number of cars per kilometer (see Fig. 1). In first approximation, following [43] we shall assume that u is continuous and that the speed s of the cars depends only on their density, say $s = s(u)$. Given any two points a, b on the highway, the number of cars between a and b therefore varies according to

$$\begin{aligned} \int_a^b u_t(x, t)\mathrm{d}x &= \frac{\mathrm{d}}{\mathrm{d}t} \int_a^b u(x, t)\mathrm{d}x \\ &= [\text{inflow at } x = a] - [\text{outflow at } x = b] \\ &= s(u(a, t)) \cdot u(a, t) - s(u(b, t)) \cdot u(b, t) \\ &= -\int_a^b [s(u)u]_x \mathrm{d}x . \end{aligned}$$

Since the above equalities hold for all a, b, one obtains the conservation law in one space dimension

$$u_t + [s(u)u]_x = 0 , \tag{3}$$

where u is the conserved quantity and $f(u) = s(u)u$ is the flux function. Based on experimental data, an appropriate flux function has the form

$$f(u) = a_1 \left(\ln \frac{a_2}{u}\right) u \qquad (0 < u \le a_2) ,$$

for suitable constants $a_1, \sim a_2$.

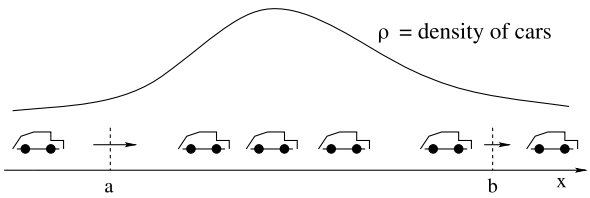

Hyperbolic Conservation Laws, Figure 1
Modelling the density of cars by a conservation law

Example 2 (Gas dynamics) The Euler equations for a compressible, non-viscous gas in Eulerian coordinates take the form

$$\begin{cases} \frac{\partial}{\partial t}\rho + \operatorname{div}(\rho v) & = 0 \\ \text{(conservation of mass)}\,, & \\ \frac{\partial}{\partial t}(\rho v_i) + \operatorname{div}(\rho v_i v) + \frac{\partial}{\partial x_i} p & = 0 \qquad i = 1, 2, 3, \\ \text{(conservation of momentum)} & \\ \frac{\partial}{\partial t}E + \operatorname{div}((E+p)v) & = 0 \\ \text{(conservation of energy)}\,. & \end{cases}$$

Here ρ is the mass density, $v = (v_1, v_2, v_3)$ is the velocity vector, E is the energy density, and p the pressure. In turn, the energy can be represented as a sum

$$E = \rho\frac{|v|^2}{2} + \rho e\,,$$

where the first term accounts for the kinetic energy while e is the internal energy density (related to the temperature). The system is closed by an additional equation $p = p(\rho, e)$, called the equation of state, depending on the particular gas under consideration [18]. Notice that here we are neglecting small viscous forces, as well as heat conductivity. Calling $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ the standard basis of unit vectors in $\mathbb{R}^3$, one has $\partial p/\partial x_i = \operatorname{div}(p\mathbf{e}_i)$. Hence all of the above equations can be written in the standard divergence form (1).

Example 3 (Isentropic gas dynamics in Lagrangian variables) Consider a gas in a tube. Particles of the gas will be labeled by a one-dimensional variable y determined by their position in a reference configuration with constant unit density. Using this Lagrangian coordinate, we denote by $u(y, t)$ the velocity of the particle y at time t, and by $v(y, t) = \rho^{-1}(y, t)$ its specific volume.

The so-called *p-system* of isentropic gas dynamics [57] consists of the two conservation laws

$$v_t - u_x = 0\,, \qquad u_t + p_x = 0\,. \tag{4}$$

The system is closed by an equation of state $p = p(v)$ expressing the pressure as a function of the specific volume. A typical choice here is $p(v) = kv^{-\gamma}$, with $\gamma \in [1, 3]$. In particular $\gamma \approx 1.4$ for air.

In general, p is a decreasing function of v. Near a constant state v_0, one can approximate p by a linear function, say $p(v) \approx p(v_0) - c^2(v - v_0)$. Here $c^2 = -p'(v_0)$. In this case the Eq. (4) reduces to the familiar wave equation

$$v_{tt} - c^2 v_{xx} = 0\,.$$

Shocks and Weak Solutions

A single conservation law in one space dimension is a first order partial differential equation of the form

$$u_t + f(u)_x = 0\,. \tag{5}$$

Here u is the *conserved quantity* while f is the *flux*. As long as the function u is continuously differentiable, using the chain rule the equation can be rewritten as

$$u_t + f'(u)u_x = 0\,, \tag{6}$$

According to (6), in the x–t-plane, the directional derivative of the function u in the direction of the vector $\mathbf{v} = (f'(u), 1)$ vanishes.

At time $t = 0$, let an initial condition $u(x, 0) = \bar{u}(x)$ be given. As long as the solution u remains smooth, it can be uniquely determined by the so-called method of characteristics. For every point x_0, consider the straight line (see Fig. 2)

$$x = x_0 + f'\big(\bar{u}(x_0)\big)t\,.$$

On this line, by (6) the value of u is constant. Hence

$$u\big(x_0 + f'\big(\bar{u}(x_0)\big)t\,, t\big) = \bar{u}(x_0)$$

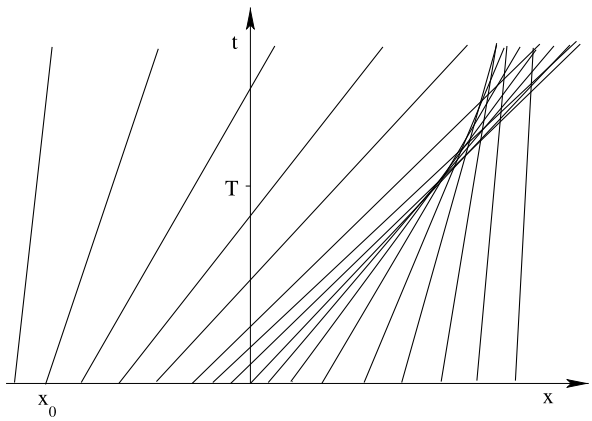

Hyperbolic Conservation Laws, Figure 2
Solving a conservation law by the method of characteristics. The function $u = u(x, t)$ is constant along each characteristic line

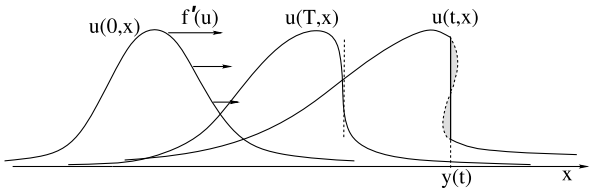

Hyperbolic Conservation Laws, Figure 3
Shock formation

for all $t \geq 0$. This allows to construct a solution up to the first time T where two or more characteristic lines meet. Beyond this time, the solution becomes discontinuous. Figure 3 shows the graph of a typical solution at three different times. Points on the graph of u move horizontally with speed $f'(u)$. If this speed is not constant, the shape of the graph will change in time. In particular, there will be an instant T at which one of the tangent lines becomes vertical. For $t > T$, the solution $u(\cdot, t)$ contains a shock. The position $y(t)$ of this shock can be determined by imposing that the total area of the region below the graph of u remains constant in time.

In order to give a meaning to the conservation law (6) when $u = u(x, t)$ is discontinuous, one can multiply both sides of the equation by a test function φ and integrate by parts. Assuming that φ is continuously differentiable and vanishes outside a bounded set, one formally obtains

$$\iint \{u\varphi_t + f(u)\varphi_x\} \mathrm{d}x\mathrm{d}t = 0 \,. \tag{7}$$

Since the left hand side of the above equation does not involve partial derivatives of u, it remains meaningful for a discontinuous function u. A locally integrable function u is defined to be a *weak solution* of the conservation law (6) if the integral identity (7) holds true for every test function φ, continuously differentiable and vanishing outside a bounded set.

Hyperbolic Systems in One Space Dimension

A system of n conservation laws in one space dimension can be written as

$$\begin{cases} \frac{\partial}{\partial t}u_1 + \frac{\partial}{\partial x}[\ f_1(u_1, \ldots, u_n)] = 0 \,, \\ \qquad\qquad \cdots \\ \frac{\partial}{\partial t}u_n + \frac{\partial}{\partial x}[\ f_n(u_1, \ldots, u_n)] = 0 \,. \end{cases} \tag{8}$$

For convenience, this can still be written in the form (5), but keeping in mind that now $u = (u_1, \ldots, u_n) \in \mathbb{R}^n$ is a vector and that $f = (f_1, \ldots, f_n)$ is a vector-valued function. As in the case of a single equation, a vector function $u = (u_1, \ldots, u_n)$ is called a *weak solution* to the system of conservation laws (8) if the integral identity (7) holds true, for every continuously differentiable test function φ vanishing outside a bounded set.

Consider the $n \times n$ Jacobian matrix of partial derivatives f at the point u:

$$A(u) \doteq Df(u) = \begin{pmatrix} \partial f_1/\partial u_1 & \cdots & \partial f_1/\partial u_n \\ & \cdots & \\ \partial f_n/\partial u_1 & \cdots & \partial f_n/\partial u_n \end{pmatrix} .$$

Using the chain rule, the system (8) can be written in the quasilinear form

$$u_t + A(u)u_x = 0 \,. \tag{9}$$

We say that the above system is *strictly hyperbolic* if every matrix $A(u)$ has n real, distinct eigenvalues, say $\lambda_1(u) < \cdots < \lambda_n(u)$. In this case, one can find a basis of right eigenvectors of $A(u)$, denoted by $r_1(u), \ldots, r_n(u)$, such that, for $i = 1, \ldots, n$,

$$A(u)r_i(u) = \lambda_i(u)r_i(u) \,, \qquad |r_i(u)| = 1 \,.$$

Example 4 Assume that the flux function is linear: $f(u) = Au$ for some constant matrix $A \in \mathbb{R}^{n\times n}$. If $\lambda_1 < \lambda_2 < \cdots < \lambda_n$ are the eigenvalues of A and $r_1, \ldots, r_n$ are the corresponding eigenvectors, then any vector function of the form

$$u(x, t) = \sum_{i=1}^{n} g_i(x - \lambda_i t)r_i$$

provides a weak solution to the system (8). Here it is enough to assume that the functions g_i are locally integrable, not necessarily continuous.

Example 3 (continued) For the system (4) describing isentropic gas dynamics, the Jacobian matrix of partial derivatives of the flux is

$$Df = \begin{pmatrix} -\partial u/\partial v & -\partial u/\partial u \\ \partial p/\partial v & \partial p/\partial u \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ p'(v) & 0 \end{pmatrix} .$$

Assuming that $p'(v) < 0$, this matrix has the two real distinct eigenvalues $\pm\sqrt{-p'(v)}$. Therefore, the system is strictly hyperbolic.

Entropy Admissibility Conditions

Given two states $u^-, u^+ \in \mathbb{R}^n$ and a speed λ, consider the piecewise constant function defined as

$$u(x, t) = \begin{cases} u^- & \text{if } x < \lambda t \,, \\ u^+ & \text{if } x > \lambda t \,. \end{cases}$$

Then one can show that the discontinuous function u is a weak solution of the hyperbolic system (8) if and only if it satisfies the Rankine–Hugoniot equations

$$\lambda(u^+ - u^-) = f(u^+) - f(u^-)\,. \tag{10}$$

More generally, consider a function $u = u(x,t)$ which is piecewise smooth in the x–t-plane. Assume that these discontinuities are located along finitely many curves $x = \gamma_\alpha(t)$, $\alpha = 1,\ldots,N$, and consider the left and right limits

$$u_\alpha^-(t) = \lim_{x\to\gamma_\alpha(t)-} u(x,t), \qquad u_\alpha^+(t) = \lim_{x\to\gamma_\alpha(t)+} u(x,t).$$

Then u is a weak solution of the system of conservation laws if and only if it satisfies the quasilinear system (9) together with the Rankine–Hugoniot equations

$$\gamma_\alpha'(t)\big(u_\alpha^+(t) - u_\alpha^-(t)\big) = f\big(u_\alpha^+(t)\big) - f\big(u_\alpha^-(t)\big)$$

along each shock curve. Here $\gamma_\alpha' = \mathrm{d}\gamma_\alpha/\mathrm{d}t$.

Given an initial condition $u(x,0) = \bar{u}(x)$ containing jumps, however, it is well known that the Rankine–Hugoniot conditions do not determine a unique weak solution. Several "admissibility conditions" have thus been proposed in the literature, in order to single out a unique physically relevant solution. A basic criterion relies on the concept of entropy: A continuously differentiable function $\eta\colon \mathbb{R}^n \mapsto \mathbb{R}$ is called an *entropy* for the system of conservation laws (8), with *entropy flux* $q\colon \mathbb{R}^n \mapsto \mathbb{R}$ if

$$D\eta(u)\cdot Df(u) = Dq(u)\,.$$

If $u = (u_1,\ldots,u_n)$ is a smooth solution of (8), not only the quantities $u_1,\ldots,u_n$ are conserved, but the additional conservation law $\eta(u)_t + q(u)_x = 0$ holds as well. Indeed,

$$D\eta(u)u_t + Dq(u)u_x = D\eta(u)[-Df(u)u_x] + Dq(u)u_x = 0.$$

On the other hand, if the solution u is not smooth but contains shocks, the quantity $\eta = \eta(u)$ may no longer be conserved. The admissibility of a shock can now be characterized by requiring that certain entropies be increasing (or decreasing) in time. More precisely, a weak solution u of (8) is said to be *entropy-admissible* if the inequality

$$\eta(u)_t + q(u)_x \le 0$$

holds in the sense of distributions, for every pair (η, q), where η is a convex entropy and q is the corresponding flux. Calling u^-, u^+ the states to the left and right of the shock, and λ its speed, the above condition implies

$$\lambda[\eta(u^+) - \eta(u^-)] \ge q(u^+) - q(u^-)\,.$$

Various alternative conditions have been studied in the literature, in order to characterize the physically admissible shocks. For these we refer to Lax [39], or Liu [46].

The Riemann Problem

Toward the construction of general solutions for the system of conservation laws (8), a basic building block is the so-called *Riemann problem* [55]. This amounts to choosing a piecewise constant initial data with a single jump at the origin:

$$u(x,0) = \begin{cases} u^- & \text{if } x < 0\,, \\ u^+ & \text{if } x > 0\,. \end{cases} \tag{11}$$

In the special case where the system is linear, i. e. $f(u) = Au$, the solution is piecewise constant in the x–t-plane. It contains $n+1$ constant states $u^- = \omega_0, \omega_1, \ldots, \omega_n = u^+$ (see Fig. 4, left). Each jump $\omega_i - \omega_{i-1}$ is an eigenvector of the matrix A, and is located along the line $x = \lambda_i t$, whose speed equals the corresponding eigenvalue λ_i.

For nonlinear hyperbolic systems of n conservation laws, assuming that the amplitude $|u^+ - u^-|$ of the jump

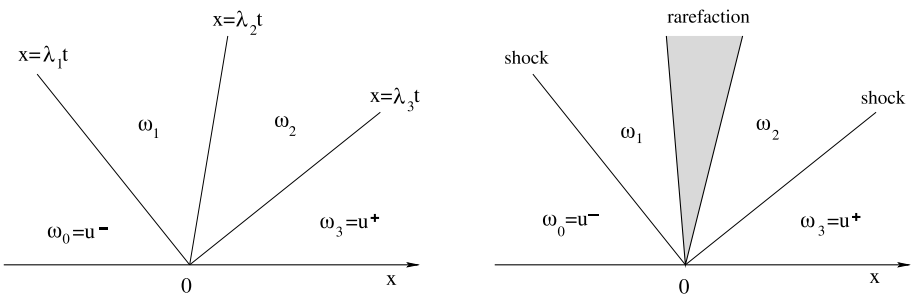

Hyperbolic Conservation Laws, Figure 4
Solutions of a Riemann problem. *Left*: the linear case. *Right*: a nonlinear example

is sufficiently small, the general solution was constructed in a classical paper of Lax [38], under the additional hypothesis

(H) For each $i = 1, \ldots, n$, the ith field is either *genuinely nonlinear*, so that $D\lambda_i(u) \cdot r_i(u) > 0$ for all u, or *linearly degenerate*, with $D\lambda_i(u) \cdot r_i(u) = 0$ for all u.

The solution is self-similar: $u(x, t) = U(x/t)$. It still consists of $n + 1$ constant states $\omega_0 = u^-, \omega_1, \ldots, \omega_n = u^+$ (see Fig. 4, right). Each couple of adjacent states ω_{i-1}, ω_i is separated either by a *shock* satisfying the Rankine Hugoniot equations, or else by a *centered rarefaction*. In this second case, the solution u varies continuously between ω_{i-1} and ω_i in a sector of the t–x-plane where the gradient u_x coincides with an i-eigenvector of the matrix $A(u)$.

Further extensions, removing the technical assumption (H), were obtained by T. P. Liu [46] and by S. Bianchini [3].

Global Solutions

Approximate solutions to a more general Cauchy problem can be constructed by patching together several solutions of Riemann problems. In the Glimm scheme [24], one works with a fixed grid in the x–t plane, with mesh sizes Δx, Δt. At time $t = 0$ the initial data is approximated by a piecewise constant function, with jumps at grid points (see Fig. 5, left). Solving the corresponding Riemann problems, a solution is constructed up to a time Δt sufficiently small so that waves generated by different Riemann problems do not interact. By a random sampling procedure, the solution $u(\Delta t, \cdot)$ is then approximated by a piecewise constant function having jumps only at grid points. Solving the new Riemann problems at every one of these points, one can prolong the solution to the next time interval $[\Delta t, 2\Delta t]$, etc...

An alternative technique for constructing approximate solutions is by wave-front tracking (Fig. 5, right). This method was introduced by Dafermos [17] in the scalar case and later developed by various authors [7,19,29]. It now provides an efficient tool in the study of general $n \times n$ systems of conservation laws, both for theoretical and numerical purposes.

The initial data is here approximated with a piecewise constant function, and each Riemann problem is solved approximately, within the class of piecewise constant functions. In particular, if the exact solution contains a centered rarefaction, this must be approximated by a *rarefaction fan*, containing several small jumps. At the first time t_1 where two fronts interact, the new Riemann problem is again approximately solved by a piecewise constant function. The solution is then prolonged up to the second interaction time t_2, where the new Riemann problem is solved, etc...

The main difference is that with the Glimm scheme one specifies a priori the nodal points where the the Riemann problems are to be solved. On the other hand, in a solution constructed by wave-front tracking the locations of the jumps and of the interaction points depend on the solution itself. Moreover, no restarting procedure is needed.

In the end, both algorithms produce a sequence of approximate solutions, whose total variation remains uniformly bounded. We recall here that the total variation of a function $u : \mathbb{R} \mapsto \mathbb{R}^n$ is defined as

$$\text{Tot.Var.}\,\{u\} \doteq \sup \sum_{i=1}^{N} |u(x_i) - u(x_{i-1})| \,,$$

where the supremum is taken over all $N \geq 1$ and all N-tuples of points $x_0 < x_1 < \cdots < x_N$. For functions of several variables, a more general definition can be found in [22]. Relying on a compactness argument, one can then show that these approximations converge to a weak solution to the system of conservation laws. Namely, one has:

Theorem 1 *Let the system of conservation laws (8) be strictly hyperbolic. Then, for every initial data $\bar{u}$ with sufficiently small total variation, the initial value problem*

$$u_t + f(u)_x = 0 \qquad u(x, 0) = \bar{u}(x) \tag{12}$$

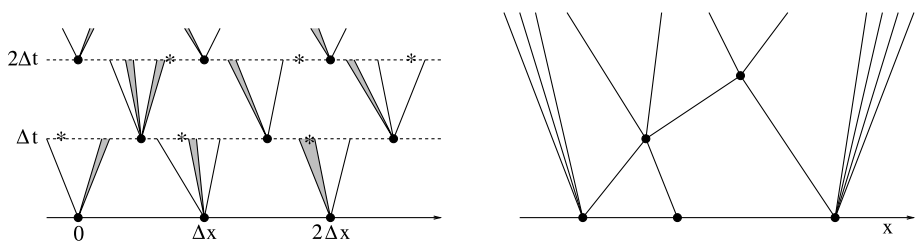

Hyperbolic Conservation Laws, Figure 5
***Left*: the Glimm scheme. *Right*: a front tracking approximation**

has a unique entropy-admissible weak solution, defined for all times $t \geq 0$.

The existence part was first proved in the famous paper of Glimm [24], under the additional hypothesis (H), later removed by Liu [46]. The uniqueness of the solution was proved more recently, in a series of papers by the present author and collaborators, assuming that all shocks satisfy suitable admissibility conditions [8,10]. All proofs are based on careful analysis of solutions of the Riemann problem and on the use of a quadratic interaction functional to control the formation of new waves. These techniques also provided the basis for further investigations of Glimm and Lax [25] and Liu [45] on the asymptotic behavior of weak solutions as $t \to \infty$.

It is also interesting to compare solutions with different initial data. In this direction, we observe that a function of two variables $u(x,t)$ can be regarded as a map $t \mapsto u(\cdot, t)$ from a time interval $[0, T]$ into a space $\mathbf{L}^1(\mathbb{R})$ of integrable functions. Always assuming that the total variation remains small, the distance between two solutions u, v at any time $t > 0$ can be estimated as

$$\|u(t) - v(t)\|_{\mathbf{L}^1} \leq L \, \|u(0) - v(0)\|_{\mathbf{L}^1} \,,$$

where L is a constant independent of time.

Estimates on the rate of convergence of Glimm approximations to the unique exact solutions are available. For every fixed time $T \geq 0$, letting the grid size $\Delta x, \Delta x$ tend to zero keeping the ratio $\Delta t/\Delta x$ constant, one has the error estimate [11]

$$\lim_{\Delta x \to 0} \frac{\left\| u^{\text{Glimm}}(T,\cdot) - u^{\text{exact}}(T,\cdot) \right\|_{\mathbf{L}^1}}{\sqrt{\Delta x} \cdot |\ln \Delta x|} = 0 \,.$$

An alternative approximation procedure involves the addition of a small viscosity. For $\varepsilon > 0$ small, one considers the viscous initial value problem

$$u^\varepsilon_t + f(u^\varepsilon)_x = \varepsilon \, u^\varepsilon_{xx} \,, \qquad u^\varepsilon(x,0) = \bar{u}(x) \,. \tag{13}$$

For initial data $\bar{u}$ with small total variation, the analysis in [6] has shown that the solutions u^ε have small total variation for all times $t > 0$, and converge to the unique weak solution of (12) as $\varepsilon \to 0$.

Hyperbolic Systems in Several Space Dimensions

In several space dimensions there is still no comprehensive theory for systems of conservation laws. Much of the literature has been concerned with three main topics: (i) Global solutions to a single conservation law. (ii) Smooth solutions to a hyperbolic system, locally in time. (iii) Particular solutions to initial or initial-boundary value problems.

Scalar Conservation Laws

The single conservation law on $\mathbb{R}^m$

$$u_t + \operatorname{div} \mathbf{f}(u) = 0$$

has been extensively studied. The fundamental works of Volpert [59] and Kruzhkov [36] have established the global existence of a unique, entropy-admissible solution to the initial value problem, for any initial data $u(x,0) = \bar{u}(x)$ measurable and globally bounded. This solution can be obtained as the unique limit of vanishing viscosity approximations, solving

$$u^\varepsilon_t + \operatorname{div} \mathbf{f}(u^\varepsilon) = \varepsilon \Delta u^\varepsilon \,, \qquad u^\varepsilon(x,0) = \bar{u} \,.$$

As in the one-dimensional case, solutions which are initially smooth may develop shocks and become discontinuous in finite time. Given any two solutions u, v, the following key properties remain valid also in the presence of shocks:

(i) If at the initial time $t = 0$ one has $u(x,0) \leq v(x,0)$ for all $x \in \mathbb{R}^m$, then $u(x,t) \leq v(x,t)$ for all x and all $t \geq 0$.
(ii) The $\mathbf{L}^1$ distance between any two solutions does not increase in time. Namely, for any $0 \leq s \leq t$ one has

$$\|u(t) - v(t)\|_{\mathbf{L}^1(\mathbb{R}^m)} \leq \|u(s) - v(s)\|_{\mathbf{L}^1(\mathbb{R}^m)} \,.$$

Alternative approaches to the analysis of scalar conservation laws were developed by Crandall [16] using nonlinear semigroup theory, and by Lions, Perthame and Tadmor [44] using a kinetic formulation. Regularity results can be found in [30].

Smooth Solutions to Hyperbolic Systems

Using the chain rule, one can rewrite the system of conservation laws (2) in the quasi-linear form

$$u_t + \sum_{\alpha=1}^{m} A^\alpha(u) u_{x_\alpha} = 0 \,. \tag{14}$$

Various definitions of hyperbolicity can be found in the literature. Motivated by several examples from mathematical physics, the system (14) is said to be *symmetrizable hyperbolic* if there exists a positive definite symmetric matrix $S = S(u)$ such that all matrices $S^\alpha(u) = SA^\alpha$ are symmetric. In particular, this condition implies that each $n \times n$ matrix $A^\alpha(u)$ has real eigenvalues and admits a basis of linearly independent eigenvectors. As shown in [23], if a system of conservation laws admits a strictly convex entropy $\eta(u)$, such that the Hessian matrix of second derivatives

$D_u^2\eta(u)$ is positive definite at every point u, then the system is symmetrizable.

A classical theorem states that, for a symmetrizable hyperbolic system with smooth initial data, the initial value problem has a unique smooth solution, locally in time. This solution can be prolonged in time up to the first time where the spatial gradient becomes unbounded at one or more points. In this general setting, however, it is not known whether the solution can be extended beyond this time of shock formation.

Special Solutions

In two space dimensions, one can study special solutions which are independent of time, so that $u(x_1, x_2, t) = U(x_1, x_2)$. In certain cases, one can regard one of the variables, say x_1 as a new time and derive a one-dimensional hyperbolic system of equations for U involving the remaining one-dimensional space variable x_2.

Another important class of solutions relates to two-dimensional Riemann problems. Here the initial data, assigned on the x_1-x_2 plane, is assumed to be constant along rays through the origin. Taking advantage of this self-similarity, the solution can be written in the form $u(x_1, x_2, t) = U(x_1/t, x_2/t)$. This again reduces the problem to an equation in two independent variables [35]. Even for the equation of gas dynamics, a complete solution to the Riemann problem is not available. Several particular cases are analyzed in [42,61].

Several other examples, in specific geometries have been analyzed. A famous problem is the reflection of a shock hitting a wedge-shaped rigid obstacle [15,51].

Numerical Methods

Generally speaking, there are three major classes of numerical methods suitable for partial differential equations: finite difference methods (FDM), finite volume methods (FVM) and finite element methods (FEM). For conservation laws, one also has semi-discrete methods, such as the method of lines, and conservative front tracking methods. The presence of shocks and the rich structure of shock interactions cause the main difficulties in numerical computations.

To illustrate the main idea, consider a uniform grid in the x–t-plane, with step sizes Δx and Δt. Consider the times $t_n = n\Delta t$ and let $I_i = [x_{i-1/2}, x_{i+1/2}]$ be a cell. We wish to compute an approximate value for the cell averages $\bar{u}_i$ over I_i. Integrating the conservation law over the rectangle $I_i \times [t_n, t_{n+1}]$ and dividing by Δx, one obtains

$$\bar{u}_i^{n+1} = \bar{u}_i^n + \frac{\Delta t}{\Delta x}[F_{i-1/2} - F_{i+1/2}] ,$$

where $F_{i+1/2}$ is the average flux

$$F_{i+1/2} = \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(u(x_{i+1/2}))\mathrm{d}t .$$

FVM methods seek a suitable approximation to this average flux $F_{i+1/2}$. First order methods, based on piecewise constant approximations, are usually stable, but contain large numerical diffusion which smears out the shock profile. High order methods are achieved by using polynomials of higher degree, but this produces numerical oscillations around the shock. The basic problem is how to accurately capture the approximated solution near shocks, and at the same time retain stability of the numerical scheme. A common technique is to use a high order scheme on regions where the solution is smooth, and switch to a lower order method near a discontinuity. Well-known methods of this type include the Godunov methods and the MUSCL schemes, wave propagation methods [40,41], the central difference schemes [52,58] and the ENO/WENO schemes [56].

The conservative front tracking methods combine the FDM/FVM with the standard front tracking [26,27]. Based on a high order FDM/FVM, the methods in addition track the location and the strength of the discontinuities, and treat them as moving boundaries. The complexity increases with the number of fronts.

In the FEM setting, the discontinuous Galerkin's methods are widely used [13,14]. The method uses finite element discretization in space, with piecewise polynomials approximation, but allows the approximation to be discontinuous at cell boundaries.

Some numerical methods can be directly extended to the multi-dimensional case, but others need to use a dimensional splitting technique, which introduces additional diffusion. The performance of numerical algorithms is usually tested with some benchmark problem, and little is known theoretically, apart from the case of a scalar conservation law. Moreover, it remains a challenging problem to construct efficient high order numerical methods for systems of conservation laws, both in one and in several space dimensions.

Future Directions

In spite of extensive research efforts, the mathematical theory of hyperbolic conservation laws is still largely incomplete.

For hyperbolic systems in one space dimension, a major challenge is to study the existence and uniqueness of solutions to problems with large initial data. In this direction, some counterexamples show that, for particular sys-

tems, solutions can become unbounded in finite time [31]. However, it is conjectured that for many physical systems, endowed with a strictly convex entropy, such pathological behavior should not occur. In particular, the so-called "p-system" describing isentropic gas dynamics (4) should have global solutions with bounded variation, for arbitrarily large initial data [60].

It is worth mentioning that, for large initial data, the global existence of solutions is known mainly in the scalar case [36,59]. For hyperbolic systems of two conservation laws, global existence can still be proved, relying on a compensated compactness argument [20]. This approach, however, does not provide information on the uniqueness of solutions, or on their continuous dependence on the initial data.

Another major open problem is to theoretically analyze the convergence of numerical approximations. Error bounds on discrete approximations are presently available only in the scalar case [37]. For solutions to hyperbolic systems of n conservation laws, proofs of the convergence of viscous approximations [6], semidiscrete schemes [4], or relaxation schemes [5] have always relied on a priori bounds on the total variation. On the other hand, the counterexample in [2] shows that in general one cannot have any a priori bounds on the total variation of approximate solutions constructed by fully discrete numerical schemes. Understanding the convergence of these discrete approximations will likely require a new approach.

At present, the most outstanding theoretical open problem is to develop a fundamental existence and uniqueness theory for hyperbolic systems in several space dimensions. In order to achieve an existence proof, a key step is to identify the appropriate functional space where to construct solutions. In the one-dimensional case, solutions are found in the space BV of functions with bounded variation. In several space dimensions, however, it is known that the total variation of an arbitrary small solution can become unbounded almost immediately [54]. Hence the space BV does not provide a suitable framework to study the problem. For a special class of systems, a positive result and a counterexample, concerning global existence and continuous dependence on initial data can be found in [1] and in [9], respectively.

Bibliography

Primary Literature

1. Ambrosio L, Bouchut F, De Lellis C (2004) Well-posedness for a class of hyperbolic systems of conservation laws in several space dimensions. Comm Part Diff Equat 29:1635–1651
2. Baiti P, Bressan A, Jenssen HK (2006) An instability of the Godunov scheme. Comm Pure Appl Math 59:1604–1638
3. Bianchini S (2003) On the Riemann problem for non-conservative hyperbolic systems. Arch Rat Mach Anal 166:1–26
4. Bianchini S (2003) BV solutions of the semidiscrete upwind scheme. Arch Ration Mech Anal 167:1–81
5. Bianchini S (2006) Hyperbolic limit of the Jin-Xin relaxation model. Comm Pure Appl Math 59:688–753
6. Bianchini S, Bressan A (2005) Vanishing viscosity solutions to nonlinear hyperbolic systems. Ann Math 161:223–342
7. Bressan A (1992) Global solutions to systems of conservation laws by wave-front tracking. J Math Anal Appl 170:414–432
8. Bressan A (2000) Hyperbolic Systems of Conservation Laws. The One Dimensional Cauchy Problem. Oxford University Press, Oxford
9. Bressan A (2003) An ill posed Cauchy problem for a hyperbolic system in two space dimensions. Rend Sem Mat Univ Padova 110:103–117
10. Bressan A, Liu TP, Yang T (1999) L^1 stability estimates for $n \times n$ conservation laws. Arch Ration Mech Anal 149:1–22
11. Bressan A, Marson A (1998) Error bounds for a deterministic version of the Glimm scheme. Arch Rat Mech Anal 142: 155–176
12. Chen GQ, Zhang Y, Zhu D (2006) Existence and stability of supersonic Euler flows past Lipschitz wedges. Arch Ration Mech Anal 181:261–310
13. Cockburn B, Shu CW (1998) The local discontinuous Galerkin finite element method for convection diffusion systems. SIAM J Numer Anal 35:2440–2463
14. Cockburn B, Hou S, Shu C-W (1990) The Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: the multidimensional case. Math Comput 54:545–581
15. Courant R, Friedrichs KO (1948) Supersonic Flow and Shock Waves. Wiley Interscience, New York
16. Crandall MG (1972) The semigroup approach to first-order quasilinear equations in several space variables. Israel J Math 12:108–132
17. Dafermos C (1972) Polygonal approximations of solutions of the initial value problem for a conservation law. J Math Anal Appl 38:33–41
18. Dafermos C (2005) Hyperbolic Conservation Laws in Continuum Physics, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 2nd edn. Springer, Berlin, pp 325–626
19. DiPerna RJ (1976) Global existence of solutions to nonlinear hyperbolic systems of conservation laws. J Differ Equ 20: 187–212
20. DiPerna R (1983) Convergence of approximate solutions to conservation laws. Arch Ration Mech Anal 82:27–70
21. Euler L (1755) Principes généraux du mouvement des fluides. Mém Acad Sci Berlin 11:274–315
22. Evans LC, Gariepy RF (1992) Measure Theory and Fine Properties of Functions. C.R.C. Press, Boca Raton, pp viii–268
23. Friedrichs KO, Lax P (1971) Systems of conservation laws with a convex extension. Proc Nat Acad Sci USA 68:1686–1688
24. Glimm J (1965) Solutions in the large for nonlinear hyperbolic systems of equations. Comm Pure Appl Math 18:697–715
25. Glimm J, Lax P (1970) Decay of solutions of systems of nonlinear hyperbolic conservation laws. Am Math Soc Memoir 101:xvii–112

26. Glimm J, Li X, Liu Y (2001) Conservative Front Tracking in One Space Dimension. In: Fluid flow and transport in porous media: mathematical and numerical treatment, (South Hadley, 2001), pp 253–264. Contemp Math 295, Amer Math Soc, Providence
27. Glimm J, Grove JW, Li XL, Shyue KM, Zeng Y, Zhang Q (1998) Three dimensional front tracking. SIAM J Sci Comput 19: 703–727
28. Goodman J, Xin Z (1992) Viscous limits for piecewise smooth solutions to systems of conservation laws. Arch Ration Mech Anal 121:235–265
29. Holden H, Risebro NH (2002) Front tracking for hyperbolic conservation laws. Springer, New York
30. Jabin P, Perthame B (2002) Regularity in kinetic formulations via averaging lemmas. ESAIM Control Optim Calc Var 8: 761–774
31. Jenssen HK (2000) Blowup for systems of conservation laws. SIAM J Math Anal 31:894–908
32. Jiang G-S, Levy D, Lin C-T, Osher S, Tadmor E (1998) High-resolution non-oscillatory central schemes with non-staggered grids for hyperbolic conservation laws. SIAM J Numer Anal 35:2147–2168
33. Jin S, Xin ZP (1995) The relaxation schemes for systems of conservation laws in arbitrary space dimensions. Comm Pure Appl Math 48:235–276
34. Kawashima S, Matsumura A (1994) Stability of shock profiles in viscoelasticity with non-convex constitutive relations. Comm Pure Appl Math 47:1547–1569
35. Keyfitz B (2004) Self-similar solutions of two-dimensional conservation laws. J Hyperbolic Differ Equ 1:445–492
36. Kruzhkov S (1970) First-order quasilinear equations with several space variables. Math USSR Sb 10:217–273
37. Kuznetsov NN (1976) Accuracy of some approximate methods for computing the weak solution of a first order quasilinear equation. USSR Comp Math Math Phys 16:105–119
38. Lax P (1957) Hyperbolic systems of conservation laws II. Comm Pure Appl Math 10:537–566
39. Lax P (1971) Shock waves and entropy. In: Zarantonello E (ed) Contributions to Nonlinear Functional Analysis. Academic Press, New York, pp 603–634
40. Leveque RJ (1990) Numerical methods for conservation laws. Lectures in Mathematics. Birkhäuser, Basel
41. Leveque RJ (2002) Finite volume methods for hyperbolic problems. Cambridge University Press, Cambridge
42. Li J, Zhang T, Yang S (1998) The two dimensional problem in gas dynamics. Pitman, Longman Essex
43. Lighthill MJ, Whitham GB (1955) On kinematic waves. II. A theory of traffic flow on long crowded roads. Proc Roy Soc Lond A 229:317–345
44. Lions PL, Perthame E, Tadmor E (1994) A kinetic formulation of multidimensional scalar conservation laws and related equations. JAMS 7:169–191
45. Liu TP (1977) Linear and nonlinear large-time behavior of solutions of general systems of hyperbolic conservation laws. Comm Pure Appl Math 30:767–796
46. Liu TP (1981) Admissible solutions of hyperbolic conservation laws. Mem Am Math Soc 30(240):iv–78
47. Liu TP (1985) Nonlinear stability of shock waves for viscous conservation laws. Mem Am Math Soc 56(328):v–108
48. Liu TP (1987) Hyperbolic conservation laws with relaxation. Comm Math Pys 108:153–175
49. Majda A (1984) Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables. Springer, New York
50. Metivier G (2001) Stability of multidimensional shocks. Advances in the theory of shock waves. Birkhäuser, Boston, pp 25–103
51. Morawetz CS (1994) Potential theory for regular and Mach reflection of a shock at a wedge. Comm Pure Appl Math 47: 593–624
52. Nessyahu H, Tadmor E (1990) Non-oscillatory central differencing for hyperbolic conservation laws J Comp Phys 87:408–463
53. Perthame B (2002) Kinetic Formulation of Conservation Laws. Oxford Univ. Press, Oxford
54. Rauch J (1986) BV estimates fail for most quasilinear hyperbolic systems in dimensions greater than one. Comm Math Phys 106:481–484
55. Riemann B (1860) Über die Fortpflanzung ebener Luftwellen von endlicher Schwingungsweite. Gött Abh Math Cl 8:43–65
56. Shu CW (1998) Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws. Advanced numerical approximation of nonlinear hyperbolic equations. (Cetraro, 1997), pp 325–432. Lecture Notes in Mathematics, vol 1697. Springer, Berlin
57. Smoller J (1994) Shock Waves and Reaction-Diffusion Equations, 2nd edn. Springer, New York
58. Tadmor E (1998) Approximate solutions of nonlinear conservation laws. In: Advanced Numerical Approximation of Nonlinear Hyperbolic Equations. Lecture Notes in Mathematics, vol 1697. (1997 C.I.M.E. course in Cetraro, Italy) Springer, Berlin, pp 1–149
59. Volpert AI (1967) The spaces BV and quasilinear equations. Math USSR Sb 2:225–267
60. Young R (2003) Isentropic gas dynamics with large data. In: Hyperbolic problems: theory, numerics, applications. Springer, Berlin, pp 929–939
61. Zheng Y (2001) Systems of Conservation Laws. Two-dimensional Riemann problems. Birkhäuser, Boston
62. Zumbrun K (2004) Stability of large-amplitude shock waves of compressible Navier–Stokes equations. With an appendix by Helge Kristian Jenssen and Gregory Lyng. Handbook of Mathematical Fluid Dynamics, vol III. North-Holland, Amsterdam, pp 311–533

Books and Reviews

Benzoni-Gavage S, Serre D (2007) Multidimensional hyperbolic partial differential equations. First-order systems and applications. Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, Oxford

Boillat G (1996) Nonlinear hyperbolic fields and waves. In: Recent Mathematical Methods in Nonlinear Wave Propagation (Montecatini Terme, 1994), pp 1–47. Lecture Notes in Math, vol 1640. Springer, Berlin

Chen GQ, Wang D (2002) The Cauchy problem for the Euler equations for compressible fluids. In: Handbook of Mathematical Fluid Dynamics, vol I. North-Holland, Amsterdam, pp 421–543

Courant R, Hilbert D (1962) Methods of Mathematical Physics, vol II. John Wiley & Sons - Interscience, New York

Garavello M, Piccoli B (2006) Traffic Flow on Networks. American Institute of Mathematical Sciences, Springfield

Godlewski E, Raviart PA (1996) Numerical approximation of hyperbolic systems of conservation laws. Springer, New York

Gurtin ME (1981) An Introduction to Continuum Mechanics. Mathematics in Science and Engineering, 158. Academic Press, New York, pp xi–265
Hörmander L (1997) Lectures on Nonlinear Hyperbolic Differential Equations. Springer, Berlin
Jeffrey A (1976) Quasilinear Hyperbolic Systems and Waves. Research Notes in Mathmatics, vol 5. Pitman Publishing, London, pp vii–230
Kreiss HO, Lorenz J (1989) Initial-boundary value problems and the Navier–Stokes equations. Academic Press, San Diego
Kröner D (1997) Numerical schemes for conservation laws. In: Wiley-Teubner Series Advances in Numerical Mathematics. John Wiley, Chichester
Landau LD, Lifshitz EM (1959) Fluid Mechanics. translated form the Russian by Sykes JB, Reid WH, Course of Theoretical Physics, vol 6. Pergamon press, London, Addison-Wesley, Reading, pp xii–536
Li T-T, Yu W-C (1985) Boundary value problems for quasilinear hyperbolic systems. Duke University Math. Series, vol 5. Durham, pp vii–325
Li T-T (1994) Global classical solutions for quasilinear hyperbolic systems. Wiley, Chichester
Lu Y (2003) Hyperbolic conservation laws and the compensated compactness method. Chapman & Hall/CRC, Boca Raton
Morawetz CS (1981) Lecture Notes on Nonlinear Waves and Shocks. Tata Institute of Fundamental Research, Bombay
Rozhdestvenski BL, Yanenko NN (1978) Systems of quasilinear equations and their applications to gas dynamics. Nauka, Moscow. English translation: American Mathematical Society, Providence
Serre D (2000) Systems of Conservation Laws, I Geometric structures, oscillations, and initial-boundry value problem, II Hyperbolicity, entropies, shock waves. Cambridge University Press, pp xii–263
Whitham GB (1999) Linear and Nonlinear Waves. Wiley-Interscience, New York

Hyperbolic Dynamical Systems

Vitor Araújo[1,2], Marcelo Viana[3]
[1] CMUP, Porto, Portugal
[2] IM-UFRJ, Rio de Janeiro, Brazil
[3] IMPA, Rio de Janeiro, Brazil

Article Outline

Glossary
Definition
Introduction
Linear Systems
Local Theory
Hyperbolic Behavior: Examples
Hyperbolic Sets
Uniformly Hyperbolic Systems
Attractors and Physical Measures
Obstructions to Hyperbolicity
Partial Hyperbolicity
Non-Uniform Hyperbolicity – Linear Theory
Non-Uniformly Hyperbolic Systems
Future Directions
Bibliography

Glossary

Homeomorphism, diffeomorphism A *homeomorphism* is a continuous map $f\colon M \to N$ which is one-to-one and onto, and whose inverse $f^{-1}\colon N \to M$ is also continuous. It may be seen as a global continuous change of coordinates. We call f a *diffeomorphism* if, in addition, both it and its inverse are smooth. When $M = N$, the iterated n-fold composition $f \circ \overset{n}{\cdots} \circ f$ is denoted by f^n. By convention, f^0 is the identity map, and $f^{-n} = (f^n)^{-1} = (f^{-1})^n$ for $n \geq 0$.

Smooth flow A flow $f^t\colon M \to M$ is a family of diffeomorphisms depending in a smooth fashion on a parameter $t \in \mathbb{R}$ and satisfying $f^{s+t} = f^s \circ f^t$ for all s, $t \in \mathbb{R}$. This property implies that f^0 is the identity map. Flows usually arise as solutions of autonomous differential equations: let $t \mapsto \phi^t(v)$ denote the solution of

$$\dot{X} = F(X)\,, \quad X(0) = v\,, \tag{1}$$

and assume solutions are defined for all times; then the family ϕ^t thus defined is a flow (at least as smooth as the vector field F itself). The vector field may be recovered from the flow, through the relation $F(X) = \partial_t \phi^t(X)\,|_{t=0}$.

C^k topology Two maps admitting continuous derivatives are said to be C^1-*close* if they are uniformly close, and so are their derivatives. More generally, given any $k \geq 1$, we say that two maps are C^k-*close* if they admit continuous derivatives up to order k, and their derivatives of order i are uniformly close, for every $i = 0, 1, \ldots, k$. This defines a topology in the space of maps of class C^k.

Foliation A foliation is a partition of a subset of the ambient space into smooth submanifolds, that one calls leaves of the foliation, all with the same dimension and varying continuously from one point to the other. For instance, the trajectories of a vector field F, that is, the solutions of Eq. (1), form a 1-dimensional foliation (the leaves are curves) of the complement of the set of zeros of F. The main examples of foliations in the context of this work are the families of stable and unstable manifolds of hyperbolic sets.

Attractor A subset Λ of the ambient space M is *invariant* under a transformation f if $f^{-1}(\Lambda) = \Lambda$, that is,

a point is in Λ if and only if its image is. Λ is invariant under a flow if it is invariant under f^t for all $t \in \mathbb{R}$. An *attractor* is a compact invariant subset Λ such that the trajectories of all points in a neighborhood U converge to Λ as times goes to infinity, and Λ is *dynamically indecomposable* (or *transitive*): there is some trajectory dense in Λ. Sometimes one asks convergence only for points in some "large" subset of a neighborhood U of Λ, and dynamical indecomposability can also be defined in somewhat different ways. However, the formulations we just gave are fine in the uniformly hyperbolic context.

Limit sets The *ω-limit set* of a trajectory $f^n(x)$, $n \in \mathbb{Z}$ is the set $\omega(x)$ of all accumulation points of the trajectory as time n goes to $+\infty$. The *α-limit set* is defined analogously, with $n \to -\infty$. The corresponding notions for continuous time systems (flows) are defined analogously. The *limit set* $L(f)$ (or $L(f^t)$, in the flow case) is the closure of the union of all $'\omega$-limit and all α-limit sets. The *non-wandering set* $\Omega(f)$ (or $\Omega(f^t)$, in the flow case) is that set of points such that every neighborhood U contains some point whose orbit returns to U in future time (then some point returns to U in past time as well). When the ambient space is compact all these sets are non-empty. Moreover, the limit set is contained in the non-wandering set.

Invariant measure A probability measure μ in the ambient space M is *invariant* under a transformation f if $\mu(f^{-1}(A)) = \mu(A)$ for all measurable subsets A. This means that the "events" $x \in A$ and $f(x) \in A$ have equally probable. We say μ is invariant under a flow if it is invariant under f^t for all t. An invariant probability measure μ is *ergodic* if every invariant set A has either zero or full measure. An equivalently condition is that μ can not be decomposed as a convex combination of invariant probability measures, that is, one can not have $\mu = a\mu_1 + (1-a)\mu_2$ with $0 < a < 1$ and μ_1, μ_2 invariant.

Definition

In general terms, a smooth dynamical system is called hyperbolic if the tangent space over the asymptotic part of the phase space splits into two complementary directions, one which is contracted and the other which is expanded under the action of the system. In the classical, so-called uniformly hyperbolic case, the asymptotic part of the phase space is embodied by the limit set and, most crucially, one requires the expansion and contraction rates to be uniform. Uniformly hyperbolic systems are now fairly well understood. They may exhibit very complex behavior which, nevertheless, admits a very precise description. Moreover, uniform hyperbolicity is the main ingredient for characterizing structural stability of a dynamical system. Over the years the notion of hyperbolicity was broadened (non-uniform hyperbolicity) and relaxed (partial hyperbolicity, dominated splitting) to encompass a much larger class of systems, and has become a paradigm for complex dynamical evolution.

Introduction

The theory of uniformly hyperbolic dynamical systems was initiated in the 1960s (though its roots stretch far back into the 19th century) by S. Smale, his students and collaborators, in the west, and D. Anosov, Ya. Sinai, V. Arnold, in the former Soviet Union. It came to encompass a detailed description of a large class of systems, often with very complex evolution. Moreover, it provided a very precise characterization of structurally stable dynamics, which was one of its original main goals.

The early developments were motivated by the problem of characterizing structural stability of dynamical systems, a notion that had been introduced in the 1930s by A. Andronov and L. Pontryagin. Inspired by the pioneering work of M. Peixoto on circle maps and surface flows, Smale introduced a class of *gradient-like* systems, having a finite number of periodic orbits, which should be structurally stable and, moreover, should constitute the majority (an open and dense subset) of all dynamical systems. Stability and openness were eventually established, in the thesis of J. Palis. However, contemporary results of M. Levinson, based on previous work by M. Cartwright and J. Littlewood, provided examples of open subsets of dynamical systems all of which have an infinite number of periodic orbits.

In order to try and understand such phenomenon, Smale introduced a simple geometric model, the now famous "horseshoe map", for which infinitely many periodic orbits exist in a robust way. Another important example of structurally stable system which is not gradient like was R. Thom's so-called "cat map". The crucial common feature of these models is hyperbolicity: the tangent space at each point splits into two complementary directions such that the derivative contracts one of these directions and expands the other, at uniform rates.

In global terms, a dynamical system is called *uniformly hyperbolic*, or Axiom A, if its limit set has this hyperbolicity property we have just described. The mathematical theory of such systems, which is the main topic of this paper, is now well developed and constitutes a main paradigm for the behavior of "chaotic" systems. In our presentation we

go from local aspects (linear systems, local behavior, specific examples) to the global theory (hyperbolic sets, stability, ergodic theory). In the final sections we discuss several important extensions (strange attractors, partial hyperbolicity, non-uniform hyperbolicity) that have much broadened the scope of the theory.

Linear Systems

Let us start by introducing the phenomenon of hyperbolicity in the simplest possible setting, that of linear transformations and linear flows. Most of what we are going to say applies to both discrete time and continuous time systems in a fairly analogous way, and so at each point we refer to either one setting or the other. In depth presentations can be found in e. g. [6,8].

The general solution of a system of linear ordinary differential equations

$$\dot{X} = AX\,, \quad X(0) = v\,,$$

where A is a constant $n \times n$ real matrix and $v \in \mathbb{R}^n$ is fixed, is given by

$$X(t) = \mathrm{e}^{tA} \cdot v\,, \quad t \in \mathbb{R}\,,$$

where $\mathrm{e}^{tA} = \sum_{n=0}^{\infty} (tA)^n/n!$. The linear flow is called *hyperbolic* if A has no eigenvalues on the imaginary axis. Then the *exponential* matrix e^A has no eigenvalues with norm 1. This property is very important for a number of reasons.

Stable and Unstable Spaces

For one thing it implies that all solutions have well-defined asymptotic behavior: they either converge to zero or diverge to infinity as time t goes to $\pm\infty$. More precisely, let

- E^{s} (*stable subspace*) be the subspace of $\mathbb{R}^n$ spanned by the generalized eigenvector associated to eigenvalues of A with negative real part.
- E^{u} (*unstable subspace*) be the subspace of $\mathbb{R}^n$ spanned by the generalized eigenvector associated to eigenvalues of A with positive real part

Then these subspaces are complementary, meaning that $\mathbb{R}^n = E^{\mathrm{s}} \oplus E^{\mathrm{u}}$, and every solution $\mathrm{e}^{tA} \cdot v$ with $v \notin E^{\mathrm{s}} \cup E^{\mathrm{u}}$ diverges to infinity both in the future and in the past. The solutions with $v \in E^{\mathrm{s}}$ converge to zero as $t \to +\infty$ and go to infinity as $t \to -\infty$, and analogously when $v \in E^{\mathrm{u}}$, reversing the direction of time.

Robustness and Density

Another crucial feature of hyperbolicity is *robustness*: any matrix that is close to a hyperbolic one, in the sense that corresponding coefficients are close, is also hyperbolic. The stable and unstable subspaces need not coincide, of course, but the dimensions remain the same. In addition, hyperbolicity if *dense*: any matrix is close to a hyperbolic one. That is because, up to arbitrarily small modifications of the coefficients, one may force all eigenvalues to move out of the imaginary axis.

Stability, Index of a Fixed Point

In addition to robustness, hyperbolicity also implies *stability*: if B is close to a hyperbolic matrix A, in the sense we have just described, then the solutions of $\dot{X} = BX$ have essentially the same behavior as the solutions of $\dot{X} = AX$. What we mean by "essentially the same behavior" is that there exists a global continuous change of coordinates, that is, a homeomorphism $h\colon \mathbb{R}^n \to \mathbb{R}^n$, that maps solutions of one system to solutions of the other, preserving the time parametrization:

$$h\big(\mathrm{e}^{tA} \cdot v\big) = \mathrm{e}^{tB} \cdot h(v) \quad \text{for all} \quad t \in \mathbb{R}\,.$$

More generally, two hyperbolic linear flows are conjugated by a homeomorphism h if and only if they have the same *index*, that is, the same number of eigenvalues with negative real part. In general, h can not be taken to be a diffeomorphism: this is possible if and only if the two matrices A and B are obtained from one another via a change of basis. Notice that in this case they must have the same eigenvalues, with the same multiplicities.

Hyperbolic Linear Systems

There is a corresponding notion of hyperbolicity for discrete time linear systems

$$X_{n+1} = CX_n\,, \quad X_0 = v\,,$$

with C a $n \times n$ real matrix. Namely, we say the system is *hyperbolic* if C has no eigenvalue in the unit circle. Thus a matrix A is hyperbolic in the sense of continuous time systems if and only if its exponential $C = \mathrm{e}^A$ is hyperbolic in the sense of discrete time systems. The previous observations (well-defined behavior, robustness, denseness and stability) remain true in discrete time. Two hyperbolic matrices are conjugate by a homeomorphism if and only if they have the same index, that is, the same number of eigenvalues with norm less than 1, and they both either preserve or reverse orientation.

Local Theory

Now we move on to discuss the behavior of non-linear systems close to fixed or, more generally, periodic trajectories. By non-linear system we understand the iteration of a diffeomorphism f, or the evolution of a smooth flow f^t, on some manifold M. The general philosophy is that the behavior of the system close to a hyperbolic fixed point very much resembles the dynamics of its linear part.

A fixed point $p \in M$ of a diffeomorphism $f\colon M \to M$ is called *hyperbolic* if the linear part $Df_p\colon T_pM \to T_pM$ is a hyperbolic linear map, that is, if Df_p has no eigenvalue with norm 1. Similarly, an equilibrium point p of a smooth vector field F is *hyperbolic* if the derivative $DF(p)$ has no pure imaginary eigenvalues.

Hartman–Grobman Theorem

This theorem asserts that if p is a hyperbolic fixed point of $f\colon M \to M$ then there are neighborhoods U of p in M and V of 0 in the tangent space T_pM such that we can find a homeomorphism $h\colon U \to V$ such that

$$h \circ f = Df_p \circ h\,,$$

whenever the composition is defined. This property means that h maps orbits of $Df(p)$ close to zero to orbits of f close to p. We say that h is a (local) *conjugacy* between the non-linear system f and its linear part Df_p. There is a corresponding similar theorem for flows near a hyperbolic equilibrium. In either case, in general h can not be taken to be a diffeomorphism.

Stable Sets

The *stable set* of the hyperbolic fixed point p is defined by

$$W^s(p) = \left\{x \in M\colon d(f^n(x), f^n(p)) \xrightarrow[n\to+\infty]{} 0\right\}\,.$$

Given $\beta > 0$ we also consider the *local stable set* of size $\beta > 0$, defined by

$$W^s_\beta(p) = \{x \in M\colon d(f^n(x), f^n(p)) \le \beta \quad \text{for all} \quad n \ge 0\}.$$

The image of W^s_β under the conjugacy h is a neighborhood of the origin inside E^s. It follows that the local stable set is an embedded topological disk, with the same dimension as E^s. Moreover, the orbits of the points in $W^s_\beta(p)$ actually converges to the fixed point as time goes to infinity. Therefore,

$$z \in W^s(p) \quad \Leftrightarrow \quad f^n(z) \in W^s_\beta(p) \quad \text{for some} \quad n \ge 0\,.$$

Stable Manifold Theorem

The stable manifold theorem asserts that $W^s_\beta(p)$ is actually a smooth embedded disk, with the same order of differentiability as f itself, and it is tangent to E^s at the point p. It follows that $W^s(p)$ is a smooth submanifold, injectively immersed in M. In general, $W^s(p)$ is not embedded in M: in many cases it has self-accumulation points. For these reasons one also refers to $W^s(p)$ and $W^s_\beta(p)$ as stable *manifolds* of p. Unstable manifolds are defined analogously, replacing the transformation by its inverse.

Local Stability

We call *index* of a diffeomorphism f at a hyperbolic fixed point p the index of the linear part, that is, the number of eigenvalues of Df_p with negative real part. By the Hartman–Grobman theorem and previous comments on linear systems, two diffeomorphisms are locally conjugate near hyperbolic fixed points if and only if the stable indices and they both preserve/reverse orientation. In other words, the index together with the sign of the Jacobian determinant form a complete set of invariants for local topological conjugacy.

Let g be any diffeomorphism C^1-close to f. Then g has a unique fixed point p_g close to p, and this fixed point is still hyperbolic. Moreover, the stable indices and the orientations of the two diffeomorphisms at the corresponding fixed points coincide, and so they are locally conjugate. This is called *local stability* near of diffeomorphisms hyperbolic fixed points. The same kind of result holds for flows near hyperbolic equilibria.

Hyperbolic Behavior: Examples

Now let us review some key examples of (semi)global hyperbolic dynamics. Thorough descriptions are available in e. g. [6,8,9].

A Linear Torus Automorphism

Consider the linear transformation $A\colon \mathbb{R}^2 \to \mathbb{R}^2$ given by the following matrix, relative to the canonical base of the plane:

$$\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\,.$$

The 2-dimensional torus $\mathbb{T}^2$ is the quotient $\mathbb{R}^2/\mathbb{Z}^2$ of the plane by the equivalence relation

$$(X_1, y_1) \sim (x_2, y_2) \quad \Leftrightarrow \quad (x_1 - x_2, y_1 - y_2) \in \mathbb{Z}^2\,.$$

Since A preserves the lattice $\mathbb{Z}^2$ of integer vectors, that is, since $A(\mathbb{Z}^2) = \mathbb{Z}^2$, the linear transformation defines an invertible map $f_A\colon \mathbb{T}^2 \to \mathbb{T}^2$ in the quotient space, which is an example of linear automorphism of $\mathbb{T}^2$. We call affine line in $\mathbb{T}^2$ the projection under the quotient map of any affine line in the plane.

The linear transformation A is hyperbolic, with eigenvalues $0 < \lambda_1 < 1 < \lambda_2$, and the corresponding eigenspaces E^1 and E^2 have irrational slope. For each point $z \in \mathbb{T}^2$, let $W_i(z)$ denote the affine line through z and having the direction of E^i, for $i = 1, 2$:

- distances along $W_1(z)$ are multiplied by $\lambda_1 < 1$ under forward iteration of f_A
- distances along $W_2(z)$ are multiplied by $1/\lambda_2 < 1$ under backward iteration of f_A.

Thus we call $W_1(z)$ *stable manifold* and $W_2(z)$ *unstable manifold* of z (notice we are not assuming z to be periodic). Since the slopes are irrational, stable and unstable manifolds are dense in the whole torus. From this fact one can deduce that the periodic points of f_A form a dense subset of the torus, and that there exist points whose trajectories are dense in T^2. The latter property is called *transitivity*.

An important feature of this systems is that its behavior is (globally) stable under small perturbations: given any diffeomorphism $g\colon \mathbb{T}^2 \to \mathbb{T}^2$ sufficiently C^1-close to f_A, there exists a homeomorphism $h\colon \mathbb{T}^2 \to \mathbb{T}^2$ such that $h \circ g = f_A \circ h$. In particular, g is also transitive and its periodic points form a dense subset of $\mathbb{T}^2$.

The Smale Horseshoe

Consider a stadium shaped region D in the plane divided into three subregions, as depicted in Fig. 1: two half disks, A and C, and a square, B. Next, consider a map $f\colon D \to D$ mapping D back inside itself as described in Fig. 1: the intersection between B and $f(B)$ consists of two rectangles, R_0 and R_1, and f is affine on the pre-image of these rectangles, contracting the horizontal direction and expanding the vertical direction.

The set $\Lambda = \cap_{n\in\mathbb{Z}} f^n(B)$, formed by all the points whose orbits never leave the square B is totally disconnected, in fact, it is the product of two Cantor sets. A description of the dynamics on Λ may be obtained through the following coding of orbits. For each point $z \in \Lambda$ and every time $n \in \mathbb{Z}$ the iterate $f^n(z)$ must belong to either R_0 or R_1. We call *itinerary* of z the sequence $\{s_n\}_{n\in\mathbb{Z}}$ with values in the set $\{0, 1\}$ defined by $f^n(z) \in R_{s_n}$ for all $n \in \mathbb{Z}$. The itinerary map

$$\Lambda \to \{0,1\}^{\mathbb{Z}}\,, \quad z \mapsto \{s_n\}_{n\in\mathbb{Z}}$$

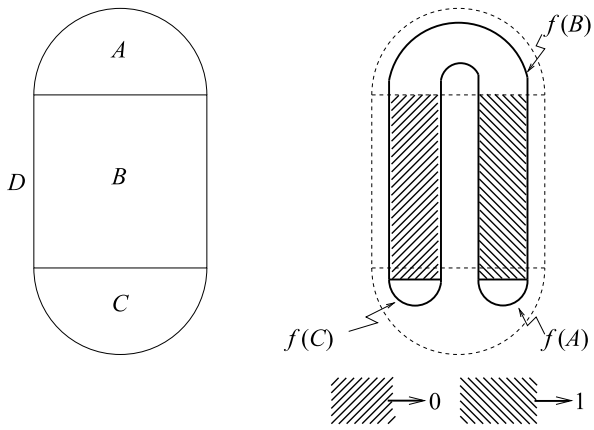

Hyperbolic Dynamical Systems, Figure 1
Horseshoe map

is a homeomorphism, and conjugates f restricted to Λ to the so-called *shift map* defined on the space of sequences by

$$\{0,1\}^{\mathbb{Z}} \to \{0,1\}^{\mathbb{Z}}\,, \quad \{s_n\}_{n\in\mathbb{Z}} \mapsto \{s_{n+1}\}_{n\in\mathbb{Z}}\,.$$

Since the shift map is transitive, and its periodic points form a dense subset of the domain, it follows that the same is true for the horseshoe map on Λ.

From the definition of f we get that distances along horizontal line segments through points of Λ are contracted at a uniform rate under forward iteration and, dually, distances along vertical line segments through points of Λ are contracted at a uniform rate under backward iteration. Thus, horizontal line segments are local stable sets and vertical line segments are local unstable sets for the points of Λ.

A striking feature of this system is the stability of its dynamics: given any diffeomorphism g sufficiently C^1-close to f, its restriction to the set $\Lambda_g = \cap_{n\in\mathbb{Z}} g^n(B)$ is conjugate to the restriction of f to the set $\Lambda = \Lambda_f$ (and, consequently, is conjugate to the shift map). In addition, each point of Λ_g has local stable and unstable sets which are smooth curve segments, respectively, approximately horizontal and approximately vertical.

The Solenoid Attractor

The *solid torus* is the product space $\mathbb{S}^1 \times \mathbb{D}$, where $\mathbb{S}^1 = \mathbb{R}/\mathbb{Z}$ is the circle and $\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}$ is the unit disk in the complex plane. Consider the map $f\colon \mathbb{S}^1 \times \mathbb{D} \to \mathbb{S}^1 \times \mathbb{D}$ given by

$$(\theta, z) \mapsto (2\theta, \alpha z + \beta e^{i\theta/2})\,,$$

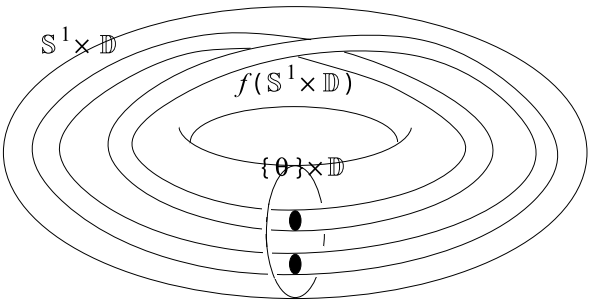

Hyperbolic Dynamical Systems, Figure 2
The solenoid attractor

$\theta \in \mathbb{R}/\mathbb{Z}$ and $\alpha, \beta \in \mathbb{R}$ with $\alpha + \beta < 1$. The latter condition ensures that the image $f(\mathbb{S}^1 \times \mathbb{D})$ is strictly contained in $\mathbb{S}^1 \times \mathbb{D}$. Geometrically, the image is a long thin domain going around the solid torus twice, as described in Fig. 2. Then, for any $n \geq 1$, the corresponding iterate $f^n(\mathbb{S}^1 \times \mathbb{D})$ is an increasingly thinner and longer domain that winds 2^k times around $\mathbb{S}^1 \times \mathbb{D}$. The maximal invariant set

$$\Lambda = \cap_{n \geq 0} f^n(\mathbb{S}^1 \times \mathbb{D})$$

is called *solenoid attractor*. Notice that the forward orbit under f of every point in $\mathbb{S}^1 \times \mathbb{D}$ accumulates on Λ. One can also check that the restriction of f to the attractor is transitive, and the set of periodic points of f is dense in Λ.

In addition Λ has a dense subset of periodic orbits and also a dense orbit. Moreover every point in a neighborhood of Λ converges to Λ and this is why this set is called an *attractor*.

Hyperbolic Sets

The notion we are now going to introduce distillates the crucial feature common to the examples presented previously. A detailed presentation is given in e.g. [8,10]. Let $f: M \to M$ be a diffeomorphism on a manifold M. A compact invariant set $\Lambda \subset M$ is a *hyperbolic set* for f if the tangent bundle over Λ admits a decomposition

$$T_\Lambda M = E^{\mathrm{u}} \oplus E^{\mathrm{s}} ,$$

invariant under the derivative and such that $\|Df^{-1} \mid E^{\mathrm{u}}\| < \lambda$ and $\|Df \mid E^{\mathrm{s}}\| < \lambda$ for some constant $\lambda < 1$ and some choice of a Riemannian metric on the manifold. When it exists, such a decomposition is necessarily unique and continuous. We call E^s the stable bundle and E^u the unstable bundle of f on the set Λ.

The definition of hyperbolicity for an invariant set of a smooth flow containing no equilibria is similar, except that one asks for an invariant decomposition $T_\Lambda M = E^{\mathrm{u}} \oplus E^0 \oplus E^{\mathrm{s}}$, where E^u and E^s are as before and E^0 is a line bundle tangent to the flow lines. An invariant set that contains equilibria is hyperbolic if and only it consists of a finite number of points, all of them hyperbolic equilibria.

Cone Fields

The definition of hyperbolic set is difficult to use in concrete situations, because, in most cases, one does not know the stable and unstable bundles explicitly. Fortunately, to prove that an invariant set is hyperbolic it suffices to have some approximate knowledge of these invariant subbundles. That is the contents of the invariant cone field criterion: a compact invariant set is hyperbolic if and only if there exists some continuous (not necessarily invariant) decomposition $T_\Lambda M = E^1 \oplus E^2$ of the tangent bundle, some constant $\lambda < 1$, and some cone field around E^1

$$C^1_a(x) = \{v = v_1 + v_2 \in E^1_x \oplus E^2_x : \|v_2\| \leq a\|v_1\|\}, \quad x \in \Lambda,$$

which is

(a) forward invariant: $Df_x(C^1_a(x)) \subset C^1_{\lambda a}(f(x))$ and
(b) expanded by forward iteration: $\|Df_x(v)\| \geq \lambda^{-1}\|v\|$ for every $v \in C^1_a(x)$

and there exists a cone field $C^2_b(x)$ around E^2 which is backward invariant and expanded by backward iteration.

Robustness

An easy, yet very important consequence is that hyperbolic sets are robust under small modifications of the dynamics. Indeed, suppose Λ is a hyperbolic set for $f: M \to M$, and let $C^1_a(x)$ and $C^2_b(x)$ be invariant cone fields as above. The (non-invariant) decomposition $E^1 \oplus E^2$ extends continuously to some small neighborhood U of Λ, and then so do the cone fields. By continuity, conditions (a) and (b) above remain valid on U, possibly for a slightly larger constant λ. Most important, they also remain valid when f is replaced by any other diffeomorphism g which is sufficiently C^1-close to it. Thus, using the cone field criterion once more, every compact set $K \subset U$ which is invariant under g is a hyperbolic set for g.

Stable Manifold Theorem

Let Λ be a hyperbolic set for a diffeomorphism $f: M \to M$. Assume f is of class C^k. Then there exist $\varepsilon_0 > 0$ and $0 < \lambda < 1$ and, for each $0 < \varepsilon \leq \varepsilon_0$ and $x \in \Lambda$, the *local stable manifold of size* ε

$$W^s_\varepsilon(x) = \{y \in M : \operatorname{dist}(f^n(y), f^n(x)) \leq \varepsilon \text{ for all } n \geq 0\},$$

and the *local unstable manifold* of size ε

$$W^u_\varepsilon(x) = \{y \in M\colon \operatorname{dist}(f^{-n}(y), f^{-n}(x)) \le \varepsilon \quad \text{for all } n \ge 0\}$$

are C^k embedded disks, tangent at x to E^{s}_x and E^{u}_x, respectively, and satisfying

- $f(W^s_\varepsilon(x)) \subset W^s_\varepsilon(f(x))$ and $f^{-1}(W^u_\varepsilon(x)) \subset W^u_\varepsilon(f^{-1}(x))$;
- $\operatorname{dist}(f(x), f(y)) \le \lambda \operatorname{dist}(x, y)$ for all $y \in W^s_\varepsilon(x)$
- $\operatorname{dist}(f^{-1}(x), f^{-1}(y)) \le \lambda \operatorname{dist}(x, y)$ for all $y \in W^u_\varepsilon(x)$
- $W^s_\varepsilon(x)$ and $W^u_\varepsilon(x)$ vary continuously with the point x, in the C^k topology.

Then, the *global stable and unstable manifolds* of x,

$$W^s(x) = \bigcup_{n\ge 0} f^{-n}\big(W^s_\varepsilon(f^n(x))\big)$$
$$\text{and } W^u(x) = \bigcup_{n\ge 0} f^{n}\big(W^u_\varepsilon(f^{-n}(x))\big)\,,$$

are smoothly immersed submanifolds of M, and they are characterized by

$$W^s(x) = \{y \in M\colon \operatorname{dist}(f^n(y), f^n(x)) \to 0 \text{ as } n \to \infty\}$$
$$W^u(x) = \{y \in M\colon \operatorname{dist}(f^{-n}(y), f^{-n}(x)) \to 0 \text{ as } n \to \infty\}\,.$$

Shadowing Property

This crucial property of hyperbolic sets means that possible small "errors" in the iteration of the map close to the set are, in some sense, unimportant: to the resulting "wrong" trajectory, there corresponds a nearby genuine orbit of the map. Let us give the formal statement. Recall that a hyperbolic set is compact, by definition.

Given $\delta > 0$, a *δ-pseudo-orbit* of $f\colon M \to M$ is a sequence $\{x_n\}_{n\in\mathbb{Z}}$ such that

$$\operatorname{dist}(x_{n+1}, f(x_n)) \le \delta \quad \text{for all } n \in \mathbb{Z}\,.$$

Given $\varepsilon > 0$, one says that a pseudo-orbit is *ε-shadowed* by the orbit of a point $z \in M$ if $\operatorname{dist}(f^n(z), x_n) \le \varepsilon$ for all $n \in \mathbb{Z}$. The *shadowing lemma* says that for any $\varepsilon > 0$ one can find $\delta > 0$ and a neighborhood U of the hyperbolic set Λ such that every δ-pseudo-orbit in U is ε-shadowed by some orbit in U. Assuming ε is sufficiently small, the shadowing orbit is actually unique.

Local Product Structure

In general, these shadowing orbits need not be inside th hyperbolic set Λ. However, that is indeed the case if Λ is a *maximal invariant set*, that is, if it admits some neighborhood U such that Λ coincides with the set of points whose orbits never leave U:

$$\Lambda = \bigcap_{n\in\mathbb{Z}} f^{-n}(U)\,.$$

A hyperbolic set is a maximal invariant set if and only if it has the local product structure property stated in the next paragraph.

Let Λ be a hyperbolic set and ε be small. If x and y are nearby points in Λ then the local stable manifold of x intersects the local unstable manifold of y at a unique point, denoted $[x, y]$, and this intersection is transverse. This is because the local stable manifold and the local unstable manifold of every point are transverse, and these local invariant manifolds vary continuously with the point. We say that Λ has *local product structure* if there exists $\delta > 0$ such that $[x, y]$ belongs to Λ for every $x, y \in \Lambda$ with $\operatorname{dist}(x, y) < \delta$.

Stability

The shadowing property may also be used to prove that hyperbolic sets are stable under small perturbations of the dynamics: if Λ is a hyperbolic set for f then for any C^1-close diffeomorphism g there exists a hyperbolic set Λ_g close to Λ and carrying the same dynamical behavior.

The key observation is that every orbit $f^n(x)$ of f inside Λ is a δ-pseudo-orbits for g in a neighborhood U, where δ is small if g is close to f and, hence, it is shadowed by some orbit $g^n(z)$ of g. The correspondence $h(x) = z$ thus defined is injective and continuous.

For any diffeomorphism g close enough to f, the orbits of x in the maximal g-invariant set $\Lambda_g(U)$ inside U are pseudo-orbits for f. Therefore the shadowing property above enables one to bijectively associate g-orbits of $\Lambda_g(U)$ to f-orbits in Λ. This provides a homeomorphism $h\colon \Lambda_g(U) \to \Lambda$ which conjugates g and f on the respective hyperbolic sets: $f \circ h = h \circ g$. Thus *hyperbolic maximal sets are structurally stable*: the persistent dynamics in a neighborhood of these sets is the same for all nearby maps.

If Λ is a hyperbolic maximal invariant set for f then its hyperbolic continuation for any nearby diffeomorphism g is also a maximal invariant set for g.

Symbolic Dynamics

The dynamics of hyperbolic sets can be described through a symbolic coding obtained from a convenient discretization of the phase space. In a few words, one partitions the

set into a finite number of subsets and assigns to a generic point in the hyperbolic set its itinerary with respect to this partition. Dynamical properties can then be read out from a shift map in the space of (admissible) itineraries. The precise notion involved is that of Markov partition.

A set $R \subset \Lambda$ is a *rectangle* if $[x, y] \in R$ for each $x, y \in R$. A rectangle is *proper* if it is the closure of its interior relative to Λ. A *Markov partition* of a hyperbolic set Λ is a cover $\mathcal{R} = \{R_1, \ldots, R_m\}$ of Λ by proper rectangles with pairwise disjoint interiors, relative to Λ, and such

$$W^u(f(x)) \cap R_j \subset f(W^u(x) \cap R_i) \quad \text{and} \quad f(W^s(x) \cap R_i) \subset W^s(f(x)) \cap R_j$$

for every $x \in \mathrm{int}_\Lambda(R_i)$ with $f(x) \in \mathrm{int}_\Lambda(R_j)$. The key fact is that *any maximal hyperbolic set Λ admits Markov partitions with arbitrarily small diameter*.

Given a Markov partition $\mathcal{R}$ with sufficiently small diameter, and a sequence $\mathbf{j} = (j_n)_{n\in\mathbb{Z}}$ in $\{1, \ldots, m\}$, there exists at most one point $x = h(\mathbf{j})$ such that

$$f^n(x) \in R_{j_n} \quad \text{for each} \quad n \in \mathbb{Z}\,.$$

We say that $\mathbf{j}$ is admissible if such a point x does exist and, in this case, we say x admits $\mathbf{j}$ as an itinerary. It is clear that $f \circ h = h \circ \sigma$, where σ is the shift (left-translation) in the space of admissible itineraries. The map h is continuous and surjective, and it is injective on the residual set of points whose orbits never hit the boundaries (relative to Λ) of the Markov rectangles.

Uniformly Hyperbolic Systems

A diffeomorphism $f\colon M \to M$ is *uniformly hyperbolic*, or satisfies the *Axiom A*, if the non-wandering set $\Omega(f)$ is a hyperbolic set for f and the set $\mathrm{Per}(f)$ of periodic points is dense in $\Omega(f)$. There is an analogous definition for smooth flows $f^t\colon M \to M$, $t \in \mathbb{R}$. The reader can find the technical details in e. g. [6,8,10].

Dynamical Decomposition

The so-called "spectral" decomposition theorem of Smale allows for the global dynamics of a hyperbolic diffeomorphism to be decomposed into elementary building blocks. It asserts that the non-wandering set splits into a finite number of pairwise disjoint *basic pieces* that are compact, invariant, and dynamically indecomposable. More precisely, the non-wandering set $\Omega(f)$ of a uniformly hyperbolic diffeomorphism f is a finite pairwise disjoint union

$$\Omega(f) = \Lambda_1 \cup \cdots \cup \Lambda_N$$

of f-invariant, transitive sets Λ_i, that are compact and maximal invariant sets. Moreover, the α-limit set of every orbit is contained in some Λ_i and so is the ω-limit set.

Geodesic Flows on Surfaces with Negative Curvature

Historically, the first important example of uniform hyperbolicity was the geodesic flow G^t on Riemannian manifolds of negative curvature M. This is defined as follows.

Let M be a compact Riemannian manifold. Given any tangent vector v, let $\gamma_v\colon \mathbb{R} \to TM$ be the geodesic with initial condition $v = \gamma_v(0)$. We denote by $\dot\gamma_v(t)$ the velocity vector at time t. Since $\|\dot\gamma_v(t)\| = \|v\|$ for all t, it is no restriction to consider only unit vectors. There is an important volume form on the unit tangent bundle, given by the product of the volume element on the manifold by the volume element induced on each fiber by the Riemannian metric. By integration of this form, one obtains the *Liouville measure* on the unit tangent bundle, which is a finite measure if the manifold itself has finite volume (including the compact case). The *geodesic flow* is the flow $G^t\colon T^1M \to T^1M$ on the unit tangent bundle T^1M of the manifold, defined by

$$G^t(v) = \dot\gamma_v(t)\,.$$

An important feature is that this flow leaves invariant the Liouville measure. By Poincaré recurrence, this implies that $\Omega(G) = T^1M$.

A major classical result in Dynamics, due to Anosov, states that *if M has negative sectional curvature then this measure is ergodic for the flow*. That is, any invariant set has zero or full Liouville measure. The special case when M is a surface, had been dealt before by Hedlund and Hopf.

The key ingredient to this theorem is to prove that the geodesic flow is uniformly hyperbolic, in the sense we have just described, when the sectional curvature is negative. In the surface case, the stable and unstable invariant subbundles are differentiable, which is no longer the case in general in higher dimensions. This formidable obstacle was overcome by Anosov through showing that the corresponding invariant foliations retain, nevertheless, a weaker form of regularity property, that suffices for the proof. Let us explain this.

Absolute Continuity of Foliations

The invariant spaces E^{s}_x and E^{u}_x of a hyperbolic system depend continuously, and even Hölder continuously, on the base point x. However, in general this dependence is not differentiable, and this fact is at the origin of several important difficulties. Related to this, the families of stable

and unstable manifolds are, usually, not differentiable foliations: although the leaves themselves are as smooth as the dynamical system itself, the holonomy maps often fail to be differentiable. By holonomy maps we mean the projections along the leaves between two given cross-sections to the foliation.

However, Anosov and Sinai observed that if the system is at least twice differentiable then these foliations are *absolutely continuous*: their holonomy maps send zero Lebesgue measure sets of one cross-section to zero Lebesgue measure sets of the other cross-section. This property is crucial for proving that any smooth measure which is invariant under a twice differentiable hyperbolic system is ergodic. For dynamical systems that are only once differentiable the invariant foliations may fail to be absolutely continuous. Ergodicity still is an open problem.

Structural Stability

A dynamical system is *structurally stable* if it is equivalent to any other system in a C^1 neighborhood, meaning that there exists a global homeomorphism sending orbits of one to orbits of the other and preserving the direction of time. More generally, replacing C^1 by C^r neighborhoods, any $r \geq 1$, one obtains the notion of C^r structural stability. Notice that, in principle, this property gets weaker as r increases.

The Stability Conjecture of Palis–Smale proposed a complete geometric characterization of this notion: for any $r \geq 1$, *C^r structurally stable systems should coincide with the hyperbolic systems having the property of strong transversality*, that is, such that the stable and unstable manifolds of any points in the non-wandering set are transversal. In particular, this would imply that the property of C^r structural stability does not really depend on the value of r.

That hyperbolicity and strong transversality suffice for structural stability was proved in the 1970s by Robbin, de Melo, Robinson. It is comparatively easy to prove that strong transversality is also necessary. Thus, the heart of the conjecture is to prove that structurally stable systems must be hyperbolic. This was achieved by Mañé in the 1980s, for C^1 diffeomorphisms, and extended about ten years later by Hayashi for C^1 flows. Thus *a C^1 diffeomorphism, or flow, on a compact manifold is structurally stable if and only if it is uniformly hyperbolic and satisfies the strong transversality condition.*

Ω-stability

A weaker property, called *Ω-stability* is defined requiring equivalence only restricted to the non-wandering set. The Ω-Stability Conjecture of Palis–Smale claims that, for any $r \geq 1$, *Ω-stable systems should coincide with the hyperbolic systems with no cycles*, that is, such that no basic pieces in the spectral decomposition are cyclically related by intersections of the corresponding stable and unstable sets.

The Ω-stability theorem of Smale states that these properties are sufficient for $C^r\Omega$stability. Palis showed that the no-cycles condition is also necessary. Much later, based on Mañé's aforementioned result, he also proved that for C^1 diffeomorphisms hyperbolicity is necessary for Ω-stability. This was extended to C^1 flows by Hayashi in the 1990s.

Attractors and Physical Measures

A hyperbolic basic piece Λ_i is a *hyperbolic attractor* if the stable set

$$W^s(\Lambda_i) = \{x \in M : \omega(x) \subset \Lambda_i\}$$

contains a neighborhood of Λ_i. In this case we call $W^s(\Lambda_i)$ the *basin* of the attractor Λ_i, and denote it $B(\Lambda_i)$. When the uniformly hyperbolic system is of class C^2, a basic piece is an attractor if and only if its stable set has positive Lebesgue measure. Thus, the union of the basins of all attractors is a full Lebesgue measure subset of M. This remains true for a residual (dense G_δ) subset of C^1 uniformly hyperbolic diffeomorphisms and flows.

The following fundamental result, due to Sinai, Ruelle, Bowen shows that, no matter how complicated it may be, the behavior of typical orbits in the basin of a hyperbolic attractor is well-defined at the statistical level: *any hyperbolic attractor Λ of a C^2 diffeomorphism (or flow) supports a unique invariant probability measure μ such that*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \varphi(f^j(z)) = \int \varphi \, \mathrm{d}\mu \tag{2}$$

for every continuous function φ and Lebesgue almost every point $x \in B(\Lambda)$. The standard reference here is [3].

Property (2) also means that the Sinai–Ruelle–Bowen measure μ may be "observed": the weights of subsets may be found with any degree of precision, as the sojourn-time of any orbit picked "at random" in the basin of attraction:

$$\mu(V) = \text{fraction of time the orbit of } z \text{ spends in } V$$

for typical subsets V of M (the boundary of V should have zero μ-measure), and for Lebesgue almost any point $z \in B(\Lambda)$. For this reason μ is called a *physical measure*.

It also follows from the construction of these physical measures on hyperbolic attractors that they depend continuously on the diffeomorphism (or the flow). This *statistical stability* is another sense in which the asymptotic behavior is stable under perturbations of the system, distinct from structural stability.

There is another sense in which this measure is "physical" and that is that μ is the zero-noise limit of the stationary measures associated to the stochastic processes obtained by adding small random noise to the system. The idea is to replace genuine trajectories by "random orbits" $(z_n)_n$, where each $z_n + 1$ is chosen ε-close to $f(z_n)$. We speak of *stochastic stability* if, for any continuous function φ, the random time average

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} \varphi(z_j)$$

is close to $\int \varphi \, \mathrm{d}\mu$ for almost all choices of the random orbit.

One way to construct such random orbits is through randomly perturbed iterations, as follows. Consider a family of probability measures ν_ε in the space of diffeomorphisms, such that each ν_ε is supported in the ε-neighborhood of f. Then, for each initial state z_0 define $z_{n+1} = f_{n+1}(z_n)$, where the diffeomorphisms f_n are independent random variables with distribution law ν_ε. A probability measure η_ε on the basin $B(\Lambda)$ is *stationary* if it satisfies

$$\eta_\varepsilon(E) = \int \eta_\varepsilon(g^{-1}(E)) \, \mathrm{d}\nu_\varepsilon(g) \, .$$

Stationary measures always exist, and they are often unique for each small $\varepsilon > 0$. Then stochastic stability corresponds to having η_ε converging weakly to μ when the noise level ε goes to zero.

The notion of stochastic stability goes back to Kolmogorov and Sinai. The first results, showing that uniformly hyperbolic systems are stochastically stable, on the basin of each attractor, were proved in the 1980s by Kifer and Young.

Let us point out that physical measures need not exist for general systems. A simple counter-example, attributed to Bowen, is described in Fig. 3: time averages diverge over any of the spiraling orbits in the region bounded by the saddle connections. Notice that the saddle connections are easily broken by arbitrarily small perturbations of the flow. Indeed, no robust examples are known of systems whose time-averages diverge on positive volume sets.

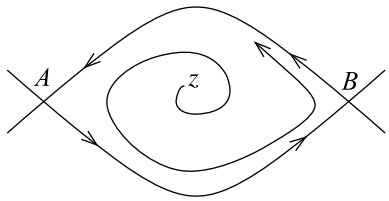

Hyperbolic Dynamical Systems, Figure 3
A planar flow with divergent time averages

Obstructions to Hyperbolicity

Although uniform hyperbolicity was originally intended to encompass a residual or, at least, dense subset of all dynamical systems, it was soon realized that this is not the case: many important examples fall outside its realm. There are two main mechanisms that yield robustly non-hyperbolic behavior, that is, whole open sets of non-hyperbolic systems.

Heterodimensional Cycles

Historically, the first such mechanism was the coexistence of periodic points with different Morse indices (dimensions of the unstable manifolds) inside the same transitive set. See Fig. 4. This is how the first examples of C^1-open subsets of non-hyperbolic diffeomorphisms were obtained by Abraham, Smale on manifolds of dimension $d \geq 3$. It was also the key in the constructions by Shub and Mañé of non-hyperbolic, yet robustly transitive diffeomorphisms, that is, such that every diffeomorphism in a C^1 neighborhood has dense orbits.

For flows, this mechanism may assume a novel form, because of the interplay between regular orbits and singularities (equilibrium points). That is, robust non-hyperbolicity may stem from the coexistence of regular and singular orbits in the same transitive set. The first, and very striking example was the geometric Lorenz attractor

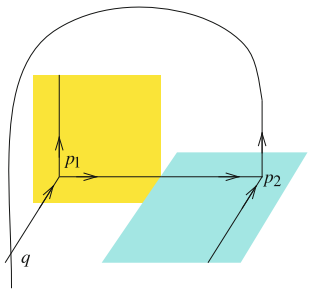

Hyperbolic Dynamical Systems, Figure 4
A heterodimensional cycle

proposed by Afraimovich, Bykov, Shil'nikov and Guckenheimer, Williams to model the behavior of the Lorenz equations, that we shall discuss later.

Homoclinic Tangencies

Of course, heterodimensional cycles may exist only in dimension 3 or higher. The first robust examples of non-hyperbolic diffeomorphisms on surfaces were constructed by Newhouse, exploiting the second of these two mechanisms: homoclinic tangencies, or non-transverse intersections between the stable and the unstable manifold of the same periodic point. See Fig. 5.

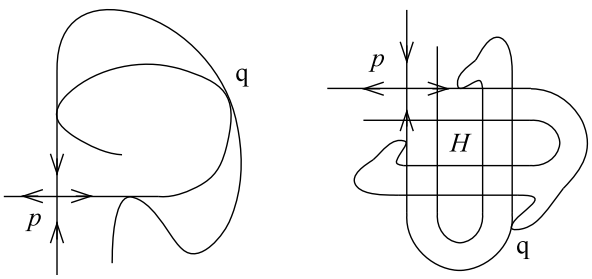

Hyperbolic Dynamical Systems, Figure 5
Homoclinic tangencies

It is important to observe that individual homoclinic tangencies are easily destroyed by small perturbations of the invariant manifolds. To construct open examples of surface diffeomorphisms with *some* tangency, Newhouse started from systems where the tangency is associated to a periodic point inside an invariant hyperbolic set with rich geometric structure. This is illustrated on the right hand side of Fig. 5. His argument requires a very delicate control of distortion, as well as of the dependence of the fractal dimension on the dynamics. Actually, for this reason, his construction is restricted to the C^r topology for $r \geq 2$. A very striking consequence of this construction is that these open sets exhibit *coexistence of infinitely many periodic attractors*, for each diffeomorphism on a residual subset. A detailed presentation of his result and consequences is given in [9].

Newhouse's conclusions have been extended in two ways. First, by Palis, Viana, for diffeomorphisms in any dimension, still in the C^r topology with $r \geq 2$. Then, by Bonatti, Díaz, for C^1 diffeomorphisms in any dimension larger or equal than 3. The case of C^1 diffeomorphisms on surfaces remains open. As a matter of fact, in this setting it is still unknown whether uniform hyperbolicity is dense in the space of all diffeomorphisms.

Partial Hyperbolicity

Several extensions of the theory of uniform hyperbolicity have been proposed, allowing for more flexibility, while keeping the core idea: splitting of the tangent bundle into invariant subbundles. We are going to discuss more closely two such extensions.

On the one hand, one may allow for one or more invariant subbundles along which the derivative exhibits mixed contracting/neutral/expanding behavior. This is generically referred to as *partial hyperbolicity*, and a standard reference is the book [5]. On the other hand, while requiring all invariant subbundles to be either expanding or contraction, one may relax the requirement of uniform rates of expansion and contraction. This is usually called *non-uniform hyperbolicity*. A detailed presentation of the fundamental results about this notion is available e. g. in [6]. In this section we discuss the first type of condition. The second one will be dealt with later.

Dominated Splittings

Let $f\colon M \to M$ be a diffeomorphism on a closed manifold M and K be any f-invariant set. A continuous splitting $T_xM = E_1(x) \oplus \cdots \oplus E_k(x)$, $x \in K$ of the tangent bundle over K is *dominated* if it is invariant under the derivative Df and there exists $\ell \in \mathbb{N}$ such that for every $i < j$, every $x \in K$, and every pair of unit vectors $u \in E_i(x)$ and $v \in E_j(x)$, one has

$$\frac{\|Df_x^\ell \cdot u\|}{\|Df_x^\ell \cdot v\|} < \frac{1}{2}\,, \tag{3}$$

and the dimension of $E_i(x)$ is independent of $x \in K$ for every $i \in \{1, \ldots, k\}$. This definition may be formulated, equivalently, as follows: there exist $C > 0$ and $\lambda < 1$ such that for every pair of unit vectors $u \in E_i(x)$ and $v \in E_j(x)$, one has

$$\frac{\|Df_x^n \cdot u\|}{\|Df_x^n \cdot v\|} < C\lambda^n \quad \text{for all} \quad n \geq 1\,.$$

Let f be a diffeomorphism and K be an f-invariant set having a dominated splitting $T_KM = E_1 \oplus \cdots \oplus E_k$. We say that the splitting and the set K are

- *partially hyperbolic* the derivative either contracts uniformly E_1 or expands uniformly E_k: there exists $\ell \in \mathbb{N}$ such that

$$\text{either} \quad \|Df^\ell \mid E_1\| < \frac{1}{2} \quad \text{or} \quad \|(Df^\ell \mid E_k)^{-1}\| < \frac{1}{2}\,.$$

- *volume hyperbolic* if the derivative either contracts volume uniformly along E_1 or expands volume uniformly

along E_k: there exists $\ell \in \mathbb{N}$ such that

$$\text{either } |\det(Df^\ell \mid E_1)| < \frac{1}{2} \quad \text{or } |\det(Df^\ell \mid E_k)| > 2\,.$$

The diffeomorphism f is *partially hyperbolic/volume hyperbolic* if the ambient space M is a partially hyperbolic/volume hyperbolic set for f.

Invariant Foliations

An crucial geometric feature of partially hyperbolic systems is the existence of invariant foliations tangent to uniformly expanding or uniformly contracting invariant subbundles: *assuming the derivative contracts E^1 uniformly, there exists a unique family $\mathcal{F}^s = \{\mathcal{F}^s(x)\colon x \in K\}$ of injectively C^r immersed submanifolds tangent to E^1 at every point of K, satisfying $f(\mathcal{F}^s(x)) = \mathcal{F}^s(f(x))$ for all $x \in K$, and which are uniformly contracted by forward iterates of f.* This is called *strong-stable foliation* of the diffeomorphism on K. Strong-unstable foliations are defined in the same way, tangent to the invariant subbundle E_k, when it is uniformly expanding.

As in the purely hyperbolic setting, a crucial ingredient in the ergodic theory of partially hyperbolic systems is the fact that strong-stable and strong-unstable foliations are absolutely continuous, if the system is at least twice differentiable.

Robustness and Partial Hyperbolicity

Partially hyperbolic systems have been studied since the 1970s, most notably by Brin, Pesin and Hirsch, Pugh, Shub. Over the last decade they attracted much attention as the key to characterizing robustness of the dynamics. More precisely, let Λ be a maximal invariant set of some diffeomorphism f:

$$\Lambda = \bigcap_{n\in\mathbb{Z}} f^n(U) \quad \text{for some neighborhood } U \text{ of } \Lambda\,.$$

The set Λ is *robust*, or *robustly transitive*, if its continuation $\Lambda_g = \cap_{n\in\mathbb{Z}} g^n(U)$ is transitive for all g in a neighborhood of f. There is a corresponding notion for flows.

As we have already seen, hyperbolic basic pieces are robust. In the 1970s, Mañé observed that the converse is also true when M is a surface, but not anymore if the dimension of M is at least 3. Counter-examples in dimension 4 had been given before by Shub. A series of results of Bonatti, Díaz, Pujals, Ures in the 1990s clarified the situation in all dimensions: robust sets always admit some dominated splitting which is volume hyperbolic; in general, this splitting needs not be partially hyperbolic, except when the ambient manifold has dimension 3.

Lorenz-like Strange Attractors

Parallel results hold for flows on 3-dimensional manifolds. The main motivation are the so-called Lorenz-like strange attractors, inspired by the famous differential equations

$$\begin{aligned} \dot{x} &= -\sigma x + \sigma y & \sigma &= 10 \\ \dot{y} &= \rho x - y - xz & \rho &= 28 \\ \dot{z} &= xy - \beta z & \beta &= 8/3 \end{aligned} \tag{4}$$

introduced by E. N. Lorenz in the early 1960s. Numerical analysis of these equations led Lorenz to realize that sensitive dependence of trajectories on the initial conditions is ubiquitous among dynamical systems, even those with simple evolution laws.

The dynamical behavior of (4) was first interpreted by means of certain geometric models, proposed by Guckenheimer, Williams and Afraimovich, Bykov, Shil'nikov in the 1970s, where the presence of strange attractors, both sensitive and fractal, could be proved rigorously. It was much harder to prove that the original Eqs. (4) themselves have such an attractor. This was achieved just a few years ago, by Tucker, by means of a computer assisted rigorous argument.

An important point is that Lorenz-like attractors cannot be hyperbolic, because they contain an equilibrium point accumulated by regular orbits inside the attractor. Yet, these strange attractors are robust, in the sense we defined above. A mathematical theory of robustness for flows in 3-dimensional spaces was recently developed by Morales, Pacifico, and Pujals. In particular, this theory shows that uniformly hyperbolic attractors and Lorenz-like attractors are the only ones which are robust. Indeed, they prove that *any robust invariant set of a flow in dimension 3 is singular hyperbolic*. Moreover, *if the robust set contains equilibrium points then it must be either an attractor or a repeller*. A detailed presentation of this and related results is given in [1].

An invariant set Λ of a flow in dimension 3 is *singular hyperbolic* if it is a partially hyperbolic set with splitting $E^1 \oplus E^2$ such that the derivative is volume contracting along E^1 and volume expanding along E^2. Notice that one of the subbundles E^1 or E^2 must be one-dimensional, and then the derivative is, actually, either norm contracting or norm expanding along this subbundle. Singular hyperbolic sets without equilibria are uniformly hyperbolic: the 2-dimensional invariant subbundle splits as the sum of the flow direction with a uniformly expanding or contracting one-dimensional invariant subbundle.

Non-Uniform Hyperbolicity – Linear Theory

In its linear form, the theory of non-uniform hyperbolicity goes back to Lyapunov, and is founded on the multiplicative ergodic theorem of Oseledets. Let us introduce the main ideas, whose thorough development can be found in e. g. [4,6,7].

The *Lyapunov exponents* of a sequence $\{A^n, n \geq 1\}$ of square matrices of dimension $d \geq 1$, are the values of

$$\lambda(v) = \limsup_{n\to\infty} \frac{1}{n} \log \|A^n \cdot v\| \tag{5}$$

over all non-zero vectors $v \in \mathbb{R}^d$. For completeness, set $\lambda(0) = -\infty$. It is easy to see that $\lambda(cv) = \lambda(v)$ and $\lambda(v + v') \leq \max\{\lambda(v), \lambda(v')\}$ for any non-zero scalar c and any vectors v, v'. It follows that, given any constant a, the set of vectors satisfying $\lambda(v) \leq a$ is a vector subspace. Consequently, there are at most d Lyapunov exponents, henceforth denoted by $\lambda_1 < \cdots < \lambda_{k-1} < \lambda_k$, and there exists a filtration $F_0 \subset F_1 \subset \cdots \subset F_{k-1} \subset F_k = \mathbb{R}^d$ into vector subspaces, such that

$$\lambda(v) = \lambda_i \quad \text{for all} \quad v \in F_i \setminus F_{i-1} ,$$

and every $i = 1, \ldots, k$ (write $F_0 = \{0\}$). In particular, the largest exponent is given by

$$\lambda_k = \limsup_{n\to\infty} \frac{1}{n} \log \|A^n\| . \tag{6}$$

One calls $\dim F_i - \dim F_{i-1}$ the *multiplicity* of each Lyapunov exponent λ_i.

There are corresponding notions for continuous families of matrices A^t, $t \in (0, \infty)$, taking the limit as t goes to infinity in the relations (5) and (6).

Lyapunov Stability

Consider the linear differential equation

$$\dot{v}(t) = B(t) \cdot v(t) , \tag{7}$$

where $B(t)$ is a bounded function with values in the space of $d \times d$ matrices, defined for all $t \in \mathbb{R}$. The theory of differential equations ensures that there exists a *fundamental matrix* A^t, $t \in \mathbb{R}$ such that

$$v(t) = A^t \cdot v_0$$

is the unique solution of (7) with initial condition $v(0) = v_0$.

If the Lyapunov exponents of the family A^t, $t > 0$ are all negative then the trivial solution $v(t) \equiv 0$ is asymptotically stable, and even exponentially stable. The stability theorem of A. M. Lyapunov asserts that, under an additional regularity condition, stability is still valid for nonlinear perturbations

$$w(t) = B(t) \cdot w + F(t, w) ,$$

with $\|F(t, w)\| \leq \text{const}\, \|w\|^{1+c}$, $c > 0$. That is, the trivial solution $w(t) \equiv 0$ is still exponentially asymptotically stable.

The regularity condition means, essentially, that the limit in (5) does exist, even if one replaces vectors v by elements $v_1 \wedge \cdots \wedge v_l$ of any lth exterior power of $\mathbb{R}^d$, $1 \leq l \leq d$. By definition, the norm of an l-vector $v_1 \wedge \cdots \wedge v_l$ is the volume of the parallelepiped determined by the vectors $v_1, \ldots, v_k$. This condition is usually tricky to check in specific situations. However, the multiplicative ergodic theorem of V. I. Oseledets asserts that, for very general matrix-valued stationary random processes, regularity is an almost sure property.

Multiplicative Ergodic Theorem

Let $f\colon M \to M$ be a measurable transformation, preserving some measure μ, and let $A\colon M \to \mathrm{GL}(d, \mathbb{R})$ be any measurable function such that $\log \|A(x)\|$ is μ-integrable. The Oseledets theorem states that Lyapunov exponents exist for the sequence $A^n(x) = A(f^{n-1}(x)) \ldots A(f(x))A(x)$ for μ-almost every $x \in M$. More precisely, for μ-almost every $x \in M$ there exists $k = k(x) \in \{1, \ldots, d\}$, a filtration

$$F_x^0 \subset F_x^1 \subset \cdots \subset F_x^{k-1} \subset F_x^k = \mathbb{R}^d ,$$

and numbers $\lambda_1(x) < \cdots < \lambda_k(x)$ such that

$$\lim_{n\to\infty} \frac{1}{n} \log \|A^n(x) \cdot v\| = \lambda_i(x) ,$$

for all $v \in F_x^i \setminus F_x^{i-1}$ and $i \in \{1, \ldots, k\}$. More generally, this conclusion holds for any vector bundle automorphism $\mathcal{V} \to \mathcal{V}$ over the transformation f, with $A_x\colon \mathcal{V}_x \to \mathcal{V}_{f(x)}$ denoting the action of the automorphism on the fiber of x.

The Lyapunov exponents $\lambda_i(x)$, and their number $k(x)$, are measurable functions of x and they are constant on orbits of the transformation f. In particular, if the measure μ is ergodic then k and the λ_i are constant on a full μ-measure set of points. The subspaces F_x^i also depend measurably on the point x and are invariant under the automorphism:

$$A(x) \cdot F_x^i = F_{f(x)}^i .$$

It is in the nature of things that, usually, these objects are *not* defined everywhere and they depend discontinuously on the base point x.

When the transformation f is invertible one obtains a stronger conclusion, by applying the previous result also to the inverse automorphism: assuming that $\log \|A(x)^{-1}\|$ is also in $L^1(\mu)$, one gets that there exists a decomposition

$$\mathcal{V}_x = E_x^1 \oplus \cdots \oplus E_x^k \,,$$

defined at almost every point and such that $A(x) \cdot E_x^i = E_{f(x)}^i$ and

$$\lim_{n\to\pm\infty} \frac{1}{n} \log \|A^n(x) \cdot v\| = \lambda_i(x) \,,$$

for all $v \in E_x^i$ different from zero and all $i \in \{1, \ldots, k\}$. These *Oseledets subspaces* E_x^i are related to the subspaces F_x^i through

$$F_x^j = \oplus_{i=1}^j E_x^i \,.$$

Hence, $\dim E_x^i = \dim F_x^i - \dim F_x^{i-1}$ is the multiplicity of the Lyapunov exponent $\lambda_i(x)$.

The angles between any two Oseledets subspaces decay sub-exponentially along orbits of f:

$$\lim_{n\to\pm\infty} \frac{1}{n} \log \operatorname{angle}\left(\bigoplus_{i\in I} E_{f^n(x)}^i \,,\ \bigoplus_{j\notin I} E_{f^n(x)}^j \right) = 0 \,,$$

for any $I \subset \{1, \ldots, k\}$ and almost every point. These facts imply the regularity condition mentioned previously and, in particular,

$$\lim_{n\to\pm\infty} \frac{1}{n} \log |\det A^n(x)| = \sum_{i=1}^k \lambda_i(x) \dim E_x^i \,.$$

Consequently, if $\det A(x) = 1$ at every point then the sum of all Lyapunov exponents, counted with multiplicity, is identically zero.

Non-Uniformly Hyperbolic Systems

The Oseledets theorem applies, in particular, when $f\colon M \to M$ is a C^1 diffeomorphism on some compact manifold and $A(x) = Df_x$. Notice that the integrability conditions are automatically satisfied, for any f-invariant probability measure μ, since the derivative of f and its inverse are bounded in norm.

Lyapunov exponents yield deep geometric information on the dynamics of the diffeomorphism, especially when they do not vanish. We call μ a *hyperbolic measure* if all Lyapunov exponents are non-zero at μ-almost every point. By *non-uniformly hyperbolic system* we shall mean a diffeomorphism $f\colon M \to M$ together with some invariant hyperbolic measure.

A theory initiated by Pesin provides fundamental geometric information on this class of systems, especially existence of stable and unstable manifolds at almost every point which form absolutely continuous invariant laminations. For most results, one needs the derivative Df to be Hölder continuous: there exists $c > 0$ such that

$$\|Df_x - Df_y\| \le \text{const} \cdot d(x, y)^c \,.$$

These notions extend to the context of flows essentially without change, except that one disregards the invariant line bundle given by the flow direction (whose Lyapunov exponent is always zero). A detailed presentation can be found in e. g. [6].

Stable Manifolds

An essential tool is the existence of invariant families of local stable sets and local unstable sets, defined at μ-almost every point. Assume μ is a hyperbolic measure. Let E_x^u and E_x^s be the sums of all Oseledets subspaces corresponding to positive, respectively negative, Lyapunov exponents, and let $\tau_x > 0$ be a lower bound for the norm of every Lyapunov exponent at x.

Pesin's stable manifold theorem states that, *for μ-almost every $x \in M$, there exists a C^1 embedded disk $W_{\text{loc}}^s(x)$ tangent to E_x^s at x and there exists $C_x > 0$ such that*

$$\operatorname{dist}(f^n(y), f^n(x)) \le C_x e^{-n\tau_x} \cdot \operatorname{dist}(y, x) \quad \text{for all } y \in W_{\text{loc}}^s(x) \,.$$

Moreover, the family $\{W_{\text{loc}}^s(x)\}$ is invariant, in the sense that $f(W_{\text{loc}}^s(x)) \subset W_{\text{loc}}^s(f(x))$ for μ-almost every x. Thus, one may define global stable manifolds

$$W^s(x) = \bigcup_{n=0}^{\infty} f^{-n}\big(W_{\text{loc}}^s(x)\big) \quad \text{for } \mu\text{-almost every } x \,.$$

In general, the local stable disks $W^s(x)$ depend only measurably on x. Another key difference with respect to the uniformly hyperbolic setting is that the numbers C_x and τ_x can not be taken independent of the point, in general. Likewise, one defines local and global unstable manifolds, tangent to E_x^u at almost every point. Most important for the applications, both foliations, stable and unstable, are absolutely continuous.

In the remaining sections we briefly present three major results in the theory of non-uniform hyperbolicity: the entropy formula, abundance of periodic orbits, and exact dimensionality of hyperbolic measures.

The Entropy Formula

The entropy of a partition $\mathcal{P}$ of M is defined by

$$h_\mu(f,\mathcal{P}) = \lim_{n\to\infty} \frac{1}{n} H_\mu(\mathcal{P}^n)\,,$$

where $\mathcal{P}^n$ is the partition into sets of the form $P = P_0 \cap f^{-1}(P_1) \cap \cdots \cap f^{-n}(P_n)$ with $P_j \in \mathcal{P}$ and

$$H_\mu(\mathcal{P}^n) = \sum_{P\in\mathcal{P}^n} -\mu(P)\log\mu(P)\,.$$

The *Kolmogorov–Sinai entropy* $h_\mu(f)$ of the system is the supremum of $h_\mu(f,\mathcal{P})$ over all partitions $\mathcal{P}$ with finite entropy. The Ruelle–Margulis inequality says that $h_\mu(f)$ is bounded above by the averaged sum of the positive Lyapunov exponents. A major result of the theorem, Pesin's entropy formula, asserts that if the invariant measure μ is smooth (for instance, a volume element) then the entropy actually coincides with the averaged sum of the positive Lyapunov exponents

$$h_\mu(f) = \int \left(\sum_{j=1}^{k} \max\{0,\lambda_j\}\right) \mathrm{d}\mu\,.$$

A complete characterization of the invariant measures for which the entropy formula is true was given by F. Ledrappier and L. S. Young.

Periodic Orbits and Entropy

It was proved by A. Katok that periodic motions are always dense in the support of any hyperbolic measure. More than that, assuming the measure is non-atomic, there exist Smale horseshoes H_n with topological entropy arbitrarily close to the entropy $h_\mu(f)$ of the system. In this context, the *topological entropy* $h(f,H_n)$ may be defined as the exponential rate of growth

$$\lim_{k\to\infty} \frac{1}{k} \log \#\{x \in H_n\colon f^k(x) = x\}\,.$$

of the number of periodic points on H_n.

Dimension of Hyperbolic Measures

Another remarkable feature of hyperbolic measures is that they are *exact dimensional*: the pointwise dimension

$$d(x) = \lim_{r\to 0} \frac{\log\mu(B_r(x))}{\log r}$$

exists at almost every point, where $B_r(x)$ is the neighborhood of radius r around x. This fact was proved by L. Barreira, Ya. Pesin, and J. Schmeling. Note that this means that the measure $\mu(B_r(x))$ of neighborhoods scales as $r^{d(x)}$ when the radius r is small.

Future Directions

The theory of uniform hyperbolicity showed that dynamical systems with very complex behavior may be amenable to a very precise description of their evolution, especially in probabilistic terms. It was most successful in characterizing structural stability, and also established a paradigm of how general "chaotic" systems might be approached. A vast research program has been going on in the last couple of decades or so, to try and build such a global theory of complex dynamical evolution, where notions such as partial and non-uniform hyperbolicity play a central part. The reader is referred to the bibliography, especially the book [2] for a review of much recent progress.

Bibliography

1. Araujo V, Pacifico MJ (2007) Three Dimensional Flows. In: XXV Brazillian Mathematical Colloquium. IMPA, Rio de Janeiro
2. Bonatti C, Díaz LJ, Viana M (2005) Dynamics beyond uniform hyperbolicity. In: Encyclopaedia of Mathematical Sciences, vol 102. Springer, Berlin
3. Bowen R (1975) Equilibrium states and the ergodic theory of Anosov diffeomorphisms. In: Lecture Notes in Mathematics, vol 470. Springer, Berlin
4. Cornfeld IP, Fomin SV, Sinaĭ YG (1982) Ergodic theory. In: Grundlehren der Mathematischen Wissenschaften Fundamental Principles of Mathematical Sciences, vol 245. Springer, New York
5. Hirsch M, Pugh C, Shub M (1977) Invariant manifolds. In: Lecture Notes in Mathematics, vol 583. Springer, Berlin
6. Katok A, Hasselblatt B (1995) Introduction to the modern theory of dynamical systems. Cambridge University Press, Cambridge
7. Mañé R (1987) Ergodic theory and differentiable dynamics. Springer, Berlin
8. Palis J, de Melo W (1982) Geometric theory of dynamical systems. An introduction. Springer, New York
9. Palis J, Takens F (1993) Hyperbolicity and sensitive-chaotic dynamics at homoclinic bifurcations. Cambridge University Press, Cambridge
10. Shub M (1987) Global stability of dynamical systems. Springer, New York

Identification of Cellular Automata

Andrew Adamatzky
Computing, Engineering and Mathematical Sciences, University of the West of England, Bristol, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Background and Basics of Identification
Identification Using Machine Learning
Identification Using Polynomial Representation
Binary Tree Representations and Genetic Programming
Identification Using Decision Trees
Identification by Immunocomputing
Application of Identification in Automatic Design of Cellular-Automata Processors
Future Directions
Bibliography

Glossary

Cellular automaton is an array of finite automata connected locally, which update their states in discrete time and at the same moments; every automaton updates its next state depending on the states of its closest neighbors.

Decision tree is a mapping from a classified set of observations about an event to the conclusion about its outcome.

Deterministic automaton has only one next state for each pair of internal and input states.

Finite automaton is an abstract machine which takes a finite number of states and transitions between the states; the machine changes its states depending on the input states.

Immunocomputing replicates principles of information processing by immune networks to perform computation.

Learning automaton modifies its transition rules depending on its past experience.

Learning classifier system is a rule-based system, a population of rules, which are processed, selected and updated using reinforcement learning techniques.

Machine learning is a subfield of artificial intelligence concerned with the design and development of algorithms and techniques that allow computers to learn – to improve automatically through experience.

Orthogonalization is subdividing a system into its distinct components.

Polynomial representation of cell-state transition rules interpret local transition rules of a cellular automaton as a Boolean or arithmetic polynomial.

Definition of the Subject

Identification of the cellular automaton is the reconstruction of cell-state transition rules from a given series of global transformations, i. e., the approximation of the minimal cellular automaton that implements given global transformations.

Introduction

The functional synthesis of finite automata problem was first formulated in the 1960s, and is as follows: given an operator on "superwords", we wish to construct a finite automaton that realizes this operator. In other words, given some language, we wish to build an automaton that represents this language [37,38]. A meta-language of regular expressions as well as algorithms for synthesis was discussed by Kleene [23]. As to the synthesis of cellular-automata-like networks, only the results obtained by [26] on the homogeneous structures with external inputs and outputs are known. The possibility of finding cellular-automata rules and initial conditions that generate a specified time series

was indicated as yet another task in cellular automaton synthesis by Voorhees [40].

Well-known methods of automata synthesis with determined structure and behavior include Markov chains and fuzzy systems for indeterministic automata, and genetic programming for deterministic automata. One of the most efficient ways is to use genetic programming in the design of automaton parts [24,25]. Thus, Andre et al. [14] used genetic programming with automatically defined functions to evolve a rule for the majority classification task for one-dimensional cellular automata. They obtained a rule with an accuracy of 82 percent.

Another way is to involve results founded on the identification of finite automata, i. e. the reconstruction of the automaton structure from the given snapshots of automaton behavior. Two pioneering works on this subject are based on the theory of experiments with finite automata [31,37]. A detailed historical overview of automata experiments can be found in a textbook on finite automata by Brauer [15]. Identification is generally very similar to the classic algorithms proposed by Trakhtenbrot and Bardzin [38], and their modifications [32].

Adamatzky [1,2,3,4,5,6,7,8,9,10,11,12,13] developed algorithms for the identification of various classes of cellular automata, i. e., approximating minimal cellular automata from a finite series of the snapshots, or configurations, recorded in the global evolution of the automaton. A basic scheme for the identification of a deterministic cellular automaton is implemented in the following steps. The minimal radius for the cell's neighborhood is chosen, and for all configurations all observable transitions in the table of cell-state transitions are collected, in the format "cell's neighborhood state at time t" $\rightarrow$ "cell's state at time $t+1$". If there are only two transitions with identical left sides but different right sides – this is a sign of indeterminism – then it is assumed the wrong neighborhood has been chosen (when identified automaton is known to be deterministic). So the radius of the neighborhood is increased and the table of cell-state transitions is collected again. Identification is complete if the state transitions for all possible states of the neighborhood are found.

Background and Basics of Identification

Cellular automaton is an array of uniform finite automata connected locally. Each finite automaton, called a *cell*, of the array takes a finite number of states. All cells update their states simultaneously, in parallel, by the same cell-state transition rule. A cell calculates its next state depending on states of its closest neighbors, known as a "cell neighborhood".

Cellular automaton can be defined by the tuple $\langle \mathbf{L}, \mathbf{Q}, u, f \rangle$, where $\mathbf{L}$ is a lattice, or an array, of cells; each cell $x \in \mathbf{L}$ takes a state from the finite set $\mathbf{Q}$; u is a cell neighborhood $u(x) = \{y \in \mathbf{L}\colon |x-y| \leq r\}$ (r is a radius, and $k = |u(x)|$ is the size of the neighborhood); $f\colon \mathbf{Q}^k \rightarrow Q$ is cell-state transition function. Let x^t be the state of cell x at time step t and $u(x)^t = (y_1^t, \ldots, y_k^t)$, the next state $x^{t+1}$of cell x is calculated as $x^{t+1} = f(u(x)^t) = f(y_1^t, \ldots, y_k^t)$. The configuration, or global configuration, of cellular automaton is an array of the cells' states.

The basic idea of identifying *deterministic* cellular automaton is as follows. Given a sequence of cellular automaton configurations $C_1 \rightarrow \ldots C_l \rightarrow C_{l+1} \rightarrow C_m$, we want to reconstruct a minimal and correct description of the cell-state transition function f. At a low level, the function f can be represented as a set, or table, of local transitions $w \rightarrow a$, where $w = \langle w_1, \ldots, w_k \rangle$ is the string of states of cell x's neighbors and $a = f(w)$. Therefore to reconstruct function f one has to select some minimal radius for the cell neighborhood, scan the given sequence $C_1 \rightarrow \ldots C_l \rightarrow C_{l+1} \rightarrow C_m$ of cellular automaton configurations, and collect all observable transitions in the form $w \rightarrow a$. Then the consistency of the set of local transitions is checked. If there are two local transitions $w \rightarrow a$ and $w \rightarrow b$, where $w \in \mathbf{Q}^k$ and $a, b \in \mathbf{Q}$, with the same left part w and different right parts, $a \neq b$, then we consider that the calculated candidate for f violates the determinism of the identified cellular automaton. This may mean that we have chosen too small of a neighborhood which does not include all neighbors influencing the cell's state transition. So, the radius of the neighborhood is increased, the sequence of the given global configurations is re-scanned, and the set of local transitions is rebuilt. The procedure is carried out until a complete set of non-contradicting transitions is calculated.

Example 1 (Identification of a one-dimensional cellular automaton) Given two configurations of a one-dimensional deterministic cellular automaton of ten cells with periodic boundary conditions, $c^t = 0011101000$ and $c^{t+1} = 0110101100$, we want to reconstruct the cell-state transition function. We start identification with a minimal neighborhood $u(x_i) = (x_{i-1}, x_{i+1})$, where $i = 1, \ldots, 10$ and collect all observable transitions $(x_{i-1}^t, x_{i+1}^t) \rightarrow x_i^{t+1}$:

$$00 \rightarrow 0$$
$$01 \rightarrow 1$$
$$11 \rightarrow 0$$
$$10 \rightarrow 1$$
$$00 \rightarrow 1\,.$$

Transitions $00 \to 0$ and $00 \to 1$ contradict each another. This means that the chosen neighborhood is insufficient in describing the conditions of local transitions. Let us include the central cell in the neighborhood: $u(x_i) = (x_{i-1}, x_i, x_{i+1})$ and collect local transitions in the form $(x_{i-1}^t, x_i^t, x_{i+1}^t) \to x_i^{t+1}$:

$$
\begin{aligned}
&000 \to 0\\
&001 \to 1\\
&010 \to 1\\
&011 \to 1\\
&100 \to 1\\
&101 \to 0\\
&110 \to 1\\
&111 \to 0\,.
\end{aligned}
$$

Now we do not have contradicting transitions, so we assume that we reconstructed the minimal cellular automaton which implements the global transformation $c' \to c''$. The automaton is identified completely. Having the cell-state transition rules, the behavior of the automaton can be investigated further, e. g. global behavior can be studied in terms of basin attraction fields (see Fig. 1).

Identification is assumed to be complete if state transitions for all possible states of the neighborhood are found. The identification is incomplete, or situational, otherwise. More algorithms and examples of identification of deterministic cellular automata, automata with memory, structurally dynamic cellular automata, and asynchronous automata are provided in [10].

Classical identification of d-dimensional cellular automata implies explicit construction of a cell-state transition, or look-up, table. Let each cell of an identified cellular automaton have q states, and each cell updates its states in discrete time depending on the states of its immediate neighbors in a neighborhood of radius r. There are q^k, possible local states, or configurations of cell neighborhood, where $k = (2r+1)^d$; so, the complete look-up table has q^k lines, each of length $k+1$. To find a minimal possible model we start by identifying the automaton with a neighborhood radius of r=1 and gradually increase it during identification. This usually leads to an enormous computational effort for the identification, e. g., time complexity $O(\sqrt{k}\, q^k)$, which becomes an unrealistic number even for small numbers of cell-states and quite a modest size of the cell neighborhood. In the rest of the article we consider several techniques aimed to reduce complexity of identification.

Identification Using Machine Learning

Tools of evolutionary computing and genetic programming have already been used extensively to evolve cellular automata aimed to perform certain range of tasks, see e. g. [19,20,22,25,29,30]. However, in these works the authors used some global measure, e. g. the density of cells at some specified state (density classification task), to select the most successful rules. In the present section we will discuss how cell-state transition rules can be determined automatically from a given sequence of cellular automaton configurations using machine learning techniques.

We consider the machine learning approach for identification of binary cellular automata. It employs accuracy-based learning classifier system designed by Bull [16]. This is a memory-less system in which the rulebase consists of N action rules, and the condition is a string of characters from the ternary alphabet {0,1,#}, where 0 and 1 are cell states, and # is a value of undefined cell-state transition. The action is also represented by a binary string. With each rule we associate a predicted payoff value p, a scalar error ε, and an estimate σ of the average size of action sets in which that rule participates. When the input message is re-

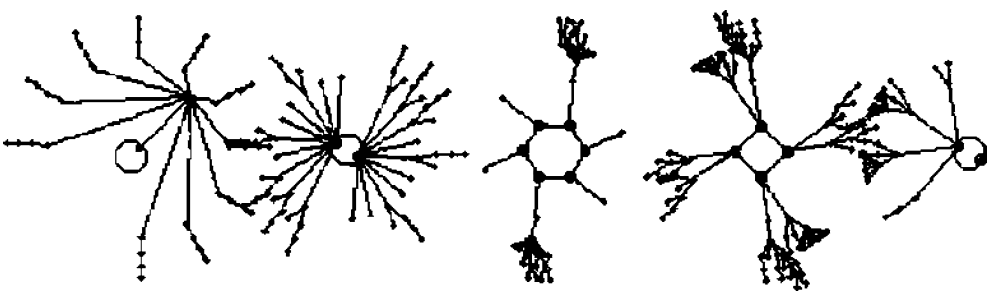

Identification of Cellular Automata, Figure 1
Global transition graphs of a one-dimensional cellular automaton, identified from the snapshot 0011101000 → 0110101100. Nodes correspond to the configurations, edges to the global transitions (produced by DDLab, www.ddlab.com)

ceived by the learning system, its rulebase is scanned, and the rules with conditions matching the input message at each position are marked as members of the current match set M. The action is selected from actions proposed by elements M. The rules proposing the selected action form the action set A.

Bull [16] uses immediate reward and reinforcement via updating the error, the niche size estimate, and the payoff estimate of each element of the current action set A using the Widrow-Hoff delta rule with learning rate β. Offsprings are produced via mutation and single-point crossover. Existing members of the rulebase are replaced randomly based on the estimated size of action set.

Let us consider an example of the identification of the one-dimensional cellular automaton with a three-cell neighborhood and two cell states, see details in [18]. Cells of the automaton update their states by the following rule: <State of the left neighbor, state of the central cell, state of the right neighbor> → new state of the central cell.

000 → 1
001 → 0
010 → 1
011 → 1
100 → 0
101 → 1
110 → 1
111 → 0

Examples of space-time snapshots of the automaton developing from random initial configurations are shown in Fig. 2.

At the beginning of identification, the learning classifier system has no information about the size of the cell neighborhood or the cell-state transition rules. The system represents automaton as a random automaton, where every cell takes its next state with probability 0.5, irrelevant to the state of the cell's neighborhood, see an example of such development in Fig. 3a.

After 2000 trials the learning classifier system extracted the rules (we also show prediction p, error ε and action set size σ estimates):

	p	ε	σ
110 → 1	: 1000.00000	: 0.00000	: 53.00009
#00 → 1	: 555.93876	: 527.68224	: 66.33652
01# → 1	: 000.00000	: 0.00000	: 76.05551
##1 → 0	: 32.45343	: 456.82484	: 50.10989

Quite an accurate generalization of 01# → 1 has already been identified, i. e., the fact that the right neighbor state is redundant for states 010 and 011 has been learned ('#' indicates a wildcard). The accurate rule 110 has also been found. The two other general rules mean the learning classifier system put the wrong states for two of the remaining five possible states. Thus the partially incorrect rule table is as shown below (contradicting transitions are underlined):

000 → 1 (rule#00 → 1)
001 → 0 (rule##1 → 0)
010 → 1 (rule01# → 1)
011 → 1 (rule01# → 1)
<u>100 → 1</u> (rule#00 → 1)
<u>101 → 0</u> (rule##1 → 0)
110 → 1 (rule110 → 1)
111 → 0 (rule##1 → 0)

Development of this partially identified cellular automaton is shown in Fig. 3b.

After ten times more steps of learning and identification, on the 20 000th trial, more accurate cell-state transition rules are discovered:

001 → 0	: 1000.00000	: 0.00000	: 19.73689
100 → 0	: 1000.00000	: 0.00000	: 23.87451
110 → 1	: 1000.00000	: 0.00000	: 25.97547
0#0 → 1	: 1000.00000	: 0.00000	: 22.78546
01# → 1	: 1000.00000	: 0.00000	: 21.89581
##1 → 0	: 526.06540	: 495.97291	: 36.40984

The table includes one more possible generalization: 0#0 → 1. The learning classifier system got the wrong rule 101 → 1 because of the persistence of the rule ##1 → 1. At this stage of identification, the cell-state transition rule table looks as follows:

000 → 1 (rule0#0 → 1)
001 → 0 (rule##1 → 0)
010 → 1 (rule01# → 1)
011 → 1 (rule01# → 1)
100 → 0 (rule100 → 0)
<u>101 → 0</u> (rule##1 → 0)
110 → 1 (rule110 → 1)
111 → 0 (rule##1 → 0)

The development of a cellular automaton governed by such transition table is shown in Fig. 3c.

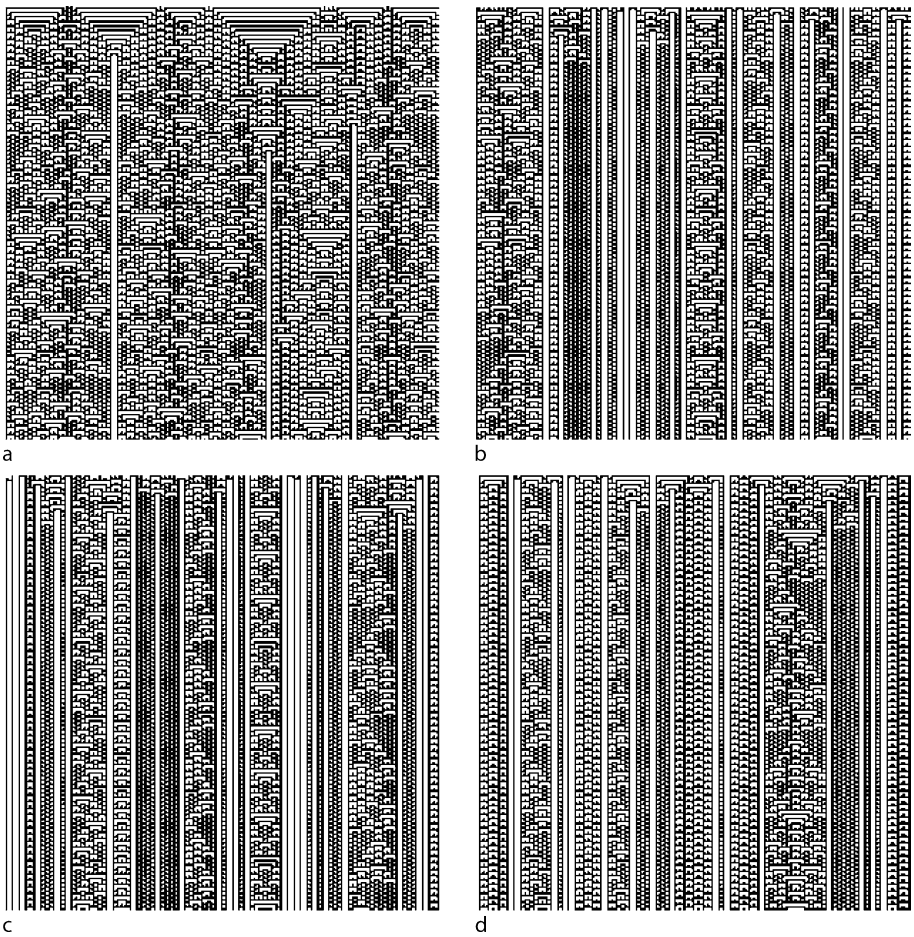

Identification of Cellular Automata, Figure 2

Example space-time configurations of a one-dimensional cellular automaton with 200 cells. Initially each cell is assigned the state 1 with probability a 0.01, b 0.3, c 0.6, and d 0.9, and state 0 with probabilities, 0.99, 0.7, 0.4 and 0.1, respectively. Time goes down. A black pixel represents state 1, and a blank pixel represents state 0 [18]

The minimum radius and complete table of cell-state transitions are computed by the learning classifier system in 100,000 steps as follows:

001 → 0	: 1000.00000	: 0.00000	: 18.78942
100 → 0	: 1000.00000	: 0.00000	: 17.33453
101 → 1	: 1000.00000	: 0.00000	: 18.21793
110 → 1	: 1000.00000	: 0.00000	: 17.65666
111 → 0	: 1000.00000	: 0.00000	: 17.89178
0#0 → 1	: 1000.00000	: 0.00000	: 20.78546
01# → 1	: 1000.00000	: 0.00000	: 21.89581

The appropriate generalization is evolved:

000 → 1	(rule0#0 → 1)
001 → 0	(rule001 → 0)
010 → 1	(rule01# → 1)
011 → 1	(rule01# → 1)
100 → 0	(rule100 → 0)
101 → 1	(rule101 → 1)
110 → 1	(rule110 → 1)
111 → 0	(rule111 → 0)

These rules give an identical space-time evolution from the same random initial configuration as the underlying cellular automaton shown in Figure 3d. More details regarding the performance and behavior of Bull's learning classified system used in the identification of cellular automata can be found in [18].

Identification Using Polynomial Representation

In developing their approaches to identify cellular automata, Billings and collegues [43,44,45,46,47] used poly-

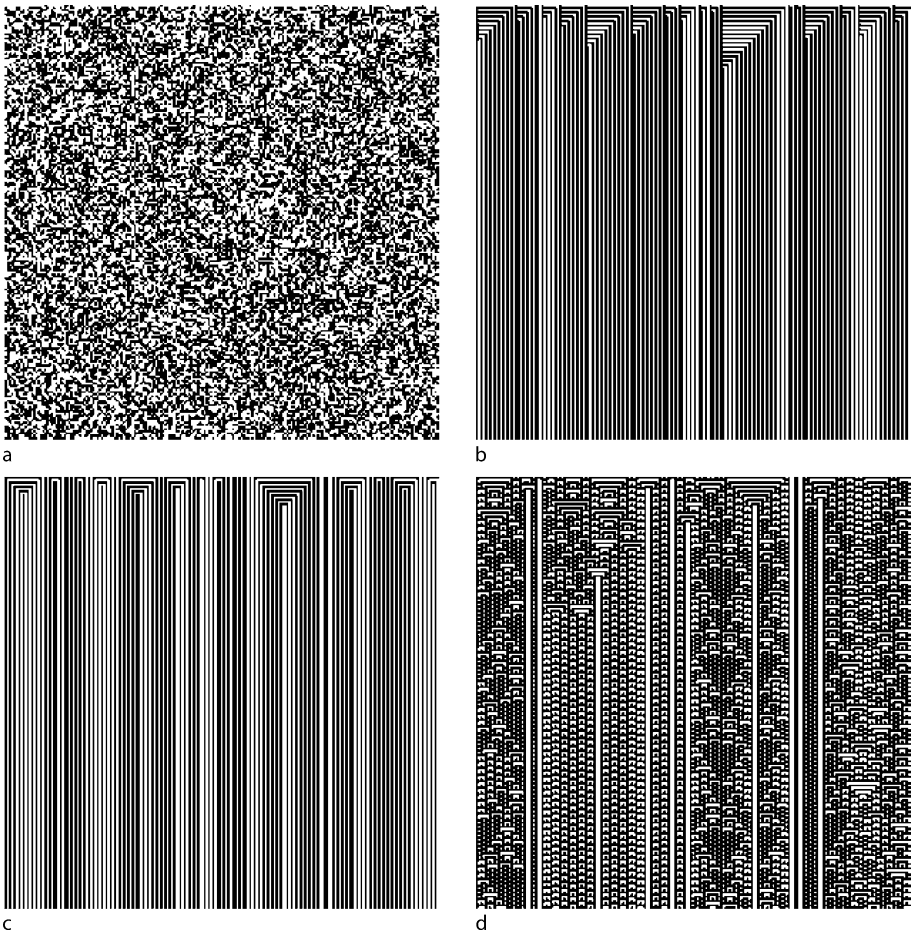

Identification of Cellular Automata, Figure 3

Examples of space-time configurations of cellular automaton (CA). a cells which update their states by the rules extracted after the 1st trial, b after the 2000th trial, c after the 20 000th trial, and d after the 100 000th trial. Initially, each cell is assigned the state 1 with probability 0.5, and is assigned the state 0 otherwise. For all rules, the CA started its development in the same random configuration. Time goes down. A black pixel represents state 1, and a blank pixel represents state 0 [18]

nomial representations of local transition rules. It is a well know and explored fact (see [40,41] for overview and further details) that cell-state transition rules of binary cellular automata can be represented as an expression from Boolean algebra, with conjunction and exclusive disjunction operations. The elements of the cell neighborhood are variables in such a logical expression. The logical expression in turn can be represented by an arithmetical polynomial with operations of multiplication and binary addition.

Yang and Billing [44] are applying orthogonalisation to derive an algorithm for detecting significant terms and estimate coefficients of the polynomial representation of local transition rules. Significant terms indicate which elements of a cell neighborhood should be necessarily taken into consideration when extracting local transitions. They are using error reduction ratios to evaluate how much each element of the cell neighborhood contributes to updating the state of the cell. The methods of selecting the correct neighborhood can also be based on statistics associated with neighborhood elements and mutual information [46].

The polynomial-representation-based-identification seems to also work for stochastic cellular automata. Thus, Yang and Billing [43] demonstrate that binary-state stochastic automata are well modeled by an integer-parametrized polynomial disturbed by noise. To detect the minimal and sufficient neighborhood one should select correct terms of the model.

Binary Tree Representations and Genetic Programming

El Yacoubi and Jacewicz [20,22] use genetic algorithms to discover and design cell-state transition functions. They are fulfilling goal-based identification tasks, where cellular automata are not extracted from a set of given configuration but rather designed from scratch to satisfy certain global conditions in their space-time evolution. We will discuss the El Yacoubi-Jacewicz approach briefly because similar ideas are explored in more depth by Maeda and Sakama [27,28], as shown in the next section.

When binary state cellular automata are concerned, a cell-state transition function is represented as a binary tree (Fig. 4), where each node corresponds to some binary logic operator, and the leaves represent cells, or elements, of the neighborhood. Genetic algorithms can then be used to evolve 'optimal' trees, which in turn give us best version of cell-state transition function, where the fitness criterion is a measure of how good the desired global dynamics can be represented by the cell-state transition function [20,22].

The local transition functions are developed in the following manner. We select the initial size of the tree, the crossover and mutation probabilities, the number of iterations, the population size, and the fitness values. Then a Boolean function is selected, and a terminal set of accepted neighborhoods, or neighborhood sizes. The random population of cell-state transition functions creates a set of configurations. For each configuration a fitness value is calculated (as some global characteristic, e. g. number of cells in certain state). Based on the fitness values, the local transition functions are selected and subject to crossover and mutation operations.

Using their approach, El Yacoubi and Jacewicz [20,22] discovered the cell-state transition rule for the uniformity problem, the synchronization problem, and the density classification problem.

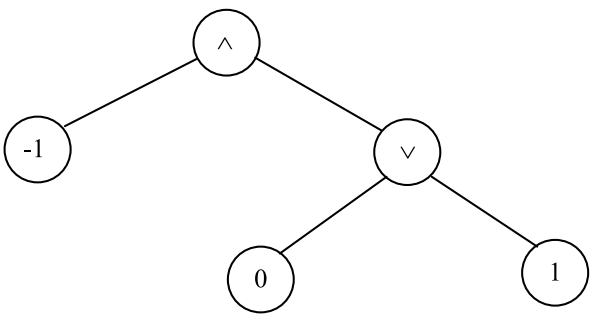

Identification of Cellular Automata, Figure 4
Binary tree representation of the cell-state transition function: $x_i^{t+1} = x_{i-1}^t \wedge (x_i^t \vee x_{i-1}^t)$

Identification Using Decision Trees

Maeda and Sakama [27,28] employ decision trees and genetic programming to identify cellular automata. Their approach is basically very similar to the identification algorithms designed by Adamatzky [10], with some improvements in computational complexity. Given a sequence of automaton configurations, local transitions are collected as evidences, which are then classified as a decision tree. A cell-state transition table is derived from the decision tree using genetic programming.

An evidence is a pair: the state of the neighborhood, and the next state of the neighborhood's central cell. The first element of each evidence is selected from a cellular automaton configuration at a time step t and, the second element at time step $t + 1$. Evidences are collected in a set. Identification starts with some minimal neighborhood. The radius of the neighborhood is increased if there are contradicting evidences, i. e. two evidences with similar first elements and different second elements. Redundant cells are removed from the final neighborhood by the standard procedure suggested in [10].

What are the classification conditions in Maeda-Sakama's identification techniques? Evidences are classified by their neighborhood patterns [27,28]. The explicit, spatial, state of the cell neighborhood is converted to if-then rules involving the enumeration of neighbors and the direct indication of their states; e. g., "If the South-West and North-East neighbors have state 1, then the central cell will take state 1". These are the classification conditions. Classification conditions are represented by condition trees, as in El Yacoubi-Jacewicz [20,22].

Genetic algorithms are applied to condition trees to find classification conditions which correctly classify all evidences. Several condition trees can be connected in one, so a condition tree is extended until it represents a classification condition which classifies all evidences.

Each node of the decision tree has a certain condition value which is computed from the condition of a classification condition of the previous node and the next state of the cell based on the classification condition. Given a neighborhood state of a cell, a decision tree returns next state of the cell.

Let us consider the following example as it is discussed by Maeda and Sakama [28]. Given a sequence of configurations (Fig. 5) for a two-dimensional cellular automaton with three cell-states and a five-cell von Neumann neighborhood, we want to extract the rules of cell-state

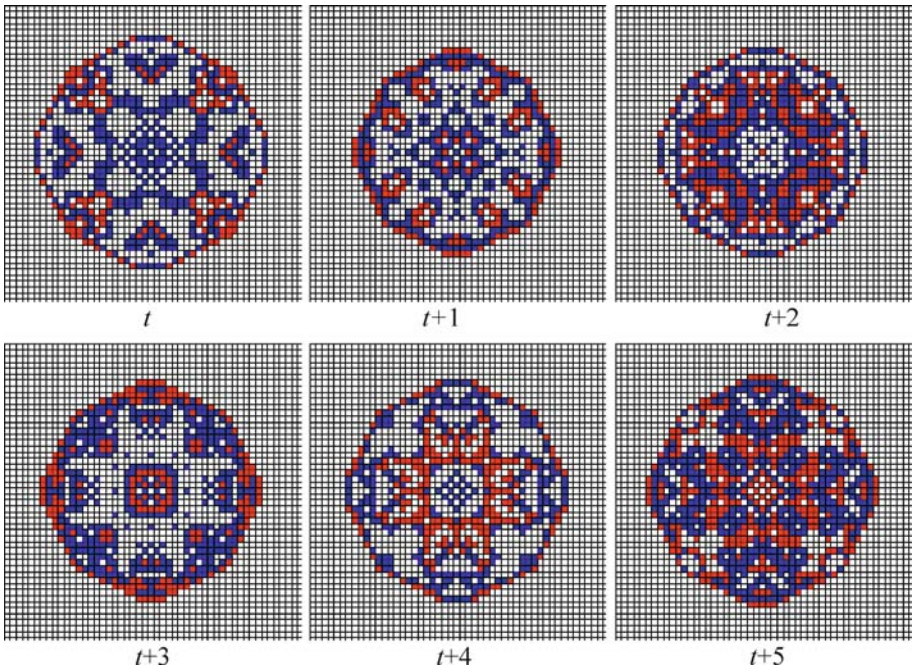

Identification of Cellular Automata, Figure 5
Given snapshots of two-dimensional cellular automaton to identify

transitions. Let us mark cells in the neighborhood according to their geographical position relative to the neighborhood's central cell: South (S), North (N), West (W), East (E) and Central (C) cells. The condition and decision trees obtained from the sequence of configurations in Fig. 5 are primitive. The condition tree T has one root, an addition operation, and six leaves: S, N, W, E, C, C (the central cell is used twice). The decision tree has one root and eight leaves, expressed by the following if-then rules (where R is the outcome of a condition in a decision tree):

$$\begin{array}{llll} \text{If} & T = 0 & \text{then} & R = 0 \\ \text{If} & T = 1 & \text{then} & R = 0 \\ \text{If} & T = 2 & \text{then} & R = 1 \\ \text{If} & T = 3 & \text{then} & R = 1 \\ \text{If} & T = 4 & \text{then} & R = 2 \\ \text{If} & T = 5 & \text{then} & R = 2 \\ \text{If} & T = 6 & \text{then} & R = 2 \\ \text{If} & T = k & \text{then} & R = 0\,, \end{array}$$

where $7 \leq k \leq 12$.

This can be minimized to the following set of rules. $P(x, t)$ is the arithmetical sum of the cell-states in a neighborhood of the cell x at the time step t, and x^{t+1} is the state of the cell x at the time step $t + 1$. Then

$$\begin{array}{l} \text{if } P(x,t) \leq 1 \text{ then } x^{t+1} = 0 \\ \quad \text{else if } P(x,t) \in \{2,3\} \text{ then } x^{t+1} = 1 \\ \qquad \text{else if } P(x,t) \in \{4,5,6\} \text{ then } x^{t+1} = 2 \\ \qquad\quad \text{if } P(x,t) \geq 7 \text{ then } x^{t+1} = 0\,. \end{array}$$

Identification by Immunocomputing

Tarakanov and Prokaev [34] represent cell-state transitions of cellular automaton by an artificial immune network. The number of possible entries in the state transition table is reduced by using apoptosis (programmed cell death, or cell suicide) and immunization [35].

A set of parameters, or state vector, is defined for each cell of the identified automaton: $X = [x_1 \ldots x_n]'$. The vector is comprised of a sequence of cell states (extracted from series of given snapshots) $c^t, c^{t-1}, \ldots, c^{t-p}$ and states of the cell's neighbors, along an identified period: $u^t, u^{t-1}, \ldots, u^{t-p}$. Vector X does not have to be defined completely.

To identify a cellular automata, Tarakanov and colleagues [34,35] employ their techniques for pattern recognition by immune-computing. This is simulated by molecular recognition, as a computation of binding energy between an antigen, n-dimensional input vector, and anti-

bodies, singular vectors of the single value decomposition of a training matrix [35]. The procedure is built of two parts: two stages of training, and recognition.

At the first stage of training, data are mapped to a formal immune network: training patterns are acquired, a training matrix is formed, and a single vector decomposition (binding energies) of the training matrix is calculated. Single vectors obtained are analogs of antibody-probes. Then the obtained data are compressed by apoptosis – if there are two neighboring cells with the same value just one cell is kept alive, and immunization – if the 'killed' cell has no neighbors with the same value as the cell has then the cell is restored. Immunization is used to correct mistakes accumulated during apoptosis.

At the second part, recognition is implemented. The cellular automaton pattern (analog of antingen) is mapped into the formal immune network. Then the nearest cell, corresponding to the pattern of the network, is detected, and the class, or local transition function of identified cellular automaton, of the nearest cell is assigned to the pattern [34,35].

Application of Identification in Automatic Design of Cellular-Automata Processors

Identification strategies can be used in developing automatic programming techniques for massively parallel cellular automaton processors [11]. Let us consider two examples of the programming of two-dimensional semi-totalistic cellular automata, with a binary set of cell states and an eight-cell neighborhood.

Example 1 (Computation of a discrete convex hull) A 45-convex hull of the subset **S** of nodes of an integer lattice is an intersection of all discrete half-places containing **S** [21,33]. A discrete 45-halfplane is the set of all such nodes (i, j) of an integer lattice that satisfy the inequality $a \cdot i + b \cdot j \leq c$ for $|a|, |b| \leq 1$ and integer c.

Given a connected set of black pixels (a cellular automaton configuration), we want to design a cellular-automaton processor which computes the discrete convex hull of this set. The given set must be mapped on to the lattice in such a manner that cells having the same indices as the nodes of the given set take state 1, whereas others remain in rest state 0. The domain grows until it reaches the shape of the convex hull. The state 1 is an absorbing state because every cell initially "excited" with state 1 remains in the state 1 forever.

In discovering the cell-state transition function, we will employ the fact that convex hull must be a stationary configuration of a cellular automaton. Therefore we take two identical configurations as shown in Fig. 6. From the transition 'given global transition source configuration' → target configuration we extracted conditions of the cell-state transition 0 → 0. It was found that a cell in state 0 remains in state 0 if it has 0, 1, 2, or 3. This means that a cell in state 0 takes state 1 if more than three of its neighbors are in state 1.

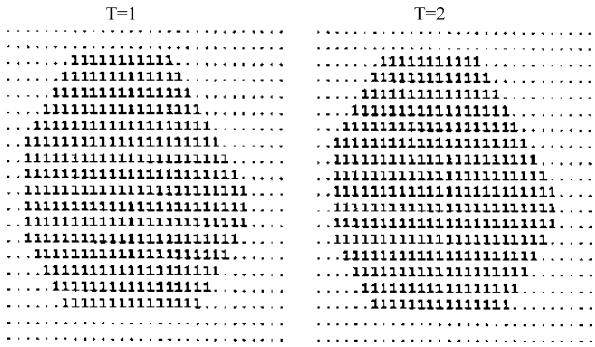

Identification of Cellular Automata, Figure 6
Source configuration ($T = 1$) and target configuration ($T = 2$) are the input for the tool of automatic programming of a cellular automaton that builds a discrete convex hull

The resultant local transition rule is as follows:

$$x^{t+1} = \begin{cases} 1\,, & (x^t = 1) \text{ or} \\ & \quad ((x^t = 0) \text{ and } (\sum_{y \in u(x)} y^t > 3)) \\ 0\,, & \text{otherwise}\,, \end{cases}$$

where x^t is the state of cell x in time step t, and $u(x)$ is a neighborhood of the cell x. The evolution of a designed cellular automaton is presented in Fig. 7.

Example 2 (Detecting the contour of a convex image) Given a convex solid image filled entirely with black pixels, we want to design a cellular automaton which computes the contour of the image. We represent a given image and its contour on the integer lattice as two separate configurations of a cellular automaton. The image is a source configuration, and the contour is a target configuration (Fig. 8).

Taking for granted a minimal eight-cell neighborhood, all observable states of the neighborhood for the transitions 1 → 1, 1 → 0 and 0 → 0 are collected. The transition 0 → 0 is unconditional. It takes place independently of the current state of the neighborhood. The two other transitions are conditional. The black pixel (cell state 1) remains black if four or five of its neighbors are black pixels; otherwise it becomes white (cell state 0). A two-dimensional, semi-totalistic cellular automaton was designed, which has an eight-cell neighborhood, two cell states, and

```
.......................
.......................
.....1.................
.................1.....
.....1.........1.......
.......1.......1.......
.......................
.......1.................
........1....1.........
.........1...1........
..........1.1.........
..1111111111.1111111111..
.........1.1..........
........1...1.........
.......1.....1........
......1.......1.......
.....1.........1......
.......................
.......................
```
T=0

```
.......................
.......................
.....1.................
.................1.....
.....1.........1.......
.......1.......1.......
.....111...............
.....1.....1...........
.....1111111111.......
.....111111111111.....
.....1111111111111111.
.....111111111111.....
.....111111111111.....
.....1111111111111111.
.....111111111111.....
.....111111111111.....
.....1111111111111111.
.....111111111111.....
```

```
T=1                            T=2
.............................  .............................
.............................  .............................
.......11111111111............  .......11111111111............
......111111111111............  ......1...........1...........
.....1111111111111............  .....1.............1..........
....11111111111111............  ....1...............1.........
...111111111111111............  ...1.................1........
..1111111111111111111.........  ..1...................1.......
..11111111111111111111........  ..1.....................1.....
..111111111111111111111.......  ..1......................1....
..1111111111111111111111......  ..1.......................1...
..11111111111111111111111.....  ..1.......................1...
..111111111111111111111111....  ..1.......................1...
..11111111111111111111111.....  ..1.....................1.....
..111111111111111111111111....  ..1.....................1.....
..1111111111111111111111111...  ..1.....................1.....
...111111111111111111111111...  ...1...................1......
....11111111111111111111......  ....1.................1.......
.....111111111111111111.......  .....1...............1........
......11111111111111111.......  ......11111111111111111.......
.............................  .............................
.............................  .............................
```

Identification of Cellular Automata, Figure 8

The source configuration ($T = 1$) and the target configuration ($T = 2$) are the input for the automatic programming of a cellular automaton that computes counter of convex image

the following function of the local transitions:

$$x^{t+1} = \begin{cases} 1\,, & (x^t = 1) \text{ and } (4 \le \sum_{y \in u(x)} y^t \le 5) \\ 0\,, & (x^t = 0) \text{ or } (x^t = 1) \text{ and} \\ & \quad (\sum_{y \in u(x)} y^t \in [0..3] \cup [6..8])\,, \end{cases}$$

where x^t is the state of cell x in time step t, and $u(x)$ is a neighborhood of cell x.

Future Directions

Theoretical studies are only powerful when they can be applied in real life. Identification of real-world spatially extended systems proved to be a non-trivial task. In computational models, the sequence of global transformations presented for identification are always discrete, noise-free, and can be easily handled by identification software systems. Nevertheless, some progress has been made.

In 1996 Adamatzky and Tolmachev [13,36] managed to identify local transitions rules of a cellular-automaton that constructs the Voronoi diagram, the tessellation of a plane from given planar set. The cell-state transition rules discovered helped to locate a proper set of chemical species, suitable for the task, and fabricate a chemical laboratory processor which approximates the Voronoi diagram [13,36].

In 2003, Bull et al. [17] proposed to undertake an experimental identification of real-world reaction-diffusion chemical systems, including the Belousov-Zhabotinsky excitable chemical medium, the gel mediated chlorite-iodide-malonic acid reaction, the inorganic gel mediated pattern forming reactions based on the reaction of copper chloride and potassium ferrocyanide, and many other systems. The ideas were independently developed by Zhao et al. [45,46,47], who approximated the local transition function of a minimal cellular automaton from snapshots of an experimental Belousov-Zhabotinsky medium [46].

Tarakanov and Prokaev [34] demonstrated how to employ identification techniques to build a cellular-automaton model of a temperature field generation using real field data from Barents Sea.

We expect that identification techniques will be applied in reconstructing local transition rules of natural spatially extended systems: interacting populations, morphogenesis, reaction-diffusion systems, homogeneous neural networks, as well in design of more efficient massively parallel processors with cellular automaton architecture.

Bibliography

Primary Literature

1. Adamatzky A (1990) Identification of probabilistic cellular automata. Izv AN SSSR Ser Tekhnicheskaya Kibern 3:95–100 (in Russian). Translated in Soviet (1990) J Comput Syst Sci 30:118–123
2. Adamatzky A (1991) Identification of fuzzy cellular automata. Autom Comput 6:75–80
3. Adamatzky A (1992) Complexity of cellular automata identification. Avtom Telemeh 9:72–86 (in Russian). Translated (1992) Complexity of identification of cellular automata. Autom Remote Control 53 9/2:1449–1458
4. Adamatzky A (1992) Identification of nonstationary cellular automata. J Comp Sci Tech 7:379–382
5. Adamatzky A (1992) On complexity of identification of nonstationary cellular automata. Izv AN SSSR Ser Tekhnicheskaya Kibern 3:74–79 (in Russian)
6. Adamatzky A (1993) Implantation of cellular automata. Appl Math Comput 55:49–71
7. Adamatzky A (1993) Identification of distributed intelligence. Izv AN SSSR Ser Tekhnicheskaya Kibern 6:359–369 (in Russian). Translated (1995) Recognition of distributed intelligence. J Comp Syst Sci Int 33:160–169
8. Adamatzky A (1994) Hierarchy of fuzzy cellular automata. J Fuzzy Sets Syst 62:167–174
9. Adamatzky A (1994) On complexity of serial simulation of cellular-automata mappings. Avtom Telemeh 3 (in Russian). Translated (1994) Complexity of sequential realisation of cellular-automata maps. Autom Remote Control 55 2/2:271–280
10. Adamatzky A (1994) Identification of cellular automata. Taylor and Francis, London
11. Adamatzky A (1997) Automatic programming of cellular automata: Identification approach. Kybernetes 26:126–135
12. Adamatzky A, Bronnikov V (1989) Identification of additive cellular automata. Izv AN SSSR Ser Tekhnicheskaya Kibern 3:200–205 (in Russian). Translated in Soviet (1990) J Comput Syst Sci 28:47–51
13. Adamatzky A, Tolmachiev D (1997) Chemical processor for computation of skeleton of planar shape. Adv Mater Opt Electron 7:535–539
14. Andre D, Bennet FH III, Koza JR (1996) Discovery by genetic programming of a cellular automata rule that is better than any known rule for the majority classification problem. In: Koza

JR (ed) Genetic programming 1996. Proceedings of the 1st Annual Conference. MIT Press, Cambridge
15. Brauer W (1984) Automaton-theorie. Teubner BG, Stuttgart
16. Bull L (2005) Two simple learning classifier systems. In: Bull L, Kovacs T (eds) Foundations of learning classifier systems. Springer, New York, pp 63–90
17. Bull L, Adamatzky A, De Lacy Costello B (2003) The automatic identification of spatially extended reaction-diffusion systems. EPSRC Proposal GR/S68798/01, University of the West of England
18. Bull L, Adamatzky A (2007) A learning classifier system approach to the identification of cellular automata. J Cell Autom (in press)
19. Das R, Crutchfield J, Mitchell M, Hanson J (1995) Evolving globally synchronized cellular automata. In: Eshelman L (ed) Proc 6th Int Conf on Genetic Algorithms. Morgan Kaufmann, New York, pp 336–343
20. El Yacoubi S, Jacewicz P (2007) A genetic programming approach to structural identification of cellular automata. J Cell Autom (in press)
21. Heijmans HJAM (1990) Iteration of morphological transformations. CWI Q 9:19–36
22. Jacewicz P, El Yacoubi S (1999) A genetic programming approach to structural identification of cellular automata. In: Proc of 3th International Conference on Parallel Processing and Applied Mathematics, Kazimierz Dolny, Poland, 1999, pp 148–157
23. Kleene SC (1956) Representation of events in nerve nets. In: Shannon CE, McCarthy J (eds) Automata Studies. Princeton University Press, Princeton pp 3–41
24. Koza JR (1994) Genetic programming II: Automatic discovery of reusable programs. MIT Press, Cambridge
25. Koza J, Bennett III F, Andre D, Keane M (1999) Genetic programming III: Darwinian invention and problem solving. Morgan Kaufmann, New York
26. Kudrjavtzev VB, Podkolzin AS, Bolotov AA (1990) The Foundations of the theory of homogeneous structures. Nauka Publishers, Moscow
27. Maeda K, Sakama C (2003) Discovery of cellular automata rules using cases. In: Proceedings of the 6th International Conference on Discovery Science. Lecture Notes in Artificial Intelligence, vol 2843. Springer, Heidelberg, p 357–364
28. Maeda K, Sakama C (2007) Identifying cellular automata rules. J Cell Autom (in press)
29. Mitchell M, Crutchfield J, Hraber P (1994) Evolving cellular automata to perform computations: Mechanisms and impediments. Phys D 75:361–391
30. Mitchell M, Hraber PT, Crutchfield JP (1993) Revisiting the edge of chaos: Evolving cellular automata to perform computations. Complex Syst 7:89–130
31. Moore EF (1956) Gedanken-experiments on sequential machines. In: Shannon CE, McCarthy J (eds) Automata studies. Princeton University Press, Princeton, pp 129–153
32. Murphy KP (1996) Passively learning finite automata. Technical Report 96-04-017, Santa-Fe Institute, Santa Fe, CA
33. Ronse C (1985) Definition of convexity and convex hulls in digital images. Bull Soc Math Belg Ser B 37:71–85
34. Tarakanov A, Prokaev A (2007) Identification of cellular automata by immunocomputing. J Cell Autom (in print)
35. Tarakanov A, Skormin V, Sokolova S (2003) Immunocomputing: Principles and applications. Springer, New York
36. Tolmachiev D, Adamatzky A (1996) Chemical processor for computation of Voronoi diagram. Adv Mater Opt Electron 6:191–196
37. Trakhtenbrot BA (1957) On operators, realizable in logical nets. DAN SSSR, Proceedings of the USSR Academy of Sciences 112:1005–1007 (in Russian)
38. Trakhtenbrot BA, Bardzin YAM (1970) Finite automata (Behaviour and synthesis). Nauka Publishers, Moscow
39. Von Neuman J (1990) Theory of self-reproducing automata. University of Illinois, Urbana
40. Voorhees B (1996) Computational analysis of one-dimensional cellular automata. World Scientific, Singapore
41. Wolfram S (1994) Cellular automata and complexity: Collected papers. Addison-Wesley, Reading
42. Wuensche A, Lesser M (1992) The global dynamics of cellular automata. Addison Wesley, Reading
43. Yang YX, Billings SA (2003) Identification of probabilistic cellular automata. IEEE Trans Syst Man and Cybern Part B 33:225–236
44. Yang YX, Billings SA (2003) Identification of the neighbourhood and CA rules from spatio-temporal CA patterns. IEEE Trans Syst Man Cybern Part B 30:332–339
45. Zhao Y, Billings SA (2006) Neighborhood detection using mutual information for the identification of cellular automata. IEEE Trans Syst Man Cybern Part B 36:473–479
46. Zhao Y, Billings SA (2007) The identification of cellular automata. J Cell Autom (in press)
47. Zhao Y, Billings SA, Routh A (2005) Identification of excitable media using cellular automata. Int J Bifur Chaos 17:153–168

Additional Reading

Adamatzky A (2001) Computation in nonlinear media and automata collectives. IoP Publishing, Bristol

Brauer W (1984) Automaton-theorie. Teubner, London

Chopard B, Droz M (1998) Cellular automata modeling of physical systems. Cambridge University Press, Cambridge

Crutchfield JP, Hanson JE (1999) Computational mechanics of cellular processes. Princeton University Press, Princeton

Narendra K, Thathachar MAL (1989) Learning automata. Prentice-Hall, New Jersey

Prokaev A, Sokolova L, Tarakanov A (2007) Using immunocomputing to forecast hydrophysical fields in the ocean. In: Int Workshop on Information Fusion and GIS (IF GIS 07), St. Petersburg, Russia (accepted for publication in LNCS)

Sipper M (1997) Evolution of parallel cellular machines. In: The cellular programming approach, lecture notes in computer science, vol 1194. Springer, Berlin

Tarakanov A, Adamatzky A (2002) Virtual clothing in hybrid cellular automata. Kybernetes (Int J Syst Cybernetics) 31:394–405

Tarakanov A, Goncharova L, Tarakanov O (2005) A cytokine formal immune network. Lecture Notes in Artificial Intelligence 3630:510–519

Tarakanov A, Prokaev A, Varnaskikh E (2007) Immunocomputing of hydroacoustic fields. Int J Unconv Comput (accepted for publication)

Tarakanov A, Skormin V, Sokolova S (2003) Immunocomputing: Principles and applications. Springer, New York

Toffoli T, Margolus N (1987) Cellular automata machines. A new environment for modeling. MIT Press, MIT Press Series in Scientific Computation, Cambridge

Weimar J (1998) Simulation with cellular automata. Logos, Berlin

Image Based State Estimation

NICHOLAS GANS, GUOQIANG HU, WARREN E. DIXON
University of Florida, Gainesville, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Pinhole Projection Model
Linear Estimation of Position and Orientation from Two Images
Extensions to the Linear Pose Estimation Algorithms
Estimation of Linear and Angular Velocity
Image Based Kalman Filter Estimation
Range Identification
Future Directions
Bibliography

Glossary

Pose The position and orientation of an object is referred to as the pose. The pose has a translation component that is an element of $\mathbb{R}^3$ (i. e., Euclidean-space), and the rotation component is an element of the special orthogonal group $SO(3) \subset \mathbb{R}^{3\times 3}$, though local mappings of rotation exist in $\mathbb{R}^3$ (e. g., Euler angles, angle/axis) or $\mathbb{R}^4$ (e. g., unit quaternions). Pose of an object has specific meaning when describing the relative position and orientation of one object to another or the position and orientation of an object at different time instances. Pose is typically used to describe the position and orientation of one reference frame with respect to another frame.

Camera-in-hand configuration The camera-in-hand configuration refers to the case when the camera is attached to a moving robotic system (e. g., held by the robot end-effector). In this case, camera pose is typically the state being estimated.

Camera-to-hand configuration The camera-to-hand configuration refers to the case when the camera is stationary and observing moving targets (e. g., a fixed camera observing a moving robot end-effector). In this case, pose of a viewed object is typically the state being estimated.

Feature point Different computer vision algorithms have been developed to search images for distinguishing features in an image (e. g., lines, points, corners, textures). The three-dimensional Euclidean-space representation of a point in the two-dimensional image plane is called a feature point. Feature points are selected based on contrast with surrounding pixels and the ability for the point to be tracked from frame to frame. Sharp corners (e. g., the windows of a building), or center of a homogenous region in the image (e. g., the center of car headlights) are good examples of feature points.

Calibration A camera has a set of parameters collectively known as calibration parameters that describe the interaction of optics and imaging surface to produce an image. The camera calibration parameters are composed of the intrinsic parameters (i. e., image center, camera scale factors, and camera magnification factor) and extrinsic parameters (i. e., relative camera pose with respect to a fixed reference frame). The process of learning the parameters for a particular camera is known as calibration. Using the calibration parameters to recover geometric data from raw image data is also referred to as calibration.

Epipolar constraint Given two images of a three-dimensional scene taken at two different poses, the Epipolar constraint restricts the positions of two-dimensional image points in both images. If the relative pose between the cameras is known and a point is identified in one image, the Epipolar constraint restricts the locations of the feature points in the second image. Conversely, if points are matched in two images, the Epipolar constraint restricts the possible relative camerea poses that could yield the two images and the relative three-dimenstional coordinates of the points in the scene. It is possible to codify the Epipolar constraint as a 3×3 matrix, known as the Essential Matrix. Given a sufficient number of points in two images, the Essential Matrix can be determined and decomposed to solve for the translation and rotation between the two camera poses. The Epipolar constraint exists even when image points have not been calibrated. in which it is codified in the Fundamental Matrix. It is possible to codify the epipolar constraint as a 3×3 matrix, known as the Essential Matrix. Given a sufficient number of points in two images, the Essential Matrix can be determined and decomposed to solve for the translation and rotation between the two camera poses. The Epipolar constraint exists even when image points have not been calibrated, in which it is codified in the Fundamental Matrix.

Homography The geometric concept of homography is a one-to-one and on-to transformation or mapping between two sets of points. In computer vision, homography generally refers to the mapping between points in two Euclidean-planes (Euclidean homogra-

phy), or to the mapping between points in two images (projective homography).

Definition of the Subject

Image-based state estimation (IBSE) is the use of image data to approximate a set of variables that determine the current conditions of a system. Image data commonly refers to visual images captured by an optical camera (such as the development in this chapter); however, images can also be synthesized from other sensors, such as SOund Navigation And Ranging (SONAR), RAdio Detection And Ranging (RADAR), LAser Detection And Ranging (LADAR) or cross-section information from Computed Tomography (CT) or Magnetic Resonance Imaging (MRI). For robotic systems that employ state-estimation, the states that are observed typically include the kinematics of a physical object such as position, orientation, and linear and angular velocity in the Euclidean-space.

The task of determining the pose and/or velocity of a moving camera observing a static scene is described as the "camera-in-hand" problem since early IBSE for robotics was achieved by a manipulator gripping the camera. The opposite task where the camera is stationary observing a moving target is called the "camera-to-hand" to denote the scenario where the camera is observing the motion of the robotic system from some fixed location.

Vision is arguably the primary environmental sensor used by human beings and many other animals. As members of a visual society, people can easily relate to information provided from a camera. Identification and interpretation of data is simple to convey in both directions of the man-machine interaction. From an engineering perspective, cameras are passive sensors (i. e., do not emit energy for sensing), and thus undetectable unlike active sensors modalities such as; SONAR, RADAR, and lasers. Passive sensing is important for surveillance and security tasks and environments where energy emitting sensors are hazardous. IBSE methods are important because they provide a means to determine information from images such as the six-dimensional pose and velocity of some identified and tracked object. Relative pose information, proximity to a target, and velocity of a target/camera are examples of common feedback information that robots require for autonomous operation.

Introduction

The use of images to determine the pose of a camera with respect to a viewed object can be traced back to at least 1913 when Kruppa [1] proved that two camera views of five Euclidean points could be used to estimate the translation and rotation separating the two camera views. However, the lack of sufficient computational resources has limited the growth of such IBSE until modern times. In recent literature, the eight-point algorithm was a landmark result introduced in [2] and [3] which uses the epipolar constraint (see Subsect. "The Essential Matrix and the Eight-Point Algorithm") to solve for camera motion and/or scene structure from eight or more feature points. Algebraically, the algorithm is an implicit function of two images of a Euclidean point and the rotation and translation between the camera views. Geometrically, the epipolar constraint restricts the set of Euclidean coordinates of a feature point between two camera views separated by a finite rotation and translation. The development of the epipolar constraint gave rise to a large body of work in reconstruction of motion and structure from multiple camera views using linear algorithms. Specifically, advances in computer vision research later reduced the number of requisite non-coplanar feature points to as few as five feature points [4,5]; however, the five-points algorithm requires solving the roots of a tenth degree polynomial and results in additional valid solutions. The planar homography (see Subsect. "The Planar Homography Algorithm") algorithm [6], which solves for motion and structure from multiple views of four or more feature points that lie in the same 3D plane, is also based on outcomes of the epipolar constraint. The linear algorithms in works such as [2,3,6] provide the geometric backbone of the IBSE methods presented in this chapter. These linear methods are attractive due to their relative simplicity and mathematical tractability. These methods are easily implemented as computer programs with limited resources such as embedded systems, and can operate much faster than thirty Hertz (considered "real time" for video) with modern desktop or laptop resources. The linear algorithms that relate multiple images also require no a priori information about the scene.

The previously described linear geometric relationships are widely used in literature and have been successfully demonstrated in a number of applications. Yet, these methods have some limitations. Like any linear technique, the approximation can give erroneous results in the presence of nonlinearities such as optical effects, normalization, and noise. Thus, in the 1990s some methods began to develop as a means to compensate for these effects using nonlinear methods. These results use linear techniques for an initial estimate of the relative geometry followed by an optimization step through gradient descent techniques (e. g., [7,8,9]), typically referred to as bundle adjustment. Accommodating for such disturbances in the geometric transformation is an important area of research, and the

reader should be aware of these methods. However, this chapter will not discuss them in detail.

Even under the assumption that the linear geometric relationships are sufficient, a significant issue that impacts IBSE is that to maintain an estimate of the motion of an object (or of the camera), the feature points must remain in the field-of-view (FOV). Motivated by this issue, researchers have investigated various methods to expand the FOV. Techniques to indefinitely relate multiple pose estimates from multiple, overlapping views, have been explored in [10,11]. Both of these papers are motivated by IBSE and control of unmanned aerial vehicles. Subsect. "Long Term Pose Estimation Through Chained Homography Decomposition" provides an example technique that can be used for camera pose estimation over large scenes where features leave the FOV. Cameras with curved lenses or curved mirrors can deliver a 360° (albeit distorted) view. Several researches have developed nonlinear observers for state estimation with such omnidirectional cameras (e. g., [12,13]).

Another challenge for IBSE is that the image-plane geometry only provides a scaled pose of a camera or object in the current image relative to a pose at reference goal image. An additional set of linear equations can deliver the depths of feature points up to an unknown scale factor. However, the unknown scaling can be a significant obstacle for autonomous robotic tasks. Subsect. "Solving for Relative Object Poses Through Knowledge of a Single Length" presents an example method that exploits some additional knowledge, such as a single known length on an object, to attach a reference frame to the object and deliver relative pose of the object with respect to the camera.

In the late 1980s, researchers began applying optimal filter techniques, namely the Kalman Filter (KF) and Extended Kalman Filter (EKF) for IBSE. Some approaches revolve around the Essential Matrix and use the EKF to refine the estimate of the Essential Matrix to determine the pose and/or velocity [14]. Other methods eschew external pose estimation schemes and use image features as inputs to the KF/EKF with pose and/or velocity as an output [15,16,17,18]. KF approaches to IBSE are discussed in Sect. "Image Based Kalman Filter Estimation".

In the 1990s and 2000s nonlinear estimators/observers were investigated as IBSE methods. Similar to the KF developments, some methods use linear pose estimation techniques to estimate velocity or pose of an object [19], while others develop estimation approaches without using the linear methods [20,21,22,23,24]. In Subsect. "Velocity Estimation Through a Nonlinear Observer" a nonlinear estimator based on the planar Homography Matrix is provided as an example that estimates velocity of a moving object given geometric knowledge of the shape of the object. In Subsect. "Nonlinear Estimator for Range Identification", an example of a nonlinear method to estimate the range (i. e., depth) to feature points is provided that exploits the additional information that the camera/object velocity is known. Given knowledge of the camera/target motion, the example Subsect. "Nonlinear Estimator for Range Identification" is developed without linear pose estimation methods.

This chapter serves as an introduction and brief survey to the field of IBSE. Several books have been written which can be recommended to provide in-depth treatment of these subjects. The bedrock methods based in epipolar geometry are covered in great detail in [25,26,27,28]. Further discussion on Kalman Filtering for IBSE is provided in [26].

Pinhole Projection Model

There are several approaches to model the projection of an object's image through a lens and imaging surface. A simple, widely used method for accurately modeling a well focused imaging system is the pinhole camera model (i. e., a perspective projection). For the pinhole camera model, a Euclidean reference frame is attached to the camera at the center of projection. The imaging surface is a plane perpendicular to the z-axis of the camera frame, and located a distance f from the center of projection. The distance f is the focal length, and is a physical characteristic of the camera and lens. The pinhole camera model is illustrated in Fig. 1.

The coordinates of a feature point in the Euclidean-space with respect to the camera frame are denoted as

$$\bar{m}(t) = [x(t), y(t), z(t)]^{\mathrm{T}} .$$

The point $\bar{m}(t)$ projects to a point $m(t)$ on the image plane, with normalized coordinates

$$m(t) = \left[\frac{x(t)}{z(t)}, \frac{y(t)}{z(t)}, 1\right]^{\mathrm{T}}$$

in the camera frame. This projection can be described as a nonlinear projection function $\Pi(\cdot)$ as

$$m = \Pi(\bar{m}) .$$

In a digital camera, the normalized coordinates of points are measured in pixel coordinates as

$$p(t) = [u(t), v(t), 1]^{\mathrm{T}} ,$$

where $u(t)$ and $v(t)$ are integers that are typically measured from a corner of the image. The relationship between the

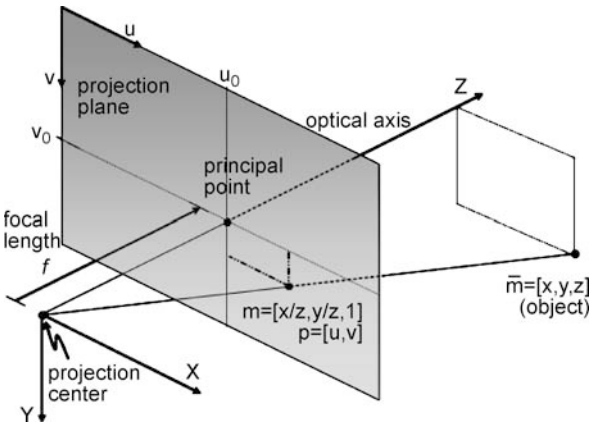

Image Based State Estimation, Figure 1
Pinhole projection model

Image Based State Estimation, Figure 2
Feature points from corner detection

Image Based State Estimation, Figure 3
Feature points from the critical points of a contour

jth image-based feature point and the corresponding Euclidean point is given as

$$p_j(t) = Am_j(t)\,, \tag{1}$$

where $A \in \mathbb{R}^{3\times 3}$ is a constant, invertible, upper-triangular camera calibration matrix given by [28]

$$A = \begin{bmatrix} f\sigma_x & f\sigma_x \cot\alpha & u_0 \\ 0 & f\sigma_y \sec\alpha & v_0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{2}$$

In (2), f is the focal length, σ_x, σ_y are the pixel size in the x and y directions of the camera frame (also the u and v directions in the image plane), α is the skew angle, which defines how rectilinear the pixels are, and u_0 and v_0 define the center of projection in pixels. Together, f, σ_x, σ_y, α, u_0, v_0 are the intrinsic calibration parameters, and can be determined through camera calibration procedures [29,30,31,32,33]. The methods presented in this chapter are based on the assumption that the camera is calibrated. Given knowledge of A, it is possible to recover the normalized coordinates of a feature point from the pixel coordinates. More complicated geometric primitives such as lines, curves and surfaces are composed of infinite sets of points $\bar{M} = \{\bar{m}_j\}$, where $\bar{m}_j(t)$ is every point in the primitive. Likewise a line, curve, or surface is described in the image by a set $M = \{m_j\}$, where $m_j = \Omega(\bar{m}_j)$ is every point in the two-dimensional primitive.

The development in this chapter will also focus on the use of points as image features for estimation. Algorithms exist for lines, curves, surfaces, etc., but points are simple to extract from an image, track between images, and to mathematically represent. Furthermore, many methods that use other primitives are simply alternative ways to solve for the structures such as the Essential or Homography Matrices, which can also be determined by just using points. Extraction of feature points in images is outside the scope of this chapter, but there are many resources available. Points in an image are often found as contrasting areas such as corners (e. g., [34,35]), as an averaged area such as the centroid of a region, as local extrema [36], as critical points of a contour, or at the intersection of lines. Figures 2–4 show examples of corner extraction from corner detection, critical points of contours and intersection of lines. There are many methods of contour extraction (cf. [37,38]) and line/curve extraction (cf. [34,39,40]).

Linear Estimation of Position and Orientation from Two Images

While techniques for estimating camera pose and scene structure from a series of images have been developed

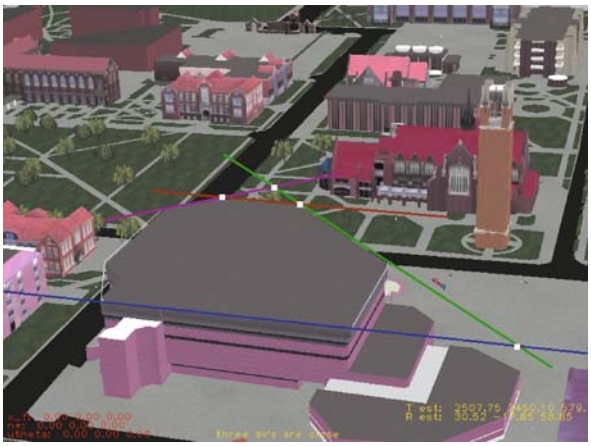

Image Based State Estimation, Figure 4
Feature points from the intersection of lines

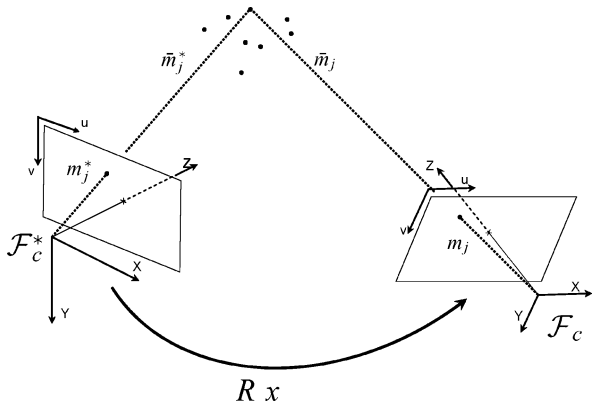

Image Based State Estimation, Figure 5
Projection of a point onto two image planes, and the resulting epipole

since the early 20th century, computational resources limited the advancement and usage of these methods until the early 1980s. The first techniques developed were linear methods, in the sense that they rely solely on linear algebra. The first of these methods involved the epipolar constraint, and is commonly referred to as the Essential Matrix or the eight-point algorithm, due to the fact that it requires at least eight points in the scene (e. g., [2,3]). A similar method was developed for planar scenes [6], and is often referred to as the Homography Matrix.

The Essential Matrix and the Eight-Point Algorithm

Consider a camera that is viewing a collection of N feature points with Euclidean coordinates denoted by

$$\bar{m}_j^* = [x_j^*, y_j^*, z_j^*]^{\mathrm{T}}\,, \quad \forall j \in \{1 \ldots N\} \tag{3}$$

expressed in the camera reference frame, denoted by $\mathcal{F}_c^*$. An image of the points is captured, resulting in a projection to a set of points in the image-plane denoted by the normalized coordinates

$$m_j^* = \left[\frac{x_j^*}{z_j^*}, \frac{y_j^*}{z_j^*}, 1\right]^{\mathrm{T}}\,, \quad \forall j \in \{1 \ldots N\}\,. \tag{4}$$

By translating the camera by $x(t)$ and rotating the camera by $R(t)$, the camera will obtain a new pose denoted by the reference frame $\mathcal{F}_c$, as illustrated in Fig. 5. The Euclidean and normalized coordinates of the collection of feature points expressed in $\mathcal{F}_c$ are defined as

$$\bar{m}_j(t) = \left[x_j(t), y_j(t), z_j(t)\right]^{\mathrm{T}}\,, \quad \forall j \in \{1 \ldots N\} \tag{5}$$

$$m_j(t) = \left[\frac{x_j(t)}{z_j(t)}, \frac{y_j(t)}{z_j(t)}, 1\right]^{\mathrm{T}}\,, \quad \forall j \in \{1 \ldots N\}\,. \tag{6}$$

The Euclidean coordinates of the feature points expressed in $\mathcal{F}_c^*$ can be related to the coordinates expressed in $\mathcal{F}_c$ through the following algebraic relationship:

$$z_j(t)m_j(t) = z_j^* R(t) m_j^* + x(t)\,, \quad \forall j \in \{1 \ldots N\}\,, \tag{7}$$

where the relationships in (3)–(7) have been used. Multiplying both sides of (7) by the skew symmetric form of $x(t)$, denoted by $[x(t)]_\times \in \mathbb{R}^{3\times 3}$, yields

$$z_j(t)[x(t)]_\times m_j(t) = z_j^*[x(t)]_\times R(t) m_j^*\,, \quad \forall j \in \{1 \ldots N\}\,, \tag{8}$$

after using the property that $[x(t)]_\times x(t)$ is equal to a vector of zeros. After multiplying both sides of (8) by $m_j^{\mathrm{T}}(t)$ and exploiting the property that

$$z_j(t) m_j^{\mathrm{T}}(t)[x(t)]_\times m_j(t) = 0$$

the essential constraint or epipolar constraint is given as

$$m_j^{\mathrm{T}}(t) E(t) m_j^*(t) = 0 \tag{9}$$

where $E(t) \in \mathbb{R}^{3\times 3}$ is known as the Essential Matrix, and is defined as

$$E(t) = [x(t)]_\times R(t)\,. \tag{10}$$

Eight or more noncoplanar feature points are required to create a set of linear equations to solve for $E(t)$. The Essential Matrix can be decomposed to recover $R(t)$ and $\lambda x(t)$, where λ is an unknown scale factor [2,3]. This decomposition only involves the use of linear algebra techniques including singular value decomposition.

The inability to recover the scale of $x(t)$ is a fundamental problem in IBSE. The inherent scale factor uncer-

tainty is due to the loss of depth information when three-dimensional Euclidean coordinates are normalized (i. e., projected to the two-dimensional image plane). Mathematically, the scale factor uncertainty is evident from the fact that there is no unique solution to (9). Furthermore, there are two possible rotations and two possible translations that satisfy (10) for a particular $E(t)$, giving four possible solutions to the decomposition, of which only one solution is physically valid. Thus, given some physical knowledge of the system, the correct solution can be selected. For further detail, see [2,3].

The epipolar constraint in (9) relates two planes of normalized Euclidean coordinates. If the calibration matrix A is unknown, the epipolar constraint can be applied to points expressed in pixel coordinates as

$$p_j(t) = F(t)p_j^* \,. \tag{11}$$

In (11) $F(t) \in \mathbb{R}^{3\times 3}$ is known as the Fundamental Matrix, and has important functions in fields of computer vision. However, the Essential Matrix must be recovered for meaningful pose reconstruction.

The Planar Homography Algorithm

Consider a camera with reference frame $\mathcal{F}_c^*$. The camera views a collection of $N \geq 4$ or more feature points lying in a plane π in front of the camera, as illustrated in Fig. 6. These points have Euclidean coordinates and normalized Euclidean coordinates defined as in (3) and (4). The plane π has a normal vector $-n^*$ with respect to $\mathcal{F}_c^*$. By translating the camera by $x(t)$ and rotating the camera by $R(t)$, the camera will obtain a new pose denoted by the reference frame $\mathcal{F}_c$. The Euclidean and normalized coordinates of the collection of feature points expressed in $\mathcal{F}_c$ are defined

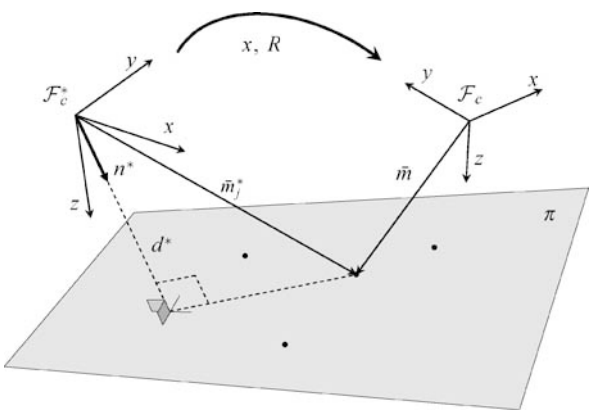

Image Based State Estimation, Figure 6
Projection of coplanar points onto two image planes

as in (5) and (6). The Euclidean coordinates of the points in two views are related by the relationship

$$\bar{m}_j(t) = R(t)\bar{m}_j^* + x(t)\,, \quad \forall j \in \{1 \ldots N\}\,. \tag{12}$$

The projective relationship

$$d^* = n^{*T}\bar{m}_j^*\,, \quad \forall j \in \{1 \ldots N\}\,, \tag{13}$$

where $d^* \in \mathbb{R}$ is the distance from the origin of $\mathcal{F}_c^*$ to the plane, can be used to rewrite (12) in terms of the normalized Euclidean coordinates as [6,41]

$$m_j(t) = \alpha_j(t)H(t)m_j^*\,. \tag{14}$$

In (14), $\alpha_j(t)$ is a time-varying ratio of the depth coordinates defined as

$$\alpha_j(t) = \frac{z_j^*}{z_j(t)}\,, \quad \forall j \in \{1 \ldots N\}\,, \tag{15}$$

and $H(t) \in \mathbb{R}^{3\times 3}$ denotes the Euclidean homography defined as

$$H(t) = R(t) + \left(\frac{x(t)}{d^*}\right) n^{*T}\,. \tag{16}$$

The matrix $H(t)$ in (16) is known by many names such as the Planar Essential Matrix, the Planar Homography Matrix, and the Euclidean Homography Matrix. In the case of digital imaging, the calibration matrix (1) is used with the Euclidean relationship in (14) to give

$$\begin{aligned} p_j &= \alpha_j AHA^{-1}p_j^* \\ &= \alpha_j G p_j^*\,. \end{aligned} \tag{17}$$

Given given a set of $N \geq 4$ image points (i. e., p_j, $j \in \{1 \ldots N\}$) matched in two images and knowledge of A, a set of linear equations can be solved to acquire $G(t)$ and recover $H(t)$. Linear decomposition methods (e. g., [6,41]) can then be applied to recover, $R(t)$, $x(t)/d^*$, n^* and $\alpha_j(t)$.

In the general case, linear decomposition methods will return four mathematically valid solutions. Two solutions are physically invalid, as $n^{*T}\bar{m}_j^* = d^* < 0$, which would place the points behind the camera; hence, there are only two physically valid solutions for the displacement. Scene knowledge, such as the expected value of n^* could be used to choose the correct solution. If such knowledge is not available, a small, arbitrary translation can be performed with respect to the frame $\mathcal{F}_c^*$, and a second Homography Matrix can be computed and decomposed. Since both homographies give a solution with respect to the constant

frame $\mathcal{F}_c^*$, one solution from each must have the same n^*. The solution with agreeing n^* is the correct solution. This method of comparing the normals is referred to as temporal coherence of the homography solution.

The translation $x(t)$ is recovered from the homography decomposition only up to the scale factor d^*, which is typically not known and may not be recovered from images alone. If $d^*(t)$ can be measured with a separate sensor, the translation $x(t)$ can be recovered. Depending on the task, additional information can be used to recover d^* and $x(t)$. For example, the camera is translated a known distance x' from $\mathcal{F}_c$ to $\mathcal{F}_c'$, and the planar homography algorithm is used to solve for the rotation R'' and scaled translation x''/d^* from $\mathcal{F}_c^*$ to $\mathcal{F}_c'$. Since $\mathcal{F}_c$ and $\mathcal{F}_c'$ are separated by pure translation, $R'' = R$, and the known scaled translations x''/d^* and x/d^* and the known translation x' are related by

$$d^* \frac{x''}{d^*} = d^* \frac{x}{d^*} + x' \,. \tag{18}$$

The constant d^* can then be determined as

$$d^* = \frac{\left(\frac{x''}{d^*} - \frac{x}{d^*}\right)^{\mathrm{T}} x'}{\left\| \left(\frac{x''}{d^*} - \frac{x}{d^*}\right) \right\|^2} \,. \tag{19}$$

The planar homography algorithm is sufficient to give orientation and (scaled) translation from the current camera frame $\mathcal{F}_c$ to a reference frame $\mathcal{F}_c^*$. The alternate case where the camera is fixed, and the planar object moves from a reference pose $\mathcal{F}_s^*$ to a new pose $\mathcal{F}_s(t)$, can also be considered in a similar framework. By solving and decomposing the Homography Matrix $H(t)$, the rotation $R_s(t)$ and scaled translation $x_s(t)/d^*$ of the object can be recovered from the perceived camera rotation $R(t)$ and translation $x(t)/d^*$ as

$$R_s(t) = R^{\mathrm{T}}(t) \tag{20}$$

$$\frac{x_s(t)}{d^*} = -R^{\mathrm{T}}(t) \frac{x(t)}{d^*} \,. \tag{21}$$

If the camera can be moved precisely while the viewed object is known to be stationary (or vice versa), then d^* can be recovered as in (18). The distance d^* can not be recovered if both the object and camera are moving with an unknown motion.

Characteristics of the Essential Matrix Algorithm and Planar Homography Matrix

The pose estimation methods that are discussed in this chapter are examples of different approaches, and each approach has different advantages and disadvantages. The characteristics of the Essential Matrix algorithm and planar homography algorithm are compared in this section to distinguish the strengths and weaknesses. Some characteristics are shared because the Essential and Planar Homography Matrices are formally related. Comparing (10) and (16) indicates that

$$E(t) = [x(t)]_\times H(t) \,. \tag{22}$$

In the general case that $x(t)$ is known only to a scale factor λ, $E(t)$ is proportional to $[x(t)]_\times H(t)$. Given $H(t)$ from four coplanar points and two additional points outside of the plane, it is possible to recover $E(t)$ [26]. Similarly, given $E(t)$ from eight non-coplanar points, it is possible to solve for the $H(t)$ matrix corresponding to any three of the eight points and their corresponding plane (see [26,42]).

A notable difference between the Essential Matrix and the Homography Matrix is that at least eight points are required to solve for the Essential Matrix using linear methods, while at least four points are necessary to solve for the Homography Matrix. The target/landmark/scene involved in the state estimation will determine how many points are available to track, and will determine which methods may be used. The Essential Matrix can not be solved if the Euclidean points $\bar{m}_j(t)$ lie in the same plane. Specifically, there is a numerical dependence on the system of equations used to solve $E(t)$, and the Essential Matrix can not be used in the case that the eight points are coplanar.

Similarly, the points used to solve for the Homography Matrix in (16) must be coplanar. Besides using linear algebra techniques to solve multiple equations, there are alternative methods to solve the Homography Matrix that may be used in cases that feature points are not coplanar. It is possible to use the Essential Matrix to recover the Homography for a subset of three points. A similar method, coined the Virtual Parallax [43], skips the initial computation of the Essential Matrix. Given eight points, three points define a plane π_s, and the remaining five or more points are projected onto π_s, delivering five virtual points. These points can be used to deliver the Homography Matrix where n^* and d^* are defined with respect to π_s. Solving $H(t)$ from $E(t)$, or using the Virtual Parallax extends the ability to use the Homography Matrix for general scenes at the cost of requiring more points, slightly increased computation cost, and possibly increased noise sensitivity.

From (10), if $x(t) = 0$, then $E(t) = 0$, even for non-zero rotation. Thus, if there is no translation, rotation cannot be recovered from the Essential Matrix. Generally, as the amount of translation becomes closer to zero, the estimate of $R(t)$ will incur greater error. For some applications where the camera or target is always translating, the Essential Matrix may be useful, but for many autonomous

robotic applications, the Essential Matrix will be insufficient and the Homography Matrix may be preferable.

A third difference is the additional information delivered from the Homography Matrix. While both methods can estimate translation $x(t)$ and rotation $R(t)$, the Homography gives the normal vector n^* which is used in Subsect. "Estimating the Pose of a Planar Object Relative to the Camera" to estimate the pose of an object. Homography decomposition also gives the ratios $\alpha_j(t)$, $\forall j \in \{1 \ldots N\}$. This ratio can be used as an error signal for control (e. g., [44,45]) or IBSE, as shown in Subsect. "Velocity Estimation Through a Nonlinear Observer".

Extensions to the Linear Pose Estimation Algorithms

The methods discussed in Subsects. "The Essential Matrix and the Eight-Point Algorithm" and "The Planar Homography Algorithm" give a foundation for pose estimation. However, as presented the methods have limitations. These methods can only estimate translation up to a scale factor, are limited to estimating camera or object pose relative to an initial pose corresponding to a fixed camera view, and are constrained to limited motions such that the viewed object remains in the FOV. This section provides examples that can be used to extend the linear pose estimation methods to overcome these deficiencies. Often these additional methods will require some additional information or sensor, such as a known length on the viewed object, a known distance to the object, or accurate odometry for the camera frame.

Long Term Pose Estimation Through Chained Homography Decomposition

The methods discussed in Subsects. "The Essential Matrix and the Eight-Point Algorithm" and "The Planar Homography Algorithm" allow for the estimation of camera motion provided the feature points used in the Essential or Homography Matrices remain in the FOV. Many tasks will require state estimation continue as objects leave and enter the field of view. Consider an unmanned air vehicle (UAV) equipped with a GPS and a camera capable of viewing a landscape. The Essential or Homography Matrix can be used to recover the motion of the plane, but the limited camera FOV and motion of the vehicle can cause observed feature points to leave the image. Motivated by the FOV issue, recent research (e. g., [10,11]) has been developed to allow a transition from estimating the camera pose from previously viewed feature points to estimating the camera pose from incoming feature points. For example, the work in [10] assumes coplanarity of all points and is focused on building and maintaining a mosaic of images. The pose estimation efforts in [11] do not assume coplanarity of all sets of feature points and uses a homography relationship to "daisy-chain" relative pose estimates together. The approach in [11] provides a mechanism to work around the FOV obstacle by indefinitely chaining together a series of pose estimates.

To further illustrate the daisy-chaining method provided in [11], consider an inertial world frame $\mathcal{F}_w$, and a frame attached to some vehicle, $\mathcal{F}_c$. Since the methods discussed in Sect. "Linear Estimation of Position and Orientation from Two Images" provide means to determine camera motion relative to a fixed camera frame, the following example is based on the assumption that the vehicle pose is known through some inertial sensor (e. g., a global positioning system (GPS)) and other sensors at some initial time t_0. That is, the translation and rotation, $x_0(t_0)$ and $R_0(t_0)$, between $\mathcal{F}_w$ and $\mathcal{F}_c(t_0)$ is known. Figure 7 illustrates this scenario with a UAV. At time t_0, the aircraft enters a GPS denied environment and relies solely on on-board sensors including a camera. Without loss of generality, the GPS unit is assumed to be fixed to the origin of the aerial vehicle's coordinate frame, and the constant pose of the camera frame is known with respect to the pose of the UAV coordinate frame. The subsequent development further assumes that the GPS is capable of delivering altitude, perhaps in conjunction with an altimeter, so that the altitude above the ground $a(t_0)$ is known, i. e., the scalar distance to the ground in the direction of gravity (see Fig. 8).

As illustrated in Fig. 7, the initial set of tracked coplanar and noncolinear feature points, denoted $p_a(t)$, are projections from the set of Euclidean points, $m_a(t)$, contained in the plane π_a. These feature points have Euclidean coordinates $\bar{m}_{aj}(t_0) \in \mathbb{R}^3 \; \forall j \in \{1 \ldots N\}$ in $\mathcal{F}_c(t_0)$. The plane π_a is perpendicular to the unit vector $n_a(t_0)$ in the camera frame, and lies at a distance $d_a(t_0)$ from the camera frame origin. At time t_1, the vehicle has some rotation $R_{01}(t_1)$ and translation $x_{01}(t_1)$ that can be determined from the images by decomposing the relationships given

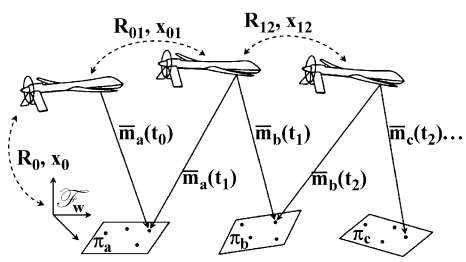

Image Based State Estimation, Figure 7
Illustration of pose estimation chaining

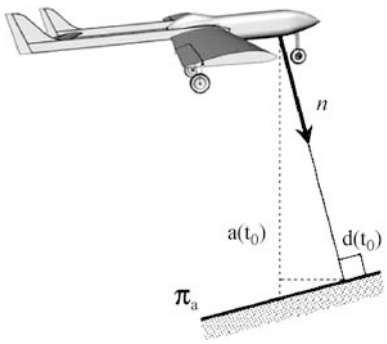

Image Based State Estimation, Figure 8
Depth estimation from altitude

in (17). For notational simplicity, the subscript j is omitted in subsequent development.

As described in Subsect. "The Planar Homography Algorithm", $R_{01}(t_1)$ and $x_{01}(t_1)/d_a(t_0)$ can be determined from two corresponding images of the feature points $p_a(t_0)$ and $p_a(t_1)$. A measurement or estimate for $d_a(t_0)$ is required to recover $x_{01}(t_1)$. This estimation is possible with distance sensors or with a priori knowledge of the geometric distances between the points in π_a. However, with an additional assumption that the plane is above the plane π_a, it is possible to estimate $d_a(t_0)$ geometrically using altitude information from the last GPS reading, range finder and/or an altimeter. From the illustration in Fig. 8, if $a(t_0)$ is the height above π_a (e. g., the slope of the ground is constant between the feature points and projection of the plane's location to the ground), then the distance $d_a(t_0)$ can be determined as

$$d_a(t_0) = n_a(t_0) \cdot a(t_0) \,. \tag{23}$$

where $n_a(t_0)$ is known from the homography decomposition.

Once $R_{01}(t_1), d_a(t_0)$, and $x_{01}(t_1)$ have been determined, the rotation $R_1(t_1)$ and translation $x_1(t_1)$ can be determined with respect to $\mathcal{F}_\mathrm{w}$ as

$$\begin{aligned} R_1 &= R_0 R_{01} \\ x_1 &= R_{01} x_{01} + x_0 \,. \end{aligned}$$

As illustrated in Fig. 7, a new collection of feature points $p_b(t)$ can be obtained that correspond to a collection of points on a planar patch denoted by π_b. At time t_2, the two images of the sets of feature points on π_b taken at times t_1 and t_2 can be used to determine $R_{12}(t_2)$ and $x_{12}(t_2)/d_b(t_1)$, which provides the rotation and scaled translation of $\mathcal{F}_\mathrm{c}$ with respect to $\mathcal{F}_\mathrm{w}$. If π_b and π_a are the same plane, then $d_b(t_1)$ can be determined as

$$d_b(t_1) = d_a(t_1) = d_a(t_0) + x_{01}(t_1) \cdot n(t_0) \,. \tag{24}$$

When π_b and π_a are the same plane $x_{12}(t_2)$ can be correctly scaled, and $R_2(t_2)$ and $x_2(t_2)$ can be computed in a similar manner as described for $R_1(t_1)$ and $x_1(t_1)$. The pose estimates can be propagated by chaining them together at each time instance without further use of GPS.

In the general case, sets of feature points on π_b and π_a are not coplanar, and (24) cannot be used to determine $d_b(t_1)$. If the sets of feature points on π_b and π_a are both visible for two or more frames, it is still possible to calculate $d_b(t)$ through geometric means. Let t_1- denote as some time before the daisy chain operation is performed, when sets of feature points on π_b and π_a are visible in the image. At time t_1-, an additional set of homography equations can be determined for the points on π_b and π_a at times t_1 and t_1-

$$m_{ai}(t_1) = \alpha_a \left(R + \frac{x n_a(t_1-)^\mathrm{T}}{d_a(t_1-)} \right) m_{ai}(t_1-) \tag{25}$$

$$m_{bi}(t_1) = \alpha_b \left(R + \frac{x n_b(t_1-)^\mathrm{T}}{d_b(t_1-)} \right) m_{bi}(t_1-) \tag{26}$$

where

$$\alpha_a = \frac{z_{ai}(t_1-)}{z_{ai}(t_1)} \quad \text{and} \quad \alpha_b = \frac{z_{ai}(t_1-)}{z_{ai}(t_1)} \,.$$

Note that R and x have the same values in Eqs. (25) and (26), but the distance and normal to the plane are different for the two sets of points, therefore the decomposition methods will give different scaled translations

$$x_a(t_1) = \frac{x(t_1)}{d_a(t_1-)} \quad \text{and} \quad x_b(t_1) = \frac{x(t_1)}{d_b(t_1-)} \,.$$

The distance $d_a(t_1-)$ can be found using (24). The translation $x(t_1)$ is solved as

$$x(t_1) = d_a(t_1-) x_a(t_1) \,.$$

The distance $d_b(t_1-)$ can then be determined as

$$d_b(t_1-) = \frac{x_b^\mathrm{T} x}{\|x_b\|^2} \,.$$

The distance $d_b(t_1)$ can then be found by using (24) with $d_b(t_1-)$ in place of $d_a(t_0)$. Additional sensors, such as an altimeter, can provide an additional estimate in the change in altitude. These estimates can be used in conjunction with (24) to update depth estimates.

Simulations To provide a practical explanation of the daisy-chaining method this section provides simulation results for an autonomous UAV flying into a zone with

denied GPS availability. For the simulation, five patches of four feature points are evenly placed 100 m apart in a straight line. For simplicity, all planar patches of feature points lie in the same plane. The objective is to perform IBSE during a maneuver of a 10 m lateral shift to the right and a 10 m longitudinal increase in altitude. This particular maneuver results in the vehicle simultaneously pitching, rolling, and yawing, while translating. For simulation purposes, the camera is considered to be mounted underneath the fuselage looking downwards. The camera model for this exercise is intended to be representative of a typical 640 × 480 lines of resolution CCD camera equipped with a 10 mm lens. To more accurately capture true system performance, pixel coordinates were rounded to the nearest integer to model errors due to camera pixilation effects (i. e., quantization noise). Furthermore, a five percent error was added to the estimated vehicle altitude to examine the robustness.

The first simulation was designed to test the accuracy of the vision-based estimation. Vision was not used in the

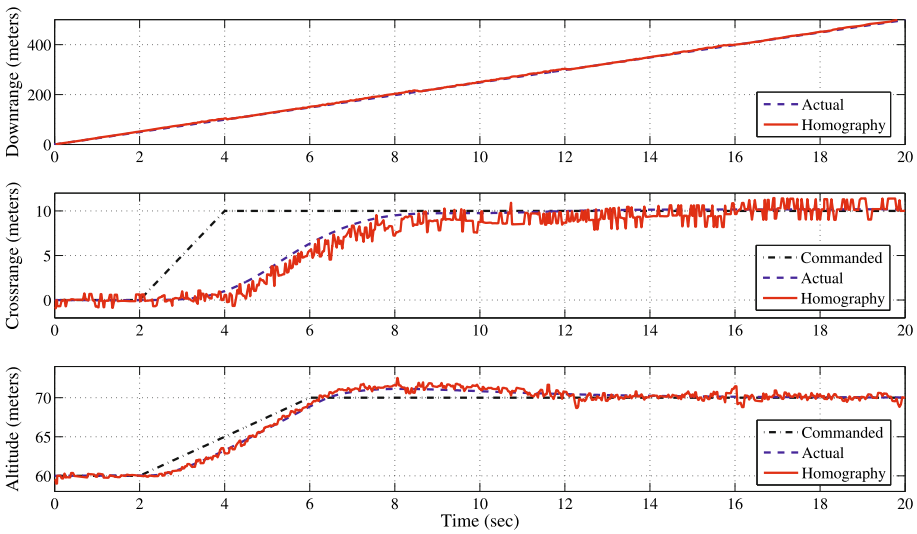

Image Based State Estimation, Figure 9
Simulation results of actual position versus estimated position when the UAV is using GPS for control

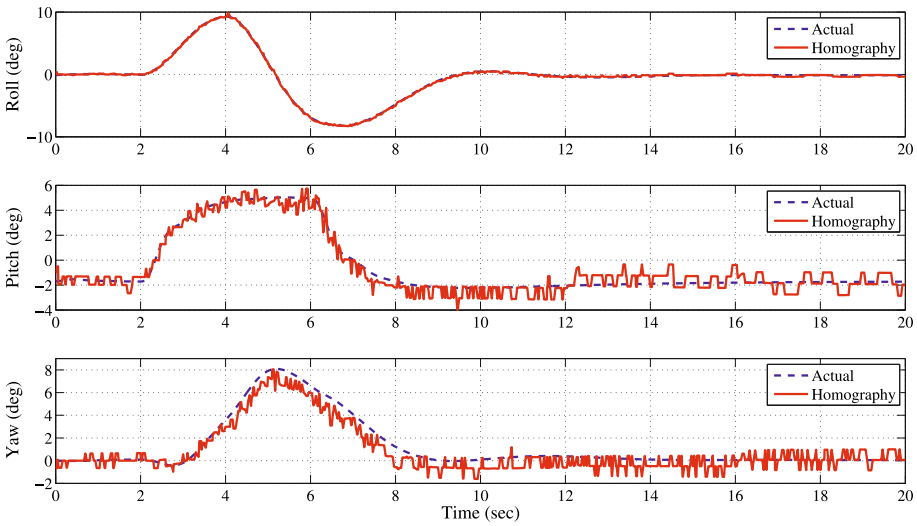

Image Based State Estimation, Figure 10
Simulation results of actual attitude versus estimated attitude when the UAV is using GPS for control

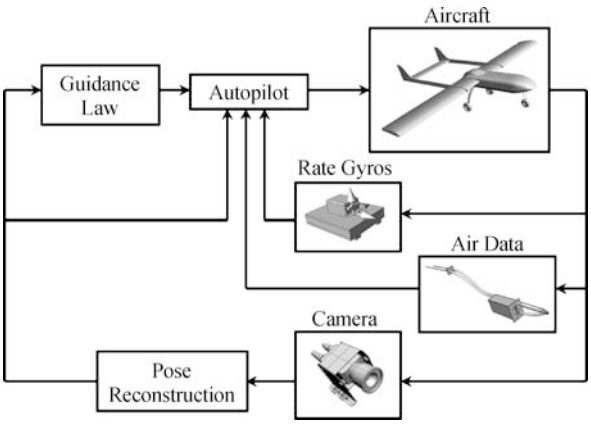

Image Based State Estimation, Figure 11
Block diagram control system architecture

feedback in this maneuver, and the estimated pose is compared to the true pose. The results of this preliminary analysis are given in Figs. 9 and 10. The effects of noise are visible but the estimated pose is accurate.

The second simulation illustrates the effects of using the IBSE as a sensor in closed-loop control. This simulation involved replacing the perfect pose measurements used in the guidance system and autopilot with the results from the IBSE. The resulting control architecture and sensor suite for this UAV is given in Fig. 11. The noise content of the estimated pose required filtering prior to being used by the autopilot to prevent the high frequency noise from being passed to the aircraft actuators. As expected, the noise occurs at 30 Hz and corresponds to the frame rate of the camera. First-order, low-pass filters (cutoff frequency as low as 4 rad/s) were used to filter the noise. The noise also prevented effective differentiation of the position and attitude and necessitated the use of rate gyros for yaw and roll damping, as depicted in Fig. 11. The air data system is also included, as shown in Fig. 11, for the initial altitude measurement, since it is more accurate for altitude than current GPS solutions. The results of the camera-in-the-loop system performing the same autonomous maneuver are given in Figs. 12 and 13.

The simulation results indicate that a camera supplemented with minimal sensors such as rate gyros and barometric altitude can be used for completely autonomous flight of a fixed wing vehicle; however, some residual oscillation effects due to noise is present in the vehicle attitude response. A majority of the noise source can directly be attributed to camera pixilation effects and the corresponding phase lag introduced by the first order filtering.

Experiments Based on the results of the simulation, a flight test experiment was conducted to establish the feasibility of the proposed IBSE method. Artificial features were placed along a stretch of a runway. A radio controlled aircraft with an onboard camera was flown over the runway. The video was overlaid with GPS data. An example of a single frame of this video is given in Fig. 14. A second GPS unit was also onboard to test inter-GPS accuracy. The use of two GPS units provides a comparison for the IBSE method, which is intended to compute GPS-like information. Video data was captured using a DV tape recorder

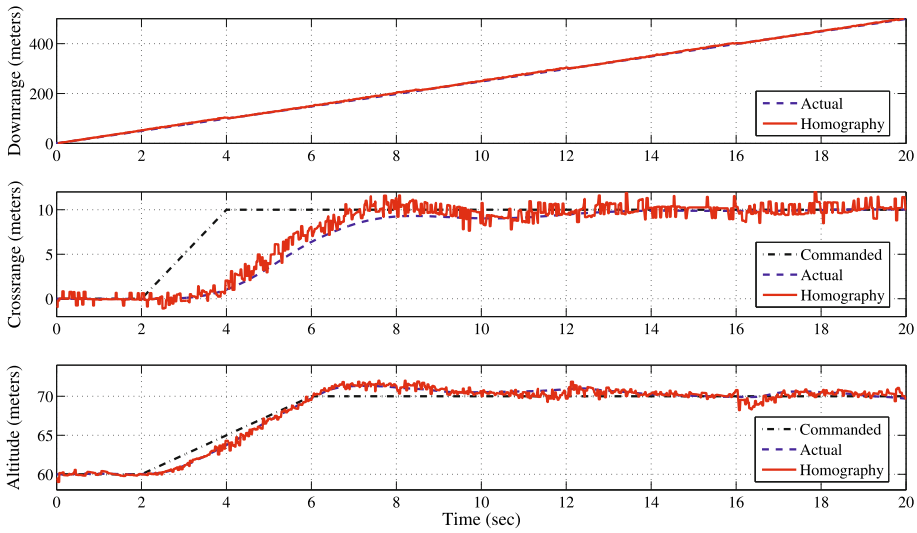

Image Based State Estimation, Figure 12
Simulation results of actual position versus estimated position when the UAV control-loop uses estimated pose information

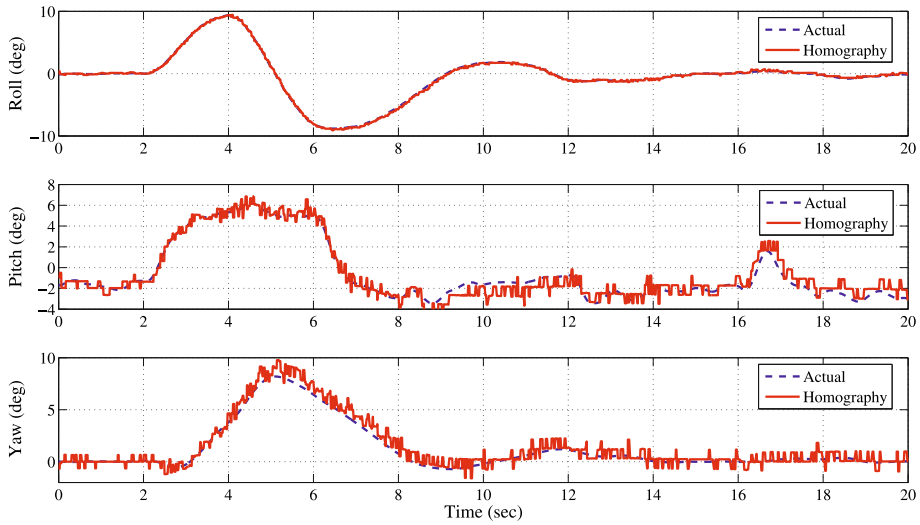

Image Based State Estimation, Figure 13
Simulation results of actual attitude versus estimated attitude when the UAV control-loop uses estimated pose information

Image Based State Estimation, Figure 14
Single video frame with GPS overlay illustrating landmarks placed along inside edge of the runway

and analyzed offline. Due to poor image quality, including focus, motion blur and interlacing of the DV video, it became necessary to extract features by hand from individual frames for this particular example. Features were extracted every sixth frame, resulting in a 5 Hz input signal.

Results of the experiment are given in Fig. 15. In the legend for Fig. 15, GPS2 represents the overlaid GPS data, and GPS1 represents the onboard data logger GPS values. The symbol '*' indicates the times when daisy-chaining was performed and pose reconstruction is performed using a new set of feature points. Significant mismatch exists between the two GPS measurements, and the IBSE results remain proportionate to the two GPS measurements. Furthermore, the estimates agree closely with GPS2 for downrange and crossrange translation, and with GPS1 for altitude translation. There is no discernible discontinuity or increased error at the daisy-chain handoff times. Note that the 5 Hz refresh rate of the vision-based estimation is also higher than the 1 Hz rate of both GPS units. The pose estimation code can be executed in real time (> 30 Hz) on a typical laptop.

Solving for Relative Object Poses Through Knowledge of a Single Length

It is often not enough to recover camera pose relative to a reference pose. This section expands the techniques in Subsect. "The Planar Homography Algorithm" to recover camera pose relative to a planar object. Recovering relative pose between the camera and multiple objects, as well as the relative poses between objects, is also possible. To achieve these results, some additional information is required to recover the scale factor. It is sufficient to know the distance to a plane, or to recover it from precise motion, as in (19). The following example is based on work in [46,47], where relative pose with respect to an object is estimated given a single known geometric length between two feature points in the Euclidean-space.

Estimating the Pose of a Planar Object Relative to the Camera The planar homography algorithm is sufficient to give rotation and scaled translation of the camera or object with respect to some reference pose. However, it is not sufficient to solve for the complete pose of the camera with respect to a viewed object or the pose of one viewed object with respect to another. The following development pro-

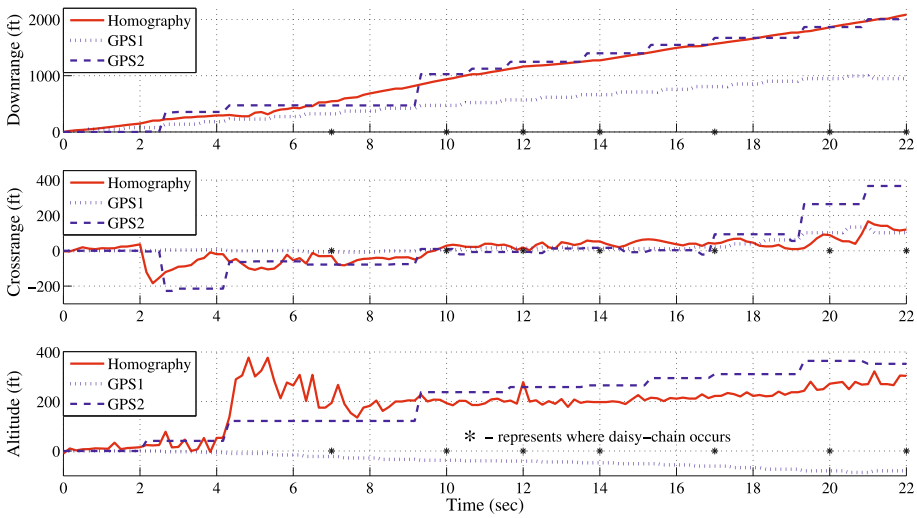

Image Based State Estimation, Figure 15
Experimental flight test results that compare the IBSE results to two GPS signals

vides an example where the planar homography algorithm can be used along with some additional information to attach a reference frame $\mathcal{F}_s$ to a planar object, with rotation and translation given with respect to the camera frame $\mathcal{F}_c$. That is, the development in this section can be used to determine the rotation and translation between $\mathcal{F}_s$ and $\mathcal{F}_c$.

Consider a camera viewing a planar object with four or more distinguishable feature points, and denote the feature point plane as π. A reference frame $\mathcal{F}_c$ is attached to the camera. If the camera and/or object move over time, a set of linear equations can be solved for the Homography Matrix $H(t)$ as in (16). The matrix $H(t)$ can be decomposed into $R(t)$, $x(t)/d^*$, and n^* as in (16), where the reference frame $\mathcal{F}_c^*$ can be taken as $\mathcal{F}_c(t_0)$, i. e., the initial frame. The normal vector of the planar object in the current frame $\mathcal{F}_c$ is given as

$$n(t) = R(t)n^* \,.$$

The goal is to attach a reference frame $\mathcal{F}_s(t)$ to the planar object. The orthonormal vectors $i_x, i_y, i_z \in \mathbb{R}^3$ that define $\mathcal{F}_s(t)$ form a rotation matrix $R(t)$ in $\mathcal{F}_c$ as

$$R_s = [i_x, i_y, i_z] \,. \tag{27}$$

Without loss of generality, the origin of $\mathcal{F}_s(t)$ is assigned to the feature point $\bar{m}_1$. The columns of $R_s(t)$ in (27) are defined as

$$i_z = -n \tag{28}$$

$$i_x = \frac{\bar{m}_2 - \bar{m}_1}{\bar{s}_1} \tag{29}$$

$$i_y = -n \times \frac{\bar{m}_2 - \bar{m}_1}{\bar{s}_1} \,, \tag{30}$$

where the constant distance between the two feature points $\bar{s}_1 = \|\bar{m}_1 - \bar{m}_2\|$ is assumed to be the extra information that can be exploited to complete the pose estimation. If $\bar{m}_1(t)$ and $\bar{m}_2(t)$ were known, then i_x and i_y can be determined from (29) and (30) since $\bar{s}_1$ is assumed to be known. To solve for $\bar{m}_1(t)$ and $\bar{m}_2(t)$, a new plane π_s' is defined with normal $-n(t)$ (so π_s' is parallel to π_s) and containing the normalized image point $m_1(t)$. A line l is defined from the origin of $\mathcal{F}_c$ through $m_2(t)$ and $\bar{m}_2(t)$. The plane π_s' intersects l at a point m_2'. The unknown distance between $m_1(t)$ and $m_2'(t)$ is s_1, as illustrated in Fig. 16.

The primitives l and π_s' are defined by the sets of points $q \in \mathbb{R}^3$ that satisfy the implicit functions

$$l = \{q \,|\, q - u m_2 = 0, \forall u \in \mathbb{R}\} \tag{31}$$

$$\pi_s' = \{q \,|\, n \cdot (q - m_1) = 0,\ q, n, m_1 \in \mathbb{R}^3\} \,. \tag{32}$$

The intersection of π_s' and l occurs when

$$u = \frac{n \cdot m_1}{n \cdot m_2} \,. \tag{33}$$

The expressions in (31) and (33) can be combined to solve for the point $q = m_2'$ as

$$m_2' = \frac{n \cdot m_1}{n \cdot m_2} m_2 \,.$$

The solution for $m_2'(t)$ is used to solve for s_1 as

$$s_1 = \left\| m_2' - m_1 \right\| \,, \tag{34}$$

and the properties of similar triangles can be used to calculate the following:

$$\frac{s_1}{\bar{s}_1} = \frac{\|m_1\|}{\|\bar{m}_1\|} = \frac{\|m_2'\|}{\|\bar{m}_2\|} \,. \tag{35}$$

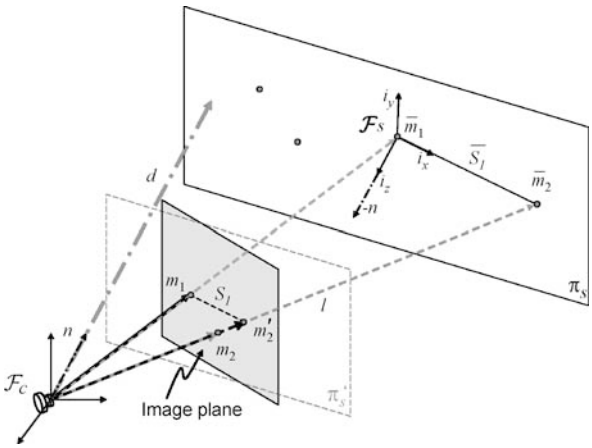

Image Based State Estimation, Figure 16
Geometric elements used to determine frame $\mathcal{F}_s$ with respect to $\mathcal{F}_c$

Since $s_1, \bar{s}_1$, $\|m_1(t)\|$, and $\|m_2'(t)\|$ are now known or measurable, (35) can be used to solve for $\|\bar{m}_1(t)\|$ and $\|\bar{m}_2(t)\|$, which can be used to recover $\bar{m}_1(t)$ and $\bar{m}_2(t)$ as

$$\bar{m}_1 = \frac{\|\bar{m}_1\|}{\|m_1\|} m_1 \,, \quad \bar{m}_2 = \frac{\|\bar{m}_2\|}{\|m_2\|} m_2 \,.$$

Solutions for i_x, i_y, and $R_s(t)$ can now be determined from (29), (30), and (27), respectively. Since $\mathcal{F}_s(t)$ is attached to $\bar{m}_1(t)$, the translation is simply given by $x_s(t) = \bar{m}_1(t)$. Furthermore, since $\bar{m}_1(t)$ can be determined, it is now possible to solve for the distance $d(t)$ as

$$d = n \cdot \bar{m}_1 \,. \tag{36}$$

If the constant length $\bar{s}_1$ is not known, but $d(t)$ is known or solved using methods such as (19), then the fact that $d(t) = z_j(t) n^{\mathrm{T}}(t) m_j(t)$ can be used to solve for the Euclidean coordinates of $\bar{m}_j(t)$ as

$$\bar{m}_j = \frac{d m_j}{n \cdot m_j} \,, \quad \forall j \in \{1 \ldots N\} \,. \tag{37}$$

Solving for the Euclidean coordinates of any two points $\bar{m}_1(t)$ and $\bar{m}_2(t)$ will allow the estimation of $\bar{s}_1$ which can be used to solve for the frame $\mathcal{F}_s(t)$.

Estimating the Pose of Multiple Planar Objects Relative to the Camera The previous development is now extended to the case of multiple planar patches and piecewise planar objects, given knowledge of only a single geometric length on a single static object in the scene. Consider a large sample of points $\mathcal{P}$ visible to the camera. These points have been grouped into k sets of coplanar points $\mathcal{P}_h \subset \mathcal{P}, \forall h \in \{1 \ldots k\}$, where all points in $\mathcal{P}_h$ lie in a plane π_h. The sets $\mathcal{P}_h$ may overlap, i. e., a point can be in more than one set.

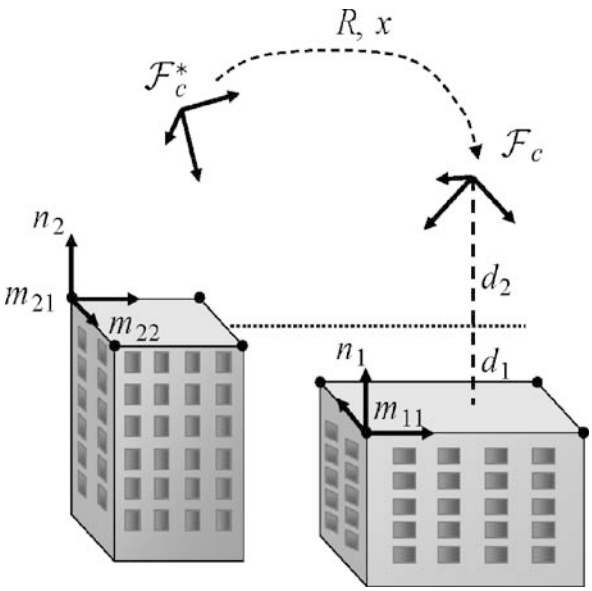

Image Based State Estimation, Figure 17
Example of multiple planar surfaces with respect to a camera

Segmenting a set of points into coplanar sets is not a trivial task. The sets can be distinguished through human interaction, scene knowledge (e. g., multiple light objects on a dark background) or various automatic methods [48,49]. This example assumes that each set $\mathcal{P}_h$ is well conditioned in the sense that it contains no more than three collinear points, as illustrated in Fig. 17.

In the following development, the camera is assumed to undergo a rotation $R(t)$ and translation $x(t)$ from a reference frame $\mathcal{F}_c^*$ to a frame $\mathcal{F}_c$. The points in each set $\mathcal{P}_h$ have coordinates

$$\begin{aligned} \bar{m}_{hj}^* &= [x_{hj}^*, y_{hj}^*, z_{hj}^*]^{\mathrm{T}}, \\ &\qquad \forall j \in \{1 \ldots N_h\}\,, \quad \forall h \in \{1 \ldots m\} \\ \bar{m}_{hj}(t) &= [x_{hj}, y_{hj}, z_{hj}]^{\mathrm{T}}, \\ &\qquad \forall j \in \{1 \ldots N_h\}\,, \quad \forall h \in \{1 \ldots m\} \end{aligned}$$

in the frames $\mathcal{F}_c^*$ and $\mathcal{F}_c$, respectively. These points project to image points with normalized coordinates m_{hj}^* and $m_{hj}(t)$ as described in Subsect. "The Planar Homography Algorithm". Each set of points in the two images are related by a homography $H_h(t)$ defined by

$$m_{hj} = \frac{z_{hj}^*}{z_{hj}} H_h \, m_{hj}^* \tag{38}$$

$$m_{hj} = \frac{z_{hj}^*}{z_{hj}} \left(R(t) + \frac{x(t)}{d_h^*} n_h^{*T} \right) m_{hj}^* \,. \tag{39}$$

Note that $R(t)$ and $x(t)$ are the same for all point sets $\mathcal{P}_h$, since all coordinate changes are due to the motion of the single camera. However, each plane π_h is different; therefore, each set of points will have different d_h^*, n_h^* and $H_h(t)$.

From m_{hj}^* and m_{hj}, it is possible to recover $H_h(t)$, n_h^*, $R(t)$, and $x_h(t) = x(t)/d_h^*$ for all $h \in \{1 \ldots m\}$. The subsequent development is based on the assumption that a single geometric length between two points in a single set is known. Without loss of generality, this length is assumed to be known in set $\mathcal{P}_1$. Given this geometric length, a reference frame $\mathcal{F}_{s1}$ can be attached to plane π_1 and the development in Subsect. "Estimating the Pose of a Planar Object Relative to the Camera" can be used to solve for $R_s(t)$, $x_s(t)$, $d_1(t)$, d_1^* and all $\bar{m}_{1j} \forall j \in \{1 \ldots N_1\}$. The translation $x(t)$ can then be recovered from d_1^* as

$$x(t) = d_1^* x_1(t) \,.$$

Given $x(t)$, each d_h^*, $\forall h \in \{2 \ldots m\}$ can be recovered from the scaled translations $x_h(t)$ as

$$d_h^* = \frac{x_h^{\mathrm{T}} x}{\|x_h\|} \,.$$

Once each d_h^* has been determined, all $\bar{m}_{hj}^*, \forall j \in \{1 \ldots N_h\}, \forall h \in \{2 \ldots m\}$ can be recovered as in (37). From knowledge of $\bar{m}_{hj}^*$, a constant length $\bar{s}_h$ between two feature points can be estimated for each plane, and the frames $\mathcal{F}_{sh}$ can be attached to the plane π_h. Thus, (27)–(30) can be used to solve for R_s and $x_s \forall h \in \{2 \ldots m\}$. Given the rotation and translation from each plane π_h to the reference camera frame $\mathcal{F}_c$, the rotation and translation between each planar patch can be recovered.

For a stationary camera viewing multiple moving planar objects, the analysis cannot be performed because there is not a common $R(t)$ or $x(t)$. If a geometric length is known on each object, then the analysis in Subsect. "Estimating the Pose of a Planar Object Relative to the Camera" can be performed for each moving plane.

Experiments of Pose Estimation of a Single Object An experiment using the moving vehicle in Fig. 18 is performed to demonstrate the method presented in Subsect. "Estimating the Pose of a Planar Object Relative to the Camera". Four bright LED arrays were fixed to the back of a truck to facilitate simple image segmentation, where the centroid of each detected array provides four feature points used to construct the Homography Matrix. Each of the four centroids is indicated in Fig. 18 by a cross and a number. The truck was equipped with a differential GPS unit to provide a reference to compare the pose estimation. The road was marked at approximately 20 f (6.1 m) intervals and the car was driven forward and stopped approximately every 20 f.

Image Based State Estimation, Figure 18
A processed video frame from the pose estimation experiment

The results of the experiment are seen in Fig. 19. The expected periodically increasing step function along the camera frame z direction is evident. Furthermore, the change in pose estimate agrees closely to the GPS Northing measurement. There is also a small periodic step increase estimated in the camera x direction. The estimate becomes noticeably degraded as the distance to the vehicle increases. This is primarily due to sensor noise. As the car moves farther from the camera, the perceived lights become dimmer and it is harder to extract the centroids. This increases the effects of pixilation (i. e., quantization noise).

Experiments of Pose Estimation of Multiple Objects Experimental results are also provided to demonstrate the efficacy of the algorithm presented in Subsect. "Estimating the Pose of Multiple Planar Objects Relative to the Camera". Two images of a cluttered desktop were taken with a camera with a resolution of 1280×1024. One geometric length, the width of the photograph frame, is known. In each image, the four corners of the picture frame, four corner points on a computer monitor and four points on a speaker give feature points. From two of the images, reference frames are attached to all three objects as seen in Fig. 20 and the location of all corner points are reconstructed in the camera frame as seen in Fig. 21.

The true pose of these objects with respect to the camera is impossible to ascertain, so two heuristic tests are performed. First the estimated dimensions of the objects are compared to hand-measured dimensions in Table 1. The maximum error is in the height of the monitor screen, with an estimate error of 5.7%. The average estimate error is 1.8%. For the second test, a third image was then taken from another viewpoint, as seen in Fig. 22. The corners of the picture frame were identified, but no other points on any other object are selected. The current location of

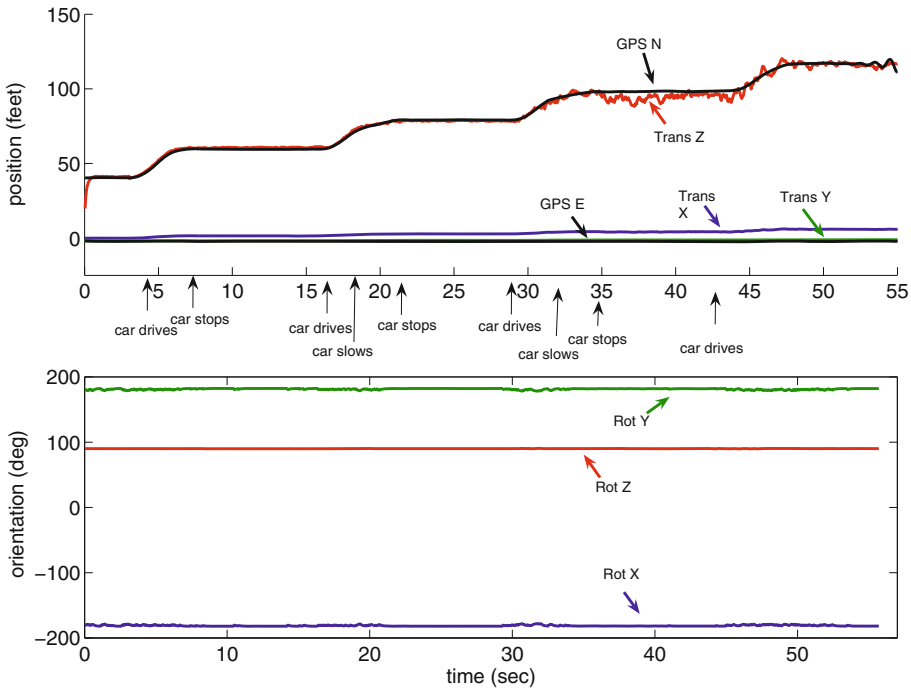

Image Based State Estimation, Figure 19
Results of experiment to estimate the pose of a moving vehicle

Image Based State Estimation, Figure 20
Image with reference frames attached to a photograph, monitor and speaker

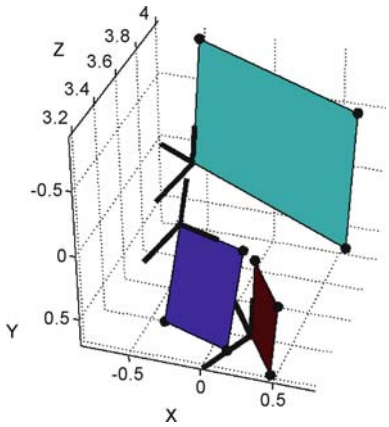

Image Based State Estimation, Figure 21
Geometric reconstruction of the photograph, monitor and speaker in the camera frame $\mathcal{F}_c$. Units are in feet

Image Based State Estimation, Table 1
Measured vs Estimated Lengths

Object	Width	Height	Est. width	Est. height
Photo	6″	8″	Known	8.03″
Screen	16″	12″	16.92″	12.04″
Speaker	4.125″	6.875″	4.19″	6.96″

the corners of the monitor and speaker in the image are estimated through the estimated relative positions solved in the previous reconstruction. From the location of the frames and points known in the camera frame, the pose of each reference frame can be determined with respect to the photograph frame, and the location of each of the object's points can be determined within its own reference frame. With the picture frame localized in the new camera frame,

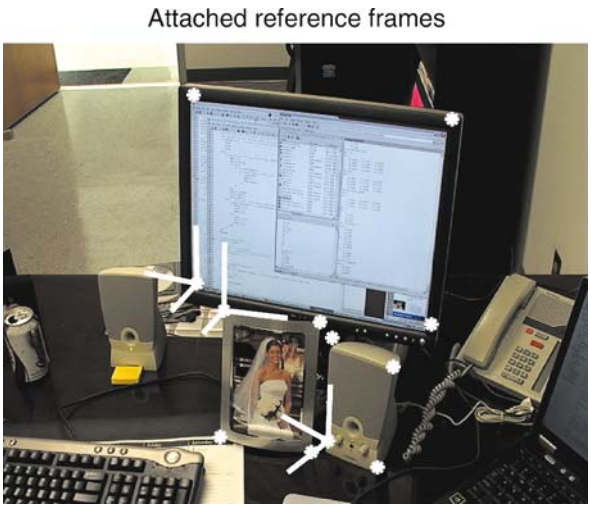

Image Based State Estimation, Figure 22
Reprojected feature points from their estimated Euclidean positions

the locations of the other frames and points are estimated and projected to their estimated locations in the third image.

Estimation of Linear and Angular Velocity

Image based estimation of velocity has a shorter history than estimation of pose. One method of velocity estimation is a continuous version of the eight-point algorithm that uses optical flow (i. e., velocity of image points). Early efforts were developed by Zhuang and Haralick [50]. Ma et al. [51] also introduced a continuous version of the eight-point algorithm that very closely resembled the classic eight-point algorithm. An optical flow based version of the planar homography algorithm is also provided in [26], which is discussed in Subsect. "Continuous Planar Homography Algorithm".

Another method of velocity estimation is to use standard pose estimation methods to generate an initial estimate that is input to functions or filters that estimate velocity. Soatto et al. used a logarithmic map of the results of the eight-point algorithm to estimate velocity [14]. Chitrakaran et al. [19] use the results of the planar homography as input to a nonlinear observer, which will be discussed in Subsect. "Velocity Estimation Through a Nonlinear Observer".

Continuous Planar Homography Algorithm

Consider a camera with reference frame $\mathcal{F}_c$ moving with linear velocity $v(t)$ and angular velocity $\omega(t)$. The camera views a collection of $N \geq 4$ or more feature points lying in a plane π in front of the camera. The plane π has normal $n(t)$ with respect to $\mathcal{F}_c$, and lies a distance $d(t)$ from the origin of $\mathcal{F}_c$ along $n(t)$. These feature points have Euclidean coordinates $\bar{m}_j(t) \in \mathbb{R}^3$ and velocities given as a function of the linear and angular velocity of the camera, v and ω, respectively,

$$\dot{\bar{m}}_j = \omega \times \bar{m}_j + v\,, \quad \forall j \in \{1 \ldots N\}\,. \tag{40}$$

The Euclidean coordinates of the coplanar feature points obey the projection relationship

$$d = n^{\mathrm{T}} \bar{m}_j\,, \quad \forall j \in \{1 \ldots N\}\,. \tag{41}$$

Substituting (41) into (40) gives

$$\dot{\bar{m}}_j = \left([\omega]_\times + \frac{v}{d} n^{\mathrm{T}}\right) \bar{m}_j\,, \quad \forall j \in \{1 \ldots N\}\,, \tag{42}$$

where $[\omega(t)]_\times \in \mathbb{R}^{3\times 3}$ is the skew-symmetric matrix form of the vector $\omega(t)$ such that $[\omega(t)]_\times \bar{m}_j(t) = \omega(t) \times \bar{m}_j(t)$. From the normalized image points $m_j(t)$ and the corresponding optical flow vectors $\dot{m}_j(t)$, the optical flow version of the Homography Matrix can be obtained as

$$\dot{m}_j = \frac{z_j}{\dot{z}_j}\left([\omega]_\times + \frac{v}{d} n^{\mathrm{T}}\right) m_j \tag{43}$$

$$= \alpha_c H_c m_j \tag{44}$$

where $\alpha_c = z_j/\dot{z}_j$ is a ratio of depth to change in depth, and $H_c = \left([\omega]_\times + (v/d)n^{\mathrm{T}}\right)$ is the Continuous Homography Matrix. Similar to the pose estimation method of Subsect. "The Planar Homography Algorithm", it is possible to solve for $\alpha_c, H_c(t), n(t), \omega(t)$ and $v(t)/d$. Similar to translation estimates, the velocity can only be solved up to the scale factor d. Also like the pose estimation case, there exist multiple mathematical solutions for $n(t), \omega(t)$ and $v(t)/d$, where two solutions can be eliminated based on the physics of the problem see [26] for further details.

Velocity Estimation Through a Nonlinear Observer

Given an image sequence (e. g., video), the pose estimation methods discussed in Sect. "Linear Estimation of Position and Orientation from Two Images" can be used to estimate velocity using established observer design methods. For example, Chitrakaran et al. [19], used homography-based techniques to design a nonlinear observer to identify an object's unknown velocity. The method in [19] requires knowledge of the object's initial pose with respect to the camera, and knowledge of a single length on the object. As demonstrated in Subsect. "Estimating the Pose of

a Planar Object Relative to the Camera" the knowledge of the single length is sufficient to recover the initial pose.

Consider an inertial frame $\mathcal{I}$ attached to a fixed camera that is viewing a planar object with an attached reference frame $\mathcal{F}_s(t)$, moving with unknown linear and angular velocities $v_s(t), \omega_s(t) \in \mathbb{R}^3$ expressed in frame $\mathcal{I}$. Four or more feature points are visible on the object at all time, and the coordinates of one point are known with respect to $\mathcal{F}_s(t)$.

The feature points have pixel coordinates in the image plane as given in (1). The extended image coordinates [52] of the image point $p_1(t)$, denoted by $p_e(t)$, seen when the camera is at a pose $\mathcal{F}_s(t)$ with respect to the viewed object are defined as

$$p_e(t) = [u_1(t), v_1(t), \ln(z_t)]^T , \tag{45}$$

where $\ln(\cdot)$ denotes the natural logarithm. The constant vector $p_e^* \in \mathbb{R}^3$ denotes the extended image coordinates of the corresponding image point seen in a reference image taken when the object is at a known reference pose $\mathcal{F}_s^*$, and is defined as

$$p_e^* = [u_1^*, v_1^*, \ln(z_1^*)]^T . \tag{46}$$

If $\mathcal{F}_s^*$ is not known, the methods of Subsect. "Estimating the Pose of a Planar Object Relative to the Camera" can be used to estimate $\mathcal{F}_s^*$ given a known geometric length.

The Homography Matrix $H(t)$ can be solved from the images of the object taken at $\mathcal{F}_s^*$ and $\mathcal{F}_s(t)$, and can be decomposed to give $R(t)$ and $\alpha_1(t)$. The translation of the object, denoted by $e_v(t) \in \mathbb{R}^3$, is defined as

$$e_v = p_e - p_e^* . \tag{47}$$

The first two elements of $e_v(t)$ are directly measured from the images, and

$$\ln(z_1(t)) - \ln(z_1^*) = -\ln(\alpha_1)$$

where $\alpha_1(t)$ is reconstructed from the homography. The time derivative of (47), is given by

$$\dot{e}_v = \dot{p}_e = \frac{\alpha_1}{z_1^*} A_e L_v \left[v_s - R[s_1]_\times R^T \omega_s\right] \tag{48}$$

where s_1 denotes the known, constant coordinates of feature point 1 in $\mathcal{F}_s(t)$. In (48), $A_e \in \mathbb{R}^{3\times3}$ is defined as

$$A_e = A - \begin{bmatrix} 0 & 0 & u_0 \\ 0 & 0 & v_0 \\ 0 & 0 & 0 \end{bmatrix} , \tag{49}$$

the Jacobian-like matrix $L_v(t) \in \mathbb{R}^{3\times3}$ is defined as

$$L_v = \begin{bmatrix} 1 & 0 & -\frac{x_1}{z_1} \\ 0 & 1 & -\frac{y_1}{z_1} \\ 0 & 0 & 1 \end{bmatrix} , \tag{50}$$

and $R(t)$ is reconstructed from the Homography Matrix.

The rotation of the object at $\mathcal{F}(t)$, with respect to the fixed coordinate system $\mathcal{F}^*$, is defined using the angle-axis representation of the rotation $R(t)$ as

$$e_\omega \triangleq u(t)\theta(t) . \tag{51}$$

In (51), $u(t) \in \mathbb{R}^3$ is a unit vector giving the axis of rotation, and $\theta(t) \in \mathbb{R}$ denotes the rotation angle about $u(t)$, where $\theta(t) \in (-\pi, \pi]$. The time derivative of (51) is given by

$$\dot{e}_\omega = L_\omega \omega_s , \tag{52}$$

where the Jacobian–like matrix $L_\omega(t) \in \mathbb{R}^{3\times3}$ is defined as

$$L_\omega = I_3 - \frac{\theta}{2}[u]_\times + \left(1 - \frac{\operatorname{sinc}(\theta)}{\operatorname{sinc}^2(\theta/2)}\right)[u]_\times^2 . \tag{53}$$

The IBSE objective is to develop an estimator that can be used to identify the velocity of an object given by

$$v(t) = \left[v_s^T(t), \omega_s^T(t)\right]^T \in \mathbb{R}^6 .$$

Combing (48) and (52), the combined linear and angular velocity error system can be determined as a function of $v(t)$ as

$$\dot{e} = \left[\dot{e}_v^T, \dot{e}_\omega^T\right]^T = Jv , \tag{54}$$

where the nonsingular Jacobian-like matrix $J(t) \in \mathbb{R}^{6\times6}$ is defined as

$$J = \begin{bmatrix} \frac{\alpha_1}{z_1^*} A_e L_v & -\frac{\alpha_1}{z_1^*} A_e L_v R[s_1]_\times R^T \\ 0 & L_\omega \end{bmatrix} . \tag{55}$$

The open-loop system in (54) can be inverted in as

$$v = J^{-1}\dot{e} \tag{56}$$

where

$$\begin{aligned} J^{-1} &= \begin{bmatrix} \frac{z_1^*}{\alpha_1} L_v^{-1} A_e^{-1} & \left(\frac{z_1^*}{\alpha_1} L_v^{-1} A_e^{-1}\right) \frac{\alpha_1}{z_1^*} A_e L_v R[s_1]_\times R^T L_\omega^{-1} \\ 0 & L_\omega^{-1} \end{bmatrix} \\ &= \begin{bmatrix} \frac{z_1^*}{\alpha_1} L_v^{-1} A_e^{-1} & R[s_1]_\times R^T L_\omega^{-1} \\ 0 & L_\omega^{-1} \end{bmatrix} , \end{aligned}$$

and

$$L_\omega^{-1} = I_3 + \frac{\theta}{2}\,\mathrm{sinc}^2\left(\frac{\theta}{2}\right)[u]_\times + (1-\mathrm{sinc}(\theta))\,[u]_\times^2 \ .$$

Based on (56), and the fact that $J^{-1}(t)$ exists, the velocity estimation goal can be expressed as the desire to estimate $\dot{e}(t)$ and map it to $v(t)$. The estimated pose is denoted as $\hat{e}(t)$, and the estimator error is defined as

$$\tilde{e} = e - \hat{e} \ . \tag{57}$$

Based on Lyapunov analysis and design methods detailed in [19] and [53], an estimate update law $\dot{\hat{e}}(t)$ is generated from the known error signal $\tilde{e}(t)$ as

$$\dot{\hat{e}} = \int_0^t (K+I_6)\tilde{e}(\tau)\,\mathrm{d}\tau + \int_0^t \rho\,\mathrm{sgn}(\tilde{e})\,\mathrm{d}\tau + (K+I_6)\tilde{e}(t) \tag{58}$$

where K and $\rho \in \mathbb{R}^{6\times 6}$ are positive definite constant diagonal gain matrices, $I_6 \in \mathbb{R}^{6\times 6}$ denotes the 6×6 identity matrix, and the notation $\mathrm{sgn}(\tilde{e})$ denotes the standard signum function applied to each element of the vector $\tilde{e}(t)$. The estimator in (58) ensures that

$$\|\tilde{e}(t)\| \ , \ \left\|\dot{\tilde{e}}(t)\right\| \to 0 \quad \text{as} \quad t \to \infty \ . \tag{59}$$

Based on (59), the relationship in (56) can be used to recover the velocity $v(t)$.

Experiments of Velocity Estimate

A simulation is presented to illustrate the performance of this type of the image-based velocity estimate. Four coplanar points are simulated undergoing a periodic motion involving translation and rotation along all directions. Figs. 23 and 24 show the true velocity, the estimated velocity and resulting estimation error.

An experiment was also performed to estimate the velocity of the truck in Subsect. "Experiments of Pose Estimation of a Single Object". A low-pass filter was applied to the output of the nonlinear estimator to mitigate the effects of signal noise. Results of the experiment are seen in Fig. 25. The GPS data was backwards differenced to provide a comparison velocity. The truck periodically accelerated and stopped without changing direction. The camera optical axis (z-axis) and the truck were bearing approximately due North. It is expected that image-based estimate of the truck velocity in direction of the camera z-axis should be very close to the GPS-based estimate of velocity North. Indeed the velocity estimation is in close agreement with the GPS estimate, though the effects of signal noise are prevalent as the vehicle moves farther away.

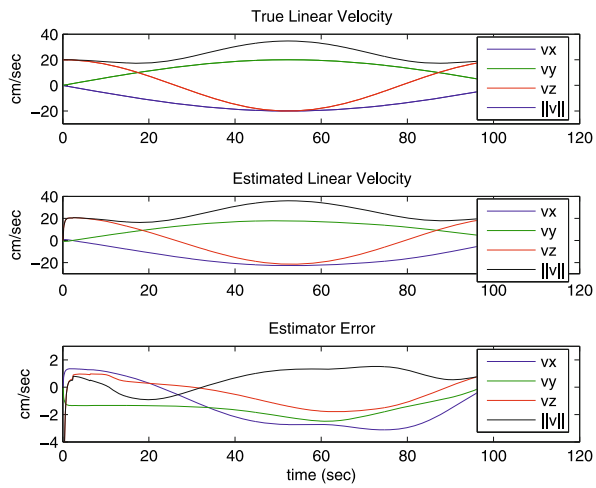

Image Based State Estimation, Figure 23
Linear velocity estimation results for simulation of nonlinear velocity observer

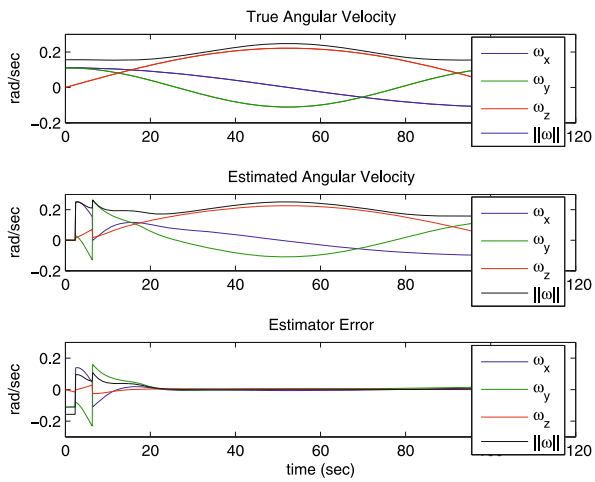

Image Based State Estimation, Figure 24
Angular velocity estimation results for simulation of nonlinear velocity observer

Image Based Kalman Filter Estimation

The KF and its variations enjoy widespread use as state estimators in many fields of engineering and science. Described simply, the KF is a linear, recursive estimator. For linear systems, the KF is an optimal estimator that gives the estimate with the least error covariance. For nonlinear systems (such as IBSE), the EKF provides an effective estimate by linearizing the state equations. A brief review of the KF/EKF is provided to introduce the notation. References for further insight into KF/EKF are provided in [54,55].

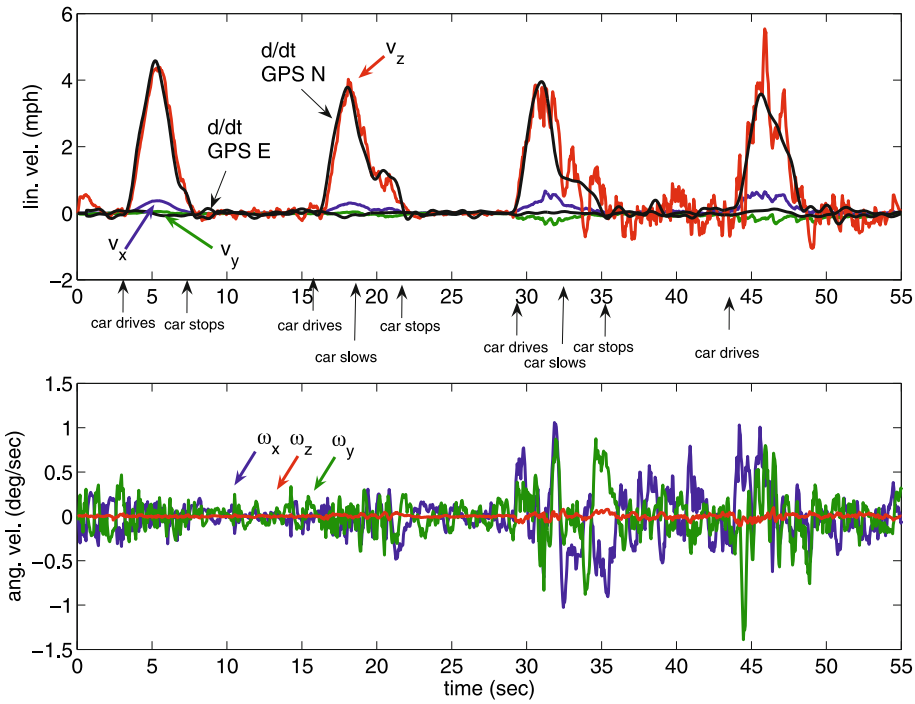

Image Based State Estimation, Figure 25
Estimated velocity of a moving vehicle

Given its relative simplicity and prominence, it is no surprise that the KF/EKF has been widely used in IBSE. KF/EKF approaches also explicitly account for signal/measurement noise, in contrast to the methods presented in Sect. "Linear Estimation of Position and Orientation from Two Images". Some IBSE approaches use pose estimation methods as in Sect. "Linear Estimation of Position and Orientation from Two Images" as an initial pose measurement that is supplied to the KF/EKF for further pose and/or velocity estimation. The advantage of this solving for the pose estimate as an input to the KF/EKF is that the system can be expressed in the linear form

$$\dot{x} = Ax\,,$$

thus the estimate is expected to be optimal. For example, Soatto et al. [14] give two Kalman filter approaches based on the Essential Matrix. The first approach solves for motion from the Essential Matrix and passes the estimation through a Kalman filter. The second approach solves for the Essential Matrix and passes the result as a measurement into a Kalman filter to be refined. However, the space of Essential Matrices is not a vector space (e. g., the sum of two Essential Matrices is not necessarily an Essential Matrix), so the output of the Kalman filter is projected to space of Essential Matrices before being decomposed to find the rotation and translation. Other methods eschew external pose estimation schemes and use image features as inputs to the KF/EKF with pose and/or velocity as output [15,16,18]. The advantage of this approach is lower computational cost and avoiding the restrictions and limitations of the pose reconstruction methods detailed in Subsect. "Characteristics of the Essential Matrix Algorithm and Planar Homography Matrix".

The state to be estimated is denoted by the vector $x(t) \in \mathbb{R}^n$ and consists of pose, orientation, and velocity signals. The system output (i. e., the signals available for measurement) consists of image information, such as feature points, and is denoted by $y(t) \in \mathbb{R}^l$. The system is expressed as a discrete-time nonlinear system as [15]

$$x[k+1] = f(x[k], w[k], k) \tag{60}$$

$$y[k] = h(x[k], v[k], k) \tag{61}$$

where $w[k]$ and $v[k]$ are uncorrelated, white, zero-mean noise signals with known covariance matrices $Q[k]$ and $R[k]$, respectively. There is varying notation and terminology associated with the KF/EKF. This chapter denotes the estimated state by $\hat{x}[k]$, the time update or a priori estimate as $\hat{x}^-[k]$, and measurement update or a posteriori estimate as $\hat{x}^+[k]$. Likewise, the a priori and a posteriori error covariance matrices are denoted by $P^-[k]$ and $P^+[k]$, respectively.

The a priori state estimates at time k are given by

$$\hat{x}^-[k] = f_{k-1}(\hat{x}[k], 0, k) \tag{62}$$

$$P^-[k] = F[k]P^+[k-1]F^{\mathrm{T}}[k] + L[k]Q[k-1]L^{\mathrm{T}}[k], \tag{63}$$

where

$$F[k] \in \mathbb{R}^{n\times n} = \frac{\partial f}{\partial x}|_{x=\hat{x}^+[k-1]} \quad \text{and}$$
$$L[k] \in \mathbb{R}^{n\times n} = \frac{\partial f}{\partial w}|_{x=\hat{x}^+[k-1]} .$$

The a posteriori state estimates at time k are given by

$$K[k] = P^-[k]H^{\mathrm{T}}[k] \cdot \left(K[k]P^-[k]H^{\mathrm{T}}[k] + M[k]R[k]M^{\mathrm{T}}[k]\right)^{-1} \tag{64}$$

$$= \hat{x}^-[k] + K[k]\left(y[k] - h(\hat{x}^-[k], 0, k)\right) \tag{65}$$

$$P^+[k] = (I - K[k]H[k])\, P^-[k]\, (I - K[k]H[k])^{\mathrm{T}} + K[k]L[k]R[k]L^{\mathrm{T}}[k]K[k] \tag{66}$$

where $K[k]$ is called the *Kalman Gain*, and

$$H[k] \in \mathbb{R}^{l\times n} = \frac{\partial h}{\partial x}|_{x=\hat{x}^-[k]} \quad \text{and}$$
$$L[k] \in \mathbb{R}^{l\times l} = \frac{\partial h}{\partial v}|_{x=\hat{x}^-[k]} .$$

Assuming the camera is stationary and viewing a rigid body undergoing general rigid body motion, a reference frame $\mathcal{F}_s$ is attached to the object with an origin at a point $\bar{m}_0(t) = [x_0(t), y_0(t), z_0(t)] \in \mathbb{R}^3$, which is not necessarily visible (e. g., it can be in the interior of the rigid body). Translation and linear velocity $v(t)$ is measured as the displacement of $\bar{m}_0(t)$, and orientation is measured as the orientation $R \in \mathrm{SO}(3)$ between the inertial camera frame and body frame. Broida et al. [15] choose to represent rotation in terms of unit quaternions, which can be mapped to rotation matrix $R(t)$ as in [56]. A unit quaternion is a four element vector denoted as

$$q(t) \triangleq [q_0(t) q_v^{\mathrm{T}}(t)]^{\mathrm{T}} \in \mathbb{R}^4 ,$$

where $q_0(t) \in \mathbb{R}, q_v(t) \in \mathbb{R}^3$, and $\|q(t)\| = 1$. The unit quaternion is computed from a rotation matrix $R(t)$ as

$$q_0 = \frac{1}{2}\sqrt{1 + tr(R)} \quad q_v = \frac{1}{2}u\sqrt{3 - tr(R)} , \tag{67}$$

with the reverse mapping given by

$$R(q) = (q_0^2 - q_v^{\mathrm{T}} q_v) I_3 + 2 q_v q_v^{\mathrm{T}} - 2 q_0 [q_v]_\times . \tag{68}$$

To facilitate the subsequent development, let the notation $R_{xj}(q), R_{yj}(q), R_{zj}(q)$ denote the first, second, or third row of $(1/z_j(t))R(q)\bar{m}_j(t)$. For instance

$$R_{xj} = \left(q_1^2 - q_2^2 - q_3^2 - q_4^2\right)\frac{x_j}{z_0} + 2(q_1q_2 - q_3q_4)\frac{y_j}{z_0} + 2(q_1q_3 + q_2q_4)\frac{z_j}{z_0} .$$

The angular velocity is given by the vector $\omega(t)$, and the derivative of the unit quaternion is given as a function of angular velocity as

$$\dot{q} = \Omega q$$
$$\Omega = \frac{1}{2}\begin{bmatrix} 0 & \omega_z & -\omega_y & \omega_x \\ -\omega_z & 0 & \omega_x & \omega_y \\ \omega_y & -\omega_x & 0 & \omega_z \\ -\omega_x & -\omega_y & -\omega_z & 0 \end{bmatrix} .$$

The state vector for the KF-based IBSE is denoted by $x[k] \in \mathbb{R}^{12+3N}$ and defined as

$$x[k] = \left[m_0^{\mathrm{T}}, \frac{v^{\mathrm{T}}}{z_0}, q^{\mathrm{T}}, \omega^{\mathrm{T}}, \frac{\bar{m}_1^{\mathrm{T}}}{z_0}, \ldots, \frac{\bar{m}_N^{\mathrm{T}}}{z_0}\right]^{\mathrm{T}} \tag{69}$$

where

$$\bar{m}_j(t) = [x_j(t), y_j(t), z_j(t)]^{\mathrm{T}} , \quad j \in \{1 \ldots N\}$$

for N points, and $m_0(t) \in \mathbb{R}^2$ is the image of the origin given by

$$m_0 = \left[\frac{x_0}{z_0}, \frac{y_0}{z_0}\right]^{\mathrm{T}} .$$

The resulting linear state-space systems can now be expressed as

$$x[k+1] = f(x[k], 0, k) = F[k]x[k]$$

where $F[k]$ is given by

$$F = \begin{bmatrix} F_1 & 0 & 0 \\ 0 & F_2 & 0 \\ F_3 & 0 & F_4 \end{bmatrix} \in \mathbb{R}^{12+3N\times 12+3N}$$

$$F_1 = \begin{bmatrix} -\frac{v_z}{z_0} & 0 & 1 & 0 & -\frac{x_0}{z_0} \\ 0 & -\frac{v_z}{z_0} & 0 & 1 & -\frac{y_0}{z_0} \\ 0 & 0 & -\frac{v_z}{z_0} & 0 & -\frac{v_x}{z_0} \\ 0 & 0 & 0 & -\frac{v_z}{z_0} & -\frac{v_y}{z_0} \\ 0 & 0 & 0 & 0 & -2\frac{v_z}{z_0} \end{bmatrix} \in \mathbb{R}^{5\times 5}$$

$$F_2 = \begin{bmatrix} 0 & \omega_z & -\omega_y & \omega_x & -x_1 & q_4 & -q_3 & q_2 \\ -\omega_z & 0 & \omega_x & \omega_y & -x_2 & q_3 & q_4 & -q_1 \\ \omega_y & -\omega_x & 0 & \omega_z & -x_3 & -q_2 & q_1 & q_4 \\ -\omega_x & -\omega_y & -\omega_z & 0 & -x_4 & -q_1 & -q_2 & -q_3 \\ 0_{3\times1} & 0_{3\times1} & 0_{3\times1} & 0_{3\times1} & 0_{3\times1} & 0_{3\times1} & 0_{3\times1} & 0_{3\times1} \end{bmatrix} \in \mathbb{R}^{8\times7}$$

$$F_3 = \begin{bmatrix} 0_{3\times4} & \frac{\bar{m}_1^{\mathrm{T}}}{z_0} \\ 0_{3\times4} & \vdots \\ 0_{3\times4} & \frac{\bar{m}_N^{\mathrm{T}}}{z_0} \end{bmatrix} \in \mathbb{R}^{3N\times5}$$

$$F_4 = -\frac{v_z}{z_0} I_{3N} \in \mathbb{R}^{3N\times3N} \tag{70}$$

where $0_{a\times b}$ is a zero matrix of size $a \times b$ and I_{3N} is a $3N \times 3N$ identity matrix.

The measurement

$$y[k] = h(x[k]) = [m_1^{\mathrm{T}}, \ldots, m_N^{\mathrm{T}}]^{\mathrm{T}}$$

is the collection of feature points where the linearization of $h(x[k])$ is denoted by $H[k]$ and given by

$$H = \begin{bmatrix} G_1 & 0 & W_1 & 0 & S_1 & 0 & \ldots & 0 \\ G_2 & 0 & W_2 & 0 & 0 & S_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ G_N & 0 & W_N & 0 & 0 & 0 & \ldots & S_N \end{bmatrix} \in \mathbb{R}^{2N\times3N}. \tag{71}$$

In (71), $G_j(q) \in \mathbb{R}^{2\times2}$ is an auxiliary variable defined as

$$G_j = \begin{bmatrix} \frac{1}{1+R_{zj}} & 0 \\ 0 & \frac{1}{1+R_{zj}} \end{bmatrix},$$

$W_j(q) \in \mathbb{R}^{2\times4}$ is given elementwise by

$$W_{j1a} = \frac{1}{(1+R_{zj})^2}\left[(1+R_{zj})\frac{\partial R_{xj}}{\partial q_a} - (x_0+R_{zj})\frac{\partial R_{zj}}{\partial q_a}\right] \quad a \in \{1,\ldots4\}$$

$$W_{j2a} = \frac{1}{(1+R_{zj})^2}\left[(1+R_{zj})\frac{\partial R_{yj}}{\partial q_a} - (x_0+R_{zj})\frac{\partial R_{zj}}{\partial q_a}\right] \quad a \in \{1,\ldots4\}$$

and $S_j(q) \in \mathbb{R}^{2\times3}$ is given elementwise by

$$S_{jba} = \frac{1}{(1+R_{zj})^2}\left[(1+R_{zj})R_{ba} - (x_0+R_{zj})R_{ba}\right], \quad b \in \{1,2\}, \quad a \in \{1,2,3\}.$$

The definitions of $x[k], F[k], H[k]$ in (69)–(71) can then be used in the recursive estimator defined in equations (62)–(66).

Range Identification

In the previous sections, the distance to the feature points is a pervasive uncertainty in the system. This section provides an example of how the range to feature points can also be estimated. The problem in range identification is to determine the depth of a set of feature points. That is, given a set of image points $m_j(t) = [x_j/z_j, y_j/z_j, 1]^{\mathrm{T}}$, the task in range identification is to determine the terms $z_j(t)$ and thus recover $\bar{m}(t) = [x_j, y_j, z_j]^{\mathrm{T}}$. In general, additional information about the target or camera motion is required to perform range identification. For example, the methods in this section require known velocity estimation of the camera or viewed target. An example problem is a mobile robot or UAV, where a combination of joint encoders, accelerometers, or inertial measurement units and a camera can be used to determine the velocity and pose relative to objects in the environment.

Several researchers have investigated the range identification problem for conventional imaging systems when the motion parameters are known. In [20], Jankovic and Ghosh developed a discontinuous observer, to exponentially identify range information of features from successive images of a camera where the object model is based on known skew-symmetric affine motion parameters. In [21], Chen and Kano generalized the object motion beyond the skew-symmetric form of [20] and developed a new discontinuous observer that exponentially forced the state observation error to be uniformly ultimately bounded. More recently, a state estimation strategy was developed in [12,24] for affine systems with known motion parameters where only a single homogeneous observation point is provided (i. e., a single image coordinate). In [57], a reduced order observer was developed to yield a semi-global asymptotic stability result for a fixed camera viewing a moving object with known motion parameters for a pinhole camera. In [23], a continuous observer is used to asymptotically identify the range information for a general affine system with known motion parameters. Recent efforts in range estimation (e. g., [12,13]) have extended this concept for use with omnidirectional cameras (i. e., cameras that use curved mirrors to achieve a 360° field of view) and uncertain motion. This section focuses on the case in where a continuous observer is used to asymptotically identify the range information for a general affine system with known motion parameters, using a pinhole camera as in [23].

Nonlinear Estimator for Range Identification

For the problems in this section, consider the scenario of a moving target with a stationary camera or the scenario of moving camera with a stationary target. In both scenarios, the relative motion dynamics can be written in the same form, although the motion parameters have different physical meanings. Specifically, for both scenarios, the affine motion dynamics can be expressed as

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \tag{72}$$

where $x(t), y(t), z(t) \in \mathbb{R}$ denote the unmeasurable task-space coordinates of the point in an inertial frame attached to the camera. The parameters $a_{i,j}(t) \in \mathbb{R}$ and $b_i(t) \forall i, j = 1, 2, 3$ denote the known motion parameters. To illustrate how the same dynamics apply to both scenarios presented, consider the case of a moving target with linear and angular velocity $v_t(t)$ and $\omega_t(t)$ measured with respect to the camera frame. A point attached to the object has coordinates $\bar{m}(t)$ in the camera frame, with relative motion to the camera described by

$$\dot{\bar{m}} = -[\omega_t]_\times \bar{m} - v_t = A\bar{m} + b\,. \tag{73}$$

For the moving camera stationary object scenario, consider a feature point attached to a stationary object. The linear and angular velocities of the target with respect to the camera (expressed in $\mathcal{F}_c$) can be written as

$$v_t = -Rv_c \quad \omega_t = -R\omega_c \tag{74}$$

where $R(t) \in SO(3)$ denotes the corresponding rotation between the camera and object frame, and $v_c(t)$ and $\omega_c(t)$ denote the linear and angular velocity of the camera, respectively. Based on (74), the relationship in (73) can be rewritten as

$$\dot{\bar{m}} = [R\omega_c]_\times \bar{m} + Rv_c = A\bar{m} + b\,. \tag{75}$$

The normalized Euclidean coordinates of a feature point (which are measurable from the image-space through the pin-hole model in (5)), are denoted by $m(t) \in \mathbb{R}^2$, and defined as

$$m \triangleq [m_1, m_2]^{\mathrm{T}} = \left[\frac{x}{z}, \frac{y}{z}\right]^{\mathrm{T}}\,. \tag{76}$$

The affine dynamics introduced in (72), and the image-space signal introduced in (76) define the perspective system [21]. After taking the time derivative of (76) and utilizing (72), the image-space trajectory of the object feature can be obtained as

$$\dot{m}_1 = \frac{a_{11}x + a_{12}y + a_{13}z + b_1}{z} - \frac{x(a_{31}x + a_{32}y + a_{33}z + b_3)}{z^2} \tag{77}$$

$$\dot{m}_2 = \frac{a_{21}x + a_{22}y + a_{23}z + b_2}{z} - \frac{y(a_{31}x + a_{32}y + a_{33}z + b_3)}{z^2}\,. \tag{78}$$

To facilitate subsequent analysis, the time derivative of the inverse of $z(t)$ is determined as

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{z}\right) = \frac{-a_{31}x - a_{32}y - a_{33}z - b_3}{z^2}\,. \tag{79}$$

By utilizing (76), the expressions given in (77)–(79) can be rewritten as

$$\dot{m}_1 = a_{13} + (a_{11} - a_{33})m_1 + a_{12}m_2 - a_{31}m_1^2 - a_{32}m_1m_2 + f_1 \tag{80}$$

$$\dot{m}_2 = a_{23} + a_{21}m_1 + (a_{22} - a_{33})m_2 - a_{32}m_2^2 - a_{31}m_1m_2 + f_2 \tag{81}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{z}\right) = -\frac{1}{z}(a_{31}m_1 + a_{32}m_2 + a_{33}) - \frac{b_3}{z^2} \tag{82}$$

where $f_1(z, m_1), f_2(z, m_2) \in \mathbb{R}$ are unmeasurable signals[1] defined as

$$f_1 \triangleq \frac{1}{z}(b_1 - b_3m_1) \tag{83}$$

$$f_2 \triangleq \frac{1}{z}(b_2 - b_3m_2)\,. \tag{84}$$

The IBSE objective in this section is to determine the unmeasurable state $z(t)$ of the perspective vision system described by (72) and (76). From (76) and the fact that $m_1(t)$ and $m_2(t)$ are measurable, it is clear that if $z(t)$ is identified, then the complete Euclidean coordinate of the feature can be determined. To achieve this objective, an observer is constructed based on the unmeasurable image-space dynamics for $m(t)$. To quantify the performance of the observer, a measurable observer estimation error signal, denoted by $e(t) \in \mathbb{R}^2$, is defined as

$$e \triangleq [e_1, e_2]^{\mathrm{T}} = [m_1 - \widehat{m}_1, m_2 - \widehat{m}_2]^{\mathrm{T}}\,, \tag{85}$$

where $\widehat{m}(t) \triangleq \left[\hat{y}_1(t), \hat{y}_2(t)\right]^{\mathrm{T}} \in \mathbb{R}^2$ denotes a subsequently designed observer signal. To facilitate the ob-

[1] The signals $f_1(z, m_1)$, and $f_2(z, m_2)$ are unmeasurable due to a dependence on the unmeasurable state $z(t)$

server design, a filtered observation error signal, denoted by $r(t) \in \mathbb{R}^2$, is designed as

$$r \triangleq [r_1, r_2]^{\mathrm{T}} = [\dot{e}_1 + \alpha_1 e_1, \dot{e}_2, +\alpha_2 e_2]^{\mathrm{T}} , \tag{86}$$

where $\alpha_1, \alpha_2 \in \mathbb{R}$ denote positive constant gains. Based on the dynamics in (80) and (81) and the definitions introduced in (85) and (86), it is clear that $r(t)$ is unmeasurable due to the fact that $\dot{m}(t)$ is a function of the unmeasurable disturbance terms $f_1(z, m_1)$ and $f_2(z, m_2)$. The subsequent development will follow the strategy given in [23], where the design estimates $f_1(z, m_1)$ and $f_2(z, m_2)$ based on the strategy that if the mismatch between the estimates and the disturbance terms $f_1(z, m_1)$ and $f_2(z, m_2)$ can be driven to zero, then $z(t)$ can be identified by exploiting the fact that $b_i(t) \forall i = 1, 2, 3$ and the states $m_1(t)$ and $m_2(t)$ are measurable. Specifically, from (83) and (84), the inverse of the square of $z(t)$ can be determined as

$$\left(\frac{1}{z}\right)^2 = \frac{f_1^2 + f_2^2}{(b_1 - b_3 m_1)^2 + (b_2 - b_3 m_2)^2} . \tag{87}$$

For (87) to be valid, it is clear that the following observability condition must be satisfied

$$(b_1 - b_3 m_1)^2 + (b_2 - b_3 m_2)^2 > 0 . \tag{88}$$

That is, $z(t)$ can be identified once the mismatch between the disturbance terms $f_1(z, m_1)$ and $f_2(z, m_2)$ and the respective estimates are driven to zero.

By taking the time-derivative of (85) the following error dynamics can be obtained for $e(t)$

$$\dot{e} = \dot{y} - \dot{\hat{y}} . \tag{89}$$

Based on the structure of (80), (81), and (89), the elements of the observer signal $\hat{y}(t)$ are designed as [23]

$$\dot{\hat{m}}_1 = a_{13} + (a_{11} - a_{33}) m_1 + a_{12} m_2 - a_{31} m_1^2 - a_{32} m_1 m_2 + \hat{f}_1 \tag{90}$$

$$\dot{\hat{m}}_2 = a_{23} + a_{21} m_1 + (a_{22} - a_{33}) m_2 - a_{32} m_2^2 - a_{31} m_1 m_2 + \hat{f}_2 \tag{91}$$

$$\dot{\hat{f}}_1 = -(k_{s1} + \alpha_1)\hat{f}_1 + \gamma_1 sgn(e_1) + \alpha_1 k_{s1} e_1 \tag{92}$$

$$\dot{\hat{f}}_2 = -(k_{s2} + \alpha_2)\hat{f}_2 + \gamma_2 sgn(e_2) + \alpha_2 k_{s2} e_2 . \tag{93}$$

Under the assumptions that $a_{i,j}(t), b_i(t) \forall i, j = 1, 2, 3$ are known motion parameters; the measured input signals $m_1(t)$ and $m_2(t)$, and that the unmeasurable state $1/z(t)$ are all bounded functions of time; that the parameters $a_{i,j}(t)$ are first order differentiable; and that the parameters $b_i(t)$ are second order differentiable, then the estimate $\hat{z}(t)$ will asymptotically converge to the true value $z(t)$.

Future Directions

The field of IBSE is still an evolving area that has achieved prominence only in the past twenty years. During this time, there have been several major shifts in research, including linear estimation methods, Kalman filtering methods and, more recently, nonlinear estimator/observer methods. However, as new methods are introduced, previous methods continue to be popular and widely used in a variety of problems. It seems likely that these methods will continue to be used, rather than giving way to any entirely new methodology.

There are several open problems attracting attention, mostly rooted in the limitations of IBSE. One open area includes camera field of view issues and moving beyond the pinhole camera model. These two problems are linked in the use of omnidirectional cameras, which have received recent attention. A second open area involves the use of multiple moving cameras and moving objects. Using methods discussed in this chapter, each camera can estimate the relative pose and velocity between itself and visible objects in its FOV, including other cameras and/or the moving platforms (e. g., a vehicle or robot) on which they are mounted. Given overlapping camera FOV's or cameras visible to each other, and communication between cameras, it is possible to relate relative pose of all objects and cameras, even if they are not in the field of view. However, it is uncertain how to maintain such a network of state estimates or how errors might propagate through such a network.

There is also research to reduce the necessary a priori knowledge needed to perform IBSE. Due to the loss of depth information in imaging, some a priori information will always be necessary to recover the correct scale of size and/or motion. Nevertheless, it is desirable to require as little extra information as possible. For example, there are efforts to eliminate complete knowledge of the body dynamics for the range identification problem discussed in Sect. "Range Identification", reducing the necessary knowledge to a subset of the motion parameters.

Bibliography

1. Kruppa E (1913) Zur Ermittlung eines Objektes aus zwei Perspektiven mit innerer Orientierung. Hölder, Wien
2. Longuet-Higgins H (1981) A computer algorithm for reconstructing a scene from two projections. Nature 293:133–135
3. Huang T, Faugeras O (1989) Some properties the e matrix in two-view motion estimation. IEEE Trans Pattern Anal Mach Intell 11:1310–1312
4. Philip J (1996) A non-iterative algorithm for determining all essential matrices corresponding to five point pairs. Photogramm Rec 15:589–99

5. Nister D (2004) An efficient solution to the five-point relative pose problem. IEEE Trans Pattern Anal Mach Intell 26:756–770
6. Faugeras O, Lustman F (1988) Motion and structure from motion in a piecewise planar environment. Int J Pattern Recognit Artif Intell 2:485–508
7. Spetsakis M, Aloimonos J (1989) Optimal motion estimation. Proc Workshop Vis Motion, Irvine, pp 229–237
8. Weng J, Ahuja N, Huang T (1993) Optimal motion and structure estimation. IEEE Trans Pattern Anal Mach Intell 15:864–884
9. Ma Y, Košecká J, Sastry S (2001) Optimization criteria and geometric algorithms for motion and structure estimation. Int J Comput Visi 44:219–249
10. Caballero F, Merino L, Ferruz J, Ollero A (2006) Improving vision-based planar motion estimation for unmanned aerial vehicles through online mosaicing. Proc IEEE Int Conf Robotics Autom, Orlando, pp 2860–2865
11. Kaiser K, Gans N, Dixon W (2007) Localization and control an aerial vehicle through chained, vision-based pose reconstruction. Proc American Control Conference, New York, pp 5934–5939
12. Ma L, Chen Y, Moore KL (2005) Range identification for perspective dynamic systems with 3D imaging surfaces. Proc American Control Conference, Portland, pp 3671–3675
13. Gupta S, Aiken D, Hu G, Dixon WE (2006) Lyapunov-based range and motion identification for a nonaffine perspective dynamic system. Proc American Control Conference, Minneapolis, pp 4471–4476
14. Soatto S, Frezza R, Perona P (1996) Motion estimation via dynamic vision. IEEE Trans Automat Control 41:393–413
15. Broida T, Chellappa R (1991) Estimating the kinematics and structure a rigid object from a sequence monocular images. IEEE Trans Pattern Anal Mach Intell 13:497–513
16. Azarbayejani A, Pentland AP (1995) Recursive estimation motion, structure, and focal length. IEEE Trans Pattern Anal Mach Intell 17:562–575
17. Kano H, Ghosh BK, Kanai H (2001) Single camera based motion and shape estimation using extended kalman filtering. Math Comput Model 34:511–525
18. Chiuso A, Favaro P, Jin H, Soatto S (2002) Structure from motion causally integrated over time. IEEE Trans Pattern Anal Mach Intell 24:523–535
19. Chitrakaran V, Dawson DM, Dixon WE, Chen J (2005) Identification a moving object's velocity with a fixed camera. Automatica 41:553–562
20. Jankovic M, Ghosh B (1995) Visually guided ranging from observations points, lines and curves via an identifier based nonlinear observer. Syst Control Lett 25:63–73
21. Chen X, Kano H (2002) A new state observer for perspective systems. IEEE Trans Automat Contr 47:658–663
22. Ma L, Chen Y, Moore K (2004) Range identification for perspective dynamic systems using linear approximation. Proc IEEE Int Conf Robotics and Automation, New Orleans, pp 1658–1663
23. Dixon WE, Fang Y, Dawson DM, Flynn TJ (2003) Range identification for perspective vision systems. IEEE Trans Automat Contr 48:2232–2238
24. Ma L, Chen Y, Moore KL (2004) Range identification for perspective dynamic system with single homogeneous observation. Proc IEEE Int Conf Robotics and Automation, New Orleans, pp 5207–5212
25. Weng J, Ahuja N, Huang T (1992) Motion and Structure from Image Sequences. Springer, New York
26. Ma Y, Soatto S, Kosecká J, Sastry S (2004) An Invitation to 3-D Vision. Springer, New York
27. Faugeras O (1993) Three-Dimensional Computer Vision. MIT Press, Cambridge
28. Hartley R, Zisserman A (2003) Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge
29. Bouguet J Complete camera calibration toolbox for matlab. http://www.vision.caltech.edu/bouguetj/calib_doc/index.html
30. Sepp W, Fuchs S The dlr camera calibration toolbox. http://www.dlr.de/rm/desktopdefault.aspx/tabid-1524/
31. Tsai R (1987) A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. IEEE Trans Robotics Automat 3:323–344
32. Tsai R (1989) Synopsis recent progress on camera calibration for 3D machine vision. MIT Press, Cambridge
33. Robert L (1996) Camera calibration without feature extraction. CVIU: Comput Vis Image Underst 63:314–325
34. Harris C, Stephens M (1988) A combined corner and edge detector. Alvey Vision Conference, vol 15, pp 247–151
35. Smith S, Brady J (1997) Susan – a new approach to low level image processing. Int J Comput Vis 23:45–78
36. Lowe D (1999) Object recognition from local scale-invariant features. Proc IEEE Int Conf Computer Vision, Toronto, pp 1150–1157
37. Beucher S, Lantuejoul C (1979) Use watersheds in contour detection. Proc Int Workshop on Image Processing, Real-Time Edge and Motion Detection/Estimation, Rennes
38. Niethammer M, Tannenbaum A, Angenent S (2006) Dynamic active contours for visual tracking. IEEE Trans Automat Control 51:562–579
39. Canny J (1986) A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell 8:679–698
40. Ballard D (1987) Generalizing the Hough transform to detect arbitrary shapes. Morgan Kaufmann Publishers, San Francisco
41. Zhang Z, Hanson A (1996) 3D reconstruction based on homography mapping. Proc ARPA Image Understanding Workshop Palm Springs CA
42. Luong Q, Faugeras O (1996) The fundamental matrix: Theory, algorithms, and stability analysis. Int J Comput Vis 17:43–75
43. Boufama B, Mohr R (1995) Epipole and fundamental matrix estimation using virtual parallax. Proc Int Conf Computer Vision, Boston, pp 1030–1036
44. Malis E, Chaumette F (2000) 2 1/2D visual servoing with respect to unknown objects through a new estimation scheme camera displacement. Int J Comput Vis 37:79–97
45. Fang Y, Dixon W, Dawson D, Chawda P (2005) Homography-based visual servo regulation mobile robots. IEEE Trans Syst Man Cybern 35:1041–1050
46. Dupree K, Gans N, Mackunis W, Dixon W (2007) Euclidean feature tracking for a rotating satellite. Proc American Control Conference, New York, pp 3874–3879
47. MacKunis W, Gans N, Kaiser K, Dixon WE (2007) Unified tracking and regulation visual servo control for wheeled mobile robots. Proc IEEE Multi-Conf Syst Control, Singapore, pp 88–93
48. Baillard C, Zisserman A (1999) Automatic reconstruction planar models from multiple views. Proc IEEE Conf Computer Vision and Pattern Recognition, Toronto, pp 559–565
49. Okada K, Kagami S, Inaba M, Inoue H (2001) Plane segment finder: algorithm, implementation and applications. Proc IEEE Int Conf Robotics and Automation, Seoul, pp 2120–2125

50. Zhuang X, Haralick R (1984) Rigid body motion and the optical flow image. Proc Int Conf Artificial Intelligence Applications, Denver, pp 366–375
51. Ma Y, Košecká J, Sastry S (2000) Linear differential algorithm for motion recovery: A geometric approach. Int J Comput Vis 36:71–89
52. Malis E, Chaumette F, Boudet S (1999) 2-1/2D visual servoing. IEEE Trans Robot Automat 15:238–250
53. Xian B, de Queiroz MS, Dawson DM (2004) A continuous control mechanism for uncertain nonlinear systems. Optimal Control, Stabilization, and Nonsmooth Analysis. Lecture Notes in Control and Information Sciences, vol 301. Springer, New York, pp 251–262
54. Sorenson H (1985) Kalman Filtering: theory and application. IEEE, New York
55. Grewal M, Andrews A (1993) Kalman filtering: theory and practice. Prentice-Hall, Upper Saddle River
56. Shuster M (1993) A survey attitude representations. J Astronaut Sci 41:439–518
57. Karagiannis D, Astolfi A (2005) A new solution to the problem range identification in perspective vision systems. IEEE Trans Automat Control 50:2074–2077

Immunecomputing

Jon Timmis[1,2]
[1] Department of Electronics, University of York, York, UK
[2] Department of Computer Science, University of York, York, UK

Article Outline

Glossary
Definition of the Subject
Introduction
What Is an Artificial Immune System?
Current Artificial Immune Systems Biology and Basic Algorithms
Alternative Immunological Theories for AIS
Emerging Methodologies in AIS
Future Directions
Bibliography

Glossary

Glossary based on [28]:

Affinity Measure or tightness of the binding between an antigen combining site and an antigenic determinant; the stronger the binding, the higher the affinity.

Antigen Any substance that when introduced into the body, is capable of inducing an immune response.

Antigen presenting cells (APC) B-cells, cells of the monocyte Lineage (including macrophages as well as dendritic cells), and various other body cells that present antigen in a form that B- and T-cells can recognize.

Antibody A soluble protein molecule produced and secreted by B-cells in response to an antigen. Antibodies are usually defined in terms of their specific binding to an antigen.

B cell White blood cells expressing immunoglobulin molecules on its surface. Also known as B-lymphocytes, they are derived from the bone marrow and develop into plasma cells that are the main antibody secretors.

Clonal selection theory A theory that states that the specificity and diversity of an immune response are the result of selection by antigen of specifically reactive clones from a large repertoire of preformed lymphocytes, each with individual specificities.

Dendritic cell Set of antigen-presenting cells (APCs) present in lymph nodes, spleen and at low levels in blood, which are particularly active in stimulating T-cells.

Lymph node Small organs of the immune system, widely distributed throughout the body and linked by lymphatic vessels.

Lymphocyte White blood cell found in blood, tissue, and in lymphoid organs.

Major histocompatability A group of genes encoding polymorphic.

Complex (MHC) Cell-surface molecules (MHC class I and II) that are involved in controlling several aspects of the immune response. MHC genes code for self-markers on all body cells and play a major role in transplantation rejection.

Pathogen A microorganism that causes disease.

T Cell White blood cell that orchestrate and/or directly participate in the immune defenses.

Definition of the Subject

Immunecomputing, or Artificial Immune Systems (AIS), has recently emerged as a computational intelligence approach that show great promise. Inspired by the complexity of the immune system, computer scientists and engineers have created systems that in some way mimic or capture certain computationally appealing properties of the immune system, with the aim of building more robust and adaptable solutions. AIS have been defined by [28] as:

> "adaptive systems, inspired by theoretical immunology and observed immune functions, principle and models, which are applied to problem solving".

However, in order to build AIS an interdisciplinary approach is required that employs modeling of immunology (both mathematical and computational) in order to understand the underlying complexity inherent within the immune system. AIS do not rival their natural counterparts, they do not exhibit the same level of complexity or even perform the same function, but they do capture essential properties of the immune systems that are making them a competitive computational intelligence paradigm.

Introduction

The immune system is a complex system that undertakes a myriad of tasks. The abilities of the immune system have helped to inspire computer scientists to build systems that *mimic*, in some way, various properties of the immune system. This field of research, Artificial Immune Systems (AIS), has seen the application of immune inspired algorithms to problems such as robotic control [72], network intrusion detection [37,67], fault tolerance [5,16], bioinformatics [20,77] and machine learning [70,71,106], to name a few. To many, trying to mimic how the immune system operates in a computer may seem an unusual thing to do, why then would people in computing wish to do this? The answer is that, from a computational point of view, the immune system has many desirable properties that they would like their computer systems to possess. These properties are such things as robustness, adaptability, diversity, scalability, multiple interactions on a variety of timescales and so on.

The origins of AIS has its roots in the early theoretical immunology work of Farmer, Perelson and Varela [35,79,101]. These works investigated a number of theoretical immune network models proposed to describe the maintenance of immune memory in the absence of antigen. Whilst controversial from an immunological perspective, these models began to give rise to an interest from the computing community. The most influential people at crossing the divide between computing and immunology in the early days were Hugues Bersini and Stephanie Forrest. It is fair to say that some of the early work by Bersini [10,11] was very well rooted in immunology, and this is also true of the early work by Forrest [36,53]. It was these works that formed the basis of a solid foundation for the area of AIS . In the case of Bersini, he concentrated on the immune network theory, examining how the immune system maintained its memory and how one might build models and algorithms mimicking that property. With regards to Forrest, her work was focused on computer security (in particular network intrusion detection) [37,55] and formed the basis of a great deal of further research by the community on the application of immune inspired techniques to computer security.

At about the same time as Forrest was undertaking her work, other researchers began to investigate the nature of learning in the immune system and how that might by used to create *machine learning* algorithms [19]. They had the idea that it might be possible to exploit mechanisms of the immune system (in particular the immune network) in learning systems, so they set about doing a proof of concept [19]. Initial results were very encouraging, and they built on their success by applying the immune ideas to the classification of DNA sequences as either promoter or non-promoter classes, [56] and the detection of potentially fraudulent mortgage applications [57].

The work of Hunt and Cook spawned more work in the area of immune network based machine learning over the next few years, notably in [91] where the Hunt and Cook system was totally rewritten, simplified and applied to unsupervised learning (very similar to cluster analysis). Concurrently, similar work was carried out by [31,32], who developed algorithms for use in function optimization and data clustering (the details of these are described in more details later in the chapter). The work of Timmis on machine learning spawned yet more work in the unsupervised learning domain, in trying to perform dynamic clustering (where the patterns in the input data move over time). This was met with some success in works such as [76,108]. At the same time, using ideas other than the immune network theory, work by [50] used immune inspired associative memory ideas to track moving targets in databases.

In the supervised learning domain, very little happened until work by [102] (later augmented in [106]) developed an immune based classifier known as AIRS. The system developed by Watkins was then adapted into a parallel and distributed learning system in [103], and has shown itself to be one of the real success stories of immune inspired learning [45,46,105].

In addition to the work on machine learning, there has been plenty of other activity in AIS over the years. To outline all the applications of AIS and developments over the past 10 years would take a long time, and there are some good review papers in the literature, thus the reader is directed those [23,28,39,95]. In addition to these works, [52] investigated the application areas AIS have been applied to, and considered the contribution AIS have made to these areas. Their survey of AIS is not exhaustive, but attempts to produce a picture of the general areas to which they have been applied. Of the 97 papers reviewed, 12 categories were identified to reflect the natural groupings of the papers. These were, in the order of most papers first:

clustering/classification, anomaly detection (e. g. detecting faults in engineering systems), computer security, numerical function optimization, combinatoric optimization (e. g. scheduling), learning, bioinformatics, image processing, robotics (e. g. control and navigation), adaptive control systems, virus detection and web mining. Hart and Timmis go on to note that these categories can be summarized into three general application areas of learning, anomaly detection and optimization.

Work in [88] details an alternative approach to the use of immune metaphors and present a deterministic immune network approach, that is in stark contrast to the work presented here. This work has shown to be exceptionally competitive when compared to other computational intelligence approaches [89,90]. Due to a growing amount of work conducted on AIS, the International Conference on Artificial Immune Systems (ICARIS) conference series was started in 2002[1] and has operated in subsequent years [12,58,78,93,98]. This is the best source of reference material to read in order to grasp the variety of application areas of AIS, and also the developments in algorithms and the more theoretical side of AIS.

This remaining article is organized as follows: in Sect. "What Is an Artificial Immune System?" first discuss the current perception of what AIS are, and we do this in terms of a simple engineering framework; in Sect. "Current Artificial Immune Systems Biology and Basic Algorithms" we provide a simple overview of the immunology that has served to inspire the development of immune inspired systems to date, and this is coupled with an outline of the basic algorithms that have been use in AIS to date; in Sect. "Alternative Immunological Theories for AIS" we turn our attention to alternative immunology (away from the more classic described in the previous section) and focus on the danger theory approach which exploits a fundamental different analogy from immunology that other AIS have done to date and discuss the cognitive immune paradigm as an alternative approach to the development of AIS; Sect. "Emerging Methodologies in AIS" reviews new approaches to the development of AIS in the context of a conceptual framework and finally in Sect. "Future Directions" we review some of the possible challenges and directions that AIS might follow.

What Is an Artificial Immune System?

In an attempt to create a common basis for AIS, work in [28] proposed the idea of a framework for engineering AIS. They argued the case for such a framework as the existence of similar frameworks in other biologically inspired approaches, such as artificial neural networks (ANN) and evolutionary algorithms (EAs), has helped considerably with the understanding and construction of such systems. For example, de Castro and Timmis [28] consider a set of artificial neurons, which can be arranged together to form an artificial neural network. In order to acquire knowledge, these neural networks undergo an adaptive process, known as learning or training, which alters (some of) the parameters within the network. Therefore, they argued that in a simplified form, a framework to design an ANN is composed of: a set of artificial neurons, a pattern of interconnection for these neurons, and a learning algorithm. Similarly, they argued that in evolutionary algorithms, there is a set of artificial chromosomes representing a population of individuals that iteratively suffer a process of reproduction, genetic variation, and selection. As a result of this process, a population of evolved artificial individuals arises. A framework, in this case, would correspond to the genetic representation of the individuals of the population, plus the procedures for reproduction, genetic variation, and selection. Therefore, they proposed that a framework to design a biologically inspired algorithm requires, at least, the following basic elements:

- A representation for the components of the system
- A set of mechanisms to evaluate the interaction of individuals with the environment and each other. The environment is usually simulated by a set of input stimuli, one or more fitness function(s), or other means
- Procedures of adaptation that govern the dynamics of the system, i. e., how its behavior varies over time

This framework can be thought of as a layered approach such as the specific framework for engineering AIS of [28] shown in Fig. 1. This framework follows the three basic elements for designing a biologically inspired algorithm just described, where the set of mechanisms for evaluation are the affinity measures and the procedures of adaptation are the immune algorithms. In order to build a system such as an AIS, one typically requires an application domain or target function. From this basis, the way in which the components of the system will be represented is considered. For example, the representation of network traffic may well be different than the representation of a real time embedded system. In AIS, the way in which something is represented is known as *shape space*. There are many kinds of shape space, such as Hamming, real valued and so on, each of which carries it own bias and should be selected with care [38]. Once the representation has been chosen, one or more affinity measures are used to quantify the interactions of the elements of the system. There are many possible affinity measures (which are partially dependent

[1] http://www.artificial-immune-systems.org

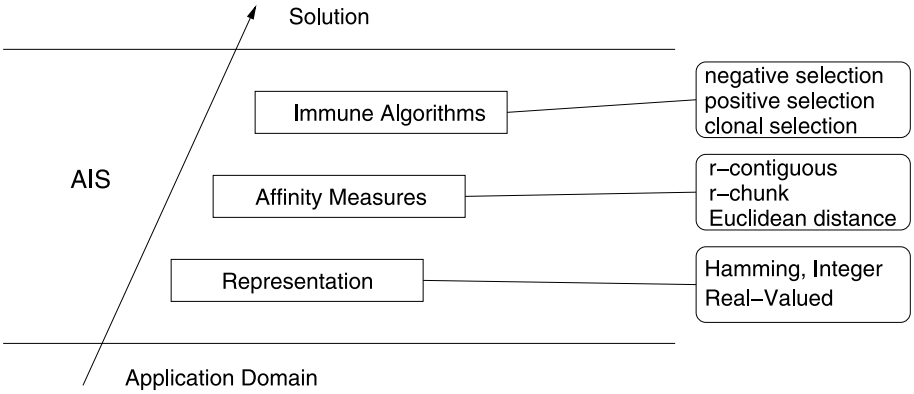

Immunecomputing, Figure 1
AIS Layered Framework adapted from [28]

upon the representation adopted), such as Hamming and Euclidean distance metrics. Again, each of these has its own bias, and the affinity function must be selected with great care, as it can affect the overall performance (and ultimately the result) of the system [38]. This was also recently shown experimentally in the case of immune networks, where the affinity function affected the overall outcome of the shape of the network [49,51] The final layer involves the use of algorithms, which govern the behavior (dynamics) of the system. Such algorithms include those based on the following immune processes: negative and positive selection, clonal selection, bone marrow, and immune network algorithms.

Current Artificial Immune Systems Biology and Basic Algorithms

The main developments within AIS, have focused on three main immunological theories: clonal selection, immune networks and negative selection. Researchers in AIS have concentrated, for the most part, on the *learning* and *memory* mechanisms of the immune system inherent in clonal selection and immune networks, and the negative selection principle for the generation of *detectors* that are capable of classifying changes in *self*. In this section, we review the immunology that has been capitalized on by the AIS community. We outline the three main immunological theories noted above that have acted as a source of inspiration. At each stage, we review a simple AIS approach that has extracted some feature from that theory. It is worth noting that, although not covered here, a large effort is currently being made in the AIS community into exploring other immune ideas and mechanisms such as danger theory and innate immunity.

Immunity

The vertebrate immune system (the one which has been used to inspire the vast majority of AIS to date) is composed of diverse sets of cells and molecules. These work in collaboration with other systems, such as the neural and endocrine, to maintain a steady state of operation within the host: this is termed *homeostasis*. The role of the immune system is typically viewed as one of protection from infectious agents such as viruses, bacteria, fungi and other parasites. On the surface of these agents are antigens that allow the identification of the invading agents (pathogens) by the immune cells and molecules, which in turn provoke an immune response. There are two basic types of immunity, innate and adaptive. Innate immunity is not directed towards specific pathogens, but against any pathogen that enter the body. The innate immune system plays a vital role in the initiation and regulation of immune responses, including adaptive immune responses. Specialized cells of the innate immune system evolved so as to recognize and bind to common molecular patterns found only in microorganisms. However, the innate immune system is by no means a complete solution to protecting the body.

Adaptive, or acquired immunity, is directed against specific invaders, with adaptive immune cells being modified by exposure to such invaders. The adaptive immune system mainly consists of lymphocytes, which are white blood cells, more specifically B and T cells. These cells aid in the process of recognizing and destroying specific substances. Any substance that is capable of generating such a response from the lymphocytes is called an antigen or immunogen. Antigens are not the invading microorganisms themselves; they are substances such as toxins or enzymes in the microorganisms that the immune system

considers foreign. Adaptive immune responses are normally directed against the antigen that provoked them and are said to be antigen-specific.

Natural Clonal Selection

The clonal selection theory (CST) [15] is the theory used to explain the basic response of the adaptive immune system to an antigenic stimulus. It establishes the idea that only those cells capable of recognizing an antigenic stimulus will proliferate, thus being selected against those that do not. Clonal selection operates on both T cells and B cells. In the case of B cells, when their antigen receptors (antibodies) bind with an antigen, the B cell becomes activated and begins to proliferate producing new B cell clones that are an exact copy of the parent B cell. The clones then undergo somatic hypermutation and produce antibodies that are specific to the invading antigen [9]. After proliferation, B cells differentiate into *plasma cells* or long-lived B *memory cells*. Plasma cells produce large amounts of *antibodies* which will attach themselves to the antigen and act as a type of *tag* for other immune cells to pick up on and remove from the system. This whole process is known as *affinity maturation*.

Memory cells help the immune system to be protective over periods of time. In the normal course of the evolution of the immune system, an organism would be expected to encounter a given antigen repeatedly during its lifetime. The initial exposure to an antigen that stimulates an adaptive immune response is handled by a small number of B cells, each producing antibodies of different affinity. Storing some high affinity antibody producing cells (memory cells) from the first infection, so as to form a large initial specific B cell sub-population for subsequent encounters, considerably enhances the effectiveness of the immune response to secondary encounters. Such a strategy ensures that both the speed and accuracy of the immune response becomes successively stronger after each infection.

Autoimmunity is the term used to describe the existence of antigen receptors that recognize the body's own molecules, or self-antigens. According to the CST, immune specificity is a property of immune receptors. When a non-self antigen is detected, a suitable immune response is elicited and the antigen is destroyed. Thus, the recognition of self-antigen is forbidden, and self-reacting receptors must be deleted.

Artificial Clonal Selection

Work in [30,32] proposes an optimization algorithm, known as CLONALG, inspired by the clonal selection process, as outlined in the previous section. Given a function F, a population of candidate solutions (antibodies) are evolved to either minimize or maximize the function. Each member of this population is a vector, in a certain shape space, which maps values to the parameters of the function F. CLONALG exploits the cloning, mutation and selection mechanisms of clonal selection, to effectively evolve a set of memory cells that contain candidate solutions to the function F.

CLONALG operates via the following procedure. A population P is initialized with random vectors, where P is set of candidate solutions for the given function. Each member of P is evaluated against the function, and the highest affinity n number are selected for cloning, where affinity can be measured as the distance to the optimal value. Clones are produced at a rate proportional to the affinity (so the better the affinity, the more clones are produced). Each clone is subject to a mutation rate, which is inversely proportional to the affinity. These clones are added to P and then the n highest affinity are selected to remain in the population. A number of low affinity members are then removed from the population and replaced with the same number of randomly generated members. This process is repeated until some convergence criteria is satisfied, or a fixed number of iterations has been performed.

Experimentally, CLONALG has been shown to perform well on standard benchmark tests for optimization problems [32]. However, it has not been reported in the literature that CLONALG itself outperforms any well known technique. Other algorithms similar to CLONALG exist in the literature, such as [65] and [20], with comparative studies showing that whilst CLONALG is effective, better results can be obtained with more specialized versions of the algorithm [20,77]. Indeed, recent work by [21] has shown that a clonal selection based algorithm using a special aging operator can perform as well as the state-of-art on certain protein folding problems.

Clonal selection based algorithms have also been developed for dynamic environments, reporting good performance [40,66,70]. CLONALG has also been adapted for simple pattern recognition problems, but the results from that work are less conclusive [107]. It has also been adapted for more sophisticated learning systems where results are very encouraging indeed for static learning [45,104,105] and for dynamic learning [80].

Immune Networks

In a landmark paper for its time, [59] proposed that the immune system is capable of achieving immunological memory by the existence of a mutually reinforcing net-

work of B cells. This network of B cells occurs due to the ability of paratopes (molecular portions of an antibody) located on B cells, to match against idiotopes (other molecular portions of an antibody) on other B cells. The binding between idiotopes and paratopes has the effect of stimulating the B cells. This is because the paratopes on B cells react to the idiotopes on similar B cells, as it would an antigen. However, to counter the reaction there is a certain amount of suppression between B cells which acts as a regulatory mechanism. This interaction of B cells due to the network, was said to contribute to a *stable* memory structure, and account for the retainment of memory cells, even in the absence of antigen. This theory was refined and formalized in successive works by [35,79] and combined with work by [13] was very influential in development of the immune network based AIS such as [56,76,96,97].

Artificial Immune Networks

Based on the work of CLONALG, an algorithm known as aiNet was proposed in [31]. aiNet is a simple extension of CLONALG (described above), but exploits interactions between B cells according to the immune network theory. The main difference between the two approaches, is that after new clones are integrated into the population, a network suppression function is employed throughout the population to remove cells that have similar affinities[2] this facilitates the maintenance of diversity within the population. Recent work by [83] has shown that aiNET performs better on data that has a more uniform underlying distribution, due to the nature of how aiNET performs the suppression. However, this work did focus on the initial version of aiNET, and there are recent variants of aiNET that seem to perform much better.

aiNet was initially designed for data clustering, but has been extended over the years, as a hierarchical clustering tool in [29] and through hybridization with fuzzy systems methods by [14]. In the last paper, aiNet was augmented to take into account an adaptive radius measure instead of a fixed radius for B cell matching. This lead to a much improved version of aiNet, being able to achieve better separation of the data, forming clusters in less time. Work by [27] adapted aiNet for multi-modal function optimization. In that paper, aiNet was also modified to be applied to the same optimization problems as CLONALG, and was shown to have greatly improved performance over CLONALG, but this is not as comparable to other clonal selection based systems [94,99]. However, it was recently identified that if careful thought was given to the optimization problem, the basic aiNet algorithm can be augmented to give significant gains in performance [3].

[2]It should be noted that this is a slight departure from the immune network theory, where both suppression and stimulation occur between cells

Negative Selection

Negative selection is a process of *selection* that takes place in the thymus gland. T cells are produced in the bone marrow and before they are released into the lymphatic system, undergo a maturation process in the thymus gland. The maturation of the T cells is conceptually very simple. T cells are exposed to self-proteins in a binding process. If this binding activates the T cell, then the T cell is killed, otherwise it is allowed into the lymphatic system. This process of *censoring* prevents cells that are reactive to *self* from entering the lymph system, thus endowing (in part) the host's immune system with the ability to distinguish between self and non-self agents.

Artificial Negative Selection

The negative selection principle inspired [36] to propose a negative selection algorithm to detect data manipulation caused by computer viruses. The basic idea is to generate a number of detectors in the complementary space and then to apply these detectors to classify new (unseen) data as self (no data manipulation) or non-self (data manipulation). In the negative selection algorithm as proposed by Forrest et al. we can define self as a set **S** of elements of length l in shape-space. Then generate a set **D** of detectors, such that each fails to match any element in **S**. With these detectors, monitor a continual data stream for any changes, by continually matching the detectors in **D** against the stream. This work spawned a great deal of investigations into the use of negative selection for intrusion detection, with early work meeting with some success [37], and this being built on in later years [7,34,54,55]. The work on negative selection has been dominate in AIS. A great deal of work has gone into investigating various alternatives of representations [26,42,44], techniques for estimating detector coverage [43,60,61,62] and applications of said technique [6,24,25,33,63,64,68,69,81] to name only a few. However, later works began to highlight certain limitations of the approach [84,85,86,87] with regards to scalability issues and applicability of the technique to classification. In addition, work in [38] outline the need to consider carefully the application domain when developing AIS, and they give particular attention to negative selection. They review the role AIS have played in the development of a number of machine learning tasks, including that of negative selection. However, Freitas and Timmis point out that there is a lack of appreciation for pos-

sible inductive bias within algorithms and positional bias within the choice of representation and affinity measures that comes from not carefully applying not only negative selection, but other algorithms as well.

Alternative Immunological Theories for AIS

Innate Immune System

In recent years there has been a growing interest in the mechanisms of innate immune system in immunology [41]. up to this point, AIS had concentrated soley on the adaptive immune systems, but the Danger theory proposed by [73,74] has caught the interest of the AIS practitioner in recent years as a compliment to the adaptive.

The Danger theory attempts to explain the nature and workings of the immune response in a way different to the more traditional clonal selection view. Matzinger criticizes this idea, as she states that observations demonstrate that it may sometimes be necessary for the body to attack itself and conversely the immune system may not attack cells it knows to be foreign (this is not possible under the classical clonal selection theory). Matzinger argues a more plausible way to describe the triggering of an immune response is a reaction to a stimulus the body considers harmful. This might be seen as a very small change but in reality this is real shift in thinking about how the immune system responds to pathogens. In essence, this model allows for foreign and immune cells to exist together, a situation impossible in the traditional standpoint. When under attack, cells dying unnaturally may release a danger signal, that disperses to cover a small area around that cell: a danger area. It is within this and only within this area that the immune system becomes active and will concentrate its attack against any antigen within it. There is still much debate in the immunological world as to whether the Danger theory is a plausible explanation for observed immune function, but it is proving to be an interesting theory none the less and if affecting the thinking in the AIS world. As we have discussed in the previous section, the idea of using the innate immune systems, in particular the danger theory, has started to become popular. The first to propose the idea was [1] in the context of using such danger theory ideas in the context of network intrusion detection. Here the authors discussed how one might introduce the notion of *danger areas* in networks which might be indicated by unusual behavior without having to define a-priori what those behaviors were (a large departure from the current way of network intrusion detection using AIS, where unusual behavior was used to train the system). This work was then extended to notably [47,48] where the dendritic cell algorithm was proposed and [8] where the idea of artificial tissue was proposed. Work in [48] describes how dendritic cells (DC), which is a cell considered to be part of the innate immune system, performs a function that in effect controls the adaptive immune response when under attack. What a DC does depends on the signals that it receives within the tissue: these might be danger signals, PAMPS (pathogenic associated molecular patterns), safe signals and inflammatory cytokines. The DC will mature into different states depending on the concentration of these signals, and the state of the DC influences the response of the T cell to which the DC is presenting the antigen. Based on these ideas, a DC inspired algorithm has been developed, and tested within a tissue environment developed in [100]. The application area was that of anomaly detection and the DC had to identify if certain types of behavior on a computer were anomalous or not: no predefined knowledge of what constitutes anomalous is required, but what is required is a definition of what constitutes dangerous behaviors for various variables. Results reported in [48] would seem to indicate that the DC algorithm is capable of identifying anomalous behavior (processes that were considered not be normal) over time. However, this is still quite preliminary work, and the system needs to be baselined against a state-of-art type system in order to fully see the contribution this approach can bring.

Cognitive Immunology

Much like the paradigm shift of danger theory, within AIS recent attention has been paid to the cognitive immune paradigm proposed by [18]. Notably work in [2] discusses how Cohen views the immune system as a cognitive system, capable of detection, cognition and decision making and argues that the primary role of the immune system is not protection (as is considered by say the clonal selection theory), but one of body maintenance. [2] say that in Cohen's view, removal of pathogen, is beneficial to the health of the body, and thus defense against pathogen is considered to be just a special case of body maintenance. In order to carry out body maintenance, the immune system must be able to detect the current state of the bodys tissues and elicit an appropriate response.

According to the clonal selection theory, immune specificity is a property of the somatically generated immune receptors of the T and B cells, which both initiates and regulates the immune response. Initiation is achieved via the binding between an antigen and a receptor that is specific to it. As stated by [2], Cohen, however, points out that immune receptors are intrinsically degenerate, i. e. they can bind more than one ligand. Immune specificity,

therefore, cannot be purely dependent on molecular binding as no one receptor can be specific to a single antigen. Instead, affinity, the strength of binding between a receptor and its ligand, is a matter of degree. Indeed, as a follow on to their work in [4] outline a simple computational model that demonstrates degenerate recognition in the context of a lymph node, and work in [75] has also investigated degeneracy in the context of pattern recognition and begun to develop high-level AIS as a result.

Emerging Methodologies in AIS

The methodology in which AIS are built has been addressed by work in [82]. This paper proposes a conceptual framework that allows for the development of more biologically grounded AIS, through the adoption of an interdisciplinary approach. As will be clear from the article, metaphors that have been employed have typically been simple, but somewhat effective. However, as proposed in [82], through greater interaction between computer scientists, engineers, biologists and mathematicians, better insights into the workings of the immune system, and the applicability (or otherwise) of the AIS paradigm will be gained. These interactions should be rooted in a sound methodology in order to fully exploit the synergy.

These interactions should be rooted in a sound methodology in order to fully exploit the synergy. The authors argue that rather than going straight from observing the biology and then to the development of an algorithm, a more principled approach is required to adequately capture the required properties of the biological system in the engineered counterpart. The methodology is one of abstraction, as seen in Fig. 2. The first step is observe the biological system through a means of experimentation and analysis. From there is is possible to create a mathematical model of the biosystem: a relatively detailed model of the system. However, these mathematical models may be too complex to solve, and therefore another level of abstraction is required to gain further insight into the interactions within the system and overall systems dynamics. Therefore, these mathematical models are then used to derive a more abstract computational model: the model can be executed and analyzed for properties that are desired in the engineered system we wish to construct, these can be encompassed into an analytical framework and design principles and high-level abstraction of algorithms and systems can be developed abstract from any application area. This is then instantiated in the application area, being tailored to the specific requirement of that application area. The result is well-grounded bio-inspired algorithm, that is understood better on a theoretical level and captures the *relevant* biological properties for the required application. As part of this process, the authors suggest a second stage of development and that is to create *meta-frameworks* which cut across a number of frameworks that have arisen as part of the initial development. At this stage it is possible to ask unifying questions across these systems that are concerned with complexity of the system, these questions are (1) Openness: how much openness is required in the system, biological systems do not stop computing, therefore should our computations stop? (2) Diversity: How much diversity is required in the system to attain the performance we require, how many different types of actors are needed? (3) Interaction: How should these agents interact? At what timescale and what should they communicate? (4) Structure: biological systems operate on a number of levels, how many levels are needed in our artificial systems, are suitable levels of a hierarchy required? and (5) Scale: Biological systems operate on vast scales, rather different from a typical immune inspired algorithm. How many actors re required in the system to achieve the desired complexity?

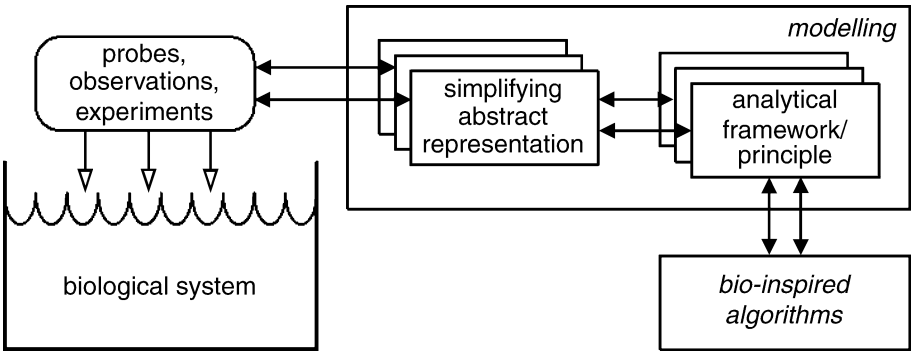

Immunecomputing, Figure 2
Conceptual Framework for the development of AIS adapted from [82]

Future Directions

With this may come a better understanding of how to apply AIS, and not fall into the traps highlighted by [38]. A recent paper by [52] highlight the fact that to date, the development of AIS has been *scattergun* i. e. many applications have been tried without a great deal of thought. Indeed, this paper provides a detailed overview of the many application areas that AIS have tried, and this will not be repeated here: the interested reader should consult that paper. The authors go on to propose a number of properties that they feel any AIS should have, and that these properties may help guide the type of application they could be applied to:

- "They will exhibit *homeostasis*.
- They will benefit from interactions between *innate* and *adaptive* immune models.
- They will consist of *multiple, interacting, communicating* components.
- Components can be easily and naturally *distributed*.
- They will be required to perform *life-long learning*" [52].

It apparent that, despite the success of some applications of AIS, all AIS to date fail to fully capture the complex operation of the immune system. What has changed is the increased scope of immunological theories that those working with AIS take inspiration from. For example, In their summaries of the future for AIS, both [39] and [52] point towards an increased emphasis on the innate and homeostatic functions of the immune system as possible areas for AIS exploitation. In addition to the increased scope of AIS, there has been a recent and healthy rise in investigating the theoretical workings of various immune algorithms [17,22,85]. In a recent position paper, [92] argues that the area of AIS has reached something of an impasse. They discusses a number of challenges to the AIS community which they believe will stimulate discussion and help move the area forward:

> "Challenge 1: To Develop Novel and Accurate Metaphors and be a Benefit to Immunology. Typically naive approaches to extracting metaphors from the immune system have been taken. This has occurred as an accident of history, and AIS has slowly drifted away from its immunological roots. Time is now ripe for greater interaction with immunologists and mathematicians to undertake specific experimentation and create useful models, all of which can be used as a basis for abstraction into powerful algorithms.
>
> Challenge 2: To Develop a Theoretical basis for AIS. Much work on AIS has concentrated on simple extraction of metaphors and direct application. Despite the creation of a framework for developing AIS, it still lacks significant formal and theoretical underpinning. AIS have been applied to a wide variety of problem domains, but a significant effort is still required to understand the nature of AIS and where they are best applied. For this, a more theoretical understanding is required.
>
> Challenge 3: To Consider the Application of AIS. Work to date in the realm of AIS has mainly concentrated on what other paradigms do, such as simple optimization, learning and the like. This has happened as an accident of history and whilst productive, the time is here to look for the killer application of AIS, or, if not that radical, then applications where the benefit of adopting the immune approach is clear.
>
> Challenge 4: To Consider the Integration of Immune and Other Systems. The immune system does not work in isolation. Therefore, attention should not only be paid to the potential of the immune system as inspiration, but also other systems with which the immune system interacts, in particular the neural and endocrine systems. This will pave the way for a greater understanding of the role and function of the immune system and develop a new breed of immune inspired algorithms." [92]

To be sure, some of these challenges are now being addressed within the community. The area of AIS is a dynamic and vibrant area of research, it is inherently interdisciplinary in nature and great lessons can be learnt between various communities and they can all benefit from successful interactions.

Bibliography

Primary Literature

1. Aickelin U, Bentley P, Cayzer S, Kim J, McLeod J (2003) Danger theory: The link between AIS and IDS? In: Timmis J, Bentley P, Hart E (eds) (2003) Proc of the 2nd International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 2787. Springer, Berlin, pp 147–155
2. Andrews PS, Timmis J (2005) Inspiration for the next generation of artificial immune systems. In: Jacob C, Pilat M, Bentley P, Timmis J (eds) (2005) Proc of the 4th International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 3627. Springer, Berlin, pp 126–138
3. Andrews PS, Timmis J (2005) On diversity and artificial immune systems: Incorporating a diversity operator into aiNet.

In: Proceedings of the International Conference on Natural and Artificial Immune Systems (NAIS05). LNCS, vol 391. Springer, Berlin, pp 293–306
4. Andrews PS, Timmis J (2006) A computational model of degeneracy in a lymph node. In: Bersini H, Carneiro J (eds) (2006) Proc of 5th International Conference on Artificial Immune Systems. LNCS. Springer, Berlin, pp 164–177
5. Ayara M (2005) An immune inspired solution for adaptable error detection in embedded systems. Ph D thesis, University of Kent
6. Ayara M, Timmis J, de Lemos R, de Castro L, Duncan R (2002) Negative selection: How to generate detectors. In: Proc of the First International Conference on Artificial Immune Systems (ICARIS-2002). University of Kent, Canterbury, pp 89–98
7. Balthrop J, Forrest S, Glickman M (2002) Revisiting lisys: Parameters and normal behavior. In: Proceedings of Congress On Evolutionary Computation (CEC). IEEE Press, pp 1045–1050
8. Bentley PJ, Greensmith J, Ujjin S (2005) Two ways to grow tissue for artificial immune systems. In: Jacob C, Pilat M, Bentley P, Timmis J (eds) (2005) Proc of the 4th International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 3627. Springer, Berlin, pp 139–152
9. Berek C, Ziegner M (1993) The maturation of the immune response. Immunol Today 14:200–402
10. Bersini H (1991) Immune network and adaptive control. In: Proceedings of the 1st European Conference on Artificial Life (ECAL). MIT Press, Cambridge, pp 217–226
11. Bersini H (1992) Reinforcement and recruitment learning for adaptive process control. In: Proc Int Fuzzy Association Conference (IFAC/IFIP/IMACS) on Artificial Intelligence in Real Time Control, pp 331–337
12. Bersini H, Carneiro J (eds) (2006) Proc of 5th International Conference on Artificial Immune Systems. LNCS, vol 4163. Springer, Berlin
13. Bersini H, Varela F (1994) The immune learning mechansims: Recruitment, reinforcement and their applications. Chapman Hall
14. Bezerra G, Barra T, de Castro LN, Von Zuben F (2005) Adaptive radius immune algorithm for data clustering. In: Jacob C, Pilat M, Bentley P, Timmis J (eds) (2005) Proc of the 4th International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 3627. Springer, Berlin, pp 290–303
15. Burnet FM (1959) The clonal selection theory of acquired immunity. Cambridge University Press, Cambridge
16. Canham RO, Tyrrell AM (2002) A multilayered immune system for hardware fault tolerance within an embryonic array. In: Timmis J, Bentley P (eds) (2002) Proc of the 1st International Conference on Artificial Immune Systems (ICARIS). University of Kent, Canterbury, pp 3–11
17. Clark E, Hone A, Timmis J (2005) A Markov chain model of the B-cell algorithm. In: Jacob C, Pilat M, Bentley P, Timmis J (eds) (2005) Proc of the 4th International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 3627. Springer, Berlin, pp 318–330
18. Cohen IR (2000) Tending Adam's garden: Evolving the cognitive immune self. Elsevier Academic Press
19. Cooke D, Hunt J (1995) Recognising promoter sequences using an artificial immune system. In: Proceedings of Intelligent Systems in Molecular Biology. AAAI Press, pp 89–97
20. Cutello V, Nicosia G, Parvone M (2004) Exploring the capability of immune algorithms: A characterisation of hypermutation operators. In: Nicosia G, Cutello V, Bentley P, Timmis J (eds) (2004) Proc of the 3rd International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 3239. Springer, Berlin, pp 263–276
21. Cutello V, Nicosia G, Pavone M, Timmis J (2007) An immune algorithm for protein structure prediction on lattice models. IEEE Trans Evol Comp 11(1):101–117
22. Cutello V, Nicosia G, Oliveto P, Romeo M (2007) On the convergence of immune algorithms. In: Proc of Foundations of Computational Intelligence. IEEE Press, pp 409–416
23. Dasgupta D (1999) Artificial Immune Systems and their Applications. Springer, Berlin
24. Dasgupta D, Forrest S (1995) Tool breakage detection in milling operations using a negative selection algorithm. Tech Rep Report, No CS95-5. Department of Computer Science, University of New Mexico
25. Dasgupta D, Majumdar NS (2002) Anomaly detection in multidimensional data using negative selection algorithm. In: Proc of Congress on Evolutionary Computation (CEC), Honolulu, Hawaii. IEEE Press, pp 1039–1044
26. Dasgupta D, Nino F (2000) A comparison of negative and positive selection algorithms in novel pattern detection. In: Proc of the IEEE International Conference on Systems, Man and Cybernetics (SMC), Nashville, 8–11 October
27. de Castro LN, Timmis J (2002) An artificial immune network for multi modal optimisation. In: Proceedings of the World Congress on Computational Intelligence WCCI, Honolulu, HI. IEEE Press, pp 699–704
28. de Castro LN, Timmis J (2002) Artificial immune systems: A new computational intelligence approach. Springer, Berlin
29. de Castro LN, Timmis J (2002) Hierarchy and convergence of immune networks: Basic ideas and preliminary results. In: Timmis J, Bentley P (eds) (2002) Proc of the 1st International Conference on Artificial Immune Systems (ICARIS). University of Kent, Canterbury, pp 231–240
30. de Castro LN, Von Zuben FJ (2000) The clonal selection algorithm with engineering applications. In: GECCO Workshop on Artificial Immune Systems and Their Applications, pp 36–37
31. de Castro LN, Von Zuben FJ (2001) aiNet: An artificial immune network for data analysis. Idea Group Publishing, pp 231–259
32. de Castro LN, Von Zuben FJ (2002) Learning and optimization using the clonal selection principle. IEEE Trans Evol Comp 6(3):239–251
33. Ebner M, Breunig H-G, Albert J (2002) On the use of negative selection in an artificial immune system. In: Proc of Genetic and Evolutionary Computation Conference (GECCO). Morgan Kaufman Publishers, San Francisco/New York, pp 957–964
34. Esponda F, Forrest S, Helman P (2004) A formal framework for positive and negative detection schemes. IEEE Trans Syst Man Cybern B 34(1):357–373
35. Farmer JD, Packard NH, Perelson AS (1986) The immune system, adaptation, and machine learning. Physica D 22:187–204
36. Forrest S, Perelson AS, Allen L, Cherukuri R (1994) Self-nonself discrimination in a computer. In: Proc IEEE Symposium on Research Security and Privacy. IEEE Press, pp 202–212
37. Forrest S, Hofmeyr S, Somayaji A (1997) Computer immunology. Comm ACM 40(10):88–96
38. Freitas A, Timmis J (2003) Revisiting the foundations of artificial immune systems: A problem oriented perspective. In:

Timmis J, Bentley P, Hart E (eds) (2003) Proc of the 2nd International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 2787. Springer, Berlin, pp 229–241
39. Garrett SM (2005) How do we evaluate artificial immune systems? Evol Comput 13(2):145–177
40. Gaspar A, Hirsbrunner B (2002) From optimization to learning in learning in changing environments: The pittsburgh immune classifier system. In: Timmis J, Bentley P (eds) (2002) Proc of the 1st International Conference on Artificial Immune Systems (ICARIS). University of Kent, Canterbury, pp 190–199
41. Germain RN (2004) An innately interesting decade of research in immunology. Nature Medicine 10:1307–1320
42. González F, Dagupta D (2003) Anomaly detection using real-valued negative selection. Genet Program Evolvable Mach 4(4):383–403
43. González F, Dasgupta D, Kozma R (2002) Combining negative selection and classification techniques for anomaly detection. In: Congress on Evolutionary Computation. IEEE, pp 705–710
44. González F, Dasgupta D, Gómez J (2003) The effect of binary matching rules in negative selection. In: Genetic and Evolutionary Computation – GECCO-2003. Lecture Notes in Computer Science, vol 2723. Springer, Chicago, pp 195–206
45. Goodman D, Boggess L, Watkins A (2002) Artificial immune system classification of multiple-class problems. In: Proc of Intelligent Engineering Systems. ASME, pp 179–184
46. Goodman D, Boggess L, Watkins A (2003) An investigation into the source of power for AIRS, an artificial immune classification system. In: Proc Int Joint Conf Neural Networks. IEEE, pp 1678–1683
47. Greensmith J, Aickelin U, Cayzer S (2005) Introducing dendritic cells as a novel immune-inspired algorithm for anomaly detection. In: Jacob C, Pilat M, Bentley P, Timmis J (eds) (2005) Proc of the 4th International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 3627. Springer, Berlin, pp 153–167
48. Greensmith J, Aickelin U, Twycross J (2006) Articulation and clarification of the dendritic cell algorithm. In: Bersini H, Coutinho A (eds) Proceedings of the 5th International Conference on Artificial Immune Systems. LNCS, vol 4163. Springer, Berlin
49. Hart E (2005) Not all balls are round: An investigation of alternative recognition-region shapes. In: Jacob C, Pilat M, Bentley P, Timmis J (eds) (2005) Proc of the 4th International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 3627. Springer, Berlin, pp 29–42
50. Hart E, Ross P (2002) Exploiting the analogy between immunology and sparse distributed memories: A system for clustering non-stationary data. In: Timmis J, Bentley P (eds) (2002) Proc of the 1st International Conference on Artificial Immune Systems (ICARIS). University of Kent, Canterbury, pp 49–58
51. Hart E, Ross P (2004) Studies on the implications of shape-space models for idiotypic networks. In: Nicosia G, Cutello V, Bentley P, Timmis J (eds) (2004) Proc of the 3rd International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 3239. Springer, Berlin, pp 413–426
52. Hart E, Timmis J (2005) Application areas of AIS: The past, the present and the future. In: Jacob C, Pilat M, Bentley P, Timmis J (eds) (2005) Proc of the 4th International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 3627. Springer, Berlin, pp 483–497
53. Hightower RR, Forrest SA, Perelson AS (1995) The evolution of emergent organization in immune system gene libraries. In: Proceedings of the 6th International Conference on Genetic Algorithms. Morgan Kaufmann, pp 344–350
54. Hofmeyr S, Forrest S (1999) Immunity by design: An artificial immune system. In: Proc of Genetic and Evolutionary Computation Conference (GECCO), pp 1289–1296
55. Hofmeyr S, Forrest S (2000) Architecture for an artificial immune system. Evol Comput 7(1):1289–1296
56. Hunt J, Cooke D (1996) Learning using an artificial immune system. J Netw Comput Appl 19:189–212
57. Hunt J, Timmis J, Cooke D, Neal M, King C (1998) JISYS: Development of an artificial immune system for real-world applications. In: Dasgupta D (ed) Artificial Immune Systems and their Applications. Springer, Berlin, pp 157–186
58. Jacob C, Pilat M, Bentley P, Timmis J (eds) (2005) Proc of the 4th International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 3627. Springer, Berlin
59. Jerne NK (1974) Towards a network theory of the immune system. Ann Immunol (Inst Pasteur) 125C:373–389
60. Ji Z, Dasgupta D (2004) Augmented negative selection algorithm with variable-coverage detectors. In: Congress on Evolutionary Computation. IEEE, pp 1081–1088
61. Ji Z, Dasgupta D (2004) Real-valued negative selection algorithm with variable-sized detectors. In: Genetic and Evolutionary Computation – GECCO-2004, Part I. Lecture Notes in Computer Science, vol 3102. Springer, Seattle, pp 287–298
62. Ji Z, Dasgupta D (2005) Estimating the detector coverage in a negative selection algorithm. In: Proceedings of Genetic and Evolutionary Computation Conference (GECCO). ACM Press, pp 281–288
63. Ji Z, Dasgupta D (2006) Applicability issues of the real-valued negative selection algorithms. In: Proceedings of Genetic and Evolutionary Computation Conference (GECCO). ACM Press, pp 111–118
64. Ji Z, Dasgupta D, Yang Z, Teng H (2006) Analysis of dental images using artificial immune systems. In: Proceedings of Congress On Evolutionary Computation (CEC). IEEE Press, pp 528–535
65. Kelsey J, Timmis J (2003) Immune inspired somatic contiguous hypermutation for function optimisation. In: Proc of Genetic and Evolutionary Computation Conference (GECCO). LNCS, vol 2723. Springer, Berlin, pp 207–218
66. Kelsey J, Timmis J, Hone A (2003) Chasing chaos. In: Proc of Congress on Evolutionary Computation (CEC). IEEE, Canberra, pp 89–98. http://www.cs.ukc.ac.uk/pubs/2002/1504
67. Kim J (2002) Integrating artificial immune algorithms for intrusion detection. Ph D thesis, UCL
68. Kim J, Bentley PJ (2001) An evaluation of negative selection in an artificial immune system for network intrusion detection. In: Proc of Genetic and Evolutionary Computation Conference (GECCO), San Francisco, USA. Morgan Kaufmann, pp 1330–1337
69. Kim J, Bentley PJ (2001) Towards an artificial immune system for network intrusion detection: An investigation of clonal selection with negative selection operator. In: Proc of Congress on Evolutionary Computation (CEC), Seoul, Korea. Morgan Kaufmann, pp 1244–1252
70. Kim J, Bentley PJ (2002) Immune memory in the dynamic clonal selection algorithm. In: Timmis J, Bentley P (eds) (2002)

Proc of the 1st International Conference on Artificial Immune Systems (ICARIS). University of Kent, Canterbury, pp 59–67
71. Knight T, Timmis J (2003) A multi-layered immune inspired machine learning algorithm. In: Lotfi A, Garibaldi M (eds) Applications and Science in Soft Computing. Springer, Berlin, pp 195–202. http://www.cs.kent.ac.uk/pubs/2003/1760
72. Krohling R, Zhou Y, Tyrrell A (2002) Evolving FPGA-based robot controllers using an evolutionary algorithm. In: Timmis J, Bentley P (eds) (2002) Proc of the 1st International Conference on Artificial Immune Systems (ICARIS). University of Kent, Canterbury, pp 41–46
73. Matzinger P (1997) An innate sense of danger. Semin Immunol 10(5):399–415
74. Matzinger P (2002) The danger model: A renewed sense of self. Science 296:301–305
75. Mendao M, Timmis J, Andrews PS, Davies M (2007) The immune system in pieces: Computational lessons from degeneracy in the immune system. In: Fogel DB (ed) Proc of Foundations of Computational Intelligence. IEEE Press, pp 394–400
76. Neal M (2002) An artificial immune system for continuous analysis of time-varying data. In: Timmis J, Bentley P (eds) (2002) Proc of the 1st International Conference on Artificial Immune Systems (ICARIS). University of Kent, Canterbury, pp 76–85
77. Nicosia G (2004) Immune algorithms for optimization and protein structure prediction. Ph D thesis, University of Catania
78. Nicosia G, Cutello V, Bentley P, Timmis J (eds) (2004) Proc of the 3rd International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 3239. Springer, Berlin
79. Perelson AS (1989) Immune network theory. Immunol Rev 110:5–36
80. Secker A, Freitas A, Timmis J (2003) AISEC: An artificial immune system for email classification. In: Proc of Congress on Evolutionary Computation (CEC). IEEE Press, pp 131–139
81. Singh S (2002) Anomaly detection using negative selection based on the r-contiguous matching rule. In: Timmis J, Bentley PJ (eds) Proceedings of the 1st International Conference on Artificial Immune Systems ICARIS. University of Kent at Canterbury, University of Kent at Canterbury Printing Unit, pp 99–106. http://www.aber.ac.uk/icaris-2002
82. Stepney S, Smith R, Timmis J, Tyrrell A, Neal M, Hone A (2006) Conceptual frameworks for artificial immune systems. Int J Unconv Comput 1(3):315–338
83. Stibor T, Timmis J (2007) An investigation into the compression quality of ainet. In: Fogel D (ed) Proc of Foundations of Computational Intelligence. IEEE Press
84. Stibor T, Bayarou KM, Eckert C (2004) An investigation of R-chunk detector generation on higher alphabets. In: Proc of Genetic and Evolutionary Computation Conference (GECCO). LNCS, vol 3102. Springer, Berlin, pp 299–307
85. Stibor T, Timmis J, Eckert C (2005) A comparative study of real-valued negative selection to statistical anomaly detection techniques. In: Jacob C, Pilat M, Bentley P, Timmis J (eds) (2005) Proc of the 4th International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 3627. Springer, Berlin, pp 262–275
86. Stibor T, Mohr P, Timmis J, Eckert C (2005) Is negative selection appropriate for anomaly detection? In: Proc of Genetic and Evolutionary Computation Conference (GECCO). ACM Press
87. Stibor T, Timmis J, Eckert C (2006) Generalization regions in hamming negative selection. In: Intelligent Information Processing and Web Mining. Advances in Soft Computing. Springer, Berlin, pp 447–456
88. Tarakanov AO, Skormin VA, Sokolova SP (2003) Immunocomputing: Principles and Applications. Springer, New York
89. Tarakanov AO, Goncharova LB, Tarakanov OA (2005) A cytokine formal immune network. In: Advances in Artificial Life, 8th European Conference, ECAL 2005, Canterbury, UK, 5–9 September 2005, pp 510–519
90. Tarakanov AO, Kvachev SV, Sukhorukov AV (2005) A formal immune network and its implementation for on-line intrusion detection. In: MMM-ACNS, pp 394–405
91. Timmis J (2000) Artificial immune systems: A novel data analysis technique inspired by the immune system. Ph D thesis, University of Wales
92. Timmis J (2007) Artificial immune systems: Today and tommorow. Natural Comput 6(1):1–18
93. Timmis J, Bentley P (eds) (2002) Proc of the 1st International Conference on Artificial Immune Systems (ICARIS). University of Kent, Canterbury
94. Timmis J, Edmonds C (2004) A comment on opt-AINet: An immune network algorithm for optimisation. In: Proc of Genetic and Evolutionary Computation Conference (GECCO). LNCS, vol 3102. Springer, Berlin, pp 308–317
95. Timmis J, Knight T (2001) Artificial immune systems: Using the immune system as inspiration for data mining. In: Abbas H, Ruhul A, Sarker A, Newton S (eds) Data Mining: A Heuristic Approach. Idea Group, pp 209–230
96. Timmis J, Neal M (2001) A resource limited artificial immune system for data analysis. Knowl Based Syst 14(3–4):121–130
97. Timmis J, Neal M, Hunt J (2000) An artificial immune system for data analysis. Biosystems 55(1/3):143–150
98. Timmis J, Bentley P, Hart E (eds) (2003) Proc of the 2nd International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 2787. Springer, Berlin
99. Timmis J, Edmonds C, Kelsey J (2004) Assessing the performance of two immune inspired algorithms and a hybrid genetic algorithm for function optimisation. In: Proc of Congress on Evolutionary Computation (CEC), vol 1. IEEE, pp 1044–1051
100. Twycross J, Aickelin U (2006) libtissue: Implementing innate immunity. In: Proc Congress on Evoutionary Computation. IEEE Press, pp499–506
101. Varela F, Coutinho A, Dupire B, Vaz N (1988) Cognitive networks: Immune, neural and otherwise. J Theor Imm 2:359–375
102. Watkins A (2001) AIRS: A resource limited artificial immune classifier. Master's thesis, Mississippi State University
103. Watkins A (2005) Exploiting immunological metaphors in the development of serial, parallel and distributed learning algorithms. Ph D thesis, University of Kent
104. Watkins A, Timmis J (2004) Exploiting parallelism inherent in AIRS, an artificial immune classifier. In: Nicosia G, Cutello V, Bentley P, Timmis J (eds) (2004) Proc of the 3rd International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 3239. Springer, Berlin, pp 427–438
105. Watkins A, Xintong B, Phadke A (2003) Parallelizing an immune-inspired algorithm for efficient pattern recognition. In: Intelligent Engineering Systems through Artificial Neural Networks: Smart Engineering System Design: Neural Networks,

Fuzzy Logic, Evolutionary Programming, Complex Systems and Artificial Life. ASME Press, pp 224–230
106. Watkins A, Timmis J, Boggess L (2004) Artificial immune recognition system (AIRS): An immune inspired supervised machine learning algorithm. Genet Program Evolvable Mach 5(3):291–318. http://www.cs.kent.ac.uk/pubs/2004/1634
107. Whitesides GM, Boncheva M (2002) Beyond molecules: Self-assembly of mesoscopic and macroscopic components. Proc Natl Acad Sci USA 99(8):4769–4774
108. Wierzchon S, Kuzelewska U (2002) Stable clusters formation in an artificial immune system. In: Timmis J, Bentley P (eds) (2002) Proc of the 1st International Conference on Artificial Immune Systems (ICARIS). University of Kent, Canterbury, pp 68–75

Books and Reviews

Dasgupta D (1999) Artificial Immune Systems and Their Applications. Springer, Berlin
Cohen I, Segal L (2001) Design Principles for the Immune System and Other Distributed Autonomous Systems. SFT, Oxford University Press
de Castro LN, Timmis J (2002) Artificial Immune Systems: A New Computational Intelligence Approach. Springer, Berlin
Tarakanov AO, Skormin VA, Sokolova SP (2003) Immunocomputing: Principles and Applications. Springer, New York
Ishida Y (2004) Immunity-Based Systems: A Design Perspective. Springer, New York

Implementation Theory

Luis C. Corchón
Departamento de Economía, Universidad Carlos III, Madrid, Spain

Article Outline

Glossary
Definition of the Subject
Introduction
Brief History of Implementation Theory
The Main Concepts
The Main Insights
Unsolved Issues and Further Research
Answers to the Questions
Acknowledgments
Bibliography

Glossary

Type of an agent All the information possessed by this agent. It may refer to the preferences of this agent and/or to the knowledge of this agent of the preferences of other agents.

State of the world Description of all information possessed by all agents.

Social choice rule A correspondence mapping the set of states of the world in the set of allocations. It represents the social objectives that the society or its representatives want to achieve.

Mechanism A list of message spaces and an outcome function mapping messages into allocations. It represents the communication and decision aspects of the organization.

Equilibrium concept A mapping (or a collection of them) from the set of states of the world into allocations yielded by equilibrium messages. This equilibrium is a game-theoretical notion of how agents behave, e. g. Nash Equilibrium, Bayesian Equilibrium, Dominant Strategies, etc.

Implementable social choice rule in an equilibrium concept (e. g. nash equilibrium) A Social Choice Rule is implementable in an equilibrium concept (e. g. Nash Equilibrium) if there is a mechanism such that for each state of the world the allocations prescribed by the Social Choice Rule and those yielded by the equilibrium concept coincide.

Definition of the Subject

Implementation theory studies which social objectives (i. e. Social Choice Rules) are compatible with the incentives of the agents (i. e. are implementable). In other words it is the systematic study of the social goals that can be achieved when agents behave strategically.

Introduction

Dear colleague;

I wrote this survey with you in mind. You are an economist doing research who would like to know why implementation is important. And by this I do not mean why some people won the Nobel Prize working in this area. I mean, what are the deep insights found by implementation theory and what applications are delivered by these tools. I propose a simple game: try to answer the following questions. If you cannot answer them, but you think they are important, read the survey. At the end of this survey, I will give you the answers. I will also tell you why I like implementation theory so much!

1. Why are agents price-takers? Is price-taking possible in economies with a finite number of agents?
2. Suppose two firms wish to merge. They claim that the merger will bring large cost reductions but some people fear that the firms just want to avoid competition. What would be your advice?

3. How should a monopoly be regulated when regulators do not know the cost function or the demand function of the monopolist?
4. How should it be determined whether or not a public facility – a road, a bridge, a stadium – should be constructed and who should pay for it?
5. Is justice possible in this world? Can we reconcile justice and self-interest?
6. Can an uninformed planner achieve better allocations than those produced by completely-informed agents in an unregulated market?
7. In competitive ice skating, the highest and lowest marks awarded by judges are discarded and the remaining are averaged. Do you think that this procedure eliminates incentives to manipulate votes?
8. What kind of policies would you advocate to fight Global Warming?

The answers to these questions are found in Sect. "Answers to the Questions" The rest of this paper goes as follows. Sect. "Brief History of Implementation Theory" is a historical introduction that can be skipped. Sect. "The Main Concepts" explains the basic model. Sect. "The Main Insights" explains the main results. Sect. "Unsolved Issues and Further Research" offers some thoughts about the future direction of the topic.

Brief History of Implementation Theory

From, at least Adam Smith on, we have assumed that agents are motivated by self-interest. We also assumed that agents interact in a market economy where prices match supply and demand. This tradition crystallized in the Arrow–Debreu-McKenzie model of General equilibrium in the 1950s. But it was quickly discovered that this model had important pitfalls other than focusing on a narrow class of economic systems: On the one hand, an extra agent was needed to set prices, the auctioneer. On the other hand agents follow rules, i. e. to take prices as given, which are not necessarily consistent with self-interest. An identical question had arisen earlier when Taylor [129] and Lange [70], following Barone [17], proposed a market socialism, where socialist managers maximize profits: Why would socialist managers choose output in the way prescribed to them (or who will provide and preserve capital in a system where the private property of such items is forbidden?)? Samuelson [113] voiced identical concern about the Lindahl solution to allocate public goods: "It is in the selfish interest of each person to give false signals". This concern gave rise later on to the golden rule of incentives – as stated by Roger Myerson [91]: "An organization must give its members the correct incentives to share information and act appropriately". Earlier, it had aroused the interest of Leonid Hurwicz, the father of Implementation theory, in economic systems other than the market. In any case it was clear that an important ingredient was missing in the theory of economic systems. This element was that not all information needed for resource allocation was transmitted by prices: Some vital items have to be transmitted by agents.

Several proposals arose to fill the gap: On the one hand, models of markets under asymmetric information, Vickrey [135], Akerlof [3], Spence [128] and Rothchild and Stiglitz [106]. On the other hand models of public intervention, like optimal taxation, Mirless [82], and mechanisms for allocating public goods, Clarke [27] and Groves [48], with the so-called Principal-Agent models somewhere in the middle. The key word was "Truthful Revelation" or "Incentive Compatibility": Truthful revelation of information must be an equilibrium strategy, either a dominant strategy, as in Clarke and Groves, or a Bayesian equilibrium as in Arrow [7] and D'Aspremont and Gerard-Varet [10]. A motivation for this procedure was provided by the "Revelation Principle", Gibbard [45], Myerson [89], Dasgupta, Hammond and Maskin [38] and Harris and Townsend [50]: If a mechanism yields certain allocations in equilibrium, telling the truth about one's characteristics must be an equilibrium as well (however, telling the truth may not be an equilibrium in the original mechanism you might have to use an equivalent direct mechanism). This result is of utmost importance and it will be thoroughly considered in Sect. "The Main Concepts" However, it was somehow misread as "there is no loss of generality in focusing on incentive compatibility". But what the revelation principle asserts is that truthful revelation is *one* of the, possibly, many equilibria. It does not say that truthful revelation is *the only* equilibrium. As we will see in some cases it is a particularly unsatisfactory way of selecting equilibria.

The paper by Hurwicz [53], popularized by Reiter [102], presented a formal structure for the study of economic mechanisms which has been followed by all subsequent papers. Maskin [76], whose first version circulated in 1977, is credited as the first paper where the problem of multiple equilibria was addressed as a part of the model and not as an afterthought, see the report of the Nobel Prize Committee [101]. Maskin studied implementation in Nash equilibrium (see Glossary). Later his results were generalized to Bayesian Equilibrium by Postlewaite and Schmeidler [100] and Palfrey and Srivastava [94,95].

Finally, Moulin [87] studied Dominance Solvability and Moore and Repullo [87] Subgame Perfect Equilibrium. The century closed with several characterizations on

what can be implemented in other equilibrium concepts: Moore and Repullo [85] in Nash Equilibrium, Palfrey and Srivastava [96] in Undominated Nash Equilibrium, Jackson [58] in Bayesian Equilibrium, Dutta and Sen [39] in Strong Equilibrium and Sjöström [125] in Trembling Hand Equilibria. With all these papers in mind, the basic aspects of implementation theory are now well understood.

The interested reader may complement the previous account with the surveys by Maskin and Sjöström [78] and Serrano [122] which cover the basic results and by Baliga and Sjöström [13] for new developments including experiments. See also Maskin [75], Moore [84], Corchón [29], Jackson [60] and Palfrey [93]. Several important applications of Implementation Theory are not surveyed here: Auctions, see Krishna [67], Contract theory, see Laffont and Martimort [69], Matching, see Roth (forthcoming) and Moral Hazard see Ma, Moore and Turnbull [73].

The Main Concepts

We divide this section into four subsections: The first describes the environment, the second deals with social objectives, the third revolves around the notion of a mechanism and the last defines the equilibrium concepts that we will use here.

The Environment

Let $I = \{1, \ldots, n\}$ be the set of agents. Let θ_i be the type of i. This includes all the information in the hands of i. Let Θ_i be agent i's type set. The set $\Theta \subset \prod_{i=1}^{n} \Theta_i$ is the set of states of the world. For each $\theta \in \Theta$ we have a feasible set $A(\theta)$ and a preference profile $R(\theta) = (R_1(\theta), \ldots, R_n(\theta))$. $R_i(\theta)$ is a complete, reflexive and transitive binary relation on $A(\theta)$. $I_i(\theta)$ denotes the corresponding indifference relation. Set $A \equiv \bigcup_{\theta \in \Theta} A(\theta)$. Let $a = (a_1, a_2, \ldots, a_n) \in A$ be an allocation, also written (a_i, a_{-i}), where $a_{-i} \equiv (a_1, a_2, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n)$.

The standard model of an exchange economy is a special case of this model: θ is an economy. $X_i(\theta) \subset \Re^k$ is the consumption set of i. $w_i(\theta) \in intX_i(\theta)$ are the endowments in the hands of i. The preferences of i are defined on $X_i(\theta)$. The set of allocations $A(\theta)$ is defined as

$$A(\theta) = \left\{ a \mid \sum_{i=1}^{n} (a_{ij} + w_{ij}(\theta)) \leq 0, \quad j = 1, 2, \ldots, k, \right.$$
$$\left. (a_{i1}, a_{i2}, \ldots, a_{ik}) \in X_i(\theta), \quad \forall i \in I \right\} .$$

A special case of an exchange economy is bilateral trading: Here there are two agents, the seller and the buyer. The seller has a unit of an indivisible good and both agents are endowed with an infinitely divisible good ("money"). Preferences are representable by linear utility functions. The type of each agent, also called her valuation, is the marginal rate of substitution between both goods. Finally, the set of types is a closed interval of the real line.

Another example is the social choice model where the set of states of the world is the Cartesian product of individual type sets, $\Theta = \prod_{i=1}^{n} \Theta_i$. The set of feasible allocations is constant. The preferences of each agent only depend on her type, for all $\theta \in \Theta$, $R_i(\theta) = R_i(\theta_i)$ all $i \in I$.

The model of public goods is a hybrid of the social choice and the exchange economy models. For a subset of goods, say $1, 2, \ldots, l$, agents receive the same bundle (these are the public goods). For goods $l + 1, \ldots, k$, agents can consume possibly different bundles.

Social Objectives

Implementation begins by asking what allocations *we* want to achieve. In this sense, implementation theory reverses the usual procedure, namely, fix a mechanism and see what the outcomes are. The theory is rather agnostic as to who is behind *we*: It could be a democratic society, it could be a dictator, a benevolent planner, etc. Formally, a correspondence $F: \Theta \twoheadrightarrow A$ such that $F(\theta) \subseteq A(\theta)$ for all $\theta \in \Theta$ will be called a *Social Choice Rule* (*SCR*). Under risk or uncertainty, allocations are state-dependent (recall the concept of contingent commodities in General Equilibrium). Thus an allocation is a single-valued function $f: \Theta \to A$. The notion of a SCR is replaced by that of a *Social Choice Set* (SCS) defined as a collection of functions mapping Θ into A. Examples of SCR are the Pareto rule, which maps every state into the set of Pareto efficient allocations for this state, the Walrasian SCR which maps every economy in the set of allocations that are a Walrasian Equilibrium for this economy, etc.

If states of the world were contractible, i. e. if they could be written in an enforceable contract specifying the allocations in each state, SCR or SCS would be directly achieved, assuming that those not complying could be punished harshly. Unfortunately, states of the world are a description of preferences and productive capabilities, being those difficult to describe and therefore easy to manipulate. Thus, we have to find another method to reach the desired allocations.

Mechanisms

If the information necessary to judge the desirability of allocations is in the hands of agents, it seems that the only way of retrieving this information is by asking them. But,

of course, agents cannot be trusted to reveal truthfully their information because they might lose by doing so. Thus the owner of a defective car will think twice about revealing the true state of the car if the price of defective cars is less than the price of reliable cars. But perhaps we may design ways in which the messages sent by different agents are checked one against the other. We may also design ways in which agents send information by indirect means, say by raising flags, making gestures, and so on and so forth. This is the idea behind the concept of a mechanism (also called a game form).

Formally, a *mechanism* is a pair (M, g) where $M \equiv \prod_1^n M_i$ is the *message space* and $g\colon M \to A$ is the outcome function. M_i denotes agent i's *message space with typical element* m_i. In some cases, i. e. when goods are indivisible, the outcome function maps M into the set of lotteries on A, denoted by $\mathcal{L}A$. In this case the outcome function yields the probability of obtaining an object. Let $m = (m_1, .., m_n) \in M$, be a list of messages, also written (m_i, m_{-i}) where m_{-i} is a list of all messages except those sent by i.

Another interpretation of a mechanism, more in tune with decentralized systems, is that messages describe contracts among agents and the outcome function is a legal system that converts contracts into allocations.

If feasible sets are state dependent we have a problem: Suppose that at θ we want to achieve allocation $a \in A(\theta)$. So there must be a message, say m such that $g(m) = a$. But what if there is another state, say θ' for which $a \notin A(\theta')$? In this case $g(m) \notin A(\theta')$. In other words, since mechanisms are not state dependent they may yield unfeasible allocations. We will postpone the discussion of this problem until Sect. "Unsolved Issues and Further Research" For the time being, let us assume that feasible sets are not state dependent.

Equilibrium

Since the messages sent by agents are tied to their incentives, it is clear that we have to use an equilibrium concept borrowed from game theory. Thus, given $\theta \in \Theta$, a mechanism (M, g) induces a game in normal form (M, g, θ). There are many "solutions" to what would constitute an equilibrium. Let us begin by considering the notion of a Nash equilibrium:

Definition 1 A message profile $m^* \in M$ is a Nash equilibrium for (M, g, θ) if, for all $i \in I$ $g(m^*)R_i(\theta)g(m_i, m^*_{-i})$ for all $m_i \in M_i$.

Let $NE(M, g, \theta)$ be the set of allocations yielded by all Nash equilibria of (M, g, θ). We now ask, given a SCR, what mechanism, if any, would produce outcomes identical to the SCR. In this sense, the mechanism is the variable of our analysis i. e. the mechanism "solves" the equation $NE(M, g, \theta) = F(\theta)$, for all $\theta \in \Theta$. Formally,

Definition 2 The SCR F is implementable in Nash equilibrium if there is a mechanism (M, g) such that, for all $\theta \in \Theta$, $NE(M, g, \theta) \neq \emptyset$ and:

1. $F(\theta) \subseteq NE(M, g, \theta)$.
2. $NE(M, g, \theta) \subseteq F(\theta)$.

The previous concept can be easily generalized. Given a mechanism (M, g) an equilibrium concept is a mapping, say $E_{(M,g)}\colon \Theta \twoheadrightarrow A$ such that $E_{(M,g)}(\theta) \subseteq A(\theta)$ for all $\theta \in \Theta$. For instance $E_{(M,g)}(\theta)$ may be the set of allocations arising from dominant strategy profiles in θ when the mechanism (M, g) is in place. The notion of implementation in an equilibrium concept easily follows. See Thomson [132] for a discussion of other concepts of implementation.

The problem is that some equilibrium concepts can not be written in the way we just described because the actions to be taken in state, say θ', depend on preferences in states other than θ'. To see this, suppose that agents attach a vector of probabilities to each possible type of the other agents, Harsanyi [51]. Denote by $q(\theta_{-i}/\theta_i)$ the vector of probabilities attached by i that other agents have types θ_{-i} given that she is of type θ_i. For simplicity assume that it is a strictly positive vector. Suppose that preferences are representable by a von Neumann-Morgenstern utility index $V_i(a, \theta)$. In this framework, (as first noticed by Vickrey [135]) a strategy for i, denoted by s_i, is no longer a message but a function from the set of types of i in the set of messages of i, namely, $s_i\colon \Theta_i \to M_i$. A strategy profile, s, is a collection of strategies, one for each agent, $s = (s_1, \ldots, s_n)$ also written as (s_i, s_{-i}). For simplicity, the next definition assumes that type sets are finite.

Definition 3 A Bayesian Equilibrium (BE) for $(M, g, R(\cdot))$ is a s^* such that for all i, $\theta \in \Theta$, and $m_i \in M_i$,

$$\sum_{\theta_{-i} \in \Theta_{-i}} q(\theta_{-i} \mid \theta_i) V_i(g(s^*(\theta)), \theta) \geq \sum_{\theta_{-i} \in \Theta_{-i}} q(\theta_{-i} \mid \theta_i) V_i(g(m_i, s^*_{-i}(\theta_{-i})), \theta)$$

Thus, an equilibrium concept – given a mechanism – is a collection of functions, denoted by $H_{(M,g)}$, such that for all $h_{(M,g)} \in H_{(M,g)}$ $h_{(M,g)}\colon \Theta \to A$. Finally, the definition of implementable SCS in BE follows.

Definition 4 The mechanism (M, g) implements a SCS F in BE if:

1. For any BE s there exists $x \in F(\theta)$, such that $g(s(\theta)) = x(\theta)$ for all $\theta \in \Theta$.
2. For any $x \in F$, there is a BE s such that $g(s(\theta)) = x(\theta)$ for all $\theta \in \Theta$.

Looking at our definitions of an implementable SCR or SCS we see that the first requirement is that all equilibria yield "good" allocations. The second requirement is that given an allocation to be implemented, there is an equilibrium "sustaining" this allocation. These two requirements bear some resemblance to the two fundamental theorems of welfare economics, namely that competitive equilibrium is efficient and that any efficient allocation can be achieved as a competitive equilibrium with the appropriate endowment redistribution. Notice that endowment redistribution is not used in the definition of implementation.

The Main Insights

We group our results here under three headings: The Revelation Principle and its consequences, Monotonicity and how to avoid it and the limits of design. We will discuss them each in turn.

The Revelation Principle and its Consequences

The definition of a mechanism is extremely abstract. No conditions have been imposed on what might constitute a message space or an outcome function. And since implementation theory considers the mechanism the variable to be found, this is an unhappy situation: we are asked to find something whose characteristics we do not know! Fortunately the revelation principle comes to the rescue by stating a necessary condition for implementation: If a single valued SCR, which we will call a Social Choice Function (SCF) is implementable, there is a *revelation mechanism* for which telling the truth is an equilibrium. A revelation mechanism (associated with a SCF) is a mechanism in which the message space for each agent is her set of types and the outcome function is the SCF. We say that a SCF is truthfully implementable or incentive compatible if truth-telling is a Bayesian equilibrium (or a dominant strategy) of the direct mechanism associated with it. The following result formally states the revelation principle:

Theorem 1 *If f is a Bayesian (resp. dominant strategy) implementable SCF, f is incentive compatible.*

Proof Let f be Bayesian implementable. Therefore, there exists a mechanism (M, g) and a Bayesian equilibrium s^* such that $g(s^*(\theta)) = f(\theta)$ for every $\theta \in \Theta$. Since $s^*(\cdot)$ is a Bayesian equilibrium, $\forall \theta \in \Theta, \forall m_i \in M$

$$\sum_{\theta_{-i} \in \Theta_{-i}} q(\theta_{-i} \mid \theta_i) V_i(g(s^*(\theta)), \theta) \geq \sum_{\theta_{-i} \in \Theta_{-i}} q(\theta_{-i} \mid \theta_i) V_i(g(m_i, s^*_{-i}(\theta_{-i})), \theta) .$$

Which implies that $\forall \theta'_i \in \Theta_i, \forall \theta_{-i} \in \Theta_{-i}$,

$$\sum_{\theta_{-i} \in \Theta_{-i}} q(\theta_{-i} \mid \theta_i) V_i(f(\theta), \theta) \geq \sum_{\theta_{-i} \in \Theta_{-i}} q(\theta_{-i} \mid \theta_i) V_i(f(\theta'_i, \theta_{-i}), \theta) .$$

The proof for the case of dominant strategies is identical. □

Theorem 1 (T.1 in the sequel) can be explained in terms of a mediator, i. e. somebody to whom you say "who you are" and who chooses the strategy that maximizes your payoffs on your behalf. Would you try to fool such a person? If you do so, you are fooling yourself because the mediator would choose a strategy that is not the best for you. Thus, the best thing for you to do is to tell the truth (providing an unexpected backing to the aphorism "honesty is the best policy"!).

Consider now the following results, due to Hurwicz [54] (who proved it for the case of $n = 2$) and to Gibbard [45] and Satterthwaite [115] respectively.

Theorem 2 *In exchange economies environments there is no SCF such that:*

1. *It is truthfully implementable in dominant strategies.*
2. *It selects individually rational allocations.*
3. *It selects efficient allocations.*
4. *Its domain includes all economies with convex and continuous preferences.*

Theorem 2′ *In social choice environments there is no SCF such that:*

1. *It is truthfully implementable in dominant strategies.*
2. *It is non-dictatorial.*
3. *Its range is A, with $\#A > 2$.*
4. *Its domain includes all possible preference profiles.*

It is clear that there are trivial SCF in which any three conditions in T.2–2′ are compatible. But T.2–2′ are very robust in the sense that they hold for small domains of economies [15,16], for weaker notions of individual rationality [110,119] and in public goods domains [71]. Moreover, assuming quasi-linear utility functions, Hurwicz and

Walker [56], building on a previous paper by Walker, proved that the set of economies for which conditions 1–3 in T.2 are incompatible is open and dense. Beviá and Corchón [20] show that these conditions are incompatible for *any* economy where utility functions are quasi-linear, strictly concave, differentiable and fulfill a very mild regularity condition. These results show that Vickrey–Clarke–Groves mechanisms fail to achieve efficient allocations in general (Vickrey–Clarke–Groves mechanisms are revelation mechanisms that work in public good economies where utility functions are quasi-linear in "money". The outcome function selects the level of public good that maximizes the sum of utilities announced by agents and the money received by each individual is the sum of the utility functions announced by all other agents. For an exposition of these mechanisms, see Green and Laffont [46]).

A proof of T.2 can be found in Serizawa [118]. Simple proofs of T.2′ can be found in Barberá [14], Benoit [18]) and Sen [117].

T.1 and 2–2′ imply that there is no mechanism implementing an efficient and individually rational (resp. nondictatorial) SCF in dominant strategies when the domain of the SCF is large enough. In other words, the revelation principle implies that the restriction to mechanisms where agents announce their own characteristic is not important when considering negative results. Thus, the Revelation principle is an appropriate tool for producing negative results. But we will see that to rely entirely on this principle when trying to implement a SCF may yield disastrous results.

A natural question to ask is what happens with the above impossibility results when we weaken the requirement of implementation in dominant strategies to that of implementation in Bayesian equilibrium. The following result, due to Myerson and Satterthwaite [92], answers this question.

Theorem 2′′ *In the bilateral trading environment there is no SCF such that:*

1. *It is truthfully implementable in Bayesian Equilibrium.*
2. *It selects individually rational allocations once agents learn their types.*
3. *It selects ex-post efficient allocations.*
4. *Its domain includes all linear utility functions with independent types distributed with positive density and the sets of types have a nonempty intersection.*

Proof (Sketch, see Krishna and Perry [66] for details) By the revenue equivalence theorem (see Klemperer [64], Appendix A), all mechanisms fulfilling conditions 2) and 3) above raise identical revenue. So it is sufficient to consider the Vickrey–Clarke–Groves which, as we remarked before, is not efficient. □

Again the weakening of any condition in T.2′′ may produce positive results [138] (Table 1 presents an illuminating discussion of this issue). For instance, suppose seller valuations are 1 or 3, and buyer valuations are 0 or 2. The mechanism fixes the price at 1.5 and a sale occurs when the valuation of the buyer is larger than the valuation of the seller. This mechanism implements truthfully a SCF satisfying 2) and 3) above. Unfortunately, it does not work when valuations are drawn from a common interval with positive densities.

But unlike T.2–2′, there are robust examples of SCF truthfully implementable in Bayesian Equilibrium when conditions 2) or 4) are relaxed. Also, inefficiency converges to zero very quickly when the number of agents increases (see [47]). This is because the equilibrium concept is now weaker and we are approaching a land where incentive compatibility has no bite, as we will see in T.3 below.

First, d'Aspremont and Gerard-Varet [9,10], and Arrow [7] showed that conditions 1)-3)-4) are compatible with individual rationality *before* agents learn their types in the domain of public goods with quasi-linear utility functions. They proposed the "expected externality mechanism" in which each agent is charged the expected externality she creates on the remaining players. Later on, Myerson [90] and Makowski and Mezzetti [74] presented incentive compatible SCF yielding ex-post efficient and individually rational allocations in the domain of exchange economies with quasi-linear preferences and more than two buyers. In Myerson [90], agents have correlated valuations. Buyers are charged even if they do not obtain the object or they may receive money and no object or even receive the object plus some money. Makoswki and Mezzetti [74] assume no correlation and that the highest possible valuation for a buyer is larger than the seller's highest possible valuation. They consider a family of mechanisms, called Second Price Auction With Seller (SPAWS), in which the highest bidder obtains the object, the seller receives the first bid and the winning buyer pays the second price. These mechanisms not only induce truthful behavior and yield ex-post efficient and individually rational allocations: For any other mechanism with these properties we can find a SPAWS mechanism yielding the same allocation.

Suppose now that information is *Non-Exclusive* in the sense that the type of each player can be inferred from the knowledge of all the other players' type. Intuition suggests that in this case, incentive compatibility has no bite whatsoever (i. e. T2′′ does not apply) since the behavior of each

player can be "policed" by the remaining players. In order to prove this, we will concentrate on an extreme, but illuminating, case of non-exclusive information, namely Nash equilibrium. In this framework, since information is complete, a direct mechanism is one where each agent announces a state of the world.

Consider the following assumption:

- (W) $\exists z \in A$ such that $\forall \theta \in \Theta, \forall a \in A, aR_i(\theta)z, \forall i \in I$.

This assumption will be called "universally worst outcome" because it postulates the existence of an allocation which is unanimously deemed as the worst. In an exchange economy this allocation would be zero consumption for everybody. Now we have the following result (Repullo [103], Matsushima [79]):

Theorem 3 *If $n = 2$ and W holds, any SCF is truthfully implementable in Nash Equilibrium. If $n > 2$, any SCF is truthfully implementable in Nash Equilibrium.*

Proof When $n = 2$ consider the following outcome function: $g(\theta', \theta') = f(\theta')\ \ \forall \theta' \in \Theta, g(\theta', \theta'') = z$ for all $\theta' \neq \theta''$. Clearly, truth is an equilibrium. When $n > 2$, consider the following outcome function: If m is such that $n - 1$ agents announce state θ', then $g(m) = f(\theta')$. Otherwise, $g(\cdot)$ is arbitrary. Clearly truth is an equilibrium as well in this case. □

The first thing to notice is the difference between the cases of two and more than two individuals. We will have more to say about this in the next section. The second is that the construction in Theorem 3 produces a large number of equilibria, and that there seems to be no good reason for individuals to coordinate in the truthful equilibria.

For instance, suppose workers can be either fit or unfit. When a profit-maximizing firm asks its employees about their characteristics, and all workers are fit, a unanimous announcement such as "we are all unfit" is an equilibrium. If fit workers are required hard work and unfit workers are asked to light work, do you think it is reasonable that workers coordinate in the truthful equilibrium? A more elaborate example was produced by Postlewaite and Schmeidler [100]: There are three agents. The first agent has no information and agents 2 and 3 are perfectly informed. The ranking of agent 1 over alternatives is the opposite of agents 2 and 3 who share the same preferences. The SCF is the top alternative of agent 1 in each state. It is intuitively clear that besides the truthful equilibria, there is another untruthful equilibrium where both informed agents lie and they are strictly better off than under truthful behavior. Again, coordination in the truthful equilibrium seems very unlikely. Thus, we have to recognize that we have a problem here. The next section will tell you how we can solve it.

Summing up, what do we learn from the results in this section?

1. When looking for an implementable SCF, a useful first test is whether this SCF yields incentives for the agents to tell the truth, see T.1. But this test is incomplete because of the existence of equilibria other than the truthful one, see T.3. These untruthful equilibria sometimes sound more plausible than the truthful one.
2. All impossibility theorems – T.2–2′–2″ – have the same structure: Truthful implementation, individual rationality/non-dictatorship, efficiency/large range of the SCR and large domain. Usually in social choice environments conditions 2 and 3 are weaker than in economic environments but the condition on the domain is stronger.
3. The classic story of the market making possible efficient allocation of resources under private information has to be revised. Private information in many cases precludes the existence of *any* mechanism achieving efficient and individually rational allocations under informational decentralization, see T.2–2′–2″ .
4. The same remarks apply to naive applications of the Coase theorem where agents are supposed to achieve Pareto efficient allocations just because they have contractual freedom (ditto about Bargaining Theory). In the parlance of Coase, private information is an important transaction cost.
5. When mechanisms with adequate properties exist, like those proposed by Arrow, d'Aspremont and Gerard-Varet, Myerson and Makowski and Mezzetti, they are not of the kind that we see in the streets. Careful design is needed. These mechanisms are tailored to specific assumptions on valuations, thus their range of applicability may be limited.

Monotonicity and How to Avoid It

We have seen that equilibria other than the truthful one are likely to arise. We have also seen that these equilibria cannot be disregarded a priori. So we have to find a way of getting rid of equilibria that do not yield desirable allocations. Under Dominant Strategies, clearly, if all preference orderings are strict, implementation and truthful implementation becomes identical, see [38], Corollary 4.1.4 ([68] presents other conditions under which this result holds. See [103] for the case where implementation and truthful implementation in dominant strategies do not coincide). For the ease of exposition we consider next Nash equilibria.

It turns out that the key to this issue in the case of Nash Equilibrium is the following monotonicity property, sometimes called Maskin monotonicity because Maskin [76] established its central relevance to implementation.

- (M) A SCR F is Monotonic if

$$\{a \in F(\theta),\ \ aR_i(\theta)b \rightarrow aR_i(\theta')b\ \ \forall i \in I\} \rightarrow a \in F(\theta')$$

Monotonicity says that if an allocation is chosen in state θ and this allocation doesn't fall in anybody's ranking in state θ', this allocation must also be chosen in θ'. We will also speak of a "monotonic transformation of preferences at θ" when the requirement $aR_i(\theta)b \rightarrow aR_i(\theta')b\ \ \forall i \in I$ is satisfied. This requirement simply says that the set of preferred allocations shrinks when we go from θ to θ'.

Monotonicity looks like a not unreasonable property, even though as we will see in a moment, there are cases in which it is incompatible with other very desirable properties. In any case the importance of monotonicity comes from the fact that it is a necessary condition for implementation in Nash Equilibrium, as proved by Maskin [76].

Theorem 4 *If a SCR is implementable in Nash Equilibrium it is Monotonic.*

Proof If F is Nash implementable, there must be a mechanism (M, g) such that $\forall a \in F(\theta)$, there is a Nash equilibrium m^*, such that $g(m^*) = a$. Since $aR_i(\theta)b \rightarrow aR_i(\theta')b\ \forall i \in I$, m^* is also a Nash Equilibrium at θ'. Since F is implementable, $a \in F(\theta')$. □

Let us now discuss the concept of monotonicity. First, the bad news. Popular concepts in voting, like Plurality, Borda Scoring and Majority Rule are not monotonic, neither is the Pareto correspondence, see Palfrey and Srivastava [96], p. 484. Even the venerable Walrasian correspondence is not monotonic! The failure of the Pareto and the Walrasian SCR to be monotonic can be amended: If preferences are strictly increasing in all goods, the Pareto SCR is monotonic in economic environments. The *Constrained Walrasian* SCR – in which consumers maximize with respect to the budget constraint *and* the availability of resources – is also monotonic. More serious is a result due to Hurwicz [55] that uses two weak conditions on a SCR defined in the domain of exchange economies.

- (L) The domain of F contains all preferences representable by linear utility functions.
- (ND) If $a \in F(\theta)$ and $aI_i(\theta)b\ \ \forall i \in I$, then $b \in F(\theta)$.

The first condition is a rather modest requirement on the richness of the domain of F. The second is a non-discrimination property which says that if everybody considers two allocations to be indifferent and one allocation belongs to the SCR then it must be the other. Now we have the following:

Theorem 5 *Let F be a SCR satisfying L and ND and such that:*

1. *It is Nash implementable.*
2. *It selects individually rational allocations.*
3. *Then, if x is a Walrasian allocation at θ, $x \in F(\theta)$.*

Proof (Sketch, see [130] for details) Take an economy θ. Let x be a Walrasian allocation for θ. Consider a new economy, called θ^L, where the marginal rates of substitution among goods are constant and equal to a vector of Walrasian prices. By individual rationality, F must select an allocation which is indifferent to x. By ND, $x \in F(\theta^L)$. Since F is Nash implementable, it satisfies M. Now since $xR_i(\theta^L)b \rightarrow xR_i(\theta)b\ \ \forall i \in I$, by M, $x \in F(\theta)$. □

Thus under weak conditions, Walrasian allocation are always in the set of those selected by a Monotonic SCR. And these allocations may fail to satisfy properties of fairness or justice as pointed out by the critics of the market. Under stronger assumptions, the converse is also true, i. e. only Walrasian allocations can be selected by a Nash implementable SCR [55]. Also, T.5 has the following unpleasant implication.

Theorem 6 *There is no SCF in exchange economies such that:*

1. *It is Nash implementable.*
2. *It selects individually rational allocations.*
3. *ND holds.*
4. *It is defined on all exchange economies.*

Proof T.5 implies that any Walrasian allocation belongs to the allocations selected by F. Since Walrasian equilibrium is not unique for some economies in the domain, hence the result. □

T.6 has a counterpart in social choice domains, Muller and Satterthwaite [88].

Theorem 6′ *There is no SCF in a social choice domain such that:*

1. *It is monotonic.*
2. *It is not dictatorial.*
3. *Its range is A with $\#A > 2$.*
4. *It is defined on all possible preferences.*

An implication of T.6–6′ is that single valued SCR are still problematic. But the consideration of multivalued SCR,

brings a new problem: The existence of several Nash equilibria. For instance, if $a, b \in F(\theta)$ with a and b being efficient allocations, agents play a kind of "Battle of the Sexes" game with no clear results. Moreover the Nash Equilibrium in mixed strategies may yield allocations outside $F(\theta)$ (the concern about mixed-strategy equilibria was first raised by Jackson [59]).

Now let us come to the good news. Firstly, the ND condition, which is essential for T.5 to hold, is not as harmless as it appears to be. For instance, it is not satisfied by the Envy-Free SCR, see Thomson [131] for a discussion. Secondly, there are perfectly reasonable SCR which are monotonic: we have already encountered the Constrained Walrasian SCR. Also any SCR selecting interior allocations in $\mathcal{L}A$ when preferences are von Neumann–Morgenstern is monotonic. In the domain of exchange economies with strictly increasing preferences, the Core and the Envy-Free SCR are also monotonic. In domains where indifference curves only cross once -the single-crossing condition – monotonicity vacuously holds. So Monotonicity, restrictive as it is, is worth a try. But before this, let us introduce a new assumption

- (NVP) A SCR f satisfies No Veto Power if $\forall \theta \in \Theta$,

$$\{aR_i(\theta)b, \quad \forall b \in A, \text{ for at least } n{-}1 \text{ agents}\} \rightarrow a \in F(\theta)$$

In other words, if there is an allocation which is top-ranked by, at least, $n-1$ agents, NVP demands that this allocation belongs to the SCR. This sounds like a reasonable property for large n. Also in exchange economies with strictly increasing preferences and more than two agents, NVP is vacuously satisfied because there is no top allocation for $n-1$ agents.

The following positive result, a relief after so many negative results, was stated and proved by Maskin [76], although his proof was incomplete.

Theorem 7 *If a SCR satisfies M and NVP is Nash implementable when $n > 2$.*

Proof (Sketch) Consider the following mechanism. $M_i = \Theta \times A \times \mathbb{N}$ where $\mathbb{N}$ is the set of natural numbers. The outcome function has three parts.

Rule 1 (Unanimity). If m is such that all agents announce the same state of the world, θ, the same allocation a with $a \in F(\theta)$ and the same integer, then $g(m) = a$.

Rule 2 (One Dissident). If there is only one agent whose message is different from the rest, this agent can choose any allocation that leaves her worse off, according to her preference as announced by others.

Rule 3 (Any other case). $a \in g(m)$ iff a was announced by the agent who announced the highest integer (ties are broken by an arbitrary rule).

Let us show that such a mechanism implements any SCR with the required conditions. Clearly if the true state is $\tilde{\theta}$, $m_i = (\tilde{\theta}, a, 1)$ with $a \in F(\tilde{\theta})$ is a Nash Equilibrium since no agent can gain by saying otherwise, so Condition 1 in the definition of Nash implementation holds. Let us now prove that Condition 2 there also holds. Suppose we have a Nash Equilibrium in Rule 1. Could it be an "untruthful" equilibrium? If so we have two cases. Either the announced preferences are a monotonic transformation of preferences at $\tilde{\theta}$, in which case, M implies that the announced allocation is also optimal at $\tilde{\theta}$. If they are not, there is an agent who can profitably deviate. Clearly, if equilibrium occurs in Rule 2, with, say, agent i as the dissident, any agent other than i can drive the mechanism to Rule 3, so it must be that all these agents are obtaining their most preferred allocation, which by NVP belongs to $F(\tilde{\theta})$. An equilibrium in Rule 3 implies that all agents are obtaining their most preferred allocation which, again by NVP belongs to $F(\tilde{\theta})$. □

The interpretation of the mechanism given in the proof of T.7 is that if everybody agrees on the state and the allocation is what the planner wants, this allocation is selected. If there is a dissident (a term due to Danilov [37]) she can make her case by choosing an allocation (a "test allocation") in her lower contour set, as announced by others. Finally, with more than one dissident, it's the jungle! Any agent can obtain her most preferred allocation by the choice of an integer. Typically, there is no equilibrium in this part of the mechanism. Notice that (M) is just used to eliminate unwanted equilibria.

The mechanism is an "augmented" revelation mechanism (a term due to Mookherjee and Reichelstein [83]), where the announcement of the state is complemented with the announcement of an allocation – this can be avoided if the SCR is single valued – and an integer. The final proof of T.7 was done independently by Williams [137], Repullo [104], Saijo [109] and McKelvey [95].

The case of two agent is more complicated because when an agent deviates from a common announcement and becomes a dissident, she converts the other agent into another dissident! As in T.3, W does the job, i. e. any SCR satisfying M, NVP and W is Nash implementable, see Moore and Repullo [85] and Dutta and Sen [40] for a full characterization. Again, the cases of two agents and more than two agents are different. In some areas of mathematics, such as statistics and differential equations the cases of two dimensions and more than two dimensions are also different. The relationship of these with the findings of implementation is not yet fully explored, see [108].

Under asymmetric information M is substituted by a – rather ugly – *Bayesian Monotonicity* (BM) condition which is a generalization of M to these environments. BM is again necessary and in conjunction with some technical conditions plus incentive compatibility, sufficient for implementation in BE. The interested reader can do no better than to read the account of these matters in [93]. It must be remarked that many well-known SCR – including Arrow–Debreu contingent commodities and some efficient SCR – do not satisfy BM and thus cannot be implemented in BE. However, the Rational Expectations Equilibria and the (interim) Envy-Free SCR satisfy BM, see Palfrey and Srivastava [94].

T.7 was the first positive finding of implementation theory. And it prompted researchers to be more ambitious: Can we implement without Monotonicity? An interesting observation, due to Matsushima [79] and Abreu and Sen [2], is that if agents have preferences representable by von Neumann–Morgenstern utility functions, *any* SCR can be "virtually implemented" in the sense that the set of allocations yielded by Nash equilibria is arbitrarily close to the set of desired allocations. This is because, as we saw before, any SCR mapping in the interior of $\mathcal{L}A$ is Monotonic. Thus allocations in the boundary can be arbitrarily approximated by allocations in the interior.

A more satisfying approach was introduced by Moore and Repullo [86] by considering subgame perfection as the solution concept. It is not possible to explain fully this approach here because it would take us too far; in particular the notion of a mechanism must be generalized to "stage mechanism". Instead, we give a result that conveys the force of subgame perfect implementation. It refers to public good economies with quasi-linear utility functions – where under dominant strategies the set of economies with inefficient outcomes is large – and with two individuals – where Nash implementability is harder to obtain.

Suppose that utility functions read $U_i = V(y, \theta_i) + m_i$ where $y \in Y \subseteq \Re$, $\theta_i \in \Theta_i$ with $\#\Theta_i < \infty$ and $m_i \in \Re$, $i = 1, 2$. The set of allocations $A = \{(y, m_1, m_2) \in Y \times \Re^2 / m_1 + m_2 \leq \omega\}$ where ω are the endowments of "money". Moore and Repullo [86] proved the following:

Theorem 8 *Any SCF is implementable in Subgame Perfect Equilibrium in the domain of economies explained above.*

Moore and Repullo proved that many SCR which could not be implemented in Nash Equilibrium can be implemented in Subgame Perfect Equilibrium. This is because subgames can be designed to kill unwanted equilibria without using monotonicity. Their result was improved upon by Abreu and Sen [2]. The problem with this approach is that the concept of subgame perfection is problematic because it requires that, no matter what has happened in past, in the remaining subgame, players are rational, even if this subgame was attained because some players made irrational choices.

The Moore–Repullo result was not only important by itself but it opened the way to the consideration of other equilibrium concepts that allow very permissive results. For instance, Palfrey and Srivastava [96] proved the following result

Theorem 8′ *Any SCR satisfying NVP is implementable in Undominated Nash Equilibrium.*

At this point, it seemed that by invoking the adequate refinement of Nash equilibrium, any SCR could be implemented. But the implementing mechanisms were getting weird and some people were beginning to get suspicious. Why and how is discussed in the next section.

Summing up the results obtained here, we have the following:

1. (Maskin) Monotonicity is a necessary and in many cases sufficient condition for implementation in Nash Equilibrium, see T.4 and 7. Similar results are obtained with Bayesian Monotonicity in Bayesian Equilibrium.
2. The Monotonicity requirements are not harmless. Many solution concepts do not satisfy it. Even worse, Monotonicity has some unpalatable consequences, see T.5–6.
3. Monotonicity can be avoided by considering stage games or refinements of Nash Equilibrium. Practically, any reasonable SCR can be implemented in this way, see T.8–8′.

The Limits of Design

So far we have assumed that there are no limits to what the designer can do. She can pick up any mechanism with no restrictions on its shape. This procedure, indeed, pushes the possibilities of design to the limit. But by doing this, we have learned a good deal about the limitations of the theory of implementation. It is fair to say that today the consensus is that there are *some extra properties* which should be considered when designing an implementing mechanism. We review here five approaches to this question.

Game-Theoretical Concerns. Jackson [59] was the first to point out that some mechanisms had unusual features from the point of view of game theory: Some subgames have no Nash equilibrium. Message spaces, which in the corresponding game become strategy spaces, are unbounded or open. Thus, in the integer game con-

sidered in T.7, if agents eliminate dominated strategies, each integer is dominated by the next highest one and no integer is undominated: Those agents who eliminate dominated strategies are unable to make a choice. These constructions eliminate unwanted equilibria, which as we saw before, is the problem with Nash implementation. Jackson illustrates his point by showing that under no restrictions on mechanisms, any SCR can be implemented in undominated strategies, a weak solution concept. Then he requires that the mechanism be *Bounded* in the following sense: whenever a strategy m_i is dominated, there is another strategy dominating m_i and which is undominated. He shows that implementation in undominated strategies with bounded mechanisms result many times in incentive compatibility, which as we saw in Sect. "The Main Concepts" is a hard requirement. This shows the bite of the boundedness assumption. However, in the case of implementation with undominated Nash equilibrium, the boundedness assumption has little impact, see [61,126]. The first of these papers introduced a related requirement, the *Best Response Property*: for every strategy played by the other agents, each agent has a best response.

Natural Mechanisms Given that we have run so far from the kind of mechanisms we are used to, it seems reasonable to ask what can be implemented by mechanisms that resemble real-life mechanisms. These mechanisms must be simple too because simplicity is an important characteristic in practice. Let us call them *Natural Mechanisms*. Dutta, Sen and Vohra [41] consider mechanisms in which messages are prices and quantities and thus resemble market mechanisms. Their approach was refined by Saijo, Tatamitani and Yamato [111] who demanded the best response property as well. They showed that several well-known SCR such as the (constrained) Walrasian, are implementable in Nash equilibrium. Beviá, Corchón and Wilkie [21] showed that in Bertrand-like market games, the Walrasian SCR is implementable in Nash and Strong equilibrium, showing that the fear of coalitions destabilizing market outcomes is, at least, partially unwarranted. Sjöström [127] considered quantity mechanisms, reminiscent of those used by Soviet planners, with negative results about what these mechanisms can achieve. In public good economies, Corchón and Wilkie [31] and Peleg [97] introduced a market mechanism implementing Lindahl allocations in Nash and Strong equilibrium. The mechanism works because Lindahl prices have to add up to the marginal cost. If an agent pretends to free ride she decreases the quantity of the public good. Here, contrary to Samuelson's dictum it is in the selfish interest of each person to give true signals. Pérez-Castrillo and Wettstein [98] offered a bidding mechanism that implements efficient allocations when choosing between a finite number of public projects. They also applied these ideas to back up Shapley value.

Credibility Another implicit assumption is that once the mechanism is in place, there is no way to stop it. Thus, if for some m, $g(m)$ is a "universally worst outcome", the planner has to deliver this allocation even if she is trying to implement a Pareto Efficient allocation. Is this a credible procedure? In many cases, if the planner is a real person it seems that she would do her best to avoid $g(m)$! Here we have two possibilities: Either we identify additional constraints on the planner that look reasonable or we jump to model the planner as a full-fledged player. The first road leads us to identify a subset of allocations of A, say X, which can never be used by the mechanism. For instance, in Chakravorty, Corch ón and Wilkie [25] X is the set of allocations that are never selected by the SCR for some state of the world, i. e. $X = \{a \in A / \nexists \theta \in \Theta, \ \ a \in F(\theta)\}$. The motivation for this definition is that it hardly seems credible that the planner can choose an allocation that is never intended to be implemented. Redefining the allocation set as $\mathbf{A} \equiv A \backslash X$ the definitions of a mechanism and an implementable SCR can be easily translated in this framework. However, depending on the domain, SCR that are monotonic when defined on A are no longer monotonic when defined on $\mathbf{A}$: For instance, the (constrained) Walrasian SCR. Thus, these SCR can not be implemented when the planner can only use allocations in $\mathbf{A}$. A weakness of this approach is that the list of reasonable constraints on allocations may be large. The second possibility drives us to model implementation as a signaling game where the planner receives signals – messages – from the agents, updates her beliefs and then chooses an allocation which maximizes her expected utility [11]. Again, some SCR that are Nash implementable, are not implementable in this framework. However, in this case there are SCR that are not Nash implementable but are implementable in this framework. This is because the model takes a basic assumption of game theory to the limit, namely, that agents know the strategies of other players. In this case, the planner knows if a report on agents' types is truthful or not before the allocation is delivered!

Renegotiation Another strong assumption is that the mechanism prescribes actions that can not be changed by agents. This contradicts experiences such as black markets where agents trade on the existing goods (Hammond [49]). A way of modeling this is to assume that

agents are able to renegotiate some allocations (Maskin and Moore [77]. Renegotiation in a different context was considered by Rubinstein and Wolinsky [107]). Assuming that agents have complete information, this is formalized by means of the concept of a reversion function. This function, say r, maps each allocation and each state of the world into a new allocation, i.e. $r\colon A \times \Theta \to A$. The reversion function induces new preferences, called *reverted preferences* (this is the "translation principle" in Maskin and Moore [77]). Notice that reverted preferences are state dependent even if preferences are not. Formally, given a reversion function r, the *reversion of* $R(\theta)$, denoted by $R^r(\theta)$ is defined as $aR_i^r(\theta)b \Leftrightarrow r(a,\theta)R_i(\theta)r(b,\theta), \forall a,b \in A, \forall i \in I$. Given a reversion function r, we can interpret that agents' preferences are the reverted preferences. Then, all definitions given before can be adapted to this case. Again, SCR that were monotonic there, are not so in this framework and vice versa. See Jackson and Palfrey [62] for applications. An extension to the case where there are several renegotiation functions is given by Amoros [6]. A weakness of this approach is that models renegotiation as a "black box".

Multiple Implementation Maskin [75] was the first to realize that the notion of implementation requires the planner to know the solution concept used by the agents to analyze the game. He proposed the notion of "Double Implementation" where a SCR was implemented at the same time in Nash and Strong equilibria. He showed that many Nash implementable SCR indeed are doubly implementable. We have seen in Point 2 above that the (constrained) Walrasian and Lindahl SCR are doubly implementable by natural mechanisms. They are also doubly implementable by abstract mechanisms, Schmeidler [116]. Double implementation also occurs with several solutions to the problem of the commons, Shin and Suh [124] and Pigouvian Taxes, Alcalde, Corchón and Moreno [4]. Yamato [139] introduced another type of double implementation by requiring implementation in Nash and Undominated Nash Equilibria [139]. He showed that in a large class of exchange economies with at least three agents, monotonicity is necessary and sufficient for double implementation. Saijo, Sjöström and Yamato [112] considered implementation in Dominant Strategies and Nash Equilibrium. Clearly, other variations of the idea of Double Implementation are possible, see Point 4 in Sect. "Unsolved Issues and Further Research" below.

Summing up, it is now clear that implementing mechanisms can not be just "anything". Their features matter. Demanding that mechanisms satisfy the best response property, be simple, not use extreme allocations, be robust to the possibility of renegotiation and implement in several equilibrium concepts makes our lives more difficult but makes our models a great deal better.

Unsolved Issues and Further Research

Implementation with State Dependent Feasible Sets

A motivation of implementation theory was to study the possibility of socialism. However, *all* the results presented in this survey refer to environments where the feasible set is given, a far cry from any kind of planning procedure. In fact, there are only a handful of papers dealing with implementation when the feasible set is unknown: Postlewaite [99] and Sertel and Sanver [123] studied manipulation of endowments. Hurwicz, Maskin and Postlewaite [57] studied implementation assuming that endowments/production possibilities can be hidden or destroyed but never exaggerated. Instead of a mechanism we have a collection of state dependent mechanisms each meant for an economy. After the mechanism is played, production capabilities are shown, e.g. endowments are put on the table. This idea was worked out in a series of papers by Hong on private good economies, Hong [52], and by Tian on public good economies, Tian and Li [134] . Serrano and Vohra [120] worked out implementation of the core and Dagan, Serrano and Volij [36] of taxation methods. And that's all folks! Why has such an important issue been almost neglected? My explanation is that the proposed mechanisms are difficult to understand. Another approach has been tried by Corchón and Triossi [35] where a reversion function takes care of restoring feasibility when messages lead to unfeasible allocations. The approach is tractable and simpler but relies on the black box of the renegotiation function.

Sociological Factors/Bounded Rationality

So far, all the solution concepts describing the behavior of agents are game-theoretical. In recent years, we have seen a host of equilibrium concepts based on "irrational" agents. It would be interesting to see what SCR can be implemented with these forms of behavior. Eliaz [42] considers "Fault Tolerant" implementation where a subset of players ("faulty players") fail to achieve their optimal strategies. Under complete information, No Veto Power and a strong form of Monotonicity are sufficient for implementation when the number of faulty players is less than $n/2-1$, $n>2$. Matsushima [80] shows that a small preference for honesty is sufficient to knock down unwanted equilibria, see also Corchón and Herrero [34].

Dynamic Implementation

The theory presented here is static but there are some papers dealing with implementation in dynamic set-ups. We mention a few: Freixas, Guesnerie and Tirole [43] studied the "Ratchet Effect", where firms underproduce for fear of being asked to do too much in the future. Kalai and Ledyard [63] showed that if the planner is sufficiently patient, every SCR is dominant-strategy implementable. Burguet [22] showed that the revelation principle does not hold when outcomes are chosen in several periods. Candel [24] proved a revelation principle in a model where a public good is produced in two periods. Finally, Cabrales [23] and Sandholm [114] studied implementation in an evolutionary setting. A related topic is that of complexity, see Conitzer and Sandholm [28]. It seems likely that a dynamic theory of incentives will bring new insights and will need new analytical tools.

Robustness Under Incomplete Information

When designing a mechanism, sometimes the planner does not know the structure of information. In this case a mechanism must implement regardless the structure of information, i. e. priors of agents, type spaces, etc. Corchón and Ortuño-Ortín [30] approached the problem by assuming that the economy is composed of "islands" and that there is complete information inside each island. A mechanism robustly implements a SCR if it does it in BE for every possible prior (compatible with the island assumption) and in Uniform Nash Equilibrium. The latter requires that an equilibrium strategy for an agent must be the best reply to what other agents in the island play and to any possible message sent by agents outside the island when they follow their equilibrium strategies (D'Aspremont and Gerard-Varet [10]). They showed that any SCR satisfying M and NVP is robustly implementable (a later contribution by Yamato [140] showed that Robust and Nash Implementation coincide in this framework). The same concern has been approached in a series of papers by Bergemann and Morris (see e. g. [19]) where they ask SCR to be implemented whatever the players' beliefs and higher order beliefs about other player' types. Artemov, Kunimoto and Serrano [8] require implementation for the payoff type space and the space of first-order beliefs about other agents' payoff types. They obtain very permissive results.

In a different vein, Koray [65], has argued that, since priors are not contractible, the regulator needs to be regulated in order to stop her from manipulating the priors. He shows that the outcomes of this game vary over a wide spectrum. Again the need of prior-free implementation is clear.

Answers to the Questions

1. Yes. We already saw in 4.3, Point 2, that "Bertrand-like" mechanisms implement the Constrained Walrasian SCR in Nash and Strong equilibrium. But this is not all: Schmeidler [116] exploited the connection between price taking – which underlies Walrasian equilibrium – and "strategy taking", which underlies Nash and Strong equilibrium and obtained double implementation by a mechanism which does not resemble the market. Implementation of the Lindahl SCR by an abstract mechanism was obtained by Walker [136] building on previous papers by Groves and Ledyard and Hurwicz. Unfortunately, these positive results turn negative when we consider Arrow–Debreu contingent commodities, Chattopadhyay, Corchón and Naeve [26] and Serrano and Vohra [121].
2. A merger affects social welfare in two ways: Positively, from cost savings and negatively, from restricting competition. The first effect is uncertain and, by now, I do not have to convince you that we should take with utmost caution all announcements made by firms concerning cost savings. Corchón and Faulí-Oller [33] show that under a condition that is fulfilled in several standard IO models, the SCR that maximize social surplus can be implemented by a dominance solvable mechanism with budget balance.
3. There is a very simple mechanism which attains maximum surplus, Loeb and Magath [72]. But in this mechanism the monopolist receives all the surplus and the demand function must be known by the planner. These points were worked out by subsequent contributions from Baron and Myerson, Lewis and Sappington, Sibley and others.
4. By now the reader should know the difficulties of implementing efficient public decisions. When information is exclusive this is impossible, even though an approximate efficient decision can be obtained when the number of agents is large. When information is complete, we have seen several examples of mechanisms implementing efficient outcomes.
5. There is no difference between implementing market or fair outcomes. Both have to pass the same tests, i. e. incentive compatibility, monotonicity and simplicity/credibility of design. In exchange economies, Thomson [133] presents a simple and elegant mechanism that implements envy-free allocations in Nash Equilibrium. In cooperative production, Corchón and Puy [32] pre-

sented a family of mechanisms that implement in Nash Equilibrium any efficient SCR where the distribution of rewards is a continuous function of efforts.

6. Yes! An uninformed planner can set up a mechanism that yields efficient outcomes in circumstances where the market yields inefficient allocations, i. e. under externalities or public goods see Point 5 in 4.3 above. All we need is non-exclusive information and that the SCR be Monotonic, the latter requirement can be skipped under refinements of Nash Equilibrium.
7. Not completely. Suppose complete information among three or more judges and that they all perceive the same quality of a given performance. Clearly, truth is an equilibrium, because if all judges minus you tell the truth you cannot change the outcome by saying something different. Unfortunately, any unanimous announcement is also an equilibrium by the same reason. Thus we are in a situation akin to T.3. Fortunately, if preferences of judges fulfill certain restrictions, full implementation of the true ranking of ice skaters is possible, because Monotonicity and No Veto Power hold so T.7 applies, Amoros, Corchón and Moreno [5]. If judges have differential information, the truth is no longer implementable as suggested by T.2″. See Gerardi, McLean and Postlewaite [44] for further insights and references on this problem.
8. ????? Do you think that we have all answers? This is just economics!! In any case, the application of Implementation to environmental policies may be a topic of great importance in the future, see Baliga and Maskin [12].
   Finally I will tell you why I like implementation theory so much. Firstly, the implementation model solves the problems of the General Equilibrium model mentioned in Sect. "Brief History of Implementation Theory", namely: 1: It models a general economic system. 2: All variables are endogenously determined by the interaction of agents. 3: Agents incentives are carefully modeled and are taken fully into account. Secondly, the theory is not based on assumptions like convexity or continuity/differentiability which, no matter how much we are used to them, are very stringent. By the way, a beautiful paper by Laffont and Maskin (referenced in their [68] survey) developed incentive compatibility in a differentiable framework.

Acknowledgments

I am grateful to Pablo Amorós, Claude d'Aspremont, Carmen Beviá, Luis Cabral, Eric Maskin, Carlos Pimienta, Socorro Puy, Tömas Sjöstrom, William Thomson, Matteo Triossi, Galina Zudenkova and an anonymous referee for helpful suggestions and to the Spanish Ministery of Education for financial support under grant SEJ2005-06167. I also thank the Department of Economics, Stern School of Business, NYU, for their hospitality while writing this survey. This survey is dedicated to Leo Hurwicz to celebrate his 90th birthday and his Nobel Prize and to the memory of those who contributed to the area and are no longer with us: Louis-André Gerard-Varet, Jean-Jacques Laffont, Richard McKelvey and Murat Sertel.

Bibliography

1. Abreu D, Sen A (1990) Virtual implementation in Nash equilibrium. Econometrica 59:997–1021
2. Abreu D, Sen A (1991) Subgame Perfect Implementation: A Necessary and Almost Sufficient Condition. J Econ Theory 50:285–299
3. Akerlof G (1970) The Market for Lemons: Qualitative Uncertainty and the Market Mechanism. Q J Econ 84:488–500
4. Alcalde J, Corchón L, Moreno B (1999) Pigouvian Taxes: A Strategic Approach. J Public Econ Theory 1 2:271–281
5. Amoros P, Corchón L, Moreno B (2002) The Scholarship Assignment Problem. Games Econ Behav 38:1–18
6. Amoros P (2004) Nash Implementation and Uncertain Renegotiation. Games Econ Behav 49:424–434
7. Arrow K (1979) The Property Rights Doctrine and Demand Revelation under Incomplete Information. Technical report No. 243, IMSSS, Stanford University, Stanford
8. Artemov G, Kunimoto T, Serrano R (2007) Robust virtual implementation with incomplete information: Towards a reinterpretation of the Wilson doctrine. W.P. 2007–06, Brown University, Providence
9. d'Aspremont C, Gerard-Varet L-A (1975) Individual Incentives and Collective Efficiency for an Externality Game with Incomplete Information. CORE DP 7519. Universite Catholique de Louvain, Louvain-la-Neuve
10. d'Aspremont C and L-A Gerard-Varet (1979) Incentives and Incomplete Information. J Public Econ 11:25–45
11. Baliga S, Corchón L, Sjöström L, Sjöström T (1997) The theory of implementation when the planner is a player. J Econ Theory 77:15–33
12. Baliga S, Maskin E (2003) Mechanism Design for the Environment. In: Maler K, Vincent JR (eds) Handbook of Environmental Economics, vol 1, ch 7, pp 305–324
13. Baliga S, Sjöström T (2007) Mechanism Design: Recent Developments. In: Blume L, Durlauf S (eds) The New Palgrave Dictionary of Economics, 2nd edn. Palgrave Macmillan, New York
14. Barberá S (1983) Strategy-proofness and pivotal voters: A direct proof of the Gibbard–Satterthwaite theorem. Int Econ Rev 24:413–417
15. Barberá S, Peleg B (1990) Strategy-proof voting schemes with continuous preferences. Soc Choice Welf 7:31–38
16. Barberá S, Sonnenschein H, Zhou L (1991) Voting by Committees. Econometrica 59(3):595–609
17. Barone E (1935) The Ministry of Production in a Collectivist State. Translated from the italian and reprinted in F. A. von Hayek Collectivist Economic Planning. Routledge and Keegan, London

18. Benoit JP (2000). The Gibbard–Satterthwaite theorem: a simple proof. Econ Lett 69:319–322
19. Bergemann D, Morris S (2005) Robust Mechanism Design. Econometrica 73:1771–1813
20. Beviá C, Corchón L (1995) On the Generic Impossibility of Truthful Behavior. Econ Theory 6:365–371
21. Beviá C, Corchón L, Wilkie S (2003) Implementation of the Walrasian Correspondence by Market Games. Rev Econ Des 7:429–442
22. Burguet R (1990) Revelation in Informational Dynamic Settings. Econ Lett 33:237–239. Corrigendum (1994) 44:451–452
23. Cabrales A (1999) Adaptive Dynamics and the Implementation Problem with Complete Information. J Econ Theory 86:159–184
24. Candel F (2004) Dynamic Provision of Public Goods. Econ Theory 23:621–641
25. Chakravorty B, Corchón L, Wilkie S (2006) Credible Implementation. Games Econ Behav 57:18–36
26. Chattopadhyay S, Corchón L, Naeve J (2000) Contingent Commodities and Implementation. Econ Lett 68:293–298
27. Clarke E (1971) Multipart pricing of Public Goods. Public Choice 11:17–33
28. Conitzer V, Sandholm T (2002) Complexity of Mechanism Design. In: Proceedings of the 18th Annual Conference on Uncertainty in Artificial Intelligence (UAI-02), Edmonton, Canada
29. Corchón L (1996) The Theory of Implementation of Socially Optimal Decisions in Economics. St. Martin's Press, New York
30. Corchón L, Ortuño-Ortín I (1995) Robust Implementation under Alternative Information Structures. Econ Des 1:159–171
31. Corchón L, Wilkie S (1996) Doubly Implementing the Ratio Correspondence by a Market Mechanism. Rev of Econ Des 2:325–337
32. Corchón L, Puy S (2002) Existence and Nash Implementation of Efficient Sharing Rules for a Commonly Owned Technology. Soc Choice Welf 19:369–379
33. Corchón L, Faulí-Oller R (2004) To Merge or not to Merge: That is the Question. Rev Econ Des 9:11–30
34. Corchón L, Herrero C (2004) A Decent Proposal. Spanish. Econ Rev 6:107–125
35. Corchón L, Triossi M (2005) Implementation with State Dependent Feasible Sets: A Renegotiation Approach. WP, Carlos Alberto Notebooks, #24, Torino
36. Dagan N, Serrano R, Volij O (1999) Feasible implementation of taxation methods. Rev Econ Des 4:57–72
37. Danilov V (1992) Implementation via Nash Equilibria. Econometrica 60(1):43–56
38. Dasgupta P, Hammond P, Maskin E (1979) The implementation of Social Choice Rules: Some Results on Incentive Compatibility. Rev Econ Stud 46:185–216
39. Dutta B, Sen A (1991) Implementation under strong equilibrium. A complete characterization. J Math Econ 20:49–67
40. Dutta B, Sen A (1991) A Necessary and Sufficient Condition for Two-Person Nash Implementation. Rev Econ Stud 58:121–128
41. Dutta B, Sen A, Vohra R (1995) Nash Implementation through Elementary Mechanisms in Economic Environments. Econ Des 1:173–204
42. Eliaz K (2002) Fault Tolerant Implementation. Rev Econ Stud 69:589–610
43. Freixas X, Guesnerie R, Tirole J (1985) The Ratchet Effect. Rev Econ Stud 52:173–191
44. Gerardi D, McLean R, Postlewaite A (2005) Aggregation of Expert Opinions. Cowles Foundation Discussion Paper # 1503, Yale
45. Gibbard A (1973) Manipulation of Voting Schemes: A General Result. Econometrica 41:587–602
46. Green J, Laffont JJ (1979) Incentives in Public Decision Making. Elsevier, Amsterdam
47. Gresik T, Satterthwaite M (1989) The Rate at Which a Simple Market Converges to Efficiency as the Number of Traders Increases. J Econ Theory 48:304–332
48. Groves T (1973) Incentives in Teams. Econometrica 41:617–631
49. Hammond P (1987) Markets as Constraints: Multilateral Incentive Compatibility in Continuum Economies. Rev Econ Stud 54:399–412
50. Harris M, Townsend R (1981) Resource Allocation Under Asymmetric Information. Econometrica 49:33–64
51. Harsanyi J (1967). Games with Incomplete Information Played by 'Bayesian' Players. Parts I, II and III. Management Science 14:159–182, 320–334 and 486–502
52. Hong L (1998) Feasible Bayesian Implementation with State Dependent Feasible Sets. J Econ Theory 80:201–221
53. Hurwicz L (1959) Optimality and Informational Efficiency in Resource Allocation Processes. In: Arrow KJ (ed) Mathematical Methods in the Social Sciences. Stanford University Press, Stanford, pp 27–46
54. Hurwicz L (1972) On Informationally Decentralized Systems. In: Radner R, McGuire CB (eds) Decision and Organization: A Volume in Honor of Jacob Marshak. Elsevier, Amsterdam, pp 297–336
55. Hurwicz L (1979) On Allocations Attainable Through Nash Equilibria. J Econ Theory 21:40–65
56. Hurwicz L, Walker M (1990) On the Generic Non-Optimality of Dominant Strategy Mechanisms. Econometrica 58(3):683–704
57. Hurwicz L, Maskin E, Postlewaite A (1995) Feasible Nash Implementation of Social Choice Rules when the Designer does not know Endowments or Production set. In: Ledyard J (ed) The Economics of Informational Decentralization: Complexity, Efficiency and Stability. Kluwer, Amsterdam
58. Jackson M (1991) Bayesian Implementation. Econometrica 59:461–477
59. Jackson M (1992) Implementation in Undominated Strategies: A Look to Bounded Mechanisms. Rev Econ Stud 59:757–775
60. Jackson MO (2001) A Crash Course in Implementation Theory. Soc Choice Welf 18:655–708
61. Jackson MO, Palfrey T, Srivastava S (1994) Undominated Nash Implementation in Bounded Mechanisms. Games Econ Behav 6:474–501
62. Jackson MO, Palfrey T (2001) Voluntary implementation. J Econ Theory 98:1–25
63. Kalai E, Ledyard J (1988) Repeated Implementation. J Econ Theory 83:308–317
64. Klemperer P (1999) Auction theory, a guide to the literature. J Econ Surv 13:227–268
65. Koray S (2005) The Need of Regulating a Bayesian Regulator. J Regul Econ 28:5–21

66. Krishna V, Perry M (1997) Efficient mechanism design. Unpublished paper, Penn State University, Pennsylvania
67. Krishna V (2002) Auction Theory. Academic Press, San Diego
68. Laffont JJ, Maskin E (1982) The theory of incentives: An overview. In: Hildenbrand W (ed) Advances in Economic Theory. 4th World Congress of the Econometric Society. Cambridge University Press, Cambridge
69. Laffont JJ, Martimort D (2001) The Theory of Incentives: The Principal-Agent Model. Princeton University Press, New Jersey
70. Lange O (1936). On the Economic Theory of Socialism. Rev Econ Stud 4:53–71, 123–142
71. Ledyard J, Roberts J (1974) On the Incentive Problem with Public Goods. Discussion Paper 116, Centre for Mathematical Studies in Economics and Management Science. Northwestern University
72. Loeb M, Magath W (1979) A Decentralized Method for Utility Regulation. J Law Econ 22:399–404
73. Ma A, Moore J, Turnbull S (1988) Stop Agents from Cheating. J Econ Theory 46:355–372
74. Makowski L, Mezzetti C (1993) The possibility of efficient mechanisms for trading an indivisible object. J Econ Theory 59:451–465
75. Maskin E (1985) The theory of implementation in Nash equilibrium: A Survey. In: Hurwicz L, Schmeidler D, Sonnenschein H (eds) Social Goals and Social Organization. Cambridge University Press, Cambridge, pp 173–204
76. Maskin E (1999) Nash Equilibrium and Welfare Optimality. (Circulating in working paper version since 1977) Rev Econ Stud 66(1):23–38
77. Maskin E, Moore J (1999) Implementation with Renegotiation. Rev Econ Stud 66:39–56
78. Maskin E, Sjöström T (2002) Implementation Theory. In: Arrow KJ, Sen AK (ed) Chapter 5 in Handbook of Social Choice Welfare. Elsevier, Amsterdam
79. Matsushima H (1988) A New Approach to the Implementation Problem. J Econ Theory 45:128–144
80. Matsushima H (2008) Role of honesty in full implementation. J Econ Theory 139:353–359
81. McKelvey R (1989) Game Forms for Nash Implementation of General Social Choice Correspondences. Soc Choice Welf 6:139–156
82. Mirless J (1971) An Exploration in the Theory of Optimum Income Taxation. Rev Econ Stud 38:175–208
83. Mookherjee D, Reichelstein S (1990) Implementation via augmented revelation mechanisms. Rev Econ Stud 57:453–475
84. Moore J (1992) Implementation, contracts and renegotiation in environments with complete information in Advances in Economic Theory. In: Laffont JJ (ed) 4th World Congress of the Econometric Society, vol I. Cambridge University Press, Cambridge, pp 182–282
85. Moore J, Repullo R (1990) Nash Implementation: A Full Characterization. Econometrica 58:1083–1089
86. Moore J, Repullo R (1988) Subgame Perfect Implementation. Econometrica 56:1191–1220
87. Moulin H (1979) Dominance Solvable Voting Schemes. Econometrica, 47(6):1337–1351
88. Muller E, Satterthwaite M (1977) The equivalence of strong positive association and strategy-proofness. J Econ Theory 14:412–418
89. Myerson R (1979) Incentive Compatibility and the Bargaining Problem. Econometrica 47:61–73
90. Myerson R (1981): Optimal auction design. Math Oper Res 6:58–73
91. Myerson R (1985) Bayesian Equilibrium and Incentive Compatibility: An Introduction. In: Hurwicz L, Schmeidler D, Sonnenschein H (eds) Social Goals and Social Organization, Chapter 8. Cambridge University Press, Cambridge
92. Myerson R, Satterthwaite MA (1983): Efficient mechanisms for bilateral trading. J Econ Theory 29:265–281
93. Palfrey TR (2002) Implementation Theory. In: Aumann RJ, Hart S (eds) Handbook of Game Theory with Economic Applications, vol III. Elsevier Science, New York, pp 2271–2326
94. Palfrey T, Srivastava S (1987) On Bayesian Implementable Allocations. Rev Econ Stud 54:193–208
95. Palfrey T, Srivastava S (1989) Implementation with Incomplete Information in Exchange Economies. Econometrica 57:115–134
96. Palfrey T, Srivastava S (1991) Nash Implementation using Undominated Strategies. Econometrica 59:479–501
97. Peleg B (1996) Double implementation of the Lindahl equilibrium by a continuous mechanism. Rev Econ Des 2(1):311–324
98. Pérez-Castrillo D, Wettstein D (2002) Choosing Wisely: A Multi-Bidding Approach. Am Econ Rev 92:1577–1587
99. Postlewaite A (1979) Manipulation via endowments. Rev Econ Stud 46:255–262
100. Postlewaite A, Schmeidler D (1986) Implementation in Differential Information Economies. J Econ Theory 39:14–33
101. Prize Committee of the Royal Swedish Academy of Sciences (2007) Mechanism Design Theory. Royal Swedish Academy of Sciences, Stockholm, October 15
102. Reiter S (1977) Information and Performance in the (New) Welfare Economics. Am Econ Rev 67:226–234
103. Repullo R (1986) On the Revelation Principle under Complete and Incomplete Information. In: Binmore K, Dasgupta P (eds) Economics Organizations as Games. Basil Blackwell, Oxford
104. Repullo R (1987) A simple proof of Maskin theorem on Nash implementation. Soc Choice Welf 4:39–41
105. Roth AE (2008) What have we learned from market design? Hahn Lecture, Econ J 118:285–310
106. Rothchild M, Stiglitz J (1976) Equilibrium in Competitive Insurance Markets: An essay on the Economics of Imperfect Information. Q J Econ 90:629–650
107. Rubinstein A, Wolinsky A (1992) Renegotiation-proof implementation and time preferences. Am Econ Rev 82:600–614
108. Saari D (1987) The Source of Some Paradoxes from Social Choice and Probability. J Econ Theory 41:1–22
109. Saijo T (1988) Strategy space reduction in Maskin's theorem. Econometrica 56:693–700
110. Saijo T (1991) Incentive Compatibility and Individual Rationality in Public Good Economies. J Econ Theory 55:103–112
111. Saijo T, Tatamitani TY, Yamato T (1996) Toward Natural Implementation. Int Econ Rev 37(4):949–980
112. Saijo T, Sjöström T, Yamato T (2007) Secure Implementation. Theor Econ 2:203–229
113. Samuelson PA (1954) The Pure Theory of Public Expenditure. Rev Econ Stat 36:387–9
114. Sandholm W (2007) Pigouvian Pricing and Stochastic Evolutionary Implementation. J Econ Theor 132:367–382

115. Satterthwaite MA (1975) Strategy-Proofness and Arrows conditions: Existence and Correspondence Theorems for Voting procedures and Social Choice Functions. J Econ Theory 10:187–217
116. Schmeidler D (1980) Walrasian Analysis via Strategic Outcome Functions. Econometrica 48:1585–1593
117. Sen A (2001) Another direct proof of the Gibbard–Satterthwaite Theorem. Econ Lett 70:381–385
118. Serizawa S (2002) Inefficiency of strategy-proof rules for pure exchange economies. J Econ Theory 106:219–241
119. Serizawa S, Weymark J (2003) Efficient Strategy-Proof Exchange and Minimum Consumption Guarantees. J Econ Theory 109:246–263
120. Serrano R, Vohra R (1997) Non Cooperative Implementation of the Core. Soc Choice Welf 14:513–525
121. Serrano R, Vohra R (2001) Some Limitation of Bayesian Virtual Implementation. Econometrica 69:785–792
122. Serrano R (2004) The Theory of Implementation of Social Choice Rules. SIAM Rev 46:377–414
123. Sertel M, Samver R (1999) Equilibrium outcomes of Lindahl-endowment pretension games. Eur J Politi Econ 15:149–162
124. Shin S, Suh S-C (1997) Double Implementation by a Simple Game Form in the Commons Problem. J Econ Theory 77:205–213
125. Sjöström T (1993) Implementation in Perfect Equilibria. Soc Choice Welf 10:97–106
126. Sjöström T (1994) Implementation in Undominated Nash Equilibrium without Integer Games. Games Econ Behav 6:502–511
127. Sjöström T (1996) Implementation by Demand Mechanisms, Econ Des 1:343–354
128. Spence M (1973) Job Market Signalling. Q J Econ 87:355–374
129. Taylor FM (1929) The Guidance of Production in a Socialist State. Am Econ Rev 19:1–8
130. Thomson W (1985) Manipulation and Implementation in Economics. Unpublished manuscript. Rochester, New York
131. Thomson W (1987) The vulnerability to manipulative behavior of economic mechanisms designed to select equitable and efficient outcomes. In: Groves T, Radner R, Reiter S (eds) Information, Incentives and Economic Mechanisms. University of Minnesota Press, Minnesota, pp 375–396
132. Thomson W (1996) Concepts of Implementation. Jpn Econ Rev 47:133–143
133. Thomson W (2005) Divide and Permute. Games Econ Behav 52:186–200
134. Tian G, Li Q (1995) On Nash-Implementation in the Presence of Withholding. Games Econ Behav 9:222–233
135. Vickrey W (1961) Counterspeculation, Auctions and Competitive Sealed Tenders. J Finance 16:8–37
136. Walker M (1981) A Simple Incentive Compatible Scheme for Attaining Lindahl Allocations. Econometrica 49:65–73
137. Williams S (1986) Realization of Nash implementation: Two Aspects of Mechanism Design. Econometrica 54:139–151
138. Williams S (1999): A Characterization of Efficient, Bayesian Incentive Compatible Mechanisms. Econ Theory 14:155–180
139. Yamato T (1993) Double Implementation in Nash and Undominated Nash Equilibria. J Econ Theory 59(2):311–323
140. Yamato T (1994) Equivalence of Nash Implementability and Robust Implementability with Incomplete Information. Soc Choice Welf 11:289–303

Infinite Dimensional Controllability

Olivier Glass
Laboratoire Jacques-Louis Lions,
Université Pierre et Marie Curie, Paris, France

Article Outline

Glossary
Definition of the Subject
Introduction
First Definitions and Examples
Linear Systems
Nonlinear Systems
Some Other Problems
Future Directions
Bibliography

Glossary

Infinite dimensional control system A infinite dimensional control system is a dynamical system whose state lies in an infinite dimensional vector space—typically a Partial Differential Equation (PDE)—and depending on some parameter to be chosen, called the control.

Exact controllability The exact controllability property is the possibility to steer the state of the system from any initial data to any target by choosing the control as a function of time in an appropriate way.

Approximate controllability The approximate controllability property is the possibility to steer the state of the system from any initial data to a state arbitrarily close to a target by choosing a suitable control.

Controllability to trajectories The controllability to trajectories is the possibility to make the state of the system join some prescribed trajectory by choosing a suitable control.

Definition of the Subject

Controllability is a mathematical problem, which consists of determining the targets to which one can drive the state of some dynamical system, by means of a control parameter present in the equation. Many physical systems such as quantum systems, fluid mechanical systems, wave propagation, diffusion phenomena, etc., are represented by an infinite number of degrees of freedom, and their evolution follows some partial differential equation (PDE). Finding active controls in order to properly influence the dynamics

of these systems generate highly involved problems. The control theory for PDEs, and among this theory, controllability problems, is a mathematical description of such situations. Any dynamical system represented by a PDE, and on which an external influence can be described, can be the object of a study from this point of view.

Introduction

The problem of controllability is a mathematical description of the general following situation. We are given an evolution system (typically a physical one), on which we can exert a certain influence. Is it possible to use this influence to make the system reach a certain state? More precisely, the system takes generally the following form:

$$\dot{y} = F(t, y, u) , \tag{1}$$

where y is a description of the state of the system, $\dot{y}$ denotes its derivative with respect to the time t, and u is the control, that is, a parameter which can be chosen in a suitable range. The standard problem of controllability is the following. Given a time $T > 0$, an initial state y_0 and a target y_1, is it possible to find a control function u (depending on the time), such that the solution of the system, starting from y_0 and provided with this function u reaches the state y_1 at time T?

If the state of the system can be described by a finite number of degrees of freedom (typically, if it belongs to an Euclidean space or to a manifold), we call the problem finite dimensional. The present article deals with the case where y belongs to an infinite-dimensional space, typically a Banach space or a Hilbert space. Hence the systems described here have an infinite number of degrees of freedom. The potential range of the applications of the theory is extremely wide: the models from fluid dynamics (see for instance [52]) to quantum systems (see [6,54]), networks of structures (see [23,44]), wave propagation, etc., are countless.

In the infinite dimensional setting, the Eq. (1) is typically a partial differential equation, where F acts as a differential operator on the function y, and the influence of u can take multiple different forms: typically, u can be an additional (force) term in the right-hand side of the equation, localized in a part of the domain; it can also appear in the boundary conditions; but other situations can clearly be envisaged (we will describe some of them).

Of course the possibilities to introduce a control problem for partial differential equations are virtually infinite. The number of results since the beginning of the theory in the 1960s has been constantly growing and has reached huge dimensions: it is, in our opinion, hopeless to give a fair general view of all the activity in the theory. This paper will only try to present some basic techniques connected to the problem of infinite-dimensional controllability.

As in the finite dimensional setting, one can distinguish between the linear systems, where the partial differential equation under view is linear (as well as the action of the control), and the nonlinear ones.

Structure of the Paper

In Sect. "First Definitions and Examples", we will introduce the problems that are under view and give some examples. The main parts of this paper are Sects. "Linear Systems" and "Nonlinear Systems", where we consider linear and nonlinear systems, respectively.

First Definitions and Examples

General Framework

We define an infinite-dimensional control system as the following data:

1. an evolution system (typically a PDE)

$$\dot{y} = F(t, y, u) ,$$

2. the unknown y is the state of the system, which is a function depending on time: $t \in [0, T] \mapsto y(t) \in \mathcal{Y}$, where the set $\mathcal{Y}$ is a functional space (for instance a Banach or a Hilbert space), or a part of a functional space,
3. a parameter u called the control, which is a time-dependent function $t \in [0, T] \mapsto u(t) \in \mathcal{U}$, where the set $\mathcal{U}$ of admissible controls is again some part of a functional space.

As a general rule, one expects that for any initial data $y_{|t=0}$ and any appropriate control function u there exists a unique solution of the system (at least locally in time). In some particular cases, one can find problems "of controllability" type for stationary problems (such as elliptic equations), see for instance [56].

Examples

Let us give two classical examples of the situation. These are two types of acting control frequently considered in the literature: in one case, the control acts as a localized source term in the equation, while in the second one, the control acts on a part of the boundary conditions. The ex-

amples below concern the wave and the heat equations with Dirichlet boundary conditions: these classical equations are reversible and irreversible, respectively, which, as we will see, is of high importance when considering controllability problems.

Example 1 Distributed control for the wave/heat equation with Dirichlet boundary conditions. We consider:

- Ω a regular domain in $\mathbb{R}^n$, which is in general required to be bounded,
- ω a nonempty open subdomain in Ω,
- the wave/heat equation is posed in $[0, T] \times \Omega$, with a localized source term in ω:

wave equation

$$\begin{cases} \Box v := \partial^2_{tt} v - \Delta v = \mathbf{1}_\omega u \,, \\ v_{|\partial\Omega} = 0 \,, \end{cases}$$

heat equation

$$\begin{cases} \partial_t v - \Delta v = \mathbf{1}_\omega u \,, \\ v_{|\partial\Omega} = 0 \,. \end{cases}$$

  - In the first case, the state y of the system is given by the couple $(v(t,\cdot), \partial_t v(t,\cdot))$, for instance considered in the space $H^1_0(\Omega) \times L^2(\Omega)$ or in $L^2(\Omega) \times H^{-1}(\Omega)$.
  - In the second case, the state y of the system is given by the function $v(t,\cdot)$, for instance in the space $L^2(\Omega)$.
  - In both cases, the control is the function u, for instance considered in $L^2([0, T]; L^2(\omega))$.

*Example 2 Boundary control of the wave/*heat *equation with Dirichlet boundary conditions.* We consider:

- Ω a regular domain in $\mathbb{R}^n$, typically a bounded one,
- Σ an open nonempty subset of the boundary $\partial\Omega$,
- the heat/wave equation in $[0, T] \times \Omega$, with nonhomogeneous boundary conditions inside Σ:

wave equation

$$\begin{cases} \Box v := \partial^2_{tt} v - \Delta v = 0 \,, \\ v_{|\partial\Omega} = \mathbf{1}_\Sigma u \,, \end{cases}$$

heat equation

$$\begin{cases} \partial_t v - \Delta v = 0, \\ v_{|\partial\Omega} = \mathbf{1}_\Sigma u \,. \end{cases}$$

  The states are the same as in the previous example, but here the control u is imposed on a part of the boundary. One can for instance consider the set of controls as $L^2([0, T]; L^2(\Sigma))$ in the first case, as $C^\infty_0((0, T) \times \Sigma)$ in the second case.

Needless to say, one can consider other boundary conditions than Dirichlet's. Let us emphasize that, while these two types of control are very frequent, these are not by far the only ones: for instance, one can consider the following "affine control": the heat equation with a right hand side $u(t)g(x)$ where the control u depends only on the time, and g is a fixed function:

$$\begin{cases} \partial_t v - \Delta v = u(t)g(x) \,, \\ v_{|\partial\Omega} = 0 \,, \end{cases}$$

see an example of this below. Also, one could for instance consider the "bilinear control," which takes the form:

$$\begin{cases} \partial_t v - \Delta v = g(x)u(t)v \,, \\ v_{|\partial\Omega} = 0 \,. \end{cases}$$

Main Problems

Now let us give some definitions of the typical controllability problems, associated to a control system.

Definition 1 A control system is said to be exactly controllable in time $T > 0$ if and only if, for all y_0 and y_1 in $\mathcal{Y}$, there is some control function $u\colon [0, T] \to \mathcal{U}$ such that the unique solution of the system

$$\begin{cases} \dot{y} = F(t, y, u) \\ y_{|t=0} = y_0 \,, \end{cases} \tag{2}$$

satisfies

$$y_{|t=T} = y_1 \,.$$

Definition 2 We suppose that the space $\mathcal{Y}$ is endowed with a metric d. The control system is said to be approximately controllable in time $T > 0$ if and only if, for all y_0 and y_1 in $\mathcal{Y}$, for any $\varepsilon > 0$, there exists a control function $u\colon [0, T] \to \mathcal{U}$ such that the unique solution of the system (2) satisfies

$$\mathrm{d}(y_{|t=T}, y_1) < \varepsilon \,.$$

Definition 3 We consider a particular element 0 of $\mathcal{Y}$. A control system is said to be zero-controllable in time $T > 0$ if and only if, for all y_0 in $\mathcal{Y}$, there exists a control function $u\colon [0, T] \to \mathcal{U}$ such that the unique solution of the system (2) satisfies

$$y_{|t=T} = 0 \,.$$

Definition 4 A control system is said to be controllable to trajectories in time $T > 0$ if and only if, for all y_0 in $\mathcal{Y}$ and any trajectory $\overline{y}$ of the system (typically but not necessarily satisfying (2) with $u = 0$), there exists a control function $u\colon [0, T] \to \mathcal{U}$ such that the unique solution of the system (2) satisfies

$$y_{|t=T} = \overline{y}(T) \,.$$

Definition 5 All the above properties are said to be fulfilled locally, if they are proved for y_0 sufficiently close to the target y_1 or to the starting point of the trajectory $\overline{y}(0)$; they are said to be fulfilled globally if the property is established without such limitations.

Remarks

We can already make some remarks concerning the problems that we described above.

1. The different problems of controllability should be distinguished from the problems of optimal control, which give another viewpoint on control theory. In general, problems of optimal control look for a control u minimizing some functional

$$J(u, y(u)) ,$$

where $y(u)$ is the trajectory associated to the control u.
2. It is important to notice that the above properties of controllability depend in a crucial way on the choice of the functional spaces $\mathcal{Y}$, $\mathcal{U}$. The approximate controllability in some space may be the exact controllability in another space. In the same way, we did not specify the regularity in time of the control functions in the above definitions: it should be specified for each problem.
3. A very important fact for controllability problems is that when a problem of controllability has a solution, it is almost never unique. For instance, if a time-invariant system is controllable regardless of the time T, it is clear that one can choose u arbitrarily in some interval $[0, T/2]$, and then choose an appropriate control (for instance driving the system to 0) during the interval $[T/2, T]$. In such a way, one has constructed a new control which fulfills the required task. The number of controls that one can construct in this way is clearly infinite. This is of course already true for finite dimensional systems.
4. Formally, the problem of interior control when the control zone ω is the whole domain Ω is not very difficult, since it suffices to consider the trajectory

$$v(t) := v_0 + \frac{t}{T}(v_1 - v_0) ,$$

to compute the left-hand side of the equation with this trajectory, and to choose it as the control. However, by doing so, one obtains in general a control with very low regularity. Note also that, on the contrary, as long as the boundary control problem is concerned, the case when Σ is equal to the whole boundary $\partial\Omega$ is not that simple.
5. Let us also point out a "principle" which shows that interior and boundary control problems are not very different. We give the main ideas.
Suppose for instance that controllability holds for any domain and subdomain Ω and ω. When considering the controllability problem on Ω via boundary control localized on Σ, one may introduce an extension $\tilde{\Omega}$ of Ω, obtained by gluing along Σ an "additional" open set Ω_2, so that

$$\tilde{\Omega} = \Omega \cup \Omega_2, \ \overline{\Omega} \cap \overline{\Omega_2} \subset \Sigma \ \text{ and } \tilde{\Omega} \text{ is regular} .$$

Consider now $\omega \subset \Omega_2$, and obtain a controllability result on $\tilde{\Omega}$ via interior control located in ω (one has, of course, to extend initial and final states from Ω to $\tilde{\Omega}$). Consider y a solution of this problem, driving the system from y_0 to y_1, in the case of exact controllability, for instance. Then one gets a solution of the boundary controllability problem on Ω, by taking the restriction of y on Ω, and by fixing the trace of y on Σ as the corresponding control (in the case of Dirichlet boundary conditions), the normal derivative in the case of Neumann boundary conditions, etc.
Conversely, when one has some boundary controllability result, one can obtain an interior control result in the following way. Consider the problem in Ω with interior control distributed in ω. Solve the boundary control problem in $\Omega \setminus \omega$ via boundary controls in $\partial\omega$. Consider y the solution of this problem. Extend properly the solution y to Ω, and as previously, compute the left hand side for the extension, and consider it as a control (it is automatically distributed in ω).
Of course, in both situations, the regularity of the control that we obtain has to be checked, and this might need a further treatment.
6. Let us also remark that for linear systems, there is no difference between controllability to zero and controllability to trajectories. In that case it is indeed equivalent to bring y_0 to $\overline{y}(T)$ or to bring $y_0 - \overline{y}(0)$ to zero. Note that even for linear systems, on the contrary, approximate controllability and exact controllability differ: an affine subspace of an infinite dimensional space can be dense without filling all the space.

Linear Systems

In this section, we will briefly describe the theory of controllability for linear systems. The main tool here is the duality between controllability and observability of the adjoint system, see in particular the works by Lions, Russell, and Dolecki and Russell [24,50,51,63,65]. This duality

is also of primary importance for finite-dimensional systems.

Let us first describe informally the method of duality for partial differential equations in two cases given in the examples above.

Two Examples

The two examples that we wish to discuss are the boundary controllability of the wave equation and the interior controllability of the heat equation. The complete answers to these problems have been given by Bardos, Lebeau and Rauch [9] for the wave equation (see also Burq and Gérard [14] and Burq [13] for another proof and a generalization), and independently by Lebeau and Robbiano [48] and Fursikov and Imanuvilov, see [35] for the heat equation. The complete proofs of these deep results are clearly out of the reach of this short presentation, but one can explain rather easily with these examples how the corresponding controllability problems can be transformed into some observability problems. These observability problems consist of proving a certain inequality. We refer for instance to Lions [51] or Zuazua [76] for a more complete introduction to these problems.

First, we notice an important difference between the two equations, which will clearly have consequences concerning the controllability problems. It is indeed well-known that, while the wave equation is a reversible equation, the heat equation is on the contrary irreversible and has a strong regularizing effect. From the latter property, one sees that it is not possible to expect an exact controllability result for the heat equation: outside the control zone ω, the state $u(T)$ will be smooth, and in particular one cannot attain an arbitrary state. As a consequence, while it is natural to seek the exact controllability for the wave equation, it will be natural to look either for approximate controllability or controllability to zero as long as the heat equation is concerned.

For both systems we will introduce the adjoint system (typically obtained via integration by parts): it is central in the resolution of the control problems of linear equations. In both cases, the adjoint system is written backward in time form. We consider our two examples separately.

Wave Equation We first consider the case of the wave equation with boundary control on Σ and Dirichlet boundary conditions on the rest of the boundary:

$$\begin{cases} \partial^2_{tt} v - \Delta v = 0\,, \\ v_{|\partial\Omega} = \mathbf{1}_\Sigma u\,, \\ (v, v_t)_{|t=0} = (v_0, v_0')\,. \end{cases}$$

The problem considered is the exact controllability in $L^2(\Omega) \times H^{-1}(\Omega)$ (recall that the state of the system is (v, v_t)), by means of boundary controls in $L^2((0,T)\times\Sigma)$.

For this system, the adjoint system reads:

$$\begin{cases} -\partial^2_{tt}\psi - \Delta\psi = 0\,, \\ \psi_{|\partial\Omega} = 0\,, \\ (\psi, \psi_t)_{|t=T} = (\psi_T, \psi_T')\,. \end{cases} \tag{3}$$

Notice that this adjoint equation is well-posed: here it is trivial since the equation is reversible.

The key argument which connects the controllability problem of the equation with the study of the properties of the adjoint system is the following duality formula. It is formally easily obtained by multiplying the equation with the adjoint state and in integrating by parts. One obtains

$$\left[\int_\Omega \psi(\cdot,x)v_t(\cdot,x)\mathrm{d}x - \int_\Omega \psi_t(\cdot,x)v(\cdot,x)\mathrm{d}x\right]_0^T = -\iint_{(0,T)\times\Sigma} \frac{\partial\psi}{\partial n}\,\mathbf{1}_\Sigma u\,\mathrm{d}t\,\mathrm{d}\sigma\,. \tag{4}$$

In other words, this central formula describes in a simple manner the jump in the evolution of the state of the system in terms of the control, when measured against the dual state.

To make the above computation more rigorous, one can consider for dual state the ψ "classical" solutions in $C^0([0,T]; H_0^1(\Omega)) \cap C^1([0,T]; L^2(\Omega))$ (these solutions are typically obtained by using a diagonalizing basis for the Dirichlet laplacian, or by using evolution semigroup theory), while the solutions of the direct problem for $(v_0, v_1, u) \in L^2(\Omega) \times H^{-1}(\Omega) \times L^2(\Sigma)$ are defined in $C^0([0,T]; L^2(\Omega)) \cap C^1([0,T]; H^{-1}(\Omega))$ via the transposition method; for more details we refer to the book by Lions [51].

Now, due to the linearity of the system and because we consider the problem of exact controllability (hence things will be different for what concerns the heat equation), it is not difficult to see that it is not restrictive to consider the problem of controllability starting from 0 (that is the problem of reaching any y_1 when starting from $y_0 := (v_0, v_0') = 0$). Denote indeed $R(T, y_0)$ the affine subspace made of states that can be reached from y_0 at time T for some control. Then calling $y(T)$ the final state of the system for $y_{|t=0}$ and $u = 0$, then one has $R(T, y_0) = y(T) + R(T, 0)$. Hence $R(T, y_0) = \mathcal{Y} \Leftrightarrow R(T, 0) = \mathcal{Y}$.

From (4), we see that reaching (v_T, v'_T) from $(0,0)$ will be achieved if and only if the relation

$$\int_\Omega \psi(T,x)v'_T \mathrm{d}x - \int_\Omega \psi'(T,x)v_T \mathrm{d}x = -\iint_{(0,T)\times\Sigma} \frac{\partial\psi}{\partial n}\mathbf{1}_\Sigma u \mathrm{d}t\,\mathrm{d}\sigma \quad (5)$$

is satisfied for all choices of (ψ_T, ψ'_T).

On the left-hand side, we have a linear form on (ψ_T, ψ'_T) in $H_0^1(\Omega)\times L^2(\Omega)$, while on the right hand side, we have a bilinear form on $((\psi_T,\psi'_T), u)$. Suppose that we make the *choice* of looking for a control in the form

$$u = \frac{\partial\overline{\psi}}{\partial n}\mathbf{1}_\Sigma\,, \quad (6)$$

for some $\overline{\psi}$ solution of (3).

Then one sees, using Riesz' theorem, that to solve this problem for $(v_T, v'_T) \in L^2(\Omega)\times H^{-1}(\Omega)$, it is sufficient to prove that the map $(\psi_T,\psi'_T)\mapsto \|\partial\psi/(\partial n)\mathbf{1}_\Sigma\|_{L^2((0,T)\times\Sigma)}$ is a norm equivalent to the $H_0^1(\Omega)\times L^2(\Omega)$ one: for some $C>0$,

$$\|(\psi_T,\psi'_T)\|_{H_0^1(\Omega)\times L^2(\Omega)} \le C\|\frac{\partial\psi}{\partial n}\mathbf{1}_\Sigma\|_{L^2((0,T)\times\Sigma)}. \quad (7)$$

This is the observability inequality to be proved to prove the controllability of the wave equation. Let us mention that the inequality in the other sense, that is, the fact that the linear map $(\psi_T,\psi'_T)\mapsto \partial\psi/(\partial n)\mathbf{1}_\Sigma$ is well-defined and continuous from $H_0^1(\Omega)\times L^2(\Omega)$ to $L^2(\partial\Omega)$ is true but not trivial: it is a "hidden" regularity result, see [66].

When this inequality is proved, a constructive way to select the control is to determine a minimum $(\overline{\psi}_T, \overline{\psi}'_T)$ of the functional

$$(\psi_T,\psi'_T)\mapsto J(\psi_T,\psi'_T) := \frac{1}{2}\iint_\Sigma \left|\frac{\partial\psi}{\partial n}\right|^2 \mathrm{d}t\,\mathrm{d}\sigma + \langle\phi_T, v_1\rangle_{H_0^1\times H^{-1}} - \langle\psi'_T, v_0\rangle_{L^2\times L^2}\,, \quad (8)$$

then to associate to $(\overline{\psi}_T, \overline{\psi}'_T)$ the solution $\overline{\phi}$ of (5), and finally to set u as in (6).

The way described above to determine a particular control—it is clear that not all controls are in the form (6)—is referred as Lions's HUM method (see [51]). This particular control can be proved to be optimal in the L^2 norm, that is, any other control answering to the controllability problem has a larger norm in $L^2((0,T)\times\Sigma)$. As a matter of fact, looking for the L^2 optimal control among those which answer to the problem is a way to justify the choice (6), see [51].

Heat Equation Now let us consider the heat equation with Dirichlet boundary conditions and localized distributed control:

$$\begin{cases} \partial_t v - \Delta v = \mathbf{1}_\omega u\,, \\ v_{|\partial\Omega} = 0\,, \\ v_{|t=0} = v_0\,. \end{cases}$$

In this case, we consider in the same way the dual problem:

$$\begin{cases} -\partial_t\phi - \Delta\phi = 0\,, \\ \phi_{|\partial\Omega} = 0\,, \\ \phi(T) = \phi_T\,. \end{cases} \quad (9)$$

Notice that this adjoint equation is well-posed. Here it is very important that the problem is formulated in a backward way: the backward in time setting compensates the opposite sign before the time derivative. It is clear that the above equation is ill-posed when considering initial data at $t=0$.

In the same way as for the wave equation, multiplying the equation by the adjoint state and integrating by parts yields, at least formally:

$$\int_\Omega \phi_T v_{|t=T}\mathrm{d}x - \int_\Omega \phi(0)v_0\mathrm{d}x = \iint_{(0,T)\times\omega}\phi u\,\mathrm{d}t\,\mathrm{d}x. \quad (10)$$

Note that standard methods yield regular solutions for both direct and adjoint equations when v_0 and ϕ_T belong to $L^2(\Omega)$, and u belongs to $L^2((0,T)\times\omega)$.

Now let us discuss the approximate controllability and the controllability to zero problems separately.

Approximate controllability. Because of linearity, the approximate controllability is equivalent to the approximate controllability starting from 0. Now the set $R(0,T)$ of all final states $v(T)$ which can be reached from 0, via controls $u\in L^2((0,T)\times\omega)$, is a vector subspace of $L^2(\Omega)$. The density of $R(0,T)$ in $L^2(\Omega)$ amounts to the existence of a nontrivial element in $(R(0,T))^\perp$. Considering ϕ_T such an element, and introducing it in (10) together with $v_0=0$, we see that this involves the existence of a nontrivial solution of (9), satisfying

$$\phi_{|(0,T)\times\omega} = 0\,.$$

Hence to prove the approximate controllability, we have to prove that there is no such nontrivial solution, that is, we have to establish a unique continuation result. In this case, this can be proved by using Holmgren's unique continuation principle (see for instance [40]), which establishes the result. Note in passing that Holmgren's theorem is a very general and important tool to prove unique continuation

results; however it requires the analyticity of the coefficients of the operator, and in many situations one cannot use it directly.

Let us also mention that as for the exact controllability of the wave equation above, and the zero-controllability for the heat equation below, one can single out a control for the approximate controllability by using a variational approach consisting of minimizing some functional as in (8): see [26,53].

Controllability to zero. Now considering the problem of controllability to zero, we see that, in order that the control u brings the system to 0, it is necessary and sufficient that for all choice of $\phi_T \in L^2(\Omega)$, we have

$$-\int_\Omega \phi(0)v_0 \mathrm{d}x = \iint_{(0,T)\times\omega} \phi u \mathrm{d}t\mathrm{d}x \,.$$

Here one would like to reason as for the wave equation, that is, make the *choice* to look for u in the form

$$u := \phi\, \mathbf{1}_\omega \,,$$

for some solution ϕ of the adjoint system. But here the application

$$N\colon \phi_T \in L^2(\Omega) \mapsto \left(\iint_{(0,T)\times\omega} \phi^2 \mathrm{d}t\mathrm{d}x\right)^{\frac{1}{2}} .$$

determines a norm (as seen from the above unique continuation result), but this norm is no longer equivalent to the usual $L^2(\Omega)$ norm (if it was, one could establish an exact controllability result!). A way to overcome the problem is to introduce the Hilbert space X obtained by completing $L^2(\Omega)$ for the norm N. In this way, we see that to solve the zero-controllability problem, it is sufficient to prove that the linear mapping $\phi_T \mapsto \phi(0)$ is continuous with respect to the norm N: for some $C > 0$,

$$\|\phi(0)\|_{L^2(\Omega)} \le C\|\phi\, \mathbf{1}_\omega\|_{L^2((0,T)\times\omega)} \,. \tag{11}$$

This is precisely the observability inequality which one has to prove to establish the zero controllability of the heat equation. It is weaker than the observability inequality when the left-hand side is $\|\phi(T)\|_{L^2(\Omega)}$ (which as we noticed is false).

When this inequality is proven, a constructive way to determine a suitable control is to determine a minimum $\overline{\phi}_T$ in X of the functional

$$\phi_T \mapsto \frac{1}{2}\iint_{(0,T)\times\omega} |\phi|^2 \mathrm{d}t\mathrm{d}x + \int_\Omega \phi(0)v_0 \mathrm{d}x \,,$$

then to associate to $\overline{\phi}_T$ the solution $\overline{\phi}$ of (3), and finally to set

$$u := \overline{\phi}\, \mathbf{1}_\omega \,.$$

(That the mapping $\overline{\phi}_T \mapsto \overline{\phi}\, \mathbf{1}_\omega$ can be extended to a mapping from X to $L^2(\omega)$ comes from the definition of X.)

So in both situations of exact controllability and controllability to zero, one has to establish a certain inequality in order to get the result; and to prove approximate controllability, one has to establish a unique continuation result. This turns out to be very general, as described in Paragraph Subsect. "Abstract Approach".

Remarks Let us mention that concerning the heat equation, the zero-controllability holds for any time $T > 0$ and for any nontrivial control zone ω, as shown by means of a spectral method—see Lebeau and Robbiano [48], or by using a global Carleman estimate in order to prove the observability property—see Fursikov and Imanuvilov [35]. That the zero-controllability property does not require the time T or the control zone ω to be large enough is natural since parabolic equations have an infinite speed of propagation; hence one should not require much time for the information to propagate from the control zone to the whole domain. The zero controllability of the heat equation can be extended to a very wide class of parabolic equations; for such results global Carleman estimates play a central role, see for instance the reference book by Fursikov and Imanuvilov [35] and the review article by Fernández-Cara and Guerrero [28]. Note also that Carleman estimates are not used only in the context of parabolic equations: see for instance [35] and Zhang [70]. Let us also mention two other approaches for the controllability of the one-dimensional heat equation: the method of moments of Fattorini and Russell (see [27] and a brief description below), and the method of Laroche, Martin and Rouchon [45] to get an explicit approximate control, based on the idea of "flatness" (as introduced by Fliess, Lévine, Rouchon and Martin [33]).

On the other hand, the controllability for the wave equation does not hold for any T or any Σ. Roughly speaking, the result of Bardos, Lebeau and Rauch [9] states that the controllability property holds if and only if every ray of the geometric optics in the domain (reflecting on the boundary) meets the control zone during the time interval $[0, T]$. In this case, it is also natural that the time should be large enough, because of the finite speed of propagation of the equation: one has to wait for the information coming from the control zone to influence the whole domain. Let us emphasize that in some cases, the geometry

of the domain and the control zone is such that the controllability property does not hold, no matter how long the time T is. An example of this is for instance a circle in which some antipodal regions both belong to the uncontrolled part of the boundary. This is due to the existence of Gaussian beams, that is, solutions that are concentrated along some ray of the geometrical optics (and decay exponentially away from it); when this ray does not meet the control zone, this contradicts in particular (7). The result of [9] relies on microlocal analysis. Note that another important tool used for proving observability inequalities is a multiplier method (see in particular [42,51,60]): however in the case of the wave equation, this can be done only in particular geometric situations.

Abstract Approach

The duality between the controllability of a system and the observability of its adjoint system turns out to be very general, and can be described in an abstract form due to Dolecki and Russell [24]. For more recent references see [46,69]. Here we will only describe a particular simplified form of the general setting of [24]. Consider the following system

$$\dot{y} = Ay + Bu \,, \tag{12}$$

where the state y belongs to a certain Hilbert space H, on which A, which is densely defined and closed, generates a strongly continuous semi-group of bounded operators. The operator B belongs to $\mathcal{L}(U;H)$, where U is also a Hilbert space. The solutions of (12) are given by the method of variation of constants so that

$$y(T) = \mathrm{e}^{tA} y(0) + \int_0^T \mathrm{e}^{(T-\tau)A} Bu(\tau)\mathrm{d}\tau \,.$$

We naturally associate with the control equation (12), the following observation system:

$$\begin{cases} \dot{z} = -A^* z \,, \\ z(T) = z_T \,, \\ c := B^* z \,. \end{cases} \tag{13}$$

The dual operator $-A^*$ generates a strongly continuous semi-group for negative times, and c in $L^2(0,T;U)$ is obtained by

$$c(t) = B^* \mathrm{e}^{(T-t)A^*} z_T \,.$$

In the above system, the dynamics of the (adjoint) state z is free, and c is called the observed quantity of the system. The core of the method is to connect the controllability properties of (12) with observability properties of (13) such as described below.

Definition 6 The system (13) is said to satisfy the unique continuation property if and only if the following implication holds true:

$$c = 0 \text{ in } [0,T] \Longrightarrow z = 0 \text{ in } [0,T] \,.$$

Definition 7 The system (13) is said to be observable if and only if there exists $C > 0$ such that the following inequality is valid for all solutions z of (13):

$$\|z(T)\|_H \le C\|c\|_{L^2(0,T;U)} \,. \tag{14}$$

Definition 8 The system (13) is said to be observable at time 0 if and only if there exists $C > 0$ such that the following inequality is valid for all solutions z of (13):

$$\|z(0)\|_H \le C\|c\|_{L^2(0,T;U)} \,. \tag{15}$$

The main property is the following.

Duality property

1. The exact controllability of (12) is equivalent to the observability of (13).
2. The zero controllability of (12) is equivalent to the observability at time 0 of (13).
3. The approximate controllability of (12) in H is equivalent to the unique continuation property for (13).

Brief explanation of the duality property. The main fact is the following duality formula

$$\begin{aligned} &\langle y(T), z(T)\rangle_H - \langle y(0), z(0)\rangle_H \\ &\qquad = \int_0^T \left\langle u(\tau), B^* \mathrm{e}^{(T-\tau)A^*} z(T)\right\rangle_U \mathrm{d}\tau \,, \end{aligned} \tag{16}$$

which in fact can be used to define the solutions of (12) by transposition.

1 and 3. Now it is rather clear that by linearity, we can reduce the problems of exact and approximate controllability to the ones when $y(0) = 0$. Now the property of exact controllability for system (12) is equivalent to the surjectivity of the operator

$$S\colon u \in L^2(0,T;U) \mapsto \int_0^T \mathrm{e}^{(T-\tau)A} Bu(\tau)\,\mathrm{d}\tau \in H \,,$$

and the approximate controllability is equivalent to the density of its range. By (16) its adjoint operator is

$$S^*\colon z_T \in H \mapsto \int_0^T B^* \mathrm{e}^{(T-\tau)A^*} z_T\,\mathrm{d}\tau \in L^2(0,T;U) \,.$$

Hence the equivalences 1 and 3 are derived from a classical result from functional analysis (see for instance the book of Brezis [12]): the range of S is dense if and only if S^* is one-to-one; S is surjective if and only if S^* satisfies for some $C > 0$:

$$\|z_T\| \leq C\|S^* z_T\| ,$$

that is, when (14) is valid.

2. In this case, still due to linearity, the zero-controllability for system (12) is equivalent to the following inclusion:

$$\text{Range}\,(\mathrm{e}^{TA}) \subset \text{Range}\,(S) .$$

In this case there is also a functional analysis result (generalizing the one cited above) which asserts the equivalence of this property with the existence of $C > 0$ such that

$$\|(\mathrm{e}^{TA})^* h\| \leq C\|S^* h\| ,$$

see Dolecki and Russell [24] and Douglas [25] for more general results.

It follows that in many situations, the exact controllability of a system is proved by establishing an observability inequality on the adjoint system. But this general method is not the final point of the theory: not only is this only valid for linear systems, but importantly as well, it turns out that these types of inequalities are in general very difficult to establish. In general when a result of exact controllability is established by using this duality, the largest part of the proof is devoted to establishing the observability inequality.

Another information given by the observability is an estimate of the size of the control. Indeed, following the lines of the proof of the correspondence between observability and controllability, one can see that one can find a control u which satisfies:

- Case of exact controllability:

$$\|u\|_{L^2(0,T;\mathcal{U})} \leq C_{\text{obs}}\|y_1 - \mathrm{e}^{TA} y_0\|_H ,$$

- Case of zero controllability:

$$\|u\|_{L^2(0,T;\mathcal{U})} \leq C_{\text{obs}}\|y_0\|_H ,$$

where in both cases C_{obs} determines a constant for which the corresponding observability inequality is true. (Obviously, not *all* the controls answering to the question satisfy this estimate). This gives an upper bound on the size of a possible control. This can give some precise estimates on the cost of the control in terms of the parameters of the problem, see for instance the works of Fernández-Cara and Zuazua [30] and Miller [58] and Seidman [66] concerning the heat equation.

Some Different Methods

Herein we will discuss two methods that do not rely on controllability/observability duality. This does not pretend to give a general vision of all the techniques that can be used in problems of controllability of linear partial differential equations. We just mention them in order to show that in some situations, duality may not be the only tool available.

Characteristics First, let us mention that in certain situations, one may use a characteristics method (see e. g. [22]). A very simple example is the first-dimensional wave equation,

$$\begin{cases} v_{tt} - v_{xx} = 0 , \\ v_{x|x=0} = 0 , \\ v_{x|x=1} = u(t) , \end{cases}$$

where the control is $u(t) \in L^2(0, T)$. This is taken from Russell [62]. The problem is transformed into

$$\frac{\partial w}{\partial t} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \frac{\partial w}{\partial x} \quad \text{with}$$

$$w := \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} := \begin{pmatrix} v_t \\ v_x \end{pmatrix} ,$$

$$w_2(t, 0) = 0 \quad \text{and} \quad w_2(t, 0) = u(t) .$$

Of course, this means that $w_1 - w_2$ and $w_1 + w_2$ are constant along characteristics, which are straight lines of slope 1 and -1, respectively (this is d'Alembert's decomposition).

Now one can deduce an explicit appropriate control for the controllability problem in $[0, 1]$ for $T > 2$, by constructing the solution w of the problem directly (and one takes the values of w_2 at $x = 1$ as the "resulting" control u).

The function w is completely determined from the initial and final values in the domains of determinacy D_1 and D_2:

$$D_1 := \{(t, x) \in [0, T] \times [0, 1] \,/\, x + t \leq 1\} \quad \text{and}$$
$$D_2 := \{(t, x) \in [0, T] \times [0, 1] \,/\, t - x \geq T - 1\} .$$

That $T > 2$ means that these two domains do not intersect. Now it suffices to complete the solution w in $D_3 := [0, T] \times [0, 1] \setminus (D_1 \cup D_2)$. For that, one chooses $w_1(0, t)$ *arbitrarily* in the part ℓ of the axis $\{0\} \times [0, 1]$ located between D_1 and D_2, that is for

$$x = 0 \quad \text{and} \quad 1 \leq t \leq T - 1 .$$

Once this choice is made, it is not difficult to solve the Goursat problem consisting of extending w in D_3: using

the symmetric role of x and t, one considers x as the time. Then the initial condition is prescribed on ℓ, as well as the boundary conditions on the two characteristic lines $x + t = 1$ and $x - t = T - 1$. One can solve elementarily this problem by using the characteristics, and this finishes the construction of w, and hence of u.

Note that as a matter of fact, the observability inequality in this (one-dimensional) case is also elementary to establish, by relying on Fourier series or on d'Alembert's decomposition, see for instance [23].

The method of characteristics described above can be generalized in broader situations, including for instance the problem of boundary controllability of one-dimensional linear hyperbolic systems

$$v_t + A(x)v_x + B(x)v = 0 ,$$

where A is a real symmetric matrix with eigenvalues bounded away from zero, and A and B are smooth; see for instance Russell [65]. As a matter of fact, in some cases this method can be used to establish the observability inequality from the controllability result (while in most cases the other implication in the equivalence is used).

Of course, the method of characteristics may be very useful for transport equations

$$\frac{\partial f}{\partial t} + v.\nabla f = g \quad \text{or} \quad \frac{\partial f}{\partial t} + \operatorname{div}(vf) = g .$$

An example of this is the controllability of the Vlasov–Poisson equation, see [36]. Let us finally mention that this method is found also to tackle directly several nonlinear problems, as we will see later.

Moments Another method which we would like to briefly discuss is the method of moments (see for instance Avdonin and Ivanov [4] for a general reference, see also Russell [65]), which can appear in many situations; in particular this method was used by Fattorini and Russell to prove the controllability of the heat equation in one space dimension, see [27]. Consider for instance the problem of controllability of the one-dimensional heat equation

$$\begin{cases} v_t - v_{xx} = g(x)u(t) , \\ v_{|[0,T]\times\{0,1\}} = 0 . \end{cases}$$

Actually in [27], much more general situations are considered: in particular it concerns more general parabolic equations and boundary conditions, and boundary controls can also be included in the discussion.

It is elementary to develop the solution in the $L^2(0,1)$ orthonormal basis $(\sin(k\pi x))_{k\in\mathbb{N}^*}$. One obtains that the state zero is reached at time $T > 0$ if and only if

$$\sum_{k>0} \mathrm{e}^{-k^2T} v_k \sin(k\pi x) = -\sum_{k>0} \int_0^T \mathrm{e}^{-k^2(T-t)} g_k u(t) \sin(k\pi x)\mathrm{d}t ,$$

where v_k and g_k are the coordinates of v_0 and g in the basis. Clearly, this means that we have to find u such that for all $k \in \mathbb{N}$,

$$\int_0^T \mathrm{e}^{-k^2(T-t)} g_k u(t)\mathrm{d}t = -\mathrm{e}^{-k^2T} v_k .$$

The classical Muntz–Szász theorem states that for an increasing family $(\lambda_n)_{n\in\mathbb{N}^*}$ of positive numbers, the family $\{\mathrm{e}^{-\lambda_n t},\ n \in \mathbb{N}^*\}$ is dense in $L^2(0,T)$ if and only if

$$\sum_{n\geq 0} \frac{1}{\lambda_n} = +\infty ,$$

and in the opposite case, the family is independent and spans a proper closed subspace of $L^2(0,T)$. Here the exponential family which we consider is $\lambda_n = n^2$ and we are in the second situation. The same method applies for other problems in which λ_n cannot be completely computed but is a perturbation of n^2; this allows us to treat a wider class of problems. Now in this situation (see e.g. [65]), one can construct in $L^2(0,T)$ a biorthogonal family to $\{\mathrm{e}^{-\lambda_n t},\ n \in \mathbb{N}^*\}$, that is a family $(p_n(t))_{n\in\mathbb{N}^*} \in (L^2(0,T))^{\mathbb{N}}$ satisfying

$$\int_0^T p_n(t)\mathrm{e}^{-\lambda_k^2 t}\mathrm{d}t = \delta_{kn} .$$

Once such a family is obtained, one has formally the following solution, under the natural assumption that $|g_k| \geq ck^{-\alpha}$:

$$u(t) = \sum_{k\in\mathbb{N}^*} -\frac{\mathrm{e}^{-\lambda_k T} v_k}{g_k} p_k(t) .$$

To actually get a control in $L^2(0,T)$, one has to estimate $\|p_k\|_{L^2}$, in order that the above sum is well-defined in L^2. In [27] it is proven that one can construct p_k in such a way that

$$\|p_k\|_{L^2(0,T)} \leq K_0 \exp(K_1\omega_k) , \quad \omega_k := \sqrt{\lambda_k} ,$$

which allows us to conclude.

Nonlinear Systems

The most frequent method for dealing with control problems of nonlinear systems is the natural one (as for the Cauchy problem): one has to linearize (at least in some sense) the equation, try to prove some controllability result on the linear equation (for instance by using the duality principle), and then try to pass to the nonlinear system via typical methods such as inverse mapping theorem, fixed point theory, iterative schemes…

As for the usual inverse mapping theorem, it is natural to hope for a local result from the controllability of the linear problem. Here we find an important difference between linear and nonlinear systems: while for linear systems, no distinction has to be made between local and global results, concerning the nonlinear systems, the two problems are really of different nature.

One should probably not expect a very general result indicating that the controllability of the linearized equation involves the local controllability of the nonlinear system. The linearization principle is a general approach which one has to adapt to the different situations that one can meet. We give below some examples that present to the reader some existing approaches.

Some Linearization Situations

Let us first discuss some typical situations where the linearized equation has good controllability properties, and one can hope to get a result (in general local) from this information. In some situations where the nonlinearity is not too strong, one can hope to get global results. As previously, the situation where the underlying linear equation is reversible and the situation when it is not have to be distinguished. We briefly describe this in two different examples.

Semilinear Wave Equation Let us discuss first an example with the wave equation. This is borrowed from the work of Zuazua [71,72], where the equation considered is

$$\begin{cases} v_{tt} - v_{xx} + f(v) = u\,\mathbf{1}_\omega \text{ in } [0,1]\,, \\ v_{|x=0} = 0,\ v_{|x=1} = 0\,, \end{cases}$$

where $\omega := (l_1, l_2)$ is the control zone and the nonlinearity $f \in C^1(\mathbb{R};\mathbb{R})$ is at most linear at infinity (see however the remark below) in the sense that for some $C > 0$,

$$|f(x)| \le C(1 + |x|) \text{ on } \mathbb{R}\,. \tag{17}$$

The global exact controllability in $H_0^1 \times L^2$ (recall that the state of the system is (v, v_t)) by means of a control in $L^2((0,T)\times\omega)$ is proved by using the following linearization technique. In [72], it is proven that the following linearized equation:

$$\begin{cases} v_{tt} - v_{xx} + a(x)v = u\,\mathbf{1}_\omega \text{ in } [0,1]\,, \\ v_{|x=0} = 0,\ v_{|x=1} = 0\,. \end{cases}$$

is controllable in $H_0^1(0,1) \times L^2(0,1)$ through $u(t) \in L^2((0,T)\times\omega)$, for times $T > 2\max(l_1, 1 - l_2)$, for any $a \in L^\infty((0,T)\times(0,1))$. The corresponding observability inequality is

$$\|(\psi_T, \psi_T')\|_{L^2(\Omega)\times H^{-1}(\Omega)} \le C_{\text{obs}} \|\psi\,\mathbf{1}_\omega\|_{L^2((0,T)\times\omega)}\,.$$

Moreover, the observability constant that one can obtain can be bounded in the following way:

$$C_{\text{obs}} \le \alpha(T, \|a\|_\infty)\,, \tag{18}$$

where α is nondecreasing in the second variable. One would like to describe a fixed-point scheme as follows. Given v, one considers the linearized problem around v, solves this problem and deduces a solution $\hat{v}$ of the problem of controllability from (v_0, v_0') to (v_1, v_1') in time T. More precisely, the scheme is constructed in the following way.

We write

$$f(x) = f(0) + xg(x)\,,$$

where g is continuous and bounded. The idea is to associate to any $v \in L^\infty((0,T)\times(0,1))$ a solution of the linear control problem

$$\begin{cases} \hat{v}_{tt} - \hat{v}_{xx} + \hat{v}g(v) = -f(0) + u\,\mathbf{1}_\omega \text{ in } [0,1]\,, \\ \hat{v}_{|x=0} = 0,\ \hat{v}_{|x=1} = 0\,. \end{cases}$$

(Note that the "drift" term $f(0)$ on the right hand side can be integrated in the final state—just consider the solution of the above system with $u = 0$ and $\hat{v}_{|t=0} = 0$ and withdraw it—or integrated in the formulation (5)).

Here an issue is that, as we recalled earlier, a controllability problem has almost never a unique solution. The method in [72] consists of selecting a particular control, which is the one of smallest L^2-norm. Taking the fact that g is bounded and (18) into account, this particular control satisfies

$$\begin{aligned} &\|u\|_{L^2(0,T)} \\ &\quad \le C(\|v_0\|_{H_0^1} + \|v_0'\|_{L^2} + \|v_1\|_{H_0^1} + \|v_1'\|_{L^2} + |f(0)|)\,, \end{aligned}$$

for some $C > 0$ independent of v.

Using the above information, one can deduce estimates for $\hat{v}$ in $C^0([0,T];L^2(0,1)) \cap L^2(0,T;H_0^1)$ independently of v, and then show by Schauder's fixed point theorem that the above process has a fixed point, which shows a global controllability result for this semilinear wave equation.

Remark 1 As a matter of fact, [72] proves the global exact controllability for f satisfying the weaker assumption

$$\lim_{|x|\to+\infty} \frac{|f(x)|}{(1+|x|)\log^2(|x|)} = 0 .$$

This is optimal since it is also proven in [72] that if

$$\liminf_{|x|\to+\infty} \frac{|f(x)|}{(1+|x|)\log^p(|x|)} > 0$$

for some $p > 2$ (and $\omega \neq (0,1)$) then the system is not globally controllable due to blow-up phenomena. To obtain this result one has to use Leray–Schauder's degree theory instead of Schauder's fixed point theorem. For analogous conditions on the semilinear heat equation, see Fernández-Cara and Zuazua [31].

Burgers Equation Now let us discuss a parabolic example, namely the local controllability to trajectories of the viscous Burgers equation:

$$\begin{cases} v_t + (v^2)_x - v_{xx} = 0 \text{ in } (0,T)\times(0,1) , \\ v_{|x=0} = u_0(t) \text{ and } v_{|x=1} = u_1(t) , \end{cases} \tag{19}$$

controlled on the boundary via u_0 and u_1. This is taken from Fursikov and Imanuvilov [34]. Now the linear result concerns the zero-controllability via boundary controls of the system

$$\begin{cases} v_t + (zv)_x - v_{xx} = 0 \text{ in } (0,T)\times(0,1) , \\ v_{|x=0} = u_0(t) \text{ and } v_{|x=1} = u_1(t) . \end{cases} \tag{20}$$

Consider X the Hilbert space composed of functions z in $L^2(0,T;H^2(0,1))$ such that $z_t \in L^2(0,1;L^2(0,1))$. In [34] it is proved that given $v_0 \in H^1(0,1)$ and $T > 0$, one can construct a map which to any z in X, associates $v \in X$ such that v is a solution of (20) which drives v_0 to 0 during time interval $[0,T]$, and moreover this map is compact from X to X. As in the previous situation, the particular control (u_0,u_1) has to be singled out in order for the above mapping to be single-valued (and compact). But here, the criterion is not quite the optimality in L^2-norm of the control. The idea is the following: first, one transforms the controllability problem for (20) from y_0 to 0 into a problem of "driving" 0 to 0 for a problem with right-hand side:

$$\begin{cases} w_t + (zw)_x - w_{xx} = f_0 \text{ in } (0,T)\times(0,1) , \\ w_{|x=0} = u_0(t) \text{ and } w_{|x=1} = u_1(t) , \end{cases} \tag{21}$$

for some f_0 supported in $(T/3, 2T/3)\times(0,1)$. For this, one introduces $\chi \in C^\infty([0,T];\mathbb{R})$ such that $\chi = 1$ during $[0,T/3]$ and $\chi = 0$ during $[2T/3,T]$, and $\hat{v}$ the solution of (20) starting from v_0 with $u_0 = u_1 = 0$, and considers $w := v - \chi\hat{v}$.

Now the operator mentioned above is the one which to z associates the solution of the controllability problem which minimizes the L^2-norm of w among all the solutions of this controllability problem. The optimality criterion yields a certain form for the solution. That the corresponding control exists (and is unique) relies on a Carleman estimate, see [34] (moreover this allows estimates on the size of w). As a matter of fact, to obtain the compactness of the operator, one extends the domain, solves the above problem in this extended domain, and then uses an interior parabolic regularity result to have bounds in smaller spaces, we refer to [34] for more details.

Once the operator is obtained, the local controllability to trajectories is obtained as follows. One considers $\overline{v}$ a trajectory of the system (19), belonging to X. Withdrawing (19) for $\overline{v}$ to (19) for the unknown, we see that the problem to solve through boundary controls is

$$\begin{cases} y_t - y_{xx} + [(2\overline{v} + y)y]_x = f_0 \text{ in } (0,T)\times(0,1) , \\ y_{|x=0} = v_0 - \overline{v}(0) \text{ and } y_{|t=T} = 0 . \end{cases}$$

Now consider $v_0 \in H^1(0,1)$ such that

$$\|v_0 - \overline{v}(0)\|_{H^1(0,1)} < r ,$$

for $r > 0$ to be chosen. To any $y \in X$, one associates the solution of the controllability problem (21) constructed above, driving $v_0 - \hat{v}(0)$ to 0, for $z := (2\hat{v} + y)$. The estimates on the solution of the control problem allow one to establish that the unit ball of X is stable by this process provided r is small enough. The compactness of the process is already proved, so Schauder's theorem allows one to conclude.

Note that it is also proved in [34] that the global approximate controllability does not hold.

Some Other Examples Let us also mention two other approaches that may be useful in this type of situation.

The first is the use of Kakutani–Tikhonov fixed-point theorem for multivalued maps (cf. for instance [68]), see in particular [38,39] and [26]. Such a technique is particularly

useful, because it avoids the selection process of a particular control. One associates to v the set $T(v)$ of all $\hat{v}$ solving the controllability problem for the equation linearized around v, with all possible controls (in a suitable class). Then under appropriate conditions, one can find a fixed point in the sense that $v \in T(v)$.

Another approach that is very promising is the use of a Nash–Moser process, see in particular the work [10] by Beauchard. In this paper the controllability of a Schrödinger equation via a bilinear control is considered. In that case, one can solve some particular linearized equation (as a matter of fact, the return method described in the next paragraph is used), but with a loss of derivative; as a consequence the approaches described above fail, but the use of Nash–Moser's theorem allows one to get a result. Note finally that in certain other functional settings, the controllability of this system fails, as shown by using a general result on bilinear control by Ball, Marsden and Slemrod [5].

The Return Method

It occurs that in some situations, the linearized equation is not systematically controllable, and one cannot hope by applying directly the above process to get even local exact controllability. The return method was introduced by Coron to deal with such situations (see in particular [18]). As a matter of fact, this method can be useful even when the linearized equation is controllable. The principle of the method is the following: find a particular trajectory $\overline{y}$ of the nonlinear system, starting at some base point (typically 0) and returning to it, such that the linearized equation around this is controllable. In that case, one can hope to find a solution of the nonlinear local controllability problem close to $\overline{y}$.

A typical situation of this is the two-dimensional Euler equation for incompressible inviscid fluids (see Coron [16]), which reads

$$\begin{cases} \partial_t y + (y.\nabla)y = -\nabla p \ \text{ in } \ \Omega\,, \\ \operatorname{div} y = 0 \ \text{ in } \ \Omega\,, \\ y.n = 0 \ \text{ on } \ \partial\Omega \setminus \Sigma\,, \end{cases} \tag{22}$$

where the unknown is the velocity field $y\colon \Omega \to \mathbb{R}^2$ (the pressure p can be eliminated from the equation), Ω is a regular bounded domain (simply connected to simplify), n is the unit outward normal on the boundary, and $\Sigma \subset \partial\Omega$ is the control zone. On Σ, the natural control which can be assigned is the normal velocity $y.n$ and the vorticity $\omega := \operatorname{curl} y := \partial_1 y^2 - \partial_2 y^1$ at "entering" points, that is points where $y.n < 0$.

Let us discuss this example in an informal way. As noticed by J.-L. Lions, the linearized equation around the null state

$$\begin{cases} \partial_t y = -\nabla p \ \text{ in } \ \Omega\,, \\ \operatorname{div} y = 0 \ \text{ in } \ \Omega\,, \\ y.n = 0 \ \text{ on } \ \partial\Omega \setminus \Sigma\,, \end{cases}$$

is trivially not controllable (even approximately). Now the main goal is to find the trajectory $\overline{y}$ such that the linearized equation near $\overline{y}$ is controllable. It will be easier to work with the vorticity formulation

$$\begin{cases} \partial_t \omega + (y.\nabla)\omega = 0 \ \text{ in } \ \Omega\,, \\ \operatorname{curl} y = \omega,\ \operatorname{div} y = 0\,. \end{cases} \tag{23}$$

In fact, one can show that assigning $y(T)$ is equivalent to assigning both $\omega(T)$ in Ω and $y(T).n$ on Σ, and since this latter is a part of the control, it is sufficient to know how to assign the vorticity of the final state.

We can linearize (23) in the following way: to y one associates $\hat{y}$ through

$$\begin{cases} \partial_t \omega + (y.\nabla)\omega = 0 \ \text{ in } \ \Omega\,, \\ \operatorname{curl} \hat{y} = \omega,\ \operatorname{div} \hat{y} = 0\,. \end{cases} \tag{24}$$

Considering Eq. (24), we see that if the flow of y is such that any point in $\overline{\Omega}$ at time T "comes from" the control zone Σ, then one can assign easily ω through a method of characteristics. Hence one has to find a solution $\overline{y}$ of the system, starting and ending at 0, and such that in its flow, all points in $\overline{\Omega}$ at time T, come from Σ at some stage between times 0 and T. Then a simple Gronwall-type argument shows that this property holds for y in a neighborhood of $\overline{y}$, hence Eq. (24) is controllable in a neighborhood of $\overline{y}$. Then a fixed-point scheme allows one to prove a controllability result locally around 0.

But the Euler equation has some time-scale invariance:

$$\begin{aligned} y(t,x) \ \text{ is a solution on } \ [0,T] &\Rightarrow y^\lambda(t,x) \\ &:= \lambda^{-1} y(\lambda^{-1} t, x) \ \text{ is a solution on } \ [0, \lambda T]\,. \end{aligned}$$

Hence given y_0 and y_1, one can solve the problem of driving λy_0 to λy_1 for λ small enough. Changing the variables, one see that it is possible to drive y_0 to y_1 in time λT, that is, in *smaller* times. Hence one deduces a global controllability result from the above local result.

As a consequence, the central part of the proof is to find the function $\overline{y}$. This is done by considering a special type of solution of the Euler equation, namely the potential solution: any $\overline{y} := \nabla\theta(t,x)$ with θ regular satisfying

$$\begin{aligned} \Delta_x \theta(t,x) = 0 \ \text{ in } \ \Omega, \quad \forall t \in [0,T], \\ \partial_n \theta = 0 \ \text{ on } \ \partial\Omega, \quad \forall t \in [0,T]\,, \end{aligned}$$

satisfies (22). In [16] it is proven that there exists some θ satisfying the above equation and whose flow makes all points at time T come from Σ. This concludes the argument.

This method has been used in various situations; see [18] and references therein. Let us underline that this method can be of great interest even in the cases where the linearized equation is actually controllable. An important example obtained by Coron concerns the Navier–Stokes equation and is given in [17] (see also [19]): here the return method is used to prove some global approximate result, while the linearized equation is actually controllable but yields in general a local controllability result (see in particular [29,35,41]).

Some Other Methods

Let us finally briefly mention that linearizing the equation (whether using the standard approach or the return method) is not systematically the only approach to the controllability of a nonlinear system. Sometimes, one can "work at the nonlinear level." An important example is the control of one-dimensional hyperbolic systems:

$$v_t + A(v)v_x = F(v),\ v\colon [0,T]\times[0,1]\to\mathbb{R}^n\,, \qquad (25)$$

via the boundary controls. In (25), A satisfies the hyperbolicity property that it possesses at every point n real eigenvalues; these are moreover supposed to be strictly separated from 0.

In the case of regular C^1 solutions, this was approached by a method of characteristics to give general local results, see in particular the works by Cirinà [15] and Li and Rao [49]. Interestingly enough, the linear tool of duality between observability and controllability has some counterpart in this particular nonlinear setting, see Li [55]. In the context of weak entropy solutions, some results for particular systems have been obtained via the ad hoc method of front-tracking, see in particular Ancona, Bressan and Coclite [2], the author [37] and references therein.

Other nonlinear tools can be found in Coron's book [18]. Let us mention two of them. The first one is power series expansion. It consists of considering, instead of the linearization of the system, the development to higher order of the nonlinearity. In such a way, one can hope to attain the directions which are unreachable for the linearized system. This has for instance applications for the Korteweg–de-Vries equation, see Coron and Crépeau [20] and the earlier work by Rosier [61]. The other one is quasi-static deformations. The general idea is to find an (explicit) "almost trajectory" $(\overline{y}(\varepsilon t), \overline{u}(\varepsilon t))$ during $[0, T/\varepsilon]$ of the control system $\dot{y} = f(y,u)$, in the sense that

$$\frac{\mathrm{d}}{\mathrm{d}t}\big[\overline{y}(\varepsilon t)\big] - f(\overline{y}(\varepsilon t), \overline{u}(\varepsilon t))$$

is of order ε (due to the "slow" motion). Typically, the trajectory $(\overline{y}, \overline{u})$ is composed of equilibrium states (that is $f(\overline{y}(\cdot), \overline{u}(\cdot)) = 0$). In such a way, one can hope to connect $\overline{y}(0)$ to a state close to $\overline{y}(T)$, and then to exactly $\overline{y}(T)$ via a local result. This was used for instance by Coron and Trélat in the context of a semilinear heat equation, see [21].

Finally, let us cite a recent approach by Agrachev and Sarychev [1] (see also [67]), which uses a generalization of the "Lie bracket" approach (a standard approach from finite-dimensional nonlinear systems), to obtain some global approximate results on the Navier–Stokes equation with a finite-dimensional (low modes) affine control.

Some Other Problems

As we mentioned earlier, many aspects of the theory have not been referred to herein. Let us cite some of them. Concerning the connection between the problem of controllability and the problem of stabilization, we refer to Lions [50], Russell [63,64,65] and Lasiecka and Triggiani [47].

A very important problem which has recently attracted great interest is the problem of numerics and discretization of distributed systems (see in particular [74,75] and references therein). The main difficulty here comes from the fact that the operations of discretizing the equation and controlling it do not commute.

Other important questions considered in particular in the second volume of Lions's book [51] are the problems of singular perturbations, homogenization, thin domains in the context controllability problems. Here the questions are the following: considering a "perturbed" system, for which we have some controllability result, how does the solution of the controllability problem (for instance associated to the L^2-optimal control) behave as the system converges to its limit? This kind of question is the source of numerous new problems.

Another subject is the controllability of equations with some stochastic terms (see for instance [8]). One can also consider systems with memory (see again [51] or [7]). Finally, let us mention that many partial differential equations widely studied from the point of view of Cauchy theory, still have not been studied from the point of view of controllability.

The reader looking for more discussion on the subject can consider the references below.

Future Directions

There are many future challenging problems for control theory. Controllability is one of the possible approaches to try to construct strategies for managing complex systems. On the road towards systems with increasing complexity, many additional difficulties have to be considered in the design of the control law: one should expect the control to take into account the possible errors of modelization, of measurement of the state, of the control device, etc. All of these should be included in the model so some robustness of the control can be expected, and to make it closer to the real world. Of course, numerics should play an important role in the direction of applications. Moreover, one should expect more and more complex systems, such as environmental or biological systems (how are regulation mechanisms designed in a natural organism?), to be approached from this point of view. We refer for instance to [32,59] for some discussions on some perspectives of the theory.

Bibliography

1. Agrachev A, Sarychev A (2005) Navier–Stokes equations: control lability by means of low modes forcing. J Math Fluid Mech 7(1):108–152
2. Ancona F, Bressan A, Coclite GM (2003) Some results on the boundary control of systems of conservation laws. Hyperbolic problems: theory, numerics, applications. Springer, Berlin, pp 255–264
3. Ancona F, Marson A (1998) On the attainable set for scalar nonlinear conservation laws with boundary control. SIAM J Control Optim 36(1):290–312
4. Avdonin SA, Ivanov SA (1995) Families of exponentials. The method of moments in controllability problems for distributed parameter systems. Cambridge Univ Press, Cambridge
5. Ball JM, Marsden JE, Slemrod M (1982) Controllability for distributed bilinear systems. SIAM J Control Optim 20:575–597
6. Bandrauk AD, Delfour MC, Le Bris C (eds) (2003) Quantum control: mathematical and numerical challenges. Papers from the CRM Workshop held at the Université de Montréal, Montréal, 6–11 October 2002. CRM Proceedings and Lecture Notes, vol 33. American Mathematical Society, Providence
7. Barbu V, Iannelli M (2000) Controllability of the heat equation with memory. Differ Integral Equ 13(10–12):1393–1412
8. Barbu V, Răşcanu A, Tessitore G (2003) Null controllability of stochastic heat equations with a multiplicative noise. Appl Math Optim 47(2):97–120
9. Barbos C, Lebeau G, Rauch J (1992) Sharp sufficient conditions for the observation, control and stabilisation of waves from the boundary. SIAM J Control Optim 305:1024–1065
10. Beauchard K (2005) Local controllability of a 1D Schrödinger equation. J Math Pures Appl 84:851–956
11. Bensoussan A, Da Prato G, Delfour MC, Mitter SK (1993) Representation and control of infinite-dimensional systems, vol I, II. Systems and Control: Foundations and Applications. Birkhäuser, Boston
12. Brezis H (1983) Analyse fonctionnelle, Théorie et applications; Collection Mathématiques Appliquées pour la Maîtrise. Masson, Paris
13. Burq N (1997) Contrôle de l'équation des ondes dans des ouverts peu réguliers. Asymptot Anal 14:157–191
14. Burq N, Gérard P (1997) Condition nécessaire et suffisante pour la contrôlabilité exacte des ondes. C R Acad Sci Paris Sér. I Math 325(7):749–752
15. Cirinà M (1969) Boundary controllability of nonlinear hyperbolic systems. SIAM J Control 7:198–212
16. Coron JM (1996) On the controllability of 2-D incompressible perfect fluids. J Math Pures Appl 75:155–188
17. Coron JM (1996) On the controllability of the 2-D incompressible Navier–Stokes equations with the Navier slip boundary conditions. ESAIM Control Optim Calc Var 1:35–75
18. Coron JM (2007) Control and Nonlinearity. Mathematical Surveys and Monographs, vol 136. American Mathematical Society, Providence
19. Coron JM, Fursikov AV (1996) Global exact controllability of the 2-D Navier–Stokes equations on a manifold without boundary. Russian J Math Phys 4:1–19
20. Coron JM, Crépeau E (2004) Exact boundary controllability of a nonlinear KdV equation with critical lengths. JEMS J Eur Math Soc 6(3):367–398
21. Coron JM, Trélat E (2004) Global steady-state controllability of one-dimensional semilinear heat equations. SIAM J Control Optim 43(2):549–569
22. Courant R, Hilbert D (1989) Methods of mathematical physics, vol II. Partial differential equations, Wiley Classics Library. Wiley, New York
23. Dàger R, Zuazua E (2006) Wave propagation, observation and control in 1-d flexible multi-structures. Mathématiques and Applications, vol 50. Springer, Berlin
24. Dolecki S, Russell DL (1977) A general theory of observation and control. SIAM J Control Optim 15(2):185–220
25. Douglas RG (1966) On majorization, factorization, and range inclusion of operators on Hilbert space. Proc Amer Math Soc 17:413–415
26. Fabre C, Puel JP, Zuazua E (1995) Approximate controllability of the semilinear heat equation. Proc Roy Soc Edinburgh Sect A 125(1):31–61
27. Fattorini HO, Russell DL (1971) Exact controllability theorems for linear parabolic equation in one space dimension. Arch Rat Mech Anal 43:272–292
28. Fernández-Cara E, Guerrero S (2006) Global Carleman inequalities for parabolic systems and applications to controllability. SIAM J Control Optim 45(4):1399–1446
29. Fernández-Cara E, Guerrero S, Imanuvilov OY, Puel JP (2004) Local exact controllability to the trajectories of the Navier–Stokes equations. J Math Pures Appl 83:1501–1542
30. Fernández-Cara E, Zuazua E (2000) The cost of approximate controllability for heat equations: The linear case. Adv Differ Equ 5:465–514
31. Fernández-Cara E, Zuazua E (2000) Null and approximate controllability for weakly blowing up semilinear heat equations. Ann Inst H Poincaré Anal Non Linéaire 17(5):583–616
32. Fernández-Cara E, Zuazua E (2003) Control Theory: History, mathematical achievements and perspectives. Boletin SEMA 26:79–140
33. Fliess M, Lévine J, Martin P, Rouchon P (1995) Flatness and de-

fect of non-linear systems: introductory theory and examples. Int J Control 61(6):1327–1361
34. Fursikov A, Imanuvilov OY (1995) On controllability of certain systems simulating a fluid flow, Flow control, Minneapolis, 1992. IMA Math Appl, vol 68. Springer, New York, pp 149–184
35. Fursikov A, Imanuvilov OY (1996) Controllability of evolution equations. Lecture Notes Series, vol 34. Seoul National University Research Institute of Mathematics Global Analysis Research Center, Seoul
36. Glass O (2003) On the controllability of the Vlasov-Poisson system. J Differ Equ 195(2):332–379
37. Glass O (2007) On the controllability of the 1-D isentropic Euler equation. J Eur Math Soc 9:427–486
38. Henry J (1978) Controllability of some non linear parabolic equations. In: Ruberti A (ed) Distributed Parameter Systems: Modelling and Identification. Proceedings of the IFIP working conference, Rome, 21–24 June 1976. Lecture Notes in Control and Information Sciences, vol 1. Springer, Berlin
39. Henry J (1977) Étude de la contrôlabilité de certaines équations paraboliques non linéaires. Thèse, Paris VI
40. Hörmander L (1983) The analysis of linear partial differential operators, vol I. Grundlehren der mathematischen Wissenschaften, vol 256. Springer, Berlin
41. Imanuvilov OY (2001) Remarks on exact controllability for the Navier–Stokes equations. ESAIM Control Optim Calc Var 6:39–72
42. Komornik V (1994) Exact controllability and stabilization, The multiplier method. RAM: Research in Applied Mathematics. Masson, Paris
43. Komornik V, Loreti P (2005) Fourier series in control theory. Springer Monographs in Mathematics. Springer, New York
44. Lagnese JE, Leugering G, Schmidt EJPG (1994) Modeling, analysis and control of dynamic elastic multi-link structures. Systems and Control: Foundations and Applications. Birkhäuser, Boston
45. Laroche B, Martin P, Rouchon P (2000) Motion planning for the heat equation. Int J Robust Nonlinear Control 10(8):629–643
46. Lasiecka I, Triggiani R (1983) Regularity of hyperbolic equations under $L^2(0, T; L^2(\Gamma))$-Dirichlet boundary terms. Appl Math Optim 10(3):275–286
47. Lasiecka I, Triggiani R (2000) Control theory for partial differential equations: continuous and approximation theories, vol I. Abstract parabolic systems and II, Abstract hyperbolic-like systems over a finite time horizon. Encyclopedia of Mathematics and its Applications, vols 74, 75. Cambridge University Press, Cambridge
48. Lebeau G, Robbiano L (1995) Contrôle exact de l'équation de la chaleur. Comm PDE 20:335–356
49. Li TT, Rao BP (2003) Exact boundary controllability for quasilinear hyperbolic systems. SIAM J Control Optim 41(6):1748–1755
50. Lions JL (1988) Exact controllability, stabilizability and perturbations for distributed systems. SIAM Rev 30:1–68
51. Lions JL (1988) Contrôlabilité exacte, stabilisation et perturbations de systèmes distribués, Tomes 1, 2. RMA 8, 9. Masson, Paris
52. Lions JL (1990) Are there connections between turbulence and controllability? In: Bensoussan A, Lions JL (eds) Analysis and Optimization of Systems. Lecture Notes Control and Inform Sci, vol 144. Springer, Berlin
53. Lions JL (1992) Remarks on approximate controllability. J Analyse Math 59:103–116
54. Le Bris C (2000) Control theory applied to quantum chemistry: some tracks, Contrôle des, systèmes gouvernés par des équations aux dérivées partielles, Nancy, 1999. ESAIM Proc 8:77–94
55. Li TT (2008) Exact boundary observability for quasilinear hyperbolic systems. ESAIM Control Optim Calc Var 14:759–766
56. Lions JL (1971) Optimal control of systems governed by partial differential equations. Grundlehren der mathematischen Wissenschaften, vol 170. Springer, Berlin
57. López A, Zuazua E (2002) Uniform null controllability for the one dimensional heat equation with rapidly oscillating periodic density. Ann IHP Analyse linéaire 19(5):543–580
58. Miller L (2006) On exponential observability estimates for the heat semigroup with explicit rates. Atti Accad Naz Lincei Cl Sci Fis Mat Natur Rend Lincei, vol 9. Mat Appl 17(4):351–366
59. Murray RM (ed) (2003) Control in an information rich world. Report of the Panel on Future Directions in Control, Dynamics, and Systems. Papers from the meeting held in College Park, 16–17 July 2000. Society for Industrial and Applied Mathematics, Philadelphia
60. Osses A (2001) A rotated multiplier applied to the controllability of waves, elasticity, and tangential Stokes control. SIAM J Control Optim 40(3):777–800
61. Rosier L (1997) Exact boundary controllability for the Korteweg-de Vries equation on a bounded domain. ESAIM Control Optim Calc Var 2:33–55
62. Russell DL (1967) On boundary-value controllability of linear symmetric hyperbolic systems. Mathematical Theory of Control. Proc Conf Los Angeles, pp 312–321
63. Russell DL (1973) A unified boundary controllability theory for hyperbolic and parabolic partial differential equations. Int Math Res Notices 52:189–221
64. Russell DL (1974) Exact boundary value controllability theorems for wave and heat processes in star-complemented regions. Differential games and control theory. Proc NSFCBMS Regional Res Conf Univ. Rhode Island, Kingston, 1973. Lecture Notes in Pure Appl Math, vol 10. Dekker, New York, pp 291–319
65. Russell DL (1978) Controllability and stabilizability theory for linear partial differential equations. Recent progress and open questions. SIAM Rev 20:639–739
66. Seidman TI (1988) How violent are fast controls? Math Control Signal Syst 1(1):89–95
67. Shirikyan A (2006) Approximate control lability of three-dimensional Navier–Stokes equations. Comm Math Phys 266(1):123–151
68. Smart DR (1974) Fixed point theorems, Cambridge Tracts in Mathematics, No 66. Cambridge University Press, New York
69. Tucsnak M, Weiss G (2009) Observation and Control for Operator Semigroups. Birkhäuser Advanced Texts. Birkhäuser, Basel
70. Zhang X (2001) Explicit observability inequalities for the wave equation with lower order terms by means of Carleman inequalities. SIAM J Cont Optim 39:812–834
71. Zuazua E (1990) Exact controllability for the semilinear wave equation. J Math Pures Appl 69(1):1–31
72. Zuazua E (1993) Exact controllability for semilinear wave equations in one space dimension. Ann Inst H Poincaré Anal Non Linéaire 10(1):109–129
73. Zuazua E (1998) Some problems and results on the controllability of Partial Differential Equations. Proceedings of the Second European Conference of Mathematics, Budapest, July

1996. Progress in Mathematics, vol 169. Birkhäuser, Basel, pp 276–311
74. Zuazua E (2002) Controllability of Partial Differential Equations and its Semi-Discrete Approximations. Discret Continuous Dyn Syst 8(2):469–513
75. Zuazua E (2005) Propagation, observation, and control of waves approximated by finite difference methods. SIAM Rev 47(2):197–243
76. Zuazua E (2006) Controllability and observability of partial differential equations: some results and open problems. In: Dafermos C, Feiresl E (eds) Handbook of differential equations: evolutionary differential equations, vol 3. Elsevier/North-Holland, Amsterdam

Information Theoretic Complexity Measures

Danail G. Bonchev
Virginia Commonwealth University, Richmond, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Information Content of Atoms
Information Content of Atomic Nuclei
Information Content of Molecules
Information Content of Networks
Future Directions
Bibliography

Glossary

Adjacency matrix A square $V \times V$ matrix, where V is the number of vertices of graph G. The matrix entries are $a_{ij} = 1$ for i, j – neighbors, and $a_{ij} = 0$, otherwise. In **undirected graphs**, the matrix is symmetric with respect to the main diagonal.

Branch A linear (path) **subgraph**, beginning with a vertex of degree higher than 2, and ending with a vertex of degree 1.

Branching The (topological) property of a graph to have branches, as well the patterns of branch connectivity.

Centrality The (topological) property of vertex (or edge) organization in a graph with respect to central vertex(es) or edge(s).

Complexity measures Quantitative assessments of systems complexity, obtained mainly by applying information theory and graph theory.

Connected graph A graph is connected, if there is a **walk** between any pair of vertices.

Cycle A **path**, the initial and end vertices of which coincide.

Cyclicity The (topological) property of a graph to have cycles, as well as the patterns of cycle connectivity.

Distance The number of edges connecting two vertices along the shortest path between them.

Eccentricity – see **Vertex eccentricity**.

Graph A mathematical structure composed of points (vertices), connected by lines (edges).

Graph center Vertex(es) in a graph with the smallest **eccentricity**.

Graph distance The sum of all **vertex-vertex distances** in the graph. Also: The sum of all **vertex distances** in the graph.

Information content A quantitative characteristic of a system calculated by using information theory. Three major types of information content have been of use: compositional, structural (mainly topological), and functional ones.

Information theory Created initially as theory of communications by Shannon in 1949, it extends the statistical thermodynamics of Boltzmann to processes involving information. Other versions of this theory have been proposed; the best known one being the Kolmogorov nonprobabilistic theory of information. One of the main applications of information theory is to define **information content** of systems.

Information theoretical descriptor (information index) A number, which is a quantitative characteristic of a system, calculated by using **information theory**.

Graph path A sequence of edges connecting two graph vertices without a repeated visit of a vertex or edge. The length of the path is equal to the number of path edges.

Molecular graph A graph, the vertices of which stand for atoms of a molecule, and edges represent chemical bonds between the atoms.

Network An interconnected system of elements of any physical nature, and the graph representing the system. The terminology used in network theory is similar to that of graph theory, vertices and edges being called *nodes* and *links*, respectively.

Subgraph A graph $G' = (V', E')$ is called a subgraph of graph $G = (V, E)$ if V' belongs to the set of vertices V and E' belongs to the edges of E.

Total adjacency of a graph The sum of all adjacency matrix entries. Alternatively, the sum of all **vertex degrees**. In undirected graphs, the total adjacency is equal to the doubled number of edges.

Undirected graph A graph in which the binary adjacency relation $a_{ij} = a_{ji} = 1$ exists for all adjacent vertices i, j. If in a graph there is at least one adjacency rela-

tion $a_{ij} = 1$, the symmetric relationship a_{ji} for which does not exist, the graph is **directed (a Digraph)**.

Vertex degree The number of the nearest neighbors of a vertex. Also, the number of edges incident to the vertex.

Vertex distance The sum of the distances from a given vertex to all other vertices in the graph.

Vertex eccentricity The largest distance from a vertex to any other vertex in the graph.

Walk A walk in a simple graph (having no multiple edges or loops) is a sequence of consecutive vertices and edges with repetitions allowed. The length of a walk is the number of edges (including repetitions).

Weighted distribution The ordered set of all values of given quantity.

Definition of the Subject

Complexity is a multifaceted concept, related to the degree of organization of systems. Patterns of complex organization and behavior are identified in all kinds of systems in nature and technology. Essential for the characterization of complexity is its quantification, the introduction of complexity measures or descriptors, following Lord Kelvin's words that science begins when we can use numbers. Historically, the first attempt to quantify complexity was based on Shannon's information theory [1], and it involved the information content as a measure of molecular complexity [2]. Fifty years later, the complexity of molecules and their interactions is assessed by a variety of methods, with information theory preserving its leading role. This article aims to review the vast area of complexity measures, based on information theory as applied to chemical and biochemical systems. Many of these measures have found application for predicting physicochemical properties and biological activities of chemical compounds, contributing thus to the development of new drugs and chemical products. The expertise accumulated has recently found a new vast area of application, the networks of biomolecules performing the basic functions of life in cells and organisms. The essence of life itself has been reformulated to incorporate as an essential component the processing of information.

Introduction

The notion of information has been first used in scientific literature in 1894 by Boltzmann, who stated that "Every piece of information obtained for a physical system is related to the decrease in the number of its possible states; therefore the increase of entropy means 'loss of information'" [3]. One may speculate that only the early death of Boltzmann, who was far ahead of his contemporaries, prevented the development of information theory at the very beginning of 20th century. Indeed, Boltzmann's statistical thermodynamics was a direct predecessor of Shannon's information theory. The entropy S of a system of N particles, distributed over t states, having $N_1, N_2, \ldots, N_t$ particles with energies $E_1, E_2, \ldots, E_t$, respectively, was related by Boltzmann to the total number W of physical states of the system, where k is a constant:

$$S = k \ln W = k \frac{N!}{N_1! N_2! \ldots N_t!} \,. \tag{1}$$

For $N \gg 1$, the approximation of Stirling turns Eq. (1) into

$$S \approx k \left(N \ln N - \sum_{i=1}^{t} N_i \ln N_i \right) \,. \tag{2}$$

Equation (2) after an elementary substitution of the constant k is identical to the Shannon equation for the entropy of information H of a message transmitted through information channels:

$$H = N \log_2 N - \sum_{i=1}^{t} N_i \log_2 N_i \,, \quad \text{bits} \,. \tag{3}$$

Shannon's information theory regards such a message as a specific set of symbols (an "outcome") selected from an ensemble of all t such sets containing the same total number of symbols N. Probabilities $p_1, p_2, \ldots, p_t$ are assigned to each outcome, the probability of the ith outcome being proportional to the number of symbols N_i it contains: $p_i = N_i/N$. Shannon's entropy of information H characterizes the uncertainty of the expected outcome. Upon a totally random transmission all outcomes are equiprobable and the entropy of information is maximal. Conversely, in case of a single outcome, $H = 0$. In the intermediate cases, the amount of information transmitted is the difference between the maximum entropy and the specific value the Shannon H-function has for the system of interest. Thus, information emerges as a measure for the eliminated outcome uncertainty.

Another more popular form of the Shannon equation defines the average entropy of information $\bar{\mathrm{H}}$ per communication symbol:

$$\begin{aligned} \bar{\mathrm{H}} &= -\sum_{i=1}^{t} p_i \log_2 p_i \\ &= -\sum_{i=1}^{t} \frac{N_i}{N} \log_2 \frac{N_i}{N} \,, \quad \text{bits/symbol} \,. \end{aligned} \tag{4}$$

One bit of information is obtained when learning the outcome of a process eliminating the uncertainty of a choice between two equally probable options. The values of the two Shannon's entropies introduced by Eqs. (3) and (4) vary within the following ranges in bits:

$$0 \leqslant H \leqslant N \,, \tag{5a}$$

$$0 \leqslant \tilde{H} \leqslant 1 \,. \tag{5b}$$

Shannon's theory soon exceeded the narrow limits of a communication theory, and was considered as an extension of Boltzmann's theory. This view was advocated by Brillouin [4], who's negentropy principle of information views information as the negative component of entropy. The second law of thermodynamics thus generalized allows only such spontaneous processes in closed systems that increase entropy and lose information.

Another important extension was the view on *structure* as a message that carries certain amount of information. That was how the notion of *information content* of a (chemical) structure emerged in the early 1950s [5,6,7,8,9]. A radical reinterpretation of the meaning of Shannon's equations (3) and (4) was proposed by Mowshowitz [10] in 1968. When applied to molecules, the H-function does not measure the average uncertainty for selecting a molecule from the ensemble of all molecules having the same number of atoms. Rather, it is the information content of the structure relative to a system of symmetry transformations that leaves the structure invariant. Bonchev [11] supported this interpretation by the argument that entropy is transformed into information by the mere process of the structure formation from its constituent elements. This structural information is conserved until the structure is destroyed, when it turns back into entropy.

Mowshowitz [10] presented his approach as a *finite probabilistic scheme*, applicable to any system having symmetry elements. The system of N interacting elements is treated as a graph, the vertices of which are partitioned into k equivalence classes, according to symmetry operations (graph automorphisms), which exchange vertices while preserving the graph adjacency.

Equivalence classes	$1, 2, \ldots, k$
Partition of elements	$N_1, N_2, \ldots, N_k$
Probability distribution	$p_1, p_2, \ldots, p_k$.

Here, $p_i = N_i/N$ is the probability of a randomly chosen element to belong to class i, which has N_i elements, and $N = \sum N_i$. The analysis of Eqs. (3) and (4), expressing now the information content I of the system, shows that information has the maximum value when each element is in a separate class, i. e., when the system has no symmetry. The information content is zero when all elements belong to a single equivalence class, i. e., when the system has no structure, due to its high symmetry. One might infer that the total and average information content, I and $\bar{I}$, could be used as complexity measures, which relate high complexity to low symmetry and larger diversity of system's elements. Low complexity (simplicity) is characterized by uniformity, resulting from high symmetry and lack of diversity.

In what follows till the end of this article, it will be shown how this information-theoretic formalism can be used to characterize the structure of atoms, molecules, and (molecular) networks. It will be demonstrated that symmetry-based information-theoretic descriptors cannot always be good complexity measures, because symmetry is a simplifying factor. A better approach will be introduced, proceeding from a weighted version of the original Mowshowitz' scheme.

Information Content of Atoms

Information Content of Chemical Elements and Nuclides

One may define the information content of atoms and their nuclei in a variety of ways proceeding from the total number of protons z, neutrons n, and electrons $e = z$ (in neutral atoms), and their different distributions [12]. A good starting point is considering the atomic structure partitioned into two substructures – a nucleus and an electron shell. The *total information content of a nuclide* is thus defined [13] as:

$$I_{\text{nuclide}} = (A+z)\log_2(A+z) - A\log_2 A - z\log_2 z \,, \tag{6}$$

where the mass number $A = z + n$. In atomic mass units, A (an integer) is approximately equal to the atom's mass (a real number).

The information content of a chemical element can then be defined [13] as the average information content of all naturally occurring nuclides of this element:

$$I_{\text{chem.element}} = \sum\nolimits_i c_i I_{\text{nuclide,i}} \,, \tag{7}$$

c_i being the abundance fraction of nuclide i.

The next step is to analyze the distribution of electrons of the atoms of chemical elements into electron shells (I_n), subshells (I_{nl}), atomic orbitals (I_{nlm}), and spin-orbitals (I_{nlmms}) [11,14,15]. In all these cases, z electrons are considered distributed into k equivalence classes having

$N_1, N_2, \ldots, N_k$ electrons, respectively. Denoting the combination of quantum numbers that defines the specific type of information content by x, one finds the corresponding total and average information content, $I(x)$ and $\bar{I}(x)$:

$$I(x) = z \log_2 z - \sum_i z_i \log_2 z_i \,, \tag{8}$$

and

$$\bar{I}(x) = -\sum_i \frac{z_i}{z} \log_2 \frac{z_i}{z} \,. \tag{9}$$

A third information function, termed *differential information content*, $\Delta I_x(z)$, is defined by the difference between the information content of the chemical element with atomic number z and combination of quantum numbers x, and those of the element with atomic number $z-1$:

$$\Delta I_x(z) = I_x(z) - I_x(z-1) \,. \tag{10}$$

The differential information content was shown to be a sensitive descriptor for the periodicity of chemical elements. As shown in Fig. 1, $\Delta I_x(z)$ has a sharp maximum in the first element of each period, or *s*-, *p*-, *d*-, and *f*-subperiod, followed by a gradual decrease to a minimum in the corresponding last element. This regular trend is best demonstrated by periods II and III. In periods IV to VII, the filling of the *d*- and *f*-subshells with delay produces the inequality:

$$\Delta I_n((n-2)f) < \Delta I_n((n-1)d) < \Delta I_n(np) < \Delta I_n(ns) \,. \tag{11}$$

The violations in the "ideal order" of filling electron *f*- and *d*-subshells, caused by the accelerated adding of $(n-1)d$- or $(n-2)f$-electrons at the cost of the decreased population of the *ns*-subshells, are also captured in the information function by sharp minima as seen in Fig. 1 for the extra *d*-electron in Cr, Cu, Nb, Pd, etc.

Information Equations for Periods and Groups of Chemical Elements

Such equations are derived proceeding from several atomic information descriptors [16,17]. Equation (12) is based on the information for electron distribution over an *nl*-subshell. It contains the period constant P, which is equal for periods I through VII to 0, 2, 19.51, 37.02, 87.75, 138.48, and 242.51 bits, and a group constants $k_n l = 1$ or 2 for groups 1 and 2, $k_n l = 1$ to 10 for groups 3 to 12, and $k_n l = 1$ to 6 for groups 13 to 18, respectively:

$$I_{nl} = z \log_2 z - P - \sum_l k_{nl} \log_2 k_{nl} \,. \tag{12}$$

Another equation for the periods and groups in the periodic table is obtained with the total information on electron distribution over atomic orbitals:

$$I_{\mathrm{AO}} = (z_0 + a + b) \log_2 (z_0 + a + b) - z_0 - b \,. \tag{13}$$

Here, b is the number of paired electrons in the atomic orbitals of the chemical element. In the ground state of the atoms of elements of groups 1, 2, and 13 through 18, the group constant a is equal to the lowest valence of the element (1, 0, 1, 2, 3, 2, 1, 0, respectively), whereas in the highest valence state it equals the old numbering of the eight main groups (1 to 8). The period constant z_0 is the atomic number of the noble gas that ends the preceding period ($z_0 = 2, 10, 18, 36, 54, 86$).

The information on electron distribution over the values of the magnetic spin quantum number ($m_s = +1/2$

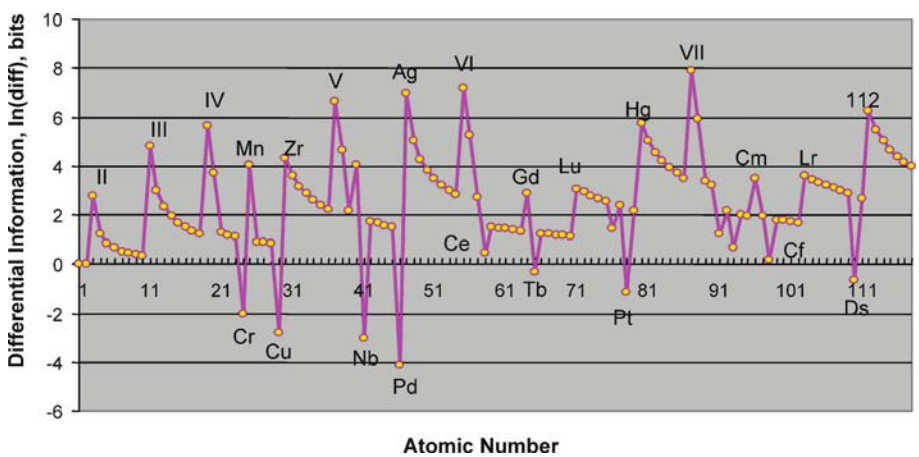

Information Theoretic Complexity Measures, Figure 1
Differential information on the distribution of atomic electrons over electron shells

and $m_s = -1/2$) provides yet another period/group equation for chemical elements:

$$\bar{I}_{ms} = -\frac{z+a}{2z}\log_2\frac{z+a}{2z} - \frac{z-a}{2z}\log_2\frac{z-a}{2z} \approx 1 - \frac{k}{(z_0+a+b)^2}\,. \quad (14)$$

Here, b and the period constant z_0 are those from Eq. (13). The group constant $k = a^2/(2\ln 2)$ includes the number of unpaired electrons a. The error introduced by the approximation used decreases with the fourth power of z and is very small.

The Pauli Exclusion Principle and Hund's First Rule Maximize the Atomic Information Content

The information equations for electron distributions in atoms provided the basis for a reinterpretation of the physical principles and rules controlling the building of the atomic electronic structure [18]. Hund's first rule, which requires maximum filling of atomic orbitals in s-, p-, d-, and f-subshells with unpaired electrons, may be interpreted as a rule demanding maximum information on atomic orbitals, I_{nlm}. This follows directly from our Eq. (13), which maximizes when the number of paired electrons is $b = 0$. The absolute maximum of the atomic information content according to Eq. (8) is reached when all $z_i = 1$. This case corresponds to electron distribution over spin–orbitals, defined by the four quantum numbers n, l, m, and m_s, and required by the Pauli exclusion principle. The Hund rule and the Pauli exclusion principle, thus, emerge as different manifestations of a more general trend requiring the maximum information content of atoms, thus characterizing atoms as structures of maximum complexity. Proceeding from quantum mechanical and group theory analysis, this trend was shown to extend to any fermionic system.

Atomic Information Descriptors as Tools for Calculating and Predicting the Properties of Atoms and Chemical Elements

Encoding a detailed description of the electronic structure of atoms, the information theoretic indices were shown to be an effective tool in quantifying the periodicity in the properties of chemical elements. They provide very high correlation with almost all atomic characteristics and physico-chemical properties of chemical elements. As an example, 21 of 23 examined properties of alkali metals have been found to correlate with the atomic information descriptors with an average correlation coefficient of 0.997, a degree of correlation far better than that with atomic number. The models derived have been applied to the prediction of a number of atomic characteristics and properties of the transactinide elements 113–120 [12,19].

Information Content of Atomic Nuclei

Information on Proton–Neutron Composition

Atomic nuclei have been characterized by the distribution of protons z and neutrons n in shells and subshells in a manner similar to the one described for electron distribution in Sect. "Information Content of Atoms". The information index on the proton–neutron composition of atomic nuclei, I_{pn}, has been found to exhibit interesting properties [20]:

$$I_{\mathrm{pn}} = A\log_2 A - z\log_2 z - n\log_2 n\,, \quad \text{bits}\,, \quad (15)$$

$$\bar{I}_{\mathrm{pn}} = -\frac{z}{A}\log_2\frac{z}{A} - \frac{n}{A}\log_2\frac{n}{A}\,. \quad (16)$$

These equations identify the symmetric nuclei having equal number of protons and neutrons, $p = z = n$, as nuclei with the highest complexity, due to their maximum information on the proton–neutron composition. The total information content of these nuclei is equal to the mass number A, whereas the mean information is exactly 1 bit. Such are the symmetric stable nuclei of the elements at the beginning of the periodic table: ^{2}H, ^{4}He, ^{6}Li, etc., up to ^{40}Ca. With the increase of the atomic number of the element, the filling of the proton energy levels in the nucleus proceeds with delay. The resulting excess of neutrons is termed *isotopic number*: $\beta = n - z = A - 2z$.

The basic patterns in the behavior of the proton–neutron information content are demonstrated in Fig. 2 for all stable nuclei up to ^{204}Hg. The symmetric light nuclei are located on the line of maximum mean information content of exactly 1 bit. With the increase in atomic mass, I_{pn} diverges more and more from its maximum, in correspondence with the delay in filling the proton energy levels. The minima in the figure correspond to the largest excess of neutrons for the series of several isotopes of the same chemical element. Conversely, the maxima after ^{40}Ca refer to the least possible excess of neutrons.

The mean I_{pn} index, defined by Eq. (16) is approximated with a sufficient accuracy (the relative error does not exceed 0.02% with exception of ^{3}He with 0.15%) by

$$\bar{I}_{\mathrm{pn}} \approx 1 - \frac{1}{2\ln 2}\times\frac{\beta^2}{A^2} = 1 - \frac{1}{2\ln 2}\times\frac{(A-2z)^2}{A^2}\,. \quad (17)$$

The information index on the proton–neutron composition of atomic nuclei is thus shown to increase with the increase in mass number A and atomic number z, and

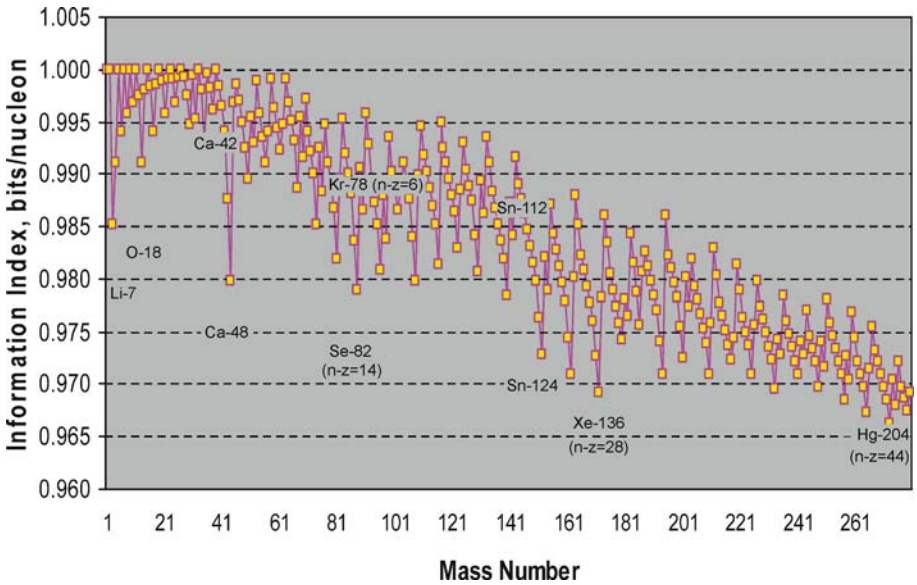

Information Theoretic Complexity Measures, Figure 2

The average information on the proton–neutron composition of the naturally occurring atomic nuclei. The horizontal line of 1 bit information content describes the symmetric light nuclei having the same number of protons and neutrons, whereas the minima and maxima stand for the isotopes of certain elements having the largest and the smallest neutron excess, respectively

to decrease with the increase in the isotopic number and the number of neutrons.

The Concept for "Defect" of Information and the Systematics of Nuclides

The total information on the proton–neutron composition of a nucleus, I_{pn}, expressed in bits according to Eq. (15), is very close to the mass number A. The difference ΔI^*_{pn} between the two quantities has been introduced [20] as "defect" of information by analogy with the defect of mass upon formation of atomic nuclei:

$$\Delta I^*_{\mathrm{pn}} = A - I_{\mathrm{pn}} \approx \frac{1}{2\ln 2} \times \frac{(A-2z)^2}{A} \,. \tag{18}$$

Since this deviation results in a decrease of the binding energy, it has been conjectured that the defect of information can be regarded as a negative component of the binding energy, E_b. In fact, ΔI^*_{pn}coincides (with a transition coefficient $k = 25.1$ MeV/bit) with the parameter of the relative symmetry of the nucleus, δ, in Weizsäcker's equation for nuclear binding energy:

$$\delta = -18.1 \times \frac{(A-2z)^2}{A} \,\mathrm{MeV} \,. \tag{19}$$

The remarkable equivalence of Eqs. (18) and (19) affords evidence for the usefulness of information theory in the study of atomic nuclei. Correlation of I_{pn} and ΔI^*_{pn} with the total nuclear binding energy has also been reported for series of nuclides with a constant isotopic number β.

The concept for a defect of information has also found application in the systematics of nuclides, introducing the new series of isodefective nuclides having the same defect of information per nucleon i [21]:

$$\begin{aligned} i &= \frac{\Delta I_{\mathrm{pn}}*}{A} = \frac{1}{2\ln 2} \times \left(\frac{\beta}{A}\right)^2 \\ &= \frac{1}{2\ln 2} \times \left(1 - 2\frac{z}{A}\right)^2 , \quad \text{bits/nucleon} \,. \end{aligned} \tag{20}$$

It can be seen from Eq. (20) that nuclei having the same relative charge z/A also have the same mean defect of information i. The new class of nuclides ("Isodefectants") thus introduced adds a fifth dimension to the Chart of the Nuclides, traditionally based on the classes of isotopes ($z = \mathrm{const}$), isotones ($n = \mathrm{const}$), isobars ($A = \mathrm{const}$), and isodifferent nuclides ($\beta = n - z = \mathrm{const}$). A portion of the Chart of the Nuclides represented with five classes of nuclides is shown in Fig. 3 with defect of information/mass number coordinates (i/A diagram). The illustration includes nuclides having mass numbers between 100 and 120, and atomic numbers between 39 (Y) and 56 (Ba).

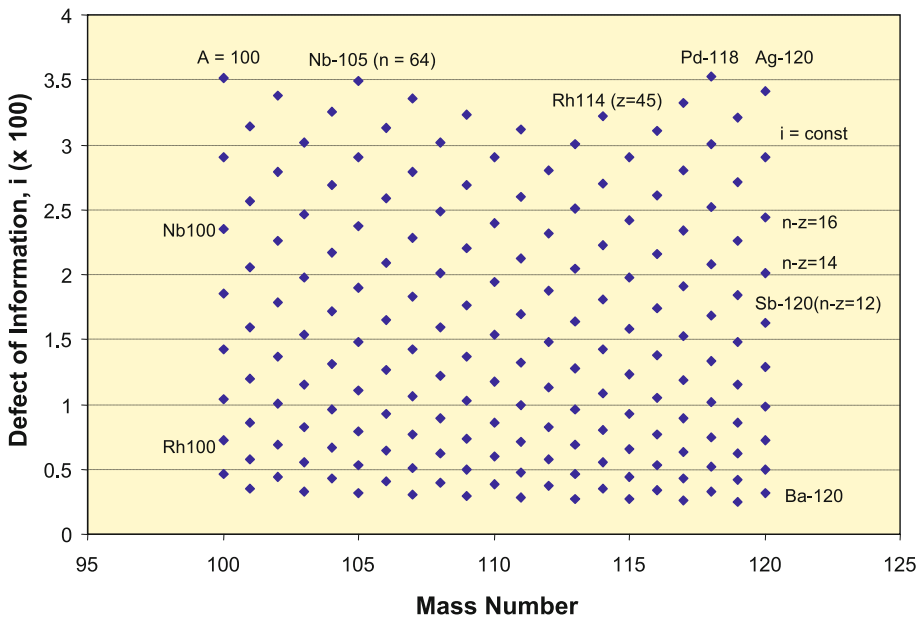

Information Theoretic Complexity Measures, Figure 3
A fragment from the representation of the Chart of the Nuclides with five classes of nuclides: isobars (e. g., Rh-100 to Nb-100), isotopes (e. g., Rh-100 to Rh-116), isotones (e. g., Nb-105 to Ba-120), isodifferents (e. g., Nb-100 to Sb-120), and isodefectants (e. g., the line with $i = 3$)

The behavior of the series of isobars, isotopes, isotones, and isodifferent nuclides in the i/A plot is described by equations for the mean defect of information i:

$$i(\text{isobars}) = c_0(1 - c_1 z)^2 \,, \tag{21a}$$

$$i(\text{isotopes}) = c_0 \left(1 - \frac{c_2}{A}\right)^2 \,, \tag{21b}$$

$$i(\text{isotones}) = c_0 \left(\frac{c_3 - z}{c_3 + z}\right)^2 \,, \tag{21c}$$

$$i(\text{isodifferents}) = c_0 \frac{c_4}{A^2} \,, \tag{21d}$$

where $c_0 = 1/2 \ln 2$, $c_1 = 2/A$, $c_2 = 2z$, $c_3 = n$, $c_4 = \beta = A - 2z$, respectively.

Predictions of Nuclear Binding Energies of Nuclides of Elements #101–108

A considerable similarity has been found in the information equations describing the nuclear and electronic structure of chemical elements [12]. This finding, along with the observed correlations of the information on the proton–neutron composition and the defect of this information, with nuclear binding energy, has prompted the search for direct relationships between nuclear and electronic energies. Equations have been derived that relate fairly well the ground state electronic energy E_e of chemical elements to the nuclear binding energy $E_b(id)$ of isodifferent series of nuclides [12]. Predictions have been made on this basis in the year 1980 for the nuclear binding energy of 45 not synthesized isotopes of elements 101–108 [22]. In 2003, these predictions have been confirmed [12] with relative error of 0.1% for 41 isotopes, which have meanwhile been synthesized. This result has confirmed the usefulness of information content as a measure of structural similarity and complexity on atomic level, and has indicated the potential existence of correlated motions of nucleons and electrons.

This concludes the analysis of information theoretic descriptors of atomic and nuclear complexity. In the remaining sections the focus will be shifted to the description of molecular structures and the biochemical networks they form in living things.

Information Content of Molecules

Early History

Several years after Shannon published his fundamental work on information theory [1] a group of US biologists proposed to define the information content of molecules and use it in assessing the complexity of living cells and organisms [2]. Several alternative definitions have been proposed. Dancoff and Quastler introduced in 1953 the *information for the kind of atoms in a molecule* [5] (renamed

Information Theoretic Complexity Measures, Table 1

Comparison of the predicted and measured values of the nuclear binding energies of some nuclides of chemical elements #101 to 108 [12]

β	z	A	E_b (pred.)	E_b (exper.)	β	z	A	E_b (pred.)	E_b (exper.)
49	103	255	1889.235	1887.465	53	106	265	1945.166	1943.199
	104	257	1899.300	1896.954		107	267	1954.800	1952.342
	105	259	1909.259	1906.121	54	102	258	1913.639	1911.128
	106	261	1919.112	1915.447		103	260	1924.599	1919.621
	107	263	1928.862	1924.336		104	262	1935.477	1930.934
50	104	258	1903.592	1904.563		105	264	1946.285	1939.257
	105	260	1912.916	1912.603		106	266	1957.026	1950.468
	106	262	1922.085	1923.259	55	102	259	1916.513	1916.569
	107	264	1931.101	1930.932		103	261	1926.384	1926.418
	108	266	1939.969	1941.345		104	263	1936.086	1936.563
51	104	259	1912.138	1910.716		105	265	1945.581	1946.224
	105	261	1921.943	1920.042		106	267	1954.913	unknown
	106	263	1931.617	1929.620	56	101	258	1910.768	1911.701
	107	265	1941.163	1938.568		102	260	1920.146	1923.139
	108	267	1950.582	1947.803		103	262	1929.240	1931.927
52	103	258	1911.461	1906.916		104	264	1938.056	1943.295
	104	260	1922.186	1918.037		105	266	1946.616	unknown
	105	262	1932.856	1926.206	57	101	259	1918.088	1917.837
	106	264	1943.456	1937.123		102	261	1928.501	1928.317
	107	266	1953.973	1944.952		103	263	1938.762	1938.413
53	103	259	1915.474	1913.955		104	265	1948.874	unknown
	104	261	1925.505	1923.949		105	267	1958.840	unknown
	105	263	1935.472	1933.417					

later as *information for atomic composition* I_{ac}, and *information for chemical composition*, I_{CC}). Atoms are not the best structural units in describing complex biomolecules like DNA and proteins, and Branson [6] calculated the *information on aminoacid composition*, I_{AAC}, of 26 proteins. Augenstine [23] added the *configurational information* of polypeptide chains in order to calculate the total information content of proteins. Rashevsky [7] has shown that even molecules composed of the same kind of atoms can have a large information content based on the atom-atom connectivity, which he termed *topological information*. All these attempts enabled the first quantitative estimates [5,24] of the information content of a living cell and a human to be about 10^{11} and 10^{25} bits, respectively. It was argued that a certain lower limit of complexity must exist even for the simplest organisms. The topological information of Rashevsky was more specifically defined by classifying the vertices of a *molecular graph* (a two-dimensional representation of the molecule) into layers of first, second, etc. neighborhoods (an idea developed in more detail much later [25,26] as neighborhood complexity of first, second, etc. order). Trucco [8] has pointed out that the equivalence neighborhood criterion of Rashevsky cannot always provide topological equivalence of two graph vertices, and redefined the topological information descriptor in terms of the vertex orbits of the automorphisms group of the graph. Trucco [9] also extended the topological characterization of the molecule by partitioning the graph edges into the edge orbits of the graph. Later, these two types of topological descriptors of molecules were more concretely renamed as vertex and edge orbit's information, respectively [27]. Other structural aspects have been taken into account by Morovitz [28], who combined the information on the chemical composition of a molecule to that on *the possible valence bonds* between the atoms in the molecule. The relation between the vertex orbit information descriptor and the complexity of graphs has been studied in detail by Mowshowitz [10], who also introduced the chromatic information content of graphs [29]. This concludes the brief early history of molecular information theoretic descriptors. Since the beginning of the 1970s, the explosive development of chemical graph theory [30,31] contributed greatly to the corresponding variety of contributions to chemical information theory [27], which will be reviewed in the rest of this section.

Application of the Shannon Equation to Sets of Elements with Complex Equivalence Criteria, and to Weighted Sets of Elements

Three major types of complexity are considered in natural sciences: compositional, structural, and functional ones. The functional complexity is essential for the characterization of biological systems ► Biological Complexity and Biochemical Information, [32], but has not been a subject of studies in chemistry. The behavior of small molecules is described sufficiently accurately not only by quantum mechanics, but also by simple concepts like functional groups, polarity, acidic and basic properties, etc. For this reason, the present article deals with the information theoretic description of molecular compositional and structural complexity. It has been shown that the finite probability scheme of Mowshowitz [10], introduced in Sect. "Introduction", provides excellent descriptors of compositional complexity. However, for structural complexity this scheme does not work well, producing predominantly information descriptors of low sensitivity, which cannot match well structural patterns of different complexity [33]. Examples illustrating this conclusion will be given further in this section.

The search for a better solution resulted in two different approaches. The insufficiency of the equivalence criterion for partitioning the set of system elements into equivalence classes has prompted the simultaneous use of two or more such criteria. A typical situation is to group the atoms in a molecule first according to their chemical nature, and to partition further the classes thus formed according to some structural criterion, e. g., atoms of the same kind having different neighborhood of atoms [7].

A more general approach [34] is based on weighted graphs (vertex-, edge-, and vertex-edge-weighted ones). Weights can represent any measured or calculated quantitative characteristics of the atoms and bonds in the molecule, such as electronegativity, bond orders, and different graph theoretical invariants like vertex and edge degrees, distances, etc. The Mowshowitz finite probabilistic scheme has, thus, been extended so as to include not only the partition of elements of the system into k classes with N_i elements in class i, and equivalence-based probabilities $p_i = N_i/\sum_i N_i = N_i/N$, but also the weight w_i of type α of this class, and the "weighted" probability ${}^w p_i = w_i/\sum_i w_i = w_i/W$:

Equivalence classes	$1, 2, \ldots, k$
Partition of elements	$N_1, N_2, \ldots, N_k$
Probability distribution	$p_1, p_2, \ldots, p_k$
Weights	$w_1, w_2, \ldots, w_k$
Probability w-distribution	${}^w p_1, {}^w p_2, \ldots, {}^w p_k$.

The Shannon equations (3) and (4) are thus modified to express the information on the weighted distribution α of the system:

$$ {}^w I(\alpha) = W \log_2 W - \sum_{i=1}^{k} N_i w_i \log_2 w_i \,, \tag{22} $$

$$ {}^w \bar{I}(\alpha) = -\sum_{i=1}^{k} N_i \frac{w_i}{W} \log_2 \frac{w_i}{W} \,. \tag{23} $$

The two weighted information descriptors from Eqs. (22) and (23) are defined within the ranges:

$$ 0 \leqslant {}^w I(\alpha) \leqslant W \log_2 W \,; \quad 0 \leqslant {}^w \bar{I}(\alpha) \leqslant \log_2 W \,, \tag{24} $$

where the lower bound corresponds to a system without structure ($w_i = W$), and the upper bound can be attained by a system having the maximum number of classes with a single element of unit weight in each class ($k = N = W$). Typical weights that might be associated with vertices of molecular graphs are vertex degree (the sum of all edges emanating from the vertex) [27] or vertex distance (the sum of distances from that vertex to all other ones), as the weighted probabilistic scheme has been introduced for the first time [34].

Information Descriptors of Compositional Complexity

As mentioned above, the first information descriptor introduced is the *information on chemical composition*, I_{CC} [5], based on the distribution of atoms into subsets containing atoms of the same chemical element. This descriptor is a measure for the molecular complexity derived from its *compositional diversity*. An example illustrating this approach to molecular complexity is the increase in diversity within the series CH_4, CH_3F, CH_2FCl, CHFClBr. The atomic compositional distributions of these molecules {1,4}, {1,1,3}, {1,1,1,2}, and {1,1,1,1,1} produce the increase in I_{CC} of this series from the simplest molecule of CH_4 to the most complex one, CHFClBr (3.61, 6.85, 9,61, 11.61 bits, respectively).

A similar approach was used to characterize the composition of complex molecules like proteins, proceeding from the distribution of their aminoacid residues [6]. Thus, the insulin molecule having 51 aminoacid residues is characterized by the following composition: Cys (6), Leu (6), Val (5), Glu (4), Gly (4), Tyr (4), Ala (3), Asn (3), Glu (3), Phe (3), Ser (3), His (2), Arg (1), Ile (1), Lys (1), Pro (1), Thr (1) . The compositional distribution 51{6,6,5,4,4,4,3,3,3,3,3,2,1,1,1,1,1} produces the aminoacid information content of insulin, $I_{AAC} = 196.89$ bits. Information theory has been widely applied in a variety of versions to assess the complexity and evolution of

proteins [35,36,37,38,39,40,41]. Mostly, these applications have been based on the original interpretation of Shannon's equations by selecting the aminoacid sequence of a protein from the set of all possible sequences of the same length. The evolution of a protein characterized quantitatively in this manner is considered also as a measure of evolution of the corresponding DNA, which encodes that protein [42,43].

The complexity of DNA and its evolution has been itself a subject of very broad studies applying information theory ► Biological Complexity and Biochemical Information, [44,45,46,47,48], which could be a subject of a separate article. Different types of DNA information content have been defined. The simplest one is the DNA nucleotide composition, proceeding from the frequency of adenine (A), cytosine (C), guanine (G) and thymine (T) in protein-coding and -noncoding sequences. The DNA and individual genes' exon/intron composition, and codon composition has also been defined. The length of exons and introns can also be taken into account either in the weighted type information descriptors (Eqs. (22) and (23)) or by using the Jensen–Shannon divergence of sequence S, $J(S)$ [49,50]

$$J(S) = H(S) - \sum_{i=1}^{k} \frac{l_i}{L} H(S_i) \,, \tag{25}$$

where $H(S)$ is the Shannon sequence entropy, and $H(S_i)$ is the entropy of the ith sequence segment of length l_i. These two approaches are applicable to all possible genome segmentations.

Another approach widely used for determining the information content of symbolic sequences is Kolmogorov's algorithmic information [51], which has been defined as the length of the sequence shortest description in some fixed description language or as the length of the shortest code generating the given sequence. A variety of constructive realizations of such coding have been developed [52,53,54] and applied to DNA sequences [55,56,57]. The mutual information has been used as a tool for predicting protein coding regions in DNA [58]. More details on complexity of DNA and proteins can be found in the article ► Biological Complexity and Biochemical Information.

Information Measures of Topological Complexity of Molecules

Adjacency-Based Information Descriptors Graph theory [59,60] offers very effective graphical and quantitative tools for representing molecular structure. Molecules are three-dimensional structures; however the two-dimensional representation of molecular graphs conserves a very important part of the structural information. Molecular graphs are composed of points (termed *vertices*) representing atoms, and connecting lines (called *edges*) representing covalent chemical bonds. Atom-atom connectivity, which is the fundament of molecular topology, is characterized by the *adjacency matrix*, a square matrix with a size determined by the number of vertices V. The entries a_{ij} of this matrix are equal to one for any pair of adjacent vertices i and j, and zero otherwise. The sum of the entries in a matrix row is in fact the number of nearest neighbors of the vertex i associated with this row, and is called *vertex degree*, a_i. The sum of all vertex degrees defines the graph *total adjacency*, $A(G)$:

$$a_i(G) = \sum_{j=1}^{V} a_{ij} \,; \tag{26a}$$

$$A(G) = \sum_{i=1}^{V} a_i = \sum_{i=1}^{V} \sum_{j=1}^{V} a_{ij} \,. \tag{26b}$$

Equations (26a), (26b) allow the generation of two distributions. The first one is the distribution of vertices in equivalence classes, according to their degrees, whereas the second one partitions the total adjacency into classes of vertex degree values. The unweighted information index for vertex degrees, which is based only on their equivalence but not the values of vertex degrees, *cannot* serve as a complexity index, in contrast to its weighted counterpart, obtained from the second distribution. This is illustrated in Fig. 4 by the comparison of the complexity of graphs **1**, **2**, and **3**. Having two elementary cycles and higher vertex connectivity, graph **3** is topologically the most complex one, whereas the linear (path) graph **1** is the least complex one. However, the *unweighted* vertex degree distributions of graphs **1** and **3**, constructed by considering only the equivalence of vertices with the same degree, are the same (one class of three and one class of two vertices), which results in equal information content of $I_{\mathrm{vd}} = 4.85$ bits. Even worse is the complexity estimate of graph **2**, all five vertices of which are equivalent, and its information content is zero, while a cyclic graph is considered more complex than an acyclic one. The *weighted* distributions take into account the different values of vertices in these three graphs, and the partitioning of their total adjacencies: 8{2,2,2,1,1}, 10{2,2,2,2,2}, and 12{3,3,2,2,2}, respectively. The basic equation (22) for total weighted information content has thus been specified [27,33] for the vertex degree distribution as:

$${}^{w}I_{v\,\mathrm{deg}} = A \log_2 A - \sum_i a_i \log_2 a_i \,. \tag{26c}$$

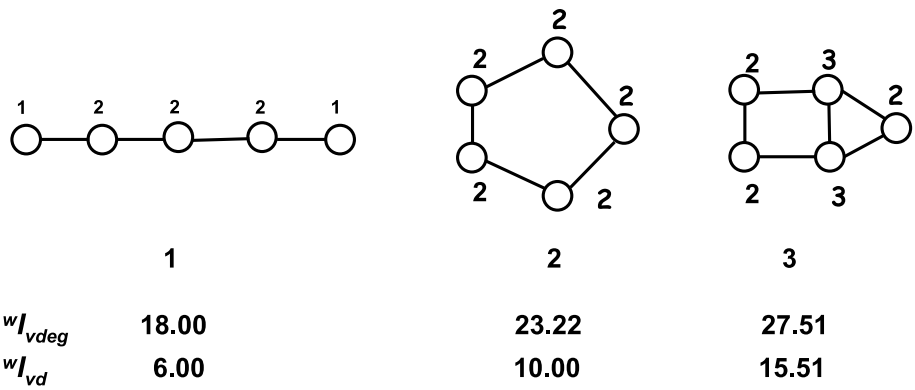

Information Theoretic Complexity Measures, Figure 4
Three graphs ordered according to their increasing topological complexity, as measured by two information theoretic complexity descriptors (Eqs. (26c), (26d)) based on the distribution of the vertex degree values. The latter are shown at each graph vertex

The information content of graphs **1–3** calculated from these distributions increases from the acyclic to the monocyclic to the bicyclic graph, and can be used as an approximate measure of graph complexity. Later study [61] has shown that the second term in Eq. (26c) is a more adequate measure of graph complexity. Denoted as $^{w}I_{\mathrm{vd}}$, and calculated by Eq. (26d),

$$I_{\mathrm{vd}} = \sum_{i=1}^{V} a_i \log_2 a_i \; ; \tag{26d}$$

its values are also shown in Fig. 4 and Fig. 7 (*vide infra*).

Different schemes have been applied in the attempts to make the unweighted information more sensitive to complex patterns in molecular structures. In the search for a better definition of vertex equivalence, Rashevsky [7] requested two equivalent vertices to have not only their first neighbors, but also their second, third, ..., kth neighbors to have the same degree. Trucco [8] has used the most rigorous definition of equivalence based on the orbits of the automorphisms group of the graph. (two vertices belong to the same graph orbit, if they can be interchange while preserving the adjacency). The vertex orbit information also fails to order correctly graphs **1–3**, according to their complexity. Thus, graph **1** has three orbits and the vertex distribution is 5{2,2,1}, in graph **2** all five vertices continue to belong the same equivalence class (orbit), and graph **3** has also three orbits like graph **1**, and they are with the same cardinality, 5{2,2,1}. Thus, the vertex orbit information, shows graphs **1** and **3** as equally complex, and graph **2** as the least complex, in contrast with the intuitive idea of topological complexity, which increases with the number of branches and cycles. Another way to a more sensitive description of molecular topology is to take into account structural elements of graphs that are more complex than the vertex. Using the equivalence of edges, [9] offers only a slight improvement, which results from the larger number of edges exceeding that of vertices in polycyclic graphs. Basak [25,26] considered the symmetry of sets of vertex neighborhoods of 1-, 2-, etc. order.

Bertz [62] made use of the equivalency of subgraphs of two adjacent edges, which he called "connections", and designed a complexity measure (termed later after his name) which has a better sensitivity to complexity features. Gordon and Kennedy [63] first used the number of such subgraphs as a measure of degree of branching of molecular skeleton. Before them, Platt [64] introduced a twice larger index obtained as the sum of all edge neighbors of each edge in the graph. In order to resolve the case with monocyclic graphs, which would always give zero equivalence-based information content no matter how large subgraphs are used, Bertz added a "size" term $n \log_2 n$ in his BI index:

$$BI = 2n \log_2 2n - \sum_{i=1}^{k} n_i \log_2 n_i \text{ bits} \; , \tag{27}$$

where n is the total number of two-edge subgraphs, k is the number of equivalence classes, and n_i is the number of such subgraphs in the ith equivalence class (connections orbits). In calculating the BI index for the three graphs, one finds for graph **1** that there are a couple of two-edge subgraphs with vertex degrees 1,2,2, and one two-edge subgraph with degrees 2,2,2, thus obtaining from Eq. (27) $BI(\mathbf{1}) = 7.55$. Graph **2** contains five identical two-edge subgraphs with vertex degrees 2,2,2, which gives $BI(\mathbf{2}) = n \log_2 n = 11.61 > BI(\mathbf{1})$. The most complex graph **3** contains nine such subgraphs: four of them with degrees 2,3,3, two with degrees 2,3,2, two with degrees 3,2,2, and one with degree 3,2,3, thus producing $BI(\mathbf{3}) = 45.06$. If a *weighted* version were used for the

graph "connections", Eq. (22) would suffice without any additional size term to produce the correct complexity ordering of these three graphs ($I_{\mathrm{conn}}(\mathbf{1}) = 16\log_2 16 - 2\times 5\log_2 5 - 6\log_2 6 = 9.29$, $I_{\mathrm{conn}}(\mathbf{2}) = 30\log_2 30 - 5\times 6\log_2 6 = 69.66$, $I_{\mathrm{conn}}(\mathbf{3}) = 68\log_2 68 - 5\times 8\log_2 8 - 4\times 7\log_2 7 = 215.34$).

Distance-Based Information Descriptors The first *weighted information theoretic indices* have been introduced by Bonchev and Trinajstić [34] in 1977 to characterize molecular branching [65,66,67,68,69,70,71] and, later, molecular cyclicity patterns [72,73,74,75] as basic components of molecular skeleton's complexity. The weighted distribution of vertex distances has been used. Graph distances are integers equal to the number of edges along the shortest path between pairs of nodes. The total *graph distance*, $D(G)$, is calculated as the sum of all entries (distances d_{ij} between pairs of vertices i and j) of the distance matrix $\boldsymbol{D(G)}$. The latter is symmetric with respect to its main diagonal. For this reason, the sum of distances in molecular graphs is frequently presented by the Wiener number [65,66], $W(G) = D(G)/2$. The sum of distances d_i from a vertex i to all other graph vertices is calculated as the sum of distance matrix' ith row entries d_{ij}:

$$d_i(G) = \sum_{j=1}^{V} d_{ij}\,; \tag{28a}$$

$$D(G) = \sum_{i=1}^{V} d_i = \sum_{i=1}^{V}\sum_{j=1}^{V} d_{ij}\,, \tag{28b}$$

where the sums run over all graph vertices V ($d_{ii} = 0$). The vertex distances thus calculated form a distribution $\{D\} \equiv \{d_{\max}, d_{\max}-1, \ldots, 2, 1\}$, which is used to define the *weighted information on graph vertex distances*, ${}^{w}I_{\mathrm{vdist}}(G)$ (also termed information on graph distance *magnitude*, ${}^{m}I_{\mathrm{vdist}}$):

$${}^{w}I_{\mathrm{vdist}}(G) = D(G)\log_2 D(G) - \sum_{i=1}^{d(\max)} N_i d_i(G)\log_2 d_i(G)\,, \tag{29}$$

$${}^{w}\bar{I}_{\mathrm{vdist}}(G) = -\sum_{i=1}^{d(\max)} N_i \frac{d_i(G)}{D(G)}\log_2\frac{d_i(G)}{D(G)}\,. \tag{30}$$

The ${}^{w}I_{\mathrm{vdist}}$ indices increase with the number of vertices, and decrease with the number, size, and more central position of the graph branches and cycles. The two opposing trends prevent the usage of these indices as more general complexity measures. Yet, they are very useful in assessing (in a reverse order) the relative complexity of isomeric molecules or graphs having the same number of vertices. Such studies have been performed to define a number of branching [65,71] and cyclicity [72,73,74,75] complexity patterns. One of the branching patterns rigorously proved, shown in Fig. 5a, is defined as follows: "Branched trees are more complex than path graphs, and less complex than star graphs". Here, the term *tree* is used for acyclic graphs, *path* is a tree without branches, and *star graphs* with V vertices have a single central point, and

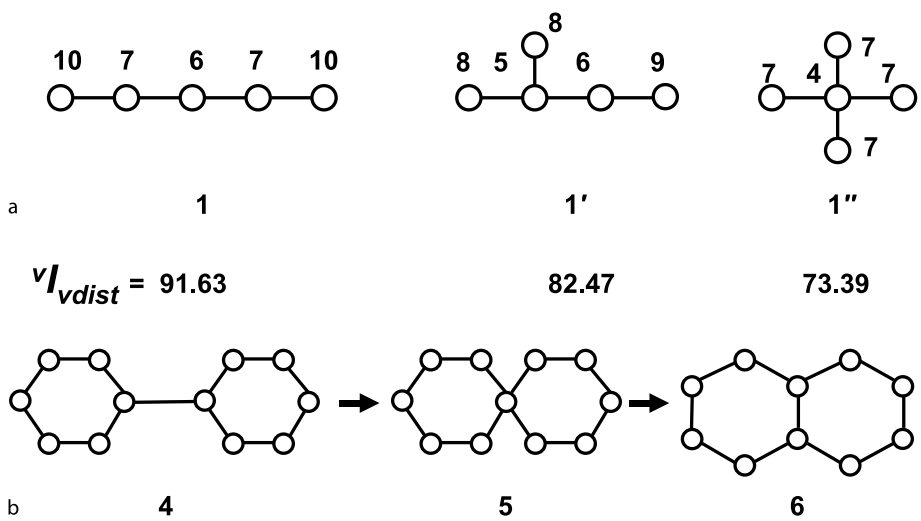

Information Theoretic Complexity Measures, Figure 5
Patterns of increasing complexity in series of a acyclic, and b cyclic graphs, as measured in a reverse order by the values of the weighted information on vertex distance distribution Eq. (30). The vertex distances shown are used in the calculations

$N-1$ branches of a unit length. Figure 5b presents a series of bicyclic structures **4–6** with complexity increasing with the stronger type of cycle connectivity, from a bridge, to a shared vertex, to a shared edge.

Centrality Information Descriptors Another topological complexity factor is *graph centrality*, the graph organization with respect to a certain central point, which can be defined in a variety of ways. Most of these have been introduced in network analysis, and will be discussed in Sect. "Information Content of Networks". Here, the analysis is focused on distance-based centrality. The general definition for a graph center [59] states that the center is the vertex with the lowest *eccentricity* $e(i)$, eccentricity being defined as the longest distance from a given vertex to any other vertex in the graph. This definition frequently classifies as central several nonequivalent graph vertices. A hierarchical definition based on several criteria has been proposed, including Harary's definition as the first criterion [76,77]. The second criterion reduces the set of central vertices to those, which have the lowest eccentricity and the smallest vertex distance, d_i. When several vertices have the same minimal eccentricity and the same minimal vertex distance, then the third criterion, requiring a minimal occurrence of the largest distance, $n_{ij}(\max)$, is used:

$$\text{Criterion 1: } e_i = \max(d_{ij}) = \min\,, \tag{31a}$$

$$\text{Criterion 2: } d_i = \sum_j d_{ij} = \min\,, \tag{31b}$$

$$\text{Criterion 3: } n_{ij}(\max) = \min\,. \tag{31c}$$

Hierarchical criteria 1–3 reduce considerably the number of nonequivalent graph centers, yet, in some cases some nonequivalent central vertices still exist. The final solution found [78] resolves effectively the problem for centric ordering of the vertex and edge orbits of the automorphism group of the graph on the basis of the iterative vertex–edge centricity (IVEC) concept:

> Central are those vertices that are incident to the most central edges and, vice versa, central are those edges that are adjacent to the most central vertices.

A similar approach has been later applied in the Google search engine algorithm [79].

Once the graph center is defined, a vertex (as well as an edge) centric distribution can be constructed, proceeding from the distance from each vertex to the center. (In case of several equivalent central vertices, the distance to the closest one is used.) The vertex centric organization is considered more complex when it is more compact, i. e., when vertices are organized in a smaller number of layers around the center or, at the same number of layers, when the outer layer(s) incorporate a smaller number of vertices. This complexity pattern matches the fact that the highly organized networks in living systems and technology are "small-world" ones [80,81], i. e., they have a rather small diameter.

The centrality patterns of complexity are illustrated in Fig. 6 with three polycyclic graphs having seven vertices each. Their ordering from the least complex to the most complex central organization is inversely proportional to the values of their distance-based information descriptors, the weighted information for the vertex distances distribution, ${}^{w}I_{\text{vdist}}$, and the weighted information for the centric vertex distribution, ${}^{w}I_{\text{vcentr}}$. Graph **7** has the most compact structure with eccentricity equal to one for the central vertex. The several central points in graphs **7** and **8** have all the same eccentricity equal to two. However, the smallest vertex distances criterion (Fig. 6a) allows one to reduce the centers to a single vertex in both graphs. The layers of centric ordering of graphs **7–9** can be traced from Fig. 6b, where vertices are labeled by their distances to the center. As seen, graph **9** is centrically the most complex having only one layer of centric neighborhood. Graphs **7** and **8** have both two layers of vertices around the center; however, graph **8** is considered more complex than graph **7**, because of the smaller number of vertices in the outer layer (1 vs. 2).

Combined Information Complexity Descriptors Based on Vertex Degrees and Distances It was shown in the foregoing that while the information measure of topological complexity based on vertex degree distribution increases in parallel with the increase in such complexity elements as branches and cycles, the measure based on the distribution of vertex distances decreases. It is logical to expect that the ratio A/D of the total adjacency and total distance of a graph could combine the two measures in a way to vary in direct proportionality of increasing complexity. A more sensitive complexity measure based on the same idea proceeds from calculating the sum of vertex ratios $b_i = a_i/d_i$, as well as from calculating the information on the distribution of these ratios, $I_{a/d}$. Thus, three complexity descriptors have been defined, and called Bourgas indices $B1$, $B2$, and $B3$, respectively [61,82,83]:

$$B1 = \frac{A}{D}\,, \tag{32a}$$

$$b_i = \frac{a_i}{b_i}\,, \tag{32b}$$

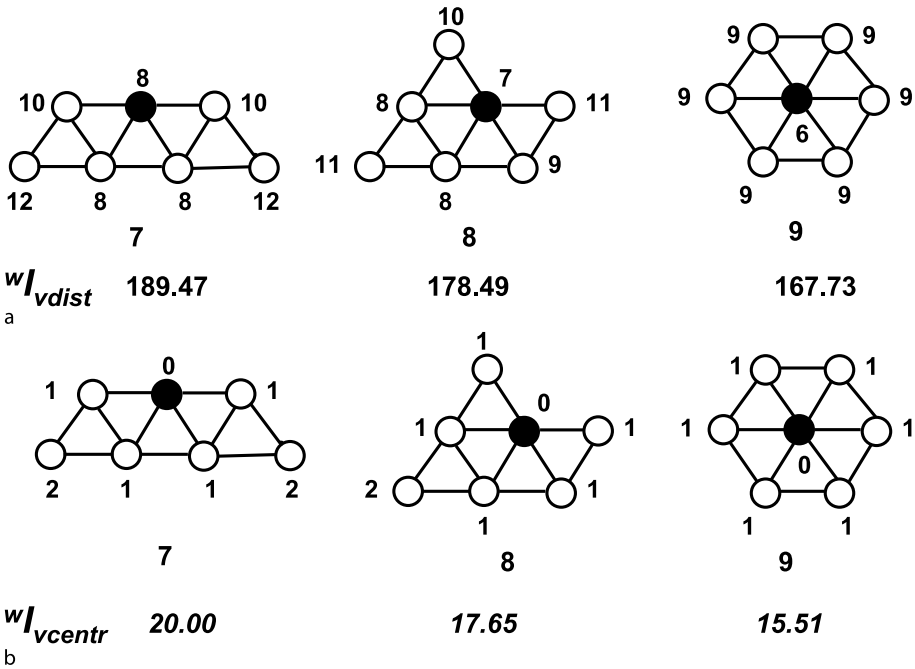

Information Theoretic Complexity Measures, Figure 6
a Three polycyclic graphs ordered according to their increasing centric complexity, and decreasing weighted information on the vertex distance distribution. The centers are presented by *full points*, and the numbers at each vertex are the vertex distances (the total distance from the vertex to all other vertices); b The same graph ordering corresponds also the decreasing weighted information on the distribution of distances from each vertex to the center (shown at each vertex)

$$B2 = \sum_{i=1}^{V} b_i = \sum_{i=1}^{V} \frac{a_i}{b_i} \,, \tag{32c}$$

$$B3 = B2 \log_2 B2 - \sum_{i=1}^{V} b_i \log_2 b_i \,. \tag{32d}$$

The Bourgas indices are shown in Fig. 7 to increase from graphs **7** to graph **8** to graph **9**, in agreement with the complexity ordering of these graphs shown in Fig. 6.

Subgraph-Based Information Complexity Descriptors The idea of using fragments of molecular graphs for a more complete representation of molecular structure emerged in the 1960s [84]. It has been developed in detail by Gordon and Kennedy [85], and has found a full realization in the molecular connectivity concept of Kier and Hall [86,87], widely applied in drug design. The design of subgraph-based complexity descriptors has been reported at the end of the 1990s [88,89,90,91,92,93,94,95,96,97,98,99,100], although an idea for such a complexity/similarity measure was proposed a decade earlier [101].

The *subgraph count*, *SC*, introduced initially under different names [88,89,90,91,92,93], proceeds from the logical assumption that the larger the number of subgraphs, the more complex the graph. This concept has been enriched by weighting each of the subgraphs with its total adjacency: *the larger the adjacency of the graph and its subgraphs, the higher the complexity*. The sum of the total adjacencies of all subgraph was termed *overall connectivity index*, *OC* (also initially named differently) [88,90,91,93,94]. Other weighting schemes, using the subgraph total distance and two simple functions of vertex degrees, have also been applied, and named as *overall Wiener index* [100] and *overall Zagreb indices* [102,103,104], respectively. Rücker and Rücker [96,97,98] used another graph invariant as a complexity measure, *the total walk count*, *TWC*, which is the total number of graph walks of all possible lengths l allowed by the graph size (A *walk* of length l on a graph is a sequence of vertices and edges such that edges can be traversed repeatedly). As in the case with subgraph count, *the larger the total walk count, the more complex the graph*. All these sophisticated measures of topological complexity have also been used to define an information-type of complexity indices [33]. For this purpose, they are considered to be partitioned into e classes, according to the number of edges e in the subgraph (^{e}SC, ^{e}OC, ^{e}OW) and for *TWC*,

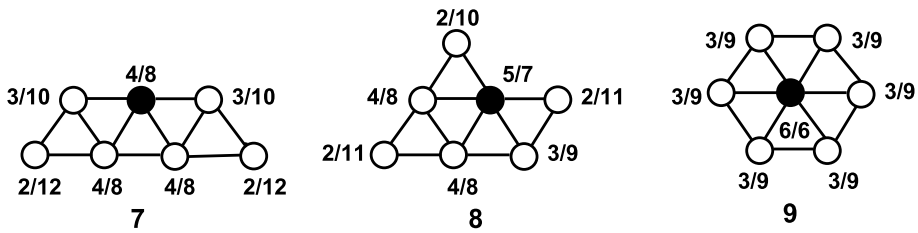

	7	8	9
B1	0.29	0.34	0.40
B2	2.43	2.61	3.00
B3	3.21	5.97	7.92

Information Theoretic Complexity Measures, Figure 7
Three polycyclic graphs in increasing complexity order, well matched by the three Bourgas complexity indices (Eqs. (32a), (32b), (32c), (32d)), combining vertex degrees and distances

into classes of walks of different length l, ${}^{l}TWC$. All these cases can be summarized by the general equations:

$$X(G) = \sum_{e=1}^{E} {}^{e}X\,; \quad \{X\} \equiv \{{}^{1}X, {}^{2}X, \ldots, {}^{E}X\}\,, \tag{33}$$

$$I_X(G) = X \log_2 X - \sum_{e=1}^{E} {}^{e}X \log_2 {}^{e}X\,. \tag{34}$$

Here, the overall topological complexity descriptor is defined as $X \equiv SC, OC, OW$, and TWC; E stands for the total number of graph edges and, for TWC, e and E are replaced by l and $L_{\max}$, respectively. The subgraph-based information descriptors are designed to increase with the increasing number of structurally complexifying elements like branches and cycles, and other, more subtle topological factors. In Fig. 8, they are compared to the information on vertex degree distribution ${}^{w}\mathbf{I}_{\mathrm{vd}}$ and the one for the combined $B3$ descriptor of vertex degree and distance distribution introduced earlier in the text.

This concludes the analysis of information theoretic descriptors for the topological complexity of molecules. More information about topological complexity descriptors not using information theory can be found in another article of this volume ► Topological Complexity of Molecules.

Information Content of Networks

Networks composed of biologically important molecules, such as DNA, RNA, proteins, and metabolites, turned in the beginning of the 21st century into a focus of major interest in computational biology and bioinformatics, as a universal language to describe the systems approach in biology. The links (edges) connecting the nodes

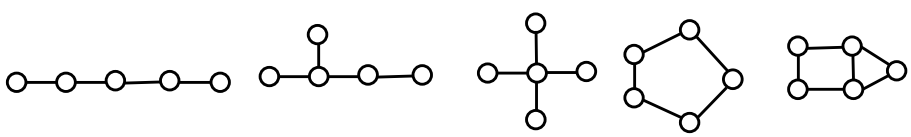

	1	10	11	2	3
${}^{w}I_{vd}$	6.00	6.75	8.00	10.00	15. 51
$B3$	2.385	2.554	2.628	3.871	4.972
I_{SC}	32.24	37.12	43.32	64 .16	138.20
I_{OC}	135.8	166.6	209.2	232.2	1136.4
I_{OW}	101.3	118.5	137.6	335.0	904.5

Information Theoretic Complexity Measures, Figure 8
The increase in topological complexity with the increase in the number of branches and cycles is precisely matched by the five information theoretic descriptors, based on vertex degrees, combination of vertex degrees and distances, subgraph count, overall connectivity, and overall Wiener subgraph distributions

(vertices) of these networks can represent a variety of interactions, from physical and chemical interaction (protein-protein interaction; posttranslational protein modification) to regulation (gene-gene, protein-gene) to co-expression (gene-gene), molecular transport (proteins), and others. The topological characterization of these dynamic evolutionary networks [61,83,105,106] proceeds from the same basic features as in characterizing molecules (which are in fact atomic networks): connectivity, distances, and subgraphs. There are, however, considerable differences in the distributions of the underlying graph invariants. While vertex degrees in molecular graphs rarely exceed the value of four, in biological networks the maximal node degree may be very high. Such highly connected nodes, termed *hubs*, are essential for the existence of the living cell; their deletion or malfunctioning is usually lethal. The distribution of vertex degrees in the complex networks in nature and technology seems to obey a scale-free law [107,108], as a consequence of which there are few hubs and many low-connected nodes. Another important topological feature of these networks is their very small diameter; the networks are "small-world" type [80]. This very compact organization is the key to life's resilience against attacks of different kinds. A signature of a complex network is its specific composition of subgraphs, called *motifs*, which are much more abundant than in random networks of the same size [109]. Larger subgraphs, termed *complexes*, *modules* and *pathways*, are responsible for the performance of specific biological functions.

The *compositional complexity* of biochemical networks is determined similarly to molecular one. One may consider, for example, a protein-protein interaction network as composed of protein complexes, which in turn include individual proteins. Similarly, a metabolic network may be regarded as being built of metabolic pathways containing individual metabolites and enzymes. The larger the diversity of network's constituents, the higher the network compositional complexity. Equations (3), (4) are of use for calculation of the compositional information content, as used for molecules.

Networks *topological complexity* also makes use of the tools introduced for assessing the topological complexity of molecules. The information descriptor of vertex degree distribution, ${}^{w}I_{\mathrm{vd}}$, and the Bourgas information index $B3$ are applied directly as defined by Eqs. (26d) and (32d), respectively, proceeding from the network connectivity and distance data. Some adaptation is needed to apply the information descriptors for network subgraphs. Due to the size of intracellular networks, which may exceed in some cases 10,000 nodes, the calculation of the subgraph-based complexity descriptors would lead to combinatorial explosion. However, it suffices for sufficiently accurate assessments of network complexity to use the distribution of small subgraphs having only three or four edges or the distribution of walks of length 3 (Fig. 9).

Equation (34) has to be slightly modified to include summing up not over all subgraphs with e edges or all walk lengths l, but at a given size (e or l) the sum has to be taken over all types i of subgraphs or walks, as illustrated in Fig. 9:

$$ {}^{e(l)}I_X(G) = {}^{e(l)}X \log_2 {}^{e(l)}X - \sum_i {}^{e(l)}X_i \log_2 {}^{e(l)}X_i \,. \quad (35) $$

Here, ${}^{e(l)}X = \sum_i {}^{e(l)}X_i$, and X can be each one of subgraph count (SC), overall connectivity (OC), overall Wiener (OW), and total walk count (TWC), as defined

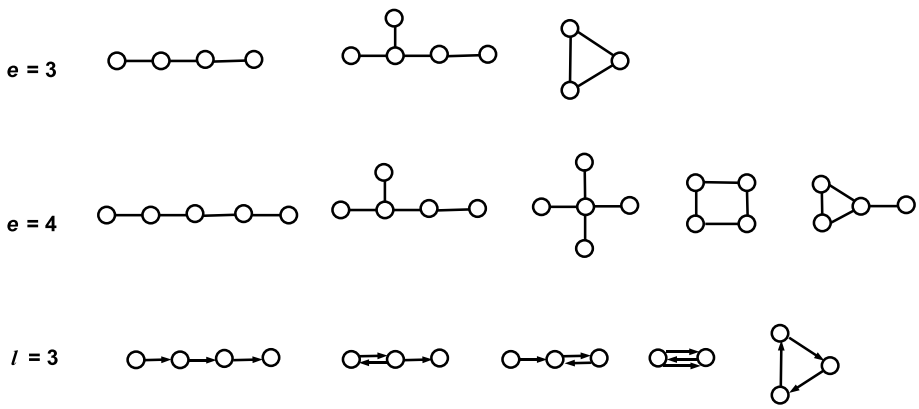

Information Theoretic Complexity Measures, Figure 9
All subgraphs with 3 and 4 edges, and all walks of length 3, recommended for assessing complexity of large size networks by Eq. (35)

in Subsect. "Subgraph-Based Information Complexity Descriptors".

The five information descriptors of network complexity shown in Fig. 8 can be applied to both undirected and directed networks. The specifics of directed graphs, however, require replacing the single-value vertex degree by in-degree and out-degree, which are defined as the total number of directed edges incoming to the vertex and outgoing from it, respectively. The $B3$ descriptor, combining vertex degrees and distances, cannot be directly applied for directed graphs, because there is no path between some pairs of vertices and the distance between such vertices is equal to infinity. A procedure has been proposed to recalculate the total distance in directed networks by taking into account the vertex accessibility [61,110].

Future Directions

The use of information theory in characterizing molecular structure and composition has been strongly influenced by the pioneering work of mathematical biologists in the 1950s [2]. In turn, the considerable expertise accumulated in chemical information theory during the last 30 years is providing nowadays its feedback to bioinformatics and computational biology by offering not only a variety of specific information descriptors for topological complexity, but also the much more sensitive weighted version of Shannon's basic equations. Along with the other developed graph theoretic complexity measures, the information complexity indices are expected to find a broad application for large-scale comparative and evolutionary studies in biology. Finding the most adequate quantitative measure of evolution is a great challenge, along with elucidating the minimum compositional, topological and functional complexity needed for the emergence of life. Practical results may be expected soon from network topology/biological function quantitative studies in the area of biomedical research, in which the information measures of network complexity will play an important role. Theoretical chemistry is also expected to benefit greatly from the exchange of ideas with network analysis in biology. The network modular and clique structure could be of interest for studies in atomic clusters, nanotechnology and crystallography. The concept of centrality, which has been very broadly developed in bioinformatics, might prove of importance to some areas of organic chemistry. The recent application of the bipartivity measure of networks [111] to the stability of fullerenes [112] and branching of molecular skeletons [113] are only among the first signs of such future developments in theoretical chemistry.

Bibliography

1. Shannon C, Weaver W (1949) Mathematical theory of communications. University of Illinois Press, Urbana
2. Quastler H (ed) (1953) Essays on the use of information theory in biology. University of Illinois Press, Urbana
3. Boltzmann L (1866) Über die mechanische Bedeutung der zweiten Hauptsatzes der Wärmetheorie. Wien Ber 53:195
4. Brillouin L (1956) Science and information theory. Academic Press, New York
5. Dancoff SM, Quastler H (1953) The information content and error rate of living things. In: Quastler H (ed) Essays on the use of information theory in biology. University of Illinois Press, Urbana
6. Branson HR (1953) Information theory and the structure of proteins. In: Quastler H (ed) Essays on the use of information theory in biology. University of Illinois Press, Urbana
7. Rashevsky N (1955) Life, information theory, and topology. Bull Math Biophys 17:229–235
8. Trucco E (1956) A note on the information content of graphs. Bull Math Biophys 18:129–135
9. Trucco E (1956) On the information content of graphs: Compound symbols; Different states for each point. Bull Math Biophys 18:237–253
10. Mowshowitz A (1968) Entropy and the complexity of graphs. I. An index of the relative complexity of a graph. Bull Math Biophys 30:175–204
11. Bonchev D (1979) Information indices for atoms and molecules. Commun Math Comput Chem (MATCH) 7:65–113
12. Bonchev D (2006) Periodicity of chemical elements and nuclides. Information theoretic analysis. In: Rouvray DH, King RB (eds) Mathematics of the periodic table. Nova Science, New York, pp 161–188
13. Bonchev D, Peev T (1973) Information theoretic study of chemical elements. Mean information content of a chemical element. God Vissh Khim–Technol Inst Bourgas 10:561–574
14. Bonchev D, Kamenska V, Kamenski D (1977) Informationsgehalt chemischer Elemente. Monatsh Chem 108:477–487
15. Bonchev D, Kamenska V (1978) Information theory in describing the electronic structure of atoms. Croat Chem Acta 51:19–27
16. Dimov D, Bonchev D (1976) Spin-information equations of the groups and periods in the periodic table of chemical elements. Commun Math Comput Chem (MATCH) 2:111–115
17. Bonchev D, Kamenska V, Tashkova C (1976) Equations for the elements in the periodic table based on information theory. Commun Math Comput Chem (MATCH) 2:117–122
18. Bonchev D (1981) Information theory interpretation of the Pauli principle and Hund rule. Intern J Quantum Chem 19:673–679
19. Bonchev D, Kamenska V (1981) Predicting the properties of the 113–120 transactinide elements. J Phys Chem 85:1177–1186
20. Bonchev D, Peev T, Rousseva B (1976) Information study of atomic nuclei. Information on proton–neutron composition. Commun Math Comput Chem (MATCH) 2:123–137
21. Rousseva B, Bonchev D (1978) A theoretic-information variant of nuclide systematics. Commun Math Comput Chem (MATCH) 4:173–192
22. Rousseva B, Bonchev D (1980) Prediction of the nuclear bind-

ing energies of the nuclides of period VII. Radiochem Radioanal Lett 45:341–346
23. Augenstine L (1953) Remarks on Pauling's protein models. In: Kastler H (ed) Essays on the use of information theory in biology. University of Illinois Press, Urbana
24. Linshitz H (1953) The information content of a bacterial cell. In: Kastler H (ed) Essays on the use of information theory in biology. University of Illinois Press, Urbana
25. Magnusson VR, Harris DK, Basac SC (1983) Topological indices based on neighborhood symmetry. In: King RB (ed) Chemical applications of topology and graph theory. Elsevier, Amsterdam, pp 178–191
26. Basak SC (1999) Information theoretic indices of neighborhood complexity and their applications. In: Devillers J, Balaban AT (eds) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach Science Publishers, Chichester, pp 563–593
27. Bonchev D (1983) Information-theoretic indices for characterization of chemical structures. Research Studies Press, Chichester
28. Morovitz H (1955) Some order-disorder considerations in living systems. Bull Math Biophys 17:81–86
29. Mowshowitz A (1968) Entropy and the complexity of graphs: IV. Entropy measures and graphical structure. Bull Math Biophys 30:533–546
30. Balaban AT (ed) (1976) Chemical applications of graph theory. Academic Press, London
31. Trinajstić N (1983) Chemical graph theory. CRC Press, Boca Raton
32. Szostak JW (2003) Functional information: Molecular messages. Nature 423:689
33. Bonchev D (2003) Shannon's information and complexity. In: Bonchev D Rouvray DH (eds) Complexity in chemistry. Mathematical Chemistry Series, vol 7. Taylor and Francis, Boca Raton, pp 155–187
34. Bonchev D, Trinajstić N (1977) Information theory, distance matrix and molecular branching. J Chem Phys 67:4517–4533; (1982) Chemical information theory. Structural aspects. Intern J Quantum Chem Symp 16:463–480
35. Yockey HP (1977) On the information content of cytochrome C. J Theor Biol 67:345–376
36. Yockey HP (1992) Information theory and molecular biology. University Press, Cambridge
37. Eigen M (1992) Steps toward life. Oxford University Press, Oxford
38. Kauffman SA (1992) Applied molecular evolution. J Theor Biol 157:1–7
39. Kauffman SA (1993) The origins of order. Oxford University Press, New York
40. Volkenstein MV (1994) Physical approaches to biological evolution. Springer, Berlin
41. Adami C (2004) Information theory in molecular biology. Phys Life Rev 1:3–22
42. Hasegawa M, Yano T-A (1975) Entropy of the genetic information and evolution. Orig Life Evol Biosph 6:219–227
43. Strait BJ, Dewey TG (1996) The Shannon information entropy of protein sequences. Biophys J 71:148–155
44. Gatlin LL (1972) Information theory and the living system. Columbia University Press, New York
45. Li W (1997) The complexity of DNA. Complexity 3:33–37
46. Schneider TD (1997) Information content of individual genetic sequences. J Theor Biol 189:427–441
47. Schneider TD (2000) Nucl Acids Res 28:2794–2799
48. Adami C (2002) What is complexity? BioEssays 24:1085–1094
49. Lin J (1991) Divergence measures based on the Shannon entropy. IEEE Trans Inf Theor 37:145–151
50. Bernaola-Galvan P, Roman-Roldan R, Oliver J (1996) Compositional segmentation and long-range fractal correlations in DNA sequences. Phys Rev E 53:5181–5189
51. Kolmogorov A (1965) Three approaches to the quantitative definition of information. Probl Inf Transm 1:4
52. Lempel A, Ziv J (1976) On the complexity of finite sequences. IEEE Trans Inf Theory IT-22:75–81
53. Gusev VD, Kulichkov VA, Chupahina OM (1991) Complexity analysis of genomes. I. Complexity and classification methods of detected structural regularities. Mol Biol (Mosk) 25:825–834
54. Chen X, Kwong S, Li MA (1999) Compression algorithm for DNA sequences and its applications to genome comparison. Genome Inform Ser Workshop Genome Inform 10:51–61
55. Ebeling W, Jimenez-Montano MA (1980) On grammars, complexity, and information measures of biological macromolecules. Math Biosci 52:53–71
56. Grumbach S, Tahi F (1994) A new challenge for compression algorithms: genetic sequences. J Inf Process Manag 30:875–886
57. Mantegna RN, Buldyrev SV, Goldberger AL, Havlin S, Peng C-K, Simons M, Stanley HE (1994) Linguistic features of noncoding DNA sequences. Phys Rev Lett 73:3169–3172
58. Grosse I, Herzel H, Buldyrev SV, Stanley HE (2000) Species independence of mutual information in coding and noncoding DNA. Phys Rev E 61:5624–5629
59. Harary F (1971) Graph theory, 2nd printing. Addison-Wesley, Reading
60. Gross JK, Yellen J (eds) (2004) Handbook of graph theory. CRC Press, Boca Raton
61. Bonchev D, Buck GA (2005) Quantitative measures of network complexity. In: Bonchev D, Rouvray DH (eds) Complexity in chemistry, biology and ecology. Springer, New York, pp 191–235
62. Bertz S (1981) The first general index of molecular complexity. J Am Chem Soc 103:3599–3601
63. Gordon M, Scantlebury GR (1964) Non-random polycondensation: Statistical theory of the substitution effect. Trans Faraday Soc 60:604–621
64. Platt JR (1952) Prediction of isomeric differences in paraffin properties. J Phys Chem 56:328–336
65. Wiener H (1947) Structural determination of paraffin boiling points. J Am Chem Soc 69:17–20
66. Wiener H (1948) Relation of the physical properties of the isomeric alkanes to molecular structure. J Phys Chem 52:1082–1089
67. Hosoya H (1971) Topological index: A newly proposed quantity characterizing the topological nature of structured isomers of saturated hydrocarbons. Bull Chem Soc Jpn 44:2332–2339
68. Lovasz L, Pelikan J (1973) On the eigenvalues of trees. Period Math Hung 3:175–182
69. Gutman I, Ruščić B, Trinajstić N, Wilcox CW Jr (1975) Graph theory and molecular orbitals. 12. Acyclic polyenes. J Chem Phys 62:3399–3405

70. Randić M (1975) On characterization of molecular branching. J Am Chem Soc 97:6609–6615
71. Bonchev D (1995) Topological order in molecules 1. Molecular branching revisited. J Mol Struct (Theochem) 336:137–156
72. Bonchev D, Mekenyan O, Trinajstić N (1980) Topological characterization of cyclic structures. Intern J Quantum Chem 17:845–893
73. Mekenyan O, Bonchev D, Trinajstić N (1979) Topological rules for spiro-compounds. Commun Math Comput Chem (MATCH) 6:93–115
74. Mekenyan O, Bonchev D, Trinajstić N (1981) On algebraic characterization of bridged polycyclic compounds. Intern J Quantum Chem 19:929–955
75. Mekenyan O, Bonchev D, Trinajstić N (1981) A topological characterization of cyclic structures with acyclic branches. Commun Math Comput Chem (MATCH) 11:145–168
76. Bonchev D, Balaban AT, Mekenyan O (1980) Generalization of the graph center concept, and derived topological indexes. J Chem Inf Comput Sci 20:106–113
77. Bonchev D (1989) The concept for the center of a chemical structure and its applications. Theochem 185:155–168
78. Bonchev D, Mekenyan O, Balaban AT (1989) An iterative procedure for the generalized graph center in polycyclic graphs. J Chem Inf Comput Sci 29:91–97
79. Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine. Proceedings of the seventh international conference on World Wide Web, Brisbane, April 1998, pp 107–117
80. Watts DJ, Strogatz SH (1998) Collective dynamics of "small-world" networks. Nature 393:440–442
81. Newman MEJ, Strogatz SH, Watts DJ (2001) Random graphs with arbitrary degree distributions and their applications. Phys Rev E 64:026118
82. Bonchev D (2005) My life-long journey in mathematical chemistry. Intern Electr J Mol Design 4:434–490
83. Bonchev D, Buck G (2007) From molecular to biological structure and back. J Chem Inform Model 47:909–917
84. Smolenski EA (1964) Graph-theory application to the calculations of structural-additive properties of hydrocarbons. Zh Fiz Khim 38:1288–1291
85. Gordon M, Kennedy JW (1973) The graph-like state of matter. Part 2. LCGI schemes for the thermodynamics of alkanes and the theory of inductive inference. J Chem Soc Faraday Trans 2 69:484–504
86. Kier LB, Hall LH (1976) Molecular connectivity in chemistry and drug research. Academic Press, New York
87. Kier LB, Hall LH (1986) Molecular connectivity in structure-activity analysis. Research Studies Press, Letchworth
88. Bonchev D (1995) Kolmogorov's information, Shannon's entropy, and topological complexity of molecules. Bulg Chem Commun 28:567–582
89. Bertz SH, Sommer TJ (1997) Rigorous mathematical approaches to strategic bonds and synthetic analysis based on conceptually simple new complexity indices. J Chem Soc, Chem Commun 2409–2410
90. Bonchev D (1997) Novel indices for the topological complexity of molecules. SAR QSAR Environ Res 7:23–43
91. Bonchev D, Seitz WA (1997) The concept of complexity in chemistry. In: Rouvray DH (ed) Concepts in chemistry: a contemporary challenge. Wiley, New York, pp 353–381
92. Bertz SH, Wright WF (1998) The graph theory approach to synthetic analysis: definition and application of molecular complexity and synthetic complexity. Graph Theory Notes New York Acad Sci 35:32–48
93. Bonchev D (1999) Overall connectivity and molecular complexity. In: Devillers J, Balaban AT (eds) Topological indices and related descriptors. Gordon and Breach, Reading, pp 361–401
94. Bonchev D (2000) Overall connectivities/topological complexities: a new powerful tool for QSPR/QSAR. J Chem Inf Comput Sci 40:934–941
95. Nicolić S, Tolić M, Trinajstić N, Bauèić I (2000) On the Zagreb indices as complexity indices. Croat Chem Acta 73:909–921
96. Rücker G, Rücker C (2000) Automatic enumeration of molecular substructures. MATCH – Commun Math Comput Chem 41:145–149
97. Rücker G, Rücker C (2000) Walk count, labyrinthicity and complexity of acyclic and cyclic graphs and molecules. J Chem Inf Comput Sci 40:99–106
98. Rücker G, Rücker C (2001) Substructure, subgraph and walk counts as measures of the complexity of graphs and molecules. J Chem Inf Comput Sci 41:1457–1462
99. Bonchev D (2001) Overall connectivity – a next generation molecular connectivity. J Mol Graph Model 5271:1–11
100. Bonchev D (2001) The overall Wiener index – a new tool for characterization of molecular topology. J Chem Inf Comput Sci 41:582–592
101. Bertz SH, Herndon WC (1986) The similarity of graphs and molecules. In: Pierce TH, Hohne BA (eds) Artificial intelligence applications to chemistry. ACS, Washington, pp 169–175
102. Bonchev D, Trinajstić N (2001) Overall molecular descriptors. 3. Overall Zagreb indices. SAR QSAR Environ Res 12:213–235
103. Nikolić S, Trinajstić N, Tolić M, Rücker G, Rücker C (2003) On molecular complexity indices. In: Bonchev D, Rouvray DH (eds) Complexity in chemistry. Mathematical chemistry series, vol 7. Taylor and Francis, London, pp 29–89
104. Bonchev D (2005) The overall topological complexity indices. Lecture series on computer and computational sciences, vol 4. Koninklijke Brill NV, Leiden, pp 1554–1557
105. Bonchev D (2003) Complexity of protein-protein interaction networks, complexes and pathways. In: Conn M (ed) Handbook of proteomics methods. Humana, New York, pp 451–462
106. Bonchev D (2004) Complexity analysis of yeast proteome network. Chem Biodivers 1:312–326
107. Barabasi A-L, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512
108. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi A-L (2000) The large-scale organization of metabolic networks. Nature 407:651–654
109. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D et al (2002) Network motifs: simple building blocks of complex networks. Science 298:824–827
110. Bonchev D (2003) On the complexity of directed biological networks. SAR QSAR Envir Res 14:199–214
111. Estrada E, Rodríguez-Velázquez JA (2005) Spectral measures of bipartitivity in complex networks. Phys Rev E 72:055510
112. Došlić T (2005) Bipartivity of fullerene graphs and fullerene stability. Chem Phys Lett 412:336–340
113. Estrada E, Rodriguez-Velazquez JA, Randić M (2006) Atomic branching in molecules. Int J Quantum Chem 106:823–832

Infrasound from Earthquakes, Tsunamis and Volcanoes

Milton Garces[1], Alexis Le Pichon[2]
[1] Infrasound Laboratory, HIGP, SOEST, University of Hawaii, Manoa, Kailua–Kona, USA
[2] CEA/DASE/LD, Bruyères-le-Châtel, France

Article Outline

Glossary
Definition of the Subject
Introduction
Infrasound Arrays
Earthquake Infrasound
Tsunami Infrasound
Volcano Infrasound
Future Directions
Concluding Remarks
Acknowledgments
Bibliography

Glossary

Infrasound atmospheric sound waves with frequencies lower than the 20 Hz hearing threshold of the human ear.

Infrasound array four or more horizontally separated identical microphones or microbarometers with precisely known locations that optimize the reception of a specified wavelength range.

Trace velocity apparent horizontal phase velocity of an acoustic arrival measured by an array.

Celerity effective propagation speed of a signal, measured from the ratio of the total range over the total travel time along the great circle path from a source to a receiver.

Tremor volcanic signal consisting of a nearly continuous oscillation of the ground and atmosphere, with durations of minutes to years. Harmonic tremor may have multiple distinct spectral peaks.

LP Long Period event. Transient volcanic signal with durations of tens of seconds to minutes and distinct spectral peaks.

Definition of the Subject

Infrasound may be radiated by earthquakes, tsunamis, and volcanoes through the displacement or rupture of Earth's surface and the subsequent flow and excitation of fluids. These complex and sometimes cataclysmic phenomena share some common physics, yet have different ways of converting energy into atmospheric sound. Signals from earthquakes, tsunamis, and volcanoes captured by the present generation of infrasound arrays are introduced in this chapter through case studies. Contemporary methods used in the analysis, interpretation, and modeling of these diverse signatures are discussed and some of the associated geophysical problems that remain unsolved are considered.

Introduction

The human ear may perceive sound in the frequency band of 20 to 20,000 cycles per second (Hz). Infrasound consists of acoustic waves in the atmosphere with frequencies lower than the 20 Hz hearing threshold of the human ear. Because of reduced acoustic attenuation and scattering in the atmosphere at long infrasonic wavelengths and the large spatial scales of the physical processes driving earthquakes, tsunamis, and volcanoes, the infrasound frequency band is well suited to the remote monitoring of these events. The ambient infrasound field at any location is rich and diverse, with sources originating from the solid Earth, the ocean, the atmosphere, space-born objects, and human activity [4,39]. These acoustic pressure waves co-exist with non-acoustic atmospheric pressure fluctuations associated with meteorological changes such as wind and frontal passages (e. g. [9]). Sound propagation paths are controlled primarily by the temperature and wind stratification in the lower, middle, and upper atmosphere [16,19,24]. The effective sound velocity of the atmosphere at a given height above the ground may be approximated as the sum of the scalar sound speed, which is proportional to the square root of temperature, and the vector wind velocity, which typically may reach magnitudes of 15–20% of the sound speed in the upper atmosphere. An acoustic waveguide may efficiently direct sound to ground-based stations, and is defined by a high sound velocity layer at the upper boundary and a lower sound velocity layer near the ground. The high temperature of the mesosphere and lower thermosphere (MLT) would always refract infrasound back to the ground, but severe attenuation above ~110 km can suppress thermospheric returns. Waveguides in the troposphere and stratosphere are expected to only transmit primarily along the downwind direction. However, observations suggest that the elevated acoustic waveguide formed by the low temperature zone in the stratosphere may routinely leak energy back to the ground through scattering and diffraction [27,28,29]. Ground cooling and low altitude winds may also produce a stable waveguide in the boundary layer near the ground surface [20,42]. A new generation of global atmospheric specifications designed

for the study of infrasound propagation has been developed by integrating multiple meteorological and upper atmosphere models [19]. Validation and further refinement of these atmospheric models is ongoing [3,43].

Infrasound Arrays

Background

Modern infrasound array technology emerged at the turn of the 21st century after the 1996 adoption of the Comprehensive Nuclear-Test-Ban Treaty and the subsequent growth of the International Monitoring System (IMS), which was designed for the detection of clandestine nuclear test explosions [55,61]. An infrasound array generally consists of four or more horizontally separated identical microphones or microbarometers with precisely known locations that optimize the reception of a specified wavelength range. Wind is the most pernicious source of incoherent noise, and infrasound's greatest vulnerability. An array that is sheltered from the wind by forest cover, snow cover, or topographical blocking will have a low noise floor, and provide high sensitivity and robust measurements. If a wind-sheltered site is not found, wavelength-specific wind noise reducing filters have to be designed for the boundary layer conditions at the array site [1,40]. IMS-type infrasound arrays have a maximum distance between sensors of 1–3 km and use sensors with over 100 dB of dynamic range and a flat frequency response between 0.02 and 20 Hz. Portable arrays often have apertures of ~100 m or less, and are thus optimized for smaller wavelengths (frequencies >1 Hz). Infrasonic sensors for portable array applications often operate within the 0.1–100 Hz frequency band, and may overlap into the audio band. Infrasound data are typically recorded with GPS time-stamped 24-bit digitizers and are often sent via digital communications to a central processing facility for real time analysis.

Basic Principles

A number of calibrated microphones, precisely timed and arranged in optimal spatial configurations as arrays, present the best design for recording infrasound. A wavefront is the spatial surface containing sound emitted from a source at the same time. An array relies on the principle that sound along a (possibly curved) wavefront has recognizable features, so that as the wavefront passes through the multiple sensors in an array it is possible to determine the time of passage of a specific waveform feature (for example, the peak of an explosive pulse). From the time of arrival of a waveform feature at each known sensor location, the direction of propagation of the incident wavefront as well as its apparent propagation speed across the array can be inferred. Once the arrival direction and speed are determined, it is possible to digitally apply time-delays or phase shifts to temporally align all the microphone waveforms along a beam of energy (beamform) to improve the ratio of the signal amplitude to the ambient noise. Based on these fundamental principles, more advanced contemporary techniques permit the extraction of waveform features with a very small signal to noise ratio [12,65].

Array Design

If two identical microphones spaced some distance from each other record identical waveforms, the two observed signals are perfectly coherent. As the microphone spacing increases beyond the characteristic wavelength of a pulse, the coherence decreases until the waveforms no longer resemble each other. Microphone arrays are designed to detect coherent signals within a specific wavelength range, and the ability of an array to accurately determine the speed and angle of arrival of coherent sound signals depends largely on the sensor distribution relative to the wavelengths of interest. A minimum of three sensors, deployed as an L, are required to discriminate incidence angle. However, a three element array has a very poor angular resolution and leaves no margin for error, as failure of a single sensor will substantially degrade its detection capability. Four or more sensors are preferable, yielding a broader frequency response and better measurement precision.

Detection, Location, and Signal Identification

Infrasonic signals measured a few kilometers from the source can vary from powerful explosions ($>10^2$ Pa) that dominate the recorded time series, to a background rumble (10^{-3} Pa) buried within the ambient sound field. A single infrasonic array can discriminate between a coherent signal arriving from the direction and height of the target source and a competing signal coming from a different angle and elevation [10]. Thus, arrays can 1) separate the coherent sound field from the incoherent ambient noise field, 2) identify and extract a specific infrasonic signal within the coherent ambient sound field (clutter) and 3) separate acoustic arrivals propagating through different waveguides yet originating from a signal source which may be stationary or moving. If two arrays are available, the target signal would be recorded at each array with an arrival angle pointing to the source and an arrival time consistent with the propagation path from the source to the receiver. The two array beams would intersect at the source, thus providing a geographic loca-

tion which would be confirmed and refined with the signal travel time information [25]. Thus two properly sited infrasonic arrays can unambiguously locate a source region and discriminate between sources. By optimizing detection with arrays, performing locations with two or more arrays, and characterizing preexisting sources of clutter, it is possible to acoustically recognize a source with high confidence. If other detection (such as seismic or imaging) and identification [36]) technologies are available, uncertainties in source location and identification drop substantially. This process of signal detection, location, and identification is routinely used successfully by the international community in the monitoring of natural and man-made events [27].

Acoustic Speed and Velocity

Infrasound studies may refer to the acoustic phase velocity, group velocity, effective sound speed, trace velocity, and the celerity of an acoustic arrival. The first three quantities depend on the sound speed and wind speed in the atmosphere along the source-receiver path. The trace velocity of a signal may be measured directly by observing the apparent horizontal speed and direction of propagation of a signal across an array. The celerity of an arrival is defined as the ratio of range over the travel time, and may be conceived as the effective, or average propagation speed over the complete propagation path. Given a known source-receiver configuration, the celerity may be computed directly. Although it is tempting to use the trace velocity for identifying a given arrival, measurement and calibration uncertainties can make this procedure rather inaccurate for distant sources. Celerity estimates from various infrasonic events with known locations suggests that 1) our knowledge of the atmosphere is presently insufficient to reliably predict all infrasonic arrivals to a station under all atmospheric conditions, and, 2) diffraction and/or scattering can feed acoustic energy to waveguides that are elevated above the ground, and these elevated waveguides may also leak sound back to the ground [10].

Analysis Method and Scope

Array processing methods allow us to extract coherent signals from the incoherent ambient sound field. One efficient and popular technique for estimating the infrasonic wave parameters is the *Progressive Multi-Channel Correlation* method (PMCC) [12]. This method, originally designed for seismic arrays, is well adapted for analyzing low-amplitude coherent waves within incoherent noise and efficient for differentiating signals of interest from background clutter [22,26]. The PMCC method was used for array processing of the signals from earthquakes, tsunamis, and volcanoes presented in this chapter.

Acoustic-gravity waves form a class of propagating atmospheric pressure signals which were studied in detail during the early Megaton-yield atmospheric nuclear tests (e. g. [32]). These waves are affected by buoyancy, have periods longer than 50 s, and have unique source and propagation characteristics that are beyond the scope of this paper. The IMS and portable arrays discussed in this chapter are tuned to higher frequencies, so are generally not designed to process acoustic gravity waves with high precision. The reader may refer to [33] for an excellent introduction to gravity and acoustic-gravity waves.

This chapter will also omit the acoustics of structural collapses such as pyroclastics flows from volcanoes, avalanches, landslides, or rockfalls. Although such signals may portent hazardous events (e. g. [79]), this family of acoustic signals are poorly understood and even less well modeled.

Earthquake Infrasound

Ground vibrations can produce sound in manifold ways. Relative to the atmosphere, earthquakes can act as distributed, supersonic acoustic sources moving at seismic or fault rupture velocities. When seismic surface waves travel through mountainous regions, the predominant source of infrasound is attributed to the re-radiation of pressure waves by topography [48,60,78]. In this case, the earthquake-induced displacement perpendicular to the ground surface can locally generate ground-coupled air waves. The local conversion from seismic waves to the sound pressure has been observed on microbarometers at regional and teleseismic distances [14,15,18,69]. Seismic to acoustic wave coupling at the ground-air interface will be enhanced when the horizontal phase velocity, or trace velocity, of the infrasonic waves and the seismic waves are matched. This type of seismoacoustic coupling can be particularly efficient in sediments and loosely consolidated volcanic environments with a low shear wave velocity [24].

The generation of infrasonic waves from the epicenter region has also been postulated [59,62]. At large infrasonic periods (50–100 s) acoustic-gravity waves from the sudden strong vertical ground displacements have been detected at distances of thousands kilometers from the origin [6,57]. This mechanism would also apply for large submarine earthquakes [30,58]. In all the aforementioned cases, an actual pressure wave is radiated from the ground into the atmosphere. This process is distinguished from microphonics, where the recorded signal is due to the sensor response to accelerations and ground level elevations

independently of any pressure changes occurring in the atmosphere [2,4,45]. In this case, the microphone acts as a seismometer.

The principles introduced in the discussion of the generation, propagation, measurement, and interpretation of earthquake infrasound are applied to the progressive case studies of the Arequipa earthquake of June 23, 2001, recorded at a range of ∼500 km, the Mongolia earthquake of November 14, 2001, recorded at a range of ∼1800 km, and the Chile earthquake of June 13, 2005, detected by multiple infrasonic arrays to a maximum range of 2300 km.

Case Study 1: M_w 8.4 Arequipa Earthquake Detected by a Single Station

On June 23, 2001, at 20:33:13 UTC, a strong earthquake measuring M_w 8.4 (NEIC) ripped along the coast of south-central Peru. The earthquake origin (16.15°S, 73.40°W, focal depth ∼30 km) was centered along the Peruvian coast about 600 km southeast of Lima and 110 km northwest of Camana (Fig. 1). The Pacific Tsunami Warning Center reported a moderate tsunami struck the Peruvian coast. Infrasonic waves associated with this event were detected for more than an hour at IMS infrasound station IS08 in La Paz, Bolivia (16.26°S, 68.45°W) [48].

Due to the relative proximity of the station (∼500 km from the epicenter), it was possible to perform relatively direct analyses of the apparent horizontal propagation speed (trace velocity) of the incident wavefield and the arrival angle of the different wave types at the array. The estimated trace velocity ranges from several kilometers per second to the sound velocity, as expected from the arrival of both seismic and infrasonic waves. Although the aperture of the IS08 array is designed for infrasonic wavelengths, array processing also yields the arrival characteristics of seismic waves. The azimuth variation for the seismic waves indicates that the rupture propagated from

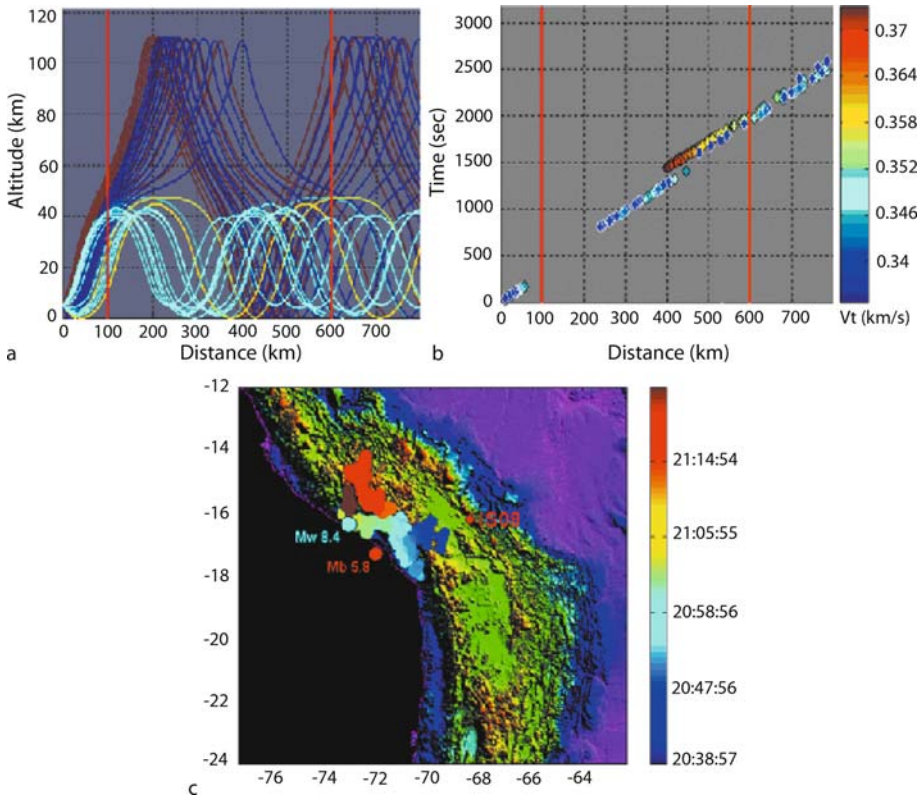

Infrasound from Earthquakes, Tsunami and Volcanoes, Figure 1

Ray traces a and travel time curves b for infrasonic waves launched almost horizontally from the epicenter area at 5 km height. The red line indicates the propagation range of the rays from the secondary sources to IS08. The color scale indicates the horizontal trace velocity of each ray. c Location of the sources of distant generation of infrasonic waves measured from 20:39 to 21:28 due to the M_w 8.4 earthquake and the m_b 5.8 aftershock (*Topography data: USGS DEM & Cornell Andes Project*). The colored dots indicate the arrival times (UTC) of the infrasonic waves at the station

the northwestern to the southeastern part of the fault at a speed of 3.3 km/s. However, the predominant source of infrasound is attributed to pressure waves radiated by the Andean Cordillera. This is consistent with the theory that the vibration of mountains can generate infrasonic waves which travel to the station along atmospheric waveguides. By performing basic inversions, the azimuth variation of the infrasonic waves can then be interpreted as a distribution of secondary sources along the highest mountain ranges. Using the azimuth and arrival time determination, the infrasonic radiation zone was estimated to be ∼100 by 400 km long.

Case Study 2: M_s 8.1 Mongolia Earthquake Detected by a Single Station

On November 14, 2001, at 09:26:10 UTC, a magnitude M_s 8.1 earthquake rattled the mountainous western Chinese region near the Qinghai-Xinjiang border. The earthquake origin (36.0°N, 90.5°E, focal depth ∼5 km) was centered along the northern margin of the Tibetan Plateau at the foot of the Kunlun Mountains where substantial surface fault ruptures have occurred before. Coherent infrasonic waves associated with this event were detected for more than one hour at a distance of 1800 km from the epicenter by IMS station I34MN in Mongolia [49].

Building on the conclusions from the Arequipa earthquake analysis, both an inverse location procedure and a complete simulation of the radiated pressure field are used to locate the distant source regions. The input parameters of the location procedure include the measured signal azimuths and arrival times as well as the origin time and coordinates of the main shock. The propagation model is based on a constant velocity of 3.3 km/s for seismic surface waves propagating from the epicenter area. The atmosphere is specified by sound velocity and wind speed profiles obtained from the time-varying MSISE-90 and HWM-93 empirical reference models [25,38], and the infrasonic wave propagation was performed using 3D ray theory [73,74].

As shown by Fig. 2, two dominant guided wave groups are predicted: (i) thermospheric arrivals refracted below 120 km for slowness ranging from 2.7 to 2.9 s/km (trace velocities of 0.345–0.37 km/s), and (ii) stratospheric ducted waves refracted below 45 km for slowness values ranging from 2.9 to 3.1 s/km (0.32–0.345 km/s). Such trapped waves can be observed when the source is located above the station [75].

The slowness distribution derived from the measured trace velocity presents a maximum between 2.85 and 2.95 s/km. These values correspond to a celerity of 0.28–0.30 km/s. The component of the wind transverse to the propagation direction deflects the rays from the original launch azimuth by $\sim 2°$. This deviation is taken into account by correcting the measured azimuths. Figure 3a reconstructs the distant source regions using a celerity of 0.29 km/s. The spatial extent of the radiating zone is estimated to be 9° in latitude and 10° in longitude. The source distributions fall into line with the Qilian range, then borders the eastern part of the Qaidam basin and join the Kunlun range. To the south of Qaidam basin, more scattered source distributions follow the Bayan Har mounts.

To verify these locations, a simulation of the radiated pressure field was performed. First, the inversion for the rupture propagation along the fault uses a slip patches model developed by [7]. Using this extended model of rupture, synthetic seismograms of surface waves are computed using a discrete wavenumber method [8,70,72] with a one-dimensional regional crust model. The source modeling displays a strong directivity, with most of the seismic energy radiated along the main strike-slip of the fault and a maximum ground velocity placed ∼300 km to the east of the epicenter [53]. To compute the acoustic radiation of the topography surrounding the fault, the topography is divided into adjacent strip-line sources radiating energy proportional to the simulated ground velocity [41].

Compared to the wavelength of the seismic surface waves (∼60 km), the area of each source element ($3 \times 3\,\mathrm{km}^2$) is small enough to consider isophase vibration. Due to the low frequencies of interest ($kL > 1$, k and L defining the acoustic wavenumber and the side of each cell, respectively), source elements radiate essentially simultaneously with a pronounced directivity. Based on this assumption, the topography is divided in adjacent strip-line sources of length L radiating energy proportional to the simulated ground velocity V_l normal to each surface element l. Considering a distance of observation R_l significantly greater than L, the Fraunhofer approximation of the Helmholtz–Huygens integral yields: $p_k(t) = iL(k\rho c)/2\pi \quad \sum_{l=1}^{N} V_l(t_l)\,\Delta h_l(\mathrm{e}^{-ikR_l})/R_l \quad [\sin(k\hat{x}_l L/2)/(k\hat{x}_l L/2)] \quad \mathrm{e}^{[-ikco(t-t_l)]}$, where t_l is the origin time of each source element, $p_k(t)$ is the predicted pressure at the arrival time $t(t = t_l + R_l/c_{\mathrm{eff}})$, $c_{\mathrm{eff}}$is the celerity in the atmosphere corresponding to the predicted wave guides (0.29 km/s), ρ is the air density, co is the sound speed, Δh_lif the height difference of the radiating surface and $\hat{x}_l = \sin(\theta_l)$is given by the angle θ_l between the outward unit normal and the source/receiver vector. Note that the sinc function in the predicted pressure reaches its peak when the surface normal and the source-receiver vector are aligned. Thus the Fraunhofer approximation provides a means of evaluating a discretized source pressure term

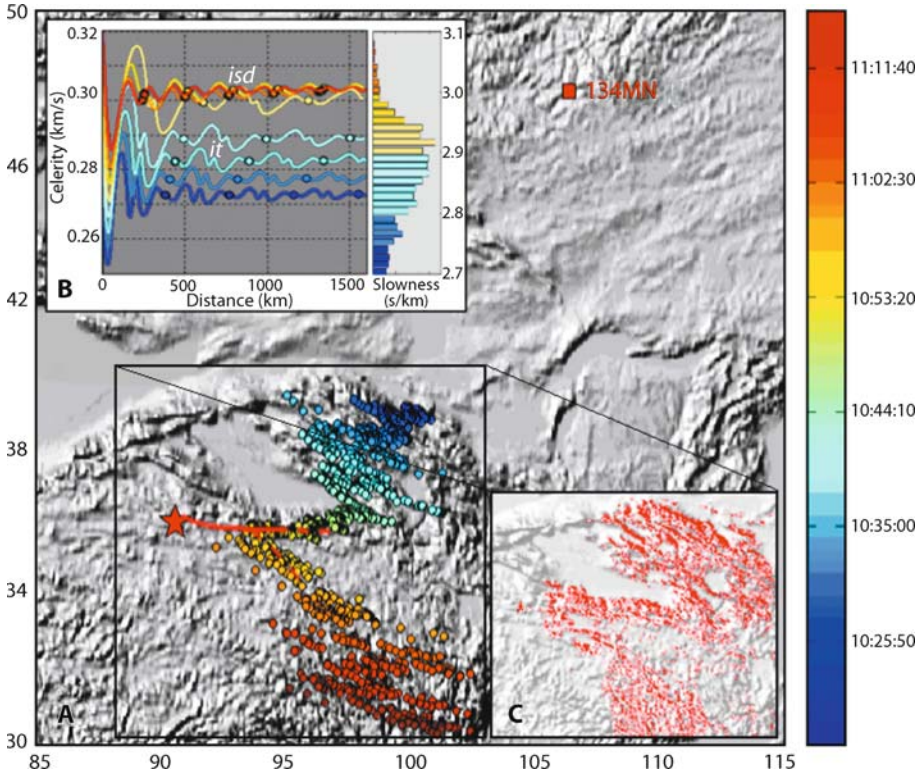

Infrasound from Earthquakes, Tsunamis and Volcanoes, Figure 2

Propagation modeling and location of distant source regions of infrasonic waves (Topography data: ETOPO30). A The colored dots indicate the sources location according the detected arrival times (UTC) of the infrasonic waves. Taking into account uncertainties due to the measurements and the propagation modeling, a maximum location error of 20 km is estimated for each dot. B Predicted celerity models versus slowness and propagation range for a source located at the main shock epicenter. The definition range of the celerity is given by the maximum of the slowness distribution derived from the measured trace velocities (Fig. 2). The circles indicate the locations of the ground reception of both ducted stratospheric (*Isd* arrivals) and thermospheric paths (*It* arrivals). C Normalized surface pressure distribution along the Kunlun fault

for radiating strip lines of topography, where each strip will yield its largest pressure contribution when its displacement is along the source-receiver direction.

Using the simulated ground velocity and this approximation of the integral formulation, the predicted surface pressure distribution matches reasonably well with those obtained with the inverse location procedure (Fig. 2). Thus, the azimuth variations and the expansion of the signal duration suggest that the Kunlun Mountains acted as sources of infrasonic waves over a radiating zone of $\sim 1000 \times 1000$ km. The maximum seismo-acoustic coupling is found to the east of the main shock epicenter, which is consistent with the seismic radiation pattern.

Case Study 3: M 7.8 Chile Earthquake Detected by Multiple Stations

On June 13, 2005, a major earthquake occurred in the mountainous section of the Tarapaca Province (North Chile) (19.93°S–69.03°W at 22:44:33 UTC, M7.8, focal depth 117 km, USGS). The epicenter was located deep under the Andes mountain range, near Chile's border with Bolivia. At large distances from the epicenter, coherent infrasonic waves were detected by IMS infrasound stations I08O-Bolivia, I09BR-Brazilia, and I41PA-Paraguay (410 km, 2300 km, and 1420 km from the epicenter, respectively). The multiple station recordings at different ranges and azimuths from the epicenter allowed a more complete reconstruction of the infrasound source regions compared to what was obtained using one single station [52].

To invert for the main source regions of infrasound, the location procedure requires the signal azimuths and arrival times measured independently by each station and the origin time and coordinates of the epicenter. A joint inversion for the source area using data from all three stations simultaneously was not used due to the pronounced directivity of the radiation pattern. Infrasound

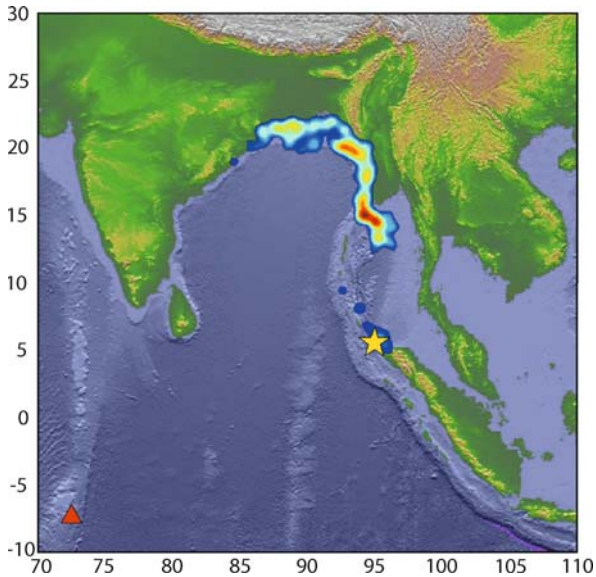

Infrasound from Earthquakes, Tsunamis and Volcanoes, Figure 3
Location distribution of distant source regions of infrasonic waves generated by the December 26, 2004 Sumatra earthquake. Color scales are normalized. The red triangle indicates the location of the IS52 station. Source locations are computed for seismic surface waves and tsunami waves originating from the maximum of coseismic slip

propagation was simulated using a spherical coordinate 3D ray theory formulation which accounts for topography and the spatio-temporal variations of the horizontal wind terms along the ray paths [17]. Atmospheric absorption is integrated using altitude dependent attenuation coefficients [66,67]. The atmospheric conditions of June 13, 2005 are described by the sound velocity and wind speed profiles provided by the time-varying Ground to Space (G2S) atmospheric specifications [19].

As in the previous case study, the slip patches model developed by [72] is used to check the association of regions radiating infrasound with areas of strong ground motion. The simplest extended elliptic source model able to explain the teleseismic seismograms is found, and the first and second order characteristics of the event (location, depth, duration, focal mechanism, and refined kinematic parameters such as spatial slip distribution on the fault and rupture velocity) are calculated from teleseismic body waves. Then, from the resulting extended source model, low frequency synthetic seismograms (period lower than 10 s) are computed on a grid in the vicinity of the epicenter using the discrete wavenumber method and a one-dimensional regional crust model [8]). Finally, the root mean square of the maximum velocity of the vertical and horizontal components of the surface waves is used to reconstruct the areas of strong ground motion.

The reconstructed source regions confirm that most of the energy is radiated by the vibration of land masses near the epicenter, which is consistent with the predicted areas of strong ground motion. No clear signal originates from the Altiplano. Southern high mountain ranges, even far from the epicenter, also generated infrasound. The Central Cordillera extending to altitudes greater than 5000 m efficiently produced infrasound in the direction of I09BR, although the predicted seismic movement is low in this region. Consistent with previous observations, these results suggest an amplification of the ground displacement caused by the topography surrounding the Altiplano. Such site effect could not be predicted from our seismic source modeling since the topography is not considered. The reconstructed source regions extend over ~800 km from the Central Cordillera to the Occidental Cordillera. The spatial extent of the radiating zones differs from one station to another, which confirms the influence of shadow zone effects for nearby stations and the directivity of the radiation. As in the previous case study, the topography is modeled as a succession of adjacent strip-line sources, so that mountain ranges radiate energy essentially simultaneously with a pronounced directivity and may generate infrasound arrivals with different azimuths. This suggests that the amount of energy radiated in the direction of the receiver and the duration of the signals also depends on the orientation of the highest mountain ranges around the station.

Summary of Earthquake Infrasound

The three case studies presented in this section build in complexity and sophistication while producing consistent results. High mountain chains rattled by large earthquakes reliably radiate infrasound and have acoustic radiation patterns that depend on the chain's orientation. Substantial contributions to the sound field are expected from steep topography, which would primarily radiate perpendicular to exposed faces. For large earthquakes occurring in mountainous regions, infrasonic measurements are valuable for the analysis of the remote effects of earthquakes and site effects over broad areas. In remote regions where there is a lack of surface motion instrumentation, infrasonic observations could lead to a rapid determination of the regions where the seismic movements are the largest.

Tsunami Infrasound

Previous study has shown that significant infrasound is produced by breaking waves (surf) and the complex interaction of open-ocean swells (microbaroms, e. g. [76]).

However, interest in tsunami infrasound emerged when IMS infrasound arrays in the Pacific and Indian Oceans recorded distinct signatures associated with the December 26, 2004 Great Sumatra-Andaman earthquake and tsunami. As in the case of mountains stirred by continental earthquakes, islands which undergo significant surface displacements during submarine earthquakes can also produce infrasound. It also appears that the initiation and propagation of a tsunami may produce low frequency sound near the epicenter as well as along coastlines and basins. In some environments, precursory sound could potentially be used for confirmation and early warning for tsunamis. This field of research is still in its infancy, and our interpretations leave much room for further development. Substantial complexity is involved in the separation of the earthquake-generated infrasound from the sound that may be produced by the genesis and propagation of the tsunami.

Sumatra Earthquake and Tsunami

The magnitude 9.1 Great Sumatra-Andaman earthquake of December 26, 2004 [63] is the largest earthquake since the 1964 magnitude 9.2 Great Alaska earthquake, and produced the deadliest tsunami in recorded history. In contrast to the Great Alaska earthquake, the Great Sumatra earthquake and tsunami were recorded globally by multiple digital sensor networks in the ground, ocean, atmosphere, and space. Further, the resulting signals were analyzed and interpreted using far more advanced computational capabilities than were available 40 years ago. Although the study of tsunami infrasound rose in the wake of the Sumatra-Andaman event, many fundamental questions on the generation of these deep sounds remain unanswered.

The clearest infrasonic signatures associated with the Sumatra event were captured by the station in Diego Garcia (IS52GB, Fig. 3), which recorded (1) seismic arrivals from the earthquake, (2) tertiary arrivals (T-phases) that propagated along sound channels in the ocean and coupled back into the ground, (3) infrasonic arrivals associated with either the tsunami generation mechanism near the seismic source or the motion of the ground above sea level, and (4) deep infrasound (with a dominant frequency lower than 0.06 Hz) coinciding with the propagation of the tsunami into the Bay of Bengal [30,51]. These signals were all recorded by the pressure sensors in the arrays. The seismic and T-phase recordings are a result of the sensitivity of the microphones to ground vibration (microphonics), whereas the infrasound arrivals correspond to sound propagating through atmospheric waveguides. Similar, but not identical arrivals were observed at Diego Garcia (range of ∼2900 km) during the March 28, 2005 Nias earthquake (M8.7, which produced a non-destructive tsunami). Yet only infrasonic arrivals were observed from the April 10, 2005 Mentawai earthquakes (M6.7 and 6.5, no reported tsunami), indicating that above-water ground motion from submarine earthquakes can produce sound in the Sumatra region. The deep infrasound from the Bay of Bengal region following the Sumatra earthquake suggests that sound can be produced by the interaction of a tsunami with coastal bathymetry.

To reconstruct the main source regions of infrasound recorded at IS52 for this event, the input parameters of the location procedure included the measured signal azimuths and arrival times, and the origin time and coordinates of the epicenter. As described in the previous section, standard velocity models were used to describe the propagation of the seismic surface waves and the propagation of infrasound through the atmosphere in the direction of IS52. However, the speed of propagation of the tsunami from the epicenter needs to be added to the inversion procedure. For the December 26, 2004 earthquake, a velocity of 3.3 km/s is used for the seismic surface waves, and a speed related to the square root of the water depth for the tsunami waves.

Using the infrasonic arrivals to IMS station in Diego Garcia and Palau, the location inversion for the high frequency infrasonic component corresponded to sound originating from the mountains and possibly from the epicenter. However, a unique characteristic of the Sumatra earthquake is that it generated large amplitude coherent waves with a dominant period of ∼30 s over four hours. The infrasonic source locations derived from the inversion procedure indicate that the tsunami also created infrasonic waves at lower frequencies when it propagated in shallower water as it reached the Bay of Bengal (Fig. 3). Even lower-frequency acoustic gravity waves, with periods of hundreds of seconds, may have been produced by the underwater ground displacement [58]. For wavelengths which are much greater than the water depth, the ocean surface displacement nearly matches the submarine ground displacement above the fault plane (e. g. [44]) and may efficiently radiate long-period pressure waves into the atmosphere.

Miyagi-Oki Earthquake and Tsunami

The events off the coast of Sumatra were ∼3000 km from the closest infrasound station in Diego Garcia. The longer ranges, coupled with the fact that all infrasound stations used in those studies were transverse to the axis of Suma-

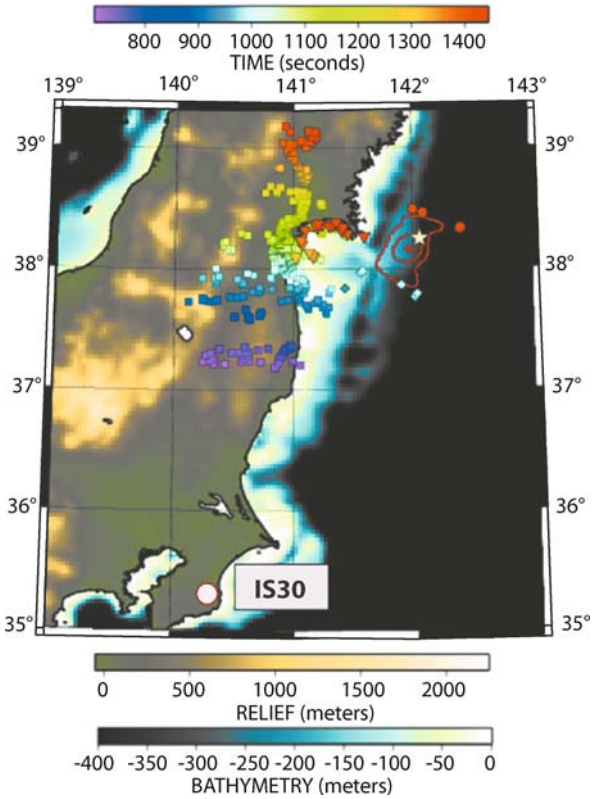

Infrasound from Earthquakes, Tsunamis and Volcanoes, Figure 4
Estimated infrasonic source locations associated with ground vibration, tsunami genesis, and the interaction of the tsunami with the coastline. The squares represent stratospheric arrivals with a celerity of 0.3 km/s. The diamonds are also stratospheric arrivals but with the celerity of 0.32 km/s predicted for that azimuth. The circles are thermospheric arrivals with a celerity of 0.27 km/s. The triangles are stratospheric arrivals with a celerity of 0.3 km/s, but with a delay time attributed to seismic seiche formation. The color of the symbols indicates the arrival time in seconds since the earthquake's origin time. The topography is from NOAA ETOPO2 data

tra, caused uncertainty in the ability to discriminate between sounds potentially produced during tsunami genesis at the ocean surface and the sounds produced by the earthquake-induced vibration of mountains and islands. In contrast, IMS infrasound station IS30 in Japan (Fig. 4) is optimally situated to recognize the different source regions of infrasound associated with the Miyagi-Oki earthquake and tsunami.

The magnitude 7.2 Miyaki-Oki earthquake occurred on August 16th, 2005 at 02:46:28 UTC. The epicenter was off the coast of Japan near Honshu (38.251°N, 142.059°E), with an estimated depth of 36 km. The earthquake caused landslides, ∼60 injuries, as well as power and transportation disruptions. A local, nondestructive tsunami was observed on the coast of northern Japan with a wave height of ∼10 cm. Because of its auspicious location, station IS30 can use the angle of arrival information derived from array processing to identify infrasonic arrivals originating from mountain chains, the earthquake epicenter (which is used as the tsunami epicenter), and the coastline of the Bay of Sendai.

In contrast with the Sumatra event, where substantial energy was observed in the deep infrasound band (0.002–0.1 Hz), most of the infrasonic energy for the smaller Miyagi-Oki event was above 0.5 Hz. The arrival azimuths at IS30 range from −5° to 28°, and do not have a well-defined temporal sequence, suggesting multiple sources, propagation paths, and possible wave types.

Using the ground to space (G2S) atmospheric profiles [19] specific to the station location and time of the event, source locations [30,51] were estimated for the infrasonic arrivals shown in Fig. 4. The first infrasonic arrivals would correspond to acoustic waves coupled to the atmosphere from the seismic vibration of land masses [49]. For ranges less than 330 km and northerly arrivals between −5° and 18° (measured clockwise from north), the time- and site-specific G2S specifications do not support thermospheric arrivals at the station, so the observed arrivals would propagate in stratospheric waveguides with a celerity of 0.3 km/s. In contrast, for arrival azimuths of ∼ 18° to 28° originating offshore, thermospheric arrivals are also supported for ranges greater than 330 km, although the celerity of stratospheric arrivals is faster.

Multiple wave propagation paths are invoked to produce the temporal and frequency distribution of the infrasonic signals observed at IS30. The squares shown in Fig. 4 correspond to seismic vibrations radiating acoustic energy into stratospheric waveguides, thereafter propagating with a celerity of 0.3 km/s, the diamonds to stratospheric arrivals with a celerity of 0.32 km/s, and the circles correspond to thermospheric arrivals with a celerity of 0.27 km/s. Inclusion of seismic speeds made a negligible difference, as it offsets the locations by <6 km. As expected, many of the arrivals originate from land, and are consistent with the predicted seismic intensity [13,71]. Of more relevance to this paper, clear stratospheric and thermospheric arrivals arrive from the perimeter of the epicenter, which is assumed to be the region of tsunami genesis. These results are consistent with the proposed infrasonic source locations obtained by [30] for the Sumatra earthquakes, and support the idea that ocean surface displacements associated with submarine earthquakes produce infrasound [58]. The possibility that vertical ocean displacements near the epicenter may radiate sound suggests that infrasound signals may be potential discriminants for tsunami genesis.

The later arrivals (shown as triangles in Fig. 4) originate from the mountain regions to the north of the Bay of Sendai when stratospheric and thermospheric arrivals are assumed. Yet, analyses of the Great Sumatra earthquake suggest that the Bay of Bengal produced deep infrasound, and the possibility that the Bay of Sendai may also act as an acoustic source is considered. The triangles in Fig. 3 contour the coastline of the northern Bay of Sendai when a combination of a delay time and stratospheric apparent propagation speeds are considered. For a celerity of 0.3 km/s, the inferred delay time is ∼250 s. This would require an acoustic generation process that requires ∼4 min to be established over a costal region ∼100 km long. The earthquake occurred in a shallow region, which corresponds to slow tsunami propagation speeds of ∼0.05 km/s. This would mean the tsunami would have traveled at most 12 km in 250 s, which is not enough time for the tsunami to reach the coastline and produce these signals. Thus the interaction of the small tsunami with the coastline is eliminated as a possible source.

A plausible ensonification mechanism is the local production of coastal waves by the earthquake, as in the generation of seismic seiches [5,56]. Low frequency radiation from enclosed bodies of water have been proposed by other authors [9,46]. For the Miyaki-Oki earthquake, substantial water displacement in shallow regions of the bay is strongly suggested by acoustic sources above the water near the coastline (blue and green arrivals in Fig. 4). Satellite imagery shows that the northern part of the Bay of Sendai has an abundance of lakes, bays, man-made harbor structures and rivers which may sustain seiches. Lower order seiche modes would take minutes to be established, but higher-order modes and coupled oscillations associated with the narrower of the volume dimension could sustain an ensonification process akin to microbarom generation from the ocean (e. g. [76]. Alternatively, the triangular symbols in Fig. 4 would should be shifted ∼75 km North of their shown location.

Summary of Tsunami Infrasound

These two case studies strongly suggest that submarine earthquakes and the water level changes they induce can produce low-frequency sound. The sound may be radiated from the ocean surface during the tsunami genesis, produced by the vibration of land masses near the epicenter, or be excited by the interaction of seismic and water waves with the coastline, shallow bathymetry, and harbors.

There is some potential for using infrasound in conjunction with other technologies for remote tsunami monitoring. The effective propagation speeds of tsunami (∼50–200 m/s) and sound waves (∼300 m/s) yield an advance warning time of at least 1.7 s/km. At 100 km, sound leads the tsunami by at least 170 s, but some of this time would be taken up by signal transmission, processing and identification, leaving less than one minute to issue an alert. However, infrasound may provide early warning in areas within shallow basins or further than a few hundred kilometers from the tsunami source region.

Volcano Infrasound

The previous sections discussed how earthquake infrasound acoustically couples solids and gases, and tsunami infrasound couples solids, liquids, and gases. Similarly, a volcano can transmit information about its current eruptive state through vibrations induced in the ground, volcanic fluids, and atmosphere. However, there is a great deal of ambiguity about the composition of the acoustically active volcanic fluids, which may vary from vesiculated magma, through gas-ash mixtures, to steam. Although a sealed volcanic conduit can produce faint sounds through earthquakes, once the volcanic plumbing breaches the surface it can broadcast infrasound quite unambiguously.

Basic Principles

The birth of volcano infrasound research may be traced to the cataclysmic 1883 eruption of Krakatoa, which triggered a series of interdisciplinary, international geophysical studies of the pressure signals produced by the volcano [77]. Barometric records observed throughout the US, Europe, and Russia, and reports of cannon-like sounds in surrounding islands (as far as Diego Garcia and Rodrigues Islands) demonstrated for the first time the ability of low frequency sound to propagate for thousands of kilometers.

Infrasonic signals originating inside magma conduits contain information about the pressure fluctuations driving eruption processes. These signals may be broadly categorized as explosions, long-period events, and tremor. Very large eruptions can also produce acoustic gravity waves [34,68]. Explosions are impulsive and have durations of seconds, long period events are more emergent than explosions, have distinct spectral peaks, and may last seconds to minutes, and tremor signals are sustained atmospheric vibrations that can persist from minutes to years. Decades ago, volcano seismology identified tremor and long-period events as signals indicative of near-surface intrusion of volcanic fluids (e. g. [11]). In contrast, the first infrasonic measurements of tremor from Sakurajima

were published in [64]. As the sensitivity of instrumentation and sophistication of analysis methods increased with advancing technology, the catalogue of known volcanic sounds has expanded and the ability to remotely detect hazardous eruptions has extended. Ongoing monitoring efforts suggest that relatively gentle Hawaiian and mild Strombolian activity may be consistently observed from a range of ∼10 km [20] and strong Strombolian, Vulcanian and Plinian eruptions from distances of tens to hundreds of kilometers [31,50,54,68].

The difficulty with modeling and interpreting volcanic sounds is that there is a lot of ambiguity in the geometry, composition, thermodynamics, and phase (solid, liquid, gas, or a mixture) of the volcanic interior preceding and during an eruption. In addition, there are many possible ways of exciting volcanic fluids into oscillation through unsteady and transient flow. As examples of these complications, some possible ways of exciting tremor signals in a magma conduit are described. Tremor signals are endemic to volcanoes, and are generally attributed to magma intrusion. Some plausible driving mechanisms of tremor signals and the physical properties of the materials inside volcanic pipes are discussed. In the following example, it is postulated that the harmonic spectrum of tremor events is caused by the acoustic resonance of the fluid within the magma conduit of Arenal volcano, Costa Rica [23]. The possible relationships between observed signals and the gas content of the melt, the physical conditions inside the volcano, and the flow dynamics of the volcanic fluids are discussed.

Case Study: Tremor Signal at Arenal Volcano

This section hopes to provide an overview of the complexities of modeling volcanic sounds by considering the frequency changes in a harmonic tremor signal recorded at Arenal volcano, Costa Rica. Figure 5 shows a section of the infrasonic tremor signal shown in [23]. This signal is illustrative of most of the tremor recorded during April–May 1997 by a three-element array of infrasonic sensors and a five-element array of seismometers [35].

The prominent features of the Arenal tremor are the harmonic character of the spectra and the oscillation of the spectral bands with time. This oscillation of the spectral bands about an equilibrium value is referred to as gliding. Since steady flow cannot generate acoustic waves, it must be the pressure, volume, and flow fluctuations of the magma injection process that are driving the infrasonic signals. It is possible that the gliding of the spectral bands is caused by time-varying changes in the flow regime or the physical properties of the material in the conduit. The resonant magma conduit model of [21] is used to explore the ramifications of having changes in the conduit length, flow velocity of a gas and in the void fraction of a magma-gas mixture.

The tremor signal shown in Fig. 5 lasts over ten minutes and is not a transient event. Attempts to model the spectra with a random continuous source, as successfully done for Pavlof volcano [24], failed at Arenal because of the sharpness of the spectral peaks and the rapid rolloff at higher frequencies. The tremor signal appears to be sustained by a source mechanism that is triggering repeatedly in time. The source region may consist of a compliant gas volume within a magma-filled conduit. A bubble may form around a corner or in a constriction in a magma-filled conduit, where fluid acceleration may generate relatively stable gas-rich pockets. These cavities will be sensitive to flow fluctuations, and may act as effective acoustic sources [21].

Lava flows at Arenal and incandescent pieces ejected during explosions strongly suggests an open magma conduit associated with the infrasonic recordings. However, there is some ambiguity on whether the acoustically active part of the conduit is filled with a magma-gas or an ash-gas mixture. The source region is placed at the lower part of the magma conduit, and it is assumed that it ensonifies the magma-gas mixture above it. Although for the Arenal signal there is evidence for a buried tremor source [35], it may also be possible to have efficient atmospheric excitation from jet flow above the vent during more powerful eruptions [31]. For the purposes of our discussion, a two-layer, open vent solution representing a slow-velocity, gas rich magma floating atop a less vesiculated layer is used, as developed in Sect. 4 of [21]. The source region at the bottom of the conduit could correspond to a constriction in the conduit, where cavitation may occur due to fluid acceleration.

The sound speed of a liquid gas mixture is a strong function of the amount of bubbles, or the void fraction, in the mixture. For high void fractions, the sound speed of a magma-gas mixture may reach sound speeds of tens of meters/second, and sound would be heavily attenuated. For a resonant fluid column of length L, $f_n = n(1 - M_2)c/(2L)$, where f_n is the resonant frequency of the nth spectral peak, M is the Mach number, or ratio of the flow speed to the sound speed, and c is the sound speed of the fluid.

Assuming a gas-rich magma inside a resonant magma conduit, the changes in the spectral peaks of the tremor signal may be attributed to changes in the amount of exsolved gas in the melt (Fig. 5), which would dramatically change the sound speed c. The tremor signal would cease

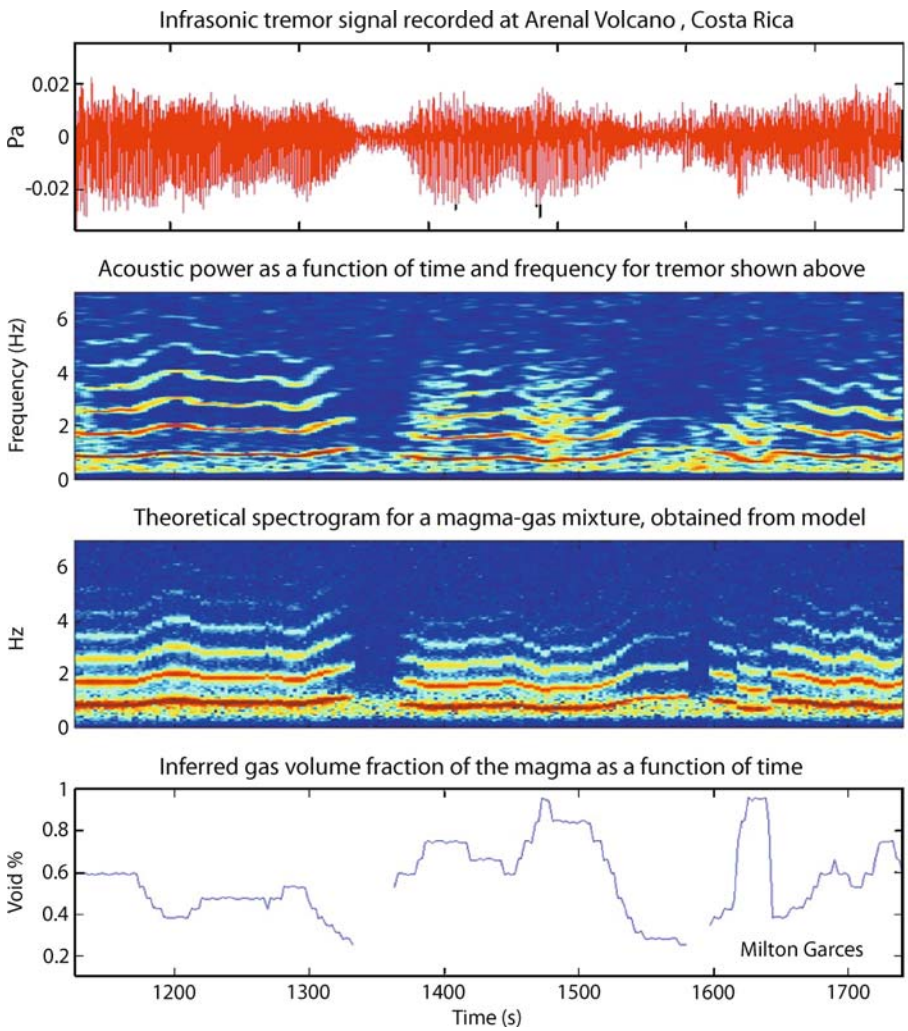

Infrasound from Earthquakes, Tsunamis and Volcanoes, Figure 5
Recorded acoustic signal, spectrogram, synthetic spectrogram, and one possible interpretation for a harmonic tremor signal observed at Arenal Volcano, Costa Rica

if the mass flux stops or becomes steady. It is unlikely that there is only gas inside the conduit because the acoustic attenuation in the gas would not be sufficient to explain the rapid amplitude decay with frequency. However, it is possible to have a gas-ash mixture, in which case it may be plausible to also explain the gliding with a change in the Mach number of the flow or a variation in the length of the resonant portion of the conduit [21].

Attenuation coefficients for dusty gases introduce an additional number of poorly characterized physical parameters, adding further ambiguity to our interpretations. It is due to these and many other complexities in modeling volcanic systems that the science of volcanology has evolved towards the synergy of geophysical observations and geological parameters [37]. Some of the primary goals of volcanic studies are to permit reliable eruption forecasting, hazard assessment, and early warning. Although it is difficult to forecast an eruption without a clear understanding of the physical processes associated with precursory activity, it is possible to rapidly identify a powerful hazardous eruption and provide early warning.

Prototype Operational System: ASHE

The Acoustic Surveillance for Hazardous Eruptions (ASHE) proof-of-concept project seeks to develop and evaluate the potential for robust, operational infrasonic remote sensing of volcanic eruptions. In contrast to

other ground-based volcano surveillance systems, the autonomous ASHE arrays are sited tens to hundreds of kilometers away from the devastation zones of erupting volcanoes. The diverse eruption signals produced by Tungurahua Volcano and captured by the ASHE arrays illustrate how acoustic remote sensing may complement seismic observations and satellite remote sensing to improve continuous monitoring of wide regions of potential eruption hazard.

The prototype ASHE stations consist of a four-element infrasound array with an aperture of ~100 m, a broadband seismic sensor, and a wind sensor [54]. A satellite dish sends the digital data in real time from the field to the Geological Survey of Canada data center in Ottawa, Canada, which distributes the data in real-time to other collaborating parties. In January of 2006 the ASHE team deployed two infrasound arrays in Ecuador [31]. These two arrays, separated by ~250 km, were sited to detect volcanic eruption in Ecuador or Southern Colombia by one of the arrays within 15 min of the eruption, and as early within 5 min for Tungurahua and Sangai volcanoes, which were within 40 km of an array.

Three distinct types of eruption signatures can be identified at Tungurahua (Fig. 6). The first and most common is low ash-producing background tremor with a dominant frequency of 1.4 Hz (Fig. 6a). During mid-May 2006, volcanic activity changed and was temporarily characterized by large explosions followed by harmonic tremor (Fig. 6b), reminiscent of the aforementioned tremor at Arenal. However, no significant ash release was observed, and this second type of eruptive regime was characterized as ash-poor. On July 14th, 2006 a large Vulcanian to sub-Plinian eruption (Fig. 6c) produced dangerous pyroclastic flows, substantial stratospheric ash clouds, and a significant increase in tremor. Between August 16th–17th, 2006 a larger Vulcanian to Plinian eruption occurred (Fig. 6d) and was characterized by lethal pyroclastic flows, a larger stratospheric ash cloud, and considerable infrasonic tremor. As the eruption progressed, the majority of acoustic energy shifted to lower frequencies (<0.5 Hz). The acoustic signatures of the two major eruptions are comparable and easily identifiable in infrasonic records due to the vast amount of energy present over a broader band of frequencies. These two eruptions injected substantial ash

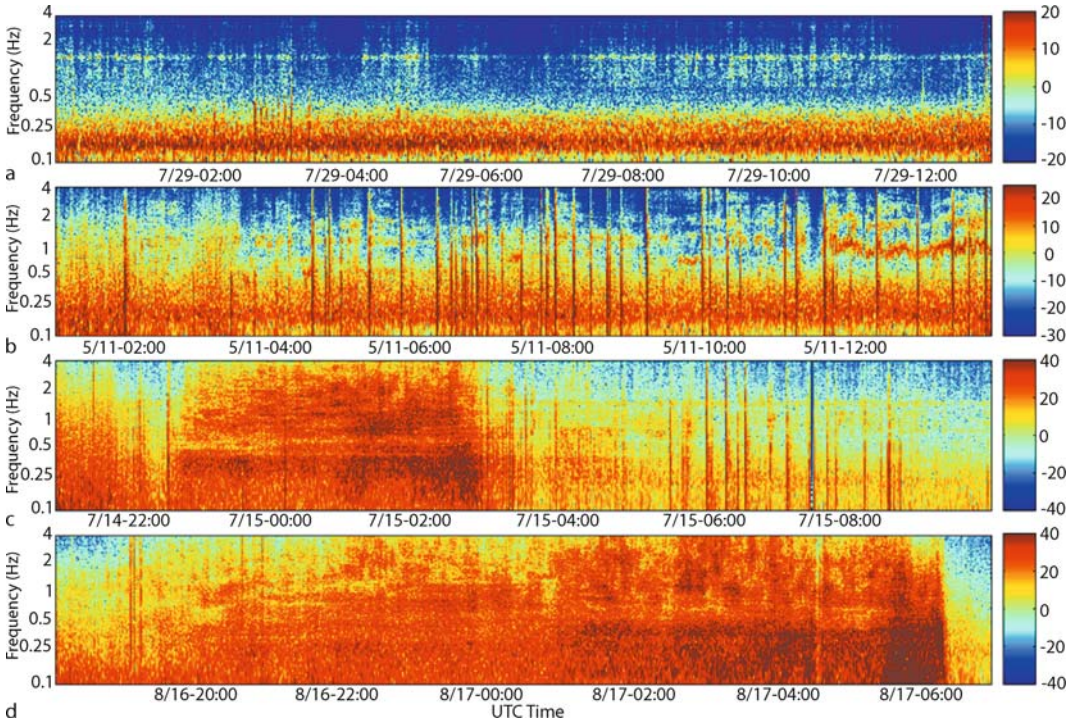

Infrasound from Earthquakes, Tsunamis and Volcanoes, Figure 6

Spectrograms for different eruption styles at Tungurahua volcano. The horizontal axis represents 13 hours of data, and the vertical scale shows frequency in a logarithmic scale from 0.1 to 4 Hz. The color denotes acoustic power in decibels referenced to 1 Pa/Hz. The ocean microbarom shows up in the upper panel two panels as a red band centered about 0.2 Hz, and is not of volcanic origin. Explosive events appear as intense vertical bands

into the stratosphere, suggesting this type of acoustic signature may be used for remote infrasonic monitoring of hazardous eruptions.

Automatic analysis techniques for eruption detection prototyped by the ASHE project produce automated low-latency email notifications coupled with more detailed real-time data products which can be used by responsible operational agencies to disseminate updated information. Based on the acoustic records captured during the Tungurahua eruptions of July and August 2006, source parameters that may be estimated during large eruptions include (but may not be limited to) the start time and duration of an ash cloud injection that could pose a hazard to international aircraft at cruising altitudes.

Summary of Volcano Infrasound

Volcanic signals incorporate many of the complexities found in earthquake and tsunami signals, and than add some because of the inherently unstable thermodynamic and hydrodynamic conditions that lead to and accompany eruptions. Because infrasound measures the excess pressure change that drives an eruption, it can be used to help unravel the dynamic physical processes driving eruptive activity. Although infrasound may be used with additional technologies in forecasting and hazard assessment, its clearest contribution is in the identification of the intensity and timing of a powerful eruption. The technology and methodology to use infrasound for acoustic remote sensing of hazardous eruptions is mature and ready to implement into operations.

Future Directions

Although infrasound has manifold applications in near-source studies, recent advances in sensor and analysis techniques have strengthened the potential to develop infrasound as a robust remote sensing tool that may cover observational gaps in distant and inhospitable environments. Although operational acoustic monitoring systems that can help notify aircraft of hazardous eruptions have been demonstrated as viable, much work is still needed to implement these systems and establish clear relationships between ash heights, infrasonic source parameters, and atmospheric conditions. Many parts of Earth's oceans remain undersampled, and there is a great potential for the application of infrasound to the detection of tsunami-genesis, the identification and tracking of severe ocean weather, and the monitoring of breaking ocean wave intensity and other coastal hazards. Much fundamental research lies ahead in the construction of functional models for these various source processes. One of the most ambitious projects the community has initiated is the passive acoustic tomography of the atmosphere using natural sources [50]. These studies, refined and extended over the next decades, may contribute to our long-term assessment of global climate change.

Concluding Remarks

There is beauty in complexity, as there is perplexity in beauty. The oscillatory interaction between the solid and fluid phases of our Earth reveals a wealth of information that can be turned into knowledge by the meticulous and persistent use of observation and modeling. Infrasound captures this solid-fluid interaction at the temporal scales of tens of seconds to tens of hertz, offering alluring glimpses into the inner workings of the natural world. The heaving ground and the frothing magmas, borne of the dark depths of Earth, oft conspire with the surging oceans to pervade the skies with deep sound. The field of infrasound strives to interpret the imperceptible yet omnipresent soundscape that permeates the air about us, so as to assuage uncertainty and perchance alleviate our dread of the natural threats posed by earthquakes, tsunamis, and volcanoes.

Acknowledgments

The authors wish to thank C. Hetzer, P. Caron, and S. McNamara for figures and analyses. Garces also extends his appreciation to M. Protti, M. Haggerty, and S. Schwartz for their contributions to the Arenal studies. D. Fee and R. Matoza provided very useful revisions to this chapter, for which we are grateful. Garces' contributions to this chapter were supported in part by the National Science Foundation (grant EAR-0609669).

Bibliography

Primary Literature

1. Alcoverro B, Le Pichon A (2005) Design and optimization of a noise reduction system for infrasonic measurements using elements with low acoustic impedance. J Acoust Soc Am 117:1717–27. doi:10.1121/1.1804966
2. Alcoverro B, Martysevich P, Starovoit Y (2005) Mechanical sensitivity of microbarometers MB2000 (DASE, France) and Chaparral 5 (USA) to vertical and horizontal ground motion. Inframatics 9:1–10 http://www.inframatics.org/. Accessed Mar 2005
3. Bhattacharyya, Bass JH, Drob D, Whitaker R, Revelle D, Sandoval T, Woodward R (2003) Description and Analysis of Infrasound and Seismic Signals Recorded from the Watusi High-explosive Experiment of September 28, 2002. SAIC techical report SAIC-03/2206
4. Bedard AJ (1971) Seismic response of infrasonic microphones. J Res Nat Bur Stand 75:41–45

5. Berninghausen WH (1969) Tsunamis and seismic seiches of Southeast Asia. BSSA 59:289–297
6. Bolt BA (1964) Seismic air waves from the great 1964 Alaskan earthquake. Nature 202:1095–1096
7. Bouchon M (1976) Teleseismic body wave radiation from a seismic source in a layered medium. Geophys J Int 47(3):515–530. doi:10.1111/j.1365-246X.1976.tb07099.x
8. Bouchon M (1981) A simple method to calculate Greens functions for elastic layered media. Bull Seism Soc Am 71:959–971
9. Bowman HS, Bedard AJ (1971) Observations of Infrasound and Subsonic Disturbances Related to Severe Weather. Geophys J Int 26(1–4):215–242. doi:10.1111/j.1365-246X.1971.tb03396.x
10. Brown D, Garcés M (2009) Ray Tracing in an Inhomogeneous Atmosphere with Winds, Handbook on Signal Processing in Acoustics. Springer (in press)
11. Chouet B (2003) Volcano Seismology. Pageoph 160:739–788
12. Cansi Y (1995) An automatic seismic event processing for detection and location: the PMCC method. Geophys Res Lett 22:1021–1024
13. Che I, Lee H, Jeon J, Kang T (2007) An analysis of the infrasound signal from the Miyagi-Oki earthquake in Japan on 16 August 2005. Earth Planets Space 59:e9–e12
14. Cook RK (1971) Infrasound radiated during the Montana earthquake of 1959 August 18. Geophys J R Astr Soc 26:191–198
15. Cook RK, Young JM (1962) Strange sounds in the atmosphere, Part II. Sound 1:25–33
16. Cox EF, Plagge HJ, Reed JW (1954) Meteorology Directs Where Blast Will Strike. Bull Am Meteorol Soc 35:95–103
17. Dessa JX, Virieux J, Lambotte S (2005) Infrasound modeling in a spherical heterogeneous atmosphere. Geophys Res Lett 32:L12808.1–5 doi:10.1029/2005GL022867
18. Donn WL, Posmentier ES (1964) Ground-coupled air waves from the great Alaskan earthquake. J Geophys Res 69:5357–5361
19. Drob DP, Picone MJ, Garces M (2003) Global morphology of infrasound partitioning. J Geophys Res 108(D21):4680. doi:10.1029/2002JD003307
20. Fee D, Garcés M (2007) Infrasonic tremor in the diffraction zone. Geophys Res Lett 34:L16826. doi:10.1029/2007GL030616
21. Garcés MA (2000) Theory of acoustic propagation in a multiphase stratified liquid flowing within an elastic-walled conduit of varying cross-sectional area. J Volcanol Geotherm Res 101:1–17
22. Garces M, Hetzer C (2003) Optimizing the Progressive Multi-Channel Correlation Detector for the Discrimination of Infrasonic Sources. In: Proceedings of the 25th Seismic Research Review, Tucson, 23–25 Sept 2003
23. Garcés MA, Hagerty MT, Schwartz SY (1998) Magma acoustics and time-varying melt properties at Arenal Volcano, Costa Rica. Geophys Res Lett 25:2293–2296
24. Garcés M, Hansen RA, Lindquist KG (1998) Traveltimes for infrasonic waves propagating in a stratified atmosphere. Geophys J Int 135:255–263
25. Garces M, Hetzer C, Lindquist K, Drob D (2002) Source Location Algorithm for Infrasonic Monitoring. 24th Annual DTRA/NNSA Seismic Research Review, Ponte Vedra, 17–19 Sept 2002
26. Garces M, Harris A, Hetzer C, Johnson J, Rowland S, Marchetti E, Okubo P (2003) Infrasonic tremor observed at Kilauea Volcano, Hawaii. Geophys Res Lett 30:2023–2027 2003
27. Garces M et al (2004) Forensic studies of infrasound from massive hypersonic sources. EOS 85(43):433
28. Garcés M, Willis M, Hetzer C, Le Pichon A, Drob D (2004) On using ocean swells for continuous infrasonic measurements of winds and temperature in the lower, middle, and upper atmosphere. Geophys Res Lett 31:L19304
29. Garcés M, Willis M, Hetzer C (2004) The Hunt for Leaky Elevated Infrasonic Waveguides. 26th Seismic Research Review, Orlando
30. Garces M, Caron P, Hetzer C, Le Pichon A, Bass H, Drob D, Bhattacharyya J (2005) Deep infrasound from the Sumatra earthquake and tsunami. EOS 86(35):317–320
31. Garces M et al (2008) An acoustic fingerprint of stratospheric ash injection. EOS 89(40):377–378
32. Georges TM (1968) Acoustic Gravity Waves in the Atmosphere. Proceedings of the ESSA-ARPA Symposium. US Government Printing Office, Washington
33. Gossard EE, Hooke WH (1975) Waves in the Atmosphere: Atmospheric Infrasound and Gravity Waves – their Generation and Propagation. Elsevier, London
34. Goerke VH, Young JM, Cook RK (1965) Infrasonic observations of the May 16, 1963, volcanic eruption on the Island of Bali. J Geophys Res 70:6017–6022
35. Hagerty M, Schwartz S, Garces M, Protti M (2000) Analysis of seismic and acoustic observations at Arenal Volcano, Costa Rica, 1995-1997. J Volcanol Geotherm Res 101:27–65
36. Ham FM, Park S (2002) A Robust Neural Network Classifier for Infrasound Events using Multiple Array. presented at WCCI-2002, (IJCNN-2002), Honolulu, 12–17 May, pp 2615–2619
37. Harris A, Ripepe M (2008) Synergy of multiple geophysical approaches to unravel explosive eruption conduit and source dynamics – A case study from Stromboli. Chemie der Erde - Geochemistry 67(1):1–35
38. Hedin AE, Biondi MA, Burnside RG, Hernandez G, Johnson RM, Killeen TL, Mazaudier C, Meriwether JW, Salah JE, Sica RJ, Smith RW, Spencer NW, Wickwar VB, Virdi TS (1996) Revised global model of upper thermospheric winds using satellite and ground-based observations. J Geophys Res 96:7657–7688
39. Hedlin M, Garces M, Bass H, Hayward C, Herrin G, Olson JV, Wilson C (2002) Listening to the Secret Sounds of Earth's Atmosphere. EOS 83:557, 564–565
40. Hedlin MAH, Alcoverro B, D'Spain G (2003) Evaluation of rosette infrasonic noise-reducing spatial filters. J Acoust Soc Am 114:1807–1820
41. Heil C, Urban M (1992) Sound fields radiated by arrayed multiple sound sources. paper presented at the 92nd Convention of the Audio Engineering Society. Preprint no 3269, Vienna, 24–27 March 1992
42. Hercz AR (1987) Fundamentals of Sound Ranging. published by Arthur R Hercz
43. Herrin E, Golden P, Negraru P, Andre W, Bass H, Garces M, Hedlin M, McKenna M, Norris D, Osborne D, Whitaker R (2006) Infrasound Calibration Explosions from Rockets Launched at White Sands Missile Range. 28th Seismic Research Review, Orlando, 19–21 Sept 2006
44. Kajiura K (1970) Tsunami source, energy and the directivity of wave radiation. Bull Earthq Res Inst 48:835–869
45. Kim TS, Hayward C, Stump B (2004) Local infrasound signals from the Tokachi-Oki earthquake. Geophys Res Lett 31:L20605 doi:10.1029/2004GL021178
46. Larson RJ, Craine LB, Thomas JE, Wilson CR (1971) Correlation of winds and geographic features with production of certain

infrasonic signals in the atmosphere. Geophys J R Astr Soc 26:201–214
47. Le Pichon A, Garcés M, Blanc E, Barthélémy M, Drob DP (2002) Acoustic propagation and atmosphere characteristics derived from infrasonic waves generated by the Concorde. J Acoust Soc Am 111:629–641
48. Le Pichon A, Guilbert J, Vega A, Garcés M, Brachet N (2002) Ground-coupled air waves and diffracted infrasound from the Arequipa earthquake of June 23, 2000. doi:10.1029/2002GL015052
49. Le Pichon A, Guilbert J, Vallée M, Dessa JX, Ulziibat M (2003) Infrasonic imaging of the Kunlun Mountains for the great 2001 China earthquake. Geophys Res Lett 30(15):1814. doi:10.1029/2003GL017581
50. Le Pichon A, Blanc E, Drob DP, Lambotte S, Dessa JX, Lardy M, Bani P, Vergniolle S (2004) Infrasound monitoring of volcanoes to probe high altitude winds. J Geophys Res 110:D13106 doi:10.1029/2004JD005587
51. Le Pichon A, Herry P, Mialle P, Vergoz J, Brachet N, Drob D, Garcés M, Ceranna L (2005) Infrasound associated with large Sumatra earthquakes and tsunami. Geophys Res Lett 32:L19802 doi:10.1029/2005GL023893
52. Le Pichon A, Mialle P, Guilbert J, Vergoz J (2006) Multistation infrasonic observations of the Chilean earthquake of 2005 June 13. Geophys J Int 167(2):838–844. doi:10.1111/j.1365-246X.2006.03190.x
53. Lin A, Fu B, Guo J, Zeng Q, Dang G, He W, Zhao Y (2002) Coseismic strike-slip and rupture length produced by the 2001 Ms 8.1 Central Kunlun earthquake. Science 296:2015–2017
54. Matoza R, Hedlin M, Garcés M (2007) An infrasound array study of Mount St. Helens. J Volcanol Geotherm Res 160:249–262
55. McKisic JM (1997) Infrasound and the infrasonic monitoring of atmospheric nuclear explosions. DOE document PL-TR-97-2123
56. McGarr A, Vorhis RC (1968) Seismic seiches from the March 1964 Alaska earthquake. US Geological Survey Professional Paper 544-E, pp E1-E43, 1 sheet, scale 1:5,000,000
57. Mikumo T (1968) Atmospheric pressure waves and tectonic deformation associated with the Alaskan earthquake of March 28, 1964. J Geophys Res 73:2009–2025
58. Mikumo T, Shibutani T, Le Pichon A, Tsuyuki T, Watada S, Garces M, Fee D, Morii W (2008) Low-Frequency Acoustic-Gravity Waves from Tectonic Deformation Associated with the 2004 Sumatra-Andaman Earthquake (Mw=9.2). J Geophys Res 113:B12402, doi:10.1029/2008JB005710
59. Mutschlecner P, Whitaker R (1998) Infrasonic observations of earthquakes. Tech Rep LA-UR-98-2689, Los Alamos Laboratory, New Mexico
60. Mutschlecner JP, Whitaker RW (2005) Infrasound from earthquakes. J Geophys Res 110:D01108. doi:10.1029/2004JD005067
61. National Academy of Sciences (2002) Technical Issues Related to the Comprehensive Nuclear Test Ban Treaty. National Academy Press, Washington, International Standard Book Number 0-309-08506-3
62. Olson JV, Wilson CR, Hansen R (2003) Infrasound associated with the 2002 Denali fault earthquake, Alaska. Geophys Res Lett 30:N0232195. doi:10.1029/2003GL018568
63. Park J, Song T-RA, Tromp J, Okal E, Stein S, Roult G, Clevede E, Laske G, Kanamori H, Davis P et al (2005) Earth's Free Oscillations Excited by the 26 December 2004 Sumatra-Andaman Earthquake. Science 308:1139–1144
64. Sakai T, Yamasato H, Uhira K (1996) Infrasound accompanying C-type tremor at Sakurajima volcano. Bull Volcanol Soc Japan 41:181–185
65. Shumway RH (2001) Detection and location capabilities of multiple infrasound arrays. Proceedings of the 23rd NNSA Research Review: Worldwide Monitoring of Nuclear Explosions, pp 160–167
66. Sutherland L, Bass H (2004) Atmospheric absorption in the atmosphere up to 160 km. J Acoust Soc Am 115:1012–1032
67. Sutherland L, Bass H (2006) Erratum: Atmospheric absorption in the atmosphere up to 160 km. J Acoust Soc Am 120:2985
68. Tahira M, Nomura M, Sawada Y, Kamo K (1996) Infrasonic and acoustic-gravity waves generated by the Mount Pinatubo eruption of June 15, 1991. In: Newhall C, Punongbayan R (ed) Fire and Mud -Eruption and Lahars of Mount Pinatubo, Philippines. Univ. Washington Press, Seattle
69. Takahashi Y, Koyama Y, Isei T (1994) In situ measured infrasound at Sapporo associated with an earthquake occurring offshore in southwest Hokkaido on July 12, 1993. J Acoust Soc Jpn 15:409–411
70. Thatcher W (1990) Order and diversity in the modes of circum-Pacific earthquake recurrence. J Geophys Res 95:2609–2623
71. Tsuda K, Steidl J, Archuleta R, Assimaki D (2006) Site-Response Estimation for the 2003 Miyagi-Oki Earthquake Sequence Considering Nonlinear Site. Response Bull Seismol Soc Am 96(4A):1474–1482. doi:10.1785/0120050160
72. Vallée M, Bouchon M (2004) Imaging coseismic rupture in far field by slip patches. Geophys J Int 156:615–630
73. Virieux J, Farra V (1991) Ray-tracing in 3D complex isotropic media: an analysis of the problem. Geophysics 56:2057–2069
74. Virieux J, Garnier N, Blanc E, Dessa JX (2004) Paraxial ray-tracing for atmospheric wave propagation. Geophys Res Lett 31:L20106. doi:10.1029/2004GL020514
75. Weber ME, Donn WL (1982) Ducted propagation of Concorde-generated shock waves. J Acoust Soc Am 71:340–347
76. Willis M, Garcés M, Hetzer C, Businger S (2004) Infrasonic observations of open ocean swells in the Pacific: Deciphering the song of the sea. Geophys Res Lett 31:L19303
77. Winchester S (2004) Krakatoa, The Day the World Exploded. Harper Perennial
78. Young JM, Greene GE (1982) Anomalous infrasound generated by the Alaskan earthquake of 28 March 1964. J Acoust Soc Am 71:334–339
79. Yamasato H (1998) Nature of infrasonic pulse accompanying low frequency earthquake at Unzen Volcano. Japan Bull Volcanol Soc Japan 43:1–13

Books and Reviews

Bouchon M, Bouin M-P, Karabulut H, Toksöz MN, Dietrich M, Rosakis AJ (2001) How fast is the rupture during an Earthquake? New insights from the 1999 Turkey Earthquakes. Geophys Res Lett 28:2723–2726

Donn WL, Balachandran NK (1981) Mount St. Helens eruption of 18 May 1980: Air waves and explosive yield. Science 213:539–541

Garcés M, Iguchi M, Ishihara K, Morrissey M, Sudo Y, Tsutsui T (1999) Infrasonic precursors to a Vulcanian eruption at Sakurajima volcano, Japan. Geophys Res Lett 26:2537–2540

Garcés MA, Hansen RA, McNutt SR, Eichelberger J (2000) Application of wave-theoretical seismoacoustic models to the interpretation of explosion and eruption tremor signals radiated by Pavlof volcano, Alaska. J Geopys Res 105:3039–3058
Hagerty M, Schwartz SY, Protti M, Garcés M, Dixon T (1997) Observations at Costa Rican volcano offers clues to causes of eruptions. EOS Trans Am Geophys Union 78:565–571
Harkrider DG (1964) Theoretical and observed acoustic-gravity waves from explosive sources in the atmosphere. J Geophys Res 69:5295–5321
Johnson JB, Aster RC, Ruiz MC, Malone SD, McChesney PJ, Lees JM, Kyle PR (2003) Interpretation and utility of infrasonic records from erupting volcanoes. J Volc Geotherm Res 121:15–63
Kamo K, Ishihara K, Tahira M (1994) Infrasonic and seismic detection of explosive eruptions at Sakurajima Volcano, Japan, and the PEGASAS-VE early warning system. In: Casadevall T (ed) Proceedings, 1st International Symposium on Volcanic Ash and Aviation Safety. 8–12 July 1991. US Geological Survey Bulletin 2047, pp 357–365
Ripepe M, Poggi P, Braun T, Gordeev E (1996) Infrasonic waves and volcanic tremor at Stromboli. Geophys Res Lett 23:181–184
Tahira M (1982) A study of the infrasonic wave in the atmosphere. II, Infrasonic waves generated by the explosions of the volcano Sakura-jima. J Meteorol Soc Jpn 60:896–907
Tahira M, Ishihara K, Iguchi M (1988) Monitoring volcanic eruptions with infrasonic waves. In: Proceedings of the Kagoshima International Conference on Volcanoes, Tokyo, 19–23 July 1998. National Institute for Research Advancement, and Kagoshima, Kagoshima Prefectural Government, p 530–533

Inspection Games

Rudolf Avenhaus[1], Morton J. Canty[2]
[1] Armed Forces University Munich, Neubiberg, Germany
[2] Institute for Chemistry and Dynamics of the Geosphere, Forschungszentrum Jülich, Jülich, Germany

Article Outline

Glossary
Definition
Introduction
Selected Inspection Models
Future Directions
Bibliography

Glossary

Noncooperative game An n-person noncooperative game in *normal* or *strategic form* is a list of actions, called *pure strategies*, for each of n players, together with a rule for specifying each player's payoff (utility) when every player has chosen a specific action. Each player seeks to maximize her own payoff.

Mixed strategy A mixed strategy for a player in a noncooperative game is a probability distribution over that player's pure strategies.

Extensive form The extensive form of a noncooperative game is a graphical representation which describes a succession of moves by different players, including chance moves, and which can handle quite intricate information patterns.

Zero-sum game A zero-sum game is a noncooperative game in which the payoffs of all players sum to zero for any specific combination of pure strategies.

Nash equilibrium A Nash equilibrium in a noncooperative game is a specification of strategies for all players with the property that no player has an incentive to deviate unilaterally from her specified strategy. A *solution* of a noncooperative game is a Nash equilibrium which is either unique, or which, for some reason, has been selected among alternative equilibria.

Saddle point A saddle point is a Nash equilibrium of a two-person zero-sum game. The *value* of the game is the (unique) equilibrium payoff to the first player.

Utility Utilities are sequences of numbers assigned to the outcomes of any strategy combination which mirror the order of preferences of each player and which fulfill the axioms of von Neumann and Morgenstern.

Deterrence In an inspection game, deterrence is said to be achieved by a Nash equilibrium in which the inspectee behaves legally, or in accordance with the agreed rule.

Inspector leadership Leadership in inspection games is a strategic concept by which, through persuasive announcement of her strategy, the inspector can achieve deterrence.

Verification Verification is the independent confirmation by an inspector of the information reported by an inspectee. It is used most commonly in the context of arms control and disarmament agreements.

Definition

Inspection games deal with the problem faced by an *inspector* who is required to control the compliance of an *inspectee* to some legal or otherwise formal undertaking. One of the best examples of an inspector, in the institutional sense, is the International Atomic Energy Agency (IAEA) which, under a United Nations mandate, verifies the activities of all States – inspectees – signatory to the Nuclear Weapons Non-proliferation Treaty. An inspection regime of this kind is a true conflict situation, even if the inspectee voluntarily submits to inspection (in multinational or bilateral treaties this is invariably the case), because the *raison d'être* of any control authority must be the assumption

that the inspected party has a real incentive to violate its commitments. The primary objective of the inspector is to *deter* the inspectee from illegal behavior or, barring this, to catch him out. It is thus natural that quantitative models of inspections should be non-cooperative games with at least two players, inspector and inspectee(s). This survey will be limited to just one inspectee, that is, we shall restrict ourselves to two-person non-cooperative inspection games.

Inspection games should be distinguished from related topics such as quality control or the prevention of random accidents, for which there are no adversaries that act strategically, or from search-and-destroy problems. The salient feature of an inspection game is that an inspector tries to prevent an inspectee from behaving illegally in terms of some commitment. The inspectee might, for example, decide not to violate, so that there is nothing to search for. In fact, deterrence is generally the inspector's highest priority. Nevertheless, a sharp distinction between, e. g., inspection games and quality control models cannot be made in all cases, as we will see in Subsect. "Illegal Production".

Introduction

Immediately after von Neumann and Morgenstern's pioneering book *Theory of Games and Economic Behavior* [48], Arms Control and Disarmament (ACD) inspections may have been analyzed game-theoretically as classified military research; this is not known for sure but may be inferred from papers published later. Non-classified work started vigorously in the early 1960s with analyses for the United States Arms Control and Disarmament Agency (ACDA). These dealt with very general ACD problems, and also with concrete problems of test ban treaty verification. In that context probably the first genuine inspection game in the open literature was the recursive game developed by Drescher [18]. Since it was seminal for later work it is presented it in some detail in Subsect. "Customs and Smugglers".

A second phase of inspection game development started around 1968 in connection with the verification of the Treaty on the Non-Proliferation of Nuclear Weapons (NPT). There was no model for this verification system, therefore new principles and tools had to be developed and analyzed, see, e. g., [11,12,24]. Because of its importance for the further development of the whole discipline of inspection games, one of the major components of NPT verification measures, namely material accountancy, will be discussed in Subsect. "Diversion of Nuclear Material".

In Economics game-theoretic work on Accounting and Auditing was begun in the late 1960s. A first survey was given by Borch [13]. Since that time papers and books have been published regularly but on a limited scale along similar lines, placing emphasis on auditing practice, see, e. g., [14,15]; Wilks and Zimbelman [51] provided an updated review of theoretical and empirical research. In economic models known as *principal-agent problems*, in which inspections of economic transactions raise the question of their most efficient design, game-theoretic methods have been applied, early surveys having been given by Baiman [9], Kanodia [28] and Dye [19].

In the last decade, new models have been developed and analyzed again in the context of recent ACD verification developments, in particular the more stringent requirements on re-negotiated NPT verification measures [27]. Whereas under the previous regime purely technical aspects like size of the nuclear fuel cycle and accuracy of the measurement systems were considered, now qualitative features such as behavior and intentions of States had to be taken into account. This required the introduction of State-specific utilities, for first analyses of this kind see, e. g., [3,29]. Independently of concrete applications there has been an ongoing interest of mathematicians in the refinement and generalization of existing models, a few examples will be given below.

Inspections cause conflicts in many real world situations. In economics, there are services of many kinds the fulfillment or payment of which has to be verified. For example one is concerned with the central problem of principal-agent relationships, where the principal, e. g., an employer, delegates work or responsibility to the agent, the employee, and chooses a payment schedule that best exploits the agent's self-interests. The agent, of course, behaves so as to maximize her own utility given the fee schedule proposed by the principal.

Environmental agreements obviously give rise to inspection problems, but these have not yet received as much attention from modelers as one might have expected (and as they might deserve). To date most methodological analyses of inspection games have been performed in the context of arms control and disarmament.

There exist previous reviews of inspection games with objectives somewhat different to those of this survey. Avenhaus et al. [7] restrict discussion to arms control and disarmament, and emphasize the historical development. Avenhaus et al. [8] stress the methodological, and in particular the mathematical aspects. Here we undertake a new approach to organize the material, one which gives more credit to the diversity of the applications and the techniques necessary for their solution. We focus on selected

game theoretic inspection models which, together with their variants and generalizations, we hope will span the full range of the subject.

Selected Inspection Models

In the following, five inspection problems are chosen to illustrate applications of inspection games together with their analysis. They are complemented with discussion of some of their variants and generalizations. In the last illustration, the special role of the leadership concept in inspection models is emphasized.

Passenger Ticket Control

A commuter on the Munich subway is, consciously or unconsciously, involved in a non-cooperative game each time he boards a train. He can buy a valid ticket or travel "schwarz", risking a fine if he is controlled (checked). In its edition of July 18th, 1996, the daily *Süddeutsche Zeitung* reported the complaint of the Munich City Treasurer to the effect that the deployment of ticket inspectors by the local transit authority (MMV) was not worthwhile, the collected fines paying for only about half the cost of the inspectors themselves.

From the game theorist's viewpoint there is obviously an optimum control intensity that would alleviate this problem: employing just a single inspector would encourage many violations and result in her collecting more than enough fines to pay for herself, but would clearly not be in the MVV's interests. Using an army of inspectors would ensure compliance, but, there now being no fines at all, would not finance the inspectors. The optimum must lie between these two extremes.

Solution The problem can be formulated as a two-person game in *normal form* involving the MVV as player 1 and the transit passenger as player 2 [1]. The pure strategies for the MVV are to control or not to control, whereas the passenger will decide whether or not to buy a valid ticket. Let f be the fare, $b > f$ the fine and $e < b$ the control costs per passenger, all in euros. The game's normal form (also called *bimatrix form*) is shown in Fig. 1.

In the figure, the pure strategies of player 1 (control/no control) are depicted as rows and those of player 2 (legal/illegal) as columns. The payoffs to player 1 for each pure strategy combination are shown in the lower left hand corners of the corresponding squares, those for player 2 in the upper right hand corners. (This simple formulation ignores MVV overhead costs and any material gain the passenger may have from his trip.)

1 \ 2	legal q	illegal $1-q$
control p	$-f$ $f-e$	$-b$ $b-e$
no control $1-p$	$-f$ f	0 0

Inspection Games, Figure 1
Normal form of the two-person game between MVV (*player 1*) and passenger (*player 2*). The *arrows* indicate the preference directions for the two players, the *horizontal arrows* for player 2, the *vertical arrows* for player 1

A solution of the game will be a *Nash equilibrium* [35], that is, a pair of strategies, called *equilibrium strategies*, with the property that neither player has an incentive to deviate unilaterally from his or her equilibrium strategy. Equivalently, the strategies are said to be *mutual best replies*. In the Figure the preference directions, i. e., the deviation incentives, are seen to be cyclical. This means that there can be no Nash equilibrium involving pure strategies. However the equilibrium concept can be generalized to involve *mixed strategies* which are probability distributions over the sets of pure strategies. In the present case, the MVV controls with some probability p and the passenger behaves legally with probability q. The expected payoffs to the two players are then given by

$$\begin{aligned} E_1(p,q) &= (f-e)pq + (b-e)p(1-q) + f(1-p)q \\ E_2(p,q) &= -fpq - bp(1-q) - f(1-p)q\,. \end{aligned} \tag{1}$$

If we designate the mixed equilibrium strategies by p^* and q^* and the equilibrium payoffs as $E_i^* = E_i(p^*, q^*)$, $i = 1, 2$, then the conditions for Nash equilibrium are

$$\begin{aligned} E_1^* &\geq E_1(p, q^*) \quad \text{for all } p \in [0,1] \\ E_2^* &\geq E_2(p^*, q) \quad \text{for all } q \in [0,1]\,. \end{aligned} \tag{2}$$

For this situation the equilibrium strategies can be determined simply by requiring that each protagonist be made

indifferent with regard to his or her own mixed strategy, see, e. g., Morris [34]. One obtains immediately

$$p^* = \frac{f}{b}\,, \qquad E_1^* = f\left(1 - \frac{e}{b}\right)\,, \qquad q^* = 1 - \frac{e}{b}\,, \quad E_2^* = -f\,. \tag{3}$$

At equilibrium the passenger behaves illegally with positive probability $1 - q^* = e/b$, that is, deterrence is not possible. We will return to this issue in Subsect. "Sharing Common Pool Resources". Nevertheless on average he enjoys the same payoff as he would receive by paying his fare every time, namely $-f$.

Remarks The average control expenditure for the MVV is ep, whereas the mean profit from collection of fines is $bp(1-q)$. The difference is $(e - b(1-q))p$. If the passenger plays his equilibrium strategy as given by Eq. (3), then

$$(e - b(1 - q^*))p = 0 \tag{4}$$

for any control probability p. The control costs are thus exactly compensated by the collected fines. It might be mentioned that the actual figures for inspection probability and income per passenger for Munich approximately satisfy (3). We may speculate that the City Treasurer's complaint arises from the fact that regular violators develop strategies to recognize and avoid inspectors. Other complications not taken into account in the model are variations in frequency, hour of day and dwelling time of passengers using the system.

There are many inspection problems which can be described with models equivalent or similar to the one presented here. Inspection of metered parking spaces provides another example. Control of the sharing of common pool resources, as discussed in the last inspection model, below, belongs to the same category.

Illegal Production

In treaties prohibiting the production, acquisition and/or proliferation of weapons of mass destruction, one is often concerned with the misuse of ostensibly legitimate production facilities. A commercial chemical plant, for example, may be used for production of forbidden precursors of chemical weapons, or a uranium enrichment facility may illegally enrich its product to weapons-grade U-235. The consequences of non-detection by an inspecting authority can be dire, and the *timeliness* of control procedures – the interval between the onset of plant misuse and its detection – may be especially important.

To illustrate, consider the following simple model. At the end of some reference time interval, for instance a calendar year or a production campaign, a major inspection takes place at a facility, one which would detect prior illegal production with certainty. Additionally, a single *interim inspection* is carried out, timed at the inspector's discretion, which will likewise detect prior violation with certainty. The interim inspection is intended to enhance the timeliness of detection should illegal activity be underway. The inspector would like to know precisely when it should take place.

Solution This example entails solution of a *zero-sum game on the unit square*. The onset of illegal production and the time of the interim inspection and are chosen on the interval strategically by the respective protagonists, plant operator and inspector. We take the payoff to the former as the time to detection of illegal production, and to the latter to be the negative of that quantity.

Representing the reference time by the interval $[0, 1]$, the operator's so-called *payoff kernel* is

$$A(y, x) = \begin{cases} y - x & \text{for } x \le y \\ 1 - x & \text{for } x \ge y\,, \end{cases} \tag{5}$$

here $x \in [0, 1]$ denotes a pure strategy for the operator, the onset of illegal production and similarly $y \in [0, 1]$ is a pure inspection strategy. Let the inspector's and operator's *mixed strategy distribution functions* be $G(y)$ and $F(x)$, respectively. $G(y)$ is the probability of an inspection taking place at time y or earlier, $F(x)$ the probability of illegal activity beginning at time x or earlier. The Nash equilibrium conditions, or, since we are dealing with a zero-sum game, the *saddle point criteria* determining the equilibrium mixed strategies G^* and F^*, are given by

$$E(G^*, x) \le E(G^*, F^*) \le E(y, F^*) \quad \text{for all } x, y \in [0, 1]\,. \tag{6}$$

Since the final inspection is certain, clearly the operator shouldn't wait too long to act. Rather, in constructing his optimal probability distribution $F^*(x)$, he might plausibly choose x randomly on an interval $[0, b]$ with $b < 1$. Consequently the inspector will not act later than b either. Let us assume that she chooses her inspection time y according to the probability density function $g^*(y)$, where

$$\int_0^b g^*(y)\mathrm{d}y = 1\,. \tag{7}$$

The expected payoff to the operator for diversion at time $x \in [0, b]$ is then

$$\begin{aligned} E(G^*, x) &= \int_0^x (1-x)g^*(y)\mathrm{d}y + \int_x^b (y-x)g^*(y)\mathrm{d}y \\ &= \int_0^x g^*(y)\mathrm{d}y + \int_x^b yg^*(y)\mathrm{d}y \\ &\quad - x\int_0^b g^*(y)\mathrm{d}y\,. \end{aligned} \tag{8}$$

But if the operator randomizes across the interval $[0, b]$ as assumed, this payoff must be constant for all $x \in [0, b]$ and equal to the value of the game, i. e., the equilibrium payoff to the operator. If this were not the case for some x in the interval, the operator would not have included it in his mixed strategy F^* in the first place. Thus the derivative of Eq. (8) with respect to x must vanish. This gives immediately

$$g^*(y) = \frac{1}{1-y} \tag{9}$$

and from Eq. (7) $b = 1 - 1/e$. The expected payoff to the operator and value of the game is then easily seen to be $E(G^*, x) = 1/e$ for all $x \in [0, b]$.

Getting the operator's optimal strategy F^* is a bit more subtle because it requires a so-called *atom* at $x = 0$, that is to say, a finite probability of starting illegal production at precisely the beginning of the interval, as well as the probability density $f^*(x)$ on the remaining half-open interval $(0, b]$. In terms of the distribution function $F^*(x)$ that characterizes this mixed strategy, the atom is $F^*(0)$ and $f^*(x)$ is the derivative of $F^*(x)$ on $(0, b]$. The operator's payoff for some inspection time $y \in [0, b]$ is

$$\begin{aligned} E(y, F^*) &= yF^*(0) + \int_0^y (y-x)f^*(x)\mathrm{d}x \\ &\quad + \int_y^b (1-x)f^*(x)\mathrm{d}x \\ &= yF^*(0) + y\int_0^y f^*(x)\mathrm{d}x + \int_y^b f^*(x)\mathrm{d}x \\ &\quad - \int_0^b xf^*(x)\mathrm{d}x \\ &= (y-1)F^*(y) + F^*(b) - \int_0^b xf^*(x)\mathrm{d}x\,. \end{aligned} \tag{10}$$

Arguing as before, if the the inspector randomizes over the interval, this expression must be constant for all $y \in [0, b]$. This is true if $(y-1)F^*(y)$ is independent of y. The requirement that $F^*(b) = 1$ then leads to

$$F^*(x) = \frac{1}{e} \cdot \frac{1}{1-x} \tag{11}$$

for $x \in [0, b]$ and $F^*(x) = 1$ for $x > b$. The atom is $F^*(0) = 1/e$, and the construction is complete. Verifying that F^* and G^* satisfy Eq. (6) is straightforward.

Remarks Owen [38] discussed the existence of Nash equilibria for continuous games on the unit square and methods for their solution. The above result was first obtained by Diamond [17], who gave a generalization to $k > 2$ inspections. Prior to Diamond's work, Derman [16] treated a somewhat similar minimax inspection problem.

Both models were motivated by reliability control problems: production units have to be inspected regularly and the earlier a failure is detected, the less costly it is for the production facility owner. Of course the production unit is not acting strategically, but in order to be on the safe side a minimax approach was chosen, which was then generalized by Diamond to give a saddle point solution. Thus we see that an approach which originally was not an inspection game according to our definition in Sect. "Definition" turned out to become one, with interesting applications such as the one discussed above.

Rothenstein and Zamir [42] extended Diamond's model with a single inspection to include errors of the first and second kind. Krieger [30] considered a time-discrete variant of the model. All variants require that both the operator and inspector commit themselves before the reference period begins. Thus if in the solution in Subsect. "Solution" the operator simply waits for the interim inspection and then violates, he will achieve an expected time to detection of

$$\int_0^b (1-x)f^*(x)\mathrm{d}x = b = 1 - \frac{1}{e} > \frac{1}{e}\,,$$

and the inspector's advantage will have evaporated. But of course $f^*(x)$ is not the inspector's equilibrium in such a sequential game. Prior commitment may be justified in some cases, but not in others. If there is no requirement for commitment, the operator may prefer to start his illegal action immediately, i. e., at the beginning of the reference period, or delay his decision until the first intermediate inspection. This situation has to be modeled as an *extensive form game*. Its time-continuous version, which also considered errors of the first and second kind, was studied by Avenhaus and Canty [4]. Surprisingly it turned out that an equilibrium strategy of the inspector is a pure strategy, contrary to the equilibrium strategies of the Diamond-type models.

Diversion of Nuclear Material

As already mentioned in the introduction, a large number of game theoretic models of inspection situations have been developed in the framework of IAEA verification activities. The basis of the IAEA inspection system is the verification of the continuing presence of fissile material in the peaceful nuclear fuel cycle of the State under consideration [26], therefore statistical measurement errors and, consequently, statistical decision theory must be taken into account.

Over a single accounting period, typically one year, we define the *material flow* as the measured net transfer of fissile material across the facility boundary, consisting of inputs (receipts of raw material) R and outputs (shipments of purified product and waste) S. If the physical inventory within the facility at the start of the period was I_0, then the *book inventory* at the end of the accounting period is defined as

$$B = I_0 + R - S = I_0 + Y \,, \tag{12}$$

where $Y = R - S$ is the net material flow into the facility.

At the end of the period a new physical inventory I_1 is taken and compared with the book inventory,

$$Z = B - I_1 = I_0 + Y - I_1 \,. \tag{13}$$

This expression, which is a random variable as a consequence of measurement error, defines the *material balance statistic*. If there are no unknown losses or diversions of material, its expectation value is

$$E(Z) = E(I_0) + E(Y) - E(I_1) = 0 \tag{14}$$

from conservation of mass. The quantity Z is commonly referred to as MUF, meaning *material unaccounted for*. Its reliable determination forms the basis for the inspector's conclusion with respect to non-diversion.

We shall focus upon a single nuclear facility and an inventory period of one year and pose the question: can the taking of *additional interim inventories* improve the detection sensitivity of the overall accountancy procedure?

Solution The additional inventories essentially define a series of shorter material balance periods, say n in all. At the beginning of the first balance period, the amount I_0 of material subject to control is measured in the facility. Then, during the ith period, $i = 1 \ldots n$, some net measured amount Y_i of material enters the area. At the end of that period the amount of material, now I_i, is again measured. The quantity

$$Z_i = I_{i-1} + Y_i - I_i \,, \quad i = 1 \ldots n \,,$$

is the material balance test statistic for the ith inventory period. Under the *null hypothesis* that no material was diverted, its expected value is, as before,

$$E_0(Z_i) = 0 \,, \quad i = 1 \ldots n \,.$$

The *alternative hypothesis* is that material is diverted from the balance area according to some specific pattern. Thus

$$E_1(Z_i) = \mu_i \,, \quad i = 1 \ldots n \,, \quad \sum_{i=1}^{n} \mu_i = \mu > 0 \,,$$

where the amount μ_i diverted in the ith period may be positive, negative or nil, while μ, the total amount of material missing, is hypothesized to be positive.

For the purpose of determining the best test procedure we now define a two-person zero-sum game, wherein the set of strategies of the inspector is the set of all possible test procedures $\{\delta_\alpha\}$, i. e., significance thresholds, for fixed false alarm probability α. The set of strategies of the operator is the set of diversion patterns $\boldsymbol{\mu} = (\mu_1 \ldots \mu_n)^{\mathrm{T}}$, $\sum_i \mu_i = \mu$. The payoff to the inspector is the probability of detection $1 - \beta(\delta_\alpha, \boldsymbol{\mu})$, where $\beta(\delta_\alpha, \boldsymbol{\mu})$ is the *second kind error probability* (= non-detection probability). A solution of the game is any strategy pair $(\delta_\alpha^*, \boldsymbol{\mu}^*)$ which satisfies the saddle point conditions

$$1 - \beta(\delta_\alpha^*, \boldsymbol{\mu}) \geq 1 - \beta(\delta_\alpha^*, \boldsymbol{\mu}^*) \geq 1 - \beta(\delta_\alpha, \boldsymbol{\mu}^*) \quad \text{for any } \delta_\alpha, \boldsymbol{\mu} \,. \tag{15}$$

With the aid of the Lemma of Neyman and Pearson, one of the most fundamental theorems in statistical decision theory, see, e. g., Rohatgi [41], we can derive the following solution first obtained by Avenhaus and Jaech [5]. Suppose that $\boldsymbol{\Sigma}$ is the covariance matrix of the multivariate normally distributed random vector $\boldsymbol{Z} = (Z_1, Z_2, \ldots Z_n)^{\mathrm{T}}$ and define $\boldsymbol{e} = (1, 1 \ldots 1)^{\mathrm{T}}$. Then the equilibrium strategies are in fact given by

$$\boldsymbol{\mu}^* = \frac{\mu}{\boldsymbol{e}^{\mathrm{T}} \cdot \boldsymbol{\Sigma} \cdot \boldsymbol{e}} \cdot \boldsymbol{\Sigma} \cdot \boldsymbol{e} \,, \tag{16}$$

and by the test δ_α^* characterized by the critical region

$$\{\boldsymbol{z} | \boldsymbol{e}^{\mathrm{T}} \cdot \boldsymbol{z} > k_\alpha\} \tag{17}$$

where k_α is determined by α. The value of the game, that is, the guaranteed probability of detection, is given by

$$1 - \beta(\delta_\alpha^*, \boldsymbol{\mu}^*) = \Phi\left(\frac{\mu}{(\boldsymbol{e}^{\mathrm{T}} \cdot \boldsymbol{\Sigma} \cdot \boldsymbol{e})^{\frac{1}{2}}} - U(1-\alpha)\right) \,, \tag{18}$$

where Φ is the normal distribution and U is its inverse.

We can demonstrate that this is the case as follows: Under the null hypothesis of legal behavior

$$1-\alpha = \mathrm{Prob}_0\left(\boldsymbol{Z}^{\mathrm{T}}\cdot\boldsymbol{e} < k_\alpha\right) = \Phi\left(\frac{k_\alpha}{(\mathrm{var}(\boldsymbol{Z}^{\mathrm{T}}\cdot\boldsymbol{e}))^{\frac{1}{2}}}\right)$$

and therefore

$$\frac{k_\alpha}{(\mathrm{var}(\boldsymbol{Z}^{\mathrm{T}}\cdot\boldsymbol{e}))^{\frac{1}{2}}} = U(1-\alpha)\,.$$

Thus the left hand side of Eq. (15) is fulfilled as equality:

$$\begin{aligned}1-\beta(\delta^*_\alpha,\boldsymbol{\mu}) &= \mathrm{Prob}_1(\boldsymbol{Z}^{\mathrm{T}}\cdot\boldsymbol{e} > k_\alpha)\\ &= 1-\Phi\left(\frac{k_\alpha - E_1(\boldsymbol{Z}^{\mathrm{T}}\cdot\boldsymbol{e})}{(\mathrm{var}(\boldsymbol{Z}^{\mathrm{T}}\cdot\boldsymbol{e}))^{\frac{1}{2}}}\right)\\ &= \Phi\left(\frac{\mu}{(\boldsymbol{e}^{\mathrm{T}}\cdot\boldsymbol{\Sigma}\cdot\boldsymbol{e})^{\frac{1}{2}}} - U(1-\alpha)\right)\,.\end{aligned}$$

As for the right hand side, the critical region which maximizes the detection probability for $\boldsymbol{\mu}^*$ and for fixed α is, according to the Lemma of Neyman and Pearson, given by

$$\left\{\boldsymbol{z}\middle|\frac{f_1(\boldsymbol{z})}{f_0(\boldsymbol{z})} > k''_\alpha\right\} = \{\boldsymbol{z}\mid \boldsymbol{\mu}^{*\mathrm{T}}\cdot\boldsymbol{\Sigma}^{-1}\cdot\boldsymbol{z} > k'_\alpha\}\,,$$

where $f_i(\boldsymbol{z})$ are the joint density functions under hypothesis i, $i=0,1$. But from Eq. (16)

$$\boldsymbol{\mu}^{*\mathrm{T}}\cdot\boldsymbol{\Sigma}^{-1}\cdot\boldsymbol{z} \propto (\boldsymbol{\Sigma}\cdot\boldsymbol{e})^{\mathrm{T}}\cdot\boldsymbol{\Sigma}^{-1}\cdot\boldsymbol{z} = \boldsymbol{e}^{\mathrm{T}}\cdot\boldsymbol{z}\,.$$

Thus δ^*_α is indeed a best reply to $\boldsymbol{\mu}^*$ and the right hand inequality (15) is fulfilled as well.

Remarks According to Eq. (17), the inspector's optimal test statistic is

$$\boldsymbol{e}^{\mathrm{T}}\cdot\boldsymbol{Z} = \sum_{i=1}^{n} Z_i = I_0 + \sum_{i=1}^{n} Y_i - I_n\,,$$

which is just the overall material balance for the entire time period involved. All of the intermediate inventories I_i, $i=1\ldots n-1$, are *ignored*. This gives a definitive answer to the question as to whether additional inventory taking can improve the sensitivity of the material balancing system for detecting diversion. The answer is no.

Satisfying as it may be from a decision theoretic point of view, this result ignores the aspect of detection time. Waiting one year or one complete production campaign before evaluating the overall material balance may be too long to meet timeliness constraints. Therefore, under the name *near real-time material accountancy*, test procedures have been discussed which indeed subdivide the year into several inventory periods (at the cost of reduced overall detection sensitivity, as was just explained). To date, it has not been possible to define or to solve satisfactorily a decision theoretic model which takes the critical time aspect into account.

The IAEA safeguards system is organized in such a way that the inspector compares the material balance data reported by the plant operators (via their national authorities) with her own findings and thereafter, if no discrepancies are found, closes the material balance with the help of the operator´s reported data. Thus, along with material accountancy, *data verification* comprises the second foundation of the IAEA safeguards system. Due to the possibility that data may be intentionally falsified to make the material balance appear to be correct, data verification again poses game theoretic problems.

Two kinds of sampling procedures have to be considered in this context. In case of identifying items or checking of seals, so-called *attribute sampling* procedures are used in which only sampling errors have to be minimized. This leads on the one hand to stratified sampling solutions similar to those found in the context of accounting and auditing. One of these solutions became well-known, at least in expert circles, under the name IAEA formula, see e. g. [2]. On the other hand, in case of quantitative destructive and non-destructive verification measurements, statistical measurement errors can no longer be avoided, leading to consideration of *variable sampling* procedures. A decision problem arises in this instance, since discrepancies between reported and independently verified data can be caused either by measurement errors or by real and intentionally generated differences (data falsification). Stewart [46] was the first to propose the so-called D-statistic for use in IAEA data verification. For one class of data consisting of N items, n of which are verified, the D-statistic is the sum of the differences of reported data X_j and independently measured data Y_j, extrapolated to the whole class population, i. e.,

$$D_1 = \frac{N}{n}\sum_{j=1}^{n}(X_j - Y_j)\,.$$

For K classes of data (for instance one class for each component of a closed material balance) the D-statistic is given by

$$D_K = \sum_{i=1}^{K}\frac{N_i}{n_i}\sum_{j=1}^{n_i}(X_{ij} - Y_{ij})\,.$$

These quantities then form the basis for the test procedure of the inspector, which goes along similar lines as outlined

before: Two hypotheses have to be formulated which permit the determination of significance thresholds for fixed false alarm probabilities and, from them, the associated detection probabilities.

Later on [3] it was proven, again using the saddle point criterion and the Lemma of Neyman and Pearson, that the use of the D-statistic is optimal for a "reasonable" class of data falsification strategies, and it was shown how the sample sizes can be determined such that they maximize the overall probability of detecting a given total falsification for a total given inspection effort.

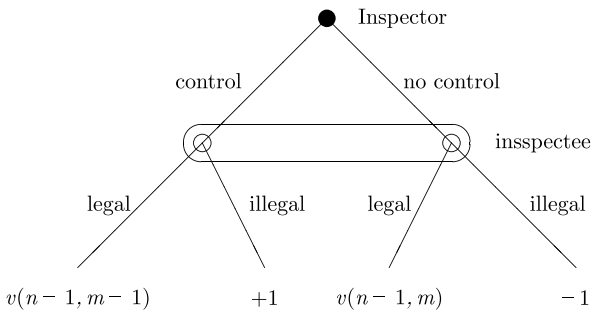

Inspection Games, Figure 2
Dresher's game in reduced extensive form

Customs and Smugglers

In the previous examples evidence of violation (illegal production, diversion) is assumed to persist: the illegal action can be detected after the fact. There are of course inspection problems where this kind of model is not appropriate. Probably the first genuine inspection game in the open literature was a recursive zero-sum game developed by Dresher [18] which treated the situation in which the violator can only be caught red-handed, that is, if the illegal act is actually in progress when an inspection occurs. It's not difficult to imagine real situations where this is the case, a much-discussed example being the customs-smuggler problem [47]. In its simplest form, a smuggler has n nocturnal opportunities to bring his goods safely across a channel. The customs office, equipped with a patrol boat, would very much like to apprehend him, but budget restrictions require that the boat can only patrol on $m < n$ nights. If a patrol coincides with a crossing attempt, the smuggler will be apprehended with certainty. Moreover the smuggler observes all patrols that take place. All that being the case, one can ask how customs should time its patrols.

Solution The game theoretic model developed by Dresher [18] at the RAND Corporation mentioned above fits this situation rather well. It illustrates nicely the special character of sequential games, and has an elegant recursive solution. We summarize it here, see von Stengel [50] for a more thorough discussion as well as some interesting variations on the same theme.

In Dresher's model there are n time periods, during each of which the inspector can decide whether or not to control the inspectee, using up one of a total of m inspections available to her if she does so. The inspectee knows at each stage the number of past inspections. He can violate at most once, but can also choose to behave legally. Detection occurs only if violation and inspection coincide in the same period. The conflicting interests of the two players are again modeled as a zero-sum game, that is, the inspectee's loss on detection is the negative of the inspector's gain. Legal action gives a payoff of nil to both players. The game is shown in reduced *extensive form*, i. e., as a decision tree, in Fig. 2.

The inspectee's *information set*, shown as an oval in the figure, encompasses both of her decision points. This is meant to imply that she doesn't know at which node she is situated when choosing her strategy.

The entries at the leaf nodes of the tree are the payoffs to the inspector. The value of the game prior to the first period is denoted $v(n, m)$. If the single violation occurs, the inspector achieves $+1$ if an inspection takes place, otherwise -1. In the latter case, the game proceeds trivially with the inspectee behaving legally (he has already violated) and the inspector inspecting or not, as she chooses. If the inspectee behaves legally, the continuation of the game has, by definition, value $v(n-1, m-1)$ to the inspector if she decided to control in the first period, otherwise value $v(n-1, m)$. These values are the corresponding payoffs to the inspector after the first period, thus giving the recursive form of the game tree shown. The game terminates after detected violation or after the n periods, in the latter case with a payoff of either 0 (legal behavior) or -1 (illegal behavior) to the inspector.

The function $v(n, m)$ is subject to two boundary conditions. If there are no periods left, and no violation has occurred,

$$v(0, m) = 0\,, \quad m \geq 0\,. \tag{19}$$

If the inspector has no inspections left, then the inspectee is aware of this and can safely violate (and will do so, since his payoff is higher):

$$v(n, 0) = -1\,, \quad n > 0. \tag{20}$$

We shall now seek an equilibrium of the game in the domain of mixed strategies. Let $p(n, m) \in [0, 1]$ be the prob-

ability with which the inspector chooses to inspect in the first period. The equilibrium choice for $p(n,m)$ makes the inspectee indifferent to legal or illegal behavior, so that he receives the same payoff $-v(n,m)$ given by

$$\begin{aligned} v(n,m) &= p(n,m)v(n-1,m-1) \\ &\quad + (1-p(n,m))v(n-1,m) \quad \text{(legal)} \\ v(n,m) &= p(n,m)(+1) + (1-p(n,m))(-1) \quad \text{(illegal)}\,. \end{aligned}$$

In a similar way the inspector chooses her probability $q(n,m)$ for violation at the first stage so as to make the inspectee indifferent as to control or no control, leading to

$$\begin{aligned} &q(n,m)v(n-1,m-1) + (1-q(n,mZ))1 \\ &\quad = q(n,m)v(n-1,m) + (1-q(n,m))(-1)\,. \end{aligned}$$

Solving these three equations for $p(n,m)$, $q(n,m)$ and $v(m,n)$ we obtain

$$p(n,m) = \frac{v(n-1,m)+1}{v(n-1,m)+2-v(n-1,m-1)}, \quad m<n \tag{21}$$

$$q(n,m) = \frac{2}{v(n-1,m)+2-v(n-1,m-1)} \tag{22}$$

and

$$v(n,m) = \frac{v(n-1,m)+v(n-1,m-1)}{v(n-1,m)+2-v(n-1,m-1)}\,. \tag{23}$$

Equation (23) along with the two boundary conditions (19) and (20) determine $v(n,m)$ and therefore $p(n,m)$ and $q(n,m)$ uniquely. The explicit solution is given as follows: Define

$$t(n,m) = \sum_{i=1}^{m} \binom{n}{i}\,, \tag{24}$$

where $\binom{n}{i}$ denotes the binomial coefficient. For $0<m<n$ the value of the game and the equilibrium strategies for both players are

$$v(n,m) = -\frac{\binom{n-1}{m}}{t(n,m)} \tag{25}$$

$$p(n,m) = \frac{t(n-1,m-1)}{t(n,m)} \tag{26}$$

$$q(m,n) = \frac{2}{\frac{\binom{n-2}{m}}{t(n-1,m)} - \frac{\binom{n-2}{m-1}}{t(n-1,m-1)} - 2}\,. \tag{27}$$

It is not quite trivial to prove that this solution satisfies its determinants. In fact one has to use the following recursive relations for $t(n,m)$:

$$\begin{aligned} t(n-1,m) &= t(n-1,m-1) + \binom{n-1}{m} \\ t(n,m) &= t(n-1,m) + t(n-1,m-1)\,. \end{aligned}$$

Remarks Since the value of the game is negative, the inspectee will behave illegally with positive probability. (If at some stage in an actual play, the inspector has as many inspections left as there are remaining opportunities, the inspectee will obviously behave legally). Dresher's model uses abstract payoffs – utilities in the sense of von Neumann and Morgenstern [48] – and therefore the zero-sum assumption becomes questionable. In some circumstances it can be argued that detected illegal action is, compared to legal action, *worse for both players*, inspector and inspectee. However, the non-zero-sum version of the game is not much more difficult to analyze and gives no additional structural insight.

Two technical remarks on Dresher's inspection game:

First of all, it works so well only because the concept of *behavioral strategies*, introduced by Kuhn [31], can be applied. Kuhn showed that in a game in extensive form with *perfect recall* (i. e., a game in which both players remember their previous moves) mixed strategies can be replaced by sequences of behavioral strategies. These are probabilities which define the actions of both players at all information sets of the game. In fact, $p(n,m)$ and $q(n,m)$ defined above are behavioral strategies.

Second, the recursive approach requires that at each decision node there follow a *subgame* (i. e. a branch of the tree) which by definition may not be connected by a joint information set to other subgames. This, however is *not* the case in Dresher's game: if the control doesn't take place, the inspector doesn't know whether the inspectee has already behaved illegally and thus whether a true recursive subgame still exists. Fortunately this doesn't matter, since if the inspectee did in fact behave illegally the game is de facto over. Whatever strategy the inspector chooses subsequently will not affect her payoff.

As indicated initially, variants and extensions of Dresher's original model probably represent the largest class of inspection games formulated and analyzed so far, and new variants are still being published, see, e. g., [25,39]. As also mentioned above, Dresher's zero sum assumption may be questioned – detected violation may still be worse for the inspector than no violation at all and, thus, negative for both players. Höpfinger [23] was the first to introduce generalized utilities and to determine the equilibria of the resulting recursive game. Von Stengel [50] considered the case that the inspectee in-

tends to violate more than once, however only once at a given stage. He was able to solve the recursive game for at most k violations on n stages, $k < n$, under the assumption that the inspector is fully informed after each stage whether or not there is a violation, even when no inspection takes place, see also [20,43]. A third category of extensions is associated with the customs-smuggler problem, see e. g. [10,21,22], where customs and the smuggler have more possibilities, e. g., customs has more than one boat, and the smugglers alternative routes. Finally, statistical errors of the first and second kind were considered, firstly by Maschler [33] and later by Rinderle [40]. However, due to the complexity of the models and the many additional assumptions which had to be made, explicit solutions have so far been found only in special cases .

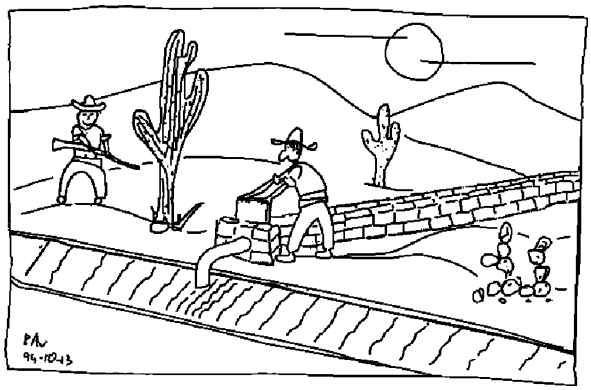

Inspection Games, Figure 3
The games irrigators play. (Reprinted from [3] with permission of Cambridge University Press)

Sharing Common Pool Resources

The sharing of so-called *common-pool resources* (see, e. g., [37]) involves an implicit inspection problem: once agreement has been reached, the parties concerned will want to make sure that the rules are followed. The following idealized common-pool resource problem illustrates this.

The Hatfields and the McCoys (who else!) farm neighboring areas of irrigated land, the Hatfields having first access to the water. They have agreed on a fair sharing procedure, but there exists a certain amount of mutual distrust. (In fact, the two clans have not been on speaking terms for 40 years.) Broadly speaking, farmer Hatfield has two strategies: to take more water than his fair share, with the twofold incentive of better crops and "doing-in" the McCoys, or to stand by the agreement. Farmer McCoy, who waits his turn for the water, has the option of controlling Hatfield at some penalty (danger to life and limb) or of simply doing nothing and taking what he gets. The strategic situation is illustrated in Fig. 3.

We may normalize the payoffs to zero for legal behavior and no control. Relative to this normalization, let e be the cost to McCoy of monitoring Hatfield's activities and a the loss he entails by not getting his fair share, but having the satisfaction of catching Hatfield out, $e > 0$, $a > 0$. Thus if Hatfield violates and McCoy monitors, McCoy's payoff is $-a$ relative to the normalization, while if he monitors but there is no violation his payoff is $-e$. Suppose, however, that McCoy's highest priority is *to keep Hatfield honest*. Then, necessarily, $a > e$. The worst outcome for McCoy is certainly undetected violation, with payoff $-c$, and so $c > a$. Hatfield's payoffs are simply $+d$ for undetected violation, and $-b$ for detected violation, $(b, d) > (0, 0)$. The game is depicted in bimatrix form in Fig. 4.

1 \ 2	take share $1-t$	take more t
no control β	0 / 0	$+d$ / $-c$
control $1-\beta$	0 / $-e$	$-b$ / $-a$

Inspection Games, Figure 4
Normal or bimatrix form of the irrigation game. *Left lower numbers* are the payoffs to player 1 (McCoy), *upper right numbers* the payoffs to player 2 (Hatfield), whereby $c > a > e$ and $(a, b, c, d, e) > (0, 0, 0, 0, 0)$. The *horizontal arrows* are incentive directions for player 2, the *vertical arrows* for player 1. The variables t and β define mixed strategies

Solution Just as in the passenger ticket control game, the preference arrows are cyclic. There is again a unique equilibrium in mixed strategies:

$$\beta^* = \frac{b}{b+d}\,, \quad t^* = \frac{e}{e+c-a}\,, \tag{28}$$

with payoffs to McCoy and Hatfield, respectively, given by

$$I_{\mathrm{M}}(\beta^*, t^*) = -e \cdot \frac{c}{e+c-a}\,, \quad I_{\mathrm{H}}(\beta^*, t^*) = 0\,. \tag{29}$$

Hatfield violates his commitment with probability $t^* > 0$ even though his payoff is the same as for behaving legally.

From a moralist's viewpoint this may not be particularly satisfactory, but Hatfield's equilibrium behavior is, given the circumstances, rational.

However it was postulated that McCoy's highest priority was to keep Hatfield honest, and there is in fact a way for him to do it. Suppose that McCoy takes the initiative and *announces credibly with what precise probability* he intends to monitor Hatfield's activities. This so-called *leadership game* can no longer be expressed as a bimatrix as in Fig. 4, because McCoy's set of pure strategies, that is, his choices of which monitoring probability $1-\beta$ he will announce in advance, is infinite. Hatfield's set of strategies on the other hand consists of all functions which assign to each value of $1-\beta$ the decision "take share" or "take more". The appropriate representation is the extensive form game shown in Fig. 5.

Due to its structure (formally, an extensive form game with *perfect information* in which all information sets are singletons) the game can be solved by *backward induction*. If this procedure leads to an equilibrium, then that equilibrium is said to be *subgame perfect*. A subgame perfect Nash equilibrium is one in which every subgame is also a Nash equilibrium.

The first step is to replace the outcome payoffs in Fig. 5 by their expected values, as shown in Fig. 6. The argument then proceeds as follows: Hatfield, knowing the probability $1-\beta$ of being controlled, decides

$$\begin{array}{lll} \text{take share} & \text{if} & 0 > -b + (b+d)\beta \\ \text{indifferent} & \text{if} & 0 = -b + (b+d)\beta \\ \text{take more} & \text{if} & 0 < -b + (b+d)\beta\,, \end{array} \tag{30}$$

since this strategy will always maximize his expected payoff. McCoy's equilibrium strategy will be shown below to be

$$\beta^* = \frac{b}{b+d}\,, \tag{31}$$

so (30) is equivalent to Hatfield's following decision:

$$\begin{array}{lll} \text{take share} & \text{if} & \beta < \beta^* \\ \text{indifferent} & \text{if} & \beta = \beta^* \\ \text{take more} & \text{if} & \beta > \beta^*\,. \end{array} \tag{32}$$

What is Hatfield's equilibrium strategy? In order to determine it, we must first define his set of strategies a little more carefully. A typical element of the set will be a recipe which tells him, for every conceivable announcement by McCoy of a value of the non-monitoring probability β, whether or not to take more than his share.

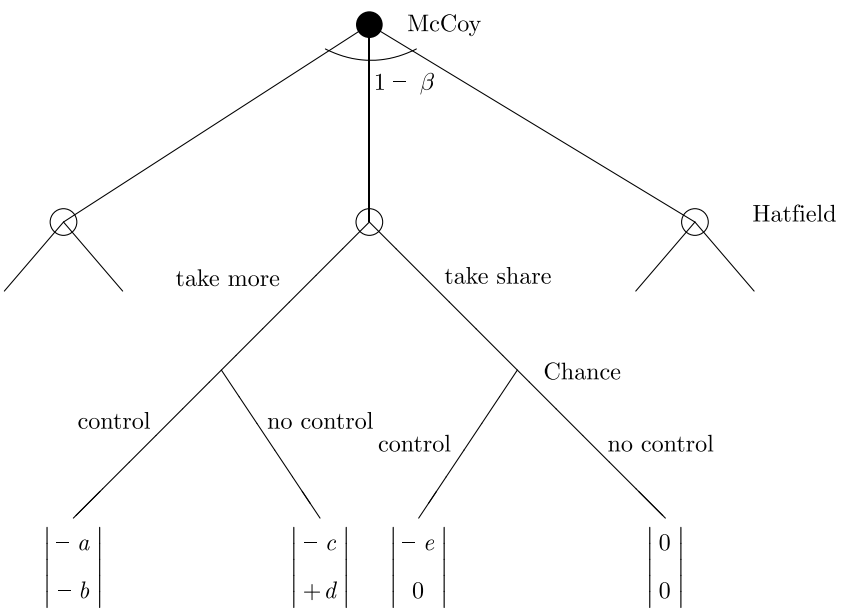

Inspection Games, Figure 5

Extensive form of the inspector leadership irrigation game. McCoy begins by choosing his monitoring probability $1-\beta$ and announces it to Hatfield, the latter then deciding whether or not to take more than his share of water. Finally, chance decides if the monitoring actually takes place. The payoffs for each possible outcome are shown at the corresponding leaf nodes, McCoy *above*, Hatfield *below*

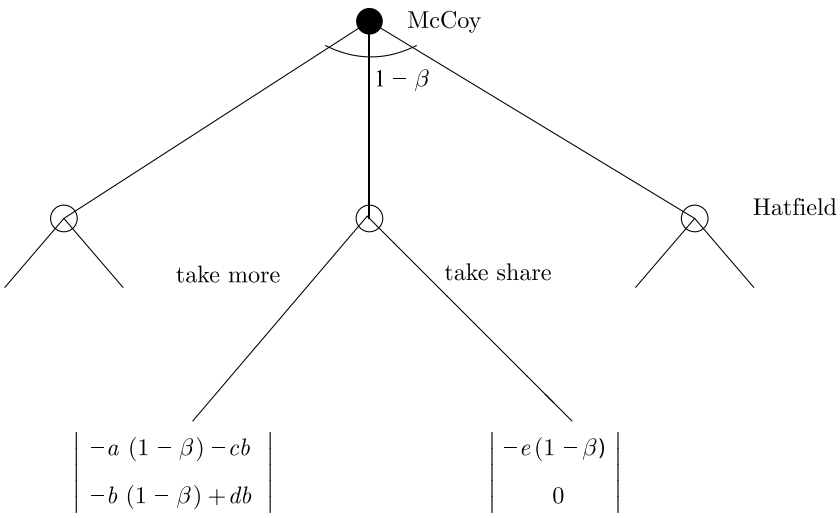

Inspection Games, Figure 6
Reduced extensive form of the inspector leadership irrigation game of Fig. 5

Recipe (30) is certainly one such, although it leaves an ambiguity for $\beta = \beta^*$. In that case Hatfield may decide to make his choice randomly, in other words to use a mixed strategy. In general, such a mixed strategy is given by a probability $t(\beta)$ for 'take more', and the complementary probability $1 - t(\beta)$ for 'take share', for all $\beta \in [0, 1]$. Hatfield's complete strategy set is therefore the set of all functions which map the unit interval onto itself, $\{t(\beta) \mid t\colon [0, 1] \mapsto [0, 1]\}$.

We now assert that Hatfield's equilibrium strategy t^* is in fact always a pure strategy, namely

$$t^*(\beta) = \begin{cases} 0 = \text{take share} & \text{for } \beta \le \beta^* \\ 1 = \text{take more} & \text{for } \beta > \beta^* \,, \end{cases} \tag{33}$$

where β^* is given by (31). This is just the conclusion we reached in Eq. (32) by backward induction, except for the case $\beta = \beta^*$. We still have to show that $t^*(\beta^*) = 0$, that is, that Hatfield stays honest at equilibrium. To do this, we now have to consider McCoy's payoffs.

The expected payoff to McCoy, as a function of β, is given by

$$I_{\mathrm{M}}(\beta, t^*) = \begin{cases} -e(1-\beta) & \text{if } \beta < \beta^* \\ -e(1-\beta^*)(1-t(\beta^*)) & \\ +(-c\beta^* - a(1-\beta^*))t(\beta^*) & \text{if } \beta = \beta^* \\ -c\beta - a(1-\beta) & \text{if } \beta > \beta^* \,. \end{cases} \tag{34}$$

It is plotted in Fig. 7. McCoy also wishes to maximize his expected payoff. For $\beta < \beta^*$, when Hatfield stays honest, it is at least $-e$ and increases with increasing β. For $\beta > \beta^*$, when Hatfield takes more than his share, it is between $-c$ and $-a$ and certainly worse for McCoy. For $\beta = \beta^*$, McCoy's payoff is something intermediate, depending on Hatfield's behavior. The argument seems to be getting circular, but we are almost done.

McCoy's equilibrium strategy, if he has one at all, has to be β^* as given by (31): for $\beta > \beta^*$ McCoy would do better by choosing alternatively a small β to make Hatfield act honestly, and for $\beta < \beta^*$ he could always do a little better by choosing a larger β, closer to β^*. So the only maximum of his payoff curve, as seen in Fig. 7, is at $\beta = \beta^*$. However, as Eq. (34) shows, the maximum exists only if Hatfield's equilibrium strategy is such that $t^*(\beta^*) = 0$. The

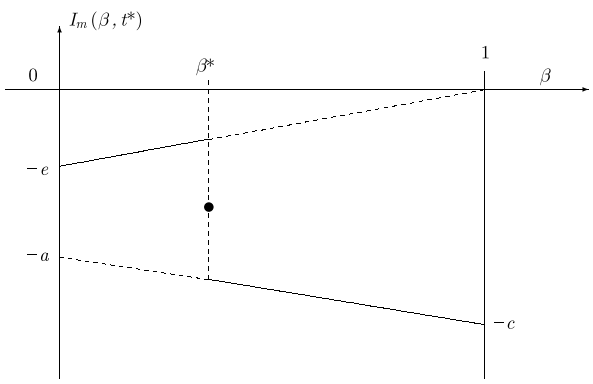

Inspection Games, Figure 7
McCoy's payoff as a function of β according to (34). The ● indicates the equilibrium payoff of the simultaneous game, Eq. (28)

unique subgame perfect equilibrium strategies must therefore be (β^*, t^*) given by (31) and (33), with payoffs

$$I_{\mathrm{M}}(\beta^*, t^*) = -c \cdot \frac{d}{b+d}\,, \quad I_{\mathrm{H}}(\beta^*, t^*) = 0\,. \tag{35}$$

Thus McCoy's equilibrium strategy is the same as before, as is Hatfield's payoff, the decisive difference being that Hatfield does not take more than his share of the water.

Remarks The leadership concept was first introduced by von Stackelberg [49] in the context of economic theory, well before game theory became a scientific discipline. In game-theoretic terminology Schelling [44] probably was the first formulate its importance:

> A strategic move is one that influences the other person's choice in a manner favorable to one's self, by affecting the other person's expectations on how one's self will behave. One constrains the partner's choice by constraining one's own behavior. The object is to set up for one's self and communicate persuasively to the other player a mode of behavior (including conditional responses to the other's behavior) that leaves the other a simple maximization problem whose solution for him is the optimum for one's self, and to destroy the other's ability to do the same.

Simaan and Cruz [45] and Wölling [52] refined the concept and formulated conditions for the existence of Nash equilibria in leadership games. Maschler was the first to apply the leadership concept to inspection games [32,33]. Later on it was widely used in the analysis of IAEA verification procedures, in particular for variable sampling inspection problems, see Avenhaus et al. [6].

The importance of deterring the inspectee from illegal behavior, or more positively, of inducing him to behave legally, depends on the specific nature of the problem. For example in the ticket control problem of Subsect. "Passenger Ticket Control", maximization of intake – fares plus fines – is no doubt the highest priority of the transit authority, even though the inspector leadership concept would work here as well, at least in theory. In principal agent models the situation is similar. In the context of arms control and disarmament, on the other hand, deterrence is fundamental: the community of States party to such an agreement have a vital interest that all members adhere to its provisions. There exists a large literature on the subject; a comprehensive survey of the leadership concept in game theory in general has been been given by Wölling [52].

Future Directions

We hope that, with our chosen examples, we have given a representative, although certainly not exhaustive, overview of the concepts and models making up the area of inspection games. At the same time we have tried to give some idea of its wide range of applications.

Of course we cannot predict future developments in the field. To be sure, there are still unsolved mathematical problems associated with present inspection models, in particular in arms control and disarmament. For example in Subsect. "Diversion of Nuclear Material" it was pointed out that near real time accountancy poses fundamental difficulties that have not yet been solved satisfactorily. Active research is proceeding and interesting results may be expected.

As mentioned at the outset, in the area of environmental control the number of published investigations is surprisingly small. With the growing awareness of the importance of international agreement on environmental protection the need for effective and efficient control mechanisms will become more and more apparent. Here we expect that the inspection game approach, especially as a means of structuring verification systems, can and will play a useful role. As Barry O'Neill concludes his examination of game theory in peace and war [36]:

> … game theory clarifies international problems exactly because they are more complicated. […] The contribution of game models is to sort out concepts and figure out what the game might be.

Bibliography

1. Avenhaus R (1997) Entscheidungstheoretische Analyse der Fahrgast-Kontrollen. Der Nahverkehr 9:27
2. Avenhaus R, Canty MJ (1989) Re-examination of the IAEA formula for stratified attribute sampling. Proc 11th ESARDA Symposium, JRC, Ispra, pp 351–356
3. Avenhaus R, Canty MJ (1996) Compliance Quantified. Cambridge University Press, Cambridge
4. Avenhaus R, Canty MJ (2005) Playing for time: a sequential inspection game. Eur J Oper Res 167(2):474–492
5. Avenhaus R, Jaech JL (1981) On subdividing material balances in time and/or space. J Inst Nucl Mater Manag IV(3):24–33
6. Avenhaus R, Okada A, Zamir S (1991) Inspector leadership with incomplete information. In: Selten R (ed) Game equilibrium models, vol IV. Springer, Heidelberg, pp 319–361
7. Avenhaus R, Canty MJ, Kilgour DM, von Stengel B, Zamir S (1996) Inspection games in arms control. Eur J Oper Res 90:383–394
8. Avenhaus R, von Stengel B, Zamir S (2002) Inspection games. In: Aumann R, Hart S (ed) Handbook of game theory. Elsevier, Amsterdam, pp 1947–1987

9. Baiman S (1982) Agency research in managerial accouting: a survey. J Account Lit 1:154–213
10. Baston VJ, Bostock FA (1991) A remark on the customs smuggler game. Nav Res Logist 41:287–293
11. Bierlein D (1968) Direkte Überwachungssysteme. Oper Res Verfahr 6:57–68
12. Bierlein D (1969) Auf Bilanzen und Inventuren basierenden Safeguards-Systeme. Oper Res Verfahr 6:36–43
13. Borch K (1990) Economics of insurance. North-Holland, Amsterdam
14. Cavasoglu H, Raghunatahan S (2004) Configuration of detection software: a comparison of decision and game theory. Decis Anal 1:131–148
15. Cook J, Nadeau L, Thomas LC (1997) Does cooperation in auditing matter? a comparison of a non-cooperative and a cooperative game model of auditing. Eur J Oper Res 103:470–482
16. Derman C (1961) On minimax surveillance schedules. Nav Res Logist Quaterly 8:415–419
17. Diamond H (1982) Minimax policies for unobservable inspections. Math Oper Res 7(1):139–153
18. Dresher M (1962) A sampling inspection problem in arms control agreements: a game theoretical analysis. Memorandum RM-2972-ARPA. RAND Corporation, Santa Monica
19. Dye RA (1986) Optimal monitoring policies in agencies. RAND J Econ 17:339–350
20. Ferguson TS, Melolidakis C (1998) On the inspection game. Nav Res Logist 45:327–334
21. Garnaev AY (1991) A generalized inspection game. Nav Res Logist 28:171–188
22. Goutal P, Garnaev A, Garnaeva G (1997) On the infiltration game. Int J Game Theory 26(2):215–221
23. Höpfinger E (1971) A game-theoretic analysis of an inspection problem. University of Karlsruhe. (unpublished manuscript)
24. Höpfinger E (1974) Zuverlässige Inspektionsstrategien. Z Wahrscheinlichkeitstheorie Verwandte Geb 31:35–46
25. Hozaki R, Kuhdoh D, Komiya T (2006) An inspection game: taking account of fulfillment probabilities of players. Nav Res Logist 53:761–771
26. IAEA (1972) The structure and content of agreements between the agency and states required in connection with the treaty on the non-proliferation of nuclear weapons. IAEA, Vienna, INF/CIRC 153 (corrected)
27. IAEA (1997) Model protocol additional to the Agreement(s) between state(s) and the international atomic energy agency for the application of safeguards. IAEA, Vienna, INF/CIRC 140
28. Kanodia CS (1985) Stochastic and moral hazard. J Account Res 23:175–293
29. Kilgour DM (1992) Site selection for on-site inspection in arms control. Arms Control 13:439–462
30. Krieger T (2008) On the asymptotic behavior of a discrete time inspection game. Math Model Anal 13(1):37–46
31. Kuhn HW (1953) Extensive games and the problem of information. In: Kuhn HW, Tucker AW (eds) Contributions to the theory of Games, vol II. Princeton University Press, Princeton, pp 193–216
32. Maschler M (1966) A price leadership method for solving the inspector's non-constant-sum game. Nav Res Logist 13:11–33
33. Maschler M (1967) The inspector´s non-constant-sum-game: its dependence on a system of detectors. Nav Res Logist 14:275–290
34. Morris P (1994) Introduction to game theory. Springer, New York
35. Nash JF (1951) Non-cooperative games. Ann Math 54:286–295
36. O'Neill B (1994) Game theory models of peace and war. In: Aumann R, Hart S (eds) Handbook of game theory. Elsevier, Amsterdam, pp 995–1053
37. Ostrom E, Gardner R, Walker J (1994) Rules, games and common pool resources. University of Michigan Press, Ann Arbor
38. Owen G (1968) Game theory. W. B. Saunders, Philadelphia
39. Pavlovic L (2002) More on the search for an infiltrator. Nav Res Logist 49:1–14
40. Rinderle K (1996) Mehrstufige sequentielle Inspektionsspiele mit statistischen Fehlern erster und zweiter Art. Kovac, Hamburg
41. Rohatgi VK (1976) An introduction to probability theory and mathematical statistics. Wiley, New York
42. Rothenstein D, Zamir S (2002) Imperfect inspection games over time. Ann Oper Res 109:175–192
43. Sakaguchi M (1994) A sequential game of multi-opportunity infiltration. Math Jpn 39:157–166
44. Schelling TC (1960) The strategy of conflict. Harvard University Press, Cambridge
45. Simaan M, Cruz JB (1973) On the Stackelberg strategy in nonzero-sum games. J Optim Theory Appl 11(5):533–555
46. Stewart KB (1971) A cost-effectiveness approach to inventory verification. Proc of the IAEA Symposium on Safeguards Techniques, vol II. International Atomic Energy Agency, Vienna, pp 387–409
47. Thomas MU, Nisgav Y (1976) An infiltration game with time-dependent payoff. Nav Res Logist 23:297–320
48. von Neumann J, Morgenstern O (1947) Theory of games and economic behavior. Princeton University Press, Princeton
49. von Stackelberg H (1934) Marktform und Gleichgewicht. Springer, Vienna
50. von Stengel B (1991) Recursive inspection games, Report No. S 9106. Computer Science Faculty, Armed Forces University Munich
51. Wilks TJ, Zimbelman MF (2004) Using game theory and strategic reasoning concepts to prevent and detect fraud. Account Horiz 18(3):173–184
52. Wölling A (2002) Das Führerschaftsprinzip bei Inspektionsspielen. Kovac, Hamburg

Intelligent Control

Clarence W. de Silva
Department of Mechanical Engineering,
University of British Columbia, Vancouver, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
Soft Computing
Fuzzy Logic and Fuzzy Sets
Composition and Inference

Fuzzy Control
Future Directions
Bibliography

Glossary

Adaptive system A system that can be automatically modified or reorganized to meet a set of performance specifications.

Artificial intelligence (AI) Computer-generated intelligence or characteristic (externally displayed) of a man-made system that behaves like a naturally intelligent system (biological system) to some degree.

Back-propagation This is a "learning by example" or "supervised learning" technique of a neural network. Using previously obtained input-output data (training data) of a system, a neural network is trained by assigning the training output data to the network output and then computing the corresponding signals in the previous layers of the network, sequentially. The error is computed and minimized in this manner.

Defuzzification The process of converting a fuzzy quantity into a crisp (non-fuzzy) quantity. One method, called the "centroid defuzzification" determines the centroid of the membership function of the fuzzy quantity and uses it as the crisp representation of the fuzzy quantity.

Evolutionary computing A non-analytical optimization technique that mimics biological evolution. Based on an optimization function (fitness function), solutions are selected by "mating" good solutions to create better solutions, and discarding poor solutions.

Expert systems Computer-based intelligent systems which take in data, match to a knowledge base, and generate inferences (solutions, advice, prescriptions, predictions, etc.) in a manner that somewhat mimics a human expert. Expert systems are developed by gathering human expertise in a specific domain and properly representing it in a computer system with a human interface.

Fuzzification The process of converting a crisp quantity into a fuzzy quantity. A membership function is determined to "fuzzify" the crisp quantity. For example, the crisp quantity may form the peak value or centriod of a membership function of sufficient width.

Fuzzy control A model-free control technique where control knowledge is represented by "if-then" statements that contain qualitative or "fuzzy" terms such as "small" and "fast" as present in human statements. System's outputs are observed (or measured) and "matched" with the control knowledge base to arrive at control actions.

Fuzzy logic A logic that is more generalized than binary crisp logic. Instead of the two states in binary logic, multiple states are possible, with a degree of overlap among states. Here, for example, the state of "warm" and the state of "hot" have some overlap, as is common with human perception.

Fuzzy set In the same manner that binary logic and Boolean sets go together, fuzzy logic and fuzzy sets are related. A fuzzy set has a non-crisp boundary. Elements that fall on the boundary have some level of presence within the set and a complementary level of presence outside the set.

Genetic algorithms A non-analytical optimization technique that mimics biological evolution. Based on an optimization function (fitness function), solutions are selected by "mating" good solutions to create better solutions, and discarding poor solutions.

Intelligent control A control approach that mimics a human who has expertise to generate a suitable control action for a particular system. A common approach uses a control knowledge base, and an inferencing mechanism which matches observed/measured information with the knowledge base to generate the control action (controller output or control input to the system).

Intelligent machine Machine with an artificial (computer-generated) brain, so that it can behave like a naturally intelligent biological system. It uses techniques of artificial intelligence (AI) for this purpose. Often, the term "machine" refers to a computer. More appropriately, the machine is a physical device that performs an engineering task, and its brain is a computer with AI software.

Knowledge-based system An artificially intelligent system that uses a knowledge base (KB) to represent expert knowledge in a particular application domain. A decision making method (inference engine) is used to "match" data or observed information with the KB to generate inferences.

Membership function A function that represents a fuzzy set. In this function, the membership of a quantity in a fuzzy set is represented by a numerical value between 0 and 1. Set elements that are clearly within the set are represented by a membership grade of 1 and elements that are clearly outside the set are represented by a membership grade of 0. Elements that lie on the set boundary have membership grades between 0 and 1, as determined by the degree of membership in the set.

Neural networks Massively parallel networks of artificial neurons that represent highly nonlinear systems or processes, without using analytical models. By adjusting the network parameters, the behavior of the network is made to resemble the actual system. It somewhat resembles the activity of a biological brain.

Probabilistic system A system that possesses a degree of randomness or uncertainty and uses methods that involve probability distribution functions to generate decisions, actions, or estimates.

Soft computing An approach that uses one or more techniques of fuzzy logic, neural, networks, evolutionary computing, and probabilistic methods to perform numerical operations by somewhat mimicking biological systems.

Definition of the Subject

An intelligent controller may be interpreted as a computer-based controller that can somewhat "emulate" the reasoning procedures of a human expert in the specific area of control, to generate the necessary control actions. Here, techniques from the field of *artificial intelligence (AI)* are used for the purpose of acquiring and representing knowledge and for generating control decisions through an appropriate reasoning mechanism. With steady advances in the field of AI, especially pertaining to the development of practical *expert systems* or *knowledge systems*, there has been a considerable interest in using AI techniques for controlling complex processes. Complex engineering systems use intelligent control to cope with situations where conventional control techniques are not effective.

Intelligent control depends on efficient ways of representing and processing the control knowledge [5]. Specifically, a knowledge base has to be developed and a technique of reasoning and making "inferences" has to be available. Knowledge-based intelligent control relies on knowledge that is gained by intelligently observing, studying, or understanding the behavior of a plant, rather than explicitly modeling the plant, to arrive at the control action. In this context, it also heavily relies on the knowledge of experts in the domain, and also on various forms of general knowledge. Modeling of the plant is implicit here. Soft computing is an important branch of study in the area of intelligent and knowledge-based systems. It has effectively complemented conventional AI in the area of machine intelligence (computational intelligence). Fuzzy logic, probability theory, neural networks, and genetic algorithms are cooperatively used in soft computing for knowledge representation and for mimicking the reasoning and decision-making processes of a human. Decision making with soft computing involves *approximate reasoning*, and is commonly used in intelligent control. This chapter presents the background theory, concepts, and development of intelligent control, with a particular emphasis on fuzzy logic control.

Introduction

Future generations of industrial machinery, plants, and decision support systems may be expected to carry out round-the-clock operation, with minimal human intervention, in manufacturing products or providing services. It will be necessary that these systems maintain consistency and repeatability of operation and cope with disturbances and unexpected variations within the system, its operating environment, and performance objectives. In essence, these systems should have the capability to accommodate rapid reconfiguration and adaptation. For example, a production machine should be able to quickly cope with variations ranging from design changes for an existing product to the introduction of an entirely new product line. This will call for tremendous flexibility and some level of autonomous operation in automated machines, which translate into a need for a higher degree of intelligence in the supporting devices. Smart systems will exhibit an increased presence and significance in a wide variety of engineering applications. Products with a "brain" are found, for example, in household appliances, consumer electronics, space technology, transportation systems, industrial processes, manufacturing systems, and services. There is clearly a need to incorporate a greater degree of intelligence and a higher level of autonomy into automated machines and systems. This will require the appropriate integration of such devices as sensors, actuators, and controllers, which themselves may have to be "intelligent" and furthermore, appropriately distributed throughout the system. Design, development, production, and operation of intelligent machines have been possible today through ongoing research and development in the field of intelligent systems and intelligent control [3,4,6,10,13].

Soft Computing

Soft computing is an important branch of study in the area of intelligent and knowledge-based systems. It has effectively complemented conventional AI in the area of machine intelligence (computational intelligence). Human reasoning is predominantly approximate, qualitative, and "soft." Humans can effectively handle incomplete, imprecise, and fuzzy information in making intelligent decisions. Fuzzy logic, probability theory, neural networks,

and genetic algorithms are cooperatively used in soft computing for knowledge representation and for mimicking the reasoning and decision-making processes of a human [5,7]. Quite effective are the mixed or hybrid techniques, which synergistically exploit the advantages of two or more of these areas. Decision making with soft computing involves *approximate reasoning*. Now we will give an introduction to the subject of soft computing. Fuzzy logic and its use in intelligent control will be covered in detail in subsequent sections.

Fuzzy Logic

Fuzzy logic is useful in representing human knowledge in a specific domain of application and in reasoning with that knowledge to make useful inferences or actions [14]. The conventional binary (bivalent) logic is crisp and allows for only two states. This logic cannot handle fuzzy descriptors, examples of which are "fast" which is a *fuzzy quantifier* and "weak" which is a *fuzzy predicate*. They are generally qualitative, descriptive, and subjective and may contain some overlapping degree of a neighboring quantity, for example, some degree of "slowness" in the case of the fuzzy quantity "fast." Fuzzy logic allows for a realistic extension of binary, crisp logic to qualitative, subjective, and approximate situations, which often exist in problems of intelligent machines where techniques of artificial intelligence are appropriate.

In fuzzy logic, the knowledge base is represented by if-then rules of fuzzy descriptors. Consider the general problem of approximate reasoning. In this case the knowledge base K is represented in an "approximate" form, for example, by a set of if-then rules with *antecedent* and *consequent* variables that are fuzzy descriptors. First, the data D are preprocessed according to

$$F_D = FP(D) \tag{1}$$

which, in a typical situation, corresponds to a data abstraction procedure called "fuzzification" and establishes the membership functions or membership grades that correspond to D. Then for a fuzzy knowledge base F_K, the fuzzy inference F_I is obtained through fuzzy-predicate approximate reasoning, as denoted by

$$F_I = F_K \circ F_D \,. \tag{2}$$

This uses a *composition* operator "$\circ$" for fuzzy matching of data (D) with the knowledge base (K), and making inferences (I) on that basis.

Fuzzy logic is commonly used in "intelligent" control of processes and machinery. In this case the inferences of a fuzzy decision making system are the control inputs to the process. These inferences are arrived at by using the process responses as the inputs (context data) to the fuzzy decision-making system.

Neural Networks

Artificial neural networks (NN) are massively connected networks of computational "neurons," and represent parallel-distributed processing structures. The inspiration for NN has come from the biological architecture of neurons in human brain. A key characteristic of neural networks is their ability to approximate arbitrary nonlinear functions. Since machine intelligence involves a special class of highly nonlinear decision making, neural networks would be effective there. Furthermore, the process of approximation of a nonlinear function (i. e., system identification) by interacting with a system and employing data on its behavior, may be interpreted as "learning." Through the use of neural networks, an intelligent system would be able to learn and perform high-level cognitive tasks. For example, an intelligent system would only need to be presented with a goal; it could achieve its objective through continuous interaction with its environment and evaluation of the responses by means of neural networks [1].

A neural network consists of a set of nodes, usually organized into layers, and connected through weight elements called synapses. At each node, the weighted inputs are summed (aggregated), thresholded, and subjected to an activation function in order to generate the output of that node. These operations are shown in Fig. 1. The analogy to the operations in a biological neuron is highlighted. Specifically, in a biological neuron, the dendrites receive information from other neurons. The soma (cell body) collects and combines this information, which is transmitted to other neurons using a channel (tubular structure) called axon. This biological analogy, apart from the abilities to learn by example, approximation of highly nonlinear functions, massive computing power, and memory, may be a root reason for inherent "intelligence" in a neural network. If the weighted sum of the inputs to a node (neuron) exceeds a threshold value w_0, then the neuron is fired and an output $y(t)$ is generated according to

$$y(t) = \Psi \sum_{i=1}^{n} w_i x_i - w_0 \,, \tag{3}$$

where x_i are neuron inputs, w_i are the synaptic weights, and $\Psi[.]$ is the activation function.

As indicated in Fig. 2, there are two main classes of neural networks known as feedforward networks (or,

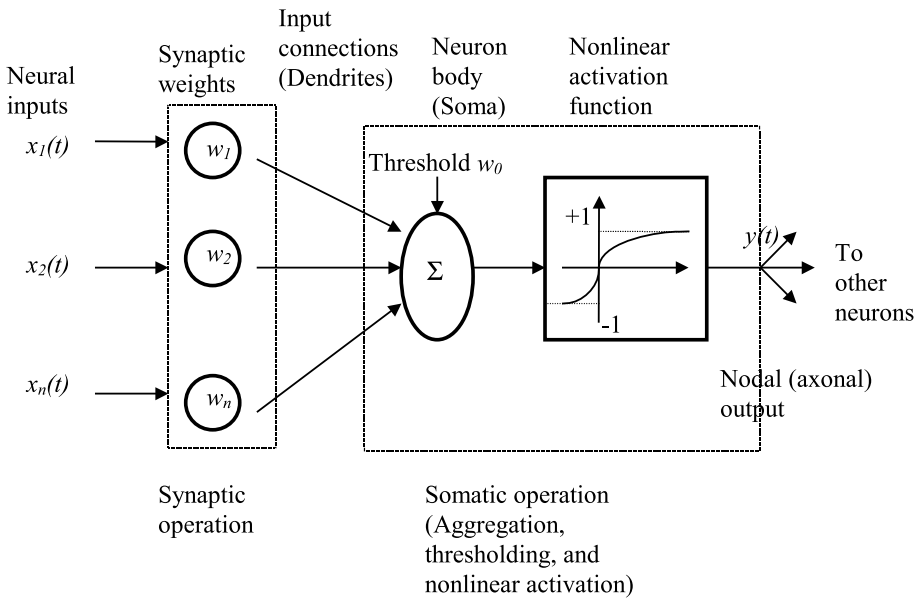

Intelligent Control, Figure 1
The operations at a node of a neural network

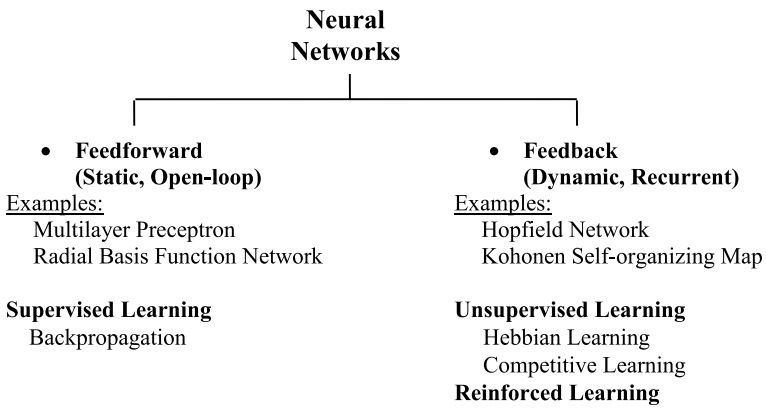

Intelligent Control, Figure 2
Classification of neural networks

static networks) and feedback networks (or, recurrent networks). In feedfoward network, an example of which is a multiplayer preceptron, the signal flow from a node to another node takes place in the forward direction only. There are no feedback paths. Figure 3 shows a multiplayer preceptron consisting of an input layer, a hidden layer, and an output layer. Another example of feedforward network is the radial basis function network. Here there are only three layers. Furthermore, only the hidden layer uses nonlinear activation functions in its nodes. These functions are called radial basis functions, the Gaussian distribution function being a popular example. These functions form the basis for the capability of the NN to approximate any nonlinear function. In a feedforward neural network, learning is achieved through example. This is known as supervised learning. Specifically, first a set of input-output data of the actual process is determined (e. g., by measurement). The input data are fed into the NN. The network output is compared with the desired output (experimental data) and the synaptic weights of the NN are adjusted using a gradient (steepest descent) algorithm until the desired output is achieved.

In a feedback NN, the outputs of one or more nodes (say, in the output layer) are fed back to one or more nodes in a previous layer (say hidden layer or input layer) or even to the same node. The feedback provides the capability of

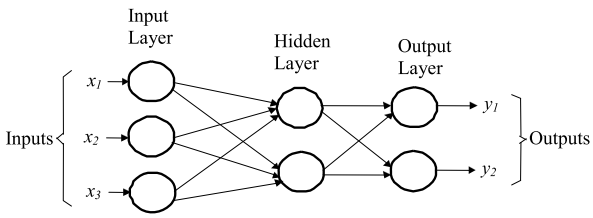

Intelligent Control, Figure 3
A feedforward neural network (Multilayer perceptron)

"memory" to the network. An example is the Hopfield network. It consists of two layers: the input layer and the Hopfield layer. Each node in the input layer is directly connected to only one node in the Hopfield layer. The outputs of the network are fed back to the input nodes via a time delay (providing memory) and synaptic weights. Nodes in the Hopfield layer have nonlinear activation functions such as sigmoidal functions.

Feedback neural networks commonly use unsupervised learning algorithms. In these learning schemes, the synaptic weights are adjusted based on the input values to the network and the reactions of individual neurons, and not by comparing the network output to the desired output data. Unsupervised learning is called self-organization (or open-loop adaptation), because the output characteristics of the network are determined internally and locally by the network itself, without any data on desired outputs. This type of learning is particularly useful in pattern classification and grouping of data. Hebbian learning and competitive learning are examples of unsupervised learning algorithms. In the Hebbian learning algorithm, the weight between a neuron and an input is strengthened (increased) if the neuron is fired by the input. In competitive learning, weights are modified to enhance a node (neuron) having the largest output. An example is the Kohonen network, which uses a winner-takes-all approach. A Kohonen network has two layers, the input layer and the output layer (Kohonen layer). In the operation of the network, the node in the Kohonen Layer with weights that most closely resemble the current input, is assigned an output of 1 (the winner), and the outputs of all the remaining nodes are set to zero. In this manner, the input nodes organize themselves according to the pattern of the input data while the output nodes compete among themselves to be activated.

Genetic Algorithms

Genetic algorithms (GA) are derivative-free optimization techniques, which can evolve through procedures analogous to biological evolution. Genetic algorithms belong to the area of evolutionary computing. They represent an optimization approach where a search is made among a set of solutions (solution space) to "evolve" a solution algorithm, which will retain the "most fit" solutions, by using a procedure that is analogous to biological evolution through natural selection, crossover, and mutation. In the present context of intelligent machines, intellectual fitness rather than physical fitness is what is important for the evolutionary process. Evolutionary computing can play an important role in the development of an optimal and self-improving intelligent machine [5,7].

Evolutionary computing has the following characteristics:

1. It is based on multiple searching points or solution candidates (population based search).
2. It uses evolutionary operations such as selection, crossover and mutation.
3. It is based on probabilistic operations.

The basic operations of a genetic algorithm are indicated in Fig. 4. The algorithm works with a population of individuals, each representing a possible solution to a given problem. Each individual is assigned a fitness score according to how good its solution is to the problem. The highly fit (in an intellectual sense) individuals are given opportunities to reproduce by crossbreeding with other individuals in the population. This produces new individuals as offspring, who share some features taken from each parent. The least fit members of the population are less likely to get selected for reproduction and will eventually die out. An entirely new population of possible solutions is produced in this manner, by mating the best individuals (i. e., individuals with best solutions) from the current

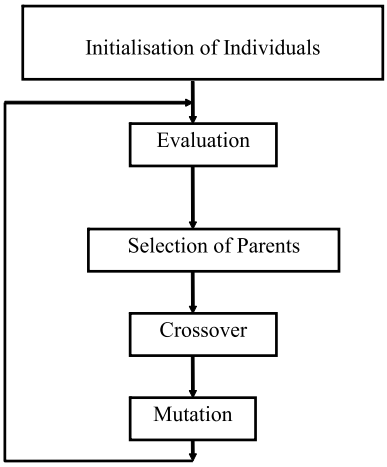

Intelligent Control, Figure 4
The key operations of a genetic algorithm

generation. The new generation will contain a higher proportion of the characteristics possessed by the "fit" members of the previous generation. In this way, over many generations, desirable characteristics are spread throughout the population, while being mixed and exchanged with other desirable characteristics, in the process. By favoring the mating of the individuals who are more fit (i. e., who can provide better solutions), the most promising areas of the search space would be exploited. A GA determines the next set of searching points using the fitness values of the current searching points, which are widely distributed throughout the searching space. It uses the mutation operation to escape from a local minimum.

Two important activities of a GA are selection and reproduction. Selection is the operation which will choose parent solutions. New solution vectors in the next generation are calculated from them. Since it is expected that better parents generate better offspring, parent solution vectors that possess higher fitness values will have a higher probability of selection. There are several methods of selection. In the method of roulette wheel selection, the probability of winning is proportional to the area rate of a chosen number on a roulette wheel. In this manner, the selection procedure assigns a selection probability to individuals in proportion to their fitness values. In the elitist strategy, the best parents are copied into the next generation. This strategy prevents the best fitness value of the offspring generation from becoming worse than that in the present generation.

During the reproductive phase of a GA, individuals are selected from the population and recombined, producing offspring, which in turn will make up the next generation. Parents are selected randomly from the population using a scheme that favors the individuals who are more fit. After two parents are selected, their chromosomes are recombined using the mechanism of crossover and mutation. Crossover takes two individuals and cuts their chromosome strings at some randomly chosen position to produce two "head" segments and two "tail" segments. The tail segments are then swapped over to produce two new full-length chromosomes. Each of the two offspring will inherit some genes from each parent. This is known as a single-point crossover. Crossover is not usually applied to all pairs of individuals that are chosen for mating. A random choice is made, where the likelihood of the crossover being applied is typically between 0.6 and 1.0. Mutation is applied individually to each child, after crossover. It randomly alters each gene at a very low probability. Mutation provides a small degree of random search and helps ensure that every point in the search space has some probability of being examined.

Probabilistic Reasoning

Uncertainty and the associated concept of probability are linked to approximation. One can justify that probabilistic reasoning should be treated within the area of soft computing [5,7]. Probabilistic approximate reasoning may be viewed in an analogous manner to fuzzy-logic reasoning, considering uncertainty in place of fuzziness as the concept of approximation that is applicable. Probability distribution/density functions are employed in place of membership functions. The formula of knowledge-based decision making that corresponds to Eq. (2), in this case, depends on the specific type of probabilistic reasoning that is employed. The *Bayesian approach* is commonly used. This may be interpreted as a classification problem.

Suppose that an observation d is made, and it may belong to one of several classes c_i. The Bayes' relation states:

$$\max_i P(c_i|d) = \frac{P(d|c_i) \bullet P(c_i)}{P(d)} , \tag{4}$$

where

$\boldsymbol{P(c_i|d)} =$ given that the observation is d, the probability that it belongs to class c_i (the a posteriori *conditional probability*)

$\boldsymbol{P(d|c_i)} =$ given that the observation belongs to the class c_i, the probability that the observation is d (the *class conditional probability*)

$\boldsymbol{P(c_i)} =$ the probability that a particular observation belongs to class c_i, without knowing the observation itself (the a priori *probability*)

$\boldsymbol{P(d)} =$ the probability that the observation is d without any knowledge of the class.

In the Bayes' decision-making approach, for a given observation (data) d, the posteriori probabilities $P(c_i|d)$ are computed for all possible classes ($i = 1, 2, \ldots, n$), using Eq. (4). The class that corresponds to the largest of these a posteriori probability values is chosen as the class of d, thereby solving the classification problem. The remaining $n - 1$ a posteriori probabilities represent the error in this decision.

Note the analogy between Eqs. (4) and (2). Specifically, $P(d)$ represents the "preprocessed" probabilistic data that correspond to the observation d. The knowledge base itself constitutes the two sets of probabilities:

1. $P(d|c_i)$ of occurrence of data d if the class (decision) is c_i, $i = 1, 2, \ldots, n$
2. $P(c_i)$ of class (decision) c_i, $i = 1, 2, \ldots, n$ without any knowledge of the observation (data) itself.

The knowledge-base matching is carried out, in this case, by computing the expression on the right side of Eq. (4)

Intelligent Control, Table 1
Techniques of computational intelligence

Technique	Characteristic	A Popular Analogy
Fuzzy Logic	Uses fuzzy rules and approximate reasoning	Human knowledge
Neural Networks	Network of massively connected nodes	Neuron structure in brain
Genetic Algorithms	Derivative -free optimization	Biological evolution
Probability	Incorporates uncertainty in predicting future events	Random action of a human
Conventional AI	Symbolic processing of information	Symbolic languages

for all possible i and then picking out the maximum value. Application areas of probabilistic decision making include forecasting, signal analysis and filtering, and parameter estimation and system identification.

Techniques of soft computing are powerful by themselves in achieving the goals of machine intelligence. Furthermore, they have a particular appeal in view of the biological analogies that exist. To summarize the biological analogies of fuzzy, neural, genetic, and probabilistic approaches: fuzzy techniques attempt to approximate human knowledge and the associated reasoning process; neural networks are a simplified representation of the neuron structure of a brain; genetic algorithms follow procedures that are crudely similar to the process of evolution in biological species; and probabilistic techniques can analyze random future action of a human. This is no accident because in machine intelligence it is the behavior of human intelligence that would be mimicked. Popular (e. g., biological) analogies and key characteristics of several techniques of machine intelligence are listed in Table 1 Conventional AI, which typically uses symbolic and descriptive representations and procedures for knowledge-based reasoning, is listed as well for completeness.

Fuzzy Logic and Fuzzy Sets

In the crisp Boolean logic, truth is represented by the state 1 and falsity is by the state 0. Boolean algebra and crisp logic have no provision for approximate reasoning. Fuzzy logic is an extension of crisp bivalent (two-state) logic in the sense that it provides a platform for handling approximate knowledge. Fuzzy logic is based on fuzzy set theory in a manner similar to how crisp bivalent logic is based on crisp set theory. Fuzzy logic provides an approximate yet practical means of representing knowledge regarding a system (e. g., describing the behavior of the system) that is too complex or ill-defined and not easy to tackle using precise mathematical means. The approach also provides a means of making inferences using that knowledge, which can be used in making correct decisions regarding the system and for carrying out appropriate actions. In particular, human-originated knowledge can be effectively handled using fuzzy logic. As the complexity of a system increases, the ability to develop precise analytical models of the system diminishes until a threshold is reached beyond which analytical modeling becomes intractable. Under such circumstances, precise model-based decision making is not practical. Fuzzy knowledge-based decision making is particularly suitable then [9].

This section introduces fuzzy logic, which uses the concept of fuzzy sets. Applications of soft computing and particularly fuzzy knowledge-based systems rely on the representation and processing of knowledge using fuzzy logic. In particular, the *compositional rule of inference*, as presented in the subsequent section, is what is applied in decision-making with fuzzy logic. Some popular applications of fuzzy decision making and intelligent control are given next.

Process Temperature Control (OMRON): This is a proportional-integral-derivative (PID) controller integrated with a fuzzy controller. When temperature control is initiated, only the PID action is present. If the temperature overshoots the set point, the fuzzy controller takes over. This control system responds well to disturbance.

Air Conditioner (Mitsubishi): Conventional air conditioning systems use on-off controllers. Specifically when the temperature drops below a preset level, the unit is automatically turned off. When the temperature rises above a preset level, the unit is turned on. The former preset value is slightly lower than the latter preset value, providing a dead zone, so that high-frequency on-off cycling (chatter) is avoided. The thermostat in the system controls the on-off action. For example, "when the temperature rises to 25°C, turn on the unit, and when the temperature falls to 20°C, turn off the unit." The Mitsubishi air conditioner is controlled by using fuzzy rules such as: "If the ambient air is getting warmer, turn the cooling power up a little; If the air is getting chilly, turn the power down moderately," etc.

The operation becomes smoother as a result. This results in less wear and tear of the air conditioner, more consistent comfortable room temperatures, less noise, and increased efficiency (energy savings).

Vacuum Cleaner (Panasonic): Characteristics of the floor and the amount of dust are sensed by an infrared sensor, and the microprocessor selects the appropriate power by fuzzy control according to these characteristics. The floor characteristics include the type and nature (hardwood, cement, tile, carpet softness, carpet thickness, etc.). The changing pattern of the amount of dust passing through the infrared sensor is established as well. The microprocessor establishes the appropriate setting of the vacuum head and the power of the motor, using a fuzzy control scheme. Red and green lamps of the vacuum cleaner show the amount of dust left on the floor.

Automatic Transmission System (Nissan, Subaru, Mitsubishi): In a conventional automatic transmission system, electronic sensors measure the vehicle speed and throttle opening, and gears are shifted based on the predetermined values of these variables. According to Nissan, this type of system is incapable of uniformly providing satisfactory control performance to a driver because it typically provides only three different shift patterns. The fuzzy control transmission senses several variables including vehicle speed and acceleration, throttle opening, the rate of change of throttle opening, engine load, and driving style. Each sensed value is given a weight, and a fuzzy aggregate is calculated to decide whether to shift gears. This controller is said to be more flexible, smooth, and efficient, providing better performance. Also, an integrated system developed by Mitsubishi uses fuzzy logic for active control of the suspension system, four-wheel drive (traction), steering, and air conditioning.

Washing Machine (Matsushita, Hitachi): The control system senses both quality and quantity of dirt, load size, and fabric type, and adjusts the washing cycle and detergent amount accordingly. Clarity of water in the washing machine is measured by light sensors. At the start of the cycle, dirt from clothes would not have yet reached the water; so light will pass through it easily. The water becomes more discolored as the wash cycle proceeds, and less light will pass through. This information is analyzed and control decisions are made using fuzzy logic.

Camcoder (Panasonic, Sanyo, Fisher, Canon): The intelligent video camera determines the best focus and lighting, particularly when several objects are present in the picture. Also, it has a digital image stabilizer to remove hand jitter. Fuzzy decision making is used in these actions. For example, the following scheme is used for image stabilization. The present image frame is compared with the previous frame from memory. A typically stationary object (e. g., house) is identified and its shift coordinates are computed. This shift is subtracted from the image to compensate for the hand jitter. A fuzzy algorithm provides a smooth control/compensation action.

Elevator Control (Fujitec, Toshiba): A fuzzy scheme evaluates passenger traffic and the elevator variables (load, speed, etc.) to determine car announcement and stopping time. This reduces waiting time and improves the efficiency and reliability of operation.

Handheld Computer (Sony): A fuzzy logic scheme reads hand-written input and interprets the characters for data entry.

Television (Sony): A fuzzy logic scheme uses sensed variables such as ambient lighting, time of day, and user profile, and adjusts such parameters as screen brightness, color, contrast, and sound.

Antilock Braking System (Nissan): The system senses wheel speed, road conditions, and driving pattern, and the fuzzy ABS determines the braking action, with skid control.

Subway Train (Sendai): A fuzzy decision scheme is used by the subway trains in Sendai, Japan, to determine the speed and stopping routine. Ride comfort and safety are used as performance requirements.

Other applications of fuzzy logic include hot water heater (Matsushita), rice cooker (Hitachi), and cement kiln (Denmark). Also see [3,4,6,8,10,11,12,13].

The theory of fuzzy logic can be developed using the concepts of fuzzy sets similar to how the theory of crisp bivalent logic can be developed using the concepts of crisp sets. Specifically, there exists an isomorphism between sets and logic. In view of this, a good foundation in fuzzy sets is necessary to develop the theory of fuzzy logic. In this section, the mathematics of fuzzy sets is presented.

A fuzzy set is a set without clear or sharp (crisp) boundaries or without binary membership characteristics. Unlike an ordinary set where each object (or, element) either belongs or does not belong to the set, a partial membership in a fuzzy set is possible. In other words, there is a "softness" associated with the membership of elements in a fuzzy set. An example of a fuzzy set could be "the set of narrow streets in Vancouver." There are streets that clearly belong to the above set, and others that cannot be considered as narrow. Since the concept of "narrow" is not precisely defined (for example, < 2 m), there will be a "gray"

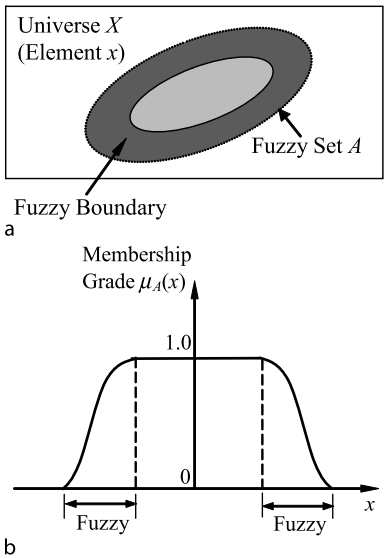

Intelligent Control, Figure 5
a Venn diagram of a fuzzy set. b The membership function of a fuzzy set

zone in the associated set where the membership is not quite obvious. As another example, consider the variable "temperature." It can take a fuzzy value (e. g., cold, cool, tepid, warm, hot). A *fuzzy value* such as "warm" is a *fuzzy-descriptor*. It may be represented by a fuzzy set because any temperature that is considered to represent "warm" belongs to this set and any other temperature does not belong to the set. Still, one cannot realistically identify a precise temperature interval (e. g., 25°C to 30°C), which is a crisp set, to represent warm temperatures.

Let X be a set that contains every set of interest in the context of a given class of problems. This is called the *universe of discourse* (or, simply universe), whose elements are denoted by x. A fuzzy set A in X may be represented by a Venn diagram as in Fig. 5a. Generally, the elements x are not numerical quantities. For analytical convenience, however, the elements x are assigned real numerical values.

Membership Function

A fuzzy set may be represented by a membership function. This function gives the grade (degree) of membership within the set, of any element of the universe of discourse. The membership function maps the elements of the universe on to numerical values in the interval [0, 1]. Specifically,

$$\mu_A(x)\colon X \to [0,1]\,, \tag{5}$$

where $\mu_A(x)$ is the membership function of the fuzzy set A in the universe in X. Stated in another way, fuzzy set if A is a set of ordered pairs:

$$A = \{(x, \mu_A(x))\,; \quad x \in X\,, \quad \mu_A(x) \in [0,1]\}\,. \tag{6}$$

The membership function $\mu_A(x)$ represents the grade of possibility that an element x belongs to the set A. It follows that a membership function is a *possibility function* and not a probability function. A membership function value of zero implies that the corresponding element is definitely not an element of the fuzzy set. A membership function value of unity means that the corresponding element is definitely an element of the fuzzy set. A grade of membership greater than 0 and less than 1 corresponds to a non-crisp (or fuzzy) membership, and the corresponding elements fall on the fuzzy boundary of the set. The closer the $\mu_A(x)$ is to 1 the more the x is considered to belong to A, and similarly the closer it is to 0 the less it is considered to belong to A. A typical membership function is shown in Fig. 5b.

Note: A crisp set is a special case of fuzzy set, where the membership function can take the two values 1 (membership) and 0 (non-membership) only. The membership function of a crisp set is given the special name *characteristic function*.

Fuzzy Logic Operations

It is well known that the "complement", "union", and "intersection" of crisp sets correspond to the logical operations NOT, OR, and AND, respectively, in the corresponding crisp, bivalent logic. Furthermore, it is known that, in the crisp bivalent logic, the union of a set with the complement of a second set represents an "implication" of the first set by the second set. Set inclusion (i. e., extracting a subset) is a special case of implication in which the two sets belong to the same universe. These operations (connectives) may be extended to fuzzy sets for corresponding use in fuzzy logic fuzzy reasoning. For fuzzy sets, the applicable connectives must be expressed in terms of the membership functions of the sets which are operated on. In view of the isomorphism between fuzzy sets and fuzzy logic, both the set operations and the logical connectives can be addressed together. Some basic operations that can be defined on fuzzy sets and the corresponding connectives of fuzzy logic are described next. Several methods are available to define the intersection and the union of fuzzy sets. The classical ones suggested by Zadeh are widely used, because of their simplicity and the analogy with crisp sets (binary logic).

Complement (Negation, NOT) Consider a fuzzy set A in a universe X. Its complement A' is a fuzzy set whose membership function is given by

$$\mu_{A'}(x) = 1 - \mu_A(x) \quad \text{for all } x \in X . \tag{7}$$

The complement in fuzzy sets corresponds to the negation (NOT) operation in fuzzy logic, just as in crisp logic, and is denoted by $\bar{A}$ where A now is a fuzzy logic proposition (or a fuzzy state).

A graphic (membership function) representation of complement of a fuzzy set (or, negation of a fuzzy state) is given in Fig. 6.

Example As an example, consider the fuzzy set of "hot temperatures." This may be represented by the membership function shown using a solid line in Fig. 7. This is the set containing all values of hot temperature in a specified universe. In fuzzy logic then, this membership function can represent the fuzzy logic state "hot" or the fuzzy statement, "the temperature is hot." The complement of the fuzzy set is represented by the dotted line in Fig. 7. This is the set containing all temperature values that are not hot, in the given universe. As before, this membership function can also represent the fuzzy logic state "not hot" or the fuzzy-logic statement "the temperature is not hot."

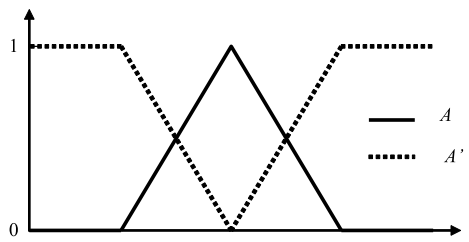

Intelligent Control, Figure 6
Fuzzy-set complement or fuzzy-logic NOT

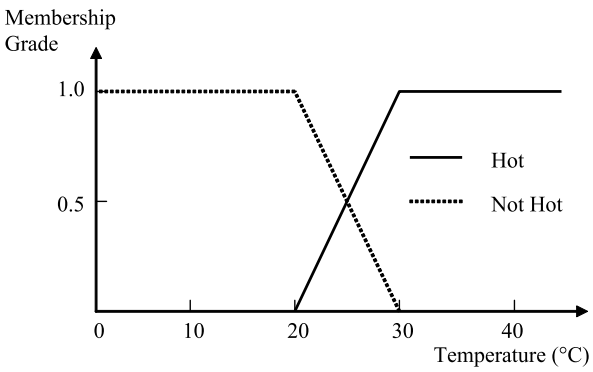

Intelligent Control, Figure 7
An example of fuzzy-logic NOT

Furthermore, the solid line can represent a fuzzy temperature value and the dotted line can represent another (complementary) fuzzy temperature value. In this manner, various concepts of fuzzy sets and fuzzy logic can be quantified by the use of membership functions.

Union (Disjunction, OR) Consider two fuzzy sets A and B in the same universe X. Their union is a fuzzy set containing all the elements from both sets, in a "fuzzy" sense. This set operation is denoted by $\cup$. The membership function of the resulting set $A \cup B$ is given by

$$\mu_{A \cup B}(x) = \max[\mu_A(x), \mu_B(x)] \quad \forall x \in X . \tag{8}$$

The union corresponds to a logical OR operation (called *Disjunction*), and is denoted by $A \vee B$, where A and B are fuzzy states or fuzzy propositions. The rationale for the use of *max* to represent fuzzy-set union is that, because element x may belong to one set or the other, the larger of the two membership grades should govern the outcome (union). Furthermore, this is consistent with the union of crisp sets. Similarly, the appropriateness of using *max* to represent fuzzy-logic operation "OR" should be clear. Specifically, since either of the two fuzzy states (or propositions) would be applicable, the larger of the corresponding two membership grades should be used to represent the outcome. A graphic (membership function) representation of union of two fuzzy sets (or, the logical combination OR of two fuzzy states in the same universe) is given in Fig. 8.

Even though set intersection is applicable to sets in a common universe, a logical "OR" may be applied for concepts in different universes. In particular, when the operands belong to different universes, orthogonal axes have to be used to represent them in a common membership function.

Intersection (Conjunction, AND) Again, consider two fuzzy set A and B in the same universe X. Their intersection is a fuzzy set containing all the elements that are common to both sets, in a "fuzzy" sense. This set operation is

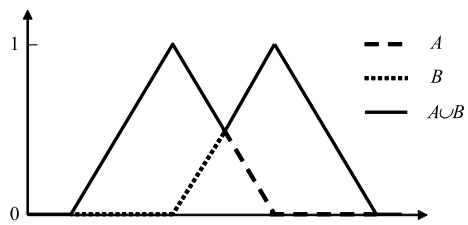

Intelligent Control, Figure 8
Fuzzy-set union or fuzzy-logic OR

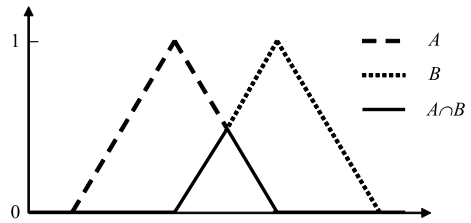

Intelligent Control, Figure 9
Fuzzy-set intersection or fuzzy-logic AND

denoted by $\cap$. The membership function of the resulting set $A \cap B$ is given by

$$\mu_{A\cap B}(x) = \min[\mu_A(x), \mu_B(x)] \quad \forall x \in X\,. \tag{9}$$

The union corresponds to a logical AND operation (called *Conjunction*), and is denoted by $A \wedge B$, where A and B are fuzzy states or fuzzy propositions. The rationale for the use of *min* to represent fuzzy-set intersection is that, because the element x must simultaneously belong to both sets, the smaller of the two membership grades should govern the outcome (intersection).

Furthermore, this is consistent with the intersection of crisp sets. Similarly, the appropriateness of using *min* to represent fuzzy-logic operation "AND" should be clear. Specifically, since both fuzzy states (or propositions) should be simultaneously present, the smaller of the corresponding two membership grades should be used to represent the outcome. A graphic (membership function) representation of intersection of two fuzzy sets (or, the logical combination AND of two fuzzy states in the same universe) is given in Fig. 9.

Implication (If-Then)

An if-then statement (a rule) is called an "implication". In a knowledge-based system, the knowledge base is commonly represented using if-then rules. In particular, a knowledge base in fuzzy logic may be expressed by a set of linguistic rules of the if-then type, containing fuzzy terms. In fact a fuzzy rule is a *fuzzy relation*. A knowledge base containing several fuzzy rules is also a relation, which is formed by combining (aggregating) the individual rules according to how they are interconnected.

Consider a fuzzy set A defined in a universe X and a second fuzzy set B defined in another universe Y. The fuzzy implication "If A then B," is denoted by $A \to B$. Note that in this fuzzy rule, A represents some "fuzzy" situation, and is the *condition* or the *antecedent* of the rule. Similarly, B represents another fuzzy situation, and is the *action* or the *consequent* of the fuzzy rule. The fuzzy rule $A \to B$ is a fuzzy relation. Since the elements of A are defined in X and the elements of B are defined in Y, the elements of $A \to B$ are defined in the *Cartesian product space* $X \times Y$. This is a two-dimensional space represented by two orthogonal axes (x-axis and y-axis), and gives the domain in which fuzzy rule (or fuzzy relation) is defined. Since A and B can be represented by membership functions, and additional orthogonal axis is needed to represent the membership grade.

Fuzzy implication may be defined (interpreted) in several ways. Two definitions of fuzzy implication are:

Method 1:

$$\mu_{A\to B}(x, y) = \min[\mu_A(x), \mu_B(y)] \quad \forall x \in X, \ \forall y \in Y, \tag{10}$$

Method 2:

$$\mu_{A\to B}(x, y) = \min[1, \{1 - \mu_A(x) + \mu_B(y)\}] \quad \forall x \in X\,, \ \forall y \in Y\,. \tag{11}$$

These two methods are approaches for obtaining the membership function of the particular fuzzy relation given by an if-then rule (implication). Note that the first method gives an expression that is symmetric with respect to A and B. This is not intuitively satisfying because "implication" is not a commutative operation (specifically, $A \to B$ does not necessarily satisfy $B \to A$). In practice, however, this method provides a good, robust result. The second method has an intuitive appeal because in crisp bivalent logic, $A \to B$ has the same truth table as [(NOT A) OR B] and hence are equivalent. Note that in Eq. (11), the membership function is upper-bounded to 1 using the *bounded sum* operation, as required (A membership grade cannot be greater than 1). The first method is more commonly used because it is simpler to use and often provides quite accurate results.

An example of fuzzy implication using the first method is shown in Fig. 10. Here the implication from a fuzzy set (a) to another fuzzy set (b) is given by (c).

Composition and Inference

If a fuzzy input is applied to a non-fuzzy (crisp) system, the output will still be fuzzy. This is known as the "extension principle," and is illustrated by an example in Fig. 11. A fuzzy system may be represented by a set of fuzzy if-then rules, which may be aggregated into a single (multivariable) membership function – a fuzzy relation. Application of a fuzzy input to a fuzzy relation in order to infer a fuzzy output is the basis of decision making in a fuzzy knowledge-based system. This idea is known as composition, and is illustrated in Fig. 12.

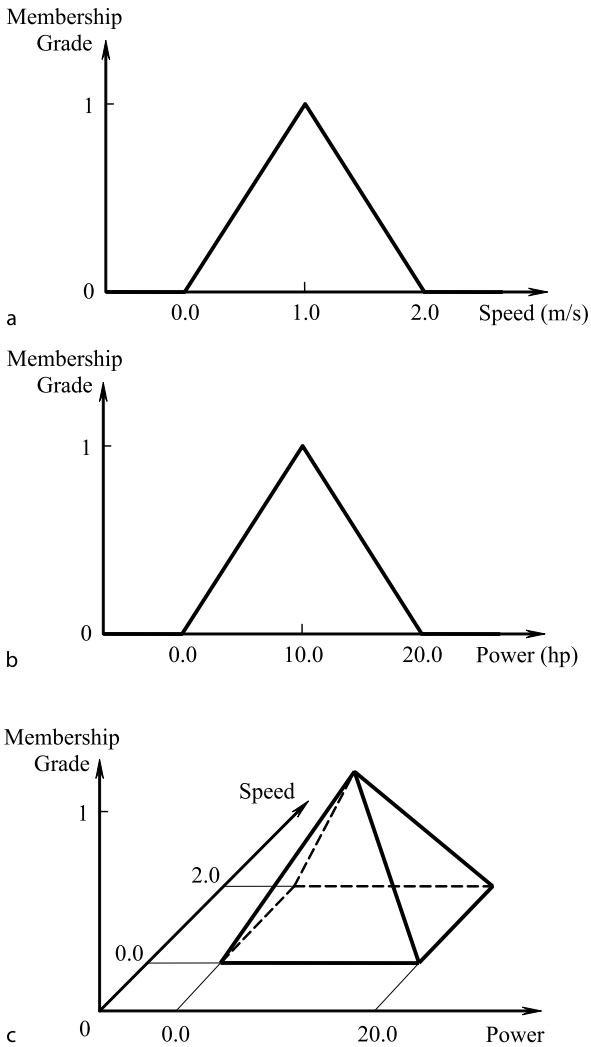

Intelligent Control, Figure 10
a Fuzzy set *A*; b Fuzzy set *B*; c Fuzzy implication *A* → *B*

Decision making using fuzzy logic is known as fuzzy inference or *approximate reasoning*. The *compositional rule of inference* is utilized for this purpose.

Composition

Consider a fuzzy relation (fuzzy set) R in the r-dimensional subspace $X_1 \times X_2 \times \ldots \times X_r$ and a second fuzzy relation (fuzzy set) S in the $(n-m+1)$-dimensional subspace $X_m \times X_{m+1} \times \ldots \times X_n$ such that $m < r+1$. Note that unlike the previous case of join, the two subspaces are never disjoint, and hence their intersection is never null (i. e., there is at least one common dimension in the subspaces R and S). The union of the two subspaces gives the overall n-dimensional space $X_1 \times X_2 \times \ldots \times X_n$. The *composition* of R and S is denoted by $R \circ S$ and is given by

$$R \circ S = \mathrm{Proj}[\mathrm{Join}(R, S); X_1, \ldots, X_{m-1}, X_{r+1}, \ldots, X_n] . \tag{12}$$

Specifically, we take the join (intersection) of the two sets R and S, and then project (max or supremum operation) the resulting fuzzy set (in the n-dimensional space) onto the subspace formed by the disjoint parts of the two subspaces in which the fuzzy sets R and S are defined. The membership function of the resulting fuzzy set is obtained from the membership functions of R and S, while noting that the operation *min* applies for *join* and the operation *supremum* applies for *projection*. Specifically,

$$\mu_{(R \circ S)} = \sup_{x_m, \ldots, x_r} [\min(\mu_R, \mu_S)] . \tag{13}$$

This is called the *sup-min composition*.

Composition can be interpreted as a matching of the common (known) parts of two fuzzy sets, and then making an inference about the disjoint (unknown) parts according to the result of the matching process. Specially, the two sets are combined and matched (through *join*) over their common subspace. The result is then projected (through *supremum*) over to the inference subspace (non-overlapping subspace), giving a fuzzy set defined over the disjoint portions of the two subspaces. This process of matching is quite analogous to matching of data with the condition part of a rule in the conventional rule-based decision making. In this regard composition plays a crucial role in fuzzy inference (fuzzy decision making) and fuzzy knowledge based systems.

As a simple example, consider a fuzzy relation R defined in the $X \times Y$ space and another fuzzy relation S defined in the $Y \times Z$ space. The composition of these two fuzzy relations is given by

$$\mu_{R \circ S}(x, z) = \sup_{y \in Y} \{\min(\mu_R(x, y), \mu_S(y, z))\} , \tag{14}$$

which is defined in the $X \times Z$ space.

Sup-Product Composition In Eq. (13) if we use the "product" operation in place of "min" we get the *sup-product composition* given by

$$\mu_{(R \bullet S)} = \sup_{X_m, \ldots, X_r} [\mu_R \times \mu_S] . \tag{15}$$

This composition is denoted by "•" and is also known as the *sup-dot composition*.

In summary, the composition operation involves:

1. Matching of data and a knowledge base, using the join (or, intersection) operation.

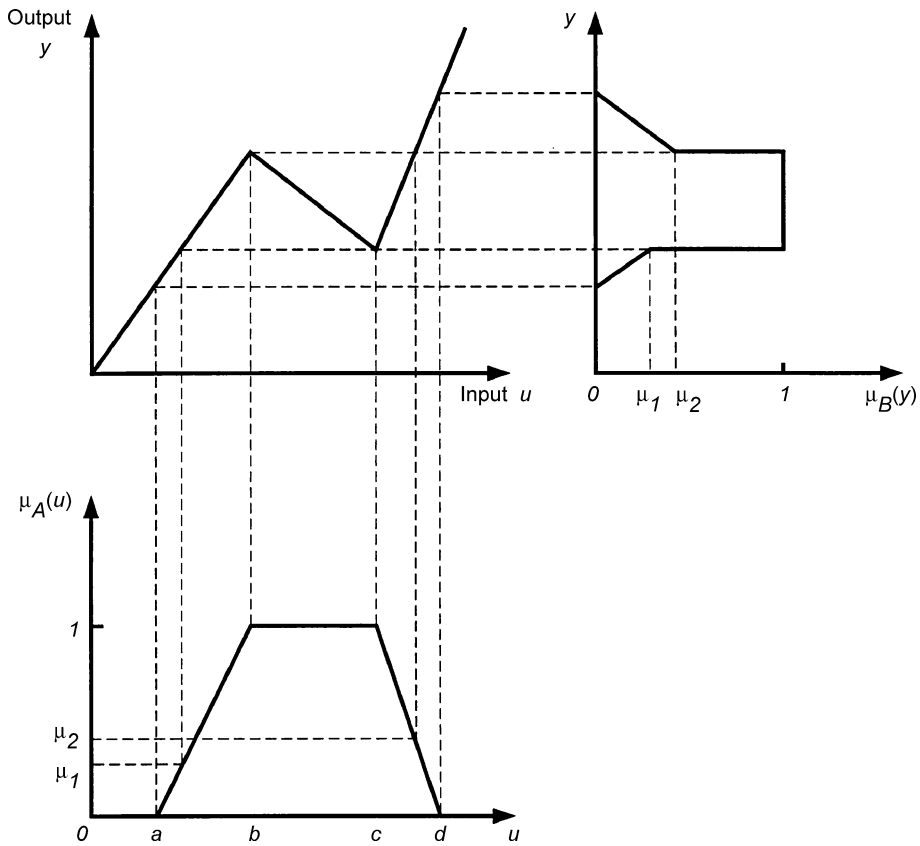

Intelligent Control, Figure 11
Fuzzy decision making using a crisp relation (extension principle)

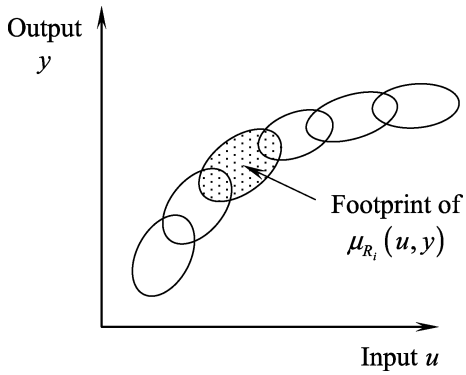

Intelligent Control, Figure 12
Fuzzy decision making using a fuzzy relation (Composition)

2. Making an inference (decision) on the variables in the subspace of the knowledge base that is outside the subspace of data (i. e., nonoverlapping subspace), using the projection (sup) operation.

Compositional Rule of Inference

In knowledge-based systems, knowledge is often expressed as rules of the form:

"**IF** condition Y_1 is y_1 **AND IF** condition Y_2 is y_2 **THEN** action C is c."

In fuzzy knowledge-based systems (e. g., fuzzy control systems), rules of this type are linguistic statements of expert knowledge in which y_1, y_2, and c are fuzzy quantities (e. g., small negative, fast, large positive). These rules are fuzzy relations that employ the fuzzy implication (IF-THEN). The collective set of fuzzy relations forms the knowledge base of the fuzzy system. Let us denote the fuzzy relation formed by this collection of rules as the fuzzy set K. This relation is an aggregation of the individual rules, and may be represented by a multivariable membership function. In a fuzzy decision making process (e. g., in fuzzy logic control), the rulebase (knowledge base) K is first collectively matched with the available data (context). Next, an inference is made on another fuzzy variable that

is represented in the knowledge base, on this basis. The matching and inference making are done using the composition operation, as discussed previously. The application of composition to make inferences in this manner is known as the *compositional rule of inference* (CRI).

For example, consider a control system. Usually the context would be the measured outputs Y of the process. The control action that drives the process is C. Typically, both these variables are crisp, but let us ignore this fact for the time being and assume them to be fuzzy, for general consideration. Suppose that the control knowledge base is denoted by R, a fuzzy relation. The method of obtaining the rule base R is analogous to model identification in conventional crisp control. Then, by applying the compositional rule of inference we get the fuzzy control action as:

$$\mu_C = \max_Y \min(\mu_Y, \mu_R) \,. \tag{16}$$

Extensions to Fuzzy Decision Making

Thus far we have considered fuzzy rules of the form:

$$\text{IF } x \text{ is } A_i \text{ AND IF } y \text{ is } B_i \text{ THEN } z \text{ is } C_i \tag{17}$$

where, A_i, B_i, and C_i are fuzzy states governing the ith rule of the rulebase. In fact this is the the Mamdani approach (Mamdani system or Mamdani model) named after the person who pioneered the application of this approach. Here, the knowledge base is represented as fuzzy protocols and represented by membership functions for A_i, B_i, and C_i, and the inference is obtained by applying the compositional rule of inference. The result is a fuzzy membership function, which typically has to be *defuzzified* for use in practical tasks.

Several variations to this conventional method are available. One such version is the *Sugeno model* (or, *Takagi–Sugeno–Kang model* or *TSK model*). Here, the knowledge base has fuzzy rules with crisp functions as the consequent, of the form

$$\text{IF } x \text{ is } A_i \text{ AND IF } y \text{ is } B_i \text{ THEN } c_i = f_i(x, y) \tag{18}$$

for Rule i, where, f_i is a crisp function of the condition variables (antecedent) x and y. Note that the condition part of this rule is the same as for the Mamdani model (17), where A_i and B_i are fuzzy sets whose membership functions are functions of x and y, respectively. The action part is a crisp function of the condition variables, however. The inference $\hat{c}(x, y)$ of the fuzzy knowledge-based system is obtained directly as a crisp function of the condition variables x and y, as follows:

For Rule i, a weighting parameter $w_i(x, y)$ is obtained corresponding to the condition membership functions, as for the Mamdani approach, by using either the "min" operation or the "product" operation. For example, using the "min" operation we form

$$w_i(x, y) = \min[\mu_{A_i}(x), \mu_{B_i}(y)] \,. \tag{19}$$

The crisp inference $\hat{c}(x, y)$is determined as a weighted average of the individual rule inferences (crisp) $c_i = f_i(x, y)$ according to

$$\hat{c}(x, y) = \frac{\sum_{i=1}^{r} w_i c_i}{\sum_{i=1}^{r} w_i} = \frac{\sum_{i=1}^{r} w_i(x, y) f_i(x, y)}{\sum_{i=1}^{r} w_i(x, y)} \tag{20}$$

where, r is the total number of rules. For any data x and y, the knowledge-based action $\hat{c}(x, y)$ can be computed from (20), without requiring any defuzzification. The Sugeno model is particularly useful when the actions are described analytically through crisp functions, as in conventional crisp control, rather than linguistically. The TSK approach is commonly used in the applications of direct control and in simplified fuzzy models. The Mamdani approach, even though popular in low-level direct control, is particularly appropriate for knowledge representation and processing in expert systems and in high-level (hierarchical) control systems.

Fuzzy Control

Fuzzy Logic Control or simply "Fuzzy Control" belongs to the class of "Intelligent Control", "Knowledge-Based Control", or "Expert Control". Fuzzy control uses knowledge-based decision making employing techniques of fuzzy logic, in determining the control actions. In this section, the basics of fuzzy logic control are presented.

There are many reasons for the practical deficiencies of conventional, crisp algorithmic control. Conventional control is typically based on "modeling" the plant that is to be controlled. If the model is not accurate the performance of the controller may not be acceptable. Furthermore, these conventional controllers rely on a complete set of data, including sensory information and parameter values, to produce control actions, and indeed, program instructions and data are conventionally combined into the same memory of the controller. If the model parameters and other necessary data are not completely known, say unexpectedly as a result of sensing problems, then appropriate estimates have to be made. Furthermore, if the available information is fuzzy, qualitative, or vague,

these "crisp" controllers that are based on such *incomplete information* will not usually provide satisfactory results. The environment with which the plant interacts may not be completely predictable either, and it is normally difficult for a crisp control algorithm to accurately respond to a condition that it did not anticipate and that it could not "understand". Also, some conventional techniques of control assume that the plant is linear and time-invariant, an assumption that hardly holds in a large majority of practical problems. Conventional control techniques that are based on "crisp" algorithms or mathematical formulas can fail when the plant to be controlled is very complex, highly nonlinear, incompletely known, or contains large process delays. Often, a control strategy that is based on "soft knowledge" such as human "experience" in operating the plant, heuristics, commonsense, or expert opinion can provide the remedy to these problems.

Basics of Fuzzy Control

Fuzzy control uses the principles of fuzzy logic-based decision making to arrive at the control actions. The decision making approach is typically based on the *compositional rule of inference* (CRI), as presented before. In essence, some information (e. g., output measurements) from the system to be controlled is matched with a knowledge base of control for the particular system, using CRI. A fuzzy rule in the knowledge base of control is generally a "linguistic relation" of the form

$$\text{IF } A_i \text{ THEN IF } B_i \text{ THEN } C_i \tag{21}$$

where, A_i and B_i are fuzzy quantities representing process measurements (e. g., process error and change in error) and C_i is a fuzzy quantity representing a control signal (e. g., change in process input). What we have is a rule-base with a set of (n) rules:

$$\begin{aligned}
&\text{Rule 1: } A_1 \text{ and } B_1 \Rightarrow C_1 \\
&\text{Rule 2: } A_2 \text{ and } B_2 \Rightarrow C_2 \\
&\qquad \dots \quad \dots \quad \dots \quad \dots \\
&\text{Rule } n\text{: } A_n \text{ and } B_n \Rightarrow C_n\,.
\end{aligned}$$

Because these fuzzy sets are related through IF-THEN implications and because an implication operation for two fuzzy sets can be interpreted as a "minimum operation" on the corresponding membership functions, the membership function of this fuzzy relation may be expressed as

$$\mu_{Ri}(a,b,c) = \min[\mu_{Ai}(a), \mu_{Bi}(b), \mu_{Ci}(c)]\,. \tag{22}$$

The individual rules in the rule-base are joined through ELSE connectives, which are OR connectives ("unions" of membership functions). Hence, the overall membership function for the complete rule-base (relation R) is obtained using the "maximum" operation on the membership functions of the individual rules; thus

$$\begin{aligned}
\mu_R(a,b,c) &= \max_i \mu_{Ri}(a,b,c) \\
&= \max_i \min[\mu_{Ai}(a), \mu_{Bi}(b), \mu_{Ci}(c)]\,. \qquad (23)
\end{aligned}$$

In this manner the membership function of the entire rule-base can be determined (or, "identified" in the terminology of conventional control) using the membership functions of the response variables and control inputs. Note that a fuzzy knowledge base is a multivariable function – a multidimensional array (a three-variable function or a dimensional array in the case of Eq. (23)) of membership function values. This array corresponds to a fuzzy control algorithm in the sense of conventional control. The control rule-base may represent linguistic expressions of experience, expertise, or knowledge of the domain experts (control engineers, skilled operators, etc.). Alternatively, a control engineer may instruct an operator (or a control system) to carry out various process tasks in the usual manner; monitor and analyze the resulting data; and learn appropriate rules of control, say by using neural networks.

Once a fuzzy control knowledge base of the form given by Eq. (23) is obtained, we need a procedure to infer control actions using process measurements, during control. Specifically, suppose that fuzzy process measurements A' and B' are available. The corresponding control inference C' is obtained using the compositional rule of inference (i. e., inference using the composition relation). The applicable relation is

$$\mu_{C'}(c) = \sup_{a,b} \min[\mu_{A'}(a), \mu_{B'}(b), \mu_R(a,b,c)]\,. \tag{24}$$

Note that in fuzzy inference, the data fuzzy sets A' and B' are jointly matched with the knowledge-base fuzzy relation R. This is a "join" operation, which corresponds to an AND operation (an "intersection" of fuzzy sets), and hence the *min* operation applies for the membership functions. For a given value of control action c, the resulting fuzzy sets are then mapped (projected) from a three-dimensional space $X \times Y \times Z$ of knowledge onto a one-dimensional space Z of control actions. This mapping corresponds to a set of OR connectives, and hence the *sup* operation applies to the membership function values, as expressed in Eq. (24).

Actual process measurements are crisp. Hence, they have to be *fuzzified* in order to apply the compositional

rule of inference. This is conveniently done by reading the grade values of the membership functions of the measurement at the specific measurement values. Typically, the control action must be a crisp value as well. Hence, each control inference C' must be *defuzzified* so that it can be used to control the process. Several methods are available to accomplish defuzzification. In the *mean of maxima* method the control element corresponding to the maximum grade of membership is used as the control action. If there is more than one element with a maximum (peak) membership value, the mean of these values is used. In the *center of gravity* (or *centroid*) *method* the centroid of the membership function of control decision is used as the value of crisp control action. This weighted control action is known to provide a somewhat sluggish, yet more robust control.

There are several practical considerations of fuzzy control that were not addressed in the above discussion. Because representation of the rule-base R by an analytical function can be somewhat unrealistic and infeasible in general, it is customary to assume that the fuzzy sets involved in R have discrete and finite universes (or at least discrete and finite support sets). As a result, process response measurements must be quantized. Hence, at the outset, a decision must be made as to the element resolution (quantization error) of each universe. This resolution governs the cardinality of the sets and in turn the size of the multidimensional membership function array of a fuzzy rule-base. It follows that the required computational effort, memory and storage requirements, and accuracy are directly affected by the quantization resolution.

Because process measurements are crisp, one method of reducing the real-time computational overhead is to precompute a decision table relating quantized measurements to crisp control actions. The main disadvantage of this approach is that it does not allow for convenient modifications (e. g., rule changes and quantization resolution adjustments) during operation. Another practical consideration is the selection of a proper sampling period in view of the fact that process responses are generally analog signals. Factors such as process characteristics, required control bandwidth, and the processing time needed for one control cycle, must be taken into account in choosing a sampling period. Scaling or gain selection for various signals in a fuzzy logic control system is another important consideration. For reasons of processing efficiency, it is customary to scale the process variables and control signals in a fuzzy control algorithm. Furthermore, adjustable gains can be cascaded with these system variables so that they may serve as tuning parameters for the controller. A proper tuning algorithm would be needed, however. A related consideration is real-time or on-line modification of a fuzzy rule-base. Specifically, rules may be added, deleted, or modified on the basis of some scheme of *learning and self-organization*. For example, using a model for the process and making assumptions such as input-output monotonicity, it is possible during control to trace and tag the rules in the rule-base that need attention. The control-decision table can be modified accordingly.

Steps of Fuzzy Logic Control

The main steps of fuzzy logic control, as described above, can be summarized in the following manner. A fuzzy knowledge base must be developed first (off-line), according to the following four steps:

1. Develop a set of linguistic control rules (protocols) that contain fuzzy variables as conditions (process outputs) and actions (control inputs to the process).
2. Obtain a set of membership functions for process output variables and control input variables.
3. Using the "fuzzy AND" operation (typically, *min*) and the "fuzzy implication" operation (typically, *min*) on each rule in Step 1, and using Step 2, obtain the multivariable rule-base function R_i (a multi-dimensional array of membership values in the discrete case) for that rule.
4. Combine (aggregate) the relations R_i using the fuzzy connectives ELSE (fuzzy OR; typically, *max*) to obtain the overall fuzzy rule-base (relation) R (a multi-dimensional array in the discrete case).

Then, control actions may be determined in real time as follows:

1. Fuzzify the measured (crisp) process variables (for example, a fuzzy singleton may be obtained by reading the grade value of the corresponding membership function at the measured value of the variable).
2. Match the fuzzy measurements obtained in Step 1 with the membership function (array in the discrete case) of the fuzzy rule-base (obtained in previous Step 4), using the compositional rule of inference.
3. Defuzzify the control inference obtained in Step 2 (for example, the mean of maxima method or the centroid method may be used here).

These steps reflect the formal procedure in FLC. There are several variations. For example, a much faster approach would be to develop a crisp decision table by combining the four steps of fuzzy knowledge base development and the first two steps of control, and using this table in

a table look-up mode to determine a crisp control action during operation. Also, hardware fuzzy processors (*fuzzy chips*) may be used to carry out the fuzzy inference at high speed. The rules, membership functions, and measured context data are generated as usual, through the use of a control "host" computer. The fuzzy processor is located in the same computer, which has appropriate interface (input/output) hardware and driver software. Regardless of all these, it is more convenient to apply the inference mechanism separately to each rule and then combine the result instead of applying it to the entire rule-base using the compositional rule of inference. This aspect is discussed next.

Fuzzy logic is commonly used in direct control of processes and machinery. In this case the inferences of a fuzzy decision making system form the control inputs to the process. These inferences are arrived at by using the process responses as the inputs (context data) to the fuzzy system. The structure of a direct fuzzy controller is shown in Fig. 13. Here, y represents the process output, u represents the control input, and R is the relation, which represents the fuzzy control knowledge base.

Example Consider the room comfort control system schematically shown in Fig. 14. The temperature (T) and humidity (H) are the process variables that are measured. These sensor signals are provided to the fuzzy logic controller, which determines the cooling rate (C) that should be generated by the air conditioning unit. The objective is to maintain a particular comfort level inside the room.

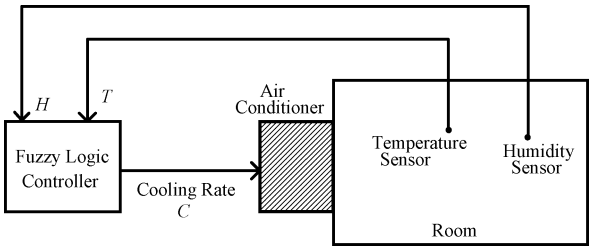

Intelligent Control, Figure 14
Comfort control system of a room

A simplified fuzzy rule-base of the comfort controller is graphically presented in Fig. 15. The temperature level can assume one of two fuzzy states *(HG, LW)*, which denote high and low, respectively, with the corresponding membership functions. Similarly, the humidity level can assume two other fuzzy states *(HG, LW)* with associated membership functions. Note that the membership functions of T are quite different from those of H, even though

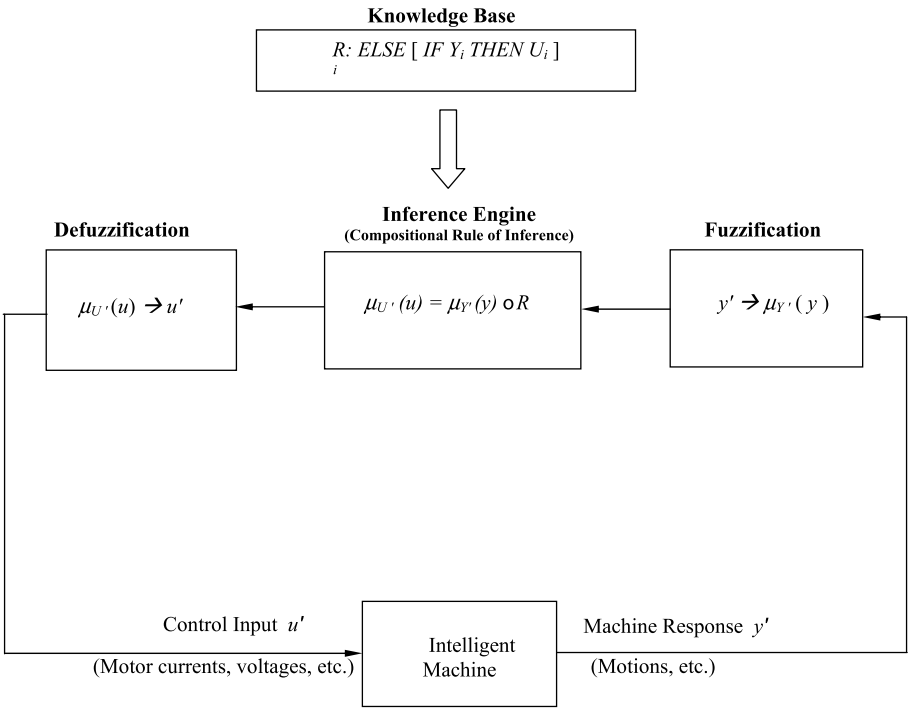

Intelligent Control, Figure 13
Structure of a direct fuzzy controller

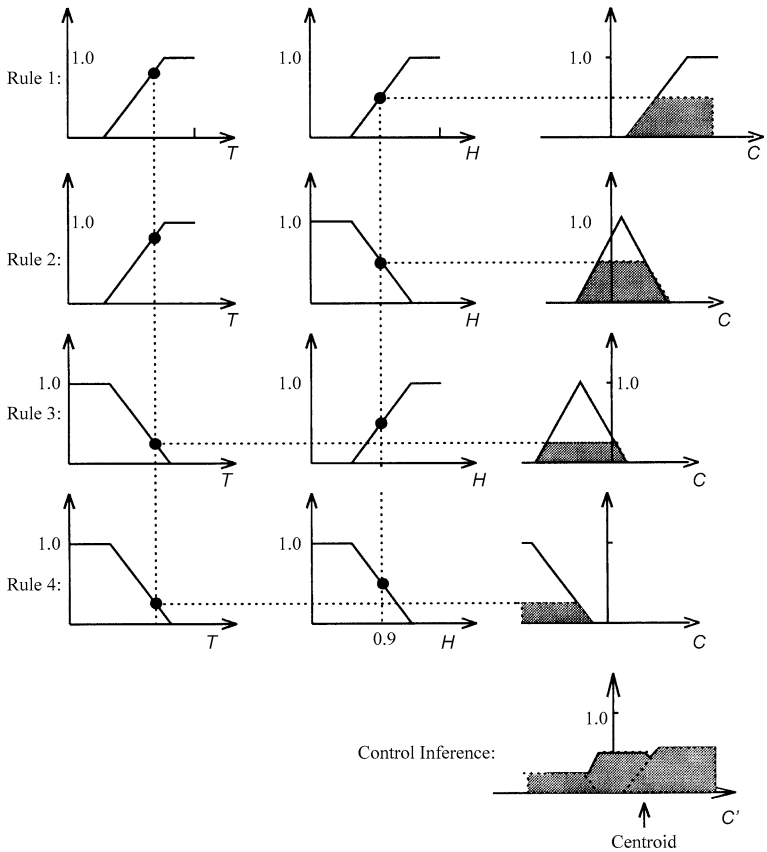

Intelligent Control, Figure 15
The fuzzy knowledge base of the comfort controller

the same nomenclature is used. There are four rules, as given in Fig. 15. The rule-base is:

Rule 1: If T is HG and H is HG then C is PH
Rule 2: else if T is HG and H is LW then C is PL
Rule 3: else if T is LW and H is HG then C is NL
Rule 4: else if T is LW and H is LW then C is NH and if

The nomenclature used for the fuzzy states is as follows:

Temperature (T)	Humidity (H)	Change in Cooling Rate (C)
HG = High	HG = High	PH = Positive High
LW = Low	LW = Low	PL = Positive Low
		NH = Negative High
		NL = Negative Low

Application of the compositional rule of inference is done here by using individual rule-based composition. Specifically, the measured information is composed with individual rules in the knowledge base and the results are aggregated to give the overall decision. For example, suppose that the room temperature is 30°C and the relative humidity is 0.9. Lines are drawn at these points, as shown in Fig. 15, to determine the corresponding membership grades for the fuzzy states in the four rules. In each rule the lower value of the two grades of process response variables is then used to clip (or modulate) the corresponding membership function of C (a *min* operation). The resulting "clipped" membership functions of C for all four rules are superimposed (a *max* operation) to obtain the control inference C' as shown. This result is a fuzzy set, and it must be defuzzified to obtain a crisp control action $\hat{c}$ for changing the cooling rate. The centroid method may be used for defuzzification.

Fuzzy Control Surface

A fuzzy controller is a nonlinear controller. A well-defined problem of fuzzy control, with analytical membership functions and fuzzification and defuzzifica-

tion methods, and well-defined fuzzy logic operators, may be expressed as a nonlinear control surface through the application of the compositional rule of inference. The advantage then is that the generation of the control action becomes a simple and very fast step of reading the surface value (control action) for given values of crisp measurement (process variables). The main disadvantage is, the controller is fixed and cannot accommodate possible improvements to control rules and membership functions through successive learning and experience. Nevertheless, this approach to fuzzy control is quite popular. A useful software tool for developing fuzzy controllers is the MATLAB® Fuzzy Logic Toolbox.

Example A schematic diagram of a simplified system for controlling the liquid level in a tank is shown in Fig. 16a. In the control system, the error (actually, correction) is given by

$$e = \text{Desired level} - \text{Actual level}\,.$$

The change in error is denoted by Δe. The control action is denoted by u, where $u > 0$ corresponds to opening the inflow valve and $u < 0$ corresponds to opening the outflow valve. A low-level direct fuzzy controller is used in this control system, with the control rule-base as given in Fig. 16b.

The membership functions for E, ΔE, and U are given in Fig. 16c. Note that the error measurements are limited to the interval $[-3a, 3a]$ and the Δ error measurements to $[-3b, 3b]$. The control actions are in the range $[-4c, 4c]$.

Following the usual steps of applying the compositional rule of inference for this fuzzy logic controller, we can develop a crisp *control surface* $u(e, \Delta e)$ for the system, expressed in the three-dimensional coordinate system $(e, \Delta e, u)$, which then can be used as a simple and fast controller. This method is described next.

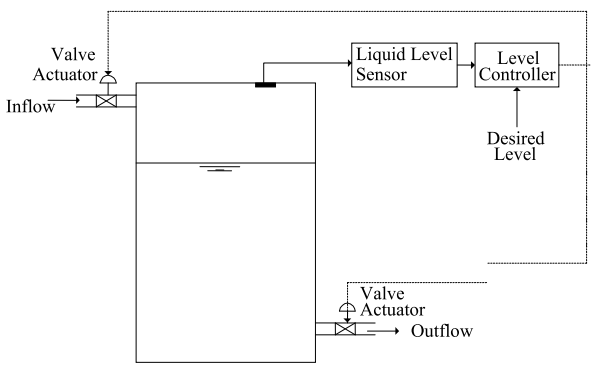

Intelligent Control, Figure 16a
Liquid level control system

E \ ΔE	NL	NS	ZO	PS	PL
NL	NL	NL	NM	NS	ZO
NS	NL	NM	NS	ZO	PS
ZO	NM	NS	ZO	PS	PM
PS	NS	ZO	PS	PM	PL
PL	ZO	PS	PM	PL	PL

Intelligent Control, Figure 16b
The control rule-base

The crisp control surface is developed by carrying out the rule-based inference for each point: $(e, \Delta e)$ in the measurement space $E \times \Delta E$, using individual rule-based inference. Specifically,

$$\mu_{U'}(u) = \max_{i,j} \min[\mu_{Ei}(e_o), \mu_{\Delta Ej}(\Delta e_o), \mu_{Uk}(u)]\,, \quad (25)$$

where

$\boldsymbol{\mu_{U'}}$	control inference membership function
$\boldsymbol{e_o}$	crisp context variable "error" defined in $[-3a, 3a]$
$\boldsymbol{\Delta e_o}$	crisp context variable "change in error" defined in $[-3b, -3b]$
$\boldsymbol{E_i}$	fuzzy states of "error"
$\boldsymbol{\Delta E_j}$	fuzzy states of "change in error"
$\boldsymbol{U_k}$	fuzzy states of "control action"
$\boldsymbol{(i, j, k)}$	possible combinations of fuzzy states of error, change in error, and control action, within the rule-base.

To find the crisp control inference (u') for a set of crisp context data $(e, \Delta e)$, the fuzzy inference $\mu_{U'}(u)$ is defuzzified using the center of gravity (centroid) method, which for the continuous case, is:

$$u' = \frac{\int\limits_{u \in U} u\mu_{U'}(u)\mathrm{d}u}{\int\limits_{u \in U} \mu_{U'}(u)\mathrm{d}u} \quad (26a)$$

or, for the discrete case, it is:

$$u' = \frac{\sum\limits_{u_i \in U} u_i \mu_{U'}(u_i)\mathrm{d}u}{\sum\limits_{u_i \in U} \mu_{U'}(u_i)\mathrm{d}u}\,, \quad (26b)$$

where $U = [-4c, 4c]$. Also, if the geometric shape of the inference is simple (e. g., piecewise linear), the centroid

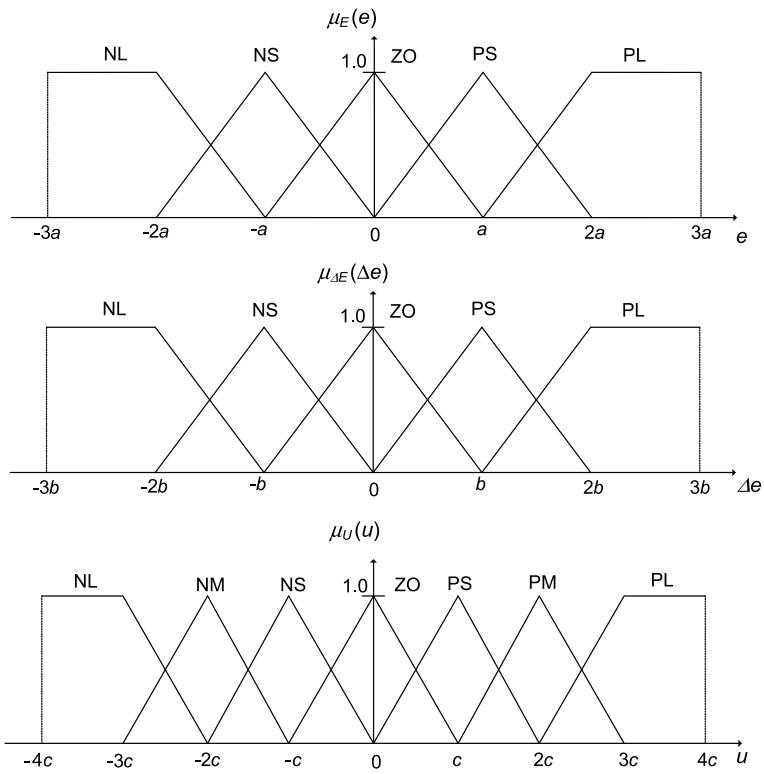

Intelligent Control, Figure 16c
The membership functions of Error, Change in Error, and Control Action

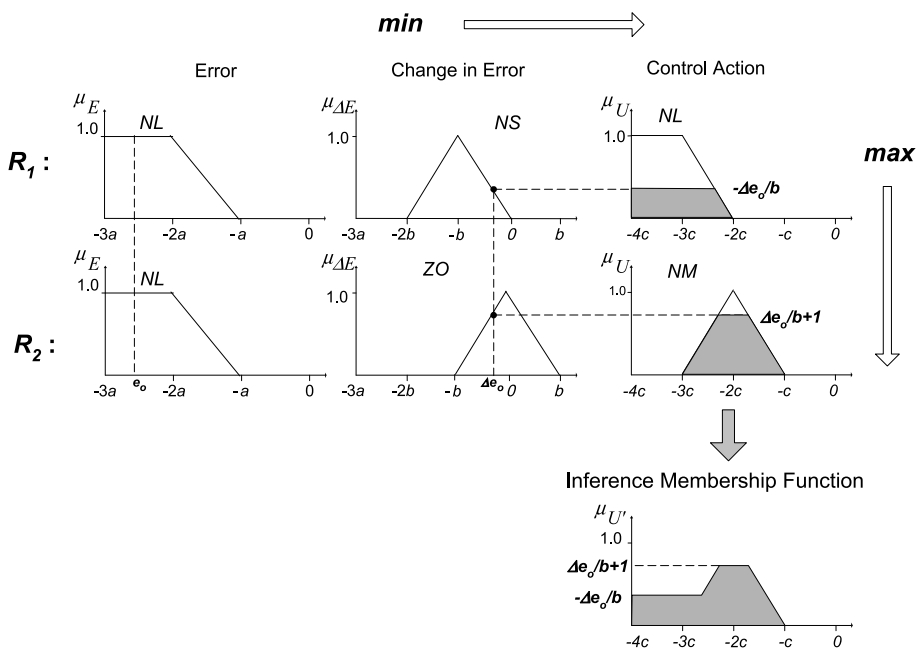

Intelligent Control, Figure 17
Individual rule-based inference for $e_o[-3a, -2a]$ and $\Delta e_o[-b/2, 0]$

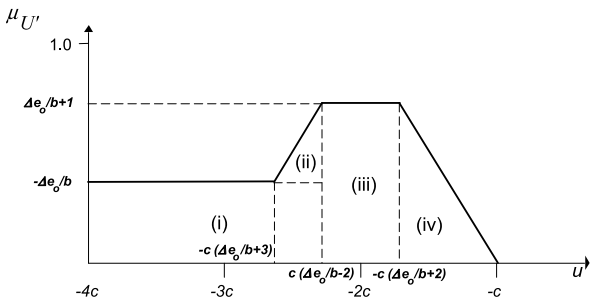

Intelligent Control, Figure 18
Sub-regions and critical points for calculation of the centroid

can be computed by the moment method; thus,

$$u' = \frac{\sum_{i=1}^{n} \text{area}_i\, m_i}{\sum_{i=1}^{n} \text{area}_i}\,, \tag{26c}$$

where

$$\begin{aligned} \text{area}_i &= \text{area of the } i\text{th sub-region} \\ m_i &= \text{distance of the centroid of the } i\text{th sub-region,} \\ &\quad\ \text{on the control axis .} \end{aligned}$$

To demonstrate this procedure, consider a set of context data $(e_o, \Delta e_o)$, where e_o is in $[-3a, -2a]$ and Δe_o is in $[-b/2, 0]$. Then, from the membership functions and the rule-base, it should be clear that only two rules are valid in this region, as given below:

$$R_1\text{: if } e \text{ is } NL \text{ and } \Delta e \text{ is } NS \text{ then } u \text{ is } NL\,,$$
$$R_2\text{: if } e \text{ is } NL \text{ and } \Delta e \text{ is } ZO \text{ then } u \text{ is } NM\,.$$

Since, in the range $[-3a, -2a]$, the membership grade of singleton fuzzification of e_o is always 1, the lower grade of the two context values is the one corresponding to the singleton fuzzification of Δe_o for both rules. Then, in applying the individual rule-based inference, the lower grade value of the two context variables is used to clip off the corresponding membership function of the control action variable U in each rule (this is a *min* operation). The resulting membership functions of U for the two applicable rules are superimposed (this is a *max* operation) to obtain the control inference U', as shown in Fig. 17.

For defuzzification, we apply the moment method to find the centroid of the resulting membership function of control inference, as shown in Fig. 18. Note that the critical points in Fig. 17 and Fig. 18 (e. g., $-\Delta e_o/b$, $-c(\Delta e_o/b+3)$, etc.) are found from the corresponding membership functions.

From the moment method, we obtain the crisp control action as a function of e and Δe. The above procedure is repeatedly applied to all possible ranges of e $[-3a, 3a]$ and Δe $[-3b, 3b]$, to obtain the complete control surface.

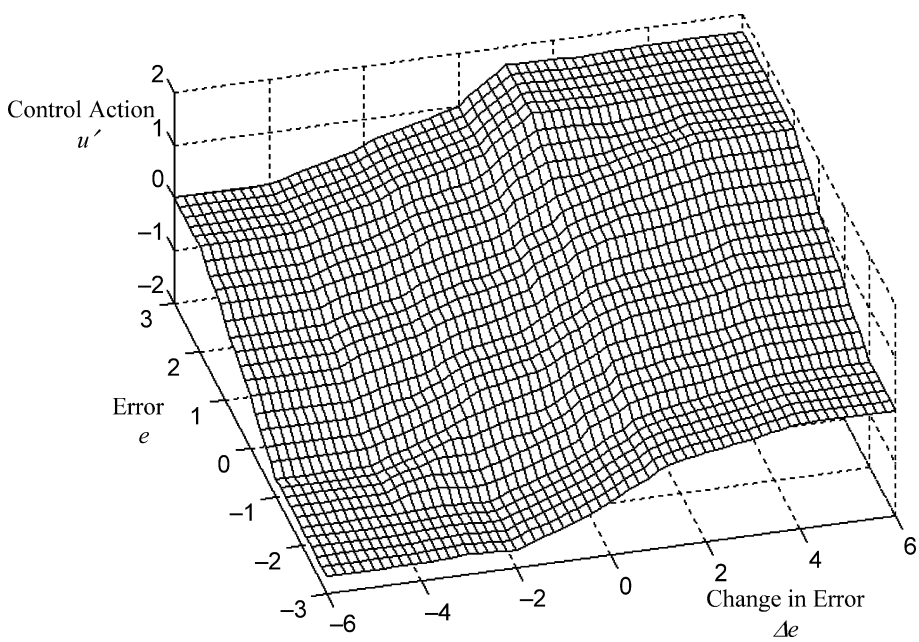

Intelligent Control, Figure 19
Control surface with $a = 1$, $b = 2$, and $c = 0.5$

Also, the procedure can be implemented in a computer program to generate a control surface. A control surface with $a = 1$, $b = 2$, and $c = 0.5$ is shown in Fig. 19.

In the present example what we have applied is in fact the Mamdani approach (Mamdani system or Mamdani model). Sugeno model (or, Takagi–Sugeno–Kang model or TSK model) could have been used as well, thereby avoiding the defuzzification step.

Future Directions

Even though a considerable effort has gone into the development of machines that somewhat mimic humans in their actions, the present generation of intelligent machines do not claim to possess all the capabilities of human intelligence; for example, common sense, display of emotions, and inventiveness. Significant advances have been made, however, in machine implementation of characteristics of intelligence such as sensory perception, pattern recognition, knowledge acquisition and learning, inference from incomplete information, inference from qualitative or approximate information, ability to handle unfamiliar situations, adaptability to deal with new yet related situations, and inductive reasoning. Much research and development would be needed in these areas, pertaining to techniques, hardware, and software before a machine could reliably and consistently possess the level of intelligence of say, a dog.

For instance, consider a handwriting recognition system, which is a practical example of an intelligent system. The underlying problem cannot be solved through simple template matching, which does not require intelligence. Handwriting of the same person can vary temporally, due to various practical shortcomings such as missing characters, errors, nonrepeatability, physiological variations, sensory limitations, and noise. It should clear from this observation that a handwriting recognition system has to deal with incomplete information and unfamiliar objects (characters), and should possess capabilities of learning, pattern recognition, and approximate reasoning, which will assist in carrying out intelligent functions of the system. Techniques of soft computing are able to challenge such needs of intelligent machines and intelligent control.

Increased attention will be paid to developments of intelligent controllers with inspiration from biological systems. In particular, hybrid systems that employ two or more soft computing approaches will become common. Various techniques of self-learning and distributed intelligent control will find an increased presence [12]. Embedded systems and integrated miniature systems with intelligent sensing, actuation, and control integrated and distributed within the same device will gain more prominence.

Bibliography

Primary Literature

1. Cao Y, de Silva CW (2006) Dynamic Modeling and Neural-Network Adaptive Control of a Deployable Manipulator System. J Guid Control Dyn 29(1):192–195
2. Cao Y, de Silva CW (2006) Supervised Switching Control of a Deployable Manipulator System. Int J Control Intell Syst 34(2):153–165
3. de Silva CW (1992) Research Laboratory for Fish Processing Automation. Int J Robotics Comput-Integr Manuf 9(1):49–60
4. de Silva CW (1997) Intelligent Control of Robotic Systems with Application in Industrial Processes. Robotics Auton Syst 21:221–237
5. de Silva CW (2003) The Role of Soft Computing in Intelligent Machines. Philos Trans Royal Soc Ser A UK 361(1809):1749–1780
6. de Silva CW, Wickramarachchi N (1997) An Innovative Machine for Automated Cutting of Fish. IEEE/ASME Trans Mechatron 2(2):86–98
7. Filippidis A, Jain LC, de Silva CW (1999) Intelligent Control Techniques. In: Jain LC, de Silva CW (eds) Intelligent Adaptive Control. CRC Press, Boca Raton
8. Goulet JF, de Silva CW, Modi VJ (2001) Hierarchical Control of a Space-Based Deployable Manipulator Using Fuzzy Logic. AIAA J Guid Control Dyn 24(2):395–405
9. Rahbari R, de Silva CW (2001) Comparison of Two Inference Methods for P-type Fuzzy Logic Control through Experimental Investigation Using a Hydraulic Manipulator. Eng Appl Artif Intell 14(6):763–784
10. Rahbari R, Leach BW, Dillon J, de Silva CW (2005) Expert System for an INS/DGPS Integrated Navigator Installed in a Bell 206 Helicopter. Eng Appl Artif Intell 18(3):353–361
11. Tang PL, Poo AN, de Silva CW (2001) Knowledge-Based Extension of Model-Referenced Adaptive Control with Application to an Industrial Process. J Intell Fuzzy Syst 10(3–4):159–183
12. Wang Y, de Silva CW (2008) A Machine Learning Approach to Multi-robot Coordination. Eng Appl Artif Intell 21(3):470–484
13. Yan GKC, de Silva CW, Wang GX (2001) Experimental Modeling and Intelligent Control of a Wood-Drying Kiln. Int J Adapt Control Signal Process 15:787–814
14. Zadeh LA (1984) Making Computers Think Like People. IEEE Spectrum 21(8):26–32

Books

Anderson TW (1984) An Introduction to Multivariate Statistical Analysis. Wiley, New York

Davis L (1991) Handbook on Genetic Algorithms. Van Nostrand–Rienhold, New York

de Silva CW (1995) Intelligent Control: Fuzzy Logic Applications. CRC Press, Boca Raton

de Silva CW (2000) Intelligent Machines: Myths and Realities. CRC Press, Boca Raton

de Silva CW (2005) Mechatronics: An Integrated Approach. Taylor & Francis/CRC Press, Boca Raton

Dubois D, Prade H (1980) Fuzzy Sets and Systems: Theory and Applications. Academic Press, Orlando
Farrell JA, Polycarpou MM (2006) Adaptive Approximation Based Control. Wiley-Interscience, New York
Forsyth R (1984) Expert Systems. Chapman & Hall, New York
Gupta MM, Rao H (1994) Neural Control Theory and Applications. IEEE Press, Piscataway
Hart A (1986) Knowledge Acquisition for Expert Systems. McGraw-Hill, New York
Jain LC, de Silva CW (1999) Intelligent Adaptive Control: Industrial Applications. CRC Press, Boca Raton
Jang JSR, Sun CTS, Mizutani E (1997) Neuro-Fuzzy and Soft Computing. Prentice Hall, Upper Saddle River
Karray FO, de Silva CW (2004) Soft Computing and Intelligent Systems Design – Theory, Tools, and Applications. Addison Wesley, New York
Klir GJ, Folger TA (1988) Fuzzy Sets, Uncertainty, and Information. Prentice Hall, Englewood Cliffs
Pao YH (1989) Adaptive Pattern Recognition and Neural Networks. Addison-Wesley, Reading
Passino KM (2005) Biomimicry for Optimization, Control, and Automation. Springer, London
Ruano AE (2005) Intelligent Control Systems Using Computational Intelligence Techniques. The Institute of Electrical Engineers, Herts
Vonk E, Jain LC, Johnson RP (1997) Automatic Generation of Neural Network Architecture Using Evolutionary Computing. World Scientific Publishing Co., Singapore

Intelligent Systems, Introduction to

James Hendler
Computer Science Department,
Rensselaer Polytechnic Institute, Troy, USA

The term "intelligent systems" has come to mean many different things in many different contexts and, like most things related to complex systems, it is hard to nail down a specific definition that is both rigorous enough to discriminate out those things which should not be included, but is loose enough to include those that are. As in defining terms like "agents" or "robots," one is able to find overly inclusive definitions, such as "an autonomously acting entity" where a thermostat in the latter case, or hard disk controller in the former, would meet the definition. On the other hand, tighten up the definition and telerobotics or Google's search bots no longer fit, despite being clearly related technologically. In the case of intelligent systems, too tight a definition of intelligence removes, say, the behaviors we see in a dog, which can seek out prey or be trained to beg for its dinner, but loosen the definition and we find ourselves talking about systems with the intelligence of a clam.

In this chapter, we are holding primarily to a tighter definition and starting to look at some of the kinds of behaviors that take us into areas traditionally associated with human intelligence. While several of the sections of this book deal with areas associated with, loosely defined, intelligent behaviors and others look specifically at aspects of Artificial Intelligence tied closely to data or a model, in this short section we pick up some of the missing pieces that help us complete the puzzle. The small number of papers in this section should not cause one to believe that there is little relation between intelligent systems and complexity, but rather that other sections of this encyclopedia necessarily included aspects of intelligence defined at some level – controlling against complexity demands it. The reader looking for other articles on Artificial Intelligence and complex systems will find them in many sections of this volume, particularly including those on:

- ► Agent Based Modeling and Simulation,
- ► Complexity and Non-linearity in Autonomous Robotics, Introduction to,
- ► Data-Mining and Knowledge Discovery, Introduction to
- ► Data-Mining and Knowledge Discovery, Neural Networks in
- ► Data-Mining and Knowledge Discovery: Case-Based Reasoning, Nearest Neighbor and Rough Sets
- ► Soft Computing, Introduction to, and
- ► Nonsmooth Analysis in Systems and Control Theory
- ► Chronological Calculus in Systems and Control Theory

However, despite the strong ties between these many subareas, there were some aspects of intelligent systems left out of these other topic areas, and this section is provided to cover these.

One such example is that of the use of mobile agents as the basis for intelligent cooperation among systems (see ► Mobile Agents). The primary need for mobility is in bandwidth-limited communications, and with the growth of modern computer networks, the area has gotten less attention of late. However, with the growing need for, and use of, sensor networks, wireless networks in noisy environments, deep-sea or space robotics, and other bandwidth-limited systems, agent-based modeling and simulation techniques can be used to model the networks, but not run on the networks themselves. Thus, an adaptive network that needs to do agent-to-agent communication for self-tuning will be seriously impacted in a bandwidth-

limited environment. Mobility can be used to help in such situations, moving small amounts of code to where the data it manipulates is stored, rather than moving the large amounts of data to where the computation could be run. Using such mobile agents bring both new capabilities, and new challenges, and as we try to increase the intelligence of systems operating in networks with disconnection, low bandwidth or high latency, such as many of the networks deployed in warfighting situations, agent mobility is regaining attention as an important area of research.

Another area that needs to be covered is that of the role of intelligent systems in the modeling and simulation area (see ► Artificial Intelligence in Modeling and Simulation). While agent-based modeling, as discussed elsewhere in this encyclopedia, is an important area, there are aspects of modeling it does not fully cover. One of these is the use of symbolic reasoning for use in validating models and simulation. Another is the actual modeling of the reasoning of other agents. For example, a baseball player recovering a ball that dropped in the outfield must be able to reason about what the base runners, who are trying to move up, and his team-mates, who are trying to hold them back, will most likely do. The player's skill in guessing the behaviors of these other agents, both competitive and cooperative, could be the difference between whether the game is won or lost.

As well as improving our capabilities in modeling and simulating systems, it is important to look at how, with the use of intelligent systems, we might better control the complex systems being modeled. Instead of trying to model a complex plant directly, in this section we look at work that takes a different approach. Instead we consider the knowledge-based control that results from observing, studying and understanding the behavior of a plant and/or the behavior of a human controlling it (see ► Intelligent Control). This area includes looking at soft-computing approaches to create an "approximate reasoning" solution that can be used for mimicking the control decisions that would be made by a human monitoring a plant, rather than for modeling the plant itself. Fuzzy logic, a particular branch of soft control has been successfully applied to the control of many complex systems, a number of which are described here. (The reader with a sense of humor may see a certain irony in many manufacturer's use of fuzzy logic to improve the picture on their camcorders and television sets, but let us leave that unexplored). Of particular import, this article outlines various ways in which the mathematical operations used in such control can be combined, allowing for the control of complex, non-linear systems that defy simpler control regimes.

Looking at very complex situations, a human operator, or at least a program simulating one, may want to look beyond soft computing and deal with the world at a higher level. One of the key abilities which separates the human from other primates and animals is our ability to learn over time to abstract away many details of the complex world in which we live, and to make plans for how to control it over time. Where planning itself can be a complex problem-solving task, learning how to abstract key aspects of situations and to apply plans to these is a critical need for dealing with complexity (see ► Learning and Planning (Intelligent Systems)). Exploring how we learn to plan is an area which has been gaining importance in the intelligent systems area as approaches which do not learn, but which apply brute force problem solving to larger and larger problems, are reaching the limits of their capabilities against the increasingly complex domains in which we wish to deploy our computational systems.

A recurring theme that arises in all of these attempts to provide intelligent behavior in evermore-complex systems environments is that of using a level of abstraction to reason not about some data or system itself, but about the meaning of the behaviors being observed. A critical aspect of performing such abstraction is the ability to represent a model of a system to the computer in a machine-readable way. The term "ontology" is used to describe this computer-based representation of the domain in which a system is trying to function.

Although ontologies have been around for a long time in AI, they have recently come back into their own in trying to help computer systems interact with one of the most complex human constructions in history – the millions of billions of dynamically changing bits of information that comprise the World Wide Web. While the Web has changed a great many aspects of our society, understanding its dynamics remains a major challenge (See Berners-Lee et al., Creating a Science of the World Wide Web, *Science*, 313(5788), August 2006, pp 769–771.) The use of ontologies, and other AI techniques, to help computers better process this massive information space is the raison d'etre of the "Semantic Web," an overview of which is also presented in this section (see ► Semantic Web).

In short, this section, despite its brevity, picks up a number of themes that arise throughout this encyclopedia. These five articles will help the reader understand how many of the themes above are connected together via the use of technologies developed in artificial intelligence labs, allowing the creation of intelligent systems that provide a key tool in our arsenal for dealing with the complexity of the natural world and/or the complex human society that has evolved to let us live in it.

Intentionality: A Naturalization Proposal on the Basis of Complex Dynamical Systems

Wolfgang Tschacher
University of Bern, Bern, Switzerland

Article Outline

Glossary
Definition of the Subject
Introduction
Intentionality and Representation
Synergetics
Discussion: Naturalization of Intentionality
Future Directions
Bibliography

Glossary

Cognition A general concept of psychology referring to all processes and structures of the mind. These comprise the processing of stimulus 'input' (i. e., perception) and the internal processing of represented information (e. g., memory functions, thinking, problem solving); the latter processes presuppose intentional features of the mind. Cognitive structures include knowledge, categories, memory, attitudes, and schemata, again intentional concepts.

Dynamical systems theory A system is any set of things (components, elements) that stand in relation to one another. If a rule or description exists that defines how the systems change over time (such as a differential equation or a mapping algorithm), the system is a dynamical system.

Intentionality A characterizing property of the mind. In contrast to physical systems, mental states have content, i. e. they 'are about something' in the sense that they contain a reference to an object, or the representation of an object. In addition to *aboutness*, intentionality demands a *functional* reference to the intentional object.

Mind–body problem The philosophical question as to if and how mind and brain/body interact. Analogously, the question if mental processes and physical processes are ontologically different.

Naturalization Explaining mental phenomena using concepts and models derived from the natural sciences. Naturalization efforts may be viewed as tools in order to develop viable simulation models of mental processes.

Definition of the Subject

In the philosophical and psychological tradition, intentionality is viewed as a characterizing property of mental (cognitive) acts. Mental acts have content, i. e. they 'are about something'. This something is called the intentional object. Intentionality may take the form of a desired state (as in, "I wish it were Friday.") or a goal (e. g., my plan for a weekend trip to the mountains). When viewing the constituents of the mind (the cognitive system) in this intentionalist manner, we stand in stark contrast to scientific descriptions of physical systems. These latter systems are material things, which are sufficiently described without reference to objects they would be about, or to states they might desire to realize. Therefore, are mental and physical systems qualitatively different with respect to intentionality? If yes, we are confronted with a dualist worldview entailing an aggravated mind–body problem. If no, a solution is demanded that can elucidate how mental phenomena may be explained avoiding intentional language or, conversely, how physical systems may show or mimic the features of intentionality. The project involved in the latter case may be named naturalization of intentionality.

Clarifying the problem of intentionality is important in several respects. First, psychology and other cognitive sciences are conceptually divided into two approaches, the phenomenological (first-person) approach and the behavioral and/or biological (third-person) approach. The conflation of first-person and third-person concepts may be viewed as a serious impediment to theorizing throughout psychology and cognitive neuroscience. Second, modern societies have a growing demand for machines and software that can function in 'intelligent' ways. Therefore, engineers of artificial knowledge-based systems need to know how intentionality can be implemented in physical information-processing machines. Third, the problem of intentionality is one of the foundational problems of the philosophy of mind and of consciousness research. Any, even if partial, solution to this problem that may be derived from a dynamical systems perspective is therefore welcome.

The phenomenology of intentional acts is well known; phenomenology yields the features of intentionality, which can indicate how closely a formal, mathematical, or physical model of intentionality approximates an understanding of the problem. These features are

- *Aboutness*, the intentional system's state must be about something in the system's environment
- *Functionality*, intentional states should be functional or instrumental with respect to what they are about

- *Mental-likeness*, in the sense that apart from being intentional, these model systems should have properties that resemble the properties of mental states. Especially, an interpreter (homunculus) who may account for missing links in the explanation of intentional mental states must not be allowed.

Introduction

If mind and matter are qualitatively different things, what is the nature of their difference? This is an elementary question of philosophy and scientific disciplines addressing the mind. This question and the potential answers to it, referred to as the mind–body problem, have a long history. What is the relationship between mind and body? If there is no essential difference between the two, or only a superficial difference, one may reach a monist understanding of the problem. The monists include idealists who posit that mind creates its world. A very distinct monist would be someone holding eliminative materialism, i. e. who would assume that the mind can be completely reduced to matter (e. g. [7]). If, on the other hand, a real difference between mind and body exists, we enter dualist notions (e. g. [41]). Dualism is confronted with intricate questions of how we may conceive of the interaction between mind and body. In the interest of the specific discussion of intentionality in this article, we will refrain from giving an account of philosophical theories of how mind and body may be associated in general.

Apart from presenting a core problem to the philosophy of mind, the mind–body problem is of interest to a wide range of contemporary scientific disciplines, especially psychology, neuroscience and computer science (artificial intelligence). In recent decades, a novel approach to cognitive science has appeared, combining dynamical systems theory with cognitive science. This dynamical approach to cognition addresses mind–body topics more or less explicitly. We will sketch this approach here because it provides the background for the ensuing treatment of intentionality.

The dynamical approach to cognition is founded on a number of studies and empirical paradigms. In various perceptual and behavioral tasks, researchers have observed a set of signatures of dynamical systems. These signatures are typically related to temporal patterns observed in the systems, especially asymptotic stability, and in many paradigmatic cases, multistability. Asymptotic stability means that a system's pattern of behavior is stable so that, in the face of an external disturbance, the system returns 'asymptotically' to this pattern. Multistability means that several such patterns may coexist in the behavioral space of a single system.

Movement coordination has been at the center of several applications. Haken, Kelso and Bunz (1985) [22] provided a model of the coordination of two limbs (e. g. the hands, or the forefingers of both hands, of a person) using equations from dynamical systems theory. The rhythmic movement of the limbs generally becomes synchronized after a short time; these synchronous movement patterns are stable with respect to external inputs, i. e. they usually return to synchrony after externally induced disruptions of movement. Furthermore, they have been shown to undergo phase transitions depending on the values of the control parameter (in the Haken–Kelso–Bunz [22] system, the velocity of movements prescribed by a metronome). Characteristic phenomena were observed in the context of phase transitions such as hysteresis and critical fluctuations. Analogous findings were reported in animal locomotion. Here, especially multistability has attracted the attention of researchers because in certain regions of control parameter space two qualitatively different limb coordination patterns frequently occur. For instance, a horse may either gallop or trot at a certain velocity. In the seemingly unrelated field of visual perception, very similar signatures of dynamical systems were found in ambiguous stimuli ([18,29]). This research on perceptual organization may be viewed as a continuation of the tradition of Gestalt psychology ([27,45]). For example, perception of apparent motion (i. e. perception of motion in the absence of real motion of the stimuli, Wertheimer [50]) can be induced presenting a stimulus, e. g. a black disk, alternately in different positions of the visual field. Certain spatial configurations of the disks allow perception of two or more qualitatively different kinds of apparent motion although identical stimuli are presented. Hence this and related paradigms create perceptual multistability. Again, the signatures of self-organizing dynamical systems (as described by synergetics) can be found in the phase transitions between the different apparent motion perceptions.

The dynamical hypothesis in cognitive science ([46,49]) proposes that cognitive agents may be modeled as dynamical systems (instead of as physical symbol systems, Newell [36]). A common denominator in these dynamical approaches is to start from elementary perception–action cycles, an idea that is also implied in the concept of embodied cognition. Clark [10] elaborated three bridging assumptions by which the numerous empirical findings can be integrated in the theory of dynamical systems. First, the assumption of continuity refers to cognition as continuous with its developmental foundations [44]. Second, 'off-line' reasoning and thinking is

viewed as continuous with on-line motor control strategies. Therefore abstract cognition may be decoupled from the actual environment but may still be working on the same dynamical principles; thinking is accordingly understood as emulated sensorimotor loops of perception–action cycles. Third, due to the dynamical hypothesis, pattern is provided not by programs but is 'soft-assembled' by a continuing process of self-organization. This latter assumption was initially formulated in the framework of theories on complex systems in the natural sciences (especially Haken's synergetics: Haken [17] and the theory of dissipative systems: [37]). The self-organization approach was successively introduced to cognitive science [21,26]. As already mentioned, prior to contemporary systems theory Gestalt psychology had developed a treatment of cognition and action which was very much akin to the dynamical systems approach, especially Köhler [27] and Lewin [30].

Cognitive science, especially its computational mainstream after behaviorist psychology turned cognitive in the 1960s, has had a tendency to start with 'higher' cognitive functions such as goals, beliefs and, in the context of goal-directed behavior, plans. This generated the symbol-grounding problem. The dynamical approach has avoided this problem arising in the computational framework [24] and has therefore proceeded in a bottom–up fashion instead. Higher cognition is assumed to emerge on the basis of elementary sensorimotor loops. Rather than focusing on symbol grounding, the dynamical view addresses symbol emergence that depends on control parameters. These control parameters comprise the ecological embedding of cognition, i. e. the environment of the cognitive agent. Therefore, the dynamical approach views cognition predominantly with reference to its embeddings, as embodied cognition [39] or situated cognition [16].

Intentionality and Representation

The dynamical view in cognitive science thus naturally leads to the concepts of embodied and situated cognition. Consequently, intentionality of mental acts can be discussed under these premises. Rather than searching, in a top–down fashion, for a fundament of experienced intentionality of the mind, the dynamical systems heuristic can be formulated differently: If the mind is conceptualized as arising from self-organization processes, is there a natural way by which emergent mental acts can be conceptualized as *being about something*?

Intentionality was introduced as a characterizing property of mental acts by Franz von Brentano, a philosopher and early psychologist. In the late 19th century Brentano was professor at the University of Vienna where Edmund Husserl, Sigmund Freud, Carl Stumpf and other later protagonists of philosophy and psychology counted among his students. According to Brentano [4], mental phenomena are always directed towards an object (the intentional object). In other words, mental states contain within themselves something else ('*intentionale Inexistenz*', i. e. intentional existence within). No physical phenomenon has such intentional content, therefore according to Brentano intentionality constitutes the distinctive feature of the mind. Many concepts of contemporary cognitive psychology are in this sense intentional. The basic concepts *goal, wish, plan* and *intention* of volitional psychology [15,28] obviously have intentional content. The same applies to *achievement*, *valence* and *need* in motivation psychology [32]. *Affects* and *emotions* are also generally about something and thus intrinsically intentional concepts.

It should be noted, however, that intentionality may not provide a sufficient and necessary condition for a state or process to be mental. Not all mental states are intentional; some emotional states (e. g. moods such as a pervasive feeling of melancholy or of serenity) do not necessarily possess intentional content because they are not about something; yet they are undoubtedly experienced mental states. Furthermore, intentionality is likely not the only property that distinguishes mental from physical systems. Many current philosophers of mind suggest that in addition to intentional content, the phenomenal content of mental states must be considered [33]. This immediately leads to the topic of consciousness, which cannot be addressed here.

A concept very closely related to intentionality is *representation*. Representation plays a central role in cognitive psychology (as schema: Neisser [35]), in philosophy of mind (as language of thought: Fodor [11]) and in artificial intelligence (as physical symbol system: Newell [36]). In all of these fields representation of knowledge is a foundational concept, yet at the same time constitutes a core problem. If a cognitive agent is to have knowledge of its environment, the obvious idea is that there must be some kind of mental map or mental model of the environment 'inside' the agent. On the basis of information thus represented, the agent would then perform cognitive actions such as memory functions, manipulations for problem solving, and the like. One may note the close analogy of representation with the intentional object.

The naive, folk-psychological intuition of an inner map or depiction of the environment on which cognitive functions can be performed is however unsatisfac-

tory [2,9]. The map concept per se is not explanatory. The reason for this is simple: if the problem of an information-processing agent is to make sense of its physical environment, the solution to this problem will not be alleviated at all by representation alone; the agent's task of making sense of the *represented* environment is just as demanding. The theory of direct perception [13] has therefore proposed that information pick-up must occur right at the moment of perception, without any representational interlude. For analogous reasons, the 'storehouse metaphor' of memory has been rejected by researchers of cognitive science [14]. Memory is likely not a passive store out of which represented items, the memory engrams, can be retrieved at a later time, but a more active, constructive process. Eyewitness research (e. g., the so-called false memory syndrome, [31]) has emphasized how modifiable and adaptive the represented contents of memory actually are. Representation-as-mapping has therefore been criticized as merely providing a pseudo solution to a deeper problem, which is likely the very problem of naturalizing intentionality.

Synergetics

We have proposed that cognitive phenomena have attributes of dynamical systems, and that higher functions and more complicated contents of the mind are constructed bottom–up from simpler components by a process of self-organization. It is the goal of this line of argument to show that intentionality can be conceptualized using this framework. Since dynamical systems theory, as well as self-organization theory, are mathematical tools that have been developed for applications in physics, biology and other natural sciences, we may thereby approach the goal of a naturalization of intentionality. The final step prior to formulation of this naturalization proposal is to introduce self-organization theory.

We will rely largely on the interdisciplinary modeling approach of synergetics that deals with complex systems, i. e. systems composed of multiple components [17,18,19]. By way of interactions, these components can produce new qualitative features on macroscopic scales. Synergetics focuses on the emergence of these new qualities, while addressing the question of whether there are general principles governing the behavior of complex systems when such qualitative changes occur. In a large class of systems, it has been shown that they become accessible to unifying mathematical and conceptual approaches. A paradigmatic system is the Bénard system (e. g. [1]), which is comprised of a layer of fluid heated with temperature T_2 from below. The temperature at the upper surface of the fluid is T_1. Beyond a critical value of $\Delta T = T_2 - T_1$ extended coordinated motion of the components of the fluid system emerge. Compared to the erratic Brownian motion of the single components, these patterns are an example of the emergence of the new qualities focused on by synergetics.

Synergetics is based on observations that the behavior of many systems is strongly determined by environmental conditions. These conditions may be divided into constant (structural) constraints (e. g. the shape of solid walls that confine fluid systems such as the Bénard system), and into further environmental conditions that energize the systems (e. g. the heat source that drives the Bénard system). In the mathematical approach, these latter driving conditions are expressed by control parameters. In many cases, control parameters take the form of externally applied gradients, which are imposed upon the system from without, such as, for instance, the difference in temperature ΔT of the Bénard experiment. The general strategy in synergetics sets out from a state of a system that is already known under a certain control parameter value. When one or several control parameters are changed, this system can become unstable and show a tendency to leave its state to develop a new structure or behavior. The system in question is described by the states of its individual components, by means of a state vector $\boldsymbol{q}$. The individual components in the Bénard system, for example, are the motions of single fluid molecules; components may also be, with respect to applications in psychology, the attributes of members within a social group or neurons in the brain.

Synergetics shows that the behavior of the system close to instability points is described and determined by few quantities, namely the order parameters. In the case of a single order parameter n of a complex system a typical equation reads

$$\frac{\mathrm{d}n}{\mathrm{d}t} = cn \tag{1}$$

where c is the 'effective' control parameter.

$$\begin{aligned} \text{For} \quad c > 0, &\quad n \text{ increases exponentially,} \\ c < 0, &\quad n \text{ decreases exponentially,} \\ c = 0, &\quad n \text{ remains constant .} \end{aligned}$$

As was mentioned, c denotes the control parameter, a relevant parameter imposed on the system from outside, i. e. from the environment of the system. The generally few order parameters enslave, i. e. determine, the behavior of the many individual components. This implies an enormous information compression, because the description of the order parameters alone, rather than of each component, suffices. In the case of the Bénard system, description of

the coordinated motion yields a much more parsimonious description of system behavior than the description of all molecular movements in the Brownian motion state of the fluid. While they are being determined by the order parameters, it is the individual components that react on the order parameters and, by so doing, even generate the latter. The relationship between order parameters and components is, therefore, founded on *circular causality*, which can explain the generally avalanche-like onset of, and transition between, macroscopic states. In other words, this theory favors neither top–down nor bottom–up modeling but claims that both processes are entangled. Order parameters, after they have been generated in this fashion, quite often exhibit very simple behavior, for instance asymptotic stability.

Obviously, the system depicted in Fig. 1 is an open system. Self-organizing systems are invariably open systems in that they depend on control parameters. In terms of thermodynamics, they are non-equilibrium systems.

Let us however first focus on closed systems, in other words, systems in thermal equilibrium. Classic thermodynamics deals with closed systems throughout. The probability of all configurations of components within the (closed) system can be estimated. When dealing with a complex system consisting of a multitude of components, a great many possible realizations of the state vector $\boldsymbol{q}$ exist, namely the number of all combinations W of the states of components. Only a small fraction of these realizations are seen as regular, well-organized patterns. The vast majority of realizations, however, will represent a state of mixture. Should ordered patterns exist as an initial condition, it is far more probable that the temporally consecutive system states will be characterized by less order. This is due to the statistical fact that the majority of possible consecutive states will be states with less order rather than states with the same, or even a higher degree, of order. Within a thermodynamics context, this touches on the concept of entropy S (disorder) in accordance with Boltzmann's statistical approach, in which S is directly related to the number of combinations W. The second law of thermodynamics states that any real closed system can only proceed in the direction of increasing entropy, hence following a maximum entropy principle. Hence, the spontaneous generation of order is highly improbable as indeed a spontaneous generation of disorder is to be expected. In other words, the emergence of pattern from a state of mixture requires explanation; the explanation is that the phenomenon of self-organization is driven by an external source, so that the premises of closed systems do not apply. Since the concept of entropy is defined only for equilibrium or close-to-equilibrium systems one may base the discussion of self-organizing systems on the concept of information [19,47].

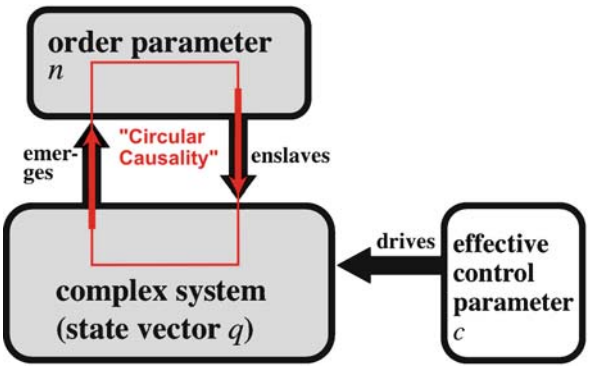

Intentionality: A Naturalization Proposal on the Basis of Complex Dynamical Systems, Figure 1
Schematic illustration of circular causality as viewed in synergetics

Some authors have applied the laws of thermodynamics in order to allow the study of self-organizing systems. The 'restated second law' of thermodynamics [42,43] addresses non-equilibrium systems, i. e. systems that are forced away from equilibrium by the application of gradients. The degree to which a system is moved away from equilibrium is measured by the gradients imposed on the system. As soon as such gradients dwell in the system's environment, the system will, as a consequence of the restated second law, "[...] utilize all avenues available to counter the applied gradients. As the applied gradients increase, so does the system's ability to oppose further movement from equilibrium" [42]. Schneider and Kay's restatement of the second law avoids some of the problems of defining entropy and entropy production by focusing on the destruction of gradients instead.

It should be kept in mind that this 'destruction of gradients' is only virtual (in analogy to the principle of virtual work in mechanics), because in open systems gradients are generally maintained by the environment. If, however, the self-organizing system and its *finite* environment act as a closed system, the gradient reduction becomes real. In other words, the effective control parameter c depends on the order parameter n,

$$c = c_0 - \alpha n \tag{2}$$

where c_0 is the control parameter prescribed from the outside and a constant. The effective control parameter (2) obeys the differential equation

$$\frac{\mathrm{d}c}{\mathrm{d}n} = -\frac{\mathrm{d}V}{\mathrm{d}c} \tag{3}$$

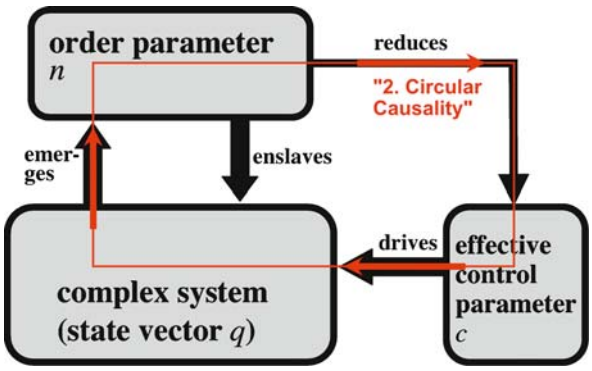

Intentionality: A Naturalization Proposal on the Basis of Complex Dynamical Systems, Figure 2
Schematic illustration of the relationship between control parameter and order parameter ('second circular causality')

where

$$V = \alpha c \tag{4}$$

is a potential and thus the right-hand side of (3) a gradient.

Discussing the time evolution of n and c according to (1) and (2), we assume that, initially, $n = n(0)$ is close to zero and $c \approx c_0 > 0$.

Thus, according to (1), n increases exponentially. As a consequence, according to (2), c decreases, and the exponential increase of n slows down. This process goes on until $c = 0$ and n reaches a time-independent, i. e. steady, state. In practice, the transition to the new state is completed while the gradient has been reduced. In the Bénard example, the coordinated motion patterns have consumed the applied temperature difference and have reduced ΔT to 0. Then the motion patterns subside and steady state remains. This relationship between emergent pattern and control parameter is thus in line with the notion of gradient destruction. The reduction of c by n establishes a second kind of *circular causality*, which is schematically illustrated in Fig. 2.

Discussion: Naturalization of Intentionality

The steps above have put us in a position where the naturalization of intentionality comes within reach. We propose that this can be achieved on the basis of the properties of self-organizing complex systems. In this section we will therefore discuss to what extent the features of intentionality (listed in Sect. "Definition of the Subject") can be approximated by such systems.

Aboutness

Intentionality implies that a system state is about something else, namely the intentional object. In terms of cognitive psychology, this process is called representation, by which a 'cognitive map' of the object is generated particularly during perceptual or memory processes.

To be capable of intentionality and representation, a minimum requirement is that the intentional system must be an open system. Many open systems can provide representations in the sense of mappings of environmental impacts. The silver particles of a light-sensitive surface of a photographic film can 'represent' the objects in front of the lens, however in a trivial, weak sense. As has been shown in the previous section, a self-organizing system can likewise 'represent' and thus generate the feature of being about something. In the latter system the order parameter is a component in the loop labeled 'second circular causality' (Fig. 2). Within this loop the complex system 'represents' an external object by the generation of an order parameter. The intentional object in this case is the external control parameter. The environmental condition described by control parameters is what self-organized patterns 'are about'.

The mechanisms of representation are clearly divergent in these two systems, as is the nature of the intentional objects. In the photographic system the mapping of the environmental objects onto the representing system is unidirectional, whereas in the self-organizing system there is continuous interaction between environmental objects and system. This circularity is illustrated in Fig. 2 and provides an important and desirable aspect of the kind of aboutness realized by such systems. Circularity guarantees that an intentional system is capable of exerting a retrograde effect back on what has been represented. This effect is generally a reduction of the gradient quantified by the control parameter.

Functionality

Intentionality must be functional in order to make representation explanatory, i. e. make representation more than just a mapping of environmental states onto the system's state.

The functionality of open systems has previously been approached from the angle of thermodynamics. Schneider and Sagan [43] pointed out that self-organizing systems maintain and increase their levels of organization by dissipating non-equilibrium gradients. If the gradient is to be kept constant, the demand on free energy (so-called exergy) that must be provided by an external source increases as the system becomes more organized. Alternatively, the efficiency of the system can be defined as the ratio of the change of work and the change of the gradient driving the system. This can be shown by findings

in simple physical systems that generate patterns; e. g. Bénard cells reduce the temperature gradient more efficiently as soon as they have generated ordered convection structures. Efficiency has thus increased in the self-organized convection regime in comparison to the linear conduction regime of the fluid. Analogous relationships are found in further self-organizing systems such as the laser when output power is plotted against input power [17]. In other words, pattern formation in these open systems is in the service of gradient reduction. The association of pattern formation with gradient reduction makes pattern formation functional.

In situations of multistability, several patterns are possible, so that each of these can be functional in reducing the gradients imposed on the system. These alternative patterns can be associated with different efficiencies. It is theoretically suggestive that the two circular loops pointed out in Fig. 1 and 2 interact and thereby create a darwinistic scenario in which a competition between microscopic modes q arises; the environmental forces c exert a selective impact on this competition of modes. From a mathematical modeling point of view it is not however clear if the optimal pattern is necessarily selected, nor under which circumstances the optimal pattern will be chosen. In some systems, such as quadruped movement coordination, it was empirically found that the specific pattern that provides the most efficient behavior (measured as the metabolic cost of transport of the animal) will be realized by the system [25]. The generalizability of such findings of optimality is not yet established.

Mental-Likeness

In this discussion we have so far found that the aboutness of intentionality and representation can be modeled by open systems. Only a subclass of open systems provides functional representations. Especially, certain self-organizing non-equilibrium systems appear to show the two features of aboutness and functionality.

It is obviously useful to reserve the predicate 'intentional' for mental systems alone. Thus, even though some physical self-organizing systems may show circular causality loops and thus stand the tests of aboutness and functionality, we would still categorize systems such as lasers, Bénard cells etc. as being not mental. We may say that such self-organizing physical systems behave 'as if they were intentional', i. e. they are proto-intentional systems. Let us finally investigate under which conditions self-organizing non-equilibrium systems may also show mental features and, with this, conclude the proposal of naturalizing intentionality. This final step is basically a discussion of the validity of the statement that self-organizing systems can show intentionality.

- *Complexity reduction* is a core hallmark of mental processes. The ability of a system to simplify, group and coordinate environmental information is a necessary premise for any system to be mental. This property is addressed in the circular causality concept of synergetics illustrated in Fig. 1. With respect to information compression, self-organizing physical systems are mental-like.
- *Stability* together with related signatures of stability (e. g. hysteresis, critical fluctuations) is a further mental property, which is empirically well founded especially in the psychology of perception. Again, the emergent order parameters of self-organizing physical systems generally show this property.
- *Autonomy* is required of intentional mental systems, i. e. mental systems must be able to function in the absence of external agents. This ability addresses the 'homunculus-problem' that has already been introduced together with the feature of functionality. We may say that generally self-organizing systems do not require external supervision for producing order parameters (therefore, *self*-organizing). This does not rule out that several autonomous systems may be nested inside one another (cf. Minsky's society of mind [34], or the notion of subsumption architecture of Brooks' [5] robotic agents).
- What is the nature of the intentional content? Which aspect of the environment is being represented by self-organizing systems? Order parameters are intentional with respect to those environmental parameters that drive the system. In other words, the intentional content is generally connected to what energizes the system. This is a satisfactory model in all those instances where mental intentionality is of a motivational character, resembling a psychological 'drive' [12]. In the introduction we specified wishes, desires, affects, intentions, goals and the like as intentional. Such intentional states can be directly modeled by self-organizing systems that act to reduce the driving parameters (in Freudian terms, drive reduction). The self-organization model, however, is less applicable whenever intentionality is of a language-like, propositional type [11]. It is quite a different task to model, for instance, the intentionality inherent in the belief that "there is a unicorn grazing in the garden", using simple self-organizing systems. Therefore, the discussion of mental-likeness of self-organizing systems remains restricted to nonlinguistic intentionality.

- The only mental systems known to date, despite the efforts of several decades of artificial intelligence research, are *neuronal networks* inside biological organisms. Therefore, can the argumentation above be applied to brain dynamics and to pattern formation in neural networks? One may then associate the gradient of c with intentionality in a neurocognitive sense. Haken [20] has discussed synchronization, i. e. self-organization, of neural nets using various mathematical frameworks. Haken and Tschacher [23] have specifically addressed the reduction of the effective control parameter on the basis of the Wilson–Cowan equations describing cortical dynamics [51]. Haken and Tschacher suggested that the general findings on circular causality in the relationship between order parameter and control parameter can be readily applied to neural networks.

In conclusion, we have argued that a formulation of intentionality is feasible on the basis of the theory of nonlinear dynamical systems. When such systems are removed from thermodynamical equilibrium they acquire the capability of producing self-organized patterns. Pattern formation consequently puts the systems in a specific relationship to environmental parameters. Owing to the accompanying circular loops, these systems show aboutness, the defining property of intentionality, as well as functionality, which is essential for making the aboutness of intentional states explanatory. Finally, it can be shown that some of these self-organizing systems are mental-like because their behavior empirically shows signatures of mental phenomena beyond the features of intentionality. We therefore conclude that intentionality can be naturalized to a considerable extent using non-equilibrium complex systems.

Some caveats have turned up during this discussion, which may provide points of departure for future research. First, it is as yet unclear if the resulting self-organized patterns obey an optimization principle. In some empirically described systems showing multistability it could be shown that the more optimal pattern wins the competition among order parameters, but the generality of such findings is yet to be corroborated. Correspondingly, Haken and Tschacher [23] have discussed 'second circular causality' in the case of only one order parameter. A mathematical model comprising $m > 1$ order parameters would be a desirable next step. Second, we found that naturalization of intentional states is achievable when these states comprise motivation, goals, intentions, or drives; these may be viewed as basic intentional states related to behavioral strivings. At the moment, however, the intentionality problem appears intractable when these intentional states are of a symbolic and propositional nature. The difficulties of modeling Fodor's language-of-thought do indeed complicate the modeling of intentionality beyond its basic form (cf. Churchland [8], p. 304). At the present time, we may consider non-linguistic intentionality as the primary problem, which in fact seems accessible. The solution to this problem might support the clarification of symbolic, secondary intentionality in the future. This restriction to non-linguistic intentionality is likely related to a further limitation. Contrary to von Brentano's supposition, mental phenomena may have to be viewed as intentional *as well as* phenomenal. Systems theory addresses solely the primary aspects of intentionality, yet not the 'hard problem' [6] of the phenomenal nature of intentionality. It does not lie within the scope of the present treatment to determine if and how the dynamical view can illuminate the hard question of consciousness.

Future Directions

The topic of intentionality has numerous implications beyond the philosophy of mind and theoretical psychology.

In psychiatry and in psychotherapy research, intentionality is a topic of considerable significance because intentional mental acts and states are often characteristically altered or disturbed during a mental disorder. Many such psychopathological conditions are found especially among the symptoms of schizophrenia, such as disorders of formal thought as well as of thought content, disorders of perception, and ego disorders. The symptoms are heterogeneous and manifold. A majority of these symptomatic alterations, however, involve changes in the cognitive coordination of the patients [40]. Recent schizophrenia research has shown that a considerable portion of the variance of psychotic symptoms can be explained by cognitive coordination measures [48]. While no generally accepted encompassing theory of schizophrenia exists, a theory of intentionality may have the potential to contribute to progress in psychopathology. It may help to link the phenomenology and neurobiology of schizophrenia and other psychiatric disorders by introducing a dynamical systems perspective.

Artificial intelligence is a completely different field, but is also confronted with intentionality as a core problem. In recent decades, the computational approach to machine intelligence has failed to a large extent; especially, no mental-like intelligence could be generated. Consequently the field has turned to the more basic tasks of designing autonomous agents and robots with rudimentary adaptivity in the real world [3,38]. In this framework, intelligence is expected to be closely associated with embodi-

ment (hence, embodied intelligence, mind–body co-evolution), rather than with symbol manipulation and the programming of symbol systems. The latter constituted the classic approach to artificial intelligence. One of the novel directions of development is using emergence principles for the design of intelligent agents or multi-agent systems; this is closely related to the view on intentionality proposed here. The engineering approach of embodied agents design and the dynamical systems proposal of intentionality have the potential of subserving each other in artificial intelligence research.

Bibliography

Primary Literature

1. Bianciardi C, Ulgiati S (1998) Modelling entropy and exergy changes during a fluid self-organization process. Ecol Modelling 110:255–267
2. Bickhard M, Terveen L (1995) Foundational issues in artificial intelligence and cognitive science. Holland, Amsterdam
3. Braitenberg V (1986) Künstliche Wesen: Verhalten kybernetischer Vehikel. Vieweg, Braunschweig
4. Brentano F von (1874) Psychologie vom empirischen Standpunkte. Duncker & Humblot, Leipzig
5. Brooks R (1991) Intelligence without Representation. Artificial Intell 47:139–159
6. Chalmers D (1996) The conscious mind. Oxford University Press, Oxford
7. Churchland P (1986) Neurophilosophy: Toward a unified science of the mind-brain. MIT Press, Cambridge
8. Churchland P (2002) Brain–Wise: Studies in neurophilosophy. MIT Press, Cambridge
9. Clancey WJ (1993) Situated action. A neurophysiological interpretation response to Vera and Simon. Cogn Sci 17:87–116
10. Clark A (1997) Being there: Putting brain, body, and world together again. MIT Press, Cambridge
11. Fodor J (1975) The language of thought. Crowell, New York
12. Freud S (1923) Das Ich und das Es [published in English (1949) The Ego and the Id. The Hogarth Press Ltd., London]
13. Gibson JJ (1979) The ecological approach to visual perception. Houghton Mifflin, Boston
14. Glenberg A (1997) What memory is for. Behav Brain Sci 20:1–19
15. Gollwitzer PM, Bargh GA (eds) (1996) The psychology of action: Linking cognition and motivation to behavior. Guilford, New York
16. Greeno J, Moore J (1993) Situativity and symbols: Response to Vera and Simon. Cogn Sci 17:49–59
17. Haken H (1977) Synergetics – An introduction. Nonequilibrium phase-transitions and self-organization in physics, chemistry and biology. Springer, Berlin
18. Haken H (1996) Principles of brain functioning: A synergetic approach to brain activity, behavior, and cognition. Springer, Berlin
19. Haken H (2000) Information and self-organization: A macroscopic approach to complex systems. Springer, Berlin
20. Haken H (2002) Brain dynamics. Synchronization and activity patterns in pulse-coupled neural nets with delays and noise. Springer, Berlin
21. Haken H, Stadler M (eds) (1990) Synergetics of cognition. Springer, Berlin
22. Haken H, Kelso J, Bunz H (1985) A theoretical model of phase transitions in human hand movements. Biol Cybern 51:347–356
23. Haken H, Tschacher W (submitted) A theoretical model of intentionality based on self-organizing systems
24. Harnad S (1990) The symbol grounding problem. Physica D 42:335–346
25. Hoyt DT, Taylor CR (1981) Gait and the energetics of locomotion in horses. Nature 292:239–240
26. Kelso JAS (1995) Dynamic Patterns: The self-organization of brain and behavior. MIT Press, Cambridge
27. Köhler W (1920) Die physischen Gestalten in Ruhe und in stationärem Zustand. Vieweg, Braunschweig
28. Kuhl J, Beckmann J (1994) Volition and personality. Action versus state orientation. Hogrefe and Huber Publishers, Göttingen
29. Leopold DA, Logothetis NK (1999) Multistable phenomena: changing views in perception. Trends Cogn Sci 3:254–264
30. Lewin K (1936) Principles of topological psychology. McGraw–Hill, New York
31. Loftus EL (2003) Our changeable memories: Legal and practical implications. Nat Rev Neurosci 4:231–234
32. McClelland DC, Atkinson JW, Clark RA, Lowell EL (1953) The achievement motive. Van Nostrand, Princeton
33. Metzinger T (2003) Being no one: The self-model theory of subjectivity. MIT Press, Cambridge
34. Minsky M (1985) Society of mind. Simon & Schuster, New York
35. Neisser U (1976) Cognition and reality. Principles and implications of cognitive psychology. Freeman, San Francisco
36. Newell A (1980) Physical symbol systems. Cogn Sci 4:135–183
37. Nicolis G, Prigogine I (1977) Self-organization in non-equilibrium systems. Wiley, New York
38. Pfeifer R, Bongard JC (2006) How the body shapes the way we think. A new view of intelligence. MIT Press, Cambridge
39. Pfeifer R, Scheier C (1999) Understanding intelligence. MIT Press, Cambridge
40. Phillips WA, Silverstein SM (2003) Convergence of biological and psychological perspectives on cognitive coordination in schizophrenia. Behav Brain Sci 26:65–138
41. Popper K, Eccles J (1977) The self and its brain. Springer, Berlin
42. Schneider ED, Kay JJ (1994) Life as a manifestation of the second law of thermodynamics. Math Comp Modelling 19:25–48
43. Schneider ED, Sagan D (2005) Into the cool. Energy flow, thermodynamics and life. University of Chicago Press, Chicago
44. Thelen E, Smith LB (1994) A dynamic systems approach to the development of cognition and action. MIT Press, Cambridge
45. Tschacher W (1997) Prozessgestalten [Processual gestalts]. Hogrefe, Göttingen
46. Tschacher W, Dauwalder J-P (eds) (2003) The dynamical systems approach to cognition. World Scientific, Singapore
47. Tschacher W, Haken H (2007) Intentionality in non-equilibrium systems? The functional aspects of self-organized pattern formation. New Ideas in Psychol 25:1–15
48. Tschacher W, Schuler D, Junghan U (2006) Reduced perception of the motion-induced blindness illusion in schizophrenia. Schizophrenia Res 81:261–267
49. van Gelder T (1998) The dynamical hypothesis in cognitive science. Behav Brain Sci 21:615–628
50. Wertheimer M (1912) Experimentelle Studien über das Sehen

von Bewegungen [Experimental studies on the perception of motion]. Z Psychol 61:165–292
51. Wilson HR, Cowan JD (1972) Excitatory and inhibitory interactions in localized populations of model neurons. Biophys J 12:1–24

Books and Reviews

Beckermann A (2001) Analytische Einführung in die Philosophie des Geistes. de Gruyter, Berlin
Carter R (2002) Consciousness. Weidenfeld & Nicolson, London
Chaisson EJ (2001) Cosmic evolution. The rise of complexity in nature. Harvard University Press, Cambridge
Ciompi L (1982) Affektlogik. Klett–Cotta, Stuttgart [published in English (1988): The psyche and schizophrenia. The bond between affect and logic. Harvard University Press, Cambridge]
Dennett DC (1987) The intentional stance. MIT Press, Cambridge
Dennett DC, Kinsbourne M (1992) Time and the observer: The where and when of consciousness in the brain. Behav Brain Sci 15:183–247
Guastello S, Koopmans M, Pincus D (eds) (2008) Chaos and complexity in psychology. Theory of nonlinear dynamical systems. Cambridge University Press, Cambridge
Kauffman S (1993) The origins of order. Self-organization and selection in evolution. Oxford University Press, New York
Kruse P, Stadler M (eds) (1995) Ambiguity in mind and nature. Springer, Berlin
Newell A, Simon HA (1972) Human problem solving. Prentice–Hall, Englewood Cliffs
Port R, van Gelder TJ (eds) (1995) Mind as motion: Explorations in the dynamics of cognition. MIT Press, Cambridge
Searle JR (1998) The rediscovery of mind. MIT Press, London
Singer W, Gray CM (1995) Visual feature integration and the temporal correlation hypothesis. Annu Rev Neurosci 18:555–586
Storch M, Cantieni B, Hüther G, Tschacher W (2006) Embodiment. Huber, Bern
Swenson R, Turvey MT (1991) Thermodynamic reasons for perception–action cycles. Ecol Psychol 3:317–348
Tschacher W, Dauwalder J-P (eds) (1999) Dynamics, synergetics, autonomous agents. Nonlinear systems approaches to cognitive psychology and cognitive science. World Scientific, Singapore
Tschacher W, Scheier C (2003) Der interaktionelle Ansatz in der Kognitionswissenschaft: Eine Positionsarbeit zu Konzepten, Methoden und Implikationen für die Psychologie [The interactional approach to cognitive science: Concepts, methods, and implications for psychology]. Zeitschrift für Psychologie 211:2–16
Varela F, Thompson E, Rosch E (eds) (1991) The embodied mind. MIT Press, Cambridge
Varela FJ (1995) Resonant cell assemblies: A new approach to cognitive function and neuronal synchrony. Biol Res 28:81–95

Interaction Based Computing in Physics

Franco Bagnoli
University of Florence, Florence, Italy

Article Outline

Glossary
Definition
Introduction: Physics and Computers
From Trajectories to Statistics and Back
Artificial Worlds
Discussions and Conclusions
Bibliography

Glossary

Nonlinear system A system composed by parts whose combined effects are different from the sum of the effects of each part.

Extended system A system composed by many parts connected by a network of interactions that may be regular (lattice) or irregular (graph).

Graph, lattice, tree A graph is set of nodes connected by links, oriented or not. If the graph is translationally invariant (it looks the same when changing nodes), it is called a (regular) lattice. A disordered lattice is a lattice with a fraction of removed links or nodes. An ordered set of nodes connected by links is called a path. A closed path not passing on the same links is a loop. A cluster is a set of connected nodes. A graph can be composed by one cluster (a connected graph) or more than one (a disconnected graph). A tree is a connected graph without loops.

Percolation The appearance of a "giant component" (a cluster that contains essentially all nodes or links) in a graph or a lattice, after adding or removing nodes or links. Below the percolation threshold the graph is partitioned into disconnected clusters, none of which contains a substantial fraction of nodes or links, in the limit of infinite number of nodes/links.

State of a system A complete characterization of a system at a given time, assigning or measuring the positions, velocities and other dynamical variables of all the elements (sometimes called a configuration). For completely discrete systems (cellular automata) of finite size, the state of the system is just a set of integer numbers, and therefore the state space is numerable.

Trajectory A sequence of states of a system, labeled with the time, i. e., a path in the state space.

Probability distribution The probability of finding a system in a given state, for all the possible states.

Mean field An approximate technique for computing the value of the observables of an extended system, neglecting correlations among parts. If necessary, the dynamics is first approximated by a stochastic process. In its simpler version, the probability of a state of the sys-

tem is approximated by the product of the probability of each component, neglecting correlations. Since the state of two components that depend on a common "ancestor" (that interact with a common node) is in general not uncorrelated, and this situation corresponds to an interaction graph with loops, the simplest mean field approximation consists in replacing the graph or the lattice of interactions with a tree.

Monte-Carlo A method for producing stochastic trajectories in the state space designed in such a way that the time-averaged probability distribution is the desired one.

Critical phenomenon A condition for which an extended system is correlated over extremely long distances.

Definition

Physics investigation is based on building models of reality: in order for a phenomenon to be *understood*, we need to represent it in our minds using a limited amount of symbols. However, it is a common experience that, even using simple "building blocks" one usually obtains systems whose behavior is quite complex. In this case one needs to develop new languages and new *phenomenological* models in order to manage this "complexity".

Computers have changed the way a physical model is studied. Computers may be used to *calculate* the properties of a very complicated model representing a real system, or to investigate *experimentally* what are the essential ingredients of a complex phenomenon. In order to carry out these explorations, several basic models have been developed, which are now used as building blocks for performing simulations and designing algorithms in many fields, from chemistry to engineering, from natural sciences to psychology. Rather than being derived from some fundamental law of physics, these blocks constitute *artificial worlds* still to be completely explored.

In this article we shall first present a pathway from Newton's laws to cellular automata and agent-based simulations, showing (some) computational approaches in classical physics. Then, we shall present some examples of *artificial worlds* in physics.

Introduction: Physics and Computers

Some sixty years ago, shortly after the end of Second World War, computers become available to scientists. Computers were used during the last years of the war for performing computations about the atomic bomb [30,41].

Up to then, the only available tool, except experiments, was paper and pencil. Starting with Newton and Leibnitz, humans discovered that continuous mathematics (i. e., differential and integral calculus) allowed to derive many consequences of a given hypothesis just by the manipulation of symbols. It seemed natural to express all quantities (e. g., time, space, mass) as continuous variables. Notice however that the idea of a continuous number is not at all "natural": one has to learn how to deal with it, while (small) integer numbers can be used and manipulated (added, subtracted) by illiterate humans and also by many animals. A point which is worth to be stressed is that any computation refers to a model of certain aspects of reality considered most important, while others are assumed to be not important

Unfortunately most of human-accessible explorations in physics are limited to almost-linear systems, or systems whose effective number of variables is quite small. On the other hand, most of naturally occurring phenomena can be "successfully" modeled only using nonlinear elements. Therefore, most of pre-computer physics is essentially linear physics, although astronomers (like other scientists) used to integrate numerically, by hand, the nonlinear equations of gravitation, in order to compute the trajectories of planets. This computation, however, was so cumbersome that no "playing" with trajectories was possible.

While analog computers have been used for integrating differential equations, the much more flexible digital computers are deterministic discrete systems. The way of working of a (serial) computer is that of a very fast automaton, that manipulates data following a program.

In order to use computers as fast calculators, scientists ported and adapted existing numerical algorithms, and developed new ones. This implied the development of techniques able to *approximate* the computations of continuous mathematics using computer algebra. However, numbers in computers are not exactly the same as human numbers, in particular they have finite (and varying) precision.

This intrinsic precision limit has deep consequences in the simulations of nonlinear systems, in particular of chaotic ones. Indeed, chaos was "numerically discovered" by Lorenz [38] after the observation that a simple approximation, a number that was retyped with fewer decimals, caused a macroscopic change in the trajectory under study.

With all their limits, computers can be fruitfully used *just* to speed-up computations that *could* eventually be performed by humans. However, since the increase in velocity is of several order of magnitude, it becomes possible to include more and more details into the basic model of the phenomenon under investigation, well beyond what would be possible with an army of "human computers". The idea of exploiting the *brute power* of fast comput-

ers has originated a fruitful line of investigation in *numerical physics* especially in the field of chemistry, biological molecules, structure of matter. The power of computers has allowed for instance to include quantum mechanical effects in the computation of the structure of biomolecules [16], and although these techniques may be targeted as "brute force", the algorithms developed are actually quite sophisticated.

However, a completely different usage of computers is possible: instead of exploiting them for performing computations on models that already proved to approximate the reality, one can use computers as "experimental" apparatus to investigate the patterns of *theoretical* models, generally non-linear. This is what Lorenz did after having found the first example of chaos in computational physics. He started simplifying his equations in order to enucleate the minimal ingredients of what would be called the *butterfly effect*.

Much earlier than Lorenz, Fermi, Pasta and Ulam (and the programmer Tsingou [20]) used one on the very first available computers to investigate the basis of statistical mechanics: how energy distributes among the oscillation modes of a chain of nonlinear oscillators [25].

Also in this case the model is simplified at its maximum, in order to put into evidence what are the fundamental ingredients of the observed pattern, and also to use all the available power of computers to increase the precision, the duration and the size of the simulation.

This simplification is even more fruitful in the study of systems with many degrees of freedom, that we may denote generically as *extended systems*. We humans are not prepared to manipulate more that a few symbols at once. So, unless there is a way of grouping together many parts (using averages, like for instance when considering the pressure of a gas as an average over extremely many particle collisions), we are in difficulties in *understanding* such systems. They may nevertheless be studied performing "experiments" on computers. Again, the idea is that of simplifying at most the original model, in order to isolate the fundamental ingredients of the observed behavior. It is therefore natural to explore systems whose *physics* is different from the usual one. These *artificial worlds* are preferably formulated in discrete terms, more suitable to be implemented in computers (see Sect. "Artificial worlds").

This line of investigation is of growing interest today: since modern computers can easily simulate thousands or millions of *elementary automata* (often called *agents*), it may be possible to design *artificial worlds* in which *artificial people* behave similarly to real humans. The rules of these worlds are not obtained from the "basic laws" of the real one, since no computer can at present simulate the behavior of all the elements of even a small portion of matter. These rules are designed so to behave similarly to the system under investigation, and to be easily implemented in digital terms. There are two main motivations (or hopes): to be able to understand real complex dynamics by studying simplified models, and to be so lucky to discover that a finely-tuned model is able to reproduce (or forecast) the behavior of its real counterpart.

This is so promising that many scientists are performing experiments on these artificial worlds, in order to extract their principal characteristics, to be subsequently analyzed possibly using paper and pencil!

In the following we shall try to elucidate some aspects of the interplay between computer and physics. In Sect. "From Trajectories to Statistics and Back", we shall illustrate possible logic pathways (in classical mechanics) leading from Newton's equations to research fields that use computers as investigative tool, like agent-based investigations of human societies. In Sect. "Artificial worlds", we shall try to present succinctly some example of "artificial worlds" that are still active research topics in theoretical physics.

From Trajectories to Statistics and Back

The outline of this Section is the following. Physics is often denoted the most "fundamental" science, and one may think that, given powerful enough computers, one should be able to reconstruct any experimental situation simply implementing the fundamental laws of physics. I would like to show that any investigation is based on models, requiring approximations and additional assumptions, and that any change of scale implies a change of model. However, an important suggestion from physics is that similar models can be used to interpret different situations, and therefore the associated computational techniques can be "reused" in different contexts. We shall follow this line from Newton's equations to agent-based modeling.

Let us assume that a working model of the reality can be built using a set of dynamical equations, for instance those of classical mechanics. We shall consider the model of a system formed by many particles, like a solid or a fluid. The state of the resulting system can be represented as a point in a high-dimensional space, since it is given by all coordinates and velocities of all particles. The evolution of the system is a trajectory in such space. Clearly, the visualization and the investigation of such a problem is challenging, even using powerful computers.

Moreover, even if we are able to compute one or many trajectories (in order to have an idea of fluctuations), this

does not imply that we have *understood* the problem. Let us consider for instance the meteorology: one is interested in the probability of rain, or in the expected wind velocity, not in forecasting the trajectories of all molecules of air. Similarly in psychology, one is interested in the expected behavior of an individual, not in computing the activity of all neurons in his brain.

Since physics is the oldest discipline that has been "quantified" into equations, it may be illuminating to follow some of the paths followed by researchers to "reduce" the complexity of a high-dimensional problem to something more manageable, or at least simpler to be simulated on a computer.

In particular, we shall see that many approaches consist in "projecting" the original space onto a limited number of dimensions, corresponding to the observables that vary in a slow and smooth way, and assuming that the rest of the dynamics is approximated by "noise"[1]. Since the resulting system is stochastic, one is interested in computing the average values of observables, over the probability distribution of the projected system.

However, the computation of the probability distribution may be hard, and so one seeks to find a way of producing "artificial" trajectories, in the projected space, designed in such a way that their probability distribution is the desired one. So doing, the problem reduces to the computation of the time-averaged values of "slow" observables. For the rest of this section, please make reference to Fig. 1.

[1] The noise can be so small, compared to the macroscopic observables that it can be neglected. In such cases, one has a deterministic, low-dimensional dynamical system, like for instance the usual models for rigid bodies, planets, etc.

Newton Laws

The success of Newton in describing the motion of one body, subjected to a static field force (say: gravitational motion of one planet, oscillation of a body attached to a spring, motion of the pendulum, etc.) clearly proved the validity of his approach, and also the validity of using simple models for dealing with natural phenomena. Indeed, the representation of a body as a point mass, the idea of massless springs and strings, the concept of force fields are all mathematical idealizations of reality.

The natural generalization of this approach is carried out in the XVIII century by Lagrange, Hamilton and many others. It brings to the mathematisation of mechanics and

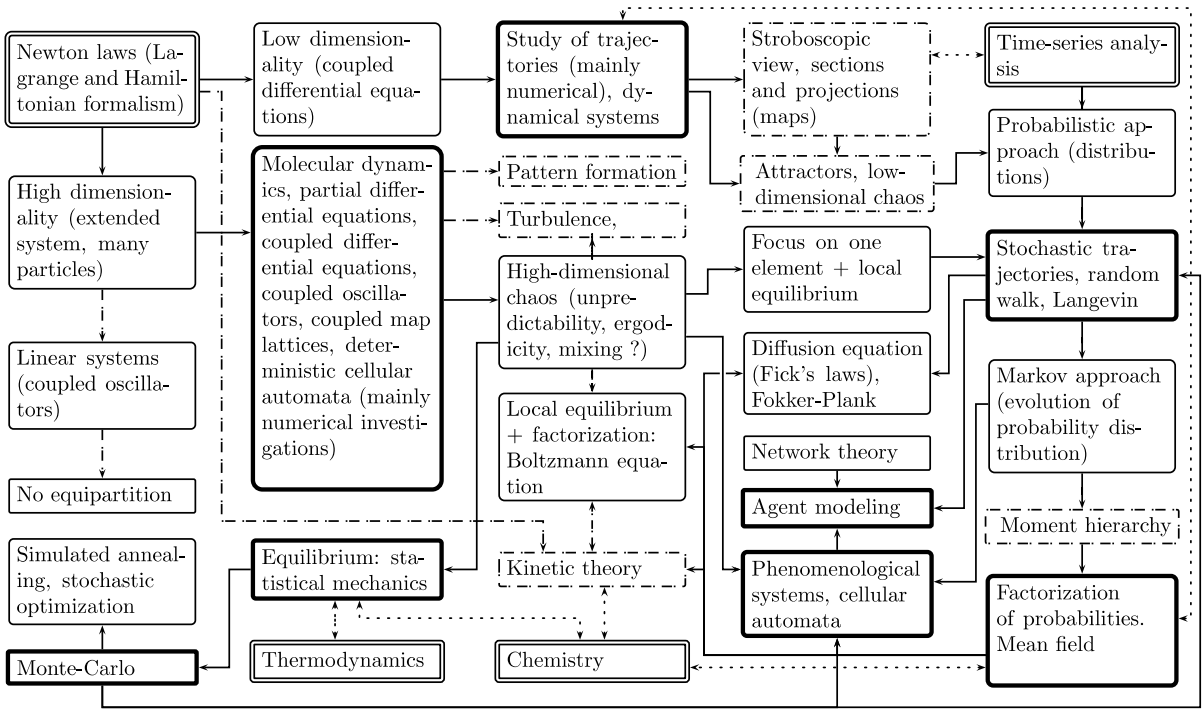

Interaction Based Computing in Physics, Figure 1
Graphical illustration of the logic path followed in this introduction. Boxes with double frame are "starting points", dashed boxes are topic that are not covered by discussion, boxes with darker frames mark topics that are investigated more in details

the derivation of *rational mechanics*. The resulting "standard" (or historical) way of modeling physical systems is that of using differential equations, i. e., a continuous description of time, space and other dynamical quantities.

From an abstract point of view, one is faced with two different options: either concentrate on systems described by a few equations (low-dimensional systems), or try to describe systems formed by many components.

Low Dimensionality

Historically, the most important problem of Newton's times was that of three bodies interacting via gravitational attraction (the Sun, the Earth and the Moon). By approximating planets with point masses, one gets a small number of coupled differential equations. This reduction of dimensionality is an example of a "scale separation": the variables that describe the motion of the planets vary slowly and smoothly in time. Other variables, for instance those that describe the oscillations of a molecule on the surface of a planet, can be approximated by a noise term so small that can be safely neglected. This approximation can also be seen as a "mean field" approach, for which one assumes that variables behave not too differently from their average. Using these techniques, one can develop models of many systems that result in the same mathematical scheme: a few coupled equations. The resulting equations may clearly have a structure quite different from that resulting from Newtonian dynamics (technically, Hamiltonian systems).

However, the reduction of the number of variables does not guarantee the simplicity of the resulting model. The problem of three gravitational bodies cannot be split into smaller pieces, and the computation of an accurate trajectory requires a computer. In general, a *nonlinear* system in a space with three or more dimensions is chaotic. This implies that there it may "react" to a small perturbation of parameters or initial conditions with large variations of its trajectory. This *sensibility* to variation implies the impossibility of predicting its behavior for long times, unless one is content with a probabilistic description.

High Dimensionality

In many cases, the "projection" operation results in a system still composed by many parts. For instance, models of nonequilibrium fluids neglect to consider the movement of the individual molecules, but still one has to deal with the values of the pressure, density and velocity in all points. In these cases one is left with a high-dimensional problem. Assuming that the "noise" of the projected dimensions can be neglected, one can either write down a large number of coupled equation (e. g., in modeling the vibration of a crystal), or use a continuous approach and describe the system using partial differential equations (e. g., the model of a fluid).

Linear Systems

In general, high and low-dimensional approaches can be systematically developed (with paper and pencil) only in the linear approximation. Let us illustrate this point for the case of coupled differential equation: if the system is linear one can write the equations using matrices and vectors. One can in principle find a (linear) transformation of variables that make the system diagonal, i. e., that reduces the problem to a set of *uncoupled* equations. At this point, one is left with (many) one-dimensional independent problems. Clearly, there are mathematical difficulties, but the path is clear. A similar approach (for instance using Fourier transforms) can be used also for dealing with partial differential equations.

The variables that result from such operations are called normal modes, because they behave independently one from the other (i. e., they correspond to orthogonal or normal directions in the state space). For instance, the linear model of a vibrating string (with fixed ends) predicts that any pattern can be described as a superposition of "modes", which are the standing oscillations with zero, one, two, … nodes (the harmonics).

However, linear systems behave in a somewhat strange way, from the point of view of thermal physics. Let us consider for instance the system composed by two *uncoupled* oscillators. It is clear that if we excite one oscillator with any amount of energy, it will remain confined to that subsystem. With normal modes, the effect is the same: any amount of energy communicated to a normal mode remains confined to that mode, if the system is completely linear. In other words, the system never forgets its initial condition.

On the contrary, the long-time behavior of "normal" systems does not depend strongly on the initial conditions. One example is the *timbre*, or "sound color" of an object. It is given by the simultaneous oscillations on many frequencies, but in general an object emits its "characteristic" sound regardless of how exactly is perturbed. This would not be true for linear systems.

Since the *distribution* of energy to all available "modes" is one of the assumptions of equilibrium statistical mechanics, which allows us to "understand" the usual behavior of matter, we arrived at an unpleasant situation: linear systems, which are so "easy" to be studied, cannot be used to ground statistical physics on mechanics.

Molecular Dynamics

Nowadays, we have computers at our disposal, and therefore we can simulate systems composed by many parts with complex interactions. One is generally interested in computing *macroscopic* quantities. These are defined as *averages* of some function of the microscopic variables (positions, velocities, accelerations, etc.) of the system. A measurement on a system implies an average, over a finite interval of time and over a large number of elementary components (say: atoms, molecules, etc.) of some quantity that depends on the microscopic state of that portion of the body.

Chaos and Probabilities

It was realized by Poincaré (with paper and pencil) and Lorenz (with one of the very first computers) that also very few (three) coupled differential equations with nonlinear interactions may give origin to complex (chaotic) behavior. In a chaotic system, a small uncertainty amplifies exponentially in time, making forecasting difficult. However, chaos may also be simple: the equations describing the trajectory of dice are almost surely chaotic, but in this case the chaos is so strong that the tiniest perturbation or uncertainty in the initial conditions will cause in a very small amount of time a complete variation of the trajectory. Our experience says that the process is well approximated by a probabilistic description. Therefore, chaos is one possible way of introducing probability in dynamics.

Chaotic behavior may be obtained in simpler models, called maps, that evolve in discrete time steps. As May showed [40], a simple map with a quadratic non-linearity (logistic map) may be chaotic. One can also model a system using coupled maps instead of a system of coupled differential equations [34]. And indeed, when a continuous system is simulated on a computer, it is always represented as an array of coupled maps.

Discretization

There is a progression of discretization from partial differential equations, coupled differential equations, coupled map lattices: from systems that are continuous in space, time and in the dynamical variables to systems that are discrete in time and space, and continuous only in the dynamical variables. The further logic step is that of studying completely discrete systems, called *cellular automata*.

Cellular automata show a wide variety of different phenomenologies. They can be considered mathematical tools, or used to model reality. In many cases, the resulting phenomenological models follows probabilistic rules, but it is also possible to use cellular automata as "building blocks". For instance, is possible to simulate the behavior of a hydrodynamical system by means of a completely discrete model, called cellular automata lattice gas [26,29].

Statistics

The investigation of chaotic extended systems proceed generally using a statistical approach. The idea is the following: any system contains a certain degree of nonlinearity, that couples otherwise independent normal modes. Therefore, (one hopes that) the initial condition is not too important for the asymptotic regime. If moreover one assumes that the motion is so chaotic that any trajectory spans the available space in a "characteristic" way (again, not depending on the initial conditions), we can use statistics to derive the "characteristic" probability distribution: the probability of finding the system in a given portion of the available space is proportional to the the time that the system spends in that region. See also the paragraph on equilibrium.

Random Walks

Another approach is that of focusing on a small part of a system, for instance a single particle. The rest of the system is approximated by "noise". This method was applied, for instance, by Einstein in the development of the simplest theory of Brownian motion, the random walk [31]. In random walks, each steps of the target particle is independent on previous steps, due to collisions with the rest of particles. Collisions, moreover, are supposed to be uncorrelated. A more sophisticated approximation consists in keeping some aspects of motion, for instance the influence of inertia or of external forces, still approximating the rest of the world by noise (which may contain a certain degree of correlation). This is known as the Langevin approach, which includes the random walk as the simplest case. Langevin equations are stochastic differential equations.

The essence of this method relies in the assumption that the behavior of the various parts of the systems is uncorrelated. This assumption is vital also for other types of approximations, that will be illustrated in the following. Notice that in the statistical mechanics approach, this assumption is not required.

In the Langevin formulation, by averaging over many "independent" realizations of the process (which in general is not the same of averaging over many particles that "simultaneously" move, due for instance to excluded volumes) one obtains the evolution equation of the probability of finding a particle in a given portion of space. This

is the Kolmogorov integro-differential equation, that in many case can be simplified, giving a differential (Fokker-Plank) equation. The diffusion equation is just the simplest case [27,56].

It is worth noticing that a similar formulation may be developed for quantum systems: the Feynman path-integral approach is essentially a Langevin formulation, and the Schroedinger equation is the corresponding Fokker-Plank equation.

Random walks and stochastic differential equations find many application in economics, mainly in stock market simulations. In this cases, one is not interested in the average behavior of the market, but rather in computing non-linear quantities over trajectories (time-series of good values, fluctuations, etc.).

Time-Series Data Analyzes

In practice, a model is never derived "ab initio", by projecting the dynamics of all the microscopic components onto a limited number of dimensions, but is constructed heuristically from observations of the behavior of a real system.

It is therefore crucial to investigate how observations are made, i. e., the analysis of a series of a time measurements. In particular, a good exercise is that of simulating a dynamical or stochastic system, analyze the resulting time-series data of a given observable, and see if one is able to reconstruct from it the relationships or the equations ruling the time evolution.

Let us consider the experimental study of a chaotic, low-dimensional system. The measurements on this system give a time series of values, that we assume discrete (which is actually the case considering experimental errors). Therefore, the output of our experiment is a series of symbols or numbers, a time-series. Let us assume that the system is stationary, i. e., that the sequence is statistically homogeneous in time. If the system is not extremely chaotic, symbols in the sequence are correlated, and one can derive the probability of observing single symbols, couples of symbols, triples of symbols and so on. There is a hierarchy in these probabilities, since the knowledge of the distribution of triples allows the computation of the distribution of couples, and so on.

It can be shown that the knowledge of the probability distribution of the infinite sequence is equivalent to the complete knowledge of the dynamics. However, this would correspond to performing an infinite number of experiments, for all possible initial conditions.

The usual investigation scheme assumes that correlations vanishes beyond a certain distances, which is equivalent to assume that the probability of observing sequences longer than that distance factorize. Therefore, one tries to model the evolution of the system by a probabilistic dynamics of symbols Sect. "Probabilistic Cellular Automata". Time-series data analysis can therefore be considered as the main experimental motivation in developing probabilistic discrete models. This can be done heuristically comparing results with observations *a posteriori*, or trying to extract the rules directly from data, like in the Markov approach.

Markov Approximation

The Markov approach, either continuous or discrete, also assumes that the memory of the system vanishes after a certain time interval, i. e., that the correlations in time series decay exponentially. In discrete terms, one tries to describe the process under study as an automata, with given transition probabilities. The main problem is: given a sequence of symbols, what is the simplest automata (hidden Markov chains [48]) that can generate that sequence with maximum "predictability", i. e., with transition probabilities that are nearest to zero or one? Again, it is possible to derive a completely deterministic automata, but in general it has a number of nodes equivalent to the length of the time-series, so it is not generalizable and has no predictability (see also Sect. "Probabilistic Cellular Automata"). On the contrary, an automata with a very small number of nodes will have typically intermediate transition probabilities, so predictability is again low (essentially equivalent to random extraction). Therefore, the good model is the result of an optimization problem, that can be studied using, for instance, Monte-Carlo techniques.

Mean-Field

Finally, from the probabilities one can compute averages of observables, fluctuations and other quantities called "moments" of the distribution. Actually, the knowledge of all moments is equivalent to the knowledge of the whole distribution. Therefore, another approach is that of relating moments at different times or different locations, truncating the recurrences at a certain level. The roughest approximation is that of truncating the relations at the level of averages, i. e., the mean field approach. It appears so natural that is is often used without realizing the implications of the approximations. For instances, chemical equations are essentially mean-field approximations of a complex phenomena.

Boltzmann Equation

Another similar approach is that of dividing an extended system into zones, and assume that the behavior of the sys-

tem in each zone is well described by a probability distribution. By disregarding correlations with other zones, one obtains the Boltzmann equation, with which many transport phenomena may be studied well beyond elementary kinetic theory. The Boltzmann equation can also be obtained from the truncation of a hierarchy of equations (BBGKY hierarchy) relating multi-particle probability distributions. Therefore, the Boltzmann equations is similar in spirit to a mean-field analysis.

Equilibrium

One of the biggest success of the stochastic approach is of course *equilibrium statistical mechanics*. The main ingredient of this approach is that of minimum information, which, in other words, corresponds to the assumption: *what is not known is not harmful*. By supposing that at equilibrium the probability distribution of the systems maximizes the information entropy (corresponding to a minimum of information on the system) one is capable of deriving the probability distribution itself and therefore the expected values of observables (ensemble averages, see Sect. "Ising"). In this way, using an explicit model, one is capable to compute the value of the parameters that appear in thermodynamics. Were it possible to show that the maximum entropy state is actually the state originated by the dynamics of a mechanical (or quantum) system, one could ground thermodynamics on mechanics. This is a long-investigated subject, dating back to Boltzmann, which is however not yet clarified. The main drawback in the derivations is about ergodicity. Roughly speaking, a system is called *ergodic* if the infinite-time average of an observable over a trajectory coincides with its average over a snapshot of infinitely many replicas. For a system with fixed energy and no other conserved quantities, a sufficient condition is that a generic trajectory passes "near" all points of the accessible phase-space. However, most systems whose behavior is "experimentally" well approximated by statistical mechanics are not ergodic. Moreover, another ingredient, the capability of *forgetting* quickly the information about initial conditions appears to be required, otherwise trajectories are strongly correlated and averages over different trajectories cannot be "mixed" together. This capability is strongly connected to the chaoticity or *unpredictability* of extended systems, but unfortunately these ingredients makes analytic approaches quite hard.

An alternative approach, due to Jaynes [33], is much more pragmatic. In essence, it says: let design a model with the ingredients that one thinks are important, and assume that what is not in the model does not affect its statistical properties. Compute the distribution that maximizes the entropy with the given constraints. Then, compare the results (averages of observables) with experiments (possibly, numerical ones). If they agree, one has captured the essence of the problem, otherwise one have to include some other ingredient and repeat the procedure. Clearly, this approach is much more general than the "dynamical" one, not considering trajectory or making assumptions about the energy, which is simply seen as a constraint. But physicists would be much more satisfied by a "microscopic" derivation of statistical physics.

In spite of this lack of strong basis, the statistical mechanics approach is quite powerful, especially for systems that can be reduced to the case of *almost* independent elements. In this situation, the system (the *partition function*) factorizes, and many computations may be performed by hand. Notice however that this behavior is in strong contrast to that of truly linear systems: the "almost" attribute indicates that actually the elements interact, and therefore share the same "temperature".

Monte-Carlo

The Monte-Carlo technique was invented for computing, with the aid of a computer, thermal averages of observables of physical systems at equilibrium. Since then, this term is often used to denote the technique of computing the average values of observables of a stochastic system by computing the time-average values over "artificial" trajectories.

In equilibrium statistical physics, one is faced by the problem of computing averages of observables over the probability distribution of the system, and since the phase space is very high-dimensional, this is in general not an easy task: one cannot simply draw *random* configurations, because in general they are so different from those "typical" of the given value of the temperature, that their statistical weight is marginal. And one does not want to revert to the original, still-more-highly-dimensional dynamical system, which typically requires powerful computers just to be followed for tiny time intervals.

First of all, one can divide (*separate*) the model into almost independent subsystems, that, due to the small residual interactions (the "almost" independency), are at the same temperature. In the typical example of a gas, the velocity components appear into the formula of energy as additive terms, i. e., they do not interact with themselves or with other variables. Therefore, they can be studied separately giving the *Maxwell distribution* of velocities. The positions of molecules, however, are linked by the potential energy (except in the case of an ideal gas), and so the

hard part of the computation is that of generating configurations. Secondly, statistical mechanics guarantees that the asymptotic probability distribution does not depend on the details of dynamics. Therefore, one is free to look for the fastest dynamics still compatible with constraints. The Monte-Carlo computation is just a set of recipes for generating such trajectories. In many problems, this approach allows to reduce the (computational) complexity of the problem of several orders of magnitude, allowing to generate "artificial" trajectories that span statistically significant configuration with small computational effort. In parallel with the generation of the trajectory, one can compute the value of several observables, and perform statistical analysis on them, in particular the computation of time averages and fluctuations.

By extension, the same terms "Monte-Carlo" is used for the technique of generating sequences of states (trajectories) given the transition probabilities, and computing averages of observables on trajectories, instead of over the probability distribution.

Stochastic Optimization

One of the most interesting applications of Monte-Carlo simulations concerns stochastic optimization via *simulated annealing*. The idea is that of exploiting an analogy between the status of a system (and its energy) and the coding of a particular procedure with corresponding cost function. The goal is that of finding the *best* solution, i. e., the global minimum of the energy given the constraints. "Easy" systems have a smooth energy landscape, shaped like a funnel, so that usual techniques like that of always choosing the displacements that locally lowers the energy (gradient descent) are successful. However, when the energy landscape is corrugated, there are many local minima where algorithms like gradient descent tend to get trapped. Methods from statistical mechanics (Monte-Carlo), on the contrary, are targeted to generating trajectories that quickly explore the "relevant" parts of the state space, i. e., those that correspond to the largest contributions to the probability distribution, that depends on the temperature, an "external" or control parameter. If the temperature is high, the probability distribution is broad and the generated trajectory does not "see" the minima of energy that are below the temperature, i. e., it can jump over and off the local minima.

By lowering the temperature, the probability distribution of system obeying statistical mechanics concentrates around minima of energy, and the Monte-Carlo trajectory does the same. The energy (or cost) function of not-extremely-complex problems is shaped in such a way that the global optimum belongs to a broad valley, so that this lowering of the temperature increases the probability of finding it.

Therefore, a sufficiently slow annealing should furnish the desired global minimum. Moreover, it is possible to "convert" constraints into energy terms. For many problems, it is difficult to generate configurations that satisfy the constraints. Let us think for instance to the problem of generating a school timetable, keeping into consideration that lessons should not last more than three consecutive hours, that a teacher or students cannot stay in two classes at the same time, that a teacher is not available on Monday, another prefers the first hours, etc. One can formulate a Monte-carlo algorithm that generates free configurations, but then weights them with a factor that depends on how many constraints are violated. At high temperature, constraints do not forbids the exploration of the state space, and therefore to try "creative" solutions. At low temperature, constraints become important. At zero temperature, the solutions with lower energy are those that satisfy all constraints, if possible, or at least the largest part of them.

In recent years, physics have dealt with extremely complex problems (e. g., spin glasses [22,42]), in which the energy landscape is extremely rough. Special techniques, based on a non-monotonous "walk" on temperature have been developed (simulated tempering [39]).

Critical Phenomena

One of the most investigated topics of statistical mechanics concerns phase transitions. This is a fascinating subject: in the vicinity of a continuous phase transitions correlation lengths diverge, and the system behave collectively, in a way which is largely independent of the details of the model. This universal behavior allows the use of extremely simplified models, that therefore can be massively simulated.

The *philosophy* of statistical mechanics may be *exported* to nonequilibrium systems: systems with absorbing states (that correspond to infinitely negative energy), driven systems (live living ones), strongly frustrated systems (that actually never reach equilibrium), etc. In general, one defines these systems in terms of transition probabilities, not in term of energy. Therefore, one cannot invoke a maximum entropy principle, and the results are less general.

However, many systems exhibit behavior reminiscent of equilibrium systems, and the same language can be used: phase transitions, correlations, susceptibilities,…

These characteristics, common to many different models, are sometimes referred as *emergent features*.

One of the most famous problems in this field is percolation: the formations of giant clusters in systems described by a local stochastic aggregation dynamics. This "basic" model has been used to describe an incredibly large range of phenomena [53].

Equilibrium and nonequilibrium phase transitions occur for a well-defined value of a control parameter. However, in nature one often observes phenomena whose characteristics resemble that of a system near a phase transition, a critical dynamics, without any *fine-tuned* parameter. For such system the term self-organized criticality has been coined [9], and they are the subject of active researches.

Networks

A recent "extension" of statistical physics is the theory of networks. Networks in physics are often regular, like the lattice of a crystal, or only slightly randomized. Characteristics of these networks are the fixed (or slightly dispersed around the mean) number of connections per node, the high probability of having connected neighbors (number of "triangles"), the large time needed to cross the network. The opposite of a regular network is a random graph, which, for the same number of connections, exhibits low number of triangles and short crossing time. The statistical and dynamical properties of systems whose connection are regular or random are generally quite different.

Watts and Strogatz [58] argued that *social networks* are never completely regular. They showed that the simple random rewiring of a small number of links in a regular network may induce the small world effect: local properties, like the number of triangles, are not affected, but large-distance ones, like the crossing time, quickly became similar to that of random graphs. Also the statistical and dynamical properties of models defined over a rewired networks are generally similar to those correlated to random graphs.

After this finding, many social networks were studied, and they revealed a yet different structure: instead of having a well-defined connectivity, many of them present a few highly-connected "hubs", and a lot of poorly-connected "leafs". The distribution of connectivity is often shaped as a power-law (or similar [43]), without a well-defined mean (scale-free networks [1]). Many of phenomenological models are presently re-examined in order to investigate their behavior over such networks. Moreover, scale-free networks cannot be "laid down", they need to be "grown" following a procedure, similar in this to fractals. It is natural therefore to include such a procedure in the model, so that not only they evolve "over" the networks, but also evolve "the" network [13].

Agents

Many of the described tools are used in the so-called agent-based modeling. The idea is that of exploiting the powerful capabilities of present computers to simulate directly a large number of *agents* that interact among them. Traditional investigations of *complex systems*, like crowds, flocks, traffic, urban models, and so on, have been performed using *homogeneous* representation: partial differential equations (i. e., mean-field), Markov equations, cellular automata, etc. In such an approach, it is supposed that each agent type is present in many identical copies, and therefore they are simulated as "macrovariables" (cellular automata), or aggregated like independent random walkers in the diffusion equation. But live elements (cells, organisms) do not behave in such a way: they are often individually unique, carry information about their own past history, and so on. With computers, we are now in the position of simulating large assemblies of individuals, possibly geographically located, like for instance humans in an urban simulation.

One of the advantages of such approach is that of offering the possibility of measuring quantities that are inaccessible to field researchers, and also to play with different scenarios. The disadvantages are the proliferation of parameters, that are often beyond experimental confirmation.

Artificial Worlds

A somewhat alternative approach to that of "traditional" computational physics is that of studying an artificial model, build with little or no direct connection with reality, trying to include only those aspect that are considered relevant. The goal is to be able to find the *simplest* system still able to exhibit the relevant features of the phenomena under investigation. The resulting models, though not directly applicable to the interpretation of experiments, may serve as interpretative tools in many different situations. For instance, the ising model was developed in the context of the modelization of magnetic systems, but is has been applied to opinion formation, social simulations, etc.

Ising

The Ising (better: Lenz-Ising) model is probably one of the most known models in statistical physics. Its history [44] is particularly illuminating in this context, even if it took

place well before the advent of computers in physics. It is also a model for which the Monte-Carlo and simulated annealing techniques are readily applied.

Let us first illustrate schematically the model. I shall present the traditional version, with the terminology that arises from the physics of magnetic systems. However, it is an interesting exercise to reformulate it in the context, for instance, of opinion formation. Let simply replace "spin up/down" with "opinion A/B", "magnetization" with "average opinion", "coupling" with "exchange of ideas", "external magnetic field" with "propaganda", and so on.

The Ising model is defined on a lattice, that can be in one, two or more dimensions, or even on a disordered graph. We shall locate a cell with an index $\boldsymbol{i}$, corresponding to the set of spatial coordinates for a regular lattice or a label for a graph. The dynamical variable $x_{\boldsymbol{i}}$ for each cell is just a binary digit, traditionally named "spin" that takes the values ± 1. We shall indicate the whole configuration as $\boldsymbol{x}$. Therefore, a lattice with N cells has 2^N distinct configurations. Each configuration $\boldsymbol{x}$ has an associated energy $E(\boldsymbol{x}) = -\sum_{\boldsymbol{i}}(H + h_{\boldsymbol{i}})x_{\boldsymbol{i}}$, where H represents the external magnetic field and $h_{\boldsymbol{i}}$ is a local magnetic field, generated by neighboring spins, $h_{\boldsymbol{i}} = \sum_{\boldsymbol{j}} J_{\boldsymbol{ij}} x_{\boldsymbol{i}}$. The coupling $J_{\boldsymbol{ij}}$ for the original Lenz-Ising model is one if $\boldsymbol{i}$ and $\boldsymbol{j}$ are nearest neighbors, and zero otherwise.

The maximum-entropy principle [33] gives the probability distribution

$$P(\boldsymbol{x}) = \frac{1}{Z} \exp\left(-\frac{E(\boldsymbol{x})}{T}\right)$$

from which averages can be computed. The parameter T is the temperature, and Z, the "partition function" is the normalization factor of the distribution.

The quantity $E(\boldsymbol{x})$ can be though as a "landscape", with low-energy configurations corresponding to valley and high-energy ones to peaks. The distribution $P(\boldsymbol{x})$ can be interpreted as the density of a gas, each "particle" corresponding to a possible realization (a replica) of the system. This *gas* concentrates in the valleys for low temperatures, and diffuses if the temperature is increased. The temperature is related to the average level of the gas.

In the absence of the local field ($J = 0$), the energy is minimized if each $x_{\boldsymbol{i}}$ is *aligned* (same sign) with H. This ordering is counteracted by thermal noise. In this case it is quite easy to obtain the average magnetization per spin (order parameter)

$$\langle x \rangle = \tanh\left(\frac{H}{T}\right),$$

which is a plausible behavior for a paramagnet. A ferromagnet however present hysteresis, i. e., it may maintain for long times (metastability) a pre-existing magnetization opposed to the external magnetic field.

With coupling turned on ($J > 0$), it may happen that the local field is strong enough to "resist" H, i. e., a compact patch of spins oriented against H may be stable, even if the energy could be lowered by flipping all them, because the flip of a single spin would rise the energy (actually, this flip may happen, but is statistically re-absorbed in short times). The fate of the patch is governed by boundaries. A spin on a boundary of a patch feels a weaker local field, since some of its neighbors are oriented in the opposite. Straight boundaries in two or more dimensions separate spins that "know" the phase they belong to, since most of their neighbors are in that phase, the spins on the edges may flip more freely. Stripes that span the whole lattice are rather stable objects, and may *resist* an opposite external field since spins that occasionally flip are surrounded by spins belonging to the opposite phase, and therefore feel a strong local field that pushes them towards the phase opposed to the external field.

In one dimensions with finite-range coupling, a single spin flip is able to create a "stripe" (perpendicularly to the the lattice dimension), and therefore can destabilize the ordered phase. This is the main reason for the absence of phase transitions in one dimension, unless the coupling extends on very large distances or some coupling is infinite (see the part on directed percolation, Sect. "Probabilistic Cellular Automata").

This model was proposed in the early 1920s by Lenz to Ising for his Ph D dissertation as a simple model of a ferromagnet. Ising studied it in one dimension, found that it shows no phase transition and concluded (erroneously) that the same happened in higher dimensions. Most of contemporaries rejected the model since it was not based on Heisenberg's quantum mechanical model of ferromagnetic interactions. It was only in the forties that it started gaining popularity as a model of cooperative phenomena, a prototype of order-disorder transitions. Finally, in 1940, Onsager [46] provided the exact solution of the two-dimensional Lenz-Ising model in zero external field. It was the first (and for many year the only) model exhibiting a non-trivial second-order transition whose behavior could be exactly computed.

Second-order transition have interested physicists for almost all the past century. In the vicinity of such transitions, the elements (say, spins) of the system are correlated up to very large distances. For instance, in the Lenz-Ising model (with coupling and more than one dimension), the high-temperature phase is disordered, and the low-temperature phase is almost completely ordered. In both these phases the connected two-points correlation

function

$$G_c(r) = \langle x_i x_{i+r} \rangle - \langle x_i \rangle^2$$

decreases exponentially, $G_c(r) \simeq \exp(-r/\xi)$, with $r = |\boldsymbol{r}|$. The length ξ is a measure of the typical size of patch of spins pointing in the same direction.

Near the critical temperature T_c, the correlation length ξ diverges like $\xi(T - T_c) \simeq (T - T_c)^{-\nu}$ ($\nu = 1$ for $d = 2$ and $\nu \simeq 0.627$ for $d = 3$, where d is the dimensionality of the system). In such case the correlation function is described by a power law

$$G_c(r) \simeq \frac{1}{r^{d-2+\eta}} ,$$

with $\eta = 1/4$ for $d = 1$ and $\eta \simeq 0.024$ for $d = 3$. This phase transition is an example of a *critical phenomenon* [12], ν and η are examples of *critical exponents*.

The divergence of the correlation length indicates that there is no characteristic scale (ξ) and therefore fluctuations of all sizes appear. In this case, the details of the interactions are not so important, so that many different models behave in the same way, for what concerns for instance the critical exponents. Therefore, models can be grouped into *universality classes*, whose details are essentially given by "robust" characteristics like the dimensionality of space and of the order parameter, the symmetries, etc.

The power-law behavior of the correlation function also indicates that if we perform a *rescaling* of the system, it would appear the same or, conversely, that one is unable to estimate the *distance* of a pattern by comparing the "typical size" of particulars. This scale invariance if typical of many natural phenomena, from clouds (whose height and size are hard to be estimated), trees and other plant elements, lungs, brain, etc.

Many examples of power laws and collective behavior can be found in natural sciences [52]. Differently from what happens in the Lenz-Ising model, in these cases there is no parameter (like the temperature) that has to be fine-tuned, so one speaks of *self-organized criticality* [9].

Since the Lenz-Ising model is so simple, exhibits a critical phase and can be exactly solved (in some cases), it has become the playground for a variety of modifications and applications to various fields. Clearly, most of modifications do no allow analytical treatment and have to be investigated numerically. The Monte-Carlo method allows to add a temporal dimension to a statistical model [35], i. e., to transform stochastic integrals into averages over fictitious trajectories. Needless to say, the Ising-Lenz model is the standard test for every Monte-Carlo beginner, and most of techniques for accelerating the convergence of averages has been developed with this model in mind [55].

Near a second-order phase transitions, a physical system exhibits *critical slowing down*, i. e., it reacts to external perturbations with an extremely slow dynamics, with a convergence time that increases with the system size. One can extend the definition of the correlation function including the time dimension: in the critical phase also the temporal correlation length diverges (as a power law). This happens also for the Lenz-Ising model using the the Monte-Carlo dynamics, unless very special techniques are used [55]. Therefore, the dynamical version of the Lenz-Ising model can be used also to investigate relaxational dynamics and how this is influenced by the characteristics of the energy landscape. In particular, if the coupling J_{ij} changes sign randomly for every couple of sites (or the field H has random sign for each site), the energy landscape becomes extremely rugged. When spins flip in order to align to the local field, they may invert the field felt by neighboring ones. This *frustration* effect is believed to be the basic mechanism of the large viscosity and memory effects of *glassy* substances [22,42].

The rough energy landscape of glassy systems is also challenging for optimization methods, like *simulated annealing* [36] and its "improved" cousin, *simulated tempering* [39]. Again, the Lenz-Ising model is the natural playground for these algorithm.

The dynamical Lenz-Ising model can be formulated such that each spin is updated in parallel [10] (with the precaution of dividing cells into sublattices, in order to keep the neighborhood of each cell fixed during updates). In this way, it is essentially a probabilistic cellular automata, as illustrated in Sect. "Probabilistic Cellular Automata".

Cellular Automata

In the same period in which traditional computation was developed, in the early fifties, John Von Neumann was interested in the logic basis of life and in particular in self-reproduction, and since the analysis of a self-reproduction automata following the rules of real physics was too difficult, he designed a playground (a cellular automaton) with just enough "physical rules" in order to make its analysis possible. It was however just a theoretical exercise, the automaton was so huge that up to now it has not yet completely implemented [57].

The idea of cellular automata is quite simple: take a lattice (or a graph) and put on each cell an automaton (all automata are equal). Each automaton exhibit its "state" (which is one out of a small number) and is programmed

to react (change state) according to the state of neighbors and its present one (the *evolution rule*). All automata update their state synchronously.

Cellular automata share many similarities with the parallel version of the Lenz-Ising model. Differently from that, their dynamics is not derived from an energy, but is defined in terms of the transition rules. These rules may be deterministic or probabilistic. In the first case (illustrated in this section), cellular automata are *fully discrete, extended dynamical systems*. Probabilistic cellular automata are illustrated in Sect. "Probabilistic Cellular Automata". The temporal evolution of deterministic cellular automata can be computed *exactly* (regardless of any approximation) on a standard computer.

Let us illustrate the simplest case, *elementary cellular automata*, in Wolfram's jargon [61]. The lattice is here one dimensional, so to identify an automaton it is sufficient to give one coordinate, say i, with $i = 1, \ldots, N$. The state of the automaton on cell i at time t is represented by a single variable, $x_i(t)$ that can take only two values, "dead/live", or "inactive/active" or 0/1. The time is also discrete, so $t = 1, 2, \ldots$.

The parallel evolution of each automaton is given by the rule

$$x_i(t+1) = f(x_{i-1}(t), x_i(t), x_{i+1}(t)) \,.$$

Since $x_i = 0, 1$, there are only eight possible combinations of the triple $\{x_{i-1}(t), x_i(t), x_{i+1}(t)\}$, from $\{0, 0, 0\}$ to $\{1, 1, 1\}$. For each of them, $f(x_{i-1}(t), x_i(t), x_{i+1}(t))$ is either zero or one, so the function f can be simply coded as a vector of eight bits, each position labeled by a different configuration of inputs. Therefore, there are only $2^8 = 256$ different elementary cellular automata, that have been studied carefully (see for instance Ref. [61]).

In spite of their simplicity, elementary cellular automata exhibit a large variety of behaviors. In the language of dynamical systems they can be "visually" classified [61] as fixed points (class-1), limit cycles (class-2) and "chaotic" oscillations (class-3). A fourth class, "complex" CA, is present with larger neighborhood or in higher dimensions. A classical example is the *Game of Life* [11]. This two-dimensional cellular automaton is based on a simple rule. A cell may be either empty (0) or "alive" (1). A living cell survives if, among its 8 nearest neighbors, there are two or three alive cells, otherwise it dies and disappears. Generation is implemented through a rule for empty cells: they may become "alive" if sourrounded by exactly three living cells. In spite of the simplicity of the rule, this automaton generates complex and long-living patterns, some of them illustrated in Fig. 2.

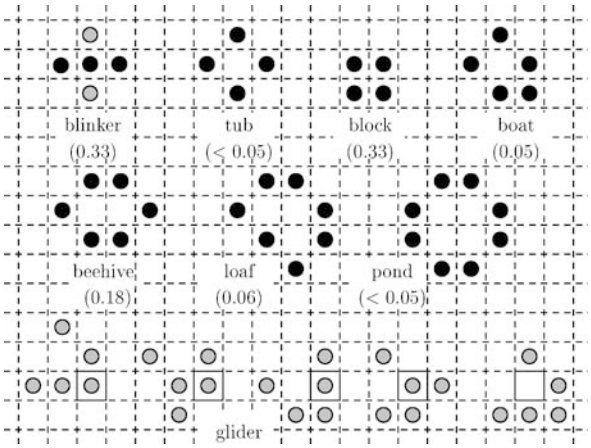

Interaction Based Computing in Physics, Figure 2
Some of the most common "animals" in the game of life, with the probability of encountering them in an asymptotic configuration [4]

Complex CA have large transients, during which interesting structures may emerge. They finally relax into class-1 automata. It has been conjectured that they are able of computation, i. e., that one can "design" an universal computer using these CA as building blocks, as has been proved to be possible with the Game of Life. Another hypothesis, again confirmed by the Game of Life, is that these automata are "near the edge" of self-organizing complexity. One can slightly "randomize" the Game of Life, allowing sometimes an exception to the rule. Let us introduce a parameter p, that measures this randomness, with the assumption that $p = 0$ is the true "life". Well, it was shown [45] that the resulting model exhibits a second-order phase transition for a value of p very near zero.

Deterministic cellular automata have been investigated as prototypes of discrete dynamical systems, in particular for what concerns the definition of chaos. Visually, one is tempted to use this word also to denote the irregular behavior of "class 3" rules. However, the usual definition of chaos involves the sensitivity to an infinitesimally small perturbation: following the time dynamics of two initially close configurations one can observe an amplification of their distance. If the initial distance (δ_0) is infinitesimal, then the distance grows exponentially for some time ($\delta(t) \simeq \delta_0 \exp(\lambda t)$), after which it tends to saturate, since the trajectories are generally bounded inside an attractor, or due to the dimensions of the accessible space. The exponent λ depends on the initial configuration, and if this behavior is observed for different portions of the trajectory, it fluctuates: a trajectory spends some time in regions of high chaoticity, after which may pass through "quiet" zones. If

one "renormalizes" periodically this distance, considering one system as the "master" and the other as a measuring device, one can accumulate good statistics, and define a Lyapunov exponent λ, that gives indications about the chaoticity of the trajectory, through a limiting procedure.

The accuracy of computation poses some problems. Since in a computer numbers are always approximate, one cannot follow "one" trajectory. The small approximations accumulates exponentially, and the computer time series actually *jumps* among neighboring trajectories. Since the Lyapunov exponent is generally not so sensible to a change of precision in computation, one can assume that the chaotic regions are rather compact and uniform, so that in general one associates a Lyapunov exponent to a system, not to an individual trajectory. Nevertheless, this definition cannot apply to completely discrete systems like cellular automata.

In any case, chaoticity is related to unpredictability. As first observed by Lorenz, and following the definition of Lyapunov exponent, the precision of an observation over a chaotic system is related to the average time for which predictions are possible. Like in weather forecasts, in order to increase the time span of a prediction one has to increase the precision of the initial measurement. In extended system, this also implies to extend the measurements over a larger area. One can also consider a "synchronization" approach. Take two replicas of a systems and let them evolve starting from different initial configurations. With a frequency and a strength that depends on a parameter q, one of these replica is "pushed" towards the other one, so to reduce their distance. Suppose that $q = 0$ is the case of no push and $q = 1$ is the case of extremal push, for which the two systems synchronize in a very short time. There should be a critical value q_c that separates these two behaviors (actually, the scenario may be more complex, with many phases [4]). In the vicinity of q_c the distance between the two replicas is small, and the distance δ grows exponentially. The critical value q_c is such that the exponential growth compensates the shrinking factor, and is therefore related to the Lyapunov exponent λ.

Finite-size cellular automata always follow periodic trajectories. Let us consider for instance a Boolean automata, of N cells. The number of possible different states is 2^N and due to determinism, once that a state has been visited twice the automata has entered a limit cycle (or a fixed point). One may have limit cycles with large basins of transient configurations (configurations that do not belong to the cycle). Many scenarios are possible. The set of different configurations may be divided in many basins, of small size (small transient) and small period, like in class 1 and 2 automata. Or one may have large basins, with long transients that lead to short cycles, like in class-4 automata. Finally, one may have one or very few large basins, with long cycles that include most of configurations belonging to the basin (small transients). This is the case of class-3 automata. For them, the typical period of a limit cycle grows exponentially, like the total number of configurations, with the system size, so that for moderately large system it is almost impossible to observe a whole cycle in a finite time. Another common characteristic of class-3 automata is that the configurations quickly decorrelates (in the sense of the correlation function) along a trajectory. If one takes into consideration as starting points two configurations that are the same but for a local difference, one observes that this difference amplifies and diffuses in class-3 automata, shrinks or remains limited in class-1 and class-2, and have an erratic transient behavior in class-4, followed by the fate of class-1 and 2. Therefore, if one considers the possibility of not knowing exactly the initial configuration of an automata, unpredictability grows with time also for such discrete systems. Actually, also (partially) continuous systems like coupled maps may exhibit this kind of behavior [4,17,19,47]. Along this line, it is possible to define an equivalent of the Lyapunov exponents for CA [7]. The synchronization procedure can be applied also to cellular automata, and it correlates well with the Lyapunov exponents [5].

An "industrial" application of cellular automata is their use for modeling gases. The *hydrodynamical equations*, like the Navier-Stokes ones, simply reflect the conservation of mass, momentum and energy (i. e., rotational, translational and time invariance) for the microscopic collision rules among particles. Since the modeling of a gas via molecular dynamics is rather cumbersome, some years ago it was proposed [26,29] to simplify drastically the microscopic dynamics using particles that may travel only along certain directions with some discrete velocities and jumping in discrete time only among nodes of a lattice. Indeed, a cellular automaton. It has been shown that their macroscopic dynamics is described by usual hydrodynamics laws (with some odd features related to the underlining lattice and finiteness of velocities) [51,60].

The hope was that these Lattice Gas Cellular Automata (LGCA) could be simulated so efficiently in hardware to make possible the investigation of turbulence, or, in other words, that they could constitute the Ising model of hydrodynamics. While they are indeed useful to investigate certain properties of gases (for instance, chemical reactions [37], or the relationship between chaoticity and equilibrium [6]), they resulted too noisy and too viscous to be useful for the investigation of turbulence. Viscosity is related to the transport of momentum in a direction perpen-

dicular to the momentum itself. If the collision rule does not "spread" quickly the particles, the viscosity is large. In LGCA there are many limitations to collisions, so that in order to lower viscosity one has to consider averages over large patches, thus lowering the efficiency of the method.

However, LGCA inspired a very interesting approximation. Let us consider a large assembly of replicas of the same system, each one starting from a different initial configuration, all compatible with the same macroscopic initial conditions. The macroscopic behavior after a certain time would be the *average* over the status of all these replicas. If one assumes a form of local equilibrium, i. e., applies the mean-field approximation for a given site, one may try to obtain the dynamics of the *average* distribution of particles, which in principle is the same of "exchanging" particles that happen to stay on the same node among replicas.

It is possible to express the dynamics of the average distribution in a simple form: it is the Lattice Boltzmann Equation (LBE) [18,54,60]. The method retains many properties of LGCA like the possibility of considering irregular and varying boundaries, and may be simulated in a very efficient way with parallel machines [54]. Differently from LGCA, there are numerical stability problems to be overcome.

Probabilistic Cellular Automata

In deterministic automata, given a local configuration, the future state of a cell is univocally determined. But let us consider the case of measuring experimentally some pattern and trying to analyze it in terms of cellular automata. In time-series analysis, it is common to perform averages over spatial patches and temporal intervals and to discretize the resulting value. For instance, this is the natural result of using a camera to record the temporal evolution of an extended system, for instance the turbulent and laminar regions of a fluid. The resulting pattern symbolically represents the dynamics of the original system, and if it is possible to extract a "rule" out of this pattern, it would be extremely interesting for the construction of a model. In general, however, one observes that sometimes a local configuration is followed by a symbol, and sometime the same local configuration is followed by another one. One should conclude that the neighborhood (the local configuration) does not univocally determines the following symbol.

One can extend the "range" of the rule, adding more neighbors farther in space and time [48]. By doing so, the "conflicts" generally reduce, but at the price of increasing the complexity of the rule. At the extremum, one could have an automaton with infinite "memory" in time and space, that perfectly reproduces the observed patterns but with almost none predictive power, since it is extremely unlucky that the same *huge* local configuration is encountered again.

So, one may prefer to limit the neighborhood to some finite extension, and accept that the rule sometimes "outputs" a symbol and sometimes another one. One defines a local transition probability $\tau(x_i(t+1)|\boldsymbol{X}_i(t))$ of obtaining a certain symbol x_i at time $t+1$ given a local configuration $\boldsymbol{X}_i$ at time t. Deterministic cellular automata correspond to the case $\tau = 0, 1$. The parallel version of the Lenz-Ising model can be re-interpreted as a probabilistic cellular automaton.

Starting from the local transition probabilities, one can build up the transition probability $T(\boldsymbol{x}|\boldsymbol{y})$ of obtaining a configuration $\boldsymbol{x}$ given a configuration $\boldsymbol{y}$ ($T(x|y)$. $T(\boldsymbol{x}|\boldsymbol{y})$ is given by the product of the local transition probabilities τ [2]. One can read the configurations $\boldsymbol{x}$ and $\boldsymbol{y}$ as indexes, so that T can be considered as a matrix. The normalization of probability corresponds to the constraint $\sum_{\boldsymbol{x}} T(\boldsymbol{x}|\boldsymbol{y}) = 1, \forall \boldsymbol{y}$.

Denoting with $P(\boldsymbol{x}, t)$ the probability of observing a given configuration $\boldsymbol{x}$ at time t, and with $\boldsymbol{P}(t)$ the whole distribution at time t, we have for the evolution of the distribution

$$\boldsymbol{P}(t+1) = T\boldsymbol{P}(t) ,$$

with the usual rules for the product of matrices and vectors. Therefore, the transition matrix T defines a Markov process, and the asymptotic state of the system is given by the eigenvalues of T. The largest eigenvalue is always 1, due to the normalization of the probability distribution, and the corresponding eigenvector is the asymptotic distribution. The theory of Markov processes says that if T is finite and irreducible, i. e., it cannot be rewritten (renumbering rows and columns) as blocks of noninteracting subspaces, like

$$T = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} ,$$

then the second eigenvalue is strictly less than one and the asymptotic state is unique. In this case the second eigenvalue determines the convergence time to the asymptotic state. For finite systems, often the matrix T is irreducible. However, in the limit of infinite size, the largest eigenvalue may become degenerate, and therefore there are more than one asymptotic state. This is the equivalent of a phase transition for Markov processes.

For the parallel Lenz-Ising model, the elements of the matrix T are given by the product of local transition rules of the Monte-Carlo dynamics. They depend on the choice

of the algorithm, but essentially have the form of exponentials of the difference in energy, divided by the temperature. Although a definitive proof is still missing, it is plausible that matrices with all elements different from zero correspond to some equilibrium model, whose transition rules can be derived from an energy function [28].

Since probabilistic cellular automata are defined in terms of the transition probabilities, one is free to investigate models that go beyond equilibrium. For instance, if some transition probability takes the value zero or one, in the language of equilibrium system this would correspond to some coupling (like the J of the Lenz-Ising model) that become infinite. This case is not so uncommon in modeling. The inverse of a transition probability corresponds to the average waiting time for the transition to occur in a continuous-time model (one may think to chemical reactions). Some transitions may have a waiting time so long with respect to the observation interval, to be practically irreversible. Therefore, probabilistic cellular automata (alongside other approaches like for instance annihilating random walks) allow the exploration of out-of-equilibrium phenomena.

One such phenomena is directed percolation, i. e., a percolation process with a special direction (time) along which links can only be crossed one-way [14]. Let think for instance to the spreading of an infection in a one-dimensional lattice, with immediate recovery (SIS model). An ill individual can infect one or both of his two neighbors but returns to the susceptible state after one step. The paths of infection (see Fig. 3) can wander in the space directions, but are directed in the time directions.

The parallel version of a directed percolation process can be mapped onto probabilistic cellular automata. The simplest case, in one spatial dimension and with just two neighbors, is called the Domany-Kinzel model [21]. It is even more general than the usual directed percolation, allowing "non-linear" interactions among sites in the neighborhood (e. g., two wet sites may have less probability of percolating than one alone).

These processes are interesting because there is an *absorbing state* [3,32], which is the dry state for the wetting phenomenon and the healthy state for the spreading of epidemics. Once the system has entered this absorbing state, it cannot exit, since the spontaneous appearing of a wet site or of an ill individual is forbidden. For finite systems, the theory of Markov chains says that the system will "encounter", sooner or later, this state, that therefore corresponds to the unique asymptotic state. For infinitely extended systems, a phase transition can occur, for which wet sites percolate for all "times", the epidemics become endemic, and so on.

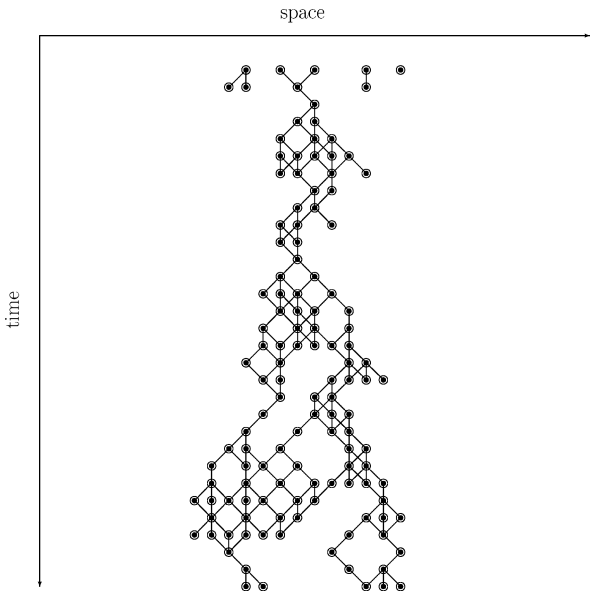

Interaction Based Computing in Physics, Figure 3
An example of a directed percolation cluster

Again these are examples of *critical phenomena*, with exponents different from the equilibrium case.

Cellular Automata and Agent-Based Simulations

Cellular automata can be useful in modeling phenomena that can be described in lattice terms. However, many phenomena requires "moving particles". Examples may be chemical reactions, ecological simulations, social models. When particles are required to obey hydrodynamics constraints, i. e., to collide conserving mass, momentum and energy, one can use lattice gas cellular automata or the lattice Boltzmann equation. However, in general one is interested in modeling just a macroscopic scale, assuming that what happens at lower level is just "noise". According with the complexity (and "intelligence") assigned to particles, one can develop models based on the concept of walkers, that move more or less randomly. From the simulation point of view, walkers are not very different from the graphs succinctly described above. In this case the identifier $\boldsymbol{i}$ is just a label, that allows to access walker's data, among which there are the coordinates of the walker (that may be continuous), its status and so on.

In order to let walker interact, one is interested in finding efficiently all walkers that are nearer than a given distance from the one under investigation. This is the same problem one is faced with when developing codes for molecular dynamics simulations [49]: scanning all walkers

in order to compute their mutual distance is a procedure that grows quadratically with the number of walkers. One "trick" is that of dividing the space in cells, i. e., to define an associated lattice. Each cell contains a list of all walkers that are locate inside it. In this way, one can directly access all walkers that are in the same or neighboring cells of the one under investigation.

Moreover, one can exploit the presence of the lattice to implement on it a "cellular automaton", that may interact with walkers, in order to simulate the evolution of "fields". Just for example, the simulation of a herd of herbivores that move according to the exhaustion of the grass, and the parallel growing of the vegetation can be modeled associating the "grass" to a cellular automaton, and the herbivores to walkers. This simulation scheme is quite flexible, allowing to implement random and deterministic displacements of moving objects or agents, continuous or discrete evolution of "cellular objects" and their interactions. Many simulation tools and games are based on this scheme [50,59,62], and they generally allow the contemporary visualization of a graphical representation of the state of the system. They are valuable didactic tool, and may be used to "experiment" with these artificial worlds. As often the case, the flexibility is paid in terms of efficiency and speed of simulation.

Discussions and Conclusions

Complex systems, like for instance human societies, cannot be simulated starting from "physical laws". This is sometimes considered a weakness of the whole idea of studying such systems from a quantitative point of view. We have tried to show that actually even the "hardest" discipline, physics, always deals with models, that have finally to be simulated on computers making various assumptions and approximations. Theoretical physics is accustomed since a long time to "extreme simplifications" of models, hoping to enucleate the "fundamental ingredients" of a complex behavior. This approach have proved to be quite rewarding for our understanding of nature.

In recent years, physics have been seen studying many fields not traditionally associated to physics: molecular biology, ecology, evolution theory, neurosciences, psychology, sociology, linguistics, and so on. Actually, the word "physics" may refer either to the classical subjects of study (mainly atomic and subatomic phenomena, structure of matter, cosmological and astronomical topics), or to the "spirit" of the investigation, that may apply to almost any discipline. This spirit is essentially that of building simplified quantitative models, composed by many elements, and study them with theoretical instruments (most of times, applying some form of mean-field treatment) and with computer simulations.

This approach has been fruitful in chemistry and molecular biology and nowadays many physical journals have sections devoted to "multidisciplinary studies". The interesting thing is that not only have physicists brought some "mathematics" into fields that are traditionally more "qualitative" (which often corresponds to "linear" modeling, plus noise), but physicists have also discovered many interesting questions to be investigated, and new models to be studied. One example is given by the current investigations about the structure of social networks, that were "discovered" by physicists in the nontraditional field of social studies.

Another contribution of physicists to this "new way" of performing investigations, is the use of networked computers. Since a long time, physicists have used computers for performing computations, storing data and diffusing information using the Internet. Actually, the concept of what is now the World Wide Web was born at CERN, as a method for sharing information among laboratories [15]. The high-energy physics experiments require a lot of simulations and data processing, and physicists (among others) developed protocols to distribute this load on a grid of networked computers. Nowadays, an European project aims to "open" grid computing to other sciences [24].

It is expected that this combination of quantitative modeling and grid computing will stimulate innovative studies in many fields. Here is a small sparse list of possible candidates:

- Theory of evolution, especially for what concerns evolutionary medicine.
- Social epidemiology, coevolution of diseases and human populations, interplay between sociology and epidemics.
- Molecular biology and drug design, again driven by medical applications.
- Psychology and neural sciences, it is expected that the "black box" of traditional psychology and psychiatry will be replaced by explicit models based on brain studies.
- Industrial and material design.
- Earth sciences, especially meteorology, vulcanology, seismology.
- Archaeology: simulation of ancient societies, reconstruction of historical and pre-historical climates.

Nowadays, the term *cellular automata* has enlarged its meaning, including any system whose elements do not move (in opposition to "agent-based modeling"). There-

fore, we now have cellular automata on non-regular lattices, non-homogeneous, with probabilistic dynamics (see Sect. "Probabilistic Cellular Automata"), etc. [23]. They are therefore considered more as a "philosophy" of modeling rather than a single tool. In some sense, cellular automata (and agent-based) modeling is opposed to the spirit of describing a phenomena using differential equations (or partial differential equations). One of the reasons is that the language of "automata" and "agents" is simpler and requires less training than that of differential equations. Another reason is that at the very end, any "reasonable" problem has to be investigated using computers, and while the implementation using discrete elements is straightforward (even if careful planning may speed-up dramatically the simulation), the computation of partially differential equations is an art in itself.

However, the final success of this approach is related to the availability of high-quality experimental data that allow to discriminate among the almost infinite number of models that can be built.

Bibliography

Primary Literature

1. Albert R, Barabasi AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
2. Bagnoli F (2000) Cellular automata. In: Bagnoli F, Ruffo S (eds) Dynamical Modeling in Biotechnologies. World Scientific, Singapore, p 1
3. Bagnoli F, Boccara N, Rechtman R (2001) Nature of phase transitions in a probabilistic cellular automaton with two absorbing states. Phys Rev E 63(4):046 116
4. Bagnoli F, Cecconi F (2001) Synchronization of non-chaotic dynamical systems. Phys Lett A 282(1–2):9–17
5. Bagnoli F, Rechtman R (1999) Synchronization and maximum Lyapunov exponents of cellular automata. Phys Rev E 59(2):R1307–R1310
6. Bagnoli F, Rechtman R (2007) Entropy and chaos in a discrete hydrodynamical system. arXiv:cond-mat/0702074
7. Bagnoli F, Rechtman R, Ruffo S (1992) Damage spreading and Lyapunov exponents in cellular automata. Phys Lett A 172:34
8. Bagnoli F, Rechtman R, Ruffo S (1994) Some facts of life. Physica A 171:249
9. Bak P, Tang C, Weisenfeld K (1987) Self-organizing criticality: An explanation of 1/*f* noise. Phys Rev A 38:364–374
10. Barkema GT, MacFarland T (1994) Parallel simulation of the ising model. Phys Rev E 50(2):1623–1628
11. Berlekamp E, Conway J, Guy R (1982) What is Life?, vol 2: Games in Particular. Academic Press, London, chap 25
12. Binney J, Dowrick N, Fisher A, Newman MEJ (1993) The Theory of Critical Phenomena. Oxford Science Publications. Clarendon Press, Oxford
13. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex networks: Structure and dynamics. Phys Rep 424(4–5):175–308
14. Broadbent S, Hammersley J (1957) Percolation processes I. Crystals and mazes. Proc Camb Philos Soc 53:629–641
15. Cailliau R (1995) A short history of the web. http://www.netvalley.com/archives/mirrors/robert_cailliau_speech.htm Accessed 25 May 2008
16. Car R, Parrinello M (1985) Unified approach for molecular dynamics and density-functional theory. Phys Rev Lett 55(22):2471–2474
17. Cecconi F, Livi R, Politi A (1998) Fuzzy transition region in a one-dimensional coupled-stable-map lattice. Phys Rev E 57(3):2703–2712
18. Chopard B, Luthi P, Masselot A, Dupuis A (2002) Cellular automata and lattice Boltzmann techniques: An approach to model and simulate complex systems. Adv Complex Syst 5(2):103–246
19. Crutchfield J, Kaneko K (1988) Are attractors relevant to turbulence? Phys Rev Lett 60(26):2715–2718
20. Daxois T, Peyrard M, Ruffo S (2005) The Fermi–Pasta–Ulam 'numerical experiment': History and pedagogical perspectives. Eur J Phys 26:S3–S11
21. Domany E, Kinzel W (1984) Equivalence of cellular automata to ising models and directed percolation. Phys Rev Lett 53(4):311–314
22. Dotsenko V (1994) An Introduction to the Theory of Spin Glasses and Neural Networks. World Scientific, Singapore
23. El Yacouby S, Chopard B, Bandini S (eds) (2006) Cellular Automata. Lecture Notes in Computer Science, vol 4173. Springer, Berlin
24. Enabling grids for e-science. http://www.eu-egee.org/ Accessed 25 May 2008
25. Fermi E, Pasta J, Ulam S (1955) Los alamos report la-1940. In: Segré E (ed) Collected papers of Enrico Fermi. University of Cicago Press, Chicago
26. Frisch U, Hasslacher B, Pomeau Y (1986) Lattice-gas automata for the navier-stokes equation. Phys Rev Lett 56(14):1505–1508
27. Gardiner CW (1994) Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences. Springer series in synergetics, vol 13. Springer, Berlin
28. Georges A, le Doussal P (1989) From equilibrium spin models to probabilistic cellular automata. J Stat Phys 54(3–4):1011–1064
29. Hardy J, Pomeau Y, de Pazzis O (1973) Time evolution of a two-dimensional classical lattice system. Phys Rev Lett 31(5):276–279
30. Harlow H, Metropolis N (1983) Computing & computers - weapons simulation leads to the computer era. Los Alamos Science 4(7):132
31. Haw M (2005) Einstein's random walk. Physics World 18:19–22
32. Hinrichsen H (1997) Stochastic lattice models with several absorbing states. Phys Rev E 55(1):219–226
33. Jaynes E (1957) Information theory and statistical mechanics. Phys Rev 106(4):620–630
34. Kaneko K (1985) Spatiotemporal intermittency in coupled map lattices. Progr Theor Phys 74(5):1033–1044
35. Kawasaki K (1972) Kinetics of Ising model. In: Domb CM, Green MS (eds) Phase Transitions and Critical Phenomena, vol 2. Academic Press, New York, p 443
36. Kirkpatrick S, Gelatt Jr CG, Vecchi MP (1983) Optimization by simulated annealing. Science 220:671–680
37. Lawniczak A, Dab D, Kapral R, Boon JP (1991) Reactive lattice gas automata. Phys D 47(1–2):132–158

38. Lorenz E (1996) History of chaos: http://library.thinkquest.org/3120/old_htdocs.1/text/fraz1.txt Accessed 25 May 2008
39. Marinari E, Parisi G (1992) Simulated tempering: A new Monte Carlo scheme. Europhys Lett 19:451–458
40. May R (1976) Simple mathematical models with very complicated dynamics. Nature 261:459–467
41. Metropolis N, Hewlett J, Rota GC (eds) (1980) A History of Computing in the Twentieth Century. Academic Press, New York
42. Mezard M, Parisi G, Virasoro MA (1987) Spin Glass Theory and Beyond. World Scientific Lecture Notes in Physics, vol 9. World Scientific, Singapore
43. Newman ME (2005) Power laws, pareto distributions and zipf's law. Contemp Phys 46:323–351
44. Niss M (2005) History of the Lenz–Ising model 1920–1950: From ferromagnetic to cooperative phenomena. Arch Hist Exact Sci 59(3):267–318
45. Nordfalk J, Alstrøm P (1996) Phase transitions near the "game of life". Phys Rev E 54(2):R1025–R1028
46. Onsager L (1944) Crystal statistics. i. a two-dimensional model with an order-disorder transition. Phys Rev 65:117–149
47. Politi A, Livi R, Oppo GL, Kapral R (1993) Unpredictable behaviour of stable systems. Europhys Lett 22(8):571–576
48. Rabiner L (1989) A tutorial on hidden markov models and selected applications in speech recognition. Proc IEEE 77(2):257–286
49. Rapaport DC (2004) The Art of Molecular Dynamics Simulation. Cambridge University Press, Cambridge
50. Repast – recursive porus agent simulation toolkit. http://repast.sourceforge.net/ Accessed 25 May 2008
51. Rothman DH, Zaleski S (2004) Lattice-Gas Cellular Automata. Monographs and Texts in Statistical Physics. Collection Alea-Saclay, Paris
52. Sornette D (2006) Critical Phenomena in Natural Sciences. Springer Series in Synergetics. Springer, Berlin
53. Stauffer D, Aharony A (1994) Introduction To Percolation Theory. Taylor Francis, London
54. Succi S (2001) The Lattice Boltzmann Equation For Fluid Dynamics and Beyond. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford
55. Swendsen R, Wang JS (1987) Nonuniversal critical dynamics in Monte Carlo simulations. Phys Rev Lett 58(2):86–88
56. van Kampen NG (1992) Stochastic Processes in Physics and Chemistry. North-Holland, Amsterdam
57. Von Neumann universal constructor. http://en.wikipedia.org/wiki/Von_Neumann_Universal_Constructor Accessed 25 May 2008
58. Watts D, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440–441
59. Wilensky U (1999) Netlogo. http://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer-Based Modeling. Northwestern University, Evanston; Accessed 25 May 2008
60. Wolf-Gladrow D (2004) Lattice-Gas Cellular Automata and Lattice Boltzmann Models: An Introduction. Lecture Notes in Mathematics, vol 1725. Springer, Berlin
61. Wolfram S (1983) Statistical mechanics of cellular automata. Rev Mod Phys 55:601–644
62. Wright W (1989) Simcity. http://simcitysocieties.ea.com/ Accessed 25 May 2008

Books and Reviews

Bungartz H-J, Mundani R-P, Frank AC (2005) Bubbles, jaws, moose tests, and more: The wonderful world of numerical simulation, Springer VideoMATH. Springer, Berlin (DVD)
Gould H, Tobochnik J, Christian W (2007) An Introduction to Computer Simulation Methods: Applications to Physical Systems. Addison-Wesley, New York
Landau RH (2005) A First Course in Scientific Computing: Symbolic, Graphic, and Numeric Modeling Using Maple, Java, Mathematica, and Fortran90. Princeton University Press, Princeton
Chopard B, Droz M (2005) Cellular automata modeling of physical systems. In: Collection Alea-Saclay: Monographs and Texts in Statistical Physics. Cambridge University Press, Cambridge
Boccara N (2004) Modeling complex systems. In: Graduate Texts in Contemporary Physics. Springer, Berlin
Deisboeck S, Kresh JY (eds) (2006) Complex systems science in biomedicine. In: Topics in Biomedical Engineering. Springer, New York
Resnick M (1994) Turtles, Termites, and Traffic Jams. Explorations in Massively Parallel Microworlds. In: Complex Adaptive Systems. MIT Press, Cambridge
Shalizi C, Cosma's Home Page http://www.cscs.umich.edu/~crshalizi/
Open Source Physics http://www.opensourcephysics.org/about/index.html

Intermittency and Localization

Gur Yaari[1,2], Dietrich Stauffer[2,3], Sorin Solomon[1,2]
[1] Institute for Scientific Interchange, Turin, Italy
[2] Racah Institute of Physics, Hebrew University, Jerusalem, Israel
[3] Institute for Theoretical Physics, Cologne University, Köln, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Logistic Systems: From Malthus Until Today
Minimal Extensions to the Classical Logistic System
Applying These Models to Real-Life Systems
Future Directions
Acknowledgments
Bibliography

Glossary

Auto-catalysis In systems that go through several reactions, the reaction is called **autocatalytic** if the reaction product is itself the catalyst for that reaction.

Exponential growth An autocatalytic reaction is usually described with a simple linear, first order differential equation. The solution for it is an exponential increasing/decreasing function.

Logistic growth If one adds a saturation term (of power two) to the linear first order differential equation which describes exponential growth, the resulting solution saturates instead of ever-lasting grow. The solution to this system is described with a **logistic curve** and the system is said to follow a **logistic growth**.

Reaction-diffusion systems Reaction-diffusion systems are mathematical models that describe how the concentration of one or more substances distributed in space changes under the influence of two processes: local reactions in which the substances are converted into each other, and diffusion which causes the substances to spread out in space.

Definition of the Subject

In this paper, we show how simple logistic growth that was studied intensively during the last 200 years in many domains of science could be extended in a rather simple way. The resulting extended model has, among other features, two very important ones: Intermittency and Localization. These features were observed repeatedly along the history of science in an enormous number of real-life systems in economics, sociology, biology, ecology and more. We suggest by this a unified theoretical umbrella that might serve in a surprising way many scientific disciplines who share similar observed patterns.

Introduction

A well known joke that many physicists like to tell during their talks in order to demonstrate the strength of simplifying the problem one has in hand is: "First, let us consider a spherical cow…". Although, no one really believes in spherical cows – the power of simplification is well accepted and appreciated by the physics community, or as Albert Einstein put it, very accurately: "Everything should be made as simple as possible, but not simpler".

There are many more such mantras like: "Keep it simple, stupid", "Kill your darlings" and "Less is more". As these lines of thought were adopted so strongly by physicist for so much time, the statement of P.W. Anderson that "More is different" made such a revolution in Science. In the paper that has this title, Anderson pushed the new scientific (inter-) discipline, now known as complexity. By introducing these new ideas, Anderson paved the way for many physicists carrying with them heavy weapons from traditional physics to start thinking and attacking many problems from a variety of scientific disciplines.

A lot of criticism about such physicists that try to cross the borders of their discipline is about *over-simplifying* real-life problems in order to be able to solve the resulting models with the tools they already have. Due to that, it is important to emphasize here that by working inside the framework of complexity one tries not to loose the minimal theoretical ingredients of the problem that are sufficient to produce the complex observed outcome. Rather than do this, one tries to study to the best of one's ability, the simplest *possible* model.

A common question that arises in the social sciences is: *Why are improbable things so frequent?* Fine-tuned irreducibly complex systems have generically a low probability to appear and highly integrated $\simeq$ arranged systems are usually artificial (often man-made) and untypical. Yet many complex systems are found lately to be self-organized. More precisely, the amount of non-generic, fine tuned and highly integrated systems is much larger in nature from what would be reasonably expected from generic stochastic estimations. It often happens that even though the range of parameters necessary for some non-trivial collective phenomenon to emerge is very narrow (or even an isolated single point out of an continuum infinite range), the phenomenon does actually take place in nature. This leads to collective objects whose properties are not explainable by the generic dynamics of their components. The explanation of the generic emergence of systems which are non-generic from the multi-agent point of view seems to be related to self-catalyzing dynamics.

As suggested by the examples above, the frequency with which we encounter non-generic situations in self-catalyzing systems is not so surprising. Consider a space of all possible systems obtainable from certain chemical and physical parts. Even if a macroscopic number of those systems are not auto-catalytic and only a very small number happen to be auto-catalytic after enough time, one of the auto-catalytic systems will eventually arise. Once this happens, the auto-catalytic system will start multiplying leading to a final (or far-future) situation in which those auto-catalytic – a priory very improbable systems – are over-represented compared with their natural probability of occurrence. Basically, this is how life spread all over Earth.

In this paper, we show how simple logistic growth that was studied intensively during the last 200 years in many domains of science could be extended in a rather simple way and with these extensions is capable to produce a collection of behaviors widely observed in an enormous number of real-life systems in economics, sociology, biology,

ecology and more. For other reviews in this direction we recommend on [4,18,38].

The paper will start with a historical overview of the use of logistic-like systems in science since its introduction by Malthus in 1798 until today. The next section will present a view of the minimal, though sufficient, extensions to the classical logistic system that are able to bring this theoretical framework, closer to reality, but auto-catalysis yet still solvable analytically in many regions of the parameter's space. Then, we will show some of the successes we had in applying this framework to real-life systems. We will finish the paper by a short fantasy trying to describe a dream about the possible usages of this powerful theoretical framework in the so called soft sciences in the future.

Logistic Systems: From Malthus Until Today

Auto-Catalysis

One of the key concepts underlying the emergence of complex macroscopic features is auto-catalysis. We therefore give at this point a provisory definition of it: auto-catalysis = self-perpetuation, $\simeq$ reproduction, $\simeq$ multiplication. As opposed to the usual stochastic systems in which the microscopic dynamics changes typically the individual microscopic quantities by additive steps (e. g. a molecule receiving or releasing a quantum of energy), the auto-catalytic microscopic dynamics involve multiplicative changes (e. g. the market worth of a company changes by a factor (index) after each elementary transaction). Such auto-catalytic microscopic rules are widespread in chemistry (under the name of auto-catalysis), biology (reproduction $\simeq$ multiplication, species perpetuation), social sciences (profit, returns, rate of growth).

The autocatalytic essence of the growth processes was formally expressed as early as 1798 by T.R. Malthus [24] who wrote a differential equation for describing the dynamics of a population of proliferating individuals:

$$\frac{\mathrm{d}W(t)}{\mathrm{d}t} = a \cdot W(t) \,. \tag{1}$$

The growth rate of the population W is proportional to W itself and parametrized by a relative growth (/proliferation) rate a. The Malthus equation can be reinterpreted to represent a very wide range of phenomena in various fields: behavior adoption in sociology, proliferation in biology, capital returns in economics, or proselytizing in politics. The (exponential) solution $\sim e^{(a\cdot t)}$ of this equation influenced much of the subsequent ideas in various fields and in particular it roused the first worries about the sustainability of growth. Malthus himself expressed great concern of the humanitarian catastrophe that unlimited population growth may lead to. However, Verhulst [39] introduced (in 1838) a nonlinear interaction term $-b \cdot W^2$ (that may represent (confrontation over) limited resources in biology, competition in economics, limited constituency in politics and finite population in sociology)

$$\frac{\mathrm{d}W(t)}{\mathrm{d}t} = a \cdot W(t) - b \cdot W^2(t) \,. \tag{2}$$

By including this term, rather than increasing indefinitely, the solution saturates at a constant asymptotic value $W \longrightarrow \frac{a}{b}$. For the following two centuries, this logistic dynamics was considered by the leading scientists as a crucial element in various fields from biology (Volterra [40]) to the everyday world of politics and economics (Lord May [26]).

Real-Life Examples

The A(utocatalysis)-Bomb The first and the most dramatic example of the macroscopic explosive power of the multi-agent auto-catalytic systems is the nuclear (atom) bomb. The simple microscopic interaction underlying it is that the U235 nucleus, when hit by a neutron splits into a few energetic fragments including neutrons:

$$n + U \longrightarrow n + n + \text{etc} \,. \tag{3}$$

On the basis of (autocatalysis equation 1) even without knowing what is a neutron or a U235 nucleus, it is clear that a macroscopic chain reaction may develop: if there are other U235 nuclei in the neighborhood, the neutrons resulting from the first (autocatalysis equation 1) may hit some of them and produce similar new reactions. Those reactions will produce more neutrons that will hit more U235 that will produce more neutrons.

The result will be a chain (or rather branching tree) of reactions in which the neutrons resulting from one generation of fission events induce a new generation of fission events by hitting new U235 nuclei. This chain reaction will go on until eventually, the entire available U235 population (of typically some 10^{26} nuclei) is exhausted and their corresponding energy is emitted: the atomic explosion. The crucial feature in the equation above, which we call auto-catalysis, is that by inputting one neutron n in the reaction one obtains two (or more) neutrons ($n + n$). The theoretical possibility of iterating it and have an exponentially increasing macroscopic number of reactions was explained in a letter from Einstein to President Roosevelt. But only the later attack on Pearl Harbor lead to the initiation of the Manhattan project and the eventual construction of the A-bomb.

It is not by chance that the basic multi-agent method (the Monte Carlo simulation algorithm used until this very day in physics applications) was invented by people (Metropolis, Rosenbluth, Rosenbluth, Teller, Teller) involved in nuclear weapons research: the multi-agent method is the best fit method to compute realistically the macroscopic effects originating in microscopic interactions!

The B-Bomb: Autocatalysis and Localization in Immunology In no field is the auto-catalysis and localization more critical than in the emergence of living organisms functions out of the elementary interactions of cells and enzymes. From the very beginning of an embryo development the problem is how to create a controlled chain reaction such that each cell (starting with the initial egg) divides into similar cells, yet spatio-temporal structures (systems and organs) emerge. Let us consider the immune system as an example. The study of the Immune System for the past half century has succeeded in characterizing the key: cells, molecules, and genes. As always in complex systems, the mere knowledge of the microscopic world is not sufficient (and, on the other hand, some details of the micros are not necessary). Understanding comes from the identification of the relevant microscopic interactions and the construction of a multi-agent simulation with which to demonstrate in detail how the complex behavior of the immune system emerges. Indeed, the immune system provides an outstanding example of the emergence of unexpectedly complex behavior from a relatively limited number of simple components interacting according to known simple rules. By simulating their interactions in computer experiments that parallel real immunology experiments, one can check and validate the various mechanisms for the emergence of collective functions in the immune system. (e. g. recognition and destruction of various threatening antigens, the oscillations characteristic to rheumatoid arthritis, the localization of diabetes 1 to pancreatic islets etc). This would allow one to design further experiments, to predict their outcome and to control the mechanisms responsible for various auto-immune diseases and their treatment.

The Tulip Bomb The tulip mania is one of the most celebrated and dramatic economic bubbles in history. It involved the rise of the tulip bulb prices in 1637 to the level of average house prices. In the same year, after an increase by a factor of 20 within a month, the market collapsed back within the next 3 months. After loosing a fortune in a similar event (triggered by the South Sea Co.) in 1720 at the London Stock, Sir Isaac Newton was quoted to say, "I can calculate the motions of the heavenly bodies, but not the madness of people."

It might seem over-ambitious to try where Newton has failed but let us not forget that we are 300 years later, have big computers and had plenty of additional opportunities to contemplate the madness of people. One finds that global macroscopic (and often catastrophic) economic phenomena are generated by reasonably simple buy and sell microscopic operations. Much attention was paid lately to the sales dynamics of marketable products. Large amounts of data has been collected describing the propagation and extent of sales of new products, yet only lately one started to study the implications of the autocatalytic multi-agent reaction-diffusion formalism in describing the underlying microscopic process [12,37,41,42].

Extensions of the Classical Logistic System

One of the great early successes of the logistic dynamics was its application to the spread of malaria in humans and mosquitos. Sir Ronald Ross was awarded the Nobel prize [30] for this work. His ideas were expressed by Lotka [19] in terms of a coupled system of two equations generalizing (2):

$$\begin{aligned} &\mathrm{d}w_1(t)/\mathrm{d}t \\ &\qquad = a_1 \cdot w_1(t) + a_{12} \cdot w_2(t) - a_{112} \cdot w_1(t) \cdot w_2(t) \\ &\mathrm{d}w_2(t)/\mathrm{d}t \\ &\qquad = a_2 \cdot w_2(t) + a_{21} \cdot w_1(t) - a_{212} \cdot w_1(t) \cdot w_2(t) \,. \end{aligned} \tag{4}$$

Lotka has studied numerically this system in order to predict the ratios between the infected mosquitoes and the infected humans and the stability of the system. Vito Volterra advocated independently the use of equations in biology and social sciences [40] and re-deduced the logistic curve by reducing the Verhulst equation (2) to a variational principle that maximized a function that he named quantity of life [19]. Later, R.A. Fisher [11] extended of (2) to spatial distributed systems and expressed it in terms of partial differential equations:

$$\frac{\partial W(\vec{x},t)}{\partial t} = a \cdot W(\vec{x},t) - b \cdot W^2(\vec{x},t) + D \cdot \nabla^2 W(\vec{x},t). \tag{5}$$

He applied this to the spread of a mutant superior gene within a population and showed that as opposed to usual diffusion, the propagation consists of a sharp frontier (Fisher wave) that advances with constant speed (rather

then proportional to $\sqrt{t}$ as in usual diffusion). Following its formulation, the mathematical study of (5) was taken over by mathematicians [17] and lead eventually a large number of physics studies (especially on the anomalous and fractal properties of the interface [1,6,14,15]).

A crucial step was then taken by Eigen [9] and Eigen and Schuster [10] who generalized the Lotka system of two equations for two populations to an arbitrary number of equations $\simeq$ populations. They used the new system in the study of the Darwinian selection and evolution in prebiotic environments. More precisely, they considered quasi-species of auto-catalytic (self reproducing RNA sequences) molecules which can undergo mutations. Each sequence i self-replicates at a rate a_i and undergoes mutations to other sequences j at rates a_{ij} . The resulting system of equations is:

$$\mathrm{d}W_i(t)/\mathrm{d}t = a_i W_i(t) + \sum_{j=1}^{N} a_{ij} W_j(t) - \sum_{j=1}^{N} a_{ji} W_i(t) - b(\overrightarrow{W(t)}, t) W_i(t) \,. \quad (6)$$

The arbitrary function $b(\overrightarrow{W(t)}, t)$ represents generically the interaction with the environment (in the specific case of ref [10] the result of replenishing and stirring the container continuously).

The extension of the logistic framework to social sciences was strongly advanced by Elliot Montroll who based a book on social dynamics on the principle that almost all the social phenomena, except in their relatively brief abnormal times obey the logistic growth [27].

An analogy that was often exploited in economics was the ecology-market metaphor (e. g. [31]) which was advanced in parallel with the more mechanical physics analogies. The connection to the logistic framework was strengthened by the evolutionary economics metaphor (e. g. [8,16,28]). This lead to the extension of (6) to economics with the a_i's representing capital $\simeq$ GDP growth rates and the a_{ij}'s representing trade, social security, mutual help or other mechanisms of wealth transfer (e. g. taxes $\simeq$ subsidies). More recently [25] the logistic dynamics was applied to the dynamics of the equities i within a personal portfolio. Then $a_i(t)$'s are interpreted as the rate of growth of the equity i (at time t) and a_{ij} as the periodic redistribution of capital between the equities by the owner of the portfolio (in order to optimize it). Stochastic generalizations of the logistic $\simeq$ Lotka–Volterra equations were studied also in a large body of mathematical literature (e. g. [17]), and in order to get meaningful results out of the model, one has to introduce the noise in a proper way that will stand for its effect in real-life systems.

The Danger of Being Mean – Simple Examples

In this subsection we argue why microscopic (i. e. agent-based) studies are needed and why simplification in the style of mean field theories can be seriously wrong.

If we deal with a small biological population, then due to random accidents it may die out completely and irreversibly. For example, poachers may kill the two surviving males of a small elephant herd which is isolated from other elephants. It does not help the herd if one shows that *on average* there is enough food and space for two adult males, two adult females, and several calves. For larger populations usually such extreme fluctuations are less probable, and the time until it happens may increase exponentially with the population size.

Also, a hurricane may sink a ship even if averaged over the whole Atlantic Ocean the absolute value of the wind speed and wave height are moderate. In a marriage a husband is supposed to be faithful to his wife and should not average his efforts to become a father over 10^9 women; at least that's what wives often demand.

A less trivial example is demography. If you want to know how many people of retirement age are there for every thousand people of working age, usually one takes into account mortalities, birth rates, and migration. Let us assume, however, that one group of the population has a higher birth rate than the rest and that this difference is given on to the following generations, either genetically or culturally. Then, if everything else is the same, the group with the higher birth rate will finally dominate in the population, and using the *average* birth rate is not correct. (Of course, if the difference is small and we want to extrapolate over less than a century, then the average birth rate is still a good approximation.) One could remedy this error by simulating the two populations together; but then there could be other inherited traits which are demographically relevant, and thus with more and finer subdivisions we finally end up with agent-based demography [5], dealing with each individual.

This explains the conceptual gap between sciences: in conditions in which only a few exceptional individuals dominate, it is impossible to explain the behavior of the collective by plausible arguments about the typical or most probable individual. In fact, in the emergence of nuclei from nucleons, molecules from atoms, DNA from simple molecules, humans from apes, there are always the atypical cases (with accidentally exceptional advantageous properties) that carry the day. This effect seems to embrace the

emergence of complex collective objects in a very wide range of disciplines from bacteria to economic enterprizes, from emergence of life and Darwinism to globalization and sustainability.

In the following section we will bring examples [33] where these effects lead to strong localization, such that the mean-field approximations give qualitatively wrong results, like predicting extinction where survival is possible. The approximations do not become good if only the population is large.

In conclusion, generic logistic ideas hinted by (2) arose for the last century in an extremely wide-ranging set of applications. For each discipline, subject and system, the variables of the model had to be interpreted in terms of the empirical observables and adapted to the relevant range of parameters and initial conditions. Once the parameters are specified, the generic framework (6) (plus noise) becomes a well defined model for a specific system. Then, one can derive from it precise predictions and confront them with the data.

Minimal Extensions to the Classical Logistic System

Here, we show how by restricting the parameter's regime of the generic framework (6) (plus noise), one ends up with a model that has a very strong prediction's power [2,3,18,29,34,35,36].

Case 1: The Generalized Lotka–Volterra System

If one considers a uniform interaction in (6), the resulting equation can be written as:

$$\mathrm{d}W_i(t)/\mathrm{d}t = a_i \cdot W_i(t) + \alpha \cdot W(t) - b(\overrightarrow{W(t)}, t)) \cdot W_i(t) \quad (7)$$

where $W(t)$ is the average value of the W_i's; then it was shown [18,29,34,35,36] that:

- The system has a steady state for the normalized quantity $X_i(t) \equiv \frac{W_i(t)}{W(t)}$.
- The steady state distribution of the X_i could be calculated analytically and the resulting distribution has the following form: $P(X) = e^{-2\alpha/XD} \cdot X^{-2-2\alpha/D}$ where D is the variance of the distribution from which the growth rates (a_i's) is drawn out of.
- The fluctuations of the average ($W(t)$) have a wide distribution with a power-law tail that is closely connected with the value of the steady state distribution $(-2 - 2\alpha/D)$.

Obviously, as there is no explicit space in this system, one cannot see localization effects. However, intermittency is very clear here: The fluctuations of the average value are enormous but changing around a fixed value. The possible interpretation of such a model are very diverse:

Income Distribution: $W_i(t)$ can represent the annual income of each individual in the society – then, the $W\alpha$ term is connected to social benefits one gets from the being part of the society, such as social security, charity and minimum wage. The a_i's stand for the relative change between this year and the previous one. $b(\overrightarrow{W(t)}, t)) \cdot W_i(t)$ then represents the overall trend of the market – periods of depression and of external investments.

Stock Market: $W_i(t)$ can represent the value of a specific stock in the stock market (at the closing time of the market for example) – then, the $W\alpha$ term is connected to correlations among the different stocks in the market. The a_i's stand again for the relative change between the value today and the previous one. $b(\overrightarrow{W(t)}, t)) \cdot W_i(t)$ represents the overall trend of the market – periods of depression and of external investments.

Population Dynamics: $W_i(t)$ can represent the number of individuals from a specific species in animals or of a specific nation in humans – then, the $W\alpha$ term is connected to immigration or mutations connecting the different populations. $b(\overrightarrow{W(t)}, t)) \cdot W_i(t)$ represents the conditions for breeding.

There are many more possible interpretations but the point is clear. For each interpretation one can argue that the uniform choice of the interaction matrix is unrealistic – of course it might be true, but as it turns out lately, the power-law prediction is very robust and can stand many different choices of this matrix.

Case 2: The AB Model

The AB Model [20,22,23,32,33] is actually a reaction-diffusion system which has two types of agents: A and B. It is a discrete system, both in space and in the fields it describes (A and B in any spatial point are natural numbers, never negative) and as such needs to be described with a set of rate equations. Then the agents may go through the following possible processes with the corresponding rates:

Diffusion: at each time step, with probabilities $D_a/2d$ and $D_b/2d$, respectively, an A or B moves to a nearest neighbor site on a d-dimensional lattice.

Reaction: at each time step, with probabilities μ and $\lambda \cdot N_A$, a single B dies or gives birth to a new B, respectively, where N_A is the number of A's in the same location.

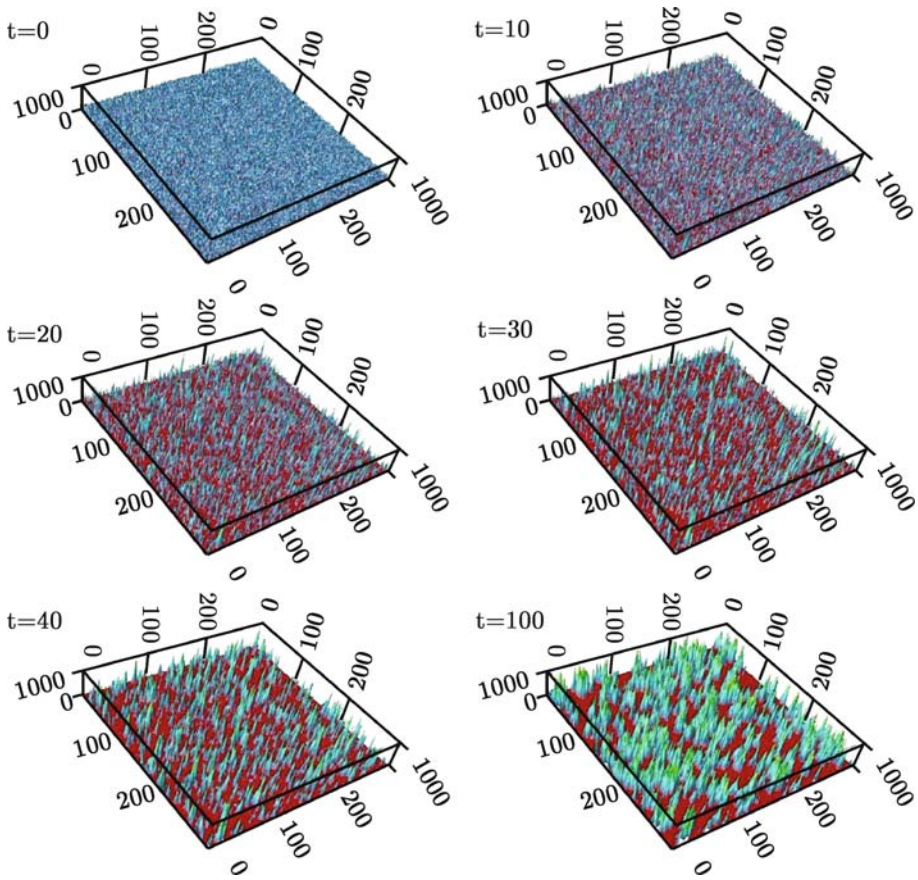

Intermittency and Localization, Figure 1
Snapshots of the AB Model in two dimensions [Log scale] in these 6 snapshots the time evolution of the AB Model is demonstrated: The two dimension lattice is set in $t = 0$ to be in a random distribution of the *B*s drawn from a Poisson distribution. The parameters are set in a way that if one looks at the mean-field approximation one will guess that the system needs to go to extinction in a relatively short time. However, due to the discreteness of the catalysts and the reactants, the *B*s adapt themselves to the rich (in food) areas and the famous island structure is formed

Naively, this system can be mapped into two partial differential equations:

$$\frac{\mathrm{d}B(x,t)}{\mathrm{d}t} = D_b \cdot \nabla^2 B(x,t) + (\lambda \cdot A(x,t) - \mu) \cdot B(x,t) \quad (8)$$

$$\frac{\mathrm{d}A(x,t)}{\mathrm{d}t} = D_a \cdot \nabla^2 A(x,t)\,. \quad (9)$$

It is tempting to say that we can solve Eq. (9) to get:

$$A(x,t) \longrightarrow n_A \quad (10)$$

in long times and then to plug it into Eq. (8) to say that depending on the parameter $m \equiv (n_A \cdot \lambda - \mu)$ the total number of *B*s will either increase exponentially (if $m > 0$) or decrease exponentially (if $m < 0$). It turns out that this mean-field treatment is totally wrong and as was shown in [33] in low enough dimensions ($\mathrm{d} \leq 2$) the *B*'s will asymptotically increase exponentially no matter what the rest of the parameters are! The intuitive explanation for this surprising result is that the *B*'s somehow adapt themselves to be localized around regions with good conditions (large number of A_i). One can see a typical snapshot of this system in Fig. 1. Another prediction [23] of this model is the intermittency of the total number of *B*'s even when one adds a saturation term similar to the second term in Eq. (2). Yet one more prediction of this model is the J-shape in the total number of *B*'s: i. e. initial decline followed by lasting exponential growth, Fig. 2.

Applying These Models to Real-Life Systems

As mentioned in Subsect. "Real-Life Examples", many real-life systems have characteristics that can be explained with the AB Model or the GLV:

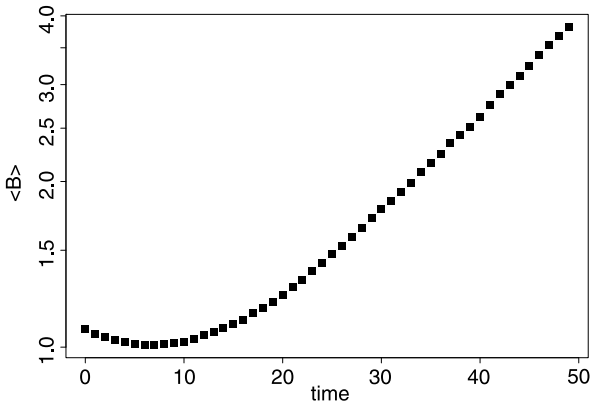

Intermittency and Localization, Figure 2

The time evolution of the average *B* number in the AB Model in two dimensions For the same parameters as Fig. 1 if one plots the average *B* number as a function of time, the above picture is revealed: a J-shape. It turns out that this shape is a ubiquitous feature for the GDP evolution of many countries that went through a major shock in their economies like the collapse of the Soviet bloc for example

In the immune system, it was shown in [21] that the body's *B* cells tend to grow in places in the genome space where they are needed (depending on the diseases existing in the system).

On the Internet, it was shown in [13] that one can use the theoretical understanding of the model in order to plan strategies that will improve the way the Internet works.

In global economics it was shown in [23] that the global economic system can be mapped into a modification of the AB Model. In [18] the power law's distribution of the income and of the returns in the stock market were compared to give extraordinary fit with reality. One more paper was just written [7] on how could decision makers gain knowledge on the economic system they are in charge of under the light of this model.

In this section we show in some more details one specific example of how one can take the predictions resulting from these types of models and apply them to real life systems: The system we will discuss is the Polish economy following the collapse of the Soviet bloc. We chose to present the system with the aid of Eq. (6). In the present application, the index i of the equations in the system (6) ranges from 1 to 2945 and labels the economy of each of the 2945 counties composing Poland. Each equation represents the evolution of the economic activity W_i of the county i. More precisely, W_i is the number of enterprizes per capita in the county i. The a_is represent endogenous growth rate of the country i and vary from county to county depending on local factors such as social capital, availability of natural resources or infrastructure. In fact the data indicate that the most important factor affecting the economic growth is the education level in the county. This dependence of a purely economical quantity on a social quantity is of great methodological importance and emphasizes in a dramatic way the importance of interdisciplinary studies (in this case economics, social science and physics). A recent work [43] led to a list of nine specific predictions resulting from the model. The data confirmed in a clear way the model predictions: Following the liberalization, the counties behaved in divergent ways: while most of the counties' economies plunged by factors of two, a few counties tripled their economic activity. This in turn lead to a quick increase in inequality between the counties.

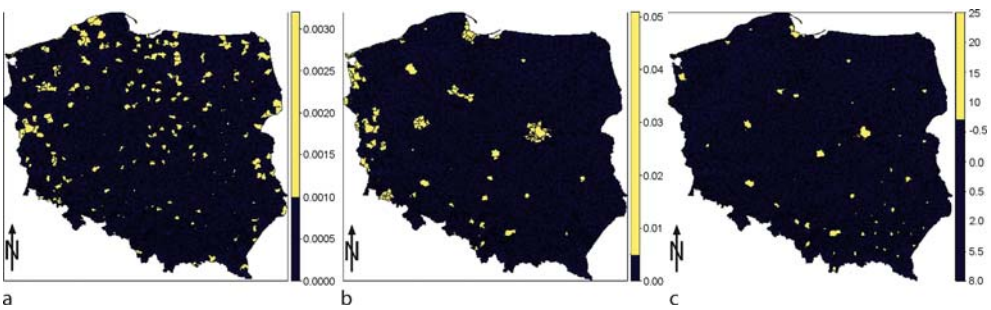

Intermittency and Localization, Figure 3

The influence of Education on Economic Activity before and after liberalization a The number of enterprizes per capita in each county in 1989. **b** The number of enterprizes per capita in each county in 1998. **c** The years of education per capita in various counties in 1988. **a** maps the number of the enterprizes per capita in the year preceding the economic transition. This initial distribution does not display any spatial pattern: is very close to a uniform random (Poisson) distribution (similar to Fig. 1 at $t = 0$). **b** After the liberalization there is a clear spatial pattern: the economic activity is concentrated around the singular growth centers which are strongly correlated with the education levels *before* the transition (**c**). In the language of the AB Model, the *A*s – represent the education level, while the *B*s represent the economic activity (enterprizes per capita)

During the preceding (socialist) regime, all counties were allocated roughly the same amount of economic activity by the central government. Thus the counties with high post-liberalization growth rate represented initially a negligible part of the country GDP and could not avert the fast global decay. However, within a couple of years, following their dramatic growth, the fast developing counties became the economic force, driving up the GDP. Moreover, their influence expanded to the neighboring regions until, eventually the entire country reached an uniform growth rate. This is not to say that the economic activity per capita equalized. Quite contrary, in the asymptotic regime in which the growth of the weak regions was due to the diffusion of economic activity from the fast developing regions, the very wide differences in GDP per capita persisted and in fact increased. One can see in Fig. 3 the spatial structure of the system, in $t = 0$ (year 1989, before the liberalization) and in 1994 and compare it with the social conditions (education level) that catalyzed the economic growth. The localization effect is very clear. In Fig. 4 one sees how the generic prediction of the J-shape resulting from the AB Model is present in all of the formerly communist countries!

Future Directions

In his science fiction novel "Foundation" (1951), Isaac Asimov was playing with the idea of having a reliable predictions of the human society under the new scientific discipline he invented and called psychohistory. In this novel he was dealing mainly with the fact that unlike other scientific disciplines – here, due to the fact that human beings

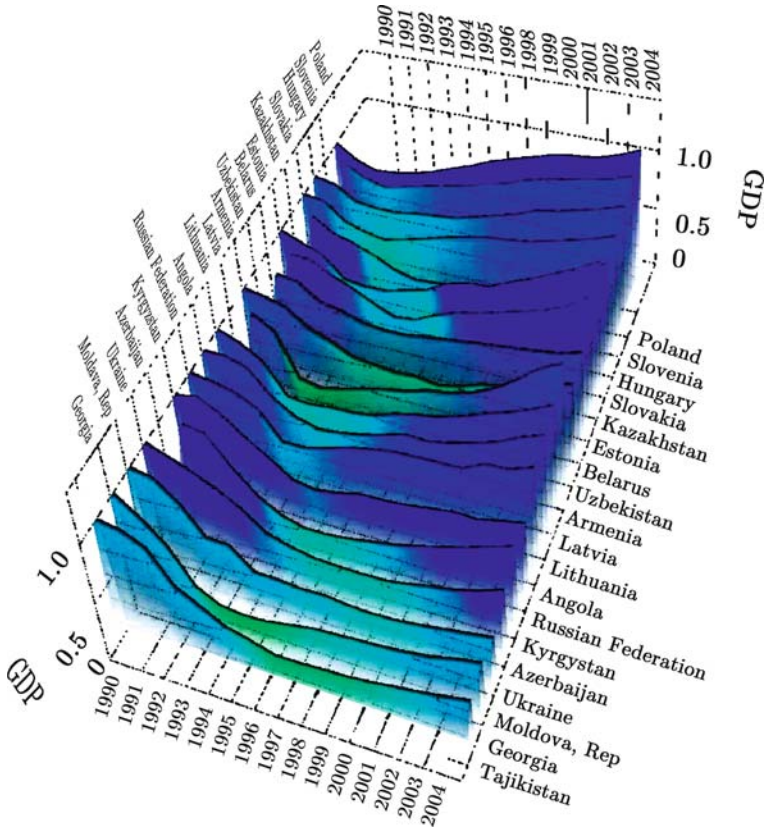

Intermittency and Localization, Figure 4

The J-curve economic recovery after the liberalization of the Soviet Block The GDPs of the Eastern European countries experienced strong decay immediately after the economic liberalization. This was generally followed by a growth period. The resulting pattern resembles the letter J which explains the name J-curve. While the magnitude of the initial decay and the time and rate of recovery varies among the various countries, the J-shape is universal. The marked departure (exponential growth) of the Polish economic activity from an exponential decay extrapolated curve indicates that the classical global logistic framework cannot explain the observed pattern. The AB Model however does! (as can be seen in Fig. 2)

are involved, they are able to read the predictions and by that changing them… Without entering this discussion, we feel that Asimov succeeded to put his finger on a very crucial point: unlike physics for example, there is no such tool that people can rely upon when trying to predict the possible future outcome of today's deeds. Many scientists that come from the natural sciences are very suspicious towards their colleagues from the soft sciences because of this reason. On the other hand the soft scientists are claiming that the problems that they deal with are far too complex to put into solvable equations. We do understand the positions of both sides, but feel that the time has come to try and close the cultural gap between the two. The methods of the accurate sciences have been improved dramatically since the availability of computer power, on the other hand the social sciences are able today (also due to computers and Internet) to measure many social indexes on a very wide scale and for many years. What is needed now is first to make efforts to quantify the observations that social scientists agree upon, and by that to create a set of so called stylized facts. After having such list of qualitative and quantitative motifs that science agree they are present in reality – the road to having models that could be validated or invalidated by comparing their theoretical predictions to reality is closer than ever. We do believe that what was described in this paper are first steps towards Asimov's fantasy. Maybe this line of research will help us understand a little better the complex nature of human society.

Acknowledgments

The present research was partially supported by the STREPs CO3 and DAPHNet of EC FP6, and by GIACS (General Integration of the Applications of Complexity in Science).

Bibliography

Primary Literature

1. Ben-Avraham D, Havlin S (2000) Diffusion and reactions in fractals and disordered systems. Cambridge University Press, Cambridge
2. Biham O, Huang ZF, Malcai O, Solomon S (2001) Long-time fluctuations in a dynamical model of stock market indices. Phys Rev E 64(2):026101
3. Biham O, Malcai O, Levy M, Solomon S (1998) Generic emergence of power law distributions and lévy-stable intermittent fluctuations in discrete logistic systems. Phys Rev E 58:1352–1358
4. Billari F, Fent T, Prskawetz A, Scheffran J (2006) Agent-based computational modelling. Physica, Heidelberg
5. Bońkowska K, Szymczak S, Cebrat S (2006) Microscopic modeling the demographic changes. Int J Mod Phys C 17:1477–1484
6. Cardy JL, Tauber U (1981) On the nonequilibrium phase transition in reaction-diffusion systems with an absorbing stationary state. Phys Rev Lett 77:4780
7. Challet D, Yaari G, Solomon S (2008) The therapy to shock-therapy: optimal dynamical policies for transition economies. Eur Phys J B
8. Ebeling W, Feistel R (1982) Physik der Selbstorganisation und Evolution. Akademie, Berlin
9. Eigen M (1971) Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften 58:465–523
10. Eigen M, Schuster P (1979) The Hypercycle. Springer, Berlin
11. Fisher RA (1937) The wave of advance of advantageous genes. Ann Eugen 7:355–369
12. Goldenberg J, Libai B, Solomon S, Jan N, Stauffer D (2000) Marketing percolation. Phys Stat Mech Appl 284:335–347
13. Goldenberg J, Shavit Y, Shir E, Solomon S (2005) Distributive immunization of networks against viruses using the 'honeypot' architecture. Nat Phys 1:184–188
14. Grassberger P (1982) On phase transitions in schlogl's second model. Z Phys B Condens Matter 47:365–374
15. Janssen HK (1981) On the nonequilibrium phase transition in reaction-diffusion systems with an absorbing stationary state. Z Phys B Condens Matter 42:151
16. Jiménez-Montano MA, Ebeling W (1980) A stochastic evolutionary model of technological change. Collect Phenom 3:107–114
17. Kesten H (1980) Random Processes in Random Environments. Springer, Heidelberg
18. Levy M, Levy H, Solomon S (2000) Microscopic simulation of financial markets: From investor behavior to market phenomena. Academic Press, New York
19. Lotka AJ (1923) Contribution to the analysis of malaria epidemiology. Am J Hyg 3:1–121
20. Louzoun Y, Shnerb NM, Solomon S (2007) Microscopic noise, adaptation and survival in hostile environments. Eur Phys J B 56:141–148
21. Louzoun Y, Solomon S, Atlan H, Cohen IR (2001) The emergence of spatial complexity in the immune system ELSEVIER. Physica A 297:242–252
22. Louzoun Y, Solomon S, Atlan H, Cohen IR (2003) Proliferation and competition in decrete biological systems. Bull Math Biol 65(3):375–396
23. Louzoun Y, Solomon S, Goldenberg J, Mazursky D (2003) World-size global markets lead to economic instability. Artif Life 9(4):357–370
24. Malthus TR (1798) An Essay on the Principle of Population. Printed for Johnson J by Bensley T. Reprinted with an introduction by Antony Frew. Harmondsworth, Middx: Penguin Books
25. Marsili M, Maslov S, Zhang YC (1998) Dynamical optimization theory of a diversified portfolio. Physica A 253:403–418
26. May RM (1976) Simple mathematical models with very complicated dynamics. Nature 261:459–467
27. Montroll EW (1978) Social dynamics and the quantifying of social forces. Proc Natl Acad Sci USA 75(10):4633–4637
28. Nelson RR, Winter SG (1982) The theory of economic development. Harvard Univ Press, Cambridge
29. Richmond P, Solomon S (2001) Power laws are disguised Boltzmann Laws. Int J Mod Phys C 12(3):333–343
30. Ross R (1911) The prevention of malaria. John Murray, London
31. Schumpeter J (1934) The theory of economic development. Harvard Univ Press, Cambridge

32. Shnerb NM, Bettelheim E, Louzoun Y, Agam O, Solomon S (2001) Adaptatioin of autocatalytic fluctuations to diffusive noise. Phys Rev E 63:21103–21108
33. Shnerb NM, Louzoun Y, Bettelheim E, Solomon S (2000) The importance of being discrete: Life always wins on the surface. Proc Natl Acad Sci USA 97(19):10322–10324
34. Solomon S (2000) Generalized Lotka Volterra (GLV) models of stock markets. In: Ballot G, Weisbuch G (eds) Applications of simulation to social sciences. Hermes Science Publications, Paris, pp 301–322
35. Solomon S, Richmond P (2001) Power laws of wealth, market order volumes and market returns. Physica A 299:(1)188–197
36. Solomon S, Richmond P (2001) Stability of Pareto–Zipf law in non-stationary economies. Computing in Economics and Finance 2001 11, Society for Computational Economics, Apr available at http://ideas.repec.org/p/sce/scecf1/11.html
37. Solomon S, Weisbuch G, de Arcangelis L, Jan N, Stauffer D (2000) Social percolation models. Physica A 277(1–2):239–247
38. Stauffer D, de Oliveira SM, de Oliveira PMC (2006) Biology, Sociology, Geology by Computational Physicists. Elsevier, Amsterdam
39. Verhulst PF (1838) Notice sur la loi que la population suit dans son accroissement. Corresp Math Phys 10:113–121
40. Volterra V (1931) Variations and Fluctuations of the Number of individuals in animal Species living together. Mc Graw Hill, NY
41. Weisbuch G, Solomon S, Stauffer D (2001) Economics with heterogeneous interacting agents. In: Lecture Notes in Economics and Mathematical Systems. Springer, Berlin, p 43
42. Yaari G, Deissenberg C, Solomon S (2006) Advertising, negative word-of-mouth and product acceptance. Eur J Econ Soc Syst 19(2):257–268
43. Yaari G, Nowak A, Rakocy K, Solomon S (2008) Microscopic study reveals the singular origins of growth. Eur Phys J B 62(4):505–513

Books and Reviews

For general books and reviews see [4,18], and [38].

Internet Topology

Yihua He, Georgos Siganos, Michalis Faloutsos
University of California, Riverside, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Data Sources and Their Limitations
Power-Laws of the Internet
Topology Generating Models and Tools
Conceptual Models for the Internet Topology
The Complete Internet Topology
Conclusion and Future Directions
Bibliography

Glossary

Autonomous system (AS) An autonomous system is a connected group of one or more IP prefixes, run by one or more network operators, which has a single and clearly defined routing policy. A unique AS number (or ASN) is allocated to each AS for identification purpose in inter-domain routing among ASes. For example, an organization, such as an ISP or a university, is an example of an AS. Some organizations may have more than one AS and thus have more than one AS number.

BGP (border gateway protocol) The Border Gateway Protocol (BGP) version 4 is the *de facto* routing protocol used in the Internet to exchange reachability information among ASes and interconnect them. The current BGP is version 4.

Degree The degree of an AS (or a node) is the number of neighbors of this AS (or node).

CCDF The complementary cumulative distribution function (CCDF) of a degree is the percentage of nodes that have a degree greater than the degree of interest.

Degree rank The degree rank of a node is its index in a list of degrees, ranked in decreasing order.

Eigenvalue Let A be an $N \times N$ matrix. If there is a vector $X \in \mathcal{R}^N \neq 0$ such that $AX = \lambda X$ for some scalar λ, then λ is called the eigenvalue of A with corresponding eigenvector X.

Definition of the Subject

Internet topology is the structure by which hosts, routers or autonomous systems (ASes) are connected to each other. The majority of existing Internet topology research focuses on the AS-level. There are three reasons for this. First, AS-level Internet topology is at the highest granularity of the Internet; other levels of Internet topology partially depend on AS-level topology. Second, the AS-level topology is relatively easy to obtain; other levels of topology are sometimes regarded as private information and are harder to get. Third, AS-level topology is not directly engineered by humans; instead, it is the aggregate result of technological and economical forces and, therefore, its origin and evolution attract considerable interest from investigators.

Research on Internet topology is driven by the explosive growth of the Internet, which has been accompanied by a wide range of inter-networking problems related to routing, resource reservation and administration. The study of algorithms and policies to address such problems often requires topological information and models. In 1999, Faloutsos et al. [27] discovered that the seemingly

random Internet topology does follow some rules: it follows power-law distributions. This finding revitalized the research on Internet topology and generated significant follow-up work.

Introduction

The Internet can be decomposed into connected subnetworks that are under separate administrative authorities, as shown in Fig. 1. These subnetworks are called *domains* or *autonomous systems* (ASes). The Internet community develops and employs different routing protocols inside an AS and between ASes. An intra-domain protocol, such as RIP, IS-IS, or OSPF, is runs within an AS, while an inter-domain protocol, such as BGP, runs between ASes. This way, the topology of the Internet can be studied at two different granularities. At the **router level**, each router is represented by a node [56], and a direct connection (either inter-domain or intra-domain) between any pair of routers is represented by an edge. At the AS level, each AS is represented by a single node [31] and each edge is an inter-domain interconnection. The study of the topology at each level is equally important.

This article focuses on AS-level Internet topology. Note that, at this level, only one edge exists between two nodes, although in practice there may be multiple connections between two ASes. This is a limitation of the nature of the data that are currently available.

There are multiple benefits from understanding the topology of the Internet. For example, we want to be able to answer questions such as the following: "What does the Internet look like?", "Are there any topological properties that don't change in time?", "How will it look a year from now?", "How can I generate Internet-like graphs for my simulations?". Modeling Internet topology is an important open problem despite the attention it has attracted recently. Paxson and Floyd consider this problem as a major reason why we don't know how to simulate the Internet [28]. An accurate topological model can have significant impact on network research. First, we can design more efficient protocols that take advantage of its topological properties. Second, we can create more accurate artificial models for simulation purposes. And third, we can derive estimates for topological parameters that are useful for the analysis of protocols and for speculation regarding the Internet topology in the future.

The rest of this article is structured as follows: We first will review how and where Internet topology information is collected in Sect. "Data Sources and Their Limitations". We also compare the pros and cons of different data sources and discuss why they have such different properties. In Sect. "Power-laws of the Internet", we will review the power-laws of the Internet, which constitute one of the most important discoveries of Internet topology. The power-laws lead to significant follow-up research in modeling the Internet topology, and we will discuss these models in Sects. "Topology Generating Models and Tools" and "Conceptual Models for the Internet Topology". Then, we present the challenges and techniques in discovering the complete Internet topology in Sect. "The Complete Internet Topology". Finally, we discuss future research directions of Internet topology in Sect. "Conclusion and Future Directions".

Data Sources and Their Limitations

We first describe the data sources for collecting AS-level Internet topology. All of these sources and methods have

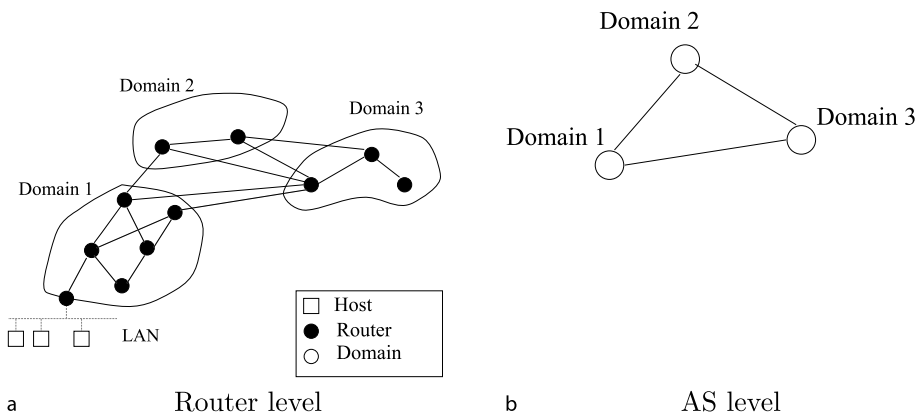

Internet Topology, Figure 1
The structure of Internet at two levels

Internet Topology, Table 1
An excerpt of an entry in typical Cisco "sh ip bgp" format BGP table dumps from BGP collector "route-views.oregon-ix.net" on May 1, 2007

Network	Next Hop	(Other Fields)	Path
2.0.0.0/24	157.130.10.233	(...)	701 1299 34211 41856 41856

their shortcomings, which will be discussed at the end of this section.

BGP Routing Tables

Border Gateway Protocol (BGP) routing table dumps are probably the most widely used resource that provides topological information. Typically, these routing table dumps are obtained from special BGP collectors, each of which connects with one or more Internet backbone routers in different ASes with special agreements. These BGP collectors do not advertise any prefixes (that is, IP blocks) to the Internet, while they are configured to receive all routes that the Internet backbone routers advertise to them. Therefore, these collectors are passive, and they have no effect on the global Internet. Periodically, each collector dumps its full routing table to Internet archives, which are usually publicly available for download.

Table 1 shows an entry of a typical Cisco "sh ip bgp" format BGP table dump from BGP collector "route-views.oregon-ix.net" on May 1, 2007. This entry indicates that, the destination network 2.0.0.0/24 can be reached via AS path "701 1299 34211 41856". Therefore, the instance of Internet topology should include four ASes (AS701, AS1299, AS34211 and AS41856) and three links (AS701–AS1299, AS1299–AS34211 and AS34211–AS41856). Typically, a BGP collector's routing table has more than a hundred thousands such entries from each peer AS; the total number of entries often exceeds several million. The number of ASes or routers of each BGP collector varies from a few to approximately one hundred. There are two well-maintained BGP routing table collector agents: Oregon Routeviews [54] and RIPE RIS [61]. Each of these agents maintains a number of BGP collectors around the world.

Besides routing tables, a BGP collector may also periodically dump routing updates received from its peers. Routing updates have a similar format to routing tables. A BGP update message displays the current route to a prefix, and therefore, a collection of BGP updates is able to reveal the dynamics of BGP routing.

Traceroute

Traceroute [36] is a tool that discovers the route that IP datagrams follow from one host to another. Traceroute takes advantage of the fact that each router decrements the TTL (Time To Live) field by 1 for each IP packet that passes through it, and each router must discard any IP packet with TTL = 0 and send an ICMP "time-exceeded" error message back to the sender of the original IP packet. The original purpose of the action is to prevent IP packets from circulating the network forever. Traceroute operates by sending IP packets to a destination with small but increasing TTL values. These packets expire at the routers along the path to the destination. Since each router along the path sends an ICMP "time-exceeded" message back to the traceroute source, the identities of these routers (or more precisely, the outgoing IP interfaces of those routers) can be discovered. Although there is no guarantees that two consecutive IP packets will traverse the same route to the same destination, most often they do.

There are several large scale measurement projects using traceroute-like probes. Skitter [66] is a part of CAIDA's [70] topology measurement project. CAIDA [70] maintains a set of (about 20) active monitors distributed around the globe. Each monitor uses a modified version of traceroute to probe a large set of IP addresses which cover nearly the whole IP address space. Rocketfuel [67] is a topology discovery project from the University of Washington. Rocketfuel uses a larger number of traceroute sources (a few hundred) from public traceroute servers as their sources. Therefore, Rocketfuel has a significantly higher number of vantage points than CAIDA. However, due to restrictions of the public traceroute servers, the rate of traceroute probing is limited in Rocketfuel. As a result, Rocketfuel is better at probing specific ISP networks rather than the whole Internet. Another promising project is NetDimes [62] from Tel Aviv University. To increase the number of vantage points, NetDimes distributes a large number of probing agents (tens of thousands) to global Internet users on a volunteer basis. These agents perform traceroutes according to the NetDimes center controls. Since the agents are mostly volunteers, coordination is still difficult when attempting to probe the Internet topology from anywhere at any given time.

All traceroute probes reflect only router-level topology. In order to obtain AS level topology, the probed IP addresses must be mapped to the ASes to which they belong. The conventional way to map an IP address to its AS number is by looking up, in the BGP routing tables, the IP block with the longest prefix match. For example, if

there is an IP address 2.0.0.18 and the longest prefix that it matches in the routing table is 2.0.0.0/24 (as shown in Table 1), then the announcing AS, which is the AS at the end of the "Path" field (AS41856 in Table 1), is the AS to which the IP 2.0.0.18 should be mapped. However, the accuracy of this method may be low in certain situations, as reported in the literature [48,49].

Internet Routing Registry (IRR)

The need for cooperation between autonomous systems is fulfilled today by the Internet Routing Registry (IRR) [54]. ASes use the Routing Policy Specification Language (RPSL) [61,66] to describe their routing policy, and router configuration files can be produced from it. At present, 55 registries exist, which form a global database from which to obtain a view of the global routing policy. Some of these registries are regional, such as RIPE or APNIC; other registries describe the policies of an autonomous system and its customers. The main uses of the IRR registries are to provide an easy way for consistent configuration of filters, and a way to facilitate the debugging of Internet routing problems. From the registered routing export and import policies, Internet topology can be extracted from IRR. For example, in Table 2, an excerpt of *aut-num* record for AS3303 in the IRR is shown. From the registered import and export policy in this excerpt, the Internet topology should include three ASes (AS3303, AS701 and AS1239), and two edges (AS3303–AS701 and AS3303–AS1239).

Data Source Comparison

BGP table dumps, especially the one from the Oregon Routeview project, are the most widely used source for Internet topology studies. An advantage of the BGP routing tables is that their link information is considered reliable.

Internet Topology, Table 2
An excerpt of the IRR in plain text format for AS3303

aut-num:	AS3303
as-name:	SWISSCOM
descr:	Swisscom Solutions Ltd
descr:	IP-Plus Internet Backbone
…	…
import:	from AS701 action pref=700; accept ANY
export:	to AS701 announce AS -SWCMGLOBAL
import:	from AS1239 action pref=700; accept ANY
export:	to AS1239 announce AS -SWCMGLOBAL
…	…

If an AS link appears in a BGP routing table dump, it is almost certain that the link exists. However, a limited number of vantage points makes it hard to discover a complete view of the AS-level topology. A single BGP routing table has the union of "shortest" or, more accurately, "preferred" paths with respect to this point of observation. As a result, such a collection will not see edges that are not on the preferred path for this point of observation. Several theoretical and experimental efforts explore the limitations of such measurements [1,21]. Furthermore, the incompleteness is statistically biased based on the type of the links: peer-to-peer links are more likely to be missing from BGP routing tables than provider-customer links, due to the selective exporting rules of BGP. Here, a provider-customer link means the two ASes incident to the edge have a provider-customer relationship, that is, one AS pays the other AS for traffic transit service. A peer-to-peer edge means the two ASes incident to the edge have a peer-to-peer relationship, that is, these two ASes have a mutual agreement that they carry traffic for each other with no or little fee. The classification of AS relationships can be inferred fairly accurately from the BGP routing tables by a number of algorithms [23,25,30,77]. The majority of the edges (approximately 80%) are provider-customer edges and most of the rest are peer-to-peer edges [33]. Typically, a peer-to-peer link can be seen only in a BGP routing table of the two peering ASes or their customers. Thus, given a peer-to-peer edge, unless a BGP collector peers with a customer of either AS incident to the edge, the edge can not be detected from the table dumps of the BGP collector. A recent work [18] discusses this limitation in depth. Thus, apart from being incomplete, the measured graph may not fairly represent the different types of links. Furthermore, BGP table dumps are likely to miss alternative and back-up paths. By definition, a router advertises only the best path to each destination, namely an IP prefix. Therefore, the back-up paths will not show up unless the primary link breaks. To address the problem, a recent effort suggests the need for actively probing backup links [19].

Previous studies [24,80] using BGP updates as a source of topological information show that by collecting BGP updates over a period of time, more AS links are visible. This is because as the topology changes, BGP updates provide transient and ephemeral route information. However, if the window of observation is long, an advertised link may cease to exist [80] by the time that we construct a topology snapshot. In other words, BGP updates may provide a superimposition of a number of different snapshots that existed at some point in time. Recently, Oliveira et al. [53] explicitly distinguished this commonly overlooked "liveness problem". Note that BGP updates are col-

lected at the same vantage points as the BGP tables in most collection sites. Naturally, topologies derived from BGP updates share the same statistical bias per link type as topologies derived from BGP routing tables: peer-to-peer links are advertised only to the peering ASes and their customers. This further limits the additional information that BGP updates can currently provide. On the other hand, over a long period of observation, BGP updates could be useful in revealing ephemeral backup links which are not visible in the Internet at large, unless the primary link breaks down. At the same time, we could obtain erroneous BGP updates which may introduce fictitious links. To tell ephemeral and erroneous links apart, we need highly targeted probes. Recently, active BGP probing [19] has been proposed for identifying backup AS links. This is a promising approach that could complement other measurement work and provide the needed capability for discovering more AS links.

By using traceroute, one can explore IP paths and then translate the IP addresses to AS numbers, thus obtaining AS paths. Similarly to BGP tables, the traceroute path information is considered reliable, since it represents the path that the packets actually traverse. On the other hand, a traceroute server explores the routing paths from its location towards the rest of the world, thus the collected data has the same limitations as BGP data in terms of completeness and link bias. One additional challenge with the traceroute data is the mapping of an IP path to an AS path, as we previously mentioned. The problem is far from trivial, and it has been the focus of several recent efforts [48,49].

The Internet Routing Registry (IRR) [35] is the union of a growing number of world-wide routing policy databases that use the Routing Policy Specification Language (RPSL). In principle, each AS should register routes to all its neighbors (that reflect the AS links between the AS and its neighbors) with this registry. IRR information is manually maintained and there is no stringent requirement for updating it. Therefore, without any processing, AS links derived from the IRR are prone to human error and could be outdated or incomplete. However, up-to-date IRR entries provide a wealth of information that could not be obtained from any other source. A recent effort [63] shows that, with careful processing of the data, we can extract a non-trivial amount of correct and useful information.

Power-Laws of the Internet

The power-laws for Internet topology were first observed by Faloutsos et al. [27], and later summarized in [64]. In those two papers, the authors showed that Internet topology at the AS level can be described efficiently with power-laws. The elegance and simplicity of the power-laws provide a novel perspective into the seemingly chaotic Internet structure.

Power-laws are expressions of the form $y \propto x^a$, where a is a constant, x and y are the measures of interest, and $\propto$ stands for "proportional to". Pareto was among the first to introduce power-laws in 1896 [57]. He used power-laws to describe distribution of income where there are a few very rich people, but most people have a low income. Another classical law, the Zipf law [81], was introduced in 1949, for the frequencies of English words and the populations of cities. More recently, power-laws have been observed in communication networks. Power-laws have been observed in traffic [20,43,59]. In addition, the topology of the World Wide Web [5,42] can be described by power-laws. Furthermore, power-laws describe the topology of peer-to-peer networks [40] and properties of multicast trees [16,60,71,76].

The initial work on power-laws [27] has generated significant follow-up work. In fact, [27] is one of the top five most cited computer science papers published in 1999 [17]. Various researchers have verified the power-law observations with different datasets [32,38,45]. In addition, significant work has been devoted to understanding the origin [50], and generating power-law topologies [8,13,37,39,50,51,55,69,78].

For the Internet topology, three power-laws have been identified: the rank power-law, the degree power-law and the eigen power-law.

Rank power-law

We empirically observe the following property for the Internet.

Power-law 1 (rank exponent) Given a graph, the degree d_v of a node v is proportional to the rank of the node r_v to the power of a constant $\mathcal{R}$:

$$d_v \propto r_v^{\mathcal{R}}$$

Definition 1 Let us sort the nodes of a graph in decreasing order of degree. We define the rank exponent $\mathcal{R}$ to be the slope of the plot, on a log–log scale, of the degrees of the nodes versus the rank of the nodes.

Figure 2 shows the (r_v, d_v) pairs ona log–log scale after the nodes in an Internet topology are sorted in decreasing order of degree d_v. The measured data are obtained from the Oregon Routeviews [54] collector and are represented by

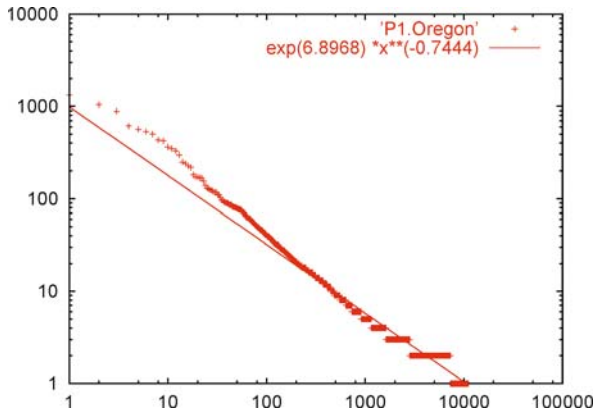

Internet Topology, Figure 2
Log-log plot of the degree d_v versus the rank r_v in the sequence of decreasing degree

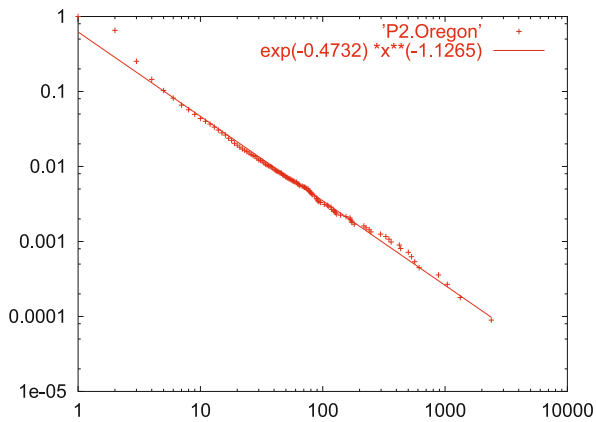

Internet Topology, Figure 3
The log–log plot of D_d versus the degree for the Oregon topologies

points, while the solid line represents the least-squares approximation. A striking observation is that the plots are approximated well by linear regression. The correlation coefficient is over 0.97 in this case. The authors of [64] also inspected more than 1000 Internet topology instances over a six year span (1997 and 2003), and they found that for every instance of the inter-domain topology, the correlation coefficient was always higher than 0.97. This linearity is unlikely to be a coincidence.

Intuitively, power-law 1 correlates the degrees of the nodes and their rank and reflects a principle of the way domains connect. Such a relationship can be used to calculate the number of edges as a function of the number of nodes for an exponent of a given rank. In fact, in a graph where power-law 1 holds, it can be shown [27] that the number of edges E of the graph can be estimated as a function of the number of nodes N and the rank exponent R as follows:

$$E = \frac{N}{2(R+1)}\left(1 - \frac{1}{N^{R+1}}\right)$$

For additional discussion on estimates using this formula, see [27].

Degree Power-Law

Power-law 2 (degree exponent) Given a graph, the CCDF D_d of a degree d is proportional to the degree to the power of a constant $\mathcal{D}$:

$$D_d \propto d^{\mathcal{D}} .$$

Definition 2 We define the degree exponent $\mathcal{D}$ to be the slope of the plot of the cumulative distribution of the degrees versus the degrees in log–log scale.

In Fig. 3 D_d is plotted versus the degree d in log–log scale. The major observation is that the plot is linear. The correlation coefficient is more than 0.996 for data obtained from Oregon Routeviews [54]. The authors in [64] found that the degree power-law holds for all the instances they inspected from 1997 to 2003, with correlation coefficient higher than 0.99.

The intuition behind this power-law is that the distribution of the degree of Internet nodes is not arbitrary. The qualitative observation is that lower degrees are more frequent. The power-law manages to quantify this observation by a single number, the degree exponent. This way, the realism of a graph can be tested with a simple numerical comparison. If a graph does not follow power-law 2, or if its degree exponent is considerably different from real exponents, it probably does not represent a realistic topology.

The exponents of rank and degree power-laws have been shown to be related [2,15]. More specifically, in a perfect power-law distribution, the exponent of the rank power-law is equal to the multiplicative inverse of the exponent of the degree power-law. However, in reality, the two exponents do not have such a perfect relationship. The discrepancy could be attributed to measurement imperfections and inaccuracies. In that regard, both the rank and the degree power-laws characterize the degree distribution from different angles, and it is useful to report both exponents when characterizing a topology.

Eigen Power-Law

Eigenvalues of a graph are the eigenvalues of its adjacency matrix. We observe the following property for the Internet graph.

Power-law 3 (eigen exponent) Given a graph, the eigenvalues λ_i are proportional to the order i to the power of a constant $\mathcal{E}$:

$$\lambda_i \propto i^{\mathcal{E}} .$$

Definition 3 We define the eigen exponent $\mathcal{E}$ to be the slope of the plot of the sorted eigenvalues versus their order in log–log scale.

In Fig. 4, the eigenvalues are plotted versus their order, in decreasing sequence, in log–log scale. The eigenvalues are shown as points in the figure, and the solid lines are approximations using a least-squares fit. Similar observations with equally high correlation coefficients were observed for all instances obtained between 1997 and 2003 [64]. The plot is practically linear with a correlation coefficient of 0.996, which constitutes an empirical power-law of the Internet topology.

Eigenvalues are fundamental graph metrics. There is a rich literature that proves that the eigenvalues of a graph are closely related to many basic topological properties such as the diameter, the number of edges, the number of spanning trees, the number of connected components, and the number of walks of a certain length between vertices, as shown in [22]. Interestingly, Mihail et al. [52] show that there is a surprising relationship between the eigen exponent and the degree exponent: the eigen exponent is approximately half of the degree exponent. In practice, the exponents adequately obey the mathematical relationship, although the match is, naturally, not perfect. All of the above suggest that the eigenvalues intimately relate to topological properties of graphs.

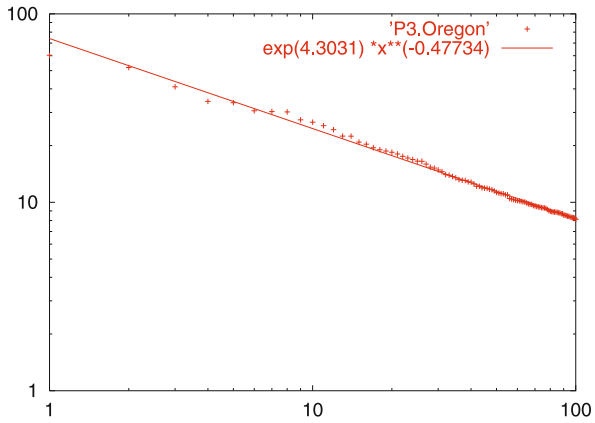

Internet Topology, Figure 4
The eigenvalues plot for the Oregon topologies

The Doubts and the Settlement

There has been a long debate on whether the degree distribution of the Internet at the AS level can be accurately characterized by power-laws [11,13]. The major concern is that by adding new edges discovered from sources other than Oregon Routeviews [54], the degree distribution of the Internet topology deviates from a perfect power-law. Interestingly, most of these new edges are of the peer-to-peer type which is underrepresented in the initial topologies as we explained earlier.

There are at least two reasons for this debate. First, this debate is partly due to the absence of a definitive statistical test. In Fig. 5a, the CCDF of node degrees is plotted for an Internet topology instance obtained from multiple resources, including Routeviews [54] and verified edges from IRR [35]. The distribution is highly skewed, and the correlation coefficient of a least square errors fitting is 98.9%. However, one could still use different statistical metrics and argue against the accuracy of the approximation [13]. Second, the answer could vary depending on which source we think is more complete and accurate, and the purpose or the required level of statistical confidence of a study. In Fig. 5b, the CCDF is plotted for an Internet topology instance obtained from the IRR after being filtered by the Nemecis tool [63]. The correlation coefficient is only 93.5%.

A recent paper [33] proposes a reconciliatory divide-and conquer approach to explain and settle the debate. The authors propose to separately model the degree distribution according to the types of the edges: provider-customer and peer-to-peer. In Fig. 5, an indicative set of degree distribution plots are shown. The graphs obtained from multiple sources (Oregon Routeviews and traceroute-verified IRR links) are plotted in the left column (Fig. 5a,c,e), and the topology obtained from Nemecis-filtered IRR are plotted in the right column (Fig. 5b,d,f). The distributions for the whole graph are shown in the top row, only the provider-customer edges in the middle row, and only the peer-to-peer edges in the bottom row. The power-law approximation in the first two rows of plots and the Weibull approximation in the bottom row of plots are shown.

The following two properties can be observed from Fig. 5: (1) The provider-customer-only degree distribution can be accurately approximated by a power-law. The correlation coefficient is 99.5% or higher in the plots of Fig. 5c,d. Note that, although the combined degree distribution of the topology in the IRR does not follow a power law as shown in Fig. 5b, its provider-customer subgraph follows a strict power law in Fig. 5d. (2) The peer-to-peer-

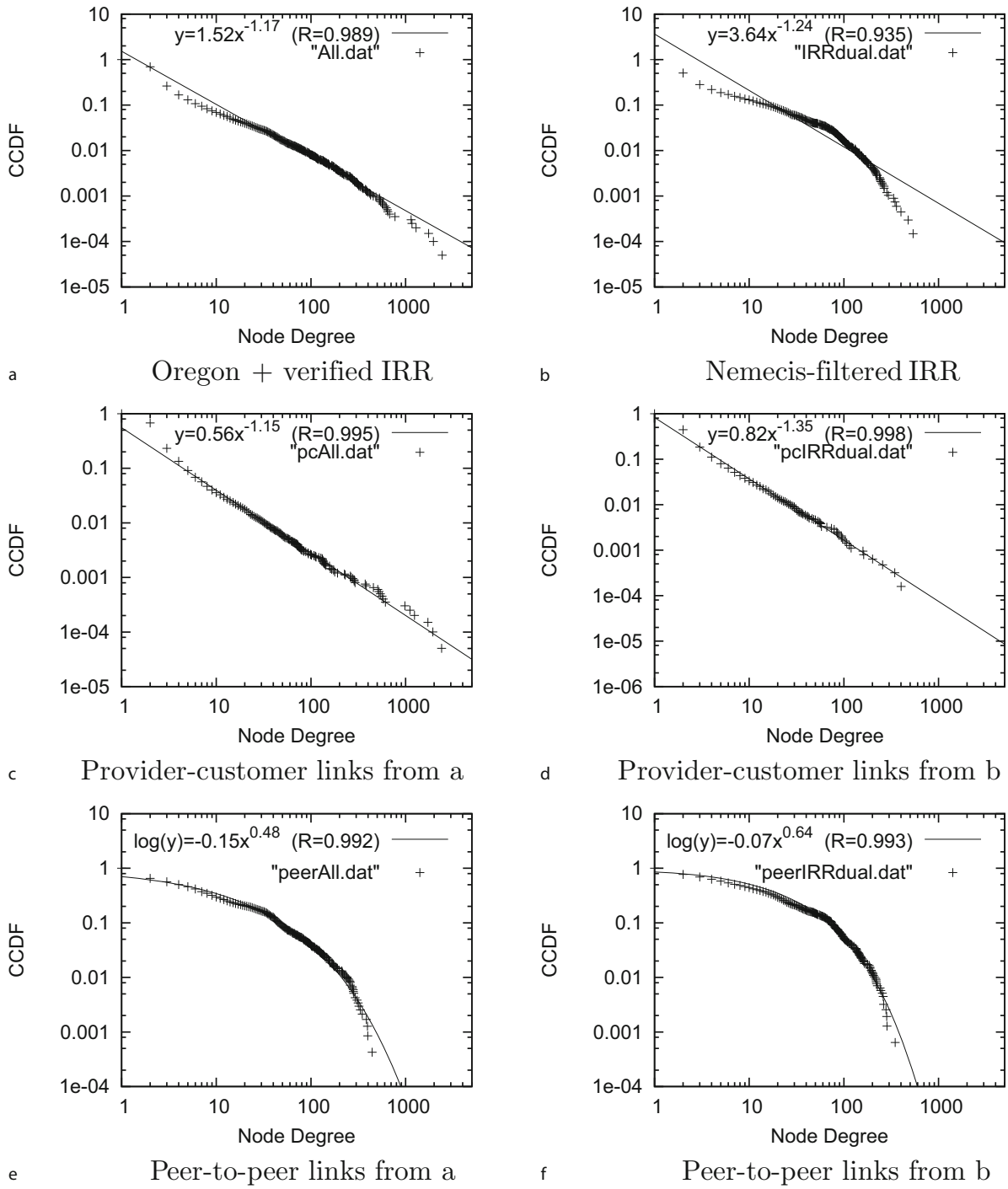

Internet Topology, Figure 5

The degree distributions of *Oregon + verified IRR* (*left*) and *Nemecis-filtered IRR* (*right*) in the *top row*, their provider-customer degree distributions in the *middle row*, and their peer-to-peer degree distributions in the *bottom row*

only degree distribution can be accurately approximated by a Weibull distribution [75]. The correlation coefficient is 99.2% or higher in the plots of Fig. 5e,f. It is natural to ask why the two distributions differ. The following could be one explanation: Power-laws are related to the rich-get-richer behavior; low degree nodes "want" to connect to high degree nodes. For provider-customer edges, this makes sense: an AS wants to connect to a high-degree provider, since that provider would likely provide shorter paths to other ASes. This is less obviously true for peer-to-peer edges. If AS1 becomes a peer of AS2, AS1 does not benefit from the other peer-to-peer edges of AS2 due to routing policies [72]: an AS normally will not carry traffic from one of its peers to its other peers. Therefore, high peer-to-peer degree does not make a node more attractive as a peer-to-peer neighbor. The validity of this explanation is still under investigation [33].

Topology Generating Models and Tools

So far, we have tried to model the learned topology as a snapshot. In this section, we present models and tools that attempt to generate Internet-like topologies.

Early Models

The simplest model is probably the **pure random** model. In this model, a set of nodes is distributed in a plane, and an edge is added between each pair of the nodes with a fixed probability p. Although this model does not explicitly attempt to reflect any structure of real networks, it is attractive for its simplicity: it is controlled by a single parameter.

The **Waxman** [73] model, on the other hand, adds edges with a probability that is a function of the distance between the nodes. Specifically, this probability for an edge between node u and node v is given by

$$P(u,v) = \alpha e^{-d/(\beta L)} , \quad (1)$$

where $0 < \alpha, \beta \leq 1$, d is the Euclidean distance from u to v, and L is the maximum distance between any two nodes. There are several variations of the Waxman model [26, 74,79].

The **transit–stub** [9] method tries to impose a more Internet-oriented hierarchical structure as follows: In this model, each routing domain in the Internet can be classified as either a *stub* domain or a *transit* domain. A domain is a stub domain if the path connecting any two nodes u or v is in that domain; transit domains do not have this restriction – in other words, transit domains carry traffic not only for its own, but also for other domains as well. In more detail, a connected random graph is first generated (by using, for example, the Waxman method described above). Each node in that graph represents an entire transit domain. Each transit domain node is expanded to form another connected random graph representing the backbone topology of that transit domain. Next, for each node in each transit domain, a number of connected random graphs are generated, representing stub domains that are attached to that transit node. Finally, some extra connectivity is added, in the form of "back-door" links between pairs of nodes, where a pair consists of a node from a transit domain and another from a stub domain, or one node from each of two different stub domains. By having nodes of different types, it is possible to generate large sparsely-connected Internet-like topologies with typically low node degrees. **GT-ITM**(Georgia Tech Internetwork Topology Models) is a tool that uses this transit–stub model.

The problem of these early models is that they do not generate power-law distributions. Medina et al. [51] tested the generated topologies from Waxman and Transit–Stub, and found both exhibit a weak presence or lack of the power-laws.

Pure Power-Law Models

Since the discovery of power-laws by Faloutsos et al. [27], the main focus of generating an Internet-like topology has shifted to matching the power-law exhibited in the Internet.

Palmer et al. [55] proposed the **PLOD** (power law out-degree) model. In this model, a degree credit is first assigned to each node in a graph with a given number of nodes. The degree distribution complies with the appropriate power-laws. Then, an edge placement loop is executed: it randomly picks two nodes and assigns an edge if they are not connected and each node still has remaining degree credit. After an edge is assigned, the degree credit of the nodes incident to the edge is deducted accordingly. The loop continues until there are no more pairs of nodes that fulfill the condition.

The concept of **PLRG** (Power Law Random Graph) was proposed by Aiello et al. [3] in the year 2000, and this model is also sometimes called **Model A**. In this model, a random graph is produced with a power-law degree distribution depending on two parameters that roughly delineate the size and density but are natural and convenient for describing a power law degree sequence. The power-law random graph model $P(\alpha, \beta)$ is described as follows: Let y be the number of nodes with degree x. $P(\alpha, \beta)$ assigns uniform probability to all graphs with $y = e^{\alpha}/x^{\beta}$, where α is the intercept and β is the (negative) slope when the de-

gree sequence is plotted on a log–log scale. After the degree distribution is defined, a set, L, which contains deg(v) distinct copies of each node v, will be formed. Then a random matching of the elements of L is chosen. For two nodes u and v, the number of edges joining u and v is equal to the number of edges in the matching of L joining copies of u to copies of v. The graph formed in the end is the PLRG.

These generators do not attempt to answer how a graph comes to have a power law degree sequence. Interestingly, these method seem to be able to match many other (but not all) topology properties of the real Internet [65].

Dynamic Growth Models

In contrast to the pure power-law models, dynamic growth models try to generate the Internet topology graph by simulating the growth of the Internet.

Barabasi and Albert [6] proposed a generic model (the **BA Model**) for network growth that relies on:

1. **Incremental growth**: The network expands continuously by the addition of new nodes.
2. **Preferential attachment**: A new node attaches preferentially to nodes that are already well connected.

In more detail, the network begins with a small number (m_0) of connected nodes. New nodes are added to the network one at a time. The probability $p(v)$ that a new node is connected to an existing node v is determined as the following:

$$p(v) = d_v / \sum_j d_j \,, \tag{2}$$

where d_v is the degree of node v and $\sum_j d_j$ is the sum of degrees of all existing nodes. In BA model, heavily linked nodes tend to quickly accumulate even more links, while nodes with only a few links are unlikely to be chosen as the destination for a new link. The new nodes have a "preference" to attach themselves to the already heavily linked nodes. This is so called "rich-get-richer" phenomenon.

The **AB** model [4] extends the BA model by adding a third operation called "rewiring". The rewiring operation consists of choosing m links randomly and rewiring each end of them according to the same preference rule used in the BA model.

Bu et al. [8] found that the graphs generated by PLRG, BA and AB models have different characteristic values from the real Internet graph in terms of path length and clustering coefficient. They proposed **GLP** (Generalized Linear Preference) [8], in which the probability p is

$$p(v) = (d_v - \beta) / \sum_j (d_j - \beta) \,, \tag{3}$$

where $\beta \in (-\infty, 0)$ is a tunable parameter. The smaller the value of β, the less preference given to high degree nodes.

All these dynamic growth models produce graphs with power-law distribution. However, it is still difficult for these models to capture every topological property of the Internet. Authors in [65] show that even GLP does not follow some hierarchical properties of the Internet.

Sampling

All aforementioned models attempt to grow a graph, an approach which we call "constructive". These methods depend on the principles of construction and the choice of parameter values. Furthermore, they often focus on matching a certain number of topology properties, while failing to match some others. At the same time, one can observe that we need small realistic graphs for simulation purposes and we do have several large measured topologies. To create small realistic graphs, Krishnamurthy et al. [41] proposed a "reductive" approach: instead of trying to construct a graph, they try to "sample" real topologies to produce a smaller graph. The idea is that the original properties, either well-known or unnoticed, can be kept during the process of reduction.

In more detail, they propose several reduction methods: **DRV** (Deletion of Random Vertex): Remove random vertexes, each with the same probability. **DRE** (Deletion of Random Edge): Remove random edges, each with the same probability. **DRVE** (Deletion of Random Vertex or Edge): Select a vertex uniformly at random, and then delete an edge chosen uniformly at random from the edges incident on this vertex. **DHYB-w** (Hybrid of DRVE/DRE): In this method, DRVE is executed with probability w and DRE is executed with probability $1 - w$, where $w \in [0, 1]$. This method was motivated by the study showing that sometimes DRVE and DRE had opposite performances with respect to different metrics.

The topologies sampled by both DRV and DRE methods are mathematically proved to follow power-law degree distribution. By comparing experimental data, the authors in [41] concluded that DHYB-0.8 is the best reduction method, and it also compares favorably to graph generation methods proposed previously in the literature. These sampling methods are successful to reduce a topology down to 30% of the original size. Beyond that the statistical confidence is found to be low.

Topology Generation Tools

Several publicly available topology generation tools exist.

BRITE (Boston university Representative Internet Topology gEnerator) [7]is a universal topology generator. It implements a single generation model that has several degrees of freedom with respect to how the nodes are placed in the plane and the properties of the interconnection method to be used. With different parameter settings, BRITE can generate either a Waxman model or a BA model.

Inet [34] is an AS-level Internet topology generator. Inet aims at reproducing the connectivity properties of Internet topologies as power-laws with additional improvements. It initially assigns node degrees from a power-law distribution and then proceeds to interconnect them using various rules. The current version, Inet-3.0, improves on previous versions by creating topologies with more accurate degree distributions and minimum vertex covers as compared to Internet topologies. Inet-3.0's topologies still do not well represent the Internet in terms of maximum clique size and clustering coefficient. These related problems stress a need for a better understanding of Internet connectivity and will be addressed in future work.

Conceptual Models for the Internet Topology

The Internet topology is large, complex and constantly changing. Even with the introduction of power-laws, which appear as a necessary though not sufficient condition for a topology to be realistic, a conceptual model of the topology [45,56,64] is still hard to obtain. Although the Internet is widely believed to be hierarchical by construction, it is too interconnected for an obvious hierarchy [69]. Several efforts to visualize the router-level topology have been made [14,66], however they can not be recreated manually and they do not provide a memorable model.

One goal here is to develop an effective conceptual model: a model that can be easily drawn by hand, while at the same time, is able to capture significant macroscopic properties. The jellyfish [65] and medusa [10] models are two conceptual models proposed for the inter-domain Internet topology.

The Jellyfish Model

The jellyfish model classifies ASes into different hierarchical layers. The highest layer is called the **Core**, which can be constructed in the following way: First, we sort all ASes in non-increasing degree order. The highest degree node is selected as the first member of the Core. Then, each AS is examined in the previous order; a node is added to the Core only if it forms a clique with the nodes already in the Core. In other words, the new node must connect to all the nodes already in the Core. The procedure stops when no more nodes can be added. The constructed Core is a clique of high-degree ASes, but not necessarily the maximal clique of the graph. The Core is a starting point to construct a jellyfish topology, and the ASes in the Core are probably the most important ASes in the Internet, since they have high degrees. The rest of the nodes are defined according to their proximity to the Core. The first **layer** is defined as all the ASes adjacent to the Core. Similarly, the second layer is defined as the non-labeled neighbors of the first layer. By repeating this procedure, six layers can be identified from the instances of Internet AS-level topology if the Core is counted as layer zero. Table 3 shows the number and percentage of ASes with each layer for three Internet topology instances at different times.

The jellyfish model also separately studies the one-degree ASes. First, the one-degree nodes are not useful in terms of connectivity to the rest of the network. Second, one-degree nodes are a large percentage of the network, and it is important to clarify and isolate their role. In fact,

Internet Topology, Table 3
Distribution of nodes in layers for three Internet instances

	Instance					
	Int-11-1997		Int-06-2000		Int-07-2003	
Layer No	Nodes	% of Nodes	Nodes	% of Nodes	Nodes	% of Nodes
Core/Layer-0	8	0.23	14	0.176	13	0.08
Layer-1	1354	44.90	3659	46.25	7330	46.27
Layer-2	1202	39.866	3090	39.05	7116	45.51
Layer-3	396	13.134	1052	13.29	1078	6.89
Layer-4	43	1.425	86	10.87	96	0.61
Layer-5	12	0.398	10	0.12	1	0.0063

Internet Topology, Table 4
Distribution of nodes in shell and hang classes

	Instance					
	Int-11-1997		Int-06-2000		Int-07-2003	
Layer ID	Nodes	% of Nodes	Nodes	% of Nodes	Nodes	% of Nodes
Core/Shell-0	8	0.23	14	0.176	13	0.08
Hang-0	465	15.42	798	10.08	1174	7.5
Shell-1	889	29.49	2861	36.16	6156	39.37
Hang-1	623	20.66	1266	16	2821	18.04
Shell-2	579	19.2	1824	23.05	4295	27.47
Hang-2	299	9.92	662	8.36	808	5.16
Shell-3	97	3.22	390	4.92	270	1.72
Hang-3	41	1.36	74	0.93	84	0.53
Shell-4	2	0.66	12	0.15	12	0.07
Hang-4	12	0.4	10	0.12	1	0.006

35%–45% of the ASes in the Internet are one-degree. In the jellyfish model, each layer is separated into two classes: a) the multiple-degree or **shell** nodes, and b) the one-degree or **hang** nodes. The one-degree nodes hanging from kth shell are referred as the kth hang class. For example, shell-0 is the Core, and its one-degree neighbors are denoted as hang-0, while the rest of the neighbors constitute shell-1. Naturally, the number of ASes in the layers, shells and hangs have the following relationship:

$$\text{Layer}_k = \text{Shell}_k + \text{Hang}_{k-1}$$

Table 4 shows the size of each group of nodes in the classification.

The conceptual the jellyfish model is described by the layer-shell-hang classification. The Core is the center of the head of the jellyfish surrounded by shells of nodes. Figure 6 shows a graphical illustration of this model. The hang nodes form the tentacles of the jellyfish. The length of the tentacle represents the concentration of one-degree neighbors for each shell.

Besides being a conceptual model, the jellyfish captures and represents concisely some fundamental properties of the Internet topology.

1. Core: The topology has a Core of highly connected important nodes, which is represented by the center of the jellyfish cap.
2. Center-heavy: Approximately 80% of the ASes are layer-1, layer-2 and layer-3 (see Table 3).
3. Node distance: Distances between ASes are small; maximum distance less than 11 hops, and 80% of the ASes are within 5 hops.
4. Edge types: Approximately 70% of the edges are between different node layers. The rest are horizontal to the hierarchy providing connectivity between nodes of the same layer.

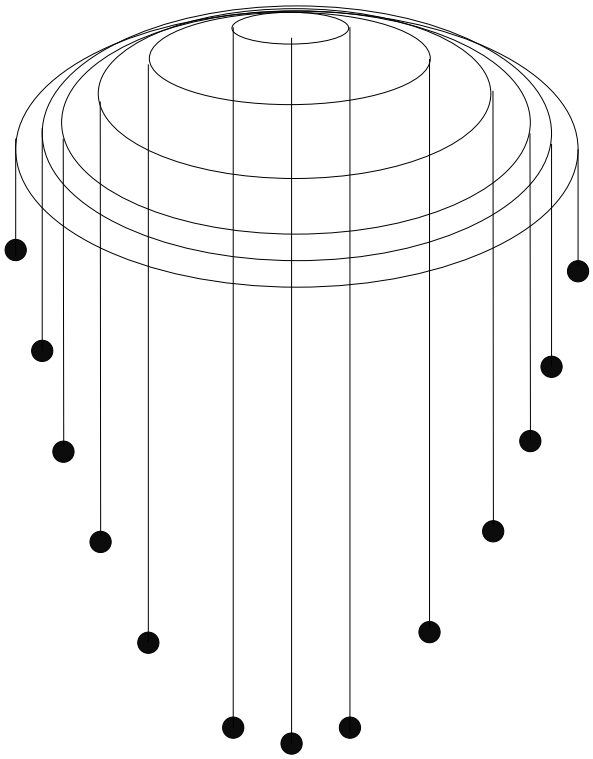

Internet Topology, Figure 6
The internet topology as a jellyfish

5. One-degree nodes: There is a non-trivial percentage (35–45%) of one-degree nodes.

Note that the jellyfish structure seems to hold over time. We observe a 5–10% (in terms of total number of nodes) change of the number of nodes in each layer over 6 years, during which the size of Internet increased dramatically (see Table 4). In the future, with more years of observation,

one would attempt to derive trends in the evolution of the jellyfish.

The Jellyfish Model Tells the Difference

As shown in the previous section, the Internet topology fits the jellyfish profile. However, not every graph can be modeled as a jellyfish. For example, if a tree with N nodes were to be modeled into a jellyfish, the number of shells would be proportional to $O(\log N)$. This does not fit into the jellyfish profile of the Internet, where the number of shells is constant at five despite the rapid growth of the number of nodes. This provides an opportunity to use the jellyfish model as a criterion of the realism of Internet-like graphs.

Siganos et al. [65] use the jellyfish model to test the GLP methodology proposed in [8], and the PLRG approach proposed in [3]. The GLP approach depends on a preferential model. On the other hand, the PLRG generator is based on an interesting theoretical model for scale-free graphs, and takes the degree distribution as a given. In [8], these two generators were compared and it was concluded that the best generator was the GLP. PLRG was shown to fail in capturing properties such as the characteristic path length and the clustering coefficient. However, by using the jellyfish model, Siganos et al. [65] were able to show that GLP does not capture the macro structure by using the jellyfish. Incidentally, PLRG seems to pass the test, although it fails with regard to other properties.

In more detail, two graphs are generated by the GLP model and the PLRG model, respectively. Both of them have similar number of nodes to an Internet topology instance in June, 2000. In Table 5, these graphs are decomposed using the jellyfish model. These results clearly show that the graph generated using the GLP methodology is qualitatively different than the Internet graph. First, the Core of the network is much bigger in GLP (21) compared to the Internet (14). Second, the number of hanging nodes (degree one) in GLP far exceeds the number of shell nodes. The ratio is approximately 70% hanging nodes to 30% shell nodes. In the case of the Internet, this ratio is the opposite. Third, the GLP topology has only up to five layers, with the fifth layer having only three members, while the real Internet has six layers. On the other hand, PLRG seems to maintain similar structure according to the jellyfish model. The only differences between PLRG and the Internet are that the clique is smaller, having only 11 nodes, and that there is a slightly smaller shell-1 and a bigger shell-2.

The Medusa Model

One problem of the jellyfish model is that the identities of the Jellyfish Core are not particularly robust when the completeness of the Internet topology is uncertain. Carmi et al. [10] found that by adding or deleting an edge, the ASes in the Jellyfish Core could change up to 25%, mostly affecting some European ASes. To address the problem, they proposed a model called *medusa*. The medusa model depends on an informative functional decomposition of the Internet ASes called k-pruning, which proceeds as follows:

First, each AS with only one neighbor is removed. The link to that neighbor along with the node is removed as well. As this pruning proceeds, further nodes with one neighbor (or fewer) may be created. They will be removed until there is no longer a one-degree AS in the remaining graph. ASes removed in this way make up what is called *1*-shell. The remaining graph is called *2*-core. Sec-

Internet Topology, Table 5
Distribution of nodes in shell and hang classes

	Instance					
	GLP		Int-06-2000		PLRG	
Layer ID	Nodes	% of Nodes	Nodes	% of Nodes	Nodes	% of Nodes
Core/Shell-0	21	0.2	14	0.176	11	0.13
Hang-0	1885	23.82	798	10.08	565	7.1
Shell-1	1672	21.13	2861	36.16	2346	29.6
Hang-1	3371	42.6	1266	16	1298	16.4
Shell-2	688	8.7	1824	23.05	2305	29.13
Hang-2	221	2.79	662	8.36	525	6.6
Shell-3	3	0.037	390	4.92	325	4.1
Hang-3	3	0.037	74	0.93	125	1.5
Shell-4	0	0	12	0.15	41	0.51
Hang-4	0	0	10	0.12	23	0.29

ond, the pruning process is repeated and is characterized by an index k. For example, when $k = 2$, all nodes with two neighbors will be removed from the 2-core, and the removed nodes in this step is called *2*-shell. The process continues, eliminating any nodes reduced to a degree of two (or fewer) by this pruning, until all nodes remaining have three or more neighbors. The remaining graph is called *3*-core. The process is repeated to identify the 3-shell and 4-core, and so on. The process stops when no further nodes remain. The last nonempty k-core provides a very robust and natural definition of the heart or nucleus of any communications network. Finally, the k-crust is defined as the union of the nodes in the 1- through k-shells, and the links that join them. Thus $k - 1$ crust is the complement of the k-core.

For small values of k, the k-crust consists of many small connected components or clusters. For sufficiently large k, the largest connected cluster of a k-crust consists of a significant fraction of the whole k-crust, while no smaller cluster contains more than a few nodes. The change occurs at a well-defined threshold value of k. There is a significant fraction of the nodes within each large-k crust which is not part of its largest cluster, and remains isolated. Thus, the AS graph can be decomposed into three distinct components as shown in Fig. 7:

1. The nucleus (the innermost k-core),
2. The giant connected component of the last crust, in which only the nucleus is left out,
3. The isolated components of the last crust, nodes forming many small clusters. These connect to the connected component of the last crust only through the nucleus.

These three classes of nodes are quite different in their functional roles within the Internet. The nucleus plays a critical role in BGP routing, since its nodes lie on a large fraction of the paths that connect different ASes. It allows redundancy in path construction, which gives immunity to multiple points of failure. The connected component of the large-k crusts could be an effective substrate on which to develop additional routing capacity for messages that do not need to circle the globe. Finally, the isolated nodes and isolated groups of nodes in the last crust essentially leave all routing up to the nodes in the nucleus of the network. Because all their message traffic passes through the nucleus, even when the destination is relatively close by, they may be contributing unnecessary load to the most heavily used portions of the Internet. The relative size of this component could be a key indicator of the evolution of the topography of the Internet.

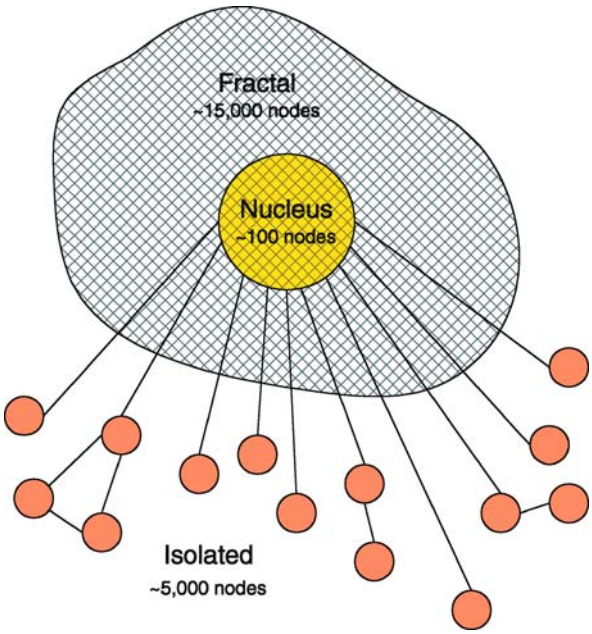

Internet Topology, Figure 7
The Internet topology as a medusa

This model can be visualized as Fig. 7. The Core of the medusa includes the most important nodes that are found in the Core and the first ring of the jellyfish's mantle. The jellyfish has relatively few rings around its Core, while the medusa's mantle is more extended and differentiated. The tendrils hanging from the jellyfish (leaf nodes) descend mostly from the Core, but also from all the other rings, while all the tendrils of the medusa are, by construction, attached to its nucleus.

The Complete Internet Topology

The accuracy of a topological model is important for effectively simulating, analyzing, and designing future protocols [28]. With an accurate Internet AS-level topology, first, one can design and analyze new inter-domain routing protocols, such as HLP [68], that take advantage of the properties of the Internet AS-level topology. Second, one can create more accurate models for simulation purposes [44]. Third, one can analyze phenomena such as the spread of viruses [29,58] more accurately. In addition, the current initiatives of rethinking and redesigning the Internet and its operation from scratch would also benefit from such an accurate Internet topology.

Toward Finding the Complete Internet Topology

Developing an accurate representation of the Internet topology at the AS level remains as a challenge despite the

recent flurry of studies [11,12,18,19,24,47,62,80]. One of the major problems is that, although the majority of ASes are represented completely in a snapshot, the edges among the ASes are not. As seen in an earlier section, each source has its own advantages, but each of them also provides an incomplete, sometimes inaccurate view of the Internet AS topology.

Recently, He et al. [33] present a systematic framework for extracting and synthesizing the AS level topology information from different sources. Instead of simply taking the union of all resources, a careful synthesis and cross-validation is performed. In addition to the sources mentioned above, they also utilize information gathered from IXPs (Internet Exchange Points), which have not received attention in terms of Internet topology discovery, although they play a major role in Internet connectivity.

He et al. identify and validate several properties of the missing AS links: (1) most of the missing AS edges are of the peer-to-peer type, (1) many of the missing AS edges from BGP tables appear in IRR, and (3) most of the missing peer-to-peer AS edges are incident at IXPs

Their work consists of four steps.

First, BGP routing tables are compared. They consider the AS edges derived from multiple BGP routing table dumps [80], and compare them to the Routeview data (OBD). The question to answer is "What is the information that the new BGP tables bring?". Table 6 lists a portion of the collection of BGP table dumps that were collected in May, 2005. One observation, here, is that about 80% of the missing links that do not appear in a single table dump (OBD) but appear in a collection of table dumps (BD) are peer-to-peer type. For example, among 8702 edges in BD but not in OBD, 7183 of them are classified as peer-to-peer type.

Second, He et al. systematically analyze the IRR data and identify topological information that seems trustworthy using the Nemecis tool [63]. They follow a conservative approach, given that the IRR may contain some out-dated and/or erroneous information. They do not accept new edges from the IRR, even after the first processing, unless they are confirmed by traceroutes using public traceroute servers. Overall, they find that the IRR is a good source of missing links. For example, they discover that more than 80% of the new edges found in the extra tables already exist in the IRR [35]. On the other hand, the IRR still has significantly more edges.

Third, He et al. provide a state-of-the-art methodology for identifying the ASes which participate at Internet Exchange Points (IXPs). An IXP is a relatively low cost solution by which an AS can peer with many other peers who are also participants at the same IXP. The exhaustive identification of IXP participants has received limited attention. Most previous work focuses on identifying the existence of IXPs. The finding here is that many of the ASes incident to the peer-to-peer edges missing from the different data sets are IXP participants. Note that even if two ASes peer at the same IXP, that does not necessarily mean there is an AS edge between these two ASes, because this totally depends on peering agreement between these two ASes. Therefore, in order to test whether or not these missing edges are indeed at the IXPs, they proceed to the next step.

Fourth, He et al. use their traceroute tool, RETRO, to verify potential edges from the IRR and IXPs. RETRO is

Internet Topology, Table 6
A collection of BGP table dumps

Route collector or Router server name	# of Nodes	# of Edges	# of edges with type inferred			edges not in OBD	edges not in OBD w/ type		
			total	p-p	p-c		total	p-p	p-c
route -views (OBD)	19843	42643	42570	5551	36766	0	0	0	0
route -views2	19837	41274	41230	4464	36514	1029	1028	835	191
route -views.eqix	19650	34889	34876	1027	33640	674	674	530	143
route -views.linx	19655	37259	37246	3246	33765	2511	2511	2188	319
route -views.isc	19753	36152	36139	1915	34004	784	783	663	118
rrc00.ripe	19770	36479	36465	1641	34605	655	654	543	111
rrc01.ripe	19640	34193	34180	1121	32855	617	617	512	105
rrc03.ripe	19737	39147	39129	3850	35042	3233	3228	2609	616
rrc05.ripe	19765	32676	32659	1122	31324	1095	1091	658	432
rrc07.ripe	19618	32811	31797	1219	30394	804	803	724	79
rrc12.ripe	19628	33841	33827	2024	31606	1611	1610	1417	193
Total (BD)	19950	51345	51259	12734	38265	8702	8689	7183	1499

a tool that collects public traceroute server configurations, sends out traceroute requests, and collects traceroute results dynamically. They confirm the existence of many potential edges identified in the previous steps. The results show that more than 94% of the RETRO-verified AS edges in the IRR indeed go through IXPs. They even discover edges that were not previously seen in either the BGP table dumps or IRR. In total, 300% more peer-to-peer links than those in the conventional BGP table dumps from Routeviews have been validated.

Towards Finding Internet Backup Links

One limitation of the previous method is that it ultimately depends on traceroute to verify the existence of a suspected edge. It is plausible that the suspected edge is a primary link, which means it exists most of the time. If a suspected edge is a backup, and it does not show unless some other links break down, it is unlikely to be seen by traceroute.

Recently, active BGP probing [19] has been proposed for identifying backup AS links. The main idea is to inject false AS path loops for an unused IP block. Since AS path loops are prohibited in inter-domain routing, BGP routers are forced to switch to backup links for this unused IP block. These links can be observed from any route collector, such as Routeviews or RIPE/RIS. This probing technique does not affect normal Internet routing because every change is restricted to the unused IP block.

In more detail, the principle of active BGP probing is the following: An active probing AS announces one of its prefixes with AS-paths including a number of other ASes. These ASes, due to loop detection, will not use or propagate the announcement. Then, if there is any alternative path available, it will show up. To avoid influencing AS-path length, the prohibited ASes are placed in an AS-set at the end of the path. For example, to stop its announcement from being propagated by ASes 1, 2, and 3, an AS (say AS12654) might announce one of its prefixes with an AS-path of [12654 {1,2,3}]. This allows AS 12654 to discover who propagates its announcements, find backup paths, and deduce the policies of other ASes with respect to its prefixes. By properly selecting the "prohibited" AS sets, one may be able to discover all backup links visible to the probing AS.

Note that, this is a promising method for discovering backup links, but so far it has been limited by the small number of probing ASes. This method and the method of discovering missing peer-to-peer links [33] are complementary to each other.

Conclusion and Future Directions

In this paper, we have surveyed current achievement in the research of Internet topology. In summary, power-law distributions constitute an undeniable property of the Internet topology, that has been consistent from 1997 to present, a period in which the size of Internet has grown eight-fold in terms of number of ASes. We examined different data sources for Internet topology research and provide some insight into their pros and cons. A number of important topology generating models and tools, as well as conceptual models of the Internet, were also studied. Finally we introduced current progress on finding the complete Internet topology.

Similar to any other science field, such as Physics, one of the ultimate goals of research is to identify the invariant laws from among seemingly uncontrolled appearance and phenomena. In terms of Internet topology, the power-law appears to be one such law emerging from of the ever-changing and growing Internet. Obviously, there are more to discover. For example, a recent work [46] examines the correlation among node degrees and finds that by replicating node degree correlation, one can essentially reproduce the original topology. Another work [12] argues that one should take into account AS relationships to study the evolution of the Internet topology. The implications of these discoveries remain to be extended.

Bibliography

Primary Literature

1. Achlioptas D, Clauset A, Kempe D, Moore C (2005) On the bias of traceroute sampling, or power-law degree distributions in regular graphs. In: STOC 2005
2. Adamic LA (2000) Zipf, power-laws, and Pareto – a ranking tutorial. http://www.parc.xerox.com/iea/
3. Aiello W, Chung F, Lu L (2000) A random graph model for massive graphs. In: Proceedings of the ACM symposium on theory of computing
4. Albert R, Barabasi A (2000) Topology of complex networks: local events and universality. Phys Rev 85:24
5. Albert R, Jeong H, Barabasi AL (1999) Diameter of the world wide web. Nature 401(6749):130
6. Barabasi A, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509
7. Brite. http://www.cs.bu.edu/brite/
8. Bu T, Towsley D (2002) On distinguishing between internet power law topology generators. In: Proceedings of the 21st Annual Joint Conference of the IEEE Computer and Communications Societies (Infocom IEEE 2002). vol 2. IEEE, pp 638–647
9. Calvert K, Doar M, Zegura EW (1997) Modeling internet topology. Communication IEEE Magazine 35(6):160–163
10. Carmi S, Havlin S, Kirkpatrick S, Shavitt Y, Shir E (2007) A model of internet topology using k-shell decomposition. Proc Natl Acad Sci USA 104:11150–11154

11. Chang H, Govindan R, Jamin S, Shenker S, Willinger W (2004) Towards capturing representative AS-level internet topologies. Comput Netw 44(6):737–755
12. Chang H, Jamin S, Willinger W (2006) To peer or not to peer: modeling the evolution of the internet's topology AS. In: Infocom IEEE 2006
13. Chen Q, Chang H, Govindan R, Jamin S, Shenker S, Willinger W (2002) The Origin of power laws in internet topologies revisited. In: Infocom IEEE 2002
14. Cheswick B, Burch H (1998) Internet mapping project. Wired Magazine 6(12):216–217. See http://cm.bell-labs.com/cm/cs/who/ches/map/index.html
15. Chou H (2000) A note on power-laws of Internet topology, e-print cs. NI/0012019, Harvard University digital library for physics and astronomy
16. Chuang J, Sirbu M (1998) Pricing multicast communications: a cost based approach. In: Proc. INET'98
17. Citeseer. http://citeseer.ist.psu.edu/articles1999.html
18. Cohen R, Raz D (2006) The internet dark matter – on the missing links in the connectivity AS map. In: Infocom IEEE 2006
19. Colitti L, Di G Battista, Patrignani M, Pissonia M, Rimondini M (2006) Investigating prefix propagation through active BGP probing. In: IEEEISCC 2006
20. Crovella M, Bestavros A (1996) Self-similarity in world wide web traffic, evidence and possible causes. In: SIGMETRICS, pp 160–169
21. Crovella M, Lakhina A, Byers JW, Matta I (2003) Sampling biases in ip topology measurements. In: Infocom IEEE 2003
22. Cvetković DM, Boob M, Sachs H (1979) Spectra of Graphs. Academic press, New York
23. di Battista G, Erlebach T, Hall A, Patrignani M, Pizzonia M, Schank T (2007) Computing the types of the relationships between autonomous systems. IEEE/ACM Trans Netw 15(2): 267–280
24. Dimitropoulos X, Krioukov D, Riley G (2005) Revisiting internet AS-level topology Discovery. In: PAM
25. Dimitropoulos X, Krioukov D, Fomenkov M, Huffaker B, Hyun Y, Claffy KC, Riley G (2007) AS relationships: inference and validation. ACM Sigcomm Comput Commun Rev (CCR) 37(1): 29–40
26. Doar M, Leslie I (1993) How bad is naive multicast routing? In: Proc. IEEEINFOCOM, pp 82–89
27. Faloutsos M, Faloutsos P, Faloutsos C (1999) On power-law relationships of the Internet topology. In: ACMSigcomm, pp 251–262
28. Floyd S, Paxson V (2001) Difficulties in simulating the Internet. IEEE Trans Netw 9(4):392–403
29. Ganesh A, Massoulie L, Towsley D (2005) The effect of network topology on the spread of epidemics. In: Infocom IEEE 2005
30. Gao L (2000) On inferring autonomous system relationships in the internet. In: Global IEEE Internet
31. Govindan R, Reddy A (1997) An analysis of internet inter-domain topology and route stability. In: Proc. IEEEINFOCOM, Kobe, Japan, 7–11 April 1997
32. Govindan R, Tangmunarunkit H (2000) Heuristics for internet map discovery. In: Proc. IEEEINFOCOM, Tel Aviv, Israel, March 2000
33. He Y, Siganos G, Faloutsos M, Krishnamurthy S (2007) A systematic framework for unearthing the missing links: measurements and impact. In: USENIXNSDI
34. Inet topology generator. http://topology.eecs.umich.edu/inet/
35. Internet routing registry. http://www.irr.net
36. Jacobson V (1995) Traceroute. Internet measurement tool. Available at ftp://ftp.ee.lbl.gov/traceroute.tar.gz
37. Jaiswal S, Rosenberg A, Towsley D (2004) Comparing the structure of power law graphs and the internet AS graph. In: ICNP 2004
38. Jamin S, Jin C, Jin Y, Raz D, Shavitt Y, Zhang L (2000) On the placement of Internet instrumentation. In: Proc. IEEEINFOCOM, Tel Aviv, Israel, March 2000
39. Jin C, Chen Q, Jamin S (2000) Inet: Internet topology generator. Technical Report UMCSE-TR-433–00, Michigan
40. Jovanovic M (2001) Modeling large-scale peer-to-peer networks and a case study of gnutella. Master thesis, University of Cincinnati
41. Krishnamurthy V, Faloutsos M, Chrobak M, Cui J, Lao L, Percus AG (2007) Sampling large internet topologies for simulation purposes. Comput Netw 51(15):4284–4302
42. Kumar R, Raghavan P, Rajagopalan S, Sivakumar D, ATomkins, Upfal E (2000) The web as a graph. In: Symposium ACM on Principles of Database Systems
43. Leland WE, Taqqu MS, Willinger W, Wilson DV (1994) On the self-similar nature of ethernet traffic. IEEE Trans Netw 2(1):1–15 (earlier version in Sigcomm'93, pp 183–193)
44. Maennel O, Feldmann A (2002) Realistic BGP traffic for test labs. In: Sigcomm ACM
45. Magoni D, Pansiot JJ (2001) Analysis of the autonomous system network topology. ACM Sigcomm Comput Commun Rev (CCR) 31(3):26–37
46. Mahadevan P, Krioukov D, Fall K, Vahdat A (2006) Systematic topology analysis and generation using degree correlations. In: Sigcom ACM
47. Mahadevan P, Krioukov D, Fomenkov M, Huffaker B, Dimitropoulos X, Claffy KC, Vahdat A (2006) The internet AS-level topology: three data sources and one definitive metric. ACM Sigcomm Comput Commun Rev (CCR) 36(1):17–26
48. Mao Z, Rexford J, Wang J, Katz R (2003) Towards an accurate AS-level traceroute tool. In: Sigcomm ACM 2003
49. Mao Z, Johnson D, Rexford J, Wang J, Katz R (2004) Scalable and accurate identification of AS-Level forwarding paths. In: Infocom IEEE 2004
50. Medina A, Matta I, Byers J (2000) On the origin of powerlaws in Internet topologies. ACM Sigcomm Comput Commun Rev CCR 30(2):18–34
51. Medina A, Lakhina A, Matta I, Byers J (2001) Brite: an approach to universal topology generation. In: MASCOTS 2001
52. Mihail M, Papadimitriou CH (2002) On the eigenvalue power law. In: Proceedings of the 6th International Workshop on Randomization and Approximation Techniques, pp 252–262
53. Oliveira R, Zhang B, Zhang L (2007) Observing the evolution of internet as topology. In: Sigcomm ACM 2007
54. Oregon Routeview Project. http://www.routeviews.org
55. Palmer CR, Stefan JG (2000) Generating network topologies that obey powerlaws. In: Proceedings of the Global Internet Symposium, GLOBECOM 2000
56. Pansiot J-J, Grad D (1998) On routes and multicast trees in the Internet. ACM Sigcomm Comput Commun Rev 28(1):41–50
57. Pareto V (1896) Cours d'economic politique. Dronz, Geneva
58. Park K, Lee H (2001) On the effectiveness of route-based packet

filtering for distributed DoS attack prevention in power-law Internets. In: Sigcomm ACM 2001
59. Paxson V, Floyd S (1995) Wide-area traffic: The failure of Poisson modeling. IEEE Trans Netw 3(3):226–244 (earlier version in Sigcomm'94, pp 257–268)
60. Philips G, Shenker S, Tangmunarunkit H (1999) Scaling of multicast trees: Comments on the chuang-sirbu scaling law. In: Sigcomm ACM 1999
61. Ripe Route Information Service. http://www.ripe.net/ris
62. Shavitt Y, Shir E DIMES: let the internet measure itself. ACM Sigcomm Comput Commun Rev (CCR) 35(5):71–74
63. Siganos G, Faloutsos M (2004) Analyzing BGP policies: methodology and tool. In: Infocom IEEE 2004
64. Siganos G, Faloutsos M, Faloutsos P, Faloutsos C (2003) Power-laws of the internet topology. IEEE Trans Netw 1(4):514–524
65. Siganos G, Tauro S, Faloutsos M (2006) Jellyfish: a conceptual model for the as internet topology. J Commun Netw 8(3): 339–350
66. Skitter. http://www.caida.org/tools/measurement/skitter/
67. Spring N, Mahajan R, Wetherall D, Anderson T (2004) Measuring ISP topologies with rocketfuel. IEEE Trans Netw 12(1):2–16
68. Subramanian L, Caesar M, Ee CT, Handley M, Mao M, Shenker S, Stoica I (2005) HLP: next a-generation interdomain routing protocol. In: Sigcomm ACM 2005
69. Tangmunarunkit H, Govindan R, Jamin S, Shenker S, Willinger W (2002) Network topology generators: degree based vs. structural. In: Sigcomm ACM 2002
70. The Cooperative Association for Internet Data Analysis. http://www.caida.org
71. van Mieghem P, Hooghiemstra G, van der Hofstad R (2001) On the efficiency of multicast. IEEE Trans Netw 9(6):719–732
72. Wang F, Gao L (2003) Inferring and characterizing internet routing policies. In: ACMIMW 2003
73. Waxman BM (1988) Routing of multipoint connections. IEEE J Sel Areas Commun 6(9):1617–1622
74. Wei L, Estrin D (1994) The trade-offs of multicast trees and algorithms. In: Proceedings of the International Conference on Computer Communications and Networks
75. Weibull Distribution (2003) NIST/SEMATECH e-handbook of statistical methods. http://www.itl.nist.gov/div898/handbook/
76. Wong T, Katz R (2000) An analysis of multicast forwarding state scalability. In: Proceedings of the International Conference on Network Protocols
77. Xia J, Gao L (2004) On the evaluation of as relationship inferences. In: Globecom IEEE 2004
78. Yook SH, Jeong H, Barabasi A (2002) Modeling the internet's large-scale topology. Proc Natl Acad Sci USA 99(21):13382–13386
79. Zegura EW, Calvert KL, Donahoo MJ (1997) A quantitative comparison of graph-based models for internetworks. IEEE Trans Netw 5(6):770–783
80. Zhang B, Liu R, Massey D, Zhang L (2005) Collecting the internet AS-level topology. ACM Sigcomm Comput Commun Rev (CCR) 2005
81. Zipf GK (1949) Human behavior and principle of least effort: An introduction to human ecology. Addison Wesley, Cambridge

Books and Reviews

Newman M, Barabasi A-L, Watts DJ (2006) The structure and dynamics of networks. Princeton University Press, Princeton

Invasion Percolation

Mark Knackstedt[1], Lincoln Paterson[2]
[1] Department of Applied Maths, RSPhysSE, Australian National University, Canberra, Australia
[2] CSIRO Petroleum, Clayton, Australia

Article Outline

Glossary
Definition of the Subject
Introduction
Classical Definitions, Algorithms and Results
Modifications to Invasion Percolation
Effect of Pore Scale Structure on IP
Effect of Correlated Heterogeneity on IP
Future Directions
Bibliography

Glossary

B	Bond number
D_b	Backbone
D_f	Fractal dimension
$D_{\min}$	Fractal dimension of minimum path
fBm	Fractional Brownian motion
fLm	Fractional Lévy motion
H	Hurst exponent
IP	Invasion percolation
NTIP	Non-trapping invasion percolation
OP	Ordinary percolation
p_c	Ordinary percolation threshold
S_r	Residual saturation
TIP	IP with trapping
ξ_w	Correlation length
ν	Percolation correlation length exponent
ϕ	Porosity
g	Field gradient
Z	Coordination number
σ	Standard deviation

Terms

Defender Fluid initially within pore space.

Invader Second fluid injected to displace defending fluid (defender).

Drainage Displacement of a wetting fluid by a non-wetting fluid.

Definition of the Subject

Invasion percolation is a simple dynamic process describing the slow displacement of one fluid by another in

a porous material. This is a common phenomena with many important applications; these include the penetration of nonaqueous polluting liquids into soil, the penetration of air into porous media such as soil, concrete, wood and ceramic powder during drying and the displacement of water from soil and rocks by gases generated by buried waste. A final important example, which is the primary focus of this review, is the accumulation during initial migration and the subsequent recovery and production of hydrocarbon reservoirs. Experiments on idealized systems have shown that the simple invasion percolation model provides a very realistic description of the slow fluid-fluid displacement processes associated with these important applications. Experiments and simulations of invasion percolation have been extended to consider the role of gravity, wettability and the complex nature of real porous materials on the dynamic immiscible displacement processes.

Introduction

Invasion percolation is a dynamic process that was proposed to describe the slow immiscible displacement of one fluid by another in a porous medium. Many porous media can be represented a network of pores (sites) connected via throats (bonds) [9,33,71]. In this representation all the pores and throats are initially filled with *defending* fluid. When a second fluid is injected very slowly into the porous medium such that the capillary forces dominate the viscous forces, the dynamics is controlled by the size of the local pore or throat. In a drainage process (the displacement of a wetting fluid by a nonwetting fluid), capillary forces are strongest at the narrowest places in the medium; in the pore throats. A drainage process is therefore represented as a series of discrete jumps across throats in which the non-wetting fluid displaces the wetting fluid via the *largest throat* (offering the least resistance to displacement). This is the version of the model originally considered by Chandler et al. [9], equivalent to bond invasion percolation. Wilkinson and Willemsen [71], who were the first to use the term *invasion percolation*, considered the process of imbibition (a non-wetting fluid being displaced by a wetting fluid) at a constant but infinitesimal (capillary dominated) flow rate. In this scenario the capillary forces are again strong in the throats, so the wetting fluid invades the throat quickly, but slows when entering the larger pores. This motion can be described by a series of discrete jumps in which at each time step the wetting fluid advances through the *smallest available pore*. This is site invasion percolation. In the absence of trapping, the bond version can be reduced to the site version through bond to site transformations. However, when trapping is introduced where regions of defending fluid are incompressible and cannot escape, the transformation becomes a correlated site-bond problem [71].

Most implementations of invasion percolation have been on a regular lattice. Sites or bonds representing pores and throats are usually assigned uniformly distributed random numbers as it is the sequencing of the events that is important, independent of the distribution. For drainage, invasion percolation is generally considered to be a good model. Although Wilkinson and Willemsen [71] used site invasion percolation for imbibition, later work by Blunt and Scher [7] has this as a special case where different alternative modes of wetting invasion are possible.

Classical Definitions, Algorithms and Results

Invasion vs. Ordinary Percolation

There are distinct differences between invasion percolation (IP) and ordinary percolation (OP). Invasion percolation starts with a well defined interface (inlet) and displaces the defending phase in a systematic way until spanning the system. In this way the concepts of history and the sequence of invading pores are naturally built into the model. The cluster generated in IP always spans a region between the injection face and the outlet face; there is no analogue to the percolation occupation probability p and there is only a single invasion cluster (no finite disconnected clusters). Figure 1 shows the spanning cluster at breakthrough of the invading phase for invasion percolation in two dimensions.

The implementation of trapping in the defending fluid introduces more differences. Monte-Carlo simulations of invasion percolation with trapping (TIP) in two dimensions [71] found that the fractal dimension of the sample

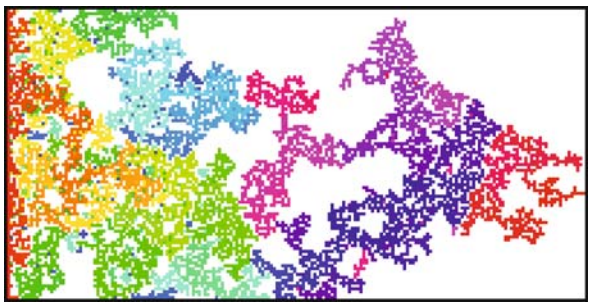

Invasion Percolation, Figure 1
Invasion percolation on 2D lattice. The invader (*colored*) enters from sites on the *left hand edge* and the defender exits from the *right hand edge*. Different colors indicate sites added within different time intervals

spanning cluster was $D = 1.82$, significantly smaller than $D = 1.896$ for OP. Simulation of non-trapping IP (NTIP) gave similar values to OP. In three dimensions no significant difference in the IP and OP models was originally observed. The difference in the 2D values was assumed to be associated with trapping and the effect of trapping thought to be negligible in 3D. The question of the universality class of IP was not conclusively established at this time.

There was extensive experimental evidence in support of the IP model of two phase flow in porous media. Lenormand and Zarcone [35] performed air drainage of oil in a transparent two-dimensional etched network of pores. Analysis of the fractal dimension of the invading phase gave $D = 1.82$ consistent with 2D simulations of TIP. In a further range of drainage experiments with a variety of wetting and non-wetting fluids including oil, air, water and different sucrose solutions Lenormand [34] showed that the results were completely consistent with an IP description of the phenomena. Stokes et al. [60] considered displacement patterns in a cell packed with unconsolidated glass beads. They found that under drainage conditions the resultant fluid displacement patterns were consistent with IP. Chen and Wada [10] used a technique of index matching fluids to visualize the fluid distributions of a quartz bead pack; again the fluid distribution patterns are reminiscent of IP. While the main application of IP has been to the description of the evolution of the interface between two immiscible fluids, IP also has applications to other problems including the characterization of optimal paths and domain walls in strongly disordered media [11,51] minimum spanning trees [12] and the simulation of the Ising model at the critical temperature [17]. Moreover, IP is one of the simplest parameter-free models which exhibits self-organized critically [58].

Methods and Algorithms

An elegant and fast algorithm for invasion percolation can be found in the book by Richard Gaylord and Paul Wellin [20], together with a detailed explanation. This algorithm, written in Mathematica, is included for convenience in Algorithms 1 und 2. As listed, this code starts on a square lattice with a single seed at 0,0 and maintains a list of potential invasion sites on the boundary as the cluster grows. It is straightforward to modify this algorithm to include boundaries and a different initial cluster.

Trapping IP results were for many years limited to small lattice sizes due to the time needed to search for the trapped regions at each time step. In the conventional algorithms the search for the trapped regions was done after every invasion event using a Hoshen–Kopelman algorithm [24,59], which traverses the whole lattice, labels all the connected regions, and then only those sites (bonds) that are connected to the outlet face are considered as potential invasion sites (bonds). A second sweep of the lattice is then done to determine which of the potential sites is to be invaded in the next time step. Thus each invasion event required $O(N^2)$ calculations and limited TIP simulations to small lattice sizes.

Sheppard et al. [53] developed more efficient algorithms for generating TIP simulations. They noted, firstly, after each invasion event only a small local change is made in the interface; implementing the global Hoshen–Kopelman search is unnecessary. Secondly, it is wasteful to traverse the entire lattice at each time step to find the most favorable site (bond) on the interface since the interface is largely static. The first problem is tackled by searching the neighbors of each newly invaded site (bond) to check for trapping. This is ruled out in almost all instances. If trapping is possible, then several simultaneous breadth first "forest-fire" searches are used to update the cluster labeling as necessary [4]. This restricts the changes to the most local region possible. Since each site (bond) can be invaded or trapped at most once during an invasion, this part of the algorithm scales as $O(N)$. The second problem (identifying the sites for invasion) was solved by storing the sites (bonds) on the fluid-fluid interface in a list, sorted according to the capillary pressure threshold (or size) needed to invade them. This list is implemented via a balanced binary search tree, so that insertion and deletion operations on the list can be performed in $\log(n)$ time, where n is the list size. The sites (bonds) that are designated as trapped using the procedures described above are removed from the invasion list. Each site (bond) is added and removed from the interface list at most once, limiting the cost of this part of the algorithm to $O[N \log(n)]$. Thus, the execution time for N sites (bonds) is dominated (for large N) by list manipulation and scales at most as $O[N \log(N)]$. This allowed multiple simulations of TIP at scales of 4000^2 in 2D and 500^3 in 3D.

Universality Class of TIP

TIP describes waterflooding processes in secondary oil recovery. If differences in the topology of the transport pathways for the bond based TIP (drainage) and site-based TIP (imbibition) exist, this has a profound effect on the conductivity of the invading phase at the breakthrough point where a sample spanning cluster first forms (water breaks through during recovery). Differences have been noted in experimental measurements on rocks under different wettability conditions. Probing this question required simu-

lating TIP processes on very large lattices. This allows one to obtain precise estimates for the fractal dimensions of the sample spanning cluster, backbone and minimal path.

In two dimensions it was found that these fractal dimensions are non-universal and vary with the coordination number Z of the lattices [53,54]. The fractal dimension of the sample spanning cluster (SSC) of lattices with low Z exhibited the standard value of $D_f = 1.82$, the fractal dimension crosses over to the value given by OP for large Z (> 6). Values for the triangular lattice ($Z = 6$) seemed to give an intermediate value close to the prediction for OP. The same trends were seen with the value of the backbone and minimal path dimensions (Table 1). These results showed that the scaling properties of TIP in 2D are lattice dependent and hence non-universal.

It was initially thought that site- and bond-based IP were identical; results on large lattices showed distinct differences in the scaling properties of site-based and bond-based TIP. Results in 3D showed that site and bond TIP were in two distinct universality classes [54]. Site TIP had the same scaling behavior as NTIP and OP. A second universality class was observed for bond TIP.

Overall the results show that while ordinary percolation, a static process, is described by a unique universality class, TIP, a dynamic phenomena does not possess a unique universality class. The difference in the topology of site-based and bond-based TIP can assist in the interpretation of experimental measurements of waterflooded rock under different wettability conditions; for an oil wet rock the conductivity of the water channels can be several orders of magnitude smaller at breakthrough than for a comparable water wet rock; this difference is consistent with the different topology of the flow paths for site vs bond based TIP.

Invasion Percolation, Table 1

The most accurate estimates of various fractal dimensions for IP in 2D and 3D, and their comparison with those of random percolation (OP) [52]

Model	D_f	D_{min}	D_b
2D			
NTIP	1.8959	1.1293	1.6422
Site TIP ($Z < 6$)	1.825	1.203	1.217
Site TIP ($Z = 6$)	1.890	1.132	1.616
Site TIP ($Z > 6$)	1.895	1.136	1.642
Bond TIP ($Z < 6$)	1.822	1.214	1.214
Bond TIP ($Z = 6$)	1.823	1.215	1.215
Bond TIP ($Z > 6$)	1.895	1.221	1.221
OP	1.895	1.1307	1.6432
3D			
Site NTIP	2.524	1.3697	1.868
Site TIP	2.524	1.3697	1.861
Bond TIP	2.524	1.458	1.458
OP	2.524	1.374	1.87

Trapping Thresholds for TIP

As the TIP cluster is a fractal at breakthrough, the initially spanning cluster has essentially zero density. However a percolation threshold can be defined for TIP; this occurs beyond the percolation point, and is associated with the disconnection of the defending phase. This second percolation threshold is reached when the defender phase no longer percolates through the system and consists only of isolated clusters. At this point the TIP process ends. In two dimensions this second threshold corresponds to the invading phase breaking through. In three dimensions the trapping does not occur until the invader has occupied a significant fraction of the pore space. In a porous rock with two immiscible fluids this second threshold is associated with the residual saturation of the defending phase; no more defending fluid can be displaced from the sample without increasing the flow rate and introducing viscous forces into the displacement process. The value of the threshold is therefore of important practical interest. Numerical estimates of the trapping threshold for TIP on a cubic lattice were given by [71] as 0.66. Percolation and trapping thresholds have been related to the mean coordination number of the lattice [23]. Simulations of ordinary percolation on a range of simple lattices has led to the approximate relationships that $p_c^{\text{bond}} = 1.5/Z$ and $p_c^{\text{site}} = 2/Z$ [52]. Galam and Mauger [19] expanded on this idea and proposed a universal formula which in 3D is given by $p_c = p_0(Z-1)^{-a}$ where p_0 and a are constants which depend on the type of percolation considered. Percolation threshold for OP and OP with trapping followed this relationship [47]. In a set of simulations on lattices of varying coordination number it was shown that the formula of [19] matches the TIP threshold of the lattices with $Z > 6$. However values of the threshold for coordination numbers $Z < 5$ [50] diverged from the prediction of [19]. The results are significant because of the application to the prediction of residual phase saturations in multiphase flow through porous media.

Modifications to Invasion Percolation

The invasion percolation model was originally developed to model two-phase displacements in porous media, with obvious application to oil recovery. However, even the early researchers in the field were aware that to be of practical significance and have utility the invasion percolation

model would need to be extended to include gravity, viscosity, and other effects.

Introduction of Gravity

Gravity was first introduced into the invasion percolation algorithm by Wilkinson [69] through the application of a simple linear weighting on the invasion thresholds in the direction of buoyancy.

An important application of invasion percolation with buoyancy is to the secondary migration of oil. Secondary migration is the slow process occurring over geological timescales where oil migrates from the source rocks where it is formed into structural or stratigraphic traps. It is in these traps where oil is found and produced, so a knowledge of secondary migration can be a key input into oil exploration. Secondary migration is a very slow process dominated by capillary and buoyancy forces, so invasion percolation is well suited to modeling this process as viscous forces can be safely ignored.

In a series of paper, researchers at the University of Oslo have explored invasion percolation as a model for secondary migration [42,43,64,66,67]. In particular they have studied the structure formed when the non-wetting fluid disintegrated into fragments [41]. They performed extensive two- and three-dimensional computer simulations and found that with a destabilizing external field on invasion percolation that displacement patterns were dominated by the growth of a single branch. This branch could be described in terms of a connected string of blobs of size ξ_w, which form a directed random walk along the direction of the field. On length scales smaller than ξ_w, the displacement patterns had the structure of invasion percolation clusters without a destabilizing field. They found that dependence of the correlation length ξ_w on the magnitude of the field gradient g is given by $\xi_w \sim |g|^{-\nu/(\nu+1)}$ (where ν is the ordinary percolation correlation length exponent) in accord with the theoretical arguments of Wilkinson [69,70].

A fundamental difference between invasion percolation and ordinary percolation is that invasion percolation generates a spanning cluster for lesser numbers of invaded sites, and this fraction of invaded sites tends toward zero as the network size increases. This is consistent with field observations; it can be very hard to detect secondary migration pathways that are only a thin filament occupying only an infinitesimal proportion of the exploration volume. Experiments conducted by Hirsch and Thompson on sandstone samples of different sizes found saturations consistent with these predictions from invasion percolation. Heterogeneity does however create local pools along the migration path, so there are occasions where parts of the path can be detected.

Pore-network simulations and concepts from invasion percolation in a gradient have been used to study the effect of gravity on the critical gas saturation in a porous medium. Critical gas saturation denotes the volume fraction of the gas phase at the onset of bulk gas flow during the depressurization of a supersaturated liquid in a porous medium. Tsimpanogiannis and Yortsos [63] found that the critical gas saturation approaches two plateau values at low and high Bond numbers B. In the intermediate region it scales as a power law of B, which for a 2D lattice is $B^{-0.91}$.

Introduction of Viscous Forces

Invasion percolation corresponds to very slow displacements where viscous forces can be ignored. However, commercial extraction of oil is often at rates where viscous forces cannot be ignored hence there is incentive to expand the model to incorporate viscous forces. Wilkinson [69] was the first to suggest how percolation may be extended to include viscous forces through the use of a mean field description of the fluid flow. His ideas were developed further by Xu et al. [72] in the limits of high and low viscosity ratios. For small viscosity ratios, they determined that displacement could be modeled by a form of gradient percolation in a stabilizing gradient, involving a particular percolation probability profile. In the opposite case, the displacement can be described by gradient percolation in a destabilizing gradient and this leads to capillary-viscous fingering.

Introduction of Wetting

Invasion percolation was originally formulated with very simple concepts of solid surfaces with uniformly wetting and non-wetting fluids. Wetting refers to the fluid that preferentially contacts the solid surface. Experiments in sedimentary rocks over time necessarily led to the development of models involving additional mechanism such as snap-off and film flow, extending basic invasion and ordinary percolation models [7]. Further work on wettability showed that many rocks exhibit mixed-wettability, with the contact angle between the fluids varying with the diverse rock mineralogy [6]. This has led to more complicated network flow models that increasingly depart from the simple original invasion percolation model [8].

Effect of Pore Scale Structure on IP

In this section we consider the application of IP to the description of multiphase fluid flow in porous media at the

pore scale. In order to produce realistic descriptions of the multiphase flow behavior of real porous materials requires one to obtain an accurate description of the morphology of the porous material. In this section we consider the importance of sizes and shapes of pores and throats and the presence of pore to throat correlations to IP. Equally important are the connection patterns of the pores and throats or the topology of the real porous network. Network topology is defined by parameters which include mean connectivity, coordination number distributions and numbers of isolated clusters. In this section we describe the topological and geometric properties of a range of porous materials obtained from 3D images and compare them to classical lattices. Calculation of IP properties illustrates the importance of accurately reproducing the topology of real porous samples.

Measurement of Pore Scale Topology and Geometry

Direct Analysis of Pore Geometry and Topology Most studies of IP in 3D before 1995 were limited to regular cubic lattices. In recent years there has been mounting evidence that real rock topologies exhibit much lower coordination numbers. Ioannidis et al. [25] measured the average coordination number $\bar{Z}$ from serial sections of a sandstone core and found $\bar{Z} = 4.1$. Bakke and Øren [5,45] developed a process-based reconstruction procedure which incorporates grain size distribution and other petrographical data obtained from 2D thin sections to build network models that are analogues of real sandstones. The average coordination number $\bar{Z}$ of the resultant pore networks was significantly less than $Z = 6$. Recently, direct measurement of a 3D pore structure has become more readily available via synchrotron X-ray computed microtomography (micro-CT) [15,16,57], conventional micro X-ray CT [2] and laser confocal scanning microscopy [18]. Coupled with skeletonization algorithms [5,36,45,62] one can extract microstructural parameters for the direct input into network models. Lindquist et al. [37] originally made measurement of the pore coordination number in equivalent network models derived from a suite of Fontainebleau sandstone samples with porosity varying from 7.5% to 22%. The average coordination number varied from $\bar{Z} = 3.37$ at $\phi = 7.5\%$ to $\bar{Z} = 3.75$ at $\phi = 22\%$. Moreover, the coordination number of the pores within each sample exhibited a broad distribution; the majority of pores were 3-connected, however some pores with $Z > 20$ were observed.

Extensive topological analysis of rock core material [55] from a range of geological settings (sandstones, carbonates) has shown that the mean coordination number Z is usually $Z < 6$. In Fig. 2a–c we show a slice of a sand pack, the partitioning of the subvolumes into over 300,000 individual pores and throats and the resultant pore network structure. The coordination number or the pore space is found to be $Z = 5.4$. The full coordination number distribution is shown in Fig. 2d. Some pores exhibit coordination numbers > 20. From the partitioning one can also directly define the pore size, throat or constriction size, shapes and tortuosity of the pore structure in 3D. Examples of distributions of these properties are given in Fig. 2e–f.

In Fig. 3 and Table 2 we illustrate and quantify the pore structure and resultant pore topologies of three other granular systems and rocks. The samples include a consolidated sandstone, an idealized multiscale material and an outcrop limestone. In Fig. 3 we show the images of network subsets. The networks in Fig. 3 contain a small subset ($< 10\%$) of the full image volume obtained from the tomogram. However the information obtained is very rich in detail. The difference in the network structure for these four samples is visually dramatic. The more complex samples (bidisperse and limestone) can exhibit pores of extremely high coordination ($Z > 100$) and large volume weighted coordination numbers $Z > 30$ reflecting the high connectivity of the larger pores. Clearly the use of a simple cubic lattice gives a poor topological description of these systems.

Implication of Pore Topology to IP Properties As discussed in Sect. "Trapping Thresholds for TIP", percolation and trapping thresholds have been related to the mean coordination number of the lattice [23]. Simulations of ordinary percolation on a range of simple lattices has led to the approximate relationships that $p_c^{\text{bond}} = 1.5/Z$ and $p_c^{\text{site}} = 2/Z$ [52] and $p_c = p_0(Z-1)^{-a}$. Results on regular lattices [50] showed that these predictions however did not give good correlations for $Z < 6$. As extensive topological analysis of rock core material from a range of ge-

Invasion Percolation, Table 2

Details of the network structure for the three samples shown in Fig. 3. Z_m gives the mean coordination number, Z_w the volume weighted mean and $Z_{\max}$ the maximal pore coordination number. The mean $(\frac{R_p}{R_t})_m$ and volume weighted $(\frac{R_p}{R_t})_w$ pore to throat aspect ratios are also given

Sample	Z_m	Z_w	$Z_{\max}$	$(\frac{R_p}{R_t})_m$	$(\frac{R_p}{R_t})_w$
Bead Pack	5.1	7.2	19	2.3	3.5
Sandstone	5.4	9.0	49	2.9	4.0
BiDisperse	3.6	31.4	227	3.0	30.3
Limestone	5.6	30.4	372	6.5	20.3

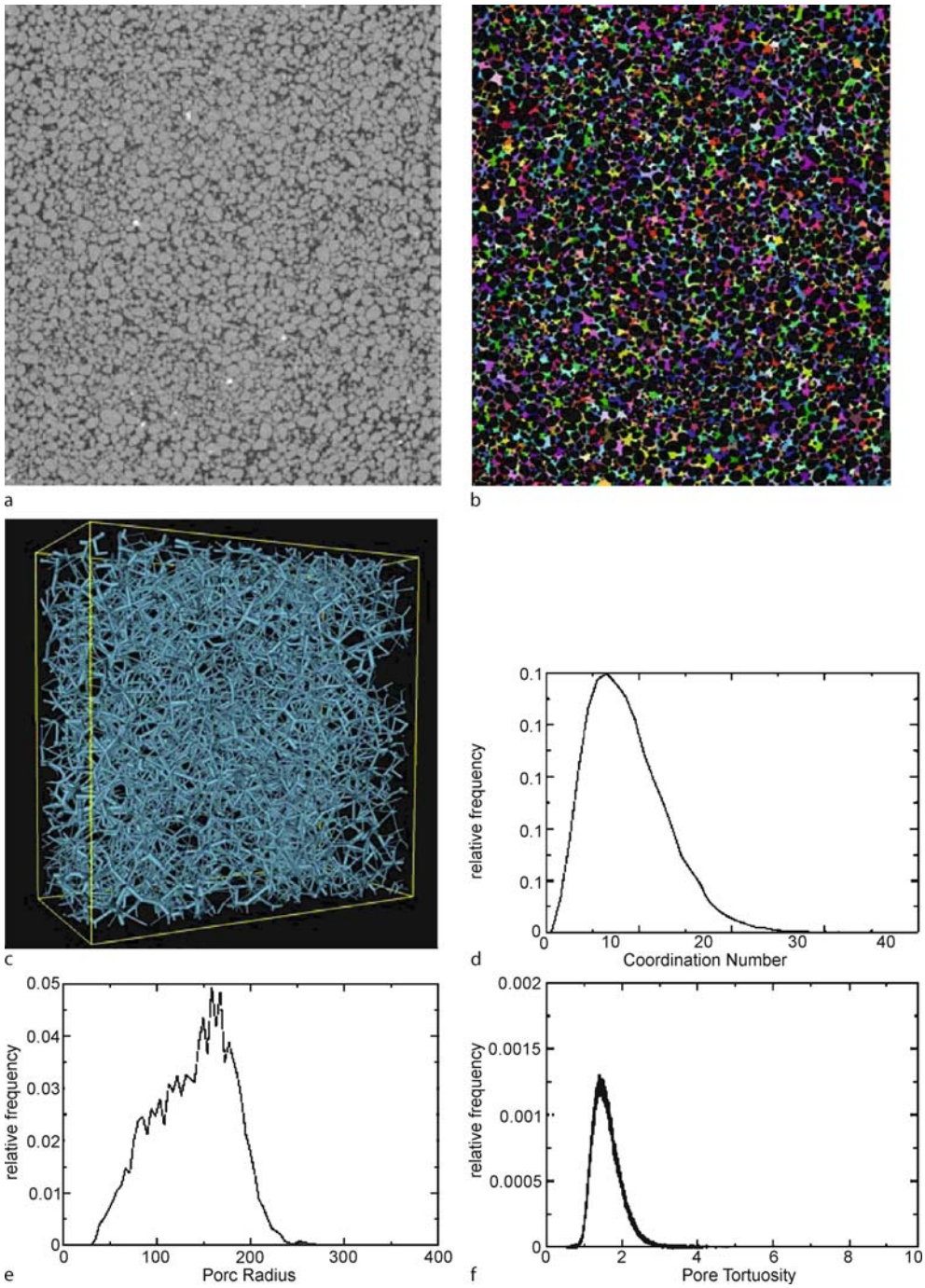

Invasion Percolation, Figure 2

a 2D slice of a 3D image of a sand pack along with the **b** results of a pore partitioning of the sample. **c** shows a small subset of the resultant network of pores and throats in 3D and **d** the coordination number distribution. **e** and **f** give geometric information including the pore size distribution and the pore tortuosity

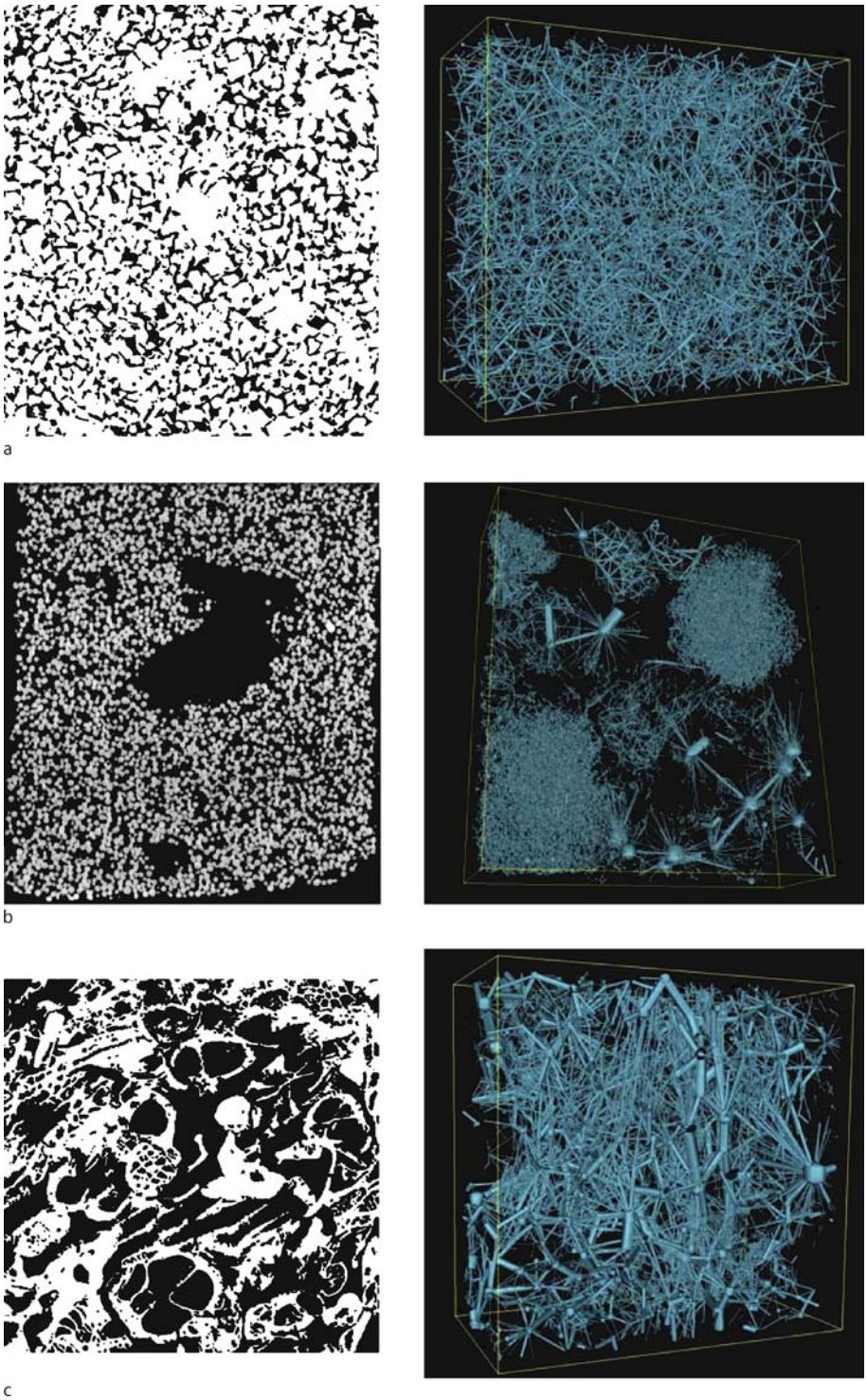

Invasion Percolation, Figure 3

Images (*left*) and networks (*right*) of four samples. a Castlegate sandstone, b bidisperse sample and c Mt. Gambier limestone. The size of the pores and throats reflects their actual size in the partitioning of the 3D image. The variation in structure across the 3 samples is dramatic

ological settings has shown that the mean coordination number Z is often $Z < 6$ and this is of importance in the prediction of residual phase saturations in porous rocks. Moreover, given the broad distributions in Z observed in real rock materials, is the mean coordination number sufficient to predict the trapping threshold on real pore network topologies? These questions have been explored recently by a number of researchers.

Mean Coordination Number Suding and Ziff [61] first showed the importance of lattice topology other than mean coordination number on percolation thresholds of OP in two dimensions. They considered 11 Archimedian lattices with identical mean coordination numbers. They showed that the mean coordination number alone was not sufficient to predict p_c. Predictions for p_c on these lattices varied from 0.55–0.80. Stochastic networks of rock material had previously [25] been generated to solely match the coordination number of rocks by diluting sites on a regular cubic lattice until the remaining connected component had the desired mean coordination number. Analysis of the trapping threshold on a diluted cubic network and Fontainebleau sandstone however gave very different thresholds. There was clearly a need to compare networks in 3D with precisely controlled topology. Sheppard developed a 4 part algorithm to generate a network with a specified coordination number distribution [56]. Use of this algorithm allowed one to generate stochastic networks with matching of mean coordination and the coordination number distribution. Sok et al. [56] generated an ordered (diamond) network ($Z = 4$) and a range of stochastic networks with an identical mean coordination number $Z = 4$ and different standard deviations in Z; $\sigma(Z) = 0.0001, 1.0, 2.0$. The trapping threshold for TIP on the lattices differed strongly from the diamond network (see Table 3). The bond TIP threshold for the diamond network and the stochastic network with $\sigma(Z) = 0.001$ varied strongly, while the site TIP threshold varied. Variation of the coordination number distribution led to strong differences in the resultant TIP thresholds. These results showed that thresholds cannot be correlated solely to dimension and coordination number of the network.

Invasion Percolation, Table 3
Threshold values for the four coordinated lattices

Regular Lattice	0.290	0.417
Stochastic: $\sigma = 0.001$	0.298	0.426
Stochastic: $\sigma = 1.0$	0.328	0.429
Stochastic: $\sigma = 2.0$	0.375	0.437

Coordination Number Distribution and Higher Order Measures Rock networks display a broad distribution of coordination number and the presence of long range topological bonds. Recent results have shown [32,56] that honoring the full coordination number distribution does lead to better prediction of the trapping IP thresholds than matching the coordination number alone. However differences have still been noted. This result shows that one might require a complete description of the network topology to accurately predict trapping thresholds. One study [56] introduced higher order topological quantities; ring size, coordination sequence and topological bond length as measures of relevance to the prediction of trapping thresholds. Comparison of these topological properties between rock networks and stochastic networks showed clear differences. A second study [32] showed that rock networks exhibited a large proportion of topological long bonds; bonds connecting two pores which are not nearest neighbors. A procedure to add topological long bonds was then developed. The introduction of topological long bonds led to a better match to the higher order measures and a marked effect on the prediction of trapping thresholds. Overall the results illustrate the importance of network topology on the accurate prediction of trapping phase thresholds.

Influence of Pore Geometry on IP Properties The resultant percolation saturations in IP are strongly dependent on the geometric properties of rock, in particular the distributions of sizes and shapes of pores [30,31,71]. For an uncorrelated system in which random numbers are assigned to each site/bond, we would expect for IP that most small/large random numbers are accepted as part of the invading cluster, while the large/small numbers remain part of the defending phase at the final saturation. Unlike OP, where one observes a sharp acceptance profile, in TIP there is a transition region [71]. From the acceptance profile and the distribution of pore volumes one can estimate the trapped saturation for various pore size distributions.

Influence of Pore Scale Geometric Correlations on IP Properties One can further illustrate the effect of local pore geometry/correlation in real rock structures by randomly rearranging pore volumes on the rock network structure. By doing this one is preserving the topology of the network; only considering the effect of randomizing the pore bodies. The effect of this rearrangement was illustrated for a set of Fontainebleau sandstone samples in [56]. The site-based trapping thresholds were STIP < 0.25 for the actual rock samples, while the rearranged sample exhibited trapping thresholds STIP > 0.45. These differences are con-

sistent with the presence of local correlated heterogeneity in the pore network at the smallest scales. A method to measure and introduce the spatial heterogeneity observed in the rock structure at a pore scale is lacking; at present the best representations of rock microstructure would be based on true 3D realizations; either from reconstruction method [5] or from direct measurement of 3D structure via X-ray computed tomography [15,16,57].

Effect of Correlated Heterogeneity on IP

In this section we consider the application of IP to the description of multiphase fluid flow in porous media at scales larger than individual pores. Early work with network models and invasion percolation concentrated on macroscopically homogeneous porous media built using independently generated random numbers to assign pore and throat sizes. Correspondingly most of the testing of the models was conducted on laboratory porous media made from relatively uniform glass spheres or from uniformly etched glass plates. However it became apparent when strategies for improved oil recovery were attempted in the field that recoveries did not match expectations, largely because of heterogeneity at all scales. This led to studies of heterogeneity and evidence that long-range correlations in properties down to the centimeter [46] and the pore scale [26,73] exist in many, if not most, porous sedimentary rocks.

Measurement of Heterogeneity

The earliest work that considered correlated substrates used two-dimensional multifractal lattices as a means of incorporating heterogeneity [39,40].

Subsequently there have been many studies that have have improved understanding of the long-range correlations in rock properties that exist from the pore scale to the kilometer scale. To describe these correlations, fractional Brownian motion (fBm) is one of the models that has been used as a model for the underlying reservoir heterogeneity [44]. Fractional Brownian motion is straightforward to generate, hence it has been an obvious candidate for studies of percolation in correlated property maps [3,13,14,48].

Invasion percolation with trapping on substrates with a spatially correlated threshold distribution resulting from the mapping of a self-affine surface has also been studied by Wagner et al. [67].

Analysis of sandstones [27] on rock samples of a few centimeters in extent suggests that correlated heterogeneity exists down to the pore scale and that at this scale correlations persist at scales up to several pore lengths. For rock samples at these scales a more appropriate model of correlated heterogeneity is one which introduces a cutoff length scale below which correlations persist and above which the system behaves like a random material.

Effect of Heterogeneity on IP

Fractal Dimension The effect of the correlated heterogeneity measured in naturally occurring sedimentary rocks on IP was first studied in 2D [40,48,67]; the fluid displacement patterns for lattices with long range correlations based on multifractal, fBm and fLm models were compared to uncorrelated networks. The cluster fractal dimension for NTIP was found to be compact for almost all systems exhibiting correlated heterogeneity while TIP exhibited fractal saturation patterns for multifractal systems and fBm models with a Hurst exponent $H < 0.5$. In all studies with fBm model correlations the cluster fractal dimension was observed to increase with $H < 0.5$; for $H > 0.5$ all clusters were compact. This result was verified later [28] on large lattices for fBm correlated lattices. Similar results were obtained on 3D lattices [28]. Other fractal dimensions were also studied; for $H < 0.5$ the fractal dimensions of the hull, minimal path, backbone and invasion front are all fractal with dimensions that depend on H [3,28,40].

The effect of correlated heterogeneity with a finite cutoff length scale for the extent of correlations (correlated heterogeneity at a smaller scale and uncorrelated above some cutoff length scale) has also been considered [28]. In these cases clusters were fractal at length scale above the cutoff length with fractal dimensions that are the same as those in normal IP models. For smaller length scales the clusters' structures are similar to those observed for fully correlated lattices.

Thresholds Studies in two dimensions have considered the effect of correlated heterogeneity on breakthrough saturations [3,14,40] – results on finite lattices showed that correlation has a significant effect on the percolation threshold and that the threshold is no longer unique but depends on the spanning rule employed [38]. Residual saturations S_r in 3D systems found that the introduction of correlation leads to a large reduction in residual saturation with increasing H [29]. For uncorrelated cubic lattices $S_r = 0.34$, while $S_r = 0.25, 0.22$ and 0.18 for fBm correlated lattices with $H = 0.2, 0.5$ and 0.8 respectively. Correlations with a finite cutoff also led to changes in S_r; in particular the residual saturation was found to exhibit a minimal value for finite cutoff lengths; the increase in saturations at large cutoff length scales was due to the possibility of trapping very large regions of the defend-

ing fluid at larger cutoffs. Small scale correlations therefore have a profound effect on resultant residuals, even at large scales. Analysis of the scaling behavior of the variance of residual saturations shows that the measurements of S_r must be made on samples that are at least ten times larger than the extent of correlated heterogeneity. Measurements on porous rock samples with extensive correlations would lead to a wide variety of S_r being measured. This has been observed experimentally [49].

Trapped Cluster Distribution Introduction of correlated heterogeneity has a strong effect on the resultant distribution of clusters of the trapped defending fluid [29]. For small-scale correlations one observes the presence of larger clusters of trapped phase. For large scale correlations one or two trapped clusters can account for much of the trapped defending phase saturation. This has important implications to recovery of fluids from rocks under tertiary displacements where a third phase is injected to further reduce the saturation of the defending fluid. The presence of larger residual clusters may make it easier to reconnect and recover the trapped phase.

Stratification and Relative Permeability Extending the work on substrates generated from fractional Brownian motion, Paterson et al. [49] simulated invasion percolation with anisotropic correlations that were introduced to simulate the stratification often apparent in sedimentary rocks. The anisotropy was created by compressing the property maps in one direction and then deleting layers so the the grid spacing in each direction is equal. In their study invasion percolation was being used as a step to calculate relative permeability in porous media. The anisotropic heterogeneity led to different sets of relative permeability curves parallel and perpendicular to the bedding direction. This was consistent with experimental observations that relative permeability at a given saturation is greater for flow parallel to the bedding when compared with flow perpendicular to the bedding.

IP in Fractured Systems

Fractured systems differ from granular porous media in that representation by a network of pores (sites) connected via throats (bonds) disappears. Instead a fracture can be represented by a single aperture of varying thickness. Nevertheless, surface roughness within the fracture creates variable capillary pressure in two-phase flow, so it is still possible to apply the invasion percolation concept.

To model flow in fractal fractures, Wagner et al. [65, 68] used invasion percolation with trapping on two-dimensional substrates with a correlated distribution of invasion thresholds. Hence their simulations are essentially two-dimensional porous-media flow with heterogeneity. They simulated displacement in the void space between one fBm surface and one plane surface, and studied how

Invasion Percolation, Algorithm 1

Mathematica algorithm for generating invasion perlocation clusters from a single starting cell. Code is from [20]. In Mathematica version 6 `Table[Random[],{4}]` can be replaced with `RandomReal[{0,1},4]`

Invasion Percolation, Algorithm 2
Mathematica code that can be used to plot clusters generated from the invasion[] function

fractal scaling behavior depends on the surface roughness. Simulations on fractures consisting of a fBm surface and its displaced replica displayed a cross-over phenomenon. For displacements longer than the correlation length the flow patterns were found to have the properties of ordinary IP clusters grown on uncorrelated substrates.

In a separate line of investigation, Glass et al. [21] used invasion percolation to model experiments in a synthetic horizontal fracture made from two plates of roughened glass in contact. The aperture field for this fracture was measured using a light absorption technique. To apply invasion percolation they analyzed the correct curvature to use for the fluid invasion steps. A subsequent study by Glass et al. [22] developed simulations for a fracture patterned on measured data from a block of welded tuff.

Invasion percolation simulations were also matched to experiments by Amundsen et al. [1]. Their experiments involved two different rough bottom plates with a smooth planar top plate. One of the bottom plates was plastic milled to a self-affine surface with a Hurst exponent of 0.8. The other bottom plate was textured glass measured using a light absorption technique like Glass et al. [21].

Future Directions

The simple invasion percolation model provides a very realistic simulation of the slow fluid-fluid displacement processes within porous materials. The utility of the model and its variants to important applications are numerous. They include the understanding of contaminant migration in soils crucial to the successful implementation of groundwater remediation strategies. The ability to predict the transport of contaminants in soils will impact on the understanding of water quality issues. After primary oil recovery, more than 50% of the original oil in place remains unrecovered; a significant volume fraction of the pore space occupied by oil and gas is unrecovered because it is bypassed in the rock by the combined effects of the natural water drive mechanism, capillary forces and rock heterogeneity. Realistic estimation of recoveries is a central problem in the development of new fields and in the development of improved oil recovery methods in existing fields. The further development of the invasion percolation model and its variants coupled with a more realistic structural characterization of the pore structure of porous materials will play a crucial role in the development of an understanding of these important problems.

Bibliography

1. Amundsen H, Wagner G, Oxaal U, Meakin P, Feder J, Jøssang T (1999) Slow two-phase flow in artificial fractures: Experiments and simulations. Water Resour Res 35:2619–2626
2. Arns CH, Sakellariou A, Senden TJ, Sheppard AP, Sok RM, Pinczewski WV, Knackstedt MA (2005) Digital core laboratory: Petrophysical analysis from 3D images. Petrophysics 46(4):260–277
3. Babadagli T (2000) Invasion percolation in correlated porous media. Physica A 285:248–258
4. Babalievski F (1998) Cluster counting: The Hoshen–Kopelman algorithm vs. spanning tree approaches. Int J Mod Phys C 9:43–60
5. Bakke S, Øren P (1997) 3-D pore-scale modelling of sandstones and flow simulations in the pore networks. SPE J 2:136–149
6. Blunt MJ (1997) Pore level modeling of the effects of wettability. SPE J 2:494–510
7. Blunt MJ, Scher H (1995) Pore-level model of wetting. Phys Rev E 52:6387–6403
8. Blunt MJ, Jackson MD, Piri M, Valvatne PH (2002) Detailed physics, predictive capabilities and macroscopic consequences for pore-network models of multiphase flow. Adv Water Resour 25:1069–1089
9. Chandler R, Koplik J, Lerman K, Willemsen J (1982) Capillary displacement and percolation in porous media. J Fluid Mech 119:249–267
10. Chen JD, Wada N (1986) Visualisation of immiscible displacement in a three dimensional transparent porous medium. Exp Fluids 4:336–338
11. Cieplak M, Maritan A, Banavar JR (1994) Optimal paths and domain walls in the strong disorder limit. Phys Rev Lett 72:2320–2323
12. Dobrin R, Duxbury P (2001) Minimum spanning trees on random networks. Phys Rev Lett 86:5076–5079

13. Du C, Xu B, Yortsos YC, Chaouche M, Rakotomalala N, Salin D (1995) Correlation of occupation profiles in invasion percolation. Phys Rev Lett 74:694–697
14. Du C, Satik C, Yortsos Y (1996) Percolation in a fractional Brownian motion lattice. AIChE J 42:2392–2394
15. Dunsmuir JH, Ferguson SR, D'Amico KL (1991) Design and operation of an imaging X-ray detector for microtomography. IOP Conf Ser 121:257–261
16. Flannery BP, Deckman HW, Roberge WG, D'Amico KL (1987) Three-dimensional X-ray microtomography. Science 237: 1439–1444
17. Franzese G, Cataudella V, Coniglio A (1998) Invaded cluster dynamics for frustrated models. Phys Rev E 57:88–93
18. Fredrich J, Menendez B, Wong TF (1995) Imaging the pore structure of geomaterials. Science 268:276–279
19. Galam S, Mauger A (1996) Universal formulas for percolation thresholds. Phys Rev E 53:2177–2180
20. Gaylord RJ, Wellin PR (1994) Computer simulations with mathematica: Explorations in complex physical and biological systems. TELOS/Springer, New York
21. Glass RJ, Nicholl MJ, Yarrington L (1998) A modified invasion percolation model for low-capillary number immiscible displacements in horizontal rough-walled fractures: Influence of local in-plane curvature. Water Resour Res 34(12):3215–3234
22. Glass RJ, Nicholl MJ, Rajaram H, Andre B (2004) Development of slender transport pathways in unsaturated fractured rock: Simulation with modified invasion percolation. Geophys Res Lett 31:L06502
23. Heiba A, Sahimi M, Scriven L, Davis H (1992) Percolation theory of two-phase relative permeability. SPE Reserv Engin 7:123–132
24. Hoshen J, Kopelman R (1976) Percolation and cluster sizes. Phys Rev B 14:3438
25. Ioannidis MA, Kwiecien MJ, Chatzis I, MacDonald IF, Dullien FAL (1997) Comprehensive pore structure characterization using 3D computer reconstruction and stochastic modeling. In: SPE Annual Technical Conference and Exhibition held in San Antonio, Texas, USA, 1997
26. Knackstedt MA, Sheppard AP, Pinczewski WV (1998) Simulation of mercury porosimetry on correlated grids: Evidence for extended correlated heterogeneity at the pore scale in rocks. Phys Rev E 58:6923–6926
27. Knackstedt MA, Sheppard AP, Pinczewski WV (1998) Simulation of mercury porosimetry on correlated grids: Evidence for extended correlated heterogeneity at the pore scale in rocks. Phys Rev E Rapid Communications 58:R6923–R6926
28. Knackstedt MA, Sahimi M, Sheppard AP (2000) Invasion percolation with long-range correlations: First-order phase transition and nonuniversal scaling properties. Phys Rev E 61:4920–4934
29. Knackstedt MA, Marrink S, Sheppard AP, Pinczewski W, Sahimi M (2001) Invasion percolation on correlated and elongated lattices: Implications for the interpretation of residual saturations in rock cores. Transp Porous Media 44:465–485
30. Larson R, Scriven LE, Davis HT (1977) Percolation theory of residual phases in porous media. Nature 268:409–413
31. Larson R, Scriven LE, Davis HT (1991) Percolation theory of two-phase flow in porous media. Chem Eng Sci 36:57–73
32. Lee J-Y, Robins V, Sok RM, Sheppard AP, Pinczewski W, Knackstedt MA (2004) Effect of topology on relative permeability. Transp Porous Media 55:21–46
33. Lenormand R, Bories S (1980) Description d'un mecanisme de connexion de liaison destin l'tude du drainage avec pigeage en milieu poreux. CR Acad Sci Paris B 291:279
34. Lenormand R, Bories S (1985) Fractal patterns from chemical dissolution. Physicochem Hydro 6:497
35. Lenormand R, Zarcone C (1985) Invasion percolation in an etched netowrk; measurement of a fractal dimension. Phys Rev Lett 54:2226–2229
36. Lindquist B, Lee SM, Coker D (1996) Medial axis analysis of void structure in three-dimensional tomographic images of porous media. J Geophys Res 101B:8297–8310
37. Lindquist WB, Venkatarangan A, Dunsmuir J, Wong TF (2000) Pore and throat size distributions measured from synchrotron X-ray tomographic images of fontainbleau sandstones. J Geophys Res 105B:21508
38. Marrink SJ, Paterson L, Knackstedt M (2000) Definition of percolation thresholds on self-affine surfaces. Physica A 280:207–214
39. Meakin P (1988) Invasion percolation and invading Eden growth on multifractal lattices. J Phys A: Math Gen 21:3501–3522
40. Meakin P (1991) Invasion percolation on substrates with correlated disorder. Physica A 173:305–324
41. Meakin P, Feder J, Frette V, Jøssang T (1992) Invasion percolation in a destabilizing gradient. Phys Rev A 46:3357–3368
42. Meakin P, Wagner G, Frette V, Feder J, Jøssang T (1995) Fractals and secondary migration. Fractals 3:799–806
43. Meakin P, Wagner G, Vedvik A, Amundsen H, Feder J, Jøssang T (2000) Invasion percolation and secondary migration: experiments and simulations. Mar Pet Geol 17:777–795
44. Molz FJ, Liu HH, Szulga J (1997) Fractional Brownian motion and fractional Gaussian noise in subsurface hydrology: A review, presentation of fundamental properties, and extensions. Water Resour Res 33:2273–2286
45. Øren P, Bakke S, Arntzen OJ (1998) Extending predictive capabilities to network models. SPE J 3:324–336
46. Painter S (2001) Flexible scaling model for use in random field simulation of hydraulic conductivity. Water Resour Res 37:1155–1163
47. Paterson L (1998) Trapping thresholds in ordinary percolation. Phys Rev E 58:7137–7140
48. Paterson L, Painter S, Knackstedt MA, Pinczewski WV (1996) Patterns of fluid flow in naturally heterogeneous rocks. Physica A 233:619–628
49. Paterson L, Painter S, Zhang X, Pinczewski WV (1998) Simulating residual saturation and relative permeability in heterogeneous formations. SPE J 3:211–218
50. Paterson L, Sheppard AP, Knackstedt MA (2002) Trapping thresholds in invasion percolation. Phys Rev E 66:056122
51. Porto M, Havlin S, Schwarzer S, Bunde A (1997) Optimal path in strong disorder and shortest path in invasion percolation with trapping. Phys Rev Lett 79:4060–4063
52. Sahimi M (1994) Applications of percolation theory, 1st edn. Taylor Francis, London
53. Sheppard AP, Knackstedt MA, Pinczewski WV, Sahimi M (1999) Invasion percolation: New algorithms and universality classes. J Phys A Lett 32:L521–L529
54. Sheppard AP, Knackstedt MA, Pinczewski WV, Sahimi M (1999) Invasion percolation: New algorithms and universality classes. J Phys A: Math Gen 32:L521–L529
55. Sheppard AP, Sok RM, Averdunk H (2005) Improved pore net-

work extraction methods. 19th International Symposium of the SCA, SCA, Toronto, 2005
56. Sok RM, Knackstedt MA, Sheppard AP, Pinczewski W, Lindquist WB, Venkatarangan A, Paterson L (2002) Direct and stochastic generation of network models from tomographic images; effect of topology on two phase flow properties. Transp Porous Media 46:345–372
57. Spanne P, Thovert J, Jacquin J, Lindquist WB, Jones K, Adler PM (1994) Synchotron computed microtomography of porous media: Topology and transports. Phys Rev Lett 73:2001–2004
58. Stark C (1991) An invasion percolation model of drainage network evolution. Nature 352:423
59. Stauffer D, Aharony A (1994) Introduction to percolation theory, 2nd edn. Taylor Francis, London
60. Stokes JP, Weitz D, Gollub J, Dougherty A, Robbins M, Chaikin P, Lindsay H (1986) Interfacial stability of immiscible displacement in a porous medium. Phys Rev Lett 57:2226–2229
61. Suding PN, Ziff RM (1999) Site percolation thresholds for archimedean lattices. Phys Rev E 60:275–283
62. Thovert J-F, Salles J, Adler P (1993) Computerised characterization of the geometry of real porous media: Their description, analysis and interpretation, J Microsc 170:65–79
63. Tsimpanogiannis IN, Yortsos YC (2004) The critical gas saturation in a porous medium in the presence of gravity. J Colloid Interface Sci 270:388–395
64. Vedvik A, Wagner G, Oxaal U, Feder J, Meakin P, Jøssang T (1998) Fragmentation transition for invasion percolation in hydraulic gradients. Phys Rev Lett 80:3065–3068
65. Wagner G, Amundsen H, Oxaal U, Meakin P, Feder J, Jøssang T (2000) Slow two-phase flow in single fractures: Fragmentation, migration, and fractal patterns simulated using invasion percolation models. Pure Appl Geophys 157:621–635
66. Wagner G, Meakin P, Feder J, Jøssang T (1997) Buoyancy-driven invasion percolation with migration and fragmentation. Physica A 245:217–230
67. Wagner G, Meakin P, Feder J, Jøssang T (1997) Invasion percolation on self-affine topographies. Phys Rev E 55:1698–1703
68. Wagner G, Meakin P, Feder J, Jøssang T (1999) Invasion percolation in fractal fractures. Physica A 264:321–337
69. Wilkinson D (1984) Percolation model of immiscible displacement in the presence of buoyancy forces. Phys Rev A 30:520–531
70. Wilkinson D (1986) Percolation effects in immiscible displacement. Phys Rev A 34:1380–1391
71. Wilkinson D, Willemsen JF (1983) Invasion percolation: A new form of percolation theory. J Phys A: Math Gen 16:3365–3376
72. Xu B, Yortsos YC, Salin D (1998) Invasion percolation with viscous forces. Phys Rev E 57:739–751
73. Yuan H (1991) Pore-scale heterogeneity from mercury porosimetry data. SPE Form Eval 6:233–242

Inverse Scattering Transform and the Theory of Solitons

Tuncay Aktosun
Department of Mathematics, University of Texas at Arlington, Arlington, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Inverse Scattering Transform
The Lax Method
The AKNS Method
Direct Scattering Problem
Time Evolution of the Scattering Data
Inverse Scattering Problem
Solitons
Future Directions
Bibliography

Glossary

AKNS method A method introduced by Ablowitz, Kaup, Newell, and Segur in 1973 that identifies the nonlinear partial differential equation (NPDE) associated with a given first-order system of linear ordinary differential equations (LODEs) so that the initial value problem (IVP) for that NPDE can be solved by the inverse scattering transform (IST) method.

Direct scattering problem The problem of determining the scattering data corresponding to a given potential in a differential equation.

Integrability A NPDE is said to be integrable if its IVP can be solved via an IST.

Inverse scattering problem The problem of determining the potential that corresponds to a given set of scattering data in a differential equation.

Inverse scattering transform A method introduced in 1967 by Gardner, Greene, Kruskal, and Miura that yields a solution to the IVP for a NPDE with the help of the solutions to the direct and inverse scattering problems for an associated LODE.

Lax method A method introduced by Lax in 1968 that determines the integrable NPDE associated with a given LODE so that the IVP for that NPDE can be solved with the help of an IST.

Scattering data The scattering data associated with a LODE usually consists of a reflection coefficient which is a function of the spectral parameter λ, a finite number of constants λ_j that correspond to the poles of the transmission coefficient in the upper half complex plane, and the bound-state norming constants whose number for each bound-state pole λ_j is the same as the order of that pole. It is desirable that the potential in the LODE is uniquely determined by the corresponding scattering data and vice versa.

Soliton The part of a solution to an integrable NPDE due to a pole of the transmission coefficient in the upper half complex plane. The term soliton was introduced by Zabusky and Kruskal in 1965 to denote a solitary wave pulse with a particle-like behavior in the solution to the Korteweg-de Vries (KdV) equation.

Time evolution of the scattering data The evolvement of the scattering data from its initial value $S(\lambda, 0)$ at $t = 0$ to its value $S(\lambda, t)$ at a later time t.

Definition of the Subject

A general theory to solve NPDEs does not seem to exist. However, there are certain NPDEs, usually first order in time, for which the corresponding IVPs can be solved by the IST method. Such NPDEs are sometimes referred to as integrable evolution equations. Some exact solutions to such equations may be available in terms of elementary functions, and such solutions are important to understand nonlinearity better and they may also be useful in testing accuracy of numerical methods to solve such NPDEs.

Certain special solutions to some of such NPDEs exhibit particle-like behaviors. A single-soliton solution is usually a localized disturbance that retains its shape but only changes its location in time. A multi-soliton solution consists of several solitons that interact nonlinearly when they are close to each other but come out of such interactions unchanged in shape except for a phase shift.

Integrable NPDEs have important physical applications. For example, the KdV equation is used to describe [14,23] surface water waves in long, narrow, shallow canals; it also arises [23] in the description of hydromagnetic waves in a cold plasma, and ion-acoustic waves in anharmonic crystals. The nonlinear Schrödinger (NLS) equation arises in modeling [24] electromagnetic waves in optical fibers as well as surface waves in deep waters. The sine-Gordon equation is helpful [1] in analyzing the magnetic field in a Josephson junction (gap between two superconductors).

Introduction

The first observation of a soliton was made in 1834 by the Scottish engineer John Scott Russell at the Union Canal between Edinburgh and Glasgow. Russell reported [21] his observation to the British Association of the Advancement of Science in September 1844, but he did not seem to be successful in convincing the scientific community. For example, his contemporary George Airy, the influential mathematician of the time, did not believe in the existence of solitary water waves [1].

The Dutch mathematician Korteweg and his doctoral student de Vries published [14] a paper in 1895 based on de Vries' Ph.D. dissertation, in which surface waves in shallow, narrow canals were modeled by what is now known as the KdV equation. The importance of this paper was not understood until 1965 even though it contained as a special solution what is now known as the one-soliton solution.

Enrico Fermi in his summer visits to the Los Alamos National Laboratory, together with J. Pasta and S. Ulam, used the computer named Maniac I to computationally analyze a one-dimensional dynamical system of 64 particles in which adjacent particles were joined by springs where the forces also included some nonlinear terms. Their main goal was to determine the rate of approach to the equipartition of energy among different modes of the system. Contrary to their expectations there was little tendency towards the equipartition of energy but instead the almost ongoing recurrence to the initial state, which was puzzling. After Fermi died in November 1954, Pasta and Ulam completed their last few computational examples and finished writing a preprint [11], which was never published as a journal article. This preprint appears in Fermi's Collected Papers [10] and is also available on the internet [25].

In 1965 Zabusky and Kruskal explained [23] the Fermi-Pasta-Ulam puzzle in terms of solitary wave solutions to the KdV equation. In their numerical analysis they observed "solitary-wave pulses", named such pulses "solitons" because of their particle-like behavior, and noted that such pulses interact with each other nonlinearly but come out of interactions unaffected in size or shape except for some phase shifts. Such unusual interactions among solitons generated a lot of excitement, but at that time no one knew how to solve the IVP for the KdV equation, except numerically. In 1967 Gardner, Greene, Kruskal, and Miura presented [12] a method, now known as the IST, to solve that IVP, assuming that the initial profile $u(x, 0)$ decays to zero sufficiently rapidly as $x \to \pm\infty$. They showed that the integrable NPDE, i. e. the KdV equation,

$$u_t - 6uu_x + u_{xxx} = 0 \,, \tag{1}$$

is associated with a LODE, i. e. the 1-D Schrödinger equation

$$-\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} + u(x,t)\,\psi = k^2\psi \,, \tag{2}$$

and that the solution $u(x, t)$ to (1) can be recovered from the initial profile $u(x, 0)$ as explained in the diagram given in Sect. "Inverse Scattering Transform". They also ex-

plained that soliton solutions to the KdV equation correspond to a zero reflection coefficient in the associated scattering data. Note that the subscripts x and t in (1) and throughout denote the partial derivatives with respect to those variables.

In 1972 Zakharov and Shabat showed [24] that the IST method is applicable also to the IVP for the NLS equation

$$iu_t + u_{xx} + 2u|u|^2 = 0 \,, \tag{3}$$

where i denotes the imaginary number $\sqrt{-1}$. They proved that the associated LODE is the first-order linear system

$$\begin{cases} \dfrac{\mathrm{d}\xi}{\mathrm{d}x} = -i\lambda\xi + u(x,t)\,\eta \,, \\ \dfrac{\mathrm{d}\eta}{\mathrm{d}x} = i\lambda\eta - \overline{u(x,t)}\,\xi \,, \end{cases} \tag{4}$$

where λ is the spectral parameter and an overline denotes complex conjugation. The system in (4) is now known as the Zakharov–Shabat system.

Soon afterwards, again in 1972 Wadati showed in a one-page publication [22] that the IVP for the modified Korteweg-de Vries (mKdV) equation

$$u_t + 6u^2 u_x + u_{xxx} = 0 \,, \tag{5}$$

can be solved with the help of the inverse scattering problem for the linear system

$$\begin{cases} \dfrac{\mathrm{d}\xi}{\mathrm{d}x} = -i\lambda\xi + u(x,t)\,\eta \,, \\ \dfrac{\mathrm{d}\eta}{\mathrm{d}x} = i\lambda\eta - u(x,t)\,\xi \,. \end{cases} \tag{6}$$

Next, in 1973 Ablowitz, Kaup, Newell, and Segur showed [2,3] that the IVP for the sine-Gordon equation

$$u_{xt} = \sin u \,,$$

can be solved in the same way by exploiting the inverse scattering problem associated with the linear system

$$\begin{cases} \dfrac{\mathrm{d}\xi}{\mathrm{d}x} = -i\lambda\xi - \dfrac{1}{2}\,u_x(x,t)\,\eta \,, \\ \dfrac{\mathrm{d}\eta}{\mathrm{d}x} = i\lambda\eta + \dfrac{1}{2}\,u_x(x,t)\,\xi \,. \end{cases}$$

Since then, many other NPDEs have been discovered to be solvable by the IST method.

Our review is organized as follows. In the next section we explain the idea behind the IST. Given a LODE known to be associated with an integrable NPDE, there are two primary methods enabling us to determine the corresponding NPDE. We review those two methods, the Lax method and the AKNS method, in Sect. "The Lax Method" and in Sect. "The AKNS Method", respectively. In Sect. "Direct Scattering" we introduce the scattering data associated with a LODE containing a spectral parameter and a potential, and we illustrate it for the Schrödinger equation and for the Zakharov–Shabat system. In Sect. "Time Evolution of the Scattering Data" we explain the time evolution of the scattering data and indicate how the scattering data sets evolve for those two LODEs. In Sect. "Inverse Scatering Problem" we summarize the Marchenko method to solve the inverse scattering problem for the Schrödinger equation and that for the Zakharov–Shabat system, and we outline how the solutions to the IVPs for the KdV equation and the NLS equation are obtained with the help of the IST. In Sect. "Solitons" we present soliton solutions to the KdV and NLS equations. A brief conclusion is provided in Sect. "Future Directions".

Inverse Scattering Transform

Certain NPDEs are classified as integrable in the sense that their corresponding IVPs can be solved with the help of an IST. The idea behind the IST method is as follows: Each integrable NPDE is associated with a LODE (or a system of LODEs) containing a parameter λ (usually known as the spectral parameter), and the solution $u(x,t)$ to the NPDE appears as a coefficient (usually known as the potential) in the corresponding LODE. In the NPDE the quantities x and t appear as independent variables (usually known as the spatial and temporal coordinates, respectively), and in the LODE x is an independent variable and λ and t appear as parameters. It is usually the case that $u(x,t)$ vanishes at each fixed t as x becomes infinite so that a scattering scenario can be created for the related LODE, in which the potential $u(x,t)$ can uniquely be associated with some scattering data $S(\lambda,t)$. The problem of determining $S(\lambda,t)$ for all λ values from $u(x,t)$ given for all x values is known as the direct scattering problem for the LODE. On the other hand, the problem of determining $u(x,t)$ from $S(\lambda,t)$ is known as the inverse scattering problem for that LODE.

The IST method for an integrable NPDE can be explained with the help of Diagram 1.

In order to solve the IVP for the NPDE, i. e. in order to determine $u(x,t)$ from $u(x,0)$, one needs to perform the following three steps:

(i) Solve the corresponding direct scattering problem for the associated LODE at $t = 0$, i. e. determine the initial scattering data $S(\lambda,0)$ from the initial potential $u(x,0)$.

$$\begin{tikzcd}
u(x,0) \arrow[r, "\text{direct scattering for LODE at } t=0"] \arrow[d, "\text{solution to NPDE}"'] & S(\lambda,0) \arrow[d, "\text{time evolution of scattering data}"] \\
u(x,t) & S(\lambda,t) \arrow[l, "\text{inverse scattering for LODE at time } t"]
\end{tikzcd}$$

Inverse Scattering Transform and the Theory of Solitons, Diagram 1
The method of inverse scattering transform

(ii) Time evolve the scattering data from its initial value $S(\lambda, 0)$ to its value $S(\lambda, t)$ at time t. Such an evolution is usually a simple one and is particular to each integrable NPDE.
(iii) Solve the corresponding inverse scattering problem for the associated LODE at fixed t, i. e. determine the potential $u(x, t)$ from the scattering data $S(\lambda, t)$.

It is amazing that the resulting $u(x, t)$ satisfies the integrable NPDE and that the limiting value of $u(x, t)$ as $t \to 0$ agrees with the initial profile $u(x, 0)$.

The Lax Method

In 1968 Peter Lax introduced [15] a method yielding an integrable NPDE corresponding to a given LODE. The basic idea behind the Lax method is the following. Given a linear differential operator $\mathcal{L}$ appearing in the spectral problem $\mathcal{L}\psi = \lambda\psi$, find an operator $\mathcal{A}$ (the operators $\mathcal{A}$ and $\mathcal{L}$ are said to form a Lax pair) such that:

(i) The spectral parameter λ does not change in time, i. e. $\lambda_t = 0$.
(ii) The quantity $\psi_t - \mathcal{A}\psi$ remains a solution to the same linear problem $\mathcal{L}\psi = \lambda\psi$.
(iii) The quantity $\mathcal{L}_t + \mathcal{L}\mathcal{A} - \mathcal{A}\mathcal{L}$ is a multiplication operator, i. e. it is not a differential operator.

From condition (ii) we get

$$\mathcal{L}(\psi_t - \mathcal{A}\psi) = \lambda(\psi_t - \mathcal{A}\psi) , \tag{7}$$

and with the help of $\mathcal{L}\psi = \lambda\psi$ and $\lambda_t = 0$, from (7) we obtain

$$\begin{aligned}\mathcal{L}\psi_t - \mathcal{L}\mathcal{A}\psi &= \lambda\psi_t - \mathcal{A}(\lambda\psi) = \partial_t(\lambda\psi) - \mathcal{A}\mathcal{L}\psi \\ &= \partial_t(\mathcal{L}\psi) - \mathcal{A}\mathcal{L}\psi = \mathcal{L}_t\psi + \mathcal{L}\psi_t - \mathcal{A}\mathcal{L}\psi ,\end{aligned} \tag{8}$$

where ∂_t denotes the partial differential operator with respect to t. After canceling the term $\mathcal{L}\psi_t$ on the left and right hand sides of (8), we get

$$(\mathcal{L}_t + \mathcal{L}\mathcal{A} - \mathcal{A}\mathcal{L})\psi = 0 ,$$

which, because of (iii), yields

$$\mathcal{L}_t + \mathcal{L}\mathcal{A} - \mathcal{A}\mathcal{L} = 0 . \tag{9}$$

Note that (9) is an evolution equation containing a first-order time derivative, and it is the desired integrable NPDE. The equation in (9) is often called a compatibility condition.

Having outlined the Lax method, let us now illustrate it to derive the KdV equation in (1) from the Schrödinger equation in (2). For this purpose, we write the Schrödinger equation as $\mathcal{L}\psi = \lambda\psi$ with $\lambda := k^2$ and

$$\mathcal{L} := -\partial_x^2 + u(x, t) , \tag{10}$$

where the notation := is used to indicate a definition so that the quantity on the left should be understood as the quantity on the right hand side. Given the linear differential operator $\mathcal{L}$ defined as in (10), let us try to determine the associated operator $\mathcal{A}$ by assuming that it has the form

$$\mathcal{A} = \alpha_3\partial_x^3 + \alpha_2\partial_x^2 + \alpha_1\partial_x + \alpha_0 , \tag{11}$$

where the coefficients α_j with $j = 0, 1, 2, 3$ may depend on x and t, but not on the spectral parameter λ. Note that $\mathcal{L}_t = u_t$. Using (10) and (11) in (9), we obtain

$$(\,)\partial_x^5 + (\,)\partial_x^4 + (\,)\partial_x^3 + (\,)\partial_x^2 + (\,)\partial_x + (\,) = 0 , \tag{12}$$

where, because of (iii), each coefficient denoted by () must vanish. The coefficient of ∂_x^5 vanishes automatically. Setting the coefficients of ∂_x^j to zero for $j = 4, 3, 2, 1$, we obtain

$$\alpha_3 = c_1, \quad \alpha_2 = c_2, \quad \alpha_1 = c_3 - \frac{3}{2}c_1 u,$$
$$\alpha_0 = c_4 - \frac{3}{4}c_1 u_x - c_2 u ,$$

with c_1, c_2, c_3, and c_4 denoting arbitrary constants. Choosing $c_1 = -4$ and $c_3 = 0$ in the last coefficient in (12) and setting that coefficient to zero, we get the KdV equation in (1). Moreover, by letting $c_2 = c_4 = 0$, we obtain the operator $\mathcal{A}$ as

$$\mathcal{A} = -4\partial_x^3 + 6u\partial_x + 3u_x . \tag{13}$$

For the Zakharov–Shabat system in (4), we proceed in a similar way. Let us write it as $\mathcal{L}\psi = \lambda\psi$, where the linear differential operator $\mathcal{L}$ is defined via

$$\mathcal{L} := i\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}\partial_x - i\begin{bmatrix}0 & u(x,t)\\ \overline{u(x,t)} & 0\end{bmatrix}.$$

Then, the operator $\mathcal{A}$ is obtained as

$$\mathcal{A} = 2i\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}\partial_x^2 - 2i\begin{bmatrix}0 & u\\ \overline{u} & 0\end{bmatrix}\partial_x - i\begin{bmatrix}-|u|^2 & u_x\\ \overline{u}_x & |u|^2\end{bmatrix}, \tag{14}$$

and the compatibility condition (9) gives us the NLS equation in (3).

For the first-order system (6), by writing it as $\mathcal{L}\psi = \lambda\psi$, where the linear operator $\mathcal{L}$ is defined by

$$\mathcal{L} := i\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}\partial_x - i\begin{bmatrix}0 & u(x,t)\\ u(x,t) & 0\end{bmatrix},$$

we obtain the corresponding operator $\mathcal{A}$ as

$$\mathcal{A} = -4\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\partial_x^3 - 6\begin{bmatrix}u^2 & -u_x\\ u_x & u^2\end{bmatrix}\partial_x - \begin{bmatrix}6uu_x & -3u_{xx}\\ 3u_{xx} & 6uu_x\end{bmatrix},$$

and the compatibility condition (9) yields the mKdV equation in (5).

The AKNS Method

In 1973 Ablowitz, Kaup, Newell, and Segur introduced [2,3] another method to determine an integrable NPDE corresponding to a LODE. This method is now known as the AKNS method, and the basic idea behind it is the following. Given a linear operator $\mathcal{X}$ associated with the first-order system $v_x = \mathcal{X}v$, we are interested in finding an operator $\mathcal{T}$ (the operators $\mathcal{T}$ and $\mathcal{X}$ are said to form an AKNS pair) such that:

(i) The spectral parameter λ does not change in time, i. e. $\lambda_t = 0$.
(ii) The quantity $v_t - \mathcal{T}v$ is also a solution to $v_x = \mathcal{X}v$, i. e. we have $(v_t - \mathcal{T}v)_x = \mathcal{X}(v_t - \mathcal{T}v)$.
(iii) The quantity $\mathcal{X}_t - \mathcal{T}_x + \mathcal{X}\mathcal{T} - \mathcal{T}\mathcal{X}$ is a (matrix) multiplication operator, i. e. it is not a differential operator.

From condition (ii) we get

$$\begin{aligned} v_{tx} - \mathcal{T}_x v - \mathcal{T}v_x &= \mathcal{X}v_t - \mathcal{X}\mathcal{T}v \\ &= (\mathcal{X}v)_t - \mathcal{X}_t v - \mathcal{X}\mathcal{T}v \\ &= (v_x)_t - \mathcal{X}_t v - \mathcal{X}\mathcal{T}v \\ &= v_{xt} - \mathcal{X}_t v - \mathcal{X}\mathcal{T}v\,. \end{aligned} \tag{15}$$

Using $v_{tx} = v_{xt}$ and replacing $\mathcal{T}v_x$ by $\mathcal{T}\mathcal{X}v$ on the left side and equating the left and right hand sides in (15), we obtain

$$(\mathcal{X}_t - \mathcal{T}_x + \mathcal{X}\mathcal{T} - \mathcal{T}\mathcal{X})v = 0\,,$$

which in turn, because of (iii), implies

$$\mathcal{X}_t - \mathcal{T}_x + \mathcal{X}\mathcal{T} - \mathcal{T}\mathcal{X} = 0\,. \tag{16}$$

We can view (16) as an integrable NPDE solvable with the help of the solutions to the direct and inverse scattering problems for the linear system $v_x = \mathcal{X}v$. Like (9), the compatibility condition (16) yields a nonlinear evolution equation containing a first-order time derivative. Note that $\mathcal{X}$ contains the spectral parameter λ, and hence $\mathcal{T}$ also depends on λ as well. This is in contrast with the Lax method in the sense that the operator $\mathcal{A}$ does not contain λ.

Let us illustrate the AKNS method by deriving the KdV equation in (1) from the Schrödinger equation in (2). For this purpose we write the Schrödinger equation, by replacing the spectral parameter k^2 with λ, as a first-order linear system $v_x = \mathcal{X}v$, where we have defined

$$v := \begin{bmatrix}\psi_x\\ \psi\end{bmatrix}, \quad \mathcal{X} := \begin{bmatrix}0 & u(x,t)-\lambda\\ 1 & 0\end{bmatrix}.$$

Let us look for $\mathcal{T}$ in the form

$$\mathcal{T} = \begin{bmatrix}\alpha & \beta\\ \rho & \sigma\end{bmatrix},$$

where the entries α, β, ρ, and σ may depend on x, t, and λ. The compatibility condition (16) yields

$$\begin{bmatrix}-\alpha_x - \beta + \rho(u-\lambda) & u_t - \beta_x + \sigma(u-\lambda) - \alpha(u-\lambda)\\ -\rho_x + \alpha - \sigma & -\sigma_x + \beta - \rho(u-\lambda)\end{bmatrix} = \begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}. \tag{17}$$

The $(1,1)$, $(2,1)$, and $(2,2)$-entries in the matrix equation in (17) imply

$$\beta = -\alpha_x + (u-\lambda)\rho, \quad \sigma = \alpha - \rho_x, \quad \sigma_x = -\alpha_x\,. \tag{18}$$

Then from the $(1,2)$-entry in (17) we obtain

$$u_t + \tfrac{1}{2}\rho_{xxx} - u_x\rho - 2\rho_x(u-\lambda) = 0\,. \tag{19}$$

Assuming a linear dependence of ρ on the spectral parameter and hence letting $\rho = \lambda\zeta + \mu$ in (19), we get

$$2\zeta_x\lambda^2 + \left(\tfrac{1}{2}\zeta_{xxx} - 2\zeta_x u + 2\mu_x - u_x\zeta\right)\lambda + \left(u_t + \tfrac{1}{2}\mu_{xxx} - 2\mu_x u - u_x\mu\right) = 0\,.$$

Equating the coefficients of each power of λ to zero, we have

$$\zeta = c_1, \quad \mu = \tfrac{1}{2}c_1 u + c_2, \quad u_t - \tfrac{3}{2}c_1 u u_x - c_2 u_x + \tfrac{1}{4}c_1 u_{xxx} = 0\,, \tag{20}$$

with c_1 and c_2 denoting arbitrary constants. Choosing $c_1 = 4$ and $c_2 = 0$, from (20) we obtain the KdV equation given in (1). Moreover, with the help of (18) we get

$$\alpha = u_x + c_3\,,$$
$$\beta = -4\lambda^2 + 2\lambda u + 2u^2 - u_{xx}\,,$$
$$\rho = 4\lambda + 2u,\ \sigma = c_3 - u_x\,,$$

where c_3 is an arbitrary constant. Choosing $c_3 = 0$, we find

$$\mathcal{T} = \begin{bmatrix} u_x & -4\lambda^2 + 2\lambda u + 2u^2 - u_{xx} \\ 4\lambda + 2u & -u_x \end{bmatrix}.$$

As for the Zakharov–Shabat system in (4), writing it as $v_x = \mathcal{X}v$, where we have defined

$$\mathcal{X} := \begin{bmatrix} -i\lambda & u(x,t) \\ -\overline{u(x,t)} & i\lambda \end{bmatrix},$$

we obtain the matrix operator $\mathcal{T}$ as

$$\mathcal{T} = \begin{bmatrix} -2i\lambda^2 + i|u|^2 & 2\lambda u + iu_x \\ -2\lambda\overline{u} + i\overline{u}_x & 2i\lambda^2 - i|u|^2 \end{bmatrix},$$

and the compatibility condition (16) yields the NLS equation in (3).

As for the first-order linear system (6), by writing it as $v_x = \mathcal{X}v$, where

$$\mathcal{X} := \begin{bmatrix} -i\lambda & u(x,t) \\ -u(x,t) & i\lambda \end{bmatrix},$$

we obtain the matrix operator $\mathcal{T}$ as

$$\mathcal{T} = \begin{bmatrix} -4i\lambda^3 + 2i\lambda u^2 & 4\lambda^2 u + 2i\lambda u_x - u_{xx} - 2u^3 \\ -4\lambda^2 u + 2i\lambda u_x + u_{xx} + 2u^3 & 4i\lambda^3 - 2i\lambda u^2 \end{bmatrix}$$

and the compatibility condition (16) yields the mKdV equation in (5).

As for the first-order system $v_x = \mathcal{X}v$, where

$$\mathcal{X} := \begin{bmatrix} -i\lambda & -\frac{1}{2}u_x(x,t) \\ \frac{1}{2}u_x(x,t) & i\lambda \end{bmatrix},$$

we obtain the matrix operator $\mathcal{T}$ as

$$\mathcal{T} = \frac{i}{4\lambda}\begin{bmatrix} \cos u & \sin u \\ \sin u & -\cos u \end{bmatrix}.$$

Then, the compatibility condition (16) gives us the sine-Gordon equation

$$u_{xt} = \sin u\,.$$

Direct Scattering Problem

The direct scattering problem consists of determining the scattering data when the potential is known. This problem is usually solved by obtaining certain specific solutions, known as the Jost solutions, to the relevant LODE. The appropriate scattering data can be constructed with the help of spatial asymptotics of the Jost solutions at infinity or from certain Wronskian relations among the Jost solutions. In this section we review the scattering data corresponding to the Schrödinger equation in (2) and to the Zakharov–Shabat system in (4). The scattering data sets for other LODEs can similarly be obtained.

Consider (2) at fixed t by assuming that the potential $u(x,t)$ belongs to the Faddeev class, i. e. $u(x,t)$ is real valued and $\int_{-\infty}^{\infty} \mathrm{d}x\,(1+|x|)\,|u(x,t)|$ is finite. The Schrödinger equation has two types of solutions; namely, scattering solutions and bound-state solutions. The scattering solutions are those that consist of linear combinations of e^{ikx} and e^{-ikx} as $x \to \pm\infty$, and they occur for $k \in \mathbf{R} \setminus \{0\}$, i. e. for real nonzero values of k. Two linearly independent scattering solutions f_l and f_r, known as the Jost solution from the left and from the right, respectively, are those solutions to (2) satisfying the respective asymptotic conditions

$$f_\mathrm{l}(k,x,t) = \mathrm{e}^{ikx} + o(1)\,, \quad f_\mathrm{l}'(k,x,t) = ik\mathrm{e}^{ikx} + o(1)\,, \quad x \to +\infty\,, \tag{21}$$

$$f_\mathrm{r}(k,x,t) = \mathrm{e}^{-ikx} + o(1)\,, \quad f_\mathrm{r}'(k,x,t) = -ik\mathrm{e}^{-ikx} + o(1)\,, \quad x \to -\infty\,,$$

where the notation $o(1)$ indicates the quantities that vanish. Writing their remaining spatial asymptotics in the form

$$f_{\rm l}(k,x,t) = \frac{e^{ikx}}{T(k,t)} + \frac{L(k,t)\,e^{-ikx}}{T(k,t)} + o(1)\,, \qquad x \to -\infty\,, \tag{22}$$

$$f_{\rm r}(k,x,t) = \frac{e^{-ikx}}{T(k,t)} + \frac{R(k,t)\,e^{ikx}}{T(k,t)} + o(1)\,, \qquad x \to +\infty\,,$$

we obtain the scattering coefficients; namely, the transmission coefficient T and the reflection coefficients L and R, from the left and right, respectively.

Let $\mathbf{C}^+$ denote the upper half complex plane. A bound-state solution to (2) is a solution that belongs to $L^2(\mathbf{R})$ in the x variable. Note that $L^2(\mathbf{R})$ denotes the set of complex-valued functions whose absolute squares are integrable on the real line $\mathbf{R}$. When $u(x,t)$ is in the Faddeev class, it is known [5,7,8,9,16,17,18,19] that the number of bound states is finite, the multiplicity of each bound state is one, and the bound-state solutions can occur only at certain k-values on the imaginary axis in $\mathbf{C}^+$. Let us use N to denote the number of bound states, and suppose that the bound states occur at $k = i\kappa_j$ with the ordering $0 < \kappa_1 < \cdots < \kappa_N$. Each bound state corresponds to a pole of T in $\mathbf{C}^+$. Any bound-state solution at $k = i\kappa_j$ is a constant multiple of $f_{\rm l}(i\kappa_j, x, t)$. The left and right bound-state norming constants $c_{{\rm l}j}(t)$ and $c_{{\rm r}j}(t)$, respectively, can be defined as

$$c_{{\rm l}j}(t) := \left[\int_{-\infty}^{\infty} dx\, f_{\rm l}(i\kappa_j, x, t)^2\right]^{-1/2}\,,$$

$$c_{{\rm r}j}(t) := \left[\int_{-\infty}^{\infty} dx\, f_{\rm r}(i\kappa_j, x, t)^2\right]^{-1/2}\,,$$

and they are related to each other through the residues of T via

$$\mathrm{Res}\,(T, i\kappa_j) = i\, c_{{\rm l}j}(t)^2\, \gamma_j(t) = i\,\frac{c_{{\rm r}j}(t)^2}{\gamma_j(t)}\,, \tag{23}$$

where the $\gamma_j(t)$ are the dependency constants defined as

$$\gamma_j(t) := \frac{f_{\rm l}(i\kappa_j, x, t)}{f_{\rm r}(i\kappa_j, x, t)}\,. \tag{24}$$

The sign of $\gamma_j(t)$ is the same as that of $(-1)^{N-j}$, and hence $c_{{\rm r}j}(t) = (-1)^{N-j}\gamma_j(t)\, c_{{\rm l}j}(t)$.

The scattering matrix associated with (2) consists of the transmission coefficient T and the two reflection coefficients R and L, and it can be constructed from $\{\kappa_j\}_{j=1}^N$ and one of the reflection coefficients. For example, if we start with the right reflection coefficient $R(k,t)$ for $k \in \mathbf{R}$, we get

$$T(k,t) = \left(\prod_{j=1}^{N} \frac{k + i\kappa_j}{k - i\kappa_j}\right) \times \exp\left(\frac{1}{2\pi i}\int_{-\infty}^{\infty} ds\, \frac{\log(1 - |R(s,t)|^2)}{s - k - i0^+}\right)\,, \qquad k \in \mathbf{C}^+ \cup \mathbf{R}\,,$$

where the quantity $i0^+$ indicates that the value for $k \in \mathbf{R}$ must be obtained as a limit from $\mathbf{C}^+$. Then, the left reflection coefficient $L(k,t)$ can be constructed via

$$L(k,t) = -\frac{\overline{R(k,t)}\, T(k,t)}{\overline{T(k,t)}}\,, \qquad k \in \mathbf{R}\,.$$

We will see in the next section that $T(k,t) = T(k,0)$, $|R(k,t)| = |R(k,0)|$, and $|L(k,t)| = |L(k,0)|$.

For a detailed study of the direct scattering problem for the 1-D Schrödinger equation, we refer the reader to [5,7,8,9,16,17,18,19]. It is important to remember that $u(x,t)$ for $x \in \mathbf{R}$ at each fixed t is uniquely determined [5,7,8,9,16,17,18] by the scattering data $\{R, \{\kappa_j\}, \{c_{{\rm l}j}(t)\}\}$ or one of its equivalents. Letting $c_j(t) := c_{{\rm l}j}(t)^2$, we will work with one such data set, namely $\{R, \{\kappa_j\}, \{c_j(t)\}\}$, in Sect. "Time Evolution of the Scattering Data" and Sect. "Inverse Scattering Problem".

Having described the scattering data associated with the Schrödinger equation, let us briefly describe the scattering data associated with the Zakharov–Shabat system in (4). Assuming that $u(x,t)$ for each t is integrable in x on $\mathbf{R}$, the two Jost solutions $\psi(\lambda,x,t)$ and $\phi(\lambda,x,t)$, from the left and from the right, respectively, are those unique solutions to (4) satisfying the respective asymptotic conditions

$$\begin{aligned} \psi(\lambda,x,t) &= \begin{bmatrix} 0 \\ e^{i\lambda x} \end{bmatrix} + o(1), \quad x \to +\infty\,, \\ \phi(\lambda,x,t) &= \begin{bmatrix} e^{-i\lambda x} \\ 0 \end{bmatrix} + o(1)\,, \quad x \to -\infty\,. \end{aligned} \tag{25}$$

The transmission coefficient T, the left reflection coefficient L, and the right reflection coefficient R are obtained

via the asymptotics

$$\psi(\lambda,x,t)=\begin{bmatrix}\dfrac{L(\lambda,t)\,\mathrm{e}^{-i\lambda x}}{T(\lambda,t)}\\[2ex] \dfrac{\mathrm{e}^{i\lambda x}}{T(\lambda,t)}\end{bmatrix}+o(1)\,,\quad x\to-\infty\,,$$
$$\phi(\lambda,x,t)=\begin{bmatrix}\dfrac{\mathrm{e}^{-i\lambda x}}{T(\lambda,t)}\\[2ex] \dfrac{R(\lambda,t)\,\mathrm{e}^{i\lambda x}}{T(\lambda,t)}\end{bmatrix}+o(1)\,,\quad x\to+\infty\,. \tag{26}$$

The bound-state solutions to (4) occur at those λ values corresponding to the poles of T in $\mathbf{C}^+$. Let us use $\{\lambda_j\}_{j=1}^N$ to denote the set of such poles. It should be noted that such poles are not necessarily located on the positive imaginary axis. Furthermore, unlike the Schrödinger equation, the multiplicities of such poles may be greater than one. Let us assume that the pole λ_j has multiplicity n_j. Corresponding to the pole λ_j, one associates [4,20] n_j bound-state norming constants $c_{js}(t)$ for $s=0,1,\ldots,n_j-1$. We assume that, for each fixed t, the potential $u(x,t)$ in the Zakharov–Shabat system is uniquely determined by the scattering data $\{R,\{\lambda_j\},\{c_{js}(t)\}\}$ and vice versa.

Time Evolution of the Scattering Data

As the initial profile $u(x,0)$ evolves to $u(x,t)$ while satisfying the NPDE, the corresponding initial scattering data $S(\lambda,0)$ evolves to $S(\lambda,t)$. Since the scattering data can be obtained from the Jost solutions to the associated LODE, in order to determine the time evolution of the scattering data, we can analyze the time evolution of the Jost solutions with the help of the Lax method or the AKNS method.

Let us illustrate how to determine the time evolution of the scattering data in the Schrödinger equation with the help of the Lax method. As indicated in Sect. "The Lax Method", the spectral parameter k and hence also the values κ_j related to the bound states remain unchanged in time. Let us obtain the time evolution of $f_{\mathrm{l}}(k,x,t)$, the Jost solution from the left. From condition (ii) in Sect. "The Lax Method", we see that the quantity $\partial_t f_{\mathrm{l}}-\mathcal{A}f_{\mathrm{l}}$ remains a solution to (2) and hence we can write it as a linear combination of the two linearly independent Jost solutions f_{l} and f_{r} as

$$\partial_t f_{\mathrm{l}}(k,x,t)-\left(-4\partial_x^3+6u\partial_x+3u_x\right)f_{\mathrm{l}}(k,x,t)=p(k,t)\,f_{\mathrm{l}}(k,x,t)+q(k,t)\,f_{\mathrm{r}}(k,x,t)\,, \tag{27}$$

where the coefficients $p(k,t)$ and $q(k,t)$ are yet to be determined and $\mathcal{A}$ is the operator in (13). For each fixed t, assuming $u(x,t)=o(1)$ and $u_x(x,t)=o(1)$ as $x\to+\infty$ and using (21) and (22) in (27) as $x\to+\infty$, we get

$$\partial_t\mathrm{e}^{ikx}+4\partial_x^3\mathrm{e}^{ikx}=p(k,t)\,\mathrm{e}^{ikx}+q(k,t)\left[\frac{1}{T(k,t)}\,\mathrm{e}^{-ikx}+\frac{R(k,t)}{T(k,t)}\,\mathrm{e}^{ikx}\right]+o(1)\,. \tag{28}$$

Comparing the coefficients of e^{ikx} and e^{-ikx} on the two sides of (28), we obtain

$$q(k,t)=0\,,\quad p(k,t)=-4ik^3\,.$$

Thus, $f_{\mathrm{l}}(k,x,t)$ evolves in time by obeying the linear third-order PDE

$$\partial_t f_{\mathrm{l}}-\mathcal{A}f_{\mathrm{l}}=-4ik^3f_{\mathrm{l}}\,. \tag{29}$$

Proceeding in a similar manner, we find that $f_{\mathrm{r}}(k,x,t)$ evolves in time according to

$$\partial_t f_{\mathrm{r}}-\mathcal{A}f_{\mathrm{r}}=4ik^3f_{\mathrm{r}}\,. \tag{30}$$

Notice that the time evolution of each Jost solution is fairly complicated. We will see, however, that the time evolution of the scattering data is very simple. Letting $x\to-\infty$ in (29), using (22) and $u(x,t)=o(1)$ and $u_x(x,t)=o(1)$ as $x\to-\infty$, and comparing the coefficients of e^{ikx} and e^{-ikx} on both sides, we obtain

$$\partial_t T(k,t)=0\,,\quad \partial_t L(k,t)=-8ik^3L(k,t),$$

yielding

$$T(k,t)=T(k,0)\,,\quad L(k,t)=L(k,0)\,\mathrm{e}^{-8ik^3t}\,.$$

In a similar way, from (30) as $x\to+\infty$, we get

$$R(k,t)=R(k,0)\,\mathrm{e}^{8ik^3t}\,. \tag{31}$$

Thus, the transmission coefficient remains unchanged and only the phases of the reflection coefficients change as time progresses.

Let us also evaluate the time evolution of the dependency constants $\gamma_j(t)$ defined in (24). Evaluating (29) at $k=i\kappa_j$ and replacing $f_{\mathrm{l}}(i\kappa_j,x,t)$ by $\gamma_j(t)f_{\mathrm{r}}(i\kappa_j,x,t)$, we get

$$f_{\mathrm{r}}(i\kappa_j,x,t)\,\partial_t\gamma_j(t)+\gamma_j(t)\,\partial_t f_{\mathrm{r}}(i\kappa_j,x,t)-\gamma_j(t)\mathcal{A}f_{\mathrm{r}}(i\kappa_j,x,t)=-4\kappa_j^3\gamma_j(t)f_{\mathrm{r}}(i\kappa_j,x,t)\,. \tag{32}$$

On the other hand, evaluating (30) at $k = i\kappa_j$, we obtain

$$\gamma_j(t)\,\partial_t f_{\rm r}(i\kappa_j, x, t) - \gamma_j(t)\,\mathcal{A} f_{\rm r}(i\kappa_j, x, t) = 4\kappa_j^3 \gamma_j(t)\, f_{\rm r}(i\kappa_j, x, t)\,. \quad (33)$$

Comparing (32) and (33) we see that $\partial_t \gamma_j(t) = -8\kappa_j^3 \gamma_j(t)$, or equivalently

$$\gamma_j(t) = \gamma_j(0)\,{\rm e}^{-8\kappa_j^3 t}\,. \quad (34)$$

Then, with the help of (23) and (34), we determine the time evolutions of the norming constants as

$$c_{{\rm l}j}(t) = c_{{\rm l}j}(0)\,{\rm e}^{4\kappa_j^3 t}\,, \quad c_{{\rm r}j}(t) = c_{{\rm r}j}(0)\,{\rm e}^{-4\kappa_j^3 t}\,.$$

The norming constants $c_j(t)$ appearing in the Marchenko kernel (38) are related to $c_{{\rm l}j}(t)$ as $c_j(t) := c_{{\rm l}j}(t)^2$, and hence their time evolution is described as

$$c_j(t) = c_j(0)\,{\rm e}^{8\kappa_j^3 t}\,. \quad (35)$$

As for the NLS equation and other integrable NPDEs, the time evolution of the related scattering data sets can be obtained in a similar way. For the former, in terms of the operator $\mathcal{A}$ in (14), the Jost solutions $\psi(\lambda, x, t)$ and $\phi(\lambda, x, t)$ appearing in (25) evolve according to the respective linear PDEs

$$\psi_t - \mathcal{A}\psi = -2i\lambda^2\psi\,, \quad \phi_t - \mathcal{A}\phi = 2i\lambda^2\phi\,.$$

The scattering coefficients appearing in (26) evolve according to

$$\begin{aligned} T(\lambda, t) &= T(\lambda, 0)\,, \\ R(\lambda, t) &= R(\lambda, 0)\,{\rm e}^{4i\lambda^2 t}\,, \\ L(\lambda, t) &= L(\lambda, 0)\,{\rm e}^{-4i\lambda^2 t}\,. \end{aligned} \quad (36)$$

Associated with the bound-state pole λ_j of T, we have the bound-state norming constants $c_{js}(t)$ appearing in the Marchenko kernel $\Omega(y, t)$ given in (41). Their time evolution is governed [4] by

$$\begin{aligned} &\big[c_{j(n_j-1)}(t) \quad c_{j(n_j-2)}(t) \quad \dots \quad c_{j0}(t)\big] \\ &\quad = \big[c_{j(n_j-1)}(0) \quad c_{j(n_j-2)}(0) \quad \dots \quad c_{j0}(0)\big]\,{\rm e}^{-4iA_j^2 t}\,, \end{aligned} \quad (37)$$

where the $n_j \times n_j$ matrix A_j appearing in the exponent is defined as

$$A_j := \begin{bmatrix} -i\lambda_j & -1 & 0 & \dots & 0 & 0 \\ 0 & -i\lambda_j & -1 & \dots & 0 & 0 \\ 0 & 0 & -i\lambda_j & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -i\lambda_j & -1 \\ 0 & 0 & 0 & \dots & 0 & -i\lambda_j \end{bmatrix}.$$

Inverse Scattering Problem

In Sect. "Direct Scattering Problem" we have seen how the initial scattering data $S(\lambda, 0)$ can be constructed from the initial profile $u(x, 0)$ of the potential by solving the direct scattering problem for the relevant LODE. Then, in Sect. "Time Evolution of the Scattering Data" we have seen how to obtain the time-evolved scattering data $S(\lambda, t)$ from the initial scattering data $S(\lambda, 0)$. As the final step in the IST, in this section we outline how to obtain $u(x, t)$ from $S(\lambda, t)$ by solving the relevant inverse scattering problem. Such an inverse scattering problem may be solved by the Marchenko method [5,7,8,9,16,17,18,19]. Unfortunately, in the literature many researchers refer to this method as the Gel'fand–Levitan method or the Gel'fand–Levitan–Marchenko method, both of which are misnomers. The Gel'fand–Levitan method [5,7,16,17,19] is a different method to solve the inverse scattering problem, and the corresponding Gel'fand–Levitan integral equation involves an integration on the finite interval $(0, x)$ and its kernel is related to the Fourier transform of the spectral measure associated with the LODE. On the other hand, the Marchenko integral equation involves an integration on the semi-infinite interval $(x, +\infty)$, and its kernel is related to the Fourier transform of the scattering data.

In this section we first outline the recovery of the solution $u(x, t)$ to the KdV equation from the corresponding time-evolved scattering data $\{R, \{\kappa_j\}, \{c_j(t)\}\}$ appearing in (31) and (35). Later, we will also outline the recovery of the solution $u(x, t)$ to the NLS equation from the corresponding time-evolved scattering data $\{R, \{\lambda_j\}, \{c_{js}(t)\}\}$ appearing in (36) and (37).

The solution $u(x, t)$ to the KdV equation in (1) can be obtained from the time-evolved scattering data by using the Marchenko method as follows:

(a) From the scattering data $\{R(k, t), \{\kappa_j\}, \{c_j(t)\}\}$ appearing in (31) and (35), form the Marchenko kernel Ω defined via

$$\Omega(y, t) := \frac{1}{2\pi}\int_{-\infty}^{\infty} {\rm d}k\, R(k, t)\,{\rm e}^{iky} + \sum_{j=1}^{N} c_j(t)\,{\rm e}^{-\kappa_j y}\,. \quad (38)$$

(b) Solve the corresponding Marchenko integral equation

$$K(x, y, t) + \Omega(x + y, t) + \int_x^{\infty} {\rm d}z\, K(x, z, t)\,\Omega(z + y, t) = 0\,, \quad x < y < +\infty\,, \quad (39)$$

and obtain its solution $K(x, y, t)$.

(c) Recover $u(x,t)$ by using

$$u(x,t) = -2\,\frac{\partial K(x,x,t)}{\partial x}\,. \tag{40}$$

The solution $u(x,t)$ to the NLS equation in (3) can be obtained from the time-evolved scattering data by using the Marchenko method as follows:

(i) From the scattering data $\{R(\lambda,t),\{\lambda_j\},\{c_{js}(t)\}\}$ appearing in (36) and (37), form the Marchenko kernel Ω as

$$\Omega(y,t) := \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}\lambda\, R(\lambda,t)\,\mathrm{e}^{i\lambda y} + \sum_{j=1}^{N}\sum_{s=0}^{n_j-1} c_{js}(t)\frac{y^s}{s!}\,\mathrm{e}^{i\lambda_j y}\,. \tag{41}$$

(ii) Solve the Marchenko integral equation

$$K(x,y,t) - \overline{\Omega(x+y,t)} + \int_x^{\infty}\mathrm{d}z \times \int_x^{\infty}\mathrm{d}s\, K(x,s,t)\,\Omega(s+z,t)\,\overline{\Omega(z+y,t)} = 0\,, \qquad x<y<+\infty\,,$$

and obtain its solution $K(x,y,t)$.

(iii) Recover $u(x,t)$ from the solution $K(x,y,t)$ to the Marchenko equation via

$$u(x,t) = -2K(x,x,t)\,.$$

(iv) Having determined $K(x,y,t)$, one can alternatively get $|u(x,t)|^2$ from

$$|u(x,t)|^2 = 2\,\frac{\partial G(x,x,t)}{\partial x}\,,$$

where we have defined

$$G(x,y,t) := -\int_x^{\infty}\mathrm{d}z\,\overline{K(x,z,t)}\,\overline{\Omega(z+y,t)}\,.$$

Solitons

A soliton solution to an integrable NPDE is a solution $u(x,t)$ for which the reflection coefficient in the corresponding scattering data is zero. In other words, a soliton solution $u(x,t)$ to an integrable NPDE is nothing but a reflectionless potential in the associated LODE. When the reflection coefficient is zero, the kernel of the relevant Marchenko integral equation becomes separable. An integral equation with a separable kernel can be solved explicitly by transforming that linear equation into a system of linear algebraic equations. In that case, we get exact solutions to the integrable NPDE, which are known as soliton solutions.

For the KdV equation the N-soliton solution is obtained by using $R(k,t)=0$ in (38). In that case, letting

$$X(x) := \begin{bmatrix}\mathrm{e}^{-\kappa_1 x} & \mathrm{e}^{-\kappa_2 x} & \dots & \mathrm{e}^{-\kappa_N x}\end{bmatrix},$$

$$Y(y,t) := \begin{bmatrix} c_1(t)\,\mathrm{e}^{-\kappa_1 y}\\ c_2(t)\,\mathrm{e}^{-\kappa_2 y}\\ \vdots\\ c_N(t)\,\mathrm{e}^{-\kappa_N y}\end{bmatrix},$$

we get $\Omega(x+y,t) = X(x)\,Y(y,t)$. As a result of this separability the Marchenko integral equation can be solved algebraically and the solution has the form $K(x,y,t) = H(x,t)\,Y(y,t)$, where $H(x,t)$ is a row vector with N entries that are functions of x and t. A substitution in (39) yields

$$K(x,y,t) = -X(x)\,\Gamma(x,t)^{-1}\,Y(y,t), \tag{42}$$

where the $N\times N$ matrix $\Gamma(x,t)$ is given by

$$\Gamma(x,t) := I + \int_x^{\infty}\mathrm{d}z\,Y(z,t)\,X(z), \tag{43}$$

with I denoting the or $N\times N$ identity matrix. Equivalently, the (j,l)-entry of Γ is given by

$$\Gamma_{jl} = \delta_{jl} + \frac{c_j(0)\,\mathrm{e}^{-2\kappa_j x + 8\kappa_j^3 t}}{\kappa_j+\kappa_l}\,,$$

with δ_{jl} denoting the Kronecker delta. Using (42) in (40) we obtain

$$u(x,t) = 2\,\frac{\partial}{\partial x}\left[X(x)\,\Gamma(x,t)^{-1}\,Y(x,t)\right] = 2\,\mathrm{tr}\,\frac{\partial}{\partial x}\left[Y(x,t)\,X(x)\,\Gamma(x,t)^{-1}\right]\,,$$

where tr denotes the matrix trace (the sum of diagonal entries in a square matrix). From (43) we see that $-Y(x,t)\,X(x)$ is equal to the x-derivative of $\Gamma(x,t)$ and hence the N-soliton solution can also be written as

$$u(x,t) = -2\,\mathrm{tr}\,\frac{\partial}{\partial x}\left[\frac{\partial\Gamma(x,t)}{\partial x}\,\Gamma(x,t)^{-1}\right] = -2\,\frac{\partial}{\partial x}\left[\frac{\frac{\partial}{\partial x}\det\Gamma(x,t)}{\det\Gamma(x,t)}\right]\,, \tag{44}$$

where det denotes the matrix determinant. When $N=1$, we can express the one-soliton solution $u(x,t)$ to the KdV equation in the equivalent form

$$u(x,t) = -2\,\kappa_1^2\,\mathrm{sech}^2\left(\kappa_1 x - 4\kappa_1^3 t + \theta\right)\,,$$

with $\theta := \log\sqrt{2\kappa_1/c_1(0)}\,.$

Let us mention that, using matrix exponentials, we can express [6] the N-soliton solution appearing in (44) in various other equivalent forms such as

$$u(x,t) = -4C\mathrm{e}^{-Ax+8A^3t}\Gamma(x,t)^{-1}A\Gamma(x,t)^{-1}\mathrm{e}^{-Ax}B\,,$$

where

$$\begin{aligned} A &:= \mathrm{diag}\{\kappa_1,\kappa_2,\ldots,\kappa_N\}\,,\\ B^\dagger &:= \begin{bmatrix}1 & 1 & \ldots & 1\end{bmatrix},\\ C &:= \begin{bmatrix}c_1(0) & c_2(0) & \ldots & c_N(0)\end{bmatrix}\,. \end{aligned} \tag{45}$$

Note that a dagger is used for the matrix adjoint (transpose and complex conjugate), and B has N entries. In this notation we can express (43) as

$$\Gamma(x,t) = I + \int_x^\infty \mathrm{d}z\,\mathrm{e}^{-zA}BC\mathrm{e}^{-zA}\mathrm{e}^{8tA^3}\,.$$

As for the NLS equation, the well-known N-soliton solution (with simple bound-state poles) is obtained by choosing $R(\lambda,t)=0$ and $n_j=1$ in (41). Proceeding as in the KdV case, we obtain the N-soliton solution in terms of the triplet A, B, C with

$$A := \mathrm{diag}\{-i\lambda_1,-i\lambda_2,\ldots,-i\lambda_N\}, \tag{46}$$

where the complex constants λ_j are the distinct poles of the transmission coefficient in $\mathbf{C}^+$, B and C are as in (45) except for the fact that the constants $c_j(0)$ are now allowed to be nonzero complex numbers. In terms of the matrices $P(x,t)$, M, and Q defined as

$$\begin{aligned} &P(x,t) :=\\ &\quad \mathrm{diag}\{\mathrm{e}^{2i\lambda_1x+4i\lambda_1^2t},\mathrm{e}^{2i\lambda_2x+4i\lambda_2^2t},\ldots,\mathrm{e}^{2i\lambda_Nx+4i\lambda_N^2t}\},\\ &M_{jl} := \frac{i}{\lambda_j-\overline{\lambda}_l}, \quad Q_{jl} := \frac{-i\overline{c}_jc_l}{\overline{\lambda}_j-\lambda_l}. \end{aligned}$$

we construct the N-soliton solution $u(x,t)$ to the NLS equation as

$$u(x,t) = -2B^\dagger[I+P(x,t)^\dagger QP(x,t)\,M]^{-1}P(x,t)^\dagger C^\dagger\,, \tag{47}$$

or equivalently as

$$u(x,t) = -2B^\dagger\mathrm{e}^{-A^\dagger x}\Gamma(x,t)^{-1}\mathrm{e}^{-A^\dagger x+4i(A^\dagger)^2t}C^\dagger\,, \tag{48}$$

where we have defined

$$\begin{aligned} \Gamma(x,t) := I + &\left[\int_x^\infty \mathrm{d}s\,(C\mathrm{e}^{-As-4iA^2t})^\dagger(C\mathrm{e}^{-As-4iA^2t})\right]\\ &\times\left[\int_x^\infty \mathrm{d}z\,(\mathrm{e}^{-Az}B)(\mathrm{e}^{-Az}B)^\dagger\right]\,. \end{aligned} \tag{49}$$

Using (45) and (46) in (49), we get the (j,l)-entry of $\Gamma(x,t)$ as

$$\Gamma_{jl} = \delta_{jl} - \sum_{m=1}^N \frac{\overline{c}_jc_l\,\mathrm{e}^{i(2\lambda_m-\overline{\lambda}_j-\overline{\lambda}_l)x+4i(\lambda_m^2-\overline{\lambda}_j^2)t}}{(\lambda_m-\overline{\lambda}_j)(\lambda_m-\overline{\lambda}_l)}\,.$$

Note that the absolute square of $u(x,t)$ is given by

$$\begin{aligned} |u(x,t)|^2 &= \mathrm{tr}\left[\frac{\partial}{\partial x}\left(\Gamma(x,t)^{-1}\frac{\partial\Gamma(x,t)}{\partial x}\right)\right]\\ &= \frac{\partial}{\partial x}\left[\frac{\frac{\partial}{\partial x}\det\Gamma(x,t)}{\det\Gamma(x,t)}\right]\,. \end{aligned}$$

For the NLS equation, when $N=1$, from (47) or (48) we obtain the single-soliton solution

$$u(x,t) = \frac{-8\overline{c}_1(\mathrm{Im}[\lambda_1])^2\,\mathrm{e}^{-2i\overline{\lambda}_1x-4i(\overline{\lambda}_1)^2t}}{4(\mathrm{Im}[\lambda_1])^2+|c_1|^2\,\mathrm{e}^{-4x(\mathrm{Im}[\lambda_1])-8t(\mathrm{Im}[\lambda_1^2])}}\,,$$

where Im denotes the imaginary part.

Future Directions

There are many issues related to the IST and solitons that cannot be discussed in such a short review. We will briefly mention only a few.

Can we characterize integrable NPDEs? In other words, can we find a set of necessary and sufficient conditions that guarantee that an IVP for a NPDE is solvable via an IST? Integrable NPDEs seem to have some common characteristic features [1] such as possessing Lax pairs, AKNS pairs, soliton solutions, infinite number of conserved quantities, a Hamiltonian formalism, the Painlevé property, and the Bäcklund transformation. Yet, there does not seem to be a satisfactory solution to their characterization problem.

Another interesting question is the determination of the LODE associated with an IST. In other words, given an integrable NPDE, can we determine the corresponding LODE? There does not yet seem to be a completely satisfactory answer to this question.

When the initial scattering coefficients are rational functions of the spectral parameter, representing the time-evolved scattering data in terms of matrix exponentials results in the separability of the kernel of the Marchenko integral equation. In that case, one obtains explicit formulas [4,6] for exact solutions to some integrable NPDEs and such solutions are constructed in terms of a triplet of constant matrices A, B, C whose sizes are $p\times p$, $p\times 1$, and $1\times p$, respectively, for any positive integer p. Some special

cases of such solutions have been mentioned in Sect. "Solitons", and it would be interesting to determine if such exact solutions can be constructed also when p becomes infinite.

Bibliography

Primary Literature

1. Ablowitz MJ, Clarkson PA (1991) Solitons, nonlinear evolution equations and inverse scattering. Cambridge University Press, Cambridge
2. Ablowitz MJ, Kaup DJ, Newell AC, Segur H (1973) Method for solving the sine–Gordon equation. Phys Rev Lett 30:1262–1264
3. Ablowitz MJ, Kaup DJ, Newell AC, Segur H (1974) The inverse scattering transform-Fourier analysis for nonlinear problems. Stud Appl Math 53:249–315
4. Aktosun T, Demontis F, van der Mee C (2007) Exact solutions to the focusing nonlinear Schrödinger equation. Inverse Problems 23:2171–2195
5. Aktosun T, Klaus M (2001) Chapter 2.2.4, Inverse theory: problem on the line, In: Pike ER, Sabatier PC (eds) Scattering. Academic Press, London, pp 770–785
6. Aktosun T, van der Mee C (2006) Explicit solutions to the Korteweg-de Vries equation on the half-line. Inverse Problems 22:2165–2174
7. Chadan K, Sabatier PC (1989) Inverse problems in quantum scattering theory. 2nd edn. Springer, New York
8. Deift P, Trubowitz E (1979) Inverse scattering on the line. Commun Pure Appl Math 32:121–251
9. Faddeev LD (1967) Properties of the *S*–matrix of the one–dimensional Schrödinger equation. Amer Math Soc Transl (Ser. 2) 65:139–166
10. Fermi E (1965) Collected papers, vol II: United States, 1939–1954. University of Chicago Press, Chicago
11. Fermi E, Pasta J, Ulam S (1955) Studies of non linear problems, I. Document LA-1940, Los Alamos National Laboratory
12. Gardner CS, Greene JM, Kruskal MD, Miura RM (1967) Method for solving the Korteweg–de Vries equation. Phys Rev Lett 19:1095–1097
13. Gel'fand IM, Levitan BM (1955) On the determination of a differential equation from its spectral function. Amer Math Soc Transl (Ser. 2) 1:253–304
14. Korteweg DJ, de Vries G (1895) On the change of form of long waves advancing in a rectangular channel and on a new type of long stationary waves. Phil Mag 39:422–443
15. Lax PD (1968) Integrals of nonlinear equations of evolution and solitary waves. Commun Pure Appl Math 21:467–490
16. Levitan BM (1987) Inverse Sturm–Liouville problems. Science VNU Press, Utrecht
17. Marchenko VA (1986) Sturm–Liouville operators and applications. Birkhäuser, Basel
18. Melin A (1985) Operator methods for inverse scattering on the real line. Commun Partial Differ Equ 10:677–766
19. Newton RG (1983) The Marchenko and Gel'fand–Levitan methods in the inverse scattering problem in one and three dimensions. In: Bednar JB, Redner R, Robinson E, Weglein A (eds) Conference on inverse scattering: theory and application. SIAM, Philadelphia, pp 1–74
20. Olmedilla E (1987) Multiple pole solutions of the nonlinear Schrödinger equation. Phys D 25:330–346
21. Russell JS (1845) Report on waves, Report of the 14th meeting of the British Association for the Advancement of Science. John Murray, London, pp 311–390
22. Wadati M (1972) The exact solution of the modified Korteweg–de Vries equation. J Phys Soc Jpn 32:1681
23. Zabusky NJ, Kruskal MD (1965) Interaction of "solitons" in a collisionless plasma and the recurrence of initial states. Phys Rev Lett 15:240–243
24. Zakharov VE, Shabat AB (1972) Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media. Soviet Phys JETP 34:62–69
25. http://www.osti.gov/accomplishments/pdf/A80037041/A80037041.pdf

Books and Reviews

Ablowitz MJ, Segur H (1981) Solitons and the inverse scattering transform. SIAM, Philadelphia

Aktosun T (2004) Inverse scattering transform, KdV, and solitons, In: Ball JA, Helton JW, Klaus M, Rodman L (ed), Current trends in operator theory and its applications. Birkhäuser, Basel, pp 1–22

Aktosun T (2005) Solitons and inverse scattering transform, In: Clemence DP, Tang G (eds) Mathematical studies in nonlinear wave propagation, Contemporary Mathematics, vol 379. Amer Math Soc, Providence, pp 47–62

Dodd RK, Eilbeck JC, Gibbon JD, Morris HC (1982) Solitons and nonlinear wave equations. Academic Press, London

Drazin PG, Johnson RS (1988) Solitons: an introduction. Cambridge University Press, Cambridge

Lamb GL Jr (1980) Elements of soliton theory. Wiley, New York

Novikov S, Manakov SV, Pitaevskii LP, Zakharov VE (1984) Theory of solitons. Consultants Bureau, New York

Scott AC, Chu FYF, McLaughlin D (1973) The soliton: a new concept in applied science. Proc IEEE 61:1443–1483

Investment Decision Making in Finance, Models of

Jørgen Vitting Andersen
Institut Non Linéaire de Nice (UMR CNRS 6618), Université de Nice-Sophia Antipolis, Valbonne, France

Article Outline

Glossary
Definition of the Subject
Introduction
Rational Decision Making
Irrational Decision Making
Investment Decision Making in a Complex Environment
Future Directions
Bibliography

Glossary

Decision making A human cognitive process which leads to a course of action among a set of choices.

Rational expectations Expectations that give the best guess of the future (the optimal forecast) using all available information.

Behavioral finance A research field that applies human and social cognitive and emotional biases to understanding how decision making affects market prices, returns and allocation of resources.

Definition of the Subject

A general definition of "decision making" is that it is a human cognitive process which leads to a course of action among a set of choices. A decision process gives rise to a choice which can be an action or an opinion. In the context of modeling investment decision making in finance, by far the majority of models assume that decision making is a reasoning process which is *rational*. Classic examples of rational decision models in finance describe how risk-willing investors should react to ensure a certain return, and have been formalized in models such as the Markovitz mean-variance formulation [26], the Capital Asset Price Model [24,34] as well as the Arbitrage Pricing Theory (all defined in this article). Rational decision making is at the cornerstone of the foundation for modern financial thinking. Nonetheless, over the last two decades or so, models of irrational reasoning have led to the appearance of a new field in finance called "Behavioral Finance" (also defined in this article). More recently, decision making in a complex environment has been a very active research field, notably in interdisciplinary approaches drawing on both physics and finance.

Introduction

Historically, *rational* decision making has been at the very core of financial models describing how one should deal with a variety of problems in finance such as, for example, portfolio allocation [26], pricing of financial assets [24,34], valuation of firms and of capital costs [27], pricing of options [8] and even (somewhat less intuitively) the creation of financial bubbles [39]. The aforementioned examples all led to Nobel prizes in economy: in 1985 (valuation of firms), in 1990 (portfolio theory and asset pricing), in 1997 (new method to price derivatives) and in 2002 (for laboratory experiments as a tool in empirical economic analysis).

The biggest advantage about the assumption of rational expectations is clearly that it allows one to progress and actually do some calculations on how to do investment decision making in finance, as illustrated by the mentioned Nobel prizes in economy. The disadvantage is, however, that it is not clear at all whether one thereby has obtained a framework that to any degree describes the reality of financial markets. As will be seen Sect. "Critiques of Markovitz Portfolio Theory and the CAPM", the strongest objection against rational expectations is the assumption that *all* market participants, if given access to a complete information set, would come to same conclusion with regard to price.

Questions about the validity of the assumption of rational expectations in decision making and the theories based upon it, had already surfaced in the 1980s. Criticism was notably raised by a succession of discoveries of anomalies, particularly the evidence of excess volatility of returns [36]. Since the reported excess volatility raised some of the very first empirically founded questions related to the efficient market theory, it became the subject of harsh academic disputes. Following the problems made clear by the puzzle of "excess volatility", competing theories of rational decision making and the efficient market surfaced in the 1990s, most notably in the field now known as behavioral finance. The emerging field of behavioral finance in turn lead to more recent theories which view financial markets as a complex system composed of many actors with different goals and often lead by irrational decision making.

Rational Decision Making

In the following, the concept of rational expectations in finance will be defined in terms of how one should determine the price of one single asset, depending on the money one is going to receive in the future for holding that asset. As will be seen, in order to know today's price of an asset, the money that one is going to receive in one year's time, two year's time etc. (called the future cash flow) has to be expressed in the present time value of money, that is, it has to be discounted. After the introduction of the concept of rational expectations has been applied to one single asset, the concept will be generalized to also take into account decision making based on risk, encountered in a situation in which one has to deal with a portfolio of assets.

Rational Expectations

According to the efficient market theory, the price of an asset at time t, P_t, should equal the expectation value conditioned on all available information, I_t, at time t of all future cash flows (coming from the asset) discounted to the present time value of money. If we call P_t^* the sum over all

future discounted cash flows, one has:

$$P_t = E_t(P_t^* | I_t) \tag{1}$$

with E_t denoting the expectation value. Rational expectations theory, therefore, defines the price as being identical to the best guess of the future (the optimal forecast) that uses all available information. This amounts to assuming that people do not make systematic errors when predicting the future, and that deviations from perfect foresight are only random. One should notice that P^* is *not* known at time t, but has to be forecast in order to assign a value to the price of the asset at time t, P_t. Note that (1) dictates that all movements in the stock market must have their origin in some new information about the fundamental value P_t^*, one of the facts of the efficient market theory that seems hard to believe in practice.

Equivalently one can write (1) as

$$P_t = \sum_{\tau=t+1}^{\infty} \frac{D(\tau)}{(1 + r(\tau))^{(\tau-t)}} \tag{2}$$

where $D(\tau)$ is the cash flow (dividends) at time τ, the factor $(1 + r(\tau))^{(\tau-t)}$ is the discount factor transforming money at time τ into present value money at time t, and $r(\tau)$ is the interest rate. That is, (2) says that the price of an asset today should be calculated from the sum over all future cash flows discounted to the present day value of money. One should notice that despite an apparently elegant and appealing explanation offered by (1),(2), the equations have profound implications that one has to counterbalance against a more realistic description of facts.

In fact, under strict assumptions of efficiency, the financial markets would be "dead" without any trades occurring. As seen from (2) the price of an asset should change only with the arrival of new information of either future dividends or interest rates. One way out of this apparent paradox is to notice that risk is *not* taken into account in the derivation above. Informationally efficient markets are an impossibility, because if markets are perfectly efficient, the return of gathering information would be nil. Therefore the degree of market inefficiency could be seen as due to the cost of gathering information as well as due to investors being risk-loving or risk-averse.

The strongest objection against rational expectations theory is, however, the assumption that *all* market participants, if given access to a complete information set, would come to same conclusion with regard to price as, e. g., illustrated by (2). This seems very far from reality, to say the least. For more detail on this discussion see Sect. "Critiques of Markovitz Portfolio Theory and the CAPM".

The remarkable fact is that rational expectations theory is the key building block, the very foundation that underpins modern portfolio and pricing theories like the CAPM, APT and versions thereof. Thinking that most risk management theories use rational expectations at their core seems somehow a paradox in that it amounts to making a very risky ansatz at the very foundation!

As noted in [30,33], since rational expectations have to be unchanged in all futures (it is without cost to change an expectation therefore if an expectation is expected to change in the future, rationality will require a revision now) rationality requires that for any $0 < j < n$:

$$E_t[E_{t+j}(P_{t+n})] = E_t(P_{t+n}) \,. \tag{3}$$

Ex post (3) takes the following form for the special case $j = 1$:

$$E_{t+1} = E_t + \epsilon_{t+1} \tag{4}$$

where ϵ_{t+1} has mean zero and covariance $\mathrm{cov}(\epsilon_{t+j}, \epsilon_t) = 0$ for $j \neq 0$. When the process ϵ is independent and identically distributed, (4) describes a random walk which is non-stationary. Therefore non-stationarity can be seen as a direct consequence of rationality [30].

Markovitz Portfolio Theory and the Capital Asset Price Model (CAPM)

One of the most fundamental tenets of economic theory and practice is that returns above the so-called riskless rate come with increased risks. This is the basis of Markovitz's portfolio theory [26] and of the Capital Asset Pricing Model (CAPM) [24,34]. Reciprocally, investors want to be compensated for taking risk, that is, they want to earn a return high enough to make them comfortable with the level of risk they are assuming.

Markovitz's portfolio theory considers investments over *one* time period, i. e., an investor selects a portfolio at time t that produces a return at time $t + 1$. It is implicitly assumed that the probability distribution function of the returns of the assets in the portfolio do not change between time t and $t + 1$. In Markovitz's model, investors are risk-averse, and when choosing among portfolios they take into account only the return of the portfolio (the mean) and the risk (estimated by the variance) of their one-period investment. To illustrate in practice how Markovitz's method works, imagine for a moment the decision an investor faces by investing a fixed amount of money in, say, the 40 assets of the French CAC40 index. One investment decision could be to put 1% of the fixed investment amount into asset 1, 3% into asset 2, ..., 4% into asset 40. Calculating the return and risk over one time period (say a year)

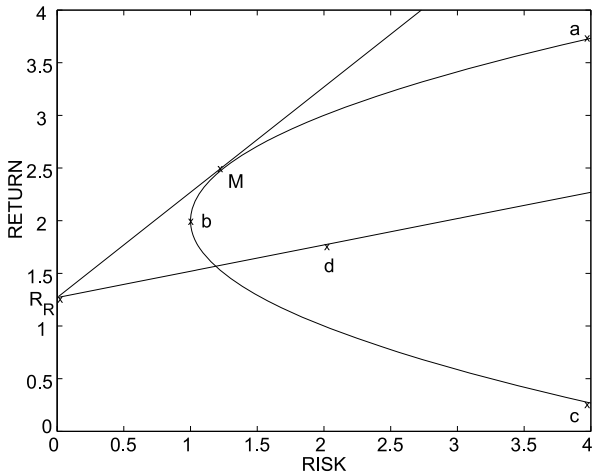

Investment Decision Making in Finance, Models of, Figure 1
Risk versus return of a portfolio

for this specific choice of portfolio would give *one* point, say the point *d*, within the parabola bounded by "*a*–*b*–*c*" in Fig. 1. Another decision choice would, of course, give rise to a different point within this risk-return diagram, and so by probing all possible decision choices one probes the area of the parabola bounded by "*a*–*b*–c". Now assuming all investors are rational, they all see the same optimal choice of investment, namely the so called "efficient frontier", which consists of the part *a*–*b* on the parabola. The part *a*–*b* on the parabola corresponds to all those portfolios which, for a given fixed value of risk, give the largest return or, equivalently, for a given fixed value of return (mean) have the smallest amount of risk (variance). Therefore, the Markovitz approach is often called the standard "mean-variance" approach.

Sharpe and Lintner added a key assumption to the work of Markovitz by assuming borrowing and lending at a risk free rate which is the same for all investors and does not depend on the amount borrowed or lent. The introduction of borrowing/lending at a risk free rate by Lintner and Sharpe turns the efficient set into a straight line [16]. If all funds are invested in the risk free asset, i. e., all funds are lent at the risk free rate R_r, one gets a portfolio at the point R_r in Fig. 1, i. e., a portfolio with zero variance and a risk free rate of return R_r. Instead, progressively investing a percentage $1 - x$ of all funds into the risky assets, one would follow the line $R_r - d$ with the point d corresponding to being invested 100% in a portfolio of risky assets. Points to the right of d correspond to investing more than 100% by borrowing at the risk free rate. Formally this can be written:

$$R_p = xR_r + (1 - x)R_d \tag{5}$$

$$E(R_p) = xR_r + (1 - x)E(R_d) \tag{6}$$

$$\sigma(R_p) = (1 - x)\sigma(R_d) \tag{7}$$

with R_p denoting the return of the portfolio, E the expectation value and σ the standard deviation. But the portfolio d in Fig. 1 is not mean-variance-efficient. In order to get a mean-variance-efficient portfolio with the largest return per unit of risk free borrowed money (the largest slope of the line), one swings the line from R_r up and to the left as far as possible, which gives the tangency portfolio M. Since all investors agree completely about the distribution of returns, they all see the same optimal solution and combine the same risky tangency portfolio M with risk free lending/borrowing. Because all investors have the same portfolio M, this has to be the value weighted market portfolio, that is, each asset's weight in the market portfolio M must be given by the total market value of all outstanding units of the asset divided by the total market value of all risky assets. This, then, in turn gives a benchmark for the returns of each individual asset, since an asset which has the same time series of returns as the market portfolio M should give exactly the same return as the market. Stated in another way:

$$E(R_i) = E(R_r) + [E(R_M) - E(R_r)]\beta_i, \quad i = 1, \ldots, N \tag{8}$$

$$\beta_i = \frac{\mathrm{cov}(R_i, R_M)}{\sigma(R_M)} \tag{9}$$

where $\mathrm{cov}(R_i, R_M)$ is the covariance of the return of asset i with the market return, and $\sigma(R_M)$ is the variance of the market return. Eq. (9) is the famous CAPM relation, which allows us to price an asset, $E(R_i)$, from the knowledge of a certain "market portfolio" and the risk free return R_r. The "market portfolio" should, in principle, include not only all traded financial assets, but also consumer durables, real estate and human capital, a task impossible to estimate in practice. Even a more limited view of the "market portfolio" as consisting only of financial assets would still mean including all quoted assets worldwide, an exercise hardly manageable in practice. Instead, a typical choice in the financial literature is to take U.S. common stocks, but this is, strictly speaking, not a "market portfolio" in the CAPM sense.

Finally, we should mention another influential asset pricing theory that also uses the assumption of rational expectations. The Arbitrage Pricing Theory (APT) was introduced by the economist Stephen Ross in 1976 [31], and can be seen as an extension of the CAPM, since the price of an asset is given by a number of economical factors (in contrast to *one* factor in the CAPM given by the market).

Specifically, the APT assumes that the returns of a given asset are determined by a number of macroeconomic factors F_j (assumed to be random variables with zero mean)

$$E(R_i) = R_r + \sum_{j=1}^{n} \beta_{i,j} P_j \tag{10}$$

$$R_i = E(R_i) + \sum_{j=1}^{n} \beta_{i,j} F_j + \epsilon_i \tag{11}$$

with P_j the risk premium of factor j, $\beta_{i,j}$ the sensitivity of asset i to factor j and ϵ_i an idiosyncratic noise term with mean zero. (11) says that the uncertainty of the return of asset i is given by a linear combination of n economic factors. In the APT, the portfolio choice is no longer unique, since different investors will hold portfolios with specific choices of $\beta_{i,j}$. In this sense, the assumptions of the APT are closer to real investment strategies, but like the CAPM it also suffers from the crucial claim of rational expectations, see the discussion in Sect. "Critiques of Markovitz Portfolio Theory and the CAPM".

Critiques of Markovitz Portfolio Theory and the CAPM

One problem within the framework of CAPM and of Markovitz's portfolio theory is the notion of risk, which is described by the standard deviation of the return. This is a one dimensional measure, whereas the fundamental decision problem in principle has an infinite number of dimensions, given only by full knowledge of the probability distribution of the portfolio returns.

In order to illustrate this point, consider Fig. 2 which illustrates a portfolio of just two assets. The topmost plot correspond to a portfolio chosen according to Markovitz's theory, whereas the plot in the middle row shows a portfolio which has been chosen so that it does not correspond to minimum variance, but rather to the minimum of higher order moments (for a detailed description see [2]). Comparing the two plots, it is clear by simple visual inspection that the Markovitz portfolio does a good job *most* of the time in ensuring that only small risks (i. e., fluctuations of the portfolio) are taken. Similarly, choosing the weights in the portfolio by minimizing higher order moments, it is clear that one opens oneself up for taking more small and intermediate size risks (seen by, in average, larger fluctuations compared to the Markovitz solution). However, and this is the important point, by minimizing the higher order moments of the portfolio distribution, one minimizes large events that takes place in the tail of the distribution. So by accepting more small and intermediate risks compared to the Markovitz solution, one avoids taking the *big* risks. Specifically, this can be seen by the band contained by the dotted lines, which are the same in the topmost and middle plots. Even though the portfolio showing the Markovitz solution is constructed so that small and intermediate risks are contained *most* of the time, it actually gives a false impression of control, since one thereby becomes susceptible to the big risks seen by the events that exceed the bounds of the dotted lines in the topmost plot. On the contrary, accepting more small and intermediate risks, as for the portfolio seen in the middle plot, one steers free of the same big risks that were encountered in Markovitz's solution. It is remarkable that, by decreasing large risks, one can, at the same time, profit from such a choice of portfolio. This is seen in the lower plot of Fig. 2, where a fourfold gain in comparison with the Markovitz solution is obtained for the portfolio that limits large risks.

The strongest objection against the CAPM and the Markovitz solution, however, lies in its foundation, since it is assumed that *all* market participants if given access to a complete information set would come to the same conclusion with regard to price. That is, there is complete agreement among investors about the joint distribution of asset returns from t to $t + 1$. In physics, such an assumption goes under the name of a "mean-field" theory, a somewhat surprising application for a system (the financial market) whose complexity by far exceeds the most complex system encountered in the realm of physics. The fact that the hypothesis of rational expectations, with its implication of "complete agreement among investors", lays the groundwork for much of the modern work done in the theory of finance, seems problematic. A recent and very clear criticism of this point is illustrated by the following excerpts from an article by Phillip Ball (editor of Nature) published in Financial Times, 29th of October 2006:

> *It is easy to mock economic theory. Any fool can see that the world of neoclassical economics, which dominates the academic field today, is a gross caricature in which every trader or company acts in the same self-interested way rational, cool, omniscient. The theory has not foreseen a single stock market crash and has evidently failed to make the world any fairer or more pleasant.*
>
> *The usual defense is that you have to start somewhere. But mainstream economists no longer consider their core theory to be a start. The tenets are so firmly embedded that economists who think it is time to move beyond them are cold-shouldered. It is a rigid dogma. To challenge these ideas is to invite blank stares of incomprehension you might as well be telling a physicist that gravity does not exist.*

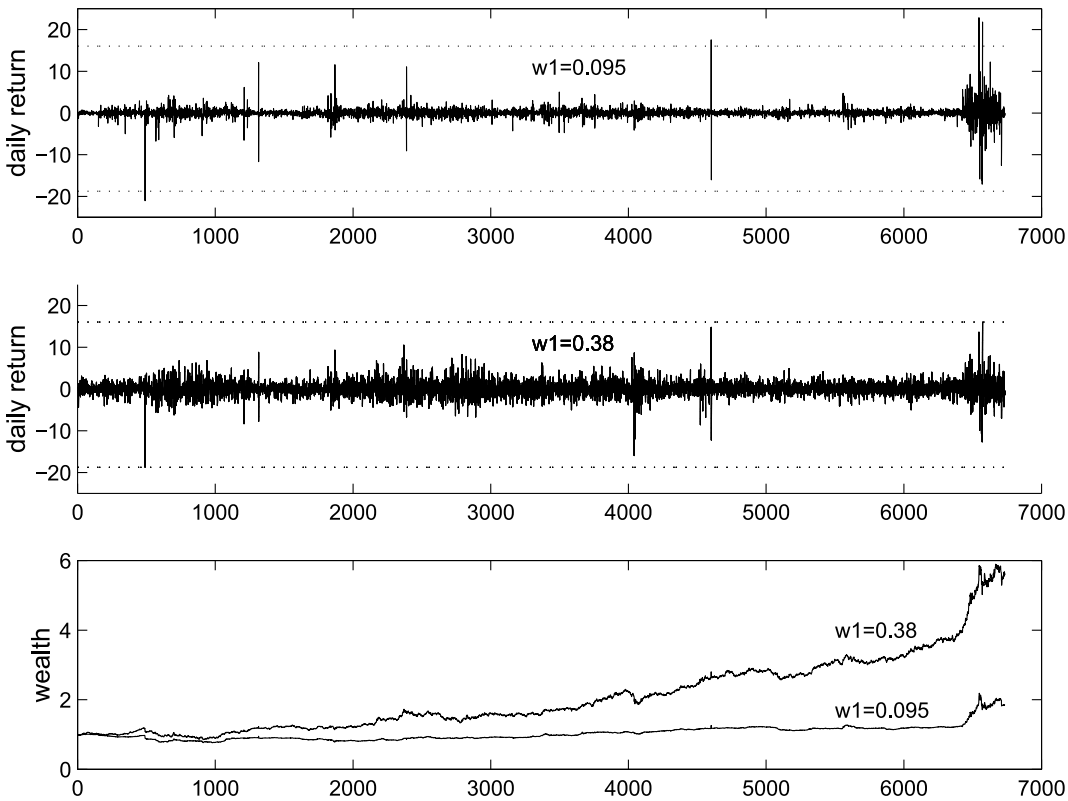

Investment Decision Making in Finance, Models of, Figure 2
Daily returns and cumulative wealth (in percentage) for a portfolio composed of two assets, Chevron stock (with weight w_1) and Malaysian ringgit (with weight $1 - w_1$). The topmost plot shows daily return for a Markovitz mean-variance portfolio ($w_1 = 0.095$), whereas the second plot shows returns for a portfolio that minimizes higher order moments (for a detailed explanation see [2]). Cumulative wealth for both choices of portfolio is shown in the bottom plot

That is disturbing because these things matter. Neoclassical idiocies persuaded many economists that market forces would create a robust post-Soviet economy in Russia (corrupt gangster economies do not exist in neoclassical theory). Neoclassical ideas favouring unfettered market forces may determine whether Britain adopts the euro, how we run our schools, hospitals and welfare system. If mainstream economic theory is fundamentally flawed, we are no better than doctors diagnosing with astrology . . .

Paul Ormerod, author of The Death of Economics, argues that one of the most limiting assumptions of neoclassical theory is that agent behavior is fixed: people in markets pursue a single goal regardless of what others do. The only way one person can influence another's choices is via the indirect effect of trading on prices. Yet it is abundantly clear that herding – irrational, copycat buying and selling – provokes market fluctuations.

There are ways of dealing with the variety and irrationality of real agents in economic theory. But not in mainstream economics journals, because the models defy neoclassical assumptions.

There is no other "science" in such a peculiar state. A demonstrably false conceptual core is sustained by inertia alone. This core, "the Citadel", remains impregnable while its adherents fashion an increasingly baroque fantasy. As Alan Kirman, a progressive economist, said: "No amount of attention to the walls will prevent the Citadel from being empty."

Irrational Decision Making

Whereas rational decision making, especially with the idea of the efficient market theory, was well established in the academic world of finance in the 1970s, questions about the validity of the theory surfaced in the 1980s, prompted by a succession of discoveries of anomalies, notably the

evidence of excess volatility of returns [36]. Since the reported excess volatility raised some of the very first empirically founded questions related to the efficient market theory, it became the subject of harsh academic disputes. The idea behind it will be explained in detail in this section. Following the problems made clear by the puzzle of "excess volatility", competing theories to rational decision making and the efficient market theory surfaced in the 1990s, most notably by the field now known as behavioral finance. The emerging field of behavioral finance, in turn, led to more recent theories such as feedback and herding, as well as models of interaction between smart money and ordinary investors.

"Excess Volatility" Created by Irrational Decision Making

To illustrate the problem of so-called "excess volatility", notice that it follows from (1) that $P_t^* = P_t + U_t$ where U_t is a forecast error. In this equation P_t is known at time t, whereas U_t, P_t^* are not known. When making a forecast at time t, one is assumed to have access to all available information. Since the forecast is assumed to be optimal, U_t therefore has to be uncorrelated with any information variable at time t. P_t, however, is itself known information at time t, so P_t and U_t must be uncorrelated. But this means that

$$\sigma^2(P_t^*) = \sigma^2(P_t) + \sigma^2(U_t) \tag{12}$$

where σ^2 denotes the variance (the standard deviation squared) of a variable. Since the variance of a variable is always positive, this means that:

$$\sigma^2(P_t^*) > \sigma^2(P_t)\,. \tag{13}$$

(13) gives a "handle" to test the efficient market theory. We can use historical price data to construct P_t^* from the actual dividend paid from time t till the present time, discounted by the actual interest rates, and then compare the variance of P_t^* to the variance of the price P_t over the same time period. In ([37]), historical data was analyzed in this manner from year $t = 1871$ till year 2002 of the Standard & Poor's Composite Stock Price Index. Clear evidence showed that (13) was violated, thus giving rise to a remarkable amount of controversy within the academic community. This so called "excess volatility" is still one of the often debated "puzzles" in finance, and gave one of the first clear empirical demonstrations that one must go beyond the rational decision making framework to gain a proper understanding of financial data.

Example of a Model That Can Not Be Solved Using Rational Expectations

It should be noted that even within a theoretical framework, some problems are simply not solvable using rational expectations. This was pointed out by B. Arthur [5] in what has been coined "the El Farol bar problem". The El Farol bar problem is a game in which N people must decide independently each week whether to show up (or not) at their favorite bar, with the constraint that the bar has only $L < N/2$ chairs. Since each person will only be happy when seated, people use last week's attendance to predict future attendance. If a person predicts that the number of people that will attend the bar is larger than L, that person will stay home. It is important to notice that the game describes adaptive behavior, as the people adapt their behavior to prior attendance. Also, it gives a clear illustration of a case where rational expectations *cannot* be used to find a solution. This can be seen as follows: suppose that a rational expectation prediction machine existed and that all agents possessed a copy of it. If, in the example given, the machine predicts that a number larger than L will attend the bar, nobody will show up, thereby negating the prediction of the prediction machine. The El Farol bar problem, in turn, led to a new parsimonious formulation of financial markets seen as a complex system and now known under the name "the Minority Game", see Sect. "Investment Decision Making in a Complex Environment.

Behavioral Finance

Decision making became a research topic in the field of psychology in the 1950s in the work of Edwards [15] as well as that of H. A. Simon, who introduced the concept of decision making based on *bounded* rationality [38]. It was, however, not until the work of Daniel Kahneman and Amos Tversky (deceased in 1996) that results from cognitive psychology found their way into economy and finance.

Prospect Theory Prospect theory, introduced by Kahneman and Tversky [22] (and for which Kahneman received the Nobel prize in 2002), takes into account the fact that people in general are unable to fully analyze situations that involve economic and probabilistic judgments. Specifically, prospect theory makes three assumptions which deviate from the standard neoclassical framework of economic rational decision making:

- When investing, people are often sensitive to some reference level. Say an investment decision is likely to be influenced by the absolute wealth possessed by an in-

vestor. The satisfaction/dissatisfaction by a gain/loss of an investor is related to the wealth of an investor and not just the absolute amount gained/lost.

- A rational person would act symmetrically with respect to gains versus losses of same size, that is, a rational person would be equally pleased/annoyed. People, however, appear to be more averse to losses than attracted by gains of the same size. This appears to be especially true for small losses/gains: even the slightest loss is conceived as very bad, whereas a small gain is not considered equally satisfying.
- People have the tendency to overweigh events that occur very rarely and to underweigh events which occur with large probabilities. E.g., people keep on buying lottery tickets despite the fact that their chance of winning is almost nil!

Anchoring and Other Irrational Beliefs Used in Decision Making Anchoring is a term used in psychology to describe the case in which human decision making relies (anchors) on just one piece of (often irrelevant) information. One of the first observations of anchoring was reported in an experiment by Tversky and Kahneman. In [41], two groups of test persons were shown to give different mean estimates of the percentage of African nations in the United Nations, depending on the specific anchor of percentage suggested by the experimenters to the two groups. Evidence for human anchoring has since been reported in many completely different domains such as e. g., customer inertia in brand switching [44] (old brand price acting as an anchor), whereas other evidence comes from studies on on-line auctions [13] (people bid more for an item the higher the "buy-now" price) and anchoring in real estate prices [28] (subjects' appraisal values depend on an arbitrary posted listing price of the house). In the context of financial markets, anchoring has been observed via the so called "disposition effect" [35], [21], which is the tendency for people to sell assets that have gained value and keep assets that have lost value. As noted in [43], conclusive tests using real market data are usually difficult because the investors' expectations, as well as individual decisions, cannot be controlled or easily observed. In experimental security trading, however, subjects were observed to sell winners and keep losers [43].

In [1], another method was analyzed which rigorously tests for anchoring in the decision making of humans when they trade in the stock market. The test, heavily inspired by interdisciplinary approaches, was introduced in a trading algorithm by L. Gil [20], which was inspired by the way biological motors work by exploiting favorable Brownian fluctuations to generate directed forces and move[1]. In [1], weekly data of the CAC40 and the Dow Jones index were analyzed to test for the claim that market participants, by actively following the price, thereby create a subjective reference (anchor) and memory of when an asset is "cheap" or "expensive". Several studies on the persistence of human memory have reported that sleep, as well as post-training wakefulness before sleep, play an important role in the offline processing and consolidation of memory [29]. It therefore makes sense to think that conscious as well as unconscious mental processes influence the judgments of people who specialize in active trading on a day-to-day basis. The out-of-sample profit from the market-neutral trading algorithm (with transaction costs taking into account) on the CAC40 index as well on the Dow Jones index, gives evidence that anchoring does, indeed, play a dominant role on the weekly pricing of the Dow Jones and CAC40 stock markets.

Examples of other human beliefs that can lead to irrational decision making are (see e. g., ([7]) for a longer discussion)

- **Overconfidence.** People are overconfident when it comes to decision making. A typical bias is that people have the tendency to ascribe any good outcomes to their own talents, while blaming bad outcomes on external circumstances.
- **Hindsight bias.** The tendency to think one predicted what had happened, whereas in reality it was only realized after the fact.
- **Framing.** Refers to the case where the *exact same problem* has different outcomes in decision making, depending on how the problem is described.
- **Confirmation bias.** The tendency to search for or interpret information in a way that confirms one's preconceptions.
- **Wishful Thinking.** Most people display unrealistically high self-esteem. Typically 90% of all car drivers consider themselves above average, whereas only 50% should think so.

Investment Decision Making in a Complex Environment

Decision making in a complex environment refers to the case in which the decision of an investor is directly influenced by the outcome of actions taken by other decision makers. A day trader in a stock market is one such

[1] Similar ideas were also introduced in [40] where it was shown how increments of uncorrelated time series can be predicted with a universal 75% probability of success.

example, since the decision of when to enter/exit a position depends on price trajectories created by other day traders. The El Farol bar game mentioned in Sect. "Example of a Model that can not be Solved Using Rational Expectations" is another example, since a person will use the prior attendance of other people in the decision making on whether or not to go to the bar.

Some of the first approaches by economists to model decision making in a complex environment were put forth by A. S. Kyle [23], J. A. Frankel and K. A. Froot [17] and J. Bradford de Long, A. Schleifer, L. H. Summers and R. J. Waldmann [10]. Kyle studied how quickly private information of a given commodity is incorporated into market prices via a dynamic model with three types of traders: insider trader, noise traders and market makers. Frankel and Froot introduced a model of the currency markets with three types of actors: fundamentalists, chartists and portfolio managers. As noted in their abstract, their motivation to go beyond the standard rational expectation theory was that "… the proportion of exchange rate movements that can be explained even *after* the fact, using contemporaneous macroeconomic variables, is disturbingly low". Bradford de long et al. came to the conclusion, in a model of rational and noise traders, that noise traders in fact can survive and dominate the market in the long run.

More recently, models of decision making in complex environments has become a major research field for people working, notably, in statistical physics. Some of the first work that sparked a lot of interest in this field was e. g., [11,25], which viewed the marketplace as a self-organizing complex system. Later work [12] extended the El Farol game to study adaptive behavior in financial markets. The Minority Game (MG) was introduced by two physicists, Y.C. Zhang and D. Challet (1997) as a "minimal" model for financial markets (it now has its own website: www.unifr.ch/econophysics [12]).

The MG describes a model in which some agents (market participants) trade in a given market by placing buy/sell orders. A decision is made to enter a position in the market based on the last m price directions of the market. As in the El Farol game, adaptation is a part of the game, since the agents always choose the best performing strategy (defined below) among a pool of strategies. As the market changes, the strategies used by the agents will change. This, in turn, will lead to new changes of the market, illustrating feedback which is thought to be an important part in real market price dynamics. The main idea behind the MG is to introduce as parsimonious a model of financial market dynamics as possible, while sorting out the most important aspects of price formation in a market. Three parameters were suggested in such a description:

N Number of agents (market participants). This parameter takes into account the assumption that in order to get a proper description of a market, one need only know the price impact of the most important players, such as, e. g., major banks, hedge funds, pension funds, brokerage houses, etc.

m "Memory" of the strategies (assumed in the simplest version of the MG to be the same for all strategies) used by the agents. The idea is that decision making is based on strategies that use the last m price movements. This corresponds to agents using technical analysis. In its simplest version the MG, therefore, considers only short time scales, where the impact of changes in the fundamental price can be ignored.

s_i Number of strategies held by agent i. Market participants try out different strategies in order to perform optimally. As the market changes, the strategy which is optimal will eventually also change.

The strategies in the MG describe how an investment decision is made depending on the way the market has performed over the last m time steps. E.g., if $m = 3$, so that only the last three time steps (say days) are used in decision making, a strategy will tell what to do in all possible cases of how the market performed. Formally, one can represent the performance of the market by assigning, say 0 to a down movement of the market, and 1 if the market moved up. So if the market went down, down, down over the last three time days, then this can be represented by a binary string (000). If, on the other hand, the market first went up, then down and then up, one can formally represent this by the binary string (101). Using such a representation, one has 2^m different price histories (i. e., 2^m different ways to describe how the market performed over the last m time steps). A strategy, then, is defined in terms of how to make a decision for each of the 2^m possible ways the market can have gone over the last m time steps. Specifically a strategy might look like:

Investment Decision Making in Finance, Models of, Table 1

Price history	Decision
000	1
001	0
010	0
011	1
100	0
101	0
110	1
111	0

This specific strategy tells you to buy if the market over the last m time steps went down, down, down, to sell if the market over the last three time steps went down, down, up, etc. However, since each price history must give a decision whether to buy or sell (i. e., 0/1) the total number of such strategies is therefore given by $S = 2^{2^m}$. This is a *very* big number even for a moderate number of time steps that one might use in the decision process. E.g., using daily market data, and considering an investment decision of whether to enter/exit (i. e., buy/sell) the market over two weeks (i. e. $m = 10$ days), the total number of possible strategies (i. e., decisions) goes like $2^{2^{10}} \approx 10^{200}$ which by far exceeds the total number ($\approx 10^{80}$) of elementary particles (i.e, quarks (the constituents of neutrons/protons), leptons, bosons) thought to exist in the whole universe! Despite the highly simplified framework of technical analysis, this huge number still gives you a hint as to the tremendous complexity a financial investment decision exhibit, and foretells some of the difficulties to be encountered when analyzing the model.

The market mechanism introduced in the MG is a minority game: strategies which took the minority action were rewarded, whereas the majority of optimal strategies that were actually used by the agents lose. At every time step, each agent uses his most successful (among the s) strategies. If one calls the decision at time t of the jth strategy of agent i for a_i^j and the optimal strategy at time t of agent i for a_i^*, then the payoff function of any given strategy is taken as:

$$u_i^j(t) = -a_i^j(t)A(t)\,, \quad A(t) = \sum_i a_i^*(t)\,. \tag{14}$$

The price dynamics, $p(t)$, is then expressed in terms of the excess demand:

$$\log p(t+1) = \log p(t) + A(t)/\lambda \tag{15}$$

with λ the "depth" (or liquidity) of the market. The MG captures some behavioral phenomena that are thought to be of importance in financial markets, such as:

- competition
- frustration
- adaptability

and

- evolution.

Since the introduction of the Minority Game many extensions have been suggested (the history of the work done on this model can be found at the website in [12]). E.g., it seems natural to model the case with agents trying to profit from market movements instead of striving always be in the minority. To do so, they will take a position at time t at, say, price $p(t)$, but will know only at time $t+1$ whether that decision meant a profit or a loss [3,19]. From (15), the return is proportional to $A(t)$, so the payoff function (14) in that case becomes $u_i^j(t) = a_i^j(t-1)A(t)$, i. e., the payoff function changes sign and is evaluated over two time steps. It turns out that the solution in that case *can* be solved using rational expectations [32], in contrast to the MG where no such solution is available. Without any constraint in buying power, it becomes rational for the agents to try to figure out what the other agents will do, i. e. the solution reflects Keynes's beauty queen contest. Interestingly bubbles in this case are created spontaneously, something missing in the MG.

Future Directions

Beyond doubt, any proper description of how decision making in finance really takes place will have to involve aspects from all three of the above-mentioned decision making contexts, that is, rational decision making, irrational decision making, and decision making in a complex environment. Also, a multi disciplinary effort seems to be an absolute necessity. Financial markets are made up of humans, not robots with perfect insight, and with the complexity and psychology often involved in human decision making, we face an enormous challenge in the development of new tools to do proper risk management.

Behavioral finance laid the first bricks for interdisciplinary approaches between psychology and finance. Econophysics has made attempts to further bridge finance with new research methods found in another field. Such approaches, where new ideas come across to an already established field, however, do not always happen without resistance from researchers using conventional methods. For the time being, the relevance of econophysics is under debate (see e. g. [6,14,18]) just as behavioral finance was when it first appeared. Whereas criticism is good and constructive, exclusion is bad. As mentioned in a discussion about econophysics by the progressive economist B. Arthur in [6]: "They tried to publish in mainstream economics journals, but were rebutted. It's a shame and a scandal that the journals haven't opened their pages to them". Talking about future directions, this does not seem like the path to take.

In general, it would be very interesting to gain firmer understanding of how aggregation of individual behavior can give rise to measurable effects in a population in general, and in financial markets in particular. It would also be interesting to model specific human traits on a micro

scale and study the emergence of a dynamics with observable or even predictable effects on a macro scale. The hope would be to reproduce in models many of the mechanisms reported at work in behavioral finance. One step in this direction was made in [4], where it was shown how consensus (called "decoupling" in [4]), and thereby predictability, could emerge due to mutual influence of the price in a commonly traded asset, among a group of agents who had initially different opinions. In [42], a very original study was done on self-referential behavior and the impact it had on overreaction and conventions in financial markets.

From the point of view of a financial market as a complex system, it would be interesting to sort out some of the important factors where new ways of investment decisions could have determining impact on the price dynamics of the market. In [45], it was suggested that the exponential growth of hedge funds which has happened since the start of the twenty-first century could have a catastrophic impact on the *long term* trends of financial markets. The explanation for this is the role of traditional market players of long-only mutual funds versus hedge funds which take both short and long positions. In [45], it was argued that financial markets since their very origin, and only till very recently, have been in a state of "broken symmetry" which favored long term growth instead of contraction. The reason for this "broken symmetry" into a long term "bull phase" is the historical, almost complete, dominance by long-only players in financial markets. Dangers connected to short trading can be seen in Fig. 3, which shows simulations of a financial market with varying percentages of market participants taking short positions. As illustrated in the figure, two predictions come out concerning the impact of the new hedge fund players in the markets: over long time horizons, one should see an increase in volatility and, more importantly, an increase in the probability for periods with "bear" markets. If, as is often speculated, one has a spillover from the financial markets into the economy (through the so called "wealth effect"), the introduction of hedge funds that accelerated from the beginning

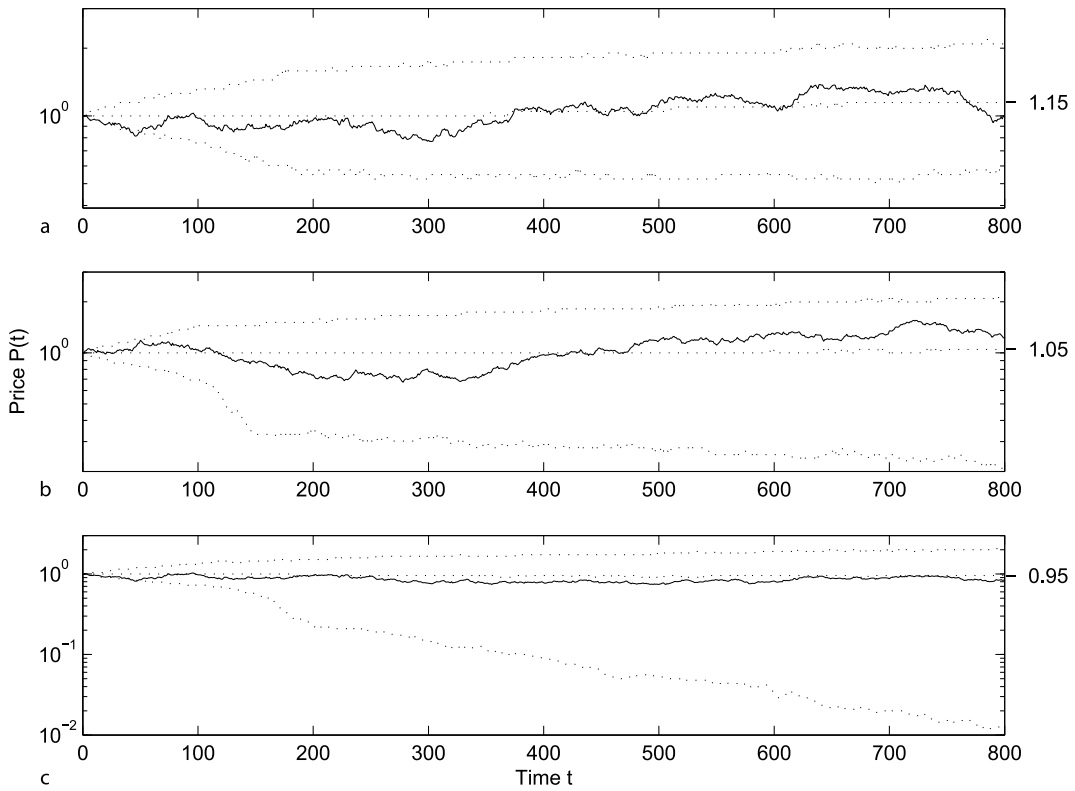

Investment Decision Making in Finance, Models of, Figure 3

***Fat solid line*: price *P*(*t*) generated from an agent based model (see [45] for a definition of the model) *Thin dotted lines* represent the 5%, 50% and 95% quantiles (from bottom to top) respectively, i. e., at every time *t*, out of the 1000 different initial configurations, only 50 got below the 5% quantile line, and similarly for the other quantile lines. a The fraction of agents allowed to take short positions, $\rho = 0$, b $\rho = 0.2$ and c $\rho = 0.4$**

of this century could have dire consequences. This message, however, is something which is difficult to communicate to researchers solidly implanted in rational expectation theory.

Bibliography

Primary Literature

1. Andersen JV (2007) Detecting anchoring in financial markets. Submitted to: Quant Financ. The article can be retrieved from the site: http://arXiv.org/abs/0705.3319. Accessed 2007
2. Andersen JV, Sornette D (2001) Have your cake and eat it too: increasing returns while lowering large risks! J Risk Rinance, p 70; Sornette D, Andersen JV, Simonetti P (2000) Minimizing volatility increases large risks. Int J Theor Appl Financ 3(3):523–535; Sornette D, Simonetti P, Andersen JV (2000) φ^q-field theory for portfolio optimization: fat tails and nonlinear correlations. Phys Rep 335:19–92
3. Andersen JV, Sornette D (2003) The $-game. Eur Phys J B 31:141
4. Andersen JV, Sornette D (2006) A Mechanism for Pockets of Predictability in Complex Adaptive Systems. Europhys Lett 70:697; The findings of the article is described in: Buchanan M (2005) Too much information. New Sci 85:32–35
5. Arthur B (1994) Bounded rationality and inductive behavior. Am Econ Rev, Papers and Proceedings 84:406
6. Ball P (2006) Culture Crash. Nature 441:686–688
7. Barberis N, Thaler R (2003) A Survey of Behavioral Finance. Handbook of the Economics of Finance. North Holland, Amsterdam, pp 1053–1128
8. Black F, Scholes M (1973) The pricing of options and corporate liabilities. J Political Econ 81:637–654; Black F (1989) How we came up with the option formula. J Portf Manag 15:4–8; Merton RC (1973) Theory of rational option pricing. Bell J Econ Manag Sci 4:141–183
9. Bofinger P, Schmidt R (2004) Should One Rely on Professional Exchange Rate Forecasts? An Empirical Analysis of Professional Forecasts for the ϵ/US $-Rate. Discussion Paper Series - Centre For Economic Policy Research London, ISSU 4235
10. Bradford de Long J, Schleifer A, Summers LH, Waldmann RJ (1990) The survival of noise traders in financial markets. J Bus 64(1):1–19
11. Caldarelli G, Marsili M, Zhang YC (1997) A prototype model of stock exchange. Europhys Lett 40:479–484
12. Challet D, Zhang Y-C (1997) Emergence of cooperation and organization in an evolutionary game. Physica A 246:407–418; To follow the literature of the MG see e. g. the special website www.unifr.ch/econophysics/minority/minority.html, and the recent book: Challet D, Marsili M, Zhang Y-C (2004) Minority Games: Interacting Agents In Financial Markets. Oxford University Press, Oxford
13. Dodonova A, Khoroshilov Y (2004) Anchoring and transaction utility: evidence from on-line auctions. Appl Econ Lett 11:307
14. Durlauf SN (2005) Complexity and empirical economics. Econ J 115:F225–F243
15. Edwards W (1954) Behavioral decision theory. Annu Rev Psychol 12:473–98
16. Fama EF, French KR (2004) The capital asset pricing model: theory and evidence. J Econ Perspect 18(3):25-46
17. Frankel JA, Froot KA (1990) Chartists, fundamentalists and the demand for dollars. In: Courakis AS, Taylor MP (eds) Private Behavior and Government Policy in Interdependent Economies. Oxford University press, Oxford, pp 73–126
18. Gallegatti M, Ken S, Lux T, Ormerod P (2006) Worrying trends in econophysics. Physica A 370:1–6
19. Giardina I, Bouchaud J-P (2003) Bubbles, crashes and intermittency in agent based models. Eur Phys J B 31:421–437
20. Gil L (2007) A simple algorithm based on fluctuations to play the market. arXiv:0705.2097
21. Heilmann K, Laeger V, Oehler A (2000) The Disposition Effect: Evidence about the Investors Aversion to Realize Losses. In: Proceedings of the 25th Annual Colloquium, Baden/Vienna, 12–16 July 2000. IAREP, Wien
22. Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. Econometrica 47:263–291; Kahneman D, Tversky A (1972) Subjective probability: A judgement of representativeness. Cogn Psychol 3:430–454; Kahneman D, Tversky A (1973) On the psychology of prediction. Psychol Rev 80:237–251
23. Kyle AS (1985) Continours auctions and insider trading. Econometrica 53(6):1315–1336
24. Lintner J (1965) The valuation of risk assets and selection of risky investments in stock portfolios and capital budgets. Rev Econ Stat 47(1):13–37
25. Lux T, Marchesi M (1999) Scaling and criticality in a stochastic multi-agent model of a financial market. Nature 397:498–500
26. Markovitz H (1959) Portfolio selection: Efficient diversification of investments. John Wiley and Sons, New York
27. Modigliani F, Miller MH (1958) The cost of capital, corporation finance and the theory of investment. Am Econ Rev 48:261–297
28. Northcraft GB, Neale MA (1987) Experts, amateurs, and real estate: an anchoring and adjustment perspective on property pricing decisions. Organ Behav Hum Decis Process 39:84
29. Peigneux P, Orban P, Balteau E, Degueldre C, Luxen A (2006) Offline Persistence of Memory-Related Cerebral Activity during Active Wakefulness. PLoS Biology 4(4):e100 doi:10.1371/journal.pbio.0040100
30. Roll R (2002) Rational infinitely lived asset prices must be nonstationary. J Bank Financ 26(6):1093–1097
31. Ross S (1976) The arbitrage theory of capital asset pricing. J Econ Theory 13(3):341–360
32. Roszczynska M, Nowak A, Kamieniarz D, Solomon S, Andersen JV (2008) Detecting speculative bubbles created in experiments via decoupling in agent based models. Available at http://arxiv.org/abs/0806.2124
33. Samuelson PA (1965) Proff that properly anticipated prices fluctuate randomly. Ind Manag Rev 6:41–49
34. Sharpe WF (1964) Capital asset prices: a theory of market equilibrium under conditions of risk. J Financ 19(3):425–442
35. Shefrin HM, Statman M (1985) The disposition to sell winners too early and rise losers too long. J Financ 40:777
36. Shiller R (1981) Do Stock Prices Move Too much to be Justified by Subsequent Changes in Dividends? Am Econ Rev 71:421–426
37. Shiller RJ (2003) From Efficient Market Theory to Behavioral Finance. J Econ Perspect, Am Econ Assoc 17(1):83–104. http://papers.ssrn.com/abstract_id=349660
38. Simon HA (1955) A behavioral model of rational choice.

Q J Econ 69(1):99–118; Simon HA (1956) Rational choice and the structure of the environment. Psychol Rev 63:129–138
39. Smith VL (1962) An experimental study of competitive market behavior. J Political Econ 70:111–137; Smith VL (1965) Experimental auction markets and the Walrasian hypothesis. J Political Econ 73:387–393; Smith VL (1976) Experimental economics: Induced value theory. Am Econ Rev, Papers and Proceedings, pp 274–279; Plott C, Smith VL (1978) An experimental examination of two exchange institutions. Rev Econ Stud 45:133–153
40. Sornette D, Andersen JV (2000) Increments of uncorrelated time series can be predicted with a universal 75% probability of success. Int J Mod Phys C 11:713
41. Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. Science 185:1124
42. Wyart M, Bouchaud J-P (2007) Self-referential behavior, overreaction and conventions in financial markets. J Econ Behav Organ 63:1–24
43. Weber M, Camerer CF (1998) The disposition effect in securities trading: An experimental analysis. J Econ Behav Organ 33:167
44. Ye G (2004) Inertia Equity: The impact of anchoring price in brand switching. SSRN-id548862
45. Andersen JV (2005) Could Short Selling Make Financial Markets Tumble? Int J Theor App Fin 8(4):509–521

Books and Reviews

Bouchaud JP, Potters M (2000) Theory of Financial Risks. Cambridge University Press, Cambridge
Johnson NF, Jefferies P, Hui PM (2003) Financial Market Complexity. Oxford University Press, Oxford
Mantegna RN, Stanley HE (2000) An Introduction to Econophysics: Correlations and Complexity in Finance. Cambridge University Press, Cambridge
McCauley JL (2004) Dynamics of Markets: Econophysics and Finance. Cambridge University Press, Cambridge
Roehner BM (2002) Patterns of Speculation: A Study in Observational Econophysics. Cambridge University Press, Cambridge
Sornette D (2003) Why Stock Markets Crash (Critical Events in Complex Financial Systems). Princeton University Press, Princeton

Isomorphism Theory in Ergodic Theory

Christopher Hoffman
Department of Mathematics, University of Washington, Seattle, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Basic Transformations
Basic Isomorphism Invariants
Basic Tools
Isomorphism of Bernoulli Shifts
Transformations Isomorphic to Bernoulli Shifts
Transformations not Isomorphic to Bernoulli Shifts
Classifying the Invariant Measures of Algebraic Actions
Finitary Isomorphisms
Flows
Other Equivalence Relations
Non-invertible Transformations
Factors of a Transformation
Actions of Amenable Groups
Future Directions
Bibliography

Glossary

Almost everywhere A property is said to hold **almost everywhere (a.e.)** if the set on which the property does not hold has measure 0.

Bernoulli shift A Bernoulli shift is a stochastic process such that all outputs of the process are independent.

Conditional measure For any measure space $(X, \mathcal{B}, \mu)$ and σ-algebra $\mathcal{C} \subset \mathcal{B}$ the conditional measure is a $\mathcal{C}$-measurable function g such that $\mu(C) = \int_C g \, d\mu$ for all $C \in \mathcal{C}$.

Coupling of two measure spaces A coupling of two measure spaces $(X, \mu, \mathcal{B})$ and $(Y, \nu, \mathcal{C})$ is a measure γ on $X \times Y$ such that $\gamma(B \times Y) = \mu(B)$ for all $B \in \mathcal{B}$ and $\gamma(X \times C) = \nu(C)$ for all $C \in \mathcal{C}$.

Ergodic measure preserving transformation A measure preserving transformation is ergodic if the only invariant sets $(\mu(A \triangle T^{-1}(A)) = 0)$ have measure 0 or 1.

Ergodic theorem The pointwise ergodic theorem says that for any measure preserving transformation $(X, \mathcal{B}, \mu)$ and T and any L^1 function f the time average

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(T^i(x))$$

converges a.e. If the transformation is ergodic then the limit is the space average, $\int f \, d\mu$ a.e.

Geodesic A geodesic on a Riemannian manifold is a distance minimizing path between points.

Horocycle A horocycle is a circle in the hyperbolic disk which intersects the boundary of the disk in exactly one point.

Invariant measure Likewise a measure μ is said to be invariant with respect to $(\mathbf{X}, T)$ provided that $\mu(T^{-1}(A)) = \mu(A)$ for all measurable $A \in \mathcal{B}$.

Joining of two measure preserving transformations A joining of two measure preserving transformations $(\mathbf{X}, T)$ and $(\mathbf{Y}, S)$ is a coupling of $\mathbf{X}$ and $\mathbf{Y}$ which is invariant under $T \times S$.

Markov shift A Markov shift is a stochastic process such that the conditional distribution of the future outputs ($\{x_n\}_{n>0}$) of the process conditioned on the last output (x_0) is the same as the distribution conditioned on all of the past outputs of the process ($\{x_n\}_{n\leq 0}$).

Measure preserving transformation A measure preserving transformation consists of a probability space ($\mathbf{X}$, T) and a measurable function $T\colon X \to X$ such that $\mu(T^{-1}(A)) = \mu(A)$ for all $A \in \mathcal{B}$.

Measure theoretic entropy A numerical invariant of measure preserving transformations that measures the growth in complexity of measurable partitions refined under the iteration of the transformation.

Probability space A probability space $\mathbf{X} = (X, \mu, \mathcal{B})$ is a measure space such that $\mu(B) = 1$.

Rational function, rational map A rational function $f(z) = g(z)/h(z)$ is the quotient of two polynomials. The degree of $f(z)$ is the maximum of the degrees of $g(z)$ and $h(z)$. The corresponding rational maps $T_f\colon z \to f(z)$ on the Riemann sphere $\mathbb{C}$ are a main object of study in complex dynamics.

Stochastic process A stochastic process is a sequence of measurable functions $\{x_n\}_{n\in\mathbb{Z}}$ (or outputs) defined on the same measure space, $\mathbf{X}$. We refer to the value of the functions as outputs.

Definition of the Subject

Our main goal in this article is to consider when two measure preserving transformations are in some sense different presentations of the same underlying object. To make this precise we say two measure preserving maps ($\mathbf{X}$, T) and ($\mathbf{Y}$, S) are **isomorphic** if there exists a measurable map $\phi\colon X \to Y$ such that

(1) ϕ is measure preserving,
(2) ϕ is invertible almost everywhere and
(3) $\phi(T(x)) = S(\phi(x))$ for almost every x.

The main goal of the subject is to construct a collection of invariants of a transformation such that a necessary condition for two transformations to be isomorphic is that the invariant be the same for both transformations. Another goal of the subject is to solve the much more difficult problem of constructing invariants such that the invariants being the same for two transformations is a sufficient condition for the transformation to be isomorphic. Finally we apply these invariants to many natural classes of transformation to see which of them are (or are not) isomorphic.

Introduction

In this article we look at the problem of determining of which measure preserving transformations are isomorphic. We look at a number of isomorphism invariants, the most important of which is the Kolmogorov–Sinai entropy. The central theorem in this field is Ornstein's proof that any two Bernoulli shifts of the same entropy are isomorphic. We also discuss some of the consequences of this theorem, which transformations are isomorphic to Bernoulli shifts as well as generalizations of Ornstein's theory.

Basic Transformations

In this section we list some of the basic classes of measure preserving transformations that we study in this article.

Bernoulli shifts Some of the most fundamental transformations are the Bernoulli shifts. A **probability vector** is a vector $\{p_i\}_{i=1}^n$ such that $\sum_{i=1}^n p_i = 1$ and $p_i \geq 0$ for all i. Let $\mathbf{p} = \{p_i\}_{i=1}^n$ be a probability vector. The **Bernoulli shift** corresponding to $\mathbf{p}$ has state space $\{1, 2, \ldots, n\}^{\mathbb{Z}}$, the shift operator $T(x)_i = x_{i+1}$. To specify the measure we only need to specify it on **cylinder sets**

$$A = \{x \in X\colon x_i = a_i \ \forall i \in \{m, \ldots, k\}\}$$

for some $m \leq k \in \mathbb{Z}$ and a sequence $a_m, \ldots, a_k \in \{1, \ldots, n\}$. The measure on cylinder sets is defined by

$$\mu\big\{x \in X\colon x_i = a_i \text{ for all } i \text{ such that } m \leq i \leq k\big\} = \prod_{i=m}^{k} p_{a_i}\,.$$

For any $d \in \mathbb{N}$ if $\mathbf{p} = (1/d, \ldots, 1/d)$ we refer to Bernoulli$_p$ as the **Bernoulli d shift**.

Markov shifts A **Markov shift** on state n symbols is defined by an $n \times n$ matrix, M, such that $\{M(i, j)\}_{j=1}^n$ is a probability vector for each i. The Markov shift is a measure preserving transformation with state space $\{1, 2, \ldots, n\}^{\mathbb{Z}}$, transformation $T(x)_i = x_{i+1}$

$$\mu\{x_1 = a_1 \mid x_0 = a_0\} = \mu\{x_1 = a_1 \mid x_0 = a_0, x_{-1} = a_{-1}, x_{-2} = a_{-2}, \ldots\}$$

for all choices of a_i, $i \leq 1$. Let $m = \{m(i)\}_{i=1}^n$ be a vector such that $Mm = m$. Then an invariant measure is defined by setting the measure on cylinder sets to be $A = \{x\colon x_0 = a_0, x_1 = a_1, \ldots, x_n = a_n\}$ is given by $\mu(A) = m(a_0) \prod_{i=1}^n M(a_{i-1}, a_i)$.

Shift maps More generally the **shift map** σ is the map $\sigma\colon \mathbb{N}^{\mathbb{Z}} \to \mathbb{N}^{\mathbb{Z}}$ where $\sigma(x)_i = x_{i+1}$ for all $x \in \mathbb{N}^{\mathbb{Z}}$ and $i \in \mathbb{Z}$. We also let σ designate the shift map on $\mathbb{N}^{\mathbb{N}}$. For each measure that is invariant under the shift map there is a corresponding measure defined on $\mathbb{N}^{\mathbb{N}}$ that is invariant under the shift map. Let μ be an invariant measure under the shift map. For any measurable set of $A \subset \mathbb{N}^{\mathbb{N}}$ we define $\tilde{A}$ on $\mathbb{N}^{\mathbb{Z}}$ by

$$\tilde{A} = \{\ldots, x_{-1}, x_0, x_1, \ldots : x_0, x_1, \ldots \in A\}\,.$$

Then it is easy to check that $\tilde{\mu}$ defined by $\tilde{\mu}(\tilde{A}) = \mu(A)$ is invariant. If the original transformation was a Markov or Bernoulli shift then refer to the resulting transformations as a **one sided Markov shifts** or **one sided Bernoulli shift** respectively.

Rational maps of the Riemann sphere We say that $f(z) = g(z)/h(z)$ is a **rational function of degree** $d \geq 2$ if both $g(z)$ and $h(z)$ are polynomials with $\max(\deg(g(z)), \deg(h(z))) = d$. Then f induces a natural action on the Riemann sphere $T_f\colon z \to f(z)$ which is a d to one map (counting with multiplicity). In Subsect. "Rational Maps" we shall see that for every rational function f there is a canonical measure μ_f such that T_f is a measure preserving transformation.

Horocycle flows The horocycle flow acts on $SL(2,\mathbb{R})/\Gamma$ where Γ is a discrete subgroup of $SL(2,\mathbb{R})$ such that $SL(2,\mathbb{R})/\Gamma$ has finite Haar measure. For any $g \in SL(2,\mathbb{R})$ and $t \in \mathbb{R}$ we define the horocycle flow by

$$h_t(\Gamma g) = \Gamma g \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix}\,.$$

Matrix actions Another natural class of actions is given by the action of matrices on tori. Let M be an invertible $n \times n$ integer valued matrix. We define $T_M\colon [0,1)^n \to [0,1)^n$ by $T_M(x)_i = M(x)_i \bmod 1$ for all i, $1 \leq i \leq n$. It is easy to check that if M is surjective then Lebesgue measure is invariant under T_M. T_M is a $|\det(M)|$ to one map. If $n = 1$ then M is an integer and we refer to the map as times M.

The $[T, T^{-1}]$ transformations Let $(\mathbf{X}, T)$ be any invertible measure preserving transformation. Let σ be the shift operator on $Y = \{-1, 1\}^{\mathbb{Z}}$. The space Y comes equipped with the Bernoulli $(1/2, 1/2)$ product measure ν.

The $[T, T^{-1}]$ transformation is a map on $Y \times X$ which preserves $\nu \times \mu$. It is defined by

$$T, T^{-1}(y, x) = (S(y), T^{y_0}(x))\,.$$

Induced transformations Let $(\mathbf{X}, T)$ be a measure preserving transformation and let $A \subset X$ with $0 < \mu(A) < 1$. The **transformation induced by** A, (A, T_A, μ_A), is defined as follows. For any $x \in A$

$$T_A(x) = T^{n(x)}(x)$$

where $n(x) = \inf\{m > 0\colon T^m(x) \in A\}$. For any $B \subset A$ we have that $\mu_A(B) = \mu(B)/\mu(A)$.

Basic Isomorphism Invariants

The main purpose of isomorphism theory is to classify which pairs of measure preserving transformation are isomorphic and which are not isomorphic. One of the main ways that we can show that two measure preserving transformation are not isomorphic is using isomorphism invariants. An **isomorphism invariant** is a function f defined on measure preserving transformations such that if $(\mathbf{X}, T)$ is isomorphic to $(\mathbf{Y}, S)$ then $f((\mathbf{X}, T)) = f((\mathbf{Y}, S))$.

A measure preserving action is said to be **ergodic** if $\mu(A) = 0$ or 1 for every A with

$$\mu\left(A \triangle T^{-1}(A)\right) = 0\,.$$

A measure preserving action is said to be **weak mixing** if for every measurable A and B

$$\frac{1}{n}\sum_{n=1}^{\infty} \left|\mu\left(A \cap T^{-n}(B)\right) - \mu(A)\mu(B)\right| = 0\,.$$

A measure preserving action is said to be **mixing** if for every measurable A and B

$$\lim_{n\to\infty} \mu\left(A \cap T^{-n}(B)\right) = \mu(A)\mu(B)\,.$$

It is easy to show that all three of these properties are isomorphism invariants and that

$$\begin{aligned}(\mathbf{X}, T) \text{ is mixing } &\Rightarrow (\mathbf{X}, T) \text{ is weak mixing}\\ &\Rightarrow (\mathbf{X}, T) \text{ is ergodic.}\end{aligned}$$

We include one more definition before introducing an even stronger isomorphism invariant. The **action of a group** (or a semigroup) G on a probability space $(X, \mu, \mathcal{B})$ is a family of measure preserving transformations $\{f_g\}_{g\in G}$ such that for any $g, h \in G$ we have that $f_g(f_h(x)) = f_{g+h}(x)$ for almost every $x \in X$. Thus any invertible measure preserving transformation $(\mathbf{X}, T)$ induces a $\mathbb{Z}$ action by $f_n(x) = T^n(x)$.

We say that a group action $(\mathbf{X}, T_g)$ is mixing of all orders if for every $n \in \mathbb{N}$ and every collection of set

$A_1, \ldots, A_n \in \mathcal{B}$ then

$$\mu\big(A_1 \cap T_{g_2}(A_2) \cap T_{g_2+g_3}(A_3) \ldots \cap T_{g_2+g_3+\ldots g_n}(A_n)\big) \to \prod_{i=1}^{n} \mu(A_i)$$

as the g_i go to infinity.

Basic Tools

Birkhoff's ergodic theorem states that for every measure preserving action the limits

$$\hat{f}(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f\left(T^k x\right)$$

exists for almost every x and if $(\mathbf{X}, T)$ is ergodic then the limit is $\bar{f} = \int f \, \mathrm{d}\mu$ [2].

A **partition** P of X is a measurable function defined on X. (For simplicity we often assume that a partition is a function to $\mathbb{Z}$ or some subset of $\mathbb{Z}$.) We write P_i for $P^{-1}(i)$. For any partition P of $(\mathbf{X}, T)$ define the partition $T^i P$ by $P \circ T^i$. Thus for invertible T this is given by $T^i P(x) = P(T^{-i}(x))$. Then define $(P)_T = \bigvee_{i\in Z} T^i P$. Thus $(P)_T$ is the smallest σ-algebra which contains $T^i(P_j)$ for all $i \in \mathbb{Z}$ and $j \in \mathbb{N}$. We say that $(P)_T$ is the **σ-algebra generated by** P. A partition P is a **generator** of $(\mathbf{X}, T)$ if $(P)_T = \mathcal{B}$. Many measure preserving transformations come equipped with a natural partition.

Rokhlin's theorem For any measure preserving transformation $(\mathbf{X}, T)$, any $\epsilon > 0$ and any $n \in \mathbb{N}$ there exists $A \subset X$ such that $T^{-i}(A) \cap T^{-j}(A)$ for all $0 \le i < j \le n$ and $\mu(\cup_{i=0}^{n} T^i(A) > 1 - \epsilon$. Moreover for any finite partition P of X we can choose A such that $\mu(P_i \cap A) = \mu(A)\mu(P_i)$ for all $i \in \mathbb{N}$ [63].

Shannon–McMillan–Breiman theorem [3,57] For any measure preserving system $(\mathbf{X}, T)$ and any $\epsilon > 0$ there exists $n \in \mathbb{N}$ and a set G with $\mu(G) > 1 - \epsilon$ with the following property. For any sequence $g_1, \ldots g_n \in \mathbb{N}$ let $g = \bigcap_{i=1}^{n} P(T^{-i}(x)) = g_i$. Then if $\mu(g \cap G) > 0$ then

$$\mu(g) \in \left(2^{h((\mathbf{X},T))-\epsilon}, 2^{h((\mathbf{X},T))+\epsilon}\right) .$$

Krieger generator theorem If $H((\mathbf{X}, T)) < \infty$ then there exists a finite partition P such that $(P)_T = \mathcal{B}$. Thus every measure preserving transformation with finite entropy is isomorphic to a shift map on finitely many symbols [33].

Measure-theoretic entropy Entropy was introduced in physics by Rudolph Clausius in 1854. In 1948 Claude Shannon introduced the concept to information theory. Consider a process that generates a string of data of length n. The entropy of the process is the smallest number h such that you can condense the data to a string of zeroes and one of length hn and with high probability you can reconstruct the original data from the string of zeroes and ones. Thus the entropy of a process is the average amount of information transmitted per symbol of the process.

Kolmogorov and Sinai introduced the concept of entropy to ergodic theory in the following way [31,32]. They defined the entropy of a partition Q is defined to be

$$H(Q) = -\sum_{i=1}^{k} \mu(Q_i) \log \mu(Q_i) .$$

The measure-theoretic entropy of a dynamical system $(\mathbf{X}, T)$ with respect to a partition $Q\colon X \to \{1, \ldots, k\}$ is then defined as

$$h(\mathbf{X}, T, Q) = \lim_{N\to\infty} \frac{1}{N} H\left(\bigvee_{n=1}^{N} T^{-n} Q\right) .$$

Finally, the measure-theoretic entropy of a dynamical system $(\mathbf{X}, T)$ is defined as

$$h((\mathbf{X}, T)) = \sup_{Q} h(\mathbf{X}, T, Q)$$

where the supremum is taken over all finite measurable partitions. A theorem of Sinai showed that if Q is a generator of $(\mathbf{X}, T)$ then $h(T) = h(T, Q)$ [73].

This shows that for every measure preserving function $(\mathbf{X}, T)$ there is an associated entropy $h(T) \in [0, \infty]$. It is easy to show from the definition that entropy is an isomorphism invariant.

We say that $(\mathbf{Y}, S)$ is a **factor** of $(\mathbf{X}, T)$ if there exists a map $\phi\colon X \to Y$ such that

(1) ϕ is measure preserving and
(2) $\phi(T(x)) = S(\phi(x))$ for almost every x.

Each factor $(\mathbf{Y}, S)$ can be associated with $\phi^{-1}(C)$, which is an invariant sub σ-algebra of $\mathcal{B}$. We say that $(\mathbf{Y}, S)$ is trivial if Y consists of only one point. We say that a transformation $(\mathbf{X}, T)$ has **completely positive entropy** if every non-trivial factor of $(\mathbf{X}, T)$ has positive entropy.

Isomorphism of Bernoulli Shifts

Kolmogorov–Sinai

A long standing open question was for which **p** *and* **q** *are Bernoulli$_p$ and Bernoulli$_q$ isomorphic. In particular are the*

Bernoulli 2 shift and the Bernoulli 3 shift isomorphic. Both of these transformations have completely positive entropy and all other isomorphism invariants which were known at the time are the same for the two transformations. The first application of the Kolmogorov–Sinai entropy was to show that the answer to this question is no.

Fix a probability vector $\mathbf{p}$. The transformation Bernoulli_p has $Q_{\mathbf{p}}\colon x \to x_0$ as a generating partition. By Sinai's theorem

$$H(\text{Bernoulli}_p) = H(Q_{\mathbf{p}}) = \sum_{i=1}^{n} -p_i \log_2(p_i)\,.$$

Thus the Bernoulli 2 shift (with entropy 1) is not isomorphic to the Bernoulli 3 shift (with entropy $\log_2(3)$).

Sinai also made significant progress toward showing that Bernoulli shifts with the same entropy are isomorphic by proving the following theorem.

Theorem 1 [72] *If $(\mathbf{X}, T)$ is a measure preserving system of entropy h and $(\mathbf{Y}, S)$ is a Bernoulli shift of entropy $h' \le h$ then $(\mathbf{Y}, S)$ is a factor of $(\mathbf{X}, T)$.*

This theorem implies that if $\mathbf{p}$ *and* $\mathbf{q}$ *are probability vectors and $H(\mathbf{p}) = H(\mathbf{q})$ then Bernoulli_p is a factor in Bernoulli_q and Bernoulli_q is a factor in Bernoulli_p. Thus we say that Bernoulli_p and Bernoulli_q are* ***weakly isomorphic****.*

Explicit Isomorphisms

The other early progress on proving that Bernoulli shifts with the same entropy are isomorphic came from Meshalkin. He considered pairs of probability vectors $\mathbf{p}$ *and* $\mathbf{q}$ *with $H(\mathbf{p}) = H(\mathbf{q})$ and all of the p_i and q_i are related by some algebraic relations. For many such pairs he was able to prove that the two Bernoulli shifts are isomorphic. In particular he proved the following theorem.*

Theorem 2 [40] *Let $\mathbf{p} = (1/4, 1/4, 1/4, 1/4)$ and $\mathbf{q} = (1/2, 1/8, 1/8, 1/8, 1/8)$. Then Bernoulli_p and Bernoulli_q are isomorphic.*

Ornstein

The central theorem in the study of isomorphisms of measure preserving transformations is Ornstein's isomorphism theorem.

Theorem 3 [46] *If $\mathbf{p}$ and $\mathbf{q}$ are probability vectors and $H(\mathbf{p}) = H(\mathbf{q})$ then the Bernoulli shifts Bernoulli_p and Bernoulli_q are isomorphic.*

To see how central this is to the field most of the rest of this article is a summary of:

(1) The proof of Ornstein's theorem,
(2) The consequences of Ornstein's theorem,
(3) The generalizations of Ornstein's theorem, and
(4) How the properties that Ornstein's theorem implies that Bernoulli shifts must have differ from the properties of every other class of transformations.

The key to Ornstein's proof was the introduction of the finitely determined property. To explain the finitely determined property we first define the Hamming distance of length n between sequences $x, y \in \mathbb{N}^{\mathbb{Z}}$ by

$$\bar{d}_n(x, y) = 1 - \frac{|\{k \in \{1, \ldots, n\}\colon x_k = y_k\}|}{n}\,.$$

Let $(\mathbf{X}, T)$ and $(\mathbf{Y}, S)$ be a measure preserving transformation and let P and Q be finite partitions of X and Y respectively.

We say that $(\mathbf{X}, T)$ and P and $(\mathbf{Y}, S)$ and Q are within δ in n distributions if

$$\sum_{(a_1,\ldots,a_n)\in\mathbb{Z}^n} \left| \mu\left(\left\{x\colon P\left(T^i(x)\right) = a_i \ \forall i = 1, \ldots, n\right\}\right) - \nu\left(\left\{y\colon Q\left(T^i(y)\right) = a_i \ \forall i = 1, \ldots, n\right\}\right) \right| < \delta\,.$$

A process $(\mathbf{X}, T)$ and P are **finitely determined** if for every ε there exist n and δ such that if $(\mathbf{Y}, S)$ and Q are such that

(1) $(\mathbf{X}, T)$ and P and $Y$and Q are within δ in n distributions and
(2) $|H(X, P) - H(Y, Q)| < \delta$

then there exists a joining γ of X and Y such that for all m

$$\int_{x,y} \bar{d}_m(x, y)\, \mathrm{d}\gamma(x, y) < \epsilon\,.$$

A transformation $(\mathbf{X}, T)$ is finitely determined if it is finitely determined for every finite partition P.

It is fairly straightforward to show that Bernoulli shifts are finitely determined. Ornstein used this fact along with the Rokhlin lemma and the Shannon–McMillan–Breiman theorem to prove a more robust version of Theorem 1.

To describe Ornstein's proof we use the a description due to Rothstein. We say that for a joining γ of (X, μ) and (Y, ν) that $P \subset_{\epsilon,\gamma} C$ if there exists a partition $\tilde{P}$ of C such that

$$\sum_i \gamma\left(P_i \triangle \tilde{P}_i\right) < \epsilon\,.$$

If $P \subset_{\epsilon,\gamma} C$ for all $\epsilon > 0$ then it is possible to show that there exists a partition $\tilde{P}$ of C such that

$$\sum_i \gamma\left(P_i \triangle \tilde{P}_i\right) = 0$$

and we write $P \subset_\gamma C$. If $P \subset_\gamma C$ then $(\mathbf{X}, T)$ is a factor of $(\mathbf{Y}, S)$ by the map ϕ that sends $y \to x$ where $P(T^i x) = \tilde{P}(S^i y)$ for all i.

In this language Ornstein proved that if $(\mathbf{X}, T)$ is finitely determined, P is a generating partition of X and $h((\mathbf{Y}, S)) \geq h((\mathbf{X}, T))$ then for every $\epsilon > 0$ the set of joinings γ such that $P \subset_{\gamma,\epsilon} C$ is an open and dense set. Thus by the Baire category theorem there exists γ such that $P \subset_\gamma C$. This reproves Theorem 1.

Moreover if $(\mathbf{X}, T)$ and $(\mathbf{Y}, S)$ are finitely determined, $h((\mathbf{X}, T)) = h((\mathbf{Y}, S))$ and P and Q are generating partition of X and Y then by the Baire category theorem there exists γ such that $P \subset_\gamma C$ and $Q \subset_\gamma \mathcal{B}$. Then the map ϕ that sends $y \to x$ where $P(T^i x) = \tilde{P}(S^i y)$ for all i is an isomorphism.

Properties of Bernoullis

Now we define the very weak Bernoulli property which is the most effective property for showing that a measure preserving transformation is isomorphic to a Bernoulli shift.

Given X and a partition P define the past of x by

$$P_{\text{past}}(x) = \left\{x' \colon T^i P\left(x'\right) = T^i P(x)\ \forall i \in \mathbb{N}\right\}$$

and denote the measure μ conditioned on $P_{\text{past}}(x)$ by $\mu|_{P_{\text{past}}(x)}$. Define

$$\bar{d}_{n,x,\mu} = \inf_\gamma \int \bar{d}_n\left(x, x'\right)\, \mathrm{d}\gamma\left(x, x'\right)\ ,$$

where the inf is taken over all γ which are couplings of $\mu|_{P_{\text{past}}(x)}$ and μ. Also define

$$\bar{d}_{n,P_{\text{past}},(\mathbf{X},T)} = \int \bar{d}_{n,x,\mu}\, \mathrm{d}\mu\ .$$

We say that $(\mathbf{X}, T)$ and P are very weak Bernoulli if for every $\epsilon > 0$ there exists n such that $\bar{d}_{n,P_{\text{past}},(\mathbf{X},T)} < \epsilon$. We say that $(\mathbf{X}, T)$ is very weakly Bernoulli if there exists a generating partition P such that $(\mathbf{X}, T)$ and P are very weak Bernoulli.

Ornstein and Weiss were able to show that the very weak Bernoulli property is both necessary and sufficient to be isomorphic to a Bernoulli shift.

Theorem 4 [45,53] *For transformations $(\mathbf{X}, T)$ the following conditions are equivalent:*

(1) $(\mathbf{X}, T)$ is finitely determined,
(2) $(\mathbf{X}, T)$ is very weak Bernoulli and
(3) $(\mathbf{X}, T)$ is isomorphic to a Bernoulli shift.

Using the fact that a transformation is finitely determined or very weak Bernoulli is equivalent to it being isomorphic to a Bernoulli shift we can prove the following theorem.

Theorem 5 [44]

(1) If $(\mathbf{X}, T^n)$ is isomorphic to a Bernoulli shift then $(\mathbf{X}, T)$ is isomorphic to a Bernoulli shift.
(2) If $(\mathbf{X}, T)$ is a factor of a Bernoulli shift then $(\mathbf{X}, T)$ is isomorphic to a Bernoulli shift.
(3) If $(\mathbf{X}, T)$ is isomorphic to a Bernoulli shift then there exists a measure preserving transformation $(\mathbf{Y}, S)$ such that $(\mathbf{Y}, S^n)$ is isomorphic to $(\mathbf{X}, T)$.

Rudolph Structure Theorem

An important application of the very weak Bernoulli condition is the following theorem of Rudolph.

Theorem 6 [67] *Let $(\mathbf{X}, T)$ be isomorphic to a Bernoulli shift, G be a compact Abelian group with Haar measure μ_G and $\sigma \colon X \to G$ be a measurable map. Then let $S \colon X \times G \to X \times G$ be defined by*

$$S(x, g) = (T(x), g + \sigma(x))\ .$$

Then $(X \times G, S, \mu \times \mu_G)$ is isomorphic to a Bernoulli shift.

Transformations Isomorphic to Bernoulli Shifts

One of the most important features of Ornstein's isomorphism theory is that it can be used to check whether specific transformations (or families of transformations) are isomorphic to Bernoulli shifts. The finitely determined property is the key to the proof of Ornstein's theorem and the proof of many of the consequences listed in Subsect. "Properties of Bernoullis". However if one wants to show a particular transformation is isomorphic to a Bernoulli shift then the very weak Bernoulli property is more useful. There have been many classes of transformations that have been proven to be isomorphic to a Bernoulli shift. Here we mention two.

The first class are the Markov chains. Friedman and Ornstein proved that if a Markov chain is mixing then it is isomorphic to a Bernoulli shift [9]. The second are automorphisms of $[0, 1)^n$. Let M be any $n \times n$ matrix with integer coefficients and $|\det(M)| = 1$. If none of the eigenvalues $\{\lambda_i\}_{i=1}^n$ of M are roots of unity then Katznelson proved that T_M is isomorphic to the Bernoulli shift with entropy $\sum_{i=1}^n \max(0, \log(\lambda_i))$ [29].

Transformations not Isomorphic to Bernoulli Shifts

Recall that a measure preserving transformation $(\mathbf{X}, T)$ has **completely positive entropy** if for every nontrivial $(\mathbf{Y}, S)$ which is a factor of $(\mathbf{X}, T)$ we have that $H((\mathbf{Y}, S)) > 0$. It is easy to check that Bernoulli shifts have completely positive entropy. It is natural to ask if the converse true? We shall see that the answer is an emphatic no. While the isomorphism class of Bernoulli shifts is given by just one number, the situation for transformations of completely positive entropy is infinitely more complicated.

Ornstein constructed the first example of a transformation with completely positive entropy which is not isomorphic to a Bernoulli shift [47]. Ornstein and Shields built upon this construction to prove the following theorem.

Theorem 7 [50] *For every $h > 0$ there is an uncountable family of completely positive entropy transformations which all have entropy h but no two distinct members of the family are isomorphic.*

Now that we see there are many isomorphic transformations that have completely positive entropy it is natural to ask if $(\mathbf{X}, T)$ is not isomorphic to a Bernoulli shift then is there any reasonable condition we can put on $(\mathbf{Y}, S)$ that implies the two transformations are isomorphic. For example if $(\mathbf{X}, T^2)$ and $(\mathbf{Y}, S^2)$ are completely positive entropy transformations which are isomorphic does that necessarily imply that $(\mathbf{X}, T)$ and $(\mathbf{Y}, S)$ are isomorphic? The answer turns out to be no [66]. We could also ask if $(\mathbf{X}, T)$ and $(\mathbf{Y}, S)$ are completely positive entropy transformations which are weakly isomorphic does that imply that $(\mathbf{X}, T)$ and $(\mathbf{Y}, S)$ are isomorphic? Again the answer is no [17]. The key insight to answering questions like this is due to Rudolph who showed that such questions about the isomorphism of transformations can be reduced to questions about conjugacy of permutations.

Rudolph's Counterexample Machine

Given any transformation $(\mathbf{X}, T)$ and any permutation π in S_n (or $S_{\mathbb{N}}$) we can define the transformation $(X^n, T_\pi, \mu^n, \mathcal{B})$ by

$$T_\pi(x_1, x_2, \ldots, x_n) = \left(T\left(x_{\pi(1)}\right), T\left(x_{\pi(2)}\right), \ldots, T\left(x_{\pi(n)}\right)\right)$$

where μ^n is the direct product of n copies of μ. Rudolph introduced the concept of a transformation having **minimal self joinings**. If a transformation has minimal self joinings then for every π it is possible to list all of the measures on X^n which are invariant under T_π.

If there exists an isomorphism ϕ between T and $(\mathbf{Y}, S)$ then there is a corresponding measure on $X \times Y$ which is supported on points of the form $(x, \phi(x))$ and has marginals μ and ν. Thus if we know all of the measures on $X \times Y$ which are invariant under $T \times S$ then we know all of the isomorphisms between $(\mathbf{X}, T)$ and $(\mathbf{Y}, S)$. Using this we get the following theorem.

Theorem 8 [65] *There exists a nontrivial transformation with minimal self joinings. For any transformation $(\mathbf{X}, T)$ with minimal self joinings the corresponding transformation T_{π_1} is isomorphic to T_{π_2} if and only if the permutations π_1 and π_2 are conjugate.*

There are two permutations on two elements, the flip $\pi_1 = (12)$ and the identity $\pi_2 = (1)(2)$. For both permutations, the square of the permutation is the identity. Thus there are two distinct permutations whose square is the same. Rudolph showed that this fact can be used to generate two transformations which are mixing that are not isomorphic but their squares are isomorphic. The following theorem gives more examples of the power of this technique.

Theorem 9 [65]

(1) There exists measure preserving transformations $(\mathbf{X}, T)$ and $(\mathbf{Y}, S)$ which are weakly isomorphic but not isomorphic.
(2) There exists measure preserving transformations $(\mathbf{X}, T)$ and $(\mathbf{Y}, S)$ which are not isomorphic but (X, T^k, μ) is isomorphic to (Y, S^k, ν) for every $k > 1$, and
(3) There exists a mixing transformation with no non trivial factors.

If $(\mathbf{X}, T)$ has minimal self joinings then it has zero entropy. However Hoffman constructed a transformation with completely positive entropy that shares many of the properties of transformations with minimal self joinings listed above.

Theorem 10 [17]

(1) There exist measure preserving transformations $(\mathbf{X}, T)$ and $(\mathbf{Y}, S)$ which both have completely positive entropy and are weakly isomorphic but not isomorphic.
(2) There exist measure preserving transformations $(\mathbf{X}, T)$ and $(\mathbf{Y}, S)$ which both have completely positive entropy and are not isomorphic but (X, T^k, μ) is isomorphic to (Y, S^k, ν) for every $k > 1$.

T, T Inverse

All of the transformations that have completely positive entropy but are not isomorphic to a Bernoulli shift described above are constructed by a process called cutting and stacking. These transformations have little inherent interest outside of their ergodic theory properties. This led many people to search for a "natural" example of such a transformation. The most natural examples are the $[T, T^{-1}]$ transformation and many other transformations derived from it. It is easy to show that the $[T, T^{-1}]$ transformation has completely positive entropy [39]. Kalikow proved that for many T the corresponding $[T, T^{-1}]$ transformation is not isomorphic to a Bernoulli shift.

Theorem 11 [24] *If $h(T) > 0$ then the $[T, T^{-1}]$ transformation is not isomorphic to a Bernoulli shift.*

The basic idea of Kalikow's proof has been used by many others. Katok and Rudolph used the proof to construct smooth measure preserving transformations on infinite differentiable manifolds which have completely positive entropy but are not isomorphic to Bernoulli shifts [27,68]. Den Hollander and Steif did a thorough study of the ergodic theory properties of $[T, T^{-1}]$ transformations where T is simple random walk on a wide family of graphs [4].

Classifying the Invariant Measures of Algebraic Actions

This problem of classifying all of the invariant measures like Rudolph did with his transformation with minimal self joinings comes up in a number of other settings. Ratner characterized the invariant measures for the horocycle flow and thus characterized the possible isomorphisms between a large family of transformations generated by the horocycle flow [59,60,61]. This work has powerful applications to number theory.

There has also been much interest in classifying the measures on $[0, 1)$ that are invariant under both times 2 and times 3. Furstenberg proved that the only closed infinite set on the circle which is invariant under times 2 and times 3 is $[0, 1)$ itself and made the following measure theoretic conjecture.

Conjecture 1 [10] *The only nonatomic measure on $[0, 1)$ which is invariant under times 2 and times 3 is Lebesgue measure.*

Rudolph improved on the work of Lyons [36] to provide the following partial answer to this conjecture.

Theorem 12 [69] *The only measure on $[0, 1)$ which is invariant under multiplication by 2 and by 3 and has positive entropy under multiplication by 2 is Lebesgue measure.*

Johnson then proved that for all relatively prime p and q that p and q can be substituted for 2 and 3 in the theorem above.

This problem can be generalized to higher dimensions by studying the actions of commuting integer matrices of determinant greater than one on tori. Katok and Spatzier [28] and Einsiedler and Lindenstrauss [5] obtained results similar to Rudolph's for actions of commuting matrices.

Finitary Isomorphisms

By Ornstein's theorem we know that there exists an isomorphism between any two Bernoulli shifts (or mixing Markov shifts) of the same entropy. There has been much interest in studying how "nice" the isomorphism can be. By this we mean can ϕ be chosen so that the map $x \to (\phi(x))_0$ is continuous and if so what is its best possible modulus of continuity?

We say that a map ϕ from $\mathbb{N}^{\mathbb{Z}}$ to $\mathbb{N}^{\mathbb{Z}}$ is **finitary** if in order to determine $(\phi(x))_0$ we only need to know finitely many coordinates of x. More precisely if for almost every x there exists $m(x)$ such that

$$\mu\Big(\big\{x' : x'_i = x_i \text{ for all } |i| \le m(x) \\ \text{and} \quad \big(\phi(x')\big)_0 \ne \phi(x)_0\big\}\Big) = 0\,.$$

We say that $m(x)$ has **finite tth moment** if $\int m(x)^t \, \mathrm{d}\mu < \infty$ and that ϕ has **finite expected coding length** if the first moment of $m(x)$ is finite.

Keane and Smorodinsky proved the following strengthening of Ornstein's isomorphism theorem.

Theorem 13 [30] *If $\mathbf{p}$ and $\mathbf{q}$ are probability vectors with $H(\mathbf{p}) = H(\mathbf{q})$ then the Bernoulli shifts Bernoulli_p and Bernoulli_q are isomorphic and there exists an isomorphism ϕ such that ϕ and ϕ^{-1} are both finitary.*

The nicest that we could hope ϕ to be is if both ϕ and ϕ^{-1} are finitary and have finite expected coding length. Schmidt proved that this happens only in the trivial case that $\mathbf{p}$ *and* $\mathbf{q}$ *are rearrangements of each other.*

Theorem 14 [70] *If $\mathbf{p}$ and $\mathbf{q}$ are probability vectors and the Bernoulli shifts Bernoulli_p and Bernoulli_q are isomorphic and there exists an isomorphism ϕ such that ϕ and ϕ^{-1} are both finitary and have finite expected coding time then $\mathbf{p}$ is a rearrangement of $\mathbf{q}$.*

The best known result about this problem is the following theorem of Harvey and Peres.

Theorem 15 [14] *If* $\mathbf{p}$ *and* $\mathbf{q}$ *are probability vectors with* $H(\mathbf{p}) = H(\mathbf{q})$ *then* $\sum_i (p_i)^2 \log(p_i) = \sum_i (q_i)^2 \log(q_i)$ *if and only if the Bernoulli shifts* Bernoulli_p *and* Bernoulli_q *are isomorphic and there exists an isomorphism* ϕ *such that* ϕ *and* ϕ^{-1} *are both finitary and have finite one half moment.*

Flows

A **flow** is a measure preserving action of the group $\mathbb{R}$ on a measure space $(\mathbf{X}, T)$. A **cross section** is any measurable set in $C \subset X$ such that for almost every x

$$0 < \inf\{t\colon T_t(x) \in C\} < \infty\,.$$

For any flow $(\mathbf{X}, T)$ and $\{T_t\}_{t\in R}$ and any cross section C we define the **return time map for** C $R\colon C \to C$ as follows. For any $x \in C$ define

$$t(x) = \inf\{t\colon T_t(x) \in C\}$$

then set $R(x) = T_{t(x)}(x)$. There is a standard method to project the probability measure μ on X to an invariant probability measure μ_C on C as well as the σ-algebra $\mathcal{B}$ on X to a σ-algebra $\mathcal{B}_C$ on C such that $(C, \mu_C, \mathcal{B}_C)$ and R is a measure preserving transformation.

First we show that there is a natural analog of Bernoulli shifts for flows.

Theorem 16 [45] *There exists a flow* $(\mathbf{X}, T)$ *and* $\{T_t\}_{t\in\mathbb{R}}$ *such that for every* $t > 0$ *the map* $(\mathbf{X}, T)$ *and* T_t *is isomorphic to a Bernoulli shift. Moreover for any* $h \in (0, \infty]$ *there exists* $(\mathbf{X}, T)$ *and* $\{T_t\}_{t\in\mathbb{R}}$ *such that* $h(T_1) = h$.

We say that such a flow $(X, \{f_t\}_{t\in\mathbb{R}}, \mu, \mathcal{B})$ is a **Bernoulli flow**.

This next version of Ornstein's isomorphism theorem shows that up to isomorphism and a change in time (considering the flow $\mathbf{X}$ and $\{T_{ct}\}$ instead of $\mathbf{X}$ and $\{T_t\}$) there are only two Bernoulli flows, one with positive but finite entropy and one with infinite entropy.

Theorem 17 [45] *If* $(\mathbf{X}, T)$ *and* $\{T_t\}_{t\in\mathbb{R}}$ *and* $(\mathbf{Y}, S)$ $\{S_t\}_{t\in\mathbb{R}}$ *are Bernoulli flows and* $h(T_1) = h(S_1)$ *then they are isomorphic.*

As in the case of actions of $\mathbb{Z}$ there are many natural examples of flows that are isomorphic to the Bernoulli flow. The first is for geodesic flows. In the 1930s Hopf proved that geodesic flows on compact surfaces of constant negative curvature are ergodic [22]. Ornstein and Weiss extended Hopf's proof to show that the geodesic flow is also Bernoulli [52].

The second class of flows comes from billiards on a square table with one circular bumper. The state space X consists of all positions and velocities for a fixed speed. The flow T_t is frictionless movement for time t with elastic collisions. This flow is also isomorphic to the Bernoulli flow [11].

Other Equivalence Relations

In this section we will discuss a number of equivalence relations between transformations that are weaker than isomorphism. All of these equivalence relations have a theory that is parallel to Ornstein's theory.

Kakutani Equivalence

We say that two transformations $(\mathbf{X}, T)$ and $(\mathbf{Y}, S)$ are **Kakutani equivalent** if there exist subsets $A \subset X$ and $B \subset Y$ such that (T_A, A, μ_A) and (S_B, B, ν_B) are isomorphic. This is equivalent to the existence of a flow $(\mathbf{X}, T)$ and $\{T_t\}_{t\in\mathbb{R}}$ with cross sections C and C' such that the return time maps of C and C' are isomorphic to $(\mathbf{X}, T)$ and $(\mathbf{Y}, S)$ respectively.

Using the properties of entropy of the induced map we have that if $(\mathbf{X}, T)$ and $(\mathbf{Y}, S)$ are Kakutani equivalent then either $h((\mathbf{X}, T)) = h((\mathbf{Y}, S)) = 0, 0 < h((\mathbf{X}, T)), h((\mathbf{Y}, S)) < \infty$ or $h((\mathbf{X}, T)) = h((\mathbf{Y}, S)) = \infty$.

In general the answer to the question of which pairs of measure preserving transformations are isomorphic is quite poorly understood. But if one of the transformations is a Bernoulli shift then Ornstein's theory gives a fairly complete answer to the question. A similar situation exists for Kakutani equivalence. In general the answer to the question of which pairs of measure preserving transformation are Kakutani equivalent is also quite poorly understood. But the more specialized question of which transformations are isomorphic to a Bernoulli shift has a more satisfactory answer.

Feldman constructed a transformation $(\mathbf{X}, T)$ which has completely positive entropy but $(\mathbf{X}, T)$ is not Kakutani equivalent to a Bernoulli shift. Ornstein, Rudolph and Weiss extended Feldman's work to construct a complete theory of the transformations that are Kakutani equivalent to a Bernoulli shift [49] for positive entropy transformations and a theory of the transformations that are Kakutani equivalent to an irrational rotation [49] for zero entropy transformations. (The zero entropy version of this theorem had been developed independently (and earlier) by Katok [26].)

They defined two class of transformations called **loosely** Bernoulli and **finitely fixed**. The definitions of these properties are the same as the definitions of very

weak Bernoulli and finitely determined except that the $\bar{d}$ metric is replaced by the $\bar{f}$ metric. For $x, y \in \mathbb{N}^{\mathbb{Z}}$ we define

$$\bar{f}_n(x, y) = 1 - \frac{k}{n}$$

where k is the largest number such that sequences $1 \le i_1 < i_2 < \cdots < i_k \le n$ and $1 \le j_1 < j_2 < \cdots < j_k \le n$ such that $x_{i_l} = y_{j_l}$ for all j, $1 \le j \le k$. (In computer science this metric is commonly referred to as the edit distance.) Note that $\bar{d}_n(x, y) \ge \bar{f}_n(x, y)$.

They proved the following analog of Theorem 5.

Theorem 18 *For transformations* $(\mathbf{X}, T)$ *with* $h((\mathbf{X}, T)) > 0$ *the following conditions are equivalent:*

(1) $(\mathbf{X}, T)$ *is finitely fixed,*
(2) $(\mathbf{X}, T)$ *is loosely Bernoulli,*
(3) $(\mathbf{X}, T)$ *is Kakutani equivalent to a Bernoulli shift and*
(4) There exists a Bernoulli flow $\boldsymbol{Y}$ *and* $\{F_t\}_{t\in\mathbb{R}}$ *and a cross section* C *such that the return time map for* C *is isomorphic to* $(\mathbf{X}, T)$.

Restricted Orbit Equivalence

Using the $\bar{d}$ metric we got a theory of which transformations are isomorphic to a Bernoulli shift. Using the $\bar{f}$ metric we got a strikingly similar theory of which transformations are Kakutani equivalent to a Bernoulli shift. Rudolph showed that it is possible to replace the $\bar{d}$ metric (or the $\bar{f}$ metric) with a wide number of other metrics and produce parallel theories for other equivalence relations. For instance, for each of these theories we get a version of Theorem 5. This collection of theories is called restricted orbit equivalence [64].

Non-invertible Transformations

The question of which noninvertible measure preserving transformations are isomorphic turns out to be quite different from the same question for invertible transformations. In one sense it is easier because of an additional isomorphism invariant.

For any measure preserving transformation $(\mathbf{X}, T)$ the probability measure $\mu|_{T^{-1}(x)}$ on $T^{-1}(x)$ is defined for almost every $x \in X$. (If $(\mathbf{X}, T)$ is invertible then this measure is trivial as $|\{T^{-1}(x)\}| = 1$ and $\mu|_{T^{-1}(x)}(T^{-1}(x)) = 1$ for almost every x.) It is easy to check that if ϕ is an isomorphism from $(\mathbf{X}, T)$ to $(\mathbf{Y}, S)$ then for almost every x and $x' \in T^{-1}(x)$ we have

$$\mu\,|_{T^{-1}(x)}\,(x') = \nu\,|_{S^{-1}(\phi(x))}\,(\phi\,(x'))\ .$$

From this we can easily see that if $\mathbf{p} = \{p_i\}_{i=1}^n$ and $\mathbf{q} = \{q_i\}_{i=1}^m$ are probability vectors then the corresponding one sided Bernoulli shifts are isomorphic only if $m = n$ and there is a permutation $\pi \in S_n$ such that $p_{\pi(i)} = q_i$ for all i. (In this case we say $\mathbf{p}$ is a **rearrangement** of $\mathbf{q}$.) If $\mathbf{p}$ is a rearrangement of $\mathbf{q}$ then it is easy to construct an isomorphism between the corresponding Bernoulli shifts. Thus the analog of Ornstein's theorem for Bernoulli endomorphisms is trivial. However we will see that there still is an analogous theory classifying the class of endomorphism that are isomorphic to Bernoulli endomorphisms.

We say that an endomorphism is **uniformly** d **to 1** if for almost every x we have that $|\{T^{-1}(x)\}| = d$ and $\mu|_{T^{-1}(x)}(y) = 1/d$ for all $y \in T^{-1}(x)$. Hoffman and Rudolph defined two classes of noninvertible transformations called **tree very week Bernoulli** and **tree finitely determined** and proved the following theorem.

Theorem 19 *The following three conditions are equivalent for uniformly* d *to 1 endomorphisms.*

(1) $(\mathbf{X}, T)$ *is tree very weak Bernoulli,*
(2) $(\mathbf{X}, T)$ *is tree finitely determined, and*
(3) $(\mathbf{X}, T)$ *is isomorphic to the one sided Bernoulli* d *shift.*

Jong extended this theorem to say that if there exists a probability vector $\mathbf{p}$ such that for almost every x the distribution of $\mu|_{T-1}(x)$ is the same as the distribution of $\mathbf{p}$ then $(\mathbf{X}, T)$ is isomorphic to the one sided Bernoulli d shift if and only if it is tree finitely determined and if and only if it is tree very weak Bernoulli [23].

Markov Shifts

We saw that mixing Markov chains are isomorphic if they have the same entropy. As we have seen there are additional isomorphism invariants for noninvertible transformations. Ashley, Marcus and Tuncel managed to classify all one sided mixing Markov chains up to isomorphism [1].

Rational Maps

Rational maps are the main object of study in complex dynamics. For every rational function $f(z) = p(z)/q(z)$ there is a nonempty compact set J_f which is called the **Julia set**. Roughly speaking this is the set of points for which every neighborhood acts "chaotically" under repeated iterations of f.

In order to consider rational maps as measure preserving transformations we need to specify an invariant measure. The following theorem of Gromov shows that for ev-

ery rational map there is one canonical measure to consider.

Theorem 20 [12] *For every f rational function of degree d there exists a unique invariant measure μ_f of maximal entropy. We have that $h(\mu_f) = \log_2 d$ and $\mu_f(J_f) = 1$.*

The properties of this measure were studied by Freire, Lopes and Mañé [8]. Mañé that analysis to prove the following theorem.

Theorem 21 [38] *For every rational function f of degree d there exists n such that $(\mathbb{C}, f^n, \mu_f)$ (where $f^n(z) = f(f(f \ldots (z)))$ is composition) is isomorphic to the one sided Bernoulli d^n shift.*

Heicklen and Hoffman used the tree very weak Bernoulli condition to show that we can always take n to be one.

Theorem 22 [16] *For every rational function f of degree $d \geq 2$ the corresponding map $((\mathbb{C}, \mu_f, \mathcal{B}), T_f)$ is isomorphic to the one sided Bernoulli d shift.*

Differences with Ornstein's Theory

Unlike Kakutani equivalence and restricted orbit equivalence which are very close parallels to Ornstein's theory, the theory of which endomorphisms are isomorphic to a Bernoulli endomorphism contains some significant differences. One of the principal results of Ornstein's isomorphism theory is that if $(\mathbf{X}, T)$ is an invertible transformation and $(\mathbf{X}, T^2)$ is isomorphic to a Bernoulli shift then $(\mathbf{X}, T)$ is also isomorphic to a Bernoulli shift. There is no corresponding result for noninvertible transformations.

Theorem 23 [19] *There is a uniformly two to one endomorphism $(\mathbf{X}, T)$ which is not isomorphic to the one sided Bernoulli 2 shift but $(\mathbf{X}, T^2)$ is isomorphic to the one sided Bernoulli 4 shift.*

Factors of a Transformation

In this section we study the relationship between a transformation and its factors. There is a natural way to associate a factor of $(\mathbf{X}, T)$ with a sub σ-algebra of $\mathcal{B}$. Let $(\mathbf{Y}, S)$ be a factor of $(\mathbf{X}, T)$ with factor map $\phi\colon X \to Y$. Then the σ-algebra associated with $(\mathbf{Y}, S)$ is $\mathcal{B}_Y = \phi^{-1}(C)$. Thus the study of factors of a transformation is the study of its sub σ-algebras.

Almost every property that we have discussed above has an analog in the study of factors of a transformation. We give three such examples. We say that two factors C and $\mathcal{D}$ of $(\mathbf{X}, T)$ are **relatively isomorphic** if there exists an isomorphism $\psi\colon X \to X$ of $(\mathbf{X}, T)$ with itself such that $\psi(C) = \mathcal{D}$. We say that $(\mathbf{X}, T)$ has **relatively completely positive entropy with respect to** C if every factor $\mathcal{D}$ which contains $\mathcal{D}$ has $h(\mathcal{D}) > h(C)$. We say that C is **relatively Bernoulli** if there exists a second factor $\mathcal{D}$ which is independent of $\mathcal{D}$ and $\mathcal{B} = C \vee \mathcal{D}$.

Thouvenot defined properties of factors called **relatively very weak Bernoulli** and **relatively finitely determined**. Then he proved an analog of Theorem 3. This says that a factor being relatively Bernoulli is equivalent to it being relatively finitely determined (and also equivalent to it being relatively very weak Bernoulli).

The **Pinsker algebra** is the maximal σ-algebra $\mathcal{P}$ such that $h(\mathcal{P}) = 0$. The Pinsker conjecture was that for every measure preserving transformation $(\mathbf{X}, T)$ there exists a factor C such that

(1) C is independent of the Pinsker algebra $\mathcal{P}$
(2) $\mathcal{B} = C \vee \mathcal{P}$ and
(3) $(\mathbf{X}, C)$ has completely positive entropy.

Ornstein found a counterexample to the Pinsker conjecture [48]. After Thouvenot developed the relative isomorphism theory he came up with the following question which is referred to as the weak Pinsker conjecture.

Conjecture 2 *For every measure preserving transformation $(\mathbf{X}, T)$ and every $\epsilon > 0$ there exist invariant σ-algebras $C, \mathcal{D} \subset \mathcal{B}$ such that*

(1) C is independent of $\mathcal{D}$
(2) $\mathcal{B} = C \vee \mathcal{D}$
(3) $(X, T, \mu, \mathcal{D})$ is isomorphic to a Bernoulli shift
(4) $h((X, T, \mu, C)) < \epsilon$.

There is a wide class of transformations which have been proven to satisfy the weak Pinsker conjecture. This class includes almost all measure preserving transformations which have been extensively studied.

Actions of Amenable Groups

All of the discussion above has been about the action of a single invertible measure preserving transformation (actions of $\mathbb{N}$ and $\mathbb{Z}$) or flows (actions of $\mathbb{R}$). We now consider more general group actions. If we have two actions S and T on a measure space (X, μ) which commute ($S(T(x)) = T(S(x))$ for almost every x) then there is an action of $\mathbb{Z}^2$ on (X, μ) given by $f_{(n,m)}(x) = S^n(T^m(x))$. A natural question to ask is do there exist a version of entropy theory and Ornstein's isomorphism theory for actions of two commuting automorphisms. More generally for each of the results discussed above we can ask what is the largest class of groups such that an analogous result is

true. It turns out that for most of the properties described above the right class of groups is discrete amenable groups.

A **Følner sequence** F_n in a group G is a sequence of subsets F_n of G such that for all $g \in G$ we have that $\lim_{n\to\infty} |g(F_n)|/|F_n|$. A countable group is **amenable** if and only if it has a Følner sequence.

For nonamenable groups it is much more difficult to generalize Birkhoff's ergodic theorem [41,42]. Lindenstrauss proved that for every discrete amenable group there is an analog of the ergodic theorem [35]. For every amenable group G and every probability vector $\mathbf{p}$ we can define a Bernoulli action of G. There are also analogs of Rokhlin's lemma and the Shannon–McMillan–Breiman theorem for actions of all discrete amenable groups [33,51,54]. Thus we have all of the ingredients to prove a version of Ornstein's isomorphism theorem.

Theorem 24 *If* $\mathbf{p}$ *and* $\mathbf{q}$ *are probability vectors and* $H(\mathbf{p}) = H(\mathbf{q})$ *then the Bernoulli actions of* G *corresponding to* $\mathbf{p}$ *and* $\mathbf{q}$ *are isomorphic.*

Also all of the aspects of Rudolph's theory of restricted orbit equivalence can be carried out for actions of amenable groups [25].

Differences Between Actions of $\mathbb{Z}$ and Actions of Other Groups

Although generalizations of Ornstein theory and restricted orbit equivalence carry over well to the actions of discrete amenable groups there do turn out to be some significant differences between the possible actions of $\mathbb{Z}$ and those of other discrete amenable groups.

Many of these have to do with the generalization of Markov shifts. For actions of $\mathbb{Z}^2$ these are called Markov random fields. By the result of Friedman and Ornstein if a Markov chain is mixing then it has completely positive entropy and it is isomorphic to a Bernoulli shift. Mixing Markov random fields can have very different properties. Ledrappier constructed a $\mathbb{Z}^2$ action which is a Markov random field and is mixing but has zero entropy [34]. Even more surprising even though it is mixing it is not mixing of all orders. The existence of a $\mathbb{Z}$ action which is mixing but not mixing of all orders is one of the longest standing open questions in ergodic theory [13].

Even if we try to strengthen the hypothesis of Friedman and Ornstein's theorem to assume that the Markov random field has completely positive entropy we will not succeed as there exists a Markov random field which has completely positive entropy but is not isomorphic to a Bernoulli shift [18].

Future Directions

In the future we can expect to see progress of isomorphism theory in a variety of different directions. One possible direction for future research is better understand the properties of finitary isomorphisms between various transformations and Bernoulli shifts described in Sect. "Finitary Isomorphisms". Another possible direction would be to find a theory of equivalence relations for Bernoulli endomorphisms analogous to the one for invertible Bernoulli transformations described in Sect. "Other Equivalence Relations".

As the subject matures the focus of research in isomorphism theory will likely shift to connections to other fields. Already there are deep connections between isomorphism theory and both number theory and statistical physics. Finally one hopes to see progress made on the two dominant outstanding conjectures in the field: Thouvenot weak Pinsker conjecture (Conjecture 2) and Furstenberg's conjecture (Conjecture 1) about measures on the circle invariant under both the times 2 and times 3 maps. Progress on either of these conjectures would invariably lead the field in exciting new directions.

Bibliography

1. Ashley J, Marcus B, Tuncel S (1997) The classification of one-sided Markov chains. Ergod Theory Dynam Syst 17(2):269–295
2. Birkhoff GD (1931) Proof of the ergodic theorem. Proc Natl Acad Sci USA 17:656–660
3. Breiman L (1957) The individual ergodic theorem of information theory. Ann Math Statist 28:809–811
4. Den Hollander F, Steif J (1997) Mixing E properties of the generalized T, T^{-1}-process. J Anal Math 72:165–202
5. Einsiedler M, Lindenstrauss E (2003) Rigidity properties of $\mathbb{Z}^d$-actions on tori and solenoids. Electron Res Announc Amer Math Soc 9:99–110
6. Einsiedler M, Katok A, Lindenstrauss E (2006) Invariant measures and the set of exceptions to Littlewood's conjecture. Ann Math (2) 164(2):513–560
7. Feldman J (1976) New K-automorphisms and a problem of Kakutani. Isr J Math 24(1):16–38
8. Freire A, Lopes A, Mañé R (1983) An invariant measure for rational maps. Bol Soc Brasil Mat 14(1):45–62
9. Friedman NA, Ornstein DS (1970) On isomorphism of weak Bernoulli transformations. Adv Math 5:365–394
10. Furstenberg H (1967) Disjointness in ergodic theory, minimal sets, and a problem in Diophantine approximation. Math Syst Theory 1:1–49
11. Gallavotti G, Ornstein DS (1974) Billiards and Bernoulli schemes. Comm Math Phys 38:83–101
12. Gromov M (2003) On the entropy of holomorphic maps. Enseign Math (2) 49(3–4):217–235
13. Halmos PR (1950) Measure theory. D Van Nostrand, New York
14. Harvey N, Peres Y. An invariant of finitary codes with finite expected square root coding length. Ergod Theory Dynam Syst, to appear

15. Heicklen D (1998) Bernoullis are standard when entropy is not an obstruction. Isr J Math 107:141–155
16. Heicklen D, Hoffman C (2002) Rational maps are *d*-adic Bernoulli. Ann Math (2) 156(1):103–114
17. Hoffman C (1999) A *K* counterexample machine. Trans Amer Math Soc 351(10):4263–4280
18. Hoffman C (1999) A Markov random field which is *K* but not Bernoulli. Isr J Math 112:249–269
19. Hoffman C (2004) An endomorphism whose square is Bernoulli. Ergod Theory Dynam Syst 24(2):477–494
20. Hoffman C (2003) The scenery factor of the $[T, T^{-1}]$ transformation is not loosely Bernoulli. Proc Amer Math Soc 131(12):3731–3735
21. Hoffman C, Rudolph D (2002) Uniform endomorphisms which are isomorphic to a Bernoulli shift. Ann Math (2) 156(1):79–101
22. Hopf E (1971) Ergodic theory and the geodesic flow on surfaces of constant negative curvature. Bull Amer Math Soc 77: 863–877
23. Jong P (2003) On the isomorphism problem of p-endomorphisms. Ph D thesis, University of Toronto
24. Kalikow SA (1982) T, T^{-1} transformation is not loosely Bernoulli. Ann Math (2) 115(2):393–409
25. Kammeyer JW, Rudolph DJ (2002) Restricted orbit equivalence for actions of discrete amenable groups. Cambridge tracts in mathematics, vol 146. Cambridge University Press, Cambridge
26. Katok AB (1975) Time change, monotone equivalence, and standard dynamical systems. Dokl Akad Nauk SSSR 223(4):789–792; in Russian
27. Katok A (1980) Smooth non-Bernoulli *K*-automorphisms. Invent Math 61(3):291–299
28. Katok A, Spatzier RJ (1996) Invariant measures for higher-rank hyperbolic abelian actions. Ergod Theory Dynam Syst 16(4):751–778
29. Katznelson Y (1971) Ergodic automorphisms of T^n are Bernoulli shifts. Isr J Math 10:186–195
30. Keane M, Smorodinsky M (1979) Bernoulli schemes of the same entropy are finitarily isomorphic. Ann Math (2) 109(2):397–406
31. Kolmogorov AN (1958) A new metric invariant of transient dynamical systems and automorphisms in Lebesgue spaces. Dokl Akad Nauk SSSR (NS) 119:861–864; in Russian
32. Kolmogorov AN (1959) Entropy per unit time as a metric invariant of automorphisms. Dokl Akad Nauk SSSR 124:754–755; in Russian
33. Krieger W (1970) On entropy and generators of measure-preserving transformations. Trans Amer Math Soc 149:453–464
34. Ledrappier F (1978) Un champ markovien peut être d'entropie nulle et mélangeant. CR Acad Sci Paris Sér A–B 287(7):A561–A563; in French
35. Lindenstrauss E (2001) Pointwise theorems for amenable groups. Invent Math 146(2):259–295
36. Lyons R (1988) On measures simultaneously 2- and 3-invariant. Isr J Math 61(2):219–224
37. Mañé R (1983) On the uniqueness of the maximizing measure for rational maps. Bol Soc Brasil Mat 14(1):27–43
38. Mañé R (1985) On the Bernoulli property for rational maps. Ergod Theory Dynam Syst 5(1):71–88
39. Meilijson I (1974) Mixing properties of a class of skew-products. Isr J Math 19:266–270
40. Meshalkin LD (1959) A case of isomorphism of Bernoulli schemes. Dokl Akad Nauk SSSR 128:41–44; in Russian
41. Nevo A (1994) Pointwise ergodic theorems for radial averages on simple Lie groups. I. Duke Math J 76(1):113–140
42. Nevo A, Stein EM (1994) A generalization of Birkhoff's pointwise ergodic theorem. Acta Math 173(1):135–154
43. Ornstein DS (1973) A *K* automorphism with no square root and Pinsker's conjecture. Adv Math 10:89–102
44. Ornstein D (1970) Factors of Bernoulli shifts are Bernoulli shifts. Adv Math 5:349–364
45. Ornstein D (1970) Two Bernoulli shifts with infinite entropy are isomorphic. Adv Math 5:339–348
46. Ornstein D (1970) Bernoulli shifts with the same entropy are isomorphic. Adv Math 4:337–352
47. Ornstein DS (1973) An example of a Kolmogorov automorphism that is not a Bernoulli shift. Adv Math 10:49–62
48. Ornstein DS (1973) A mixing transformation for which Pinsker's conjecture fails. Adv Math 10:103–123
49. Ornstein DS, Rudolph DJ, Weiss B (1982) Equivalence of measure preserving transformations. Mem Amer Math Soc 37(262). American Mathematical Society
50. Ornstein DS, Shields PC (1973) An uncountable family of *K*-automorphisms. Adv Math 10:63–88
51. Ornstein D, Weiss B (1983) The Shannon–McMillan–Breiman theorem for a class of amenable groups. Isr J Math 44(1): 53–60
52. Ornstein DS, Weiss B (1973) Geodesic flows are Bernoullian. Isr J Math 14:184–198
53. Ornstein DS, Weiss B (1974) Finitely determined implies very weak Bernoulli. Isr J Math 17:94–104
54. Ornstein DS, Weiss B (1987) Entropy and isomorphism theorems for actions of amenable groups. J Anal Math 48:1–141
55. Parry W (1981) Topics in ergodic theory. Cambridge tracts in mathematics, vol 75. Cambridge University Press, Cambridge
56. Parry W (1969) Entropy and generators in ergodic theory. WA Benjamin, New York
57. Petersen K (1989) Ergodic theory. Cambridge studies in advanced mathematics, vol 2. Cambridge University Press, Cambridge
58. Pinsker MS (1960) Dynamical systems with completely positive or zero entropy. Dokl Akad Nauk SSSR 133:1025–1026; in Russian
59. Ratner M (1978) Horocycle flows are loosely Bernoulli. Isr J Math 31(2):122–132
60. Ratner M (1982) Rigidity of horocycle flows. Ann Math (2) 115(3):597–614
61. Ratner M (1983) Horocycle flows, joinings and rigidity of products. Ann Math (2) 118(2):277–313
62. Ratner M (1991) On Raghunathan's measure conjecture. Ann Math (2) 134(3):545–607
63. Halmos PR (1960) Lectures on ergodic theory. Chelsea Publishing, New York
64. Rudolph DJ (1985) Restricted orbit equivalence. Mem Amer Math Soc 54(323). American Mathematical Society
65. Rudolph DJ (1979) An example of a measure preserving map with minimal self-joinings, and applications. J Anal Math 35:97–122
66. Rudolph DJ (1976) Two nonisomorphic *K*-automorphisms with isomorphic squares. Isr J Math 23(3–4):274–287
67. Rudolph DJ (1983) An isomorphism theory for Bernoulli free *Z*-skew-compact group actions. Adv Math 47(3):241–257
68. Rudolph DJ (1988) Asymptotically Brownian skew products

give non-loosely Bernoulli *K*-automorphisms. Invent Math 91(1):105–128
69. Rudolph DJ (1990) $\times 2$ and $\times 3$ invariant measures and entropy. Ergod Theory Dynam Syst 10(2):395–406
70. Schmidt K (1984) Invariants for finitary isomorphisms with finite expected code lengths. Invent Math 76(1):33–40
71. Shields P (1973) The theory of Bernoulli shifts. Chicago lectures in mathematics. University of Chicago Press, Chicago
72. Sinaĭ JG (1962) A weak isomorphism of transformations with invariant measure. Dokl Akad Nauk SSSR 147:797–800; in Russian
73. Sinaĭ J (1959) On the concept of entropy for a dynamic system. Dokl Akad Nauk SSSR 124:768–771; in Russian
74. Thouvenot J-P (1975) Quelques propriétés des systèmes dynamiques qui se décomposent en un produit de deux systèmes dont l'un est un schéma de Bernoulli. Conference on ergodic theory and topological dynamics, Kibbutz, Lavi, 1974. Isr J Math 21(2–3):177–207; in French
75. Thouvenot J-P (1975) Une classe de systèmes pour lesquels la conjecture de Pinsker est vraie. Conference on ergodic theory and topological dynamics, Kibbutz Lavi, 1974. Isr J Math 21(2–3):208–214; in French

J

Jamming of Granular Matter

Bulbul Chakraborty[1], Bob Behringer[2]
[1] Department of Physics, Brandeis University, Waltham, USA
[2] Department of Physics, Duke University, Durham, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Statistical Framework of Jamming
Jamming Phase Diagram
Force Fluctuations, Dynamical Heterogeneities, Length and Time Scales in Granular Flows Approaching Jamming
Force Distributions, Length and Time Scales in Jammed States near the Jamming Transition
Isostaticity and Point J
Future Directions
Bibliography

Glossary

Granular matter Material such as sand or sugar, which is composed of independent, macroscopic particles characterized by short range interactions that do not conserve energy. Energy is lost to excitations of internal degrees of freedom, and is then unavailable for macroscopic flow.

Supercooled liquid A liquid cooled below its freezing point by avoiding crystallization.

Shear deformation Deformation of a material in which internal clusters or layers slide past each other.

Couette flow Flow between two surfaces one of which is moving with respect to the other.

Definition of the Subject

A jammed material is defined as one that is structurally disordered but, unlike a fluid, possesses an yield stress. In the field of traditional condensed matter physics, such a material would be called an amorphous solid. The broader use of "jammed" extends this concept to non-traditional materials such as granular systems, foams and colloids. Jamming is, similarly, the extension of the concept of freezing to the transition from a fluid state to a jammed state. Understanding jamming in granular systems is important from technological, environmental, and basic science perspectives. Jamming of grains in silos causes catastrophic failures. Avalanches are examples of unjamming, which we need to understand prevent and control. The phenomenon of jamming in granular matter poses fundamental challenges in basic science because there is no known framework leading from the microscopic, grain level interactions to the macroscopic properties that reflect collective behavior.

Jamming in granular matter is intimately related to stress propagation, and the nature of jamming will depend on whether the material is under shear or isotropic compression. It will also depend on whether there is sustained motion with the grains having a finite kinetic energy or if the system is at rest and being slowly deformed. In the following sections, we explore different types of jamming in granular matter.

Introduction

Experiments point to many similarities between the jamming transition in granular materials [1,2,3] and other similar non-equilibrium transitions such as that in driven foams, emulsions and gels. The common unifying features are: (a) these are purely non-equilibrium transitions since the flowing state can be maintained only by external driving, (b) the systems exhibit anomalously slow dynamics as the transition is approached and (c) there are no obvious structural signatures of the transition. By structural, we refer to the geometric arrangement of particles. In addition,

we will see that there is increasing evidence for a divergence in correlation functions for forces in the granular case. Features (b) and (c) are, in fact, similar to a transition in a thermal system; that of a supercooled liquid undergoing the glass transition. The origin of the slow dynamics and the relationship between time and length scales has been studied extensively in the context of the glass transition in supercooled liquids, and it has been shown that the dynamics becomes spatially heterogeneous as the glass transition is approached [4,5,6,7]. There is a time scale and a length scale associated with dynamical heterogeneities and these have been measured in experiments [7] and simulations [8]. Inspired by this framework, dynamical heterogeneities have been explored in experiments on granular systems approaching jamming and a length and time scale associated with the jamming transition has been identified [9,10,11].

In granular systems, there is a type of heterogeneity arising from stress transmission which is distinct from the heterogeneity of particle displacements and density fluctuations that are common to both supercooled liquids and granular matter. Stress heterogeneities in flowing granular matter has been studied in experiments on Couette shear flow [12,13,14,15,16,17,18] and in hopper flow [19]. Simulations have a played a seminal role in the developments leading up to the dynamical heterogeneities framework in supercooled liquids. Likewise, simulations of granular flows can play a role in clarifying the relationship between heterogeneity in different dynamical variables. For instance, a simulation of grains flowing through a hopper (a long, vertical tube with a narrow opening at the bottom), has identified/characterized both stress and displacement heterogeneities. However, a clear relationship between the two has not been identified [20]. Indeed, there may be some reason to think that there is anti-correlation between the strong force structures of force chains, and the freedom needed for density fluctuations. Dynamical heterogeneities have been explored in simulations of other types of dense granular flows [21,22,23].

The similarities in dynamics between supercooled liquids and granular systems have raised the intriguing possibility of "universality" in the jamming transition. Universality is a concept associated with continuous phase transitions in thermal systems such as the liquid-gas transition or the Curie point in ferromagnets [24]. In these systems, universality, owes its origin to the presence of a diverging length-scale at the transition. Microscopic properties of the system can be coarse-grained because of the presence of this diverging length scale and universality classes can be specified by properties such as the symmetry of the system and the spatial dimensionality. These concepts are still evolving in the context of granular systems. However, there are an increasing number of studies that have identified a growing, possibly diverging, length scale.

Statistical Framework of Jamming

The ubiquitous presence of force fluctuations, well-defined statistical distributions and similarities to thermal systems observed in granular systems close to jamming, have led researchers to look for a statistical framework that can describe the spatial and temporal fluctuations in static and quasi-static (weakly driven) granular flows.

Experiments and simulations suggest that the phenomenon of jamming has the flavor of a phase transition. Phase transitions can be broadly divided into two classes: (a) Equilibrium and (b) Non-equilibrium. The former encompasses transitions such as the liquid-solid transition, the superfluid and superconducting transitions, and describes transitions between states which are in thermal equilibrium. Non-equilibrium phase transitions are commonly observed in pattern-forming systems such as chemical reactions, a well known example being the Belousov–Zhabotinsky reaction [25], or in fluid flows, such as Raleigh–Bénard convection. The framework of equilibrium statistical mechanics is well established and can be used to analyze and understand equilibrium phase transitions. In contrast, there is no universal framework for nonequilibrium phase transitions even though there are many specific transitions which are well understood [26,27]. In many cases where non-equilibrium transitions occur, the systems are locally close to thermal equilibrium, but not globally so. It is then possible to use hydrodynamic-like descriptions involving gradients in relevant quantities, as well as transport models. Transitions in such systems often involve spatially extended modes which become unstable or stable for different control parameters. In the usual case, it is the most unstable mode which dominates, and thermal fluctuations play a very limited role.

An interesting question is whether a statistical approach can be used to understand jamming and other phenomena in granular matter. In such an approach, one would like to know the ensemble of states that are consistent with the external constraints. In the granular case, there is some evidence to suggest that fluctuations are important, particulary near jamming unlike the typical non-equilibrium transition discussed above.

Granular Packings

A static granular packing is in mechanical equilibrium if every grain satisfies the conditions of force and torque bal-

ance. At the microscopic level, the grains interact via contact forces, which are frictional. For rigid grains, friction leads to a microscopic indeterminacy of the forces since the tangential force magnitude at a contact has to be *less* than the normal force: $|f_t| \leq \mu f_n$, where μ is the friction coefficient.

There is a broader class of indeterminacy which has to do with the fact that there are many packing configurations that are consistent with a set of macroscopic constraints. For example, if one shakes a box of marbles and examines the positions of the marbles after the shaking has stopped, the positions and the contact network will be different each time one stops shaking. This is analogous to the situation of systems in thermal equilibrium: many microscopic configurations of the oxygen molecules in a closed room occur while the temperature and total volume remain fixed. Of course, in the granular case, there is no internal 'shaking' mechanism.

The different configurations of systems in thermal equilibrium (such as the oxygen molecules in a room) have a given probability of occurring at a given temperature. This probability is well known from equilibrium statistical mechanics and is the Boltzmann distribution: $e^{-\beta E}$. Here $\beta = 1/k_B T$ is the inverse of the temperature T multiplied by the Boltzman constant k_B, and E is the energy of the configuration. The probability of finding a configuration is, therefore, completely determined by its energy and does not depend on the particular dynamics or the history. This is the essence of systems in thermal equilibrium and leads to their simplicity.

Edward's Ensemble

The configurations of a box of marbles at the end of a shaking experiment are certainly not in thermal equilibrium. Marbles are large objects, and they dissipate energy as heat when they interact. Ordinary thermal fluctuations are not relevant to a box of marbles; one has to shake them to move them and generate new states. These end states are, however, static. A natural question to ask is whether one can make a priori prediction of the probability of occurrence of a particular configuration, as we can for systems in thermal equilibrium. The first theoretical approach to address this question was that of the Edwards ensemble which asserts that the total volume (V) of a mechanically stable grain packing plays the role of energy, and that there is a temperature-like quantity called the compactivity [28,29,30,31]. The mechanically stable grain packings of infinitely rigid objects have been termed blocked states [31], and are related to strictly jammed states [32]. The basic hypothesis underlying the Edwards ensemble is the microcanonical [24] hypothesis that all blocked states with the same total volume V are equally likely. This hypothesis has been tested in numerical simulations, in experiments, and in some exactly solvable models [33,34,35,36,37,38]. The consensus seems to be that the hypothesis is not universally valid but holds under some conditions. The Edwards hypothesis involves the definition of the Edwards entropy: $S(V) = \ln \Omega(V)$, where $\Omega(V)$ is the density of blocked states, that is, the number of blocked states between V and $V + \delta V$. Evaluating this function has been a challenging problem, although progress has been made recently [39].

Force Ensembles

The Edwards picture was originally based on the assumption of completely rigid particles. However, all physical particles have finite stiffness, and in the usual case, interact via frictional contact forces. The conditions of finite vs. infinite stiffness and frictionless vs. frictional interactions has some important consequences for granular packings. Relaxing the infinite stiffness constraint introduces couplings between the positions of grains and the contact forces. Below we discuss an ensemble for frictionless, spherical particles.

A mechanically stable packing of frictionless, deformable, spherical particles have to satisfy the equations of force balance. In addition, there is force law relating the positions of the particles to the forces [40,41]. For grains interacting through purely repulsive, short-range, forces, there are dN equations of force balance for N grains in a space of dimensionality d:

$$\text{force} - \text{balance} \quad dN \text{ eqs}: \sum_j F_{ij} \frac{\mathbf{r}_{ij}}{|\mathbf{r}_{ij}|} = 0 \tag{1}$$

$$\text{force} - \text{law} \quad \langle z \rangle N/2 \text{ eqs}: \; F_{ij} = f(\mathbf{r}_{ij}) \tag{2}$$

Here $\langle z \rangle$ is the average number of contacts per grain, F_{ij} is the magnitude of the contact force between grains i and j, the angles of the contacts being fixed by the geometry, and $f(\mathbf{r}_{ij})$ is a function specifying the inter-grain force law. For a given geometry, (i. e. fixed $\frac{\mathbf{r}_{ij}}{|\mathbf{r}_{ij}|}$), the equations of force balance involve $\langle z \rangle N/2$ unknowns. The number of force-balance equations cannot be greater than the number of unknowns, otherwise the forces are overdetermined, and, therefore, $\langle z \rangle N/2 \geq dN$.

For rigid, non-deformable grains, the force law becomes a constraint on the positions of the grains [41]. There is one constraint for each contact, which leads to $\langle z \rangle N/2$ equations for the dN positions, the unknowns. Since the number of constraint equations have to be larger

than the number of unknowns, $\langle z \rangle N/2 \leq dN$. The force law and the force balance constraints can, therefore be satisfied only if $\langle z \rangle = z_{\rm iso} = 2d$. This enumeration argument applies only to disordered packings, since for ordered, crystalline packings, angles of the lines connecting the grains are not independent, and therefore, not all the constraint equations are linearly independent. Considering such disordered packings, the enumeration argument leads to the conclusion that packings of rigid, isotropic, frictionless particles have to be *isostatic*; they have to have $\langle z \rangle = z_{\rm iso}$.

For deformable particles, mechanically stable packings can exist for $\langle z \rangle \geq z_{\rm iso}$. If the particles are very stiff, the magnitude of the forces change a lot for small changes in the separation between grains. There is, therefore, an effective separation of scales [40,42] in Eqs. (1) and (2), and one can consider the ensemble of forces which satisfy force-balance for a fixed geometry of the packing, i. e., a packing with fixed grain positions. The properties of such force-ensembles have been studied [40,42,43,44,45,46,47], and it has been shown that $P(F)$, the probability distribution function (PDF) of contact forces, evolves to a exponential form as the particles are made increasingly rigid [40,42]. In addition, work on sheared and isotropically compressed packings have shown that there are more extended spatial correlations of the forces in the force ensembles for sheared packings [48].

The force ensemble approach captures many features of packings near Point J. However, the approach to this point involves coupling between the geometry and forces, as observed in simulations [49]. Below, we discuss a generalized ensemble that allows for the coupling of these two, and applies to spherical and non-spherical grains with and without friction.

Generalized Ensembles

In a more recent development, it has been shown that the "equally likely" hypothesis of Edwards and the Microcanonical ensemble of thermal systems [24] is not essential for the definition of a temperature-like quantity, and a much weaker condition of factorizability of distributions is sufficient [50,51,52]. The necessary conditions for being able to define a temperature-like variable and a statistical ensemble based on this variable are (a) the existence of a physical quantity that is conserved by the natural dynamics of the system (in thermal systems energy is conserved but in dissipative granular media, it is not) and (b) that the frequency of finding different states with the same value of the conserved quantity is factorizable: $\omega_{\nu_1+\nu_2} = \omega_{\nu_1}\,\omega_{\nu_2}$. The latter condition implies that if one creates a configuration by bringing together two configurations $_1$ and $_2$, then the frequency of occurrence of this joint configuration is a product of the frequency ω_ν of the individual configurations.

In the context of mechanically stable packings of grains (soft, deformable or rigid), a conserved quantity that has been identified is the force-moment tensor [53,54,55] which is related to the Cauchy stress tensor:

$$\hat{\sigma} = (1/V) \sum_{ij} \vec{r}_{ij} \vec{F}_{ij} \tag{3}$$

The summation in Eq. (3) is over all contacts $\{ij\}$ in an assembly of grains, occupying a volume V, with contact vectors $\vec{r}_{ij}$ and contact forces $\vec{F}_{ij}$ (for grains with friction $\vec{F}_{ij}$ does not lie along the direction of $\vec{r}_{ij}$).

The microscopic force moment tensor for a grain is given by:

$$\hat{\kappa}_i = \sum_j \vec{r}_{ij} \vec{F}_{ij}\,, \tag{4}$$

where the sum is over all grains j that contact grain i. For grains in mechanical equilibrium, it can be shown, using a generalized Stoke's theorem [28,56,57], that the *total* force moment tensor

$$\hat{\Sigma} = \sum_i \kappa_i$$

becomes a boundary integral for systems with open boundaries and is a topologically conserved quantity for systems with periodic boundary conditions [53]. In two dimensions, this property can be explicitly demonstrated by introducing a set of auxiliary variables [55] which are the analog of the vector potential in electromagnetism. The connection to electromagnetism is natural if one remembers that for mechanically stable packings, the divergence of the stress tensor is zero, just as the divergence of the magnetic field is zero in electromagnetism.

Given the topologically conserved nature and/or the strict boundary sensitivity of $\hat{\Sigma}$, the phase space of all mechanically stable packings can be grouped into sectors characterized by the value of $\hat{\Sigma}$. Thinking of systems with periodic boundary conditions, the topologically conserved nature implies that packings in different sectors are completely disconnected by any natural dynamics, and therefore, $\hat{\Sigma}$ is the type of conserved quantity that meets criterion (a) of the previous paragraph. If we now assume criterion (b), then a statistical ensemble can be constructed to describe the end states of processes such as shaking grains. The role of energy is played by $\hat{\Sigma}$ which is an extensive

quantity (scales with system size) and the analog of temperature is a tensorial quantity, $\hat{\alpha}$ (In the context of infinitely, rigid grains, this intensive variable has been called Angoricity by Edwards [28]). This tensor is defined by:

$$\hat{\alpha}(\hat{\Sigma}) = \frac{\partial Z_0(\hat{\Sigma})}{\partial \hat{\Sigma}} \tag{5}$$

where

$$Z_0(\hat{\Sigma}) = \sum_{\nu} \omega_\nu \delta(\hat{\Sigma}_\nu - \hat{\Sigma})$$

and the sum is over all mechanically stable packings ν.

Tests of the Stress-Based Ensemble

The ability to define an ensemble does not necessarily mean that real granular packings conform to this ensemble. The assumption of factorizabilty could break down or the frequency of occurrence of a packing, ω_ν could be history-dependent making the temperature or the ensemble concept not very useful. It is, therefore, essential to check the predictions of this ensemble against experiments and simulations.

This has been done for simulations of frictionless grain packings in two dimensions [53] and it has been shown to work remarkably well. As a result of this comparison, an equation of state analogous to thermal equations of state, has been derived for frictionless grain packing approaching Point J [53], that is consistent with the field theory constructed for Point J [57].

Two applications of the stress-based ensemble to experiments probing the jamming transition will be discussed below.

Jamming Phase Diagram

About a decade ago, Liu and Nagel proposed that the jamming transitions in all systems, including the glass transition in supercooled liquids can be described by one unified framework through the jamming phase diagram and a special point in this phase diagram referred to as Point J [58]. The proposed phase diagram, shown in Fig. 1, delineates the boundaries of the jammed region in a phase space spanned by a temperature, density, and a *load* axis. The load axis captures the non-equilibrium aspects of these materials. For ideal spherical particles, the transition at Point J is a purely density-driven change, occurring at a critical packing fraction, ϕ_c between an amorphous solid state of dry granular packings and a state where the packings fall apart under any external stress [49]. Recent experiments in granular packings have verified the existence of this, transition [59], as discussed in the section on experiments, and a theoretical description of this specific point is also emerging [56,60,61].

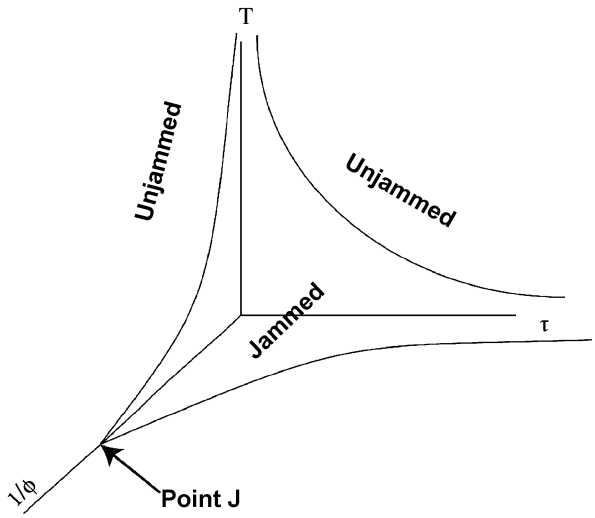

Jamming of Granular Matter, Figure 1
Jamming Phase Diagram proposed by Liu and Nagel. Adapted from [49]

In recent years, a theoretical framework has been developed [62,63,64] that describes the complete phenomenology of the glass transition in supercooled liquids. There are similarities between this framework and the properties of Point J, but whether or not these diverse phenomena can be unified under the umbrella of universality is a subject of intense current research. The stress-based ensemble approach described earlier has been used to construct a coarse-grained field theory of Point J and the jamming transition [56,57]. Since field theories and scaling ideas based on them underlies universality in thermal phase transitions, the extension of this approach to granular materials hold out the promise of identifying different universality classes in the context of jamming.

The special properties of Point J are related to isostaticity [49]. The discussion in Sect. "Statistical Framework of Jamming" analyzes isostaticity for frictionless, spherical grains. In general, isostatic packings are those special ones where the number of contacts is just enough to provide mechanical stability [41]. It has become increasingly evident [65] that isostaticity plays very different roles in spherical vs. non-spherical particles, which raises the question of whether Point J is special to frictionless spheres and disks. Also, it is not clear how generic the properties of this special point are, and whether or not it controls the behavior as one moves away from it in parameter space.

The jamming phase diagram of Liu and Nagel provides a framework for exploring the phenomenon of jamming in granular matter. Since temperature is not a relevant variable in the granular case, the phase space of interest would generically be the loading-density plane. The mechanical failure (or unjamming) of granular packings can occur because of (i) a vanishing bulk modulus (instability with respect to volume fluctuations) or (ii) a vanishing shear modulus (instability with respect to fluctuations of the shear stress). The state at Point J is an especially fragile one since both conditions (i) and (ii) are met. The stress state of the packing at this point is isotropic. In many instances, however, a granular material can be irreversibly deformed through the application of shear, and this process is also frequently referred to as unjamming. Specifically, for large enough shear stresses, a solid-like granular material will dilate, and fail in a process that typically leads to a locally weak and dilated region known as a shear band. This type of process has interesting analogues in the plastic failure of disordered glassy materials, in foams, in colloids and perhaps elsewhere. The tendency of densely packed granular materials to dilate under shear was discovered by Reynolds [66], and the adjustment of the density under shear is incorporated into critical state models of soil mechanics [67]. Although under shear, granular materials yield, and in some sense are 'unjammed', this process typically occurs under non-zero pressure as well as non-zero shear stress. Hence, 'unjamming' due to shear differs from the isotropic-stress state that occurs at Point J. In fact, the vast majority of experiments associated with jamming/unjamming are of this latter shear-induced process. There is compelling reason, discussed below, to think that plastic failure by shearing has a qualitatively different behavior than what occurs at Point J. For instance, jammed states also depend on whether they are isotropic or not, as shown by Majmudar and Behringer [68].

Given the dilatancy effects, and the differences between isotropically and anisotropically stressed states, we organize our discussion of jamming around a phase space spanned by the components of the Cauchy stress tensor (Eq. (3)). This is a symmetric tensor, and in two dimensions we can, therefore, work with the isotropic pressure, P, and the shear stress τ. $P = \sigma_1 + \sigma_2$, and $\tau = \sigma_1 - \sigma_2$, where the σ_i are the principal stresses, i. e. eigenvalues of the stress tensor. Since the principal stresses are positive for granular materials, the physical space is enclosed by the $|\tau| = P$ and $P = 0$ lines. In fact, shear failure typically occurs for $|\tau| < P$. Point J lies at $P = 0\,, \tau = 0$. Along the $\tau = 0$ axis, simulations provide strong evidence of Point J having critical properties with, in particular, at least one diverging length scale [49].

If $\tau \neq 0$ then the jamming behavior could be quite different from that characteristic of Point J. Here, a useful analogy to equilibrium phase transitions may be the difference between a critical point, where a line of first-order phase transitions terminates, such as the ferromagnetic Curie Point which occurs at zero magnetic field and at a characteristic temperature T_c, or a tricritical point, which separates a region of second-order transitions from a region of first-order transitions. This type of a critical point arises in multicomponent, thermodynamic systems and a well-known example is superfluidity in a mixture of He^3 and He^4. For a high enough concentration of He^3, the mixture undergoes a first order transition into two phases, and only one, the He^4-rich phase is superfluid [69]. A characteristic of a tricritical point is the existence of two diverging length scales, and in particular, there is a higher dimensional phase space that describes the various transitions. We present this discussion here purely as an illustrative example of a phase-transition scenario in which the signatures depend on what parameters are varied. We use the phase diagram of Fig. 2 to provide a framework for differentiating between the avenues available for approaching the jamming transition. In much of the literature, no distinction has been made, and we believe that has led to confusion. For each possible path, the transition can also be approached from the jammed side or the unjammed side. If $\tau = 0$, then the unjammed state is fluid-like in the sense that the system cannot respond elastically to any applied stress.

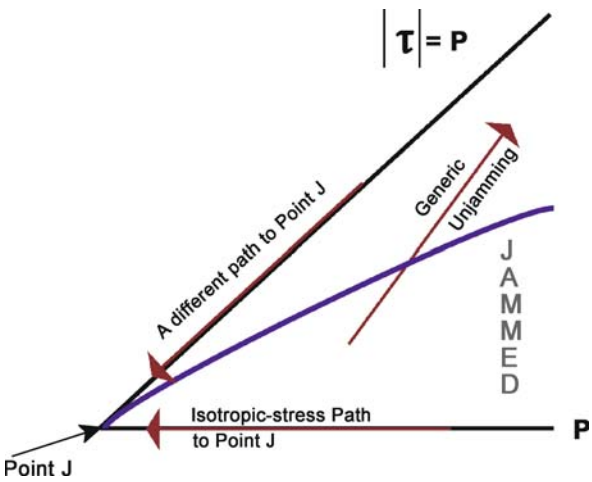

Jamming of Granular Matter, Figure 2
Jamming Phase Diagram in the space of components of the Cauchy stress tensor. The diagram shows a schematic in the $P - |\tau|$ plane, with *arrows* marking possible avenues of unjamming

Below, we first provide an overview of the signatures of jamming observed as the transition is approached from the fluid-like side (Sect. "Force Fluctuations, Dynamical Heterogeneities, Length and Time Scales in Granular Flows Approaching Jamming"), we then discuss the properties of jammed states under both shear and isotropic compression, as the jamming transition is approached from the jammed side (Sect. "Force Distributions, Length and Time Scales in Jammed States near the Jamming Transition"). In Sect. "Isostaticity and Point J", we summarize the experimental and theoretical description of Point J and the associated length and time scales. Related areas of research that we do not include in this article at all, are the phenomenon of aging, non-equilibrium dynamics, and generalized elasticity for granular matter. We refer the interested reader to articles listed in the bibliography.

Force Fluctuations, Dynamical Heterogeneities, Length and Time Scales in Granular Flows Approaching Jamming

In this section, we review a number of physical and numerical experiments which give insight into the nature of the jamming process. Each approach has advantages and disadvantages. Numerics, typically consisting of direct simulations of the equations of motions for a set of particles, has the advantage that it is possible to easily explore parameter space. This approach is referred to as Molecular Dynamics (MD) or as Discrete Element Simulations (DEM), depending on the community. The disadvantage of simulation is that it is difficult to achieve an accurate representation of friction (often friction is neglected), and in addition, simulations are typically done only for spheres in 3D or disks in 2D (which we will refer to as isotropic particles). Experiments do not suffer from issues of accuracy of the interaction law, but it is very difficult to obtain quantitative data except at the boundary of 3D systems. Alternatively, it is possible to obtain information using quasi-2D systems, such as collections of disks, something that is also done frequently for simulations. Two-dimensional systems for both experiments and simulations can yield excellent quantitative information, but may leave open the question of generalizability to higher dimension. With these caveats, we turn to a review of both numerical and physical experiments that probe the jamming transition from the fluid-like side. Granular flows can be broadly divided into two categories, (a) inertial flows dominated by collisions between grains, and (b) quasistatic flows that involve extended contacts and slow evolution from one static, mechanically stable state to another.

Recent results from numerical simulations by Lois et al. [70,71] are of particular interest here. These authors studied shear flow of rigid particles which they modeled using contact dynamics, a version of MD/DEM for rigid particles. A particularly interesting discovery coming from these studies was the existence of a divergent correlation length for forces, associated with the formation of clusters, as the packing fraction approached ϕ_c from below. The presence of a divergent length scale reinforces the idea that jamming is actually a critical transition.

Evidence for critical properties also come from experiments on hopper flow by Longhi et al. [19]. These experiments involved a quasi-2D flow of monodisperse spheres out of a hopper-shaped container. Necessarily, the presence of flow implies that the states studied here were unjammed. As the particles moved by, a small high speed force gauge mounted near the outlet of the hopper yielded the impulse of individual collisions. The collisions could be clearly identified because of a large separation of scales between the duration of contact, and time between contacts. The distributions of these impulses, at large impulses were roughly exponentially distributed, and independent of the flow rate. At small impulses, the distributions evolve with flow rate with a relative increase of small impulse events as the flow rate decreases. A more interesting statistic is the distribution of τ's, the time between successive collisions. As the jamming transition was approached, these distributions approached a powerlaw, $P(\tau) \propto \tau^{-3/2}$.

An MD simulation of hard disks, in a geometry designed to mimic the experiments of Longhi et al. [19], provided an explanation for the evolution of the impulse distribution in terms of increasing velocity correlations [72,73]. These correlations also led to a qualitative explanation of the changes in $P(\tau)$ [72], which occur because the average time between collisions *decreases* as the flow velocity decreases while at the same time the time taken for a particle to fall through its own diameter increases. This leads to a large separation of time scales and a range of times over which the dynamics is scale invariant. The simulations provided strong indications that these physical effects owe their origin to the existence of chains of frequently-colliding particles. In addition, it was shown that these collision chains could be interpreted as stress chains with the stress measured through momentum transfer [20, the counterpart of force chains in this system of flowing hard particles. Fig. 3 is a snapshot of a simulation illustrating the stress chains for two different flow rates, as controlled by the width of the opening at the bottom of the hopper. More recent work on this system of hard disks has provided evidence for the existence of

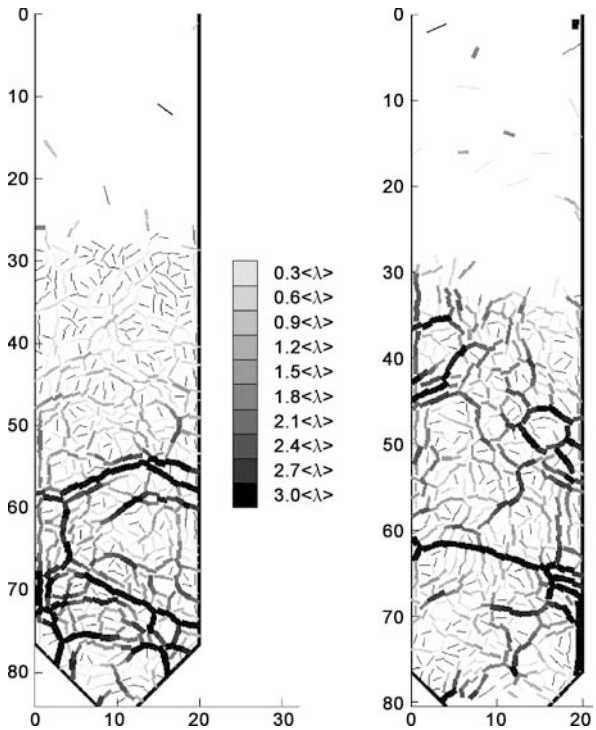

Jamming of Granular Matter, Figure 3

Stress chains. The grayscale plot is constructed by scaling the collisional stress by the average collisional stress ($\langle\lambda\rangle$) obtained by coarse graining over many collisions. Close to jamming, the coarse-graining time is much shorter than the time taken by a particle to fall through its own diameter [72]

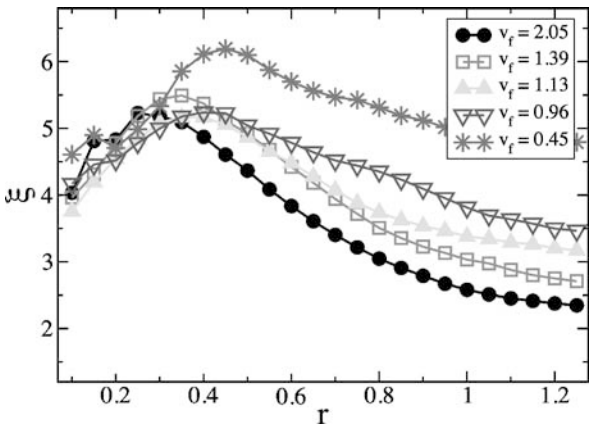

Jamming of Granular Matter, Figure 4

Lengths scale, ξ, associated with the heterogeneity of dynamics as a function of the mean square displacement, r, of particles in MD simulations of a two-dimensional hopper flow [20]. This length scale, extracted by analyzing the spatial correlation between displacements of particles [74,75], peaks at a characteristic displacement which can be related to a characteristic time at which the dynamics is most heterogeneous. This plot is qualitatively similar to ξ vs. time plots observed in the experiments of Durian et al. on air-driven beads, in the experiments of Dauchot et al. on particles driven by oscillating shear (see text), and in supercooled liquids

dynamical heterogeneities identified through the displacements of grains [72].

A length scale can be extracted from measurements of spatio-temporal correlations of these heterogeneities. As in supercooled liquids, there is a length scale that has a peak at a characteristic time. Both length and time scales depend on the average flow velocity, growing as the system approaches jamming. The increase in length, however, is not as dramatic as in supercooled liquids or in experiments on air-driven granular beads [9]. As an example of the characteristic shapes of the length versus time curves observed in all of these systems, supercooled liquids, inertial and quasistatic granular flows, we show a set of data from the MD simulations of hopper flow in Fig. 4 [20].

Pouliquen et al. have carried out a series of experiments on flow down chutes/inclined planes [11,76]. These studies have shown that there is a range of inclination angles, θ, for the chute for which it is possible to obtain steady state flow, and that the depth of the flowing layer, h, depends on θ. The flow stops if h is smaller then an inclination-angle-dependent function, $h_{\text{stop}}(\theta)$. Data for the depth-averaged velocity, u can then be expressed in terms of a flow rule, $u/(gh)^{1/2} = \beta h/h_{\text{stop}}(\theta)$, where β is a material-dependent $O(1)$ constant. Following a suggestion by Ertas and Halsey [77], Pouliquen [11] explored the possibility that there might be observable spatial correlations for the surface velocity which would signal the onset of jamming. Indeed, such correlations exist, and they grow as the inclination angle of the chute is lowered, i. e. as the system approaches the jammed state.

Kolb et al. [78] have studied the response of a jammed 2D layer to a localized cyclic displacement. This process was carried out quasi-statically, and in such a way as to locally unjam the system. More specifically, the experiment consisted of a bidisperse layer of disks, inclined from the vertical, so that the system was normally jammed. An intruder particle having a diameter corresponding to the larger of the two grain sizes, was displaced periodically in time by small amounts. These local displacements led to a much longer range set of displacements, part of which was reversible, but part of which was irreversible. This long-range response is suggestive of a critical-like response for states near jamming.

Drocco et al. [79] have used MD to to simulate the behavior of a 2D particle that is pushed through originally static bidisperse packings of similar particles. The pack-

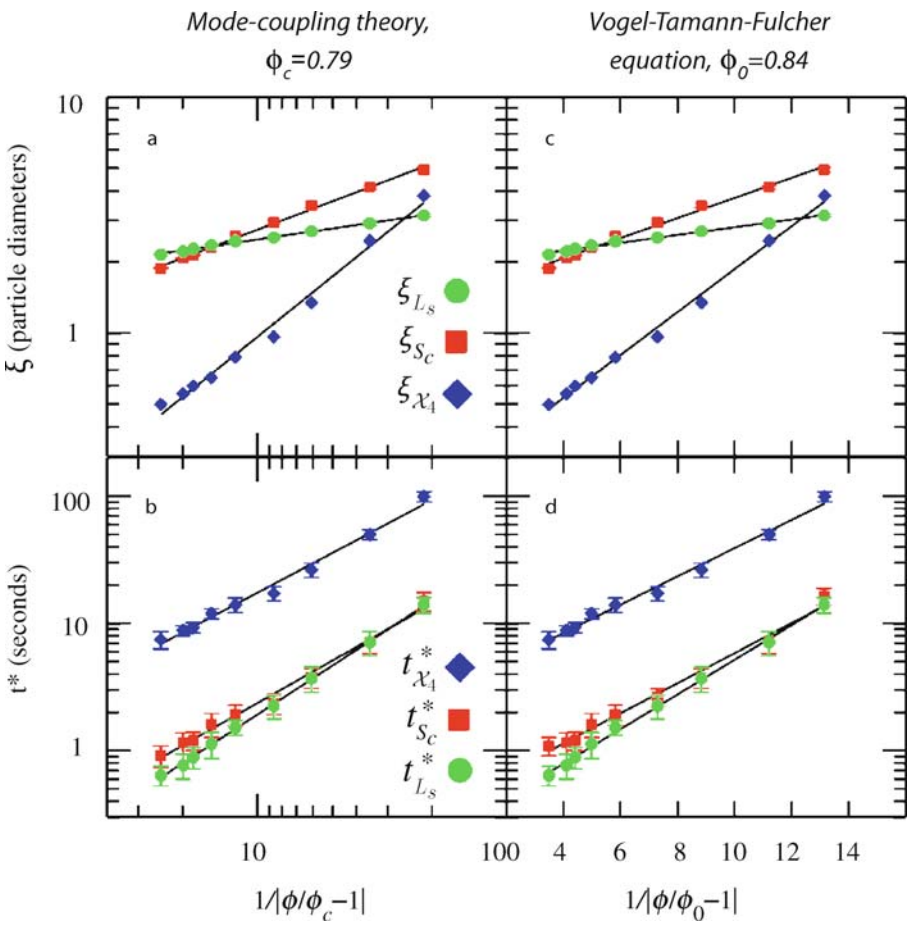

Jamming of Granular Matter, Figure 5

Data for length (ξ) and time (t^*) showing an increase in these scales associated with dynamical heterogeneities in a system of air-driven beads. Reprinted from [9]

ings reside in a square geometry with periodic boundary conditions, and are prepared at packing fractions below the jamming value for this system. The authors first focus on the intruder velocity, which becomes increasingly intermittent as $\phi \to \phi_c$ from below. To characterize the intermittency, Drocco et al. use multifractal analysis to compute moments, $M(q) = \int \mathrm{d}v P(v) v^{-q} \propto (\phi_c - \phi)^{\tau(q)}$, and present evidence for multiscaling. Of particular relevance here is a demonstration that as $\phi \to \phi_c$, the number of disks, n, that move in response to a fixed pushing distance for the intruder grows rapidly. These authors use data for the number of moving disks vs. ϕ to estimate a correlation length, which diverges as $\xi \propto (\phi_c \phi)^{-\nu}$, with $\nu = 0.71 \pm 0.12$. These numerical studies have an interesting parallel with experimental studies by Geng and Behringer [80] which yielded the force needed to push a disk through a channel of other disks in the jammed state. These data yielded a pushing force that also became increasingly intermittent on approach to ϕ_c, in this case from above, and a mean force that vanished as $(\phi - \phi_c)^\alpha$ with $\alpha = 1.5$.

Dauchot et al. [10] have studied correlations in systems of particles in 2D that are subject to small-amplitude oscillating shear. In this case, simple shear was applied, meaning that a rectangular sample was alternately tilted into a parallelogram to the right, then left. Dauchot et al. were particularly interested in characterizing the spatio-temporal fluctuations of systems near jamming. An image was obtained each time that the system was returned to the rectangular state, and the location of each particle was then determined. From these data, the authors first calculated a measure of individual particle motion, the self-intermediate scattering function, $F_s = \langle \hat{F}_s(k,t) \rangle = N^{-1} \Sigma_j \langle \exp[-ik(r_j(t) - r_j(0))] \rangle$. The results for $F_s(k,t) \propto \exp[-(t/\tau(k))^{\beta(k)}$ indicate Brownian diffusion at small k, but a different behavior at large k,

i. e. smaller spatial scales, where β is smaller than unity: a stretched exponential. They attribute this behavior to dynamical heterogeneities, and they further quantify the nature of these dynamical heterogeneities by analyzing length and time scales associated with fluctuations of $\hat{F}_s(k, t)$, as has been done for supercooled liquids. The analysis identifies a moderate length scale associated with the fluctuations, in agreement with the type of length scales that have been observed in supercooled liquids.

Durian et al. [9] have studied dynamical heterogeneities in a system of air-driven granular beads as a function of the packing density. Using multiple measures of lengths associated with dynamical heterogeneities, these authors demonstrate that there are length and time scales, which grow with the approach to Point J. They also show that the increase in these scales follows the Vogel–Fulcher–Tamann form: a form that is associated with supercooled liquids approaching the glass transition [81]. Figure 5 shows length and time scales extracted from different measures of dynamical heterogeneities as a function of the packing fraction. The fits are to the predictions of Mode-Coupling-Theory of the glass transitions and to the Vogel–Fulcher–Tamann form, which implies an exponential divergence at a packing fraction ϕ_0. Interestingly, the value of ϕ_0 is very close to the ϕ_c associated with Point J [9].

Force Distributions, Length and Time Scales in Jammed States near the Jamming Transition

An interesting premise is that the distribution of interparticle contact forces, $P(F)$, can provide insight into the nature of the global granular state of a system. In particular, it may be useful as a tool to distinguish jammed from unjammed or nearly unjammed states. It is also an important indicator of the overall stress state of the system, as demonstrated in experiments [68,82] and theory [83,84]. Since the transition to a deforming state in a dense granular material has often been obtained by shearing, there have been a number of studies that have addressed this case [12,13,14,16,17,18,85,86]. In particular, Howell et al. [13,14] have directly probed the transition that occurs when the density of a sheared granular sample is reduced to the point where stresses vanish. Other studies have probed the nature of steady granular shear [12,16, 18,85,86].

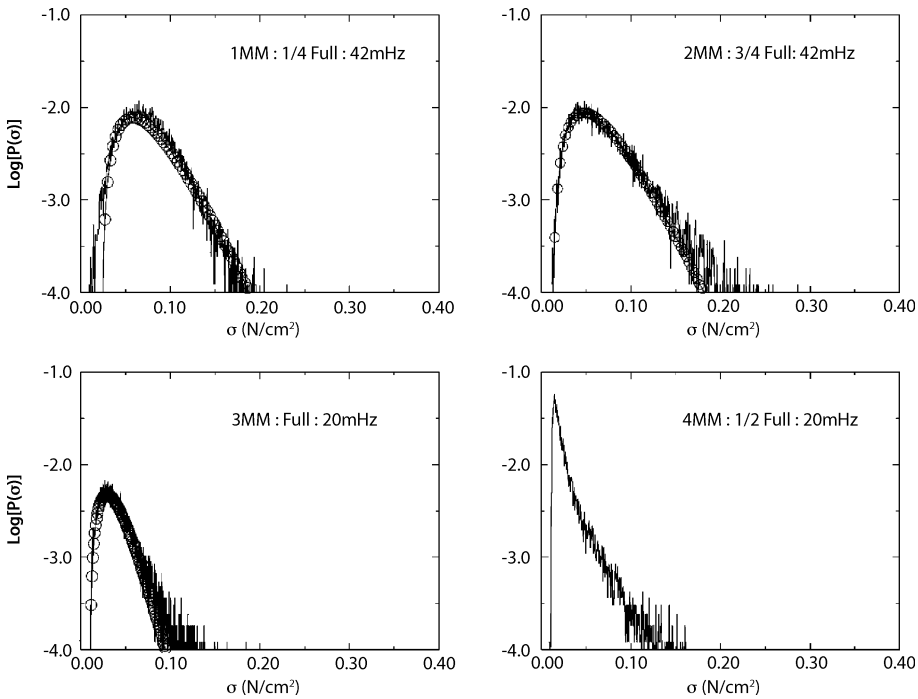

Jamming of Granular Matter, Figure 6

Distributions of forces measured at the base of a continuously sheared layer of glass spheres, for different sphere diameters of 1 mm, 2 mm, 3 mm and 4 mm, as indicated, and a fixed detector diameter of 1.0 cm. The other numbers give the depth of the granular layer as a fraction of the full height, 4.1 cm, and the shearing rate [12]

We discuss below several recent experimental studies which help elucidate this point. Here, we focus first on sheared, and then on isotropicically confined systems.

Sheared States

Fluctuations, their distributions and their correlations, are interesting for a number of reasons. First, they can reveal the nature of short and long range behavior. In addition, they provide a useful test of model behavior, and they help to identify key parameters. One of the earliest experimental characterization of force fluctuations in dense granular materials was by Miller et al. [12]. These experiments involved the shearing of an annular layer of glass spheres. A rough plate sheared the top of the annular layer and the pressure was measured at the bottom of the layer using a force gauge of fixed area. These studies are typical of many quasi-static shearing experiments. The system remains close to a static state, so that if the shear stress is reduced, the system typically reverts to a static state with only small changes in the particle configurations. However, enough shear is applied during steady-state motion that the system unjams intermittently, presumably associated with a short-term weakening of the shear modulus. In some sense, this type of motion straddles the jammed-unjammed boundary, first on one side, and then on the other. In addition, it is typical that a region of the material weakens and forms what is called a shear band. Because a shear band has lower shear strength than other regions, it tends to persist as a localized region where most of the shear deformation occurs. This process is typified by local structural rearrangements and frequently with large force fluctuations. The fluctuations in the experiments of Miller et al. showed a kind of rate-independence. They also yielded information on the force and pressure distributions, Fig. 6. The force detector was of fixed diameter, so by varying the particle diameter, it was possible to tune the number of particles contacting the detector from a small number, to ~ 100 particles. For the largest particles, this provided, at least roughly, information on $P(F)$. For the smallest particles, one might expect that the force fluctuations would be substantially averaged out, yielding something more like the mean pressure. We show data for the PDF's of the measured forces/stresses, in Fig. 6. If the forces acting on the detector from individual particles were uncorrelated, then one would expect that this would narrow as $N^{1/2}$, i. e. as the square root of the number of contacting particles, and evolve towards a gaussian. In fact, there is much less narrowing than would occur for uncorrelated forces from individual particles.

Two recent studies of 3D continuously sheared systems provide additional insight [16,17,18,86]. Both systems yield PDF's of forces at the boundaries, and it is interesting to contrast the two sets of results.

Corwin et al. [87] have recently presented extensive experiments using a photoelastic technique to measure forces at the bottom boundary of a cylindrical container filled with spherical particles and sheared from above. In the steady state, particles in the bottom layer located beyond a characteristic radius, R_b, exhibited dynamic shear flow, specifically, a shear band. Particles located closer to the center than this radius remained in a permanently jammed state and moved as a solid body. Corwin et al. obtained P(F)'s, with a particular focus on the region above R_b. These distributions, Fig. 7, show an enhancement for small F, and qualitatively resemble those obtained by Howell et al. for 2D shear flow, as discussed

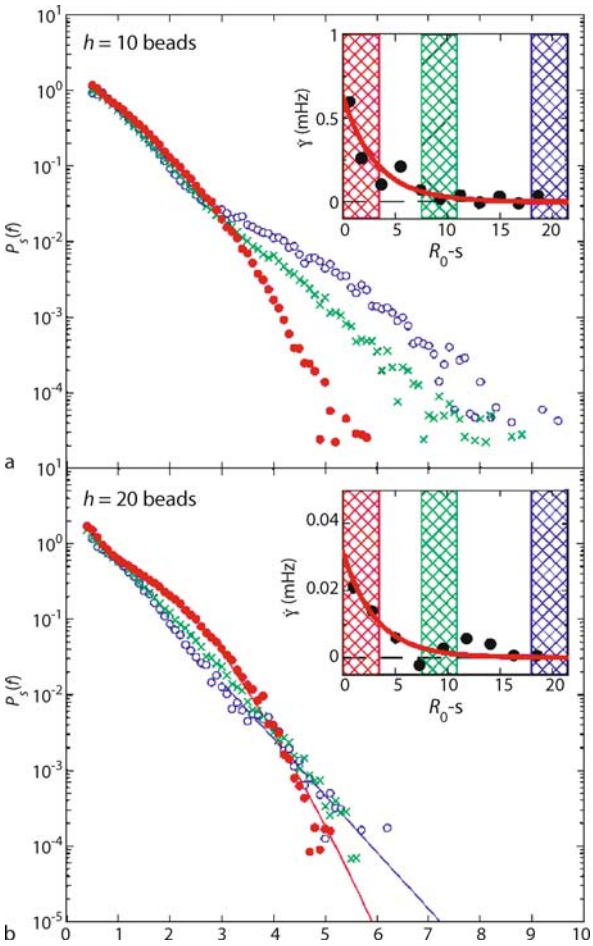

Jamming of Granular Matter, Figure 7
P(F) in sheared and non-sheared zones. Reprinted from [87]

below. Corwin et al. proposed that these distributions can be related to an effective temperature, T_{eff}, by noting that the distribution of forces and the radial correlation function, assuming forces described by a potential, $V(r)$, should satisfy $P(F)dF = G(r)dr$, where $G(r)dr$ is the radial distribution function. They then drew on the small-r thermal result $G(r) \propto r^{-2}\exp[-V(r)\beta]$ where $\beta = 1/(k_B T)$ (in this case, T should be replaced by T_{eff}). This argument leads to the prediction that at low forces, one should expect $P(F) \propto F^{-1/3}$, a prediction that describes their data well. The resulting effective temperatures obtained from this analysis are insensitive to shear rate and also to the height of the granular layer.

Other recent studies of sheared 3D particles, this time in an annular channel, were carried out by Daniels and Behringer [16,17,18]. In these experiments, the overall geometry was similar to Miller et al. [12] but in addition, the bottom boundary of the experiment was shaken with peak accelerations that typically exceeded the acceleration of gravity. The shaking motion allowed the system to find not only a jammed state, but actually a crystalline state. At high enough shear, the crystal melted, producing a shear-banded, disordered, and flowing state. This transition was clearly evident in $P(F)$. Here, F is the force exerted by particles on a transducer that is roughly three particle diameters across. In the crystalline state, $P(F)$, Fig. 8, had an overall envelope that was roughly gaussian, but that had two peaks which were induced by the vibrational motion. In the melted, disordered state, $P(F)$ resembled the distributions of Miller et al. As the transition to the ordered state was approached from the disordered state, the tail of $P(F)$ became increasingly extended. Associated with this long tail were increasingly large volume fluctuations. In fact, the variance of volume fluctuations and the various moments, including the kurtosis of the P(F), showed a cusp at the transition. This is particularly interesting in the present discussion, since the volume variance should be proportional to the derivative of the Edwards entropy (discussed in Sect. "Statistical Framework of Jamming") with respect to compactivity, i. e. the Edwards ensemble analogue of a specific heat.

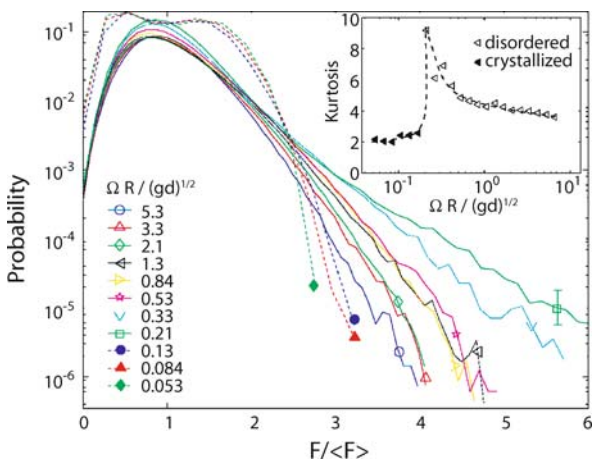

Jamming of Granular Matter, Figure 8
Distributions for the pressure/force at the bottom of an annular layer of spheres that is sheared at a rate Ω, from above, and that is vibrated from below with a peak acceleration of $\Gamma = 2.0$, in units of g. R is the mean radius of the annulus, and d is the particle diameter. The two-peaked distributions are in the crystalline phase, and the single-peaked distributions are in the disordered phase. The *inset* shows the Kurtosis of the distributions as a function of a scaled shear rate. The cusp in the kurtosis is caused by the stretching out of the distributions as Ω is decreased from a large value where the state is disordered

A relatively early demonstration of jamming properties was the work of Howell et al. [13,14] These experments consisted of photoelastic disks which were sheared in an annular Couette geometry, as in the sketch of Fig. 9. The boundaries of the apparatus consisted of a fixed outer ring, and in inner wheel whose rotations provided the shear. The particles were confined in the horizontal directions by the ring and wheel, and they rested on a smooth powder-lubricated horizontal Plexiglas sheet. Data were obtained in a steady state after the system had been sheared for roughly 30 minutes. The low shearing rate meant that that the process was quasi-static and hence always close to equilibrium. In particular, if the shearing was stopped, the majority of the force chains remained unchanged following an initial relaxation. Over long times, more of the force chains decayed, but that process was logarithmically slow [85]. In these experiments, the average force per particle (roughly a measure of the particle-scale pressure) was obtained by calibrating an applied compression to the photoelastic response of the particles. These experiments, as well as those of Daniels and Behringer [18,88], showed an increase of the mean stress with the shear rate that varied linearly as the logarithm of the shear rate. The logarithmic strengthening has been shown to be consistent with predictions of the stress-based ensemble, described in Sect. "Statistical Framework of Jamming" [89].

Several observations concerning these experiments are noteworthy. First, these experiments yielded the force distributions for the mean force on a particle, and also for averages over multiple particles. Second, they showed a kind of critical slowing down as the packing fraction, ϕ, approached the jamming transition from above, and third, they showed that a characteristic length scale associated with the mean length of force chains grew as $\phi \to \phi_c$ from above. We reproduce these results in Figs. 10 and 11 [13].

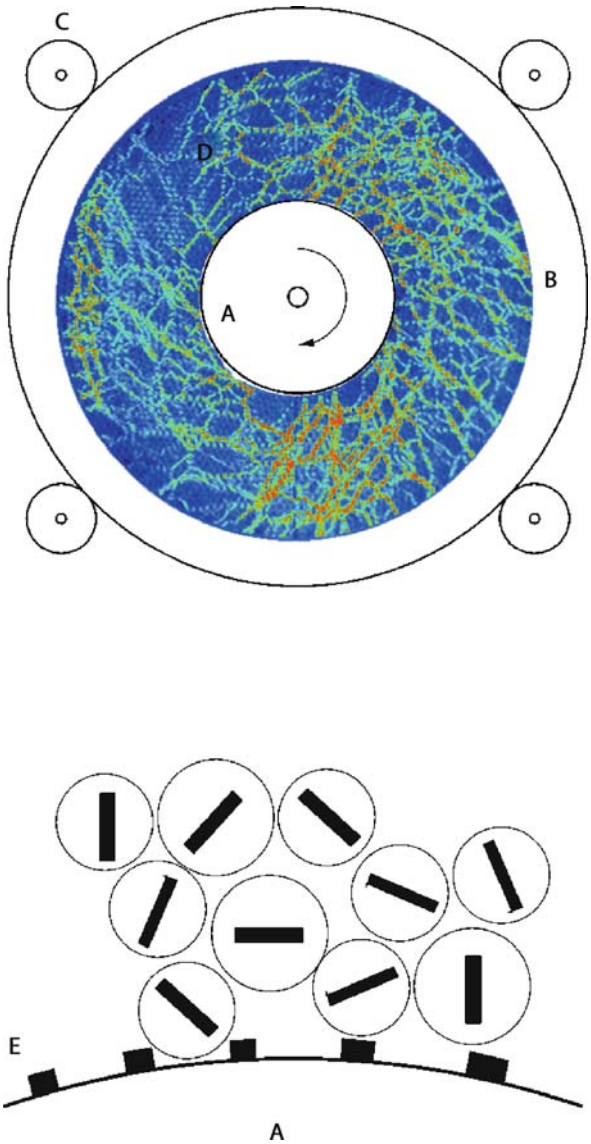

Jamming of Granular Matter, Figure 9

Sketch of the basic two-dimensional Couette shear apparatus, with an overlaid image showing the heterogeneous force structure that occurs. Photoelastic particles are confined within an annular region whose boundaries are a rotating inner wheel, and a fixed outer ring. Also shown is a blow-up sketch indicating the rough nature of the shearing wheel, and the fact that the photoelastic disks are marked with small bars for tracking purposes. In the false-color image from this experiment, *red* corresponds to particles experiencing a large force, and *blue* 0 to particles experiencing a small force [13,14]

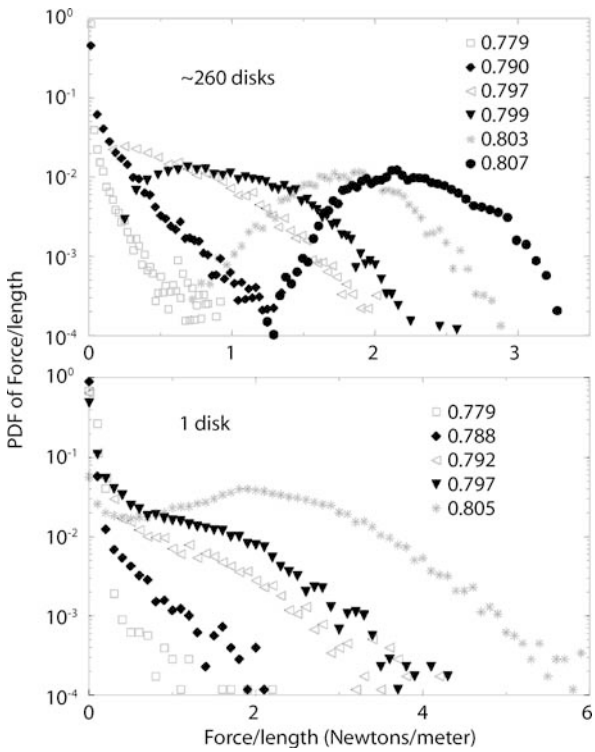

Jamming of Granular Matter, Figure 10

PDF's from the 2D Couette experiment for single particles (bottom) and for collections of particles lying within an extended region. Numbers indicate the packing fraction. This system has a critical value of $\phi = 0.776$ where stresses vanish. Note that the distributions of forces for single particles develops a maximum for large packing fractions. The distributions for forces for collections of roughly 260 particles in a contiguous region show a much stronger peak as ϕ increases, and no peak when ϕ is just above ϕ_c. This suggests that correlations in force grow as $\phi \to \phi_c$ [13]

The distributions for the single particle forces depend on the distance to ϕ_c. For ϕ close to ϕ_c, the distributions fall off roughly exponentially with large force, although there is an extra density of low force states. These distributions are nominally rather similar to single particle force distributions which were observed for particles at the boundaries of the 3D sheared system by Corwin et al. [86]. For larger ϕ, the distributions from the 2D shear experiments developed a peak which gradually migrated to larger force as the density grew. The distributions for the force on collections of particles also depend on the density. Close to jamming, they are similar to the single-particle distributions; at larger density, they evolve to a more gaussian shape. The multi-particle-force distributions are then interesting on several accounts. First, they provide a way to compare to earlier 3D pressure measurements (i. e. force measurements involving multiple particles) and they also give some sense of force correlations. That is, if the force on a given particle were uncorrelated with the force on other nearby particles, then the distribution for the forces

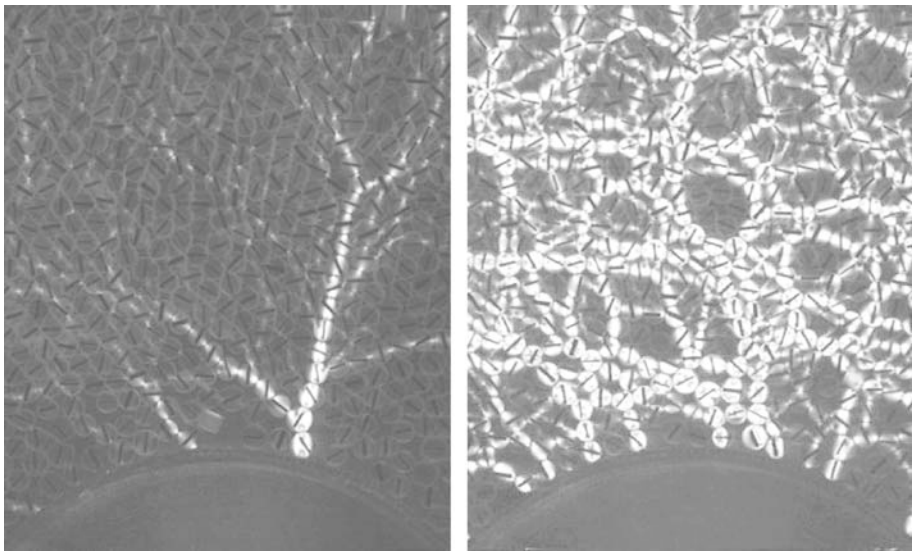

Jamming of Granular Matter, Figure 11
Photoelastic images from the Couette shear experiment at a low (*left*, $\phi - \phi_c = 0.001$) and high (*right*, $\phi - phi_c = 0.031$) distance from the unjamming point of this system. Note the long filimentary chains near ϕ_c and the tangled network far from ϕ_c [13,14]

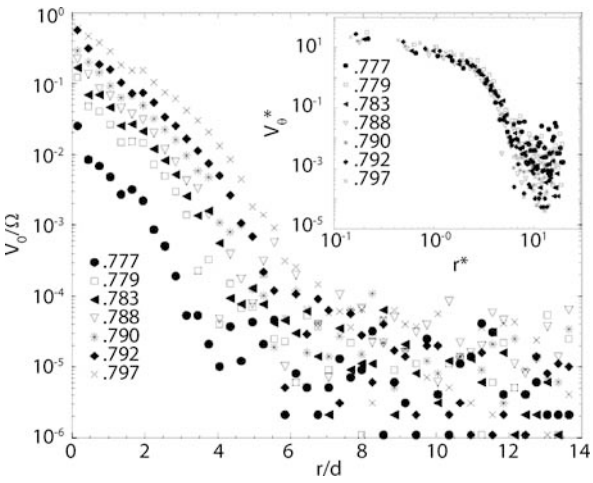

Jamming of Granular Matter, Figure 12
Data for the mean azimuthal velocity profiles vs. radial distance from the shearing wheel for the 2D Couette experiment. These data are normalized by the angular speed, Ω, of the shearing wheel. *Inset* shows that these data collapse onto a single curve when scaled a function $\propto (\phi - \phi_c)^1$ [13]

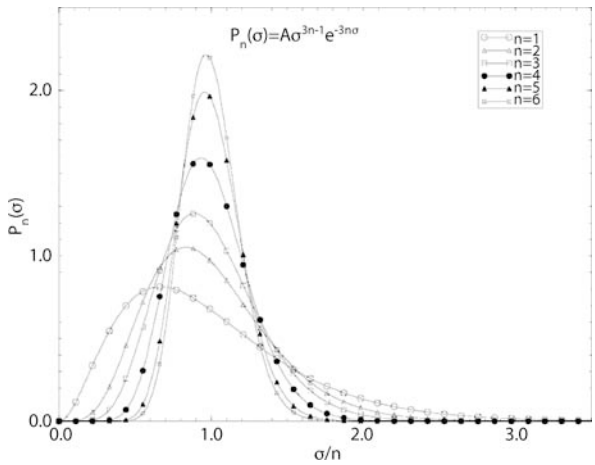

Jamming of Granular Matter, Figure 13
Computed PDF of the force per particle exerted by collections of n particles whose forces are uncorrleated and are drawn from exponential distributions. The absence of correlations leads to the expected narrowing by $n^{-1/2}$

experienced by a collection of particles would typically evolve to a guassian as the number of particles in the collection increases. This is consistent with what is observed for larger ϕ, but not with what is observed for ϕ near ϕ_c. Hence, by inference, there exist correlations for the particle-scale forces for near-critical ϕ's. To make this idea somewhat more precise, we show in Fig. 13 what happens to the PDF for the collective force for groups of particles acting on a detector if a) the forces exerted by each particle are chosen from q-model like distribution [103], and b) the forces exerted by any one particle are uncorrelated with any of the others. Specifically, the distribution narrows as $N^{-1/2}$.

The results from the 2D Couette shear for the velocity profiles vs. ϕ and also data for the characteristic force chain length both suggest that the system becomes increasingly inhomogeneous, in terms of force transmission as $\phi \to \phi_c$ from above. In particular, the typical force chain length clearly grows, although it does not necessarily diverge, in these experiments on approach to ϕ_c. In fact, given our discussion above concerning a $P - \tau$ phase diagram, we might well expect that in shear experiments one approaches point J along a path which lies above the P axis. Figure 11 contrasts a photoelastic image from Howell et al. [13] taken well above ϕ_c with an image for ϕ just slightly above ϕ_c. In the higher ϕ case, the force network is more nearly homogeneous, in contrast to the case just above ϕ_c. For the latter case, only a handful of particles are visibly in a force chain. (Here, some caution must be taken, since there is a force below which the photoelastic image is too weak to be resolved.) Nevertheless, only intermittently is there a strong enough contact at the shearing wheel so that a particle is entrained by the shearing wheel. This leads to critical slowing down which is evident in the mean velocity profiles, Fig. 12.

Isotropic vs. Anistropic Stress States

As discussed above, the nature of jamming depends, in what appears to be a significant way, on the presense or absence of shear stresses. The experiments and simulation discussed above have all involved sheared states. Here, we consider simulations and experiments which are for isotropic systems, and which focus on an approach to jamming that involves the dense/jammed side.

Numerical Simulations The phenomenon of jamming in isotropically stressed states has been extensively investigated through numerical simulations of frictionless, spherical grains in both two and three dimensions [49,90]. This work includes in particular, a series of MD simulations by O'Hern, Silbert, Liu and Nagel [49,91,92]. This series of simulations led to the introduction of the existence of a critical point at zero shear and a critical density, ϕ_c: Point J. The simulations probed the nature of the packings of spherical grains (disks in two dimensions), which interact via a short-range, soft-repulsive potential, meaning that there are no long-range interactions, and that there are purely repulsive normal forces which come into play when the grains are compressed. The packings were obtained by quenching from initial random states with a packing fraction $\phi = NV_p/V$, where N grains each with volume V_p were packed in a box of volume V. For polydispersed assemblies of grains, the definition of the packing fraction can be easily generalized. An interesting protocol that was developed in trying to identify Point J was that of inflation or deflation, in which the pressure of a quenched state was measured, and the radius of the grains were scaled up or down by small factors until the pressure was zero (within a numerical tolerance). For each initial configuration (ν) the zero pressure state corresponded to a particular packing fraction, ϕ_ν^c. At this packing fraction, the grains just touched and if the packing fraction was reduced below this, no mechanically stable states could be obtained. The crucial findings of these simulations are summarized below:

1. If properties such as pressure, shear modulus, and average number of contacts of the packings are measured as a function of $\phi_\nu - \phi_\nu^c$, then the properties are found to be independent of ν.
2. The pressure, bulk modulus and shear modulus all go to zero as some power of $\phi_\nu - \phi_\nu^c$. The exponent of the power law for the pressure and the bulk modulus follows from the force law but the shear modulus has a non-trivial exponent that indicates non-affine displacements [49,93].
3. The distribution of ϕ_ν^c becomes narrower with increasing system size and the mean approaches the Random Close Packing value [32].
4. Above ϕ_c, the average number of contact follows a power law: $\langle z \rangle - z_{\text{iso}} \propto (\phi - \phi_c)^{1/2}$.
5. The vibrational density of states exhibits increasing weight at low frequencies as Point J is approached, and the lowest frequency below which the packing behaves as a normal elastic material approaches zero [94].
6. Two different diverging lengths scales can be identified from analysis of the vibrational spectrum. These two length scales diverge with *different* exponents [95], and therefore indicate the existenc of multiple diverging correlation lengths at Point J.
7. The force distribution, $P(F)$ is sensitive to the distance from ϕ_c, and approaches an exponential as the packing fraction approaches ϕ_c

Experiments Regarding physical experiments for the isotropic case, we are aware of only one set of results, namely data by Majmudar and Behringer [68]. These results depend crucially on using photoelastic techniques in a much more precise way to obtain contact forces [68,82]. Hence, we discuss briefly how this approach works. The basic idea is the following. For given contact forces acting on a disk, the stress field within the disk is determined. The stress field in turn determines the photoelastic response. Specifically, for a ray that traverses a slab of

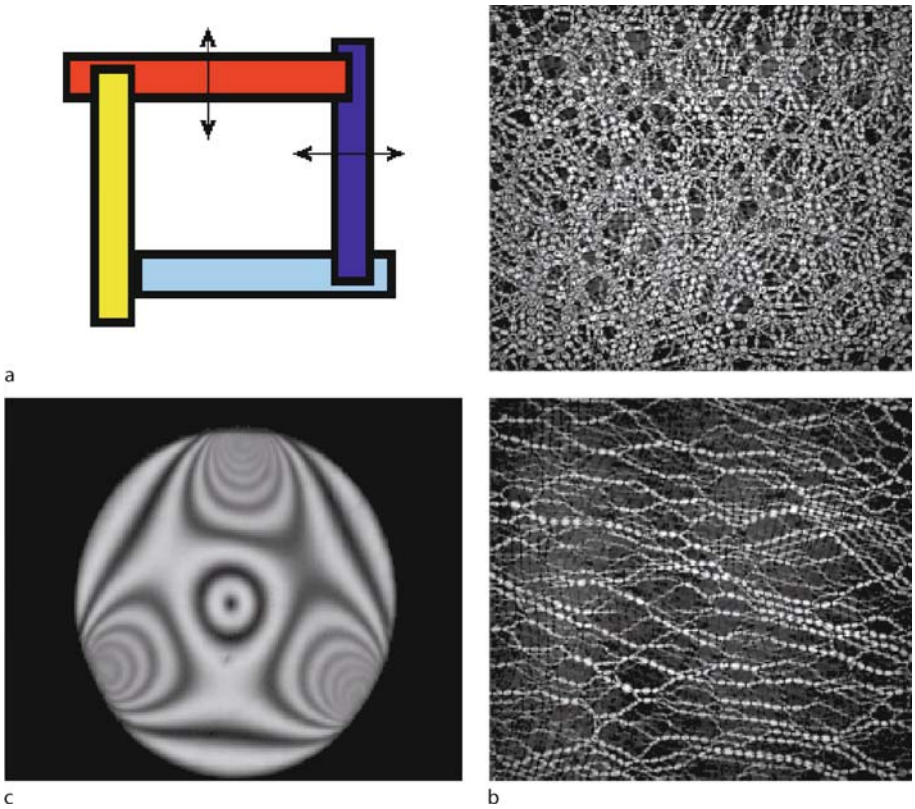

Jamming of Granular Matter, Figure 14
Biaxial experiment. a Sketch illustrating the fact that two walls of the experiment are moveable, allowing the creation of arbitrary (rectangular) deformations and stress states. b Two different stress states produced by top: isotropic compression, and bottom: pure shear, i. e. compression in one direction and equal dilation in the opposite. c A closeup image of a single photoelastic particle showing the detailed photoelastic pattern that is used to deduce contact forces [68]

photoelastic material of thickness L that is sandwiched between crossed circular polarizers, the transmitted intensity is $I = I_o \sin^2[(\sigma_2 - \sigma_1)CL/\lambda]$. Here, σ_2 and σ_1 are the principal stresses in the plane of the slab (or disk), C, the stress optic coefficient is a property of the material, and λ is the wavelength of the light with intensity I_o. Thus, contact forces determine stresses which in turn determine the photoelastic pattern. The idea is to carry out the inverse of this process. Specifically, from the photoelastic pattern, determine the applied forces. This can be done efficiently in the case of a disk because there is a closed form relation for the stress field generated by any number of contact forces.

MB have developed such an inverse process and applied it to systems of photoelastic disks that have been prepared in well defined stress/deformation states. The particularly important aspect of these experiments is that they have yielded the only experimental data, to our knowledge, for contact forces inside a granular sample, although a number of other experiments have yielded forces at boundaries. This point is important; theoretical studies by Snoeijer et al. [83,84] indicate that force distributions for contacts at boundaries and in the interior of a sample differ substantially.

Figure 14 shows a rough schematic of a biaxial tester (or biax) and resulting photoelastic response images for different types of applied stresses [68]. The purpose of the biax is to prepare states of well controlled strain or stress. The biax is constructed so that the space between opposing pairs of walls can be set to any convenient value. The particles in this sample are bidisperse disks, with about 80% having a smaller diameter (0.8cm) and 20% having a larger diameter (0.9 cm). Bidisperse packings of this sort typically have no long-range crystalline order, although they tend to show hexagonal bond order.

Manifestly, the force chain structures that result from pure shear vs. isotropic compression, Fig. 14b, are substantially different. Pure shear induces force chains that extend nearly straight and uninterrupted over the compressional

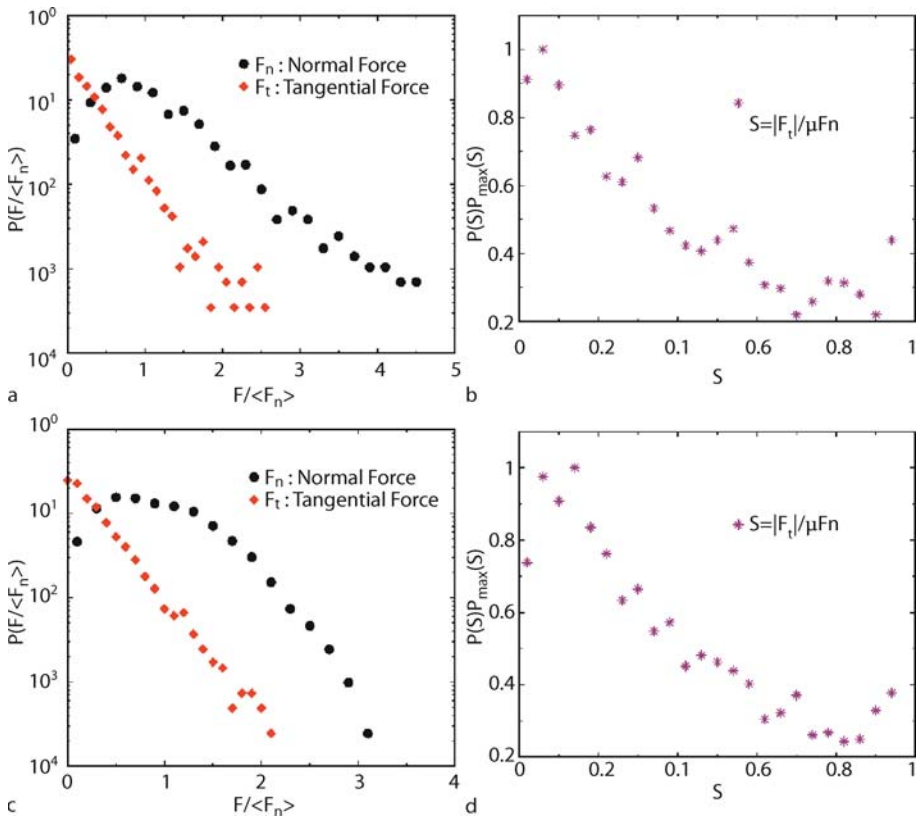

Jamming of Granular Matter, Figure 15

Distributions of forces for pure shear (*top*) and isotropic compression (*bottom*). Parts a and c give distributions of normal (F_n) and tangential (F_t, frictional) forces, all normalised by the appropriate mean normal force, $\langle F_n \rangle$. Parts b and d indicate the mobilization of friction, where the variable S indicates fully mobilized (at the point of slipping) friction when $S = 1$ [68]

direction of the biax. This is intuitively reasonable, since the pure shear state is achieved by compressing, hence strengthening contacts in one direction, while dilating the sample, thus weakening contacts, in the perpendicular direction. By contrast, for an isotropically compressed sample, the force network consists of a dense tangle of short force chain segments.

Various statistical measures reflect the fact that shear and isotropic compression yield qualitatively different states. We consider first, the distributions of contact forces, $P(F)$, which are given in Fig. 15. For convenience, forces are normalized by $\langle F_n \rangle$, the mean normal force. The figures show separate distributions for the normal and tangential contact forces. Although the distributions for the tangential forces are always exponentially distributed, regardless of the stress state, the same is not true for the normal force distribution. These clearly reflect the stress state, with a state of pure shear showing a roughly exponential fall-off at large F, and with a state of isotropic compression showing a much more rapid fall-off with F. In fact, these distributions are consistent with recent entropy-based arguments by Snoeijer et al. [83] and by Tighe et al. [96]. In both cases, the distributions for the normal forces show a peak at a force comparable to the mean. Perhaps a better measure of the difference between the shear and isotropic deformations is given by the force correlation function, which is significantly longer range along the force chain direction of the sheared case, than any direction for the isotropic case [68]. We show these results in Fig. 16.

Isostaticity and Point J

Isostaticity and Random Close Packing

This section is devoted to a geometrical picture which offers a perspective that is complementary to the statistical ensemble picture. The geometrical approach is focused on mathematically precise descriptions of packing of infinitely rigid particles [32,65]. In this statistical geometric picture, packings are characterized by the type of jamming: local, collective or strict. In the hierarchy of jammed states,

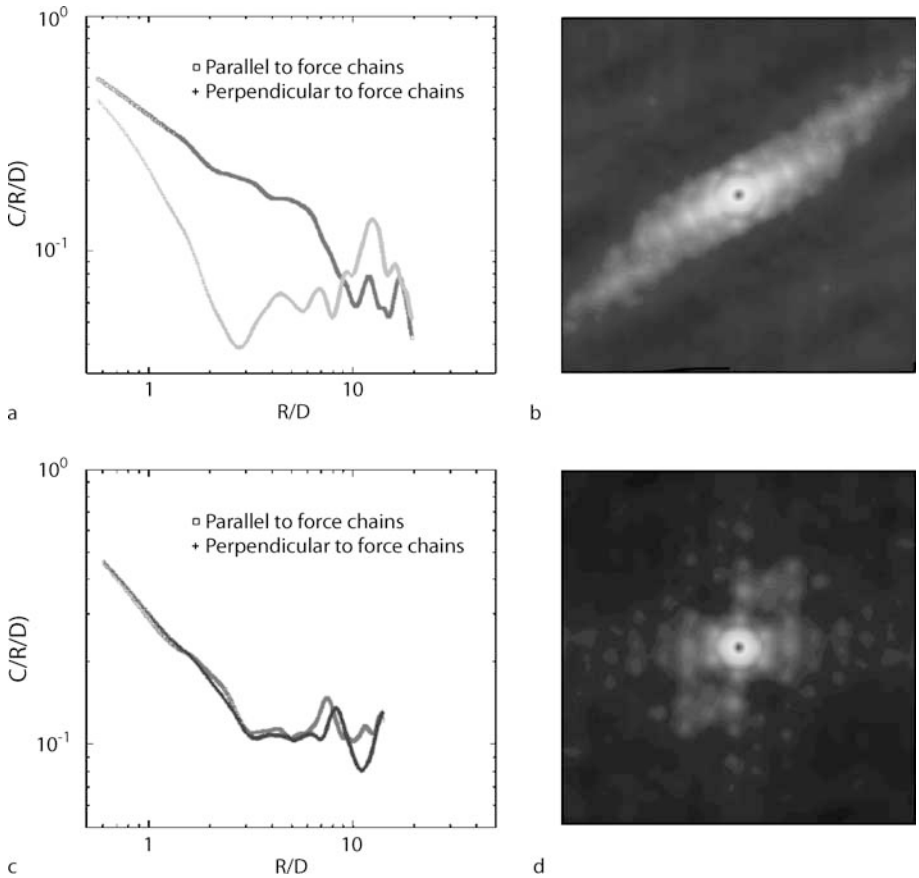

Jamming of Granular Matter, Figure 16

Force-force correlation functions obtained from 2D experiments. To compute these correlation functions, we compute the average force magnitude, $\langle F(r + r')F(r)\rangle$, acting on a particle. Since the system may be anisotropic, the average over r' must retain the angular information. We contrast data for a system that has been subjected to pure shear, top, with one that has been subject to isotropic compression, bottom. (See part b of Fig. 14). The right side of the figure shows a grayscale representation of the data. For the shear case, the correlation function has a roughly power-law decay (to the limits of the system size) for correlations along the strong force chain direction, and a rapid drop off for the transverse direction. For the isotropically compressed case, the correlations are identical in all directions, and decay rapidly [68]

locally jammed states are ones where each particle in the system is locally trapped by its neighbors. Collectively jammed states are ones where all finite subsets of particles are trapped by their neighbors, and strictly jammed states are collectively jammed configurations for which no global, volume-nonincreasing deformations are possible [32]. In addition, the degree of order in the packings is characterized by an order parameter or an order metric [97]. The emphasis in this approach is to understand the existence of packings, not their statistical weight in any dynamical protocol. The geometrical approach applies to infinitely rigid bodies which is an idealization of real particles, but the classification into jamming categories and a precise understanding of the nature of packings which can exist as the packing fraction is varied provides a framework onto which the statistical decsription can be superposed. There are a few special points on a phase diagram in the order-packing fraction space: (a) the maximally random jammed state (MRJ) is an extremal point corresponding to the least ordered, strictly jammed state, (b) the random-loose-packed (RLP) state is another extremal point corresponding to the lowest density at which a strictly jammed state can exist. This RLP state is more ordered than the MRJ state and is extremely fragile [97]. In contrast to the concept of random close packing (RCP) [32,97], the MRJ state is protocol-independent leading to a mathematical framework for studying randomeness in hard particle packings. The definition of MRJ does, however, depend on the choice of the order parameter, and further work needs to be done to explore appropriate order parameters.

Isostatic Packings are Marginal

It has become increasingly clear through theoretical analysis and simulations that isostatic packing of frictionless, isotropic particles are marginally stable. Simulations show that the vibrational spectrum of these packings have many soft modes and that the density of states of zero-frequency modes become non-zero as Point J (which is isostatic) is approached [49]. There is a diverging length scale associated with these vibrational modes, and this length scales as $1/(z - z_{\rm iso})$ [61,94]. The theoretical explanation for this divergence is rooted in the vanishing degrees of freedom as the isostatic point is approached from the jammed, hyperstatic side [60,61,93]. Stability analysis of the packing of deformable, isotropic grains, also shows that there is a marginal stability line along which, the pressure of the packing obeys $p/B = (z - z_{\rm iso})^2$, where p is the pressure and B is the bulk modulus of the packing [60,93]. Packings with fewer than the number of contacts specified by this relation are unstable. The question of whether isostacity and marginality always go hand in hand is an interesting question that is beginning to be explored [65,98].

Simulations and Theory

As discussed in Sect. "Isotropic vs. Anistropic Stress States", simulations of deformable, spherical grains [49,91] exhibit a special packing fraction, ϕ_c, below which it is impossible to construct a jammed packing, and interestingly, for large systems, ϕ_c approaches the RCP value. Recent theoretical work based on the isostatic nature of Point J [60,93], have provided an explanation for the diverging correlation length and the scaling relation between pressure and the number of contacts. A theory based on a minimum number of contacts needed for *local* stability is the K-core percolation model of Schwarz, Chayes and Liu [99]. This theory focusses on the mixed first-second order nature of the transition at Point J in terms of the order parameter, $\langle z\rangle$, which approaches $z_{\rm iso}$ with an exponent close to 1/2 as $\phi \to \phi_c^+$, but which is predicted to be zero for ϕ infinitesimally below ϕ_c. The meanfield exponents of this model agree with the exponents associated with Point J in the numerical simulations.

A field theory of jammed grain packings in two dimensions, and close to Point J has been constructed using the generalized stress-based ensemble (Sect. "Statistical Framework of Jamming"), and a fluctuating field related to the Airy Stress function [55]. This field theory [56], identifies two order parameters associated with the transition at Point J; the Airy stress function and the deviation of the number of contacts from the isostatic value. The field theory predicts a transition with a diverging length scale. Recent work [57] has focussed on refining this field theory by comparing its predictions to simulations of frictionless disks [49].

The stress-based ensemble described in Sect. "Statistical Framework of Jamming" can be used to understand the exponential form of $P(F)$, if one uses the fact that the packings at ϕ_c are isostatic. The starting point of this calculation is the canonical partition function of the stress-based ensemble: $Z(\alpha) = \sum_\nu {\rm e}^{-\alpha\Gamma_\nu}$, where α is the counterpart of the inverse temperature and $\Gamma = \sum_{ij} |r_{ij}|F_{ij}$ [53]. At ϕ_c, the spheres or disks are just touching, and therefore, the separations $|r_{ij}|$ can be replaced by the diameter of the grains. The sum over grain configurations, ν, therefore, involves a sum over the F_{ij}'s and the angles of the contact vectors. The isostatic point is special in that there is a one-to-one correspondence between geometry and forces. This means that for a chosen set of $\{F_{ij}\}$, there is only one geometry characterized by a set of contact angles that can be mechanically stable. The partition function then becomes:

$$Z(\alpha) = \sum_{\{F_{ij}\}} {\rm e}^{-\sum_{ij}\alpha F_{ij}} = \left[\int_0^\infty {\rm d}F {\rm e}^{-\alpha F}\right]^{z_{\rm iso}N/2} \tag{6}$$

The integral is easily performed, and the result can be used to relate α to the average contact force, $\langle F\rangle$: $\alpha = (z_{\rm iso}/2)\langle F\rangle$. Since the forces on different contacts are completely independent of each other at this isostatic point, as seen from Eq. (6), it follows that $P(F) \propto {\rm e}^{-\alpha F} = {\rm e}^{-z_{\rm iso}F/(2\langle F\rangle)}$, a pure exponential. This is a consequence of isostaticity. Away from the isostatic point, the correspondence between geometry and a set of contact forces is no longer one-to-one and distributions will acquire a non-exponential character, in agreement with observations that $P(F)$ of crystalline packings are more Gaussian than disordered packings, as discussed in Sect. "Force Distributions, Length and Time Scales in Jammed States near the Jamming Transition".

Analysis of simulations of frictionless disks, using this generalized ensemble led to a definitive functional form for the distribution of $\Gamma_m = \sum_{i=1}^m \sum_j r_{ij}F_{ij}$, the pressure of m particles inside an assembly of N grains, integrated over the volume occupied by the grains (Γ is known as the internal virial). The distribution $P_m(\Gamma_m)$ depends only on $x = N\Gamma_m/\Gamma_N$, where Γ_N is the internal virial of the whole assembly of N grains, and the functional form is [53]:

$$P_m(x) \propto x^{\rm ma}{\rm e}^{-ax}\ , \tag{7}$$

where $a = 2 + (\langle z\rangle - z_{\rm iso})^2$ This form indicates that the distribution would approach a Gaussian with a width which narrows as $\sqrt{m}$, in x space. In the space of the

unscaled variable Γ_m, however, the distribution could acquire a non-trivial shape and scaling with m if $\Gamma_N \to 0$, or equivalently, $\alpha \to \infty$. Since this is what happens as $\phi \to \phi_c$, one can expect non-trivial scaling of the pressure distribution. Equation (7) also shows that the exponential tail is maximally stretched at $\langle z \rangle = z_{\rm iso}$. Evidence for this may appear in the distributions of Fig. 8, where just at the transition, the exponential tail has become extremely extended.

There are many existing models for the exponential distribution of forces. In one view [87,91], the exponential distribution is related to the form of the pair correlation function $G(r)$, a function that measures the number of particles separated by a distance r, at small r. Using generic properties of $G(r)$ and the limit where grains are barely compressed, leads to a $P(F) \propto e^{-F^\kappa}$, where the exponent κ depends on the force law. In this approach, the exponential distribution is not related to isostaticity. This point of view was developed through the discussion of the experimental results of Corwin et al. [87]. A different set of approaches have considered the origin of the unidirectional transmission of forces (force chains) [100,101,102] by proposing a new class of constitutive relations arising from constraints on the stress tensor due to geometry of the packings. This approach has a natural connection to the generalized ensemble [28,54] and isostaticity [28,41,55]. An earlier, heuristic model, the q-model led to unidirectional propagation of forces, and an exponential distribution of contact forces [103]. More recently, Snoeijer et al. [83] and Tighe et al. [96] have predicted force distributions using the force ensemble approach (Sect. "Statistical Framework of Jamming"). The connection between the stress-based ensemble and force chains, away from isostaticity is still being explored.

Experimental Observations

Direct experimental investigation of Point J, approached from the jammed side, and the behavior in its vicinity have been more limited than computational and theoretical studies. We are aware of only one experiment by Majmudar et al. [59] that has directly probed Point-J from the jammed side. This work [59] has provided detailed experimental data for the mean contact number, Z, and the pressure, P vs. the packing fraction, ϕ_c. Hence, it is possible now to make direct comparison between a physical experiment and theory/model results. In these experiments, photoelastic disks were confined in the biaxial tester discussed above. The photoelastic approach allows the experimental determination of contact forces between pairs of particles or between particles and a boundary.

Figure 17 shows results for $Z(\phi)$ and $P(\phi)$. The data for Z show a fairly rapid increase in Z with increasing ϕ, but not a truly sharp discontinuity. However, assuming a reasonable choice of ϕ_c, the data indicate an increase of Z above the critical value which varies as $Z - Z_c \propto (\phi - \phi_c)^\alpha$ with an exponent $\alpha = 0.55 \pm 0.05$. This result is in agreement with the simulation of frictionless grains [49]. The pressure also follows a power-law in $\phi - \phi_c$ with an exponent of 1.1 ± 0.05, which is in agreement with Silbert et al. [49]. Finally, it is possible to compare to the stress-ensemble predictions of Henkes and Chakraborty by eliminating α in their predictions for $P(\alpha)$ and $Z(\alpha)$ to obtain an expression for $P(Z)$. The data, Fig. 18, are also in reasonable agreement with this prediction.

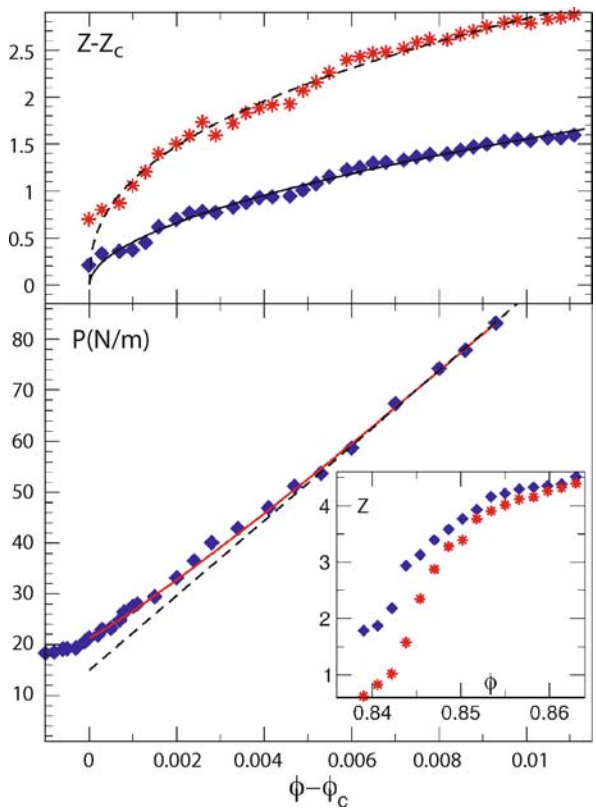

Jamming of Granular Matter, Figure 17
Data for the pressure and for the mean contact number per particle, Z vs. ϕ. The *inset* shows that Z vs. ϕ over a larger range than the *top* part of the figure. Data including rattlers as shown by asterisks, and data without rattlers by diamonds. Note that Z rises rapidly but not discontinuously at jamming. Power-law fits of P and $Z - Z_c$ vs. $\phi - \phi_c$ above jamming and for reasonable ϕ_c show expected exponents of 0.55 ± 0.05 for $Z - Z_c$ and 1.1 for P [59]

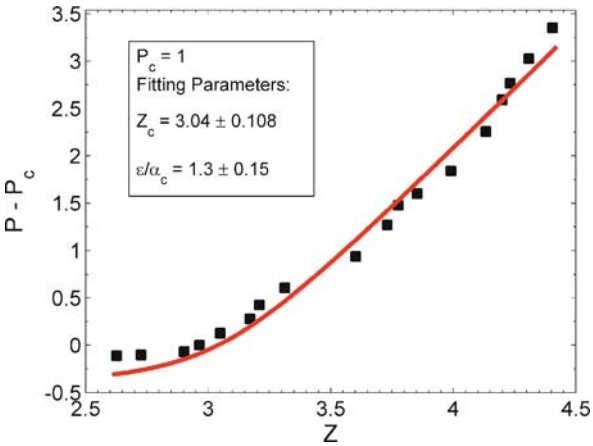

Jamming of Granular Matter, Figure 18
Data of Majmudar et al. for *P* vs. *Z* showing the generally good agreement with the predictions of Henkes and Chakraborty [59]

Future Directions

From the combined experimental, theoretical, and simulation studies of the last decade, a picture of the nature of jamming in granular matter is beginning to emerge, although many questions remain. In our discussion, we distinguish jamming/unjamming under isotropic and anisotropic stress conditions. Many recent simulations have considered the isotropic approach to Point J. A number of experiments have explored jamming/unjamming, although most of these involved the succession of failures that occur under shear and away from Point J, which corresponds to processes with a heuristically 'first-order' character. We believe that it is important to distinguish these various cases. To our knowledge, no one has considered the approach to Point J along an anisotropic stress path.

There are clear indications of heterogeneities in granular matter. These have both a spatial and a temporal aspect, and they become increasingly important with the approach to jamming. They have been identified on both the fluid-like and jammed side of the transition.

One aspect of these heterogeneities is concerned with forces and their transmission. A clear manifestation is in force chains, which can show strong correlations in the sheared state, but very short range correlations in the isotropic stress state. In dynamic processes, force chains continually rearrange, leading to strong force fluctuations. Although force chains are clearly visible in a variety of 2D experiments that are carried out above jamming, incipient stress chains have been observed in simulations of hopper flow on the fluid-like side [20], and there is evidence of arch formation and stress chains in experiments studying jamming in hopper flow [104,105,106]. An open question concerns the role that such structures may play if Point J is approached along a path of anisotropic stress.

A different type of heterogeneity in granular flows is related to the mobilities of grains [9,10]. This situation is similar to heterogeneities in supercooled liquids.

The analogies between jamming and thermal phase transitions prompts questions about the extent of the similarities between the two. For instance, is there an order parameter associated with the transition? Studies of the special type of jamming at Point J suggest that pressure and the deviation of the number of contacts from the isostatic value are the two, possibly coupled, candidates for the order parameter.

The generic jamming/unjamming transition that takes place in the presence of non-zero shear stress is characterized by intermittent dynamics and is reminiscent of trap-like dynamics, where the shearing causes the system to "hop" from one jammed state to another and jamming occurs when the average time scale for exploring all the traps diverges [89,107]. The framework of soft-glassy-rheology [108] provides an approach to calculating the stress response of granular packings approaching jamming. In this picture, the jammed state is an intrinsically non-equilibrium state, which exhibits aging, meaning a slow evolution of its properties with time. The issues raised here are much the same as those arising in the context of the glass transition, and involve understanding the relation between fluctuations and response [109], through the definition of an effective temperature. This is an area of active research [110,111,112,113,114].

Understanding the jamming of non-spherical grains is much less advanced than that of spherical grains. In this regard, the geometrical approach of Torquato and Stillinger [65] is being applied to non-spherical grains. Experiments have probed the issue of isostaticity in non-spherical grains [115,116,117]. The stress-based ensemble described in this article is a promising approach for analyzing these issues.

In the context of frictional and frictionless packings, a question that needs to be addressed is what aspects of the jamming transition are sensitive to the presence of friction. Simulations and theories of jamming have mainly focussed on frictionless grains. The statistical ensemble approaches, however, are general and provide an avenue for a unified framework.

These observations raise many questions:

1. How are stress chains related to the the mobility of grains? Since force chains tend to lock grains in place, it may well be that there is anti-correlation between force chains and mobility.

2. Is there a basic dynamical principle leading to the occurrence of stress chains, or are they more strongly influenced by geometry?
3. What if any, is the relationship between the length and time scales on the jammed side to those observed on the flowing side?
4. Are incipient chains on the fluid-like side responsible for jamming as the transition is approached? Is jamming caused by a divergence of the lifetime of these chains?
5. Is there an extension to jamming in the presence of shear? That is, how does the path in P-τ space affect jamming?
6. What is the nature of jamming for nonspherical particles, for systems with a broad range of particle sizes, or for particles that interact with cohesive forces?

The advances in experimental techniques which have led to the measurement of grain-level forces, has been a tremendous boon for theorists. Simulations have led to detailed descriptions of fluctuations which provide means for testing theoretical frameworks. The recent advance in the statistical ensemble approach to jamming, the geometrical characterizations of packings of rigid grains, combined with the detailed experimental and simulation studies, is expected to answer many of the remaining questions and lead to a fuller understanding of the phenomenon of jamming.

Bibliography

1. Jaeger H, Nagel S, Behringer R (1996) Granular solids, liquids, and gases. Rev Mod Phys 68:1259–1273
2. Jaeger H, Nagel S, Behringer R (1996) The physics of granular materials. Phys Today 49:32–38
3. Kadanoff LP (1999) Built upon sand: Theoretical ideas inspired by granular flows. Rev Mod Phys 71:435
4. Weeks ER, Crocker JC, Levitt AC, Schofield A, Weitz DA (2000) Three-dimensional direct imaging of structural relaxation near the colloidal glass transition. Science 287:627
5. Donati C et al (1998) Stringlike cooperative motion in a supercooled liquid. Phys Rev Lett 80:2338
6. Donati C, Glotzer S, Poole P, Kob W, Plimpton S (1999) Spatial correlations of mobility and immobility in a glass-forming lennard-jones liquid. Phys Rev E 60:3107
7. Berthier L et al (2005) Direct experimental evidence of a growing length scale accompanying the glass transition. Science 310:1797–1800
8. Lacevic N, Starr F, Schroder T, Glotzer S (2003) Growing correlation length on cooling below the onset of caging in a simulated glass-forming liquid. J Chem Phys 119:7372
9. Keys AS, Abate AR, Glotzer SC, Durian DJ (2007) Direct observation of growing dynamical length scales and prediction of the jamming transition from dynamical heterogeneity in a granular system. Nat Phys 3:260
10. Dauchot O, Marty G, Biroli G (2005) Dynamical heterogeneity close to the jamming transition in a sheared granular material Phys Rev Lett 95:265701
11. Pouliquen O (2004) Velocity correlations in dense granular flows. Phys Rev Lett 93:248001
12. Miller B, O'Hern C, Behringer RP (1996) Stress fluctuations for continously sheared granular materials. Phys Rev Lett 77: 3110
13. Howell D, Behringer RP (1999) Fluctuations in a 2d granular Couette experiment: A critical transition. Phys Rev Lett 82: 5241
14. Veje C, Howell D, Behringer RP (1999) Kinematics of a two-dimensional granular couette experiment at the transition to shearing. Phys Rev E 59:739–745
15. Geng J, Behringer RP, Reydellet G, Clément E (2003) Green's function measurements of force transmission in 2d granular materials. Physica D 182:274
16. Daniels KE, Behringer RP (2005) Hysteresis and competition between disorder and crystallization in sheared and vibrated granular flow. Phys Rev Lett 94:168001
17. Daniels KE, Behringer RP (2005) Characterization of a freezing/melting transition in a vibrated and sheared granular medium. In: Garcia-Rojo R, Herrmann HJ, McNamara S (eds) Powders, Grains 05 357–360 Balkema, Rotterdam
18. Daniels KE, Behringer RP (2006) Characterization of a freezing/melting transition in a vibrated and sheared granular medium. J Stat Mech 7:P07018
19. Longhi E, Easwar N, Menon N (2002) Large force fluctuations in a flowing granular medium. Phys Rev Lett 89:0455011–0455014
20. Ferguson A, Chakraborty B (2007) Spatially heterogenous dynamics in dense, driven granular flows. Europhys Lett 78
21. Silbert LE, Ertas D, Grest GS, Halsey TC, Levine D (2002) Analogies between granular jamming and the liquid-glass transition. Phys Rev E 65:051307
22. Silbert L (2005) Temporally heterogeneous dynamics in granular flows. cond-mat/0501617
23. Silbert LE, Grest GS, Brewster R, Levine AJ (2007) Rheology and contact lifetimes in dense granular flows. Phys Rev Lett 99:068002
24. Chandler D (1987) Introduction to Modern Statistical Mechanics. Oxford University Press, New York
25. Dolnik M, Zhabotinsky A, Epstein I (1996) Modulated standing waves in a short reaction-diffusion system. J Physical Chem 100:6604–6607
26. Dhar D (2006) Theoretical studies of self-organized criticality. Phys A-Stat Mech Appl 369:29–70
27. Schmittmann B, Zia RKP (1995) Statistical Mechanics of Driven Diffusive Systems. In: Domb C, Lebowitz J (eds) Phase Transitions and Critical Phenomena, vol 17. Academic Press, New York
28. Edwards SF, Blumenfeld R (2007) The thermodynamics of granular materials. In: Mehta A (ed) Physics of Granular Materials, Cambridge University Press, Cambridge
29. Edwards SF, Grinev DV (1999) Statistical mechanics of stress transmission in disordered granular arrays. Phys Rev Lett 82:5397
30. Edwards SF, Oakeshott RBS (1989) Theory of powders. Phys A 157:1080
31. Edwards SF, Grinev DV (2001) Jamming and Rheology: Con-

strained Dynamics on Microscopic and Macroscopic Scales. Taylor, New York
32. Donev A, Torquato S, Stillinger F, Connelly R (2004) Jamming in hard sphere and disk packings. J Appl Phys 95:989–999
33. Barrat A, Kurchan J, Loreto V, Sellitto M (2001) Edwards' measures: A thermodynamic construction for dense granular media and glasses. Phys Rev E 6305:0513011–05130114
34. Barrat A, Kurchan J, Loreto V, Sellitto M (2000) Edwards' measures for powders and glasses. Phys Rev Lett 85:5034
35. Coniglio A, Fierro A, Nicodemi M (2002) Probability distribution of inherent states in models of granular media and glasses. Eur Phys J E 9:219
36. Coniglio A, Fierro A, Nicodemi M (2001) Applications of the statistical mechanics of inherent states to granular media. Phys a-Stat Mech Appl 302:193
37. Kurchan J (2001) Recent theories of glasses as out of equilibrium systems. Comptes Rendus Acad Sci Ser Iv Phys Astrophys 2:239–247
38. Makse H, Kurchan J (2002) Testing the thermodynamic approach to granular matter with a numerical model of a decisive experiment. Nature 415:614–617
39. Blumenfeld R, Edwards SF (2003) Granular entropy: Explicit calculations for planar assemblies. Phys Rev Lett 90:1143031–1143034
40. Snoeijer JH, Vlugt TJH, van Hecke M, van Saarloos W (2003) Force network ensemble: a new approach to static granular matter. Phys Rev Lett 91:072303
41. Tkachenko A, Witten T (1999) Stress propagation through frictionless granular material Phys Rev E 60:687
42. Snoeijer J, Vlugt T, van Hecke M, van Saarloos W (2004) Force network ensemble: A new approach to static granular matter. Phys Rev Lett 92:0543021–0543024
43. Snoeijer J, Ellenbroek W, Vlugt T, van Hecke M (2006) Sheared force networks: Anisotropies, yielding, and geometry. Phys Rev Lett 96:09800191–0980014
44. van Eerd ART, Ellenbroek WG, van Hecke M, Snoeijer JH, Vlugt TJH (2007) Tail of the contact force distribution in static granular materials. Phys Rev E 75:0603021–0603024
45. Snoeijer J, van Hecke M, Somfai E, van Saarloos W (2003) Force and weight distributions in granular media: Effects of contact geometry. Phys Rev E 67:0303021–0303024
46. Snoeijer J, van Hecke M, Somfai E, van Saarloos W (2004) Packing geometry and statistics of force networks in granular media. Phys Rev E 70:0113011–01130115
47. Snoeijer J, Vlugt T, Ellenbroek W, van Hecke M, van Leeuwen J (2004) Ensemble theory for force networks in hyperstatic granular matter. Phys Rev E 70:0613061–06130616
48. Tighe B, Socolar J, Schaeffer D, Mitchener W, Huber M (2005) Force distributions in a triangular lattice of rigid bars. Phys Rev E 72:0313061–03130610
49. O'Hern CS, Silbert LE, Liu AJ, Nagel SR (2003) Jamming at zero temperature and zero applied stress: The epitome of disorder. Phys Rev E 68:011306
50. Bertin E, Dauchot O, Droz M (2006) Definition and relevance of nonequilibrium intensive thermodynamic parameters. Phys Rev Lett 96:1206011–1206014
51. Bertin E, Dauchot O, Droz M (2005) Nonequilibrium temperatures in steady-state systems with conserved energy. Phys Rev E 71:0461401–04614014
52. Bertin E, Dauchot O, Droz M (2004) Temperature in nonequilibrium systems with conserved energy. Phys Rev Lett 93:2306011–2306014
53. Henkes S, O'Hern CS, Chakraborty B (2007) Entropy and temperature of a static granular assembly: An ab initio approach. Phys Rev Lett 99:0380021–0380024
54. Blumenfeld R (2007) On entropic characterization of granular materials. To appear in: Aste T, Tordesiillas A, Matteo T (eds) Lecture Notes. World Scientific, Singapore
55. Ball RC, Blumenfeld R (2002) Stress field in granular systems: Loop forces and potential formulation. Phys Rev Lett 88:115505
56. Henkes S, Chakraborty B (2005) Jamming as a critical phenomenon: A field theory of zero-temperature grain packings. Phys Rev Lett 95:1980021–1980024
57. Henkes S, Chakraborty B (2008) Field theory, soft modes and the nature of jamming the critical point. Phys Rev E (to appear)
58. Liu AJ, Nagel SR (1998) Jamming is not just cool anymore. Nature 396:21
59. Majmudar TS, Sperl M, Luding S, Behringer RP (2007) Jamming transition in granular systems. Phys Rev Lett 98:058001
60. Wyart M (2005) On the rigidity of amorphous solids. Ann Phys 30:1–96
61. Wyart M, Nagel S, Witten T (2005) Geometric origin of excess low-frequency vibrational modes in weakly connected amorphous solids. Europhys Lett 72:486–492
62. Kirkpatrick TR, Thirumalai D, Wolynes PG (1989) Scaling concepts for the dynamics of viscous liquids near an ideal glassy state. Phys Rev A 40:1045
63. Xia X, Wolynes PG (2001) Fragilities of liquids predicted from the random first order transition theory of glasses. Phys Rev Lett 86:5526
64. Bouchaud J, Biroli G (2004) On the adam-gibbs-kirkpatrick-thirumalai-wolynes scenario for the viscosity increase in glasses. J Chem Phys 121:7347–7354
65. Donev A, Connelly R, Stillinger FH, Torquato S (2007) Underconstrained jammed packings of nonspherical hard particles: Ellipses and ellipsoids. Phys Rev E 75:0513041–05130432
66. Reynolds O (1885) On the dilatancy of media composed of rigid particles in contact. Philos Mag Ser 5 50–20:469
67. Nedderman RM (1992) Statics and kinematics of granular materials. Cambridge University Press, Cambridge
68. Majmudar TS, Behringer RP (2005) Contact force measurements and stress-induced anisotropy in granular materials. Nature 435:1079–1082
69. Blume M, Emery V, Griffiths RB (1971) Ising model for the lambda transition... Phys Rev A 4:1071
70. Lois G, Lemaitre A, Carlson JM (2007) Spatial force correlations in granular shear flow. i. numerical evidence. Phys Rev E 76:0213021–02130212
71. Lois G, Lemaitre A, Carlson JM (2007) Spatial force correlations in granular shear flow. ii. theoretical implications. Phys Rev E 76:0213031–02130314
72. Ferguson A, Chakraborty B (2006) Stress and large-scale spatial structures in dense, driven granular flows. Phys Rev E 73:0113031–0113037
73. Ferguson A, Fisher B, Chakraborty B (2004) Impulse distributions in dense granular flows: Signatures of large-scale spatial structures. Europhys Lett 66:277–283
74. Hurley M, Harrowell P (1995) Kinetic structure of a two-dimensional liquid. Phys Rev E 52:1694–1698

75. Perera D, Harrowell P (1996) Kinetic structure of a two-dimensional liquid. Phys Rev E 54:1652
76. Pouliquen O (1999) Scaling laws in granular flows down rough inclined planes. Phys Fluid 11:542–548
77. Ertas D, Halsey T (2002) Granular gravitational collapse and chute flow. Europhys Lett 60:931–937
78. Kolb E, Cviklinski J, Lanuza J, Clauding P, Clement E (2004) Reorganizationof a dense granular assembly: The unjamming response function. Phys Rev E 69:031306
79. Drocco JA, Hastings MB, Reichardt CJO, Reichardt C (2005) Multiscaling at point j: Jamming is a critical phenomenon. Phys Rev Lett 95:088001
80. Geng J, Behringer RP (2005) Slow drag in two-dimensional granular media. Phys Rev E 71:011302
81. Ediger MD, Angell CA, Nagel SR (1996) Supercooled liquids and glasses. J Phys Chem 100:13200
82. Majmudar TS, Behringer RP (2005) Contact forces and stress induced anisotropy. In: Garcia-Rojo R, Herrmann HJ, McNamara S (eds) Powders and Grains. 65 AA Balkema, Leiden
83. Snoeijer JH, van Hecke M, Somfai E, van Saarloos W (2004) Force and weight distributions in granulr media: Effects of contact geometry. Phys Rev E 67:030302
84. Snoeijer JH, Vlugt TJH, van Hecke M, van Saarloos W (2004) Force network ensemble: A new approach to static granular matter. Phys Rev Lett 92:054302
85. Hartley RR, Behringer RP (2003) Logarithmic rate dependence of force networks in sheared granular materials. Nature 421:928
86. Corwin EI, Jaeger HM, Nagel SR (2005) Structural signature of jamming in granular media. Nature 435:1075–1078
87. Corwin E, Jaeger H, Nagel S (2005) Structural signature of jamming in granular media. Nature 435:1075–1078
88. Daniels KE, Behringer RP (2005) Hysteresis and competition between disorder and crystallization in sheared and vibrated granular flow: Hysteresis and competition between disorder and crystallization in sheared and vibrated granular flow. Phys Rev Lett 94:168001
89. Behringer RP, Chakraborty B, Henkes S, Hartley RR (2008) Why do granular materials stiffen with shear rate? a test of novel stress-based statistics. Phys Rev Lett (to appear)
90. Makse H, Johnson D, Schwartz L (2000) Packing of compressible granular materials. Phys Rev Lett 84:4160–4163
91. O'Hern C, Langer S, Liu A, Nagel S (2001) Force distributions near jamming and glass transitions. Phys Rev Lett 86:111
92. O'Hern C, Langer S, Liu A, Nagel S (2002) Random packings of frictionless particles. Phys Rev Lett 88:075507
93. Ellenbroek WG (2007) Response of Granular Media near the Jamming Transition. Ph.D thesis, Leiden University
94. Wyart M, Silbert LE, Nagel SR, Witten TA (2005) Effects of compression on the vibrational modes of marginally jammed solids. Phys Rev E 72:051306
95. Silbert LE, Liu AJ, Nagel SR (2005) Vibrations and diverging length scales near the unjamming transition. Phys Rev Lett 95:098301
96. Tighe BP, Socolar JES, Schaeffeer DG, Mitchener WG, Huber ML (2005) Force distributions in a triangular lattice of rigid bars. Phys Rev E 72:031306
97. Torquato S, Truskett T, Debenedetti P (2000) Is random close packing of spheres well defined? Phys Rev Lett 84:2064–2067
98. Krzakala F, Kurchan J (2007) Landscape analysis of constraint satisfaction problems. Phys Rev E 76:0210021–02100213
99. Schwartz J, Liu A, Chayes L (2006) The onset of jamming as the sudden emergence of an infinite k-core cluster. Europhys Lett 73:560–566
100. Cates M, Wittmer J, Bouchaud J, Claudin P (1998) Jamming, force chains, and fragile matter. Phys Rev Lett 81:1841–1844
101. Cates M, Wittmer J, Bouchaud J, Claudin P (1998) Development of stresses in cohesionless poured sand. Philos Trans Royal Soc London Ser A-Math Phys Eng Sci 356:2535–2560
102. Cates ME, Wittmer JP, Bouchaud JP, Claudin P (1999) Jamming and static stress transmission in granular materials. Chaos 9:511–522
103. Coppersmith S, Liu C, Majumdar S, Narayan O, Witten T (1996) Model for force fluctuations in bead packs. Phys Rev E 53:4673–4685
104. To K (2005) Jamming transition in two-dimensional hoppers and silos. Phys Rev E Stat Nonlin Soft Matter Phys 71:060301
105. Zuriguel I, Garcimartin A, Maza D, Pugnaloni LA, astor JM (2005) Jamming during the discharge of granular matter from a silo. Phys Rev E Stat Nonlin Soft Matter Phys 71:051303
106. Easwar N Private Communication
107. Monthus C, Bouchaud J-P (1996) Models of traps and glass phenomenology. J Phys A 29:3847
108. Sollich P (1998) Rheological constitutive equation for a model of soft glassy materials. Phys Rev E 58:738
109. Bouchaud J, Cugliandolo L, Kurchan J, Mezard M (1996) Mode-coupling approximations, glass theory and disordered systems. Physica A 226:243–273
110. Song C, Wang P, Potiguar F, Makse H (2005) Experimental and computational studies of jamming. J Phys-Condens Matter 17:S2755–S2770
111. Song C, Wang P, Makse H (2005) Experimental measurement of an effective temperature for jammed granular materials. Proc National Acad Sci USA 102:2299–2304
112. Potiguar F, Makse H (2006) Effective temperature and jamming transition in dense, gently sheared granular assemblies. Euro Phys J E 19:171–183
113. Ono I et al (2002) Effective temperatures of a driven system near jamming. Phys Rev Lett 89:0957031–0957034
114. O'Hern C, Liu A, Nagel S (2004) Effective temperatures in driven systems: Static versus time-dependent relations. Phys Rev Lett 93:1657021–1657024
115. Desmond K, Franklin SV (2006) Jamming of three-dimensional prolate granular materials. Phys Rev E Stat Nonlin Soft Matter Phys 73:031306
116. Man W et al (2005) Experiments on random packings of ellipsoids. Phys Rev Lett 94:198001
117. Blouwolff J, Fraden S (2006) The coordination number of granular cylinders. Europhys Lett 76:1095

Books and Reviews

Blumenfeld R (2004) Stresses in isostatic granular systems and emergence of force chains. Phys Rev Lett 93:108301

Bouchaud JP (2003) Granular media: Some ideas from statistical physics. In: Barrat JL, Dalibard J, Feigelman M, Kurchan J (eds) Slow relaxations and nonequilibrium dynamics in condensed matter. Springer, Berlin

Bouchaud JP, Claudin P, Levine D, Otto M (2001) Force chain splitting in granular materials: A mechanism for large-scale pseudo-elastic behaviour. Eur Phys J E 4:451–457

Coniglio A, Fierro A, Herrmann H, Nicodemi M (eds) (2004) Unifying concepts in granular media and glasses, 1st edn. Elsevier, Amsterdam
Mehta A (1994) Granular matter: an interdisciplinary approach. Springer, New York
Halsey TC, Mehta A (2002) Challenges in granular physics. World Scientific, Singapore
Herrmann HJ, Hovi JP, Luding S (1998) Physics of dry granular media. In: NATOASI series. Series E, Applied sciences, vol 350. Kluwer Academic, Dordrecht
Jiang Y, Liu M (2007) A brief review of "granular elasticity": why and how far is sand elastic? Eur Phys JE Soft Matter 22:255–260
Krimer DO, Pfitzner M, Brauer K, Jiang Y, Liu M (2006) Granular elasticity: general considerations and the stress dip in sand piles. Phys Rev E Stat Nonlin Soft Matter Phys 74:061310
Otto M, Bouchaud JP, Claudin P, Socolar JES (2003) Anisotropy in granular media: classical elasticity and directed-force chain network. Phys Rev E Stat Nonlin Soft Matter Phys 67:031302
Ovarlez G, Fond C, Clement E (2003) Overshoot effect in the janssen granular column: a crucial test for granular mechanics. Phys Rev E Stat Nonlin Soft Matter Phys 67:060302
Torquato S, Donev A, Stillinger F (2003) Breakdown of elasticity theory for jammed hard-particle packings: conical nonlinear constitutive theory. Int J Solid Struct 40:7143–7153

Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of

Karin A. Dahmen[1], Yehuda Ben-Zion[2]
[1] Department of Physics, University of Illinois at Urbana-Champaign, Urbana, USA
[2] Department of Earth Sciences, University of Southern California, Los Angeles, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Models
Theoretical Results
Summary
Future Directions
Acknowledgments
Bibliography

Glossary

Mean field theory A theoretical approximation with an interaction field that has constant strength and infinite range. In mean field approximation every domain interacts equally strongly with every other domain, regardless of their relative distance.

Critical point A (phase transition) point in the parameter space of a physical system where the length scale characteristic of its structure, called the correlation length ξ, becomes infinite and the system displays power law scaling behavior on all available scales. The associated critical power law exponents are universal, i. e. they are independent of the microscopic details of the system.

Universality Power law scaling exponents and scaling functions near a critical point are the same for a class of systems, referred to as universality class, independent of the microscopic details. Universal aspects typically depend only on a few basic physical attributes, such as symmetries, range of interactions, dimensions, and dynamics.

Tuning parameters Parameters such as disorder, temperature, pressure, driving force etc. that span phase diagrams. Critical values of the tuning parameters describe critical points of the phase diagrams.

Renormalization group (RG) A set of mathematical tools and concepts used to describe the change of physics with the observation scale. Renormalization Group techniques can be used to identify critical points of a system as fixed points under a coarse graining transformation, and to calculate the associated critical power law exponents and the relevant tuning parameters. They can also be used to determine what changes to the system will leave the scaling exponents unchanged, and thus to establish the extent of the associated universality class of the critical point.

Earthquake quantities The most common form of earthquake data consists of seismic catalogs that list the time, location, and size of earthquakes in a given space-time domain. The size of earthquakes is usually specified by magnitudes associated with spectral amplitudes of seismograms at a given frequency and site-instrument conditions. The seismic potency and moment provide better physical characterizations for the overall size of earthquakes. Additional important quantities are the geometry of faulting (e. g., strike slip), stress drop at the source region, and radiated seismic energy.

Seismic potency A physical measure for the size of earthquakes given by the integral of slip over the rupture area during a seismic event.

Seismic moment A physical measure of earthquakes given by the rigidity at the source region times the seismic potency.

Strike slip fault A style of faulting involving pure horizontal tangential motion, predicted for situations where the maximum and minimum principal stresses are both horizontal. Prominent examples include the

San Andreas fault in California, the Dead Sea transform in the Levant and the North Anatolian fault in Turkey.

Definition of the Subject

Observations indicate that earthquakes and avalanches in magnetic systems (Barkhausen Noise) exhibit broad regimes of power law size distributions and related scale-invariant quantities. We review results of simple models for earthquakes in heterogeneous fault zones and avalanches in magnets that belong to the same universality class, and hence have many similarities. The studies highlight the roles of tuning parameters, associated with dynamic effects and property disorder, and the existence of several general dynamic regimes. The models suggest that changes in the values of the tuning parameters can modify the frequency size event statistics from a broad power law regime to a distribution of small events combined with characteristic system size events (characteristic distribution). In a certain parameter range, the earthquake model exhibits mode switching between both distributions. The properties of individual events undergo corresponding changes in different dynamic regimes. Universal scaling functions for the temporal evolution of individual events provide similar predictions for the earthquake and magnet systems. The theoretical results are generally in good agreement with observations. Additional developments may lead to improved understanding of the dynamics of earthquakes, avalanches in magnets, and the jerky response to slow driving in other systems.

Introduction

Global Statistics and Power Law Scaling

Earthquakes occur in a broad spectrum of sizes, ranging from unnoticeable tremors to catastrophic events. While short term earthquake prediction is still beyond reach, understanding the statistics of earthquakes might facilitate longer term prediction of large earthquakes and statistical estimates of seismic hazard. Gutenberg and Richter [29] found that the frequency of observed regional and global earthquakes versus magnitude forms a regular function over a very large range of scales (see Fig. 1). When the measure for the earthquake size is the seismic potency or moment (see "Glossary"), the frequency size statistics of regional and global earthquakes follow a power law distribution. Precise definitions and details on the seismic potency and moment are given in [1] and [3]. (In this paper we assume a unit nominal rigidity and will therefore use potency and moment interchangeably.) Omori [57] found

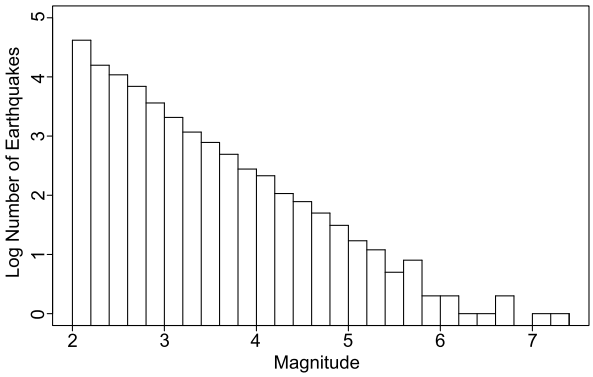

Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Figure 1
Histogram of earthquakes with magnitude 2.0 or larger recorded by the Southern California network during 1984–2002. The earthquake catalog is available at http://www.data.scec.org/research/SHLK.html

that the decay rate of aftershocks with time follows a power law distribution. One would expect that there might be a simple explanation for why earthquakes occur in a broad range of sizes and follow regular statistical patterns!

In the last two decades it has become increasingly evident that there are many other systems that respond to slowly changing external conditions with events on extremely large ranges of scales ("crackling noise"). An example of particular interest here involves magnets, which respond to a slowly varying external field by changing their magnetization in a series of bursts or "avalanches" called Barkhausen Noise. Just like earthquakes, these avalanches come in many sizes, ranging from microscopic to macroscopic and are distributed according to a regular function over the entire range. The spectra of the source time function of earthquakes is approximately flat up to a corner frequency related to the rupture size, followed by a power law decay at higher frequencies [1,3]. Similarly, the spectra of the number of spins flipping per time during an avalanche in magnets has high frequency power law decay with a low frequency roll off [68]. For certain values of tuning parameters, earthquake and magnet quantities are associated with scale invariant functions (power laws). In such cases each individual magnetic avalanche or earthquake slip has fractal spatial structure (see Fig. 2 for magnets and Fig. 5a for earthquakes). Other systems with similar "collective events" of all available sizes include, among others, superconductors, charge density waves, and group decision making [66].

While there are several interesting recent reviews, pointing out the similarities between systems with power law event size distributions, the goal of this paper is to de-

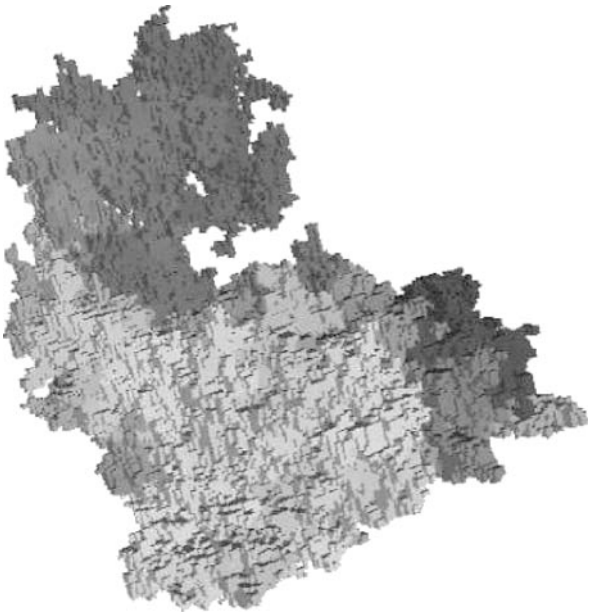

Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Figure 2
Fractal spatial structure of a medium sized avalanche of 282 785 domain flips in the 3 dimensional random field Ising model [66]. Fractal structures and power laws are characteristic of systems at their critical point. The shading represents time of the domain flips: the first domains to flip are at the right end of the avalanche, the last towards the left. The short range of the ferromagnetic interactions causes the avalanche to be spatially connected (see [66])

velop in detail some of the connections and analysis methods in earthquake and magnetic systems. Expanding on some of our earlier results, we focus especially on the role of disorder and dynamic changes in the strength threshold as potential tuning parameters to drive the system toward power law scaling behavior or away from it. Table 1 summarizes some of the similarities between magnets and earthquakes that are discussed in this review.

In Sect. "Models" we describe several magnet and earthquake models that are simple enough to make the connections transparent and easy to recognize. In Sect. "Theoretical Results" we review theoretical results obtained from these models and their comparison to experimental or observational data. Finally in Sect. "Summary" we summarize the results and discuss future work, both observationally and theoretically, that can help to improve our understanding of the dynamics of earthquakes and magnets.

Models

Models for Barkhausen Noise in Magnets

Hysteresis and avalanches in disordered magnetic materials have been modeled using several variants of the nonequilibrium, zero-temperature random-field Ising model (RFIM), which is one of the simplest models of magnetism, with applications far beyond magnetic systems (for a review, see [66], and also [15,56,58]). In contrast to some other hysteresis models, like the Preisach model [44] and the Stoner–Wohlfarth model [33], where interactions between the individual hysteretic units (grains) are not included and collective behavior in the form of avalanches is not addressed, in the RFIM the inter grain coupling is an essential feature and cause for hysteresis and avalanche effects.

The Random Field Ising Model (RFIM) The *equilibrium* RFIM was originally introduced to study disordered magnetic materials in thermal equilibrium. We study the *nonequilibrium* version, to model hysteresis and

Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Table 1
Some scaling features that are similar for magnets and earthquakes. More details on the various properties are given in the indicated sections

	Earthquake system	Magnetic system
Frequency size statistics	Power law near criticality, characteristic distribution away from criticality (Sect. "Results on the Monotonic Version of the Model")	Same as earthquakes (Sect. "Results on the Monotonic Version of the Model")
Scaling of source shape functions	Parabola for moment rate shape of events with fixed duration T in simulations, scaling function skewed to the left for observational data (Sect. "Moment Rate Shapes for Monotonic Models")	Same as earthquakes (Sect. "Moment Rate Shapes for Monotonic Models")
Spatial properties of individual events	Fractal near criticality, compact away from criticality (Sect. "Non-monotonic Models" and Fig. 5)	Same as earthquakes (Sects. "Non-monotonic Models"" and "Phase Diagram"" and Fig. 2)
Spectral decay of source function of individual events	Flat up to a corner frequency followed by power law decay (Sect. "Introduction")	Same as earthquakes (Sect. "Introduction")

avalanches observed *far from thermal equilibrium*. Even though the model is a toy version of the microscopic details in a magnet, near the critical point it correctly describes the large scale behavior of systems with the same general properties such as symmetries, dimensions, interaction ranges and dynamics [15], as follows from renormalization group arguments.

In the RFIM, to each site i in a simple cubic lattice is assigned a variable s_i, a so called "spin", which can take two different values, $s_i = +1$ ("up") or $s_i = -1$ ("down"). (This corresponds to a real magnet where a crystal anisotropy prefers the magnetic moments or elementary domains, represented by the spins, to point along a certain easy axis.) Each spin interacts with its nearest neighbors on the lattice through a positive exchange interaction, J_{nn}, which favors parallel alignment. (For the behavior on large scales the exact range of the microscopic interaction is irrelevant, so long as it is finite.) Some variations of the RFIM also include *long range* interactions due to the demagnetizing field and the dipole-dipole interactions. A general form of the Hamiltonian can be written as [37]

$$\mathcal{H} = -\sum_{\mathrm{nn}} J_{\mathrm{nn}} s_i s_j - \sum_i H s_i - \sum_i h_i s_i + \sum_i \frac{J_{\mathrm{inf}}}{N} s_i - \sum_{\{i,j\}} J_{\mathrm{dipole}} \frac{3\cos(\theta_{ij}) - 1}{r_{ij}^3} s_i s_j \,, \quad (1)$$

where H is the homogeneous external magnetic driving field, h_i is a local, uncorrelated random field, that models the disorder in the system, J_{inf} is the strength of an infinite range demagnetizing field, N is the total number of spins in the system, and J_{dipole} is the strength of the dipole-dipole interactions. The power laws of generated events are independent of the particular choice for the distribution $\rho(h_i)$ of random fields, for a large variety of distributions. Usually a Gaussian distribution of random fields is used, with a standard deviation ("disorder") R. As a simple approximation the model is studied at zero temperature, far from equilibrium, to describe materials with sufficiently high barriers to equilibration, so that temperature fluctuations are negligible on experimental time scales. As the magnetic field is adiabatically slowly changed between $H = -\infty$ to $H = +\infty$ two different *local* dynamics have been considered:

(1) in the first ("bulk") dynamics, each spin s_i flips while decreasing its own energy. We have studied this dynamics for the original RFIM without long range interactions, i. e. for $J_{\mathrm{inf}} = J_{\mathrm{dipole}} = 0$ [15,58]. This dynamics allows for *both* domain nucleation (when a spin s_i surrounded by equal valued spins flips in the opposite direction), *and* for domain wall motion (when a spin flips on the surface of a preexisting cluster of uniform spins in a background of opposite valued spins). A spin flip can trigger neighboring (or more generally, coupled) spins to flip as well, leading to an avalanche of spin flips, analogous to a real Barkhausen pulse. During an avalanche the external field is kept constant until the avalanche is finished, in accordance with the assumed adiabatic limit. The model is completely deterministic – two successive sweeps through the hysteresis loop produce the exact same sequence of avalanches (since the temperature is set to zero). This dynamics may be appropriate to describe for example hard magnetic materials with strong anisotropies. The analogue earthquake system may be associated with fault regions or fault networks that have strong geometrical and material heterogeneities.

(2) The second dynamics is a "front propagation dynamics" in which only the spins on the edge of an existing front (interface between up and down spins) flip if that decreases their energy. This dynamics can be used to model soft magnetic materials with a single or several noninteracting advancing domain walls and negligible new domain nucleation, due to antiferromagnetic demagnetizing fields. The front propagation model without long range interactions ($J_{\mathrm{inf}} = J_{\mathrm{dipole}} = 0$) was originally introduced by Robbins et al. to model fluids invading porous media [43]. The analogue earthquake system for this case may be associated with a single fault zone.

Simple Models for Inhomogeneous Earthquake Faults

Much of the previous work on simple earthquake models has involved variants of the Burridge–Knopoff (or "sliderblock") model, in which complex behavior is generated in a system with many degrees of freedom, and where inertia, friction laws and inherent discreteness play important roles [11,38,61]. These systems appear to exhibit power-law statistics over some range with a cutoff beyond some magnitude, and with most of the slip occurring in larger system-size events. However the understanding of the origin of the power law behavior is limited. Our approach here is to obtain an analytic understanding of a class of models and then to add in various additional features by analytic scaling arguments using tools from the theory of phase transition and the renormalization group, aided by numerical studies. There are interesting related studies using tools from statistical physics [12,63]. Some studies suggest that the power law scaling is connected to a spinodal [34]. Various cellular automata models have also been used for modeling earthquakes [40]. Rather than reviewing a large number of models, we will focus on a subgroup

of models, that we found particularly well suited to clarify the connection between earthquake and magnetic systems with a jerky response to slowly changing external conditions.

The Ben-Zion and Rice Model A representative of the class of models that we consider is a model developed originally by Ben-Zion and Rice [2,5,6], referred to below as the BZR model. The model assumes that a narrow irregular strike-slip fault zone of horizontal length L and vertical depth W may be represented by an array of $N \sim LW$ cells in a two dimensional planar region of length L and width W, with long range interaction, abrupt transitions in the threshold dynamics during failure, and constitutive parameters that vary from cell to cell to model the disorder (offsets etc.) of the fault zone structure (Fig. 3).

The cells represent brittle patches on the interface between two tectonic blocks that move with slow transverse velocity v in the x direction at a great distance from the fault. The interaction between cells during slip events is governed by 3-D elasticity and falls off with a distance r from the failure zone as $1/r^3$. The cells remain stuck while the stress τ_i on each cell is increased gradually as a result of the external loading which grows adiabatically (that is we take the limit $v \to 0$). When the stress on a cell i reaches its local failure threshold $\tau_{s,i}$, the cell slips until the stress is reduced to its local arrest stress $\tau_{a,i}$. Both failure stress and arrest stress are distributed according to some bounded probability distribution. The stress drop resulting from a cell failure is redistributed to the other cells according to the long range elastic stress transfer function. The resulting stress increase on the other cells can cause some of them to slip as well, leading to an avalanche of cell slips, or a model earthquake. A review of extensive numerical simulations with various versions of the BZR model, in relation to observed features of seismicity, criticality, and other dynamic regimes, is given in [75].

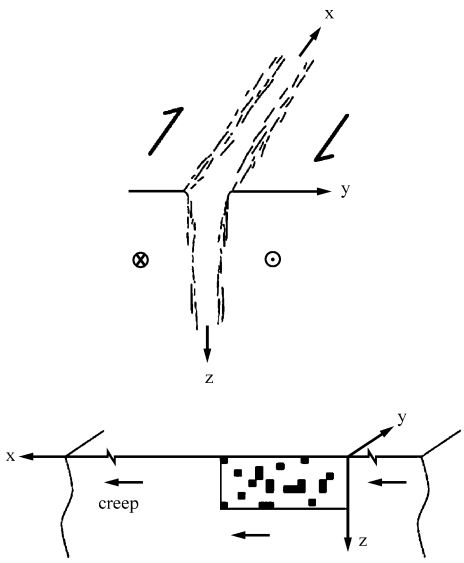

Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Figure 3
Illustration of the Ben-Zion and Rice (BZR) model: projection of a 3D fault zone (*top*) onto a 2D interface embedded in a 3D elastic halfspace (*bottom*). The geometrical inhomogeneities of the physical fault zone are modeled by spatially varying constitutive parameters of the brittle patches (see [46])

Dynamical Weakening The model includes dynamic weakening effects during the failure process [2,5,6]: after an initial slip in an earthquake, the strength of a failed cell is reduced to a *dynamical* value:

$$\tau_{d,i} \equiv \tau_{s,i} - \epsilon(\tau_{s,i} - \tau_{a,i}), \tag{2}$$

with $0 \le \epsilon \le 1$ parametrizing the relative importance of the dynamical weakening in the system. This weakening represents the transition from static friction to dynamic friction during the rupture. The strength of a failed cell remains at its dynamic value throughout the remainder of the earthquake. In the time intervals between earthquakes all failure thresholds heal back to their static value $\tau_{s,i}$.

Dynamical Strengthening The model can be expanded further to include dynamic strengthening represented by $\epsilon < 0$. Multidisciplinary observations indicate [7] that brittle failure of rock has an initial transient phase associated with strengthening, distributed deformation, and creation of new structures. Detailed frictional studies also show an initial strengthening phase associated with the creation of a new population of asperity contacts [3,19]. In mean field studies of our model (Fig. 3) discussed in Sect. "Results on Aftershocks", we associate $\epsilon < 0$ with regions off the main fault segments that are in an early deformation stage. The events that are triggered as the failure stresses are lowered back in the following weakening period are referred to as *aftershocks*. The Omori law [3,69,70] is obtained if we assume that the increased failure stress thresholds $\tau_{f,i}$ are slowly lowered with time as $\log(t)$ towards their earlier static values $\tau_{s,i}$, and that the stresses are distributed over a wide range of values [46].

Related General Continuum Equations of Motion The above model is a special case of a more general class of models for infinite systems driven by a constant drive force F [27]. We consider general equations of motion of

the form:

$$\eta \partial u(\mathbf{r}, t)/\partial t = F + \sigma(\mathbf{r}, t) - f_R[u(\mathbf{r}, t), \mathbf{r}, \{u(\mathbf{r}, t' < t)\}] \quad (3)$$

where

$$\sigma(\mathbf{r}, t) = \int_{-\infty}^{t} \mathrm{d}t' \int \mathrm{d}^d r' J(\mathbf{r} - \mathbf{r}', t - t') \cdot [u(\mathbf{r}', t') - u(\mathbf{r}, t)] \quad (4)$$

is the stress and f_R is a quenched random "pinning" force crudely representing inhomogeneities in the friction, asperities, stepovers etc., which in general can depend on the local past history (e. g. as in velocity dependent friction). The dynamical variables $u(\mathbf{r}, t)$ are assumed to represent the discontinuity across the fault plane in the component of the displacement in the direction of slip. The dynamics depend on the local history dependence of the pinning force, the stress transfer function $J(\mathbf{r}, t)$, and the coefficient η that represents the fault impedance. (In an elastic medium, the impedance depends on mass density, the elastic parameters, and directional parameters [1].) Equation (3) can be considered a continuum description of the rules of the BZR model. Integrating out the degrees of freedom due to the bulk material on either side of the d=2 dimensional fault plane leaves us with effective long range static stress transfer: $J_s(r) \equiv \int \mathrm{d}t J(r, t) \sim 1/r^{d+\Gamma} \sim 1/r^3$. For a planar fault in an elastic half space, $d = 2$ and $\Gamma = 1$ [2,6]. The correlations in f_R are generally assumed to be short-range in u and $\mathbf{r}$. (For results on the BZR model with long range correlations in the disorder, see [2,27,77] and Sect. "Theoretical Results".) In a version of the BZR earthquake model with a constant driving force F, the loading may be replaced by driving through a weak spring with spring constant $K \sim 1/L$ coupled to the slowly moving continents far away (i. e. replacing F in Eq. (3) by $F(\mathbf{r}, t) = K[vt - u(\mathbf{r}, t)]$, with $v \to 0$).

Monotonic Models Substantial simplifications occur if f_R is history independent and $J(\mathbf{r}, t) \geq 0$ for all $(\mathbf{r}, t)$; leading to *monotonic* models [27]. Related monotonic models have been studied extensively in various other contexts [25,53]. Examples include elastic depinning models for contact lines, vortex lines, liquids invading porous materials, and elastic charge density waves. Their crucial simplifying feature is that the steady state velocity $\overline{v} \equiv \langle \partial u/\partial t \rangle$ is a history independent function of F [48]. In the context of the BZR model this corresponds to the case with zero weakening ($\epsilon = 0$) and non-negative J. A crucial feature of monotonic models is that the slip profile $\Delta u(\mathbf{r})$ of a quake is *independent of the dynamics* [48]. However, several interesting dynamical issues discussed below are associated with the effects left out of the monotonic models that can make this feature break down.

Non-monotonic Models (a) *Weakening*: We first consider including some weakening effects of sections which have already slipped in a given quake. This is best studied in the discrete model. In analogy to the dynamic weakening in the BZR model discussed above, we choose [27]

$$f_R = \tilde{f}_R[u(\mathbf{r}), \mathbf{r}]\{1 - \epsilon \Theta[u(\mathbf{r}, t) - u(\mathbf{r}, t - T)]\} \quad (5)$$

with T a cutoff time much longer than the duration of the largest quakes, but much smaller than the interval between the quakes. Here $\Theta(x)$ is the Heavyside step function. As mentioned, the case $\epsilon > 0$ represents the difference between static and dynamic friction. The effects of small weakening ($\epsilon > 0$) can be analyzed perturbatively (see Sect. "Theoretical Results").
(b) *Stress Pulses*: A similar but more subtle effect can be caused by stress pulses that result from non-positive $J(\mathbf{r}, t)$; these arise naturally when one includes elastodynamic effects. We consider

$$J(\mathbf{r}, t) \sim \frac{\delta\left(t - \frac{r}{c}\right)}{r^{d+\Gamma}} + \frac{\alpha \delta'\left(t - \frac{r}{c}\right)}{c r^{d+\gamma}} \quad (6)$$

with c the sound speed, $\delta(t)$ the Dirac delta distribution, and $\delta'(t) = \mathrm{d}\delta(t)/\mathrm{d}t$. The scalar approximation to elasticity in a half space corresponds to $d = 2$, $\Gamma = 1$, $\gamma = 0$, and $\alpha = 1$ [27]. If a region slips forward, the stress at another point first has a short pulse at the sound arrival time from the second term in Eq. (6), and then settles down to its smaller static value, i. e. it is non-monotonic. The magnitude of these stress pulses and their duration is set by various aspects of the models; for example larger η in Eq. (3) implies weaker stress pulses as the local motion will be slower.

Theoretical Results

Both the magnet and earthquake models of the previous section are capable of producing a large range of power law scaling of event sizes, and related scale invariant quantities in response to a slowly varying driving force or field. This section highlights similarities between these different physical systems and attempts to explain them.

The Universality Class of the BZR Model

We first review results for the simplified monotonic case, starting with scaling relations for driving with fixed force

and far field plate motion and continuing with moment rate shapes. We then discuss additional results associated with non monotonic versions of the model, including mode-switching and aftershocks.

Results on the Monotonic Version of the Model

General results: Depinning transition As mentioned above, substantial simplifications occur for the monotonic version of the model, i. e. if f_R is history independent and $J(\mathbf{r}, t) \geq 0$ for all $(\mathbf{r}, t)$. In [25,27,53] it is shown that for F greater than a critical force F_c the displacement grows continuously in a "sliding state" for which the mean velocity $\overline{v} \equiv \langle \partial u/\partial t \rangle \sim (F - F_c)^\beta$. Here β is a universal exponent that is independent of the microscopic details of the system. It only depends on a few fundamental properties, such as symmetries, spatial dimensions d, range of interactions, etc. [53]. Long time dynamic properties such as β depend in addition on the small ω dependence of $J(\mathbf{q}, \omega)$ [54].

For F less than the critical force F_c, the mean velocity is $\overline{v} \equiv 0$. If F is adiabatically slowly increased towards F_c, the system moves from one metastable configuration to another by a sequence of "quakes" of various sizes. The "quakes" can be characterized by their radius R, the d-dimensional area A which slips (by more than some small cutoff), their potency or moment $M \equiv \int_A \mathrm{d}^d \mathbf{r} \Delta u(\mathbf{r})$, a typical displacement $\Delta u \sim M/A$, and a duration τ. The critical force F_c marks a second order phase transition point. Such phase transitions are typically associated with power law scaling behavior.

In the class of earthquake models with long range interactions along the fault involving the static stress transfer $J_s(r) \equiv \int \mathrm{d}t J(r, t) \sim 1/r^3$, the equations are very similar to those of a model for contact line depinning studied in ref. [25]. Using renormalization group methods it was shown in [25] that for a physical two dimensional interface (or "fault") in a 3 dimensional elastic half space, these long range interactions are so long that the scaling behavior near F_c is correctly described by mean field theory (up to logarithmic corrections, since $d = 2$ is the "upper critical dimension"). The main assumption in mean field theory is that the spatial and temporal fluctuations in the displacement field $u(\mathbf{r}, t)$ are so small that the local displacement $u(\mathbf{r}, t)$ can be replaced by a time dependent spatial average $u(t)$, which then needs to be determined self consistently from the behavior of the neighboring regions that contribute to the stress at a chosen point $\mathbf{r}$ [26]. The same mean field equations are obtained when the long range interaction is approximated to be constant in space $J(\mathbf{r}, t) = J_{mft}(t)/(LW)$. With this approximation Eqs. (3) and (4) become

$$\eta \partial u(\mathbf{r}, t)/\partial t = F + \sigma_{mft}(\mathbf{r}, t) - f_R[u(\mathbf{r}, t), \mathbf{r}, \{u(\mathbf{r}, t' < t)\}] \quad (7)$$

where

$$\sigma_{mft}(\mathbf{r}, t) = \int_{-\infty}^{t} \mathrm{d}t' J_{mft}(t - t')[u(t') - u(\mathbf{r}, t)] \quad (8)$$

and the self consistency requirement is

$$\int u(\mathbf{r}, t) \mathrm{d}^2 r/(LW) = u(t) \quad (9)$$

Many scaling exponents and scaling functions can be calculated exactly in mean field theory by solving these simplified equations of the model. In [15,26,42], several illustrative examples are given for solving similar self consistent mean field theories. There are various approaches that one may use, ranging from numerical simulations to analytical expansion and scaling analysis near a phase transition point where universal power law scaling occurs. The approach of choice to solve the mean field equations depends on the quantity under consideration. To obtain exact results for the scaling behavior of the frequency size statistics of earthquake or avalanche events, a fairly simple approach is to use a discrete version of the model in which we treat the fault as a discrete set of dislocation patches, coupled to a mean displacement and an external driving force that slowly increases with time. (The stress τ_i at each patch is given by Eq. (16) of Sect. "Mode–Switching" below.) As shown in [17], the sequence that describes the distance from failure of the rescaled stress variables resembles a biased random walk. The scaling behavior of the resulting random walk is known exactly from the literature. Using this mapping it then becomes straightforward to derive universal scaling predictions for the mean field earthquake frequency size distribution [17].

Furthermore, as shown in [13,25], their (and thus also our) model have the same scaling behavior as a front propagation model for a two dimensional domain wall in a soft magnet with long range dipolar magnetic interactions, driven by a slowly changing external field (see Sect. "Models"). A flipping spin in the magnet model corresponds to a slipping dislocation patch in the earthquake model. The long range elastic interactions in the earthquake model are similar to the long range dipolar magnetic interactions in the magnet model. The driven two dimensional magnetic domain wall in the (three dimensional) magnet model corresponds to the driven two dimensional earthquake fault

in a three dimensional elastic half space. Since the scaling behavior of the earthquake model and that of [25] and [73] are identical, we may simply copy their results and translate them into quantities that can be extracted from seismic data. Using tools from phase transitions, such as the renormalization group (RG), near the critical force the following scaling results were derived by [13,25,53,54,73] and others:

$$\Delta u \sim R^{\zeta} \ ,$$
$$A \sim R^{d_f} \ \text{ with } \ d_f \leq 2 \ \text{ a fractal dimension} \ ,$$
$$M \sim R^{d_f+\zeta} \ ,$$
$$\text{and } \ \tau \sim R^{z} \ .$$

The differential distribution $P(M)$ of moments M is shown in [25,54] and [26,27] to scale as

$$P(M)\mathrm{d}M \sim \mathrm{d}M/M^{1+B}\rho_{\infty}(M/\hat{M}) \tag{10}$$

with ρ_{∞} a universal scaling function which decays exponentially for large argument. The cutoff $\hat{M}$ for large moments is characterized by a correlation length – the largest likely radius – $\xi \sim 1/(F_c - F)^{\nu}$ with $\hat{M} \sim \xi^{d_f+\zeta}$.

In the same references it is shown that in mean-field theory, $B = 1/2$, $1/\nu = 1$, $z = 1$ and the quakes are fractal with displacements of order the range of correlations in $f_R(u)$, i. e. $\zeta = 0$.

These mean-field exponents are valid for a $d = 2$ dimensional planar fault in a three dimensional elastic half space [7], since the physical fault operates at the upper critical dimension. As usual, at the upper critical dimension, there are logarithmic corrections to mean-field results. Using renormalization group methods one can calculate these corrections [27] and finds barely fractal quakes with $A \sim R^2/\ln R$ so that the fraction of the area slipped decreases only as $1/\ln r$ away from the "hypocenter". The typical slip is $\Delta u \sim (\ln R)^{1/3}$ so that $M \sim R^2/(\ln R)^{2/3}$. The scaling form of $P(M)$ is the same as Eq. (10) with the mean-field ρ_{∞}, although for $M \ll \hat{M}$, $P(M) \sim (\ln M)^{1/3}/M^{3/2}$ so that B will be virtually indistinguishable from 1/2 [27]. A similar form of moment distribution and exponent value $B = 1/2$ were obtained also for a critical stochastic branching model [71].

More realistic driving of a fault We now consider more realistic drive and finite-fault-size effects. As mentioned, driving the fault by very slow motion far away from the fault is roughly equivalent to driving it with a weak spring, i. e. replacing F in Eq. (3) by $F(\mathbf{r}, t) = K[vt - u(\mathbf{r}, t)]$. With $v \to 0$ the system must then operate with the spring stretched to make $F(\mathbf{r}, t) \lesssim F_c$ at least on average, to ensure $\overline{v} = 0$; depending on the stiffness of the spring, it will actually operate just below F_c, as shown below. If in constant force drive the force is increased by a small amount ΔF, the average resulting slip per area, $\langle \Delta u \rangle \equiv \sum_i \Delta u_i/(LW)$ is given by the total potency/moment per total area $M \equiv \int \mathrm{d}^d\mathbf{r}\Delta u(\mathbf{r})/(LW)$. The total moment per area observed in response to a small force increase equals the number $n\Delta F$ of earthquakes per area that are triggered by the increase ΔF, multiplied with the average observed moment of a single earthquake $\langle M \rangle = \int MP(M)\mathrm{d}M$. The result is

$$\langle \Delta u \rangle = n\Delta F \int MP(M)\mathrm{d}M \tag{11}$$

where n is the number of quakes per unit area per force increase ΔF. It has been shown that $n(F)$ is non-singular at F_c [53], so it can be treated like a constant near F_c. Plugging in Eq. (10) and the scaling laws written above and below that equation, we obtain

$$\langle \Delta u \rangle \sim \Delta F \xi^{(2\tilde{\Gamma}+\zeta)(1-B)} \sim \Delta F \xi \tag{12}$$

for our case where mean field results can be used for the critical exponents. For consistency, we must have in steady state with the spring drive, $Kv\Delta t = \Delta F = K\Delta u$ so that the system will operate with a correlation length $\xi \sim 1/K^{1/\tilde{\Gamma}}$, i. e. $1/K$ for our case. For a fault section with linear dimensions of order L, drive either from uniformly moving fault boundaries or from a distance $\sim L$ perpendicularly away from the fault plane will be like $K \sim 1/L$ so the power-law quake distribution will extend out to roughly the system size $\xi \sim L$. For smaller quakes, i. e. $R \ll L$, the behavior will be the same as in the infinite system with constant F drive, but the cutoff of the distribution of moments will be like Eq. (10) with a different cutoff function ρ that depends on the shape of the fault, how it is driven, and the boundary conditions.

We have tested these conclusions numerically by simulating the BZR model, which is a discrete space, time, and displacement version of a monotonic Eq. (3), with quasistatic stress transfer appropriate for an elastic half space [2,5]. The slip, u, is purely in the horizontal direction along the fault and $f_R[u(\mathbf{r})]$ is a series of equal height spikes with spacings which are a random function of $\mathbf{r}$. When $\sigma(\mathbf{r}, t) > f_R[u(\mathbf{r}, t]$, $u(\mathbf{r})$ jumps to the next spike. This provides a way of implementing the random stress drops of the BZR model. The boundary conditions on the bottom and sides are uniform creep or slip – ($u = vt$) with infinitesimal v – and stress free on the top (Fig 3). The statistics of the moments of the quakes are shown by the triangles in Fig. 4. Although the uncertainties are appreciable, relatively good agreement is found with the prediction $B = 1/2$. One typical large quake is illustrated in

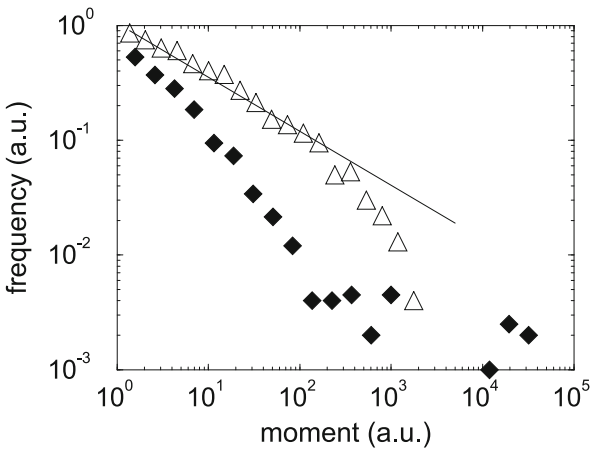

Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Figure 4
Histograms of moments for a simulation of a rectangular fault with 32 x 128 cells for the discrete monotonic quasistatic model (with arbitrary units (a.u.)). Triangles: without dynamical weakening ($\epsilon = 0$). Diamonds: with dynamic weakening of $\epsilon = 0.95$. (ϵ is defined in Eq. (5).) The straight line indicates the predicted slope $B = 1/2$ (from [27])

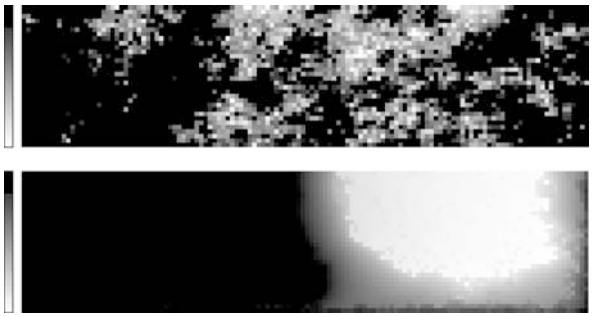

Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Figure 5
Distribution of horizontal slip, *u*, along a fault with 32 x 128 cells for a *single* large quake event. Lighter shading represents larger slip during the quake. *Top*: almost fractal quake with a total moment of 1750 (and 1691 cells failing) for the monotonic model without any dynamical effects ($\epsilon = 0$). *Bottom*: "crack like" quake with a total moment of 16922 (and 2095 cells failing) for the model with dynamic weakening ($\epsilon = 0.95$). In both cases the system is driven by horizontally creeping fault boundaries (*sides* and *bottom*) while the top boundary is free (from [27])

Fig. 5 (top); it appears almost fractal as predicted, and tends to stay away from the bottom and sides due to the specific loading that we chose. The ratios of the moments of quakes to their areas have been studied and found to grow only very slowly with the area, as predicted from the logarithmic corrections listed below Eq. (3). This is in striking contrast to earthquakes in conventional crack models which are compact (Fig. 5 (bottom)) and have $\Delta u \sim R$ (i. e. $\zeta = 1$), so that $M/A \sim \sqrt{A}$. As discussed by [8], however, the scaling $M \sim A$ appears to be consistent with observational results for small earthquakes which presumably propagate and are arrested in rough stress fields. More observational data on the scaling of the moment M with the slipping area A for smaller earthquakes would be highly desirable to test this prediction more precisely.

Because the system is at its critical dimension, the cutoff function ρ of the moment distribution appropriate to the boundary conditions, as well as various aspects of the shapes and dynamics of quakes can be computed using tools from the theory of phase transitions [26,27]. For quasistatic stress transfer, $J(\mathbf{r}, t) \sim \delta(t)/r^3$, in the infinite system the quake durations are found to scale as $\tau \sim R^z$ with $z = 1$ for a $d = 2$ dimensional fault, with logarithmic corrections [25]. (A more physical dynamics with sound-travel-time delay has slower growth of the quakes with $z = 1$ in all dimensions.) Due to the geometrical disorder included in the model, in either case the growth will be very irregular – including regions starting and stopping – in contrast to crack models and what is often assumed in seismological analysis of earthquakes on more regular faults. Similar fractal-like quakes were simulated by Zöller et al. [75,76], for a quasi-dynamic version of the BZR model that includes stress redistribution with a finite communication speed.

Moment Rate Shapes for Monotonic Models In both magnet and earthquake models it has been shown that there are not just universal scaling exponents but also some experimentally accessible universal scaling functions [66]. By comparing theoretical predictions for these functions to experiments or observations, one can often test models much more accurately than by merely comparing a finite set of discrete exponents. Two such functions were first discovered for Barkhausen Noise in magnets [47,66]. The analogy between magnets and earthquakes then lead to the development of the corresponding functions for earthquakes. For slowly driven magnets, consider the time history $V(t)$ of the number of domains flipping per unit time (Barkhausen train). It is called V because it is usually measured as a voltage in a pickup coil. An example of a Barkhausen train for a single avalanche is shown in Fig. 6.

The voltage function $V(t)$ in magnets is the analogue of the moment rate $\mathrm{d}m/\mathrm{d}t(t)$, or the slip per unit time for earthquakes. Recent analysis allowed researchers to obtain the moment rate $\mathrm{d}m_0(t)/\mathrm{d}t$, during the propagation of earthquake rupture for hundreds of large seismic events recorded on global networks [9,31]. The moment rates

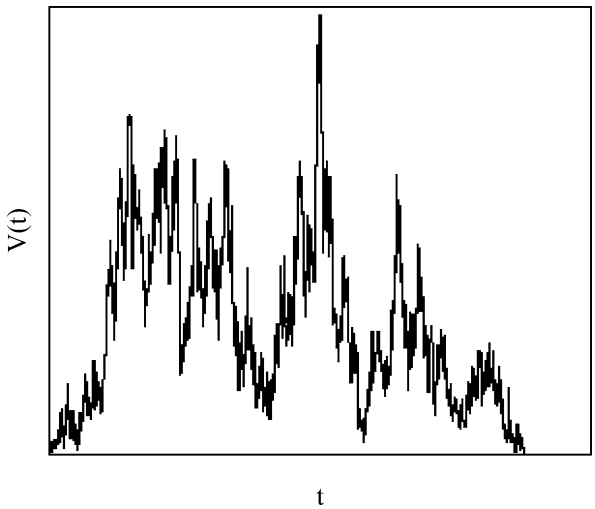

Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Figure 6
Voltage train of a typical large avalanche. Note that the voltage fluctuates drastically and the avalanche nearly stopped several times (from [37]). The analogous moment rate time trace for earthquakes (though measured with lower resolution) is shown in the right inset marked "RAW" of Fig. 8

shown below are derived from inversions of teleseismically recorded seismograms on a global seismic network [62]. (The frequency-moment distribution, $D(M_0) \sim M_0^{-1-\beta}$ of the observed data [9] has three decades of scaling and an exponent of $\beta = 1/2 \pm 0.05$, in close agreement with the BZR model near $\epsilon = 0$ [46].)

For both magnets and earthquakes there are large fluctuations in $V(t)$ and $\mathrm{d}m/\mathrm{d}t(t)$ respectively (Fig. 6). However averaging the signal over many avalanches, leads to typical shapes. Figure 7 shows the average over all avalanches of fixed duration T, $\langle V\rangle(T,t)$ obtained from simulations of two variants of the RFIM (a), and from three different Barkhausen noise experiments (b). Figure 8 shows $\langle \mathrm{d}m/\mathrm{d}t\rangle(T,t)$ obtained for the BZR earthquake model and derived from earthquake observations respectively. The renormalization group and scaling theory [66] predict that for a self similar system at a critical point with power law size and duration distributions for avalanches, there are self similar average avalanche profiles. As shown in [46,66] one finds

$$\langle \mathrm{d}m/\mathrm{d}t\rangle(T,t) \sim T^{b'} g(t/T) \tag{13}$$

where the function $g(x)$ is a universal scaling prediction and $b' \equiv 1/(\sigma\nu z) - 1 = 1$ for the BZR earthquake model (as obtained from mean field theory). The corresponding value for b' for magnets in three dimensions is smaller – the values used for the corresponding collapses can be read off for the different versions of the RFIM from the caption of Fig. 7.

Based on universality one would expect these theoretical predictions to agree with experimental results, apart from an overall shift in time and voltage or moment rate scales. For the moment rate of earthquakes this means

$$\langle \mathrm{d}m/\mathrm{d}t\rangle_{\mathrm{observation}}(T,t) = A\langle \mathrm{d}m/\mathrm{d}t\rangle_{\mathrm{theory}}(T/B, t/B) \tag{14}$$

for some rescaling factors A and B, and similarly for the average voltage $\langle V\rangle(t,T)$ in magnets. In both cases the theory predicts a symmetric looking profile. The mean field prediction for $g(x)$ is in fact a parabola [46,66] – the theoretical prediction thus is that events grow as quickly as they decay. As seen in Figs. 7b and 8 the experimental/observational profile in both cases, however, appear skewed – the real events tend to grow more quickly than they decay! A similar asymmetry has also been observed in avalanches associated with plastic deformation [39].

For magnets this apparent disagreement has been resolved by taking greater account of a microscopic detail involving eddy currents that had been neglected by previous models. Eddy currents are transient current loops that arise in conducting magnets in response to the reorientation of a magnetic domain. These currents temporarily prevent neighboring domains from being triggered to realign in the same direction in an avalanche of domain reversals. The eddy currents decay after a microscopic time τ given by the resistance of the material. Their delay effect thus also decays after a time τ. If the avalanche duration is large compared to τ this effect is negligible and the mean profile approaches the predicted symmetrical shape (see Fig. 7c and d).

The source of asymmetry in the mean moment rate profile may be similar for earthquakes [16]. It has been suggested that triggering delays – arising from a noticeable earthquake nucleation time, or an increase in the failure threshold during the formation of new cracks and subsequent weakening as rock damage increases – could be responsible for aftershocks that often follow large earthquakes [46]. On long time scales a large mainshock with smaller aftershocks can be seen as a similar asymmetry to that seen in magnets, possibly with a similar explanation.

There is a second scaling function that may be extracted from the same data: Fig. 9 shows the average over all earthquakes of fixed total moment $M\langle \mathrm{d}m/\mathrm{d}t\rangle(M,t)$, both for observations and the BZR model prediction. As shown in [46,66] the theory predicts

$$\langle \mathrm{d}m/\mathrm{d}t\rangle(M,t) \sim M^{1/2} q(t/M^{1/2}) \tag{15}$$

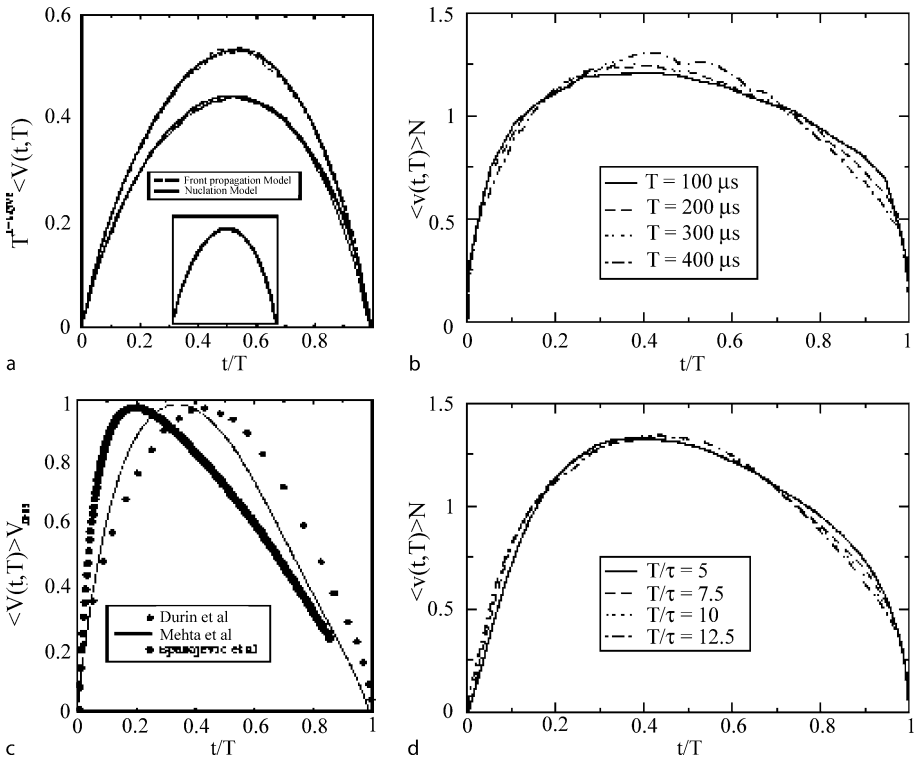

Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Figure 7

a Theoretical average avalanche shape scaling functions for fixed avalanche durations T **denoted with** $g(t/T)$ **in the text, for the nucleation and the front propagation RFIM [47]. The overall height is non universal, the curves for the two models are otherwise extremely similar. The front propagation model has** $1/\sigma\nu z = 1.72$ **and, the nucleation model has** $1/\sigma\nu z = 1.75$ **in this collapse. The** *inset* **shows the two curves rescaled to the same (non universal) height: the two curves are quantitatively different, but far more similar one to another than either is to the experimental curve in b. b Experimental average pulse shapes from three different experiments for fixed pulse duration, as measured by three different groups [21,23,47,67]. Notice that both theory curves are much more symmetric than those of the experiments. Notice also that the three experiments do not agree. At first this result represented a serious challenge to the idea about universality of the dynamics of crackling noise [66]. c Pulse shape asymmetry experiment [72]. Careful experiments show a weak but systematic duration dependence in the collapse of the average Barkhausen pulse shape. The longer pulses (larger avalanches) are systematically more symmetric (approaching the theoretical prediction). d Pulse shape asymmetry theory [72]. Incorporating the delay effects of eddy currents into the theoretical model produces a similar systematic effect. The non-universal effects of eddy currents are in principle irrelevant for extremely large avalanches (from [64])**

where the universal scaling function $q(x) = Ax\exp -Bx^2/2$ and the universal exponents are obtained from the mean field theory for the BZR earthquake model. A comparison between prediction and observational results for this scaling function is shown in Fig. 9.

Clearly, more data, especially for small earthquakes, are needed to decrease the statistical error bars of the observational data and determine the degree of agreement between theory and observations. An alternative scaling approach to moment rate data was given by [31] and a comparison between both approaches is discussed in [46].

Non-Monotonic Models We first consider including weakening of the cell failure threshold by an amount ϵ for sections which have already slipped in a given quake. This crudely models the difference in static versus dynamic friction (see Sect. "Models", Eq. (5)). In between quakes all thus weakened thresholds heal back to their static strength. The effects of small weakening can be analyzed perturbatively.

With $\epsilon = 0$, consider a quake of diameter R_1 ($\ll L$ *or* ξ), with moment M_1 and area A_1: i. e. A_1 sites have slipped. If a small ϵ is turned on at the end of the quake, all slipped sites that are within ϵ of slipping will now slip again – this will be $N_2^{ex} \sim \epsilon A_1$ sites. The simplest justifiable guess is that each of these will cause an approximately independent secondary quake. The total moment of these secondary quakes will be dominated by the largest one, so the extra moment will be $M_2^{ex} \sim (\epsilon A_1)^{1/B}$. (For a very large or

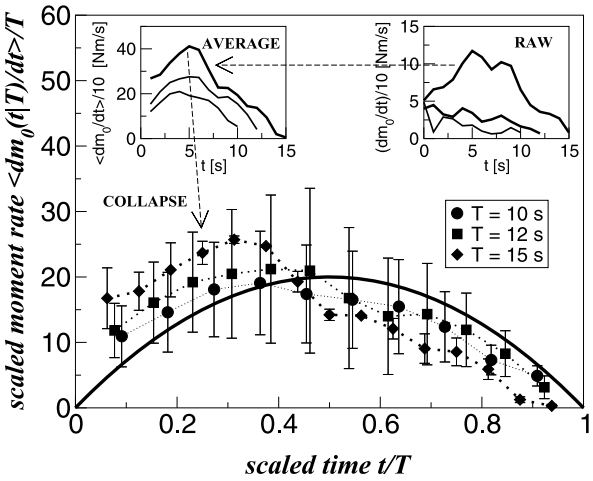

Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Figure 8

A collapse of averaged earthquake pulse shapes, $\langle dm_0(t|M_0)/dt\rangle$ with a duration of T (seconds) within 10% (given in legend), is shown. The collapse was obtained using the mean field scaling relation [37]: $\langle dm_0(t|T)/dt\rangle \sim g(t/T)$. In order to obtain each collapsed pulse shape, two to ten earthquakes were averaged for each value of T. In our mean field theory the universal scaling function is $g_{mf}(x) = Ax(1-x)$ with $x = t/T$. We plot this functional form (bold curve) with $A = 80$. Note the apparent asymmetry to the left in the observed data while the theoretical curve is symmetric around its maximum. *Inset*: The raw data and the averaged data (before collapsed) (from [46])

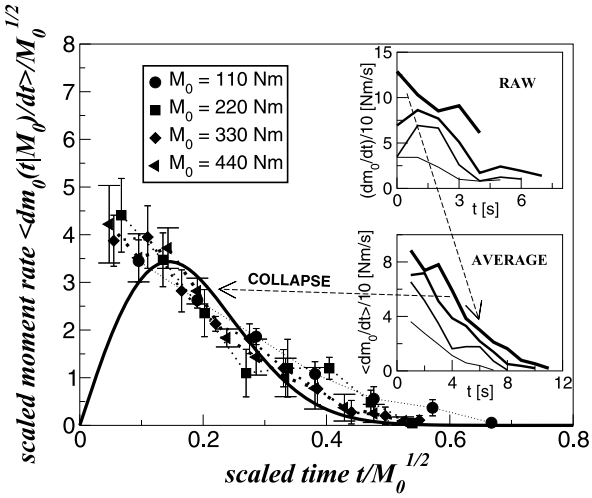

Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Figure 9

A collapse of averaged earthquake pulse shapes, $\langle dm_0(t|M_0)/dt\rangle$, with the size of the moment M_0 in Newton meters within 10% of each size given in the legend respectively. In order to obtain each collapsed moment rate shape, five to ten earthquakes were averaged for each value of M_0. The collapse was obtained using the mean field scaling relation [27]: $\langle dm_0(t|M_0)/dt\rangle/M_0^{1/2} \sim f(t/M_0^{1/2})$. In our mean field theory the universal scaling function is $f_{mf}(x) = Axe^{-Bx^2/2}$ where $x = t/M_0^{1/2}$. We plot this functional form (bold curve) with $A = 4$ and $B = 4.9$. *Inset*: The raw data and the averaged data (before collapsed) (see [46])

infinite fault this is obtained from $1 = N_2^{ex}\int_{M_2^{ex}}^{\infty} P(M)\mathrm{d}M$, and inserting Eq. (10).) If $M_2^{ex} \ll M_1$ this process can continue but will not increase the total moment substantially. If $M_2^{ex} \sim M_1$, however, the process can continue with a larger area A_2 and hence a larger M^{ex}, leading to a catastrophic runaway event. From the above exponent relations and scaling laws we obtain $B = 1/2$ and $A \sim M$, so that for any ϵ, for large enough M_1, $M_1 \gtrsim M_D \sim \epsilon^{-2}$, M_2^{ex} will be comparable to M_1 and the quake will become much larger (runaway). In the force driven infinite system for $F \lesssim F_c$, quakes of size ξ will runaway and become infinite if $\xi > \epsilon^{-1}$. Since $\xi \sim (F - F_c)^{-\nu}$ and $1/\nu = 1$, this will occur for $F_c - F < C_w\epsilon$ with some constant C_w. This result is very intuitive and justifies a posteriori the assumptions leading to it: Since on slipping, the random pinning forces, f_R in a region are reduced by order ϵ, the effective critical force F_c for continuous slip will have been reduced by order ϵ; thus if $F > F_c(\epsilon) = F_c - C_w\epsilon$, the mean velocity $\overline{v}$ will be nonzero. A similar effect can be caused by stress pulses associated with Eq. (6). By considering which of the sites in a long quake with $\alpha = 0$ can be caused to slip further by such stress pulses one finds that runaway will occur for $M \geq M_D \sim \alpha^{-4}$ for the physical case [27]. This has been checked in $d = 1$ with $\Gamma = 1$ and $\gamma = 0$, finding the predicted reduced critical force $F_c(\alpha) \sim F_c - C_p\alpha^2$ as shown in Fig. 10 [27]. These 1-d simulations also reveal a hysteretic $\overline{v}(F)$ curve in finite systems. This is expected to also occur with the model with weakening discussed above. Related higher dimensional systems are discussed in [60] and in [30].

We can now understand what should happen with either weakening or stress pulses in finite systems driven with a weak spring or with slowly moving boundaries. As the system is loaded, quakes of increasing size are observed. If the system is small enough that it cannot sustain quakes with $M > M_D(\epsilon, \alpha)$, i. e. even events within the power law scaling regime of the event size distribution, with $M \leq M_D(\epsilon, \alpha)$, are system spanning, then the behavior will not be much different from the monotonic case with $\epsilon = \alpha = 0$. In both cases there is a power law event size distribution all the way to the largest events, that are determined by the system size. This will occur if the dominant linear system size L is less than the maximum possible linear extent of an earthquake that does not become a runaway event: $L < R_D(\epsilon, \alpha) \sim M_D^{1/2} \sim \max(C_\alpha/\alpha^2, C_\epsilon/\epsilon)$ with appropriate coefficients C_α, C_ϵ, which will depend on

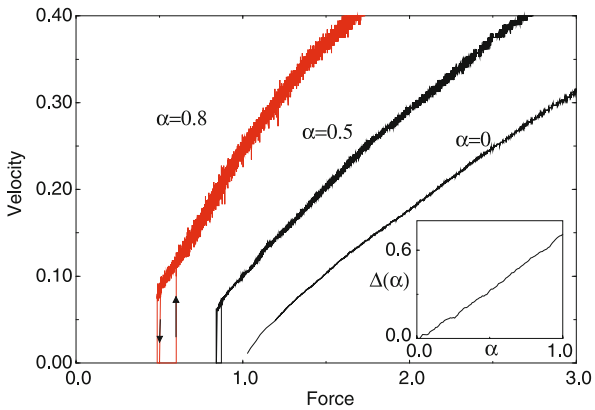

Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Figure 10

Mean velocity vs. force for one dimensional system with a non-monotonic kernel $J(x,t) = \delta(t-x)/x^2 + \alpha\delta'(t-x)/x$ for $\alpha = 0.8, 0.5, 0$. A spring or boundary loaded system will traverse the hysteresis loops in the direction indicated. *Inset*: the threshold force, $F_c^{\uparrow}(\alpha)$, on increasing the load; $\Delta_\alpha = [1 - F_c^{\uparrow}(\alpha)/F_c^{\uparrow}(\alpha = 0)]^{1/2}$ is plotted vs. α (from [27])

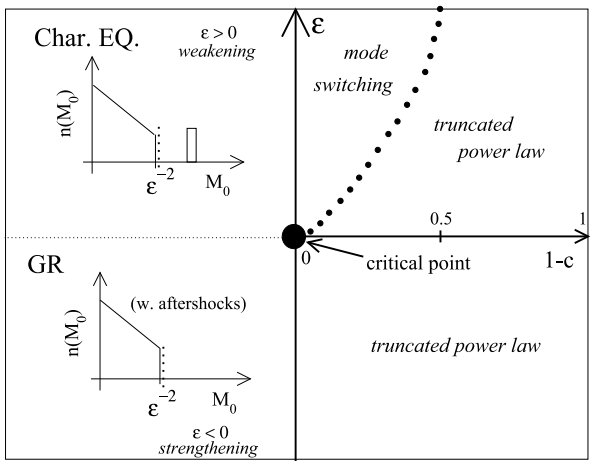

Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of, Figure 11

Phase Diagram of the BZR model, described in the text. The range $\epsilon > 0$ represents dynamic weakening, while $\epsilon < 0$ represents strengthening. The parameter $1 - c$ quantifies the deviation from stress conservation in the mean field approximation of the model

the amount of randomness in the fault. On the other hand, if $L > R_D$, quakes of size of order R_D will runaway and most of the system will slip, stopping only when the load has decreased enough to make the loading forces less than the *lower* end of the hysteresis loop in $\overline{v}(F)$ (as in Fig. 10).

Because of the tendency of regions that have already slipped to slip further, and the consequent buildup of larger stresses near the boundaries of the slipped regions, large events in systems with dynamic weakening will be much more crack like than in monotonic models, probably with $\Delta u \sim L$. Statistics of quakes with weakening, ϵ, reasonably large, but no stress pulses ($\alpha = 0$) are shown in Fig. 4 and in [2,5,6]; note the absence of quakes with intermediate moments. A typical large event in this case is shown in Fig. 5b; it appears to be crack-like.

In this section we have shown that simple models of heterogeneous faults – with the dimensionality and long-range elastic interactions properly included – can give rise to either power-law statistics of earthquake moments or a distribution of small events combined with characteristic system size events. Which behavior – or intermediate behavior – obtains is found to depend on a number of physical properties such as frictional weakening and dynamic stress transfer, analogs of which should definitely exist in real systems. In the power-law-regime the conventionally defined Gutenberg–Richter exponent $b \equiv 3B/2$ [3] is is found to be $b = 3/4$. This is close to the observed b-value of global strike-slip earthquakes at depth less than 50 km [28].

Mode–Switching In [17] the mean field approximation of infinite range elastic interaction in the BZR model with $N = LW$ (with $W \sim L$) geometrically equal cells on the fault, is used to write the local stress τ_i on cell i as

$$\begin{aligned}\tau_i &= J/N \sum_j (u_j - u_i) + K_L(vt - u_i) \\ &= J\bar{u} + K_L vt - (K_L + J)u_i \,, \end{aligned} \tag{16}$$

where u_i is the total fault offset of cell i in the horizontal (x) direction, $\bar{u} = (\sum_j u_j)/N$, J/N is the elastic coupling between cells in the mean-field approximation, and K_L is the effective loading stiffness of the bulk material surrounding the fault patch. Instead of the loading spring stiffness K_L, a conservation parameter $c \equiv J/(K_L + J)$ is introduced, which equals the fraction of the stress drop of the failing cell, that is retained in the system after the slip. There, it is shown that for the physical loading spring stiffness $K_L \sim 1/L$, one has $1 - c \sim O(1/\sqrt{N})$. A value $c < 1$ for a large system would be physically realized if the external drive is closer to the fault than its linear extent. To be precise, mean field theory only gives the correct physical scaling behavior near the critical point at zero weakening $\epsilon \to 0$ and for $c \to 1$. In [17] it is shown, however, that in a certain parameter regime for $\epsilon > 0$ and $0.5 < c < 1$ indicated in the phase diagram of Fig. 11 one finds a mode switching behavior between Gutenberg–Richter statistics and characteristic earthquake statistics. Similar mode switching behavior has also been seen in

a more realistic three dimensional model for coupled evolution of earthquakes and faults [4,41], and in numerical simulations with the BZR model that includes elastic stress transfer [76]. In the mean field BZR model, the activity switching results from episodic global reorganization of the mode of strain energy release of the fault system, reflected in a "configurational entropy" of stress states on the fault [17]. This is associated with a statistical competition between a tendency of a synchronized behavior leading to clusters of large earthquakes and the characteristic earthquake distribution, and a tendency for disordered response leading to Gutenberg–Richter type statistics without a preferred event size. Mode switching happens when these two opposite tendencies are roughly equal in strength. Some possible observational evidence for mode switching in earthquake data are discussed in [4].

Results on Aftershocks As mentioned in Sect. "Simple Models for Inhomogeneous Earthquake Faults", we associate regions off the main fault segments that are in an early deformation stage with dynamic strengthening $\epsilon < 0$. To capture basic aspects of brittle deformation on such regions in the three-dimensional volume around the main fault (Fig. 3), we change the model as follows: when any cell i slips during an earthquake, and thereby reduces its stress by $\Delta\tau_i \equiv \tau_{f,i} - \tau_{a,i}$, the failure stress $\tau_{f,j}$ of *every* cell $j = 1, \ldots, N$ is *strengthened* by an amount $|\epsilon|\Delta\tau_i/N$. Once the earthquake is complete, the failure stress of each cell is slowly lowered back to its original value. This represents in a simple way the brittle deformation that occurs during an earthquake in the off-fault regions, which are first in a strengthening regime, compared to the main fault, and then have a weakening process. The events that are triggered as the failure stresses are lowered in the weakening period are referred to as *aftershocks*. The occurrence of aftershocks in this version of the model for off-fault regions is in agreement with the observation that a large fraction of observed aftershocks typically occur in off-fault regions [70]. For this version of the model with $\epsilon < 0$, both the primary earthquakes (i. e., mainshocks) and the triggered aftershocks are distributed according to the Gutenberg–Richter distribution, up to a cutoff moment scaling as $1/\epsilon^2$. Assuming that the increased failure stress thresholds $\tau_{f,i}$ are slowly lowered with time as $\log(t)$ towards their earlier static values $\tau_{s,i}$, and that the stresses are distributed over a wide range of values, we show analytically in [46] that the temporal decay of aftershock rates at long times is proportional to $1/t$, as in the modified Omori law $\Delta N/\Delta t K/(t+c)^p$ with $p = 1$ [3,69,70], where N is the cumulative number of aftershocks, t is the time after the mainshock, and K, c, and p are empirical constants.

Remarkably, the long length scale behavior of this model can be shown [45] to be the same as the behavior of the mean field BZR model given in Eq. (16) with an added "antiferroelastic" term $(-|\epsilon|J\bar{u})$:

$$\tau_i = J\bar{u} + K_L v t - (K_L + J)u_i - |\epsilon|J\bar{u}\,. \tag{17}$$

In Eq. (17) every time a cell fails, it slips by an amount Δu_i that leads to stress loading of the other cells, lessened by $|\epsilon|J\Delta u_i/N$ compared to our original model (Eq. (16)). On the other hand, in the global strengthening model (described above) when a cell slips the failure stresses of all cells are strengthened by $|\epsilon|J\Delta u_i/N$. On long length scales the global strengthening of the failure stress has equivalent effects on the earthquake statistics as the dissipation of the redistributed stress, up to corrections of order $O(1/N)$, so the scaling behavior for large events of both models are the same. Moreover, Eq. (17) can be rewritten as:

$$\tau_i = J[1 - |\epsilon|][\bar{u} - u_i] + K_L v t - [K_L + J|\epsilon|]u_i\,. \tag{18}$$

We can now absorb $|\epsilon|$ by defining $J' = J(1 - |\epsilon|)$ and $K'_L = K_L + J|\epsilon|$. Rewriting Eq. (18) with the new definitions, and dropping the $|\epsilon|$ contribution in $[K'_L - J|\epsilon|]vt$ since $v \to 0$, we find:

$$\tau_i = J'\bar{u} + K'_L v t - (K'_L + J')u_i\,. \tag{19}$$

Therefore we recover Eq. (16) with $J \to J'$ and $K_L \to K'_L$. This amounts to changing the stress conservation parameter c (from reference [17]). For Eq. (19):

$$c = J'/(K'_L + J') = 1 - |\epsilon| \tag{20}$$

where $K_L \to 0$ since we are concerned with the adiabatic limit. We also know (from reference [17]) that the cutoff S_{cf} for the Gutenberg–Richter distribution scales as $S_{cf} \sim 1/(1-c)^2$. Thus, from Eq. (20) we find that the cutoff for Eq. (17) will scale as $\sim 1/|\epsilon|^2$.

Mapping to Single Interface Magnet Model The mean field version of the single interface magnet model with infinite range antiferromagnetic interactions is given by [22,74]:

$$\dot{h}_i(t) = J[\bar{h} - h_i(t)] + H(t) - k\bar{h} + \eta_i(h) \tag{21}$$

where $h_i(t)$ is the position of the domain wall, $H(t)$ is the external driving field, k is the coefficient of the antiferromagnetic term, and $\eta_i(h)$ is the pinning field. In the paper by Fisher et al. [27] it has been shown that the scaling behavior on long length scales resulting from Eq. (5), without the $-|\epsilon|J\bar{u}$ term, is same as that of Eq. (21) without the

antiferromagnetic term $-k\bar{h}$. Furthermore, upon inspection we see the following correspondence between the single interface magnet model (Eq. (21)), and the mean field earthquake model (Eq. (17)):

$$-k\bar{h} \Longleftrightarrow -|\epsilon|J\bar{u} \tag{22}$$

In other words, the coefficient of the antiferromagnetic term k plays the same role in the magnet model (Eq. (21)), as the coefficient of strengthening $|\epsilon|J$ does in the earthquake model (Eq. (17)).

Summary

Phase Diagram

The regimes with various statistics produced by the model are summarized by the phase diagram given in Fig. 11. The range $\epsilon > 0$ corresponds to "mature" localized faults with a weakening rheology and characteristic earthquake statistics. The value $\epsilon = 0$ corresponds to "immature" strongly inhomogeneous fault zones and fault networks with power law statistics and scale invariant rupture properties. The range $\epsilon < 0$ corresponds to the fracture and fault networks around large rupture zones, characterized by strengthening due to the creation of new structures and associated emerging aftershocks. The right side of the diagram summarizes the mean field theory results on mode switching described in Sect. "Mode–Switching". The left side of the phase diagram resembles the phase diagram for avalanches in the nucleation RFIM for magnets [65]. There, too, increasing the disorder from small to large (compared to the ferromagnetic coupling between the individual domains) drives the system from a characteristic avalanche size distribution to a truncated power law, with a disorder induced critical point separating the two regimes.

It may be surprising that the discussed simple BZR model can capture many of the essential general features of earthquake statistics (or other systems with avalanches, such as driven magnetic domain walls). This can be understood through the renormalization group [10,66], a powerful mathematical tool to coarse grain a system and extract its effective behavior on long space-time scales. Many microscopic details of a system are averaged out under coarse graining, and universal aspects of the behavior on long scales depend only on a few basic properties such as symmetries, dimensions, range of interactions, weakening/strengthening, etc. When a model correctly captures those basic features, the results provide proper predictions for statistics, critical exponents, and universal scaling functions near the critical point. Consequently, many models that are in the same universality class lead to the same statistics and exponents [10,17,27,66].

Conclusions

The phenomenology of earthquakes and avalanches in magnets exhibit a number of power law distributions and scale-invariant functions (Table 1). In search of basic model ingredients that can explain these results, we have focused on models that are rich enough to produce a diversity of observed features, while being simple enough to allow analytical predictions on long spatio-temporal scales. For the earthquake system we use the BZR model for a heterogeneous fault with threshold dynamics and long range stress-transfer interactions [2,5,6]. For the magnet system we use variants of the RFIM model with threshold dynamics and both short range and long range interactions [37,65,66,74]. In both classes of models, changes in the property disorder and dynamic effects lead to different dynamic regimes (Fig. 11). For different ranges of parameters, the earthquake model produces fractal and crack-like slip functions, power law frequency-size statistics, characteristic earthquake distribution, mode switching, and aftershocks. Similar features are found with the magnet models. We discussed two universal scaling functions of moment rates near criticality as a stronger test of the theory against observations than mere scaling exponents that have large error bars. As in magnetic systems, we find that our analysis for earthquakes provides a good overall agreement between theory and observations, but with a potential discrepancy in one particular universal scaling function for mean moment-rate shapes at fixed duration. The discrepancy has an interesting precedent in the context of avalanches in magnetic systems, and has been explained there in terms of non-universal time retardation effects due to eddy currents. Similar retardation effects may be due to triggering delays or strengthening effects that are responsible for aftershocks in earthquake faults. More observational data, in particular on small earthquakes would be needed to test some of the predictions in detail.

Future Directions

We have highlighted some interesting connections between earthquake and magnet systems with a jerky response to a slowly varying driving force. Future useful studies include analysis of factors controlling nucleation processes, transitions to instabilities and final event sizes, along with a more detailed analysis of the effects of geometrical heterogeneities in the fault structure on the statistics of earthquakes. Additional observational data, particularly for small earthquakes, are needed to test predictions for the scaling of the earthquake duration and rupture area with moment, and for accurately testing our mean field predictions for moment rate shapes. Developing analytical

corrections to the mean field earthquake models can provide additional important insights. Testing similar ideas in other systems with crackling noise would improve and deepen our understanding of universal behavior in disordered nonequilibrium systems.

Acknowledgments

We thank Daniel S. Fisher, James R. Rice, James P. Sethna, Michael B. Weissman, Deniz Ertas, Matthias Holschneider, Amit Mehta, Gert Zöller and many others for very helpful discussions. K.D. acknowledges support from the National Science Foundation, the NSF funded Materials Computation Center, and IBM. YBZ acknowledges support from the National Science Foundation, the United States Geological Survey, and the Southern California Earthquake Center.

Bibliography

Primary Literature

1. Aki K, Richards PG (2002) Quantitative Seismology, 2nd edn. University Science Books, Sausalito
2. Ben-Zion Y (1996) Stress slip and earthquakes in models of complex single-fault systems incorporating brittle and creep deformations. J Geophys Res 101:5677–5706
3. Ben-Zion Y (2003) Appendix 2, Key Formulas in Earthquake Seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part B. Academic Press, San Diego, pp 1857–1875
4. Ben-Zion Y, Dahmen K, Lyakhovsky V, Ertaş D, Agnon A (1999) Self Driven Mode Switching of Earthquake Activity on a Fault System. Earth Planet Sci Lett 172(1–2):11–21
5. Ben-Zion Y, Rice JR (1993) Earthquake failure sequences along a cellular fault zone in a three-dimensional elastic solid containing asperity and nonasperity regions. J Geophys Res 98:14109–14131
6. Ben-Zion Y, Rice JR (1995) Slip patterns and earthquake populations along different classes of faults in elastic solids. J Geophys Res 100:12959–12983
7. Ben-Zion Y, Sammis CG (2003) Characterization of Fault Zones. Pure Appl Geophys 160:677–715
8. Ben-Zion Y, Zhu L (2002) Potency-magnitude Scaling Relations for Southern California Earthquakes with $1.0 < ML < 7.0$. Geophys J Int 148:F1-F5
9. Bilek SL (2001) Earthquake rupture processes in circum-Pacific subduction zones. Ph D thesis, University of California
10. Binney JJ, Dowrick NJ, Fisher AJ, Newman MEJ (1993) The theory of critical phenomena. Oxford University Press
11. Carlson JM, Langer JS, Shaw BE (1994) Dynamics of earthquake faults. Rev Mod Phys 66:658–70, and references therein
12. Chen K, Bak P, Obukhov SP (1991) Phys Rev A 43:625
13. Cizeau P, Zapperi S, Durin G, Stanley HE (1997) Phys Rev Lett 79:4669–4672
14. Cowie PA, Vanette C, Sornette D (1993) J Geophys Res 98:21809
15. Dahmen K (1995) Hysteresis, Avalanches, and Disorder Induced Critical Scaling: A Renormalization Group Approach, Ph D Thesis, Cornell University
16. Dahmen K (2005) Nature Physics 1:13–14
17. Dahmen K, Ertaş D, Ben-Zion Y (1998) Gutenberg–Richter and Characteristic Earthquake behavior in a Simple Mean-Field Model of Heterogeneous Faults. Phys Rev E 58:1494–1501
18. Dahmen KA, Sethna JP (1996) Hysteresis, Avalanches, and Disorder Induced Critical Scaling: A Renormalization Group Approach. Phys Rev B 53:14872
19. Dieterich JH (1979) J Geophys Res 84:2161–2168
20. Dieterich JH (1981) Amer Geophys Union Monog 24:103–120
21. Durin G, Zapperi S (2000) Scaling exponents for barkhausen avalanches in polycrystalline and amorphous ferromagnets. Phys Rev Lett 84:4705–4708
22. Durin G, Zapperi S (2001) J Magn Mat 1085:242–245
23. Durin G, Zapperi S (2002) Low field hysteresis in disordered ferromagnets. Phys Rev B 65:144441
24. Ertaş D, Kardar M (1994) Critical dynamics of contact line depinning. Phys Rev E 49:R2532–5
25. Ertaş D, Kardar M (1994) Phys Rev E 49:R2532. (1994) Phys Rev Lett 73:1703
26. Fisher DS (1998) Phys Rep 301:113
27. Fisher DS, Dahmen K, Ramanathan S, Ben-Zion Y (1997) Phys Rev Lett 78:4885–4888
28. Frohlich C, Davis SD (1993) J Geophys Res 98:631
29. Gutenberg B, Richter CF (1954) Seismicity of Earth and Associated Phenomena. Princeton Univ. Press, Princeton
30. Hillers G, Mai PM, Ben-Zion Y, Ampuero J-P (2007) Statistical Properties of Seismicity Along Fault Zones at Different Evolutionary Stages. Geophys J Int 169:515 V533. doi: 10.1111/j.1365-246X.2006.03275.x
31. Houston H (2001) Influence of depth, focal mechanism, and tectonic setting on the shape and duration of earthquake source time functions. J Geophys Res 106(B6):11137–11150
32. Ji H, Robbins MO (1992) Percolative, self-affine, and faceted domain growth in random three-dimensional magnets. Phys Rev B 46:14519–27
33. Jiles D (1991) Introduction to Magnetism and Magnetic Materials. Chapman and Hall
34. Klein W, Rundle JB, Ferguson CD (1997) Scaling and Nucleation in Models of Earthquake Faults. Phys Rev Lett 78:3793–3796
35. Koiller B, Ji H, Robbins MO (1992) Fluid wetting properties and the invasion of square networks. ibid 45:7762–7
36. Koiller B, Ji H, Robbins MO (1992) Effect of disorder and lattice type on domain-wall motion in two dimensions. Phys Rev B 46:5258–65
37. Kuntz MC, Sethna JP (2000) Phys Rev B 62:11699–11708
38. Langer JS, Carlson JM, Myers CR, Shaw BE (1996) Slip complexity in dynamic models of earthquake faults. Proc Natl Acad Sci 93:3825–3829
39. Laurson L, Alava MJ (2006) $1/f$ noise and avalanche scaling in plastic deformation. Phys Rev E 74:066106
40. Lomnitz-Adler J (1993) Automaton models of seismic fracture: constraints imposed by the magnitude-frequency relation. J Geophys Res 98:17745–17756
41. Lyakhovsky V, Ben-Zion Y, and Agnon A (2001) Earthquake Cycle, Fault Zones, and Seismicity Patterns in a Rheologically Layered Litosphere. J Geophys Res 106:4103–4120
42. Marchetti MC, Middleton AA, Prellberg T (2000) Viscoelastic

Depinning of Driven Systems: Mean-Field Plastic Scallops. Phys Rev Lett 85:1104–1107
43. Martys N, Robbins MO, Cieplak M (1991) Scaling relations for interface motion through disordered media: application to two-dimensional fluid invasion. Phys Rev B 44:12294–306
44. Mayergoyz ID (1991) Mathematical Models of Hysteresis. Springer
45. Mehta AP (2005) Ph D Thesis, University of Illinois at Urbana Champaign
46. Mehta AP, Dahmen KA, Ben-Zion Y (2006) Universal mean moment rate Profiles of earthquake ruptures. Phys Rev E 73:056104
47. Mehta AP, Mills AC, Dahmen KA, Sethna JP (2002) Phy Rev E 65:46139/1–6
48. Middleton AA (1992) Phys Rev Lett 68:670
49. Miltenberger P, Sornette D, Vanette C (1993) Phys Rev Lett 71:3604
50. Myers CR, Sethna JP (1993) Collective dynamics in a model of sliding charge-density waves. I. Critical behavior. Phys Rev B 47:11171–93
51. Myers CR, Sethna JP (1993) Collective dynamics in a model of sliding charge-density waves. II. Finite-size effects. Phys Rev B 47:11194–203
52. Narayan O, Fisher DS (1992) Critical behavior of sliding charge-density waves in $4-\epsilon$ dimensions, Phys Rev B 46:11520–49
53. Narayan O, Fisher DS (1992) Dynamics of sliding charge-density waves in $4-\epsilon$ dimensions. Phys Rev Lett 68:3615–8
54. Narayan O, Fisher DS (1993) Threshold critical dynamics of driven interfaces in random media. Phys Rev B 48:7030–42
55. Narayan O, Middleton AA (1994) Avalanches and the renormalization group for pinned charge-density waves. Phys Rev B 49:244
56. Nattermann T (1997) Theory of the Random Field Ising Model. In: Young AP (ed) Spin Glasses and Random Fields. World Scientific, Singapore
57. Omori F (1894) On the aftershocks of earthquakes. J Coll Sci Imp Univ Tokyo 7:111–200
58. Perković O, Dahmen K, Sethna JP (1995) Avalanches, Barkhausen Noise, and Plain Old Criticality. Phys Rev Lett 75:4528–31
59. Perković O, Dahmen K, Sethna JP (1999) Disorder-Induced Critical Phenomena in Hysteresis: Numerical Scaling in Three and Higher Dimensions. Phys Rev B 59:6106–19
60. Ramanathan S, Fisher DS (1998) Phys Rev B 58:6026
61. Rice JR, Ben-Zion Y (1996) Slip complexity in earthquake fault models. Proc Natl Acad Sci 93:3811–3818
62. Ruff LJ, Miller AD (1994) Pure Appl Geophys 142:101
63. Schwarz JM, Fisher DS (2001) Depinning with Dynamic Stress Overshoots: Mean Field Theory. Phys Rev Lett 87:096107/1–4
64. Sethna JP (2006) Les Houches Summer School notes. Crackling Noise and Avalanches: Scaling, Critical Phenomena, and the Renormalization Group. e-print at http://xxx.lanl.gov/pdf/cond-mat/0612418
65. Sethna JP, Dahmen K, Kartha S, Krumhansl JA, Roberts BW, Shore JD (1993) Hysteresis and Hierarchies: Dynamics of Disorder Driven First Order Phase Transformations. Phys Rev Lett 70:3347
66. Sethna JP, Dahmen KA, Myers CR (2001) Nature 410:242–250
67. Spasojevic D, Bukvic S, Milosevic S, Stanley HE (1996) Barkhausen noise: Elementary signals. power laws, and scaling relations. Phys Rev E 54:2531–2546
68. Travesset A, White RA, Dahmen KA (2002) Phys Rev B 66:024430
69. Utsu T (2002) Statistical features of seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part A. pp 719–732
70. Utsu Y, Ogata Y, Matsu'uara RS (1995) The centenary of the Omori Formula for a decay law of aftershock activity. J Phys Earth 43:1–33
71. Vere-Jones D (1976) A branching model for crack propagation. Pure Appl Geophys 114(4):711–726
72. Zapperi S, Castellano C, Calaiori F, Durin G (2005) Signature of effective mass in crackling-noise asymmetry. Nature Phys 1:46–49
73. Zapperi S, Cizeau P, Durin G, Stanley HE (1998) Dynamics of a ferromagnetic domain wall: Avalanches, depinning transition, and the Barkhausen effect. Phys Rev B 58:6353–66
74. Zapperi S, Cizeau P, Durin G, Stanley HE (1998) Phys Rev B 58(10):6353–6366
75. Zöller G, Hainzl S, Ben-Zion Y, Holschneider M (2009) Critical states of seismicity: From models to practical seismic hazard estimates. In: Encyclopedia of Complexity and System Science
76. Zöller G, Holschneider M, Ben-Zion Y (2004) Quasi-static and Quasi-dynamic modeling of earthquake failure at intermediate scales. Pure Appl Geophys 161:2103–2118
77. Zöller G, Holschneider M, Ben-Zion Y (2005) The role of heterogeneities as a tuning parameter of earthquake dynamics. Pure Appl Geophys 162:1027 V1049. doi: 10.1007s00024-004-2660-9

Joinings in Ergodic Theory

Thierry de la Rue
Laboratoire de Mathématiques Raphaël Salem,
CNRS – Université de Rouen,
Saint Étienne du Rouvray, France

Article Outline

Glossary
Definition of the Subject
Introduction
Joinings of Two or More Dynamical Systems
Self-Joinings
Some Applications and Future Directions
Bibliography

Glossary

Disjoint measure-preserving systems The two measure-preserving dynamical systems $(X, \mathcal{A}, \mu, T)$ and $(Y, \mathcal{B}, \nu, S)$ are said to be *disjoint* if their only joining is the product measure $\mu \otimes \nu$.

Joining Let I be a finite or countable set, and for each $i \in I$, let $(X_i, \mathcal{A}_i, \mu_i, T_i)$ be a measure-preserving dynamical system. A *joining* of these systems is a probability measure on the Cartesian product $\prod_{i\in I} X_i$, which has the μ_i's as marginals, and which is invariant under the product transformation $\bigotimes_{i\in I} T_i$.

Marginal of a probability measure on a product space Let λ be a probability measure on the Cartesian product of a finite or countable collection of measurable spaces $\left(\prod_{i\in I} X_i, \bigotimes_{i\in I} \mathcal{A}_i\right)$, and let $J = \{j_1, \ldots, j_k\}$ be a finite subset of I. The *k-fold marginal* of λ on $X_{j_1}, \ldots, X_{j_k}$ is the probability measure μ defined by:

$$\forall A_1 \in \mathcal{A}_{j_1}, \ldots, A_k \in \mathcal{A}_{j_k}\,,$$
$$\mu(A_1 \times \cdots \times A_k) := \lambda\left(A_1 \times \cdots \times A_k \times \prod_{i\in I\setminus J} X_i\right).$$

Markov intertwining Let $(X, \mathcal{A}, \mu, T)$ and $(Y, \mathcal{B}, \nu, S)$ be two measure-preserving dynamical systems. We call *Markov intertwining* of T and S any operator $P\colon L^2(X,\mu) \to L^2(Y,\nu)$ enjoying the following properties:

- $PU_T = U_S P$, where U_T and U_S are the unitary operators on $L^2(X,\mu)$ and $L^2(Y,\nu)$ associated respectively to T and S (i. e. $U_T f(x) = f(Tx)$, and $U_S g(y) = g(Sy)$).
- $P\mathbb{1}_X = \mathbb{1}_Y$,
- $f \geq 0$ implies $Pf \geq 0$, and $g \geq 0$ implies $P^* g \geq 0$, where P^* is the adjoint operator of P.

Minimal self-joinings Let $k \geq 2$ be an integer. The ergodic measure-preserving dynamical system T has *k-fold minimal self-joinings* if, for any ergodic joining λ of k copies of T, we can partition the set $\{1, \ldots, k\}$ of coordinates into subsets $J_1, \ldots, J_\ell$ such that

1. For j_1 and j_2 belonging to the same J_i, the marginal of λ on the coordinates j_1 and j_2 is supported on the graph of T^n for some integer n (depending on j_1 and j_2);
2. For $j_1 \in J_1, \ldots, j_\ell \in J_\ell$, the coordinates $j_1, \ldots, j_\ell$ are independent.

We say that T has *minimal self-joinings* if T has k-fold minimal self-joinings for every $k \geq 2$.

Off-diagonal self-joinings Let $(X, \mathcal{A}, \mu, T)$ be a measure-preserving dynamical system, and S be an invertible measure-preserving transformation of $(X, \mathcal{A}, \mu)$ commuting with T. Then the probability measure Δ_S defined on $X \times X$ by

$$\Delta_S(A \times B) := \mu(A \cap S^{-1}B) \tag{1}$$

is a 2-fold self-joining of T supported on the graph of S. We call it an *off-diagonal self-joining of T*.

Process in a measure-preserving dynamical systems Let $(X, \mathcal{A}, \mu, T)$ be a measure-preserving dynamical system, and let $(E, \mathcal{B}(E))$ be a measurable space (which may be a finite or countable set, or $\mathbb{R}^d$, or $\mathbb{C}^d \ldots$). For any E-valued random variable ξ defined on the probability space $(X, \mathcal{A}, \mu)$, we can consider the stochastic process $(\xi_i)_{i\in\mathbb{Z}}$ defined by

$$\xi_i := \xi \circ T^i\,.$$

Since T preserves the probability measure μ, $(\xi_i)_{i\in\mathbb{Z}}$ is a *stationary process*: For any ℓ and n, the distribution of $(\xi_0, \ldots, \xi_\ell)$ is the same as the probability distribution of $(\xi_n, \ldots, \xi_{n+\ell})$.

Self-joining Let T be a measure-preserving dynamical system. A *self-joining* of T is a joining of a family $(X_i, \mathcal{A}_i, \mu_i, T_i)_{i\in I}$ of systems where each T_i is a copy of T. If I is finite and has cardinal k, we speak of a *k-fold* self-joining of T.

Simplicity For $k \geq 2$, we say that the ergodic measure-preserving dynamical system T is *k-fold simple* if, for any ergodic joining λ of k copies of T, we can partition the set $\{1, \ldots, k\}$ of coordinates into subsets $J_1, \ldots, J_\ell$ such that

1. for j_1 and j_2 belonging to the same J_i, the marginal of λ on the coordinates j_1 and j_2 is supported on the graph of some $S \in C(T)$ (depending on j_1 and j_2);
2. for $j_1 \in J_1, \ldots, j_\ell \in J_\ell$, the coordinates $j_1, \ldots, j_\ell$ are independent.

We say that T is *simple* if T is k-fold simple for every $k \geq 2$.

Definition of the Subject

The word *joining* can be considered as the counterpart in ergodic theory of the notion of *coupling* in probability theory (see e. g. [48]): Given two or more processes defined on different spaces, what are the possibilities of embedding them together in the same space? There always exists the solution of making them independent of each other, but interesting cases arise when we can do this in other ways. The notion of joining originates in ergodic theory from pioneering works of H. Furstenberg [14], who introduced the fundamental notion of *disjointness*, and D.J. Rudolph, who laid the basis of joining theory in his article on minimal self-joinings [41]. It has today become an essential tool

in the classification of measure-preserving dynamical systems and in the study of their intrinsic properties.

Introduction

A central question in ergodic theory is to tell when two measure-preserving dynamical systems are essentially the same, i. e. when they are *isomorphic*. When this is not the case, a finer analysis consists of asking what these two systems could share in common: For example, do there exist stationary processes which can be observed in both systems? This latter question can also be asked in the following equivalent way: Do these two systems have a *common factor*? The arithmetical flavor of this question is not fortuitous: There are deep analogies between the arithmetic of integers and the classification of measure-preserving dynamical systems, and these analogies were at the starting point of the study of joinings in ergodic theory.

In the seminal paper [14] which introduced the concept of joinings in ergodic theory, H. Furstenberg observed that two operations can be done with dynamical systems: We can consider the *product* of two dynamical systems, and we can also take a *factor* of a given system. Like the multiplication of integers, the product of dynamical systems is commutative, associative, it possesses a neutral element (the trivial single-point system), and the systems S and T are both factors of their product $S \times T$. It was then natural to introduce the property for two measure-preserving systems to be *relatively prime*. As far as integers are concerned, there are two equivalent ways of characterizing the relative primeness: First, the integers a and b are *relatively prime* if their unique positive common factor is 1. Second, a and b are *relatively prime* if, each time both a and b are factors of an integer c, their product ab is also a factor of c. It is a well-known theorem in number theory that these two properties are equivalent, but this was not clear for their analog in ergodic theory. Furstenberg reckoned that the second way of defining relative primeness was the most interesting property in ergodic theory, and called it *disjointness* of measure-preserving systems (we will discuss precisely in Subsect. "From Disjointness to Isomorphy" what the correct analog is in the setting of ergodic theory). He also asked whether the non-existence of a non-trivial common factor between two systems was equivalent to their disjointness. He was able to prove that disjointness implies the impossibility of a non-trivial common factor, but not the converse. And in fact, the converse turns out to be false: In 1979, D.J. Rudolph exhibited a counterexample in his paper introducing the important notion of *minimal self-joinings*. The relationships between disjointness and the lack of common factor will be presented in details in Sect. "Joinings and Factors".

Given two measure-preserving dynamical systems S and T, the study of their disjointness naturally leads one to consider all the possible ways these two systems can be both seen as factors of a third system. As we shall see, this is precisely the study of their *joinings*. The concept of joining turns out to be related to many important questions in ergodic theory, and a large number of deep results can be stated and proved inside the theory of joinings. For example, the fact that the dynamical systems S and T are isomorphic is equivalent to the existence of a special joining between S and T, and this can be used to give a joining proof of Krieger's finite generator theorem, as well as Ornstein's isomorphism theorem (see Sect. "Joinings Proofs of Ornstein's and Krieger's Theorems"). As it already appears in Furstenberg's article, joinings provide a powerful tool in the classification of measure-preserving dynamical systems: Many classes of systems can be characterized in terms of their disjointness with other systems. Joinings are also strongly connected with difficult questions arising in the study of the convergence almost everywhere of non-conventional averages (see Sect. "Joinings and Pointwise Convergence").

Amazingly, a situation in which the study of joinings leads to the most interesting results consists of considering two or more *identical* systems. We then speak of the *self-joinings* of the dynamical system T. Again, the study of self-joinings is closely related to many ergodic properties of the system: Its mixing properties, the structure of its factors, the transformations which commute with T, and so on... We already mentioned minimal self-joinings, and we will see in Sect. "Minimal Self-Joinings" how this property may be used to get many interesting examples, such as a transformation with no root, or a process with no non-trivial factor. In the same section we will also discuss a very interesting generalization of minimal self-joinings: the property of being *simple*.

The range of applications of joinings in ergodic theory is very large; only some of them will be given in Sect. "Some Applications and Future Directions": The use of joinings in proving Krieger's and Ornstein's theorems, the links between joinings and some questions of pointwise convergence, and the strong connections between the study of self-joinings and Rohlin's famous question on multifold mixing, which was first posed in 1949 [39].

Joinings of Two or More Dynamical Systems

In the following, we are given a finite or countable family $(X_i, \mathcal{A}_i, \mu_i, T_i)_{i \in I}$ of measure-preserving dynamical

systems: T_i is an invertible measure-preserving transformation of the standard Borel probability space $(X_i, \mathcal{A}_i, \mu_i)$. When it is not ambiguous, we shall often use the symbol T_i to denote both the transformation and the system.

A *joining* λ of the T_i's (see the definition in the Glossary) defines a new measure-preserving dynamical system: The product transformation

$$\bigotimes_{i\in I} T_i : (x_i)_{i\in I} \longmapsto (T_i x_i)_{i\in I}$$

acting on the Cartesian product $\prod_{i\in I} X_i$, and preserving the probability measure λ. We will denote this big system by $\left(\bigotimes_{i\in I} T_i\right)_\lambda$. Since all marginals of λ are given by the original probabilities μ_i, observing only the coordinate i in the big system is the same as observing only the system T_i. Thus, each system T_i is a factor of $\left(\bigotimes_{i\in I} T_i\right)_\lambda$, via the homomorphism π_i which maps any point in the Cartesian product to its ith coordinate.

Conversely, if we are given a measure-preserving dynamical system (Z, C, ρ, R) admitting each T_i as a factor via some homomorphism $\varphi_i : Z \to X_i$, then we can construct the map $\varphi : Z \to \prod_{i\in I} X_i$ sending z to $(\varphi_i(z))_{i\in I}$. We can easily check that the image of the probability measure ρ is then a joining of the T_i's.

Therefore, studying the joinings of a family of measure-preserving dynamical system amounts to studying all the possible ways these systems can be seen together as factors in another big system.

The Set of Joinings

The set of all joinings of the T_i's will be denoted by $J(T_i, i \in I)$. Before anything else, we have to observe that this set is never empty. Indeed, whatever the systems are, the product measure $\bigotimes_{i\in I} \mu_i$ always belongs to this set. Note also that any convex combination of joinings is a joining: $J(T_i, i \in I)$ is a convex set.

The set of joinings is turned into a compact metrizable space, equipped with the topology defined by the following notion of convergence: $\lambda_n \xrightarrow[n\to\infty]{} \underline{\lambda}$ if and only if, for all families of measurable subsets $(A_i)_{i\in I} \in \prod_{i\in I} \mathcal{A}_i$, finitely many of them being different from X_i, we have

$$\lambda_n\left(\prod_{i\in I} A_i\right) \xrightarrow[n\to\infty]{} \underline{\lambda}\left(\prod_{i\in I} A_i\right). \tag{2}$$

We can easily construct a distance defining this topology by observing that it is enough to check (2) when each of the A_i's is chosen in some countable algebra C_i generating the σ-algebra $\mathcal{A}_i$. We can also point out that, when the X_i's are themselves compact metric spaces, this topology on the set of joinings is nothing but the restriction to $J(T_i, i \in I)$ of the usual weak* topology.

It is particularly interesting to study *ergodic* joinings of the T_i's, whose set will be denoted by $J_e(T_i, i \in I)$. Since any factor of an ergodic system is itself ergodic, a necessary condition for $J_e(T_i, i \in I)$ not to be empty is that all the T_i's be themselves ergodic. Conversely, if all the T_i's are ergodic, we can prove by considering the ergodic decomposition of the product measure $\bigotimes_{i\in I} \mu_i$ that ergodic joinings do exist: Any ergodic measure appearing in the ergodic decomposition of some joining has to be itself a joining. This result can also be stated in the following way:

Proposition 1 *If all the T_i's are ergodic, the set of their ergodic joinings is the set of extremal points in the compact convex set $J(T_i, i \in I)$.*

From Disjointness to Isomorphy

In this section, as in many others in this article, we are focusing on the case where our family of dynamical systems is reduced to two systems. We will then rather call them S and T, standing for $(Y, \mathcal{B}, \nu, S)$ and $(X, \mathcal{A}, \mu, T)$. We are interested here in two extremal cases for the set of joinings $J(T, S)$. The first one occurs when the two systems are as far as possible from each other: They have nothing to share in common, and therefore their set of joinings is reduced to the singleton $\{\mu \otimes \nu\}$: This is called the *disjointness* of S and T. The second one arises when the two systems are isomorphic, and we will see how this property shows through $J(T, S)$.

Disjointness Many situations where disjointness arises were already given by Furstenberg in [14]. Particularly interesting is the fact that classes of dynamical systems can be characterized through disjointness properties. We list here some of the main examples of disjoint classes of measure-preserving systems.

Theorem 2

1. *T is ergodic if and only if it is disjoint from every identity map.*
2. *T is weakly mixing if and only if it is disjoint from any rotation on the circle.*
3. *T has zero entropy if and only if it is disjoint from any Bernoulli shift.*
4. *T is a K-system if and only if it is disjoint from any zero-entropy system.*

The first result is the easiest, but is quite important, in particular when it is stated in the following form: If λ is a join-

ing of T and S, with T ergodic, and if λ is invariant by $T \times \mathrm{Id}$, then $\lambda = \mu \otimes \nu$.

The second, third and fourth results were originally proved by Furstenberg. They can also be seen as corollaries of the theorems presented in Sect. "Joinings and Factors", linking the non-disjointness property with the existence of a particular factor.

Both the first and the second results can be derived from the next theorem, giving a general spectral condition in which disjointness arises. The proof of this theorem can be found in [49]. It is a direct consequence of the fact that, if f and g are square-integrable functions in a given dynamical system, and if their spectral measures are mutually singular, then f and g are orthogonal in L^2.

Theorem 3 *If the reduced maximum spectral types of T and S are mutually singular, then T and S are disjoint.*

As we already said in the introduction, disjointness was recognized by Furstenberg as the most pertinent way to define the analog of the arithmetic property "a and b are relatively prime" in the context of measure-preserving dynamical systems. We must however point out that the statement

(i) S and T are disjoint

is, in general, strictly stronger than the straightforward translation of the arithmetic property:

(ii) Each time both S and T appear as factors in a third dynamical system, then their product $S \times T$ also appears as a factor in this system.

Indeed, contrary to the situation in ordinary arithmetic, there exist non-trivial dynamical systems T which are isomorphic to $T \times T$: For example, this is the case when T is the product of countably many copies of a single non-trivial system. Now, if T is such a system and if we take $S = T$, then S and T do not satisfy statement (i): A non-trivial system is never disjoint from itself, as we will see in the next Section. However they obviously satisfy the statement (ii).

A correct translation of the arithmetic property is the following: S and T are disjoint if and only if, each time T and S appear as factors in some dynamical system through the respective homomorphisms π_T and π_S, $T \times S$ also appears as a factor through a homomorphism $\pi_{T\times S}$ such that $\pi_X \circ \pi_{T\times S} = \pi_T$ and $\pi_Y \circ \pi_{T\times S} = \pi_S$, where π_X and π_Y are the projections on the coordinates in the Cartesian product $X \times Y$ (see the diagram below).

$$\begin{tikzcd}
 & Z \arrow[dl, "\pi_T"'] \arrow[d, "\pi_{T\times S}"] \arrow[dr, "\pi_S"] & \\
X & X \times Y \arrow[l, "\pi_X"] \arrow[r, "\pi_Y"'] & Y
\end{tikzcd}$$

Joinings and Isomorphism We first introduce some notations: For any probability measure λ on a measurable space, let '$A \overset{\lambda}{=} B$' stand for '$\lambda(A\Delta B) = 0$'. Similarly, if $\mathcal{C}$ and $\mathcal{D}$ are σ-algebras of measurable sets, we write '$\mathcal{C} \overset{\lambda}{\subset} \mathcal{D}$' if, for any $C \in \mathcal{C}$, we can find some $D \in \mathcal{D}$ such that $C \overset{\lambda}{=} D$, and by '$\mathcal{C} \overset{\lambda}{=} \mathcal{D}$' we naturally mean that both $\mathcal{C} \overset{\lambda}{\subset} \mathcal{D}$ and $\mathcal{D} \overset{\lambda}{\subset} \mathcal{C}$ hold.

Let us assume now that our two systems S and T are isomorphic: This means that we can find some measurable one-to-one map $\varphi\colon X \to Y$, with $T(\mu) = \nu$, and $\varphi \circ T = S \circ \varphi$. With such a φ, we construct the measurable map $\psi : X \to X \times Y$ by setting

$$\psi(x) := \big((x, \varphi(x)\big) .$$

Let Δ_φ be the image measure of μ by ψ. This measure is supported on the graph of φ, and is also characterized by

$$\forall A \in \mathcal{A}\,, \forall B \in \mathcal{B}\,, \Delta_\varphi(A \times B) = \mu(A \cap \varphi^{-1}B) . \quad (3)$$

We can easily check that, φ being an isomorphism of T and S, Δ_φ is a joining of T and S. And this joining satisfies very special properties:

- For any measurable $A \subset X$, $A \times Y \overset{\Delta_\varphi}{=} X \times \varphi(A)$;
- Conversely, for any measurable $B \subset Y$, $X \times B \overset{\Delta_\varphi}{=} \varphi^{-1}(B) \times Y$.

Thus, in the case where S and T are isomorphic, we can find a special joining of S and T, which is supported on the graph of an isomorphism, and which identifies the two σ-algebras generated by the two coordinates. What is remarkable is that the converse is true: The existence of an isomorphism between S and T is characterized by the existence of such a joining, and we have the following theorem:

Theorem 4 *The measure-preserving dynamical systems S and T are isomorphic if and only if there exists a joining λ of S and T such that*

$$\{X, \emptyset\} \otimes \mathcal{B} \overset{\lambda}{=} \mathcal{A} \otimes \{Y, \emptyset\} . \quad (4)$$

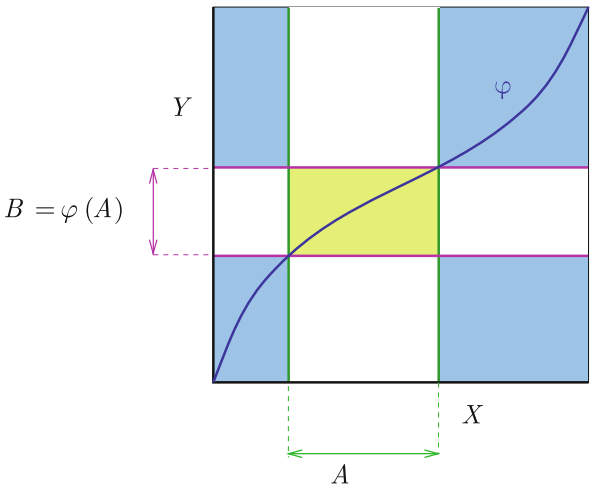

Joinings in Ergodic Theory, Figure 1
The joining Δ_φ identifies the sets $A \times Y$ and $X \times \varphi(A)$

When λ is such a joining, it is supported on the graph of an isomorphism of T and S, and both systems are isomorphic to the joint system $(T \otimes S)_\lambda$.

This theorem finds nice applications in the proof of classical isomorphism results. For example, it can be used to prove that two discrete-spectrum systems which are spectrally isomorphic are isomorphic (see [49] or [6]). We will also see in Sect.“ Joinings Proofs of Ornstein's and Krieger's Theorems” how it can be applied in the proofs of Krieger's and Ornstein's deep theorems.

Consider now the case were T and S are no longer isomorphic, but where S is only a factor of T. Then we have a factor map $\pi : X \to Y$ which has the same properties as an isomorphism φ, except that it is not one-to-one (π is only onto). The measure Δ_π, constructed in the same way as Δ_φ, is still a joining supported on the graph of π, but it does not identify the two σ-algebras generated by the two coordinates anymore: Instead of condition (4), Δ_π only satisfies the weaker one:

$$\{X, \emptyset\} \otimes \mathcal{B} \overset{\Delta_\pi}{\subset} \mathcal{A} \otimes \{Y, \emptyset\} . \tag{5}$$

The existence of a joining satisfying (5) is a criterion for S being a factor of T.

For more details on the results stated in this section, we refer the reader to [6].

Joinings and Factors

The purpose of this section is to investigate the relationships between the disjointness of two systems S and T, and the lack of a common factor. The crucial fact which was pointed out by Furstenberg is that the existence of a common factor enables one to construct a very special joining of S and T: The *relatively independent joining over this factor*.

Let us assume that our systems S and T share a common factor (Z, C, ρ, R), which means that we have measurable onto maps $\pi_X : X \to Z$ and $\pi_Y : Y \to Z$, respectively sending μ and ν to ρ, and satisfying $\pi_X \circ T = R \circ \pi_X$ and $\pi_Y \circ S = R \circ \pi_Y$. We can then consider the joinings supported on their graphs $\Delta_{\pi_X} \in J(T, R)$ and $\Delta_{\pi_Y} \in J(S, R)$, as defined in the preceding section. Next, we construct a joining λ of the three systems S, T and R. Heuristically, λ is the probability distribution of the triple (x, y, z) when we first pick z according to the probability distribution ρ, then x and y according to their conditional distribution knowing z in the respective joinings Δ_{π_X} and Δ_{π_Y}, but independently of each other. More precisely, λ is defined by setting, for all $A \in \mathcal{A}$, $B \in \mathcal{B}$ and $C \in C$

$$\lambda(A \times B \times C) := \int_C \mathbb{E}_{\Delta_{\pi_X}}[\mathbb{1}_{x \in A}|z]\mathbb{E}_{\Delta_{\pi_Y}}[\mathbb{1}_{y \in B}|z]\,\mathrm{d}\rho(z) . \tag{6}$$

Observe that the two-fold marginals of λ on $X \times Z$ and $Y \times Z$ are respectively Δ_{π_X} and Δ_{π_Y}, which means that we have $z = \pi_X(x) = \pi_Y(y)$ λ-almost surely. In other words, we have identified in the two systems T and S the projections on their common factor R. The two-fold marginal of λ on $X \times Y$ is itself a joining of T and S, which we call the *relatively independent joining* over the common factor R. This joining will be denoted by $\mu \otimes_R \nu$. (Be careful: The projections π_X and π_Y are hidden in this notation, but we have to know them to define this joining.) From (6), we immediately get the formula defining $\mu \otimes_R \nu$:

$$\forall A \in \mathcal{A}\,, \forall B \in \mathcal{B}\,,\ \mu \otimes_R \nu(A \times B) := \int_Z \mathbb{E}_{\Delta_{\pi_X}}[\mathbb{1}_{x \in A}|z]\mathbb{E}_{\Delta_{\pi_Y}}[\mathbb{1}_{y \in B}|z]\,\mathrm{d}\rho(z) . \tag{7}$$

This definition of the relatively independent joining over a common factor can easily be extended to a finite or countable family of systems sharing the same common factor.

Note that $\mu \otimes_R \nu$ coincides with the product measure $\mu \otimes \nu$ if and only if the common factor is the trivial one-point system. We therefore get the following result:

Theorem 5 *If S and T have a non-trivial common factor, then these systems are not disjoint.*

As we already said in the introduction, Rudolph exhibited in [41] a counterexample showing that the converse is not

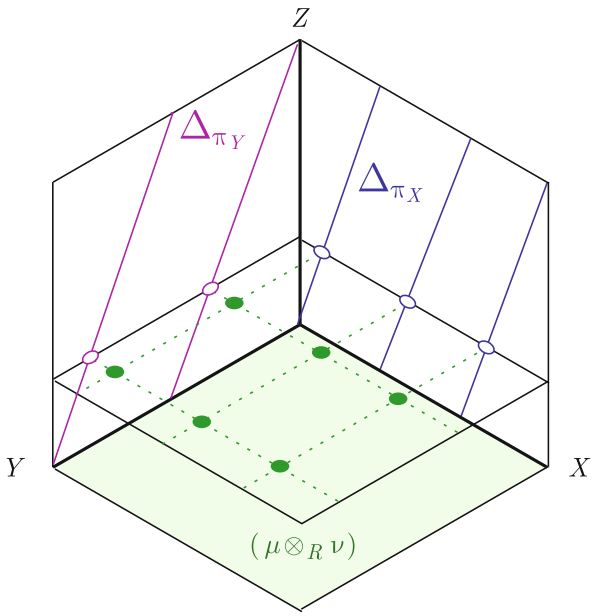

Joinings in Ergodic Theory, Figure 2
The relatively independent joining $\mu \otimes_R \nu$ and its disintegration over Z

true. There exists, however, an important result, which was published in [20,29] allowing us to derive some information on factors from the non-disjointness of two systems.

Theorem 6 *If T and S are not disjoint, then S has a non-trivial common factor with some joining of a countable family of copies of T.*

This result leads to the introduction of a special class of factors when some dynamical system T is given: For any other dynamical system S, call *T-factor* of S any common factor of S with a joining of countably many copies of T. If (Z, C, ρ, R) is a T-factor of S and $\pi : Y \to Z$ is a factor map, we say that the σ-algebra $\pi^{-1}(C)$ is a *T-factor σ-algebra* of S. Another way to state Theorem 6 is then the following: If S and T are not disjoint, then S has a non-trivial T-factor. In fact, an even more precise result can be derived from the proof of Theorem 6: For any joining λ of S and T, for any bounded measurable function f on X, the factor σ-algebra of S generated by the function $\mathbb{E}_\lambda[f(x)|y]$ is a T-factor σ-algebra of S.

With the notion of T-factor, Theorem 6 has been extended in [31] in the following way, showing the existence of a special T-factor σ-algebra of S concentrating anything in S which could lead to a non-trivial joining between T and S.

Theorem 7 *Given two measure-preserving dynamical systems $(X, \mathcal{A}, \mu, T)$ and $(Y, \mathcal{B}, \nu, S)$, there always exists a maximum T-factor σ-algebra of S, denoted by $\mathcal{F}_T$.*

Under any joining λ of T and S, the σ-algebras $\mathcal{A} \otimes \{\emptyset, Y\}$ and $\{\emptyset, X\} \otimes \mathcal{B}$ are independent conditionally to the σ-algebra $\{\emptyset, X\} \otimes \mathcal{F}_T$.

Theorem 6 gives a powerful tool to prove some important disjointness results, such as those stated in Theorem 2. These results involve properties of dynamical systems which are stable under the operations of taking joinings and factors. We will call these properties *stable properties*. This is, for example, the case of the zero-entropy property: We know that any factor of a zero-entropy system still has zero entropy, and that any joining of zero-entropy systems also has zero entropy. In other words, T has zero entropy implies that any T-factor has zero entropy. But the property of S being a K-system is precisely characterized by the fact that any non-trivial factor of S has positive entropy. Hence a K-system S cannot have a non-trivial T-factor if T has zero entropy, and is therefore disjoint from T. The converse is a consequence of Theorem 5: If S is not a K-system, then it possesses a non-trivial zero-entropy factor, and therefore there exists some zero-entropy system from which it is not disjoint.

The same argument also applies to the disjointness of discrete-spectrum systems with weakly mixing systems, since a discrete spectrum is a stable property, and weakly mixing systems are characterized by the fact that they do not have any discrete-spectrum factor.

Markov Intertwinings and Composition of Joinings

There is another way of defining joinings of two measure-preserving dynamical systems involving operators on L^2 spaces, mainly put to light by Ryzhikov (see [47]): Observe that for any joining $\lambda \in J(T, S)$, we can consider the operator $P_\lambda : L^2(X, \mu) \to L^2(Y, \nu)$ defined by

$$P_\lambda(f) := \mathbb{E}_\lambda\left[f(x)|y\right] .$$

It is easily checked that P_λ is a *Markov intertwining* of T and S. Conversely, given any Markov intertwining P of T and S, it can be shown that the measure λ_P defined on $X \times Y$ by

$$\lambda_P(A \times B) := \langle P\mathbb{1}_A, \mathbb{1}_B\rangle_{L^2(Y,\nu)}$$

is a joining of T and S.

This reformulation of the notion of joining is useful when joinings are studied in connection with spectral properties of the transformations (see e. g. [22]). It also provides us with a convenient setting to introduce the *composition* of joinings: If we are given three dynamical systems $(X, \mathcal{A}, \mu, T)$, $(Y, \mathcal{B}, \nu, S)$ and (Z, C, ρ, R), a joining $\lambda \in J(T, S)$ and a joining $\lambda' \in J(S, R)$, the composition of the Markov intertwinings P_λ and $P_{\lambda'}$ is easily seen

to give a third Markov intertwining, which itself corresponds to a joining of T and R denoted by $\lambda \circ \lambda'$. When $R = S = T$, i. e. when we are speaking of *2-fold self-joinings* of a single system T (*cf.* next section), this operation turns $J(T,T) = J^2(T)$ into a semigroup. Ahn and Lemańczyk [1] have shown that the subset $J^2_e(T)$ of ergodic two-fold self-joinings is a sub-semigroup if and only if T is semisimple (see Sect. "Simple Systems").

Self-Joinings

We now turn to the case where the measure-preserving dynamical systems we want to join together are all copies of a single system T. For $k \geq 2$, any joining of k copies of T is called a *k-fold self-joining of T*. We denote by $J^k(T)$ the set of all k-fold self-joinings of T, and by $J^k_e(T)$ the subset of *ergodic* k-fold self-joinings.

Self-Joinings and Commuting Transformations

As soon as T is not the trivial single-point system, T is never disjoint from itself: Since T is obviously isomorphic to itself, we can always find a two-fold self-joining of T which is not the product measure by considering self-joinings supported on graphs of isomorphisms (see Sect. "Joinings and Isomorphism"). The simplest of them is obtained by taking the identity map as an isomorphism, and we get that $J^2(T)$ always contains the *diagonal measure* $\Delta_0 := \Delta_{\mathrm{Id}}$.

In general, an isomorphism of T with itself is an invertible measure-preserving transformation S of $(X, \mathcal{A}, \mu)$ which commutes with T. We call *commutant* of T the set of all such transformations (it is a subgroup of the group of automorphisms of $(X, \mathcal{A}, \mu)$), and denote it by $C(T)$. It always contains, at least, all the powers T^n, $n \in \mathbb{Z}$.

Each element S of $C(T)$ gives rise to a two-fold self-joining Δ_S supported on the graph of S. Such self-joinings are called *off-diagonal self-joinings*. They also belong to $J^k_e(T)$ if T is ergodic.

It follows that properties of the commutant of an ergodic T can be seen in its ergodic joinings. As an example of application, we can cite Ryzhikov's proof of King's *weak closure theorem* for rank-one transformations[1]. Rank-one measure-preserving transformations form a very important class of zero-entropy, ergodic measure-preserving transformations. They have many remarkable properties, among which the fact that their commutant is reduced to the weak limits of powers of T. In other words, if T is rank one, for any $S \in C(T)$ there exists a subsequence of integers (n_k) such that,

$$\forall A \in \mathcal{A}\,,\ \mu\left(T^{-n_k} A \Delta S^{-1} A\right) \xrightarrow[k\to\infty]{} 0\,. \tag{8}$$

King proved this result in 1986 [26], using a very intricate coding argument. Observing that (8) was equivalent to the convergence, in $J^2(T)$, of $\Delta_{T^{n_k}}$ to Δ_S, Ryzhikov showed in [45] that King's theorem could be seen as a consequence of the following general result concerning two-fold self-joinings of rank-one systems:

Theorem 8 *Let T be a rank-one measure-preserving transformation, and $\lambda \in J^2_e(T)$. Then there exist $t \geq 1/2$, a subsequence of integers (n_k) and another two-fold self-joining λ' of T such that*

$$\Delta_{T^{n_k}} \xrightarrow[k\to\infty]{} t\lambda + (1-t)\lambda'\,.$$

Minimal Self-Joinings

For any measure-preserving dynamical system T, the set of two-fold self-joinings of T contains at least the product measure $\mu \otimes \mu$, the off-diagonal joinings Δ_{T^n} for each $n \in \mathbb{Z}$, and any convex combination of these. Rudolph [41] discovered in 1979 that we can find systems for which there are no other two-fold self-joinings than these obvious ones. When this is the case, we say that T has *2-fold-minimal self-joinings*, or for short: $T \in \mathrm{MSJ}(2)$. It can be shown (see e. g. [42]) that, as soon as the underlying probability space is not atomic (which we henceforth assume), two-fold minimal self-joinings implies that T is weakly mixing, and therefore that $\mu \otimes \mu$ and Δ_{T^n}, $n \in \mathbb{Z}$, are *ergodic* two-fold self-joinings of T. That is why two-fold minimal self-joinings are often defined by the following:

$$T \in \mathrm{MSJ}(2) \iff J^2_e(T) = \{\mu \otimes \mu\} \cup \{\Delta_{T^n}, n \in \mathbb{Z}\}\,. \tag{9}$$

Systems with two-fold minimal self-joinings have very interesting properties. First, since for any S in $C(T)$, Δ_S belongs to $J^2_e(T)$, we immediately see that the commutant of T is reduced to the powers of T. In particular, it is impossible to find a square root of T, i. e. a measure-preserving S such that $S \circ S = T$. Second, the existence of a non-trivial factor σ-algebra of T would lead, via the relatively independent self-joining over this factor, to some ergodic two-fold self-joining of T which is not in the list prescribed by (9). Therefore, any factor σ-algebra of a system with two-fold minimal self-joinings must be either the trivial σ-algebra $\{\emptyset, X\}$ or the whole σ-algebra $\mathcal{A}$. This has the remarkable consequence that if ξ is any random

[1] An introduction to finite-rank transformations can be found e. g. in [32]; we also refer the reader to the quite complete survey [12].

variable on the underlying probability space which is not almost-surely constant, then the process $(\xi \circ T^n)_{n\in\mathbb{Z}}$ always generates the whole σ-algebra $\mathcal{A}$. This also implies that T has zero entropy, since positive-entropy systems have many non-trivial factors.

The notion of two-fold minimal self-joinings extends for any integer $k \geq 2$ to *k-fold minimal self-joinings*, which roughly means that there are no other k-fold ergodic self-joinings than the "obvious" ones: Those for which the k coordinates are either independent or just translated by some power of T (see the Glossary for a more precise definition). We denote in this case: $T \in \mathrm{MSJ}(k)$. If T has k-fold minimal self-joinings for all $k \geq 2$, we simply say that T has *minimal self joinings*.

Rudolph's construction of a system with two-fold minimal self-joinings [41] was inspired by a famous work of Ornstein [35], giving the first example of a transformation with no roots. It turned out that Ornstein's example is a mixing rank-one system, and all mixing rank-one systems were later proved by J. King [27] to have *2-fold minimal self-joinings*. This can also be viewed as a consequence of Ryzhikov's Theorem 8. Indeed, in the language of joinings, the mixing property of T translates as follows:

$$T \text{ is mixing} \iff \Delta_{T^n} \xrightarrow[|n|\to\infty]{} \mu \otimes \mu \,. \tag{10}$$

Therefore, if in Theorem 8 we further assume that T is mixing, then either the sequence (n_k) we get in the conclusion is bounded, and then λ is some Δ_{T^n}, or it is unbounded and then $\lambda = \mu \otimes \mu$.

$T \in \mathrm{MSJ}(k)$ obviously implies $T \in \mathrm{MSJ}(k')$ for any $2 \leq k' \leq k$, but the converse is not known. The question whether two-fold minimal self-joinings implies k-fold minimal self-joinings for all k is related to the important open problem of *pairwise independent joinings* (see Sect. "Pairwise-Independent Joinings"). But the latter problem is solved for some special classes of systems, in particular in the category of mixing rank-one transformations. It follows that, if T is mixing and rank one, then T has minimal self-joinings.

In 1980, Del Junco, Rahe and Swanson proved that Chacon's transformation also has minimal self-joinings [9]. This well-known transformation is also a rank-one system, but it is not mixing (it had been introduced by R.V. Chacon in 1969 [4] as the first explicit example of a weakly mixing transformation which is not mixing). For another example of a transformation with two-fold minimal self-joinings, constructed as an exchange map on three intervals, we refer to [7].

The existence of a transformation with minimal self-joinings has been used by Rudolph as a wonderful tool to construct a large variety of striking counterexamples, such as

- A transformation T which has no roots, while T^2 has roots of any order,
- A transformation with a cubic root but no square root,
- Two measure-preserving dynamical systems which are weakly isomorphic (each one is a factor of the other) but not isomorphic…

Let us now sketch the argument showing that we can find two systems with no common factor but which are not disjoint: We start with a system T with minimal self-joinings. Consider the direct product of T with an independent copy T' of itself, and take the *symmetric factor* S of $T \otimes T'$, that is to say the factor we get if we only look at the non-ordered pair of coordinates $\{x, x'\}$ in the Cartesian product. Then S is surely not disjoint from T, since the pair $\{x, x'\}$ is not independent of x. However, if S and T had a non-trivial common factor, then this factor should be isomorphic to T itself (because T has minimal self-joinings). Therefore we could find in the direct product $T \otimes T'$ a third copy $\tilde{T}$ of T, which is measurable with respect to the symmetric factor. In particular, $\tilde{T}$ is invariant by the flip map $(x, x') \mapsto (x', x)$, and this prevents $\tilde{T}$ from being measurable with respect to only one coordinate. Then, since $T \in \mathrm{MSJ}(3)$, the systems T, T' and $\tilde{T}$ have no choice but being independent. But this contradicts the fact that $\tilde{T}$ is measurable with respect to the σ-algebra generated by x and x'. Hence, T and S have no non-trivial common factor.

We can also cite the example given by Glasner and Weiss [21] of a pair of horocycle transformations which have no nontrivial common factor, yet are not disjoint. Their construction relies on the deep work by Ratner [38], which describes the structure of joinings of horocycle flows.

Simple Systems

An important generalization of two-fold minimal self-joinings has been proposed by William A. Veech in 1982 [51]. We say that the measure-preserving dynamical system T is *two-fold simple* if it has no other ergodic two-fold self-joinings than the product measure $\{\mu \otimes \mu\}$ and joinings supported on the graph of a transformation $S \in C(T)$. (The difference with MSJ(2) lies in the fact that $C(T)$ may contain other transformations than the powers of T.) It turns out that simple systems may have non-trivial factors, but the structure of these factors can be explicitly described: They are always associated with some compact subgroup of $C(T)$. More precisely, if K is a com-

pact subgroup of $C(T)$, we can consider the factor σ-algebra

$$\mathcal{F}_K := \{A \in \mathcal{A} : \forall S \in K\,, A = S(A)\}\ ,$$

and the corresponding factor transformation $T_{|\mathcal{F}_K}$ (called a *group factor*). Then Veech proved the following theorem concerning the structure of factors of a two-fold simple system.

Theorem 9 *If the dynamical system T is two-fold simple, and if $\mathcal{F} \subset \mathcal{A}$ is a non-trivial factor σ-algebra of T, then there exists a compact subgroup K of the group $C(T)$ such that $\mathcal{F} = \mathcal{F}_K$.*

There is a natural generalization of Veech's property to the case of k-fold self-joinings, which has been introduced by A. Del Junco and D.J. Rudolph in 1987 [10] (see the precise definition of simple systems in the Glossary). In their work, important results concerning the structure of factors and joinings of simple systems are proved. In particular, they are able to completely describe the structure of the ergodic joinings between a given simple system and any ergodic system (see also [18,49]). Recall that, for any $r \geq 1$, the *symmetric factor* of $T^{\otimes r}$ is the system we get if we observe the r coordinates of the point in X^r and forget their order. This is a special case of group factor, associated with the subgroup of $C(T^{\otimes r})$ consisting of all permutations of the coordinates. We denote this symmetric factor by $T^{\langle r\rangle}$.

Theorem 10 *Let T be a simple system and S an ergodic system. Assume that λ is an ergodic joining of T and S which is different from the product measure. Then there exists a compact subgroup K of $C(T)$ and an integer $r \geq 1$ such that*

- *$(T_{|\mathcal{F}_K})^{\langle r\rangle}$ is a factor of S,*
- *λ is the projection on $X \times Y$ of the relatively independent joining of $T^{\otimes r}$ and S over their common factor $(T_{|\mathcal{F}_K})^{\langle r\rangle}$.*

If we further assume that the second system is also simple, then in the conclusion we can take $r = 1$. In other words, ergodic joinings of simple systems S and T are either the product measure or relatively independent joinings over a common group factor. This leads to the following corollary:

Theorem 11 *Simple systems without non-trivial common factor are disjoint.*

As for minimal self-joining, it is not known in general whether two-fold simplicity implies k-fold simplicity for all k. This question is studied in [19], where sufficient spectral conditions are given for this implication to hold. It is also proved that any three-fold simple weakly mixing transformation is simple of all order.

Relative Properties with Respect to a Factor

In fact, Veech also introduced a weaker, "relativized", version of the two-fold simplicity. If $\mathcal{F} \subset \mathcal{A}$ is a non-trivial factor σ-algebra of T, let us denote by $J^2(T, \mathcal{F})$ the two-fold self-joinings of T which are "constructed over $\mathcal{F}$", which means that their restriction to the product σ-algebra $\mathcal{F} \otimes \mathcal{F}$ coincides with the diagonal measure. (The relatively-independent joining over $\mathcal{F}$ is the canonical example of such a joining.) For the conclusion of Theorem 9 to hold, it is enough to assume only that the ergodic elements of $J^2(T, \mathcal{F})$ be supported on the graph of a transformation $S \in C(T)$. This is an important situation where the study of $J^2(T, \mathcal{F})$ gives strong information on the way $\mathcal{F}$ is embedded in the whole system T, or, in other words, on the relative properties of T with respect to the factor $T|_{\mathcal{F}}$. A simple example of such a relative property is the *relative weak mixing with respect to $\mathcal{F}$*, which is characterized by the ergodicity of the relatively-independent joining over $\mathcal{F}$ (recall that weak-mixing is itself characterized by the ergodicity of the direct product $T \otimes T$).

For more details on this subject, we refer the reader to [30]. We also wish to mention the generalization of simplicity called *semisimplicity* proposed by Del Junco, Lemańczyk and Mentzen in [8], which is precisely characterized by the fact that, for any $\lambda \in J^2_e(T)$, the system $(T \otimes T)_\lambda$ is a relatively weakly mixing extension of T.

Some Applications and Future Directions

Joinings Proofs of Ornstein's and Krieger's Theorems

We have already seen that joinings could be used to prove isomorphisms between systems. This fact found a nice application in the proofs of two major theorems in ergodic theory: Ornstein's isomorphism theorem [34], stating that two Bernoulli shifts with the same entropy are isomorphic, and Krierger's finite generator theorem [28], which says that any dynamical system with finite entropy is isomorphic to the shift transformation on a finite-valued stationary process. The idea of this joining approach to the proofs of Krieger's and Ornstein's theorems was originally due to Burton and Rothstein, who circulated a preliminary report on the subject which was never published [3]. The first published and fully detailed exposition of these proofs can be found in Rudolph's book [42] (see also in Glasner's book [18]).

In fact, Ornstein's theorem goes far more beyond the isomorphism of two given Bernoulli shifts: It also gives

a powerful tool for showing that a specific dynamical system is isomorphic to a Bernoulli shift. In particular, Ornstein introduced the property for an ergodic stationary process to be *finitely determined*. We shall not give here the precise definition of this property (for a complete exposition of Ornstein's theory, we refer the reader to [36]), but simply point out that Bernoulli shifts and mixing Markov chains are examples of finitely determined processes. Rudolph's argument to show Ornstein's theorem via joinings makes use of Theorem 4, and of the topology of $J(T,S)$.

Theorem 12 (Ornstein's Isomorphism Theorem) *Let T and S be two ergodic dynamical systems with the same entropy, and both generated by finitely determined stationary processes. Then the set of joinings of T and S which are supported on graphs of isomorphisms forms a dense G_δ in $J_e(T,S)$.*

Krieger's theorem is not as easily stated in terms of joinings, because it does not refer to the isomorphism of two specific systems, but rather to the isomorphism of one given system with some other system which has to be found. We have therefore to introduce a larger set of joinings: Given an integer n, we denote by Y_n the set of double-sided sequences taking values in $\{1,\dots,n\}$. We consider on Y_n the shift transformation S, but we do not determine yet the invariant measure. Now, for a specific measure-preserving dynamical system T, consider the set $J(n,T)$ of all possible joinings of T with some system (Y_n,ν,S), when ν ranges over all possible shift-invariant probability measures on Y_n. $J(n,T)$ can also be equipped with a topology which turns it into a compact convex metric space, and as soon as T is ergodic, the set $J_e(n,T)$ of *ergodic* elements of $J(n,T)$ is not empty. In this setting, Krieger's theorem can be stated as follows:

Theorem 13 (Krieger's Finite Generator Theorem) *Let T be an ergodic dynamical system with entropy $h(T) < \log_2 n$. Then the set of $\lambda \in J(n,T)$ which are supported on graphs of isomorphisms between T and some system (Y_n,ν,S) forms a dense G_δ in $\overline{J_e(n,T)}$.*

Since any system of the form (Y_n,ν,S) obviously has an n-valued generating process, we obtain as a corollary that T itself is generated by an n-valued process.

Joinings and Pointwise Convergence

The study of joinings is also involved in questions concerning pointwise convergence of (non-conventional) ergodic averages. As an example, we present here the relationships between disjointness and the following well-known open problem: Given two commuting measure-preserving transformations S and T acting on the same probability space $(X,\mathcal{A},\mu)$, is it true that for any f and g in $L^2(\mu)$, the sequence

$$\left(\frac{1}{n}\sum_{k=0}^{n-1} f(T^k x)g(S^k x)\right)_{n>0} \tag{11}$$

converges μ-almost surely?

It turns out that disjointness of T and S is a sufficient condition for this almost-sure convergence to hold. Indeed, let us first consider the case where T and S are defined on a priori different spaces $(X,\mathcal{A},\mu)$ and $(Y,\mathcal{B},\nu)$ respectively, and consider the ergodic average in the product

$$\frac{1}{n}\sum_{k=0}^{n-1} f(T^k x)g(S^k y)\,, \tag{12}$$

which can be viewed as the integral of the function $f\otimes g$ with respect to the empirical distribution

$$\delta_n(x,y) := \frac{1}{n}\sum_{k=0}^{n-1}\delta_{(T^k x, S^k y)}\,.$$

We can always assume that T and S are continuous transformations of compact metric spaces (indeed, any measure-preserving dynamical system is isomorphic to such a transformation on a compact metric space; see e. g. [15]). Then the set of probability measures on $X\times Y$ equipped with the topology of weak convergence is metric compact. Now, here is the crucial point where joinings appear: If T and S are ergodic, we can easily find subsets $X_0\subset X$ and $Y_0\subset Y$ with $\mu(X_0)=\nu(Y_0)=1$, such that for all $(x,y)\in X_0\times Y_0$, any cluster point of the sequence $(\delta_n(x,y))_{n>0}$ is automatically a joining of T and S. (We just have to pick x and y among the "good points" for the ergodic theorem in their respective spaces.) When T and S are disjoint, there is therefore only one possible cluster point to the sequence $\delta_n(x,y)$ which is $\mu\otimes\nu$. This ensures that, for continuous f and g, (12) converges to the product of the integrals of f and g as soon as (x,y) is picked in $X_0\times Y_0$. The subspace of continuous functions being dense in L^2, the classical ergodic maximal inequality (see [17]) ensures that, for any f and g in $L^2(\mu)$, (12) converges for any (x,y) in a rectangle of full measure $X_0\times Y_0$.

Coming back to the original question where the spaces on which T and S act are identified, we observe that with probability one, x belongs both to X_0 and Y_0, and therefore the sequence (11) converges.

The existence of a rectangle of full measure in which the sequence of empirical distributions $(\delta_n(x,y))_{n>0}$ always converges to some joining has been studied in [31] as

a natural generalization of the notion of disjointness. This property was called *weak disjointness* of S and T, and it is indeed strictly weaker than disjointness, since there are examples of transformations which are weakly disjoint from themselves.

There are other situations in which joinings can be used in the study of everywhere convergence, among which we can cite Rudolph's joinings proof of Bourgain's return time theorem [43].

Joinings and Rohlin's Multifold Mixing Question

We have already seen that the property of T being mixing could be expressed in terms of two-fold self-joinings of T (see (10)). Rohlin proposed in 1949 [39] a generalization of this property, called *multifold mixing*: The measure-preserving transformation T is said to be *k-fold mixing* if $\forall A_1, A_2, \ldots, A_k \in \mathcal{A}$,

$$\lim_{n_2,n_3,\ldots,n_k\to\infty} \mu(A_1 \cap T^{-n_2}A_2 \cap \cdots \cap T^{-(n_2+\cdots+n_k)}A_k) = \prod_{i=1}^{k} \mu(A_i)\,.$$

Again, this definition can easily be translated into the language of joinings: T is k-fold mixing when the sequence $(\Delta_{T^{n_2},\ldots,T^{n_2+\cdots+n_k}})$ converges in $J^k(T)$ to $\mu^{\otimes k}$ as $n_2, \ldots, n_k$ go to infinity, where $(\Delta_{T^{n_2},\ldots,T^{n_2+\cdots+n_k}})$ is the obvious generalization of Δ_{T^n} to the case of k-fold self-joinings. The classical notion of mixing corresponds in this setting to two-fold mixing. (We must point out that Rohlin's original definition of k-fold mixing involved $k+1$ sets, thus the classical mixing property was called *1-fold mixing*. However it seems that the convention we adopt here is now used by most authors, and we find it more coherent when translated in the language of multifold self-joinings.)

Obviously, *3-fold mixing* is stronger than *2-fold mixing*, and Rohlin asked in his article whether the converse is true. This question is still open today, even though many important works have dealt with it and supplied partial answers. Most of these works directly involve self-joinings via the argument exposed in the following section.

Pairwise-Independent Joinings Let T be a two-fold mixing dynamical system. If T is not three-fold mixing, $(\Delta_{T^n,T^{n+m}})$ does not converge to the product measure as n and m go to ∞. By compactness of $J^3(T)$, we can find subsequences (n_k) and (m_k) such that $(\Delta_{T^{n_k},T^{n_k+m_k}})$ converges to a cluster point $\lambda \neq \mu \otimes \mu \otimes \mu$. However, by two-fold mixing, the three coordinates must be pairwise independent under λ. We therefore get a three-fold self-joining λ with the unusual property that λ has pairwise independent coordinates, but λ is not the product measure.

In fact, systems with this kind of pairwise-independent but non-independent three-fold self-joining are easy to find (see e. g. [6]), but the examples we know so far are either periodic transformations (which cannot be counterexamples to Rohlin's question since they are not mixing!), or transformations with positive entropy. But using an argument provided by Thouvenot, we can prove that, if there exists a two-fold mixing T which is not three-fold mixing, then we can find such a T in the category of zero-entropy dynamical systems (see e. g. [5]). Therefore, a negative answer to the following question would solve Rohlin's multifold mixing problem:

Question 14 Does there exist a zero-entropy, weakly mixing dynamical system T with a self-joining $\lambda \in J^3(T)$ for which the coordinates are pairwise independent but which is different from $\mu \otimes \mu \otimes \mu$?

(non) existence of such pairwise-independent joinings is also related to the question of whether MSJ(2) implies MSJ(3), or whether two-fold simplicity implies three-fold simplicity. Indeed, any counter-example to one of these implication would necessarily be of zero entropy, and would possess a pairwise-independent three-fold self-joining which is not the product measure.

Question 14 has been answered by B. Host and V. Ryzhikov for some special classes of zero-entropy dynamical systems.

Host's and Ryzhikov's Theorems The following theorem, proved in 1991 by B. Host [23] (see also [18,33]), establishes a spectacular connection between the spectral properties of a finite family of dynamical systems and the non-existence of a pairwise-independent, non-independent joining:

Theorem 15 (Host's Theorem on Singular Spectrum) *Let $(X_i, \mathcal{A}_i, \mu_i, T_i)_{1\le i\le r}$ be a finite family of measure-preserving dynamical systems with purely singular spectrum. Then any pairwise-independent joining $\lambda \in J(T_1, \ldots, T_r)$ is the product measure $\mu_1 \otimes \cdots \otimes \mu_r$.*

Corollary 16 *If a dynamical system with singular spectrum is two-fold mixing, then it is k-fold mixing for any $k \ge 2$.*

The multifold-mixing problem for rank-one measure-preserving systems was solved in 1984 by S. Kalikow [25], using arguments which do not involve the theory of joinings. In 1993, V. Ryzhikov [46] extended Kalikow's result to finite-rank systems, by giving a negative answer to Question 14 in the category of finite-rank mixing systems:

Theorem 17 (Ryzhikov's Theorem for Finite Rank Systems) *Let T be a finite-rank mixing transformation, and $k \geq 2$. Then the only pairwise-independent k-fold self-joining of T is the product measure.*

Corollary 18 *If a finite-rank transformation is two-fold mixing, then it is k-fold mixing for any $k \geq 2$.*

Future Directions

A lot of important open questions in ergodic theory involve joinings, and we already have cited several of them: Joinings are a natural tool when we want to deal with some problems of pointwise convergence involving several transformations (see Sect. "Joinings and Pointwise Convergence"). Their use is also fundamental in the study of Rohlin's question on multifold mixing. As far as this latter problem is concerned, we may mention a recent approach to Question 14: Start with a transformation for which some special pairwise-independent self-joining exists, and see what this assumption entails. In particular, we can ask under which conditions there exists a pairwise-independent three-fold self-joining of T under which the third coordinate is a function of the two others. It has already been proven in [24] that if this function is sufficiently regular (continuous for some topology), then T is periodic or has positive entropy. And there is strong evidence leading to the conjecture that, when T is weakly mixing, such a situation can only arise when T is a Bernoulli shift of entropy $\log n$ for some integer $n \geq 2$. A question in the same spirit was raised by Ryzhikov, who asked in [44] under which conditions we can find a factor of the direct product $T \times T$ which is independent of both coordinates.

There is also a lot of work to do with joinings in order to understand the structure of factors of some dynamical systems, and how different classes of systems are related. An example of such a work is given in the class of Gaussian dynamical systems, i. e. dynamical systems constructed from the shift on a stationary Gaussian process: For some of them (which are called *GAG*, from the French *Gaussien à Autocouplages Gaussiens*), it can be proven that any ergodic self-joining is itself a Gaussian system (see [29,50]), and this gives a complete description of the structure of their factors. This kind of analysis is expected to be applicable to other classes of dynamical systems. In particular, Gaussian joinings find a nice generalization in the notion of *infinitely divisible joinings*, studied by Roy in [40]). These ID joinings concern a wider class of dynamical systems of probabilistic origin, among which we can also find Poisson suspensions. The counterpart of Gaussian joinings in this latter class are *Poisson joinings*, which have been introduced by Derriennic, Frączek, Lemańczyk and Parreau in [11]. As far as Poisson suspensions are concerned, the analog of the GAG property in the Gaussian class can also be considered, and a family of Poisson suspension for which the only ergodic self-joinings are Poisson joinings has been given recently by Parreau and Roy [37]. In [11], a general joining property is described: T satisfies the *ELF property* (from the French: *Ergodicité des Limites Faibles*) if any joining which can be obtained as a limit of off-diagonal joinings $\Delta_{T^{n_k}}$ is automatically ergodic. It turns out that this property is satisfied by any system arising from an infinitely divisible stationary process (see [11,40]). It is proven in [11] that ELF property implies disjointness with any system which is two-fold simple and weakly mixing but not mixing. ELF property is expected to give a useful tool to prove disjointness between dynamical systems of probabilistic origin and other classes of systems (see e. g. [13] in the case of $\mathbb{R}$-action for disjointness between ELF systems and a class of special flows over irrational rotations).

Many other questions involving joinings have not been mentioned here. We should at least cite *filtering problems*, which were one of the motivations presented by Furstenberg for the introduction of the disjointness property in [14]. Suppose we are given two real-valued stationary processes (X_n) and (Y_n), with their joint distribution also stationary. We can interpret (X_n) as a signal, perturbed by a noise (Y_n), and the question posed by Furstenberg is: Under which condition can we recover the original signal (X_n) from the observation of $(X_n + Y_n)$? Furstenberg proved that it is always possible if the two processes (X_n) and (Y_n) are integrable, and if the two measure-preserving dynamical systems constructed as the shift of the two processes are disjoint. Furstenberg also observed that the integrability assumption can be removed if a stronger disjointness property is satisfied: A perfect filtering exists if the system T generated by (X_n) is *doubly disjoint* from the system S generated by (Y_n), in the sense that T is disjoint from any ergodic self-joining of S. Several generalizations have been studied (see [2,16]), but the question of whether the integrability assumption of the processes can be removed is still open.

Bibliography

1. Ahn Y-H, Lemańczyk M (2003) An algebraic property of joinings. Proc Amer Math Soc 131(6):1711–1716 (electronic)
2. Bułatek W, Lemańczyk M, Lesigne E (2005) On the filtering problem for stationary random $\mathbb{Z}^2$-fields. IEEE Trans Inform Theory 51(10):3586–3593
3. Burton R, Rothstein A (1977) Isomorphism theorems in ergodic theory. Technical report, Oregon State University
4. Chacon RV (1969) Weakly mixing transformations which are not strongly mixing. Proc Amer Math Soc 22:559–562

5. de la Rue T (2006) 2-fold and 3-fold mixing: why 3-dot-type counterexamples are impossible in one dimension. Bull Braz Math Soc (NS) 37(4):503–521
6. de la Rue T (2006) An introduction to joinings in ergodic theory. Discret Contin Dyn Syst 15(1):121–142
7. del Junco A (1983) A family of counterexamples in ergodic theory. Isr J Math 44(2):160–188
8. del Junco A, Lemańczyk M, Mentzen MK (1995) Semisimplicity, joinings and group extensions. Studia Math 112(2):141–164
9. del Junco A, Rahe M, Swanson L (1980) Chacon's automorphism has minimal self-joinings. J Analyse Math 37:276–284
10. del Junco A, Rudolph DJ (1987) On ergodic actions whose self-joinings are graphs. Ergod Theory Dynam Syst 7(4):531–557
11. Derriennic Y, Frączek K, Lemańczyk M, Parreau F (2008) Ergodic automorphisms whose weak closure of off-diagonal measures consists of ergodic self-joinings. Colloq Math 110:81–115
12. Ferenczi S (1997) Systems of finite rank. Colloq Math 73(1): 35–65
13. Frączek K, Lemańczyk M (2004) A class of special flows over irrational rotations which is disjoint from mixing flows. Ergod Theory Dynam Syst 24(4):1083–1095
14. Furstenberg H (1967) Disjointness in ergodic theory, minimal sets, and a problem in Diophantine approximation. Math Syst Theory 1:1–49
15. Furstenberg H (1981) Recurrence in ergodic theory and combinatorial number theory. In: M.B. Porter Lectures. Princeton University Press, Princeton
16. Furstenberg H, Peres Y, Weiss B (1995) Perfect filtering and double disjointness. Ann Inst H Poincaré Probab Stat 31(3):453–465
17. Garsia AM (1970) Topics in almost everywhere convergence. In: Lectures in Advanced Mathematics, vol 4. Markham Publishing Co, Chicago, IL
18. Glasner E (2003) Ergodic theory via joinings. In: Mathematical Surveys and Monographs, vol 101. American Mathematical Society, Providence
19. Glasner E, Host B, Rudolph DJ (1992) Simple systems and their higher order self-joinings. Isr J Math 78(1):131–142
20. Glasner E, Thouvenot J-P, Weiss B (2000) Entropy theory without a past. Ergod Theory Dynam Syst 20(5):1355–1370
21. Glasner S, Weiss B (1983) Minimal transformations with no common factor need not be disjoint. Isr J Math 45(1):1–8
22. Goodson GR (2000) Joining properties of ergodic dynamical systems having simple spectrum. Sankhyā Ser A 62(3):307–317, Ergodic theory and harmonic analysis (Mumbai, 1999)
23. Host B (1991) Mixing of all orders and pairwise independent joinings of systems with singular spectrum. Isr J Math 76(3):289–298
24. Janvresse É, de la Rue T (2007) On a class of pairwise-independent joinings. Ergod Theory Dynam Syst (to appear)
25. Kalikow SA (1984) Twofold mixing implies threefold mixing for rank one transformations. Ergod Theory Dynam Syst 4(2): 237–259
26. King J (1986) The commutant is the weak closure of the powers, for rank-1 transformations. Ergod Theory Dynam Syst 6(3):363–384
27. King J (1988) Joining-rank and the structure of finite rank mixing transformations. J Anal Math 51:182–227
28. Krieger W (1970) On entropy and generators of measure-preserving transformations. Trans Amer Math Soc 149:453–464
29. Lemańczyk M, Parreau F, Thouvenot J-P (2000) Gaussian automorphisms whose ergodic self-joinings are Gaussian. Fund Math 164(3):253–293
30. Lemańczyk M, Thouvenot J-P, Weiss B (2002) Relative discrete spectrum and joinings. Monatsh Math 137(1):57–75
31. Lesigne E, Rittaud B, and de la Rue T (2003) Weak disjointness of measure-preserving dynamical systems. Ergod Theory Dynam Syst 23(4):1173–1198
32. Nadkarni MG (1998) Basic ergodic theory. In: Birkhäuser Advanced Texts: Basler Lehrbücher, 2nd edn. Birkhäuser, Basel
33. Nadkarni MG (1998) Spectral theory of dynamical systems. In: Birkhäuser Advanced Texts: Basler Lehrbücher. Birkhäuser, Basel
34. Ornstein DS (1970) Bernoulli shifts with the same entropy are isomorphic. Adv Math 4:337–352
35. Ornstein DS (1972) On the root problem in ergodic theory. In: Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Univ. California, Berkeley, 1970/1971, vol II: Probability theory. Univ California Press, Berkeley, pp 347–356
36. Ornstein DS (1974) Ergodic theory, randomness, and dynamical systems. In: James K Whittemore Lectures in Mathematics given at Yale University, Yale Mathematical Monographs, vol 5. Yale University Press, New Haven
37. Parreau F, Roy E (2007) Poisson joinings of Poisson suspension. Preprint
38. Ratner M (1983) Horocycle flows, joinings and rigidity of products. Ann of Math 118(2):277–313
39. Rohlin VA (1949) On endomorphisms of compact commutative groups. Izvestiya Akad Nauk SSSR Ser Mat 13:329–340
40. Roy E (2007) Poisson suspensions and infinite ergodic theory. Ergod Theory Dynam Syst(to appear)
41. Rudolph DJ (1979) An example of a measure preserving map with minimal self-joinings, and applications. J Anal Math 35:97–122
42. Rudolph DJ (1990) The title is Fundamentals of Measurable Dynamics: Ergodic Theory on Lebesgue Spaces. Oxford University Press, New York
43. Rudolph DJ (1994) A joinings proof of Bourgain's return time theorem. Ergod Theory Dynam Syst 14(1):197–203
44. Ryzhikov VV (1992) Stochastic wreath products and joinings of dynamical systems. Mat Zametki 52(3):130–140, 160
45. Ryzhikov VV (1992) Mixing, rank and minimal self-joining of actions with invariant measure. Mat Sb 183(3):133–160
46. Ryzhikov VV (1993) Joinings and multiple mixing of the actions of finite rank. Funktsional Anal Prilozhen 27(2):63–78, 96
47. Ryzhikov VV (1993) Joinings, wreath products, factors and mixing properties of dynamical systems. Izv Ross Akad Nauk Ser Mat 57(1):102–128
48. Thorisson H (2000) Coupling, stationarity, and regeneration. In: Probability and its Applications (New York). Springer, New York
49. Thouvenot J-P (1995) Some properties and applications of joinings in ergodic theory. In: Ergodic theory and its connections with harmonic analysis, Alexandria, 1993. London Math Soc Lecture Note Ser, vol 205. Cambridge Univ Press, Cambridge, pp 207–235
50. Thouvenot J-P (1987) The metrical structure of some Gaussian processes. In: Proceedings of the conference on ergodic theory and related topics, II, Georgenthal, 1986. Teubner-Texte Math, vol 94. Teubner, Leipzig, pp 195–198
51. Veech WA (1982) A criterion for a process to be prime. Monatsh Math 94(4):335–341

K

Knowledge Discovery: Clustering

PAVEL BERKHIN[1], INDERJIT S. DHILLON[2]
[1] eBay Inc., San Jose, USA
[2] University of Texas, Austin, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Partitioning Relocation Clustering
Hierarchical Clustering
Spectral Clustering
Graph Clustering
Clusters as Dense Connected Components
Clustering Transactional Data
Co-clustering
Scalability
Other Aspects of Clustering
Future Directions
Bibliography

Glossary

BIRCH A scalable clustering solution that first assembles an in-memory balanced tree of micro-clusters (called clustering features) containing sufficient statistics for the data.

CACTUS Clustering algorithm for transactional data based on the idea of co-occurrence.

CHAMELEON Hierarchical clustering algorithm using a very fine partitioning of a sparsified connectivity graph followed by agglomeration phase.

Cluster A subset of the data that consists of similar objects.

CLIQUE Clustering algorithm for high-dimensional data that tries to find high density low-dimensional segments using an inductive process similar to the Apriori algorithm.

Co-clustering Clustering methodology that along with grouping points, also groups attributes that have similar distributions among data points.

Connectivity matrix A matrix of similarities or dissimilarities between data points, which gives pairwise connectivity information that is used in agglomerative, spectral and graph clustering algorithms.

CURE Important scalable hierarchical clustering algorithm that uses a fixed number of points as cluster representatives.

DBSCAN Density-based partitioning that utilizes a definition of density-connectivity and a core point (a point whose ϵ-neighborhood has sufficiently many points).

Density-based partitioning Clustering algorithm that tries to identify clusters with dense connected components of arbitrary shape.

GRACLUS A super-fast graph clustering algorithm that optimizes weighted graph cuts.

Graph clustering Clustering of graph nodes; instead of a concept of distance, a concept of connectivity based on graph edges is used.

Grid methods Clustering algorithms that find relatively high populated segments in an underlying attribute space and then assemble clusters from adjacent segments.

Hierarchical clustering Represents the data in the form of a tree dendrogram whose leaves correspond to individual points and nodes to clusters of different granularities.

***k*-means** Partitioning relocation clustering algorithm used in many applications that deals with numerical data; represents a cluster with its centroid or mean.

***k*-medoid methods** Partitioning relocation algorithm (e. g. CLARANS) that represents a cluster by one of its points, called medoid.

Linkage metric Metric used in hierarchical clustering in conjunction with the Lance-Williams updating formula to compute similarity between two subsets.

Probabilistic clustering Clustering algorithm that associates with each cluster a particular probability distribution whose parameters are fitted by the algorithm.

Spectral clustering Clustering algorithm that uses eigenvectors or singular vectors for data or graph clustering.

Definition of the Subject

Data that we find in scientific and business applications usually does not fit a particular parameterized probability distribution. In other words, the data is *complex*. Knowledge discovery starts with exploration of this complexity in order to find inconsistencies, artifacts, errors, etc in the data. After data is cleaned, it is usually still extremely complex. Descriptive data mining deals with comprehending and reducing this complexity. Clustering is a premier methodology in descriptive unsupervised data mining.

Clustering is the division of N data points $X = \{x_1, x_2, \ldots, x_N\}$ into k disjoint groups C_i. Each group, called *cluster*, is required to consist of points that are similar to one another and dissimilar to points in other groups:

$$X = C_1 \bigcup \cdots \bigcup C_k \bigcup C_{\text{outliers}}, C_i \bigcap C_j = 0, \quad i \neq j.$$

A cluster could represent an important subset of the data such as a galaxy in astronomical data or a segment of customers in marketing applications. Clustering is important as a fundamental technology to reduce data complexity and to find data patterns in an unsupervised fashion. It is universally used as a first technology of choice in data exploration.

Introduction

Classic clustering algorithms have existed for a long time. Contemporary clustering faces many challenges, such as (a) sheer size of modern data sets, (b) objects with many attributes, (c) attributes of different types, and (d) unstructured data or data of complex structure. These challenges have led to the emergence of powerful and broadly applicable clustering methods. First, we should ask ourselves: does a set of axioms exists that would result in a consistent clustering framework? A sobering answer was given by Jon Kleinberg [49]: under natural assumptions, clustering cannot be axiomatized. General references on clustering include [7,25,28,33,39,40,43,44,47,62]. From a statistical standpoint, clustering relates to a traditional multivariate statistical estimation.

System complexity provides another fruitful way of looking at clustering. As a result of clustering, data complexity is reduced to a small number of clusters. More precisely, to transmit data we can transmit (1) k cluster "prototypes"; (2) each point's cluster ID and a relatively short encoding that describes deviation of a point from its cluster "prototype". This connection to data compression in used in image processing (*vector quantization* [32]). Other clustering applications include scientific data analysis (astronomy), biochemistry and medicine, information retrieval and text mining, spatial database applications, sequence and heterogeneous data analysis, web applications, marketing, user segmentation, fraud detection, and many others.

We use the following notation. Dataset X consists of data points (*objects*, *instances*, *cases*, etc.) $x_i = (x_{i1}, \cdots, x_{id})$, $i = 1:N$, in d-dimensional attribute space A, $x_{il} \in A_l$, $l = 1:d$. This point-by-attribute data format conceptually corresponds to a $N \times d$ matrix and is used by the majority of algorithms. A simplest subset in A, a *segment*, is a direct Cartesian product of sub-ranges. Some clustering algorithms indeed use segments as building blocks for clusters. Data in other formats, such as variable length sequences and heterogeneous data, are not uncommon.

Partitioning Relocation Clustering

If you know nothing about clustering, the k-means algorithm is what you should start with and many people never go beyond it in their practice.

The name comes from representing each of the k clusters C_j by the mean (or weighted average) c_j of its points, the so-called *centroid*. While this representation does not work well with categorical attributes, it makes sense from a geometric and statistical perspective for numerical attributes. K-Means tries to minimize the objective function equal to the sum of the squares of L_2-norm errors between the points and the corresponding centroids:

$$E(C) = \sum_{j=1:k} \sum_{x_i \in C_j} \left\| x_i - c_j \right\|^2 .$$

The above can be thought of as the within-cluster variance. It turns out that the total data variance can be expressed as the sum of the within-cluster variance and the between-cluster variance:

$$\sum_{i=1}^{N} \|x_i - c\|^2 = \sum_{j=1:k} \sum_{x_i \in C_j} \left\| x_i - c_j \right\|^2 + \sum_{j=1:k} n_j \left\| c_j - c \right\|^2 ,$$

where $n_j = |C_j|$. Thus while k-means explicitly tries to minimize the within-cluster variance, it implicitly maximizes the between-cluster variance. By examining the k-means objective carefully, it is easy to show that k-means is restricted to separating clusters by linear separators, i. e., by hyperplanes.

Up to a constant, $E(C)$ can be recognized as the negative of the log-likelihood for a normally distributed mixture model with uniform variance (points in C_j are distributed as $N(c_j, 1)$). Therefore, the k-means algorithm is related to a general probabilistic framework. This suggests generalizations: fit not only means, but variances, or hyper-ellipsoidal clusters.

The squared L_2-distance has many unique properties. For example, $E(C)$ equals the sum of pair-wise errors within all the clusters:

$$E'(C) = \frac{1}{2} \sum_{j=1:k} \sum_{x_i, y_i \in C_j} \|x_i - y_i\|^2 .$$

Other dissimilarity measures can also be used with a k-means like algorithm. Given a dissimilarity measure $d(x, y)$, the representative vector of a cluster can be defined as:

$$z_j = \text{argmin}_z \sum_{x_i \in C_j} d(x_i, z) .$$

For the squared L_2 distance, z_j simply equals the mean vector c_j. It turns out that the same result is true, i. e., the mean vector is the representative vector for a much larger class of dissimilarity measures called Bregman divergences [4]. These divergences are not always symmetric and they do not obey the triangle inequality but they have many other desirable properties.

The exact optimum for $E(C)$ cannot be computed, but two versions of k-means iterative optimization converging to a local minimum are known. The first version is similar to the EM algorithm. It makes two-step major iterations: (1) reassign points to their nearest centroids; (2) recompute centroids of newly assembled groups. Iterations continue until a stopping criterion is achieved. The result is independent of data ordering, and straightforward parallelization can be applied:

Initialize centroids c_i, $i = 1 : k$
Until convergence is achieved do
  for each $x \in X$ ** Step 1: Reassign points
    $I(x) = \text{argmin}\,\{\|c_i - x\|\}$
  for each $1 = 1 : k$ ** Step 2: Recompute centroids
    $c_i = mean\,\{x \in X : I(x) = i\}$

The second version, tries to readjust centroids as soon as reassignment happens. It is not obvious that it is computationally feasible, but it is: in fact, the computational complexity of both versions is the same. This second version can actually result in a better optimum, but it depends on the ordering of points, and is somewhat more difficult to implement:

Initialize centroids c_i, $i = 1 : k$
Until convergence is achieved do
  for each $x \in X$ ** Iterate over all points
    let $x \in C_i$, *and let* $d_i > 0$ *denote the change in* $E(C)$ *on deleting* x *from* C_i
    for each $j \neq i$ *let* $-d_j$ *be the change in* $E(C)$ *on adding* x *to* C_j
    $j = \text{argmax}\,\{d_i - d_j\}$
    if $d_i - d_j > 0$ *reassign* x *to cluster* C_j *and recompute* c_i *and* c_j

The popularity of the k-means algorithm is well deserved: it is easily understood, easily implemented, and based on the firm foundation of analysis of variances. It also has shortcomings:

- During the reassignment stage a cluster can become empty or unbalanced
- The computed local optimum may be quantitatively and qualitatively much worse than the global optimum
- Initialization of centroids is crucial (see [10] for suggestions)
- Choice of k is unclear
- The process is sensitive to outliers
- Algorithm lacks scalability
- Only numerical attributes are covered in a straightforward manner

K-means is the most popular example of a family of clustering algorithms called *Partitioning Relocation*. Algorithms of this family try to reassign points from one potential cluster to another in order to achieve some objective through iterative optimization. For example, in *Probabilistic Clustering* the data is represented as a mixture of k models whose parameters we want to reconstruct. Each model is expressed as a probability distribution – it turns out that a rich family of probability distributions, namely, the exponential family is in one-to-one correspondence with Bregman divergences [4], that were discussed earlier. The exponential family includes multivariate Gaussians, the Poisson, Bernoulli and exponential distributions.

Clusters discovered by probabilistic clustering are conveniently *interpretable*.

We assume that data points are generated (a) by randomly picking a model j (cluster) with probability τ_j, $j = 1 : k$, and (b) by drawing a point x from a corresponding distribution. Maximization of log-likelihood $\log(L)$,

where

$$L = \Pr(X|C) = \prod_{i=1:N} \sum_{j=1:k} \tau_j \Pr\left(x_i \middle| C_j\right),$$

is achieved through *Expectation-Maximization* (EM) iterations, similar to used in k-means: (E) recompute membership probabilities, (M) estimate model parameters (maximize likelihood). Advantages of this method are:

- Points of complex structure can be handled (heterogeneous data, dynamic sequences) by appropriate probabilistic modeling
- Iterations can be stopped and resumed; intermediate model is available
- Clusters have conceptual meaning
- Number of parameters, in particular k, can be addressed within the Bayesian framework

An example of probabilistic clustering is the algorithm AUTOCLASS [15] that covers a broad variety of probability distributions. including Bernoulli, Poisson, Gaussian, and log-normal distributions. Beyond fitting a particular fixed mixture model, AUTOCLASS extends the search to different models and different values of k.

Another *Partitioning Relocation* family, namely k-medoids methods, represents a cluster by one of its points called a *medoid*. This is an easy solution: any attribute type can be handled, no guess on a probability distributions is required, clusters are subsets of points close to respective medoids, and the objective function is defined as the averaged dissimilarity measure between a point and its medoid.

An example is the CLARANS algorithm (Clustering Large Applications based upon RANdomized Search) [56] that deals with spatial databases. In CLARANS a search over a graph whose nodes are subsets of k medoids (points) is performed. Two nodes are connected by an edge if they differ by exactly one medoid. CLARANS is extended to large databases in [26]; this extension relies heavily on data indexing.

Hierarchical Clustering

Hierarchical clustering combines data points into clusters, those clusters into larger clusters, and so forth, creating a hierarchy. A tree representing this hierarchy of clusters is known as a *dendrogram*. Individual data points are the tree leaves. A dendrogram allows the exploration of data at different levels of granularity. An *agglomerative* hierarchical clustering algorithm starts with one-point (singleton) clusters and recursively merges two or more of the most similar clusters. A *divisive* hierarchical clustering starts with a single cluster containing all data points and recursively splits that cluster into appropriate sub-clusters. The process may be terminated when a stopping criterion (frequently, the requested number k of clusters) is achieved. The advantages of hierarchical clustering include flexibility regarding the level of granularity, ease of handling any form of similarity or distance, and applicability to any attribute type. On the negative side, hierarchical clustering faces the difficulty of choosing the right stopping criteria and most hierarchical algorithms do not revisit (intermediate) clusters once they are constructed.

To explore the topic further, consider agglomerative clustering. An $N \times N$ matrix of distances or similarities between points, called *connectivity* matrix, is used to find closest singleton data points to merge together. To merge or split subsets of points rather than individual points, the distance between individual points has to be generalized to the distance between subsets. Such a derived proximity measure is called a *linkage metric*. It has a significant impact on hierarchical algorithms, because it reflects a particular concept of *closeness* and *connectivity*. Important inter-cluster linkage metrics include *single link*, *average link*, and *complete link*. The linkage metric between two subsets of nodes is computed by applying an aggregator operation Op to pairs of dissimilarities between nodes in the first subset C_1 and nodes in the second subset C_2:

$$d(C_1, C_2) = Op\{d(x, y),\ x \in C_1,\ y \in C_2\}.$$

Examples of Op include minimum (single link, algorithm SLINK [61], $Op = \min$), average (average link, Voorhees' method [64], $Op = \text{Avr}$), or maximum (complete link, algorithm CLINK [19], $Op = \max$). All these linkage metrics can be derived from the Lance–Williams updating formula [51],

$$d\left(C_i \bigcup C_j, C_k\right) = a(i)d(C_i, C_k) + a(j)d(C_j, C_k) + b \cdot d(C_i, C_j) + c\left|d(C_i, C_k) - d(C_j, C_k)\right|$$

with coefficients a, b, c depending on a particular linkage metric. This formula helps to avoid actual computing of all pairwise dissimilarities that makes a definition computationally feasible, see [18].

Numerous connections to graph theory exist. SLINK, for example, is related to the problem of finding the Euclidean minimal spanning tree [67] and has $O(N^2)$ complexity. More importantly, when the $N \times N$ connectivity matrix is sparsified, graph methods directly dealing with the connectivity graph can be used. In particular, the hierarchical divisive MST (Minimum Spanning Tree) algorithm is based on partitioning the graph [44].

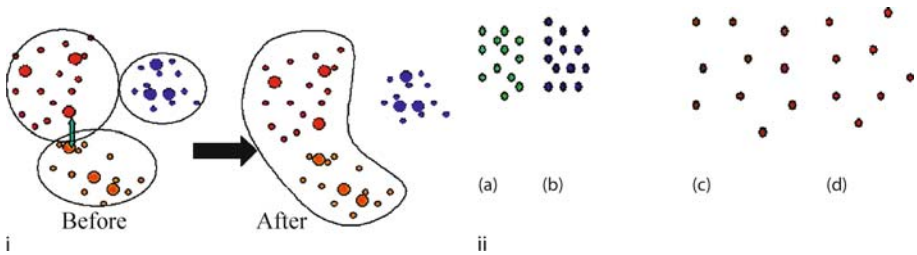

Knowledge Discovery: Clustering, Figure 1
Agglomeration of clusters of arbitrary shapes. i Algorithm CURE ii Algorithm CHAMELEON

Guha et al. [36] introduced the hierarchical agglomerative clustering algorithm CURE (Clustering Using REpresentatives). This algorithm has a number of novel and important features. CURE is the first hierarchical clustering algorithm that was designed with a scalability requirement in mind. It also takes special steps to handle outliers and to provide labeling in the assignment stage. In CURE a cluster is represented by a fixed number, c, of points scattered around it. Selecting representatives scattered around a cluster allows to cover non-spherical shapes. The distance between two clusters used in the agglomerative process is the minimum of distances between two scattered representatives and combines ideas of single and average link closeness. Agglomeration continues until the requested number k of clusters is achieved. CURE employs one additional trick: the original selected scattered points are shrunk to the geometric centroid of the cluster by a user-specified factor α. Shrinkage decreases the impact of outliers; outliers happen to be located further from the cluster centroid than the other scattered representatives. CURE is capable of finding clusters of different shapes and sizes. Figure 1(i) illustrates agglomeration in CURE. Three clusters, each with three representatives, are shown before and after the merge and shrinkage. The two closest representatives are connected. The algorithm CURE works with numerical attributes (particularly, low dimensional spatial data). It is complemented by the algorithm ROCK that handles categorical attributes.

The hierarchical algorithm CHAMELEON, [45], uses a sparsified connectivity graph G: edges corresponding to the K most similar points to any given point are preserved, and the rest are pruned. CHAMELEON performs both the steps of partitioning and agglomeration. It first partitions the data in small tight clusters by using the HMETIS library. Then it agglomerates these small micro-clusters taking into account local measures of connectivity and closeness. The CHAMELEON algorithm does not depend on assumptions about the data model, and has been shown to find clusters of different shapes, densities, and sizes in two-dimensional space. CHAMELEON has complexity $O(Nm + N\log(N) + m^2\log(m))$, where m is the number of micro-clusters built during the first partitioning phase. Figure 1ii clarifies the difference between CHAMELEON and CURE – it shows a choice of four clusters (a)–(d) for a merge. While CURE would merge clusters (a) and (b), CHAMELEON makes the intuitively better choice of merging (c) and (d).

Finally, k-way partitions, when available, naturally lead to divisive hierarchical algorithms. For example, k-means can be used to first divide the data into k clusters, and then it can be recursively applied to each of the k-clusters to yield a hierarchical divisive partitioning. The singular value decomposition (SVD) yields a spectral hierarchical divisive clustering method for document collections called PDDP (Principal Direction Divisive Partitioning) [9]. PDDP is an algorithm that uses the SVD to repeatedly bisect the data into two clusters. In our notation, point x_i is a document, its lth attribute corresponds to a word (*index term*), and x_{il} is a measure (e. g. TF-IDF) of the frequency of term l in document x_i. PDDP begins with the SVD of the matrix

$$(X - \bar{x}\mathrm{e}^{\mathrm{T}}), \quad \text{where} \quad \bar{x} = \frac{1}{N}\sum_{i=1:N} x_i, \mathrm{e} = (1, \ldots, 1)^{\mathrm{T}}.$$

PDDP bisects data in Euclidean space by a hyperplane that passes through the data centroid $\bar{x}$ and is orthogonal to the singular vector with the largest singular value.

Spectral Clustering

As mentioned above, the PDDP method uses spectral information, namely, singular vectors for data clustering. In general, the methods of spectral clustering use such spectral information for clustering. Spectral clustering can be viewed as first constructing an $N \times N$ *connectivity matrix* (explicitly or implicitly) between all data points. As an example, the (i, j) entry of the connectivity matrix may be formed to be $\mathrm{e}^{-\|x_i - x_j\|^2/2\sigma^2}$, or $(x_i - \bar{x})^{\mathrm{T}}(x_j - \bar{x})$ (as done implicitly in PDDP). Then the eigenvectors of the connectivity matrix are computed, typically only the leading few.

From these eigenvectors, the clustering of the data can be extracted in a myriad of ways, as discussed below.

Although spectral clustering was introduced to the data mining and machine learning communities recently [55], it has had a long history in graph clustering problems that arise in a variety of applications described in Sect. "Graph Clustering". The connection to graph clustering is not a surprise since the entries of the connectivity matrix may be viewed as weights of edges between vertices that correspond to data points.

There are three main computational issues in spectral clustering. Construction of the connectivity matrix might take time quadratic in the number of data items, which might be prohibitive for large-scale applications. The connectivity matrix is large in size if the number of data points is large, and can often be sparsified, i. e., many of the entries can be thresholded to zero. Computation of the eigenvectors of such a large, sparse symmetric matrix, is done by typically invoking the Lanczos algorithm [57], which can again be a computational bottleneck if many eigenvectors are desired. A PDDP type recursive spectral bisection algorithm is more effective. An important issue in spectral clustering is to obtain the clusters once the eigenvectors are computed. If r eigenvectors $v_1, v_2, \ldots, v_r$ of an $N \times N$ connectivity matrix are computed, then the ith components $v_1(i), v_2(i), \ldots, v_r(i)$ may be viewed as the reduced dimensional representation of the ith data point. In this case, a simple way to obtain clusters from the eigenvectors is to run k-means on the reduced dimensional representations of the data points [55]. More complicated methods for obtaining the clusters can be used that perform better [68].

Graph Clustering

Graph clustering, also called graph partitioning, is applicable when the data is presented in the form of a graph, for example the link structure of the web, or a social network. The graph clustering problem is to partition or cluster the nodes of the given graph, such that the connectivity between partitions is minimized. The most popular measure of connectivity is the sum of the crossing edges or *cut* between the partitions. The edges of the graph have weights that reflect similarities between the vertices, and the cut is defined to be the sum of the weights of the crossing edges. The minimum cut problem is solvable in polynomial time, however the graph clustering problem has additional (explicit or implicit) constraints on the sizes of the partitions. For example, the partitions might be constrained to be equal in size [48], or a weighted cut objective, such as ratio-cut [14] or normalized-cut [60] might need to be minimized. These constraints make the graph clustering problem NP-hard.

Graph clustering has been employed in many applications, for example, in circuit layout, partitioning the workload among processors for parallel processing, image segmentation, and of course, in the analysis of networks that arise in data mining. Since the problem is NP-hard, several heuristics are employed to try and solve this important problem. Some of the early successful approaches include the greedy search heuristic of Kernighan and Lin [48]. Another approach to graph clustering is based on the idea of graph flows. A survey of this research is presented in [52].

It turns out that the graph clustering objective may be written as a quadratic programming objective with discrete constraints on the variables. However, if the variables are relaxed to be real-valued, then eigenvectors of a symmetric matrix called the Graph Laplacian can be shown to exactly solve the relaxed quadratic problem. Consider the adjacency matrix A of an undirected graph which is defined to be a matrix with entries $A(i, j) = 1$ if there is an edge between vertices i and j and zero otherwise. The Graph Laplacian L equals $D - A$, where D is a diagonal matrix whose ith diagonal entry equals the sum of all entries in the ith row of A. Hence the sum of each row of L is zero, which implies that the vector of all 1's is an eigenvector of L with eigenvalue 0. The eigenvector of L corresponding to the next smallest eigenvalue is often called the Fiedler vector [29] and since it is orthogonal to the all 1's eigenvector, it has positive as well as negative entries. The graph bi-partition can now be obtained by placing the vertices with positive entries in the Fiedler vector in one partition, while the ones with negative entries are placed in the other partition. Such spectral methods have a long and rich history in graph clustering, dating back to the early 1970s [23,29,37], and have been used to optimize various weighted graph clustering objectives, such as ratio cut in circuit layout [14] and normalized cut in image segmentation [60]. One of the reasons for the success of spectral methods is that these methods provide a globally optimal solution to the relaxed problem, and thus provide a good global heuristic to solve the actual graph clustering problem.

Spectral graph clustering can be computationally expensive, especially when many eigenvectors of a large, sparse matrix need to be computed in order to directly give a k-way cut. An alternative class of algorithms is made up of multilevel graph clustering algorithms, which are attractive, efficient and powerful alternatives. In multilevel algorithms, the input graph is repeatedly coarsened level by level until only a small number of nodes remain. An initial clustering is performed on the coarsened graph, and

then this clustering is refined as the graph is uncoarsened level by level. These methods are extremely fast and give high-quality partitions. However, earlier multilevel methods, such as Metis [46] and Chaco [41], force clusters to be of nearly equal size, and are all based on optimizing the Kernighan-Lin objective [48]. In graph clustering for data mining problems, there is no reason why clusters should be of the same size. A recently developed multilevel graph clustering method, called GRACLUS (GRAph CLUStering) [22] removes the restriction of equal cluster sizes. In fact, this multilevel algorithm is able to optimize for several weighted spectral clustering objectives, such as ratio cut and normalized cut, without having to compute any eigenvectors. GRACLUS exploits a mathematical equivalence between two objectives – the seemingly different clustering objectives of weighted graph cuts, and weighted kernel k-means (an enhanced version of k-means) are shown to be equivalent in [22]. Using this equivalence, GRACLUS employs the weighted kernel k-means algorithm during the refinement phase to directly optimize various weighted graph cuts. Furthermore, this multilevel algorithm does not require any extra memory, which is in contrast to spectral methods that compute k eigenvectors for k-way cuts and require $O(Nk)$ storage, where N is the number of vertices. Thus multilevel algorithms are scalable to much larger data sets than standard spectral methods, and offer a state-of-the-art solution to the graph clustering problem.

Clusters as Dense Connected Components

In this section, a dataset X will be divided (with the exception of outliers) into dense connected components that will serve as its natural clusters. But how can the concepts of *density* and *connectivity* be defined for discrete data? Three approaches to answering this question exist: *Density-Based Partitioning*, *Density Functions*, and *Grid Methods*.

The first approach, *Density-Based Partitioning*, pins these concepts to a particular point $x \in X \subset A$.

Density and connectivity are defined in terms of a point's nearest neighbors. A cluster grows in any direction that density leads it to. Therefore, density-based algorithms are capable of discovering clusters of arbitrary shapes and they provide a natural protection against outliers. Figure 2 illustrates some cluster shapes that present problems for partitioning relocation clustering (e. g., k-means), but are handled properly by density-based algorithms. Finally, density-based algorithms are scalable.

These outstanding properties come along with certain inconveniences:

Knowledge Discovery: Clustering, Figure 2
Irregular shapes

- A dense cluster consisting of two adjacent areas with significantly different densities (both higher than a threshold) is not very informative
- Clusters lack interpretability
- Density-based partitioning primarily works with low-dimensional spatial data (though generalizations exist).

To explain the last point, note that scalability is achieved through usage of some sort of data index (such as an R^*-tree). Index construction is a topic of active research. Classic indices are effective only with low-dimensional data, since index complexity increases exponentially with data dimension. An excellent introduction to density-based methods is contained in [39].

As a representative of its class, consider the algorithm DBSCAN (Density Based Spatial Clustering of Applications with Noise) [27]. Two input parameters ϵ and MinPts are used to introduce:

1. An *ϵ-neighborhood* $N_\epsilon(x) = \{y \in X \mid \mathrm{dist}(x, y) \leq \epsilon\}$ of the point x,
2. A *core object*, which is a point with $|N_\epsilon(x)| \geq \mathrm{MinPts}$,
3. A notion of a point y *density-reachable* from a core object x (a sequence of core objects between x and y exists such that each object in the sequence belongs to an ϵ-neighborhood of its predecessor),
4. A definition of *density-connectivity* between two points x and y (they should be density-reachable from a common core object).

Density-connectivity is an equivalence relation. All the points reachable from core objects can be factorized into maximal connected components serving as clusters. The core points are *internal* points. The non-core points inside a cluster represent its *boundary*. The points not connected to any core point are declared to be outliers, and are not covered by any cluster. For low-dimensional spatial data, the theoretical complexity of DBSCAN is $O\left(N \log(N)\right)$. Experiments confirm a slightly super-linear runtime.

Selection of two parameters ϵ and MinPts presents a problem. Besides, no choice allows the fitting of data that

has different densities in different localities as frequently happens. Algorithms OPTICS (Ordering Points To Identify the Clustering Structure) [3] and DBCLASD (Distribution Based Clustering of Large Spatial Databases) [66] rectify DBSCAN to address these issues.

The second approaches to definition of density and connectivity is based on *Density Functions* and is due to Hinneburg and Keim [42]. They proposed the algorithm DENCLUE (DENsity-based CLUstEring). In this approach definitions are no longer pinned to a point $x \in X$, but to a point in the underlying attribute space $x \in A$. Data affects density indirectly through so-called density functions.

DENCLUE uses a *density function*

$$f^D(x) = \sum_{y \in D(x)} f(x, y)$$

that is the superposition of several *influence functions*, for example, $f(x, y) = \theta\left(\|x - y\| / \sigma\right)$ (equals to one, if the distance between x and y is less than or equal to σ) or Gaussian influence functions $f(x, y) = \exp(-\|x - y\|^2/\sigma^2)$. This provides a high level of generality: the first example leads to DBSCAN, and the second to k-means clustering. Both examples depend on the parameter σ.

Restricting the summation to $D(x) = \{y \in X : \|x - y\| < k\sigma\} \subset X$ enables a practical implementation. DENCLUE concentrates on local maxima of density function called *density-attractors* and uses a gradient hill-climbing technique to find them. In addition to *center-defined* clusters, *arbitrary-shape* clusters are defined as unions of local shapes along sequences of neighbors whose local densities are no less than a prescribed threshold.

Applications include high dimensional multimedia and molecular biology data. While no clustering algorithm could have less than $O(N)$ complexity, the runtime of DENCLUE scales with N sub-linearly! The explanation is that though all the points are fetched, the bulk of the analysis in the clustering stage involves only points in highly populated areas.

The third approach, *Grid-Based Methods*, suggests exploiting structure of the underlying attribute space A and partitioning it rather than data X. Data partitioning is induced by a point's membership in segments resulting from space partitioning, while space partitioning is based on grid-characteristics accumulated from input data. One advantage of this indirect handling (data $\rightarrow$ grid-data $\rightarrow$ space-partitioning $\rightarrow$ data-partitioning) is that it is independent of the data ordering and it works with numerical and categorical attributes (actually numerical attributes are binned).

To limit the amount of computations, multi-rectangular segments (direct Cartesian product of individual attribute sub-ranges) are considered. The elementary segment whose sides correspond to single-bins is called a *unit*. To some extent, the grid-based methodology reflects a technical point of view: it contains both partitioning and hierarchical algorithms and is used as an intermediate step in many clustering algorithms including the just described algorithm DENCLUE (that uses grids at its initial stage), and the important algorithm CLIQUE presented later.

Few other examples are considered below.

The algorithm BANG [58] summarizes data over the segments. The segments are stored in a grid-directory incorporating different scales. Adjacent segments are considered neighbors. If a common face has maximum dimension they are called nearest neighbors. More generally, neighbors of degree (dimension of a common face) between 0 and $d - 1$ can be defined. The density of a segment is defined as a ratio between the number of points in

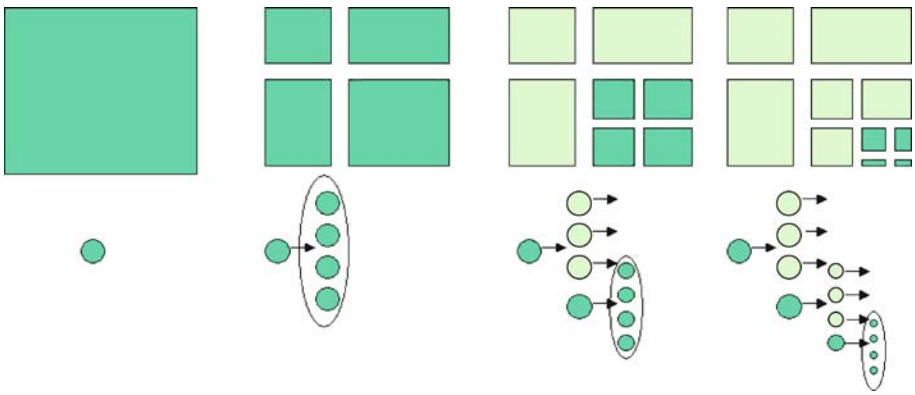

Knowledge Discovery: Clustering, Figure 3
Algorithm STING

that segment and its volume. From the grid-directory, a hierarchical clustering (a dendrogram) is calculated directly.

The algorithm STING (STatistical INformation Grid-based method) [65] works with numerical attributes (spatial data) and is designed to facilitate "region oriented" queries. STING assembles summary statistics in a hierarchical tree of nodes that are grid-cells. Figure 3 illustrates the proliferation of cells in 2D space and the construction of the corresponding tree. Each cell has four (default) children.

We want to briefly mention two other algorithms that work with numerical data and are based on certain mathematical ideas. The algorithm WaveCluster [59] supports a multi-resolution technique. It uses *wavelet* transforms and has several salient properties: (1) it can work with relatively high-dimensional data, (2) it successfully handles outliers, and (3) it has $O(N)$ complexity. The second algorithm, FC (Fractal Clustering) [6], utilizes another mathematical concept, *Hausdorff Fractal Dimension* (HFD). Its implementation relies on the hierarchy of grids: FC scans the full data incrementally attempting to add an incoming point to a cluster so as to minimally increase its HFD. The FC algorithm has its pros: (1) incremental structure, (2) suspendable nature, (3) ability to discover clusters of irregular shapes, (4) $O(N)$ complexity. Its cons: dependency on (1) data ordering, (2) on cluster initialization, and (3) on some input parameters.

Clustering Transactional Data

In this section we consider clustering of categorical data consisting of transactions, finite sets of elements (items) from a common item universe. Market basket data is typical transaction data. Co-occurrence of elements is the major idea in clustering transactional data. For example, one well-known similarity measure between two transactions is the Jaccard coefficient $\mathrm{sim}(T_1, T_2) = |T_1 \bigcap T_2| / |T_1 \bigcup T_2|$. Unfortunately, two random transactions rarely have elements in common.

We start with the algorithm CACTUS (Clustering Categorical Data Using Summaries), [30], that exemplifies the idea of co-occurrence. Values a, b of two different attributes are *strongly connected* if the number of transactions containing both a and b is larger than expected under an independence assumption by a user-defined margin α. This definition is extended to subsets A, B of two different attributes (each value pair $a \in A, b \in B$ has to be strongly connected), to segments (each 2D projection is strongly connected), and to the similarity of a pair of values of a single attribute via connectivity to other attributes. A cluster is defined as the maximal strongly connected segment (a Cartesian product of attribute subsets) having at least α times more elements than expected from the segment under the attribute independence assumption. Only two scans of data are required by this fast and scalable algorithm.

Han et al. exploited another idea [38] especially relevant when the item universe is large: (1) cluster items in item subsets C_j (*association rules* and *hyper-graph* machineries are used) and then (2) assign transactions T to a subset with highest similarity $\left|T \bigcap C_j\right| / \left|C_j\right|$.

Finally, consider the elegant algorithm STIRR (Sieving Through Iterated Reinforcement) [34] that deals with co-occurrence phenomenon for d-dimensional transactional records, *tuples* (e. g. tables of car sales with d fields). With each value (e. g. *Honda*) of one of d fields (e. g. *Manufacturer*) we associate a node v. Nodes v_1 and v_2 are close when many tuples containing them have a lot of co-occurring values (e. g., the highest sale month *August*). To formalize this concept, *STIRR* introduces a functional analysis concept of a *dynamical system*: with each node it associates a weight w_v and a transformation $w' = \Phi(w)$ over defined weights. This transformation computes w'_v as a sum of some *recombination* of weights w_b corresponding to the other $d-1$ values b over all tuples containing v (the overall weight is then renormalized). A fixed point $w = \Phi(w)$ can be achieved in several iterations. In addition to the idea of co-occurrence encapsulated in a weight propagation process, *STIRR* uses ideas of *spectral clustering* to perform the actual partitioning using non-principal eigenvectors.

Co-clustering

To explain the idea, consider text clustering, where each document is represented by a bag of its term frequencies (e. g. TF-IDF). In our usual case-by-attribute representation such textual data will result in a very large and sparse rectangular matrix with as many rows as number of documents and as many columns as number of terms, which is typically very high. Meanwhile, some terms are very similar to each other in the sense that they are used in the same subsets of documents. It would be handy to group such words together. Not only would a grouping of terms reduce dimensionality, but it would also eliminate certain randomness in preferential usage of a term from the same group by a particular document. In reality, this process is no different than merging documents that have similar term distributions. This is why the described idea is known as *co-clustering*.

For now, let us assume that we are dealing with a large sparse rectangular numerical non-negative data X. One

approach to co-clustering in text mining is based on minimal cuts partitioning for bipartite graphs using SVD [20]. Another popular approach is based on *distributional* or *informational* clustering. After a row normalization (non-negative elements sum to one), a data point can be viewed as a probability distribution (e. g. term distribution within a document). The same is true about every column attribute. Two attributes (two columns in matrix X) with exactly the same probability distributions are identical for the purpose of data mining, and so, one can be deleted. Attributes having distributions with small Kullback-Leibler (KL) distance (a common statistical measure to compare probability distributions) can still be grouped together without much impact on the information contained in the data. In addition, a natural derived attribute, the mixed distribution (a normalized sum of two columns), is now available to represent the group. This process can be generalized to more than two attributes. The grouping simplifies the original matrix X into the compressed form $\bar{X}$. In fact, our observation about impact on information can be quantified: information reduction $R = I(X) - I(\bar{X})$, where $I(X)$ is the mutual information contained in X[17], and is exactly equal to a sum of KL distances between original attributes assembled into a group and total group distribution (similarly to the objective function $E(C)$ used before in Sect. "Partitioning Relocation Clustering".

Now it is obvious that the intuitive idea of grouping similar attributes corresponds exactly to a minimization of information loss resulting in data compression. Moreover, R is symmetric with respect to columns and rows. So both dimensions can be gradually agglomerated under control of a single measure.

The above idea serves as a framework to several *co-clustering* developments, such as the *SimplifyRelation* algorithm [8] and *Information-Theoretic Co-clustering method* [21]. A natural question arises – how should co-clustering be done when the data matrix cannot be viewed as a probability distribution, for example, gene expression data that contains positive as well as negative entries. Recently, the theory and practice of co-clustering has been significantly advanced by extension to generalized distance measures known as *Bregman divergences* [5]. This advancement follows by using the important notion of *Bregman Information* and associating a co-clustering with a matrix approximation.

Scalability

Scalability challenges to clustering are much more severe than to predictive mining, with respect to computing time, to memory, and to the clustering concept itself in case of high dimensionality. We only reflect on a few of the most influential ideas without any references to many others. The first idea is to *squash* data into some summaries (*sufficient statistics*) [24] that, in the case of clustering, was first implemented in the BIRCH algorithm (Balanced Iterative Reduction and Clustering using Hierarchies) [70]. Many other developments (e. g. [11,31]) followed in its footsteps.

BIRCH pre-processes data into small tight micro-clusters, called *Cluster Features* (CF) accumulating zero, first, and second moments of CF members. CF are viewed as leaves in a height-balanced tree that resides in memory. They can later be used to build clusters via the algorithm of choice, for example hierarchical or k-means. The emphasis is shifted to a pre-processing phase of building a high-quality CF tree that is controlled by some parameters (as *branching factor*). When a tree reaches the assigned memory size, it is rebuilt with somewhat less tight CF. Outliers are saved in an auxiliary file. The overall complexity of BIRCH is $O(N)$ and it only takes one or two scans of data (which is not assumed to reside in memory).

Many algorithms (e. g. CLARANS) use old-fashioned sampling without rigorous statistical reasoning. *Sampling* was advanced to a new level with introduction of Hoeffding or Chernoff bounds [54]. In a nutshell, independently of the distribution for a real-valued random variable Y, $0 \le Y \le R$, the average of its n independent observations lies within ϵ of the actual mean

$$\left| \bar{Y} - \frac{1}{n} \sum_{j=1:n} Y_j \right| \le \epsilon$$

with probability $1 - \delta$ as soon as $\epsilon = \sqrt{R^2 \ln(1/\delta)/2n}$.

For example, these bounds are used in the clustering algorithm CURE [36] mentioned previously.

Another challenge is related to a situation when the dimensionality d is high.

Clustering in high-dimensional spaces presents two difficulties:

(1) *irrelevant attributes* – while in predictive mining they are easily discarded, irrelevant attributes can render the clustering task hopeless; their inspection in high dimension is nothing but easy;
(2) *curse of dimensionality* – in high-dimensional space, distance to a furthest point and to a closest point becomes on average undistinguishable; this is purely geometric phenomenon making overall proximity clustering very suspicious.

Some algorithms adapt to high dimensionality better than others. For example, the algorithm CACTUS adjusts well

because it defines a cluster only in terms of a cluster's 2D projections. However, as a rule of thumb, the performance of classic clustering algorithms can degrade gradually with dimension.

A proposed cure against *the curse of dimensionality* is *subspace clustering*. It is exemplified by the algorithm CLIQUE (Clustering In QUEst) [2]. CLIQUE works with high-dimensional numerical data and combines many ideas. It starts with the units (elementary rectangular cells) in low-dimensional subspaces. Only units whose densities exceed a threshold τ are retained (this is similar to grid-based clustering). A bottom-up approach of finding such units is applied, starting from units of dimension one. An inductive step to move from a dimension $q-1$ to a dimension q involves a self-joint over common $q-2$ faces. This is similar to the Apriori-reasoning in association rules. All the subspaces are sorted by their coverage and lesser-covered subspaces are pruned. A cut point between retained and pruned subspaces is selected based on the Minimum Description Length (MDL) principle. A cluster is defined as a maximal set of connected dense units. Effectively, CLIQUE selects several subspaces, each representing different perspective. The result is a series of cluster systems in different subspaces. Unfortunately, the result is not a partition, since the systems overlap.

There have been many follow-up improvements to subspace clustering. On a theoretical side, a more transparent criterion for subspace selection based on the concept of *entropy* was suggested in [16]. A more advanced inductive unit generation that, in addition, utilizes adaptive grids is used in the algorithm MAFIA (Merging of Adaptive Finite Intervals) [35]. MAFIA starts with one data pass to construct *adaptive grids* in each dimension. Many (1000) bins are used to compute histograms by reading blocks of data into memory. The bins are then merged to come up with a smaller number of adaptive variable-size bins than CLIQUE. The algorithm uses a parameter α, called the cluster dominance factor, to select bins that are α-times more densely populated than average. These variable-size bins are $q = 1$ candidate dense units (CDUs). Then MAFIA proceeds recursively to higher dimensions (every time a data scan is involved). Unlike CLIQUE, when constructing a new q-CDU, MAFIA tries two $(q-1)$-CDUs as soon as they share any (not only the first dimensions) $(q-2)$-face. This creates an order of magnitude more candidates. The algorithm PROCLUS (PROjected CLUstering) [1] explores pairs consisting of a data subset $C \subset X$ and a subspace in an attribute space A. A subset-subspace pair is called a projected cluster, if a projection of C onto the corresponding subspace is a tight cluster. Unlike with CLIQUE, projected clusters do not overlap.

Other Aspects of Clustering

Due to limited space, we have only touched on major clustering approaches. Now without any details we want to present a reader with a brief account of a few other clustering techniques.

The first important direction of thought is assessing clustering results. Different measures (e. g. *Silhouette* and *Partition* coefficients) are available, but not very useful. Usually assessment is done based purely on the objectives of the application. Another important research thread tries to answer the question of what is the optimal number k of clusters to build. Different statistical criteria have been developed such as BIC, AIC, MDL, and others. From the system complexity angle, clustering provides simplified data description. Given k cluster "prototypes" (e. g. centroids), data can be very roughly described by assigning each data point its cluster ID. Such very economical date compression is lossy, but in principle residuals between data points and their cluster centroids are tightly distributed and so takes relatively small amount of information to transmit. The more clusters k we introduce, the smaller the residuals will be, in the extreme each point coinciding with cluster centroid. However, more and more cluster centroids are needed to be transmitted upfront. So, the optimal k may be found from such considerations.

In real-world applications customers are rarely interested in unconstrained solutions. Clusters are frequently subjected to some problem-specific limitations that make them suitable for particular business actions. Finding clusters satisfying certain limitations is the subject of active research [40].

Consider, for example, clustering algorithm COD (Clustering with Obstructed Distance) [63]. The problem is easily illustrated by Fig. 4, where we show the difference in constructing three clusters in the absence of any obstacle (left) and in the presence of a river with a bridge (right).

Soft cluster assignments (assigning a point to several clusters with some probabilities) make sense if the k-means objective function is modified to incorporate "fuzzy errors" (similar to EM)

$$E'(C) = \sum_{i=1:N} \sum_{j=1:k} \left\| x_i - c_j \right\|^2 \omega_{ij}^2$$

The probabilities ω are defined based on Gaussian models. This makes the objective function differentiable with respect to means and allows application of general *gradient decent* methods presented in [53] in a context of

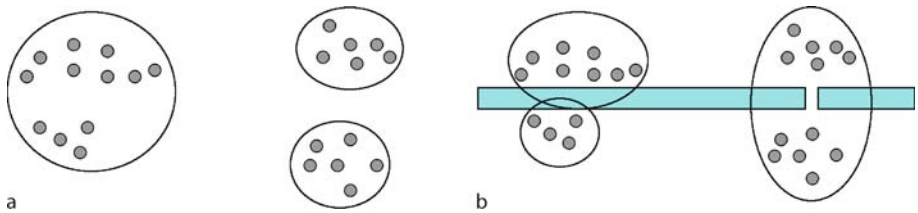

Knowledge Discovery: Clustering, Figure 4
Clustering with Obstructed Distance (COD) clustering. a No obstacles b River with the bridge

vector quantization. Another soft-assignment leads to so-called *harmonic means* [69]. Another iterative approach is used in *Self-Organizing Map* (SOM) [50].

Future Directions

Clustering is an important and well-studied problem. As discussed earlier, many algorithms exist that are applicable in different scenario. However, the theory of clustering is far from satisfactory. Even a simple and often used algorithm such as k-means does not have a performance guarantee. There have been some efforts to modify k-means, but none of these modified algorithms are simultaneously efficient and provably optimal. Thus a very important future direction of research is to develop a sound theory of clustering, that includes algorithms that have satisfactory performance guarantees. There have been a few efforts to axiomatize clustering, such as [49], however more such efforts are needed.

For the practitioner too, the variety of clustering algorithms is bewildering. An immensely beneficial tool will be an expert system for clustering. That is, a system that queries the user about various aspects of the data, the target application and the target clustering and then suggests appropriate clustering formulations and algorithms. Ideally, such a system would also provide a justification (relevant features) for the suggested algorithm. Besides an expert system, high quality public domain software would be invaluable.

Bibliography

1. Aggarwal CC, Wolf JL, Yu PS, Procopiuc C, Park JS (1999) Fast algorithms for projected clustering. In: Proceedings ACM SIGMOD International Conference on Management of Data, Philadelphia, pp 61–72
2. Agrawal R, Gehrke J, Gunopulos D, Raghavan P (1998) Automatic subspace clustering of high dimensional data for data mining applications. In: Proceedings of ACM SIGMOD, Seattle, pp 94–105
3. Ankerst M, Breunig MM, Kriegel HP, Sander J (1999) Optics: Ordering points to identify the clustering structure. In: Proceedings ACM SIGMOD International Conference on Management of Data, Philadelphia, pp 49–60
4. Banerjee A, Merugu S, Dhillon I, Ghosh J (2005) Clustering with Bregman divergences. J Mach Learn Res (JMLR) 6:1705–1749
5. Banerjee A, Dhillon I, Ghosh J, Merugu S, Modha D (2007) A generalized maximum entropy approach to Bregman co-clustering and matrix approximation. J Mach Learn Res(JMLR) 8:1919–1986
6. Barbara D, Chen P (2000) Using the fractal dimension to cluster datasets. In: Proceedings of the 6th ACM SIGKDD, Boston, pp 260–264
7. Berkhin P (2005) A survey of clustering data mining techniques. In: Kogan J, Nicholas C, Teboulle M (eds) Grouping Multidimensional Data. Springer
8. Berkhin P, Becher JD (2002) Learning simple relations: Theory and applications. In: Proceeding Second SIAM International Conference on Data Mining, Arlington, pp 420–436
9. Boley DL (1998) Principal direction divisive partitioning. Data Min Knowl Discov 2(4):325–344
10. Bradley P, Fayyad U (1998) Refining initial points for k-means clustering. In: Shavlik J (ed) Proceedings of the Fifteenth International Conference on Machine Learning (ICML). AAAI Press, San Francisco, pp 91–99
11. Bradley P, Fayyad U, Reina C (1998) Scaling clustering algorithms to large databases. In: Proceedings Fourth International Conference on Knowledge Discovery and Data Mining. AAAI Press, San Francisco
12. Cadez I, Gaffney S, Smyth P (2000) A general probabilistic framework for clustering individuals. Tech Rep UCI-ICS 00-09. University of California, Irvine
13. Cadez I, Smyth P, Mannila H (2001) Probabilistic modeling of transactional data with applications to profiling, visualization, and prediction. In: Proceedings of the 7th ACM SIGKDD, San Francisco, pp 37–46
14. Chan P, Schlag M, Zien J (1994) Spectral k-way ratio cut partitioning. IEEE Trans CAD-Integr Circuits Syst 13:1088–1096
15. Cheeseman P, Stutz J (1996) Bayesian classification (autoclass): Theory and results. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds) Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, pp 153–180
16. Cheng C, Fu A, Zhang Y (1999) Entropy-based subspace clustering for mining numerical data. In: Proceedings of the 5th ACM SIGKDD, San Diego, pp 84–93
17. Cover TM, Thomas JA (1991) Elements of Information Theory. Wiley, New York
18. Day W, Edelsbrunner H (1984) Efficient algorithms for agglomerative hierarchical clustering methods. J Classif 1(7):7–24

19. Defays D (1977) An efficient algorithm for a complete link method. Comp J 20:364–366
20. Dhillon IS (2001) Co-clustering documents and words using bipartite spectral graph partitioning. In: Proceedings of the The Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2001). Also appears as: UT CS Technical Report # TR 2001-05, March, 1999
21. Dhillon IS, Mallela S, Kumar R (2003) A divisive information-theoretic feature clustering algorithm for text classification. J Mach Learn Res (JMLR): Var Feature Sel Special Issue 3:1265–1287
22. Dhillon IS, Guan Y, Kulis B (2007) Weighted graph cuts without eigenvectors: A multilevel approach. IEEE Trans Pattern Anal Mach Intell 29(11):1944–1957
23. Donath WE, Hoffman AJ (1973) Lower bounds for the partitioning of graphs. IBM J Res Dev 17:422–425
24. DuMouchel W, Volinsky C, Johnson T, Cortes C, Pregibon D (1999) Squashing flat files flatter. In: Proceedings of the 5th ACM SIGKDD, San Diego, pp 6–15
25. Dubes R (1993) Cluster analysis and related issues. In: Chen C, Pau L, Wang P (eds) Handbook of Pattern Recognition and Computer Vision. World Scientific, River Edge, pp 3–32
26. Ester M, Kriegel HP, Xu X (1995) A database interface for clustering in large spatial databases. In: Proceedings of the 1st ACM SIGKDD, Montreal, pp 94–99
27. Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd ACM SIGKDD, Portland, pp 226–231
28. Everitt B (1993) Cluster Analysis, 3rd edn. Edward Arnold, London
29. Fiedler M (1975) A property of eigenvectors of non-negative symmetric matrices and its application to graph theory. Czech Math J 25:619–633
30. Ganti V, Gehrke J, Ramakrishnan R (1999) CACTUS-clustering categorical data using summaries. In: Proceedings of the 5th ACM SIGKDD, San Diego, pp 73–83
31. Ganti V, Ramakrishnan R, Gehrke J, Powell A, French J (1999) Clustering large datasets in arbitrary metric spaces. In: Proceedings of the 15th ICDE, Sydney, pp 502–511
32. Gersho A, Gray RM (1992) Vector quantization and signal compression. Kluwer, Dordrecht
33. Ghosh J (2003) Scalable clustering methods for data mining. In: Ye N (ed) Handbook of Data Mining. Lawrence Ealbaum Assoc, pp 247–277
34. Gibson D, Kleinberg J, Raghavan P (1998) Clustering categorical data: An approach based on dynamic systems. In: Proceedings of the 24th International Conference on Very Large Databases, New York, pp 311–323
35. Goil S, Nagesh H, Choudhary A (1999) MAFIA: Efficient and scalable subspace clustering for very large data sets. Tech Rep CPDC-TR-9906-010. Northwestern University, Evanston
36. Guha S, Rastogi R, Shim K (1998) CURE: A clustering algorithm for large databases. ACM SIGMOD International Conference on Management of Data, pp 73–84
37. Hall KM (1970) A r-dimensional quadratic placement algorithm. Manag Sci 11(3):219–229
38. Han EH, Karypis G, Kumar V, Mobasher B (1997) Clustering based on association rule hypergraphs. In: SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, (SIGMOD-DMKD'97)
39. Han J, Kamber M (2001) Data Mining: Concepts and Techniques. Morgan Kaufmann, San Mateo
40. Han J, Kamber M, Tung AKH (2001) Spatial clustering methods in data mining: A survey. In: Miller H, Han J (eds) Geographic Data Mining and Knowledge Discovery. Taylor and Francis
41. Hendrickson B, Leland R (1993) A multilevel algorithm for partitioning graphs. Tech Rep SAND93-1301, Sandia National Laboratories, Albuquerque
42. Hinneburg A, Keim D (1998) An efficient approach to clustering large multimedia databases with noise. In: Proceedings of the 4th ACM SIGKDD, New York, pp 58–65
43. Jain A, Murty M, Flynn PJ (1999) Data clustering: A review. ACM Comp Surv 31(3):264–323
44. Jain AK, Dubes RC (1988) Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs
45. Karypis G, Han EH, Kumar V (1999) CHAMELEON: Hierarchical clustering using dynamic modeling. IEEE Comput 32(8):68–75
46. Karypis G, Kumar V (1999) A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM J Sci Comput 20(1):359–392
47. Kaufman L, Rousseeuw P (1990) Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York
48. Kernighan B, Lin S (1970) An efficient heuristic procedure for partitioning graphs. Bell Syst Tech J 49(2):291–307
49. Kleinberg J (2002) An impossibility theorem for clustering. In: Proc Neural Info. Processing Systems (NIPS 15)
50. Kohonen T (1990) The self-organizing map. Proc IEEE 9:1464–1479
51. Lance G, Williams W (1967) A general theory of classification sorting strategies. Comp J 9:373–386
52. Leighton T, Rao S (1999) Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. J ACM 46(6):787–832
53. Marroquin J, Girosi F (1993) Some extensions of the k-means algorithm for image segmentation and pattern classification. Tech Rep AI Memo, MIT, Cambridge, p 1390
54. Motwani R, Raghavan P (1995) Randomized Algorithms. Cambridge University Press, Cambridge
55. Ng A, Jordan M, Weiss Y (2001) On spectral clustering: Analysis and an algorithm. In: Proc Neural Info Proc Syst (NIPS), Vancouver
56. Ng R, Han J (1994) Efficient and effective clustering methods for spatial data mining. In: Proc of the 20th Int Conf on Very Large Data Bases (VLDB), Santiago, pp 144–155
57. Parlett B (1997) The Symmetric Eigenvalue Problem, 2nd edn. SIAM, Philadelphia
58. Schikuta E, Erhart M (1997) The BANG-clustering system: grid-based data analysis. In: Proceeding of Advances in Intelligent Data Analysis, Reasoning about Data, 2nd International Symposium, London, pp 513–524
59. Sheikholeslami G, Chatterjee S, Zhang A (1998) Wavecluster: A multi-resolution clustering approach for very large spatial databases. In: Proceedings of the 24th Conference on VLDB, New York, pp 428–439
60. Shi J, Malik J (2000) Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intell 22(8):888–905
61. Sibson R (1973) SLINK: An optimally efficient algorithm for the single link cluster method. Comput J 16:30–34
62. Spath H (1980) Cluster Analysis Algorithms. Ellis Horwood, Chichester

63. Tung A, Hou J, Han J (2001) Spatial clustering in the presence of obstacles. In: Proceedings of the 17th ICDE, Heidelberg, pp 359–367
64. Voorhees EM (1986) Implementing agglomerative hierarchical clustering algorithms for use in document retrieval. Inf Proc Manag 22(6):465–476
65. Wang W, Yang J, Muntz R (1997) STING: a statistical information grid approach to spatial data mining. In: Proceedings of the 23rd Conference on VLDB, Athens, pp 186–195
66. Xu X, Ester M, Kriegel HP, Sander J (1998) A distribution-based clustering algorithm for mining large spatial datasets. In: Proceedings of the 14th ICDE, Orlando, pp 324–331
67. Yao A (1982) On constructing minimum spanning trees in k-dimensional space and related problems. SIAM J Comput 11(4):721–736
68. Yu SX, Shi J (2003) Multiclass spectral clustering. In: International Conference on Computer Vision, Nice
69. Zhang B (2001) Generalized k-harmonic means – dynamic weighting of data in unsupervised learning. In: Proceedings of the 1st SIAM International Conference on Data Mining, Chicago
70. Zhang T, Ramakrishnan R, Livny M (1997) BIRCH: A new data clustering algorithm and its applications. J Data Mining Knowl Discov 1(2):141–182

Kolmogorov–Arnold–Moser (KAM) Theory

Luigi Chierchia
Dipartimento di Matematica, Università "Roma Tre", Roma, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Kolmogorov Theorem
Arnold's Scheme
The Differentiable Case: Moser's Theorem
Future Directions
A The Classical Implicit Function Theorem
B Complementary Notes
Bibliography

Glossary

Action-angle variables A particular set of variables $(y, x) = ((y_1, \ldots, y_d), (x_1, \ldots, x_d))$, x_i ("angles") defined modulus 2π, particularly suited to describe the general behavior of an integrable system.

Fast convergent (Newton) method Super-exponential algorithms, mimicking Newton's method of tangents, used to solve differential problems involving small divisors.

Hamiltonian dynamics The dynamics generated by a Hamiltonian system on a symplectic manifold, i. e., on an even-dimensional manifold endowed with a symplectic structure.

Hamiltonian system A time reversible, conservative (without dissipation or expansion) dynamical system, which generalizes classical mechanical systems (solutions of Newton's equation $m_i \ddot{x}_i = f_i(x)$, with $1 \le i \le d$ and $f = (f_1, \ldots, f_d)$ a conservative force field); they are described by the flow of differential equations (i. e., the time t map associating to an initial condition, the solution of the initial value problem at time t) on a symplectic manifold and, locally, look like the flow associated with the system of differential equation $\dot{p} = -H_q(p, q)$, $\dot{q} = H_p(p, q)$ where $p = (p_1, \ldots, p_d)$, $q = (q_1, \ldots, q_d)$.

Integrable Hamiltonian systems A very special class of Hamiltonian systems, whose orbits are described by linear flows on the standard d-torus: $(y, x) \to (y, x + \omega t)$ where (y, x) are action-angle variables and t is time; the ω_i's are called the "frequencies" of the orbit.

Invariant tori Manifolds diffeomorphic to tori invariant for the flow of a differential equation (especially of Hamiltonian differential equations); establishing the existence of tori invariant for Hamiltonian flows is the main object of KAM theory.

KAM Acronym from the names of Kolmogorov (Andrey Nikolaevich Kolmogorov, 1903–1987), Arnold (Vladimir Igorevich Arnold, 1937) and Moser (Jürgen K. Moser, 1928–1999), whose results, in the 1950's and 1960's, in Hamiltonian dynamics, gave rise to the theory presented in this article.

Nearly-integrable Hamiltonian systems Hamiltonian systems which are small perturbations of an integrable system and which, in general, exhibit a much richer dynamics than the integrable limit. Nevertheless, KAM theory asserts that, under suitable assumptions, the majority (in the measurable sense) of the initial data of a nearly-integrable system behaves as in the integrable limit.

Quasi-periodic motions Trajectories (solutions of a system of differential equations), which are conjugate to linear flow on tori.

Small divisors/denominators Arbitrary small combinations of the form $\omega \cdot k := \sum_{j=1}^{d} \omega_i k_i$ with $\omega = (\omega_1, \ldots, \omega_d) \in \mathbb{R}^d$ a real vector and $k \in \mathbb{Z}^d$ an integer vector different from zero; these combinations arise in the denominators of certain expansions appearing in the perturbation theory of Hamiltonian systems, mak-

ing (when $d > 1$) convergent arguments very delicate. Physically, small divisors are related to "resonances", which are a typical feature of conservative systems.

Stability The property of orbits of having certain properties similar to a reference limit; more specifically, in the context of KAM theory, stability is normally referred to as the property of action variables of staying close to their initial values.

Symplectic structure A mathematical structure (a differentiable, non-degenerate, closed 2-form) apt to describe, in an abstract setting, the main geometrical features of conservative differential equations arising in mechanics.

Definition of the Subject

KAM theory is a mathematical, quantitative theory which has as its primary object the persistence, under small (Hamiltonian) perturbations, of typical trajectories of integrable Hamiltonian systems. In integrable systems with bounded motions, the typical trajectory is quasi-periodic, i. e., may be described through the linear flow $x \in \mathbb{T}^d \to x + \omega t \in \mathbb{T}^d$ where $\mathbb{T}^d$ denotes the standard d-dimensional torus (see Sect. "Introduction" below), t is time, and $\omega = (\omega_1, \ldots, \omega_d) \in \mathbb{R}^d$ is the set of frequencies of the trajectory (if $d = 1$, $2\pi/\omega$ is the *period* of the motion).

The main motivation for KAM theory is related to stability questions arising in celestial mechanics which were addressed by astronomers and mathematicians such as Kepler, Newton, Lagrange, Liouville, Delaunay, Weierstrass, and, from a more modern point of view, Poincaré, Birkhoff, Siegel, ...

The major breakthrough in this context, was due to Kolmogorov in 1954, followed by the fundamental work of Arnold and Moser in the early 1960s, who were able to overcome the formidable technical problem related to the appearance, in perturbative formulae, of arbitrarily small divisors[1]. Small divisors make the use of classical analytical tools (such as the standard Implicit Function Theorem, fixed point theorems, etc.) impossible and could be controlled only through a "fast convergent method" of Newton-type[2], which allowed, in view of the super-exponential rate of convergence, counterbalancing the divergences introduced by small divisors.

Actually, the main bulk of KAM theory is a set of *techniques* based, as mentioned, on fast convergent methods, and solving various questions in Hamiltonian (or generalizations of Hamiltonian) dynamics. By now, there are excellent reviews of KAM theory – especially Sect. 6.3 of [6] and [60] – which should complement the reading of this article, whose main objective is not to review but rather to explain the main fundamental ideas of KAM theory. To do this, we re-examine, in modern language, the main ideas introduced, respectively, by the founders of KAM theory, namely Kolmogorov (in Sect. "Kolmogorov Theorem"), Arnold (in Sect. "Arnold's Scheme") and Moser (Sect. "The Differentiable Case: Moser's Theorem").

In Sect. "Future Directions" we briefly and informally describe a few developments and applications of KAM theory: this section is by no means exhaustive and is meant to give a non technical, short introduction to some of the most important (in our opinion) extensions of the original contributions; for more detailed and complete reviews we recommend the above mentioned articles Sect. 6.3 of [6] and [60].

Appendix A contains a quantitative version of the classical Implicit Function Theorem.

A set of technical notes (such as notes 17, 18, 19, 21, 24, 26, 29, 30, 31, 34, 39), which the reader not particularly interested in technical mathematical arguments may skip, are collected in Appendix B and complete the mathematical expositions. Appendix B also includes several other complementary notes, which contain either standard material or further references or side comments.

Introduction

In this article we will be concerned with Hamiltonian flows on a symplectic manifold $(\mathcal{M}, \mathrm{d}y \wedge \mathrm{d}x)$; for general information, see, e. g., [5] or Sect. 1.3 of[6]. Notation, main definitions and a few important properties are listed in the following items.

(a) As symplectic manifold ("phase space") we shall consider $\mathcal{M} := B \times \mathbb{T}^d$ with $d \geq 2$ (the case $d = 1$ is trivial for the questions addressed in this article) where: B is an open, connected, bounded set in $\mathbb{R}^d$; $\mathbb{T}^d := \mathbb{R}^d/(2\pi\mathbb{Z}^d)$ is the standard flat d-dimensional torus with periods[3] 2π

(b) $\mathrm{d}y \wedge \mathrm{d}x := \sum_{i=1}^d \mathrm{d}y_i \wedge \mathrm{d}x_i$, $(y \in B,\ x \in \mathbb{T}^d)$ is the standard symplectic form[4]

(c) Given a real-analytic (or smooth) function $H\colon \mathcal{M} \to \mathbb{R}$, the *Hamiltonian flow governed by* H is the one-parameter family of diffeomorphisms $\phi_H^t\colon \mathcal{M} \to \mathcal{M}$, which to $z \in \mathcal{M}$ associates the solution at time t of the differential equation[5]

$$\dot{z} = J_{2d} \nabla H(z)\,, \quad z(0) = z\,, \tag{1}$$

where $\dot{z} = \mathrm{d}z/\mathrm{d}t$, J_{2d} is the standard symplectic $(2d \times 2d)$-matrix

$$J_{2d} = \begin{pmatrix} 0 & -\mathbb{1}_d \\ \mathbb{1}_d & 0 \end{pmatrix}\,,$$

$\mathbb{1}_d$ denotes the unit $(d \times d)$-matrix, 0 denotes a $(d \times d)$ block of zeros, and ∇ denotes gradient; in the symplectic coordinates $(y, x) \in B \times \mathbb{T}^d$, equations (1) reads

$$\begin{cases} \dot{y} = -H_x(y, x) \\ \dot{x} = H_y(y, x) \end{cases}, \qquad \begin{cases} y(0) = y \\ x(0) = x \end{cases} \tag{2}$$

Clearly, the flow ϕ_H^t is defined until $y(t)$ eventually reaches the border of B.

Equations (1) and (2) are called the *Hamilton's equations* with *Hamiltonian* H; usually, the symplectic (or "conjugate") variables (y, x) are called *action-angle* variables[6]; the number d (= half of the dimension of the phase space) is also referred to as "the number of degrees of freedom[7]".

The Hamiltonian H is constant over trajectories $\phi_H^t(z)$, as it follows immediately by differentiating $t \to H(\phi_H^t(z))$. The constant value $E = H(\phi_H^t(z))$ is called the energy of the trajectory $\phi_H^t(z)$.

Hamilton equations are left invariant by *symplectic* (or "canonical") change of variables, i. e., by diffeomorphisms on $\mathcal{M}$ which preserve the 2-form $\mathrm{d}y \wedge \mathrm{d}x$; i. e., if $\phi\colon (y, x) \in \mathcal{M} \to (\eta, \xi) = \phi(y, x) \in \mathcal{M}$ is a diffeomorphism such that $\mathrm{d}\eta \wedge \mathrm{d}\xi = \mathrm{d}y \wedge \mathrm{d}x$, then

$$\phi \circ \phi_H^t \circ \phi^{-1} = \phi_{H \circ \phi^{-1}}^t \,. \tag{3}$$

An equivalent condition for a map ϕ to be symplectic is that its Jacobian ϕ' is a *symplectic matrix*, i. e.,

$$\phi'^T J_{2d} \phi' = J_{2d} \tag{4}$$

where J_{2d} is the standard symplectic matrix introduced above and the superscript T denotes matrix transposition.

By a (generalization of a) theorem of Liouville, the Hamiltonian flow is symplectic, i. e., the map $(y, x) \to (\eta, \xi) = \phi_H^t(y, x)$ is symplectic for any H and any t; see Corollary 1.8, [6].

A classical way of producing symplectic transformations is by means of *generating functions*. For example, if $g(\eta, x)$ is a smooth function of $2d$ variables with

$$\det \frac{\partial^2 g}{\partial \eta \partial x} \neq 0 \,,$$

then, by the Implicit Function Theorem (IFT; see [36] or Sect. "A The Classical Implicit Function Theorem" below), the map $\phi\colon (y, x) \to (\eta, \xi)$ defined implicitly by the relations

$$y = \frac{\partial g}{\partial x} \,, \quad \xi = \frac{\partial g}{\partial \eta} \,,$$

yields a local symplectic diffeomorphism; in such a case, g is called the generating function of the transformation ϕ. For example, the function $\eta \cdot x$ is the generating function of the identity map.

For general information about symplectic changes of coordinates, generating functions and, in general, about symplectic structures we refer the reader to [5] or [6].

(d) A solution $z(t) = (y(t), x(t))$ of (2) is a *maximal quasi-periodic solution* with frequency vector $\omega = (\omega_1, \ldots, \omega_d) \in \mathbb{R}^d$ if ω is a rationally-independent vector, i. e.,

$$\exists n \in \mathbb{Z}^d \text{ s.t. } \omega \cdot n := \sum_{i=1}^d \omega_i n_i = 0 \implies n = 0 \,, \tag{5}$$

and if there exist smooth (periodic) functions $v, u\colon \mathbb{T}^d \to \mathbb{R}^d$ such that[8]

$$\begin{cases} y(t) = v(\omega t) \\ x(t) = \omega t + u(\omega t) \,. \end{cases} \tag{6}$$

(e) Let ω, u and v be as in the preceding item and let U and ϕ denote, respectively, the maps

$$\begin{cases} U\colon \theta \in \mathbb{T}^d \to U(\theta) := \theta + u(\theta) \in \mathbb{T}^d \\ \phi\colon \theta \in \mathbb{T}^d \to \phi(\theta) := (v(\theta), U(\theta)) \in \mathcal{M} \end{cases}$$

If U is a smooth diffeomorphism of $\mathbb{T}^d$ (so that, in particular[9] $\det U_\theta \neq 0$) then ϕ is an embedding of $\mathbb{T}^d$ into $\mathcal{M}$ and the set

$$\mathcal{T}_\omega = \mathcal{T}_\omega^d := \phi(\mathbb{T}^d) \tag{7}$$

is an embedded d-torus invariant for ϕ_H^t and on which the motion is conjugated to the linear (Kronecker) flow $\theta \to \theta + \omega t$, i. e.,

$$\phi^{-1} \circ \phi_H^t \circ \phi(\theta) = \theta + \omega t \,, \quad \forall \theta \in \mathbb{T}^d \,. \tag{8}$$

Furthermore, the invariant torus $\mathcal{T}_\omega$ is a graph over $\mathbb{T}^d$ and is *Lagrangian*, i. e., ($\mathcal{T}_\omega$ has dimension d and) the restriction of the symplectic form $\mathrm{d}y \wedge \mathrm{d}x$ on $\mathcal{T}_\omega$ vanishes[10].

(f) In KAM theory a major role is played by the numerical properties of the frequencies ω. A typical assumption is that ω is a (homogeneously) *Diophantine vector*: $\omega \in \mathbb{R}^d$ is called Diophantine or (κ, τ)-Diophantine if, for some constants $0 < \kappa \leq \min_i |\omega_i|$ and

$\tau \geq d-1$, it verifies the following inequalities:

$$|\omega \cdot n| \geq \frac{\kappa}{|n|^\tau}\,, \quad \forall n \in \mathbb{Z}^d \setminus \{0\}\,, \tag{9}$$

(normally, for integer vectors n, $|n|$ denotes $|n_1|+\cdots+|n_d|$, but other norms may well be used). We shall refer to κ and τ as the Diophantine constants of ω. The set of Diophantine numbers in $\mathbb{R}^d$ with constants κ and τ will be denoted by $\mathcal{D}^d_{\kappa,\tau}$; the union over all $\kappa > 0$ of $\mathcal{D}^d_{\kappa,\tau}$ will be denoted by $\mathcal{D}^d_\tau$ and the union over all $\tau \geq d-1$ of $\mathcal{D}^d_\tau$ will be denoted by $\mathcal{D}^d$. Basic facts about these sets are[11]: if $\tau < d-1$ then $\mathcal{D}^d_\tau = \emptyset$; if $\tau > d-1$ then the Lebesgue measure of $\mathbb{R}^d \setminus \mathcal{D}^d_\tau$ is zero; if $\tau = d-1$, the Lebesgue measure of $\mathcal{D}^d_\tau$ is zero but its intersection with any open set has the cardinality of $\mathbb{R}$.

(g) The tori $\mathcal{T}_\omega$ defined in (e) with $\omega \in \mathcal{D}^d$ will be called *maximal KAM tori* for H.

(h) A Hamiltonian function $(\eta,\xi) \in \mathcal{M} \to H(\eta,\xi)$ having a maximal KAM torus (or, more generally, a maximal invariant torus as in (e) with ω rationally independent) $\mathcal{T}_\omega$, can be put into the form[12]

$$K(y,x) := E + \omega \cdot y + Q(y,x)\,,$$
$$\text{with } \partial_y^\alpha Q(0,x) = 0, \quad \forall \alpha \in \mathbb{N}^d, \quad |\alpha| \leq 1\,; \tag{10}$$

compare, e. g., Sect. 1 of [59]; in the variables (y,x), the torus $\mathcal{T}_\omega$ is simply given by $\{y=0\} \times \mathbb{T}^d$ and E is its (constant) energy. A Hamiltonian in the form (10) is said to be in *Kolmogorov normal form*.
If

$$\det\langle \partial_y^2 Q(0,\cdot)\rangle \neq 0\,, \tag{11}$$

(where the brackets denote an average over $\mathbb{T}^d$ and ∂_y^2 the Hessian with respect to the y-variables) we shall say that the Kolmogorov normal form K in (10) is *non-degenerate*; similarly, we shall say that the KAM torus $\mathcal{T}_\omega$ for H is non-degenerate if H can be put in a non-degenerate Kolmogorov normal form.

Remark 1

(i) A classical theorem by H. Weyl says that the flow

$$\theta \in \mathbb{T}^d \to \theta + \omega t \in \mathbb{T}^d\,, \quad t \in \mathbb{R}$$

is dense (ergodic) in $\mathbb{T}^d$ if and only if $\omega \in \mathbb{R}^d$ is rationally independent (compare [6], Theorem 5.4 or Sect. 1.4 of [33]). Thus, trajectories on KAM tori fill them densely (i. e., pass in any neighborhood of any point).

(ii) In view of the preceding remark, it is easy to see that if ω is rationally independent, $(y(t),x(t))$ in (6) is a solution of (2) if and only if the functions v and u satisfy the following quasi-linear system of PDEs on $\mathbb{T}^d$:

$$\begin{cases} D_\omega v = -H_x(v(\theta), \theta + u(\theta)) \\ \omega + D_\omega u = H_y(v(\theta), \theta + u(\theta)) \end{cases} \tag{12}$$

where D_ω denotes the directional derivative $\omega \cdot \partial_\theta = \sum_{i=1}^d \omega_i \frac{\partial}{\partial \theta_i}$.

(iii) Probably, the main motivation for studying quasi-periodic solutions of Hamiltonian systems on $\mathbb{R}^d \times \mathbb{T}^d$ comes from perturbation theory of *nearly-integrable* Hamiltonian systems: a completely integrable system may be described by a Hamiltonian system on $\mathcal{M} := B(y_0,r) \times \mathbb{T}^d \subset \mathbb{R}^d \times \mathbb{T}^d$ with Hamiltonian $H = K(y)$ (compare Theorem 5.8, [6]); here $B(y_0,r)$ denotes the open ball $\{y \in \mathbb{R}^d : |y-y_0| < r\}$ centered at $y_0 \in \mathbb{R}^d$. In such a case the Hamiltonian flow is simply

$$\phi_K^t(y,x) = \big(y, x + \omega(y)t\big)\,,$$
$$\omega(y) := K_y(y) := \frac{\partial K}{\partial y}(y)\,. \tag{13}$$

Thus, if the "frequency map" $y \in B \to \omega(y)$ is a diffeomorphism (which is guaranteed if $\det K_{yy}(y_0) \neq 0$, for some $y_0 \in B$ and B is small enough), in view of (f), for almost all initial data, the trajectories (13) belong to maximal KAM tori $\{y\} \times \mathbb{T}^d$ with $\omega(y) \in \mathcal{D}^d$.
The main content of (classical) KAM theory, in our language, is that, *if the frequency map $\omega = K_y$ of a (real-analytic) integrable Hamiltonian $K(y)$ is a diffeomorphism, KAM tori persist under small (smooth enough) perturbations of K*; compare Remark 7–(iv) below.
The study of the dynamics generated by the flow of a one-parameter family of Hamiltonians of the form

$$K(y) + \varepsilon P(y,x;\varepsilon)\,, \quad 0 < \varepsilon \ll 1\,, \tag{14}$$

was called by H. Poincaré *le problème général de la dynamique*, to which he dedicated a large part of his monumental *Méthodes Nouvelles de la Mécanique Céleste* [49].

(iv) A big chapter in KAM theory, strongly motivated by applications to PDEs with Hamiltonian structure

(such as nonlinear wave equation, Schrödinger equation, KdV, etc.), is concerned with quasi-periodic solutions with $1 \le n < d$ frequencies, i. e., solutions of (2) of the form

$$\begin{cases} y(t) = v(\omega t) \\ x(t) = U(\omega t)\,, \end{cases} \tag{15}$$

where $v\colon \mathbb{T}^n \to \mathbb{R}^d$, $U\colon \mathbb{T}^n \to \mathbb{T}^d$ are smooth functions, $\omega \in \mathbb{R}^n$ is a rationally independent n-vector. Also in this case, if the map U is a diffeomorphism onto its image, the set

$$\mathcal{T}^n_\omega := \{(y,x) \in \mathcal{M}\colon \\ y = v(\theta)\,,\ x = U(\theta)\,,\ \theta \in \mathbb{T}^n\} \tag{16}$$

defines an invariant n-torus on which the flow ϕ^t_H acts by the linear translation $\theta \to \theta + \omega t$. Such tori are normally referred to as *lower dimensional tori*.

Even though this article will be mainly focused on "classical KAM theory" and on maximal KAM tori, we will briefly discuss lower dimensional tori in Sect. "Future Directions".

Kolmogorov Theorem

In the 1954 International Congress of Mathematicians, in Amsterdam, A.N. Kolmogorov announced the following fundamental (for the terminology, see (f), (g) and (h) above).

Theorem 1 (Kolmogorov [35]) *Consider a one-parameter family of real-analytic Hamiltonian functions on $\mathcal{M} := B(0,r) \times \mathbb{T}^d$ given by*

$$H := K + \varepsilon P \quad (\varepsilon \in \mathbb{R})\,, \tag{17}$$

where: (i) K is a non-degenerate Kolmogorov normal form (10)–(11); (ii) $\omega \in \mathcal{D}^d$ is Diophantine. Then, there exists $\varepsilon_0 > 0$ and for any $|\varepsilon| \le \varepsilon_0$ a real-analytic symplectic transformation $\phi_*\colon \mathcal{M}_* := B(0,r_*) \times \mathbb{T}^d \to \mathcal{M}$, for some $0 < r_* < r$, putting H in non-degenerate Kolmogorov normal form, $H \circ \phi_* = K_*$, with $K_* := E_* + \omega \cdot y' + Q_*(y', x')$. Furthermore[13], $\|\phi_* - \mathrm{id}\|_{C^1(\mathcal{M}_*)}$, $|E_* - E|$, and $\|Q_* - Q\|_{C^1(\mathcal{M}_*)}$ are small with ε.

Remark 2

(i) From Theorem 1 it follows that the torus

$$\mathcal{T}_{\omega,\varepsilon} := \phi_*(0, \mathbb{T}^d)$$

is a maximal non-degenerate KAM torus for H and the H-flow on $\mathcal{T}_{\omega,\varepsilon}$ is analytically conjugated (by ϕ_*) to the translation $x' \to x' + \omega t$ with the *same frequency* vector of $\mathcal{T}_{\omega,0} := \{0\} \times \mathbb{T}^d$, while the energy of $\mathcal{T}_{\omega,\varepsilon}$, namely E_*, is in general different from the energy E of $\mathcal{T}_{\omega,0}$. The idea of keeping the frequency fixed is a key idea introduced by Kolmogorov and its importance will be made clear in the analysis of the proof.

(ii) In fact, *the dependence upon ε is analytic* and therefore the torus $\mathcal{T}_{\omega,\varepsilon}$ is an analytic deformation of the unperturbed torus $\mathcal{T}_{\omega,0}$ (which is invariant for K); see Remark 7–(iii) below.

(iii) Actually, Kolmogorov not only stated the above result but gave also a precise outline of its proof, which is based on a *fast convergent "Newton" scheme*, as we shall see below; compare also [17].

The map ϕ_* is obtained as

$$\phi_* = \lim_{j\to\infty} \phi_1 \circ \cdots \circ \phi_j\,,$$

where the ϕ_j's are (ε-dependent) symplectic transformations of $\mathcal{M}$ closer and closer to the identity. It is enough to describe the construction of ϕ_1; ϕ_2 is then obtained by replacing $H_0 := H$ with $H_1 = H \circ \phi_1$ and so on.

We proceed to analyze the scheme of Kolmogorov's proof, which will be divided into three main steps.

Step 1: Kolmogorov Transformation

The map ϕ_1 is close to the identity and is generated by

$$g(y', x) := y' \cdot x + \varepsilon\left(b \cdot x + s(x) + y' \cdot a(x)\right)$$

where s and a are (respectively, scalar and vector-valued) real-analytic functions on $\mathbb{T}^d$ with zero average and $b \in \mathbb{R}^d$: setting

$$\begin{aligned} \beta_0 &= \beta_0(x) := b + s_x\,, \\ A &= A(x) := a_x \quad \text{and} \\ \beta &= \beta(y', x) := \beta_0 + Ay'\,, \end{aligned} \tag{18}$$

($s_x = \partial_x s = (s_{x_1}, \ldots, s_{x_d})$ and a_x denotes the matrix $(a_x)_{ij} := \frac{\partial a_i}{\partial x_j}$) ϕ_1 is implicitly defined by

$$\begin{cases} y = y' + \varepsilon\beta(y', x) := y' + \varepsilon(\beta_0(x) + A(x)y') \\ x' = x + \varepsilon a(x)\,. \end{cases} \tag{19}$$

Thus, for ε small, $x \in \mathbb{T}^d \to x + \varepsilon a(x) \in \mathbb{T}^d$ defines a diffeomorphism of $\mathbb{T}^d$ with inverse

$$x = \varphi(x') := x' + \varepsilon\alpha(x'; \varepsilon)\,, \tag{20}$$

for a suitable real-analytic function α, and ϕ_1 is explicitly given by

$$\phi_1 : (y', x') \to \begin{cases} y = y' + \varepsilon\beta\left(y', \varphi(x')\right) \\ x = \varphi(x') \,. \end{cases} \tag{21}$$

Remark 3

(i) Kolmogorov transformation ϕ_1 is actually the composition of two "elementary" symplectic transformations: $\phi_1 = \phi_1^{(1)} \circ \phi_1^{(2)}$ where $\phi_1^{(2)} : (y', x') \to (\eta, \xi)$ is the symplectic lift of the $\mathbb{T}^d$-diffeomorphism given by $x' = \xi + \varepsilon a(\xi)$ (i. e., $\phi_1^{(2)}$ is the symplectic map generated by $y' \cdot \xi + \varepsilon y' \cdot a(\xi)$), while $\phi_1^{(1)} : (\eta, \xi) \to (y, x)$ is the angle-dependent action translation generated by $\eta \cdot x + \varepsilon(b \cdot x + s(x))$; $\phi_1^{(2)}$ acts in the "angle direction" and will be needed to straighten out the flow up to order $O(\varepsilon^2)$, while $\phi_1^{(1)}$ acts in the "action direction" and will be needed to keep the frequency of the torus fixed.

(ii) The inverse of ϕ_1 has the form

$$(y, x) \to \begin{cases} y' = M(x)y + c(x) \\ x' = \phi(x) \end{cases} \tag{22}$$

with M a $(d \times d)$-invertible matrix and ϕ a diffeomorphism of $\mathbb{T}^d$ (in the present case $M = (\mathbb{1}_d + \varepsilon A(x))^{-1} = \mathbb{1}_d + O(\varepsilon)$ and $\phi = \mathrm{id} + \varepsilon a$) and it is easy to see that the symplectic diffeomorphisms of the form (22) form a subgroup of the symplectic diffeomorphisms, which we shall call *the group of Kolmogorov transformations*.

Determination of Kolmogorov transformation Following Kolmogorov, we now try to determine b, s and a so that the "new Hamiltonian" (better: "the Hamiltonian in the new symplectic variables") takes the form

$$H_1 := H \circ \phi_1 = K_1 + \varepsilon^2 P_1 \,, \tag{23}$$

with K_1 in the Kolmogorov normal form

$$K_1 = E_1 + \omega \cdot y' + Q_1(y', x') \,, \quad Q_1 = O(|y'|^2) \,. \tag{24}$$

To proceed we insert $y = y' + \varepsilon\beta(y', x)$ into H and, after some elementary algebra and using Taylor formula, we find[14]

$$\begin{aligned} H(y' + \varepsilon\beta, x) &= E + \omega \cdot y' + Q(y', x) \\ &\quad + \varepsilon Q'(y', x) + \varepsilon F'(y', x) + \varepsilon^2 P'(y', x) \end{aligned} \tag{25}$$

where, defining

$$\begin{cases} Q^{(1)} := Q_y(y', x) \cdot (a_x y') \\ Q^{(2)} := [Q_y(y', x) - Q_{yy}(0, x)y'] \cdot \beta_0 \\ \qquad = \displaystyle\int_0^1 (1-t) Q_{yyy}(ty', x)y' \cdot y' \cdot \beta_0 \mathrm{d}t \\ Q^{(3)} := P(y', x) - P(0, x) - P_y(0, x)y' \\ \qquad = \displaystyle\int_0^1 (1-t) P_{yy}(ty', x)y' \cdot y' \mathrm{d}t \\ P^{(1)} := \dfrac{1}{\varepsilon^2}[Q(y' + \varepsilon\beta, x) - Q(y', x) \\ \qquad\qquad - \varepsilon Q_y(y', x) \cdot \beta] \\ \qquad = \displaystyle\int_0^1 (1-t) Q_{yy}(y' + t\varepsilon\beta, x)\beta \cdot \beta \mathrm{d}t \\ P^{(2)} := \dfrac{1}{\varepsilon}[P(y' + \varepsilon\beta, x) - P(y', x)] \\ \qquad = \displaystyle\int_0^1 P_y(y' + t\varepsilon\beta, x) \cdot \beta \mathrm{d}t \,, \end{cases} \tag{26}$$

(recall that $Q_y(0, x) = 0$) and denoting the ω-directional derivative

$$D_\omega := \sum_{j=1}^d \omega_j \frac{\partial}{\partial x_j}$$

one sees that $Q' = Q'(y', x)$, $F' = F'(y', x)$ and $P' = P'(y', x)$ are given by, respectively

$$\begin{cases} Q'(y', x) := Q^{(1)} + Q^{(2)} + Q^{(3)} = O(|y'|^2) \\ F'(y', x) := \omega \cdot b + D_\omega s + P(0, x) \\ \qquad + \left\{D_\omega a + Q_{yy}(0, x)\beta_0 + P_y(0, x)\right\} \cdot y' \\ P' := P^{(1)} + P^{(2)} \,, \end{cases} \tag{27}$$

where $D_\omega a$ is the vector function with kth entry $\sum_{j=1}^d \omega_j \frac{\partial a_k}{\partial x_j}$; $D_\omega a \cdot y' = \omega \cdot (a_x y') = \sum_{j,k=1}^d \omega_j \frac{\partial a_k}{\partial x_j} y'_k$; recall, also, that $Q = O(|y|^2)$ so that $Q_y = O(y)$ and $Q' = O(|y'|^2)$.

Notice that, as an intermediate step, we are considering H as a function of mixed variables y' and x (and this causes no problem, as it will be clear along the proof).

Thus, recalling that x is related to x' by the (y'-independent) diffeomorphism $x = x' + \varepsilon\alpha(x'; \varepsilon)$ in (21), we see that in order to achieve relations (23)–(24), we have to determine b, s and a so that

$$F'(y', x) = \text{const} \,. \tag{28}$$

Remark 4

(i) F' is a first degree polynomial in y' so that (28) is equivalent to

$$\begin{cases} \omega \cdot b + D_\omega s + P(0,x) = \text{const}\,, \\ D_\omega a + Q_{yy}(0,x)\beta_0 + P_y(0,x) = 0\,. \end{cases} \tag{29}$$

Indeed, the second equation is necessary to keep the torus frequency fixed and equal to ω (which, as we shall see in more detail later, is a key ingredient introduced by Kolmogorov).

(ii) In solving (28) or (29), we shall encounter differential equations of the form

$$D_\omega u = f\,, \tag{30}$$

for some given function f real-analytic on $\mathbb{T}^d$. Taking the average over $\mathbb{T}^d$ shows that $\langle f\rangle = 0$, and we see that (30) can be solved only if f has vanishing mean value

$$\langle f\rangle = f_0 = 0\,;$$

in such a case, expanding in Fourier series[15], one sees that (30) is equivalent to

$$\sum_{\substack{n\in\mathbb{Z}^d\\ n\neq 0}} i\omega\cdot n u_n e^{in\cdot x} = \sum_{\substack{n\in\mathbb{Z}^d\\ n\neq 0}} f_n e^{in\cdot x}\,, \tag{31}$$

so that the solutions of (30) are given by

$$u = u_0 + \sum_{\substack{n\in\mathbb{Z}^d\\ n\neq 0}} \frac{f_n}{i\omega\cdot n} e^{in\cdot x}\,, \tag{32}$$

for an arbitrary u_0. Recall that for a continuous function$f$ over $\mathbb{T}^d$ to be analytic it is necessary and sufficient that its Fourier coefficients f_n decay exponentially fast in n, i. e., that there exist positive constants M and ξ such that

$$|f_n| \le M e^{-\xi|n|}\,, \quad \forall n\,. \tag{33}$$

Now, since $\omega \in \mathcal{D}^d_{\kappa,\tau}$ one has that (for $n \neq 0$)

$$\frac{1}{|\omega\cdot n|} \le \frac{|n|^\tau}{\kappa} \tag{34}$$

and one sees that if f is analytic so is u in (32) (although the decay constants of u will be different from those of f; see below)

Summarizing, *if f is real-analytic on $\mathbb{T}^d$ and has vanishing mean value f_0, then there exists a unique real-analytic solution of (30) with vanishing mean value, which is given by*

$$D_\omega^{-1} f := \sum_{\substack{n\in\mathbb{Z}^d\\ n\neq 0}} \frac{f_n}{i\omega\cdot n} e^{in\cdot x}\,; \tag{35}$$

all other solutions of (30) are obtained by adding an arbitrary constant to $D_\omega^{-1} f$ as in (32) with u_0 arbitrary.

Taking the average of the first relation in (29), we may the determine the value of the constant denoted const, namely,

$$\text{const} = \omega\cdot b + P_0(0) := \omega\cdot b + \langle P(0,\cdot)\rangle\,. \tag{36}$$

Thus, by (ii) of Remark 4, we see see that

$$s = -D_\omega^{-1}\left(P(0,x) - P_0(0)\right) = -\sum_{\substack{n\in\mathbb{Z}^d\\ n\neq 0}} \frac{P_n(0)}{i\omega\cdot n} e^{in\cdot x}\,, \tag{37}$$

where $P_n(0)$ denote the Fourier coefficients of $x \to P(0,x)$; indeed s is determined only up to a constant by the relation in (29) but *we select the zero-average solution*. Thus, s has been completely determined.

To solve the second (vector) equation in (29) we first have to require that the left hand side (l.h.s.) has vanishing mean value, i. e., recalling that $\beta_0 = b + s_x$ (see (18)), we must have

$$\langle Q_{yy}(0,\cdot)\rangle b + \langle Q_{yy}(0,\cdot)s_x\rangle + \langle P_y(0,\cdot)\rangle = 0\,. \tag{38}$$

In view of (11) this relation is equivalent to

$$b = -\langle Q_{yy}(0,\cdot)\rangle^{-1}\left(\langle Q_{yy}(0,\cdot)s_x\rangle + \langle P_y(0,\cdot)\rangle\right)\,, \tag{39}$$

which uniquely determines b. Thus β_0 is completely determined and the l.h.s. of the second equation in (29) has average zero; thus its unique zero-average solution (again zero-average of a is required as a normalization condition) is given by

$$a = -D_\omega^{-1}\left(Q_{yy}(0,x)\beta_0 + P_y(0,x)\right)\,. \tag{40}$$

Finally, if $\varphi(x') = x' + \varepsilon\alpha(x';\varepsilon)$ is the inverse diffeomorphism of $x \to x + \varepsilon a(x)$ (compare (20)), then, by Taylor's formula,

$$Q(y',\varphi(x')) = Q(y',x') + \varepsilon\int_0^1 Q_x(y', x' + \varepsilon\alpha t)\cdot\alpha\,\mathrm{d}t\,.$$

In conclusion, we have

Proposition 1 *If ϕ_1 is defined in (19)–(18) with s, b and a given in (37), (39) and (40) respectively, then (23) holds with*

$$\begin{cases} E_1 := E + \varepsilon\widetilde{E} \\ \qquad \widetilde{E} := \omega \cdot b + P_0(0) \\ Q_1(y', x') := Q(y', x') + \varepsilon\widetilde{Q}(y', x') \\ \qquad \widetilde{Q} := \displaystyle\int_0^1 Q_x(y', x' + t\varepsilon\alpha) \cdot \alpha dt + Q'(y', \varphi(x')) \\ P_1(y', x') := P'(y', \varphi(x')) \end{cases} \tag{41}$$

with Q' and P' defined in (26), (27) and φ in (20).

Remark 5 The main technical problem is now transparent: because of the appearance of the *small divisors* $\omega \cdot n$ (which may become arbitrarily small), the solution $D_\omega^{-1} f$ is *less regular* than f so that the approximation scheme cannot work on a fixed function space. To overcome this fundamental problem – which even Poincaré was unable to solve notwithstanding his enormous efforts (see, e. g., [49]) – three ingredients are necessary:

(i) To set up a Newton scheme: this step has just been performed and it has been summarized in the above Proposition 1; such schemes have the following features: they are "quadratic" and, furthermore, after one step one has reproduced the initial situation (i. e., the form of H_1 in (23) has the same properties of H_0). It is important to notice that the new perturbation $\varepsilon^2 P_1$ is proportional to the *square* ε; thus, if one could iterate j times, at the jth step, would find

$$H_j = H_{j-1} \circ \phi_j = K_j + \varepsilon^{2^j} P_j \,. \tag{42}$$

The appearance of the exponential of the exponential of ε justifies the term "super-convergence" used, sometimes, in connection with Newton schemes.

(ii) One needs to introduce a *scale of Banach function spaces* $\{\mathcal{B}_\xi : \xi > 0\}$ with the property that $\mathcal{B}_{\xi'} \subset \mathcal{B}_\xi$ when $\xi < \xi'$: the generating functions ϕ_j will belong to $\mathcal{B}_{\xi_j}$ for a suitable decreasing sequence ξ_j;

(iii) One needs to control the small divisors at each step and this is granted by Kolmogorov's idea of keeping the frequency *fixed* in the normal form so that one can systematically use the Diophantine estimate (9).

Kolmogorov in his paper very neatly explained steps (i) and (iii) but did not provide the details for step (ii); in this regard he added: "*Only the use of condition* (9) *for proving the convergence of the recursions, ϕ_j, to the analytic limit for the recursion ϕ_* is somewhat more subtle*". In the next paragraph we shall introduce classical Banach spaces and discuss the needed straightforward estimates.

Step 2: Estimates

For $\xi \le 1$, we denote by $\mathcal{B}_\xi$ the space of function $f\colon B(0,\xi) \times \mathbb{T}^d \to \mathbb{R}$ analytic on

$$W_\xi := D(0,\xi) \times \mathbb{T}_\xi^d \,, \tag{43}$$

where

$$\begin{aligned} D(0,\xi) &:= \{y \in \mathbb{C}^d : |y| < \xi\} \quad \text{and} \\ \mathbb{T}_\xi^d &:= \{x \in \mathbb{C}^d : |\mathrm{Im} x_j| < \xi\}/(2\pi\mathbb{Z}^d) \end{aligned} \tag{44}$$

with finite sup-norm

$$\|f\|_\xi := \sup_{D(0,\xi)\times\mathbb{T}_\xi^d} |f| \,, \tag{45}$$

(in other words, $\mathbb{T}_\xi^d$ denotes the complex points x with real parts $\mathrm{Re} x_j$ defined modulus 2π and imaginary part $\mathrm{Im} x_j$ with absolute value less than ξ).

The following properties are elementary:

(P1) $\mathcal{B}_\xi$ equipped with the $\|\cdot\|_\xi$ norm is a Banach space;

(P2) $\mathcal{B}_{\xi'} \subset \mathcal{B}_\xi$ when $\xi < \xi'$ and $\|f\|_\xi \le \|f\|_{\xi'}$ for any $f \in \mathcal{B}_{\xi'}$;

(P3) if $f \in \mathcal{B}_\xi$, and $f_n(y)$ denotes the n-Fourier coefficient of the periodic function $x \to f(y,x)$, then

$$|f_n(y)| \le \|f\|_\xi \mathrm{e}^{-|n|\xi} \,, \quad \forall n \in \mathbb{Z}^d, \quad \forall y \in D(0,\xi) \,. \tag{46}$$

Another elementary property, which together with **(P3)** may found in any book of complex variables (e. g., [1]), is the following "Cauchy estimate" (which is based on Cauchy's integral formula):

(P4) let $f \in \mathcal{B}_\xi$ and let $p \in \mathbb{N}$ then there exists a constant $B_p = B_p(d) \ge 1$ such that, for any multi-index $(\alpha, \beta) \in \mathbb{N}^d \times \mathbb{N}^d$ with $|\alpha| + |\beta| \le p$ (as above for integer vectors α, $|\alpha| = \sum_j |\alpha_j|$) and for any $0 \le \xi' < \xi$ one has

$$\|\partial_y^\alpha \partial_x^\beta f\|_{\xi'} \le B_p \|f\|_\xi (\xi - \xi')^{-(|\alpha|+|\beta|)} \,. \tag{47}$$

Finally, we shall need estimates on $D_\omega^{-1} f$, i. e., on solutions of (30):

(P5) Assume that $x \to f(x) \in \mathcal{B}_\xi$ has a zero average (all above definitions may be easily adapted to functions depending only on x); assume that $\omega \in \mathcal{D}_{\kappa,\tau}^d$ (recall Sect. "Introduction", point (f)), and let $p \in \mathbb{N}$. Then, there exist constants $\bar{B}_p = \bar{B}_p(d,\tau) \ge 1$ and $k_p = k_p(d,\tau) \ge 1$ such that, for any multi-index $\beta \in \mathbb{N}^d$ with $|\beta| \le p$ and for any $0 \le \xi' < \xi$ one has

$$\|\partial_x^\beta D_\omega^{-1} f\|_{\xi'} \le \bar{B}_p \frac{\|f\|_\xi}{\kappa} (\xi - \xi')^{-k_p} \,. \tag{48}$$

Remark 6

(i) A proof of (48) is easily obtained observing that by (35) and (46), calling $\delta := \xi - \xi'$, one has

$$\begin{aligned}\|\partial_x^\beta D_\omega^{-1} f\|_{\xi'} &\le \sum_{\substack{n\in\mathbb{Z}^d\\ n\neq 0}} \frac{|n|^{|\beta|}|f_n|}{|\omega\cdot n|} e^{\xi'|n|} \\ &\le \|f\|_\xi \sum_{\substack{n\in\mathbb{Z}^d\\ n\neq 0}} \frac{|n|^{|\beta|+\tau}}{\kappa} e^{-\delta|n|} \\ &= \frac{\|f\|_\xi}{\kappa}\delta^{-(|\beta|+\tau+d)} \\ &\quad\cdot \sum_{\substack{n\in\mathbb{Z}^d\\ n\neq 0}} [\delta|n|]^{|\beta|+\tau} e^{-\delta|n|}\delta^d \\ &\le \mathrm{const}\frac{\|f\|_\xi}{\kappa}(\xi-\xi')^{-(|\beta|+\tau+d)}\,,\end{aligned}$$

where the last estimate comes from approximating the sum with the Riemann integral

$$\int_{\mathbb{R}^d} |y|^{|\beta|+\tau} e^{-|y|} dy\,.$$

More surprising (and much more subtle) is that (48) holds with $k_p = |\beta| + \tau$; such an estimate has been obtained by Rüssmann [54,55]. For other explicit estimates see, e. g., [11] or [12].

(ii) If $|\beta| > 0$ it is not necessary to assume that $\langle f\rangle = 0$.

(iii) Other norms may be used (and, sometimes, are more useful); for example, rather popular are Fourier norms

$$\|f\|'_\xi := \sum_{n\in\mathbb{Z}^d} |f_n| e^{\xi|n|}\,; \tag{49}$$

see, e. g., [13] and references therein.

By the hypotheses of Theorem 1 it follows that there exist $0 < \xi \le 1$, $\kappa > 0$ and $\tau \ge d-1$ such that $H \in \mathcal{B}_\xi$ and $\omega \in \mathcal{D}^d_{\kappa,\tau}$. Denote

$$T := \langle Q_{yy}(0,\cdot)\rangle^{-1}\,, \quad M := \|P\|_\xi\,. \tag{50}$$

and let $C > 1$ be a constant such that[16]

$$|E|, |\omega|, \|Q\|_\xi, \|T\| < C \tag{51}$$

(i. e., each term on the l.h.s. is bounded by the r.h.s.); finally, fix

$$0 < \delta < \xi \quad \text{and define } \bar\xi := \xi - \frac{2}{3}\delta\,,\ \xi' := \xi - \delta\,. \tag{52}$$

The parameter ξ' will be the size of the domain of analyticity of the new symplectic variables (y', x'), domain on which we shall bound the Hamiltonian $H_1 = H \circ \phi_1$, while $\bar\xi$ is an intermediate domain where we shall bound various functions of y' and x.

*By **(P4)** and **(P5)**, it follows that there exist constants $\bar c = \bar c(d,\tau,\kappa) > 1$, $\bar\mu \in \mathbb{Z}_+$ and $\bar\nu = \bar\nu(d,\tau) > 1$ such that*[17]

$$\begin{cases} \|s_x\|_{\bar\xi}, |b|, |\widetilde{E}|, \|a\|_{\bar\xi}, \|a_x\|_{\bar\xi}, \|\beta_0\|_{\bar\xi}, \|\beta\|_{\bar\xi}, \\ \qquad \|Q'\|_{\bar\xi}, \|\partial^2_{y'} Q'(0,\cdot)\|_0 \le \bar c C^{\bar\mu}\delta^{-\bar\nu} M =: \bar L\,, \\ \|P'\|_{\bar\xi} \le \bar c C^{\bar\mu}\delta^{-\bar\nu} M^2 = \bar L M\,. \end{cases} \tag{53}$$

The estimate in the first line of (53) allows us to construct, for ε small enough, the symplectic transformation ϕ_1, whose main properties are collected in the following

Lemma 1 *If $|\varepsilon| \le \varepsilon_0$ and ε_0 satisfies*

$$\varepsilon_0 \bar L \le \frac{\delta}{3}\,, \tag{54}$$

then the map $\psi_\varepsilon(x) := x + \varepsilon a(x)$ has an analytic inverse $\varphi(x') = x' + \varepsilon\alpha(x';\varepsilon)$ such that, for all $|\varepsilon| < \varepsilon_0$,

$$\|\alpha\|_{\xi'} \le \bar L \quad \textit{and} \quad \varphi = \mathrm{id} + \varepsilon\alpha\colon \mathbb{T}^d_{\xi'} \to \mathbb{T}^d_{\bar\xi}\,. \tag{55}$$

Furthermore, for any $(y', x) \in W_{\bar\xi}$, $|y' + \varepsilon\beta(y', x)| < \xi$, so that

$$\begin{aligned}&\phi_1 = \big(y' + \varepsilon\beta(y', \varphi(x')), \varphi(x')\big) : W_{\xi'} \to W_\xi\,, \quad \textit{and} \\ &\|\phi_1 - \mathrm{id}\|_{\xi'} \le |\varepsilon|\bar L\,;\end{aligned} \tag{56}$$

finally, the matrix $\mathbb{1}_d + \varepsilon a_x$ is, for any $x \in \mathbb{T}^d_{\bar\xi}$, invertible with inverse $\mathbb{1}_d + \varepsilon S(x;\varepsilon)$ satisfying

$$\|S\|_{\bar\xi} \le \frac{\|a_x\|_{\bar\xi}}{1 - |\varepsilon|\|a_x\|_{\bar\xi}} < \frac{3}{2}\bar L\,, \tag{57}$$

so that ϕ_1 defines a symplectic diffeomorphism.

The simple proof[18] of this statement is based upon standard tools in mathematical analysis such as the contraction mapping theorem or the inversion of close-to-identity matrices by Neumann series (see, e. g., [36]).

From the Lemma and the definition of P_1 in (41), it follows immediately that

$$\|P_1\|_{\xi'} \le \bar L\,. \tag{58}$$

Next, by the same technique used to derive (53), one can easily check that

$$\|\widetilde{Q}\|_{\xi'}\,,\ \ 2C^2\|\partial^2_{y'}\widetilde{Q}(0,\cdot)\|_0 \le cC^\mu\delta^{-\nu}M = L\,, \tag{59}$$

for suitable constants $c \geq \bar{c}$, $\bar{\mu} \geq \mu$, $\bar{\nu} \geq \nu$ (the factor $2C^2$ has been introduced for later convenience; notice also that $L \geq \bar{L}$). But, then, if

$$\varepsilon_0 L := \varepsilon_0 c C^\mu \delta^{-\nu} M \leq \frac{\delta}{3} , \tag{60}$$

there follows that[19] $\|\widetilde{T}\| \leq L$; this bound, together with (53), (59), (56), and (58), shows that

$$\begin{cases} |\tilde{E}|, \|\widetilde{Q}\|_{\xi'}, \|\widetilde{T}\|, \|\phi_1 - \mathrm{id}\|_{\xi'} \leq L \\ \|P_1\|_{\xi'} \leq LM ; \end{cases} \tag{61}$$

provided (60) holds (notice that (60) implies (54)).

One step of the iteration has been concluded and the needed estimates obtained. The idea is to iterate the construction infinitely many times, as we proceed to describe.

Step 3: Iteration and Convergence

In order to iterate Kolmogorov's construction, analyzed in Step 2, so as to construct a sequence of symplectic transformations

$$\phi_j \colon W_{\xi_{j+1}} \to W_{\xi_j} , \tag{62}$$

closer and closer to the identity, and such that (42) hold, the first thing to do is to choose the sequence ξ_j: this sequence must be convergent, so that $\delta_j = \xi_j - \xi_{j+1}$ has to go to zero rather quickly. Inverse powers of δ_j (which, at the jth step will play the role of δ in the previous paragraph) appear in the smallness conditions (see, e. g., (54)): this "divergence" will, however, be beaten by the super-fast decay of ε^{2^j}.

Fix $0 < \xi_* < \xi$ (ξ_* will be the domain of analyticity of ϕ_* and K_* in Theorem 1 and, for $j \geq 0$, let

$$\begin{cases} \xi_0 := \xi \\ \delta_0 := \dfrac{\xi - \xi_*}{2} \end{cases} \qquad \begin{cases} \delta_j := \dfrac{\delta_0}{2^j} \\ \xi_{j+1} := \xi_j - \delta_j = \xi_* + \dfrac{\delta_0}{2^j} \end{cases} \tag{63}$$

and observe that $\xi_j \downarrow \xi_*$. With this choice[20], Kolmogorov's algorithm can be iterated infinitely many times, provided ε_0 is small enough. To be more precise, let c, μ and ν be as in (59), and define

$$C := 2 \max \left\{ |E|, |\omega|, \|Q\|_\xi, \|T\|, 1 \right\} . \tag{64}$$

Smallness Assumption: *Assume that $|\varepsilon| \leq \varepsilon_0$ and that ε_0 satisfies*

$$\varepsilon_0 DB \|P\|_\xi \leq 1 \qquad \text{where } D := 3c\delta_0^{-(\nu+1)} C^\mu , \quad B := 2^{\nu+1} ; \tag{65}$$

notice that the constant C in (64) satisfies (51) and that (65) implies (54). Then the following claim holds.

Claim C: *Under condition (65) one can iteratively construct a sequence of Kolmogorov symplectic maps ϕ_j as in (62) so that (42) holds in such a way that $\varepsilon^{2^j} P_j$, $\Phi_j := \phi_1 \circ \phi_2 \circ \cdots \circ \phi_j$, E_j, K_j, Q_j converge uniformly on W_{ξ_*} to, respectively, 0, ϕ_*, E_*, K_*, Q_*, which are real-analytic on W_{ξ_*} and $H \circ \phi_* = K_* = E_* + \omega \cdot y + Q_*$ with $Q_* = O(|y|^2)$. Furthermore, the following estimates hold for any $|\varepsilon| \leq \varepsilon_0$ and for any $i \geq 0$:*

$$|\varepsilon|^{2^i} M_i := |\varepsilon|^{2^i} \|P_i\|_{\xi_i} \leq \frac{(|\varepsilon| DBM)^{2^i}}{DB^{i+1}} , \tag{66}$$

$$\|\phi_* - \mathrm{id}\|_{\xi_*}, |E - E_*|, \|Q - Q_*\|_{\xi_*}, \|T - T_*\| \leq |\varepsilon| DBM , \tag{67}$$

where $T_* := \langle \partial_y^2 Q_*(0, \cdot) \rangle^{-1}$, showing that K_* is non-degenerate.

Remark 7

(i) From Claim **C** Kolmogorov Theorem 1 follows at once. In fact we have proven the following quantitative statement: *Let $\omega \in \mathcal{D}^d_{\kappa,\tau}$ with $\tau \geq d - 1$ and $0 < \kappa < 1$; let Q and P be real-analytic on $W_\xi = D^d(0, \xi) \times \mathbb{T}^d_\xi$ for some $0 < \xi \leq 1$ and let $0 < \theta < 1$; let T and C be as in, respectively, (50) and (64). There exist $c_* = c_*(d, \tau, \kappa, \theta) > 1$ and positive integers $\sigma = \sigma(d, \tau)$, μ such that if*

$$|\varepsilon| \leq \varepsilon_* := \frac{\xi^\sigma}{c_* \|P\|_\xi C^\mu} \tag{68}$$

then one can construct a near-to-identity Kolmogorov transformation (Remark 3–(ii)) $\phi_* \colon W_{\theta\xi} \to W_\xi$ such that the thesis of Theorem 1 holds together with the estimates

$$\|\phi_* - \mathrm{id}\|_{\theta\xi}, |E - E_*|, \|Q - Q_*\|_{\theta\xi}, \|T - T_*\| \leq \frac{|\varepsilon|}{\varepsilon_*} = |\varepsilon| c_* \|P\|_\xi C^\mu \xi^{-\sigma} . \tag{69}$$

(The correspondence with the above constants being: $\xi_* = \theta\xi, \delta_0 = \xi(1-\theta)/2, \sigma = \nu + 1, D = 3c(2/(1-\theta))^{\nu+1} C^\mu$, $c_* = 3c(4/(1-\theta))^{\nu+1}$).

(ii) From Cauchy estimates and (67), it follows that $\|\phi_* - \mathrm{id}\|_{C^p}$ and $\|Q - Q_*\|_{C^p}$ are small for any p (small in $|\varepsilon|$ but not uniformly in[21] p).

(iii) All estimates are uniform in ε, therefore, from Weierstrass theorem (compare note 18) it follows that ϕ_* and K_* are *analytic in ε in the complex ball of radius ε_0*.

Power series expansions in ε were very popular in the nineteenth and twentieth centuries[22], however convergence of the formal ε-power series of quasi-periodic solutions was proved for the first time only in the 1960s thanks to KAM theory [45]. Some of this matter is briefly discussed in Sect. "Future Directions" below.

(iv) **The Nearly-Integrable Case** In [35] it is pointed out that Kolmogorov's Theorem easily yields the existence of many KAM tori for nearly-integrable systems (14) for $|\varepsilon|$ small enough, provided K is non-degenerate in the sense that

$$\det K_{yy}(y_0) \neq 0 \,. \tag{70}$$

In fact, without loss of generality we may assume that $\omega := H_0'$ is a diffeomorphism on $B(y_0, 2r)$ and $\det K_{yy}(y) \neq 0$ for all $y \in B(y_0, 2r)$. Furthermore, letting $B = B(y_0, r)$, fixing $\tau > d-1$ and denoting by ℓ_d the Lebesgue measure on $\mathbb{R}^d$, from the remark in note 11 and from the fact that ω is a diffeomorphism, there follows that there exists a constant $c_\#$ depending only on d, τ and r such that

$$\ell_d(\omega(B)\backslash\mathcal{D}^d_{\kappa,\tau}), \ell_d(\{y \in B\colon \omega(y) \notin \mathcal{D}^d_{\kappa,\tau}\}) < c_\#\kappa \,. \tag{71}$$

Now, let $B_{\kappa,\tau} := \{y \in B\colon \omega(y) \in \mathcal{D}^d_{\kappa,\tau}\}$ (which, by (71) has Lebesgue measure $\ell_d(B_{\kappa,\tau}) \geq \ell_d(B) - c_\#\kappa$), then for any $\bar{y} \in B_{\kappa,\tau}$ we can make the trivial symplectic change of variables $y \to \bar{y} + y$, $x \to x$ so that K can be written as in (10) with

$$E := K(\bar{y}) \,, \quad \omega := K_y(\bar{y}) \,,$$
$$Q(y,x) = Q(y) := K(y) - K(\bar{y}) - K_y(\bar{y}) \cdot y \,,$$

(where, for ease of notation, we did not change names to the new symplectic variables) and $P(\bar{y} + y, x)$ replacing (with a slight abuse of notation) $P(y, x)$. By Taylor's formula, $Q = O(|y|^2)$ and, furthermore (since $Q(y,x) = Q(y)$, $\langle \partial_y^2 Q(0,x)\rangle = Q_{yy}(0) = K_{yy}(\bar{y})$, which is invertible according to out hypotheses. Thus K is Kolmogorov non-degenerate and Theorem 1 can be applied yielding, for $|\varepsilon| < \varepsilon_0$, a KAM torus $\mathcal{T}_{\omega,\varepsilon}$, with $\omega = K_y(\bar{y})$, for each $\bar{y} \in B_{\kappa,\tau}$. Notice that the measure of initial phase points, which, perturbed, give rise to KAM tori, has a small complementary bounded by $c_\#\kappa$ (see (71)).

(v) In the nearly-integrable setting described in the preceding point, the union of KAM tori is usually called the **Kolmogorov set**. It is not difficult to check that the dependence upon $\bar{y}$ of the Kolmogorov transformation ϕ_* is Lipschitz[23], implying that the measure of the complementary of Kolmogorov set itself is also bounded by $\hat{c}_\#\kappa$ with a constant $\hat{c}_\#$ depending only on d, τ and r.
Indeed, the estimate on the measure of the Kolmogorov set can be made more quantitative (i. e., one can see how such an estimate depends upon ε as $\varepsilon \to 0$). In fact, revisiting the estimates discussed in **Step 2** above one sees easily that the constant c defined in (53) has the form[24]

$$c = \hat{c}\kappa^{-4} \,. \tag{72}$$

where $\hat{c} = \hat{c}(d,\tau)$ depends only on d and τ (here the Diophantine constant κ is assumed, without loss of generality, to be smaller than one). Thus the smallness condition (65) reads $\varepsilon_0\kappa^{-4}\bar{D} \leq 1$ with some constant $\bar{D}$ independent of κ: such condition is satisfied by choosing $\kappa = (\bar{D}\varepsilon_0)^{1/4}$ and since $\hat{c}_\#\kappa$ was an upper bound on the complementary of Kolmogorov set, we see that *the set of phase points which do not lie on KAM tori may be bounded by a constant times* $\sqrt[4]{\varepsilon_0}$. Actually, it turns that this bound is not optimal, as we shall see in the next section: see Remark 10.

(vi) The proof of claim **C** follows easily by induction on the number j of the iterative steps[25].

Arnold's Scheme

The first detailed proof of Kolmogorov Theorem, in the context of nearly-integrable Hamiltonian systems (compare Remark 1–(iii)), was given by V.I. Arnold in 1963.

Theorem 2 (Arnold [2]) *Consider a one-parameter family of nearly-integrable Hamiltonians*

$$H(y,x;\varepsilon) := K(y) + \varepsilon P(y,x) \quad (\varepsilon \in \mathbb{R}) \tag{73}$$

with K and P real-analytic on $\mathcal{M} := B(y_0, r) \times \mathbb{T}^d$ (endowed with the standard symplectic form $dy \wedge dx$) satisfying

$$K_y(y_0) = \omega \in \mathcal{D}^d_{\kappa,\tau} \,, \quad \det K_{yy}(y_0) \neq 0 \,. \tag{74}$$

Then, if ε is small enough, there exists a real-analytic embedding

$$\phi\colon \theta \in \mathbb{T}^d \to \mathcal{M} \tag{75}$$

close to the trivial embedding (y_0, id), *such that the d-torus*

$$\mathcal{T}_{\omega,\varepsilon} := \phi\left(\mathbb{T}^d\right) \tag{76}$$

is invariant for H and

$$\phi_H^t \circ \phi(\theta) = \phi(\theta + \omega t)\,, \tag{77}$$

showing that such a torus is a non-degenerate KAM torus for H.

Remark 8

(i) The above Theorem is a corollary of Kolmogorov Theorem 1 as discussed in Remark 7–(iv).
(ii) Arnold's proof of the above Theorem *is not based* upon Kolmogorov's scheme and is rather different in spirit – although still based on a Newton method – and introduces several interesting technical ideas.
(iii) Indeed, the iteration scheme of Arnold is more classical and, from the algebraic point of view, easier than Kolmogorov's, but the estimates involved are somewhat more delicate and introduce a logarithmic correction, so that, in fact, the smallness parameter will be

$$\epsilon := |\varepsilon|(\log|\varepsilon|^{-1})^{\rho} \tag{78}$$

(for some constant $\rho = \rho(d,\tau) \geq 1$) rather than $|\varepsilon|$ as in Kolmogorov's scheme; see, also, Remark 9–(iii) and (iv) below.

Arnold's Scheme

Without loss of generality, one may assume that K and P have analytic and bounded extension to $W_{r,\xi}(y_0) := D(y_0,r) \times \mathbb{T}^d_\xi$ for some $\xi > 0$, where, as above, $D(y_0,r)$ denotes the complex ball of center y_0 and radius r. We remark that, in what follows, the analyticity domains of actions and angles play a different role

The Hamiltonian H in (73) admits, for $\varepsilon = 0$ the (KAM) invariant torus $\mathcal{T}_{\omega,0} = \{y_0\} \times \mathbb{T}^d$ on which the K-flow is given by $x \to x + \omega t$. Arnold's basic idea is to find a symplectic transformation

$$\phi_1\colon W_1 := D(y_1,r_1)\times\mathbb{T}^d_{\xi_1} \to W_0 := D(y_0,r)\times\mathbb{T}^d_\xi\,, \tag{79}$$

so that $W_1 \subset W_0$ and

$$\begin{cases} H_1 := H\circ\phi_1 = K_1 + \varepsilon^2 P_1\,, \quad K_1 = K_1(y)\,, \\ \partial_y K_1(y_1) = \omega\,, \quad \det\partial_y^2 K_1(y_1) \neq 0 \end{cases} \tag{80}$$

(with abuse of notation we denote here the new symplectic variables with the same name of the original variables; as above, dependence on ε will, often, not be explicitly indicated). In this way the initial set up is reconstructed and, for ε small enough, one can iterate the scheme so as to build a sequence of symplectic transformations

$$\phi_j\colon W_j := D(y_j,r_j)\times\mathbb{T}^d_{\xi_j} \to W_{j-1} \tag{81}$$

so that

$$\begin{cases} H_j := H_{j-1}\circ\phi_j = K_j + \varepsilon^{2^j} P_j\,, \quad K_j = K_j(y)\,, \\ \partial_y K_j(y_j) = \omega\,, \quad \det\partial_y^2 K_j(y_j) \neq 0\,. \end{cases} \tag{82}$$

Arnold's transformations, as in Kolmogorov's case, are closer and closer to the identity, and the limit

$$\begin{aligned} \phi(\theta) &:= \lim_{j\to\infty} \Phi_j(y_j,\theta)\,, \\ \Phi_j &:= \phi_1\circ\cdots\circ\phi_j\colon W_j \to W_0\,, \end{aligned} \tag{83}$$

defines a real-analytic embedding of $\mathbb{T}^d$ into the phase space $B(y_0,r)\times\mathbb{T}^d$, which is close to the trivial embedding (y_0, id); furthermore, the torus

$$\mathcal{T}_{\omega,\varepsilon} := \phi(\mathbb{T}^d) = \lim_{j\to\infty} \Phi_j(y_j,\mathbb{T}^d) \tag{84}$$

is invariant for H and (77) holds as announced in Theorem 2. Relation (77) follows from the following argument. The radius r_j will turn out to tend to 0 but in a much slower way than $\varepsilon^{2^j} P_j$. This fact, together with the rapid convergence of the symplectic transformation Φ_j in (83) implies

$$\begin{aligned} \phi_H^t\circ\phi(\theta) &= \lim_{j\to\infty} \phi_H^t\left(\Phi_j(y_j,\theta)\right) \\ &= \lim_{j\to\infty} \Phi_j\circ\phi_{H_j}^t(y_j,\theta) \\ &= \lim_{j\to\infty} \Phi_j(y_j,\theta+\omega t) \\ &= \phi(\theta+\omega t) \end{aligned} \tag{85}$$

where: the first equality is just smooth dependence upon initial data of the flow ϕ_H^t together with (83); the second equality is (3); the third equality is due to the fact that (see (82)) $\phi_{H_j}^t(y_j,\theta) = \phi_{K_j}^t(y_j,\theta) + O(\varepsilon^{2^j}\|P_j\|) = (y_j,\theta+\omega t) + O(\varepsilon^{2^j}\|P_j\|)$ and $O(\varepsilon^{2^j}\|P_j\|)$ goes very rapidly to zero; the fourth equality is again (83).

Arnold's Transformation

Let us look for a near-to-the-identity transformation ϕ_1 so that the first line of (80) holds; this transformation will be determined by a generating function of the form

$$y'\cdot x + \varepsilon g(y',x)\,, \quad \begin{cases} y = y' + \varepsilon g_x(y',x) \\ x' = x + \varepsilon g_{y'}(y',x)\,. \end{cases} \tag{86}$$

Inserting $y = y' + \varepsilon g_x(y', x)$ into H, one finds

$$H(y' + \varepsilon g_x, x) = K(y') + \varepsilon \left[K_y(y') \cdot g_x + P(y', x)\right] + \varepsilon^2 \left(P^{(1)} + P^{(2)}\right) \quad (87)$$

with (compare (26))

$$\begin{aligned} P^{(1)} &:= \frac{1}{\varepsilon^2}[K(y' + \varepsilon g_x) - K(y') - \varepsilon K_y(y') \cdot g_x] \\ &= \int_0^1 (1-t) K_{yy}(y' + t\varepsilon g_x, x)\, g_x \cdot g_x \mathrm{d}t \\ P^{(2)} &:= \frac{1}{\varepsilon}[P(y' + \varepsilon g_x, x) - P(y', x)] \\ &= \int_0^1 P_y(y' + t\varepsilon g_x, x) \cdot g_x \mathrm{d}t\,. \end{aligned} \quad (88)$$

Remark 9

(i) The (naive) idea is to try determine g so that

$$K_y(y') \cdot g_x + P(y', x) = \text{function of } y' \text{ only}, \quad (89)$$

however, such a relation is impossible to achieve. First of all, by taking the x-average of both sides of (89) one sees that the "function of y' only" has to be the mean of $P(y', \cdot)$, i. e., the zero-Fourier coefficient $P_0(y')$, so that the *formal* solution of (89), is (by Fourier expansion)

$$\begin{cases} g = \displaystyle\sum_{n \neq 0} \frac{-P_n(y')}{iK_y(y') \cdot n} \mathrm{e}^{in \cdot x}\,, \\ K_y(y') \cdot g_x + P(y', x) = P_0(y')\,. \end{cases} \quad (90)$$

But, (at difference with Kolmogorov's scheme) the frequency $K_y(y')$ is *a function of the action* y' and since, by the Inverse Function Theorem (Appendix "A The Classical Implicit Function Theorem"), $y \to K_y(y)$ is a local diffeomorphism, it follows that, in any neighborhood of y_0, there are points y such that $K_y(y) \cdot n = 0$ for some[26] $n \in \mathbb{Z}^d$. Thus, in any neighborhood of y_0, some divisors in (90) will actually vanish and, therefore, *an analytic solution g cannot exist*[27].

(ii) On the other hand, since $K_y(y_0)$ is rationally independent, it is clearly possible (simply by continuity) to control a finite number of divisors in a suitable neighborhood of y_0, more precisely, for any $N \in \mathbb{N}$ one can find $\bar{r} > 0$ such that

$$K_y(y) \cdot n \neq 0\,, \qquad \forall y \in D(y_0, \bar{r})\,, \quad \forall 0 < |n| \leq N\,; \quad (91)$$

the important quantitative aspects will be shortly discussed below.

(iii) Relation (89) is also one of the main "identity" in *Averaging Theory* and is related to the so-called *Hamilton–Jacobi equation*. Arnold's proof makes such a theory rigorous and shows how a Newton method can be built upon it in order to establish the existence of invariant tori. In a sense, Arnold's approach is more classical than Kolmogorov's.

(iv) When (for a given y and n) it occurs that $K_y(y) \cdot n = 0$ one speaks of an (exact) *resonance*. As mentioned at the end of point (i), in the general case, *resonances are dense*. This represents the main problem in Hamiltonian perturbation theory and is a typical feature of *conservative systems*. For generalities on Averaging Theory, Hamilton–Jacobi equation, resonances etc. see, e. g., [5] or Sects. 6.1 and 6.2 of [6].

The key (simple!) idea of Arnold is to split the perturbation into two terms

$$P = \hat{P} + \check{P} \quad \text{where} \quad \begin{cases} \hat{P} := \displaystyle\sum_{|n| \leq N} P_n(y) \mathrm{e}^{in \cdot x} \\ \check{P} := \displaystyle\sum_{|n| > N} P_n(y) \mathrm{e}^{in \cdot x} \end{cases} \quad (92)$$

choosing N so that

$$\check{P} = O(\varepsilon) \quad (93)$$

(this is possible because of the fast decay of the Fourier coefficients of P; compare (33)). Then, for $\varepsilon \neq 0$, (87) can be rewritten as follows

$$H(y' + \varepsilon g_x, x) = K(y') + \varepsilon \left[K_y(y') \cdot g_x + \hat{P}(y', x)\right] + \varepsilon^2 \left(P^{(1)} + P^{(2)} + P^{(3)}\right) \quad (94)$$

with $P^{(1)}$ and $P^{(2)}$ as in (88) and

$$P^{(3)}(y', x) := \frac{1}{\varepsilon} \check{P}(y', x)\,. \quad (95)$$

Thus, letting[28]

$$g = \sum_{0 < |n| \leq N} \frac{-P_n(y')}{iK_y(y') \cdot n} \mathrm{e}^{in \cdot x}\,, \quad (96)$$

one gets

$$H(y' + \varepsilon g_x, x) = K_1(y') + \varepsilon^2 P'(y', x) \quad (97)$$

where

$$\begin{aligned} K_1(y') &:= K(y') + \varepsilon P_0(y')\,, \\ P'(y', x) &:= P^{(1)} + P^{(2)} + P^{(3)}\,. \end{aligned} \quad (98)$$

Now, by the IFT (Appendix "A The Classical Implicit Function Theorem"), for ε small enough, the map $x \to x + g_{y'}(y', x)$ can be inverted with a real-analytic map of the form

$$\varphi(y', x'; \varepsilon) := x' + \varepsilon\alpha(y', x'; \varepsilon) \tag{99}$$

so that Arnold's symplectic transformation is given by

$$\phi_1 : (y', x') \to \begin{cases} y = y' + \varepsilon g_x\left(y', \varphi(y', x'; \varepsilon)\right) \\ x = \varphi(y', x'; \varepsilon) \\ \quad = x' + \varepsilon\alpha(y', x', ; \varepsilon) \end{cases} \tag{100}$$

(compare (21)). To finish the construction, observe that, from the IFT (see Appendix "A The Classical Implicit Function Theorem" and the quantitative discussion below) it follows that there exists a (unique) point $y_1 \in B(y_0, \bar{r})$ so that the second line of (80) holds, provided ε is small enough.

In conclusion, the analogue of Proposition 1 holds, describing Arnold's scheme:

Proposition 2 *If ϕ_1 is defined in (100) with g given in (96) (with N so that (93) holds) and φ given in (99), then (80) holds with K_1 as in (98) and $P_1(y', x') := P'(y', \varphi(y', x'))$ with P' defined in (98), (95) and (88).*

Estimates and Convergence

If f is a real-analytic function with analytic extension to $W_{r,\xi}$, we denote, for any $r' \le r$ and $\xi' \le \xi$,

$$\|f\|_{r',\xi'} := \sup_{W_{r',\xi'}(y_0)} |f(y, x)| ; \tag{101}$$

furthermore, we define

$$T := K_{yy}(y_0)^{-1} , \quad M := \|P\|_{r,\xi} , \tag{102}$$

and assume (without loss of generality)

$$\kappa < 1 , \ r < 1 , \ \xi \le 1 , \qquad \max\{1, \|K_y\|_r, \|K_{yy}\|_r, \|T\|\} < C , \tag{103}$$

for a suitable constant C (which, as above, will not change during the iteration).

We begin by discussing how N and $\bar{r}$ depend upon ε. From the exponential decay of the Fourier coefficients (33), it follows that, choosing

$$N := 5\delta^{-1}\lambda , \quad \text{where } \lambda := \log|\varepsilon|^{-1} , \tag{104}$$

then

$$\|\check{P}\|_{r,\xi-\frac{\delta}{2}} \le |\varepsilon| M \tag{105}$$

provided

$$|\varepsilon| \le \text{const}\, \delta \tag{106}$$

for a suitable[29] const = const(d).

The second key inequality concerns the control of the small divisors $K_y(y') \cdot n$ appearing in the definition of g (see (96)), in a neighborhood $D(y_0, \bar{r})$ of y_0: this will determine the size of $\bar{r}$.

Recalling that $K_y(y_0) = \omega \in \mathcal{D}^d_{\tau,\kappa}$, by Taylor's formula and (9), one finds, for any $0 < |n| \le N$ and any $y' \in D(y_0, \bar{r})$,

$$\begin{aligned} |K_y(y') \cdot n| &= \left|\omega \cdot n + \left(K_y(y') - K_y(y_0)\right) \cdot n\right| \\ &\ge |\omega \cdot n| \left(1 - \frac{\|K_{yy}\|_r}{|\omega \cdot n|} |n| \bar{r}\right) \\ &\ge \frac{\kappa}{|n|^\tau} \left(1 - \frac{C}{\kappa} |n|^{\tau+1} \bar{r}\right) \\ &\ge \frac{\kappa}{|n|^\tau} \left(1 - \frac{C}{\kappa} N^{\tau+1} \bar{r}\right) \\ &\ge \frac{1}{2} \frac{\kappa}{|n|^\tau} , \end{aligned} \tag{107}$$

provided $\bar{r} \le r$ satisfies also

$$\bar{r} \le \frac{\kappa}{2CN^{\tau+1}} \overset{(104)}{=} \frac{\kappa}{2 \cdot 5^{\tau+1} C (\delta^{-1}\lambda)^{\tau+1}} . \tag{108}$$

Equation (107) allows us to easily control Arnold's generating function g. For example:

$$\begin{aligned} \|g_x\|_{\bar{r},\xi-\frac{\delta}{2}} &= \sup_{D(y_0,\bar{r}) \times \mathbb{T}^d_{\xi-\frac{\delta}{2}}} \left| \sum_{0<|n|\le N} \frac{n P_n(y')}{K_y(y') \cdot n} e^{in \cdot x} \right| \\ &\le \sum_{0<|n|\le N} \frac{\sup_{D(y_0,r)} |P_n(y')|}{|K_y(y') \cdot n|} |n| e^{(\xi-\frac{\delta}{2})|n|} \\ &\le \sum_{n \in \mathbb{Z}^d} M \frac{2|n|^{\tau+1}}{\kappa} e^{-\frac{\delta}{2}|n|} \\ &\le \text{const} \frac{M}{\kappa} \delta^{-(\tau+1+d)} , \end{aligned} \tag{109}$$

where "const" denotes a constant depending on d and τ only; compare also Remark 6–(i).

Let us now discuss, from a quantitative point of view, how to choose the new "center" of the action variables y_1, which is determined by the requirements in (80). Assuming that

$$\bar{r} \le \frac{r}{2} \tag{110}$$

(allowing the use of Cauchy estimates for y-derivatives of K or P in $D(y_0, \bar{r})$), it is not difficult to see that the quantitative IFT of Appendix "A The Classical Implicit Function Theorem" implies that there exists a unique $y_1 \in D(y_0, \bar{r})$ such that (80) holds and, furthermore

$$|y_1 - y_0| \le 4CMr^{-1}|\varepsilon| \,, \tag{111}$$

and

$$\begin{aligned} \partial_y^2 K_1(y_1) &:= K_{yy}(y_1) + \varepsilon \partial_y^2 P_0(y_1) \\ &=: T^{-1}(\mathbb{1}_d + A) \end{aligned} \tag{112}$$

with a matrix A satisfying

$$\|A\| \le 10C^3 M|\varepsilon| \le \frac{1}{2} \tag{113}$$

provided[30]

$$\begin{cases} 8C^2 \dfrac{\bar{r}}{r} \le 1 \,, \\ 8CM\bar{r}^{-2}|\varepsilon| \le 1 \end{cases} \tag{114}$$

Equation (113) shows that $\partial_y^2 K(y_1)$ is invertible (Neumann series) and that[31]

$$\partial_y^2 K_1(y_1)^{-1} = T + \varepsilon \widetilde{T} \,, \quad \|\widetilde{T}\| \le 20C^3 M \,. \tag{115}$$

Finally, notice that the second conditions in (114) and (111) imply that $|y_1 - y_0| < \bar{r}/2$ so that

$$D(y_1, \bar{r}/2) \subset D(y_0, \bar{r}) \,. \tag{116}$$

Now, all the estimating tools are set up and, writing

$$\begin{aligned} K_1 &:= K + \varepsilon \widetilde{K} = K + \varepsilon P_0(y') \,, \\ y_1 &:= y_0 + \varepsilon \tilde{y} \,, \end{aligned} \tag{117}$$

one can easily prove (along the lines that led to (53)) the following estimates, where as in Sect. "Kolmogorov Theorem", $\bar{\xi} := \xi - \frac{2}{3}\delta$ and $\bar{r}$ is as above:

$$\begin{cases} \dfrac{\|g_x\|_{\bar{r},\bar{\xi}}}{r}, \|g_y\|_{\bar{r},\bar{\xi}}, \|\widetilde{K}_y\|_{\bar{r}}, \|\widetilde{K}_{yy}\|, |\tilde{y}|, \|\widetilde{T}\| \\ \qquad \le c\kappa^{-2} C^{\mu} \delta^{-\nu} \lambda^{\rho} M =: L \,, \\ \|P'\|_{\bar{r},\bar{\xi}} \le c\kappa^{-2} C^{\mu} \delta^{-\nu} \lambda^{\rho} M^2 = LM \,, \end{cases} \tag{118}$$

where $c = c(d, \tau) > 1$, $\mu \in \mathbb{Z}_+$, ν and ρ are positive integers depending on d and τ. Now, by[32] Lemma 1 and (118), one has that map $x \to x + \varepsilon g_y(y', x)$ has, for any $y' \in D_{\bar{r}}(y_0)$, an analytic inverse $\varphi = x' + \varepsilon\alpha(x'; y', \varepsilon) =: \varphi(y', x')$ on $\mathbb{T}^d_{\bar{\xi} - \frac{\delta}{3}}$ provided (54), with $\bar{L}$ replaced by L in (118), holds, in which case (55) holds (for any $|\varepsilon| \le \varepsilon_0$ and any $y' \in D_r(y_0)$). Furthermore, under the above hypothesis, it follows that[33]

$$\begin{cases} \phi_1 := \big(y' + \varepsilon g_x(y', \varphi(y', x')), \varphi(y', x')\big) : \\ \qquad W_{\bar{r}/2, \xi - \delta}(y_1) \to W_{r,\xi}(y_0) \\ \|\phi_1 - \mathrm{id}\|_{\bar{r}/2, \xi - \delta} \le |\varepsilon| L \,. \end{cases} \tag{119}$$

Finally, letting $P_1(y', x') := P'(y', \varphi(y', x'))$ one sees that P_1 is real-analytic on $W_{\bar{r}/2, \bar{\xi} - \delta}(y_1)$ and bounded on that domain by

$$\|P_1\|_{\bar{r}/2, \xi - \delta} \le LM \,. \tag{120}$$

In order to iterate the above construction, we fix $0 < \xi_* < \xi$ and set

$$\begin{aligned} C &:= 2\max\{1, \|K_y\|_r, \|K_{yy}\|_r, \|T\|\} \,, \\ \gamma &:= 3C \,, \\ \delta_0 &:= \frac{(\gamma - 1)(\xi - \xi_*)}{\gamma} \,; \end{aligned} \tag{121}$$

ξ_j and δ_j as in (63) *but with* δ_0 *as in* (121); we also define, for any $j \ge 0$,

$$\begin{aligned} \lambda_j &:= 2^j \lambda = \log \varepsilon_0^{-2^j} \,, \\ r_j &:= \frac{\kappa}{4 \cdot 5^{\tau+1} C(\delta_j^{-1} \lambda_j)^{\tau+1}} \,; \end{aligned} \tag{122}$$

(this part is adapted from **Step 3** in Sect. "Kolmogorov Theorem"; see, in particular, (103)). With such choices it is not difficult to check that the iterative construction may be carried out infinitely many times yielding, as a byproduct, Theorem 2 with ϕ real-analytic on $\mathbb{T}^d_{\xi_*}$, provided $|\varepsilon| \le \varepsilon_0$ with ε_0 satisfying[34]

$$\begin{cases} \varepsilon_0 \le \mathrm{e}^{-\beta} & \text{with} \quad \beta := \dfrac{\delta_0}{5} \left(\dfrac{\kappa}{Cr}\right)^{\frac{1}{\tau+1}} \\ \varepsilon_0 DB\|P\|_{\xi} \le 1 & \text{with} \quad D := 3c\kappa^{-2} \delta_0^{-(\nu+1)} C^{\mu} \,, \\ & \qquad\quad B := \gamma^{\nu+1} (\log \varepsilon_0^{-1})^{\rho} \,. \end{cases} \tag{123}$$

Remark 10 Notice that the power of κ^{-1} (the inverse of the Diophantine constant) in the second smallness condition in (123) is two, which implies (compare Remark 7–(v)) that the measure of the complement of the Kolmogorov set may be bounded by a constant times $\sqrt{\varepsilon_0}$. This bound is optimal as the trivial example $(y_1^2 + y_2^2)/2 +$

$\varepsilon\cos(x_1)$ shows: the Hamiltonian is integrable and the phase portrait shows that the separatrices of the pendulum $y_1^2/2 + \varepsilon\cos x_1$ bound a region of area $\sqrt{|\varepsilon|}$ with no KAM tori (as the librational curves within such region are not graphs over the angles).

The Differentiable Case: Moser's Theorem

J.K. Moser, in 1962, proved a perturbation (KAM) Theorem, in the framework of area-preserving twist mappings of an annulus[35] $[0,1]\times\mathbb{S}^1$, for integrable analytic systems perturbed by a C^k perturbation, [42] and [43]. Moser's original setup corresponded to the Hamiltonian case with $d=2$ and the required smoothness was C^k with $k=333$. Later, this number was brought down to 5 by H. Rüssmann, [53].

Moser's original approach, similarly to the approach that led J. Nash to prove its theorem on the smooth embedding problem of compact Riemannian manifolds [48], is based on a smoothing technique (via convolutions), which re-introduces at each step of the Newton iteration a certain number of derivatives which one loses in the inversion of the small divisor operator.

The technique, which we shall describe here, is again due to Moser [46] but is rather different from the original one and is based on a quantitative analytic KAM Theorem (in the style of statement in Remark 7–(i) above) in conjunction with a characterization of differentiable functions in terms of functions, which are real-analytic on smaller and smaller complex strips; see [44] and, for an abstract functional approach, [65], [66]. By the way, this approach, suitably refined, leads to optimal differentiability assumptions (i. e., the Hamiltonian may be assumed to be C^ℓ with $\ell > 2d$); see, [50] and the beautiful exposition [59], which inspires the presentation reported here.

Let us consider a Hamiltonian $H = K + \varepsilon P$ (as in (17)) with K a real-analytic Kolmogorov normal form as in (10) with $\omega\in\mathcal{D}^d_{\kappa,\tau}$ and Q real-analytic; P is assumed to be a $C^\ell(\mathbb{R}^d,\mathbb{T}^d)$ function with $\ell=\ell(d,\tau)$ to be specified later[36].

Remark 11 The analytic KAM theorem, we shall refer to is the quantitative Kolmogorov Theorem as stated in Remark 7–(i) above, with (69) strengthened by including in the left hand side of (69) also[37] $\|\partial(\phi_*-\mathrm{id})\|_{\theta\xi}$ and $\|\partial(Q-Q_*)\|_{\theta\xi}$ (where "∂" denotes, here, "Jacobian" with regard to (y,x) for $(\phi_*-\mathrm{id})$ and "gradient" for $(Q-Q_*)$).

The analytic characterization of differentiable functions, suitable for our purposes, is explained in the following two lemmata[38]

Lemma 2 (Jackson, Moser, Zehnder) *Let $f\in C^l(\mathbb{R}^d)$ with $l>0$. Then, for any $\xi>0$ there exists a real-analytic function $f\colon X^d_\xi := \{x\in\mathbb{C}^d\colon |\mathrm{Im}x_j|<\xi\}\to\mathbb{C}$ such that*

$$\begin{cases} \sup\limits_{X^d_\xi}|f_\xi| \le c\|f\|_{C^0}\,, \\ \sup\limits_{X^d_{\xi'}}|f_\xi - f_{\xi'}| \le c\|f\|_{C^l}\xi^l\,, \quad \forall\, 0<\xi'<\xi\,, \end{cases} \tag{124}$$

where $c=c(d,l)$ is a suitable constant; if f is periodic in some variable x_j, so is f_ξ.

Lemma 3 (Bernstein, Moser) *Let $l\in\mathbb{R}_+\backslash\mathbb{Z}$ and $\xi_j := 1/2^j$. Let $f_0=0$ and let, for any $j\ge1$, f_j be real analytic functions on $X^d_j := \{x\in\mathbb{C}^d\colon |\mathrm{Im}x_j|<2^{-j}\}$ such that*

$$\sup_{X^d_j}|f_j - f_{j-1}| \le A2^{-jl} \tag{125}$$

for some constant A. Then, f_j tends uniformly on $\mathbb{R}^d$ to a function $f\in C^l(\mathbb{R}^d)$ such that, for a suitable constant $C=C(d,l)>0$,

$$\|f\|_{C^l(\mathbb{R}^d)} \le CA\,. \tag{126}$$

Finally, if the f_i's are periodic in some variable x_j then so is f.

Now, denote by $X_\xi = X^d_\xi\times\mathbb{T}^d\subset\mathbb{C}^{2d}$ and define (compare Lemma 2)

$$P^j := P_{\xi_j}\,, \quad \xi_j := \frac{1}{2^j}\,. \tag{127}$$

Claim M: *If $|\varepsilon|$ is small enough and if $\ell>\sigma+1$, then there exists a sequence of Kolmogorov symplectic transformations $\{\Phi_j\}_{j\ge0}$, $|\varepsilon|$-close to the identity, and a sequence of Kolmogorov normal forms K_j such that*

$$H_j\circ\Phi_j = K_{j+1} \ \text{ on } \ W_{\xi_{j+1}} \tag{128}$$

where

$$\begin{aligned} &H_j := K + \varepsilon P^j \\ &\Phi_0 = \phi_0 \quad\text{and}\quad \Phi_j := \Phi_{j-1}\circ\phi_j\,, (j\ge1) \\ &\phi_j\colon W_{\xi_{j+1}}\to W_{\alpha\xi_j}\,,\ \ \Phi_{j-1}\colon W_{\alpha\xi_j}\to X_{\xi_j}\,, \\ &\qquad j\ge1 \ \text{ and } \ \alpha := \frac{1}{\sqrt{2}}\,, \\ &\sup_{x\in\mathbb{T}^d_{\xi_{j+1}}} |\Phi_j(0,x)-\Phi_{j-1}(0,x)| \le \mathrm{const}|\varepsilon|2^{-(\ell-\sigma)j}\,. \end{aligned} \tag{129}$$

The proof of Claim **M** follows easily by induction[39] from Kolmogorov's Theorem (compare Remark 11) and Lemma 2.

From Claim **M** and Lemma 3 (applied to $f(x) = \Phi_j(0,x) - \Phi_0(0,x)$ and $l = \ell - \sigma$, which may be assumed not integer) it then follows that $\Phi_j(0,x)$ converges in the C^1 norm to a C^1 function $\phi\colon \mathbb{T}^d \to \mathbb{R}^d \times \mathbb{T}^d$, which is ε-close to the identity, and, because of (128),

$$\begin{aligned}\phi(x+\omega t) &= \lim \Phi_j(0, x+\omega t)\\ &= \lim \phi_H^t \circ \Phi_j(0,x) = \phi_H^t \circ \phi(x) \qquad (130)\end{aligned}$$

showing that $\phi(\mathbb{T}^d)$ is a C^1 KAM torus for H (note that the map ϕ is close to the trivial embedding $x \to (y,x)$).

Future Directions

In this section we review in a schematic and informal way some of the most important developments, applications and possible future directions of KAM theory. For exhaustive surveys we refer to [9], Sect. 6.3 of [6] or [60].

1. *Structure of the Kolmogorov set and Whitney smoothness*
   The Kolmogorov set (i. e., the union of KAM tori), in nearly-integrable systems, tends to fill up (in measure) the whole phase space as the strength of the perturbation goes to zero (compare Remark 7–(v) and Remark 10). A natural question is: what is the global geometry of KAM tori?
   It turns out that KAM tori smoothly interpolate in the following sense. *For ε small enough, there exists a C^∞ symplectic diffeomorphism ϕ_* of the phase space $\mathcal{M} = B \times \mathbb{T}^d$ of the nearly-integrable, non-degenerate Hamiltonians $H = K(y) + \varepsilon P(y,x)$ and a Cantor set $C_* \subset B$ such that, for each $y' \in C_*$, the set $\phi_*^{-1}(\{y'\} \times \mathbb{T}^d)$ is a KAM torus for H*}; in other words, the Kolmogorov set is a smooth, symplectic deformation of the fiber bundle $C_* \times \mathbb{T}^d$. Still another way of describing this result is that *there exists a smooth function $K_*\colon B \to \mathbb{R}$ such that $(K + \varepsilon P) \circ \phi_*$ and K_* agree, together with their derivatives, on $C_* \times \mathbb{T}^d$*: we may, thus, say that, in general, nearly-integrable Hamiltonian systems are integrable on Cantor sets of relative big measure.
   Functions defined on closed sets which admit C^k extensions are called *Whitney smooth*; compare [64], where H. Whitney gives a sufficient condition, based on Taylor uniform approximations, for a function to be Whitney C^k.
   The proof of the above result – given, independently, in [50] and [19] in, respectively, the differentiable and the analytic case – follows easily from the following lemma[40]:

Lemma 4 *Let $C \subset \mathbb{R}^d$ a closed set and let $\{f_j\}$, $f_0 = 0$, be a sequence of functions analytic on $W_j := \cup_{y \in C} D(y, r_j)$. Assume that $\sum_{j\ge1} \sup_{W_j} |f_j - f_{j-1}| r_j^{-k} < \infty$. Then, f_j converges uniformly to a function f, which is C^k in the sense of Whitney on C.*

   Actually, the dependence upon the angles x' of ϕ_* is analytic and it is only the dependence upon $y' \in C_*$ which is Whitney smooth ("anisotropic differentiability", compare Sect. 2 in [50]).
   For more information and a systematic use of Whitney differentiability, see [9].
2. *Power series expansions*
   KAM tori $\mathcal{T}_{\omega,\varepsilon} = \phi_\varepsilon(\mathbb{T}^d)$ of nearly-integrable Hamiltonians correspond to quasi-periodic trajectories $z(t;\theta,\varepsilon) = \phi_\varepsilon(\theta + \omega t) = \phi_H^t(z(0;\theta,\varepsilon))$; compare items (d) and (e) of Sect. "Introduction" and Remark 2–(i) above. While the *actual* existence of such quasi-periodic motions was proven, for the first time, only thanks to KAM theory, the *formal* existence, in terms of formal ε-power series[41] was well known in the nineteenth century to mathematicians and astronomers (such as Newcombe, Lindstedt and, especially, Poincaré; compare [49], vol. II). Indeed, formal power solutions of nearly-integrable Hamiltonian equations are not difficult to construct (see, e. g., Sect. 7.1 of [12]) but *direct proofs* of the convergence of the series, i. e., proofs not based on Moser's "indirect" argument recalled in Remark 7–(iii) but, rather, based upon direct estimates on the kth ε-expansion coefficient, are quite difficult and were carried out only in the late eighties by H. Eliasson [27]. The difficulty is due to the fact that, in order to prove the convergence of the Taylor–Fourier expansion of such series, one has to recognize compensations among huge terms with different signs[42]. After Eliasson's breakthrough based upon a semi-direct method (compare the "Postscript 1996" at p. 33 of [27]), fully direct proofs were published in 1994 in [30] and [18].
3. *Non-degeneracy assumptions*
   Kolmogorov's non-degeneracy assumption (70) can be generalized in various ways. First of all, Arnold pointed out in [2] that the condition

$$\det \begin{pmatrix} K_{yy} & K_y \\ K_y & 0 \end{pmatrix} \neq 0\,, \qquad (131)$$

   (this is a $(d+1)\times(d+1)$ symmetric matrix where last column and last row are given by the $(d+1)$-vector $(K_y, 0)$) which is independent from condition (70),

is also sufficient to construct KAM tori. Indeed, (131) may be used to construct *iso-energetic* KAM tori, i. e., tori on a *fixed energy level*[43] E.

More recently, Rüssmann [57] (see, also, [58]), using results of Diophantine approximations on manifolds due to Pyartly [52], formulated the following condition (the "Rüssmann non-degeneracy condition"), which is essentially necessary and sufficient for the existence of a positive measure set of KAM tori in nearly-integrable Hamiltonian systems: *the image* $\omega(B) \subset \mathbb{R}^d$ *of the unperturbed frequency map* $y \to \omega(y) := K_y(y)$ *does not lie in any hyperplane passing through the origin*. We simply add that one of the prices that one has to pay to obtain these beautiful general results is that one cannot fix the frequency ahead of time.

For a thorough discussion of this topic, see Sect. 2 of [60].

4. *Some physical applications*

We now mention a short (and non-exhaustive) list of important physical application of KAM theory. For more information, see Sect. 6.3.9 of [6] and references therein.

4.1. *Perturbation of classical integrable systems*

As mentioned above (Remark 1–(iii)), one of the main original motivations of KAM theory is the perturbation theory for nearly-integrable Hamiltonian systems. Among the most famous classical integrable systems we recall: one-degree-of freedom systems; Keplerian two-body problem, geodesic motion on ellipsoids; rotations of a heavy rigid body with a fixed point (for special values of the parameters: Euler's, Lagrange's, Kovalevskaya's and Goryachev–Chaplygin's cases); Calogero–Moser's system of particles; see, Sect. 5 of [6] and [47].

A first step, in order to apply KAM theory to such classical systems, is to explicitly construct action-angle variables and to determine their analyticity properties, which is in itself a technically non-trivial problem. A second problem which arises, especially in Celestial Mechanics, is that the integrable (transformed) Hamiltonian governing the system may be highly degenerate (*proper degeneracies* – see Sect. 6.3.3, B of [6]), as is the important case of the planetary n-body problem. Indeed, the first complete proof of the existence of a positive measure set of invariant tori[44] for the planetary $(n+1)$ problem (one body with mass 1 and n bodies with masses smaller than ε) has been published only in 2004 [29]. For recent reviews on this topic, see [16].

4.2. *Topological trapping in low dimensions*

The general 2-degree-of-freedom nearly-integrable Hamiltonian exhibits a kind of particularly strong stability: the phase space is 4-dimensional and the energy levels are 3-dimensional; thus KAM tori (which are two-dimensional and which are guaranteed, under condition (131), by the iso-energetic KAM theorem) *separate* the energy levels and orbits lying between two KAM tori will remain forever trapped in the invariant region. In particular the evolution of the action variables stays forever close to the initial position ("total stability").

This observation is originally due to Arnold [2]; for recent applications to the stability of three-body problems in celestial mechanics see [13] and item **4.4** below.

In higher dimension this topological trapping is no longer available, and in principle nearby any point in phase space it may pass an orbit whose action variables undergo a displacement of order one ("Arnold's diffusion"). A rigorous complete proof of this conjecture is still missing[45].

4.3. *Spectral Theory of Schrödinger operators*

KAM methods have been applied also very successfully to the spectral analysis of the one-dimensional Schrödinger (or "Sturm–Liouville") operator on the real line $\mathbb{R}$

$$L := -\frac{d^2}{dt^2} + v(t)\,, \quad t \in \mathbb{R}\,. \tag{132}$$

If the "potential" v is bounded then there exists a unique self-adjoint operator on the real Hilbert space $\mathcal{L}^2(\mathbb{R})$ (the space of Lebesgue square-integrable functions on $\mathbb{R}$) which extends L above on C_0^2 (the space of twice differentiable functions with compact support). The problem is then to study the spectrum $\sigma(L)$ of L; for generalities, see [23].

If v is periodic, then $\sigma(L)$ is a continuous band spectrum, as it follows immediately from Floquet theory [23]. Much more complicated is the situation for quasi-periodic potentials $v(t) := V(\omega t) = V(\omega_1 t, \ldots, \omega_n t)$, where V is a (say) real-analytic function on $\mathbb{T}^n$, since small-divisor problems appear and the spectrum can be nowhere dense. For a beautiful classical exposition, see [47], where, in particular, interesting connections with mechanics are discussed[46]; for deep developments of generalization of Floquet theory to quasi-periodic Schrödinger operators ("reducibility"), see [26] and [7].

4.4. *Physical stability estimates and break-down thresholds*

KAM Theory is perturbative and works if the parameter ε measuring the strength of the perturbation is small enough. It is therefore a fundamental question: *how small ε has to be in order for KAM results to hold*. The first concrete applications were extremely discouraging: in 1966, the French astronomer M. Hénon [32] pointed out that Moser's theorem applied to the restricted three-body problem (i. e., the motion of an asteroid under the gravitational influence of two unperturbed primary bodies revolving on a given Keplerian ellipse) yields existence of invariant tori if the mass ratio of the primaries is less than[47] 10^{-52}. Since then, much progress has been made and very recently, in [13], it has been shown via a computer-assisted proof[48], that, for a restricted-three body model of a subsystem of the Solar system (namely, Sun, Jupiter and Asteroid Victoria), KAM tori exist for the "actual" physical values (in that model the Jupiter/Sun mass ratio is about 10^{-3}) and, in this mathematical model – thanks to the trapping mechanism described in item **4.2** above – they trap the actual motion of the subsystem.

From a more theoretical point of view, we notice that, (compare Remark 2–(ii)) KAM tori (with a fixed Diophantine frequency) are analytic in ε; on the other hand, it is known, at least in lower dimensional settings (such as twist maps), that above a certain critical value KAM tori (curves) cannot exist ([39]). Therefore, there must exist a critical value $\varepsilon_c(\omega)$ ("breakdown threshold") such that, for $0 \le \varepsilon < \varepsilon_c\omega$, the KAM torus (curve) $\mathcal{T}_{\omega,\varepsilon}$ exists, while for $\varepsilon > \varepsilon_c(\omega)$ does not. The mathematical mechanism for the breakdown of KAM tori is far from being understood; for a brief review and references on this topic, see, e. g., Sect. 1.4 in [13].

5. *Lower dimensional tori*

In this item we consider (very) briefly, the existence of quasi-periodic solutions with a number of frequencies smaller than the number of degrees of freedom[49]. Such solutions span *lower dimensional* (non Lagrangian) *tori*. Certainly, this is one of the most important topics in modern KAM theory, not only in view of applications to classical problems, but especially in view of extensions to infinite dimensional systems, namely PDEs (Partial Differential Equations) with a Hamiltonian structure; see, item **6** below. For a recent, exhaustive review on lower dimensional tori (in finite dimensions), we refer the reader to [60].

In 1965 V.K. Melnikov [41] stated a precise result concerning the persistence of *stable* (or "elliptic") lower dimensional tori; the hypotheses of such results are, now, commonly referred to as "Melnikov conditions". However, a proof of Melnikov's statement was given only later by Moser [45] for the case $n = d - 1$ and, in the general case, by H. Eliasson in [25] and, independently, by S.B. Kuksin [37]. The *unstable* ("partially hyperbolic") case (i. e., the case for which the lower dimensional tori are linearly unstable and lie in the intersection of stable and unstable Lagrangian manifolds) is simpler and a complete perturbation theory was already given in [45], [31] and [66] (roughly speaking, the normal frequencies to the torus do not resonate with the inner (or "proper") frequencies associated with quasi-periodic motion). Since then, Melnikov conditions have been significantly weakened and much technical progress has been made; see [60], Sects. 5, 6 and 7, and references therein.

To illustrate a typical situation, let us consider a Hamiltonian system with $d = n + m$ degrees of freedom, governed by a Hamiltonian function of the form

$$H(y, x, v, u; \xi) = K(y, v, u; \xi) + \varepsilon P(y, x, v, u; \xi), \quad (133)$$

where $(y, x) \in \mathbb{T}^n \times \mathbb{R}^n$, $(v, u) \in \mathbb{R}^{2m}$ are pairs of standard symplectic coordinates and ξ is a real parameter running over a compact set $\Pi \subset \mathbb{R}^n$ of positive Lebesgue measure[50]; K is a Hamiltonian admitting the n-torus

$$\mathcal{T}_0^n(\xi) := \{y = 0\} \times \mathbb{T}^n \times \{v = u = 0\}, \quad \xi \in \Pi,$$

as invariant linearly stable invariant torus and is assumed to be in the normal form:

$$K = E(\xi) + \omega(\xi) \cdot y + \frac{1}{2} \sum_{j=1}^{m} \Omega_j(\xi)(u_j^2 + v_j^2). \quad (134)$$

The ϕ_K^t flow decouples in the linear flow $x \in \mathbb{T}^n \to x + \omega(\xi)t$ times the motion of m (decoupled) harmonic oscillators with characteristic frequencies $\Omega_j(\xi)$ (sometimes referred to as *normal frequencies*). Melnikov's conditions (in the form proposed in [51]) reads as follows: assume that ω is a Lipschitz homeomorphism; let $\Pi_{k,l}$ denote the "resonant parameter set" $\{\xi \in \Pi : \omega(\xi) \cdot k + \Omega \cdot (\xi) = 0\}$ and assume

$$\begin{cases} \Omega_i(\xi) > 0, \quad \Omega_i(\xi) \neq \Omega_j(\xi), \quad \forall \xi \in \Pi, \forall i \neq j \\ \text{meas } \Pi_{k,l} = 0, \quad \forall k \in \mathbb{Z}^n \backslash \{0\}, \quad \forall l \in \mathbb{Z}^m : |l| \le 2. \end{cases} \quad (135)$$

Under these assumptions and if $|\varepsilon|$ is small enough, there exists a (Cantor) subset of parameters $\Pi_* \subset \Pi$ of posi-

tive Lebesgue measure such that, to each $\xi \in \Pi_*$, there corresponds a n-dimensional, linearly stable H-invariant torus $\mathcal{T}_\varepsilon^n(\xi)$ on which the H flow is analytically conjugated to $x \to x + \omega_*(\xi)t$ where ω_* is a Lipschitz homeomorphism of Π_* assuming Diophantine values and close to ω.

This formulation has been borrowed from [51], to which we refer for the proof; for the differentiable analog, see [22].

Remark 12 The small-divisor problems arising in the perturbation theory of the above lower dimensional tori are of the form

$$\omega \cdot k - l \cdot \Omega\,, \quad |l| \le 2\,, \quad |k| + |l| \neq 0\,, \tag{136}$$

where one has to regard the normal frequency Ω as *functions of the inner frequencies* ω and, at first sight, one has – in J. Moser words – a *lack-of-parameter* problem. To overcome this intrinsic difficulty, one has to give up full control of the inner frequencies and construct, iteratively, n-dimensional sets (corresponding to smaller and smaller sets of ξ-parameters) on which the small divisors are controlled; for more motivations and informal explanations on lower dimensional small divisor problems, see, Sects. 5, 6 and 7 of [60].

6. *Infinite dimensional systems*

As mentioned above, the most important recent developments of KAM theory, besides the full applications to classical n-body problems mentioned above, is the successful extension to infinite dimensional settings, so as to deal with certain classes of partial differential equations carrying a Hamiltonian structure. As a typical example, we mention the non-linear wave equation of the form

$$\begin{aligned} u_{tt} - u_{xx} + V(x)u &= f(u)\,, \\ f(u) = O(u^2)\,, \quad 0 < x < 1\,, &\quad t \in \mathbb{R}\,. \end{aligned} \tag{137}$$

These extensions allowed, in the pioneering paper [63], establishing the existence of small-amplitude quasi-periodic solutions for (137), subject to Dirichlet or Neumann boundary conditions (on a finite interval for odd and analytic nonlinearities f); the technically more difficult periodic boundary condition case was considered later; compare [38] and references therein.

A technical discussion of these topics goes far beyond the scope of the present article and, for different equations, techniques and details, we refer the reader to the review article [38].

A The Classical Implicit Function Theorem

Here we discuss the classical Implicit Function Theorem for complex functions from a quantitative point of view. The following Theorem is a simple consequence of the Contraction Lemma, which asserts that a contraction Φ on a closed, non-empty metric space[51] X has a unique fixed point, which is obtained as $\lim_{j\to\infty} \Phi^j(u_0)$ for any[52] $u_0 \in X$. As above, $D^n(y_0, r)$ denotes the ball in $\mathbb{C}^n$ of center y_0 and radius r.

Theorem 3 (Implicit Function Theorem) *Let*

$$\begin{aligned} F\colon (y,x) \in D^n(y_0,r) \times D^m(x_0,s) \subset \mathbb{C}^{n+m} \\ \to F(y,x) \in \mathbb{C}^n \end{aligned}$$

be continuous with continuous Jacobian matrix F_y; assume that $F_y(y_0, x_0)$ is invertible and denote by T its inverse; assume also that

$$\begin{aligned} \sup_{D(y_0,r)\times D(x_0,s)} \|\mathbb{1}_n - TF_y(y,x)\| &\le \frac{1}{2}\,, \\ \sup_{D(x_0,s)} |F(y_0,x)| &\le \frac{r}{2\|T\|}\,. \end{aligned} \tag{138}$$

Then, all solutions $(y,x) \in D(y_0,r)\times D(x_0,s)$ of $F(y,x) = 0$ are given by the graph of a unique continuous function $g\colon D(x_0,s) \to D(y_0,r)$ satisfying, in particular,

$$\sup_{D(x_0,s)} |g| \le 2\|T\| \sup_{D(x_0,s)} |F(y_0,\cdot)|\,. \tag{139}$$

Proof Let $X = C(D^m(x_0,s), D^n(y_0,r))$ be the closed ball of continuous function from $D^m(x_0,s)$ to $D^n(y_0,r)$ with respect to the sup-norm $\|\cdot\|$ (X is a non-empty metric space with distance $d(u,v) := \|u - v\|$) and denote $\Phi(y;x) := y - TF(y,x)$. Then, $u \to \Phi(u) := \Phi(u,\cdot)$ maps $C(D^m(x_0,s))$ into $C(\mathbb{C}^m)$ and, since $\partial_y\Phi = \mathbb{1}_n - TF_y(y,x)$, from the first relation in (138), it follows that $u \to \Phi(u)$ is a contraction. Furthermore, for any $u \in C(D^m(x_0,s), D^n(y_0,r))$,

$$\begin{aligned} |\Phi(u) - y_0| &\le |\Phi(u) - \Phi(y_0)| + |\Phi(y_0) - y_0| \\ &\le \frac{1}{2}\|u - y_0\| + \|T\|\|F(y_0,x)\| \\ &\le \frac{1}{2}r + \|T\|\frac{r}{2\|T\|} = r\,, \end{aligned}$$

showing that $\Phi\colon X \to X$. Thus, by the Contraction Lemma, there exists a unique $g \in X$ such that $\Phi(g) = g$, which is equivalent to $F(g,x) = 0\ \forall x$. If $F(y_1,x_1) = 0$ for some $(y_1,x_1) \in D(y_0,r) \times D(x_0,s)$, it follows that $|y_1 - g(x_1)| = |\Phi(y_1;x_1) - \Phi(g(x_1),x_1)| \le \alpha|y_1 - g(x_1)|$,

which implies that $y_1 = g(x_1)$ and that all solutions of $F = 0$ in $D(y_0, r) \times D(x_0, s)$ coincide with the graph og g. Finally, (139) follows by observing that $\|g - y_0\| = \|\Phi(g) - y_0\| \leq \|\Phi(g) - \Phi(y_0)\| + \|\Phi(y_0) - y_0\| \leq \frac{1}{2}\|g - y_0\| + \|T\|\|F(y_0, \cdot)\|$, finishing the proof. □

Additions:

(i) If F is periodic in x or/and real on reals, then (by uniqueness) so is g;
(ii) If F is analytic, then so is g (Weierstrass Theorem, since g is attained as uniform limit of analytic functions);
(iii) The factors 1/2 appearing in the right-hand sides of (138) may be replaced by, respectively, α and β for any positive α and β such that $\alpha + \beta = 1$.

Taking $n = m$ and $F(y, x) = f(y) - x$ for a given $C^1(D(y_0, r), \mathbb{C}^n)$ function, one obtains the

Theorem 4 (Inverse Function Theorem) *Let $f\colon y \in D^n(y_0, r) \to \mathbb{C}^n$ be a C^1 function with invertible Jacobian $f_y(y_0)$ and assume that*

$$\sup_{D(y_0,r)} \|\mathbb{1}_n - Tf_y\| \leq \frac{1}{2}\,, \quad T := f_y(y_0)^{-1}\,, \tag{140}$$

then there exists a unique C^1 function $g\colon D(x_0, s) \to D(y_0, r)$ with $x_0 := f(y_0)$ and $s := r/(2\|T\|)$ such that $f \circ g(x) = \mathrm{id} = g \circ f$.

Additions analogous to the above also hold in this case.

B Complementary Notes

1 Actually, the first instance of a small divisor problem solved analytically is the linearization of the germs of analytic functions and is due to C.L. Siegel [61].

2 The well-known Newton's tangent scheme is an algorithm, which allows us to find roots (zeros) of a smooth function f in a region where the derivative f' is bounded away from zero. More precisely, if x_n is an "approximate solution" of $f(x) = 0$, i. e., $f(x_n) := \varepsilon_n$ is small, then the next approximation provided by Newton's tangent scheme is $x_{n+1} := x_n - f(x_n)/f'(x_n)$ [which is the intersection with x-axis of the tangent to the graph of f passing through $(x_n, f(x_n))$] and, in view of the definition of ε_n and Taylor's formula, one has that $\varepsilon_{n+1} := f(x_{n+1}) = \frac{1}{2}f''(\xi_n){\varepsilon_n}^2/(f'(x_n)^2$ (for a suitable ξ_n) so that $\varepsilon_{n+1} = O(\varepsilon_n^2) = O(\varepsilon_1^{2^n})$ and, in the iteration, x_n will converge (at a super-exponential rate) to a root $\bar{x}$ of f. This type of extremely fast convergence will be typical in the analyzes considered in the present article.

3 The elements of $\mathbb{T}^d$ are equivalence classes $x = \bar{x} + 2\pi\mathbb{Z}^d$ with $\bar{x} \in \mathbb{R}^d$. If $x = \bar{x} + 2\pi\mathbb{Z}^d$ and $y = \bar{y} + 2\pi\mathbb{Z}^d$ are elements of $\mathbb{T}^d$, then their distance $d(x, y)$ is given by $\min_{n\in\mathbb{Z}^d} |\bar{x} - \bar{y} + 2\pi n|$ where $|\cdot|$ denotes the standard euclidean norm in $\mathbb{R}^n$; a smooth (analytic) function on $\mathbb{T}^d$ may be viewed as ("identified with") a smooth (analytic) function on $\mathbb{R}^d$ with period 2π in each variable. The torus $\mathbb{T}^d$ endowed with the above metric is a real-analytic, compact manifold. For more information, see [62].

4 A symplectic form on an (even dimensional) manifold is a closed, non-degenerate differential 2-form. The symplectic form $\alpha = \mathrm{d}y \wedge \mathrm{d}x$ is actually *exact* symplectic, meaning that $\alpha = \mathrm{d}(\sum_{i=1} y_i \mathrm{d}x_i)$. For general information see [5].

5 For general facts about the theory of ODE (such as Picard theorem, smooth dependence upon initial data, existence times, ...) see, e. g., [23].

6 This terminology is due to that fact the the x_j are "adimensional" angles, while analyzing the physical dimensions of the quantities appearing in Hamilton's equations one sees that $\dim(y) \times \dim(x) = \dim H \times \dim(t)$ so that y has the dimension of an energy (the Hamiltonian) times the dimension of time, i. e., by definition, the dimension of an action.

7 This terminology is due to the fact that a classical mechanical system of d particles of masses $m_i > 0$ and subject to a potential $V(q)$ with $q \in A \subset \mathbb{R}^d$ is governed by a Hamiltonian of the form $\sum_{j=1}^d p_j^2/2m_j + V(q)$ and d may be interpreted as the (minimal) number of coordinates necessary to physically describe the system.

8 To be precise, (6) should be written as $y(t) = v(\pi_{\mathbb{T}^d}(\omega t))$, $x(t) = \pi_{\mathbb{T}^d}(\omega t + u(\pi_{\mathbb{T}^d}(\omega t)))$ where $\pi_{\mathbb{T}^d}$ denotes the standard projection of $\mathbb{R}^d$ onto $\mathbb{T}^d$, however we normally omit such a projection.

9 As standard, U_θ denotes the $(d \times d)$ Jacobian matrix with entries $(\partial U_i)/(\partial\theta_j) = \delta_{ij} + (\partial u_i)/(\partial\theta_j)$.

10 For generalities, see [5]; in particular, a Lagrangian manifold $L \subset \mathcal{M}$ which is a graph over $\mathbb{T}^d$ admits a "generating function", i. e., there exists a smooth function $g\colon \mathbb{T}^d \to \mathbb{R}$ such that $L = \{(y, x)\colon y = g_x(x), x \in \mathbb{T}^d\}$.

11 Compare [54] and references therein. We remark that, if $B(\omega_0, r)$ denote the ball in $\mathbb{R}^d$ of radius r centered at ω_0 and fix $\tau > d - 1$, then one can prove that the Lebesgue measure of $B(y_0, r)\backslash\mathcal{D}^d_{\kappa,\tau}$ can be bounded by $c_d\kappa r^{d-1}$ for a suitable constant c_d depending only on d; for a simple proof, see, e.g, [21].

12 The sentence "can be put into the form" means "there exists a symplectic diffeomorphism $\phi\colon (y, x) \in \mathcal{M} \to$

$(\eta, \xi) \in \mathcal{M}$ such that $H \circ \phi$ has the form (10)"; for multi-indices α, $|\alpha| = \alpha_1 + \cdots + \alpha_d$ and $\partial_y^\alpha = \partial_{y_1}^{\alpha_1} \cdots \partial_{y_d}^{\alpha_d}$; the vanishing of the derivatives of a function $f(y)$ up to order k in the origin will also be indicated through the expression $f = O(|y|^{k+1})$.

13 **Notation:** If A is an open set and $p \in \mathbb{N}$, then the C^p-norm of a function $f: x \in A \to f(x)$ is defined as $\|f\|_{C^p(A)}$: $\sup_{|\alpha| \le p} \sup_A |\partial_x^\alpha f|$.

14 **Notation:** If f is a scalar function f_y is a d-vector; f_{yy} the Hessian matrix $(f_{y_i y_j})$; f_{yyy} the symmetric 3-tensor of third derivatives acting as follows: $f_{yyy} a \cdot b \cdot c := \sum_{i,j,k=1}^d (\partial^3 f)/(\partial y_i \partial y_j \partial y_k) a_i b_j c_k$.

15 **Notation:** If f is (a regular enough) function over $\mathbb{T}^d$, its Fourier coefficients are defined as $f_n := \int_{\mathbb{T}^d} f(x) e^{-in\cdot x} dx/(2\pi)^d$; where, as usual, $i = \sqrt{-1}$ denotes imaginary unit; for general information about Fourier series see, e. g., [34].

16 The choice of norms on finite dimensional spaces ($\mathbb{R}^d$, $\mathbb{C}^d$, space of matrices, tensors, etc.) is not particularly relevant for the analysis in this article (since changing norms will change d-depending constants); however for matrices, tensors (and, in general, linear operators), it is convenient to work with the "operator norm", i. e., the norm defined as $\|L\| = \sup_{u \ne 0} \|Lu\|/\|u\|$, so that $\|Lu\| \le \|L\| \|u\|$, an estimate, which will be constantly be used; for a general discussion on norms, see, e. g., [36].

17 As an example, let us work out the first two estimates, i. e., the estimates on $\|s_x\|_{\bar\xi}$ and $|b|$: actually these estimates will be given on a larger intermediate domain, namely, $W_{\bar\xi - \delta/3}$, allowing to give the remaining bounds on the smaller domain $W_{\bar\xi}$ (recall that W_s denotes the complex domain $D(0,s) \times \mathbb{T}_s^d$). Let $f(x) := P(0,x) - \langle P(0,\cdot)\rangle$. By definition of $\|\cdot\|_\xi$ and M, it follows that $\|f\|_\xi \le \|P(0,x)\|_\xi + \|\langle P(0,\cdot)\rangle\|_\xi \le 2M$. By **(P5)** with $p = 1$ and $\xi' = \xi - \delta/3$, one gets

$$\|s_x\|_{\xi - \frac{\delta}{3}} \le \bar B_1 \frac{2M}{\kappa} 3^{k_1} \delta^{-k_1} ,$$

which is of the form (53), provided $\bar c \ge (\bar B_1 2 \cdot 3^{k_1})/\kappa$ and $\bar\nu \ge k_1$. To estimate b, we need to bound first $|Q_{yy}(0,x)|$ and $|P_y(0,x)|$ for real x. To do this we can use Cauchy estimate: by **(P4)** with $p = 2$ and, respectively, $p = 1$, and $\xi' = 0$, we get

$$\|Q_{yy}(0,\cdot)\|_0 \le m B_2 C \xi^{-2} \le m B_2 C \delta^{-2} , \quad \text{and}$$
$$\|P_y(0,x)\|_0 \le m B_1 M \delta^{-1} ,$$

where $m = m(d) \ge 1$ is a constant which depend on the choice of the norms, (recall also that $\delta < \xi$). Putting these bounds together, one gets that $|b|$ can be bounded by the r.h.s. of (53) provided $\bar c \ge m(B_2 \bar B_1 2 \cdot 3^{k_1} \kappa^{-1} + B_1)$, $\mu \ge 2$ and $\bar\nu \ge k_1 + 2$. The other bounds in (53) follow easily along the same lines.

18 We sketch here the **proof of Lemma 1**. The defining relation $\psi_\varepsilon \circ \varphi = \mathrm{id}$ implies that $\alpha(x') = -a(x' + \varepsilon\alpha(x'))$, where $\alpha(x')$ is short for $\alpha(x';\varepsilon)$ and that equation is a fixed point equation for the non-linear operator $f: u \to f(u) := -a(\mathrm{id} + \varepsilon u)$. To find a fixed point for this equation one can use a standard contraction Lemma (see [36]). Let Y denote the closed ball (with respect to the sup-norm) of continuous functions $u: \mathbb{T}_{\xi'}^d \to \mathbb{C}^d$ such that $\|u\|_{\xi'} \le \bar L$. By (54), $|\mathrm{Im}(x' + \varepsilon u(x'))| < \xi' + \varepsilon_0 \bar L < \xi' + \delta/3 = \bar\xi$, for any $u \in Y$, and any $x' \in \mathbb{T}_{\xi'}^d$; thus, $\|f(u)\|_{\xi',\varepsilon_*} \le \|a\|_{\bar\xi} \le \bar L$ by (53), so that $f: Y \to Y$; notice that, in particular, this means that f sends periodic functions into periodic functions. Moreover, (54) implies also that f is a contraction: if $u, v \in Y$, then, by the mean value theorem, $|f(u) - f(v)| \le \bar L |\varepsilon| |u - v|$ (with a suitable choice of norms), so that, by taking the sup-norm, one has $\|f(u) - f(v)\|_{\xi'} < \varepsilon_0 \bar L \|u - v\|_{\xi'} < \frac{1}{3}\|u - v\|_{\xi'}$ showing that f is a contraction. Thus, there exists a unique $\alpha \in Y$ such that $f(\alpha) = \alpha$. Furthermore, recalling that the fixed point is achieved as the uniform limit $\lim_{n\to\infty} f^n(0)$ $(0 \in Y)$ and since $f(0) = -a$ is analytic, so is $f^n(0)$ for any n and, hence, by Weierstrass Theorem on the uniform limit of analytic function (see [1]), the limit α itself is analytic. In conclusion, $\varphi \in \mathcal{B}_{\xi'}$ and (55) holds.

Next, for $(y', x) \in W_{\bar\xi}$, by (53), one has $|y' + \varepsilon\beta(y', x)| < \bar\xi + \varepsilon_0 \bar L < \bar\xi + \delta/3 = \xi$ so that (56) holds. Furthermore, since $\|\varepsilon a_x\|_{\bar\xi} < \varepsilon_0 \bar L < 1/3$ the matrix $\mathbb{1}_d + \varepsilon a_x$ is invertible with inverse given by the "Neumann series" $(\mathbb{1}_d + \varepsilon a_x)^{-1} = \mathbb{1}_d + \sum_{k=1}^\infty (-1)^k (\varepsilon a_x)^k =: \mathbb{1}_d + \varepsilon S(x;\varepsilon)$, so that (57) holds. The proof is finished.

19 From (59), it follows immediately that $\langle \partial_{y'}^2 Q_1(0,\cdot)\rangle = \langle \partial_y^2 Q(0,\cdot)\rangle + \varepsilon \langle \partial_{y'}^2 \widetilde Q(0,\cdot)\rangle = T^{-1}(\mathbb{1}_d + \varepsilon T \langle \partial_{y'}^2 \widetilde Q(0,\cdot)\rangle) =: T^{-1}(\mathbb{1}_d + \varepsilon R)$ and, in view of (51) and (59), we see that $\|R\| < L/(2C)$. Therefore, by (60), $\varepsilon_0 \|R\| < 1/6 < 1/2$, implying that $(1 + \varepsilon R)$ is invertible and $(\mathbb{1}_d + \varepsilon R)^{-1} = \mathbb{1}_d + \sum_{k=1}^\infty (-1)^k \varepsilon^k R^k =: 1 + \varepsilon D$ with $\|D\| \le \|R\|/(1 - |\varepsilon|\|R\|) < L/C$. In conclusion, $T_1 = (1 + \varepsilon R)^{-1} T = T + \varepsilon D T =: T + \varepsilon \widetilde T$, $\|\widetilde T\| \le \|D\| C \le (L/C) C = L$.

20 Actually, there is quite some freedom in choosing the sequence $\{\xi_j\}$ provided the convergence is not too fast; for general discussion, see, [56], or, also, [10] and [14].

21 In fact, denoting by B_* the real d-ball centered at 0 and of radius $\theta\xi_*$ for $\theta \in (0,1)$, from Cauchy estimate (47) with $\xi = \xi_*$ and $\xi' = \theta\xi_*$, one has $\|\phi_* -$

$\mathrm{id}\|_{C^p(B_*\times\mathbb{T}^d)} = \sup_{B_*\times\mathbb{T}^d}\sup_{|\alpha|+|\beta|\le p}|\partial_y^\alpha\partial_x^\beta(\phi_*-\mathrm{id})| \le \sup_{|\alpha|+|\beta|\le p}\|\partial_y^\alpha\partial_x^\beta(\phi_*-\mathrm{id})\|_{\theta\xi_*} \le B_p\|\phi_*-\mathrm{id}\|_{\xi_*}1/(\theta\xi_*)^p \le \mathrm{const}_p|\varepsilon|$ with $\mathrm{const}_p := B_pDBM1/(\theta\xi_*)^p$. An identical estimate holds for $\|Q_*-Q\|_{C^p(B_*\times\mathbb{T}^d)}$.

22 Also very recently ε-power series expansions have been shown to be a very powerful tool; compare [13].

23 A function $f\colon A\subset\mathbb{R}^n\to\mathbb{R}^n$ is Lipschitz on A if there exists a constant ("Lipschitz constant") $L>0$ such that $|f(x)-f(y)|\le L|x-y|$ for all $x,y\in A$. For a general discussion on how Lebesgue measure changes under Lipschitz mappings, see, e. g., [28]. In fact, the dependence of ϕ_* on $\bar{y}$ is much more regular, compare Remark 11.

24 In fact, notice that inverse powers of κ appear through (48) (inversion of the operator D_ω), therefore one sees that the terms in the first line of (53) may be replaced by $\tilde{c}\kappa^{-2}$ (in defining a one has to apply the operator D_ω^{-1} twice) but then in $P^{(1)}$ (see (26)) there appears $\|\beta\|^2$, so that the constant c in the second line of (53) has the form (72); since $\kappa<1$, one can replace in (53) c with $\hat{c}\kappa^{-4}$ as claimed.

25 **Proof of Claim C** Let $H_0:=H$, $E_0:=E$, $Q_0:=Q$, $K_0:=K$, $P_0:=P$, $\xi_0:=\xi$ and let us assume (*inductive hypothesis*) that we can iterate the Kolmogorov transformation j times obtaining j symplectic transformations $\phi_{i+1}\colon W_{\xi_{i+1}}\to W_{\xi_i}$, for $0\le i\le j-1$, and j Hamiltonians $H_{i+1}=H_i\circ\phi_{i+1}=K_i+\varepsilon^{2^i}P_i$ real-analytic on W_{ξ_i} such that

$$\begin{aligned}&|\omega|,|E_i|,\|Q_i\|_{\xi_i},\|T_i\|<C\,,\\&|\varepsilon|^{2^i}L_i:=|\varepsilon|^{2^i}cC^\mu\delta_0^{-\nu}2^{\nu i}M_i\le\frac{\delta_i}{3}\,,\\&\qquad\forall\,0\le i\le j-1\,.\end{aligned}\tag{$*$}$$

By $(*)$, Kolmogorov iteration (**Step 2**) can be applied to H_i and therefore all the bounds described in paragraph **Step 2** hold (having replaced $H,E,\dots,\xi,\delta,H',E',\dots,\xi'$ with, respectively, $H_i,E_i,\dots,\xi_i,\delta_i,H_{i+1},E_{i+1},\dots,\xi_{i+1}$); in particular (see (61)) one has, for $0\le i\le j-1$ (and for any $|\varepsilon|\le\varepsilon_0$),

$$\begin{cases}|E_{i+1}|\le|E_i|+|\varepsilon|^{2^i}L_i\,,\\ \quad\|Q_{i+1}\|_{\xi_{i+1}}\le\|Q_i\|_{\xi_i}+|\varepsilon|^{2^i}L_i\,,\\ \quad\|\phi_{i+1}-\mathrm{id}\|_{\xi_{i+1}}\le|\varepsilon|^{2^i}L_i\\ M_{i+1}\le M_iL_i\end{cases}\tag{C.1}$$

Observe that the definition of D, B and L_I, $|\varepsilon|^{2^j}L_j(3C\delta_j^{-1}) =: DB^j|\varepsilon|^{2^j}M_j$, so that $L_i<DB^iM_i$, thus by the second line in (C.1), for any $0\le i\le j-1$, $|\varepsilon|^{2^{i+1}}M_{i+1} < DB_i(M_i|\varepsilon|^{2^i})^2$, which iterated, yields (66) for $0\le i\le j$. Next, we show that, thanks to (65), $(*)$ holds also for $i=j$ (and this means that Kolmogorov's step can be iterated an infinite number of times). In fact, by $(*)$ and the definition of C in (64): $|E_j|\le|E|+\sum_{i=0}^{j-1}\varepsilon_0^{2^i}L_i\le|E|+\frac{1}{3}\sum_{i\ge0}\delta_i<|E|+\frac{1}{6}\sum_{i\ge1}2^{-i}<|E|+1<C$. The bounds for $\|Q_i\|$ and $\|T_i\|$ are proven in an identical manner. Now, by $(66)_{i=j}$ and (65), $|\varepsilon|^{2^j}L_j(3\delta_j^{-1}) = DB^j|\varepsilon|^{2^j}M_j \le DB^j(DB\varepsilon_0M)^{2^j}/(DB^{j+1}) \le 1/B < 1$, which implies the second inequality in $(*)$ with $i=j$; the proof of the induction is finished and one can construct an *infinite sequence* of Kolmogorov transformations satisfying $(*)$, (C.1) and (66) *for all* $i\ge0$. To check (67), we observe that $|\varepsilon|^{2^i}L_i = \delta_0/(3\cdot2^i)DB^i|\varepsilon|^{2^i}M_i \le (1/2^{i+1})(|\varepsilon|DBM)^{2^i} \le (|\varepsilon|DBM/2)^{i+1}$ and therefore $\sum_{i\ge0}|\varepsilon|^{2^i}L_i\le\sum_{i\ge1}(|\varepsilon|DBM/2)^i\le|\varepsilon|DBM$. Thus, $\|Q-Q_*\|_{\xi_*}\le\sum_{i\ge0}\|\tilde{Q}_i\|_{\xi_i}\le|\varepsilon|^{2^i}L_i\le|\varepsilon|DBM$; and analogously for $|E-E_*|$ and $\|T-T_*\|$. To estimate $\|\phi_*-\mathrm{id}\|_{\xi_*}$, observe that $\|\Phi_i-\mathrm{id}\|_{\xi_i}\le\|\Phi_{i-1}\circ\phi_i-\phi_i\|_{\xi_i}+\|\phi_i-\mathrm{id}\|_{\xi_i}\le\|\Phi_{i-1}-\mathrm{id}\|_{\xi_{i-1}}+|\varepsilon|^{2^i}L_i$, which iterated yields $\|\Phi_i-\mathrm{id}\|_{\xi_i}\le\sum_{k=0}^i|\varepsilon|^{2^k}L_k\le|\varepsilon|DBM$: taking the limit over i completes the proof of (67) and the proof of Claim **C**.

26 In fact, observe: (i) given any integer vector $0\ne n\in\mathbb{Z}^d$ with $d\ge2$, one can find $0\ne m\in\mathbb{Z}^d$ such $n\cdot m=0$; (ii) the set $\{tn\colon t>0 \text{ and } n\in\mathbb{Z}^d\}$ is dense in $\mathbb{R}^d$; (iii) if U is a neighborhood of y_0, then $K_y(U)$ is a neighborhood of $\omega=K_y(y_0)$. Thus, by (ii) and (iii), in $K_y(U)$ there are infinitely many points of the form tn with $t>0$ and $n\in\mathbb{Z}^d$ to which correspond points $y(t,n)\in U$ such that $K_y(y(t,n))=tn$ and for any of such points one can find, by (i), $m\in\mathbb{Z}$ such that $m\cdot n=0$, whence $K_y(y(t,n))\cdot m=tn\cdot m=0$.

27 This fact was well known to Poincaré, who based on the above argument his non-existence proof of integral of motions in the general situation; compare Sect. 7.1.1, [6].

28 Compare (90) but observe, that, since $\hat{P}$ is a trigonometric polynomial, in view of Remark 9–(ii), g in (96) defines a real-analytic function on $D(y_0,\bar{r})\times\mathbb{T}^d_{\xi'}$ with a suitable $\bar{r}=\bar{r}(\varepsilon)$ and $\xi'<\xi$. Clearly it is important to see explicitly how the various quantities depend upon ε; this is shortly discussed after Proposition 2.

29 In fact: $\|\check{P}\|_{r,\xi-\delta/2}\le M\sum_{|n|>N}\mathrm{e}^{-|n|\delta/2}\le M\mathrm{e}^{-(\delta/4)N}\sum_{|n|>N}\mathrm{e}^{-|n|\delta/4}\le M\mathrm{e}^{-(\delta/4)N}\sum_{|n|>0}\mathrm{e}^{-|n|\delta/4}\le \mathrm{const}\,M\mathrm{e}^{-(\delta/4)N}\delta^{-d}\le|\varepsilon|M$ if (106) holds and N is taken as in (104).

30 Apply the IFT of Appendix "A The Classical Implicit Function Theorem" to $F(y,\eta) := K_y(y) + \eta\partial_y P_0(y) - K_y(y_0)$ defined on $D^d(y_0,\bar r)\times D^1(0,|\varepsilon|)$: using the mean value theorem, Cauchy estimates and (114), $\|\mathbb{1}_d - TF_y\| \le \|\mathbb{1}_d - TK_{yy}\| + |\varepsilon|\|\partial_y^2 P_0\| \le \|T\|\|K_{yyy}\|\bar r + \|T\||\varepsilon|\|\partial_y^2 P_0\| \le C^2 2\bar r/r + C|\varepsilon|4/r^2 M \le \frac14 + \frac18 < \frac12$; also: $2\|T\|\|F(y_0,\eta\| = 2\|T\||\eta|\partial_y P_0(y_0)\| < 2C|\varepsilon|M2/r \le 2CM\bar r^{-1}|\varepsilon| < \frac14\bar r$ (where last inequality is due to (114)), showing that conditions (138) are fulfilled. Equation (111) comes from (139) and (113) follows easily by repeating the above estimates.

31 Recall note 18 and notice that $(\mathbb{1}_d + A)^{-1} = \mathbb{1}_d + D$ with $\|D\| \le \|A\|/(1-\|A\|) \le 2\|A\| \le 20C^3M|\varepsilon|$, where last two inequalities are due to (113).

32 Lemma 1 can be immediately extended to the y'-dependent case (which appear as a dummy parameter) as far as the estimates are uniform in y' (which is the case).

33 By (118) and (54), $|\varepsilon|\|g_x\|_{\bar r,\bar\xi} \le |\varepsilon|rL \le r/2$ so that, by (116), if $y' \in D_{\bar r/2}(y_1)$, then $y' + \varepsilon g_x(y',\varphi(y',x')) \in D_r(y_0)$.

34 The first requirement in (123) is equivalent to require that $r_0 \le r$, which implies that if $\bar r$ is *defined* as the r.h.s. of (108), then $\bar r \le r/2$ as required in (110). Next, the first requirement in (114) at the $(j+1)$th step of the iteration translates into $16C^2 r_{j+1}/r_j \le 1$, which is satisfied, since, by definition, $r_{j+1}/r_j = (1/(2\gamma))^{\tau+1} \le (1/(2\gamma))^2 = 1/(36C^2) < 1/(16C^2)$. The second condition in (114), which at the $(j+1)$th step, reads $2CM_j r_{j+1}^{-2}|\varepsilon|^{2^j}$ is implied by $|\varepsilon|^{2^j}L_j \le \delta_j/(3C)$ (corresponding to (54)), which, in turn, is easily controlled along the lines explained in note 25.

35 An area-preserving twist mappings of an annulus $A = [0,1]\times\mathbb{S}^1$, ($\mathbb{S}^1 = \mathbb{T}^1$), is a symplectic diffeomorphism $f = (f_1,f_2)\colon (y,x)\in A \to f(y,x)\in A$, leaving invariant the boundary circles of A and satisfying the twist condition $\partial_y f_2 > 0$ (i. e., f twists clockwise radial segments). The theory of area preserving maps, which was started by Poincaré (who introduced such maps as section of the dynamics of Hamiltonian systems with two degrees of freedom), is, in a sense, the simplest nontrivial Hamiltonian context. After Poincaré the theory of area-preserving maps became, in itself, a very rich and interesting field of Dynamical Systems leading to very deep and important results due to Herman, Yoccoz, Aubry, Mather, etc; for generalities and references, see, e. g., [33].

36 It is not necessary to assume that K is real-analytic, but it simplify a little bit the exposition. In our case, we shall see that ℓ is related to the number σ in (66). We recall the definition of Hölder norms: If $\ell = \ell_0 + \mu$ with $\ell_0 \in \mathbb{Z}_+$ and $\mu \in (0,1)$, then $\|f\|_{C^\ell} := \|f\|_{C^\ell} + \sup_{|\alpha|=\ell_0}\sup_{0<|x-y|<1}|\partial^\alpha f(x) - \partial^\alpha f(y)|/|x-y|^\mu$; $C^\ell(\mathbb{R}^d)$ denotes the Banach space of functions with finite C^ℓ norm.

37 To obtain these new estimates, one can, first replace θ by $\sqrt{\theta}$ and then use the remark in the note 21 with $p = 1$; clearly the constant σ has to be increased by one unit with respect to the constant σ appearing in (69).

38 For general references and discussions about Lemma 2 and 3, see, [44] and [65]; an elementary detailed proof can be found, also, in [15].

39 **Proof of Claim M** The first step of the induction consists in constructing $\Phi_0 = \phi_0$: this follows from Kolmogorov's Theorem (i. e., Remark 7–(i) and Remark 11) with $\xi = \xi_1 = 1/2$ (assume, for simplicity, that Q is analytic on W_1 and note that $|\varepsilon|\|P^0\|_{\xi_1} \le |\varepsilon|\|P\|_{C^0}$ by the first inequality in (124)). Now, assume that (128) and (129) holds together with $C_i < 4C$ and $\|\partial(\Phi_i - \mathrm{id})\|_{\alpha\xi_{i+1}} < (\sqrt2 - 1)$ for $0 \le i \le j$ ($C_0 = C$ and C_i are as in (64) for, respectively, $K_0 := K$ and K_i). To determine ϕ_{j+1}, observe that, by (128), one has $H_{j+1}\circ\Phi_{j+1} = (K_{j+1} + \varepsilon P_{j+1})\circ\phi_{j+1}$ where $P_j := (P^{j+1} - P^j)\circ\Phi_j$, which is real-analytic on $W_{\alpha\xi_{j+1}}$; thus we may apply Kolmogorov's Theorem to $K_{j+1} + \varepsilon P_{j+1}$ with $\xi = \alpha\xi_{j+1}$ and $\theta = \alpha$; in fact, by the second inequality in (124), $\|P_{j+1}\|_{\alpha\xi_{j+1}} \le \|P^{j-1} - P^j\|_{X_{j+1}} \le c\|P\|_{C^\ell}\xi_{j+1}^\ell$ and the smallness condition (66) becomes $|\varepsilon|D\xi_{j+1}^{\ell-\sigma}$ (with $D := c_* c\|P\|_{C^\ell}(4C)^b 2^{\sigma/2}$), which is clearly satisfied for $|\varepsilon| < D^{-1}$. Thus, ϕ_{j+1} has been determined and (notice that $\alpha^2\xi_{j+1} = \xi_{j+1}/2 = \xi_{j+2}$) $\|\phi_{j+1} - \mathrm{id}\|_{\xi_{j+2}}$, $\partial(\|\phi_{j+1} - \mathrm{id})\|_{\xi_{j+2}} \le |\varepsilon|D\xi_{j+1}$. Let us now check the domain constraint $\Phi_j\colon W_{\alpha\xi_{j+1}} \to X_{\xi_{j+1}}$. By the inductive assumptions and the real-analyticity of Φ_j, one has that, for $z \in W_{\alpha\xi_{j+1}}$, $|\mathrm{Im}\Phi_j(z)| = |\mathrm{Im}(\Phi_j(z) - \Phi_j(\mathrm{Re}z))| \le |\Phi_j(z) - \Phi_j(\mathrm{Re}z)| \le \|\partial\Phi_j\|_{\alpha\xi_{j+1}}|\mathrm{Im}z| \le (1 + \|\partial(\Phi_i - \mathrm{id})\|_{\alpha\xi_{i+1}})\alpha\xi_{j+1} < \sqrt2\alpha\xi_{j+1} = \xi_{j+1}$ so that $\Phi_j\colon W_{\alpha\xi_{j+1}} \to X_{\xi_{j+1}}$. The remaining inductive assumptions in (129) with j replaced by $j+1$ are easily checked by arguments similar to those used in the induction proof of Claim **C** above.

40 See, e. g., the Proposition at page 58 of [14] with $g_j = f_j - f_{j-1}$. In fact, the lemma applies to the Hamiltonians H_j and to the symplectic map ϕ_j in (82) in Arnold's scheme with W_j in (81) and taking $C = C_* := \{y' = \lim_{j\to\infty} y_j(\omega)\colon \omega \in B \cap K_y^{-1}(\mathcal{D}_{\kappa,\tau}^d)\}$ and $y_j(\omega) := y_j$ is as in (82).

41 A formal ε-power series quasi-periodic trajectory, with rationally-independent frequency ω, for a nearly-integrable Hamiltonian $H(y,x;\varepsilon) := K(y) + \varepsilon P(y,x)$

is, by definition, a sequence of functions $\{z_k\} := (\{v_k\}, \{u_k\})$, real-analytic on $\mathbb{T}^d$ and such that $D_\omega z_k = J_{2d}\pi_k(\nabla H(\sum_{j=0}^{k-1}\varepsilon^j z_j))$ where $\pi_k(\cdot) := \frac{1}{k!}\partial_\varepsilon^k(\cdot)|_{\varepsilon=0}$; compare Remark 1–(ii) above.

42 In fact, Poincaré was not at all convinced of the convergence of such series: see chapter XIII, n° 149, entitled "Divergence des séries de M. Lindstedt", of his book [49].

43 Equation (70) guarantees that the map from y in the $(d-1)$-dimensional manifold $\{K = E\}$ to the $(d-1)$-dimensional real projective space $\{\omega_1 : \omega_2 : \cdots : \omega_d\} \subset \mathbb{RP}^{d-1}$ (where $\omega_i = K_{y_i}$) is a diffeomorphism. For a detailed proof of the "iso-energetic KAM Theorem", see, e. g., [24].

44 Actually, it is not known if such tori are KAM tori in the sense of the definitions given above!

45 The first example of a nearly-integrable system (with two parameters) exhibiting Arnold's diffusion(in a certain region of phase space) was given by Arnold in [4]; a theory for "a priori unstable systems" (i. e., the case in which the integrable system carries also a partially hyperbolic structure) has been worked out in [20] and in recent years a lot of literature has been devoted to study the "a priori unstable" case and to to try to attack the general problem (see, e. g., Sect. 6.3.4 of [6] for a discussion and further references). We mention that J. Mather has recently announced a complete proof of the conjecture in a general case [40].

46 Here, we mention briefly a different and very elementary connection with classical mechanics. To study the spectrum $\sigma(L)$ (L as above with a quasi-periodic potential $V(\omega_1 t, \ldots, \omega_n t)$) one looks at the equation $\ddot{q} = (V(\omega t) - \lambda)q$, which is the q-flow of the Hamiltonian ϕ_H^t $H = H(p, q, I, \varphi; \lambda) := p^2/2 + [\lambda - V(\varphi)]q^2/2$ where $(p, q) \in \mathbb{R}^2$ and $(I, \varphi) \in \mathbb{R}^n \times \mathbb{T}^n$ w.r.t. the standard form $\mathrm{d}p \wedge \mathrm{d}q + \mathrm{d}I \wedge \mathrm{d}\varphi$ and λ is regarded as a parameter. Notice that $\dot{\varphi} = \omega$ so that $\varphi = \varphi_0 + \omega t$ and that the (p, q) decouples from the I-flow, which is, then, trivially determined one the (p, q) flow is known. Now, the action-angle variables(J, θ) for the harmonic oscillator $p^2/2 + \lambda q^2/2$ are given by $J = r^2/\sqrt{\lambda}$ and (r, θ) are polar coordinates in the $(p, \sqrt{\lambda}q)$-plane; in such variables, H takes the form $H = \omega \cdot I + \sqrt{\lambda}J - V(\varphi)/\sqrt{\lambda}\sin^2\theta$. Now, if, for example V is small, this Hamiltonian is seen to be a perturbation of $(n+1)$ harmonic oscillator with frequencies $(\omega, \sqrt{\lambda})$ and it is remarkable that one can provide a KAM scheme, which preserves the linear-in-action structure of this Hamiltonian and selects the (Cantor) set of values of the frequency $\alpha = \sqrt{\lambda}$ for which the KAM scheme can be carried out so as to conjugate H to a Hamiltonian of the form $\omega \cdot I + \alpha J$, proving the existence of (generalized) quasi-periodic eigen-functions. For more details along these lines, see [14].

47 The value 10^{-52} is about the proton-Sun mass ratio: the mass of the Sun is about $1.991 \cdot 10^{30}$ kg, while the mass of a proton is about $1.672 \cdot 10^{-21}$ kg, so that (mass of a proton)/(mass of the Sun) $\simeq 8.4 \cdot 10^{-52}$.

48 "Computer-assisted proofs" are mathematical proofs, which use the computers to give rigorous upper and lower bounds on chains of long calculations by means of so-called "interval arithmetic"; see, e. g., Appendix C of [13] and references therein.

49 Simple examples of such orbits are equilibria and periodic orbits: in such cases there are no small-divisor problems and existence was already established by Poincaré by means of the standard Implicit Function Theorem; see [49], Volume I, chapter III.

50 Typically, ξ may indicate an initial datum y_0 and y the distance from such point or (equivalently, if the system is non-degenerate in the classical Kolmogorov sense) $\xi \to \omega(\xi)$ might be simply the identity, which amounts to consider the unperturbed frequencies as parameter; the approach followed here is that in [51], where, most interestingly, m is allowed to be ∞.

51 I. e., a map $\Phi : X \to X$ for which $\exists 0 < \alpha < 1$ such that $d(\Phi(u), \Phi(v)) \le \alpha d(u, v)$, $\forall u, v \in X$, $d(\cdot, \cdot)$ denoting the metric on X; for generalities on metric spaces, see, e. g., [36].

52 $\Phi^j = \Phi \circ \cdots \circ \Phi$ j-times. In fact, let $u_j := \Phi^j(u_0)$ and notice that, for each $j \ge 1$ $d(u_{j+1}, u_j) \le \alpha d(u_j, u_{j-1}) \le \alpha^j d(u_1, u_0) =: \alpha^j\beta$, so that, for each $j, h \ge 1$, $d(u_{j+h}, u_j) \le \sum_{i=0}^{h-1} d(u_{j+i+1}, u_{j+i}) \le \sum_{i=0}^{h-1} \alpha^{j+i}\beta \le \alpha^j\beta/(1-\alpha)$, showing that $\{u_j\}$ is a Cauchy sequence. Uniqueness is obvious.

Bibliography

Primary Literature

1. Ahlfors LV (1978) Complex analysis. An introduction to the theory of analytic functions of one complex variable, 3rd edn. International Series in Pure and Applied Mathematics. McGraw–Hill, New York, xi+331 pp
2. Arnold VI (1963) Proof of a Theorem by Kolmogorov AN on the invariance of quasi-periodic motions under small perturbations of the Hamiltonian. Russ Math Surv 18:13–40
3. Arnold VI (1963) Small divisor problems in classical and Celestial Mechanics. Russ Math Surv 18:85–191
4. Arnold VI (1964) Instability of dynamical systems with many degrees of freedom (Russian). Dokl Akad Nauk SSSR 156:9–12
5. Arnold VI (1974) Mathematical methods of classical mechanics series: graduate texts in mathematics, vol 60. Originally published by Nauka, Moscow, 1974. 2nd edn. 1989. Corr. 4th printing, 1997, pp XVI,509

6. Arnold VI, Kozlov VV, Neishtadt AI (ed) (2006) Mathematical Aspects of Classical and Celestial Mechanics, Dynamical Systems III, 3rd edn. Series: Encyclopedia of Mathematical Sciences, vol 3. Springer, Berlin
7. Avila A, Krikorian R (2006) Reducibility or nonuniform hyperbolicity for quasiperiodic Schrödinger cocycles. Ann Math 164:911–940
8. Broer HW (2004) KAM theory: the legacy of AN Kolmogorov's 1954 paper. Comment on: "The general theory of dynamic systems and classical mechanics". (French in: Proceedings of the International Congress of Mathematicians, Amsterdam, 1954, vol 1, pp 315–333, Erven P, Noordhoff NV, Groningen, 1957). Bull Amer Math Soc (N.S.) 41(4):507–521 (electronic)
9. Broer HW, Huitema GB, Sevryuk MB (1996) Quasi-periodic motions in families of dynamical systems. Order amidst chaos. Lecture Notes in Mathematics, vol 1645. Springer, Berlin
10. Celletti A, Chierchia L (1987) Rigorous estimates for a computer-assisted KAM theory. J Math Phys 28:2078–2086
11. Celletti A, Chierchia L (1988) Construction of analytic KAM surfaces and effective stability bounds. Commun Math Phys 118:119–161
12. Celletti A, Chierchia L (1995) A Constructive Theory of Lagrangian Tori and Computer-assisted Applications. In: Jones CKRT, Kirchgraber U, Walther HO (eds) Dynamics Reported, vol 4 (New Series). Springer, Berlin, pp 60–129
13. Celletti A, Chierchia L (2007) KAM stability and celestial mechanics. Mem Amer Math Soc, 187(878), viii+134 pp
14. Chierchia L (1986) Quasi-periodic Schrödinger operators in one dimension, Absolutely continuous spectra, Bloch waves and integrable Hamiltonian systems. Quaderni del Consiglio Nazionale delle Ricerche, Firenze, Italia, 127 pp http://www.mat.uniroma3.it/users/chierchia/REPRINTS/CNR86.pdf
15. Chierchia L (2003) KAM Lectures. In: Dynamical Systems. Part I: Hamiltonian Systems and Celestial Mechanics. Pubblicazioni della Classe di Scienze, Scuola Normale Superiore, Pisa, pp 1–56. Centro di Ricerca Matematica "Ennio De Giorgi"
16. Chierchia L (2006) KAM theory and Celestial Mechanics. In: Françoise J-P, Naber GL, Tsou ST (eds) Encyclopedia of Mathematical Physics, vol 3. Elsevier, Oxford, pp 189–199. ISBN 978-0-1251-2666-3
17. Chierchia L (2008) A.N. Kolmogorov's 1954 paper on nearly-integrable Hamiltonian systems. A comment on: On conservation of conditionally periodic motions for a small change in Hamilton's function [Dokl Akad Nauk SSSR (N.S.) 98:527–530 (1954)]. Regul Chaotic Dyn 13(2):130–139
18. Chierchia L, Falcolini C (1994) A direct proof of a theorem by Kolmogorov in Hamiltonian systems. Ann Sc Norm Su Pisa IV(XXI):541–593
19. Chierchia L, Gallavotti G (1982) Smooth prime integrals for quasi-integrable Hamiltonian systems. Il Nuovo Cimento 67(B)277–295
20. Chierchia L, Gallavotti G (1994) Drift and diffusion in phase space. Ann Inst Henri Poincaré, Phys Théor 60:1–144; see also Erratum, Ann Inst Henri Poincaré, Phys Théor 68(1):135 (1998)
21. Chierchia L, Perfetti P (1995) Second order Hamiltonian equations on $\mathbb{T}^\infty$ and almost-periodic solutions. J Differ Equ 116(1):172–201
22. Chierchia L, Qian D (2004) Moser's theorem for lower dimensional tori. J Differ Equ 206:55–93
23. Coddington EA, Levinson N (1955) Theory of Ordinary Differential Equations. McGraw–Hill, New York. 14+429 pp
24. Delshams A, Gutierrez P (1996) Effective stability and KAM theory. J Differ Equ 128(2):415–490
25. Eliasson L (1988) Perturbations of stable invariant tori for Hamiltonian systems. Ann Scuola Norm Sup Pisa Cl Sci 15:115–147
26. Eliasson LH (1992) Floquet solutions for the 1-dimensional quasi-periodic Schrödinger equation. Comm Math Phys 146:447–482
27. Eliasson LH (1996) Absolutely convergent series expansion for quasi-periodic motions. Math Phys Electron J 2(Paper 4):33 (electronic); see also: Report 2–88, Dept. of Math., University of Stockholm (1988)
28. Evans LC, Gariepy RF (1992) Measure theory and fine properties of functions. Studies in Advanced Mathematics. CRC Press, Boca Raton, pp viii+268
29. Féjoz J (2004) Démonstration du 'théorème d'Arnold' sur la stabilité du système planétaire (d'après Herman). (French) Ergod Theor Dyn Syst 24(5):1521–1582
30. Gallavotti G (1994) Twistless KAM tori, quasi flat homoclinic intersections and other cancellations in the perturbation series of certain completely integrable Hamiltonian systems. A review. Rev Math Phys 6:343–411
31. Graff S (1974) On the continuation of stable invariant tori for Hamiltonian systems, J Differ Equ 15:1–69
32. Hénon M (1966) Explorationes numérique du problème restreint IV: Masses egales, orbites non periodique. Bull Astronom 3(1, fasc. 2):49–66
33. Katok A, Hasselblatt B (1995) Introduction to the modern theory of dynamical systems. In: Encyclopedia of Mathematics and its Applications, vol 54, Cambridge University Press, Cambridge, xviii+802 pp
34. Katznelson Y (2004) An introduction to harmonic analysis, 3rd edn. Cambridge Mathematical Library. Cambridge University Press, Cambridge, pp xviii+314 pp
35. Kolmogorov AN (1954) On the conservation of conditionally periodic motions under small perturbation of the Hamiltonian. Dokl Akad Nauk SSR 98:527–530
36. Kolmogorov AN, Fomin SV (1999) Elements of the Theory of Functions and Functional Analysis. Courier Dover Publications, 288 pp
37. Kuksin SB (1988) Perturbation theory of conditionally periodic solutions of infinite-dimensional Hamiltonian systems and its applications to the Korteweg–de Vries equation. (Russian) Mat Sb (N.S.) 136,178(3):396–412,431; translation in Math USSR-Sb 64(2):397–413
38. Kuksin SB (2004) Fifteen years of KAM for PDE. Geometry, topology, and mathematical physics. Am Math Soc Transl Ser 212(2):237–258
39. Mather J (1984) Nonexistence of invariant circles. Erg Theor Dyn Syst 4:301–309
40. Mather JN (2003) Arnold diffusion. I. Announcement of results. (Russian) Sovrem Mat Fundam Napravl 2:116–130 (electronic); translation in J Math Sci 124(5):5275–5289
41. Melnikov VK (1965) On certain cases of conservation of almost periodic motions with a small change of the Hamiltonian function. (Russian) Dokl Akad Nauk SSSR 165:1245–1248
42. Moser JK (1961) A new technique for the construction of solutions of nonlinear differential equations. Proc Nat Acad Sci USA 47:1824–1831

43. Moser JK (1962) On invariant curves of area-preserving mappings of an annulus. Nach Akad Wiss Göttingen, Math Phys Kl II 1:1–20
44. Moser J (1966) A rapidly convergent iteration method and non-linear partial differential equations. Ann Scuola Norm Sup Pisa 20:499–535
45. Moser J (1967) Convergent series expansions for quasi-periodic motions. Math Ann 169:136–176
46. Moser J (1970) On the construction of almost periodic solutions for ordinary differential equations. In: Proc. Internat. Conf. on Functional Analysis and Related Topics, Tokyo, 1969. University of Tokyo Press, Tokyo, pp 60–67
47. Moser J (1983) Integrable Hamiltonian systems and spectral theory, Lezioni Fermiane. [Fermi Lectures]. Scuola Normale Superiore, Pisa, iv+85 pp
48. Nash J (1956) The imbedding problem for Riemannian manifolds. Ann Math 63(2):20–63
49. Poincaré H (1892) Les méthodes nouvelles de la Mécanique céleste. Tome 1: Gauthier–Villars, Paris, 1892, 385 pp; tome 2: Gauthier–Villars, Paris 1893, viii+479 pp; tome 3: Gauthier–Villars, Paris, 1899, 414 pp
50. Pöschel J (1982) Integrability of Hamiltonian sytems on Cantor sets. Comm Pure Appl Math 35:653–695
51. Pöschel J (1996) A KAM-theorem for some nonlinear PDEs Ann Scuola Norm Sup Pisa Cl Sci 23:119–148
52. Pyartli AS (1969) Diophantine approximations of submanifolds of a Euclidean space. Funktsional Anal Prilozhen 3(4):59–62 (Russian). English translation in: Funct Anal Appl 3:303–306
53. Rüssmann H (1970) Kleine Nenner. I. Über invariante Kurven differenzierbarer Abbildungen eines Kreisringes. (German) Nachr Akad Wiss Göttingen Math-Phys Kl II 1970:67–105
54. Rüssmann H (1975) On optimal estimates for the solutions of linear partial differential equations of first order with constant coefficients on the torus. In: Moser J (ed) Dynamical Systems, Theory and Applications. Lecture Notes in Physics vol 38. Springer, Berlin, pp 598–624
55. Rüssmann H (1976) Note on sums containing small divisors. Comm Pure Appl Math 29(6):755–758
56. Rüssmann H (1980) On the one-dimensional Schrödinger equation with a quasiperiodic potential. In: Nonlinear dynamics, Internat. Conf., New York, 1979, Ann New York Acad Sci, vol 357. New York Acad Sci, New York, pp 90–107
57. Rüssmann H (1989) Nondegeneracy in the perturbation theory of integrable dynamical systems, Number theory and dynamical systems (York, 1987). London Math Soc. Lecture Note Ser, vol 134. Cambridge University Press, Cambridge, pp 5–18
58. Rüssmann H (2001) Invariant tori in non-degenerate nearly integrable Hamiltonian systems. Regul Chaot Dyn 6(2):119–204
59. Salamon D (2004) The Kolmogorov–Arnold–Moser theorem. Math Phys Electron J 10(Paper 3):37 (electronic)
60. Sevryuk MB (2003) The Classical KAM Theory and the dawn of the twenty-first century. Mosc Math J 3(3):1113–1144
61. Siegel CL (1942) Iteration of analytic functions. Ann Math 43:607–612
62. Spivak M (1979) A Comprehensive Introduction to Differential Geometry, vol 1, 2nd edn. Publish or Perish Inc, Wilmington
63. Wayne CE (1990) Periodic and quasi-periodic solutions for nonlinear wave equations via KAM theory. Commun Math Phys 127:479–528
64. Whitney H (1934) Analytic extensions of differentiable functions defined in closed sets. Trans Amer Math Soc 36(1):63–89
65. Zehnder E (1975) Generalized implicit function theorems with applications to some small divisor problems, vol I. Comm Pure Appl Math 28:91–14
66. Zehnder E (1976) Generalized implicit function theorems with applications to some small divisor problems, vol II. Comm Pure Appl Math 29(1):49–111

Books and Reviews

Arnold VI, Novikov SP (eds) (1985) Dynamical Systems IV, Symplectic Geometry and its Applications. In: Encyclopaedia of Mathematical Sciences, vol 4, Volume package Dynamical Systems. Original Russian edition published by VINITI, Moscow. 2nd expanded and rev. edn., 2001

Berretti A, Celletti A, Chierchia L, Falcolini C (1992) Natural Boundaries for Area-Preserving Twist Maps. J Stat Phys 66:1613–1630

Bost JB (1984/1985) Tores invariants des systèmes dynamiques hamiltoniens. Séminaire Bourbaki, exp. 639

Bourgain J (1997) On Melnikov's persistency problem. Math Res Lett 4(4):445–458

Broer HW, Sevryuk MB KAM theory: Quasi-periodicity in dynamical systems. In: Handbook of Dynamical Systems. Elsevier, to appear

Celletti A, Chierchia L (2006) KAM tori for *N*-body problems: a brief history. Celest Mech Dynam Astronom 95(1–4):117–139

Chierchia L, Falcolini C (1996) Compensations in small divisor problems. Comm Math Phys 175:135–160

Gallavotti G (1983) The Elements of Mechanics. Springer, New York

Giorgilli A, Locatelli U (1997) On classical series expansion for quasi-periodic motions. MPEJ 3(5):1–25

Herman M (1986) Sur les courbes invariantes pour les difféomorphismes de l'anneau. Astérisque 2:248

Herman M (1987) Recent results and some open questions on Siegel's linearization theorems of germs of complex analytic diffeomorphisms of $\mathbf{C}^n$ near a fixed point. In: VIIIth International Congress on Mathematical Physics, Marseille, 1986. World Sci. Publishing, Singapore, pp 138–184

Herman M (1989) Inégalités a priori pour des tores Lagrangiens invariants par des difféomorphismes symplectiques. Publ Math IHÈS 70:47–101

Hofer H, Zehnder E (1995) Symplectic invariants and Hamiltonian dynamics. The Floer memorial volume. Progr. Math., 133. Birkhäuser, Basel, pp 525–544

Kappeler T, Pöschel J (2003) KdV & KAM. Springer, Berlin, XIII,279 pp

Koch H (1999) A renormalization group for Hamiltonians, with applications to KAM tori. Ergod Theor Dyn Syst 19:475–521

Kuksin SB (1993) Nearly integrable infinite dimensional Hamiltonian systems. Lecture Notes in Mathematics, vol 1556. Springer, Berlin

Lanford OE III (1984) Computer assisted proofs in analysis. Phys A 124:465–470

Laskar J (1996) Large scale chaos and marginal stability in the solar system. Celest Mech Dyn Astronom 64:115–162

de la Llave R (2001) A tutorial on KAM theory. Smooth ergodic theory and its applications. In: Proc. Sympos. Pure Math., vol 69, Seattle, WA, 1999. Amer Math Soc, Providence, pp 175–292

de la Llave R, González A, Jorba À, Villanueva J (2005) KAM theory without action-angle variables. Nonlinearity 18(2):855–895

de la Llave R, Rana D (1990) Accurate strategies for small divisor problems. Bull Amer Math Soc (N.S.) 22:(1)85–90
Lazutkin VF (1993) KAM theory and semiclassical approximations to eigenfunctions. Springer, Berlin
MacKay RS (1985) Transition to chaos for area-preserving maps. Lect Notes Phys 247:390–454
Meyer KR, Schmidt D (eds) (1991) Computer Aided Proofs in Analysis. Springer, Berlin
Moser J (1973) Stable and random motions in dynamical systems. With special emphasis on celestial mechanics. Hermann Weyl Lectures, the Institute for Advanced Study, Princeton, N. J Annals of Mathematics Studies, No. 77. Princeton University Press, Princeton; University of Tokyo Press, Tokyo
Pöschel J (2001) A lecture on the classical KAM theorem. Smooth ergodic theory and its applications. In: Proc Sympos Pure Math, vol 69 Seattle,WA, 1999. Amer Math Soc, Providence, pp 707–732
Salamon D, Zehnder E (1989) KAM theory in configuration space. Comment Math Helvetici 64:84–132
Sevryuk MB (1995) KAM-stable Hamiltonians. J Dyn Control Syst 1(3):351–366
Sevryuk MB (1999) The lack-of-parameters problem in the KAM theory revisited, Hamiltonian systems with three or more degrees of freedom (S'Agaró, 1995). NATO Adv Sci Inst Ser C Math Phys Sci, vol 533. Kluwer, Dordrecht, pp 568–572
Siegel CL, Moser JK (1971) Lectures on Celestial Mechanics. Springer, Berlin
Weinstein A (1979) Lectures on symplectic manifolds, 2nd edn., CBMS Regional Conference Series in Mathematics, vol 29. American Mathematical Society, Providence
Xu J, You J, Qiu Q (1997) Invariant tori for nearly integrable Hamiltonian systems with degeneracy. Math. Z 226(3):375–387
Yoccoz J-C (1992) Travaux de Herman sur les tores invariants. Séminaire Bourbaki, 1991/92, Exp. 754. Astérisque 206:311–344
Yoccoz J-C (1995) Théorème de Siegel, pôlynomes quadratiques et nombres de Brjuno. Astérisque 231:3–88

Korteweg–de Vries Equation (KdV), Different Analytical Methods for Solving the

Yu-Jie Ren[1,2,3]
[1] Department of Mathematics, Shantou University, Shantou, People's Republic of China
[2] Department of Mathematics and Physics, Dalian Polytechnic University, Dalian, People's Republic of China
[3] Beijing Institute of Applied Physics and Computational Mathematics, Beijing, People's Republic of China

Article Outline

Glossary
Definition of the Subject
Introduction
The Generalized Hyperbolic Function–Bäcklund Transformation Method and Its Application in the (2 + 1)-Dimensional KdV Equation
The Generalized F-expansion Method and Its Application in Another (2 + 1)-Dimensional KdV Equation
The Generalized Algebra Method and Its Application in (1 + 1)-Dimensional Generalized Variable – Coefficient KdV Equation
A New Exp-N Solitary-Like Method and Its Application in the (1 + 1)-Dimensional Generalized KdV Equation
The Exp-Bäcklund Transformation Method and Its Application in (1 + 1)-Dimensional KdV Equation
Future Directions
Acknowledgment
Bibliography

Glossary

Analytical method for solving an equation The method for obtaining the exact solutions of an equation.

Solitons The special solitary waves which retain their original shapes and speeds after collision and exhibit only a small overall phase shift.

ε ± 1

t the time

(x, y) Cartesian coordinates of a point

∂_x^{-1} an indefinite integrate operator $\int \mathrm{d}x$.

Definition of the Subject

This paper presents the analytical methods for obtaining the exact solutions of the Korteweg–de Vries Equation (KdV equation).

The KdV equation and its exact solutions can describe and explain many physical problems. In addition, it is a typical, relatively simple and classical equation among the many nonlinear equations in physics. Much of the literature of nonlinear equation theory customarily uses solving soliton solutions of the KdV equation as an example to introduce the nonlinear theory, method and character of soliton solutions. During the last five decades, the construction of exact solution for a wide class of nonlinear equations has been an exciting and extremely active area of research. This includes the most famous nonlinear example of the KdV equation.

In the family of the KdV equations, a well known KdV equation is expressed in its simplest form as [1,2,3,4]:

$$u_t(x,t) + \alpha u(x,t) u_x(x,t) + \beta u_{xxx}(x,t) = 0 \,. \qquad (1)$$

where the coefficient of the nonlinear term α and the coefficient of the dispersive term β are independent of x and t.

The KdV equation was derived in 1885 by the two scientists, Korteweg and de Vries [1], to describe long wave propagation on shallow water. Its importance comes from its many applications in weakly nonlinear and weakly dispersive physical systems and its interesting mathematical properties. It can be considered as a paradigm in nonlinear science. The equation is so powerful it has been used to simulate the red spots on Jupiter. Moreover, it describes non-linear waves in rotating fluids [5], the giant ocean waves known as Tsunami, and other aspects of solitary waves in plasma. The giant internal waves in the interior of the ocean arising from temperature differences which may destroy marine vessels can be also described by the powerful KdV equation. It possesses infinitely many generalized symmetries and infinitely many integrals of motion [6], and can be viewed as a completely integrable Hamiltonian system and solved by the inverse scattering transform. It also provides resources for studying integrability of nonlinear differential (and difference) equations, serving as a typical model of integrable equations. More remarkable is that various physically important solutions to the KdV equation can be presented explicitly in a simple way, among which are solitons, rational solutions, positons and negatons.

The KdV equation has been widely investigated and developed continually in recent decades. Several generalizations of the KdV equation have found applications in many areas, including quantum field theory, plasma physics, solid-state physics, liquid-gas bubble mixtures, and anharmonic crystals. For instance, the classical generalized Korteweg–de Vries equation (gKdV) has the following form [7,8,9]

$$u_t(x,t) + \alpha u^p(x,t)u_x(x,t) + \beta u_{xxx}(x,t) = 0\,. \quad (2)$$

where the coefficient of the nonlinear term α and the coefficient of the dispersive term β are independent of x and t.

The more generalized form of the KdV equation is defined by parameters (l, p) as follows [10,11,12,13,14]

$$\begin{aligned} u_t(x,t) = u^{l-2}(x,t)u_x(x,t) + \alpha[&2u^p(x,t)u_{xxx}(x,t) \\ &+ 4pu^{p-1}(x,t)u_x(x,t)u_{xx}(x,t) \\ &+ p(p-1)u^{p-2}(x,t)u_x^3(x,t)]\,. \end{aligned} \quad (3)$$

Propagation of weakly nonlinear long waves in an inhomogeneous waveguide is governed by a variable-coefficient KdV equation of the form [15]

$$u_t(x,t) + 6u(x,t)u_x(x,t) + B(t)u_{xxx}(x,t) = 0\,. \quad (4)$$

where $u(x,t)$ is the wave amplitude, t the propagation coordinate, x the temporal variable and $B(t)$ is the local dispersion coefficient.

The applicability of the variable-coefficient KdV equation (4) arises in many areas of physics as, for example, the description of the propagation of gravity-capillary and interfacial-capillary waves, internal waves and Rossby waves [15]. In order to study the propagation of weakly nonlinear, weakly dispersive waves in the inhomogeneous media, Eq. (4) is rewritten as follows [16]

$$u_t(x,t) + 6A(t)u(x,t)u_x(x,t) + B(t)u_{xxx}(x,t) = 0\,. \quad (5)$$

which includes a variable nonlinearity coefficient $A(t)$.

In studying long waves, tides, solitary waves, and related phenomena, one is led to an equation of the form [17]

$$\begin{aligned} u_t(x,t) + (m+1)(m+2)u^m(x,t)u_x(x,t) \\ + u_{xxx}(x,t) = f(x,t)\,. \end{aligned} \quad (6)$$

where $f(x,t)$ is a given function and $m = 1, 2, \ldots$, with $u(x,t), u_x(x,t), u_{xx}(x,t) \to 0$ as $|x| \to +\infty$. This equation is referred to as a generalized Korteweg–de Vries equation in [15].

The extended Korteweg–de Vries (eKdV) equation [18]

$$\begin{aligned} u_t(x,t) + \alpha u(x,t)u_x(x,t) + \beta u^2(x,t)u_x(x,t) \\ + \delta u_{xxx}(x,t) = 0\,. \end{aligned} \quad (7)$$

incorporates both quadratic and cubic nonlinearities, and serves as a useful model in the propagation of long waves in physical oceanography.

A system of two coupled Korteweg–de Vries equations in [4] is given by the following equation:

$$\begin{aligned} u_{it}(x,t) + \sum_{h,k} A_i^{hk} u_h(x,t)u_{kx}(x,t) \\ + \sum_k A_i^k u_{kxxx}(x,t) = 0\,, \quad i,k,h = 1,2\,. \end{aligned} \quad (8)$$

The $(2+1)$-dimensional KdV equation is presented by the following equation [19]:

$$\begin{cases} u_t(x,y,t) - 3v(x,y,t)u_x(x,y,t) \\ \qquad -3v_x(x,y,t)u(x,y,t) + u_{xxx}(x,y,t) = 0\,, \\ u_x(x,y,t) - v_y(x,y,t) = 0\,. \end{cases} \quad (9)$$

Another $(2+1)$-dimensional KdV equation in [20] is written the following form

$$u_t(x,y,t) - 4u(x,y,t)u_y(x,y,t) - 4u_x(x,y,t)\partial_x^{-1}u_y(x,y,t) - u_{xxy}(x,y,t) = 0\,. \tag{10}$$

where ∂_x^{-1} is an indefinite integrate operator $\int \mathrm{d}x$, possessing some interesting coherent structures.

In addition, the complex coupled KdV equation in [49] is written the following form

$$\begin{cases} u_t(x,t) = \frac{1}{2}[u_{xxx}(x,t) - 6u(x,t)u_x(x,t)] \\ \qquad\qquad\qquad +3(|\,v(x,t)\,|^2)_x\,, \\ v_t(x,t) = -v_{xxx}(x,t) + 3u(x,t)v_x(x,t)\,. \end{cases} \tag{11}$$

Introduction

It is well known that nonlinear complex physical phenomena are related to nonlinear partial differential equations (NLPDEs), which are involved in many fields from physics to biology, chemistry, mechanics, etc. The investigation of exact solutions of NLPDEs as mathematical models of phenomena can help one to better understand a variety of phenomena. In the past several decades, many analytical methods for obtaining exact solutions of NLPDEs have been presented, such as the inverse scattering method, Hirota's bilinear method [21], the Bäcklund transformation [22], the Painleve expansion [23], the tanh function method [24], the sine-cosine method [25], the homogenous balance method [26], the homotopy perturbation method [27], the variational method [28], asymptotic methods [29], non-perturbative methods [30], the exp-function method [31], the Adomian Pade approximation [32], the Jacobi elliptic function expansion method [33], the F-expansion method [34], the Weierstrass semi-rational expansion method [35], the unified rational expansion method [36], the algebraic method [37,50,51], the auxiliary equation method [38], and so on.

In recent years, based on the ideas of unification methods, algorithm realization and mechanization for solving NLPDEs, we have improved and presented some analytical methods for obtaining exact solutions of NLPDEs, such as the generalized hyperbolic function – the Bäcklund transformation method [39], the method for constructing higher order and higher dimension Bäcklund transformation [40], the generalized hyperbolic function – the Riccati method [40], the generalized F-expansion method [41,42], the extended generalized algebraic method [43,44], the Exp-Bäcklund method [45,46], the Exp-N soliton-like method [47,48], the extended sine-cosine method [52], and so on. The present article is motivated by the desire to introduce and make use of our works published in [39,40,41,42,43,44,45,46,47,48] to construct more general exact solutions, which contain not only the results obtained by using the methods [1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,49,50,51] and Ablowitz and Clarkson 1991, Dodd et al. 1984, Drazin and Johnson 1989, Fan 2004, Guo 1996, Liu and Liu 2000 Miura 1976, and Miurs 1978, but also a series of new and more general exact solutions. For illustration, we apply our analytical methods above to the different types of the KdV equations, and successfully obtain many new and more general exact solutions.

The remainder of this article is organized as follows. In Sect. "The Generalized Hyperbolic Function–Bäcklund Transformation Method and Its Application in the $(2+1)$-Dimensional KdV Equation", in order to construct the unification solutions of the hyperbolic function solution and exponential function solution of NLPDEs, we first find a new theory of generalized hyperbolic functions which includes two new definitions of generalized hyperbolic functions and generalized hyperbolic function transformations and their properties. Then we present a new higher order and higher dimension Bäcklund transformation method and a new generalized hyperbolic function – Bäcklund transformation method. The validity of the methods are tested by their application in the $(2+1)$-dimensional KdV equation (9). In Sect. "The Generalized F-expansion Method and Its Application in Another $(2+1)$-Dimensional KdV Equation", we present a new and more general formal transformation and a new generalized F-expansion method and apply them to another $(2+1)$-dimensional KdV equation (10). In Sect. "The Generalized Algebra Method and Its Application in $(1+1)$-Dimensional Generalized Variable – Coefficient KdV Equation", in order to develop the algebraic method [50,51] for constructing the traveling wave solutions of NLPDEs, we first present two new and general transformations, two new theorems and their proofs, by using Maple, and a new mechanization method to find the exact solutions of a first-order nonlinear ordinary differential equations with any degree. The validity of the method is tested by its application in the first-order nonlinear ordinary differential equation with six degrees. Next, we present a new and more general transformation and a new generalized algebra method and apply them to the $(1+1)$-dimensional generalized variable – coefficient KdV equation (5). In Sect. "A New Exp-N Solitary-like Method and Its Application in the $(1+1)$-Dimensional Generalized KdV Equation", in order to develop the Exp-function method [31], we present two new generic trans-

formations, a new Exp-N solitary-like method, and its algorithm. In addition, we apply our methods to construct new exact solutions of the $(1+1)$-dimensional classical generalized KdV (gKdV) equation (2); In Sect. "The the Exp-Bäcklund Transformation Method and Its Application in $(1+1)$-Dimensional KdV Equation", a new Exp-Bäcklund transformation method and its algorithm is presented to find more exact solutions of NLPDEs. We choose the $(1+1)$-dimensional KdV equation (1) to illustrate the effectiveness and convenience of our algorithm. As a result, we obtain more general exact solutions including non-traveling wave solutions and traveling wave solutions. In addition, the long-term behavior of the non-traveling wave solutions and traveling wave solutions are illustrated by some Figures. In Sect. "Future Directions", future directions for Solving NLPDEs are given.

The Generalized Hyperbolic Function–Bäcklund Transformation Method and Its Application in the (2 + 1)-Dimensional KdV Equation

In this section, in order to construct the unification solutions of the hyperbolic function solution and the exponential function solution of NLPDEs, we establish a new theory of generalized hyperbolic functions, which includes two new definitions of generalized hyperbolic functions and generalized hyperbolic function transformations and their properties that we first presented in [39,40]. Then we present a new higher order and higher dimension Bäcklund transformation method and a new generalized hyperbolic function-Bäcklund transformation method that we first presented in [39,40]. With the aid of symbolic computation, we choose the $(2+1)$-dimensional Korteweg–de Vries equations to illustrate the validity and advantages of the methods. As a result, many new and more general exact non-traveling waves are obtained.

The Definition and Properties of Generalized Hyperbolic Functions

In [39], we first defined the following new functions which named generalized hyperbolic functions and studied the properties of these functions for constructing new exact solutions of NLPDEs.

Definition 1 Suppose that ξ is an independent variable, p, q and k are constants.

The **generalized hyperbolic sine function** is

$$\sinh_{pqk}(\xi) = \frac{pe^{k\xi} - qe^{-k\xi}}{2} \tag{12}$$

generalized hyperbolic cosine function is

$$\cosh_{pqk}(\xi) = \frac{pe^{k\xi} + qe^{-k\xi}}{2} \tag{13}$$

generalized hyperbolic tangent function is

$$\tanh_{pqk}(\xi) = \frac{pe^{k\xi} - qe^{-k\xi}}{pe^{k\xi} + qe^{-k\xi}} \tag{14}$$

generalized hyperbolic cotangent function is

$$\coth_{pqk}(\xi) = \frac{pe^{k\xi} + qe^{-k\xi}}{pe^{k\xi} - qe^{-k\xi}} \tag{15}$$

generalized hyperbolic secant function is

$$\operatorname{sech}_{pqk}(\xi) = \frac{2}{pe^{k\xi} + qe^{-k\xi}} \tag{16}$$

generalized hyperbolic cosecant function is

$$\operatorname{csch}_{pqk}(\xi) = \frac{2}{pe^{k\xi} - qe^{-k\xi}} \tag{17}$$

the above six kinds of functions are said **generalized hyperbolic functions**.

Thus we can prove the following theory of generalized hyperbolic functions on the basis of Definition 1.

Theorem 1 *The generalized hyperbolic functions satisfy the following relations:*

$$\cosh^2_{pqk}(\xi) - \sinh^2_{pqk}(\xi) = pq\,, \tag{18}$$

$$1 - \tanh^2_{pqk}(\xi) = pq \cdot \operatorname{sech}^2_{pqk}(\xi)\,, \tag{19}$$

$$1 - \coth^2_{pqk}(\xi) = -pq \cdot \operatorname{csch}^2_{pqk}(\xi)\,, \tag{20}$$

$$\operatorname{sech}_{pqk}(\xi) = \frac{1}{\cosh_{pqk}(\xi)}\,, \tag{21}$$

$$\operatorname{csch}_{pqk}(\xi) = \frac{1}{\sinh_{pqk}(\xi)}\,, \tag{22}$$

$$\tanh_{pqk}(\xi) = \frac{\sinh_{pqk}(\xi)}{\cosh_{pqk}(\xi)}\,, \tag{23}$$

$$\coth_{pqk}(\xi) = \frac{\cosh_{pqk}(\xi)}{\sinh_{pqk}(\xi)}\,, \tag{24}$$

$$\tanh_{pqk}(\xi) = \frac{1}{\coth_{pqk}(\xi)}\,, \tag{25}$$

$$\sinh_{pqk}(-\xi) = -\sinh_{qpk}(\xi)\,, \tag{26}$$

$$\cosh_{pqk}(-\xi) = \cosh_{qpk}(\xi)\,, \tag{27}$$

$$\tanh_{pqk}(-\xi) = -\tanh_{qpk}(\xi)\,, \tag{28}$$

$$\coth_{pqk}(-\xi) = -\coth_{qpk}(\xi)\,, \tag{29}$$

$$\mathrm{sech}_{pqk}(-\xi) = \mathrm{sech}_{qpk}(\xi)\,, \tag{30}$$

$$\mathrm{csch}_{pqk}(-\xi) = -\mathrm{csch}_{qpk}(\xi)\,, \tag{31}$$

$$\sinh_{11k}(\xi-\eta) = \frac{1}{pq}[\sinh_{qpk}(\xi)\cdot\cosh_{pqk}(\eta) - \cosh_{pqk}(\xi)\cdot\sinh_{qpk}(\eta)]\,, \tag{32}$$

$$\sinh_{p^2q^2k}(\xi+\eta) = \sinh_{qpk}(\xi)\cdot\cosh_{pqk}(\eta) + \cosh_{pqk}(\xi)\cdot\sinh_{qpk}(\eta)\,, \tag{33}$$

$$\cosh_{p^2q^2k}(\xi+\eta) = \cosh_{qpk}(\xi)\cdot\cosh_{pqk}(\eta) + \sinh_{pqk}(\xi)\cdot\sinh_{qpk}(\eta)\,, \tag{34}$$

$$\cosh_{11k}(\xi-\eta) = \frac{1}{pq}[\cosh_{qpk}(\xi)\cdot\cosh_{pqk}(\eta) - \sinh_{pqk}(\xi)\cdot\sinh_{qpk}(\eta)]\,, \tag{35}$$

$$\sinh_{q^2p^2k}(2\xi) = 2\sinh_{pqk}(\xi)\cdot\cosh_{pqk}(\xi)\,, \tag{36}$$

$$\cosh_{p^2q^2k}(2\xi) = \sinh^2_{qpk}(\xi) + \cosh^2_{qpk}(\xi)\,. \tag{37}$$

$$\frac{\tanh_{qpk}(\xi)+\tanh_{qpk}(\eta)}{1-\tanh_{qpk}(\xi)\tanh_{qpk}(\eta)} = \frac{\sinh_{p^2q^2k}(\xi+\eta)}{p\cdot q\cdot\cosh_{11k}(\xi-\eta)}\,, \tag{38}$$

$$\frac{\tanh_{qpk}(\xi)-\tanh_{qpk}(\eta)}{1+\tanh_{qpk}(\xi)\tanh_{qpk}(\eta)} = \frac{p\cdot q\cdot\sinh_{11k}(\xi-\eta)}{\cosh_{p^2q^2k}(\xi+\eta)}\,. \tag{39}$$

For simplicity, only a sample of these are proved here.

Proof By (18) and (21), we obtain

$$\begin{aligned}1-\tanh^2_{pqk}(\xi) &= 1-\frac{\sinh^2_{pqk}(\xi)}{\cosh^2_{pqk}(\xi)}\\ &= \frac{\cosh^2_{pqk}(\xi)-\sinh^2_{pqk}(\xi)}{\cosh^2_{pqk}(\xi)}\\ &= \frac{pq}{\cosh^2_{pqk}(\xi)} = pq\cdot\mathrm{sech}^2_{pqk}(\xi)\,.\end{aligned}$$

Similarly, we can prove other conclusions in Theorem 1.□

Theorem 2 *The derivative formulae of generalized hyperbolic functions are as follows:*

$$\frac{d(\sinh_{pqk}(\xi))}{d\xi} = k\cdot\cosh_{pqk}(\xi)\,, \tag{40}$$

$$\frac{d(\cosh_{pqk}(\xi))}{d\xi} = k\cdot\sinh_{pqk}(\xi)\,, \tag{41}$$

$$\frac{d(\tanh_{pqk}(\xi))}{d\xi} = kpq\cdot\mathrm{sech}^2_{pqk}(\xi)\,, \tag{42}$$

$$\frac{\mathrm{d}(\coth_{pqk}(\xi))}{d\xi} = -kpq\cdot\mathrm{csch}^2_{pqk}(\xi)\,, \tag{43}$$

$$\frac{d(\mathrm{sech}_{pqk}(\xi))}{d\xi} = -k\cdot\mathrm{sech}_{pqk}(\xi)\cdot\tanh_{pqk}(\xi)\,, \tag{44}$$

$$\frac{d(\mathrm{csch}_{pqk}(\xi))}{d\xi} = -k\cdot\mathrm{csch}_{pqk}(\xi)\cdot\coth_{pqk}(\xi)\,. \tag{45}$$

Proof According to (40) and (41), we can get

$$\begin{aligned}\frac{\mathrm{d}(\tanh_{pqk}(\xi))}{\mathrm{d}\xi} &= \left(\frac{\sinh_{pqk}(\xi)}{\cosh_{pqk}(\xi)}\right)'\\ &= \frac{(\sinh_{pqk}(\xi))'\cosh_{pqk}(\xi)-\sinh_{pqk}(\xi)(\cosh_{pqk}(\xi))'}{\cosh_{\mathrm{pqk}}{}^2(\xi)}\\ &= \frac{k\cdot\cosh_{pqk}(\xi)\cdot\cosh_{pqk}(\xi)-\sinh_{pqk}(\xi)\cdot(k\cdot\sinh_{pqk}(\xi))}{\cosh^2_{pqk}(\xi)}\\ &= kpq\cdot\mathrm{sech}^2_{pqk}(\xi)\,.\end{aligned}$$

□

Similarly, we can prove other differential coefficient formulae in Theorem 2.

Remark 1 We see that when $p=1, q=1, k=1$ in (12)–(17), new generalized hyperbolic functions $\sinh_{pqk}(\xi)$, $\cosh_{pqk}(\xi)$, $\tanh_{pqk}(\xi)$, $\coth_{pqk}(\xi)$, $\mathrm{sech}_{pqk}(\xi)$ and $\mathrm{csch}_{pqk}(\xi)$ degenerate as hyperbolic functions $\sinh(\xi)$, $\cosh(\xi)$, $\tanh(\xi)$, $\coth(\xi)$, $\mathrm{sech}(\xi)$ and $\mathrm{csch}(\xi)$, respectively. In addition, when $p=0$ or $q=0$ in (12)–(17), $\sinh_{pqk}(\xi)$, $\cosh_{pqk}(\xi)$, $\mathrm{sech}_{pqk}(\xi)$, $\mathrm{csch}_{pqk}(\xi)$, $\tanh_{pqk}(\xi)$ and $\coth_{pqk}(\xi)$ degenerate as exponential functions $\frac{1}{2}p\mathrm{e}^{k\xi}$, $\pm\frac{1}{2}q\mathrm{e}^{-k\xi}$, $2p\mathrm{e}^{-k\xi}$, $\pm 2q\mathrm{e}^{k\xi}$ and ± 1, respectively.

A New Higher Order and Higher Dimension Bäcklund Transformation Method to Construct an Auto–Bäcklund Transformation of the (2 + 1)-Dimensional KdV Equation

In this section, we obtain an auto-Bäcklund transformation of the following (2 + 1)-dimensional KdV equation by making use of the method for constructing higher order and higher dimension Bäcklund transformations which we presented in [39,40] via symbolic computation.

The $(2+1)$-dimensional KdV equations [19] take the form

$$\begin{cases} u_t(x,y,t) - 3v(x,y,t)u_x(x,y,t) \\ \qquad -3v_x(x,y,t)u(x,y,t) + u_{xxx}(x,y,t) = 0\,, \\ u_x(x,y,t) - v_y(x,y,t) = 0\,. \end{cases} \tag{46}$$

The auto-Bäcklund transformations of Eqs. (46) have the following forms:

$$\begin{cases} u(x,y,t) = \sum_{j_1=0}^{m_1} u_{j_1}(x,y,t) f^{j_1-m_1}(x,y,t)\,, \\ v(x,y,t) = \sum_{j_2=0}^{m_2} v_{j_2}(x,y,t) f^{j_2-m_2}(x,y,t)\,, \end{cases} \tag{47}$$

where $f(x,y,t), u_{j_1}(x,y,t), j_1 = 0,1,2,\ldots,m_1-1$ and $v_{j_2}(x,y,t), j_2 = 0,1,2,\ldots,m_2-1$ are all differential functions to be determined later, and $u_{m_1}(x,y,t), v_{m_2}(x,y,t)$ are the trivial seed solutions of Eqs. (46).

By balancing the highest order linear term and nonlinear terms in Eqs. (46), we get $m_1 = 2$ and $m_2 = 2$, and (47) has the following formal

$$\begin{cases} u(x,y,t) = u_0(x,y,t) f^{-2}(x,y,t) \\ \qquad\qquad + u_1(x,y,t) f^{-1}(x,y,t) + u_2(x,y,t)\,, \\ v(x,y,t) = v_0(x,y,t) f^{-2}(x,y,t) \\ \qquad\qquad + v_1(x,y,t) f^{-1}(x,y,t) + v_2(x,y,t)\,, \end{cases} \tag{48}$$

We take the trivial seed solutions as

$$u_2 = u_2(x,y,t), v_2 = v_2(x,y,t)\,. \tag{49}$$

With the aid of Maple symbolic computation software, substituting (48) and (49) into (46), and collecting all terms with $f^{-i}(x,y,t)$, $i = 0,1,2,\ldots$, we obtain

$$\begin{aligned} &\left(-24u_0 f_x^3 + 12u_0 v_0 f_x\right) f^{-5} + \left(12u_1 f_x^3 + 36 f_x^2 f_y f_{x^2}\right. \\ &\quad \left. + 18v_1 f_x^2 f_y + 24 f_x^3 f_{xy}\right) f^{-4} + \cdots = 0\,, \end{aligned} \tag{50}$$

$$\begin{aligned} &\left(-2u_0 f_x + 2v_0 f_y\right) f^{-3} + \left(u_{0x} - u_1 f_x + v_1 f_y - v_{0y}\right) \\ &\quad f^{-2} + \left(u_{1x} - v_{1y}\right) f^{-1} + u_{2x} - v_{2y} = 0\,, \end{aligned} \tag{51}$$

Setting the coefficient of f^{-5} in (50) and f^{-3} in (51) to be zero, we obtain a differential equation

$$-24u_0 f_x^3 + 12u_0 v_0 f_x = 0\,, \tag{52}$$

$$-2u_0 f_x + 2v_0 f_y = 0\,, \tag{53}$$

which has the solution

$$u_0 = 2f_x f_y, \quad v_0 = 2f_x^2\,. \tag{54}$$

Setting the coefficients of f^{-4} in (50) and f^{-2} in (51) to be zero, we obtain a differential equation

$$12u_1 f_x^3 + 36 f_x^2 f_y f_{x^2} + 18v_1 f_x^2 f_y + 24 f_x^3 f_{xy} = 0\,, \tag{55}$$

$$u_{0x} - u_1 f_x + v_1 f_y - v_{0y} = 0\,, \tag{56}$$

from which we can get the following expression:

$$u_1 = -2f_{yx}, \quad v_1 = -2f_{xx}\,. \tag{57}$$

By (48), (49), (54) and (57), we obtain an auto-Bäcklund transformation of Eqs. (46)

$$\begin{aligned} u &= -2\partial_{yx} \ln(f(x,y,t)) + u_2(x,y,t)\,, \\ v &= -2\partial_{xx} \ln(f(x,y,t)) + v_2(x,y,t)\,, \end{aligned} \tag{58}$$

where $u_2(x,y,t), v_2(x,y,t)$ are the trivial seed solutions of Eqs. (46), $f = f(x,y,t)$ satisfies the Eqs. (52), (53), (55), (56) and the following equations

$$\begin{aligned} &-u_0 v_{1x} - u_1 v_{0x} + 2u_1 v_1 f_x + 2u_2 v_0 f_x - 2/3 u_0 f_t \\ &- 2/3 u_0 f_{x^3} + 2u_1 f_x f_{x^2} - 2u_{0x} f_{x^2} \\ &+ 2u_{1x} f_x^2 - 2u_{0x^2} f_x - v_0 u_{1x} - v_1 u_{0x} \\ &+ 2v_2 u_0 f_x = 0\,, \\ &u_{0x^3} - u_1 f_{x^3} - 3u_0 v_{2x} - 3u_1 v_{1x} - 3u_2 v_{0x} \\ &+ 3u_2 v_1 f_x - 3u_{1x} f_{x^2} - u_1 f_t - 3v_0 u_{2x} - 3v_1 u_{1x} \\ &- 3v_2 u_{0x} + 3v_2 u_1 f_x + u_{0t} - 3u_{1x^2} f_x = 0\,. \end{aligned} \tag{59}$$

The Generalized Hyperbolic Function–Bäcklund Transformation Method and Its Application in the (2 + 1)-Dimensional KdV Equation

We first give the definition of a generalized hyperbolic function transformation [40] as follows:

Definition 2 If a transformation includes the generalized hyperbolic functions, then the transformation is defined to be a generalized hyperbolic function transformation.

In this section, we will use the generalized hyperbolic function–Bäcklund transformation method to seek new exact solutions of Eqs. (46). We take $f(x,y,t)$ in (48), (49), (54), (57) and (59) as being of a new form which is the following generalized hyperbolic function transformation

$$\begin{aligned} f(x,y,t) &= H(t) \\ &\quad + \sum_{i=1}^{n} K_i(t) \cdot F_i(\xi_i(x,y,t)) \cdot G_i(\eta_i(x,y,t))\,, \end{aligned} \tag{60}$$

where $F_i(\xi_i)$ and $G_i(\eta_i)$ may take any two generalized hyperbolic functions among $\sinh_{pqk}(\xi), \cosh_{pqk}(\xi)$,

$\tanh_{pqk}(\xi)$, $\coth_{pqk}(\xi)$, $\mathrm{sech}_{pqk}(\xi)$and $\mathrm{csch}_{pqk}(\xi)$. And then (60) has twenty–one types of general forms. For example

$$f(x,y,t) = H(t) + \sum_{i=1}^{n} K_i(t)\cdot \mathrm{sech}_{p_{1i}q_{1i}k_{1i}}(\xi_i(x,y,t)) \cdot \tanh_{p_{2i}q_{2i}k_{2i}}(\eta_i(x,y,t)), \quad (61)$$

$$f(x,y,t) = H(t) + \sum_{i=1}^{n} K_i(t)\cdot \mathrm{csch}^2_{p_iq_ik_i}(\xi_i(x,y,t)), \quad (62)$$

We see that when $p = 1, q = 1, k = 1$ in (12)–(12), new generalized hyperbolic function $\sinh_{pqk}(\xi)$, $\cosh_{pqk}(\xi)$, $\tanh_{pqk}(\xi)$, $\coth_{pqk}(\xi)$, $\mathrm{sech}_{pqk}(\xi)$ and $\mathrm{csch}_{pqk}(\xi)$ degenerate as hyperbolic functions $\sinh(\xi)$, $cosh(\xi)$, $\tanh(\xi)$, $\coth(\xi)$, $\mathrm{sech}(\xi)$ and $\mathrm{csch}(\xi)$ respectively. Therefore (3.49) includes twenty-one types of hyperbolic function forms. For example

$$f(x,y,t) = H(t) + \sum_{i=1}^{n} K_i(t)\cdot \sinh(\xi_i(x,y,t)) \cdot \mathrm{sech}(\eta_i(x,y,t)), \quad (63)$$

and thirty-six types of the omnibus forms of generalized hyperbolic functions and hyperbolic functions. For example

$$f(x,y,t) = H(t) + \sum_{i=1}^{n} K_i(t)\cdot \mathrm{sech}(\xi_i(x,y,t)) \cdot \mathrm{csch}_{p_{2i}q_{2i}k_{2i}}(\eta_i(x,y,t)), \quad (64)$$

In addition, when $p = 0$ or $q = 0$ in (12)–(17), $\sinh_{pqk}(\xi)$, $\cosh_{pqk}(\xi)$, $\mathrm{sech}_{pqk}(\xi)$, $\mathrm{csch}_{pqk}(\xi)$, $\tanh_{pqk}(\xi)$ and $\coth_{pqk}(\xi)$ degenerate as exponential functions $\frac{1}{2}pe^{k\xi}$, $\pm\frac{1}{2}qe^{-k\xi}$, $2pe^{-k\xi}$, $\pm 2qe^{k\xi}$ and ± 1 respectively.

For example, we take $F_i(\xi_i(x,y,t)) = \tanh_{p_{1i}q_{1i}k_{1i}}(\xi_i)$, $G_i(\eta_i(x,y,t)) = \sinh_{p_{2i}q_{2i}k_{2i}}(\eta_i)$. Then (60) has the new form

$$f(x,y,t) = H(t) + \sum_{i=1}^{n} K_i(t)\tanh_{p_1q_1k_1}(\xi_i) \mathrm{sech}_{p_2q_2k_2}(\eta_i), \quad (65)$$

where $\xi_i = \alpha_i(t)x+\beta_i(t)y+r_i(t)$, $\eta_i = \mu_i(t)x+\nu_i(t)y+l_i(t)$, $(i = 1,2,\ldots,n)$, and $H(t)$, $K_i(t)$, $\alpha_i(t)$, $\beta_i(t)$, $r_i(t)$, $\mu_i(t)$, $\nu_i(t)$ and $l_i(t)$ are the functions of t, p_{1i}, q_{1i}, k_{1i}, p_{2i}, q_{2i}, k_{2i} are the constants, and $\sum_{i=1}^{n} K_i^2(t) \neq 0 (i = 0,1,2,\ldots,n)$.

We take the initial solution of Eqs. (46) as $u_2(x,y,t) = u_2(t)$, and $v_2(x,y,t) = v_2(t)$ in (58) for convenience. So based on the Bäcklund transformation (58), the generalized hyperbolic function transformation (65), and

$$\frac{d^2(\tanh_{pqk}(\xi))}{d\xi^2} = -2k^2pq\cdot\mathrm{sech}^2_{pqk}(\xi)\cdot\tanh_{pqk}(\xi), \quad (66)$$

$$\frac{d^2(\mathrm{sech}_{pqk}(\xi))}{d\xi^2} = k^2\cdot\mathrm{sech}_{pqk}(\xi)\cdot(2\tanh^2_{pqk}(\xi)-1), \quad (67)$$

we can get the following new multi-soliton-like solution which we call a generalized hyperbolic function solution of Eq. (46):

$$\begin{aligned} u = {} & \frac{2\sum_{i=1}^{n} p_1q_1k_1k_2K_i(t)\nu_i(t)\alpha_i(t)\tanh_{p_2q_2k_2}(\eta_i)\,\mathrm{sech}^2_{p_1q_1k_1}(\xi_i)\,\mathrm{sech}_{p_2q_2k_2}(\eta_i)}{H(t)+\sum_{i=1}^{n} K_i(t)\tanh_{p_1q_1k_1}(\xi_i)\,\mathrm{sech}_{p_2q_2k_2}(\eta_i)} \\ & + \frac{2\sum_{i=1}^{n} k_2^2K_i(t)\mu_i(t)\nu_i(t)\tanh_{p_1q_1k_1}(\xi_i)\,\mathrm{sech}_{p_2q_2k_2}(\eta_i)(1-2\tanh^2_{p_2q_2k_2}(\eta_i))}{H(t)+\sum_{i=1}^{n} K_i(t)\tanh_{p_1q_1k_1}(\xi_i)\,\mathrm{sech}_{p_2q_2k_2}(\eta_i)} \\ & + \frac{2\sum_{i=1}^{n} p_1q_1k_1k_2K_i(t)\beta_i(t)\mu_i(t)\,\mathrm{sech}^2_{p_1q_1k_1}(\xi_i)\tanh_{p_2q_2k_2}(\eta_i)\,\mathrm{sech}_{p_2q_2k_2}(\eta_i)}{H(t)+\sum_{i=1}^{n} K_i(t)\tanh_{p_1q_1k_1}(\xi_i)\,\mathrm{sech}_{p_2q_2k_2}(\eta_i)} \\ & \frac{2\sum_{i=1}^{n} K_i(t)(p_1q_1k_1\beta_i(t)\,\mathrm{sech}^2_{p_1q_1k_1}(\xi_i)-k_2\nu_i(t)\tanh_{p_1q_1k_1}(\xi_i)\tanh_{p_2q_2k_2}(\eta_i))\,\mathrm{sech}_{p_2q_2k_2}(\eta_i)}{H(t)+\sum_{i=1}^{n} K_i(t)\tanh_{p_1q_1k_1}(\xi_i)\,\mathrm{sech}_{p_2q_2k_2}(\eta_i)} \\ & \frac{\sum_{i=1}^{n} K_i(t)(p_1q_1k_1\alpha_i(t)\,\mathrm{sech}^2_{p_1q_1k_1}(\xi_i)-k_2\mu_i(t)\tanh_{p_1q_1k_1}(\xi_i)\tanh_{p_2q_2k_2}(\eta_i))\,\mathrm{sech}_{p_2q_2k_2}(\eta_i)}{H(t)+\sum_{i=1}^{n} K_i(t)\tanh_{p_1q_1k_1}(\xi_i)\,\mathrm{sech}_{p_2q_2k_2}(\eta_i)} \\ & + u_2, \end{aligned} \quad (68)$$

$$\begin{aligned} v = {} & \frac{\sum_{i=1}^{n} 4p_1q_1k_1\alpha_i(t)K_i(t)\,\mathrm{sech}^2_{p_1q_1k_1}(\xi_i)\,\mathrm{sech}_{p_2q_2k_2}(\eta_i)(k_1\alpha_i(t)\tanh_{p_1q_1k_1}(\xi_i)+k_2\mu_i(t)\tanh_{p_2q_2k_2}(\eta_i))}{H(t)+\sum_{i=1}^{n} K_i(t)\tanh_{p_1q_1k_1}(\xi_i)\,\mathrm{sech}_{p_2q_2k_2}(\eta_i)} \\ & + \frac{2\sum_{i=1}^{n} K_i(t)k_2^2(\mu_i(t))^2\tanh_{p_1q_1k_1}(\xi_i)\,\mathrm{sech}_{p_2q_2k_2}(\eta_i)(1-2\tanh^2_{p_2q_2k_2}(\eta_i))}{H(t)+\sum_{i=1}^{n} K_i(t)\tanh_{p_1q_1k_1}(\xi_i)\,\mathrm{sech}_{p_2q_2k_2}(\eta_i)} \\ & \frac{\left(\sum_{i=1}^{n} K_i(t)\,\mathrm{sech}_{p_2q_2k_2}(\eta_i)(p_1q_1k_1\alpha_i(t)\,\mathrm{sech}^2_{p_1q_1k_1}(\xi_i)-k_2\mu_i(t)\tanh_{p_1q_1k_1}(\xi_i)\tanh_{p_2q_2k_2}(\eta_i))\right)^2}{\left(H(t)+\sum_{i=1}^{n} K_i(t)\tanh_{p_1q_1k_1}(\xi_i)\,\mathrm{sech}_{p_2q_2k_2}(\eta_i)\right)^2} \\ & + v_2, \end{aligned} \quad (69)$$

where $\xi_i = \alpha_i(t)x+\beta_i(t)y+r_i(t)$, $\eta_i = \mu_i(t)x+\nu_i(t)y+l_i(t)$.

For example, when we take $n = 2, H, K_i, p_i, q_i, k_i, \alpha_i, \beta_i$ and r_i, $i = 1, 2$ are all constants, then (65) becomes

$$\begin{aligned} f(x,y,t) &= H + K_1 \tanh_{p_1q_1k_1}(\alpha_1 x + \beta_1 y + r_1) \\ &\cdot \operatorname{sech}_{p_1q_1k_1}(\alpha_1 x + \beta_1 y + r_1) + K_2 \tanh_{p_2q_2k_2} \\ &\cdot (\alpha_2 x + \beta_2 y + r_2) \operatorname{sech}_{p_2q_2k_2}(\alpha_2 x + \beta_2 y + r_2)\,, \end{aligned} \tag{70}$$

With the help of symbolic computation, substituting (70) and $u_2(x,y,t) = u_2(t), v_2(x,y,t) = v_2(t)$ into (52), (53), (55), (56) and (59), we can get $H, K_i, p_i, q_i, k_i, \alpha_i, \beta_i$ and r_i, $i = 1, 2$ respectively as follows:

Case 1

$$\begin{aligned} &u_2 = C_1\,, \\ &v_2 = \frac{C_1(\beta_2{}^2\alpha_1{}^2 + \beta_1{}^2\alpha_2{}^2 + \beta_1\alpha_2\beta_2\alpha_1)}{3\beta_1\beta_2(\beta_1\alpha_2 + \beta_2\alpha_1)}\,, \\ &p_1 = p_1\,, \quad p_2 = p_2\,, \quad q_1 = 0\,, \quad q_2 = 0\,, \\ &k_2 = \varepsilon\frac{\sqrt{\beta_1\beta_2(\beta_1\alpha_2 + \beta_2\alpha_1)C_1}(2\beta_1\alpha_2 + \beta_2\alpha_1)}{\beta_1\beta_2(\beta_1\alpha_2 + \beta_2\alpha_1)\alpha_2} \quad \text{or} \\ &k_2 = \varepsilon\frac{\sqrt{\beta_1\beta_2(\beta_1\alpha_2 + \beta_2\alpha_1)C_1}(2\beta_1\alpha_2 + \beta_2\alpha_1)}{\beta_1\beta_2(\beta_1\alpha_2 + \beta_2\alpha_1)\alpha_2}\,, \\ &k_1 = \varepsilon\frac{\sqrt{\beta_1\beta_2(\beta_1\alpha_2 + \beta_2\alpha_1)C_1}(\beta_1\alpha_2 + 2\beta_2\alpha_1)}{(\beta_1\beta_2{}^2\alpha_1 + \beta_1{}^2\beta_2\alpha_2)\alpha_1}\,, \\ &H = H\,, \quad \alpha_1 = \alpha_1\,, \quad \alpha_2 = \alpha_2\,, \\ &\beta_1 = \beta_1\,, \quad \beta_2 = \beta_2\,, \quad \gamma_1 = \gamma_1\,, \quad \gamma_2 = \gamma_2\,, \\ &K_1 = K_1\,, \quad K_2 = K_2\,. \end{aligned} \tag{71}$$

Case 2

$$\begin{aligned} &p_2 = p_2\,, \quad u_2 = C_1\,, \\ &v_2 = \frac{C_1(\beta_2{}^2\alpha_1{}^2 + \beta_1{}^2\alpha_2{}^2 + \beta_1\alpha_2\beta_2\alpha_1)}{3\beta_1\beta_2(\beta_1\alpha_2 + \beta_2\alpha_1)}\,, \\ &p_1 = p_1\,, \quad q_1 = 0\,, \\ &k_1 = \varepsilon\frac{\sqrt{\beta_1\beta_2(\beta_1\alpha_2 + \beta_2\alpha_1)C_1}(\beta_1\alpha_2 + 2\beta_2\alpha_1)}{(\beta_1\beta_2{}^2\alpha_1 + \beta_1{}^2\beta_2\alpha_2)\alpha_1}\,, \\ &q_2 = q_2\,, \quad k_2 = 0\,, \quad H = H\,, \quad \alpha_1 = \alpha_1\,, \quad \alpha_2 = \alpha_2\,, \\ &\beta_1 = \beta_1\,, \quad \beta_2 = \beta_2\,, \quad \gamma_1 = \gamma_1\,, \quad \gamma_2 = \gamma_2\,, \\ &K_1 = K_1\,, \quad K_2 = K_2\,. \end{aligned} \tag{72}$$

Case 3

$$\begin{aligned} &p_2 = 0\,, \quad u_2 = C_1\,, \\ &v_2 = \frac{C_1(\beta_2{}^2\alpha_1{}^2 + \beta_1{}^2\alpha_2{}^2 + \beta_1\alpha_2\beta_2\alpha_1)}{3\beta_1\beta_2(\beta_1\alpha_2 + \beta_2\alpha_1)}\,, \\ &p_1 = p_1\,, \quad q_1 = 0\,, \quad K_1 = K_1\,, \quad K_2 = K_2\,, \\ &k_1 = \varepsilon\frac{\sqrt{\beta_1\beta_2(\beta_1\alpha_2 + \beta_2\alpha_1)C_1}(\beta_1\alpha_2 + 2\beta_2\alpha_1)}{(\beta_1\beta_2{}^2\alpha_1 + \beta_1{}^2\beta_2\alpha_2)\alpha_1}\,, \\ &q_2 = q_2\,, \quad k_2 = 0\,, \quad H = H\,, \quad \alpha_1 = \alpha_1\,, \quad \alpha_2 = \alpha_2\,, \\ &\beta_1 = \beta_1\,, \quad \beta_2 = \beta_2\,, \quad \gamma_1 = \gamma_1\,, \quad \gamma_2 = \gamma_2\,. \end{aligned} \tag{73}$$

Case 4

$$\begin{aligned} &k_2 = \varepsilon\frac{\sqrt{\beta_1\beta_2(\beta_1\alpha_2 + \beta_2\alpha_1)C_1}(2\beta_1\alpha_2 + \beta_2\alpha_1)}{\beta_1\beta_2(\beta_1\alpha_2 + \beta_2\alpha_1)\alpha_2} \quad \text{or} \\ &k_2 = -\varepsilon\frac{\sqrt{\beta_1\beta_2(\beta_1\alpha_2 + \beta_2\alpha_1)C_1}(2\beta_1\alpha_2 + \beta_2\alpha_1)}{\beta_1\beta_2(\beta_1\alpha_2 + \beta_2\alpha_1)\alpha_2}\,, \\ &p_2 = 0\,, \quad u_2 = C_1\,, \\ &v_2 = \frac{C_1(\beta_2{}^2\alpha_1{}^2 + \beta_1{}^2\alpha_2{}^2 + \beta_1\alpha_2\beta_2\alpha_1)}{3\beta_1\beta_2(\beta_1\alpha_2 + \beta_2\alpha_1)}\,, \\ &p_1 = p_1\,, \quad q_1 = 0\,, \\ &k_1 = \varepsilon\frac{\sqrt{\beta_1\beta_2(\beta_1\alpha_2 + \beta_2\alpha_1)C_1}(\beta_1\alpha_2 + 2\beta_2\alpha_1)}{(\beta_1\beta_2{}^2\alpha_1 + \beta_1{}^2\beta_2\alpha_2)\alpha_1}\,, \\ &q_2 = q_2\,, \quad H = H\,, \quad \alpha_1 = \alpha_1\,, \quad \alpha_2 = \alpha_2\,, \\ &\beta_1 = \beta_1\,, \quad \beta_2 = \beta_2\,, \quad \gamma_1 = \gamma_1\,, \quad \gamma_2 = \gamma_2\,, \\ &K_1 = K_1\,, \quad K_2 = K_2\,. \end{aligned} \tag{74}$$

Case 5

$$\begin{aligned} &k_2 = \varepsilon\frac{\sqrt{\beta_1\beta_2(\beta_1\alpha_2 + \beta_2\alpha_1)C_1}(2\beta_1\alpha_2 + \beta_2\alpha_1)}{\beta_1\beta_2(\beta_1\alpha_2 + \beta_2\alpha_1)\alpha_2} \quad \text{or} \\ &k_2 = -\varepsilon\frac{\sqrt{\beta_1\beta_2(\beta_1\alpha_2 + \beta_2\alpha_1)C_1}(2\beta_1\alpha_2 + \beta_2\alpha_1)}{\beta_1\beta_2(\beta_1\alpha_2 + \beta_2\alpha_1)\alpha_2}\,, \\ &v_2 = \frac{C_1(\beta_2{}^2\alpha_1{}^2 + \beta_1{}^2\alpha_2{}^2 + \beta_1\alpha_2\beta_2\alpha_1)}{3\beta_1\beta_2(\beta_1\alpha_2 + \beta_2\alpha_1)}\,, \\ &k_1 = -\varepsilon\frac{\sqrt{\beta_1\beta_2(\beta_1\alpha_2 + \beta_2\alpha_1)C_1}(\beta_1\alpha_2 + 2\beta_2\alpha_1)}{(\beta_1\beta_2{}^2\alpha_1 + \beta_1{}^2\beta_2\alpha_2)\alpha_1}\,, \\ &H = H\,, \quad \alpha_1 = \alpha_1\,, \quad \alpha_2 = \alpha_2\,, \\ &\beta_1 = \beta_1\,, \quad \beta_2 = \beta_2\,, \quad \gamma_1 = \gamma_1\,, \quad \gamma_2 = \gamma_2\,, \\ &q_2 = 0\,, \quad p_1 = 0\,, \quad q_1 = q_1\,, \\ &u_2 = C_1\,, \quad p_2 = p_2\,, \\ &K_1 = K_1\,, \quad K_2 = K_2\,. \end{aligned} \tag{75}$$

Case 6

$$\begin{aligned}&p_1 = 0\,, \quad k_2 = k_2\,, \quad k_1 = k_1\,, \quad v_2 = v_2\,,\\ &u_2 = C_1\,, \quad p_2 = p_2\,, \quad q_1 = 0\,, \quad q_2 = q_2\,,\\ &H = H\,, \quad \alpha_1 = \alpha_1\,, \quad \alpha_2 = \alpha_2\,,\\ &\beta_1 = \beta_1\,, \quad \beta_2 = \beta_2\,, \quad \gamma_1 = \gamma_1\,, \quad \gamma_2 = \gamma_2\,,\\ &K_1 = K_1\,, \quad K_2 = K_2\,.\end{aligned} \tag{76}$$

Case 7

$$\begin{aligned}&p_2 = 0\,, \quad q_2 = 0\,, \quad p_1 = 0\,, \quad q_1 = q_1\,, \quad v_2 = C_1\,,\\ &k_2 = \varepsilon\frac{\sqrt{3C_1}}{\alpha_2} \text{ or } k_2 = -\varepsilon\frac{\sqrt{3C_1}}{\alpha_2}\,, \quad k_1 = \varepsilon\frac{\sqrt{3C_1}}{\alpha_1}\,,\\ &u_2 = 0\,, \quad H = H\,, \quad \alpha_1 = \alpha_1\,, \quad \alpha_2 = \alpha_2\,,\\ &\beta_1 = \beta_1\,, \quad \beta_2 = \beta_2\,, \quad \gamma_1 = \gamma_1\,, \quad \gamma_2 = \gamma_2\,,\\ &K_1 = K_1\,, \quad K_2 = K_2\,.\end{aligned} \tag{77}$$

Case 8

$$\begin{aligned}&v_2 = C_2\,, \quad p_2 = 0\,, \quad k_1 = \varepsilon\frac{\sqrt{3\beta_1(C_2\beta_1 + C_1\alpha_1)}}{\beta_1\alpha_1} \text{ or}\\ &k_1 = -\varepsilon\frac{\sqrt{3\beta_1(C_2\beta_1 + C_1\alpha_1)}}{\beta_1\alpha_1}\,, \quad q_1 = q_1\,, \quad u_2 = C_1\,,\\ &k_2 = \varepsilon\frac{\sqrt{3\beta_2(C_2\beta_2 + C_1\alpha_2)}}{\beta_2\alpha_2}\,, \quad q_2 = 0\,, \quad p_1 = 0\,,\\ &H = H\,, \quad \alpha_1 = \alpha_1\,, \quad \alpha_2 = \alpha_2\,,\\ &\beta_1 = \beta_1\,, \quad \beta_2 = \beta_2\,, \quad \gamma_1 = \gamma_1\,, \quad \gamma_2 = \gamma_2\,,\\ &K_1 = K_1\,, \quad K_2 = K_2\,.\end{aligned} \tag{78}$$

Case 9

$$\begin{aligned}&v_2 = C_2\,, \quad p_2 = 0\,, \quad p_1 = 0\,, \quad q_1 = q_1\,,\\ &k_1 = 0\,, \quad k_2 = \varepsilon\frac{\sqrt{3\beta_2(C_2\beta_2 + C_1\alpha_2)}}{\beta_2\alpha_2}\,,\\ &u_2 = C_1\,, \quad q_2 = q_2\,,\\ &H = H\,, \quad \alpha_1 = \alpha_1\,, \quad \alpha_2 = \alpha_2\,,\\ &\beta_1 = \beta_1\,, \quad \beta_2 = \beta_2\,, \quad \gamma_1 = \gamma_1\,, \quad \gamma_2 = \gamma_2\,,\\ &K_1 = K_1\,, \quad K_2 = K_2\,.\end{aligned} \tag{79}$$

Case 10

$$\begin{aligned}&v_2 = C_2\,, \quad p_2 = 0\,, \quad k_1 = \varepsilon\frac{\sqrt{3\beta_1(C_2\beta_1 + C_1\alpha_1)}}{\beta_1\alpha_1}\,,\\ &p_1 = 0\,, \quad q_1 = q_1\,, \quad k_2 = 0\,, \quad u_2 = C_1\,,\\ &q_2 = q_2\,, \quad H = H\,, \quad \alpha_1 = \alpha_1\,, \quad \alpha_2 = \alpha_2\,,\\ &\beta_1 = \beta_1\,, \quad \beta_2 = \beta_2\,, \quad \gamma_1 = \gamma_1\,, \quad \gamma_2 = \gamma_2\,,\\ &K_1 = K_1\,, \quad K_2 = K_2\,.\end{aligned} \tag{80}$$

Case 11

$$\begin{aligned}&p_2 = p_2\,, \quad q_2 = 0\,, \quad p_1 = 0\,, \quad q_1 = q_1\,,\\ &k_2 = \varepsilon\frac{\sqrt{3C_1}}{\alpha_2}\,, \quad v_2 = C_1 \text{ or } k_2 = -\varepsilon\frac{\sqrt{3C_1}}{\alpha_2}\,,\\ &v_2 = C_1\,, \quad k_1 = \varepsilon\frac{\sqrt{3C_1}}{\alpha_1}\,, \quad u_2 = 0\,,\\ &H = H\,, \quad \alpha_1 = \alpha_1\,, \quad \alpha_2 = \alpha_2\,,\\ &\beta_1 = \beta_1\,, \quad \beta_2 = \beta_2\,, \quad \gamma_1 = \gamma_1\,, \quad \gamma_2 = \gamma_2\,,\\ &K_1 = K_1\,, \quad K_2 = K_2.\end{aligned} \tag{81}$$

Case 12

$$\begin{aligned}&p_2 = p_2\,, \quad q_2 = 0\,, \quad p_1 = 0\,, \quad q_1 = q_1\,,\\ &k_2 = \varepsilon\frac{\sqrt{3C_1}}{\alpha_2} \text{ or } k_2 = -\varepsilon\frac{\sqrt{3C_1}}{\alpha_2}\,, \quad v_2 = C_1\,,\\ &k_1 = \varepsilon\frac{\sqrt{3C_1}}{\alpha_1}\,,\\ &u_2 = 0\,, \quad K_1 = K_1\,, \quad K_2 = K_2\,,\\ &H = H\,, \quad \alpha_1 = \alpha_1\,, \quad \alpha_2 = \alpha_2\,,\\ &\beta_1 = \beta_1\,, \quad \beta_2 = \beta_2\,, \quad \gamma_1 = \gamma_1\,, \quad \gamma_2 = \gamma_2\,.\end{aligned} \tag{82}$$

Case 13

$$\begin{aligned}&v_2 = C_2\,, \quad p_2 = 0\,, \quad k_1 = \varepsilon\frac{\sqrt{3\beta_1(C_2\beta_1 + C_1\alpha_1)}}{\beta_1\alpha_1}\,,\\ &q_2 = 0\,, \quad p_1 = 0\,, \quad q_1 = q_1\,, \quad k_2 = k_2\,, \quad u_2 = C_1\,,\\ &H = H\,, \quad \alpha_1 = \alpha_1\,, \quad \alpha_2 = \alpha_2\,,\\ &\beta_1 = \beta_1\,, \quad \beta_2 = \beta_2\,, \quad \gamma_1 = \gamma_1\,, \quad \gamma_2 = \gamma_2\,,\\ &K_1 = K_1\,, \quad K_2 = K_2\,.\end{aligned} \tag{83}$$

Case 14

$$\begin{aligned}&p_2 = 0\,, \quad v_2 = \varepsilon\frac{C_1\alpha_2}{\beta_2}\,, \quad k_1 = k_1\,, \quad q_2 = 0\,,\\ &p_1 = 0\,, \quad q_1 = q_1\,, \quad k_2 = 0\,, \quad u_2 = C_1\,,\\ &H = H\,, \quad \alpha_1 = \alpha_1\,, \quad \alpha_2 = \alpha_2\,,\\ &\beta_1 = \beta_1\,, \quad \beta_2 = \beta_2\,, \quad \gamma_1 = \gamma_1\,, \quad \gamma_2 = \gamma_2\,,\\ &K_1 = K_1\,, \quad K_2 = K_2\,.\end{aligned} \tag{84}$$

Case 15

$$\begin{aligned}&v_2 = C_2\,, \quad p_2 = 0\,, \quad k_2 = \varepsilon\frac{\sqrt{3\beta_2(C_2\beta_2 + C_1\alpha_2)}}{\beta_2\alpha_2}\,,\\ &k_1 = k_1\,, \quad q_2 = 0\,, \quad p_1 = 0\,, \quad q_1 = q_1\,,\\ &u_2 = C_1\,, \quad H = H\,, \quad \alpha_1 = \alpha_1\,, \quad \alpha_2 = \alpha_2\,,\\ &\beta_1 = \beta_1\,, \quad \beta_2 = \beta_2\,, \quad \gamma_1 = \gamma_1\,, \quad \gamma_2 = \gamma_2\,,\\ &K_1 = K_1\,, \quad K_2 = K_2\,.\end{aligned} \tag{85}$$

Case 16

$$\begin{aligned}
&p_2 = 0\,, \quad k_1 = k_1\,, \quad q_2 = 0\,, \quad p_1 = 0\,,\\
&q_1 = q_1\,, \quad v_2 = v_2\,, \quad k_2 = k_2\,, \quad u_2 = C_1\,,\\
&H = H\,,\ \alpha_1 = \alpha_1\,,\ \alpha_2 = \alpha_2\,,\\
&\beta_1 = \beta_1\,,\ \beta_2 = \beta_2\,,\ \gamma_1 = \gamma_1\,,\ \gamma_2 = \gamma_2\,,\\
&K_1 = K_1\,,\ K_2 = K_2\,.
\end{aligned} \tag{86}$$

Case 17

$$\begin{aligned}
&p_2 = 0\,, \quad v_2 = \varepsilon\frac{C_1\alpha_2}{\beta_2} \ \text{ or } \ v_2 = -\varepsilon\frac{C_1\alpha_2}{\beta_2}\,,\\
&k_2 = 0\,,\ k_1 = \varepsilon\frac{\sqrt{-3\beta_2\beta_1 C_1(\beta_1\alpha_2 - \beta_2\alpha_1)}}{\beta_2\beta_1\alpha_1}\,,\\
&q_1 = 0\,, \quad p_1 = p_1\,,\ u_2 = C_1\,,\ q_2 = q_2\,,\\
&H = H\,,\ \alpha_1 = \alpha_1\,,\ \alpha_2 = \alpha_2\,,\\
&\beta_1 = \beta_1\,,\ \beta_2 = \beta_2\,,\ \gamma_1 = \gamma_1\,,\ \gamma_2 = \gamma_2\,,\\
&K_1 = K_1\,,\ K_2 = K_2\,.
\end{aligned} \tag{87}$$

Case 18

$$\begin{aligned}
&p_2 = p_2\,, \quad v_2 = \varepsilon\frac{C_1\alpha_2}{\beta_2} \ \text{ or } \ v_2 = -\varepsilon\frac{C_1\alpha_2}{\beta_2}\,,\\
&k_2 = 0\,,\ k_1 = \frac{\sqrt{-3\beta_2\beta_1 C_1}(\beta_1\alpha_2 - \beta_2\alpha_1)}{\beta_2\beta_1\alpha_1}\,,\\
&q_1 = 0\,, \quad p_1 = p_1\,,\ u_2 = C_1\,, \quad q_2 = q_2\,,\\
&H = H\,,\ \alpha_1 = \alpha_1\,,\ \alpha_2 = \alpha_2\,,\\
&\beta_1 = \beta_1\,,\ \beta_2 = \beta_2\,,\ \gamma_1 = \gamma_1\,,\ \gamma_2 = \gamma_2\,,\\
&K_1 = K_1\,,\ K_2 = K_2\,.
\end{aligned} \tag{88}$$

Case 19

$$\begin{aligned}
&p_2 = 0\,, \quad q_1 = q_1\,, \quad k_1 = 0\,, \quad v_2 = \varepsilon\frac{C_1\alpha_1}{\beta_1} \ \text{ or}\\
&v_2 = -\varepsilon\frac{C_1\alpha_1}{\beta_1}\,,\\
&k_2 = -\varepsilon\frac{\sqrt{3\beta_2\beta_1 C_1(\beta_1\alpha_2 - \beta_2\alpha_1)}}{\beta_2\beta_1\alpha_2}\,, \quad p_1 = p_1\,,\\
&u_2 = C_1\,, \quad q_2 = q_2\,,\\
&H = H\,,\ \alpha_1 = \alpha_1\,,\ \alpha_2 = \alpha_2\,,\\
&\beta_1 = \beta_1\,,\ \beta_2 = \beta_2\,,\ \gamma_1 = \gamma_1\,,\ \gamma_2 = \gamma_2\,,\\
&K_1 = K_1\,,\ K_2 = K_2\,.
\end{aligned} \tag{89}$$

Case 20

$$\begin{aligned}
&p_2 = 0\,, \quad v_2 = C_2\,, \quad k_2 = \varepsilon\frac{\sqrt{3\beta_2(C_2\beta_2 + C_1\alpha_2)}}{\beta_2\alpha_2}\\
&\text{or } \ k_2 = -\varepsilon\frac{\sqrt{3\beta_2(C_2\beta_2 + C_1\alpha_2)}}{\beta_2\alpha_2}\,,\\
&k_1 = \varepsilon\frac{\sqrt{3\beta_1(C_2\beta_1 + C_1\alpha_1)}}{\beta_1\alpha_1}\,, \quad q_2 = 0\,,\ q_1 = 0\,,\\
&p_1 = p_1\,, \quad u_2 = C_1\,,\\
&H = H\,,\ \alpha_1 = \alpha_1\,,\ \alpha_2 = \alpha_2\,,\\
&\beta_1 = \beta_1\,,\ \beta_2 = \beta_2\,,\ \gamma_1 = \gamma_1\,,\ \gamma_2 = \gamma_2\,,\\
&K_1 = K_1\,,\ K_2 = K_2\,.
\end{aligned} \tag{90}$$

Case 21

$$\begin{aligned}
&v_2 = C_1\,, \quad k_2 = \varepsilon\frac{\sqrt{3C_1}}{\alpha_2} \ \text{ or } \ k_2 = -\varepsilon\frac{\sqrt{3C_1}}{\alpha_2}\,,\\
&k_1 = -\varepsilon\frac{\sqrt{3C_1}}{\alpha_1}\,, \quad u_2 = 0\,, \quad p_2 = 0\,, \quad q_1 = 0\,,\\
&p_1 = p_1\,, q_2 = q_2\,,\\
&H = H\,,\ \alpha_1 = \alpha_1\,,\ \alpha_2 = \alpha_2\,,\\
&\beta_1 = \beta_1\,,\ \beta_2 = \beta_2\,,\ \gamma_1 = \gamma_1\,,\ \gamma_2 = \gamma_2\,,\\
&K_1 = K_1\,,\ K_2 = K_2\,.
\end{aligned} \tag{91}$$

Case 22

$$\begin{aligned}
&p_2 = 0\,, \quad q_1 = q_1\,, \quad k_2 = \varepsilon\frac{\sqrt{3\beta_2(C_2\beta_2 + C_1\alpha_2)}}{\beta_2\alpha_2}\\
&\text{or } \ k_2 = -\varepsilon\frac{\sqrt{3\beta_2(C_2\beta_2 + C_1\alpha_2)}}{\beta_2\alpha_2}\,,\\
&k_1 = \varepsilon\frac{\sqrt{3}\sqrt{\beta_1(C_2\beta_1 + C_1\alpha_1)}}{\beta_1\alpha_1}\,, \quad v_2 = C_2\,,\\
&q_2 = 0\,, \quad p_1 = p_1\,, \quad u_2 = C_1\,,\\
&H = H\,,\ \alpha_1 = \alpha_1\,,\ \alpha_2 = \alpha_2\,,\\
&\beta_1 = \beta_1\,,\ \beta_2 = \beta_2\,,\ \gamma_1 = \gamma_1\,,\ \gamma_2 = \gamma_2\,,\\
&K_1 = K_1\,,\ K_2 = K_2\,.
\end{aligned} \tag{92}$$

Case 23

$$\begin{aligned}
&p_2 = 0\,, \quad q_1 = q_1\,, \quad k_2 = \varepsilon\frac{\sqrt{3\beta_2(C_2\beta_2 + C_1\alpha_2)}}{\beta_2\alpha_2}\,,\\
&v_2 = C_2\,, \quad k_1 = 0\,, \quad p_1 = p_1\,, \quad u_2 = C_1\,,\\
&q_2 = q_2\,,\ H = H\,,\ \alpha_1 = \alpha_1\,,\ \alpha_2 = \alpha_2\,,\\
&\beta_1 = \beta_1\,,\ \beta_2 = \beta_2\,,\ \gamma_1 = \gamma_1\,,\ \gamma_2 = \gamma_2\,,\\
&K_1 = K_1\,,\ K_2 = K_2\,.
\end{aligned} \tag{93}$$

Case 24

$$\begin{aligned}
&p_2 = p_2\,, \quad v_2 = C_1\,, \quad k_1 = \varepsilon\frac{\sqrt{3C_1}}{\alpha_1} \text{ or}\\
&k_1 = -\varepsilon\frac{\sqrt{3C_1}}{\alpha_1}\,, \quad u_2 = 0\,, \quad q_2 = 0\,, \quad q_1 = 0\,,\\
&p_1 = p_1\,, \quad k_2 = -\varepsilon\frac{\sqrt{3C_1}}{\alpha_2}\,,\\
&H = H\,, \ \alpha_1 = \alpha_1\,, \ \alpha_2 = \alpha_2\,,\\
&\beta_1 = \beta_1\,, \ \beta_2 = \beta_2\,, \ \gamma_1 = \gamma_1\,, \ \gamma_2 = \gamma_2\,,\\
&K_1 = K_1\,, \ K_2 = K_2\,.
\end{aligned} \tag{94}$$

Case 25

$$\begin{aligned}
&p_2 = 0\,, \quad k_2 = k_2\,, \ k_1 = k_1\,, \quad v_2 = v_2\,,\\
&q_1 = q_1\,, \quad q_2 = 0\,, \quad p_1 = p_1\,, \quad u_2 = C_1\,,\\
&H = H\,, \ \alpha_1 = \alpha_1\,, \ \alpha_2 = \alpha_2\,,\\
&\beta_1 = \beta_1\,, \ \beta_2 = \beta_2\,, \ \gamma_1 = \gamma_1\,, \ \gamma_2 = \gamma_2\,,\\
&K_1 = K_1\,, \ K_2 = K_2\,.
\end{aligned} \tag{95}$$

We get the following new solution (called generalized hyperbolic function solution) of Eq. (46):

$$\begin{aligned}
u_1 = {}& 2\frac{\begin{array}{c}K_1k_1\beta_1(t)\,\mathrm{sech}_{p_1q_1k_1}(\xi)(1-2p_1q_1\\ \mathrm{sech}^2_{p_1q_1k_1}(\xi)) + K_2k_2\beta_2(t)\,\mathrm{sech}_{p_2q_2k_2}(\eta)\\ (1-p_2q_2(1+p_2q_2)\,\mathrm{sech}^2_{p_2q_2k_2}(\eta))\end{array}}{\begin{array}{c}H + K_1\tanh_{p_1q_1k_1}(\xi)\,\mathrm{sech}_{p_1q_1k_1}(\xi)\\ +K_2\tanh_{p_2q_2k_2}(\eta)\,\mathrm{sech}_{p_2q_2k_2}(\eta)\end{array}}\\
& \frac{\begin{array}{c}K_1k_1\alpha_1(t)\,\mathrm{sech}_{p_1q_1k_1}(\xi)(1-2p_1q_1\,\mathrm{sech}^2_{p_1q_1k_1}(\xi))\\ +K_2k_2\alpha_2(t)\,\mathrm{sech}_{p_2q_2k_2}(\eta)(1-p_2q_2(1+p_2q_2)\\ \mathrm{sech}^2_{p_2q_2k_2}(\eta))\end{array}}{\begin{array}{c}H + K_1\tanh_{p_1q_1k_1}(\xi)\,\mathrm{sech}_{p_1q_1k_1}(\xi)\\ +K_2\tanh_{p_2q_2k_2}(\eta)\,\mathrm{sech}_{p_2q_2k_2}(\eta)\end{array}}\\
& -2\frac{\begin{array}{c}{k_1}^2K_1\alpha_1(t)\beta_1(t)\tanh_{p_1q_1k_1}(\xi)\,\mathrm{sech}_{p_1q_1k_1}(\xi)\\ (1-6p_1q_1\,\mathrm{sech}^2_{p_1q_1k_1}(\xi))\end{array}}{\begin{array}{c}H + K_1\tanh_{p_1q_1k_1}(\xi)\,\mathrm{sech}_{p_1q_1k_1}(\xi)\\ +K_2\tanh_{p_2q_2k_2}(\eta)\,\mathrm{sech}_{p_2q_2k_2}(\eta)\end{array}}\\
& -2\frac{\begin{array}{c}{k_2}^2K_2\alpha_2(t)\beta_2(t)\tanh_{p_2q_2k_2}(\eta)\,\mathrm{sech}_{p_2q_2k_2}(\eta)\\ (1-4p_2q_2(1+p_2q_2)\,\mathrm{sech}^2_{p_2q_2k_2}(\eta))\end{array}}{\begin{array}{c}H + K_1\tanh_{p_1q_1k_1}(\xi)\,\mathrm{sech}_{p_1q_1k_1}(\xi)\\ +K_2\tanh_{p_2q_2k_2}(\eta)\,\mathrm{sech}_{p_2q_2k_2}(\eta)\end{array}} + u_2\,,
\end{aligned} \tag{96}$$

$$\begin{aligned}
v_1 = {}& 2\frac{\begin{array}{c}(K_1k_1\alpha_1(t)\,\mathrm{sech}_{p_1q_1k_1}(\xi)(1-2p_1q_1\\ \mathrm{sech}^2_{p_1q_1k_1}(\xi)) + K_2k_2\alpha_2(t)\,\mathrm{sech}_{p_2q_2k_2}(\eta)\\ (1-p_2q_2(1+p_2q_2)\,\mathrm{sech}^2_{p_2q_2k_2}(\eta)))^2\end{array}}{\begin{array}{c}(H + K_1\tanh_{p_1q_1k_1}(\xi)\,\mathrm{sech}_{p_1q_1k_1}(\xi)\\ +K_2\tanh_{p_2q_2k_2}(\eta)\,\mathrm{sech}_{p_2q_2k_2}(\eta))^2\end{array}}\\
& -2\frac{\begin{array}{c}{k_1}^2K_1(\alpha_1(t))^2\tanh_{p_1q_1k_1}(\xi)\,\mathrm{sech}_{p_1q_1k_1}(\xi)\\ (1-6p_1q_1\,\mathrm{sech}^2_{p_1q_1k_1}(\xi))\end{array}}{\begin{array}{c}H + K_1\tanh_{p_1q_1k_1}(\xi)\,\mathrm{sech}_{p_1q_1k_1}(\xi)\\ +K_2\tanh_{p_2q_2k_2}(\eta)\,\mathrm{sech}_{p_2q_2k_2}(\eta)\end{array}}\\
& -2\frac{\begin{array}{c}K_2{k_2}^2(\alpha_2(t))^2\tanh_{p_2q_2k_2}(\eta)\,\mathrm{sech}_{p_2q_2k_2}(\eta)\\ (1-2p_2q_2(2+p_2q_2)\,\mathrm{sech}^2_{p_2q_2k_2}(\eta))\end{array}}{\begin{array}{c}H + K_1\tanh_{p_1q_1k_1}(\xi)\,\mathrm{sech}_{p_1q_1k_1}(\xi)\\ +K_2\tanh_{p_2q_2k_2}(\eta)\,\mathrm{sech}_{p_2q_2k_2}(\eta)\end{array}} + v_2\,,
\end{aligned} \tag{97}$$

where $\xi = \alpha_1(t)x + \beta_1(t)y + r_1(t)$, $\eta = \alpha_2(t)x + \beta_2(t)y + r_2(t)$, here $H, K_i, p_i, q_i, k_i, \alpha_i, \beta_i$ and $r_i, i = 1, 2$ satisfy (71)–(95), respectively.

The Generalized F-expansion Method and Its Application in Another (2 + 1)-Dimensional KdV Equation

In the section, we will introduce the main steps of the generalized F-expansion method for constructing exact solutions of NLPDEs which we first presented in [40,41,42]. Then we will make use of the method to find new exact solutions of the (2 + 1)-dimensional KdV equation.

Summary of the Generalized F-expansion Method

In the following we would like to outline the main steps of our general method (called the generalized F-expansion method) in [40,41,42].

Step 1 For a given nonlinear partial differential equation system with some physical fields $u_i(t, x_1, x_2, \ldots, x_m)$, $(i = 1, 2, \ldots, n)$ in $m + 1$ independent variables $t, x_1, x_2, \ldots, x_m$,

$$\begin{cases}
F_1(u_1, \ldots, u_n, u_{1t}, \ldots, u_{nt}, u_{1x_1}, \ldots, u_{nx_m},\\
\qquad u_{1tx_1}, \ldots, u_{ntx_m}, \ldots) = 0\,,\\
F_2(u_1, \ldots, u_n, u_{1t}, \ldots, u_{nt}, u_{1x_1}, \ldots, u_{nx_m},\\
\qquad u_{1tx_1}, \ldots, u_{ntx_m}, \ldots) = 0\,,\\
\ldots\ldots\ldots\ldots,\\
F_n(u_1, \ldots, u_n, u_{1t}, \ldots, u_{nt}, u_{1x_1}, \ldots, u_{nx_m},\\
\qquad u_{1tx_1}, \ldots, u_{ntx_m}, \ldots) = 0\,.
\end{cases} \tag{98}$$

We first consider the following general formal solutions of the Eqs. (98)

$$u_i(t, x_1, x_2, \ldots, x_m) = u_i(\omega)\,, \quad \omega = \omega(x)\,, \quad (i = 1, 2, \ldots, n)\,, \tag{99}$$

where $x = (x_1, x_2, x_3, \ldots, x_m, t)$ and $\omega(x)$ is a function to be determined later. For example, when $n = 2$, we may take $\omega(x) = p(x_2, t)x_1 + q(x_2, t)$ for convenience, here $p(x_2, t)$ and $q(x_2, t)$ are functions to be determined later.

Step 2 We introduce new and more general formal transformations of the Eq. (98), if available, in the forms which we first presented in [42]:

$$\begin{cases} u_1(x) = a_{1_0}(x) + \sum_{i_1=1}^{k_1}(f(\omega(x)))^{i_1-1}\Big(a_{i_1}(x) \\ \quad f(\omega(x)) + b_{i_1}(x)g(\omega(x)) + \frac{c_{i_1}(x)}{f(\omega(x))} + \frac{d_{i_1}(x)}{g(\omega(x))}\Big)\,, \\ u_2(x) = a_{2_0}(x) + \sum_{i_2=1}^{k_2}(f(\omega(x)))^{i_2-1}\Big(a_{i_2}(x) \\ \quad f(\omega(x)) + b_{i_2}(x)g(\omega(x)) + \frac{c_{i_2}(x)}{f(\omega(x))} + \frac{d_{i_2}(x)}{g(\omega(x))}\Big)\,, \\ \ldots\ldots\ldots\ldots, \\ u_n(x) = a_{n_0}(x) + \sum_{i_n=1}^{k_n}(f(\omega(x)))^{i_n-1}\Big(a_{i_n}(x) \\ \quad f(\omega(x)) + b_{i_n}(x)g(\omega(x)) + \frac{c_{i_n}(x)}{f(\omega(x))} + \frac{d_{i_n}(x)}{g(\omega(x))}\Big)\,. \end{cases} \tag{100}$$

where $a_{i_j}^2(x) + b_{i_j}^2(x) + c_{i_j}^2(x) + d_{i_j}^2(x) \neq 0$, $a_{j_0}(x)$, $a_{i_j}(x)$, $b_{i_j}(x)$, $c_{i_j}(x)$, $d_{i_j}(x)(j = 1, 2, \ldots, n, i_j = 1, 2, \ldots, k_j)$ and $\omega(x)$ are all functions to be determined later, $k_j(j = 1, 2, \ldots, n)$ is an integer which is determined by balancing the highest order derivative terms with the nonlinear terms in the given Eqs. (98), and the new variables $f(\omega) = f(\omega(x))$ and $g(\omega) = g(\omega(x))$ satisfy the following relations:

$$\begin{cases} f'^2(\omega) = l_1 f^4(\omega) + m_1 f^2(\omega) + n_1 \\ g'^2(\omega) = l_2 g^4(\omega) + m_2 g^2(\omega) + n_2 \\ g^2(\omega) = \frac{l_1 f^2(\omega)}{l_2} + \frac{m_1 - m_2}{3l_2} \\ n_1 = \frac{{m_1}^2 - {m_2}^2 + 3l_2 n_2}{3l_1} \end{cases} \tag{101}$$

Step 3 Determine $k_j(j = 1, 2, \ldots, n)$ in (100) by balancing the highest nonlinear terms and the highest-order partial differential terms in the given Eq. (98). If k_j is a nonnegative integer, then we first make the transformation $u_j(x) = v_j^{k_j}(x)$.

Step 4 Substitute (100) into the Eq. (98) with (101) and collecting coefficients of polynomials of $f(\omega)$, $g(\omega)$, $\sqrt{l_1 f^4(\omega) + m_1 f^2(\omega) + n_1}$, with the aid of Maple, then setting each coefficient to zero to get a set of over-determined partial differential equations with respect to $l_1, l_2, m_1, m_2, n_1, n_2, a_{j_0}(x), a_{i_j}(x), b_{i_j}(x), c_{i_j}(x), d_{i_j}(x)(j = 1, 2, \ldots, n, i_j = 1, 2, \ldots, k_j)$ and $\omega(x)$.

Step 5 Solving the over-determined partial differential equations with Maple, we then can determine $a_{j_0}(x), a_{i_j}(x), b_{i_j}(x), c_{i_j}(x), d_{i_j}(x)(j = 1, 2, \ldots, n, i_j = 1, 2, \ldots, k_j)$ and $\omega(x)$.

Step 6 Selecting the proper value of parameters $l_1, l_2, m_1, m_2, n_1, n_2$ to determine the corresponding the solutions of the Jacobi elliptic functions $f(\omega)$ and $g(\omega)$ in (101). The relations between the parameters and their corresponding Jacobi elliptic functions are known and are given in the following Table 1.

Remark 2 Relations between values of $l_1, m_1, n_1, l_2, m_2, n_2$ and corresponding $f(\omega)$, $g(\omega)$ in ODEs (101) are in the following Table 1.

Step 7 By using the results obtained in the above step, we can derive a series of generalized solutions, such as different kinds of Jacobi elliptic function solutions in terms of Remark 2. Finally, substituting $a_{j_0}(x), a_{i_j}(x), b_{i_j}(x), c_{i_j}(x), d_{i_j}(x)(j = 1, 2, \ldots, n, i_j = 1, 2, \ldots, k_j)$ and $\omega(x)$ into the generalized solutions with the corresponding solutions of $f(\omega)$, $g(\omega)$, we can get the the new exact solutions of the given Eq. (98).

Remark 3 As is known to all, when $k \to 1$, the Jacobi elliptic functions can degenerate as hyperbolic functions, and when $k \to 0$, the Jacobi elliptic functions degenerate as trigonometric functions. So by this method we can obtain many other new exact solutions to Eq. (98).

Remark 4 The main advantages of the generalized F-expansion method are simpler, more powerful, and more convenient than the Jacobi elliptic function expansion method and the F-expansion method. First of all, with the aid of nonlinear ordinary differential equations (ODEs) (101), one only needs to calculate the functions $f(\omega)$ and $g(\omega)$ which are the solutions of the ODEs (101), instead of calculating the Jacobi elliptic functions one by one. Secondly, the values of the coefficients $l_1, m_1, n_1, l_2, m_2, n_2$ of the ODEs (101) can be selected so that the corresponding solutions of the coupled functions $f(\omega)$ and $g(\omega)$ are Jacobi elliptic functions in the above Table 1. The relations between the coefficients and the corresponding Jacobi elliptic function solutions are known and are given in Remark 2. Thus, in terms of Remark 2, one can simul-

Korteweg–de Vries Equation (KdV), Different Analytical Methods for Solving the, Table 1
The ODEs (101) and Jacobi Elliptic Functions Relation between values of l_i, m_i, n_i $(i = 1, 2)$ and corresponding $f(\omega)$ and $g(\omega)$ in ODEs (101)

l_1	m_1	n_1	l_2	m_2	n_2	$f(\omega)$	$g(\omega)$
k^2	$-1-k^2$	1	$-k^2$	$2k^2-1$	$1-k^2$	$sn(\omega)$	$cn(\omega)$
k^2	$-1-k^2$	1	-1	$2-k^2$	$-1+k^2$	$sn(\omega)$	$dn(\omega)$
$-k^2$	$2k^2-1$	$1-k^2$	-1	$2-k^2$	k^2-1	$cn(\omega)$	$dn(\omega)$
$1-k^2$	$2k^2-1$	$-k^2$	$1-k^2$	$2-k^2$	1	$nc(\omega)$	$sc(\omega)$
1	$2-k^2$	$1-k^2$	1	$2k^2-1$	$-k^2(1-k^2)$	$cs(\omega)$	$ds(\omega)$
1	$-1-k^2$	k^2	$1-k^2$	$2k^2-1$	$-k^2$	$dc(\omega)$	$nc(\omega)$
1	$-1-k^2$	k^2	1	$2-k^2$	$1-k^2$	$ns(\omega)$	$cs(\omega)$
1	$-1-k^2$	k^2	1	$2k^2-1$	$-k^2(1-k^2)$	$ns(\omega)$	$ds(\omega)$
k^2	$-1-k^2$	1	$-1+k^2$	$2-k^2$	-1	$cd(\omega)$	$nd(\omega)$
$-k^2(1-k^2)$	$2k^2-1$	1	$-1+k^2$	$2-k^2$	-1	$sd(\omega)$	$nd(\omega)$
$-k^2(1-k^2)$	$2k^2-1$	1	k^2	$-1-k^2$	1	$sd(\omega)$	$cd(\omega)$
$1-k^2$	$2-k^2$	1	1	$-1-k^2$	k^2	$sc(\omega)$	$dc(\omega)$

taneously obtain more periodic wave solutions expressed by various Jacobi elliptic functions. Thirdly, we present a new transformations (100) that are more general than the transformations of the Jacobi elliptic function expansion method and the F-expansion method.

The Generalized F-expansion Method to Find the Exact Solutions of the (2 + 1)-Dimensional KdV Equation

In this section, we will make use of our method [40,41,42] and symbolic computation to find new exact solutions of the $(2+1)$-dimension KdV equation.

The $(2+1)$-dimension KdV equation in [20] is written in the following form

$$u_t(x,y,t) - 4u(x,y,t)u_y(x,y,t) - 4u_x(x,y,t)\partial_x^{-1} u_y(x,y,t) - u_{xxy}(x,y,t) = 0\,. \quad (102)$$

where ∂_x^{-1} is an indefinite integrate operator $\int \mathrm{d}x$, possessing some interesting coherent structures.

We discuss the similar solution of (102) in its potential form $(u = v_x)$,

$$v_{xt}(x,y,t) - v_{xxxy}(x,y,t) - 4v_x(x,y,t)v_{xy}(x,y,t) - 4v_{xx}(x,y,t)v_y(x,y,t) = 0\,. \quad (103)$$

By balancing the highest nonlinear terms and the highest-order partial derivative terms in (103), we suppose (103) to have the following formal solution:

$$v = c_0(y,t) + c_1(y,t)g(\omega) + c_2(y,t)f(\omega) + \frac{c_3(y,t)}{g(\omega)} + \frac{c_4(y,t)}{f(\omega)}\,, \quad (104)$$

where $\omega = \alpha(y,t)x + p(y,t) + q(t)$, $\alpha(y,t)$, $p(y,t)$, $q(t)$, $c_0(y,t)$, $c_1(y,t)$, $c_2(y,t)$, $c_3(y,t)$ and $c_4(y,t)$ are all functions to be determined later. And the new variables $f(\omega)$ and $g(\omega)$ satisfy the relation (101).

Substitute (104) into the Eq. (103) with (101). Collect coefficients of polynomials of $f(\omega), g(\omega)$, $\sqrt{l_1 f^4(\omega) + m_1 f^2(\omega) + n_1}$, with the aid of Maple, then set each coefficient to zero to get a set of over-determined partial differential equations with respect to $l_1, l_2, m_1, m_2, n_1, n_2, \alpha(y,t), p(y,t), c_i(y,t), (i = 0,1,2,3,4)$ and $q(t)$ as follows:

$$\begin{aligned}
&120(\alpha(y,t))^2 c_4(y,t)c_1(y,t)\left(\frac{\partial}{\partial y}\alpha(y,t)\right) x{m_1}^4{m_2}^2 \\
&+ 168(\alpha(y,t))^2 c_4(y,t)l_2c_3(y,t)\left(\frac{\partial}{\partial y}p(y,t)\right) {m_1}^4 m_2 \\
&+ 120(\alpha(y,t))^2 c_4(y,t)c_1(y,t)(\frac{\partial}{\partial y}p(y,t)){m_2}^2{m_1}^4 \\
&+ 48(\alpha(y,t))^2 c_4(y,t)l_2c_3(y,t)\left(\frac{\partial}{\partial y}\alpha(y,t)\right) x{m_1}^2{m_2}^3 \\
&- 80(\alpha(y,t))^2 c_4(y,t)c_1(y,t)(\frac{\partial}{\partial y}\alpha(y,t))x{m_2}^3{m_1}^3 \\
&- 72(\alpha(y,t))^2 c_4(y,t)c_1(y,t)\left(\frac{\partial}{\partial y}p(y,t)\right) {m_1}^5 m_2 \\
&+ 24(\alpha(y,t))^2 c_4(y,t)c_1(y,t)\left(\frac{\partial}{\partial y}\alpha(y,t)\right) x{m_2}^5 m_1 \\
&+ \cdots = 0\,, \\
&23328c_2(y,t)(\alpha(y,t))^3 l_2{l_1}^5\left(\frac{\partial}{\partial y}p(y,t)\right) m_2 \\
&+ 1458c_2(y,t)\alpha(y,t)l_2{l_1}^5\frac{\mathrm{d}}{\mathrm{d}t}q(t) - 37908c_2(y,t)
\end{aligned}$$

$$
\begin{aligned}
&(\alpha(y,t))^3 l_2 {l_1}^5 \left(\frac{\partial}{\partial y} p(y,t)\right) m_1 + 23328 c_2(y,t) \\
&(\alpha(y,t))^3 l_2 {l_1}^5 \left(\frac{\partial}{\partial y} \alpha(y,t)\right) x m_2 + 1458 c_2(y,t) \\
&\alpha(y,t) l_2 {l_1}^5 \left(\frac{\partial}{\partial t} \alpha(y,t)\right) x \\
&- 5832(\alpha(y,t))^2 \left(\frac{\partial}{\partial y} c_0(y,t)\right) l_2 {l_1}^5 c_2(y,t) \\
&- 37908 c_2(y,t)(\alpha(y,t))^3 l_2 {l_1}^5 \left(\frac{\partial}{\partial y} \alpha(y,t)\right) x m_1 \\
&+ \cdots = 0 \\
&- 60(\alpha(y,t))^3 c_1(y,t) {m_1}^3 \left(\frac{\partial}{\partial y} \alpha(y,t)\right) x {m_2}^3 l_1 \\
&- 540(\alpha(y,t))^3 c_1(y,t) l_2 n_2 \left(\frac{\partial}{\partial y} p(y,t)\right) {m_2}^4 l_1 \\
&+ 2160(\alpha(y,t))^3 c_1(y,t) l_2 n_2 \left(\frac{\partial}{\partial y} \alpha(y,t)\right) x m_1 {m_2}^3 l_1 \\
&- 3240(\alpha(y,t))^3 c_1(y,t) l_2 n_2 \left(\frac{\partial}{\partial y} \alpha(y,t)\right) x {m_1}^2 {m_2}^2 l_1 \\
&+ 2160(\alpha(y,t))^3 c_1(y,t) l_2 n_2 \left(\frac{\partial}{\partial y} \alpha(y,t)\right) x {m_1}^3 m_2 l_1 \\
&- 1296 c_3(y,t)(\alpha(y,t))^3 {l_2}^2 n_2 \left(\frac{\partial}{\partial y} \alpha(y,t)\right) x l_1 {m_2}^3 \\
&+ 180 c_3(y,t)(\alpha(y,t))^3 l_2 l_1 {m_1}^3 \left(\frac{\partial}{\partial y} \alpha(y,t)\right) x {m_2}^2 \\
&+ \cdots = 0 , \\
&\cdots
\end{aligned}
\tag{105}
$$

Because there are so many over-determined partial differential equations, only a few of them are shown here for convenience. With the aid of Maple symbolic computation software, we solve the over-determined partial differential equations, and we get the following solutions of $\alpha(y,t), p(y,t), q(t), c_0(y,t), c_1(y,t), c_2(y,t), c_3(y,t)$ and $c_4(y,t)$:

Case 1

$$
\begin{aligned}
&c_4(y,t) = \frac{C_1}{F_1(t)} , \quad p(y,t) = F_2(t) , \quad c_3(y,t) = \frac{C_2}{F_1(t)} , \\
&\alpha(y,t) = F_1(t) , \quad c_2(y,t) = \frac{C_3}{F_1(t)} , \quad c_1(y,t) = \frac{C_4}{F_1(t)} , \\
&q(t) = \int 4F_1(t)F_3(t) - \left(\frac{\mathrm{d}}{\mathrm{d}t} F_1(t)\right) x \\
&\quad - \frac{\mathrm{d}}{\mathrm{d}t} F_2(t)\mathrm{d}t + C_5 , \quad c_0(y,t) = F_3(t)y + F_4(t) ,
\end{aligned}
\tag{106}
$$

where C_i, $i = 1,2,3,4,5$ are arbitrary constants and F_i, $i = 1,2,3,4$ are all arbitrary functions with respect to variable t.

Case 2

$$
\begin{aligned}
&c_2(y,t) = -\frac{48 C_1 l_1}{F_1(t)(m_1 - m_2)} , \\
&q(t) = \int 4F_1(t)F_3(t) - \left(\frac{\mathrm{d}}{\mathrm{d}t} F_1(t)\right) x \\
&\quad - \frac{\mathrm{d}}{\mathrm{d}t} F_2(t)\mathrm{d}t + C_4 , \\
&\alpha(y,t) = F_1(t) , \quad c_1(y,t) = \frac{C_3}{F_1(t)} , \\
&c_4(y,t) = \frac{C_1}{F_1(t)} , \quad p(y,t) = F_2(t) , \\
&c_3(y,t) = \frac{C_2}{F_1(t)} , \quad c_0(y,t) = F_3(t)y + F_4(t) ,
\end{aligned}
\tag{107}
$$

where C_i, $i = 1,2,3,4$ are arbitrary constants and F_i, $i = 1,2,3,4$ are all arbitrary functions with respect to variable t.

Case 3

$$
\begin{aligned}
&c_2(y,t) = -\frac{48 C_1 l_1}{F_1(t)(m_1 - m_2)} , \quad c_4(y,t) = \frac{C_1}{F_1(t)} , \\
&p(y,t) = F_2(t) , \quad c_3(y,t) = \frac{C_2}{F_1(t)} , \quad \alpha(y,t) = F_1(t) , \\
&c_0(y,t) = F_3(t)y + F_4(t) , \quad c_1(y,t) = 0 , \\
&q(t) = \int 4F_1(t)F_3(t) - \left(\frac{\mathrm{d}}{\mathrm{d}t} F_1(t)\right) x \\
&\quad - \frac{\mathrm{d}}{\mathrm{d}t} F_2(t)\mathrm{d}t + C_3 ,
\end{aligned}
\tag{108}
$$

where C_i, $i = 1,2,3,4$ are arbitrary constants and F_i, $i = 1,2,3,4$ are all arbitrary functions with respect to variable t.

Case 4

$$
\begin{aligned}
&c_3(y,t) = c_3(y,t) , \quad c_4(y,t) = c_4(y,t) , \\
&q(t) = q(t) , \quad c_0(y,t) = c_0(y,t) , \\
&p(y,t) = p(y,t) , \quad c_1(y,t) = c_1(y,t) , \\
&c_2(y,t) = c_2(y,t) , \quad \alpha(y,t) = 0 ,
\end{aligned}
\tag{109}
$$

where C_1 is an arbitrary constant and $c_3(y,t), c_4(y,t), q(t), c_0(y,t), p(y,t), c_1(y,t), c_2(y,t)$ are all arbitrary functions with respect to variable y and t.

Select the proper value of parameters $l_1, l_2, m_1, m_2, n_1, n_2$ to determine the corresponding Jacobi elliptic functions $f(\omega)$ and $g(\omega)$ in (101). The relations between the

parameters and the corresponding Jacobi elliptic function are known and given in Table 1, above.

By using the results obtained in the above step, we can derive a series of generalized solutions, such as different kinds of Jacobi elliptic functions solutions. Finally, by substituting (106)–(109) into the generalized solutions with the corresponding solutions of $f(\omega)$ and $g(\omega)$, respectively, we can get the the new exact solutions of the given Eqs. (103).

Type 1 If we select $l_1 = k^2, m_1 = -(1+k^2), n_1 = 1, l_2 = -k^2, m_2 = (2k^2-1), n_2 = (1-k^2)$, corresponding to (101), we can get $f = sn(\omega), g = cn(\omega)$.

Thus we get the following Jacobi elliptic function solutions of the equation (103):

$$v_1(x,y,t) = c_0(y,t) + c_1(y,t)cn(\omega) + c_2(y,t)sn(\omega) + \frac{c_3(y,t)}{cn(\omega)} + \frac{c_4(y,t)}{sn(\omega)}, \quad (110)$$

where $\omega = \alpha(y,t)x + p(y,t) + q(t), \alpha(y,t), p(y,t), q(t), c_0(y,t), c_1(y,t), c_2(y,t), c_3(y,t)$ and $c_4(y,t)$ satisfy (106)–(109) respectively.

For example, when we select $\alpha(y,t), p(y,t), q(t), c_0(y,t), c_1(y,t), c_2(y,t), c_3(y,t)$ and $c_4(y,t)$ to satisfy (106), namely Case 1, we can easily get the following Jacobi elliptic function solution of Eq. (103):

$$v_2(x,y,t) = F_3(t)y + F_4(t) + \frac{C_4 cn(\omega)}{F_1(t)} + \frac{C_3 sn(\omega)}{F_1(t)} + \frac{C_2}{F_1(t)cn(\omega)} + \frac{C_1}{F_1(t)sn(\omega)}, \quad (111)$$

where $F_i, i = 1,2,3,4$ are arbitrary functions of t, $C_i (i = 1,2,3,4,5)$ are arbitrary constants, and

$$\omega = F_1(t)x + F_2(t) + \int 4F_1(t)F_3(t) - \left(\frac{\mathrm{d}}{\mathrm{d}t}F_1(t)\right)x - \frac{\mathrm{d}}{\mathrm{d}t}F_2(t)\mathrm{d}t + C_5 .$$

When $k \to 1, sn(\omega) \to \tanh(\omega), cn(\omega) \to \mathrm{sech}(\omega)$, so we get degenerative soliton-like solutions from the solution (111):

$$v_3(x,y,t) = F_3(t)y + F_4(t) + \frac{C_4\,\mathrm{sech}(\omega)}{F_1(t)} + \frac{C_3\tanh(\omega)}{F_1(t)} + \frac{C_2}{F_1(t)\,\mathrm{sech}(\omega)} + \frac{C_1}{F_1(t)\tanh(\omega)}. \quad (112)$$

When $k \to 0, sn(\omega) \to \sin(\omega), cn(\omega) \to \cos(\omega)$, so we get the degenerative trigonometric function solutions from the solution (111):

$$v_4(x,y,t) = F_3(t)y + F_4(t) + \frac{C_4\cos(\omega)}{F_1(t)} + \frac{C_3\sin(\omega)}{F_1(t)} + \frac{C_2}{F_1(t)\cos(\omega)} + \frac{C_1}{F_1(t)\sin(\omega)}. \quad (113)$$

So from Case 1, we can obtain Jacobi elliptic function solutions, degenerative soliton-like solutions and trigonometric function solutions of Eq. (103). We can also get many other solutions if we make use of Case 2–Case 4. Therefore, through selecting $l_1 = k^2, m_1 = -(1+k^2), n_1 = 1, l_2 = -k^2, m_2 = (2k^2-1), n_2 = (1-k^2)$, we can get families of new exact solutions of Eq. (103).

Type 2 If we select $l_1 = -k^2, m_1 = (2k^2-1), n_1 = 1-k^2, l_2 = -1, m_2 = (2-k^2), n_2 = k^2-1$, corresponding to (101), we can get $f = cn(\omega), g = dn(\omega)$.

So we get the following Jacobi elliptic function solutions of Eq. (103):

$$v_5(x,y,t) = c_0(y,t) + c_1(y,t)dn(\omega) + c_2(y,t)cn(\omega) + \frac{c_3(y,t)}{dn(\omega)} + \frac{c_4(y,t)}{cn(\omega)}, \quad (114)$$

where $\omega = \alpha(y,t)x + p(y,t) + q(t), \alpha(y,t), p(y,t), q(t), c_0(y,t), c_1(y,t), c_2(y,t), c_3(y,t)$ and $c_4(y,t)$ satisfy (106)–(109), respectively.

For example, when we select $\alpha(y,t), p(y,t), q(t), c_0(y,t), c_1(y,t), c_2(y,t), c_3(y,t)$ and $c_4(y,t)$ to satisfy (107), namely Case 2, we can easily get the following Jacobi elliptic function solution of Eq. (103):

$$v_6(x,y,t) = F_3(t)y + F_4(t) + \frac{C_3 dn(\omega)}{F_1(t)} - 48\frac{C_1 l_1 cn(\omega)}{F_1(t)(m_1-m_2)} + \frac{C_2}{F_1(t)dn(\omega)} + \frac{C_1}{F_1(t)cn(\omega)}. \quad (115)$$

where $F_i(t), i = 1,2,3,4$ are arbitrary functions of t, $C_i (i = 1,2,3,4)$ are arbitrary constants, and

$$\omega = F_1(t)x + F_2(t) + \int 4F_1(t)F_3(t) - \left(\frac{\mathrm{d}}{\mathrm{d}t}F_1(t)\right)x - \frac{\mathrm{d}}{\mathrm{d}t}F_2(t)\mathrm{d}t + C_4 .$$

When $k \to 1, cn(\omega) \to \mathrm{sech}(\omega), dn(\omega) \to \mathrm{sech}(\omega)$, so we get a degenerative soliton-like solution from the solu-

tion (115):

$$v_7(x,y,t) = F_3(t)y + F_4(t) + \frac{C_3\,\mathrm{sech}(\omega)}{F_1(t)} - 48\frac{C_1 l_1\,\mathrm{sech}(\omega)}{F_1(t)(m_1-m_2)} + \frac{C_1+C_2}{F_1(t)\,\mathrm{sech}(\omega)}\,, \quad (116)$$

When $k \to 0$, $cn(\omega) \to cos(\omega)$, $dn(\omega) \to 1$, so we get the degenerative trigonometric function solutions from the solution (115):

$$v_8 = F_3(t)y + F_4(t) + \frac{C_2+C_3}{F_1(t)} - 48\frac{C_1 l_1\cos(\omega)}{F_1(t)(m_1-m_2)} + \frac{C_1}{F_1(t)\cos(\omega)}\,. \quad (117)$$

where $F_i(t)$, $i = 1,2,3,4$ are arbitrary functions of t and $C_i (i = 1,2,3,4)$ are arbitrary constants.

So from Case 2, we we can get the Jacobi elliptic function solutions, degenerative soliton-like solutions and trigonometric function solutions of Eq. (103). We can also get some other solutions if we make use of other cases. Therefore, through selecting $l_1 = -k^2$, $m_1 = (2k^2-1)$, $n_1 = 1-k^2$, $l_2 = -1$, $m_2 = (2-k^2)$, $n_2 = k^2-1$, we can get families of new exact solutions of Eq. (103).

Type 3 If we select $l_1 = 1$, $m_1 = (2-k^2)$, $n_1 = 1-k^2$, $l_2 = 1$, $m_2 = (2k^2-1)$, $n_2 = -k^2(1-k^2)$, corresponding to (101), we can get $f = cs(\omega)$, $g = ds(\omega)$.

So we get the following Jacobi elliptic function solutions of Eq. (103):

$$v_9(x,y,t) = c_0(y,t) + c_1(y,t)ds(\omega) + c_2(y,t)cs(\omega) + \frac{c_3(y,t)}{ds(\omega)} + \frac{c_4(y,t)}{cs(\omega)}\,, \quad (118)$$

where $\omega = \alpha(y,t)x + p(y,t) + q(t)$, $\alpha(y,t)$, $p(y,t)$, $q(t)$, $c_0(y,t)$, $c_1(y,t)$, $c_2(y,t)$, $c_3(y,t)$ and $c_4(y,t)$ satisfy (106)–(109), respectively.

For example, when we select $\alpha(y,t)$, $p(y,t)$, $q(t)$, $c_0(y,t)$, $c_1(y,t)$, $c_2(y,t)$, $c_3(y,t)$ and $c_4(y,t)$ to satisfy (108), namely Case 3, we can easily get the following Jacobi elliptic function solution of Eq. (103):

$$v_{10}(x,y,t) = F_3(t)y + F_4(t) - 48\frac{C_1 l_1 cs(\omega)}{F_1(t)(m_1-m_2)} + \frac{C_2}{F_1(t)ds(\omega)} + \frac{C_1}{F_1(t)cs(\omega)}\,, \quad (119)$$

where $F_i(t)$, $i = 1,2,3,4$ are arbitrary functions of t, $C_i (i = 1,2,3,4)$ are arbitrary constants, and

$$\omega = F_1(t)x + F_2(t)q(t) = \int 4F_1(t)F_3(t) - \left(\frac{\mathrm{d}}{\mathrm{d}t}F_1(t)\right)x - \frac{\mathrm{d}}{\mathrm{d}t}F_2(t)\mathrm{d}t + C_3\,.$$

When $k \to 1$, $cs(\omega) \to \mathrm{sech}(\omega)\coth(\omega)$, $ds(\omega) \to \mathrm{sech}(\omega)\coth(\omega)$, so we get a degenerative soliton-like solution from the solution (119):

$$v_{11}(x,y,t) = F_3(t)y + F_4(t) - 48\frac{C_1 l_1\,\mathrm{sech}(\omega)\coth(\omega)}{F_1(t)(m_1-m_2)} + \frac{C_1+C_2}{F_1(t)\,\mathrm{sech}(\omega)\coth(\omega)}\,. \quad (120)$$

When $k \to 0$, $cs(\omega) \to cot(\omega)$, $ds(\omega) \to \csc(\omega)$, so we get the triangular function solution from the solution (119):

$$v_{12}(x,y,t) = F_3(t)y + F_4(t) - 48\frac{C_1 l_1\cot(\omega)}{F_1(t)(m_1-m_2)} + \frac{C_2}{F_1(t)\csc(\omega)} + \frac{C_1}{F_1(t)\cot(\omega)}\,. \quad (121)$$

So from Case 3, we can get Jacobi elliptic function solutions, degenerative soliton-like solutions and trigonometric function solutions of Eq. (103). We can also get many other solutions if we make use of other cases. Therefore, through selecting $l_1 = 1$, $m_1 = (2-k^2)$, $n_1 = 1-k^2$, $l_2 = 1$, $m_2 = (2k^2-1)$, $n_2 = -k^2(1-k^2)$, we can get families of new exact solutions of Eq. (103).

Type 4 If we select $l_1 = 1-k^2$, $m_1 = (2k^2-1)$, $n_1 = -k^2$, $l_2 = (1-k^2)$, $m_2 = (2-k^2)$, $n_2 = 1$, corresponding to (101), we can get $f = nc(\omega)$, $g = sc(\omega)$.

So we get the following Jacobi elliptic function solutions of Eq. (103):

$$v_{13}(x,y,t) = c_0(y,t) + c_1(y,t)sc(\omega) + c_2(y,t)nc(\omega) + \frac{c_3(y,t)}{sc(\omega)} + \frac{c_4(y,t)}{nc(\omega)}\,, \quad (122)$$

where $\omega = \alpha(y,t)x + p(y,t) + q(t)$, $\alpha(y,t)$, $p(y,t)$, $q(t)$, $c_0(y,t)$, $c_1(y,t)$, $c_2(y,t)$, $c_3(y,t)$ and $c_4(y,t)$ satisfy (106)–(109), respectively.

So, in the same way, we can get many other Jacobi elliptic function solutions of Eq. (103). When $k \to 1$, or $k \to 0$, we can also get families of new exact solutions of Eq. (103).

Type 5 If we select $l_1 = k^2$, $m_1 = -(1+k^2)$, $n_1 = 1$, $l_2 = -1$, $m_2 = (2-k^2)$, $n_2 = -(1-k^2)$, corresponding to (101), we can get $f = sn(\omega)$, $g = dn(\omega)$.

So we get the following Jacobi elliptic function solutions of Eq. (103):

$$v_{14}(x,y,t) = c_0(y,t) + c_1(y,t)dn(\omega) + c_2(y,t)sn(\omega) + \frac{c_3(y,t)}{dn(\omega)} + \frac{c_4(y,t)}{sn(\omega)}\,, \quad (123)$$

where $\omega = \alpha(y,t)x + p(y,t) + q(t), \alpha(y,t), p(y,t), q(t), c_0(y,t), c_1(y,t), c_2(y,t), c_3(y,t)$ and $c_4(y,t)$ satisfy (106)–(109), respectively.

So, in the same way, we can get many other Jacobi elliptic function solutions of Eq. (103). When $k \to 1$, or $k \to 0$, we can also get families of new exact solutions of Eq. (103).

Type 6 If we select $l_1 = 1, m_1 = -(1+k^2), n_1 = k^2, l_2 = 1-k^2, m_2 = (2k^2-1), n_2 = -k^2$, corresponding to (101), we can get $f = dc(\omega), g = nc(\omega)$.

So we get the following Jacobi elliptic function solutions of Eq. (103):

$$v_{15}(x,y,t) = c_0(y,t) + c_1(y,t)nc(\omega) + c_2(y,t)dc(\omega) + \frac{c_3(y,t)}{nc(\omega)} + \frac{c_4(y,t)}{dc(\omega)}, \quad (124)$$

where $\omega = \alpha(y,t)x + p(y,t) + q(t), \alpha(y,t), p(y,t), q(t), c_0(y,t), c_1(y,t), c_2(y,t), c_3(y,t)$ and $c_4(y,t)$ satisfy (106)–(109), respectively.

So, in the same way, we can get many other Jacobi elliptic function solutions of Eq. (103). When $k \to 1$, or $k \to 0$, we can also get families of new exact solutions of Eq. (103).

Type 7 If we select $l_1 = 1, m_1 = -(1+k^2), n_1 = k^2, l_2 = 1, m_2 = 2-k^2, n_2 = 1-k^2$, corresponding to (101), we can get $f = ns(\omega), g = cs(\omega)$.

So we get the following Jacobi elliptic function solutions:

$$v_{16}(x,y,t) = c_0(y,t) + c_1(y,t)cs(\omega) + c_2(y,t)ns(\omega) + \frac{c_3(y,t)}{cs(\omega)} + \frac{c_4(y,t)}{ns(\omega)}, \quad (125)$$

where $\omega = \alpha(y,t)x + p(y,t) + q(t), \alpha(y,t), p(y,t), q(t), c_0(y,t), c_1(y,t), c_2(y,t), c_3(y,t)$ and $c_4(y,t)$ satisfy (106)–(109), respectively.

So, in the same way, we can get many other Jacobi elliptic function solutions of Eq. (103). When $k \to 1$, or $k \to 0$, we can also get families of new exact solutions of Eq. (103).

Type 8 If we select $l_1 = 1, m_1 = -(1+k^2), n_1 = k^2, l_2 = 1, m_2 = 2k^2-1, n_2 = -k^2(1-k^2)$, corresponding to (101), we can get $f = ns(\omega), g = ds(\omega)$.

So we get the following Jacobi elliptic function solutions:

$$v_{17}(x,y,t) = c_0(y,t) + c_1(y,t)ds(\omega) + c_2(y,t)ns(\omega) + \frac{c_3(y,t)}{ds(\omega)} + \frac{c_4(y,t)}{ns(\omega)}, \quad (126)$$

where $\omega = \alpha(y,t)x + p(y,t) + q(t), \alpha(y,t), p(y,t), q(t), c_0(y,t), c_1(y,t), c_2(y,t), c_3(y,t)$ and $c_4(y,t)$ satisfy (106)–(109), respectively.

So, in the same way, we can get many other Jacobi elliptic function solutions of Eq. (103). When $k \to 1$, or $k \to 0$, we can also get families of new exact solutions of Eq. (103).

Type 9 If we select $l_1 = k^2, m_1 = -(1+k^2), n_1 = 1, l_2 = -(1-k^2), m_2 = 2-k^2, n_2 = -1$, corresponding to (101), we can get $f = cd(\omega), g = nd(\omega)$.

So we get the following Jacobi elliptic function solutions of Eq. (103):

$$v_{18}(x,y,t) = c_0(y,t) + c_1(y,t)nd(\omega) + c_2(y,t)cd(\omega) + \frac{c_3(y,t)}{nd(\omega)} + \frac{c_4(y,t)}{cd(\omega)}, \quad (127)$$

where $\omega = \alpha(y,t)x + p(y,t) + q(t), \alpha(y,t), p(y,t), q(t), c_0(y,t), c_1(y,t), c_2(y,t), c_3(y,t)$ and $c_4(y,t)$ satisfy (106)–(109), respectively.

So, in the same way, we can get many other Jacobi elliptic function solutions of Eq. (103). When $k \to 1$, or $k \to 0$, we can also get families of new exact solutions of Eq. (103).

Type 10 If we select $l_1 = -k^2(1-k^2), m_1 = 2k^2-1, n_1 = 1, l_2 = -(1-k^2), m_2 = 2-k^2, n_2 = -1$, corresponding to (101), we can get $f = sd(\omega), g = nd(\omega)$.

So we get the following Jacobi elliptic function solutions of Eq. (103):

$$v_{19}(x,y,t) = c_0(y,t) + c_1(y,t)nd(\omega) + c_2(y,t)sd(\omega) + \frac{c_3(y,t)}{nd(\omega)} + \frac{c_4(y,t)}{sd(\omega)}, \quad (128)$$

where $\omega = \alpha(y,t)x + p(y,t) + q(t), \alpha(y,t), p(y,t), q(t), c_0(y,t), c_1(y,t), c_2(y,t), c_3(y,t)$ and $c_4(y,t)$ satisfy (106)–(109), respectively.

So, in the same way, we can get many other Jacobi elliptic function solutions of Eq. (103). When $k \to 1$, or $k \to 0$, we can also get families of new exact solutions of Eq. (103).

Type 11 If we select $l_1 = -k^2(1-k^2)$, $m_1 = 2k^2 - 1$, $n_1 = 1$, $l_2 = k^2$, $m_2 = -(1+k^2)$, $n_2 = 1$, corresponding to (101), we can get $f = sd(\omega)$, $g = cd(\omega)$.

So we get the following Jacobi elliptic function solutions of Eq. (103):

$$v_{20}(x,y,t) = c_0(y,t) + c_1(y,t)cd(\omega) + c_2(y,t)sd(\omega) + \frac{c_3(y,t)}{cd(\omega)} + \frac{c_4(y,t)}{sd(\omega)}, \quad (129)$$

where $\omega = \alpha(y,t)x + p(y,t) + q(t)$, $\alpha(y,t)$, $p(y,t)$, $q(t)$, $c_0(y,t)$, $c_1(y,t)$, $c_2(y,t)$, $c_3(y,t)$ and $c_4(y,t)$ satisfy (106)–(109), respectively.

So, in the same way, we can get many other Jacobi elliptic function solutions of Eq. (103). When $k \to 1$, or $k \to 0$, we can also get families of new exact solutions of Eqs. (103).

Type 12 If we select $l_1 = 1 - k^2$, $m_1 = 2 - k^2$, $n_1 = 1$, $l_2 = 1$, $m_2 = -(1+k^2)$, $n_2 = k^2$, corresponding to (101), we can get $f = sc(\omega)$, $g = dc(\omega)$.

So we get the following Jacobi elliptic function solutions:

$$v_{21}(x,y,t) = c_0(y,t) + c_1(y,t)dc(\omega) + c_2(y,t)sc(\omega) + \frac{c_3(y,t)}{dc(\omega)} + \frac{c_4(y,t)}{sc(\omega)}, \quad (130)$$

where $\omega = \alpha(y,t)x + p(y,t) + q(t)$, $\alpha(y,t)$, $p(y,t)$, $q(t)$, $c_0(y,t)$, $c_1(y,t)$, $c_2(y,t)$, $c_3(y,t)$ and $c_4(y,t)$ satisfy (106)–(109), respectively.

So, in the same way, we can get many other Jacobi elliptic function solutions of Eq. (103). When $k \to 1$, or $k \to 0$, we can also get families of new exact solutions of Eq. (103).

Remark 5 From the details above, it is very easy to see that we have gotten many new exact solutions of the $(2+1)$-dimensional KdV equation (103). The solutions we get include Jacobi elliptic doubly periodic function solutions, soliton-like solutions and triangular function solutions. These solutions are more general than the solutions which the extended F-expansion method gets, and they are more abundant than the solutions in [34]. Besides, our method is more convenient than the method in [34]. Among the solutions, the arbitrary functions imply that these solutions have rich local structures. In this paper, we have provided some figures to describe the character of the new exact solutions of Eq. (103).

The Generalized Algebra Method and Its Application in (1 + 1)-Dimensional Generalized Variable – Coefficient KdV Equation

In this section, with the aid of the Maple symbolic computation system, we will develop the algebraic methods [50,51] for constructing the traveling wave solutions of NLPDEs and present the following new methods and theorems:

1. A new and general transformation, a new theorem and its proof by using Maple, are presented in [40].
2. A new mechanization method to find the exact solutions of a first-order nonlinear ordinary differential equation with any degree. The validity and reliability of the method are tested by its application to the first-order nonlinear ordinary differential equation with six degrees, eight degrees, ten degrees, and twelve degrees in [40,43,44].
3. A general transformation, a new generalized algebraic method and their algorithms are suggested based on a nonlinear ordinary differential equation with any degree. The (1 + 1)-Dimensional Generalized Variable – Coefficient KdV Equations are chosen to illustrate our algorithm so that more families of new exact solutions are obtained, which contain both non-traveling and traveling wave solutions in [40,43,44].

Recently, much research work has been concentrated on the various extensions and applications of the algebraic method with computerized symbolic computation [50,51]. The algebraic method can obtain many types of traveling wave solutions based on a nonlinear ordinary differential equation with four degrees and the following theorem:

Theorem 3 *The following nonlinear ordinary differential equation with four degrees*

$$\frac{d}{d\xi}\phi(\xi) = \sqrt{c_0 + c_1\phi(\xi) + c_2\phi^2(\xi) + c_3\phi^3(\xi) + c_4\phi^4(\xi)}, \quad c_i = constant, \quad i = 0,1,2,3,4 \quad (131)$$

admits many kinds of fundamental solutions, some of which are listed as follows.

Case 1 If $c_0 = c_1 = c_3 = 0$, Eq. (131) admits: a bell shaped solitary wave solution

$$\phi(\xi) = \sqrt{-\frac{c_2}{c_4}}\,\mathrm{sech}\left(\sqrt{c_2}\xi\right), \quad c_2 > 0, \quad c_4 < 0, \quad (132)$$

a triangular type solution

$$\phi(\xi) = \sqrt{-\frac{c_2}{c_4}} sec\left(\sqrt{-c_2}\xi\right), \quad c_2 < 0, \quad c_4 > 0, \quad (133)$$

a rational polynomial type solution

$$\phi(\xi) = -\frac{1}{\sqrt{c_4}\xi}, \quad c_2 = 0, \quad c_4 > 0. \quad (134)$$

Case 2 If $c_0 = d_2^2/(4c_4)$, $c_1 = c_3 = 0$, Eq. (131) admits a kink shaped solitary wave solution

$$\phi(\xi) = \sqrt{-\frac{c_2}{2c_4}} \tanh\left(\sqrt{-\frac{c_2}{2}}\xi\right), \quad c_2 < 0, \quad c_4 > 0, \quad (135)$$

a triangular type solution

$$\phi(\xi) = \sqrt{\frac{c_2}{2c_4}} \tan\left(\sqrt{\frac{c_2}{2}}\xi\right), \quad c_2 > 0, \quad c_4 > 0, \quad (136)$$

a rational polynomial type solution

$$\phi(\xi) = -\frac{1}{\sqrt{c_4}\xi}, \quad c_2 = 0, \quad c_4 > 0. \quad (137)$$

Case 3 If $c_1 = c_3 = 0$, Eq. (131) admits three Jacobian elliptic functions type solutions

$$\phi(\xi) = \sqrt{-\frac{c_2 m^2}{c_4(2m^2-1)}} cn\left(\sqrt{\frac{c_2}{2m^2-1}}\xi\right), \quad c_0 = \frac{d_2^2 m^2(m^2-1)}{c_4(2m^2-1)^2}, \quad c_2 > 0, \quad (138)$$

$$\phi(\xi) = \sqrt{-\frac{c_2 m^2}{c_4(m^2+1)}} sn\left(\sqrt{-\frac{c_2}{m^2+1}}\xi\right), \quad c_0 = \frac{d_2^2 m^2}{c_4(m^2+1)^2}, \quad c_2 < 0, \quad (139)$$

and

$$\phi(\xi) = \sqrt{-\frac{c_2}{c_4(2-m^2)}} dn\left(\sqrt{\frac{c_2}{2-m^2}}\xi\right), \quad c_0 = \frac{d_2^2(1-m^2)}{c_4(2-m^2)^2}, \quad c_2 > 0. \quad (140)$$

Case 4 If $c_0 = c_1 = c_4 = 0$, Eq. (131) admits two bell shaped solitary wave solutions

$$\phi(\xi) = -\frac{c_2}{c_3} \operatorname{sech}^2\left(\frac{\sqrt{c_2}}{2}\xi\right), \quad c_2 > 0, \quad (141)$$

and

$$\phi(\xi) = -\frac{c_2}{c_3} \operatorname{sech}^2\left(\frac{\sqrt{-c_2}}{2}\xi\right), \quad c_2 < 0, \quad (142)$$

a rational polynomial type solution

$$\phi(\xi) = \frac{1}{c_3\xi^2}, \quad c_2 = 0, \quad (143)$$

Case 5 If $c_3 > 0$, $c_4 = 0$, Eq. (131) admits a Weiertrass elliptic functions type solution

$$\phi(\xi) = \wp\left(\frac{\sqrt{c_3}}{2}\xi, -\frac{4c_1}{c_3}, -\frac{4c_0}{c_3}\right). \quad (144)$$

Let's specifically see how the algebraic method works. For a given nonlinear differential equation, say, in two variables x, t

$$F(u, u_t, u_x, u_{xx}, u_{xt}, u_{tt}, \ldots) = 0 \quad (145)$$

where F is a polynomial function with respect to the indicated variables or some function which can be reduced to a polynomial function by using some transformations.

By using the traveling wave transformation

$$u = u(\xi), \xi = x - \lambda t, \quad (146)$$

Eq. (145) is reduced to an ordinary differential equation with constant coefficients

$$G(U, U', U'', U''', \ldots) = 0 \quad (147)$$

A transformation was presented by Fan [5] in the form

$$u(x) = A_0 + \sum_{i=1}^{n} A_i \phi^{i-1}(\xi), \quad (148)$$

with the new variable $\phi(\xi)$ satisfying Eq. (131), where A_0, A_i, d_j are constants.

Substituting (148) into (147) along with (131), we can determine the parameter n in (148). And then by substituting (148) with the concrete n into (147) and equating the coefficients of these terms $\omega^i \omega'^j (i = 0, 1, 2, \ldots; j = 0, 1)$, we obtain a system of algebraic equations with respect to other parameters A_0, A_i, d_j, λ. By solving the system, if available, we may determine these parameters. Therefore we establish a transformation (147) between (146) and (148). If we know the solutions of (148), then we can obtain the solutions of (147) (or (145)) by using (146).

A New Transformation and a New Theorem

In Yu-Jie Ren's Ph.D Dissertation of Dalian University of Technology [40], in order to develop the above algebraic method, we first presented the following new transformation (150) and new Theorem 4. Then we proved the theorem by means of the Maple computer algebraic system.

Theorem 4 ([40]) *Suppose n and m are any integers, a first-order nonlinear ordinary differential equation with any degree in the form of*

$$\frac{d}{d\xi}\phi(\xi) = \varepsilon\sqrt{c_0 + \sum_{i=1}^{r} c_i \cdot \phi^i(\xi)}\,,$$
$$c_i = consts\,,\quad i = 1, 2, \ldots, r\,. \tag{149}$$

If we substitute the following new transform into Eq. (149)

$$\phi(\xi) = (g(\xi))^{\frac{n}{m}}\,, \tag{150}$$

then Eq. (149) can be transformed into an ordinary differential equation

$$\left(\frac{d}{d\xi}g(\xi)\right)^2 = m^2 n^{-2}\sum_{i=0}^{r} c_i (g(\xi))^{2+\frac{n(i-2)}{m}}\,. \tag{151}$$

We give the proof of Theorem 4 [40] by using Maple as follows:

Proof

Step 1 Importing the following Maple program at the Maple Command Window

$$eq := \mathrm{diff}(phi(xi), xi)^2 - \mathrm{sum}(c[i] * \mathrm{phi}(xi)^i, i = 0..r);$$

Eq. (149) is displayed at a computer screen (after implementing the above program) as follows:

$$eq := \left(\frac{\mathrm{d}}{\mathrm{d}\xi}\phi(\xi)\right)^2 - \sum_{i=0}^{r} c_i (\phi(\xi))^i\,.$$

Step 2 Importing the following Maple program at the Maple Command Window

$$eq := \mathrm{subs}(phi(xi) = (g(xi))^{(}n/m), eq)\,;$$

the following result is displayed at the screen (after running the above program):

$$eq := \left((g(\xi))^{\frac{n}{m}}\right)^2 n^2 \left(\frac{\mathrm{d}}{\mathrm{d}\xi}g(\xi)\right)^2 m^{-2}(g(\xi))^{-2} - \sum_{i=0}^{r} c_i \left((g(\xi))^{\frac{n}{m}}\right)^i\,.$$

Step 3 Importing the following Maple program at the Maple Command Window

$$eq := \mathrm{simplify}(eq)\,;\quad eq := \mathrm{expand}(eq)\,;\quad eq := \mathrm{numer}(eq)\,;$$

the following result is displayed at the screen (after running the above program):

$$eq := \left((g(\xi))^{\frac{n}{m}}\right)^2 n^2 \left(\frac{\mathrm{d}}{\mathrm{d}\xi}g(\xi)\right)^2 - \sum_{i=0}^{r} c_i \left((g(\xi))^{\frac{n}{m}}\right)^i m^2 (g(\xi))^2\,.$$

Step 4 Importing the following Maple program at the Maple Command Window

$$eq := \mathrm{subs}(\mathrm{diff}(g(xi), xi) = G(xi), g(xi) = g, eq)\,;$$
$$eq := eq * (g^{(}n/m))^{(} - 2)\,;\quad eq := \mathrm{simplify}(eq)\,;$$
$$eq := n^2 * G(xi)^2 - (\mathrm{sum}(c[i] * g^{(}(-2 * (n-m)/m) + i * n/m), i = 0..r)) * m^2\,;$$
$$eq := \mathrm{simplify}(eq);$$

the following result is displayed at the screen (after running the above program):

$$eq := n^2 (G(\xi))^2 - g^{-2\frac{n-m}{m}} \sum_{i=0}^{r} c_i \left(g^{\frac{n}{m}}\right)^i m^2$$
$$eq := n^2 (G(\xi))^2 - \sum_{i=0}^{r} c_i g^{\frac{-2n+2m+ni}{m}} m^2\,.$$

Step 5 Importing the following Maple program at the Maple Command Window

$$eq := \mathrm{subs}(g = g(xi), G(xi) = \mathrm{diff}(g(xi), xi), eq)\,;$$

the following result is displayed at the screen (after running the above program):

$$eq := n^2 \left(\frac{\mathrm{d}}{\mathrm{d}\xi}g(\xi)\right)^2 - \sum_{i=0}^{r} c_i (g(\xi))^{\frac{-2n+2m+ni}{m}} m^2\,. \tag{152}$$

We can reduce Eq. (152) to (151).

□

Remark 6 The above transformation (150), Theorem 4, and its proof by means of Maple were first presented by us in Yu-Jie Ren's Ph.D Dissertation of Dalian University of Technology [40].

A New Mechanization Method to Find the Exact Solutions of a First-Order Nonlinear Ordinary Differential Equation with any Degree Using Maple and Its Application

According to Theorem 4 in Subsect. "The the Exp-Bäcklund Transformation Method and Its Application in (1 + 1)-Dimensional KdV Equation", in [40] we first presented the following mechanization method to find the exact solutions of a first-order nonlinear ordinary differential equation with any degree. The validity and reliability of our method are tested by its application to a first-order nonlinear ordinary differential equation with six degrees [40]. Now, we simply describe our mechanization method as follows:

Step 1 Import the following Maple program at the Maple Command Window

$$eq := n^2 * \mathrm{diff}(g(xi), xi)^2 - \mathrm{sum}(c[i] *g(xi)^{(}(-2*n+2*m+n*i)/m), i = 0..r)*m^2;$$

Eq. (151) is displayed at a computer screen (after implementing the above program) as follows

$$eq := n^2 \left(\frac{\mathrm{d}}{\mathrm{d}\xi} g(\xi)\right)^2 - \sum_{i=0}^{r} c_i (g(\xi))^{\frac{-2n+2m+ni}{m}} m^2$$

Step 2 Import the following Maple programs with some values of the degree r and parameter (n, m) in new transform (150) at the Maple Command Window. For example,

$$eq621 := \mathrm{subs}(r = 6, m = 2, n = 1, eq)\,; \qquad eq621 := \mathrm{simplify}(eq621)\,;$$

the following result is displayed at the screen (after running the above program):

$$\begin{aligned} eq621 := {} & \left(\frac{\mathrm{d}}{\mathrm{d}\xi} g(\xi)\right)^2 - 4c_0 g(\xi) \\ & - 4c_1 (g(\xi))^{3/2} - 4c_2 (g(\xi))^2 - 4c_3 (g(\xi))^{5/2} \\ & - 4c_4 (g(\xi))^3 - 4c_5 (g(\xi))^{7/2} - 4c_6 (g(\xi))^4 \,. \end{aligned} \tag{153}$$

Step 3 According to the output results in Step 2, we choose the coefficients of ci, $i = 1, 2, \ldots, m(< n)$ to be zero. Then we import their Maple program. For example, we import the Maple program as follows:

$$eq246 := \mathrm{subs}(c[0] = 0, c[1] = 0, c[3] = 0, c[5] = 0, eq621)\,;$$

the following result is displayed at the screen (after running the above program):

$$\begin{aligned} eq246 := {} & \left(\frac{\mathrm{d}}{\mathrm{d}\xi} g(\xi)\right)^2 \\ & - 4c_2 (g(\xi))^2 - 4c_4 (g(\xi))^3 - 4c_6 (g(\xi))^4 \,. \end{aligned} \tag{154}$$

Step 4 Import the Maple program for solving the output equation in Step 3. For example, we import the Maple program for solving the output Eq. (154) as follows:

$$\mathrm{dsolve}(eq246, g(xi));$$

the following formal solutions of (154) are displayed at the screen (after running the above program):

$$g_1(\xi) = 1/2 \frac{-c_4 - \sqrt{c_4^2 - 4c_6 c_2}}{c_6}\,, \tag{155}$$

$$g_2(\xi) = 1/2 \frac{-c_4 + \sqrt{c_4^2 - 4c_6 c_2}}{c_6}\,, \tag{156}$$

$$g_3(\xi) = 4 \left(\mathrm{e}^{C_1 \sqrt{c_2}}\right)^2 c_2 \left(\mathrm{e}^{\xi \sqrt{c_2}}\right)^{-2} \left(-4c_6 c_2 + \frac{\left(\mathrm{e}^{C_1 \sqrt{c_2}}\right)^4}{\left(\mathrm{e}^{\xi \sqrt{c_2}}\right)^4} - 2\frac{\left(\mathrm{e}^{C_1 \sqrt{c_2}}\right)^2 c_4}{\left(\mathrm{e}^{\xi \sqrt{c_2}}\right)^2} + c_4^2\right)^{-1}, \tag{157}$$

$$g_4(\xi) = -4 \left(\mathrm{e}^{\xi \sqrt{c_2}}\right)^2 c_2 \left(\mathrm{e}^{C_1 \sqrt{c_2}}\right)^{-2} \left(4c_6 c_2 - \frac{\left(\mathrm{e}^{\xi \sqrt{c_2}}\right)^4}{\left(\mathrm{e}^{C_1 \sqrt{c_2}}\right)^4} + 2\frac{\left(\mathrm{e}^{\xi \sqrt{c_2}}\right)^2 c_4}{\left(\mathrm{e}^{C_1 \sqrt{c_2}}\right)^2} - c_4^2\right)^{-1}. \tag{158}$$

Import the Maple program for reducing the above solutions of (154). For example, we import the Maple program for reducing the (157) and (158) as follows:

$$g3 := \mathrm{simplify}(g3); g4 := \mathrm{simplify}(g4)\,;$$

the following results are displayed at a computer screen (after running the above program):

$$g_3(\xi) = 4 \frac{\mathrm{e}^{2\sqrt{c_2}(C_1+\xi)} c_2}{-4c_6 c_2 \mathrm{e}^{4\xi\sqrt{c_2}} + \mathrm{e}^{4C_1\sqrt{c_2}} - 2\mathrm{e}^{2\sqrt{c_2}(C_1+\xi)} c_4 + c_4^2 \mathrm{e}^{4\xi\sqrt{c_2}}}\,, \tag{159}$$

$$g_4(\xi) = 4 \frac{e^{2\sqrt{c_2}(C_1+\xi)} c_2}{-4c_6c_2e^{4C_1\sqrt{c_2}} + e^{4\xi\sqrt{c_2}} - 2e^{2\sqrt{c_2}(C_1+\xi)} c_4 + c_4{}^2 e^{4C_1\sqrt{c_2}}} . \tag{160}$$

Step 5 We discuss the above solutions under different conditions in Step 3.

For example, we discuss the solutions under different conditions $c_4^2 - 4c_6c_2 < 0$ or $c_4^2 - 4c_6c_2 > 0$ or $c_4^2 - 4c_6c_2 = 0$ or $c_2 < 0$ or $c_2 > 0$ or $c_2 = 0$. When

$$c_2 > 0, 4c_6c_2 - c_4^2 < 0 , \tag{161}$$

we may import the following Maple program:

```
assume(c[2] > 0, (4 * c[2] * c[6] − c[4]^2) < 0);
eq246jie := dsolve(eq246, g(xi));
```

six solutions of a first-order nonlinear ordinary differential equation with six degrees under condition (161), which includes two new solutions, are displayed at a computer screen (after running the above program). Here we omit them due to the length of our article.

Importing the following Maple program for reducing the two new solutions above,

```
eq246jie1 := subs(2 * xi * sqrt(c[2])
    − 2 *_C 1 * sqrt(c[2]) = eta, eq246jie1);
eq246jie2 := subs(2 * xi * sqrt(c[2])
    − 2 *_C 1 * sqrt(c[2]) = eta, eq246jie2);
```

the following results are displayed at a computer screen (after running the above program):

$$(\phi_{11}(\xi))^2 = -2\frac{c_2\left(-c_4{}^2 + (\tanh(\eta))^2 c_4{}^2 - \tanh(\eta)\sqrt{c_4{}^2 - 4c_6c_2}\sqrt{c_4{}^2((\tanh(\eta))^2 - 1)}\right)}{\left(4c_6(\tanh(\eta))^2 c_2 - c_4{}^2\right) c_4} , \tag{162}$$

$$(\phi_{12}(\xi))^2 = -2\frac{c_2\left(-c_4{}^2 + (\tanh(\eta))^2 c_4{}^2 + \tanh(\eta)\sqrt{c_4{}^2 - 4c_6c_2}\sqrt{c_4{}^2((\tanh(\eta))^2 - 1)}\right)}{(4c_6(\tanh(\eta))^2 c_2 - c_4{}^2)c_4} . \tag{163}$$

Importing the following Maple program for reducing (162) and (163):

```
eq246jie3 := subs(sqrt(c[4]^2 * (tanh(eta)^2 − 1))
= I * abs(c[4] * sech(eta)), −c[4] + tanh(eta)^2 * c[4]
= − sech(eta)^2 * c[4], eq246jie3);
eq246jie4 := subs(sqrt(c[4]^2 * (tanh(eta)^2 − 1))
= I * abs(c[4] * sech(eta)), −c[4] + tanh(eta)^2 * c[4]
= − sech(eta)^2 * c[4], eq246jie4);
phi[1](xi) := epsilon * (eq246jie3)^(1/2);
phi[2](xi) := epsilon * (eq246jie4)^(1/2);
```

the following results are displayed at a computer screen (after running the above program):

$$\phi_1(\xi) = \varepsilon\sqrt{\frac{2c_2\left(c_4^2\,\mathrm{sech}^2(\eta) + i|c_4\,\mathrm{sech}(\eta)|\tanh(\eta)\sqrt{c_4{}^2 - 4c_6c_2}\right)}{c_4\left(4c_6\tanh^2(\eta)c_2 - c_4{}^2\right)}} , \tag{164}$$

$$\phi_2(\xi) = \varepsilon\sqrt{\frac{2c_2\left(c_4{}^2\,\mathrm{sech}^2(\eta) - i|c_4\,\mathrm{sech}(\eta)|\tanh(\eta)\sqrt{c_4{}^2 - 4c_6c_2}\right)}{c_4\left(4c_6\tanh^2(\eta)c_2 - c_4{}^2\right)}} . \tag{165}$$

where $\eta = 2\sqrt{c_2}(\xi - C_1)$.

Step 6 We need to further discuss the results found by sometimes adding new conditions so that the results are simpler in form. For example, we can import the following Maple program for getting rid of the absolute value sign in the above results:

```
assume(c[4] * sech(eta) < 0);
eq246jie3 := 2 * c[2] * (c[4]^2 * sech(eta)^2 + I
    * tanh(eta) * sqrt(c[4]^2 − 4 * c[6] * c[2])
    * abs(c[4] * sech(eta)))/(4 * c[6] * tanh(eta)^2
    * *c[2] − c[4]^2)/c[4];
eq246jie4 := 2 * c[2] * (c[4]^2 * sech(eta)^2 − I
    * tanh(eta)sqrt(c[4]^2 − 4 * c[6] * c[2]) * abs(c[4]*
    sech(eta)))/(4 * c[6] * tanh(eta)^2
    * c[2] − c[4]^2)/c[4]; phi[1, 1](xi) := epsilon
    * (eq246jie3)^(1/2);
phi[1, 2](xi) := epsilon * (eq246jie4)^(1/2);
```

the following results are displayed at a computer screen (after running the above program):

$$\phi_{1,1}(\xi) = \varepsilon\sqrt{\frac{2c_2\,\mathrm{sech}(\eta)\left(c_4\,\mathrm{sech}(\eta) - i\tanh(\eta)\sqrt{{c_4}^2 - 4c_6c_2}\right)}{4c_6(\tanh(\eta))^2c_2 - {c_4}^2}}, \tag{166}$$

$$\phi_{1,2}(\xi) = \varepsilon\sqrt{\frac{2c_2\,\mathrm{sech}(\eta)\left(c_4\,\mathrm{sech}(\eta) + i\tanh(\eta)\sqrt{{c_4}^2 - 4c_6c_2}\right)}{4c_6(\tanh(\eta))^2c_2 - {c_4}^2}}. \tag{167}$$

Importing the following Maple program for reducing and discussing above results:

```
assume(c[4] * sech(eta) > 0);
eq246jie3 := 2 * c[2] * (c[4]^2 * sech(eta)^2 + I
    * tanh(eta) * sqrt(c[4]^2 - 4 * c[6] * c[2])*
    abs(c[4] * sech(eta)))/(4 * c[6] * tanh(eta)^2
    * c[2] - c[4]^2)/c[4];
eq246jie4 := 2 * c[2] * (c[4]^2 * sech(eta)^2
    - I * tanh(eta) * sqrt(c[4]^2 - 4 * c[6] * c[2])*
    abs(c[4] * sech(eta)))/(4 * c[6] * tanh(eta)^2
    * c[2] - c[4]^2)/c[4];
phi[2, 1](xi) := epsilon * (eq246jie3)^(1/2);
phi[2, 2](xi) := epsilon * (eq246jie4)^(1/2);
```

the following results are shown at a computer screen (after running the above program):

$$\phi_{2,1}(\xi) = \varepsilon\sqrt{\frac{2c_2\,\mathrm{sech}(\eta)\left(c_4\,\mathrm{sech}(\eta) + i\tanh(\eta)\sqrt{{c_4}^2 - 4c_6c_2}\right)}{4c_6(\tanh(\eta))^2c_2 - {c_4}^2}}, \tag{168}$$

$$\phi_{2,2}(\xi) = \varepsilon\sqrt{\frac{2c_2\,\mathrm{sech}(\eta)\left(c_4\,\mathrm{sech}(\eta) - i\tanh(\eta)\sqrt{{c_4}^2 - 4c_6c_2}\right)}{4c_6(\tanh(\eta))^2c_2 - {c_4}^2}}. \tag{169}$$

By using this method, we obtained some new types of general solution of a first-order nonlinear ordinary differential equation with six degrees and presented the following theorem in [40,43,44].

Theorem 5 *The following nonlinear ordinary differential equation with six degrees*

$$\frac{d\phi(\xi)}{d\xi} = \varepsilon\sqrt{\begin{array}{c} c_0 + c_1\phi(\xi) + c_2\phi^2(\xi) + c_3\phi^3(\xi) \\ +c_4\phi^4(\xi) + c_5\phi^5(\xi) + c_6\phi^6(\xi) \end{array}}, \quad c_i = constant, \quad i = 0, 1, 2, 3, 4, 5, 6 \tag{170}$$

admits many kinds of fundamental solutions which depend on the values and constraints between c_i, $i = 0, 1, 2, 3, 4, 5, 6$.

Some of the solutions are listed in the following sections.

Case 1. If $c_0 = c_1 = c_3 = c_5 = 0$, Eq. (170) admits the following constant solutions

$$\phi_{1,2}(\xi) = \varepsilon\sqrt{\frac{-c_4 + \sqrt{{c_4}^2 - 4c_6c_2}}{2c_6}}, \tag{171}$$

$$\phi_{3,4}(\xi) = \varepsilon\sqrt{\frac{-c_4 - \sqrt{{c_4}^2 - 4c_6c_2}}{2c_6}}, \tag{172}$$

and the following exponential type solutions

$$\phi_{5,6}(\xi) = 2\varepsilon\sqrt{\frac{e^{2\sqrt{c_2}(\xi+C_1)}c_2}{\begin{array}{c} -4c_6c_2e^{4C_1\sqrt{c_2}} + e^{4\xi\sqrt{c_2}} \\ -2e^{2\sqrt{c_2}(\xi+C_1)}c_4 + {c_4}^2e^{4C_1\sqrt{c_2}} \end{array}}}, \tag{173}$$

$$\phi_{7,8}(\xi) = 2\varepsilon\sqrt{\frac{e^{2\sqrt{c_2}(\xi+C_1)}c_2}{\begin{array}{c} -4c_6c_2e^{4\xi\sqrt{c_2}} + e^{4C_1\sqrt{c_2}} \\ -2e^{2\sqrt{c_2}(\xi+C_1)}c_4 + {c_4}^2e^{4\xi\sqrt{c_2}} \end{array}}}. \tag{174}$$

When we take different values and constraints of $4c_2c_6 - {c_4}^2$, the solutions (5.43) and (5.44) can be written in different formats as follows:

Case 1.1. If $4c_2c_6 - {c_4}^2 < 0$, Eq. (170) admits the following tanh-sech hyperbolic type solutions

$$\phi_{1,2}(\xi) = \varepsilon\sqrt{\frac{2c_2\,\mathrm{sech}(\eta)\left[c_4\,\mathrm{sech}(\eta) + \varepsilon i\sqrt{{c_4}^2 - 4c_6c_2}\tanh(\eta)\right]}{4c_6c_2\tanh^2(\eta) - {c_4}^2}}, \quad c_2 > 0, \tag{175}$$

$$\phi_{3,4}(\xi) = \varepsilon\sqrt{\frac{2c_2\,\mathrm{sech}(\eta)\Big[c_4\,\mathrm{sech}(\eta) - \varepsilon i\sqrt{{c_4}^2 - 4c_6c_2}\tanh(\eta)\Big]}{4c_6c_2\tanh^2(\eta) - {c_4}^2}}\,, \quad c_2 > 0\,, \tag{176}$$

and the following tan-sec triangular type solutions

$$\phi_{5,6}(\xi) = \varepsilon\sqrt{\frac{2c_2\sec(\zeta)\Big[-c_4\sec(\zeta) + \varepsilon\sqrt{{c_4}^2 - 4c_6c_2}\tan(\zeta)\Big]}{4c_6c_2\tan^2(\zeta) + {c_4}^2}}\,, \quad c_2 < 0\,, \tag{177}$$

$$\phi_{7,8}(\xi) = \varepsilon\sqrt{\frac{2c_2\sec(\zeta)\left[-c_4\sec(\zeta) - \varepsilon\sqrt{{c_4}^2 - 4c_6c_2}\tan(\zeta)\right]}{4c_6c_2\tan^2(\zeta) + {c_4}^2}}\,, \quad c_2 < 0\,, \tag{178}$$

where $\eta = 2\sqrt{c_2}(\xi - C_1)$ and C_1 is any constant, $\zeta = 2\sqrt{-c_2}(\xi - C_1)$ and C_1 is any constant.

Case 1.2. If $4c_2c_6 - {c_4}^2 > 0$, Eq. (170) admits the following sinh-cosh hyperbolic type solutions

$$\phi_{9,10}(\xi) = \varepsilon\sqrt{\frac{2c_2\left[c_4 + \varepsilon\sqrt{4c_2c_6 - {c_4}^2}\sinh(\eta)\right]}{4c_2c_6\sinh^2(\eta) - {c_4}^2\cosh^2(\eta)}}\,, \quad c_2 > 0\,, \tag{179}$$

$$\phi_{11,12}(\xi) = \varepsilon\sqrt{\frac{2c_2\left[c_4 - \varepsilon\sqrt{4c_2c_6 - {c_4}^2}\sinh(\eta)\right]}{4c_2c_6\sinh^2(\eta) - {c_4}^2\cosh^2(\eta)}}\,, \quad c_2 > 0\,, \tag{180}$$

and the following sin-cos triangular type solutions

$$\phi_{13,14}(\xi) = \varepsilon\sqrt{\frac{2c_2\left[-c_4 + \varepsilon i\sqrt{4c_2c_6 - {c_4}^2}\sin(\zeta)\right]}{4c_6c_2\sin^2(\zeta) + {c_4}^2\cos^2(\zeta)}}\,, \quad c_2 < 0 \tag{181}$$

$$\phi_{15,16}(\xi) = \varepsilon\sqrt{\frac{2c_2\left[-c_4 - \varepsilon i\sqrt{4c_6c_2 - {c_4}^2}\sin(\zeta)\right]}{4c_6c_2\sin^2(\zeta) + {}_4{}^2\cos^2(\zeta)}}\,, \quad c_2 < 0\,; \tag{182}$$

where $\eta = 2\sqrt{c_2}(\xi - C_1)$, $\zeta = 2\sqrt{-c_2}(\xi - C_1)$ and C_1 is any constant.

Case 1.3. If $4c_2c_6 - {c_4}^2 = 0$, Eq. (170) admits the following tanh hyperbolic type solutions

$$\phi_{17,18,19,20}(\xi) = \varepsilon\sqrt{\frac{2c_2}{\mathrm{e}^{-\varepsilon 2\sqrt{c_2}(\xi - C_1)} - c_4}} = \varepsilon\sqrt{\frac{2c_2\left(1 + \varepsilon\tanh\left(\sqrt{c_2}(\xi - C_1)\right)\right)}{1 - c_4 - \varepsilon(1 + c_4)\tanh\left(\sqrt{c_2}(\xi - C_1)\right)}}\,, \quad c_2 > 0\,, \tag{183}$$

and the following tan triangular type solutions

$$\phi_{21,22,23,24}(\xi) = \varepsilon\sqrt{\frac{2c_2}{\mathrm{e}^{-\varepsilon 2i\sqrt{-c_2}(\xi - C_1)} - c_4}} = \varepsilon\sqrt{\frac{2c_2\left(1 + \varepsilon\tan\left(i\sqrt{-c_2}(\xi - C_1)\right)\right)}{1 - c_4 - \varepsilon(1 + c_4)\tan\left(i\sqrt{-c_2}(\xi - C_1)\right)}}\,, \quad c_2 < 0 \tag{184}$$

where C_1 is any constant.

Case 2. If $c_5 = c_6 = 0$, Eq. (170) admits the solutions in Theorem 6.1.

Remark 7 By using our mechanization method using Maple to find the exact solutions of a first-order nonlinear ordinary differential equation with any degree, we can obtain some new types of general solution of a first-order nonlinear ordinary differential equation with $r(= 7, 8, 9, 10, 11, 12, \ldots)$ degree [40]. We do not list the solutions here in order to avoid unnecessary repetition.

Summary of the Generalized Algebra Method

In this section, based on a first-order nonlinear ordinary differential equation with any degree (149) and its the exact solutions obtained by using our mechanization method via Maple, we will develop the algebraic methods [50,51] for constructing the traveling wave solutions and present a new generalized algebraic method and its algorithms [40, 43,44]. The KdV equation is chosen to illustrate our algorithm so that more families of new exact solutions are obtained which contain both non-traveling solutions and traveling wave solutions. We outline the main steps of our generalized algebraic method as follows:

Step 1. For given nonlinear differential equations, with some physical fields $u_i(t, x_1, x_2, \ldots, x_m)$, $(i = 1, 2, \ldots, n)$

in $m+1$ independent variables $t, x_1, x_2, \ldots, x_m$,

$$F_j(u_1, \ldots, u_n, u_{1,t}, \ldots, u_{n,t}, u_{1,x_1}, \ldots, u_{n,x_m}, u_{1,tt}, \ldots, u_{n,tt}, u_{1,tx_1}, \ldots u_{n,tx_m}, \ldots) = 0\,, \quad j = 1, 2, \ldots, n\,, \tag{185}$$

where $j = 1, 2, \ldots, n$.

By using the following more general transformation, which we first present here,

$$\begin{aligned} &u_i(t, x_1, x_2, \ldots, x_m) = U_i(\xi)\,, \\ &\xi = u_i(t, x_1, x_2, \ldots, x_m) = U_i(\xi)\,, \\ &\xi = \alpha_0(t) + \sum_{i=1}^{m-1} \alpha_i(x_i, x_{i+1}, \ldots, x_m, t) \cdot \beta_i(x_i)\,, \end{aligned} \tag{186}$$

where $\alpha_0(t), \alpha_i(x_i, x_{i+1}, \ldots, x_m, t)$ and $\beta_i(x_i)$, $i = 1, 2, \ldots, m-1$ are functions to be determined later. For example, when $n = 1$, we may take

$$\xi = \alpha_0(t) + \alpha_1(x_1 t) \cdot \beta_1(x_1)\,,$$

where $\alpha_0(t), \alpha_1(x_1 t)$ and $\beta_1(x_1)$ are undetermined functions.

Then Eq. (185) is reduced to nonlinear differential equations

$$G_j(U_1, \ldots, U_n, U_1', \ldots, U_n', U_1'', \ldots, U_n'', \ldots) = 0\,, \quad j = 1, 2, \ldots, n\,, \tag{187}$$

where $G_j (j = 1, 2, \ldots, n)$ are all polynomials of $U_i (i = 1, 2, \ldots, n)$, $\alpha_0(t), \alpha_i(x_i, x_{i+1}, \ldots, x_m, t)$, $\beta_i(x_i)$, $i = 1, 2, \ldots, m-1$ and their derivatives. If G_k of them is not a polynomial of $U_i (i = 1, 2, \ldots, n)$, $\alpha_i(x_i, x_{i+1}, \ldots, x_m, t)$, $\beta_i(x_i)$, $i = 1, 2, \ldots, m-1$, $\alpha_0(t)$ and their derivatives, then we may use new variable $v_i(\xi) (i = 1, 2, \ldots, n)$ which makes G_k become a polynomial of $v_i(\xi) (i = 1, 2, \ldots, n)$, $\alpha_0(t), \alpha_i(x_i, x_{i+1}, \ldots, x_m, t)$ and $\beta_i(x_i)$, $i = 1, 2, \ldots, m-1$ and their derivatives. Otherwise the following transformation will fail to seek solutions of Eq. (185).

Step 2. We introduce a new variable $\phi(\xi)$ which is a solution of the following ODE

$$\frac{d\phi(\xi)}{d\xi} = \varepsilon \sqrt{\begin{array}{c} c_0 + c_1\phi(\xi) + c_2\phi^2(\xi) + c_3\phi^3(\xi) \\ +c_4\phi^4(\xi) + \cdots + c_r\phi^r(\xi) \end{array}}\,, \quad r = 0, 1, 2, 3, \ldots\,. \tag{188}$$

Then the derivatives with respect to the variable ξ become the derivatives with respect to the variable ϕ.

Step 3. By using the new variable ϕ, we expand the solution of Eq. (185) in the form:

$$U_i = a_{i,0}(X) + \sum_{k=1}^{n_i} (a_{i,k}(X)\phi^k(\xi(X)) + b_{i,k}(X)\phi^{-k}(\xi(X)))\,. \tag{189}$$

where $X = (x_1, x_2, \ldots, x_m, t)$, $\xi = \xi(X)$, $a_{i,0}(X)$, $a_{i,k}(X), b_{i,k}(X) (i = 1, 2, \ldots, n; k = 1, 2, \ldots, n_i)$ are all differentiable functions of X to be determined later.

Step 4. In order to determine $n_i (i = 1, 2, \ldots, n)$ and r, we may substitute (188) into (187) and balance the highest derivative term with the nonlinear terms in Eq. (187). By using the derivatives with respect to the variable ϕ, we can obtain a relation for n_i and r, from which the different possible values of n_i and r can be obtained. These values lead to the series expansions of the solutions for Eq. (185).

Step 5. Substituting (189) into the given Eq. (185) and collecting coefficients of polynomials of ϕ^k, ϕ^{-k}, and $\phi^i\phi^{-k}\sqrt{\sum_{j=1}^{r} c_j\phi^j(\xi)}$, with the aid of Maple, then setting each coefficient to zero, we will get a system of over-determined partial differential equations with respect to $a_{i,0}(X), a_{i,k}(X), b_{i,k}(X)$, $i = 1, 2, \ldots, n; k = 1, 2, \ldots, n_i c_j$, $j = 0, 1, \ldots, r\alpha_0(t), \alpha_i(x_i, x_{i+1}, \ldots, x_m, t)$ and $\beta_i(x_i)$, $i = 1, 2, \ldots, m-1$.

Step 6. Solving the over-determined partial differential equations with Maple,then we can determine $a_{i,0}(X)$, $a_{i,k}(X), b_{i,k}(X)$, $i = 1, 2, \ldots, n$; $k = 1, 2, \ldots, n_i)c_j$, $j = 0, 1, \ldots, r\alpha_0(t), \alpha_i(x_i, x_{i+1}, \ldots, x_m, t)$ and $\beta_i(x_i)$, $i = 1, 2, \ldots, m-1$.

Step 7. From the constants $a_{i,0}(X), a_{i,k}(X), b_{i,k}(X) (i = 1, 2, \ldots, n; k = 1, 2, \ldots, n_i)c_j (j = 0, 1, \ldots, r)$, $\alpha_0(t)$, $\alpha_i(x_i, x_{i+1}, \ldots, x_m, t)$ and $\beta_i(x_i)$, $i = 1, 2, \ldots, m-1$ obtained in Step 6 to Eq. (188), we can then obtain all the possible solutions.

Remark 8 When $c_5 = c_6 = 0$ and $b_{i,k} = 0$, Eq. (170) and transformation (189) just become the ones used in our previous method [50,51]. However, if $c_5 \neq 0$ or $c_6 \neq 0$, we may obtain solutions that cannot be found by using the methods [50,51]. It should be pointed out that there is no method to find all solutions of nonlinear PDEs. But our method can be used to find more solutions of nonlinear PDEs, and with the exact solutions obtained by using our mechanization method via Maple, we will develop the algebraic methods [50,51] for constructing the travel-

ing wave solutions and present a new generalized algebraic method and its algorithms [40,43,44].

Remark 9 By the above description, we find that our method is more general than the method in [50,51]. We have improved the method [50,51] in five aspects: First, we extend the ODE with four degrees (131) into the ODE with any degree (188) and get its new general solutions by using our mechanization method via Maple [40,43,44]. Second, we change the solution of Eq. (185) into a more general solution (189) and get more types of new rational solutions and irrational solutions. Third, we replace the traveling wave transformation (146) in [50,51] by a more general transformation (186). Fourth, we suppose the coefficients of the transformation (186) and (189) are undetermined functions, but the coefficients of the transformation (146) in [50,51] are all constants. Fifth, we present a more general algebra method than the method given in [50,51], which is called the generalized algebra method, to find more types of exact solutions of nonlinear differential equations based upon the solutions of the ODE (188). This can obtain more general solutions of the NPDEs than the number obtained by the method in [50,51].

The Generalized Algebra Method to Find New Non-traveling Waves Solutions of the (1 + 1)-Dimensional Generalized Variable-Coefficient KdV Equation

In this section, we will make use of our generalized algebra method and symbolic computation to find new non-traveling waves solutions and traveling waves solutions of the following (1 + 1) – dimensional Generalized Variable – Coefficient KdV equation [16].

Propagation of weakly nonlinear long waves in an inhomogeneous waveguide is governed by a variable – coefficient KdV equation of the form [15]

$$u_t(x,t) + 6u(x,t)u_x(x,t) + B(t)u_{xxx}(x,t) = 0\,. \quad (190)$$

where $u(x,t)$ is the wave amplitude, t the propagation coordinate, x the temporal variable and $B(t)$ is the local dispersion coefficient.

The applicability of the variable-coefficient KdV equation (190) arises in many areas of physics as, for example, for the description of the propagation of gravity-capillary and interfacial-capillary waves, internal waves and Rossby waves [15]. In order to study the propagation of weakly nonlinear, weakly dispersive waves in inhomogeneous media, Eq. (190) is rewritten as follows [16]

$$u_t(x,t)+6A(t)u(x,t)u_x(x,t)+B(t)u_{xxx}(x,t) = 0. \quad (191)$$

which now has a variable nonlinearity coefficient $A(t)$. Here, Eq. (191) is called a (1 + 1)-dimensional generalized variable-coefficient KdV (gvcKdV) equation.

In order to find new non-traveling waves solutions and traveling waves solutions of the following (1 + 1)-dimensional gvcKdV equation (191) by using our generalized algebra method and symbolic computation, we first take the following new general transformation, which we first present here

$$u(x,t) = u(\xi)\,, \quad \xi = \alpha(x,t)\beta(t) - r(t)\,, \quad (192)$$

where $\alpha(x,t)$, $\beta(t)$ and $r(t)$ are functions to be determined later.

By using the new variable $\phi = \phi(\xi)$ which is a solution of the following ODE

$$\frac{\mathrm{d}\phi(\xi)}{\mathrm{d}\xi} = \varepsilon\sqrt{\begin{array}{c} c_0 + c_1\phi(\xi) + c_2\phi^2(\xi) + c_3\phi^3(\xi) \\ +c_4\phi^4(\xi) + c_5\phi^5(\xi) + c_6\phi^6(\xi) \end{array}}\,. \quad (193)$$

we expand the solution of Eq. (191) in the form [40,43]:

$$u = a_0(X)+\sum_{i=1}^{n}(a_i(X)\phi^i(\xi(X))+b_i(X)\phi^{-i}(\xi(X))) \quad (194)$$

where $X = (x,t)$, $a_0(X), a_i(X), b_i(X)$ $(i = 1,2,\ldots,n)$ are all differentiable functions of X to be determined later.

Balancing the highest derivative term with the nonlinear terms in Eq. (191) by using the derivatives with respect to the variable ϕ, we can determine the parameter $n = 2$ in (194). In addition, we take $a_0(X) = a_0(t)$, $a_i(X) = a_i(t), b_i(X) = b_i(t), i = 1,2$ in (194) for simplicity, then substituting them and (192) into (194) along with $n = 2$, leads to:

$$u(x,t) = a_0(t) + a_1(t)\phi(\xi) + a_2(t)(\phi(\xi))^2 + \frac{b_1(t)}{\phi(\xi)} + \frac{b_2(t)}{(\phi(\xi))^2}\,. \quad (195)$$

where $\xi = \alpha(x,t)\beta(t) - r(t)$, and $\alpha(x,t)$, $\beta(t)$, $r(t)$, $a_0(t)$, $a_i(t)$ and $b_i(t)$, $i = 1,2$ are all differentiable functions of x or t to be determined later.

By substituting (195) into the given Eq. (191) along with (193) and the derivatives of ϕ, and collecting coefficients of polynomials of ϕ^k and $\phi^\tau\sqrt{\sum_{j=1}^{6} c_j\phi^j(\xi)}$, with the aid of Maple, then setting each coefficient to zero, we will get a system of over-determined partial differential equations with respect to $A(t), B(t), \alpha(x,t), \beta(t), r(t)$,

$a_0(t)$, $a_i(t)$ and $b_i(t)$, $i = 1, 2$ as follows:

$$
\begin{aligned}
&21B(t)(\beta(t))^3 - \varepsilon a_1(t)\left(\frac{\partial}{\partial x}\alpha(x,t)\right)^3 c_4 c_6 + 18A(t) \\
&- \varepsilon\left(\frac{\partial}{\partial x}\alpha(x,t)\right)\beta(t)a_1(t)a_2(t)c_6 - 3B(t)(\beta(t))^3 \\
&- \varepsilon b_1(t)\left(\frac{\partial}{\partial x}\alpha(x,t)\right)^3 {c_6}^2 = 0\,, \\
&12b_2(t) - \varepsilon\left(\frac{\mathrm{d}}{\mathrm{d}t}r(t)\right)c_2 - 8B(t)(\beta(t))^3 \\
&- \varepsilon b_2(t)\left(\frac{\partial}{\partial x}\alpha(x,t)\right)^3 {c_2}^2 - 2b_2(t) \\
&- \varepsilon\left(\frac{\partial}{\partial t}\alpha(x,t)\right)\beta(t)c_2 - 2B(t)\beta(t) \\
&- \varepsilon b_2(t)\left(\frac{\partial^3}{\partial x^3}\alpha(x,t)\right)c_2 - 2b_2(t) \\
&- \varepsilon\alpha(x,t)\left(\frac{\mathrm{d}}{\mathrm{d}t}\beta(t)\right)c_2 - 6A(t) \\
&- \varepsilon\left(\frac{\partial}{\partial x}\alpha(x,t)\right)\beta(t)(b_1(t))^2 c_2 \\
&- 12A(t) - \varepsilon\left(\frac{\partial}{\partial x}\alpha(x,t)\right)\beta(t)a_0(t)b_2(t)c_2 - 12A(t) \\
&- \varepsilon\left(\frac{\partial}{\partial x}\alpha(x,t)\right)\beta(t)(b_2(t))^2 c_4 = 0\,, \\
&b_1(t) - \varepsilon\left(\frac{\mathrm{d}}{\mathrm{d}t}r(t)\right)c_6 - a_1(t) - \varepsilon\left(\frac{\mathrm{d}}{\mathrm{d}t}r(t)\right)c_4 \\
&+ 6A(t) - \varepsilon\left(\frac{\partial}{\partial x}\alpha(x,t)\right)\beta(t)a_2(t)b_1(t)c_4 \\
&+ a_1(t) - \varepsilon\alpha(x,t)\left(\frac{\mathrm{d}}{\mathrm{d}t}\beta(t)\right)c_4 - b_1(t) \\
&- \varepsilon\alpha(x,t)\left(\frac{\mathrm{d}}{\mathrm{d}t}\beta(t)\right)c_6 + 6A(t) \\
&- \varepsilon\left(\frac{\partial}{\partial x}\alpha(x,t)\right)\beta(t)a_0(t)a_1(t)c_4 - 6A(t) \\
&- \varepsilon\left(\frac{\partial}{\partial x}\alpha(x,t)\right)\beta(t)a_0(t)b_1(t)c_6 \\
&- 4B(t)(\beta(t))^3 - \varepsilon b_1(t)\left(\frac{\partial}{\partial x}\alpha(x,t)\right)^3 c_2 c_6 \\
&+ B(t)\beta(t) - \varepsilon a_1(t)\left(\frac{\partial^3}{\partial x^3}\alpha(x,t)\right)c_4 \\
&+ a_1(t) - \varepsilon\left(\frac{\partial}{\partial t}\alpha(x,t)\right)\beta(t)c_4 - b_1(t) \\
&- \varepsilon\left(\frac{\partial}{\partial t}\alpha(x,t)\right)\beta(t)c_6 - 6A(t) \\
&- \varepsilon\left(\frac{\partial}{\partial x}\alpha(x,t)\right)\beta(t)a_1(t)b_2(t)c_6 + 7B(t)(\beta(t))^3 \\
&- \varepsilon a_1(t)\left(\frac{\partial}{\partial x}\alpha(x,t)\right)^3 c_2 c_4 + 18A(t) \\
&- \varepsilon\left(\frac{\partial}{\partial x}\alpha(x,t)\right)\beta(t)a_1(t)a_2(t)c_2 - B(t)\beta(t) \\
&- \varepsilon b_1(t)\left(\frac{\partial^3}{\partial x^3}\alpha(x,t)\right)c_6 = 0\,, \qquad (196) \\
&\cdots\cdots\cdots\cdots
\end{aligned}
$$

because there are so many over-determined partial differential equations, only a few of them are shown here for convenience. Solving the over-determined partial differential equations with Maple, we have the following solutions.

Case 1

$$
\begin{aligned}
&A(t) = A(t)\,, \quad B(t) = B(t)\,, \quad \alpha(x,t) = F_1(t)\,, \\
&\beta(t) = \beta(t)\,, \quad a_2(t) = C_1\,, \quad a_1(t) = C_2\,, \\
&b_2(t) = C_3\,, a_0(t) = C_5\,, \\
&r(t) = \int F_1(t)\frac{\mathrm{d}}{\mathrm{d}t}\beta(t) + \left(\frac{\mathrm{d}}{\mathrm{d}t}F_1(t)\right)\beta(t)\mathrm{d}t + C_6\,, \\
&b_1(t) = C_4\,,
\end{aligned} \qquad (197)
$$

where $A(t)$, $B(t)$, $\beta(t)$, $F_1(t)$ are arbitrary functions of t, and C_i, $i = 1, 2, 3, 4, 5$ are arbitrary constants.

Case 2

$$
\begin{aligned}
&b_1(t) = 0\,, \quad A(t) = A(t)\,, \quad B(t) = B(t)\,, \\
&\alpha(x,t) = F_1(t)\,, \quad \beta(t) = \beta(t)\,, \quad a_2(t) = C_1\,, \\
&a_0(t) = C_4\,, \quad a_1(t) = C_2\,, \\
&r(t) = \int F_1(t)\frac{\mathrm{d}}{\mathrm{d}t}\beta(t) + \left(\frac{\mathrm{d}}{\mathrm{d}t}F_1(t)\right)\beta(t)\mathrm{d}t + C_5\,, \\
&b_2(t) = C_3\,,
\end{aligned} \qquad (198)
$$

where $A(t)$, $B(t)$, $\beta(t)$, $F_1(t)$ are arbitrary functions of t, and C_i, $i = 1, 2, 3, 4$ are arbitrary constants.

Case 3

$$
\begin{aligned}
&\beta(t) = 0\,, \quad A(t) = A(t)\,, \quad B(t) = B(t)\,, \\
&a_2(t) = C_1\,, \quad a_1(t) = C_2\,, \quad b_2(t) = C_3\,, \\
&a_0(t) = C_5\,, \quad b_1(t) = C_4\,, \quad r(t) = C_6\,, \\
&\alpha(x,t) = \alpha(x,t)\,,
\end{aligned} \qquad (199)
$$

where $\alpha(x, t)$ are arbitrary functions of x and t, $A(t)$, $B(t)$ are arbitrary functions of t, and C_i, $i = 1, 2, 3, 4, 5$ are arbitrary constants.

Case 4

$$\begin{aligned}
&b_2(t) = 0\,, \quad A(t) = A(t)\,, \quad B(t) = B(t)\,,\\
&\alpha(x,t) = F_1(t)\,, \quad \beta(t) = \beta(t)\,, \quad a_2(t) = C_1\,,\\
&a_0(t) = C_4\,, a_1(t) = C_2\,,\\
&r(t) = \int F_1(t)\frac{\mathrm{d}}{\mathrm{d}t}\beta(t) + \left(\frac{\mathrm{d}}{\mathrm{d}t}F_1(t)\right)\beta(t)\mathrm{d}t + C_5\,,\\
&b_1(t) = C_3\,,
\end{aligned} \tag{200}$$

where $A(t)$, $B(t)$, $F_1(t)$, $\beta(t)$ are arbitrary functions of t, and C_i, $i = 1, 2, 3, 4, 5$ are arbitrary constants.

Case 5

$$\begin{aligned}
&b_2(t) = 0\,, \quad \beta(t) = 0\,, \quad A(t) = A(t)\,,\\
&B(t) = B(t)\,, \quad a_2(t) = C_1\,, \quad a_0(t) = C_4\,,\\
&a_1(t) = C_2\,, \quad r(t) = C_5\,, \quad b_1(t) = C_3\,,\\
&\alpha(x,t) = \alpha(x,t)\,,
\end{aligned} \tag{201}$$

where $\alpha(x, t)$ are arbitrary functions of x and t, $A(t)$, $B(t)$ are arbitrary functions of t, and C_i, $i = 1, 2, 3, 4, 5$ are arbitrary constants.

Case 6

$$\begin{aligned}
&b_1(t) = \frac{C_2 c_4}{c_6}\,, \quad \beta(t) = 0\,, \quad b_2(t) = 0\,,\\
&A(t) = A(t)\,, \quad B(t) = B(t)\,, \quad a_2(t) = C_1\,,\\
&a_1(t) = C_2\,, \quad r(t) = C_4\,, \quad a_0(t) = C_3\,,\\
&\alpha(x,t) = \alpha(x,t)\,,
\end{aligned} \tag{202}$$

where $\alpha(x, t)$ are arbitrary functions of x and t, $A(t)$, $B(t)$ are arbitrary functions of t, and C_i, $i = 1, 2, 3, 4$ are arbitrary constants.

Case 7

$$\begin{aligned}
&B(t) = B(t)\,, \quad b_1(t) = 0\,, \quad a_0(t) = C_3\,,\\
&A(t) = 0\,, \quad b_2(t) = C_2\,, \quad a_1(t) = 0\,, \quad a_2(t) = 0\,,\\
&\alpha(x,t) = F_1(t)x + F_2(t)\,, \quad \beta(t) = \frac{C_1}{F_1(t)}\,,\\
&r(t) = \int \frac{C_1, \left(F_1(t)\frac{\mathrm{d}}{\mathrm{d}t}F_2(t) - \left(\frac{\mathrm{d}}{\mathrm{d}t}F_1(t)\right) F_2(t) + 4c_2 B(t) {C_1}^2 (F_1(t))^2\right)}{(F_1(t))^2}\mathrm{d}t + C_4\,,
\end{aligned} \tag{203}$$

where $F_1(t)$, $B(t)$ are arbitrary functions of t, and C_i, $i = 1, 2, 3, 4$ are arbitrary constants.

Case 8

$$\begin{aligned}
&B(t) = B(t)\,, \quad b_1(t) = 0\,, \quad \alpha(x,t) = F_1(t)\,,\\
&\beta(t) = \beta(t)\,, \quad A(t) = 0\,, \quad a_1(t) = 0\,,\\
&a_2(t) = 0\,,\\
&r(t) = \int F_1(t)\frac{\mathrm{d}}{\mathrm{d}t}\beta(t) + \left(\frac{\mathrm{d}}{\mathrm{d}t}F_1(t)\right)\beta(t)\mathrm{d}t + C_3\,,\\
&a_0(t) = C_2\,, \quad b_2(t) = C_1\,,
\end{aligned} \tag{204}$$

where $A(t)$, $B(t)$, $\beta(t)$, $F_1(t)$ are arbitrary functions of t, and C_i, $i = 1, 2, 3$ are arbitrary constants.

Case 9

$$\begin{aligned}
&A(t) = A(t)\,, \quad B(t) = B(t)\,, \quad b_1(t) = 0\,,\\
&\alpha(x,t) = F_1(t)\,, \quad \beta(t) = \beta(t)\,,\\
&a_2(t) = C_1\,, \quad a_0(t) = C_3\,, \quad b_2(t) = C_2\,,\\
&a_1(t) = 0\,,\\
&r(t) = \int F_1(t)\frac{\mathrm{d}}{\mathrm{d}t}\beta(t) + \left(\frac{\mathrm{d}}{\mathrm{d}t}F_1(t)\right)\beta(t)\mathrm{d}t + C_4\,,
\end{aligned} \tag{205}$$

where $A(t)$, $B(t)$, $\beta(t)$, $F_1(t)$ are arbitrary functions of t, and C_i, $i = 1, 2, 3, 4$ are arbitrary constants.

Case 10

$$\begin{aligned}
&A(t) = A(t)\,, \quad B(t) = B(t)\,, \quad b_1(t) = 0\,,\\
&\alpha(x,t) = F_1(t)\,, \quad \beta(t) = \beta(t)\,,\\
&a_2(t) = C_1\,, \quad b_2(t) = 0\,, \quad a_1(t) = 0\,,\\
&r(t) = \int F_1(t)\frac{\mathrm{d}}{\mathrm{d}t}\beta(t) + \left(\frac{\mathrm{d}}{\mathrm{d}t}F_1(t)\right)\beta(t)\mathrm{d}t + C_3\,,\\
&a_0(t) = C_2\,,
\end{aligned} \tag{206}$$

where $A(t)$, $B(t)$, $\beta(t)$, $F_1(t)$ are arbitrary functions of t, and C_i, $i = 1, 2, 3, 4$ are arbitrary constants.

Case 11

$$\begin{aligned}
&A(t) = A(t)\,, \quad B(t) = B(t)\,, \quad b_1(t) = 0\,,\\
&a_2(t) = C_1\,, \quad b_2(t) = 0\,, \quad a_1(t) = 0\,,\\
&\beta(t) = 0\,, \quad r(t) = C_3\,,\\
&\alpha(x,t) = \alpha(x,t)\,, \quad a_0(t) = C_2\,,
\end{aligned} \tag{207}$$

where $\alpha(x, t)$ are arbitrary functions of x and t, $A(t)$, $B(t)$ are arbitrary functions of t, and C_i, $i = 1, 2$ are arbitrary constants. Because there are so many solutions, only a few of them are shown here for convenience.

So we get new general forms of solutions of equations (191):

$$u(x,t) = a_0(t) + a_1(t)\phi(\alpha(x,t)\beta(t) - r(t)) + a_2(t)(\phi(\alpha(x,t)\beta(t) - r(t)))^2 + \frac{b_1(t)}{\phi(\alpha(x,t)\beta(t) - r(t))} + \frac{b_2(t)}{(\phi(\alpha(x,t)\beta(t) - r(t)))^2}. \tag{208}$$

where $\alpha(x,t)$, $\beta(t)$, $r(t)$, $a_0(t)$, $a_i(t)$, $b_i(t)$, $i = 1,2$ satisfy (197)–(207) respectively, and the variable $\phi(\alpha(x,t)\beta(t) - r(t))$ takes the solutions of the Eq. (193). For example, we may take the variable $\phi(\alpha(x,t)\beta(t) - r(t))$ as follows:

Type 1. If $4c_2c_6 - c_4{}^2 < 0$, corresponding to Eq. (193), $\phi(\alpha(x,t)\beta(t) - r(t))$ is taken, we get the following four tanh-sech hyperbolic type solutions

$$\phi_{1,2}((\alpha(x,t)\beta(t) - r(t))) = \varepsilon\sqrt{\frac{2c_2\,\mathrm{sech}(\eta)[c_4\,\mathrm{sech}(\eta) + \varepsilon i\sqrt{c_4{}^2 - 4c_6c_2}\tanh(\eta)]}{4c_6c_2\tanh^2(\eta) - c_4{}^2}}, \quad c_2 > 0, \tag{209}$$

$$\phi_{3,4}((\alpha(x,t)\beta(t) - r(t))) = \varepsilon\sqrt{\frac{2c_2\,\mathrm{sech}(\eta)[c_4\,\mathrm{sech}(\eta) - \varepsilon i\sqrt{c_4{}^2 - 4c_6c_2}\tanh(\eta)]}{4c_6c_2\tanh^2(\eta) - c_4{}^2}}, \quad c_2 > 0, \tag{210}$$

and the following tan-sec triangular type solutions

$$\phi_{5,6}((\alpha(x,t)\beta(t) - r(t))) = \varepsilon\sqrt{\frac{2c_2\sec(\zeta)[-c_4\sec(\zeta) + \varepsilon\sqrt{c_4{}^2 - 4c_6c_2}\tan(\zeta)]}{4c_6c_2\tan^2(\zeta) + c_4{}^2}}, \quad c_2 < 0, \tag{211}$$

$$\phi_{7,8}((\alpha(x,t)\beta(t) - r(t))) = \varepsilon\sqrt{\frac{2c_2\sec(\zeta)[-c_4\sec(\zeta) - \varepsilon\sqrt{c_4{}^2 - 4c_6c_2}\tan(\zeta)]}{4c_6c_2\tan^2(\zeta) + c_4{}^2}}, \quad c_2 < 0, \tag{212}$$

where $\eta = 2\sqrt{c_2}((\alpha(x,t)\beta(t) - r(t)) - C_1)$ and C_1 is any constant, $\zeta = 2\sqrt{-c_2}((\alpha(x,t)\beta(t) - r(t)) - C_1)$ and C_1 is any constant.

Type 2. If $4c_2c_6 - c_4{}^2 > 0$, corresponding to Eq. (193), $\phi(\alpha(x,t)\beta(t) - r(t))$ is taken, we get the following four sinh-cosh hyperbolic type solutions

$$\phi_{9,10}((\alpha(x,t)\beta(t) - r(t))) = \varepsilon\sqrt{\frac{2c_2[c_4 + \varepsilon\sqrt{4c_2c_6 - c_4{}^2}\sinh(\eta)]}{4c_2c_6\sinh^2(\eta) - c_4{}^2\cosh^2(\eta)}}, \quad c_2 > 0, \tag{213}$$

$$\phi_{11,12}((\alpha(x,t)\beta(t) - r(t))) = \varepsilon\sqrt{\frac{2c_2[c_4 - \varepsilon\sqrt{4c_2c_6 - c_4{}^2}\sinh(\eta)]}{4c_2c_6\sinh^2(\eta) - c_4{}^2\cosh^2(\eta)}}, \quad c_2 > 0, \tag{214}$$

and the following sin-cos triangular type solutions

$$\phi_{13,14}((\alpha(x,t)\beta(t) - r(t))) = \varepsilon\sqrt{\frac{2c_2[-c_4 + \varepsilon i\sqrt{4c_2c_6 - c_4{}^2}\sin(\zeta)]}{4c_6c_2\sin^2(\zeta) + c_4{}^2\cos^2(\zeta)}}, \quad c_2 < 0 \tag{215}$$

$$\phi_{15,16}((\alpha(x,t)\beta(t) - r(t))) = \varepsilon\sqrt{\frac{2c_2[-c_4 - \varepsilon i\sqrt{4c_6c_2 - c_4{}^2}\sin(\zeta)]}{4c_6c_2\sin^2(\zeta) + {}_4{}^2\cos^2(\zeta)}}, \quad c_2 < 0, \tag{216}$$

where $\eta = 2\sqrt{c_2}((\alpha(x,t)\beta(t) - r(t)) - C_1)$, $\zeta = 2\sqrt{-c_2}((\alpha(x,t)\beta(t) - r(t)) - C_1)$,and C_1 is any constant.

Type 3. If $4c_2c_6 - c_4{}^2 = 0$, corresponding to Eq. (193), $\phi(\alpha(x,t)\beta(t) - r(t))$ is taken, we get the following two sech-tanh hyperbolic type solutions $4c_2c_6 - c_4{}^2 = 0$, Eq. (170) admits the following tanh hyperbolic type solutions

$$\phi_{17,18,19,20}(\alpha(x,t)\beta(t) - r(t)) = \varepsilon\sqrt{\frac{2c_2\{1 + \varepsilon\tanh[\sqrt{c_2}(\alpha(x,t)\beta(t) - r(t) - C_1)]\}}{1 - c_4\varepsilon(1 + c_4)\tanh[\sqrt{c_2}(\alpha(x,t)\beta(t) - r(t) - C_1)]}}, \quad c_2 > 0, \tag{217}$$

and the following tan triangular type solutions

$$\phi_{21,22,23,24}(\alpha(x,t)\beta(t) - r(t)) = \varepsilon\sqrt{\frac{2c_2\{1 + \varepsilon\tan[i\sqrt{-c_2}(\alpha(x,t)\beta(t) - r(t) - C_1)]\}}{1 - c_4 - \varepsilon(1 + c_4)\tan[i\sqrt{-c_2}(\alpha(x,t)\beta(t) - r(t) - C_1)]}}, \quad c_2 < 0, \tag{218}$$

where C_1 is any constant.

Substituting (209)–(218) and (197)–(207) into (208), respectively, we get many new irrational solutions and rational solutions of combined hyperbolic type or triangular type solutions of Eq. (191).

For example, when we select $A(t), B(t), \alpha(x,t), \beta(t), r(t), a_0(t), a_i(t)$ and $b_i(t), i = 1, 2$ to satisfy Case 1, we can easily get the following irrational solutions of combined hyperbolic type or triangular type solutions of Eq. (191):

$$\begin{aligned} u(x,t) &= C_5 + C_2\phi(F_1(t)\beta(t) - r(t)) \\ &+ C_1(\phi(F_1(t)\beta(t) - r(t)))^2 \\ &+ \frac{C_4}{\phi(F_1(t)\beta(t) - r(t))} + \frac{C_3}{(\phi(F_1(t)\beta(t) - r(t)))^2}, \end{aligned} \tag{219}$$

where $A(t), B(t), \beta(t), F_1(t)$ are arbitrary functions of t, C_i, $i = 1, 2, 3, 4, 5$ are arbitrary constants, and

$$r(t) = \int F_1(t)\frac{\mathrm{d}}{\mathrm{d}t}\beta(t) + \left(\frac{\mathrm{d}}{\mathrm{d}t}F_1(t)\right)\beta(t)\mathrm{d}t + C_6 .$$

Substituting (209)–(218) into (219), respectively, we get the following new irrational solutions and rational solutions of combined hyperbolic type or triangular type solutions of Eq. (191).

Case 1. If $4c_2c_6 - {c_4}^2 < 0, c_2 > 0$, corresponding to (209), (210) and (197), Eq. (191) admits the following four tanh-sech hyperbolic type solutions

$$\begin{aligned} u_{1,2}(x,t) &= C_5 \\ &+ \frac{C_4}{\varepsilon\sqrt{\frac{2c_2\,\mathrm{sech}(\eta)[c_4\,\mathrm{sech}(\eta) \quad + \varepsilon i\sqrt{c_4^2 - 4c_6c_2}\tanh(\eta)]}{4c_6c_2\tanh^2(\eta) - {c_4}^2}}} \\ &+ \frac{C_3(4c_6c_2\tanh^2(\eta) - {c_4}^2)}{2c_2\,\mathrm{sech}(\eta)[c_4\,\mathrm{sech}(\eta) + \varepsilon i\sqrt{{c_4}^2 - 4c_6c_2}\tanh(\eta)]} \\ &+ C_2\varepsilon\sqrt{\frac{2c_2\,\mathrm{sech}(\eta)[c_4\,\mathrm{sech}(\eta) + \varepsilon i\sqrt{{c_4}^2 - 4c_6c_2}\tanh(\eta)]}{4c_6c_2\tanh^2(\eta) - {c_4}^2}} \\ &+ \frac{2c_2C_1\,\mathrm{sech}(\eta)[c_4\,\mathrm{sech}(\eta) + \varepsilon i\sqrt{{c_4}^2 - 4c_6c_2}\tanh(\eta)]}{4c_6c_2\tanh^2(\eta) - {c_4}^2}, \end{aligned} \tag{220}$$

and

$$\begin{aligned} u_{3,4}(x,t) &= C_5 \\ &+ \frac{C_3(4c_6c_2\tanh^2(\eta) - {c_4}^2)}{2c_2\,\mathrm{sech}(\eta)[c_4\,\mathrm{sech}(\eta) - \varepsilon i\sqrt{{c_4}^2 - 4c_6c_2}\tanh(\eta)]} \\ &+ \frac{C_4}{\varepsilon\sqrt{\frac{2c_2\,\mathrm{sech}(\eta)[c_4\,\mathrm{sech}(\eta) - \varepsilon i\sqrt{{c_4}^2 - 4c_6c_2}\tanh(\eta)]}{4c_6c_2\tanh^2(\eta) - {c_4}^2}}} \\ &+ C_2\varepsilon\sqrt{\frac{2c_2\,\mathrm{sech}(\eta)[c_4\,\mathrm{sech}(\eta) - \varepsilon i\sqrt{{c_4}^2 - 4c_6c_2}\tanh(\eta)]}{4c_6c_2\tanh^2(\eta) - {c_4}^2}} \\ &+ \frac{2c_2C_1\,\mathrm{sech}(\eta)[c_4\,\mathrm{sech}(\eta) - \varepsilon i\sqrt{{c_4}^2 - 4c_6c_2}\tanh(\eta)]}{4c_6c_2\tanh^2(\eta) - {c_4}^2}, \end{aligned} \tag{221}$$

where $\eta = 2\sqrt{c_2}((\alpha(x,t)\beta(t) - r(t)) - C_1)$, $A(t), B(t), \beta(t), F_1(t)$ are arbitrary functions of t, C_i, $i = 1, 2, 3, 4, 5$ are arbitrary constants, and

$$r(t) = \int F_1(t)\frac{\mathrm{d}}{\mathrm{d}t}\beta(t) + \left(\frac{\mathrm{d}}{\mathrm{d}t}F_1(t)\right)\beta(t)\mathrm{d}t + C_6 .$$

Case 2. If $4c_2c_6 - {c_4}^2 < 0, c_2 < 0$, corresponding to (211), (212) and (197), Eq. (191) admits the following four tan-sec triangular type solutions

$$\begin{aligned} u_{5,6}(x,t) &= C_5 \\ &+ \frac{C_4}{\varepsilon\sqrt{\frac{2c_2\,\mathrm{sec}(\zeta)[-c_4\,\mathrm{sec}(\zeta) + \varepsilon\sqrt{{c_4}^2 - 4c_6c_2}\tan(\zeta)]}{4c_6c_2\tan^2(\zeta) + {c_4}^2}}} \\ &+ \frac{C_3(4c_6c_2\tan^2(\zeta) + {c_4}^2)}{2c_2\,\mathrm{sec}(\zeta)[-c_4\,\mathrm{sec}(\zeta) + \varepsilon\sqrt{{c_4}^2 - 4c_6c_2}\tan(\zeta)]} \\ &+ C_2\varepsilon\sqrt{\frac{2c_2\,\mathrm{sec}(\zeta)[-c_4\,\mathrm{sec}(\zeta) + \varepsilon\sqrt{{c_4}^2 - 4c_6c_2}\tan(\zeta)]}{4c_6c_2\tan^2(\zeta) + {c_4}^2}} \\ &+ \frac{2c_2C_1\,\mathrm{sec}(\zeta)[-c_4\,\mathrm{sec}(\zeta) + \varepsilon\sqrt{{c_4}^2 - 4c_6c_2}\tan(\zeta)]}{4c_6c_2\tan^2(\zeta) + {c_4}^2}, \end{aligned} \tag{222}$$

and

$$\begin{aligned} u_{7,8}(x,t) &= C_5 \\ &+ \frac{C_4}{\varepsilon\sqrt{\frac{2c_2\,\mathrm{sec}(\zeta)[-c_4\,\mathrm{sec}(\zeta) - \varepsilon\sqrt{{c_4}^2 - 4c_6c_2}\tan(\zeta)]}{4c_6c_2\tan^2(\zeta) + {c_4}^2}}} \end{aligned}$$

$$+\frac{C_3(4c_6c_2\tan^2(\zeta)+{c_4}^2)}{2c_2\sec(\zeta)[-c_4\sec(\zeta)-\varepsilon\sqrt{{c_4}^2-4c_6c_2}\tan(\zeta)]}$$
$$+C_2\varepsilon\sqrt{\frac{2c_2\sec(\zeta)[-c_4\sec(\zeta)-\varepsilon\sqrt{{c_4}^2-4c_6c_2}\tan(\zeta)]}{4c_6c_2\tan^2(\zeta)+{c_4}^2}}$$
$$+\frac{2c_2C_1\sec(\zeta)[-c_4\sec(\zeta)-\varepsilon\sqrt{{c_4}^2-4c_6c_2}\tan(\zeta)]}{4c_6c_2\tan^2(\zeta)+{c_4}^2}, \tag{223}$$

where $\zeta=2\sqrt{-c_2}((\alpha(x,t)\beta(t)-r(t))-C_1)$ and the rest of the parameters are the same as with Case 1.

Case 3. If $4c_2c_6-{c_4}^2>0, c_2>0$, corresponding to (213), (214) and (197), Eq. (191) admits the following four sinh-cosh hyperbolic type solutions

$$u_{9,10}(x,t)$$
$$=C_5+C_2\varepsilon\sqrt{\frac{2c_2[c_4+\varepsilon\sqrt{4c_2c_6-{c_4}^2}\sinh(\eta)]}{4c_2c_6\sinh^2(\eta)-{c_4}^2\cosh^2(\eta)}}$$
$$+\frac{2c_2C_1[c_4+\varepsilon\sqrt{4c_2c_6-{c_4}^2}\sinh(\eta)]}{4c_2c_6\sinh^2(\eta)-{c_4}^2\cosh^2(\eta)}$$
$$+\frac{C_4}{\varepsilon\sqrt{\frac{2c_2[c_4+\varepsilon\sqrt{4c_2c_6-{c_4}^2}\sinh(\eta)]}{4c_2c_6\sinh^2(\eta)-{c_4}^2\cosh^2(\eta)}}}$$
$$+\frac{2c_2C_3[c_4+\varepsilon\sqrt{4c_2c_6-{c_4}^2}\sinh(\eta)]}{4c_2c_6\sinh^2(\eta)-{c_4}^2\cosh^2(\eta)}, \tag{224}$$

and

$$u_{11,12}(x,t)$$
$$=C_5+C_2\varepsilon\sqrt{\frac{2c_2[c_4+\varepsilon\sqrt{4c_2c_6-{c_4}^2}\sinh(\eta)]}{4c_2c_6\sinh^2(\eta)-{c_4}^2\cosh^2(\eta)}}$$
$$+\frac{2c_2C_1[c_4+\varepsilon\sqrt{4c_2c_6-{c_4}^2}\sinh(\eta)]}{4c_2c_6\sinh^2(\eta)-{c_4}^2\cosh^2(\eta)}$$
$$+\frac{C_4}{\varepsilon\sqrt{\frac{2c_2[c_4+\varepsilon\sqrt{4c_2c_6-{c_4}^2}\sinh(\eta)]}{4c_2c_6\sinh^2(\eta)-{c_4}^2\cosh^2(\eta)}}}$$
$$+\frac{2c_2C_3[c_4+\varepsilon\sqrt{4c_2c_6-{c_4}^2}\sinh(\eta)]}{4c_2c_6\sinh^2(\eta)-{c_4}^2\cosh^2(\eta)}, \tag{225}$$

where $\eta=2\sqrt{c_2}((\alpha(x,t)\beta(t)-r(t))-C_1)$ and the rest of the parameters are the same as with Case 1.

Case 4. If $4c_2c_6-{c_4}^2>0, c_2<0$, corresponding to (215), (216) and (197), Eq. (191) admits the following four sin-cos triangular type solutions

$$u_{13,14}(x,t)$$
$$=C_5+\frac{C_4}{\varepsilon\sqrt{\frac{2c_2[-c_4+\varepsilon i\sqrt{4c_2c_6-{c_4}^2}\sin(\zeta)]}{4c_6c_2\sin^2(\zeta)+{c_4}^2\cos^2(\zeta)}}}$$
$$+\frac{C_3(4c_6c_2\sin^2(\zeta)+{c_4}^2\cos^2(\zeta))}{2c_2[-c_4+\varepsilon i\sqrt{4c_2c_6-{c_4}^2}\sin(\zeta)]}$$
$$+C_2\varepsilon\sqrt{\frac{2c_2[-c_4+\varepsilon i\sqrt{4c_2c_6-{c_4}^2}\sin(\zeta)]}{4c_6c_2\sin^2(\zeta)+{c_4}^2\cos^2(\zeta)}}$$
$$+\frac{2c_2C_1\left[-c_4+\varepsilon i\sqrt{4c_2c_6-{c_4}^2}\sin(\zeta)\right]}{4c_6c_2\sin^2(\zeta)+{c_4}^2\cos^2(\zeta)}, \tag{226}$$

and

$$u_{15,16}(x,t)$$
$$=C_5+\frac{C_4}{\varepsilon\sqrt{\frac{2c_2\left[-c_4-\varepsilon i\sqrt{4c_2c_6-{c_4}^2}\sin(\zeta)\right]}{4c_6c_2\sin^2(\zeta)+{c_4}^2\cos^2(\zeta)}}}$$
$$+\frac{C_3(4c_6c_2\sin^2(\zeta)+{c_4}^2\cos^2(\zeta))}{2c_2\left[-c_4-\varepsilon i\sqrt{4c_2c_6-{c_4}^2}\sin(\zeta)\right]}$$
$$+C_2\varepsilon\sqrt{\frac{2c_2\left[-c_4-\varepsilon i\sqrt{4c_2c_6-{c_4}^2}\sin(\zeta)\right]}{4c_6c_2\sin^2(\zeta)+{c_4}^2\cos^2(\zeta)}}$$
$$+\frac{2c_2C_1\left[-c_4-\varepsilon i\sqrt{4c_2c_6-{c_4}^2}\sin(\zeta)\right]}{4c_6c_2\sin^2(\zeta)+{c_4}^2\cos^2(\zeta)}, \tag{227}$$

where $\zeta=2\sqrt{-c_2}((\alpha(x,t)\beta(t)-r(t))-C_1)$ and the rest of the parameters are the same as with Case 1.

Case 5. If $4c_2c_6-{c_4}^2=0, c_2>0$, corresponding to (217) and (197), Eq. (191) admits the following four tanh hyperbolic type solutions

$$u_{17,18,19,20}(x,t)$$
$$=C_5+C_2-\varepsilon\sqrt{\frac{2c_2\left[1+\varepsilon\tanh\left(\frac{\eta}{2}\right)\right]}{1-c_4-\varepsilon(1+c_4)\tanh\left(\frac{\eta}{2}\right)}}$$
$$+\frac{2c_2C_1\left[1+\varepsilon\tanh\left(\frac{\eta}{2}\right)\right]}{1-c_4-\varepsilon(1+c_4)\tanh\left(\frac{\eta}{2}\right)}$$
$$+\frac{C_4}{-\varepsilon\sqrt{\frac{2c_2[1+\varepsilon\tanh(\frac{\eta}{2})]}{1-c_4-\varepsilon(1+c_4)\tanh(\frac{\eta}{2})}}}$$
$$+\frac{C_3\left[1-c_4-\varepsilon(1+c_4)\tanh\left(\frac{\eta}{2}\right)\right]}{2c_2\left[1+\varepsilon\tanh\left(\frac{\eta}{2}\right)\right]}, \tag{228}$$

where $\eta=2\sqrt{c_2}((\alpha(x,t)\beta(t)-r(t))-C_1)$ and the rest of the parameters are the same as with Case 1.

Case 6. If $4c_2c_6 - c_4{}^2 = 0, c_2 < 0$, corresponding to (218) and (197), Eq. (191) admits the following two sectan triangular type solutions

$$\begin{aligned} &u_{21,22,23,24}(x,t) \\ &= C_5 + C_2 - \varepsilon\sqrt{\frac{2c_2\left[1+\varepsilon\tan\left(\frac{i\zeta}{2}\right)\right]}{1-c_4-\varepsilon(1+c_4)\tan\left(\frac{i\zeta}{2}\right)}} \\ &\quad - \frac{2c_2C_1\left[1+\varepsilon\tan\left(\frac{i\zeta}{2}\right)\right]}{1-c_4-\varepsilon(1+c_4)\tan\left(\frac{i\zeta}{2}\right)} \\ &\quad + \frac{C_4}{-\varepsilon\sqrt{\frac{2c_2\left[1+\varepsilon\tan\left(\frac{i\zeta}{2}\right)\right]}{1-c_4-\varepsilon(1+c_4)\tan\left(\frac{i\zeta}{2}\right)}}} \\ &\quad - \frac{C_3\left(1-c_4-\varepsilon(1+c_4)\tan\left(\frac{i\zeta}{2}\right)\right)}{2c_2\left[1+\varepsilon\tan\left(\frac{i\zeta}{2}\right)\right]}, \end{aligned} \tag{229}$$

where $\zeta = 2\sqrt{-c_2}((\alpha(x,t)\beta(t) - r(t)) - C_1)$ and the rest of the parameters are the same as with Case 1.

Remark 10 We may further generalize (194) as follows:

$$\begin{aligned} u &= a_0(X) \\ &+ \sum_{i=0}^{m} \frac{\sum_{r_{i1}+\cdots+r_{in}=i} a_{r_{i1},\ldots,r_{ni}}(X) G_{1i}^{r_{i1}}(\xi_{1i}(X))\ldots G_{ni}^{r_{ni}}(\xi_{ni}(X))}{\left[\sum_{l_{i1}+\cdots+l_{in}=i} b_{l_{i1},\ldots,l_{ni}}(X) G_{1i}^{l_{i1}}(\xi_{1i}(X))\ldots G_{ni}^{l_{ni}}(\xi_{ni}(X)) + c_i(X)\right]^{\tau}}, \end{aligned} \tag{230}$$

where $\sum_{l_{i1}+\cdots+l_{in}=i} b^2_{l_{i1},\ldots,l_{ni}}(X) + c_i^2(X) \neq 0$, $a_0(X)$, $\xi_{1i}(X), \ldots, \xi_{ni}(X)$, $a_{r_{i1},\ldots,r_{ni}}(X)$, $b_{l_{i1},\ldots,l_{ni}}(X)$ and $c_i(X)$, $i = 0, 1, 2, \ldots, m$ are all differentiable functions to be determined later, τ is a constant, $G_{1i}(\xi_{1i}(X)), \ldots, G_{ni}(\xi_{ni}(X))$ are all $\phi(\xi_{ki}(X))$ or $\phi^{-1}(\xi_{ki}(X))$ or some derivatives $\phi^{(j)}(\xi_{ki}(X)), k = 1, 2, \ldots, n; i = 0, 1, 2, \ldots, m; j = \pm 1, \pm 2, \ldots$. We can get many new explicit solutions of Eq. (191).

A New Exp-N Solitary-Like Method and Its Application in the (1 + 1)-Dimensional Generalized KdV Equation

In this section, in order to develop the Exp-function method [31], we present two new generic transformations, a new Exp-N solitary-like method, and its algorithm [40, 47]. In addition, we apply our method to construct new exact solutions of the (1 + 1)-dimensional classical generalized KdV(gKdV) equation.

Summary of the Exp-N Solitary-Like Method

In the following we would like to outline the main steps of our Exp-N solitary-like method [40,47]:

Step 1. For a given NLEE system with some physical fields $u_m(t, x_1, x_2, \ldots, x_{n-1}), (m = 1, 2, \ldots, n)$ in n independent variables $t, x_1, x_2, \ldots, x_{n-1}$,

$$\begin{aligned} &F_m(u_1, u_2, \ldots, u_n, u_{1,t}, u_{2,t}, \ldots, u_{n,t}, \\ &u_{1,x_1}, u_{2,x_1}, \ldots, u_{n,x_1}, u_{1,tt}, u_{2,tt}, \ldots, u_{n,tt}, \ldots) = 0, \\ &\qquad (m = 1, 2, \ldots, n). \end{aligned} \tag{231}$$

We introduce a new generic transformation

$$\begin{cases} u_m(t, x_1, \ldots, x_{n-1}) = u_m(\xi_1, \xi_2) \quad (m = 1, 2, \ldots, n), \\ \xi_j = p_{0j}(t) + \\ \sum_{i=1}^{n-1} p_{ij}(t, x_1, x_2, \ldots, x_{i-1}) \cdot q_{ij}(x_i), \quad j = 1, 2, \end{cases} \tag{232}$$

where we let $p_{1j}(t, x_0) = p_{1j}(t)$, here $p_{ij}(t, x_1, x_2, \ldots, x_{i-1}), q_{ij}(x_i)$, $j = 1, 2$; $i = 1, 2, \ldots, n-1$ and $p_{0j}(t)$ are functions to be determined later. Then the nonlinear partial differentials of Eq. (231) are reduced to a nonlinear partial differential equation with respect to $\xi_j, j = 1, 2, p_{ij}(t, x_1, x_2, \ldots, x_{i-1}), q_{ij}(x_i), j = 1, 2$; $i = 1, 2, \ldots, n-1$ and $p_{0j}(t)$.

Step 2. Let $X = (t, x_1, \ldots, x_{n-1})$. Then we introduce a new transformation in terms of Exp-function rational formal expansion in the following form based on the idea of the Exp-function method [41] by the new generic transformation (232).

$$u_m(\xi_1, \xi_2) = g_m(X) + \frac{\sum_{j=-c}^{d} a_{mj}(X) exp(j\xi_1)}{\sum_{k=-p}^{q} b_{mk}(X) exp(k\xi_2)}, \quad (m = 1, 2, \ldots, n), \tag{233}$$

where c, d, p and q are any positive integers, $\xi_1, \xi_2, g_m(X)$, $a_{mj}(X), j = -c, \ldots, d$, and $b_{mk}(X), k = -p, \ldots, q; m = 1, 2, \ldots, n$ are functions to be determined later.

Step 3. We take a numerical value of c, d, p and q, substitute (233) into the partial differential equations obtained in Step 1, and then set all coefficients of $\exp(k\eta), k = 0, 1, 2, \ldots$ to zero to get an over-determined partial differential equation with respect to $g_m(X)$, $a_{mj}(X), j = -c, \ldots, d$, $b_{mk}(X), k = -p, \ldots, q; m = 1, 2, \ldots, n$, $p_{ij}(t, x_1, x_2, \ldots, x_{i-1}), q_{ij}(x_i), j = 1, 2; i = 1, 2, \ldots, n-1$ and $p_{0j}(t)$.

Step 4. Solving the over-determined partial differential equations in Step 3 by use of Maple, we would end up with explicit expressions for $g_m(X), a_{mj}(X), j = -c, \ldots, d$, $b_{mk}(X), k = -p, \ldots, q$; $m = 1, 2, \ldots, n$, $p_{ij}(t, x_1, x_2, \ldots, x_{i-1}), q_{ij}(x_i), j = 1, 2; i = 1, 2, \ldots, n-1$ and $p_{0j}(t)$.

Step 5. Substituting the results obtained in Step 4 into (232) and (233), we can then obtain new rational solutions of Exp-functions as follows

$$u_m(X) = g_m(X) + \frac{\sum_{j=-c}^{d} a_{mj}(X) \exp\left(j\left(p_{01}(t) + \sum_{i=1}^{n-1} p_{i1}(t, x_1, x_2, \ldots, x_{i-1}) \cdot q_{i1}(x_i)\right)\right)}{\sum_{k=-p}^{q} b_{mk}(X) \exp\left(k\left(p_{02}(t) + \sum_{i=1}^{n-1} p_{i2}(t, x_1, x_2, \ldots, x_{i-1}) \cdot q_{i2}(x_i)\right)\right)}, \quad (m = 1, 2, \ldots, n) \tag{234}$$

Step 6. If we again take a numerical value of c, d, p and q and repeat Step 3, Step 4, and Step 5, then we can obtain other rational solutions of Exp-functions Eq. (231). Now, repeat the process. We can obtain a family of new N solitary-like solutions of Eq. (231).

Remark 11 The new transformations (232) and (233) proposed here are more general than the transformation in the Exp-function method [31]. We have obtained many families of non-traveling waves solutions and traveling waves solutions of the $(2+1)$-dimensional generalized KdV–Burgers equation by using the two transformations in [40,47].

Remark 12 Our methods are more powerful, simpler, and more convenient than the Exp-function method [31]. The Exp-function method [31] is used only to obtain a traveling waves solution of lower dimensional NLPDE because it makes use of the homogeneous balance method. In this paper we improve the method to be more powerful such that it can be used to obtain many families of non-traveling waves solutions and traveling waves solutions of higher dimensional NLPDEs because we don't use the homogeneous balance method and make use of our transformations (232) and (233).

The Application of the Exp-N Solitary-Like Method in the (1 + 1)-Dimensional Generalized KdV Equation

In this section, we will present a new transformation, and then make use of the transformation and our method via symbolic computation to find the solutions of the $(1+1)$-dimensional classical generalized Korteweg–de Vries (gKdV) equation.

The $(1+1)$-dimensional gKdV equation has the following form [7,8,9]:

$$u_t(x,t) + \alpha u^p(x,t) u_x(x,t) + \beta u_{xxx}(x,t) = 0\,, \tag{235}$$

where the coefficient of the nonlinear term α and the coefficient of the dispersive term β are independent of x and t.

According to the above steps, to seek traveling wave solutions of Eq. (235), we present the following new transformation here. When $p \geq 2$

$$u(x,t) = f^{\frac{2}{p}}(x,t)\,. \tag{236}$$

By substituting (236) into Eq. (235), we get the following equation:

$$\begin{aligned} &p^2 f^2(x,t) f_t(x,t) + \alpha p^2 f^4(x,t) f_x(x,t) \\ &+ 4\beta f_x^3(x,t) + 6p\beta f(x,t) f_x(x,t) f_{xx}(x,t) \\ &- 6p\beta f_x^3(x,t) + \beta p^2 f_{xxx}(x,t) f^2(x,t) \\ &- 3\beta p^2 f(x,t) f_x(x,t) f_{xx}(x,t) + 2\beta p^2 f_x^3(x,t) = 0\,. \end{aligned} \tag{237}$$

We introduce a new generic transformation [40,47]

$$f(x,t) = g + \frac{\sum_{i=-m_1}^{m_2} k_i e^{i(\alpha_1 x + \beta_1 t + \gamma_1)}}{\sum_{j=-n_1}^{n_2} h_j e^{j(\alpha_2 x + \beta_2 t + \gamma_2)}}\,, \tag{238}$$

where m_1, m_2, n_1, n_2 are any positive integers, g, $\alpha_1, \beta_1, \alpha_2, \beta_2, \gamma_1, \gamma_2$, $k_i, i = -m_1, \ldots, m_2$, and $h_j, j = -n_1, \ldots, n_2$ are all constants to be determined later.

With the aid of Maple, substituting (238) into (237) and setting $\xi = \alpha_1 x + \beta_1 t + \gamma_1, \eta = \alpha_2 x + \beta_2 t + \gamma_2$, yields the following equation (because the equation is so long, just part of the equation is shown here for simplification)

$$\begin{aligned} &\alpha p^2 \left(\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^4 \alpha_1 \sum_{i=-m_1}^{m_2} k_i i e^{i\xi} \sum_{j=-n_1}^{n_2} h_j e^{j\eta} \\ &- p^2 \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^4 \sum_{i=-m_1}^{m_2} k_i e^{i\xi} \beta_2 \\ &\quad \sum_{j=-n_1}^{n_2} h_j j e^{j\eta} \beta\, p^2 g^2 \\ &+ {\alpha_1}^3 \sum_{i=-m_1}^{m_2} k_i i^3 e^{i\xi} \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^5 g^2 \\ &- \beta\, p^2 \sum_{i=-m_1}^{m_2} k_i e^{i\xi} {\alpha_2}^3 \\ &\quad \sum_{j=-n_1}^{n_2} h_j j^3 e^{j\eta} \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^4 g^2 \\ &+ \alpha\, p^2 g^4 \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^5 \alpha_1 \sum_{i=-m_1}^{m_2} k_i i e^{i\xi} \\ &+ 6\,\beta\, p {\alpha_1}^3 \sum_{i=-m_1}^{m_2} k_i i e^{i\xi} \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^4 \\ &\quad \sum_{i=-m_1}^{m_2} k_i i^2 e^{i\xi} g \end{aligned}$$

$$+ 6\beta p \left(\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^2 \alpha_2{}^3 \sum_{j=-n_1}^{n_2} h_j j e^{j\eta}$$
$$\sum_{j=-n_1}^{n_2} h_j j^2 e^{j\eta} \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^2 g$$
$$- 12\beta p \left(\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^2 \alpha_2{}^3 \left(\sum_{j=-n_1}^{n_2} h_j j e^{j\eta}\right)^3$$
$$g \sum_{j=-n_1}^{n_2} h_j e^{j\eta}$$
$$+ 12\beta \alpha_1 \sum_{i=-m_1}^{m_2} k_i i e^{i\xi} \sum_{j=-n_1}^{n_2} h_j e^{j\eta}$$
$$\left(\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^2 \alpha_2{}^2 \left(\sum_{j=-n_1}^{n_2} h_j j e^{j\eta}\right)^2$$
$$- 12\beta \alpha_1{}^2 \left(\sum_{i=-m_1}^{m_2} k_i i e^{i\xi}\right)^2 \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^2$$
$$\sum_{i=-m_1}^{m_2} k_i e^{i\xi} \alpha_2 \sum_{j=-n_1}^{n_2} h_j j e^{j\eta}$$
$$+ 4\alpha p^2 g \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^2 \left(\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^3 \alpha_1$$
$$\sum_{i=-m_1}^{m_2} k_i i e^{i\xi}$$
$$- 4\alpha p^2 g^3 \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^3 \left(\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^2 \alpha_2$$
$$\sum_{j=-n_1}^{n_2} h_j j e^{j\eta}$$
$$+ 6\alpha p^2 g^2 \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^3 \left(\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^2 \alpha_1$$
$$\sum_{i=-m_1}^{m_2} k_i i e^{i\xi}$$
$$- 2p^2 \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^3 \left(\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^2 \beta_2$$
$$\sum_{j=-n_1}^{n_2} h_j j e^{j\eta} g$$
$$- 3\beta p^2 \alpha_1{}^3 \sum_{i=-m_1}^{m_2} k_i i e^{i\xi} \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^4$$
$$\sum_{i=-m_1}^{m_2} k_i i^2 e^{i\xi} g$$
$$- 4\alpha p^2 g \sum_{j=-n_1}^{n_2} h_j e^{j\eta} \left(\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^4$$
$$\alpha_2 \sum_{j=-n_1}^{n_2} h_j j e^{j\eta}$$
$$+ 2\beta p^2 \alpha_1{}^3 \sum_{i=-m_1}^{m_2} k_i i^3 e^{i\xi} \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^4$$
$$g \sum_{i=-m_1}^{m_2} k_i e^{i\xi}$$
$$- 6\beta p^2 \sum_{i=-m_1}^{m_2} k_i e^{i\xi} \alpha_2{}^3 \left(\sum_{j=-n_1}^{n_2} h_j j e^{j\eta}\right)^3$$
$$g^2 \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^2$$
$$+ 3\beta p^2 \left(\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^3 \alpha_2{}^3 \sum_{j=-n_1}^{n_2} h_j j e^{j\eta}$$
$$\sum_{j=-n_1}^{n_2} h_j j^2 e^{j\eta} \sum_{j=-n_1}^{n_2} h_j e^{j\eta}$$
$$+ 6\beta p \left(\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^3 \alpha_2{}^3 \sum_{j=-n_1}^{n_2} h_j j e^{j\eta}$$
$$\sum_{j=-n_1}^{n_2} h_j j^2 e^{j\eta} \sum_{j=-n_1}^{n_2} h_j e^{j\eta}$$
$$+ 6\beta p^2 \alpha_1 \sum_{i=-m_1}^{m_2} k_i i e^{i\xi} \alpha_2{}^2 \left(\sum_{j=-n_1}^{n_2} h_j j e^{j\eta}\right)^2$$
$$\left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^3 g^2$$
$$- 3\beta p^2 \alpha_1 \sum_{i=-m_1}^{m_2} k_i i e^{i\xi} \alpha_2{}^2 \sum_{j=-n_1}^{n_2} h_j j^2 e^{j\eta}$$
$$\left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^4 g^2$$
$$+ 6\beta p \alpha_1{}^3 \sum_{i=-m_1}^{m_2} k_i i e^{i\xi} \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^3$$
$$\sum_{i=-m_1}^{m_2} k_i i^2 e^{i\xi} \sum_{i=-m_1}^{m_2} k_i e^{i\xi}$$
$$- 6\beta p \alpha_1{}^3 \left(\sum_{i=-m_1}^{m_2} k_i i e^{i\xi}\right)^3 \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^3$$
$$- 2\beta p^2 \left(\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^3 \alpha_2{}^3 \left(\sum_{j=-n_1}^{n_2} h_j j e^{j\eta}\right)^3$$
$$- 6\beta p \left(\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^3 \alpha_2{}^3 \left(\sum_{j=-n_1}^{n_2} h_j j e^{j\eta}\right)^3$$
$$+ 2\beta p^2 \alpha_1{}^3 \left(\sum_{i=-m_1}^{m_2} k_i i e^{i\xi}\right)^3 \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^3$$
$$- 6\alpha p^2 g^2 \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^2 \left(\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^3$$
$$\alpha_2 \sum_{j=-n_1}^{n_2} h_j j e^{j\eta} +$$
$$6\beta p \alpha_1{}^2 \left(\sum_{i=-m_1}^{m_2} k_i i e^{i\xi}\right)^2 \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^2$$
$$\sum_{i=-m_1}^{m_2} k_i e^{i\xi} \alpha_2 \sum_{j=-n_1}^{n_2} h_j j e^{j\eta}$$
$$+ 9\beta p^2 \left(\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^2 \alpha_2{}^3 \sum_{j=-n_1}^{n_2} h_j j e^{j\eta}$$
$$\sum_{j=-n_1}^{n_2} h_j j^2 e^{j\eta} \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^2 g$$
$$- 6\beta p \left(\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^2 \alpha_2 \sum_{j=-n_1}^{n_2} h_j j e^{j\eta}$$
$$\alpha_1{}^2 \sum_{i=-m_1}^{m_2} k_i i^2 e^{i\xi} \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^2$$
$$+ 6\beta p^2 \alpha_1{}^2 \left(\sum_{i=-m_1}^{m_2} k_i i e^{i\xi}\right)^2 \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^3$$
$$\alpha_2 \sum_{j=-n_1}^{n_2} h_j j e^{j\eta} g$$
$$- 6\beta p \alpha_1 \sum_{i=-m_1}^{m_2} k_i i e^{i\xi} \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^2$$
$$\left(\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^2 \alpha_2{}^2 \sum_{j=-n_1}^{n_2} h_j j^2 e^{j\eta}$$
$$+ 6\beta p \alpha_1 \sum_{i=-m_1}^{m_2} k_i i e^{i\xi} \sum_{j=-n_1}^{n_2} h_j e^{j\eta}$$
$$\left(\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^2 \alpha_2{}^2 \left(\sum_{j=-n_1}^{n_2} h_j j e^{j\eta}\right)^2$$
$$- 2\beta p^2 \left(\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^2 \alpha_2{}^3 \sum_{j=-n_1}^{n_2} h_j j^3 e^{j\eta}$$
$$\left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^3 g$$
$$+ 4\alpha p^2 g^3 \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^4 \sum_{i=-m_1}^{m_2} k_i e^{i\xi}$$
$$\alpha_1 \sum_{i=-m_1}^{m_2} k_i i e^{i\xi}$$
$$+ 2p^2 \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^4 \beta_1 \sum_{i=-m_1}^{m_2} k_i i e^{i\xi}$$
$$g \sum_{i=-m_1}^{m_2} k_i e^{i\xi}$$
$$- 6\beta p \alpha_1 \sum_{i=-m_1}^{m_2} k_i i e^{i\xi} \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^3$$
$$\sum_{i=-m_1}^{m_2} k_i e^{i\xi} \alpha_2{}^2 \sum_{j=-n_1}^{n_2} h_j j^2 e^{j\eta} g$$
$$- 6\beta p^2 \left(\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^2 \alpha_2{}^3 \left(\sum_{j=-n_1}^{n_2} h_j j e^{j\eta}\right)^3$$
$$g \sum_{j=-n_1}^{n_2} h_j e^{j\eta}$$
$$- 3\beta p^2 \alpha_1{}^3 \sum_{i=-m_1}^{m_2} k_i i e^{i\xi} \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^3$$
$$\sum_{i=-m_1}^{m_2} k_i i^2 e^{i\xi} \sum_{i=-m_1}^{m_2} k_i e^{i\xi}$$
$$- 3\beta p^2 \alpha_1 \sum_{i=-m_1}^{m_2} k_i i e^{i\xi} \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^3$$
$$\sum_{i=-m_1}^{m_2} k_i e^{i\xi} \alpha_2{}^2 \sum_{j=-n_1}^{n_2} h_j j^2 e^{j\eta} g$$
$$+ 24\beta p \alpha_1 \sum_{i=-m_1}^{m_2} k_i i e^{i\xi} \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^2$$
$$\sum_{i=-m_1}^{m_2} k_i e^{i\xi} \alpha_2{}^2 \left(\sum_{j=-n_1}^{n_2} h_j j e^{j\eta}\right)^2 g$$
$$- 6\beta p \sum_{i=-m_1}^{m_2} k_i e^{i\xi} \alpha_2 \sum_{j=-n_1}^{n_2} h_j j e^{j\eta} \alpha_1{}^2$$

$$
\begin{aligned}
&\textstyle\sum_{i=-m_1}^{m_2} k_i i^2 e^{i\xi} \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^3 g \\
&- 3\beta p^2 \textstyle\sum_{i=-m_1}^{m_2} k_i e^{i\xi} \alpha_2 \sum_{j=-n_1}^{n_2} h_j j e^{j\eta} {\alpha_1}^2 \\
&\textstyle\sum_{i=-m_1}^{m_2} k_i i^2 e^{i\xi} \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^3 g \\
&- 3\beta p^2 \alpha_1^2 \textstyle\sum_{i=-m_1}^{m_2} k_i i^2 e^{i\xi} \alpha_2 \sum_{j=-n_1}^{n_2} h_j j e^{j\eta} \\
&\left(\textstyle\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^4 g^2 \\
&- 12\beta p {\alpha_1}^2 \left(\textstyle\sum_{i=-m_1}^{m_2} k_i i e^{i\xi}\right)^2 \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^3 \\
&\alpha_2 \textstyle\sum_{j=-n_1}^{n_2} h_j j e^{j\eta} g \\
&+ 6\beta p^2 \textstyle\sum_{i=-m_1}^{m_2} k_i e^{i\xi} {\alpha_2}^3 \sum_{j=-n_1}^{n_2} h_j j e^{j\eta} \\
&\textstyle\sum_{j=-n_1}^{n_2} h_j j^2 e^{j\eta} \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^3 g^2 \\
&+ p^2 \left(\textstyle\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^3 \beta_1 \sum_{i=-m_1}^{m_2} k_i i e^{i\xi} \\
&\left(\textstyle\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^2 \\
&+ \beta p^2 {\alpha_1}^3 \textstyle\sum_{i=-m_1}^{m_2} k_i i^3 e^{i\xi} \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^3 \\
&\left(\textstyle\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^2 \\
&- \beta p^2 \left(\textstyle\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^3 {\alpha_2}^3 \sum_{j=-n_1}^{n_2} h_j j^3 e^{j\eta} \\
&\left(\textstyle\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^2 \\
&- 4\beta \left(\textstyle\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^3 {\alpha_2}^3 \left(\sum_{j=-n_1}^{n_2} h_j j e^{j\eta}\right)^3 \\
&+ 4\beta {\alpha_1}^3 \left(\textstyle\sum_{i=-m_1}^{m_2} k_i i e^{i\xi}\right)^3 \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^3 \\
&- p^2 \left(\textstyle\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^2 \left(\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^3 \\
&\beta_2 \textstyle\sum_{j=-n_1}^{n_2} h_j j e^{j\eta} \\
&- \alpha p^2 \left(\textstyle\sum_{i=-m_1}^{m_2} k_i e^{i\xi}\right)^5 \alpha_2 \sum_{j=-n_1}^{n_2} h_j j e^{j\eta} + p^2
\end{aligned}
$$

$$
\begin{aligned}
&\left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^5 \beta_1 \sum_{i=-m_1}^{m_2} k_i i e^{i\xi} g^2 \\
&- \alpha p^2 g^4 \left(\sum_{j=-n_1}^{n_2} h_j e^{j\eta}\right)^4 \sum_{i=-m_1}^{m_2} k_i e^{i\xi} \alpha_2 \sum_{j=-n_1}^{n_2} h_j j e^{j\eta} \\
&= 0 . \qquad (239)
\end{aligned}
$$

where m_1, m_2, n_1, n_2 are all any positive integers which we may take arbitrarily.

For example, we can take $n_2 = 2, m_2 = 2, n_1 = 1$, $m_1 = 1$ and substitute them into (239), and then set all coefficients of $exp(k\eta)$ and $exp(j\xi)$, $j, k = 0, 1, 2, \ldots$ to zero to get an over-determined equation with respect to $h_{-1}, h_0, h_1, h_2, k_0, k_1, k_2, k_{-1}, \alpha_1, \beta_1, \alpha_2, \beta_2$ and g. Solving the over-determined equations by use of Maple, we obtain the following solutions of h_{-1}, $h_0, h_1, h_2, k_0, k_1, k_2, k_{-1}, \alpha_1, \beta_1, \alpha_2, \beta_2$ and g:

Case 1.

$$
\begin{aligned}
&\beta_2 = -\frac{4\beta {\alpha_2}^3}{p^2} , \quad h_1 = \frac{p^2 \alpha {k_0}^2}{8{\alpha_2}^2 \beta h_{-1}(3p + 2 + p^2)} , \\
&k_{-1} = 0 , \quad g = 0 , \gamma_2 = \gamma_2 , \quad \alpha_2 = \alpha_2 , \\
&k_0 = k_0 , \quad \alpha_1 = \alpha_1 , \quad \beta_1 = \beta_1 , \quad \gamma_1 = \gamma_1 , \\
&h_{-1} = h_{-1} , \quad h_0 = 0 , \quad k_2 = 0 , \quad k_1 = 0 , \\
&\qquad h_2 = 0 , \qquad (240)
\end{aligned}
$$

where $\alpha, \beta, \alpha_2, \gamma_2, \alpha_2, k_0, \alpha_1, \beta_1, \gamma_1$ and k_0 are any constants.

Case 2.

$$
\begin{aligned}
&\gamma_2 = \gamma_2 , \quad k_0 = k_0 , \quad \alpha_1 = \alpha_1 , \quad \beta_1 = \beta_1 , \\
&\gamma_1 = \gamma_1 , \quad k_2 = k_2 , \quad k_1 = k_1 , \quad h_{-1} = 0 , \\
&\beta_2 = \beta_2 , \quad \alpha_2 = \alpha_2 , \quad k_{-1} = k_{-1} , \quad h_2 = 0 , \\
&\qquad h_0 = 0 , \quad h_1 = 0 , \quad g = g , \qquad (241)
\end{aligned}
$$

where $\alpha, \beta, \gamma_2, k_0, \alpha_1, \beta_1, \gamma_1, k_2, k_1, \beta_2, \alpha_2, k_{-1}$ and g are any constants.

Case 3.

$$
\begin{aligned}
&\alpha_2 = -2\alpha_1 , \quad \beta_2 = -2\beta_1 , \quad \gamma_2 = \gamma_2 , \quad \alpha_1 = \alpha_1 , \\
&\beta_1 = \beta_1 , \quad \gamma_1 = \gamma_1 , \quad h_{-1} = h_{-1} , \quad h_0 = 0 , \\
&k_1 = 0 , \quad h_2 = 0 , \quad k_{-1} = 0 , \quad k_0 = 0 , \\
&\qquad h_1 = 0 , \quad k_2 = k_2, g = g , \qquad (242)
\end{aligned}
$$

where $\alpha, \beta, \alpha_1, \beta_1, \gamma_2, \beta_1, \gamma_1, h_{-1}, k_2$ and g are any constants.

Case 4.

$$
\begin{aligned}
&\gamma_2 = \gamma_2 , \quad \alpha_2 = \alpha_2 , \quad \alpha_1 = \alpha_1 , \quad \beta_1 = \beta_1 , \\
&\gamma_1 = \gamma_1 , \quad h_{-1} = h_{-1} , \quad k_2 = 0 , \quad k_1 = 0 , \\
&k_{-1} = 0 , \quad \beta_2 = \beta_2 , \quad h_1 = h_1 , \quad h_2 = h_2 , \\
&\qquad h_0 = h_0 , \quad k_0 = 0 , \quad g = g , \qquad (243)
\end{aligned}
$$

where $\gamma_2, \alpha_2, \alpha_1, \beta_1, \gamma_1, h_{-1}, \beta_2, h_1, h_2$ and g are any constants.

Case 5.

$$
\begin{aligned}
&\gamma_2 = \gamma_2, k_0 = k_0, \alpha_1 = \alpha_1 , \quad \beta_1 = \beta_1 , \\
&\gamma_1 = \gamma_1 , \quad h_{-1} = h_{-1} , \quad k_2 = 0 , \quad k_1 = 0 , \\
&k_{-1} = 0 , \quad h_1 = h_1 , \quad h_2 = h_2 , \quad h_0 = h_0 , \\
&\qquad \alpha_2 = 0, \beta_2 = 0 , \quad g = g , \qquad (244)
\end{aligned}
$$

where $\alpha, \beta, \gamma_2, k_0, \alpha_1, \beta_1, \gamma_1, h_{-1}, h_1, h_2, h_0$ and g are any constants.

Case 6.

$$\beta_2 = -\beta_1\,, \quad \alpha_2 = -\alpha_1\,, \quad \gamma_2 = \gamma_2\,, \quad \alpha_1 = \alpha_1\,,$$
$$\beta_1 = \beta_1\,, \quad \gamma_1 = \gamma_1\,, \quad h_{-1} = h_{-1}\,, \quad h_0 = 0\,,$$
$$k_2 = 0\,, \quad h_2 = 0\,, \quad k_{-1} = 0\,, \quad k_0 = 0\,,$$
$$h_1 = 0\,, \quad g = g\,, \quad k_1 = k_1\,, \tag{245}$$

where $\alpha, \beta, \beta_1, \alpha_1, \gamma_2, \gamma_1, h_{-1}, g$ and k_1 are any constants.

Case 7.

$$\gamma_2 = \gamma_2\,, \quad \gamma_1 = \gamma_1\,, \quad h_{-1} = h_{-1}\,, \quad k_2 = 0\,,$$
$$k_1 = 0\,, \quad \beta_2 = \beta_2\,, \quad \alpha_2 = \alpha_2\,, \quad k_{-1} = k_{-1}\,,$$
$$\beta_1 = \beta_2\,, \quad h_2 = 0\,, \quad h_0 = 0\,, \quad h_1 = 0\,,$$
$$\alpha_1 = \alpha_2\,, \quad k_0 = 0\,, \quad g = g\,, \tag{246}$$

where $\alpha, \beta, \gamma_2, \gamma_1, h_{-1}, \beta_2, \alpha_2, k_{-1}, \alpha_1$ and g are any constants.

Case 8.

$$\gamma_2 = \gamma_2\,, \quad k_0 = k_0\,, \quad \gamma_1 = \gamma_1\,, \quad h_0 = h_0\,,$$
$$\beta_1 = 0\,, \quad \alpha_1 = 0\,, \quad k_2 = k_2\,, \quad k_1 = k_1\,,$$
$$h_{-1} = 0\,, \quad \beta_2 = \beta_2\,, \quad \alpha_2 = \alpha_2\,, \quad k_{-1} = k_{-1}\,,$$
$$h_2 = 0\,, \quad h_1 = 0\,, \quad g = g\,, \tag{247}$$

where $\alpha, \beta, \gamma_2, k_0, \gamma_1, h_0, k_2, k_1, \beta_2, \alpha_2, k_{-1}$ and g are any constants.

Case 9.

$$\gamma_2 = \gamma_2\,, \quad \gamma_1 = \gamma_1\,, \quad k_2 = 0\,, \quad k_1 = 0\,,$$
$$\alpha_1 = -\alpha_2\,, \quad \beta_1 = -\beta_2\,, \quad g = 0\,, \quad h_1 = h_1\,,$$
$$h_{-1} = 0\,, \quad \beta_2 = \beta_2\,, \quad \alpha_2 = \alpha_2\,, \quad k_{-1} = k_{-1}\,,$$
$$h_2 = 0\,, \quad h_0 = 0\,, \quad k_0 = 0\,, \tag{248}$$

where $\alpha, \beta, \gamma_2, \gamma_1, \alpha_2, \beta_2, h_1, \beta_2, \alpha_2$ and k_{-1} are any constants.

Case 10.

$$\gamma_2 = \gamma_2, \beta_1 = \beta_1, \gamma_1 = \gamma_1\,, \quad k_2 = 0\,,$$
$$k_1 = 0\,, \quad \alpha_1 = -\alpha_2, h_1 = h_1, \beta_2 = -\beta_1\,,$$
$$h_{-1} = 0\,, \alpha_2 = \alpha_2\,, \quad k_{-1} = k_{-1}, h_2 = 0\,,$$
$$h_0 = 0, k_0 = 0, g = g\,, \tag{249}$$

where $\alpha, \beta, \gamma_2, \beta_1, \gamma_1, \alpha_2, h_1, k_{-1}$ and g are any constants.

Case 11.

$$\beta_1 = \frac{1}{2}\beta_2\,, \quad \gamma_2 = \gamma_2\,, \quad \gamma_1 = \gamma_1\,, \quad k_1 = 0\,,$$
$$k_{-1} = 0\,, \quad h_1 = h_1\,, \quad \alpha_2 = 0\,, \quad \alpha_1 = 0\,,$$
$$k_2 = k_2\,, \quad h_{-1} = 0\,, \quad \beta_2 = \beta_2\,, \quad h_2 = 0\,,$$
$$h_0 = 0\,, \quad k_0 = 0\,, \quad g = g\,, \tag{250}$$

where $\beta_1 = \frac{1}{2}\beta_2, \alpha, \beta, \gamma_2, \gamma_1, h_1, k_2, \beta_2$ and g are any constants.

Case 12.

$$\beta_2 = 2\beta_1\,, \quad \alpha_1 = \frac{1}{2}\alpha_2\,, \quad \gamma_2 = \gamma_2\,, \quad \gamma_1 = \gamma_1\,,$$
$$k_1 = 0\,, \quad k_{-1} = 0\,, \quad h_1 = h_1\,, \quad \alpha_2 = \alpha_2\,,$$
$$\beta_1 = \beta_1\,, \quad k_2 = k_2\,, \quad h_{-1} = 0\,, \quad h_2 = 0\,,$$
$$h_0 = 0\,, \quad k_0 = 0\,, \quad g = g\,, \tag{251}$$

where $\alpha, \beta, \gamma_2, \gamma_1, h_1, \alpha_2, \beta_1, k_2$ and g are any constants.

Case 13.

$$\beta_1 = -2\beta_2\,, \quad h_1 = 0\,, \quad k_2 = 0\,, \quad \alpha_2 = -1/2\alpha_1\,,$$
$$g = 0\,, \quad h_2 = h_2\,, \quad k_{-1} = k_{-1}\,, \quad \alpha_1 = \alpha_1\,,$$
$$\gamma_2 = \gamma_2\,, \quad \gamma_1 = \gamma_1\,, \quad k_1 = 0\,, \quad h_{-1} = 0\,,$$
$$\beta_2 = \beta_2\,, \quad h_0 = 0\,, \quad k_0 = 0\,, \tag{252}$$

where $\alpha, \beta, \beta_2, \alpha_1, h_2, \gamma_2, \gamma_1, \beta_2$ and k_{-1} are any constants.

Case 14.

$$\alpha_1 = 2\alpha_2\,, \quad \beta_1 = 2\beta_2\,, \quad \alpha_2 = \alpha_2\,, \quad h_1 = 0\,,$$
$$k_2 = 0, h_2 = h_2\,, \quad \gamma_2 = \gamma_2\,, \quad \gamma_1 = \gamma_1\,,$$
$$k_{-1} = 0\,, \quad k_1 = k_1\,, \quad h_{-1} = 0\,, \quad \beta_2 = \beta_2\,,$$
$$h_0 = 0\,, \quad k_0 = 0\,, \quad g = g\,, \tag{253}$$

where $\alpha, \beta, \alpha_2, \beta_2, h_2, \gamma_2, \gamma_1, k_1, \beta_2$ and g are any constants.

Case 15.

$$\beta_1 = -2\beta_2\,, \quad h_1 = 0\,, \quad k_2 = 0\,, \quad \alpha_2 = -1/2\alpha_1\,,$$
$$g = 0\,, \quad h_2 = h_2\,, \quad k_{-1} = k_{-1}\,, \quad \alpha_1 = \alpha_1\,,$$
$$\gamma_2 = \gamma_2\,, \quad \gamma_1 = \gamma_1\,, \quad k_1 = 0\,, \quad h_{-1} = 0\,,$$
$$\beta_2 = \beta_2\,, \quad h_0 = 0\,, \quad k_0 = 0\,, \tag{254}$$

where $\alpha, \beta, \beta_2, \alpha_1, h_2, \alpha_1, \gamma_2, \gamma_1, \beta_2$ and k_{-1} are any constants.

Case 16.

$$\beta_1 = 2\beta_2\,, \quad h_1 = 0\,, \quad k_2 = 0\,, \quad h_2 = h_2\,,$$
$$\gamma_2 = \gamma_2\,, \quad \gamma_1 = \gamma_1\,, \quad k_{-1} = 0\,, \quad k_1 = k_1\,,$$
$$\alpha_1 = 2\alpha_2\,, \quad \alpha_2 = \alpha_2\,, \quad h_{-1} = 0\,, \quad \beta_2 = \beta_2\,,$$
$$h_0 = 0\,, \quad k_0 = 0\,, \quad g = g\,, \tag{255}$$

where $\alpha, \beta, \beta_2, \gamma_2, \gamma_1, k_1, \alpha_2, \beta_2$ and g are any constants.

Case 17.

$$\alpha_1 = -2\alpha_2\,, \quad h_1 = 0\,, \quad k_2 = 0\,, \quad h_2 = h_2\,,$$
$$\gamma_2 = \gamma_2\,, \quad \gamma_1 = \gamma_1\,, \quad \alpha_2 = \alpha_2\,, \quad \beta_1 = -2\beta_2\,,$$
$$k_{-1} = k_{-1}\,, \quad k_1 = 0\,, \quad h_{-1} = 0\,, \quad \beta_2 = \beta_2\,,$$
$$h_0 = 0\,, \quad k_0 = 0\,, \quad g = g\,, \tag{256}$$

where $\alpha, \beta, \alpha_2, h_2, \gamma_2, \gamma_1, \beta_2, k_{-1}, \beta_2$ and g are any constants.

Case 18.

$$h_2 = h_2\,, \quad \gamma_2 = \gamma_2\,, \quad \gamma_1 = \gamma_1\,, \quad h_1 = 0\,,$$
$$k_{-1} = 0\,, \quad \alpha_2 = \alpha_2\,, \quad \alpha_1 = \alpha_2\,, \quad k_2 = k_2\,,$$
$$\beta_1 = \beta_2\,, \quad k_1 = 0\,, \quad h_{-1} = 0\,, \quad \beta_2 = \beta_2\,,$$
$$h_0 = 0\,, \quad k_0 = 0\,, \quad g = g\,, \tag{257}$$

where $\alpha, \beta, h_2, \gamma_2, \gamma_1, \alpha_2, k_2, \beta_2$ and g are any constants.

Case 19.

$$h_1 = 0\,, \quad k_0 = k_0\,, \quad h_0 = h_0\,, \quad h_2 = 0\,,$$
$$\beta_1 = 0\,, \quad \alpha_1 = 0\,, \quad \gamma_2 = \gamma_2\,, \quad \gamma_1 = \gamma_1\,,$$
$$k_1 = k_1\,, \quad \alpha_2 = \alpha_2\,, \quad k_2 = k_2\,, \quad k_{-1} = k_{-1}\,,$$
$$h_{-1} = 0\,, \quad \beta_2 = \beta_2\,, \quad g = g\,, \tag{258}$$

where $\alpha, \beta, k_0, h_0, \gamma_2, \gamma_1, k_1, \alpha_2, k_2, k_{-1}, \beta_2$ and g are any constants.

Substituting (240)–(258) into (236) and (238), respectively, we can obtain the exact solutions of Eq. (235).

$$u(x,t) = \sqrt[2/p]{g + \frac{k_0 + 2\cosh_{k_1,k_{-1},1}(\alpha_1 x + \beta_1 t + \gamma_1) + k_2 e^{2(\alpha_1 x + \beta_1 t + \gamma_1)}}{h_0 + 2\cosh_{h_1,h_{-1},1}(\alpha_2 x + \beta_2 t + \gamma_2) + h_2 e^{2(\alpha_2 x + \beta_2 t + \gamma_2)}}}\,, \tag{259}$$

where $\cosh_{k_1,k_{-1},1}(\alpha_1 x + \beta_1 t + \gamma_1)$ and $\cosh_{h_1,h_{-1},1}(\alpha_2 x + \beta_2 t + \gamma_2)$ are all the generalized hyperbolic cosine functions, here $k_i, h_i, i = 0, \pm 1, 2; \alpha_j, \beta_j, \gamma_j, j = 1, 2$ and g satisfy (240)–(258), respectively.

For example, if we substitute (240) into (236) and (238), we obtain the following solutions of Eq. (235).

$$u_1(x,t) = \sqrt[2/p]{\frac{k_0 + 2\cosh_{0,0,1}(\alpha_1 x + \beta_1 t + \gamma_1)}{2\cosh_{h_1,h_{-1},1}(\alpha_2 x - 4\frac{\beta\alpha_2{}^3 t}{p^2} + \gamma_2)}}\,, \tag{260}$$

where $\cosh_{0,0,1}(\alpha_1 x + \beta_1 t + \gamma_1)$ and $\cosh_{h_1,h_{-1},1}(\alpha_2 x - 4(\beta\alpha_2{}^3 t)/p^2 + \gamma_2)$ are all the generalized hyperbolic cosine function, here $h_1 = (p^2\alpha k_0{}^2)/(8\alpha_2{}^2\beta h_{-1}(3p + 2 + p^2))$, and $\alpha, \beta, \alpha_2, \gamma_2, \alpha_2, k_0, \alpha_1, \beta_1, \gamma_1, k_0$ are any constants.

Remark 13 When $m_1, m_2, n_1, n_2 = 1, 2, \ldots$, we can obtain a family of N solitary-like solutions of Eq. (235).

In summary, we suggest two new generic transformations, a new Exp–N solitary-like method and its algorithm, to find new non-traveling waves solutions and traveling waves solutions of higher dimensional NLPDEs by our transformations. The suggested method is more powerful than the method proposed by Dr. He [31] to seek more exact solutions of higher dimensional NLPDEs. The $(1+1)$-dimensional generalized KdV equation is chosen to illustrate our algorithm such that new exact solutions are found.

The Exp-Bäcklund Transformation Method and Its Application in (1 + 1)-Dimensional KdV Equation

In this section, a new method and its algorithm, which is called Exp-Bäcklund transformation method, is presented to find more exact solutions of NLPDEs based on the idea of the Exp-function method [31]. We choose the $(1+1)$-dimensional KdV equations to illustrate the effectiveness and convenience of our algorithm. As a result, many new solutions are obtained.

Summary of the Exp-Bäcklund Transformation Method

In the following we would like to outline the main steps of our method:

Step 1. For a given NLEEs, with $u_i(t, x_1, x_2, \ldots, x_{n-1})$ $(i = 1, 2, \ldots, n)$ in n independent variables $t, x_1, x_2, \ldots, x_{n-1}$,

$$F_j(u_1, \ldots, u_n, u_{1,t}, \ldots, u_{n,t}, u_{1,x_1}, \ldots, u_{n,x_{n-1}}, u_{1,tt}, \ldots, u_{n,tt}, u_{1,tx_1}, \ldots u_{n,tx_{n-1}}, \ldots) = 0\,, \tag{261}$$

where $j = 1, 2, \ldots, n$.

We obtain an auto-Bäcklund transformation of Eqs. (261) using our method of constructing auto-Bäcklund transformation (if it exist)

$$u_i(t, x_1, \ldots, x_{n-1}) = \sum_{j=0}^{m_i} u_{ij}(t, x_1, \ldots, x_{n-1}) f^{j-m_i}(t, x_1, \ldots, x_{n-1}) , \quad i = 1, 2, \ldots, n . \quad (262)$$

where $u_{ij}(t, x_1, \ldots, x_{n-1}), i = 1, 2, \ldots, n; j = 0, 1, 2, \ldots, m_i - 1$ are differentiable functions which have already been obtained. But $f(t, x_1, \ldots, x_{n-1})$ are any differentiable functions, $u_{i,m_i}, i = 1, 2, \ldots, n$ are the seed solutions of Eq. (261), and $m_i, i = 1, 2, \ldots, n$ are positive integers which have already been obtained.

Step 2. Substituting (262) into Eq. (261), and setting the coefficients of these terms $1/(f^k(t, x, y)), k = 0, 1, 2, \ldots$ to zero yields a set of over-determined partial differential equations with respect to $f(t, x, y), u_{i,m_i}, i = 1, 2, \ldots, n$.

Step 3. We consider the variable in the form

$$\eta_i = p_{i0}(t) + \sum_{k=1}^{n-1} p_{ik}(t, x_1, \ldots, x_{k-1}) \cdot x_k , \quad i = 1, 2 \quad (263)$$

and make the transformation

$$f(t, x_1, \ldots, x_{n-1}) = f(\eta_1, \eta_2) . \quad (264)$$

Then we express the $f(t, x_1, \ldots, x_{n-1})$ in (262) and (264) to be the following forms

$$f(\eta_1, \eta_2) = \frac{\sum_{n=-c}^{d} a_n exp(n\eta_1)}{\sum_{m=-p}^{q} b_m exp(m\eta_2)} , \quad (265)$$

where $\sum_{k=0}^{n-1} p_k \neq 0$, c, d, p and q are any positive integers, $a_n, n = -c, \ldots, d$, $b_m, m = -p, \ldots, q$, and $p_{ik}, k = 0, 1, 2, \ldots, n-1$are functions or constants to be determined later.

Step 4. If we substitute (262)–(265) into the over-determined partial differential equations in Step 2 and yield a set of over-determined partial differential equations with respect to $exp(k\eta_i), i = 1, 2; k = 0, \pm 1, \pm 2, \ldots$, $a_n, n = -c, \ldots, d$, $b_m, m = -p, \ldots, q$, $u_{i,m_i}, i = 1, 2, \ldots, n$, and $p_{ik}, k = 0, 1, 2, \ldots, n-1$. Then we collect the coefficients of the polynomials with respect to $exp(k\eta_i), i = 1, 2; k = 0, \pm 1, \pm 2, \ldots$ and set each coefficient to zero, we will get a system of over-determined partial differential of $a_n, n = -c, \ldots, d$, $b_m, m = -p, \ldots, q$, and $p_{ik}, k = 0, 1, 2, \ldots, n-1$, $u_{i,m_i}, i = 1, 2, \ldots, n$.

Step 5. We solve the over-determined partial differential in Step 4 and obtain $a_n, n = -c, \ldots, d$, $b_m, m = -p, \ldots, q$, $u_{i,m_i}, i = 1, 2, \ldots, n$, and $p_{ik}, k = 0, 1, 2, \ldots, n-1$.

Step 6. Substituting the results obtained in Step 5 into (262)–(265), then we can obtain a family of rational solutions of Exp-functions as follows

$$u_i(t, x_1, \ldots, x_{n-1}) = \sum_{j=0}^{m_i} u_{ij}(t, x_1, \ldots, x_{n-1}) \left(\frac{\sum_{n=-c}^{d} a_n exp(np_{10}(t) + n\sum_{k=1}^{n-1} p_{1k}(t, x_1, \ldots, x_{k-1}) \cdot x_k)}{\sum_{m=-p}^{q} b_m exp(mp_{20}(t) + m\sum_{k=1}^{n-1} p_{2k}(t, x_1, \ldots, x_{k-1}) \cdot x_k)} \right)^{j-m_i} , \quad i = 1, 2, \ldots, n , \quad (266)$$

where c, d, p and q are any positive integers.

The Application of the Exp-Bäcklund Transformation Method in (1 + 1)-Dimensional KdV Equation

In this section, we will make use of our Exp-Bäcklund transformation method [40,46] and symbolic computation to find the exact solutions of the following (1 + 1)-dimensional KdV equation [1,2,3,4]

$$u_t(x, t) + \alpha u(x, t) u_x(x, t) + \beta u_{xxx}(x, t) = 0 , \quad (267)$$

where the coefficient of the nonlinear term α and the coefficient of the dispersive term β is independent of x and t.

Set the auto-Bäcklund transformation of Eq. (267) to have the following form:

$$u(x, t) = \sum_{j_1=0}^{m} u_j(x, t) f^{j-m}(x, t) , \quad (268)$$

where $f(x, t)$ and $u_j(x, t), j = 0, 1, 2, \ldots, m-1$ are all differential functions to be determined later, and $u_m(x, t)$ is the trivial seed solution of Eq. (267).

By balancing the highest order linear term and nonlinear terms in Eq. (267), we get $m = 2$ and (268) has the

following formal

$$u(x,t) = u_0(x,t)f^{-2}(x,t)+u_1(x,t)f^{-1}(x,t)+u_2(x,t), \tag{269}$$

We take the trivial seed solution as

$$u_2 = u_2(x,t)\,. \tag{270}$$

With the aid of Maple symbolic computation software, substituting (269) and (270) into (267), and collecting all terms with $f^{-i}(x,t)$, $i = 0,1,2,\ldots$, we obtain

$$\begin{aligned}
&[-2\alpha u_0^2(x,t)f_x(x,t) - 24\beta u_0(x,t)f^3(x,t)]\frac{1}{f^5}\\
&+\{-\beta[-18u_0(x,t)f_x(x,t)f_{xx}(x,t) - 18u_{0x}(x,t)f_x^2(x,t)\\
&\qquad + 6u_1(x,t)f_x^3(x,t)]\\
&\quad + \alpha u_0(x,t)[u_{0x}(x,t) - u_1(x,t)f_x(x,t)]\\
&\quad - 2\alpha u_1(x,t)u_0(x,t)f_x(x,t)\}\frac{1}{f^4}\\
&+\{-\beta[-6u_1(x,t)f_x(x,t)f_{xx}(x,t)\\
&\qquad + 6u_{0xx}(x,t)f_x(x,t) + 2u_0(x,t)f_{xxx}(x,t)\\
&\qquad + 6u_{0x}(x,t)f_{xx}(x,t) - 6u_{1x}(x,t)f_x^2(x,t)]\\
&\quad - 2u_0(x,t)f_t(x,t) + \alpha u_0(x,t)u_{1x}(x,t)\\
&\quad + \alpha u_1(x,t)[u_{0x}(x,t) - u_1(x,t)f_x(x,t)]\\
&\quad - 2\alpha u_2(x,t)u_0(x,t)f_x(x,t)\}\frac{1}{f^3}\\
&+\{-\beta[-u_{0xxx}(x,t) + 3u_{1x}(x,t)f_{xx}(x,t)\\
&\qquad + 3u_{1xx}(x,t)f_x(x,t) + u_1(x,t)f_{xxx}(x,t)]\\
&\quad + \alpha u_0(x,t)u_{2x}(x,t) + \alpha u_1(x,t)u_{1x}(x,t)\\
&\quad + \alpha u_2(x,t)[u_{0x}(x,t) - u_1(x,t)f_x(x,t)]\\
&\quad + u_{0t}(x,t) - u_1(x,t)f_t(x,t)\}\frac{1}{f^2}\\
&+(u_{1t}(x,t) + \alpha u_1(x,t)u_{2x}(x,t) + \alpha u_2(x,t)u_{1x}(x,t)\\
&+\beta u_{1xxx}(x,t))\frac{1}{f} + u_{2t}(x,t) + \alpha u_2(x,t)u_{2x}(x,t)\\
&+\beta u_{2xxx}(x,t) = 0\,,
\end{aligned} \tag{271}$$

Setting the coefficient of f^{-5} in (271) to be zero, we obtain a differential equation

$$-2\alpha u_0^2(x,t)f_x(x,t) - 24\beta u_0(x,t)f^3(x,t) = 0\,, \tag{272}$$

which has solution

$$u_0(x,t) = -\frac{12\beta}{\alpha}f_x^2(x,t)\,. \tag{273}$$

Setting the coefficient of f^{-4}in (271) to be zero, we obtain a differential equation

$$\begin{aligned}
&-\beta[-18u_0(x,t)f_x(x,t)f_{xx}(x,t)\\
&\qquad - 18u_{0x}(x,t)f_x^2(x,t) + 6u_1(x,t)f_x^3(x,t)]\\
&+\alpha u_0(x,t)[u_{0x}(x,t) - u_1(x,t)f_x(x,t)]\\
&-2\alpha u_1(x,t)u_0(x,t)f_x(x,t) = 0\,,
\end{aligned} \tag{274}$$

we can get the following expressions:

$$u_1(x,t) = \frac{12\beta}{\alpha}f_{xx}(x,t)\,. \tag{275}$$

By (273), (275), (270) and (269), we obtain an auto-Bäcklund transformation of Eq. (267) as follows

$$u(x,t) = -\frac{12\beta}{\alpha}\left(\frac{f_{xx}(x,t)}{f(x,t)} - \frac{f_x^2(x,t)}{f^2(x,t)}\right) + u_2(x,t) \tag{276}$$

where $u_2(x,t)$ is a seed solution of Eq. (267), $f(x,t)$ is any function that satisfies Eq. (272), Eq. (274), and the following equations

$$\begin{aligned}
&-\beta[-6u_1(x,t)f_x(x,t)f_{xx}(x,t) + 6u_{0xx}(x,t)f_x(x,t)\\
&\qquad + 2u_0(x,t)f_{xxx}(x,t) + 6u_{0x}(x,t)f_{xx}(x,t)\\
&\qquad - 6u_{1x}(x,t)f_x^2(x,t)]\\
&-2u_0(x,t)f_t(x,t) + \alpha u_0(x,t)u_{1x}(x,t)\\
&+\alpha u_1(x,t)[u_{0x}(x,t) - u_1(x,t)f_x(x,t)]\\
&-2\alpha u_2(x,t)u_0(x,t)f_x(x,t) = 0\,,\\
&-\beta[-u_{0xxx}(x,t) + 3u_{1x}(x,t)f_{xx}(x,t)\\
&\qquad + 3u_{1xx}(x,t)f_x(x,t) + u_1(x,t)f_{xxx}(x,t)]\\
&+\alpha u_0(x,t)u_{2x}(x,t) + \alpha u_1(x,t)u_{1x}(x,t)\\
&+\alpha u_2(x,t)[u_{0x}(x,t) - u_1(x,t)f_x(x,t)] + u_{0t}(x,t)\\
&-u_1(x,t)f_t(x,t) = 0\,,\\
&u_{1t}(x,t) + \alpha u_1(x,t)u_{2x}(x,t)\\
&+\alpha u_2(x,t)u_{1x}(x,t) + \beta u_{1xxx}(x,t) = 0\,.
\end{aligned} \tag{277}$$

We take $f(x,t)$ to be of the following form

$$f(x,t) = \frac{\sum_{i=-m_1}^{m_2} k_i(t)e^{i(p_1(t)x+q_1(t))}}{\sum_{j=-n_1}^{n_2} h_j(t)e^{j(p_2(t)x+q_2(t))}}\,, \tag{278}$$

where n_1, n_2, m_1 and m_2 are any positive integers, $p_l(t), q_l(t), l = 1,2; k_i(t), i = -m_1,\ldots,m_2$ and $h_j(t)$, $j = -n_1,\ldots,n_2$ are the coefficients to be determined later.

By substituting (278) into the given Eq. (277) with (273) and (275), and then collecting the coefficients of the polynomials of $e^{i(p_1(t)x+q_1(t))}$ and $e^{j(p_2(t)x+q_2(t))}$, $i, j = 0, 1, 2, \ldots$, then setting each coefficient to zero, we will get a system of over-determined partial differential equations with respect to $k_i(t)$, $i = -m_1, \ldots, m_2$; $h_j(t)$, $j = -n_1, \ldots, n_2$; $p_k(t), q_k(t)$, $k = 1, 2$ and $u_2(x, t)$. We solve the over-determined partial differential equations and obtain many families of solutions of $k_i(t), i = -m_1, \ldots, m_2; h_j(t)$, $j = -n_1, \ldots, n_2; p_k(t), q_k(t), k = 1, 2$ and $u_2(x, t)$. Finally, substituting them into (276) with (278), we can obtain the following solutions of Eq. (267)

$$
\begin{aligned}
u = {} & \\
& - \frac{12\beta\left(-\sum_{i=-m_1}^{m_2} k_i(t) i p_1(t) e^{i\xi_1} \sum_{j=-n_1}^{n_2} h_j(t) e^{j\xi_2} + \sum_{i=-m_1}^{m_2} k_i(t) e^{i\xi_1} \sum_{j=-n_1}^{n_2} h_j(t) j p_2(t) e^{j\xi_2}\right)^2}{\alpha\left(\sum_{j=-n_1}^{n_2} h_j(t) e^{j\xi_2}\right)^2 \left(\sum_{i=-m_1}^{m_2} k_i(t) e^{i\xi_1}\right)^2} \\
& - \frac{12\beta {p_1}^2(t)\left(\sum_{i=-m_1}^{m_2} k_i(t) i e^{i\xi_1}\right)^2}{\alpha\left(\sum_{i=-m_1}^{m_2} k_i(t) e^{i\xi_1}\right)^2} \\
& + \frac{12\beta {p_1}^2(t)\sum_{i=-m_1}^{m_2} k_i(t) i^2 e^{i\xi_1}}{\alpha\sum_{i=-m_1}^{m_2} k_i(t) e^{i\xi_1}} \\
& + \frac{12\beta {p_2}^2(t)\left(\sum_{j=-n_1}^{n_2} h_j(t) j e^{j\xi_2}\right)^2}{\left(\sum_{j=-n_1}^{n_2} h_j(t) e^{j\xi_2}\right)^2 \alpha} \\
& - \frac{12\beta {p_2}^2(t)\sum_{j=-n_1}^{n_2} h_j(t) j^2 e^{j\xi_2}}{\sum_{j=-n_1}^{n_2} h_j(t) e^{j\xi_2}\alpha} + u_2(t)\,,
\end{aligned}
\tag{279}
$$

where n_1, n_2, m_1 and m_2 are any positive integers, and $\xi_1 = p_1(t)x + q_1(t)$ and $\xi_2 = p_2(t)x + q_2(t)$.

For example, when we take $n_1 = 1, n_2 = 1, m_1 = 1$ and $m_2 = 1$, substituting (278) into the given Eq. (277) with (273) and (275), and then collecting the coefficients of the polynomials of $e^{i(p_1(t)x+q_1(t))}$ and $e^{j(p_2(t)x+q_2(t))}$, i, $j = 0, 1, 2, \ldots$, then setting each coefficient to zero, we will get a system of over-determined partial differential equations with respect to $p_l(t), q_l(t), l = 1, 2; k_i(t), i = -1, 0, 1$ and $h_j(t), j = -1, 0, 1$ as follows

$$
\begin{aligned}
& - p_2(t)k_0(t)h_1^4(t)k_{-1t}(t) \\
& - p_2(t)k_{-1}(t)h_1^4(t)k_{0t}(t) \\
& + 4\beta p_2(t)k_0(t)h_1^4(t)p_1^3(t)k_{-1}(t) \\
& + 2p_2(t)k_{-1}(t)h_1^3(t)k_0(t)h_{1t}(t) \\
& + p_2(t)k_0(t)h_1^4(t)k_{-1}(t)q_{1t}(t) \\
& + p_1(t)k_{-1}(t)h_1^4(t)k_0(t)q_{2t}(t) \\
& + 2u_2(x,t)\alpha p_1(t)k_{-1}(t)h_1^4(t)p_2(t)k_0(t) \\
& + 2u_2(x,t)\alpha p_2^2(t)k_{-1}(t)h_1^4(t)k_0(t) \\
& + 2p_2(t)k_{-1}(t)h_1^4(t)k_0(t)q_{2t}(t) \\
& + p_1(t)k_{-1}(t)h_1^3(t)k_0(t)h_{1t}(t) \\
& - p_1(t)k_{-1}(t)h_1^4(t)k_{0t}(t) \\
& + 4\beta p_1(t)k_{-1}(t)h_1^4(t){p_2}^3(t)k_0(t) \\
& + 2\beta p_2^4(t)k_{-1}(t)h_1^4(t)k_0(t) \\
& + 6\beta p_2^2(t)k_0(t)h_1^4(t)p_1^2(t)k_{-1}(t) = 0\,,
\end{aligned}
$$

$$
\begin{aligned}
& p_1(t)k_1(t)h_1^4(t)k_{0t}(t) - 4\beta p_2(t)k_0(t)h_1^4(t)p_1^3(t)k_1(t) \\
& - 4\beta p_1(t)k_1(t)h_1^4(t)p_2^3(t)k_0(t) \\
& + 2p_2(t)k_0(t)h_1^4(t)k_1(t)q_{2t}(t) + 2\beta p_2^4(t)k_1(t)h_1^4(t)k_0(t) \\
& + 2u_2(x,t)\alpha p_2^2(t)k_0(t)h_1^4(t)k_1(t) \\
& - 2u_2(x,t)\alpha p_1(t)k_1(t)h_1^4(t)p_2(t)k_0(t) \\
& - p_2(t)k_1(t)h_1^4(t)k_{0t}(t) + 6\beta {p^2}_2(t)k_0(t)h_1^4(t)p_1^2k_1(t) \\
& + 2p_2(t)k_0(t)h_1^3(t)k_1(t)h_{1t}(t) \\
& - p_1(t)k_1(t)h_1^3(t)k_0(t)h_{1t}(t) - p_2(t)k_0(t)h_1^4(t)k_1(t)q_{1t}(t) \\
& - p_2(t)k_0(t)h_1^4(t)k_{1t}(t) \\
& - p_1(t)k_1(t)h_1^4(t)k_0(t)q_{2t}(t)4\beta p_1(t)k_1(t)h_{-1}^4(t)p_2^3(t)k_0(t) \\
& + p_2(t)k_0(t)h_{-1}^4(t)k_{1t}(t) + p_2(t)k_0(t)h_{-1}^4(t)k_1(t)q_{1t}(t) \\
& + 2p_2(t)k_0(t)h_{-1}^4(t)k_1(t)q_{2t}(t) \\
& + p_1(t)k_1(t)h_{-1}^4(t)k_0(t)q_{2t}(t) \\
& + 6\beta p_2^2(t)k_0(t)h_{-1}^4(t)p_1^2(t)k_1(t) \\
& + 2u_2(x,t)\alpha p_1(t)k_1(t)h_{-1}^4(t)p_2(t)k_0(t) \\
& - p_1(t)k_1(t)h_{-1}^3(t)k_0(t)h_{-1t}(t) \\
& + 2u_2(x,t)\alpha p_2^2(t)k_0(t)h_{-1}^4(t)k_1(t) \\
& - 2p_2(t)k_0(t)h_{-1}^3(t)k_1(t)h_{-1t}(t) \\
& + 2\beta p_2^4(t)k_1(t)h_{-1}^4(t)k_0(t) \\
& + 4\beta p_2(t)k_0(t)h_{-1}^4(t)p_1^3k_1(t) + p_1(t)k_1(t)h_{-1}^4(t)k_{0t}(t) \\
& + p_2(t)k_1(t)h_{-1}^4(t)k_{0t}(t) = 0\,, \qquad (280) \\
& \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned}
$$

We solve the over-determined partial differential equations and obtain the following many families of solutions:

Case 1

$$u_1(x,t) = -\frac{12\beta(k_{-1}(t))^2(p_1(t))^2(e^{-p_1(t)x-q_1(t)})^2}{\alpha(k_{-1}(t)e^{-p_1(t)x-q_1(t)}+C_1)^2} + \frac{12\beta k_{-1}(t)(p_1(t))^2 e^{-p_1(t)x-q_1(t)}}{\alpha(k_{-1}(t)e^{-p_1(t)x-q_1(t)}+C_1)} + u_2(t)\,, \quad (281)$$

where $q_1(t) = \int \frac{\frac{d}{dt}k_{-1}(t)-(p_1(t))^3 k_{-1}(t)\beta - p_1(t)u_2(t)\alpha k_{-1}(t)}{k_{-1}(t)}dt + C_3$, $k_{-1}(t), p_1(t)$ are any functions of t, and α, β, C_i, $i = 1, 3$ are any constants.

Case 2

$$u_2(x,t) = -\frac{12\beta(C_1p_2(t)+k_{-1}(t)(p_2(t)-p_1(t))e^{-\xi_1} + C_2(p_2(t)+p_1(t))e^{\xi_1})^2}{\alpha(k_{-1}(t)e^{-\xi_1}+C_1+C_2e^{\xi_1})^2} + \frac{12\beta(C_1(p_2(t))^2 + C_2(p_2(t)+p_1(t))^2e^{\xi_1} + (p_1(t)-p_2(t))^2k_{-1}(t)e^{-\xi_1})}{\alpha(k_{-1}(t)e^{-\xi_1}+C_1+C_2e^{\xi_1})}\,, \quad (282)$$

where $\xi_1 = p_1(t)x + q_1(t)$, $k_{-1}(t), q_1(t), p_1(t)$ are any functions of t, and α, β, C_i, $i = 1, 2$ are any constants.

Case 3

$$u_3(x,t) = -\frac{12\beta(k_{-1}(t)(p_2(t)-p_1(t))e^{-\xi_1} + k_1(t)(p_2(t)+p_1(t))e^{\xi_1} + C_1p_2(t))^2}{\alpha(k_{-1}(t)e^{-\xi_1}+C_1+k_1(t)e^{\xi_1})^2} + \frac{12\beta(C_1(p_2(t))^2 + k_1(t)(p_2(t)+p_1(t))^2e^{\xi_1} + (p_1(t)-p_2(t))^2k_{-1}(t)e^{-\xi_1})}{\alpha(k_{-1}(t)e^{-\xi_1}+C_1+k_1(t)e^{\xi_1})}\,, \quad (283)$$

where $\xi_1 = p_1(t)x + q_1(t)$, $k_{-1}(t), k_1(t), q_1(t), p_1(t)$ are any functions of t, and α, β, C_1 are any constants.

Case 4

$$u_4(x,t) = -\frac{12\beta(p_1(t))^2(k_1(t))^2(e^{p_1(t)x+q_1(t)})^2}{\alpha(C_1+k_1(t)e^{p_1(t)x+q_1(t)})^2} + \frac{12\beta(p_1(t))^2k_1(t)e^{p_1(t)x+q_1(t)}}{\alpha(C_1+k_1(t)e^{p_1(t)x+q_1(t)})} + u_2(t)\,, \quad (284)$$

where $q_1(t) = -\int \frac{\beta(p_1(t))^3k_1(t)+u_2(t)\alpha p_1(t)k_1(t)+\frac{d}{dt}k_1(t)}{k_1(t)}dt + C_3$, $k_1(t), p_1(t)$ are the functions of t, and α, β, C_i, $i = 1, 3$ are any constants.

Case 5

$$u_5(x,t) = \frac{2C_1h_0(t)(u_2(t)\alpha - 6\beta(p_2(t))^2)e^{-\xi_2} + u_2(t)\alpha C_1^2e^{-2\xi_2} + u_2(t)\alpha(h_0(t))^2}{(C_1e^{-\xi_2}+h_0(t))^2\alpha}\,, \quad (285)$$

where $\xi_2 = p_2(t)x + q_2(t)$, $h_0(t), p_2(t), q_2(t)$ are the functions of t, and α, β, C_1 are any constants.

Case 6

$$u_6(x,t) = -\frac{2C_1h_0(t)(6\beta(p_2(t))^2 - u_2(t)\alpha)e^{-\xi_2} + 2C_1h_1(t)(24\beta(p_2(t))^2 - u_2(t)\alpha)}{(C_1e^{-\xi_2}+h_0(t)+h_1(t)e^{\xi_2})^2\alpha} - \frac{2h_0(t)h_1(t)(6\beta(p_2(t))^2 - u_2(t)\alpha)e^{\xi_2} - u_2(t)\alpha((h_1(t))^2+C_1^2)e^{-2\xi_2} - u_2(t)\alpha(h_0(t))^2}{(C_1e^{-\xi_2}+h_0(t)+h_1(t)e^{\xi_2})^2\alpha}\,, \quad (286)$$

where $\xi_2 = p_2(t)x + q_2(t)$, $h_0(t), h_1(t), q_1(t), q_2(t)$ are the functions of t, and α, β, C_1 are any constants.

Case 7

$$u_7(x,t) = \frac{2h_0(t)h_1(t)(u_2(t)\alpha - 6\beta(p_2(t))^2)e^{\xi_2} + 2h_0(t)h_{-1}(t)(u_2(t)\alpha - 6\beta(p_2(t))^2)e^{-\xi_2}}{(h_{-1}(t)e^{-\xi_2}+h_0(t)+h_1(t)e^{\xi_2})^2\alpha} + \frac{u_2(t)\alpha((h_{-1}(t))^2e^{-2\xi_2}+(h_1(t))^2e^{2\xi_2}) - 48\beta h_{-1}(t)(p_2(t))^2h_1(t) + u_2(t)\alpha(2h_{-1}(t)h_1(t)+(h_0(t))^2)}{(h_{-1}(t)e^{-\xi_2}+h_0(t)+h_1(t)e^{\xi_2})^2\alpha}\,, \quad (287)$$

where $\xi_2 = p_2(t)x + q_2(t)$, $h_{-1}(t), h_0(t), h_1(t), q_1(t), q_2(t)$ are the functions of t, and α, β are any constants.

Case 8

$$u_8(x,t) = \frac{2C_1k_0(t)(6\beta(p_1(t))^2 + u_2(t)\alpha)e^{-\xi_1} + u_2(t)\alpha C_1^2e^{-2\xi_1} + u_2(t)\alpha(k_0(t))^2}{\alpha(C_1e^{-\xi_1}+k_0(t))^2}\,, \quad (288)$$

where $\xi_1 = p_1(t)x + q_1(t)$, $k_0(t), q_1(t), q_2(t)$ are the functions of t, and α, β, C_1 are any constants.

Case 9

$$u_9(x,t) = \frac{u_2(t)\alpha((k_0(t))^2 + 2C_1k_1(t)) + 48\beta C_1(p_1(t))^2k_1(t) + u_2(t)\alpha(k_1(t))^2e^{2\xi_1}}{\alpha(C_1e^{-\xi_1}+k_0(t)+k_1(t)e^{\xi_1})^2} + \frac{2k_0(t)k_1(t)(6\beta(p_1(t))^2 + u_2(t)\alpha)e^{\xi_1} + 2C_1k_0(t)(6\beta(p_1(t))^2 + u_2(t)\alpha)e^{-\xi_1} + u_2(t)\alpha C_1^2e^{-2\xi_1}}{\alpha(C_1e^{-\xi_1}+k_0(t)+k_1(t)e^{\xi_1})^2}\,, \quad (289)$$

where $\xi_1 = p_1(t)x + q_1(t)$, $k_0(t), k_1(t), q_1(t), q_2(t)$ are the functions of t, and α, β, C_1 are any constants.

Case 10

$$u_{10}(x,t) = -\frac{12\beta(p_1(t))^2(k_1(t))^2(e^{p_1(t)x+q_1(t)})^2}{\alpha(k_0(t)+k_1(t)e^{p_1(t)x+q_1(t)})^2} + \frac{12\beta(p_1(t))^2k_1(t)e^{p_1(t)x+q_1(t)}}{\alpha(k_0(t)+k_1(t)e^{p_1(t)x+q_1(t)})} + u_2(t)\,, \quad (290)$$

where $q_1(t) = \int \frac{k_0(t)\frac{d}{dt}k_1(t) - k_1(t)\frac{d}{dt}k_0(t) + k_0(t)k_1(t)\cdot(p_1(t))^3\beta + u_2(t)\alpha p_1(t)k_1(t)k_0(t)}{k_0(t)k_1(t)}dt + C_2$, $k_0(t), k_1(t), q_2(t)$ are the functions of t, and α, β, C_2 are any constants.

Case 11

$$u_{11}(x,t) = -\frac{12\beta(k_{-1}(t)(p_2(t)-p_1(t))e^{-\xi_1} + k_1(t)(p_2(t)+p_1(t))e^{\xi_1} + k_0(t)p_2(t))^2}{\alpha(k_{-1}(t)e^{-\xi_1}+k_0(t)+k_1(t)e^{\xi_1})^2} + \frac{12\beta(k_0(t)(p_2(t))^2 + k_1(t)(p_2(t)+p_1(t))^2e^{\xi_1} + (p_1(t)-p_2(t))^2k_{-1}(t)e^{-\xi_1})}{\alpha(k_{-1}(t)e^{-\xi_1}+k_0(t)+k_1(t)e^{\xi_1})} + u_2(t)\,, \quad (291)$$

where $\xi_1 = p_1(t)x + q_1(t)$, $k_{-1}(t), k_0(t), k_1(t), q_1(t), q_2(t)$ are the functions of t, and α, β are any constants.

Case 12

$$u_{12}(x,t) = -\frac{12\beta(k_{-1}(t))^2(p_1(t))^2(e^{-p_1(t)x-q_1(t)})^2}{\alpha(k_{-1}(t)e^{-p_1(t)x-q_1(t)}+k_0(t))^2} + \frac{12k_{-1}(t)(p_1(t))^2e^{-p_1(t)x-q_1(t)}\beta}{\alpha(k_{-1}(t)e^{-p_1(t)x-q_1(t)}+k_0(t))} + u_2(t)\,, \quad (292)$$

where $q_1(t) = \int \frac{k_0(t)\frac{d}{dt}k_{-1}(t) - u_2(t)\alpha p_1(t)k_{-1}(t)k_0(t) - k_0(t)(p_1(t))^3k_{-1}(t)\beta - k_{-1}(t)\frac{d}{dt}k_0(t)}{k_0(t)k_{-1}(t)}dt + C_2$, $k_{-1}(t), k_0(t), q_2(t)$ are the functions of t, and α, β, C_2 are any constants.

Remark 14 From the above application, it is easily seen that our method is more powerful, more convenient, and more general than the method presented by He [31]. Firstly, our method is convenient to obtain the exact solutions of higher order and higher dimensional NLEEs because we do not use the homogeneous balance method to get the value of c, d, p, q in (265). Secondly, we can obtain more general exact solutions including non-traveling wave solutions and traveling wave solutions by using our method. For example, the solution (282) of Eq. (267) is a more general exact solution including the non-traveling wave solutions with constant coefficients, the non-traveling wave solutions with variable coefficients, the traveling wave solutions with constant coefficients, and the traveling wave solutions with variable coefficients of Eq. (267).

Example 1 The solution (282) of Eq. (267) includes abundant non-traveling wave solutions with variable coefficients. For example, when we take $C_1 = 1/2, C_2 = 1$, $p_2(t) = \cos(t^2), u_2(t) = 1/2, p_1(t) = \tanh^2(t), q_1(t) = \sin(t^4), k_{-1}(t) = \cosh(t), k_0(t) = \cos(t^2), \beta = 2, \alpha = 3$, we can obtain a non-traveling wave solution $u_{2,1}(x,t)$ with variable coefficients as follows:

$$u_{2,1}(x,t) = \frac{8\cos^2(t^2)(2\cosh(t)e^{-\xi_1}+1+2e^{\xi_1}) + 32\tanh^2(t)\cos(t^2)(e^{\xi_1}-\cosh(t)e^{-\xi_1}) + 16\tanh^4(t)(\cosh(t)e^{-\xi_1}+e^{\xi_1})}{2\cosh(t)e^{-\xi_1}+1+2e^{\xi_1}} - 8\frac{(2\tanh^2(t)(e^{\xi_1}-\cosh(t)e^{-\xi_1}) + \cos(t^2)(2\cosh(t)e^{-\xi_1}+1+2e^{\xi_1}))^2}{(2\cosh(t)e^{-\xi_1}+1+2e^{\xi_1})^2} + 1/2\,, \quad (293)$$

where $\xi_1 = \tanh^2(t)x + \sin(t^4)$.

Some figures of the long-term behavior of the $u_{2,1}(x,t)$ from $\{(x,t)||x| < 0.1, |t| < 0.1, x \in R, t \in R\}$ to $\{(x,t)||x| < 10^9, |t| < 10^9, x \in R, t \in R\}$ are shown by following Fig. 1 and Fig. 2. From Fig. 1 and Fig. 2, it is easy to find the global existence, asymptotic behaviors as $t \to +\infty$ and $|x| \to +\infty$ and the scattering properties of the solution $u_{2,1}(x,t)$ of Eq. (267). It is worth noting that the shape of $u_{2,1}(x,t)$ degenerates gradually from a set of waves to a set of solitons on a trigonal curve as the $|x|$ and $|t|$ in the $u_{2,1}(x,t)$ continue to increase.

Example 2 The solution (282) of Eq. (267) includes abundant non-traveling wave solutions with constant coefficients. For example, when we take $k_{-1}(t) = 1/2, \beta = 2, \alpha = 3, k_0(t) = 2, C_2 = 1, C_1 = 1/2, p_1(t) = 1/2, p_2(t) = 1/3, u_2(t) = 1/2$, and $q_1(t) = \sin(t^4)$, we can obtain a non-traveling wave solution $u_{2,2}(x,t)$ with constant coefficients as follows:

$$u_{2,2}(x,t) = -\frac{2(e^{-x/2-\sin(t^4)} - 10e^{x/2+\sin(t^4)} - 2)^2}{9(e^{-x/2-\sin(t^4)}+1+2e^{x/2+\sin(t^4)})^2} + \frac{8+100e^{x/2+\sin(t^4)}+2e^{-x/2-\sin(t^4)}}{9e^{-x/2-\sin(t^4)}+9+18e^{x/2+\sin(t^4)}} + 1/2\,. \quad (294)$$

Some figures of the long-term behavior of the $u_{2,2}(x,t)$ from $\{(x,t)||x| < 0.1, |t| < 0.1, x \in R, t \in R\}$

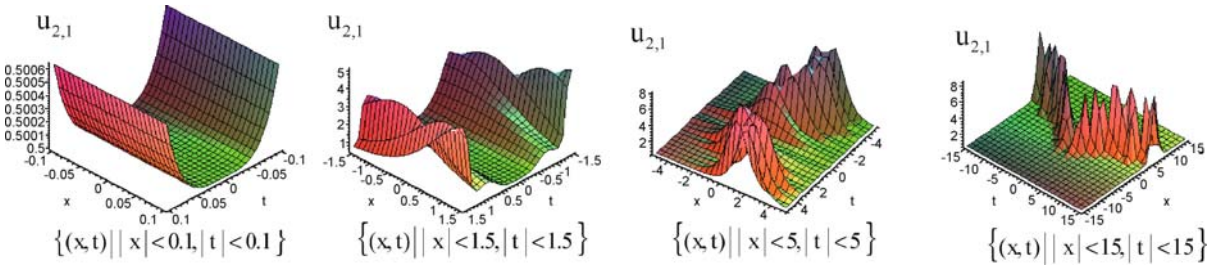

Korteweg–de Vries Equation (KdV), Different Analytical Methods for Solving the, Figure 1

The evolution of a non-traveling wave solution $u_{2,1}(x,t)$ with variable coefficients of Eq. (267) from $\{(x,t)||x| < 0.1, |t| < 0.1, x \in R, t \in R\}$ to $\{(x,t)||x| < 15, |t| < 15, x \in R, t \in R\}$

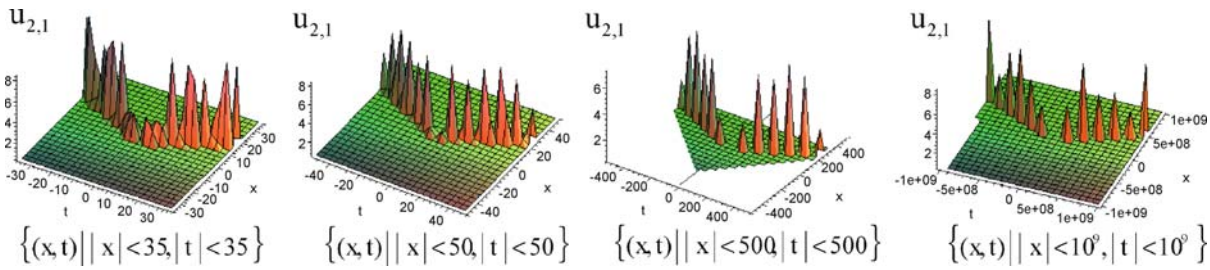

Korteweg–de Vries Equation (KdV), Different Analytical Methods for Solving the, Figure 2

The evolution of a non-traveling wave solution $u_{2,1}(x,t)$ with variable coefficients of Eq. (267) from $\{(x,t)||x| < 35, |t| < 35, x \in R, t \in R\}$ to $\{(x,t)||x| < 10^9, |t| < 10^9, x \in R, t \in R\}$

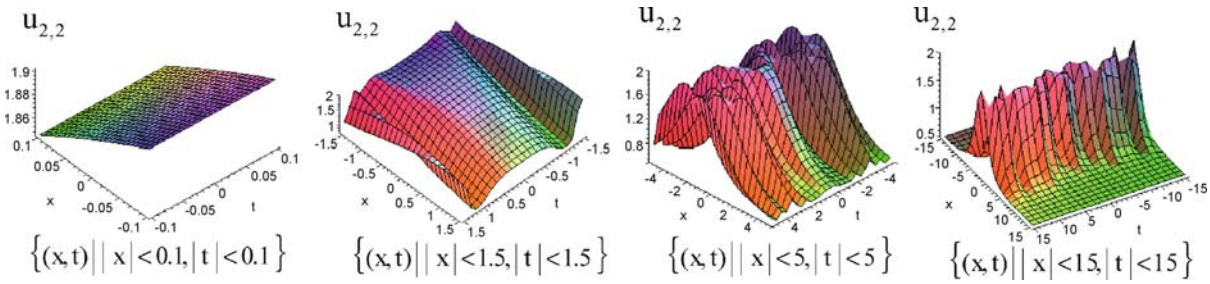

Korteweg–de Vries Equation (KdV), Different Analytical Methods for Solving the, Figure 3

The evolution of a non-traveling wave solution $u_{2,2}(x,t)$ with constant coefficients of Eq. (267) from $\{(x,t)||x| < 0.1, |t| < 0.1, x \in R, t \in R\}$ to $\{(x,t)||x| < 15, |t| < 15, x \in R, t \in R\}$

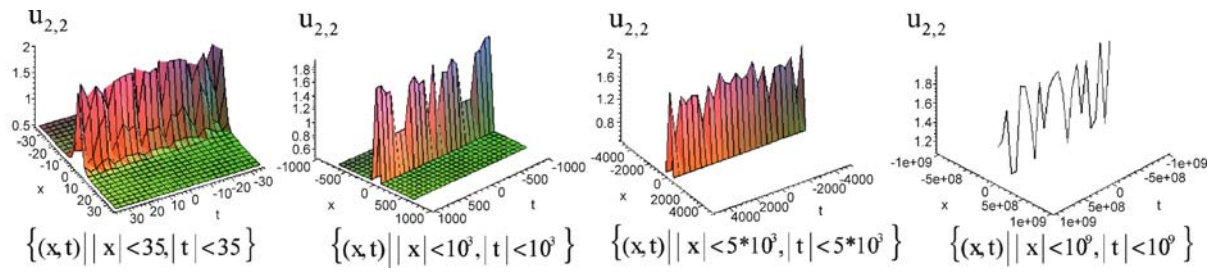

Korteweg–de Vries Equation (KdV), Different Analytical Methods for Solving the, Figure 4

The evolution of a non-traveling wave solution $u_{2,2}(x,t)$ with constant coefficients of Eq. (267) from $\{(x,t)||x| < 35, |t| < 35, x \in R, t \in R\}$ to $\{(x,t)||x| < 10^9, |t| < 10^9, x \in R, t \in R\}$

to $\{(x,t)||x|<10^9, |t|<10^9, x\in R, t\in R\}$ are shown by following Fig. 3 and Fig. 4.

Example 3 The solution (282) of Eq. (267) includes abundant traveling wave solutions with constant coefficients. For example, when $q_1(t)=\sin(t^4)$ in (294) is replaced by $q_1(t)=t/3$, we can obtain a traveling wave solution $u_{2,3}(x,t)$ with constant coefficients of Eq. (267) as follows:

$$u_{2,3}(x,t)=-\frac{2(\mathrm{e}^{-x/2-t/3}-10\mathrm{e}^{x/2+t/3}-2)^2}{9(\mathrm{e}^{-x/2-t/3}+1+2\mathrm{e}^{x/2+t/3})^2}+\frac{8+100\mathrm{e}^{x/2+t/3}+2\mathrm{e}^{-x/2-t/3}}{9\mathrm{e}^{-x/2-t/3}+9+18\mathrm{e}^{x/2+t/3}}+1/2\,. \quad (295)$$

Some figures of the long-term behavior of the $u_{2,3}(x,t)$ from $\{(x,t)||x|<0.1, |t|<0.1, x\in R, t\in R\}$ to $\{(x,t)||x|<423, |t|<423, x\in R, t\in R\}$ are shown by following Fig. 5 and Fig. 6. It is worth noting that the shape of the $u_{2,3}(x,t)$ degenerates gradually from a wave to a row of solitons on a straight line as the $|x|$ and $|t|$ in $u_{2,3}(x,t)$ continue to increase.

Example 4 The solution (282) of Eq. (267) includes abundant traveling wave solutions with variable coefficients. For example, when $p_1(t)=\tanh^2(t), q_1(t)=\sin(t^4)$ in (293) is replaced by $p_1(t)=1/2, q_1(t)=t/3$, we can obtain a traveling wave solution $u_{2,4}(x,t)$ with variable coefficients of Eq. (267) as follows:

$$u_{2,4}(x,t)=\frac{1+12\cosh(t)\mathrm{e}^{-x/2-t/3}+12\mathrm{e}^{x/2+t/3}+4\mathrm{e}^{x+2t/3}+72\cosh(t)+4\cosh^2(t)\mathrm{e}^{-x-2t/3}}{2(2\cosh(t)\mathrm{e}^{-x/2-t/3}+1+2\mathrm{e}^{x/2+t/3})^2}\,. \quad (296)$$

Some figures of the long-term behavior of the $u_{2,4}(x,t)$ from $\{(x,t)||x|<0.1, |t|<0.1, x\in R, t\in R\}$ to $\{(x,t)||x|<10^9, |t|<10^9, x\in R, t\in R\}$ are shown by following Fig. 7 and Fig. 8. From Fig. 7 and Fig. 8, it is easy to find the global existence, asymptotic behaviors as $t\to+\infty$ and $|x|\to+\infty$ and the scattering properties of the solution $u_{2,4}(x,t)$ of Eq. (267). It is worth noting that the shape of $u_{2,4}(x,t)$ degenerates gradually from a trigonal wave to a set of solitons on a trigonal curve as $|x|$ and $|t|$ in $u_{2,4}(x,t)$ continue to increase. But the plane has become a scraggly surface over $\{(x,t)||x|<10^9, |t|<10^9, x\in R, t\in R\}$, namely, the $u_{2,4}(x,t)$ appears a mild disorder.

Remark 15 From the above long-term behavior of the $u_{2,1}(x,t), u_{2,2}(x,t), u_{2,3}(x,t), u_{2,4}(x,t)$, we find easily that the stabilization of the traveling wave solutions $u_{2,3}(x,t)$ with constant coefficients is best as $t\to+\infty$ and $|x|\to+\infty$. In addition, we still discover that the longterm behavior of a solution, which includes the global existence, asymptotic behaviors as $t\to+\infty$ and $|x|\to+\infty$, and the scattering properties of the solution, depends mainly on the functions that constitute the solution and the selection of all arbitrary constants and the parameters in the solution. We choose other nonlinear evolution equations [39,40,45,47] to test and obtain similar results. Therefore, we devise a new guess as follows.

R-guess The long-term behavior of a solution for a given NLEEs (261), which includes the global existence, asymptotic behaviors as $t\to+\infty$ and $|x|\to+\infty$, and the scattering properties of the solution, depends mainly on the functions that constitute the solution and the selection of all arbitrary constants and parameters in the solution.

Future Directions

In recent years, directly searching for exact solutions of NLPDEs has become more and more attractive partly due to the availability of computer symbolic systems like Maple or mathematica which allow us to perform complicated and tedious algebraic calculation on a computer, as well as helping us to find new exact solutions of NLPDEs in mathematical physics. Although many powerful analytical methods for solving NLPDEs have been presented, the methods fail to satisfy the developmental needs in physics, mechanics, chemistry, biology, computer science, etc. There is much new work to be done on analytical methods for solving NLPDEs. The main future directions are as follows:

1. Researching new and uniform analytical methods based on the ideas of unification methods, algorithm realization and mechanization for solving NLPDEs (we have done some new work in Sect. "The Generalized Hyperbolic Function–Bäcklund Transformation Method and Its Application in the (2 + 1)-Dimensional KdV Equation" of this article).
2. Researching mechanization methods and developing their computer software for showing the long-playing traveling state of the exact solutions of NLPDEs (we have done some new work in [40]).
3. Developing and improving existing analytical methods, computer algebraic systems, and software so that analytical methods can be actualized with complete mechanization by their computer software at a familiar computer (we have done some new work in [39,40,41,42,43,44,45,46,47,48]).

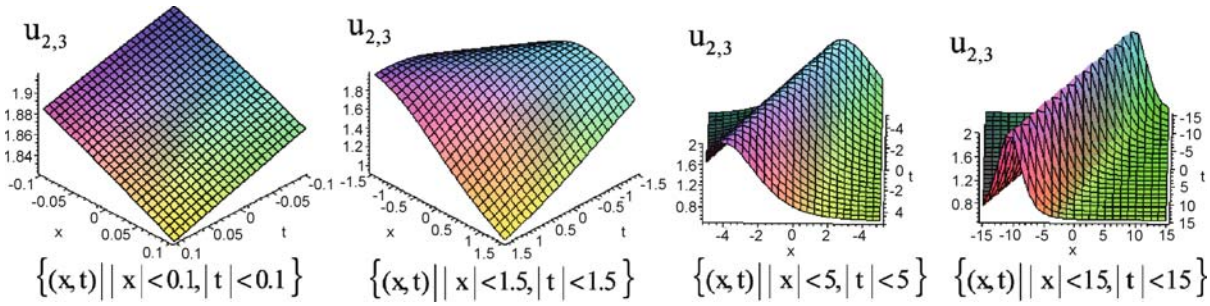

Korteweg–de Vries Equation (KdV), Different Analytical Methods for Solving the, Figure 5

The evolution of a traveling wave solution $u_{2,3}(x, t)$ with constant coefficients of Eq. (267) from $\{(x, t)||x| < 0.1, |t| < 0.1, x \in R, t \in R\}$ to $\{(x, t)||x| < 15, |t| < 15, x \in R, t \in R\}$

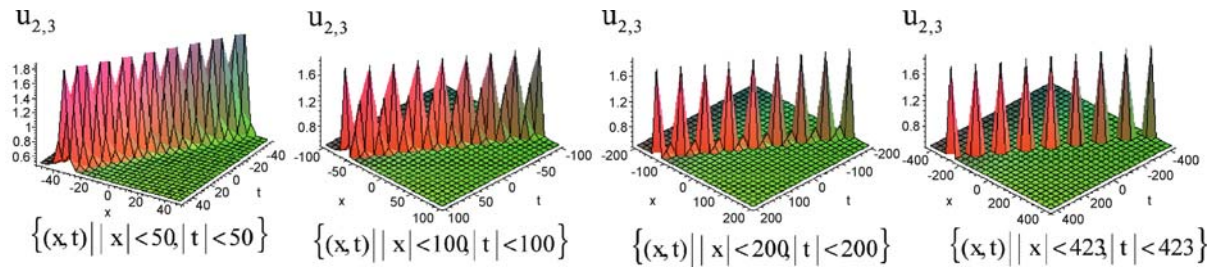

Korteweg–de Vries Equation (KdV), Different Analytical Methods for Solving the, Figure 6

The evolution of a traveling wave solution $u_{2,3}(x, t)$ with constant coefficients of Eq. (267) from $\{(x, t)||x| < 50, |t| < 50, x \in R, t \in R\}$ to $\{(x, t)||x| < 423, |t| < 423, x \in R, t \in R\}$

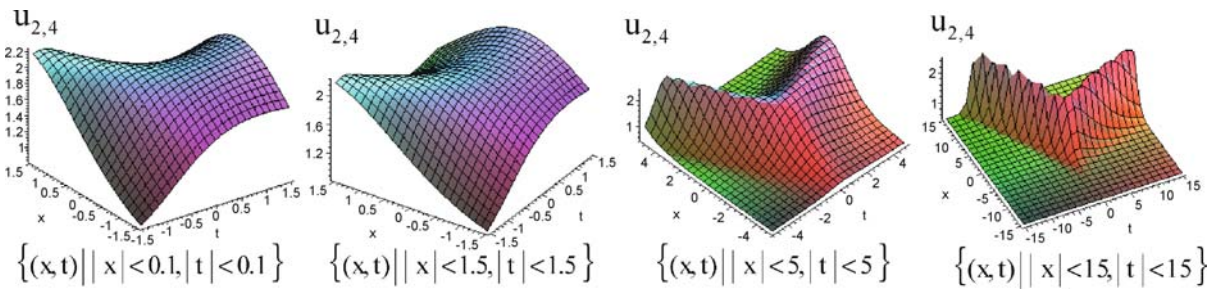

Korteweg–de Vries Equation (KdV), Different Analytical Methods for Solving the, Figure 7

The evolution of a traveling wave solution $u_{2,4}(x, t)$ with variable coefficients of Eq. (267) from $\{(x, t)||x| < 0.1, |t| < 0.1, x \in R, t \in R\}$ to $\{(x, t)||x| < 15, |t| < 15, x \in R, t \in R\}$

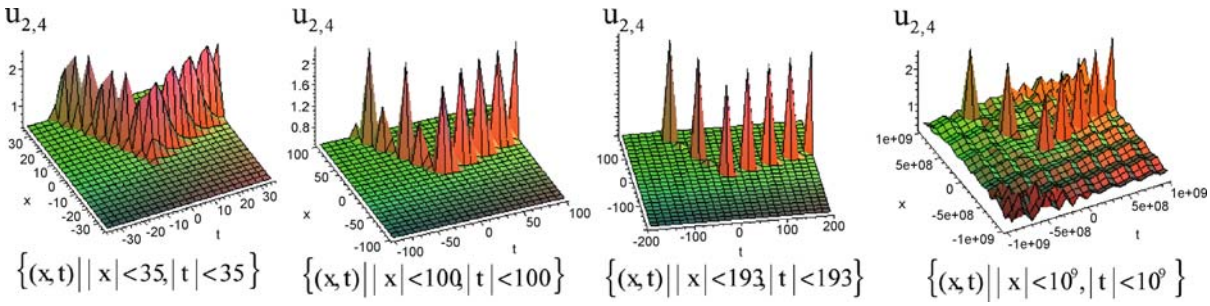

Korteweg–de Vries Equation (KdV), Different Analytical Methods for Solving the, Figure 8

The evolution of a traveling wave solution $u_{2,4}(x, t)$ with variable coefficients of Eq. (267) from $\{(x, t)||x| < 35, |t| < 35, x \in R, t \in R\}$ to $\{(x, t)||x| < 10^9, |t| < 10^9, x \in R, t \in R\}$

4. Researching new analytical and numerical complex methods and their computer software for solving those NLPDEs that can't make use of analytical methods. The complex methods can be actualized with complete mechanization by their computer software at a familiar computer (we have done some new work in [40,53]).
5. One can calculate the solutions of NLPDEs as easily as counting or addition by using a small calculator with the development of mathematics and computer science (we have done some new work in [39,40,41,42,43,44,45,46,47,48,52,53,54] and Ren 2055 and 2007, and Ren und Sun 2004).

Acknowledgment

The author would like to express his gratitude to all the people whose efforts contributed to the various editions of this paper. One reviewer who made useful recommendations for this paper is Professor M. Atef Helal at Mathematics Department, Cairo University, Egypt. The author is grateful to Professor M. Atef Helal for his time and valuable suggestions. The author thanks Editor-in-Chief Robert Meyers of the Encyclopedia of Complexity and System Science, Professor M.S. El Naschie at Frankfurt Institute for the advancement of Theoretical Physics, Frankfurt, Germany, Professor Hong-Qing Zhang at Department of Applied Mathematics, Dalian University of Technology, Dalian, PR China, and Professor Abdul-Majid Wazwaz at Department of Mathematics and Computer Science, Saint Xavier University, Chicago, USA, for their help. This work has been supported partially by a National Key Basic Research Project of China (2004 CB 318000).

Bibliography

Primary Literature

1. Korteweg DJ, de Vries G (1895) On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves. Philos Mag Ser 5 39:422–443
2. Khaters AH, El-Kalaawy H, Helal MA (1997) Two new classes of exact solutions for the KdV equation via Bäcklund transformations. Chaos Solitons Fractals 8(12):1901–1909
3. Helal MA (2001) A Chebyshev spectral method for solving Korteweg–de Vries equation with hydrodynamical application. Chaos Solitons Fractals 12:943–950
4. Brugarino T, Sciacca M (2008) A direct method to find solutions of some type of coupled Korteweg–de Vries equations using hyperelliptic functions of genus two. Phys Lett A 372(11):1836–1840
5. Leibovitch S (1970) Weakly nonlinear waves in rotating fluids. J Fluid Mech 42:803–825
6. RM Miura, CS Gardner, MD Kruskal (1968) Korteweg–de Vries equation and generalizations. II Existence of conservation laws and constants of motion. J Math Phys 9:1204–1209
7. MA Helal, MS Mehanna (2007) A comparative study between two different methods for solving the general Korteweg–de Vries equation (GKdV). Chaos Solitons Fractals 33:725–739
8. Hamdia S, Enrighta WH, Schiesserb WE, Gottliebc JJ (2005) Exact solutions and conservation laws for coupled generalized Korteweg–de Vries and quintic regularized long wave equations. Nonlinear Anal 63:1425–1434
9. Sarma J (2008) Solitary wave solution of higher-order Korteweg–de Vries equation. Chaos Solitons Fractals, in press
10. Karpman VI (1998) Evolution of radiating solitons described by the fifth-order Korteweg–de Vries type equations. Phys Lett A 244:394–396
11. Hassan MM (2004) Exact solitary wave solutions for a generalized KdV–Burgers equation. Chaos Solitons Fractals 19:1201–1206
12. Khare A, Cooper F (1993) One-parameter family of soliton solutions with compact support in a class of generalized Korteweg–de Vries equation. Phys Rev E 48(6):4843–4
13. Doďan K (2005) An application for the higher order modified KdV equation by decomposition method. Commun Nonlinear Sci Num Simul 10:693–702
14. Shen J, Xu W (2007) Travelling wave solutions in a class of generalized Korteweg–de Vries equation. Chaos Solitons Fractals 34:1299–1306
15. Clarke S, Malomed BA, Grimshaw R (2002) Dispersion management for solitons in a Korteweg–de Vries system. Chaos 12(1):8–15
16. Houria T (2007) Energy transfer in a dispersion-managed Korteweg–de Vries system. Math Comput Simul 76:283–292
17. Kaya D, Aassila M (2002) An application for a generalized KdV equation by the decomposition method. Physic Lett A 299:201–206
18. Shek ECM, Chow KW (2008) The discrete modified Korteweg–de Vries equation with non-vanishing boundary conditions: Interactions of solitons. Chaos Solitons Fractals 36:296–302
19. Zhang S (2007) A generalized auxiliary equation method and its application to the $(2+1)$-dimensional KdV equations. Appl Math Comput 188:1–6
20. Song L, Zhang H (2007) A new Korteweg–de Vries equation-based sub-equation method and its application to the $(2+1)$-dimensional Korteweg–de Vries equation. Appl Math Comput 187:1368–1372
21. Hirota R (1971) Exact solution of the Korteweg–de Vries equation for multiple collisions of solitons. Phys Rev Lett 27:1192–1194
22. Wadati M, Sanuki H, Konno K (1975) Relationships among inverse method, Bäcklund transformation and an infinite number of conservation laws. Prog Theor Phys 53:419–436
23. Weiss J, Tabor M, Carnevale G (1983) The Painlevel property for partial differential equations. J Math Phys 24:522–526
24. Fan EG, Zhang HQ (1998) A note on the homogeneous balance method. Phys Lett A 246:403–406
25. Wazwaz AM (2005) The tanh and the sine-cosine methods for the complex modified KdV and the generalized KdV equations. Comput Math Appl 45:1101–1112
26. Wang ML (1996) Exact solution for a compound KdV–Burgers equations. Phys Lett A 213:279–287
27. He JH (2005) Homotopy perturbation method for bifurcation of nonlinear problems. Int J Nonlinear Sci Numer Simul 6:207–208

28. He JH (2004) Variational principles for some nonlinear partial differential equations with variable coefficients. Chaos Solitons Fractals 19:847–851
29. He JH (2006) Some asymptotic methods for strongly nonlinear equations. Int J Modern Phys B 20:1141–1199
30. He JH (2006) Non-perturbative Methods for Strongly Nonlinear Problems, Dissertation. Verlag im Internet, Berlin
31. He JH, Wu XH (2006) Exp-function method for nonlinear wave equations. Chaos Solitons Fractals 30:700–708
32. Abassy TA, El-Tawil MA, Saleh HK (2004) The solution of KdV and mKdV equations using adomian pade approximation. Int J Nonlinear Sci Numer Simul 5:327–340
33. Liu SK, Fu ZT, Liu SD, Zhao Q (2001) Jacobi elliptic function expansion method and periodic wave solutions of nonlinear wave equations. Phys Lett A 289:69–74
34. Zhou YB, Wang ML, Wang YM (2003) Periodic wave solutions to a coupled KdV equations with variable coefficients. Phys Lett A 308:31–36
35. Chen Y, Yan ZY (2006) Weierstrass semi-rational expansion method and new doubly periodic solutions of the generalized Hirota–Satsuma coupled KdV system. Appl Math Comput 177:85–91
36. Chen Y, Wang Q (2006) A unified rational expansion method to construct a series of explicit exact solutions to nonlinear evolution equations. Appl Math Comput 177:396–409
37. Zhang S, Xia TC (2006) A further improved extended Fan sub-equation method and its application to the (3 + 1)-dimensional Kadomstev–Petviashvili equation. Phys Lett A 356:119–123
38. Xu GQ, Li ZB (2005) Exact traveling wave solutions of the Whitham–Broer–Kaup and Broer–Kaup–Kupershmidt equations. Chaos Solitons Fractals 24:549–556
39. Ren YJ, Zhang HQ (2006) New generalized hyperbolic function and Auto-Bäcklund transformation to find new exact solutions of the (2 + 1)-dimensional NNV equation. Phys Lett A 357:438–448
40. Ren YJ (2007) Some investigation on the methods for finding solutions of nonlinear evolution equations and realization of mathematics mechanization, Yujie Ren' Ph.D Dissertation. Dalian University of Technology, Dalian, p 39–50
41. Ren YJ, Zhang HQ (2006) A generalized F-expansion method to find abundant families of Jacobi Elliptic Function solutions of the (2 + 1)-dimensional Nizhnik–Novikov–Veselov equation. Chaos Solitons Fractals 27(4):959–979
42. Ren YJ, Liu ST, Zhang HQ (2006) On a Generalized Improved F-expansion method. Commun Theor Phys 45(1):15–28
43. Ren YJ, Liu ST, Zhang HQ (2007) A new generalized algebra method and its application in the (2 + 1) dimensional Boiti–Leon–Pempinelli equation. Chaos Solitons Fractals 32(5):1655–1665
44. Ren YJ, Zhang DH, Cheng F, Zhang HQ (2008) New generalized method to construct new non-traveling waves solutions and traveling waves solutions of K-D equations. Int J Comput Math 1:1–16
45. Ren YJ, He JH, Zhang HQ (2008) A new Exp-Bäcklund transformation method and its application. Chaos Solitons Fractals (accepted)
46. Ren YJ (2007) A new Exp-Bäcklund method. Ph D Dissertation. Dalian University of Technology, p 158–171
47. Ren YJ, He JH, Zhang HQ (2008) A new Exp-N solitary-like method and its application in the (2 + 1)-dimensional generalized KdV–Burgers equation. Chaos Solitons Fractals (accepted)
48. Ren YJ (2007) A new the Exp-N soliton-like method. Ph D Dissertation. Dalian University of Technology, p 172–178
49. Wu YT, Geng XG, Hu XB, Zhu SM (1999) A generalized Hirota–Satsuma coupled Korteweg–de Vries equation and Miura transformations. Phys Lett A 255:259–264
50. Fan EG, Dai HH (2003) A direct approach with computerized symbolic computation for finding a series of traveling waves to nonlinear equations. Computer Physics Communications 153:17–21
51. Yan ZY (2004) An improved algebra method and its applications in nonlinear wave equations. Chaos Solitons Fractals 21:1013–1021
52. Wang DS, Ren YJ, Zhang HQ (2005) Further extended sih-cosh and sin-cos methods and new non-travelng wave solutions of the (2 + 1)-dimensional dispersive long wave equation. Appl Math E-notes 5:157–163
53. Ren YJ, Li YJ, Song FG, Zhang HQ (2006) A new compound operation method of the numerical solutions and analytic solution of partial differential equation. International Conf. on Multimedia, Information Technology and its Applications, Dalian, China, p 328–330
54. Ren YJ, Song FG, Li YJ, Zhang HQ (2006) A mechanization method on the mixed operation of the numerical solutions, analytic solution, and showing their figures simultaneously for differential equation. International Conf. on Multimedia, Information Technology and its Applications, Dalian, China, p 276–278

Books and Reviews

Ablowitz MJ, Clarkson PA (1991) Soliton, Nonlinear Evolution Equations and Inverse Scattering. Cambridge University Press, New York

Dodd RK, Eilbeck JC, Gibbon JD, Morris HC (1984) Solitons and nonlinear wave equations. Academic Press, London

Drazin PG, Johnson RS (1989) Solitons: an introduction. Cambridge University Press, Cambridge

Fan EG (2004) Integrable System and Computer Algebra. Science Press, Beijing

Guo BL (1996) Nonlinear Evolution Equation [M]. Shanghai Science and Technology Press, Shanghai

Liu SK, Liu SD (2000) Nonlinear Equations in Physics. Peking University Press, Beijing

Miura RM (ed) (1976) Backlaund Transformation C. Springer Lecture Notes in Mathematica, vol 515. Springer, New York

Miurs MR (1978) Bäcklund Transformation. Springer, Berlin

Ren YJ (2005) Advanced Mathematics – Multivariable Calculus and Experiments (MATLAB edition). China Machine Press, Beijing

Ren YJ (2007) Numerical Analysis and its MATLAB Realization. Higher Education Press, Beijing

Ren YJ, Sun WH (2004) Advanced Mathematics – Unitary Calculus and Experiments (MATLAB edition). China Machine Press, Beijing

Rogerscand Shaduick WR (1982) Backland Transformation and their Applications. Academic Press, New York

Whitham GB (1974) Linear and nonlinear waves. Wiley, New York

Zhuo LY et al (2000) Nonlinear Physics Threoy and Application [M]. Science Press, Peking

Korteweg–de Vries Equation (KdV), History, Exact *N*-Soliton Solutions and Further Properties of the

Yi Zang
Department of Mathematics,
Zhejiang Normal University, Jinhua, China

Article Outline

Glossary
Definition of the Subject
Introduction
Inverse Scattering Transform for the KdV Equation
Exact *N*-soliton Solutions of the KdV Equation
Further Properties of the KdV Equation
Future Directions
Bibliography

Glossary

Soliton A soliton is a solitary wave which asymptotically preserves its shape and velocity upon nonlinear interaction with other solitary waves, or more generally, with another (arbitrary) localized disturbance.

Korteweg–de Vries equation In mathematics, the Korteweg-de Vries (KdV) equation is a mathematical model of waves on shallow water surfaces. It is particularly famous as the prototypical example of an exactly solvable model, that is, a nonlinear partial differential equation whose solutions can be exactly and precisely specified.

Inverse scattering transform The identification of the KdV equation as an isospectral flow of the Schrödinger operator enabled Gardner, Greene, Kruskal and Miura (GGKM) to devise a method of solving the KdV equation (with 'rapidly decreasing' boundary conditions), called the inverse scattering or inverse spectral transform (IST). This is a direct generalization of the Fourier transform used to solve linear equations and can be represented by essentially the same scheme.

Hirota bilinear method In 1971, Hirota developed a direct method for finding N-soliton solutions of nonlinear evolution equations, also named Hirota bilinear method. The key step is to transform the equation into a bilinear form, from which we can get soliton solutions successively by means of a kind of perturbational technique.

Definition of the Subject

Among integrable equations is the celebrated KdV equation, which serves as a model equation governing weakly nonlinear long waves whose phase speed attains a simple maximum for waves of infinite length. It motivates us to explore beauty hidden in nonlinear differential (and difference) equations. The equation is named for Diederik Korteweg and Gustav de Vries. The remarkable and exceptional discovery of the inverse scattering transform is one of important developments in the field of applied mathematics, which comes from the study of the KdV equation. There are various algebraic and geometric characteristics that the KdV equation possesses, for example, infinitely many symmetries and infinitely many conserved densities, the Lax representation, bi-Hamiltonian structure, loop group, and the Darboux-Bäcklund transformation. More significantly, many physically important solutions to the KdV equation can be presented explicitly through a simple, specific form, called the Hirota bilinear form.

Introduction

A nice story about the history and the underlying physical properties of KdV equation can be found at an Internet page of the Herriot-Watt University in Edinburgh (Scotland). The following text is taken from that page:

Over one hundred and fifty years ago, which conducting experiments to determine the most efficient design for canal boats, a young Scottish engineer named John Scott Russell (1808–1882) made a remarkable scientific discovery. Here his original text as he described it in Russell [1]:

I believe I shall best introduce this phenomenon by describing the circumstances of my own first acquaintance with it. I was observing the motion of a boat which was rapidly drawn along a narrow channel by a pair of horses, when the boat suddenly stopped-not so the mass of water in the channel which it had put in motion; it accumulated round the prow of the vessel in a state of violent agitation, then suddenly leaving it behind, rolled forward with great velocity, assuming the form of a large solitary elevation, a rounded, smooth and well-defined heap of water, which continued its course along the channel apparently without change of form or diminution of speed. I followed it on horseback, and overtook it still rolling on at a rate of some eight or nine miles an hour, preserving its original figure some thirty feet long and a foot to a foot and a half in height. Its height gradually diminished, and after a chase of one or two miles I lost it in the windings of the channel. Such, in the month of August 1834, was my first chance interview with that singular and beautiful phenomenon which I have called the Wave of Translation, a name which it now very generally bears.

He followed this observation with a number of experiments during which he determined the shape of a solitary wave to be that of a $sech^2()$ function. He also determined the relationship of the speed of the wave to its amplitude. At that time there was no equation describing such water waves and having such a solution. Thus John Scott Russell discovered the solution to an as yet unknown equation!

Further investigations were undertaken by Airy, Stokes, Boussinesq and Rayleigh in an attempt to understand this phenomenon. Boussinesq and Rayleigh independently obtained approximate descriptions of the solitary wave; Boussinesq derived a one-dimensional nonlinear evolution equation, which now bears his name, in order to obtain his results.

These investigations provoked much lively discussion and controversy as to whether the inviscid equations of water waves would possess such solitary wave solutions. The issue was finally resolved by Korteweg and de Vries [2]. After performing a Galilean and a variety of scaling transformations, the KdV equation can be written in simplified form:

$$u_t - 6uu_x - u_{xxx} = 0\,, \tag{1}$$

where subscripts denote partial differentiations. In particular, if we now assume a solution in the form of a traveling wave $u(x,t) = f(x+ct)$, then Eq. (1) can be integrated: imposing the boundary conditions at large distances that $u(x,t)$ tends to 0 sufficiently fast as $x \longrightarrow \pm\infty$, we find the exact solution

$$u(x,t) = \frac{c}{2}\mathrm{sech}^2\frac{1}{2}\sqrt{c}(x+ct+\delta)\,, \tag{2}$$

where δ is the phase. This clearly represents the solitary wave observed by John Scott Russell and shows that the peak amplitude is exactly half the speed. Thus larger solitary waves have greater speeds. This suggests a numerical experiment: start with two solitary wave solutions, with centers well separated and the larger to the right. Initially, with negligible overlap, they will evolve *independently* as solitary wave solutions. However, the larger, faster one will start to overtake the smaller and the non linearity will play a significant role. For most dispersive evolution equations these solitary waves would scatter in elastically and lose 'energy' to radiation. Not so for the KdV equation: after a fully nonlinear interaction, the solitary waves reemerge, retaining their identities (same speed and form), suffering nothing more than a phase shift (modified δ's, representing a displacement of their centers). It was after a similar numerical experiment that Kruskal and Zabusky [3] coined the name 'soliton', to reflect the particle-like behavior of the solitary waves under interaction.

It was not until the mid 1960's when applied scientists began to use modern digital computers to study nonlinear wave propagation that the soundness of Russell's early ideas began to be appreciated. He viewed the solitary wave as a self-sufficient dynamic entity, a "thing" displaying many properties of a particle. From the modern perspective it is used as a constructive element to formulate the complex dynamical behavior of wave systems throughout science: from hydrodynamics to nonlinear optics, from plasmas to shock waves, from tornado's to the Great Red Spot of Jupiter, from the elementary particles of matter to the elementary particles of thought. For a more detailed and technical account of the solitary wave, see [4,5].

Inverse Scattering Transform for the KdV Equation

We now briefly discuss the inverse scattering transform for the KdV Eq. (1). First consider the Miura map $u = -v_x - v^2 + \lambda$, which can be viewed as a Riccati equation for v and thus linearized by the substitution $v = \psi_x/\psi$, giving

$$L\psi \equiv \psi_{xx} + u\psi = \lambda\psi\,. \tag{3}$$

This is the time-independent Schrödinger equation with $u(x,t)$ playing the role of potential and λ the energy. It is important to realize that *t is not the time of the time-independent Schrödinger equation*. We think of x as the spatial variable and t as a parameter. Considered as a Sturm–Liouville eigenvalue problem it is natural to ask how λ and ψ change with t as $u(x,t)$ evolves from some initial state according to the KdV equation. GGKM [6,7] discovered the remarkable fact that the (discrete part of the) spectrum necessarily remains constant in 'time' while the corresponding wave functions ψ evolve according to a very simple *linear differential equation*.

Today (following Lax [8]) we usually take the opposite route. We postulate that ψ evolves through a linear differential equation:

$$\psi_t = P\psi\,. \tag{4}$$

Eq. (3) and (4) form an overdetermined system, whose integrability conditions can be written:

$$L_t = [P,L] = PL - LP\,. \tag{5}$$

A consequence of (5) is that all eigenvalues corresponding to the function $u(x,t)$ remain constant. When P is given by:

$$P = -\left(4\partial^3 + 6u\partial + 3u_x\right)\,, \quad \text{(where } \partial \text{ is a differential operator)} \tag{6}$$

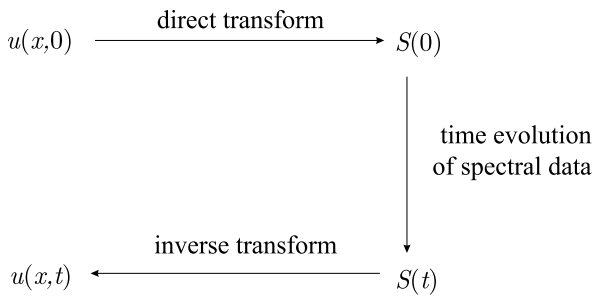

Korteweg–de Vries Equation (KdV), History, Exact *N*-Soliton Solutions and Further Properties of the, Figure 1
Inverse Scattering Transform

(5) reduces to the KdV Eq. (1). Since, under these conditions, the spectrum of L remains constant, the KdV equation is referred to as an *isospectral flow*. Eq. (5) is called the *Lax representation* of the KdV equation.

The identification of the KdV equation as an isospectral flow of the Schrödinger operator enabled GGKM to devise a method of solving the KdV equation (with 'rapidly decreasing' boundary conditions), called the inverse scattering or inverse spectral transform (IST). This is a direct generalization of the Fourier transform used to solve linear equations and can be represented by essentially the same scheme (Fig. 1).

For these boundary conditions, the solutions of (3) are characterized by their asymptotic properties as $x \to \pm\infty$. For a given potential function, the continuous and discrete spectrum are treated separately. Corresponding to the continuous spectrum ($\lambda = -k^2$) the solutions are asymptotically oscillatory, characterized by two coefficients:

$$\psi \sim \begin{cases} T(k)\mathrm{e}^{-ikx} & x \to -\infty \\ \mathrm{e}^{-ikx} + R(k)\mathrm{e}^{ikx} & x \to +\infty \end{cases} \tag{7}$$

subject to the condition $|R|^2 + |T|^2 = 1$. The constants $R(k)$ and $T(k)$ are respectively called the *reflection and transmission* coefficients, from their quantum mechanical interpretation. Under mild conditions on the potential function, the Schrödinger operator L has only a finite number of discrete eigenvalues $\{\kappa_n^2\}$. The corresponding eigenfunctions are square integrable and are the 'bound states' of quantum mechanics with asymptotic properties (for $\kappa_n > 0$):

$$\psi_n \sim \begin{cases} \widetilde{c}_n \mathrm{e}^{\kappa_n x} & x \to -\infty \\ c_n \mathrm{e}^{-\kappa_n x} & x \to +\infty \end{cases} . \tag{8}$$

The *direct scattering transform* constructs the quantities $\{T(k), R(k), \kappa_n, c_n\}$ from a given potential function. The important inversion formula were derived by Gel'fand and Levitan in 1955 [9]. These enable the potential u to be constructed out of the spectral or scattering data $S = \{R(k), \kappa_n, c_n\}$. This is considerably more complicated than the Inverse Fourier Transform, involving the solution of a nontrivial integral equation, whose kernel is built out of the scattering data (see [10,11,12,13] for descriptions of this).

To solve the KdV equation we first construct the scatting data $S(0)$ from the initial condition $u(x, 0)$. As a consequence of (4) (with an additional constant) with the given boundary conditions, the scatting data evolves in a very simple way. Indeed, we can give explicit formula:

$$R(k, t) = R(k, 0)\mathrm{e}^{8ik^3 t}, \quad c_n(t) = c_n(0)\mathrm{e}^{-4\kappa_n^3 t} . \tag{9}$$

Using the inverse scattering transform on the scattering data $S(t)$, we obtain the potential $u(x, t)$ and thus the solution to the initial value problem for the KdV equation.

This process cannot be carried out *explicitly* for arbitrary initial data, although, in this case, it gives a great deal of information about the solution $u(x, t)$. However, whenever the reflection coefficient is zero, the kernel of Gel'fand-Levitan integral equation becomes separable and explicit solutions can be found. It is in this way that the N-soliton solution is constructed by IST from the initial condition:

$$u(x, 0) = N(N + 1)\mathrm{sech}^2 x . \tag{10}$$

The general formula for the multi-soliton solution is given by:

$$u(x, t) = 2(\ln \det M)_{xx} , \tag{11}$$

where M is a matrix built out of the discrete scattering data.

Exact *N*-soliton Solutions of the KdV Equation

Besides the IST, there are several analytical methods for obtaining solutions of the KdV equation, such as Hirota bilinear method [14,15,16], Bäcklund transformation [17], Darboux transformation [18], and so on. The existence of such analytical methods reflects a rich algebraic structure of the KdV equation. In Hirota's method, we transform the equation into a bilinear form, from which we can get soliton solutions successively by means of a kind of perturbational technique. The Bäcklund transformation is also employed to obtain solutions from a known solution of the concerned equation. In what follows, we will mainly discuss the Hirota bilinear method to derive the N-soliton solutions of the KdV equation.

It is well known that Hirota developed a direct method for finding N-soliton solutions of nonlinear evolution equations. In particular, we shall discuss the KdV bilinear form

$$D_x\left(D_t + D_x^3\right) f \cdot f = 0\,, \tag{12}$$

by the dependent variable transformation

$$u(x,t) = 2(\ln f)_{xx}\,. \tag{13}$$

Here the Hirota bilinear operators are defined by

$$D_x^m D_t^n a{\cdot}b = (\partial_x - \partial_{x'})^m (\partial_t - \partial_{t'})^n a(x,t) b(x',t')|_{x'=x,t'=t}\,. \tag{14}$$

We expand f as power series in a parameter ε

$$f(x,t) = 1 + f^{(1)}\varepsilon + f^{(2)}\varepsilon^2 + \cdots + f^{(j)}\varepsilon^j + \cdots\,. \tag{15}$$

Substituting (15) into (12) and equating coefficients of powers of ε gives the following recursion relations for the $f^{(n)}$

$$\varepsilon\colon f^{(1)}_{xxxx} + f^{(1)}_{xt} = 0\,, \tag{16}$$

$$\varepsilon^2\colon f^{(2)}_{xxxx} + f^{(2)}_{xt} = -\frac{1}{2}(D_x D_t + D_x^4) f^{(1)} \cdot f^{(1)}\,, \tag{17}$$

$$\varepsilon^3\colon f^{(3)}_{xxxx} + f^{(3)}_{xt} = -\left(D_x D_t + D_x^4\right) f^{(1)} \cdot f^{(2)}\,, \tag{18}$$

and so on. N-soliton solutions of the KdV equation are found by assuming that $f^{(1)}$ has the form

$$f^{(1)} = \sum_{j=1}^{N} \exp(\eta_j), \quad \eta_j = k_j x - \omega_j t + x_{j0}\,, \tag{19}$$

and k_j, ω_j and x_{j0} are constants, provided that the series (15) truncates.

For $N = 1$, we take

$$f^{(1)} = \exp(\eta_1),$$

and by solving (16) we find that

$$f^{(n)} = 0, \quad \text{for } n = 2, 3, \ldots\,.$$

Therefore we have

$$f_1 = 1 + \exp(\eta_1), \quad \omega_1 = -k_1^3\,,$$

and

$$u(x,t) = \frac{k_1^2}{2}\operatorname{sech}^2 \frac{1}{2}\left(k_1 x - k_1^3 t + x_{10}\right)\,. \tag{20}$$

For $N = 2$, the two-soliton solution for the KdV equation is similarly obtained from

$$u(x,t) = 2(\ln f_2)_{xx}\,,$$

where

$$f_2 = 1 + \exp(\eta_1) + \exp(\eta_2) + \exp(\eta_1 + \eta_2 + A_{12})\,. \tag{21}$$

Frequently the N-soliton solutions are obtained as follows

$$f_N = \sum_{\mu=0,1} \exp\left(\sum_{j=1}^{N} \mu_j \eta_j + \sum_{1\le j<l}^{N} \mu_j \mu_l A_{jl}\right)\,, \tag{22}$$

where the first $\sum_{\mu=0,1}$ means a summation over all possible combinations of $\mu_j = 0, 1$ and $\sum_{1\le j<l}^{N}$ means a summation over all possible pairs (j, l) chosen from the set $\{1, 2, \ldots, N\}$, with the condition that $j < l$.

Further Properties of the KdV Equation

Conservation Laws

It was this numerical evidence which prompted Kruskal and his co-workers in Princeton to analytically investigate the KdV equation. Initially, Miura investigated local conservation laws:

$$\partial_t T + \partial_x F = 0\,, \tag{23}$$

where T and F are polynomials in $u(x,t)$ and its x-derivatives and where ∂_t and ∂_x denote total derivatives. With appropriate boundary conditions this leads to a conserved quantity (constant of the motion). Integrating (23) with respect to x we get:

$$\partial_t \int_A^B T + [F]_A^B = 0\,. \tag{24}$$

Under periodic (in x) boundary conditions with $A - B$ an integer multiple of the period or with $u(x,t)$ rapidly decreasing as $x \to \pm\infty$ and $(A, B) = (-\infty, +\infty)$, the square bracket in (24) vanishes and we have the constant of motion $\partial_t \int_A^B T dx$. The quantities T and F are respectively called conserved density and flux. Each T is, in fact, only determined up to an exact x-derivative, so defines an equivalence class of conserved densities:

$$T \sim T + \partial_x S\,, \tag{25}$$

since this only adds $\partial_t S$ to F and leaves the value of $\int T dx$ unchanged. For the KdV equation the first three are:

$$\begin{aligned} T_0 &= u\,, & F_0 &= -u_{xx} - 3u^2, \\ T_1 &= \frac{1}{2}u^2\,, & F_1 &= -uu_{xx} + \frac{1}{2}u_x^2 - 2u^3\,, \\ T_2 &= u^3 - \frac{1}{2}u_x^2\,, & F_2 &= u_x u_{xxx} - \frac{1}{2}u_{xx}^2 \\ & & &\quad - 3u^2 u_{xx} + 6uu_x^2 - \frac{9}{2}u^4\,. \end{aligned} \tag{26}$$

The first of these is just the equation itself. These three conservation laws have same physical interpretation, so it was no surprise that they exist. However, Miura discovered several more by direct calculation and was led to the conjecture that there should be *infinitely many*.

In order to ascertain whether the KdV equation was the only such equation with so many conservation laws Miura investigated equations of the form:

$$u_t = u_{xxx} + 6u^n u_x\,, \tag{27}$$

and found that for $n = 1$ and $n = 2$ (and *only* these values) there existed many conservation laws. With a slight change in notation the second of these, called the modified KdV (MKdV) equation, can be written:

$$v_t = v_{xxx} - 6v^2 v_x = \left(v_{xx} - 2v^3\right)_x\,. \tag{28}$$

The first few conservation laws correspond to conserved densities and fluxes:

$$\begin{aligned} \tilde{T}_{-1} &= v\,, & \tilde{F}_{-1} &= -v_{xxx} + 2v^3, \\ \tilde{T}_0 &= v^2\,, & \tilde{F}_0 &= -2vv_{xx} + v_x^2 + 3v^4, \\ \tilde{T}_1 &= \frac{1}{2}\left(v_x^2 + v^4\right)\,, & \tilde{F}_1 &= -v_x v_{xxx} + \frac{1}{2}v_{xx}^2 \\ & & &\quad - 2v^3 v_{xx} + 6v^2 v_x^2 + 2v^6\,. \end{aligned} \tag{29}$$

The Lax Hierarchy

In [8] Lax reformulated GGKM's discovery [6,7] of the isospectral nature of the KdV equation in algebraic form:

$$\left.\begin{aligned} L\psi \equiv (\partial^2 + u)\psi &= \lambda\psi \\ \psi_t = P\psi \equiv -\left(4\partial^3 + 6u\partial + 3u_x\right)\psi & \\ \lambda_t &= 0 \end{aligned}\right\} \Rightarrow L_t = [P, L] \equiv PL - LP\,. \tag{30}$$

This led Lax to an interesting generalization:

$$\left.\begin{aligned} L\psi \equiv (\partial^2 + u)\psi &= \lambda\psi \\ \psi_{t_m} = P_{(m)}\psi \equiv \partial^{2m+1} + \sum_{i=0}^{2m-1} b_i \partial^i & \\ \lambda_t &= 0 \end{aligned}\right\} \Rightarrow L_{t_m} = [P_{(m)}, L]\,. \tag{31}$$

The integrability condition $L_{t_m} = [P_{(m)}, L]$ is explicitly written as:

$$u_{t_m} = (2m+1)u_x - 2b_{2m-1x}\partial^{2m} + \cdots + (P_{(m)}u - Lb_0)\,. \tag{32}$$

Equating coefficients of $\partial^i, i = 0, \ldots, 2m$, gives us $2m + 1$ equations, from which we deduce the $2m$ coefficients $b_0, \ldots, b_{2m-1}$:

$$\begin{aligned} b_{2m-1} &= \frac{1}{2}(2m+1)u, \\ b_{2m-2} &= \frac{1}{4}(2m+1)(2m-1)u_x, \ldots \end{aligned} \tag{33}$$

together with the isospectral flow:

$$u_{t_m} = P(m)u - Lb_0\,. \tag{34}$$

Nontrivial flows only exist for odd-order equations, since the adjoint of the integrability condition implies $P^\dagger = -P$. There exists an infinite hierarchy of such isospectral flows, the first three of which are:

$$u_{t_0} = u_x\,, \tag{35}$$

$$u_{t_1} = \frac{1}{4}(u_{xxx} + 6uu_x)\,, \tag{36}$$

$$u_{t_2} = \frac{1}{16}(u_{xxxxx} + 10uu_{xxx} + 20u_x u_{xx} + 30u^2 u_x)\,, \tag{37}$$

corresponding respectively to operators:

$$P_{(0)} = \partial\,, \tag{38}$$

$$P_{(1)n} = \partial^3 + \frac{3}{4}(u\partial + \partial u)\,, \tag{39}$$

$$\begin{aligned} P_{(2)} = \partial^5 &+ \frac{5}{4}(u\partial^3 + \partial^3 u) \\ &+ \frac{5}{16}((3u^2 - uxx)\partial + \partial(3u^2 - u_{xx}))\,. \end{aligned} \tag{40}$$

Future Directions

In mathematics, the KdV equation is a mathematical model of waves on shallow water surfaces. It is particu-

larly famous as the prototypical example of an exactly solvable model, that is, a nonlinear partial differential equation whose solutions can be exactly and precisely specified. The solutions in turn are the prototypical examples of solitons; these may be found by means of the inverse scattering transform. The mathematical theory behind the KdV equation is rich and interesting, and, in the broad sense, is a topic of active mathematical research.

The KdV equation has several connections to physical problems. In addition to being the governing equation of the string in the Fermi–Pasta–Ulam problem in the continuum limit, it approximately describes the evolution of long, one-dimensional waves in many physical settings, including shallow-water waves with weakly nonlinear restoring forces, long internal waves in a density-stratified ocean, ion-acoustic waves in a plasma, acoustic waves on a crystal lattice, and more.

The scientists' interest for analytical solutions of the KdV equation stems from the fact that in applying numerical methods to nonlinear partial differential equations, the KdV equation is well suited as a test object, since having an analytical solution statements can be made on the quality of the numerical solution in comparing the numerical result to the exact result. More significantly, the existence of several analytical methods reflects a rich algebraic structure of the KdV equation. It is hoped that the study of the KdV equation could further assist in understanding, identifying and classifying nonlinear integrable differential equations and their exact solutions.

Bibliography

Primary Literature

1. Russell JS (1845) Report on Waves. Proc Roy Soc, Edinburgh, pp 421–445
2. Korteweg DJ, de Vries G (1895) On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves. Philos Mag Serl 39(5):422–443
3. Zabusky NJ, Kruskal MD (1965) Interaction of Solitons in a Collisionless Plasma and the recurrence of initial States. Phys Rev Lett 15:240–243
4. Bullough RK (1988) "The wave par excellence", the solitary, progressive great wave of equilibrium of the fluid – an early history of the solitary wave. In: Lakshmanan M (ed) solitons. Springer Series in Nonlinear Dynamics. Springer, New York, Berlin, Heidelberg etc. pp 150–281
5. Scott AC, Chu FY, Mclaughlin DW (1973) The Soliton: A New Concept in Applied Science. Proc IEEE 61(10):1443–1483
6. Gardner CS, Greene JM, Kruskal MD, Miura RM (1967) Method for solving the Korteweg–de Vries equation. Phys Rev Lett 19:1095–1097
7. Gardner CS, Greene JM, Kruskal MD, Miura RM (1974) The Korteweg–de Vries equation and generalizations VI. Methods for exact solution. Commun Pure Appl Math 27:97–133
8. Lax PD (1968) Integrals of nonlinear equations of evolution and solitary waves. Commun. Pure Appl. Math 21:467–490
9. Gel'fand IM, Levitan BM (1951) On the determination of a differential equation from its spectral function. Izv Akad Nauk SSR Ser Math 15:309–360
10. Ablowitz MJ, Clarkson PA (1991) Solitons, Nonlinear Evolution Equations and Inverse Scattering. London Math Soc, vol 149, CUP, Cambridge
11. Fordy AP (1990) Soliton Theory: A Survey of Results. MUP, Manchester
12. Newell AC (1985) Solitons in Mathematics and Physics. SIAM, Philadelphia
13. Novikov SP, Manakov SV, Pitaevskii LP, Zakharov VE (1984) Theory of Solitons. Plenum, NY
14. Hirota R (1971) Exact soliton of the KdV equation for multiple collisions of solitons. Phys Rev Lett 27:1192–1194
15. Hirota R, Satsuma J (1977) Nonlinear evolution equation generated from the Bäcklund transformation for the Boussinesq equation. Prog Theor Phys Suppl 57:797–807
16. Hirota R (1980) Direct methods in soliton theory. In: Bullough RK, Caudrey PJ (ed) Topics of Modern Physics, vol 17. Springer, New York, pp 157–176
17. Wahlquist HD, Estabrook FB (1973) Bäcklund transformations for solitons of the Korteweg–de Vries equation. Phys Rev Lett 31:1386–1390
18. Matveev VB, Salle MA (1991) Darboux Transformations and Solitons. Springer, Berlin

Books and Reviews

Ablowitz MJ, Clarkson PA (1991) Solitons, Nonlinear Evolution Equations and Inverse Scattering. Cambridge University Press, New York

Ablowitz MJ, Segur H (1981) Solitons and the Inverse Scattering Transform. SIAM, USA

Drazin PG, Johnson RS (1989) Solitons: an Introduction. Cambridge University Press, New York

Konopelchenko BG (1993) Solitons in Multidimensions: Inverse Spectral Transform Method. World Scientific, Singapore

Rogers C, Shadwick WR (1982) Bäcklund Transformations and Their Application. Academic Press, New York

Whitham GB (1974) Linear and Nonlinear Waves. Wiley, New York

Korteweg–de Vries Equation (KdV) and Modified Korteweg–de Vries Equations (mKdV), Semi-analytical Methods for Solving the

Doğan Kaya[1,2]

[1] Department of Mathematics, Firat University, Elazig, Turkey

[2] Dean of Engineering Faculty, Ardahan University, Ardahan, Turkey

Article Outline

Glossary
Definition of the Subject
Introduction
An Analysis of the Semi-analytical Methods and Their Applications
Future Directions
Bibliography

Glossary

Korteweg–de Vries equation The classical nonlinear equations of interest usually admit for the existence of a special type of the traveling wave solutions, which are either solitary waves or solitons.

Modified Korteweg–de Vries This equation is a modified form of the classical KdV equation in the nonlinear term.

Soliton This concept can be regarded as solutions of nonlinear partial differential equations.

Exact solution A solution to a problem that contains the entire physics and mathematics of a problem, as opposed to one that is approximate, perturbative, closed, etc.

Adomian decomposition method, Homotopy analysis method, Homotopy perturbation method and Variational iteration method These are some of the semi-analytic/numerical methods for solving ODE or PDE in literature.

Definition of the Subject

In this study, some semi-analytical/numerical methods are applied to solve the Korteweg–de Vries (KdV) equation and the modified Korteweg–de Vries (mKdV) equation, which are characterized by the solitary wave solutions of the classical nonlinear equations that lead to solitons. Here, the classical nonlinear equations of interest usually admit for the existence of a special type of the traveling wave solutions which are either solitary waves or solitons. These approaches are based on the choice of a suitable differential operator which may be ordinary or partial, linear or nonlinear, deterministic or stochastic. It does not require discretization, and consequently massive computation.

In this scheme the solution is performed in the form of a convergent power series with easily computable components. This section is particularly concerned with the Adomian decomposition method (ADM) and the results obtained are compared to those obtained by the variational iteration method (VIM), homotopy analysis method (HAM), and homotopy perturbation method (HPM). Some numerical results of these particular equations are also obtained for the purpose of numerical comparisons of those considered approximate methods. The numerical results demonstrate that the ADM is relatively accurate and easily implemented.

Introduction

In this part, a certain nonlinear partial differential equation is introduced which is characterized by the solitary wave solutions of the classical nonlinear equations that lead to solitons [8,9,11,64,70]. The classical nonlinear equations of interest usually admit for the existence of a special type of the traveling wave solutions which are either solitary waves or solitons. In this study, a few solutions arising from the semi-analytical work on the Korteweg–de Vries (KdV) equation and modified Korteweg–de Vries (mKdV) equation will be reviewed.

In this work, a brief history of the above mentioned nonlinear KdV equation is given and how this type of equation leads to the soliton solutions is then presented as an introduction to the theory of solitons. This theory is an important branch of applied mathematics and mathematical physics. In the last decade this topic has become an active and productive area of research, and applications of the soliton equations in physical cases have been considered. These have important applications in fluid mechanics, nonlinear optics, ion plasma, classical and quantum fields theories, etc.

The best introduction for this study may be the Scottish naval engineer J. Scott Russell's seminal 1844 report to the Royal Society. In the time of the eighteen century, Scott Russell's report was called "Report on Waves" [57]. Around forty years later, starting from Scott Russell's experimental observations in 1934, the theoretical scientific work of Lord Rayleigh and Joseph Boussinesq around 1870 [56] independently confirmed Russell's prediction and derived formula (1) from the equation of motion for an inviscid incompressible liquid. They also gave the solitary wave profile $z = u(x, t)$ as

$$u(x, t) = a \operatorname{sech}^2 \left[\beta(x - Ut)\right] \tag{1}$$

where $\beta^2 = 3a \div \left\{4h^2(h + a)\right\}$ any $a > 0$. They drew the solitary wave profile as in the following

Finally, two Dutch scientists, Korteweg and de Vries, developed a nonlinear partial differential equation to model the propagation of shallow water waves applicable to situation in 1895 [43]. This work was really what Scott Russell fortuitously witnessed. This famous classical equation is known simply as the KdV equation. Korteweg and

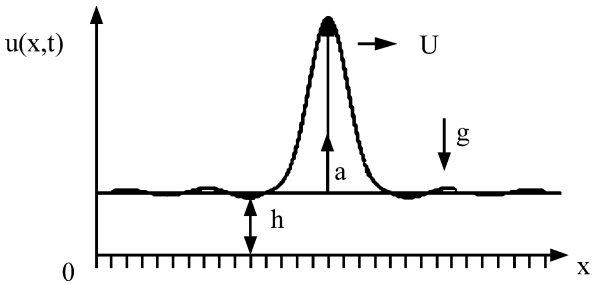

Korteweg–de Vries Equation (KdV) and Modified KdV (mKdV), Semi-analytical Methods for Solving the, Figure 1
A solitary wave

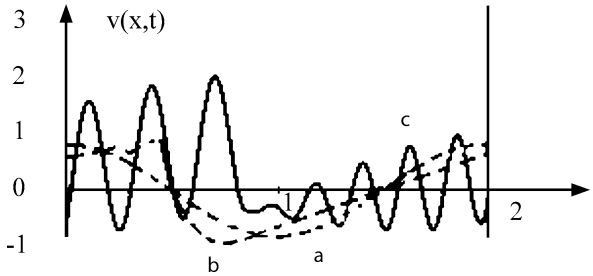

Korteweg–de Vries Equation (KdV) and Modified KdV (mKdV), Semi-analytical Methods for Solving the, Figure 2
Development of solitary waves: a initial profile at $t = 0$, b profile at $t = \pi^{-1}$, and c wave profile at $t = (3.6)\pi^{-1}$ [13]

de Vries published a theory of shallow water waves which reduced Russell's observations to its essential features. The nonlinear classical dispersive equation was formulated by Korteweg and de Vries in the form

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} + \varepsilon\frac{\partial^3 u}{\partial x^3} + \gamma\, u\frac{\partial u}{\partial x} = 0\,, \tag{2}$$

where c, ε, γ are physical parameters. This equation now plays a key role in soliton theory.

After setting the theory and applications of the KdV solitary waves, scientists should know a numerical model for the such nonlinear equations. This model was constructed and published as a Los Alamos Scientific Laboratory Report in 1955 by Fermi, Pasta, and Ulam (FPU). This was a numerical model of a discrete nonlinear mass–spring system [13]. The natural question arises of how the energy would be distributed among all modes in this nonlinear discrete system. The energy would be distributed uniformly among all modes in accordance with the principle of the equipartition of energy stated by Fermi, Pasta and Ulam. This model follows the form of a coupled nonlinear ordinary differential equation system, where w_i is a function of t, and this system is written as

$$\frac{m}{k}\frac{\mathrm{d}^2 w_i}{\mathrm{d}t^2} = (w_{i+1} + w_{i-1} - 2w_i) + \alpha\left[(w_{i+1} - w_i)^2 - (w_i - w_{i-1})^2\right] \tag{3}$$

where w_i is the displacement of the ith mass from the equilibrium position, $i = 1, 2, \ldots, n$, $w_0 = w_n = 0$, k is the linear spring constant and α (> 0) measures the strength of nonlinearity.

In the late 1960s, Zabusky and Kruskal numerically studied, and then analytically solved, the KdV equation [71] from the result of the FPU experiment inspiration. They came to the result that the stable pulse-like waves could exist in a problem described by the KdV equation from their numerical study. A remarkable aspect of the discovered solitary waves is that they retain their shapes and speeds after a collision. After the observation of Zabusky and Kruskal, they called to these waves "solitons", because the character of the waves had a particle-like nature. In fact, they considered the initial value problem for the KdV equation in the form

$$v_t + v v_x + \delta\, v_{xxx} = 0\,, \tag{4}$$

where $\delta = (h/\ell)^2$, ℓ is a typical horizontal length scale, with the initial condition

$$v(x, 0) = \cos\,\pi\, x, \qquad 0 \le x \le 2 \tag{5}$$

and the periodic boundary conditions with period 2, so that $v(x, t) = v(x + 2, t)$ for all t. Their numerical study with $\sqrt{\delta} = 0.022$ produced a lot of new interesting results, which are shown as follows.

In 1967, this numerical development was placed on a settled mathematical basis with C.S. Gardner, J.M. Greene, M.D. Kruskal and R.M. Miura's discovery of the inverse-scattering-transform method [14,15]. This way of calculation was an ingenious method for finding the exact solution of the KdV equation. Almost simultaneously, Miura [51] and Miura et al. [52] formulated another ingenious method to derive an infinite set of conservation laws for the KdV equations by introducing the so-called Miura transformation. Subsequently, Hirota [28,29,30] constructed analytical solutions of the KdV equation which provide the description of the interaction among N solitons for any positive integral N.

The soliton concept can be regarded with solutions of nonlinear partial differential equations. The soliton solution of a nonlinear equation is usually used as single wave. If there are several soliton solutions, these solutions are

called "solitons." On the other hand, if a soliton separates infinitely from other soliton, this soliton is a single wave. Besides, a single wave solution cannot be a sech^2 function for equations different from nonlinear equations, such as the KdV equation. But this solution can be sech or $\tan^{-1}(e^{\alpha x})$.

At this stage, one could ask what the definition of a soliton solution is. It is not easy to define the soliton concept. Wazwaz [64] describes solitons as solutions of nonlinear differential equations such that

1. A long and shallow water wave should not lose its permanent forms;
2. A long and shallow water wave of the solution is localized, which means that either the solutions decay exponentially to zero such as the solitons admitted by the KdV equation, or approach a constant at infinity such as the solitons provided by the SG equation;
3. A long and shallow water wave of the solution can interact with other solitons and still preserve its character.

There is also a more formal definition of this concept, but these definitions require substantial mathematics [12]. On the other hand, the phenomena of the solitons use not quite the same as have stated above three properties in all expression of the solutions. For example, the concept referred to as "light bullets" in nonlinear optics are often called solitons despite losing energy during interaction. This idea can be found at an internet web page of the Simon Fraser University British Columbia, Canada [50].

Nonlinear phenomena play a crucial role in applied mathematics and physics. Calculating exact and numerical solutions, particularly traveling wave solutions, of nonlinear equations in mathematical physics play an important role in soliton theory [8,70]. It has recently become more interesting to obtain exact solutions of nonlinear partial differential equations utilizing symbolical computer programs (such as Maple, Matlab, and Mathematica) which facilitate complex and tedious algebraical computations. It is also important to find exact solutions of nonlinear partial differential equations. These equations exist because of the mathematical models of complex physical phenomenon that arise in engineering, chemistry, biology, mechanics and physics. Various effective methods have been developed to understand the mechanisms of these physical models, to help physicians and engineers and to ensure knowledge for physical problems and its applications.

Many explicit exact methods have been introduced in literature [8,9,11,31,32,33,42,62,64,65,66,67,69,70]. Some of them are: Bäcklund transformation, Cole–Hopf transformation, Generalized Miura Transformation, Inverse Scattering method, Darboux transformation, Painleve method, similarity reduction method, tanh (and its variations) method, homogeneous balance method, Exp-function method, sine–cosine method and so on. There are also many numerical methods implemented for nonlinear equations [16,17,27,34,35,37,38,39,40,54,61]. Some of them are: finite difference methods, finite elements method, Sinc–Galerkin method and some approximate or semi-analytic/numerical methods such as Adomian decomposition method, variational analysis method, homotopy analysis method, homotopy perturbation method and so on.

An Analysis of the Semi-analytical Methods and Their Applications

The recent years have seen a significant development in the use of various methods to find the numerical and semi-analytical/numerical solution of a linear or nonlinear, deterministic or stochastic ODE or PDE. In this work, we will only discuss the nonlinear PDE case of the problem.

Before giving the semi-analytical and numerical implementations of the considered methods, may draw some conclusions from Liao's book [46] and the recent new papers [1,7,10] for comparisons of the ADM, HAM, HPM and VIM. ADM can be applied to solve linear or nonlinear ordinary and partial differential equations, no matter whether they contain small/large parameters, and thus is rather general. In some cases, the Adomian decomposition series converge rapidly. However, this method has some restrictions, e. g. if you are solving nonlinear equation you have to sort out Adomian's polynomials. In the same case, this needs to take more calculations. Liao [46] pointed out that "in general, convergence regions of power series are small, thus acceleration techniques are often needed to enlarge convergence regions. This is mainly due to the fact that a power series is often not an efficient set of base functions to approximate a nonlinear problem, but unfortunately ADM does not provide us with freedom to use different base functions. Like the artificial small parameter method and the d-expansion method, ADM itself also does not provide us with a convenient way to adjust the convergence region and the rate of approximation solutions".

Many authors [1,7,10] made their own studies in which they gave implementations of equations by using HAM, HPM, and then found numerical solutions. Their results show that: (i) Liao's HAM can produce much better approximations than the previous solutions for nonlinear differential equations. (ii) They compared the approximations of these two methods and they observe that although

HAM is faster than HPM (see Figs. 1–6 in [10]), both of the methods converge to the exact solution quite fast. HAM and HPM are in some cases similar to each other; e. g., the obtained solution function of the equation is shorter if HPM is used instead of HAM. The corresponding results converge more rapidly if HAM is used. (iii) The other advantage of the HAM is that it gives the flexibility to choose an auxiliary parameter h to adjust and control the convergence and its rate for the solutions series and the defined different functions which originate from the nature of the considered problems [1]. (iv) If one compares the approximated numerical solution with the corresponding exact solution by using HAM and HPM, one can see that when the small parameter h is increased the error of the first method is less than the second one.

Chowdhury and co worker [7], show that their obtained numerical results by the 5-term HAM are exactly the same as the ADM solutions and HPM solutions for the special case of the auxiliary parameter $h = -1$ and auxiliary function $H(x) = 1$. Because of this conclusion, they admitted that the HPM and the ADM is a special case of HAM.

Adomian Decomposition Method

The aim of the present section is to give an outline and implementation of the Adomian decomposition method (ADM) for nonlinear wave equations, and to obtain analytic and approximate solutions which are obtained in a rapidly convergent series with elegantly computable components by this method. The approach is based on the choice of a suitable differential operator which may be ordinary or partial, linear or nonlinear, deterministic or stochastic [4,5,6,63,64]. It allows one to obtain a decomposition series solution of the equation which is calculated in the form of a convergent power series with easily computable components.

The inhomogeneous problem is quickly solved by observing the self-canceling "noise" terms, where the sum of the components vanishes in the limit. Many tests which model problems from mathematical physics, linear and nonlinear, are discussed to illustrate the effectiveness and the performance of the ADM. Adomian and Rach [5] and Wazwaz [63] have investigated the phenomena of the self-canceling "noise" terms where the sum of the components vanishes in the limit. An important observation was made that the "noise" terms appear for nonhomogenous cases only. Further, it was formally justified that if terms in u_0 are canceled by terms in u_1, even though u_1 includes further terms, then the remaining non-canceled terms in u_1 constitute the exact solution of the equation.

ADM is valid for ordinary and partial differential equations, whether or not they contain small/large parameters, and thus is rather general. Moreover, the Adomian approximation series converge quickly. However, this method has some restrictions. Approximates solutions given by ADM often contain polynomials. In general, convergence regions of power series are small, thus acceleration techniques are often needed to enlarge convergence regions. This is mainly due to the fact that power series is often not an efficient set of base functions to approximate a nonlinear problem, but unfortunately ADM does not provide us with freedom to use different base functions.

An outline of the method will be given here in order to obtain analytic and approximate solutions by using the ADM. Considering a generalized KdV equation [41]

$$u_t + \alpha\, u^m u_x + u_{xxx} = 0\,, \quad u(x,0) = g(x)\,, \tag{6}$$

where α and $m > 0$ are constants. An operator form of this equation can be written as

$$L_t\,(u) + \alpha N u + L_{xxx}\,(u) = 0\,, \tag{7}$$

where $L_t \equiv \partial/(\partial t)$, $Nu = u^m u_x$ and $L_{xxx} \equiv \partial^3/(\partial x^3)$. It is assumed that L_t^{-1} is an integral operator given by $L_t^{-1} \equiv \int_0^t (.)\mathrm{d}t$. Operating with the integral operator L_t^{-1} on both sides of (7) with

$$L_t^{-1} L_t\,(u) = -\alpha L_t^{-1}\,(Nu) - L_t^{-1} L_{xxx}\,(u)\ . \tag{8}$$

Therefore, it follows that

$$u(x,t) = u(x,0) - \alpha L_t^{-1}\,(Nu) - L_t^{-1} L_{xxx}\,(u)\ .$$

We find that the zeroth component is obtained by

$$u_0 = u(x,0)\,, \tag{9}$$

which is defined by all terms that arise from the initial condition and from integrating the source term and decomposing the unknown function $u(x,t)$ as a sum of components defined by the decomposition series

$$u(x,t) = \sum_{n=0}^{\infty} u_n(x,t)\,. \tag{10}$$

The nonlinear term $u^m u_x$ can be decomposed into the infinite series of polynomials given by

$$Nu = u^m u_x = \sum_{n=0}^{\infty} A_n\,,$$

where the components $u_i(x,t)$ will be determined recurrently, and the A_n polynomials are the so-called Adomian polynomials [64] of $u_0, u_1, u_2, \cdots, u_i$ defined by

$$A_n = \frac{1}{n!}\left[\frac{\mathrm{d}^n}{\mathrm{d}\lambda^n}\Phi\left(\sum_{k=1}^{\infty}\lambda^k u_k\right)\right]\Bigg|_{\lambda=0}, \quad n \geq 0. \quad (11)$$

Substituting (11) and (9) into (8) gives rise to

$$u_{n+1} = -L_t^{-1}(A_n) - L_t^{-1}L_{\mathrm{xxx}}(u_n), \quad n \geq 0 \quad (12)$$

where L_t^{-1} is the previously given integration operator. The solution $u(x,t)$ must satisfy the requirements imposed by the initial conditions. The decomposition method provides a reliable technique that requires less work if compared with the traditional techniques. To give a clear overview of the methodology, the following examples will be discussed.

Example 1 Let's consider the mKdV equation (6) for the value of $m = 1$, which is called the classical KdV equation, and take this equation with the following initial value condition [36],

$$u(x,0) = \frac{3\lambda}{\alpha} - \frac{3\lambda}{\alpha}\tanh^2\left(\frac{\sqrt{\lambda}}{2}x\right), \quad (13)$$

and then basically Eq. (6) is taken in an operator form exactly in the same manner as the form of the Eq. (6) and using (9) to find the zeroth component of $u_0 = u(x,0)$. Taking into consideration (12) with (11) and one can calculate the rest of the terms of the series with the aid of Mathematica and then write these terms to the Eq. (10), yielding

$$\begin{aligned}
u(x,t) = {} & \\
& \frac{1}{4\alpha}\left\{3t^2\lambda^4\left(-2+\cosh\left(x\sqrt{\lambda}\right)\right)\operatorname{sech}^4\left(\frac{x\sqrt{\lambda}}{2}\right)\right\} \\
& + \frac{1}{64\alpha}\left\{t^4\lambda^7\left(-33-26\cosh\left(x\sqrt{\lambda}\right)\right.\right. \\
& \left.\left.+\cosh\left(2x\sqrt{\lambda}\right)\right)\operatorname{sech}^6\left(\frac{x\sqrt{\lambda}}{2}\right)\right\} \\
& + \frac{1}{8\alpha}\left\{t^3\lambda^{\frac{11}{2}}\operatorname{sech}^5\left(\frac{x\sqrt{\lambda}}{2}\right)\left(-11\sinh\left(\frac{x\sqrt{\lambda}}{2}\right)\right.\right. \\
& \left.\left.+\sinh\left(\frac{3x\sqrt{\lambda}}{2}\right)\right)\right\} \\
& + \frac{1}{640\alpha}\left\{t^5\lambda^{\frac{7}{2}}\operatorname{sech}^7\left(\frac{x\sqrt{\lambda}}{2}\right)\left(302\sinh\left(\frac{x\sqrt{\lambda}}{2}\right)\right.\right. \\
& \left.\left.-57\sinh\left(\frac{3x\sqrt{\lambda}}{2}\right)+\sinh\left(\frac{5x\sqrt{\lambda}}{2}\right)\right)\right\} \\
& + \frac{1}{\alpha}\left\{9t\lambda^{\frac{5}{2}}\operatorname{sech}^2\left(\frac{x\sqrt{\lambda}}{2}\right)\tanh\left(\frac{x\sqrt{\lambda}}{2}\right)\right\} \\
& - \frac{1}{\alpha}\left\{6t\lambda^{\frac{5}{2}}\operatorname{sech}^4\left(\frac{x\sqrt{\lambda}}{2}\right)\tanh\left(\frac{x\sqrt{\lambda}}{2}\right)\right\} \\
& - \frac{1}{\alpha}\left\{6t\lambda^{\frac{5}{2}}\operatorname{sech}^2\left(\frac{x\sqrt{\lambda}}{2}\right)\tanh^3\left(\frac{x\sqrt{\lambda}}{2}\right)\right\} \\
& + \frac{1}{\alpha}\left\{3\lambda\left(1-\tanh^2\left(\frac{x\sqrt{\lambda}}{2}\right)\right)\right\} + \cdots
\end{aligned}$$

Example 2 Consider the mKdV equation (6) with the values of $m = 2$ and $\alpha = 6$, it has the traveling wave solution which can be obtained subject to the initial condition [41]

$$u(x,0) = \sqrt{c}\operatorname{sech}\left[k+\sqrt{c}x\right], \quad (14)$$

for all $c \geq 0$, where k is an arbitrary constant.

Again, to find the solution of this equation, simply take the equation in a operator form exactly in the same manner as the form of Eq. (6) and use (12) to find the zeroth component of $u_0 = u(x,0)$ and obtain sequential terms by using (12) with (11) to determine the other individual terms of the decomposition series with the aid of Mathematica, to get

$$\begin{aligned}
u(x,t) = {} & \sqrt{c}\operatorname{sech}\left(k+\sqrt{c}x\right) \\
& + \frac{1}{4}c^{\frac{7}{2}}t^2\left(-3+\cosh\left(2k+2\sqrt{c}x\right)\right)\operatorname{sech}^3\left(k+\sqrt{c}x\right) \\
& + \frac{1}{192}c^{\frac{13}{2}}t^4\left(115-76\cosh\left(2k+2\sqrt{c}x\right)\right. \\
& \left.+\cosh\left(4k+4\sqrt{c}x\right)\right)\operatorname{sech}^5\left(k+\sqrt{c}x\right) \\
& + \frac{1}{23040}\left\{c^{\frac{19}{2}}t^6\left(-11774+10543\cosh\left(2k+2\sqrt{c}x\right)\right.\right. \\
& \left.-722\cosh\left(4k+4\sqrt{c}x\right)+\cosh\left(6k+6\sqrt{c}x\right)\right) \\
& \left.\times\operatorname{sech}^7\left(k+\sqrt{c}x\right)\right\} \\
& + \frac{1}{4}c^5t^3\operatorname{sech}^4\left(k+\sqrt{c}x\right)\left(-23\sinh\left(k+\sqrt{c}x\right)\right. \\
& \left.+\sinh\left(3k+3\sqrt{c}x\right)\right)
\end{aligned}$$

$$+\frac{1}{1920}\Big\{c^8t^5\operatorname{sech}^6\left(k+\sqrt{c}x\right)\Big(1682\sinh\left(k+\sqrt{c}x\right)$$
$$-237\sinh\left(3k+3\sqrt{c}x\right)+\sinh\left(5k+5\sqrt{c}x\right)\Big)\Big\}$$
$$+\frac{1}{322560}\Big\{c^{11}t^7\operatorname{sech}^8\left(k+\sqrt{c}x\right)$$
$$\Big(-259723\sinh\left(k+\sqrt{c}x\right)+60657\sinh\left(3k+3\sqrt{c}x\right)$$
$$-2179\sinh\left(5k+5\sqrt{c}x\right)+\sinh\left(7k+7\sqrt{c}x\right)\Big)\Big\}$$
$$+c^2t\operatorname{sech}^3\left(k+\sqrt{c}x\right)\tanh\left(k+\sqrt{c}x\right)$$
$$+c^2t\operatorname{sech}\left(k+\sqrt{c}x\right)\tanh^3\left(k+\sqrt{c}x\right)+\cdots$$

Homotopy Analysis Method In this section, the homotopy analysis method (HAM) [44,46] is considered. HAM has been constructed and successfully implemented as an approximate and numerical solution for many types of nonlinear problems [18,19,45,46,47,48,49,58,59,60] and the references cited therein. A very nice explanation of the basic ideas of the HAM, its relationships with other analytic techniques, and some of its applications in science and engineering are given in Liao's book [46]. There are many groups of methods similar to HAM in the literature which are implemented nonlinear problems. Most of these groups of methods are in principle based on a Taylor series in an embedding parameter. If one could guess the initial function and the auxiliary linear operator well, then one can get very good approximations in a few terms, especially for a small value of the variable of the series.

For the purpose of illustration of the HAM [44], the mKdV equation is written in the operator form as

$$L(u)+\alpha\, u^m u_x+u_{xxx}=0\,, \tag{15}$$

where L is a linear operator: $L\equiv\partial/(\partial t)$. Equation (15) can be written in a nonlinear operator form as

$$N\left[\varphi\left(x,t;q\right)\right]=\frac{\partial\varphi\left(x,t;q\right)}{\partial t}+\alpha\varphi^m\left(x,t;q\right)\frac{\partial\varphi\left(x,t;q\right)}{\partial x}+\frac{\partial^3\varphi\left(x,t;q\right)}{\partial x^3}\,, \tag{16}$$

where $q\in[0,1]$ is an embedding parameter and $\varphi\left(x,t;q\right)$ is a function.

From $u(x,0)=U_0(x)$, $-\infty<x<\infty$, it is straightforward to express the solution u by a set of base functions

$$\{e_n(x)\,t^n,\ \ n\geq 0\}\,,$$

where $e_n(x)$ as a coefficient is a function with respect to x. This provides us with the so-called Rule of Solution Expression.

Following Liao's method [44,46], let $u(x,0)=U_0(x)$ indicate an initial guess of the exact solution u, $h\neq 0$, an auxiliary parameter, $H(x,t)\neq 0$ an auxiliary function. A zero-order deformation equation is constructed as

$$\left(1-q\right)L\left[\varphi\left(x,t;q\right)-u_0(x,t)\right]=qhH(x,t)\,N\left[\varphi\left(x,t;q\right)\right]\,, \tag{17}$$

with the initial condition

$$\varphi\left(x,0;q\right)=U_0(x)\,. \tag{18}$$

When $q=0$ and 1, the above equation has the solution

$$\varphi(x,t;0)=u_0(x,t) \tag{19}$$

and

$$\varphi(x,t;1)=u(x,t) \tag{20}$$

respectively.

Assume the auxiliary function $H(x,t)$ and the auxiliary parameter h are properly chosen so that $\varphi\left(x,t;q\right)$ can be expressed by the Taylor series

$$\varphi\left(x,t;q\right)=u_0(x,t)+\sum_{n=1}^{\infty}u_n(x,t)\,q^n\,, \tag{21}$$

where

$$u_n(x,t)=\frac{1}{n!}\frac{\partial^n\varphi\left(x,t;q\right)}{\partial q^n}\bigg|_{q=0} \tag{22}$$

and that the above series is convergent at $q=1$. Equations (19) and (20) then yield

$$u(x,t)=u_0(x,t)+\sum_{n=1}^{\infty}u_n(x,t)\,. \tag{23}$$

For the sake of simplicity, define the vectors

$$\vec{u}_n(x,t)=\{u_0(x,t),u_1(x,t),\ldots,u_n(x,t)\} \tag{24}$$

differentiating the zero-order deformation Eq. (17) n times with respect to the embedding parameter q, then setting $q=0$, and finally dividing by $n!$, the nth-order deformation equation is written as

$$L\left[u_n(x,t)-\chi_n u_{n-1}(x,t)\right]=hH(x,t)\,R_n\left[\vec{u}_{n-1}(x,t)\right]\,, \tag{25}$$

where

$$R_n\left[\vec{u}_{n-1}(x,t)\right] = \frac{1}{(n-1)!}\left\{\frac{\partial^{n-1}}{\partial q^{n-1}} N\left[\sum_{m=0}^{\infty} u_m(x,t)\, q^m\right]\right\}\Bigg|_{q=0} \tag{26}$$

and

$$\chi_n = \begin{cases} 0, & n \le 1 \\ 1, & n > 1 \end{cases} \tag{27}$$

with the initial condition

$$u_n(x,0) = 0, \qquad n \ge 1\,. \tag{28}$$

Therefore the nth order approximation of $u(x,t)$ is given by

$$u(x,t) \approx u_0(x,t) + \sum_{m=1}^{N} u_m(x,t)\,. \tag{29}$$

Example 1 Let's consider the mKdV equation (6) for the value of $m = 1$, which is called the classical KdV equation, and take this equation with the following initial value condition

$$u(x,0) = u_0 = \frac{3\lambda}{\alpha}\left(1 - \tanh^2\left(\frac{\sqrt{\lambda}}{2}x\right)\right)\,. \tag{30}$$

All related formulae are the same as those given from (25),

$$R_n\left[u_{n-1}\right] = \frac{\partial u_{n-1}}{\partial t} + \alpha \sum_{i=0}^{n-1} u_i \frac{\partial u_{n-1-i}}{\partial x} + \frac{\partial^3 u_{n-1}}{\partial x^3}\,, \tag{31}$$

using (24) and (25) with (31) and the initial function (30), three terms of the series (29) can be calculated as

$$\begin{aligned}
u(x,t) = {} & \frac{3\lambda}{\alpha}\left(1 - \tanh^2\left(\frac{\sqrt{\lambda}}{2}x\right)\right) \\
& - \frac{3ht\lambda^{\frac{5}{2}}\operatorname{sech}^2\left(\frac{x\sqrt{\lambda}}{2}\right)\tanh\left(\frac{x\sqrt{\lambda}}{2}\right)}{\alpha} \\
& + \frac{1}{4\alpha}\left\{3ht\lambda^{\frac{5}{2}}\operatorname{sech}^4\left(\frac{x\sqrt{\lambda}}{2}\right)\right. \\
& \times\left(-2 + \cosh\left(x\sqrt{\lambda}\right)\right) - 2(1+h)\sinh\left(x\sqrt{\lambda}\right)\Bigg\} \\
& + \frac{1}{8\alpha}\left\{ht\lambda^{\frac{5}{2}}\operatorname{sech}^2\left(\frac{x\sqrt{\lambda}}{2}\right)\right\{12h(1+h)t\lambda^{\frac{3}{2}} \\
& \times\left(-2 + \cosh\left(x\sqrt{\lambda}\right)\right)\operatorname{sech}^2\left(\frac{x\sqrt{\lambda}}{2}\right) \\
& - h^2t^2\lambda^3\operatorname{sech}^3\left(\frac{x\sqrt{\lambda}}{2}\right) \\
& \times\left(-11\sinh\left(\frac{x\sqrt{\lambda}}{2}\right) + \sinh\left(\frac{3x\sqrt{\lambda}}{2}\right)\right) \\
& -24(1+h)^2\tanh\left(\frac{x\sqrt{\lambda}}{2}\right)\Bigg\}\Bigg\} \\
& + \frac{1}{64\alpha}\left\{ht\lambda^{\frac{5}{2}}\operatorname{sech}^2\left(\frac{x\sqrt{\lambda}}{2}\right)\right. \\
& \times\left\{144h(1+h)^2t\lambda^{\frac{3}{2}}\left(-2 + \cosh\left(x\sqrt{\lambda}\right)\right)\right. \\
& \times\operatorname{sech}^2\left(\frac{x\sqrt{\lambda}}{2}\right) \\
& + h^3t^3\lambda^{\frac{9}{2}}\left(33 - 26\cosh\left(x\sqrt{\lambda}\right) + \cosh\left(2x\sqrt{\lambda}\right)\right) \\
& \times\operatorname{sech}^4\left(\frac{x\sqrt{\lambda}}{2}\right) - 24h^2(1+h)t^2\lambda^3 \\
& \times\operatorname{sech}^3\left(\frac{x\sqrt{\lambda}}{2}\right) \\
& \times\left(-11\sinh\left(\frac{x\sqrt{\lambda}}{2}\right) + \sinh\left(\frac{3x\sqrt{\lambda}}{2}\right)\right) \\
& \left.\left. - 192(1+h)^3\tanh\left(\frac{x\sqrt{\lambda}}{2}\right)\right\}\right\} + \cdots
\end{aligned}$$

Example 2 In this example, the mKdV equation (6) is considered for the values of $m = 2$ and $\alpha = 6$. It has the traveling wave solution which can be obtained subject to the initial condition

$$u(x,0) = \sqrt{c}\operatorname{sech}\left[k + \sqrt{c}x\right]. \tag{32}$$

For implementation of the HAM for this particular equation with the initial function (32), all related formulae are the same as those given from (15) and (25) with the following $R_n\left[u_{n-1}\right]$:

$$R_n\left[u_{n-1}\right] = \frac{\partial u_{n-1}}{\partial t} + \alpha \sum_{i=0}^{n-1}\sum_{j=0}^{i} u_j u_{i-j} \frac{\partial u_{n-1-i}}{\partial x} + \frac{\partial^3 u_{n-1}}{\partial x^3}\,. \tag{33}$$

Using (24) and (25) with (33) by with the initial function (32), we have

$$\begin{aligned}
u(x,t) &= \sqrt{c}\,\mathrm{sech}\left[k+\sqrt{c}x\right] \\
&- c^2 ht\,\mathrm{sech}\left[k+\sqrt{c}x\right]\tanh\left[k+\sqrt{c}x\right] \\
&+ \frac{1}{4}c^2 ht\,\mathrm{sech}\left[k+\sqrt{c}x\right] \\
&\times\left(2c^{\frac{3}{2}}ht\left(1-2\,\mathrm{sech}^2\left(k+\sqrt{c}x\right)\right)\right. \\
&\left.-4(1+h)\tanh\left(k+\sqrt{c}x\right)\right) \\
&+ \frac{1}{24}c^2 ht\,\mathrm{sech}\left(k+\sqrt{c}x\right) \\
&\times\left(12c^{\frac{3}{2}}h(1+h)t\left(-3+\cosh\left(2\left(k+\sqrt{c}x\right)\right)\right)\right) \\
&\times\mathrm{sech}^2\left(k+\sqrt{c}x\right)+c^3h^2t^2\,\mathrm{sech}^3\left(k+\sqrt{c}x\right) \\
&\times\left(23\sinh\left(k+\sqrt{c}x\right)-\sinh\left(3\left(k+\sqrt{c}x\right)\right)\right) \\
&\qquad -24(1+h)^2\tanh\left(k+\sqrt{c}x\right)+\cdots
\end{aligned}$$

which are the three terms of the approximate series solution (29).

Homotopy Perturbation Method

In 1998, Liao's early idea to construct the one-parameter family of equations was introduced by He's so-called "homotopy perturbation method" [20,21,23,24], which is only a special case of the homotopy analysis method [45]. But this homotopy perturbation method is almost the same as Liao's early one-parameter family equation which can be found in [44]. There are some minor differences, and detailed discussions and proofs can be seen in literature [20,21,23,24]. The authors in [20,21,23,24], have proved that these two methods lead to the same solutions of the considered equation for high-order approximations. This is also mainly because of the Taylor series properties. In conclusion of this, nothing is new in He's idea, except the new name "homotopy perturbation method" [58].

To illustrate HPM, consider the following nonlinear differential equation

$$A(u)-f(r)=0\,,\quad r\in\Omega \tag{34}$$

with the boundary condition

$$B\left(u,\frac{\partial u}{\partial n}\right)=0\,,\quad r\in\Gamma\,, \tag{35}$$

where $A(u)$ is written as follows:

$$A(u)=L(u)+N(u)\ . \tag{36}$$

A is a general differential operator, B is a boundary operator, $f(r)$ is a given analytic right hand side function, and Γ is the boundary of the domain Ω. The operator A can be generally divided into two parts L and N, where L is linear and N is the nonlinear term. So, Eq. (34) can be rewritten as follows:

$$L(u)+N(u)-f(r)=0\,. \tag{37}$$

By the homotopy technique [45], a homotopy $v(r,p):\Omega\times[0,1]\to\Re$ is obtained which satisfies

$$\begin{aligned}
&\mathrm{H}(v,p) \\
&=(1-p)\left[L(v)-L(u_0)\right]+p\left[A(v)-f(r)\right] \\
&\qquad =0\,,\quad p\in[0,1]\,,\quad r\in\Omega\,,
\end{aligned} \tag{38}$$

where $p\in[0,1]$ is an embedding parameter, and u_0 is an initial approximation of Eq. (34) which satisfies the boundary conditions. Obviously, from (38) the result will be

$$\begin{aligned}
H(v,0)&=L(v)-L(u_0)\quad=0\,,\\
H(v,1)&=A(v)-f(r)\quad=0\,,
\end{aligned}$$

changing process of p from zero to unity is just that of $v(r,p)$ from $u_0(r)$ to $u(r)$. In topology, this is called deformation, and $L(v)-L(u_0)$ and $A(v)-f(r)$ are called homotopy.

We consider v as the following:

$$v=v_0+pv_1+p^2v_2+p^3v_3+\cdots\,. \tag{39}$$

According to HPM, the best approximation solution of Eq. (37) can be explained as a series of the power of p,

$$u=\lim_{p\to1}v=v_0+v_1+v_2+v_3+\cdots\,. \tag{40}$$

The above convergence is given in [21]. Some results have been discussed in [20,23,24].

Example 1 Here, the mKdV equation (6) is again considered for the value of $m=1$, which is called classical KdV equation; first we construct a homotopy as follows:

$$(1-p)\left[\dot{Y}-\dot{u_0}\right]+p\left[\dot{Y}+\alpha YY'+Y'''\right]=0\,, \tag{41}$$

where

$$\dot{Y}=\frac{\partial Y}{\partial t}\,,\ Y'=\frac{\partial Y}{\partial x}\,,\ Y'''=\frac{\partial^3 Y}{\partial x^3}\quad\text{and}\quad p\in[0,1]\ .$$

With the initial approximation

$$Y_0=u_0=\frac{3\lambda}{\alpha}-\frac{3\lambda}{\alpha}\tanh^2\left(\frac{\sqrt{\lambda}}{2}x\right)\,,$$

suppose the solution of Eq. (41) has the form:

$$Y = Y_0 + pY_1 + p^2Y_2 + p^3Y_3 + \dots = \sum_{n=0}^{\infty} p^n Y_n(x,t). \quad (42)$$

Substituting Eq. (42) into Eq. (41) and arranging the coefficients of "p" powers, we have

$$p^0:\ \dot{Y}_0 - \dot{u}_0 = 0, \quad (43)$$

$$p^1:\ \dot{Y}_1 + \dot{u}_0 + \alpha Y_0 Y_0' + Y_0''' = 0, \quad (44)$$

$$p^2:\ \dot{Y}_2 + \alpha Y_0 Y_1' + \alpha Y_1 Y_0' + Y_1''' = 0, \quad (45)$$

$$p^3:\ \dot{Y}_3 + \alpha Y_0 Y_2' + \alpha Y_1 Y_1' + \alpha Y_2 Y_0' + Y_2''' = 0, \quad (46)$$

$$\vdots$$

and finally using Mathematica, the solutions of the equation can be obtained as follows:

$$Y_0 = \frac{3\lambda}{\alpha} - \frac{3\lambda}{\alpha}\tanh^2\left(\frac{\sqrt{\lambda}}{2}x\right), \quad (47)$$

$$Y_1 = \frac{3t\lambda^{\frac{5}{2}}}{\alpha}\operatorname{sech}^2\left(\frac{x\sqrt{\lambda}}{2}\right)\tanh\left(\frac{x\sqrt{\lambda}}{2}\right) \quad (48)$$

$$Y_2 = \frac{3t^2\lambda^4}{4\alpha}\left(-2+\cosh\left(x\sqrt{\lambda}\right)\right)\operatorname{sech}^4\left(\frac{x\sqrt{\lambda}}{2}\right), \quad (49)$$

$$Y_3 = \frac{3t^3\lambda^{\frac{11}{2}}}{8\alpha}\operatorname{sech}^5\left(\frac{x\sqrt{\lambda}}{2}\right) \times\left(-11\sinh\left(\frac{x\sqrt{\lambda}}{2}\right)+\sinh\left(\frac{3x\sqrt{\lambda}}{2}\right)\right) \quad (50)$$

$$\vdots$$

The above terms of the series (42) could be calculated. When the series (42) is considered with the terms (43)–(46) and supposing $p = 1$, an approximate series solution of the considered KdV equation is obtained as follows:

$$u(x,t) = \frac{3\lambda}{\alpha} - \frac{3\lambda}{\alpha}\tanh^2\left(\frac{\sqrt{\lambda}}{2}x\right) + \frac{3t\lambda^{\frac{5}{2}}}{\alpha}\operatorname{sech}^2\left(\frac{x\sqrt{\lambda}}{2}\right)\tanh\left(\frac{x\sqrt{\lambda}}{2}\right) + \frac{3t^2\lambda^4}{4\alpha}\left(-2+\cosh\left(x\sqrt{\lambda}\right)\right)\operatorname{sech}^4\left(\frac{x\sqrt{\lambda}}{2}\right) + \frac{3t^3\lambda^{\frac{11}{2}}}{8\alpha}\operatorname{sech}^5\left(\frac{x\sqrt{\lambda}}{2}\right) \left(-11\sinh\left(\frac{x\sqrt{\lambda}}{2}\right)+\sinh\left(\frac{3x\sqrt{\lambda}}{2}\right)\right) + \cdots$$

Example 2 Let's consider the mKdV equation (6) with the values of $m = 2$ and $\alpha = 6$. For this method's implementation, a homotopy is constructed as follows:

$$(1-p)\left[\dot{Y} - \dot{u}_0\right] + p\left[\dot{Y} + 6Y^2Y' + Y'''\right] = 0. \quad (51)$$

With the initial approximation

$$Y_0 = u_0 = \sqrt{c}\operatorname{sech}\left[k + \sqrt{c}x\right],$$

suppose the solution of Eq. (41) has the form:

$$Y = Y_0 + pY_1 + p^2Y_2 + p^3Y_3 + \dots = \sum_{n=0}^{\infty} p^n Y_n(x,t). \quad (52)$$

Substituting Eq. (52) into Eq. (51) and arranging the coefficients of "p" powers, we have

$$p^0:\ \dot{Y}_0 - \dot{u}_0 = 0, \quad (53)$$

$$p^1:\ \dot{Y}_1 + \dot{u}_0 + 6Y_0^2Y_0' + Y_0''' = 0, \quad (54)$$

$$p^2:\ \dot{Y}_2 + 12Y_0Y_1Y_0' + 6Y_0^2Y_1' + Y_1''' = 0, \quad (55)$$

$$p^3:\ \dot{Y}_3 + 12Y_0Y_2Y_0' + 6Y_1^2Y_0' + 12Y_0Y_1Y_1' + Y_2''' = 0 \quad (56)$$

$$\vdots$$

Finally, using Mathematica, the few terms of the series solution of the Eq. (51) can be obtained as follows:

$$u(x,t) = \sqrt{c}\operatorname{sech}\left(k+\sqrt{c}x\right) + \frac{1}{4}c^{\frac{7}{2}}t^2\left(-3+\cosh\left(2\left(k+\sqrt{c}x\right)\right)\right) \times\operatorname{sech}^3\left(k+\sqrt{c}x\right) + \frac{1}{24}c^5t^3\operatorname{sech}^4\left(k+\sqrt{c}x\right) \times\left(-23\sinh\left(k+\sqrt{c}x\right)+\sinh\left(3\left(k+\sqrt{c}x\right)\right)\right) + c^2t\operatorname{sech}\left(k+\sqrt{c}x\right)\tanh\left(k+\sqrt{c}x\right) + \cdots$$

Variational Iteration Method

In this section, a kind of semi-analytical technique is considered for a non-linear problem called the variational iteration method (VIM) [20,22,25]. It is proposed by He [20], and the method is based on the use of restricted variations and correction functionals. In this technique, a correction functional is defined by a general Lagrange multiplier using the ingenuity of variational theory. This method is implemented to get approximate solutions for many linear and non-linear problems in physics and mathematics [2,3,26,53,55,68]. When the method is implemented on the differential equation, it does not require the presence of small parameters and the solution of the equation is obtained as a sequence of iterates. The method does not require that the nonlinearities be differentiable with respect to the dependent variable and its derivatives [2,3,26,53,55,68].

To illustrate the VIM, let's consider the nonlinear KdV equation (6):

$$Lu + Nu = 0 \,, \tag{57}$$

where L and N are linear and nonlinear operators, respectively. In [2,3,26,53,55,68], He proposed the VIM where a correction functional for Eq. (57) can be written as:

$$u_{n+1}(x,t) = u_n(x,t) + \int_0^t \lambda \left\{ Lu_n(x,\tau) + \tilde{N}u_n(x,\tau) \right\} \mathrm{d}\tau \,, \quad n \geq 0 \,, \tag{58}$$

where λ is a general Lagrange multiplier [20], which can be identified optimally via the variational theory, and $\tilde{u}_n$is a restricted variation, which means $\delta\tilde{u}_n = 0$. It is required first to determine the Lagrangian multiplier λ that will be identified optimally via integration by parts. The successive approximations $u_{n+1}(x,t)$, $n \geq 0$, of the solution $u(x,t)$ will be readily obtained upon using the Lagrangian multiplier obtained and by using any selective function u_0. The initial values $u(x,0)$ and $u_t(x,0)$ are usually used for the selective zeroth approximation u_0. Having λ determined, then several approximations $u_j(x,t)$, $j \geq 0$, can be determined. Consequently, the solution is given by

$$u = \lim_{n\to\infty} u_n \,. \tag{59}$$

In what follows, the VIM will apply to two nonlinear KdV equations with similar initial functions to illustrate the strength of the method and to compare the given above methods.

Example 1 Let's consider the mKdV equation (6) for the value of $m = 1$ with the following initial function

$$u(x,0) = \frac{3\lambda}{\alpha} - \frac{3\lambda}{\alpha}\tanh^2\left(\frac{\sqrt{\lambda}}{2}x\right) \,, \tag{60}$$

where λ, α are arbitrary constant. The correction functional for this equation is

$$u_{n+1}(x,t) = u_n(x,t) + \int_0^t \lambda \Big\{ \big(u_n(x,\tau)\big)_\tau + \frac{\alpha}{2}\big(\tilde{u}_n^2(x,\tau)\big)_x + \big(u_n(x,\tau)\big)_{xxx} \Big\} \mathrm{d}\tau \,, \quad n \geq 0 \,, \tag{61}$$

where λ is the general Lagrange multiplier [20] whose optimal value is found using variational theory. u_0 is an initial solution with or without unknown parameters. In the case of no unknown parameters, u_0 should satisfy initial-boundary conditions and $\tilde{u}_n^2$ is the restricted variation [22], i. e., $\tilde{u}_n^2 = 0$. To find the optimal value of λ, we have

$$\delta u_{n+1}(x,t) = \delta u_n(x,t) + \delta \int_0^t \lambda \Big\{ \big(u_n(x,\tau)\big)_\tau + \frac{\alpha}{2}\big(\tilde{u}_n^2(x,\tau)\big)_x + \big(u_n(x,\tau)\big)_{xxx} \Big\} \mathrm{d}\tau \,, \tag{62}$$

and this yields the stationary conditions

$$\lambda'(\tau) = 0 \,, \quad 1 + \lambda(\tau)|_{\tau=t} = 0 \,.$$

This in turn gives $\lambda = -1$ therefore, Eq. (51) can be written as following

$$u_{n+1}(x,t) = u_n(x,t) - \int_0^t \Big\{ \big(u_n(x,\tau)\big)_\tau + \frac{\alpha}{2}\big(\tilde{u}_n^2(x,\tau)\big)_x + \big(u_n(x,\tau)\big)_{xxx} \Big\} \mathrm{d}\tau \,. \tag{63}$$

Substituting the value of $m = 1$ into the mKdV equation (6) with an initial value (60), and then substituting this into Eq. (63) and using Mathematica, the solutions of the Eq. (6) with initial value (60) can be obtained as follows:

$$u(x,t) = \frac{1}{560\alpha\left(1 + \cosh\left(x\sqrt{\lambda}\right)\right)}$$

$$\times \left\{ 105\lambda \left\{ 32 + t^4\lambda^6 \operatorname{sech}^2\left(\frac{x\sqrt{\lambda}}{2}\right) \right.\right.$$
$$\times \left\{ 88 + 4\left(-149 + 6t^2\lambda^3\right) \operatorname{sech}^2\left(\frac{x\sqrt{\lambda}}{2}\right) \right.$$
$$+ \left(1095 - 124t^2\lambda^3\right) \operatorname{sech}^4\left(\frac{x\sqrt{\lambda}}{2}\right)$$
$$+ 198\left(-3 + t^2\lambda^3\right) \operatorname{sech}^6\left(\frac{x\sqrt{\lambda}}{2}\right)$$
$$\left.\left. -99t^2\lambda^3 \operatorname{sech}^8\left(\frac{x\sqrt{\lambda}}{2}\right)\right\}\right\} - 4t\lambda^{\frac{5}{2}} \left\{ -140\left(6 + t^2\lambda^3\right)\right.$$
$$+ 9t^2\lambda^3 \operatorname{sech}^2\left(\frac{x\sqrt{\lambda}}{2}\right) \left\{ 140 - 141t^2\lambda^3 \right.$$
$$- 35\left(4 + 3t^2\lambda^3\right) \operatorname{sech}^2\left(\frac{x\sqrt{\lambda}}{2}\right)$$
$$+ \left(497t^2\lambda^3 - 80t^4\lambda^6\right) \operatorname{sech}^4\left(\frac{x\sqrt{\lambda}}{2}\right)$$
$$+ 20t^2\lambda^3\left(-21 + 20t^2\lambda^3\right) \operatorname{sech}^6\left(\frac{x\sqrt{\lambda}}{2}\right)$$
$$- 630t^4\lambda^6 \operatorname{sech}^8\left(\frac{x\sqrt{\lambda}}{2}\right)$$
$$\left.\left. + 315t^4\lambda^6 \operatorname{sech}^{10}\left(\frac{x\sqrt{\lambda}}{2}\right)\right\}\right\}$$
$$\left.\times \tanh\left(\frac{x\sqrt{\lambda}}{2}\right)\right\} + \cdots$$

Example 2 Again the mKdV equation (6) is considered with the values of $m = 2$ and $\alpha = 6$ and the following initial function

$$u(x, 0) = \sqrt{c} \operatorname{sech}\left[k + \sqrt{c}x\right] . \tag{64}$$

The correction functional for this equation is

$$u_{n+1}(x, t) = u_n(x, t) - \int_0^t \left\{ \left(u_n(x, \tau)\right)_\tau + 2\left(\tilde{u}_n^3(x, \tau)\right)_x + \left(u_n(x, \tau)\right)_{xxx} \right\} d\tau , \quad n \geq 0 . \tag{65}$$

Substituting the value of $m = 2$ into the mKdV equation (6) with initial value (64), and then substituting this into Eq. (65) and using Mathematica, the solutions of the considered equation with initial value (64) can be obtained as follows:

$$u(x, t) = \frac{1}{4} \operatorname{sech}\left(k + \sqrt{c}x\right)$$
$$\times \left\{ 2c^{\frac{7}{2}} t^2 \left(1 - 2\operatorname{sech}^2\left(k + \sqrt{c}x\right)\right) \right.$$
$$+ 2c^5 t^3 \operatorname{sech}^5\left(k + \sqrt{c}x\right)$$
$$\times \left(-17 \sinh\left(k + \sqrt{c}x\right) + 3\sinh\left(3\left(k + \sqrt{c}x\right)\right)\right)$$
$$+ 3c^{\frac{13}{2}} t^4 \left(-3 + \cosh\left(2\left(k + \sqrt{c}x\right)\right)\right)$$
$$\times \operatorname{sech}^4\left(k + \sqrt{c}x\right) \tanh^2\left(k + \sqrt{c}x\right)$$
$$\left. + 4\left(\sqrt{c} + c^2 t \tanh\left(k + \sqrt{c}x\right)\right)\right\} + \cdots$$

We have just written the series solution for the mKdV as three terms rather than four terms.

Numerical Experiments For numerical comparisons purposes, two KdV equations are considered. The first one is the classical KdV and the second one is the modified KdV equation. The formula of numerical results for ADM, HAM and HPM are given as follows

$$\lim_{n\to\infty} \phi_n = u(x, t)$$
$$\text{where} \quad \phi_n(x, t) = \sum_{k=0}^{n} u_k(x, t) , \quad n \geq 0 . \tag{66}$$

The formula of numerical results for VIM is as (59). The recurrence relations of the methods are given as in (12), (25), (43)–(46) and (58), respectively.

Moreover, the decomposition series of the KdV equation's solutions generally converge very rapidly in real physical problems [5,6,64]. Results were obtained about the speed of convergence of this method, providing us methods to solve linear and nonlinear functional equations. The convergence of the HAM and VIM are numerically shown in some works [3,55] and references there in. Here, how all these methods are converged to their corresponding exact solutions has been proved.

Numerical approximations show a high degree of accuracy and in most cases of ϕ_n, the n-term approximation is accurate for quite low values of n. The solutions are very rapidly convergent by utilizing the ADM, VIM, HPM and HAM. The obtained numerical results justify the advantage of these methodologies, and even with few terms the approximation is accurate. Furthermore, as the all these methods do not require discretization of the variables, i. e.

Korteweg–de Vries Equation (KdV) and Modified KdV (mKdV), Semi-analytical Methods for Solving the, Table 1
Comparison between the absolute error of the solution of KdV equation (6) ($m = 1$) at various values of t and $x = 0.5$ for $\alpha = 1$, $\lambda = 0.01$ in ADM, VIM and HAM ($h = -1$)

t	$x = 0$			$x = 5$		
	ADM	HAM	VIM	ADM	HAM	VIM
0.1	2.9925×10^{-8}	2.9925×10^{-8}	$3. \times 10^{-8}$	0.0000130999	0.0000145274	0.0000145274
0.2	1.197×10^{-7}	1.197×10^{-7}	1.2×10^{-7}	0.0000261535	0.0000291008	0.000029101
0.3	2.69323×10^{-7}	2.69323×10^{-7}	2.69998×10^{-7}	0.0000391608	0.0000437201	0.0000437206
0.4	4.78795×10^{-7}	4.78795×10^{-7}	4.79995×10^{-7}	0.0000521216	0.0000583853	0.0000583862
0.5	7.48113×10^{-7}	7.48113×10^{-7}	7.49987×10^{-7}	0.000065036	0.0000730962	0.0000730976
0.6	1.07727×10^{-6}	1.07727×10^{-6}	1.07997×10^{-6}	0.0000779036	0.0000878527	0.0000878548
0.7	1.46628×10^{-6}	1.46628×10^{-6}	1.46995×10^{-6}	0.0000907246	0.000102655	0.000102658
0.8	1.91512×10^{-6}	1.91512×10^{-6}	1.91992×10^{-6}	0.000103499	0.000117502	0.000117506
0.9	2.42379×10^{-6}	2.42379×10^{-6}	2.42987×10^{-6}	0.000116226	0.000132395	0.0001324
1.	2.9923×10^{-6}	2.9923×10^{-6}	2.9998×10^{-6}	0.000128906	0.000147333	0.000147339

t	$x = 10$			$x = 15$		
	ADM	HAM	VIM	ADM	HAM	VIM
0.1	0.0000207071	0.0000229046	0.0000229046	0.0000216022	0.0000238682	0.0000238682
0.2	0.0000413971	0.000045826	0.0000458261	0.0000432119	0.0000477289	0.0000477289
0.3	0.0000620701	0.000068764	0.0000687642	0.0000648289	0.0000715819	0.0000715819
0.4	0.0000827257	0.0000917187	0.000091719	0.0000864532	0.0000954272	0.0000954271
0.5	0.000103364	0.00011469	0.00011469	0.000108085	0.000119265	0.000119265
0.6	0.000123985	0.000137677	0.000137678	0.000129723	0.000143094	0.000143094
0.7	0.000144588	0.000160681	0.000160682	0.000151369	0.000166916	0.000166916
0.8	0.000165173	0.0001837	0.000183702	0.000173022	0.00019073	0.000190729
0.9	0.000185741	0.000206736	0.000206738	0.000194682	0.000214535	0.000214535
1.	0.00020629	0.000229787	0.00022979	0.000216348	0.000238333	0.000238332

Korteweg–de Vries Equation (KdV) and Modified KdV (mKdV), Semi-analytical Methods for Solving the, Table 2
Comparison between the absolute error of the solution for Eq. (6) ($m = 1$) by HAM and HPM at various values of t and $x = 0.5$ with different values of h for HAM, in fact HPM is the value of $h = -1$ for HAM

t	HPM	HAM ($h = -1$)	HAM ($h = -1.4$)	HAM ($h = -0.9$)	HAM ($h = -0.6$)
0.1	1.60354×10^{-6}	1.60354×10^{-6}	1.60838×10^{-6}	1.60346×10^{-6}	1.59877×10^{-6}
0.2	3.26676×10^{-6}	3.26676×10^{-6}	3.27654×10^{-6}	3.26662×10^{-6}	3.25727×10^{-6}
0.3	4.98966×10^{-6}	4.98966×10^{-6}	5.00446×10^{-6}	4.98946×10^{-6}	4.97551×10^{-6}
0.4	6.77222×10^{-6}	6.77222×10^{-6}	6.79214×10^{-6}	6.77196×10^{-6}	6.75346×10^{-6}
0.5	8.61444×10^{-6}	8.61444×10^{-6}	8.63955×10^{-6}	8.61411×10^{-6}	8.59111×10^{-6}
0.6	0.0000105163	0.0000105163	0.0000105467	0.0000105159	0.0000104885
0.7	0.0000124777	0.0000124777	0.0000125135	0.0000124773	0.0000124455
0.8	0.0000144988	0.0000144988	0.0000145401	0.0000144984	0.0000144621
0.9	0.0000165795	0.0000165795	0.0000166263	0.000016579	0.0000165384
1.	0.0000187197	0.0000187197	0.0000187722	0.0000187192	0.0000186744

time and space, it is not effected by computation round off errors and one is not faced with the necessity of large computer memory and time.

In order to verify numerically whether the proposed methodologies lead to higher accuracy, the numerical solutions can be evaluated using the n-term approximation (66) and (59). The numerical results of the Eq. (6) (for KdV $m = 1$ and for mKdV $m = 2$) for the various values of x and t are illustrated in Tables 1 and 3 KdV and mKdV, respectively. These tabulated results show the differences

Korteweg–de Vries Equation (KdV) and Modified KdV (mKdV), Semi-analytical Methods for Solving the, Table 3
Comparison between the absolute error of the solution of Eq. (6) ($m = 2$) for $c = 1$, $k = -7$

$\lvert u_{\text{Exact}} - \varphi_3 \rvert$ for ADM					
$t \lvert x$	0.1	0.2	0.3	0.4	0.5
0.1	8.23236×10^{-9}	1.29161×10^{-7}	6.41349×10^{-7}	1.98865×10^{-6}	4.76447×10^{-6}
0.2	9.098×10^{-9}	1.42742×10^{-7}	7.08789×10^{-7}	2.19776×10^{-6}	5.26547×10^{-6}
0.3	1.00546×10^{-8}	1.57752×10^{-7}	7.83317×10^{-7}	2.42885×10^{-6}	5.81914×10^{-6}
0.4	1.11118×10^{-8}	1.74338×10^{-7}	8.65678×10^{-7}	2.68424×10^{-6}	6.43099×10^{-6}
0.5	1.228×10^{-8}	1.92667×10^{-7}	9.56694×10^{-7}	2.96645×10^{-6}	7.10715×10^{-6}

$\lvert u_{\text{Exact}} - u_3 \rvert$ for VIM					
$t \lvert x$	0.1	0.2	0.3	0.4	0.5
0.1	8.19516×10^{-9}	1.2858×10^{-7}	6.38476×10^{-7}	1.97978×10^{-6}	4.74334×10^{-6}
0.2	9.04779×10^{-9}	1.41958×10^{-7}	7.0491×10^{-7}	2.18579×10^{-6}	5.23696×10^{-6}
0.3	9.98686×10^{-9}	1.56692×10^{-7}	7.78082×10^{-7}	2.4127×10^{-6}	5.78065×10^{-6}
0.4	1.10203×10^{-8}	1.72908×10^{-7}	8.58611×10^{-7}	2.66243×10^{-6}	6.37904×10^{-6}
0.5	1.21566×10^{-8}	1.90738×10^{-7}	9.47155×10^{-7}	2.93702×10^{-6}	7.03703×10^{-6}

$\lvert u_{\text{Exact}} - \varphi_3 \rvert$ for HPM $\equiv$ HAM with value of $h = -1$					
$t \lvert x$	0.1	0.2	0.3	0.4	0.5
0.1	8.19516×10^{-9}	1.2858×10^{-7}	6.38476×10^{-7}	1.97978×10^{-6}	4.74334×10^{-6}
0.2	9.04779×10^{-9}	1.41958×10^{-7}	7.0491×10^{-7}	2.18579×10^{-6}	5.23696×10^{-6}
0.3	9.98686×10^{-9}	1.56692×10^{-7}	7.78082×10^{-7}	2.4127×10^{-6}	5.78065×10^{-6}
0.4	1.10203×10^{-8}	1.72908×10^{-7}	8.58611×10^{-7}	2.66243×10^{-6}	6.37904×10^{-6}
0.5	1.21566×10^{-8}	1.90738×10^{-7}	9.47155×10^{-7}	2.93702×10^{-6}	7.03703×10^{-6}

$\lvert u_{\text{Exact}} - \varphi_3 \rvert$ for HAM with value of $h = -0.9$					
$t \lvert x$	0.1	0.2	0.3	0.4	0.5
0.1	2.1809×10^{-9}	1.26484×10^{-7}	1.18355×10^{-7}	1.28924×10^{-7}	5.68147×10^{-7}
0.2	2.41029×10^{-9}	1.39787×10^{-7}	1.30802×10^{-7}	1.42489×10^{-7}	6.27919×10^{-7}
0.3	2.66382×10^{-9}	1.54489×10^{-7}	1.44558×10^{-7}	1.57481×10^{-7}	6.93984×10^{-7}
0.4	2.94403×10^{-9}	1.70737×10^{-7}	1.59761×10^{-7}	1.74053×10^{-7}	7.67005×10^{-7}
0.5	3.25373×10^{-9}	1.88694×10^{-7}	1.76562×10^{-7}	1.9237×10^{-7}	8.47719×10^{-7}

of the absolute errors between the exact solution and the numerical solution for the ADM, VIM and HAM. One can conclude that the values on the interval $0 \le x \le 15$ and for some small values of t all the considered methods give very close numerical solutions to a corresponding KdV equation. If one closely looks at Tables 1 and 3 then it is evident that the numerical results are almost the same for a few terms of the series solutions, such as the first three terms. The calculation for the later terms of the series has stopped because the scheme VIM does not work very well in the sense of computer time. This problem needs to be approached in a completely different way. Moreover, it has also been numerically proved that HPM is a special case of the HAM which is tabulated in Table 2 for the various values of h for interval $-1.4 \le h \le 0.6$ and similar constant values as in Table 1. It is numerically shown that when we take the value of $h = -1$ in the formula HAM, then (numerically) HAM $\equiv$ HPM in Table 2. It is also our experience that the choice of the h value depends on the chosen equation and the solution of the equation, too. This can be seen in Tables 2 and 3.

Overall, one can see that closeness of the approximate solutions and exact solutions of the all methods give almost equal absolute values of error, except for the approximate solutions from HAM. The numerical solutions with a special value of $h = -0.9$ for the HAM perform very well, as one can see in Table 3. It is also a valuable opportunity for us to use HAM among the considered methods because it gives the users flexibility to choose different values of h, depending on which value is necessary for the problem.

The graphs of the numerical values of the exact and approximate solutions of the classical KdV equation are

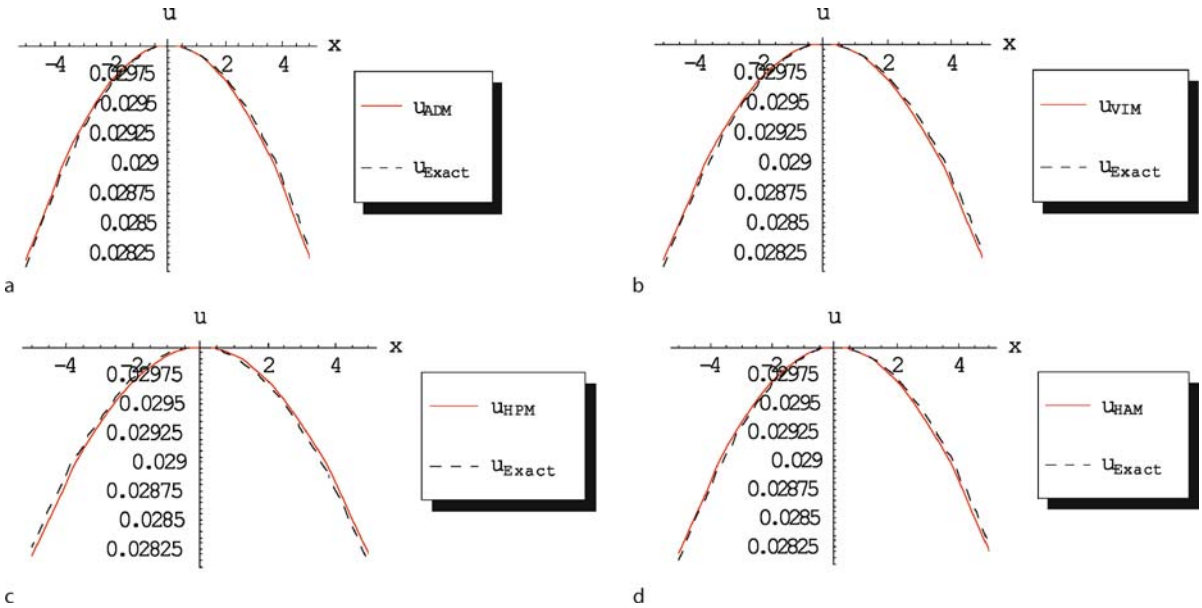

Korteweg–de Vries Equation (KdV) and Modified KdV (mKdV), Semi-analytical Methods for Solving the, Figure 3
A comparison by the ADM, VIM, HPM and HAM with exact solution of Eq. (6) with ($m = 1$) at $t = 0.5$ for $\alpha = 1$, $\lambda = 0.01$ in a–c and $\alpha = 1$, $\lambda = 0.01$, $h = -1$ in d, respectively

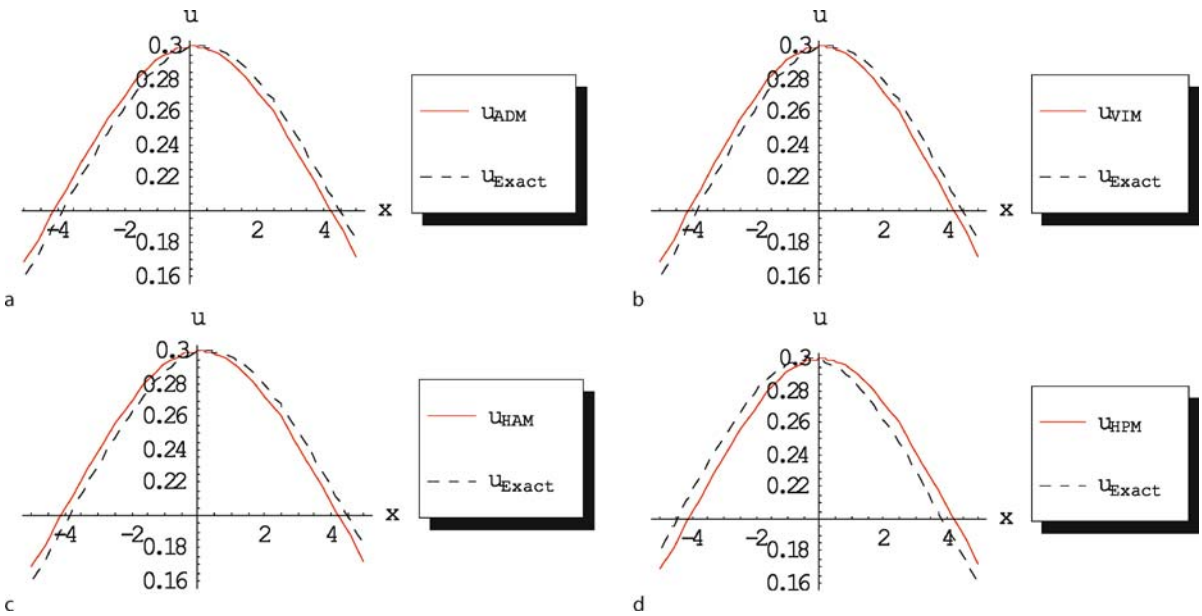

Korteweg–de Vries Equation (KdV) and Modified KdV (mKdV), Semi-analytical Methods for Solving the, Figure 4
A comparison by the ADM, VIM, HAM and HPM with exact solution of Eq. (6) with ($m = 1$) at $t = 0.5$ for $\alpha = 1$, $\lambda = 0.1$ in a–c and $\alpha = 1$, $\lambda = 0.1$, $h = -1$ in d, respectively

depicted in Figs. 3–6. In Figs. 3–5, and numerical results are given for the KdV equation and corresponding exact solution by using the considered methods. It is to be noted that only four terms were used in evaluating the approximate solutions with different values of the constant $\lambda = 0.01,\ 0.1,\ 1$ in order to get how close the approximate solutions were to the exact solution. A very good approximation to the actual solution of the equation was

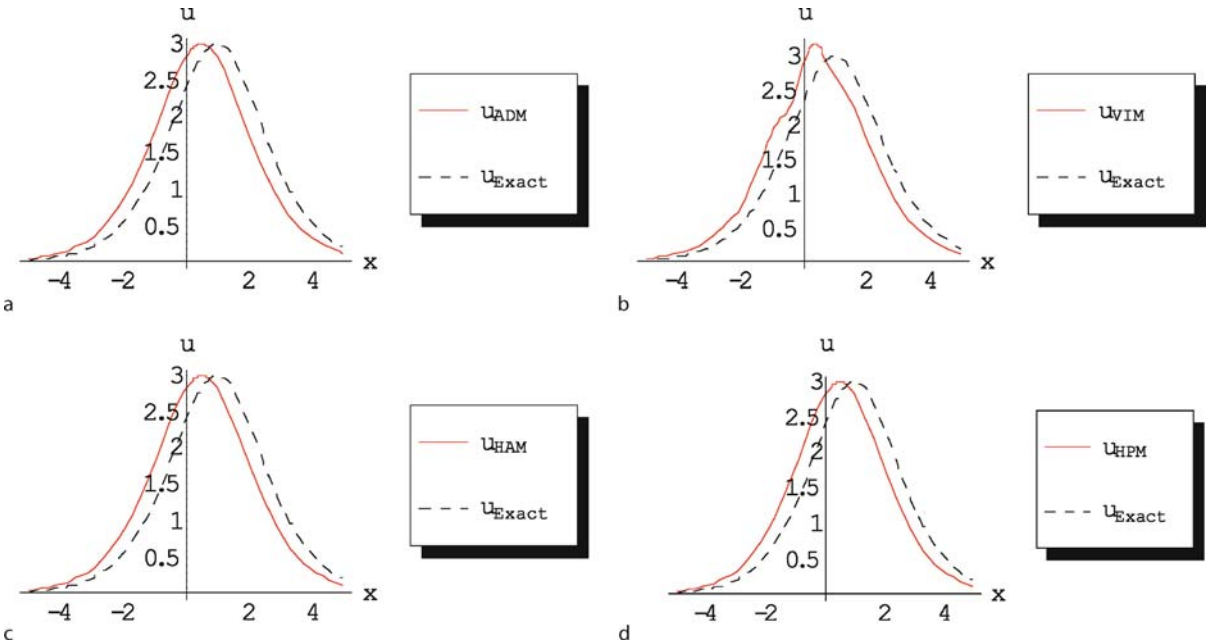

Korteweg–de Vries Equation (KdV) and Modified KdV (mKdV), Semi-analytical Methods for Solving the, Figure 5
A comparison by the ADM, VIM, HAM and HPM with exact solution of Eq. (6) with ($m = 1$) at $t = 0.5$ for $\alpha = 1,\ \lambda = 1$ in a–c and $\alpha = 1,\ \lambda = 1,\ h = -1$ in d, respectively

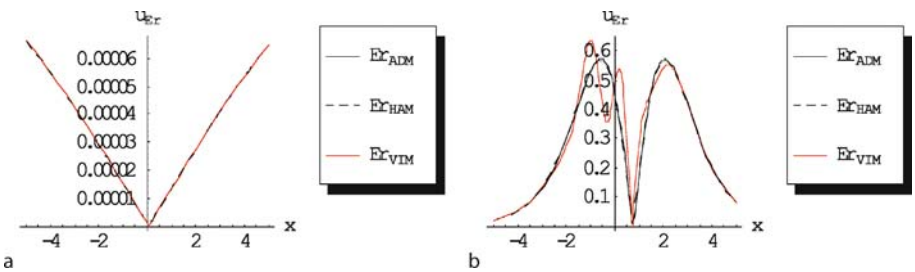

Korteweg–de Vries Equation (KdV) and Modified KdV (mKdV), Semi-analytical Methods for Solving the, Figure 6
These two graphs is a comparison of $u_{Er} = |u_{Exact} - u_{Approx}|$ for ADM, HAM and VIM of the solutions of Eq. (6) with ($m = 1$) at $t = 0.5$ for $\alpha = 1,\ \lambda = 0.01,\ h = -1$ in a and $\alpha = 1,\ \lambda = 1,\ h = -1$ in b, respectively

achieved by only using the four terms of the series derived above (59) and (66) for all considered methods with the value of $\lambda = 0.01$ which are depicted in Fig. 3. This is because of the nature of the series methods. The closeness of the numerical results for approximate solutions and the exact solution diverges for the other value of $\lambda = 0.1$ in all considered methods which are depicted in Fig. 4. This divergence is very clear in Fig. 5 for the results of the all the methods. But a sharp divergence can be seen in the numerical results of VIM especially at the value of $x = 0.5$, in Fig. 5b. In fact, these are illustrated in Tables 1–2 and Figs. 3–4. It is evident that the overall errors can be made smaller by adding new terms of the series (59) and (66).

One can see from Fig. 6a, the absolute values of the numerical results for approximate solutions and exact solutions are very small at the value of $0 \le x \le 1$ but other than this values of the x neither positive nor negative values of the x gives small errors for all considered methods with the value of $\lambda = 0.01$. However, in all considered methods with the value of $\lambda = 1$ a relatively small absolute error is given at the value of $0 \le x \le 1$, but all the methods are given biggest absolute error at a value of x around ± 2. The VIM is poorly performed at the value of x around -2,

but besides this all the other methods are performed relatively well at the other values of the x, which can be seen in Fig. 4b.

Lastly, the clear conclusion can be draw from the numerical results that the ADM and HAM algorithms provide highly accurate numerical solutions without spatial discretizations for nonlinear partial differential equations. It is also worth noting that the advantage of the approximation of the series methodologies displays a fast convergence of the solutions. The illustrations show the dependence of the rapid convergence depends on the character and behavior of the solutions just as in a closed form solutions. Finally, it can be pointed out that, for given equations with initial values $u(x, 0)$, the corresponding analytical and numerical solutions are obtained according to the recurrence relations (12), (25), (43)–(46) and (58) using Mathematica package version of Mathematica 4 in PC computer.

Future Directions

Nonlinear phenomena play a crucial role in applied mathematics and physics. Furthermore, when an original nonlinear equation is directly calculated, the solution will preserve the actual physical characters of solutions. Explicit solutions to the nonlinear equations are of fundamental importance. Various effective methods have been developed to understand the mechanisms of these physical models, to help physicists and engineers, and to ensure knowledge for physical problems and their applications.

In the future, the scientist will be very busy doing many works via applications of the nonlinear evolution equations in order to obtain exact and numerical solutions. One has to find out more efficient and effective methods to solve the nonlinear problems in applied mathematical areas of constructing either an exact solution or a numerical solution. Scientists have long way to go in this nonlinear study.

Bibliography

Primary Literature

1. Abbasbandy S (2006) The application of homotopy analysis method to nonlinear equations arising in heat transfer. Phys Lett A 360:109–13
2. Abdou MA, Soliman AA (2005) New applications of variational iteration method. Phys D 211:1–8
3. Abdou MA, Soliman AA (2005) Variational iteration method for solving Burger's and coupled Burger's equations. J Comput Appl Math 181:245–251
4. Adomian G (1988) A Review of the Decomposition Method in Applied Mathematics. J Math Anal Appl 135:501–544
5. Adomian G, Rach R (1992) Noise Terms in Decomposition Solution Series. Comput Math Appl 24:61–64
6. Adomian G (1994) Solving Frontier Problems of Physics: The decomposition method. Kluwer, Boston
7. Chowdhury MSH, Hashim I, Abdulaziz O (2009) Comparison of homotopy analysis method and homotopy-perturbation method for purely nonlinear fin-type problems. Commun Nonlinear Sci Numer Simul 14:371–378
8. Debtnath L (1997) Nonlinear Partial Differential Equations for Scientist and Engineers. Birkhauser, Boston
9. Debtnath L (2007) A brief historical introduction to solitons and the inverse scattering transform – a vision of Scott Russell. Int J Math Edu Sci Techn 38:1003–1028
10. Domairry G, Nadim N (2008) Assessment of homotopy analysis method and homotopy perturbation method in non-linear heat transfer equation. Intern Commun Heat Mass Trans 35:93–102
11. Drazin PG, Johnson RS (1989) Solutions: An Introduction. Cambrige University Press, Cambridge
12. Edmundson DE, Enns RH (1992) Bistable light bullets. Opt Lett 17:586
13. Fermi E, Pasta J, Ulam S (1974) Studies of nonlinear problems. American Math Soci, Providence, pp 143–156
14. Gardner CS, Greene JM, Kruskal MD, Miura RM (1967) Method for solving the Korteweg–de Vries equation. Phys Rev Lett 19:1095–1097
15. Gardner CS, Greene JM, Kruskal MD, Miura RM (1974) Korteweg–de Vries equation and generalizations, VI, Methods for exact solution. Commun Pure Appl Math 27:97–133
16. Geyikli T, Kaya D (2005) An application for a Modified KdV equation by the decomposition method and finite element method. Appl Math Comp 169:971–981
17. Geyikli T, Kaya D (2005) Comparison of the solutions obtained by B-spline FEM and ADM of KdV equation. Appl Math Comp 169:146–156
18. Hayat T, Abbas Z, Sajid M (2006) Series solution for the upper-convected Maxwell fluid over a porous stretching plate. Phys Lett A 358:396–403
19. Hayat T, Sajid M (2007) On analytic solution of thin film flow of a fourth grade fluid down a vertical cylinder. Phys Lett A 361:316–322
20. He JH (1997) A new approach to nonlinear partial differential equations. Commun Nonlinear Sci Numer Simul 2:203–205
21. He JH (1999) Homotopy perturbation technique. Comput Math Appl Mech Eng 178:257–262
22. He JH (1999) Variation iteration method – a kind of non-linear analytical technique: some examples. Int J Nonlinear Mech 34:699–708
23. He JH (2000) A coupling method of homotopy technique and a perturbation technique for nonlinear problems. Int J Nonlinear Mech 35:37–43
24. He JH (2003) Homotopy perturbation method a new nonlinear analytical technique. Appl Math Comput 135:73–79
25. He JH (2006) Some asymptotic methods for strongly nonlinear equations. Int J Mod Phys B 20:1141–1199
26. He JH, Wu XH (2006) Construction of solitary solution and compacton-like solution by variational iteration method. Chaos Solitons Fractals 29:108–113
27. Helal MA, Mehanna MS (2007) A comparative study between two different methods for solving the general Korteweg–de Vries equation. Chaos Solitons Fractals 33:725–739

28. Hirota R (1971) Exact solution of the Korteweg–de Vries equation for multiple collisions of solitons. Phys Rev Lett 27:1192–1194
29. Hirota R (1973) Exact N-solutions of the wave equation of long waves in shallow water and in nonlinear lattices. J Math Phys 14:810–814
30. Hirota R (1976) Direct methods of finding exact solutions of nonlinear evolution equations. In: Miura RM (ed) Bäcklund Transformations. Lecture Notes in Mathematics, vol 515. Springer, Berlin, pp 40–86
31. Inan IE, Kaya D (2006) Some Exact Solutions to the Potential Kadomtsev–Petviashvili Equation. Phys Lett A 355:314–318
32. Inan IE, Kaya D (2006) Some Exact Solutions to the Potential Kadomtsev–Petviashvili Equation. Phys Lett A 355:314–318
33. Inan IE, Kaya D (2007) Exact Solutions of the Some Nonlinear Partial Differential Equations. Phys A 381:104–115
34. Kaya D (2003) A Numerical Solution of the Sine–Gordon Equation Using the Modified Decomposition Method. Appl Math Comp 143:309–317
35. Kaya D (2006) The Exact and Numerical Solitary-wave Solutions for Generalized Modified Boussinesq Equation. Phys Lett A 348:244–250
36. Kaya D Some Methods for the Exact and Numerical Solutions of Nonlinear Evolution Equations. Paper presented at the 2nd International Conference on Mathematics: Trends and Developments (ICMTD), Cairo, Egypt, Dec 2007, pp 27–30 (in press)
37. Kaya D, Al-Khaled K (2007) A Numerical Comparison of a Kawahara Equation. Phys Lett A 363:433–439
38. Kaya D, El-Sayed SM (2003) An Application of the Decomposition Method for the Generalized KdV and RLW Equations. Chaos Solitons Fractals 17:869–877
39. Kaya D, El-Sayed SM (2003) Numerical Soliton-like Solutions of the Potential Kadomstev–Petviashvili Equation by the Decomposition Method. Phys Lett A 320:192–199
40. Kaya D, El-Sayed SM (2003) On a Generalized Fifth Order KdV Equations. Phys Lett A 310:44–51
41. Kaya D, Inan IE (2005) A Convergence Analysis of the ADM and an Application. Appl Math Comp 161:1015–1025
42. Khater AH, El-Kalaawy OH, Helal MA (1997) Two new classes of exact solutions for the KdV equation via Bäcklund transformations. Chaos Solitons Fractals 8:1901–1909
43. Korteweg DJ, de Vries H (1895) On the change of form of long waves advancing in a rectangular canal and on a new type of long stationary waves. Philos Mag 39:422–443
44. Liao SJ (1992) The proposed homotopy analysis technique for the solution of nonlinear problems. Ph D Thesis, Shanghai Jiao Tong University
45. Liao SJ (1995) An approximate solution technique not depending on small parameters: A special example. Int J Nonlinear Mech 30:371
46. Liao SJ (2003) Beyond Perturbation: Introduction to Homotopy Analysis Method. Chapman & Hall/CRC Press, Boca Raton
47. Liao SJ (2004) On the homotopy analysis method for nonlinear problems. Appl Math Comp 147:499
48. Liao SJ (2004) On the homotopy analysis method for nonlinear problems. Appl Math Comput 147:499–513
49. Liao SJ (2005) Comparison between the homotopy analysis method and homotopy perturbation method. Appl Math Comp 169:1186–1194
50. Light Bullet Home Page, http://www.sfu.ca/~renns/lbullets.html
51. Miura RM (1968) Korteweg–de Vries equations and generalizations, I; A remarkable explicit nonlinear transformations. J Math Phys 9:1202–1204
52. Miura RM, Gardner CS, Kruskal MD (1968) Korteweg–de Vries equations and generalizations, II; Existence of conservation laws and constants of motion. J Math Phys 9:1204–1209
53. Momani S, Abuasad S (2006) Application of He's variational iteration method to Helmholtz equation. Chaos Solitons Fractals 27:1119–1123
54. Polat N, Kaya D, Tutalar HI (2006) A Analytic and Numerical Solution to a Modified Kawahara Equation and a Convergence Analysis of the Method. Appl Math Comp 179:466–472
55. Ramos JI (2008) On the variational iteration method and other iterative techniques for nonlinear differential equations. Appl Math Comput 199:39–69
56. Rayleigh L (1876) On Waves. Lond Edinb Dublin Philos Mag 5:257
57. Russell JS (1844) Report on Waves. In: Proc 14th meeting of the British Association for the Advancement of Science, BAAS, London
58. Sajid M, Hayat T (2008) Comparison of HAM and HPM methods in nonlinear heat conduction and convection equations, Nonlinear Analysis. Real World Appl 9:2296–2301
59. Sajid M, Hayat T (2008) The application of homotopy analysis method for thin film flow of a third order fluid. Chaos Solitons Fractal 38:506–515
60. Sajid M, Hayat T, Asghar S (2007) Comparison between HAM and HPM solutions of thin film flows of non-Newtonian fluids on a moving belt. Nonlinear Dyn 50:27–35
61. Shawagfeh N, Kaya D (2004) Series Solution to the Pochhammer–Chree Equation and Comparison with Exact Solutions. Comp Math Appl 47:1915–1920
62. Ugurlu Y, Kaya D Exact and Numerical Solutions for the Cahn–Hilliard Equation. Comput Math Appl
63. Wazwaz AM (1997) Necessary Conditions for the Appearance of Noise Terms in Decomposition Solution Series J Math Anal Appl 5:265–274
64. Wazwaz AM (2002) Partial Differential Equations: Methods and Applications. Balkema, Rotterdam
65. Wazwaz AM (2007) Analytic study for fifth-order KdV-type equations with arbitrary power nonlinearities. Comm Nonlinear Sci Num Sim 12:904–909
66. Wazwaz AM (2007) A variable separated ODE method for solving the triple sine-Gordon and the triple sinh-Gordon equations. Chaos Solitons Fractals 33:703–710
67. Wazwaz AM (2007) The extended tanh method for abundant solitary wave solutions of nonlinear wave equations. App Math Comp 187:1131–1142
68. Wazwaz AM (2007) The variational iteration method for solving linear and nonlinear systems of PDEs. Comput Math Appl 54:895–902
69. Wazwaz AM, Helal MA (2004) Variants of the generalized fifth-order KdV equation with compact and noncompact structures. Chaos Solitons Fractals 21:579–589
70. Whitham GB (1974) Linear and Nonlinear Waves. Wiley, New York
71. Zabusky NJ, Kruskal MD (1965) Interactions of solitons in collisionless plasma and the recurrence of initial states. Phys Rev Lett 15:240–243

Books and Reviews

The following, referenced by the end of the paper, is intended to give some useful for further reading.

For another obtaining of the KdV equation for water waves, see Kevorkian and Cole (1981); one can see the work of the Johnson (1972) for a different water-wave application with variable depth, for waves on arbitrary shears in the work of Freeman and Johnson (1970) and Johnson (1980) for a review of one and two-dimensional KdV equations. In addition to these; one can see the book of Drazin and Johnson (1989) for some numerical solutions of nonlinear evolution equations. In the work of the Zabusky, Kruskal and Deam (F1965) and Eilbeck (F1981), one can see the motion pictures of soliton interactions. See a comparison of the KdV equation with water wave experiments in Hammack and Segur (1974).

For further reading of the classical exact solutions of the nonlinear equations can be seen in the works: the Lax approach is described in Lax (1968); Calogero and Degasperis (1982, A.20), the Hirota's bilinear approach is developed in Matsuno (1984), the Bäckland transformations are described in Rogers and Shadwick (1982); Lamb (1980, Chap. 8), the Painleve properties is discussed by Ablowitz and Segur (1981, Sect. 3.8), In the book of Dodd, Eilbeck, Gibbon and Morris (1982, Chap. 10) can found review of the many numerical methods to solve nonlinear evolution equations and shown many of their solutions.

Korteweg–de Vries Equation (KdV), Some Numerical Methods for Solving the

Mustafa Inc
Department of Mathematics, Fırat University,
Elazığ, Turkey

Article Outline

Glossary
Definition of the Subject
Introduction
Some Numerical Methods for Solving the Korteweg–de Vries (KdV) Equation
The Adomian Decomposition Method (ADM)
The Homotopy Analysis Method (HAM)
The Variational Iteration Method (VIM)
The Homotopy Perturbation Method (HPM)
Numerical Applications and Comparisons
Conclusions and Discussions
Future Directions
Bibliography

Glossary

Absolute error When a real number x is approximated by another number x^*, the error is $x - x^*$. The absolute error is $|x - x^*|$.

Adomian decomposition method The method was introduced and developed by George Adomian. The method proved to be powerful, effective, and can easily handle a wide class of linear or nonlinear differential equations.

Adomian polynomials It is well known now that the ADM suggests that the unknown linear function u may be presented by the decomposition series $u = \sum_{n=0}^{\infty} u_n$, the nonlinear term $F(u)$, such as $u^2, u^3, \sin u, e^u, u_x^2$, etc. can be expressed by an infinite series of the so-called Adomian polynomials A_n.

Finite difference method The method handles the differential equation by replacing the derivatives in the equation with difference quotients.

Homotopy analysis method The HAM was introduced by Shi-Jun Liao in 1992. For details see Sect. "The Variational Iteration Method (VIM)".

Homotopy perturbation method The HPM was introduced by Ji-Huan He in 1999. For details see Sect. "Numerical Applications and Comparisons".

Korteweg–de Vries equation This is one of the simplest and most useful nonlinear model equations for solitary waves.

Lagrange multiplier The multiplier in the functional should be chosen such that its correction solution is superior to its initial approximation (trial function) and is the best within the flexibility of trial function, accordingly we can identify the multiplier by variational theory.

Initial conditions The PDEs mostly arise to govern physics phenomena such as heat distribution, wave propagation phenomena. Most of the PDEs depend on the time t. The initial values of the dependent variable u at the starting time $t = 0$ should be prescribed.

Solitons It is interesting to point out that there is no precise definition of a soliton. However, a soliton can be defined as a solution of a nonlinear partial differential equation.

Variational iteration method The VIM was first proposed by Ji-Huan He. This method has been employed to solve a large variety of linear and nonlinear problems with approximations converging rapidly to accurate solutions.

Definition of the Subject

In this work, the Adomian decomposition method (ADM), the homotopy analysis method (HAM), the variational iteration method (VIM), the homotopy perturbation method (HPM) and the explicit finite difference method (EFDM) are implemented to investigate numer-

ical solutions of the Korteweg–de Vries (KdV) equation with initial condition. The five methods are compared and it is shown that the HAM and EFDM are more efficient and effective than the ADM, the VIM and the HPM, and also converges to its exact solution more rapidly. The HAM contains the auxiliary parameter $\hbar$ which provides us with a simple way to adjust and control the convergence region of solution series.

Introduction

John Scott Russell first experimentally observed the solitary wave, a long water wave without change in shape, on the Edinburgh-Glasgow Canal in 1834. He called it the great wave of translation and, then, reported his observations at the British Association in his 1844 paper "Report on waves" [1].

After 60 years from this discovery, the two scientists D.J. Korteweg and G. de Vries formulated a mathematical model equation to provide an explanation of the phenomenon observed by Russell [2]. They derived the famous equation, the Korteweg–de Vries (KdV) equation, for the propagation of waves in one dimension on the surface of water. This is one of the simplest and most useful nonlinear model equations for solitary waves, and it represents the longtime evolution of wave phenomena in which the steepening effect of the nonlinear term is counterbalanced by dispersion. The KdV equation is a nonlinear partial differential equation of third order.

Modern developments in the theory and applications of the KdV equation began with the seminal work published as a Los Alamos Scientific Laboratory Report in 1955 by Fermi, Pasta and Ulam (FPU) on a numerical model of a discrete nonlinear mass-spring system [3]. This curious result of the FPU experiment inspired Zabusky and Kruskal [4] marked the birth of the new concept of the soliton, a name intended to signify particle like quantities. They found that stable pulse like waves could exist in a system described by the KdV equation. A remarkable quality of these solitary waves was that they could collide with each other and yet preserve their shapes and speeds after the collision. This means that a collision between KdV solitary waves is elastic. Subsequently, Zabusky [5] confirmed, numerically, the actual physical interaction of two solitons, and Lax [6] gave a rigorous analytical proof that the identities of two distinct solitons are preserved through the nonlinear interaction governed by the KdV equation. Subsequently, a paper by Gardner et al. [7] demonstrated that it was possible to write many exact solutions to the equation by using ideas from scattering theory. They discovered the first integrable nonlinear partial differential equation. Gardner et al. [8] and Hirota [9,10] constructed analytical solutions of the KdV equation that provide the description of the interaction among N solitons for any positive integer N. Experimental confirmation of solitons and their interactions has been provided successfully by Zabusky and Galvin [11], Hammack and Segur [12], and Weidman and Maxworthy [13]. During the past fifteen years a rather complete mathematical description of solitons has been developed.

The nondispersive nature of the soliton solutions to the KdV equation arises not because the effects of dispersion are absent, but because they are balanced by nonlinearities in the system [14,15,16,17]. The KdV equation [2,7,18] arises in several areas of nonlinear physics such as hydrodynamics, plasma physics, etc.

During the last five decades or so an exciting and extremely active area of research has been devoted to the construction of exact and numerical solutions for a wide class of nonlinear equations. This includes the most famous nonlinear one of Korteweg and de-Vries.

Exact solutions of the KdV equation have been used for many powerful methods such as the application of the Jacobian elliptic function [19,20,21], the powerful inverse scattering transform [22], the Painleve analysis [23], the Backlund transformations method [24], the Lie group theoretical methods [25], the direct algebraic method [26], the tanh-method [27], the sine-cosine method [28], the homogeneous balance method [29], the Riccati expansion method with constant coefficient [30] and the mapping method [31]. In addition, Sawada and Kotera [32], Rosales [33], Whitham [34], Wadati and Sawada [35,36] have all employed perturbation techniques. Recently, Helal and El-Eissa [37], Khater et al. [38,39], Helal [40] have studied analytically and numerically some physical problems that lead to the KdV equation and the soliton solutions have been obtained [17].

The KdV equation is a generic equation for the study of weakly nonlinear long waves. The KdV type of equation has been an important class of nonlinear evolution equations with numerous applications in physical sciences and engineering fields. For example, in plasma physics, these equations give rise to the ion acoustic solitons [41]; in geophysical fluid dynamics, they describe a long wave in shallow seas and deep oceans [42,43]. Their strong presence is exhibited in cluster physics, super deformed nuclei, fission, thin film, radar and rheology [44,45], optical-fiber communications [46] and superconductors [47].

Thus many analytical solutions of the KdV equation are found, and their existence and uniqueness have been studied for a certain class of initial functions [14]. The numerical solutions of the KdV equation are essen-

tial because of solutions which are not analytically available. Many methods have been proposed for numerical treatment of the KdV equation for the various boundary and initial conditions. For appropriate initial conditions, Gradner et al. [7] have shown the existence and uniqueness of solutions of the KdV equation. For the KdV equation, several quite successful numerical methods have been available such as spectral/pseudo spectral methods [48,49], finite-difference methods and Fourier spectral methods developed by many authors from both the theoretical and computational points of view [50,51]. The exponential finite-difference method is used to solve the KdV equation with the initial and boundary conditions by Bahadır [52]. Jain et al. [53] developed a numerical method for solving the KdV equation by using splitting method and quintic spline approximation technique. Soliman [54] worked numerical solutions for the KdV equation based on the collocation method using septic splines as element shape functions were set up. Frauendiener and Klein [55] presented the hyperelliptic theta-functions with spectral methods for numerical solutions of the KP and KdV equations. Bhatta and Bhatti [56] studied numerical solution of the KdV equation by using modified Bernstein polynomials. Helal and Mehanna [57] presented a comparative study between the Adomian decomposition method and the finite-difference method for solving the general KdV equation. Kutluay et al. [58] used the heat balance integral method to the KdV equation prescribed by appropriate homogenous boundary conditions and a set of initial conditions to obtain its approximate analytical solutions at small times. An analytical-numerical method was applied to the KdV equation with a variant of boundary and initial conditions to obtain its numerical solutions at small times by Özer and Kutluay [59]. Recently, Dehghan and Shokri [60] proposed a numerical scheme to solve the third-order nonlinear KdV equation using collocation points and approximating the solution using multiquadric radial basis function. Dag and Dereli [61] presented numerical solution of the KdV equation by using the meshless method based on the collocation with radial basis functions.

Some Numerical Methods for Solving the Korteweg–de Vries (KdV) Equation

The Korteweg–de Vries (KdV) equation has the following form:

$$u_t - 6uu_x + u_{xxx} = 0\,, \tag{1}$$

subject to initial condition

$$u(x,0) = f(x)\,, \tag{2}$$

where $u(x,t)$ is a differentiable function and $f(x)$ is bounded. We shall assume that the solution $u(x,t)$, along with its derivatives, tends to zero as $|x| \to \infty$. The second and third terms uu_x and u_{xxx} represent the nonlinear convection and dispersion effects, respectively. The solitons of the KdV equation arise as a balance between nonlinear convection and dispersion terms. The nonlinear effect causes the steepening of the waveform [61], while the dispersion effect makes the waveform spread. Due to the competition of these two effects, a stationary waveform (solitary wave) exists. The solution is obtained among the nonlinear and dispersion by stability. The soliton solutions of nonlinear wave equations shown by Wadati [62,63] have the following properties:

- Some certain waves propagate which does not change its special behavior.
- Reginal waves do not lose their properties and also they are stable towards the collisions.

The basic purpose of this work is to approach the KdV Eq. (1) with initial condition (2) differently, by using four semi inverse methods and the EFDM. In this work, we will use the Adomian decomposition method (ADM), the homotopy analysis method (HAM), the variational iteration method (VIM), the homotopy perturbation method (HPM) and the explicit finite difference method (EFDM). The results are compared with those obtained using the five methods. More details for the five methods can be found in next sections.

The Adomian Decomposition Method (ADM)

The ADM was first introduced by Adomian in the beginning of [64,65,66]. The method is useful for obtaining both a closed form and the explicit solution and numerical approximations of linear or nonlinear differential equations and it is also quite straight forward for writing computer codes. This method has been applied to obtain a formal solution to a wide class of stochastic and deterministic problems in science and engineering involving algebraic, differential, integro-differential, differential delay, integral and partial differential equations [67,68,69,70,71,72,73,74,75,76]. The convergence of ADM for partial differential equations was presented by Cherruault [77]. Application and convergence of this method for nonlinear partial differential equations are found in [78,79,80,81].

In general, it is necessary to construct the solution of the problems in the form of a decomposition series solution. In the simplest case, the solution can be developed as a Taylor series expansion about the function, not the point

at which the initial condition and integration of the right-hand side function of the problem are determined (the first term u_0 of the decomposition series for $n \geq 0$). The sum of the $u_0, u_1, u_2, \ldots$ terms are simply the decomposition series

$$u(x,t) = \sum_{n=0}^{\infty} u_n(x,t) \ . \tag{3}$$

Suppose that the differential equation operator, including both linear and nonlinear terms, can be formed as

$$Lu + Ru + Nu = F(x,t) \ , \tag{4}$$

with initial condition

$$u(x,0) = g(x) \ , \tag{5}$$

where L is the higher-order derivative which is assumed to be invertible, R is a linear differential operator of order less than L, N is the nonlinear term and $F(x,t)$ is a source term. We next apply the inverse operator L^{-1} to both sides of Eq. (4) and using the given condition (5) to obtain

$$u(x,t) = g(x) + f(x,t) - L^{-1}(Ru) - L^{-1}(Nu) \ , \tag{6}$$

where the function $f(x,t)$ represents the terms arising from integrating the source term $F(x,t)$ and from using the given conditions, all are assumed to be prescribed. The nonlinear term can be written as

$$Nu = \sum_{n=0}^{\infty} A_n \ , \tag{7}$$

where A_n is the Adomian polynomials. These polynomials are defined as

$$A_n = \frac{1}{n!}\frac{\mathrm{d}^n}{\mathrm{d}\lambda^n}\left[N\left(\sum_{n=0}^{\infty}\lambda^k u_k(x,t)\right)\right]_{k=0} \ , \quad n = 0, 1, 2, \ldots \tag{8}$$

for example

$$\begin{aligned} A_0 &= N(u_0) \ , \\ A_1 &= u_1 N'(u_0) \ , \\ A_2 &= u_2 N'(u_0) + \frac{1}{2}u_1 N''(u_0) \ , \\ A_3 &= u_3 N'(u_0) + u_1 u_2 N''(u_0) + \frac{1}{6}u_1^3 N'''(u_0) \ , \end{aligned} \tag{9}$$

and so on, the other polynomials can be constructed in a similar way [66]. As indicated before, Adomian method defines the solution u by an infinite series of components given by Eq. (4) and the components $u_0, u_1, u_2, \ldots$ are usually recurrently determined. Thus, the formal recursive relation is defined by

$$\begin{cases} u_0(x,t) & = g(x) + f(x,t) \ , \\ u_{n+1}(x,t) & = -L^{-1}(Ru_n) - L^{-1}(Nu_n) \ , \quad n \geq 0 \ , \end{cases} \tag{10}$$

which are obtained for all components of u. As a result, the terms of the series $u_0, u_1, u_2, \ldots$ are identified and the exact solution may be entirely determined by using the approximation

$$u(x,t) = \lim_{n\to\infty} \varphi_n(x,t) \ , \tag{11}$$

where

$$\varphi_n(x,t) = \sum_{k=0}^{n-1} u_k(x,t) \ , \tag{12}$$

or

$$\begin{cases} \varphi_0 = u_0 \ , \\ \varphi_1 = u_0 + u_1 \ , \\ \varphi_2 = u_0 + u_1 + u_2 \ , \\ \vdots \\ \varphi_n = u_0 + u_1 + u_2 + \ldots + u_{n-1}, \quad n \geq 0 \ . \end{cases} \tag{13}$$

The Homotopy Analysis Method (HAM)

The HAM was developed in 1992 by Liao in [82,83,84,85]. This method has been successfully applied by many authors [86,87,88,89,90,91]. The HAM contains the auxiliary parameter $\hbar$ which provides us with a simple way to adjust and control the convergence region of solution series for large or small values of x and t.

We consider the following differential equation

$$N[u(x,t)] = 0 \ , \tag{14}$$

where N is a nonlinear differential operator, x and t denote independent variables, $u(x,t)$ is an unknown function. By means of the HAM, one first constructs a zero-order deformation equation

$$(1-p)\,\pounds\left[\phi(x,t;p) - u_0(x,t)\right] = p\,\hbar\,N\left[\phi(x,t;p)\right] \ , \tag{15}$$

where $p \in [0,1]$ is the embedding parameter, $\hbar \neq 0$ is a non-zero auxiliary parameter, is an auxiliary linear operator, $u_0(x,t)$ is an initial guess of $u(x,t)$ and $\phi(x,t;p)$

is a unknown function. It is important that one has great freedom to choose auxiliary things in the HAM. Obviously, when $p = 0$ and $p = 1$, it holds

$$\phi(x,t;0) = u_0(x,t)\ , \quad \phi(x,t;1) = u(x,t)\ , \tag{16}$$

respectively. The solution $\phi(x,t;p)$ varies from the initial guess $u_0(x,t)$ to the solution $u(x,t)$. Expanding $\phi(x,t;p)$ in Taylor series about the embedding parameter

$$\phi(x,t;p) = u_0(x,t) + \sum_{m=1}^{+\infty} u_m(x,t)\ p^m\ , \tag{17}$$

where

$$u_m(x,t) = \frac{1}{m!}\frac{\partial^m \phi(x,t;p)}{\partial p^m}\bigg|_{p=0}\ . \tag{18}$$

The convergence of the series (17) depends upon the auxiliary parameter $\hbar$. If it is convergent at $p = 1$, one has

$$u(x,t) = u_0(x,t) + \sum_{m=1}^{+\infty} u_m(x,t)\ . \tag{19}$$

According to the definition (19), the governing equation can be deduced from the zero-order deformation Eq. (15). Define the vector

$$\vec{u}_n = \{u_0(x,t)\ , u_1(x,t)\ , \ldots , u_n(x,t)\}\ .$$

Differentiating Eq. (15) m-times with respect to the embedding parameter p and then setting $p = 0$ and finally dividing them by $m!$, we have the so-called mth-order deformation equation

$$\pounds\left[u_m(x,t) - \chi_m\ u_{m-1}(x,t)\right] = \hbar\ \Re_m\left(\vec{u}_{m-1}\right)\ , \tag{20}$$

where

$$\Re_m\left(\vec{u}_{m-1}\right) = \frac{1}{(m-1)!}\frac{\partial^{m-1} N\left[\phi(x,t;p)\right]}{\partial p^{m-1}}\bigg|_{p=0}\ , \tag{21}$$

and

$$\chi_m = \begin{cases} 0, & m \le 1\ , \\ 1, & m > 1\ . \end{cases} \tag{22}$$

It should be emphasized that $u_m(x,t)$ for $m \ge 1$ is governed by the nonlinear Eq. (20) with the linear boundary conditions that come from the original problem, which can be easily solved by symbolic computation software such as Maple and Mathematica.

The Variational Iteration Method (VIM)

The VIM was first proposed by He [92,93,94,95,96]. This method has been employed to solve a large variety of linear and nonlinear problems with approximations converging rapidly to accurate solutions. The idea of VIM is to construct a correction functional by a general Lagrange multiplier. The multiplier in the functional should be chosen such that its correction solution is superior to its initial approximation (trial function) and is the best within the flexibility of the trial function, accordingly we can identify the multiplier by variational theory. The initial approximation can be freely chosen with possible unknowns, which can be determined by imposing the boundary/initial conditions.

We consider the following general differential equation:

$$Lu + Nu = g\ , \tag{23}$$

where L and N are linear and nonlinear operators respectively, and $g(t)$ is the source inhomogeneous term.

According to the VIM, the terms of a sequence $\{u_n\}$ are constructed such that this sequence converges to the exact solution. u_ns are calculated by a correction functional as follows:

$$u_{n+1}(x,t) = u_n(x,t) + \int_0^t \lambda\left\{Lu_n(\xi) + N\widetilde{u_n}(\xi) - g(\xi)\right\} d\xi\ , \quad n \ge 0\ , \tag{24}$$

where λ is the general Lagrange multiplier, which can be identified optimally by the variational theory, the subscript n denotes the nth approximation and $\widetilde{u_n}$ considered as a restricted variation, i. e. $\delta\widetilde{u_n} = 0$. For linear problems, the exact solution can be obtained by only one iteration step due to the fact that the Lagrange multiplier can be exactly identified. In nonlinear problems, in order to determine the Lagrange multiplier in a simple manner, the nonlinear terms have to be considered as restricted variations. Consequently, the exact solution may be obtained by using

$$u(x,t) = \lim_{n\to\infty} u_n(x,t)\ . \tag{25}$$

This method has been employed to solve a large variety of linear and nonlinear problems with approximations converging to accurate solutions. This technique was used to find the solutions of partial differential equations, ordinary differential equations, boundary-value problems and integral equations by many authors [97,98,99,100,101,102,103,104,105].

The Homotopy Perturbation Method (HPM)

The HPM was developed by He [106,107,108,109]. In recent years, the applications of HPM in nonlinear problems has been devoted to scientists and engineers [110,111,112,113].

We consider the following general nonlinear differential equation:

$$A\left(y\right) - f\left(r\right) = 0\,, \quad r \in \Omega\,, \tag{26}$$

with boundary conditions

$$B\left(y, \partial y/\partial n\right) = 0\,, \quad r \in \Gamma\,, \tag{27}$$

where A is a general differential operator, B is a boundary operator, $f\left(r\right)$ is a known analytic function and Γ is the boundary of the domain Ω. The operator A can be generally divided into two parts L and N, where L is linear and N is nonlinear. Therefore, Eq. (26) can be written as follows:

$$L\left(y\right) + N\left(y\right) - f\left(r\right) = 0\,. \tag{28}$$

We construct a homotopy of Eq. (26) $y\left(r,p\right) : \Omega \times [0,1] \to R$ which satisfies

$$\begin{aligned} H\left(y,p\right) &= \left(1-p\right)\left[L\left(y\right) - L\left(y_0\right)\right] + p\left[A\left(y\right) - f\left(r\right)\right] \\ &= 0\,, \quad r \in R\,, \end{aligned} \tag{29}$$

which is equivalent to

$$\begin{aligned} H\left(y,p\right) &= L\left(y\right) - L\left(y_0\right) + pL\left(y_0\right) + p\left[A\left(y\right) - f\left(r\right)\right] \\ &= 0\,, \end{aligned} \tag{30}$$

where $p \in [0,1]$ is an embedding parameter and y_0 is an initial approximation which satisfies the boundary conditions. It follows from Eqs. (29) and (30) that

$$\begin{aligned} &H\left(y,0\right) = L\left(y\right) - L\left(y_0\right) = 0 \quad \text{and} \\ &H\left(y,1\right) = A\left(y\right) - f\left(r\right) = 0\,. \end{aligned} \tag{31}$$

Thus, the changing process of p from 0 to 1 is just that of $y\left(r,p\right)$ from $y_0\left(r\right)$ to $y\left(r\right)$. In topology this is called deformation and $L\left(y\right) - L\left(y_0\right)$ and $A\left(y\right) - f\left(r\right)$ are called homotopic. Here the embedding parameter is introduced much more naturally, unaffected by artificial factors; further it can be considered as a small parameter for $0 \le p \le 1$. So it is very natural to assume that the solution of (29) and (30) can be expressed as

$$y\left(x\right) = u_0\left(x\right) + pu_1\left(x\right) + p^2u_2\left(x\right) + \cdots\,. \tag{32}$$

According to HPM, the approximate solution of (26) can be expressed as a series of the power of p, i. e.,

$$y = \lim_{p\to 1} y\left(x\right) = u_0\left(x\right) + u_1\left(x\right) + u_2\left(x\right) + \cdots\,. \tag{33}$$

The convergence of series (33) has been proved by He in [106].

Numerical Applications and Comparisons

We now consider the initial value problem associated with the KdV equation:

$$\begin{cases} u_t - 6uu_x + u_{xxx} = 0\,, \\ u\left(x,0\right) = -\frac{c}{2}\sec h^2\left(\frac{\sqrt{c}}{2}x\right)\,, \end{cases} \tag{34}$$

with $u\left(x,t\right)$ is a sufficiently smooth function and c is the velocity of the wavefront.

In this section, we apply the above described four methods and the explicit finite difference method on the KdV Eq. (34) so that the comparisons are made numerically.

The ADM for Eq. (14)

According to the ADM [64,65,66], Eq. (34) can be written in an operator form

$$\begin{cases} Lu = 6uu_x - u_{xxx}\,, \\ u\left(x,0\right) = f\left(x\right) = -\frac{c}{2}\sec h^2\left(\frac{\sqrt{c}}{2}x\right)\,, \end{cases} \tag{35}$$

where the differential operator L is

$$L \equiv \frac{\partial}{\partial t}\,, \tag{36}$$

and the inverse operator L^{-1} is an integral operator given by

$$L^{-1}\left(\circ\right) = \int_0^t \left(\circ\right)\mathrm{d}t\,. \tag{37}$$

The ADM [64,65,66] assumes a series solution for the unknown function $u\left(x,t\right)$ given by Eq. (3) and the nonlinear term $Nu = uu_x$ can be decomposed into a infinite series of polynomials given by Eq. (5). Operating with the integral operator (37) on both sides of (35) and using the initial condition, we get

$$u\left(x,t\right) = f\left(x\right) + L^{-1}\left(6uu_x - u_{xxx}\right)\,. \tag{38}$$

Substituting (3) and (5) into the functional Eq. (38) yields

$$\sum_{n=0}^{\infty} u_n(x,t) = f(x) + L^{-1}\left[6\left(\sum_{n=0}^{\infty} A_n\right) - \left(\sum_{n=0}^{\infty} u_n\right)_{xxx}\right]. \tag{39}$$

The ADM admits that the zeroth component $u_0(x,t)$ be identified by all terms that arise from the initial conditions and from the source terms if exist, and as a result, the remaining components $u_n(x,t)$, $n \geq 1$ can be determined by using the recurrence relation:

$$\begin{cases} u_0(x,t) = f(x) = -\frac{c}{2}\sec h^2\left(\frac{\sqrt{c}}{2}x\right), \\ u_{k+1}(x,t) = L^{-1}\left(6A_k - (u_k)_{xxx}\right), \quad k \geq 0, \end{cases} \tag{40}$$

where A_k, $k \geq 0$ are Adomian polynomials that represent the nonlinear term (uu_x) and given by

$$\begin{cases} A_0 = u_0 u_{0x}, \\ A_1 = u_{0x}u_1 + u_0 u_{1x}, \\ A_2 = u_1 u_{1x} + u_{0x}u_2 + u_0 u_{2x}, \\ A_3 = u_1 u_{2x} + u_{1x}u_2 + u_{0x}u_3 + u_0 u_{3x}. \end{cases} \tag{41}$$

Thus, some of the symbolically computed components are found as:

$$\begin{aligned} u_0(x,t) &= -\frac{c}{2}\sec h^2\left(\frac{\sqrt{c}}{2}x\right), \\ u_1(x,t) &= -\frac{c^{5/2}}{2}\sec h^2\left(\frac{\sqrt{c}}{2}x\right)\tanh\left(\frac{\sqrt{c}}{2}x\right)t, \\ u_2(x,t) &= -\frac{c^4}{8}\left(\cosh\left[\sqrt{c}x\right]-2\right)\sec h^4\left(\frac{\sqrt{c}}{2}x\right)t^2, \\ u_3(x,t) &= -\frac{c^{11/2}}{48}\sec h^5\left(\frac{\sqrt{c}}{2}x\right) \\ &\quad \cdot\left(-11\sinh\left(\frac{\sqrt{c}}{2}x\right) + \sinh\left(\frac{3\sqrt{c}}{2}x\right)\right)t^3, \\ u_4(x,t) &= -\frac{c^7}{384}\sec h^6\left(\frac{\sqrt{c}}{2}x\right) \\ &\quad \cdot\left[33 - 26\cosh\left[\sqrt{c}x\right] + \cosh\left[2\sqrt{c}x\right]\right]t^4, \end{aligned}$$

and so on. In this manner the other components of the decomposition series can be easily obtained using any symbolic program. The solution in a series for is given by

$$\begin{aligned} u(x,t) &= \sum_{n=0}^{\infty} u_n(x,t) \\ &= u_0(x,t) + u_1(x,t) + u_2(x,t) + u_3(x,t) + \cdots, \\ &= -\frac{c}{2}\sec h^2\left(\frac{\sqrt{c}}{2}x\right) \\ &\quad -\frac{c^{5/2}}{2}\sec h^2\left(\frac{\sqrt{c}}{2}x\right)\tanh\left(\frac{\sqrt{c}}{2}x\right)t \\ &\quad -\frac{c^4}{8}\left(\cosh\left[\sqrt{c}x\right]-2\right)\sec h^4\left(\frac{\sqrt{c}}{2}x\right)t^2 \\ &\quad -\frac{c^{11/2}}{48}\sec h^5\left(\frac{\sqrt{c}}{2}x\right) \\ &\quad \cdot\left(-11\sinh\left(\frac{\sqrt{c}}{2}x\right) + \sinh\left(\frac{3\sqrt{c}}{2}x\right)\right)t^3 \\ &\quad + \cdots, \end{aligned} \tag{42}$$

so that the exact solution

$$u(x,t) = -\frac{c}{2}\sec h^2\left(\frac{\sqrt{c}}{2}(x-ct)\right), \tag{43}$$

is readily obtained.

The convergence analysis of the ADM applied to the KdV Eq. (35) without the initial condition has been conducted by Helal and Mehanna [55].

The HAM for Eq. (34)

We will apply the HAM to the Eq. (34) to illustrate the strength of the method and to establish exact solution for this Eq. (34). We choose the linear operator

$$\pounds\left[\phi(x,t;p)\right] = \frac{\partial\phi(x,t;p)}{\partial t}, \tag{44}$$

with property

$$\pounds\left[k\right] = 0, \tag{45}$$

where k is a constant. We now define a nonlinear operator

$$N\left[\phi(x,t;p)\right] = \frac{\partial\phi(x,t;p)}{\partial t} - 6\phi(x,t;p)\frac{\partial\phi(x,t;p)}{\partial x} + \frac{\partial^3\phi(x,t;p)}{\partial x^3}. \tag{46}$$

We construct the zeroth-order deformation equation

$$(1-p)\,\pounds\left[\phi\left(x,t;p\right)-u_0\left(x,t\right)\right]=p\,\hbar\,N\left[\phi\left(x,t;p\right)\right]. \tag{47}$$

For $p=0$ and $p=1$, we can write

$$\begin{cases}\phi\left(x,t;0\right)=u\left(x,0\right)=u_0\left(x,t\right),\\ \phi\left(x,t;1\right)=u\left(x,t\right).\end{cases} \tag{48}$$

So we obtain mth-order deformation equation

$$\pounds\left[u_m\left(x,t\right)-\chi_m\,u_{m-1}\left(x,t\right)\right]=\hbar\,\Re_m\left(\vec{u}_{m-1}\right), \tag{49}$$

where

$$\Re_m\left(\vec{u}_{m-1}\right)=\frac{\partial\phi_{m-1}\left(x,t;p\right)}{\partial t}-6\sum_{n=0}^{m-1}\phi_n\left(x,t;p\right)\frac{\partial\phi_{m-1-n}\left(x,t;p\right)}{\partial x}+\frac{\partial^3\phi_{m-1}\left(x,t;p\right)}{\partial x^3}. \tag{50}$$

Now the solution of the mth-order deformation Eq. (49) for $m\geq 1$ become

$$u_m\left(x,t\right)=\chi_m\,u_{m-1}\left(x,t\right)+\hbar\,\pounds^{-1}\left[\Re_m\left(\vec{u}_{m-1}\right)\right]. \tag{51}$$

Thus we get the following terms by using (51) for Eq. (34):

$$u_0\left(x,t\right)=-\frac{c}{2}\sec h^2\left(\frac{\sqrt{c}}{2}x\right),$$

$$u_1\left(x,t\right)=\frac{c^{5/2}}{2}\hbar\sec h^2\left(\frac{\sqrt{c}}{2}x\right)\tanh\left(\frac{\sqrt{c}}{2}x\right)t,$$

$$u_2\left(x,t\right)=-\frac{c^{5/2}}{2}\hbar t\sec h^4\left(\frac{\sqrt{c}}{2}x\right)\left[c^{3/2}\hbar t\cosh\left(\sqrt{c}x\right)-2\left(c^{3/2}\hbar t+(1+\hbar)\sinh\left(\sqrt{c}x\right)\right)\right],$$

$$\begin{aligned}u_3(x,t)=&\frac{c^{5/2}}{48}\hbar t\sec h^2\left(\frac{\sqrt{c}}{2}x\right)\Bigg[c^3\hbar^2t^2\sec h^3\left(\frac{\sqrt{c}}{2}x\right)\\&\cdot\sinh\left(\frac{3\sqrt{c}}{2}x\right)+24\hbar(1+\hbar)\tanh\left(\frac{\sqrt{c}}{2}x\right)\\&+\sec h^2\left(\frac{\sqrt{c}}{2}x\right)\Bigg(-12c^{3/2}\hbar(1+\hbar)t\cosh(\sqrt{c}x)\\&+12(1+\hbar)\sinh(\sqrt{c}x)\\&+c^{3/2}\hbar t\left(24+24\hbar-11c^{3/2}\hbar t\tanh\left(\frac{\sqrt{c}}{2}x\right)\right)\Bigg)\Bigg],\end{aligned}$$

and so on. In this manner the other components of the decomposition series can be easily obtained using any symbolic program.

Liao [82,83,84,85] showed that whatever a solution series converges to it will be one of the solutions of considered problem. Liao [82,83,84,85] showed the approximate solutions obtained by the HAM to be controlled by the auxiliary parameter and the rate of convergence of $\hbar$.

It is noted that the approximate solutions converge at $-2.1\leq\hbar\leq-1.1$ for 5th-order approximation and at $-4\leq\hbar\leq 2$ for 10th-order approximation for the KdV Eq. (34) when $x=10$, $t=0.5$, $c=1$ and $c=5$ (see Figs. 2 and 3).

The VIM for Eq. (34)

To solve the KdV Eq. (34) by means of the VIM, we construct a correction functional which reads as

$$u_{n+1}\left(x,t\right)=u_n\left(x,t\right)+\int_0^t\lambda\left(\frac{\partial u_n\left(x,\tau\right)}{\partial\tau}-N\widetilde{u_n}\left(x,\tau\right)+\frac{\partial^3\widetilde{u_n}\left(x,\tau\right)}{\partial x^3}\right)\mathrm{d}\tau, \tag{52}$$

where λ is the general Lagrange multiplier [109] whose optimal value is found using variational theory, $u_0\left(x,t\right)$ is an initial approximation which must be chosen suitable and $N\widetilde{u_n}$ is the restricted variation, i. e., $\delta\widetilde{u_n}=0$ [107]. To find the optimal values of λ we have

$$\delta u_{n+1}\left(x,t\right)=\delta u_n\left(x,t\right)+\delta\int_0^t\lambda\left(\frac{\partial u_n\left(x,\tau\right)}{\partial\tau}\right)\mathrm{d}\tau, \tag{53}$$

that results

$$\delta u_{n+1}\left(x,t\right)=\delta u_n\left(x,t\right)(1+\lambda)-\int_0^t\delta u_n\left(x,t\right)\lambda'\,\mathrm{d}\tau, \tag{54}$$

which yields

$$\lambda'\left(\tau\right)=0\,|_{\tau=t},\,1+\lambda\left(\tau\right)=0\,|_{\tau=t}. \tag{55}$$

Thus we have

$$\lambda\left(t\right)=-1, \tag{56}$$

and we obtain the following iteration formula

$$u_{n+1}(x,t) = u_n(x,t) - \int_0^t \left(\frac{\partial u_n(x,\tau)}{\partial \tau} - 6u_n(x,\tau)\frac{\partial u_n(x,\tau)}{\partial x} + \frac{\partial^3 u_n(x,\tau)}{\partial x^3} \right) \mathrm{d}\tau \,. \quad (57)$$

Now using (57) we can find the solution of the KdV Eq. (34) as a convergent sequence. We get the following components:

$$u_0(x,t) = -\frac{c}{2}\sec h^2\left(\frac{\sqrt{c}}{2}x\right),$$

$$u_1(x,t) = -\frac{c}{2}\sec h^3\left(\frac{\sqrt{c}}{2}x\right)\left[\cosh\left(\frac{\sqrt{c}}{2}x\right) + c^{3/2}t\sinh\left(\frac{\sqrt{c}}{2}x\right)\right],$$

$$\begin{aligned} u_2(x,t) = -\frac{c}{32}\sec h^7\left(\frac{\sqrt{c}}{2}x\right)\Big[&-2\left(4c^3t^2-5\right) \\ &\cdot\cosh\left(\frac{\sqrt{c}}{2}x\right) + \left(5-c^3t^2\right)\cosh\left(\frac{3\sqrt{c}}{2}x\right) \\ &+\cosh\left(\frac{5\sqrt{c}}{2}x\right) + c^3t^2\cosh\left(\frac{5\sqrt{c}}{2}x\right) \\ &+4c^{3/2}t\sinh\left(\frac{\sqrt{c}}{2}x\right) - 60c^{9/2}t^3\sinh\left(\frac{\sqrt{c}}{2}x\right) \\ &+6c^{3/2}t\sinh\left(\frac{3\sqrt{c}}{2}x\right) + 12c^{9/2}t^3\sinh\left(\frac{3\sqrt{c}}{2}x\right) \\ &+2c^{3/2}t\sinh\left(\frac{5\sqrt{c}}{2}x\right)\Big], \end{aligned}$$

and so on. In the same manner the rest of components of the iteration formulae (57) can be obtained using any symbolic packages.

So we obtain

$$u(x,t) = \lim_{n\to\infty} u_n(x,t) = -\frac{c}{2}\sec h^2\left(\frac{\sqrt{c}}{2}(x-ct)\right). \quad (58)$$

The HPM for Eq. (34)

To investigate the approximate solution of Eq. (34), we first construct a homotopy as follows:

$$(1-p)\left(\dot{Y}-\dot{u}_0\right) + p\left(\dot{Y} - 6YY' + Y'''\right) = 0\,, \quad (59)$$

where "primes" denote differentiation with respect to x, and "dot" denotes differentiation with respect to t. In order to obtain the unknowns of $Y_i(x,t)$, $i = 1,2,3,4$, we must construct and solve the following system which includes four equations with three unknowns, considering the initial condition of $Y(x,0) = u(x,0)$ and having the initial approximation of Eq. (34):

$$\begin{aligned} p^0:&\quad \dot{Y}_0 - \dot{u}_0 = 0\,,\\ p^1:&\quad \dot{Y}_1 + \dot{u}_0 - 6Y_0Y_0' + Y_0''' = 0\,,\\ p^2:&\quad \dot{Y}_2 - 6Y_0Y_1' - 6Y_1Y_0^0 + Y_1''' = 0\,,\\ p^3:&\quad \dot{Y}_3 - 6Y_0Y_2' - 6Y_1Y_1^{2\prime} - 6Y_2Y_0' + Y_2''' = 0\,. \end{aligned} \quad (60)$$

Thus, we will obtain:

$$u(x,t) = \lim_{p\to 1} Y_k(x,t)\,, \quad k = 0,1,2,3,\cdots. \quad (61)$$

To calculate the terms of the homotopy series (61) for $u(x,t)$, we substitute the initial condition into the systems (60) and finally using Mathematica, the solutions of the equation can be obtained as follows:

$$\begin{aligned} u_0(x,t) &= -\frac{c}{2}\sec h^2\left(\frac{\sqrt{c}}{2}x\right),\\ u_1(x,t) &= -\frac{c^{5/2}}{2}\sec h^2\left(\frac{\sqrt{c}}{2}x\right)\tanh\left(\frac{\sqrt{c}}{2}x\right)t\,, \end{aligned} \quad (62)$$

$$\begin{aligned} &u_2(x,t)\\ &= \frac{c^{5/2}}{2}t^2\sec h^4\left(\frac{\sqrt{c}}{2}x\right)\left[-c^{3/2}\cosh\left(\sqrt{c}x\right)+2c^{3/2}\right], \end{aligned}$$

and so on. In this manner the other components can be easily obtained. Substituting Eq. (62) into Eq. (61):

$$\begin{aligned} u(x,t) &= Y_0(x,t) + Y_1(x,t) + Y_2(x,t) + \cdots\\ &= -\frac{c}{2}\sec h^2\left(\frac{\sqrt{c}}{2}x\right)\\ &\quad -\frac{c^{5/2}}{2}\sec h^2\left(\frac{\sqrt{c}}{2}x\right)\tanh\left(\frac{\sqrt{c}}{2}x\right)t\\ &\quad -\frac{c^4}{2}t^2\sec h^4\left(\frac{\sqrt{c}}{2}x\right)\left[\cosh\left(\sqrt{c}x\right)-2\right]+\cdots. \end{aligned} \quad (63)$$

The EFDM for Eq. (34)

The finite difference methods are the most frequently used and universally applicable. These methods are approximate in the sense that the derivatives at a point are approximated by difference quotients over a small interval [114]. In order to obtain a finite difference replacement of the KdV Eq. (34) the region to be examined is divided into equal rectangular meshes with sides Δx and Δt parallel to

the x- and t-axes respectively. The function $u(x,t)$ is approximated by $u_i^n = u(i\Delta x, n\Delta t)$ where i and n are integers and $i = n = 0$ is the origin. Let us define

$$U_t\,|_{(i,n)} = \frac{U_i^{n+1} - U_i^{n-1}}{2\Delta t}\,, \tag{64}$$

$$U_x\,|_{(i,n)} = \frac{U_{i+1}^n - U_{i-1}^n}{2\Delta x}\,, \tag{65}$$

$$U_{xxx}\,|_{(i,n)} = \frac{1}{2(\Delta x)^3}\left[U_{i+2}^n - 2U_{i+1}^n + 2U_{i-1}^n - U_{i-2}^n\right]. \tag{66}$$

The explicit scheme computes the value of the numerical solution at the forward time step in terms of known values at the previous time step. An explicit scheme for solving the KdV equation produced originally by Zabusky and Galvin [11] is centered in time and space. Substituting (64)–(66) into the KdV Eq. (34) with

$$U\,|_{(i,n)} = \frac{1}{3}\left(U_{i-1}^n + U_i^n + U_{i+1}^n\right)\,, \tag{67}$$

leads to

$$\begin{aligned} U_i^{n+1} &= U_i^{n-1} + 2\Delta t\left(U_{i-1}^n + U_i^n + U_{i+1}^n\right)\left(U_{i+1}^n - U_{i-1}^n\right) \\ &\quad - \frac{\Delta t}{(\Delta x)^3}\left(U_{i+2}^n - 2U_{i+1}^n + 2U_{i-1}^n - U_{i-2}^n\right)\,. \end{aligned} \tag{68}$$

For the initial step, we use a scheme which is forward in time and centered in space

$$\begin{aligned} U_i^1 &= U_i^0 + \Delta t\left(U_{i-1}^0 + U_i^0 + U_{i+1}^0\right)\left(U_{i+1}^0 - U_{i-1}^0\right) \\ &\quad - \frac{\Delta t}{(\Delta x)^3}\left(U_{i+2}^0 - 2U_{i+1}^0 + 2U_{i-1}^0 - U_{i-2}^0\right)\,. \end{aligned} \tag{69}$$

It is clear that Eq. (68) is a three-level scheme of time, i. e. in order to obtain U_i at the time level $n+1$, we need the following values of U_{i-2}, U_{i-1}, U_{i+1} and U_{i+2} at the previous time level n in addition to the value of U_i at the time level $n-1$. The explicit difference scheme (68) has second order accuracy in Δt and Δx as the truncation error is $O(\Delta t)^2 + O(\Delta x)^2$ and is also consistent with KdV equation. A stability analysis of the nonlinear numerical scheme (68) using Fourier mode method is not easy to handle unless it is assumed that U, in the nonlinear term, is locally constant. This is equivalent to replacing the term (67) in Eq. (68) by U^*. This linearized scheme for the KdV Eq. (34) has stability condition [115]:

$$\frac{\Delta t}{\Delta x}\left[-6\left|U^*\right| + \frac{4}{(\Delta x)^2}\right] \le 1\,. \tag{70}$$

In Table 7, h is x-direction mesh size and k is t-direction time step.

Conclusions and Discussions

In this paper, the ADM, HAM, VIM, HPM and EFDM were used for the KdV Eq. (34) with initial conditions. We made comparisons of the numerical results obtained by using ADM, HAM, VIM, HPM and EFDM. The approximate solutions to the equation has been also calculated by using the ADM, HAM, VIM, HPM and EFDM without any need to use transformation techniques and linearization of the equation. The ADM, HAM and HPM avoids the difficulties and massive computational work by determining analytic solutions of the nonlinear equations. The main advantage of HAM, VIM and HPM are to overcome the difficulty arising in calculating Adomian's polynomials in the ADM. But the main disadvantage of VIM and EFDM, in general, very big terms and the consuming time to compute it is big, so we need a large computer memory and time. In addition, it is difficult and takes a lot of time to calculate the terms after the fifth term in the VIM for the KdV Eq. (14). It is convenient to examine Tables 1, 2, 3, 4, 5, 6, 7 and Figs. 1, 3, 4, 5 for comparing exact solution and numerical solutions which one gets by using these five methods. Tables and figures are calculated for these methods by using five terms and $c = 1$, except Tables 2 and 4. The results which are obtained by the EFDM are better than the results which are obtained by the other four (semi-inverse) methods for five terms (see Tables 1, 3, 5, 6, 7. If you increase the numbers of terms in the semi-inverse methods, we can see that the results are better (see Tables 2 and 4. Obtaining serial solutions with the semi-inverse methods are easier and more efficient than pure numerical methods (the finite difference methods, the finite element methods, B-spline methods, etc.). Finding numerical results by the EFDM are difficult and take time. The present methods give nearly the same results for $c = 1$ and large values of x but the HAM is the more sensitive method than the others. Because it is possible to obtain a rapidly convergent solution for large and small values of x and t by choosing convenient helper the auxiliary parameter $\hbar$. For this it is enough to consider Figs. 1, 3, 4, 5. Finally, we point out that, for given equations with initial value, the corresponding analytical and numerical solutions are obtained according to the recurrence equations by using Mathematica.

Future Directions

The Korteweg–de Vries equation, which is the most important equation of mathematical physics, has solitons and

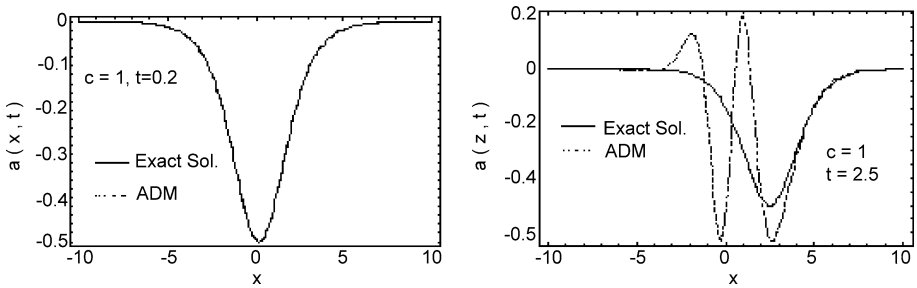

Korteweg–de Vries Equation (KdV), Some Numerical Methods for Solving the, Figure 1
The comparison of the ADM (φ_5) and the exact solutions for Eq. (34) at $c = 1$

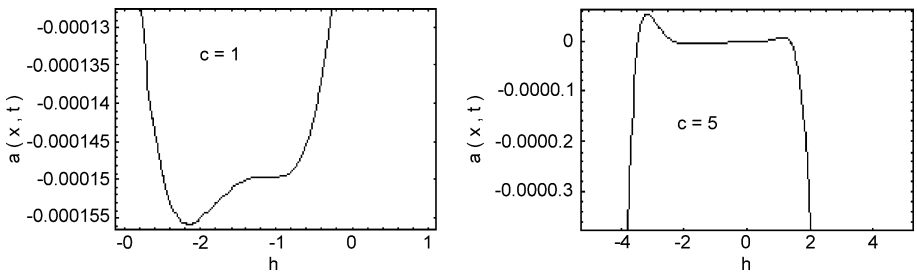

Korteweg–de Vries Equation (KdV), Some Numerical Methods for Solving the, Figure 2
The $\hbar$-curves for $x = 10$, $t = 0.5$, $c = 1$ for 5th-order approximation and $x = 10$, $t = 0.5$, $c = 5$ for 10th-order approximation of $u(x, t)$, respectively

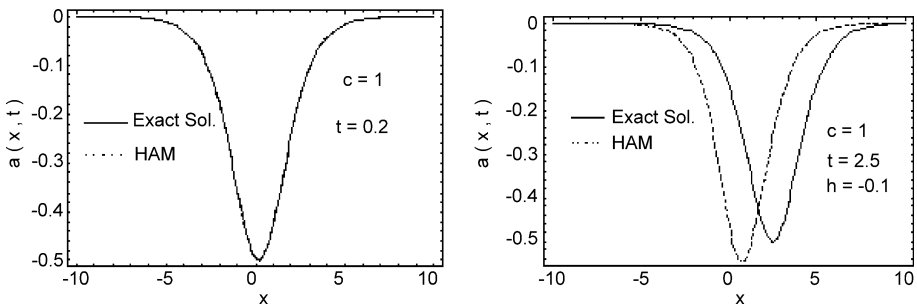

Korteweg–de Vries Equation (KdV), Some Numerical Methods for Solving the, Figure 3
The comparison of the HAM (5th-order approximation) and the exact solutions for Eq. (34) at $c = 1$

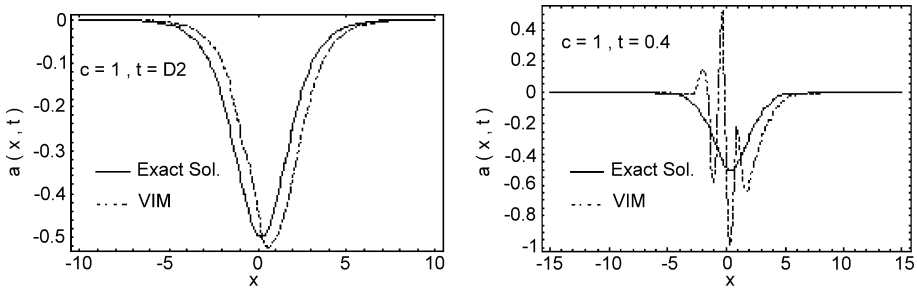

Korteweg–de Vries Equation (KdV), Some Numerical Methods for Solving the, Figure 4
The comparison of the VIM (five iterations) and the exact solutions for Eq. (34) at $c = 1$

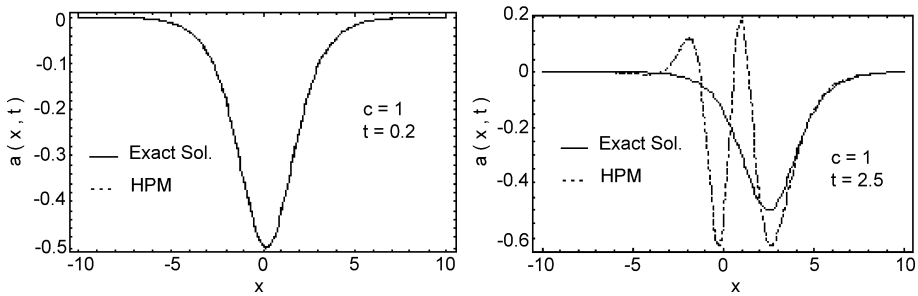

Korteweg–de Vries Equation (KdV), Some Numerical Methods for Solving the, Figure 5
The comparison of the HPM (five iterations) and the exact solutions for Eq. (34) at $c = 1$

Korteweg–de Vries Equation (KdV), Some Numerical Methods for Solving the, Table 1
Error between the ADM (φ_5) and exact solutions for the KdV Eq. (14) at $c = 1$

t/x	10	15	20	25	30
0.1	2.00679×10^{-4}	1.35233×10^{-6}	9.11171×10^{-9}	6.13942×10^{-11}	4.13671×10^{-13}
0.2	2.22178×10^{-4}	1.49452×10^{-6}	1.00712×10^{-8}	6.78512×10^{-11}	4.57177×10^{-13}
0.3	2.45102×10^{-4}	1.65169×10^{-6}	1.11292×10^{-8}	7.49865×10^{-11}	5.05255×10^{-13}
0.4	2.70872×10^{-4}	1.82535×10^{-6}	1.22991×10^{-8}	8.28713×10^{-11}	5.58381×10^{-13}
0.5	2.99336×10^{-4}	2.01722×10^{-6}	1.35919×10^{-8}	9.15815×10^{-11}	6.17071×10^{-13}

Korteweg–de Vries Equation (KdV), Some Numerical Methods for Solving the, Table 2
Error between the ADM (φ_{10}) and exact solutions for the KdV Eq. (34) at $c = 5$

t/x	10	15	20	25	30
0.1	2.00679×10^{-8}	1.35233×10^{-13}	9.11171×10^{-18}	6.13942×10^{-23}	4.1367×10^{-28}
0.2	2.22178×10^{-8}	1.49452×10^{-12}	1.00712×10^{-18}	6.78512×10^{-23}	4.5717×10^{-27}
0.3	2.45102×10^{-7}	1.65169×10^{-12}	1.11292×10^{-17}	7.49865×10^{-22}	5.0525×10^{-27}
0.4	2.70872×10^{-7}	1.82535×10^{-12}	1.22991×10^{-17}	8.28713×10^{-22}	5.5838×10^{-26}
0.5	2.99336×10^{-6}	2.01722×10^{-11}	1.35919×10^{-16}	9.15815×10^{-21}	6.1707×10^{-26}

Korteweg–de Vries Equation (KdV), Some Numerical Methods for Solving the, Table 3
Error between the HAM (5th-order approximation) and exact solutions for the KdV Eq. (34) at $c = 1$ and $\hbar = -0.5$

t/x	10	15	20	25	30
0.1	1.94277×10^{-4}	1.30915×10^{-6}	8.82101×10^{-9}	5.94355×10^{-11}	4.00473×10^{-13}
0.2	2.08022×10^{-4}	1.40179×10^{-6}	9.44517×10^{-9}	6.36412×10^{-11}	4.28811×10^{-13}
0.3	2.22925×10^{-4}	1.50222×10^{-6}	1.01219×10^{-8}	6.82009×10^{-11}	4.59534×10^{-13}
0.4	2.39102×10^{-4}	1.61125×10^{-6}	1.08565×10^{-8}	7.31508×10^{-11}	4.92886×10^{-13}
0.5	2.56683×10^{-4}	1.72975×10^{-6}	1.16549×10^{-8}	7.85304×10^{-11}	5.29134×10^{-13}

Korteweg–de Vries Equation (KdV), Some Numerical Methods for Solving the, Table 4
Error between the HAM (10th-order approximation) and exact solutions for the KdV Eq. (34) at $c = 5$ and $\hbar = 0.5$

t/x	10	15	20	25	30
0.1	8.12684×10^{-9}	1.13334×10^{-13}	1.58053×10^{-18}	2.20415×10^{-23}	3.0738×10^{-28}
0.2	3.62126×10^{-8}	5.05121×10^{-13}	7.04272×10^{-18}	9.82156×10^{-23}	1.3696×10^{-27}
0.3	8.86781×10^{-7}	1.23668×10^{-12}	1.72463×10^{-17}	2.40512×10^{-22}	3.3541×10^{-27}
0.4	1.87306×10^{-7}	2.61211×10^{-12}	3.64277×10^{-17}	5.08009×10^{-22}	7.0845×10^{-27}
0.5	5.02718×10^{-7}	7.01076×10^{-12}	9.77698×10^{-17}	1.36347×10^{-21}	1.9014×10^{-26}

Korteweg–de Vries Equation (KdV), Some Numerical Methods for Solving the, Table 5
Error between the VIM (5 terms) and exact solutions for the KdV Eq. (34) at $c = 1$

t/x	10	15	20	25	30
0.1	2.33263×10^{-4}	1.57189×10^{-6}	1.05913×10^{-8}	7.13638×10^{-11}	4.24233×10^{-13}
0.2	2.99151×10^{-4}	2.01592×10^{-6}	1.35833×10^{-8}	9.15216×10^{-11}	6.16666×10^{-13}
0.3	3.81873×10^{-4}	2.57322×10^{-6}	1.73383×10^{-8}	1.16824×10^{-10}	7.87156×10^{-13}
0.4	4.84308×10^{-4}	3.26302×10^{-6}	2.19864×10^{-8}	1.48141×10^{-10}	9.98164×10^{-13}
0.5	6.09593×10^{-4}	4.10597×10^{-6}	2.76657×10^{-8}	1.86417×10^{-10}	1.25602×10^{-12}

Korteweg–de Vries Equation (KdV), Some Numerical Methods for Solving the, Table 6
Error between the HPM (five iterations) and exact solutions for the KdV Eq. (34) at $c = 1$

t/x	10	15	20	25	30
0.1	2.00679×10^{-4}	1.35232×10^{-6}	9.11171×10^{-9}	6.13942×10^{-11}	4.13671×10^{-13}
0.2	2.21782×10^{-4}	1.49452×10^{-6}	1.00734×10^{-8}	6.78513×10^{-11}	4.57177×10^{-13}
0.3	2.45102×10^{-4}	1.65169×10^{-6}	1.11296×10^{-8}	7.49865×10^{-11}	5.05255×10^{-13}
0.4	2.70871×10^{-4}	1.82535×10^{-6}	1.22991×10^{-8}	8.28716×10^{-11}	5.58381×10^{-13}
0.5	2.99337×10^{-4}	2.01722×10^{-6}	1.35919×10^{-8}	9.15815×10^{-11}	6.17071×10^{-13}

Korteweg–de Vries Equation (KdV), Some Numerical Methods for Solving the, Table 7
Error between the EFDM and exact solutions for the KdV Eq. (34) at $c = 1$, $h = 0.05$ and $k = 0.025$

t/x	10	15	20	25	30
0.1	6.15156×10^{-6}	4.14842×10^{-8}	2.79520×10^{-10}	1.88339×10^{-12}	1.26902×10^{-14}
0.2	7.54177×10^{-6}	5.09160×10^{-8}	3.43074×10^{-10}	2.31161×10^{-12}	1.55755×10^{-14}
0.3	9.14755×10^{-6}	6.19126×10^{-8}	4.17173×10^{-10}	1.70489×10^{-12}	1.89396×10^{-14}
0.4	1.10086×10^{-5}	7.46977×10^{-8}	5.03358×10^{-10}	3.39162×10^{-12}	1.38607×10^{-14}
0.5	1.35102×10^{-5}	8.94343×10^{-8}	6.03380×10^{-10}	4.06554×10^{-12}	2.73934×10^{-14}

solitary wave solutions. It was solved by using analytical and numerical methods by many authors. We have mentioned these analytical and numerical methods in the introduction. The theoretical studies on the KdV equation have been nearly completed, but we may find new soliton solutions by producing new analytical methods or expanding the old methods. It is also possible to perform same operations on numerical methods. We may show the interaction of KdV solitons with each other by numerical.

Bibliography

Primary Literature

1. Russell JS (1844) Reports on waves. Report of the 14th Meeting of the British Association for the Advancement of Science, London, pp 311–390
2. Korteweg DJ, de Vries G (1895) On the change of form of long waves advancing in a rectangular canal and on a new type of long stationary waves. Phil Mag 39:422–443
3. Fermi E, Pasta J, Ulam S (1955) Studies of nonlinear problems I. Los Alamos Report LA-1940, Los Alamos, New Mexico
4. Zabusky NJ, Kruskal MD (1965) Interaction of solitons in a collisionless plasma and the recurrence on initial states. Phys Rev Lett 15:240–243
5. Zabusky NJ (1967) A synergetic approach to problems of nonlinear dispersive wave propagation and interaction. In: Ames WF (ed) Proceedings of the Symposium on Nonlinear Partial Differential Equations. Academic Press, Boston, pp 223–258
6. Lax PD (1968) Integrals of nonlinear equations of evolution and solitary waves. Commun Pure Appl Math 21:467–490
7. Gardner CS, Greene JM, Kruskal MD, Miura RM (1967) Method for solving the KdV equation. Phys Rev Lett 19:1095–1097
8. Gardner CS, Greene JM, Kruskal MD, Miura RM (1974) Korteweg–de Vries equation and generalizations VI methods for exact solutions. Commun Pure Appl Math 27:97–133
9. Hirota R (1971) Exact solution of the Korteweg–de Vries equation for multiple collisions of solitons. Phys Rev Lett 27:1192–1194
10. Hirota R (1973) Exact envelope-soliton solutions of a nonlinear wave equation. J Math Phys 14:805–809
11. Zabusky NJ, Galvin CJ (1971) Shallow water waves, the Korteweg–de Vries equation and solitons. J Fluid Mech 47:811–824
12. Hammack JL, Segur H (1974) The Korteweg–de Vries equation and water waves, Part 2, comparison with experiments. J Fluid Mech 65:289–314
13. Weidman PD, Maxworthy T (1978) Experiments on strong interactions, between solitary waves. J Fluid Mech 85:417–431
14. Dodd RK, Eilbeck JC, Gibbon JD, Morris HC (1982) Solitons and Nonlinear Wave Equations. Academic Press, New York

15. Lombdahl PS (1984) What is a soliton? Los Alamos Sci 10:27–31
16. Debnath L (1997) Nonlinear Partial Differential Equations for Scientists and Engineers. Birkhauser, Boston
17. Helal MA (2002) Soliton solution of some nonlinear partial differential equations and its applications in fluid mechanics. Chaos Solitons Fractals 13:1917–1929
18. Davidson RC (1972) Methods in Nonlinear Plasma Theory. Academic Press, New York
19. Lamb GL (1980) Elements of Soliton Theory. Wiley, New York
20. Drazin PG, Johnson RS (1989) Solitons: An Introduction. Cambridge University Press, New York
21. Drazin PG (1992) Nonlinear Systems. Cambridge University Press, New York
22. Khater AH, Helal MA, El-Kalaawy OH (1998) Backlund transformations: exact solutions for the KdV and Calogero-Degasperis-Fokas mKdV equations. Math Meth Appl Sci 21:719–731
23. Khater AH, El-Sabbagh MF (1989) The pain level property and coordinates transformations. Nuovo Cimento B 104:123–129
24. Wahlquist HD, Estabrook FB (1973) Backlund transformation for solutions of the KdV equation. Phys Rev Lett 31:1386–1390
25. Olver PJ (1986) Applications of Lie Groups to Differential equations Graduate Texts in Mathematics. Springer, Berlin
26. Hereman W, Korpel A, Banerjee PP (1985) A general physical approach to solitary wave construction from linear solutions. Wave Motion 7:283–289
27. Malfliet W (1992) Solitary wave solutions of nonlinear wave equations. Am J Phys 60:650–654
28. Yan CT (1996) New explicit solitary wave solutions and periodic wave solutions for Whitham-Broer-Kaup equation in shallow water. Phys Lett 224:77–84
29. Wang ML (1996) Exact solutions for a compound KdV-Burgers equation. Phys Lett 213(5–6):279–287
30. Yan ZY, Zhang HQ (2001) A simple transformation forn nonlinear waves. Phys Lett 285:355–362
31. Peng YZ (2003) Exact solutions for some nonlinear partial differential equations. Phys Lett 314(5–6):401–408
32. Sawada K, Kotera T (1974) A method for finding N-solitons of the KdV equation and KdV-like equation. Prog Theor Phys 51:1355–1367
33. Rosales RR (1978) Exact solutions of some nonlinear evolution equations. Stud Appl Math 59:117–157
34. Whitham GB (1927) Linear and Nonlinear Waves. Pure and Applied Mathematics. Wiley/Interscience, New York
35. Wadati M, Sawada K (1980) New representations of the soliton solution for the KdV equation. J Phys Soc Jpn 48:312–318
36. Wadati M, Sawada K (1980) Application of the trace method to the modified KdV equation. J Phys Soc Jpn 48:319–326
37. Helal MA, El-Eissa HN (1996) Shallow water waves and KdV equation. PUMA 7:263–282
38. Khater AH, El-Kalaawy, Helal MA (1997) Two new classes of exact solutions for the KdV equation via Backlund transformations. Chaos Solitons Fractals 8:1901–1909
39. Khater AH, Helal MA, Seadawy AR (2000) General soliton solutions of an n-dimensional nonlinear Schrödinger equations. Nuovo Cimento B 115:1303–1311
40. Helal MA (2001) Chebyshev spectral method for solving KdV equation with hydro-dynamical application. Chaos Solitons Fractals 12:943–950
41. Das G, Sarma J (1999) A new mathematical approach for finding the solitary waves in dusty plasma. Phys Plasmas 6:4394–4397
42. Osborne A (1995) The inverse scattering transform: Tools for the nonlinear fourier analysis and filtering of ocean surfaces waves. Chaos Solitons Fractals 5:2623–2637
43. Ostrovsky L, Stepanyants YA (1989) Do internal solitons exist in the ocean? Rev Geophys 27:293–310
44. Ludu A, Draayer JP (1998) Nonlinear modes of liquid drops as solitary waves. Phy Rev Lett 80:2125–2128
45. Reatto L, Galli D (1999) What is a ROTON? Int J Modern Phys B 13:607–616
46. Turitsyn S, Aceves A, Jones C, Zharnitsky V (1998) Average dynamics of the optical soliton in communication lines with dispersion management: Analytical results. Phys Rev E 58:48–51
47. Coffey MW (1996) Nonlinear dynamics of vortices in ultraclean type-II superconductors: Integrable wave equations in cylindrical geometry. Phys Rev B 54:1279–1285
48. Gou BY, Shen J (2001) On spectral approximations using modified Legendre rational functions: Application to the KdV equation on the half line. Indiana Univ Math J 50:181–204
49. Huang WZ, Sloan DM (1992) The pseudo spectral method for third-order differential equations. SIAM J Numer Anal 29:1626–1647
50. Man HP, Sun WW (2000) A Legendre-Petrov-Galerkin and Chebyshev collocation methods for third-order differential equations. SIAM J Numer Anal 38:1425–1438
51. Man HP, Sun WW (2001) Optimal error estimates of the Legendre-Petrov-Galerkin method for the KdV equation. SIAM J Numer Anal 39:1380–1394
52. Bahadır AR (2005) Exponential finite-difference method applied to the KdV equation for small times. Appl Math Comput 160:675–682
53. Jain PC, Shankar R, Bhardwaj D (1997) Numerical solution of the KdV equation. Chaos Solitons Fractals 8:943–951
54. Soliman AA (2004) Collocation solution of the KdV equation using septic splines. Int J Comput Math 81:325–331
55. Frauendiener J, Klein C (2006) Hyperelliptic theta-functions and spectral methods: KdV and KP solutions. Lett Math Phys 76:249–267
56. Bhatta DD, Bhatti MI (2006) Numerical solution of KdV equation using modified Bernstein polynomials. Appl Math Comput 174:1255–1268
57. Helaln MA, Mehanna MS (2007) A comparative study between two different methods for solving the GKdV equation. Chaos Solitons Fractals 33:729–739
58. Kutluay S, Bahadır AR, Özdeş A (2000) A small time solutions for the KdV equation. Appl Math Comput 107:203–210
59. Özer S, Kutluay S (2005) An analytical numerical method for solving the KdV equation. Appl Math Comput 164:789–797
60. Dehghan M, Shokri A (2007) A numerical method for KdV equation using collocation and radial basis functions. Nonlinear Dyn 50:111–120
61. Dağ I, Dereli Y (2008) Numerical solutions of KdV equation using radial basis functions. Appl Math Model 32:535–546
62. Wadati M (1973) The modified Kortweg–de Vries equation. J Phys Soc Jpn 34:1289–1296
63. Wadati M (2001) Introduction to solitons. Pramana J Phys 57(5–6):841–847
64. Adomian G (1989) Nonlinear Stochastic Systems and Applications to Physics. Kluwer, Boston

65. Adomian G (1994) Solving Frontier Problems of Physics: The Decomposition Method. Kluwer, Boston
66. Adomian G (1998) A review of the decomposition method in applied mathematics. J Math Anal Appl 135:501–544
67. Lesnic D, Elliott L (1999) The decomposition approach to inverse heat conduction. J Math Anal Appl 232:82–98
68. Wazwaz AM (2002) Partial Differential Equations: Methods and Applications. Balkema Publication, Lisse
69. Wazwaz AM (2003) An analytical study on the third-order dispersive partial differential equation. Appl Math Comput 142:511–520
70. Cherruault Y (1998) Modéles et Méthodes Math ématiques pour les Sciences du Vivant. Presses Universitaires de France, Paris
71. Dehghan M (2004) Application of ADM for two-dimensional parabolic equation subject to nonstandard boundary specification. Appl Math Comput 157:549–560
72. Dehghan M (2004) The solution of a nonclassic problem for one-dimensional hyberbolic equation using the decomposition procedure. Int J Comput Math 81:979–989
73. Babolian E, Biazar J, Vahidi AR (2004) The decomposition method applied to systems of Fredholm integral equations of the second kind. Appl Math Comput 148:443–452
74. El-Sayed MS (2002) The modified decomposition method for solving nonlinear algebraic equations. Appl Math Comput 132:589–597
75. Inc M (2006) On numerical Jacobi elliptic function solutions of (1+1)-dimensional dispersive long wave equation by the decomposition method. Appl Math Comput 173:372–382
76. Inc M (2007) Exact solutions with solitary patterns for the Zakharov-Kuznetsov equations with fully nonlinear dispersion. Chaos Solitons Fractals 33:1783–1790
77. Cherruault Y (1990) Convergence of Adomian's decomposition method. Math Comput Modelling 14:83–86
78. Ngarhasta N, Some B, Abboui K, Cherruault Y (2002) New numerical study of Adomian method applied to a diffusion model. Kybernetes 31:61–75
79. Mavoungou T, Cherruault Y (1992) Convergence of Adomian's method and applications to nonlinear partial differential equations. Kybernetes 21:13–25
80. Inc M (2006) On numerical soliton solution of the KK equation and convergence analysis of the decomposition method. Appl Math Comput 172:72–85
81. Hashim I, Noorani MS, Al-Hadidi MRS (2006) Solving the generalized Burgers-Huxley equation using the Adomian decomposition method. Math Comput Model 43:1404–1411
82. Liao SJ (2002) An analytic approximation of the drag coefficient for the viscous flow past a sphere. Int J Nonlinear Mech 37:1–18
83. Liao SJ (2003) Beyond Perturbation: Introduction to the Homotopy Analysis Method. Champan, Boca Raton
84. Liao SJ (2003) On the analytic solution of magnetohydrodynamic flows non-Newtonian fluids over a stretching sheet. J Fluid Mech 488:189–212
85. Liao SJ (2005) A new branch of solutions of boundary-layer flows over an impermeable stretched plate. Int J Heat Mass Transfer 48:2529–2539
86. Sajid M, Hayat T, Asghar S (2006) On the analytic solution of the steady flow of a fourth grade fluid. Phys Lett 355:18–26
87. Hayat T, Abbas Z, Sajid M (2006) Series solution for the upper-convected Maxwell fluid over a porous stretching plate. Phys Lett 358:396–403
88. Hayat T, Khan M (2005) Homotopy solutions for a generalized second-grade fluid past a porous plate. Nonlinear Dynamics 42:395–405
89. Abbasbandy S (2006) The application of homotopy analysis method to nonlinear equations arising in heat transfer. Phys Lett 360:109–113
90. Abbasbandy S (2007) The application of homotopy analysis method to solve a generalized Hirota-Satsuma coupled KdV equation. Phys Lett 361:478–483
91. Inc M (2007) On exact solution of Laplace equation with Dirichlet and Neumann boundary conditions by the homotopy analysis method. Phys Lett 365:412–415
92. He JH (1997) A new approach to nonlinear partial differential equations. Commun Nonlinear Sci Numer Simul 2:230–235
93. He JH (1997) Variational iteration method for delay differential equations. Commun Nonlinear Sci Numer Simul 2:235–236
94. He JH (1999) Variational iteration method a kind of nonlinear analytical technique: Some examples. Int J Nonlinear Mech 34:699–708
95. He JH (2000) Variational iteration method for autonomous ordinary differential systems. Appl Math Comput 114:115–123
96. He JH (2006) Some asymptotic methods for strongly nonlinear equations. Int J Modern Phys B 20:1141–1199
97. Abdou MA, Soliman AA (2005) Variational iteration method for solving Burger's and coupled Burger's equations. J Comput Appl Math 181:245–251
98. Abdou MA, Soliman AA (2005) New applications of variational iteration method. Phys D 211:1–8
99. Bildik N, Konuralp A (2006) The use of variational iteration method, differential transform method and Adomian decomposition method for solving different types of nonlinear partial differential equations. Int J Nonlinear Sci Numer Simul 7:65–70
100. Draganescu G (2006) Application of a variational iteration method to linear and nonlinear viscoelastic models with fractional derivatives. J Math Phys 47:082902
101. Tatari M, Dehghan M (2007) He's variational iteration method for computing a control parameter in a semi-linear inverse parabolic equation. Chaos Solitons Fractals 33:671–677
102. Wazwaz AM (2007) The variational iteration method for exact solutions of Laplace equation. Phys Lett 363:260–262
103. Tatari M, Dehghan M (2007) Solution of problems in calculus of variations via He's variational iteration method. Phys Lett 362:401–406
104. Inc M (2007) An approximate solitary wave solution with compact support for the modified KdV equation. Appl Math Comput 184:631–637
105. Inc M (2007) Exact and numerical solutions with compact support for nonlinear dispersive K(m,p) equations by the variational iteration method. Physica 375:447–456
106. He JH (1999) Homotopy perturbation technique. Comput Meth Appl Mech Eng 178:257–262
107. He JH (2005) Application of homotopy perturbation method to nonlinear wave equations. Chaos Solitons Fractals 26:695–700
108. He JH (2005) Homotopy perturbation method for bifurcation of nonlinear problems. Int J Nonlinear Sci Numer Simul 6:207–208

109. He JH (2006) Non-perturbative Methods for Strongly Nonlinear Problems. Dissertation. de-Verlag im Internet GmbH, Berlin
110. Ganji DD, Rajabi A (2006) Assesment of homotopy perturbation and perturbation methods in heat radiation equations. Int Commun Heat Mass Transf 33:391–400
111. Siddiqui AM, Mahmood R, Ghori QK (2006) Homotopy perturbation method for thin film flow of a fourth grade fluid down a vertical cylinder. Phys Lett 352:404–410
112. Ganji DD, Rafei M (2006) Solitary wave solutions for a generalized Hirota-Satsuma coupled KdV equation by homotopy perturbation method. Phys Lett 356:131–137
113. Chowdhury MSH, Hashim I (2007) Solutions of a class of singular second-order IVPs by homotopy-perturbation method. Phys Lett 356(5–6):439–447
114. Smith GD (1987) Numerical Solution of Partial Differential Equation: Finite Difference Methods. Oxford University Press, New York
115. Vilegenthart AC (1971) On finite difference methods for the Korteweg–de Vries equations. J Eng Math 5(2):137–155

Books and Reviews

Abbasbandy S (2007) An approximation solution of a nonlinear equation with Riemann–Liouville's fractional derivatives by He's variational iteration method. J Comp Appl Math 207: 53–58

Abbasbandy S (2007) A new application of He's variational iteration method for quadratic Riccati differential equation by using Adomian's polynomials. J Comp Appl Math 207:59–63

Deeba E, Khuri SA, Xie S (2000) An algorithm for solving boundary value problems. J Comp Phys 159:125–138

El-Sayed SM (2002) The modified decomposition method for solving nonlinear algebraic equations. Appl Math Comput 132:589–597

Ganji DD, Sadighi A (2007) Application of homotopy-perturbation and variational iteration methods to nonlinear heat transfer and porous media equations. J Comp Appl Math 207:24–34

Hayat T, Sajid M (2007) Analytic solution for axisymmetric flow and heat transfer of a second grade fluid past a stretching sheet. Int J Heat Mass Transf 50:75–84

Kaya D (2005) An application for the higher order modified KdV equation by decomposition method. Commun Nonlinear Sci Numer Simul 10(6):693–702

Kaya D, Aassila M (2002) An application for a generalized KdV equation by the decomposition method. Phys Lett 299(2):201–206

Khelifa S, Cherruault Y (2000) New results for the Adomian method. Kybernetes 29(3):332–354

Öziş T, Yıldırım A (2007) A note on He's homotopy perturbation method for van der Pol oscillator with very strong nonlinearity. Chaos Solitons Fractals 34(3):989–991

Wang C, Wu Y, Wu W (2005) Solving the nonlinear periodic wave problems with the Homotopy Analysis Method. Wave Motion 41(4):329–337

Wazwaz AM (1997) A First Course in Integral Equations. World Scientific, London

Wu Y, Wang C, Liao S (2005) Solving the one-loop soliton solution of the Vakhnenko equation by means of the Homotopy analysis method. Chaos Solitons Fractals 23(5):1733–1740

L

Learning in Games

JOHN NACHBAR
Department of Economics, Washington University, St. Louis, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Deterministic Learning
Stochastic Learning
Future Directions
Bibliography

Glossary

Bayesian learning In repeated games, a model in which each player best responds to her prior, which is a probability distribution over her opponent's behavior strategies.

Behavior strategy In a repeated game, a behavior strategy for player i gives, for each possible date and each possible history of play in the game up to that date, a probability distribution over i's actions next period. This includes the possibility that player i may play some action for certain.

Belief learning In repeated games, a model in which players best respond to prediction rules.

Prediction rule In a two-player repeated game, a *deterministic prediction rule* gives a probability distribution over the opponent's actions next period as a deterministic function of the history of the game. In a *stochastic* prediction rule, the distribution over the opponent's actions can depend on history probabilistically. A deterministic prediction rule that gives player 2's forecast about player 1's actions is formally equivalent to a behavior strategy for player 1.

Repeated games A repeated game is an extensive form game representing a repeated strategic interaction. The interaction being repeated is called the *stage game*. In a *discounted* repeated game, payoffs in the repeated game are a geometrically weighted sum of the payoffs each period from the stage game. The weight on period t payoffs is δ^t, where $\delta \in (0, 1)$ is the *discount factor*. A player who is patient has a discount factor close to 1.

Definition of the Subject

In the context of this article, *learning* refers to a particular class of dynamic game theoretic models. In models in this class, players are "rational" but not necessarily in equilibrium: players forecast, possibly inaccurately, the future behavior of their opponents and optimize, or ε optimize, with respect to their forecasts. Two objectives are to model out-of-equilibrium behavior by sophisticated players and to understand when, or whether, play might converge to equilibrium.

Learning models are a branch of a larger literature on out-of-equilibrium behavior in dynamic games. In other branches of the literature, players are modeled as "adaptive". Players do not forecast and they do not optimize. Rather, they follow some other form of behavioral rule, such as imitation, regret minimization, or reinforcement. Learning models, especially those that attempt to capture sophisticated behavior, are most appropriate in settings where players have a good understanding of their strategic environment and where the stakes are high enough to make forecasting and optimization worthwhile. For surveys of evolutionary/adaptive models, see ► Evolutionary Game Theory and [11,73,74].

Introduction

This essay surveys research on learning in games. The subject is as old as game theory itself, dating to the work of Cournot in the early nineteenth century. But, with some important exceptions, research on learning in games was largely dormant until the late 1980s.

I have divided the survey into two parts. The first part covers *deterministic* learning, in which each player optimizes with respect to a prediction rule that gives a probability distribution over the opponent's next period actions as a deterministic function of the history of the game to date. Cournot's original dynamic model falls into this category. The second part briefly surveys some of the literature on *stochastic* learning, in which each player optimizes with respect to a prediction rule that is a probabilistic function of history. This survey expands and updates my earlier survey [58]. For a complementary survey, with a different emphasis and more detail on some topics, see [28].

One can formulate learning within the context of very general dynamic games. For this survey, however, I focus on the most studied case: discounted infinitely repeated games with perfect monitoring (at each stage, players know the sequence of action profiles chosen at previous stages). In Subsect. "Convergence to What Sort of Equilibrium?", I discuss an important variant: repeated extensive form games in which the perfect monitoring assumption is violated. Also for simplicity, I focus mostly on two-player games.

Deterministic Learning

Classical Learning Models

I begin by introducing two "classical" learning models, Cournot's best response dynamics and fictitious play. In both models, players are unsophisticated: they cannot learn to forecast even simple patterns in opponent behavior, such as deterministic alternation between two stage games actions. But Cournot best response and fictitious play are historically important, easy to describe, relatively tractable, heavily studied, often used as building blocks in more elaborate dynamic models (e. g. [65]), and can be used to illustrate some of the major themes in the literature.

Cournot's Best Response Dynamics Following Cournot [12], consider two mineral spring owners, A and B, who are each trying to make money by selling water. Suppose that the profit to firm A is $(13 - Q^A - Q^B)\,Q^A - Q^A$, where Q^i is the output in one period by firm i, $13 - Q^A - Q^B$ gives the market price, and output costs 1 per unit, hence total cost to firm A is Q^A. The quantity Q^A that maximizes A's profit depends on Q^B, and so in order to choose Q^A, firm A must forecast Q^B. In a model of sophisticated behavior, firm A would want to forecast not only Q^B next period but also how B might respond in subsequent periods. But, for the moment, assume that A is myopic, meaning that A chooses Q^A to maximize next period's expected profit. Then A should set,

$$Q^A = \frac{12 - \mathbb{E}[Q^B]}{2}\,,$$

where $\mathbb{E}[Q^B]$ is A's expectation of Q^B.

What should $\mathbb{E}[Q^B]$ be? In Cournot's story, A puts probability one on B continuing to produce next period whatever B produced this period, hence $\mathbb{E}[Q^B_{t+1}] = Q^B_t$. Cournot argued that if both firms adjust in this way, then eventually Q^A and Q^B converge to the same number, Q^*, which must satisfy,

$$Q^* = \frac{12 - Q^*}{2}\,,$$

hence

$$Q^* = 4\,.$$

In modern terminology, the symmetric output profile $(4, 4)$ is the Nash equilibrium of the single period version of this game. A generalized version of Cournot's mineral springs game has become the benchmark model in the economic field of Industrial Organization. Cournot not only identified the Nash equilibrium of this game, but he also provided a dynamic rationale for why the Nash equilibrium might arise. Cournot, in his exposition, was actually more interested in the dynamic story than in the equilibrium.

Fictitious Play Fictitious play and its variants, first introduced by Brown [10] in the 1950s, is the single most heavily studied learning model. The term "fictitious play" refers to the idea that players would run through the fictitious play dynamic in their minds, prior to the start of actual play, as a way of forecasting their opponent's behavior once play begins. In recent years, however, fictitious play has usually been interpreted as a model of actual behavior in real time.

In its modern form, fictitious play dictates that each player predicts that the probability that her opponent will play an action, say L, next period is a weighted sum of an initial probability on L and the frequency with which L has been chosen to date. The weight on the frequency is $t/(t + k)$, where t is the number of periods thus far and $k > 0$ is a fixed parameter. The larger is k, the longer the initial probability on L dominates the player's forecast about L.

In fictitious play, players are *as if* myopic: for any k and any discount factor, each player effectively believes he cannot influence his opponent's future behavior (see the discussion in Subsect. "Belief Learning and Bayesian Learning"), and so it is optimal to maximize current pe-

riod payoff, ignoring the future. As this observation suggests, forecasting under fictitious play is extremely naive. Fictitious play is, in fact, mathematically very close to regret minimization, an adaptive learning model; see [35] and [36]. To justify fictitious play for sophisticated players, [21] suggests thinking of environments in which there is a large population of players randomly matched (an environment normally associated with evolutionary models), rather than a standard repeated game; see also [29].

Belief Learning and Bayesian Learning

In general, *belief learning* refers to any learning model in which players best respond (or ε-best respond) to a deterministic prediction rule. Both the Cournot dynamic and fictitious play are examples of belief learning. An alternative approach to modeling learning is *Bayesian*. Recall that in a repeated game, a *behavior strategy* gives, for every history, a probability over the player's stage game actions next period. In a Bayesian learning model, each player chooses a behavior strategy that best responds (or ε-best responds) to a *prior*, a probability distribution over the opponent's behavior strategies. A basic observation in the theory of deterministic learning is that belief learning and Bayesian learning are mathematically equivalent.

Explicitly, player 1's prediction rule about player 2 is mathematically identical to a behavior strategy for player 2. Thus, any belief learning model is equivalent to a Bayesian model in which each player optimizes with respect to a prior that places probability one on her prediction rule, now reinterpreted as the opponent's behavior strategy.

Conversely, any Bayesian model is equivalent to a belief learning model. Explicitly, for any prior over player 2's behavior strategies there is a degenerate prior, assigning probability one to a particular behavior strategy, that is equivalent in the sense that both priors induce the same distributions over play in the game, no matter what behavior strategy player 1 herself adopts. This is a form of Kuhn's theorem [50]; for a version applicable to repeated games, see [4]. I refer to the behavior strategy used in the degenerate prior as a *reduced form* of the original prior. Thus, any Bayesian model is equivalent to a Bayesian model in which each player's prior places probability one on the reduced form, and any such Bayesian model is equivalent to a belief learning model.

To illustrate, consider fictitious play. For simplicity, consider a stage game with just two actions, L and R. By an i. i. d. strategy for player 2, I mean a behavior strategy in which player 2 plays L with probability q, independent of history. Thus, if $q = 1/2$ then player 2 always randomizes 50:50 between L and R. Fictitious play is equivalent to a degenerate Bayesian model in which each player places probability one on the fictitious play prediction rule, and one can show that this is equivalent in turn to a non-degenerate Bayesian model in which the belief is represented as a beta distribution over q; see, for example, [25]. The uniform distribution over q, for example, corresponds to taking the initial probability of L to be 1/2 and the parameter k to be 2. Thus, fictitious play can be thought of as a Bayesian model in which each player is certain that the other is playing an i. i. d. strategy, but isn't sure which i. i. d. strategy. Note that since each player believes the other is i. i. d., each believes she has no influence on the others future behavior. Hence, as noted earlier, myopic optimization is optimal for any discount factor.

Likewise, the Cournot prediction rule is Bayesian. In particular, one can form the degenerate prior from the Cournot prediction rule. To my knowledge, the Cournot rule has no compelling non-degenerate Bayesian interpretation.

The fact that both the Cournot dynamic and fictitious play can be interpreted as Bayesian underscores the fact that there is no presumption that Bayesian players are sophisticated. It is possible for a Bayesian to have an "idiotic" prior. I take up the issue of sophisticated learning below, especially in Subsect. "Sophisticated Learning".

What Should "Convergence" Mean?

The primary research question in the learning literature is whether play converges to equilibrium. One of the characteristics of the literature is a running dialog about what "converges to equilibrium" ought to mean.

Convergence in What Game? For concreteness, suppose that we are looking for convergence to Nash equilibrium; similar comments apply to convergence to other forms of equilibrium, such as correlated equilibrium.

In the models that are the focus of this essay, learning takes place within the context of a repeated game. Therefore, it is natural to look for convergence to Nash equilibria of the repeated game. Repeated play of a stage game Nash equilibrium always constitutes a Nash equilibrium of the repeated game, but typically there are also other types of repeated game equilibria.

First, even if players do play stage game Nash equilibria, they need not play the *same* stage game Nash equilibrium in every period. As an example, consider the repeated Battle of the Sexes, a stage game of which is given in Table 1. Alternation between (a, a) in odd periods and (b, b) in even periods, both of which are stage game

Learning in Games, Table 1
Battle of the Sexes

	a	b
a	8, 10	0, 0
b	0, 0	10, 8

Learning in Games, Table 2
Matching Pennies

	a	b
a	1, −1	−1, 1
b	−1, 1	1, −1

Nash equilibria, is a perfectly reasonable repeated game Nash equilibrium.

Second, if the discount factor is close enough to 1, the folk theorem implies that there will be many repeated game Nash equilibrium, and some of these may involve play of actions that, in the stage game, are not part of any stage game equilibrium (for more on the folk theorem, see [30]). The standard example is the repeated prisoner's dilemma, in which the repeated game can have Nash equilibria that sustain cooperation, even though cooperation is strictly dominated in the stage game.

More generally, learning stories take place in the context of a larger dynamic game. Learning may lead to Nash equilibrium play in the continuation of this larger game, but Nash equilibrium in the continuation of the larger game need not correspond to Nash equilibrium-like play in a component or projection of that game.

Convergence in What Sense? Until the 1990s, most of the work on learning focused on whether the empirical marginal frequencies of realized play converged to a Nash equilibrium of the stage game.

As an illustration, consider repeated Matching Pennies, the stage game for which is given in Table 2. The Nash equilibrium of repeated Matching Pennies, for any discount factor, calls for players to randomize 50:50 in every period, regardless of history. Under convergence of the empirical marginal frequencies, play is said to converge to Nash equilibrium if each player plays a half of the time. This criterion is satisfied if play alternates deterministically between (a, a) and (b, b). In contrast, a Nash equilibrium play path looks random; in particular, with high probability, over any finite set of dates, the four pure outcomes (a, a), (a, b), (b, a), and (b, b) occur about equally often. Convergence of the empirical marginal frequencies is thus a very weak convergence criterion.

A somewhat tougher criterion is convergence of the empirical *joint* frequencies. In the matching pennies example, this would require that each of the four pure action profiles gets played 1/4 of the time. This convergence criterion eliminates the previous example but it is still extremely weak. It would, for example, allow convergence to the deterministic sequence (a, a), (a, b), (b, a), (b, b), (a, a), … to count as convergence to a Nash equilibrium. In contrast, in the Nash equilibrium, by the strong law of large numbers, (a, a) gets played at dates $t = 1, 5, 9, \ldots$ only 1/4 of the time. These difficulties with defining convergence in terms of empirical frequencies were first emphasized by Fudenberg and Kreps; see [21] and also the related paper by Jordan [43].

Choice of convergence standard is obviously a matter of balance. Choose too weak a standard, as in the examples above, and convergence is arguably not meaningful. Choose too strong a standard and one gets impossibility results, even though positive results are still available for other, weaker but arguably still satisfactory, forms of convergence. For a stark example, consider the Battle of the Sexes game of Table 1. If players start in the repeated game Nash equilibrium in which they play (a, a) in odd periods and (b, b) in even periods then play converges (trivially) to Nash equilibrium play in any continuation game. But note that, strictly speaking, we get a *different* Nash equilibrium depending on whether the starting date of the continuation game is odd or even. So a convergence standard that requires that play be close to the *same* repeated game Nash equilibrium in every continuation game yields non-convergence, even in this trivial example. For a more subtle example illustrating the tradeoffs in choosing the right convergence standard, see Subsect. "Payoff Uncertainty".

Loosely, the strongest form of convergence that one can generally hope for is that, to an outside observer, play over finite continuation histories looks asymptotically like play of some repeated game Nash equilibrium or ε-Nash equilibrium. Again, one must allow for the possibility of convergence to the play of different Nash equilibria in different continuation games, even along the same play path.

Convergence to What Sort of Equilibrium? Although my focus here is primarily on Nash equilibrium, other solution concepts are often of interest. Obvious alternatives are rationalizable strategy profiles and correlated equilibria and both have received attention (for example [7,16], and [61]).

Two other variants of Nash equilibrium, less common in standard game theory texts, have also proved important in the literature. First, it is often natural to assume that

players ε optimize rather than exactly optimize. If players only ε optimize then convergence will be to an ε-Nash equilibrium rather than to an exact Nash equilibrium. It is common in the literature to model myopic players as using a logit selection from the stage game's ε-best response correspondence. In this case, ε-Nash equilibrium in the stage game takes the form of a McKelvey and Palfrey *quantile response equilibrium* (QRE) [52], a solution concept that has become important in the experimental game theory literature (see, for example, [32]). While any Nash equilibrium is an ε-Nash equilibrium, ε-Nash equilibria are typically not Nash equilibria, although the difference is small if ε is small.

A second important variant arises in repeated games in which the stage game has a non-trivial extensive form. In such settings, the perfect monitoring assumption may be untenable. In particular, it may make sense to assume instead that while players observe the outcome of the stage game, they do not observe the full strategy profile for the stage game.

To take a concrete example, consider the repeated ultimatum bargaining game, in which, in the stage game, player A makes an offer in the form of an integer $x \in [0, 100]$ and player B either rejects, yielding a payoff profile of $(0, 0)$, or accepts, yielding a payoff profile of $(x, 100 - x)$. It may make sense to assume that while player A can observe player B's action (accept or reject) in response to the actual offer, she cannot observe player B's entire stage game strategy, since this would mean observing how B would have responded to every other possible offer.

For this setting, Fudenberg and Levine [22] propose *self-confirming equilibrium* (SCE) as an alternative to Nash equilibrium. There are actually several forms of SCE, depending on the setting and the information of the players. The basic idea is that in a SCE of the stage game, players are optimizing with respect to predictions about their opponents that are correct along the stage game play path. Predictions about how the opponent would behave off the stage game play path, however, may be wrong. SCE bears a family resemblance to Hahn's conjectural equilibrium [33] and Kalai and Lehrer's subjective equilibrium [46].

For players to learn to make correct predictions about stage game behavior off the play path, they must experiment with alternative play. Fudenberg and Levine show, in the context of a large population, steady-state learning model [23], that if players are sufficiently patient then they have enough incentive to experiment, and thus will learn to play a stage game Nash equilibrium. In other settings, or if players are not sufficiently patient, it may be plausible to expect players to end up at SCE that are far from any Nash equilibrium. Fudenberg and Levine suggest that this may explain some real world behaviors that may otherwise seem at variance with standard game theory predictions. A provocative example can be found in [27].

Learning in Games, Table 3
The Shapley game

	a	b	c
a	1, 0	0, 0	0, 1
b	0, 1	1, 0	0, 0
c	0, 0	0, 1	1, 0

Finally, some other solution concepts have received attention, notably Samet and Kalai's persistent retracts [48], Basu and Weibull's CURB (closed under rational behavior) sets [5], and Voorneveld's minimal prep sets [70] and [71]. These solution concepts tend to arise most readily in stochastic learning models, which I discuss in Sect. "Stochastic Learning".

Convergence in Classical Learning Models

Neither Cournot best response dynamics nor fictitious play exhibit universal convergence, by which I mean convergence for all stage games. It is easy to construct non-convergence examples for Cournot best response, even for games with pure strategy equilibria. For fictitious play, there are classes of games that cause convergence problems even when one considers only weak forms of convergence, such as convergence of the empirical marginal distributions. The classic example of a problem game, due to Shapley [69], is given in Table 3. In this game, if the prediction rules concentrate initial probability on, say, (b, a), then play cycles, starting at (a, a) then moving to (a, c), then to (c, c) and so on. Moreover the cycle becomes slower and slower, so that not even the empirical marginal frequencies converge to the unique Nash equilibrium of the stage game, which has each player randomize $(1/3, 1/3, 1/3)$. See also [43]. Hart and Mas-Colell [37] prove a kind of generalization of the Shapley example: for a large class of dynamics that include a continuous time version of fictitious play as a special case, one can always find a stage game and initial conditions for which convergence to stage game equilibrium play fails.

On a more positive note, Cournot best response dynamics do converge to pure strategy stage game equilibrium if the stage game is solvable by the iterated deletion of strictly dominated strategies. See [7,55], and, for a more general class of models [53].

As for fictitious play, there are ε optimizing variants of fictitious play that yield convergence to approximate stage game equilibrium play (possibly mixed) for all zero sum

games, all games with an interior ESS (evolutionarily stable strategy), and all common interest games, in addition to all games that are strict dominance solvable, with the approximation closer the smaller is ε. Somewhat weaker convergence results are available for supermodular games. These claims follow from results in [39], which builds on [6] and [21], and which also provides additional references to the large literature on fictitious play.

The use of ε optimization, or something analogous (such as small, privately observed, payoff shocks in each period), is necessary. In a standard repeated game setting, for almost every stage game, standard fictitious play, with exact optimization, yields pure continuation play paths (see [21]). Thus, in games like repeated matching pennies, standard fictitious play cannot give convergence to an equilibrium play path. This is an instance of a recurring modeling theme in the learning literature: it is often easier to get positive convergence results for *vp* optimization than for exact optimization.

Kalai–Lehrer Learning

Kalai and Lehrer [45] (hereafter KL) take a Bayesian perspective and ask what conditions on priors are sufficient to give convergence to equilibrium, or approximate equilibrium, play. I find it helpful to characterize KL, and related papers, in the following way. Say that a prior profile (giving a prior for each player) has the *learnable best response property* (LBR) if there is a profile of best response (or ε-best response) strategies (LBR strategies) such that, if the LBR strategies are played, then each player learns to predict the play path.

A player *learns to predict the play path* if her prediction of next period's play is asymptotically as good as if she knew her opponent's behavior strategy. If the behavior strategies call for randomization then players accurately predict the distribution over next period's play rather than the realization of next period's play. For example, consider a 2×2 game in which player 1 has stage game actions T and B and player 2 has stage game actions L and R. If player 2 is randomizing 50:50 every period and player 1 learns to predict the play path, then for every ε there is a time, which depends on the realization of player 2's strategy, after which player 1's next period forecast puts the probability of L within ε of 1/2. (This statement applies to a set of play paths that arises with probability one with respect to the underlying probability model; I gloss over this sort of complication both here and below.)

If LBR holds, and players are using their LBR strategies, then, asymptotically, the continuation play path is an approximate equilibrium play path of the continuation repeated game. The exact sense in which play converges to equilibrium play depends on the strength of learning and of optimization. See KL and also [66] (both of which focus on exact optimization) and [60] (which considers ε optimization).

KL, building on work in the probability literature on the merging of measures, notably Blackwell–Dubins [8], show that a strong form of LBR holds if beliefs satisfy an absolute continuity condition: each player assigns positive probability to any (measurable) set of play paths that has positive probability given the players' actual strategies. A strong sufficient condition for this is that each player assigns positive, even if extremely low, prior probability to her opponent's actual strategy, a condition that KL call *grain of truth*.

Universal Convergence

The work described thus far leaves open whether one can write down *any* deterministic learning model, even an implausible one, that exhibits universal convergence, that is, a learning model that, for a given stage game form (giving stage game actions but not payoff functions), gives convergence to Nash equilibrium play in the repeated game for all (or at least for generic) discount factors and stage game payoff functions.

Implicit in most work on this question is an assumption that the learning model is "uncoupled" in the sense of [37]: player i's prior over her opponent's repeated game strategies does not depend on the specification of her opponent's stage game payoffs. Cournot best response and fictitious play, as typically employed, are both examples of uncoupled learning models. One can get convergence in a "coupled" model simply by having the players play a Nash equilibrium of the repeated game. So, some degree of uncoupling is needed to avoid triviality. On the other hand, full uncoupling is a strong assumption from the perspective of sophisticated learning, since it effectively rules out introspective reasoning about ones opponent. And full uncoupling presumably makes convergence more difficult to achieve. Indeed, the main result of [37] is an impossibility theorem on convergence for certain classes of adaptive learning models.

For belief learning models, there are two competing intuitions on (uncoupled) universal convergence. One intuition is that universal convergence, even to ε equilibrium, is impossible because universal learnability is impossible: for any prior, there are opposing strategies that a player will fail to forecast, even approximately. One can show this via a diagonalization argument along the lines of [63] and it also follows as a corollary of results discussed in Sub-

sect. "Sophisticated Learning". In fact, for any given prior, the set of strategies one can learn to forecast is small, in a sense that can be made precise. A classic reference on the difficulty of prediction, in the context of general stochastic processes, is [20]. For a useful up to date survey, see the literature review in [1].

The competing intuition is that universal learnability is not necessary for convergence. The fact that both players are engaged in a game provides structure that conceivably, could force posteriors to be correct, at least along the path of play, even if priors are fundamentally wrong. Something like this happens in fictitious play. In the standard Bayesian interpretation of fictitious play (see Subsect. "Belief Learning and Bayesian Learning"), each player is certain that her opponent is playing an i. i. d. strategy even though neither is, because an i. i. d. strategy is not optimal. Yet, in many games, fictitious play generates behavior that looks asymptotically i. i. d. and play does converge to that of a Nash equilibrium.

Based on the first intuition and also on other negative results like those in [19] and [37], there has been a consensus that universal convergence is impossible for deterministic learning. Recently, however, Noguchi has reported a positive result [60]. Noguchi argues that, starting from any pair of priors, one can construct augmented priors with the property that (a) if a player could have learned to predict a strategy under the old prior, then she can also learn to predict that strategy under the new prior and (b), for ε optimization, a form of LBR holds, hence convergence to ε-equilibrium play obtains, where ε can be made arbitrarily small. One could take the original priors to be the fictitious play priors. Or one could be more ambitious and build various forms of pattern recognition into the original priors.

Sophisticated Learning

A number of papers investigate classes of prediction rules that exhibit desirable properties, such as the ability to detect certain kinds of patterns in opponent behavior. Important examples include [3,24,26,67].

In [57], I consider the issue of sophistication from a Bayesian perspective. For simplicity, focus on 2 player games. Fix a profile of priors and a subset of behavior strategies for each player and consider the following criteria for these strategy subsets.

Learnability For any strategy profile drawn from the strategy subsets, both players learn to prediction the play path.

Richness Informally, richness requires that if a behavior strategy is included in one of the strategy subsets then certain variations on that strategy must be included as well. Richness, called CSP in [57], is satisfied automatically if the strategy subsets consist of all strategies satisfying a standard complexity bound, the same bound for both players. Thus richness holds if the subsets consist of all strategies with k-period memory, or all strategies that are automaton implementable, or all strategies that are Turing implementable, and so on.

Consistency Each player's subset contains a best response to her belief.

The motivating idea is that, for priors to be considered sophisticated, a necessary (but not sufficient) condition is that the priors can be interpreted as probability distributions over strategy subsets satisfying these criteria. [57] studies whether any such priors exist.

Consider, for example, the Bayesian interpretation of fictitious play in which priors are probability distributions over the i. i. d. strategies. The set of i. i. d. strategies satisfies learnability and richness. But for any stage game in which neither player has a weakly dominant action, the i. i. d. strategies violate consistency: any player who is optimizing will not be playing i. i. d.

As shown in [57], this feature of Bayesian fictitious play extends to all Bayesian learning models. For large classes of repeated games, for *any* profile of priors and for *any* strategy subsets satisfying richness, if learnability holds then consistency fails.

Let me make a few remarks. First, since the set of all strategies satisfies richness and consistency, it follows that for any profile of priors there is a strategy profile that the players will not learn to predict. This can also be shown directly by a diagonalization argument along the lines of [13] and [63]. The impossibility result of [57] can be viewed as a game theoretic version of [13]. For a description of what subsets *are* learnable, see [59].

Second, suppose that the strategy subsets are generated by some standard definition of complexity, the same for both players. Then, as noted above, richness holds. Suppose further that there are priors for which learnability holds. This will be the case, for example, for Turing implementable strategies, since the set of such strategies is countable. Then, for such priors, for large classes of repeated games, consistency fails: best responses must violate the complexity bound.

Third, if one constructs a Bayesian learning model satisfying learnability and consistency then LBR (see Subsect. "Kalai–Lehrer Learning") holds and, if players play their LBR strategies, play converges to equilibrium play. This identifies a potentially attractive class of Bayesian

models in which convergence obtains. The impossibility result says, however, that if learnability and consistency hold, then player beliefs must be partially equilibrated in the sense of, in effect, excluding some of the strategies required by richness.

Fourth, the theorem is robust. It holds under ε optimization, for ε small. It holds even for fairly weak definitions of prediction (definitions that allow occasional but persistent forecasting errors, for example). And while learnability, as defined, requires prediction for any strategy profile drawn from the strategy subsets, this is only for ease of exposition. A variation of the result says that if the subsets are rich then for any strategy profile for which prediction holds, either that profile is not (ε) optimal or it is not contained in the strategy subsets.

Last, consistency is not *necessary* for convergence. See Subsect. "Universal Convergence". The impossibility result is a statement about the ability to construct Bayesian models with certain properties; it is not a statement about convergence to equilibrium per se.

Payoff Uncertainty

Suppose that, at the start of the repeated game, each player is privately informed of his or her stage game payoff function, which remains fixed throughout the course of the repeated game. Refer to player i's stage game payoff function as her *payoff type*. Assume that the joint distribution over payoff functions is independent (to avoid correlation issues that are not central to my discussion) and commonly known.

Each player can condition her behavior strategy in the repeated game on her realized payoff type. A mathematically correct way of representing this conditioning is via distributional strategies; see [54].

For any prior about player 2, now a probability distribution over player 2's distributional strategies, and given the probability distribution over player 2's payoff types, there is a behavior strategy for player 2 in the repeated game that is equivalent in the sense that it generates the same distribution over play paths. Again, this is essentially Kuhn's theorem.

Say that a player *learns to predict the play path* if her forecast of next period's play is asymptotically as good as if she knew the reduced form of her opponent's distributional strategy. This definition specializes to the previous one if the distribution over types is degenerate. If distributional strategies are in Nash equilibrium (also known in this context as a Bayesian Nash equilibrium), then, in effect, each player is optimizing with respect to a degenerate belief that puts probability one on her opponent's actual distributional strategy, and in this case players trivially learn to predict the path of play.

One can define LBR (see Subsect. "Kalai–Lehrer Learning") for distributional strategies and, as in the payoff certainty case, one can show that LBR implies convergence to Nash equilibrium play in the repeated game with payoff types.

More interestingly, Nash equilibrium play in the repeated game with payoff types implies convergence to Nash equilibrium play of the *realized* repeated game – the repeated game determined by the realized type profile. This line of research was initiated by Jordan in [42]. Other important papers include KL [45] (see Subsect. "Kalai–Lehrer Learning"), [41,44], and [62] (which studies recurring rather than repeated games).

Suppose first that the realized type profile has positive probability. In this case, if a player learns to predict the play path then, as shown by KL, her forecast is asymptotically as good as if she knew both her opponent's distributional strategy *and* her opponent's realized type. LBR then implies that actual play, meaning the distribution over play paths generated by the realized behavior strategies, converges to equilibrium play of the realized repeated game. For example, suppose that the type profile for matching pennies gets positive probability. In the unique equilibrium of repeated matching pennies, players randomize 50:50 in every period. Therefore, LBR implies that if the matching pennies type profile is realized then each player's behavior strategy in the realized repeated game involves 50:50 randomization asymptotically.

If the distribution over types admits a continuous density, so that no type profile receives positive probability, then convergence is more complicated. Suppose that players are myopic and that the realized stage game is like matching pennies, with a unique and fully mixed equilibrium. Given myopia, the unique equilibrium of the realized repeated game calls for repeated play of the stage game equilibrium. In particular, it calls for players to randomize. It is not hard to show, however, that in a type space game with a continuous density, exact optimization calls for each player to play a pure strategy for almost every realized type (this is a generalization of a point made in the context of fictitious play in Subsect. "Convergence in Classical Learning Models"). Thus, for almost every realized type profile in a neighborhood of a game like matching pennies, actual play (again meaning the distribution over play paths generated by the realized behavior strategies) cannot converge to Nash equilibrium play of the realized repeated game, *even if the distributional strategies are in Nash equilibrium*. Foster and Young [18] provide a generalization for non-myopic players.

This impossibility result is not robust to weakening optimization to ε optimization; the positive convergence results for ε-optimizing fictitious play provide one illustration, see Subsect. "Convergence in Classical Learning Models". A more subtle point, however, is that, even for exact optimization, a form of convergence obtains that, while weaker than convergence of actual play, is still very strong.

For simplicity, assume that each player knows the others distributional strategy and that these strategies form a (Bayesian) Nash equilibrium. Then to an outsider, for almost any type profile, observed play looks asymptotically like Nash equilibrium play in the realized repeated game (this follows from the main theorem in [62]). In particular, in a neighborhood of a game like matching pennies, for almost any type profile, observed play looks random. Since, by [18], actual play in this setting cannot converge to equilibrium, and in particular cannot be random, the implication is that convergence to equilibrium involves a form of purification in the sense of [34], a point that has been emphasized by [41] and [62]. To a player in the game, opponent behavior likewise looks random because, even though she knows her opponent's distributional strategy, she does not know her opponent's type. As play proceeds, each player in effect learns more about her opponent's type, but never enough to zero in on her opponent's realized, pure, behavior strategy.

Finally, the difficulties with characterizing sophistication in Bayesian learning models extend to models with payoff uncertainty, with learnability, richness, and consistency redefined in terms of distributional strategies; see [56].

Stochastic Learning

For a concrete example of a stochastic prediction rule, consider Young's variation on fictitious play [72]. In standard fictitious play, a player gets to observe the entire history of the game to date. In a bounded memory version of fictitious play, a player gets to observe only the last, say, s periods. In Young's variation on bounded memory fictitious play, a player observes only an $\ell < s$-period sample of the last s periods, with the sample drawn according to a probability distribution that assigns positive probability to every possible sample of length ℓ. If the player's forecast about period $t+1$ depends on her information in period t, then her prediction rule will be stochastic: her forecast will depend on the realization of the date t sample. With a stochastic prediction rule, a player at the beginning of the game cannot compute her own future forecast as a function of history.

A range of stochastic learning models have been studied, nearly all for myopic optimization. In some cases (e. g. [19]), the stochastic prediction rules have a quasi-Bayesian interpretation: most of the time, players optimize with respect to fixed prediction rules, as in a Bayesian model, but occasionally players switch to new prediction rules, implicitly abandoning their priors.

In this last part of the survey, I discuss two of the main themes in the stochastic learning literature. One is familiar: convergence to equilibrium. The other is new: calibration.

Convergence in Stochastic Learning Models

Although details can differ substantially from model to model, convergence results for stochastic learning models often have a similar feel. The stochastic nature of the prediction rule introduces an element of randomness into behavior. This randomness can cause players to stumble upon close-to-equilibrium play in the stage game and players, once near equilibrium, tend to stay near equilibrium, if not forever then at least for relatively long periods of time.

One thread of the literature on convergence in stochastic learning models poses specific models and investigates what sort of solution concepts emerge. In the case of Young's stochastic fictitious play model [72], the stochastic dynamics serve as a selection mechanism: learning tends to concentrate on some stage game Nash equilibrium but not others, even distinguishing between multiple strict Nash equilibrium, a phenomenon that also appears in many stochastic evolutionary models. In other stochastic learning models, convergence is sometimes not to Nash equilibrium but instead to sets containing Nash equilibria, such as CURB sets or minimal prep sets; see [40,49], and [64].

A second thread of the literature asks whether it is possible to construct *any* stochastic learning model that yields universal convergence to Nash equilibrium. The focus on stochastic prediction rules was motivated in part by the failure to find universal convergence results using deterministic prediction rules; see also Subsect. "Universal Convergence". The seminal paper on global convergence with stochastic prediction rules is [19]. For more recent work, see [31] and [38].

Calibration

In [17], Foster and Vohra showed that if one allows stochastic prediction rules then one can get approximate calibration on *every* realized history. That is, no matter what strategy the opponent plays, the prediction rule will approximately satisfy a statistical condition called calibra-

tion – and one can make the approximation to exact calibration as close as one likes. Calibration, in turn, implies a weak form of convergence to correlated equilibrium; see [16].

Calibration is a weak condition, so weak that it is possible for a prediction rule to be calibrated without it being accurate in any reasonable sense; see [47]. Nevertheless, calibration is sufficiently strong that a simple diagonalization argument shows that no deterministic prediction rule can satisfy it even approximately [63]. Foster and Vohra's discovery that one can get approximate calibration using stochastic prediction was startling.

Foster and Vohra's calibration result has spawned another literature, distinct from its application to learning, on whether it is possible for an observer to distinguish between an expert (someone who knows the true form of some stochastic process) and a charlatan (someone who uses the Foster and Vohra prediction rule, or one of its more elaborate cousins, to pass the observer's tests). Papers in this currently very active area of current research include [2,14,15,51,68].

Future Directions

Directions for future work, both for the learning literature and, more broadly, for the general literature on out-of-equilibrium dynamics in games, include the following.

Benchmark learning One of the goals of the literature is to establish a benchmark model of sophisticated learning. [57] sheds doubt on whether this is possible, but the desiderata considered in [57] are not based on axiomatic, decision theoretic criteria. The question of whether it is possible to build an axiomatic theory of sophisticated learning is unexplored.

Environments This survey has focused on repeated two-player games, but many other learning environments and patterns of interaction are possible. One can, for example, study environments in which players are linked through a network as in [9], or environments in which players are behaviorally heterogeneous. The possibilities are endless. The task is to enrich the scope of research without simply generating an ever expanding catalog of models, which leads to . . . .

Empirical testing One objective of the learning literature is to understand how real people behave in real games. Work tying learning models to actual behavior is underway [11] but still in its infancy. The appropriateness of a model may depend on the environment and in at least in some cases behavioral heterogeneity may be relevant.

Application An understanding of how players behave outside of equilibrium will enrich the application of game theoretic models. One example concerns institutional design, an instance of which is the choice of game for selling goods (a particular type of auction, for instance). Different institutions may have similar Nash equilibria but very different dynamic properties, and the dynamic properties could therefore play a role in the choice of institution.

Bibliography

1. Al-Najjar N (2007) Decision makers as statisticians. Department of Managerial Economics and Decision Sciences, Kellogg School of Management, Northwestern University, Evanston
2. Al-Najjar N, Weinstein J (2008) Comparative testing of experts. Econometrica 76(3):541–559
3. Aoyagi M (1996) Evolution of beliefs and the Nash equilibrium of normal form games. J Econ Theory 70:444–469
4. Aumann R (1964) Mixed and behaviour strategies in infinite extensive games. In: Dresher M, Shapley LS, Tucker AW (eds) Advances in game theory. Annals of Mathematics Studies, vol 52. Princeton University Press, Princeton, pp 627–650
5. Basu K, Weibull AW (1991) Strategy subsets closed under rational behavior. Econ Lett 36:141–146
6. Benaim M, Hirsch M (1999) Mixed equilibria arising from fictitious play in perturbed games. Games Econom Behav 29: 36–72
7. Bernheim BD (1984) Rationalizable strategic behavior. Econometrica 52(4):1007–1028
8. Blackwell D, Dubins L (1962) Merging of opinions with increasing information. Ann Math Stat 38:882–886
9. Blume LE (1993) The statistical mechanics of strategic interaction. Game Econ Behav 5(3):387–424
10. Brown GW (1951) Iterative solutions of games by fictitious play. In: Koopmans TJ (ed) Activity analysis of production and allocation. John Wiley, New York, pp 374–376
11. Camerer C (2008) Behavioral game theory. In: Blume LE, Durlauf SN (eds) The new Palgrave dictionary of eocnomics, 2nd edn. MacMillan, New York
12. Cournot A (1838) Researches into the mathematical principles of the theory of wealth. Kelley, New York; translation from the French: Bacon NT (1960)
13. Dawid AP (1985) The impossibility of inductive inference. J Am Stat Assoc 80(390):340–341
14. Dekel E, Feinberg Y (2006) Non-Bayesian testing of an expert. Rev Econ Stud 73:893–906
15. Feinberg Y, Stewart C (2008) Testing multiple forecasters. Econometrica 76(3):561–582
16. Foster D, Vohra R (1997) Calibrated learning and correlated equilibrium. Game Econ Behav 21:40–55
17. Foster D, Vohra R (1998) Asymptotic calibration. Biometrika 85:379–390
18. Foster D, Young P (2001) On the impossibility of predicting the behavior of rational agents. Proc Natl Acad Sci 98: 12848–12853
19. Foster D, Young P (2003) Learning, hypothesis testing, and Nash equilibrium. Game Econ Behav 45:73–96

20. Freedman D (1965) On the asymptotic behavior of Bayes estimates in the discrete case ii. Ann Math Stat 36:454–456
21. Fudenberg D, Kreps D (1993) Learning mixed equilibria. Game Econ Behav 5(3):320–367
22. Fudenberg D, Levine DK (1993) Self-confirming equilibrium. Econometrica 61(3):523–545
23. Fudenberg D, Levine DK (1993) Steady state learning and Nash equilibrium. Econometrica 61(3):547–574
24. Fudenberg D, Levine DK (1995) Universal consistency and cautious fictitious play. J Econ Dyn Control 19:1065–1089
25. Fudenberg D, Levine DK (1998) Theory of learning in games. MIT Press, Cambridge
26. Fudenberg D, Levine DK (1999) Conditional universal consistency. Game Econ Behav 29:104–130
27. Fudenberg D, Levine DK (2006) Superstition and rational learning. Am Econ Rev 96:630–651
28. Fudenberg D, Levine DK (2008) Learning and equilibrium. Harvard University, Cambridge
29. Fudenberg D, Takahashi S (2007) Heterogeneous beliefs and local information in stochastic fictitious play. Harvard University, Cambridge
30. Fudenberg D, Tirole J (1991) Game theory. MIT Press, Cambridge
31. Germano F, Lugosi G (2007) Global Nash convergence of Foster and Young's regret testing. Game Econ Behav 60:135–154
32. Goeree JK, Holt CA (2001) Ten little treasures of game theory and ten intuitive contradictions. Am Econ Rev 91:1402–1422
33. Hahn F (1977) Exercises in conjectural equilibrium. Scand J Econ 79:210–226
34. Harsanyi J (1973) Games with randomly disturbed payoffs: A new rationale for mixed-strategy eequilibrium points. Int J Game Theory 2:1–23
35. Hart S, Mas-Colell A (2000) A simple adaptive procedure leading to correlated equilibrium. Econometrica 68(5):1127–1150
36. Hart S, Mas-Colell A (2001) A general class of adaptive strategies. J Econ Theory 98:26–54
37. Hart S, Mas-Colell A (2003) Uncoupled dynamics do not lead to Nash equilibrium. Am Econ Rev 93:1830–1836
38. Hart S, Mas-Colell A (2006) Stochastic uncoupled dynamics and Nash equilibrium. Game Econ Behav 57:286–303
39. Hofbauer J, Sandholm W (2002) On the global convergence of stochastic fictitious play. Econometrica 70(6):2265–2294
40. Hurkens S (1995) Learning by forgetful players. Game Econ Behav 11:304–329
41. Jackson M, Kalai E (1999) False reputation in a society of players. J Econ Theory 88(1):40–59
42. Jordan JS (1991) Bayesian learning in normal form games. Game Econ Behav 3:60–81
43. Jordan JS (1993) Three problems in learning mixed-strategy Nash equilibria. Game Econ Behav 5(3):368–386
44. Jordan JS (1995) Bayesian learning in repeated games. Game Econ Behav 9:8–20
45. Kalai E, Lehrer E (1993) Rational learning leads to Nash equilibrium. Econometrica 61(5):1019–1045
46. Kalai E, Lehrer E (1993) Subjective equilibrium in repeated games. Econometrica 61(5):1231–1240
47. Kalai E, Lehrer E, Smorodinsky R (1999) Calibrated forecasting and merging. Game Econ Behav 18(1–2):151–169
48. Kalai E, Samet D (1984) Persistent equilibria in strategy form games. Int J Game Theory 13:129–144
49. Kets W, Voorneveld M (2005) Learning to be prepared. Tilburg University, Tilburg
50. Kuhn HW (1964) Extensive games and the problem of information. In: Dresher M, Shapley LS, Tucker AW (eds) Contributions to the theory of games, vol II. Annals of Mathematics Studies, vol 28. Princeton University Press, Princeton, pp 193–216
51. Lehrer E (2001) Any inspection is manipulable. Econometrica 69(5):1333–1347
52. McKelvey R, Palfrey T (1995) Quantal response equilibrium for normal form games. Game Econ Behav 10:6–38
53. Milgrom P, Roberts J (1991) Adaptive and sophisticated learning in repeated normal form games. Game Econ Behav 3: 82–100
54. Milgrom P, Weber R (1985) Distributional strategies for games with incomplete information. Math Oper Res 10:619–632
55. Moulin H (1984) Dominance-solvability and cournot stability. Math Soc Sci 7(1):83–102
56. Nachbar JH (2001) Bayesian learning in repeated games of incomplete information. Soc Choice Welf 18(2):303–326
57. Nachbar JH (2005) Beliefs in repeated games. Econometrica 73:459–480
58. Nachbar JH (2008) Learning and evolution in games: Belief learning. In: Blume LE, Durlauf SN (eds) The new Palgrave dictionary of eocnomics, 2nd edn. Palgrave MacMillan, New York
59. Noguchi Y (2005) Merging with a set of probability measures: A characterization. Kanto Gakuin University, Tokyo
60. Noguchi Y (2007) Bayesian learning with bounded rationality: Convergence to ε-Nash equilibrium. Kanto Gakuin University, Tokyo
61. Nyarko Y (1994) Bayesian learning leads to correlated equilibria in normal form games. Econ Theory 4:821–841
62. Nyarko Y (1998) Bayesian learning and convergence to Nash equilibria without common priors. Econ Theory 11(3):643–655
63. Oakes D (1985) Self-calibrating priors do not exist. J Am Stat Assoc 80(390):339
64. Sanchirico CW (1996) A probabilisitc model of learning in games. Econometrica 64(6):1375–1393
65. Sandholm W (2007) Evolution in Bayesian games ii: Stability of purified equilibrium. J Econ Theory 136:641–667
66. Sandroni A (1998) Necessary and sufficient conditions for convergence to Nash equilibrium: The almost absolute continuity hypothesis. Game Econ Behav 22:121–147
67. Sandroni A (2000) Reciprocity and cooperation in repeated coordination games: The principled-player approach. Game Econ Behav 32(2):157–182
68. Sandroni A (2003) The reproducible properties of correct forecasts. Int J Game Theory 32(1):151–159
69. Shapley L (1962) On the nonconvergence of fictitious play. Tech Rep RM-3026 RAND
70. Voorneveld M (2004) Persistent retracts and preparation. Game Econ Behav 51:228–232
71. Voorneveld M (2004) Preparation. Game Econ Behav 48: 403–414
72. Young PH (1993) The evolution of conventions. Econometrica 61(1):57–84
73. Young PH (2008) Adaptive heuristics. In: Blume LE, Durlauf SN (eds) The new Palgrave dictionary of eocnomics, 2nd edn. Palgrave MacMillan, New York
74. Young PH (2008) Stochastic adaptive dynamics. In: Blume LE, Durlauf SN (eds) The new Palgrave dictionary of eocnomics, 2nd edn. Palgrave MacMillan, New York

Learning and Planning (Intelligent Systems)

Ugur Kuter
Institute of Advanced Computer Studies and Department of Computer Science, University of Maryland, College Park, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Artificial Intelligence Planning, Traditionally
Machine Learning Methods in Planning
Learning Domain Models
Learning Search-Control Knowledge
Learning Operational Planning Procedures
Reinforcement Learning
Summary and Future Directions
Acknowledgments
Bibliography

Glossary

Automated planning Automated Planning is the problem of generating a sequence of actions for an initial configuration of the world that, when executed, produces a final configuration that satisfies a specified set of goal conditions.

Planning knowledge Almost all of successful AI planning systems developed in recent years use some sort of *search-control knowledge* for effective planning. Sometimes the planning knowledge is specified in terms of *domain-independent heuristics*, i. e., heuristics intended for use in many different planning domains, and sometimes they are specified in terms of *domain-specific control knowledge* (i. e., tailored to a specific problem domain).

Learning for problem-solving and planning Learning for planning is the process of acquiring auxiliary knowledge related to a planning problem, which can be used by a planning system to understand better the underlying planning domain and to control its search for generating plans in that domain.

Definition of the Subject

The ability to produce knowledge about past experiences and exploit that knowledge in an operational context in later problem-solving and planning sessions is an important attribute of any intelligent system, human- or AI-based alike. Automated Planning and Learning is the research paradigm that focuses on the development of intelligent systems and technologies that combine the ability to make decisions and generate courses of actions (i. e., plans) with the capability to reason and produce knowledge about past experiences, future problems that the system needs to tackle, and strategies about how to tackle them.

Probably the first work that laid a formal treatment for this combination is the early planning system STRIPS [17], developed in early 1970s. The STRIPS planning system was an evidence that planning and learning are usually two pieces of an intelligent system, where the knowledge acquired via learning is used to enhance the problem-solving and planning experience later. Following its vision, many approaches have been developed in order to bring the techniques from the machine-learning community into planning. Most notably, *explanation-based learning (EBL)* and *inductive learning* techniques have been the most popular to contribute to AI planning.

The common objective of using learning for planning is to develop and use automated techniques to produce some sort of planning knowledge that is used to improve the performance of a planner. Sometimes the learned knowledge describes search strategies for "how to plan" (i. e., "what a planner should do in order to generate solutions"), and thus, guides the planner to focus only on the relevant parts of the planning problem and generate correct plans effectively. Sometimes the learned knowledge specifies the potential failures a planner could encounter during its search and their explanations. The planner uses such failure knowledge in order not to repeat similar mistakes over and over again. Other times, the planning knowledge is about the dynamics of the underlying environment; e. g., knowledge about the possible unexpected effects of the actions when executed in different circumstances. In all cases, the learned knowledge serves as a guidance to a planner in generating solutions to the planning problems.

Perhaps the biggest obstacle for developing planning and learning systems has been the so-called *utility problem*; that is, the trade-off between the time and resources spent for learning a sound and complete body of knowledge for planning and the cost of exploiting that knowledge during planning, which sometimes is shown to overshadow the effort spent in planning itself. Because of this problem, regrettably, the most important and remarkable developments in automated planning in recent years did not involve any learning component; instead, recent planning approaches include very efficient planning techniques that use controlled search with domain-specific

and/or domain-independent heuristics, constraint-satisfaction techniques for reasoning with time and resources, and model-checking based planning algorithms.

A key challenge for most of the recent planning approaches is that they require a domain expert to provide some sort of planning knowledge to the system, thus, replacing the functionality of a learning component by expert-level knowledge given as input to the planner. In many realistic planning problems, however, such planning knowledge may not be completely available; this is partly because it is very hard to compile such knowledge due to the complexities in the domains, e. g., evacuation and rescue operations, and it is partly because there is no expert to provide it, e. g., space operations. As the practicality of the planning, problem-solving and decision-making systems has been expanded to such complex, interesting, real-world challenges over the years, however, the fellowship of planning and learning has become once again important for the development of such systems.

Introduction

Traditional Artificial Intelligence (AI) planning, also known as *classical planning*, is the problem of generating a sequence of actions for an initial configuration of the world that, when executed, produces a final configuration that satisfies a specified set of goal conditions. The mathematical formulation of classical planning assumes that the environment must contain finitely many objects, and configurations of those objects describe the states of the environment. A classical planner always knows what is true and what is false in a state of the world. Planner's actions have deterministic outcomes; i. e., when executed, an action has a single and an instantaneous effect on the state of the world. Furthermore, the planner's actions are the only cause of change in the world. Thus, the world evolves in discrete and deterministic time steps when a plan (i. e., a sequence of deterministic actions) is executed.

It has been both theoretically and experimentally demonstrated that classical planning is a hard problem [14]. Over the years, some of the most impressive recent advances in classical AI planning are based on the use of "planning knowledge" for organizing the search space and guiding the search during planning. Sometimes the planning knowledge is specified in terms of *domain-independent heuristics*, i. e., heuristics intended for use in many different planning domains, and sometimes they are specified in terms of *domain-specific control knowledge* (i. e., tailored to a specific problem domain). In any case, this is usually in the form of auxiliary knowledge about the dynamics of the planning domain (i. e., semantic descriptions of possible actions and changes in the world), certain properties of the objectives of possible planning problems in that domain, and how to guide the planning process to solve those planning problems.

A key challenge for most of the successful planners is the requirement of a domain expert to provide the sort of background planning knowledge mentioned above. In many realistic planning domains, however, such planning knowledge may not be completely available; this is partly because it is very hard for the experts to compile such knowledge due to the complexities in the domains and it is partly because there is limited access to an expert to provide it. Here are several examples of such real-world planning domains and the currently-deployed planning techniques/systems in those domains:

- **Space Operations** Probably the most important success story of AI and AI planning technologies is their use in space operations. MAPGEN, the Mixed-initiative Activity Plan Generation tool, of NASA is a crucial part of the Mars Exploration Rover (MER) ground planning systems. The motivation for the MAPGEN system is to better support users that need to rapidly build activity plans that have to satisfy complex rules and meet the resource limits [2]. MAPGEN's planning component is based on the paradigm of constraint-based planning, where activities and states are described by predicate statements that hold over temporal intervals. The interval time-points and the predicate parameters are represented by variables connected by constraints. This approach supports a variety of complex planning constructs, including: activities with temporal durations, states that expire, exogenous events, complex constraints on parameters, temporal constraints linking activities and states, and subgoaling rules with conditions and disjunctions.
  Currently, MAPGEN uses mixed-initiative techniques (i. e., human-in-the-loop during the planning process) and automated reasoning methods in order to acquire its constraint-based planning knowledge. For the times the human expert is not available, learning such safety and operational constraints would be important and useful complement of the existing planning technologies in MAPGEN, since a planning system that can learn such knowledge to develop ways on how to safely and correctly operate in the world holds great promise for success.
- **Rescue and Evacuation Operations** Rescue and Evacuation operations [6,37,38] is the task of planning how to evacuate groups of people who may be in danger. Generally in these situations, there will be an incom-

plete description of the environment, in the form of standard requirements and operating procedures. Formulating an evacuation plan can be quite complex: typically there will be hundreds of tasks to be carried out. These tasks will depend on a wide range of factors: sources of danger, available resources, geography, weather predictions, political issues, and so forth. Complete information about the current state will never be available; the planning must include dynamic information gathering, and plans must be formulated with an incomplete world state. For such a problem, the planning must be done by a human expert or under the supervision of a human expert. It is unrealistic to expect that a planning system could produce good plans by itself, and awed evacuation plans could yield dire consequences.

Rescue and evacuation plans often require detailed analysis and knowledge about previous experiences. Learning lessons from past experiences of an evacuation planning system and from the information gathered from external knowledge sources will be crucial, especially when the human experts are not accessible and when the evacuation problem is too complex to be solved and managed by humans alone.

- **Vehicle Movement Planning** Planning the movements of vehicles, such as ships in naval operations and Unmanned Aerial Vehicles (UAVs) in surveillance missions, is important both commercially and militarily [11,19]. The state space and action space are effectively infinite: states include positions and velocities of vehicles, and actions correspond to movements of the vehicles along various routes. Since movements of different vehicles may occur concurrently, it is important to make sure they do not interfere with each other. The outcomes and durations of the actions cannot be known with certainty, because of factors such as weather, currents, and the behavior of the vehicles' operators. Elaborate simulation tools are available to aid in planning ship movements but the planning is still mostly done manually [1].

All of the domains above require some sort of planning knowledge to be used during the planning and execution process, to understand and exploit the underlying dynamics and properties of the environment in which the planning is taking place, and the ability to use past experiences in order to guide future decisions in solving problems. Over the years, several different machine-learning paradigms have been used to support automated planning. The common objective of using learning for planning is to develop and use automated techniques to learn some knowledge that is used to improve the performance of a planner. Many different techniques have been developed with this objective, including *learning macro-operators*, e. g., [5,36], *learning search control knowledge* [16,32,34], *learning of task hierarchies* [8,42,44] and *learning plan abstraction* [3,26].

Using explanation-based learning techniques have been investigated in mid-90s [25] in order to speed up the planning process where there is no planning knowledge available to give to the planner. These techniques mainly focused on observing the failures that occur during planning and generalize them into conditions that the planner should avoid during its search. These generalized conditions are then used for search-space pruning in the later stages of the planning process. Existing works on learning macro-operators (e. g., [5,36]) also falls in the category of speed-up learning, as do work on learning search control knowledge (e. g., [16,32,35]).

Recently, several architectures have been proposed to learn hierarchical planning knowledge from a collection of plan traces and from a given action model [8,42,44]. [42]'s X-Learn, for example, uses inductive generalization to learn task decomposition constructs, which relate subgoals and conditions for decomposing those goals into their subgoals. By grouping goals in this way, task models are learned that lead to significant speed-ups in problem-solving. As another example, [8,40] achieves the same objective via problem-solving techniques similar to explanation-based learning. Two other recent studies [23,51] proposed algorithms to learn the applicability conditions of task-decomposition constructs, given the hierarchical relationships between the tasks.

This paper describes some basic concepts in automated planning systems where machine learning has been of assistance in order to produce planning knowledge. We start with a summary of traditional automated planning paradigm of AI and an overview of analytical machine-learning methods that have been deemed suitable for integration with planning systems. Then, we present an analysis of four different forms of approaches that have integrated planning and learning techniques to achieve different objectives. We conclude with our remarks on the lessons learned in the Planning and Learning research so far and with some possible future directions.

Artificial Intelligence Planning, Traditionally

Traditional Artificial Intelligence (AI) planning, also known as *classical planning*, is the problem of generating a sequence of actions that, when executed in an initial configuration of the world, produces a final configu-

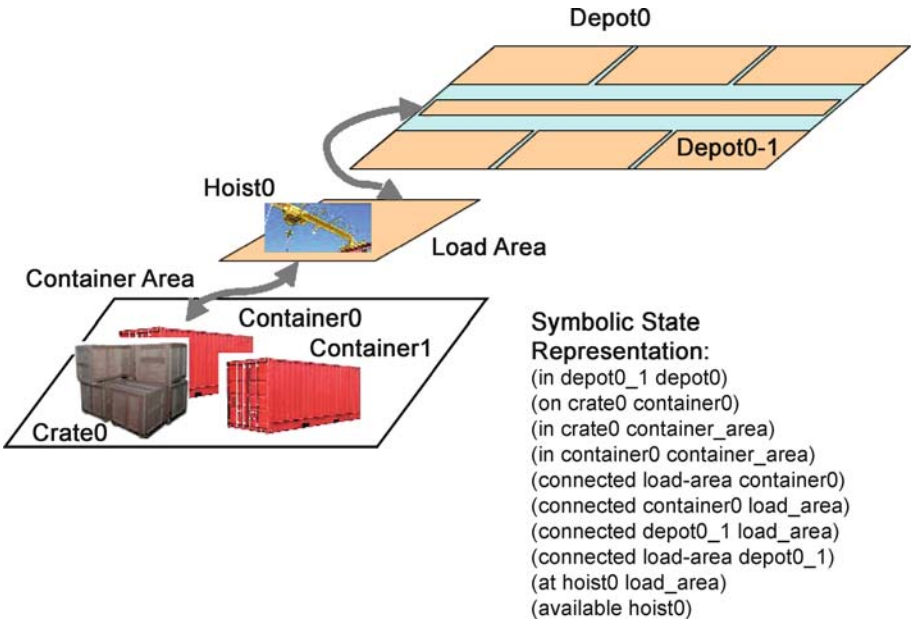

Learning and Planning (Intelligent Systems), Figure 1
A state description in the Simple-Storage domain

ration that satisfies a specified set of goal conditions. The mathematical formulation of classical planning assumes that the environment must contain finitely many objects, and configurations of those objects describe the states of the environment. More formally, a *state* is a collection of logical ground atoms. Classical planners usually make the *closed-world assumption*: the set of logical ground atoms in a state describe the facts that are true about the environment, whereas any logical atom that does not appear in the state is assumed to be false. Thus, a classical planner always knows what is true and what is false in a state of the world.

As an example, consider a simple transportation planning domain, a simple variant of a benchmark planning domain from the International Planning Competition (IPC) at 2006 [18], that we are going to use for illustration purposes here and throughout this paper. The original domain at the IPC is called *Storage*, and we will call our version as *Simple-Storage* in this paper. This domain involves moving a certain number of crates from some containers to some depots by a hoist. Each crate has a weight known to the planner. Each depot area can contain a number of crates in stacks, provided that in a stack, a heavier crate must always be placed lower than a lighter one. Inside a depot, each hoist can move according to a specified spatial map connecting different areas of the depot. Once the hoist is holding a crate for delivery and moved into a depot, it cannot move outside of the depot without dropping the crate.

Figure 1 illustrates a snapshot state of the world in this domain. This state specifies a depot and several storage areas in it, a hoist, a crate, two containers, and the spatial map of the depot areas. In particular, the state describes what is true in the world at this instance; e. g., the *crate*0 is in the *container*0, there is an empty slot in *depot*0 at area *depot*0–1, the *hoist* is empty and available for use in the loading area, and so on. It also specifies information about what is not true in the world; e. g., we know that *crate*0 is not in *container*1, we know that the hoist is not in any depot or container area, and so forth.

A classical planner's actions have always deterministic outcomes; i. e., when executed, an action has a single and an instantaneous effect on the state of the world. Actions are specified via an *action schema*, or equivalently, through a set of *planning operators*. A planning operator is an expression of the following form (*head, preconditions, effects*), where *head* is an expression of the form $(name, arg_1, \ldots, arg_k)$ such that *name* is a symbol denoting the name of the operator and $arg_1, \ldots, arg_k$ are the *arguments* of the planning operator: each argument a_i is either a variable or a constant symbol. The *preconditions* and *effects* of the operator are specified as logical formulas that can contain both positive and negative literals. A negative literal in the effects of the operator specifies that the literal will become false in the state of the world after the operator is applied (i. e., the atom will be deleted from the state). A positive literal in the effects becomes true in the world after the operator is applied.

For example in our *Simple-Storage* domain, we will specify five different planning operators for this planning domain: a pick-up operator for lifting a crate by a hoist, a put-down operator for dropping a crate by a hoist, a move operator for moving a hoist inside and outside a depot, and a locate operator for moving a hoist from one area of a depot to another one. The following is an abstract description of the pick-up operator in a Lisp-like notation:

```
operator for pick-up
  :parameters (?c – crate ?a1 – area ?a2 – area ?p – place)
  :precondition (and (connected ?a1 ?a2) (at ?h ?a2)
   (available ?h)(on ?c ?a1) (in ?a1 ?p))
  :effect (and (not (on ?c ?a1)) (clear ?a1)
   (not (available ?h))(holding ?h ?c) (not (in ?c ?p))) .
```

An *action* is a ground instance of a planning operator, where all of the variable symbols that appear in the operator definition are replaced by constant symbols. We will denote the preconditions and effects of an action a by $\mathsf{Pre}(a)$ and $\mathsf{Eff}(a)$, respectively.

An action a is applicable in a state s, if the preconditions of a are satisfied in s; i. e., if $s \models \mathrm{pre}(a)$. The *result* of applying an action a in a state s, $\mathrm{result}(s, a)$, is the state s' that is produced by removing every negative literal in $\mathsf{Eff}(a)$ from s and by adding every positive literal in $\mathsf{Eff}(a)$. A *plan* is a sequence (i. e., totally-ordered set) of actions. The *result* of applying a plan $\langle a_1, \ldots, a_k\rangle$ in a state s is the state $s' = \mathrm{result}(\mathrm{result}(\ldots, \mathrm{result}(\mathrm{result}(s, a_1), a_2), \ldots), a_k)$. We denote this by $s' = \mathrm{result}(s, \langle a_1, \ldots, a_k\rangle)$ for simplicity of the notation.

As the reader may have noticed already, the planner's actions are the only cause of change in the world dynamics of classical planning. Also, it follows from the above definitions that classical planning assumes that the world evolves only in discrete and deterministic time steps when a plan (i. e., a sequence of deterministic actions) is executed.

A *classical planning problem* is defined as a tuple (s_0, G, O) where s_0 is the initial state (a symbolic representation of the state of the world at the time that the plan executor will begin executing its plan), G is a set of goal states (a symbolic representation of the desired state or states the executor will end executing the plan), and O is the set of planning operators. A *solution* to a classical planning problem (s_0, G, O) is a plan $\langle a_1, \ldots, a_k\rangle$ such that $\mathrm{result}(s_0, \langle a_1, \ldots, a_k\rangle) \in G$.

Solution methods for solving classical planning problems are categorized as *state-space search* and *plan-space search* techniques:

State-Space Planning State-space search techniques perform a search over the state space. If the search is done in a forward-chaining fashion, then the planner starts from the initial state of the planning problem, and it successively generates new states by applying actions, until a goal state is reached. In backward-chaining state-space search, the planner starts from a goal state and applies successively the inverse of the actions until it reaches to the initial state.

Plan-Space Planning In plan-space planning, the planner starts with the initial plan, and successively adds new actions and fixes the flaws in the plan, until no flaw is left and the plan is a solution to the current planning problem. The initial plan consists of only two special actions, start and finish. The start action does not have any preconditions and its effects encode the initial state of the planning problem. The finish action does not have any effects and its preconditions specify a logical condition that encodes the goals of the planning problem. Any precondition that is not established in a partial plan is called a flaw, and the planner successively enhances the initial plan by adding new actions whose effects establish the preconditions of the existing actions in the plan. This process continues until no flaws left in the plan, which then encodes the solution of the input planning problem.

Several variants of state-space and plan-space search paradigms have been developed in automated planning over the years. A key issue with any planning algorithm that uses either of the search paradigms is that the amount for the search that the planner performs could be huge, and therefore, the planner could take a lot of time before it returns a solution plan. Some of the most impressive advances in the recent years are based on various techniques for organizing the search space and guiding the search. Sometimes, these techniques are *domain independent*, i. e., they can work in any planning domain, and sometimes they are *domain-specific*, i. e., they exploit characteristics of the underlying planning domain. When *domain-specific planning knowledge* is used for planning, sometimes the planners themselves are domain-specific, i. e., they work only in a particular domain such as process planning [21] or bridge [46]. In other cases, the planner is *domain-configurable*, i. e., it consists of a domain-independent planning engine, plus a language for writing domain-specific problem-solving information.

For modeling and learning structured planning knowledge about a problem domain, one of the best-known formalisms is *Hierarchical Task Networks (HTNs)*. An HTN planner formulates a plan by decomposing tasks (i. e., symbolic representations of activities to be performed) into smaller and smaller subtasks until tasks are reached that can be performed directly. The basic idea was devel-

oped in the mid-70s [45,47], and the formal underpinnings were developed in the mid-90s [13].

An *HTN planning problem* description consists of the initial state and the goal task network (a set of tasks to be performed, along with some constraints over those tasks that must be satisfied). A solution to an HTN planning problem is a *plan*, as in classical planning; i. e., a sequence of actions that, when executed in the initial state, perform the desired tasks.

In order to generate solutions for the planning problems, an HTN planner uses an *HTN domain description* that contains mainly two kinds of knowledge artifacts: *methods* and *operators*.[1] The operators are like the planning operators used in any classical planner. The action symbols of these operators are designated as *primitive tasks* (i. e., tasks that we know how to perform directly). Any task that does not correspond to an operator name is a *non-primitive* task.

Each method is a prescription for how to accomplish a non-primitive task by decomposing it into subtasks (which may be either primitive or non-primitive tasks). A method consists of four elements: (1) the task that the method can be used to accomplish, (2) the set of preconditions which must be satisfied for the method to be applicable, (3) the subtasks to accomplish, (4) any constraints on the bindings of objects to variables in the above and on the orderings of the subtasks to be accomplished.

For example, consider the task of moving a collection of crates from their containers to their designated depot areas. Here is an informal description of a method we might use for the task move_crate in the Simple-Storage domain:

method move_crate_by_hoist for the task move_crate:
- *applicability conditions:* the hoist is in working order and is present at the location of the container, the crate does not have any other crate on top of it
- *subtasks:* (1) pick up the crate, (2) move the hoist to the designated depot, (3) locate the depot area, and (4) put down the crate
- *constraints:* do subtask (1) before subtask (2), do subtask (2) before subtask (3), and do subtask (3) before subtask (4).

Machine Learning Methods in Planning

The most commonly investigated machine-learning paradigms that were used in automated planning systems are *explanation-based learning* and *inductive learning*. We will overview these paradigms below; the subsequent sections will discuss some examples of their integration with planning and describe why these paradigms have been the most popular among researchers so far.

Explanation-Based Learning (EBL) EBL is an *analytical learning* method, where an explanation-based learner uses prior, common-sense background knowledge about the problem-solving domain and deductive reasoning over that background knowledge to augment the information originated from a certain number of input training examples. The background knowledge in EBL helps the learning process in several ways, described below.

First, the deductive reasoning in EBL over the background knowledge and the training examples specifies those features of the hypothesis space that are relevant to the underlying learning problem. The objective of a learning problem is to derive an explanation (i. e., a logical proof) of a given target concept. For example, a target concept in our Simple-Storage planning problem described earlier could be a simple or complex logical sentence such as "it is safe to place the crate x on top of the crate y in the depot", or in a space-operations domain, it could be "any day without any sandstorms on Mars is good for the rover Opportunity to explore the Victoria's Crater".

Given the target concept (i. e., the hypothesis concept), an EBL algorithm attempts to generate a logical proof of that concept using the training examples and the background theory. The background domain theory specifies any additional, general and common-sense background knowledge about the underlying problem. For example, the background theory for stacking containers above might include rules as follows:

$$SafeCheck(x, y) \leftarrow Crate(x), Crate(y), Stackable(x, y).$$

$$Stackable(x, y) \leftarrow Weight(x, wx), Weight(y, wy), wx < wy.$$

A training example would most naturally be modeled as the values of certain attributes of domain and the target concepts; for example, in the Simple-Storage domain, a training example would contain information about the weights of the crates such as $Weight(Crate\,0,10)$ and $Weight(Crate\,1,20)$. From this information, the EBL system deduces that $SafetyCheck(Crate\,0, Crate\,1)$ is true.

[1] HTN planners may have other kinds of knowledge artifacts as well. For example, the SHOP2 planner [39] has axioms that can be used to infer conditions about the current state.

If there is no logical proof of the target concept generated by deductive reasoning over the background theory and the training examples, then the algorithm discards the hypothesis concept. Otherwise, each concept that appears in the proof is relevant to the hypothesis and each concept that does not appear there is not relevant. Note that, with this deductive proof-generation capability, an EBL algorithm can discover relevant concepts to the target that cannot be generated from the training examples alone, which is an important advantage of EBL over other learning paradigms as this allows EBL to generate more general hypotheses to learning problems.

Deductive reasoning over the input background knowledge in EBL also allows a learning algorithm to focus on only the relevant portions of the hypothesis space, ignoring the rest. As a result, the complexity of the hypothesis space and the learning process is usually much less than that in other learning paradigms that cannot exploit such background knowledge.

A key drawback of EBL that surfaces in some cases is that the background knowledge must be complete and perfect in order for EBL to work and generate correct hypotheses as described above. The reason for this is that the deductive proof-generation in EBL cannot differentiate between perfect and imperfect knowledge, and even if it could, the deductive proof-generation in EBL would not be able to handle it. Thus, if the input knowledge contains errors or it is incomplete, then there is no guarantee that the hypotheses found by EBL algorithms will be correct solutions to the learning problems. There has been some attempts to alleviate this issue with EBL, and we are going to touch those works below when we are discussing *Inductive Learning* methods.

Inductive Learning Inductive learning focuses on the problem of producing (i. e., inducing) general hypotheses from specific training examples. Inductive learners generalize from the observed examples by identifying features that empirically distinguish positive examples from negative ones. Version-space learning, decision-tree learning, inductive logic programming, neural networks, and genetic algorithms are all examples of the inductive learning paradigm.

The primary difference between inductive learning and explanation-based learning is the use of background knowledge in the latter. In inductive learning, a learner selects a hypothesis from the given hypothesis space that fits best to the input training examples $(\langle x_1, f(x_1)\rangle, \ldots, \langle x_n, f(x_n)\rangle)$ in order to model an unknown target concept f. For example, consider our target concept *safe stacking* from above. In addition to this target concept, we must also specify all of the relevant attributes in the planning domain for an inductive learner to learn this concept. In this case, let's suppose the following is the set of those relevant attributes:

- *Holding*: whether the hoist is holding a crate or not;
- *WeightTop*: whether the crate to be placed on top has light, medium or heavy weight;
- *WeightBottom*: whether the crate to be placed at the bottom has light, medium or heavy weight;

Some training examples for inductively learning our safety-stacking concept from the above attributes will be as follows:

Holding = *True*, *WeightTop* = *Light*,
  WeightBottom = *Heavy*, *SafeStack* = *True*
Holding = *True*, *WeightTop* = *Light*,
  WeightBottom = *Medium*, *SafeStack* = *True*
Holding = *True*, *WeightTop* = *Heavy*,
  WeightBottom = *Medium*, *SafeStack* = *False*
Holding = *True*, *WeightTop* = *Heavy*,
  WeightBottom = *Light*, *SafeStack* = *False*
Holding = *True*, *WeightTop* = *Heavy*,
  WeightBottom = *Heavy*, *SafeStack* = *True*

From these examples, an inductive learner will produce the following definition of the SafetyStack concept:

Holding = *True* AND
  IF *WeightTop* = *Light* THEN *SafeStack* = *True*; ELSE
  IF *WeightTop* = *Heavy* AND *WeightBottom* = *Heavy*
    THEN *SafeStack* = *True*; ELSE
  IF *WeightTop* = *Heavy* THEN *SafeStack* = *False*;

The reasoning behind this learned concept is as follows. First of all, in every training example the learner has seen *Holding* is *true*, and therefore, it appears as such in the definition of the target concept. The first and second examples specify the case in which $WeightTop = Light$ for two different values for *WeightBottom*. Thus, the inductive learner generalizes from the value of the latter attribute for the case *WeightTop* = *Light* and *SafetyStack* = *True*. The third and the fourth examples causes the same type of generalization for the case we have *WeightTop* = *Heavy* and *SafetyStack* = *False*. Finally, the remaining if-condition in the above learned concept is originated from the last example.

As learning is based on only the input training examples, the learned hypothesis can only be an approximation of the target concept in inductive learning. The

learning algorithms rely on training examples that are sufficiently many and rich in content, and lacking any further information or rich training samples, they assume that the best hypothesis regarding the unseen examples is the hypothesis that best fits the observed examples. This strong assumption about the training examples is, in some cases, a key practical limit on inductive learning algorithms.

In addition to the above machine-learning approaches, there are two other approaches that have been incorporated with automated planning. Case-based learning is an *instance-based learning* method, where given some training examples, a CBL algorithm simply stores the training examples as *cases* in order to be retrieved, adapted, and used in later problem-solving sessions. CBL uses complex, rich, and relational symbolic representations to represent its cases regarding the training examples. The retrieval and combination of cases to solve to the current training example may rely on knowledge-based reasoning and search-intensive problem-solving methods. Given such complex logical descriptions of already-seen examples, a CBL algorithm attempts to map those descriptions to a newly-seen example and generate an explanation/solution for that example. Note that the methodology of CBL is in contrast to explanation-based and inductive learning, which produce a general approximation of a target concept/hypothesis from training examples and sometimes background knowledge. Generalizing beyond the already-seen examples are postponed in CBL until new ones are received.

Reinforcement Learning (RL) emerged as another paradigm where the learning and the planning processes are modeled under the same framework (rather than a combination of the two in some fashion), with a focus on non-classical planning domains where actions may have multiple, non-deterministic outcomes and the objective is to produce a mapping from states of the world to actions that optimizes some utility function. We will come back to RL later in this paper in Sect. "Reinforcement Learning" and discuss it in more detail as a separate topic, as it is a quite different planning and learning approach than the way above learning paradigms have been usually investigated in the context of automated planning.

Learning Domain Models

An important assumption of traditional automated planning is that the planning algorithms require complete and correct descriptions of all of the possible planning operators in a planning domain; otherwise, the plan they generate does not guarantee to be a correct solution to planning problems. In many practical problems, however, it may not be possible to compile the complete models of the actions due to the uncertainties in the world, incomplete information about the semantics of the actions, and other practical reasons. Thus, it is important to have the ability to learn action models in order to produce complete semantics for them and plan with them.

The primary approach for learning action models is by observation. Intuitively, a learner is demonstrated information about the semantics of the actions (i. e., about the preconditions and effects), allowed for generalizing from that information, and practice with the generalized knowledge to discover mistakes and correct hypothesis. This process continues until the learned action models provide failure-free planning for the planning problems in a domain.

An observation for an action model can be understood in terms of the state-transitions induced by that action: let s be a state in which the action a is applicable and suppose applying a in s produces s'. Then, the tuple (s, a, s') is an observation of the action model; i. e., the states s and s' give information about the preconditions of the action and its effects. Note that we are not assuming anything about the action here; we do not necessarily know a's preconditions and effects; we only observed that a was applied in s and produced s'.

The observations over the action models could be given in terms of solution plans as training examples, or in terms of single-action executions in the world. The first time an observation is obtained on the action a, the learner makes the entire state s as the preconditions of a and the entire state s' as its effects. If the learner is provided with a domain theory that specifies information about what is not true in those states, it is also possible for the learner to exploit that theory to hypothesize negated preconditions and effects for a as well. As the learning procedure sees more observations for the action a, it generalizes from those initial preconditions and effects.

At any point before the learning finishes, the acquired planning operators are not necessarily complete or correct in the following ways. First, the learned preconditions for a planning operator could be *over-specific* or it could be *over-general*. In this case, the preconditions of the planning operator will have redundant and perhaps incorrect conditions and a planner that uses this operator will have to achieve all of those unnecessary preconditions during planning. In some cases, i. e., if the planner a plan-space planner, this does force the planning system to do unnecessary search by attempting to insert actions to establish those preconditions and so on. The entire process is deviated from the actual objectives of the planning problem in hand. In a state-space planner, the unnecessary

conditions may make the operator inapplicable in certain states, which again, induces unnecessary search effort. Furthermore, unnecessary preconditions can make many solvable problems unsolvable because it may not be possible to achieve some of them.

If the preconditions are over-general, then this increases the branching factor of the search space of the planner (since a planning operator may become applicable in a state that it should not), degrading the performance of the planner significantly.

When some effects of an operator are not learned, then a planner generates incorrect states during its search for a plan. This makes the later decisions on the applicability of the operator even more problematic, since, for example, an operator that should not be applicable in a state may become applicable and incorrect states with over-general preconditions of operators makes the branching factor even larger.

As a result of the aforementioned factors, a learning algorithm that produces action models by observation usually require several examples of such models in order to refine its learned knowledge.

Learning Search-Control Knowledge

As we discussed in Sect. "Artificial Intelligence Planning, Traditionally", a planning algorithm is essentially a search procedure over a space of planning constructs, such as world states in state-space planning and partially-ordered plans in plan-space planning. At each point during its search, the planner must make some decisions and choose among many alternative ways of how to expand its search. Sometimes those choices leads to failures and the planner backtracks in order to try other alternatives. Other times, the choices lead to solution plans; sometimes, those choices produce good quality plans very efficiently, sometimes they are not efficient, and sometimes they result in poor solutions.

Each decision that the planner makes among alternative choices during the search is a learning opportunity to enhance the planner search-control functionality. If a planner is able to understand and learn good choices against bad ones and its mistakes that result in failures, then the planner does not necessarily repeat those mistakes in later problems, if equipped with sufficient learning and planning capabilities. In the following, we are going to discuss such learning and planning capabilities for controlled search during planning.

Explanation-Based Methods

Probably the most common approach to learning control knowledge is to use analytical learning techniques where the systems learns by problem solving, and the most popular analytical method is Explanation-Based Learning (EBL). As we described in Sect. "Machine Learning Methods in Planning", an EBL system learns by analyzing explanations of problem-solving behaviors; in particular, the system builds an explanation for why an example is a member of some target concept. When used to learn planning control knowledge, the example usually corresponds to a decision made by the planner during its search and the target concept is explaining why the decision was a good one or a bad one. The explanation built by EBL can be used to construct a control rule which can then be used to help make future planning decisions that were similar to the original example.

During planning, a planner's problem-solver goes through several critical decision points. In a plan-space planner, for example, these decision points correspond to the choice of which causal links should be established and how, and threat selection and resolution (including promotion, demotion, and separation). In a forward-chaining state-space planner, those decisions correspond to the choice of an action applicable to the current state among all of the applicable ones.

The most common learning opportunity during planning is when the system reaches an impasse; i. e., when the planner reaches to a decision point, but fails to make one. This may happen when there are no applicable actions in the current state during the search or there is no way to resolve a threat on a causal link in a partial plan. At such an impasse, the system that exploits an EBL component can attempt to generate an explanation of the failure that produced the impasse. The failures depend on the type of planning being done in the system. When a plan-space planner reaches an impasse, the failure(s) can be explained in terms of the inconsistency of the ordering and binding constraints associated with the partial plan, or in terms of an unsatisfiable open condition. In a state-space planner, the failures could be understood in terms of the unsatisfiable preconditions of the inapplicable action(s).

As an illustration of how search failures are used as explanations, consider the search tree of a forward – chaining state – space planner as shown in Fig. 2. The search node that is marked as a red box in this tree is the point of impasse, since there are no actions applicable in that state. In our Simple-Storage planning problem, for example, the red box could represent a state that specifies a crate that is in a depot area, another crate that is held by the hoist, and the hoist is in that particular depot area. In this configuration, suppose the crate on the hoist, say crate2, is heavier than the one in the depot area . Therefore, the action (put-

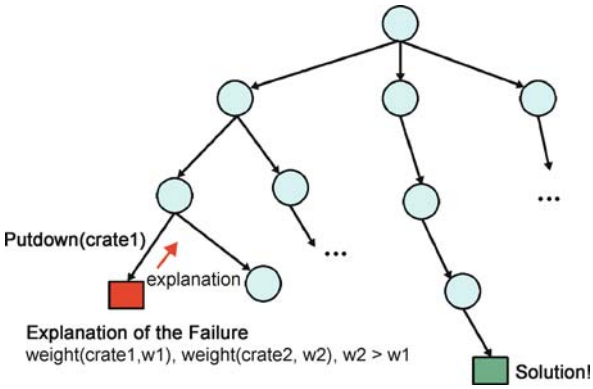

Learning and Planning (Intelligent Systems), Figure 2
The example of a search tree generated by a forward-chaining state-space planner

down crate2) is not applicable. The explanation for this failure can be understood by

$$weight(crate\,1, w\,1), weight(crate\,2, w\,2), w\,2 > w\,1\ .$$

While backtracking, the planner propagates back this explanation to the parent node of the failed one and through that node, the rest of of the search node to be expanded. When this explanation is generated and propagated, the system uses this explanation as a search-control rule and it never attempts to move the first crate on top of the second in the later search process.

Learning from failures in the way described above is the simplest form of using EBL for producing planning control knowledge. Another possible source for learning during planning is the successful search traces that a planner generates. For example, consider again the search tree shown in Fig. 2. The decision points in the search trace that leads to an optimal solution to the input planning problem, as marked in the figure, could be used for learning control rules that prefers these decisions or their generalizations over the other search traces in the figure. In order for an EBL component to learn somewhat more effective control rules than the sort described above, the system must be provided a domain theory that includes rules for deriving relevant features of the planning problems and the search process, as well as templates for the control rules to be learned by the system.

As an example, suppose the domain theory for the Simple-Storage domain includes the following control-rule template:

$$\textbf{select action}\ (\text{move}, ?hoist, ?from, ?to) \leftarrow COND_1, \ldots, COND_k\ ,$$

where each $COND_i$, $i = 1, \ldots, k$, specifies a template for any logical formula that can be evaluated given a state and any combination of concepts that cannot directly inferred from the state, but could be learned using the domain theory. For example, suppose the planner successfully applied the action move(hoist0, load-area, depot0-1-1) in a state consists of the following:

$$\begin{aligned}
&in(depot\,0_1, depot\,0), in(crate\,1, depot\,0_1),\\
&\quad holding(crate\,0)\\
&in(container\,0, container_area),\\
&\quad connected(load_area, container_area)\\
&connected(container_area, load_area),\\
&\quad connected(depot\,0_1, load_area)\\
&connected(load_area, depot\,0_1), at(hoist\,0, load_area)
\end{aligned}$$

Then, with the above domain theory, an EBL component of the system will learn a control rule that is similar to the following:

$$\begin{aligned}
&\textbf{select action}\ (\text{move}, hoist\,0, load_area, depot\,0_1)\\
&\qquad \leftarrow holding(crate\,0), at(hoist\,0, load_area),\\
&\qquad\quad at(crate\,1, depot\,0_1),\\
&\qquad\quad SafeStack(crate\,0, crate\,1)\ ,
\end{aligned}$$

where the definition of the concept $SafeStack(crate\,0, crate\,1)$ is learned via the domain theory given in Sect. "Machine Learning Methods in Planning".

The search-control rules using an EBL approach as above can be learned either during planning or after planning. In the former case, the planner propagates its explanations of failures while backtracking to the upper levels of the search tree and forms the control rules based on the domain theory during this propagation. The control rules generated in this fashion could then be used in the later branches of the search tree and in the other planning problems in the same planning domain. In the latter case, the planner generates the entire search tree that includes both the failed and successful search traces and post-processes the generated tree in order to derived the search-control rules. These rules are then used in other planning problems. Learning via backtracking has the advantage of not requiring to wait until the entire search tree is generated: the learner could produce rules in an incremental fashion and use those rules as they are generated during planning. Learning after planning, on the other hand, cannot produce and use any control rules until the search tree is constructed for an input problem. However, as the search tree contains more information in terms of comparisons between failed and successful decisions and optimal and suboptimal search traces, which is not usually available during a backtracking search, this approach can learn more expressive and powerful control rules.

Unfortunately due to its reliance on an human expert specified domain theory and only a few examples, standard EBL can often produce complex, sometimes overly-specific and sometimes overly-general control rules that cannot be utilized very well in new planning problems [4,32]. This situation is commonly known as the utility problem, where even though the learned rules are correct, the cost of testing their applicability to new planning situations often outweighs their savings. Another problem is that EBL methods are difficult to apply in domains where it is hard to construct a complete and tractable domain theory. These drawbacks usually make it difficult to successfully apply EBL methods to real-world problems.

Inductive Methods

Inductive learning techniques can acquire planning control rules by looking at examples of positive and negative planning decisions. These examples are usually found by analyzing a solution to a planning problem and by solving a set of training problems and extracting examples based on good and bad planning decisions. Positive learning examples for inductive methods are those planning decisions that a hypothesized search-control rule correctly classifies – i. e., the control rule should not eliminate the planning decisions that yield to solutions and guides the planner to take that decisions. Negative examples, on the other hand, are those planning decisions that does not yield to a solution but enforced by the control rule on the planner. In the former case, if the control rule eliminates a good decision, then this means that the rule is over-specific: in that case, an inductive learner generalizes the rule in order to classify all positive examples. In the latter, the case the negative example describes a case where the rule is overly-general since it accepts a decision it is not supposed to. In this case, the learning component specialized this rule by adding extra constraints while correctly classifying all positive and negative examples.

The difference between inductive methods and EBL is that the latter is a deduction based method: EBL methods use a expert-provided domain theory in order to produce search-control knowledge as explained above. Inductive methods, on the other hand, build control rules by performing an inductive search through the space of possible rules, where the objective is to produce search-control knowledge that cover some positive planning decisions but do not cover any negative planning decisions.

The inductive search is usually guided by a domain independent bias that helps the learning component to choose between several search-control rules that applies to the same or very similar planning decisions. For example, both of the following rules describe the situations to the planner in which it should prefer a particular crate over another:

1. **select action** (move, *hoist* 0, *load_area*, *depot* 0_1)
   $\leftarrow$ *holding*(*crate* 0), *at*(*hoist* 0, *load_area*),
   at(*crate* 1, *depot* 0_1),
   SafeStack(*crate* 0, *crate* 1).
2. **select action** (move, *hoist* 0, *load_area*, *depot* 0_1)
   $\leftarrow$ *holding*(*crate* 0), *at*(*hoist* 0, *load_area*),
   at(*crate* 1, *depot* 0_1),
   in(*container* 0, *container_area*),
   connected(*load_area*, *container* 0),
   connected(*container* 0, *load_area*),
   connected(*depot* 0_1, *load_area*),
   connected(*load_area*, *depot* 0_1),
   SafeStack(*crate* 0, *crate* 1) .

The first rule is much more simpler than the second one since the latter includes some redundant features of the decision on the crate, which can also be inferred from the first rule. One very-commonly used bias is Occam's razor which prefers simple rules such as the first one above over more complex ones.

A main advantage of inductive learning techniques over other learning methods such as EBL is that they tend to learn very general control knowledge since they examine a number of different examples. This quality is important for building useful general control knowledge, which can be more effectively applied to new planning problems. Unfortunately, inductive techniques have several disadvantages as well. Perhaps, most importantly, inductive methods usually require large numbers of examples to acquire effective control information. Also, these methods can quickly become computationally intractable since they often search through large amounts of information when building control rules. These disadvantages often prevent inductive methods from being successfully used for planning problems.

However, inductive learning of search-control knowledge can be combined with EBL methods to develop systems that could address the issues with the both approaches when exploited independently. One way to combine the two approaches is to use EBL methods to generate a first draft of search-control knowledge. As we discussed above, EBL methods usually produce very specific search-control knowledge given the expert-provided domain theory. Furthermore, the learning mostly depends on this input domain theory, which might be incorrect and in-

complete. In such cases, EBL methods could produce over-general knowledge, or even incorrect control rules. However, this knowledge can then be refined that knowledge via inductive generalization and specialization over the set of training examples.

Learning Operational Planning Procedures

The techniques described in the previous section enhance the performance of a planner by learning search-control knowledge about what to do and what not to do during a problem-solving session and using that knowledge on later problems in order to organize the search space and not to do similar mistakes all over again. Despite the effectiveness of these techniques, in many domains, the performance of a planner can be improved even more greatly by inferring and exploiting operational knowledge about the domain structure that is not explicitly encoded in the formulation input planning problems (i. e., it is not encoded in terms of the planning operators, initial and the goal state).

As many human skills are procedural and hierarchical in nature, with complex procedures defined in terms of simpler ones, it is usually preferable to describe the structural domain planning knowledge in a form that conforms to a human's understanding of the planning problems and domains. In the simplest form, such knowledge is described in terms of *macro-operators*, i. e., sequences of planning operators (i. e., actions) that are grouped together since all or most solution plans in the planning domain require those actions to be executed in that sequence. More expressive planning knowledge is described in multiple hierarchies and usually in terms of HTN methods discussed in Sect. "Artificial Intelligence Planning, Traditionally". In the rest of this section, we are going to present these two forms of of domain planning knowledge and describe basic methods for how to learn such knowledge and use in planning.

Learning Macro-Operators

Informally, a *macro-operator* is perhaps the simplest form of planning knowledge about how a planning domain works and what characteristics that the planning problems in that domain exhibit in terms of the actions they contain. A macro-operator specifies, therefore, a grouping of the standard planning operators such that the operators (or more precisely, the actions that are the ground instantiations of those planning operators) are executed together in all or most of the solution plans in the domain.

Let O be a set of planning operators in a classical planning domain. Then, formally, a *macro-operator* is a tuple of the form $(O', \prec, \sigma)$ where $O' \subseteq O$ is a multi-set of classical planning operators, $\prec$ is a partial-ordering relation over the operators in O', and σ is a variable-binding substitution over the variables that appear in the operators of O'. Note that the set of planning operators O' may contain more than one instance of the same planning operator.

A planning process that involves macro-operators has usually two components: learning a set of macro-operators for a domain, and afterwards, using them during planning. The solution methods for learning macro-operators first generate solutions to a number of planning problems in the underlying domain, and then, process the structural properties of those solutions in order to infer correct and useful operator groupings as macros. The simplest way of realizing this learning approach is to identify patterns of actions in those solution plans. For example, consider our example Simple-Storage planning domain and suppose we have the following plan:

$$\langle move(hoist\,0, depot\,0_1, load_area), move(hoist\,0, load_area, container\,0), pickup(crate\,0, container\,0, container_area), move(hoist\,0, container\,0, load_area), move(hoist\,0, load_area, depot\,0_1), putdown(crate\,0, depot\,0_1, depot\,0)\rangle .$$

In two places in this plan, there are sequences of move operations for moving the hoist between the load area, depot, and the container. This pattern is a good candidate as a macro-operator because we could expect that any solution plan in this domain would involve a hoist performing similar move operations between a load area, a container, and a depot.

A more complicated and better approach would be to use structural semantic information in those example plans in order to learn the macro-operators. The basic structural semantic property in all planning domains is the causal dependencies between the actions in the solutions. A causal dependency is the relationship between two actions in the plan, where an effect of one of the actions establish a precondition of the other. A causal dependency is *positive* when an action a produces an effect p and the action a' has p as one of its preconditions. A causal dependency is *negative* if a produces $\neg p$ and p is a precondition of a'. Note that the causal dependencies above are very similar to the causal links that we described for plan-space planning previously in Sect. "Artificial Intelligence Planning, Traditionally".

Figure 3 shows the graph of causal dependencies within this plan. This graph reveals certain patterns in the solution that may not be trivially extracted from a pattern-based learning approach as the one above. For example, the causal dependency between the actions pickup(*crate* 0,

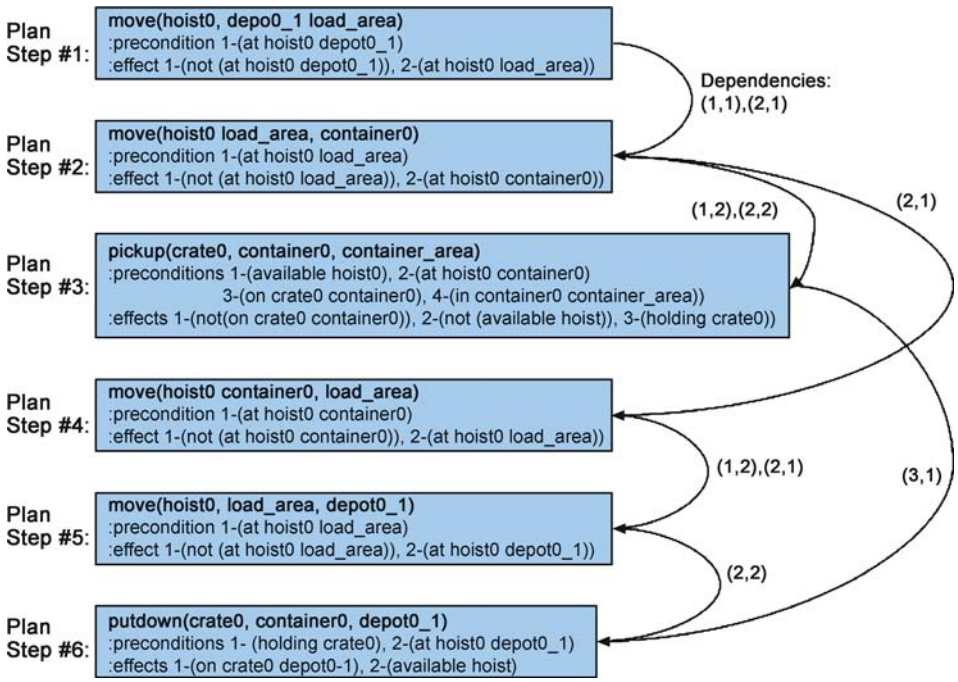

Learning and Planning (Intelligent Systems), Figure 3

A simple example showing the causal dependencies between the preconditions and effects of the actions in a solution plan. Each precondition and effect of an action (i.e., plan step above) is labeled by positive numbers 1, 2, ... A dependency between two actions a_1 and a_2 is shown as a pair of numbers (x, y) where x is the label of a effect of a_1 and y is the label of the precondition of a_2

container 0, *container_area*) and move(*hoist* 0, *container* 0, *load_area*) due to fact that both actions operate the same hoist is not necessarily clear from a syntactic analysis of the above plan. Similar is true for the dependency between the actions move(*hoist* 0, *load_area*, *depot* 0_1) and putdown(*crate* 0, *depot* 0_1, *depot* 0). Such dependencies help a macro-operator learner to produce better groupings of actions such as the following one:

$$\langle pickup(?crate, ?container, ?area), move(?hoist, ?container, ?load_area), move(?hoist, ?load_area, ?depot_area), putdown(?crate, ?depot_area, ?depot)\rangle.$$

Typical solution methods for learning such macro-operators in our Simple-Storage domain involve searching through the subgraphs of the causal-dependency graph of a solution plan and identifying patterns in those subgraphs. Then, the learning algorithm compares those patterns with those generated from other solution plans in the domain and generalizes into a set of macro-operators for the domain. Those macro-operators are then added into the set of planning operators O originally specified from the domain. This way, the macro-operators are treated by a planning algorithm as regular operators, ensuring the soundness and completeness of the planning algorithm.

The learned macro-operators are used for speeding up the search in a planning algorithm by causing jump-aheads in the search space, reducing the amount of work done by the search and the amount of backtracking. It has been experimentally demonstrated in the International Planning Competitions that learning and using macros in classical planners increases the performances of the planners significantly, allowing them to solve larger planning problems faster [5,12]. One important issue with using macro-operators, however, is that the search for macros over the possible subgraphs of solution dependencies could generate many macro-operators, most of which are redundant and some of which are not useful for planning. Thus, typical learning methods for macro-operators include heuristic selection and filtering capabilities to prune the set of generated macro-operators and use a small, but heuristically most useful ones in later planning.

Learning HTN Methods

Macro-operator learning is the simplest form of learning procedural domain planning knowledge that is described by observed groupings of actions that an executor performs in the plans. Although macro-operators can improve the performance of a planner as described above,

their expressivity in terms of the planning knowledge they specify is limited. For example, it is not possible to say and learn "if there is a crate that needs to be moved to a depot and there is a hoist at the crate's current location, move the crate with that hoist; otherwise, use a hoist that is the nearest one to the crate's location" in terms of a macro operator.

More specifically, macro-operators can only specify domain structures that is observable in the solution plans; they cannot include complex operational planning procedures for the domain that are not clearly explicit in the observations. Furthermore, macro-operators do not include any preconditions that does not appear in the original planning operators already; the conditions under which a macro-operator is applicable in a state is usually determined by the causal dependencies and the preconditions of the operators involved in that macro. This means that the applicability conditions of a macro-operator are sound, but do not specify any additional strategical knowledge about when and how to apply that macro operator.

Learning the kinds of strategical planning knowledge mentioned above requires more expressive languages in which such knowledge could be formally written. One of the most successful approaches that address such expressive has been languages that allows for specifying procedural and hierarchical planning domain structures, such as Hierarchical Task Networks, and their variants. There has been several recent techniques have been developed in order to learn HTNs or certain portions of HTN knowledge; we are going to present the basic ideas of this paradigm below.

As an illustration of the kinds of structured knowledge we are aiming to learn, recall our example of an HTN method for moving a crate from its container to its designated depot. We copy the example from Sect. "Artificial Intelligence Planning, Traditionally" here:

method move_crate_by_hoist for the task move_crate:
- *applicability conditions:* the hoist is in working order and is present at the location of the container, the crate does not have any other crate on top of it
- *subtasks:* (1) pick up the crate, (2) move the hoist to the designated depot, (3) locate the depot area, and (4) put down the crate
- *constraints:* do subtask (1) before subtask (2), do subtask (2) before subtask (3), and do subtask (3) before subtask (4).

There are several issues that a learner must address while learning the above components of a method. Here is a summary:

- A learner must produce/hypothesize the structural (i. e., hierarchical and procedural) organization of knowledge. This involves, in an HTN framework, determining the structure of the HTN methods, i. e., the task a method can decompose and into which subtasks it does so.
- The applicability conditions under which a method should be invoked.
- Any binding and ordering constraints regarding the subtasks and the variable symbols in the method.

Typical solution techniques for learning such HTN methods involve a sort of explanation-based learning (EBL) methodology. As we described in Sect. "Machine Learning Methods in Planning", any EBL-based learning method requires a domain theory to be provided to the learning for deductive proof generation process. For HTN-method learning, the input domain theory also needs to augmented with some *operational information* regarding a planning domain. For example, consider the following rules for our *Simple-Storage* problem:

Move_Crate (?*crate*, ?*from_loc*, ?*to_depot*)
$\leftarrow$ *Crate*(?*crate* 2), *in*(?*crate* 2, ?*to_depot*),
SafeStack(?*crate*, ?*crate* 2),
PickedUp(?*crate*, ?*from_loc*).

PickedUp (?*crate*, ?*from_loc*)
$\leftarrow$ pickup(?crate, ?container, ?area),
move(?*hoist*, ?*container*, ?*load_area*),
move(?*hoist*, ?*load_area*, ?*depot_area*),
putdown(?*crate*, ?*depot_area*, ?*depot*).

Note that the above rules contain facts that can be evaluated directly in a state such as *in*(?*crate* 2, ?*to_depot*), high-level domain concepts that could be learned by EBL such as *SafeStack*, and descriptions of actions and higher-level activities that could be performed in the domain. For example, the first rule above specifies a higher-level activity called Move_Crate. It further specifies some conditions under which this activity could be accomplished in the body of that rules such as the learned concept *SafeStack*. The second rule specifies a concept that corresponds to a grouping of primitive actions that can be executed in the world. Note that the body of this rule is essentially a specification of a macro-operator that could also be learned by the techniques described in the earlier. The *PickedUp* concept is the crucial piece that allows to learn a hierarchy, as it appears both in the head of the second rule and in the body of the first one.

An operational activity specification allows an EBL-based learning method to accomplish several things. The rules in the specification provide a general common-sense description of associating the activities in the domain with the factual world state predicates. These predicates could be used to specify the goals that an activity could be expected to accomplished when performed in the world, or other side effects associated with the activity.

Second, as the proof tree of a deduction engine in an EBL-based learner will use and include information from task specification rules, the proof itself gives significant and strong clues about the structure of an HTN method and hierarchy of the methods. In other words, a learning algorithm for HTNs uses the proof tree to generate the relationships between tasks and their subtasks, as well as the applicability conditions associated with those relationships using the proof trees generated by the EBL mechanisms. These proof trees are further refined to produce the HTN methods by identifying the tasks in those trees and their subtasks, and organizing the subtrees to generate the applicability conditions and the variable-bindings regarding them.

Note that the above learning schema only produces simple HTN structures, where the only constraints over the orderings of the subtasks of a task is mostly totally-ordered, as dictated by the proof tree. Furthermore, generally HTN methods could encode complex groupings of subtasks (that is, more formally, HTNs are capable of modeling context-free languages) and complex recursion over those tasks. Depending on the problems, the practical implementations of the learning schema described above could learn some simple forms of recursion and methods that encode a bit more complex than regular languages, but not structural knowledge as expressive as standard HTNs. Extending this learning schema remains an important research direction in automated planning and learning.

Reinforcement Learning

Reinforcement Learning (RL) focuses on the problem of learning to choose actions for possible situations that an intelligent system could encounter while acting and interacting with the environment. The objective of a reinforcement learning is to use rewards and punishments (i. e., reinforcements) it receives from the environment while acting, in order to produce a *policy* (i. e., a mapping from situations to actions) that maximizes some utility function defined in terms of the reinforcements. In this respect, reinforcement learning naturally diverges from the other machine-learning paradigms, esp., the ones we have mentioned in this paper, since RL does not learn from any problem knowledge either in the form of training examples or in the form of background theory. Instead, the emphasis in RL is to learn from solely interactions with the world in order to achieve an explicitly stated goal.

When seen as a form of interleaved planning and execution paradigm, RL is a very effective form of generating plans (i. e., policies) in non-classical domains. In non-classical domains, actions may have multiple possible outcomes and the planner usually does not know which outcome will occur in the world when the action is executed. The planning process is usually characterized as trial-and-error search and the interactive learning built into the core of RL provides a way for reasoning about such multiple outcomes of the actions, since the learner executes and observes the effects of the actions immediately and weights that execution with other actions that might have been executed in terms of the rewards and punishments it receives.

Furthermore, interleaving planning and execution does not require that a RL algorithm know the underlying domain model, i. e., the preconditions and the effects of the actions. Whether an action is successfully applied in a situation is given through the reward mechanisms and the RL algorithm learns the effects of the actions via its observations of the world after the execution of the action. Thus, in this respect, RL can also be seen as a learning paradigm domain understanding since the outcome policies specify knowledge about how the domain works. Due to this property, most researchers believe that reinforcement learning is very suitable for robotic applications, systems control, games, and other applications.

Typical solution methods for RL problems, such as Q-Learning [50], are based on dynamic programming and simulation techniques, and they usually require exploring all or most of the state space in order to generate optimal or near-optimal policies. In complex problems, the sizes of the state spaces are usually prohibitive, and as a result, RL algorithm usually have efficiency problems in such settings. One way to alleviate this problem is to incorporate the aforementioned learning methods for planning into RL algorithms in order to learn within RL some sort of planning knowledge about what to do and what not to do by using background knowledge. An example of this approach is the integration of explanation-based learning techniques with RL, as described in [10].

Summary and Future Directions

The ability to produce knowledge about past experiences and exploit that knowledge in an operational context in

later problem-solving and planning sessions is an important attribute of any intelligent system, human- or AI-based alike. Automated Planning and Learning is the research paradigm that focuses on the development of intelligent systems and technologies that combine the ability to make decisions and generate courses of actions (i. e., plans) with the capability to reason and produce knowledge about past experiences, future problems that the system needs to tackle, and strategies about how to tackle them.

We have described the basic concepts in automated planning systems where machine learning has been of assistance in order to produce planning knowledge. We presented first a summary of traditional automated planning paradigm of AI and an overview of analytical machine-learning methods that have been deemed suitable for integration with planning systems. Then, we presented an analysis of four different forms of approaches that have integrated planning and learning techniques to achieve different objectives.

Learning domain models from observation has been investigated in several works such as [28,30,31,48,49,52,53]. Our formulation for learning domain models in this paper is partly based on the early work described in [48] and partly an analysis of all the other works on this topic. Among those that are more advanced than our formulation here, Levine and DeJong [30] describes a different approach to learn domain models that is based on explanation-based learning techniques. The work in [28] describes a way not only to learn action models, but also some temporal characteristics of the actions such as the duration it takes an action when executed. It is important to note that all these recent developments in learning domain models go beyond the expressivity of the earlier approaches, and this holds great promise for their applicability in realistic applications.

Explanation-based learning has been probably the most popular machine-learning method applied in automated planning, as evidenced by the works reported in [4,7,9,20,24,25,27,32,33] and others. Inductive learning, on the other hand, has been less investigated in planning, probably due to the correctness and reliability of the produced knowledge when there are not many training examples, compared to EBL's deductive reasoning guarantees. Examples of systems that used inductive learning for planning include [4,15,22,31,43] and others. Many of these inductive approaches, however, include an EBL component in the learning process in one way or another to reduce the uncertainty with the training examples. Thus, induction is usually used to overcome the limitations of EBL to produce over-specific or over-general knowledge, and to produce knowledge that is reliable based on a domain theory even if there are not sufficiently many training examples.

Existing works on learning macro-operators (e. g., [5,36]) are sometimes categorized as speed-up learning mentioned above, as do work on learning search control knowledge (e. g., [16,32,35]). Recently, several architectures have been proposed to learn hierarchical planning knowledge from a collection of plan traces and from a given action model [8,42,44]. Our formulations of learning macro-operators and operational procedures (i. e., HTNs) described in this paper are based on the previous works [5,8,36,40] on the topic.

Although our description of learning macro-operators and planning procedures is mostly based on EBL methods, rather than inductive learning, there has also been some advances on the latter approach for learning operational planning procedures. For example, [42]'s X-Learn, for example, uses inductive generalization to learn task decomposition constructs, which relate subgoals and conditions for decomposing those goals into their subgoals. By grouping goals in this way, task models are learned that lead to significant speed-ups in problem-solving. As another example, [8,40] achieves the same objective via problem-solving techniques similar to explanation-based learning. Two other recent studies [23,51] proposed algorithms to learn the applicability conditions of task-decomposition constructs, given the hierarchical relationships between the tasks.

Despite all the approaches described in this paper that have been very successful in learning for planning, we believe there are still interesting planning problems that machine-learning has not yet or just started to contribute. Here are several examples of such problems:

Learning for Multi-Planner Systems

There has been a growing recognition that sophisticated, *multi*-planner systems are needed for full automation in real-world applications involving complex, uncertain, and time-directed situations. In those situations, extensive planning knowledge is required but is difficult to obtain. This is partly because of the complexities in the environments, e. g., rescue and evacuation operations, and it is partly because there is often no expert to provide it, e. g., space operations. In such complex domains, a planning system that can learn such knowledge to develop ways on how to safely and correctly operate in the world holds great promise for success.

A key challenge of automated planning, when *multiple heterogeneous planners*, is "safety learning"; i. e., the problem of automated extraction of some safety con-

straints to identify and ensure the correct behavior of the entire system. A general-purpose use of the learned safety constraints further motivates a hybrid approach, which involves combinations of the learning methods described in this paper, such as inductive and explanation-based learning. It also requires the ability to maximize the information from the world, as such information is not available and reliable in most real life situations. For this latter requirement, Bayesian Learning approaches could be used in automated planning, although examples of Bayesian Learning for planning are rare.

Planning and Learning Under Uncertainty

All of the learning methods and opportunities we described in the paper for automated planning involve traditional planning problems, where the actions are deterministic and the state information is always complete and correct. In complex planning domains, however, there is usually a huge amount of uncertainty where an action may have a large number of possible outcomes, or even we may not know how many different possible things is going to happen when we execute an action.

We described Reinforcement Learning (RL) as a learning method for planning in planning domains that involve such uncertainties. Although RL can learn mappings from states to actions in non-traditional planning domains, these mappings are usually solutions to planning problems, rather than planning knowledge about a domain. Domain-specific knowledge has been used to search in reinforcement learning to address efficiency problems [10,41]. These approaches are based on hierarchical abstraction techniques that are somewhat similar to HTN planning. Given a domain, the hierarchical abstraction of that domain is analogous to an instance of the decomposition tree that an HTN planner might generate. However, the abstractions must be supplied in advance by the user, rather than being generated on-the-fly by the HTN planner. An important future direction is therefore to investigate learning methodologies for HTNs in the context of reinforcement learning in non-traditional planning domains.

Another approach for handling uncertainty in planning domains is to interleave the planning and learning process with the execution of some the planned actions in order to obtain key information about the world. This information can be then used to determine how to proceed with the planning and the learning processes. The interleaving of planning, learning, and execution has been investigated for mostly domains without uncertainties as in the works of [29,40]. Our description of learning HTNs for planning in this paper were based on these works. However, learning HTNs as in these works are very simple and not expressive enough to deal with uncertainty, and thus, it is still an open research area.

Acknowledgments

This work was supported by DARPA's Transfer Learning and Integrated Learning programs. The opinions in this paper are those of the author and do not necessarily reflect the opinions of the funders.

Bibliography

1. Aha DW (2002) Plan deconfliction, repair, and authoring in EDSS. Technical report, Progress report, Naval Research Laboratory
2. Ai-Chang M, Bresina J, Charest L, Hsu J, Jnsson AK, Kanefsky B, Maldague P, Morris P, Rajan K, Yglesias J (2003) Mapgen planner: Mixed-initiative activity planning for the mars exploration rover mission. In: Printed Notes of ICAPS 03 System demos
3. Bergmann R, Wilke W (1996) On the role of abstraction in case-based reasoning. In: European Workshop on Case-Based Reasoning (EWCBR-96), pp 28–43
4. Borrajo D, Veloso M (1997) Lazy incremental learning of control knowledge for efficiently obtaining quality plans. Artif Intell Rev 11(1–5):371–405
5. Botea A, Müller M, Schaeffer J (2005) Learning partial-order macros from solutions. In: International Conference on Automated Planning and scheduling (ICAPS-05). AAAI Press, Menlo Park, pp 231–240
6. Burstein M, Brinn M, Cox M, Hussain T, Laddaga R, McDermott D, McDonald D, Tomlinson R (2007) An architecture and language for the integrated learning of demonstrations. In:Burstein M, Hendler J (eds) AAAI Workshop Acquiring Planning Knowledge via Demonstration. AAAI Press, Menlo Park, pp 6–11
7. Chien SA (1989) Using and refining simplifications: Explanation-based learning of plans in intractable domains. In: International Joint Conference on Artificial Intelligence (IJCAI-89), Detroit. Morgan Kaufmann, San Francisco, pp 590–595
8. Choi D, Langley P (2005) Learning teleoreactive logic programs from problem solving. In: International Conference Inductive Logic Programming (ILP-05), Bonn. Springer, New York, pp 51–68
9. DeJong GF, Mooney R (1986) Explanation-based learning: An alternative view. Mach Learn 1(2):145–176
10. Dietterich TG (2000) Hierarchical reinforcement learning with the MAXQ value function decomposition. JAIR 13:227–303
11. DOT (1999) An assessment of the US marine transportation system, a report to congress. Technical report, US Department of Transportation, p 103
12. Edelkamp S, Hoffmann J (2004) International planning competition. http://ipc.icaps-conference.org
13. Erol K, Hendler J, Nau DS (1996) Complexity results for hierarchical task-network planning. AMAI 18:69–93
14. Erol K, Nau DS, Subrahmanian VS (1995) Complexity, decidability and undecidability results for domain-independent planning. Artif Intell 76(1–2):75–88

15. Estlin TA, Mooney RJ (1997) Learning to improve both efficiency and quality of planning. In: IJCAI, pp 1227–1233, Nagoya, Japan. Morgan Kaufmann, San Francisco
16. Fern A, Yoon S, Givan R (2004) Learning domain-specific control knowledge from random walks. In: ICAPS, Whistler. AAAI Press, Menlo Park, pp 191–199
17. Fikes RE, Nilsson NJ (1971) STRIPS: A new approach to the application of theorem proving to problem solving. Artif Intell 2:189–208
18. Gerevini A, Dimopoulos Y, Haslum P, Saetti A (2006) International planning competition. http://zeus.ing.unibs.it/ipc-5/
19. Goldman R (2004) Adapting research planners for applications. In: ICAPS workshop on Connecting Planning Theory with Practice
20. Gratch J, DeJong G (1992) Composer: A probabilistic solution to the utility problem in speed-up learning. In: AAAI, San Jose. AAAI Press, Menlo Park, pp 235–240
21. Hebbar K, Smith SJJ, Minis I, Nau DS (1996) Plan-based evaluation of designs for microwave modules. In: Proc ASME Design Technical Conference, Irvine, August 1996
22. Huang Y, Kautz H, Selman B (2000) Learning declarative control rules for constraint-based planning. In: International Conference on Machine Learning (ICML-00). Morgan Kaufmann, San Francisco, pp 415–422
23. Ilghami O, Nau DS, Muñoz-Avila H, Aha DW (2005) Learning preconditions for planning from plan traces and HTN structure. Comput Intell 21(4):388–413
24. Kambhampati S (2000) Planning graph as (dynamic) CSP: Exploiting EBL, DDB and other CSP techniques in Graphplan. JAIR 12:1–34
25. Kambhampati S, Katukam S, Qu Y (1996) Failure driven dynamic search control for partial order planners: An explanation based approach. Artif Intell 88(1–2):253–315
26. Knoblock CA (1993) Generating abstraction hierarchies: An automated approach to reducing search in planning. Kluwer, Norwell
27. Laird J, Newell A, Rosenbloom P (1987) SOAR: An architecture for general intelligence. Artif Intell 33(1):1–67
28. Lanchas J, Jimenez S, Fernandez F, Borrajo D (2007) Learning action durations from executions. In: Proceedings of the ICAPS-07 Workshop on AI Planning and Learning, Providence, 22 Sep 2007
29. Langley P, Choi D (2006) Learning recursive control programs from problem solving. J Mach Learn Res 7:493–518
30. Levine G, DeJong GF (2006) Explanation-based acquisition of planning operators. In: International Conference on Automated Planning and Scheduling (ICAPS-06), The English lave District, 6–10 June 2006. AAAI Press, Menlo Park, pp 152–161
31. Matthew D, Oates T, Cohen PR (2000) Learning planning operators in real-world, partially observable environments. In: AIPS, Breckenridge, 14–17 April 2000. AAAI Press, Menlo Park
32. Minton S (1988) Learning effective search control knowledge: An explanation-based approach. Technical Report TR CMU-CS-88-133, School of Computer Science, Carnegie Mellon University
33. Mitchell T, Keller R, Kedar-Ceballi S (1986) Explanation-based generalization: A unifying view. Mach Learn 1(1):47–80
34. Mitchell TM (1977) Version spaces: A candidate elimination approach to rule learning. In: IJCAI. AAAI Press, Cambridge, pp 305–310
35. Mitchell TM (1997) Machine learning. McGraw-Hill, New York
36. Mooney RJ (1988) Generalizing the order of operators in macro-operators. In: Machine Learning, International Conference on Machine Learning (ML-88), Ann Arbor, 12–14 June 1988. Morgan Kaufmann, San Francisco, pp 270–283
37. Muñoz-Avila H, Aha DW, Breslow L, Nau DS (1999) HICAP: An interactive case-based planning architecture and its application to non-combatant evacuation operations. In: AAAI/IAAI Proceedings, Orlando, 18–22 July 1999. AAAI Press, Menlo Park, pp 870–875
38. Muñoz-Avila H, Breslow LA, Aha DW, Nau DS (1998) Description and functionality of NEODocTA. Technical Report AIC-96-005, Naval Research Laboratory, Navy Center for Applied Research in Artificial Intelligence
39. Nau DS, Cao Y, Lotem A, Muñoz-Avila H (1999) SHOP: Simple hierarchical ordered planner. In: International Joint Conference on Artificial Intelligence (IJCAI-99), Stockholm, 31 July–6 Aug 1999. Morgan Kaufmann, San Francisco, pp 968–973
40. Nejati N, Langley P, Konik T (2006) Learning hierarchical task networks by observation. In: Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, 25–29 June 2006. ACM Press, New York
41. Parr R (1998) Hierarchical control and learning for Markov decision processes. University of California, Berkeley
42. Reddy C, Tadepalli P (1997) Learning goal-decomposition rules using exercises. In: International Conference on Machine Learning (ICML-97), Nashville, 8–12 July 1997. Morgan Kaufmann, San Francisco, pp 278–286
43. Reddy C, Tadepalli P (1999) Learning horn definitions: Theory and application to planning. New Gener Comput 17(1): 77–98
44. Ruby D, Kibler DF (1991) SteppingStone: An empirical and analytic evaluation. In: AAAI, Anaheim, July 1991. Morgan Kaufmann, San Francisco, pp 527–531
45. Sacerdoti E (1975) The non-linear nature of plans. In: International Joint Conference on Artificial Intelligence (IJCAI-75), Tiblisi, pp 206–214
46. Smith SJJ, Nau DS, Throop T (1998) Computer bridge: A big win for AI planning. AI Mag 19(2):93–105
47. Tate A (1977) Generating project networks. In: International Joint Cinference on Artificial Intelligence (IJCAI-77), Cambridge, 22–25 Aug 1977, pp 888–893
48. Wang X (1994) Learning by observation and practice: A framework for automatic acquisition of planning operators. In: AAAI
49. Wang X (1994) Learning planning operators by observation and practice. In: International Conference on AI Planning Systems (AIPS-94), Chicago, June 1994. AAAI Press, Mento Park, pp 335–340
50. Watkins CJCH, Dayan P (1992) Q-learning. Mach Learn 8(3–4):279–292
51. Xu K, Munoz-Avila H (2005) A domain-independent system for case-based task decomposition without domain theories. In: AAAI, Pittsburgh, July 2005. AAAI Press, Menlo Park, pp 234–240
52. Yang Q, Wu K, Jiang Y (2005) Learning actions models from plan examples with incomplete knowledge. In: International Conference on Automated Planning and Scheduling (ICAPS-05), Monterey, June 2005. AAAI Press Menlo Park, pp 241–250
53. Yang Q, Wu K, Jiang Y (2007) Learning action models from plan examples using weighted max-sat. Artif Intell 171(2–3): 107–143

Learning, System Identification, and Complexity

M. VIDYASAGAR
Tata Consultancy Services, Software Units Layout, Hyderabad, India

Article Outline

Glossary
Definition of the Subject
Introduction
Least Squares Curve Fitting – A Precursor
System Identification
Statistical Learning Theory
The Complexity of Learning
System Identification as a Learning Problem
Future Directions
Acknowledgment
Bibliography

Glossary

Accuracy A significant parameter in statistical learning theory that describes the error between the current estimate and the truth.

Algorithmic complexity The computational complexity of computing a hypothesis in a learning problem, on the basis of observations.

ARMA Auto-Regressive Moving Average. It is a specific kind of model for a dynamical system.

Confidence An estimate of the confidence with which we can make a particular statement, that is, a bound on the probability that the statement is false.

PAC Probably Approximately Correct. It is a widely used model for statistical learning.

Sample complexity The number of samples required to learn to a specific accuracy and confidence.

Vapnik–Chervonenkis dimension A combinatorial parameter that captures the "richness" of a set of concepts to be learnt.

Definition of the Subject

Definitions

$A \Delta B$ Symmetric difference between the sets A and B.

A_m Algorithm to be applied to m labeled inputs.

C A concept class, that is, a collection of subsets.

$d_P(T, H_m)$ Generalization error with the target set T and the hypothesis H_m.

H_m Hypothesis generated by the algorithm A_m.

$H_m(T; \mathbf{x})$ Hypothesis generated when the target set is T and the sample is $\mathbf{x}$.

$I_T(\cdot)$ Indicator function of the set T.

$J(\theta)$ Cost function (often, though not always, the sum of squares).

$M(\epsilon, C, d_P)$ ϵ-packing number of the concept class C with the pseudometric d_P.

$N(\epsilon, C, d_P)$ ϵ-covering number of the concept class C with the pseudometric d_P.

$r(m, \epsilon)$ Learning rate function; depends on the number of inputs m and the accuracy ϵ.

θ Parameter that characterizes the model used in identification.

$\hat{\theta}_t$ Estimate of θ at time t.

Θ Set of possible parameters θ.

$\mathbf{u}_t$ Input to the system to be identified at time t.

$\mathbf{y}_t$ Output of the system to be identified at time t.

System identification refers to the problem of fitting a model, from within a given set of models, to a series of observations (or measurements). By tradition, system identification is *recursive*, that is, the data is given one observation at a time, and ideally the next model is obtained as a function of the previously fit model and the current observation. One consequence of having a recursive identification procedure is that, once we have used the latest observation to update the model, the observation can be discarded, because the latest model embodies the effect of all the past observations.

Statistical learning is a closely related problem in which it is assumed that the data at hand is generated by a "true but unknown" target concept (or function), and the objective is to fit a model to the data at hand. Unlike in system identification, statistical learning is generally not recursive. In other words, at any given point in time, *the entire set of past observations* is often used to do the model fitting.

Perhaps the major difference between system identification theory and statistical theory is that the latter is heavily focused on so-called "finite time estimates." In other words, by tradition most theorems in system identification theory are *asymptotic in nature*, and tell us what happens as the number of observations approaches infinity. In contrast, by its very nature, statistical learning theory is devoted to the derivation of explicit upper bounds on just far the current model is from the unknown target. These bounds are a function of the number of observations, and the desired accuracy and confidence parameters (these are defined precisely later on).

Introduction

One of the earliest known examples of system identification is "curve fitting," which is the problem of fitting

a curve from within a given family to set of points in the plane (or some higher dimensional space). The problem of "curve fitting" goes back to antiquity, with some of the major contributions being made two centuries ago by Gauss and Laplace. System identification is a natural extension of curve fitting to the situation where the models are not "static" functions, but rather dynamical systems. The dynamical nature of the models introduces additional complexities into the theory.

In a learning problem, the issue of "complexity" is central, whether it is "sample complexity" or "algorithmic complexity." The former refers to the number of samples needed to achieve a desired level of accuracy and confidence, and the latter refers to the difficulty of computing a model on the basis of the given set of observations. Both types of complexity are studied in this article.

The difference between system identification and statistical learning is one of emphasis more than anything else. Thus it is not surprising that it is possible to view system identification as a problem in statistical learning. In this article, this connection is brought out explicitly.

We begin the article by reviewing the problem of curve fitting using a least squares optimization criterion, and then show how the basic ideas can be extended to the problem of system identification. Then we introduce the problem of PAC (probably approximately correct) learning. Then we discuss the issues of complexity. Then we show how identification can be viewed as a kind of learning problem. We conclude with some directions for future research.

Least Squares Curve Fitting – A Precursor

Many of the issues that underlie system identification can be readily understood in the context of the much simpler problem of "curve fitting" using a least squares error criterion.

Suppose we are given a series of pairs of vectors $(\mathbf{x}_i, \mathbf{y}_i)$, $i = 1, \ldots, l$, where $\mathbf{x}_i \in \mathbb{R}^n, \mathbf{y}_i \in \mathbb{R}^m \forall i$. The objective is to fit the data with a linear relationship of the form $\mathbf{y} = A\mathbf{x}$, where $A \in \mathbb{R}^{m\times n}$, such that the least squares error criterion

$$J(A) = \sum_{i=1}^{l} \| \mathbf{y}_i - A\mathbf{x}_i \|_2^2$$

is minimized, where $\| \cdot \|_2$ denotes the Euclidean norm

$$\| \mathbf{v} \|_2 = (\mathbf{v}^t\mathbf{v})^{1/2} .$$

Using the notation

$$Y := [\mathbf{y}_1 \ldots \mathbf{y}_l] \in \mathbb{R}^{m\times l}, \; X := [\mathbf{x}_1 \ldots \mathbf{x}_l] \in \mathbb{R}^{n\times l} ,$$

we can easily establish that the optimal choice of the matrix A is given by

$$A^* = CD^{-1} , \tag{1}$$

where

$$C = YX^t , \quad D = XX^t .$$

This formula presupposes that the matrix D is nonsingular and therefore invertible. Since

$$D = XX^t = \sum_{i=1}^{l} \mathbf{x}_i\mathbf{x}_i^t ,$$

it is easy to see that D is nonsingular if and only if the set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_l\}$ contains n linearly independent vectors. This leads to the first observation, namely: Unless there is sufficient data, the curve-fitting problem does not have a unique solution.

Now suppose that the data is in fact generated by a "true but unknown" linear relationship, where the data is corrupted by noise. In other words, suppose that

$$\mathbf{y}_i = A_0\mathbf{x}_i + \mathbf{v}_i ,$$

where A_0 is the "true" matrix, and $\{\mathbf{v}_i\}$ is a sequence of random vectors that are pairwise independent. For the moment, suppose that the vectors $\mathbf{x}_i$ are deterministic and not random. Then the measurements are described by

$$Y = AX + V ,$$
$$\text{where } \; V = [\mathbf{v}_1 \ldots \mathbf{v}_l] .$$

Now from this formula and (1), the following facts follow readily.

1. Since the measurement noise vectors are random, the optimal choice of A after l measurements, call it A_l, is also random.
2. If the measurement noise vectors $\mathbf{v}_i$ are all zero mean, then the optimal choice A_l is *unbiased*, that is, the expected value of A_l equals the true matrix A_0 for each value of l.
3. If the measurement noise vectors are normally distributed around zero, then A_l is normally distributed around the true value A_0.
4. Under very broad conditions, for example if the noise vectors are a sequence of i.i.d. (independent, identically distributed) vectors with finite variance, then the estimate A_l converges almost surely to the true value A_0 as $l \to \infty$.

We shall see that many of the same features are present in the system identification problem as well.

System Identification

System identification refers to the problem of fitting a dynamical system model to a given sequence of temporal (i. e., time-spaced) measurements. Thus one is given a set of input-output pairs $\{(u_t, y_t), t \geq 0\}$, as well as a family of models $\{\boldsymbol{h}(\theta), \theta \in \Theta\}$. The idea is to choose the "best possible" model from within this family to represent the data.

System identification is a vast and well-developed subject. In a brief essay such as this one, it is difficult to do more than to scratch the surface of the topic. The reader is referred to the book [9] for a very authoritative and comprehensive treatment.

Problem Formulation

Suppose one is given a sequence of input-output measurements $\{(\mathbf{u}_t, \mathbf{y}_t), t \geq 0\}$, where $\mathbf{u}_t$ is the input to the system at time t, and $\mathbf{y}_t$ is the output of the system at time t. The family of models $\{h(\theta), \theta \in \Theta\}$ can consist of linear recursive relationships, or nonlinear mappings. Perhaps the simplest class of models is the so-called ARMA (Auto-Regressive Moving Average) model, in the form

$$\mathbf{y}_t = \sum_{i=1}^{k} A_i(\theta)\mathbf{y}_{t-i} + B_i(\theta)\mathbf{u}_{t-i} \,, \tag{2}$$

where $A_i(\theta), B_i(\theta)$ are matrices of appropriate dimension that depend on the parameter $\theta \in \Theta$, where Θ is a subset of some Euclidean space. Note that the input $\mathbf{u}_t$ is not necessarily "white" noise, but could be a combination of a deterministic input and "colored" noise. The relationship (2) can be rewritten in the form

$$[I - A(\theta) - B(\theta)]\phi_t = \mathbf{0} \,, \tag{3}$$

where

$$A(\theta) = [A_1(\theta) \ldots A_k(\theta)],$$

$$B(\theta) = [B_1(\theta) \ldots B_k(\theta)], \phi_t = \begin{bmatrix} \mathbf{y}_t \\ \mathbf{y}_{t-1} \\ \vdots \\ \mathbf{y}_{t-k} \\ \mathbf{u}_{t-1} \\ \vdots \\ \mathbf{u}_{t-k} \end{bmatrix} .$$

To estimate the parameter θ_t at time t, it is common to choose it so as to minimize the accumulated least-squares error. Since we wish to fit a model of the form (3) to the data, the vector

$$\begin{aligned} \mathbf{e}_l(\theta) &= [I - A(\theta) - B(\theta)]\phi_l \\ &= \mathbf{y}_l - \sum_{i=1}^{k} A_i(\theta)\mathbf{y}_{l-i} + B_i(\theta)\mathbf{u}_{l-i} \end{aligned}$$

gives a goodness of fit of the model to the data, when the parameter is chosen as θ. Therefore we can define the cost function at time t as

$$\begin{aligned} J_t(\theta) &= \frac{1}{t}\sum_{l=k}^{t} \| \mathbf{e}_t \|_2^2 \\ &= \frac{1}{t}\sum_{l=k}^{t} \left\| \left[\mathbf{y}_l - \sum_{i=1}^{k} A_i(\theta)\mathbf{y}_{l-i} + B_i(\theta)\mathbf{u}_{l-i} \right] \right\|_2^2 \end{aligned}$$

as a scalar measure of how good a model parameter θ is. It is easy to see that the normalizing factor $1/t$ is inserted to ensure that the quantity J_t remains well behaved as $t \to \infty$. Thus it is logical to choose the estimate $\hat{\theta}_t$ such that

$$\hat{\theta}_t = \arg\min_{\theta \in \Theta} J_t(\theta) \,.$$

Note that $\hat{\theta}$ is our estimate of the parameter θ, based on the observations up to time t.

The above description applies when the model is a linear ARMA model. However, the approach itself is far more general. Define the vector ϕ_t as above, and suppose that the family of models consists of a set of relationships of the form

$$\mathbf{f}(\theta, \phi_t) = \mathbf{0} \,.$$

If $\mathbf{f}(\theta, \cdot)$ is a linear relationship of the form (3) for each choice of the parameter $\theta \in \Theta$, then we get the familiar parametrized ARMA model. As before, given a sequence of observations, the error at time t is defined as

$$\mathbf{e}_t(\theta) = \mathbf{f}(\theta, \phi_t) \,,$$

while the accumulated least squares error criterion is defined as before, as

$$J_t(\theta) = \frac{1}{t}\sum_{l=k}^{t} \| \mathbf{e}_t \|_2^2 \,.$$

In case the models are linear, choosing the least-squares error criterion leads to a very elegant solution of the optimization problem. But in case the models are not linear, there is no particular advantage in choosing the least-squares error criterion. Instead, one can choose some "loss" function $\ell \colon \mathbb{R}_+ \to \mathbb{R}_+$ such that $\ell(\cdot)$ is monoton-

ically increasing, and choose

$$J_t(\theta) = \frac{1}{t}\sum_{l=k}^{t} \ell(\| \mathbf{e}_t \|_2)$$

as the cost function at time t. Once again, the estimate $\hat{\theta}_t$ is chosen as the minimizer of $J_t(\theta)$ as θ varies over Θ.

Suppose now that the data comes from a "true but unknown" system whose output measurements are corrupted by measurement noise. Specifically, suppose it is the case that

$$\mathbf{y}_t = \sum_{i=1}^{k} A_i(\theta_{\text{true}})\mathbf{y}_{t-i} + B_i(\theta_{\text{true}})\mathbf{u}_{t-i} + \mathbf{v}_t \ ,$$

where the measurement noise $\mathbf{v}_t$ is an i.i.d. zero-mean sequence. For convenience, let us define

$$\mathbf{y}_t^0 = \sum_{i=1}^{k} A_i(\theta_{\text{true}})\mathbf{y}_{t-i} + B_i(\theta_{\text{true}})\mathbf{u}_{t-i}$$

to be the corresponding output sequence of the same system, but without any measurement noise. Because the current output $\mathbf{y}_t$ depends not only on the current and past inputs $\mathbf{u}_{t-1}$ but also the past outputs $\mathbf{y}_{t-i}$, the difference $\mathbf{y}_t - \mathbf{y}_t^0$ is a function of not just the current measurement noise $\mathbf{v}_t$, but also of the past measurement noise vectors $\mathbf{v}_{t-i}$. As a result, even if the measurement noise $\mathbf{v}_t$ consists of independent random variables, the dynamical nature of the system ensures that the difference between the actual measured output $\mathbf{y}_t$ and the ideal, noiseless measurement $\mathbf{y}_t^0$ is a sequence of *dependent* random variables. In other words, the dynamical nature of the system ensures that we cannot any longer make an analogy with the "static" least squares curve-fitting problem of the previous section.

Behavior of Identification Algorithms

We have seen that a widely used approach to system identification is to formulate the error criterion $J_t(\theta)$, and at each instant of time, choose the model parameter $\hat{\theta}_t$ so as to minimize $J_t(\theta)$. As time goes on, this approach generates a sequence of estimated parameters $\{\hat{\theta}_t\}$. Now we can ask three successively stronger questions, as follows:

First, as more and more data is provided to the identification algorithm, does the estimation error between the outputs of the identified model and the actual time series approach the minimum possible estimation error achievable by any model within the given model class? To clarify this question, let us define the quantity $J(\theta)$ as the limit as $t \to \infty$ of the cost function $J_t(\theta)$. It is easy to see that the limit exists under very mild conditions. Now define

$$J^* = \min_{\theta \in \Theta} J(\theta)$$

to be the best possible error measure. The question that is being asked is: Does the quantity $J(\hat{\theta}_t)$ approach this minimum value? In other words, is the performance of the identification algorithm asymptotically optimal?

Second, does the identified model converge to "the set of best possible models" within the given model class? In other words, suppose that the minimum possible estimation error is achievable by one or more "best possible models," and let Θ^* the set of all θ such that $J(\theta) = J^*$. The question is: does the distance between the estimated vector $\hat{\theta}_t$ and the set Θ^* approach zero at $t \to \infty$? Note that the question, as posed, permits the sequence $\{\hat{\theta}_t\}$ not to have a limit in the conventional sense, but to keep bouncing around nearer and nearer the set Θ^*. The difference between the first question and the second question is this. The first question requires only that the scalar $J(\hat{\theta}_t)$ should converge to the optimal value J^*, while the second question requires that the estimated parameter vector $\hat{\theta}_t$ should converge to the set of minimizers Θ^*.

Third, assuming that the data is generated by a "true" system whose output is corrupted by measurement noise, does the identified model converge to the "true" system? In other words, if both the true system and the family of models are parametrized by a vector of parameters, does the estimated parameter vector converge to the true parameter vector? The difference between the second question and the third question is the following: In the second question, it is *not* assumed that the data is generated by a "true" system corresponding to a "true" parameter θ_{true} belonging to the set Θ. Accordingly, in the second question, all that we ask is that the sequence of estimates $\{\hat{\theta}_t\}$ should converge to the set of minimizers of the function J. In contrast, in the third question, we add the assumption that the data to which the identification algorithm is applied has a specific structure, namely that it is generated by a true but unknown parameter vector θ_{true}. The requirement in the third question is that the sequence of estimates $\{\hat{\theta}_t\}$ should converge to θ_{true} as $t \to \infty$.

From a technical standpoint, Questions 2 and 3 are easier to answer than Question 1. Since identification is often carried out recursively, the output of the identification algorithm is a sequence of estimates $\{\theta_t\}_{t \geq 1}$. Suppose that we are able to show that Question 1 has an affirmative answer, i. e., that $J(\theta_t) \to J^*$, where J^* denotes the minimum possible estimation error. In such a case, with some additional assumptions it is possible to answer both Questions 2 and 3 in the affirmative.

Traditionally a positive answer to Question 2 is assured by assuming that Θ is a *compact* set, which in turn ensures that the sequence $\{\theta_t\}$ contains at least one convergent subsequence. For convenience, let us relabel this subsequence again as $\{\theta_t\}$. If the answer to Question 1 is "yes," if θ^* is any limit point of the sequence, and if $J(\theta)$ is continuous (or at worst, lower semi-continuous) with respect to θ, then it readily follows that $J(\theta_t) \to J^*$. In other words, the model $h(\theta^*)$ is an "optimal" fit to the data among the family $\{h(\theta), \theta \in \Theta\}$. Coming now to Question 3, suppose θ_{true} is the parameter of the "true" system, and let h_{true} denote the "true" system. In order for Question 3 to have an affirmative answer, the true system h_{true} must belong to the model family $\{h(\theta), \theta \in \Theta\}$; otherwise we cannot hope that $h(\theta_t)$ will converge to h_{true}. The traditional way to ensure that $\theta_{\text{true}} = \theta^*$ is to assume that the input to the true system is "persistingly exciting" or "sufficiently rich," so that the only way for $h(\theta^*)$ to match the performance of h_{true} is to have $\theta^* = \theta_{\text{true}}$.

Because of the arguments presented above, the main emphasis in system identification theory has been to study conditions to ensure that Question 1 has an affirmative answer, i. e., that the identification algorithm is asymptotically optimal. This question can be readily addressed using the following "universal" approach. We have defined $J(\theta)$ as the limit as $t \to \infty$ of the quantity $J_t(\theta)$. As remarked earlier, this limit exists under very mild conditions, *for each fixed* θ. Thus it can be said without much fuss that $J_t(\theta) \to J(\theta)$ as $t \to \infty$, again *for each fixed* θ. However, if a stronger property holds, namely that $J_t(\theta) \to J(\theta)$ as $t \to \infty$, *uniformly with respect* θ, then Question 1 has an affirmative answer. This is the approach adopted in [8]. In many ways, this approach to answering Question 1 is the common thread that binds system identification theory to learning theory. An alternate approach to Question 1 based on ergodic theory can be found in [5,6].

The fourth question that can be asked is: Suppose that the estimated parameter sequence $\{\hat{\theta}_t\}$ converges to the true parameter θ_{true}. As we have already seen, since the data that acts as the input to the identification algorithm is generated at random, the resulting estimates $\hat{\theta}_t$ are also random. Thus we may wish to know, not merely that the random variable $\hat{\theta}_t$ converges to θ_{true} in some probabilistic sense (say in probability, or almost surely), but also, the asymptotic distribution of the estimate $\hat{\theta}_t$ around the true value θ_{true}. To make this point clear, let us return to the linear least-squares curve-fitting problem. We know that if the measurement noise consists of i.i.d. gaussian variables, then the optimal estimate also has a gaussian distribution centered around the true value, and that the variance around the true value goes to zero at a rate that can be quantified fairly precisely. When we address this question in the context of dynamical systems however, the situation is considerably more complex, and one is obliged to use very sophisticated mathematics in order to analyze the asymptotic behavior of the estimated parameter sequence. Nevertheless, it is possible to say something about the asymptotic behavior of the estimates. The reader is invited to consult Chap. 9 of [9] for details.

To keep the exposition simple, we have discussed only one possible approach to system identification. Minimizing some kind of generalized least squares error criterion is only one way to choose an estimated parameter $\hat{\theta}_t$. There are other popular methods, such as the "instrument variables" method. In the setting of linear systems especially, it is possible to evolve quite different approaches based on frequency domain (instead of time domain) methods, based on correlating the input and output variables. Here again, the book [9] is an excellent reference. Another excellent reference is [10].

Statistical Learning Theory

Problem Formulation

A parallel theme to the problem of system identification is statistical learning. This kind of learning theory comes in many varieties, and the type of learning studied here is known as Probably Approximately Correct (PAC) learning theory. Even within PAC learning theory, there are two aspects, namely: statistical and computational. The present section discusses the statistical aspects of PAC learning, while the next section discusses the computational aspects.

Suppose one is trying to identify an unknown set on the basis of observation. The notion of PAC (probably approximately correct) learnability gives mathematical formalization of this intuitive idea. Suppose X is a set, and that C is a collection of subsets of X. Suppose also that P is a probability measure on the set X, which may or may not be known ahead of time to the learner. The learning problem is formulated as follows: There is a fixed but unknown concept $T \in C$, called the **target** concept. The objective is to "learn" the target concept on the basis of observation, consisting of i.i.d. samples $x_1, \ldots, x_m \in X$ drawn in accordance with P. For each sample x_j, an "oracle" tells the learner whether or not $x_j \in T$; equivalently, the oracle returns the value of $I_T(x_j)$, where $I_T(\cdot)$ is the indicator function of T. Thus, after m samples have been drawn, the information available to the learner consists of the "labeled multisample"

$$[(x_1, I_T(x_1)), \ldots, (x_m, I_T(x_m))] \in [X \times \{0,1\}]^m .$$

The objective is to construct a suitable approximation to the unknown target concept T on the basis of the labeled sample, using an appropriate algorithm. For the purposes of the present discussion, an "algorithm" is merely an indexed family of maps $\{A_m\}_{m\geq 1}$, where

$$A_m : [X \times \{0,1\}]^m \to C .$$

Thus an algorithm merely maps sets of labeled multisamples into elements of the concept class C. Suppose m i.i.d. samples have been drawn, and define

$$H_m(T;\mathbf{x}) = A_m[(x_1, I_T(x_1)), \ldots, (x_m, I_T(x_m))] .$$

The set $H_m(T;\mathbf{x})$ is referred to as the **hypothesis** generated by the algorithm. When there is no danger of confusion, one can abbreviate $H_m(T;\mathbf{x})$ by H_m.

In order to analyze how close the hypothesis H_m is to the unknown target set T, we need a *quantitative* measure of the disparity. Given two sets $A, B \subseteq X$, define their symmetric difference $A\Delta B$ by

$$A\Delta B = (A \cap B^c) \cup (A^c \cap B) = (A \cup B) - (A \cap B) .$$

Thus $A\Delta B$ consists of those points that belong to precisely one of the sets (and not the other). Now define the "generalization error"

$$d_P(T, H_m) = P(T\Delta H_m) .$$

Then it is easy to see that $d_P(T, H_m)$ is precisely the probability that a randomly chosen test sample is *misclassified* by the hypothesis H_m.

Roughly speaking, the algorithm $\{A_m\}$ can be said to "learn" the target concept T if the distance $d_P(T, H_m)$ approaches zero as $m \to \infty$. However, since the samples $\mathbf{x}_i$ are random, the resulting hypothesis H_m is also "random" and the generalization error $d_P(T, H_m)$ is also random. Define

$$r(m,\epsilon) = \sup_{T\in C} P^m\{\mathbf{x} \in X^m : d_P[T, H_m(T;\mathbf{x})] > \epsilon\} . \quad (4)$$

The quantity $r(m,\epsilon)$ is called the "learning rate" function.

Definition 1 The algorithm $\{A_m\}$ is said to be **probably approximately correct (PAC)** if $r(m,\epsilon) \to 0$ as $m \to \infty$ for each $\epsilon > 0$. The concept class C is said to be **PAC learnable** if there exists an algorithm that is PAC.

Thus the algorithm $\{A_m\}$ is PAC if, for each target concept T, the distance $d_P(T, H_m)$ converges to zero in probability as $m \to \infty$, and moreover, the rate of convergence is *uniform* with respect to the unknown target concept T.

Note that, if an algorithm is PAC, then for each pair of numbers $\epsilon, \delta > 0$, there exists a corresponding integer $m_0 = m_0(\epsilon, \delta)$ such that

$$P^m\{\mathbf{x} \in X^m : d_P[T, H_m(T;\mathbf{x})] > \epsilon\} \leq \delta , \quad \forall m \geq m_0 , \ \forall T \in C . \quad (5)$$

In other words, if at least m_0 i.i.d. samples are drawn and the algorithm is applied to the resulting labeled multisample, it can be said with confidence at least $1-\delta$ that the hypothesis H_m produced by the algorithm is within a distance ϵ of the true but unknown target concept T. In the above expression, the quantity ϵ is referred to as the "accuracy" of the algorithm, while δ is referred to as the "confidence." The integer $m_0(\epsilon, \delta)$ is called the "sample complexity" of the algorithm. In all but the simplest cases, one can only obtain upper bounds for $m_0(\epsilon, \delta)$, but this is sufficient.

The above statement provides some motivation for the nomenclature "PAC." If the algorithm $\{A_m\}$ is PAC, the hypothesis H_m produced by the algorithm need not *exactly* equal T; rather, H_m is only an *approximately* correct version of T, in the sense that $d_P(H_m, T) \leq \epsilon$. Even this statement is only *probably* true, with a confidence of at least $1-\delta$.

The notion of PAC learnability is usually credited to Valiant (see [12]), who showed that a certain collection of Boolean formulas is PAC learnable; however, the term "PAC learnability" seems to be due to [1]. It should be mentioned that Valiant had some additional requirements in his formulation beyond the condition that $r(m,\epsilon) \to 0$ as $m \to \infty$. In particular, he required that the mapping A_m be computable in polynomial time. These additional requirements are, in some sense, what distinguish computational (PAC) learning theory from statistical (PAC) learning theory. The references on computational learning theory in the bibliography go into computational issues in greater depth.

PAC learning is closely related to an older topic in probability theory known as empirical process theory, which is devoted to a study of the behavior of empirical means as the number of samples becomes large. One of the standard references in the field, namely [13], is actually about empirical process theory. In the book [15] PAC learning theory is treated side by side with the uniform convergence problem, and it is shown that each has something to offer to the other. However, in the present write-up we do not discuss the uniform convergence problem. Finally, the book [16] removes the assumption that the learning samples must be independent, and replaces it with the weaker assumption that the learning samples form a so-called mixing process.

Conditions for PAC Learnability

In this section we summarize the main results that provide necessary and sufficient conditions for a collection of sets to be PAC learnable. A central role is played by a combinatorial parameter known as the Vapnik–Chervonenkis (VC)-dimension. This notion is implicitly contained in the paper [14].

Definition 2 Let $\mathcal{A}$ be a collection of subsets of X. A set $S = \{x_1, \ldots, x_n\} \subseteq X$ is said to be **shattered** by $\mathcal{A}$ if, for every subset $B \subseteq S$, there exists a set $A \in \mathcal{A}$ such that $S \cap A = B$. The **Vapnik–Chervonenkis dimension** of $\mathcal{A}$, denoted by VC-dim($\mathcal{A}$), equals the largest integer n such that there exists a set of cardinality n that is shattered by $\mathcal{A}$.

One can think of the VC-dimension of $\mathcal{A}$ as a measure of the "richness" of the collection $\mathcal{A}$. A set S is "shattered" by $\mathcal{A}$ if $\mathcal{A}$ is rich enough to distinguish between all possible subsets of S. Note that S is shattered by $\mathcal{A}$ if (and only if) one can "pick off" every possible subset of S by intersecting S with an appropriately chosen set $A \in \mathcal{A}$. In this respect, perhaps "completely distinguished" or "completely discriminated" would be a better term than "shattered." However, the latter term has by now become standard in the literature. Note that, if $\mathcal{A}$ has finite VC-dimension, say d, then it is not rich enough to distinguish all subsets of *any* set containing $d + 1$ elements or more; but it is rich enough to distinguish all subsets of *some* set containing d elements (but not necessarily *all* sets of cardinality d).

From the definition it is clear that, in order to shatter a set S of cardinality n, the concept class $\mathcal{A}$ must contain at least 2^n distinct concepts. Thus an immediate corollary of the definition is that every finite collection of sets has finite VC-dimension, and that

$$\text{VC-dim}(\mathcal{A}) \leq \lg(|\mathcal{A}|) .$$

The importance of the VC-dimension in PAC learning theory arises from the following theorem, which is due to Blumer et al.; see [3]. In the statement of the theorem, we use the notion of a "consistent" algorithm. An algorithm is said to be "consistent" if it perfectly reproduces the training data. In other words, an algorithm is consistent if, for every labeled multisample $((x_1, I_T(x-1), \ldots, (x_m, I_T(x_m))$, we have that

$$I_{H_m}(x_i) = I_T(x_i) \forall i ,$$

where H_m is the hypothesis produced by the algorithm.

Theorem 1 *Suppose the concept class C has finite VC-dimension; then C is PAC-learnable for every probability measure over X. Let d denote the VC-dimension of C, and let $\{A_m\}$ be any consistent algorithm. Then the algorithm is PAC; moreover, the quantity $r(m, \epsilon)$ defined in (4) is bounded by*

$$r(m, \epsilon) \leq 2 \left(\frac{2em}{d} \right)^d 2^{-m\epsilon/2} , \tag{6}$$

for every probability measure over X. Any consistent algorithm learns to accuracy ϵ and confidence δ provided at least

$$m \geq \max \left\{ \frac{8d}{\epsilon} \lg \frac{8e}{\epsilon}, \frac{4}{\epsilon} \lg \frac{2}{\delta} \right\} \tag{7}$$

samples are used.

Thus Theorem 1 shows that the finiteness of the VC-dimension of the concept class C is a sufficient condition for so-called "distribution-free" PAC learnability. Moreover, (7) gives an estimate of the sample complexity, that is, the number of samples that suffice to achieve an accuracy of ϵ and a confidence of δ. Note that the sample complexity estimate is $O(1/\epsilon \log(1/\epsilon))$ for fixed δ, and $O(\log(1/\delta))$ for fixed ϵ. This is often summarized by the statement "confidence is cheaper than accuracy."

There is a converse to Theorem 1. It can be shown that, in order for a concept class C to be PAC learnable for *every* probability measure on the set X, and for the learning rate to be uniformly bounded with respect to the underlying probability measure, the concept class C must have finite VC-dimension.

Theorem 1 provides not only a condition for PAC-learnability but also relates the learning rate to the VC-dimension of the concept class. Thus it is clearly worthwhile to determine upper bounds for the VC-dimension of various concept classes. Over the years, several researchers have studied this problem, using a wide variety of approaches, ranging from simple counting arguments to extremely sophisticated mathematical reasoning based on advanced ideas such as algebraic topology, model theory of real numbers, and so on. Since the list of relevant references would be too long, the reader is referred instead to Chap. 10 of [15,16] which contain a derivation of many of the important known results together with citations of the original sources.

The Complexity of Learning

In this section, we discuss briefly the complexity of a learning problem. Complexity can arise in two ways: sample complexity, and algorithmic complexity. Sample complexity refers to the rate at which the number $m_0(\epsilon, \delta)$ grows with respect to $1/\epsilon$ and $1/\delta$, where $m_0(\epsilon, \delta)$ is the smallest value of m_0 such that (5) holds. Algorithmic complexity refers to the complexity of computing the hypothesis H_m

as a function of the number of samples of m, Each kind of complexity has its own interesting features.

Sample Complexity

Let us begin by discussing the notions of covering numbers, and packing numbers. Given a concept class C consisting of subsets of a given set X and a probability measure P on X, let us observe that the quantity d_P defines a pseudometric on the collection of (measurable) subsets of X, as follows: Given two subsets $A, B \subseteq X$, define

$$d_P(A, B) = P(A \Delta B) ,$$

where $A \Delta B$ is the symmetric difference between the sets A and B. Note that d_P is only a pseudometric because $d_P(A, B) = 0$ whenever $A \Delta B$ is a set of measure zero – it need not be the empty set (meaning that $A = B$). However, this is a minor technicality, and we can still refer to d_P as defining a "distance" between subsets of X.

Given the concept class C, the probability measure P (which in turn defines the pseudometric d_P), and a number ϵ, we say that a finite collection $\{C_1, \ldots, C_n\}$ where each $C_i \in C$ is an **ϵ-cover** of C if, for every $A \in C$, we have $d_P(A, C_i) \leq \epsilon$ for at least one index i. To put it in other words, the collection $\{C_1, \ldots, C_n\}$ is an ϵ-cover of C under the pseudometric d_P if every set $A \in C$ is within a distance of ϵ from at least one of the sets C_i. The **ϵ-covering number** $N(\epsilon, C, d_P)$ is defined to be the smallest number n such that there exists an ϵ-cover of cardinality n. Note that, in some cases, an ϵ-cover may not exist, in which case we take the ϵ-covering number to be infinity.

The next, and related, notion is the ϵ-packing number of a concept class. A finite collection $\{C_1, \ldots, C_m\}$ where each $C_i \in C$ is said to be **ϵ-separated** if $d_P(C_i, C_j) > \epsilon$ $\forall i \neq j$. In other words, the finite collection $\{C_1, \ldots, C_m\}$ is ϵ-separated if the concepts are pairwise at a distance of more than ϵ from each other. The **ϵ-packing number** $M(\epsilon, C, d_P)$ is defined to be the largest number m such that there exists an ϵ-separated collection of cardinality m. It is not difficult to show that

$$\begin{aligned} M(2\epsilon, C, d_P) &\leq N(\epsilon, C, d_P) \\ &\leq M(\epsilon, C, d_P) \quad \forall \epsilon > 0 . \end{aligned}$$

See for instance Lemma 2.2 of [16].

The next two theorems, due to Benedek and Itai [2], are fundamental results on sample complexity.

Theorem 2 *Suppose C is a given concept class, and let $\epsilon > 0$ be specified. Then C is PAC-learnable to accuracy provided the covering number $N(\epsilon/2, C, d_P)$ is finite. Moreover, there exists an algorithm that is PAC with sample complexity*

$$m_0(\epsilon, \delta) \leq \frac{32}{\epsilon} \ln \frac{N(\epsilon/2, C, d_P)}{\delta} .$$

Thus a concept class with a finite covering number $N(\epsilon/2, C, d_P)$ is learnable to accuracy ϵ. The next theorem provides a kind of converse.

Theorem 3 *Suppose C is a given concept class, and let $\epsilon > 0$ be specified. Then any algorithm that is PAC to accuracy ϵ requires at least $\lg M(2\epsilon, C, d_P)$ samples, where $M(2\epsilon, C, d_P)$ denotes the 2ϵ-packing number of the concept class C with respect to the pseudometric d_P.*

One consequence of the theorem is that a concept class C is learnable to accuracy ϵ only if $M(2\epsilon, C, d_P)$ is finite. This fact is used to construct examples of "non-learnable" concept classes; see for instance Example 6.10 in [16]. However, the question of interest here is whether there exists a concept class that is learnable, but for which the sample complexity $m_0(\epsilon, \delta)$ grows faster than any polynomial in $1/\epsilon$.

In Sect. 3.3 in [7], the author constructs a concept class with the property that

$$M(1/n, C, d_P) \geq 2^{2^{c \cdot n^{1/4}}} ,$$

where c is some constant. Thus it follows from Theorem 3 that the sample complexity satisfies

$$m_0(\epsilon/2, \delta) \geq 2^{c \cdot (1/\epsilon)^{1/4}} , \quad \forall \epsilon .$$

So it is after all possible to have a situation where a concept class is learnable, but the sample complexity grows faster than any polynomial in the reciprocal accuracy $1/\epsilon$.

Algorithmic Complexity

We have seen from Theorem 1 that if a concept class C has finite VC-dimension, then every *consistent* algorithm is PAC. In the original paper [12], one studies not a single concept class C, but rather an indexed family of concept classes $\{C_n\}$, for every integer n. The objective is to determine whether there exists an algorithm that PAC-learns each concept class, such that the computational complexity of learning C_n is polynomial in n. Note that we are *not* referring here to the sample complexity. In most problems in computer science, each concept class C_n is finite and therefore automatically has finite VC-dimension. Rather, we are referring to the difficulty of constructing a consistent hypothesis, and the computational complexity of this task as a function of n. In the original paper [12], it is shown that the problem of learning 3-CNF (3 conjunctive

normal forms) is PAC learnable in this more restrictive sense. Note that 3-CNF in n Boolean variables consists of all conjunctions of clauses in which each clause is a disjunction of no more than three Boolean variables or their negations. For instance, if $n = 5$, then

$$(x_1 \vee \neg x_2 \vee x_5) \wedge (x_2 \vee x_3 \vee \neg x_4) \wedge (\neg x_1 \vee \neg x_3) \wedge (\neg x_2 \vee \neg x_3 \vee \neg x_5)$$

belongs to 3-CNF. In this instance, Valiant shows that it is possible to construct a consistent hypothesis in polynomial time in n.

Shortly after the publication of this paper, a very simple example was produced of a learning problem in which it is NP-hard to construct a consistent hypothesis. Specifically, consider the family of all conjunctions of at most k clauses, where each clause is a disjunction of various literals or their negations (without limit). For instance, if $n = 5$, then

$$(x_1 \vee \neg x_2 \vee \neg x_3 \vee x_5) \wedge (\neg x_1 \vee x_3 \vee \neg x_4 \vee \neg x_5) \wedge (\neg x_1 \vee x_2 \vee x_4 \vee \neg x_5)$$

is an example of this kind of formula. In this case, it can be shown that, for each fixed k and n, the VC-dimension of the concept class is polynomial in k and n. Thus it follows from Theorem 1 that the *sample complexity* of learning is polynomial. However, it is shown in [11] that the *computational complexity* of finding a consistent hypothesis is intractable unless $P = NP$. In contrast, if each such Boolean formula is expressed as a 3-DNF, then it is possible to construct a consistent hypothesis in polynomial time with respect to n.

System Identification as a Learning Problem

The Need for a Quantitative Identification Theory

By tradition, identification theory is *asymptotic* in nature. In other words, the theory analyzes what happens as the number of samples approaches infinity. However, there are situations where it is desirable to have *finite time*, or nonasymptotic, estimates of the rate at which the output of the identification process converges to the best possible model. For instance, in "indirect" adaptive control, one first carries out an identification of an unknown system, and after a finite amount of time has elapsed, designs a controller based on the current model of the unknown system. If one can quantify just how close the identified model is to the true but unknown system *after a finite number of samples*, then these estimates can be combined with robust control theory to ensure that indirect adaptive control leads to closed-loop stability.

Among the first papers to state that the derivation of finite-time estimates is a desirable property in itself are Weyer, Williamson and Mareels; see [20,21]. In those papers, it is assumed that the time series to which the system identification algorithm is applied consists of so-called "finitely-dependent" data; in other words, it is assumed that the time series can be divided into blocks that are independent. In a later paper [19], the author replaced the assumption of finite dependence by the assumption that the time series is β-mixing. However, he did not derive conditions under which a time series is β-mixing. In a still later paper [4], the authors study the case where the data is generated by general linear systems (Box–Jenkins models) driven by i.i.d. Gaussian noise, and the model family also belongs to the same class. In this paper, the authors say that they are motivated by the observation that "signals generated by dynamical systems are not β-mixing in general". This is why their analysis is limited to an extremely restricted class. However, their statement that dynamical systems do not generate β-mixing sequences is simply incorrect. Specifically, the all the systems studied in [4] are themselves β-mixing! In another paper [18], the present author has shown that *practically any exponentially stable system driven by i.i.d. noise with bounded variance* is β-mixing. Hence, far from being restrictive, the results of [19] have broad applicability, though he did not show it at the time.

Problem Formulation

System identification viewed as a learning problem can be stated as follows: One is given a time series $\{(u_t, y_t)\}$, where u_t denotes the input to the unknown system at time t, and y_t denotes the output at time t. We have that $u_t \in U \subseteq \mathbb{R}^l$, $y_t \in Y \subseteq \mathbb{R}^k$ for all t, where for simplicity it is assumed that both U and Y are compact sets. One is also given a family of models $\{h(\theta), \theta \in \Theta\}$ parametrized by a parameter vector θ belonging to a set Θ. Usually Θ is a subset of a finite-dimensional Euclidean space. At time t, the data available to the modeler consists of all the past measurements until time $t-1$. Based on these measurements, the modeler chooses a parameter θ_t, with the objective of making the best possible prediction of the next measured output y_t. The method of choosing θ_t is called the identification algorithm.

To make this formulation a little more precise, let us introduce some notation. Define $\mathcal{U} := \prod_{-\infty}^{\infty} U$, and define $\mathcal{Y}$ analogously. Note that the input sequence $\{u_t\}$ belongs to the set $\mathcal{U}$, while the output sequence belongs to $\mathcal{Y}$. Let $\mathcal{U}_{-\infty}^{0}$ denote the one-sided infinite cartesian product $\mathcal{U}_{-\infty}^{0} := \prod_{-\infty}^{0} U$, and for a given two-sided infinite se-

quence $\mathbf{u} \in \mathcal{U}$, define

$$\mathbf{u}_t := (u_{t-1}, u_{t-2}, u_{t-3}, \dots) \in \mathcal{U}^0_{-\infty} .$$

Thus the symbol $\mathbf{u}_t$ denotes the infinite past of the input signal $\mathbf{u}$ at time t. The family of models $\{h(\theta), \theta \in \Theta\}$ consists of a collection of maps $h(\theta), \theta \in \Theta$, where each $h(\theta)$ maps $\mathcal{U}^0_{-\infty}$ into Y. Thus, at time t, the quantity $h(\theta) \cdot \mathbf{u}_t =: \hat{y}_t(\theta)$ is the "predicted" output if the model parameter is chosen as θ. The quality of this prediction is measured by a "loss function" $\ell\colon Y \times Y \to [0, 1]$. Thus $\ell(y_t, h(\theta) \cdot \mathbf{u}_t)$ is the loss we incur if we use the model $h(\theta)$ to predict the output at time t, and the actual output is y_t.

Since we are dealing with a time series, all quantities are *random*. Hence, to assess the quality of the prediction made using the model $h(\theta)$, we should take the *expected value* of the loss function $\ell(y_t, h(\theta) \cdot \mathbf{u}_t)$ with respect to the law of the time series $\{(u_t, y_t)\}$, which we denote by $\tilde{P}_{\mathbf{u},\mathbf{y}}$. Thus we define the objective function

$$J(\theta) := E\left[\ell(y_t, h(\theta) \cdot \mathbf{u}_t), \tilde{P}_{\mathbf{u},\mathbf{y}}\right] . \tag{8}$$

Note that $J(\theta)$ depends solely on θ and nothing else.

A key observation at this stage is that the probability measure $\tilde{P}_{\mathbf{u},\mathbf{y}}$ is *unknown*. This is because, if the statistics of the time series are known ahead of time, then there is nothing to identify!

Definition 3 System Identification Problem: Given the time series $\{(u_t, y_t)\}$ with *unknown* law $\tilde{P}_{\mathbf{u},\mathbf{y}}$, construct if possible an iterative algorithm for choosing θ_t as a function of t, in such a way that

$$J(\theta_t) \to \inf_{\theta \in \Theta} J(\theta) .$$

A General Result

Since the law $\tilde{P}_{\mathbf{u},\mathbf{y}}$ of the time series is not known, the objective function $J(\theta)$ cannot be computed exactly. To circumvent this difficulty, one replaces the "true" objective function $J(\cdot)$ by an "empirical approximation," as defined next. For each $t \geq 1$ and each $\theta \in \Theta$, define the empirical error

$$\hat{J}_t(\theta) := \frac{1}{t} \sum_{i=1}^{t} \ell[y_i, h(\theta) \cdot \mathbf{u}_i] . \tag{9}$$

If we could choose $\ell(y, z) = \| y - z \|^2$, then

$$\hat{J}_t(\theta) = \frac{1}{t} \sum_{i=1}^{t} \| y_i - \hat{y}_i(\theta) \|^2$$

would be the average cumulative mean-squared error between the actual output y_i and the predicted error $\hat{y}_i(\theta)$, from time 1 to time t. Note that the least squares error is in general unbounded, whereas the loss function ℓ is supposed to map into the bounded interval $[0, 1]$. To circumvent this technical difficulty, we can choose

$$\hat{J}_t(\theta) = \frac{1}{t} \sum_{i=1}^{t} \eta\big(\| y_i - \hat{y}_i(\theta) \|^2 \big) ,$$

where $\eta(\cdot)$ is any kind of saturating nonlinear function, for instance $\eta(s) = \tanh(s)$. This corresponds to the choice $\ell(y, z) = \tanh(\| y - z \|^2)$.

Note that, unlike the quantity $J(\theta)$, the function $\hat{J}(\theta)$ *can* be computed on the basis of the available data. Hence, in principle at least, it is possible to choose θ_t so as to minimize the "approximate" (but computable) objective function $\hat{J}(\theta)$ in the hope that, by doing so, we will somehow minimize the "true" (but uncomputable) objective function $J(\theta)$.

The next theorem gives some sufficient conditions for this approach to work. Specifically, if the estimates $\hat{J}_t(\theta)$ converge *uniformly* to the correct values $J(\theta)$ as $t \to \infty$, where the uniformity is with respect to $\theta \in \Theta$, then the above approach works. The importance of this uniform convergence property was recognized very early on in identification theory. In particular, Caines [5,6] uses ergodic theory to establish this property, whereas Ljung [8] takes a more direct approach.

Theorem 4 *At time t, choose θ_t^* so as to minimize $\hat{J}_t(\theta)$; that is,*

$$\theta_t^* = \arg\min_{\theta \in \Theta} \hat{J}_t(\theta) .$$

Let

$$J^* := \inf_{\theta \in \Theta} J(\theta) .$$

Define the quantity

$$q(t, \epsilon) := \tilde{P}_{\boldsymbol{u},\boldsymbol{y}}\{\sup_{\theta \in \Theta} |\hat{J}_t(\theta) - J(\theta)| > \epsilon\} . \tag{10}$$

Suppose it is the case that $q(t, \epsilon) \to 0$ as $t \to \infty, \forall \epsilon > 0$. Then

$$\tilde{P}_{\boldsymbol{u},\boldsymbol{y}}\{J(\theta_t^*) > J^* + \epsilon\} \to 0 \quad \text{as } t \to \infty, \forall \epsilon > 0 . \tag{11}$$

In other words, the quantity $J(\theta_t^*)$ converges to the optimal value J^* in probability, with respect to the measure $\tilde{P}_{\boldsymbol{u},\boldsymbol{y}}$.

Corollary 1 *Suppose that $q(t, \epsilon) \to 0$ as $t \to \infty, \forall \epsilon > 0$. Given $\epsilon, \delta > 0$, choose $t_0 = t_0(\epsilon, \delta)$ such that*

$$q(t, \epsilon) < \delta \forall t \geq t_0(\epsilon, \delta) . \tag{12}$$

Then

$$\tilde{P}_{u,y}\{J(\theta_t^*) > J^* + \epsilon\} < \delta \forall t \geq t_0(\epsilon/3, \delta) . \quad (13)$$

Proofs of the theorem and corollary can be found in [17].

A Result on the Uniform Convergence of Empirical Means

The property of Theorem 4 does indeed hold in the commonly studied case where y_t is the output of a "true" system corrupted by additive noise, and the loss function ℓ is the squared error, provided the model family and the true system satisfy some fairly reasonable conditions, namely: each system is BIBO (bounded input, bounded output) stable, all systems in the model family have exponentially decaying memory, and finally, an associated collection of input-output maps has finite P-dimension. Note that the P-dimension is a generalization of the VC-dimension to the case of maps assuming values in a bounded interval $[0, 1]$. By identifying a set with its support function, we can think of the VC-dimension as a combinatorial parameter associated with binary-valued maps. The P-dimension is defined in Chap. 4 of [15,16]

Define the collection of functions $\mathcal{H}$ mapping $\mathcal{U}$ into $\mathbb{R}$ as follows:

$$g(\theta) := \mathbf{u} \mapsto \| (f - h(\theta)) \cdot \mathbf{u}_0 \|^2 : \mathcal{U} \to \mathbb{R} ,$$

$$\mathcal{G} := \{g(\theta) : \theta \in \Theta\} .$$

Now the various assumptions are listed.

A1. There exists a constant M such that

$$|g(\theta) \cdot \mathbf{u}_0| \leq M, \forall \theta \in \Theta, \mathbf{u} \in \mathcal{U} .$$

This assumption can be satisfied, for example, by assuming that the true system and each system in the family $\{h(\theta), \theta \in \Theta\}$ is BIBO stable (with an upper bound on the gain, independent of θ), and that the set U is bounded (so that $\{u_t\}$ is a bounded stochastic process).

A2. For each integer $k \geq 1$, define

$$g_k(\theta) \cdot \mathbf{u}_t := g(\theta) \cdot (u_{t-1}, u_{t-2}, \ldots, u_{t-k}, 0, 0, \ldots) . \quad (14)$$

With this notation, define

$$\mu_k := \sup_{\mathbf{u} \in \mathcal{U}} \sup_{\theta \in \Theta} |(g(\theta) - g_k(\theta)) \cdot u_0| .$$

Then the assumption is that μ_k is finite for each k and approaches zero as $k \to \infty$. This assumption essentially means that each of the systems in the model family has decaying memory (in the sense that the effect of the values of the input at the distant past on the current output becomes negligibly small).

A3. Consider the collection of maps $\mathcal{G}_k = \{g_k(\theta) : \theta \in \Theta\}$, viewed as maps from U^k into $\mathbb{R}$. For each k, this family $\mathcal{G}_k$ has finite P-dimension, denoted by $d(k)$.

Now we can state the main theorem.

Theorem 5 *Define the quantity $q(t, \epsilon)$ as in (10) and suppose Assumptions A1 through A3 are satisfied. Given an $\epsilon > 0$, choose $k(\epsilon)$ large enough that $\mu_k \leq \epsilon/4$ for all $k \geq k(\epsilon)$. Then for all $t \geq k(\epsilon)$ we have*

$$q(t, \epsilon) \leq 8k(\epsilon) \left(\frac{32e}{\epsilon} \ln \frac{32e}{\epsilon} \right)^{d(k(\epsilon))} \cdot \exp(-\lfloor t/k(\epsilon) \rfloor \epsilon^2/512M^2) , \quad (15)$$

where $\lfloor t/k(\epsilon) \rfloor$ denotes the largest integer part of $t/k(\epsilon)$.

Theorem 6 *Let all notation be as in Theorem 5. Then, in order to ensure that the current estimate θ_t satisfies the inequality $J(\theta_t) \leq J^* + \epsilon$ (i. e., is ϵ-optimal) with confidence $1 - \delta$, it is enough to choose the number of samples t large enough that*

$$\lfloor t/k(\epsilon) \rfloor \geq \frac{512M^2}{\epsilon^2} \left[\ln \left(\frac{24k(\epsilon)}{\delta} \right) + d(k(\epsilon)) \ln \left(\frac{32e}{\epsilon} \right) + d(k(\epsilon)) \ln \ln \left(\frac{32e}{\epsilon} \right) \right] . \quad (16)$$

Bounds on the P-Dimension

In order for the estimate in Theorem 5 to be useful, it is necessary for us to derive an estimate for the P-dimension of the family of functions defined by

$$\mathcal{G}_k := \{g_k(\theta) : \theta \in \Theta\} , \quad (17)$$

where $g_k(\theta) : U^k \to \mathbb{R}$ is defined by

$$g_k(\theta)(\mathbf{u}) := \| (f - h(\theta)) \cdot \mathbf{u}_k \|^2 ,$$

where

$$\mathbf{u}_k := (\ldots, 0, u_k, u_{k-1}, \ldots, u_1, 0, 0, \ldots) .$$

Note that, in the interests of convenience, we have denoted the infinite sequence with only k nonzero elements as $u_k, \ldots, u_1$ rather than $u_0, \ldots, u_{1-k}$ as done earlier. Clearly this makes no difference. In this section, we state and prove such an estimate for the commonly occurring case where each system model $h(\theta)$ is an ARMA model where the parameter θ enters linearly. Specifically, it is

supposed that the model $h(\theta)$ is described by

$$x_{t+1} = \sum_{i=1}^{l} \theta_i \, \phi_i(x_t, u_t), \;\; y_t = x_t \,, \tag{18}$$

where $\theta = (\theta_1, \ldots, \theta_l) \in \Theta \subseteq \mathbb{R}^l$, and each $\phi_i(\cdot,\cdot)$ is a polynomial of degree no larger than r in the components of x_t, u_t.

Theorem 7 *With the above assumptions, we have that*

$$\begin{aligned} P\text{-}dim(\mathcal{G}_k) &\le 9l + 2l \lg[2(r^{k+1} - 1)/(r-1)] \\ &\approx 9l + 2lk \lg(2r) \quad \text{if} \quad r > 1 \,. \end{aligned} \tag{19}$$

In case $r = 1$ so that each system is linear, the above bound can be simplified to

$$P\text{-}dim(\mathcal{G}_k) \le 9l + 2l \lg(2k) \,. \tag{20}$$

It is interesting to note that the above estimate is *linear* in both the number of parameters l and the duration k of the input sequence $\mathbf{u}$, but is only logarithmic in the degree of the polynomials ϕ_i. In the practically important case of linear ARMA models, even k appears inside the logarithm.

Future Directions

At present, system identification theory is a mature, widely used subject. Statistical learning theory is newer and still needs to find its feet. In particular, the various estimates for sample complexity given here are deemed to be "too conservative" to be of practical use. Thus considerable research needs to be carried out on the issue of sample complexity. The interplay between system identification theory and statistical learning theory perhaps offers the greatest scope for interesting new work. At present, statistical learning theory allows us to derive *finite time estimates* of the disparity between the current estimates of a system, and the true but unknown system. However, these estimates do not always translate readily into a metric "distance" between the true but unknown system and the current estimate. In order to be useful in the context of robust controller design, it is necessary to derive estimates for some kind of metric distance. This is an important topic for future research.

Acknowledgment

The author thanks Erik Weyer for his careful reading and comments on an earlier draft of this article.

Bibliography

Primary Literature

1. Angluin D (1987) Queries and concept learning. Mach Learn 2:319–342
2. Benedek GM, Itai A (1988) Learnability by fixed distributions. First Workshop on Computational Learning Theory. Morgan-Kaufmann, San Mateo, pp 80–90
3. Blumer A, Ehrenfeucht A, Haussler D, Warmuth M (1989) Learnability and the Vapnik–Chervonenkis dimension. J ACM 36(4):929–965
4. Campi MC, Weyer E (2002) Finite sample properties of system identification methods. IEEE Trans Auto Control 47:1329–1334
5. Caines PE (1976) Prediction error identification methods for stationary stochastic processes. IEEE Trans Auto Control AC-21:500–505
6. Caines PE (1978) Stationary linear and nonlinear system identification and predictor set completeness. IEEE Trans Auto Control AC-23:583–594
7. Hammer B (1999) Learning recursive data, Math of Control. Signals Syst 12(1):62–79
8. Ljung L (1978) Convergence analysis of parametric identification methods. IEEE Transa Auto Control AC-23:77–783
9. Ljung L (1999) System Identification: Theory for the User. Prentice-Hall, Englewood Cliffs
10. Pintelon R, Schoukens J (2001) System Identification: A frequency domain approach. IEEE press, Piscataway
11. Pitt L, Valiant L (1988) Computational limits on learning from examples. J ACM 35(4):965–984
12. Valiant LG (1984) A theory of the learnable. Comm ACM 27(11):1134–1142
13. Vapnik VN (1982) Estimation of Dependences Based on Empirical Data. Springer, Heidelberg
14. Vapnik VN, Chervonenkis AY (1971) On the uniform convergence of relative frequencies to their probabilities. Theory Prob Appl 16(2):264–280
15. Vidyasagar M (1997) A Theory of Learning and Generalization. Springer, London
16. Vidyasagar M (2003) Learning and Generalization: With Applications to Neural Networks and Control Systems. Springer, London
17. Vidyasagar M, Karandikar RL (2001) System identification: A learning theory approach. Proc IEEE Conf on Decision and Control, Orlando, pp 2001–2006
18. Vidyasagar M, Karandikar RL (2001) A learning theory approach to system identification and stochastic adaptive control. In: Calafiore G, Dabbene F (eds) Probabilistic and Randomized Algorithms for Design Under Uncertainty. Springer, Heidelberg, pp 265–302
19. Weyer E (2000) Finite sample properties of system identification of ARX models under mixing conditions. Automatica 36(9):1291–1299
20. Weyer E, Williamson RC, Mareels I (1996) Sample complexity of least squares identification of FIR models. Proc 13th World Congress of IFAC, San Francisco, pp 239–244
21. Weyer E, Williamson RC, Mareels I (1999) Finite sample properties of linear model identification. IEEE Trans Auto Control 44(7):1370–1383

Books and Reviews

Anthony M, Biggs N (1992) Computational Learning Theory. Cambridge University Press, Cambridge

Devroye L, Györfi L, Lugosi G (1996) A probabilistic theory of pattern recognition. Springer, New York

Haussler D (1992) Decision theoretic generalizations of the PAC

model for neural net and other learning applications. Info Comput 100:78–150
Kearns M, Vazirani U (1994) Introduction to Computational Learning Theory. MIT Press, Cambridge
Kulkarni SR, Vidyasagar M (2001) Learning decision rules under a family of probability measures. IEEE Trans Info Thy
Natarajan BK (1991) Machine Learning: A Theoretical Approach. Morgan-Kaufmann, San Mateo
van der Vaart AW, Wallner JA (1996) Weak convergence and Empirical Processes. Springer, New York
Vapnik AN, Chervonenkis AY (1981) Necessary and sufficient conditions for the uniform convergence of means to their expectations. Theory Prob Appl 26(3):532–553
Vapnik VN (1995) The Nature of Statistical Learning Theory. Springer, New York
Vapnik VN (1998) Statistical Learning Theory. Wiley, New York
Weyer E, Campi MC (2002) Non-asymptotic confidence ellipsoids for the least squares error estimate. Automatica 38:1529–1547

Levy Statistics and Anomalous Transport: Levy Flights and Subdiffusion

Ralf Metzler[1], Aleksei V. Chechkin[2], Joseph Klafter[3,4]
[1] Physics Department, Technical University of Munich, Garching, Germany
[2] Institute for Theoretical Physics NSC KIPT, Kharkov, Ukraine
[3] School of Chemistry, Tel Aviv University, Tel Aviv, Israel
[4] Freiburg Institute for Advanced Studies (FRIAS), University of Freiburg, Freiburg, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Lévy Flights
Subdiffusion and the Fractional Fokker–Planck Equation
Future Directions
Acknowledgments
Bibliography

Glossary

Anomalous diffusion Deviations of the linear time dependence $\langle x^2(t)\rangle = 2Kt$, of the mean squared displacement in absence of an external bias, in the form of a power-law: $\langle x^2(t)\rangle = 2K_\zeta t^\zeta$. Here, K_ζ is the anomalous diffusion constant of dimension $\mathrm{cm}^2/\mathrm{sec}^\zeta$. In the range $0 < \zeta < 1$, we deal with subdiffusion, whereas $\zeta > 1$ describes superdiffusion. Lévy flights have a diverging mean squared displacement.

Continuous time random walk A theory that describes a random motion by assigning each jump a jump length x and a waiting time t elapsing between two successive jumps, drawn from the two probability densities $\lambda(x)$ and $\psi(t)$, respectively. The two densities λ and ψ fully specify the probability density function $P(x,t)$ describing the random process. In Fourier–Laplace space, the propagator follows as $P(k,u) = u^{-1}\psi(u)/[1-\psi(u)\lambda(k)]$.

Fourier and Laplace transforms Integral transforms that aid in convenient solution of Linear partial differential equations. The definitions of the fractional operators used in the text correspond to Laplace and Fourier convolutions, such that these transformations also become useful there. The Laplace transform of a function $f(t)$ is defined as

$$f(u) \equiv \mathcal{L}\{f(t);u\} = \int_0^\infty f(t)\mathrm{e}^{-ut}\mathrm{d}t\,,$$

while the Fourier transform of $g(x)$ reads

$$g(k) \equiv \mathcal{F}\{g(x);k\} = \int_{-\infty}^\infty g(x)\mathrm{e}^{ikx}\mathrm{d}x\,.$$

Note that we denote the transform of a function by explicit dependence on the respective variable. For $u = 0$ and $k = 0$ the Laplace and Fourier transform are the averages of the function f or g, respectively. Tauberian theorems ascertain relations between the original function and its transform. For instance, the small u behavior $f(u) \sim 1-(u\tau)^\alpha$ implies long time scaling $f(t) \sim \tau^\alpha/t^{1+\alpha}$. See [35,46] for details.

Fractional differintegration The multiple derivative of an integer power is $\mathrm{d}^m x^n/\mathrm{d}x^m = n!/(n-m)!x^{n-m}$ for $m \le n$. The result is zero if $m > n$. Replacing the factorials by the Γ-function, one can generalize this relation to ${}_0D_x^q x^p = x^{p-q}\Gamma(1+p)/\Gamma(1+p-q)$. In particular, this includes the fractional differentiation of a constant, ${}_0D_x^q 1 = x^{-q}/\Gamma(1-q)$, that no longer vanishes. Fractional differentiation was first mentioned by Leibniz in a letter to de l'Hospital in 1695. The Riemann–Liouville and Riesz-Weyl fractional operators used in the following are a straightforward generalization of the Cauchy multiple integral, followed by regular differentiation.

Lévy flight A special type of continuous time random walk. A Lévy flight's waiting time distribution is narrow, for instance, Poissonian with $\psi(t) = \tau^{-1}\exp(-t/\tau)$, and the resulting dynamics therefore Markovian. The jump length distribution of a Lévy

flight is long-tailed: $\lambda(x) \sim \sigma^{\mu}|x|^{-1-\mu}$, with $0 < \mu < 2$, such that no second moment exists. The resulting PDF is a Lévy stable law with Fourier transform $P(k,t) = \exp(-K^{(\mu)}t|k|^{\mu})$.

Lévy stable laws The central limit theorem states that the properly normalized sum of independent, identically distributed random variables with finite variance converges to a Gaussian limit distribution. A generalization of this theorem exists for the case with infinite variance, namely, the generalized central limit theorem. The related distributions are the Lévy stable laws, whose density in the simplest case have the characteristic function $p(k) = \int_{-\infty}^{\infty} p(x)\exp(ikx)\mathrm{d}x$ (Fourier transform) of the form $p(k) = \exp(-c|k|^{\mu})$, with $0 < \mu < 2$. In direct space, this corresponds to a power-law asymptotic $p(x) \simeq c|x|^{-1-\mu}$. In the limit $\mu = 2$, the universal Gaussian distribution is recovered.

Lévy walks In contrast to Lévy flights, Lévy walks possess a finite mean squared displacement, albeit having a broad jump length distribution. This is possible by the introduction of a time penalty for long jumps through a coupling $\lambda(x)p(t|x)$ between waiting times and jump lengths, such that long jumps involve a longer waiting time. For instance, a δ-coupling of the form $\frac{1}{2}\lambda(x)\delta(|x| - vt)$ is often chosen, such that v plays the role of a velocity.

Strange kinetics Often, deviations from exponential relaxation patterns $\exp(-t/\tau)$ and regular Brownian diffusion are observed. Instead, non-exponential relaxation, for instance, of the stretched exponential form $\exp(-[t/\tau]^{\alpha})$ $(0 < \alpha < 1)$ or of the inverse power-law form $t^{-\zeta}$, is observed, or anomalous diffusion behavior is found.

Weak ergodicity breaking In a system with a broadly distributed waiting time with diverging characteristic time scale, a particle can get stuck at a certain position for a long time. For instance, for a waiting time distribution of the form $\psi(t) \sim \tau^{\alpha}t^{-1-\alpha}$, the probability of not moving until time t scales as $\tau^{\alpha}t^{-\alpha}$, that is, the probability of not moving is therefore appreciable even for long times. A particle governed by such a $\psi(t)$ even at stationarity does not equally explore different domains of phase space.

Definition of the Subject

Classical Brownian motion, characterized by a mean squared displacement

$$\langle x^2(t)\rangle = 2Kt\,, \tag{1}$$

growing linearly in time in absence of an external bias, is the paradigm for random motion. Note that we restrict our discussion to one dimension. Brownian motion quantifies the jittery motion of coal dust particles observed by Dutchman Jan Ingenhousz in 1785 [49], the zigzagging of pollen grains in solution reported by Robert Brown in 1827 [16], and possibly the dance of dust particles in the beam of sunlight in a stairwell so beautifully immortalized in the famed poem by Lucretius [17]. In general, Brownian motion occurs in simple, sufficiently homogeneous systems such as simple liquids or gasses. In the continuum limit, Brownian motion is governed by the diffusion equation

$$\frac{\partial P(x,t)}{\partial t} = K\frac{\partial^2}{\partial x^2}P(x,t) \tag{2}$$

for the probability density function (PDF) $P(x,t)$ describing the probability $P(x,t)\mathrm{d}x$ of finding the particle at a position in the interval $x,\ldots,x+\mathrm{d}x$ at time t. For a point-like initial condition $P(x,0) = \delta(x)$, the solution becomes the celebrated Gaussian PDF

$$P(x,t) = \frac{1}{\sqrt{4\pi Kt}}\exp\left(-\frac{x^2}{4Kt}\right)\,. \tag{3}$$

The diffusion constant K fulfills the Einstein–Stokes relation $K = k_BT/(m\eta)$, where $k_BT = RT/N_A$ is the Boltzmann energy at temperature T, R is the gas constant, N_A the Avogadro number, m is the mass of the test particle, and η the friction coefficient. This relation between microscopic and macroscopic quantities was used for the determination of the Avogadro number by Perrin [90]. Examples of Perrin's recorded random walks and the jump length distribution constructed from the data for a time increment of 30 seconds are shown in Fig. 1. Figure 2 for two different ambient pressures displays time traces of the Brownian motion of a small mirror in air that led to an amazingly accurate determination of the Avogadro number by Kappler [51].

However, in many systems deviations from linear behavior (1) are observed [14,25,55,78,80,107,122]. These deviations can assume the power-law form

$$\langle x^2(t)\rangle = 2K_{\zeta}t^{\zeta}\,. \tag{4}$$

In such systems, for $0 < \zeta < 1$, we observe subdiffusion. Prominent examples of subdiffusion include charge carrier transport in amorphous semiconductors [92,101], diffusion of chemicals in subsurface aquifers [102], the motion of beads in actin gels [121], motion in chaotic maps [8,123], or the subdiffusion of biomacromolecules or granules in cells [41,110]; just to name a few (compare [80]

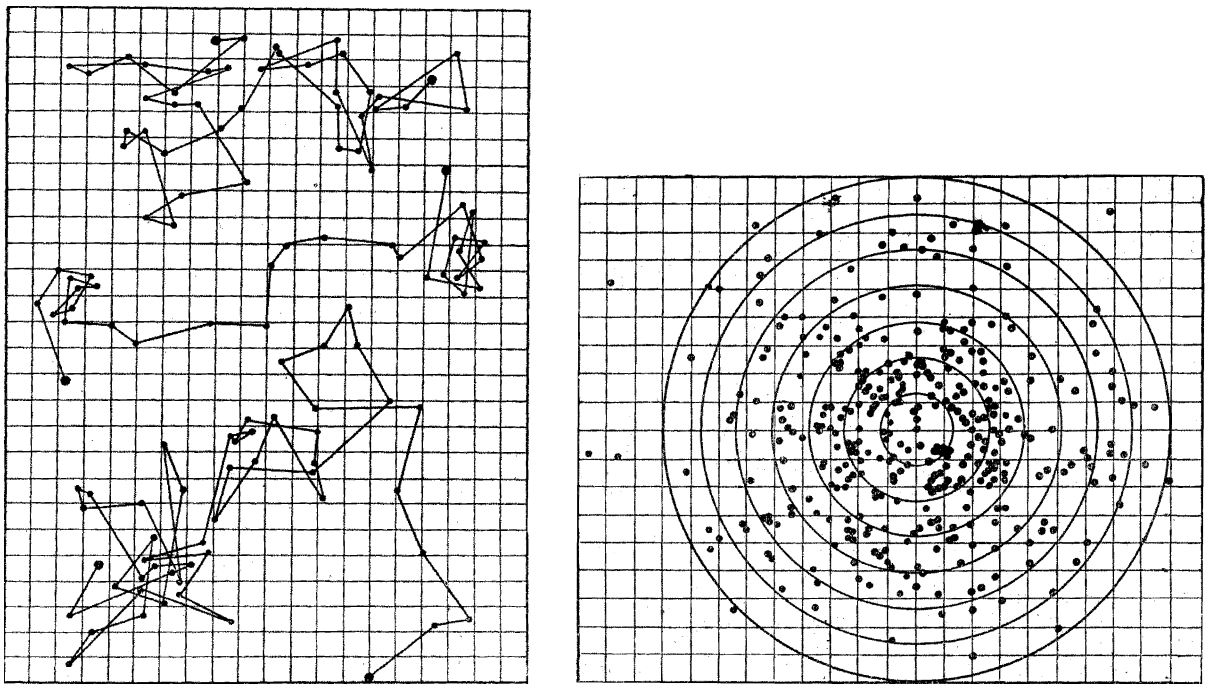

Levy Statistics and Anomalous Transport: Levy Flights and Subdiffusion, Figure 1
Random walk trajectories of putty particles in water recorded by Perrin [91]. *Left*: three designs obtained by tracing a small grain of putty at intervals of 30 sec. *Right*: the starting point of each motion event is shifted to the origin. The figure illustrates the continuum approach of the jump length distribution if only a large number of jumps is considered

for more details). Subdiffusion of this type is characterized by a long-tailed waiting time PDF $\psi(t) \simeq t^{-1-\alpha}$ $(0 < \alpha < 1)$, corresponding to the time-fractional diffusion equation

$$\frac{\partial P(x,t)}{\partial t} = K_\alpha \, {}_0D_t^{1-\alpha} \frac{\partial^2}{\partial x^2} P(x,t) \,, \tag{5}$$

with the fractional Riemann–Liouville operator [89,93,99]

$${}_0D_t^{1-\alpha} P(x,t) = \frac{1}{\Gamma(\alpha)} \frac{\partial}{\partial t} \int_0^t \frac{P(x,t')}{(t-t')^{1-\alpha}} \mathrm{d}t' \,. \tag{6}$$

From the latter definition, it becomes apparent that subdiffusion corresponds to a slowly decaying memory integral in the dynamical equation for $P(x,t)$.

The lack of a characteristic time scale (that is, the divergence of $\int_0^\infty t\psi(t)\mathrm{d}t$) of the waiting time PDF $\psi(t) \simeq t^{-1-\alpha}$ no longer permits us to distinguish microscopic and macroscopic events. The result is that the diffusing particle can get stuck at a certain position for very long times, quantified by the sticking probability $\phi(t) = \int_t^\infty \psi(t')\mathrm{d}t'$ of not moving. As detailed below the PDF $P(x,t)$ exhibits characteristic cusps at the location where the particle was initially released and, in the presence of an external drift, growing asymmetry. Moreover, a process whose measurement begins at $t = 0$ depends on the preparation time at some prior time, the so-called ageing effect [8,86]. Another effect in this context is that a particle no longer evenly distributes in a certain space, such that despite the existence of stationary states a weak ergodicity breaking occurs [11,13,45,65,66].

Contrasting with a subdiffusing particle are Lévy flights (LFs), that are based on a waiting time PDF with a finite characteristic time but a jump length distribution $\lambda(x) \simeq |x|^{-1-\mu}$ $(0 < \mu < 2)$ with diverging variance $\int_{-\infty}^{\infty} |x|^2 \lambda(x)\mathrm{d}x$. The diffusion equation for an LF becomes generalized to the space-fractional diffusion equation

$$\frac{\partial P(x,t)}{\partial t} = K^{(\mu)} \frac{\partial^\mu}{\partial |x|^\mu} P(x,t) \,, \tag{7}$$

where the fractional Riesz–Weyl operator is defined through its Fourier transform $\mathcal{F}\{\partial^\mu/\partial|x|^\mu \, P(x,t)\} = -|k|^\mu \, P(k,t)$. In Fourier space, one therefore recovers immediately the characteristic function $P(k,t) = \exp(-K^{(\mu)}|k|^\mu t)$ of a symmetric Lévy stable law. As can be seen from Eq. (7) LFs are Markovian processes. However, their trajectory now has a fractal dimension $d_f = \mu$ and

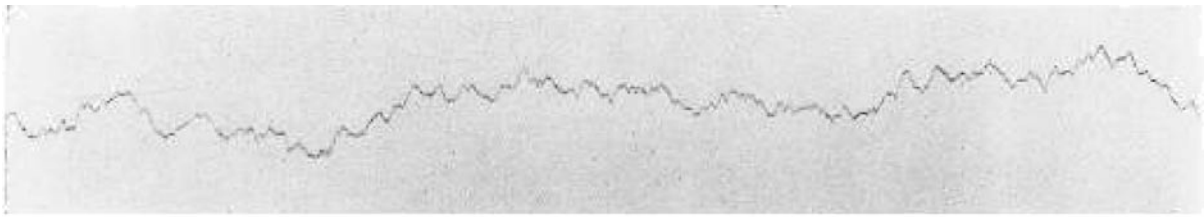

Registrieraufnahme der Brownschen Bewegung (natürlich Größe)
Direktionskraft 9,428 × 10^{-9} abs. Einh. Trägheitsmoment: 1×10^{-3} abs. Einh. Abstand Spiegel–Kamera: 72,1 cm.
Zeitmarke: 30 sec dx = 1 mm. a) Atmosphärendruck. Temperatur 13°C

Registrieraufnahme der Brownschen Bewegung (natürlich Größe)
Direktionskraft 9,428 × 10^{-9} abs. Einh. Trägheitsmoment: 1×10^{-7} abs. Einh. Abstand Spiegel–Kamera: 72,1 cm.
Zeitmarke: 30 sec dx = 1 mm. b) 1×10^{-3}mm Hg. Temperatur 13°C

Levy Statistics and Anomalous Transport: Levy Flights and Subdiffusion, Figure 2
Erratic behavior of Brownian motion made visible in a high-precision measurement (including Kappler's original figure captions). Data from an Edelmann recorder, obtained by Kappler in 1931 [51]. Kappler monitored the Brownian motion of a small mirror (surface approx. 1mm^2), suspended from a fine quartz thread (several cm long and some μm thick). The mean squared of the torsional displacement, $\overline{\varphi^2}$, follows the relation $D\overline{\varphi^2} = k_B T$, where D is the directional force of the suspension [51]. The facsimiles show two different realizations. From his data, Kappler obtained the Avogadro–Loschmidt number $N_L = 60.59 \cdot 10^{22} \pm 1\%$, to a remarkable accuracy

is characterized by local search changing with long excursions (see Fig. 3). This property has been shown to be a better search strategy than Brownian motion, as it oversamples less [64,116]. In fact, animals like albatross [115], spider monkeys [95], jackals [4], and even plankton [114] and bacteria [62] were claimed to follow Lévy search strategies. LFs also occur in diffusion in energy space [124], or in optical lattices [52]. Due to the clustering nature of their trajectory, LFs also exhibit a form of ergodicity breaking [68].

Introduction

For sums of independent, identically distributed random variables with proper normalization to the sample size, the generalized central limit theorem guarantees the convergence of the associated probability density to a Lévy stable density even though the variance of these random variables diverges [35,40,46,63,112]. Well-known examples of Lévy stable densities are the one-sided (defined for $x \geq 0$) Lévy–Smirnov distribution

$$f_{1/2,-1/2}(x) = \sqrt{\frac{1}{2\pi x^3}} \exp\left(-\frac{1}{2x}\right) , \tag{8}$$

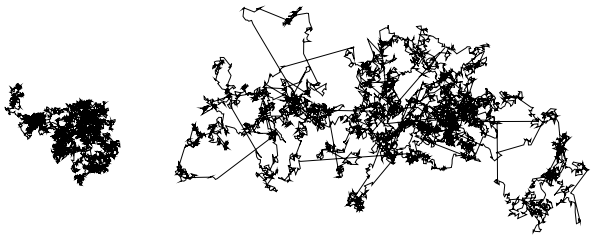

Levy Statistics and Anomalous Transport: Levy Flights and Subdiffusion, Figure 3
Trajectories of Brownian motion (*left*) and a Lévy flight of index $\mu = 1.5$ (*right*), both with the same number ($\simeq$ 7000) of steps. The long sojourns and clustering appearance of the Lévy flight are distinct

related to the first passage time density of a Gaussian random walk process of passing the origin (see below), and the Cauchy (or Lorentz) distribution

$$f_{1,0}(x) = \frac{1}{\pi\left(1 + x^2\right)} . \tag{9}$$

In general, a Lévy stable density is defined through its

characteristic function of the PDF $f(x)$

$$\varphi(z) \equiv \mathcal{F}\{f(x)\} = \int_{-\infty}^{\infty} f_{\mu,\beta}(x) \mathrm{e}^{ikx} \mathrm{d}x \tag{10}$$

where

$$\log \varphi(z) = -|z|^{\mu} \exp\left\{ i \frac{\pi\beta}{2} \mathrm{sign}(z) \right\} , \tag{11}$$

for $\mu \neq 1$. Here, the skewness (or asymmetry) parameter β is restricted to the following region:

$$|\beta| \leq \begin{cases} \mu , & \text{if } 0 < \mu < 1 \\ 2 - \mu , & \text{if } 1 < \mu < 2 . \end{cases} \tag{12}$$

For $\beta = 0$, the corresponding Lévy stable density is symmetric around $x = 0$, while for $\beta = -\mu$ and $0 < \mu < 1$, it is one-sided. In general, a Lévy stable density follows the power-law asymptotic behavior

$$f_{\mu,\beta}(x) \sim \frac{A_{\mu,\beta}}{|x|^{1+\mu}} , \quad \mu < 2 , \tag{13}$$

with $A_{\mu,\beta}$ being a constant, such that for all Lévy stable densities with $\mu < 2$ the variance diverges

$$\langle x^2 \rangle = \infty . \tag{14}$$

Conversely, all fractional moments $\langle |x|^{\delta} \rangle < \infty$ for all $0 < \delta < \mu \leq 2$. From the above definitions it is obvious that the Lévy stable density $f_{2,0}$ corresponds to the Gaussian normal distribution

$$f_{2,0}(x) = \sqrt{\frac{1}{4\pi}} \exp\left(-\frac{1}{4} x^2 \right) \tag{15}$$

possessing finite moments of any order. In this limit, the generalized central limit theorem coincides with the more traditional, and universal, central limit theorem.

Brownian motion has traditionally been employed as the dominant model of choice for random noise in continuous-time systems, due to its remarkable statistical properties and its amenability to mathematical analysis. However, Brownian motion is just a single example of the Lévy family. Moreover, it is a very special and somewhat unrepresentative member of this family. Amongst the Lévy family, the Brownian member is the only motion with continuous sample-paths. All other motions have discontinuous trajectories, exhibiting jumps. Moreover, the Lévy family is characterized by selfsimilar motions. Brownian motion is the only selfsimilar Lévy motion possessing finite variance, while all other selfsimilar Lévy motions have an infinite variance.

Random processes whose spatial coordinates x or clock times t are distributed according to a Lévy stable density exhibit anomalies, that is, no longer follow the laws of Brownian motion. Consider a continuous time random walk process defined in terms of the jump length and waiting time distributions $\lambda(x)$ and $\psi(t)$ [87,101]. That is, each jump event of this random walk is characterized by a jump length x drawn from the distribution λ, and the time t between two jump events is distributed according to ψ. (Note that an individual jump is supposed to occur spontaneously.) In absence of an external bias, continuous time random walk theory connects $\lambda(x)$ and $\psi(t)$ with the probability distribution $P(x, t)\mathrm{d}x$ to find the random walker at a position in the interval $(x, x + \mathrm{d}x)$ at time t. In Fourier–Laplace space, $P(k, u) \equiv \mathcal{F}\{\mathcal{L}\{P(x, t); t \to u\}; x \to k\}$, this relation reads [54]

$$P(k, u) = \frac{1 - \psi(u)}{u} \frac{1}{1 - \lambda(k)\psi(u)} , \tag{16}$$

where $\mathcal{L}\{f(t)\} \equiv \int_0^{\infty} \exp(-ut) f(t) \mathrm{d}t$. We here neglect potential complications due to ageing effects. The following cases can be distinguished:

(i) $\lambda(x)$ is Gaussian with variance σ^2 and $\psi(t) = \delta(t - \tau)$. Then, to leading order in k^2 and u, respectively, one obtains $\lambda(k) \simeq 1 - \sigma^2 k^2$ and $\psi(u) \simeq 1 - u\tau$. From relation (16), one recovers the Gaussian probability density $P(x, t) = \sqrt{1/(4\pi K t)} \exp\{-x^2/(4Kt)\}$ with diffusion constant $K = \sigma^2/\tau$. The corresponding mean squared displacement grows linearly with time; see Eq. (1). This case corresponds to the continuum limit of regular Brownian motion. Note that here and in the following, we restrict the discussion to one dimension.

(ii) Assume $\lambda(x)$ still to be Gaussian, while for the waiting time distribution $\psi(t)$ we choose a one-sided Lévy stable density with stable index $0 < \alpha < 1$. Consequently, $\psi(u) \simeq 1 - (u\tau)^{\alpha}$, and the characteristic waiting time $\int_0^{\infty} t\psi(t)\mathrm{d}t$ diverges. Due to this lack of a time scale separating microscopic (single jump events) and macroscopic (on the level of $P(x, t)$) scales, $P(x, t)$ is no longer Gaussian, but is given by a more complex H-function [78,80,103]. In Fourier space, however, one finds the quite simple analytical form [78]

$$\begin{aligned} P(k, t) &= E_{\alpha}\left(-K_{\alpha} k^2 t^{\alpha}\right) \\ &= \sum_{0}^{\infty} \frac{\left(-K_{\alpha} k^2 t^{\alpha}\right)^n}{\Gamma(1 + \alpha n)} \end{aligned} \tag{17}$$

in terms of the Mittag–Leffler function. This generalized relaxation function of the Fourier modes turns

from an initial stretched exponential (KWW) behavior

$$P(k,t) \sim 1 - K_\alpha k^2 t^\alpha/\Gamma(1+\alpha) \sim \exp\left\{-K_\alpha k^2 t^\alpha/\Gamma(1+\alpha)\right\} \quad (18)$$

to a terminal power-law behavior [78]

$$P(k,t) \sim \left(K_\alpha k^2 t^\alpha \Gamma(1-\alpha)\right)^{-1} . \quad (19)$$

In the limit $\alpha \to 1$, it reduces to the traditional exponential $P(k,t) = \exp(-Kk^2 t)$ with a finite characteristic waiting time. Also the mean squared displacement changes from a linear to a power-law time dependence

$$\langle x^2(t)\rangle = 2K_\alpha t^\alpha/\Gamma(1+\alpha) , \quad (20)$$

with $K_\alpha = \sigma^2/\tau_0^\alpha$. This is the case of *subdiffusion*. We note that in x, t space the dynamical equation is the fractional diffusion equation [103]. In the presence of an external potential, it generalizes to the time-fractional Fokker–Planck equation [78,80,84], see also below.

(iii) Finally, take $\psi(t) = \delta(t-\tau)$ sharply peaked, but $\lambda(x)$ of Lévy stable form with index $0 < \mu < 2$. The resulting process is Markovian, but with diverging variance. It can be shown that the fractional moments scale as [82]

$$\langle |x(t)|^\delta\rangle \propto \left(K^{(\mu)} t\right)^{\delta/\mu} , \quad (21)$$

where $K^{(\mu)} = \sigma^\mu/\tau$. The upper index μ is chosen to distinguish $K^{(\mu)}$ from the subdiffusion constant K_μ. Note that the dimension of $K^{(\mu)}$ is $\mathrm{cm}^\mu/\mathrm{sec}$. From Eq. (16) one can immediately obtain the Fourier image of the associated probability density function,

$$P(k,t) = \exp\left\{-K^{(\mu)}|k|^\mu t\right\} . \quad (22)$$

From Eq. (11) this is simply a symmetric Lévy stable density with stable index μ, and this type of random walk process is most aptly coined a Lévy flight. A Lévy flight manifestly has regular exponential mode relaxation and is in fact Markovian. However, the modes in position space are no longer sharply localized like in the Gaussian or subdiffusive case. Instead, individual modes bear the hallmark of a Lévy stable density: diverging variance. We will see below how the presence of steeper-than-harmonic external potentials causes a finite variance of the Lévy flight, although a power-law form of the probability density remains.

In the remainder of this paper, we deal with the physical and mathematical properties of Lévy flights and subdiffusion. While we will be concerned mostly with the overdamped case, in the last section we will address the dynamics in velocity space in the presence of Lévy noise, in particular, the question of diverging variance of Lévy flights.

Lévy Flights

Underlying Random Walk Process

To derive the dynamic equation of a Lévy flight in the presence of an external force field $F(x) = -\mathrm{d}V(x)/\mathrm{d}x$, we pursue two different routes. One starts with a generalized version of the continuous time random walk (compare [85] for details; a slightly different derivation is presented in [10]).

To include the local asymmetry of the jump length distribution due to the force field $F(x)$ [72,85], we introduce the generalized transfer kernel $\Lambda(x,x') = \lambda(x-x')\,[A(x')\Theta(x-x') + B(x')\Theta(x'-x)]$ (and therefore $\Lambda(x,x') = \Lambda(x';x-x')$). As in standard random walk theory (compare [117]), the coefficients $A(x)$ and $B(x)$ define the local asymmetry for jumping left and right, depending on the value of $F(x)$. Here, $\Theta(x)$ is the Heaviside jump function. With the normalization $\int \Lambda(x',\Delta)\,\mathrm{d}\Delta = 1$, the fractional Fokker–Planck equation (FFPE) ensues [85]:

$$\frac{\partial}{\partial t}P(x,t) = \left(-\frac{\partial}{\partial x}\frac{F(x)}{m\eta} + K^{(\mu)}\frac{\partial^\mu}{\partial|x|^\mu}\right)P(x,t) . \quad (23)$$

Remarkably, the presence of the Lévy stable $\lambda(x)$ affects only the diffusion term, while the drift term remains unchanged and additive [36,37,85]. The fractional spatial derivative $\partial^\mu/\partial|x|^\mu$ of Riesz–Weyl type represents an integrodifferential operator defined through

$$\frac{\partial^\mu}{\partial|x|^\mu}P(x,t) = \frac{-1}{2\cos(\pi\mu/2)\Gamma(2-\mu)} \frac{\partial^2}{\partial x^2}\int_{-\infty}^{\infty}\frac{P(x',t)}{|x-x'|^{\mu-1}} , \quad (24)$$

for $1 < \mu < 2$, and a similar form for $0 < \mu < 1$ [22,93,99]. In Fourier space, for all $0 < \mu \le 2$ the simple relation

$$\mathcal{F}\left\{\frac{\partial^\mu}{\partial|x|^\mu}P(x,t)\right\} = -|k|^\mu P(k,t) \quad (25)$$

holds. In the Gaussian limit $\mu = 2$, all relations above reduce to the familiar second-order derivatives in x and thus the corresponding $P(x,t)$ is governed by the standard Fokker–Planck equation.

The FFPE (23) can also be derived from the Langevin equation [22,36,37,50]

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -\frac{1}{m\gamma}\frac{\mathrm{d}V(x)}{\mathrm{d}x} + \xi_\mu(t)\,, \tag{26}$$

driven by white Lévy stable noise $\xi_\mu(t)$, defined through $L(\Delta t) = \int_t^{t+\Delta t} \xi_\mu\left(t'\right)\mathrm{d}t'$ being a symmetric Lévy stable density of index μ with characteristic function $p(k,\Delta t) = \exp\left(-K^{(\mu)}|k|^\mu \Delta t\right)$ for $0 < \mu \le 2$. As with standard Langevin equations, $K^{(\mu)}$ denotes the noise strength, m is the mass of the diffusing (test) particle, and γ is the friction constant characterizing the dissipative interaction with the bath of surrounding particles.

A subtle point about the FFPE (23) is that it does not uniquely define the underlying trajectory [105]; however, starting from our definition of the process in terms of the stable jump length distribution $\lambda(x) \sim |x|^{-1-\mu}$, or its generalized pendant $\Lambda(x,x')$, the FFPE (23) truly represents a Lévy flight in the presence of the force $F(x)$. The strongly non-local character of Lévy flights poses certain difficulties when non-trivial boundary conditions are involved, as shown below.

Propagator and Symmetries

In absence of an external force, $F(x) = 0$, the exact solution of the FFPE is readily obtained as the Lévy stable density $P(k,t) = \exp\left(-K^{(\mu)}|k|^\mu t\right)$ in Fourier space. Back-transformed to position space, an analytical solution is given in terms of the Fox H-function [50,78,120]

$$P(x,t) = \frac{1}{\mu|x|} H^{1,1}_{2,2}\left[\frac{|x|}{\left(K^{(\mu)}t\right)^{1/\mu}}\,\middle|\,\begin{matrix}(1,1/\mu),(1,1/2)\\(1,1),(1,1/2)\end{matrix}\right], \tag{27}$$

from which the series expansion

$$P(x,t) = \frac{1}{\mu(K^{(\mu)}t)^{1/\mu}} \sum_{\nu=0}^{\infty} \frac{\Gamma([1+\nu]/\mu)}{\Gamma([1+\nu]/2)\Gamma(1-[1+\nu]/2)} \times \frac{(-1)^\nu}{\nu!}\left(\frac{|x|}{(K^{(\mu)}t)^{1/\mu}}\right)^\nu \tag{28}$$

derives. For $\mu = 1$, the propagator reduces to the Cauchy Lévy stable density

$$P(x,t) = \frac{1}{\pi\left(K^{(1)}t + x^2/[K^{(1)}t]\right)}\,. \tag{29}$$

We plot the time evolution of $P(x,t)$ for the Cauchy case $\mu = 1$ in Fig. 4 in comparison to the limiting Gaussian case $\mu = 2$.

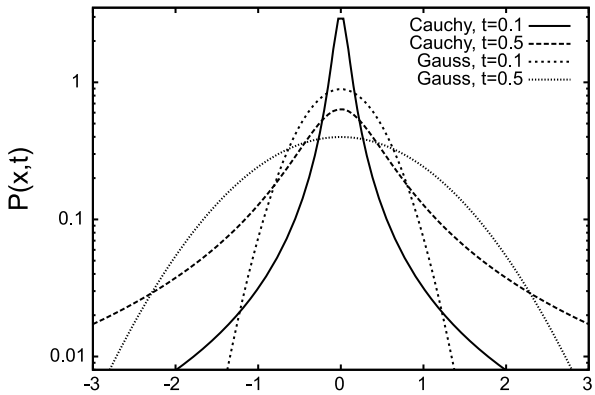

Levy Statistics and Anomalous Transport: Levy Flights and Subdiffusion, Figure 4

Cauchy distribution ($\mu = 1$) for two times in comparison to the Gaussian ($\mu = 2$). We chose $K^{(1)} = K = 1$. Note that the Cauchy distribution is narrower at the origin, and after crossing the Gaussian falls off in the much slower power-law fashion

Due to the point symmetry of the FFPE (23) for $F(x) = 0$, the propagator $P(x,t)$ is invariant under change of sign, and it is monomodal, that is, it has its global maximum at $x = 0$, the point where the initial distribution $P(x,0) = \delta(x)$ was launched at $t = 0$. Interestingly, the latter property is lost in the case of strongly confined Lévy flights discussed below. Due to their Markovian character, Lévy flights also possess a Galilei invariance [73,78]. Thus, under the influence of a constant force field $F(x) = F_0$, the solution of the FFPE can be expressed in terms of the force-free solution by introducing the wave variable $x - F_0 t$, to obtain

$$P_{F_0}(x,t) = P_0\left(x - \frac{F_0 t}{m\gamma}, t\right)\,. \tag{30}$$

This result follows from the FFPE (23), that in Fourier domain becomes [50]

$$\frac{\partial}{\partial t}P(k,t) = \left(-ik\frac{F_0}{m\gamma} - K^{(\mu)}|k|^\mu\right)P(k,t)\,, \tag{31}$$

with solution

$$P(k,t) = \exp\left(-\left[ik\frac{F_0}{m\gamma} + K^{(\mu)}|k|^\mu\right]t\right)\,. \tag{32}$$

By the translation theorem of the Fourier transform, Eq. (30) results. We show an example of the drift superimposed on the dispersional spreading of the propagator in Fig. 5.

Presence of External Potentials

Harmonic Potential In an harmonic potential $V(x) = \frac{1}{2}\lambda x^2$, an exact form for the characteristic function can

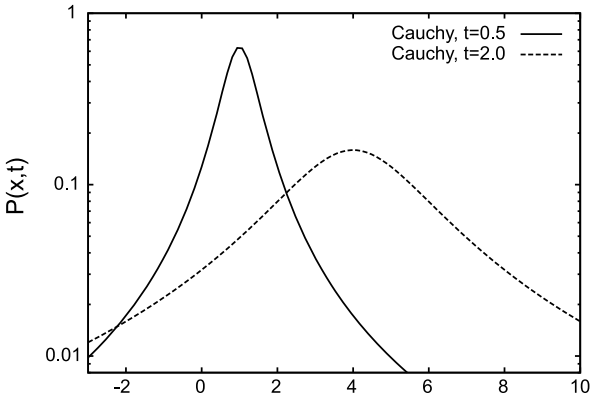

Levy Statistics and Anomalous Transport: Levy Flights and Subdiffusion, Figure 5
Cauchy distribution with $K^{(1)} = 1$ advected along a field $F_0/(m\gamma) = 2$, for different times

be found. Thus, from the corresponding FFPE in Fourier space,

$$\frac{\partial}{\partial t}P(k,t) = -\frac{\lambda}{m\gamma}k\frac{\partial}{\partial k}P(k,t) - K^{(\mu)}|k|^{\mu}P(k,t)\,, \quad (33)$$

by the method of characteristics one obtains

$$P(k,t) = \exp\left(-\frac{m\gamma K^{(\mu)}|k|^{\mu}}{\lambda\mu}\left[1-\mathrm{e}^{-\mu\lambda t/(m\gamma)}\right]\right) \quad (34)$$

for an initially central δ-peak, $P(x,0) = \delta(x)$ [50]. This is simply the characteristic function of a Lévy stable density with time-varying width. For short times, $1-\exp(-\mu\lambda t/[m\gamma]) \sim \mu\lambda t/[m\gamma]$ grows linearly in time, such that $P(k,t) \sim \exp\left(-K^{(\mu)}|k|^{\mu}t\right)$ as for a free Lévy flight. At long times, the stationary solution defined through

$$P_{\mathrm{st}}(k) = \exp\left(-\frac{m\gamma K^{(\mu)}|k|^{\mu}}{\lambda\mu}\right), \quad (35)$$

is reached. Notably, it has the same stable index μ as the driving Lévy noise. By separation of variables, a summation formula for $P(x,t)$ can be obtained similarly to the solution of the Ornstein–Uhlenbeck process in the presence of white Gaussian noise, but with the Hermite polynomials replaced by H-functions [50].

We note that in the Gaussian limit $\mu = 2$, the stationary solution, by necessity, must match the Boltzmann distribution corresponding to $\exp\left(-k_{\mathrm{B}}Tk^2/[2\lambda]\right)$. This requires that the Einstein–Stokes relation $K = k_{\mathrm{B}}T/(m\gamma)$ be fulfilled [113]. One might therefore speculate whether, for a system driven by external Lévy noise, a generalized Einstein–Stokes relation should hold, as was established for the subdiffusive case [78,84]. We will come back to this point below. As will be shown now, in steeper than harmonic external potentials, the stationary form of $P(x,t)$ even leaves the basin of attraction of Lévy stable densities.

Steeper than Harmonic Potentials To investigate the behavior of Lévy flights in potentials that are steeper than the harmonic case considered above, we introduce the non-linear oscillator potential

$$V(x) = \frac{a}{2}x^2 + \frac{b}{4}x^4\,, \quad (36)$$

that can be viewed as a next order approximation to a general confining, symmetric potential. It turns out that the resulting process differs from the above findings if a suitable choice of the ratio a/b is made. For simplicity, we introduce dimensionless variables through

$$x \to \frac{x}{x_0}\,; \quad t \to \frac{t}{t_0}\,; \quad a \to \frac{at_0}{m\gamma}\,, \quad (37)$$

where

$$x_0 = \left(\frac{m\gamma K^{(\mu)}}{b}\right)^{1/(2+\mu)}\,; \quad t_0 = \frac{x_0^{\mu}}{K^{(\mu)}}\,, \quad (38)$$

arriving at the FFPE

$$\frac{\partial}{\partial t}P(k,t) + |k|^{\mu}P(k,t) = \left(k\frac{\partial^3}{\partial k^3} - ak\frac{\partial}{\partial k}\right)P(k,t)\,. \quad (39)$$

Consider first the simplest case of a quartic oscillator with $a = 0$ in the presence of Cauchy noise ($\mu = 1$). In this limit, the stationary solution can be obtained exactly, yielding the expression

$$P_{\mathrm{st}}(x) = \frac{1}{\pi}\frac{1}{1-x^2+x^4} \quad (40)$$

plotted in Fig. 6. Two distinct new features are revealed in comparing the free Lévy flight with the Lévy flight in an harmonic potential: (1) Instead of the maximum at $x = 0$, one observes two maxima positioned at

$$x_m = \pm\sqrt{1/2}\,; \quad (41)$$

at $x = 0$, we find a local minimum. (2) There occurs a power-law asymptote

$$P_{\mathrm{st}}(x) \sim \frac{1}{\pi x^4} \quad (42)$$

for $x \gg 1$; consequently, this stationary solution no longer represents a Lévy stable density, and the associated mean squared displacement is finite, $\langle x^2\rangle < \infty$.

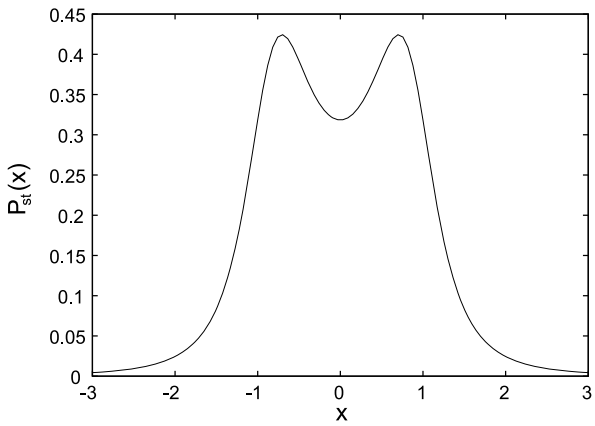

Levy Statistics and Anomalous Transport: Levy Flights and Subdiffusion, Figure 6
Bimodal stationary probability density $P_{st}(x)$ from Eq. (40). The maxima are at $\pm\sqrt{1/2}$

A more detailed analysis of Eq. (39) reveals [21,22], that (i) the bimodality of $P(x, t)$ occurs only if the amplitude of the harmonic term, a, is below a critical value a_c; (ii) for general μ, the asymptotic behavior is $P_{st}(x) \sim \pi^{-1} \sin(\pi\mu/2)\Gamma(\mu)|x|^{-\mu-3}$; (iii) and there exists a finite bifurcation time t_c at which the initially monomodal form of $P(x, t)$ acquires a zero curvature at $x = 0$, before settling in the terminal bimodal form.

Interestingly, in the more general power-law behavior

$$V(x) = \frac{|x|^c}{c} , \tag{43}$$

the turnover from monomodal to bimodal form of $P(x, t)$ occurs exactly when c becomes larger than 2. The harmonic potential is therefore a limiting case when the solution of the FFPE still belongs to the class of Lévy stable densities and follows the generalized central limit theorem. This is broken in a superharmonic (steeper than harmonic) potential. The corresponding bifurcation time t_c is finite for all $c > 2$ [22]. An additional effect appears when $c > 4$: there exists a transient trimodal state when the relaxing $\delta(x)$-peak overlaps with the forming humps at $x = \pm x_m$. At the same time, the variance is finite, if only $c > 4 - \mu$, following from the asymptotic stationary solution

$$P_{st}(x) \sim \frac{\sin(\pi\mu/2)\Gamma(\mu)}{\pi|x|^{\mu+c-1}} . \tag{44}$$

Details of the asymptotic behavior and the bifurcations can be found in [21,22]. From a reverse engineering point of view, Lévy flights in confining potentials are studied in [31].

First Passage and First Arrival of Lévy Flights

One might naively expect that a jump process of Lévy type, whose variance diverges (unless confined in a steep potential) may lead to ambiguities when boundary conditions are introduced, such as an absorbing boundary at finite x. Indeed, it is conceivable that for a jump process with extremely long jumps, it becomes ambiguous how to properly define the boundary condition: should the test particle be absorbed when it arrives exactly *at* the boundary, or when it crosses it *anyplace* during a non-local jump?

This question is trivial in the case of a narrow jump length distribution: all steps are small, and the particle cannot jump across a point (in the continuum limit considered herein). For such processes, one enforces a Cauchy boundary condition $P(0, t) = 0$ at the point $x = 0$ of the absorbing boundary, removing the particle once it hits the barrier after starting at x_0, where the dynamics are governed by Eq. (23) with $F(x) = 0$. Its solution can easily be obtained by standard methods, for instance, the method of images. This is completely equivalent to considering the *first arrival* to the point $x = 0$, expressed in terms of the diffusion equation with sink term:

$$\frac{\partial}{\partial t}\mathcal{P}(x, t) = K\frac{\partial^2}{\partial x^2}\mathcal{P}(x, t) - p_{fa}(t)\delta(x) , \tag{45}$$

defined such that $\mathcal{P}(0, t) = 0$. Note that the quantity $\mathcal{P}$ is no longer normalized, as probability decays to zero; for this reason, we use the notation $\mathcal{P}$. From Eq. (45) by integration we obtain the survival probability

$$\mathcal{S}(t) = \int \mathcal{P}(x, t)\mathrm{d}x \tag{46}$$

with $\mathcal{S}(0) = 1$ and $\lim_{t\to\infty} \mathcal{S}(t) = 0$. Then, the first arrival density becomes

$$p_{fa}(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{S}(t) . \tag{47}$$

Equation (45) can be solved by standard methods (determining the homogeneous and inhomogeneous solutions). It is then possible to express $\mathcal{P}(x, t)$ in terms of the propagator $P(x, t)$, the solution of Eq. (23) with $F(x) = 0$ with the same initial condition, $P(x, 0) = \delta(x - x_0)$ and natural boundary conditions. One obtains

$$P(0, t) = \int_0^t p_{fa}(\tau)P(x_0, t - \tau)\,\mathrm{d}\tau , \tag{48}$$

such that the first arrival density corresponds to the waiting time distribution to jump from x_0 to 0 (or, vice versa, since the problem is symmetric). In Laplace space, this relation takes on the simple algebraic form $p_{fa}(u)$

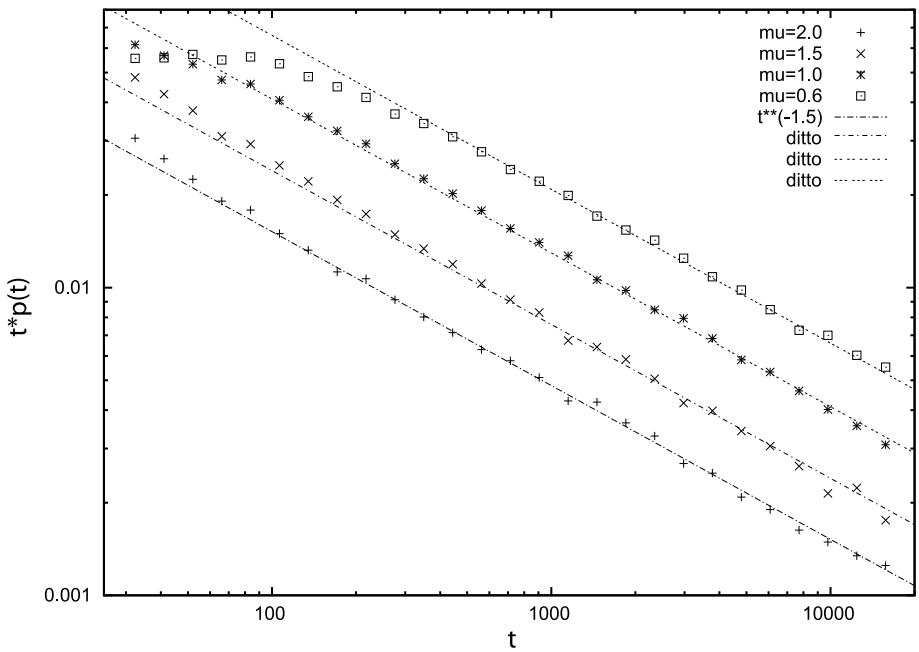

Levy Statistics and Anomalous Transport: Levy Flights and Subdiffusion, Figure 7
First passage density for various stable indices μ and initial position $x_0 = 10.0$ away from the absorbing boundary the universal $\sim t^{-3/2}$ scaling is distinct. Note that we plot $tp(t)$ in the y-axis

$= P(0,u)/P(x_0,u)$. Both methods, the explicit boundary value problem and the first arrival problem for Gaussian processes, produce the well-known first passage (or arrival) density of Lévy–Smirnov type (8),

$$p(t) = p_{\rm fa}(t) = \frac{x_0}{\sqrt{4\pi K t^3}} \exp\left(-\frac{x_0^2}{4Kt}\right) \sim \frac{x_0}{\sqrt{4\pi K t^3}}\,, \qquad (49)$$

with the asymptotic power-law decay $p(t) \sim t^{-3/2}$, such that no mean first passage time exists [46,97].

Long-tailed jump length distributions of Lévy stable form, however, endow the test particle with the possibility to jump across a certain point repeatedly. The first arrival necessarily becomes less efficient. Indeed, as shown in [20], the standard result (49) is generalized to

$$p_{\rm fa}(t) \sim C(\mu)\frac{x_0^{\mu-1}}{\left(K^{(\mu)}\right)^{1-1/\mu}\, t^{1-1/\mu}}\,, \quad \text{as } t \to \infty \qquad (50)$$

with $C(\mu) = \mu\Gamma(2-\mu)\Gamma(2-1/\mu)\sin(\pi[2-\mu]/2)\sin^2(\pi/\mu)/(\pi^2[\mu-1])$, and $1 < \mu \le 2$ [20]. The long-time decay $\sim t^{-2+1/\mu}$ is slower than in (49).

One might naively assume that the first passage problem for Lévy flights (the particle is removed once it crosses the boundary) should be more efficient, that is, the first passage density $p(t)$ should decay more quickly, than for a narrow jump length distribution. However, as we have a symmetric jump length distribution $\lambda(x)$, the long outliers characteristic for these Lévy flights can occur both toward and away from the absorbing barrier. From this point of view it is not totally surprising to see the simulations result in Fig. 7, that clearly indicate a universal asymptotic decay $\sim t^{-3/2}$, exactly as for the Gaussian case.

In fact, for all Markovian processes with a *symmetric* jump length distribution, the Sparre Andersen theorem [35,96,108,109] proves, without knowing any details about $\lambda(x)$, the asymptotic behavior of the first passage time density universally follows $p(t) \sim t^{-3/2}$. The details of the specific form of $\lambda(x)$ enter only the prefactor and the pre-asymptotic behavior. A special case of the Sparre Andersen theorem was proved in [38] in which the particle is released at $x_0 = 0$ at time $t = 0$, and after the first jump an absorbing boundary is installed at $x = 0$. This latter case was simulated extensively in [125]. From a fractional diffusion equation point of view, it was shown in [20] that the fractional operator $\partial^\mu/\partial|x|^\mu$ needs to be modified to account for the fact that $\mathcal{P}(x,t) \equiv 0$ beyond the absorbing boundary. The fractional diffusion equation in the presence of the absorbing boundary therefore becomes [20]

$$\frac{\partial}{\partial t}\mathcal{P}(x,t) = \frac{K_\mu}{\kappa}\frac{\partial^2}{\partial x^2}\int_0^\infty \frac{\mathcal{P}(x',t)}{|x-x'|^{\mu-1}}\mathrm{d}x'\,, \qquad (51)$$

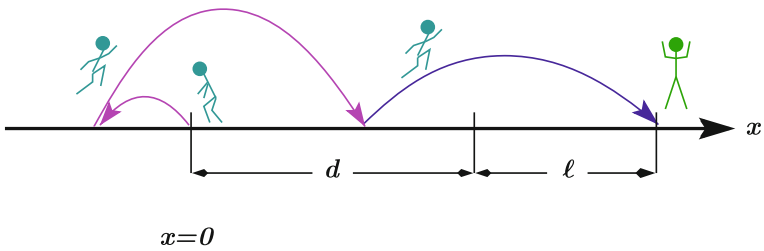

Levy Statistics and Anomalous Transport: Levy Flights and Subdiffusion, Figure 8
Schematic of the leapover problem: the random walker starts at $x = 0$ and after a number of jumps crosses the point $x = d$, overshooting it by a distance ℓ. For narrow jump length distributions, each jump is so small that crossing of the point d is equal to arriving at this point

where $\kappa = 2\Gamma(2-\mu)\,|\cos(\pi\mu/2)|$, such that the first term on the right hand side no longer represents a Fourier convolution. An approximate solution with Cauchy boundary condition reveals $p(u) \sim 1 - cu^{1/2}$, where c is a constant, indeed leading to the Sparre Andersen behavior $p(t) \sim t^{-3/2}$. The asymptotically exact result for the first passage time density of a Lévy flight on a semi-infinitive domain was derived in [58]:

$$p(t) \sim \frac{x_0^{\mu/2}}{\mu\sqrt{\pi K^{\mu}\Gamma(/mu/29)}} * t^{-3/2} .$$

This also demonstrates that the method of images no longer applies when Lévy flights are considered, for the images solution

$$\mathcal{P}_{\rm im}(x,t) = P(x-x_0,t) - P(x+x_0,t) \tag{52}$$

would be governed by the *full* fractional diffusion equation, and not Eq. (51), so the result for the first passage density, $p(t) \sim t^{-1-1/\mu}$ would decay faster than the Sparre Andersen universal behavior. A detailed discussion of the applicability of the method of images is given in terms of a subordination argument in [105]. We emphasize that this subtle failure of the method of images has been overlooked in literature previously [39,88], and care should therefore be taken when working with results based on such derivations. We also note that the method of images works in cases of subdiffusion, as the step length is narrow [75].

Leapover Properties of Lévy Flights

The statistics of first passage times is a classical concept for quantifying processes in which it is of interest when the dynamic variable crosses a certain threshold value for the first time. For processes with broad jump length distributions, another quantity is of interest, namely, the statistics of the first passage leapovers, that is, the distance ℓ by which the random walker overshoots the threshold value $x = d$ in a single jump (see Fig. 8). Surprisingly, for symmetric LFs with jump length distribution $\lambda(x) \sim |x|^{-1-\mu}$ $(0 < \mu < 2)$, the distribution of leapover lengths across x_0 is distributed like $p_l(\ell) \sim \ell^{-1-\mu/2}$, that is, it is much broader than the original jump length distribution. In contrast, for one-sided LFs jumps, the scaling of $p_l(\ell)$ bears the same index μ. Information on the leapover behavior is important to the understanding of how far search processes of animals for food or of proteins for their specific binding site along DNA overshoot their targets, or to better define stock market strategies determining when to buy or sell a certain relative to a given threshold price.

Using a general theorem for the first passage times and leapovers for homogeneous processes with independent increments, leapover properties for completely symmetric and fully asymmetric (one-sided) LFs are derived in [57]. The general case is considered in [59]. The basic results are as follows: For a completely symmetric LF with index μ, the distribution of first passage leapover lengths ℓ for a particle originally released a distance d away from the boundary reads (in scaled units)

$$p_l(\ell) = \frac{\sin(\pi\mu/2)}{\pi}\,\frac{d^{\mu/2}}{\ell^{\mu/2}(d+\ell)} . \tag{53}$$

The validity of this result is confirmed by extensive simulations – for more details refer to [57,58,59]. Note that p_l is normalized. In the limit $\mu \to 2$, p_l tends to zero if $\ell \neq 0$ and to infinity at $\ell = 0$, corresponding to the absence of leapovers in the Gaussian continuum limit. However, for $0 < \mu < 2$ the leapover PDF follows an asymptotic power-law with index $\mu/2$, and is thus broader than the original jump length PDF $\lambda(x)$ with index μ. This is a remarkable finding: while λ for $1 < \mu < 2$ has a finite characteristic length $\langle|x|\rangle$, this always diverges for $p_l(\ell)$ irrespective of μ.

In contrast, the result for completely asymmetric LFs has the form [31,57]

$$p_l(q) = \langle e^{-q\ell} \rangle = \frac{\sin(\pi\mu)}{\pi} \int_0^\infty d\ell e^{-q\ell} \frac{d^\mu}{\ell^\mu (d+\ell)} , \quad (54)$$

leading to the leapover PDF

$$p_l(\ell) = \frac{\sin(\pi\mu)}{\pi} \frac{d^\mu}{\ell^\mu (d+\ell)} , \quad (55)$$

which corresponds to the result obtained in [32,33] from a different method. Thus, for the one-sided LF, the scaling of the leapover is exactly the same as for the jump length distribution, namely, with exponent μ. This result compares favorably with simulations [57,59].

Kramers Problem for Lévy Flights

Many physical and chemical problems are related to the crossing of an energetic barrier driven by thermal fluctuations, including such problems as dissociation of molecules, nucleation processes, or the escape from an external, confining potential of finite height [44]. A particular example of barrier crossing in a double-well potential driven by Lévy noise was proposed for a long time series of paleoclimatic data [30]. Further cases where the crossing of a potential barrier driven by Lévy noise is of interest are in the theory of plasma devices [19], among others [80].

To investigate the detailed behavior of barrier crossing under the influence of external Lévy noise, we choose the rather generic double well shape

$$V(x) = -\frac{a}{2}x^2 + \frac{b}{4}x^4 . \quad (56)$$

Integrating the Langevin Equation (26) with white Lévy noise, we find an exponential decay of the survival density in the initial well:

$$p(t) = \frac{1}{T_c} \exp\left(-\frac{t}{T_c}\right) , \quad (57)$$

as demonstrated in Fig. 9. Lévy flight processes being Markovian, this is not surprising, since the mode relaxation is exponential [78,80]. More interesting is the question of how the mean escape time T_c behaves as a function of the characteristic noise parameters D and μ. In the regular Kramers problem with Gaussian driving noise the Arrhenius-type activation $T_c = C\exp(h/D)$ is followed, where h is the barrier height, and the prefactor C includes details of the potential. In the case of Lévy noise, a power-law form

$$T_c(\mu, D) = \frac{C(\mu)}{\left(K^{(\mu)}\right)^{\alpha(\mu)}} \quad (58)$$

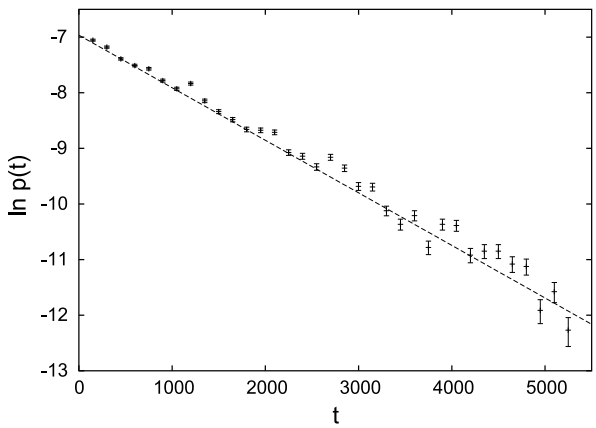

Levy Statistics and Anomalous Transport: Levy Flights and Subdiffusion, Figure 9
Probability density function $p(t)$ of barrier crossing times for $\mu = 1.0$ and $D = 10^{-2.5} \approx 0.00316$. The *dashed line* is a fit to Eq. (57) with mean escape time $T_c = 1057.8 \pm 17.7$

was assumed and corroborated by extensive simulations [23]. Detailed investigations [26] show that the scaling exponent $\alpha(\mu) = 1$ for all μ is strictly smaller than 2. As already proposed in [29] and derived in [48] in a somewhat different model, this means that, apart from a prefactor, the Lévy flight is insensitive to the external potential for barrier crossing, as confirmed by simulations [26]. Note that in comparison to [23], values of μ in the range $(0, 1)$ are also included. For large values of D, deviations from scaling are observed: eventually it will take only a single jump to cross the barrier when $D \to \infty$. Detailed studies show, indeed, that eventually the unit time step is reached, that is, $T_c \to 1$.

More on the "Pathology"

Despite their mathematical foundations in the generalized central limit theorem and their broad use in the sciences and beyond as description for statistical quantities, and despite the existence of systems (for instance, the diffusion on a polymer in chemical space mediated by jumps where the polymer loops back on itself [15,64,106]), the divergence of the fluctuations of Lévy processes is sometimes considered a pathology. This was already put forward by West and Seshadri [119], who pointed out that a Lévy flight in velocity space would be equivalent to a diverging kinetic energy. Here, we show that higher order dissipation effects lead to natural cutoffs in Lévy processes.

At higher velocities the friction experienced by a moving body starts to depend on the velocity itself [12]. Such non-linear friction is known from the classical Riccati equation $M dv(t)/dt = Mg - Kv(t)^2$ for the fall of

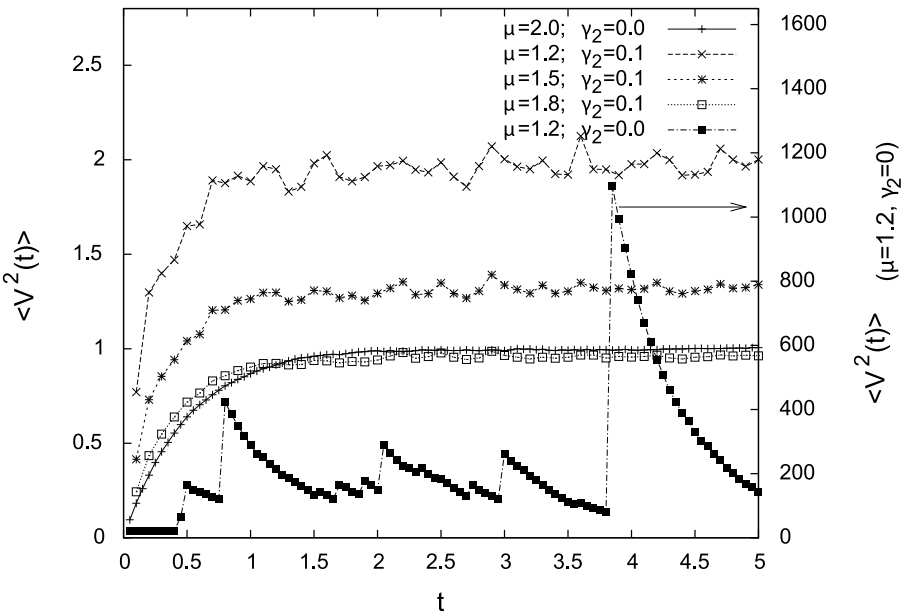

Levy Statistics and Anomalous Transport: Levy Flights and Subdiffusion, Figure 10
Variance $\langle V^2(t)\rangle$ as function of time t, with the quartic term set to zero, $\gamma_4 = 0$ and $\gamma_0 = 1.0$ for all cases. The variance is finite for the cases $\mu = 2.0, \gamma_2 = 0.0$; $\mu = 1.2, \mu = 1.5, \gamma_2 = 0.1$; and $\mu = 1.8, \gamma_2 = 0.1$. These correspond to the left ordinate. For the case $\mu = 1.2, \gamma_2 = 0.0$, the variance diverges, strong fluctuations are visible; note the large values of this curve corresponding to the right ordinate

a particle of mass M in a gravitational field with acceleration g [28], or autonomous oscillatory systems with a friction that is non-linear in velocity [3,12]. The occurrence of a non-constant friction coefficient $\gamma(V)$ leading to a non-linear dissipative force $-\gamma(V)V$ was highlighted in Klimontovich's theory of non-linear Brownian motion [56]. It is therefore natural that higher order, non-linear friction terms also occur in the case of Lévy processes.

We consider the velocity-dependent dissipative non-linear form (necessarily an even function) [24]

$$\gamma(V) = \gamma_0 + \gamma_2 V^2 + \ldots + \gamma_{2N} V^{2N} \quad \therefore \gamma_{2N} > 0 \quad (59)$$

for the friction coefficient of the Lévy flight in velocity space as governed by the Langevin equation

$$\mathrm{d}V(t) + \gamma(V)V(t)\mathrm{d}t = \mathrm{d}L(t) \quad (60)$$

with the constant friction $\gamma_0 = \gamma(0)$. $L(t)$ is again the μ-stable Lévy noise defined in terms of a characteristic function $p^*(\omega, t) = \mathcal{F}\{p(L,t)\} \equiv \int_{-\infty}^{\infty} p(L,t)\exp(i\omega L)\,\mathrm{d}L$ of the form $p^*(\omega, t) = \exp\left(-K^{(\mu)}|\omega|^\mu t\right)$ [63,100,112], where $K^{(\mu)}$ of dimension cm$^\mu$/sec is the generalized diffusion constant. This is equivalent to the fractional Fokker–Planck equation [19,22,36,37,78,80]

$$\frac{\partial P(V,t)}{\partial t} = \frac{\partial}{\partial V}\Big(V\gamma(V)P\Big) + K^{(\mu)}\frac{\partial^\mu P}{\partial |V|^\mu}\,. \quad (61)$$

As we showed before by the example of Lévy flights in position space, the presence of the first higher order correction, $\gamma_2 V^2$ in the friction coefficient $\gamma(V)$ rectifies the Lévy motion such that the asymptotic power-law becomes steeper and the variance finite. When even higher order corrections are taken into consideration, higher order moments also become finite. Figure 10 shows an example for the second moment.

The effect on the velocity distribution of the process defined by Eqs. (60) and (61) for higher order corrections is demonstrated in Fig. 11 for the stationary limit, $P_{\mathrm{st}}(V) = \lim_{t\to\infty} P(V,t)$. While for smaller V, the character of the original Lévy stable behavior is preserved (the original power-law behavior, that is, persists to intermediately large V), for even larger V the corrections due to the dissipative non-linearity are visible in the transition(s) to steeper slope(s).

These dissipative non-linearities remove the divergence of the kinetic energy from the measurable subsystem of the random walker. In ideal mathematical language, the surrounding bath provides an infinite amount of energy

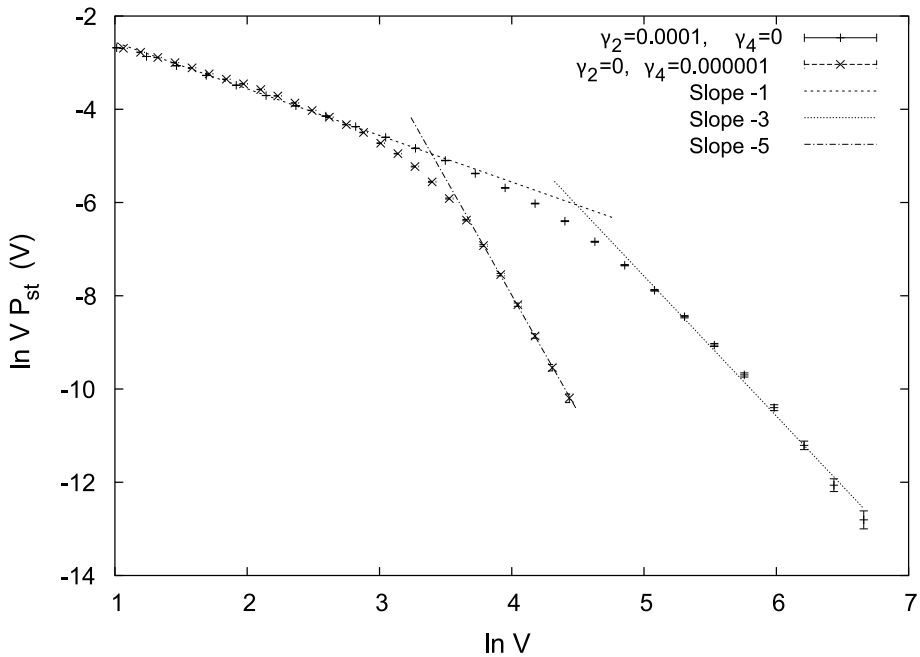

Levy Statistics and Anomalous Transport: Levy Flights and Subdiffusion, Figure 11
Stationary PDF $P_{st}(V)$ for $\gamma_0 = 1.0$ and (i) $\gamma_2 = 0.0001$ and $\gamma_4 = 0$; and (ii) $\gamma_2 = 0$ and $\gamma_4 = 0.000001$; with $\mu = 1.0$. The *lines* indicate the slopes − 1, − 3, and − 5

through the Lévy noise, and coupling via non-linear friction dissipates an infinite amount of energy into the bath, thereby introducing a natural cutoff in the kinetic energy distribution of the random walker subsystem. Physically, such divergencies are not expected, but correspond to the limiting procedure of large numbers in probability theory. We showed that both statements can be reconciled and that Lévy processes are, indeed, physical.

Also, Gaussian continuum diffusion exhibits non-physical features, possibly the most prominent being the infinite propagation speed inherent in the parabolic nature of the diffusion equation: even at very short times after system preparation in, say, a state $P(x, 0) = \delta(x)$, there has already arrived a finite portion of probability at large x. This problem can be corrected by changing from the diffusion to the Cattaneo (telegrapher's) equation. Still, for most purposes, the uncorrected diffusion equation is used. Or, one often uses natural boundary conditions even though the system under consideration is finite, since one might not be interested in the behavior at times when a significant portion of probability has reached the boundaries. In a similar sense, we showed that "somewhere out in the wings" Lévy flights are naturally cut off by dissipative non-linear effects. However, instead of introducing artificial cutoffs, knowing that for all purposes Lévy flights are a good quantitative description, and therefore meaningful, we use "pure" Lévy stable laws in physical models.

Bi-fractional Transport Equations

The coexistence of long-tailed forms for both jump length and waiting time PDFs was investigated within the CTRW approach in [126], discussing in detail the laminar-localized phases in chaotic dynamics. In the framework of fractional transport equations, combination of the waiting time PDF $\psi(t) \sim t^{-1-\alpha}$ $(0 < \alpha < 1)$ and jump length PDF $\lambda(x) \sim |x|^{-1-\mu}$ leads to a dynamical equation with fractional derivatives in respect to both time and space [67, 69,82,118]:

for $\alpha \leq 1$ and [82]

$$\frac{\partial}{\partial t} P(x, t) = K_\alpha^{(\mu)} \, {}_0D_t^{1-\alpha} \frac{\partial^\mu}{\partial |x|^\mu} P(x, t) \tag{62}$$

for $1 < \alpha \leq 2$.

As long as the condition $1 \leq \alpha \leq \mu \leq 2$ is met, both exponents can be chosen within the entire range [69]. For $\mu = 2$, in particular, this equation covers both sub- and superdiffusion up to ballistic motion, the latter corresponding to the wave equation $\alpha = \mu = 2$ [74]. A closed form solution of Eq. (62) can be found in terms of Fox's H-functions, [82], where some special cases permitting elementary solutions are also considered. Bi-fractional diffusion equations were also discussed in Refs. [5,7,42,47,98,111]. A bi-fractional Fokker–Planck equation with a power-law

dependence $\propto |x|^{-\theta}$ ($\theta \in \mathbb{R}$) of the diffusion coefficient was studied in [34,61].

Lévy Walks

Lévy walks correspond to the spatiotemporally coupled version of continuous time random walks. The waiting time and jump length PDFs, that is, are no longer decoupled but appear as conditional in the form $\psi(x,t) \equiv \lambda(x)p(t|x)$ (or $\psi(t)\tilde{p}(x|t)$) [54]. In particular, through the coupling $p(t|x) = \frac{1}{2}\delta\left(|x| - vt^{\nu}\right)$, one introduces a generalized velocity v, which penalizes long jumps such that the overall process, the Lévy walk, attains a finite variance and a PDF with two spiky fronts successively exploring space [53,123]. Thus, Lévy walks have similar properties to generalized Cattaneo/telegraphers' equation-type models [27,73,81]. As we focus here on the properties of transport processes governed by the Langevin equation under Lévy noise, we only briefly introduce Lévy walks.

On the basis of fractional equations, formulations were obtained for the description of Lévy walks in the presence of non-trivial external force fields, with the same restriction to lower order moments in respect to a Lévy walk process [9,83]. Recently, however, a coupled fractional equation was reported [104], which describes a force-free Lévy walk exactly. Thus, it was shown that the fractional version of the material derivative $\partial/\partial t \pm \partial/\partial x$,

$$d_{\pm}^{\beta} P(x,t) \equiv {}_0D_t^{\beta} P(x \pm t, t)\,, \tag{63}$$

defined in Fourier–Laplace space through

$$\mathcal{F}\left\{\mathcal{L}\left\{d_{\pm}^{\beta} f(x,t); u\right\}; k\right\} \equiv (u \pm ik)^{\beta} f(k,u) \tag{64}$$

(where $\mathcal{F}$ acts on x and $\mathcal{L}$ on t) replaces the uncoupled fractional time operators – see also the detailed discussion of Lévy walk processes in [123]. Although one may argue for certain forms [104], there is so far no derivation for the incorporation of general external force fields in the coupled formalism. We note that a very similar fractional approach to Lévy walks was suggested in [70].

Subdiffusion and the Fractional Fokker–Planck Equation

Physical Foundation of Subdiffusion

The stochastic motion of a Brownian particle of mass m is described by the Langevin equation [18,60,113]

$$m\frac{\mathrm{d}^2 x}{\mathrm{d}t^2} = -m\eta_1 v + F(x) + m\Gamma(t)\,, \quad v = \frac{\mathrm{d}x}{\mathrm{d}t}\,, \tag{65}$$

where $F(x)$ is an external force field, and η_1 is the friction coefficient. The erratic bombardment through the surrounding bath molecules is described by the fluctuating noise $\Gamma(t)$. To properly describe Brownian motion, $\Gamma(t)$ must be chosen δ-correlated (white) and Gaussian distributed. This is, the time averages of $\Gamma(t)$ are: $\overline{\Gamma(t)} = 0$, and $\overline{\Gamma(t)\Gamma(t')} = D\delta(t-t')$, where D is the noise strength. After averaging the fluctuations, the velocity moments become [18]

$$\langle \Delta v \rangle = \left(\eta v - \frac{F(x)}{m}\right)\Delta t\,,$$
$$\langle (\Delta v)^2 \rangle = \frac{2\eta k_{\mathrm{B}} T}{m}\Delta t + \mathcal{O}\left([\Delta t]^2\right)\,. \tag{66}$$

Both are proportional to Δt. For regular Brownian motion, these increments are used as expansion coefficients in the Chapman–Kolmogorov equation [18].

Subdiffusion now comes about through a so-called trapping scenario. Trapping describes the occasional immobilization of the test particle for a waiting time distributed according to the distribution $\psi(t)$. Here we assume that the particle leaves the trap with the same velocity it had prior to immobilization (this condition can be relaxed). In between trapping events, the particle is assumed to follow the regular Langevin equation such that each motion event on average lasts for a mean time τ^*. Choosing $\psi(t)$ with a finite characteristic waiting time one can show that this trapping scenario indeed preserves Brownian motion. However, once the characteristic waiting time diverges, it can be shown by application of the generalized central limit theorem that the occurrence of a large number of trapping events leads to subdiffusion and the fractional Klein–Kramers equation for the joint PDF $P(x,v,t)$ [71,76,77]. Integrating out the velocity coordinate, the fractional Fokker–Planck equation emerges for the PDF $P(x,t)$:

$$\frac{\partial P(x,t)}{\partial t} = {}_0D_t^{1-\alpha}\frac{\partial}{\partial x}\left(-\frac{F(x)}{m\eta_\alpha} + K_\alpha\frac{\partial}{\partial x}\right)P(x,t)\,. \tag{67}$$

where η_α is generalized friction coefficient of dimension $[\eta_\alpha] = \mathrm{sec}^{\alpha-2}$.

In mathematical terms the trapping scenario corresponds to a subordination principle: Trapping events cause a transformation of the internal step time of the random walk to a different observation time, corresponding to the relation

$$P(x,t) = \int_0^{\infty} \mathcal{E}_\alpha(s,t)P_1(x,s)\mathrm{d}s\,, \tag{68}$$

where $P_1(x,t)$ is the solution of the regular Brownian Fokker–Planck equation with $\alpha = 1$. Explicit forms of

the kernel $\mathcal{E}_\alpha(s,t)$ are known in terms of Fox' H-function [78]. In the Laplace domain, the kernel $\mathcal{E}_\alpha(s,t)$ has the comparatively simple form

$$\mathcal{E}_\alpha(s,u) = \frac{\eta_\alpha}{\eta_1 u^{1-\alpha}} \exp\left(-\frac{\eta_\alpha}{\eta_1} u^\alpha s\right) . \tag{69}$$

Note that this transformation guarantees the existence and positivity of the solution $P(x,t)$ of the fractional Fokker–Planck equation only if the corresponding Brownian problem possesses a well-defined solution. We note that recently an alternative approach to continuous time random walk subdiffusion in terms of δ-noise spike trains was presented [94].

Linear Response and Fluctuation-Dissipation Relation

From the fractional Fokker–Planck Equation (67) it is straightforward to prove that in the presence of a constant field F_0 the linear response relation

$$\langle x(t)\rangle_{F_0} = \frac{k_B T}{2} \langle x^2(t)\rangle_{F=0} \tag{70}$$

is fulfilled [78,84]. This is due to the fact that the waiting time distribution $\psi(t)$ is independent of force. Given the comparatively large traps required to create the long-tailed form of $\psi(t)$, this assumption appears reasonable for values of F_0 that are not too large. Experimentally, the linear response relation (70) was verified [43].

The stationary solution of the fractional Fokker–Planck Equation (67) is the standard Boltzmann–Gibbs equilibrium, $\lim_{t\to\infty} P(x,t) = \mathcal{N} \exp\{-V(x)/[k_B T]\}$, where $\mathcal{N}$ is a normalization constant [84]. This result is immediately obvious from the concept of subordination which changes the temporal spacing of events but preserves causal mode relaxation. As $F(x) = -V'(x)$, from the stationary solution of Eq. (67) by comparison with the Boltzmann–Gibbs form, one recovers the generalized Einstein–Stokes relation [84]

$$K_\alpha = \frac{k_B T}{m\eta_\alpha} , \tag{71}$$

reflecting the preservation of the fluctuation-dissipation relation for the subdiffusion process. An experimental verification of this relation was reported by [2].

Propagator

In the absence of an external force, the fractional diffusion equation is solved by Fox's H-function [78],

$$P(x,t) = \frac{1}{\sqrt{4K_\alpha t^\alpha}} H^{1,0}_{1,1}\left[\frac{|x|}{\sqrt{K_\alpha t^\alpha}} \,\middle|\, \begin{matrix}(1-\frac{\alpha}{2},\frac{\alpha}{2})\\(0,1)\end{matrix}\right] , \tag{72}$$

an equivalent form to the original solution by [103]. Its series expansion reads [78]

$$P(x,t) = \frac{1}{\sqrt{4K_\alpha t^\alpha}} \sum_{n=0}^{\infty} \frac{(-1)^n}{n!\Gamma(1-\alpha[n+1]/2)} \left(\frac{x^2}{K_\alpha t^\alpha}\right)^{n/2} , \tag{73}$$

and asymptotically reaches the stretched Gaussian form

$$P(x,t) \sim \frac{1}{\sqrt{4\pi K_\alpha t^\alpha}} \sqrt{\frac{1}{2-\alpha}} \times \left(\frac{2}{\alpha}\right)^{(1-\alpha)/(2-\alpha)} \left(\frac{|x|}{\sqrt{K_\alpha t^\alpha}}\right)^{-(1-\alpha)/(2-\alpha)} \times \exp\left(-\frac{2-\alpha}{2}\left(\frac{\alpha}{2}\right)^{\alpha/(2-\alpha)} \left[\frac{|x|}{\sqrt{K_\alpha t^\alpha}}\right]^{1/(1-\alpha/2)}\right) , \tag{74}$$

valid for $|x| \gg \sqrt{K_\alpha t^\alpha}$. The latter corresponds to the known result from continuous time random walk theory [53,123]. The H-function simplifies if the exponent α is a rational number. For instance, for $\alpha = 1/2$, it can be rewritten in terms of the Meijer G-function

$$P(x,t) = \frac{1}{\sqrt{8\pi^3 K_{1/2} t^{1/2}}} G^{3,0}_{0,3}\left[\left(\frac{x^2}{16K_{1/2}t^{1/2}}\right)^2 \,\middle|\, \begin{matrix}\\0,\frac{1}{4},\frac{1}{2}\end{matrix}\right] , \tag{75}$$

that is found in symbolic mathematics packages such as Mathematica. Using the representation in Fig. 12, we show the subdiffusive propagator with its pronounced cusps at the site of the initial condition, in comparison to the smooth Gaussian propagator of Brownian motion.

Boundary Value and First Passage Time Problems

In the presence of reflecting, absorbing, or mixed, boundary conditions, the narrow jump length distribution of subdiffusive processes makes it possible to use analogous techniques to calculate the propagator as known from normal diffusion. Thus, separation into eigenmodes or the method of images can be applied, the only difference entering through time-dependence of the modes. A very convenient way is to employ the Brownian result for a given geometry, subordinating that result to obtain the behavior for a subdiffusing particle, according to Eq. (68).

In the presence of an absorbing boundary an important quantity characterizing the dynamics of the system is the survival probability

$$\mathcal{S}_\alpha(t) = \int_{\mathbb{D}} \mathcal{P}(x,t)\mathrm{d}x , \tag{76}$$

where $\mathbb{D}$ denotes the interval over which $\mathcal{P}(x,t)$ is defined. The initial value of the survival probability is $\mathcal{S}_\alpha(0) = 1$,

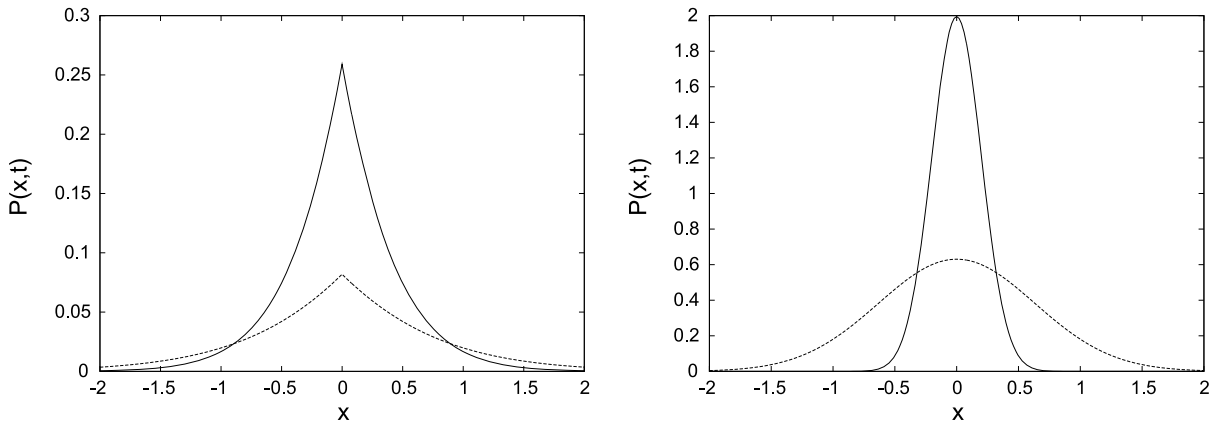

Levy Statistics and Anomalous Transport: Levy Flights and Subdiffusion, Figure 12
Propagator $P(x, t)$ of the fractional diffusion equation with $\alpha = 1/2$ (*left*) and the normal diffusion equation (*right*) for consecutive times

and it decays to zero for long times. From $\mathcal{S}_\alpha(t)$, we can define the first passage time density

$$p_\alpha(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{S}_\alpha(t)\,. \tag{77}$$

The mean first passage time is

$$\mathcal{T} = \int_0^\infty p_\alpha(t)t\mathrm{d}t\,. \tag{78}$$

Note that from Eqs. (77) and (68), we obtain the relation

$$p_\alpha(u) = p_1\left(u^\alpha\right) \tag{79}$$

between the subdiffusive and Brownian results in Laplace space. Thus, these are connected by a relation similar to the subordination (68), with kernel $\exp\left(-su^\alpha\right)$ instead of $\mathcal{E}_\alpha(s, u)$ [79,80].

For the three most prominent cases of first passage time problems, we obtain the following subdiffusive generalizations:

(i) For subdiffusion in the semi-infinite domain with an absorbing wall at the origin and initial condition $P(x, 0) = \delta(x - x_0)$, it was found that [75]

$$p(t) \sim \frac{x_0}{|\Gamma(-\alpha/2)|K_\alpha^{1/2}}t^{-1-\alpha/2}\,, \tag{80}$$

that is, the decay becomes a flatter power-law than in the Markovian case where $p(t) \sim t^{-3/2}$.

(ii) Subdiffusion in the semi-infinite domain in the presence of an external bias towards the absorbing boundary falls off faster, but still in power-law manner [6,79,102]:

$$p(t) \sim t^{-1-\alpha}\,. \tag{81}$$

In strong contrast to the biased Brownian case, we now end up with a process whose characteristic time scale diverges. This is exactly the mirror of the multiple trapping model, that is, the classical motion events become repeatedly interrupted such that immobilization time dominates the process. In contrast, for $\alpha = 1$ the result

$$p_1(t) = \frac{x_0}{\sqrt{4\pi K_\alpha t^3}}\exp\left(-\frac{(x_0 - Vt)^2}{4Kt}\right) \tag{82}$$

is valid, producing the classical form $\mathcal{T} = x_0/V$ for the mean first passage time.

(iii) Subdiffusion in a finite box [75]:

$$p(t) \sim t^{-1-\alpha}\,, \tag{83}$$

that is, this process leads to the same scaling behavior for longer times as found for the biased semi-infinite case (ii).

The latter two results should be compared to the classical Scher–Montroll finding for the first passage time density of biased motion in a finite system of size L with absorbing boundary conditions. In that case, the first passage time density exhibits two power-laws

$$p(t) \sim \begin{cases} t^{\alpha-1}\,, & t < \tau \\ t^{-1-\alpha}\,, & t > \tau \end{cases} \tag{84}$$

the sum of whose exponents equals -2 [92,101,102]. Here, τ is a system size dependent time scale [92].

Fractional Ornstein–Uhlenbeck Process

The Ornstein–Uhlenbeck process corresponds to the motion in an harmonic potential $V(x) = \frac{1}{2}m\omega^2x^2$ giving rise

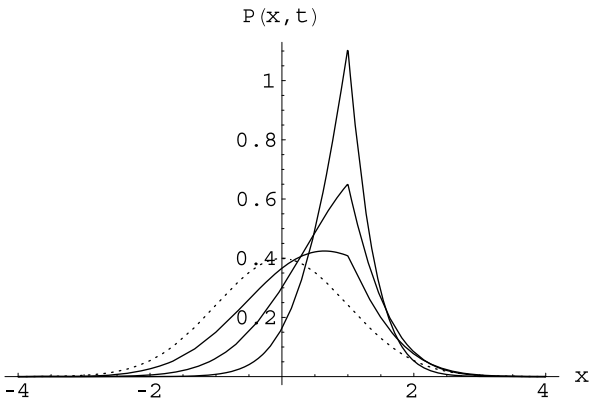

Levy Statistics and Anomalous Transport: Levy Flights and Subdiffusion, Figure 13
Time evolution of the PDF of the fractional Ornstein–Uhlenbeck process ($\alpha = 1/2$). The initial condition was chosen as $\delta(x-1)$. Note the strongly persistent cusp at the location of the initial peak. Dimensionless times: 0.02, 0.2, 2. The *dashed line* corresponds to the Boltzmann equilibrium

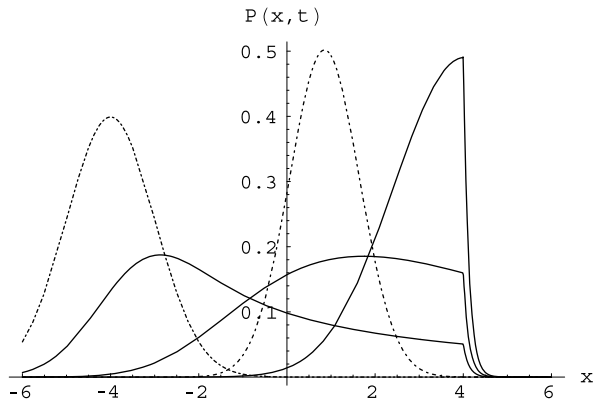

Levy Statistics and Anomalous Transport: Levy Flights and Subdiffusion, Figure 14
Time evolution of the PDF of the fractional Ornstein–Uhlenbeck process with superposed constant force of dimensionless strength $V = -4$ ($\alpha = 1/2$). The initial condition was chosen as $\delta(x-4)$. Dimensionless times: 0.02, 0.2, 2. The *dashed line* corresponds to the Brownian solution at times 0.5 and 50 (in essence, the stationary state). Again, note the cusps due to the initial condition, causing a strongly asymmetric shape of the PDF in contrast to the Gaussian nature of the Brownian counterpart

to the restoring force field $F(x) = -m\omega^2 x$, that is, to the dynamical equation

$$\frac{\partial}{\partial t}P(x,t) = {}_0D_t^{1-\alpha}\left(\frac{\partial}{\partial x}\frac{\omega^2 x}{\eta_\alpha} + K_\alpha\frac{\partial^2}{\partial x^2}\right)P(x,t)\,. \quad (85)$$

From separation of variables and the definition of the Hermite polynomials [1], one finds the series solution for the fractional Fokker–Planck equation with the Ornstein–Uhlenbeck potential [78,84],

$$P(x,t) = \sqrt{\frac{m\omega^2}{2\pi k_B T}}\sum_{n=0}^{\infty}\frac{1}{2^n n!}E_\alpha\left(-\frac{n\omega^2 t^\alpha}{\eta_\alpha}\right) H_n\left(\frac{\sqrt{m}\omega x_0}{\sqrt{2k_B T}}\right)\times H_n\left(\frac{\sqrt{m}\omega x}{\sqrt{2k_B T}}\right)\exp\left(-\frac{m\omega^2 x^2}{2k_B T}\right) \quad (86)$$

plotted in Fig. 13. Individual spatial eigenmodes follow the ordinary Hermite polynomials of increasing order, while their temporal relaxation is of Mittag–Leffler form, with decreasing internal time scale $(\eta_\alpha/[n\omega^2])^{1/\alpha}$. Numerically, the solution (86) is somewhat cumbersome to treat. In order to plot the PDF $P(x,t)$ in Fig. 13, it is preferable to use the closed form solution (we use dimensionless variables)

$$P(x,t) = \frac{1}{\sqrt{2\pi\left(1-\mathrm{e}^{-2t}\right)}}\exp\left(-\frac{\left(x-x_0\mathrm{e}^{-t}\right)^2}{2\left(1-\mathrm{e}^{-2t}\right)}\right) \quad (87)$$

of the Brownian case, and the transformation (68) to construct the fractional analogue.

Figure 13 shows the distinct cusps at the position of the initial condition at $x_0 = 1$. The relaxation to the final Gaussian Gibbs–Boltzmann PDF can be seen from the sequence of three consecutive times. Only at stationarity does the cusp give way to the smooth Gaussian shape of the equilibrium PDF. By adding an additional linear drift V to the harmonic restoring force, the drift term in the FFPE (67) changes to $-\partial(x-V)P(x,t)/\partial x$, and the exponential in expression (87) takes on the form $\exp(-[x-V-(x_0-V)\mathrm{e}^{-t}]/[2(1-\mathrm{e}^{-2t})])$. As displayed in Fig. 14, the strong persistence of the initial condition causes a highly asymmetric shape in the PDF, whereas the Brownian solution shown in dashed lines retains its symmetric Gaussian profile.

Let us finally address the moments of the fractional Ornstein–Uhlenbeck process, Eq. (86). These can be readily obtained either from the Brownian result with the integral transformation (68), or from integration $\int \mathrm{d}x x^n \cdot$ of the FFPE (67). For the first and second moments one obtains:

$$\langle x(t)\rangle = x_0 E_\alpha\left(-\frac{\omega^2 t^\alpha}{\eta_\alpha}\right) \quad (88)$$

and

$$\langle x(t)^2\rangle = x_{\mathrm{th}}^2 + \left(x_0^2 - x_{\mathrm{th}}^2\right)E_\alpha\left(-\frac{2\omega^2 t^\alpha}{\eta_\alpha}\right), \quad (89)$$

respectively. The first moment starts off at the initial position, x_0, and then falls off in a Mittag–Leffler pattern,

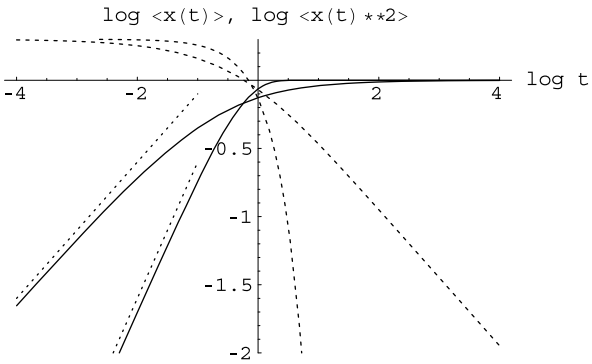

Levy Statistics and Anomalous Transport: Levy Flights and Subdiffusion, Figure 15
First ($x_0 = 2$, *dashed line*) and second ($x_0 = 0$, *full line*) moment of the fractional Ornstein–Uhlenbeck process ($\alpha = 1/2$), in comparison to the Brownian case. $\log_{10}$—$\log_{10}$ scale. The *dotted straight lines* show the initial (sub)diffusive behavior with slopes 1/2 and 1, in the special case $x_0 = 0$ chosen for the second moment

reaching the terminal inverse power-law $\sim t^{-\alpha}$. The second moment turns from the initial value x_0^2 to the thermal value $x_{\text{th}}^2 = k_B T/(m\omega^2)$. In the special case $x_0 = 0$, the second moment measures initial force-free diffusion due to the initial exploration of the flat apex of the potential. We graph the two moments in Fig. 15 in comparison to their Brownian counterparts.

Future Directions

Anomalous diffusion is becoming widely recognized in a variety of fields. Apart from the anomalous spreading of tracers and the consequences for the propagator, additional questions such as ageing and weak ergodicity breaking become important for the understanding of experiments and their modeling on complex systems. At the same time, the theory of anomalous processes is expanding. For instance, regarding questions on the weak ergodicity breaking, the calculation of multipoint moments or the correct introduction of cutoffs are currently being worked on, to complete the world *Beyond Brownian Motion* [55].

Acknowledgments

RM acknowledges partial financial support through the Natural Sciences and Engineering Research Council (NSERC) of Canada and the Canada Research Chairs program of the Government of Canada. AVC acknowledges partial support from the Deutsche Forschungsgemeinschaft (DFG).

Bibliography

1. Abramowitz M, Stegun I (1972) Handbook of Mathematical Functions. Dover, New York
2. Amblard F, Maggs AC, Yurke B, Pargellis AN, Leibler S (1996) Subdiffusion and anomalous local viscoelasticity in actin networks. Physical Review Letters 77:4470–4473. Erratum: Physical Review Letters 81:1136–1136. Reply: Physical Review Letters 81:1135–1135
3. Andronow AA, Chaikin CE, Lefschetz S (1949) Theory of oscillations. University Press, Princeton
4. Atkinson RPD, Rhodes CJ, Macdonald DW, Anderson RM (2002) Scale-free dynamics in the movement patterns of jackals. OIKOS 98:134–140
5. Baeumer B, Benson DA, Meerschaert MM (2005) Advection and dispersion in time and space. Phys A-Stat Mech Appl 350:245–262
6. Barkai E (2001) Fractional Fokker–Planck equation, solution, and application. Phys Rev E 63:046118
7. Barkai E (2002) CTRW pathways to the fractional diffusion equation. Chem Phys 284:13–27
8. Barkai E (2006) Aging in subdiffusion generated by a deterministic dynamical system. Phys Rev Lett 96:104101
9. Barkai E, Silbey RJ (2000) Fractional Kramers equation. J Phys Chem B 104:3866–3874
10. Barkai E, Metzler R, Klafter J (2000) From continuous time random walks to the fractional Fokker–Planck equation. Phys Rev E 61:132–138
11. Bel G, Barkai E (2006) Weak ergodicity breaking in the continuous-time random walk. Phys Rev Lett 94:240602
12. Bogoliubov NN, Mitropolsky YA (1961) Asymptotic methods in the theory of non-linear oscillations. Hindustan Publ. Corp., Delhi (distr. by Gordon & Breach, New York)
13. Bouchaud JP (1992) Weak ergodicity breaking and aging in disordered-systems. J de Phys 2:1705–1713
14. Bouchaud JP, Georges A (1990) Anomalous diffusion in disordered media: statistical mechanisms, models and physical applications. Phys Rep 195:127–293
15. Brockmann D, Geisel T (2003) Particle dispersion on rapidly folding random heteropolymers. Phys Rev Lett 91:048303
16. Brown R (1828) A brief account of microscopical observations made in the months of June, July and August, 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. Philos Mag 4:161–173
17. Carus TL (1975) De rerum natura (50 BCE): On the nature of things. Harvard University Press, Cambridge
18. Chandrasekhar S (1943) Stochastic problems in physics and astronomy. Revs Mod Phys 15:1–89
19. Chechkin AV, Gonchar VY, Szydlowski M (2002) Fractional kinetics for relaxation and superdiffusion in a magnetic field. Phys of Plasmas 9:78–88
20. Chechkin AV, Metzler R, Gonchar VY, Klafter J, Tanatarov LV (2003) J Phys A-Math Gen 36:L537–L544
21. Chechkin AV, Klafter J, Gonchar VY, Metzler R, Tanatarov LV (2003) Phys Rev E 67:010102(R)
22. Chechkin AV, Gonchar VY, Klafter J, Metzler R, Tanatarov LV (2004) J Stat Phys 115:1505–1535
23. Chechkin AV, Gonchar VY, Klafter J, Metzler R (2005) Europhys Lett 72:348–354

24. Chechkin AV, Gonchar VY, Klafter J, Metzler R (2005) Phys Rev E 72:010101
25. Chechkin AV, Gonchar VY, Klafter J, Metzler R (2006) Fundamentals of Lévy flight processes. Adv Chem Phys 133B:439–496
26. Chechkin AV, Sliusarenko AYU, Metzler R, Klafter J (2007) Barrier crossing driven by Lévy noise: Universality and the role of noise intensity. Phys Rev E 75:041101
27. Compte A, Metzler R (1997) The generalized Cattaneo equation for the description of anomalous transport processes. J Phys A-Math Gen 30:7277–7289
28. Davis H (1962) Introduction to nonlinear differential and integral equations. Dover Publications, New York
29. Ditlevsen PD (1999) Anomalous jumping in a double-well potential. Phys Rev E 60:172–179
30. Ditlevsen PD (1999) Observation of alpha-stable noise induced millennial climate changes from an ice-core record. Geophys Res Lett 26:1441–1444
31. Eliazar I, Klafter J (2003) Lévy-driven Langevin systems: Targeted stochasticity. J Stat Phys 111:739–768
32. Eliazar I, Klafter J (2004) On the first passage of one-sided Levy motions. Phys A-Stat Mech Appl 336:219–244
33. Eliazar I, Koren T, Klafter J (2007) Searching circular DNA strands. J Phys-Condens Matter 19:065140
34. Fa KS, Lenzi EK (2003) Power law diffusion coefficient and anomalous diffusion: Analysis of solutions and first passage time. Phys Rev E 67:061105
35. Feller W (1971) An Introduction to Probability Theory and its Appl, vol 2. Wiley, New York
36. Fogedby HC (1994) Lévy flights in random-environments. Phys Rev Lett 73:2517–2520
37. Fogedby HC (1998) Lévy flights in quenched random force fields. Phys Rev 58:1690–1712
38. Frisch U, Frisch H (1995) Universality of escape from a half space for symmetrical random walks. In: Shlesinger MF, Zaslavsky GM, Frisch U (eds) Lévy flights and related topics in physics. Lecture Notes in Physics, vol 450:262–268. Springer, Berlin
39. Gitterman M (2000) Mean first passage time for anomalous diffusion. Phys Rev E 62:6065–6070
40. Gnedenko BV, Kolmogorov AN (1954) Limit Distributions for Sums of Random Variables. Addison-Wesley, Reading
41. Golding I, Cox EC (2006) Physical nature of bacterial cytoplasm. Phys Rev Lett 96:098102
42. Gorenflo R, Mainardi F, Moretti D, Pagnini G, Paradisi P (2002) Fractional diffusion: probability distributions and random walk models. Phys A-Stat Mech Appl 305:106–112
43. Gu Q, Schiff EA, Grebner S, Wang F, Schwarz R (1996) Non-Gaussian transport measurements and the Einstein relation in amorphous silicon. Phys Rev Lett 76:3196–3199
44. Hänggi P, Talkner P, Borkovec M (1990) Reaction-rate theory – 50 years after Kramers. Revs Mod Phys 62:251–341
45. He Y, Burov S, Metzler R, Barkai E (2008) Random Time-Scale Invariant Diffusion and Transport Coefficients. Phys Rev Lett 101:058101
46. Hughes BD (1995) Random Walks and Random Environments, vol 1. Oxford University Press, Oxford
47. Hughes BD (2002) Anomalous diffusion, stable processes, and generalized functions. Phys Rev E 65:035105
48. Imkeller P, Pavlyukevich I (2006) Lévy flights: transitions and meta-stability. J Phys A-Math Gen 39: L237–L246
49. Ingenhousz J (1785) Nouvelles expériences et observations sur divers objets de physique. T. Barrois le jeune, Paris
50. Jespersen S, Metzler R, Fogedby HC (1999) Lévy flights in external force fields: Langevin and fractional Fokker–Planck equations and their solutions. Phys Rev E 59:2736–2745
51. Kappler E (1931) Versuche zur Messung der Avogadro–Loschmidtschen Zahl aus der Brownschen Bewegung einer Drehwaage. Ann Phys 11:233–256
52. Katori H, Schlipf S, Walther H (1997) Anomalous dynamics of a single ion in an optical lattice. Phys Rev Lett 79:2221–2224
53. Klafter J, Zumofen G (1994) Probability distributions for continuous-time random-walks with long tails. J Phys Chem 98:7366–7370
54. Klafter J, Blumen A, Shlesinger MF (1987) Stochastic pathway to anomalous diffusion. Phys Rev A 35:3081–3085
55. Klafter J, Shlesinger MF, Zumofen G (1996) Beyond Brownian motion. Phys Today 49:33–39
56. Klimontovich YUL (1992) Turbulent motion and the structure of chaos: a new approach to the statistical theory of open systems. Kluwer, Dordrecht
57. Koren T, Lomholt MA, Chechkin AV, Klafter J, Metzler R (2007) Leapover lengths and first passage time statistics for Lévy flights. Phys Rev Lett 99:160602
58. Koren T, Chechkin AV, Klafter J (2007) On the first passage time and leapover properties of Lévy motion. Phys A 379: 10–22
59. Koren T, Lomholt MA, Chechkin AV, Klafter J, Metzler R Unpublished
60. Langevin P (1908) The theory of Brownian movement. C R Hebd Seances Acad Sci 146:530–533
61. Lenzi EK, Mendes RS, Fa KS, Malacarne LC, da Silva LR (2003) Anomalous diffusion: Fractional Fokker–Planck equation and its solutions. J Math Phys 44:2179–2185
62. Levandowsky M, White BS, Schuster FL (1997) Random movements of soil amebas. Acta Protozool 36:237–248
63. Lévy P (1954) Théorie de l'addition des variables aléatoires. Gauthier-Villars, Paris
64. Lomholt MA, Ambjornsson T, Metzler R (2005) Optimal target search on a fast-folding polymer chain with volume exchange. Phys Rev Lett 95:260603
65. Lomholt MA, Zaid IM, Metzler R (2007) Subdiffusion and weak ergodicity breaking in the presence of a reactive boundary. Phys Rev Lett 98:200603
66. Lubelski A, Sokolov IM, Klafter J (2008) Non ergodicity Mimics Inhomoneneity in Single Particle Tracking. Phys Rev Lett 100:250602
67. Luchko Y, Gorenflo R (1998) Scale-invariant solutions of a partial differential equation of fractional order. Fractional Calc Appl Anal 1:63–78
68. Lutz E (2004) Power-law tail distributions and nonergodicity. Phys Rev Lett 93:190602
69. Mainardi F, Luchko Y, Pagnini G (2001) The fundamental solution of the space-time fractional diffusion equation. Fractional Calc Appl Anal 4:153–192
70. Meerschaert MM, Benson DA, Scheffler HP, Becker-Kern P (2002) Governing equations and solutions of anomalous random walk limits. Phys Rev E 66:060102
71. Metzler R (2000) Generalized Chapman–Kolmogorov equation: A unifying approach to the description of anomalous transport in external fields. Phys Rev E 62:6233–6245

72. Metzler R (2001) Non-homogeneous random walks, generalized master equations, fractional Fokker–Planck equations, and the generalized Kramers–Moyal expansion. Eur Phys J B 19:249–258
73. Metzler R, Compte A (2000) Generalized diffusion-advection schemes and dispersive sedimentation: A fractional approach. J Phys Chem B 104:3858–3865
74. Metzler R, Klafter J (2000) Accelerating Brownian motion: A fractional dynamics approach to fast diffusion. Europhys Lett 51:492–498
75. Metzler R, Klafter J (2000) Boundary value problems for fractional diffusion equations. Phys A-Stat Mech Appl 278:107–125
76. Metzler R, Klafter J (2000) From a generalized Chapman–Kolmogorov equation to the fractional Klein–Kramers equation. J Phys Chem B 104:3851–3857
77. Metzler R, Klafter J (2000) Subdiffusive transport close to thermal equilibrium: From the Langevin equation to fractional diffusion. Phys Rev E 61:6308–6311
78. Metzler R, Klafter J (2000) The random walk's guide to anomalous diffusion: a fractional dynamics approach. Phys Rep 339:1–77
79. Metzler R, Klafter J (2003) When translocation dynamics becomes anomalous. Biophys J 85:2776–2779
80. Metzler R, Klafter J (2004) The restaurant at the end of the random walk: recent developments in the description of anomalous transport by fractional dynamics. J Phys A-Math Gen 37: R161–R208
81. Metzler R, Nonnenmacher TF (1998) Fractional diffusion, waiting-time distributions, and Cattaneo-type equations. Phys Rev E 57:6409–6414
82. Metzler R, Nonnenmacher TF (2002) Space- and time-fractional diffusion and wave equations, fractional Fokker–Planck equations, and physical motivation. Chem Phys 284: 67–90
83. Metzler R, Sokolov IM (2002) Superdiffusive Klein–Kramers equation: Normal and anomalous time evolution and Levy walk moments. Europhys Lett 58:482–488
84. Metzler R, Barkai E, Klafter J (1999) Anomalous diffusion and relaxation close to thermal equilibrium: A fractional Fokker–Planck equation approach. Phys Rev Lett 82:3563–3567
85. Metzler R, Barkai E, Klafter J (1999) Deriving fractional Fokker–Planck equations from a generalized master equation. Europhys Lett 46:431–436
86. Monthus C, Bouchaud JP (1996) Models of traps and glass phenomenology. J Phys A-Math Gen 29:3847–3869
87. Montroll EW, Weiss GH (1965) Random Walks on Lattices. II. J Math Phys 6:167–181
88. Montroll EW, West BJ (1979) In: Montroll EW, Lebowitz JL (eds) Fluctuation phenomena. North-Holland, Amsterdam, pp 61–176
89. Oldham KB, Spanier J (1974) The fractional calculus. Academic Press, New York
90. Perrin J (1908) Molecular agitation and the Brownian movement. C R Hebd Seances Acad Sci 146:967–970
91. Perrin J (1909) Brownian motion and molecular reality. Ann Chem Phys 18:5–114
92. Pfister G, Scher H (1978) Dispersive (non-Gaussian) transient transport in disordered solids. Adv in Phys 27:747–798
93. Podlubny I (1998) Fractional differential equations. Academic Press, San Diego
94. Priyatinska A, Saichev AI, Woyczynski WA (2005) Models of anomalous diffusion: the subdiffusive case. Phys A 349:375–420
95. Ramos-Fernandez G, Mateos JL, Miramontes O, Cocho G, Larralde H, Ayala-Orozco B (2004) Lévy walk patterns in the foraging movements of spider monkeys (Ateles geoffroyi). Behav Ecol Sociobiol 55:223–230
96. Redner S (2001) A guide to first-passage processes. Cambridge University Press, Cambridge
97. Risken H (1989) The Fokker-Planck equation. Springer, Berlin
98. Saichev AI, Zaslavsky GM (1997) Fractional kinetic equations: solutions and applications. Chaos 7:753–764
99. Samko SG, Kilbas AA, Marichev OI (1993) Fractional Integrals and Derivatives, Theory and Applications. Gordon and Breach, New York
100. Samorodnitsky G, Taqqu MS (1994) Stable non-gaussian random processes: Stochastic models with infinite variance. Chapman and Hall, New York
101. Scher H, Montroll EW (1975) Anomalous transit-time dispersion in amorphous solids. Phys Rev B 12:2455–2477
102. Scher H, Margolin G, Metzler R, Klafter J, Berkowitz B (2002) The dynamical foundation of fractal stream chemistry: the origin of extremely long retention times. Geophys Res Lett 29:1061
103. Schneider WR, Wyss W (1989) Fractional diffusion and wave equations. J Math Phys 30:134–144
104. Sokolov IM, Metzler R (2003) Towards deterministic equations for Levy walks: The fractional material derivative. Phys Rev E 67:010101
105. Sokolov IM, Metzler R (2004) Non-uniqueness of the first passage time density of Levy random processes. J Phys A-Math Gen 37: L609–L615
106. Sokolov IM, Mai J, Blumen A (1997) Paradoxical diffusion in chemical space for nearest-neighbor walks over polymer chains. Phys Rev Lett 79:857–860
107. Sokolov IM, Klafter J, Blumen A (2002) Fractional kinetics. Phys Today 55:48–54
108. Sparre Andersen E (1953) On the fluctuations of sums of random variables. Math Scand 1:263–285
109. Sparre Andersen E (1954) On the fluctuations of sums of random variables II. Math Scand 2:194–223
110. Tolić-Norrelytke I, Munteanu EL, Thon G, Berg-Sørensen K (2004) Anamalous Diffusion in Living Yeast Cells. Phys Rev Lett 93:078102
111. Uchaikin VV (2002) Multidimensional symmetric anomalous diffusion. Chem Phys 284:507–520
112. Uchaikin VV, Zolotarev VM (1999) Chance and stability. Stable distributions and their applications. VSP, Utrecht
113. van Kampen NG (1981) Stochastic Processes in Physics and Chemistry. North-Holland, Amsterdam
114. Visser AW, Thygesen UH (2003) Random motility of plankton: diffusive and aggregative contributions. J Plankton Res 25:1157–1168
115. Viswanathan GM, Afanasyev V, Buldyrev SV, Murphy EJ, Prince PA, Stanley HE (1996) Lévy flight search patterns of wandering albatrosses. Nature 381:413–415
116. Viswanathan GM, Buldyrev SV, Havlin S, da Luz MGE, Raposo EP, Stanley HE (1999) Optimizing the success of random searches. Nat 401:911–914
117. Weiss GH (1994) Aspects and Applications of the Random Walk. North-Holland, Amsterdam

118. West BJ, Nonnenmacher TF (2001) An ant in a gurge. Phys Lett A 278:255–259
119. West BJ, Seshadri V (1982) Linear systems with Lévy fluctuations. Phys A 113:203–216
120. West BJ, Grigolini P, Metzler R, Nonnenmacher TF (1997) Fractional diffusion and Levy stable processes. Phys Rev E 55:99–106
121. Wong IY, Gardel ML, Reichman DR, Weeks ER, Valentine MT, Bausch AR, Weitz DA (2004) Anomalous diffusion probes microstructure dynamics of entangled F-actin networks. Phys Rev Lett 92:178101
122. Zaslavsky GM (2002) Chaos, fractional kinetics and anomalous transport. Phys Rep 371:461–580
123. Zumofen G, Klafter J (1993) Scale-invariant motion in intermittent chaotic systems. Phys Rev E 47:851–863
124. Zumofen G, Klafter J (1994) Spectral random walk of a single molecule. Chem Phys Lett 219:303–309
125. Zumofen G, Klafter J (1995) Absorbing boundary in one-dimensional anomalous transport. Phys Rev E 51:2805–2814
126. Zumofen G, Klafter J (1995) Laminar-localized-phase coexistence in dynamical systems. Phys Rev E 51:1818–1821

Linear and Non-linear Fokker–Planck Equations

Till D. Frank
University of Connecticut, Storrs, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Time–Dependent Solutions and First Order Statistics
Markov Property, 2nd Order and Higher Order Statistics
Chapman–Kolmogorov Equation, Kramers–Moyal Expansion and Drift-Diffusion Estimates
Martingales
From Martingales to Strongly Nonlinear Fokker–Planck Equations
Examples
Linear Nonequilibrium Thermodynamics
Summary and Future Directions
Bibliography

Glossary

Linear Here, linear with respect to a probability density.

Nonlinear Here, nonlinear with respect to a probability density.

Markov process Process for which it is sufficient to have information about the presence in order to make best predictions about the future. Additional information about the past will not improve the predictions.

Martingale process $\hat{Z}$ Process for which the future mean value $\langle Z(t+\Delta t)\rangle$ of a set of realizations $Z^{(i)}$ that is passing at presence t through a certain common state z is the state z: $\langle Z(t+\Delta t)\rangle_{Z(t)=z} = z$. Additional information about states z' visited at times t' prior to t is irrelevant.

Definition of the Subject

Let $\hat{X}$ denote a stochastic process defined on the space Ω and the time interval $[t_0, \infty]$, where t_0 denotes the initial time of the process. We assume that the process $\hat{X}$ can be described in terms of a random variable $X \in \Omega$. More precisely, let $X(t)$ denote the time-dependent evolution of the random variable X for $t \geq t_0$. Then, we assume that the process $\hat{X}$ can be described in terms of the infinitely large set of realizations $X^{(i)}(t)$ of $X(t)$ with $i = 1, 2, \ldots$. The realizations $i = 1, 2, \ldots$ constitute a statistical ensemble. At every time t the probability density P of $\hat{X}$ can be computed from the realizations $X^{(i)}(t)$, that is, from the ensemble by means of

$$P(x,t) = \langle \delta(x - X(t)) \rangle \ , \tag{1}$$

where $\langle\cdot\rangle$ denotes ensemble averaging and $\delta(\cdot)$ is the delta function. We assume that at time t_0 the process is distributed like u. That is, the function $u(x)$ describes the initial probability density of $\hat{X}$ and we have $P(x, t_0) = u(x)$. In general, the evolution of P depends on how the process is distributed at initial time t_0. In order to emphasize this point we will use in what follows the notation $P(x, t; u)$. That is, we interpret Eq. (1) as a conditional probability density with the constraint given by the initial distribution u:

$$P(x,t;u) = \langle \delta(x - X(t)) \rangle_{\langle \delta(x - X(t_0)) \rangle = u(x)} \ . \tag{2}$$

We may also say that we study a family of stochastic processes [80]. Each family member has a label or name which is given by u. For example, consider three experiments in which the evolution of dust particles in the air is observed for Gaussian, Lévy, and Cauchy initial distributions, respectively. It is known that dust particles perform a so-called Brownian random walk. So we would distinguish the three members $\hat{X}_1, \hat{X}_2, \hat{X}_3$ of our family of Brownian walk processes by the names of their initial distributions: Gauss, Lévy, and Cauchy.

Let us consider a stochastic process $\hat{X}$ whose evolution of its probability density P is defined by a partial differential equation of the form

$$\frac{\partial}{\partial t} P(x,t;u) = \left[-\frac{\partial}{\partial x} D_1(x,t) + \frac{\partial^2}{\partial x^2} D_2(x,t) \right] P(x,t;u) \ , \tag{3}$$

where D_1 and D_2 are functions of state x and time t. The functions D_1 and D_2 are referred to as drift- and diffusion coefficients and constitute the Fokker–Planck operator

$$L^0(x,t) = -\frac{\partial}{\partial x} D_1(x,t) + \frac{\partial^2}{\partial x^2} D_2(x,t) \,. \tag{4}$$

The evolution equation (3) is linear with respect to P. In this sense Eq. (3) is a linear partial differential equation. Irrespective of this feature the coefficients D_1 and D_2 may depend in a highly nonlinear fashion on the state x. For example, we may have $D_1 = -x + x^3$.

For appropriately chosen coefficients D_1 and D_2 Eq. (3) describes the probability density P of a Markov process. In this case Eq. (3) is referred to as a Fokker–Planck equation. More precisely, if $\hat{X}$ is a Markov diffusion process whose probability density P is defined by Eq. (3) then Eq. (3) is called a Fokker–Planck equation. Note that roughly speaking a Markov *diffusion* process is a Markov process characterized by a partial differential operator that can be truncated after the second-order partial derivative (see Sect. "Kramers–Moyal Expansion"). In order to distinguish between linear and nonlinear Fokker–Planck equations, we will use the phrase 'linear Fokker–Planck equation' instead of 'Fokker–Planck equation'.

Let us generalize Eq. (3) by assuming that the drift- and diffusion coefficients depend on the probability density P. In this case, Eq. (3) becomes

$$\frac{\partial}{\partial t} P(x,t;u) = \left[-\frac{\partial}{\partial x} D_1(x,t,P(x,t;u)) + \frac{\partial^2}{\partial x^2} D_2(x,t,P(x,t;u)) \right] P(x,t;u) \,. \tag{5}$$

Likewise, the operator (4) is generalized to

$$L(x,t,P(x,t;u)) = -\frac{\partial}{\partial x} D_1(x,t,P(x,t;u)) + \frac{\partial^2}{\partial x^2} D_2(x,t,P(x,t;u)) \,. \tag{6}$$

Equation (5) is nonlinear with respect to $P(x,t;u)$. Since the structure of the differential operator in the bracket of Eq. (5) is equivalent to the structure of the differential operator (4), evolution equations of the form (5) are frequently called nonlinear Fokker–Planck equations. In this context it is important to realize that the phrase 'nonlinear Fokker–Planck equation' does not necessarily imply that we are dealing with a Markov process. The phrase 'nonlinear Fokker–Planck equation' simply means that we are dealing with a nonlinear partial differential equation involving a partial differential operator that exhibits the structure of a Fokker–Planck operator.

Linear and Non-linear Fokker–Planck Equations, Table 1
Definition of linear and nonlinear Fokker–Planck equations based on structure, existence of solutions, and Markov property

Fokker–Planck equations	Linear	Nonlinear
Structure	Eq. (3) –(4)	Eq. (5) –(6)
Solutions	exist	do not necessarily exist
Corresponding processes are Markov processes?	yes	maybe

Note again that if an evolution equation of the form (3) is referred to as a Fokker–Planck equation then it is tacitly assumed that there exists a stochastic process defined by that equation and that this process is a Markov process. Table 1 summarizes how to define linear and nonlinear Fokker–Planck equations by means of structure, existence of solutions, and Markov property.

Linear Fokker–Planck equations are an indispensable tool to describe stochastic processes in a variety of disciplines, see Fig. 1. The theoretical concept of Markov diffusion processes related to linear Fokker–Planck equations is well-established. Researchers, applied scientist, technicians, research and development engineers in general and financial engineers in particular are usually aware that the particular linear Fokker–Planck model they are using belongs to class of Markov models. That is, the world of linear Fokker–Planck equations is closed and connected.

Nonlinear Fokker–Planck equations are used in a variety of fields that are as diverse as the application fields of linear Fokker–Planck equations. Unfortunately, so far, there is no well-established theory connecting all kinds of nonlinear Fokker–Planck equations. There is not even an academic consent about how to define them at all. This is why in Table 1 we used a very general and less constraining definition for nonlinear Fokker–Planck equations. Concepts of nonlinear Fokker–Planck equations are often developed for particular purposes and are not put into other contexts. That is, theoretical results and other achievements are often tailored to serve special needs and are not discussed in a larger framework. Even worse, so far, a well-established link between linear and nonlinear Fokker–Planck equations that applies to the variety of nonlinearities found in the literature does not exist. In sum, the world of nonlinear Fokker–Planck equations is disconnected. Different types of nonlinear Fokker–Planck equations and different application fields of nonlinear Fokker–Planck equations are often not related to each other and nonlinear Fokker–Planck equations are only loosely connected with their linear 'relatives', see Fig. 1. For example, it will most likely take years before a group of liquid crystal scientist will apply a new approach to solve numerically

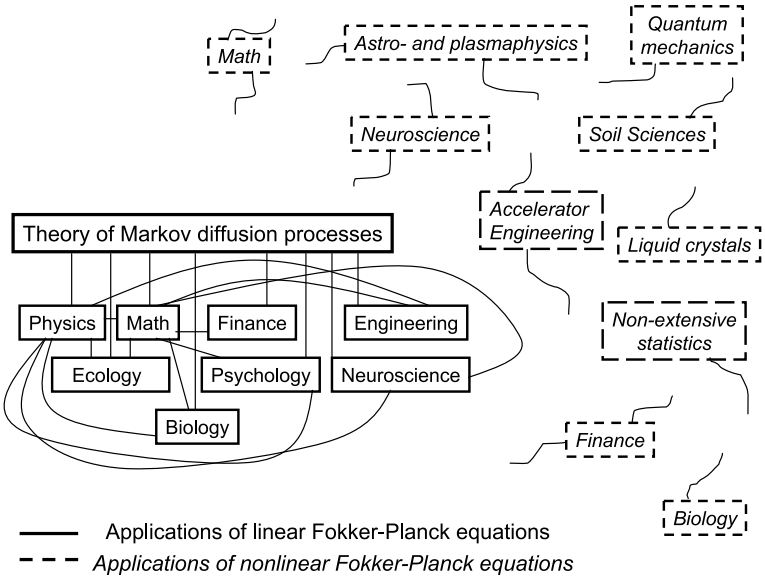

Linear and Non-linear Fokker–Planck Equations, Figure 1
Connected and disconnected applications of linear and nonlinear Fokker–Planck equations

a nonlinear Fokker–Planck equations that was discovered by a group of accelerator scientists. In addition, most likely this new solution method will not be discussed in the context of linear Fokker–Planck equations at all.

Therefore, there is a need for developing an unifying approach to nonlinear Fokker–Planck equations that involves the concept of linear Fokker–Planck equations and applies to all types of nonlinearities and in doing so applies to all kinds of scientific disciplines. Some first efforts in this regard have been made previously [3,31,32,69,70,71,80,85,87,88,127,128,199,200]. In the following sections we will review these efforts, present them in a consistent way, and in doing so make a further effort into this direction.

Introduction

Linear and nonlinear Fokker–Planck equations are widely used to describe stochastic phenomenon, see Fig. 1.

Linear Fokker–Planck equations [98,108,184] have been introduced by Fokker [66] and Planck [176]. In physics, linear Fokker–Planck equations have been used for example to describe Brownian motion, that is, the diffusion of dust particles in air or fluid layers [180]. Linear Fokker–Planck equations have been applied in engineering sciences for example to describe fluctuations in electronic circuits [98]. Linear Fokker–Planck equations have been frequently used in chemistry to model stochastic aspects of chemical reactions [107,228]. In finance, one of the most important applications of the Fokker–Planck theory is option pricing by means of the so-called Black–Scholes model [172]. Linear Fokker–Planck equations of biology systems [101] have been concerned for example with so-called Brownian motors [106,181]. Population diffusion [167] and group behavior [60,157,193] in ecological systems and stochastic neuronal processing [116] are further examples of application fields of linear Fokker–Planck equations. In psychology linear Fokker–Planck models have been proposed for decision making [19,179] and group behavior [194].

Many applications of nonlinear Fokker–Planck equation are related to several benchmark models: the Desai–Zwanzig model [52], the liquid crystal model, [55,115], the Kuramoto–Shinomoto–Sakaguchi model [3,143,207], the Vlasov model, and the nonlinear diffusion equation [12,173]. Let us highlight some of these benchmark models.

Desai–Zwanzig Model

The Desai–Zwanzig model

$$\frac{\mathrm{d}}{\mathrm{d}t}P(x,t;u) = -\frac{\partial}{\partial x}\left[h(x) - \kappa\left(x - \int xP(x,t;u)\,\mathrm{d}x\right) + Q\frac{\partial}{\partial x}\right]P(x,t;u) \quad (7)$$

for $\kappa, Q > 0$ has been proposed by Desai and Zwanzig [52] and Kometani and Shimizu [139] to study collective phenomena in self-organizing systems.

- A Lyapunov functional approach to the Desai–Zwanzig model has been introduced by Shiino [195,196] and since then has found several generalizations [31,32,47,68,80,88,89,127,128,135,192,197,198,199,200]. With such a Lyapunov functional at hand, the stability of stationary probability densities, collective phenomena and bifurcations can be studied by means of Lyapunov's direct method.
- The original Desai–Zwanzig model and Various modifications of it have been discussed [45,146,148].
- The additive noise term in Eq. (7) has been replaced by a multiplicative noise term [117] in order to study the interplay between the nonlinearity and the multiplicative noise [17,159,239].
- Fluctuation-dissipation theorems for stochastic processes described by the Desai–Zwanzig model have been derived [59,76,158].
- The Desai–Zwanzig model has frequently been used as a mean field approximation of spatially distributed systems with diffusive coupling. By means of such a mean field approximation analytical result have been derived and compared with numerical simulations [96,97,225,226,227].

Liquid Crystal Model

The nonlinear Fokker–Planck equation proposed by Hess [115] and Doi and Edwards [55] reads

$$\frac{\partial}{\partial t}P(\mathbf{x},t;u) = D_r\mathbf{L}\cdot\left\{\mathbf{L} + \frac{1}{kT}\left[\mathbf{L}\,e(\mathbf{x},P)\right]\right\} P(\mathbf{x},t;u) \quad (8)$$

with $D_r, k, T > 0$ and $\mathbf{L} = \mathbf{x} \times \partial/\partial\mathbf{x}$. The function $e(\mathbf{x}, P)$ describes the self-consistent potential of the Maier–Saupe mean field force. For processes $\hat{X}$ that exhibit cylindrical symmetry $e(\mathbf{x}, P)$ reads

$$e(\theta, P) = -U_0 kT \frac{3\cos^2\theta - 1}{2}\left\langle\frac{3\cos^2\theta - 1}{2}\right\rangle, \quad (9)$$

where θ is related to the unit vector $\mathbf{x}$ by $\mathbf{x} = (\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta)$. Equation (8) and generalization of it have been extensively studied in the literature [64,65,118,120,121,145] (see also [163] in general and Sect. 6.3.2 in [163] in particular). We will return to this model in Sect. "Liquid Crystal Model".

Winfree and Kuramoto Model

Winfree's seminal studies on synchronization among animal populations [234,235] supported the interest in the nonlinear Fokker–Planck equation

$$\frac{\partial}{\partial t}P(x,t;u) = \left\{-\frac{\partial}{\partial x}\left[h(x) - \kappa \cdot \int \sin(x-y)P(y,t;u)\,\mathrm{d}y\right] + Q\frac{\partial^2}{\partial x^2}\right\} P(x,t;u)\,, \quad (10)$$

that has been proposed by Kuramoto and co-workers [143]. In Eq. (10) $h(x)$ is a 2π-periodic function and κ, $Q > 0$.

- While the Kuramoto–Shinomoto–Sakaguchi model involves an interaction term $\int a(x,y)P(y,t)\,\mathrm{d}y$, the model originally proposed by Winfree exhibits a coupling term of the form $\int a(y)P(y,t)\,\mathrm{d}y$. Models of this latter kind have also been address in [8,147,178].
- The Kuramoto–Shinomoto–Sakaguchi model describes an ensemble of phase oscillators. The eigenfrequencies of the phase oscillators do not occur in Eq. (10) because Eq. (10) describes an ensemble of phase oscillators exhibiting the same eigenfrequency ω. In this case the common eigenfrequency ω can then be eliminated by means of a variable transformation into a rotating frame [80]. However, in general, we may think of ensembles of coupled phase oscillators with different eigenfrequencies. In this context, the question arises as to what extent oscillators with different eigenfrequencies synchronize their behavior [2,7,21,38,141,174,188,209].
- Coupled phase oscillator models of the form (10) have been used to describe associative memories [237,238].
- Just as for the Desai–Zwanzig model, the interplay between multiplicative noise and the nonlinearity of the Kuramoto–Shinomoto–Sakaguchi model has been investigated in several studies [136,170,182,183].
- The *sine*-coupling term in Eq. (10) has been replaced by higher-order coupling functions $\sin(2z)$, $\sin(3z)$, ... [6,40,41,111,143]. In this context, Daido proposed the so-called order function [40,41] that generalizes the notion of cluster phases and cluster amplitudes [143]. This order function has also been related to experimental data [240].
- The Kuramoto–Shinomoto–Sakaguchi has found clinical applications in the context of Parkinsonian disease [216,217,218] (see also [215]).

Vlasov–Fokker–Planck Model

Vlasov–Fokker–Planck models frequently describe particle systems with electromagnetic interactions between charged particles. A typical examples of a Vlasov–Fokker–

Planck equation is shown here [13,137]:

$$\frac{\partial}{\partial t}P(\mathbf{v},t;u) = -\sum_{i=1}^{3}\frac{\partial}{\partial v_i}D_i(\mathbf{v},P)P + \sum_{i,k=1}^{3}\frac{\partial^2}{\partial v_i \partial v_k}D_{ik}(\mathbf{v},P)P\,. \quad (11)$$

Equation (11) involving the drift- and diffusion-coefficients

$$D_i(\mathbf{v},P) = a\frac{\partial}{\partial v_i}\int_{\Omega}\frac{P(\mathbf{v}',t;u)}{|\mathbf{v}-\mathbf{v}'|}\,\mathrm{d}^3v'\,,$$

$$D_{ik}(\mathbf{v},P) = b\frac{\partial^2}{\partial v_i v_k}\int_{\Omega}\left|\mathbf{v}-\mathbf{v}'\right|P(\mathbf{v}',t;u)\,\mathrm{d}^3v'\,. \quad (12)$$

Time-dependent solutions [5,149,160,187,213] for Eq. (12) and generalization of Eq. (12) that account for additional drift forces [26,62] have been studied. In particular, numerical methods using short-time propagators (see Sect. "Short-Time Propagator") have been developed for Vlasov–Fokker–Planck equations of the form (12) [56,57,205]. Such nonlinear Vlasov–Fokker–Planck equations play important roles in plasma physics [13,137,160] and astrophysics [16,144]. In general, astrophysical problems often require a stochastic description in terms of nonlinear Fokker–Planck equations [33,200,206]. Finally note that Vlasov–Fokker–Planck models have been used in accelerator physics and accelerator engineering to examine instabilities in particle beams [72,82,113,114,203,211,230].

Nonlinear Diffusion Equation, Nonextensive Thermostatistics, and Semi–classical Descriptions of Quantum Systems

The nonlinear diffusion equation [12,173] reads

$$\frac{\partial}{\partial t}P(x,t;u) = -\frac{\partial}{\partial x}h(x)P(x,t;u) + \frac{\partial^2}{\partial x^2}D(P(x,t;u))\,, \quad (13)$$

where $D(P)$ is a diffusion coefficient that depends on the probability density $P(x,t)$ of $X(t)$. In the original version of the nonlinear diffusion equation, the drift coefficient $h(x)$ vanishes and the diffusion coefficient is proportional to a power of P. In general, there might be a more complicated dependence of D on P [37,43].

- Since fluid flow through porous materials is a important application of the nonlinear diffusion equation, nonlinear diffusion plays a crucial role in soil sciences [14]. In biology, nonlinear diffusion equations of the form (13) seem to capture particular aspects of population diffusion [104,167].
- The nonlinear diffusion equation (13) provides a link to stochastic processes subjected to Tsallis's nonextensive thermostatistics [1,219,220,221]. For $D(P) \propto P^q$ Eq. (13) becomes

$$\frac{\partial}{\partial t}P(x,t;u) = -\frac{\partial}{\partial x}h(x)P(x,t;u) + Q\frac{\partial^2}{\partial x^2}P^q(x,t;u)\,. \quad (14)$$

  Plastino and Plastino showed that stationary distributions of Eq. (14) correspond to canonical distribution that can be derived in a nonextensive framework [177]. Equation (14) has turned out to be a testbed for various analytical and numerical studies [22,31,32,35,58,200,222]. Alternative nonlinear Fokker–Planck equation related to the Tsallis statistics have been derived from master equations in [39,161]. In addition, Eq. (14) has more recently discussed in finance in the context of a generalized Black–Scholes model for option pricing [23,24,25,229] and fat tail distributions [36,156].
- The nonlinear diffusion equation (13) is also related to semi-classical descriptions of quantum mechanical systems. For an appropriate choice of D nonlinear Fokker–Planck equations for Fermi–Bose and Einstein–Dirac systems have been derived from Eq. (13) [80,85]. Alternative forms of nonlinear Fokker–Planck equations have been derived from quantum mechanical Boltzmann equations [127,129,130], on the basis of energy balance equations [223,224], and by means of other techniques [75,79,125,206]. We will return to semi-classical quantum mechanical descriptions in Sect. "Semi-classical Description of Quantum Systems".

In addition nonlinear Fokker–Planck equations have been turned out to be useful models to describe stochastic aspects of Josephson arrays [105,233], Landau damping [210], arrays of semi-conductor lasers [142], charge density waves [20], and neurons coupled by Hodgkin–Huxley equations [109,110].

Stochastic systems composed of different kinds of interacting subsystems or species have been modeled in terms of multivariate nonlinear Fokker–Planck equations [44,95,119]. For example, the collective behavior of coupled relaxation oscillators has been studied [236]. Networks of neural oscillators as defined by the Wilson–Cowan model, [191], the two-dimensional Morris–Lecar system [109], the FitzHugh–Nagumo equations [112,126,169], and the Hidmarsh–Rose equations [186] have been studied.

The dynamics of mean field coupled phase oscillators under the impact of inertia effects [4] and related models

have attracted considerable attention. Bridge vibrations induced by pedestrian walking have discussed in this context recently [61,208]. Models for circadian rythms have been examined [42].

Solutions of the Kadar–Parisi–Zhang equation have been examined by means of nonlinear Fokker–Planck equations [99,151]. In doing so, the growth of surfaces and roughening phenomena have been studied.

Wetting processes [51], interacting Brownian motors [15,189], and spatially distributed phase oscillators [133,134] have been analyzed by means of the nonlinear Fokker–Planck perspective.

In the mathematical literature, a seminal study on nonlinear Fokker–Planck equations of the Burgers equation type was due to McKean Jr. [153]. In particular, the convergence of stochastic processes described by multivariate linear Fokker–Planck equations to processes described by nonlinear Fokker–Planck equations [29,45,46,53,54,67,94,103,124,153,154,155,162,168,175,185] and martingales of stochastic processes defined by nonlinear Fokker–Planck equations have been addressed [54,67,94,102,103,124,154,155,168]. Moreover, the propagation of molecular chaos has been studied [20,154,155]. The convergence of transient solutions of nonlinear Fokker–Planck equations to stationary ones has been examined by means of functionals that are similar to the Lyapunov functionals introduced by Shiino (see above) [9,10,11,27,28]. In addition, from a purely mathematical perspective nonlinear Fokker–Planck equations should be consider as nonlinear parabolic partial differential equations that have been discussed in several textbooks [91].

In what follows we will show that there is a common theoretical framework that unifies most of the aforementioned studies on nonlinear Fokker–Planck equations and includes the theory of linear Fokker–Planck equations as a special case. This common theoretical framework is rooted in the notion of Markov processes and martingales.

Time–Dependent Solutions and First Order Statistics

Linear Case

Equation (3) defines the evolution of P given an initial distribution $u(x)$. The norm of the probability density P equals unity for all times provided that the norm of $u(x)$ equals unity. That is, if $\int_\Omega u(x)\mathrm{d}x = 1$ holds we have $\int_\Omega P(x,t;u)\mathrm{d}x = 1$ for $t \geq t_0$. We can see this by integrating Eq. (3) with respect to x. For appropriate boundary conditions it can be shown by partial integration that the right hand side vanishes which implies that $\mathrm{d}[\int_\Omega P(x,t;u)\,\mathrm{d}x]/\mathrm{d}t = 0$ holds. The formal solution of Eq. (3) reads

$$P(x,t;u) = \exp\left\{\int_{t_0}^{t} \mathrm{d}z\, L^0(x,z)\right\} u(x)\,, \tag{15}$$

where L^0 is defined in Eq. (4). Equation (15) can be used to solve Eq. (3) numerically (see Vol. 1, Sect. 6.5 in [108]). Let t_n denote a discrete time point $t_n = t_0 + n\Delta t$ with $n = 0, 1, 2, \ldots$, where Δt is the interval of a single time step and should be small. Let us define $P_n(x;u) = P(x,t_n;u)$. Then, we have

$$P_{n+1}(x;u) = \left\{1 + \Delta t L^0(x,t_n)\right\} P_n(x;u) \tag{16}$$

with $P_0 = u(x)$. If $\hat{X}$ corresponds to an autonomous processes, then the coefficient D_1 and D_2 do not depend on t. In this case P_n can be expressed in terms of u as

$$P_n(x;u) = \left[1 + \Delta t L^0(x)\right]^n u(x)\,. \tag{17}$$

Numerical solutions converge to exact solutions in the limit $\Delta t \to 0$.

Nonlinear Case

For appropriately chosen drift- and diffusion coefficients Eq. (5) exhibits time-dependent solutions P. By analogy with the linear case, these solutions are normalized to unity provided that appropriate boundary conditions hold and that the initial probability density is normalized to unity. Solutions of Eq. (5) are formally defined by

$$P(x,t;u) = \exp\left\{\int_{t_0}^{t} \mathrm{d}z\, L(x,z,P(x,z;u))\right\} u(x)\,. \tag{18}$$

The time-dependent solutions P can be computed numerically by analogy to the linear case discussed above. That is, the probability densities $P_n(x;u) = P(x,t_n;u)$ on the discrete time grid $t_0, t_0 + \Delta t, t_0 + 2\Delta t, \cdots$ can be computed from

$$P_{n+1}(x;u) = \{1 + \Delta t L(x,t_n,P_n(x;u))\}\, P_n(x;u) \tag{19}$$

with $P_0 = u(x)$ and $n = 0, 1, 2, \ldots$. If drift- and diffusion coefficients do not explicitly depend on t, we find that the operator L still depends on t because it depends on the time-dependent solution P that in turn depends on t. Consequently, it is not trivial to generalize Eq. (17) to the nonlinear case. If the drift- and diffusion coefficients do not depend explicitly on time t and the process converges to a stationary one, then the nonlinear Fokker–Planck operator L does not depend on time. This implies that the stationary probability density P_{st} satisfies

$$P_{\mathrm{st}} = \{1 + \Delta t L(x,P_{\mathrm{st}})\}\, P_{\mathrm{st}} + O(\Delta t^2)\,. \tag{20}$$

Finally, note that we do not necessarily need to define the

formal solution with respect to the initial probability density u as in Eq. (18). We can solve the nonlinear Fokker–Planck equation on the time interval $[t_0, t]$ by splitting the solution in two intervals $[t_0, t']$ and $[t', t]$. Then, we obtain

$$P(x, t; u) = \exp\left\{\int_{t'}^{t} \mathrm{d}z\, L(x, z, P(x, z; u))\right\} P(x, t'; u)\,. \quad (21)$$

Equation (21) can be solved iteratively by means of Eq. (19) yielding a mapping $T_{\Delta t}\colon P(x, t; u) = T_{t-t'}[P(x, t'; u)]$ with $\Delta t = t - t'$.

Markov Property, 2nd Order and Higher Order Statistics

Conditional Probability Densities

Let $p(x)$ define the probability density of the time-dependent random variable X at time t given that X assumed at earlier times $t', t'', t''', \ldots$ with $t \geq t' > t'' > t''' > \ldots$ particular values $x', x'', x''', \ldots$. Then p is defined by

$$p = \langle \delta(x - X(t))\rangle_{X(t')=x', X(t'')=x'', X(t''')=x''', \ldots}\,. \quad (22)$$

In order to point out the information that we need to compute p, we write

$$p = p(x, t|x', t'; x'', t''; x''', t''', \ldots)\,. \quad (23)$$

A conditional probability density is a relation that gives us estimates about future events and tells us what we need to know in order to be able to calculate these estimates. In our example given by Eq. (23) we see that we need the information of random values X at times $t' > t'' > t''' > \cdots$ in order to make a prediction about the statistics or probability density of X at time t. Alternatively, we may say that the conditional probability density depends on a list of variables. In the context of Markov processes this alternative viewpoint however gives raise to a problem that will be discussed below.

If $\hat{X}$ is a Markov process then the information about the stochastic process available at one particular time t' is sufficient to make predictions about the future $t \geq t'$. Adding more information about how the process evolved in the past before t' does not improve these predictions. That is, the information of events at time t' are sufficient to make statistical estimates about events at time $t \leq t'$. An alternative definition of a Markov process is that a Markov process exhibits a conditional probability density $p(x, t|\cdot)$ that depends only on one time point prior to t. That is, according to the first definition we look from time t' into the future, whereas according to the second definition we look in the opposite direction: we look from time t into the past.

For example, in order to describe the probability density $p(T)$ of the temperature T in Boston (USA) on December 1st, 2007, given that on November 1st, 2007, the temperature was 2 degrees Celsius and on October 1st, 2007, the temperature was 3 degrees Celsius, we would define the conditional probability density $p(T,\ \text{Dec 1st 2007} \mid T = 2,\ \text{Nov 1st 2007};\ T = 3,\ \text{Oct 1st 2007})$. If the temperature T as a function of time t is a Markov process, it is sufficient to know the temperature at November 1st in order to compute the probability density $p(T)$ at December 1st. For example, we would obtain the same function $p(T)$ for the conditions (i) and (ii) with (i) $T = 2°C$ on Nov 1st and $T = 3°C$ on Oct 1st and (ii) $T = 2°C$ on Nov 1st and $T = 5°C$ on Oct 1st. That is, we would have

$$\begin{aligned} &p(T,\ \text{Dec 1st 2007} \mid T = 2,\ \text{Nov 1st 2007};\\ &\qquad\qquad T = 3,\ \text{Oct 1st 2007})\\ &= p(T,\ \text{Dec 1st 2007} \mid T = 2,\ \text{Nov 1st 2007};\\ &\qquad\qquad T = 5,\ \text{Oct 1st 2007})\\ &= p(T,\ \text{Dec 1st 2007} \mid T = 2,\ \text{Nov 1st 2007})\,. \end{aligned} \quad (24)$$

The information about the October temperature is irrelevant. In this sense the conditional probability density would depend on the November temperature but would not depend on the October temperature.

A problem that arises in the context of the definition of Markov process is as follows. Suppose that there is a purely deterministic dynamical aspect involved in a stochastic process. In our example about Boston temperatures we may think of the annual period changes of the temperature that are related to the annual changes in distance and declination angle between earth and sun. Let us assume that distance and declination angle change periodically in a purely deterministic fashion such that the distance and declination angle at November 1st can be computed from the distance and declination angle at January 1st by a simple one-to-one mapping. Then the question arises: does the temperature in Boston on December 1st depend on the distance and declination angle of November 1st as suggested by $p(T,\ \text{Dec 1st 2007} \mid T = 25,\ \text{Nov 1st 2007})$ or does it depend on the distance and declination angle of January 1st. In the former case we have a Markov conditional probability density. In the latter case we would need to write p like

$$\begin{aligned} &p(T,\ \text{Dec 1st 2007} \mid T = 25,\ \text{Nov 1st 2007};\\ &\qquad\qquad \text{distance and angle},\ \text{Jan 1st 2007}) \end{aligned}$$

indicated that we are dealing with a non-Markovian process. The situation becomes even worse if we take into consideration that the earth-sun distance and the declination

angle at January 1st 2008 can be computed from the information known at November 1st using our simple one-to-one mapping. Therefore, we may say that the temperature estimate for December 1st, 2007, depends on a future event, namely, the earth-sun distance and declination angle given at January 1st, 2008. The conditional probability density would assume the form

$$p(T,\ \text{Dec 1st 2007} \mid T = 25,\ \text{Nov 1st 2007};\\ \text{distance and angle},\ \text{Jan 1st 2008})\,,$$

which would suggest again that we are dealing with a non-Markovian and – to a certain extent – non-causal process. We can solve this problem by realizing that purely deterministic relationships in time that represent external driving forces are irrelevant for the distinction between Markov and non-Markov processes. We can completely determine such external driving force by a parameter set $\{t_0, A_1, A_2, \dots\}$ that describe the initial state of the driving forces. Although this initial state is related to the initial time t_0 of the Markov process the conditional probability density does not actually depend (i. e. it does not explicitly depend) on t_0. Likewise the conditional probability density does not actually depend on the parameters $\{A_1, A_2, \dots\}$. The information that we have at time t' includes the information about the driving force at time t' and therefore the information about the driving force at all times $t \in [t_0, \infty)$. Consequently, the information at time t' is sufficient to predict how the driving force will evolve in the future at times $t \le t'$. There is no need to assess information about events prior to t' or information about events that will happen in the future at times larger than t' in order to determine the evolution of the deterministic driving force.

Let us summarize. A stochastic process $\hat{X}$ is called a Markov process if information about the process at time t' is sufficient to make predictions about future events. This implies that the conditional probability density p defined in Eq. (23) can be simplified like

$$p(x, t \mid x', t'; x'', t''; x''', t''', \dots) = p(x, t \mid x', t')\,. \quad (25)$$

Note that we may say that p depends only on the state x' related to the time t' in the sense that the information at time t' is sufficient to predict how X will be distributed at time $t \le t'$. We may say it is sufficient to *best* predict future events where *best* refers to the fact that adding additional information about the past does not improve our predictions.

Let us illustrate this issue by another example. Let $p(x, t \mid X = \theta)$ denote the probability density of X at time t given that X equals the function θ in the interval $[t_0, t']$ with $t' \le t$. If X describes a Markov process we have

$$\begin{aligned} p(x, t \mid X = \theta) &= p(x, t \mid \theta(t'), t') \\ &= p(x, t \mid x', t') \end{aligned} \quad (26)$$

with $x' = \theta(t')$.

Linear Fokker–Planck Equations

As mentioned in Sect. "Definition of the Subject", linear Fokker–Planck equations describe Markov processes [98, 184]. Markov processes related to linear Fokker–Planck equations of the form (4) have conditional probability densities defined by

$$\frac{\partial}{\partial t} p(x, t|x', t') = L^0(x, t)\, p(x, t|x', t') \quad (27)$$

with $\lim_{t \to t'} p(x, t|x', t') = \delta(x - x')$. The conditional probability density p is also called the fundamental solution or Green's function of the Fokker–Planck equation (3). In general a stochastic process $\hat{X}$ is completely defined in terms of the joint probability density

$$\begin{aligned} P(x_n, t_n; x_{n-1}, t_{n-1}; \dots; x_0, t_0) = \langle \delta(x_n - X(t_n)) \\ \cdot \delta(x_{n-1} - X(t_{n-1})) \cdots \delta(x_0 - X(t_0)) \rangle\,, \end{aligned} \quad (28)$$

where n can assume arbitrarily large integer numbers. In particular, if $\hat{X}$ is a Markov process then this joint probability density can be computed from p and u like

$$\begin{aligned} P(\cdot) &= p(x_n, t_n|x_{n-1}, t_{n-1}) \\ &\cdot p(x_{n-1}, t_{n-1}|x_{n-2}, t_{n-2}) \cdots p(x_1, t_1|x_0, t_0) u(x_0)\,. \end{aligned} \quad (29)$$

Consequently, the linear Fokker–Planck equation (3) defines completely a Markov process via the associated evolution equation (27) and the initial distribution u.

In particular, the time-dependent probability densities $P(x, t; u)$ and $P(x, t'; u)$ with $t \ge t'$ are related to each other by means of a *linear* functional

$$P(x, t; u) = \int_\Omega p(x, t|x', t') P(x', t'; u)\, \mathrm{d}x'\,. \quad (30)$$

That is, the Green's function p induces a functional that is linear with respect to $P(x', t'; u)$.

Langevin Equations of Linear Fokker–Planck Equations

The stochastic trajectories $X(t)$ of the Markov process $\hat{X}$ defined by Eq. (3) can be computed from the Ito–Langevin equation [34,98,184]

$$\frac{\mathrm{d}}{\mathrm{d}t} X(t) = D_1(X(t), t) + \sqrt{D_2(X(t), t)}\, \Gamma(t)\,, \quad (31)$$

where $\Gamma(t)$ denotes a Langevin force normalized to the

delta function like $\langle \Gamma(t)\Gamma(t')\rangle = 2\delta(t-t')$. From the Langevin equation (31) it follows again that we are dealing with a Markov process. Information about one reference time t' is sufficient to compute the future behavior of the trajectory $X(t)$ with $t \geq t'$. On a discrete time grid the stochastic trajectories or realizations of $\hat{X}$ can be computed iteratively like [138,184]

$$X_{n+1} = X_n + \Delta t\, D_1(X_n, t_n) + \sqrt{\Delta t\, D_2(X_n, t_n)}\,\epsilon_n \quad (32)$$

with $X(t_n) = X_n$, $t_n = t_0 + n\Delta t$, and $n = 0, 1, 2, \ldots$ Here, ϵ_n are independent Gaussian distributed random numbers with vanishing mean and variance 2. That is, we have $\langle \epsilon_n \rangle = 0$ and $\langle \epsilon_n \epsilon_m \rangle = 2\delta_{nm}$, where δ_{nm} is the Kronecker symbol. Finally, the probability density W of ϵ_n at every step n is given by

$$W(\epsilon_n) = \frac{1}{\sqrt{4\pi}} \exp\left\{-\frac{\epsilon_n^2}{4}\right\} . \quad (33)$$

Note that we do not necessarily need to start the iteration scheme at t_0. The scheme (32) can be started at any time step n. Moreover, in order to compute the subsequent time steps it is sufficient to have information about the random variable X at time t_n. Consequently, the sequence $X_n, X_{n+1}, X_{n+2}, \ldots$ computed from Eq. (32) describes a trajectory of a Markov process.

Strongly Nonlinear Fokker–Planck Equations

In Sect. "Definition of the Subject" we pointed out that there is some kind of asymmetry in the definition of linear and nonlinear Fokker–Planck equations. While a linear Fokker–Planck equation defines a stochastic process, a nonlinear Fokker–Planck equation defines at best the evolution of a probability density $P(x, t)$. That is, if solutions of Eq. (5) exist for $u \in U$, then Eq. (5) defines the evolution of first-order statistical properties of a stochastic process $\hat{X}$ such as the time-dependent probability density, the mean and the variance of the process $\hat{X}$. In any case, Eq. (5) does not define second- and higher-order statistical quantities such as correlation functions and conditional probability densities [77]. In particular, the time-dependent solutions P of Eq. (5) in general can not be used to construct Green's functions of Markov processes because they do not necessarily correspond to Green's functions of Markov processes [73]. Note that this is not a peculiarity of stochastic processes defined by nonlinear Fokker–Planck equations. In fact, time-dependent solutions P of linear Fokker–Planck equations Eq. (3) involving explicitly time-dependent coefficients D_1 and D_2 do not necessarily correspond to Green's functions. Mathematically speaking, let $P(x, t; u = \delta(x - x_0))$ denote the probability density of a process $\hat{X}$ defined by a non-autonomous linear Fokker–Planck equation or by a nonlinear Fokker–Planck equation and let $p(x, t|x', t')$ denote the conditional probability density of that process $\hat{X}$, then we have [73]

$$P\big(x, t; u = \delta(x - x')\big)_{t_0 = t'} \text{ is not necessarily equivalent to } p\big(x, t \mid x', t'\big) , \quad (34)$$

where $P(x, t; u = \delta(x - x'))_{t_0=t'}$ means that we take the time-dependent solution $P(x, t; u = \delta(x - x_0))$ and replace in this expression x_0 by x' and t_0 by t'.

Let us return to the issue how to define a stochastic process $\hat{X}$ on the basis of a nonlinear Fokker–Planck equation (5). In order to do so, we need to define appropriate constraints such that out of all possible stochastic processes that exhibit a time-dependent probability density P defined by Eq. (5) one particular process is selected. In what follows, we will discuss one particular set of constraints [77]. As we will see the stochastic processes thus defined exhibit the Markov property.

Let U denote a set of initial probability densities u. That is, U is a set of functions or a space of functions. Let $P(x, t; u)$ denote the solution of the nonlinear Fokker–Planck equation

$$\frac{\partial}{\partial t} P(x, t; u) = L(x, t, P(x, t; u))P(x, t; u) \quad (35)$$

with

$$L(x, t, P(x, t; u)) = -\frac{\partial}{\partial x} D_1(x, t, P(x, t; u)) + \frac{\partial^2}{\partial x^2} D_2(x, t, P(x, t; u)) \quad (36)$$

for an initial distribution $u \in U$. Let us introduce the associated drift- and diffusion coefficients $\tilde{D}_1$ and $\tilde{D}_2$ by

$$\begin{aligned} \tilde{D}_1(x, t; u) &= D_1(x, t, P(x, t; u)) , \\ \tilde{D}_2(x, t; u) &= D_2(x, t, P(x, t; u)) . \end{aligned} \quad (37)$$

That is, for any $u \in U$ Eq. (35) is solved analytically or by numerical iteration (19). The solution is substituted into the drift- and diffusion coefficients D_1 and D_2. The coefficients thus obtained are the functions $\tilde{D}_1$ and $\tilde{D}_2$ associated to D_1 and D_2. Let us assume that for all $u \in U$ the evolution equation

$$\frac{\partial}{\partial t} p(x, t|x', t') = \left[-\frac{\partial}{\partial x} D_1(x, t; u) + \frac{\partial^2}{\partial x^2} D_2(x, t; u) \right] \cdot p(x, t|x', t') \quad (38)$$

has a fundamental solution or Green's function. Then, this solution p and its corresponding initial distribution u define a Markov process. In [77] nonlinear Fokker–

Planck equations that induce evolution equations (38) with fundamental solutions were called strongly nonlinear Fokker–Planck equations. Note that nonlinear Fokker–Planck equations (5) do not necessarily exhibit the property of being strongly nonlinear. Note also that in some applications it might be worth to define carefully the set U of initial probability densities u such that a nonlinear Fokker–Planck equation under consideration becomes strongly nonlinear.

As indicated above the time-dependent probability densities P of a nonlinear Fokker–Planck equation depend on the initial distribution u. Likewise, the associated coefficients $\tilde{D}_1$ and $\tilde{D}_2$ depend on u. As a result the conditional probability densities $p(x, t|x', t')$ depend on u as well. For this reason, the notation $p(x, t|x', t'; u)$ has been suggested. Unfortunately, this notation is likely to cause confusion because one might think that p depends not only on the time t' but also on the initial time t_0 which seems to be incompatible with the notion of a Markov conditional probability density [83,152]. In fact, this confusion results from the second alternative way to define Markov processes that has been discussed above. The evolution of the function $P(x, t; u)$ is a purely deterministic one. That is, $P(x, t; u)$ represent a deterministic driving force for the purpose of computing the conditional probability density. The distribution u is just a parameter which determines the initial value of this driving force. In this context note again that the conditional probability density of a Markov process in general depends on parameters and in particular can depend on the initial time t_0 and other parameters $\{A_1, A_2, \dots\}$ that define the initial state of a driving force. Consequently, the notation $p(x, t|x', t'; u)$ does not imply a contradiction with the notion of a Markov process. For example, the non-autonomous Langevin equation

$$\frac{\partial}{\partial t}X(t) = -\gamma X(t) - A\cos(\omega(t - t_0)) + \sqrt{Q}\Gamma(t) \quad (39)$$

with $\gamma, A, \omega, Q > 0$ defines a Markov process that is driven by a harmonic force $-A\cos(\omega(t - t_0))$. That is, the harmonic force has amplitude A at the beginning of the process. The conditional probability density of that process depends on the parameter γ, ω, Q but also on the parameters t_0 and A which correspond to the initial amplitude and time. We have (see Sect. 3.7.3 in [80])

$$p = p(x, t \mid x', t', \gamma, \omega, Q, A, t_0)\,. \quad (40)$$

Nevertheless, in what follows we will develop a slightly different notation for conditional probability densities p of strongly nonlinear Fokker–Planck equations that is in line with the first definition of Markov processes discussed above and will be helpful to elucidate that the functions p reflect indeed Markov processes.

Let us exploit first the fact that if Eq. (35) is a strongly nonlinear Fokker–Planck equations then time-dependent solutions $P(x, t; u)$ of Eq. (35) exist for $u \in U$ and are related to their initial probability densities u by a one-to-one mapping T_t. That is, for every t we have $P(x, t; u) = T_t[u(x)]$. For an explicit construction of the map T_t see for example Eq. (18). Likewise, we have $P(x, t'; u) = T_{t'}[u(x)]$. Using the inverse of T we can map u to P like $u(x) = T_{t'}^{-1}[P(x, t; u)]$. Substituting these expressions into $p(x, t \mid x', t'; u)$, we obtain $p(x, t \mid x', t'; T_{t'}^{-1}[P(x, t'; u)]$. This result demonstrates that the information about the stochastic process $\hat{X}$ at time t' is sufficient to predict the future at $t > t'$. We can regard the conditional probability density p as a function that does not depend explicitly on u but it depends explicitly on the state of the driving force P at time t'. In line with this remark, we introduce the conditional probability densities $p(x, t \mid x', t', P(x', t'; u))$.

Let us dwell on the interpretation of the conditional probability density $p(x, t \mid x', t', P(x', t'; u))$. To this end, we need to discuss briefly the notion of a particular conditional averaging that is important in this context and will become important later on as well. Let us assume that we make observations of realizations of a stochastic process $\hat{X}$ for which the following two conditions hold: (i) $X(t') = x'$ and (ii) the ensemble of all realization is distributed like P at time t'. Next, we average across all observations that we make under these conditions. In doing so, we average under the constraints

$$X(t') = x' \text{ and } \langle\delta(x' - X(t'))\rangle = P(x', t'; u)\,. \quad (41)$$

In order to indicate that such a structured constraint should hold, we will use the notation

$$(\cdot)_{X(t')=x';\,\langle\delta(x'-X(t'))\rangle=P(x',t';u)}\,. \quad (42)$$

The conditional probability density $p(x, t \mid x', t', P(x', t'; u))$ is a special case in which the delta function is averaged under the constraint (41). We have

$$p(x, t \mid x', t', P(x', t'; u)) = \langle\delta(x - X(t))\rangle_{X(t')=x';\,\langle\delta(x'-X(t'))\rangle=P(x',t';u)}\,. \quad (43)$$

Summarizing the results we have derived so far, we see that strongly nonlinear Fokker–Planck equations define Markov processes whose

- time-dependent probability densities $P(x, t; u)$ are defined by Eq. (35)

- and conditional probability densities $p(x,t|x',t',P(x',t';u))$ are defined by

$$\frac{\partial}{\partial t}p(x,t\mid x',t',P') = L(x,t,P)\,p(x,t\mid x',t',P') \quad (44)$$

with L given by Eq. (36), $P = P(x,t;u)$, and $P' = P(x',t';u)$. Note that by multiplying Eq. (44) with $P(x',t';u)$ and integrating with respect to x' we get Eq. (35) which in turn defines the evolution of $P(x,t;u)$. Consequently, Eq. (44) defines both the evolution of $P(x,t;u)$ and $p(x,t\mid x',t',P')$. Note also that the solution of Eq. (44) formally reads

$$p(x,t\mid x',t',P(x',t';u)) = \exp\left\{\int_{t'}^{t}\mathrm{d}z\,L(x,z,P(x,z;u))\right\}\delta(x-x') \quad (45)$$

and depends on the evolution of $P(x,z;u)$ for $z\in[t',t]$. In fact, as indicated above p depends only on $P(x',t',u)$. To see this recall that the formal solution (21) can be obtained by means of the iterative method (19) such that we can write $P(x,z;u) = T_{z-t'}[P(x,t';u)]$. Substituting this solution into Eq. (45) we get

$$p(x,t\mid x',t',P(x',t';u)) = \exp\left\{\int_{t'}^{t}\mathrm{d}z\,L\big(x,z,T_{z-t'}[P(x,t';u)]\big)\right\}\delta(x-x')\,. \quad (46)$$

In addition, we find that the solution (45) does not explicitly depend on u.

We arrive at the following conclusion: conditional probability densities $p(x,t|\cdot)$ of Markov processes described by strongly nonlinear Fokker–Planck equations depend only on the value of individual realizations at one prior time $t'\le t$ and on the probability density P defined by all realizations at the very same prior time t'.

Equation (45) can be simplified for stationary Markov processes with operators L that do not depend explicitly on time t. Then the conditional probability density in the stationary case can be computed from

$$p(x,t\mid x',t',P_{\mathrm{st}}(x')) = \exp\left\{(t-t')\,L(x,P_{\mathrm{st}}(x))\right\}\delta(x-x')\,. \quad (47)$$

where $P_{\mathrm{st}}(x)$ denotes a stationary probability density out of a set of stationary probability densities defined by $LP_{\mathrm{st}} = 0$. Note that in this context P_{st} plays the role of an initial distribution u.

Just as in the linear case, the conditional probability density p in combination with the initial distribution u completely defines the stochastic process $\hat{X}$. In particular, the joint probability density $P(x_n,t_n;x_{n-1},t_{n-1};\ldots;x_0,t_0)$ can be computed from p and u like

$$P(\cdot) = p(x_n,t_n\mid x_{n-1},t_{n-1},P_{n-1}) \cdot p(x_{n-1},t_{n-1}\mid x_{n-2},t_{n-2},P_{n-2})\cdots \cdots p(x_1,t_1\mid x_0,t_0,u)u(x_0)\,, \quad (48)$$

with $P_n = P(x_n,t_n;u)$, $P_{n-1} = P(x_{n-1},t_{n-1};u)$, and so on.

In particular, the time-dependent probability densities $P(x,t;u)$ and $P(x,t';u)$ with $t\ge t'$ are related to each other by means of a *nonlinear* functional

$$P(x,t;u) = \int_{\Omega} p(x,t\mid x',t',P(x',t';u))\,P(x',t';u)\,\mathrm{d}x'\,, \quad (49)$$

where p is defined by Eq. (46). That is, the Green's function p induces a functional that is nonlinear with respect to $P(x',t';u)$.

Langevin Equations of Strongly Nonlinear Fokker–Planck Equations

The stochastic trajectories $X(t)$ of the Markov process $\hat{X}$ defined by Eq. (44) can be computed from two-layered Langevin equations (see Sect. 3.4 in [80]) or alternatively from the self-consistent Ito–Langevin equation

$$\frac{\mathrm{d}}{\mathrm{d}t}X(t) = D_1(X(t),t,P(X(t),t;u)) + \sqrt{D_2(X(t),t,P(X(t),t;u))}\,\Gamma(t)\,, \quad (50)$$

where $\Gamma(t)$ denotes the Langevin force introduced earlier. Note that the expression $P(X(t),t;u)$ means that the function $P(x,t;u)$ is evaluated at the state x that is given by the random variable X at time t. That is, we may write $P(X(t),t;u) = P(x,t;u)|_{x=X(t)}$. From the Langevin equation (50) we can read off that we are dealing with a Markov process. Information about one reference time t' in terms of the state $X(t')$ of a realization and the distribution of the ensemble as given by the probability density $P(x,t';u)$ is sufficient to compute the future behavior of the trajectory $X(t)$ with $t\ge t'$.

The Langevin equation (50) may be implemented on a computer using the iterative scheme

$$X_{n+1} = X_n + \Delta t\,D_1(X_n,t_n,P(X_n,t_n;u)) + \sqrt{\Delta t\,D_2(X_n,t_n,P(X_n,t_n;u))}\,\epsilon_n \quad (51)$$

with $X(t_n) = X_n$, $t_n = t_0 + n\Delta t$, $n = 0,1,2,\ldots$, and ϵ_n given as statistically independent Gaussian distributed random numbers with vanishing mean and variance 2 (see

above). The expression $P(X_n, t; u)$ can be computed from the realizations generated by the iteration scheme (51). Let $X_n^{(i)}$ denote the ith realization at time step n. Then, the stochastic trajectories $X(t)$ can numerically be computed by simulating an ensemble of realizations $i = 1, \ldots, N$ like

$$X_{n+1}^{(i)} = X_n^{(i)} + \Delta t\, D_1(X_n^{(i)}, t_n, P_n(X_n^{(i)})) + \sqrt{\Delta t\, D_2(X_n^{(i)}, t_n, P_n(X_n^{(i)}))}\, \epsilon_n^{(i)} , \quad (52)$$

where $\epsilon_n^{(i)}$ are statistically independent Gaussian random numbers with respect to both indices n and i and P_n is computed from the set $\{X_n^{(1)}, \ldots, X_n^{(N)}\}$ of realizations using standard kernel estimators. For example, we may use

$$P_n(x) = \frac{1}{Ns\sqrt{2\pi}} \sum_{i=1}^{N} \exp\left\{-\frac{(x - X_n^{(i)})^2}{2s^2}\right\} , \quad (53)$$

with $s = N^{-1/5}\sigma_e(t_n)$ where $\sigma_e(t_n)$ is the standard deviation of the empirical ensemble $\{X_n^{(1)}, \ldots, X_n^{(N)}\}$ [80,84, 204]. Just as in the case of Langevin equations of linear Fokker–Planck equations, the scheme (52) can be started at any time step n provided that we have information about P_n and X_n. In particular, if we start at a step $n > 0$ we see that the information about the initial distribution is irrelevant. Consequently, the sequence $X_n, X_{n+1}, X_{n+2}, \ldots$ computed from the time-discrete Langevin equation (52) related to the nonlinear Fokker–Planck equation (44) describes a trajectory of a Markov process.

Finally, note that self-consistent Langevin equations can be evaluated analytically in order to determine second-order statistical properties of a stochastic process defined by a strongly nonlinear Fokker–Planck equation [22,135].

Short-Time Propagator

The Green's function for short time intervals is frequently called the short-time propagator and can be derived from the time-discrete Ito-Langevin (51). Equation (51) relates the random variable ϵ_n that is distributed like $W(\epsilon_n)$, see Eq. (33), to the random variable X_{n+1}. In general, if X_{n+1} is a function of ϵ_n then the probability density $W'(x_{n+1})$ of X_{n+1} is given by

$$W'(x_{n+1}) = W(\epsilon_n)\frac{\mathrm{d}\epsilon_n}{\mathrm{d}x_{n+1}} . \quad (54)$$

In particular, if X_{n+1} is computed from ϵ_n for a particular value x_n and probability density P, then we obtain the short-time conditional probability density

$$p_s(x_{n+1} \mid x_n, P(x_n, t_n; u)) = W(\epsilon_n)\frac{\mathrm{d}\epsilon_n}{\mathrm{d}x_{n+1}} . \quad (55)$$

Equation (51) can be transformed into

$$\epsilon_n = \frac{X_{n+1} - X_n + \Delta t\, D_1(X_n, t_n, P(X_n, t_n; u))}{\sqrt{\Delta t\, D_2(X_n, t_n, P(X_n, t_n; u))}} . \quad (56)$$

Substituting Eq. (56) into Eq. (55) we obtain

$$p_s(x_{n+1} \mid x_n, P(x_n, t_n; u)) = \frac{\exp\left\{-\dfrac{[x_{n+1} - x_n + \Delta t\, D_1(x_n, t_n, P(x_n, t_n; u))]^2}{4\Delta t\, D_2(x_n, t_n, P(x_n, t_n; u))}\right\}}{\sqrt{4\pi\Delta t\, D_2(x_n, t_n, P(x_n, t_n; u))}} . \quad (57)$$

Using the time-continuous framework and the replacements $n \to t'$, $x_n \to x'$, $n+1 \to t = t' + \Delta t$, $x_{n+1} \to x$ and likewise $P(x_n, t_n; u) \to P(x', t'; u) = P'$, we obtain the short-time propagator (see Sect. 2.8.1 in [80])

$$p_s(x, t \mid x', t', P') = \frac{\exp\left\{-\dfrac{[x - x' + \Delta t\, D_1(x', t', P')]^2}{4\Delta t\, D_2(x', t', P')}\right\}}{\sqrt{4\pi\Delta t\, D_2(x', t', P')}} . \quad (58)$$

The short-time propagator has originally been proposed by Wehner and Wolfer [232] and can be used to solve nonlinear Fokker–Planck equations numerically [56,57,205]. To this end, the short-time propagator is substituted into Eq. (48) and subsequently Eq. (48) is integrated over all variables $x_{n-1}, \ldots, x_0$. Thus we obtain $P(x, t_n; u)$ for $t_n = n\Delta t$. In the context of linear Fokker–Planck equation the construction of solutions by means of short-time propagators is referred to as path integral approach [98,108, 184]. We will return to a similar path integral approach in Sect. "Semi-classical Description of Quantum Systems".

Expectation values of functions f can be computed from (58) like

$$\langle f(X(t))\rangle_{X(t')=x';\,\langle\delta(x'-X(t'))\rangle=P(x',t';u)} = \int_\omega f(x) p_s(x, t \mid x', t', P')\,\mathrm{d}x , \quad (59)$$

which holds for small intervals $\Delta t = t - t'$. The short-time propagator illustrates again the Markov property of solutions of the strongly nonlinear Fokker–Planck equation (44). The information about x' and P' at time t' is sufficient to make predications in terms of expectation values that the stochastic process will assume at time $t = t' + \Delta t$.

Chapman–Kolmogorov Equation, Kramers–Moyal Expansion and Drift-Diffusion Estimates

Linear Fokker–Planck equation can be derived using the Kramers–Moyal expansion of the Chapman–Kolmogorov equation [98,184]. The definition of the expansion coeffi-

cients in turn can be used to estimate the Kramers–Moyal coefficients in general and the drift- and diffusion-coefficients of linear Fokker–Planck equations in particular from experimental data [92,93]. We will show in this section that if a stochastic process defined by a nonlinear Fokker–Planck equation can be embedded into a Markov process using the concept of strongly nonlinear Fokker–Planck equations, then we can proceed as in the linear case. Taking a slightly different perspective, we may say that there are Markov processes that involve conditional probability densities of the form $p(x, t \mid x', t', P(x', t'; u))$ and can be characterized in terms of generalized Kramers–Moyal expansion coefficients.

Chapman–Kolmogorov Equation

Let $\hat{X}$ denote a stochastic Markov process with conditional probability density $p(x, t \mid x', t', P(x', t'; u))$. Then as discussed in the previous section the joint probability $P(x, t; x', t'; x'', t''; u)$ can be expressed by

$$P(x, t; x', t'; x'', t''; u) = p(x, t \mid x', t', P(x', t'; u)) \cdot p(x', t' \mid x'', t'', P(x'', t''; u))\, P(x'', t''; u)\,. \quad (60)$$

Integrating with respect to x' and dividing by $P(x'', t''; u)$ yields the generalized Chapman–Kolmogorov equation

$$p(x, t|x'', t'', P(x'', t''; u)) = \int_\Omega p(x, t|x', t', P(x', t'; u)) \cdot p(x', t' \mid x'', t'', P(x'', t''; u))\, \mathrm{d}x'\,. \quad (61)$$

Note that in what follows we will use the notation

$$\begin{aligned} P &= P(x, t; u)\,, \\ P' &= P(x', t'; u)\,, \\ P'' &= P(x'', t''; u)\,. \end{aligned} \quad (62)$$

If we need to express probability densities P different from those listed in Eq. (62), we will write down if necessary their arguments explicitly. For example, we will write $P(x, t'; u)$ to express the probability density $\langle \delta(x - X(t'))\rangle$ for a stochastic process $\hat{X}$ with initial distribution u.

Using the notation of Eq. (62), we can write the joint probability density (60) like

$$P(x, t; x', t'; x'', t''; u) = p(x, t \mid x', t', P')p(x', t' \mid x'', t'', P'')P'' \quad (63)$$

and the generalized Chapman–Kolmogorov equation (61) becomes

$$p(x, t \mid x'', t'', P'') = \int_\Omega p(x, t \mid x', t', P')p(x', t' \mid x'', t'', P'')\, \mathrm{d}x'\,. \quad (64)$$

Kramers–Moyal Expansion

In this section the Kramers–Moyal expansion for linear Fokker–Planck equations as discussed in [184] will be generalized to the nonlinear case. Consider the conditional probability density $p(x, t|x', t', P')$ for $t = t' + \Delta t$. Then, we have

$$p(x, t' + \Delta t \mid x', t', P') = \int_\Omega \delta(y - x)p(y, t' + \Delta t \mid x', t', P')\, \mathrm{d}y\,. \quad (65)$$

The variables x and x' denote arbitrary states in Ω. However, let us consider next states y that are close to x' such that $\epsilon = y - x'$ is small. Using $y - x = \epsilon + x' - x$, we obtain

$$p(x, t' + \Delta t \mid x', t', P') = \int_\Omega \delta(x' - x + \epsilon) \cdot p(x' + \epsilon, t' + \Delta t \mid x', t', P')\, \mathrm{d}\epsilon\,. \quad (66)$$

Use

$$\delta(x'-x+\epsilon) = \delta(x'-x)+\sum_1^\infty \frac{\epsilon^n}{n!}\left(\frac{\partial}{\partial x'}\right)^n \delta(x'-x). \quad (67)$$

Then, Eq. (66) becomes

$$p(x, t' + \Delta t|x', t', P') = \delta(x' - x) + \sum_1^\infty \int_\Omega \mathrm{d}\epsilon \frac{\epsilon^n}{n!} \cdot p(x'+\epsilon, t'+\Delta t|x', t', P')\left(\frac{\partial}{\partial x'}\right)^n \delta(x' - x)\,. \quad (68)$$

Multiplying Eq. (68) with $p(x', t'|x'', t'', P'')$ and integrating with respect to x' yields on the left hand side

$$\begin{aligned} \mathrm{LHS} &= \int_\Omega p(x, t' + \Delta t|x', t', P')p(x', t'|x'', t'', P'')\, \mathrm{d}x' \\ &= p(x, t' + \Delta t|x'', t'', P'') \end{aligned} \quad (69)$$

and on the right hand side

$$\begin{aligned} &\mathrm{RHS} \\ &= p(x, t'|x'', t'', P'') \\ &\quad + \sum_1^\infty \int_\Omega \mathrm{d}x' \int_\Omega \mathrm{d}\epsilon \frac{\epsilon^n}{n!} p(x' + \epsilon, t' + \Delta t|x', t', P') \\ &\quad \cdot p(x', t'|x'', t'', P'') \frac{\partial^n \delta(x'-x)}{\partial x'^n} \\ &= p(x, t'|x'', t'', P'') \\ &\quad + \sum_1^\infty \int_\Omega \mathrm{d}x' \delta(x' - x) \frac{\partial^n}{\partial x'^n}(-1)^n \\ &\quad \cdot \int_\Omega \mathrm{d}\epsilon \frac{\epsilon^n}{n!} p(x' + \epsilon, t' + \Delta t|x', t', P')p(x', t'|x'', t'', P'') \end{aligned}$$

$$= p(x, t'|x'', t'', P'')$$
$$+\sum_{1}^{\infty}\left(-\frac{\partial}{\partial x}\right)^n \int_\Omega \mathrm{d}\epsilon \frac{\epsilon^n}{n!} p(x+\epsilon, t'+\Delta t\,|\,x, t', P(x, t'; u))$$
$$\cdot\, p(x, t'|x'', t'', P'')\,. \tag{70}$$

Note that we used the Chapman–Kolmogorov equation (61) in order to evaluate the left hand side (69) and we used Eq. (61) as well as partial integration in order to evaluate the right hand side (70). Let us define the moments $M_n(x, t, \Delta t, P(x, t; u))$ by

$$M_n(x, t, \Delta t, P) = \int_\Omega \mathrm{d}\epsilon \frac{\epsilon^n}{n!} p(x+\epsilon, t+\Delta t|x, t, P) \tag{71}$$

or using $\epsilon + x = z$ by

$$M_n(x, t, \Delta t, P) = \int_\Omega \mathrm{d}z \frac{(z-x)^n}{n!} p(z, t+\Delta t|x, t, P)\,. \tag{72}$$

Combining the left and right hand sides given by Eqs. (69) and (70), respectively, we obtain

$$p(x, t'+\Delta t\,|\,x'', t'', P'') = p(x, t'\,|\,x'', t'', P'') + \sum_{1}^{\infty}\left(-\frac{\partial}{\partial x}\right)^n$$
$$\cdot\, M_n(x, t', \Delta t, P(x, t'; u))\, p(x, t'|x'', t'', P'')\,. \tag{73}$$

To improve readability let us replace t' by t and subsequently t'' by t'. Thus, we obtain

$$p(x, t+\Delta t\,|\,x', t', P') = p(x, t\,|\,x', t', P') + \sum_{1}^{\infty}\left(-\frac{\partial}{\partial x}\right)^n$$
$$\cdot\, M_n(x, t, \Delta t, P)\, p(x, t|x', t', P')\,. \tag{74}$$

This is the time-discrete version of the Kramers–Moyal expansion of the generalized Chapman–Kolmogorov equation (64). Note that M_n depends on $P(x, t; u)$, whereas p depends on $P(x', t'; u)$. Next, we define the Kramers–Moyal coefficients

$$D_n(x, t, P) = \lim_{\Delta t\to 0} \frac{M_n}{\Delta t} = \lim_{\Delta t\to 0} \frac{1}{\Delta t} \int_\Omega \mathrm{d}x \frac{(z-x)^n}{n!}$$
$$\cdot\, p(z, t+\Delta t\,|\,x, t, P)\,. \tag{75}$$

Dividing Eq. (74) by Δt and taking the limiting case $\Delta t \to 0$, Eq. (74) becomes the time-continuous generalized Kramers–Moyal expansion

$$\frac{\partial}{\partial t} p(x, t\,|\,x', t', P')$$
$$= \sum_{1}^{\infty}\left(-\frac{\partial}{\partial x}\right)^n D_n(x, t, P) p(x, t\,|\,x', t', P')\,. \tag{76}$$

Note that by generalizing the Kramers–Moyal expansion to the nonlinear case we found immediately that the coefficients D_n depend on $P(x, t; u)$, whereas the conditional probability density p depends on $P(x', t'; u)$. Note also that in the special case $D_n = 0$ for $n \geq 3$ the Kramers–Moyal expansion (76) yields the nonlinear Fokker–Planck equation (44). Note also that since we have $M_n(\Delta t = 0) = 0$ for all n Kramers–Moyal coefficients can also be defined by

$$D_n(x, t, P) = \left.\frac{\partial M_n}{\partial \Delta t}\right|_{\Delta t=0}$$
$$= \int_\Omega \mathrm{d}x \frac{(z-x)^n}{n!} \left.\frac{\partial}{\partial u} p(z, u|x, t, P)\right|_{u=t}\,. \tag{77}$$

Drift-Diffusion Estimates

The definition of the Kramers–Moyal coefficients can be exploited to extract the drift- and diffusion coefficients of nonlinear Fokker–Planck equations from time series data. Accordingly, the drift coefficient D_1 and the diffusion coefficient D_2 are defined by

$$D_1(x, t, P) = \lim_{\Delta t\to 0} \frac{1}{\Delta t} \int_\Omega \mathrm{d}x (z-x) p(z, t+\Delta t|x, t, P)\,,$$
$$D_2(x, t, P) = \lim_{\Delta t\to 0} \frac{1}{2\Delta t} \int_\Omega \mathrm{d}x \frac{(z-x)^2}{2} p(z, t+\Delta t|x, t, P)\,. \tag{78}$$

The limiting case Δt may be approximated by the smallest time step that is available in the data set:

$$D_1(x, t, P) \approx \frac{1}{\Delta t} \int_\Omega \mathrm{d}x (z-x)\, p(z, t+\Delta t\,|\,x, t, P)\,,$$
$$D_2(x, t, P) \approx \frac{1}{2\Delta t} \int_\Omega \mathrm{d}x \frac{(z-x)^2}{2} p(z, t+\Delta t|x, t, P)\,, \tag{79}$$

where Δt denotes now the sampling interval between two data points. Note that on the basis of the alternative definition (77) higher-order approximations can also be defined [171]. The conditional averages can be approximated by empirical conditional averages computed from a finite set of realizations $X^{(1)}(t), X^{(2)}(t), \ldots, X^{(N)}(t)$. Thus, we obtain

$$D_1(x, t, P) \approx \frac{1}{\Delta t} \frac{1}{\sum_{i\in I(t,x)} 1}$$
$$\cdot \sum_{i\in I(t,x)} [X^{(i)}(t+\Delta t) - X^{(i)}(t)]\,,$$
$$D_2(x, t, P) \approx \frac{1}{2\Delta t} \frac{1}{\sum_{i\in I(t,x)} 1}$$
$$\cdot \sum_{i\in I(t,x)} [X^{(i)}(t+\Delta t) - X^{(i)}(t)]^2\,, \tag{80}$$

where $I(t,x)$ is the set of indices i for which $X^{(i)}(t) \approx x$. In the case of Markov processes described by linear Fokker–Planck equations the argument P in the coefficients can be dropped and the above drift-diffusion estimates reduce to the estimates proposed in [92,93] that have recently found many applications [18,122,212,231]. For Markov processes described by nonlinear Fokker–Planck equations we need to compute the conditional averages for different probability densities P. To this end, we may vary the initial distribution u of a stochastic process. For a stochastic process with a particular distribution of X at time t we will obtain the coefficients D_1 and D_2 only for that particular distribution. Using the kernel estimate method mentioned above, we obtain

$$D_1\left(x,t,P \approx \frac{1}{Ns\sqrt{2\pi}}\sum_{i=1}^{N}\exp\left\{-\frac{(x-X^{(i)}(t))^2}{2s^2}\right\}\right) \approx \frac{1}{\Delta t}\frac{1}{\sum_{i\in I(t,x)}1}\sum_{i\in I(t,x)}[X^{(i)}(t+\Delta t)-X^{(i)}(t)] \quad (81)$$

and

$$D_2\left(x,t,P \approx \frac{1}{Ns\sqrt{2\pi}}\sum_{i=1}^{N}\exp\left\{-\frac{(x-X^{(i)}(t))^2}{2s^2}\right\}\right) \approx \frac{1}{2\Delta t}\frac{1}{\sum_{i\in I(t,x)}1}\sum_{i\in I(t,x)}[X^{(i)}(t+\Delta t)-X^{(i)}(t)]^2 \quad (82)$$

with $s = N^{-1/5}\sigma_e(t_n)$, where $\sigma_e(t_n)$ is the standard deviation of the empirical ensemble $\{X^{(1)}(t),\ldots,X^{(N)}(t)\}$. Note that in general the Kramers–Moyal coefficients of Markov processes induced by conditional probability densities of the form $p(x,t|x',t',P')$ can be estimated using

$$D_n\left(x,t,P \approx \frac{1}{Ns\sqrt{2\pi}}\sum_{i=1}^{N}\exp\left\{-\frac{(x-X^{(i)}(t))^2}{2s^2}\right\}\right) \approx \frac{1}{n!\Delta t}\frac{1}{\sum_{i\in I(t,x)}1}\sum_{i\in I(t,x)}[X^{(i)}(t+\Delta t)-X^{(i)}(t)]^n\,. \quad (83)$$

Alternatively, parametric estimate methods may be used. For example, we may be interested in estimating the exponent q of a Markov process defined by the Plastino–Plastino model (see Sect. "Nonextensive Systems" below)

$$\frac{\partial}{\partial t}p(x,t|x',t',P') = \left[\frac{\partial}{\partial x}\gamma x + Q\frac{\partial^2}{\partial x^2}P(x,t;u)^{q-1}\right]p(x,t|x',t',P') \quad (84)$$

with $\gamma, Q, q > 0$. Then, the diffusion coefficient $D_2(P) = QP^{q-1}$ involves the parameter Q and q. Using Eq. (82) and taking the logarithm, we get

$$\ln Q + (q-1)\ln\left\{\frac{1}{Ns\sqrt{2\pi}}\sum_{i=1}^{N}\exp\left\{-\frac{(x-X^{(i)}(t))^2}{2s^2}\right\}\right\} \approx \ln\left\{\frac{1}{2\Delta t}\frac{1}{\sum_{i\in I(t,x)}1}\sum_{i\in I(t,x)}[X^{(i)}(t+\Delta t)-X^{(i)}(t)]^2\right\}\,. \quad (85)$$

For example, at a particular time t Eq. (85) can be evaluated for different states x_i. In that case Eq. (85) assumes the form $\ln Q + (q-1)A_1(x_i) = A_2(x_i)$. Then, the expressions $\ln Q$ and $q-1$ (and in doing so the parameters Q and q) can be estimated from a linear regression [90].

Martingales

Let $Z(t)$ denote a functional of a stochastic process $\hat{X}$ defined for $t \geq t_0$. In this chapter we will put $t_0 = 0$. Then, Z is a martingale of $\hat{X}$ if

$$\langle Z(t)\rangle_{X=\theta} = Z(t') \quad (86)$$

holds for $t \geq t'$, where θ is a realization of the random variable X on the interval $[0,t']$ (see Sect. 1.3 in [131]). That is, the constraint $X = \theta$ means $X(s) = \theta(s)$ holds for $s \in [0,t']$. Roughly speaking, a martingale is a random variable for which the best predictor of its future mean value is the present value. With regard to Eq. (86) the prediction of the future mean value is $\langle Z(t)\rangle$, whereas the present value of Z is $Z(t')$. Alternatively, we may say that the information at one time t' about the value of the martingale Z is sufficient to predict the mean value of the martingale Z for future times $t \geq t'$. Note that this alternative point of view is closely related with the first definition of Markov processes discussed in the previous section.

For linear Fokker–Planck equations there is a close link between martingales and Markov property. Accordingly, a stochastic process is a Markov process defined by a linear Fokker–Planck equation with drift and diffusion coefficients D_1 and D_2 if and only if a particular random variable Z that involves the Fokker–Planck operator is a martingale (see Sect. 15.1 in [132]). In the mathematical literature this link has also be studied in the context of nonlinear Fokker–Planck equations [54,67,94,102,103,124,154,155,168].

Our aim in this section is to make the martingale approach more accessible to scientists working in physics, applied mathematicians, and related disciplines. To this end, we will in what follows illustrate this link between martingales and Markov processes defined by strongly nonlinear Fokker–Planck equation by means of standard techniques frequently used in physics.

Theorem 1 *Let $\hat{X}$ be a stochastic process with initial probability density $u(x)$ and conditional probability density*

$$p(x,t|x',t';P') = \langle \delta(x - X(t))\rangle_{X(t')=x';\langle\delta(x'-X(t'))\rangle=P(x',t';u)} \,. \quad (87)$$

Then, $\hat{X}$ is a Markov process defined by the nonlinear Fokker–Planck equation

$$\frac{\partial}{\partial t}p(x,t|x',t';P') = L(x,t,P)p(x,t|x',t';P') \quad (88)$$

with

$$L(x,t,P) = -\frac{\partial}{\partial x}D_1(x,t,P) + \frac{\partial^2}{\partial x^2}D_2(x,t,P) \quad (89)$$

if and only if $Z(t)$ defined by

$$Z(t) = f(X(t)) - \int_0^t L_B f[X(z),z,P]\,dz \quad (90)$$

with

$$L_B(x,t,P) = D_1(x,t,P)\frac{\partial}{\partial x} + D_2(x,t,P)\frac{\partial^2}{\partial x^2} \quad (91)$$

is a martingale of X for smooth functions f. In the context of linear Fokker–Planck equations the operator L_B is the Fokker–Planck backwards operator [98,184]. Note that in our context we refer to f as a smooth function if it has continuous second-order derivatives. Note also that above and in what follows we will frequently use the notation (54). Note finally that in the above theorem the notion $L_B f[X(z),z,P]$ should be interpreted like

$$L_B f[X(z),z,P] = L_B(X(z),z,P)f(X(z)) = \{L_B(x,t,P)f(x)\}_{x=X(z),t=z} \,. \quad (92)$$

That is, first we carry out the differentiations defined by the operator L_B. Subsequently, we replace in the result the state variable x by the value of the random variable X at time z. Moreover we replace t by z. Let us prove the theorem in two parts.

From Strongly Nonlinear Fokker–Planck Equations to Martingales

Let us prove in this section that a Markov process defined by a strongly nonlinear Fokker–Planck equation exhibits the martingale Z. To this end we first compute the conditional mean of the random variable Z defined in Eq. (90). Thus, we obtain

$$\begin{aligned}\langle Z(t)\rangle_{X=\theta} &= \langle f(X(t))\rangle_{X=\theta} - \int_0^t \mathrm{d}z\,\langle L_B f[X(z),z,P]\rangle_{X=\theta} \\ &= \langle f(X(t))\rangle_{X=\theta} - \int_{t'}^t \mathrm{d}z\,\langle L_B f[X(z),z,P]\rangle_{X=\theta} \\ &\quad - \int_0^{t'} \mathrm{d}s L_B f[\theta(s),s] \,. \end{aligned} \quad (93)$$

The Markov property implies that the constraints can be relaxed. That is, for every functional $g(t)$ of $X(t)$ with $t \geq t'$ we have $\langle g(t)\rangle_{X=\theta} = \langle g(t)\rangle_{X(t')=x';P'}$, where x' is given by $x' = \theta(t')$. We have indicated here that the average may depend on how the process is distributed at time t'. Consequently, Eq. (93) becomes

$$\begin{aligned}\langle Z(t)\rangle_{X=\theta} &= \langle f(X(t))\rangle_{X(t')=x';\langle\delta(x'-X(t'))\rangle=P'} \\ &- \int_{t'}^t \mathrm{d}z\,\langle L_B f[X(z),z,P]\rangle_{X(t')=x',\langle\delta(x'-X(t'))\rangle=P'} \\ &\quad - \int_0^{t'} \mathrm{d}s L_B f[\theta(s),s,P] \,. \end{aligned} \quad (94)$$

Multiplying the Fokker–Planck equation (88) with $f(x)$ and integrating with respect to x, we obtain

$$\frac{\partial}{\partial t}\int_\Omega f(x)p(x,t|x',t';P')\,\mathrm{d}x = \int_\Omega f(x)L(x,t,P)p(x,t|x',t';P')\,\mathrm{d}x \,. \quad (95)$$

By means of partial integration we find that $\int_\Omega f(x)\,L(x,t,P)\,p(x,t\,|\,x',t';P')\,\mathrm{d}x = \int_\Omega p(x,t\,|\,x',t';P')\,L_B(x,t,P)\,f(x)\,\mathrm{d}x$. As a result, Eq. (95) can be transformed into

$$\begin{aligned}\frac{\partial}{\partial t}\int_\Omega f(x)p(x,t|x',t',P')\,\mathrm{d}x &= \int_\Omega p(x,t|x',t',P')L_B(x,t,P)f(x)\,\mathrm{d}x \\ &= \langle L_B(x,t,P)f(x)\rangle_{X(t')=x';P'} \,. \end{aligned} \quad (96)$$

Using Eq. (96), we obtain

$$\langle L_B f[X(z),z]\rangle_{X(t')=x';\langle\delta(x-X(t'))\rangle=P'} = \frac{\partial}{\partial z}\int_\Omega f(x)p(x,z|x',t';P')\,\mathrm{d}x \,. \quad (97)$$

Consequently, the following integral transformation holds:

$$\begin{aligned} I &= \int_{t'}^t \mathrm{d}z\,\langle L_B f[X(z),z,P]\rangle_{X(t')=x';P'} \\ &= \int_{t'}^t \mathrm{d}z\frac{\partial}{\partial z}\int_\Omega \mathrm{d}x\,f(x)p(x,z|x',t',P') \\ &= \langle f(X(t)\rangle_{X(t')=x';P'} - f(x') \,. \end{aligned} \quad (98)$$

Substituting Eq. (98) into Eq. (94), we get

$$\langle Z(t)\rangle_{X=\theta} = f(x') - \int_0^{t'} \mathrm{d}s L_B f[\theta(s), s, P] \,. \tag{99}$$

By definition the function $Z(t')$ for $X(s) = \theta(s)$ given in $s \in [0, t']$ reads

$$Z(t') = f(x') - \int_0^{t'} \mathrm{d}s\, L_B f[\theta(s), s, P] \,. \tag{100}$$

Consequently, we have our final result

$$\langle Z(t)\rangle_{X=\theta} = Z(t') \tag{101}$$

and the proof is completed.

From Martingales to Strongly Nonlinear Fokker–Planck Equations

Let us prove next that the martingale (90) defines a Markov process of a strongly nonlinear Fokker–Planck equation. Evaluating Eq. (90) by analogy to Eq. (93) gives us

$$\langle Z(t)\rangle_{X=\theta} = \langle f(X(t))\rangle_{X=\theta} - \int_{t'}^{t} \mathrm{d}z\, \langle L_B f[X(z), z, P]\rangle_{X=\theta} - \int_0^{t'} \mathrm{d}s\, L_B f[\theta(s), s, P] \,. \tag{102}$$

Substituting this result into Eq. (86) and substituting Eq. (100) into Eq. (86), we see that Eq. (86) becomes

$$\langle f(X(t))\rangle_{X=\theta} = f(x') + \int_{t'}^{t} \mathrm{d}z\, \langle L_B f[X(z), z]\rangle_{X=\theta} \,. \tag{103}$$

Equation (103) can equivalently be written as

$$\int_\Omega f(x) p(x, t|X=\theta)\, \mathrm{d}x = f(x') + \int_{t'}^{t} \mathrm{d}z \int_\Omega \mathrm{d}x\, p(x, z|X=\theta) L_B f[x, z, P] \tag{104}$$

with $p(x, z|X=\theta) = \langle \delta(x - X(z))\rangle_{X=\theta}$. Using partial integration, we can show that the operator L_B and the differential operator L are related to each other like

$$\int_\Omega \mathrm{d}x\, p(x, z|X=\theta) L_B f[x, z, P] = \int_\Omega \mathrm{d}x\, f(x) L(x, z, P) p(x, z|X=\theta) \,. \tag{105}$$

Substituting this result into Eq. (104) yields

$$0 = \int_\Omega \mathrm{d}x\, f(x) \left\{ p(x, t|X=\theta) - \delta(x - x') - \int_{t'}^{t} \mathrm{d}z\, L(x, z, P) p(x, z|X=\theta) \right\} \,. \tag{106}$$

This holds for arbitrary smooth functions f. Since f is arbitrary, the expression in the brackets $\{\cdot\}$ of Eq. (106) must vanish and we obtain

$$p(x, t|X=\theta) = \delta(x - x') + \int_{t'}^{t} \mathrm{d}z\, L(x, z, P) p(x, z|X=\theta) \,. \tag{107}$$

Differentiating Eq. (107) with respect to t gives us

$$\frac{\partial}{\partial t} p(x, t|X=\theta) = L(x, t, P) p(x, t|X=\theta) \,. \tag{108}$$

Multiplying with the probability density $P(X=\theta)$ and performing a functional integration with respect to the path θ, we obtain

$$\frac{\partial}{\partial t} P(x, t) = L(x, t, P) P(x, t) \,. \tag{109}$$

The formal solutions of Eqs. (108) and (109) read

$$p(x, t|X=\theta) = \exp\left\{ \int_{t'}^{t} \mathrm{d}z\, L(x, z, P) \right\} \delta(x - x') \tag{110}$$

and

$$P(x, t) = \exp\left\{ \int_{t'}^{t} \mathrm{d}z\, L(x, z, P) \right\} P(x, t') \,. \tag{111}$$

We see that a solution of Eq. (108) under the initial condition $p(x, t|X=\theta) = \delta(x - x')$ for $t \to t'$ with $x' = \theta(t')$ only depends on $\theta(t')$ but does not depend on $\theta(s)$ for $s < t'$. Consequently, X is a Markov process. However, L depends on P. From Eq. (110) it is clear that the conditional probability density p depends on the time-dependent probability density P for $z \in [t', t]$. Since $P(x, t; u)$ for $t \geq t'$ can be computed from $P(x, t'; u)$ as shown in Eq. (111) we conclude that p depends only on $P(x, t'; u)$ and does not depend on the evolution of P on the whole interval $[t', t]$. Therefore, we have $p(x, t|X=\theta) = p(x, t|x', t', P')$. Substituting this result into Eq. (108), we see that Eq. (108) becomes a strongly nonlinear Fokker–Planck equation

$$\frac{\partial}{\partial t} p(x, t|x', t'; P') = L(x, t, P) p(x, t|x', t'; P') \,. \tag{112}$$

Examples

Shimizu–Yamada Model

The Shimizu–Yamada model [201,202] corresponds to the Desai–Zwanzig model (7) for a linear single-particle force $h(x) = -\gamma x$. The evolution of the conditional probability

density p is defined by

$$\frac{\partial}{\partial t}p(x,t\mid x',t',P') = \left[\frac{\partial}{\partial x}\gamma x + \kappa\left(x - \int_\Omega xP(x,t;u)\,\mathrm{d}x\right) + Q\frac{\partial^2}{\partial x^2}\right] \cdot p(x,t\mid x',t',P') \quad (113)$$

with $\Omega = \mathbb{R}$ and $\gamma, \kappa, Q > 0$. Multiplying Eq. (113) with $P(x',t';u)$ and integrating with respect to x' yields the evolution equation for $P(x,t;u)$:

$$\frac{\partial}{\partial t}P(x,t;u) = \left[\frac{\partial}{\partial x}\gamma x + \kappa\left(x - \int_\Omega xP(x,t;u)\,\mathrm{d}x\right) + Q\frac{\partial^2}{\partial x^2}\right] \cdot P(x,t;u)\,. \quad (114)$$

See also [77] and Sect. 3.10 in [80]. From Eq. (114) it follows that the mean value $m(t) = \int_\Omega xP(x,t;u)\,\mathrm{d}x$ decays exponentially like

$$m(t) = m(t_0)\exp\{-\gamma(t-t_0)\} \quad (115)$$

with $m(t_0) = \int_\Omega xu(x)\,\mathrm{d}x$. Substituting Eq. (115) into Eqs. (113) and (114), we realize that a solution $P(x,t;u)$ and a Green's function p exist for any initial probability density $u(x)$. Therefore, the Shimizu–Yamada model is a strongly nonlinear Fokker–Planck equation and describes a Markov process.

It can be shown that the conditional probability density $p(x,t\mid x',t';u)$ reads (see [77] and Sect. 3.10 in [80])

$$p(x,t\mid x',t';u) = \frac{\exp\left\{-\dfrac{[x-g(t,t',t_0,u)-x'm(t,t')]^2}{2K(t,t')}\right\}}{\sqrt{2\pi K(t,t')}} \quad (116)$$

with

$$m(t,t') = \exp\left\{-(\gamma+\kappa)(t-t')\right\}\,, \quad (117)$$

$$K(t,t') = \frac{Q}{\gamma+\kappa}\left[1+\exp\left\{-2(\gamma+\kappa)(t-t')\right\}\right]\,, \quad (118)$$

and

$$g = \left[\exp\{-\gamma(t-t_0)\} - \exp\{-(\gamma+\kappa)t + \gamma t_0 + \kappa t'\}\right] \cdot \int_\Omega xu(x)\,\mathrm{d}x\,. \quad (119)$$

The mean value $m(t)$ acts as a self-organized driving force or time-dependent order parameter of the stochastic process. Since there is a one-to-one mapping of $m(t)$ to $m(t')$ with $t' < t$, we can eliminate the parameter u in $p(x,t|x',t';u)$ as argued in Sect. "Strongly Nonlinear Fokker–Planck Equations". Substituting Eq. (115) into Eq. (119), we obtain

$$g(t,t',P(x,t';u)) = \exp\{-\gamma(t-t')\} \cdot \left[1-\exp\{-\kappa)(t-t')\}\right]\int_\Omega xP(x,t';u)\,\mathrm{d}x \quad (120)$$

or

$$g(t,t',\langle X(t')\rangle)\exp\{-\gamma(t-t')\} \cdot \left[1-\exp\{-\kappa)(t-t')\}\right]\langle X(t')\rangle\,. \quad (121)$$

Consequently, the conditional probability density $p(x,t\mid x',t',P')$ reads

$$p(x,t\mid x',t';P') = \frac{\exp\left\{-\dfrac{[x-g(t,t',\langle X(t')\rangle)-x'm(t,t')]^2}{2K(t,t')}\right\}}{\sqrt{2\pi K(t,t')}}\,. \quad (122)$$

Dynamic Takatsuji Model

The dynamic Takatsuji model for the conditional probability density p is defined by

$$\frac{\partial}{\partial t}p(x,t\mid x',t';P') = \left[\frac{\partial}{\partial x}(\gamma+c)x - \sqrt{c}\tanh\left(\sqrt{c}\int_\Omega xP(x,t;u)\,\mathrm{d}x\right) + Q\frac{\partial^2}{\partial x^2}\right] \cdot p(x,t\mid x',t';P') \quad (123)$$

with $x \in \Omega = \mathbb{R}$ and $c, Q > 0$, $\gamma \in \mathbb{R}$. Likewise, the probability density $P(x,t;u)$ satisfies

$$\frac{\partial}{\partial t}P(x,t;u) = \left[\frac{\partial}{\partial x}(\gamma+c)x - \sqrt{c}\tanh\left(\sqrt{c}\int_\Omega xP(x,t;u)\,\mathrm{d}x\right) + Q\frac{\partial^2}{\partial x^2}\right] \cdot P(x,t;u)\,. \quad (124)$$

For details see [78,214]. From Eq. (124) it follows that the first moment $M_1(t) = \langle X\rangle$ can be computed from

$$\frac{\mathrm{d}}{\mathrm{d}t}M_1(t) = -(\gamma+c)M_1 + \sqrt{c}\tanh[\sqrt{c}M_1(t)]\,. \quad (125)$$

For arbitrary initial distributions u solutions of $M_1(t)$ exist and are smooth functions of t. Substituting these solutions into Eqs. (123) and (124), we see that solutions of Eqs. (123) and (124) in terms of Green's functions p and probability densities P exist as well. Consequently, the dynamic Takatsuji model belongs to the class of strongly nonlinear Fokker–Planck equations and describes a Markov process.

Since $p(x,t|x',t',P')$ depends on P', the expected mean value of $X(t)$ of realizations that assume the value x' at time t' depends on the distribution of the ensemble at time t'. Let us illustrate this issue. The conditional mean value under consideration reads

$$\langle X(t)\rangle_{X(t')=x',P'} = \int x p(x,t \mid x',t',P')\,\mathrm{d}x\,. \quad (126)$$

Multiplying Eq. (123) with x and integrating with respect to x, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle X(t)\rangle_{X(t')=x',P'} = -(\gamma+c)\langle X(t)\rangle_{X(t')=x',P'} + \sqrt{c}\tanh[\sqrt{c}M_1(t)]\,. \quad (127)$$

The solution reads

$$\langle X(t)\rangle_{X(t')=x',P'} = x'\exp\{-(\gamma+c)(t-t')\} + \sqrt{c}\int_{t'}^{t}\tanh[\sqrt{c}M_1(z)]\,\mathrm{d}z\,, \quad (128)$$

where $M_1(z)$ is the solution of Eq. (125) for initial value $M_1(t') = \int_\Omega xP(x,t';u)\,\mathrm{d}x$. Let I denote the integral $I = \sqrt{c}\int_{t'}^{t}\tanh[\sqrt{c}M_1(z)]\,\mathrm{d}z$. Then, I depends on $M_1(t')$, c, γ, t' and t: $I = I(t,t',M_1(t'),c,\gamma)$. Consequently, Eq. (128) can be cast into the form

$$\langle X(t)\rangle_{X(t')=x',P'} = x'\exp\{-(\gamma+c)(t-t')\} + I(t,t',\int x'P(x',t';u)\,\mathrm{d}x',c,\gamma) \quad (129)$$

or

$$\langle X(t)\rangle_{X(t')=x',P'} = x'\exp\{-(\gamma+c)(t-t')\} + I(t,t',\langle X(t')\rangle,c,\gamma)\,. \quad (130)$$

Equation (130) illustrates that in order to predict future conditional mean values of a Takatsuji process $\hat{X}$ at times t it is sufficient to have at one time $t' \le t$ information about the state value x' of a realization of $\hat{X}$ and the mean value $\langle X(t')\rangle$ of all realizations of $\hat{X}$.

Note that the trajectories $Z(t)$ of martingale processes $\hat{Z}$ induced by the Takatsuji process $\hat{X}$ are given by

$$Z(t) = f(X) - \int_0^t \mathrm{d}s\left\{\left[-(\gamma+c)X(s) + \sqrt{c}\tanh(\sqrt{c}\langle X(s)\rangle\right]\frac{\partial f}{\partial X(s)} + Q\frac{\partial^2 f}{\partial X^2(s)}\right\} \quad (131)$$

for arbitrary smooth functions f. We can exploit these martingale processes in order to compute conditional expectations. For example, for $f(y) = y$ from the martingale property (86) it follows that

$$\langle X(t)\rangle_{X(t')=x',P'} = x' - \int_{t'}^{t}\mathrm{d}s\,(\gamma+c)\,\langle X(s)\rangle - \sqrt{c}\tanh[\sqrt{c}\,\langle X(s)\rangle]\,. \quad (132)$$

Differentiating this relation with respect to t we obtain Eq. (127) again and so we can compute the conditional expectation (130).

Liquid Crystal Model

Liquid crystals exhibit nematic-isotrope phase transitions [30,48,50]. At high temperatures the liquid crystal macromolecules exhibit an orientational disorder. The liquid crystal is said to be in the isotropic phase. Below a critical temperature the macromolecules show some degree of orientational order. The degree of orientational order is often measured by the Maier–Saupe order parameter S [150].

A nonlinear Fokker–Planck equation that describes the stochastic behavior of the liquid crystal in the isotropic and nematic phases and can to a certain extent also describe the phase transition was proposed by Doi and Edwards [55] and Hess [115] and is shown above in Eq. (8). Equation (8) describes the random walk of a orientation of liquid crystal molecules, where the orientation is given by a vector $\mathbf{x}$ that points to the surface of a unit sphere. For liquid crystals with an axial symmetry the liquid crystal model can be simplified. The simplified model describes the random walk of the molecule alignment with the symmetry axis. The random variable is defined on $X \in \Omega = [0,1]$. For sake of simplicity we will extend the range of definition to the interval $\Omega = [-1,1]$ and require that distributions are symmetric. For $X = 0$ the molecule has an orientation perpendicular to the symmetry axis. If $X = 1$ or $X = -1$ the molecule points exactly in the direction of the symmetry axis. In this symmetric case, the probability density P of X satisfies [64]

$$\frac{\partial}{\partial t}P(x,t;u) = \frac{\partial}{\partial x}(1-x^2) \cdot\left[-\frac{9}{2}\kappa x\left(\int x^2P(x,t;u)\,\mathrm{d}x - \frac{1}{3}\right) + D_r\frac{\partial}{\partial x}\right]P(x,t;u) \quad (133)$$

with $\kappa, D_r > 0$. Equation (133) as well as the original equation (8) are regarded as description for an ensemble of macromolecules that perform rotational Brownian motion [55]. Since Brownian motion is a Markov process it is reasonable to construct on the basis of Eq. (133) a model for a many-body system that exhibits a Markov process. In line with our discussion in Sect. "Markov Property, 2nd

Order and Higher Order Statistics" we assume that the conditional probability density p satisfies [81]

$$\frac{\partial}{\partial t}p(x,t|x',t',P') = \frac{\partial}{\partial x}(1-x^2) \cdot\left[-\frac{9}{2}\kappa x\left(\int x^2 P(x,t;u)\,\mathrm{d}x - \frac{1}{3}\right) + D_r\frac{\partial}{\partial x}\right] \cdot p(x,t|x',t',P')\,. \quad (134)$$

Note that the expression in the bracket $(\cdot)$ is related to the Maier–Saupe order parameter which reads in the symmetric case

$$S(t) = \frac{1}{2}\left(3\int x^2 P(x,t;u)\,\mathrm{d}x - 1\right)\,. \quad (135)$$

Due to the boundary conditions $X \in [-1,1]$ the order parameter S and consequently the bracket $(\cdot)$ is bounded. This implies that solutions P and p of Eqs. (133) and (134) exist and that the liquid crystal model (133)–(134) describes a Markov process. The self-consistent Ito–Langevin equation of this Markov process reads [81]

$$\frac{\mathrm{d}}{\mathrm{d}t}X(t) = \frac{9\kappa}{2}\left(1-X(t)^2\right)X(t)\left(\left\langle X(t)^2\right\rangle - \frac{1}{3}\right) - 2D_r X(t) + \sqrt{D_r(1-X(t)^2)}\,\Gamma(t)\,. \quad (136)$$

Trajectories $Z(t)$ of martingale processes $\hat{Z}$ of the liquid crystal model are defined by [83]

$$Z(t) = f(X(t)) - \int_0^t \mathrm{d}s\left[\,\cdot\,\right]f(X(s)) \quad (137)$$

with

$$\left[\,\cdot\,\right] = \left[\left(\frac{9\kappa}{2}(1-X(s)^2)X(s)\left(\left\langle[X(s)]^2\right\rangle - \frac{1}{3}\right) - 2D_r X(s)\right)\frac{\partial}{\partial X(s)} + D_r\frac{\partial^2}{\partial X^2(s)}\right]. \quad (138)$$

In the stationary case the short-time autocorrelation function $C(\Delta t) = \langle X(t)X(t+\Delta t)\rangle_{\mathrm{st}}$ reads [81]

$$C(\Delta t) = \frac{2S+1}{3} - \frac{2D_r(1-S)}{3}\Delta t + O(\Delta t^2)\,, \quad (139)$$

where S denotes the order parameter (see above) in the stationary case. That is, we have $S = (3\left\langle X^2\right\rangle_{\mathrm{st}} - 1)/2$. Consequently, C depends on S. This has important implications for the hysteresis loop of the nematic-isotropic phase transition. Let us assume that if we decrease the temperature of a liquid crystal we find the transition from the isotrope to the nematic phase with $S = 0 \to S > 0$ at the critical temperature $T_{c,\mathrm{low}}$. In contrast, if we increase the temperature of a liquid crystal we find the transition from the nematic to the isotropic phase with $S > 0 \to S = 0$ at the the slightly higher critical temperature $T_{c,\mathrm{high}}$. Then, in the temperature interval $[T_{c,\mathrm{low}}, T_{c,\mathrm{high}}]$ the liquid crystal exhibit two autocorrelation functions

$$C_{\mathrm{isotrope}}(\Delta t) = \frac{1}{3} - \frac{2D_r}{3}\Delta t\,, \quad (140)$$

$$C_{\mathrm{nem}}(\Delta t) = \frac{2S(T)+1}{3} - \frac{2D_r\,(1-S(T))}{3}\Delta t \quad (141)$$

which hold up to terms of order Δt^2. Equations (140)–(141) illustrate that we are dealing with a system that exhibits two kinds of Markov processes that we may label or name 'isotropic' and 'nematic'. The modeling approach by means of strongly nonlinear Fokker–Planck equations indicates that these Markov processes are just different members of a family of Markov processes that naturally emerge in the self-organized liquid crystal. That is, the two Markov processes are not related to two different systems but they are just two different faces of the very same self-organizing many-body system.

Let us compute the conditional mean value of molecules that are perpendicular to the symmetry axis. To this end, we consider the random walk of the orientation angle ϕ defined by $X(t) = \sin\phi(t)$. Using the Stratonovich–Langevin equation of Eq. (134) (see [81]), we obtain a self-consistent Langevin equation for ϕ:

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi = \frac{9}{4}\kappa\sin(2\phi(t))\left(\left\langle\sin^2(\phi(t))\right\rangle - \frac{1}{3}\right) - D_r\tan\phi(t) + D_r\Gamma(t)\,. \quad (142)$$

For short time intervals $\Delta t = t - t'$ and appropriate small noise amplitudes D_r we assume that $\phi(t) \approx 0$ if $\phi(t') \approx 0$. Linearizing Eq. (142) at $\phi = 0$ yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi(t) = \left\{\frac{9}{4}\kappa\left(\left\langle\sin^2(\phi(t))\right\rangle - \frac{1}{3}\right) - D_r\right\}\phi(t) + D_r\Gamma(t)\,. \quad (143)$$

The conditional expectation value $\int \phi\, p(\phi,t|\phi',t',P)\mathrm{d}\phi$ for short time intervals Δt can then be computed from Eq. (143) simply by averaging both side of Eq. (143) under the constraint $\phi(t') = \phi'$ and $\phi(t')$ distributed like P'. Thus, we obtain

$$\langle\phi(t)\rangle_{\phi(t')=\phi',P'} = \phi'\left[1 + \Delta t\left\{\frac{9}{4}\kappa\left(\left\langle\sin^2(\phi(t'))\right\rangle - \frac{1}{3}\right) - D_r\right\}\right] \quad (144)$$

or

$$\langle\phi(t)\rangle_{\phi(t')=\phi',P'} = \phi'\left[1+\Delta t\left\{\frac{3}{2}\kappa\, S(t') - D_r\right\}\right] . \tag{145}$$

These estimates hold for small intervals Δt, sufficiently small noise amplitudes D_r, and orientation angles $\phi' \approx 0$. Again, in line with our general discussion in the preceding sections we see that the conditional expectation $\langle\phi(t)\rangle_{\phi(t')=\phi',P'}$ can be computed provided that for $t' < t$ the distribution of $\phi(t')$ or at least the order parameter $S(t')$ is known and the angle ϕ' is selected.

Semi-classical Description of Quantum Systems

A stochastic treatment of semi-classical quantum systems by means of nonlinear Fokker–Planck equations that can be cast into the form of Eqs. (5) and (6) has been proposed and analyzed in several studies [28,31,85,125,127,129,130]. Accordingly, a Fermi or Bose particle with mass 1 that moves in a one-dimensional space with velocity v exhibits in the stationary case a Fermi–Dirac or Bose–Einstein distribution of the kinetical energy $E_{\text{kin}} = v^2/2$. The free diffusion of the particle can be described by the nonlinear Fokker–Planck equations [85]

$$\frac{\partial}{\partial t}P(v,t;u) = \frac{\partial}{\partial v}\gamma v\left[1 \mp P(v,t,u)\right]P(v,t;u) + Q\frac{\partial^2}{\partial v^2}P(v,t;u) , \tag{146}$$

where the upper sign holds for Fermi particles, the lower for Bose particles. The parameters γ and Q represent damping and fluctuation strength and are related to the temperature T by the fluctuation dissipation theorem $Q/\gamma = 1/(k_B T)$, where k_B is the Boltzmann constant. The stationary probability density $P_{\text{st}}(v)$ of Eq. (146) reads

$$P_{\text{st}}(v) = \frac{1}{\exp\{(E_{\text{kin}} - \mu)/(k_B T)\} \pm 1} , \tag{147}$$

where μ is a normalization constant that can be interpreted as chemical potential. The transient solution $P(v, t; u)$ can be obtained by solving the integral equation [83, 155]

$$P(v,t;u) = \int_\Omega \mathrm{d}v_0\, G_B(t,t_0,v,v_0)u(v_0,t_0) + \int_{t_0}^t \mathrm{d}s \cdot \int_\Omega \mathrm{d}v'\, G_B(t,s,v,v')\gamma\frac{\partial}{\partial v'}\left[1 \mp P(v',s;u)\right]P(v',s;u) \tag{148}$$

with

$$G_B(t,t',v,v') = \frac{1}{\sqrt{2\pi Q(t-t')}}\exp\left\{-\frac{(v-v')^2}{2Q(t-t')}\right\} , \tag{149}$$

where G_B is the Gaussian propagator of Brownian motion. In the limit $t \to \infty$ the transient solution $P(v, t; u)$ approaches $P_{\text{st}}(v)$ [88,127]. Equation (148) is a useful description for numerical approaches. Using partial integration Eq. (148) can be written in the form

$$P(v,t;u) = \int_\Omega \mathrm{d}v_0\, G_B(t,t_0,v,v_0)u(v_0,t_0) + \int_{t_0}^t \mathrm{d}s \int_\Omega \mathrm{d}v'\, G_B(t,s,v,v')\gamma\frac{v-v'}{Q(t-s)} \cdot \left[1 \mp P(v',s;u)\right] P(v',s;u) . \tag{150}$$

This integral relation can be solved iteratively. In contrast to the iterative procedure discussed in Sect. "Time–Dependent Solutions and First Order Statistics", there is no need to compute derivatives. That is, we are dealing with some kind of path integral approach here that is similar to the numerical path integral approach involving short-time propagators, see Sect. "Short-Time Propagator".

In order to describe quantum particles that exhibits a Markov process we may exploit the approach outlined in Sect. "Markov Property, 2nd Order and Higher Order Statistics". Accordingly, the Markov conditional probability density of the quantum particle satisfies

$$\frac{\partial}{\partial t}p(v,t \mid v',t',P') = \left\{\frac{\partial}{\partial v}\gamma v\left[1 \mp P(v,t;u)\right] + Q\frac{\partial^2}{\partial v^2}\right\} \cdot p(v,t \mid v',t',P') \tag{151}$$

and the self-consistent Langevin equation reads

$$\frac{\mathrm{d}}{\mathrm{d}t}v(t) = -\gamma v(t)(1 \mp P(v(t),t;u)) + \sqrt{Q}\,\Gamma(t) . \tag{152}$$

From a martingale perspective we see that stochastic trajectories $v(t)$ induce for arbitrary smooth functions f the martingale $\hat{Z}$ with trajectories

$$Z(t) = f(v(t)) - \int_0^t \mathrm{d}s \cdot \left[-\gamma v(s)\left[1 \mp P(v(s),s;u)\right]\frac{\partial}{\partial v} + Q\frac{\partial^2}{\partial v^2}\right] f(v)\Big|_{v=v(s)} . \tag{153}$$

The Markov short-time propagator for $t = t' + \Delta t$ reads

$$p(v,t|v',t',P') = \sqrt{\frac{1}{2\pi Q\Delta t}}\exp\left\{-\frac{\left[v - v' + \Delta t\gamma v'[1 \mp P']\right]^2}{2Q\Delta t}\right\} , \tag{154}$$

and can be computed from the information about the distribution P' of $v(t')$ and the state v that was observed for particular realizations of the process $\hat{v}$. In the stationary case p reads for small time intervals $\Delta t = t - t'$:

$$p(v, t \mid v', t'; P_{\mathrm{st}}(v')) = \sqrt{\frac{1}{2\pi Q \Delta t}} \exp\left\{ -\frac{\left[v - v' + \Delta t \gamma v'(1 \mp P_{\mathrm{st}}(v'))\right]^2}{2Q\Delta t} \right\} \tag{155}$$

with P_{st} defined by Eq. (147).

Nonextensive Systems

Nonextensive thermostatistical systems have been related to the Tsallis entropy [1,219]

$$S_q = \frac{1}{q-1} \int_{\Omega} \left[P(v)^q - P(v)\right] \mathrm{d}v \,, \tag{156}$$

where q measures the degree of nonextensivity. Diffusion processes in nonextensive thermostatistical systems can be regarded as generalized Ornstein–Uhlenbeck processes that satisfy the nonlinear Fokker–Planck equation [177] (see also [22,31,32,35,58,85,86,200,222])

$$\frac{\partial}{\partial t} P(v, t; u) = \frac{\partial}{\partial v} \gamma v P(x, t; u) + Q \frac{\partial^2}{\partial v^2} P(v, t; u)^q \,, \tag{157}$$

where v is the velocity of a particle with mass 1 that moves in one spatial dimension. In the asymptotic domain $P(v, t; u)$ approaches a stationary Tsallis distribution

$$P_{\mathrm{st}}(v) = \frac{D_{\mathrm{st}}}{\left[1 + \gamma(1-q)v^2/[2qQ\, D_{\mathrm{st}}^{q-1}]\right]^{1/(1-q)}} \tag{158}$$

for $q \in (1/3, 1)$ with $D_{\mathrm{st}} = [\gamma/(2qQz_q^2)]^{1/(1+q)}$ and $z_q = \sqrt{\pi/(1-q)}\Gamma[(1+q)/[2(1-q)]]/\Gamma[1/(1-q)]$ [86]. The process described by Eq. (157) is said to evolve in a nonextensive thermodynamic framework because its stationary probability density (158) maximizing the entropy measure (156). As discussed in Sect. "Markov Property, 2nd Order and Higher Order Statistics" the Markov conditional probability density p satisfies

$$\frac{\partial}{\partial t} p(v, t \mid v', t', P') = \frac{\partial}{\partial x} \gamma v\, p(v, t \mid v', t', P') + Q \frac{\partial^2}{\partial v^2} P(v, t; u)^{q-1}\, p(v, t \mid v', t', P') \,. \tag{159}$$

The self-consistent Langevin equation of the Markov diffusion process reads (see also [22])

$$\frac{\mathrm{d}}{\mathrm{d}t} v(t) = -\gamma v(t) + \sqrt{QP(v(t), t; u)^{q-1}}\, \Gamma(t) \,. \tag{160}$$

Any stochastic path $v(t)$ computed from Eq. (160) yields for arbitrary smooth functions f the martingale

$$Z(t) = f(v(t)) - \int_0^t \mathrm{d}s \left[-\gamma v(s) \frac{\partial}{\partial v(s)} + QP(v(s), s; u)^{q-1} \frac{\partial^2}{\partial v(s)^2} \right] f(v(s)) \,. \tag{161}$$

The autocorrelation $C = \langle v(t) v(t') \rangle$ in the transient domain for $u(v_0) = \delta(v - v_0)$ reads [74]

$$C(t, t', v_0, t_0) = M_2(t', t_0, v_0) \exp\{-\gamma(t - t')\} \tag{162}$$

with

$$\begin{aligned} M_2(t', t_0, v_0) &= K(t', t_0) + M_1^2(t', t_0, v_0) \,, \\ K(t', t_0) &= \frac{1}{3q-1} \left[\frac{2qQ[z_q]^{(1-q)}}{\gamma} \cdot (1 - \exp\{-(1+q)\gamma(t' - t_0)\}) \right]^{2/(1+q)} \,, \\ M_1(t', t_0, v_0) &= v_0 \exp\{-\gamma(t' - t_0)\} \,. \end{aligned} \tag{163}$$

The autocorrelation function C depends on t_0. This is not in contradiction with the Markov property of the underlying process as discussed in Sect. "Markov Property, 2nd Order and Higher Order Statistics". In particular, we may eliminating the initial condition. Then, Eq. (162) simply reads

$$C(t, t') = \left\langle v^2(t') \right\rangle \exp\{-\gamma(t - t')\} \,. \tag{164}$$

and holds for arbitrary initial probability densities u.

Linear Nonequilibrium Thermodynamics

Linear and nonlinear Fokker–Planck equations alike can be approached from the principles of linear nonequilibrium thermodynamics [49,100,140]. For stochastic processes to which linear nonequilibrium thermodynamics applies the probability density $P(x, t; u)$ of a process evolves such that the free energy functional $F[P]$ decreases as a function of time t. More precisely, following a study by Compte and Jou [35] it has been proposed that P satisfies the nonlinear Fokker–Planck equations of the form [32,70,80,190]

$$\frac{\partial}{\partial t} P = \frac{\partial}{\partial x} P \tilde{M} \frac{\partial}{\partial x} \frac{\delta F}{\delta P} \,, \tag{165}$$

where $\tilde{M}$ is an appropriately defined mobility coefficient and $\delta F/\delta P$ denotes the variational derivative of F. Note that this thermodynamic approach is closely related to the GENERIC approach developed in [63,123,164,165,166].

For example, the Desai–Zwanzig model (8) can be expressed in terms of Eq. (165) for $\tilde{M} = 1$ and [80,196]

$$F = \langle V \rangle + U_{\mathrm{MF}} - QS_{\mathrm{BGS}} \,. \tag{166}$$

Here, V is the potential of the force h (i. e. we have $V(x) = -\int h(x)\,\mathrm{d}x$), U_{MF} is the mean field energy given by $U_{\mathrm{MF}} = -\kappa\sigma^2/2$ (where σ^2 is the variance of the process) and S_{BGS} is the Boltzmann–Gibbs-Shannon entropy

$$S_{\mathrm{BGS}} = -\int_{\Omega} P(x,t;u) \ln P(x,t;u)\,\mathrm{d}x \,, \tag{167}$$

where we have put the Boltzmann constant equal to unity. The liquid crystal model (8) can be written as Eq. (165) with [81]

$$F = -\frac{\kappa}{2}S^2 - D_r S_{\mathrm{BGS}} \,, \tag{168}$$

where S is the Maier–Saupe order parameter (135) and $\tilde{M} = 1 - x^2$ and $-\kappa S^2/2$ is the Maier–Saupe mean field energy. The Kuramoto–Shinomoto–Sakaguchi model (10) can equivalently be expressed in term of Eq. (165) and $\tilde{M} = 1$ and see Sect. 5.4 in [80]

$$F = \langle V \rangle - \frac{\kappa}{2}r^2 - QS_{\mathrm{BGS}} \,, \tag{169}$$

where r is the cluster phase defined by $r = |\langle \exp\{-iX(t)\}\rangle|$ and $-\kappa r^2/2$ is a measure for the mean field energy among the phase oscillators described by the model. The Takatsuji model (124) involves a constant mobility coefficient $\tilde{M} = 1$ and the free energy functional [80]

$$F = \frac{\gamma + c}{2}\left\langle X^2 \right\rangle - \ln\cosh\left(\sqrt{c}\,\langle X \rangle\right) - QS_{\mathrm{BGS}} \,. \tag{170}$$

The Plastino–Plastino model (14) related to the nonextensive Tsallis entropy (156) is given by Eq. (165) and $\tilde{M} = 1$ with [80]

$$F = \langle V \rangle - QS_q \,, \tag{171}$$

where V is the potential of the gradient force h. For an appropriate choice of $\tilde{M}$ the quantum mechanical nonlinear Fokker–Planck equations (146) can be cast into the form Eq. (165) with

$$F = \langle V \rangle - QS_{\mathrm{FD,BE}} \,, \tag{172}$$

where $S_{\mathrm{FD,BE}}$ is the quantum mechanical entropy of the Fermi–Dirac or Bose–Einstein statistics and V is the potential of the function $h(x)$ again. For details see [80,85].

From the perspective of linear nonequilibrium thermodynamics linear and nonlinear Fokker–Planck equations can be distinguished by means of the thermodynamic flux [35,70,80,83]

$$J = -\tilde{M}P\frac{\partial}{\partial x}\frac{\delta F}{\delta P} \,. \tag{173}$$

Note that in this approach the thermodynamic flux is equivalent to the probability current [80]. As can be seen from Eq. (172) the flux is associated to the free energy F on the one hand and the evolution equation (165) on the other hand by

$$\frac{\partial}{\partial t}P = -\frac{\partial}{\partial x}J \,. \tag{174}$$

If J is linear with respect to P, then the corresponding Fokker–Planck equation is linear with respect to P as well. If J is nonlinear with respect to P, then we are dealing with a nonlinear Fokker–Planck equation. The question whether J is linear or not is answered by nature herself [83]. For the Brownian particle motion we have $\tilde{M} = \gamma$ and

$$F = \frac{1}{2}\left\langle v^2 \right\rangle - QS_{\mathrm{BGS}} \,, \tag{175}$$

which yields

$$J = -\gamma v P - Q\frac{\partial}{\partial x}P \,. \tag{176}$$

J is linear and the corresponding Fokker–Planck equation is linear as well. For self-organizing systems J is nonlinear because F involves a mean field energy term that is nonlinear with respect to P see, for example, Eqs. (166), (168), (169), and (170). Likewise, for quantum and nonextensive systems we find that J is nonlinear because F involves quantum and nonextensive entropies such as $S_{\mathrm{FD,BE}}$ and S_q, see, for example, Eqs. (171) and (172).

Summary and Future Directions

From the previous discussion in Sect. "Linear Nonequilibrium Thermodynamics", it is clear that modeling approaches based on nonlinear Fokker–Planck equations are rooted in the theory of collective phenomena and self-organization, on the one hand, and in the theory of quantum mechanical and nonextensive systems, on the other. In contrast, linear Fokker–Planck equations are tailored to address the stochastic properties of systems composed of non-interacting subsystems when equating the material subsystem ensemble with the ensemble of statistical realizations. Note that – of course – linear Fokker–Planck equations can also be applied to discuss stochastic properties of self-organizing systems. However, in such cases either the stochastic behavior of order parameters by means of low dimensional linear Fokker–Planck equations is

discussed [108] or linear high-dimensional or even functional Fokker–Planck equations are involved [98].

We showed in Sect. "Markov Property, 2nd Order and Higher Order Statistics" that both linear and nonlinear Fokker–Planck equations exhibit Green's functions and Langevin equations. The fact that a nonlinear evolution equation can give rise to a Green's function may be counter-intuitive because Green's functions are associated with linearity. In fact, the evolution equation of the Green's function p is linear with respect to p. The nonlinearity is in the evolution equation for the time-dependent probability density P but not in the evolution equation of the Green's function p. In this context, we would like to re-iterate what we pointed out in Sect. "Strongly Nonlinear Fokker–Planck Equations": time-dependent solutions P do not necessarily correspond to Green's function p [73].

In the mathematical literature the theory of Markov processes that involve conditional probability densities of the form $p(x, t|x', t', P')$ has been discussed for decades (see references in Sect. "Definition of the Subject"). In line with these studies we suggest to refer to Markov processes with conditional probability densities of the form $p(x, t|x', t', P')$ as *nonlinear* Markov processes or *nonlinear* families of Markov processes [77]. Likewise, we suggest to refer to Markov processes whose conditional probability densities do not depend on P' as *linear* Markov processes or *linear* families of Markov processes. Using this terminology we would say that strongly nonlinear Fokker–Planck equations describe nonlinear Markov diffusion processes and vice versa nonlinear Markov diffusion processes can be expressed in terms of strongly nonlinear Fokker–Planck equations.

In physics and related disciplines the relevance of nonlinear Markov processes has to be explored in the future. That is, except for mathematics the theory of Markov process constructed from conditional probability measures of the form $p(x, t|x', t', P')$ is still in its infancy. The Chapman–Kolmogorov equation and the Kramers–Moyal expansion presented in Sect. "Markov Property, 2nd Order and Higher Order Statistics" provide promising departure points for future studies in this regard.

In the present study we pointed out that there are a few 'umbrella' concepts that apply to linear and nonlinear Fokker–Planck equations alike: the concepts of Markov diffusion processes, martingales, and linear nonequilibrium thermodynamics. Therefore, future studies that follow the ideas of the present study could change the state of the art illustrated in Fig. 1 into a scenario as shown in Fig. 2. In doing so, a closely connected world of linear and nonlinear Fokker–Planck equations that is governed by a small set of powerful principles could emerge.

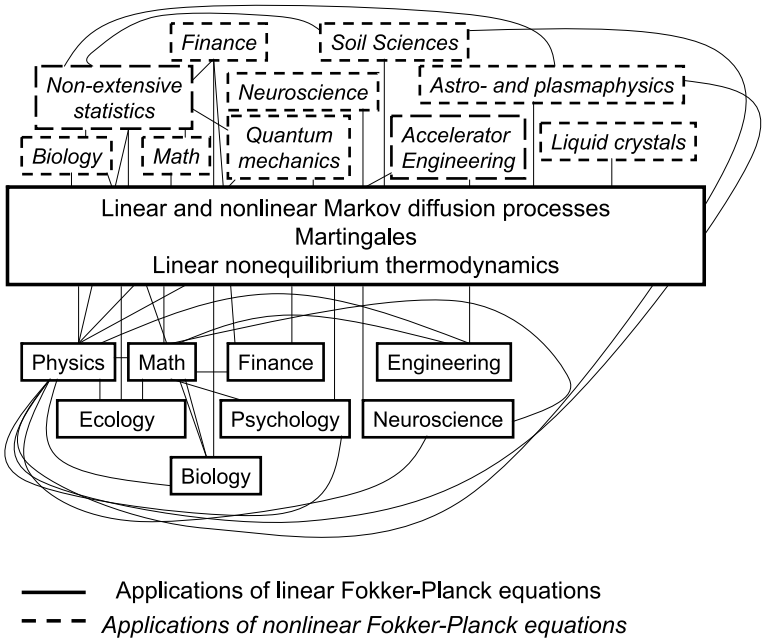

Linear and Non-linear Fokker–Planck Equations, Figure 2
Connected applications of linear and nonlinear Fokker–Planck equations

Bibliography

1. Abe S, Okamoto Y (2001) Nonextensive statistical mechanics and its applications. Springer, Berlin
2. Acebron JA, Bonilla LL, De Leo S, Spigler R (1998) Phys Rev E 57:5287
3. Acebron JA, Bonilla LL, Vicente CJP, Ritort F, Spigler R (2005) Rev Mod Phys 77:137
4. Acebron JA, Spigler R (1998) Phys Rev Lett 81:2229
5. Allen EJ, Victory HD Jr (1994) Physica A 209:318
6. Aonishi T, Okada M (2002) Phys Rev Lett 88:024102
7. Arenas A, Perez Vincente CJ (1994) Phys Rev E 50:949
8. Ariaratnam JT, Strogatz SH (2001) Phys Rev Lett 86:4278
9. Arnold A, Bonilla LL, Markowich P (1996) Transp Theor Stat Phys 25:733
10. Arnold A, Markowich P, Toscani G (2000) Transp Theor Stat Phys 29:571
11. Arnold A, Markowich P, Toscani G, Unterreiter A (2001) Commun Part Diff Eq 26:43
12. Aronson DG (1986) In: Dobb A, Eckmann B (eds) Nonlinear diffusion problems – Lecture notes in mathematics, vol 1224. Springer, Berlin, pp 1–46
13. Balescu R (1975) Equilibrium and nonequilibrium statistical mechanics. Wiley, New York
14. Barenblatt GI, Entov VM, Ryzhik VM (1990) Theory of fluid flows through natural rocks. Kluwer Academic Publisher, Dordrecht
15. Becker V, Engel A (2007) Phys Rev E 75:031118
16. Binney J, Tremaine S (1987) Galatic dynamics. Princeton University Press, Princeton, New Jersey
17. Birner T, Lippert K, Müller R, Kühnel A, Behn U (2002) Phys Rev E 65:046110
18. Bödeker HU, Röttger MC, Liehr A, Frank TD, Friedrich R, Purwins HG (2003) Phys Rev E 67:056220
19. Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD (2006) Psychol Rev 113:700
20. Bonilla LL (1987) J Stat Phys 46:659
21. Bonilla LL, Vicente CJP, Rubi JM (1993) Stat J Phys 70:921
22. Borland L (1998) Phys Rev E 57:6634
23. Borland L (2002) Phys Rev Lett 89:098701
24. Borland L (2008) Anomalous fluctuations in complex systems: plasmas, fluids and financial markets. In: Riccardi C, Roman HE (eds) Special Review Book for Research Singpost, Transworld Research Network, KeralaMadison, 2008, in press
25. Borland L, Bouchaud JP (2004) Quantitative finance 4:499
26. Bychenkov VY, Rozmus W, Tikhonchuk VT (1995) Phys Rev Lett 75:4405
27. Carillo JA, Jüngel A, Markowich PA, Toscani G, Unterreiter A (2001) Monatshefte für Mathematik 113:1
28. Carrillo JA, Rosado J, Salvarani F (2008) Appl Math Letters 21:148
29. Cepa E, Lepingle D (1997) Probab Theory Relat Fields 107:429
30. Chandrasekhar S (1977) Liquid crystals. Cambridge University Press, Cambridge
31. Chavanis PH (2003) Phys Rev E 68:036108
32. Chavanis PH (2004) Physica A 340:57
33. Chavanis PH, Rosier C, Sire C (2002) Phys Rev E 66:036105
34. Coffey WT, Kalmykov YP, Waldron JT (2004) The Langevin equation. World Scientific, New Jersey
35. Compte A, Jou D (1996) Phys J Math A Gen 29:4321
36. Cortines AAG, Riera R (2007) Physica A 377:181
37. Crank J (1975) The mathematics of diffusion. Clarendon Press, Oxford
38. Crawford JD (1995) Phys Rev Lett 74:4341
39. Curado EMF, Nobre FD (2003) Phys Rev E 67:021107
40. Daido H (1996) Physica D 91:24
41. Daido H (1996) Phys Rev Lett 77:1406
42. Daido H (2001) Phys Rev Lett 87:048101
43. Daly E, Porporato A (2004) Phys Rev E 70:056303
44. Dano S, Hynne F, De Monte S, d'Ovidio F, Sorensen PG, Westerhoff H (2001) Faraday Discuss 120:261
45. Dawson DA (1983) J Stat Phys 31:29
46. Dawson DA (1993) Lecture notes in mathematics 1541. In: Dawson DA, Maisonneuve B, Spencer J (eds) Springer, Berlin, pp 6–16
47. Dawson DA, Gärtner J (1989) Large deviations, free energy functional and quasi-potential for a mean field model of interacting diffusions (American Mathematical Society, Providence)
48. de Gennes P, Prost J (1993) The physics of liquid crystals. Clarendon Press, Oxford
49. de Groot SR, Mazur P (1962) Non-equilibrium thermodynamics. North-Holland, Amsterdam
50. de Jeu WH (1980) Physical properties of liquid crystalline materials. Gordon and Beach, New York
51. de los Santos F, Telo da Gamma MM, Munoz MA (2003) Phys Rev E 67:021607
52. Desai RC, Zwanzig R (1978) Stat J Phys 19:1
53. Ding X (1994) Adv Appl Prob 26:1022
54. Djehiche B, Kaj I (1995) Ann Probab 23:1414
55. Doi M, Edwards SF (1988) The theory of polymer dynamics. Clarendon Press, Oxford
56. Donoso JM, Salgado JJ (2006) J Phys A 39:12587
57. Donoso JM, Salgado JJ, Soler M (2005) J Phys A 38:9145
58. Drazer G, Wio HS, Tsallis C (2000) Phys Rev E 61:1417
59. Drozdov AN, Morillo M (1996) Phys Rev E 54:931
60. Ebeling W, Sokolov IM (2004) Statistical thermodynamics and stochastic theory of nonequilibrium systems. World Scientific, Singapore
61. Eckhardt B, Ott E, Strogatz SH, Abrams DM, McRobie A (2007) Phys Rev E 75:021110
62. Epperlein EM, Rickard GJ, Bell AR (1988) Phys Rev Lett 61:2453
63. Espanol P, Serrano M, Öttinger HC (1999) Phys Rev Lett 83:4542
64. Felderhof BU (2003) Physica A 323:88
65. Fialkowiski M, Hess S (2000) Physica A 282:65
66. Fokker AD (1914) Annalen der Physik 43:810
67. Fontbona J (2003) J Funct Analysis 200:198
68. Frank TD (2001) Phys Lett A 280:91
69. Frank TD (2001) Phys Lett A 290:93
70. Frank TD (2002) Phys Lett A 305:150
71. Frank TD (2002) Physica A 310:397
72. Frank TD (2003) Phys Lett A 319:173
73. Frank TD (2003) Physica A 320:204
74. Frank TD (2004) Eur Phys J B 37:139
75. Frank TD (2004) J Phys A 37:3561
76. Frank TD (2004) Phys Lett A 329:475
77. Frank TD (2004) Physica A 331:391
78. Frank TD (2004) Physica D 195:229
79. Frank TD (2005) Math Comput Mod 42:1057
80. Frank TD (2005) Nonlinear Fokker–Planck equations: Fundamentals and applications. Springer, Berlin

81. Frank TD (2005) Phys Rev E 72:041703
82. Frank TD (2006) Phys Rev ST-AB 9:084401
83. Frank TD (2007) Physica A 382:453
84. Frank TD (2008) Physica A 387:773
85. Frank TD, Daffertshofer A (1999) Physica A 272:497
86. Frank TD, Daffertshofer A (2000) Physica A 285:351
87. Frank TD, Daffertshofer A (2001) Physica A 292:392
88. Frank TD, Daffertshofer A (2001) Physica A 295:455
89. Frank TD, Daffertshofer A, Peper CE, Beek PJ, Haken H (2001) Physica D 150:219
90. Frank TD, Friedrich R (2005) Physica A 347:65
91. Friedman A (1969) Partial differential equations. Holt, Rinehart and Winston, Inc., New York
92. Friedrich R, Peinke J (1997) Phys Rev Lett 78:863
93. Friedrich R, Siegert S, Peinke J, Lück S, Seifert M, Lindemann M, Raethjen J, Deuschl G, Pfister G (2000) Phys Lett A 271:217
94. Gärtner J (1988) Math Nachr 137:197
95. Gang H, Haken H, Fagen X (1996) Phys Rev Lett 77:1925
96. Garcia-Ojalvo J, Parrondo JMR, Sancho JM, van den Broeck C (1996) Phys Rev E 54:6918
97. Garcia-Ojalvo J, Sancho JM (1999) Noise in spatially extended systems. Springer, New York
98. Gardiner CW (1997) Handbook of stochastic methods, 2nd edn. Springer, Berlin
99. Giada L, Marsili M (2000) Phys Rev E 62:6015
100. Glansdorff P, Prigogine I (1971) Thermodynamic theory of structure, stability, and fluctuations. Wiley, New York
101. Goel NS, Dyn NR (1974) Stochastic models in biology. Academic Press, New York
102. Graham C (1990) Appl Math Optim 22:75
103. Greven A (2005) Interacting stochastic systems. In: Deuschel J, Greven A (eds) Springer, Berlin, pp 209–246
104. Gurtin ME, MacCamy RC (1977) Mathematical Biosciences 33:35
105. Hadley P, Beasley MR, Wiesenfeld K (1988) Phys Rev B 38:8712
106. Hänggi P, Marchesoni F, Nori F (2005) Annalen der Physik 14:51
107. Hänggi P, Talkner P, Borkovec M (1990) Rev Mod Phys 62:251
108. Haken H (2004) Synergetics: Introduction and advanced topics. Springer, Berlin
109. Han SK, Kurrer C, Kuramoto Y (1995) Phys Rev Lett 75:3190
110. Hansel D, Mato G, Meunier C (1993) Europhysics Letters 23:367
111. Hansel D, Mato G, Meunier C (1993) Phys Rev E 48:3470
112. Hasegawa H (2003) Phys Rev E 67:041903
113. Heifets S (2001) Phys Rev ST-AB 4:044401
114. Heifets S (2003) Phys Rev ST-AB 6:080701
115. Hess S (1976) Naturforschung Z A 31:1034
116. Holden AV (1976) Models of the stochastic activity of neurons. In: Lecture notes in biomathematics, vol 12. Springer, Berlin
117. Horsthemke W, Lefever R (1984) Noise-induced transitions. Springer, Berlin
118. Hütter M, Karlin IV, Öttinger HC (2003) Phys Rev E 68:016115
119. Ichiki A, Ito H, Shiino M (2007) Physica E 40:402
120. Ilg P, Karlin IV, Öttinger HC (1999) Phys Rev E 60:5783
121. Ilg P, Kröger M, Hess S (2005) Phys Rev E 71:051201
122. Jafari GR, Fazeli SM, Ghasemi F, Allaei SMV, Tabar MRR, Zad AI, Kavei G (2002) Phys Rev Lett 91:226101
123. Jelic A, Hütter M, Öttinger HC (2006) Phys Rev E 74:041126
124. Jourdain B (2000) Methodology and computing in applied probability 2:69
125. Kadanoff LP (2000) Statistical physics: statics, dynamics and renormalization. World Scientific, Singapore
126. Kanamaru T, Horita T, Okabe Y (2001) Phys Rev E 64:031908
127. Kaniadakis G (2001) Phys Lett A 288:283
128. Kaniadakis G (2001) Physica A 296:405
129. Kaniadakis G, Quarati P (1993) Phys Rev E 48:4263
130. Kaniadakis G, Quarati P (1994) Phys Rev E 49:5103
131. Karlin S, Taylor HM (1975) A first course in stochastic processes. Academic Press, New York
132. Karlin S, Taylor HM (1981) A second course in stochastic processes. Academic Press, New York
133. Kawamura Y (2007) Phys Rev E 76:047201
134. Kawamura Y, Nakao H, Kuramoto Y (2007) Phys Rev E 75:036209
135. Kharchenko DO, Kharchenko VO (2005) Physica A 354:262
136. Kim S, Park SH, Ryu CS (1997) Phys Rev Lett 78:1616
137. Klimontovich YL (1986) Statistical physics. Harwood Academic Publ, New York
138. Kloeden PE, Platen E (1992) The numerical solution of stochastic differential equations. Springer, Berlin
139. Kometani K, Shimizu H (1975) J Stat Phys 13:473
140. Kondepudi D, Prigogine I (1998) Modern thermodynamics. Wiley, New York
141. Kostur M, Luczka J, Schimansky-Geier L (2002) Phys Rev E 65:051115
142. Kozyreff G, Vladimirov AG, Mandel P (2000) Phys Rev Lett 85:3809
143. Kuramoto Y (1984) Chemical oscillations, waves, and turbulence. Springer, Berlin
144. Lancellotti C, Kiessling M (2001) Astrophys J 549:L93
145. Larson RG, Öttinger HC (1991) Macromolecules 24:6270
146. Li JH, Huang ZQ, Xing DY (1998) Phys Rev E 58:2838
147. Li JH, Hänggi P (2001) Phys Rev E 64:011106
148. Lo CF (2005) Phys Lett A 336:141
149. MacDonald WM, Rosenbluth MN, Chuck W (1957) Phys Rev 107:350
150. Maier W, Saupe A (1958) Z Naturforschung A 13:564
151. Marsili M, Bray AJ (1996) Phys Rev Lett 76:2750
152. McCauley JL, Gunaratne GH, Bassler KE (2006) Physica A 382:445
153. McKean HP Jr (1969) Propagation of Chaos for a class of nonlinear parabolic equations. In: Aziz AK (ed) Lectures in differential equations, vol II. Van Nostrand Reinhold Company, New York, pp 177–193
154. Meleard S (1996) Asymptotic behavior of some interacting particle systems: McKean-Vlasov and Boltzmann models. In: Graham C, Kurtz TG, Meleard S, Potter PE, Pulvirenti M, Talay D (eds) Probabilistic models for nonlinear partial differential equations. Springer, Berlin, pp 42–95
155. Meleard S, Coppoletta SR (1987) Stoch Process Appl 26:317
156. Michael F, Johnson MD (2003) Physica A 320:525
157. Mikhailov AS, Zanette DH (1999) Phys Rev E 60:4571
158. Morillo M, Gomez-Ordonez J, Casado JM (1995) Phys Rev E 52:316
159. Müller R, Lippert K, Kühnel A, Behn U (1997) Phys Rev E 56:2658
160. Nicholson DR (1983) Introduction to plasma theory. Wiley, New York
161. Nobre FD, Curado EMF, Rowlands G (2004) Physica A 334:109
162. Oelschläger K (1989) Probab Th Rel Fields 82:565

163. Öttinger HC (1996) Stochastic processes in polymeric fluids. Springer, Berlin
164. Öttinger HC (2005) Beyond equilibrium thermodynamics. Wiley, New Jersey
165. Öttinger HC (2007) Phys Rev Lett 99:130602
166. Öttinger HC, Grmela M (1997) Phys Rev E 56:6633
167. Okubo A, Levin SA (2001) Diffusion and ecological problems: modern perspectives. Springer, Berlin
168. Overbeck L (1996) Ann Probab 24(2):743
169. Park K, Lai YC, Liu Z, Nachman A (2004) Phys Lett A 326:391
170. Park SH, Kim S (1996) Phys Rev E 53:3425
171. Patanarapeelert K, Frank TD, Friedrich R, Beek PJ, Tang IM (2006) Phys Lett A 360:190
172. Paul W, Baschnagel J (1999) Stochastic processes: from physics to finance. Springer, Berlin
173. Peletier LA (1981) In: Amann H, Bazley N, Kirchgässner K (eds)(1981) Applications of nonlinear analysis in the physical science. Pitman Advanced Publishing Program, Boston, pp 229–241
174. Pikovsky A, Rosenblum M, Kurths J (2001) Synchronization: a universal concept in nonlinear sciences. Cambridge University Press, Cambridge
175. Pilipenko AY (2005) Urkainian Math J 57:1507
176. Planck M (1917) Sitzungsber. Preuss Akad Wissens, Berlin, p 324
177. Plastino AR, Plastino A (1995) Physica A 222:347
178. Quinn DA, Rand RH, Strogatz SH (2007) Phys Rev E 75:036218
179. Ratcliff R, Gomez P, McKoon G (2004) Psychol Rev 111:159
180. Reif F (1965) Fundamentals of statistical and thermal physics. McGraw-Hill Book Company, New York
181. Reimann P (2002) Phys Rep 361:57
182. Reimann P, Kawai R, Van den Broeck C, Hänggi P (1999) Europhysics Letters 45:545
183. Reimann P, Van den Broeck C, Kawai R (1999) Phys Rev E 60:6402
184. Risken H (1989) The Fokker–Planck equation – Methods of solution and applications. Springer, Berlin
185. Rogers LCG, Shi Z (1993) Probab Theroy Relat Fields 95:555
186. Rosenblum MG, Pikovsky AS (2004) Phys Rev Lett 92:114102
187. Rosenbluth M, MacDonald WM, Judd DL (1957) Phys Rev 107:1
188. Sakaguchi H (1988) Prog Theor Phys 79:39
189. Savel'ev S, Marchesoni F, Nori F (2003) Phys Rev Lett 91:010601
190. Scarfone MA, Wada T (2007) Physica A 384:305
191. Schuster HG, Wagner P (1990) Biol Cybern 64:77
192. Schwämmle V, Nobre FD, Curado E (2007) Phys Rev E 76:041123
193. Schweitzer F (2003) Brownian agents and active particles. Springer, Berlin
194. Schwämmle V, Gonzahles MC, Moreira AA, Andrade JS, Hermann HJ (2007) Phys Rev E 75:066108
195. Shiino M (1985) Phys Lett A 112:302
196. Shiino M (1987) Phys Rev A 36:2393
197. Shiino M (2001) J Math Phys 42:2540
198. Shiino M (2002) J Math Phys 43:2654
199. Shiino M (2002) J Phys Soc 40:1037
200. Shiino M (2003) Phys Rev E 67:056118
201. Shimizu H (1974) Prog Theor Phys 52:329
202. Shimizu H, Yamada T (1972) Prog Theor Phys 47:350
203. Shobuda Y, Hirata K (2001) Phys Rev E 64:067501
204. Silverman BW (1986) Density estimation for statistics and data analysis. Chapman and Hall, London
205. Soler M, Martinez FC, Donoso JM (1992) Stat J Phys 69:813
206. Sopik J, Sire C, Chavanis PH (2006) Phys Rev E 74:011112
207. Strogatz SH (2000) Physica D 143:1
208. Strogatz SH, Abrams D, McRobie A, Eckhardt B, Ott E (2005) Nature 438:43
209. Strogatz SH, Mirollo RE (1991) Stat J Phys 63:613
210. Strogatz SH, Morillo RE, Matthews PC (1992) Phys Rev Lett 68:2730
211. Stupakov GV, Breizman BN, Pekker MS (1997) Phys Rev E 55:5976
212. Sura P, Barsugli J (2002) Phys Lett A 305:304
213. Takai M, Akiyama H, Takeda S (1981) J Phys Soc Japan 50:1716
214. Takatsuji M (1975) Biol Cybern 17:207
215. Tass PA (1999) Phase resetting in medicine and biology – Stochastic modelling and data analysis. Springer, Berlin
216. Tass PA (2001) Europhysics Letters 55:171
217. Tass PA (2003) Biol Cybern 89:81
218. Tass PA (2006) Biol Cybern 94:58
219. Tsallis C (1988) Stat J Phys 52:479
220. Tsallis C (1997) Phys World 10(7):42
221. Tsallis C (2004) Physica D 193:3
222. Tsallis C, Bukman DJ (1996) Phys Rev E 54:R2197
223. Tsekov R (1995) J Phys A Math Gen 28:L557
224. Tsekov R (2001) Int J Mol Sci 2:66
225. van den Broeck C, Parrondo JMR, Armero J, Hernandez-Machado A (1994) Phys Rev E 49:2639
226. van den Broeck C, Parrondo JMR, Toral R (1994) Phys Rev Lett 73:3395
227. van den Broeck C, Parrondo JMR, Toral R, Kawai R (1997) Phys Rev E 55:4084
228. van Kampen NG (1981) Stochastic processes in physics and chemistry. North-Holland, Amsterdam
229. Vellekoop M, Nieuwenhuis H (2007) Quantitative finance 7:563
230. Venturini M, Warnock R (2002) Phys Rev Lett 89:224802
231. Waechter M, Riess F, Schimmel T, Wendt U, Peinke J (2004) Eur Phys J B 41:259
232. Wehner MF, Wolfer WG (1987) Phys Rev A 35:1795
233. Wiesenfeld K, Colet P, Strogatz SH (1996) Phys Rev Lett 76:404
234. Winfree AT (1967) Theor J Biol 16:15
235. Winfree AT (2001) The geometry of biological time, 2nd edn. Springer, Berlin
236. Yamaguchi Y, Shimizu H (1984) Physica D 11:213
237. Yamana M, Shiino M, Yoshioka M (1999) Phys J A Math Gen 32:3525
238. Yoshioka M, Shiino M (2000) Phys Rev E 61:4732
239. Zaikin AA, Garcia-Ojalvo J, Schimansky-Geier L, Kurths J (2002) Phys Rev Lett 88:010601
240. Zhai Y, Kiss IZ, Daido H, Hudson JL (2005) Physica D 205:57

Link Analysis and Web Search

Johannes Bjelland,
Geoffrey Canright,
Kenth Engø-Monsen
Telenor R&I, Fornebu, Norway

Article Outline

Glossary
Definition of the Subject
Introduction
Structure of Directed Graphs
Approaches to Web Link Analysis
The 'Sink Problem'
Sensitivity of Link-Analysis Scores
Personalization of Link Analysis Scores
Peer-to-Peer Search and Link Analysis
Future Directions
Bibliography

Glossary

Graph A set of vertices connected by directed or undirected edges.

Web graph The World Wide Web represented as a graph. The webpages are vertices and hyperlinks are edges. The Web graph is called a *directed graph* since the hyperlinks are directed links from one page to another.

Adjacency matrix A matrix A which represents the structure of a graph. For the Web graph, with directed links, A is typically defined such that $A_{ij} = A_{i \to j} = 1$ if page i points to page j, and 0 otherwise.

Link analysis In Link Analysis the relations between nodes in a graph are studied. In large complex networks link analysis is a tool to analyze a node's relations to all other nodes in the network. Based on the link structure of the graph, several node-properties can be found, for instance: Is the node well connected to the rest of the network? Can node A be reached from node B? How many edges separate A from B? Is a node central in its network? Is it contained in a sub-network? With link analysis these and other node properties are found. In short, link analysis helps us to extract a node's role in a graph.

Web-page ranking Link analysis is used to measure the quality of a webpage. A link is interpreted as a recommendation. Link analysis scores are used by search engines when hits from queries (Web pages) are presented to the user in the form of a ranked list.

PageRank The most well-known Web-page ranking algorithm, developed at Stanford University by Larry Page and Sergey Brin. Page and Brin are the founders of Google, which employs the PageRank-algorithm.

Personalization of ranking-scores When the web-page link analysis score is calculated, some pages may be boosted according to criteria other than just the link-structure of the Web graph. For example, pages relevant to a certain topic (which is preferred by the user) might be boosted to receive a better ranking.

Link-spamming A method where a spammer tries to boost a webpage's rank in a search engine hit list. This is done by constructing several pages (a 'link spam farm') and manipulating the link-structure in such a way that a targeted page will get a boost in its link analysis score, and hence a higher ranking in a search engine hit list.

Crawling A *crawler* (or *spider*) is a program that maps the contents of the World Wide Web by starting at some initial webpage(s) and jump from page to page via hyperlinks in a systematic manner. All pages reachable from the startpage(s) by clicking hyperlinks are then visible to the crawler. From each page visited, hyperlinks and useful information is extracted. Search engines use webcrawlers to build their searchable databases (*index*). The part of the WWW which has been visited by a crawler and stored in a search-engine's database is called the *indexable Web* or the *visible Web*.

Node indegree (outdegree) The number of links pointing to (from) a node. The idea is readily generalized to a set of nodes such as an SCC (see below).

Strongly connected component (SCC) A maximal set of nodes in a directed graph, such that there is a directed path from any node in the SCC to any other.

Source (sink) SCC An SCC with zero indegree (outdegree).

Definition of the Subject

The World Wide Web consists of a vast amount of information in the form of web-pages. Almost all subjects are covered somewhere in the billions of pages reachable from your computer. To access this enormous library from our computers is a wonderful thing – but the tremendous size of the web is also a problem. With an ever-growing number of pages covering all sorts of subjects, it becomes very hard to find the one, relevant piece of information you need. Web link analysis is a tool which assists us with this task. It helps us to rank web documents, and so to find the most important documents among the huge amount of irrelevant information available online. The input to the web link analysis is a network of hyperlinked documents, where the nodes are documents, and the directed edges are hyperlinks from one page to another. The output is a vector containing a numerical score which represents the importance of each page in the Web graph. A typical use of this score is to assist ranking of a search engine query result. The most well-known example is the search engine

Google, which uses a web-link analysis algorithm developed by Brin and Page [68]. In this article we will present an overview of the principal existing methods used to calculate importance scores for Web pages, using the Web graph as input. We will discuss the theory behind the different approaches, compare them to each other, and discuss some selected practical and technical problems with Web link analysis.

Introduction

Since the early days of the World Wide Web (WWW) there has been an explosion in the number of hyperlinked documents available online. Lawrence and Giles [56] estimated the publicly indexable web to contain about 800 million pages in 1999. In 2005 the number was estimated to be over 11.5 billion pages [39]. The task of sorting and organizing all this information is an impossible task for humans – machines must somehow be taught to do this job. Web link analysis is just one of the ways that machines help us to find the needle in the enormous haystack that is the World Wide Web.

Consider a reader who has just gotten a machine to help her to find a document of interest. More precisely: the reader *hopes* that the machine has delivered one or more documents of interest. This is an extremely common situation in the Information Age; and it presents interesting challenges – some of which we will address in this article. The reader normally has two types of criteria in mind for a 'good' document. First, it should be *relevant* to the user's current focus or need – for example the purchase of new cars in Malaysia. Second, the document should be of high *quality*. Clearly, each of these criteria is important for the reader. Furthermore, each is difficult for a machine to evaluate; but the second one – quality – is perhaps the most difficult. After all, human readers can debate endlessly what constitutes or exemplifies high quality, and what does not; how then can a machine say something useful about this question?

Web link analysis (or simply 'link analysis') – which we often abbreviate simply as LA – is a tool which helps us to calculate the quality of web pages. Consider two web pages A and B. The basic idea behind LA is that if page A contains a hyperlink to page B, this is interpreted as a recommendation of page B by page A. This interpretation is plausible since many (but not all) hyperlinks are laid down by human writers – typically the author of the pointing page. By using a hyperlink, the author is sending the readers away to another page. If one assumes the author has read this page and found the contents interesting enough to hyperlink it, this would increase our subjective idea of the quality of the targeted page. Two of the most well-known LA algorithms that use this approach are PageRank [68] and HITS [52]. These and other algorithms will be discussed in Sect. "Approaches to Web Link Analysis".

We note briefly here that there are close connections between Web link analysis and the older field termed *bibliometrics* – in particular, a subfield called *citation analysis*. The latter involves the analysis of the "web of citations" for scholarly publications, with the aim of estimating the importance or impact of each article. Citations (like hyperlinks) are one-way links, so that the citation graph is a directed graph. However – unlike the case for hyperlinks – citations which point "forward in time" are very rare, so that the citation graph is more treelike than the Web graph. Garfield [37] defines the 'impact factor' for a publication in a way that basically counts the inlinks (citations) for a given period. Hence this approach is close to one in Web link analysis which we will call *'link popularity'* below. For a good discussion of more sophisticated, whole-graph approaches to citation analysis, see [52].

The WWW contains billions of pages connected with hyperlinks. A page's LA score can be regarded as a form of *collective recommendation* from all the other pages. High LA score means that the page has a high degree of collective recommendation. It is then perhaps more precise to replace the term 'quality' with 'importance' [19] or 'authority' [52] on the WWW.

A common way to retrieve information from the large and complex WWW is to submit a query to a search engine. Lawrence and Giles [56] found that search-related pages ranked among the top 10 most visited sites on the web. Also, from studying the most frequently used search words (queries), one can see that many users use search engines to replace bookmarks – typing in the name or (part or even all of) the url of a well known page. Thus, search engines are a normal part of life for many of us. A search engine contains a searchable database *(index)* of extracted information from pages found on the World Wide Web. The largest search engines have indexed billions of web-pages [39]. If a query is not too obscure, the user will be presented with a large list of hits more or less relevant to the query. For instance, a query 'car' returns around 3 billion hits from the Yahoo search engine at the time of writing this document [88]. The huge majority of the hits are useless for the person submitting the search. Some hits may even be spam – pages constructed with the purpose to get as many hits as possible. (Search engine spamming is discussed in Sect. "Sensitivity of Link-Analysis Scores".)

For a human, it is impossible to browse billions of

search engine hits to find the most relevant and high-quality pages. The importance score from link analysis is used in the process of ranking the search results presented to the user, according to their quality or importance on the web. This importance score is typically combined with a *text relevance score* which evaluates each hit with respect to its relevance to the query. Hits are then ranked according to the combined score, so that pages that are both important and relevant appear first on the hit list. An example of a successful implementation of LA in a search engine is the earlier mentioned PageRank used by Google. The PageRank of a web page is Google's way of measuring how important this page is, as compared to all other pages in the enormous Web.

In this article we present an overview of web search and link analysis. In link analysis it is useful to define the World Wide Web (WWW) as a network. The web-documents are nodes and the hyperlinks are edges. This network is then represented by a directed graph. Properties of general directed graphs, and also of the Web graph, are discussed in Sect. "Structure of Directed Graphs". A subsection of Sect. "Structure of Directed Graphs" is devoted to the 'Sink Problem' – a numerical problem that may arise when calculating scores for groups of pages which basically have many inlinks but no outlinks. In Sect. "Approaches to Web Link Analysis", existing approaches to web link analysis are discussed. Among these are the already mentioned and well-known methods PageRank and HITS, plus two other methods 'Link popularity' and 'SALSA' [58], and also a recent algorithm called 'T-Rank' [15,16]. In Sect. "The 'Sink Problem'", the 'sink problem' is revisited and discussed in relation to the methods described in Sect. "Approaches to Web Link Analysis". Another topic that has gotten considerable study is the sensitivity of the link analysis scores. How do network topology changes, sink remedies and spam affect the scores? This is discussed in Sect. "Sensitivity of Link-Analysis Scores". Then, in Sect. "Personalization of Link Analysis Scores" we discuss approaches to personalizing link analysis scores, and in Sect. "Peer-to-Peer Search and Link Analysis" we look at distributed approaches to link analysis, including the application to peer-to-peer search. Finally, we sum up with a look at some future directions of search and link analysis in Sect. "Future Directions".

Structure of Directed Graphs

In this section we briefly review the structure of directed graphs – which is different in important ways from that of undirected graphs. We also briefly present the structure of the Web graph: a directed, dynamic, largely self-organized graph, composed of Web documents connected via hyperlinks. The Web graph is the raw material which is used by Web link analysis algorithms to generate LA scores.

General

As earlier mentioned, the World Wide Web is a network of pages connected by hyperlinks. In mathematics and computer science such a network is called a graph. The webpages are nodes and the hyperlinks are edges. Since the hyperlinks have a direction, the webgraph is a directed graph. In this subsection a few useful basic graph-theoretic terms will be explained. These and several other properties of graphs are described in textbooks on graph theory and in Web graph articles, for instance [13,21], and [34].

We start with the simpler case – an undirected (or symmetric) graph. Undirected links are termed 'symmetric' because an undirected link $\mathbf{A} - \mathbf{B}$ implies a path from A to B *and* from B to A, and hence is equivalent to $\mathbf{A} \leftrightarrow \mathbf{B}$. The *degree* of a node in an undirected graph is the number of edges it has. We typically study *connected* graphs – i. e., graphs which do not decompose into several pieces, without removing one or more links. Connected symmetric graphs are very simple: for *any* two nodes i and j in the graph, there is a path from i to j. The *distance* from i to j is the number of edges we have to follow to reach j starting at i.

Directed graphs are built up of one-way links: $\mathbf{A} \rightarrow \mathbf{B}$. This statement does not exclude the possibility of reciprocal one-way links, i. e., that both $\mathbf{A} \rightarrow \mathbf{B}$ and $\mathbf{A} \leftarrow \mathbf{B}$ are present; however, neither is this guaranteed. The above definition of a connected graph may also be applied to directed graphs. This is equivalent to the more conventional definition – that a directed graph is connected if, when all links are made symmetric, there is a path from i to j for any two nodes i and j in the graph.

A node in a directed graph has both edges pointing out from it and edges pointing in. The number of edges pointing respectively out and in is called the node's *outdegree* and *indegree*. The directionality of the links gives directed graphs features not seen in symmetric (undirected) graphs. For example, in general, when we do *not* ignore the direction of the links, there is *not* necessarily a (directed) path from i to j, for any two nodes i and j in a connected directed graph. This is clear from the 'simplest possible' directed graph, the two-node graph $\mathbf{A} \rightarrow \mathbf{B}$: there is a path from A to B, but not from B to A.

This single (negative) statement distinguishes directed graphs from undirected graphs. Now we build from this property, in a positive way. One can find a set C of nodes in a directed graph, such that, for any two nodes i and j

in the set C, there is a directed path from i to j. This set is called *strongly connected*; and if we make this set as large as possible – without killing its property of being strongly connected – then we get a set (subgraph) which is called a *strongly connected component* or SCC. SCCs are very useful in thinking about directed graphs. Any directed graph may be uniquely decomposed into its strongly connected components – that is, every node belongs uniquely to one and only one SCC. Furthermore, given two SCCs C1 and C2, any links connecting the two are either (i) all in the same direction, e. g. **C1→C2**, or (ii) nonexistent. A two-way connection such as **C1↔C2** is impossible, since then the two SCCs collapse to one. For the same reason, there are no cyclic connections of the form **C1→C2→C3→…→C1**, since the entire cycle comprises in fact a single SCC.

Thus, movement along directed paths in a directed graph takes two forms: 'circulation' from any node to any other *within* an SCC; and one-way or 'downhill' flow *between* SCCs. This overall one-way flow in directed graphs, which is found at the coarse (inter-SCC) scale (see Fig. 1), is quite distinct from the purely circulatory nature of connected undirected graphs; and it has interesting consequences. For example: unless the directed graph is composed of a single SCC, there must always be one or more SCCs which have no inlinks from other SCCs; these are called *source* SCCs. Similarly, there must be one or more 'ends' to the downhill flow – SCCs without outlinks, termed *sink* SCCs.

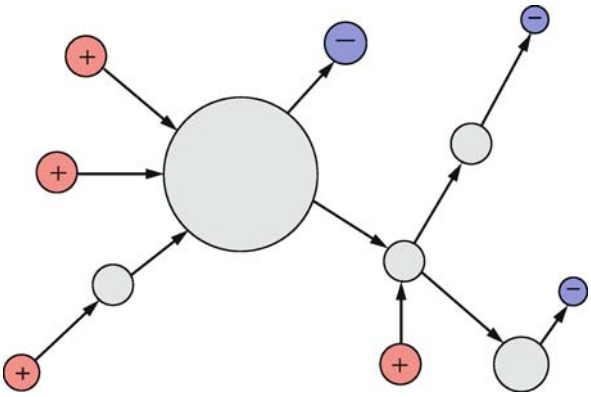

Link Analysis and Web Search, Figure 1
Generic structure of a directed graph. Each strongly connected component or SCC (see text) is indicated with a *circle* here. The (one-way) direction of links between SCCs is indicated with *arrows*. We see that the *arrows* take one from 'source' SCCs (marked with '+'), via (typically) intermediate SCCs, to 'sink' SCCs (marked with '−')

If we now recall the real, concrete Web graph, our intuition might lead us to say that 'most' Web pages participate in some kind of large-scale community of pages, with circulation from any page to any other. We will see that this intuitive picture is not particularly misleading. Yet, at the same time, we reiterate that because the Web graph is a directed graph, which is almost certainly not a single SCC – it *must* have one or more sets of pages for which there are no inlinks to the set. These are the source SCCs of the Web graph. Similarly, the Web graph is guaranteed to have one or more sink SCCs – sets of pages such that a reader who arrives to the set via a hyperlink will not find any path leading out again. Furthermore, we cannot expect just one single, almost-all-encompassing intermediate SCC lying between these sources and sinks. The Web graph is built up by billions of uncoordinated individual acts; and its structure (as we will see in the next subsection) reflects its anarchic and self-organized nature.

The Web Graph

The Web graph is a directed graph. The edges are hyperlinks which point from page A to page B. Using a normal web-browser, there is no way to jump from page B to A if the hyperlink points from A to B. The Web graph is also extremely huge, and is continuously growing and changing. It expands by the addition of nodes (webpages) and of new edges (hyperlinks). Anyone can add pages and links to the WWW, so the structure of the Web graph becomes very complex. The Web graph has grown into a massive world-wide 'organism' [32]. An interesting feature of the Web graph in relation to its complex structure is the indegree-distribution. As the age of a page increases, the page might become more and more popular, such that other pages start linking to it. Studies have shown that the indegree-distribution of the Web graph follows a *power-law distribution* [9] and [52]. A discrete random variable X follows a power-law distribution if the probability of X taking value x is proportional to $x^{-\gamma}$, with $\gamma > 0$ being the power-law exponent. Barabasi and Albert [9] found that this feature was a consequence of new nodes constantly being added, and that new vertices are preferentially attached to already well-connected nodes.

Broder et al. [21] studied a set of 200 million webpages with 1.5 billion links and confirmed the power-law scaling of indegrees, with $\gamma = 2.1$. In addition they found that the Web graph has a structure schematically described by Fig. 2. This figure is now generally accepted as describing the structure of the World Wide Web.

The structure in Fig. 2 has been called a 'bow tie' – for

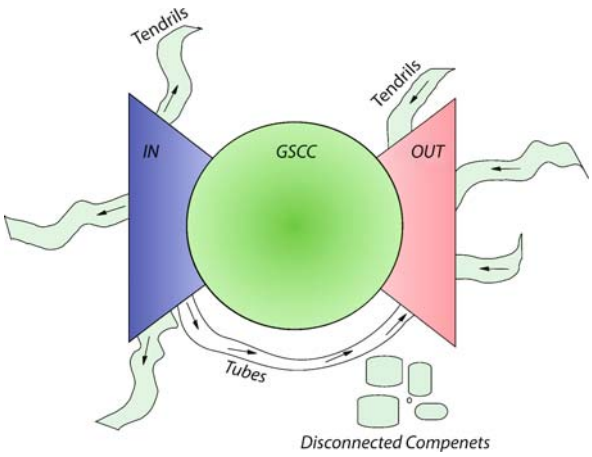

Link Analysis and Web Search, Figure 2
The bow-tie structure of the WWW as found by Broder et al., reproduced after [21]

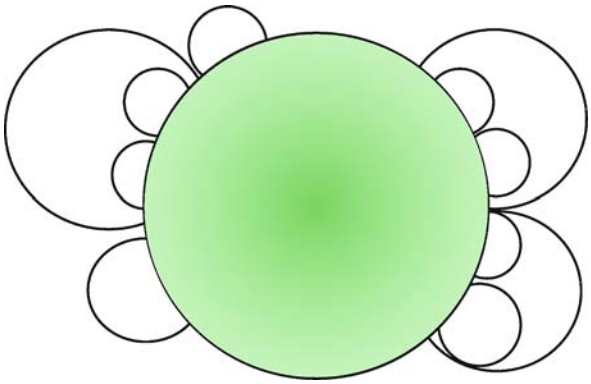

Link Analysis and Web Search, Figure 3
Daisy. The structure of the WWW as found by Donato et al. The IN and OUT set are highly fragmented and lie close (in hops) to the highly connected CORE. Figure reproduced after [32]

obvious reasons. The 'knot' of the bow-tie, the CORE, is a giant SCC (GSCC), which lies *centrally* (in an important sense) in the Web graph. The IN component represents another huge set of nodes. A Web surfer who starts in the IN component may reach the giant SCC from a node in IN; but no node in IN is reachable from the GSCC. IN should be thought of as many SCCs, many of which (but not all) are source SCCs. Similarly, nodes in OUT may be reached from the GSCC – but there is no path from any node in OUT into the GSCC. Also, we see that some paths which leave IN do not reach the GSCC; instead they terminate in 'out-Tendrils' to which IN is linked. OUT has its corresponding 'in-Tendrils' (which must include some source SCCs).

Still we have not exhausted the possible categories of Web page positions in the Web graph. We see from Fig. 2 that there are paths from IN to OUT which never reach the GSCC. Nodes in such paths lie in the 'Tubes'. Finally, there are pages which are not even connected to the large, connected structure which is composed of IN, GSCC, OUT, Tendrils, and Tubes. These pages are noted as 'Disconnected components' in Fig. 2.

The corresponding sizes of the components in the connected, 200M subset of the WWW were found to be: 44M in the IN-set, 56M in the GSCC, 44M in the Tendrils and 44M in the OUT-set. In the same way as the distributions of indegrees, the sizes of SCCs inside the main components were also found to follow a power-law distribution.

In Donato et al. [32] the inner structure of the "bow-tie" was studied, and a refined picture of the Web graph was suggested. They found that the main components (IN, OUT, GSCC=CORE, TENDRILS) have differences on a finer level. IN and OUT were found to be highly fragmented, while the CORE was well interconnected. No large SCC was found in the IN and OUT components. Most of the nodes contained in these sets belong to very small SCCs, and lie within a short distance from the GSCC. The resulting picture that emerges from [32] is shown in Fig. 3. We see that the IN and OUT sets are fragmented into small and poorly interconnected SCCs, hanging from the large and well-connected GSCC.

The CORE of the Web graph is highly interconnected and has "Small World" properties [3,6]. Graphs found in several man-made systems have these properties. For example the power grid of the western US, the collaboration graph of film actors and the neural network of the worm Caenorhabditis elegans, all share the property of being "small-world" networks [3]. A 'small world' graph is highly clustered (so that the idea of 'neighborhoods' makes sense), while at the same time the minimum distance between any two nodes in the network is short [85]. Albert et al. estimate an average of 19 clicks between any two pages of the WWW (at that time estimated to contain $8 * 10^8$ pages), based on a crawl of the nd.edu site [6]. Adamic finds an average distance of 4.2 between any sites in a GSCC containing about 65 000 sites [3]. Using a 200M nodes crawl of the WWW, Broder et al. finds an average distance of roughly 16 [21]. This means that even if the Web graph is large and complex, the diameter is small, and a huge amount of information is in principle reachable within a few clicks.

To implement a successful web link analysis, the link structure of the web has to be mapped.

Crawling is used to build large indexes of the WWW. These indexes are used by search engines, when a query is submitted, to find the hits which contain the search terms.

In addition, the web link structure is obtained by crawling, and stored. A crawler starts at some page (or set of pages), collects the information and outlinks on this page, and continues to the next by following one of the collected outlinks. Given the directed "bow-tie" structure of the Web graph, its size and its complexity, crawling is not a trivial task. A crawl starting on a page in the OUT set will never find pages in the GSCC or IN, since the crawler only sees the outlinks on any given page. More generally, a crawl starting in any SCC will only be able to reach "downstream" SCC's (see discussion on SCC's and Fig. 1).

Knowledge of the webgraph's structure is used to improve crawling strategies and to create effective crawlers [63]. Results from [78] imply that the structure of the obtained graph might depend on the crawler used.

In summary: Figs. 2 and 3 reveal the Web graph as a somewhat wild and woolly community of Web pages, which is built up by the combined, but largely uncoordinated, actions of humans and machines. It is the task of Web link analysis to extract meaningful and useful information from this structure – one which, furthermore, is constantly growing and changing. This is a formidable task; and yet, remarkably enough, it is not impossible.

The 'Sink Problem'

If we view hyperlinks as recommendations, then there are clearly pages which receive little or no collective recommendation from the Web graph. For example, pages lying in source SCCs may recommend one another; but no page from 'outside' this set recommends any of these. Similarly, IN is in some sense more weakly recommended than OUT. The point is that the asymmetric nature of hyperlinks seems to favor 'selfish' pages (and sets of pages) which manage to accumulate recommendations, without giving out recommendations. Put differently: the arrows themselves in Fig. 1 show that there is a real tendency for 'recommending weight' to accumulate on the 'downstream' side. In some cases (as we will see below) this tendency takes the extreme form that 'upstream' SCCs may receive *no* weight (and hence a score of zero), even though they are far from being isolated, and may be, by most criteria, important.

We will call this generic tendency the 'sink problem'. We use the word 'problem' because some approaches to link analysis give poor or useless results unless some corrective measure is applied – a 'sink remedy'. We will look at the sink problem in more detail in Sect. "Sensitivity of Link-Analysis Scores", after we have described several distinct link analysis algorithms in the next Section.

Approaches to Web Link Analysis

In the previous section we have sketched the structure of the Web graph. We see that this structure is heavily influenced by the one-way nature of the hyperlinks. In this section, we examine many of the studied approaches to link analysis. The problem is simply stated: given a page p in the Web graph, how can we generate a meaningful 'importance score' for p, which is based on the *position* of p in the entire graph?

Link Popularity

Here the idea is very simple: a Web page is a good page if it has many pages pointing to it – and hence recommending it. Therefore, one can simply count the inlinks to a given Web page to get a quantitative measure of its quality or importance. The resulting score k^{in} is commonly termed the 'link popularity' of the page – since it is entirely determined by the number of 'votes' given to a page.

This idea is intuitively reasonable. Also, it is very easy to calculate: once the full Web graph is obtained, the 'invisible' inlinks become fully visible, and the counting of these is a computational task which can be done very easily. The obtaining of k_i^{in} for every page i is equivalent to multiplying a matrix times a vector once. To see this, let A be the *adjacency matrix* of the Web graph, such that $A_{ij} = A_{i\to j} = 1$ if i points to j, and 0 otherwise. Then

$$k_i^{\text{in}} = \sum_j A_{j\to i} = (A^{\mathrm{T}}\mathbf{1})_i \,. \tag{1}$$

Here A^{T} is the transpose of A, and $\mathbf{1}$ is a vector of ones.

The idea of link popularity is simple, and it has a certain utility. However, link popularity is particularly vulnerable to the problem of *link spam*, by which 'phony' pages and links are used to boost the importance scores of 'real' pages. Clearly, it is easy and cheap to set up a large number of fake pages which all point to a given target page T – thus giving T a high (and meaningless) link popularity score. What is clearly needed is some LA approach which takes into consideration the *quality* of the 'voting' pages responsible for the high popularity score. The approaches described in the following subsections attempt to do this, in distinct ways.

We conclude this subsection by mentioning *anchor text scoring*. This is a hybrid of link popularity and text relevance: a page P gets a high anchor text score, relative to the search term 'bla', if there are many other pages with inlinks to P, and those other pages associate the anchor text 'bla' with the links to P. It is intuitively reasonable that a short text, which is used by a page P_{out} to describe a page P_{in}, may be a good succinct description of the

linked-to page P_{in} (but not always – see the discussion on 'Google bombing' below).

HITS

The fundamental idea that may be expected to correct for the effects of link spam – largely, if not completely – is to insist that the *quality* of the pointing pages should be taken into account in assigning the importance of the pointed-to page. This means, among other things, that worthless dummy pointing pages should contribute nothing to the score of the pointed-to page.

This fundamental idea forms the foundation for all the other methods for LA that we will discuss. The idea clearly can be expressed in many ways; but it is important to note that, since the quality of the pointing pages is used to find the quality of the pointed-to pages, one finds oneself in fact in possession of a circular definition. We may express this circularity most succinctly as follows:

Good pages are pointed to by good pages. (2)

The HITS approach of Jon Kleinberg [52] is an implementation of the idea embodied in (2), but in a form in which every document gets *two* types of importance scores – one for the extent to which it is a good '*hub*', and one expressing the extent to which it is a good '*authority*'. These two roles may be expressed (again circularly) in terms of each other, as follows:

Good hubs point to good authorities. (3)

Good authorities are pointed to by good hubs. (4)

Note that (3) and (4) are redundant – that is, (4) is simply a restatement of (3). Together, these statements amount to a circular definition of two kinds of quality: 'authoritativeness' and 'hubness'. Kleinberg's idea was that – just as with people – some documents may be viewed as good sources to consult, because of the special knowledge that they have about some theme. These documents are then the authorities. Similarly, other documents may be the best places to go if one wants to find out where are the best authorities. These 'other' documents are hubs – they are like the friend that does not know so much about repairing cars, but does know where to find a good mechanic.

Kleinberg proposed that we view all Web pages as having – to some degree – both of these functions. We can quantify their two roles – hub and authority – by translating (3) and (4) into equations. We let $h(i)$ be the hub score of node i, and $a(i)$ its authority score. Then we relate the two as follows:

$$h(i) = (\text{const}) \times \sum_j A_{i\to j}\, a(j) = (\text{const}) \times (Aa)_i \quad (5)$$

$$a(i) = (\text{const}) \times \sum_j A_{j\to i}\, h(j) = (\text{const}) \times (A^{\mathrm{T}}h)_i\,. \quad (6)$$

Here we have used the symbol a to represent the vector of authority scores, while h is the vector of hub scores. We insert an undetermined constant (const) in each equation because, without it, a consistent solution is in general impossible – as we shall soon see. Equations (5) and (6) are simply the expression of (3) and (4), respectively, in mathematical form. For example, in (5), we estimate the hub score of page i by simply using the authority scores of all pages j that i points to – consistent with the verbal expression (3).

Equations (5) and (6) may be rewritten as

$$h = \left(\frac{1}{c_h}\right) Aa = \left(\frac{1}{|Aa|}\right) Aa \quad (7)$$

$$a = \left(\frac{1}{c_a}\right) A^{\mathrm{T}}h = \left(\frac{1}{|A^{\mathrm{T}}h|}\right) A^{\mathrm{T}}h\,. \quad (8)$$

Here (following Kleinberg) we choose the constant, in each case, to give a vector of unit length as a result. Kleinberg in fact presents the process of assigning authority and hub scores as an iterative application of (7) and (8), and shows that this iterative process converges. Calling the converged score vectors a^* and h^*, we find from (7) and (8) that these must be solutions to

$$\begin{aligned} h^* &= \left(\frac{1}{c_h^*}\right) Aa^* = \left(\frac{1}{c_h^*}\right) A \left(\frac{1}{c_a^*}\right) A^{\mathrm{T}}h^* \\ &= \left(\frac{1}{c_h^* c_a^*}\right) AA^{\mathrm{T}}h^* \end{aligned} \quad (9)$$

$$\begin{aligned} a^* &= \left(\frac{1}{c_a^*}\right) A^{\mathrm{T}}h^* = \left(\frac{1}{c_a^*}\right) A^{\mathrm{T}} \left(\frac{1}{c_h^*}\right) Aa^* \\ &= \left(\frac{1}{c_a^* c_h^*}\right) A^{\mathrm{T}}Aa^*\,. \end{aligned} \quad (10)$$

Thus we see that the hub and authority score vectors h^* and a^* must be *eigenvectors* of, respectively, the matrices

$$\boldsymbol{H} = AA^{\mathrm{T}} \quad (11)$$

and

$$\boldsymbol{A} = A^{\mathrm{T}}A\,. \quad (12)$$

We note that both the Hub matrix $\boldsymbol{H}$ and the Authority matrix $\boldsymbol{A}$ are symmetric matrices. Also, it is clear from

our discussion of Eqs. (5) and (6) that right multiplication by the matrix A sends weights "backwards" – i. e., against the arrows, hence towards the pointing node – while right multiplication with A^{T} sends the scores forwards – with the arrows, towards the pointed-to node.

We can see this by examining a typical matrix element of (for example) $\boldsymbol{H}$:

$$\boldsymbol{H}_{ij} = \left(AA^{\mathrm{T}}\right)_{ij} = \sum_k A_{ik}A^{\mathrm{T}}_{kj} = \sum_k A_{i\to k}A_{k\leftarrow j}\,. \quad (13)$$

Thus, when $\boldsymbol{H}$ acts (by right multiplication) on a vector h of candidate hub scores, it will send scores from a node j, first forward to j's outlinks k (which are thus candidate authority nodes), then backwards to these nodes' inlinks – for example, i – which is thus rated for its role as a hub. In short, $\boldsymbol{H}$ allows nodes to send one another (via two hops) their hub scores; and a node j with a high hub score (which thus "thinks" it points to good authorities) will send (via action of $\boldsymbol{H}$) larger hub weight to other nodes i which point to the same nodes (good authority nodes) as j does. This sending is iterated until convergence – at which point all nodes agree on their respective hub scores. The same logic applies to authority scores; one simply exchanges, in the above discussion, authority ↔ hub, forward ↔ backward, pointing ↔ pointed-to, etc.

As noted after Eqs. (9) and (10), the converged scores h^* and a^* are eigenvectors of (respectively) $\boldsymbol{H}$ and $\boldsymbol{A}$. However, we know that these matrices, for a Web graph with N nodes, will be $N \times N$ matrices, which may have as many as N distinct eigenvectors [38]. Thus we must ask: which eigenvector do we want?

The answer is simple. As noted by Kleinberg, the iterative process of sending scores will converge (almost always) to a unique vector, called the *principal eigenvector* of the matrix in question ($\boldsymbol{H}$ or $\boldsymbol{A}$). The principal eigenvector is that corresponding to the largest eigenvalue. For nonnegative matrices such as $\boldsymbol{H}$ or $\boldsymbol{A}$ (and also for A), the largest eigenvalue is positive, and the principal eigenvector is nonnegative (positive or zero) at every node Theorem 2.7 in [83]. In fact, as long as the graph represented by $\boldsymbol{H}$ or $\boldsymbol{A}$ is connected, the principal eigenvector will be strictly positive at each node – because these matrices are symmetric. Hence we can expect a positive, nonzero hub and authority score for each node as a result of this process.

Kleinberg designed the HITS method to work on subgraphs of the WWW. That is, he assumed that one is given (from some search engine) a hit list that is large, because the topic of the search is not very narrow. The hit list likely includes many good authorities, but also many uninteresting pages. To find the good authorities (and hubs) for the given topic, he proposed to take the top-ranked t pages from the hit list (thus forming the *root set* of pages), and then to augment this root set by including (i) all pages pointed to by the root set, (ii) some or all pages pointing to the root set, and (iii) all links internal to the resulting, augmented, *start set*. Kleinberg argued that the resulting start set should contain most of the good authorities for the topic in question. Because of this property of using hyperlinks to 'zero in' on a topic, the method acquired the name Hypertext-Induced Topic Selection or HITS.

We note here that Kleinberg's choice of using a topic-focused subgraph – built up from a hit list – is not logically necessary for the application of the mathematics of the HITS method, as defined by Eqs. (9) and (10). We can think of this choice (subgraph taken from a hit list) as a *'real-time'* approach, since it requires that LA be performed *after* the generation of a hit list (thus, in response to a query, and so in real time). Alternatively, one can apply the HITS method (as is done with PageRank, see below) to the *entire* Web graph. When the whole graph is used we call the approach an *'offline'* method, i. e., one which is carried out in background mode, and which gives query-independent results. Thus, the reader should keep in mind that any method for LA (mapping a directed graph to a score) may be applied either as a real-time or as an offline approach.

The key ideas of HITS are by now (we hope) clear:

(i) Define the quality of Web pages in a circular way – that is, in terms of the quality of other Web pages.
(ii) Use the link structure of the Web graph to express these definitions as equations.
(iii) The resulting equations call for the principal eigenvector of a matrix; this eigenvector gives the desired quality scores.

These three basic steps underlie all of the methods we will discuss, beyond the most basic method of link popularity.

PageRank

Now we discuss a method which was originally devised to be used on the entire Web graph. The PageRank method [68] gives a single score for each Web page – its PageRank. This score is best regarded as a kind of authority score, in that it is generated (like link popularity) from inlinks.

We begin with the basic concept as expressed in (2). However, we will only consider one type of "goodness" or quality score; hence we need not expand (2) to (3)+(4).

Instead we express (2) mathematically as:

$$p_i = \sum_{j\to i}\left(\frac{1}{k_j^{\text{out}}}\right)p_j = \sum_j\left(\frac{1}{k_j^{\text{out}}}\right)A^{\text{T}}_{ij}p_j$$
$$= \sum_j \left(A^{\text{T}}\right)^{(\text{col})}_{ij} p_j\,. \tag{14}$$

The first equality in (14) tells us that – as prescribed by (2) – node i gets its PageRank score p_i from all nodes j pointing to i. However, it also says that nodes j do not send their full quality score (weight) to i; instead, sending nodes j must divide their weight by the number of outlinks k_j^{out} which they have. Thus, the PageRank approach, as expressed in (14), is rather "democratic": every page can only "vote" once (using its current score). That is, a good page cannot give its full weight to each of its outlinks, because the *total* weight it sends must be conserved – hence, divided up among its outlinks. This is expressed in another way by saying that the forward-sending matrix A^{T} is *normalized*. More precisely, $(A^{\text{T}})^{(\text{col})}$ is a *column-normalized* version of A^{T}, since the entries in each (nonzero) column are adjusted so as to sum to one.

Note that there is no constant prefactor in (14). The reason is that we know the constant before we do any calculation; and it is one. We can (as before) reformulate (14) into an eigenvector equation:

$$p = \left(A^{\text{T}}\right)^{(\text{col})} p\,; \tag{15}$$

but in this case we know that the (principal) eigenvalue is 1 – because of the normalization property of $(A^{\text{T}})^{(\text{norm})}$ [13].

On the face of it the PageRank approach (14)–(15) appears to be simpler than the HITS approach: just one equation, giving one set of scores p. This simplicity has a price however.

We recall that the HITS approach gives a positive score (actually, two positive scores) to every page, as long as the effective graph ($\boldsymbol{H}$ or $\boldsymbol{A}$) is connected. This nice positive definite property is guaranteed for the principle eigenvector of any strongly connected graph, and hence comes from the fact that the effective links (in $\boldsymbol{H}$ or $\boldsymbol{A}$) are symmetric. PageRank uses the simpler operator $(A^{\text{T}})^{(\text{col})}$,which is *not* in general symmetric. This means that, even given a connected Web graph, one can expect in general exactly zero weight for a large fraction of all the pages. This is an undesirable outcome: it means that, for the set of nodes with exactly zero weight, (i) one is forced to regard them as having zero quality, and (ii) LA provides one with no basis whatsoever for comparing them.

One can in fact show [16,17] that all pages in the Web graph – except those lying in sink SCCs – will get zero weight in the principal eigenvector of $(A^{\text{T}})^{(\text{col})}$. This fact agrees with our intuition: all SCCs lose weight to downstream SCCs, except sinks (which have no downstream SCCs). Thus we have come up against the 'sink problem' mentioned earlier – and in a rather severe form that clearly justifies its name.

Brin and Page [19] offer their own solution to this sink problem. The normalization of their forward operator allows them to think of the score p_i as a *probability* – more precisely, the probability that a random walker, following outlinks from one page to another, will (in the limit of many hops) visit page i. They then allow this "random Web surfer" to hop (again at random) from any page to any other, with probability ε/N. To avoid violating weight (i. e., probability) conservation, they correspondingly adjust the probability of following an outlink from j to be $(1-\varepsilon)(1/k_j^{\text{out}})$. In terms of matrices, this amounts to adding an all-to-all 'random surfer matrix' (or 'teleportation matrix') $\hat{R}$ to the given matrix $(A^{\text{T}})^{(\text{col})}$, and weighting the sum to retain probability conservation. The PageRank eigenvalue equation then becomes

$$p = \left((1-\varepsilon)\left(A^{\text{T}}\right)^{(\text{col})} + \varepsilon\hat{R}\right)p\,. \tag{16}$$

The all-to-all matrix $\hat{R}$ (with $\hat{R}_{ij} = 1/N$) makes the effective network which is represented by the sum $((1-\varepsilon)(A^{\text{T}})^{(\text{col})} + \varepsilon\hat{R})$ a single SCC. Hence there are no sinks in this effective network, and thus no sink problem: all entries in p are positive definite.

We note that the term $\hat{R}p$ in (16) can be simplified: $\hat{R}p = (1/N)\left[\sum_i p_i\right]\mathbf{1} \equiv v$ (where $\mathbf{1}$ is a vector of ones). That is, it gives a vector v (the 'teleportation vector') which is known, because $\sum_i p_i = 1$ is known. Because the teleportation matrix is completely unbiased, the teleportation vector is also unbiased, or 'flat'. We will look at applications for a biased teleportation vector in Sect. "Personalization of Link Analysis Scores".

We note that the factor $(1-\varepsilon)$ in (16) is often called the *'damping factor'*. This reflects the fact that a common approach to computing PageRank scores is the *power method* [48] – in which one finds the principal eigenvector of a matrix by repeating the operation (matrix) × (vector) until convergence. One finds that convergence is faster for smaller values of the damping factor [55]; this is related to the fact that the eigenvalue gap (discussed below) increases with ε, and hence with decreasing damping factor. Brin and Page [19] report obtaining good convergence after 50–100 iterations of the power method, with $\varepsilon = 0.85$. Langville and Meyer [55] discuss the practical consequences (in terms of score accuracy) of these choices.

Kleinberg [52] also used the power method to find hub and authority scores. Here – because the principal eigenvalue is larger than 1 – it is necessary to periodically rescale the length of the converging vector, as one iterates (matrix) × (vector); otherwise the vector's length will grow arbitrarily large.

SALSA

The 'SALSA' algorithm (Stochastic Approach for Link-Structure Analysis) was investigated by Lempel and Moran [58]. In the SALSA approach, the normalized Hub operator is defined as

$$\widetilde{\boldsymbol{H}} = A^{(\mathrm{row})}(A^{\mathrm{T}})^{(\mathrm{col})} \tag{17}$$

and the normalized Authority operator is correspondingly

$$\widetilde{\boldsymbol{A}} = (A^{\mathrm{T}})^{(\mathrm{col})}A^{(\mathrm{row})} \,. \tag{18}$$

Here $A^{(\mathrm{row})}$ is the row-normalized (i. e., by indegree) adjacency matrix, and $(A^{\mathrm{T}})^{(\mathrm{col})}$ is the column-normalized transpose which we have seen before in the PageRank algorithm. The logic is thus (we hope) clear: SALSA seeks to combine the two-level approach of HITS with the normalization of PageRank. For example, (18) says that authority weights are sent via two hops: first backwards to hub candidates, using $A^{(\mathrm{row})}$ – with the row (indegree) normalization ensuring that each candidate authority can only use its full authority weight once in 'voting' for candidate hubs. Then $(A^{\mathrm{T}})^{(\mathrm{col})}$ allows these candidate hubs to 'vote' once for new candidate authorities, by dividing the weight they have gotten among their outlinks.

The SALSA approach clearly offers a new set of basic choices for link analysis. However – remarkably – the eigenvectors of the SALSA operators $\widetilde{\boldsymbol{H}}$ and $\widetilde{\boldsymbol{A}}$ can be proven [58] to give the same scores as simple link popularity. That is, to within vector normalization, the score of a node in the dominant eigenvector of $\widetilde{\boldsymbol{A}}$ is its indegree, and the score of a node in the dominant eigenvector of $\widetilde{\boldsymbol{H}}$ is its outdegree.

This result is very similar to a familiar result for undirected graphs. That is, it is known [62] that, when A is symmetric (as is the case for undirected graphs), the dominant eigenvector of the column-normalized $A^{(\mathrm{col})}$ is (except again for vector normalization) simply the node degree. Since column normalization gives weight conservation, one can also interpret the (vector normalized) scores as probabilities. Hence, this known result for undirected graphs says that, in the limit of a very long random walk, the probability of visiting each node is proportional to the node degree. Similarly, the SALSA results say that a random walk which always follows a forward hop with a backward hop (and vice versa) will (in the limit of very long time) reach a node after a forward hop with probability proportional to that node's indegree, and will reach a node after a backward hop with a probability proportional to that node's outdegree.

This result is even more remarkable in light of the fact that the naïve generalization of these ideas to the one-level, normalized case (PageRank) does *not* work. That is, it is not the case that, in the dominant eigenvector of $(A^{\mathrm{T}})^{(\mathrm{col})}$, each node's score is proportional to that node's indegree.

We can offer an intuitive understanding of all these results, based on a simple idea. Suppose we have a special directed graph satisfying the following constraint:

$$k_i^{\mathrm{in}} = k_i^{\mathrm{out}} \,, \quad \forall i \,. \tag{19}$$

That is, every node in the graph has (number of inlinks) = (number of outlinks). Now we show that, for this graph, the vector $\vec{k} = \vec{k}^{\mathrm{in}} = \vec{k}^{\mathrm{out}}$ is the dominant eigenvector of $(A^{\mathrm{T}})^{(\mathrm{col})}$. Suppose that each node i has weight k_i at some step of the weight passing process. We then apply $(A^{\mathrm{T}})^{(\mathrm{col})}$ to send weights again; this makes each node send out exactly one unit of weight over each outlink. At the receiving end, each node j receives exactly $k_j^{\mathrm{in}} = k_j$ units of weight. Thus the weight distribution is unchanged by the weight passing – it is an eigenvector. That is, the node degree gives the eigenvector (hence LA score), for this special graph, for the PageRank operator without teleportation. We then note that every undirected graph always satisfies Eq. (19) – hence we see that the node degree is also the principal eigenvector for a normalized symmetric graph. Finally we consider the SALSA approach. Suppose every node i has weight k_i^{in} after a Forward iteration. Then the next hop (Backward) will send exactly one unit of weight (Backward) over each of i's inlinks – for every node i – so that each receiving node j receives exactly k_j^{out} units. It is easy to see that the next Forward operation takes us back to our starting point. Hence again we get the (in- and out-) degree as the eigenvector of (respectively) $\widetilde{\boldsymbol{A}}$ and $\widetilde{\boldsymbol{H}}$.

T-Rank

The three approaches presented above (HITS, PageRank, and SALSA) are by far the most well studied approaches to Web link analysis. Now we attempt to place all three (and a new one) in a simple, common picture.

HITS and SALSA were first proposed as a *postprocessing* (real-time) form of LA: one first generates a hit list using conventional text relevance analysis, and then generates and analyzes a topic-focused subgraph from this hit list. PageRank, in contract, is a preprocessing, 'offline' form of LA, in which the entire Web graph is analyzed,

and the resulting LA (PageRank) scores are *search-independent* – they are 'global'.

Hence we have a fundamental choice to make in choosing a LA approach: global and offline, or postprocessing on a topic-focused subgraph. This choice clearly has important implications, both in terms of the results obtained from LA, and in terms of practical considerations. In this review we will be essentially 'agnostic' with regard to this choice. That is, we focus primarily on the mathematical structure of the LA algorithm, and on the consequences of that structure, without addressing in detail the question of whether the algorithm is applied to a subgraph or to the entire Web graph.

The mathematical structure of HITS is simply stated. The approach is 'two-level' – in that each type of score (authority and hub) is passed from node to node in two hops – and it does not use matrix normalization. SALSA is also two-level, but uses matrix normalization, so that the weight passing is weight conserving. Finally, PageRank is a one-level approach – authority scores are passed directly from a node to its neighbors, in a single hop – and it uses matrix normalization.

Thus, the fundamental mathematical choices distinguishing these three approaches boil down to two binary choices: one- or two-level, and normalized or not.

We summarize this very simple analysis in Table 1. An alert reader will note that two binary choices generate four possible approaches; and we have included, in Table 1, a very recently studied approach (*'T-Rank'*) in the upper right-hand box [15,16]. T-Rank is a non-normalized, one-level approach. In other words, the vector of T-Rank scores t is taken to be the principal eigenvector of A^{T}:

$$A^{\mathrm{T}} t = \Lambda t \,, \tag{20}$$

where Λ is the principal eigenvalue of A^{T}. T-Rank has also been studied with the teleportation sink remedy, in which case we have

$$\Lambda' t = \left(A^{\mathrm{T}} + \varepsilon \hat{R}\right) t \,. \tag{21}$$

Here there is no need for a factor $(1 - \varepsilon)$, since there is no attempt to normalize the composite matrix $(A^{\mathrm{T}} + \varepsilon \hat{R})$.

Link Analysis and Web Search, Table 1

A simple tabulation of some fundamental choices in link analysis, including a recently studied approach (T-Rank) which seeks to fill the 'hole' in the upper right

	Normalized	Non-normalized
1-level Forward	PageRank	T-Rank
2-level (compound) H and A	SALSA	HITS

The T-Rank approach has received very little study, especially in comparison to the 'big three'. However we see that it is, in an a priori sense, equally worthy of study: a thorough understanding of link analysis should include an understanding of all the options in Table 1. Hence we point out here the existence of T-Rank – both logically and in [16]. The latter paper includes a preliminary comparison of all four entries in Table 1; a principal (tentative) conclusion from this comparison is that changing columns has a more profound effect on the results than does changing rows. In other words, the choice of whether or not to normalize the weight passing has a stronger effect on the nature of the eigenvector than does the choice of one- versus two-level weight passing. These effects were quantified in [16] in terms of the strength of the 'tightly-knit community' (TKC) effect [58] – in which a relatively small elite of mutually reinforcing pages dominates the scores – and in terms of the tendencies of the scores to correlate strongly with node degree. More specifically, the right column of Table 1 (non-normalized approaches) tended to give a stronger TKC effect, while the normalized approaches (left column) tended to show a stronger correlation of LA scores with node degree (which was of course complete for SALSA).

Another benefit of seeing LA in terms of Table 1 is that one can begin to imagine *intermediate* approaches. For example, the choice to normalize or not need not be strictly binary – one can instead imagine, in the top row of Table 1, a continuous tuning of the degree of normalization, from full normalization (PageRank) to no normalization (T-Rank). A simple (non-unique) interpolation formula which accomplishes this is given in [16]. Here it is also shown that the strength of the TKC effect may be continuously varied, by varying the strength of the normalization.

Of course, the picture given in Table 1 is highly simplified, and the reality is not so tidy. We mention for example the study of Ding et al. [31], who also placed the big three approaches in a common framework. This framework however was distinct from that of Table 1, and T-Rank was not mentioned. The tendency of LA scores to correlate with node degree was also noted here.

Another feature which is missing from the Table is the choice of Forward or Backward propagation. This choice is irrelevant for the two-level approaches (which use both types of propagation, but in alternation), but it is significant for the one-level approaches. Ding et al. [31] mention using one-level backward propagation in order to generate a kind of hub score without resorting to a two-level approach. Mathematically, this involves simply using the adjacency matrix A – normalized, or not – rather than its

transpose. This one-level hub-score approach has received little study. Mathematically, it is the same as the authority approach, since a matrix and its transpose are distinguished only in the details. In practice, of course, with the real Web graph generating the matrix, the choice of forward or backward propagation is significant.

The 'Sink Problem'

We have argued above that, in some cases, the sink problem can have a strong and undesirable effect on LA scores. We revisit this question in this section, examining each of the approaches discussed in Sect. "Approaches to Web Link Analysis".

HITS and SALSA

We can discuss HITS and SALSA simultaneously here, because each of these approaches solves the sink problem in the same fashion – i. e., by building an effective, two-hop graph which is symmetric. Symmetric graphs have no sinks or sources, hence no sink problem.

This solution is not necessarily problem-free however, because the resulting effective graph may be disconnected [66]. This can happen if, for example, two subgraphs are not reachable, one from the other, via a forward hop followed by a backward hop (i. e., by application of the Hub operator). The possibility of the effective graph being disconnected is discussed in [58]; here there is offered a simple solution, with weights distributed among the disconnected components according to the number of nodes in each component.

PageRank

The PageRank approach as expressed in Eq. (16) has a sink remedy built in; and typically, by the term 'PageRank' one means this 'remedied' combination of normalized forward weight passing – using $(A^{\mathrm{T}})^{(\mathrm{col})}$ – and teleportation – using $\hat{R}$. In the absence of $\hat{R}$ (i. e., with $\varepsilon = 0$), the PageRank approach has a severe sink problem, as noted in [8,14,17], and [16]. This may be understood rather simply: the PageRank eigenvector (with $\varepsilon = 0$) represents the asymptotic (in time) probability distribution for a random walker which moves over the directed links without help of the random-hop option. Clearly, such walks become 'stuck' in the sinks in the limit of large time. Hence, in the absence of any sink remedy, the eigenvector will have zero weight everywhere on the graph except in sink SCCs.

For this reason, PageRank is only practical for nonzero ε. Much effort has been expended [55] towards the goal of finding the 'best' value for this parameter. Clearly, a compromise must be taken between two extremes, neither of which is attractive: for $\varepsilon = 0$, all weight is in sinks, while for $\varepsilon = 1$, the eigenvector is uniform and carries no information – all nodes are equally important. It is perhaps remarkable that eigenvectors for values of ε lying between these two meaningless extremes can in fact give highly useful and meaningful information about the status of Web pages in the Web graph.

The sink problem impacts PageRank in another way. We consider a 'dangling page' D which has only inlinks and no outlinks. This page D is thus a single-node sink SCC. Furthermore, as long as there is at least one such dangling node in the entire Web graph, the graph's adjacency matrix is not normalizable: page D is represented by a column of zeroes in A^{T}, and this column cannot be made to sum to one in $(A^{\mathrm{T}})^{(\mathrm{col})}$. Such dangling nodes represent (in the language of Bianchini et al. [14]) a source of 'energy loss': weight which is sent to them simply vanishes in the next iteration. Dangling nodes do occur in the Web graph, so that the problem is not academic. One idea [68] is to remove them; this idea has the corresponding problem that removing some dangling nodes may in turn create new ones. A common solution is to implement a 'back button': one allows for the possibility (with suitable normalization) of a hop to *any* other node from a dangling page D – which thus ceases to be a sink. With the back-button solution in place, there are no zero columns in the modified A^{T}, and so normalization is possible.

T-Rank

As noted above, the non-normalized, one-level approach has received little study. Some limited results may be found in [16]. Here it is shown that the sink problem is much less severe when the normalization constraint is dropped, in that the eigenvector – *without* sink remedy – has positive weight in the GSCC, and in the OUT set. This is in fact guaranteed to be the case, given the (almost certainly true) assumption that the GSCC has the largest eigenvalue of all SCCs. Also, numerical results are presented which suggest that adding a random teleportation matrix with weight ε has little effect on the eigenvector. In short: there is no evidence that $\varepsilon = 0$ is a singular point for T-Rank, as it is for PageRank.

It is also shown in [16] that two novel types of sink remedy are possible for the non-normalized approach (and that neither is possible with the normalized approach). One of them, termed 'source pumping', has the unique property that it is the *only* way to ensure a nonzero eigenvector over the entire graph, without making the graph strongly connected. We note that sink remedies are

of interest for T-Rank because, without some such remedy, the IN set, TENDRILS, and TUBES have exactly zero weight.

Sensitivity of Link-Analysis Scores

There are at least three real-world problems which complicate the application of link analysis to the Web. First, there is the fact that one is always working with incomplete information – it is impossible to crawl the entire Web. Secondly – in the case of one-level approaches – one normally applies a sink remedy such as the teleportation operator. This changes the (incompletely) known Web graph into another. Third, there is the very real likelihood that many of the links and even pages in the Web graph are placed with intent to influence the link analysis score – that is, there is "link spam".

It is thus important to examine how various link analysis approaches are affected by these three kinds of perturbations. We take them up in this Section, in order, starting with the incomplete information.

Sensitivity to Small Topology Changes

The Web graph which is input for LA differs from the "real" Web graph in that it lacks many existing pages and links. Furthermore, this input graph has links and pages which no longer exist, due to imperfect updating. Hence we must ask, how important is this source of error? More specifically, how much does the change from the "real" graph W to the measured graph W' affect the scores and rankings obtained from LA?

Early work in this direction was done by Ng et al. [66,67]. Here HITS and PageRank were examined for stability against small topology changes. In [66] it was shown that the change in PageRank scores is bounded above by $2\Sigma/\varepsilon$, where Σ is the sum of PageRank scores of the pages suffering the modification, and ε is the teleportation weight. Thus, for "reasonable" ε and for perturbations avoiding the most important pages, the bound is encouraging. In the same paper, the stability of HITS was related to its eigengap – the difference between the principal eigenvalue and the second largest. Experimental results implied that HITS was more sensitive than PageRank (with $\varepsilon = 0.2$). Thus we see a stabilizing effect from the teleportation operator. Motivated by this, [67] proposed a "randomized HITS" approach (HITS plus random teleportation), and gave evidence that it is more stable than the original version.

In [57], and in [14], the bound of Ng et al. is tightened to $2(1-\varepsilon)\Sigma/\varepsilon$. Also, [57] studied the stability of SALSA, finding it stable as long as the effective graph is connected, and proposed a stabilized form of SALSA, using (again) random teleportation.

Borodin et al. [18] studied similarities between, and stability of, HITS, variations on HITS, and node degree (INDEGREE), in a very general framework, and found that almost all algorithms (including HITS) are unstable by their definition and assumptions. Donato et al. [32] examined again the stability of HITS, for a class of graphs termed product graphs, and found that (under some assumptions) HITS scores were both stable, and very close to node degree (authority $\approx$ indegree). The product graph model allows for the possibility of realistic structures such as a power law.

Lempel and Moran [58] point out that stability of ranking is not equivalent to stability of scoring – since, in many cases, small changes in scores can lead to large changes in rank. They found further that SALSA is rank stable on the class of graphs for which it is scoring stable, while PageRank and HITS are not.

We emphasize here that we have not sought to give an exhaustive guide to the literature in this section. There exist many definitions of stability, and many possible perturbations. However, some simple facts are clear. The eigengap for PageRank can be tuned by tuning ε, and is in fact equal to ε for uniform teleportation [44]. Thus a stable behavior can be practically ensured for PageRank. Basic HITS in contrast has no such tuning parameter and so cannot guarantee a large enough gap – unless of course teleportation is also added here. SALSA is stable in the sense that it is local – so that local topology changes cannot affect the entire eigenvector. Finally, T-Rank has not been studied in regard to this question.

Sensitivity to Sink Remedies

Now we consider the sensitivity of PageRank to variation of the teleportation probability ε. Here we find a clear and simple picture: PageRank scores become unstable as one approaches the 'un-remedied' limit $\varepsilon \to 0$. This conclusion is already implicit in the results of Ng et al., cited above. In addition, Langville and Meyer [55] pointed out that the eigengap for PageRank vanishes as $\varepsilon \to 0$. A bound for derivatives of PageRank scores with respect to ε was presented; this bound blows up as $\varepsilon \to 0$. Boldi et al. [17] found closed forms for all orders of derivatives for PageRank scores, and graphically demonstrated the instability of the scores as $\varepsilon \to 0$. Similar results were found by Avrachenkov et al. [8].

All of these results are consistent with the fact (discussed above) that the PageRank distribution is pathological at $\varepsilon = 0$ – along with the well known fact that

it gives sensible results for ε not too small. The distribution of scores is also pathological (uniform) as $\varepsilon \to 1$. As expressed in [17]: "… PageRank oscillates between a meaningless uniform distribution [$\varepsilon \to 1$] and a meaningless distribution concentrated mostly in irrelevant nodes [$\varepsilon \to 0$]." This makes it all the more remarkable that PageRank gives, seemingly, stable and useful results when ε is held away from these limits.

Sensitivity to Spam

Because getting noticed has a monetary value, and because link analysis (applied to Web search) helps determine which pages are noticed, it is also true that understanding (and being able to manipulate) link analysis is worth money. It is also a good research problem. For these two reasons, the related arts of committing and of fighting *link spam* have received a great deal of attention.

The problem was discussed already in Sect 6.1 in [68] – in which use of a nonuniform teleportation vector was proposed. In another early work, Davison [30] discusses using machine learning to classify links and to recognize "nepotistic" (spam) links.

A fairly standard model for link spam was given in Gyöngyi and Garcia-Molina [40]. One has a set of *target pages* for which it is desired to boost the scores, and a set of *boosting pages* whose main or only function is to help the target pages. Also, the resulting *spam farm* may in addition exploit the benefit gained from laying in *hijacked links*. These are links from "real" (external to the spam farm) Web pages, but they are pages of a type in which the spammer is able to place outlinks pointing to the spam farm. (Examples of hijackable pages are unmoderated discussion fora and directories.) This work found (assuming PageRank for LA) optimal structures for a spam farm with a single target page and k boosting pages, and also for an alliance of two such spam farms. Bianchini et al. [14] presented similar results, but using a more general model for a spam farm. Du et al. [35] present a modification of the simple optimal rules of Gyöngyi and Garcia-Molina [40], obtained by relaxing an assumption of the earlier authors.

Gyöngyi and Garcia-Molina [41] presented a useful taxonomy of types of Web spam. (Note the name of the conference in this reference, underscoring the coming to maturity of the field!) They also presented some experimental data indicating that (i) about 7–18% of Web pages, in the years 2002–3, were spam pages, and (ii) this fraction seemed to be growing over time. In addition, there is a limited discussion of spamming HITS, in which it is clear that boosting a hub score is easier than boosting an authority score (as it is easier to choose one's outlinks than one's inlinks). An earlier paper [42] proposes the TrustRank algorithm for combating spam. Here, reputable "seed" pages are used to bias the teleportation vector in a PageRank approach, thus allowing the trustedness of the seeds to propagate throughout the distribution. (A version of TrustRank may apparently be tested live at http://trustrank.org/.) A similar method is described in Benczúr et al. [11]. Here, suspicious pages are recognized by their non-typical distribution of inlinks, and are then penalized by (again) biasing the teleportation vector. The difference here is that one punishes "bad" pages, rather than rewarding "good" ones. Taking this idea even further, one can compute a "BadRank", based on the principle that pages pointing to "bad" pages cannot be very "good". BadRank is thus a reverse propagation of PageRank, with seeds in the teleportation vector being known bad pages. BadRank has appeared on the Web in the form of rumors (http://pr.efactory.de/e-pr0.shtml), and has been studied by Liang et al. [61].

Zhang et al. [89] studied spam in the form of *collusion*: coordinated link structure manipulation by a group of targets (similar to the alliances mentioned above, but more general). For some simple models of collusion, it was found that the boost to be expected (in PageRank) is proportional to $(1/\varepsilon)$ – consistent with the early result of Ng et al. [66,67]. This suggested in turn a novel spam recognition mechanism: colluding nodes should be recognizable by an abnormally high sensitivity of their scores. The proposed remedy is again to bias the probabilities in the teleportation vector.

Another variant of spam is termed 'link bombing' or 'Google bombing' [2]. Here there is normally a coordinated use of anchor text, in order to boost the ranking of a target with respect to the anchor text in the boosting links. A famous example is the boosting of a web-biography of the US President with respect to the text "miserable failure". Although the purpose here is negative, clearly the same mechanism can be used to give boosts that the target would desire. Adali et al. show that uncoordinated individual attacks give optimal bombing.

The spam farm described above (with boosting pages and target pages) falls under the more general category of 'sybil strategies' [58]. In a 'sybil attack' on a *reputation system*, an individual entity masquerades as multiple simultaneous identities, in order to enhance its own reputation. Cheng and Friedman [26] show that no symmetric reputation function can be sybilproof. Here 'symmetric' means invariant under a relabeling of the agents (nodes, Web pages). This result is intuitively reasonable, since 'sybilproof' is a strong criterion (unlimited duplication is allowed), and since symmetric reputation systems (such as

PageRank) are essentially anonymous – i. e., dependent solely on the link structure. Cheng and Friedman [27] examine the manipulability of PageRank by sybil attacks, and find bounds on the attainable boost which are linear in the number of sybils – and hence consistent with the results of Garcia-Molina [40] and Bianchini et al. [14].

A recent paper [10] examines the possible gains in spam fighting by combining a variety of techniques. Included among these is a new one, Truncated PageRank, in which the normal damping factor $(1 - \varepsilon)$ is set to zero for neighbors within T hops. This approach is motivated by the fact that most spam links are close to the target page of the spam farm; also, it is consistent with the finding of Zhang et al. [89] that spam pages are more sensitive to damping factor. The authors report a spam detection rate of 80%, using the best found combination of (many) methods. This rate is improved further in the work of Castillo et al. [23] – to over 88% detection, at a cost of about 6% false positives. Castillo et al. use a machine-learning approach which combines both link attributes (in fact, all of the attributes used by Becchetti et al. [10]) and attributes based on examining the *content* of pages. Regarding the link attributes, Castillo et al. report evidence supporting the idea that spam and non-spam pages are not well interlinked: they find that spam pages are mostly linked to by other spam pages, while non-spam pages are mostly linked to by other non-spam pages. This statistical tendency was used to give a significant improvement in spam detection.

Clearly, Web link spam has received a great deal of attention. Much of this is no doubt based on the perception that Google's PageRank is and remains an important factor in its (highly successful and influential) approach to ranking. There is no sign of a "silver bullet" against link spam; and it seems likely that the fight between link spammers and anti-link-spam warriors will continue for as long as Web link analysis is useful.

Personalization of Link Analysis Scores

Web search engines, and in particular the approaches to Web link analysis discussed so far, have a "one size fits all" approach to ranking search results. Obviously, it is of interest to be able to skew the rankings, based on the preferences of the individual user. In this section we discuss some approaches to personalizing Web link analysis. We note that the teleportation vector is also commonly known as the "personalization vector", because it presents an obvious and simple mechanism (already pointed out in [68]) for skewing scores obtained from the PageRank algorithm. Also, we note that, since the skewing is normally expressed in terms of the content of the preferred pages, personalization is often equivalent to topic-sensitive scoring.

Rafiei and Mendelzon [74] propose methods – one a modification of PageRank, and one a modification of HITS – for assessing each page's reputation with respect to a topic. In each case, a biased random walk is implemented; in the PageRank case, the modification is small, while for HITS, one modifies the distribution of Forward hops to bias for the topic of interest. Chang et al. [25] also modify HITS, but in a different way. Here one or more preferred pages are given as a target, and the weights of the adjacency matrix are modified by adding the gradient towards the target.

A related (link reweighting) approach was studied in Richardson and Domingos [75]; here both the teleportation links and the hyperlinks of a PageRank approach are reweighted according to their relevance to a given target query (topic). Hence one obtains Query-dependent PageRank. It is proposed to precompute such scores for all one-word queries in the lexicon; and some discussion is given of the resulting scaling problems. Haveliwala [43] proposes instead to precompute such scores for a chosen set of "basis topics", and then to resolve any query in terms of these basis topics (16 in number in the implementation). Jeh and Widom [46] offer a more sophisticated approach in the same spirit, in which a set of partial vectors are precomputed, while the basis vectors are efficiently reconstructed as needed, such that the effective number of basis vectors can be large (tens of thousands). Haveliwala et al. [45] present a nice, concise comparison of the latter two approaches [43,46], along with the 'BlockRank' approach of [49].

The work by Fogaras et al. [36] offers a solution which allows for 'full personalization', i. e., personalization for *any* possible interest profile. This seemingly intractable goal is achieved by providing only an approximate solution, but for a complete set of basis vectors. The latter are generated by performing N random walks starting from each page in the Web graph, and the approximation is due to using only a finite N. The resulting approach requires storage which is linear in the size of the Web graph. Fogaras et al. found good precision for their method on a graph with 80 million nodes, using $N = 1000$.

We see in the above approaches a variety of ways to compute personalized (i. e., topic-sensitive) link analysis scores. There is no lack of ideas, and some good approaches to the scaling problem have been offered. We are not however aware of any commercial activity along these lines.

Another approach to personalizing Web search is to *decentralize* it. This allows for the possibility of local views

of the Web, biased by local interests, and represented by profiles which can act as guides for queries. We discuss this approach – with the scope limited to link analysis – in the next section.

Peer-to-Peer Search and Link Analysis

There are at least two motivations for considering distributed link analysis. One is the need to ameliorate the severe scaling problems – particularly those associated with crawling – that accompany a centralized approach to the vast Web. This could be done for example by spreading the load among Web servers. Another motivation is the desire to build a *peer-to-peer (P2P) search engine*, in which a distributed set of users share the work of crawling, indexing, query routing, and (not least) link analysis. These two scenarios are not necessarily exclusive. Furthermore, in the case of link analysis – and at the level of detail we present in this section – there is little difference between these scenarios; each boils down to seeking to compute global LA scores in a distributed fashion. A short and readable introduction to the world of P2P search may be found in Nejdl [64]; see also Part VI in Steinmetz and Wehrle [80].

Early work by Sankaralingam et al. [76,77] emphasized the need to solve the eigenvector equations (here PageRank) in an asynchronous fashion. This means in turn that convergence of iterative methods, such as the power method, is no longer guaranteed. In Sankaralingam et al. [77] it is pointed out that the PageRank system with its eigengap bounded from below satisfies the necessary and sufficient conditions for asynchronous convergence. The method employed was asynchronous iterated passing of updated weights (scores), with local estimates of convergence. We note that, in this work, it is assumed that all local nodes can find the hosts for every outlink that they store, by using hashed URLs as page IDs, and a structured distributed hash table assigning IDs to nodes. Shi et al. [79] present a similar approach, in which weights are handled differently according to whether they come from "outside" or "inside" the host. They show theoretically that the distributed system converges monotonically, and show experimentally that it converges to the same result as the centralized one. Chirita et al. [28] present a related approach, with the new feature that the distributed PageRank calculation is personalized.

Aberer and Wu [1,86] present a distributed approach, in which *rankings*, rather than scores, are aggregated from the local level to the global level. The local rankings are obtained using PageRank on the local graph. In Wu and Aberer [86], one first computes SiteRank (using the PageRank algorithm on the 'site graph'), and then (based partially on this) computes the local 'DocRank' for each document. The final global ranking for a document is obtained using the ranking aggregation method. In [87], it is shown that the distributed two-level approach converges to the same result as the equivalent centralized version. Wang and DeWitt [84] present a closely related two-level approach. Also, Zhu et al. [90] use a two-level approach, but based on iterative aggregation-disaggregation; they are able to prove convergence for this approach. We note that none of these works address the problem of asynchronous computation.

The 'JXP' method is presented in [69,70,71,72], and [73]. Here the peers do not initially know of any links or pages other than those they themselves host. They then learn about links to their own pages by random meetings with other peers. An iterated PageRank calculation is run, in which no other topology information is retained by any peer; instead, the rest of the Web graph is modeled using a 'world node' W. Monotonic convergence to the global PageRank vector is proven in [70], with the assumption that all peers know the total number of pages. [71] addresses the problem of cheating peers, while [72] show that informed (rather than random) strategies for meetings can enhance the rate of convergence.

Finally, we mention three, rather distinct, recent works. Jelasity et al. [47] present a general analysis of asynchronous distributed eigenvector calculations, including graphs whose spectral radius is larger than 1. They present methods for controlling the length of the iterated vector, and for estimating the principal eigenvalue ρ (when it is not known to be 1). While they do not prove convergence for the $\rho > 1$ case, they present numerical experiments showing good convergence, even in the face of severe noise (message delay and message loss).

Costache et al. [29] present a distributed ranking system for a collaborative work environment, in which users have privacy concerns which lead them to provide only incomplete information to other users. Furthermore, the link analysis (ObjectRank) deviates from PageRank in that links are weighted, based on metadata about the linked nodes. Several approaches are presented for handling partial information exchange. Interestingly, approaches using a "world node" are judged to give the best results in terms of respecting privacy without significant loss of accuracy.

Kempe and McSherry [50] present a decentralized approach for finding the k principal eigenvectors of a weighted symmetric matrix. The approach uses a decentralized form of orthogonal iteration in order to both normalize and orthogonalize the k "working vectors"; and a proof of convergence is given. Regarding application to Web link analysis, we note that PageRank and T-Rank rely

on asymmetric matrices, and also on some form of teleportation, so that this approach may not be used for these types of LA without further work. The HITS matrices are symmetric and weighted, so that the method may be applied directly to these matrices. One would however need a distributed protocol for computing the elements of these matrices – a not insuperable problem.

Thus we see that distributed LA – with main focus on PageRank – has been well studied, and a number of promising approaches have been identified. Perhaps the most interesting questions for this line of work are not rooted in the link analysis itself, but are rather about the future viability of decentralized search scenarios.

Future Directions

In this final section we offer some brief and speculative discussion about future directions. Many of the topics have appeared above; a few have not.

Personalization

We have seen that there exist nice research results pointing the way to personalizing Web search via personalizing link analysis scores. Google offers personalized search (see http://googleblog.blogspot.com/2007/02/personally-speaking.html), but it is not clear whether this personalization uses personalized LA or not. We note further that, whenever a machine is asked to read the mind of a user, there is some danger that the result is as much irritating as satisfying. Hence we offer this question: is there a future for personalized Web search? If so, will it involve personalized link analysis?

P2P

The arguments for peer-to-peer search are similar to those for other peer-to-peer systems: tapping the resources of billions of idling devices (technical), breaking free of centralized controls (political), and empowering the individual (humanistic). One vision is that each peer can represent a living, curious person, who has a personalized index based on her own interests, and a network of related peers based on related interests. In short, there are many motivations offered for the idea of P2P search. These visions have as their counterpart the seemingly innocuous and satisfying situation that users get what they want, instantly, and for free, with current centralized systems such as Google. Thus we find the same question as for personalization: is there a future for these visionary ideas? And, again, we expect that the answer will depend as much on human factors as on technical aspects.

T-Rank

The LA literature is dominated by the "big three". PageRank (as employed by Google) appears to have great commercial success. The HITS approach was developed in the context of the IBM CLEVER project (see http://alme1.almaden.ibm.com/cs/k53/clever.html and Chakrabarty et al. [24]), and stimulated a great deal of further work. Methods which build on the HITS approach are apparently in use [54] by the search engine Teoma (now Ask.com – see http://about.ask.com/en/docs/about/ask_technology.shtml). SALSA (as we have seen) gives results equivalent to scoring by in- or outdegree; the former (link popularity) is used by a number of search engines.

We have seen that placing the big three in a simple table suggests a fourth alternative, termed T-Rank. T-Rank appears to fit the same kind of technical niche as does PageRank – that is, it seems best suited for offline, global importance scoring. T-Rank has only recently appeared, and it is too early to say whether it represents a viable alternative to PageRank. Once the one-level non-normalized approach is considered, however, several new possibilities arise: interpolation – allowing for continuous tuning between full normalization and no normalization, with the resulting tuning of other properties – and novel sink remedies.

Mobile Search

The number of mobile devices is exploding, and has already surpassed the number of PCs. Mobile users also want information; but their priorities tend to be different from those of stationary users (more active, less patient), and the mobile devices themselves present new technical limitations and possibilities (small screen, lower bandwidth, GPS, etc.). How will Web search change as it goes mobile? And most importantly for our own theme, what will happen to link analysis? Will the "mobile Web" be simply a transcoding of today's Web, with hyperlinks intact? Perhaps the kinds of documents mobile users want to see will be so different from today's Web as to amount to a new Web – and one without hyperlinks.

Link Analysis Without Hyperlinks

All of our discussion thus far is based on the assumption that one has a Web graph which is built up from hyperlinks. There are many information domains which have few or no hyperlinks. Also, there is no guarantee that the present, unmanaged, system of collectively generated hyperlinks will persist in a form that makes link analysis useful. We need only consider the explosive growth of

user-generated content in the form of pictures and videos: whether they are placed in the 'visible Web' or not, they are unlikely to ever have the kind of rich hyperlink structure that has made LA work so well on the Web. Thus it is of interest to consider the idea of link analysis in the absence of hyperlinks. The beauty of link analysis is that it places a document in *context* – something that (we believe) is easier for a machine to do than it is to evaluate a document's *content*. Therefore, one should look at methods for placing a document in the context of the entire set of documents (or a chosen subset), by building links between the given documents and analyzing the resulting graph. An obvious choice is to use some kind of distance or similarity measure between pairs of documents; such measures give weighted, symmetric links – in contract to the unweighted, directed hyperlinks of the Web graph. A more elaborate approach, involving merging a Web graph (with hyperlinks) with a structured ontology, may be found in Stoyanovich et al. [81]. Here, the ontology and the Web graph together form a larger directed graph, to which a two-level authority propagation approach is applied; but the ideas here also have application in scenarios where the hyperlinks are few or absent.

Link Analysis Without the Web

Web link analysis has motivated and generated a great deal of understanding of, and insight into, the structure and properties of directed graphs. We believe that there exist possibilities for exporting of the ideas and methods presented here to fields which do not involve Web pages at all, but do involve directed graphs in a significant way. One possibility is the spreading of computer and mobile-phone viruses using address books: the graph on which the spreading takes place is directed, due to the one-way nature of such address lists [51,65]. One example of such cross-fertilization may be found in [22].

Another area in which asymmetric links play a role is in the related fields of social network analysis and innovation diffusion. It is common in these fields to treat links as symmetric, but in many cases this is far from realistic. Once asymmetry is taken into account, then most of the (eigenvector-based) notions of centrality need to be generalized. In such cases, HITS, PageRank, and T-Rank – both in the Forward or 'authority' sense, and in the Backward or 'hub' sense – might offer useful tools for quantifying and understanding the roles of the nodes in these networks.

Bibliography

Primary Literature

1. Aberer K, Wu J (2003) A framework for decentralized ranking in web information retrieval. In: Web technologies and applications. Lecture Notes in computer science, vol 2642. Springer, Berlin, p 594
2. Adali S, Liu S, Magdon-Ismail M (2005) Optimal link bombs are uncoordinated. In: Proceedings of the first International workshop on adversarial information retrieval on the web (AIRWeb) Chiba, May 2005, pp 58–69; available online at http://airweb.cse.lehigh.edu/2005/proceedings.pdf
3. Adamic LA (1999) The small world web. In: Lecture notes in computer science, vol 1696. Proceedings of the third European conference on research and advanced technology for digital libraries. Springer, London, pp 443–452
4. Aktas MS, Nacar MA, Menczer F (2006) Using hyperlink features to personalize web search. In: Advances in web mining and web usage analysis. Lecture notes in computer science, vol 3932. Springer, Berlin
5. Aktas M, Nacar M, Menczer F (2004) Personalizing PageRank based on domain profiles. In: Proceedings of the KDD workshop on web mining and web usage analysis (WebKDD) Seattle, August 22–25 2004. Springer, Berlin
6. Albert R, Jeong H, Barabasi AL (1999) Diameter of World-Wide Web. Nature 410:130–131
7. Altman A, Tennenholtz M (2007) An Axiomatic Approach to Personalized Ranking Systems. In: Veloso MM (ed) Proceedings of the 20th International joint conference on artificial intelligence (IJCAI) Hyderabad, January 6–12 2007. AAAI Press, pp 1187–1192
8. Avrachenkov K, Litvak N, Pham KS (2007) Distribution of PageRank mass among principle components of the web. In: Bonato A, Chung FRK (eds) Proceedings of the 5th International Workshop on algorithms and models for the web-graph (WAW) San Diego, December 11–12 2007. Lecture notes in computer science, vol 4863. Springer
9. Barabasi AL, Albert R (1999) Emergence of Scaling in Random Networks. Science 286:509–512
10. Becchetti L, Castillo C, Donato D, Leonardi S, Baeza-Yates R (2006) Link-based characterization and detection of web spam. In: Proceedings of the second International workshop on adversarial information retrieval on the web (AIRWeb) Seattle, August 10 2006, available online at http://airweb.cse.lehigh.edu/2006/proceedings.pdf
11. Benczúr A, Csalogány K, Sarlós T, Uher M (2005) Spamrank: fully automatic link spam detection. In: Proceedings of the first International workshop on adversarial information retrieval on the web (AIRWeb) Chiba, May 2005, pp 58–69; available online at http://airweb.cse.lehigh.edu/2005/proceedings.pdf, held in conjunction with WWW2005
12. Berberich K, Bedathur SJ, Weikum G, Vazirgiannis M (2007) Comparing apples and oranges: Normalized pagerank for evolving graphs. In: Williamson CL, Zurko ME, Patel-Schneider PF, Shenoy PJ (eds) Proceedings of the 16th International conference on World Wide Web (WWW) Banff, May 8–12 2007. ACM, pp 1145–1146
13. Berman A, Plemmons RJ (1979) Nonnegative Matrices in the Mathematical Sciences. Academic Press, New York
14. Bianchini M, Gori M, Scarselli F (2005) Inside PageRank. ACM Trans Inter Tech 5(1):92–128
15. Bjelland J, Canright GS, Engø-Monsen K (2008) Web Link Analysis: estimating a document's importance from its context. Telektronikk 1:95–113; also in http://delis.upb.de/paper/DELIS-TR-0631.pdf

16. Bjelland J, Burgess M, Canright G, Engø-Monsen K (2008) Eigenvectors of Directed Graphs and Importance Scores: Dominance, T-Rank, and Sink Remedies. DELIS technical report TR-0629, http://delis.upb.de/paper/DELIS-TR-0629.pdf
17. Boldi P, Santini M, Vigna S (2005) PageRank as a function of the damping factor. In: Proc. of the Fourteenth International World Wide Web Conference, Chiba. ACM Press, New York
18. Borodin A, Rosenthal JS, Roberts GO, Tsaparas P (2005) Link Analysis Ranking: Algorithms, Theory and Experiments. ACM Trans Internet Technol (TOIT) 5(1):131–152
19. Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine. In: Proceedings of the seventh international conference on World Wide Web. Comput Netw ISDN Syst 30(1–7):107–117
20. Brinkmeier M (2006) PageRank revisited. ACM TOIT 6(3): 257–279
21. Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata S, Tomkins A, Wiener J (2000) Graph Structure in the web. Comput Netw 33:309–320
22. Canright G, Engø-Monsen K (2007) A pumping model for the spreading of computer viruses. In: Bandara AK, Burgess M (eds) Inter-domain management. Proceedings of the first International conference on autonomous infrastructure, management and security (AIMS) Oslo, June 21–22 2007. Lecture notes in computer science, vol 4543. Springer, pp 133–144
23. Castillo C, Donato D, Gionis A, Murdock V, Silvestri F (2007) Know your neighbors: Web spam detection using the web topology. In: Proceedings of the 30th annual International ACM SIGIR conference on research and development in information retrieval (SIGIR) Amsterdam, July 23–27 2007. ACM, New York, pp 423–430
24. Chakrabarti S, Dom B, Gibson D, Kleinberg J, Kumar SR, Raghavan P, Rajagopalan S, Tomkins A (1999) Hypersearching the web. Sci Am 280(6)
25. Chang H, Cohn D, McCallum A (2000) Creating customized authority lists. In: Langley P (ed) Proceedings of the 17th International conference on machine learning (ICM) Stanford University, Stanford, June 29 – July 2 2000. Morgan Kaufmann, pp 127–134
26. Cheng A, Friedman E (2005) Sybilproof reputation mechanisms. In: Proceeding of the 2005 ACM SIGCOMM Workshop on Economics of Peer-To-Peer Systems. Philadelphia, 22 August. P2PECON '05. ACM, New York, pp 128–132
27. Cheng A, Friedman EJ (2006) Manipulability of PageRank under Sybil strategies. In: Proceedings of the first workshop on the economics of networked systems (NetEcon) Ann Arbor, June 11 2006, available at http://www.cs.duke.edu/nicl/netecon06/papers/proceedings.pdf
28. Chirita P, Nejdl W, Scurtu O (2004) Knowing where to search: Personalized search strategies for peers in P2P networks. SIGIR Workshop on Peer-to-Peer Information Retrieval, Sheffield
29. Costache S, Nejdl W, Paiu R (2007) Personalizing PageRank-Based Ranking over Distributed Collections. Advanced Information Systems Engineering. In: Proceedings of the 19th International Conference, CAiSE 2007, Trondheim, 11–15 June. Springer, Berlin, pp 111–126
30. Davison BD (2000) Recognizing nepotistic links on the Web. In: Artificial Intelligence for Web Search. AAAI Press, Menlo Park, pp 23–28
31. Ding C, He X, Husbands P, Zha H, Simon HD (2002) PageRank, HITS and a unified framework for link analysis. In: Proceedings of the 25th Annual international ACM SIGIR Conference on Research and Development in information Retrieval, Tampere, 11–15 August. SIGIR '02. ACM, New York, pp 353–354
32. Donato D, Leonardi S, Millozzi S, Tsaparas P (2005) Mining the inner structure of the Web graph. In: Proc of the 8th WebDB. J Phys A Math Theor 41:224017
33. Donato D, Leonardi S, Tsaparas P (2005) Stability and Similarity of Link Analysis Ranking Algorithms. ICALP, Lisbon
34. Donato D, Laura L, Leonardi S, Millozzi S (2007) The Web as a graph: How far we are. ACM Trans Internet Technol 7:1
35. Du Y, Shi Y, Zhao X (2007) Using spam farm to boost PageRank. In: Proceedings of the 3rd international Workshop on Adversarial information Retrieval on the Web, Banff, 08 May. AIRWeb '07 vol 215. ACM, New York, pp 29–36
36. Fogaras D, Rácz B, Csalogány K, Sarlós T (2005) Towards Scaling Fully Personalized PageRank: Algorithms, Lower Bounds, and Experiments. Internet Math 2(3):333–358
37. Garfield E (1972) Citation analysis as a tool in journal evaluation. Science 178:471–479
38. Golub GH, Van Loan CF (1996) Matrix Computations, 3rd edn. Johns Hopkins University Press, Baltimore
39. Gulli A, Signorini A (2005) The indexable web is more than 11.5 billion pages. In: WWW 2005 Conference Proceedings, Chiba. ACM Press, New York, pp 902–903
40. Gyöngyi Z, Garcia-Molina H (2005) Link spam alliances. In: Proceedings of the 31st international Conference on Very Large Data Bases, Trondheim, 30 August–02 September. Very Large Data Bases. VLDB Endowment, pp 517–528
41. Gyöngyi Z, Garcia-Molina H (2005) Web spam taxonomy. In: First International Workshop on Adversarial Information Retrieval on the Web, AIRWeb, Chiba
42. Gyöngyi Z, Garcia-Molina H, Pedersen J (2004) Combating web spam with trustrank. In: Proceedings of the Thirtieth International Conference on Very Large Data Bases, VLDB, Toronto, pp 576–587. Morgan Kaufmann Publishers, San Francisco
43. Haveliwala TH (2002) Topic-sensitive PageRank. Proc. of WWW Conf, Honolulu. ACM Press, New York, pp 517–526
44. Haveliwala TH, Kamvar SD (2003) The second eigenvalue of the Google matrix. Stanford University Technical Report 2003-20
45. Haveliwala T, Kamvar S, Jeh G (2003) An analytical comparison of approaches to personalizing PageRank. Technical report 2003-35, Stanford University. http://citeseer.ist.psu.edu/haveliwala03analytical.html
46. Jeh G, Widom J (2003) Scaling personalized web search. In: WWW '03: Proceedings of the 12th international conference on World Wide Web, New York. ACM Press, New York, pp 271–279
47. Jelasity M, Canright G, Engø-Monsen K (2007) Asynchronous distributed power iteration with gossip-based normalization. In: Kermarrec A-M, Bougé L, Priol T (eds) Euro-Par 2007. Lecture Notes in Computer Science, vol 4641. Springer, Berlin, pp 514–525
48. Kamvar SD, Haveliwala TH, Manning CD, Golub GH (2003) Extrapolation methods for accelerating PageRank computations. In: Proc 12th Intl Conf World Wide Web. ACM Press, New York, pp 261–270
49. Kamvar SD, Haveliwala TH, Manning CD, Golub GH (2003) Exploiting the block structure of the web for computing pagerank. Technical Report 2003–17, Stanford University
50. Kempe D, McSherry F (2008) A decentralized algorithm for spectral analysis. J Comput Syst Sci 74(1):70–83

51. Kephart JO, White SR (1991) Directed-graph epidemiological models of computer viruses. In: Proceedings of the IEEE symposium on security and privacy, Oakland, May 1991. IEEE Computer Society Press, Piscataway, pp 343–359
52. Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5):604–632
53. Kleinberg JM, Kumar R, Raghavan P, Rajagopalan S, Tomkins AS (1999) The web as a graph: measurements, models and methods. In: Proceedings of the 19th ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems (PODS) Dallas, May 15–18 2000. ACM, New York, pp 1–10
54. Langville AN, Meyer CD (2004) The Use of Linear Algebra by Web Search Engines. IMAGE Newsl 33:2–6
55. Langville AN, Meyer CD (2005) Deeper Inside PageRank. Internet Math 1(3):335–380
56. Lawrence S, Giles CL (1999) Accessibility of information on the web. Nature 400:107–109
57. Lee HC, Borodin A (2003) Perturbation of the hyper-linked environment. In: Warnow T, Zhu B (eds) Proceedings of the 9th annual International conference on computing and combinatorics (COCOON) Big Sky, July 25–28 2003. Lecture notes in computer science, vol 2697. Springer, pp 272–283
58. Lempel R, Moran S (2000) The stochastic approach for link-structure analysis (SALSA) and the TKC effect. Comput Netw 33(1–6):387–401
59. Lempel R, Moran S (2005) Rank Stability and Rank Similarity of Link-Based Web Ranking Algorithms in Authority Connected Graphs. Inf Retr 8:245–264
60. Levine BN, Shields C, Margolin NB (2006) A Survey of Solutions to the Sybil Attack. Tech report 2006–052, University of Massachusetts Amherst, Amherst
61. Liang C, Ru L, Zhu X, R-SpamRank: A Spam Detection Algorithm Based on Link Analysis. J Comput Inf Syst, to appear
62. Motwani R, Raghavan P (1996) Randomized Algorithms. ACM Press, New York
63. Najork M, Wiener JL (2001) Breadth-first search crawling yields high-quality pages. In: Proc 10th Intl WWW Conf. Elsevier Science, Hong-Kong, pp 114–118
64. Nejdl W (2004) How to Build Google2Google – An (Incomplete) Recipe. The Semantic Web – ISWC 2004, 1–5. Lecture Notes in Computer Science, vol 3298. Springer, Berlin
65. Newman MEJ, Forrest S, Balthrop J (2002) Email networks and the spread of computer viruses. Phys Rev E 66:35–101
66. Ng A, Zheng A, Jordan M (2001) Link analysis, eigenvectors and stability. In: Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence. Seattle, pp 903–910
67. Ng AY, Zheng AX, Jordan MI (2001) Stable algorithms for link analysis. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM Press, New York, pp 258–266
68. Page L, Brin S, Motwani R, Winograd T (1998) The PageRank citation ranking: Bringing order to the web. Technical Report, Stanford University, Stanford
69. Parreira JX, Weikum G (2005) JXP: Global authority scores in a P2P network. In: Doan AH, Neven F, McCann R, Bex GJ (eds) Proceedings of the 8th International workshop on the web and databases (WebDB) Baltimore, collocated with ACM SIGMOD/PODS 2005, June 16–17 2005. ACM Press, New York, pp 31–36
70. Parreira JX, Donato D, Michel S, Weikum G (2006) Efficient and decentralized PageRank approximation in a peer-to-peer web search network. In: Dayal U, Whang K, Lomet D, Alonso G, Lohman G, Kersten M, Cha SK, Kim Y (eds) Proceedings of the 32nd international Conference on Very Large Data Bases, Seoul, 12–15 September. Very Large Data Bases. VLDB Endow, pp 415–426
71. Parreira JX, Donato D, Castillo C, Weikum G (2007) Computing trusted authority scores in peer-to-peer web search networks. In: Proceedings of the 3rd international Workshop on Adversarial information Retrieval on the Web, Banff, 8 May. AIRWeb '07, vol 215. ACM, New York, pp 73–80
72. Parreira JX, Michel S, Weikum G (2007) p2pDating: Real life inspired semantic overlay networks for Web search. Inf Process Manag 43(3):643–664
73. Parreira XP, Castillo C, Donato D, Michel S, Weikum G (2008) The Juxtaposed approximate PageRank method for robust PageRank approximation in a peer-to-peer web search network. VLDB J 17(2). Springer, New York
74. Rafiei D, Mendelzon AO (2000) What is this page known for? Computing Web page reputations. In: Proceedings of the 9th International world wide web conference on computer networks. Int J Comput Telecommun Netw. North-Holland, Amsterdam, pp 823–835
75. Richardson M, Domingos P (2002) The intelligent surfer: Probabilistic combination of link and content information in pagerank. Adv Neural Inf Process Syst 14:1441–1448
76. Sankaralingam K, Sethumadhavan S, Browne JC (2003) Distributed Pagerank for P2P Systems. In: Proceedings of the 12th IEEE International symposium on high performance distributed computing (HPDC) Seattle, June 22–24 2003. IEEE Computer Society, Washington, pp 58–68
77. Sankaralingam K, Yalamanchi M, Sethumadhavan S, Browne JC (2003) Pagerank Computation and Keyword Search on Distributed Systems and P2P Networks. J Grid Comput 1(3): 291–307
78. Serrano MA, Maguitman A, Boguna M, Fortunato S, Vespignani A (2007) Decoding the structure of the WWW: A comparative analysis of web crawls. ACM Trans Web 1(2):10
79. Shi S, Yu J, Yang GW, Wang DX (2003) Distributed Page Ranking in Structured P2P Networks. In: Proceedings of the 32nd International conference on parallel processing (ICPP) Kaohsiung, October 6–9 2003. IEEE Computer Society, Washington, pp 179–186
80. Steinmetz R, Wehrle K (2005) Peer-to-peer Systems and Applications. Springer, Heidelberg
81. Stoyanovich J, Bedathur SJ, Berberich K, Weikum G (2007) EntityAuthority: Semantically enriched graph-based authority propagation. In: Proceedings of the 10th International workshop on the web and databases (WebDB) Beijing, June 15 2007. ACM, New York
82. Sydow M (2004) Random surfer with back step. In: Proceedings of the 13th International world wide web conference on alternate track papers and posters (WWW Alt) New York, May 19–21 2004. ACM Press, New York, pp 352–353
83. Varga RS (1962) Matrix iterative Analysis. Prentice-Hall, Upper Saddle River, USA
84. Wang Y, DeWitt D (2004) Computing pagerank in a distributed internet search system. In: Nascimento MA, Özsu MT, Kossmann D, Miller RJ, Blakeley JA, Schiefer KB (eds) Proceedings of the 30th International conference on very large data bases (VLDB) Toronto, August 31 – September 3 2004. VLDB Endowment, pp 420–431

85. Watts D, Strogatz S (1998) Collective dynamics of 'small world' networks. Nature 393:440–442
86. Wu J, Aberer K (2004) Using SiteRank for decentralized computation of Web document ranking. Adaptive Hypermedia and adaptive Web-based Systems. LNCS, vol 3137. Springer, Berlin, pp 265–274
87. Wu J, Aberer K (2005) Using a Layered Markov Model for Distributed Web Ranking Computation. In: Proceedings of the 25th IEEE International Conference on Distributed Computing Systems. ICDCS 10(10):533–542
88. Yahoo.com (2008) Search query 'car' at http://www.yahoo.com. Accessed 13 Feb 2008
89. Zhang H, Goel A, Govindan R, Mason K, van Roy B (2004) Making eigenvector-based reputation systems robust to collusion. In: Proceedings of the third Workshop on Web Graphs (WAW). Lecture Notes in Computer Science, vol 3243. Springer, Rome, pp 92–104
90. Zhu Y, Ye S, Li X (2005) Distributed pagerank computation based on iterative aggregation-disaggregation methods. In: Proceedings of the 14th ACM International conference on information and knowledge management (CIKM) Bremen, October 31 – November 5 2005. ACM Press, New York, pp 578–585

Books and Reviews

Bianchini M, Gori M, Scarselli F (2005) Inside PageRank. ACM Trans Inter Tech 5(1):92–128
Chakrabarti S (2002) Mining the Web: Analysis of Hypertext and Semi Structured Data. Morgan Kaufmann, San Francisco
Langville AN, Meyer CD (2005) Deeper Inside PageRank. Internet Math 1(3):335–380
Langville AN, Meyer CD (2006) Google's PageRank and Beyond: The Science of Search Engine Rankings. Princeton University Press, Princeton
Thelwall M (2004) Link Analysis: An Information Science Approach. Academic Press, Amsterdam

Logic and Geometry of Agents in Agent-Based Modeling

Samson Abramsky
Oxford University Computing Laboratory, Oxford, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Towards a Logic and Geometry of Interaction
Further Directions
Bibliography

Glossary

Agent A computational or biological entity which can perform actions which affect its environment, and observe actions performed by the environment or other agents.

Agent-based modeling Modeling a complex system in terms of multiple agents interacting with each other.

Interaction A pattern of actions in a multi-agent system. Each action is performed by some agent, and may be observed by others.

Information flow The process whereby information held by one agent or part of a system is transferred to another, possibly in some modified form. Generally mediated by interaction: information flow is caused by agents performing actions, and other agents observing those actions.

Compositionality Describing a complex system or object in a structured fashion, as built up by applying certain operations hierarchically, starting from a stock of basic types of system. Compositional definitions of functions of systems are those which respect this hierarchical structure.

Linear logic A sub-structural logic in which the operations of copying and deleting premises are not allowed in general.

Combinators Basic operations in a function algebra. Combinatory logic is a variable-free formulation of functional computation.

Definition of the Subject

Agent-based Modeling has become of increasing importance in Computer Science, and also in mathematical modeling and simulation. The idea is that the behavior of a complex system can be described as arising from the interaction of multiple agents, with each other and with the environment, using simple local rules. It is widely recognized that building a sound and widely applicable theory for such systems will require an inter-disciplinary approach, and the development of new mathematical and computational concepts.

In this article, agents and interaction will be studied from the perspective of Logic and Computer Science. It will be shown how ideas about Logical Dynamics, Games and Geometry of Interaction, which have been developed over the past two decades, lead towards a structural theory of agents and interaction. This provides a basis for powerful logical methods such as compositionality, types and high-level calculi, which have proved so fruitful in Computer Science, to be applied in this domain.

This approach should be contrasted with the more familiar approaches to agent-based modeling using primitive agents such as cellular automata. The main focus of such approaches is systems modeling, where the agents are essentially discrete counterparts to traditional PDE or ODE models of dynamical systems. Ultimately, these ap-

proaches should be seen as complementary, and the goal is to combine them. In the present article, the emphasis is on presenting some basic ideas of the novel approach to complex systems using logical and compositional methods.

Introduction

In this article, a logical and geometric perspective on the information flow arising from agent interaction will be described. This is a novel and distinctive approach within the emerging field of agent-based modeling. While it is still in its formative stages, there are already some key structural insights which it offers which will be fundamental to any deep and comprehensive theory of agents.

Firstly, a fundamental methodological point will be discussed, which has played a crucial rôle in Computer Science for several decades, but has yet to achieve the recognition in general scientific modeling which it deserves: the importance of **compositionality**.

Compositionality

A methodological principle from Computer Science (and Logic) of **major** potential importance for mathematical modeling throughout the sciences.

- **Traditional approach**: Whole-system (monolithic) analysis of given systems. A key rôle is played by structuring templates, e. g. "Find the Hamiltonian".
- **Compositional approach**: Start with a fixed set of basic (simple) building blocks, and **constructions** for building new (in general more complex) systems out of given sub-systems, and build up the required complex system with these.

More formally, compositionality can be expressed **algebraically**:

$$S = \omega(S_1, \ldots, S_n) .$$

The system S is described as being built up from sub-systems $S_1, \ldots, S_n$ by the operation ω. There is also a **logical** perspective:

$$\frac{S_1 \models \phi_1, \ldots, S_n \models \phi_n}{\omega(S_1, \ldots, S_n) \models \phi}$$

(Read $S \models \phi$ as "system S satisfies the property ϕ"). Here **properties** ϕ of the compound system S can be inferred by verifying properties $\phi_1, \ldots, \phi_n$ for the simpler sub-systems $S_1, \ldots, S_n$.

Some Key Points

- The compositional view of complex systems sees them as built up, not just by one-level composition of basic agents: $S = \|_{i \in I}\, a_i$, which is the usual scenario in current agent-based modeling and simulation, but **hierarchically**:

  $$S = \omega_1(\omega_2(a_1, a_2, a_3), \omega_1(a_4, a_5)) .$$

  It tracks the properties of the sub-compound systems all the way up (or down) the tree of syntax.
- The **available repertoire** of system constructors is also an important aspect of the modeling here, leading to questions of **expressiveness** and **functional completeness**.
- This paradigm has played a major role in Computer Science over the past four decades, and is already starting to be applied in
  - quantum computing (quantum programming languages)
  - biological modeling (process calculi)
  - business modeling (idem)

  and will surely be applied more widely and deeply, in economics as well as physical and biological sciences.
- This kind of modeling carries in its train a range of powerful **analytical techniques**: types, semantics, verification, model-checking, etc.

Computation as Interaction

The second key point to be emphasized is the emergence of **interaction** as a key notion in computation, and increasingly in a wide range of scientific fields, whether in physics (quantum information), biology (at a range of scales and levels, from molecular interactions to evolutionary theory), economics (game theory), linguistics (dialogical analysis), etc. Moreover, these developments are feeding back into logic and philosophy.

Note that interaction happens **between agents**—a theory of interaction amounts to an **agent dynamics**.

Changing Views of Computation The scene can be set by recalling how perspectives on computation have changed since the first computers appeared. The early practice of computing can be pictured as in Fig. 1. This is the era of stand-alone machines and programs: computers are served by an elite priesthood, and have only a narrow input-output interface with the rest of the world.

First-Generation Models of Computation These models live on the existing intellectual inheritance from discrete mathematics and logic. **Time** and **processes** lurk in the background, but are largely suppressed.

Given this limited vision of computing, there is a very natural abstraction of computation, in which programs are

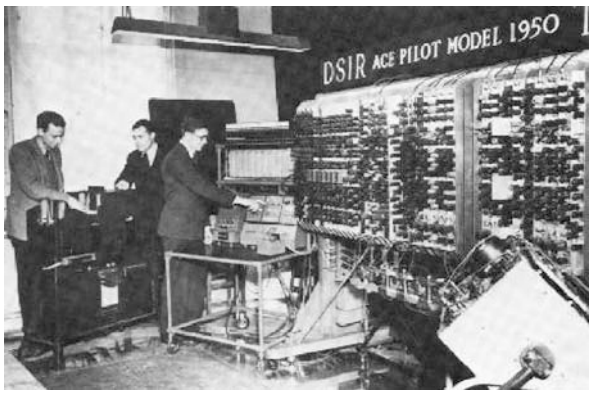

Logic and Geometry of Agents in Agent-Based Modeling, Figure 1
Computing "in the isolation ward"

seen as computing **functions** or **relations** from inputs to outputs.

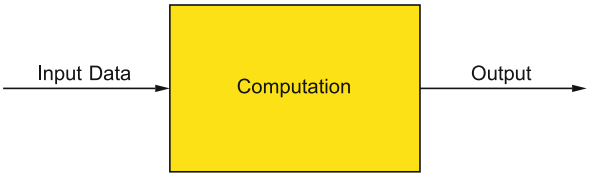

Computation in the Age of the Internet Instead of isolated systems, with rudimentary interactions with their environment, the standard unit of description or design becomes a **process** or **agent**, the essence of whose behavior is **how it interacts** with its environment.

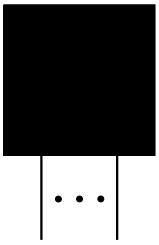

The interaction between the System and the Environment can be conceptualized as a **two-person game**. A program specifying how the System should behave in the face of all possible actions by the Environment is then a **strategy** for the player corresponding to the System.

Interaction Complex behavior arises as the global effect of a **system** of **interacting agents** (or processes).

The key building block is the agent. The key operation is **interaction**—plugging agents together so that they interact with each other.

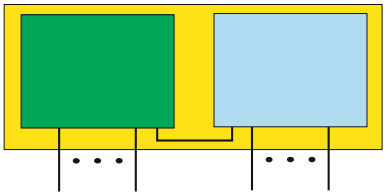

Who is the System? Who is the Environment? This **symmetry** between System and Environment carries a first clue that there is some structure here; it will lead to a key **duality**, and a deep connection to logic.

Towards a Compositional Approach to Complex Systems This conceptual model works at all "scales" :

- Macro-scale: processes in operating systems, software agents on the Internet, transactions.
- Micro-scale: how programs are implemented (subroutine call-return protocols, register transfer) all the way down into hardware.

It is applicable both to **design** (synthesis) and to **description** (analysis); to **artificial** and to **natural** information-processing systems.

There are of course large issues lurking here, e. g. in the realm of "Complex Systems": **emergent behavior** and even **intelligence**. Is is helpful, or even feasible, to understand this complexity **compositionally**? New conceptual tools, new theories, are needed to help us analyze and synthesize these systems, to help us to **understand** and to **build**.

Towards a Logic and Geometry of Interaction

Towards a "Logic of Interaction"

Specifying and reasoning about the behavior of computer programs takes us into the realm of logic. For the first-generation models, logic could be taken "as it was"—static and timeless. For the second-generation models, getting an adequate account—a genuine "logic of interaction"—may require a fundamental reconceptualization of logic itself. This radical revision of the view of logic is happening anyway—prompted partly by the applications, and partly by ideas arising within logic.

The Static Conception of Logic The usual "static" notion of tautology is as "a statement which is vacuously true because it is compatible with all states of affairs"

$$A \vee \neg A \,.$$

"It is raining **or** it is not raining"—truth-functional semantics. This is illustrated (subversively) in Fig. 2. But what could a **dynamic notion of tautology** look like?

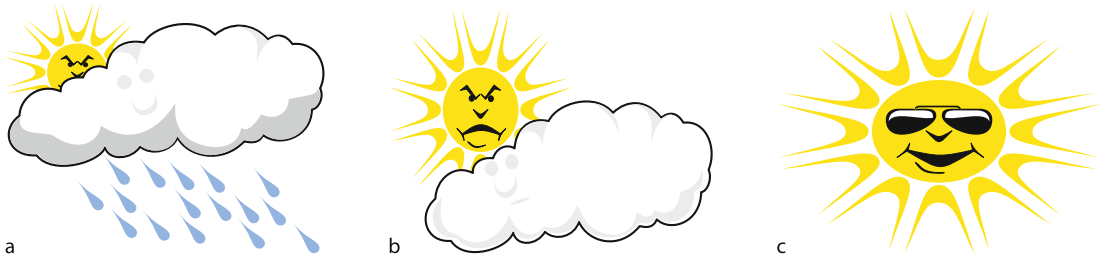

Logic and Geometry of Agents in Agent-Based Modeling, Figure 2
Tertium non datur?

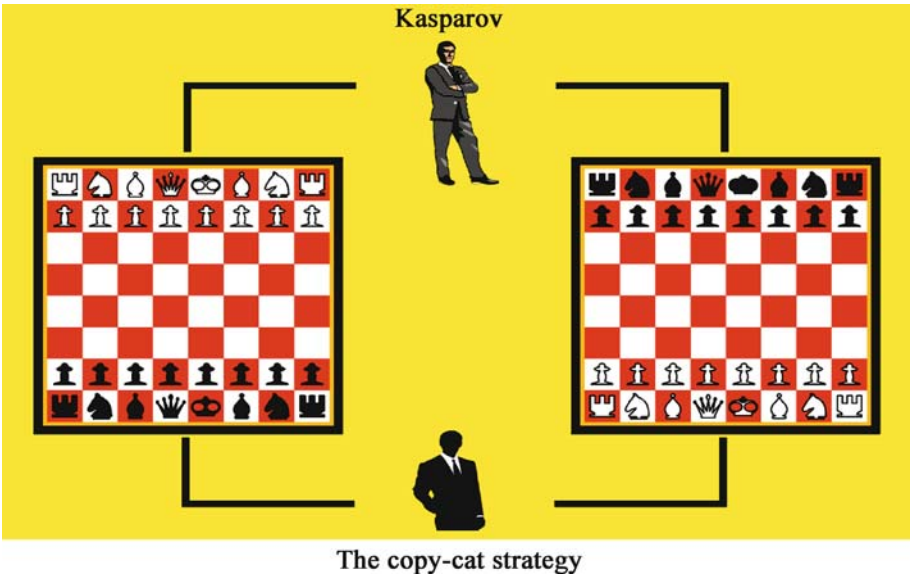

Logic and Geometry of Agents in Agent-Based Modeling, Figure 3
How to beat a Grandmaster

The Copy-Cat Strategy Consider the following little fable, illustrated by Fig. 3:

> How to beat an International Chess Grandmaster by the power of pure logic

The idea is to rely on logic, rather than on any talent at Chess. We arrange to play two games of Chess with the grandmaster, say Gary Kasparov, once as White and once as Black. Moreover, we so arrange matters that we start with the game in which we play as Black. Kasparov makes his opening move; we respond by playing the **same** move in the **other** game—this makes sense, since we are playing as White there. Now Kasparov responds (as Black) to our move in that game; and we copy that response back in the first game. We simply proceed in this fashion, copying the moves that our opponent makes in one board to the other board. The net effect is that **we play the same game twice—once as White, and once as Black**. (We have essentially made Kasparov play against himself). Thus, whoever wins that game, we can claim a win in one of our games against Kasparov! (Even if the game results in a stalemate, we have done as well as Kasparov over the two games—surely still a good result!) Of course, this idea has nothing particularly to do with Chess. It can be applied to any two-person game of a very general form. The use of Chess-boards to illustrate the discussion will continue, but this underlying generality should be kept in mind.

What are the salient features which can be extracted from this example?

A dynamic tautology There is a sense (which will shortly be made more precise) in which the copy-cat strategy can be seen as a **dynamic version** of the tautology $A \vee \neg A$. Note, indeed, that an essential condition for being able to play the copy-cat is that the rôles of the two players are inter-changed on one board as compared to the other. Note also the disjunctive quality of the argument that we must win in one or other of the two games. But the copy-cat strategy is a **dynamic**

process: a two-way channel which maintains the correlation between the plays in the two games.

Conservation of information flow The copy-cat strategy does not **create** any information; it reacts to the environment in such a way that information is conserved. It ensures that exactly the same information flows out to the environment as flows in from it. Thus one gets a sense of logic appearing in the form of **conservation laws for information dynamics**.

The power of copying Another theme which appears here, and of which more will be seen later, concerns the surprising power of simple processes of copying information from one place to another. Indeed, as shall eventually be seen, such processes are **computationally universal**.

The geometry of information flow From a dynamical point of view, the copy-cat strategy realizes a channel between the two game boards, by performing the **actions** of copying moves. But there is also some implicit **geometry** here. Indeed, the very idea of two boards laid out side by side appeals to some basic underlying spatial structure. In these terms, the copy-cat channel can also be understood geometrically, as creating a graphical link between these two spatial locations. These two points of view are complementary, and link the logical perspective to powerful ideas arising in modern geometry and mathematical physics.

Further evidence that the copy-cat strategy embodies more substantial ideas than might at first be apparent, can be obtained by varying the scenario. Consider now the case where we play against Kasparov on **three boards**; one as Black, two as White.

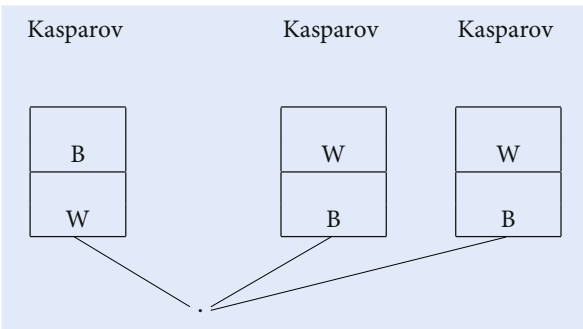

Does the Copy-Cat strategy still work here? In fact, it can easily be seen that it does **not**. Suppose Kasparov makes an opening move m_1 in the left-hand board where he plays as White; we copy it to the board where we play as White; he responds with m_2; and we copy m_2 back to the board where Kasparov opened. So far, all has proceeded as in our original scenario. But now Kasparov has the option of playing a **different** opening move, m_3 say, in the rightmost board. We have no idea how to respond to this move; nor can we copy it anywhere, since the board where we play as White is already "in use". This shows that these simple ideas already lead us naturally to the setting of a **resource-sensitive** logic, in which in particular the Contraction Rule, which can be expressed as $A \to A \wedge A$ (or equivalently as $\neg A \vee (A \wedge A)$) cannot be assumed to be valid.

What about the other obvious variation, where we play on two boards as White, and one as Black?

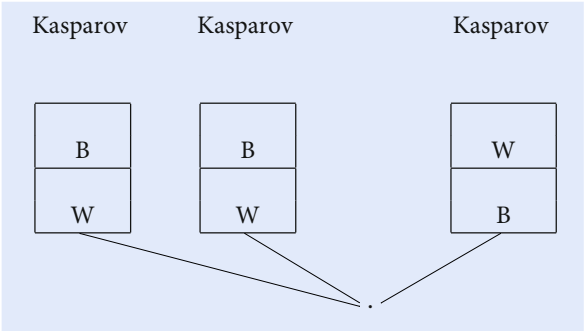

It seems that the copy-cat strategy **does** still work here, since we can simply ignore one of the boards where we play as White. However, a geometrical property of the original copy-cat strategy has been lost, namely a **connectedness** property, that information flows to every part of the system. This at least calls the corresponding logical principle of Weakening, which can be expressed as $A \wedge A \to A$, (or equivalently as $\neg A \vee \neg A \vee A$) into question.

These remarks indicate that we are close to the realm of Linear Logic and its variants; and, mathematically, to the world of monoidal (rather than cartesian) categories.

Game Semantics These ideas find formal expression in **Game Semantics**. Games play the role of:

- Interface types for computation modules
- Propositions with dynamic content.

In particular, 2-person games capture the duality of:

- Player vs. Opponent
- System vs. Environment.

Agents are Strategies In this setting, agents or processes can be modeled as **strategies** for playing the game. These strategies **interact** by playing against each other. A notion of correctness is obtained which is **logical** in character in terms of the idea of **winning** strategy—one which is guaranteed to reach a successful outcome however the environment behaves. This in a sense replaces (or better, **refines**) the logical notion of "truth": winning strategies are the dynamic version of tautologies (more accurately, of their **proofs**).

Building Complex Systems by Combining Games It will now be seen how games can be combined to produce more complex behaviors while retaining control over the interface. This provides a basis for the **compositional** understanding of systems of interacting agents—understanding the behavior of a complex system in terms of the behavior of its parts. This is crucial for both analysis and synthesis, i. e. for both description and design. These operations for building games can be seen as (dynamic forms of) "type constructors" or "logical connectives". (The underlying logic here will in fact be Linear Logic).

Duality—"Linear Negation" $A^{\perp}$—interchange rôles of Player and Opponent (reflecting the symmetry of interaction).

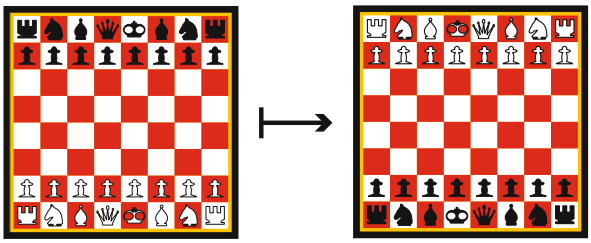

Note that, with this interpretation, negation is involutive:

$$A^{\perp\perp} = A\,.$$

Tensor—"Linear Conjunction"

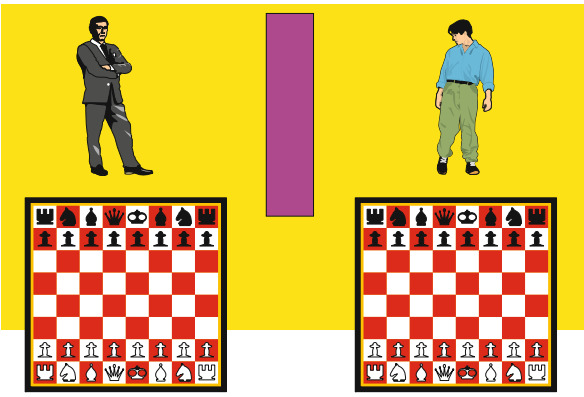

$A \otimes B$

The idea here is to combine the two game boards into one system, **without any information flow between the two sub-systems**. (This is the significance of the "wall" separating the two players, who will be referred to as Gary (Kasparov) and Nigel (Short)). This connective has a conjunctive quality, since we must independently be able to play (and to win) in each conjunct. Note however, that there is no constraint on information flow for the environment, as it plays against this compound system.

Par—"Linear Disjunction"

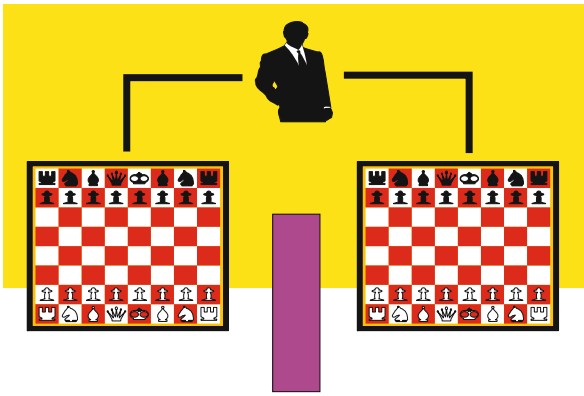

$A \parr B$

In this case, there are two boards, but one player (who shall be referred to as the Copy-Cat), indicating that we **do** allow information flow for this player between the two game boards. This for example allows information revealed in one game board by the Opponent to be used against him on the other game board—as exemplified by the copy-cat strategy. However, note that the wall appears on the environment's side now, indicating that the environment is constrained to play separately on the two boards, with no communication between them.

Thus there is a De Morgan duality between these two connectives, mediated by the Linear negation:

$$(A \otimes B)^{\perp} = A^{\perp} \parr B^{\perp}\,, \qquad (A \parr B)^{\perp} = A^{\perp} \otimes B^{\perp}\,.$$

The idea is that on one side of the mirror of duality (Player/System for the Tensor, Opponent/Environment for the Par), there is the constraint of no information flow, while on the other side, there is information flow.

The Copy-Cat strategy can now be reconstrued in logical terms:

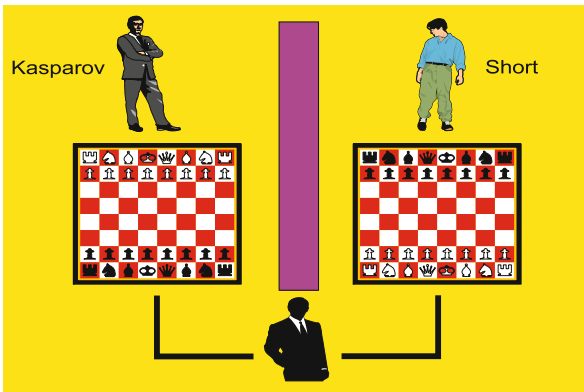

It can be seen that it is indeed a winning strategy for $A^{\perp} \parr A$. Moreover, $A \multimap B$ ("Linear implication") can be defined by

$$A \multimap B \equiv A^{\perp} \parr B\,,$$

(*cf.* $A \supset B \equiv \neg A \vee B$.) Then the copy-cat strategy becomes the canonical proof of the most basic tautology of all: $A \multimap A$.

The information flow possibilities of Par receive a more familiar logical interpretation in terms of the Linear implication; namely, that information about the antecedent can be used in proving the consequent (and conversely with respect to their negations, if proof by contraposition is considered).

Thus an entire "linearized" logical structure opens up before us, with a natural interpretation in terms of the dynamics of information flow.

Interaction A key step in the development now arises: the modeling of **interaction** itself. Constructors create "potentials" for interaction; the operation of plugging modules together so that they can communicate with each other **releases** this potential into **actual computation**. Consider the diagram shown in Fig. 4. Here two separate sub-systems are shown, each with a compound structure, expressed by the **logical types of their interfaces**. What these types tell us is that these systems are **composable**; in particular, the **output type** of the first system, namely B, matches the input type of the second system. Note that this "logical plug-compatibility" makes essential use of the duality, just as the copy-cat strategy did. What makes Gary (the player for the first system) a fit partner for interaction with Nigel (the player for the second system), is that they have **complementary views** of their locus of interaction, namely B. Gary will play in this type "positively", as Player (he sees it as B), while Nigel will play "negatively", as Opponent (he sees it as $B^{\perp}$). Thus each will become part of the environment of the other—part of the potential environment of each will be realized by the other, and hence part of the **potential** behavior of each will become **actual** interaction.

This leads to a dynamical interpretation of the fundamental operation of **composition**, in mathematical terms:

$$\frac{A \xrightarrow{\text{Gary}} B \xrightarrow{\text{Nigel}} C}{A \xrightarrow{\text{Gary;Nigel}} C}$$

or of the **Cut rule**, in logical terms:

$$\text{Cut:}\quad \frac{\vdash \Gamma, A \vdash A^{\perp}, \Delta}{\Gamma, \Delta}$$

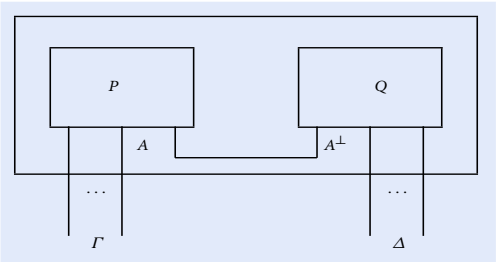

Composition as Interaction

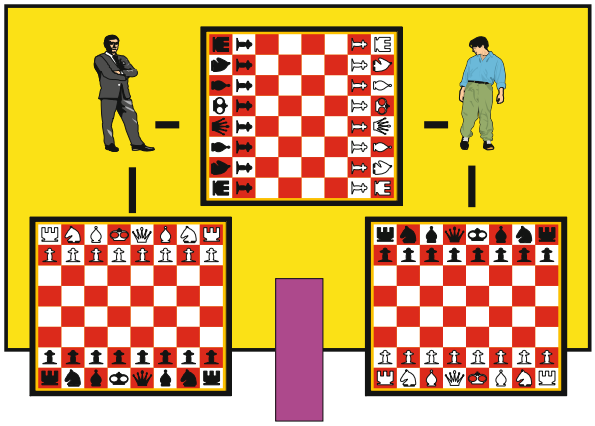

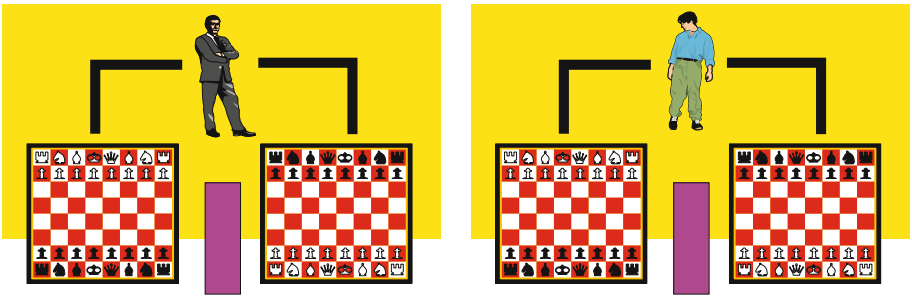

$A^{\perp} \parr B \equiv A \multimap B\,,\quad B^{\perp} \parr C \equiv B \multimap C\,.$

Logic and Geometry of Agents in Agent-Based Modeling, Figure 4
Two composable systems

The picture here shows the new system formed by plugging together the two sub-systems. The "external interface" to the environment now shows just the left hand board A as input, and the right hand board C as output. The Cut formula B is hidden from the environment, and becomes the locus of interaction inside the black box of the system. Suppose that the Environment makes some move m in C. This is visible only to Nigel, who as a strategy for $B \multimap C$ has a response. Suppose this response m_1 is in B. This is a move by Nigel as Player in $B^{\perp}$, hence appears to Gary as a move by Opponent in B. Gary as a strategy for $A \multimap B$ has a response m_2 to this move. If this response is again in B, Nigel sees it as a response by the environment to his move, and will have a response again; and so on. Thus there is a sequence of moves $m_1, \ldots, m_k$ in B, ping-ponging back and forth between Nigel and Gary. If, eventually, Nigel responds to Gary's last move by playing in C, or Gary responds to Nigel's last move by playing in A, then this provides the response of the **composed strategy** Gary; Nigel to the original move m. Indeed, all that is visible to the Environment is that it played m, and eventually some response appeared, in A or C.

Moreover, if both Nigel and Gary are winning strategies, then so is the composed strategy; and the composed strategy will not get stuck forever in the internal ping-pong in B. To see this, suppose for a contradiction that it did in fact get stuck in B. Then there would be an infinite play in B following the winning strategy Gary for **Player** in B, and the **same** infinite play following the winning strategy Nigel for Player in $B^{\perp}$, hence for **Opponent** in B. Hence the same play would count as a win for both Player and Opponent. This yields the desired contradiction.

Discussion Game Semantics in the sense discussed in this section has had an extensive development over the past decade and a half, with a wealth of applications to the semantics of programming languages, type theories and logics [9,11,12,13,14,16,31]. More recently, there has been an algorithmic turn, and some striking applications to verification and program analysis [2,17,24,38].

From the point of view of the general analysis of Information, there are the following promising lines of development:

- Game semantics provides a promising arena for exploring the combination of quantitative and qualitative theories of information. In particular, it provides a setting for quantifying information flow between agents. Important quantitative questions can be asked about **rate of information flow** through a strategy (representing a program, or a proof); how can a system gain **maximum** information from its environment while providing **minimal** information in return; robustness in the presence of **noise**, etc.
- As in the discussion of the copy-cat strategy, there is an intuition of logical principles arising as **conservation laws for information flow**. (And indeed, in the case of Multiplicative Linear Logic, the proofs correspond exactly to "generalized copy-cat strategies"). Can this intuition be developed into a fully-fledged theory? Can logical principles be **characterized** as those expressing the conservation principles of this information flow dynamics?
- There is also the hope that the more structured setting of game semantics will usefully constrain the exuberant variety of possibilities offered by process algebra, and allow a sharper exploration of the logical space of possibilities for information dynamics. This has already been borne out in part, by the success of game semantics in exploring the space of programming language semantics. It has been possible to give crisp characterizations of the "shapes" of computations carried out within certain **programming disciplines**: including purely functional programming [16,31], stateful programming [11,12], general references [15], programming with non-local jumps and exceptions [35,36], non-determinism [29], probability [21], concurrency [25,26], names [18], polymorphism [10,30] and more. See [13] for an overview (now rather out of date). There has also been a parallel line of development of giving **full completeness** results for a range of logics and type theories, characterizing the "space of proofs" for a logic in terms of informatic or geometric constraints which pick out those processes which are proofs for that logic [9,14,19,20,22,37]. This allows a new look at such issues as the boundaries between classical and constructive logic, or the fine structure of polymorphism and second-order quantification.
- This also gives some grounds for optimism that what **computational processes** are can be captured—in a "machine-independent", and moreover "geometrical", non-inductive way—, **without** referring back to Turing machines or any other explicit machine model.
- In the same spirit as for computability, can **polynomial-time computation** and other complexity classes be characterized in such terms?

Emergent Logic: The Geometry of Information Flow

The geometric structure inherent in interaction and information flow, and its connections to the dynamic view of interaction, will now be considered. Recent work has shown how common structure arises in a variety of contexts: in **logic** (geometric representations of proofs);

geometry (diagram algebras, especially the Temperley–Lieb algebra, with connections to knot theory and geometric topology); **computation** (especially functional computation); and **quantum mechanics** (quantum information protocols, exploiting the information flow inherent in quantum entanglement).

A rapid tour will be made through these disparate areas, focusing on how common structure arises. The exposition will rely extensively on **graphical calculi** which provide an intuitive and visually appealing window onto the various formalisms to be encountered. These calculi have a substantial mathematical content, founded on the diagrammatic approach to tensor categories; further details can be found in the references.

Logic Firstly Multiplicative Linear Logic [27], the logic of the linear connectives $\otimes$, $\parr$, $(_)^\perp$ which have already been encountered, will be considered, as a basic paradigmatic example. A key insight [27] is that the essential information in a proof in this system is given by a pairwise matching of the occurrences of positive and negative literals in the sequent—a **proof structure**. E.g., the two possible proof structures for the sequent $\alpha^\perp \parr \alpha^\perp, \alpha \otimes \alpha$ are:

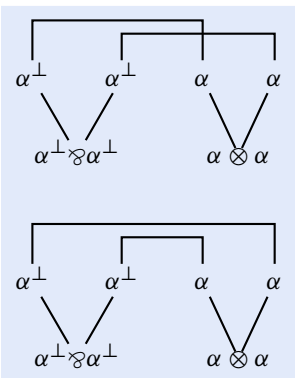

A **geometric** view of proof structures, as certain graphs obtained from the forest of formula trees in the sequent by drawing arcs—"axiom links"—between the paired literal occurrences. Alternatively, a more **dynamic** view can be taken, and proof structures can be represented as **involutions** on the set of literal occurrences in the sequent. These must be **literal preserving**, i. e. an occurrence of a literal l must be mapped to an occurrence of its dual $l^\perp$. Such functions represent a flow of information around the sequent. They can be viewed as **copy-cat strategies**.

Every sequent proof determines a proof structure. The fundamental question is: **which proof structures arise from sequent proofs**? A first answer is **geometric** (or topological) in character. For each proof structure, a **switching graph** can be obtained by deleting, for each occurrence of a subformula $A \parr B$, exactly one of the arcs $A - A \parr B - B$ connecting it to its immediate subformulas. If all such switching graphs are trees, the proof structure is said to be a **proof net**.

A second form of characterization is **interactive**. An orthogonality relation can be defined, between permutations on the set of occurrences of literals in the sequent $\Gamma : f \perp g$ if fg is cyclic. The idea is that f is a candidate proof net, while g is an attempted counter-proof—a passage through the literals induced by a choice of switching graph. Note that the alternating composition of f and g expresses **interaction** between f and g, thought of as **strategy** and **counter-strategy**. It generates a path along which information flows around the system.

A semantics of MLL proofs can be given by specifying, for each formula A, a set S of permutations on the set of literal occurrences $|A|$, such that $S = S^{\perp\perp}$, where $S^\perp = \{g \mid \forall f \in S.\ f \perp g\}$. For a literal, the unique permutation (the identity) is specified.

$$S(A \otimes B) = \{f + g \mid f \in S(A) \wedge g \in S(B)\}^{\perp\perp}$$
$$S(A \parr B) = S(A^\perp \otimes B^\perp)^\perp .$$

Here $f + g$ is disjoint union of permutations, expressing the absence of information flow (or **information independence**).

Theorem 1 (Sequentializability [27] and Full Completeness [9]) *Let f be a literal-respecting involution on $|\Gamma|$. The following are equivalent: (i) f is the permutation assigned to a sequent proof of Γ; (ii) f is a proof net; (iii) $f \in S(\Gamma)$.*

This shows that **the geometric and interactive characterizations of the space of proofs coincide**.

Further Developments From a computational perspective, the multiplicative connectives embody **concurrency** and **causal independence**. The scope of the enterprise is greatly expanded when the other **levels of connective** in Linear Logic are incorporated:

Additives: The additive conjunction and disjunction allow **causality**, **conflict** and **case analysis/conditionals** to be expressed. The interaction between the additive and multiplicative levels is rather subtle. The theory sketched above for the multiplicatives has in large part been lifted to Multiplicative Additive Linear Logic [14], but a number of key questions and issues needed for a deeper analysis remain to be investigated.

Exponentials: The multiplicative fragment only allows linear time computation to be expressed (under the Curry–Howard paradigm). For a full analysis of com-

putationally expressive systems, it is necessary to allow for **copying** and **deleting**, as regulated by the **exponential connectives** of Linear Logic. Existing results have extended the multiplicative theory to various systems of typed λ-calculus (corresponding to various forms of Intuitionistic Multiplicative Exponential Linear Logic), and have begun to investigate systems which, by constraining the exponential types in certain natural ways, capture significant **complexity classes**, especially **PTIME**.

Diagram Algebras It will now be indicated how this apparently very specialized corner of Proof Theory in fact connects directly to a broad topic arising in Representation Theory and Knot Theory, with connections to Mathematical Physics. On the one hand, some structure will be **lost**, by obliterating the distinction between $\otimes$ and $\parr$; this corresponds to moving from **$*$-autonomous** to **compact closed** categories. This means that the formula tree structure can be dispensed with altogether; it is simply a matter of connecting up literal occurrences, which shall be drawn as "joining up the dots". Motivation: compact closed categories show up in many contexts of interest!

On the other hand, rather than one-sided sequents, general arrows or two-sided sequents will be represented diagrammatically. This means arrows can be represented as

$$A_1 \otimes \cdots \otimes A_n \longrightarrow B_1 \otimes \cdots \otimes B_m \ ,$$

where each A_i and B_j is a literal, by involutions on $\{1, \ldots, n\} + \{1, \ldots, m\}$, which are literal-preserving in the extended sense that **opposite** literals are connected in the domain or in the codomain, while occurrences of the **same** literal in the domain and the codomain are connected . An advantage of this representation is that composition is expressed very transparently, by "stacking" arrows.

Example The composition of

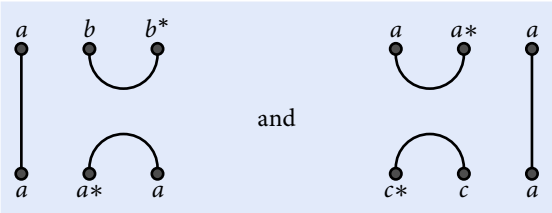

is given by

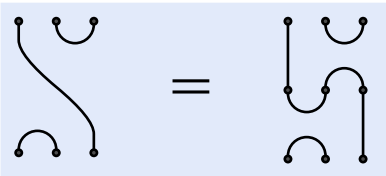

Temperley–Lieb Algebra The **Temperley–Lieb algebra** played a central role in the **Jones polynomial invariant of knots** [32] and ensuing developments. It was originally presented, rather forbiddingly, in terms of abstract generators and relations [32]. It was recast in beautifully elementary and conceptual terms by Louis Kauffman as a **planar diagram algebra** [33].

Generators:

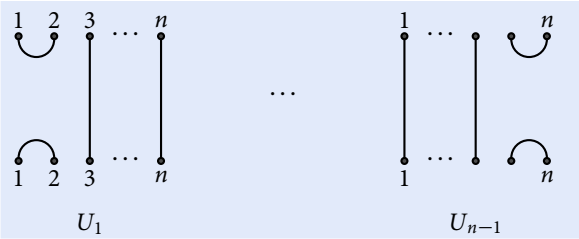

Relations:

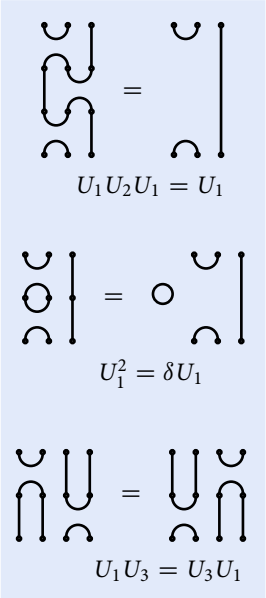

The general form of an element of the algebra (actually of the basic multiplicative monoid: the algebra is then constructed freely over this as the "monoid algebra") is obtained by "joining up the dots" in a **planar** fashion. Multiplication xy is defined by identifying the bottom row of x with the top row of y. In general loops may be formed—these are "scalars", which can float freely across these figures, represented symbolically by δ above.

How does this connect to knots? A key conceptual insight is due to Kauffman, who saw how to recast the Jones polynomial in elementary combinatorial form in terms of his **bracket polynomial**. The basic idea of the bracket polynomial is expressed by the following equation:

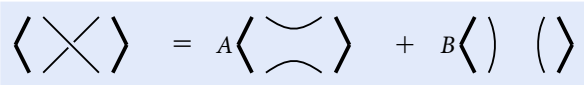

Each over-crossing in a knot or link is evaluated to a weighted sum of the two possible planar smoothings. With suitable choices for the coefficients A and B (as Laurent polynomials), this is invariant under the second and third Reidemeister moves. With an ingenious choice of normalizing factor, it becomes invariant under the first Reidemeister move—and yields the Jones polynomial! What this means algebraically is that the braid group has a representation in the Temperley–Lieb algebra—the above bracket evaluation shows how the basic generators of the braid group are mapped into the Temperley–Lieb algebra. Every knot arises as the closure of a braid; the invariant arises by mapping the open braid into the Temperley–Lieb algebra, and taking the trace there.

Moreover, it turns out that this connection can itself carry interesting information between the Computer Science ideas and the geometry and algebra. Indeed, using Computer Science methods it is possible to give the first **direct presentation** (no quotients) of the Temperley–Lieb algebra, using logical methods. In fact, the elements of the Temperley–Lieb algebra are completely determined by the relations they induce on the "dots"; and **planarity** can be characterized using only the ordering relations on the two rows of dots. Moreover, the multiplication of the algebra can be described as a form of Cut Elimination, using the methods developed in the "Geometry of Interaction" [1, 8,28]. This exactly corresponds to characterizing the **geometric composition of diagrams** as above in terms of an **information flow dynamics**, just as in the case of Multiplicative Linear Logic. This also shows that **planarity is an invariant of Cut Elimination**, and raises interesting questions about computational expressiveness under topological constraints such as planarity, and the computational significance of braiding.

Applicative Computation This ideas in turn apply directly to **applicative computation**, offering the same combination of geometric/diagrammatic and information-flow tools of analysis. This is illustrated firstly with a **planar combinator**

$$\mathbf{B} \equiv \lambda x.\lambda y.\lambda z.\, x(yz) \colon (B \to C) \to (A \to B) \to (A \to C)$$

characterized by the equation

$$\mathbf{B}abc = a(bc)\,.$$

The interpretation of the open term

$$x \colon B \to C\,,\quad y \colon A \to B\,,\quad z \colon A \vdash x(yz) \colon C$$

is as follows:

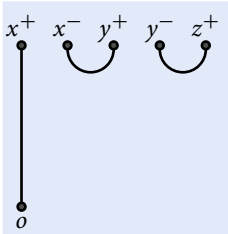

Here x^+ is the output of x, and x^- the input, and similarly for y. The output of the whole expression is o. When the variables are abstracted, the following caps-only diagram is obtained:

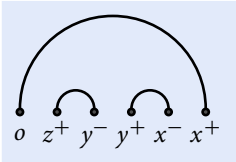

Now consider an application $\mathbf{B}abc$:

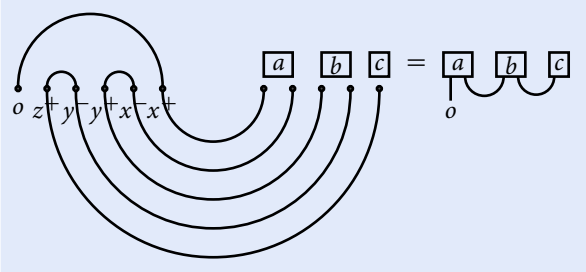

If the constraint on planarity in the diagrams is relaxed, a similar representation can be given of the **commutation combinator**:

$$\mathbf{C} \equiv \lambda x.\,\lambda y.\,\lambda z.\, xzy \colon$$
$$(A \to B \to C) \to B \to A \to C$$

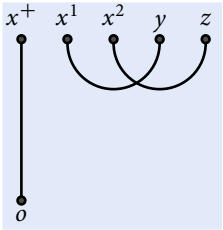

More generally, the **Brauer algebra** (1931) arises if the planarity condition on the TL algebra is removed. This plays an important role in the representation theory of the Orthogonal group ('Schur–Weyl duality'), which is now part of a whole genre of 'diagram algebras' in Representation Theory.

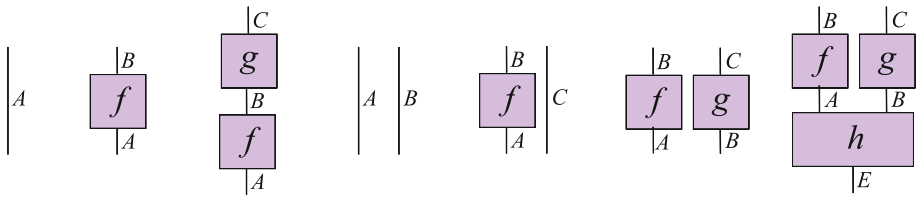

Logic and Geometry of Agents in Agent-Based Modeling, Figure 5
Graphical calculus for monoidal categories

With **BCI** combinators one can interpret **Linear λ-calculus**. The Kelly–Laplaza construction of the free compact closed category [34] can be retrieved by a straightforward generalization of these ideas.

Quantum Computation The **same** graphical calculus and underlying algebraic structure which have been tracked through logic, geometry and computation has been applied to quantum information and computation, yielding an incisive analysis of **quantum information flow**, and powerful and illuminating methods for reasoning about quantum informatic processes and protocols [6]. This "strongly compact closed graphical calculus" can be seen as a very substantial 2-dimensional extension of Dirac's *bra-ket* notation [23]. In the graphical calculus, physical processes are depicted by boxes, and the inputs and outputs of these boxes are labeled by *types* which indicate the kind of system on which these boxes act, e. g. one qubit, several qubits, classical data, etc.: see Fig 5. Algebraically, these correspond to $1_A : A \to A$, $f : A \to B$, $g \circ f$, $1_A \otimes 1_B$, $f \otimes 1_C$, $f \otimes g$, $(f \otimes g) \circ h$, respectively. (The convention in these diagrams is that the 'upward' vertical direction represents progress of time.)

Kets, Bras and Scalars A special role is played by boxes with either no input or no output, corresponding to **states** and **costates** respectively (cf. Dirac's kets and bras [23]), which are depicted by triangles. **Scalars** then arise naturally by composing these elements (cf. inner-product or Dirac's bra-ket):

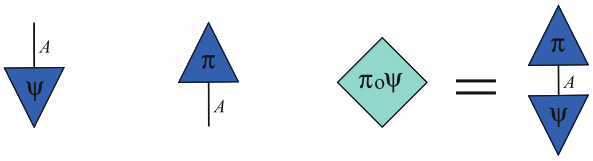

Bell States and Costates The cups and caps which have already appeared in their various guises as axiom and cut links, or in abstraction and application, now take on the rôle of **Bell states and costates**, the fundamental building blocks of quantum entanglement. (Mathematically, they arise as the transpose and co-transpose of the identity, which exist in any finite-dimensional Hilbert space by "map-state duality"). They are represented with triangles encasing the cups and caps to emphasize their **operational character** in the physical interpretation: they represent **preparation and test** of the Bell state.

The formation of **names** and **conames** of arrows (i. e. map-state and map-costate duality) is conveniently depicted thus:

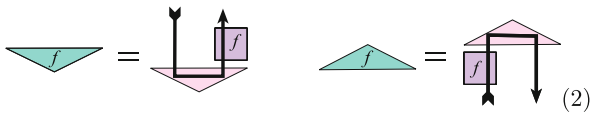

(2)

The key lemma in exposing the quantum information flow in (bipartite) entangled quantum systems [6]:

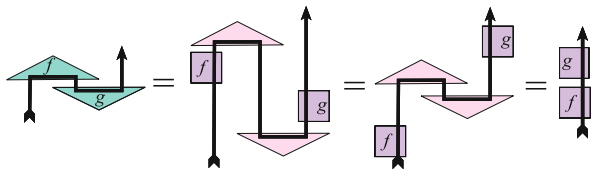

Note in particular the interesting phenomenon of "apparent reversal of the causal order" . While on the left, physically, the state labeled g is firstly prepared, and then the costate labeled f is applied, the global effect is *as if* f itself is applied first, and only then g.

Derivation of Quantum Teleportation This is the most basic application of compositionality in action. The basic quantum mechanical potential for teleportation can be read off immediately from the geometry of Bell states and

costates:

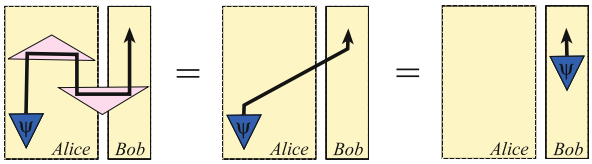

This is not quite the whole story, because of the non-deterministic nature of measurements. But it suffices to introduce a unitary correction. Using the lemma, the full description of teleportation becomes:

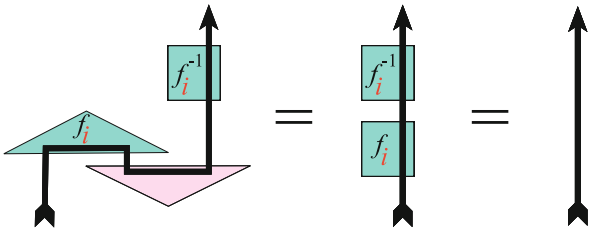

Further Directions

This article has described what is still very much an emerging area of research, rather than surveying an established field. Based on the progress which has already been made, a number of promising directions for future work are apparent, which may lead to important contributions to the general study of agent-based systems:

- The way in which, as illustrated in the previous subsection, the **same** structures arise in a wide diversity of situations, strongly suggests that a Logic of Interaction can be developed, which combines genuine depth with wide applicability.
- Further study of the links between logic and geometry, and of the significance of geometric constraints on informational contexts, looks promising.
- The program of full completeness, and giving equivalent geometric and interactive characterizations of constrained type systems which capture important complexity classes, may open up new insights into currently intractable problems.
- The connections between informatic and physical structures and constraints which have been found in the work on quantum information are promising both as new directions in the foundations of quantum mechanics and the quantum/classical boundary, and in elucidating fundamental structures of informatics.

Bibliography

Primary Literature

1. Abramsky S (1996) Retracing some paths in process algebra. In: Proc of CONCUR 96. Lectures notes in computer science, vol 1119. Springer, Berlin, pp 1–17
2. Abramsky S (2002) Algorithmic game semantics: a tutorial introduction. In: Proof and system-reliability. Kluwer, Dordrecht
3. Abramsky S (2004) High-level methods for quantum computation and information. In: Proc of the 19th annual IEEE symposium on logic in computer science. IEEE Computer Science Press, Los Alamitos pp 410–414
4. Abramsky S (2005) Abstract scalars, loops, and free traced and strongly compact closed categories. In: Proc of CALCO 2005. Springer lecture notes in computer science, vol 3629. Springer, Berlin, pp 1–31
5. Abramsky S (2007) Temperley–Lieb algebras and geometry of interaction. In: Chen G, Kauffman L, Lamonaco S (eds) Mathematics of quantum computing and technology. Chapman and Hall/CRC, Boca Raton, pp 413–458
6. Abramsky S, Coecke B (2004) A categorical semantics of quantum protocols. In: Proc of the 19th Annual IEEE symposium on logic in computer science: LICS 2004. IEEE Computer Society, Los Alamitos, pp 415–425
7. Abramsky S, Coecke B (2005) Abstract physical traces. Theory Appl Categ 14:111–124
8. Abramsky S, Jagadeesan R (1994) New foundations for the geometry of interaction. Inform Comput 111:53–119
9. Abramsky S, Jagadeesan R (1994) Games and full completeness for multiplicative linear logic. J Symb Log 59:543–574
10. Abramsky S, Jagadeesan R (2005) A game semantics for generic polymorphism. Ann Pure Appl Log 133:3–37
11. Abramsky S, McCusker G (1997) Linearity, sharing and state. In: O'Hearn P, Tennent RD (eds) Algol-like languages. Birkhauser, Basel, pp 317–348
12. Abramsky S, McCusker G (1999) Full abstraction for idealized Algol with passive expressions. Theor Comput Sci 227: 3–42
13. Abramsky S, McCusker G (1999) Game semantics. In: Computational logic: Proc of the 1997 Marktoberdorf Summer School. Springer, Berlin, pp 1–56
14. Abramsky S, Mellies P-A (1999) Concurrent games and full completeness. In: Proc of the 14th international symposium on logic in computer science. Computer Society Press of the IEEE, Los Alamitos, pp 431–442
15. Abramsky S, Honda K, McCusker G (1998) A fully abstract game semantics for general references. In: Proc of the 13th international symposium on logic in computer science. Computer Society Press of the IEEE, Los Alamitos, pp 334–344
16. Abramsky S, Jagadeesan R, Malacaria P (2000) Full abstraction for PCF. Inform Comput 163:409–470
17. Abramsky S, Ghica DR, Murawski AS, Ong C-HL (2004) Applying game semantics to compositional software modeling and verification. In: Proc TACAS'04. LNCS, vol 2988. pp 421–435
18. Abramsky S, Ghica DR, Murawski AS, Stark IDB, Ong C-HL (2004) Nominal games and full abstraction for the nu-calculus. In: Proc LICS'04. IEEE Computer Society Press, Los Alamitos, pp 150–159
19. Blute R, Scott PJ (1998) The Shuffle Hopf Algebra and noncommutative full completeness. J Symb Log 63(4):1413–1436
20. Blute R, Hamano M, Scott PJ (2005) Softness of hypercoherences and MALL full completeness. Ann Pure Appl Log 131(1–3):1–63

21. Danos V, Harmer R (2002) Probabilistic game semantics. ACM Trans Comput Log 3(3):359–382
22. Devarajan H, Hughes D, Plotkin G, Pratt V (1999) Full completeness of the multiplicative linear logic of Chu spaces. In: Proc of the 14th Annual IEEE symposium on logic in computer science. pp 234–242
23. Dirac PAM (1947) Principles of quantum mechanics. Oxford University Press, Oxford
24. Ghica DR, McCusker G (2000) Reasoning about idealized algol using regular languages. In: Proc ICALP'00. LNCS, vol 1853. pp 103–116
25. Ghica DR, Murawski AS (2004) Angelic semantics of fine-grained concurrency. In: Proc FOSSACS'04. LNCS, vol 2987. pp 211–225
26. Ghica DR, Murawski AS (2006) Compositional model extraction for higher-order concurrent programs. In: Proc TACAS'06. LNCS
27. Girard J-Y (1987) Linear logic. Theor Comput Sci 50(1):1–102
28. Girard J-Y (1989) Geometry of interaction I: interpretation of system F. In: Ferro R et al (eds) Logic Colloquium '88. Elsevier, Amsterdam, pp 221–260
29. Harmer R, McCusker G (1999) A fully abstract game semantics for finite nondeterminism. In: Proc 14th Annual IEEE symposium on logic in computer science. IEEE Computer Society Press, Los Alamitos
30. Hughes D (2000) Hypergame semantics: full completeness for system F. D Phil Mathematical Sciences. Oxford University, Oxford
31. Hyland JME, Ong C-HL (2000) On full abstraction for PCF: i. models, observables and the full abstraction problem, ii. dialogue games and innocent strategies, iii. a fully abstract and universal game model. Informat Comput 163:285–408
32. Jones VFR (1985) A polynomial invariant for links via von Neumann algebras. Bull Amer Math Soc 129:103–112
33. Kauffman LH (1990) An invariant of regular isotopy. Trans Amer Math Soc 318(2):417–471
34. Kelly GM, Laplaza ML (1980) Coherence for compact closed categories. J Pure Appl Algebra 19:193–213
35. Laird J (1997) Full abstraction for functional languages with control. In: Proc of the 12th Annual symposium on logic in computer science, LICS '97. Extended abstract
36. Laird J (2001) A fully abstract games semantics of local exceptions. In: Proc of the 16th Annual symposium on logic in computer science, LICS '01. Extended abstract
37. Loader R (1994) Models of Lambda calculi and linear logic. D Phil thesis. Oxford University, Oxford
38. Murawski AS, Ong C-HL, Walukiewicz I (2005) Idealized Algol with ground recursion and DPDA equivalence. In: Proc ICALP'05. LNCS, vol 3580. pp 917–929

Books and Reviews

Girard J-Y, Lafont Y, Taylor P (1989) Proof and types. Cambridge Tracts in Theoretical Computer Science
Hindley JR, Seldin JP (1986) Introduction to combinators and the λ-calculus. Cambridge University Press, Cambridge
Kauffman LH (1994) Knots in physics. World Scientific Press, Singapore
Troelstra AS (1992) Lectures on linear logic. Center for the Study of Language and Information Lecture Notes No. 29

Lyapunov–Schmidt Method for Dynamical Systems

André Vanderbauwhede
Department of Pure Mathematics and Computer Algebra, Ghent University, Gent, Belgium

Article Outline

Glossary
Definition of the Subject
Introduction
Lyapunov–Schmidt Method for Equilibria
Lyapunov–Schmidt Method in Discrete Systems
Lyapunov–Schmidt Method for Periodic Solutions
Lyapunov–Schmidt Method in Infinite Dimensions
Outlook
Bibliography

Glossary

Bifurcation A change in the local or global qualitative behavior of a system under the change of one or more parameters; the critical parameter values at which the change takes place are called bifurcation values.

Saddle-node bifurcation A bifurcation where two equilibria, usually one stable and the other unstable, merge and then disappear. A similar bifurcation can also happen with periodic orbits.

Pitchfork bifurcation A bifurcation where an equilibrium loses stability and at the same time two new stable equilibria emerge; this type of bifurcation frequently appears in systems with symmetry.

Period-doubling bifurcation A bifurcation where a fixed point of a mapping (discrete system) loses stability and a stable period-two point emerges. In a continuous system this corresponds to a periodic orbits losing stability while a new periodic orbit, with approximately twice the period of the original one, bifurcates.

Equivariant system A system with symmetries; these symmetries form a group of (usually linear) transformations commuting with the vectorfield or the mapping generating the system.

Reversible system A symmetric system where some of the symmetries involve a reversal of the time.

Definition of the Subject

The mathematical description of the state and the evolution of systems used to model all kind of applications requires in many cases a large or even an infinite number of

variables. In contrast, a number of basic phenomena observed in such systems (in particular bifurcation phenomena) are known to depend only on a very small number of variables. A simple example is the Hopf bifurcation, where the system changes from an equilibrium state to a periodic regime when the parameters cross a certain critical boundary, a transition which in a typical system essentially only involves two of the state variables. There are several ways in which the original system can be reduced to a much smaller system which captures the phenomenon one wants to analyze. There is a dynamical approach, usually in the form of some kind of center manifold reduction, where one tries to find a subsystem (invariant under the full system flow) which contains the core of the bifurcation under consideration. The Lyapunov–Schmidt method takes a rather different point of view, by concentrating purely on the objects under study, such as equilibria or periodic solutions, setting up an equation for these objects in an appropriate space, and then performing a reduction on this equation without any further reference to the original dynamics. This is done by splitting both the unknown and the equation in two parts, solving one of the resulting partial equations and substituting the solutions in the second equation. The outcome is typically a low-dimensional set of algebraic equations – usually called bifurcation equations – the solutions of which correspond in a one-to-one way with the chosen solutions of the original system. The analysis of these bifurcation equations (using appropriate techniques) allows to prove existence and multiplicity results and to describe bifurcation scenarios.

The Liapunov–Schmidt reduction method has a long history, going back to the early 20th century. It has been used in many different forms, and sometimes under different names, to a large variety of problems both within and outside the strict domain of dynamical systems and bifurcation theory: continuous and discrete systems, ordinary and partial differential equations, functional differential equations, integral equations, variational problems, and so on. Within the theory of dynamical systems the method has been widely used (and remains to be used) for the study of bifurcations of equilibria and periodic orbits.

Introduction

The Lyapunov–Schmidt method is, as its name suggests, a method, a technique or procedure which can be applied to a wide class of problems and equations, ranging from local bifurcation problems for equilibria and periodic solutions in finite-dimensional systems to global existence and multiplicity results for solutions of infinite-dimensional problems. The common theme in each of these applications is that of a *reduction*: the original equation is partly solved, and substitution of the partial solution in the remaining equations gives a *reduced equation* of a lower dimension than the original one. In many cases – but there are notorious exceptions – the reduced problem is finite-dimensional, with low dimension (one, in the best case).

Lyapunov–Schmidt Method – An Overview

The basic idea behind the Lyapunov–Schmidt reduction method is a rather simple splitting algorithm which can in very general terms be described as follows. The starting point is an equation

$$F_\lambda(x) = 0 \tag{1}$$

for an unknown x belonging to some space X, and depending on some parameter λ. The mapping F_λ takes the space X into a space Y; in simple cases $Y = X$, and typically X is a subspace of Y (although this is not necessary for the method to work). For example, $X = Y$ may be the phase space, and (1) the condition for $x \in X$ to be an equilibrium of a given dynamical system on X. Or X and Y may be spaces of periodic mappings, while (1) represents the condition for such periodic mapping to be a solution of a given system.

The main ingredients for the splitting are two projections: a projection P of the space X onto a subspace U, and a projection Q of the space Y onto a subspace W; in most applications the subspaces U and W are finite-dimensional. The projection P is used to split the variable x:

$$x = Px + (I - P)x = u + v\,,$$

with $u = Px \in U$ and $v = (I - P)x \in V = (I - P)(X)$. The projection Q allows to split the Eq. (1) into a set of two equations, $(I - Q)F_\lambda(x) = 0$ and $QF_\lambda(x) = 0$. Bringing the two splittings together shows that (1) is equivalent to the system

$$(I - Q)F_\lambda(u + v) = 0\,, \quad QF_\lambda(u + v) = 0\,. \tag{2}$$

Of course these equations heavily depend on the choices for the projections P and Q. The main hypothesis which allows to perform a reduction is that the projections P and Q should be such that the first equation in (2) has for each (u, λ) in a certain domain a unique solution $v = v^*_\lambda(u)$. Depending on the particular problem this solution $v^*_\lambda(u)$ will be obtained by different techniques, but mostly some form of the contraction principle or the implicit function theorem is used. Substituting the solution of the first equation of (2) into the second equation results in the following

reduced equation, usually called the *bifurcation* or *determining equation*:

$$G_\lambda(u) = QF_\lambda\big(u + v_\lambda^*(u)\big) = 0\,. \tag{3}$$

Under the condition that both U and W are finite-dimensional this represents a finite set of equations for a finite number of unknowns. Every solution (u,λ) of (3) corresponds to a solution $(x,\lambda) = (u + v_\lambda^*(u),\lambda)$ of (1), and conversely, every solution (x,λ) of (1) (in the appropriate domain) can be written as $x = u + v_\lambda^*(u)$ for some solution (u,λ) of (3).

The main difficulty in analyzing the Eq. (3) consists in the fact that typically the solution $v_\lambda^*(u)$ of the first equation of (2) is not explicitly known, and therefore also the mapping G_λ (which takes U into W) is not explicitly known. It is therefore important that the set-up (i. e. the projections P and Q used in the reduction) should be such that at least some qualitative properties of v_λ^* and G_λ, such as continuity and differentiability, can be obtained; when the method is used to study local bifurcation problems also a good approximation of these mappings, such as for example a few terms in their Taylor expansion, is required to obtain relevant conclusions.

When the method is used to study (local) bifurcation phenomena one generally assumes that a particular solution (x_0,λ_0) of (1) is known, and one wants to obtain the full solution set of (1) close to this given solution. In such case a convenient choice for the projections P and Q is related to the linearization $L_0 = DF_{\lambda_0}(x_0)$ of the mapping F_λ: the projection P should be such that $U = \mathrm{N}(L_0) = \{x \in X \mid L_0 x = 0\}$, while Q should be a projection of Y onto a subspace W which is complementary to the range $\mathrm{R}(L_0) = \{A_0 x \mid x \in X\}$ of L_0, i. e. $(I-Q)(Y) = \mathrm{R}(L_0)$. With such choice, and under appropriate smoothness conditions for the mapping F_λ, the existence of a unique solution $v = v_\lambda^*(u)$ of the first equation in (2) is guaranteed by the implicit function theorem for all (u,λ) near $(u_0,\lambda_0) = (Px_0,\lambda_0)$. Also the approximation of $v_\lambda^*(u)$ and $G_\lambda(u)$ by Taylor expansions works quite well in such case.

Lyapunov–Schmidt Method – An Example

As a simple but illustrative example on how the method works in practice consider the second order scalar ordinary differential equation

$$\ddot y + \alpha\dot y + y + g(y,\dot y) = 0\,, \tag{4}$$

with $\alpha \in \mathbb{R}$ a parameter and $g(y,\dot y)$ a smooth function of two variables such that $g(y,\dot y) = \mathrm{O}((|y| + |\dot y|)^2)$ as $(y,\dot y) \to (0,0)$. This equation has the equilibrium solution $y(t) \equiv 0$ for all values of the parameter; the linearization at this equilibrium,

$$\ddot y + \alpha\dot y + y = 0\,, \tag{5}$$

has for $\alpha = 0$ a two-dimensional space of periodic solutions given by $\{y(t) = a\cos t + b\sin t \mid (a,b) \in \mathbb{R}^2\}$. One may then ask which of these periodic solutions survive when the linear Eq. (5) is replaced by the non-linear Eq. (4). Since (5) is an approximation of (4) only for small $(y,\dot y)$, and since (5) has periodic solutions only for $\alpha = 0$, the question can be stated more precisely as follows: determine, for all small values of α, all small periodic solutions of (4). An immediate problem which arises is that different periodic solutions may have different periods; indeed, the determination of the period is part of the problem. In the Lyapunov–Schmidt approach this is handled by setting $y(t) = x((1+\tau)t)$, resulting in the new equation

$$(1+\tau)^2\ddot x + \alpha(1+\tau)\dot x + x + g(x,(1+\tau)\dot x) = 0\,; \tag{6}$$

2π-periodic solutions of (6) correspond to $2\pi(1+\tau)^{-1}$-periodic solutions of (4). The new problem is then to determine, for all small values of (τ,α), all small 2π-periodic solutions of (6). Observe that this problem is symmetric with respect to phase shifts: if $x(t)$ is a 2π-periodic solution, then so is $x(t+\phi)$, for each ϕ.

To solve this problem a possible 2π-periodic solution $x(t)$ is written in the form

$$x(t) = a\cos t + b\sin t + v(t) = \rho\cos(t+\phi) + v(t)\,,$$

with $v(t)$ a twice continuously differentiable function such that

$$\int_0^{2\pi} v(t)\cos t\,\mathrm{d}t = \int_0^{2\pi} v(t)\sin t\,\mathrm{d}t = 0\,. \tag{7}$$

Taking the symmetry with respect to phase shifts into account one can set $\phi = 0$, such that $x(t)$ takes the form

$$x(t) = \rho\cos t + v(t)\,, \quad \text{with } v \in V^{(2)}\,; \tag{8}$$

here $V^{(2)}$ denotes the space of 2π-periodic C^2-functions satisfying (7), while $V^{(0)}$ will denote the space of continuous such functions. It is an easy exercise to verify that for each $\tilde v \in V^{(0)}$ the non-homogeneous linear equation

$$\ddot v + v = \tilde v(t) \tag{9}$$

has a unique solution $v \in V^{(2)}$, given by

$$v(t) = \int_0^t \sin(t-s)\tilde v(s)\mathrm{d}s + \frac{1}{2\pi}\int_0^{2\pi} \sin(t-s)s\tilde v(s)\mathrm{d}s\,.$$

For each $(\rho, v, \tau) \in \mathbb{R} \times V^{(2)} \times \mathbb{R}$ the term $g(\rho \cos t + v(t), (1+\tau)(-\rho \sin t + \dot{v}(t)))$ appearing in (6) is decomposed as

$$A(\rho, v, \tau) \cos t + B(\rho, v, \tau) \sin t + \tilde{g}(\rho, v, \tau)(t) ,$$

with

$$A(\rho, v, \tau) = \frac{1}{\pi} \int_0^{2\pi} g(\rho \cos t + v(t), (1+\tau)(-\rho \sin t + \dot{v}(t))) \cos t \mathrm{d}t ,$$

$$B(\rho, v, \tau) = \frac{1}{\pi} \int_0^{2\pi} g(\rho \cos t + v(t), (1+\tau)(-\rho \sin t + \dot{v}(t))) \sin t \mathrm{d}t ,$$

and $\tilde{g}(\rho, v, \tau) \in V^{(0)}$. The Eq. (6) then splits into two equations, namely

$$[-(2\tau + \tau^2)\rho + A(\rho, v, \tau)] \cos t + \left[\alpha(1+\tau)\rho + B(\rho, v, \tau)\right] \sin t = 0 \quad (10)$$

and

$$(1+\tau)^2 \ddot{v} + \alpha(1+\tau)\dot{v} + v + \tilde{g}(\rho, v, \tau) = 0 . \quad (11)$$

Both equations are satisfied for $(\rho, v) = (0, 0)$. Using the solvability property of (9) mentioned above in combination with the implicit function theorem proves that the Eq. (11) has for each sufficiently small (ρ, τ, α) a unique small solution $v = v^*(\rho, \tau, \alpha) \in V^{(2)}$, with $v^*(0, \tau, \alpha) = 0$. Bringing this solution in (10) gives two scalar equations

$$\begin{aligned} &-(2\tau + \tau^2)\rho + A(\rho, v^*(\rho, \tau, \alpha), \tau) = 0 \\ \text{and} \quad &\alpha(1+\tau)\rho + B(\rho, v^*(\rho, \tau, \alpha), \tau) = 0 . \end{aligned} \quad (12)$$

It is easy to verify that $A(\rho, v^*(\rho, \tau, \alpha), \tau)$ and $B(\rho, v^*(\rho, \tau, \alpha), \tau)$ are of the order $\mathrm{O}(\rho^2)$ as $\rho \to 0$, such that

$$\begin{aligned} &A(\rho, v^*(\rho, \tau, \alpha), \tau) = \rho \tilde{A}(\rho, \tau, \alpha) \\ \text{and} \quad &B(\rho, v^*(\rho, \tau, \alpha), \tau) = \rho \tilde{B}(\rho, \tau, \alpha) , \end{aligned}$$

with $\tilde{A}(0, \tau, \alpha) = \tilde{B}(0, \tau, \alpha) = 0$, and (12) reduces for non-trivial solutions (i. e. solutions with $\rho \neq 0$) to

$$\begin{aligned} &-(2\tau + \tau^2) + \tilde{A}(\rho, \tau, \alpha) = 0 \\ \text{and} \quad &\alpha(1+\tau) + \tilde{B}(\rho, \tau, \alpha) = 0 . \end{aligned} \quad (13)$$

For small (ρ, τ, α) these equations can be solved by the implicit function theorem for $(\tau, \alpha) = (\tau^*(\rho), \alpha^*(\rho))$, with $(\tau^*(0), \alpha^*(0)) = (0, 0)$. This means that for each small ρ one can find a parameter value $\alpha = \alpha^*(\rho)$ such that (4) has a periodic solution with "amplitude" equal to $|\rho|$ and given by

$$x^*(\rho)(t) = \rho \cos((1+\tau^*(\rho))t) + v^*(\rho, \tau^*(\rho), \alpha^*(\rho))((1+\tau^*(\rho))t) ;$$

the minimal period of this solution is $2\pi(1+\tau^*(\rho))^{-1}$. In fact, the periodic orbits corresponding to ρ and $-\rho$ coincide, as can be seen from the following. The phase shift symmetry together with $-\rho \cos t = \rho \cos(t + \pi)$ and the uniqueness of the solution $v^*(\rho, \tau, \alpha)$ of (11) implies that $v^*(-\rho, \tau, \alpha)(t) = v^*(\rho, \tau, \alpha)(t + \pi)$, from which one can easily deduce that $A(-\rho, v^*(-\rho, \tau, \alpha), \tau) = -A(\rho, v^*(\rho, \tau, \alpha), \tau)$, $B(-\rho, v^*(-\rho, \tau, \alpha), \tau) = -B(\rho, v^*(\rho, \tau, \alpha), \tau)$ and $(\tau^*(-\rho), \alpha^*(-\rho)) = (\tau^*(\rho), \alpha^*(\rho))$. Hence

$$x^*(-\rho)(t) = x^*(\rho)(t + \pi(1+\tau^*(\rho))^{-1}) .$$

A further consequence is that $\tau^*(\rho)$ and $\alpha^*(\rho)$ are in fact functions of ρ^2, i. e. $\tau^*(\rho) = \hat{\tau}(\rho^2) = c\rho^2 + \text{h.o.t.}$ and $\alpha^*(\rho) = \hat{\alpha}(\rho^2) = d\rho^2 + \text{h.o.t.}$

It is interesting to notice the difference in treatment between the *auxiliary equation* (11) and the *bifurcation equations* (12). The auxiliary equation (11) is solved by a direct application of the implicit function theorem, while the same implicit function theorem can only be used to solve the bifurcation equations after some manipulation, which in this example simply consists in dividing by ρ to "eliminate" the trivial solution branch.

A further remark is that when the nonlinearity g in (4) does not depend on $\dot{y}$ (i. e. $g = g(y)$) then $\alpha^*(\rho) \equiv 0$ and we have what is usually called a *vertical bifurcation*. For the example this means that the equation $\ddot{y} + \alpha\dot{y} + y + g(y) = 0$ can have periodic solutions only for $\alpha = 0$, and that all small solutions of $\ddot{y} + y + g(y) = 0$ are indeed periodic. To see how this well known fact fits into the Lyapunov–Schmidt reduction observe that the equation $\ddot{y} + y + g(y) = 0$ is *reversible*, in the sense that if $y(t)$ is a solution then so is $y(-t)$. Using this reversibility one can show that $v^*(\rho, \tau, 0)(-t) = v^*(\rho, \tau, 0)(t)$; therefore $g(\rho \cos t + v^*(\rho, \tau, 0)(t))$ is an even function of t, and $B(\rho, v^*(\rho, \tau, 0), \tau) = 0$ for all (ρ, τ). This immediately implies that $\alpha^*(\rho) \equiv 0$, as claimed.

The foregoing example forms a very simple case of a *Hopf bifurcation*; this bifurcation will be treated in a more general context in Subsect. "Hopf Bifurcation".

Lyapunov–Schmidt Method – Historical Background

The Lyapunov–Schmidt method has been initiated in the early 20th century by A.M. Lyapunov [48] and E. Schmidt [55] in their (independent) studies of nonlinear integral equations. Later the method was used in diverse

forms in the work of (among many others) Cacciopoli [8], Cesari [9,10,11], Shimizu [56], Cronin [19], Bartle [5], Hale [26], Lewis [46], Vainberg and Trenogin [59], Antosiewicz [2], Rabinowitz [51], Bancroft, Hale and Sweet [4], Mawhin [49], Hall [33,34], Crandall and Rabinowitz [17], etc. Expositions on the method were given by Friedrichs [21], Nirenberg [50], Krasnosel'skii [43,44], Hale [27,28,30], Vainberg and Trenogin [60], and others. In particular the work of Cesari and Hale, who used the name *alternative method*, has put the method on a strong functional analytic basis. Apparently the name *Lyapunov–Schmidt method* was first introduced in the paper [59] of Vainberg and Trenogin; it has been used systematically for this type of approach since the seventies. In the seventies and the eighties the method was used extensively in a wide variety of bifurcation studies, not only in finite-dimensional systems but also in delay equations [29], reaction-diffusion equations [18,20], boundary value problems [17,52], and so on. In particular it forms together with group representation theory and singularity theory the main cornerstone for equivariant bifurcation theory, the study of bifurcation problems in the presence of symmetry (see e. g. the monographs by Sattinger [54], Vanderbauwhede [61], Golubitsky, Stewart and Schaeffer [23,24], and Chossat and Lauterbach [13], and the more recent theory of bifurcations in cell networks initiated by Golubitsky, Pivato and Stewart [25]). Later the method was extended and adapted to study the bifurcation of periodic orbits in Hamiltonian systems [63], reversible systems [41,42] and mappings [15,62], and there have been several numerical implementations [3,6,38]. Lin's method to study bifurcations near homoclinic and heteroclinic connections, based on the work of Lin [47] and developed by Sandstede [53] and Knobloch [40], is maybe from a technical point of view not a pure example of a Lyapunov–Schmidt reduction, but it certainly has all the flavors from it. In a different direction the method was used in combination with topological methods, in particular degree theory, to prove all kind of existence and multiplicity results and to obtain more global bifurcation results; a few references where one can read more about this are [22,37] and [39]. A recent account of the use of the Lyapunov–Schmidt method by mainly Russian authors is given in [57]. In recent years the Lyapunov–Schmidt method has been used to study (among many other topics) the existence of transition layers in singularly perturbed reaction-diffusion equations [31,32,58], and, in combination with a Newton scheme to solve small divisor problems, the existence of periodic and quasi-periodic solutions of nonlinear Schrödinger equations in one or more space dimensions [7,16].

The subsequent sections give a more detailed account of the method as it is used in the local bifurcation theory for equilibria and periodic orbits of finite-dimensional systems.

Lyapunov–Schmidt Method for Equilibria

The easiest example to illustrate the Lyapunov–Schmidt method is to consider the equilibria of a finite-dimensional parameter-dependent system of the form

$$\dot{x} = f(x, \lambda)\,, \tag{14}$$

with $x \in \mathbb{R}^n$ and $f\colon \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n$ a smooth mapping. These equilibria are given by the solutions of the equation

$$f(x, \lambda) = 0\,. \tag{15}$$

Let (x_0, λ_0) be any solution of this equation; by a translation one can without loss of generality assume that $(x_0, \lambda_0) = (0, 0)$, and hence $f(0, 0) = 0$. The problem is to describe *all* solutions of (15) which are close to the given solution $(0, 0)$.

Let A_0 be the Jacobian $(n \times n)$-matrix $D_x f(0, 0)$. If A_0 is invertible, that is, if the nullspace $\mathrm{N}(A_0)$ is trivial, then by the implicit function theorem the Eq. (15) has for each small parameter value λ a unique small solution $x = x^*(\lambda) \in \mathbb{R}^n$. This means that the equilibrium $x = 0$ for $\lambda = 0$ can be continued for all nearby parameter values λ. Therefore one is mainly interested in the case where A_0 is not invertible and $\mathrm{N}(A_0)$ nontrivial, and that is when the Lyapunov–Schmidt method can be used to reduce the Eq. (15), as follows.

Let P be the orthogonal projection in $\mathbb{R}^n$ onto $\mathrm{N}(A_0)$. This projection allows to split a vector $x \in \mathbb{R}^n$ as $x = Px + (I - P)x = u + v$, with $u = Px \in \mathrm{N}(A_0)$ and $v = (I - P)x \in V = \mathrm{N}(A_0)^\perp$ (the orthogonal complement of $\mathrm{N}(A_0)$). It is known from elementary algebra that $\mathrm{R}(A_0) = \mathrm{N}(A_0^T)^\perp$, where A_0^T is the transpose of A_0, and that $\dim \mathrm{N}(A_0) = \dim \mathrm{N}(A_0^T)$. Hence, if Q is the orthogonal projection of $\mathbb{R}^n$ onto $W = \mathrm{N}(A_0^T)$, then $(I - Q)(\mathbb{R}^n) = \mathrm{R}(A_0)$ and $\mathbb{R}^n = W \oplus \mathrm{R}(A_0)$. Bringing the splitting $x = u + v$ into (15) and using Q to split (15) into $Qf(x, \lambda) = 0$ and $(I - Q)f(x, \lambda)$ results in the following system of two equations, equivalent to (15):

$$\begin{aligned} &\tilde{f}(u, v, \lambda) = Qf(u + v, \lambda) = 0 \\ \text{and}\quad &\hat{f}(u, v, \lambda) = (I - Q)f(u + v, \lambda) = 0\,. \end{aligned} \tag{16}$$

Clearly $\tilde{f}(0, 0, 0) = 0$ and $\hat{f}(0, 0, 0) = 0$; also, $D_v \hat{f}(0, 0, 0)$ is equal to the restriction of $(I - Q)A_0$ to V, which is an invertible linear operator from V onto $\mathrm{R}(A_0)$. From the implicit function theorem one can then conclude

that the *auxiliary equation* $\hat{f}(u,v,\lambda)=0$ has, for each sufficiently small $(u,\lambda)\in U\times\mathbb{R}^k$, a unique small solution $v=v^*(u,\lambda)\in V$, with $v^*\colon U\times\mathbb{R}^k\to V$ a smooth mapping such that $v^*(0,0)=0$. Moreover, differentiating the identity $\hat{f}(u,v^*(u,\lambda),\lambda)=0$ and using the fact that $A_0u=0$ for all $u\in U$ gives $(I-Q)A_0D_uv^*(0,0)=0$, and hence $D_uv^*(0,0)=0$. Substitution of the solution $v=v^*(u,\lambda)$ in the first equation of (16) gives the bifurcation equation

$$\begin{aligned} g(u,\lambda) &= \tilde{f}(u,v^*(u,\lambda),\lambda) \\ &= Qf(u+v^*(u,\lambda),\lambda)=0\,. \end{aligned} \tag{17}$$

The bifurcation mapping $g\colon U\times\mathbb{R}^k\to W$ is smooth, and direct calculation shows that $g(0,0)=0$ and $D_ug(0,0)=0$ (just use the fact that $QA_0x=0$ for all $x\in\mathbb{R}^n$).

In order to come to conclusions about the equilibria of (14) one has now to study the bifurcation Eq. (17). Two remarks about this equation are in order here. First, since the solution $v=v^*(u,\lambda)$ is only implicitly defined, the mapping $v^*(u,\lambda)$ and hence also $g(u,\lambda)$ are usually not explicitly known; however, it is in principle possible to obtain the Taylor expansions of these mappings around the point $(u,\lambda)=(0,0)$. Second, since $D_ug(0,0)=0$ it is not possible to use the implicit function theorem to solve (17) for u (or even part of u) as a function of the parameter λ.

Lyapunov–Schmidt Reduction at a Simple Zero Eigenvalue

In the case of just one single parameter (i. e. $k=1$ and $\lambda\in\mathbb{R}$) one can argue as follows. If at some parameter value one has an equilibrium at which the linearization is invertible, then, as explained before, this equilibrium can be continued for nearby parameter values. This leads to one-parameter branches of equilibria of the form $\{(x^*(\lambda),\lambda)\mid\lambda\in I\}$, with $I\subset\mathbb{R}$ some interval. This continuation stops at parameter values $\lambda=\lambda_0$ where the linearization $A_\lambda=D_xf(x^*(\lambda),\lambda)$ becomes singular, i. e. when zero is an eigenvalue of A_{λ_0}. For one-parameter families of matrices, such as $\{A_\lambda\mid\lambda\in I\}$, this typically happens in such a way that zero will be a *simple eigenvalue* of A_{λ_0}, which means that $\dim\mathrm{N}(A_{\lambda_0})=1$ and $\mathbb{R}^n=\mathrm{N}(A_{\lambda_0})\oplus\mathrm{R}(A_{\lambda_0})$. Shifting $(x^*(\lambda_0),\lambda_0)$ to the origin and using an appropriate linear transformation this motivates the following hypothesis:

$$\begin{aligned} f(0,0)=0 \quad\text{and}\quad A_0x=(0,\hat{A}_0v), \\ \forall x=(u,v)\in\mathbb{R}\times\mathbb{R}^{n-1}=\mathbb{R}^n\,, \end{aligned} \tag{18}$$

where (as before) $A_0=D_xf(0,0)$ and where $\hat{A}_0$ is an *invertible* $(n-1)\times(n-1)$-matrix. With this set-up the Lyapunov–Schmidt reduction is straightforward: both P and Q are the projections in $\mathbb{R}^n$ onto the first component, the decomposition $x=(u,v)$ used in (18) is precisely the one needed to apply the reduction, and the resulting bifurcation equation is a single scalar equation

$$g(u,\lambda)=f_1(u,v^*(u,\lambda),\lambda)=0\,. \tag{19}$$

As in the general case $g(0,0)=0$ and $\partial g/\partial u(0,0)=0$. Depending on some further hypotheses this bifurcation equation leads to different bifurcation scenario's which are discussed in the next subsections. It should be noted that under appropriate conditions the bifurcation function $g(u,\lambda)$ also contains some stability information for the equilibria of (14) corresponding to solutions (u,λ) of (19) (see e. g. [23]): such equilibrium is stable if $\partial g/\partial u(u,\lambda)<0$, and unstable if $\partial g/\partial u(u,\lambda)>0$.

Saddle-Node Bifurcation

Just keeping the lowest order terms in u and λ the bifurcation function $g(u,\lambda)$ can be written as

$$\begin{aligned} g(u,\lambda) &= \alpha u^2+\beta\lambda+\text{h.o.t.}\,, \\ \alpha &= \frac{1}{2}\frac{\partial^2 f_1}{\partial u^2}(0,0,0)\,,\quad \beta=\frac{\partial f_1}{\partial\lambda}(0,0,0)\,. \end{aligned} \tag{20}$$

Assuming that $\beta\neq 0$ one can solve (19) for λ as a function of u to obtain

$$\lambda=\lambda^*(u)=-\alpha\beta^{-1}u^2+\text{h.o.t.} \tag{21}$$

If also $\alpha\neq 0$ then one has, in case $\alpha\beta>0$, for $\lambda<0$ two equilibria, one stable and one unstable, which merge for $\lambda=0$ and then disappear for $\lambda>0$; in case $\alpha\beta<0$ one has the opposite scenario. This is called a *saddle-node* or *limit point* bifurcation.

Pitchfork Bifurcation

Suppose that the vectorfield $f(x,\lambda)$ appearing in (14) has some symmetry properties, for example $f(-x,\lambda)=-f(x,\lambda)$. Then clearly one has for each parameter value $\lambda\in\mathbb{R}$ the *trivial* equilibrium $x^*(\lambda)=0$; setting (as before) $A_\lambda=D_xf(0,\lambda)$ one will typically find isolated parameter values where zero is a simple eigenvalue of A_λ. Assuming that $\lambda=0$ is such a critical parameter value, and working through the Lyapunov–Schmidt reduction, it is not hard to see that $v^*(-u,\lambda)=-v^*(u,\lambda)$, $g(-u,\lambda)=-g(u,\lambda)$ and $g(0,\lambda)=0$. Then clearly both coefficients α and β appearing in (20) will be zero, and the conditions for a saddle-node bifurcation are not satisfied. Instead, one

can write the bifurcation function $g(u,\lambda)$ in the form

$$g(u,\lambda) = u\tilde{g}(u,\lambda)\,, \quad \text{with } \tilde{g}(-u,\lambda) = \tilde{g}(u,\lambda) \text{ and } \tilde{g}(0,0) = 0\,; \tag{22}$$

hence the non-trivial solutions of (19) (these are solutions with $u \neq 0$) are obtained by solving the equation

$$\tilde{g}(u,\lambda) = 0\,. \tag{23}$$

Further calculations show that

$$\tilde{g}(u,\lambda) = \gamma u^2 + \delta\lambda + \text{h.o.t.}\,, \quad \gamma = \frac{1}{6}\frac{\partial^3 f_1}{\partial u^3}(0,0,0)\,, \quad \delta = \frac{\partial^2 f_1}{\partial u \partial \lambda}(0,0,0)\,. \tag{24}$$

Under the condition $\delta \neq 0$ it is possible to solve (23) for λ as a function of u, giving

$$\lambda = \lambda^*(u)\,, \quad \text{with } \lambda^*(0) = 0, \lambda^*(-u) = \lambda^*(u) \text{ and } \lambda^*(u) = -\gamma\delta^{-1}u^2 + \text{h.o.t.} \tag{25}$$

When also $\gamma \neq 0$ this results in the following bifurcation scenario when $\delta > 0$ and $\gamma < 0$: the zero equilibrium is stable for $\lambda < 0$ and becomes unstable for $\lambda > 0$; at $\lambda = 0$ two branches of symmetrically related non-trivial equilibria bifurcate from the trivial branch of equilibria; these nontrivial equilibria exist for $\lambda > 0$ and are stable. For other signs of γ and δ similar scenario's hold. This type of bifurcation is known as a *pitchfork bifurcation*.

Transcritical Bifurcation

The arguments used to obtain the pitchfork bifurcation remain partly valid in the following situation which is not necessarily related to symmetry. Suppose that the system (14) has a *given* equilibrium $x^*(\lambda)$ for each $\lambda \in \mathbb{R}$; such equilibrium may be a consequence of symmetry properties (see the case of a pitchfork bifurcation) or of some other property of the system. For example, in population models one always has the zero equilibrium; and in oscillation systems there is usually some basic equilibrium which persists for all parameter values. By a parameter dependent translation one can always arrange that $x^*(\lambda) = 0$ and hence $f(0,\lambda) = 0$ for all λ. Assuming as before that zero is a simple eigenvalue of $A_0 = D_x f(0,0)$ and applying the Lyapunov–Schmidt method one finds that

$$v^*(0,\lambda) = 0, \quad g(0,\lambda) = 0\,, \quad \forall\lambda \in \mathbb{R} \Rightarrow g(u,\lambda) = u\tilde{g}(u,\lambda)\,, \tag{26}$$

with $\tilde{g}(u,\lambda)$ a smooth function such that $\tilde{g}(0,0) = 0$. So again non-trivial equilibria are obtained by solving the Eq. (23), in which now the function $\tilde{g}(u,\lambda)$ can be written as

$$\tilde{g}(u,\lambda) = \eta u + \delta\lambda + \text{h.o.t.}\,, \quad \eta = \frac{1}{2}\frac{\partial^2 f_1}{\partial u^2}(0,0,0)\,, \quad \delta = \frac{\partial^2 f_1}{\partial u \partial \lambda}(0,0,0)\,. \tag{27}$$

Assuming that $\delta \neq 0$ one can solve (23) for λ as a function of u, resulting in

$$\lambda = \lambda^*(u) = -\eta\delta^{-1}u + \text{h.o.t.} \tag{28}$$

When also $\eta \neq 0$ this represents a *transcritical bifurcation*: for each $\lambda \neq 0$ there exists next to the trivial equilibrium at zero also a non-trivial equilibrium which approaches zero as λ tends to zero. For the same parameter value $\lambda \neq 0$ the trivial and the nontrivial equilibria have opposite stability properties, and there is an *exchange of stability* at $\lambda = 0$. For example, if $\delta > 0$ then the trivial equilibrium is stable for $\lambda < 0$ and unstable for $\lambda > 0$, while the non-trivial equilibrium is unstable for $\lambda < 0$ and stable for $\lambda > 0$.

A remark about the condition $\delta \neq 0$ used to obtain the pitchfork and the transcritical bifurcation is in order here. In both cases it was assumed that zero is a simple eigenvalue of $A_0 = A_{\lambda=0}$, where $A_\lambda = D_x f(0,\lambda)$. This simple eigenvalue can be continued for small values of λ, i. e. the matrix A_λ has, for sufficiently small values of λ, a simple real eigenvalue $\mu(\lambda)$ which depends smoothly on λ and is such that $\mu(0) = 0$. It is then not very hard to prove that the condition $\delta \neq 0$ is equivalent to the *transversality condition* $\mu'(0) \neq 0$ which states that the simple eigenvalue $\mu(\lambda)$ crosses zero with non-zero speed when λ crosses zero. Similar transversality conditions are needed to obtain most of the elementary bifurcations which appear in continuous systems such as (14) or in discrete systems such as considered in the next section.

A further remark has to do with the condition that zero should be a simple eigenvalue of $A_0 = D_x f(0,0)$. In certain cases this condition can be satisfied by restricting a larger system to an appropriate invariant subspace. This happens in particular when one studies bifurcations in equivariant systems; in many such systems the simple zero eigenvalue can be obtained by restricting the full system to invariant subsystems consisting of solutions which are invariant under certain isotropy subgroups. Usually the symmetry also forces the origin to be an equilibrium. Pitchfork or transcritical bifurcations observed for such invariant subsystems form then special cases of the *Equivariant Branching Lemma* [12,61]; see [61] and [24] for more details.

Lyapunov–Schmidt Method in Discrete Systems

The discrete analogue of the continuous system (14) has the form

$$x_{j+1} = f(x_j, \lambda)\,, \quad j = 0, 1, 2, \dots\,, \tag{29}$$

again with $x_j \in \mathbb{R}^n$ $(j \in \mathbb{N})$ and $f\colon \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n$ smooth. For a given value λ of the parameter the equilibria of (29) are given by the *fixed points* of $f_\lambda = f(\cdot, \lambda)$, i. e. by the solutions of the equation

$$f(x, \lambda) = x\,. \tag{30}$$

The continuation and bifurcation of such fixed points under a change of the parameter can be studied by the same techniques as the ones used in Sect. "Lyapunov–Schmidt Method for Equilibria": it is sufficient to replace $f(x, \lambda)$ by $F(x, \lambda) = f(x, \lambda) - x$. At a given solution $(x_0, \lambda_0) \in \mathbb{R}^n \times \mathbb{R}^k$ of (29) the linearization $L_0 = D_x F(x_0, \lambda_0)$ of $F_{\lambda_0} = F(\cdot, \lambda_0)$ is given by $L_0 = A_0 - I_n$, where $A_0 = D_x f(x_0, \lambda_0)$ and I_n is the identity matrix on $\mathbb{R}^n$. Therefore, if 1 is not an eigenvalue of A_0 then the fixed point x_0 can be continued for all λ sufficiently close to λ_0, and if 1 is a simple eigenvalue of A_0 (and $k = 1$) then, depending on some further conditions, one may have at $\lambda = \lambda_0$ a saddle-node, a pitchfork bifurcation or a transcritical bifurcation of fixed points.

Discrete systems such as (29) arise for example frequently in mathematical biology and mathematical economics. A special situation arises when the mapping f in (29) is a Poincaré map constructed near a given periodic orbit Γ_0 of a continuous parameter-dependent autonomous $(n+1)$-dimensional system (see e. g. [45]); in that case a fixed point of f_λ corresponds to a periodic orbit of the original continuous system, usually with the origin corresponding to Γ_0. The periodic orbit Γ_0 has always 1 as a Floquet multiplier; if this multiplier is *simple*, then 1 is not an eigenvalue of the linearization at the origin of the Poincaré map, and the periodic orbit Γ_0 can be continued for nearby parameter values. If $k = 1$ (a single parameter) and the multiplier 1 of Γ_0 has multiplicity two, then 1 will be a simple eigenvalue of the linearization of the Poincaré map, and one will typically obtain a saddle-node of fixed points of the Poincaré map, corresponding to a saddle-node of periodic orbits in the original continuous system. In particular cases such as the presence of symmetry one may also see a transcritical or pitchfork bifurcation of periodic orbits.

Periodic Points in Discrete Systems

Next to the bifurcation of fixed points for (29) one can also study the bifurcation of *periodic points* from fixed points. For any $\lambda \in \mathbb{R}^k$ and any integer $q \geq 1$ a q-periodic point of $f_\lambda = f(\cdot, \lambda)$ is a point $x \in \mathbb{R}^n$ such that the orbit $\{x_j \mid j \in \mathbb{N}\}$ of (29) starting at $x_0 = x$ has the property that $x_{j+q} = x_j$ for all $j \in \mathbb{N}$. In case f_λ is a Poincaré map associated to a periodic solution with minimal period $T_0 > 0$ of a continuous system then such q-periodic point corresponds to a periodic orbit of the continuous system with minimal period near qT_0. Such periodic solution is called a *subharmonic solution* (when $q \geq 2$); the bifurcation of periodic points from fixed points for discrete systems such as (29) is therefore strongly related to subharmonic bifurcation in continuous systems.

The periodicity condition $x_{j+q} = x_j$ (for all $j \in \mathbb{N}$) is equivalent to the condition that $x_q = x_0 = x$, or to

$$f_\lambda^q(x) = x, \qquad f_\lambda^q = f_\lambda \circ f_\lambda \circ \cdots \circ f_\lambda \quad (q \text{ times})\,. \tag{31}$$

Hence a q-periodic point of f_λ is a fixed point of f_λ^q (the q-th iterate of f_λ), and therefore one could in principle study the bifurcation of q-periodic points of f_λ by studying the bifurcation of fixed points of f_λ^q. Observe that a fixed point of f_λ is automatically also a fixed point of f_λ^q, for any $q \geq 1$.

However, there is a better approach which takes explicitly into account an important property of (31), namely that if x is a solution then so are all the other points of the orbit $\{x_j = f_\lambda^j(x) \mid j \in \mathbb{N}\}$ generated by x. The basic idea is the following: instead of considering all orbits of (29) and then finding all those which are q-periodic, one considers all q-periodic sequences with elements in $\mathbb{R}^n$ and then tries to determine which ones are actually orbits of the system (29). To work this out one has to introduce (for the chosen value of q) the space

$$\mathcal{O}_q = \left\{z = (x_j)_{j\in\mathbb{Z}} \mid x_j \in \mathbb{R}^n,\ x_{j+q} = x_j,\ \forall j \in \mathbb{Z}\right\}. \tag{32}$$

This *orbit space* is finite-dimensional, with dimension qn. Any map $g\colon \mathbb{R}^n \to \mathbb{R}^n$ can be lifted to a map $\hat{g}\colon \mathcal{O}_q \to \mathcal{O}_q$ by setting

$$\hat{g}(z) = (g(x_j))_{j\in\mathbb{Z}}\,, \quad \forall z = (x_j)_{j\in\mathbb{Z}} \in \mathcal{O}_q\,.$$

The *shift operator* is the linear mapping $\sigma\colon \mathcal{O}_q \to \mathcal{O}_q$ defined by

$$(\sigma \cdot z)_j = x_{j+1}\,, \quad \forall j \in \mathbb{Z}\,, \quad \forall z = (x_j)_{j\in\mathbb{Z}} \in \mathcal{O}_q\,.$$

Observe that $\sigma^q = I_{\mathcal{O}_q}$, and hence σ generates a $\mathbb{Z}_q$-action on $\mathcal{O}_q$ ($\mathbb{Z}_q$ is the cyclic group of order q, i. e. $\mathbb{Z}_q = \mathbb{Z}/q\mathbb{Z}$). Also $\hat{g} \circ \sigma = \sigma \circ \hat{g}$ for any $g\colon \mathbb{R}^n \to \mathbb{R}^n$. And finally, identifying $\mathcal{O}_q$ with $(\mathbb{R}^n)^q$ and using the standard scalar product it is easily seen that $\langle \sigma \cdot z_1, \sigma \cdot z_2\rangle = \langle z_1, z_2\rangle$ for all $z_1, z_2 \in \mathcal{O}_q$, i. e. σ is orthogonal and $\sigma^T = \sigma^{-1}$.

A q-periodic sequence $z = (x_j)_{j\in\mathbb{Z}} \in \mathcal{O}_q$ forms an orbit under the iteration of f_λ if and only if $f_\lambda(x_j) = x_{j+1}$ for all $j \in \mathbb{Z}$, that is, if and only if

$$\hat{f}_\lambda(z) = \sigma \cdot z \,. \tag{33}$$

The advantage of this equation over the equivalent Eq. (31) is that (33) is $\mathbb{Z}_q$-equivariant: both sides of the equation commute with σ, and if $(z, \lambda) \in \mathcal{O}_q \times \mathbb{R}^k$ is a solution, then so is $(\sigma^j \cdot z, \lambda)$ for any $j \in \mathbb{Z}$. This equivariance plays an important role in the bifurcation analysis for q-periodic points of (29) since it puts strong restrictions on the possible form of the bifurcation equation. The $\mathbb{Z}_q$-equivariance of (33) also fits nicely together with the Lyapunov–Schmidt reduction, as will be shown next.

Lyapunov–Schmidt Reduction for Periodic Points

Assume that (30) has a solution $(x_0, \lambda_0) \in \mathbb{R}^n \times \mathbb{R}^k$; as before one can without loss of generality translate this solution to the origin and assume that $f(0,0) = 0$. Consider then, for a given integer $q \geq 1$, the following problem:

(P_q) Determine, for all small values of the parameter λ, all small q-periodic points of f_λ.

The bifurcation equation for this problem can be obtained by a Lyapunov–Schmidt reduction near the solution $(z, \lambda) = (0, 0)$ of (33), as follows. Linearization of (33) at $(0, 0)$ gives the equation

$$\hat{A}_0 \cdot z = \sigma \cdot z \,, \quad A_0 = D_x f(0,0) \,, \tag{34}$$

and the corresponding linear operator $L_0 = \hat{A}_0 - \sigma \colon \mathcal{O}_q \to \mathcal{O}_q$. The nullspace $\mathrm{N}(L_0)$ consists of the q-periodic orbits of the linear discrete system $x_{j+1} = A_0 x_j$ and can be described as follows. Let

$$U = \mathrm{N}\big(A_0^q - I_n\big) \,, \quad S = A_0\big|_U \colon U \to U$$
$$\text{and} \quad \zeta \colon U \to \mathcal{O}_q \,,\ u \mapsto \zeta(u) = (S^j u)_{j\in\mathbb{Z}} \,. \tag{35}$$

Then $\mathrm{N}(L_0) = \zeta(U)$, and ζ is an isomorphism between U and $\mathrm{N}(L_0)$. Observe that the subspace U of $\mathbb{R}^n$ is the sum of the eigenspaces of A_0 corresponding to all eigenvalues $\mu \in \mathbb{C}$ of A_0 for which $\mu^q = 1$, i. e. all eigenvalues μ which are q-th roots of unity. A further observation is that the linear operator S generates a $\mathbb{Z}_q$-action on U (since $S^q = I_U$ by definition of U and S), and that

$$\zeta(Su) = \sigma \cdot \zeta(u) \,, \quad \forall u \in U \,. \tag{36}$$

The description of $\mathrm{R}(L_0) = \mathrm{N}(L_0^T)^\perp$ and the necessary projections can be somewhat simplified by assuming that all eigenvalues of A_0 which are q-th roots of unity are *semi-simple*, an assumption which implies that $\mathbb{R}^n = \mathrm{N}(A_0^q - I_n) \oplus \mathrm{R}(A_0^q - I_n) = U \oplus Y$; writing $x \in \mathbb{R}^n$ as $x = (u, y)$, with $u \in U$ and $y \in Y$, and constructing an appropriate basis of $\mathbb{R}^n$ one can then put A_0 in the following (diagonal) form:

$$A_0 x = A_0(u, y) = (Su, \tilde{A}_0 y) \,,$$
$$\forall x = (u, y) \in U \times Y = \mathbb{R}^n \,, \tag{37}$$

where S is orthogonal, i. e. $S^T S = I_U$, and where $\tilde{A}_0 \colon Y \to Y$ is such that $\tilde{A}_0^q - I_Y$ is invertible. Under these assumptions (see [62] for the general case) it is easily seen that $\mathrm{N}(L_0^T) = \mathrm{N}(\widehat{A}_0^T - \sigma^T)$ is equal to $\mathrm{N}(L_0)$ (use the fact that $\sigma^T = \sigma^{-1}$ and $S^T = S^{-1}$). Therefore $\mathrm{R}(L_0)$ is the orthogonal complement in $\mathcal{O}_q = (\mathbb{R}^n)^q$ of $\mathrm{N}(L_0) = \zeta(U)$, and every $z \in \mathcal{O}_q$ has a unique decomposition of the form

$$z = \zeta(u) + v \,, \quad \text{with } u \in U \text{ and } v \in V = \mathrm{R}(L_0) \,. \tag{38}$$

In particular, going back to (33) and replacing x by $\zeta(u) + v$, one can write

$$\hat{f}_\lambda(\zeta(u) + v) = \zeta(g(u, v, \lambda)) + h(u, v, \lambda) \,,$$

with $g \colon U \times V \times \mathbb{R}^k \to U$ and $h \colon U \times V \times \mathbb{R}^k \to V$ smooth mappings such that $g(0,0,0) = 0$, $h(0,0,0) = 0$, $D_u g(0,0,0) = S$, $D_v g(0,0,0) = 0$, $D_u h(0,0,0) = 0$, $D_v h(0,0,0) = \hat{A}_0\big|_V$, and

$$\begin{aligned} g(Su, \sigma \cdot v, \lambda) &= S g(u, v, \lambda) \,, \\ h(Su, \sigma \cdot v, \lambda) &= \sigma \cdot h(u, v, \lambda) \,. \end{aligned} \tag{39}$$

This last property is a consequence of (36), the $\mathbb{Z}_q$-equivariance of $\hat{f}_\lambda$, and the fact that σ leaves the decomposition $\mathcal{O}_q = \zeta(U) \oplus V$ invariant since σ commutes with L_0.

It follows from the foregoing that the Eq. (33) can be split in the two equations

$$g(u, v, \lambda) = Su \quad \text{and} \quad h(u, v, \lambda) = \sigma \cdot v \,. \tag{40}$$

Since $h(0,0,0) = 0$, $D_v h(0,0,0) = \hat{A}_0\big|_V$ and $L_0 = \hat{A}_0 - \sigma$ is invertible on $V = \mathrm{R}(L_0)$, the second equation in (40) can, locally near $(u, v, \lambda) = (0, 0, 0)$, be solved by the implicit function theorem for $v = v^*(u, \lambda)$; the mapping $v^* \colon U \times \mathbb{R}^k$ is smooth, with $v^*(0,0) = 0$ and $D_u v^*(0,0) = 0$. Also, (39) and the uniqueness of solutions given by the implicit function theorem imply that

$$v^*(Su, \lambda) = \sigma \cdot v^*(u, \lambda) \,. \tag{41}$$

Substituting the solution $v = v^*(u, \lambda)$ into the first equation of (40) gives the bifurcation equation

$$g^*(u, \lambda) = Su \,,$$
$$\text{with } g^*(u, \lambda) = g(u, v^*(u, \lambda), \lambda) \,. \tag{42}$$

This is an equation on the subspace U of $\mathbb{R}^n$; the mapping $g^*\colon U \times \mathbb{R}^k \to U$ is smooth, with $g^*(0,0) = 0$, $D_u g^*(0,0) = S$ and

$$g^*(Su,\lambda) = Sg^*(u,\lambda)\,, \quad \forall (u,\lambda) \in U \times \mathbb{R}^k\,. \tag{43}$$

This last property ensures that the bifurcation equation (42) is $\mathbb{Z}_q$-equivariant, i. e. commutes with the $\mathbb{Z}_q$-action generated by S on U, and if $(u,\lambda) \in U \times \mathbb{R}^k$ is a solution then so is $(\tilde{u},\lambda)$ for any $\tilde{u}$ belonging to the $\mathbb{Z}_q$-orbit $\{S^j u \mid j \in \mathbb{Z}\}$ generated by u.

Period-Doubling Bifurcation

Suppose that $f(0,0) = 0$ and that $A_0 = D_x f(0,0)$ has -1 as a simple eigenvalue, while $+1$ is not an eigenvalue. Then, taking $q = 2$ in the foregoing, $\dim\, U = 1$, $Su = -u$ for all $u \in U$, and the bifurcation equation (42) is a scalar equation:

$$g^*(u,\lambda) = -u\,, \quad \text{with } \frac{\partial g^*}{\partial u}(0,0) = -1 \quad \text{and } g^*(-u,\lambda) = -g^*(u,\lambda)\,. \tag{44}$$

Due to the symmetry the discussion of this bifurcation equations runs completely analogous to the one in Subsect. "Pitchfork Bifurcation" and leads, for $k = 1$ and under appropriate genericity conditions, to a pitchfork bifurcation. However, the interpretation of the bifurcation diagram is different in this case: the trivial solution $u = 0$ of (42) corresponds to a fixed point $x_0(\lambda)$ of f_λ which changes stability as λ crosses zero, while the symmetric pairs $\pm\hat{u}(\lambda)$ of nontrivial solutions which exist either for $\lambda < 0$ or for $\lambda > 0$ correspond to the two points $\hat{x}(\lambda)$ and $f_\lambda(\hat{x}(\lambda))$ of a period-two orbit of f_λ. Such bifurcation is known as a *period-doubling bifurcation* or a *flip bifurcation*. In case f_λ represents a Poincaré map near a given (nontrivial) periodic orbit with minimal period T_0 of a continuous system such period-doubling bifurcation means the following: the given periodic orbit persists for nearby parameter values, with minimal period T_λ near T_0, but when the parameter crosses the critical value there bifurcates a branch of periodic solutions with minimal period near $2T_0$; so also in the continuous system one sees a period-doubling.

Bifurcation of q-Periodic Points for $q \geq 3$

Next let $q \geq 3$ and assume that $f(0,0) = 0$ and that A_0 has a pair of simple eigenvalues

$$\exp(\pm i\theta_0)\,, \quad \text{with } \theta_0 = \frac{2\pi p}{q}, \quad \gcd(p,q) = 1\,. \tag{45}$$

Assume also that A_0 has no other eigenvalues μ with $\mu^q = 1$; in particular $+1$ and -1 are no eigenvalues. Then $\dim\, U = 2$ and the linear operator S in (37) has the matrix form

$$S = \begin{pmatrix} \cos\theta_0 & -\sin\theta_0 \\ \sin\theta_0 & \cos\theta_0 \end{pmatrix}.$$

Identifying $u = (u_1,u_2) \in U = \mathbb{R}^2$ with $z = u_1 + \mathrm{i}z_2 \in \mathbb{C}$ the expression (37) for A_0 takes the more explicit form

$$A_0 x = A_0(z,y) = \left(\mathrm{e}^{\mathrm{i}\theta_0} z, \tilde{A}_0 y\right), \quad \forall x = (z,y) \in \mathbb{C} \times Y = \mathbb{R}^n\,. \tag{46}$$

The Lyapunov–Schmidt reduction for q-periodic points as explained in Subsect. "Lyapunov–Schmidt Reduction for Periodic Points" leads to a complex bifurcation equation

$$g^*(z,\lambda) = \mathrm{e}^{\mathrm{i}\theta_0} z\,, \quad \text{with } g^*(z,0) = \mathrm{e}^{\mathrm{i}\theta_0} z + \mathrm{O}(|z|^2) \quad \text{and } g^*\left(\mathrm{e}^{\mathrm{i}\theta_0} z,\lambda\right) = \mathrm{e}^{\mathrm{i}\theta_0} g^*(z,\lambda)\,. \tag{47}$$

It can be shown (see e. g. [24] or [62]) that the $\mathbb{Z}_q$-equivariance of g^* implies that

$$g^*(z,\lambda) = \left(e^{\mathrm{i}\theta_0} + \phi_1(|z|^2,\lambda)\right)z + \phi_2(|z|^2,\lambda)\bar{z}^{q-1} + \mathrm{O}(|z|^{q+1}), \tag{48}$$

with ϕ_1 and ϕ_2 complex valued functions and $\phi_1(0,0) = 0$. The bifurcation equation (47) has for each $\lambda \in \mathbb{R}^k$ the trivial solution $z = 0$, corresponding to the continuation of the fixed point $x = 0$ at $\lambda = 0$ (remember that it was assumed that $+1$ is not an eigenvalue of A_0). To obtain nontrivial solutions one can multiply (47) by $\bar{z}$, set $z = \rho e^{\mathrm{i}\theta}$ and divide by ρ^2; the resulting equation has the form

$$\tilde{g}(\rho,\theta,\lambda) = \phi_1(\rho^2,\lambda) + \phi_2(\rho^2,\lambda)\rho^{q-2}\mathrm{e}^{-\mathrm{i}q\theta} + \mathrm{O}(\rho^q) = 0\,, \tag{49}$$

with $\tilde{g}(\rho,\theta+\theta_0,\lambda) = \tilde{g}(\rho,\theta,\lambda)$ and $\tilde{g}(-\rho,\theta+\pi,\lambda) = \tilde{g}(\rho,\theta,\lambda)$. The standard way to handle (49) is to assume $k = 2$ (i. e. $\lambda = (\lambda_1,\lambda_2) \in \mathbb{R}^2$) together with the transversality condition

$$\frac{\partial(\Re\phi_1, \Im\phi_1)}{\partial(\lambda_1,\lambda_2)}(0,0) \neq 0\,. \tag{50}$$

Then (49) can be solved for $\lambda = \lambda_q^*(\rho,\theta)$, with $\lambda_q^*(0,\theta,\lambda) = 0$, $\lambda_q^*(\rho,\theta+\theta_0,\lambda) = \lambda_q^*(\rho,\theta,\lambda)$ and $\lambda_q^*(-\rho,\theta+\pi,\lambda) = \lambda_q^*(\rho,\theta,\lambda)$. Both the transversality condition and the solution set $\{(\rho,\theta,\lambda_q^*(\rho,\theta)) \mid (\rho,\theta) \in \mathbb{R}^2\}$ require some further explanation.

Since $\exp(\pm\theta_0)$ are simple eigenvalues of A_0 it follows that these eigenvalues can be smoothly continued for nearby parameter values, i. e. A_λ has for small λ a pair

of simple eigenvalues $\{\mu^*(\lambda), \bar{\mu}^*(\lambda)\}$ near $\exp(\pm\theta_0)$. The transversality condition (50) means that the mapping $\lambda \mapsto \mu^*(\lambda)$ is a local diffeomorphism near $\lambda = 0$. Under this condition the solution $\lambda = \lambda_q^*(\rho, \theta)$ of (49) implies that for each λ in the range

$$\mathcal{A}_q = \{\lambda_q^*(\rho, \theta) \mid (\rho, \theta) \in \mathbb{R}^2\}$$

the mapping f_λ has at least one q-periodic orbit near $x = 0$. For $q \geq 5$ the set $\mathcal{A}_q$ typically forms a horn-like domain in the λ-plane, with tip at the origin; higher values of q correspond to sharper tips. Indeed, it follows from (48) that

$$\lambda_q^*(\rho, \theta) = \rho^2\alpha(\rho^2) + \rho^{q-2}e^{-q\theta}\beta(\rho^2) + \mathrm{O}(\rho^q) ,$$

for some complex valued functions $\alpha(\rho^2)$ and $\beta(\rho^2)$; if $\alpha(0) \neq 0$ this shows that the "thickness" of $\mathcal{A}_q$ at a distance d from the origin is of the order $d^{(q-2)/2}$ (for $q = 3$ and $q = 4$ the situation is different, see e. g. [62] or [36]). The horn-like regions $\mathcal{A}_q$ are known as *Arnol'd tongues*. For parameter values λ in the interior of $\mathcal{A}_q$ and close to $\lambda = 0$ the mapping f_λ has two q-periodic orbits which annihilate each other in a saddle-node bifurcation when the parameter crosses the boundary of $\mathcal{A}_q$. Moreover, the hypotheses imply that f_λ undergoes a *Neimark–Sacker bifurcation*: when λ crosses the curve defined by $|\mu^*(\lambda)| = 1$ the stability of the fixed point changes and an invariant curve bifurcates from the fixed point. The Arnol'd tongue $\mathcal{A}_q$ is contained in the region where an invariant curve exists, and the two periodic orbits are contained in this invariant curve; on the curve, one periodic orbit is stable, the other unstable.

When f_λ represents a Poincaré map the bifurcation picture as described in the foregoing can be lifted to the original continuous system, leading to the bifurcation of invariant tori and subharmonics. For more details see e. g. [36] or [45]. For a discussion of the case where the mappings f_λ are reversible see [15].

Lyapunov–Schmidt Method for Periodic Solutions

The ideas described in Sect. "Lyapunov–Schmidt Method in Discrete Systems and in particular in Subsect. "Lyapunov–Schmidt Reduction for Periodic Points" can be adapted for the treatment of the bifurcation of periodic orbits from equilibria in continuous systems (14); however, there are a few complications which arise for this continuous case. First, the orbit space consists of continuous periodic functions defined on the real line, hence this orbit space is infinite-dimensional; this necessitates a more careful handling of splittings and projections. Second, when looking for periodic solutions of (14) the (minimal) period is one of the unknowns of the problem. As will be seen in the treatment which follows the finite symmetry group $\mathbb{Z}_q$ which plays an important role in the bifurcation analysis of periodic orbits in discrete systems will be replaced by the continuous circle group S^1 in the case of continuous systems.

In order to get a more precise description of the problem, assume that $f(0, 0) = 0$ and consider the linearized equation

$$\dot{x} = A_0 x , \tag{51}$$

(with $A_0 = D_x f(0, 0)$ as before). The solutions have the form $x(t) = \mathrm{e}^{A_0 t}x_0$ with $x_0 \in \mathbb{R}^n$; such solution is T_0-periodic (for some $T_0 > 0$) if $x_0 \in \mathrm{N}(\mathrm{e}^{A_0 T_0} - I_n)$. In order to exclude a possible bifurcation of equilibria one should require that zero is not an eigenvalue of A_0, which entails that without further loss of generality one can also assume that $f(0, \lambda) = 0$ for all λ. By a time rescale one can put $T_0 = 2\pi$. The nullspace $\mathrm{N}(\mathrm{e}^{2\pi A_0} - I_n)$ is then given by the direct sum of the (real) eigenspaces corresponding to all pairs of eigenvalues of the form $\pm ik$ $(k = 1, 2, \ldots)$. A further simplification in the formulation of the Lyapunov–Schmidt reduction occurs if one assumes that all such eigenvalues are semisimple, leading to the following hypotheses about (14) which will be assumed further on:

(H) $f(0, \lambda) = 0$ for all λ, $A_0 = D_x f(0, 0)$ is invertible, and 1 is a semisimple eigenvalue of $\mathrm{e}^{2\pi A_0}$, i. e. $\mathbb{R}^n = \mathrm{N}(\mathrm{e}^{2\pi A_0} - I_n) \oplus \mathrm{R}(\mathrm{e}^{2\pi A_0} - I_n) = U \oplus Y$, where $U = \mathrm{N}(\mathrm{e}^{2\pi A_0} - I_n)$ and $Y = \mathrm{R}(\mathrm{e}^{2\pi A_0} - I_n)$.

By an appropriate choice of basis one can then more explicitly assume that

$$A_0 x = A_0(u, y) = (Su, \tilde{A}_0 y) , \qquad \forall x = (u, y) \in U \times Y = \mathbb{R}^n , \tag{52}$$

with $S\colon U \to U$ and $\tilde{A}_0\colon Y \to Y$ linear and invertible, S anti-symmetric ($S^T = -S$), $\mathrm{e}^{2\pi S} = I_U$ and $\mathrm{N}(\mathrm{e}^{2\pi \tilde{A}_0} - I_Y) = \{0\}$. The space of 2π-periodic solutions of (51) is then given by $\{\mathrm{e}^{St}u \mid u \in U\}$. The question one wants to answer is then which of these periodic solutions survive when we turn on the nonlinearities and change parameters. More precisely, the problem can be formulated as follows:

(P) Determine, for all $(T, \lambda) \in \mathbb{R} \times \mathbb{R}^k$ near $(2\pi, 0)$, all small T-periodic solutions of (14).

Lyapunov–Schmidt Reduction for the Problem (P)

Analogically to the treatment for discrete systems in Sect. "Lyapunov–Schmidt Method in Discrete Systems" the first step should be to reformulate (P) in an appropriate orbit space; to do so one has to fix the period, which is done as

follows. Let $\hat{x}(t)$ be a T-periodic solution of (14), define $\tau \in \mathbb{R}$ by $(1+\tau)T = 2\pi$, and set $\tilde{x}(t) = \hat{x}((1+\tau)^{-1}t)$. Then $\tilde{x}(t)$ is a 2π-periodic solution of the equation

$$(1+\tau)\dot{x} = f(x,\lambda)\,. \tag{53}$$

Conversely, if $\tilde{x}(t)$ is a 2π-periodic solution of (53) then $\hat{x}(t) = \tilde{x}((1+\tau)t)$ is a T-periodic solution of (14), with $T = 2\pi(1+\tau)^{-1}$. Therefore the problem (P) is equivalent to

(P*) Determine, for all small $(\tau,\lambda) \in \mathbb{R}\times\mathbb{R}^k$, all small 2π-periodic solutions of (53).

The orbit space for this problem is the space $C^0_{2\pi}$ of all continuous 2π-periodic mappings $x\colon \mathbb{R} \to \mathbb{R}^n$; this is a Banach space when one uses the norm $\|x\|^0_{2\pi} = \sup\{\|x(t)\| \mid t \in \mathbb{R}\}$. On $C^0_{2\pi}$ one can define a natural S^1-action (where $S^1 = \mathbb{R}/2\pi\mathbb{R}$ is the circle group) by the *phase shift operator*

$$\Sigma\colon S^1 \times C^0_{2\pi} \to C^0_2\pi,\ (\phi,x) \mapsto \Sigma_\phi(x)\,, \quad \text{with } \Sigma_\phi(x)(t) = x(t+\phi)\,. \tag{54}$$

The infinitesimal generator of this S^1-action is defined on the dense subspace $C^1_{2\pi}$ of continuously differentiable 2π-periodic mappings $x\colon \mathbb{R} \to \mathbb{R}^n$ and is given by

$$D_t\colon C^1_{2\pi} \to C^0_{2\pi},\ x \mapsto D_t x = \left.\frac{\mathrm{d}}{\mathrm{d}\phi}\right|_{\phi=0} \Sigma_\phi(x)\,, \quad (D_t x)(t) = \dot{x}(t),\ \forall (x,t) \in C^1_{2\pi}\times\mathbb{R}\,. \tag{55}$$

The subspace $C^1_{2\pi}$ is itself a Banach space when one uses the norm $\|x\|^1_{2\pi} = \|x\|^0_{2\pi} + \|D_t x\|^0_{2\pi}$; the space $C^1_{2\pi}$ is continuously imbedded in $C^0_{2\pi}$. Also, any continuous mapping $g\colon \mathbb{R}^n \to \mathbb{R}^n$ can be lifted to a mapping on $C^0_{2\pi}$ by defining

$$\hat{g}\colon C^0_{2\pi} \to C^0_{2\pi},\ x \mapsto \hat{g}(x)\,, \quad \text{with } \hat{g}(x)(t) = g(x(t))\,, \quad \forall (x,t) \in C^0_{2\pi}\times\mathbb{R}\,.$$

Moreover, such a lifting commutes with the shift operator: $\hat{g}\circ\Sigma_\phi = \Sigma_\phi \circ \hat{g}$ for all $\phi \in S^1$.

Using the foregoing definitions the problem (P*) translates into finding all small solutions $(x,\tau,\lambda) \in C^1_{2\pi}\times\mathbb{R}\times\mathbb{R}^k$ of the equation

$$\widehat{f}_\lambda(x) = (1+\tau)D_t x\,. \tag{56}$$

(Observe the analogy and the differences with (33)). This equation is S^1-equivariant: both sides commute with Σ_ϕ for all $\phi \in S^1$; therefore, if (x,τ,λ) is a solution, then so is $(\Sigma_\phi(x),\tau,\lambda)$ for all $\phi \in S^1$.

Next define $F\colon C^1_{2\pi}\times\mathbb{R}\times\mathbb{R}^k \to C^0_{2\pi}$ by $F(x,\tau,\lambda) = \widehat{f}_\lambda - (1+\tau)D_t$; then $F(0,\tau,\lambda) = 0$, and the linearization $L_0 = D_x F(0,0,0)\colon C^1_{2\pi} \to C^0_{2\pi}$ is given by $L_0 = \widehat{A}_0 - D_t$. The nullspace $\mathrm{N}(L_0) \subset C^1_{2\pi}$ is the space of all 2π-periodic solutions of (51) which was described before; hence there exists an isomorphism $\zeta\colon U \to \mathrm{N}(L_0)$ given by

$$\zeta(u)(t) = \mathrm{e}^{St}u\,, \quad \forall u \in U,\ \forall t \in \mathbb{R}\,. \tag{57}$$

It follows from the definition of U and $S = A_0\big|_U$ that S is the infinitesimal generator of an S^1-action on U given by

$$(\phi,u) \in S^1\times U \mapsto \mathrm{e}^{S\phi}u \in U\,.$$

Also, $\zeta(e^{S\phi}u) = \Sigma_\phi\zeta(u)$ and $\zeta(Su) = D_t\zeta(u)$ for all $(\phi,u) \in S^1\times U$.

In order to describe $\mathrm{R}(L_0)$ one has to find those $z \in C^0_{2\pi}$ for which the non-homogeneous linear 2π-periodic differential equation

$$\dot{x} = A_0 x + z(t) \tag{58}$$

has a 2π-periodic solution. The *Fredholm alternative* for this problem (see e.g. [28] or [61]) states that this will be the case if and only if

$$\int_0^{2\pi} \langle z(t), x(t)\rangle \mathrm{d}t = 0\,, \quad \forall x \in C^1_{2\pi}\colon\ -\dot{x} = A_0^T x\,.$$

Stated differently: if we introduce on $C^0_{2\pi}$ a scalar product $\langle\cdot,\cdot\rangle_{C^0_{2\pi}}$ by

$$\langle x,\hat{x}\rangle_{C^0_{2\pi}} = \int_0^{2\pi} \langle x(t),\hat{x}(t)\rangle \mathrm{d}t\,, \quad \forall x\,,\ \hat{x} \in C^0_{2\pi}\,,$$

and if we define the *adjoint operator* $L_0^*\colon C^1_{2\pi} \to C^0_{2\pi}$ by $L_0^* = \widehat{A}_0^T + D_t$, then $\mathrm{R}(L_0) = \mathrm{N}(L_0^*)^\perp$. It follows easily from (52) that $\mathrm{N}(L_0^*) = \mathrm{N}(L_0)$, and therefore $\mathrm{R}(L_0) = \mathrm{N}(L_0)^\perp$ and $C^0_{2\pi} = \mathrm{N}(L_0) \oplus \mathrm{R}(L_0)$. Moreover, $\mathrm{N}(L_0)$ is finite-dimensional and $\mathrm{R}(L_0)$ is a closed subspace of $C^0_{2\pi}$; these properties imply that $L_0\colon C^1_{2\pi} \to C^0_{2\pi}$ is a *Fredholm operator of index zero* (see the next section for the definition).

One can associate the orthogonal splitting $C^0_{2\pi} = \mathrm{N}(L_0) \oplus \mathrm{R}(L_0) = \zeta(U) \oplus V$ (with $V = \mathrm{R}(L_0)$) to a continuous projection $P\colon C^0_{2\pi} \to C^0_{2\pi}$ such that $\mathrm{R}(P) = \zeta(U)$ and $\mathrm{N}(P) = V$; explicitly $P = \zeta\circ\tilde{P}$, with $\tilde{P}\colon C^0_{2\pi} \to U$ given by

$$\tilde{P}(x) = \frac{1}{2\pi}\int_0^{2\pi} \mathrm{e}^{-Ss}u(s)\mathrm{d}s\,, \quad \forall x = x(\cdot) = (u(\cdot),y(\cdot)) \in C^0_{2\pi}\,. \tag{59}$$

Observe that $\tilde{P}\circ\Sigma_\phi = \mathrm{e}^{S\phi}\circ\tilde{P}$ and $P\circ\Sigma_\phi = \Sigma_\phi\circ P$ for every $\phi \in S^1$, and that the splitting $C^0_{2\pi} = \zeta(U) \oplus V$ is invariant under the S^1-action; as a consequence the infinitesimal generator D_t of the S^1-action maps $V \cap C^1_{2\pi}$ into V. Every $x \in C^1_{2\pi}$ can then be written as $x = \zeta(u)+v$, with $u = \tilde{P}(x) \in U$ and $v = (I-P)x \in V \cap C^1_{2\pi}$; in

a similar way $\widehat{f}_\lambda(\zeta(u)+v) = \zeta(g(u,v,\lambda)) + h(u,v,\lambda)$, with $g(u,v,\lambda) = \tilde{P}\hat{f}_\lambda(\zeta(u)+v) \in U$ and $h(u,v,\lambda) = (I-P)\hat{f}_\lambda(\zeta(u)+v) \in V$; the mappings $g\colon U\times V\times\mathbb{R}^k \to U$ and $h\colon U\times V\times\mathbb{R}^k \to V$ are smooth, with $g(0,0,\lambda)=0$, $h(0,0,\lambda)=0$, and satisfy the equivariance relations

$$g\big(\mathrm{e}^{S\phi}u, \Sigma_\phi v, \lambda\big) = \mathrm{e}^{S\phi}g(u,v,\lambda) \quad \text{and}$$
$$h\big(\mathrm{e}^{S\phi}u, \Sigma_\phi v, \lambda\big) = \Sigma_\phi(h(u,v,\lambda))\,, \quad \forall\phi\in S^1\,.$$

Using the foregoing decompositions the Eq. (56) splits into two equations

$$\begin{aligned} &g(u,v,\lambda) = (1+\tau)Su \\ \text{and}\quad &h(u,v,\lambda) = (1+\tau)D_t v\,, \end{aligned} \tag{60}$$

to be solved for small $(u,v,\tau,\lambda)\in U\times(V\cap C^1_{2\pi})\times\mathbb{R}\times\mathbb{R}^k$.

Since $h(0,0,\lambda) = 0$ and the restriction of L_0 to $V\cap C^1_{2\pi}$ is an isomorphism of $V\cap C^1_{2\pi}$ onto V the second equation in (60) can be solved for $v = v^*(u,\tau,\lambda)$, with $v^*\colon U\times\mathbb{R}\times\mathbb{R}^k \to V\cap C^1_{2\pi}$ a smooth mapping, $v^*(0,\tau,\lambda)=0$, $D_u v^*(0,0,0)=0$ and $v^*(e^{S\phi}u,\tau,\lambda) = \Sigma_\phi(v^*(u,\tau,\lambda))$ for all $\phi\in S^1$. Bringing this solution into the first equation of (60) gives the bifurcation equation

$$\begin{aligned} &g^*(u,\tau,\lambda) = (1+\tau)Su\,, \\ &\qquad\text{with}\ \ g^*(u,\tau,\lambda) = g(u,v^*(u,\tau,\lambda),\lambda)\,. \end{aligned} \tag{61}$$

The mapping $g^*\colon U\times\mathbb{R}\times\mathbb{R}^k \to U$ is smooth, with $g^*(0,\tau,\lambda)=0$, $D_u g^*(0,\tau,0) = S$ and

$$g^*(\mathrm{e}^{S\phi}u,\tau,\lambda) = \mathrm{e}^{S\phi}g^*(u,\tau,\lambda)\,, \quad \forall\phi\in S^1\,. \tag{62}$$

This S^1-equivariance of g^* implies that for each solution $(u,\tau,\lambda)\in U\times\mathbb{R}\times\mathbb{R}^k$ of (62) also $(e^{S\phi}u,\tau,\lambda)$ is a solution, for each $\phi\in S^1$. Solving (61) near the origin and working back through the reduction allows to solve the problems (P*) and (P).

The method outlined in this subsection can be adapted for special types of systems, such as Hamiltonian or reversible systems, and can also be generalized to the case where the linearization A_0 has some nilpotencies (non-semisimple eigenvalues); see e. g. [41,63] and [42] for more details. In these papers there is also established a link between the Lyapunov–Schmidt reduction and normal form theory.

Hopf Bifurcation

The simplest possible case in which the reduction of Subsect. "Lyapunov–Schmidt Reduction for the Problem (P)" can be applied is when A_0 has $\pm i$ as simple eigenvalues, while there are no other eigenvalues of the form $\pm ki$, with $k\in\mathbb{N}$ (this is called a *non-resonance condition*). Then $\dim U = 2$ and S can be brought into the matrix form

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Identifying $u=(u_1,u_2)\in U=\mathbb{R}^2$ with $z=u_1+iu_2\in\mathbb{C}$ the linear operator A_0 gets the form (compare with (52))

$$\begin{aligned} &A_0x = A_0(z,y) = (iz, \tilde{A}_0 y)\,, \\ &\qquad\qquad \forall x=(z,y)\in\mathbb{C}\times\mathbb{R}^{n-2} = \mathbb{R}^n\,. \end{aligned} \tag{63}$$

The bifurcation equation is then a single complex equation

$$g^*(z,\tau,\lambda) = (1+\tau)iz\,, \tag{64}$$

with $g^*\colon\mathbb{C}\times\mathbb{R}\times\mathbb{R}^k\to\mathbb{C}$ such that $g^*(0,\tau,\lambda)=0$, $g^*(z,\tau,0) = \mathrm{i}z + \mathrm{O}(|z|^2)$ and

$$g^*(\mathrm{e}^{\mathrm{i}\phi}z,\tau,\lambda) = \mathrm{e}^{\mathrm{i}\phi}g^*(z,\tau,\lambda)\,, \quad \forall\phi\in S^1\,. \tag{65}$$

This equivariance property of g^* implies (see e. g. [24]) that one can write

$$g^*(z,\tau,\lambda) = \tilde{g}(|z|^2,\tau,\lambda)z\,, \quad \text{with}\ \ \tilde{g}(0,\tau,0) = i\,. \tag{66}$$

The Eq. (64) has for each (τ,λ) the trivial solution $z=0$, corresponding to the equilibrium $x=0$ of (14) Using (66) the nontrivial solutions of (64) are given by the solutions of the equation

$$\tilde{g}(|z|^2,\tau,\lambda) = (1+\tau)i\,. \tag{67}$$

The solutions of this equation come in circles $\{(\mathrm{e}^{\mathrm{i}\phi}\rho, \tau,\lambda) \mid \phi\in S^1\}$; periodic solutions of (14) corresponding to different points on such circle are related to each other by phase shift, and therefore each circle of solutions of (67) corresponds to a unique periodic orbit of (14).

The standard way to handle (67) is to assume $k=1$ (i. e. $\lambda\in\mathbb{R}$) and

$$\Re\,\frac{\partial\tilde{g}}{\partial\lambda}(0,0,0) \neq 0\,. \tag{68}$$

Then (67) can be split in real and imaginary parts and solved by the implicit function theorem for $(\tau,\lambda) = (\tau^*(|z|^2),\lambda^*(|z|^2))$, with $(\tau^*(0),\lambda^*(0)) = (0,0)$. For each sufficiently $\rho>0$ there is a parameter value $\lambda=\lambda^*(\rho^2)$ such that the vectorfield f_λ has a small periodic orbit of "amplitude" ρ and with period $2\pi(1+\tau^*(\rho^2))^{-1}$. Such bifurcation of periodic orbits from an equilibrium is known as a *Hopf bifurcation*, or *Andronov–Hopf bifurcation*, or even *Poincaré–Andronov–Hopf bifurcation*. The papers [35] and [1] contain some of the original work on this topic; an extensive discussion on the history of Hopf bifurcation together with many more references can be found in Chap. 9 of [14].

Since the eigenvalues $\pm i$ of A_0 were assumed to be simple they can be continued to eigenvalues $\alpha(\lambda) \pm i\beta(\lambda)$ of $A_\lambda = D_x f_\lambda(0)$, for small values of λ. It is not hard to prove that the condition (68) is equivalent to

$$\alpha'(0) \neq 0\,; \tag{69}$$

this transversality condition means that as λ crosses zero the pair of eigenvalues $\alpha(\lambda) \pm i\beta(\lambda)$ of the linearization A_λ at the equilibrium $x = 0$ crosses the imaginary axis with non-zero transversal speed.

Similarly as for equilibria the requirement for Hopf bifurcation to have a pair of *simple* eigenvalues crossing the imaginary axis can sometimes be achieved by restricting (56) to periodic loops with some particular *space-time symmetries*. Working out this idea leads to the *Equivariant Hopf Bifurcation Theorem*. There exists a massive literature developing and applying this theorem; for more details see the basic reference [24].

Lyapunov–Schmidt Method in Infinite Dimensions

The Hopf bifurcation problem as explained in the preceding section leads to an Eq. (56) in the infinite-dimensional orbit space $C^0_{2\pi}$. Also bifurcation problems associated to boundary value problems for partial differential equations typically lead to equations in infinite-dimensional spaces (see e. g. [39] for some examples). The aim of this section is to briefly sketch a basic abstract setting which, under the appropriate conditions, allows to reduce these infinite-dimensional problems to finite-dimensional ones.

Consider an equation

$$F(x, \lambda) = 0\,, \tag{70}$$

with $F\colon X \times \mathbb{R}^k \to Y$ a smooth parameter-dependent mapping between two Banach spaces X and Y. Assuming that $x = 0$ is a solution for $\lambda = 0$, one can ask the question to what extend this solution persists for nearby values of λ, i. e. one wants to find all solutions (x, λ) of (70) near $(0, 0)$. The first step in solving this local problem consists in considering the linearized equation

$$A_0 x = 0\,, \quad \text{with } A_0 = D_x F(0, 0)\,, \tag{71}$$

and to determine the nullspace and range of A_0. The following hypothesis about A_0 allows to proceed with the Lyapunov–Schmidt reduction:

(F) The continuous linear operator $A_0\colon X \to Y$ is a *Fredholm operator*:

- (i) the nullspace $\mathrm{N}(A_0)$ is finite-dimensional;
- (ii) the range $\mathrm{R}(A_0)$ is closed in Y;
- (iii) the range $\mathrm{R}(A_0)$ has finite codimension in Y.

The *index* of such Fredholm operator is given by the number $\dim \mathrm{N}(A_0) - \operatorname{codim}\mathrm{R}(A_0)$; in many applications this index is zero, i. e. $\operatorname{codim}\mathrm{R}(A_0) = \dim \mathrm{N}(A_0)$.

Under the condition (F) there exist continuous linear projections $P\colon X \to X$ and $Q\colon Y \to Y$ such that

$$\mathrm{R}(P) = \mathrm{N}(A_0) \quad \text{and} \quad \mathrm{N}(Q) = \mathrm{R}(A_0)\,, \tag{72}$$

resulting in the topological decompositions $X = \mathrm{N}(A_0) \oplus \mathrm{N}(P) = U \oplus V$ and $Y = \mathrm{R}(Q) \oplus \mathrm{R}(A_0) = W \oplus \mathrm{R}(A_0)$, with $U = \mathrm{N}(A_0)$, $V = \mathrm{N}(P)$ and $W = \mathrm{R}(Q)$. The restriction of A_0 to V is an isomorphism from V onto $\mathrm{R}(A_0)$. Replacing $x \in X$ by $x = Px + (I - P)x = u + v$ in (70), and splitting the equation itself using Q results in the following system equivalent to (70):

$$QF(u + v, \lambda) = 0\,, \quad (I - Q)F(u + v, \lambda) = 0\,. \tag{73}$$

The second of these equations is sometimes called the *auxiliary equation*; the corresponding mapping $F_{\mathrm{aux}}\colon U \times V \times \mathbb{R}^k \to \mathrm{R}(A_0)$ defined by $F_{\mathrm{aux}}(u, v, \lambda) = (I - Q)F(u + v, \lambda)$ is such that $F_{\mathrm{aux}}(0, 0, 0) = 0$, $D_u F_{\mathrm{aux}}(0, 0, 0) = 0$ and $D_v F_{\mathrm{aux}}(0, 0, 0) = (I - Q)A_0\big|_V$. Since this last operator is an isomorphism from V onto $\mathrm{R}(A_0)$ one can apply the implicit function theorem to solve the auxiliary equation in the neighborhood of $(u, v, \lambda) = (0, 0, 0)$ for $v = v^*(u, \lambda)$. The mapping $v^*\colon U \times \mathbb{R}^k \to V$ is smooth, with $v^*(0, 0) = 0$ and $D_u v^*(0, 0) = 0$. Substituting $v = v^*(u, \lambda)$ into the first equation of (73) we obtain the bifurcation equation

$$G(u, \lambda) = QF(u + v^*(u, \lambda), \lambda) = 0\,. \tag{74}$$

The bifurcation mapping $G\colon U \times \mathbb{R}^k \to W$ is smooth, with $G(0, 0) = 0$, $D_u G(0, 0) = 0$ and $D_\lambda G(0, 0) = QD_\lambda F(0, 0)$. The further analysis of the bifurcation Eq. (74) will of course strongly depend on the particularities of the problem, such as the dimensions of U and W, possible symmetries, etc.

A well known and widely applied particular case is the by now classical theorem of Crandall and Rabinowitz [17] on bifurcation from simple eigenvalues. The hypotheses for this result are the following:

(CR) $F\colon X \times \mathbb{R} \to Y$ is smooth, with $F(0, \lambda) = 0$ for all $\lambda \in \mathbb{R}$, and:

- (i) the nullspace of $D_x F(0, 0)$ is one-dimensional;
- (ii) the range of $D_x F(0, 0)$ is closed and has codimension one in Y;
- (iii) $D_\lambda D_x F(0, 0) \cdot u_0 \notin \mathrm{R}(D_x F(0, 0))$, where $u_0 \neq 0$ belongs to $\mathrm{N}(D_x F(0, 0))$.

Under these hypotheses $\dim U = \dim W = 1$ in the foregoing, and therefore $U = \{\rho u_0 \mid \rho \in \mathbb{R}\}$ and $W = \{\eta w_0 \mid \eta \in \mathbb{R}\}$ for some non-zero elements $u_0 \in \mathrm{N}(D_x F(0, 0))$ and $w_0 \notin \mathrm{R}(D_x F(0, 0))$. Also $v^*(0, \lambda) = 0$ and $G(0, \lambda) = 0$

for all $\lambda \in \mathbb{R}$; writing $G(\rho u_0, \lambda) = g(\rho, \lambda) w_0$ it follows that also $g(0, \lambda) = 0$ for all λ. Hence

$$g(\rho, \lambda) = \rho \tilde{g}(\rho, \lambda), \quad \text{with } \tilde{g}(\rho, \lambda) = \int_0^1 \frac{\partial g}{\partial \rho}(s\rho, \lambda) \mathrm{d}s .$$

Clearly $\tilde{g}(0, 0) = 0$, since $\tilde{g}(0, 0) w_0 = \frac{\partial g}{\partial \rho}(0, 0) w_0 = D_u G(0, 0) \cdot u_0$ and $D_u G(0, 0) = 0$. Moreover,

$$\begin{aligned} \frac{\partial \tilde{g}}{\partial \lambda}(0, 0) w_0 &= \frac{\partial^2 g}{\partial \lambda \partial \rho}(0, 0) w_0 = D_\lambda D_u G(0, 0) \cdot u_0 \\ &= Q D_\lambda D_x F(0, 0) \cdot u_0 \neq 0 , \end{aligned}$$

by the hypothesis (CR)-(iii), and so $\frac{\partial \tilde{g}}{\partial \lambda}(0, 0) \neq 0$. Nonzero solutions of the bifurcation equation $g(\rho, \lambda) = 0$ are given by the solutions of $\tilde{g}(\rho, \lambda) = 0$, and using the implicit function theorem this last equation can be solved for $\lambda = \lambda^*(\rho)$, with $|\rho|$ sufficiently small and $\lambda^*(0) = 0$. This gives, under the hypotheses (CR), a unique branch

$$\begin{aligned} \{(x^*(\rho), \lambda^*(\rho)) =| \, 0 < |\rho| < \rho_0\} , \\ \text{with } x^*(\rho) = \rho u_0 + v^*(\rho u_0, \lambda^*(\rho)) , \end{aligned}$$

of nontrivial solutions of (70) bifurcation from the trivial branch $\{(0, \lambda) \mid \lambda \in \mathbb{R}\}$ at $\lambda = 0$.

Outlook

The Liapunov–Schmidt method is a well established reduction technique which can and has been used on a wide variety of problems. The preceding sections just give a small glimpse of its use in local bifurcation theory; the references given in the introduction will point the reader to a whole plethora of other applications. The basics of the method are rather simple, but depending on the mappings and the spaces one is using a number of technical conditions (sometimes easy, sometimes more difficult) need to be verified before the reduction can be carried out. In view of the steady increase in applications which the method has seen over the last half century one can safely say that the Lyapunov–Schmidt method will remain a valuable tool for future research in bifurcation theory, dynamical systems theory and nonlinear analysis.

Bibliography

Primary Literature

1. Andronov A, Leontovich FA (1938) Sur la théorie de la variation de la structure qualitative de la division du plan en trajectoires. Dokl Akad Nauk 21:427–430
2. Antosiewicz HA (1966) Boundary value problems for nonlinear ordinary differential equations. Pacific J Math 17:191–197
3. Ashwin P, Böhmer K, Mei Z (1993) A numerical Lyapunov–Schmidt method for finitely determined systems. In: Allgower E, Georg K, Miranda R (eds) Exploiting Symmetry in Applied and Numerical Analysis. Lectures in Appl Math, vol 29. AMS, Providence, pp 49–69
4. Bancroft S, Hale JK, Sweet D (1968) Alternative problems for nonlinear functional equations. J Differ Equ 4:40–56
5. Bartle R (1953) Singular points of functional equations. Trans Amer Math Soc 75:366–384
6. Böhmer K (1993) On a numerical Lyapunov–Schmidt method for operator equations. Computing 53:237–269
7. Bourgain J (1998) Quasi-periodic Solutions of Hamiltonian Perturbations of 2D Linear Schrödinger Equations. Ann Math 148(2):363–439
8. Cacciopoli R (1936) Sulle correspondenze funzionali inverse diramata: teoria generale e applicazioni ad alcune equazioni funzionali non lineari e al problema di Plateau. Atti Accad Naz Lincei Rend Cl Sci Fis Mat Natur Sez I 24:258–263, 416–421
9. Cesari L (1940) Sulla stabilità delle soluzioni dei sistemi di equazioni differenziali lineari a coefficienti periodici. Atti Accad Ital Mem Cl Fis Mat Nat 6(11):633–692
10. Cesari L (1963) Periodic solutions of hyperbolic partial differential equations. In: Intern. Symposium on Nonlinear Differential Equations and Nonlinear Mechanics. Academic Press, New York, pp 33–57
11. Cesari L (1964) Functional analysis and Galerkin's method. Michigan Math J 11:385–414
12. Cicogna C (1981) Symmetry breakdown from bifurcations. Lett Nuovo Cimento 31:600–602
13. Chossat P, Lauterbach R (2000) Methods in Equivariant Bifurcations and Dynamical Systems. Advanced Series in Nonlinear Dynamics, vol 15. World Scientific, Singapore
14. Chow S-N, Hale JK (1982) Methods of Bifurcation Theory. Grundlehren der Mathematischen Wissenschaften, vol 251. Springer, New York
15. Ciocco M-C, Vanderbauwhede A (2004) Bifurcation of Periodic Points in Reversible Diffeomorphisms. In: Elaydi S, Ladas G, Aulbach B (eds) New Progress in Difference Equations – Proceedings ICDEA2001, Augsburg. Chapman & Hall, CRC Press, pp 75–93
16. Craig W, Wayne CE (1993) Newton's method and periodic solutions of nonlinear wave equations. Comm Pure Appl Math 46:1409–1498
17. Crandall MG, Rabinowitz PH (1971) Bifurcation from simple eigenvalues. J Funct Anal 8:321–340
18. Crandall MG, Rabinowitz PH (1978) The Hopf bifurcation theorem in infinite dimensions. Arch Rat Mech Anal 67:53–72
19. Cronin J (1950) Branch points of solutions of equations in Banach spaces. Trans Amer Math Soc 69:208–231; (1954) 76:207–222
20. Da Prato G, Lunardi A (1986) Hopf bifurcation for fully nonlinear equations in Banach space. Ann Inst Henri Poincaré 3:315–329
21. Friedrichs KO (1953–54) Special Topics in Analysis. Lecture Notes. New York University, New York
22. Gaines R, Mawhin J (1977) Coincidence Degree, and Nonlinear Differential Equations. Lect Notes in Math, vol 568. Springer, Berlin
23. Golubitsky M, Schaeffer DG (1985) Singularities and Groups in Bifurcation Theory, vol 1. Appl Math Sci, vol 51. Springer, New York
24. Golubitsky M, Stewart IN, Schaeffer DG (1988) Singularities and

Groups in Bifurcation Theory, vol 2. Appl Math Sci, vol 69. Springer, New York
25. Golubitsky M, Pivato M, Stewart I (2004) Interior symmetry and local bifurcation in coupled cell networks. Dyn Syst 19:389–407
26. Hale JK (1954) Periodic solutions of non-linear systems of differential equations. Riv Mat Univ Parma 5:281–311
27. Hale JK (1963) Oscillations in nonlinear systems. McGraw-Hill, New York
28. Hale JK (1969) Ordinary Differential Equations. Wiley, New York
29. Hale JK (1977) Theory of Functional Differential Equations. Appl Math Sciences, vol 3. Springer, New York
30. Hale JK (1984) Introduction to dynamic bifurcation. In: Salvadori L (ed) Bifurcation Theory and Applications. Lecture Notes in Math, vol 1057. Springer, Berlin, pp 106–151
31. Hale JK, Sakamoto K (1988) Existence and stability of transition layers. Japan J Appl Math 5:367–405
32. Hale JK, Sakamoto K (2005) A Lyapunov–Schmidt method for transition layers in reaction-diffusion systems. Hiroshima Math J 35:205–249
33. Hall WS (1970) Periodic solutions of a class of weakly nonlinear evolution equations. Arch Rat Mech Anal 39:294–332
34. Hall WS (1970) On the existence of periodic solutions of the equation $D_{tt}u + (-1)^p D_x^{2p} u = \varepsilon f(t, x, u)$. J Differ Equ 7:509–526
35. Hopf E (1942) Abzweigung einer periodischen Lösung von einer stationären Lösung eines Differentialsystems. Ber Math Phys Kl Sachs Acad Wiss Leipzig 94:1–22. English translation in: Marsden J, McCracken M (1976) Hopf Bifurcation and its Applications. Springer, New York
36. Iooss G, Joseph D (1981) Elementary Stability and Bifurcation Theory. Springer, New York
37. Ize J (1976) Bifurcation theory for Fredholm operators. Mem AMS 174
38. Jepson A, Spence A (1989) On a reduction process for nonlinear equations. SIAM J Math Anal 20:39–56
39. Kielhöfer H (2004) Bifurcation Theory, An Introduction with Applications to PDEs. Appl Math Sciences, vol 156. Springer, New York
40. Knobloch J (2000) Lin's method for discrete dynamical systems. J Differ Eq Appl 6:577–623
41. Knobloch J, Vanderbauwhede A (1996) A general reduction method for periodic solutions in conservative and reversible systems. J Dyn Diff Eqn 8:71–102
42. Knobloch J, Vanderbauwhede A (1996) Hopf bifurcation at k-fold resonances in conservative systems. In: Broer HW, van Gils SA, Hoveijn I, Takens F (eds) Nonlinear Dynamical Systems and Chaos. Progress in Nonlinear Differential Equations and Their Applications, vol 19. Birkhäuser, Basel, pp 155–170
43. Krasnosel'skii MA (1963) Topological Methods in the Theory of Nonlinear Integral Equations. Pergamon Press, Oxford
44. Krasnosel'skii MA, Vainiko GM, Zabreiko PP, Rutitskii YB, Stesenko VY (1972) Approximate solutions of operator equations. Wolters-Noordhoff, Groningen
45. Kuznetsov YA (1998) Elements of Applied Bifurcation Theory, 2nd edn. Appl Math Sci, vol 112. Springer, New York
46. Lewis DC (1956) On the role of first integrals in the perturbations of periodic solutions. Ann Math 63:535–548
47. Lin XB (1990) Using Melnikov's method to solve Silnikov's problems. Proc Roy Soc Edinb 116A:295–325
48. Lyapunov AM (1906) Sur les figures d'équilibre peu différentes des ellipsoï, des d'une masse liquide homogène dotées d'un mouvement de rotation. Zap Akad Nauk St Petersb, pp 1–225; (1908), pp 1–175; (1912), pp 1–228; (1914), pp 1–112
49. Mawhin J (1969) Degré topologique et solutions périodiques des systèmes différentiels non linéaires. Bull Soc Roy Sci Liège 38:308–398
50. Nirenberg L (1960–61) Functional Analysis. Lecture Notes. New York University, New York
51. Rabinowitz PH (1967) Periodic solutions of nonlinear hyperbolic partial differential equations. Comm Pure Appl Math 20:145–206
52. Rabinowitz PH (1976) A Survey of Bifurcation Theory. In: Cesari L, Hale JK, LaSalle J (eds) Dynamical Systems, An International Symposium, vol 1. Academic Press, New York, pp 83–96
53. Sandstede B (1993) Verzweigungstheorie homokliner Verdopplungen. Ph D Thesis, University of Stuttgart, IAAS Report No. 7
54. Sattinger DH (1979) Group Theoretic Methods in Bifurcation Theory. Lecture Notes in Math, vol 762. Springer, Berlin
55. Schmidt E (1908) Zur Theorie des linearen und nichtlinearen Integralgleichungen, 3 Teil. Über die Auflösung der nichtlinearen Integralgleichungen und die Verzweigung ihrer Lösungen. Math Ann 65:370–399
56. Shimizu T (1948) Analytic operations and analytic operational equations. Math Japon 1:36–40
57. Sidorov N, Loginov B, Sinitsyn A, Falaleev M (2002) Lyapunov–Schmidt Methods in Nonlinear Analysis and Applications. Mathematics and Its Applications, vol 550. Kluwer, Dordrecht
58. Taniguchi M (1995) A Remark on Singular Perturbation Methods via the Lyapunov–Schmidt Reduction. Publ RIMS, Kyoto Univ 31:1001–1010
59. Vainberg MM, Trenogin VA (1962) The method of Lyapunov and Schmidt in the theory of nonlinear differential equations and their further development. Russ Math Surv 17:1–60
60. Vainberg MM, Trenogin VA (1974) Theory of Branching of Solutions of Non-Linear Equations. Noordhoff, Leyden
61. Vanderbauwhede A (1982) Local bifurcation and symmetry. Research Notes in Mathematics, vol 75. Pitman, London
62. Vanderbauwhede A (2000) Subharmonic bifurcation at multiple resonances. In: Elaydi S, Allen F, Elkhader A, Mughrabi T, Saleh M (eds) Proceedings of the Mathematics Conference, Birzeit, August 1998. World Scientific, Singapore, pp 254–276
63. Vanderbauwhede A, Van der Meer JC (1995) A general reduction method for periodic solutions near equilibria in Hamiltonian systems. In: Langford WF, Nagata W (eds) Normal Forms and Homoclinic Chaos. Fields Institute Comm. AMS, Providence, pp 273–294

Books and Reviews

Alligood KT, Sauer T, Yorke JA (1997) Chaos, An Introduction to Dynamical Systems. Springer, New York

Chicone C (2006) Ordinary Differential Equations with Applications, 2nd edn. Texts in Applied Math, vol 34. Springer, New York

Govaerts W (2000) Numerical Methods for Bifurcations of Dynamical Equilibria. SIAM Publications, Philadelphia

Hale JK, Koçak H (1991) Dynamics and Bifurcations. Texts in Applied Mathematics, vol 3. Springer, New York

Ma T, Wang S (2005) Bifurcation Theory and Applications. World Scientific Series on Nonlinear Science, vol A53. World Scientific, Singapore

M

Machine Learning, Ensemble Methods in

Sašo Džeroski, Panče Panov, Bernard Ženko
Jožef Stefan Institute, Ljubljana, Slovenia

Article Outline

Glossary
Definition of the Subject
Introduction
Learning Ensembles
Frequently Used Ensemble Methods
Future Directions
Bibliography

Glossary

Attribute (*also* feature *or* variable) An attribute is an entity that defines a property of an object (or example). It has a domain defined by its type which denotes the values that can be taken by an attribute (e. g., nominal or numeric). For example, apples can have attributes such as weight (with numeric values) and color (with nominal values such as red or green).

Example (*also* instance *or* case) An example is a single object from a problem domain of interest. In machine learning, examples are typically described by a set of attribute values and are used for learning a descriptive and/or predictive model.

Model (*also* classifier) In machine learning, a model is a computer program that attempts to simulate a particular system or its part with the aim of gaining insight into the operation of this system, or to observe its behavior. Strictly speaking, a classifier is a type of model that performs a mapping from a set of unlabeled examples to a set of (discrete) classes. However, in machine learning the term classifier is often used as a synonym for model.

Learning (*also* training) set A learning set is a set of examples that are used for learning a model or a classifier. Examples are typically described in terms of attribute values and have a corresponding output value or class.

Testing set A testing set is a set of examples that, as opposed to examples from the learning set, have not been used in the process of model learning; they are also called unseen examples. They are used for evaluating the learned model.

Ensemble An ensemble in machine learning is a set of predictive models whose predictions are combined into a single prediction. The purpose of learning ensembles is typically to achieve better predictive performance.

Definition of the Subject

Ensemble methods are machine learning methods that construct a set of predictive models and combine their outputs into a single prediction. The purpose of combining several models together is to achieve better predictive performance, and it has been shown in a number of cases that ensembles can be more accurate than single models. While some work on ensemble methods has already been done in the 1970s, it was not until the 1990s, and the introduction of methods such as *bagging* and *boosting*, that ensemble methods started to be more widely used. Today, they represent a standard machine learning method which has to be considered whenever good predictive accuracy is demanded.

Introduction

Most machine learning techniques deal with the problem of learning predictive models of data. The data are usually given as a set of examples where examples represent objects or measurements. Each example can be described in terms of values of several (independent) variables, which are also referred to as attributes, features, inputs or predictors (for example, when talking about cars, possible attributes include the manufacturer, number of

seats, horsepower of a car, etc.). Associated with each example is a value of a dependent variable, also referred to as class, output or outcome. The class is some property of special interest (such as the price of the car). The typical machine learning task is to learn a model using a learning data set with the aim of predicting the value of the class for unseen examples (in our car example this would mean that we want to predict the price of a specific car based on its properties). There exist a number of methods, developed within machine learning and statistics, that solve this task more or less successfully (cf., [21,31,43]). Sometimes, however, the performance obtained by these methods (we will call them simple or base methods) is not sufficient.

One of the possibilities to improve predictive performance are ensemble methods, which in the literature are also referred to as multiple classifier systems, committees of classifiers, classifier fusion, combination or aggregation. The main idea is that, just as people often consult several sources when making an important decision, the machine learning model that takes into account several aspects of the problem (or several submodels) should be able to make better predictions. This idea goes in line with the principle of multiple explanations first proposed by the Greek philosopher Epicurus (cf., [28]), which says that for an optimal solution of a concrete problem we have to take into consideration all the hypotheses that are consistent with the input data. Indeed, it has been shown that in a number of cases ensemble methods offer better predictive performance than single models. The performance improvement comes at a price, though. When we humans want to make an informed decision we have to make an extra effort, first to find additional viewpoints on the subject, and second, to compile all this information into a meaningful final decision. The same holds true for ensemble methods; learning the entire set of models and then combining their predictions is computationally more expensive than learning just one simple model. Let us present some of the reasons why ensemble methods might still be preferred over simple methods [33].

Statistical Reasons

As already mentioned, we learn a model on the learning data, and the resulting model can have more or less good predictive performance on these learning data. However, even if this performance is good, this does not guarantee good performance also on the unseen data. Therefore, when learning single models, we can easily end up with a bad model (although there are evaluation techniques that minimize this risk). By taking into account several models and averaging their predictions we can reduce the risk of selecting a very bad model.

Very Large or Very Small Data Sets

There exist problem domains where the data sets are so large that it is not feasible to learn a model on the entire data set. An alternative and sometimes more efficient approach is to partition the data into smaller parts, learn one model for each part, and combine the outputs of these models into a single prediction.

On the other hand, there exist also many domains where the data sets are very small. As a result, the learned model can be unstable, i. e., it can drastically change if we add or remove just one or two examples. A possible remedy to this problem is to draw several overlapping subsamples from the original data, learn one model for each subsample, and then combine their outputs.

Complex Problem Domains

Sometimes, the problem domain we are modeling is just too complex to be learned by a single learning method. For illustration only, let us assume we are trying to learn a model to discriminate between examples with class '+' and examples with class '–', and the boundary between the two is a circle. If we try to solve this problem using a method that can learn only linear boundaries we will not be able to find an adequate solution. However, if we learn a set of models where each model approximates only a small part of the circular boundary, and then combine these models in an appropriate way, the problem can be solved even with a linear method.

Heterogeneous Data Sources

In some cases, we have data sets from different sources where the same type of objects are described in terms of different attributes. For example, let us assume we have a set of treated cancer patients for which we want to predict whether they will have a relapse or not. For each patient different tests can be performed, such as gene expression analyzes, blood tests, CAT scans, etc., and each of these tests results in a data set with different attributes. It is very difficult to learn a single model with all these attributes. However, we can train a separate model for each test and then combine them. In this way, we can also emphasize the importance of a given test, if we know, for example, that it is more reliable than the others.

In the remainder of the article we first describe the process of learning ensembles and then give an overview of some of the commonly used methods. We conclude with

a discussion on potential impacts of ensemble methods on the development of other science areas.

Learning Ensembles

Ensembles of models are sets of (simple) models whose outputs are combined, for instance with majority voting, into a single output or prediction. The problem of learning ensembles attracts a lot of attention in the machine learning community [10], since it is often the case that predictive accuracy of ensembles is better than that of their constituent (base) models. This has also been confirmed by several empirical studies [2,11,15] for both classification (predicting a nominal variable) and regression (predicting a numeric variable) problems. In addition, several theoretical explanations have been proposed to justify the effectiveness of some commonly used ensemble methods [1,27,38].

The learning of ensembles consists of two steps. In the first step we have to learn the base models that make up the ensemble. In the second step we have to figure out how to combine these models (or their predictions) into a single coherent model (or prediction). We will now look more closely into these two steps.

Generating Base Models

When learning base models it makes sense to learn models that are diverse. Combining identical or very similar models clearly does not improve the predictive accuracy of base models. Moreover, it only increases the computational cost of the final model. By diverse models we mean models that make errors on different learning examples, so that when we combine their predictions in some smart way, the resulting prediction will be more accurate. Based on this intuition, many diversity measures have been developed with the purpose of evaluating and guiding the construction of ensembles. However, despite considerable research in this area, it is still not clear whether any of these measures can be used as a practical tool for constructing better ensembles [30]. Instead, several more or less ad hoc approaches are used for generating diverse models. We can group these approaches roughly into two groups. In the first case, the diversity of models is achieved by modifying the learning data, while in the second case, diverse models are learned by changing the learning algorithm.

The majority of ensemble research has focused on methods from the first group, i. e., methods that use different learning data sets. Such data sets can be obtained by resampling techniques such as bootstrapping [14], where learning sets are drawn randomly with replacement from the initial learning data set; this is the approach used in bagging [3] and random forests [5]. An alternative approach is used in boosting [37]. Here we start with a model that is learned on the initial data set. We identify learning examples for which this model performs well. Now we decrease the weights of these examples, since we wish for the next members of the ensemble to focus on examples misclassified by the first model. We iteratively repeat this procedure until enough base models are learned. Yet another approach to learn diverse base models is taken by the random subspaces method [22] where, instead of manipulating examples in the learning set, we each time randomly select a subset of attributes used for describing the learning set examples. These methods are typically coupled with unstable learning algorithms such as decision trees [6] or neural networks [36], for which even a small change in the learning set can produce a significantly different model.

Ensemble methods from the second group, which use different learning algorithms, use two major approaches for achieving diversity. First, if we use a base learning algorithm that depends on some parameters, diverse models can be learned by changing the values of these parameters. Again, because of their instability, decision trees and neural networks are most often employed here. A special case are randomized learning algorithms, where the outcome of learning depends on a seed used for the randomization. The second possibility is to learn each base model with a completely different learning algorithm altogether; for example, we could combine decision trees, neural networks, support vector machines and naive Bayes models into a single ensemble; this approach is used in stacking [44].

Combining Base Models

Once we have generated a sufficiently diverse set of base models, we have to combine them so that a single prediction can be obtained from the ensemble. In general, we have two options, model *selection* or model *fusion* (please note that in the literature a somewhat different definition of these two terms is sometimes used, e. g., [29]). In model selection, we evaluate the performance of all base models, and simply use predictions of the best one as predictions of the ensemble. This approach cannot be strictly regarded as an ensemble method since in the end we are using only one base model for prediction. On one hand, this can be seen as an advantage from the viewpoint that the final model is simpler, more understandable and can be executed fast. On the other hand, it is obvious that the performance of such an ensemble cannot be better than the performance of the best base model. While this seems like a serious drawback it turns out that constructing ensembles that are

more accurate than a selected best base model can be a very hard task [13].

In model fusion, we really combine the predictions of all base models into a prediction of the ensemble. By far the most common method for combining predictions is voting; it is used in bagging [3], boosting [37], random forests [5] and many variations of these methods. Voting is a relatively simple combining scheme and can be applied to predictions with nominal or numeric values, or probability distributions over these. A different approach is adopted in stacking [44]. As opposed to voting, where the combining scheme is known in advance and is fixed, stacking tries to learn a so called *meta model* in order to combine base predictions as efficiently as possible. The meta model is learned on data where examples are described in terms of the predictions of the base models and the dependent variable is the final prediction of the ensemble. There are, of course, many other possibilities for combining models, including custom combining schemes specifically tailored for a given problem domain. In the next section we describe some of the most frequently used ensemble methods in more detail.

Frequently Used Ensemble Methods

The use of different schemes for base models generation and their combination, as briefly mentioned in the previous section, gives rise to a large number of possible ensemble methods. We describe here a few of them that are most common, with the exception of the best base model selection approach, which is very straightforward and does not need an additional description.

Voting

Strictly speaking, voting is not an ensemble method, but a method for combining base models, i. e., it is not concerned with the generation of the base models. Still, we include it in this selection of ensemble methods because it can be used for combining models regardless of how these models have been constructed. As mentioned before, voting combines the predictions of base models according to a static voting scheme, which does not depend on the learning data or on the base models. It corresponds to taking a linear combination of the models. The simplest type of voting is the plurality vote (also called majority vote), where each base model casts a vote for its prediction. The prediction that collects most votes is the final prediction of the ensemble. If we are predicting a numeric value, the ensemble prediction is the average of the predictions of the base models.

A more general voting scheme is weighted voting, where different base models can have different influence on the final prediction. Assuming we have some information on the quality of the base models' predictions (provided by the models themselves or through some background knowledge), we can put more weight on the predictions coming from more trustworthy models. Weighted voting predicting nominal values simply means that vote of each base model is multiplied by its weight and the value with the most weighted votes becomes the final prediction of the ensemble. For predicting numeric values we use a weighted average. If d_i and w_i are the prediction of the ith model and its weight, the final prediction is calculated as $Y = \sum_{i=1}^{b} w_i d_i$. Usually we demand that the weights are nonnegative and normalized: $w_i \geq 0\,, \forall i\,; \sum_{i=1}^{b} w_i = 1$.

Another interesting aspect of voting is that, because of its simplicity, it allows for some theoretical analyzes of its efficiency. For example, when modeling a binary problem (a problem with two possible values, e. g., *positive* and *negative*) it has been shown that, if we have an ensemble with independent base models each with success probability (accuracy) greater than 1/2, i. e., better than random guessing, the accuracy of the ensemble increases as the number of base models increases (cf., [20,37,41]).

Bagging

Bagging (short for bootstrap aggregation) [3] is a voting method where base models are learned on different variants of the learning data set which are generated with bootstrapping (bootstrap sampling) [14]. Bootstrapping is a technique for sampling with replacement; from the initial learning data set we randomly select examples for a new learning (sub)set, where each example can be selected more than once. If we generate a set with the same number of examples as the original learning set, the new one will on average contain only 63.2% different examples from the original set, while the remaining 36.8% will be multiple copies. This technique is often used for estimating properties of a variable, such as its variance, by measuring those properties on the samples obtained in this manner.

Using these sampled sets, a collection of base models is learned and their predictions are combined by simple majority voting. Such an ensemble often gives better results than its individual base models because it combines the advantages of individual models. Bagging has to be used together with an unstable learning algorithm (e. g., decision trees or neural networks), where small changes in the learning set result in largely different classifiers. Another benefit of the sampling technique is that it is less likely

Input: Learning set S, Ensemble size B
Output: Ensemble E

$E = \emptyset$
for $i = 1$ to B **do**
  $S^i = \text{BootstrapSample}(S)$
  $C^i = \text{ConstructBaseModel}(S^i)$
  $E = E \cup \{C^i\}$
end for
return E

Machine Learning, Ensemble Methods in, Algorithm 1
Learning ensembles with bagging

that (many) outliers in the learning set show up also in the bootstrap sample. As a result, base models and the ensemble as a whole should be less sensitive to data outliers. The bagging algorithm is presented in Algorithm 1. Bagging can be used both for classification and regression problems. In the case of regression the individual predictions are combined by averaging.

Boosting

Boosting [15] comprises a whole family of similar methods that, just as bagging, use voting to combine the predictions of base models learned by a single learning algorithm. The difference between the two approaches is that in bagging the complementarity of the constructed base models is left to chance, while in boosting we try to generate complementary base models by learning subsequent models, taking into account the mistakes of previous models. The procedure starts by learning the first base model on the entire learning set with equally weighted examples. For the next base models, we want them to correctly predict the examples that have not been correctly predicted by previous base models. Therefore, we increase the weights of these examples (or decrease the weights of the correctly predicted examples) and learn a new base model. We stop learning new base models when some stopping criterion is satisfied (like when the accuracy of the new base model is less then or equal to 0.5). The prediction of the ensemble is obtained by weighted voting, where more weight is given to more accurate base models; the weights of all classifiers that vote for a specific class are summed and the class with the highest total vote is predicted.

An interesting property of some boosting methods is that they provide a theoretical guarantee of the accuracy [15,26]. We can show that the predictive error of the ensemble on the learning data quickly decreases as we increase the number of base models within the ensemble. The only precondition for error decrease is that the error of the individual members of the ensemble is less than 0.5. For binary classification problems this condition is usually easy to fulfill. While the guarantee of a small error on the learning set is not a guarantee of a small error on unseen examples, boosting methods are known to frequently improve the predictive performance of the base algorithms [39]. Just as bagging, boosting should also be used together with unstable learning methods such as decision trees or neural networks. The most widely used boosting method is AdaBoost.M1 [15] presented in Algorithm 2 (together with the exact example reweighting scheme used in this algorithm), which was designed for learning with binary classification problems. Nevertheless, there exist also modifications of the original method that work on classification problems with more than two possible values (multiclass) [40] and even on regression problems [34,35]. An alternative name often used for boosting methods is *arcing* (adaptively resample and combine) [4], although strictly speaking, boosting methods are a subset of arcing methods, i. e., boosting methods are the ones for which it can be shown that they can achieve an arbitrarily small error on the learning data set.

Input: Learning set S, Ensemble size B
Output: Ensemble E

$E = \emptyset$
$W = \text{AssignEqualWeights}(S)$
for $i = 1$ to B **do**
  $C^i = \text{ConstructModel}(S, W)$
  $Err = \text{ApplyModel}(C^i, S)$
  if $(Err = 0) \vee (Err \geq 0.5)$ **then**
    TerminateModelGeneration
    return E
  end if
  for $j = 1$ to $\text{NumberOfExamples}(S)$ **do**
    if $\text{CorrectlyClassified}(S_j, C^i)$ **then**
      $W_j = W_j \frac{Err}{1-Err}$
    end if
  end for
  $W = \text{NormalizeWeights}(W)$
  $E = E \cup \{C^i\}$
end for
return E

Machine Learning, Ensemble Methods in, Algorithm 2
The AdaBoost.M1 algorithm for learning ensembles with boosting

Input: Learning set S, Ensemble size B,
Proportion of attributes considered f
Output: Ensemble E

$E = \emptyset$
for $i = 1$ to B **do**
$S^i = \text{BootstrapSample}(S)$
$C^i = \text{BuildRandomTreeModel}(S^i, f)$
$E = E \cup \{C^i\}$
end for
return E

Machine Learning, Ensemble Methods in, Algorithm 3
Learning random forests

Random Forests

Random forests [5] is a method for combining models learned with a randomized version of a decision tree algorithm. Random forests can be seen as an implementation of bagging in which each model is learned with a modified version of the CART decision tree algorithm [6]; namely, when searching for an optimal attribute split in a tree, rather than considering all possible splits, only a small subset of randomly selected splits is tested (i. e., a random subset of attributes), and the best one is chosen from this subset. There are two sources of diversity when learning the trees, and both are random: the selection of a bootstrap sample for learning each tree, and the selection of attributes to on which to split at every node of the tree. Random forests are a robust and typically very accurate ensemble method applicable to classification and regression problems. The algorithm for learning random forests is presented in Algorithm 3.

Stacking

Stacking or stacked generalization [44] is a method for combining heterogeneous base models, i. e., models learned with different learning algorithms such as the nearest neighbor method, decision trees, naive Bayes, etc. Base models are not combined with a fixed scheme such as voting, but rather an additional model called *meta* (or *level 1*) model is learned and used for combining base (or *level 0*) models. The procedure has two steps. First, we generate the meta learning data set using the predictions of the base models. Second, using the meta learning set we learn the meta model which can combine predictions of base models into a final prediction.

Let $L_1, \ldots, L_N$ be the base learning algorithms, and S be the learning data set, which consists of examples $\mathbf{s}_i = (\mathbf{x}_i, y_i)$, i. e., pairs of attribute vectors $\mathbf{x}_i$ and their classifications y_i. Generation of the meta learning data set is done using a leave-one-out, or in general, a K-fold cross-validation procedure. The initial learning set S with n examples is split into K proper subsets S_k of roughly equal size and class value distribution. For each of the subsets a group of base models $C_1^k, C_2^k, \ldots, C_N^k$ is learned ($C_j^k = L_j(S - S_k), \forall j = 1, \ldots, N, \forall k = 1, \ldots, K$). These models are now used for predicting examples that were not included in their learning set: $\hat{y}_i^j = C_j^k(x_i)$, $x_i \in S_k$. These predictions are collected into a meta learning set S^m. Each example from the original learning set S has a corresponding example in S^m of the form $\mathbf{s}_i^m = (\hat{\mathbf{y}}_i, y_i) = ((\hat{y}_i^1, \ldots, \hat{y}_i^N), y_i)$. The attributes of the meta learning set are therefore the predictions of the base models ($\hat{y}_i^j$), while the class value is the true class value from the original data set (y_i). In the second step, a meta learning algorithm L^m is applied to this meta learning set. When predicting a value of an unseen example, we first collect the predictions of the base models which are then given to the meta model that combines them into a final prediction. The stacking algorithm is presented in Algorithm 4. The performance of stacking highly depends on the attributes used in the meta learning set (we have only described the simplest option above) and the meta learning algorithm used for learning the meta model (cf. [13,42]).

Random Subspace Method

The random subspace method (RSM) [22] is an ensemble method somewhat similar to bagging. However, while in bagging the diversity of base models is achieved by sampling examples from the initial learning data set, in RSM the diversity is achieved by sampling attributes from the learning set. Let each learning example X_i in the learning set S be a p-dimensional vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$. RSM randomly selects p^* attributes from S, where $p^* < p$. By this, we obtain the p^* dimensional random subspace of the original p-dimensional attribute space. Therefore, the modified training set $\tilde{S} = (\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_n)$ consists of p^*-dimensional learning examples $\tilde{\mathbf{x}}_i = (x_{j1}, x_{i2}, \ldots, x_{ip^*})$ $(i = 1, 2, \ldots, n)$. Afterwards, base models are learned from the random subspaces $\tilde{S}^j$ (of the same size), $j = 1, 2, \ldots, B$, and they are combined by voting to obtain a final prediction. Typically, p^* is equal for all base models. The RSM algorithm is presented in Algorithm 5.

The RSM benefits from using random subspaces for learning base models and from their aggregation. When the number of learning examples is relatively small as com-

Input: Learning set S, Number of folds for meta data generation K, Base and meta learning algorithms $\{L_1, L_2, \ldots, L_N\}$, L^m
Output: Ensemble E

$E = \emptyset$
$\{S_1, S_2, \ldots, S_K\} = \text{SplitData}(S, K)$
$S^m = \emptyset$
for $k = 1$ to K **do**
  for $j = 1$ to N **do**
    $C_j^k = L_j(S - S_k)$
  end for
  $S_k^m = \bigcup_{x_i \in S_k} \{(C_1^k(x_i), C_2^k(x_i), \ldots, C_N^k(x_i), y_i)\}$
end for
$S^m = \bigcup_{k=1}^{K} S_k^m$
$C^m = L^m(S^m)$
$\{C_1, C_2, \ldots, C_N\} = \{L_1(S), L_2(S), \ldots, L_N(S)\}$
$E = (\{C_1, C_2, \ldots, C_N\}, C_m)$
return E

Machine Learning, Ensemble Methods in, Algorithm 4
Learning ensembles with stacking

Input: Training examples S, Number of subspaces B, Dimension of subspaces p^*
Output: Ensemble E

$E = \emptyset$
for $j = 1$ to B **do**
  $\tilde{S}^j = \text{SelectRandomSubspace}(S, p^*)$
  $C_j = \text{ConstructModel}(\tilde{S}^j)$
  $E = E \cup \{C_j\}$
end for
return E

Machine Learning, Ensemble Methods in, Algorithm 5
Learning ensembles with the random subspace method

pared to the dimensionality of the data, learning models in random subspaces alone may solve the small sample problem. In this case the subspace dimensionality is smaller than in the original attribute space, while the number of learning objects remains the same. When the data set has many redundant attributes, one may obtain better models in random subspaces than in the original attribute space. The combined decision of such models may be superior to a single model constructed on the original learning set in the complete attribute space.

The RSM was originally developed to be used with decision trees, but the methodology can also be used to improve the performance of other unstable learning methods (e. g., rule sets, neural networks, etc.). The RSM is expected to perform well when there is a certain redundancy in the data attribute space [22]. It has been noticed that the performance of the RSM is affected by the problem complexity (attribute efficiency, length of class boundary, etc.) [23]. When applied to decision trees, the RSM is superior to a single decision tree and may outperform both bagging and boosting [22].

Other Methods

Mixture of Experts Models The combination of the base learners can be governed by a supervisor learner, that selects the most appropriate element of the ensemble on the basis of the available input data. This idea led to the mixture of experts methods [24], where a gating network performs the division of the input space and small neural networks perform the effective calculation at each assigned region separately. An extension of this approach is the hierarchical mixture of experts method, where the outputs of the different experts are non-linearly combined by different supervisor gating networks hierarchically organized [25]. Cohen and Intrator extended the idea of constructing local simple base learners for different regions of the input space, searching for appropriate architectures that should be locally used and for a criterion to select a proper unit for each region of input space [8,9].

Error Correcting Output Codes Error-correcting output codes (ECOC) [12] is an ensemble method for im-

proving the performance of classification algorithms in multiclass learning problems. Let us note that some machine learning algorithms (e. g., standard support vector machines) work only with two class problems. In order to apply such algorithms to a multiclass problem it has to be decomposed into several independent two-class problems; the algorithm is run on each of them and the outputs of the resulting binary models are combined. The error-correcting output codes method enables us to efficiently combine the outputs of such models.

As already mentioned, we have binary base models with possible outputs (-1 or $+1$), and there exists a code matrix W of size $K*B$ whose K rows are the binary codes of classes in terms of B base models C_j. This code matrix allows us to define a multiclass classification problem in terms of two-class classification problems. The problem here is that if there is an error with one of the base models, there will be a misclassification because the class code words are so similar. The ECOC approach sets the B beforehand and then tries to find such a code matrix W that the distances between rows, and at the same time the distances between the columns, are as large as possible in terms of the Hamming distance [19]. The ECOC can be written as a voting scheme where the entries of W, w_{ij} are considered as vote weights $y_i = \sum_{j=}^{B} w_{ij} d_j$. As a final prediction the class with the highest y_i is chosen.

Future Directions

Recent and future research directions in ensemble methods that are likely to have high impact on data mining and other areas of science focus along the following topics: Combinations of different sources of diversity; Understanding and interpretation of ensembles; Understanding and explaining in more basic terms why ensembles perform better that individual models.

Random forests [5], one of the most successful ensemble approaches, combine two sources of diversity of the base models: Variations in the learning data set (achieved through different bootstrap samples, as in bagging) and a randomized base-level learning algorithm. Another recent approach [32] combines the bagging way of sampling with the random subspaces way of randomly selecting subsets of the original set of attributes. This approach has the advantage of being applicable in conjunction with a variety of base-level learning algorithms that do not need to be randomized.

We have provided some intuition of why ensembles work better than individual models in terms of the diversity of the base models. More fundamental explanations are produced in the bias-variance analysis framework: Roughly speaking, the error of a learning algorithm can be divided into a part due to the functional form used by the algorithm (bias) and a part that is due to the instability of the algorithms (variance). Bagging and random forests reduce the variance part. Boosting reduces mainly the bias part, but also the variance part. Finally, boosting can also be viewed as an incremental forward stagewise regression procedure with regularization (Lasso penalty), which maximizes the margin between the two classes, much like the approach of support vector machines [18].

While ensembles typically perform better than a single model, they do have an important disadvantage: They are more complex and difficult (if not impossible) to interpret. Recent research has addressed this issue in several ways. Some approaches produce an estimate of the relative importance of the attributes as an explanation, for example partial dependency plots [16] and the attribute ranking approach based on bagging and random forests. The approach of Caruana [7] is to construct a single model that approximates the behavior of the ensemble: This is done by generating examples, classifying them with the ensemble, and learning a single model from the resulting learning set. Finally, a recent approach [17] builds rule ensembles, where small (and understandable) sets of rules are preferred through regularization.

Bibliography

Primary Literature

1. Allwein EL, Schapire RE, Singer Y (2000) Reducing multiclass to binary: a unifying approach for margin classifiers. J Mach Learn Res 1:113–141
2. Bauer E, Kohavi R (1999) An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Mach Learn 36(1–2):105–139
3. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
4. Breiman L (1998) Arcing classifiers. Ann Stat 26(3):801–849
5. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
6. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth and Brooks, Monterey
7. Buciluǎ C, Caruana R, Niculescu-Mizil A (2006) Model compression. In: Proc. of the 12th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '06). ACM, New York, pp 535–541
8. Cohen S, Intrator N (2000) A hybrid projection based and radial basis function architecture. In: Proc. of the 1st international workshop on multiple classifier systems (MCS '00). Springer, Berlin, pp 147–156
9. Cohen S, Intrator N (2001) Automatic model selection in a hybrid perceptron/radial network. In: Proc. of the 2nd international workshop on multiple classifier systems (MCS '01). Springer, Berlin, pp 440–454
10. Dietterich TG (1997) Machine-learning research: four current directions. AI Mag 18(4):97–136

11. Dietterich TG (2000) Ensemble methods in machine learning. In: Proc. of the 1st international workshop on multiple classifier systems (MCS '00). Springer, Berlin, pp 1–15
12. Dietterich TG, Bakiri G (1995) Solving multiclass learning problems via error-correcting output codes. J Artif Intell Res 2:263–286
13. Džeroski S, Ženko B (2004) Is combining classifiers with stacking better than selecting the best one? Mach Learn 54(3):255–273
14. Efron B (1979) Bootstrap methods: Another look at the jackknife. Ann Stat 7(1):1–26
15. Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: Saitta L (ed) Machine learning: Proc. of the 13th international conference (ICML '96). Morgan Kaufmann, San Francisco, pp 148–156
16. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):11189–1232
17. Friedman JH, Popescu BE (2005) Predictive learning via rule ensembles. Technical report, Stanford University, Department of Statistics
18. Friedman JH, Hastie T, Tibshirani RJ (1998) Additive logistic regression: a statistical view of boosting. Technical report, Stanford University, Department of Statistics
19. Hamming RW (1950) Error detecting and error correcting codes. Bell Syst Tech J 26(2):147–160
20. Hansen LK, Salamon P (1990) Neural network ensembles. IEEE Trans Pattern Anal Mach Intell 12(10):993–1001
21. Hastie T, Tibshirani RJ, Friedman JH (2001) The elements of statistical learning. Springer Series in Statistics. Springer, Berlin
22. Ho TK (1998) The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 20(8):832–844
23. Ho TK (2000) Complexity of classification problems and comparative advantages of combined classifiers. In: Kittler J, Roli F (eds) Proc. of the 1st international workshop on multiple classifier systems (MCS '00), vol 1857. Springer, Berlin, pp 97–106
24. Jacobs RA (1995) Methods for combining experts' probability assessments. Neural Comput 7(5):867–888
25. Jordan MI, Jacobs RA (1992) Hierarchies of adaptive experts. In: Moody JE, Hanson S, Lippmann RP (eds) Advances in Neural Information Processing System (NIPS). Morgan Kaufmann, San Mateo, pp 985–992
26. Kearns MJ, Vazirani UV (1994) An introduction to computational learning theory. MIT Press, Cambridge
27. Kittler J, Hatef M, Duin RPW, Matas J (1998) On combining classifiers. IEEE Trans Pattern Anal Mach Intell 20(3):226–239
28. Kononenko I, Kukar M (2007) Machine learning and data mining: introduction to principles and algorithms. Horwood, Chichester
29. Kuncheva LI (2004) Combining pattern classifiers: methods and algorithms. Wiley-Interscience, Hoboken
30. Kuncheva LI, Whitaker CJ (2003) Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach Learn 51(2):181–207
31. Mitchell T (1997) Machine Learning. McGraw-Hill, New York
32. Panov P, Džeroski S (2007) Combining bagging and random subspaces to create better ensembles. In: Proc. of 7th international symposium on intelligent data analysis (IDA '07), vol 4723. Lecture notes in computer science. Springer, Berlin, pp 118–129
33. Polikar R (2006) Ensemble based systems in decision making. IEEE Circuits Syst Mag 6(3):21–45
34. Rätsch G, Demiriz A, Bennett KP (2002) Sparse regression ensembles in infinite and finite hypothesis spaces. Mach Learn 48(1–3):189–218
35. Ridgeway G, Madigan D, Richardson T (1999) Boosting methodology for regression problems. In: Heckerman D, Whittaker J (eds) Proc. of the 7th international workshop on artificial intelligence and statistics. Morgan Kaufmann, San Francisco, pp 152–161
36. Rosenblatt F (1962) Principles of neurodynamics: perceptron and the theory of brain mechanisms. Spartan Books, Washington
37. Schapire RE (1990) The strength of weak learnability. Mach Learn 5(2):197–227
38. Schapire RE (1999) A brief introduction to boosting. In: Proc. of the 6th international joint conference on artificial intelligence. Morgan Kaufmann, San Francisco, pp 1401–1406
39. Schapire RE (2001) The boosting approach to machine learning: an overview. In: MSRI workshop on nonlinear estimation and classification, Berkeley, CA, 2001
40. Schapire RE, Singer Y (1999) Improved boosting using confidence-rated predictions. Mach Learn 37(3):297–336
41. Schapire RE, Freund Y, Bartlett P, Lee WS (1998) Boosting the margin: a new explanation for the effectiveness of voting methods. Ann Stat 26(5):1651–1686
42. Ting KM, Witten IH (1999) Issues in stacked generalization. J Artif Intell Res 10:271–289
43. Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco
44. Wolpert DH (1992) Stacked generalization. Neural Netw 5(2):241–259

Books and Reviews

Brown G Ensemble learning bibliography. http://www.cs.man.ac.uk/~gbrown/ensemblebib/index.php. Accessed 26 March 2008

Weka 3: Data mining software in Java. http://www.cs.waikato.ac.nz/ml/weka/. Accessed 26 March 2008

Macroeconomics, Non-linear Time Series in

James Morley
Washington University, St. Louis, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Types of Nonlinear Models
Business Cycle Asymmetry
Future Directions
Bibliography

Glossary

Nonlinear time series in macroeconomics A field of study in economics pertaining to the use of statistical analysis of data in order to make inferences about nonlinearities in the nature of aggregate phenomena in the economy.

Time series A collection of data corresponding to the values of a variable at different points of time.

Linear Refers to a class of models for which the dependence between two random variables can be completely described by a fixed correlation parameter.

Nonlinear Refers to the class of models for which the dependence between two random variables has a more general functional form than a linear equation and/or can change over time.

Structural change A change in the model describing a time series, with no expected reversal of the change.

Level Refers to a definition of the business cycle that links the cycle to alternation between phases of expansion and recession in the level of economic activity.

Deviations Refers to a definition of the business cycle that links the cycle to transitory deviations of economic activity from a trend level.

Fluctuations Refers to a definition of the business cycle that links the cycle to any short-run changes in economic activity.

Deepness A characteristic of a process with a skewed unconditional distribution.

Steepness A characteristic of a process with a skewed unconditional distribution for its first-differences.

Sharpness A characteristic of a process for which the probability of a peak when increasing is different than the probability of a trough when decreasing.

Time reversibility The ability to substitute $-t$ and t in the equations of motion for a process without changing the process.

Markov-switching models Models that assume the prevailing regime governing the conditional distribution of a variable or variables being modeled depends on an unobserved discrete Markov process.

Self-exciting threshold models Models that assume the prevailing regime governing the conditional distribution of a variable or variables being modeled is observable and depends on whether realized values of the time series being modeled exceed or fall below certain "threshold" values.

Nuisance parameters Parameters that are not of direct interest in a test, but influence the distribution of a test statistic.

Pivotal Refers to the invariance of the distribution of a test statistic with respect to values of parameters in the data generating process under the null hypothesis.

Size Probability of false rejection of a null hypothesis in repeated experiments.

Power Probability of correct rejection of a null hypothesis in repeated experiments.

Definition of the Subject

Nonlinear time series in macroeconomics is a broad field of study in economics. It refers to the use of statistical analysis of data to make inferences about nonlinearities in the nature of aggregate phenomena in the economy. This analysis is relevant for forecasting, the formulation of economic policy, and the development and testing of macroeconomic theories.

Introduction

In macroeconomics, the primary aggregate phenomenon is the flow of total production for the entire economy over the course of a year, which is measured by real gross domestic product (GDP). A collection of data corresponding to the values of a variable such as real GDP at different points of time is referred to as a *time series*. Figure 1 presents the time series for US real GDP for each year from 1929 to 2006.

Time series analysis employs stochastic processes to explain and predict the evolution of a time series. In particular, a process captures the idea that different observations are in some way related to each other. The relationship can simply be that the observations behave as if they are drawn from random variables with the same distribution. Or the relationship can be that the distribution assumed to generate one observation depends on the values of other observations. Either way, a relationship implies that the observations can be used jointly to make inferences about the parameters describing the distributions (a.k.a. "estimation").

Within the context of time series in macroeconomics, the terms "linear" and "nonlinear" typically refer to classes of models for processes, although other meanings arise in the literature. For the purposes of this survey, a model that assumes the dependence between two random variables in a process can be completely captured by a fixed correlation parameter is said to be *linear*. A very basic example of a linear time series model is the workhorse first-order autoregressive (AR(1)) model:

$$y_t = c + \phi y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{ i.i.d. } (0, \sigma^2)\,, \tag{1}$$

where $|\phi| < 1$. In words, the random variable y_t that generates the observation in period t is a linear function of

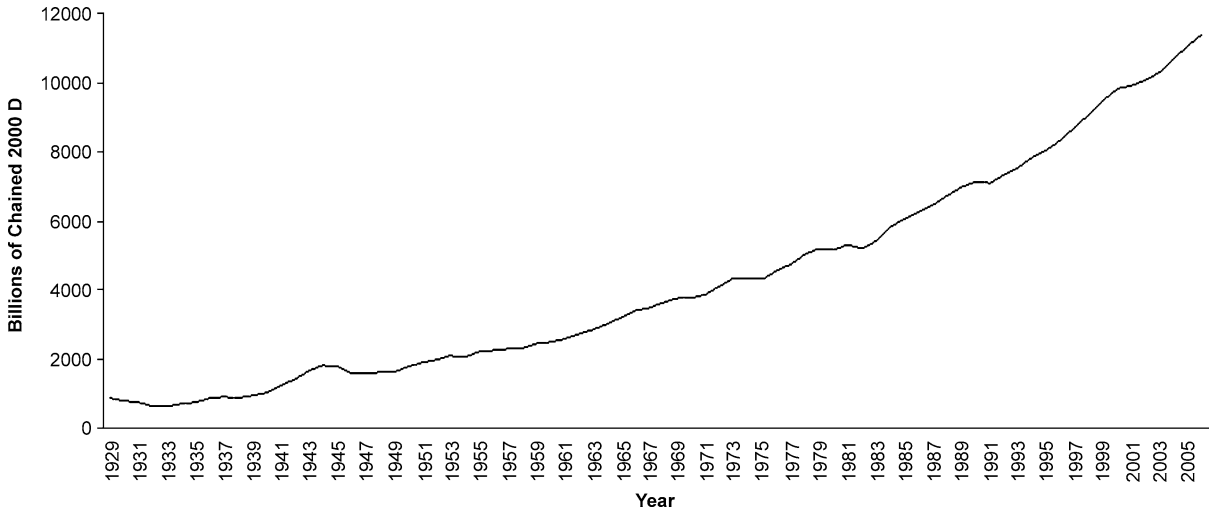

Macroeconomics, Non-linear Time Series in, Figure 1
US real GDP 1929–2006 (Source: St. Louis Fed website)

the random variable y_{t-1} that generates the observation in period $t-1$. The process $\{y_t\}_{-\infty}^{\infty}$ is stochastic because it is driven by random "shocks", such as ε_t in period t. These shocks have the same distribution in every period, meaning that, unlike with y_t and y_{t-1}, the distribution of ε_t does not depend on the value of ε_{t-1} or, for that matter, any other shock in any other period (hence the "i.i.d." tag, which stands for "independently and identically distributed"). It is straightforward to show that the correlation between y_t and y_{t-1} is equal to ϕ and this correlation describes the entire dependence between the two random variables. Indeed, for the basic AR(1) model, the dependence and correlation between any two random variables y_t and y_{t-j}, for all t and j, depends only on the fixed parameter ϕ according to the simple function ϕ^j and, given $|\phi| < 1$, the process has finite memory in terms of past shocks. For other time series models, the functions relating parameters to correlations (i. e., "autocorrelation generating functions") are generally more complicated, as are the restrictions on the parameters to ensure finite memory of shocks. However, the models are still linear, as long as the parameters and correlations are fixed.

In contrast to the linear AR(1) model in (1) and other models with fixed correlations, any model that allows for a more general functional form and/or time variation in the dependence between random variables can be said to be *nonlinear*. This nomenclature is obviously extremely open-ended and examples are more revealing than general definitions. Fortunately, macroeconomics provides many examples, with "nonlinear" typically used to describe models that are closely related to linear models, such as the AR(1) model, but which relax one or two key assumptions in order to capture some aspect of the data that cannot be captured by a linear model. The focus of this survey is on these types of nonlinear models.

It should be mentioned at the outset that, in addition to nonlinear models, "nonlinear time series" evokes nonparametric and semiparametric methods (e. g., neural networks). These methods tend to be data intensive and so find more use in finance and other fields where sample sizes are larger than in macroeconomics. "Nonlinear time series" also evokes the development and application of tests for nonlinearity. However, these are the purview of econometrics, not macroeconomics. Thus, tests for nonlinearity will only be discussed in the context of applications that are particularly relevant to the choice of appropriate models for macroeconomic data.

Types of Nonlinear Models

Starting with the linear AR(1) model in (1), there are many ways to introduce nonlinearities. An obvious way is to consider a nonlinear specification for the relationship between the random variables in the model. For example, consider the simple bilinear model:

$$y_t = c + \phi y_{t-1} + \varepsilon_t + \theta(\varepsilon_{t-1} \cdot y_{t-1}) \,, \qquad \varepsilon_t \sim \text{ i.i.d. } (0, \sigma^2) \,. \quad (2)$$

See Granger and Andersen [57] and Rao and Gabr [139] on bilinear models. In macroeconomics at least, there are relatively few applications of bilinear models, although

see Peel and Davidson [119], Rothman [128], and Hristova [71].

A more typical approach to introducing nonlinearities in macroeconomics is to allow one (or more) of the parameters in a linear model to be driven by its own process. For example, in a macroeconomics paper that was motivated in part by bilinear models, Engle [46] assumed the squares of shocks (i. e., ε_t^2) follow an AR process, with the implication that the conditional variance of y_t is no longer a constant parameter. Given an AR(1) assumption for ε_t^2, the conditional variance is

$$E_{t-1}\left[\sigma_t^2\right] = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 \,, \tag{3}$$

where $E_{t-1}[\,]$ is the conditional expectations operator, with expectations formed using information available in period $t-1$. Engle [46] applied this "autoregressive conditional heteroskedasticity" (ARCH) model to U.K. inflation, although in subsequent research, it has mostly been applied to financial time series. In particular, asset returns tend to display little dependence in the mean, but high positive dependence in terms of the variance (a.k.a. "volatility clustering"), which is exactly what the ARCH model was designed to capture. Beyond Engle's original paper, ARCH models have found little use in macroeconomics, although Bansal and Yaron [4] have recently attempted to resolve the so-called "equity premium puzzle" in part by assuming that US aggregate consumption growth follows a GARCH(1,1) process that generalizes Engle's original ARCH process. However, Ma [104] shows that estimates supporting a GARCH(1,1) model for aggregate consumption growth are due to weak identification, with an appropriate confidence interval suggesting little or no conditional heteroskedasticity. Weak identification is also likely a problem for the earlier application of GARCH models to macroeconomic variables by French and Sichel [49]. In general, because most macroeconomic data series are highly aggregated, the central limit theorem is relevant, at least in terms of eliminating "fat tails" due to volatility clustering that may or may not be present at the microeconomic level or at higher frequencies than macroeconomic data are typically measured.

The ARCH model begs the question of why not consider a stochastic process directly for the variance, rather than for the squares of the shocks. The short answer is a practical one. A model with "stochastic volatility" is more difficult to estimate than an ARCH model. In particular, it can be classified as a state-space model with an unobserved non-Gaussian volatility process that has a nonlinear relationship to the observable time series being modeled. In the simple case of no serial correlation in the underlying series (e. g., no AR dynamics), a stochastic volatility model can be transformed into a linear state-space model for the squares of the series, although the model still has non-Gaussian errors. However, the lack of serial correlation means that this simple version of the model would be more appropriate for applications in finance than macroeconomics. In any event, while the Kalman filter can be employed to help estimate linear Gaussian state-space models, it is less suitable for non-Gaussian state-space models and not at all suitable for nonlinear state-space models. Recent advances in computing power have made simulation-based techniques (the Gibbs sampler and the so-called "particle filter") available to estimate such models, but these techniques are far from straightforward and are highly computationally intensive. See Kim, Shephard, and Chib [88] and Chib, Nardari, and Shephard [21] on estimation of stochastic volatility models via the Gibbs sampler and particle filtering. Meanwhile, such models have rarely been applied to macroeconomic data due to the lack of interesting volatility dynamics discussed above.

To the extent that stochastic volatility models have been applied in macroeconomics, the focus has been on capturing structural change (i. e., permanent variation) in volatility rather than volatility clustering. For example, Stock and Watson [138] investigate the so-called "Great Moderation" using a stochastic volatility model and confirm the findings reported in Kim and Nelson [77] and McConnell and Perez-Quiros [107] that there was a permanent reduction in the volatility of US real GDP growth in the mid-1980s (see also [82,116,132]). This change in volatility is fairly evident in Fig. 2, which presents the time series for US real GDP growth for each quarter from 1947:Q2 to 2006:Q4.

Yet, while it is sometimes merely a matter of semantics, it should be noted that "structural change" is a distinct concept from "nonlinearity". In particular, *structural change* can be thought of as a change in the model describing a time series, where the change is permanent in the sense that it is not expected to be reversed. Then, if the underlying structure of each model is linear, such as for the AR(1) model in (1), there is nothing particularly "nonlinear" about structural change. On the other hand, Bayesian analysis of structural change blurs the distinction between structural change and nonlinearity. In particular, it treats parameters as random variables for the purposes of making inferences about them. Thus, the distinction between a model that allows "parameters" to change according to a stochastic process and a collection of models with the same structure, but different parameters, is essentially a matter of taste, even if the former setup is clearly nonlinear, while the latter is not. For example, consider the classic time-varying parameter model (see, for example [29]).

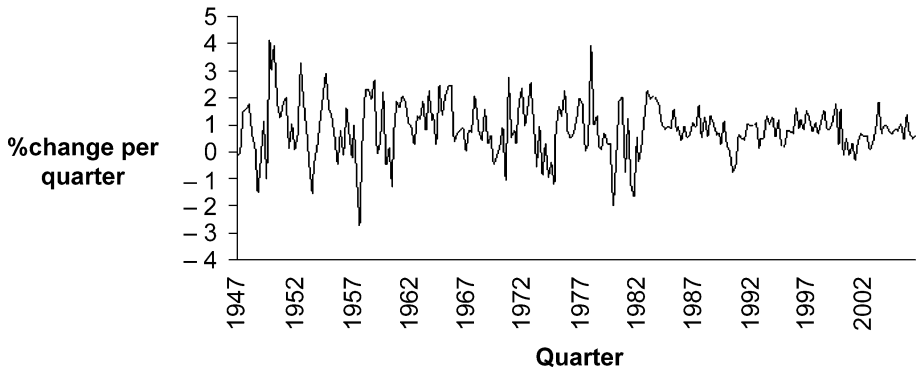

Macroeconomics, Non-linear Time Series in, Figure 2
US real GDP growth 1947–2006 (Source: St. Louis Fed website)

Macroeconomics, Non-linear Time Series in, Figure 3
US inflation 1960–2006 (Source: St. Louis Fed website)

Like the stochastic volatility model, it assumes a stochastic process for the parameters in what would, otherwise, be a linear process. Again, starting with the AR(1) model in (1) and letting $\beta = (c, \phi)'$, a time-varying parameter model typically assumes that the parameter vector evolves according to a multivariate random walk process:

$$\beta_t = \beta_{t-1} + v_t, \quad v_t \sim \text{ i.i.d. } (0, \Sigma) . \tag{4}$$

Because the time-varying parameter model treats the evolution of parameters as a stochastic process, it is clearly a nonlinear model. At the same time, its application to data provides an inherently Bayesian investigation of structural change in the relationships between dependent and independent variables, where those relationships may, in fact, be linear. In general, then, analysis of structural change in linear relationships should be considered an example of nonlinear time series analysis when nonlinear models, such as stochastic volatility models or time-varying parameter models, are used in the analysis, but structural change should not be thought of as nonlinear in itself.

In terms of macroeconomics, time-varying parameter models have recently been used to consider structural change in vector autoregressive (VAR) models of the US economy. Cogley and Sargent [26] employ such a model to argue that US inflation dynamics have changed considerably in the postwar period. Based on Sims' [135] critique that evidence for structural change in time-varying parameters may be the spurious consequence of ignoring heteroskedasticity in the error processes for a VAR model, Cogley and Sargent [27] augment their time-varying parameter model with stochastic volatility and find that their results are robust. Primiceri [123] employs a structural VAR with time-varying parameters and stochastic volatility and also finds evidence of structural changes in inflation dynamics, although he questions the role of monetary policy in driving these changes. Whether these structural changes are evident in Fig. 3, which displays US consumer price inflation for each month from 1960:M1 to 2006:M12, is debatable. However, it is fairly clear that a basic AR process with constant parameters would be an inadequate model for inflation.

It is worth mentioning that there is a simpler time-varying parameter model that has seen considerable use in macroeconomics. It is the unobserved components (UC) model used for trend/cycle decomposition. A standard version of the model has the following form:

$$y_t = \tau_t + c_t \,, \tag{5}$$

$$\tau_t = \mu + \tau_{t-1} + \eta_t, \quad \eta_t \sim \text{ i.i.d.N } \left(0, \sigma_\eta^2\right) \,, \tag{6}$$

$$\phi(L)c_t = \varepsilon_t, \quad \varepsilon_t \sim \text{ i.i.d.N } \left(0, \sigma_\varepsilon^2\right) \,, \tag{7}$$

where $\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$, the roots of $\phi(z) = 0$ lie outside the unit circle, and $\text{corr}\,(\eta_t, \varepsilon_t) = \rho_{\eta\varepsilon}$. It is possible to think of the UC model as a time-varying parameter model in which the unconditional mean of the process is equal to the trend τ_t, meaning that it undergoes structural change, rather than remaining constant, as it does for the AR(1) process described by (1). A glance at the upward trajectory of real GDP in Fig. 1 makes it clear that a basic AR process would be an extremely bad model for the time series. Indeed, Morley, Nelson, and Zivot [114] applied the model in (5)–(7) to 100 times the natural logarithms of US real GDP under the assumption that the lag order $p = 2$ and with no restrictions on the correlation between η_t and ε_t and found that most of the variation in log real GDP was due to the trend rather than the AR cycle c_t (note that natural logarithms are more appropriate for time series modeling than the raw data in Fig. 1 because the "typical" scale of variation for real GDP is more closely linked to percentage changes than to absolute changes). Yet, while the UC model can be thought of as a time-varying parameter model, it is not, in fact, nonlinear. In particular, the UC model for log real GDP is equivalent to an autoregressive moving-average (ARMA) model for the first differences of log real GDP. Likewise, the AR(1) model in (1) may be very sensible for real GDP growth in Fig. 2, even though it would be a bad model for real GDP in Fig. 1. In general, if it is possible to transform a time series, such as going from Fig. 1 to Fig. 2, and employ a linear model for the transformed series, then the time series analysis involved is linear. Likewise, under this formulation, the simple version of the stochastic volatility model for a series with no serial correlation also falls under the purview of linear time series analysis. Only time-varying parameter and stochastic volatility models that cannot be transformed into linear representations are nonlinear.

Of course, the semantics over "linear" and "nonlinear" are hardly important on their own. What is important is whether structural change is mistaken for recurring changes in parameters or vice versa. In terms of structural VAR models for the US economy, Sims and Zha [136] argue that when parameters are allowed to undergo large, infrequent changes, rather than the smaller, more continuous changes implied by a time-varying parameter model, there is no evidence for changes in dynamic structure of postwar macroeconomic data. Instead, there are only a few large, infrequent changes in the variance of shocks. Furthermore, among the models that assume some change in dynamics, their Bayesian model comparison favors a model in which only the monetary policy rule changes. Among other things, these findings have dramatic implications for the Lucas [100,101] critique, which suggests that correlations between macroeconomic variables should be highly sensitive to changes in policy, thus leaving successful forecasting to "structural" models that capture optimizing behavior of economic agents, rather than "reduced-form" models that rely on correlations between macroeconomic structures. The results in Sims and Zha [136] suggest that the Lucas critique, while an interesting theoretical proposition with the virtue of being empirically testable, is not, in fact, supported by the data.

From the point of view of time series analysis, an interesting aspect of the Sims and Zha [136] paper and earlier papers on structural change in the US economy by Kim and Nelson [77] and McConnell and Perez-Quiros [107] is that they consider nonlinear regime-switching models that allow for changes in parameters to be recurring. That is, while the models can capture structural change, they do not impose it. Using univariate regime-switching models of US real GDP growth, Kim and Nelson [77] and McConnell and Perez-Quiros [107] find a one-time permanent reduction in output growth volatility in 1984. However, using their regime-switching VAR model, Sims and Zha [136] find that a small number of volatility regimes recur multiple times in the postwar period. In terms of the earlier discussion about the lack of volatility dynamics in macroeconomic data, this finding suggests that there are some volatility dynamics after all, but these dynamics correspond to less frequent changes than would be implied by ARCH or a continuous stochastic volatility process. More generally, the allowance for recurring regime switches is relevant because time series models with regime switches have been the most successful form of nonlinear models in macroeconomics. However, for reasons discussed in the next section, regime-switching models are typically employed to capture changing dynamics in measures of economic activity over different phases of the business cycle, rather than structural change in inflation or recurring changes in shock variances.

To summarize this section, there are different types of nonlinear time series models employed in macroeconom-

ics. While models that assume a nonlinear specification for the relationship between observable variables exist (e. g., the bilinear model), they are rarely used in practice. By contrast, models that allow some parameters to undergo changes over time are much more common in macroeconomics. The examples discussed here are ARCH models, stochastic volatility models, time-varying parameter models, and regime-switching models. When examining structural change, there is a conceptual question of whether the analysis is "linear" or "nonlinear". However, as long as the process of structural change is an explicit part of the model (e. g., the time-varying parameter model), and excluding cases where it is possible to transform the model to have a linear representation (e. g., the UC model to an ARMA model), the analysis can be thought of as nonlinear. Meanwhile, time series analysis of recurring regime switches is unambiguously nonlinear. As discussed in the next section, nonlinear regime-switching models come in many versions and have found wide use in macroeconomics modeling business cycle asymmetry.

Business Cycle Asymmetry

The topic of business cycle asymmetry is broad and the literature on it extensive. As a result, it is useful to divide the discussion in this section into four areas: i) concepts of business cycle asymmetry and their relationships to nonlinearity; ii) nonlinear models of business cycle asymmetry; iii) evidence for nonlinear forms of business cycle asymmetry; and iv) the relevance of nonlinear forms of business cycle asymmetry for macroeconomics.

Concepts

Notions of business cycle asymmetry have a long tradition in macroeconomics. Classic references to the idea that recessions are shorter, sharper, and generally more volatile than expansions are Mitchell [109], Keynes [72], and Burns and Mitchell [13]. For example, in his characteristic style, John Maynard Keynes writes, "…the substitution of a downward for an upward tendency often takes place suddenly and violently, whereas there is, as a rule, no such sharp turning point when an upward is substituted for a downward tendency." (see p. 314 in [72]). Similarly, albeit more tersely, Wesley Mitchell writes, "…the most violent declines exceed the most considerable advances. The abrupt declines usually occur in crises; the greatest gains occur in periods of revival… Business contractions appear to be a briefer and more violent process than business expansions." (see p. 290 in [109]). Milton Friedman also saw business cycle asymmetry in the form of a strong relationship between the depth of recession and the strength of a recovery, with no corresponding relationship between the strength of an expansion with the severity of the subsequent recession (see [50,51]).

The link between business cycle asymmetry and nonlinearity depends, in part, on the definition of "business cycle". Harding and Pagan [67] discuss three possible definitions that are presented here using slightly modified terminology. Based on the work of Burns and Mitchell [13], the first definition is that the business cycle is the alternation between phases of expansion and recession in the *level* of economic activity. The second definition, which is often left implicit when considered, is that the business cycle represents transitory *deviations* in economic activity from a permanent or "trend" level. The third definition, which is also often only implicitly considered, is that the business cycle corresponds to any short-run *fluctuations* in economic activity, regardless of whether they are permanent or transitory.

Under the "level" definition of the business cycle, there is nothing inherently nonlinear about asymmetry in terms of the duration of expansions and recessions. Positive drift in the level of economic activity implies longer expansions than recessions, even if the underlying process is linear. Even asymmetry in the form of relative sharpness and steepness of a recession alluded to in the above quote from Keynes does not necessarily indicate nonlinearity. Again, given positive drift, an outright decline in economic activity only occurs when there are large negative shocks to the underlying process, while an expansion occurs for all positive shocks and small negative shocks. Thus, a recession is likely to look like a relatively sharp reversal in the level. Furthermore, with positive serial correlation in growth, such as implied by a linear AR(1) process as in (1) with $\phi > 0$, recessions will appear steeper than expansions due to the dynamic effects of large negative shocks. On the other hand, as discussed in more detail later, nonlinear models are much more successful than linear models at reproducing business cycle asymmetry in the form of a strong link between recessions and their recoveries versus a weak link between expansions and subsequent recessions noted by Friedman [50].

Under the "deviations" definition of the business cycle, asymmetry is closely linked to nonlinearity. While it is possible for asymmetry in the independent and identical distribution of the underlying shocks to generate asymmetry in a linear process, any persistence in the process would severely dampen the asymmetries in the unconditional distribution. Thus, under the assumption that the transitory component of economic activity is at least somewhat persistent, asymmetries such as differences in the durations of positive and negative deviations from trend or rel-

ative sharpness and steepness in negative deviations compared to positive deviations are more suggestive of nonlinear dynamics (i. e., changing correlations) than underlying asymmetric shocks.

Under the "fluctuations" definition of the business cycle, the link between nonlinearity and asymmetry also depends on the relative roles of shocks and dynamics in generating asymmetries. However, because growth rates are less persistent than most measures of the transitory component of economic activity and because they mix together permanent and transitory shocks that may have different means and variances, it is quite plausible that asymmetry in the distribution of shocks is responsible for asymmetry in growth rates. Of course, nonlinear dynamics are also a plausible source of asymmetry for growth rates.

In terms of asymmetries, it is useful to consider the formal classifications developed and discussed in Sichel [133], McQueen and Thorley [108], Ramsey and Rothman [124], Clements and Krolzig [24], and Korenok, Mizrach, and Radchenko [95] of "deepness", "steepness", and "sharpness". Following Sichel [133], a process is said to have *deepness* if its unconditional distribution is skewed and *steepness* if the distribution of its first-differences is skewed. Following McQueen and Thorley [108], a process is said to have *sharpness* if the probability of a peak occurring when it has been increasing is different than the probability of a trough occurring when it has been decreasing. However, despite these definitions, the different types of asymmetries are most easily understood with visual examples.

Figure 4 presents an example of a simulated time series with deepness, with the distance from peak of the cycle to the mean less than the distance from the mean to trough of the cycle (see [124], for the details of the process generating this time series). In addition to deepness, the series appears to display sharpness in recessions, with the peak of the cycle more rounded than the trough, although the fact that the simulated series is deterministic means it cannot be directly related to the definition of sharpness in McQueen and Thorley [108] mentioned above. Meanwhile, there is no steepness because the slope from peak to trough is the same magnitude as the slope from trough to peak.

As discussed in Ramsey and Rothman [124], these different types of asymmetry can be classified in two broader categories of "time reversible" and "time irreversible". *Time reversibility* means that the substitution of $-t$ for t in the equations of motion for a process leaves the process unchanged. The upward drift that is present in many macroeconomic time series (such as real GDP) is clearly time irreversible. More generally, the issue of time reversibility is relevant for determining whether business cycle asymmetry corresponds to deepness and sharpness, which are time reversible, or steepness, which is time irreversible. For example, the time series in Fig. 4 can be flipped on the vertical axis without any resulting change. Thus, it is time reversible. By contrast, consider the simulated time series with "steepness" in Fig. 5. The series is generated from a regime-switching process with asymmetric shocks across two regimes and different persistence for shocks in each regime. In this case, flipping the series on the vertical axis would produce flat inclines and steep declines. Thus, it is time irreversible.

The relevance of the distinction between time reversible and time irreversible processes is obvious from Fig. 6, which presents the time series for the US civilian unemployment rate for each month from 1960:M1 to 2006:M12. The inclines are steep relative to the declines. Thus, there is a clear visual suggestion of the steepness form of asymmetry. Indeed, the modern literature on business cycle asymmetry begins with Neftçi's [115] investi-

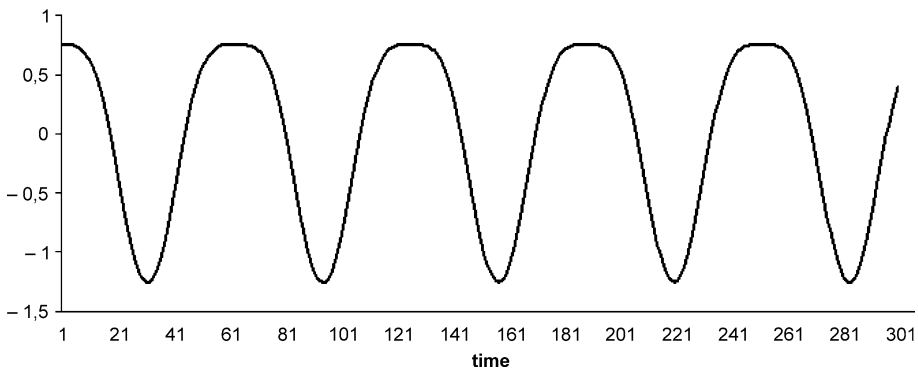

Macroeconomics, Non-linear Time Series in, Figure 4
A "deep" cycle (Source: Author's calculations based on Ramsey and Rothman [124])

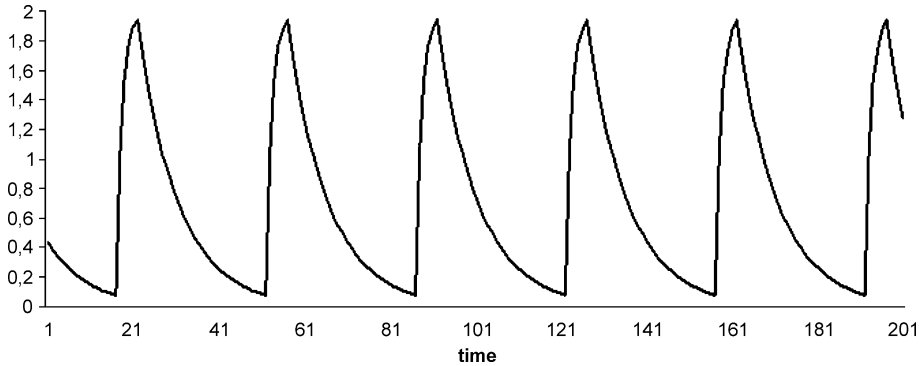

Macroeconomics, Non-linear Time Series in, Figure 5
A "steep" cycle (Source: Author's calculations)

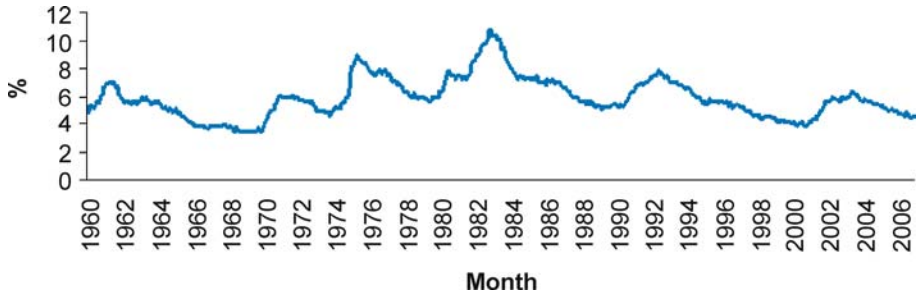

Macroeconomics, Non-linear Time Series in, Figure 6
US civilian unemployment rate 1960–2006 (Source: St. Louis Fed website)

gation of this issue using a nonlinear regime-switching model in which the prevailing "business cycle" regime in a given period is assumed to depend on a discrete Markov process driven by whether the US unemployment rate is rising or falling in that period. Given the link to the first differences of the unemployment rate, his finding that the continuation probabilities for the two regimes are different, with declines more likely to persist than increases, provides formal support for the presence of the steepness forms of asymmetry in the unemployment rate (also, see [127]). It should also be noted that, while not related to time irreversibility, the different continuation probabilities also directly imply sharpness.

Models

The subsequent literature on regime-switching models in macroeconomics can be usefully divided into two categories that are both related to Neftçi's [115] model. First, *Markov-switching models* assume that the prevailing regime depends on an unobserved discrete Markov process. The main distinction from Neftçi [115] is that the Markov process is unobserved (hence, these models are sometimes referred to as a "hidden Markov models"). Second, *self-exciting threshold models* assume that the prevailing regime is observable and depends on whether realized values of the time series being modeled exceed or fall below certain "threshold" values, much like the regime in Neftçi's [115] model depends on whether the change in the unemployment rate was positive or negative.

Hamilton [59] is the seminal paper in terms of Markov-switching models. His model has a basic AR structure, like in (1), but for the first-differences of the time series of interest:

$$\phi(L)\left(\Delta y_t - \mu_t\right) = \varepsilon_t, \quad \varepsilon_t \sim \text{ i.i.d. } (0, \sigma^2)\,, \tag{8}$$

where Δy_t is 100 times the change in the natural logarithm of real Gross National Product (GNP). The only difference from a linear AR model is that the mean follows a stochastic process:

$$\mu_t = \mu_1 \cdot I\left(S_t = 1\right) + \mu_2 \cdot I\left(S_t = 2\right)\,, \tag{9}$$

with the indicator function $I(S_t = j)$ equal to 1 if $S_t = j$ and 0 otherwise and $S_t = \{1, 2\}$ following an unobserved discrete Markov state variable that evolves according to

the following fixed transition matrix:

$$\begin{bmatrix} p_{11} & p_{21} \\ p_{12} & p_{22} \end{bmatrix},$$

where $p_{ij} \equiv \Pr[S_t = j | S_{t-1} = i]$ and the columns sum to one.

There are two aspects of Hamilton's [59] model that should be mentioned. First, while the demeaned specification is equivalent to a regression model specification (e. g., (1)) in the linear setting, with $\mu = c/(1-\phi)$, the two specifications are no longer equivalent in the nonlinear setting. In particular, if the intercept c were switching instead of the mean μ, then past regime switches would be propagated by the AR dynamics (see [61], for an example of such a model). By contrast, with μ switching, there is a clear separation between the "nonlinear" dynamics due to the evolution of the state variable (which does alter the correlations between Δy_t and its lags) and the "linear" dynamics due to the ε_t shocks and the AR parameters. Second, in order to eliminate arbitrariness in the labeling of states, it is necessary to impose a restriction such as $\mu_1 > \mu_2$, which corresponds to higher mean growth in state 1 than in state 2. Furthermore, given the application to output growth, if $\mu_1 > 0$ and $\mu_2 < 0$, the states 1 and 2 can be labeled "expansion" and "recession", respectively.

Hamilton's [59] paper had a big impact on econometrics and macroeconomics for two reasons. First, it included an elegant filter that could be used to help estimate Markov-switching models via maximum likelihood and, along with a smoother, calculate the posterior distribution of the unobserved state variable (filters and smoothers are recursive algorithms that make inferences about unobserved state variables, with filters considering only information available at the time the state variable is realized and smoothers incorporating any subsequent available information). Second, the resulting posterior probability of the "recession" regime corresponded closely to the National Bureau of Economic Research (NBER) dating of recessions. The NBER dating is based on non-structural and subjective analysis of a wide variety of indicators. The official line from its website is "The NBER does not define a recession in terms of two consecutive quarters of decline in real GDP. Rather, a recession is a significant decline in economic activity spread across the economy, lasting more than a few months, normally visible in real GDP, real income, employment, industrial production, and wholesale-retail sales." (www.nber.org/cycles/cyclesmain.html). Thus, it is, perhaps, remarkable that a simple time series model using only information in real GNP could find such similar dates for recessions. Of course, as emphasized by Harding and Pagan [66], a simple rule like "two consecutive quarters of decline in real GDP" also does extremely well in matching the NBER recessions, regardless of NBER claims that it is not following such a rule. Yet, more important is the notion implied by Hamilton's [59] results that the NBER is identifying a meaningful structure in the economy, rather than simply reporting (sometimes with considerable lag) that the economy had an episode of prolonged negative growth. Specifically, "recession" appears to be an indicator of a different state for the dynamics of the economy, rather than a label for particular realizations of linear process. (As an aside, the fact that the popular press pays so much attention to NBER pronouncements on recessions also supports the idea that it is identifying a meaningful macroeconomic structure).

Numerous modifications and extensions of Hamilton's [59] model have been applied to macroeconomic data. For example, while estimates for Hamilton's [59] model imply that the linear ε_t shocks have large permanent effects on the level of real GDP, Lam [96] considers a model in which the only permanent shocks to real GNP are due to regime switches. Despite this very different assumption, he also finds that the regime probabilities implied by his model correspond closely to NBER dating of expansions and recessions. Kim [74] develops a filter that can be used for maximum likelihood estimation of state-space models with Markov-switching parameters and confirms the results for Lam's [96] model. Motivated by Diebold and Rudebusch's [38] application of Hamilton's [59] model to the Commerce Department's coincident index of economic activity instead of measures of aggregate output such as real GNP or real GDP, Chauvet [19] employs Kim's [74] filter to estimate an unobserved components model of a coincident index using Hamilton's [59] model as the specification for its first differences. Other multivariate extensions include Kim and Yoo [87], Ravn and Sola [125], Kim and Nelson [76], Kim and Murray [75], Kim and Piger [81], Leamer and Potter [97], Camacho [14], and Kim, Piger, and Startz [84]. The general theme of these studies is that the multivariate information, such as coincident indicators or aggregate consumption and investment, helps to strongly identify the nonlinearity in economic activity, with regimes corresponding even more closely to NBER dates than for univariate analysis based on real GNP or real GDP.

In terms of business cycle asymmetry, an important extension of Hamilton's [59] model involves allowing for three regimes to capture three phases of the business cycle: "expansion", "recession", and "recovery" (see [134]). Papers with three-regime models include Boldin [8], Clements and Krolzig [23], and Leyton and Smith [98].

The specification in Boldin [8] modifies the time-varying mean in Hamilton's [59] model as follows:

$$\mu_t = \mu_1 \cdot I(S_t = 1) + \mu_2 \cdot I(S_t = 2) + \mu_3 \cdot I(S_t = 3)\,, \quad (10)$$

where $S_t = \{1, 2, 3\}$ has fixed transition matrix:

$$\begin{bmatrix} p_{11} & 0 & p_{31} \\ p_{12} & p_{22} & 0 \\ 0 & p_{23} & p_{33} \end{bmatrix}.$$

The zeros in the transition matrix restrict the state sequence to follow the pattern of $\{S_t\} = \cdots 1 \to 2 \to 3 \to 1 \cdots$. Given the normalization $\mu_1 > \mu_2$, the restriction on the transitional matrix implies that the economy goes from expansion to recession to recovery and back to expansion. While there is no restriction on μ_3, Boldin [8] finds it is greater than μ_1, which means that the third regime corresponds to a high-growth recovery. As discussed in Clements and Krolzig [24], this third regime allows for steepness in output growth, while the basic two-regime Hamilton [59] model can only capture deepness and sharpness (the two are inextricably linked for a two-regime model) in growth. Note, however, from the definitions presented earlier, deepness in growth implies steepness the level of output.

It is possible to capture high-growth recoveries without resorting to three regimes. For example, Kim and Nelson [79] develop an unobserved components model that assumes two regimes in the transitory component of US real GDP. A slightly simplified version of their model is given as follows:

$$y_t = \tau_t + c_t\,, \quad (11)$$

$$\tau_t = \mu + \tau_{t-1} + \eta_t, \quad \eta_t \sim \text{i.i.d.N}\left(0, \sigma_\eta^2\right)\,, \quad (12)$$

$$\phi(L)c_t = \lambda \cdot I(S_t = 2) + \varepsilon_t, \quad \varepsilon_t \sim \text{i.i.d.N}\left(0, \sigma_\varepsilon^2\right)\,, \quad (13)$$

where y_t is 100 times log real GDP, $S_t = \{1, 2\}$ is specified as in Hamilton's [59] model, and state 2 is identified as the recession regime by the restriction $\lambda < 0$ (see [112,113], on the need for and implications of this restriction). Unlike Morley, Nelson, and Zivot [114], Kim and Nelson [79] impose the restriction that $\rho_{\eta\varepsilon} = 0$ in estimation, which they conduct via approximate maximum likelihood using the Kim [74] filter. As with Hamilton [59] and Lam [96], the regimes correspond closely to NBER-dated expansions and recessions. However, because the regime switching is in the transitory component only, the transition from state 1 to state 2 corresponds to a downward "pluck" in economic activity that is followed by a full recovery to trend after the transition from state 2 to state 1. Kim and Nelson [79] motivate their model as nesting Friedman's [50,51] plucking model, which assumes output cannot exceed a ceiling level, but is occasionally plucked below full capacity by recessionary shocks resulting from activist monetary policy. In line with Friedman's observations, Kim and Nelson's [79] model relates the strength of a recovery to the severity of the preceding recession, with no corresponding link between the strength of an expansion and the severity of a recession (see also [2,134,150]). Notably, the transitory component for their estimated model achieves the trifecta of business cycle asymmetries in the form of deepness, steepness, and sharpness.

Another model that captures three phases of the business cycle with only two underlying regimes is the "bounceback" model of Kim, Morley, and Piger [83]. The model modifies the time-varying mean in Hamilton's [59] model as follows:

$$\mu_t = \mu_1 \cdot I(S_t = 1) + \mu_2 \cdot I(S_t = 2) + \lambda \cdot \sum_{j=1}^{m} I\left(S_{t-j} = 2\right), \quad (14)$$

where the number of lagged regimes to consider in the third term on the right hand side of (14) is determined by the discrete "memory" parameter m, which is estimated to be six quarters for US postwar quarterly real GDP. Given the restriction $\mu_1 > \mu_2$, the third term can be interpreted as a pressure variable that builds up the longer a recession persists (up to m periods, where $m = 6$ quarters is long enough to capture all postwar recessions) and is motivated by the "current depth of recession" variable of Beaudry and Koop [6] discussed later. Then, if $\lambda > 0$, growth will be above μ_1 for up to the first six quarters of an expansion. That is, there is a post-recession "bounceback" effect, as in Kim and Nelson's [79] plucking model. Meanwhile, the specification in (14) can be thought of as a "u-shaped recession" version of the model because the pressure variable starts mitigating the effects of a recession the longer the regime persists. Morley and Piger [111] consider a slightly modified "v-shaped recession" version of the model that assumes the pressure variable only affects growth after the recession ends, thus producing a sharper trough:

$$\mu_t = \mu_1 \cdot I(S_t = 1) + \mu_2 \cdot I(S_t = 2) + \lambda \cdot \sum_{j=1}^{m} I(S_t = 1) \cdot I\left(S_{t-j} = 2\right). \quad (15)$$

This version of the model is identical to Hamilton's [59] model in all but the first m periods of an expansion. Finally, Morley and Piger [113] consider a "depth" version

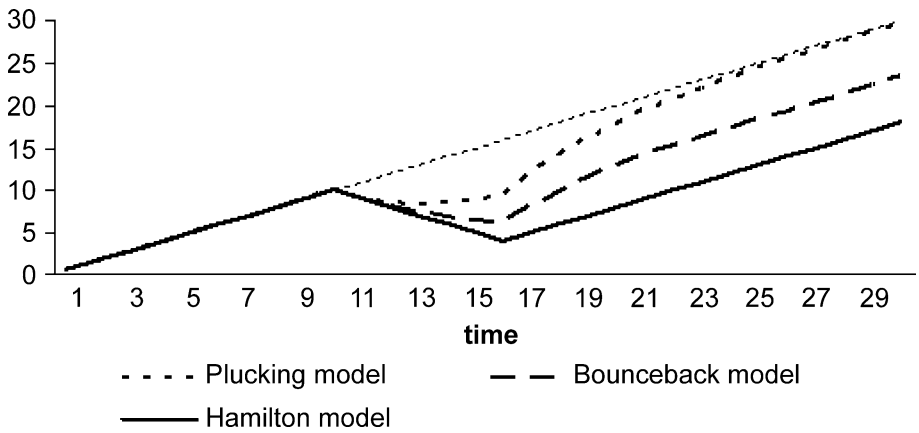

Macroeconomics, Non-linear Time Series in, Figure 7
Simulated paths for "Output" (Source: Author's calculations)

of the model that relates the pressure variable to both the length and severity of a recession:

$$\mu_t = \mu_1 \cdot I(S_t = 1) + \mu_2 \cdot I(S_t = 2) + \lambda \cdot \sum_{j=1}^{m} \left(\mu_1 - \mu_2 - \Delta y_{t-j}\right) \cdot I\left(S_{t-j} = 2\right). \quad (16)$$

In this case, the post-recession bounceback effect depends on the relative severity of a recession. Regardless of the specification, the estimated bounceback effect for US real GDP based on maximum likelihood estimation via the Hamilton [59] filter is large (see [83,111,113]).

While Kim, Morley, and Piger's [83] bounceback model can capture "plucking" dynamics, there is no restriction that regime switches have only transitory effects. Instead, the model nests both the Hamilton [59] model assumption that recessions have large permanent effects in the case that $\lambda = 0$ and Kim and Nelson's [79] "plucking" model assumption that recessions have no permanent effects in the case that $\lambda = (\mu_1 - \mu_2)/m$ (for the specification in (14)). Figure 7 presents examples of simulated time series for the plucking model, the bounceback model, and the Hamilton model. In each case, "output" is subject to a recession regime that lasts for six periods. For the plucking model, output returns to the level it would have been in the absence of the recession. For the Hamilton model, output is permanently lower as a result of the recession. For the bounceback model, recessions can have permanent effects, but they will be less than for the Hamilton model if $\lambda > 0$ (indeed, if $\lambda > (\mu_1 - \mu_2)/m$, the long-run path of the economy can be increased by recessions, a notion related to the "creative destruction" hypothesis of Schumpeter [131]). In practice, Kim, Morley, and Piger [83] find a very small negative long-run impact of US recessions, providing support for the plucking model dynamics and implying considerably lower economic costs of recessions than the Hamilton model.

Another extension of Hamilton's [59] model involves relaxing the assumption that the transition probabilities for the unobserved state variable are fixed over time (see [39]). Filardo [48] considers time-varying transition probabilities for a regime-switching model of industrial production growth where the transition probabilities depend on leading indicators of economic activity. Durland and McCurdy [40] allow the transition probabilities for real GNP growth to depend on the duration of the prevailing regime. DeJong, Liesenfeld, and Richard [34] allow the transition probabilities for real GDP growth depend on an observed "tension index" that is determined by the difference between recent growth and a "sustainable" rate that corresponds to growth in potential output. Kim, Piger, and Startz [84] allow for a dynamic relationship between multiple unobserved discrete state variables in a multivariate setting and find that regime-switches in the permanent component of economic activity tend to lead regime-switches in the transitory component when the economy heads into recessions.

The distinction between Markov-switching models and threshold models is blurred somewhat by time-varying transition probabilities. A standard demarcation is that Markov-switching models typically assume the discrete state variables driving changes in regimes are exogenous, while threshold models allow for endogenous switching. However, this exogenous/endogenous demarcation is less useful than it may at first appear. First, as is always the problem in macroeconomics, it is unlikely that the variables affecting time-varying transition prob-

abilities are actually strictly exogenous, even if they are predetermined. Second, Kim, Piger and Startz [85] have developed an approach for maximum likelihood estimation of Markov-switching models that explicitly allow for endogenous switching. In terms of macroeconomics, Sinclair [137] applies their approach to estimate a version of the regime-switching UC model in (11)–(13) for US real GDP that allows for a non-zero correlation between the regular shocks η_t and ε_t, as in Morley, Nelson, and Zivot [114], as well as dependence between these shocks and the unobserved state variable S_t that generates downward plucks in output. She finds that permanent shocks are more important than suggested by Kim and Nelson's [79] estimates. However, she confirms the importance of the plucking dynamic, with a test supporting the standard exogeneity assumption for the discrete Markov-switching state variable.

Another demarcation that would seem to provide a possible means of distinguishing between Markov-switching models and threshold models arises from the fact that, starting from an AR specification, threshold models typically extend the basic model by allowing for changes in AR parameters, while, as discussed earlier, Markov-switching models typically extend the model by allowing for changes in the mean. However, this demarcation is also less useful than it may at first appear since Markov-switching models have alternative representations as autoregressive processes (see [59]). Furthermore, some threshold models assume constant AR parameters (e. g., [120]). In particular, regardless of presentation, both types of models capture nonlinear dynamics in the conditional mean.

The more general and useful demarcation between Markov-switching models and threshold models is that the prevailing regime is unobservable in the former, while it is observed in the latter. Meanwhile, the observable regimes in threshold models make it feasible to consider more complicated transitions between regimes than Markov switching models. In particular, it is possible with a threshold model to allow a mixture of regimes to prevail in a given time period.

Tong [145] introduced the basic threshold autoregressive (TAR) model. In a "self-exciting" TAR model, the autoregressive coefficient depends on lagged values of the time series. For example, a simple two-regime AR(1) TAR model is given as follows:

$$y_t = c + \phi^{(1)} \cdot I\left(y_{t-m} < \tau\right) \cdot y_{t-1} + \phi^{(2)} \cdot I\left(y_{t-m} \geqslant \tau\right) \cdot y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{i.i.d.}(0, \sigma^2), \tag{17}$$

where $\phi^{(1)}$ and $\phi^{(2)}$ are the AR(1) parameters associated with the two regimes, τ is the threshold, and m is the discrete delay parameter. A variant of the basic TAR model that allows multiple regimes to prevail to different degrees is the smooth transition autoregressive (STAR) model (see [18,58,140,142]). For STAR models, the indicator function is replaced by transition functions bounded between zero and one. The STAR model corresponding to (17) is

$$y_t = c + \phi^{(1)} \cdot F_1\left(y_{t-m}|\tau, \gamma\right) \cdot y_{t-1} + \phi^{(2)} \cdot F_2\left(y_{t-m}|\tau, \gamma\right) \cdot y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{i.i.d.}\left(0, \sigma^2\right), \tag{18}$$

where $F_2(y_{t-m}\,|\tau, \gamma) = 1 - F_1(y_{t-m}\,|\tau, \gamma)$ and γ is a parameter that determines the shape of the transition function (in general, the larger γ, the closer the STAR model is to the TAR model). The two most popular transition functions are exponential (ESTAR) and logistic (LSTAR). The exponential transition function is

$$F_1^e = 1 - \exp\left(-\gamma(y_{t-m} - \tau)^2\right), \quad \gamma > 0, \tag{19}$$

while the logistic transition function is

$$F_1^l = \left[1 + \exp\left(-\gamma(y_{t-m} - \tau)\right)\right]^{-1}, \quad \gamma > 0. \tag{20}$$

For STAR models the transition functions are such that the prevailing autoregressive dynamics are based on a weighted average of the autoregressive parameters for each regime, rather than reflecting only one or the other, as in TAR models.

In terms of macroeconomics, both TAR and STAR models have been employed to capture business cycle asymmetry. A key question is what observed threshold might be relevant. On this issue, a highly influential paper is Beaudry and Koop [6]. Related to the notion discussed above that recessions represent a meaningful macroeconomic structure, they consider whether real GDP falls below a threshold defined by its historical maximum. Specifically, they define a "current depth of recession" (CDR) variable as follows:

$$\text{CDR}_t = \max\left\{y_{t-j}\right\}_{j \geqslant 0} - y_t. \tag{21}$$

Figure 8 presents the current depth of recession using US real GDP for each quarter from 1947:Q1 to 2006:Q4.

Beaudry and Koop [6] augment a basic linear ARMA model of US real GNP growth with lags of the CDR variable. They find that the inclusion of the CDR variable implies much less persistence for large negative shocks than

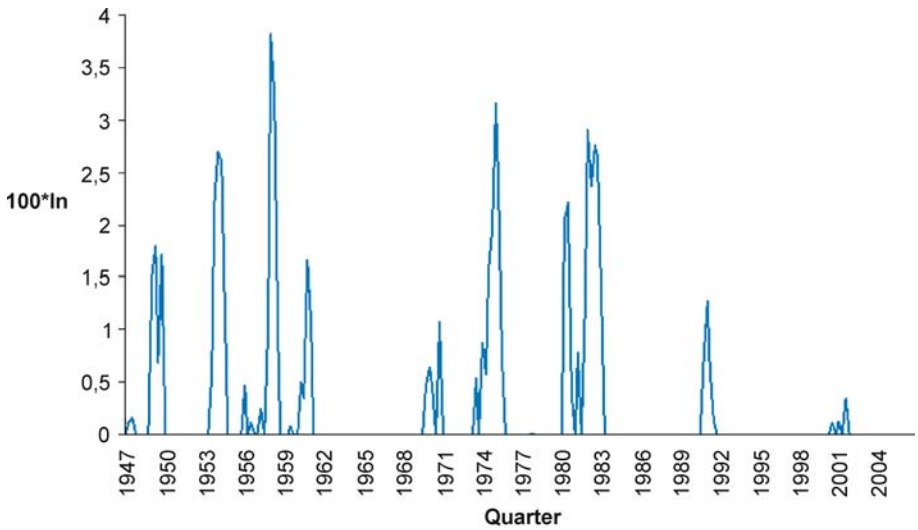

Macroeconomics, Non-linear Time Series in, Figure 8
Current depth of recession 1947–2006 (Source: Author's calculations based on Beaudry and Koop [6])

for small negative shocks or positive shocks. The asymmetry in terms of the response of the economy to shocks corresponds closely to the idea discussed earlier that deep recessions produce strong recoveries. Indeed, the Beaudry and Koop [6] paper provided a major motivation for most of the extensions of Hamilton's [59] model discussed earlier that allow for high-growth recoveries.

In terms of threshold models in macroeconomics, Beaudry and Koop [6] initiated a large literature. Tiao and Tsay [144], Potter [121], and Clements and Krolzig [23] consider two-regime TAR models with the threshold either fixed at zero or estimated to be close to zero. Pesaran and Potter [120] and Koop and Potter [91] consider a three-regime TAR model (with many restrictions for tractability) that incorporates the CDR variable and an "overheating" (OH) variable reflecting cumulated growth following large positive shocks. Specifically, a simple homoskedastic, AR(1) version of Pesaran and Potter's [120] "floor and ceiling" model is given as follows:

$$\Delta y_t = c + \phi \Delta y_{t-1} + \lambda_1 \mathrm{CDR}_{t-1} + \lambda_2 \mathrm{OH}_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2)\,, \tag{22}$$

where

$$\mathrm{CDR}_t = -(\Delta y_t - \tau_F) \cdot F_t \cdot (1 - F_{t-1}) + (\mathrm{CDR}_{t-1} - \Delta y_t) \cdot F_t \cdot F_{t-1}\,, \tag{23}$$

$$F_t = I\left(\Delta y_t < \tau_F\right) \cdot (1 - F_{t-1}) + I\left(\mathrm{CDR}_{t-1} - \Delta y_t > 0\right) \cdot F_{t-1}\,, \tag{24}$$

$$\mathrm{OH}_t = (\mathrm{OH}_{t-1} + \Delta y_t - \tau_C) \cdot C_t\,, \tag{25}$$

$$C_t = (1 - F_t) \cdot I\left(\Delta y_t > \tau_C\right) \cdot I\left(\Delta y_{t-1} > \tau_C\right). \tag{26}$$

The indicator variable $F_t = \{0, 1\}$ denotes whether the economy is in the "floor" regime, while $C_t = \{0, 1\}$ denotes whether the economy is in the "ceiling" regime. The CDR variable is the same as in (20) if the threshold $\tau_F = 0$. Thus, a high-growth post-recession recovery is implied by $\lambda_1 > 0$. In particular, with $\tau_F = 0$, the "floor" regime is activated when real GDP falls below its historical maximum at the onset of a recession and remains activated until output recovers back to its pre-recession level. The OH variable captures whether real GDP is above a sustainable level based on the threshold level τ_C of growth. A capacity-constraint effect is implied by $\lambda_2 < 0$. Note, however, that the "ceiling" regime that underlies the OH variable can be activated only when the "floor" regime is off, ruling out the possibility that a high-growth recovery from the trough of a recession is labeled as "overheating". There is also a requirement of two consecutive quarters of fast growth above the threshold level τ_C in order to avoid labeling a single quarter of fast growth as "overheating". Meanwhile, a heteroskedastic version of the model allows the variance of the shocks to evolve as follows:

$$\sigma_t^2 = \sigma_1^2 \cdot I\left(F_{t-1} + C_{t-1} = 0\right) + \sigma_2^2 F_{t-1} + \sigma_3^2 C_{t-1}\,. \tag{27}$$

Also, in a triumph of controlled complexity, Koop and Potter [92] develop a multivariate version of this model, discussed later.

A related literature on STAR models of business cycle asymmetry includes Teräsvirta and Anderson [143], Teräsvirta [141], van Dijk and Franses [148], and Öcal and Osborn [117]. Similar to the development of Markov-switching models and TAR models, van Dijk and Franses [148] develop a multi-regime STAR model and find evidence for more than two regimes in economic activity. Likewise, using U.K. data on industrial production, Öcal and Osborn [117] find evidence for three regimes corresponding to recessions, normal growth, and high growth. Rothman, van Dijk, and Franses [130] develop a multivariate STAR model to examine nonlinearities in the relationship between money and output.

While there are many different nonlinear models of economic activity, it should be noted that, in a general sense, Markov-switching models and threshold models are close substitutes for each other in terms of their abilities to forecast (see [23]) and their abilities to capture business cycle asymmetries such as deepness, steepness, and sharpness (see [24]). On the other hand, specific models are particularly useful for capturing specific asymmetries and, as discussed next, for testing the presence of nonlinear dynamics in macroeconomic time series.

Evidence

While estimates for regime-switching models often imply the presence of business cycle asymmetries, it must be acknowledged that the estimates may be more the consequence of the flexibility of nonlinear models in fitting the data than any underlying nonlinear dynamics. In the regime-switching model context, an extreme example of over-fitting comes from a basic i.i.d. mixture model. If the mean and variance are allowed to be different across regimes, the sample likelihood will approach infinity as the estimated variance approaches zero in a regime for which the estimated mean is equal to a sample observation. (It should be noted, however, that the highest local maximum of the sample likelihood for this model produces consistent estimates of the model parameters. See [73]). Thus, it is wise to be skeptical of estimates from nonlinear models and to seek out a correct sense of their precision. Having said this, the case for nonlinear dynamics that correspond to business cycle asymmetries is much stronger than it is often made out to be, although it would be a mistake to claim the issue is settled.

From the classical perspective, the formal problem of testing for nonlinearity with regime-switching models is that the models involve *nuisance parameters* that are not identified under the null hypothesis of linearity, but influence the distributions of test statistics. For example, Hamilton's [59] model outlined in (8)–(9) collapses to a linear AR model if $\mu_1 = \mu_2$. However, under this null hypothesis, the two independent transition probabilities p_{11} and p_{22} in the transition matrix will no longer be identified (i. e., they can take on different values without changing the fit of the model). The lack of identification of these nuisance parameters is referred to as the Davies [32] problem and it means that test statistics of the null hypothesis such as a t-statistic or a likelihood ratio (LR) statistic will not have their standard distributions, even asymptotically. An additional problem for Markov-switching models is that the null hypothesis of linearity often corresponds to a local maximum for the likelihood, meaning that the score is identically zero for some parameters, thus violating a standard assumption in classical testing theory. The problem of an identically zero score is easily seen by noting that one of the fundamental tests in classical statistics, the Lagrange multiplier (LM) test, is based on determining whether the score is significantly different than zero when imposing the null hypothesis in a more general model. For Hamilton's [59] model, the scores are zero for $\mu_d = \mu_2 - \mu_1$, p_{11}, and p_{22}. Again, identically zero scores imply nonstandard distributions for a t-statistic or an LR statistic. In practice, these nonstandard distributions mean that, if researchers were to apply standard critical values, they would over-reject linearity.

Hansen [61] derives a bound for the asymptotic distribution of a likelihood ratio statistic in the setting of unidentified nuisance parameters and identically zero scores. The bound is application-specific as it depends on the covariance function of an empirical process associated with the likelihood surface in a given setting (i. e., it is model and data dependent). The distribution of the empirical process can be obtained via simulation. In his application, Hansen [61] tests linearity in US real GNP using Hamilton's [59] model. His upper bound for the p-value of the likelihood ratio test statistic is far higher than conventional levels of significance. Thus, he is unable to reject linearity with Hamilton's [59] model. However, when he proposes an extended version of the model that assumes switching in the intercept and AR coefficients, rather than the mean as in (8)–(9), he is able to reject linearity with an upper bound for the p-value of 0.02.

In a subsequent paper, Hansen [62] develops a different method for testing in the presence of unidentified nuisance parameters that yields an exact critical value rather than an upper bound for a p-value. Again, the method requires simulation, as the critical value is model and data dependent. However, this approach assumes nonzero scores and is, therefore, more appropriate for testing threshold models than Markov-switching models. In his

application for this approach, Hansen [62] tests linearity in US real GNP using Potter's [121] TAR model mentioned earlier (see also [17,63,146,147], on testing TAR models and [140], on testing STAR models). Referring back to the TAR model in (17), the threshold τ and the delay parameter m are unidentified nuisance parameters under the null of linearity (i. e., the case where the AR parameters and any other parameters that are allowed to switch in the model are actually the same across regimes). Hansen [62] finds that the p-values for a variety of test statistics are above conventional levels of significance, although the p-value for the supLM (i. e., the largest LM statistic for different values of the nuisance parameters) under the hypothesis of homoskedastic errors is 0.04, thus providing some support for nonlinearity.

Garcia [53] reformulates the problem of testing for Markov-switching considered in Hansen [61] by proceeding as if the score with respect to the change in Markov-switching parameters (e. g., $\mu_d = \mu_2 - \mu_1$ for Hamilton's [59], model) is not identically zero and examining whether the resulting asymptotic distribution for a likelihood ratio test statistic is approximately correct. The big advantage of this approach over Hansen [61] is that the distribution is no longer sample-dependent, although it is still model-dependent. Also, it yields an exact critical value instead of an upper bound for the p-value. Garcia [53] reports asymptotic critical values for some basic Markov-switching models with either no linear dynamics or a mild degree of AR(1) linear dynamics ($\phi = 0.337$) and compares these to critical values based on a simulated distribution of the LR statistic under the null of linearity and a sample size of 100. He finds that his asymptotic critical values are similar to the simulated critical values for the simple models, suggesting that they may be approximately correct despite the problem of an identically zero score. The asymptotic critical values are considerably smaller than the simulated critical values in the case of Hamilton's [59] model with an AR(4) specification, although this is perhaps due to small sample issues rather than approximation error for the asymptotic distribution. Regardless, even with the asymptotic critical values, Garcia [53] is unable to reject linearity for US real GNP using Hamilton's [59] model at standard levels of significance, although the p-value is around 0.3 instead of the upper bound of around 0.7 for Hansen [61].

It is worth mentioning that the simulated critical values in Garcia's [53] study depend on the values of parameters used to simulate data under the null hypothesis. That is, the LR statistic is not *pivotal*. Thus, the approach of using the simulated critical values to test linearity would correspond to a parametric bootstrap test (see [105,106], for excellent overviews of bootstrap methods). The use of bootstrap tests (sometimes referred to as Monte Carlo tests, although see MacKinnon [105,106], for the distinction) for Markov-switching models has been limited (although see [96], for an early example) for a couple of reasons. First, the local maximum at the null hypothesis that is so problematic for asymptotic theory is also problematic for estimation. While a researcher is likely to re-estimate a nonlinear model using different starting values for the parameters when an optimization routine converges to this or another local maximum in an application, it is harder to do an exhaustive search for the global maximum for each bootstrap sample. Thus, the bootstrapped critical value may be much lower than the true critical value (note, however, that Garcia's [53], bootstrapped critical values were considerably higher than his asymptotic critical values). Second, given the unidentified nuisance parameters, the test statistic may not even be asymptotically pivotal. Thus, it is unclear how well the bootstrapped distribution approximates the true finite sample distribution. Despite this, bootstrap tests have often performed better in terms of *size* (the probability of false rejection of the null hypothesis in repeated experiments) than asymptotic tests in the presence of unidentified nuisance parameters. For example, Diebold and Chen [37] consider Monte Carlo analysis of bootstrap and asymptotic tests for structural change with an unknown breakpoint that is a nuisance parameter and find that the bootstrap tests perform well in terms of size and better than the asymptotic tests. Enders, Falk, and Siklos [44] find that bootstrap and asymptotic tests both have size problems for TAR models, although bootstrap LR tests perform better than the asymptotic tests or other bootstrap tests. In terms of testing for nonlinearity with Markov-switching models, Kim, Morley, and Piger [83] bootstrap the distribution of the LR statistic testing linearity for the bounceback model discussed above and reject linearity with a p-value of less than 0.01. The local maximum problem is addressed by conducting a grid search across transition probabilities.

In a recent paper, Carrasco, Hu, and Ploberger [15] develop an information matrix-type test for Markov-switching that is asymptotically optimal and only requires estimation under the null of no Markov-switching (their null allows for other forms of nonlinearity such as ARCH). At this point, there is little known about the finite sample properties of the test. However, Carrasco, Hu, and Ploberger [15] show that it has higher *power* (probability of correct rejection of the null hypothesis in repeated experiments) than Garcia's [53] approach for a basic Markov-switching model with no autoregressive dynamics. Hamilton [60] applies Carrasco, Hu, and

Ploberger's [15] method to test for Markov switching in the US unemployment rate (he also provides a very helpful appendix describing how to conduct the test). The null hypothesis is a linear AR(4) model with student t errors. The alternative is an AR(4) with student t errors where the intercept is Markov-switching with three regimes. The test statistic is 26.02, while the 5 percent critical value is 4.01. Thus, linearity can be rejected for the unemployment rate. Meanwhile, the estimated Markov-switching model implies asymmetry in the form of steepness (the unemployment rate rises above its average more quickly than it falls below its average rate).

In contrast to Markov-switching models or threshold models, Beaudry and Koop's [6] ARMA model with the CDR variable provides a very simple test of nonlinearity. In particular, for their preferred specification, Beaudry and Koop [6] find support for nonlinearity with a t-statistic of 3.39 for the CDR variable. Hess and Iwata [68] question the significance of this statistic on the basis of Monte Carlo analysis. However, the data generating process in their Monte Carlo study assumed no drift in the simulated "output" series, meaning that the simulated CDR variable behaves much like a unit root process. By contrast, given drift, the CDR variable can be expected to revert to zero over a fairly short horizon, as it does in the real world (see Fig. 8). Elwood [43] develops an unobserved components model with a threshold process for the transitory component and argues that there is no evidence for asymmetry in the responses to positive and negative shocks. However, his model does not confront the key distinction between large negative shocks versus other shocks that Beaudry and Koop [6] address directly with the inclusion of the CDR variable in their model. A more fundamental issue is whether the CDR variable is merely a proxy for another variable such as the unemployment rate or interest rates and the apparent nonlinearity is simply the result of an omitted variable. However, as discussed in more detail later, the results in Clarida and Taylor [22] and Morley and Piger [113] suggest that Beaudry and Koop's [6] model is capturing a nonlinear dynamic that is fundamentally different than what would be implied by any linear model.

Hess and Iwata [69] provide a more formidable challenge to Beaudry and Koop's [6] model, and, indeed, to many of the regime-switching models discussed earlier, by examining the relative abilities of linear and nonlinear models to reproduce particular features of US real GDP. This alternative form of model evaluation is related to encompassing tests for non-nested models (see [110], on encompassing tests and [9], on the use of encompassing tests to evaluate Markov-switching models). In particular, Hess and Iwata [69] simulate data from a variety of models of output growth, including an AR(1) model, an ARMA(2,2) model, Beaudry and Koop's [6] model, Potter's [121] two-regime TAR model, Pesaran and Potter's [120] "floor and ceiling" model, Hamilton's [59] two-regime Markov-switching model, and a three-regime Markov-switching model with restrictions on the transition matrix as in Boldin [8]. They then consider whether the simulated data for each model can successfully reproduce "business cycle" features in terms of the duration and amplitude of expansions and recessions. Their definition of the business cycle is related to the level of real GDP. However, they label any switch between positive and negative growth, no matter how short-lived, to be a business cycle turning point. For US real GDP, their approach identifies twice as many turning points as reported by the NBER. Under this definition, Hess and Iwata [69] find that the linear AR(1) model is better than the nonlinear models at reproducing the duration and amplitude of "expansions" and "recessions" in US real GDP.

Harding and Pagan [65] and Engel, Haugh, and Pagan [45] confirm Hess and Iwata's [69] findings of little or no "value-added" for nonlinear models over linear models using a business cycle dating procedure that more closely matches NBER dates. The procedure is a quarterly version of an algorithm by Bry and Broschan [12] and identifies recessions as being related to two consecutive quarters of decline in real GDP. In terms of nonlinear models, Engel, Haugh, and Pagan [45] move beyond Hess and Iwata [69] by considering van Dijk and Franses' [149] version of the floor and ceiling model with ARCH errors, Kim, Morley, and Piger's [83] bounceback model, and DeJong, Liesenfeld, and Richard's [34] tension index model. Meanwhile, Clements and Krolzig [25] find that multivariate two-regime Markov-switching models provide little improvement over linear models in capturing business cycle features.

However, beyond the issue of how to define a business cycle, the major question in the literature on reproducing business cycle features is which features to consider. Galvão [52], Kim, Morley, and Piger [83], and Morley and Piger [111] examine the ability of linear and nonlinear models to capture high-growth recoveries that are related to the severity of the preceding recessions, which is the asymmetry emphasized by Friedman [50], Wynne and Balke [150], Sichel [134], and Balke and Wynne [2]. When considering this feature, there is strong support for Kim and Nelson's [79] plucking model and Kim, Morley, and Piger's [83] bounceback model over linear models. Interestingly, the three-regime Markov-switching model does not reproduce this feature. In particular, even though it implies high-growth recoveries, the fixed transition prob-

abilities mean that the strength of the recovery is independent of the severity of the preceding recession. However, the strong support for the plucking model and bounceback model over linear models when considering the relationship between recessions and their recoveries represents a major reversal of the earlier findings for linear models by Hess and Iwata [69] and others.

In terms of directly testing business cycle asymmetries, DeLong and Summers [35] consider a nonparametric test for steepness in real GNP and unemployment rates for eight countries (including the US). In particular, they test for skewness in output growth rates and changes in unemployment rates. With the exception of changes in the US unemployment rate, the measures of economic activity produce no statistically significant evidence of skewness, although the point estimates are generally large and negative for output growth and large and positive for the unemployment rates. Of course, given that the nonparametric test of skewness is unlikely to have much power for the relatively small sample sizes available in macroeconomics, it is hard to treat the non-rejections as particularly decisive. In a more parametric setting, Goodwin [56] considers a likelihood ratio test for sharpness using Hamilton's [59] model. Applying the model and test to real GNP for eight countries (including the US), he is able to reject non-sharpness in every country except Germany. In a more general setting, Clements and Krolzig [24] develop tests of deepness, steepness, and sharpness that are conditional on the number of regimes. For a three-regime model, they are able to reject the null hypotheses of no steepness and no sharpness in US real GDP growth, although the test results are somewhat sensitive to the sample period considered. Meanwhile, Ramsey and Rothman [124] develop a test of time reversibility and find that many measures of economic activity are irreversible and asymmetric, although the nature of the irreversibility does not always provide evidence for nonlinearity.

In addition to classical tests of nonlinear models and the encompassing-style approach discussed above, there are two other approaches to testing nonlinearity that should be briefly mentioned: nonparametric tests and Bayesian model comparison. In terms of nonparametric tests, there is some evidence for nonlinearity in macroeconomic time series. For example, Brock and Sayers [11] apply the nonparametric test for independence (of "prewhitened" residuals using a linear AR model) developed by Brock, Dechert, and Schienkman [10] and are able to reject linearity for the US unemployment rate and industrial production. However, as is always the case with such general tests, it is not clear what alternative is being supported (i. e., is it nonlinearity in the conditional mean or time-variation in the conditional variance?). Also, again, the nonparametric approach is hampered in macroeconomics by relatively small sample sizes. In terms of Bayesian analysis, there is some support for nonlinearity related to business cycle asymmetry using Bayes factors for multivariate models (see [80]). Bayes factors correspond to the posterior odds of one model versus another given equal prior odds. In essence, they compare the relative abilities of two models to predict the data given the stated priors for the model parameters. Obviously, Bayes factors can be sensitive to these priors. However, given diffuse priors, they have a tendency to favor more tightly parametrized models, as some of the prior predictions from the more complicated models can be wildly at odds with the data. Thus, because the findings in favor of nonlinear models correspond to relatively more complicated models, evidence for nonlinearity using Bayes factors is fairly compelling.

Relevance

Even accepting the presence of nonlinear dynamics related to business cycle asymmetry, there is still a question of economic relevance. Following the literature, the case can be made for relevance in three broad, but related areas: forecasting, macroeconomic policy, and macroeconomic theory.

In terms of forecasting, the nonlinear time series models discussed earlier directly imply different conditional forecasts than linear models. Beaudry and Koop's [6] model provides a simple example with a different implied persistence for large negative shocks than for other shocks. By contrast, linear models imply that the persistence of shocks is invariant to their sign or size. Koop, Pesaran, and Potter [94] develop "generalized impulse response functions" to examine shock dynamics for nonlinear models. Their approach involves simulating artificial time series both in the presence of the shock and in the absence of the shock, holding all else (e. g., other shocks) equal, and comparing the paths of the two simulated time series. This simulation can be done repeatedly for different values of other shocks in order to integrate out their impact on the difference in conditional expectations of the time series implied by presence and absence of a shock. Clarida and Taylor [22] use related simulated forecasts to carry out the Beveridge–Nelson (BN) decomposition (see [7]) for US real GNP using Beaudry and Koop's [6] model. The BN decomposition produces estimates of the permanent and transitory components of a time series based on long-horizon conditional forecasts. Importantly, the estimated cycle (under the "deviations" definition of the business cycle)

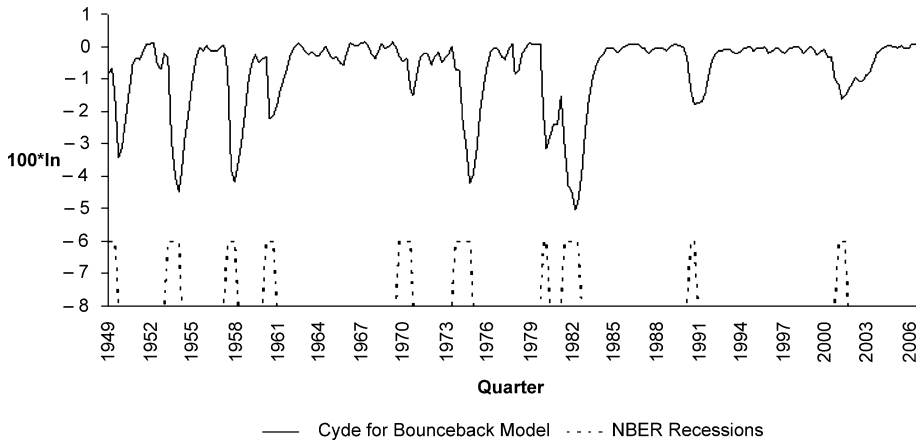

Macroeconomics, Non-linear Time Series in, Figure 9
"Bounceback" cycle and NBER recessions (Source: Author's calculations based on Morley and Piger [113], and NBER website)

for Beaudry and Koop's [6] model displays deepness that would be difficult to replicate with any linear forecasting model, even with multivariate information. Thus, there is a direct sense in which Beaudry and Koop's [6] model is not just approximating a linear multivariate model.

In a recent paper, Morley and Piger [112] develop an extension of the BN decomposition that produces optimal (in a "minimum mean squared error" sense) estimates of the cyclical component of an integrated time series when the series can be characterized by a regime-switching process such as for a Markov-switching model with fixed transition probabilities. The approach, which is labeled the "regime-dependent steady-state" (RDSS) decomposition, extracts the trend by constructing a long-horizon forecast conditional on remaining in a particular regime (hence, "regime-dependent"). In Morley and Piger [113], the RDSS decomposition is applied to US real GDP using the "depth" version of Kim, Morley, and Piger's [83] bounceback model given by (8) and (16). Figure 9 presents the estimated cycle for a version of the model with a structural break in σ^2, μ_1, and μ_2 in 1984:Q2 to account for the Great Moderation. The figure also displays an indicator variable for NBER-dated recessions for each quarter from 1949:Q2 to 2006:4. (For visual ease, the indicator variable is -8 in expansions and -6 in recessions).

There are three particularly notable features of the cycle in Fig. 9. First, there is a close correspondence between the big negative movements in it and the NBER-dated periods of recession. Thus, in practice, there is a direct relationship between the level and deviations definitions of the business cycle discussed earlier. Also, this correspondence directly implies that the NBER is identifying a meaningful macroeconomic structure (i. e., it is capturing a phase that is closely related to large movements in the transitory component of economic activity), rather than merely noting negative movements in economic activity. Second, it is fairly evident from Fig. 9 that the cycle displays all three business cycle asymmetries in the form of deepness, steepness, and sharpness. Third, the unconditional mean of the cycle is negative. As discussed in Morley and Piger [113], this finding stands in contrast to cyclical estimates for all linear models, whether univariate or multivariate.

The negative mean of the cycle in US real GDP has strong implications for the potential benefits of macroeconomic stabilization policy. Lucas [102,103] famously argued that the elimination of all business cycle fluctuations would produce a benefit equivalent to less than one-tenth of one percent of lifetime consumption. One reason for this extraordinarily low estimate is that his calculation assumes business cycle fluctuations are symmetric. However, as discussed in DeLong and Summers [36], Cohen [28], Barlevy [5], and Yellen and Akerlof [151], a nonzero mean cyclical component of economic activity directly implies that stabilization policies, if effective, could raise the average level of output and lower the average level of the unemployment rate. In this setting, the potential benefits of stabilization policy are much larger than calculated by Lucas [102,103]. (In deference to Milton Friedman and his plucking model, it is worth mentioning that the optimal "stabilization" policy might be a passive rule that prevents policymakers from generating recessionary shocks in the first place. Regardless, the point is that, given a negative mean for the cycle in real GDP, the costs of business cycles are high and can be affected by policy).

A related issue is asymmetry in terms of the effects of macroeconomic policy on economic activity. For example,

DeLong and Summers [36] and Cover [31] find that negative monetary policy shocks have a larger effect on output than positive shocks of the same size (the so-called "pushing on a string" hypothesis). This form of asymmetry represents a third type of nonlinearity in macroeconomics beyond structural change and business cycle asymmetry, although it is clearly related to business cycle asymmetry. Indeed, Garcia and Schaller [54] and Lo and Piger [99] consider Markov-switching models and find that asymmetry in the effects of monetary policy shocks is more closely related to whether the economy is in an expansion or a recession, rather than whether the shock was positive or negative. In particular, positive shocks can have large effects on output, but only in recessions. There is an obvious link between this result, which is suggestive of a convex short-run aggregate supply curve rather than the "pushing on a string" hypothesis, and the business cycle displayed in Fig. 9, which is also highly suggestive of a convex short-run aggregate supply curve.

In addition to the implications for more traditional theoretical notions in macroeconomics such as the shape of the short-run aggregate supply curve, the findings for business cycle asymmetry are important for modern macroeconomic theory because dynamic stochastic general equilibrium (DSGE) models are often evaluated and compared based on their ability to generate internal propagation that matches what would be implied by linear AR and VAR models of US real GDP (see, for example [126]). These linear models imply a time-invariant propagation structure for shocks, while the business cycle presented in Fig. 9 suggests that theory-based models should instead be evaluated on their ability to generate levels of propagation that vary over business cycle regimes, at least if they are claimed to be "business cycle" models.

Future Directions

There are several interesting avenues for future research in nonlinear time series in macroeconomics. However, two follow directly from the findings on nonlinearities summarized in this survey. First, in terms of structural change, it would be useful to determine whether the process of change is gradual or abrupt and the extent to which it is predictable. Second, in terms of business cycle asymmetries, it would be useful to pin down the extent to which they reflect nonlinearities in conditional mean dynamics, conditional variance dynamics, and/or the contemporaneous relationship between macroeconomic variables.

The issue of whether structural change is gradual or abrupt is only meaningful when structural change is thought of as a form of nonlinearity in a time series model. In particular, formal classical tests of structural change based on asymptotic theory make no distinction between whether there are many small change or a few large changes. All that matters is the cumulative magnitude of changes over the long horizon (see [42], on this point). Of course, a time-varying parameter model and a regime-switching model with permanent changes in regimes can fit the data in very different ways in finite samples. Thus, it is possible to use finite-sample model comparison (e. g., Bayes factors) to discriminate between these two behaviors. It is even possible to use a particle filter to estimate a nonlinear state-space model that nests large, infrequent changes and small, frequent changes (see [90]). In terms of predicting structural change, Koop and Potter [93] develop a flexible model that allows the number of structural breaks in a given sample and the duration of structural regimes to be stochastic processes and discuss estimation of the model via Bayesian methods.

The issue of the relative importance of different types of recurring nonlinearities is brought up by the findings in Sims and Zha [136], discussed earlier, that there are no changes in the conditional mean dynamics, but only changes in the conditional variance of shocks for a structural VAR model of the US economy. Likewise, in their multivariate three-regime TAR model, Koop and Potter [92] consider a VAR structure, and find that a linear VAR structure with heteroskedastic errors is preferred over a "vector floor and ceiling" structure for the conditional mean dynamics. The question is how to reconcile these results with the large body of evidence supporting nonlinearity in conditional mean dynamics discussed at length in this survey. A short answer is that VAR models are highly parametrized in terms of the conditional mean. Thus, it may be hard to identify regime shifts or nonlinear forms of time-variation in conditional means using a VAR model, even if they are present. On the other hand, even for their nonlinear model, Koop and Potter [92] find stronger evidence for nonlinearity in the contemporaneous relationship between variables than in the conditional mean dynamics. Meanwhile, in terms of multivariate analysis, consideration of more parsimonious factor models has typically increased the support for nonlinear models over linear models (e. g. [80]). Thus, a full comparison of different types of nonlinearity in the context of a parsimonious nonlinear model would be useful.

Another important avenue for future research in macroeconomics is an increased integration of the findings in nonlinear time series into macroeconomic theory. In terms of structural change, there has been considerable progress in recent years. In particular, some of the papers on changes in policy regimes discussed earlier (e. g. [123,

136]) can be classified as "theory-oriented" given their consideration of structural VAR models. Another nonlinear time series paper on changing policy regimes with a structural model is Owyang and Ramey [118], which considers the interaction between regime switching in the Phillips curve and the policy rule. Meanwhile, Fernández-Villaverde and Rubio-Ramírez [47] and King [89] directly incorporate structural change (of the gradual form) in theory-based DSGE models, which they proceed to estimate with the aid of particle filters. In terms of Bayesian analysis of the sources of the Great Moderation, Chauvet and Potter [20] and Kim, Nelson, and Piger [82] consider disaggregated data (in a joint model and separately, respectively) and find that the decline in volatility of economic activity is a broadly-based phenomenon, rather than corresponding to particular sectors, while Kim, Morley, and Piger [86] employ structural VAR models and find that the decline in volatility cannot be explained by changes in aggregate demand shocks, monetary policy shocks, or the response of the private sector or policymakers to shocks.

In terms of the integration of business cycle asymmetries into macroeconomic theory, there has been less progress in recent years, perhaps due the obviously greater difficulty in modeling endogenous regime switching than in simply assuming exogenous structural change. However, the theoretical literature contains some work on asymmetries. In particular, mechanisms for regime switching in the aggregate data that have been considered in the past include spillovers and strategic complementarities [41], animal spirits [70], a history-dependent selection criterion in an economy with multiple Nash equilibria corresponding to different levels of productivity [30], and intertemporal increasing returns [1]. However, Potter [122] notes that, while these mechanisms can generate regime switching in the aggregate data, they cannot explain asymmetry in the form of high-growth recoveries following large negative shocks. He proposes a model with Bayesian learning and an information externality (see [16]) that can generate such dynamics. Meanwhile, in terms of business cycle asymmetry more generally, obvious mechanisms are investment irreversibilities [55] and capacity constraints [64]. More promisingly for future developments in macroeconomic theory, there is a growing empirical literature on the sources of business cycle asymmetries. For example, Korenok, Mizrach, and Radchenko [95] use disaggregated data and find that asymmetries are more pronounced in durable goods manufacturing sectors than nondurable goods manufacturing sectors (also see [129]) and appear to be related to variation across sectors in credit conditions and reliance on raw material inventories, while they do not appear to be related to oil price shocks [33] or adjustment costs [3].

Bibliography

Primary Literature

1. Acemoglu D, Scott A (1997) Asymmetric business cycles: Theory and time-series evidence. J Monet Econ 40:501–533
2. Balke NS, Wynne MA (1996) Are deep recessions followed by strong recoveries? Results for the G-7 countries. Appl Econ 28:889–897
3. Ball L, Mankiw NG (1995) Relative price changes as aggregate supply shocks. Q J Econ 110:161–193
4. Bansal R, Yaron A (2004) Risks for the long run: A potential resolution of asset pricing puzzles. J Financ 59:1481–1509
5. Barlevy G (2005) The cost of business cycles and the benefits of stabilization. Econ Perspect 29:32–49
6. Beaudry P, Koop G (1993) Do recessions permanently change output? J Monet Econ 31:149–163
7. Beveridge S, Nelson CR (1981) A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the business cycle. J Monet Econ 7:151–174
8. Boldin MD (1996) A check on the robustness of Hamilton's Markov switching model approach to the economic analysis of the business cycle. Stud Nonlinear Dyn Econom 1:35–46
9. Breunig R, Najarian S, Pagan A (2003) Specification testing of Markov-switching models. Oxf Bull Econ Stat 65:703–725
10. Brock WA, Dechert WD, Scheinkman JA (1996) A test of independence based on the correlation dimension. Econom Rev 15:197–235
11. Brock WA, Sayers C (1988) Is the business cycle characterized by deterministic chaos? J Monet Econ 22:71–90
12. Bry G, Boschan C (1971) Cyclical analysis of time series: Selected procedures and computer programs. NBER, New York
13. Burns AF, Mitchell WA (1946) Measuring Business Cycles. NBER, New York
14. Camacho M (2005) Markov-switching stochastic trends and economic fluctuations. J Econ Dyn Control 29:135–158
15. Carrasco M, Hu L, Ploberger W (2007) Optimal test for Markov switching. Working Paper
16. Chalkley M, Lee IH (1998) Asymmetric business cycles. Rev Econ Dyn 1:623–645
17. Chan KS (1991) Percentage points of likelihood ratio tests for threshold autoregression. J Royal Stat Soc Ser B 53:691–696
18. Chan KS, Tong H (1986) On estimating thresholds in autoregressive models. J Tim Ser Analysis 7:179–190
19. Chauvet M (1998) An econometric characterization of business cycle dynamics with factor structure and regime switches. Int Econ Rev 39:969–996
20. Chauvet M, Potter S (2001) Recent changes in the US business cycle. Manch Sch 69:481–508
21. Chib S, Nardari F, Shephard N (2002) Markov chain Monte Carlo methods for stochastic volatility models. J Econom 108:281–316
22. Clarida RH, Taylor MP (2003) Nonlinear permanent-temporary decompositions in macroeconomics and finance. Econ J 113:C125–C139
23. Clements MP, Krolzig HM (1998) A comparison of the forecast

performance of Markov-switching and threshold autoregressive models of US GNP. Econ J 1:C47–C75
24. Clements MP, Krolzig HM (2003). Business cycle asymmetries: Characterization and testing based on Markov-switching autoregressions. J Bus Econ Stat 21:196–211
25. Clements MP, Krolzig HM (2004) Can regime-switching models reproduce the business cycle features of US aggregate consumption, investment and output? Int J Financ Econ 9:1–14
26. Cogley T, Sargent TJ (2001) Evolving post-World War II US inflation dynamics. In: Bernanke BS, Rogoff K (eds) NBER Macroeconomics Annual 2001. MIT Press, Cambridge, pp 331–373
27. Cogley T, Sargent TJ (2005) Drift and volatilities: Monetary policies and outcomes in the post WW II US. Rev Econ Dyn 8:262–302
28. Cohen D (2000) A quantitative defense of stabilization policy. Federal Reserve Board Finance and Economics Discussion Series. Paper 2000-34
29. Cooley TF, Prescott EC (1976) Estimation in the presence of stochastic parameter variation. Econometrica 44:167–184
30. Cooper R (1994) Equilibrium selection in imperfectly competitive economies with multiple equilibria. Econ J 104:1106–1122
31. Cover JP (1992) Asymmetric effects of positive and negative money-supply shocks. Q J Econ 107:1261–1282
32. Davies RB (1977) Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 64:247–254
33. Davis SJ, Haltiwanger J (2001) Sectoral job creation and destruction responses to oil price changes. J Monet Econ 48:468–512
34. DeJong DN, Liesenfeld R, Richard JF (2005) A nonlinear forecasting model of GDP growth. Rev Econ Stat 87:697–708
35. DeLong JB, Summers LH (1986) Are business cycles symmetrical? In: Gordon RJ (ed) The American Business Cycle. University of Chicago Press, Chicago, pp 166–179
36. DeLong B, Summers L (1988) How does macroeconomic policy affect output? Brook Papers Econ Activity 2:433–480
37. Diebold FX, Chen C (1996) Testing structural stability with endogenous breakpoint: A size comparison of analytic and bootstrap procedures. J Econ 70:221–241
38. Diebold FX, Rudebusch GD (1996) Measuring business cycles: A modern perspective. Rev Econ Stat 78:67–77
39. Diebold FX, Lee JH, Weinbach G (1994) Regime switching with time-varying transition probabilities. In: Hargreaves C (ed) Nonstationary Time Series Analysis and Cointegration. Oxford University Press, Oxford, pp 283–302
40. Durland JM, McCurdy TH (1994) Duration-dependent transitions in a Markov model of US GNP growth. J Bus Econ Stat 12:279–288
41. Durlauf SN (1991) Multiple equilibria and persistence in aggregate fluctuations. Am Econ Rev Pap Proc 81:70–74
42. Elliott G, Müller U (2006) Efficient tests for general persistent time variation in regression coefficients. Rev Econ Stud 73:907–940
43. Elwood SK (1998) Is the persistence of shocks to output asymmetric? J Monet Econ 41:411–426
44. Enders W, Falk BL, Siklos P (2007) A threshold model of real US GDP and the problem of constructing confidence intervals in TAR models. Stud Nonlinear Dyn Econ 11(3):4
45. Engel J, Haugh D, Pagan A (2005) Some methods for assessing the need for non-linear models in business cycles. Int J Forecast 21:651–662
46. Engle RF (1982) Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50:987–1007
47. Fernández-Villaverde J, Rubio-Ramírez JF (2007) Estimating macroeconomic models: A likelihood approach. Rev Econ Stud 54:1059–1087
48. Filardo AJ (1994) Business-cycle phases and their transitional dynamics. J Bus Econ Stat 12:299–308
49. French MW, Sichel DE (1993) Cyclical patterns in the variance of economic activity. J Bus Econ Stat 11:113–119
50. Friedman M (1964) Monetary Studies of the National Bureau, the National Bureau Enters Its 45th Year. 44th Annual Report. NBER, New York, pp 7–25; Reprinted in: Friedman M (1969) The Optimum Quantity of Money and Other Essays. Aldine, Chicago, pp 261–284
51. Friedman M (1993) The "plucking model" of business fluctuations revisited. Econ Inq 31:171–177
52. Galvão AB (2002) Can non-linear time series models generate US business cycle asymmetric shape? Econ Lett 77:187–194
53. Garcia R (1998) Asymptotic null distribution of the likelihood ratio test in Markov switching models. Int Econ Rev 39:763–788
54. Garcia R, Schaller H (2002) Are the effects of interest rate changes asymmetric? Econ Inq 40:102–119
55. Gilchrist S, Williams JC (2000) Putty-clay and investment: A business cycle analysis. J Political Econ 108:928–960
56. Goodwin TH (1993) Business-cycle analysis with a Markov-switching model. J Bus Econ Stat 11:331–339
57. Granger CWJ, Andersen AP (1978) An Introduction to Bilinear Time Series Models. Vandenhoek and Ruprecht, Göttingen
58. Granger CWJ, Teräsvirta T (1993) Modelling Nonlinear Economic Relationships. Oxford University Press, Oxford
59. Hamilton JD (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57:357–384
60. Hamilton JD (2005) What's real about the business cycle? Fed Reserve Bank St. Louis Rev 87:435–452
61. Hansen BE (1992) The likelihood ratio test under nonstandard conditions: Testing the Markov switching model of GNP. J Appl Econ 7:S61–S82
62. Hansen BE (1996) Inference when a nuisance parameter is not identified under the null hypothesis. Econometrica 64:413–430
63. Hansen BE (1997) Inference in TAR models. Stud Nonlinear Dyn Econom 2:1–14
64. Hansen GD, Prescott EC (2005) Capacity constraints, asymmetries, and the business cycle. Rev Econ Dyn 8:850–865
65. Harding D, Pagan AR (2002) Dissecting the cycle: A methodological investigation. J Monet Econ 49:365–381
66. Harding D, Pagan AR (2003) A Comparison of Two Business Cycle Dating Methods. J Econ Dyn Control 27:1681–1690
67. Harding D, Pagan AR (2005) A suggested framework for classifying the modes of cycle research. J Appl Econom 20:151–159
68. Hess GD, Iwata S (1997) Asymmetric persistence in GDP? A deeper look at depth. J Monet Econ 40:535–554
69. Hess GD, Iwata S (1997) Measuring and comparing business-cycle features. J Bus Econ Stat 15:432–444

70. Howitt P, McAfee RP (1992) Animal spirits. Am Econ Rev 82:493–507
71. Hristova D (2005) Maximum likelihood estimation of a unit root bilinear model with an application to prices. Stud Nonlinear Dyn Econom 9(1):4
72. Keynes JM (1936) The General Theory of Employment, Interest, and Money. Macmillan, London
73. Kiefer NM (1978) Discrete parameter variation: Efficient estimation of a switching regression model. Econometrica 46:413–430
74. Kim CJ (1994) Dynamic linear models with Markov switching. J Econom 60:1–22
75. Kim CJ, Murray CJ (2002) Permanent and transitory components of recessions. Empir Econ 27:163–183
76. Kim CJ, Nelson CR (1998) Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime switching. Rev Econ Stat 80:188–201
77. Kim CJ, Nelson CR (1999) State-Space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications. MIT Press, Cambridge
78. Kim CJ, Nelson CR (1999) Has the US economy become more stable? A Bayesian approach based on a Markov-switching model of the business cycle. Rev Econ Stat 81:608–616
79. Kim CJ, Nelson CR (1999) Friedman's plucking model of business fluctuations: Tests and estimates of permanent and transitory components. J Money Credit Bank 31:317–34
80. Kim CJ, Nelson CR (2001) A Bayesian approach to testing for Markov-switching in univariate and dynamic factor models. Int Econ Rev 42:989–1013
81. Kim CJ, Piger JM (2002) Common stochastic trends, common cycles, and asymmetry in economic fluctuations. J Monet Econ 49:1181–1211
82. Kim CJ, Nelson CR, Piger J (2004) The less-volatile US economy: A Bayesian investigation of timing, breadth, and potential explanations. J Bus Econ Stat 22:80–93
83. Kim CJ, Morley J, Piger J (2005) Nonlinearity and the permanent effects of recessions. J Appl Econom 20:291–309
84. Kim CJ, Piger J, Startz R (2007) The dynamic relationship between permanent and transitory components of US business cycles. J Money Credit Bank 39:187–204
85. Kim CJ, Piger J, Startz R (2008) Estimation of Markov regime-switching regression models with endogenous switching. J Econom 143:263–273
86. Kim CJ, Morley J, Piger J (2008) Bayesian Counterfactual Analysis of the Sources of the Great Moderation. J Appl Econom 23:173–191
87. Kim M-J, Yoo J-S (1995) New index of coincident indicators: A multivariate Markov switching factor model approach. J Monet Econ 36:607–630
88. Kim S, Shephard N, Chib S (1998) Stochastic volatility: Likelihood inference and comparison with ARCH models. Rev Econ Stud 65:361–393
89. King TB (2006) Dynamic equilibrium models with time-varying structural parameters. Working Paper
90. King TB, Morley J (2007) Maximum likelihood estimation of nonlinear, non-Gaussian state-space models using a multistage adaptive particle filter. Working Paper
91. Koop G, Potter S (2003) Bayesian analysis of endogenous delay threshold models. J Bus Econ Stat 21:93–103
92. Koop G, Potter S (2006) The vector floor and ceiling model. In: Milas C, Rothman P, Van Dijk D (eds) Nonlinear Time Series Analysis of Business Cycles. Elsevier, Amsterdam, pp 97–131
93. Koop G, Potter S (2007) Estimation and forecasting in models with multiple breaks. Rev Econ Stud 74:763–789
94. Koop G, Pesaran MH, Potter S (1996) Impulse response analysis in nonlinear multivariate models. J Econometrics 74:119–148
95. Korenok O, Mizrach B, Radchenko S (2009) A note on demand and supply factors in manufacturing output asymmetries. Macroecon Dyn (forthcoming)
96. Lam PS (1990) The Hamilton model with a general autoregressive component: Estimation and comparison with other models of economic time series. J Monet Econ 26:409–432
97. Leamer EE, Potter SM (2004) A nonlinear model of the business cycle. Working Paper
98. Leyton AP, Smith D (2000) A further note of the three phases of the US business cycle. Appl Econ 32:1133–1143
99. Lo MC, Piger J (2005) Is the response of output to monetary policy asymmetric? Evidence from a regime-switching coefficients model. J Money Credit Bank 37:865–887
100. Lucas RE (1972) Econometric testing of the natural rate hypothesis. In: Eckstein O (ed) Econometrics of Price Determination. US Federal Reserve Board, Washington DC, pp 50–59
101. Lucas RE (1976) Econometric policy evaluation: A critique. In: Brunner K, Meltzer A (eds) The Phillips Curve and Labor Markets, vol 1. Carnegie-Rochester Ser Public Policy, pp 19–46
102. Lucas RE (1987) Models of Business Cycles. Basil Blackwell, Oxford
103. Lucas RE (2003) Macroeconomic Priorities. Am Econ Rev 93:1–14
104. Ma J (2007) Consumption persistence and the equity premium puzzle: New evidence based on improved inference. Working paper
105. MacKinnon J (2002) Bootstrap inference in econometrics. Can J Econ 35:615–645
106. MacKinnon J (2006) Bootstrap methods in econometrics. Econ Rec 82:S2–S18
107. McConnell MM, Quiros GP (2000) Output fluctuations in the United States: What has changed since the early 1980s? Am Econ Rev 90:1464–1476
108. McQueen G, Thorley SR (1993) Asymmetric business cycle turning points. J Monet Econ 31:341–362
109. Mitchell WA (1927) Business Cycles: The Problem and Its Setting. NBER, New York
110. Mizon GE, Richard JF (1986) The encompassing principle and its application to non-nested hypotheses. Econometrica 54:657–678
111. Morley J, Piger J (2006) The Importance of Nonlinearity in Reproducing Business Cycle Features. In: Milas C, Rothman P, Van Dijk D (eds) Nonlinear Time Series Analysis of Business Cycles. Elsevier, Amsterdam, pp 75–95
112. Morley J, Piger J (2008) Trend/cycle decomposition of regime-switching processes. J Econom (forthcoming)
113. Morley J, Piger J (2008) The asymmetric business cycle. Working Paper
114. Morley JC, Nelson CR, Zivot E (2003) Why are the Beveridge-Nelson and unobserved-components decompositions of GDP so different? Rev Econ Stat 85:235–243
115. Neftçi SH (1984) Are economic time series asymmetric over the business cycle? J Political Econ 92:307–328

116. Niemira MP, Klein PA (1994) Forecasting Financial and Economic Cycles. Wiley, New York
117. Öcal N, Osborn DR (2000) Business cycle non-linearities in UK consumption and production. J Appl Econom 15:27–44
118. Owyang MT, Ramey G (2004) Regime switching and monetary policy measurement. J Monet Econ 51:1577–1198
119. Peel D, Davidson J (1998) A non-linear error correction mechanism based on the bilinear model. Econ Lett 58:165–170
120. Pesaran MH, Potter SM (1997) A floor and ceiling model of US output. J Econ Dyn Control 21:661–695
121. Potter SM (1995) A nonlinear approach to US GNP. J Appl Econ 10:109–125
122. Potter SM (2000) A nonlinear model of the business cycle. Stud Nonlinear Dyn Econom 4:85–93
123. Primiceri GE (2005) Time varying structural vector autogressions and monetary policy. Rev Econ Stud 72:821–852
124. Ramsey JB, Rothman P (1996) Time irreversibility and business cycle asymmetry. J Money Credit Bank 28:1–21
125. Ravn MO, Sola M (1995) Stylized facts and regime changes: Are prices procyclical? J Monet Econ 36:497–526
126. Rotemberg JJ, Woodford M (1996) Real-business-cycle Models and the forecastable movements in output, hours, and consumption. Am Econ Rev 86:71–89
127. Rothman P (1991) Further Evidence on the Asymmetric Behavior of Unemployment Rates Over the Business Cycle. J Macroeconom 13:291–298
128. Rothman P (1998) Forecasting asymmetric unemployment rates. Rev Econ Stat 80:164–168
129. Rothman P (2008) Reconsideration of Markov chain evidence on unemployment rate asymmetry. Stud Nonlinear Dyn Econo 12(3):6
130. Rothman P, van Dijk D, Franses PH (2001) A multivariate STAR analysis of the relationship between money and output. Macroeconom Dyn 5:506–532
131. Schumpeter J (1942) Capitalism, socialism, and democracy. Harper, New York
132. Sensier M, van Dijk D (2004) Testing for volatility changes in US macroeconomic time series. Rev Econ Stat 86:833–839
133. Sichel DE (1993) Business cycle asymmetry: A deeper look. Econ Inq 31:224–236
134. Sichel DE (1994) Inventories and the three phases of the business cycle. J Bus Econ Stat 12:269–277
135. Sims CA (2001) Comment on Sargent and Cogley's: Evolving Post-World War II US Inflation Dynamics. In: Bernanke BS, Rogoff K (eds) NBER Macroeconomics Annual 2001. MIT Press, Cambridge, pp 373–379
136. Sims CA, Zha T (2006) Were there regime switches in US monetary policy? Am Econ Rev 96:54–81
137. Sinclair TM (2008) Asymmetry in the business cycle: Friedman's plucking model with correlated innovations. Working Paper
138. Stock JH, Watson MW (2002) Has the business cycle changed and why? In: Gertler M, Rogoff K (eds) NBER Macroeconomics Annual 2002. MIT Press, Cambridge, pp 159–218
139. Subba Rao T, Gabr MM (1984) An Introduction to Bispectral Analysis and Bilinear Time Series Models. Lecture Notes in Statistics, vol 24. Springer, New York
140. Teräsvirta T (1994) Specification, estimation, and evaluation of smooth transition autoregressive models. J Am Stat Assoc 89:208–218
141. Teräsvirta T (1995) Modeling nonlinearity in US Gross National Product 1889–1987. Empir Econ 20:577–598
142. Teräsvirta T (1998) Modelling economic relationships with smooth transition regressions. In: Ullah A, Giles DEA (eds) Handbook of Applied Economic Statistics. Marcel Dekker, New York, pp 507–552
143. Teräsvirta T, Anderson HM (1992) Characterizing nonlinearities in business cycles using smooth transition autoregressive models. J Appl Econom 7:S119–S136
144. Tiao GC, Tsay RS (1994) Some advances in non-linear and adaptive modeling in time- series analysis. J Forecast 13:109–131
145. Tong H (1978) On a threshold model. In: Chen CH (ed) Pattern Recognition and Signal Processing. Sijhoff and Noordhoff, Amsterdam, pp 575–586
146. Tsay RS (1989) Testing and modeling threshold autoregressive processes. J Am Stat Assoc 84:231–240
147. Tsay RS (1998) Testing and modeling multivariate threshold processes. J Am Stat Assoc 93:1188–1202
148. van Dijk D, Franses PH (1999) Modeling multiple regimes in the business cycle. Macroeconom Dyn 3:311–340
149. van Dijk D, Franses PH (2003) Selecting a nonlinear time series model using weighted tests of equal forecast accuracy. Oxf Bull Econ Stat 65:727–744
150. Wynne MA, Balke NS (1992) Are deep recessions followed by strong recoveries? Econ Lett 39:183–189
151. Yellen JL, Akerlof GA (2006) Stabilization policy: A reconsideration. Econ Inq, pp 44:1–22

Books and Reviews

Davidson R, MacKinnon JG (2004) Econometric Theory and Methods. Oxford University Press, Oxford
Diebold FX (1998) The past, present, and future of macroeconomic forecasting. J Econ Perspectives 12:175–192
Engle R (2001) GARCH 101: The use of ARCH/GARCH models in applied econometrics. J Econc Perspectives 15:157–168
Franses PH (1998) Time Series Models for Business and Economic Forecasting. Cambridge University Press, Cambridge
Hamilton JD (1994) State-space models. In: Engle RF, McFadden DL (eds) Handbook of Econometrics, vol 4. Elsevier, Amsterdam, pp 041–3080
Hamilton JD (1994) Time Series Analysis. Princeton University Press, Princeton
Koop G (2003) Bayesian Econometrics. Wiley, Chichester
Teräsvirta T, Tjøstheim D, Granger CWJ (1994) Aspects of modeling nonlinear time series. In: Engle RF, McFadden DL (eds) Handbook of Econometrics, vol 4. Elsevier, Amsterdam, pp 2919–2957
Tsay RS (2005) Analysis of Financial Time Series. Wiley, Hoboken

Manipulating Data and Dimension Reduction Methods: Feature Selection

Huan Liu, Zheng Zhao
Computer Science and Engineering, Arizona State University, Tempe, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Basics of Feature Selection
Supervised Feature Selection
Unsupervised Feature Selection
Some Recent Research Development
Future Directions
Bibliography

Glossary

Feature It is a dimension of some multi-variate data. It is also called "attribute" or "variable". A feature can have continuous, discrete, or nominal values.

Feature selection It is a task of selecting some features from a given set of features according to some performance criterion.

Supervised learning It is a machine learning task that learns patterns (or hypotheses, models) from labeled data.

Unsupervised learning It is a machine learning task that learns patterns (or hypotheses, models) from unlabeled data.

Semi-supervised learning It is a machine learning task that learns patterns (or hypotheses, models) from partially labeled data.

Search It is to find a solution in a search space which contains all possible solutions. There exist a large body of search algorithms ranging from uninformed to informed.

Ranking It is to order concerned items according to some metric. Ranked features indicate their relevance or importance.

Evaluation It is an important task that measures the quality of some machine learning task, e. g., feature selection.

Definition of the Subject

Feature selection is the study of algorithms for reducing dimensionality of data for various purposes. One of the most common purposes is to improve *machine learning* performance. The other purposes include simplifying data description, streamlining data collection, improving comprehensibility of the learned models, and helping gain insight through learning.

The objective of feature selection is to remove irrelevant and/or redundant features and retain only relevant features. *Irrelevant features* can be removed without affecting learning performance. *Redundant features* are a type of irrelevant features. The distinction is that a redundant feature implies the co-presence of another feature; individually, each feature is relevant, but the removal of either one will not affect learning performance.

As a plethora of data are generated in every possible means with the exponential decreasing costs of data storage and computer processing power, data dimensionality increases on a scale beyond imagination in cases ranging from transactional data to high-throughput data. In many fields such as medicine, health care, Web search, and bioinformatics, it is imperative to reduce high dimensionality such that efficient data processing and meaningful data analysis can be conducted in order to mine nuggets from high-dimensional, massive data.

Introduction

For a dataset with N features and M dimensions (or features, attributes), feature selection aims to reduce M to M' and $M' \leq M$. For example, in bioinformatics, some gene expression dataset can have 40,000 genes, and biologists often want to reduce that large number to a manageable figure, say, a few dozens or a couple of hundreds. It is an important and widely used approach to dimensionality reduction. Another effective approach is *feature extraction*. One of the key distinctions of the two approaches lies at their outcomes. Assuming we have four features F_1, F_2, F_3, F_4, if both approaches result in 2 features, the 2 *selected* features are a subset of 4 original features (say, F_1, F_3), but the 2 *extracted* features are some combination of 4 original features (e. g., $F'_1 = \sum a_i F_i$ and $F'_2 = \sum b_i F_i$ where a_i, b_i are some constants). Feature selection is commonly used or preferred in applications where original features need to be retained. Some examples are document categorization, medical diagnosis and prognosis, gene-expression profiling. We focus our discussion on feature selection.

Broadly speaking, two factors matter most for effective learning: (1) the number of features (M), and (2) the number of instances (N). For a fixed M, a larger N means more constraints, and the resulting correct hypothesis is expected to be more reliable. For a fixed N, a decreased M amounts to a significantly increased number of instances. Consider the following *thought experiment* for a binary domain of a binary classification problem: F_1, F_2, F_3, F_4 are binary and class C is also binary (e. g., positive or negative). If the training data is of 4 instances ($N = 4$), it is only a quarter of all possible number of instances ($2^4 = 16$). The size of the hypothesis space is $2^{2^4} = 65,536$. If only two features are relevant, the size of the hypothesis space becomes $2^{2^2} = 16$, an exponential re-

duction of the hypothesis space. Now, the only available 4 instances might suffice for perfect learning if there is no duplicate instance in the reduced training data with two features. And a resulting model of 4 features can also be more complex than that of 2 features. Hence, feature selection can effectively reduce the hypothesis space, or virtually increase the number of training instances, and help create a compact model.

An unnecessarily complex learning model subjects itself to over-searching an excessively large hypothesis space. Its consequence is that the learned hypothesis *overfits* the training data and may not perform well when applying the learned model to the unseen data. Another way of describing the relationship between N and M in the context of learning is the so-called **curse of dimensionality** [2], the need for the *exponential* increase in data size associated with *linearly* adding additional dimensions to a multi-dimensional space [10]; or the concept of proximity becomes blurry in a high-dimensional space [22], resulting in degrading learning performance. Theoretically, the reduction of dimensionality can eventuate the exponential shrinkage of hypothesis space [15].

Basics of Feature Selection

Feature selection algorithms can be studied in various perspectives [21]. We choose to discuss four aspects in terms of the outcomes of feature selection (subsets or ranked lists), evaluation of feature selection algorithms, and an intuitive way of categorizing feature selection algorithms (supervised vs. unsupervised).

Searching for Feature Subsets

Selecting M' out of M features where $M' \leq M$ can be conveniently defined as a problem of search [27]. Starting with an initial state, search techniques use an explicit search tree generated by a successor function. Depending on search techniques, the search tree can define the search space (in case of exhaustive search) and can be some paths from the initial state to the goal (solution) state (in case of heuristic search). The search space generated by exhaustive search can be $O(b^d)$ where b is a branching factor and d is the depth where a solution can be found. First, we introduce the concept of search direction. In the case of searching M' out M features, one can start searching in various ways, or search direction. If one starts with an empty feature set and adds subsequent features into it one (or a few) at a time, it is called *forward search*; if one starts with a full set and removes irrelevant features one at a time, it is *backward search*. Other search directions are bi-directional, random or stochastic. Take the example of forward selection. Assuming that we have 4 features, do not know how many are relevant, but in effect 2 of them are relevant, we start with finding the best 1 out of 4 features and check if the one best feature can satisfy some performance criterion indicating we have found sufficient and necessary relevant features. If not, we need to find the best 2 out of 4 features. The search continues until we either exhaust all possibilities $\left(\binom{4}{1}, \binom{4}{2}, \text{and } \binom{4}{3}\right)$, or we find a smaller number of relevant features (2 in this example). Hence, the big O analysis of time complexity is $O(2^M)$. When M is large, it is evident that exhaustive search is infeasible.

It is necessary to resort to alternative search strategies. One commonly adopted strategy is *sequential selection*. Sequential forward selection (SFS) retains the features selected in the previous rounds, and sequential back selection (SBS) keeps out the removed features in future rounds. For example, if SFS selects F_1 among F_1, F_2, F_3, F_4 in the first round, F_1 will always remain selected. In other words, for the final two selected features, F_1 is one of them. Conversely, if F_4 is removed by SBS in the first round, it can be certain that F_4 will not be included in the final two selected features. Illustrative examples of sequential forward and backward search of are depicted in Fig. 1. The sequential search eliminates the need of retrospect-

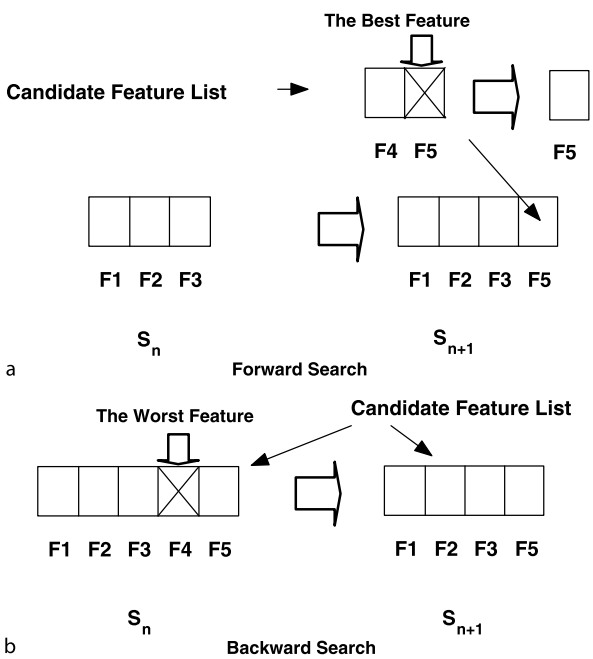

Manipulating Data and Dimension Reduction Methods: Feature Selection, Figure 1
Sequential forward and backward search in feature subset selection. a SFS to add F_5. b SBS to remove F_4

ing and thus likely makes early commitments to selected or removed features which might turn out to be suboptimal. Given the time constraint, exhaustive research for high-dimensional data can never be a practical option. The question is whether alternative search strategies such as sequential selection are sensible. As we know, given relatively few data points (instances) with high dimensionality, exhaustive search can lead to over-searching, resulting in selecting those features that might not truly relevant features if more instances were available. The reason behind it is similar to that we discuss earlier in Sect. "Introduction" about over-searching a large hypothesis space.

Now we turn our attention to another two pertinent issues: (1) the measures of feature quality evaluation, and (2) search stopping criteria. Let us assume that our purpose of feature selection is to improve learning performance (say, a classifier's accuracy). Our obvious choice of measuring feature quality is whether selected features can help improve a learning model's performance (e. g., a classifier's accuracy), and the search stops when no improvement of estimated accuracy is observed. What we just described is a so-called *wrapper* model for feature selection [14] in which a learning model's performance serves as the feature-quality measure. The shortcomings of a wrapper model are (1) the features are selected for improving the chosen learning model (e. g., a classifier), and thus might be idiosyncratic to that particular classifier, or lack of generality; in other words, for a different classifier, it is necessary to reselect features, and (2) every time when a new feature subset is considered, a new classifier needs to be built to obtain a new estimate of accuracy; this might involve some cross validation procedure such as 10-fold, 5×2-fold cross validation for reliable estimation; and thus it can be time consuming. A *filter* model of feature selection does not employ a learning model (say a classifier) in evaluating the quality of selected features. Instead, it uses some intrinsic data properties in determining how good some features are. One measure is the correlation between a feature (or a feature subset) with the class. Its gist is that the least correlated feature is removed first. Depending on how the search and comparison are performed, the selection or removal process may iterate until some stopping criterion is met. We present one example in Sect. "Supervised Feature Selection" when providing some specifics about supervised feature selection algorithms.

A gamut of stopping criteria can be employed to help determine when the selection or removal process should stop. One intuitive criterion is *the number of selected features*. Sometimes, this number is not known a priori. Another intuitive stopping criterion is that the selection or removal process converges. For example, the further iterations do not change the feature weights. If it converges very slowly, one can specify a cap for the number of iterations. For different feature quality measures [18], disparate stopping criteria should be used. Taking consistency as the measure of feature quality, one can remove features until the minimum inconsistency allowed is exceeded. It is clear that selecting a stopping criterion is not as straightforward as it seems. However, an existing feature selection algorithm is often coupled with a fixed stopping criterion. That significantly alleviates the need for a user. In addition, the domain knowledge and application requirements should be considered in the selection of feature selection algorithms and stopping criteria.

Ranking Features

Feature selection can be attained via feature ranking as well. The basic idea is to assign weights to features according to some weight updating principle. When the weight updating does not result in any weight change, features can be sorted according to their weights, and top ranked features are then selected. The two key components of feature ranking algorithms are (1) weight updating, and (2) when to stop. A simple ranking algorithm can have linear time complexity in terms of the number of features. It measures the relationship between each feature and the class attribute, and assigns bigger weights to those features having stronger relationships. For example, one can use correlation to measure the relationship between a continuous feature and the class attribute. For a discrete feature and the class, one can use *information gain* [24] as an example. One can calculate the importance of each feature with respect to the class and rank features accordingly. The advantage of this ranking approach is that it is fast. Two obvious shortcomings are (1) each feature's importance is measured independently of each other, and (2) redundant features cannot be removed as they likely have similar rankings. More sophisticated ranking algorithms have been developed. One idea is to use the boundary between the instances of different classes to determine how important features are. This treatment considers all features together in evaluating feature importance. A popular feature ranking algorithm (ReliefF) is presented in Sect. "Supervised Feature Selection" to illustrate its working details.

Feature Selection Evaluation

The end results of feature selection are a subset of selected features. Presumably, the selected feature subset should contain all relevant features. Besides, different feature selection algorithms eventuate differed results. It is incumbent to evaluate whether feature selection is effective in

one of the intended purposes we outline in the beginning. Evaluation can be made simple if we know the true, relevant features. However, the dilemma is if we know what the relevant features are, we really do not need to perform feature selection; and when we need to select features, we often have no inkling about what features are relevant. That presents a challenge to the evaluation of feature selection. The evaluation of feature selection can be addressed in two tasks. One task is whether feature selection helps or not. The second task is which algorithm is better given two feature selection algorithms. If we allow a dummy feature selection algorithm that simply selects all features, we end up having only the second task – which algorithm is better given two feature selection algorithms.

A realistic issue of evaluation for feature selection now is which algorithm should be adopted, given an application or a data set. Without knowing what relevant features are, we need to find out if the two algorithms are similar or different. One commonly used evaluation method is to check if the selected features can lead to better classification accuracy (when the training data has class labels). It consists of two basic steps. First, features subsets are selected by both feature selection algorithms (A_1 and A_2), named by F_1 and F_2. Second, employ a classifier (e. g., k-NN, SVM, Decision Tree Induction, etc.) to induce models with F_1 and F_2, and compare their corresponding accuracy rates. One feature selection algorithm is better than the other if its feature subset results in better accuracy. Depending on how accuracy is estimated, specific procedures may vary and it may be necessary to integrate the two steps. For example, if a 10-fold cross validation method is used for accuracy estimation, one may consider insert the feature selection in the cross-validation process. The ramifications of so doing and alternatives are discussed in [25].

Another relevant issue about feature selection for classification is the correct use of training data which is used in both phases of feature selection and classification. The use of the same training data can cause the so-called *feature selection bias* [28]. Theoretically, the data used in the two phases should not be correlated. However, almost without exceptions, researchers and practitioners employ the same training data for both feature selection and classification learning. The study reported in [28] investigates the effect of feature selection bias, provides concrete evidence about the minimal effect of feature selection bias in classification learning, and illustrates why so. Research in [1,30] suggests that in medicine and especially in bioinformatics, a feature selection bias could have a significant impact that can have serious consequences when applied as a clinical test. Further research on feature selection bias in terms of various performance measures should be encouraged.

So far, our discussion is based on the data with class labels. This is because the corresponding supervised learning (e. g., classification) is one of the most familiar types in machine learning and data mining. Equipped with basic concepts of feature selection, we are ready to consider various types of feature selection algorithms.

Categories of Feature Selection Algorithms

Two main types of machine learning algorithms are supervised and unsupervised based on the availability of labeling information in the data. Supervised learning is for data with class labels, and unsupervised learning is mainly for data without class labels. Feature selection algorithms can be similarly divided into supervised and unsupervised groups. In the early days of feature selection research, the majority of feature selection algorithms are supervised (refer to many algorithms surveyed in [5]). Since late 90s, the demand for unsupervised feature selection increases as data evolve with the rapid growth of computer generated data, text/Web data, and high-throughput data in genomics and proteomics [29]. Many unsupervised feature selection algorithms have been developed [6,21]. Having class labels or not necessitates the different ways of evaluating feature quality for supervised and unsupervised feature selection. Lack of class labels can make the feature-quality evaluation less clear-cut because in essence, unsupervised learning (or feature selection) is a less constrained problem than supervised one. More often than not, an unsupervised feature selection algorithm adopts a wrapper model approach as it relies on an unsupervised learning algorithm to help determine if some features are relevant or not. We elaborate the two types of feature selection algorithms below with illustrative examples.

Supervised Feature Selection

To demonstrate the workings of supervised feature selection, we present two representative filter algorithms for subset selection and feature ranking below. The subset selection algorithm is called FCBF (Fast Correlation-Based Filter). It is a search-based algorithm and designed to eliminate both irrelevant and redundant features [31]. It extends the usual feature-class correlation model by adding pair-wise feature correlation in determining a feature's relevance. After a feature is determined to be selected, other features are checked to see their necessity to remain in the selected list of features. The concept is illustrated in Fig. 2. There are 6 features ordered according to their individual class correlations. Starting with feature F_1, one check whether the remaining features are still needed as follows: if their class correlations are smaller than their correlations

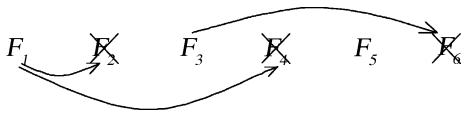

Manipulating Data and Dimension Reduction Methods: Feature Selection, Figure 2
The key concept of FCBF

```
Input: X, Y, δ
Output: S_best – the selected feature list
for i = 1 to N do
    calculate SU_{i,c} for F_i;
    append F_i to S'_list, if SU_{i,c} > δ;
end
order S'_list in descending SU_{i,c} value;
F_j = getFirstElement(S'_list);
while F_j ≠ NULL do
    F_i = getNextElement(S'_list, F_j);
    if F_i ≠ NULL then
        while F_i == NULL do
            if SU_{i,j} ≤ SU_{i,c}, remove F_i from
            S'_list;
            F_i = getNextElement(S'_list, F_i);
        end
    end
    F_j = getNextElement (S'_list, F_j);
end
S_best = S'_list;
```

Manipulating Data and Dimension Reduction Methods: Feature Selection, Algorithm 1
FCBF – Fast Correlation-Based Filter

with F_1, these features can be removed. In the figure, F_2 and F_4's class correlations are smaller than their correlations with F_1 (thus F_2 and F_4 are redundant with respect to F_1), hence, they can be removed. Likewise, F_6 is removed due to its high correlation with F_3. After selection, F_1, F_3 and F_5 are retained. If simply using these features' class correlations to determine relevance, F_1, F_2 and F_3 will be selected assuming we need 3 features. The algorithm FCBF is given in Algorithm 1. It contains two major parts: (1) ranking individual features based on their individual class-correlations such that the feature with the highest class-correlation is ranked first; and (2) going through the ranked list one feature at a time and removing its redundant features, and when every feature in the list is considered, the remaining features are selected features.

The second supervised feature selection algorithm is ReliefF [26] which has significantly extended the work

```
Input: X, Y
Output: w – the vector of feature weight
w ← 0;
for i = 1 to m do
    randomly select an instance R_i;
    find k nearest hits H_j;
    for each class C ≠ class(R_i) do
        from class C find k nearest misses M_j(C);
    end
end
for each feature F_i do
    w(F_i): = w(F_i) − Σ_{j=1}^{k} DIFF(A,R_i,H_j)/(m·k)
    + Σ_{C≠class(R_i)} ( P(C)/(1−P(class(R_i))) )
    × Σ_{j=1}^{k} DIFF(A,R_i,M_j(C))/(m·k) )
end
```

Manipulating Data and Dimension Reduction Methods: Feature Selection, Algorithm 2
ReliefF

in [16]. The basic idea is that a relevant feature should contribute to differentiate two nearby instances of different classes. When this idea is implemented algorithmically, each feature is associated with a weight which is adjusted according to the the local geometry reflected by the nears-hits and near-misses of sampled instances as shown in Algorithm 2. Given an instance, its near-hit is a nearby instance of the same class, and its near-miss is a nearby instance of a different class. The two main parts are (1) for every instance, find their near-hits and near-misses; and (2) for each feature, adjust its weight according to their contributions in terms of near-hits and near-misses. *DIFF*() is a distance function (examples are hamming or Euclidean distances).

When the number of instances is huge, ReliefF can simply employ some sampling techniques [9] to select a small sample from the data, instead of the whole available data. Therefore, ReliefF can be a very efficient algorithm.

Unsupervised Feature Selection

Feature selection is also applied in unsupervised learning [7]. It has gained much attention in the recent years. Most data collected are without class labels since labeling data can incur huge costs. The basic principle of unsupervised learning is to cluster data such that similar objects (instances) are grouped together and dissimilar objects are separated. For data of high-dimensionality, distance cal-

culation can be a big problem due to the curse of dimensionality. One idea is to find features that can promote the data separability. A rudimentary idea of unsupervised feature selection can be implemented as follows: if one can find a sufficient and necessary subset of features for each cluster of data points, the union of these subsets may also be sufficient and necessary. Since finding clusterings is not a trivial matter, various approaches to unsupervised feature selection have been proposed. A recent comprehensive survey of unsupervised feature selection algorithms can be found in [6]. The goal of unsupervised feature selection can be defined as finding the smallest feature subset that best uncovers "interesting natural" clusters from data according the chosen criterion (see [7]). A variant of unsupervised feature selection is subspace clustering. It explores the fact that in a high-dimensional space, clusters can often be found in various subspaces of very low dimensionality. Some subspace clustering algorithms are reviewed in [23].

An unsupervised feature selection algorithm, Laplacian Score, is proposed in [11]. It is briefly described here. Let X be a given date set, and S the matrix recording similarity among instances. $\mathbb{G}(V, E)$ denote the undirected graph constructed from $\mathbb{S}$, where V is the vertex set, and E is the edge set. The ith vertex v_i of $\mathbb{G}$ corresponds to $\mathbf{x}_i \in X$ and there is an edge between each vertex pair (v_i, v_j), where the weight w_{ij} is determined by $\mathbb{S}$, $w_{ij} = S_{ij}$. Given $\mathbb{G}$, its *adjacency matrix* W is defined as $W(i, j) = w_{ij}$. Let $\mathbf{d}$ denote the vector: $\mathbf{d} = \{d_1, d_2, \ldots, d_n\}$, where $d_i = \sum_{k=1}^{n} w_{ik}$, the *degree matrix* D of the graph $\mathbb{G}$ is defined by: $D(i, j) = d_i$ if $i = j$, and 0 otherwise. Given the adjacency matrix W and the degree matrix D of $\mathbb{G}$, the *Laplacian matrix* L and the *normalized Laplacian matrix* $\mathcal{L}$ are defined as:

$$L = D - W; \quad \mathcal{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} . \tag{1}$$

Given a graph $\mathbb{G}$, the Laplacian matrix of $\mathbb{G}$ is a linear operator on feature vectors $\mathbf{f} = (f_1, f_2, \ldots, f_n)^T \in \mathbb{R}^n$:

$$\langle \mathbf{f}, L\mathbf{f} \rangle = \mathbf{f}^T L \mathbf{f} = \frac{1}{2} \sum_{v_i \sim v_j} w_{ij} (x_i - x_j)^2 . \tag{2}$$

The equation quantifies how much $\mathbf{f}$ varies locally or how "smooth" it is over $\mathbb{G}$. More specifically, the smaller the value of $< \mathbf{f}, L\mathbf{f} >$, the smoother the vector $\mathbf{f}$ on $\mathbb{G}$. A smooth vector $\mathbf{f}$ assigns similar values to the instances that are close to each other on $\mathbb{G}$, thus it is consistent with the graph structure. Based on this concept, the Laplacian Score algorithm for unsupervised feature selection is proposed as shown in Algorithm 3.

Input: X, Y, δ
Output: $\mathbf{w}$ – the vector of feature weight
for *each feature* F_i **do**
  $\hat{\mathbf{f}}_i = \mathbf{f}_i - \frac{\mathbf{f}_i^T D \mathbf{1}}{\mathbf{1}^T D \mathbf{1}} \mathbf{1}$; $w(F_i) = \frac{\hat{\mathbf{f}}_i^T L \hat{\mathbf{f}}_i}{\hat{\mathbf{f}}_i^T D \hat{\mathbf{f}}_i}$;
end

Manipulating Data and Dimension Reduction Methods: Feature Selection, Algorithm 3
Laplacian Score: An unsupervised feature selection algorithm

Some Recent Research Development

As research in data processing and analysis advances, feature selection research also evolves in various ways. We present three examples to illustrate some recent studies of feature selection. One example is an attempt to unify supervised and unsupervised feature selection via so-called *spectral feature selection*. Another is of *semi-supervised feature selection* that can take advantage of both labeled and unlabeled data. The third example is an introduction of dealing with *feature interaction* in feature selection. In Sect. "Future Directions", we will discuss some future directions of feature selection research.

Spectral Feature Selection

There seems a chasm between supervised and unsupervised feature selection as one works with class labels and the other does not. A framework for feature selection enables us to (1) jointly study supervised and unsupervised feature selection algorithms, (2) gain a deeper understanding of some existing successful algorithms, and (3) derive novel algorithms with better performance. A unified framework is proposed in [35] based on spectral graph theory [4], and it is shown that existing powerful algorithms such as ReliefF (supervised) and Laplacian Score (unsupervised) are special cases of the proposed framework.

Both supervised and unsupervised feature selection can be viewed as an effort to select features that are consistent with the target concept. In supervised learning, the target concept is related to class affiliation, while in unsupervised learning the target concept is usually related to the innate structures of the data. Essentially, in both cases, the target concept is related to dividing instances into well separable subsets according to different definitions of the separability. Pairwise instance similarity is widely used in both supervised and unsupervised learning to describe the relationships among instances. Given a set of pairwise instance similarities $\mathbb{S}$, the separability of the instances can

be studied by analyzing the spectrum of the graph induced from $\mathbb{S}$. A unified framework for feature selection using the spectrum of the graph induced from $\mathbb{S}$ is proposed in [35]. By designing different $\mathbb{S}$'s, the unified framework can produce families of algorithms for both supervised and unsupervised feature selection. For example, without using the class information, a popular similarity measure is the *RBF* kernel function:

$$S_{ij} = e^{-\frac{\|x_1 - x_2\|^2}{(2\delta^2)}} . \tag{3}$$

Using the class labels, the similarity can be defined by:

$$S_{ij} = \begin{cases} \frac{1}{n_l}, & y_i = y_j = l \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where n_l denotes the number of instances in class l.

One can apply L on a feature vector to measure its consistency with the graph structure. In [35], three functions are proposed to evaluate features so that feature selection can be performed in three steps: (1) building similarity set $\mathbb{S}$ and constructing its graph representation (Line 1–3); (2) evaluating features using the spectrum of the graph (Line 4–6); and (3) ranking features in descending order in terms of feature relevance, where features selection is accomplished by choosing the desired number of features from the returned feature list (Line 7–8). The framework is named as *SPEC*, stemming from the *SPEC*trum decomposition of $\mathcal{L}$.

The connections of the framework to unsupervised and supervised feature selection algorithms (Laplacian Score and ReliefF) are shown via both analysis and experiments in [35]. The framework can also be used to systematically derive novel spectral algorithm by using different $\mathbb{S}$, $\gamma(\cdot)$ and ranking functions $\widehat{\varphi}(\cdot)$.

Input: X, $\gamma(\cdot)$, k, $\widehat{\varphi} \in \{\widehat{\varphi}_1, \widehat{\varphi}_2, \widehat{\varphi}_3\}$ – ranking functions
Output: SF_{SPEC} – the ranked feature list
1 construct $\mathbb{S}$, the similarity set from X (and Y);
2 construct graph G from $\mathbb{S}$;
3 build W, D and L from G;
4 **for** *each feature vector* $\boldsymbol{f}_i$ **do**
5 $\quad \widehat{\boldsymbol{f}}_i \leftarrow \frac{D^{\frac{1}{2}} \boldsymbol{f}_i}{\|D^{\frac{1}{2}} \boldsymbol{f}_i\|}$; $SF_{SPEC}(i) \leftarrow \widehat{\varphi}(F_i)$;
6 **end**
7 ranking SF_{SPEC} in ascending order for $\widehat{\varphi}_1$ and $\widehat{\varphi}_2$, or descending order for $\widehat{\varphi}_3$;
8 return SF_{SPEC};

Manipulating Data and Dimension Reduction Methods: Feature Selection, Algorithm 4
SPEC

Semi-Supervised Feature Selection

While we are inundated with data, but labeled data is costly to obtain. Nowadays it is also common to have a data set with huge dimensionality but small labeled-sample size. The data sets of this kind present a serious challenge, the so-called *small labeled-sample problem*, to supervised feature selection. In other words, when the labeled sample size is too small to carry sufficient information about the target concept, supervised feature selection algorithms might fail by either unintentionally removing many relevant features or selecting irrelevant features. Unsupervised feature selection algorithms can be an alternative in this case, as they are able to use large amounts of unlabeled data. However, as these algorithms do not use label information, important information about labels is left out, and this can downgrade the performance of unsupervised feature selection algorithms. Under the assumption that labeled and unlabeled data are sampled from the same population of the target concept, using both labeled and unlabeled data is expected to better estimate feature relevance. The task of learning from mixed labeled and unlabeled data is of semi-supervised learning [3]. A *semi-supervised feature selection* algorithm is proposed in [34] to rank features through a regularization framework, in which a feature's relevance is evaluated by its fitness with both labeled and unlabeled data.

Supervised and unsupervised feature selection methods require to measure feature relevance, but in different ways. Therefore the key for designing an effective semi-supervised feature selection algorithm is to ensure that the relevance of a feature can be evaluated by both labeled and unlabeled data in a natural way. The basic idea is illustrated in Fig. 3. A feature vector $\mathbf{f}_i$ is first transformed into a cluster indicator so that each element $f_{i_j}, (j = 1, 2, \ldots, n)$ of $\mathbf{f}_i$ indicates the affiliation of the corresponding instance $\mathbf{x}_j$. The fitness of the cluster indicator can be evaluated by two factors: (1) separability of unlabeled data – whether the cluster structures formed are well separable; and (2) separability of labeled data – whether the cluster structures formed is consistent with the given label information. The ideal case is that all labeled data in each cluster are of the same class.

It is clear that the spectral feature selection framework (*SPEC*) naturally captures both types of separability. The total separability can be measured through a regularization

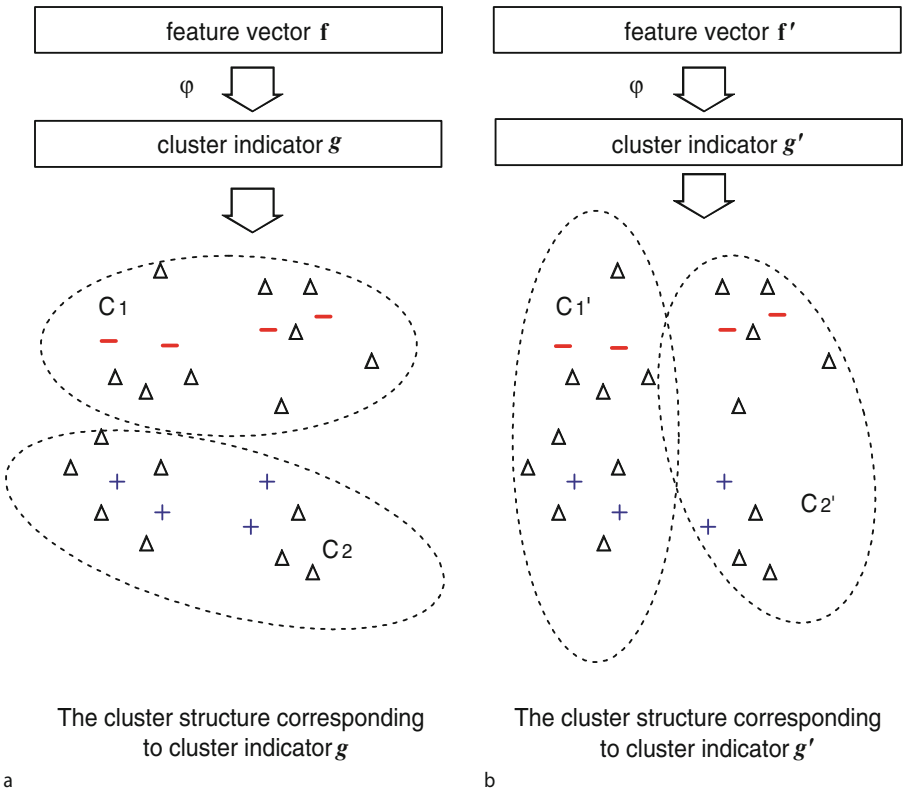

Manipulating Data and Dimension Reduction Methods: Feature Selection, Figure 3
The basic idea for comparing the fitness of cluster indicators according to both labeled and unlabeled data for semi-supervised feature selection. "–" corresponds to instances of negative class, "+" to those of positive class, and "Δ" to unlabeled instances

in the following form:

$$\lambda\varphi_{\mathbb{S}_u}(F_i) + (1-\lambda)\varphi_{\mathbb{S}_s}(F_i) \tag{5}$$

where λ is the regularization parameter about which part (the supervised or the unsupervised part) is more important. $\varphi_{\mathbb{S}_s}(\cdot)$ and $\varphi_{\mathbb{S}_u}(\cdot)$ are the evaluation functions for supervised part and unsupervised part respectively. The evaluation function for the supervised part is based on normalized mutual information which is applied on the cluster indicator derived from a feature and the class label. Given the regularization framework, a semi-supervised feature selection algorithm, *sSelect* is proposed in [34] in Algorithm 5. More detailed analysis and experimental results of *sSelect* can be found in [32].

Input: X, Y_L, λ, k
Output: $SF_{sSelect}$, the ranked feature list
construct k-neighborhood graph G from X;
build W, $\mathbf{d}$ and L from G;
for *each feature vector* $\mathbf{f}_i$ **do**
  construct $\mathbf{g}_i$ from $\mathbf{f}_i$ using φ;
  calculate s_i, the score of F_i using Eq. (5);
end
$SF_{sSelect} \leftarrow$ ranking F_i in descending order;
return $SF_{sSelect}$;

Manipulating Data and Dimension Reduction Methods: Feature Selection, Algorithm 5
***sSelect* for Semi-supervised Feature Selection**

Interacting Features

Feature interaction presents another challenge to feature selection. A feature by itself may have little correlation with the target concept, but when it is combined with some other features, they can be strongly correlated with the target concept. Unintentional removal of these features can result in poor classification performance. Existing efficient feature selection algorithms usually assume feature independence [5]. Because of the irreducible nature of feature interactions, these algorithms either cannot select interacting features or can only handle low-order interactions

(2- or 3-way). The *XOR* problem is a good example of two interacting features (F_1 and F_2). Considered individually, the correlation between F_1 and the class C (similarly for F_2 and C) is zero, measured by mutual information. Hence, F_1 or F_2 is irrelevant when each is individually evaluated. However, if we combine F_1 with F_2, they are strongly relevant in defining the target concept of *XOR*. An intrinsic character of feature interaction is its irreducibility [13], i. e., a feature could lose its relevance due to the absence of its interacting feature(s).

In [12], the authors suggest to use interaction gain as a practical heuristic for detecting attribute interaction. Using interaction gain, their algorithms can detect if datasets have 2-way (one feature and the class) and 3-way (two features and the class) interactions. They further provide in [13] a justification of the interaction information, and replace the earlier notion of 'high' and 'low' with statistical significance and illustrate the significant interactions in the form of interaction graph. Handling feature interaction can be computationally intractable. An efficient feature selection algorithm, *INTERACT*, is proposed in [33] to handle feature interaction.

The three key components of *INTERACT* are: (1) consistency based feature relevance measure: *c-contribution*. *C-contribution* of a feature F_i is a function of $\mathbf{F} - \{F_i\}$, where $\mathbf{F}$ is the set of features for D. *C-contribution* of a feature is an indicator about how significantly the elimination of that feature will affect data consistency. (2) Simplified backward elimination. To remove k out of n features, we start from the end of the ranked list of features to check if a feature's *c-contribution* is below δ: if it is, the feature is removed, otherwise, it is retained. On one hand, the simplified backward elimination has a time complexity of $O(k)$, instead of $O(kn)$ as in the normal backward elimination [18]. On the other hand, cooperating with *c-contribution*, the simplified backward elimination works as well as normal backward elimination. It is clear that backward elimination plusing *c-contribution* allows a feature to be evaluated with all features it potentially interacts with. (3) A hash table data structure for accelerating the evaluation of *c-contribution*. *INTERACT* also using *symmetrical uncertainty* (SU) as a preprocessing step to rank feature in an descending order such that the (heuristically) most relevant feature is positioned at the beginning of the list. The preprocessing step helps overcome the so called feature order problem of *c-contribution*.

The pseudocode of *INTERACT* is listed in Algorithm 6. Given a full set with N features and a class attribute Y, it finds a feature subset S_{best} for the class concept. The algorithm consists of two major parts. In the first part (lines 1–6), the features are ranked in descending order

```
Input:  F_1, F_2, ..., F_N, the full feature set;
        Y, the class label;
        δ, a predefined threshold;
Output: S_best, the best subset;
// Ranking
1  S_list = NULL;
2  for i=1 to N do
3      calculate SU_{F_i,y} for F_i;
4      append F_i to S_list;
5  end
6  order S_list in descending values of SU_{i,y};
// Feature Eliminating
7  counter = N;
8  repeat
9      F = S_list[counter];
10     p = c-contribution(F, S_list);
11     if p ≤ δ then
12         remove F from S_list;
13     end
14     counter = counter − 1;
15 until counter = 0 ;
16 S_best = S_list;
17 return S_best;
```

Manipulating Data and Dimension Reduction Methods: Feature Selection, Algorithm 6
Algorithm INTERACT

based on their *SU* values. In the second part (lines 7–16), features are evaluated one by one starting from the end of the ranked feature list. $S_{\text{list}}[i]$ returns the feature in the position i of the list, S_{list}. If *c-contribution* of a feature is less than δ, the feature is removed, otherwise it is selected. And the counter is decreased and pointed to next unchecked feature preceding the F in the ranked feature list (line 14). The algorithm continues until all features in the list are checked. δ is a predefined threshold ($0 < \delta < 1$). Features with their *c-contribution* $< \delta$ are considered immaterial and removed. A large δ is associated with a high probability of removing relevant features. In practice a proper δ can be found by cross-validation.

Future Directions

Feature selection finds its application in many fields such as image processing, computer vision, pattern recognition, bioinformatics, text categorization, information retrieval and extraction, data fusion, Web mining, to name some examples. The research of feature selection evolves and de-

velops rapidly to answer the pressing needs arising from the real world [17]. One challenge is to deal with large numbers of instances. In addition to sampling methods, another approach is to rely on search to find representative instances to represent the data used for feature selection. One such approach is presented in [20] in which data is first organized into bins according to its intrinsic properties, and representative instances are sampled from respective bins for feature selection. Another challenge is how to allow a domain expert to interject her/his expert knowledge in solving real world problems. That is, the computing system serves mainly as an intelligent assistant that can request advice from a human expert. An example of involving domain experts in the decision process is reported in [8]. The objectives of such assistant systems are to mitigate the human expert's effort in data intensive processing tasks so that the expert can have ample time and focus on solving more complex but fewer problems. Active learning plays an instrumental role. Feature selection can also take advantage of such a mechanism in efficiently integrating human expertise.

The frontier of feature selection is expanding incessantly in various directions. On one hand, feature selection develops from supervised to unsupervised and to semi-supervised methods; on the other hand, new feature selection methods emerge such as causal feature selection, relational feature selection, and sequential feature selection. Some latest research and new directions have been showcased in a recent book on computational methods of feature selection [19]. Included exemplar methods are randomized feature selection, active learning of feature relevance, ensembles with variable independent probing, weighting methods, local methods, and non-myopic feature quality evaluation. Representative outstanding applications of feature selection are text and bioinformatics covering Bayesian feature scoring, aggressive feature ranking, feature generation for biological sequence classification, ensemble methods for biomarker identification, model building and feature selection with genomic data. A large body of feature selection work can also be accessed in various journals and conferences, as well as specialized workshops.

Bibliography

Primary Literature

1. Ambroise C, McLachlan GJ (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci USA 99(10):6562–6566
2. Bellman R (1961) Adaptive Control Processes: A Guided Tour. Princeton University Press, Princeton
3. Chapelle O, Scholkopf B, Zien A (2006) Semi-Supervised Learning. MIT Press, Cambridge
4. Chung F (1997) Spectral Graph Theory. American Mathematical Society, Providence
5. Dash M, Liu H (1997) Feature selection methods for classifications. Intell Data Anal: Int J 1(3):131–156
6. Dy J (2007) Unsupervised feature selection. In: Liu H, Motoda H (eds) Computational Methods of Feature Selection. Chapman Hall/CRC Press, Boca Raton
7. Dy JG, Brodley CE (2004) Feature selection for unsupervised learning. J Mach Learn Res 5:845–889
8. Foschi PG, Liu H (2004) Active learning for detecting a spectrally variable subject in color infrared imaginery. Pattern Recognit Lett 25(13):1509–1517
9. Gu B, Hu F, Liu H (2001) Sampling: Knowing Whole from Its Part. Kluwer, Boston, pp 21–38
10. Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning. Springer, New York
11. He X, Cai D, Niyogi P (2005) Laplacian score for feature selection. In: Weiss Y, Schölkopf B, Platt J (eds) Advances in Neural Information Processing Systems 18. MIT Press, Cambridge
12. Jakulin A, Bratko I (2003) Analyzing attribute dependencies. In: Lavrač N, Gamberger D, Blockeel H, Todorovski L (eds) Proc. of Principles of Knowledge Discovery in Data (PKDD), vol 2838 of LNAI. Springer, New York, September 2003, pp 229–240
13. Jakulin A, Bratko I (2004) Testing the significance of attribute interactions. In: ICML '04: Twenty-first international conference on Machine learning.
14. John GH, Kohavi R, Pfleger K (1994) Irrelevant feature and the subset selection problem. In: Cohen WW, Hirsh H (eds) Machine Learning: Proceedings of the Eleventh International Conference. Rutgers University, New Brunswick, pp 121–129
15. Kearns MJ, Vazirani UV (1994) An Introduction to Computational Learning Theory. MIT Press, Cambridge
16. Kira K, Rendell LA (1992) The feature selection problem: Traditional methods and a new algorithm. In: Proceedings of the Tenth National Conference on Artificial Intelligence. AAAI Press/MIT Press, Menlo Park, pp 129–134
17. Liu H (2005) Evolving feature selection. IEEE Intell Syst 20(6):64–76
18. Liu H, Motoda H (1998) Feature Selection for Knowledge Discovery & Data Mining. Kluwer, Boston
19. Liu H, Motoda H (eds) (2007) Computational Methods of Feature Selection. Chapman Hall/CRC Press, Boca Raton
20. Liu H, Motoda H, Yu L (2004) A selective sampling approach to active feature selection. Artif Intell J 159(1–2):49–74
21. Liu H, Yu L (2005) Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowl Data Engin 17(3):1–12
22. Mitchell TM (1997) Machine Learning. McGraw-Hill, Columbus
23. Parson L, Haque E, Liu H (2004) Subspace clutering for high dimensional data – a review. SIGKDD Explorations
24. Quinlan JR (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann, San Fransisco
25. Refaeilzadeh P (2007) An analysis of feature selection evalaution methods. Master's thesis, Arizona State University, Computer Science and Engineering
26. Robnik-Sikonja M, Kononenko I (2003) Theoretical and empirical analysis of Relief and ReliefF. Mach Learn 53:23–69
27. Russell S, Norvig P (2003) Artificial Intelligence: A Modern Approach. Prentice Hall, Upper Saddle River
28. Singhi S, Liu H (2006) Feature subset selection bias for clas-

sification learning. In: International Conference on Machine Learning. Pittsburgh, USA, 2006
29. Smyth P, Pregibon D, Faloutsos C (2002) Data-driven evolution of data mining algorithms. Commun Assoc Comput Mach 45(8):33–37
30. Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. BMC Bioinf 7:91
31. Yu L, Liu H (2004) Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res 5(Oct):1205–1224
32. Zhao Z, Liu H (2006) Semi-supervised feature selection via spectral analysis. Technical Report TR-06–022, Computer Science and Engineering Department
33. Zhao Z, Liu H (2007) Searching for interacting features. In: Proceedings of IJCAI – International Joint Conference on AI. Hyderabad, January 2007
34. Zhao Z, Liu H (2007) Semi-supervised feature selection via spectral analysis. In: Proceedings of SIAM International Conference on Data Mining (SDM-07). Minneapolis
35. Zhao Z, Liu H (2007) Spectral feature selection for supervised and unsupervised learning. In: Proceedings of International Conference on Machine Learning. Corvallis

Books and Reviews

Dash M, Liu H (1997) Feature selection methods for classifications. Intell Data Anal: Int J 1(3):131–156
Dy JG, Brodley CE (2004) Feature selection for unsupervised learning. J Mach Learn Res 5:845–889
Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
Liu H, Motoda H (1998) Feature Selection for Knowledge Discovery & Data Mining. Kluwer, Boston
Liu H, Motoda H (eds) (1998) Feature Extraction, Construction and Selection: A Data Mining Perspective. Kluwer, Boston
Liu H, Motoda H (eds) (2007) Computational Methods of Feature Selection. Chapman Hall/CRC Press, Boca Raton
Liu H, Yu L (2005) Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. Knowl Data Engin 17(3):1–12
Parson L, Haque E, Liu H (2004) Subspace clutering for high dimensional data – a review. SIGKDD Explorations

Market Games and Clubs

Myrna Wooders
Department of Economics, Vanderbilt University, Nashville, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Transferable Utility Games; Some Standard Definitions
A Market
Market-Game Equivalence
Equivalence of Markets and Games with Many Players
Cores and Approximate Cores
Nonemptiness and Convergence of Approximate Cores of Large Games
Shapley Values of Games with Many Players
Economies with Clubs
With a Continuum of Players
Other Related Concepts and Results
Some Remarks on Markets and More General Classes of Economies
Conclusions and Future Directions
Bibliography

Glossary

Game A (cooperative) game (in characteristic form) is defined simply as a finite set of players and a function or correspondence ascribing a worth (a non-negative real number, interpreted as an idealized money) to each nonempty subset of players, called a group or coalition.

Payoff vector A payoff vector is a vector listing a payoff (an amount of utility or money) for each player in the game.

Core The core of a game is the set (possibly empty) of feasible outcomes – divisions of the worths arising from coalition formation among the players of the game – that cannot be improved upon by any coalition of players. core

Totally balanced game A game is totally balanced if the game and every subgame of the game (a game with player set taken as some subset of players of the initially given game) has a nonempty core.

Market A market is defined as a private goods economy in which all participants have utility functions that are linear in (at least) one commodity (money).

Shapley value The Shapley value of a game is feasible outcome of a game in which all players are assigned their expected marginal contribution to a coalition when all orders of coalition formation are equally likely.

Pregame A pair, consisting of a set of player types (attributes or characteristics) and a function mapping finite lists of characteristics (repetitions allowed) into the real numbers. In interpretation, the pregame function ascribes a worth to every possible finite group of players, where the worth of a group depends on the numbers of players with each characteristic in the group. A pregame is used to generate games with arbitrary numbers of players.

Small group effectiveness A pregame satisfies small group effectiveness if almost all gains to collective activities can be realized by cooperation only within arbitrarily small groups (coalitions) of players.

Per capita boundedness A pregame satisfies per capita boundedness if the supremum of the average worth of any possible group of players (the per capita payoff) is finite.

Asymptotic negligibility A pregame satisfies asymptotic negligibility if vanishingly small groups can have only negligible effects on per capita payoffs.

Market games A market game is a game derived from a market. Given a market and a group of agents we can determine the total utility (measured in money) that the group can achieve using only the endowments belonging to the group members, thus determining a game.

Club A club is a group of agents or players that forms for the purpose of carrying out come activity, such as providing a local public good.

An economy We use the term 'economy' to describe any economic setting, including economies with clubs, where the worth of club members may depend on the characteristics of members of the club, economies with pure public goods, local public goods (public goods subject to crowding and/or congestion), economies with production where what can be produced and the costs of production may depend on the characteristics of the individuals involved in production, and so on. A *large economy* has many participants.

Price taking equilibrium A price taking equilibrium for a market is a set of prices, one for each commodity, and an allocation of commodities to agents so that each agent can afford his part of the allocation, given the value of his endowment.

Definition of the Subject

The equivalence of markets and games concerns the relationship between two sorts of structures that appear fundamentally different – markets and games. Shapley and Shubik [60] demonstrates that: (1) games derived from markets with concave utility functions generate totally balanced games where the players in the game are the participants in the economy and (2) every totally balanced game generates a market with concave utility functions. A particular form of such a market is one where the commodities are the participants themselves, a labor market for example.

But markets are very special structures, more so when it is required that utility functions be concave. Participants may also get utility from belonging to groups, such as marriages, or clubs, or productive coalitions. It may be that participants in an economy even derive utility (or disutility) from engaging in processes that lead to the eventual exchange of commodities. The question is when are such economic structures equivalent to markets with concave utility functions.

This paper summarizes research showing that a broad class of large economies generate balanced market games. The economies include, for example, economies with clubs where individuals may have memberships in multiple clubs, with indivisible commodities, with nonconvexities and with non-monotonicities. The main assumption are: (1) that an option open to any group of players is to break into smaller groups and realize the sum of the worths of these groups, that is, essential superadditivity is satisfied and: (2) relatively small groups of participants can realize almost all gains to coalition formation.

The equivalence of games with many players and markets with many participants indicates that relationships obtained for markets with concave utility functions and many participants will also hold for diverse social and economic situations with many players. These relationships include: (a) equivalence of the core and the set of competitive outcomes; (b) the Shapley value is contained in the core or approximate cores; (c) the equal treatment property holds – that is, both market equilibrium and the core treat similar players similarly. These results can be applied to diverse economic models to obtain the equivalence of cooperative outcomes and competitive, price taking outcomes in economies with many participants and indicate that such results hold in yet more generality.

Introduction

One of the subjects that has long intrigued economists and game theorists is the relationship between games, both cooperative and noncooperative, and economies. Seminal works making such relationships include Shubik [67], Debreu and Scarf [22], Aumann [4], Shapley and Shubik [60,62] and Aumann and Shapley [7], all connecting outcomes of price-taking behavior in large economies with cores of games. See also Shapley and Shubik [63] and an ongoing stream of papers connecting strategic behavior to market behavior. Our primary concern here, however, is not with the equivalence of outcomes of solution concepts for economies, as is Debreu and Scarf [22] or Aumann [6] for example, but rather with equivalences of the *structures* of markets and games. Solution concepts play some role, however, in establishing these equivalences and in understanding the meaning of the equivalence of markets and games.

In this entry, following Shapley and Shubik [60], we focus on markets in which utility functions of participants are quasi-linear, that is, the utility function u of a partici-

pant can be written as $u(x, \xi) = \widehat{u}(x) + \xi$ where $x \in \mathbb{R}^L_+$ is a commodity bundle, $\xi \in \mathbb{R}$ is interpreted as money and $\widehat{u}$ is a continuous function. Each participant in an economy has an endowment of commodities and, without any substantive loss of generality, it is assumed that no money is initially endowed. The price of money is assumed equal to one. A price taking equilibrium for a market then consists of a price vector $p \in \mathbb{R}^L$ for the commodities.and an assignment of commodities to participants such that: the total amounts of commodities assigned to participants equals the total amount of commodities with which participants are endowed and; given prices, each participant can afford his assignment of commodities and no participant, subject to his budget constraint, can afford a preferred commodity bundle.

We also treat games with side payments, alternatively called games with transferable utility or, in brief, TU games. Such a game consists of a finite set N of players and a worth function that assigns to each group of players $S \subset N$ a real number $v(S) \in \mathbb{R}_+$, called the worth of the group. In interpretation, $v(S)$ is the total payoff that a group of players can realize by cooperation. A central game-theoretic concept for the study of games is the core. The core consists of those divisions of the maximal total worth achievable by cooperation among the players in N so that each group of players is assigned at least its worth. A game is balanced if it has a nonempty core and totally balanced if all subgames of the game have nonempty cores. A subgame of a game is simply a group of players $S \subset N$ and the worth function restricted to that group and the smaller groups that it contains.

Given a market any feasible assignment of commodities to the economic participants generates a total worth of each group of participants. The worth of a group of participants (viewed as players of a game) is the maximal total utility achievable by the members of the group by allocating the commodities they own among themselves. In this way a market generates a game – a set of players (the participants in the economy) and a worth for each group of players.

Shapley and Shubik [60] demonstrate that any market where all participants have concave, monotonic increasing utility functions generates a totally balanced game and that any totally balanced game generates a market, thus establishing an equivalence between a class of markets and totally balanced cooperative games. A particular sort of market is canonical; one where each participant in the market is endowed with one unit of a commodity, his "type". Intuitively, one might think of the market as one where each participant owns one unit of himself or of his labor.

In the last twenty years or so there has been substantial interest in broader classes of economies, including those with indivisibilities, nonmonotonicities, local public goods or clubs, where the worth of a group depends not only on the private goods endowed to members of the group but also on the characteristics of the group members. For example, the success of the marriage of a man and a woman depends on their characteristics and on whether their characteristics are complementary. Similarly, the output of a machine and a worker using the machine depends on the quality and capabilities of the machine and how well the abilities of the worker fit with the characteristics of the machine – a concert pianist fits well with an high quality piano but perhaps not so well with a sewing machine. Or how well a research team functions depends not only on the members of the team but also on how well they interact. For simplicity, we shall refer to these economies as club economies. Such economies can be modeled as cooperative games.

In this entry we discuss and summarize literature showing that economies with many participants are approximated by markets where all participants have the same concave utility function and for which the core of the game is equivalent to the set of price-taking economic equilibrium payoffs. The research presented is primarily from Shubik and Wooders [65], Wooders [92] and earlier papers due to this author. For the most recent results in this line of research we refer the reader to Wooders [93,94,95]. We also discuss other related works throughout the course of the entry. The models and results are set in a broader context in the conclusions.

The importance of the equivalence of markets and games with many players relates to the hypothesis of perfect competition, that large numbers of participants leads to price-taking behavior, or behavior "as if" participants took prices as given. Von Neumann and Morgenstern perceived that even though individuals are unable to influence market prices and cannot benefit from strategic behavior in large markets, large "coalitions" might form. Von Neumann and Morgenstern write:

> It is neither certain nor probable that a mere increase in the number of participants might lead *in fine* to the conditions of free competition. The classical definitions of free competition all involve further postulates besides this number. E.g., it is clear that if certain great groups of individuals will – for any reason whatsoever– act together, then the great number of participants may not become effective; the decisive exchanges may take place directly between large "coalitions", few in number and not be-

> tween individuals, many in number acting independently. … Any satisfactory theory … will have to explain when such big coalitions will or will not be formed –i. e., when the large numbers of participants will become effective and lead to more or less free competition.

The assumption that small groups of individuals cannot affect market aggregates, virtually taken for granted by von Neumann and Morgenstern, lies behind the answer to the question they pose. The results presented in this entry suggest that the great number of participants will become effective and lead to more or less free competition when *small* groups of participants cannot significantly affect market outcomes. Since all or almost all gains to collective activities can be captured by relatively small groups, large groups gain no market power from size; in other words, large groups are inessential. That large groups are inessential is equivalent to small group effectiveness [89]. A remarkable feature of the results discussed in this essay is they are independent of any particular economic structure.

Transferable Utility Games; Some Standard Definitions

Let (N, v) be a pair consisting of a finite set N, called a *player set*, and a function v, called a *worth function*, from subsets of N to the real numbers $\mathbb{R}$ with $v(\phi) = 0$. The pair (N, v) is a *TU game* (also called a game with side payments). Nonempty subsets S of N are called *groups* (of players) and the number of members of the group S is given by $|S|$. Following is a simple example.

Example 1 A glove game: Suppose that we can partition a player set N into two groups, say N_1 and N_2. In interpretation, a member of N_1 is endowed with a right-hand (RH) glove and a member of N_2 is endowed with a left-hand (LH) glove. The worth of a pair of gloves is \$1, and thus the worth of a group of players consisting of player $i \in N_1$ and player $j \in N_2$ is \$1. The worth of a single glove and hence of a one-player group is \$0. The worth of a group $S \subset N$ is given by $v(S) = \min\{|S \cap N_1|, |S \cap N_2|\}$. The pair (N, v) is a game.

A *payoff vector* for a game (N, v) is a vector $\overline{u} \in \mathbb{R}^N$. We regard vectors in finite dimensional Euclidean space $\mathbb{R}^T$ as functions from T to $\mathbb{R}$, and write $\overline{u}_i$ for the ith component of $\overline{u}$, etc. If $S \subset T$ and $\overline{u} \in \mathbb{R}^T$, we shall write $\overline{u}_S := (\overline{u}_i : i \in S)$ for the restriction of $\overline{u}$ to S. We write 1_S for the element of $\mathbb{R}^S$ all of whose coordinates are 1 (or simply 1 if no confusion can arise.) A payoff vector $\overline{u}$ is *feasible for a group* $S \subset N$ if

$$\overline{u}(S) \stackrel{\text{def}}{=} \sum_{i \in S} \overline{u}^i \leq \sum_{k=1}^{K} v(S^k) \tag{1}$$

for some partition $\{S^1, \ldots, S^K\}$ of S.

Given $\varepsilon \geq 0$, a payoff vector $\overline{u} \in \mathbb{R}^N$ is in the *weak ε-core* of the game (N, v) if it is feasible and if there is a group of players $N^0 \subset N$ such that

$$\frac{|N \backslash N^0|}{|N|} \leq \varepsilon \tag{2}$$

and, for all groups $S \subset N^0$,

$$\overline{u}(S) \geq v(S) - \varepsilon |S| \tag{3}$$

where $|S|$ is the cardinality of the set S. (It would be possible to use two different values for epsilon in expressions (2) and (3). For simplicity, we have chosen to take the same value for epsilon in both expressions.) A payoff vector $\overline{u}$ is in the *uniform ε-core* (or simply in the *ε-core*) if if is feasible and if (3) holds for *all* groups $S \subset N$. When $\varepsilon = 0$, then both notions of ε-cores will be called simply the *core*.

Example 1 (continued) The glove game (N, v) described in Example 1 has the happy feature that the core is always nonempty. For the game to be of interest, we will suppose that there is least one player of each type (that is, there is at least one player with a RH glove and one player with a LH glove). If $|N_1| = |N_2|$ any payoff vector assigning the same share of a dollar to each player with a LH glove and the remaining share of a dollar to each player with a RH glove is in the core. If there are more players of one type, say $|N_1| > |N_2|$ for specificity, then any payoff vector in the core assigns \$1 to each player of the scarce type; that is, players with a RH glove each receive 0 while players with a LH glove each receive \$1.

Not all games have nonempty cores, as the following example illustrates.

Example 2 (A simple majority game with an empty core) Let $N = \{1, 2, 3\}$ and define the function v as follows:

$$v(S) = \begin{cases} 0 \text{ if } |S| = 1\,, \\ 1 \text{ otherwise}\,. \end{cases}$$

It is easy to see that the core of the game is empty. For if a payoff vector $\overline{u}$ were in the core, then it must hold that for any $i \in N$, $\overline{u}_i \geq 0$ and for any $i, j \in N$, $\overline{u}_i + \overline{u}_j \geq 1$. Moreover, feasibility dictates that $\overline{u}_1 + \overline{u}_2 + \overline{u}_3 \leq 1$. This is impossible; thus, the core is empty.

Before leaving this example, let us ask whether it would be possible to subsidize the players by increasing the payoff to the total player set N and, by doing so, ensure that the core of the game with a subsidy is nonempty. We leave it to the reader to verify that if $v(N)$ were increased to \$3/2 (or more), the new game would have a nonempty core.

Let (N, v) be a game and let $i, j \in N$. Then players i and j are *substitutes* if, for all groups $S \subset N$ with $i, j \notin S$ it holds that

$$v(S \cup \{i\}) = v(S \cup \{j\}) .$$

Let (N, v) be a game and let $\overline{u} \in \mathbb{R}^N$ be a payoff vector for the game. If for all players i and j who are substitutes it holds that $\overline{u}_i = \overline{u}_j$ then $\overline{u}$ has the *equal treatment property*. Note that if there is a partition of N into T subsets, say $N_1, \ldots, N_T$, where all players in each subset N_t are substitutes for each other, then we can *represent* $\overline{u}$ by a vector $\overline{\overline{u}} \in \mathbb{R}^T$ where, for each t, it holds that $\overline{\overline{u}}_t = \overline{u}_i$ for all $i \in N_t$.

Essential Superadditivity

We wish to treat games where the worth of a group of players is independent of the total player set in which it is embedded and an option open to the members of a group is to partition themselves into smaller groups; that is, we treat games that are *essentially superadditive*. This is built into our the definition of feasibility above, (1). An alternative approach, which would still allow us to treat situations where it is optimal for players to form groups smaller than the total player set, would be to assume that v is the "superadditive cover" of some other worth function v'. Given a not-necessarily-superadditive function v', for each group S define $v(S)$ by:

$$v(S) = \max \sum v'(S^k) \tag{4}$$

where the maximum is taken over all partitions $\{S^k\}$ of S; the function v is the *superadditive cover* of v'. Then the notion of feasibility requiring that a payoff vector $\overline{u}$ is feasible only if

$$\overline{u}(N) \leq v(N) , \tag{5}$$

gives an equivalent set of feasible payoff vectors to those of the game (N, v') with the definition of feasibility given by (1).

The following Proposition may be well known and is easily proven. This result was already well understood in Gillies [27] and applications have appeared in a number of papers in the theoretical literature of game theory; see, for example (for $\varepsilon = 0$) Aumann and Dreze [6] and Kaneko and Wooders [33]. It is also well known in club theory and the theory of economies with many players and local public goods.

Proposition 1 *Given $\varepsilon \geq 0$, let (N, v') be a game. A payoff vector $\overline{u} \in R^N$ is in the weak, respectively uniform, ε-core of (N, v') if and only if it is in the weak, respectively uniform, ε-core of the superadditive cover game, say (N, v), where v is defined by (4).*

A Market

In this section we introduce the definition, from Shapley and Shubik [60], of a market. Unlike Shapley and Shubik, however, we do not assume concavity of utility functions. A *market* is taken to be an economy where all participants have continuous utility functions over a finite set of commodities that are all linear in one commodity, thought of as an "idealized" money. Money can be consumed in any amount, possibly negative. For later convenience we will consider an economy where there is a finite set of types of participants in the economy and all participants of the same type have the same endowments and preferences.

Consider an economy with $T + 1$ types of commodities. Denote the set of participants by

$$N = \{(t, q) : t = 1, \ldots, T, \quad \text{and } q = 1, \ldots, n_t\} .$$

Assume that all participants of the same type, (t, q), $q = 1, \ldots, n_t$ have the same utility functions given by

$$\widehat{u}_t(y, \xi) = u_t(y) + \xi$$

where $y \in \mathbb{R}^T_+$ and $\xi \in \mathbb{R}$. Let $a^{tq} \in \mathbb{R}^T_+$ be the *endowment* of the (t, q)th player of the first T commodities. The *total endowment* is given by $\sum_{(t,q)\in N} a^{tq}$. For simplicity and without loss of generality, we can assume that no participant is endowed with any nonzero amount of the $(T + 1)^{th}$ good, the "money" or medium of exchange. One might think of utilities as being measured in money. It is because of the transferability of money that utilities are called "transferable".

Remark 1 Instead of assuming that money can be consumed in negative amounts one might assume that endowments of money are sufficiently large so that no equilibrium allocates any participant a negative amount of money. For further discussion of transferable utility see, for example, Bergstrom and Varian [9] or Kaneko and Wooders [34] .

Given a group $S \subset N$, a *S-allocation of commodities* is a set

$$\left\{ (y^{tq}, \xi^{tq}) \in \mathbb{R}^T_+ \times \mathbb{R} : \sum_{(t,q)\in S} y^{tq} \le \sum_{(t,q)\in S} a^{tq} \text{ and } \sum_{(t,q)\in S} \xi^{tq} \le 0 \right\} ;$$

that is, a S-allocation is a redistribution of the commodities owned by the members of S among themselves and monetary transfers adding up to no more than zero. When $S = N$, a S-allocation is called simply an *allocation*.

With the price of the $(T+1)^{th}$ commodity ξ set equal to 1, a *competitive outcome* is a price vector p in $\mathbb{R}^T$, listing prices for the first T commodities, and an allocation $\{(y^{tq}, \xi^{tq}) \in \mathbb{R}^T \times \mathbb{R} : (t,q) \in N\}$ for which

$$\begin{aligned}
&\text{(a) } u_t(y^{tq}) - p \cdot (y^{tq} - a^{tq}) \ge u_t(\widehat{y}) - p \cdot (\widehat{y} - a^{tq}) \\
&\text{for all } \widehat{y} \in \mathbb{R}^T_+, (t,q) \in N , \\
&\text{(b) } \textstyle\sum_{(t,q)\in N} y^{tq} = \sum_{(t,q)} a^{tq} = \overline{y} , \\
&\text{(c) } \xi^{tq} = p \cdot (y^{tq} - a^{tq}) \quad \text{for all} \quad (t,q) \in N \quad \text{and} \\
&\text{(d) } \textstyle\sum_{(t,q)\in N} \xi^{tq} = 0 .
\end{aligned} \tag{6}$$

Given a competitive outcome with allocation $\{(y^{tq}, \xi^{tq}) \in \mathbb{R}^T_+ \times \mathbb{R} : (t,q) \in N\}$ and price vector p, the *competitive payoff to the* $(t,q)^{th}$ *participant* is $u(y^{tq}) - p \cdot (y^{tq} - a^{tq})$. A *competitive payoff vector* is given by

$$(u(y^{tq}) - p \cdot (y^{tq} - a^{tq}) : (t,q) \in N) .$$

In the following we will assume that for each t, all participants of type t have the same endowment; that is, for each t, it holds that $a^{tq} = a^{tq'}$ for all $q, q' = 1, \ldots, n_t$. In this case, every competitive payoff has the equal treatment property;

$$u_t(y^{tq}) - p \cdot (y^{tq} - a^{tq}) = u_t(y^{tq'}) - p \cdot (y^{tq'} - a^{tq'})$$

for all q, q' and for each t. It follows that a competitive payoff vector can be represented by a vector in $\mathbb{R}^T$ with one component for each player type.

It is easy to generate a game from the data of an economy. For each group of participants $S \subset N$, define

$$v(S) = \max \sum_{tq\in S} u_t(y^{tq}, \xi^{tq})$$

where the maximum is taken over the set of S-allocations. Let (N, v) denote a game derived from a market.

Under the assumption of concavity of the utility functions of the participants in an economy, Shapley and Shubik [60] show that a competitive outcome for the market exists and that the competitive payoff vectors are in the core of the game. (Since [22], such results have been obtained in substantially more general models of economies.)

Market-Game Equivalence

To facilitate exposition of the theory of games with many players and the equivalence of markets and games, we consider games derived from a common underlying structure and with a fixed number of types of players, where all players of the same type are substitutes for each other.

Pregames

Let T be a positive integer, to be interpreted as a number of player types. A *profile* $s = (s_1, \ldots, s_T) \in \mathbf{Z}^T_+$, where $\mathbf{Z}^T_+$ is the T-fold Cartesian product of the non-negative integers $\mathbf{Z}_+$, describes a group of players by the numbers of players of each type in the group. Given profile s, define the *norm* or *size* of s by

$$\|s\| \stackrel{\text{def}}{=} \sum_t s_t ,$$

simply the total number of players in a group of players described by s. A *subprofile of a profile* $n \in \mathbf{Z}^T_+$ is a profile s satisfying $s \le n$. A *partition of a profile* s is a collection of subprofiles $\{s^k\}$ of n, not all necessarily distinct, satisfying

$$\sum_k s^k = s .$$

A partition of a profile is analogous to a partition of a set except that all members of a partition of a set are distinct.

Let Ψ be a function from the set of profiles $\mathbf{Z}^T_+$ to $\mathbb{R}_+$ with $\Psi(0) = 0$. The value $\Psi(s)$ is interpreted as the total payoff a group of players with profile s can achieve from collective activities of the group membership and is called the *worth of the profile* s.

Given Ψ, define a worth function Ψ^*, called the *superadditive cover* of Ψ, by

$$\Psi^*(s) \stackrel{\text{def}}{=} \max \sum_k \Psi(s^k) ,$$

where the maximum is taken over the set of all partitions $\{s^k\}$ of s . The function Ψ is said to be *superadditive* if the worth functions Ψ and Ψ^* are equal.

We define a *pregame* as a pair (T, Ψ) where $\Psi : \mathbf{Z}^T_+ \to \mathbb{R}_+$. As we will now discuss, a pregame can be used

to generate multiple games. To generate a game from a pregame, it is only required to specify a total player set N and the numbers of players of each of T types in the set. Then the pregame can be used to assign a worth to every group of players contained in the total player set, thus creating a game.

A *game determined by the pregame* (T, Ψ), which we will typically call a *game* or a *game with side payments*, is a pair $[n; (T, \Psi)]$ where n is a profile. A *subgame* of a game $[n; (T, \Psi)]$ is a pair $[s; (T, \Psi)]$ where s is a subprofile of n.

With any game$[n; (T, \Psi)]$ we can associate a game (N, v) in the form introduced earlier as follows: Let

$$N = \{(t, q) : t = 1, \ldots, T \text{ and } q = 1, \ldots, n_t\}$$

be a *player set* for the game. For each subset $S \subset N$ define the *profile of* S, denoted by prof$(S) \in \mathbf{Z}_+^T$, by its components

$$\text{prof}(S)_t \stackrel{\text{def}}{=} \left|\{S \cap \{(t', q) : t' = t \text{ and } q = 1, \ldots, n_t\}\right|$$

and define

$$v(S) \stackrel{\text{def}}{=} \Psi(\text{prof}(S)) .$$

Then the pair (N, v) satisfies the usual definition of a game with side payments. For any $S \subset N$, define

$$v^*(S) \stackrel{\text{def}}{=} \Psi^*(\text{prof}(S)) .$$

The game (N, v^*) is the *superadditive cover of* (N, v).

A *payoff vector* for a game (N, v) is a vector $\overline{u} \in \mathbb{R}^N$. For each nonempty subset S of N define

$$\overline{u}(S) \stackrel{\text{def}}{=} \sum_{(t,q) \in S} \overline{u}^{tq} .$$

A payoff vector $\overline{u}$ is *feasible for* S if

$$\overline{u}(S) \leq v^*(S) = \Psi^*(\text{prof}(S)) .$$

If $S = N$ we simply say that the payoff vector $\overline{u}$ is *feasible* if

$$\overline{u}(N) \leq v^*(N) = \Psi^*(\text{prof}(N)) .$$

Note that our definition of feasibility is consistent with essential superadditivity; a group can realize at least as large a total payoff as it can achieve in any partition of the group and one way to achieve this payoff is by partitioning into smaller groups.

A payoff vector $\overline{u}$ satisfies the *equal-treatment property* if $\overline{u}^{tq} = \overline{u}^{tq'}$ for all $q, q' \in \{1, \ldots, n_t\}$ and for each $t = 1, \ldots, T$.

Let $[n, (T, \Psi)]$ be a game and let β be a collection of subprofiles of n. The collection is a *balanced collection of subprofiles of* n if there are positive real numbers γ_s for $s \in \beta$ such that $\sum_{s \in \beta} \gamma_s s = n$. The numbers γ_s are called *balancing weights*. Given real number $\varepsilon \geq 0$, the game $[n; (T, \Psi)]$ is *ε-balanced* if for every balanced collection β of subprofiles of n it holds that

$$\Psi^*(n) \geq \sum_{s \in \beta} \gamma_s \left(\Psi(s) - \varepsilon \|s\|\right) \tag{7}$$

where the balancing weights for β are given by γ_s for $s \in \beta$. This definition extends that of Bondareva [13] and Shapley [56] to games with player types. Roughly, a game is (ε) balanced if allowing "part time" groups does not improve the total payoff (by more than ε per player). A game $[n; (T, \Psi)]$ is *totally balanced* if every subgame $[s; (T, \Psi)]$ is balanced.

The *balanced cover game* generated by a game $[n; (T, \Psi)]$ is a game $[n; (T, \Psi^b)]$ where

1. $\Psi^b(s) = \Psi(s)$ for all $s \neq n$ and
2. $\Psi^b(n) \geq \Psi(n)$ and $\Psi^b(n)$ is as small as possible consistent with the nonemptiness of the core of $[n; (T, \Psi^b)]$.

From the Bondareva–Shapley Theorem it follows that $\Psi^b(n) = \Psi^*(n)$ if and only if the game $[n; (T, \Psi)]$ is balanced (ε-balanced, with $\varepsilon = 0$).

For later convenience, the notion of the balanced cover of a pregame is introduced. Let (T, Ψ) be a pregame. For each profile s, define

$$\Psi^b(s) \stackrel{\text{def}}{=} \max_{\beta} \sum_{g \in \beta} \gamma_g \Psi(g) , \tag{8}$$

where the maximum is taken over all balanced collections β of subprofiles of s with weights γ_g for $g \in \beta$. The pair (T, Ψ^b) is called the *balanced cover pregame* of (T, Ψ). Since a partition of a profile is a balanced collection it is immediately clear that $\Psi^b(s) \geq \Psi^*(s)$ for every profile s.

Premarkets

In this section, we introduce the concept of a premarket and re-state results from Shapley and Shubik [60] in the context of pregames and premarkets.

Let $L + 1$ be a number of types of commodities and let $\{\widehat{u}_t(y, \xi) : t = 1, \ldots, T\}$ denote a finite number of functions, called *utility functions*, of the form

$$\widehat{u}_t(y, \xi) = u_t(y) + \xi ,$$

where $y \in \mathbb{R}_+^L$ and $\xi \in \mathbb{R}$. (Such functions, in the literature of economics, are commonly called *quasi-linear*).

Let $\{a^t \in \mathbb{R}^L_+ : t = 1, \ldots, T\}$ be interpreted as a set of *endowments*. We assume that $u_t(a^t) \geq 0$ for each t. For $t = 1, \ldots, T$ we define $c^t \stackrel{\text{def}}{=} (u_t(\cdot), a^t)$ as a *participant type* and let $\mathbb{C} = \{c^t : t = 1, \ldots, T\}$ be the set of participant types. Observe that from the data given by $\mathbb{C}$ we can construct a market by specifying a set of participants N and a function from N to $\mathbb{C}$ assigning endowments and utility functions – types – to each participant in N. A *premarket* is a pair $(T, \mathbb{C})$.

Let $(T, \mathbb{C})$ be a premarket and let $s = (s_1, \ldots, s_T) \in \mathbf{Z}^T_+$. We interpret s as representing a group of economic participants with s_t participants having utility functions and endowments given by c^t for $t = 1, \ldots, T$; for each t, that is, there are s_t participants in the group with type c^t. Observe that the data of a premarket gives us sufficient data to generate a pregame. In particular, given a profile $s = (s_1, \ldots, s_T)$ listing numbers of participants of each of T types, define

$$W(s) \stackrel{\text{def}}{=} \max \sum_t s_t u_t(y^t)$$

where the maximum is taken over the set $\{y^t \in \mathbb{R}^L_+ : t = 1, \ldots, T$ and $\sum_t s_t y^t = \sum_t a^t y^t\}$. Then the pair (T, W) is a *pregame generated by the premarket*.

The following Theorem is an extension to premarkets or a restatement of a result due to Shapley and Shubik [60].

Theorem 1 *Let $(T, \mathbb{C})$ be a premarket derived from economic data in which all utility functions are concave. Then the pregame generated by the premarket is totally balanced.*

Direct Markets and Market-Game Equivalence

Shapley and Shubik [60] introduced the notion of a direct market derived from a totally balanced game. In the direct market, each player is endowed with one unit of a commodity (himself) and all players in the economy have the same utility function. In interpretation, we might think of this as a labor market or as a market for productive factors, (as in [50], for example) where each player owns one unit of a commodity. For games with player types as in this essay, we take the player types of the game as the commodity types of a market and assign all players in the market the same utility function, derived from the worth function of the game.

Let (T, Ψ) be a pregame and let $[n; (T, \Psi)]$ be a derived game. Let $N = \{(t, q) : t = 1, \ldots, T$ and $q = 1, \ldots, n_t$ for each $t\}$ denote the set of players in the game where all participants $\{(t', q) : q = 1, \ldots, n_{t'}\}$ are of type t' for each $t' = 1, \ldots, T$. To construct the direct market generated by a derived game $[n; (T, \Psi)]$, we take the commodity space as $\mathbb{R}^T_+$ and suppose that each participant in the market of type t is endowed with one unit of the tth commodity, and thus has endowment $\mathbf{1}_t = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^T_+$ where "1" is in the tth position. The total endowment of the economy is then given by $\sum n_t \mathbf{1}_t = n$.

For any vector $y \in \mathbb{R}^T_+$ define

$$u(y) \stackrel{\text{def}}{=} \max \sum_{s \leq n} \gamma_s \Psi(s) \,, \tag{9}$$

the maximum running over all $\{\gamma_s \geq 0 : s \in \mathbf{Z}^T_+,\ s \leq n\}$ satisfying

$$\sum_{s \leq n} \gamma_s s = y \,. \tag{10}$$

As noted by Shapley and Shubik [60], but for our types case, it can be verified that the function u is concave and one-homogeneous. This does not depend on the balancedness of the game $[n; (T, \Psi)]$. Indeed, one may think of u as the "balanced cover of $[n; (T, \Psi)]$ extended to $\mathbb{R}^T_+$". Note also that u is superadditive, independent of whether the pregame (T, Ψ) is superadditive. We leave it to the interested reader to verify that if Ψ were not necessarily superadditive and Ψ^* is the superadditive cover of Ψ then it holds that $\max \sum_{s \leq n} \gamma_s \Psi(s) = \max \sum_{s \leq n} \gamma_s \Psi^*(s)$.

Taking the utility function u as the utility function of each player $(t, q) \in N$ where N is now interpreted as the set of participants in a market, we have generated a market, called the *direct market*, denoted by $[n, u; (T, \Psi)]$, from the game $[n; (T, \Psi)]$.

Again, the following extends a result of Shapley and Shubik [60] to pregames.

Theorem 2 *Let $[n, u; (T, \Psi)]$ denote the direct market generated by a game $[n; (T, \Psi)]$ and let $[n; (T, u)]$ denote the game derived from the direct market. Then, if $[n; (T, \Psi)]$ is a totally balanced game, it holds that $[n; (T, u)]$ and $[n; (T, \Psi)]$ are identical.*

Remark 2 If the game $[n; (T, \Psi)]$ and every subgame $[s, (T, \Psi)]$ has a nonempty core – that is, if the game is 'totally balanced'– then the game $[n; (T, u)]$ generated by the direct market is the initially given game $[n; (T, \Psi)]$. If however the game $[n; (T, \Psi)]$ is not totally balanced then $u(s) \geq \Psi(s)$ for all profiles $s \leq n$. But, whether or not $[n; (T, \Psi)]$ is totally balanced, the game $[n; (T, u)]$ is totally balanced and coincides with the totally balanced cover of $[n; (T, \Psi)]$.

Remark 3 Another approach to the equivalence of markets and games is taken by Garratt and Qin [26], who define a class of direct lottery markets. While a player can

participate in only one coalition, both ownership of coalitions and participation in coalitions is determined randomly. Each player is endowed with one unit of probability, his own participation. Players can trade their endowments at market prices. The core of the game is equivalent to the equilibrium of the direct market lottery.

Equivalence of Markets and Games with Many Players

The requirement of Shapley and Shubik [60] that utility functions be concave is restrictive. It rules out, for example situations such as economies with indivisible commodities. It also rules out club economies; for a given club structure of the set of players – in the simplest case, a partition of the total player set into groups where collective activities only occur within these groups – it may be that utility functions are concave over the set of alternatives available within each club, but utility functions need not be concave over all possible club structures. This rules out many examples; we provide a simple one below.

To obtain the result that with many players, games derived from pregames are market games, we need some further assumption on pregames. If there are many substitutes for each player, then the simple condition that *per capita payoffs are bounded* – that is, given a pregame (T, Ψ), that there exists some constant K such that $\frac{\Psi(s)}{\|s\|} < K$ for all profiles s – suffices. If, however, there may be 'scarce types', that is, players of some type(s) become negligible in the population, then a stronger assumption of 'small group effectiveness' is required. We discuss these two conditions in the next section.

Small Group Effectiveness and Per Capita Boundedness

This section discusses conditions limiting gains to group size and their relationships. This definition was introduced in Wooders [83], for NTU, as well as TU, games.

PCB A pregame (T, Ψ) satisfies *per capita boundedness* (PCB) if

$$PCB: \quad \sup_{s \in \mathbf{Z}_+^T} \frac{\Psi(s)}{\|s\|} \text{ is finite} \tag{11}$$

or equivalently,

$$\sup_{s \in \mathbf{Z}_+^T} \frac{\Psi^*(s)}{\|s\|} \text{ is finite .}$$

It is known that under the apparently mild conditions of PCB and essential superadditivity, in general games with many players of each of a finite number of player types and a fixed distribution of player types have nonempty approximate cores; Wooders [81,83]. (Forms of these assumptions were subsequently also used in Shubik and Wooders [69,70]; Kaneko and Wooders [35]; and Wooders [89,91] among others.) Moreover, under the same conditions, approximate cores have the property that most players of the same type are treated approximately equally ([81,94]; see also Shubik and Wooders [69]). These results, however, either require some assumption ruling out 'scarce types' of players, for example, situations where there are only a few players of some particular type and these players can have great effects on total feasible payoffs. Following are two examples. The first illustrates that PCB does not control limiting properties of the per capita payoff function when some player types are scarce.

Example 3 ([94]) Let $T = 2$ and let (T, Ψ) be the pregame given by

$$\Psi(s_1, s_2) = \begin{cases} s_1 + s_2 & \text{when } s_1 > 0 \\ 0 & \text{otherwise .} \end{cases}$$

The function Ψ obviously satisfies PCB. But there is a problem in defining $\lim \Psi(s_1, s_2)/s_1 + s_2$ as $s_1 + s_2$ tends to infinity, since the limit depends on how it is approached. Consider the sequence (s_1^ν, s_2^ν) where $(s_1^\nu, s_2^\nu) = (0, \nu)$; then $\lim \Psi(s_1^\nu, s_2^\nu)/s_1^\nu + s_2^\nu = 0$. Now suppose in contrast that $(s_1^\nu, s_2^\nu) = (1, \nu)$; then $\lim \Psi(s_1^\nu, s_2^\nu)/s_1^\nu + s_2^\nu = 1$. This illustrates why, to obtain the result that games with many players are market games either it must be required that there are no scarce types or some some assumption limiting the effects of scarce types must be made. We return to this example in the next section.

The next example illustrates that, with only PCB, uniform approximate cores of games with many players derived from pregames may be empty.

Example 4 ([94]) Consider a pregame (T, Ψ) where $T = \{1, 2\}$ and Ψ is the superadditive cover of the function Ψ' defined by:

$$\Psi'(s) \stackrel{\text{def}}{=} \begin{cases} |s| & \text{if } s_1 = 2 \,, \\ 0 & \text{otherwise .} \end{cases}$$

Thus, if a profile $s = (s_1, s_2)$ has $s_1 = 2$ then the worth of the profile according to Ψ' is equal to the total number of players it represents, $s_1 + s_2$, while all other profiles s have worth of zero. In the superadditive cover game the worth of a profile s is 0 if $s_1 < 2$ and otherwise is equal to s_2 plus the largest even number less than or equal to s_1.

Now consider a sequence of profiles $(s^\nu)_\nu$ where $s_1^\nu = 3$ and $s_2^\nu = \nu$ for all ν. Given $\varepsilon > 0$, for all sufficiently

large player sets the uniform ε-core is empty. Take, for example, $\varepsilon = 1/4$. If the uniform ε-core were nonempty, it would have to contain an equal-treatment payoff vector.[1] For the purpose of demonstrating a contradiction, suppose that $u^\nu = (u_1^\nu, u_2^\nu)$ represents an equal treatment payoff vector in the uniform ε-core of $[s^\nu; (T, \Psi)]$. The following inequalities must hold:

$$\begin{aligned} 3u_1^\nu + \nu u_2^\nu &\leq \nu + 2\,, \\ 2u_1^\nu + \nu u_2^\nu &\geq \nu + 2, \text{ and} \\ u_1^\nu &\geq \tfrac{3}{4}\,. \end{aligned}$$

which is impossible. A payoff vector which assigns each player zero is, however, in the weak ε-core for any $\varepsilon > \frac{1}{\nu+3}$. But it is not very appealing, in situations such as this, to ignore a relatively small group of players (in this case, the players of type 1) who can have a large effect on per capita payoffs. This leads us to the next concept.

To treat the scarce types problem, Wooders [88,89,90] introduced the condition of small group effectiveness (SGE). SGE is appealing technically since it resolves the scarce types problem. It is also economically intuitive and appealing; the condition defines a class of economies that, when there are many players, generate competitive markets. Informally, SGE dictates that *almost all* gains to collective activities can be realized by relatively small groups of players. Thus, SGE is exactly the sort of assumption required to ensure that multiple, relatively small coalitions, firms, jurisdictions, or clubs, for example, are optimal or near-optimal in large economies.

A pregame (T, Ψ) satisfies *small group effectiveness, SGE*, if:

$$SGE: \quad \begin{array}{c} \text{For each real number } \varepsilon > 0, \\ \text{there is an integer } \eta_0(\varepsilon) \\ \text{such that for each profile } s, \\ \text{for some partition } \{s^k\} \text{ of } s \text{ with} \\ \|s^k\| \leq \eta_0(\varepsilon) \text{ for each subprofile } s^k, \text{ it holds that} \\ \Psi^*(s) - \sum_k \Psi(s^k) \leq \varepsilon \|s\|\,; \end{array} \tag{12}$$

given $\varepsilon > 0$ there is a group size $\eta_0(\varepsilon)$ such that the loss from restricting collective activities within groups to groups containing fewer that $\eta_0(\varepsilon)$ members is at most ε per capita [88].[2]

[1] It is well known and easily demonstrated that the uniform ε-core of a TU game is nonempty if and only if it contains an equal treatment payoff vector. This follows from the fact that the uniform ε-core is a convex set.

[2] Exactly the same definition applies to situations with a compact metric space of player types, c.f. Wooders [84,88].

SGE also has the desirable feature that if there are no 'scarce types' – types of players that appear in vanishingly small proportions– then SGE and PCB are equivalent.

Theorem 3 ([91] With 'thickness,' SGE = PCB) *(1) Let (T, Ψ) be a pregame satisfying SGE. Then the pregame satisfies PCB.*

(2) Let (T, Ψ) be a pregame satisfying PCB. Then given any positive real number ρ, construct a new pregame (T, Ψ_ρ) where the domain of Ψ_ρ is restricted to profiles s where, for each $t = 1, \ldots, T$, either $\frac{s_t}{\|s\|} > \rho$ or $s_t = 0$. Then (T, Ψ_ρ) satisfies SGE on its domain.

It can also be shown that small groups are effective for the attainment of nearly all feasible outcomes, as in the above definition, if and only if small groups are effective for improvement – any payoff vector that can be significantly improved upon can be improved upon by a small group (see Proposition 3.8 in [89]).

Remark 4 Under a stronger condition of *strict* small group effectiveness, which dictates that $\eta(\varepsilon)$ in the definition of small group effectiveness can be chosen independently of ε, stronger results can be obtained than those presented in this section and the next. We refer to Winter and Wooders [80] for a treatment of this case.

Remark 5 (On the importance of taking into account scarce types) Recall the quotation from von Neumann and Morgenstern and the discussion following the quotation. The assumption of per capita boundedness has significant consequences but is quite innocuous – ruling out the possibility of average utilities becoming infinite as economies grow large does not seem restrictive. But with only per capita boundedness, even the formation of small coalitions can have significant impacts on aggregate outcomes. With small group effectiveness, however, there is no problem of either large or small coalitions acting together – large coalitions cannot do significantly better then relatively small coalitions.

Roughly, the property of large games we next introduce is that relatively small groups of players make only "asymptotic negligible" contributions to per-capita payoffs of large groups. A pregame (Ω, Ψ) satisfies *asymptotic negligibility* if, for any sequence of profiles $\{f^\nu\}$ where

$$\begin{aligned} &\|f^\nu\| \to \infty \text{ as } \nu \to \infty, \\ &\sigma(f^\nu) = \sigma(f^{\nu'}) \quad \text{for all} \quad \nu \quad \text{and} \quad \nu' \quad \text{and} \\ &\lim_{\nu\to\infty} \tfrac{\Psi^*(f^\nu)}{\|f^\nu\|} \quad \text{exists}\,, \end{aligned} \tag{13}$$

then for any sequence of profiles $\{\ell^\nu\}$ with

$$\lim_{\nu\to\infty} \frac{\|\ell^\nu\|}{\|f^\nu\|} = 0\,, \tag{14}$$

it holds that

$$\begin{aligned} &\lim_{\nu\to\infty} \frac{\Psi^*\|f^\nu+\ell^\nu\|}{\|f^\nu+\ell^\nu\|} \text{ exists, and} \\ &\lim_{\nu\to\infty} \frac{\Psi^*\|f^\nu+\ell^\nu\|}{\|f^\nu+\ell^\nu\|} = \lim_{\nu\to\infty} \frac{\Psi^*(f^\nu)}{\|f^\nu\|} . \end{aligned} \tag{15}$$

Theorem 4 ([89,95]) *A pregame (T, Ψ) satisfies SGE if and only if it satisfies PCB and asymptotic negligibility*

Intuitively, asymptotic negligibility ensures that vanishingly small percentages of players have vanishingly small effects on aggregate per-capita worths. It may seem paradoxical that SGE, which highlights the importance of relatively small groups, is equivalent to asymptotic negligibility. To gain some intuition, however, think of a marriage model where only two-person marriages are allowed. Obviously two-person groups are (strictly) effective, but also, in large player sets, no two persons can have a substantial affect on aggregate per-capita payoffs.

Remark 6 Without some assumptions ensuring essential superadditivity, at least as incorporated into our definition of feasibility, nonemptiness of approximate cores of large games cannot be expected; superadditivity assumptions (or the close relative, essential superadditivity) are heavily relied upon in all papers on large games cited. In the context of economies, superadditivity is a sort of monotonicity of preferences or production functions assumption, that is, superadditivity of Ψ implies that for all $s, s' \in \mathbf{Z}_+^T$, it holds that $\Psi(s + s') \geq \Psi(s) + \Psi(s')$. Our assumption of small group effectiveness, SGE, admits nonmonotonicities. For example, suppose that 'two is company, three or more is a crowd,' by supposing there is only one commodity and by setting $\Psi(2) = 2$, $\Psi(n) = 0$ for $n \neq 2$. The reader can verify, however, that this example satisfies small group effectiveness since $\Psi^*(n) = n$ if n is even and $\Psi^*(n) = n - 1$ otherwise. Within the context of pregames, requiring the superadditive cover payoff to be approximately realizable by partitions of the total player set into relatively small groups is the weakest form of superadditivity required for the equivalence of games with many players and concave markets.

Derivation of Markets from Pregames Satisfying SGE

With SGE and PCB in hand, we can now derive a premarket from a pregame and relate these concepts.

To construct a limiting direct premarket from a pregame, we first define an appropriate utility function. Let (T, Ψ) be a pregame satisfying SGE. For each vector x in $\mathbb{R}_+^T$ define

$$U(x) \stackrel{\text{def}}{=} \|x\| \lim_{\nu\to\infty} \frac{\Psi^*(f^\nu)}{\|f^\nu\|} \tag{16}$$

where the sequence $\{f^\nu\}$ satisfies

$$\begin{aligned} &\lim_{\nu\to\infty} \frac{f^\nu}{\|f^\nu\|} = \frac{x}{\|x\|} \\ &\text{and} \\ &\|f^\nu\| \to \infty . \end{aligned} \tag{17}$$

Theorem 5 ([84,91]) *Assume the pregame (T, Ψ) satisfies small group effectiveness. Then for any $x \in \mathbb{R}_+^T$ the limit (16) exists. Moreover, $U(\cdot)$ is well-defined, concave and 1-homogeneous and the convergence is uniform in the sense that, given $\varepsilon > 0$ there is an integer η such that for all profiles s with $\|s\| \leq \eta$ it holds that*

$$\left| U\left(\frac{s}{\|s\|}\right) - \frac{\Psi^*(s)}{\|s\|} \right| \leq \varepsilon .$$

From Wooders [91] (Theorem 4), if arbitrarily small percentages of players of any type that appears in games generated by the pregame are ruled out, then the above result holds under per capita boundedness [91] (Theorem 6). As noted in the introduction to this paper, for the TU case, the concavity of the limiting utility function, for the model of Wooders [83] was first noted by Aumann [5]. The concavity is shown to hold with a compact metric space of player types in Wooders [84] and is simplified to the finite types case in Wooders [91].

Theorem 5 follows from the facts that the function U is superadditive and 1-homogeneous on its domain. Since U is concave, it is continuous on the interior of its domain; this follows from PCB. Small group effectiveness ensures that the function U is continuous on its entire domain [91](Lemma 2).

Theorem 6 ([91]) *Let (T, Ψ) be a pregame satisfying small group effectiveness and let (T, U) denote the derived direct market pregame. Then (T, U) is a totally balanced market game. Moreover, U is one-homogeneous, that is, $U(\lambda x) = \lambda U(x)$ for any non-negative real number λ.*

In interpretation, T denotes a number of types of players/commodities and U denotes a utility function on $\mathbb{R}_+^T$. Observe that when U is restricted to profiles (in $\mathbf{Z}_+^T$), the pair (T, U) is a pregame with the property that every game $[n; (T, U)]$ has a nonempty core; thus, we will call (T, U) the *premarket generated by the pregame* (T, Ψ). That every game derived from (T, U) has a nonempty core is a consequence of the Shapley and Shubik [60] result that market games derived from markets with concave utility functions are totally balanced.

It is interesting to note that, as discussed in Wooders (Section 6 in [91]), if we restrict the number of commodities to equal the number of player types, then the utility

function U is *uniquely* determined. (If one allowed more commodities then one would effectively have 'redundant assets'.) In contrast, for games and markets of fixed, finite size, as demonstrated in Shapley and Shubik [62], even if we restrict the number of commodities to equal the number of player types, given any nonempty, compact, convex subset of payoff vectors in the core, it is possible to construct utility functions so that this subset coincides with the set of competitive payoffs. Thus, in the Shapley and Shubik approach, equivalence of the core and the set of price-taking competitive outcomes for the direct market is only an artifact of the method used there of constructing utility functions from the data of a game and is quite distinct from the equivalence of the core and the set of competitive payoff vectors as it is usually understood (that is, in the sense of Debreu and Scarf [22] and Aumann [4]. See also Kalai and Zemel [31,32] which characterize the core in multi-commodity flow games.

Cores and Approximate Cores

The concept of the core clearly was important in the work of Shapley and Shubik [59,60,62] and is also important for the equivalence of games with many players and market games. Thus, we discuss the related results of nonemptiness of approximate cores and convergence of approximate cores to the core of the 'limit' – the game where all players have utility functions derived from a pregame and large numbers of players. First, some terminology is required. A vector p is a *subgradient at x* of the concave function U if $U(y) - U(x) \leq p \cdot (y - x)$ for all y. One might think of a subgradient as a bounding hyperplane. To avoid any confusion it might be helpful to note that, as Mas-Colell [46] remarks: " Strictly speaking, one should use the term *subgradient* for convex functions and *supergradient* for concave. But this is cumbersome", (p. 29–30 in [46]).

For ease of notation, equal-treatment payoff vectors for a game $[n; (T, \Psi)]$ will typically be represented as vectors in $\mathbb{R}^T$. An *equal-treatment payoff vector*, or simply a *payoff vector* when the meaning is clear, is a point $\overline{x}$ in $\mathbb{R}^T$. The tth component of $\overline{x}$, $\overline{x}_t$, is interpreted as the payoff to each player of type t. The feasibility of an equal-treatment payoff vector $\overline{x} \in \mathbb{R}^T$ for the game $[n; (T, \Psi)]$ can be expressed as:

$$\Psi^*(n) \geq \overline{x} \cdot n \,.$$

Let $[n; (T, \Psi)]$ be a game determined by a pregame (T, Ψ), let ε be a non-negative real number, and let $\overline{x} \in \mathbb{R}^T$ be a (equal-treatment) payoff vector. Then $\overline{x}$ is in the *equal-treatment ε-core* of $[n; (T, \Psi)]$ or simply "in the ε-core" when the meaning is clear, if $\overline{x}$ is feasible for $[n; (T, \Psi)]$ and

$$\Psi(s) \leq \overline{x} \cdot s + \varepsilon \|s\| \text{ for all subprofiles } s \text{ of } n \,.$$

Thus, the equal-treatment ε-core is the set

$$C(n; \varepsilon) \stackrel{\text{def}}{=} \{\overline{x} \in \mathbb{R}^T_+ : \Psi^*(n) \geq \overline{x} \cdot n \quad \text{and} \quad \Psi(s) \leq \overline{x} \cdot s + \varepsilon \|s\| \quad \text{for all subprofiles} \quad s \text{ of } n\} \,. \tag{18}$$

It is well known that the ε-core of a game with transferable utility is nonempty if and only if the equal-treatment ε-core is nonempty.

Continuing with the notation above, for any $s \in \mathbb{R}^T_+$, let $\Pi(s)$ denote the set of subgradients to the function U at the point s;

$$\Pi(s) \stackrel{\text{def}}{=} \{\pi \in \mathbb{R}^T : \pi \cdot s = U(s) \text{ and } \pi \cdot s' \geq U(s') \text{ for all } s' \in \mathbb{R}^T_+\} \,. \tag{19}$$

The elements in $\Pi(s)$ can be interpreted as equal-treatment core payoffs to a limiting game with the mass of players of type t given by s_t. The core payoff to a player is simply the value of the one unit of a commodity (himself and all his attributes, including endowments of resources) that he owns in the direct market generated by a game. Thus $\Pi(\cdot)$ is called *the limiting core correspondence* for the pregame (T, Ψ). Of course $\Pi(\cdot)$ is also the limiting core correspondence for the pregame (T, U).

Let $\widehat{\Pi}(n) \subset \mathbb{R}^T$ denote *equal-treatment core of the market game* $[n; (T, u)]$:

$$\widehat{\Pi}(n) \stackrel{\text{def}}{=} \{\pi \in \mathbb{R}^T : \pi \cdot n = u(n) \quad \text{and} \quad \pi \cdot s \geq u(s) \quad \text{for all} \quad s \in \mathbf{Z}^T_+, s \leq n\} \,. \tag{20}$$

Given any player profile n and derived games $[n; (T, \Psi)]$ and $[n; (T, U)]$ it is interesting to observe the distinction between the equal-treatment core of the game $[n; (T, U)]$, denoted by $\widehat{\Pi}(n)$, defined by (20), and the set $\Pi(n)$ (that is, $\Pi(\overline{x})$ with $\overline{x} = n$). The definitions of $\Pi(n)$ and $\widehat{\Pi}(n)$ are the same except that the qualification "$s \leq n$" in the definition of $\widehat{\Pi}(n)$ does not appear in the definition of $\Pi(n)$. Since $\Pi(n)$ is the *limiting* core correspondence, it takes into account arbitrarily large coalitions. For this reason, for any $\overline{x} \in \Pi(n)$ and $\widehat{x} \in \widehat{\Pi}(n)$ it holds that $\overline{x} \cdot n \geq \widehat{x} \cdot n$. A simple example may be informative.

Example 5 Let (T, Ψ) be a pregame where $T = 1$ and $\Psi(n) = n - \frac{1}{n}$ for each $n \in \mathbb{Z}_+$, and let $[n; (T, \Psi)]$ be a derived game. Then $\Pi(n) = \{1\}$ while $\widehat{\Pi}(n) = \{(1 - \frac{1}{n^2})\}$.

The following Theorem extends a result due to Shapley and Shubik [62] stated for games derived from pregames.

Theorem 7 ([62]) *Let $[n;(T,\Psi)]$ be a game derived from a pregame and let $[n,u;(T,\Psi)]$ be the direct market generated by $[n;(T,\Psi)]$. Then the equal-treatment core $\widehat{\Pi}(n)$ of the game $[n;(T,u)]$ is nonempty and coincides with the set of competitive price vectors for the direct market $[n,u;(T,\Psi)]$.*

Remark 7 Let (T,Ψ) be a pregame satisfying PCB. In the development of the theory of large games as models of competitive economies, the following function on the space of profiles plays an important role:

$$\lim_{r\to\infty}\frac{\Psi^*(rf)}{r}\,;$$

see, for example, Wooders [81] and Shubik and Wooders [69]. For the purposes of comparison, we introduce another definition of a limiting utility function. For each vector x in $\mathbb{R}^T_+$ with rational components let $r(x)$ be the smallest integer such that $r(x)x$ is a vector of integers. Therefore, for each rational vector x, we can define

$$\hat{U}(x)\stackrel{\text{def}}{=}\lim_{\nu\to\infty}\frac{\Psi^*(\nu r(x)x)}{\nu r(x)}\,.$$

Since Ψ^* is superadditive and satisfies per capita boundedness, the above limit exists and $\hat{U}(\cdot)$ is well-defined. Also, $\hat{U}(x)$ has a continuous extension to any closed subset strictly in the interior of $\mathbb{R}^T_+$. The function $\hat{U}(x)$, however, may be discontinuous at the boundaries of $\mathbb{R}^T_+$. For example, suppose that $T=2$ and

$$\Psi^*(k,n)=\begin{cases}k+n & \text{when } k>0\\ 0 & \text{otherwise}\,.\end{cases}$$

The function Ψ^* obviously satisfies PCB but does not satisfy SGE. To see the continuity problem, consider the sequences $\{x^\nu\}$ and $\{y^\nu\}$ of vectors in $\mathbb{R}^2_+$ where $x^\nu=(\frac{1}{\nu},\frac{\nu-1}{\nu})$ and $y^\nu=(0,\nu)$. Then $\lim_{\nu\to\infty}x^\nu=\lim_{\nu\to\infty}y^\nu=(0,1)$ but $\lim_{\nu\to\infty}\hat{U}(x^\nu)=1$ while $\lim_{\nu\to\infty}\hat{U}(y^\nu)=0$. SGE is precisely the condition required to avoid this sort of discontinuity, ensuring that the function U is continuous on the boundaries of $\mathbb{R}^T_+$.

Before turning to the next section, let us provide some additional interpretation for $\widehat{\Pi}(n)$. Suppose a game $[n;(T,\Psi)]$ is one generated by an economy, as in Shapley and Shubik [59] or Owen [50], for example. Players of different types may have different endowments of private goods. An element π in $\widehat{\Pi}(n)$ is an equal-treatment payoff vector in the core of the balanced cover game generated by $[n;(T,\Psi)]$ and can be interpreted as listing prices for player types where π_t is the price of a player of type t; this price is a price for the player himself, *including* his endowment of private goods.

Nonemptiness and Convergence of Approximate Cores of Large Games

The next Proposition is an immediate consequence of the convergence of games to markets shown in Wooders [89,91] and can also be obtained as a consequence of Theorem 5 above.

Proposition 2 (Nonemptiness of approximate cores) *Let (T,Ψ) be a pregame satisfying SGE. Let ε be a positive real number. Then there is an integer $\eta_1(\varepsilon)$ such that any game $[n;(T,\Psi)]$ with $\|n\|\geq\eta_1(\varepsilon)$ has a nonempty uniform ε-core.*

(Note that no assumption of superadditivity is required but only because our definition of feasibility is equivalent to feasibility for superadditive covers.)

The following result was stated in Wooders [89]. For more recent results see Wooders [94].

Theorem 8 ([89] Uniform closeness of (equal-treatment) approximate cores to the core of the limit game) *Let (T,Ψ) be a pregame satisfying SGE and let $\Pi(\cdot)$ be as defined above. Let $\delta>0$ and $\rho>0$ be positive real numbers. Then there is a real number ε^* with $0<\varepsilon^*$ and an integer $\eta_0(\delta,\rho,\varepsilon^*)$ with the following property: for each positive $\varepsilon\in(0,\varepsilon^*]$ and each game $[f;(T,\Psi)]$ with $\|f\|>\eta_0(\delta,\rho,\varepsilon^*)$ and $f_t/\|f\|\geq\rho$ for each $t=1,\ldots,T$, if $C(f;\varepsilon)$ is nonempty then both*

$$\text{dist}[C(f;\varepsilon),\Pi(f)]<\delta \text{ and } \text{dist}[C(f;\varepsilon),\widehat{\Pi}(f)]<\delta\,,$$

where 'dist' is the Hausdorff distance with respect to the sum norm on $\mathbb{R}^T$.

Note that this result applies to games derived from diverse economies, including economies with indivisibilities, nonmonotonicities, local public goods, clubs, and so on.

Theorem 8 motivates the question of whether approximate cores of games derived from pregames satisfying small group effectiveness treat players most of the same type nearly equally. The following result, from Wooders [81,89,93] answers this question.

Theorem 9 *Let (T,Ψ) be a pregame satisfying SGE. Then given any real numbers $\gamma>0$ and $\lambda>0$ there is a positive real number ε^* and an integer ρ such that for each $\varepsilon\in[0,$*

ε^*] and for every profile $n \in \mathbb{Z}_+^T$ with $\|n\|_1 > \rho$, if $x \in \mathbb{R}^N$ is in the uniform ε-core of the game $[n, \Psi]$ with player set

$$N = \{(t,q) : t = 1, \ldots, T \text{ and, for each } t, q = 1, \ldots, n_t\}$$

then, for each $t \in \{1, \ldots, T\}$ with $\frac{n_t}{\|n\|_1} \geq \frac{\lambda}{2}$ it holds that

$$|\{(t,q) : |x^{tq} - z_t| > \gamma\}| < \lambda n_t\} ,$$

where, for each $t = 1, \ldots, T$,

$$z_t = \frac{1}{n_t} \sum_{q=1}^{n_t} x^{tq} ,$$

the average payoff received by players of type t.

Shapley Values of Games with Many Players

Let (N, v) be a game. The *Shapley value* of a superadditive game is the payoff vector whose ith component is given by

$$SH(v, i) = \frac{1}{|N|} \sum_{J=0}^{|N|-1} \frac{1}{\binom{|N|-1}{J}} \sum_{\substack{S \subset N \setminus \{i\} \\ |S| = J}} \left[v(S \cup \{i\}) - v(S) \right] .$$

To state the next Theorem, we require one additional definition. Let (T, Ψ) be a pregame. The pregame satisfies *boundedness of marginal contributions* (BMC) if there is a constant M such that

$$|\Psi(s + 1_t) - \Psi(s)| \leq M$$

for all vectors $1_t = (0, \ldots, 0, 1_{t^{th}\text{ place}}, 0, \ldots 0)$ for each $t = 1, \ldots, T$. Informally, this condition bounds marginal contributions while SGE bounds average contributions. That BMC implies SGE is shown in Wooders [89]. The following result restricts the main Theorem of Wooders and Zame [96] to the case of a finite number of types of players.

Theorem 10 ([96]) *Let (T, Ψ) be a superadditive pregame satisfying boundedness of marginal contributions. For each $\varepsilon > 0$ there is a number $\delta(\varepsilon) > 0$ and an integer $\mu(\varepsilon)$ with the following property:*

If $[n, (T, \Psi)]$ is a game derived from the pregame, for which $n_t > \mu(\varepsilon)$ for each t, then the Shapley value of the game is in the (weak) ε-core.

Similar results hold within the context of private goods exchange economies (cf., Shapley [55], Shapley and Shubik [60], Champsaur [17], Mas-Colell [43], Cheng [18] and others). Some of these results are for economies without money but all treat private goods exchange economies with divisible goods and concave, monotone utility functions. Moreover, they all treat either replicated sequences of economies or convergent sequences of economies. That games satisfying SGE are asymptotically equivalent to balanced market games clarifies the contribution of the above result. In the context of the prior results developed in this paper, the major shortcoming of the Theorem is that it requires BMC. This author conjectures that the above result, or a close analogue, could be obtained with the milder condition of SGE, but this has not been demonstrated.

Economies with Clubs

By a club economy we mean an economy where participants in the economy form groups – called clubs – for the purposes of collective consumption and/or production collectively with the group members. The groups may possibly overlap. A club structure of the participants in the economy is a covering of the set of players by clubs. Providing utility functions are quasi-linear, such an economy generates a game of the sort discussed in this essay. The worth of a group of players is the maximum total worth that the group can achieve by forming clubs. The most general model of clubs in the literature at this point is Allouch and Wooders [1]. Yet, if one were to assume that utility functions were all quasi-linear and the set of possible types of participants were finite. the results of this paper would apply.

In the simplest case, the utility of an individual depends on the club profile (the numbers of participants of each type) in his club. The total worth of a group of players is the maximum that it can achieve by splitting into clubs. The results presented in this section immediately apply. When there are many participants, club economies can be represented as markets and the competitive payoff vectors for the market are approximated by equal-treatment payoff vectors in approximate cores. Approximate cores converge to equal treatment and competitive equilibrium payoffs. A more general model making these points is treated in Shubik and Wooders [65]. For recent reviews of the literature, see Conley and Smith [19] and Kovalenkov and Wooders [38].[3]

Coalition production economies may also be viewed as club economies. We refer the reader to Böhm [12], Son-

[3] Other approaches to economies with clubs/local public goods include Casella and Feinstein [15], Demange [23], Haimanko, O., M. Le Breton and S. Weber [28], and Konishi, Le Breton and Weber [37]. Recent research has treated clubs as networks.

dermann [73], Shubik and Wooders [70], and for a more recent treatment and further references, Sun, Trockel and Yang [74]).

Let us conclude this section with some historical notes. Club economies came to the attention of the economics profession with the publication of Buchanan [14]. The author pointed out that people care about the numbers of other people with whom they share facilities such as swimming pool clubs. Thus, there may be congestion, leading people to form multiple clubs. Interestingly, much of the recent literature on club economies with many participants and their competitive properties has roots in an older paper, Tiebout [77]. Tiebout conjectured that if public goods are 'local' – that is, subject to exclusion and possibly congestion – then large economies are 'market-like'. A first paper treating club economies with many participants was Pauly [51], who showed that, when all players have the same preferred club size, then the core of economy is nonempty if and only if all participants in the economy can be partitioned into groups of the preferred size. Wooders [82] modeled a club economy as one with local public goods and demonstrated that, when individuals within a club (jurisdiction) are required to pay the same share of the costs of public good provision, then outcomes in the core permit heterogeneous clubs if and only if all types of participants in the same club have the same demands for local public goods and for congestion. Since these early results, the literature on clubs has grown substantially.

With a Continuum of Players

Since Aumann [4] much work has been done on economies with a continuum of players. It is natural to question whether the asymptotic equivalence of markets and games reported in this article holds in a continuum setting. Some such results have been obtained.

First, let $N = [01]$ be the 0,1 interval with Lesbegue measure and suppose there is a partition of N into a finite set of subsets $N_1, \ldots, N_T$ where, in interpretation, a point in N_t represents a player of type t. Let Ψ be given. Observe that Ψ determines a payoff for any finite group of players, depending on the numbers of players of each type. If we can aggregate partitions of the total player set into finite coalitions then we have defined a game with a continuum of players and finite coalitions.

For a partition of the continuum into finite groups to 'make sense' economically, it must preserve the relative scarcities given by the measure. This was done in Kaneko and Wooders [35]. To illustrate their idea of measurement consistent partitions of the continuum into finite groups, think of a census form that requires each three-person household to label the players in the household, #1, #2, or #3. When checking the consistency of its figures, the census taker would expect the numbers of people labeled #1 in three-person households to equal the numbers labeled #2 and #3. For consistency, the census taker may also check that the number of first persons in three-person households in a particular region is equal to the number of second persons and third persons in three person households in that region. It is simple arithmetic. This consistency should also hold for k-person households for any k. Measurement consistency is the same idea with the work "number" replaced by "proportion" or "measure".

One can immediately apply results reported above to the special case of TU games of Kaneko–Wooders [35] and conclude that games satisfying small group effectiveness and with a continuum of players have nonempty cores and that the payoff function for the game is one-homogeneous. (We note that there have been a number of papers investigating cores of games with a continuum of players that have came to the conclusion that non-emptiness of exact cores does not hold, even with balancedness assumptions, cf., Weber [78,79]). The results of Wooders [91], show that the continuum economy must be representable by one where all players have the same concave, continuous one-homogeneous utility functions. Market games with a continuum of players and a finite set of types are also investigated in Azriel and Lehrer [3], who confirm these conclusions.)

Other Related Concepts and Results

In an unpublished 1972 paper due to Edward Zajac [97], which has motivated a large amount of literature on 'subsidy-free pricing', cost sharing, and related concepts, the author writes:

> "A fundamental idea of equity in pricing is that 'no consumer group should pay higher prices than it would pay by itself…'. If a particular group is paying a higher price than it would pay if it were severed from the total consumer population, the group feels that it is subsidizing the total population and demands a price reduction".

The "dual" of the cost allocation problem is the problem of surplus sharing and subsidy-free pricing.[4] Tauman [75] provides a excellent survey. Some recent works treating cost allocation and subsidy free-pricing include

[4] See, for example Moulin [47,48] for excellent discussions of these two problems.

Moulin [47,48]. See also the recent notion of "Walras' core" in Qin, Shapley and Shimomura [52].

Another related area of research has been into whether games with many players satisfy some notion of the Law of Demand of consumer theory (or the Law of Supply of producer theory). Since games with many players resemble market games, which have the property that an increase in the endowment of a commodity leads to a decrease in its price, such a result should be expected. Indeed, for games with many players, a Law of Scarcity holds – if the numbers of players of a particular type is increased, then core payoffs to players of that type do not increase and may decrease. (This result was observed by Scotchmer and Wooders [54]). See Kovalenkov and Wooders [38,41] for the most recent version of such results and a discussion of the literature. Laws of scarcity in economies with clubs are examined in Cartwright, Conley and Wooders [16].

Some Remarks on Markets and More General Classes of Economies

Forms of the equivalence of outcomes of economies where individuals have concave utility functions but not necessarily linear in money. These include Billera [10], Billera and Bixby [11] and Mas-Colell [42]. A natural question is whether the results reported in this paper can extend to nontransferable utility games and economies where individuals have utility functions that are not necessarily liner in money. So far the results obtained are not entirely satisfactory. Nonemptiness of approximate cores of games with many players, however, holds in substantial generality; see Kovalenkov and Wooders [40] and Wooders [95].

Conclusions and Future Directions

The results of Shapley and Shubik [60], showing equivalence of structures, rather than equivalence of outcomes of solution concepts in a fixed structure (as in [4], for example) are remarkable. So far, this line of research has been relatively little explored. The results for games with many players have also not been fully explored, except for in the context of games, such as those derived from economies with clubs, and with utility functions that are linear in money.

Per capita boundedness seems to be about the mildest condition that one can impose on an economic structure and still have scarcity of per capita resources in economies with many participants. In economies with quasi-linear utilities (and here, I mean economies in a general sense, as in the glossary) satisfying per capita boundedness and where there are many substitutes for each type of participant, then as the number of participants grows, these economies resemble or (as if they) *are* market economies where individuals have continuous, and monotonic increasing utility functions. Large groups cannot influence outcomes away from outcomes in the core (and outcomes of free competition) since large groups are not significantly more effective than many small groups (from the equivalence, when each player has many close substitutes, between per capita boundedness and small group effectiveness).

But if there are not many substitutes for each participant, then, as we have seen, per capita boundedness allows small groups of participants to have large effects and free competition need not prevail (cores may be empty and price-taking equilibrium may not exist). The condition required to ensure free competition in economies with many participants, without assumptions of "thickness", is precisely small group effectiveness.

But the most complete results relating markets and games, outlined in this paper, deal with economies in which all participants have utility functions that are linear in money and in games with side payments, where the worth of a group can be divided in any way among the members of the group without any loss of total utility or worth. Nonemptiness of approximate cores of large games without side payments has been demonstrated; see Wooders [83,95] and Kovalenkov and Wooders [40]. Moreover, it has been shown that when side payments are 'limited' then approximate cores of games without side payments treat similar players similarly [39].

Results for *specific economic structures*, relating cores to price taking equilibrium treat can treat situations that are, in some respects, more general. A substantial body of literature shows that certain classes of club economies have nonempty cores and also investigates price-taking equilibrium in these situations. Fundamental results are provided by Gale and Shapley [25], Shapley and Shubik [61], and Crawford and Kelso [21] and many more recent papers. We refer the reader to Roth and Sotomayor [53] and to ► Two-Sided Matching Models, by Ömer and Sotomayor in this encyclopedia. A special feature of the models of these papers is that there are two sorts of players or two sides to the market; examples are (1) men and women, (2) workers and firms, (3) interns and hospitals and so on.

Going beyond two-sided markets to clubs in general, however, one observes that the positive results on nonemptiness of cores and existence of price-taking equilibria only holds under restrictive conditions. A number of recent contributions however, provide specific economic models for which, when there are many participants in the economy, as in exchange economies it holds that price-

taking equilibrium exists, cores are non-empty, and the set of outcomes of price-taking equilibrium are equivalent to the core. (see, for example, [1,2,24,85,92]).

Bibliography

1. Allouch N, Wooders M (2008) Price taking equilibrium in economies with multiple memberships in clubs and unbounded club sizes. J Econ Theor 140:246–278
2. Allouch N, Conley JP, Wooders M (2008) Anonymous price taking equilibrium in Tiebout economies with a continuum of agents: Existence and characterization. J Math Econ. doi:10.1016/j.jmateco.2008.06.003
3. Azrieli Y, Lehrer E (2007) Market games in large economies with a finite number of types. Econ Theor 31:327–342
4. Aumann RJ (1964) Markets with a continuum of traders. Econometrica 32:39–50
5. Aumann RJ (1987) Game theory. In: Eatwell J, Milgate M, Newman P (eds) The New Palgrave: A Dictionary of Economics. Palgrave Macmillan, Basingstoke
6. Aumann RJ, Dreze J (1974) Cooperative games with coalition structures. Int J Game Theory 3:217–37
7. Aumann RJ, Shapley S (1974) Values of Non-Atomic Games. Princeton University Press, Princeton
8. Bennett E, Wooders M (1979) Income distribution and firm formation. J Comp Econ 3:304–317. http://www.myrnawooders.com/
9. Bergstrom T, Varian HR (1985) When do market games have transferable utility? J Econ Theor 35(2):222–233
10. Billera LJ (1974) On games without side payments arising from a general class of markets. J Math Econ 1(2):129–139
11. Billera LJ, Bixby RE (1974) Market representations of n-person games. Bull Am Math Soc 80(3):522–526
12. Böhm V (1974) The core of an economy with production. Rev Econ Stud 41:429–436
13. Bondareva O (1963) Some applications of linear programming to the theory of cooperative games. Problemy kibernetiki 10 (in Russian, see English translation in Selected Russian papers in game theory 1959–1965, Princteon University Press, Princeton
14. Buchanan J (1965) An economic theory of clubs. Economica 33:1–14
15. Casella A, Feinstein JS (2002) Public goods in trade on the formation of markets and jurisdictions. Intern Econ Rev 43:437–462
16. Cartwright E, Conley J, Wooders M (2006) The Law of Demand in Tiebout Economies. In: Fischel WA (ed) The Tiebout Model at 50: Essays in Public Economics in honor of Wallace Oates. Lincoln Institute of Land Policy, Cambridge
17. Champsaur P (1975) Competition vs. cooperation. J Econ Theory 11:394–417
18. Cheng HC (1981) On dual regularity and value convergence theorems. J Math Econ 8:37–57
19. Conley J, Smith S (2005) Coalitions and clubs; Tiebout equilibrium in large economies. In: Demange G, Wooders M (eds) Group Formation in Economies; Networks, Clubs and Coalitions. Cambridge University Press, Cambridge
20. Conley JP, Wooders M (1995) Hedonic independence and taste-homogeneity of optimal jurisdictions in a Tiebout economy with crowding types. Ann D'Econ Stat 75/76:198–219
21. Crawford VP, Kelso AS (1982) Job matching, coalition formation, and gross substitutes. Econornetrica 50:1483–1504
22. Debreu G, Scarf H (1963) A limit theorem on the core of an economy. Int Econ Rev 4:235–246
23. Demange G (1994) Intermediate preferences and stable coalition structures. J Math Econ 1994:45–48
24. Ellickson B, Grodal B, Scotchmer S, Zame W (1999) Clubs and the market. Econometrica 67:1185–1218
25. Gale D, Shapley LS (1962) College admissions and the stability of marriage. Am Math Mon 69:9–15
26. Garratt R, Qin C-Z (1997) On a market for coalitions with indivisible agents and lotteries, J Econ Theor 77(1):81–101
27. Gillies DB (1953) Some theorems on *n*-person games. Ph.D Dissertation, Department of Mathematics. Princeton University, Princeton
28. Haimanko O, Le Breton M, Weber S (2004) Voluntary formation of communities for the provision of public projects. J Econ Theor 115:1–34
29. Hildenbrand W (1974) Core and Equilibria of a Large Economy. Princeton University Press, Princeton
30. Hurwicz L, Uzawa H (1977) Convexity of asymptotic average production possibility sets. In: Arrow KJ, Hurwicz L (eds) Studies in Resource Allocation Processes. Cambridge University Press, Cambridge
31. Kalai E, Zemel E (1982) Totally balanced games and games of flow. Math Oper Res 7:476–478
32. Kalai E, Zemel E (1982) Generalized network problems yielding totally balanced games. Oper Res 30:998–1008
33. Kaneko M, Wooders M (1982) Cores of partitioning games. Math Soc Sci 3:313–327
34. Kaneko M, Wooders M (2004) Utility theories in cooperative games. In: Handbook of Utility Theory vol 2, Chapter 19. Kluwer Academic Press, Dordrecht, pp 1065–1098
35. Kaneko M, Wooders M (1986) The core of a game with a continuum of players and finite coalitions; the model and some results. Math Soc Sci 12:105–137. http://www.myrnawooders.com/
36. Kannai Y (1972) Continuity properties of the core of a market. Econometrica 38:791–815
37. Konishi H, Le Breton M, Weber S (1998) Equilibrium in a finite local public goods economy. J Econ Theory 79:224–244
38. Kovalenkov A, Wooders M (2005) A law of scarcity for games. Econ Theor 26:383–396
39. Kovalenkov A, Wooders M (2001) Epsilon cores of games with limited side payments: nonemptiness and equal treatment. Games Econ Behav 36(2):193–218
40. Kovalenkov A, Wooders M (2003) Approximate cores of games and economies with clubs. J Econ Theory 110:87–120
41. Kovalenkov A, Wooders M (2006) Comparative statics and laws of scarcity for games. In Aliprantis CD, Matzkin RL, McFadden DL, Moore JC, Yannelis NC (eds) Rationality and Equilibrium: A Symposium in Honour of Marcel K. Richter. Studies in Economic Theory Series 26. Springer, Berlin, pp 141–169
42. Mas-Colell A (1975) A further result on the representation of games by markets. J Econ Theor 10(1):117–122
43. Mas-Colell A (1977) Indivisible commodities and general equilibrium theory. J Econ Theory 16(2):443–456
44. Mas-Colell A (1979) Competitive and value allocations of large exchange economies. J Econ Theor 14:307–310
45. Mas-Colell A (1980) Efficiency and decentralization in the pure theory of public goods. Q J Econ 94:625–641

46. Mas-Colell A (1985) The Theory of General Economic Equilibrium. Economic Society Publication No. 9. Cambridge University Press, Cambridge
47. Moulin M (1988) Axioms of Cooperative Decision Making. Econometric Society Monograph No. 15. Cambridge Press, Cambridge
48. Moulin H (1992) Axiomatic cost and surplus sharing. In: Arrow K, Sen AK, Suzumura K (eds) Handbook of Social Choice and Welfare, 1st edn, vol 1, chap 6. Elsevier, Amsterdam, pp 289–357
49. von Neumann J, Morgenstern O (1953) Theory of Games and Economic Behavior. Princeton University Press, Princeton
50. Owen G (1975) On the core of linear production games. Math Program 9:358–370
51. Pauly M (1970) Cores and clubs. Public Choice 9:53–65
52. Qin C-Z, Shapley LS, Shimomura K-I (2006) The Walras core of an economy and its limit theorem. J Math Econ 42(2):180–197
53. Roth A, Sotomayer M (1990) Two-sided Matching; A Study in Game-theoretic Modeling and Analysis. Cambridge University Press, Cambridge
54. Scotchmer S, Wooders M (1988) Monotonicity in games that exhaust gains to scale. IMSSS Technical Report No. 525, Stanford University
55. Shapley LS (1964) Values of large games -VII: A general exchange economy with money. Rand Memorandum RM-4248-PR
56. Shapley LS (1967) On balanced sets and cores. Nav Res Logist Q 9:45–48
57. Shapley LS (1952) Notes on the N-Person game III: Some variants of the von-Neumann-Morgenstern definition of solution. Rand Corporation research memorandum, RM-817:1952
58. Shapley LS, Shubik M (1960) On the core of an economic system with externalities. Am Econ Rev 59:678–684
59. Shapley LS, Shubik M (1966) Quasi-cores in a monetary economy with nonconvex preferences. Econometrica 34:805–827
60. Shapley LS, Shubik M (1969) On market games. J Econ Theor 1:9–25
61. Shapley LS, Shubik M (1972) The Assignment Game 1; The core. Int J Game Theor 1:11–30
62. Shapley LS, Shubik M (1975) Competitive outcomes in the cores of market games. Int J Game Theor 4:229–237
63. Shapley LS, Shubik M (1977) Trade using one commodity as a means of payment. J Political Econ 85:937–68
64. Shubik M (1959) Edgeworth market games. In: Luce FR, Tucker AW (eds) Contributions to the Theory of Games IV, Annals of Mathematical Studies 40. Princeton University Press, Princeton, pp 267–278
65. Shubik M, Wooders M (1982) Clubs, markets, and near-market games. In: Wooders M (ed) Topics in Game Theory and Mathematical Economics: Essays in Honor of Robert J Aumann. Field Institute Communication Volume, American Mathematical Society, originally Near Markets and Market Games, Cowles Foundation, Discussion Paper No. 657
66. Shubik M, Wooders M (1983) Approximate cores of replica games and economies: Part II Set-up costs and firm formation in coalition production economies. Math Soc Sci 6:285–306
67. Shubik M (1959) Edgeworth market games. In: Luce FR, Tucker AW (eds) Contributions to the Theory of Games IV, Annals of Mathematical Studies 40, Princeton University Press, Princeton, pp 267–278
68. Shubik M, Wooders M (1982) Near markets and market games. Cowles Foundation Discussion Paper No. 657, on line at http://www.myrnawooders.com/
69. Shubik M, Wooders M (1982) Clubs, markets, and near-market games. In: Wooders M (ed) Topics in Game Theory and Mathematical Economics: Essays in Honor of Robert J Aumann. Field Institute Communication Volume, American Mathematical Society, originally Near Markets and Market Games, Cowles Foundation, Discussion Paper No. 657
70. Shubik M, Wooders M (1983) Approximate cores of replica games and economies: Part I Replica games, externalities, and approximate cores. Math Soc Sci 6:27–48
71. Shubik M, Wooders M (1983) Approximate cores of replica games and economies: Part II Set-up costs and firm formation in coalition production economies. Math Soc Sci 6:285–306
72. Shubik M, Wooders M (1986) Near-markets and market-games. Econ Stud Q 37:289–299
73. Sondermann D (1974) Economics of scale and equilibria in coalition production economies. J Econ Theor 8:259–291
74. Sun N, Trockel W, Yang Z (2008) Competitive outcomes and endogenous coalition formation in an n-person game. J Math Econ 44:853–860
75. Tauman Y (1987) The Aumann–Shapley prices: A survey. In: Roth A (ed) The Shapley Value: Essays in Honor of Lloyd S Shapley. Cambridge University, Cambridge
76. Tauman Y, Urbano A, Watanabe J (1997) A model of multiproduct price competition. J Econ Theor 77:377–401
77. Tiebout C (1956) A pure theory of local expenditures. J Political Econ 64:416–424
78. Weber S (1979) On ε-cores of balanced games. Int J Game Theor 8:241–250
79. Weber S (1981) Some results on the weak core of a nonsidepayment game with infinitely many players. J Math Econ 8:101–111
80. Winter E, Wooders M (1990) On large games with bounded essential coalition sizes. University of Bonn Sondeforschungsbereich 303 Discussion Paper B-149. on-line at http://www.myrnawooders.com/ Intern J Econ Theor (2008) 4:191–206
81. Wooders M (1977) Properties of quasi-cores and quasi-equilibria in coalition economies. SUNY-Stony Brook Department of Economics Working Paper No. 184, revised (1979) as A characterization of approximate equilibria and cores in a class of coalition economies. State University of New York Stony Brook Economics Department. http://www.myrnawooders.com/
82. Wooders M (1978) Equilibria, the core, and jurisdiction structures in economies with a local public good. J Econ Theor 18:328–348
83. Wooders M (1983) The epsilon core of a large replica game. J Math Econ 11:277–300, on-line at http://www.myrnawooders.com/
84. Wooders M (1988) Large games are market games 1. Large finite games. C.O.R.E. Discussion Paper No. 8842 http://www.myrnawooders.com/
85. Wooders M (1989) A Tiebout Theorem. Math Soc Sci 18:33–55
86. Wooders M (1991) On large games and competitive markets 1: Theory. University of Bonn Sonderforschungsbereich 303 Discussion Paper No. (B-195, Revised August 1992). http://www.myrnawooders.com/
87. Wooders M (1991) The efficaciousness of small groups and the approximate core property in games without side payments. University of Bonn Sonderforschungsbereich 303 Discussion Paper No. B-179. http://www.myrnawooders.com/

88. Wooders M (1992) Inessentiality of large groups and the approximate core property; An equivalence theorem. Econ Theor 2:129–147
89. Wooders M (1992) Large games and economies with effective small groups. University of Bonn Sonderforschingsbereich 303 Discussion Paper No. B-215.(revised) in Game-Theoretic Methods in General Equilibrum Analysis. (eds) Mertens J-F, Sorin S, Kluwer, Dordrecht. http://www.myrnawooders.com/
90. Wooders M (1993) The attribute core, core convergence, and small group effectiveness; The effects of property rights assignments on the attribute core. University of Toronto Working Paper No. 9304
91. Wooders M (1994) Equivalence of games and markets. Econometrica 62:1141–1160. http://www.myrnawooders.com/n
92. Wooders M (1997) Equivalence of Lindahl equilibria with participation prices and the core. Econ Theor 9:113–127
93. Wooders M (2007) Core convergence in market games and club economics. Rev Econ Design (to appear)
94. Wooders M (2008) Small group effectiveness, per capita boundedness and nonemptiness of approximate cores. J Math Econ 44:888–906
95. Wooders M (2008) Games with many players and abstract economies permitting differentiated commodities, clubs, and public goods (submitted)
96. Wooders M, Zame WR (1987) Large games; Fair and stable outcomes. J Econ Theor 42:59–93
97. Zajac E (1972) Some preliminary thoughts on subsidization. presented at the Conference on Telecommunications Research, Washington

Marketing: Complexity Modeling, Theory and Applications in

Jacob Goldenberg[1], Daniel Shapira[2]
[1] School of Business Administration, Hebrew University, Jerusalem, Israel
[2] Department of Business Administration, Guilford Glazer School of Business and Management, Ben–Gurion University, Beer Sheva, Israel

Article Outline

Glossary
Definition of the Subject
Introduction
Markets Are Complex
Innovation Growth Processes in Marketing
Blazing Saddles – a Simple Case for Complexity Modeling
The Complexity Approach in the Social Sciences – the Basic Idea
Running the ABM Algorithm
Back to the Saddle
Agent Based Modeling and Aggregate Diffusion Models
The Lost Dimension: Using Spatial Analysis to Predicting New Product Success
A Universal Framework for Modeling the Market Penetration Dynamics
Future Directions
Bibliography

Glossary

All terms in this glossary relate mainly to the case of Innovation.

Aggregate models Models that use market level data of adoption, with less if any emphasis on individuals' data.

Consumer behavior Consumers make decisions and choices according to preferences, personalities but mainly according to some heuristics and rules. In Marketing, the field that investigates this behavior is defined consumer behavior.

Diffusion of innovation A research field in marketing that deals with the dissemination of a new product in the marketplace.

Early adopters Similar to the innovators, they like innovations and they also not risk averse. Contrary to innovators they are interested in product advantages and new benefits. on average they are estimated as 13.5%.

Early and late majority Together they are estimated as 66%. they adopt the dominant designs, products that are compatible, bugs free, reliable, user friendly, and after the price has been stabled on a reasonable level.

Innovation A new product or service that provides new benefits, typically by new product or process or technology. Innovation can be Radical (extremely new) or incremental (moderately new).

Innovators First adopters, typically consist of 2.5% of the population, interested mainly in the technology and new features, and less interested in the advantages. They are not risk averse, and they are not concerned from products with bugs.

Laggards Individuals who do not like to adopt innovation and prefer to avoid it if possible (estimated as 16% of the population).

Marketing efforts A firm's activity to increase awareness of the new product and the propensity to try it.

Penetration and product life cycle Product life cycle is a general model that describes the "life" of a product by flat growth of sales followed by a take off in which sales are increased exponentially until a peak of sales and decline until the product "dies". Product life cycle can be examined by sales or by units sold. penetration is measured by counting the first adoption on each individual.

Saddle A dual peak penetration pattern that is characterized by an initial peak of sales, followed by a slump, and then a recovery until a second (much larger) peak is obtained.

Word of mouth (w-o-m) Interaction between consumers in which information and recommendation is passed. There are other paths of information transfer that fall under this classification as well like imitation in which no conversation is needed.

Definition of the Subject

Adoption of innovation is one of the most interesting fields in marketing, with high monetary importance as more than 50% of firms profits are influenced from new products performances. Innovations are also an important factor for society as it influences many aspects in life.

Because adoption of an innovation involves risks, learning periods, and sometimes adaptations, individuals tend to consult with their friends and peers before making an adoption decision. The interactions between large numbers of consumers is one of the trademarks of complex system. Innovation markets are complex systems, and the complexity approach can become an important tool, both scientifically and practically.

Introduction

Having been established as a scientific field only as recently as the late 1960s, marketing can be considered a young discipline in the social sciences. Nevertheless in a relatively short time, this field had rapidly matured into a fascinating, highly interdisciplinary field with its own rigorous foundations. The reason why the complexity approach, which emerged so many years ago, is now only in its initial stages in the field of marketing, might be because it is a wide interdisciplinary topic, without too many years of heritage.

In this chapter we cover some of the benefits of complexity approaches for improving our understanding of marketing systems. We point out real-world applications that can be derived from this approach, and demonstrate that marketing also offers a unique contribution to complexity paradigm development.

Broadly speaking, research in marketing has traditionally followed one of two main approaches: the first is a behavioral one (micro perspective), and the second is more aggregate level (macro perspective). In the behavioral approach, the unit of analysis is the individual, or more precisely, the consumer. Studies adopting this approach employ psychological research methods to study choices, decision making processes, and other forms of consumer behavior. Aggregate outcomes are less interesting to researchers conducting this type of research. In the second approach, the unit of analysis is a market – a collection of individuals whose behavior is aggregated. This group of studies generally employs research methods that are similar to economic and other mathematical models. The individual consumer is typically ignored.

Naturally this is a very coarse classification. But it is interesting to see that scholars in both groups participate in conferences that focus on issues that belong to "their" group, and seldom collaborate with scholars from the second group. Papers that include both focuses are rare. It seems that the tools are part of the definition and selection of the research problems to face.

A complexity approach requires knowledge from both sub-fields of marketing, in order to combine individual and aggregate levels relatively easier, and offer innovative insights that may be otherwise overlooked.

Yet even preceding the issue of *feasibility* of applying complexity models in marketing is the question of the justification for such an approach. Are markets complex systems? Fortunately for researchers who constantly seek for new adventures, the answer is an absolute "yes". Consumers interact – they talk to each other, they make recommendations to each other, they express their opinions (both positive and negative), and they influence the people in their social networks. In fact, various studies have shown that interactions between consumers can sometimes have an effect that is 10 times stronger than marketing efforts such as advertising or promotions. Interactions between consumers are sometimes termed "word of mouth (w-o-m)" or "the internal force", but we believe that they should more aptly be viewed as "magma forces": they are invisible, they act below the surface, but they are extremely powerful.

Complexity in marketing deals with understanding these invisible forces.

For the sake of convenience, in this chapter we focus on Innovation Adoption (also labeled as Diffusion of Innovation, Growth Processes). In our view, this is a rich and fascinating phenomenon, and one in which the benefits of complexity modeling have been demonstrated.

Markets Are Complex

The environment in which firms operate fits the definition of a complex system quite well: Customers, employees, partners, suppliers, and other stakeholders interact with each other, exchange information, and adapt their behavior in response to actions by the firm and other network peers. This network of individuals has features that are

similar to those of other complex systems, such as colonies of bacteria, flocks of birds, genes, or neural networks.

As in other complex systems, a set of relatively simple individual interactions can result in surprising (even counter intuitive), non-linear (and sometime unstable) dynamics. As a result, marketing managers find it difficult to predict the consequences of changes in a firm's environment or strategy, such as introducing a new product, creating a lateral alliance, or changing the way service personnel operate and interact with the firm's customers. Unfortunately, conventional decision-making tools such as analytical models – often built on static assumptions to preserve their parsimony – are less relevant where players in the marketing exchange arena constantly interact with each other and adapt their behavior. While managers can often measure and predict behavior and information on the individual level, how the multi-player mix ultimately transforms into the aggregate firm or market level largely remains a mystery. Take as an example the phenomenon of consumer word-of-mouth. Given its centrality in adoption decision-making [16,17], one would expect to find abundant marketing literature that explicitly models this process with the aim of helping managers understand the role of word of mouth in market growth and evolution. However, unlike other types of marketing communications, such as advertising, sales promotion, or sales force management, where significant attention has been given to assessing aggregate effects on sales, little is known about how word-of-mouth aggregates to impact sales levels. Papers that uncover the determinants of word of mouth have only recently appeared in marketing literature. Similarly, while efforts are made to understand inter-organizational information exchange among employees – for example, information acquisition and utilization in new product alliances [15] – marketers have yet to better understand how individual- or network-level behavior aggregates to the firm or market level.

Innovation adoption is, in our view, one of the most interesting cases in marketing, with a tremendous impact on many social aspects, and one which demonstrates how general complexity modeling can be utilized to improve our understanding of and ability to predict how markets work.

Innovation Growth Processes in Marketing

Diffusion of Innovation

Much of the knowledge that has accumulated in recent decades on new product adoption is of a practical, prescriptive nature. One such maxim in marketing, supported by PDMA survey findings [18], is that a new product which offers a solution to a problem has a greater chance of success than one which offers a superior marketing mix that resolves no problem.

Despite the cumulative wealth of marketing experience available to manufacturers and marketers of new products, the rate of new product failures is distressingly high [7], even discounting innovations which fail to satisfy any genuine consumer need. When we consider the cumulative wealth of marketing experience available to manufacturers and marketers of new products, this fact alone is sufficient to undermine confidence in our own marketing savvy.

Actually, we are probably making fairly sophisticated use of the knowledge we have – it is the knowledge that we have yet to acquire that is working against our best efforts. New findings on how consumers communicate marketing information to each other may yield a wealth of valuable information for marketing professionals hoping to gain acceptance for new innovations they introduce to the market. Recent developments now allow us to investigate as yet unexplored dimensions of the phenomenon of word-of-mouth, whose significance in the adoption of new products has been recognized for several decades. Qualitatively new findings on this fascinating and hitherto elusive marketing factor, the inherent complexity of the factors involved, as well as the changes in their relative significance, may hold the key to our understanding of the new product development process and, more specifically, of why so many new products fail.

To highlight the unique contribution of complexity modeling for w-o-m research, it is necessary to review the current state of our knowledge on this topic. We can trace the origin of this concept to several sources, including the classic formulation of the product growth curve. This model of economical growth, embraced with enthusiasm by theoreticians and marketers alike, is based on the work of economist and demographer Thomas Robert Malthus, who formulated the first equation describing the dynamics of auto-catalytically proliferating individuals in 1798. A correction to Malthus' equation, offered by P.F. Verhuulst in 1838, offered a more realistic growth pattern with saturation terms. The models used in marketing today are almost identical to Verhuulst's logistic equation.

Two social sciences researchers took up this basic notion to establish a broad framework for the process of innovation adoption. In 1962, Everett Rogers published the first edition of his book "Diffusion of Innovation". Through case study analysis, Rogers uncovered the fundamental sociological structure describing how adopters of new products interact with each other; influence each other's adoption and rejection decisions, and ultimately affect the diffusion of innovations in a market. Several

years later, Frank Bass [2] developed a quantitative model of diffusion, based on a model taken from the field of epidemiology. Extending the biological analogy, Rogers established the critical role of interpersonal contact in the diffusion of new products as one of the most basic assumptions of his model. In this model, consumers are either infected (i. e., adopt the new product) spontaneously or through contact with other consumers (previous purchasers of the product). He assumed, for example, that "infection" through contact is more prevalent than spontaneous infection (as represented by effects such as advertising and marketing efforts).

Together, the premises of Bass and Rogers created the foundation for the product growth curve (visually similar to the product life cycle and also known as the adoption curve or the diffusion curve). The product growth curve represents the introduction of an innovation into a market as a progression along an inevitably successful path, in which early adopters stimulate subsequent product acceptance by the main market. The pivotal role of early adopters is based on their basic traits and social roles, bolstered by a host of marketing activities, including advertising. Indeed, early adopters are conceptualized on the basis of their unique contribution to other population groups in the form of product-related information. The information that early adopters communicate – either in the form of data injected into the market by marketers and actively sought or passively received by consumers; or as information passed directly or indirectly from one consumer to another – reduces uncertainty in subsequent consumer decision making. In addition to their ability to reduce the perceived risks associated with an innovation by communicating product-related information, early adopters also instigate a process of social emulation and provide the basis for social legitimization.

Rogers's book triggered huge interest in the social aspects of innovation adoption, and Bass' paper demonstrated the first steps toward a model of this phenomenon. Both Everett Rogers and Frank Bass lived long enough to see the impact of their work on the field of innovation adoption and its evolution as a strong research discipline. In fact, the Bass model [2] remains the best-known and most widely used model in diffusion research today. Since its publication in *Management Science*, it inspired many other models and it is considered a bench mark with which new models are judged.

The Bass Model

Assume a market with potential M where $N(t)$ is the cumulative number of adopters and $n(t)$ is the adoption rate at time t. At each point in time, new adopters join the market as a result of external and international influences. *External influences* are the activities of firms in the market, mainly advertising, and they are represented by a constant P. *Internal influences*, which are interactions between consumers (such as word of mouth), are represented by a constant Q. In the original paper by Bass, as well as later interpretations that regarded diffusion as a theory of communications, Q represented word-of-mouth communications. According to current interpretations, the internal coefficient Q. represents consumer interdependencies.

The Bass model postulates that the hazard function of the adoption process, which represents the percentage of new adopters of all potential adopters, is determined by the sum of these two influences. Specifically:

$$\frac{n(t)}{M - N(t)} = P + Q\left(\frac{N(t)}{M}\right) , \tag{1}$$

where the total internal influence (i. e. the total word-of-mouth effect) is proportional to the fraction of the market potential that already adopted the new product. Therefore, setting the problem to the continuous limit, the number of new adopters at time t can be described by the following differential equation:

$$\begin{aligned}\frac{\mathrm{d}N(t)}{\mathrm{d}t} &= \big(M - N(t)\big)\left(p + \frac{Q}{M}N(t)\right)\\ &= MP + (Q - P)N(t) - \frac{Q}{M}\big(N(t)\big)^2 .\end{aligned} \tag{2}$$

Equation (2) is a non-linear first-order differential equation, and can be solved analytically. Given the initial condition $N(0) = 0$ (so that the time $t = 0$ is the launching time of the new product), the solution for Eq. (1) takes the form

$$N(t) = M\frac{1 - \mathrm{e}^{-(P+Q)t}}{1 + \left(\frac{Q}{P}\right)\mathrm{e}^{-(P+Q)t}} . \tag{3}$$

This equation reflects time-dependent growth governed by 3 parameters P, Q, and M. Its general shape is a regular S shape (logistic) graph.

The Bass model parameters P, Q, and M can be estimated from adoption data, usually by using non-linear least squares [21]. Numerous studies have estimated these parameters in various industries, and found that the average values of P and, Q for durable goods were $P = 0.03$, and $Q = 0.38$ (Sultan, Farley, & Lehmann 1990). Figure 1 presents real data of VCR penetration along with a Bass curve that fits the data.

The diffusion models are similar to the Bass models, and contribution to our understanding of the marketing forces that govern innovation adoption and growth

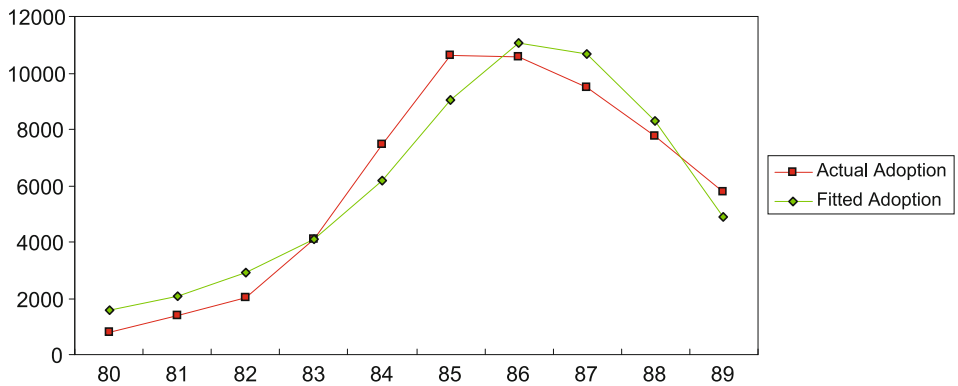

Marketing: Complexity Modeling, Theory and Applications in, Figure 1
Estimated Bass model from penetration data of VCR in US

of markets, and, at the same time, they enable forecasting of penetration in forthcoming periods by estimating the model parameter based on initial data. These two advantages were probably the main reasons behind the widespread appeal of diffusion models. However, this aggregate approach involves several problems, including:

1. Need for adequate database for forecasting: Even the Bass model, which is the simplest diffusion model, requires a minimum periods to estimate its three parameters. This implies that reliable predictions can be performed only after several years of data have been collected, which is sometimes too late for an innovation marketer. One approach that circumvents this requirement is the use of higher resolution data by switching to quarterly or even monthly data. However, the volatility of the data also causes estimation problems (we will return to this issue later).
2. Limited local fit: At the beginning of the growth process when the number of actual adopters is much smaller than the total market potential (i. e. $N(t) \ll M$), the Bass model is reduced to a simple growth process of the following form:

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} \approx MP + (Q - P)N(t) \,.$$

   Since the market potential M cannot be separated from the product MP, predictions that are based on curve fitting at the early stages of the adoption process become inaccurate and even erroneous. On the other hand, in subsequent stages of the growth process, the market is closer to exhausting its potential, and the actual number of adopters $N(t)$ is on the same scale as total market potential M. Consequently, the quadratic term $Q\left(\frac{N(t)}{M}\right)N(t)$ which appears in the Bass model is no longer negligible and fitting and forecasting are quite satisfactory.
3. Simplistic model assumptions: As the most simple of all diffusion models, the Bass model is based on the most naive assumptions, as a result of which it is quite synthetic. For example, the Bass model assumes homogeneity, implying that all the individuals are equally affected by marketing efforts and peer recommendations, and that all social ties are of the same intensity. In fact, according to the Bass model, everyone meets and talks with everyone. Bass parameters P, Q, and M do not change over time, and social network, according to these assumption, do not exist. This world is too simplistic – social life is much richer and more complex. Even more recent models based on aggregate modeling have succeeded in capturing only part of this complexity.

The original model captured a truth and was admittedly appealing in its simplicity. While related theories matured, significant deviations from the basic model were revealed. It has become increasingly obvious that the growth of contemporary markets and products is driven by a host of intricately interwoven factors, which the classic adoption curve fails to reflect, and therefore fails to predict.

Today we have access to information on advertising investments, the number of coupons distributed, and the number of sales promotion personnel hired. We are able to measure the inputs and outputs of various elements in our distribution channel and even evaluate the long-term contribution of each one. However, since we lack understanding of the weight consumers attribute to information, recommendations or warnings they receive from friends and acquaintances in their purchase decisions, the effect of w-o-m tends to catch us unprepared. Emmanuel Rosen [17]

offered some surprising data confirming the strong impact of word-of-mouth on sales. For example, 65% of all PDA purchasers reported learning about their hand-held organizer from their friends and 70% of all Americans select their physician based on personal recommendations.

It is not surprising that the implications of this factor have remained unnoticed for so long. The unique feature of word of mouth (w-o-m) is the almost surreptitious nature of the process it generates beneath the surface of the market and marketing data. The word of mouth effect becomes observable only when its ultimate results become accessible for analysis, in the form of sales data. The process of transmitting and communicating information from individual to individual and the manner in which individual consumers are differentially affected remain a mystery to market researchers.

One of the causes of this gap in knowledge is the underlying complexity of the w-o-m process. W-o-m develops in a social system that may be described as "an adaptive complex system", i. e., a system that consists of a large number of individual entities which interact with each other (in a manner that is sometimes indiscernible), ultimately generating collective (macro) behavior [23]. Although the interactions in many adaptive systems may in themselves be simple, the magnitude of the system's scale admits the emergence of patterns which are hard to predict, hard to track empirically, and often almost impossible to model analytically.

Complexity has emerged as a distinct field of research in recent decades to address a wide range of phenomena occurring in such systems. Originated in physics and biology, the study of complexity has recently moved into the social sciences [10], including economics [19], management [1], and marketing [8]. Let us explore the benefits of this approach for marketing modeling.

We begin with the *chasm* and *saddle* – a simple case of complexity in which two sub-markets interact with each other phenomenon.

Blazing Saddles – a Simple Case for Complexity Modeling

The Dual Market phenomenon is a simple case that neatly illustrates the significant role that word-of-mouth plays in the adoption process. The Dual Market model is offered as a recent extension of the classic adoption curve, purporting to reflect a more accurate representation of new product development by attributing the high failure rate of new products to communications (or the lack thereof) between individual consumers. The model, also known as Moore's Chasm Theory [13], is supported by a wealth of anecdotal evidence. Ascribing a large number of new product failures to the marketers' disregard of an important twist on the classic adoption theory, Moore cautions marketers to recognize a communication barrier between early adopters and the main market, which results in the formation of two distinct markets rather a single, general market in which consumers are classified by "adopter category". The success of a new product critically depends on adoption rates in *both* these markets. According to Moore, marketing failure is avoided by the realization that w-o-m communication among consumers cannot not be taken for granted as inevitably sweeping the adoption curve from product introduction to take-off and success.

This intuitively appealing idea has been confirmed by empirical evidence in [8]. It is important to look at sales data and see whether the chasm (defined in the model as an expression of the lack of communications between the two sub-markets) is just another factor contributing to market failure or an immanent mechanism. Indeed, when sales data for several innovative products were tracked over time, it became clear that new product adoption takes place in two semi-consecutive loci – two different consumer markets, with distinct consumer attributes. Specifically, using an information bank containing data on a large number of innovative products in the consumer electronics industry, this study found that approximately one-third of the growth patterns of these new products involved an initial peak, giving rise to a trough of sufficient depth and duration to preclude random fluctuations, followed by sales which eventually exceeded the initial peak – a pattern which is termed a 'saddle'. When the inherent differences in the reception of new products by these two markets are sufficiently large, a lag occurs between the adoption patterns of the early market and the main market, creating two distinct sales peaks, rather than the single, classic Bass diffusion pattern.

Figure 2 is a graphic illustration of the relation of the dual market model to the saddle phenomenon. Figure 3 presents a real case of the pc sales to consumers

From a managerial point of view, this phenomenon warrants attention because a significant and unexpected decline in sales in the relatively early stages of a product's life cycle may erroneously cast doubt on product viability. Thus, identifying the conditions underlying the occurrence of a saddle may prevent premature withdrawal of potentially successful new products. This is especially true for high-tech and similar innovative products, since firms typically continue R&D and product improvements after market launch, increasing early sales fluctuations. Evidence of a decline in sales leads to a sudden, unexpected drop in the cash flow just when firms are in the invest-

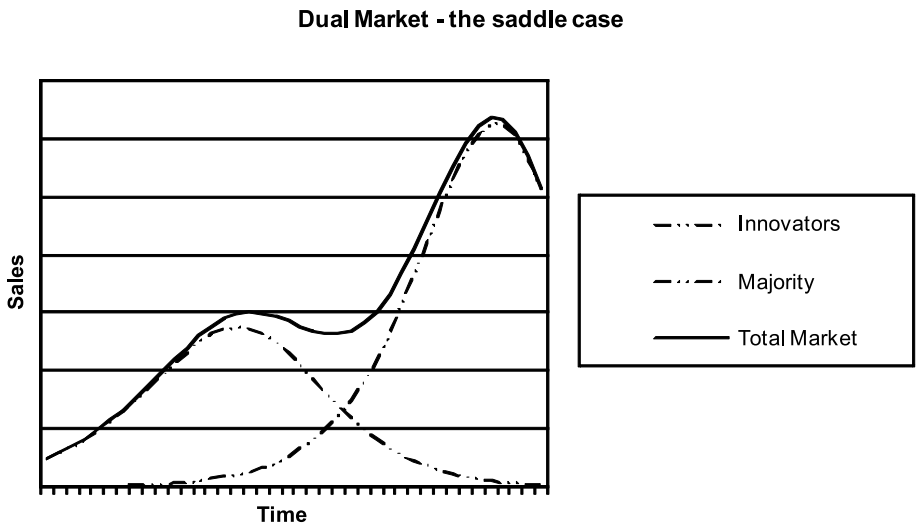

Marketing: Complexity Modeling, Theory and Applications in, Figure 2
The dual market underlying the Saddle mechanism

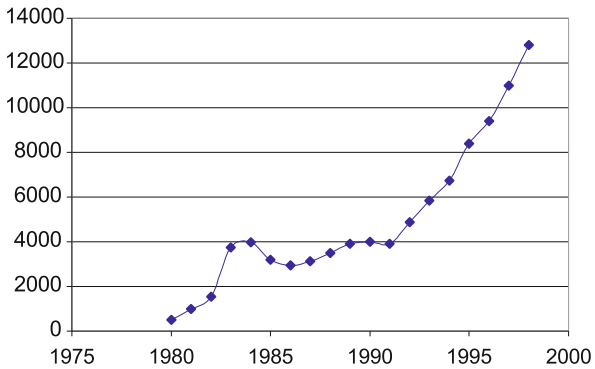

Marketing: Complexity Modeling, Theory and Applications in, Figure 3
Saddle in personal computers penetration

ment-intensive stage of simultaneously launching and improving the new product. The saddle's contribution in informing managerial decision-making at such a crucial stage of the new product launch is magnified if we take into account that saddles may be even more ubiquitous than our statistics show. Using data on products that ultimately survived the drop in sales and endured, a saddle pattern was evident in approximately one third of the cases studied. Is a traditional marketing approach insufficient to reveal this phenomenon? Muller and Yogev [13] followed by Van den Bulte and Joshi [22] modeled the same saddle phenomenon using aggregate modeling and combined close form solutions while Goldenberg, Libai and Muller [8] used a complexity approach. Although in this case model performances in terms of predictions and understanding the phenomenon are similar, in our view, both approaches are required, both have advantages, and together they allow more understanding and managerial advantages.

However let us remember that this was a simple case which involves only two subsystems that interact with each other is a very simple way. The chasm, and its consequence, the saddle are not a genuine complex system. However this phenomenon was one of the first cases that was addressed through complexity modeling tools, and it conveniently illustrates how complexity modeling can be applied to marketing systems. This framework is extended in the following sections.

The Complexity Approach in the Social Sciences – the Basic Idea

Complexity models have more names than principles. One of the first applications of complexity modeling in social sciences was coined *cellular automata*. The conception of *cellular automata* is typically attributed to John von Neumann as a formal model of a self-reproducing biological organism. The history of the use of cellular automata in a variety of disciplines has been well documented (e. g., [20]) and will not be repeated here (Cellular automata are models of computation that can generate complex aggregate behaviors using a limited number of simple individual-level rules. Cellular automata were publicly recognized when proposed by John von Neumann as formal models of self-reproducing organisms. However, the first attempt to understand complex system behavior can be traced to the Ising model, first proposed in 1924 by

Ernst Ising. Despite its deceptively simple appearance, this consequently well-established model explains and predicts deterministic phenomena in nature. The Ising model attempts to imitate a process in which individual elements (e. g., atoms, animals, protein folds, biological membranes, and social elements) modify their behavior in response to the behavior of other individuals in their vicinity).

Recently *Agent based modeling* (ABM) was developed as a more general framework for complexity applications. Basically one can think of ABM as a grid of cells. In its simplest state this is a one-dimensional grid, but most applications, including the ones we present here, are two-dimensional matrices. In the ABM environment, time is discrete and at each point in time, a cell can assume one of a finite number of possible states. A cell can change its state each period in response to (or, as a function of) the state of its surrounding cells. The algorithm by which cells change their state is usually called "local rule" or "transition rule". The collection of all states at a given point in time is called the "global state". In each period, application of the local rule of the ABM to a cell changes the global state of the matrix. In the simple version of ABM, local rules are deterministic: A global state determines the next global state with certainty. However, one can also use *stochastic* ABM, in which the state of the cells changes based on some probability function, which is also a function of the state of the cells around it.

Consider the example of the diffusion of innovations, which we will use in this chapter to demonstrate the uses of ABM for marketing applications. Diffusion theory as well as most diffusion modeling efforts in marketing, suggest that the process in which a social system adopts an innovation is largely based on interactions among potential adopters. The transition from a potential adopter to an adopter is attributed to two information sources: *External* sources are *unrelated* to the number of previous adopters, and include advertising, sales force, and other marketing efforts, as well as mass media. *Internal* sources, on the other hand, are the previous adopters of the innovation who can influence potential adopters by spreading word of mouth and functioning as role models of imitation.

Using the basic principles of ABM it is easily shown how the combination of external and internal sources creates an aggregate adoption pattern in a given social system. We define an innovation diffusion ABM model consisting of three components:

1. A **matrix** representing adoption by individual consumers
2. **Relationships** among the individuals
3. **Transition rules** representing adoption probabilities

The matrix is a two-dimensional array of cells. Each cell, representing a potential consumer, can accept one of two states: "0", representing a potential consumer who has *not* adopted the innovation, and "1", representing a consumer who *has* adopted the new product. In addition, irreversibility of transition is assumed, so that consumers cannot "un-adopt" after they have adopted.

The model assumes that potential consumers interact with all individuals in their immediate "neighborhood". Hence in the above matrix, any individual who is not on the edge of the matrix has eight neighbors. While this seems an obvious assumption, other kinds of neighborhoods have been defined (e. g., [20]).

Transition rules define the conditions needed to convert individuals from state "0" to state "1" as a result of information. Transition rules involve two types of information sources:

1. *External Factors*: A generic transition rule of this type is: Some probability p exists, such that in a given time period, external influence mechanisms (such as advertising or mass media) will cause the individual to adopt the innovative product.
2. *Internal Factors*: A generic transition rule of this type is: Some probability q exists, that during a given time period, an individual will be affected by an interaction with *a single* other individual in its immediate neighborhood who has already adopted the product.

Running the ABM Algorithm

The following step-by-step outline describes how the ABM algorithm is applied:

Period 0: This is the initial condition, wherein no individual has yet adopted the product (all cells have a value of 0).

Period 1: The probability functions prob(t) are applied to each consumer. Obviously, in this period, advertising is the only source of information available because, by definition, word of mouth needs consumers who have already adopted the product to initiate the process. Hence the probability of adoption in this term is p. A random number U is drawn from a uniform distribution in the range [0, 1], representing the probability of individuals to become influenced by advertising. If $U < \text{prob}(t)$, the consumer moves from non-adopter to adopter state (receiving the value of 1). Otherwise the consumer remains a non-adopter.

Period 2: The individuals who have adopted the product initiate the word-of-mouth process by interacting and communicating with other consumers. Probabili-

Period 0: Before Launch Time

0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Period 1: At Introduction

0	0	0	0	0	0	0	1
0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0
0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0

Period 15: Toward Peak in Sales

0	0	0	0	0	0	1	1
0	0	1	0	0	0	1	1
0	1	1	0	1	1	1	1
0	1	1	1	0	1	1	1
0	0	0	1	0	1	0	1
0	1	0	0	0	1	1	0
0	0	1	1	0	0	0	0
1	0	0	1	0	0	1	0
0	0	0	1	0	0	0	0

Period 25: At Maturity

1	1	0	1	1	1	1	1
1	1	1	1	1	0	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1
0	1	1	1	1	1	0	1
1	1	0	1	1	1	1	1
1	0	1	1	1	1	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1

Marketing: Complexity Modeling, Theory and Applications in, Figure 4
Illustration of the ABM process

ties are realized as in step 1, and the random number is drawn so that when $U < \text{prob}(t)$, the consumer moves from non-adopter to adopter.

Period *n*: This process is repeated until a certain percentage (e. g., 95%) of the total market turns into adopters.

It is easy to see that in this basic diffusion ABM model, the time-dependent (non-cumulative) individual probability of adoption, prob(t), given that the individual has not yet adopted, is based on the following binomial formula:

$$\text{prob}(t) = 1 - (1-p)(1-q)^{k(t)}\,, \tag{4}$$

where $k(t)$ is the number of previous adopters with whom the individual maintains interactions in the matrix. The logic is that $1-p$ is the probability that the individual has not been influenced to adopt by the external force, and $(1-q)^{k(t)}$ is the probability that she was not affected by any w-o-m. This general ABM process is performed computationally by running a stochastic process wherein at each period, the individual probability of adoption is determined by equations such as Eq. (4). The results for a particular realization of the stochastic process are depicted in Fig. 4.

Tracking the total number of "1"s in each period over time reveals the cumulative adoption of the product over time, and when taking into account the difference between periods, non-cumulative adoption is also evident. In the simple case shown above, marketers can examine, for example, how strategies that affect the individual-level adoption parameters p and q influence the aggregate adoption curve.

Back to the Saddle

Using data on a large number of innovative products in the consumer electronics industry, one can find that a considerable percentage of the cases followed a similar pattern: an initial peak, giving rise to a trough of sufficient depth and duration to exclude random fluctuations, followed by sales which eventually exceeded the initial peak. This pattern, the *saddle*, is explained by the dual-market phenomenon that treats the *early market* adopters and *main market* adopters as sufficiently different to warrant differential treatment as two separate markets. If these two segments – an early market and a main market – communicate at different rates, and if the difference is pronounced, then overall sales to the two markets will exhibit a temporary decline during the intermediate stage.

ABM can be mobilized to gain deeper understanding of the dual market phenomenon. The market is divided into two main groups:

1. The *early market* (indexed by i), and
2. The *main market* (indexed by m)

Each group has distinct external (p_i and p_m) and internal probabilities of adoption (q_{ii} and q_{mm}), corresponding to

the distinct nature of these markets. In addition, a cross-market effect parameter q_{im} was defined, to reflect some degree of communications and influence between early adopters and the main market.

ABM-generated data offers the flexibility required to generate different sets of data needed to prove the main point in question. In this study, if one wishes to prove that cross-market communications are the main determinant of the saddle phenomenon, one can generate multiple would-be worlds, differentiated by different cross-market communication values. For each value of this parameter, a number of different worlds are created by manipulating the values of the rest of the parameters. ABM can confirm that the saddle phenomenon is a natural by-product of the dual-market assumption, and illuminate the circumstances under which saddles form.

The models show good fit with the data [8], and can also be used to formulate predictions. However as we already mentioned, the saddle may be too simple to be called a genuine complex system phenomenon. Dividing the market to two sub-markets explains an important dynamics of innovation, but the complexity approach offers much more.

Agent Based Modeling and Aggregate Diffusion Models

The basic model described so far is similar in nature to the basic Bass (1969) model, which constitutes the foundation for most aggregate diffusion modeling in marketing, as well as much of the strategic thought on new product marketing (see [11], for a review). However, while the Bass model describes a process that occurs on the aggregate level, developments in the agent based model (ABM) takes place at the level of the individual level. While both methods can be used to model the growth of a new product, ABM offers the advantage of more flexible dynamics that avoid the typical oversimplifications of aggregate assumptions through the following principles:

Individual-level assumptions: Using ABM, researchers model the growth process without having to make aggregate-level assumptions that are not theory-based. For example, the Bass model assumes that the hazard rate of adoption is a simple linear function of both external and internal effects. However, this assumption lacks any theoretical foundation, and the examination of other functions can be equally justified.

Simplifying assumptions: Most diffusion models include a number of implicit or explicit simplifying assumptions that were originally introduced to facilitate analytical solutions [12]. These assumptions include a binary adoption state (adopted/did not adopt), constant market potential, a single adoption event by a single adopting unit, probability functions are universal (applied equally to all individuals), and a non-changing innovation. While extensions to the Bass model have tried to relax some of these assumptions (see [11]), these attempts have generally been limited and performed piecemeal, thus preserving many of the assumptions of the basic model. The ABM diffusion model, on the other hand, can relax all of the assumptions mentioned above in a tractable way. The two studies cited below relax several of the original assumptions, and we are confident that many others will follow.

The Lost Dimension: Using Spatial Analysis to Predicting New Product Success

Analyzes of new products in marketing literature has been typically time-dependent. Studies have sought to describe, explain, and predict how product sales grow over time. Although patterns of product growth *in space* can also be of crucial interest to marketers, very few formal attempts have been to model growth over time and space due to the complexities involved in such modeling. Because ABM modeling explicitly takes into account the spatial connection among potential adopters, it is well-suited for spatial analysis.

One of the main problems associated with early-period forecasting of new product success is the lack of sufficient sales data required for reliable predictions. A study by Garber et al. [4], attempted to enhance prediction reliability by utilizing information related to the *spatial adoption* of a new product. According to diffusion theory, word of mouth and imitation may play a significant role in the success of many innovative products. Because traditional word-of-mouth spread is often associated with some level of geographic proximity between the parties involved, one can expect the formation of non-uniform "clusters" of adopters. Alternatively, when market reaction to external factors is an overall reluctance to adopt the new product, word-of-mouth effect is expected to be significantly reduced, leading to a more uniform pattern of sales (assuming that there are no external reasons for clustering). Below we explain in greater detail how adoption patterns may help marketers formulate reliable early-period predictions of new product success.

Figure 5 describes the spatial growth of two products in an ABM environment. Figure 5a presents a simulated product adoption process in six discrete time periods (8,

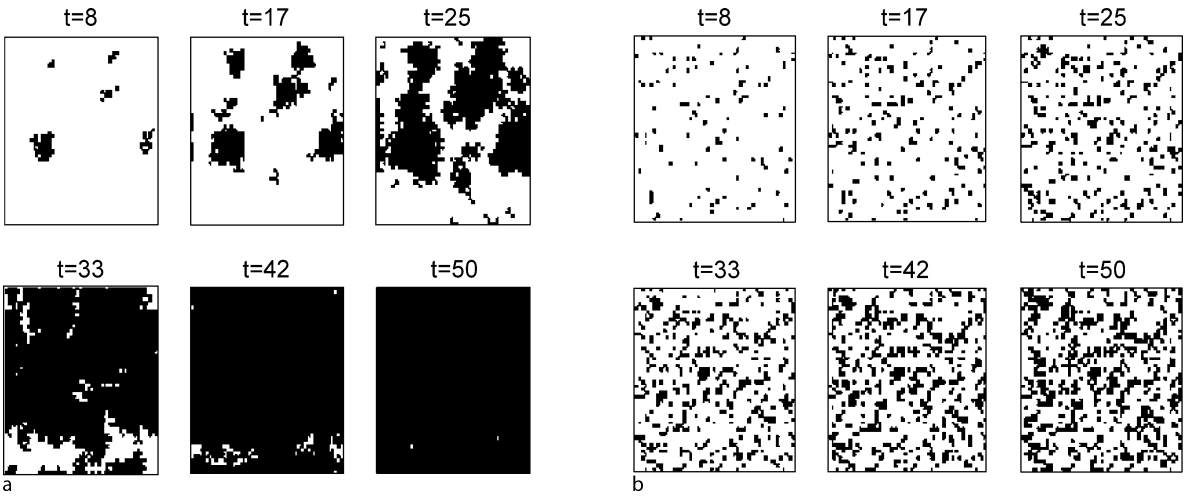

Marketing: Complexity Modeling, Theory and Applications in, Figure 5
Spatial Adoption of Two Products: a Successful Product (clustered) b Failed Product (uniformly distributed)

17, 25, 33, 42, and 50) over a certain rectangular geographical area. The product is clearly adopted in clusters. In comparison, in Fig. 5b presents the adoption pattern of a second product, whose distribution of adopters in the same geographical area is relatively uniform.

ABM was used in this case in a number of ways. First, it was used as a graphic tool to visually identify the phenomenon. Second, ABM was used to examine various scenarios and the conditions under which clusters are formed. Finally, it was used to examine the feasibility of using a distance measure (in this case, cross-entropy) to predict long-term success from early-period data. Following ABM results for this study, the early-period spatial analysis method was successfully applied to actual product data, and was found to generate correct predictions for the new product success of 15 out of 17 new products (that included 9 success and 8 failures).

We turn now to develop a more universal model that can be tailored to a wide range of applications based on [9].

A Universal Framework for Modeling the Market Penetration Dynamics

Can complexity modeling support a general (perhaps universal) modeling framework which takes into account both individual level behavior as well as aggregate results? Such a framework would bridge between individual-level and aggregate-level point of views. It will hence be a suitable framework for applying novel theories which are rooted in the individual consumer behavior analysis to explore the collective behavior of the whole market. In this section, we present a universal modeling platform for synthesizing individual and aggregate levels, using the most general case of new product adoption dynamics, and a specific focus on the dynamics of the initial purchase of a new product. Although we assume that an individual can adopt the product only once, this framework supports a straightforward generalization to repeat purchase cases as well.

Our unit of analysis is the individual consumer rather than the market segments or the entire market, as is typical of aggregate approaches. Let s_i be a binary variable that represents the state of adoption of potential customer$i$. All individuals in the market are potential adopters. That is, $s_i(t)$ takes the value 1 if customer i adopted the innovation before time t, and 0 otherwise. We define the vector $\vec{S}(t) = [s_1(t), s_2(t), \ldots, s_M(t)]$ as the market state vector at time t, where M is the market potential. A customer status change from potential adopter to actual adopter is based on the transition from $s_i = 0$ to $s_i = 1$.

As the transition process between potential adopter to actual adopter is stochastic in nature, we can define time-dependent probabilities $F_i(t)\Delta t$ $(i = 1, 2, \ldots, M)$, where $F_i(t)\Delta t$ is the probability that individual i will adopt the innovation within the time interval between t and $t + \Delta t$. These probabilities are determined by the market conditions which affect individuals' adoption decisions. Thus we can identify the transition rate $F_i(t)$ with applied market forces on potential adopter i, where more intense market force leads to a greater transition rate. In general market forces depend on the entire history of market dynamics. The entire history, denoted by Ω_t, may include current and the past market state vectors, $\{(\vec{S}(t), \vec{S}(t - \Delta t), \ldots, \vec{S}(0))\}$, the social network topology, intensities of word-of-mouth interactions among individuals, and the affect of

external factors (such as advertising and promotion activities) on different consumers.

Since our interest is exploring the dynamic penetration of a new product in the marketplace, we define a transition index for a potential customer i as

$$\Delta s_i(t) = s_i(t + \Delta t) - s_i(t)\,, \tag{5}$$

where given the entire history of the market dynamics Ω_t:

$$\Delta s_i(t) = \begin{cases} 1 & \text{with probability } F_i(t\,|\Omega_t)\Delta t \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

The transition indices $\Delta s_i(t)$ are stochastic variables by definition. If we set the time interval Δt to a very small value, which limits the probability of potential adopter communications, the resulting word-of-mouth effect to induce purchasing becomes negligible. In that case, the transition indices can be considered independent (Also note that the independence assumption does not imply that individuals who simultaneously purchase the new product cannot have mutual influence. Such an influence is encapsulated in the entire history of the market dynamics Ω_t).

The first step in integrating the micro and the macro points of view is to define the non-cumulative penetration of the new product. This is the number of individuals who have adopted the innovation within a given short time interval Δt, subsequent to a certain time t, and can be expressed as the sum of all individual transition indices, such that

$$\Delta N(t) = \sum_{i=1}^{M} \Delta s_i(t)\,. \tag{7}$$

The non-cumulative penetration is a stochastic variable in the sense that if we were able to reconstruct the current market status and repeatedly run the new product adoption process, each rerun would generate different results. Under the assumption that simultaneous purchases are uncorrelated within a short time interval Δt, and given the entire history of the market dynamics Ω_t, the non-cumulative penetration is a sum of independent binomially distributed variables. Thus, it can be shown that [9]

$$\Delta N(t) = F(t, \Omega_t)\Delta t + \varepsilon_t \tag{8}$$

where

$$F(t, \Omega_t) = \sum_i F_i(t, \Omega_t)$$

is the net market force applied to entire potential market and ε_t is uncorrelated noise with mean zero which satisfies

$$\mathrm{E}\big(\varepsilon_t^2\,|\Omega_t\big) \cong \mathrm{E}(\Delta N(t)\,|\Omega_t)$$

($E(\cdot\,|\Omega_t)$ denotes a conditional excepted value, given Ω_t the entire history of the market dynamics at time t.)

Equation (8) is in fact an "equation of motion of penetration dynamics" where modeling the net market force becomes, in a sense, a marketing engineering challenge (Marketing engineering is a relatively new concept (see http://www.mktgeng.com/) of education and implementation framework of marketing models. Here, we propose to extend this view and include development and tailor models using complexity framework in a marketing engineering framework).

In many respects, this assignment is analogous to typical problems in classical mechanics where one aims to retrieve the spatial motion of physical bodies by modeling the applied force on those bodies and solving the appropriate equations of motion provided by Newton's second law of motion. Yet, while the dynamical equations of classical mechanics are deterministic, the penetration dynamics is stochastic by nature.

Furthermore, actual net market force F is defined as a function of Ω_t, the entire history of the market dynamics that also contains individual-level information. Since such data are seldom available, the applicability of such a modeling approach is significantly restricted. Nevertheless, we can use Eq. (8) (non-cumulative penetration), which we recast as follows to reflect partial information:

$$\Delta N(t) = \hat{F}(t, \hat{\Omega}_t)\Delta t + u(t) \tag{9}$$

where $\hat{F}(t, \hat{\Omega}_t)$ is the model of the net market force which is used to estimate actual net market force F on the basis of the partial information $\hat{\Omega}_t$. In contrast, the term $u(t)$ denotes the actual noise, or stochasticity, of the process. The actual noise contains all the relevant micro-level information that has not been (or cannot be) modeled. Since in general $E(u(t)\,|\Omega_t) \neq 0$ and $E(u(t)u(t')\,|\Omega_t) \neq 0$ for different time indexes t and t', the actual noise in the non-cumulative penetration usually becomes biased and correlated, resulting in strong coupling effects evident in the penetration curve. These effects can be interpreted as large fluctuations or trends in the adoption rates. That is, modeling the actual net market force F exhibits the following trade-off: The simpler and hence the less micro-informative the model of the net market force $\hat{F}$, the more significant the impact of the noise u in producing large fluctuations and trends in the penetration data.

On the other hand, by zooming in to a unit of analysis which is a single consumer (rather than market segments or the entire market), and to more granular (e. g., daily) instead of smoothed (e. g., annual) data, we are able to obtain a more accurate modeling of the net market force. How-

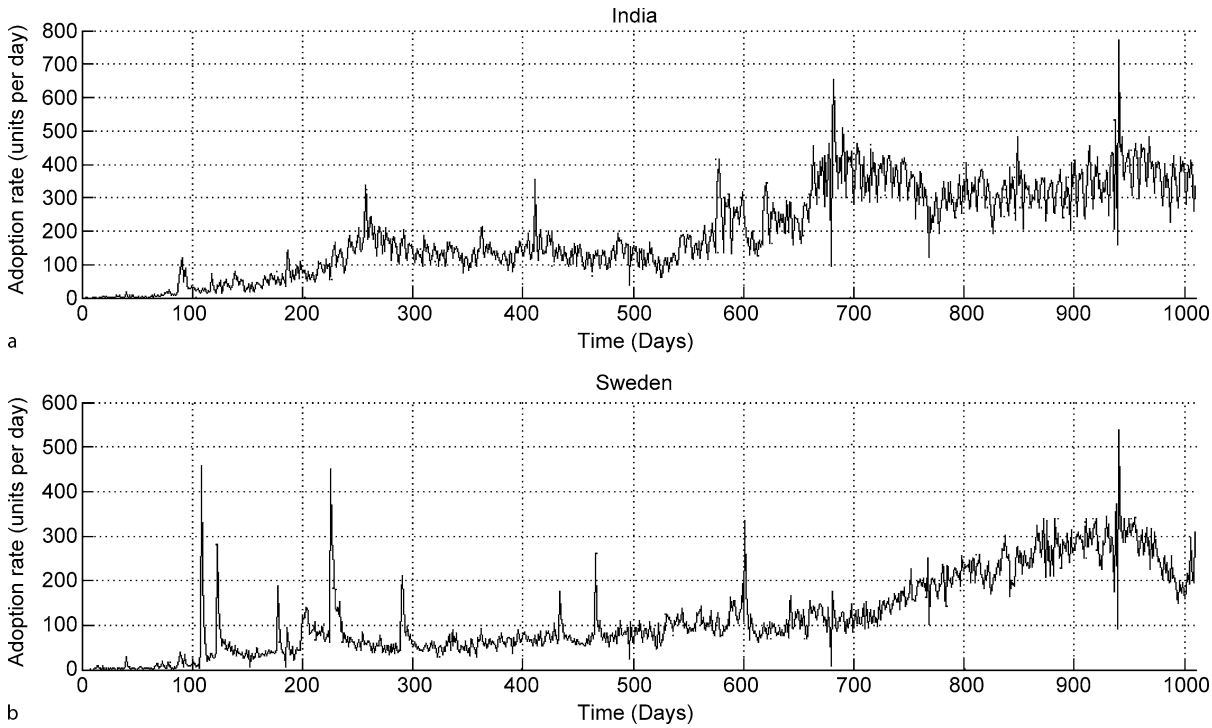

Marketing: Complexity Modeling, Theory and Applications in, Figure 6
Actual daily penetration of an e-mail software product in India and Sweden

ever, such models may become very difficult, if not impossible, to formulate and solve.

The stochastic term $u(t)$ incorporates all the effects that have not been modeled, including heterogeneity among individual consumers, social networks topology, dynamical changes of external and internal influences. These effects may cause significant deviations from the pattern of product life cycle, predicted by traditional models (e. g., the Bass model). This phenomenon is illustrated in Fig. 6 which presents the actual daily penetration of an e-mail software product in two different markets (India and Sweden). Although a primary tendency of increasing adoption rates is evident in all cases (as expected in the initial phase of a standard new product growth process at the beginning of the product life cycle), secondary movements and large fluctuations can also be observed. Naturally, in order to explicate these secondary movements and large fluctuations one may use a more realistic model $\hat{F}$ of the net market force F that includes additional and more detailed information, rather than rely on total adoption rates.

This particular example is not merely a theoretical case but has significant practical implications as well. Often, monthly or weekly sales curves of trends, patterns and changes of the curve are characterized by "noise", a term often used (in most fields) to reflect measurement tool error. This is what we see in Fig. 6. The "noise" in sales data is typically handled by data smoothing and the use of larger time frames of analysis (i. e., quarters or years) in aggregate models. However, as measurement tools and information technology developme and allow increasingly precise measurements, we can posit that what we see in granular data is *not* noise, but simply the true face of growth. In particular, we can identify "ripples that ride on small waves". Goldenberg, Lowengart and Shapira [9] argued that the volatility may surprisingly contain valuable information. Accordingly, they propose to "step inside the noise" and use granular data instead of the more commonly used smoothed quarterly or annual data. Analysis of the structure of the "waves" yields information that improves the accuracy of post-launch predictions at much earlier stages (i. e., immediately following product introduction). By analyzing highly fluctuating daily data, changes in sales patterns can self-emerge as a direct consequence of the stochastic nature of the process. This study demonstrated how to "stepping into the noise-like data" and treat it as information, through a Kalman–Filter-based tracker, to trace and significantly improve predictions.

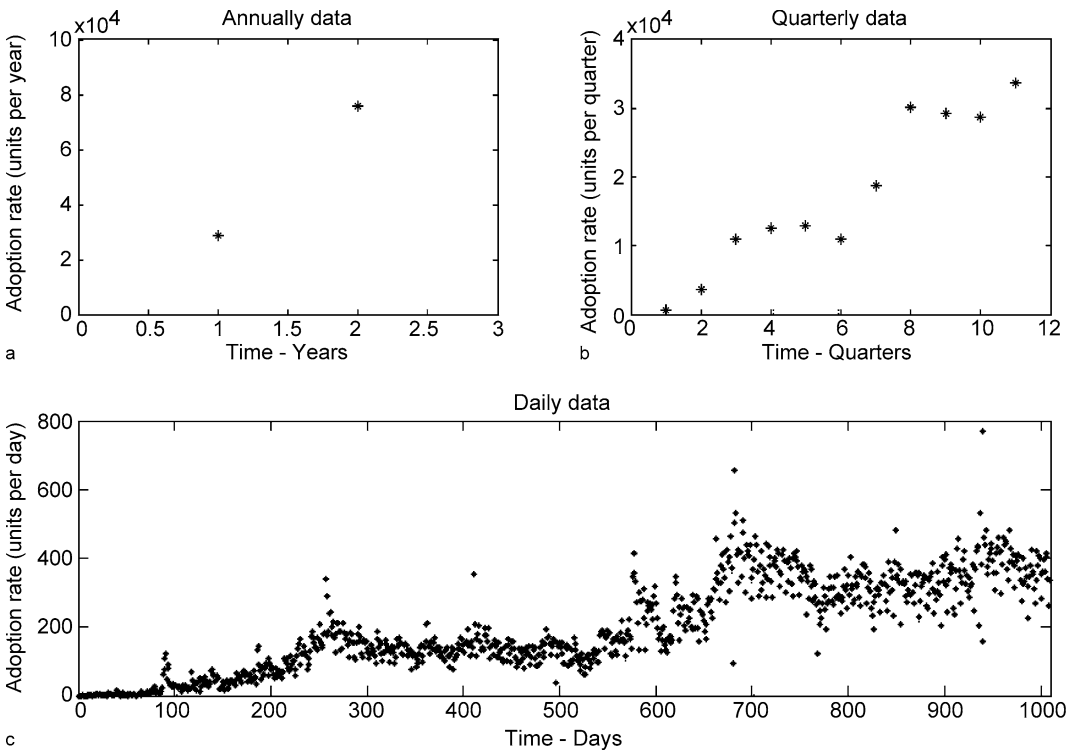

Marketing: Complexity Modeling, Theory and Applications in, Figure 7
Actual penetration data of a software product on an a annual, b quarterly, and c daily scale

Selecting the best time resolution for data analysis is not a trivial decision. Annual data (Fig. 7a) offers only two data points that scantly provide any insight, while daily data (Fig. 7c) involves high fluctuations that may appear to be noise. Possible compromises including accumulating data over several years, or using quarterly data (Fig. 7) which is less noisy than the daily data yet offers sufficient data points to make forecasts based on smoothing and fitting. Is this our best option for analysis? Goldenberg, Lowengart and Shapira [9] show that the patterns observed in Fig. 7c are an inherent component of the diffusion process of new products, and better predictions result at very early stages of the product introduction process if granular data (instead of smoothing) is used. More specifically, in few examined cases it was possible to provide forecasting already 50 days after launch two quarters ahead. This is a valuable information for firms that have to make plans, and such forecasting are not provided by regular diffusion models (because they use smoothed data, such annual one). When comparing the model with 3 diffusion models the fit of the forecasts using the zooming in approach was sometimes twice better.

Through the aggregation of a microscopic approach, marketers may obtain insights to the collective behavior of their target market, by resolving practical and operational tradeoffs imposed by factual constraints and by the firm's business plan and objectives. In this context, one can think of the role of a "sales engineer" who is familiar with his or her target market and can evaluate the specific market forces that are to be substituted in the sales equation of motion in order to derive the dynamics of the new product consuming process. The universal framework based on micro-level data, offers a natural platform to profile collective market behavior by applying advanced theories in marketing that are usually rooted in the analysis of individual consumer behavior.

The universal framework for modeling market penetration dynamics involves an inter-disciplinary approach that combines complex system analysis methods with advanced marketing theories. It can also be applied to other problems from different disciplines that involve the dissemination of information over social networks, including collective decision making, the spread of computer viruses via the internet, or the evolution of species in ecological environments, to name only a few.

Illustration of Penetration Dynamics Modeling Using the Universal Framework

To illustrate the use of the universal framework for modeling market penetration dynamics, we consider the most simple case of a homogeneous and stationary market. Let $p\Delta t$ be the probability that a potential adopter is persuaded by an external influence (e. g. advertising or firm's promotion activities) to adopt the new product within a short time interval Δt; let $q\Delta t$ be the probability that a potential adopter interacts with an actual adopter and is affected by his or her word of mouth to adopt the innovation within a short time interval Δt, and let $N(t)$ be the total number of actual adopters at time t. Assuming that word of mouth communications and external effects are orthogonal, we can model the applied market force on any potential adopter i as follows:

$$\begin{aligned}\hat{F}_i(t)\Delta t &= 1-(1-p\Delta t)(1-q\Delta t)^{Nt}\\ &\cong (p+qN(t))\Delta t\,. \end{aligned} \tag{A1}$$

In this case, the net market force is given by the product of the number of potential adopters and the applied market force on each potential adopter. Namely,

$$\begin{aligned}F(t,\Omega_t) &\approx \hat{F}(\hat{\Omega}_t)\\ &= (M-N(t))\left(p+\frac{Q}{M}N(t)\right)\,, \end{aligned} \tag{A2}$$

where M is the total market potential, $P=p$ and $Q=Mq$. Partial information $\hat{\Omega}_t$ on which our model of the net market force is based, includes $N(t)$ the total number of actual adopters at time t as well as the constant parameters P, Q and M that denote market potential, external influences, and word of mouth effects respectively (Note that our model is not explicitly time-dependent). After modeling the net market force, we can now use Eq. (9) to describe the dynamics of new product growth, as follows:

$$\begin{aligned}\Delta N(t) &= \left((M-N(t))\left(p+\frac{Q}{M}N(t)\right)\right)\Delta t\\ &\quad + u(t)\,. \end{aligned} \tag{A3}$$

As a special case, we can set Eq. (A3) to the continuous limit, while neglecting the stochastic effects of the process to obtain the following ordinary differential equation:

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = (M-N(t))\left(p+\frac{Q}{M}N(t)\right) \tag{A4}$$

which is the Bass Eq. (1) [2].

Future Directions

An important question is how robust the approach is to errors resulting from data that is not at the industry level. While the last demonstration used penetration data of a unique product, and the data can be considered a good proxy for an industry level, there are other studies that are a case of partial data (firm level). The results indicate that despite such a limitation, the proposed model works well, and its predictions are relatively high.

The proposed framework and modeling approach lay the groundwork for further research. Extending the current model to account for repeat purchase goods is a natural step that can provide meaningful insight into the emergence of sales trends. Another avenue for future research might focus on identification of opinion leaders and their effect on other consumers through social ties. Another important force is the resistance to innovation, in which a negative word of mouth is disseminated and competes again the positive one. The two diffusions interact and the outcome of this battle can determine the faith of the new product. Modeling the two forces seems to be doable. Finally, the spatial dimension can be more sophistically utilized if information is used to develop a spatial allocation of marketing resources.

Bibliography

1. Anderson P (1999) Complexity theory and organization science. Organ Sci 10(3):216–232
2. Bass FM (1969) A new product growth model for consumer durables. Manag Sci 15:215–227
3. Bass's Basement (2008) Bass's Basement Research Institute, Austin. http://www.frankmbass.org/. Accessed 2008
4. Garber T, Goldenberg J, Libai B, Muller E (2002) From density to destiny: using spatial analysis for early prediction of new product success. Marketing Science Institute, Cambridge
5. Goldenberg J, Libai B, Muller E (2001) Talk of the network: a complex system look at the underlying process of word of mouth. Mark Lett 12:209–221
6. Goldenberg J, Libai B, Muller E (2001) Using complex systems analysis to advance marketing theory development: modeling heterogeneity effects on new product growth through stochastic cellular automata. Acad Mark Sci Rev 1(9), special issue on Emergent and Co-evolutionary Processes in Marketing:1–9
7. Goldenberg J, Lehmann DR, Mazursky D (2001) The idea itself and the circumstances of its emergence as predictors of new product success. Manag Sci 47(1):69–84
8. Goldenberg J, Libai B, Muller E (2002) Riding the saddle: how cross-market communications can create a major slump in sales. J Mark 66(2):1–16
9. Goldenberg J, Lowengart O, Shapira D (2008) Zooming in. Mark Sci (forthcoming)
10. Hegselmann R (1998) Modeling social dynamics by cellular automata. In: Liebrand WBG, Nowak A, Hegselmann R (eds) Computer Modeling of Social Processes. Sage, London

11. Mahajan V, Muller E, Wird Y (2000) New-Product Diffusion Models. Kluwer, Dordrecht
12. Mahajan V, Peterson RA (1979) Integrating Time and Space in Technological Substitution Models. Technol Forecast Soc Change 14:231–241
13. Moore GA (1991) Crossing the chasm. HarperBusiness, New York
14. Muller E, Yogev G (2006) When does the majority become a majority? Empirical analysis of the time at which main market adopters purchase the bulk of our sales. Technol Forecast Soc Change 73(10):1107–1120
15. Rindfleisch A, Moorman C (2001) The acquisition and utilization of information in new product alliances: a strength-of-ties perspective. J Mark 65(2):1–18
16. Ragers EM (1995) The Diffusion of Innovations, 4th edn. Free Press, New York
17. Rosen E (2000) The anatomy of buzz: How to create word of mouth marketing. Random House, New York
18. Rosenau MD et al (eds) (1996) PDMA Handbook of new product development. Wiley, New York
19. Rosser JB (1999) On the complexities of complex economic dynamics. J Econ Perspect 13(4):169–192
20. Sarkar P (2000) A brief history of cellular automata. ACM Comput Surv 32(1):80–107
21. Srinivasan V, Mason CH (1986) Nonlinear least square estimation of new product diffusion models. Mark Sci 5:169–178
22. Van den Bulte C, Joshi YV (2007) New Product Diffusion with Influentials and Imitators. Marketing Sci 26(3):400–421
23. Waldrop MM (1992) Complexity. Touchstone Books, New York

Market Microstructure

Clara Vega, Christian S. Miller[1]
Division of International Finance,
Board of Governors of the Federal Reserve System,
Washington DC, USA

Article Outline

Glossary
Definition of the Subject
Market Structures
Inventory Models
Information-Based Models
Easley and O'Hara's Model
Kyle Model
Empirical Market Microstructure
Future Directions
Bibliography

[1]The concepts in this paper in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or of any other person associated with the Federal Reserve System.

Glossary

Ask price Price at which a trader is willing to sell an asset. The most competitive ask price in a financial market or best ask is the lowest price offered by a seller.

Bid price Price at which a trader is willing to buy an asset. The most competitive bid price or best bid in a financial market is the highest price offered by a buyer.

Limit order Orders placed by market participants contingent upon the realization of a certain price in the market. In other words, traders will identify a maximum or minimum price at which they are willing to buy or sell a specific quantity of a particular asset.

Market order Order to buy or sell a particular asset immediately at current market prices.

Market structure The way in which trade occurs within a particular market. Institutions have constructed idiosyncratic guidelines to dictate how transactions can take place, so generalizing one trading structure to model all markets is quite difficult, if impossible.

Order flow is the cumulative flow of signed transactions over a time period, where each transaction is signed positively or negatively depending on whether the initiator of the transaction (the non-quoting counterparty) is buying or selling, respectively. By definition, in any market, the quantity purchased of an asset equals the quantity sold of the same asset. The key is to sign the transaction volume from the perspective of the initiator of the transaction.

Bid-ask spread The difference between the highest bid price and the lowest ask price. This difference, or spread, constitutes part of the cost of trading.

Definition of the Subject

Market microstructure is a field of study in economics that examines the way in which assets are traded and priced under different trading mechanisms, e. g., single-price call auction, dealer markets, limit-order book markets, hybrid markets, etc., and under different trading environments, e. g., perfect information environments (complete markets) compared to asymmetric information environments (incomplete markets). While much of economics abstracts from the market structure and market frictions the microstructure literature specializes in understanding them and the effects they may have on asset prices and quantities traded. Even though economic theorists assume a frictionless economy to prove powerful theorems about the efficiency of a decentralized market system, the market structure and market frictions can be very important. Ignoring them may lead researchers and policy makers to wrong conclusions. For example, in a Walrasian world with per-

fect information and no transaction costs, prices efficiently aggregate information when trading is organized as a single-price call auction with large numbers of traders. However, most securities markets are not single-price call auctions as several studies show that this trading mechanism may be optimal when uncertainty about the fundamental value of the asset is high, but it is not optimal at other times. Furthermore, in the 1970's the economics of information literature argued that allowing for imperfect information could overturn the central implication of the complete-markets model, that competitive, decentralized markets yield economically efficient results.

Market Structures

A large part of the market microstructure field consists of developing models to describe the behavior of individuals acting according to the guidelines of various trading institutions, and to study how trading quantities and prices in various markets arise given a particular set of assumptions. Thus, we start with a short description of the common market structures. It is outside the scope of this article to detail the myriad rules that govern various financial markets. It is also counter-productive because trading systems are in a continuous process of structural changes generated by research, competition, and technological innovations. Instead we present a general outline of the guidelines that dictate the way in which assets trade and the effects these rules may have on asset prices and quantities traded.

Auctions

Auctions are order-driven trading mechanism, i. e., investors submit their orders before observing the transaction price. In contrast, investors in a quote-driven trading mechanism obtain firm price quotations from dealers prior to order submission (these price quotations usually depend on the size of the order). Auctions can be continuous or periodic. An example of a continuous auction is the automated limit order book, which consists of a sequence of bilateral transactions at possibly different prices (we describe limit-order books in more detail below). In contrast, a periodic or call auction is characterized by multilateral transactions. Periodic or batch systems, such as the single-price call auction, are used to set opening prices in several exchanges, e. g., NYSE, Tokyo Stock Exchange, etc. In these markets limit orders and market-on-open orders are collected overnight. At the beginning of the trading day the specialist chooses the price that enables the largest number of orders to be executed. Stock exchanges use call auctions to fix opening prices because uncertainty about fundamentals is larger at the opening than during regular trading hours. Indeed, Madhavan [36] provides a theoretical argument for batch markets as a way to reduce market failures caused by information asymmetries. Another example of a periodic auction market is the primary market for US Treasury securities. These securities are sold through sealed-bid single-price auctions at pre-determined dates announced by the US Treasury Department (before November 1998 the Treasury auctioned securities through multiple-price or discriminatory auctions).

Limit Order Markets

Limit order books are the most widespread conduit to facilitate trade in financial markets; at least one limit order book exists in most continuous (liquid) security markets (see p. 10 in [29]). In such markets, traders submit their bid and ask orders, and the order book(s) process these orders, comparing them to already existing orders to establish whether any matches can be made. These pre-existing, unfilled limit orders comprise the limit order book. Various rules dictate how and when limit orders are acted upon Parlour and Seppi [42]. Generally price and then time determine priority of execution. For instance, a limit order to sell an asset for $50 will take precedence over an order to sell at $52. If two limit orders are priced the same, then the first limit order submitted is the first order executed.

Sometimes traders may request the execution of a market order; this order is immediately executed at the best price available. A problem can arise if the quantity designated in the market order is larger than the quantity available at the best price available on the limit book. Different exchanges have different rules to deal with the left-over quantity. In the NYSE, the excess quantity "walks the book", meaning that the market order achieves partial executions at progressively worse prices until the order is filled. This process results in the partial execution of market orders at less than desirable prices for the order-issuing trader. In contrast, in the Euronext system and the Paris Bourse, if the quantity in the market order exceeds the quantity at the best price, the unfilled part of the market order is transformed into a limit order, requesting execution on the remaining quantity at the same execution price.

Various other rules regarding the execution of limit orders exist. For example, traders can post orders with an expiration time, i. e., the limit order is canceled if it is not executed within a given time frame. This prevents limit orders to be "picked off" by investors who receive updated public or private information. Traders can also hide part of the order they submit to the limit order book, these are called "iceberg" orders.

Exchanges vary in the degree of transparency of the limit order books. The automated limit-order-book system used by the Toronto Stock Exchange and the Paris Bourse are among the most transparent systems. They offer continuous trading and the public display of current and away limit orders (an open book limit-order system). The NYSE has shifted from a close limit order book policy (although specialists made the book available to traders on the floor at their own discretion) to making the content of the limit-order book public. In January 24, 2002 the service OpenBook was introduced. This service provides information about depth in the book in real time at each price level for all securities to subscribers either directly from the NYSE or through data vendors such as Reuters and Bloomberg. Boehmer et al. [10] empirically examine the effect of increased transparency in the NYSE and Goettler et al. [24] numerically solves a dynamic model of limit orders in which agents arrive randomly and may trade one share in an open electronic limit order market.

Single Dealer Markets

It is a market where one dealer (market maker or specialist) stands ready to buy at his bid quote and sell at his offer quote. In this environment, incoming orders are necessarily market orders (in contrast to limit orders). The customer either buys (sells) at the dealer's offer (bid) or chooses not to trade. Dealer markets are less transparent than open book limit order markets (only the best-bid and best-ask price are known to the customer in a dealer market, while the entire depth of the market is visible in an open limit-order book). In reality there are very few pure single-dealer markets. The NYSE is sometimes mistakenly labeled as a single-dealer market, but it is a hybrid system with both limit-order and single-dealer features. Equity trading is centered on the stock specialist, who is assigned particular stocks in which to make a market. Each listed security has a single specialist, and all trading on the exchange must go through the specialist. The specialist receives market orders (orders for immediate execution) and limit orders (orders contingent on price, quantity and time), so that specialists do not enjoy monopoly power because they compete against the limit order book. If a market order comes to buy, the specialist can either match it with the best sell limit order or if he offers a lower price, he can take the other side. Examples of pure single-dealer markets are foreign exchange markets in developing countries with fixed exchange rates, where all orders must be routed through a single dealer – the central bank.

Multiple Dealer Markets

Competition in this environment is brought through multiple dealers. In a centralized market, quotes from many dealers are available on a screen (NASDAQ) or on the floor of an exchange (like a futures trading pit: the Chicago Board of Trade, the New York Mercantile Exchange, and the Chicago Mercantile Exchange). In a decentralized market trading occurs over-the-counter rather than through an organized exchange. The foreign exchange market, government bond's secondary market and corporate bond markets are good examples of decentralized multiple-dealer markets. There is less transparency in these markets than in a centralized multiple-dealer one because not all dealer quotes are observable. As a result, there can be simultaneous transactions that occur at different prices. The main mechanism that mitigates the dealer's monopoly power is the fact that the inter-

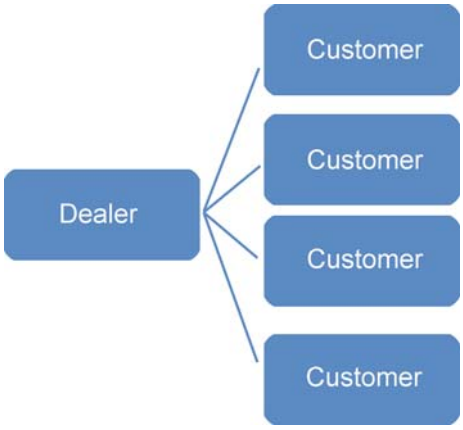

Market Microstructure, Figure 1
Single dealer market

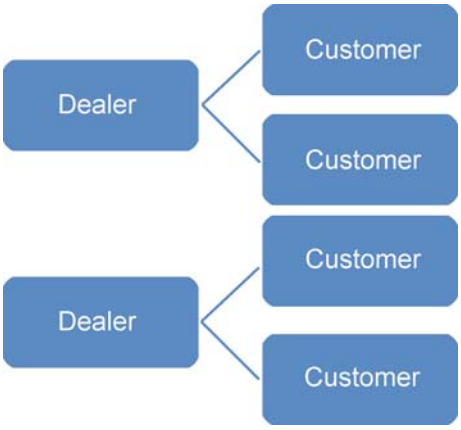

Market Microstructure, Figure 2
Multiple dealer market

action between a dealer and a customer is repeated over time. Dealers have an incentive to keep their reputation in quoting reasonable bid and ask prices so that the customer does not go to another dealer. In particular, dealers are concerned about losing large customers, so that small customers have less bargaining power. Competition in these markets and pressure from regulators has also forced a shift from voice-based brokers to electronic brokers, who provide a higher level of transparency. For example, recently the Bond Market Association responded to SEC pressure for more transparency in the corporate bond market by setting up a single reporting system for investment grade bonds Viswanathan and Wang [47]. For a detailed description of how the foreign exchange market and the government bond market operate please refer to Lyons [35] and Fabozzi and Fleming [21], respectively.

Inter-Dealer Markets

In addition to dealer-customer interactions, inter-dealer trading is very important. Ho and Stoll [31] suggest that risk-sharing is the main reason for inter-dealer trading. The incoming orders that a particular dealer receives are rarely balanced, so that the dealer is left with an undesired short or long position. To balance their inventory the dealer can sell to or buy from other dealers. The dealer can do so by either contacting another dealer directly or through a broker. The benefit of going through a broker is that they provide anonymity and the cost is the broker fee. In addition, brokers offer dealers electronic trading platforms that help the flow of information. These screens typically post: (i) the best bid and offer prices of several dealers, (ii) the associated quantities bid or offered, (iii) transaction prices, and (iv) transaction size. Common brokers in the secondary government bond market are ICAP's BrokerTec, Cantor Fitzgerald/eSpeed, Garban-Intercapital, Hilliard Farber, and Tullett Liberty. The main electronic brokers in the major spot markets (JPY/USD, Euro/USD, CHF/USD and GBP/USD currency pairs) are EBS and Dealing 2000-2, a dealer-broker Reuter product (Dealing 2000-1 is the Reuter product for direct inter-dealer trading). It is worth noting that EBS and Dealing 2000-2 typically conduct trades via a limit order book, while Reuters D2000-1 is a sequential trading system (an outside customer trades with dealer 1 who trades with dealer 2 who trades with dealer 3 and so on; hence it is often referred to as "hot potato" trading). In the equity markets inter-dealer trading is also common. On the NASDAQ market, dealers can trade with each other on the SuperSoes system, the SelectNet system and on electronic crossing networks (ECNs) like Instinet. In equity markets like the London Stock Exchange, inter-dealer trading constitutes about 40% of the total volume Viswanathan and Wang [47], while in the foreign exchange market and the US government bond market inter-dealer trading far exceeds public trades. Inter-dealer trading accounts for about 85% Lyons [35] of the trading volume in the foreign exchange market and about 99% Viswanathan and Wang [47] in the US government bond market. Two-thirds of the transactions in the US government bond market are handled by inter-dealer brokerage firms and the remaining one-third is done via direct interactions between the primary dealers listed in Table 1. For more details on inter-dealer trading please refer to Viswanathan and Wang [47].

Market Microstructure, Table 1
Primary Government Securities Dealers as of Nov. 30, 2007. Source: Federal Reserve Bank of New York
http://www.newyorkfed.org/markets/pridealers_listing.html

BNP Paribas Securities Corp.
Banc of America Securities LLC
Barclays Capital Inc.
Bear, Stearns & Co., Inc.
Cantor Fitzgerald & Co.
Citigroup Global Markets Inc.
Countrywide Securities Corporation
Credit Suisse Securities (USA) LLC
Daiwa Securities America Inc.
Deutsche Bank Securities Inc.
Dresdner Kleinwort Wasserstein Securities LLC.
Goldman, Sachs & Co.
Greenwich Capital Markets, Inc.
HSBC Securities (USA) Inc.
J.P. Morgan Securities Inc.
Lehman Brothers Inc.
Merrill Lynch Government Securities Inc.
Mizuho Securities USA Inc.
Morgan Stanley & Co. Incorporated
UBS Securities LLC.

In the next section we present a few of the basic models that are employed in the market microstructure literature.

Inventory Models

The first theoretical models in the market microstructure field were inventory models; however information-based models have come to dominate the field because the former describe temporary price deviations around the equilibrium price, while the later describe permanent price changes. The main idea of inventory models is captured by Smidt [44] who argued that dealers, or market makers

in general, are not simply passive providers of immediacy, but actively adjust the bid-ask spread in response to fluctuation in their inventory levels. Though dealers' main responsibility is to facilitate trade in an asset market, they set prices to realize rapid inventory turnover and to prevent the accumulation of significant positions on one side of the market. The consequence of this paradigm is a price that may diverge from the expected value of an asset if a dealer is long or short relative to a desired (target) inventory, which would result in temporary price movements over various (short-term) periods of time. How "short-term" these deviations are differs across studies. Data on specialists' inventories is scarce, but studies have been successful in showing that inventories play an important role in intraday trading and a recently published paper by Hendershot and Seasholes [30] shows that inventory considerations affect prices beyond intraday trading. Hendershott and Seasholes argument is that market makers are willing to provide liquidity as long as they are able to buy (sell) at a discount (premium) relative to future prices. Hence, large inventories of the market maker should coincide with large buying or selling pressure, which cause prices to subsequently reverse (e. g., Amihud and Mendelson [2] and Grossman and Miller [26] provide inventory models that lead to reversals). But the reversal of prices does not have to be immediate, in fact, they document that reversals can take as long as 12-days.

Inventory models assume that there is no asymmetric information. Fluctuations in market prices, therefore, results solely from dealers' decisions about the positions of their inventory. Dealers' hold sub-optimal portfolios, bare a cost for maintaining inventories – holding assets for the purpose of providing liquidity to the market exposes them to risk. Consequently, market makers receive compensation (i. e., bid-ask spread) for incurring the transaction costs entailed in managing their inventories.

Various texts, including O'Hara [41], present different inventory models. The discussion below will focus on one such model—the model presented by Garman [22] that inaugurated the field of market microstructure and builds on Smidt [44] idea. As O'Hara notes, aspects of basic inventory models, such as the assumption of perfect information, are not realistic; however, it is still useful to review basic models' assumptions about the functioning of asset markets to isolate the various ways in which the behavior of market makers can influence asset prices.

Garman's Model

The expected value of the asset or the equilibrium price is equal to the price at which quantity demanded equals quantity supplied at a particular period in time. Let's label this price p^*. Garman (16) shows that it is optimal for the market maker to charge two different prices. One price, p_a, the ask price, at which he will fill orders wishing to buy the stock, and another price, p_b, the bid price, at which he will fill order wishing to sell the stock. These prices will not necessarily straddle the equilibrium price, p^*, i. e., $p_b > p^* > p_a$. By being willing to take profits in the form of stock inventory increases, the market maker can artificially inflate prices by maintaining the inequality $p_b > p_a > p^*$. In no case, however, will the market maker be able to set both prices below p^* without ultimately running out of inventory. Furthermore, if the market maker sets both prices equal to each other, equal to the equilibrium price, i. e., $p_b = p^* = p_a$, then the market maker will fail with certainty (i. e., the market maker will either run out of inventory or cash with probability equal to 1). In what follows we describe briefly how the model works and we ask the reader to refer to the original paper for more details. Garman considered two market clearing mechanisms: a dealer structure and a double auction mechanism; however, we will focus on the dealer structure only.

Garman conceived the dealer as a monopolist; he alone receives orders from traders, determines asset prices, and facilitates trade. In making the market, the dealer engages in optimizing behavior by maximizing his expected profit per unit of time while avoiding bankruptcy or failure, which is defined as depleting his inventory or losing all of his money. The dealer sets an ask price and a bid price at the beginning of trading, and investors submit their orders after observing the dealer's bid and ask quote. The arrivals of orders to buy and sell the asset are independent stochastic processes that are distributed according to a Poisson distribution. The dealer, therefore, runs a chance of failing since he must ensure liquidity—selling part of his inventory or buying a particular asset as determined by the arrival rate of buyers and sellers.

Assuming a Poisson arrival rate necessitates that (i) many agents are interacting in the market, (ii) these agents issue orders independently without consideration of others' behavior, (iii) no one agent can issue an infinite number of orders in a finite period, and (iv) no subset of agents can dominate order generation, which precludes the existence of private information. It requires that the order flow be stochastic without being informative about future market or price movements.

Garman's model is based on two equations—one that determines the dealer's cash, $I_c(t)$, at time t and one that determines the dealer's inventory of the asset, $I_s(t)$, at time t. At time 0, the dealer holds $I_c(0)$ units of cash and $I_s(0)$ of stock. Inventories at any point in time can be rep-

resented as follows:

$$I_c(t) = I_c(0) + p_a N_a(t) - p_b N_b(t) ,$$

$$I_s(t) = I_s(0) + N_b(t) - N_a(t) ,$$

where $N_a(t)$ is the number of executed buy orders at time t, $N_b(t)$ is the number of executed sell orders at time t, p_a is the ask price and p_b is the bid price for a stock. Using these equations, Garman sets forth to determine how a dealer can avoid market failure or bankruptcy (i. e., $I_c(t)$ or $I_s(t) = 0$). Preventing this situation from occurring is the main goal of dealers in setting asset prices. Garman [22] provides a detailed explanation for determining when failure will occur, but for the purpose of this article it is enough to skip to the main conclusion. In order to avoid market failure, dealers must set p_a and p_b to satisfy both equations:

$$p_a \lambda_a(p_a) > p_b \lambda_b(p_b) \quad \text{and}$$

$$\lambda_b(p_b) > \lambda_a(p_a)$$

where $\lambda_a(p_a)$ is the probability of stock leaving the dealer's inventory and $\lambda_b(p_b)$ is the probability of stock being added to the dealers inventory. Simultaneously satisfying these equations requires that the dealer set p_a above p_b. In other words, a spread must be in place in order for the dealer to avoid bankruptcy or market failure, though the market maker still faces a positive probability of failure.

Various inventory models exist that explain the presence of the bid-ask spread. Although Garman's approach focuses on the threat of market failure to explain the disparity in bid and ask prices, other explanations such as dealers' market power or risk aversion have also been proposed by theorists (see p. 51 in O'Hara [41]). Though the dissimilarities among inventory models are many, the common theme that links these models together is the complex balancing problem faced by the dealer who must moderate random deviations in inflows and outflows of cash and assets. Over the long run the flow of orders had no effect on asset prices, but the dealers' attempt to recalibrate their positions in response to the random stochastic order flows causes price fluctuation in the short run.

Information-Based Models

One implication of the inventory approach discussed in the previous section is that inventory costs determine the bid-ask spread. Beginning with an insightful paper by Bagehot [9], a new theory emerged to explain bid-ask spreads that did not rely on inventory costs, but rather posited an important role for information. These information-based models used insights from the theory of adverse selection to demonstrate how, even in competitive markets without explicit inventory costs, spreads would exist. In what follow we describe three information-based models to illustrate the insights gained from adopting an information-based approach to studying market interactions.

Copeland and Galai's Model

Copeland and Galai [14] were first to construct a formalized model incorporating information costs. Similar to Garman's inventory model the agents in the model are dealers and traders. In contrast to Garman's model, there is more than one dealer and there are two types of traders: informed and uninformed. Informed traders know the true value of the asset, P, and uninformed or liquidity traders trade for exogenous reasons to the value of the asset (e. g., immediate consumption needs). The existence of uninformed traders that trade for non-speculative reasons is ubiquitous in the literature. This assumption is necessary because for information to be valuable informed traders need to be anonymous. If traders known to possess superior knowledge could easily be identified, then no one would agree to trade with them. This is the so called no-trade equilibrium described in Milgrom and Stokey [39].

The trader arrival process is exogeneously determined and is independent of the price change process. This is the same assumption as in Garman's model, but this assumption is not harmless in the presence of informed traders as it appears likely that informed trader behavior would depend on what they know about the true value of the asset relative to what the market thinks. This aspect of the problem is not resolved in Copeland and Galai's paper, but other authors relax this assumption and allow the number of informed traders in the market to be endogenously determined. However, the main contribution of Copeland and Galai's paper is to show that even in the presence of competitive dealers, the mere presence of informed traders implies that the bid-ask spread will be positive. The dealer knows the stochastic process that generates prices, $f(P)$, knows the probability that the next trader is informed, π_1, and knows the elasticity of demand of uninformed and informed traders. With this information the objective of the dealer is to choose a bid-ask spread that maximizes his profits. If the dealer sets the bid-ask spread too wide, he loses expected revenues from uninformed traders, but reduces potential losses to informed traders. On the other hand, if he establishes a spread which is too narrow, the probability of losses incurring to informed traders increases, but is offset by potential revenues from liquid-

ity traders. His optimal bid-ask spread is determinted by a tradeoff between expected gains from liquidity trading and expected losses to informed trading.

The timing of the model is as follows. A trader arrives to the trading post, the dealer offers a quote, and the "true" price, P, is revealed immediately after the trade. An uninformed trader will buy an asset with probability π_{BL}, sell an asset with probability π_{SL}, and decide not to trade with probability π_{NL}. (The "L" in this notation reflects the fact that Copeland and Galai refer to uninformed traders as liquidity traders.) Because informed traders know the true value of P, their decisions to buy, sell, or refrain from trade are based on strategies that maximize their profit.

Dealers at any instant will trade with informed traders with probability π_1 and can expect to lose:

$$\int_{P_A}^{\infty}(P-P_A)f(P)\mathrm{d}P+\int_{0}^{P_B}(P_B-P)f(P)\mathrm{d}P\,,$$

where P_A and P_B are the ask and bid prices quoted by the dealer, and P is the "true" value of the asset. Dealers at any instant will trade with uninformed traders with probability $1-\pi_1$ and can expect to gain:

$$\pi_{BL}(P_A-P)+\pi_{SL}(P-P_B)+\pi_{NL}(0)$$

Because the dealer does not know whether individual trades are with informed or uninformed traders, the dealers' objective function is the product of π_1 and the first equation added to the product of $1-\pi_1$ and the second equation. The dealers' optimal bid and ask prices result from this maximization problem. If the prices are negative, however, the market closes.

Not all informed traders who arrive at the marketplace will trade. Informed traders who believe the quoted price by the dealer will fall between P_A and P_B will not trade. Hence, the elasticity of demand by informed traders with respect to the bid-ask spread interval is implicit in the limits of integration in the equation above. The dealers revenue comes from those liquidity traders who are willing to pay P_A-P or $P-P_B$ as a price for immediacy. The authors assume that the likelihood that a liquidity trader will consummate trade declines as the bid-ask spread increases, in other words, the liquidity traders elasticity of demand is implicit in the probabilities that liquidity traders will either buy the asset, sell it or not trade.

The framework described above can include competition by incorporating a zero-profit constraint into the dealers problem. The most important result is that even with risk neutral, competitive dealers, the bid-ask spread is positive. The size of the spread will depend on the particular elasticities of the traders' demand functions, and the arrival rate of informed and uninformed traders. As long as there is a positive probability that some trader is informed, the spread will not be zero.

This model, however, is a static one-trade framework and as such it does not allow trade itself to convey information. The model we describe in the next section captures the dynamic aspect of trading and introduces the concept of trade as signals of information.

Easley and O'Hara's Model

What follows is a brief summary of the model; for an extensive discussion of the structure of the model please refer to Easley and O'Hara [16].

The game consists of three players, liquidity traders, informed traders and a market maker. All players are risk neutral, there are no transactions costs, and there is no discounting by traders. The no-discounting assumption is reasonable since agents are optimizing their behavior over one day. Liquidity traders buy or sell shares of the asset for reasons that are exogenous to the model and each buy and sell order arrives to the market according to an independent Poisson distribution with a daily arrival rate equal to ε. The probability that an information event occurs is α, in which case the probability of bad news is δ and the probability of good news is $(1-\delta)$. If an information event occurs, the arrival rate of informed traders is μ. Informed traders trade for speculative reasons; if they receive good news (the current asset price is below the liquidation value of the asset) they buy one share of the asset, if they receive bad news they sell one share of the asset.

On days with no information events, which occur with probability $(1-\alpha)$, the arrival rate of buy orders is ε and the arrival rate of sell orders is ε as well. The model can be parametrized so that the arrival rate of liquidity buyers and sellers is different. However, the numbers of trades for certain stocks from 2000 on are very large, particularly for Nasdaq stocks, and as a result the parameter estimates suffer from a truncation error. To minimize this problem, it is useful to set the arrival rates of liquidity sellers and buyers equal to each other, so that one can factor out a common factor in the likelihood function as in Easley, Engle, O'Hara, and Wu [19]. Figure 3, represents a diagram of how the model works.

Thus, the total amount of transactions on non-information days is 2ε with the number of buys approximately equal to the number of sells. On a bad information event day, which occurs with probability $\alpha\delta$, we observe more sells than buys. To be precise, the arrival rate of buy orders is ε and the arrival rate of sell orders is $\varepsilon+\mu$. In contrast,

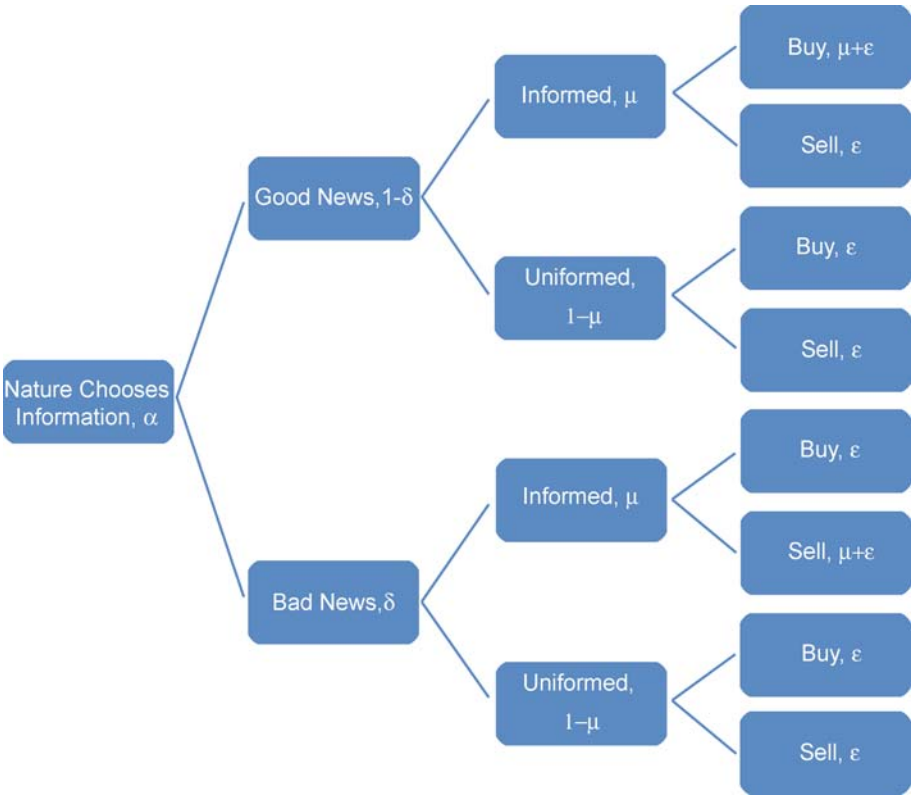

Market Microstructure, Figure 3
The tree diagram of the trading process [16]

on a good information event day, which occurs with probability $\alpha(1-\delta)$, we observe more buys than sells, i. e., the arrival rate of buy orders is $\varepsilon+\mu$ and the arrival rate of sell orders is ε.

Easley and O'Hara [16] define PIN as the estimated arrival rate of informed trades divided by the estimated arrival rate of all trades during a pre-specified period of time. Formally,

$$\mathrm{PIN} = \frac{\hat{\alpha}\hat{\mu}}{\hat{\alpha}\hat{\mu}+2\hat{\varepsilon}} .$$

One can estimate all four parameters, $\theta = \{\varepsilon, \mu, \alpha, \delta\}$, by maximizing the likelihood function

$$L(\theta|M) = \prod_{t=1}^{T} L(\theta|B_t, S_t)$$

where B_t is the number of buys and S_t is the number of sells on day t. Assuming days are independent, the likelihood of observing the history of buys and sells $\{M = (B_t, S_t)\}_{t=1}^{T}$ over T days is just the product of the daily likelihoods,

$$\begin{aligned} L(\theta|M) = {} & \alpha\delta e^{-(2\varepsilon+\mu)} \frac{\varepsilon^B (\varepsilon+\mu)^S}{B!S!} \\ & + \alpha(1-\delta) e^{-(2\varepsilon+\mu)} \frac{(\varepsilon+\mu)^B \varepsilon^S}{B!S!} \\ & + (1-\alpha)\delta e^{-(2\varepsilon)} \frac{\varepsilon^{B+S}}{B!S!} \end{aligned}$$

where T is equal to the time frame the researcher is interested in, e. g., Vega [46] choses 40 trading days before an earnings announcement is released, Easley, O'Hara, and Paperman [17] also use 40 trading days to estimate PIN, while Easley, Hvidkjaer, and O'Hara [18] use one calendar year to estimate PIN. The more trading days one uses to estimate PIN the more accurately one will measure information-based trading. Hence, one should check for robustness different estimation windows.

While all the parameters are identified and the likelihood function is differentiable, there is no closed-form solution to the four $(\varepsilon, \mu, \alpha, \delta)$ first-order conditions. Nev-

ertheless, the arrival rate of liquidity traders ε can be interpreted as the daily average number of transactions during the estimation window. The parameter μ reflects the abnormal or unusual number of transactions. The parameter α is equal to the proportion of days characterized by an abnormal level of transactions. The parameter δ is equal to the number of days with an abnormal number of sells divided by the number of days with an abnormal level of transactions.

To calculate the daily number of buys and sells most authors use the Lee and Ready [33] algorithm for NYSE- and AMEX-listed stocks and Ellis, Michaely, and O'Hara's [20] suggested variation of the Lee and Ready algorithm for Nasdaq-listed stocks. Odders–White [40], Lee and Radhakrishna [34], and Ellis, Michaely, and O'Hara [20] evaluate how well the Lee and Ready algorithm performs and they find that the algorithm is from 81% to 93% accurate, depending on the sample period and stocks studied. Thus the measurement error is relatively small.

To estimate the model using US stock market data most researchers use bid quotes, ask quotes, and transaction prices from the Institute for the Study of Securities Markets (ISSM) and the Trade and Quotes (TAQ) database. ISSM data contains tick-by-tick data covering the NYSE and AMEX trades between 1983 to 1992 and NASDAQ trades from 1987 to 1992, while TAQ data covers the sample period from 1993 to the present.

Vega [46] plots the time series of the parameter estimates in addition to the PIN measure averaged across all stocks in the sample. It is evident in that plot that the parameters ε and μ are not stationary. These parameters are related to the trading frequency, hence they are upwards-trending as the number of transactions has increased over the years. In contrast, the estimates of δ, α, and PIN are stationary over the years.

Vega [46] also shows average quarterly bivariate correlations of firm characteristics and PIN. PIN is most highly correlated with log market value with a bivariate correlation coefficient equal to -0.481. The cross-sectional range of -0.70 to -0.32 over the 64 periods implies that across stocks within the same quarter, PIN is negatively correlated with the firm capital size. To test this hypothesis formally Vega [46] first calculate Mann–Whitney test statistics for all periods. Then she tests the hypothesis that the sample of large firms has the same median PIN as the sample of small firms against the alternative hypothesis that they have different medians. In untabulated results she finds that she can soundly reject the null hypothesis in favor of the alternative for 60 out of the 64 periods she analyzes.

The negative relation between private information and firm size is consistent with both previous empirical studies that use PIN as an informed trading measure and Diamond and Verrecchia [15] who assert that asymmetric information is largest for small firms.

Next we present the Kyle Model, which is a workhorse within the market microstructure literature.

Kyle Model

In this information model, an auctioneer determines a price after all traders, uninformed and informed, submit their orders. Besides the risk-neutral market-maker, there is also one risk-neutral informed trader and multiple uninformed traders, who do not issue strategic orders. The market makers are unable to distinguish orders emanating from informed traders from those issued by uninformed traders. Informed traders understand this lack of transparency and can use it for their own advantage.

In the Kyle model there is just one risky asset that is traded over one period. This period of time consists of four distinct phases. First, the informed trader (and only the informed trader) observes the value V of the risky asset's payoff at the end of the period. V is a normally distributed random variable with mean zero and variance equal to σ_v^2. Second, market orders from the informed trader as well as the uninformed traders are submitted to the auctioneer, who is unaware of the end-of-period payoff of the asset, V. The market orders from the informed trader can be represented by D^I, and the market orders from the uninformed traders collectively can be referred to as D^U, which is a normally distributed random variable independent of V with mean zero and variance σ_u^2. If D^U is positive, then uninformed traders are buying on net. Conversely, uninformed traders are selling the asset on net, if D^U is negative. Though the informed trader knows V, he does not know D^U prior to submitting his orders. Effectively this precludes the informed trader from conditioning on the market-clearing price, as it is usual in a rational expectations model.

Once receiving these orders, the auctioneer determines P, the market clearing price. Kyle assumes free entry into the auctioneering market and therefore the auctioneer has no monopoly power, so that he earns zero profits and P is determined by the following equation:

$$P = E[V|D^I + D^U] .$$

To arrive at a value for P, the auctioneer only takes into account the sum of the orders issued by the informed trader and the uninformed traders: $D^I + D^U$. P depends on the sum of the orders because he cannot differentiate between

the orders issued by the informed trader from the rest. Note that D^U is an exogenous variable, but D^I depends on the informed trader's trading strategy. The informed trader knows that his order has some effect on the price created by the auctioneer. Since he is risk neutral, the informed trader will seek to maximize his expected profit. He accomplishes this goal by considering each possible value of V and choosing the value of D^I that maximizes:

$$E[D^I(V-P)|V]\,.$$

These two equations illustrate that the auctioneer's strategy for setting the asset's price depends on D^I while the informed trader's strategy for setting D^I depends on his perceived effect of D^I on P.

Kyle first conjectures general functions for the pricing rule and the informed trader's demand, then he solves for the parameters assuming the informed trader maximizes his profits conditioning on his information set, i. e. $D^I = \operatorname{argmax} E[D^I(V-P)|V]$ and the market maker sets prices equal to $P = E[V|D^I + D^U]$.

Although the proof is not shown here, in equilibrium the market maker will choose a price such that

$$\mathrm{P} = \lambda(\mathrm{D}^I + \mathrm{D}^U)$$

and the informed trader will choose D^I such that

$$\mathrm{D}^I = \beta V$$

where λ and β are positive coefficients that solely depend on σ_v^2, the variance of V, the normally distributed random variable for the asset's payoff, and σ_u^2 the variance of D^U, the normally distributed random variable for the uninformed traders' orders. The exact expressions (not derived here) for λ and β are:

$$\lambda = \frac{1}{2}\sqrt{\frac{\sigma_v^2}{\sigma_u^2}}$$

$$\beta = \sqrt{\frac{\sigma_u^2}{\sigma_v^2}}\,.$$

If λ has a high value, then order flow has a high impact on prices, and we say that the particular asset is not very liquid. B, on the other hand, is rather low, which is interpreted as informed traders issuing less aggressive orders in an effort to minimize the impact of their own trades on price.

Empirical Market Microstructure

As transaction-by-transaction or high frequency data from a variety of sources has become available, empirical market microstructure has grown extensively. Most papers use high frequency data to predict transaction costs, estimate limit-order book models for intraday trading strategies, and estimate the liquidity of the market. There are a few papers, though, that do not estimate market microstructure models per se, but use high frequency data to answer questions relevant to the asset pricing field, corporate finance field, and economics in general. For example, Andersen et al. [6] use intraday data to obtain better measures of the volatility of asset prices, Chen, Goldstein and Jiang [13] estimate the PIN measure to answer questions relevant to corporate finance, and Andersen et al. [8] and [7] use intraday data to better identify the effect macroeconomic news announcements have on asset prices.

In what follows we describe the most commonly used empirical estimations of liquidity or adverse selection costs. The most general measure of adverse selection costs that does not assume a particular economic model is Hasbrouck [27]. He assumes the quote midpoint is the sum of two unobservable components,

$$q_t = m_t + s_t$$

where m_t is the efficient price, i. e., the expected security value conditional on all time-t public information, and s_t is a residual term that is assumed to incorporate transient microstructure effects such as inventory control effects, price discreteness, and other influences that cause the observed midquote to temporarily deviate from the efficient price. As such Hasbrouck [27] further assumes that $\mathrm{E}[\mathrm{s}_t] = 0$ and that it is covariance stationary which implies that microstructure imperfections do not cumulate over time, i. e., $\mathrm{E}_t[\mathrm{s}_{t+k}] \to \mathrm{E}[\mathrm{s}_{t+k}] = 0$ as $\to \infty$. The efficient price evolves as a random walk,

$$m_t = m_{t-1} + w_t \tag{1}$$

where $\mathrm{E}[\mathrm{w}_t] = 0$, $\mathrm{E}[\mathrm{w}_t^2] = \sigma_w^2$, $\mathrm{E}[\mathrm{w}_t\mathrm{w}_\tau] = 0$ for $t \neq \tau$ and w_t is also covariance stationary. The innovations, w_t, reflect updates to the public information set including the most recent trade. The market's signal of private information is the current trade innovation defined as $x_t - \mathrm{E}[x_t|\Phi_{t-1}]$, where Φ_{t-1} is the public information set at time $t-1$. The impact of the trade innovation on the efficient price innovation is $\mathrm{E}[\mathrm{w}_t|x_t - \mathrm{E}[x_t|\Phi_{t-1}]]$. Hence, two measures of information asymmetry, or trade informativeness, that Hasbrouck [27] proposes are:

$$\mathrm{Var}(E[w_t|x_t - E[x_t|\Phi_{t-1}]]) = \sigma_{w,x}^2$$

an absolute measure of trade informativeness and

$$R_w^2 = \frac{\mathrm{Var}(E[w_t|x_t - E[x_t|\Phi_{t-1}]])}{\mathrm{Var}(w_t)} = \frac{\sigma_{w,x}^2}{\sigma_w^2}$$

a relative measure of trade informativeness. The random walk decomposition, Eq. (1), on which these measures are based is unobservable. However, we can estimate $\sigma^2_{w,x}$ and σ^2_w using a vector autoregressive (VAR) model,

$$r_t = \sum_{i=1}^{\infty} a_i r_{t-i} + \sum_{i=0}^{\infty} b_i x_{t-i} + v_{1,t}$$
$$x_t = \sum_{i=1}^{\infty} c_i r_{t-i} + \sum_{i=0}^{\infty} d_i x_{t-i} + v_{2,t}$$

where $r_t = q_t - q_{t-1}$ is the change in the quote midpoint, and x_t is an indicator variable that takes values $\{-1, +1\}$ whether the trade was seller-initiated or buyer-initiated according to the Lee and Ready [33] algorithm. Some papers also consider taking signed volume (number of transactions times shares traded) rather than signed transactions, but empirical evidence shows that what is important is the number of transactions not the number of shares traded.

Hasbrouck [27] estimates the VAR system using OLS. Wold's representation theorem states that any covariance-stationary process possesses a vector moving average (VMA) representation of infinite order, i. e. $\{r_t, x_t\}$ can be written as an infinite distributed lag of white noise, called the Wold representation or VMA. The minimum and maximum daily number of transactions among all the equities varies greatly, so the researcher has to set truncation points for each individual stock separately. Rather than use the Akaike and SIC information criteria to determine the optimal lag length, the purpose of the VAR estimation above is to get rid off all serial correlation. Once the lag lengths are set we can estimate the following VMA representation:

$$r_t = \sum_{t=1}^{N} a_i^* v_{1,t-1} + \sum_{t=0}^{N} b_i^* v_{2,t-1}$$
$$x_t = \sum_{t=1}^{N} c_i^* v_{1,t-1} + \sum_{t=0}^{N} \mathrm{d}_i^* v_{2,t-1} \,.$$

Hence the trade-correlated component of the variance is equal to

$$\hat{\sigma}^2_{w,x} = \left(\sum_{t=0}^{N} b_i^*\right) \Omega \left(\sum_{t=0}^{N} b_i^{*\prime}\right) + \left(1 + \sum_{t=1}^{N} a_i^*\right)^2 \sigma_1^2 \,,$$

where $\Omega = \mathrm{Var}(v_{1,t}, v_{2,t})$ and $\sigma_1^2 = \mathrm{Var}(v_{1,t})$ the variance of the random-walk component is

$$\hat{\sigma}^2_{w} = \left(\sum_{t=0}^{N} b_i^*\right) \Omega \left(\sum_{t=0}^{N} b_i^{*\prime}\right) \,.$$

Some Estimation Considerations

The VAR and VMA systems described above are not standard autoregressive models, in the sense that the index t is not a wall-clock index, but an event index, i. e., it is incremented whenever a trade occurs or a quote is revised. The choice between an event index and a wall-clock index depends on the goals of the analysis. If the analysis involves a single security, an event index is better than a wall-clock index because the process is more likely to be covariance stationary in event time than in wall-clock Hasbrouck (see p. 90 in [29]). However, when conducting a cross-sectional analysis or estimating a pooled regression, comparability across securities becomes the dominant consideration and one may want to adopt a wall-clock time in estimating the above equations.

Hasbrouck (see p. 39 in [29]) also points out that the overnight return will almost certainly have different properties than the intraday return and he suggests that one should drop the overnight return.

All told, researchers that use Hasbrouck's measure of adverse selection costs to test important economic hypothesis should feel very uncomfortable if their results depended on the way they estimate the VAR equations. As a robustness check researchers should estimate the VAR using different specifications, i. e., wall-clock time as opposed to event-time indexes, and researchers should sample quotes at different frequencies.

Madhavan, Richardson and Roomans Model

Similar to Hasbrouck [27], Madhavan, Richardson and Roomans [38] model the efficient price, m, as a random walk, in contrast they include an order flow innovation term,

$$m_t = m_{t-1} + \theta(x_t - E[x_t|x_{t-1}]) + \varepsilon_t \tag{2}$$

where θ measures the permanent price impact of order flow and ε_t is the public information innovation. The transaction price, p, is equal to the efficient price plus a stochastic rounding error term, ξ, and a market makers' cost per share of supplying liquidity, ϕ, i. e. compensation for order processing costs, inventory costs etc.

$$p_t = m_t + \phi x_t + \xi_t \tag{3}$$

Combining Eq. (2) and Eq. (3) we obtain,

$$m_t = p_t - p_{t-1} - (\phi + \theta)x_t + (\phi + \rho\theta)x_{t-1}$$

where ρ is the first-order autocorrelation of the signed

trade variable, x_t. Then, the measure of permanent price impact, θ, alongside the temporary price impact of order flow, ϕ, the autocorrelation of signed trades, ρ, the unconditional probability that a transaction occurs within the quoted spread λ, and a constant, β, can be estimated using GMM applied to the following moment conditions:

$$E\begin{pmatrix} x_t x_{t-1} - x_{t-1}^2 \rho \\ |x_t| - (1-\lambda) \\ m_t - \beta \\ (m_t - \beta) x_t \\ (m_t - \beta) x_{t-1} \end{pmatrix} = 0$$

Future Directions

Hasbrouck [29] on page 7 lists a few outstanding significant questions in market microstructure. To this list we add two particularly important issues. First, the recent availability of good quality high frequency data has made it possible for researchers to answer a wide range of questions. This new data, though, also raises questions. In our opinion, it is imperative for researchers to determine under what circumstances more data is better. Some papers in the *realized volatility* literature have made headway in this direction by determining optimal sampling frequencies to estimate the volatility of assets with different liquidity. Future research should investigate what is the optimal frequency in estimating adverse selection costs and in event studies – studies that investigate the impact of public announcements on prices and trading in the hours surrounding its release. Second, most empirical and theoretical studies assume that trades affect prices, but prices do not affect trades (see, for example, Hasbrouck's VAR specification). Theory provides means of understanding why causality runs from trades to prices – trades are correlated with private information, so that trades cause asset price changes, with the underlying private information being the primitive cause. However, a more realistic setting is that in which there are heterogeneous beliefs and prices partially reveal other agent's information so that there is a learning process. Future research should relax the assumed exogeneity of trades. Finally, two productive areas of research are (i) the investigation of microstructure issues in fixed income markets, and (ii) studies that link microstructure to other areas in finance such as asset pricing and corporate finance.

Readings

Various economists have written books and articles about the field of market microstructure. This article is a short survey and here we compile an non-exhaustive list of publications that provide more comprehensive reviews of the literature: [12,25,29,35,37,41], and [42]. Martin Evans and Richard K Lyons have also written a useful manuscript entitled "Frequently Asked Questions About the Micro Approach to FX," even though its main focus is on foreign exchange markets it is also applicable to the market microstructure literature in general.

Bibliography

1. Akerlof GA (1970) The market for lemons: Quality uncertainty and the market mechanism. Q J Econ 84:488–500
2. Amihud Y, Mendelson H (1980) Dealership market: Market making with inventory. J Financ Econ 8:31–53
3. Amihud Y, Mendelson H (1987) Trading mechanisms and stock returns: An empirical investigation. J Financ 42:533–553
4. Amihud Y, Mendelson H (1991) Volatility, efficiency and trading: Evidence from the Japanese stock market. J Financ 46:1765–1790
5. Amihud Y, Mendelson H, Murgia M (1990) Stock market microstructure and return volatility: Evidence from Italy. J Bank Financ 14:423–440
6. Andersen T, Bollerslev T, Diebold FX, Labys P (2003) Modelling and forecasting realized volatility. Econometrica 71:579–626
7. Andersen T, Bollerslev T, Diebold FX, Vega C (2003) Micro effects of macro announcements: Real-time price discovery in foreign exchange. Am Econ Rev 93:38–62
8. Andersen T, Bollerslev T, Diebold FX, Vega C (2007) Real-time price discovery in stock, bond, and foreign exchange markets. J Int Econ 73:251–277
9. Bagehot W (pseudonum for Jack Treynor) (1971) The only game in town. Financ Analysts J 27:12–14
10. Boehmer, Saar, Yu (2004) Lifting the veil: An analysis of pretrade transparency at the NYSE. J Financ 60:783–815
11. Brunnermeier MK (2001) Asset pricing under asymmetric information. Oxford University Press, Oxford
12. Bruno B, Glosten LR, Spatt C (2005) Market microstructure: A survey of microfundations, empirical results, and policy implications. J Financ Mark 8:217–264
13. Chen Q, Goldstein I, Jiang W (2007) Price Informativeness and Investment Sensitivity to Stock Price. Rev Financ Stud 20:619–650
14. Copeland T, Galai D (1983) Information effects and the bid-ask spread. J Financ 38:1457–1469
15. Diamond DW, Verrecchia RE (1991) Liquidity and the cost of capital. J Financ 46:1325–1359
16. Easley D, O'Hara M (1992) Time and the process of security price adjustment. J Financ 47:577–604
17. Easley D, Kiefer MN, O'Hara M, Paperman JB (1996) Liquidity, information, and infrequently traded stocks. J Financ 51:1405–1436
18. Easley D, Hvidkjaer S, O'Hara M (2002) Is information risk a determinant of asset returns? J Financ 57:2185–2221
19. Easley D, Engle RF, O'Hara M, Wu L (2008) Time varying arrival rates of informed and uninformed trades. J Financ Econ 6:171–207
20. Ellis K, Michaely R, O'Hara M (2000) The accuracy of trade classi-

fication rules: Evidence from NASDAQ. J Financ Quant Analysis 35:529–551
21. Fabozzi FJ, Fleming MJ (2004) In: Fabozzi FJ (ed) US treasury and agency securities, 6th edn. McGraw Hill, pp 175–196
22. Garman (1976) Market microstructure. J Financ Econ 3:257–275
23. Glosten L, Milgrom P (1985) Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders. J Financ Econ 13:71–100
24. Goettler R, Parlour C, Rajan U (2007) Microstructure effects and asset pricing. Working Paper, UC Berkeley
25. Goodhart AE, O'Hara M (1997) High frequency data in financial markets: Issues and applications. J Empir Financ 4:74–114
26. Grossman S, Miller M (1988) Liquidity and market structure. J Financ 43:617–637
27. Hasbrouck J (1991) The summary informativeness of stock trades: An econometric analysis. Rev Financ Stud 4:571–595
28. Hasbrouck J (1992) Using the TORQ database. NYU Working Paper
29. Hasbrouck J (2007) Empirical market microstructure: The institutions, economics, and econometrics of securities trading. Oxford University Press, New York
30. Hendershott T, Seasholes M (2007) Market maker inventories and stock prices. Am Econ Rev Pap Proc 97:210–214
31. Ho T Stoll H (1983) The dynamics of dealer markets under competition. J Financ 38:1053–1074
32. Jensen MC (1978) Some anomalous evidence regarding market efficiency. J Financ Econ 6:95–102
33. Lee MC, Ready MJ (1991) Inferring trade direction from intraday data. J Financ 46:733–746
34. Lee CM, Radhakrishna B (2000) Inferring investor behavior: Evidence from TORQ data. J Financ Mark 3:83–111
35. Lyons R (2001) The microstructure approach to exchange rates. MIT Press, Cambridge
36. Madhavan (1992) Trading mechanisms in securities markets. J Financ 47:607–642
37. Madhavan A (2000) Market microstructure: A survey. J Financ Mark 3:205–258
38. Madhavan A, Richardson M, Roomans M (1997) Why do security prices change? A transaction-level analysis of NYSE stocks. Rev Financ Stud 10:1035–1064
39. Milgrom P, Stokey N (1982) Information, trade, and common knowledge. J Econ Theor 26:17–27
40. Odders-White E (2000) On the occurrence and consequences of inaccurate trade classification. J Financ Mark 3:259–286
41. O'Hara M (1995) Market microstructure theory. Blackwell Publishers, Oxford
42. Parlour CA, Seppi DJ (2008) Limit order markets: A survey. In: Boot AWA, Thakor AV (eds) Handbook of financial intermediation and banking
43. Schwert WG (2003) Anomalies and market efficiency. In: Harris M, Stulz RM, Constantinides GM (eds) Handbook of the economics of finance, vol 15. North-Holland
44. Smidt (1971) The road to an efficient stock market. Financ Analysts J 27:18–20; pp 64–69
45. Stoll H, Whaley R (1990) Stock market structure and volatility. Rev Financ Stud 3:37–71
46. Vega C (2005) Stock price reaction to public and private information. J Financ Econ 82:103–133
47. Viswanathan S, Wang J (2004) Inter-dealer trading in financial markets. J Bus 77:49–75

Market Microstructure, Foreign Exchange

Carol Osler
Brandeis International Business School, Brandeis University, Waltham, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Institutional Structure
Intraday Dynamics
Returns and Volatility
Order Flow and Exchange Rates, Part I: Liquidity and Inventories
Order Flow and Exchange Rates, Part II: Information
Price Discovery in Foreign Exchange
Summary and Future Directions
Bibliography

Glossary

Barrier options Options that either come into existence or disappear when exchange rates cross pre-specified levels. Barriers can be triggered by price rises or declines and reaching a barrier can either extinguish or create an option. An "up-and-out call," for example, is a call option that disappears if the exchange rate rises above a certain level. A "down-and-in put," by contrast, is created if the exchange rate falls to a certain level.

Bid-ask spread The difference between the best (lowest) price at which one can buy an asset (the ask) and the best (highest) price at which one can sell it (the bid). In quote-driven markets both sides of the spread are set by one dealer. In order-driven markets, the "best bid and the best offer" (BBO) are likely to be set by different dealers at any point in time.

Brokers Intermediaries in the interbank foreign exchange market that match banks willing to buy with banks willing to sell at a given price. Two electronic brokerages – EBS (Electronic Broking Service) and Reuters – now dominate interbank trading in the major currencies. In other currencies voice brokers still play an important role.

Call markets Financial markets that clear periodically rather than continuously. During a specified time interval, agents submit orders listing how much they are willing to buy or sell at various prices. At the end of the

interval a single price is chosen at which all trades will take place. The price is chosen to maximize the amount traded and is essentially the intersection of the supply and demand curves revealed by the submitted orders.

Clearing The administration process that ensures an individual trade actually takes place. The amounts and direction are confirmed by both parties and bank account information is exchanged.

Corporate (or commercial) customers One of the two main groups of end-users in the foreign exchange market. Includes large multinational corporations, middle-market corporations, and small corporations. Their demand is driven almost entirely by international trade in goods and services, since traders at these firms are typically not permitted to speculate spot and forward markets.

Covered interest arbitrage A form of riskless arbitrage involving the spot market, the forward market, and domestic and foreign deposits.

Dealership market See Quote-driven markets.

Delta-hedge A delta-hedge is designed to minimize first-order price risk in a given position. That is, small price changes should change the agent's overall position by only a minimal amount (ideally zero). A delta-hedge gets its name from an option's "delta," which is the first derivative of the option's price with respect to the price of the underlying asset. To delta-hedge a long call (put) option position, the agent takes a short (long) position in the underlying asset equal in size to the option's delta times the notional value of the option.

Expandable limit order An order whose quantity can be expanded if it is crossed with a market order for a larger quantity.

Financial customers One of the two main groups of end-users in the foreign exchange market. Includes hedge funds and other highly-leveraged investors, institutional investors such as mutual funds, pension funds, and endowments, multilateral financial institutions such as the World Bank or the IMF, broker-dealers, and regional banks.

Feedback trading The practice of trading in response to past returns. Positive-feedback trading refers to buying (selling) after positive (negative) returns. Negative-feedback trading refers to selling (buying) after positive (negative) returns.

Foreign exchange dealers Intermediaries in the foreign exchange market who stand ready, during trading hours, to provide liquidity to customers and other dealers by buying or selling currency. Salespeople manage relationships with clients; interbank traders manage the inventory generated by customer sales, and also speculate on an extremely high-frequency basis, by trading with other banks; proprietary traders speculate on a lower-frequency basis in currency and other markets.

Forward market Currencies traded in forward markets settle after more than two trading days (and infrequently after less than two trading days).

FX Foreign Exchange.

Limit order See "Order-driven markets."

Long position A long position arises when an agent owns an asset outright.

Market order See "Order-driven markets."

Order flow Buy-initiated transactions minus sell-initiated transactions over a given period. Since customers are always the initiators, their order flow is just customer purchases minus customer sales. In the interdealer market, a dealer initiates a trade if s/he places a market order with a broker or if the dealer calls out to another dealer.

Order-driven markets Also known as "limit-order markets." Asset markets in which participants can both supply liquidity or demand it, as they choose. Liquidity suppliers place limit orders, which specify an amount the agent is willing to trade, the direction, and the worst acceptable price. A limit buy order in the euro-dollar market, for example, might specify that the agent is willing to buy up to $2 million at $1.2345 or less. These limit orders are placed into a "limit-order book," where they remain until executed or canceled. Agents demanding liquidity place "market" orders, which state that the agent wishes to trade a specified amount immediately at whatever price is required to fulfill the trade. Market orders are executed against limit orders in the book, beginning with the best-priced limit order and, if necessary, moving to limit orders with successively less attractive prices. The foreign exchange interdealer markets for major currencies are dominated by two electronic limit-order markets, one run by EBS and the other run by Reuters.

Overconfidence A human tendency to have more confidence in oneself than is justified. Humans tend to overestimate their own personal and professional success ("hubris") and that they overestimate the accuracy of their judgments ("miscalibration").

Over-the-counter market See quote-driven market.

Picking-off risk The risk that a limit order will be executed against a better informed trader, leaving the limit-order trade with a loss.

Price-contingent orders Orders that instruct a dealer to transact a specified amount at market prices once a currency has traded at a pre-specified price. There

are two types: stop-loss orders and take-profit orders. Stop-loss orders instruct the dealer to sell (buy) if the rate falls (rises) to the trigger rate. Take-profit orders instruct the dealer to sell (buy) if the price rises (falls) to the trigger rate.

Quote-driven markets Also known as "dealership markets" or "over-the-counter markets." An asset market in which dealers provide immediate liquidity to those needing it. During trading hours the dealers commit to trade at any time but at prices they quote. The price at which they are willing to buy, the "bid," is always no greater – and usually lower – than the price at which they are willing to sell, the "ask." Foreign exchange dealers transact with end-users in a quote-driven market.

Settlement The process by which funds actually change hands in the amounts and direction indicated by a trade.

Short position A short position arises when an agent sells an asset, possibly before actually owning the asset. A "short position in euros" could arise if a dealer starts with zero inventory and then sells euros. The dealer could keep the short euro inventory overnight, but will typically close the position out at the end of the trading day by buying the equivalent amount of euros. Note that the overall bank will not have a negative inventory position, since the bank maintains balances in every currency it trades. Someone "short euros in the forward market" would have entered into a forward contract to sell euros in the future.

Slippage The concurrent effect of a given trade on price.

Stop-loss orders See "Price-contingent orders."

Spot market Currencies traded in the spot market settle after two trading days (except for transactions between the US and Canadian dollars).

Swaps A swap in the foreign exchange market is analogous to a repo in the money market. One counterparty agrees to buy currency A in exchange for currency B from another counterparty in the spot market, and simultaneously agrees to sell currency A back to the same counterparty, and buy back currency B, at a future date. The spot transaction is at the spot rate, the forward transaction is at the forward rate.

Take-profit orders See "Price-contingent orders."

Technical Trading Trading based on technical analysis, an approach to forecasting asset-price movements that relies exclusively on historical prices and trading volume. In foreign exchange, the absence of frequent volume figures limits the information basis to past prices. Notably, technical forecasts do not rely on economic analysis. Nonetheless, many technical trading strategies have been demonstrated to be profitable in currency markets, even after considering transaction costs and risk.

Trading volume The value of transactions during a given time period.

Triangular arbitrage Between every three currencies A, B, and C there are three bilateral exchange rates. Triangular arbitrage is a way to make riskless profits if the A-per-B exchange rate does not equal the C-per-B exchange rate multiplied by the A-per-C exchange rate.

Definition of the Subject

"Foreign exchange microstructure" is the study of the currency trading process and high-frequency exchange-rate determination. The field is also called "the new microeconomics of exchange rates." Research in this area began in the late-1980s, when it became clear after many years of floating rates that traditional, macro-based exchange-rate models were not able to explain short-run dynamics. Research accelerated in the mid-1990s as currency trading systems became sufficiently automated to provide useful data.

Introduction

Foreign exchange microstructure research, or the study of the currency trading process, is primarily motivated by the need to understand exchange-rate dynamics at short horizons. Exchange rates are central to almost all international economic interactions – everything from international trade to speculation to exchange-rate policy. The dominant exchange-rate models of recent decades, meaning specifically the monetary model and the intertemporal optimizing models based on Obstfeld and Rogoff [149], come from macro tradition. These have some value relative to horizons of several years, but they have made little headway in explaining exchange rate dynamics at shorter horizons [69,116,134]. Shorter horizons are arguably of greater practical relevance.

As elucidated by Kuhn in his seminal analysis of scientific progress (1970), the emergence of major anomalies typically leads researchers to seek an alternative paradigm. Currency microstructure research embodies the search for a new paradigm for short-run exchange-rate dynamics.

The search for an alternative paradigm has focused on the currency trading process for a number of reasons. First, it is widely held that macroeconomic models are enhanced by rigorous "microfoundations" in which agent behavior is carefully and accurately represented. A rigorous microfoundation for exchange rates will require a thorough understanding of the currency trading process.

Researchers are also motivated to study currency trading by evident contradictions between the way currency markets actually work and the way exchange-rate determination is represented in macro-based models. As Charles Goodhart remarked of his time as adviser at the Bank of England, "I could not help but observe that some of the features of the foreign exchange … market did not seem to tally closely with current theory …" (p. 437 in [81]). To others with first-hand experience of the trading world, it seemed natural "to ask whether [the] empirical problems of the standard exchange-rate models … might be solved if the structure of foreign exchange markets was to be specified in a more realistic fashion" (p. 3 in [72]).

The emergence of currency-market research in recent years also reflects a confluence of forces within microstructure. By the mid-1990s, microstructure researchers had studied equity trading for over a decade, thereby creating a foundation of theory and a tradition of rigorous analysis. Meanwhile, technological advances at foreign-exchange dealing banks made it possible to access high-frequency transactions data. Currency markets – huge and hugely influential – were a logical new frontier for microstructure research.

Currency microstructure research – like all microstructure research – embodies the conviction that economic analysis should be based solidly on evidence. As articulated by Charles Goodhart, arguably the founder of this discipline, "economists cannot just rely on assumption and hypotheses about how speculators and other market agents may operate in theory, but should examine how they work in practice, by first-hand study of such markets" (p. 437 in [81]). Most papers in this area are empirical, and those that include theory almost always confront the theory with the data. The literature includes quite a few dealer surveys, reflecting a widespread appreciation of practitioner input. This survey, like the literature, emphasizes evidence.

Institutional Structure

This section describes the institutional structure of the foreign exchange market.

Basics

Foreign exchange trading is dispersed throughout the day and around the world. Active trading begins early in Asia, continues in Europe, peaks when both London and New York are open, and finally tapers off after London traders leave for the day. There is an "overnight" period during which trading is relatively thin, but it lasts only the few hours between the end of trading in London (around 19 GMT) and early trading in Sydney (around 22 GMT). In terms of geography, currency trading takes place in almost every big major city around the world, though there are major trading centers. These major centers are Singapore, Sydney, and Tokyo in Asia, London in Europe, and New York in North America.

Foreign exchange trading is an intensely competitive business. Price is one dimension of competition, but there are many others. When it evaluates trading institutions each year, *Euromoney* considers their pricing consistency, strategies and ideas for trading in options, and innovative hedging solutions [55]. Customer relations are also critically important. As in many industries, good customer relations are fostered by personal attention from salespeople and by perks for good customers, such as sports tickets and elegant feasts.

Unlike trading in stocks, bonds, and derivatives, trading in currency markets is essentially unregulated. There is no government-backed authority to define acceptable trading practices, nor is there a self-regulating body. Local banking authorities are limited to regulating the structure of trading operations: they typically require, for example, that clearing and settlement are administratively separate from trading. Any attempt to regulate trading itself, however, would encourage dealers to move elsewhere, an undesirable outcome since foreign exchange is an attractive industry – it pays high salaries and generates little pollution. In the absence of regulation, certain practices that are explicitly illegal in other markets, such as front-running, are not only legal but common in foreign exchange.

Market Size Spot and forward trading volume in all currencies is worth around $1.4 trillion per day [9]. If foreign exchange swap contracts are included, daily trading is roughly twice as large, at $3.2 trillion. By either figure, foreign exchange is the largest market in the world. Trading on the New York Stock Exchange (NYSE), for example, is on the order of $0.050 trillion per day [145], while daily trading in the US Treasury market, possibly the world's second-largest market, is on the order of $0.20 trillion [67]. Spot and forward trading, on which FX microstructure research has consistently focused, has grown rapidly for many years – average yearly growth since 1992 has been nine percent, and since 2004 has been 18 percent.

The vast bulk of foreign exchange trading involves fewer than ten currencies. The US dollar is traded most actively [9] due to its role as the market's "vehicle currency": to exchange almost any non-dollar currency for any other requires one to convert the first currency into dollars and then trade out of dollars into the second currency. The value of US dollars traded in spot and forward

markets is roughly $1.2 trillion per day, over 86 percent of total traded value. Of course, two currencies are involved in every transaction so the total traded value every day is twice the day's trading volume. The euro accounts for 37 percent of all trading, a staggering $518 billion per day. The yen and the UK pound each account for a further sixteen percent of traded value. The next tier of currencies, comprising the Swiss franc, the Australian dollar and the Canadian dollar, accounts for eighteen percent of traded value. The remaining 150 or so of the world's convertible currencies account for merely thirty percent of traded value.

Only the dollar, the euro, and the yen are liquid throughout the trading day. Liquidity in most other currencies is concentrated during locally-relevant segments of the day. The Swedish krone, for example, is liquid only during European trading hours.

Quotation Conventions Each exchange rate is quoted according to market convention: dollar-yen is quoted as yen per dollar, euro-dollar is quoted as dollars per euro, etc. Trade sizes are always measured in units of the base (denominator) currency and the price is set in terms of the numerator currency. In euro-dollar, for example, where the euro is the base currency, a customer asking to trade "ten million" would be understood to mean ten million euros and the dealer's quotes would be understood to be dollars per euro. The minimum tick size is usually on the order of one basis point, though it is technically one "pip," meaning one unit of the fifth significant digit for the exchange rate as conventionally quoted. Examples will be more helpful: in euro-dollar, where the exchange rate is currently around $1.5000, one tick is $0.0001; for dollar-yen, where current exchange rates are roughly ¥ 110.00/$, one tick is ¥ 0.01.

The average trade size is on the order of $3 million [18]; trades of $50,000 or less are considered "tiny." Thus the average foreign exchange trade is roughly the same size as normal "block" (large) trades on the NYSE [125], which makes it large relative to the overall average NYSE trade. The average foreign exchange trade is smaller, however, than the average trade in the US Treasury market, where average interdealer trades vary from $6 to $22 million depending on maturity [67].

A Two-Tiered Market

The foreign exchange market has two segments or "tiers." In the first tier, dealers trade exclusively with customers. In the second tier, dealers trade primarily with each other. The interdealer market forms the market's core in the sense that customer prices are all based on the best available interdealer prices.

Interdealer trading in spot and forward markets now accounts for 38 percent of all trading [9]. This is down sharply from its 57 percent share in 1998, a change often ascribed to rapid consolidation in the industry. The current share is comparable to the share of interdealer trading on the London Stock Exchange, which was most recently estimated to be between 25 and 35 percent [163]. It is lower, however, than the share of interdealer trading in the US Treasury market, which was 68 percent in October 2007 [66].

The Customer Market The customer foreign exchange market is quote-driven, meaning that liquidity is provided by professional market makers. As in most such markets, currency dealers are under no formal obligation to provide liquidity, unlike specialists on the NYSE. Failing to provide liquidity on demand, however, could be costly to a dealer's reputation so dealers are extremely reliable. The market functioned smoothly even during the crisis of September 11, 2001. Spreads widened, as would be expected given the heightened uncertainty, but market makers stayed at their desks and trading continued uninterrupted [135].

The customer market is fairly opaque. Quotes and transactions are the private information of the two parties involved, the customer and the dealer. Unlike stock and bond markets, which publish trading volume daily, aggregate figures for customer trading volume are published only once every three years e. g.[9]. The lack of transparency is intensified by the tendency for large customer trades, meaning those over around $25 million, to be split into multiple smaller trades. Splitting trades, which is a way to minimize market impact and thus execution costs [16], also characterizes the London Stock Exchange [165], among other markets. Trade-splitting makes it more difficult for a dealer to know how much a customer actually intends to trade. Dealers like to know when customers are trading large amounts, since large trades move the market.

Dealers divide their customers into two main groups, and structure their sales force accordingly. The first group, financial customers, is dominated by asset managers but also includes non-dealing banks, central banks, and multilateral financial institutions. The asset managers, in turn, are divided into "leveraged investors," such as hedge funds and commodity trading associations (CTAs), and "real money funds," such as mutual funds, pension funds, and endowments. Financial customers account for 40 percent of foreign exchange trading [9], sharply higher than their 22 percent share in 1998 [9].

The second group of customers, referred to as "corporates," are commercial firms that purchase currency as part of ongoing real production activities or for financial purposes such as dividend payments or foreign direct investment. The share of such commercial trading has been steady at roughly twenty-percent for a decade [9]. Commercial customers tend to be the mainstay of profitability for smaller banks [136]. Financial customers, by contrast, tend to make bigger transactions and thus gravitate to bigger banks [154].

The customers listed above are all institutions. Unlike equity markets, where the trading of individuals for their own account can account for half of all trading, retail trading has historically been tiny in foreign exchange. The participation of individuals has been discouraged by large average trade sizes and by the need to establish lines of credit with dealing banks.

Though customer trading has historically been carried out over the telephone, trading over electronic communication networks is growing rapidly, spurred by the advent of new technologies [11]. Formal figures are not available, but dealers estimate informally that these new networks now account for over one fifth of all customer transactions. Major dealers run single-bank proprietary networks through which they are connected to individual customers. The biggest networks, however, are managed independently. Some of these multi-bank e-portals, such as FXAll, permit customers to get multiple quotes simultaneously. FXAll has appealed primarily to commercial customers, which have historically paid relatively wide spreads on average (as discussed later), since it has brought them enhanced pre-trade transparency, intensified competition among dealers and, according to dealers, smaller spreads. Other multi-bank e-portals, such as FXConnect or Hotspot FXi, focus on financial customers and are valued because they permit "straight-through processing" (STP), meaning fully automated clearing and settlement. STP handles back office functions far more efficiently than the traditional manual approach in part because it reduces the opportunity for human error. Another type of network, such as Oanda.com, target individuals trading for their own account, permitting them to trade with no more than a Paypal account. Though such retail trading has grown rapidly in the current century, dealers report that it does not yet affect market dynamics.

The Interdealer Market In the foreign exchange interbank market there are no designated liquidity providers. At every moment a dealing bank can choose whether to supply liquidity or demand it. A dealer needing liquidity can, of course, call another dealer and request a quote. Until the mid-1990s such "direct dealing" accounted for roughly half of all interdealer trading [36], while the other half of interdealer trading was handled by voice brokers – essentially limit-order markets in which people match the orders. During this period the best indication of the market price was often indicative quotes posted on Reuters' "FXFX" screen.

The structure of interdealer trading changed dramatically after the introduction of electronic brokerages in 1992. In the major currencies, electronic brokerages not only took over from the voice brokers but also gained market share relative to direct dealing. Electronic Broking Service (EBS) now dominates in euro and yen while Reuters, the other major electronic brokerage, dominates in sterling. As the electronic brokerages took over, their best posted bid and offer quotes became the benchmark for market prices. By the end of the 1990s, voice brokers were important only in the "exotic" (relatively illiquid) currencies for which electronic brokers are unavailable. The speed of this transition reflects the intensity of competition in this market.

EBS and Reuters share a common, uncomplicated structure. Standard price-time priority applies. Hidden orders are not permitted. Limit orders are not expandable. Orders must be for integer amounts (in millions). Trading is anonymous in the sense that a counterparty's identity is revealed only when a trade is concluded. Dealers pay commissions on limit orders as well as market orders, though the commission on limit orders is smaller.

These markets have moderate pre- and post-trade transparency relative to most other limit-order markets. With respect to pre-trade information, price information is limited to the five best bid and offer quotes, and depth information is limited to total depth at the quotes unless it exceeds $20 million (which it usually does during active trading hours). The only post-trade information is a listing of transaction prices. The exchanges do not publish any trading volume figures.

Automated (program) trading on the electronic brokerages was introduced in 2004. Trading was restricted to dealers until 2006, but now certain hedge funds are permitted to trade on EBS. These shifts are reported to be a major source of the surge in trading between dealers and their financial customers since 2004 [9].

Objectives and Constraints

To construct exchange-rate models with well-specified microfoundations it is critical to know the objectives and constraints of major market participants. It is also critical to know the constraints that determine equilibrium.

Dealers' Objectives and Constraints Dealers are motivated by profits according to the conscious intent of their employers. Half or more of their annual compensation comes in the form of a bonus which depends heavily on their individual profits [153]. Profits are calculated daily and reviewed monthly by traders and their managers.

Dealers are constrained by position and loss limits which are, in turn, management's response to rogue trader risk, meaning the risk that traders will incur immense losses [43,81]. A single rogue trader can bring down an entire institution: Nick Leeson brought down Barings Bank in the early 1990s by losing $1.4 billion; John Rusnack brought down Allfirst Bank by losing $700 million. Such catastrophes could not occur in the absence of an information asymmetry that plagues every trading floor: management cannot know each trader's position at all times. Traders are technically required to record their profits and losses faithfully and in a timely manner, but as losses mount they sometimes resort to falsifying the trading record. Position- and loss-limits are intended to minimize the risk that losses mushroom to that point. Intraday position limits begin at around $5 million for junior traders, progress to around $50 million for proprietary traders, and can be far higher for executive managers. Data presented in Oberlechner and Osler [148] suggests that intraday limits average roughly $50 million. Overnight position limits are a fraction of intraday limits, and loss limits are a few percent of position limits.

Profit-maximization for dealers involves inventory management, speculation, and arbitrage. We review these activities in turn.

Inventory management Foreign exchange dealers manage their own individual inventory positions [18,81], tracking them in a "deal blotter" or on "position cards" [120]. Large dealers as well as small dealers typically choose to end the day "flat," meaning with zero inventory, and generally keep their inventory close to zero intraday as well. Average intraday inventory levels are $1 to $4 million in absolute value and account for less than five percent of daily trading activity [18,154]. Though these absolute levels far exceed the $0.1 million median inventory level of NYSE specialists [98], the NYSE inventories are much larger relative to daily trading (24 percent).

Dealers generally eliminate inventory positions quickly. The half-life of an inventory position is below five minutes for highly active dealers and below half an hour for less active dealers [18,155]. Fast inventory mean-reversion has also been documented for futures traders [130], but standard practice in other markets often differs markedly. On the NYSE, for example, the half-life of inventory averages over a week [127]. Even on the London Stock Exchange, which has an active interdealer market like foreign exchange, inventory half-lives average 2.5 trading days [85].

Foreign exchange dealers in the major currencies generally prefer to manage their inventory via interdealer trades, rather than waiting for customer calls. In consequence, recent studies of dealer practices find no evidence of inventory-based price shading to customers, e. g.[154]. This distinguishes currency dealers from those in some equity markets [127] and bond markets [51]. Currency dealers also do not shade prices to other dealers in response to inventory accumulation [18]. Instead, dealers wishing to eliminate inventory quickly choose more aggressive order strategies [18,154].

Speculation Foreign exchange dealers speculate actively in the interdealer market [81]. Indeed, according to a dealer cited in Cheung and Chinn [36], "[d]ealers make the majority of their profit on rate movement, not spread" (p. 447). Consistent with this, Bjønnes and Rime [18] find that speculative profits are the dominant source of dealer profitability at the good-sized bank they analyze. Dealers' speculative positions are based on information gathered from customers, from professional colleagues at other banks, and from real-time news services.

Arbitrage Some dealers also engage in arbitrage across markets, such as triangular arbitrage or covered interest arbitrage. The associated software originally just identified the arbitrage opportunities, but by now it can actually carry out the trades. Arbitrage opportunities, though typically short-lived, arise frequently and occasionally provide sizeable profits (2).

Customers' Objectives and Constraints The three main types of customers are active traders, meaning levered funds and proprietary traders; real-money funds; and commercial firms.

Active Currency Traders The objectives and constraints of active currency traders are in some ways consistent with those assigned to international investors in standard academic models. These groups are motivated by profits: proprietary traders are motivated by an annual bonus; hedge fund managers receive a share of the firm's net asset value growth in [169]. Further, their risk-taking is constrained since active currency traders, like dealers, face position limits. Notice, however, that active currency traders are not motivated by consumption and they do not care about consumption risk. Indeed, there is no reason to ex-

pect the objectives of financial market participants to be aligned with those of consumers. It is agency problems that drive a wedge between the objectives of consumers and traders in foreign exchange: the institutions that employ the traders have to align the traders' incentives with those of shareholders under conditions of asymmetric information, with the result that consumption is irrelevant. Agency problems have been shown to be of overwhelming importance in understanding financial management at corporations. It would appear risky to assume that agency problems do not exist at currency-management firms.

Active currency traders also differ, however, from the academic image of the international investor. The speculative horizons of active currency traders typically range from a day to a month – longer than a dealer's intraday horizon but still short by macro standards. Further, these traders rarely take positions in assets with fixed supplies, such as bonds or equities. Instead, they rely on forwards, other derivatives, or possibly deposits, which are in flexible supply. This seemingly simple observation may unlock a longstanding puzzle in international macro, the apparent irrelevance of bond supplies for exchange rates. Under the standard assumption that speculative agents invest in bonds (an asset with fixed supply) bond supplies should influence exchange rates. Since bonds are not widely used by active currency speculators, however, the irrelevance of bond supplies seems natural.

Common speculative strategies among active currency traders are based on (i) forward bias, (ii) anticipated trends or trend reversals, and (iii) anticipated macro news.

Real-Money Managers Most managers of real money funds do conform to the academic image of an international investor in terms of their investment horizon and their assets of choice: they take positions for a month or more and generally invest in bonds or equities. These managers do not, however, conform to that image in a separate, critical dimension: real-world real money managers generally ignore the currency component of their return. According to Taylor and Farstrup [178], who survey the currency management business,

> there are key participants in foreign exchange markets … that are not always seeking profit derived *from their currency positions*. … [I]n this category are international equity managers. While some managers factor in currency considerations as they go about picking foreign stocks, most are attempting to add value through stock, sector, and region bets rather than currency plays (p. 10, italics in original).

The decision not to forecast the currency component of returns is sometimes justified by pointing to the well-known inability of macro-based exchange-rate models to forecast more accurately than a random walk [134]. Further information about financial customers is presented in Sager and Taylor [169].

Note that all speculative positions are constrained in currency markets. In exchange-rate models this would be consistent with the assumption that speculators are risk averse. It would not, however, be consistent with the assumption that deviations from purchasing power parity or uncovered interest parity are instantaneously eliminated by infinite trading. This may help explain why macroeconomic evidence of long standing shows that these parity conditions do not hold over short-to-medium horizons.

Commercial Customers With only rare exceptions, commercial firms do not take overtly speculative positions in spot and forward foreign exchange markets. Goodhart [81] estimates that less than five percent of large corporate customers will speculate in the forward market, and dealers report that zero middle-market or small corporations speculate in that way. Indeed, many firms explicitly prohibit their trading staff – often administrators with other responsibilities besides trading – from engaging in such transactions. Rogue trader risk is one key motivation for this choice. To impede the deception that enables rogue trading, firms that permit speculation must "separate the front office from the back office," meaning they must prohibit traders from confirming or settling their own trades. This requires a separate staff to handle these functions [65]. The firms must also hire "compliance officers" to ensure that controls on the trading process are being observed faithfully (Federal Reserve Bank of New York, Best Practice 48). Since the vast majority of commercial firms need to trade only infrequently to carry out their real-side business, these heavy staffing requirements make speculative trading prohibitively expensive.

Another powerful reason why corporate customers avoid overt speculation is that it can raise corporate tax burdens. In the US, at least, profits from overtly speculative positions are accounted for differently from gains designed to offset losses on existing business exposures, with the result that speculative profits are taxed more heavily. If a treasurer wishes to speculate, s/he can do so at a lower cost by redistributing the firm's assets and liabilities around the world. Goodhart [81] lists additional reasons why corporate customers generally do not speculate in spot and forward markets.

The presence of non-financial customers provides a natural source of heterogeneity in the motivations for

currency trading. Such heterogeneity is critical for modeling asset prices, and may thus be critical for the functioning of asset markets [142,143]. When all agents are rational speculators it is hard to find reasons why speculators would trade with each other. If the price is away from its fundamental value both agents should insist on taking the profitable side of any trade, which is impossible. If the price is at its equilibrium, however, there is no profit to be gained from trading.

In the foreign exchange market, commercial firms necessarily have different trading motivations from speculators. Speculative agents primarily care about currencies as a store of value and commercial traders primarily care about currencies as a medium of exchange. Thus the existence of high trading volumes is less difficult to explain in foreign exchange than in, say, equity markets. (In bond markets, an alternative trading motivation may be provided by insurers and others engaged in duration matching.)

To generate trading volume in models of equity markets, financial modelers typically introduce "liquidity traders" or "noise traders" [22,115], typically modeled as a pure random variable and verbally assigned some motivation for trading. For liquidity traders the motivation is exogenous portfolio rebalancing; for noise traders the motivation is often speculation based on misinformation [22]. Neither motivation is fully satisfactory to the profession, however. Portfolio rebalancing is not sufficient to account for observed trading volumes and the professional preference for assuming rationality is not well-served by the noise trader concept. In foreign exchange markets, commercial traders provide rational trading partners for rational speculators.

Constraints on Exchange Rates The institutional features outlined in this section reveal a key constraint on exchange rates. On most days the amount of currency purchased by end-users must (roughly) equal the amount sold by end-users. Though dealers stand ready to provide liquidity intraday, the fact that they generally go home flat means that the dealing community, as a whole, does not provide overnight liquidity. Within a day, the net purchases of any end-user group must ultimately be absorbed by the net sales of some other end-user group. The exchange rate is presumably the mechanism that adjusts to induce end-users to supply the required liquidity.

This same explicit constraint can be found in financial markets known as "call markets" (see glossary), where a single price is chosen to match the amount bought to the amount sold. Prominent call markets include the opening markets on the NYSE and the Paris Bourse.

The very real constraint that end-user purchases equal end-user sales over a trading day differs dramatically from the exchange-rate equilibrium condition common to standard macroeconomic models. That condition is, in essence, that money demand equals money supply. The evidence does not support the relevance of aggregate money demand/supply to day-to-day exchange-rate determination [153].

Intraday Dynamics

This section provides descriptive information about trading volume, volatility, and spreads on an intraday basis.

Intraday Patterns in Volume, Volatility, and Spreads

Trading volume, volatility, and interdealer spreads all vary according to strong intraday patterns that differ in certain key respects from corresponding patterns in bond and equity markets. Figure 1a and b shows these patterns for euro-dollar and dollar-yen, based on EBS trade and quote data over the period 1999–2001 [103].

As in other markets, trading volume (measured here by the number of interbank deals) and volatility move together. As Asian trading opens (around hour 22) they both rise modestly from overnight lows, after which they follow a crude U-shape pattern during Asian trading hours and then another U-shape during the London hours. They both peak for the day as London is closing and New York traders are having lunch and then decline almost monotonically, reaching their intraday low as Asian trading opens early in the New York evening.

Some back-of-the envelope figures may help make these trading-volume patterns concrete. In Ito and Hashimoto's 1999–2001 EBS database there were roughly eight trades per minute in euro-dollar and six trades in dollar-yen [103]. Together with the seasonal patterns, this suggests that overnight interdealer trading was on the order of one or fewer trades per minute while peak trading (outside of news events) was on the order of 10 (JPY) to 25 (EUR) trades per minute. Current interdealer trading activity would be substantially larger, reflecting subsequent market growth.

Bid-ask spreads almost perfectly mirror the pattern of volume and volatility. They are highest during the overnight period, and then decline as trading surges at the Asian open. As trading and volatility follow their double-U pattern during Asian and London trading hours, spreads follow the inverse pattern: they rise-then-fall during Asian trading and then rise-then-fall once again during the London morning. After London closes, spreads rise roughly monotonically to their overnight peaks.

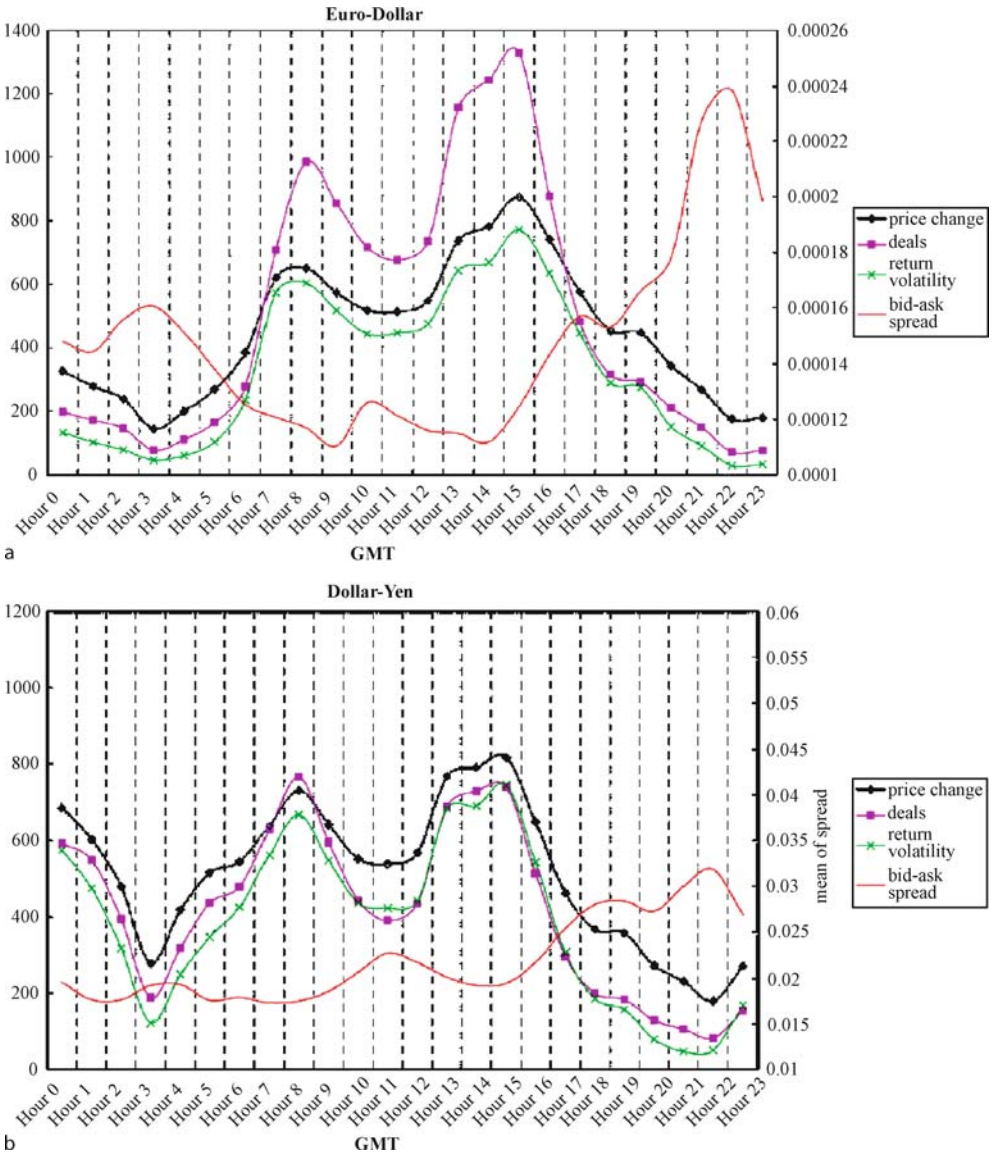

Market Microstructure, Foreign Exchange, Figure 1

Intraday Patterns for Volume, Volatility, Spreads, and the Number of Price Changes. Figures are calculated from tick-by-tick EBS trade and quote data during winter months during 1999–2001. Seasonal patterns are only slightly different in summer. (Source: [103]). Greenwich Mean Time

Conventional interdealer spreads, as reported in Cheung and Chinn [36], average three basis points in eurodollar and dollar-yen, the two most active currency pairs. In sterling-dollar and dollar-swiss, the next two most active pairs, these averaged five basis points. Dealers in both the US [36] and the UK [37] report that the dominant determinant of spreads is the market norm. One important reason spreads widen is thin trading and a hectic market. Another important reason is market uncertainty [36], which is often associated with volatility. Since volatility also increases inventory risk, it makes sense that volatility and spreads have been shown to be positively related [23,92,105].

This tendency for interdealer spreads to move inversely from volume and volatility is consistent with predictions from two conceptual frameworks. Hartmann [92] explains the relationship in terms of fixed operating costs, such as the costs of maintaining a trading floor and of

acquiring real-time information. When trading volume is high these costs can easily be covered with small spread, and vice versa, so long as the extra volume is dominated by uninformed traders. The same explanation could also apply at the intraday horizon.

Admati and Pfleiderer [1] develop an asymmetric information model consistent with some of the key properties just noted. In their model, discretionary uninformed traders (who can time their trades) choose to trade at one time since this brings low adverse selection costs to dealers and thus low spreads. The low spreads encourage informed traders to trade at the same time and the information they bring generates volatility. Overall, this model predicts that trading volume and volatility move in parallel and both move inversely with spreads, consistent with the patterns in major foreign exchange markets.

In most equity and bond markets, spreads move in parallel with trading volume and volatility, rather than inversely, with all three following an intraday (single) U-shape. Notably, a similar U-shape characterizes interdealer foreign exchange markets in smaller markets, such as Russia's electronic interdealer market for rubles, which only operate for a few hours every day [141]. In Taipei's interdealer market, which not only has fixed opening and closing times but also closes down for lunch, spreads follow a double-U-shape: they begin the day high, tumble quickly, and then rise somewhat just before lunch; after lunch they follow roughly the same pattern [76]. This contrast suggests that there is a connection between fixed trading hours and this U-shape for spreads.

Madhavan et al. [128] provide evidence that high spreads at the NYSE open reflect high adverse-selection risk, since information has accumulated overnight. High spreads at the close, by contrast, reflect high inventory risk, according to their evidence, since dealers cannot trade until the market re-opens the next morning. In less-liquid foreign exchange markets, such as those for emerging market currencies, the overnight period is relatively long and there is little overnight liquidity, so similar patterns may arise. The failure of interdealer spreads in major currencies to follow the pattern observed in equity and bond markets need not imply, however, that adverse selection is irrelevant in the interdealer markets. In the major currencies, the overnight period is short and liquid (relative to other assets), so adverse-selection risk may not rise as sharply as the market opens and inventory risk may not rise as sharply as the overnight period approaches. In this case adverse selection could be relevant but subordinate to other factors, such as Hartmann's fixed operating costs.

Weekends are a different story, since foreign exchange trading largely ceases from about 21 GMT on Fridays until 21 GMT on Sundays. The previous analysis suggests that foreign exchange spreads might be particularly wide on Monday mornings in Tokyo and Friday afternoons in New York. There is support for the first of these implications: Ito and Hashimoto [103] provide tentative evidence that spreads are indeed exceptionally wide on Monday mornings in Tokyo.

Minute-by-minute data show that volume and volatility spike sharply at certain specific times of day [12]. In the New York morning there are spikes at 8:20, 8:30, 10 and 11 am, reflecting the opening of derivatives exchanges, the release of US macro news, standard option expiration times, and the WM/Reuters fixing (at 4 pm London time; this is a price at which many banks guarantee to trade with customers), respectively. Further spikes occur at 2 pm, and 8 pm New York time, reflecting the closing of derivatives exchanges and Japanese news releases, respectively. The timing of these spikes differs slightly in summer when daylight saving time is adopted in the UK and the US but not Japan.

The high trading that typically accompanies macro news releases represents a further dimension on which the markets differ from the features assumed in macro-based exchange-rate models. In macro-based models all agents have rational expectations and all information is public. The release of macro news causes everyone's expectations to be revised identically so the price moves instantly to reflect the new expectation without associated trading volume.

Feedback Trading

The data provide substantial evidence of both positive and negative feedback trading in foreign exchange. Sager and Taylor [169] find evidence for positive feedback trading in interdealer order flow using Granger-causality tests applied to the Evans and Lyons [58] daily data. Marsh and O'Rourke [131] and Bjønnes et al. [18] find evidence for negative feedback trading in semi-daily commercial-customer order flow but not in corresponding financial-customer order flow. Daniélsson and Love [44] find evidence of feedback trading in transaction-level interdealer trading data.

Feedback trading can greatly influence asset-price dynamics. For example, Delong et al. [45] show that in the presence of positive-feedback traders, the common presumption that rational speculators stabilize markets is turned on its head, and rational speculators intensify market booms and busts instead. Negative-feedback traders, by contrast, tend to dampen volatility.

There are at least three important sources of feedback trading in currency markets: technical trading, options

hedging, and price-contingent orders. We discuss each in turn.

Technical Trading Technical trading is widespread in foreign exchange markets. Taylor and Allen [180] show that 90 percent of chief dealers in London rely on technical signals. Cheung and Chinn [36] find that technical trading best characterizes thirty percent of trading behavior among US dealers and the fraction has been rising. Similar evidence has emerged for Germany [137] and Hong Kong (Lui and Mole 1998).

Trend-following technical strategies generate positive-feedback trading. Froot and Ramadorai [74] present evidence for positive-feedback trading among institutional investors: their results indicate that, for major currencies vs. the dollar, a one standard deviation shock to current returns is associated with an 0.29-standard-deviation rise in institutional-investor order flow over the next thirty days.

Contrarian technical strategies generate negative feedback. For example, technical analysts claim that "support and resistance" levels are points at which trends are likely to stop or reverse, so one should sell (buy) after rates rise (fall) to a resistance (support) level. Support and resistance levels are a day-to-day topic of conversation among market participants, and most major dealing banks provide active customers with daily lists of support and resistance levels.

Option Hedging Option hedging also generates both positive- and negative-feedback trading. To illustrate, consider an agent who buys a call option on euros. If the intent is to speculate on volatility, the agent will minimize first-order price risk (delta-hedge) by opening a short euro position. Due to convexity in the relationship between option prices and exchange rates, the short hedge position must be modestly expanded (contracted) when the euro appreciates (depreciates). The dynamic adjustments therefore bring negative-feedback trading for the option holder and, by symmetry, positive-feedback trading for the option writer.

Barrier options – which either come into existence or disappear when exchange rates cross pre-specified levels – can trigger either positive- or negative-feedback trading and the trades can be huge. Consider an "up-and-out call," a call that disappears if the exchange rate rises above a certain level. If the option is delta-hedged it can trigger substantial positive-feedback trading when the barrier is crossed: since the short hedge position must be eliminated, the rising exchange rate brings purchases of the underlying asset. The entire hedge is eliminated all at once, however, so the hedge-elimination trade is far larger than the modest hedge adjustments associated with plain-vanilla options. Many market participants pay close attention to the levels at which barrier options have been written, and make efforts to find out what those levels are. Related option types, such as Target Resumption Notes (TARNs), also trigger substantial feedback trading but tend to spread it out.

Price-Contingent Orders Price-contingent customer orders are the third important source of feedback trading in foreign exchange. These are conditional market orders, in which the dealer is instructed to transact a specified amount at market prices once a trade takes place at a pre-specified exchange-rate level. There are two types: stop-loss orders and take-profit orders. Stop-loss orders instruct the dealer to sell (buy) if the rate falls (rises) to the trigger rate, thereby generating positive-feedback trading. By contrast, take-profit orders instruct the dealer to sell (buy) if the price rises (falls) to the trigger rate, thereby generating negative-feedback trading.

Take-profit orders are often used by non-financial customers that need to purchase or sell currency within a given period of time. Their option to wait is valuable due to the volatility of exchange rates. They can avoid costly monitoring of the market and still exploit their option by placing a take-profit order with a dealer. Financial customers also use take-profit orders in this way. Stop-loss orders, as their name implies, are sometimes used to ensure that losses on a given position do not exceed a certain limit. The limits are frequently set by traders' employers but can also be self-imposed to provide "discipline." Stop-loss orders can also be used to ensure that a position is opened in a timely manner if a trend develops quickly. Savaser [171] finds that stop-loss order placement intensifies prior to major macro news releases in the US.

One might imagine that these orders would tend to offset each other, since rising rates trigger stop-loss buys and take-profit sales, and vice versa. However, as discussed in Osler [151,152], differences between the clustering patterns of stop-loss and take-profit orders reduce the frequency of such offsets. Take-profit orders tend to cluster just on big round numbers: Stop-loss orders are less concentrated on the round numbers and more concentrated just beyond them (meaning above (below) the round number for stop-loss buy (sell) orders).

Since stop-loss and take-profit orders cluster at different points, offsets are limited and these orders create noticeable non-linearities in exchange-rate dynamics [151,152]. The presence of stop-loss orders, for example, substantially intensifies the exchange-rate's reaction to macro news releases [171]. Likewise, the tendency of take-

profit orders to cluster at the round numbers increases the likelihood that trends reverse at such levels. This is consistent with the technical prediction, introduced earlier, that rates tend to reverse course at support; and resistance levels. Finally, the tendency of stop-loss orders to cluster just beyond the round numbers brings a tendency for exchange rates to trend rapidly once they cross round numbers. This is consistent with another technical prediction, that rates trend rapidly after a trading-range break out.

Market participants often report that stop-loss orders are responsible for fast intraday exchange-rate trends called "price cascades." In a downward cascade, for example, an initial price decline triggers stop-loss sell orders that in turn trigger further declines, which in turn trigger further stop-loss sell orders, etc. Upward cascades are equally possible: since every sale of one currency is the purchase of another, there are no short-sale constraints and market dynamics tend to be fairly symmetric in terms of direction (most notably, there is no equivalent to the leverage effect). Dealers report that price cascades happen relatively frequently – anywhere from once per week to many times per week. Osler [152] provides evidence consistent with the existence of such cascades.

News Announcements

Macro news announcements typically generate a quick surge in currency trading volume and volatility. As shown in Fig. 2a and b, which are taken from Chaboud et al. [31], volume initially surges within the first minute by an order of magnitude or more. Dealers assert that the bulk of the exchange-rate response to news is often complete within ten seconds [36].

Carlson and Lo [29] closely examines one macro announcement, the timing of which was unanticipated. They show that in the first half-minute spreads widened and in the second half-minute trading surged and the price moved rapidly. Chaboud et al. [31] shows that after the first minute volume drops back substantially, but not completely, in the next few minutes. The remaining extra volume then disappears slowly over the next hour. The response of returns to news is particularly intense after a period of high volatility or a series of big news surprises [48,54], conditions typically interpreted as heightened uncertainty.

The US macro statistical releases of greatest importance are the GDP, the unemployment rate, payroll employment, initial unemployment claims, durable goods orders, retail sales, the NAPM index, consumer confidence, and the trade balance [3]. Strikingly, money supply releases have little or no effect on exchange rates [3,25,36,62], consistent with the observation above that aggregate money supply and demand seem unimportant for short run exchange-rate dynamics.

Statistical releases bring a home-currency appreciation when they imply a strong home economy. A positive one-standard deviation surprise to US employment, which is released quite soon after the actual employment is realized, appreciates the dollar by 0.98 percent. For GDP, which is released with a greater lag, a positive one-standard deviation surprise tends to appreciate the dollar by 0.54 percent [3]. Responses are driven by associated anticipations of monetary policy: anything that implies a stronger economy or higher inflation leads investors to expect higher short-term interest rates [13] and thus triggers a dollar appreciation, and vice versa.

Federal Reserve announcements following FOMC meetings do not typically elicit sharp increases in trading volume and volatility [13]. Instead, FOMC announcements bring only a small rise in trading volume (Fig 2c) and tend to reduce exchange-rate volatility [33]. This suggests that Federal Reserve policy shifts are generally anticipated, which is encouraging since that institution prefers not to surprise markets.

Unanticipated changes in monetary policy do affect exchange rates. Fratscher [133] finds that an unanticipated 25 basis-point rise in US interest rates tends to appreciate the dollar by 4.2 percent. Kearns and Manners [108], who analyze other Anglophone countries, find that a surprise 25 basis-point interest-rate rise tends to appreciate the home currency by only 38 basis points. Kearns and Manners also note a more subtle dimension of response: If the policy shift is expected merely to accelerate an already-anticipated interest-rate hike, the exchange-rate effect is smaller (only 23 basis points, on average) than if the shift is expected to bring consistently higher interest rates over the next few months (43 basis points on average).

Evidence presented in Evans and Lyons [59] suggests that exchange rates overshoot in responses to news announcements. For some types of news, between a tenth and a quarter of the initial response is typically reversed over the four consecutive days. The reversals are most pronounced for US unemployment claims and the US trade balance. This contrasts strikingly with the well-documented tendency for the initial stock-price response to earnings announcements to be amplified after the first day, a phenomenon known as "post-earnings announcement drift" (Kothari [113] provides a survey). Nonetheless, over-reaction to fundamentals has been documented repeatedly for other financial assets [10,26,173].

Exchange-rate responses to a given macro news statistic can vary over time, as dealers are well aware [36]. Dur-

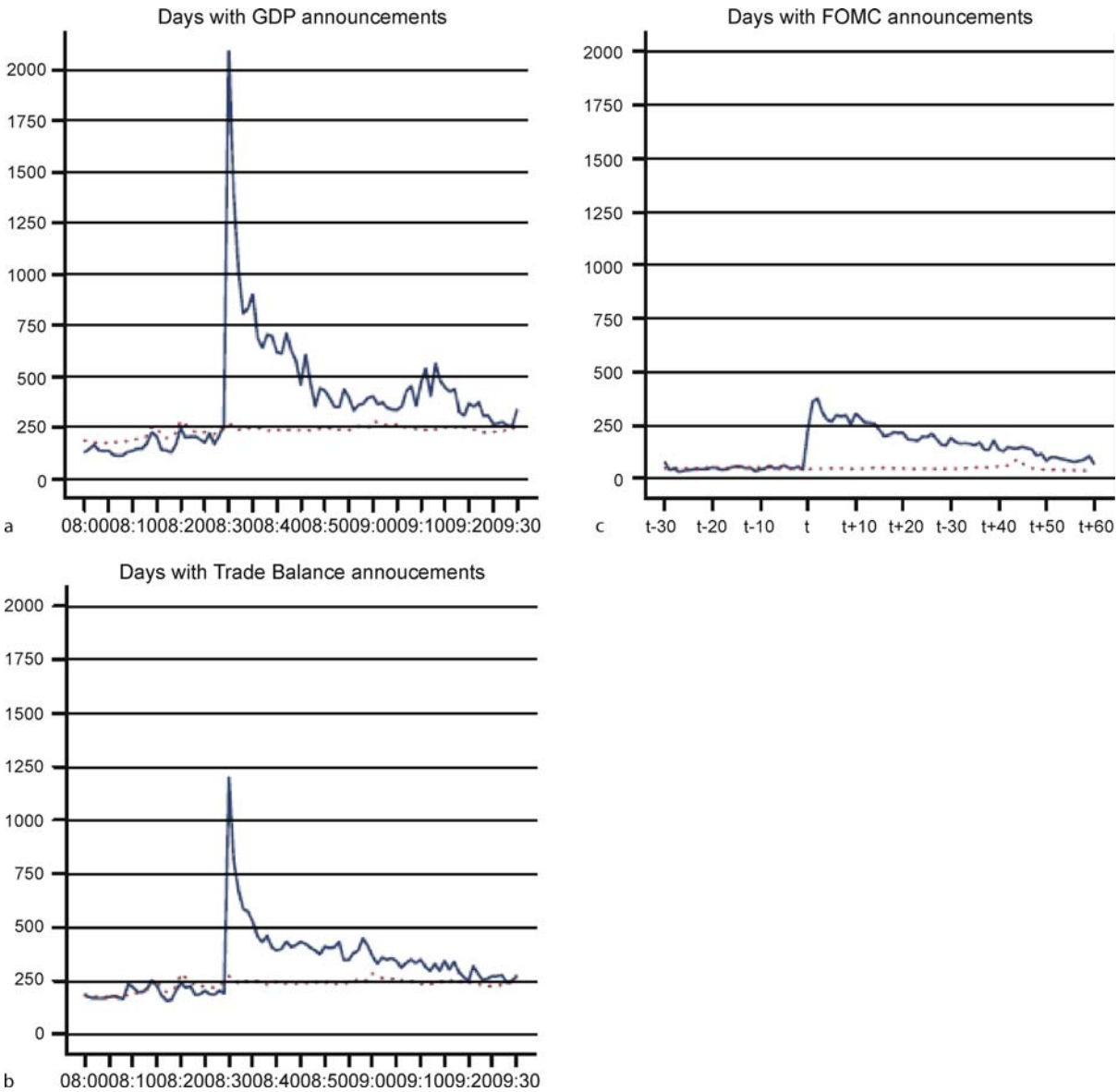

Market Microstructure, Foreign Exchange, Figure 2

Minute-by-minute trading volume, euro-dollar, around US scheduled macro news announcements. Based on tick-by-tick EBS trade data over 1999–2004. Trading volume relative to the intraday average. Source: [31]. Eastern Standard Time

ing the early 1980s, for example, the dollar responded fairly strongly to money supply announcements which, as noted above, is no longer the case. This shift appears to have been rational since it reflected public changes in Federal Reserve behavior: in the early 1980s the Fed claimed to be targeting money supply growth, a policy it has since dropped. The possibility that such shifts are not entirely rational is explored in Bachetta and van Wincoop [7]. Cheung and Chinn [36] provide further discussion of how and why the market's focus shifts over time. Using daily data, Evans and Lyons [60] find little evidence of such shifting during the period 1993–1999. This could reflect the masking of such effects in their daily data or it could indicate that such shifting was modest during those years of consistent economic expansion and consistent monetary policy structure.

Information relevant to exchange rates comes from many more sources than macroeconomic statistical releases. Trading volume and volatility are triggered by official statements, changes in staffing for key government

positions, news that demand for barrier options is rising or falling, reports of stop-loss trading, even rumors [48,147]. As documented in Dominguez and Panthaki [48], much of the news that affects the market is non-fundamental.

Numerous asymmetries have been documented in the responses to news. The effects of US macro announcements tend to be larger than the effect of non-US news [59,82]. Ehrmann and Fratzscher [54] attribute this asymmetry, at least in part, to the tendency for non-US macroeconomic statistical figures to be released at unscheduled times and with a greater lag. Ehrmann and Fratzscher also shows that exchange rates respond more to weak than strong European news, and Andersen et al. [3] report a similar pattern with respect to US announcements. This asymmetry is not well understood.

Carlson and Lo [29] shows that many interdealer limit orders are not withdrawn upon the advent of unexpected macro news. This might seem surprising, since by leaving the orders dealers seem to expose themselves to picking-off risk. It may not be the dealers themselves, however, that are thus exposed. The limit orders left in place may be intended to cover take-profit orders placed by customers, so the customer may be the one exposed to risk.

To be concrete: suppose a customer places a take-profit order to buy 5 at 140.50 when the market is at 140.60. The dealer can ensure that he fills the order at exactly the requested price by placing a limit order to buy 5 at 140.50 in the interdealer market. Suppose news is then released implying that the exchange rate should be 140.30. The dealer loses nothing by leaving the limit order in place: the customer still gets filled at the requested rate of 140.50.

This interpretation may appear to push the mystery back one step, because now the customer is buying currency at 140.50 when the market price of 140.30 would be more advantageous. Why wouldn't customers change their orders upon the news release, or withdraw them beforehand? This could reflect a rational response of customers to the high costs of monitoring the market intraday. Indeed, as noted earlier it is to avoid those costs that customers place orders in the first place. The Customers that choose not to monitor the market may not even be aware of the news.

Returns and Volatility

This section describes the basic statistical properties of returns and order flow.

Returns

Major exchange rates are often described as following a random walk, since it has long been well-documented that daily returns to major exchange rates vis-à-vis the dollar are not autocorrelated and are almost entirely unpredictable. The random walk description is technically inaccurate, of course, since the variance of returns can indeed be forecast: it is statistically more accurate to describe the exchange rate as a martingale. (Further, at the highest frequencies returns are slightly negatively autocorrelated, as shown in Table 1 [33]). Whatever the nomenclature, the fact that current exchange rates provide better forecasts than standard fundamentals-based models [134] has long been a source of pessimism about exchange-rate theory in general.

Market Microstructure, Foreign Exchange, Table 1
Autocorrelation of high-frequency returns. High-frequency autocorrelation of DEM returns, using Reuters indicative quotes over the period 1 October, 1992 through 30 September, 1993. Source: [33]

	5 min	10 min	15 min	30 min	Hourly
$\rho(1)$	−0.108	−0.093	−0.085	−0.066	−0.018
$\rho(2)$	−0.019	−0.030	−0.018	0.008	0.006
$\rho(3)$	−0.011	−0.002	0.006	0.024	−0.018

Though the unconditional autocorrelation of daily returns is approximately zero, the conditional autocorrelation is not. Research has long shown that trend-following technical trading rules are profitable in major exchange rates [140]. Though returns to these rules seems to have declined in recent years, more subtle strategies remain profitable on a risk-adjusted basis [35]. Markov switching models also have predictive power for exchange rate returns [46,50], though the switching variables must include more than mean returns [117].

Daily returns are correlated across currencies, as one might expect given exchange-rate responses to news. The correlation between daily euro-dollar and sterling-dollar returns, for example, is 70 percent, while correlations between these European exchange rates and dollar-yen are smaller: both are 46 percent [12].

It has long been recognized that short-horizon exchange-rate returns are leptokurtotic. Kurtosis in euro-dollar returns, for example, is 24, 19, and 14 at the fifteen-minute, half-hour, and one hour horizons, respectively, all significantly higher than the level of three associated with the normal distribution [154]. Even at the two-day horizon kurtosis is still statistically significantly above three, though it has declined to five. These figures need not be constant. Osler and Savaser [154] demonstrate that a number of properties of price contingent orders impart high kurtosis to the distribution of returns. These properties include: high kurtosis in the orders' own size distribution, intraday seasonals in the execution of these orders; and

the clustering patterns in their trigger rates described earlier. Stop-loss orders can also contribute to high kurtosis by contributing to price cascades. This analysis suggests that changes in market reliance on price-contingent orders could bring changes in the distribution of returns.

Within the overall distribution of returns there seems to have been a shift during the 1990s from the smallest returns, meaning those within one standard deviation of the mean, towards returns between one and five standard deviations [34]. The frequency of the most extreme returns, however, showed no trend.

Volatility

Unlike returns, volatility exhibits strong autocorrelation. As shown in Table 2, the first-order autocorrelation for daily volatility is typically above 0.50 and remains above 0.40 for at least a week. Evidence suggests that volatility is so persistent as to be fractionally integrated [12].

As recommended by Baillie and Bollerslev [8], volatility is typically captured with a GARCH(1,1) model or a close variant. Table 2b gives illustrative results from Chang and Taylor [33] showing that the AR component of the volatility process dominates (coefficients above 0.90) but the MA component is still significant. The MA component becomes increasingly important as the time horizon is shortened, though it remains subordinate. Table 2b also provides results suggesting that the double exponential distribution may fit return volatility better than the normal distribution. The thickness-of-tails parameter, "v," is two for the normal distribution but lower for the double exponential: estimates place it closer to unity than two.

Ederington and Lee [53] show, using 10-minute futures data for the DEM over July 3, 1989 through September 28, 1993, that the GARCH(1,1) model tends to underestimate the influence of the most recent shock and also shocks at long lags. These effects are captured better with an ARCH formulation that includes the lagged one-hour, one-day, and one-week return shock:

$$h_t = \alpha_0 + \sum_{i=1,6} \alpha_i \varepsilon_{t-i}^2 + \alpha_7 \varepsilon_{\text{hour}}^2 + \alpha_8 \varepsilon_{\text{day}}^2 + \alpha_9 \varepsilon_{\text{week}}^2 ,$$

where h_t is estimated conditional volatility and ε_t is the shock to returns. These authors also find that daily and intraday seasonal patterns in volatility become fairly unimportant after controlling for announcements and ARCH effects. They conclude that "much of the time-of-day patterns and day-of-the-week patterns are due to announcement patterns" (p. 536).

Volatility usually rises upon news announcements, consistent with the analysis presented in III.C [53], but it can fall: Chang and Taylor [33] find that US Federal Reserve news reduces volatility. This is consistent with the earlier finding that Fed news does not induce much extra trading. Volatility, like returns, can behave asymmetrically. Chang and Taylor [33] show that, during 1992, the volatility of dollar-mark was sensitive to US macro news but insensitive to German macro news. Such asymmetries need not be stable over time: Hashimoto [94] shows that asymmetries in the behavior of volatility changed dramatically around the Japanese bank failures of late 1997.

It is often hypothesized that volatility persistence derives from persistence in the flow of information, based on two premises: (i) volatility moves in parallel with trading volume, and (ii) trading volume is persistent because the advent of news is persistent. There is evidence to support both of these premises. Volatility and volume move together in most financial markets and foreign exchange is no exception, as shown in Fig. 1. Foreign exchange trading volume and volatility also move together at longer horizons [18,75]. Evidence also indicates persistence in the news process. Chang and Taylor [33], who count news releases on the Reuters real-time information system, find that autocorrelation in the number of news items is 0.29 at the one-hour horizon.

Market Microstructure, Foreign Exchange, Table 2
Strong autocorrelation in return volatility. a Daily realized volatilities constructed from five-minute returns based on Reuters indicative quote, July 1, 1987-December 31, 1993. Source: [160]. b Illustrative GARCH results assuming the normal distribution or the double-exponential distribution. Complete Reuters indicative quote for DEM, October 1992 through September 1993. Source: [33]

a	USD/DEM	USD/JPY	USD/GBP
$\rho(1)$	0.62	0.64	0.63
$\rho(2)$	0.52	0.53	0.54
$\rho(3)$	0.48	0.47	0.50
$\rho(4)$	0.45	0.44	0.47
$\rho(5)$	0.46	0.43	0.48

b	Hourly	30 Minutes	15 Minutes	5 Minutes
Normal Dist.				
α	0.045	0.035	0.098	0.100
	(3.83)	(4.36)	(8.32)	(13.82)
β	0.932	0.953	0.853	0.864
	(48.33)	(79.74)	(38.53)	(75.89)
Double-Exponential Dist.				
α	0.053	0.054	0.106	
	(5.07)	(4.86)	(4.97)	
β	0.930	0.936	0.878	
	(59.91)	(66.01)	(26.64)	
v	1.173	1.123	1.128	
	(41.71)	(52.14)	(58.82)	

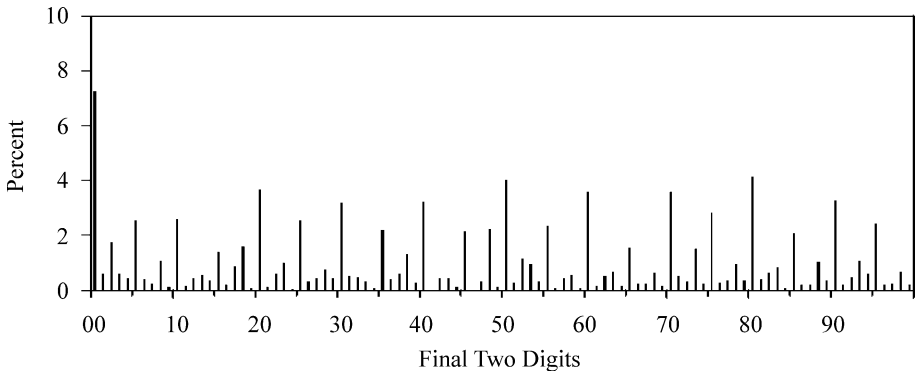

Market Microstructure, Foreign Exchange, Figure 3
Stop-loss and take-profit orders tend to be placed at round numbers. Data comprise the complete order book of the Royal Bank of Scotland in euro- dollar, sterling-dollar, and dollar-yen during the period September 1, 1999 through April 11, 2000. Chart shows the frequency with trigger rates ended in the 100 two-digit combinations from 00 to 99. Source: [151]

There is, however, little empirical evidence that directly traces volatility persistence in foreign exchange to news persistence. In fact, the only direct evidence on this point suggests that other factors are more important than news. Berger et al. [12] finds that persistence in news is primarily relevant to shorter-term volatility dynamics while long-run persistence in volatility is captured primarily by the low-frequency persistence in price impact, meaning the impact on exchange-rates of order flow. Figure 6, taken from Berger et al. [12], shows that daily price-impact coefficients for euro-dollar varied quite a bit during 1999–2004, and the series displays strong persistence at low frequencies. Further tests show that trading volume has modest explanatory power even after controlling for order flow.

Implied volatilities from exchange-traded options contracts have also been studied. Kim and Kim [111] find that implied volatilities in futures options are heavily influenced by volatility in the underlying futures price itself. They are not strongly influenced by news, and the few macro news releases that matter tend to reduce implied volatilities. Their analysis also indicates that implied volatilities tend to be lower on Mondays and higher on Wednesdays, though the pattern is not strong enough to generate arbitrage trading profits after transaction costs. Two studies show that daily volatility forecasts can be improved by using intraday returns information in addition to, or instead of, implied volatilities [132,160].

Order Flow and Exchange Rates, Part I: Liquidity and Inventories

Customer currency demand usually must net to around zero on trading days, as discussed earlier, and exchange-rate adjustment seems likely to be the mechanism that induces this outcome. If one group of customers decides to purchase foreign currency over the day, on net, the currency's value must rise to bring in the required liquidity supply from another group of customers. This implies, crudely, a relationship between net liquidity demand and exchange-rate returns.

To identify this relationship empirically one must distinguish liquidity-demand trades from liquidity-supply trades on a given day. We cannot simply look at trading volume or, equivalently, total buys or total sells, since it is the motivation behind the trades that matters. Instead we need to compare the purchases and sales of liquidity consumers. If they buy more than they sell then rates should rise to induce overnight liquidity supply and vice versa. The concept of "order flow" or, equivalently, "order imbalances," which we examine next, can be viewed as a measure of net liquidity demand.

Interdealer Order Flow

In the interdealer market we identify liquidity demanders with either (i) those placing market orders or (ii) those calling other dealers to trade directly. When using transaction data from a broker, order flow is calculated as market buy orders minus market sell orders; when using direct dealing data, order flow is calculated as dealer-initiated buy trades minus dealer-initiated sell trades.

Evans and Lyons [58] were the first to show that interdealer order flow has substantial explanatory power for concurrent daily exchange-rate returns, a result that has been replicated in numerous studies [56,97]. Benchmark results are provided in Berger et al. [12], which has the advantage of a relatively long dataset. That paper shows that

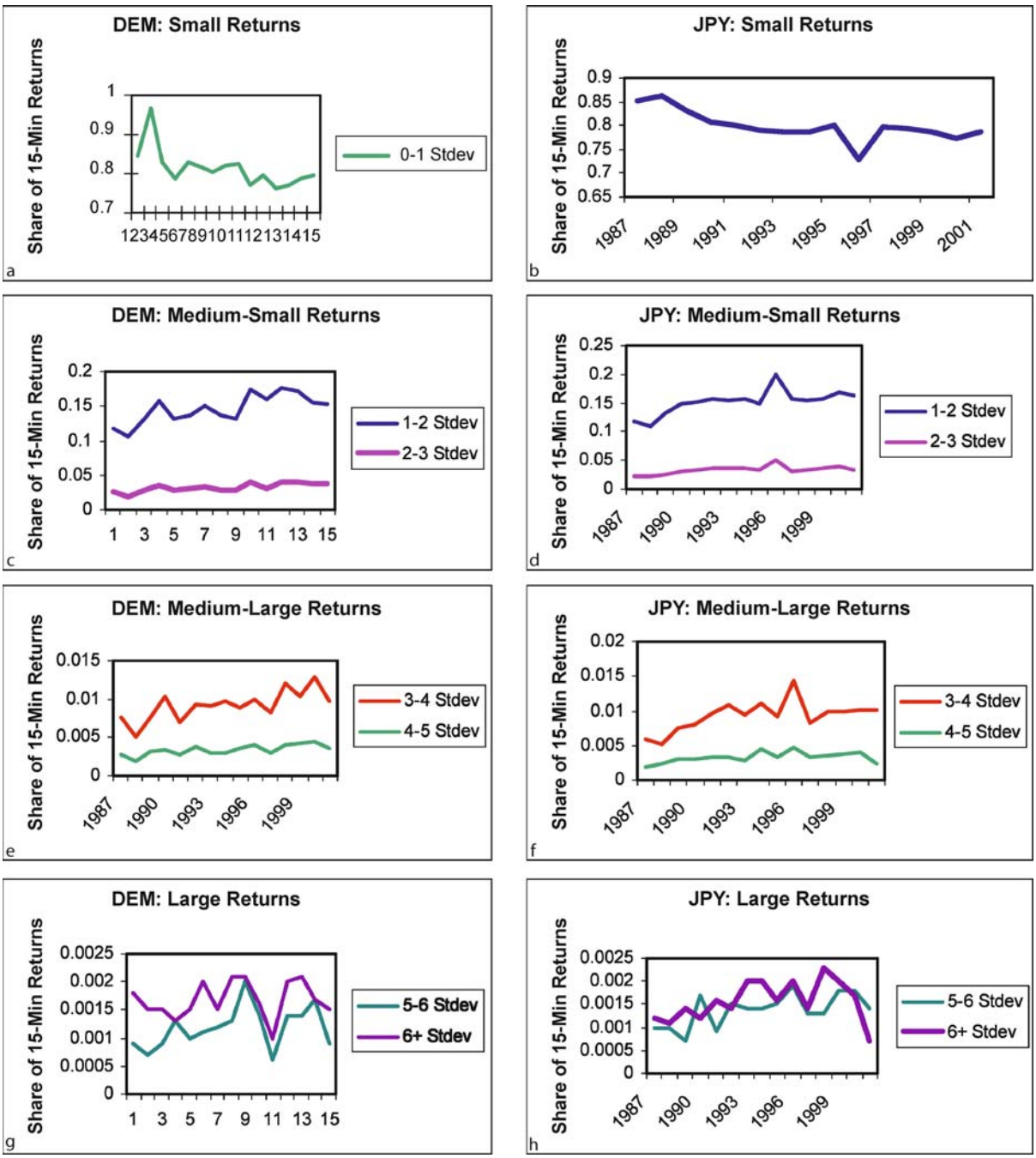

Market Microstructure, Foreign Exchange, Figure 4

Frequency distribution of returns has shifted. Data comprise tick-by-tick Reuters indicative quotes over 1987–2001. Source: [34]

the raw correlation between daily returns and interdealer order flow is 65 percent for euro-dollar, 42 percent for sterling-dollar, and 49 percent for dollar-yen. Berger et al. estimates that an extra $1 billion in order flow in a given day appreciates the euro, the pound, and the yen by roughly 0.40 percent, with R^2s in the vicinity of 0.50. By contrast, it is well known that the explanatory power of standard fundamental variables is typically well below 0.10 [58].

Evans and Lyons [58] and Rime, Sarno, and Sojli [166] find that the overall explanatory power of interdealer order flow for returns can be substantially increased by including order flow from other currencies. In Evans and Lyons [58], which uses daily interbank order flows for seven currencies against the dollar over four months in 1996, the joint explanatory power averages 65 percent and ranges as high as 78 percent.

Since feedback trading is ubiquitous in foreign exchange, one must consider the possibility that these correlations represent reverse causality – that returns are in fact driving order flow. Two studies investigate this possibility. Using daily data, Evans and Lyons [63] find that the influence of order flow on price survives intact after controlling for feedback effects; using transactions data, Daniélsson and Love [44] find that the estimated influence becomes even stronger after controlling for feedback trading.

Dealers have long recognized the importance of currency flows in driving exchange rates, and have said as much in surveys. In Gehrig and Menkhoff's survey [77], for example, over 86 percent of dealers said they rely on analysis of flows in carrying out their responsibilities. Indeed, the influence of order flow on exchange rates is a critical assumption in their trading strategies, as illustrated in the following debate over optimal management of stop-loss orders.

A dealer with a large stop-loss buy order could begin filling the order after the exchange-rate rises to the trigger price. Since the order-filling trades themselves will drive the price up, however, the average price paid will exceed the trigger rate, to the customer's disadvantage. The dealer could, alternatively, begin filling the order before the rate hits the trigger price. The buy trades will push the price up through the trigger rate and the average fill price will be closer to the trigger rate. The risk here is that the exchange rate bounces back down below the trigger rate, in which case the customer could justly complain of getting inappropriately "stopped out."

The key observation here is that the pros and cons of both strategy options are driven by the impact of order flow. Dealers do not view this as an hypothesis or as an assumption. To them it is something they know, in the same sense that one "knows" that the sun will disappear below the horizon at the end of the day (pace Hume). Dealers see order flow influence price too often and too consistently to question it.

The estimated price impact of interdealer order flow varies according to order size, time of day, and time horizon. Price impact has a concave relationship to size [155], consistent with evidence from equity markets [93,104]. This may reflect order splitting and other dealer strategies for minimizing the impact of large trades [17]. At the daily horizon, the price impact is linearly related to order flow, which makes sense since splitting a large trade into smaller individual transactions rarely takes more than a few hours. On an intraday basis, the price impact of interdealer order flow is inversely related to trading volume and volatility, as shown for dollar-yen in Fig. 7 [12]. As discussed earlier, spreads have a similarly inverse relation to trading volume and volatility (Fig. 1). This suggests, logically enough, that price impact is heavily influenced by spreads: when spreads widen, a given-sized transaction has a bigger price impact. Alternatively, however a third factor could be at work: depth. Depth presumably varies inversely with spreads and positively with trading volume intraday. Unfortunanely, information on depth is as yet almost nonexistent.

As time horizons lengthen the price impact of interdealer order flow declines monotonically [12]. For the euro, an extra $1 billion in order flow is estimated to bring an appreciation of 0.55 at the one-minute horizon but only 0.20 percent at the three-month horizon (Fig. 5, left). The explanatory power of interdealer order flow also varies with horizon but in a rising-falling pattern. The R^2 is 0.36 at the one-minute horizon, reaches 0.50 at the 30-minute horizon, stays fairly constant to the one-week horizon, and then falls sharply to about 0.17 percent at the two-month horizon (Fig. 5, right). Even at 17 percent, however, the explanatory power of order flow at three months is substantially higher than has been achieved with other approaches. A similar pattern is found in Froot and Ramadorai, using institutional investor order flow, though they find a peak at roughly one month rather than one week [74]. They attribute the initial rise to positive-feedback trading.

The positive relation between interdealer order flow and exchange rates could be influenced by inventory effects as well as the liquidity effects described above. Inventory effects were, in fact, the first connection between order flow and asset prices to be analyzed in the broader microstructure literature, e. g. [177]. Dealers that provide liquidity to other dealers are left with an inventory position and thus inventory risk. Dealers charge a spread which compensates them for this risk. The spread, in itself, generates a positive relationship between order flow and returns: prices typically rise to the ask price upon buy orders and fall to the bid price upon sell orders.

Customer Order Flow

Order flow in the customer market is measured as customer-initiated buy trades minus customer-initiated sell

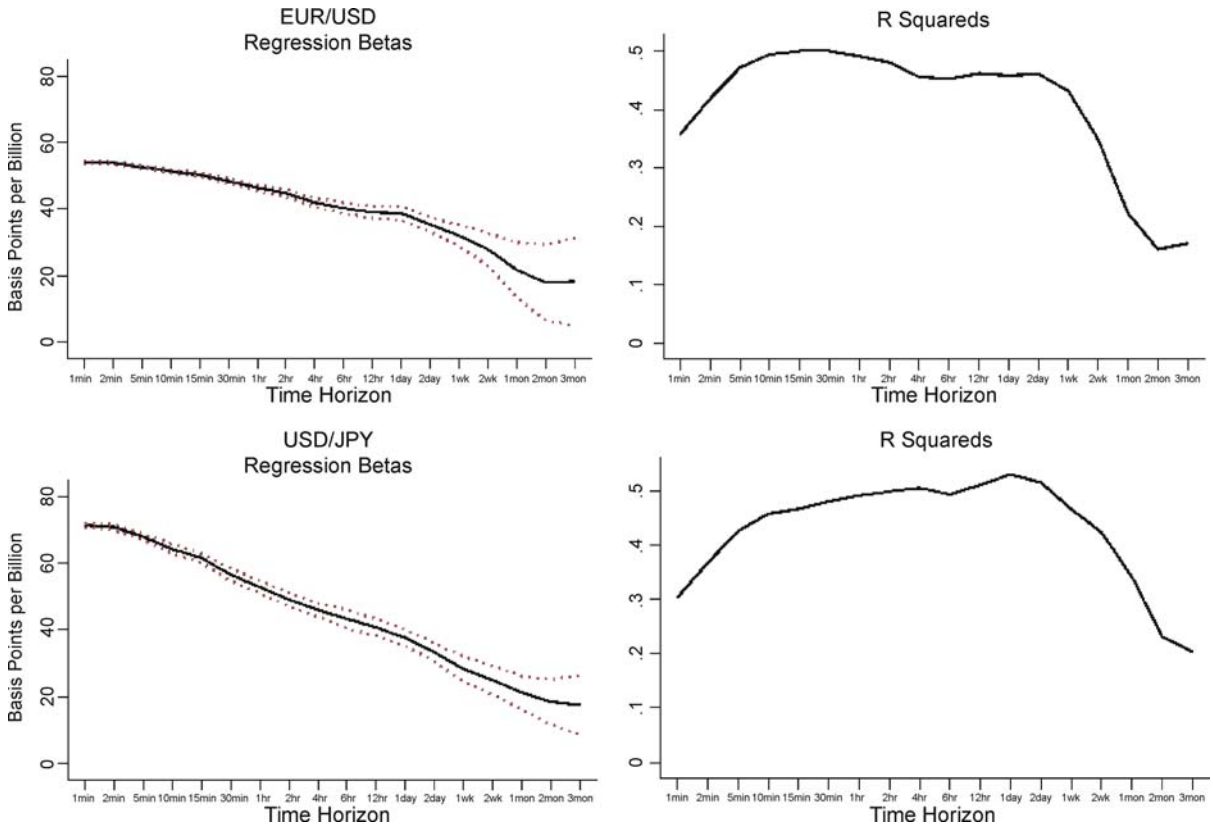

Market Microstructure, Foreign Exchange, Figure 5

Response of returns to order flow at various horizons. Charts on the left show beta coefficients from regressions of returns on contemporaneous interdealer order flow for time horizons ranging from one minute to three months. Charts on the right show coefficients of determination from those same regressions. Underlying data comprise minute-by-minute EBS transaction and quote records from 1999–2004. [12]

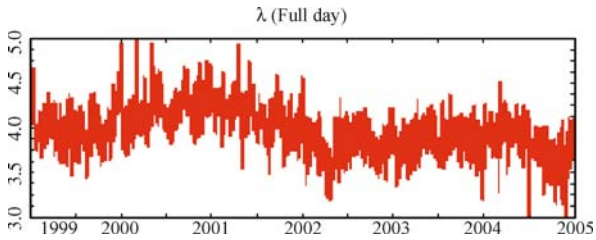

Market Microstructure, Foreign Exchange, Figure 6

Daily price impact coefficients for euro-dollar, 1999–2004. Underlying data comprise minute-by-minute EBS transaction and quote records from 1999–2004. Source: [12]

trades. This is consistent with a liquidity interpretation on a trade-by-trade basis, since each customer effectively demands instantaneous liquidity from their dealer. Customer order flow, however, is not ideally suited to measuring customer net liquidity demand at daily or longer horizons. If a customer is coming to the market in response to an exchange-rate change, then the customer may be demanding liquidity from its own dealer at that instant while effectively supplying liquidity to the overall market.

This distinction proves critical when interpreting the empirical relation between daily customer order flow and exchange rates. There should be a positive relation between daily order flow and returns for customer groups that typically demand overnight liquidity. An increase in their demand for foreign currency, for example, should induce a rise in the value of foreign currency to elicit the required overnight supply. Implicit in that story, however, is a *negative* relation between order flow and returns for customer groups that typically supply overnight liquidity.

Researchers have documented repeatedly that, at the daily horizon, financial-customer order flow is positively related to returns while commercial-customer order flow is negatively related to returns. Confirming evidence is found in Lyons' [122] study of monthly customer order flows at Citibank; in Evans and Lyons [61] study of daily

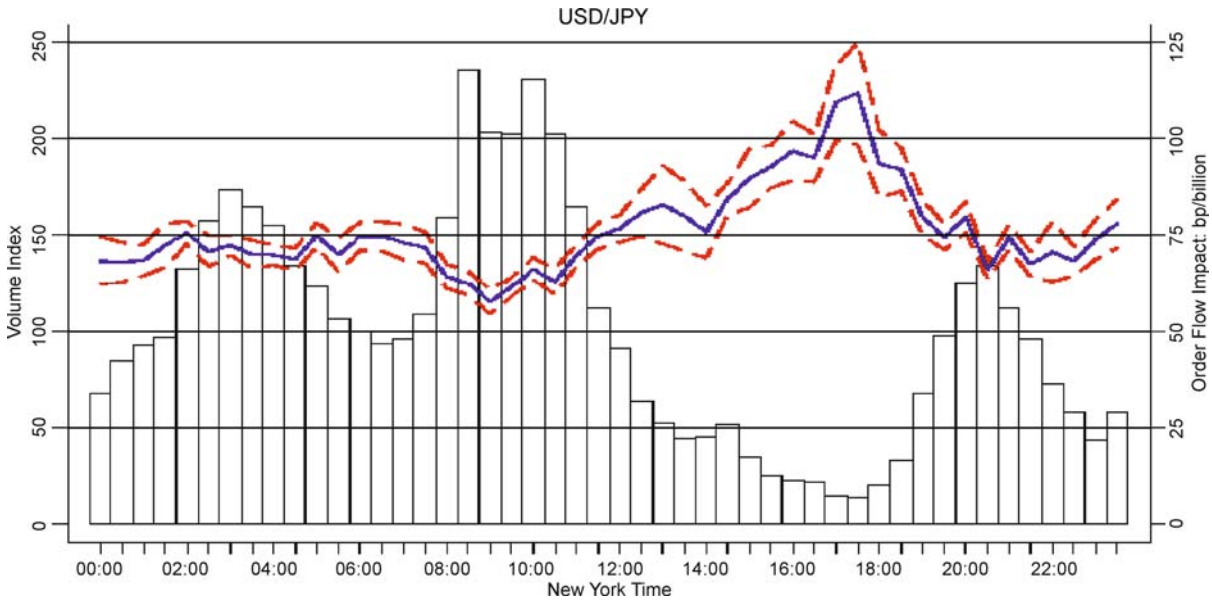

Market Microstructure, Foreign Exchange, Figure 7

Intraday Regression Betas and Average Trading Volume. Figure is based on the following regression: $\Delta s_t = \alpha + \beta \mathrm{OF}_t + \eta_t$, where Δs_t is the return and OF_t is contemporaneous order flow. Regressions based on one-minute EBS trade data from 1999–2004 are run separately for each half hour of the trading day. Line shows estimated coefficients with standard error bands. Bars show order flow measured relative to the days' average (day's average set at 100). Source: [12]

Market Microstructure, Foreign Exchange, Table 4

Autocorrelation coefficients for the number of exchange-rate relevant news items, 1 October 1992 through 30 September, 1993. Reuters News data. Source: [33]

	Hourly	30 Min	15 Min	10 Min	5 Min
$\rho(1)$	0.27	0.22	0.34	0.09	0.06
$\rho(2)$	0.29	0.16	0.12	0.09	0.04
$\rho(3)$	0.22	0.15	0.11	0.08	0.05

and weekly customer flows at the same bank; in Marsh and O'Rourke's [131] analysis of daily customer data from the Royal Bank of Scotland, another large dealing bank; and in Bjønnes et al. [18] comprehensive study of trading in Swedish kroner, and in Osler et al.'s [154] study of a single dealer at a medium-sized bank. The pattern is typically examined using cointegration analysis where the key relationship is between exchange-rate levels and cumulative order flow.

This pattern suggests that financial customers are typically net consumers of overnight liquidity while commercial customers are typically net suppliers. More direct evidence that commercial customers effectively supply overnight liquidity, on average, comes from evidence that commercial-customer order flow responds to lagged returns, rising in response to lower prices and vice versa. Marsh and O'Rourke [131] show this with daily data from the Royal Bank of Scotland. Bjønnes et al. [18] show this using comprehensive trading data on the Swedish krone sampled twice daily.

It is easy to understand why financial customers would demand liquidity: presumably they are speculating on future returns based on some information that is independent of past returns. Indeed, the identification of financial customers with speculation is explicit in Klitgaard and Weir's [112] study of currency futures markets. The IMM requires the agents they deem large speculators to report their positions on a weekly basis. Klitgaard and Weir show that their weekly position-changes are strongly correlated with concurrent exchange-rate returns. "[B]y knowing the actions of futures market speculators over a given week, an observer would have a 75 percent likelihood of correctly guessing an exchange-rate's direction over that same week" (p. 17).

It is not so immediately obvious why commercial customers would supply overnight liquidity, since our first image of a liquidity supplier is a dealer. Dealers supply intraday liquidity knowingly and are effectively passive in their trades with customers. By contrast, commercial customers are not supplying liquidity either knowingly or passively.

Market Microstructure, Foreign Exchange, Table 5

Order flow carries information about exchange-rate fundamentals. The table shows the R^2 statistics and associated marginal significance levels for the ability of daily customer order flow at Citibank during the period 1994 to 2001 to forecast upcoming announcements of key macro variables. Source: [57]

	US Output Growth				German Output Growth			
Forecasting Variables	1 Mo.	2 Mo.	1 Qtr.	2 Qtrs.	1 Mo.	2 Mo.	1 Qtr.	2 Qtrs.
Output	0.002	0.003	0.022	0.092	0.004	0.063	0.069	0.006
	(0.607)	(0.555)	(0.130)	(0.087)	(0.295)	(0.006)	(0.009)	(0.614)
Spot Rate	0.001	0.005	0.005	0.007	0.058	0.029	0.003	0.024
	(0.730)	(0.508)	(0.644)	(0.650)	(0.002)	(0.081)	(0.625)	(0.536)
Order Flows	0.032	0.080	0.189	0.246	0.012	0.085	0.075	0.306
	(0.357)	(0.145)	(0.002)	(0.000)	(0.806)	(0.227)	(0.299)	(0.000)
All	0.052	0.086	0.199	0.420	0.087	0.165	0.156	0.324
	(0.383)	(0.195)	(0.011)	(0.000)	(0.021)	(0.037)	(0.130)	(0.000)

Commercial customers are, instead, just responding to changes in relative prices in order to maximize profits from their core real-side businesses. Suppose the foreign currency depreciates. Domestic firms note that their foreign inputs are less expensive relative to domestic inputs and respond by importing more, raising their demand for the foreign currency. This effect, a staple of all international economic analysis, has been well-documented empirically at horizons of a quarter or longer, e. g. [5]. On an intraday basis this effect is often evident in the behavior of Japanese exporting firms, which hire professional traders to manage their vast dollar revenues. These traders monitor the market intraday, selling dollars whenever the price is attractive. The vast majority of commercial customers need to buy or sell currency only occasionally so they can't justify hiring professional traders. They can use take-profit orders, however, to achieve the same goal, since this effectively enlists their dealers to monitor the market for them. At the Royal Bank of Scotland take-profit orders are 75 (83) percent of price-contingent orders placed by large corporations (middle-market) corporations [155], but only 53 percent of price-contingent orders overall.

The evidence to date suggests the following crude portrait of day-to-day liquidity provision in foreign exchange (a portrait first articulated in [18]). Financial customers tend to demand liquidity from their dealers, who supply it on an intraday basis. The dealing community as a whole, however, does not provide overnight liquidity. Instead, commercial customers supply the required overnight liquidity, drawn to the market by new, more attractive prices. Sager and Taylor [169] distinguish between "push" customers, who demand liquidity, and "pull" customers, who respond to price changes by providing liquidity. The market structure just outlined effectively identifies financial customers as short-run push customers and commercial customers as short-run pull customers.

This picture is extremely preliminary and will doubtless change as new evidence arrives. There is, for example, no theoretical or institutional reason why commercial customers must exclusively supply overnight liquidity or financial customers exclusively demand it. To the contrary, there are good theoretical reasons why the roles could sometimes be reversed. A change in commercial currency demand could result from forces outside the currency market, such as a war-induced rise in domestic economic activity, rather than a response to previous exchange-rate changes. In this case commercial end-users would consume liquidity rather than supplying it.

Speculative demand could also respond to changes in exchange-rate levels. Indeed, rational speculators are the *only* overnight liquidity suppliers in the widely-respected Evans and Lyons [58] model. In these models the trading day begins when agents arrive with arbitrary liquidity demands. The agents trade with their dealers, leaving the dealers with unwanted inventory. Dealers then trade with each other, redistributing their aggregate inventory but not reducing it. At the end of the trading day dealers sell the unwanted inventory to rational investors who are induced to supply the required liquidity by a change in the exchange rate. If the initial liquidity demanders have sold foreign currency, for example, the currency's value declines thus raising the risk premium associated with holding the currency. This encourages the risk-averse investors to take bigger positions in foreign assets, and as they enact the portfolio shift financial order flow is positive.

The Evans–Lyons scenario is necessarily simple. In a model with many assets, negative-feedback trading among financial customers requires that the currency has no perfect substitutes [88]. This condition holds in foreign exchange since exchange rates generally have low correlation with each other and with equities. For the negative feedback trading to be finite it is also required that specu-

lators are risk-averse and/or face constraints on their trading. Though currency speculators appear to have a fairly high risk tolerance, their trading is always administratively constrained, as discussed earlier. The prevalence of contrarian technical trading strategies, such as those based on support and resistance levels, provides a further reason to expect negative-feedback trading among financial customers.

Despite these reasons to expect negative-feedback trading among financial customers, the evidence for it is thin and mixed. Financial agents do place a hefty share of take-profit orders [155], so a liquidity response from them is a fact. But their liquidity response may not be substantial relative to the overall market. Bjønnes et al. [18] study of trade in Swedish kroner and Marsh and O'Rourke's [131] study of customer trades at the Royal Bank of Scotland both find no sensitivity of financial order flow to lagged returns.

The influence of order flow on exchange rates described in this section works through liquidity effects. The broader microstructure literature refers to this influence in terms of "downward-sloping demand," highlighting that the demand for the asset has finite, rather than infinite, elasticity. Downward-sloping demand could explain why Froot and Ramadorai [74] find that the initial influence of institutional investor order flow disappears after roughly a year. Institutional investors – indeed, all speculative agents – have to liquidate positions to realize profits. When the positions are initially opened, the associated order flow could move the exchange rate in one direction; when the positions are liquidated the reverse order flow could move the exchange rate in the reverse direction.

Finite elasticity of demand is the underlying reason for exchange-rate movements in Hau and Rey's [96] model of equity and currency markets. Carlson et al. [30] develop a related exchange-rate model in which financial and commercial traders can be both liquidity suppliers and liquidity demanders. This model, which takes its critical structural assumptions directly from the microstructure evidence, predicts that financial (commercial) order flow is positively (negatively) related to concurrent returns, consistent with the evidence. It also predicts that these relations are reversed in the long run, consistent with evidence in Fan and Lyons [64] and Froot and Ramadorai [74]. Investors in the model have no long-run effect on exchange rates because they ultimately liquidate all their positions. Since commercial agents dominate long-run exchange rates, fundamentals such as prices and economic activity are important in the long run even though the may not dominate in the short run. In addition to being consistent with the microstructure evidence, this model is also consistent with most of the major puzzles in international macroeconomics, including: the apparent disconnect between exchange-rates and fundamentals, the increase in real-exchange-rate volatility upon the advent of floating rates, the short-run failure and long-run relevance of purchasing power parity, and the short-run failure of uncovered interest parity.

Order Flow and Exchange Rates

The influence of order flow on exchange rates is another aspect of the foreign exchange market that "does not seem to tally closely with current theory … " [81]. The equilibrium exchange rate in standard models adjusts to ensure that domestic and foreign money supplies equal corresponding money demands. The currency purchases or sales that accompany portfolio adjustments are not modeled and are considered unimportant. Indeed, order flow per se cannot be calculated in these models since they assume continuous purchasing power parity and/or continuous uncovered interest parity.

The contrast between microstructural reality and standard models is especially clear when we examine the mechanism through which news affects exchange rates. In macro-based models, the public release of information generates an immediate revision of shared expectations of future exchange rates, which in turn brings an immediate exchange-rate adjustment that requires no trading. Trading is unlikely, in fact, since no rational speculator would trade at any other price. Thus order flow in these models has no role in the exchange-rate adjustment to news.

The evidence shows, however, that order flow is the main conduit through which news influences exchange rates. Roughly two thirds of the influence of news on exchange-rate levels and volatility comes from the associated order flow [63,118]. During the "once-in-a-generation yen volatility" of 1998, "order flow [was the] most important … source of volatility," according to the investigation of Cai et al. [25], even more important than news and central bank intervention.

Reassuringly, the idea that order flow affects exchange rates is a natural extension of an important lesson learned after the advent of floating rates in the 1970s.

> [E]xchange rates should be viewed as prices of durable assets determined in organized markets (like stock and commodity exchanges) in which current prices reflect the market's expectations concerning present and future economic conditions relevant for determining the appropriate values of these durable assets, and in which price changes are

> largely unpredictable and reflect primarily new information that alters expectations concerning these present and future economic conditions (p. 726 in [73]).

There has long been extensive evidence that order flow influences price in stock markets [38,101,174]. In bond markets the evidence emerged later, due to constraints on data availability, but is nonetheless substantial [24,67,106,156,175,176]. Since exchange rates are asset prices they should be determined like other asset prices and thus order flow should be influential.

Order Flow and Exchange Rates, Part II: Information

So far we have considered two reasons why order flow could affect exchange rates: liquidity effects and inventory risk. This section considers a third and critically important reason: order flow carries private information.

The information hypothesis is suggested by evidence showing that much of the exchange-rate response to order flow is permanent. Payne [157], who decomposes returns into permanent and transitory components consistent with Hasbrouck [93], finds that "the permanent component accounts for … one quarter of all return variation" (p. 324). A permanent effect is implicit in Evans and Lyons' [58] evidence that order flow has strong explanatory power for daily exchange-rate returns, since daily returns are well described as a random walk. A permanent relation is also suggested by the finding, noted earlier, that cumulative order flow is cointegrated with exchange rates [18,110]. A permanent relation between order flow and price is not consistent with the inventory analysis presented earlier. A permanent relation is consistent with liquidity effects if the shifts in liquidity demand or supply are permanent. A permanent relation is inevitable, however, if order flow carries private fundamental information.

The influence of private fundamental information on asset prices was originally analyzed in equity-inspired models [79,115], which begin with the observation that sometimes customers often have private information about an asset's true value that dealers do not share. Since an informed customer only buys (sells) when the dealer's price is too low (high), dealers typically lose when they trade with such customers. To protect themselves from this adverse selection, dealers charge a bid-ask spread, ensuring that profits gained from trading with uninformed customers balance the inevitable losses from trading with informed customers [39]. Rational dealers ensure that their prices reflect the information communicated by a customer's choice to buy or sell [52,79]. Prices are "regret-free" in the sense that a dealer would not wish s/he had charged a higher (lower) price after learning that the customer wishes to buy (sell). Due to the spread, prices rise when informed customers buy and fall when informed customers sell. Meanwhile, others update their conditional expectation of the asset's true value and adjust their trades and quotes accordingly. Ultimately the information becomes fully impounded in price. Since the information is fundamental, the effect is permanent.

Types of Information

Private fundamental information in the foreign exchange market is likely to be structurally different from private fundamental information in a stock market. The fundamental determinants of a firm's value include many factors about which there can naturally be private information, such as management quality, product quality, and a competitor's strength. The fundamental determinants of a currency's value, by contrast are macroeconomic factors such as economic activity, interest rates, and aggregate price levels, most of which are revealed publicly.

The foreign exchange literature implicitly elaborates multiple different interpretations of the private information customers might bring to the market. These vary along three dimensions: (i) whether the information comes from commercial customers, real-money funds, or leveraged investors; (ii) whether the information is fundamental; and (iii) whether the information is passively or actively acquired. Though these three dimensions provide eight conceivable information categories, only some of these appear to be relevant for research. For example, only a small minority of the thousands of non-financial firms around the world would ever attempt to acquire either fundamental or non-fundamental information before trading. The four categories that seem likely to be important, based on the current literature, are discussed below.

Fundamental Information Passively Acquired by Commercial Customers Information about exchange-rate fundamentals may be "dispersed" among customers without being under their control. This hypothesis is most closely associated with Evans and Lyons:

> The dispersed information we have in mind in fact characterizes most variables at the center of exchange rate modeling, such as output, money demand, inflation, [and] consumption preferences … These variables are not realized at the macro level, but rather first as dispersed micro realizations, and only later aggregated by markets and/or governments. For some of these measures, such as risk

> preferences and money demands, government aggregations of the underlying micro-level shocks do not exist, leaving the full task of aggregation to markets. For other variables, government aggregations exist, but publication lags underlying realizations by 1–4 months, leaving room for market-based aggregation in advance of publication ([61], p. 3).

For concreteness, suppose the economy is expanding rapidly and in consequence commercial firms are all trading actively. Each individual firm might not recognize the generality of its experience but a dealer could potentially see the high economic activity reflected in his commercial-customer order flow. This information would provide the dealer with a signal of GDP concurrent with its realization and thus prior to the associated statistical release.

Fundamental Information Passively Acquired by Financial Customers A variant of the dispersed information hypothesis postulates that the relevant fundamentals concern capital markets as well as the real economy. For example, high demand from institutional investors might indicate that risk aversion is low [58,61,122]. It is not clear whether structural features of financial markets should be considered fundamental, in part because the definition of the term fundamental is not entirely clear. It is clear, however, that any fundamental factor should be relevant to long run equilibrium. Certain structural features of financial markets, like risk appetite, seem likely to influence long-run international macro variables such as international net asset positions (the US net asset position has changed sign but once since 1970), and these in turn seem likely to influence exchange rates. So it seems that some deep financial-market parameters are fundamental, or at least represent some intermediate category between fundamental and non-fundamental.

Fundamental Information Actively Sought by Customers Certain financial customers – typically leveraged investors – forecast exchange rates by combining existing public information with their own economic insights. For example, many such agents attempt to profit from the big returns associated with macro statistical releases by generating private forecasts of upcoming announcements. These customers thus actively generate private fundamental information, rather than passively reflecting information that arises as a normal part of their business. This actively-acquired information could also be reflected in customer order flow, so dealers could still generate their own private signals by observing it. Dealers often report that currency demand is highly correlated within certain types of leveraged investors, permitting them to infer information from observing the trades of just one or a few of these investors.

Indirect evidence for the existence of actively-acquired information comes from Marsh and MacDonald [124]. They find, in a sample of exchange-rate forecasts, that a major cause of forecast heterogeneity "is the idiosyncratic interpretation of widely available information, and that this heterogeneity translates into economically meaningful differences in forecast accuracy" (p. 665). They also find that heterogeneity is a significant determinant of trading volume, consistent with predictions in the literature that diversity of price forecasts generates trading [91,107,181,182].

Non-fundamental Information Some speculative traders may respond to non-fundamental information, like noise traders. Others could respond to non-fundamental hedging needs, as suggested in Bacchetta and van Wincoop [7]. Evidence for the relevance of non-fundamental information is provided in Osler [152], Dominguez and Panthaki [48], and Cao, Evans and Lyons [27]. If the information in order flow is not fundamental it is likely to have only a transitory influence on rates.

Trades based on non-fundamental information may be informative to dealers even if they have only a transitory impact on the market, since dealers speculate at such high frequencies. Indeed, Goodhart [81] insists that dealers rely on nothing but non-fundamental information: dealers' "speculative activities are not based on any consideration of longer-term fundamentals. … And to repeat, … the extremely large-scale, very short-term speculative activity in this market by the individual traders … is *not* based on a long-term future view of economic fundamentals" (pp. 456–457, italics in the original) Consistent with this, US dealers assert that the high-frequency returns on which they focus are unrelated to fundamentals [36]. For example, "at the intraday horizon, PPP has no role according to 93 percent of respondents" (p. 465).

The Evidence: Order Flow Does Carry Information

The evidence indicates fairly clearly that some foreign exchange order flow carries private information. For example, Bjønnes, Osler, and Rime [21] show statistically that banks with the most customer business have an information advantage in the interdealer market, a proposition that dealers themselves certainly support [36,81].

The broader microstructure literature identifies location, specifically proximity to relevant decision-makers, as another potential source of information advantage in fi-

Market Microstructure, Foreign Exchange, Table 6

Net purchases for banks in four size categories. The table considers net purchases – the number of purchases minus the number of sales – for four groups of banks vis-à-vis a Scandinavian bank during one week of 1998. Table shows how these net purchases are correlated with contemporaneous returns and with net purchases for other bank categories. All numbers with absolute value over 0.24, 0.28, or 0.36 are significant at the 10 percent, 5 percent, and 1 percent level, respectively. Source: [21]

	Return	Biggest (Rank 1–20)	Big Rank (21–50)	Small (Rank 51–100)	Smallest (Rank > 100)
Return	1.00				
Biggest	0.55***	1.00			
Big	0.26*	0.29**	1.00		
Small	–0.43***	–0.66***	–0.28**	1.00	
Smallest	–0.44***	–0.79***	–0.32***	0.41***	1.00

nancial markets [40,95,129]. Location also appears to be relevant in foreign exchange. Covrig and Melvin [41] find that order flow from Japan tends to lead movements in dollar-yen. Menkhoff and Schmeling [141] find that location affects the information content of interbank trades in the market for rubles. Their analysis indicates that trades originating from the two major financial centers, Moscow and St. Petersburg, have a permanent price impact while trades originating from six peripheral cities do not. D'Souza [49] shows that "trades are most informative when they are initiated in a local country or in major foreign exchange centers of London and New York."

If order flow carries exchange-rate relevant information then one should be able to use it to forecast exchange rates. Studies consistently find that *customer* order flow has predictive power for exchange rates. Evans and Lyons [60] find that daily customer order flow at Citibank has forecasting power for exchange-rate returns at horizons up to one month. Gradojevic and Yang [83] finds that customer and interbank order flow in the Canadian dollar market jointly have forecasting power for exchange rates. They also conclude that a non-linear forecasting structure, specifically an artificial neural network, is superior to linear approaches. Both Evans and Lyons [60] and Gradojevic and Yang [83] conclude that return forecasts are improved when customer order flow is disaggregated according to customer type, which suggests that some participants are more informed than others. Curiously, Rosenberg and Traub [168] provide evidence that *futures* order flow has predictive power for near-term spot returns. This raises the possibility that some informed investors choose to trade in futures markets.

Studies of the forecasting power of *interdealer* order flow arrive at mixed conclusions. Sager and Taylor [170] examine the predictive power of daily interdealer order flow series, including two heavily filtered commercially available order flow series, and the raw interdealer flows examined in Evans and Lyons [58]. They estimate single-equation regressions including order flow and interest differentials as independent variables. Measuring performance in terms of root mean squared error they find that these series do not outperform the random walk when information on future fundamentals is unavailable. In contrast, Rime et al. [166] find that interdealer order flow does outperform the random walk in predicting exchange rates one day ahead. Using three exchange rates (euro-dollar, dollar-yen, sterling-dollar) and associated Reuters (broker) order flow for one year they create forecasts based on what is, in essence, a structural VAR. They use the forecasts to create portfolios of the currencies. For forecast horizons ranging from 14 to 24 hours, the portfolios' Sharpe ratios range from 0.44 to 2.24 and average 1.59. Sharpe ratios for the random walk model and a UIP-based model are generally much lower.

What kind of information is carried by order flow? Evidence is consistent with the presence of both passively-acquired and actively-acquired fundamental information. Evans and Lyons [61] show that Citibank customer order flow has substantial predictive power for US and German GDP growth, inflation, and money growth at horizons ranging up to six months. The results are especially strong at longer horizons, where regressions using only order flow forecast between 21 percent and 58 percent of changes in the fundamental variables. (By contrast, regressions using only the lagged dependent variable or the spot rate generally forecast less than 10 percent.) This suggests that customer order flow concurrently reflects macro fundamentals and that the information may be passively acquired.

Evidence also suggests that order flow carries actively-acquired information about upcoming macro events and news releases. Froot and Ramadorai [74] show that State Street Corporation's institutional-investor flows have significant predictive power for changes in real interest rates at horizons up to thirty days. This would appear to be actively-acquired information.

Rime et al. [166] provide evidence that order flow carries information about upcoming macro news releases. Using thirty different news statistics (fifteen from the US, six from Europe, nine from the UK), the authors run the following regression:

$$\mathrm{Ann}^{ki}_{\mathrm{Thurs}+j} - E_{\mathrm{Thurs}}\mathrm{Ann}^{k}_{\mathrm{Thurs}+j} = \theta \sum_{i=1}^{j} \mathrm{OrderFlow}_{\mathrm{Thurs}+i} + \psi_{\mathrm{Thurs}+j} \,.$$

On the left is the news "surprise" for announcement-type k ($k = 1, 2, \ldots, 30$), meaning the difference between the announced figure and the median survey forecast for that announcement. On the right is cumulative interdealer order flow for the period between the survey and the announcement. The estimated relationships are generally quite strong: reported coefficients of determination range up to 0.91 and average 0.45. Since the news releases all lag the realization of the underlying macro aggregate by a month or more, the order flow would not reflect concurrent macro developments but instead appears to have been actively acquired.

This evidence suggests a strong focus on upcoming announcements among speculative agents, a focus that is quite evident in the market. Dealer communication with active customers includes regular – often daily – information on upcoming releases and extensive discussion of the macro context relevant for interpreting these releases. The agents that speculate on such announcements are typically leveraged investors.

Further support for the view that some private information is actively acquired in foreign exchange comes from Osler and Vandrovych [155]. They consider the information in price-contingent orders at the Royal Bank of Scotland with the agents placing those orders disaggregated into eight groups: leveraged investors, institutional investors, large corporations, middle-market corporations, broker-dealers, other banks, the bank's own spot dealers, and the bank's own exotic options desk. The price impact of executed orders, measured as the post-execution return over horizons ranging from five minutes to one week, is evaluated for the three major currency pairs. Results show that orders from leveraged investors have a strong and lasting impact while orders from institutional investors have little or no impact. Consistent with the possible dominance of levered investors, further evidence indicates financial order flow carries more information than commercial order flow, at least at short horizons [28,64,[154].

In short, the evidence is consistent with the hypothesis that customer order flow carries information about macro aggregates that is aggregated by dealers and then reflected in interdealer order flow. The evidence suggests that the customers acquire their information actively and perhaps passively as well.

The Evidence: Is the Information Really Fundamental?

Not all researchers are convinced that the information in foreign exchange order flow is fundamental. Berger et al. [12] highlight their findings (reported earlier) that the long-run price impact of interdealer order flow is smaller than the initial impact, and that explanatory power also declines at longer time horizons. They comment:

> The findings … are consistent with an interpretation of the association between exchange rate returns and order flow as reflecting principally a temporary – although relatively long-lasting – liquidity effect. They are also perhaps consistent with a behavioral interpretation … But our results appear to offer little support to the idea that order flow has a central role in driving long-run fundamental currency values – the 'strong flow-centric' view (p. 9).

Bacchetta and van Wincoop [7] suggest that this interpretation of the result may be more pessimistic than necessary regarding the relevance of fundamental information in order flow. Their model indicates that this pattern would be predicted when order flow reflects both fundamental and non-fundamental information. "In the short run, rational confusion plays an important role in disconnecting the exchange rate from observed fundamentals. Investors do not know whether an increase in the exchange rate is driven by an improvement in average private signals about future fundamentals or an increase in [non-fundamentals]. This implies that [non-fundamentals] have an amplified effect on the exchange rate …" (p. 554)

Evidence presented in Froot and Ramadorai [74] also suggests that the connection from order flow to exchange rates is transitory though long-lasting. Their institutional-flows dataset is large enough to permit a rigorous analysis of order flow and returns at horizons of a year or more (it extends from mid-1994 through early 2001 and covers 18 different currencies vs. the dollar), far longer than horizons considered in most other papers. Like Berger et al. [12], they find that the positive short-run correlation between order flow and returns peaks and then declines. Their correlation estimates reach zero at about 300 trading days and then become statistically negative. The authors note: "[O]ne can interpret the facts as suggesting that any impact of flows on currencies is transitory … [and] any information contained in flows is not about intrinsic value

per se (p. 1550)." Since this conclusion is based initially on crude correlations, the authors also undertake a sophisticated VAR decomposition of returns into permanent and transitory components, the results of which lead to the same overall conclusion. This finding cannot be explained in terms of the Bacchetta and van Wincoop [7] insights, since these do not imply the ultimate disappearance of the effect.

Could institutional-investor order flow carry information about macro fundamentals and yet have zero price impact after a year? It was suggested earlier that these observations are consistent when liquidity effects drive the connection from order flow to exchange rates. If real-money funds have roughly a one-year average investment horizon, then the initial upward impact of any, say, purchases – whether or not motivated by fundamental information – would ultimately be offset by a downward impact when the positions are unwound, leaving a zero impact at the one-year horizon. It is also worth noting that Froot and Ramadorai [74] analyze only institutional order flow. As noted earlier, institutional investors typically ignore the currency component of returns when making portfolio allocations, so one would not expect their order flow to have a permanent relation with exchange rates. The trades of other customers might still carry information.

Order flow could also have a transitory influence if exchange-rate expectations are not fully rational, as noted by both Berger et al. [12] and Froot and Ramadorai [74]. A tendency for professional exchange-rate forecasts to be biased and inefficient has been frequently documented [123]. This could explain why exchange rates apparently overreact to certain macro announcements [60]. As in Keynes's beauty contest, short-term traders could profit by correctly anticipating news and how other market participants will react to it, whether or not the reaction to news is rational.

The potential relevance of the behavioral perspective is underscored by extensive evidence for imperfect rationality among currency dealers presented in Oberlechner [147]. Indeed, dealers themselves typically claim that short-run dynamics are driven in part by "excess speculation" [36]. One potential source of excess speculative trading is overconfidence, a human tendency towards which has been extensively documented by psychologists [159]. Odean [150] shows that when agents overestimate the accuracy of their information – a common manifestation of overconfidence – they trade excessively and thereby generate excess volatility. Oberlechner and Osler [148] show, based on a sample of over 400 North American dealers, that currency dealers do not escape the tendency towards overconfidence. Further, they find that overconfident dealers are not driven out of the market: overconfidence is unrelated to a dealer's rank or trading longevity. This suggests that overconfidence may be a permanent structural feature of currency markets.

Information as an Incomplete Explanation

It is important to recognize that "information" is at best a partial explanation for the influence of order flow on exchange rates. An appeal to "information" quickly becomes circular in the absence of a successful economic model of the underlying connections between fundamentals and exchange rates.

This point is best clarified by illustration. Suppose a speculator expects a soon-to-be-released trade balance statistic to be higher than generally expected. According to the information hypothesis, three things happen: (i) the speculator evaluates whether a higher trade balance implies a stronger or weaker home currency and then trades accordingly; (ii) the associated order flow reveals to dealers whether the currency is over- or undervalued; (iii) as more dealers learn the information, it becomes progressively impounded in the exchange rate.

The information research just summarized concentrate on parts (ii) and (iii) of this story. But part (i) is also critical: Speculators must somehow evaluate the implications of the trade balance for the exchange rate in order to choose a position. To accomplish this, the speculator might rely on a model of how fundamentals and exchange rates are connected. But that model cannot itself rely on the information hypothesis without becoming circular: The information hypothesis asserts that exchange rates are determined by order flow because order flow carries information; circularity arises if the information in the order flow is that order flow determines exchange rates, which are determined by information. The speculator might alternatively ignore fundamentals and rely instead on a model of how other people think about fundamentals influence exchange rates. But of course this version of Keynes' beauty contest is equally prone to circularity.

The good news is that models intended to analyze the deep connections between fundamentals and exchange rates can now be based on more than just "assumption and hypotheses" [81]. Instead, they can have well-specified microfoundations based on our new understanding of the structure of currency markets and the exchange-rate determination process. Indeed, in the philosophical outlook of Karl Popper [161], reliance on the best available information is a key test of a model's scientific validity.

Price Discovery in Foreign Exchange

Research so far indicates that order flow influences exchange rates at least in part because it carries information brought to the market by customers. Research has also begun to clarify the exact mechanism through which the information becomes embodied in exchange rates.

Adverse Selection and Customer Spreads

Researchers have tended to assume that the price discovery process in foreign exchange conforms to the process discussed earlier in which adverse selection is key. This view of price discovery has been extensively elaborated in theoretical work, e. g., [100], and many of its predictions are fulfilled in the NYSE [14,89,158].

For structural reasons, this price discovery mechanism cannot apply directly to the foreign exchange market. The mechanism assumes a one-tier market, in which dealers interact only with customers, while foreign exchange is a two-tier market, in which dealers trade with customers in the first tier and trade with each other in the second tier. While this need not imply that adverse selection is entirely irrelevant, it does mean, at a minimum, that the framework needs adjustment before it can be relevant.

Empirical evidence shows that some of the key predictions of adverse selection do not hold in foreign exchange. The framework predicts, for example, that customer spreads are widest for the trades most likely to carry information, which would be large trades and trades with financial customers. The reverse is true, however. Osler et al. [154] analyzes the euro-dollar transactions of a single dealer over four months in 2001 and finds that customer spreads are smaller for large trades and for financial customers. The authors test three other implications of adverse selection, none of which gain support.

Further evidence for an inverse relationship between customer spreads and trade size is provided in Ding [47], which analyzes customer trading on a small electronic communication network. Direct evidence that spreads are narrowest for customer trades that carry the most information comes from Ramadorai [162], which analyzes daily flows through State Street's global custody operations. He finds that asset managers with the greatest skill in predicting (risk-adjusted) returns pay the smallest spreads. Overall it appears that adverse selection does not drive spreads in the customer foreign exchange market.

Adverse selection could, nonetheless, be an important determinant of spreads in the *interdealer* market. Information definitely appears to be asymmetric in that market [21], and the evidence is consistent with the hypothesis that spreads include a significant adverse selection component. Adverse-selection models predict two possible relations between trades and spreads. First, quoted spreads could widen with trade size if trade size is considered informative [52,78,126]. Evidence consistent with this prediction is presented in Lyons [120], but he examined a dealer who exclusively traded in the interdealer market, a form of trading that may no longer exist; later dealer studies fail to confirm this prediction [18,185]. It is possible, however, that trade direction is considered informative even while trade size is not, in which case spreads could still include a significant adverse selection component [99]. This is especially likely in limit-order markets, where the liquidity supplier (limit-order trader) often determines trade size, rather than the liquidity demander (market-order trader). Bjønnes and Rime [18] find strong evidence that trade direction is considered informative in the interdealer market and that adverse selection thereby influences interdealer spreads.

What Drives Customer Spreads?

The apparent irrelevance of adverse selection in the foreign exchange customer market raises an important question: What does drive customer spreads? It appears that structural factors may be at play, since spreads are also widest for the least informed trades in other two-tier markets, including the London Stock Exchange [86], the US corporate bond market [80], and the US municipal bond markets [84,90].

Osler et al. [154] reviews three hypotheses suggested in the broader microstructure literature that could explain this pattern in foreign exchange markets. First, the pattern could reflect the existence of fixed operating costs, which can be covered by a small spread on a large trade or a large spread on a small trade.

Fixed operating costs cannot, however, explain why commercial customers pay higher spreads than financial customers. The "strategic dealing" hypothesis suggests that dealers are strategically subsidizing informed-customer trades in order to gather information they can exploit during later interdealer trading [144,154].

Commercial customers could also pay higher spreads under the "market power" hypothesis of Green et al. [84]. This suggests that dealers have transitory market power relative to customers that do not carefully evaluate their execution quality or who do not know market conditions at the time they trade. Commercial customers in the foreign exchange market tend to be relatively unsophisticated: they are less familiar with standard market practice and typically do not monitor the market on an intraday ba-

sis. This may give dealers greater flexibility to extract wider spreads.

Price Discovery in Foreign Exchange

If adverse selection does not describe the price discovery process in foreign exchange, what does? Osler et al. [154] propose an alternative price discovery mechanism consistent with the foreign exchange market's two-tier structure. The mechanism focuses on how dealers choose to offload the inventory accumulated in customer trades. Dealers typically use limit orders to control inventory [18], but not always. Existing theory highlights important determinants of this choice [71,87]: market orders provide speedy execution at the cost of the bid-ask spread, while limit orders provide uncertain execution at an uncertain time but earn the bid-ask spread if execution does take place. This trade-off creates incentives such that market orders are more likely when a dealer's inventory is high, consistent with evidence in Bjønnes and Rime [18] and Osler et al. [154]. It also implies that a dealer should be more likely to place a market order after trading with an informed customer than after trading with an uninformed customer.

To clarify the logic of this second inference, suppose that an informed customer buys from a dealer that previously had zero inventory. That dealer will have three reasons to place a market order in the interdealer market: (i) information that exchange-rate is likely to rise; (ii) a non-zero (and therefore risky) inventory position; and (iii) information that his (short) inventory position is likely to lose value because prices are likely to rise. In consequence, after an informed customer buy transaction the dealer is relatively likely to place a market buy order. This raises the traded price, consistent with the customer's information.

After an uninformed customer purchase, by contrast, a dealer has only one reason to place a market order: risky inventory. If the dealer places a limit order rather than a market order then the uninformed-customer purchase would tend to be associated with negative downward returns, as the limit buy order is executed against a market sell.

One key testable implication of this proposed price discovery mechanism is that the likelihood of an interbank market order is higher after trades that are relatively likely to carry information, specifically financial-customer trades and large trades. Osler et al. [154] finds support for this implication using a probit analysis of their dealer's own trading choices. This indicates that the conditional probability that the dealer places an interbank market order is 9.5 percent for small commercial-customer trades and almost twice as high, at 18.5 percent, after small financial-customer trades. After large financial-customer trades – the most informed of all – the corresponding likelihood is 40.2 percent.

This proposed price discovery mechanism is consistent with much of the empirical evidence discussed so far. For example, it is consistent with the signs of the cointegrating relationships between returns and order flow: positive for financial customers, negative for commercial customers, positive for dealers. The positive cointegration between financial order flow and returns indicates that financial order flow carries fundamental information. The positive cointegration between interdealer order flow and returns suggests that dealers' market orders reflect the information in their customer order flow. The negative cointegration between commercial order flow and returns could also be an outcome of the price discovery hypothesis: if dealers place limit orders after trades with commercial customers (and if commercial customers are indeed relatively uninformed) then a commercial-customer buy will be reflected in an interdealer market sell order, with an associated price decline.

The mechanism is also consistent with Rime et al.'s [166] demonstration that interdealer order flow has strong predictive power for upcoming macro statistical releases, together with other evidence suggesting that leveraged investors bring the most information to the market. If leveraged investors are the most informed customers, then under this price discovery hypothesis interdealer order flow will reflect that group's trades. Since interdealer order flow has strong predictive power for upcoming macro releases, the implication is that leveraged investors devote much effort to forecasting those releases.

Summary and Future Directions

The currency microstructure evidence summarized here provides many new insights about the economics of the currency market and thus the economics of exchange-rate determination. The field thus merits its alternative moniker, "the new microeconomics of exchange rates."

The new evidence reveals that the proximate cause of most exchange-rate dynamics is order flow, which can be interpreted as net liquidity demand. The critical role of order flow is not, of course, in itself an economic explanation for exchange-rate dynamics. Recognizing this, the new literature provides evidence for three economic mechanisms through which order flow could influence exchange rates: inventory effects, liquidity effects, and information.

The information mechanism raises a critical question: What information is carried by order flow? The informa-

tion apparently originates with customers; dealers then see it reflected in their customer order flow. Some of the information may be dispersed, passively-acquired information about concurrent fundamentals. Some of the information appears to be actively-acquired information about upcoming macro news releases, with the most informative order flow coming from leveraged investors. Some of the information may be non-fundamental.

The literature also investigates the precise mechanism through which a customer's private information becomes reflected in exchange rates. This price discovery mechanism appears to differ strikingly from price discovery on the NYSE, a difference that could reflect a key structural difference across markets: foreign exchange dealers can trade with each other as well as with customers, but the NYSE has no interdealer market.

The literature addresses many questions of importance to researchers in microstructure per se. For example, what determines spreads in foreign exchange? Customer spreads in foreign exchange behave entirely differently from those on, say, the NYSE. On the NYSE, market makers try to protect themselves from informed traders and, if possible, they charge informed traders wider spreads. By contrast, foreign exchange dealers actively court the business of informed traders by quoting them narrow spreads. This could reflect the ability of currency dealers to trade with each other. Currency dealers seek trades with informed customers because the customers' order flow provides information the dealers can exploit in subsequent interdealer trades.

Our knowledge of this market still has big gaps, of course, which provide many fascinating questions for future research. A partial list includes the following:

1. Why do interdealer spreads vary inversely with trading volume and volatility? Does this pattern reflect fixed operating costs, the optimal bunching of liquidity traders, or something else?
2. What determines intraday variations in the price impact of order flow? While it looks like this is strongly influenced by the intraday pattern in interdealer spreads, there is little hard evidence on this point. What other factors might matter?
3. What determines longer-horizon variation in the price impact of order flow? The relevance of this question is enhanced, of course, by the evidence that variation in price impact contributes importantly to the persistence of volatility.
4. There is bound to be substantially more variation across types of financial customers, and across types of corporate customers, than has yet been identified. How much technical trading is there? What fraction of international investors disregard the currency component of returns when choosing portfolio allocations? Is this fraction changing?
5. There is still much to learn about the nature of the information provided by order flow, how dealers perceive that information, and how dealers use that information. Dealers claim they don't seek and don't use fundamental information but the evidence reveals that much of the information moving through the market is, in fact, related to fundamentals.
6. How strong are inventory, liquidity effects, and information effects in determining the connection between order flow and exchange rates?

Even when these questions have been addressed, however, the larger question – the question that originally motivated foreign exchange microstructure research – will still remain. In dealing with this question the foreign exchange microstructure researchers have followed Karl Popper's [161] agenda for scientific inquiry in its purest form. According to his philosophical perspective, good scientists produce evidence that "falsifies" existing paradigms and then create new paradigms consistent with all the evidence, old and new. The new evidence revealed by currency microstructure has falsified many aspects of traditional macro-based models while shedding new light on the economics of exchange-rate determination.

To develop the next generation of exchange-rate models, researchers now have at their disposal an extensive body of knowledge about how exchange rates are actually determined. This information brings with it the ability – and the responsibility – to construct models with well-specified microfoundations. A rigorous, empirically-relevant paradigm for short-run exchange-rate dynamics is much closer than it was a decade ago.

Bibliography

1. Admati AR, Pfleiderer P (1988) A Theory of Intraday Patterns: Volume and Price Variability. Rev Financial Stud 1:3–40
2. Akram FQ, Rime D, Sarno P (2005) Arbitrage in the Foreign Exchange Markets: Turning on the Microscope. Norges Bank Working Paper 2005-12
3. Andersen TG, Bollerslev T, Francis DX, Vega C (2003) Micro Effects of Macro Announcements: Real-Time Price Discovery in Foreign Exchange. Am Econ Rev 93:38–62
4. Anderson TG, Bollerslev T, Diebold FX, Vega C (2003) Micro Effects of Macro Announcements: Real-Time Price Discovery in Foreign Exchange. Am Econ Rev 93:38–62
5. Artus JR, Knight MD (1984) Issues in the Assessment of the Exchange Rates of Industrial Countries, Occasional Paper 29. International Monetary Fund, Washington, D.C.

6. Austin MP, Bates RG, Dempster MAH, Williams SN (2004) Adaptive Systems for Foreign Exchange Trading. Quant Finance 4:C37–45
7. Bacchetta P, van Wincoop E (2005) Can Information Heterogeneity Explain the Exchange Rate Determination Problem? Am Econ Rev 96:552
8. Baillie R, Bollerslev T (1989) The Message in Daily Exchange Rates: A Conditional Variance Tail. J Bus Econ Stat 7:297–305
9. Bank for International Settlements (2007) Triennial Central Bank Survey of Foreign Exchange and Derivatives Trading Activity. Basle
10. Barberis N, Thaler R (2002) A Survey of Behavioral Finance. NBER Working Paper No. 9222
11. Barker W (2007) The Global Foreign Exchange Market: Growth and Transformation. Bank Can Rev Autumn:3–12
12. Berger D, Chaboud A, Hjalmarsson E, Howorka E (2006) What Drives Volatility Persistence in the Foreign Exchange Market? Board of Governors of the Federal Reserve System, International Finance Discussion Papers No. 862
13. Berger D, Chaboud A, Chernenko S, Howorka E, Wright J (2006) Order Flow and Exchange Rate Dynamics in Electronic Brokerage System Data. Board of Governors of the Federal Reserve System, International Finance Discussion Papers No. 830
14. Bernhardt D, Hughson E (2002) Intraday trade in Dealership Markets. Euro Econ Rev 46:1697–1732
15. Bertsimas D (1998) Optimal Control of Execution Costs. J Financ Mark 1:1–50
16. Bertsimas D, Andrew WL (1998) Optimal Control of Execution Costs. J Financial Markets 1:1–50
17. Bertsimas, Lo A (1997)
18. Bjønnes GH, Rime D (2005) Dealer Behavior and Trading Systems in Foreign Exchange Markets. J Financial Econ 75:571–605
19. Bjønnes GH, Rime D, Solheim HOA (2005) Liquidity Provision in the Overnight Foreign Exchange Market. J Int Money Finance 24:175–196
20. Bjønnes GH, Rime D, Solheim HOA (2005) Volume and Volatility in the FOREIGN EXCHANGE Market: Does it Matter Who You Are? In: De Grauwe P (ed) Exchange Rate Modeling: Where Do We Stand? MIT Press, Cambridge
21. Bjønnes G, Osler C, Rime D (2007) Asymmetric Information in the Interdealer Foreign Exchange Market. Presented at the Third Annual Conference on Market Microstructure, Budapest, Hungary, 15 Sept 2007
22. Black, Fischer (1986) Noise. Finance 41:529–543
23. Bollerslev T, Melvin M (1994) Bid-ask Spreads and Volatility in the Foreign Exchange Market. J Int Econ 36:355–372
24. Brandt M, Kavajecz K (2005) Price Discovery in the US Treasury Market: The Impact of Order Flow and Liquidity on the Yield Curve. J Finance 59:2623–2654
25. Cai J, Cheung YL, Raymond SKL, Melvin M (2001) Once-in-a Generation Yen Volatility in 1998: Fundamentals, Intervention, and Order Flow. J Int Money Finance 20:327–347
26. Campbell JY, Shiller RJ (1988) Stock Prices, Earnings, and Expected Dividends. J Finance 43:661–676
27. Cao H, Evans M, Lyons KR (2006) Inventory Information. J Bus 79:325–363
28. Carpenter A, Wang J (2003) Sources of Private Information in FX Trading. Mimeo, University of New South Wales
29. Carlson JA, Melody L (2006) One Minute in the Life of the DM/US$: Public News in an Electronic Market. J Int Money Finance 25:1090–1102
30. Carlson JA, Dahl C, Osler C (2008) Short-Run Exchange-Rate Dynamics: Theory and Evidence. Typescript, Brandeis Univ
31. Chaboud AP, Chernenko SV, Howorka E, Iyer KRS, Liu D, Wright JH (2004) The High-Frequency Effects of US Macroeconomic Data Releases on Prices and Trading Activity in the Global Interdealer Foreign Exchange Market. Federal Reserve System Board of Governors, International Finance Discussion Papers Number 823
32. Chakravarty S (2000) Stealth Trading: Which Traders' Trades Move Stock Prices? J Financial Econ 61:289–307
33. Chang Y, Taylor SJ (2003) Information Arrivals and Intraday Exchange Rate Volatility. Int Financial Mark Inst Money 13:85–112
34. Chaboud A, Weinberg S (2002) Foreign Exchange Markets in the 1990s: Intraday Market Volatility and the Growth of Electronic Trading. B.I.S. Papers No. 12
35. Chaunzwa MJ (2006) Investigating the Economic Value Added of More Advanced Technical Indicators. Typescript, Brandeis University
36. Cheung YW, Chinn MD (2001) Currency Traders and Exchange Rate Dynamics: A Survey of the US Market. J Int Money Finance 20:439–471
37. Cheung YW, Chinn MD, Marsh I (2004) How Do UK-based Foreign Exchange Dealers Think Their Market Operates? Int J Finance Econ 9:289–306
38. Chordia T, Roll R, Subrahmanyam A (2002) Order Imbalance, Liquidity, and Market Returns. J Financial Econ 65:111–130
39. Copeland T, Galai D (1983) Information Effects on the Bid-Ask Spread. J Finance 38:1457–1469
40. Coval JD, Moskowitz TJ (2001), The Geography of Investment: Informed Trading and Asset Prices. J Political Econ 109(4):811–841
41. Covrig V, Melvin M (2002) Asymmetric Information and Price Discovery in the FX Market: Does Tokyo Know More About the Yen? J Empir Financ 9:271–285
42. Covrig V, Melvin M (2005) Tokyo Insiders and the Informational Efficiency of the Yen/Dollar Exchange Rate. Int J Finance Econ 10:185–193
43. Cross S (1998) All About … The Foreign Exchange Market in the United States. Federal Reserve Bank of New York. http://www.newyorkfed.org/education/addpub/usfxm/
44. Daniélsson J, Love R (2006) Feedback Trading. Int J Finance Econ 11:35–53
45. Delong B, Shleifer A, Summers L, Waldmann RJ (1990) Positive Feedback Investment Strategies and Destabilizing Rational Speculation. J Finance 45:379–395
46. Dewachter H (2001) Can Markov Switching Models Replicate Chartist Profits in the Foreign exchange Market? J Int Money Financ 20:25–41
47. Ding L (2006) Market Structure and Dealer's Quoting Behavior in the Foreign Exchange Market. Typescript, University of North Carolina at Chapel Hill
48. Dominguez K, Panthaki F (2006) What Defines News in Foreign Exchange Markets? J Int Money Finance 25:168–198
49. D'Souza C (2007) Where Does Price Discovery Occur in FX Markets? Bank of Canada Working Paper 2007-52
50. Dueker M, Neely CJ (2007) Can Markov Switching Models Predict Excess Foreign Exchange Raturns? J Bank Finance 31:279–296

51. Dunne P, Hau H, Moore M (2007) A Tale of Two Plattforms: Interdealer and Retail Quotes in the European Bond Markets. Presented at the Third Annual Conference on Microstructure, Magyar Bank, Budapest, September 2007
52. Easley D, O'Hara M (1987) Price Trade Size, and Information in Securities Markets. J Financial Econ 19:69–90
53. Ederington L, Jae HL (2001) Intraday Volatility in Interest-Rate and Foreign-Exchange Markets: ARCH, Announcement, and Seasonality Effects. J Futures Mark 21:517–552
54. Ehrmann M, Fratzscher M (2005) Exchange Rates and Fundamentals: New Evidence from Real-Time Data. J Int Money Finance 24:317–341
55. Euromoney (2006) FX Poll. http://www.euromoney.com/article.asp?ArticleID=1039514
56. Evans M (2002) FX Trading and Exchange Rate Dynamics. J Finance 57:2405–2448
57. Evans M, Lyons RK (2002) Information Integration and FX Trading. J Int Money Financ 21:807–831
58. Evans M, Lyons KR (2002) Order Flow and Exchange Rate Dynamics. J Political Econ 110(2002):170–180
59. Evans M, Lyons KR (2005) Do Currency Markets Absorb News Quickly? J Int Money Financ 24:197–217
60. Evans M, Lyons KR (2005) Meese-Rogoff Redux: Micro-Based Exchange-Rate Forecasting. Am Econ Rev Papers Proc 95:405–414
61. Evans M, Lyons KR (2007) Exchange-Rate Fundamentals and Order Flow. NBER Working Paper 13151
62. Evans M, Lyons KR (2008) How is Macro News Transmitted to Exchange Rates? J Financ Econ, forthcoming
63. Evans M, Lyons KR (2008) How is Macro News Transmitted to Exchange Rates? forthcoming, Journal of Financial Econmics
64. Fan M, Lyons RK (2003) Customer Trades and Extreme Events in Foreign exchange. In: Paul Mizen (ed) Monetary History, Exchange Rates and Financial Markets: Essays in Honour of Charles Goodhart. Edward Elgar, Northampton, pp 160–179
65. Federal Reserve Bank of New York (2004) Managing Operational Risk in Foreign Exchange
66. Federal Reserve Bank of New York (2007) http://www.newyorkfed.org/xml/gsds_transactions.html
67. Fleming M (2003) Measuring Treasury Market Liquidity. Federal Reserve Bank of New York. Econ Policy Rev 9:83–108
68. Fleming M, Mizrach B (2007) The Microstructure of a USTreasury ECN: The BrokerTec Platform. Typescript, Federal Reserve Bank of New York
69. Flood RP, Taylor MP (1996) Exchange-Rate Economics: What's Wrong with the Conventional Macro Approach. In: Jeffrey A Frankel, Galli G, Giovannini A (eds) The Microstructure of Foreign Exchange Markets. University of Chicago Press, Chicago, pp. 261–301
70. Foster DF, Viswanathan S (1993) Variations in Trading Volume, Return Volatility, and Trading Costs: Evidence on Recent Price Formation Models. J Finance 3:187–211
71. Foucault T (1999) Order Flow Composition and Trading Costs in a Dynamic Limit Order Market. J Financial Mark 2:99–134
72. Frankel JA, Galli G, Giovannini A (1996) Introduction. In: Frankel JA, Galli G, Giovannini A (eds) The Microstructure of Foreign Exchange Markets. University of Chicago Press, Chicago pp. 1–15
73. Frenkel JA, Mussa ML (1985) Asset markets, exchange rates, and the balance of payments. NBER working paper 1287
74. Froot K, Ramadorai Tarun (2005) Currency Returns, Intrinsic Value, and Institutional-Investor Flows. Finance 55:1535–1566
75. Galati G (2000) Trading Volume, Volatility, and Spreads in Foreign Exchange Markets: Evidence from Emerging Market Countries. BIS Working Paper 93
76. Gau YF (2005) Intraday Volatility in the Taipei Foreign Exchange Market. Pac Basin Finance J 13:471–487
77. Gehrig T, Menkhoff L (2004) The Use of Flow Analysis in Foreign Exchange: Exploratory Evidence. J Int Money Finance 23:573–594
78. Glosten L (1989) Insider Trading, Liquidity, and the Role of the Monopolist Specialist. J Bus 62:211–235
79. Glosten LR, Milgrom PR (1985) Bid, Ask, and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders. J Financial Econ 14:71–100
80. Goldstein MA, Hotchkiss ES, Sirri ER (2007) Transparency and Liquidity: A Controlled Experiment on Corporate Bonds. Rev Financial Stud 20:235–273
81. Goodhart C (1988) The Foreign Exchange Market: A Random Walk with a Dragging Anchor. Economica 55:437–60
82. Goodhart CAE, Hall SG, Henry SGB, Pesaran B (1993) News Effects in High-Frequency Model of the Sterling-Dollar Exchange Rate. J Appl Econ 8:1–13
83. Gradojevic N, Yang J (2006) Non-Linear, Non-Parametric, Non-Fundamental Exchange Rate Forecasting. J Forecast 25:227–245
84. Green RC, Hollifield B, Schurhoff N (2007) Financial Intermediation and the Costs of Trading in an Opaque Market. Rev Financial Stud 20:275–314
85. Hansch O, Naik N, Viswanathan S (1998) Do Inventories Matter in Dealership Markets? Some Evidence from the London Stock Exchange. J Finance 53:1623–1656
86. Hansch O, Naik N, Viswanathan S (1999) Preferencing, Internalization, Best Execution, and Dealer Profits. J Finance 54:1799–1828
87. Harris L (1998) Optimal Dynamic Order Submission Strategies in Some Stylized Trading Problems. Financial Mark Inst Instrum 7:1–75
88. Harris L, Gurel E (1986) Price and Volume Effects Associated with Changes in the S&P 500 List: New Evidence for the Existence of Price Pressures. J Finance 41:815–29
89. Harris L, Hasbrouck J (1996) Market vs. Limit orders: The SuperDOT Evidence on Order Submission Strategy. J Financial Quant Anal 31:213–231
90. Harris LE, Piwowar MS (2006) Secondary Trading Costs in the Municipal Bond Market. J Finance 61:1361–1397
91. Harris M, Raviv A (1993) Differences of Opinion Make a Horse Race. Rev Financial Stud 6:473–506
92. Hartmann P (1999) Trading Volumes and Transaction Costs in the Foreign Exchange Market: Evidence from Daily Dollar-Yen Spot Data. J Bank Finance 23:801–824
93. Hasbrouck J (1991) Measuring the Information Content of Stock Trades. J Finance 46:179–220
94. Hashimoto Y (2005) The Impact of the Japanes Banking Crisis on the Intraday FOREIGN EXCHANGE Market in Late 1997. J Asian Econ 16:205–222
95. Hau H (2001) Location Matters: An Examination of Trading Profits. J Finance 56(3):1959–1983
96. Hau H, Rey H (2003) Exchange Rates, Equity Prices, and Capital Flows. Rev Financial Stud 19:273–317
97. Hau H, Killeen W, Moore M (2002) How has the Euro Changed

the Foreign Exchange Market? Econ Policy, issue 34, pp 151–177
98. Hendershott T, Seasholes M (2006) Market Maker Inventories and Stock Prices. Presented at the Second Annual Microstructure Workshop, Ottawa, Canada, Oct 2006
99. Huang RD, Stoll HR (1997) The Components of the Bid-Ask Spread: A General Approach. Rev Financial Stud 10:995–1034
100. Holden CW, Subrahmanyam A (1992) Long-Lived Private Information and Imperfect Competition. J Finance 47:247–270
101. Holthausen RW, Leftwich RW, Mayers D (1990) Large-Block Transactions, the Speed of Response, and Temporary and Permanent Stock-Price Effects. J Financial Econ 26:71–95
102. Ito T (1990) Foreign Exchange Rate Expectations: Micro Survey Data. Am Econ Rev 80:434–449
103. Ito T, Hashimoto Y (2006) Intraday Seasonality in Activities of the Foreign Exchange Markets: Evidence from the Electronic Broking System. J Jap Int Econ 20:637–664
104. Jones CM, Kaul G, Lipson ML (1994) Transactions, Volume and Volatility. Rev Financial Stud 7:631–651
105. Jorion P (1996) Risk and Turnover in the Foreign Exchange Market. In: Frankel JA, Galli G, Giovaninni A (eds) The Microstructure of Foreign Exchange Markets. University of Chicago Press, Chicago, pp 19–37
106. Jovanovic B, Rousseau PL (2001) Liquidity Effects in the Bond Market. Federal Reserve Bank of Chicago Econ Perspect 25:17–35
107. Kandel E, Pearson ND (1995) Differential Interpretation of Public Signals and Trade in Speculative Markets. J Political Econ 103:831–872
108. Kearns J, Manners P (2006) The Impact of Monetary Policy on the Exchange Rate: A Study Using Intraday Data. International Journal of Central Banking 2:175–183
109. Killeen W, Lyons RK, Moore M (2005) Fixed versus Flexible: Lessons from EMS Order Flow. Forthcoming, J Int Money Finance
110. Killeen W, Lyons RK, Moore M (2006) Fixed versus Flexible: Lessons from EMS Order Flow. J Int Money Financ. 25:551–579
111. Kim M, Kim M (2001) Implied Volatility Dynamics in the Foreign Exchange Markets. J Int Money Finance 22:511–528
112. Klitgaard T, Weir L (2004) Exchange Rate Changes and Net positions of Speculators in the Futures Market. Federal Reserve Bank of New York Econ Policy Rev 10:17–28
113. Kothari SP (2001) Capital Markets Research in Accounting. J Account Econ 31:105–31
114. Kuhn TS (1970) The Structure of Scientific Revolutions, 2nd. edn. University of Chicago Press, Chicago
115. Kyle A (1985) Continuous Auctions and Insider Trading. Econometrica 53:1315–1335
116. Lane PR (2001) The New Open Economy Macroeconomics: A Survey. J Int Econ 54:235–266
117. LeBaron B (1998) Technical Trading Rules and Regime Shifts in Foreign Exchange. In Acar E, Satchell S (eds) Advanced Trading Rules. Butterworth-Heinemann, pp. 5–40
118. Love R, Payne R (2003) Macroeconomic News, Order Flows, and Exchange Rates. Typescript, London School of Economics
119. Lui Yu-Hon, Mole D (1998) The Use of Fundamental and Technical Analyses by Foreign Exchange Dealers: Hong Kong Evidence. J Int Money Financ 17:535–45
120. Lyons RK (1995) Tests of Microstructural Hypotheses in the Foreign Exchange Market. J Financ Econ 39:321–351
121. Lyons RK (1997) A Simultaneous Trade Model of the Foreign Exchange Hot Potato. J Int Econ 42:275–98
122. Lyons RK (2001) The Microstructure Approach to Exchange Rates. MIT Press, Cambridge and London
123. MacDonald R (2000) Expectations Formation and Risk in Three Financial Markets: Surveying What the Surveys Say. J Econ Surv 14:69–100
124. MacDonald R, Marsh I (1996) Currency Forecasters are Heterogeneous: Confirmation and Consequences. J Int Money Finance 15:665–685
125. Madhavan AN, Cheng M (1997) In Search of Liquidity: Block Trades in the Upstairs and Downstairs Markets. Rev Financial Stud 10:175–203
126. Madhavan AN, Smidt S (1991) A Bayesian Model of Intraday Specialist Pricing. J Financial Econ 30:99–134
127. Madhavan AN, Smidt S (1993) An Analysis of Changes in Specialist Inventories and Quotations. J Finance 48:1595–1628
128. Madhavan A, Richardson M, Roomans M (1997) Why Do Security Prices change? A Transaction-Level Analysis of NYSE Stocks. Rev Financial Stud 10:1035–1064
129. Malloy C (2005) The Geography of Equity Analysis. J Finance 60(2):719–756
130. Manaster S, Mann SC (1996) Life in the Pits: Competitive Market Making and Inventory Control. Rev Financial Stud 9:953–975
131. Marsh I, O'Rourke C (2005) Customer Order Flow and Exchange Rate Movements: Is There Really Information Content? Presented at the Norges Bank Conference on Equity and Currency Microstructure, Oslo
132. Martens M (2001) Forecasting Daily Exchange Rate Volatility Using Intraday Returns. J Int Money Financ 20:1–23
133. Fratscher M (2007) US Shocks and Global Exchange Rate Configurations. Presented at the NBERIFM meetings, Cambridge, MA, 10 July 2007
134. Meese RA, Rogoff K (1983) Empirical Exchange Rate Models of the Seventies: Do They Fit Out of Sample? J Int Econ 14:3–24
135. Mende A (2006) 09/11 and the USD/EUR Exchange Market. Appl Financial Econ 16:213–222
136. Mende A, Menkhoff L (2006) Profits and Speculation in Intra-Day Foreign Exchange Trading. J Financial Mark 9:223–245
137. Menkhoff L (1997) Examining the Use of Technical Currency Analysis. Int J Finance Econ 2:307–318
138. Menkhoff L (2001) Importance of Technical and Fundamental Analysis in Foreign Exchange Markets. Int J Finance Econ 6:81–93
139. Menkhoff L, Gehrig T (2006) Extended Evidence on the Use of Technical Analysis in Foreign Exchange. Int J Finance Econ 11:327–38
140. Menkhoff L, Taylor MP (2006) The Obstinate Passion of Foreign Exchange Professionals: Technical Analysis. University of Warwick, Department of Economics, The Warwick Economics Research Paper Series (TWERPS)
141. Menkhoff L, Osler C, Schmeling M (2007) Order-Choice Dynamics under Asymmetric Information: An Empirical Analysis. Typescript, Brandeis University
142. Milgrom P, Stokey N (1982) Information, Trade, and Common Knowledge. J Econ Theory 26:17–27
143. Morris S (1984) Trade with Heterogeneous Prior Beliefs and Asymmetric Information. Econometrica 62:1327–1347
144. Naik NY, Neuberger A, Viswanathan S (1999) Trade Disclosure

Regulation in Markets With Negotiated Trades. Rev Financial Stud 12:873–900
145. New York Stock Exchange (2007) Historical Facts and Statistics. www.nysedata.com/nysedata/Default.aspx?tabid=115
146. Oberlechner T (2001) Evaluation of Currencies in the Foreign Exchange Market: Attitudes and Expectations of Foreign Exchange Traders. Z Sozialpsychologie 3:180–188
147. Oberlechner T (2004) The Psychology of the Foreign Exchange Market. Wiley, Chichester
148. Oberlechner T, Osler C (2007) Overconfidence in Currency Markets. Typescript, Brandeis International Business School
149. Obstfeld M, Rogoff K (1995) Exchange Rate Dynamics Redux. J Political Econ 3:624–660
150. Odean T (1998) Volume, Volatility, Price, and Profit: When All Traders Are Above Average. J Finance 53:1887–1934
151. Osler CL (2003) Currency Orders and Exchange-Rate Dynamics: An Explanation for the Predictive Success of Technical Analysis. J Finance 58:1791–1819
152. Osler CL (2005) Stop-Loss Orders and Price Cascades in Currency Markets. J Int Money Finance 24:219–41
153. Osler CL (2006) Macro Lessons from Microstructure. Int J Finance Econ 11:55–80
154. Osler CL, Savaser T (2007) The Microstructure of Extreme Exchange-Rate Returns. Typescript, Brandeis International Business School
155. Osler CL, Vandrovych V (2007) Which Customers Bring Information to the in Foreign Exchange Market? Typescript, Brandeis International Business School
156. Pasquariello P, Vega C (2005) Informed and Strategic Order Flow in the Bond Markets, Working Paper
157. Payne R (2003) Informed trade in Spot Foreign Exchange Markets: An Empirical Investigation. J Int Econ 61:307–329
158. Peterson MA, Sirri ER (2003) Order Preferencing and Market Quality on US Equity Exchanges. Rev Financial Stud 16:385–415
159. Plous S (1993) The psychology of judgment and decision making. McGraw-Hill, New York
160. Pong S, Shackleton MB, Taylor SJ, Xu X (2004) Forecasting Currency Volatility: A Comparison of Implied Volatilities and AR(FI)MA Models. J Bank Financ 28:2541–2563
161. Popper K (1959) The Logic of Scientific Discovery. Routledge, United Kingdom
162. Ramadorai T (2006) Persistence, performance, and prices in foreign exchange markets. Oxford University Working Paper
163. Reiss PC, Werner IM (1995) Transaction Costs in Dealer Markets: Evidence from the London Stock Exchange. In: Andrew L (ed) The Industrial Organization and Regulation of the Securities Industry. University of Chicago Press, Chicago
164. Reiss PC, Werner IM (1998) Does risk sharing motivate interdealer trading? J Finance 53:1657–1703
165. Reiss PC, Werner IM (2004) Anonymity, adverse selection, and the sorting of interdealer trades. Rev Financial Stud 18:599–636
166. Rime D, Sarno L, Sojli E (2007) Exchange-rate Forecasting, Order Flow, and Macro Information. Norges Bank Working Paper 2007-2
167. Rogoff KS (1996) The Purchasing Power Parity Puzzle. J Econ Lit 34:647–668
168. Rosenbert JV, Traub LG (2006) Price Discovery in the Foreign Currency Futures and Spot Market. Typescript, Federal Reserve Bank of New York
169. Sager M, Taylor MP (2006) Under the Microscope: The Structure of the Foreign Exchange Market. Int J Finance Econ 11:81–95
170. Sager M, Taylor MP (2008) Commercially Available Order Flow Data and Exchange Rate Movements: Caveat Emptor. J Money Credit Bank 40:583–625
171. Savaser T (2006) Exchange Rate Response to Macro News: Through the Lens of Microstructure. Presented at the Bank of Canada Workshop on Equity and Currency Microstructure, October 2006
172. Savaser T (2007) Stop-Loss Orders and Macro News in Currency Markets. Presented at the Second Annual Microstructure Workshop, Ottawa, Canada, Oct 2006
173. Shiller R (1981) Do Stock Prices Move Too Much to Be Jusdtified by Subseuqent Changes in Dividends? Am Econ Rev 71:421–435
174. Shleifer A (1986) Do Demand Curves Slope Down? J Finance 41:579–90
175. Simon DP (1991) Segmentation in the Treasury Bill Market: Evidence from Cash Management Bills. J Financial Quant Anal 26:97–108
176. Simon DP (1994) Further Evidence on Segmentation in the Treasury Bill Market. J Bank Finance 18:139–151
177. Stoll H (1978) The Supply of Dealer Services in Securities Markets. J Finance 33:1133–1151
178. Taylor A, Farstrup A (2006) Active Currency Management: Arguments, Considerations, and Performance for Institutional Investors. CRA Rogers Casey International Equity Research, Darien Connecticut
179. Taylor MP (2002) Purchasing Power Parity and the Real Exchange Rate. IMF Staff Papers 48:65–105
180. Taylor MP, Allen H (1992) The Use of Technical Analysis in the Foreign Exchange Market. J Int Money Finance 11:304–314
181. Varian HR (1985) Divergence of Opinion in Complete Markets: A Note. J Finance 40:309–317
182. Varian HR (1989) Differences of Opinion in Financial Markets. In: Stone CC (ed) Financial Risk: Theory, Evidence, and Implications. Federal Reserve Bank of St. Louis, pp. 3–37
183. Wudunn S (1995) Japanese Delayed Letting US Know of Big Bank Loss. New York Times, 10 Oct
184. Yan B, Zivot E (2007) The Dynamics of Price Discovery. Typescript, University of Washington
185. Yao JM (1998) Market making in the interbank foreign exchange market. Stern School of Business, New York University Working Paper S-98

Mathematical Basis of Cellular Automata, Introduction to

Andrew Adamatzky
University of the West of England, Bristol, UK

A cellular automaton is a discrete universe with discrete time, discrete space and discrete states. Cells of the universe are arranged into regular structures called lattices or arrays. Each cell takes a finite number of states and updates its states in a discrete time, depending on the states

of its neighbors, and all cells update their states in parallel. Cellular automata are mathematical models of massively parallel computing; computational models of spatially extended non-linear physical, biological, chemical and social systems; and primary tools for studying large-scale complex systems.

Cellular automata are ubiquitous; they are objects of theoretical study and also tools of applied modeling in science and engineering. Purely for ease of representation, one can roughly split articles in this section into three groups: cellular automata theory, cellular automata models of computation and cellular automata models of natural phenomena. Many topics, however, belong to several groups.

We recommend that the reader begin with articles on history and modern analysis of classifying cellular automata based on internal characteristics of their cell-state transition functions, development of cellular automata configurations in space and time, and decidability of the cellular automata (see ► Identification of Cellular Automata). Studies in dynamical behavior are essential in progressing cellular automata theory. They include topological dynamics, for example, in relation to symbolic dynamics, surjectivity, and permutations (see ► Topological Dynamics of Cellular Automata); chaos, entropy and decidability of cellular automata behavior (see ► Chaotic Behavior of Cellular Automata), and insights into cellular automata as dynamical systems with invariant measures (see ► Ergodic Theory of Cellular Automata).

Self-reproducing patterns and gliders are amongst the most remarkable features of cellular automata. Certain cellular automata can reproduce configurations of cell-states, for example, the von Neumann universal constructor, and thus can be used in designs of self-replicating hardware (see ► Self-Replication and Cellular Automata). Gliders are translating oscillators, or traveling patterns, of non-quiescent states, for example, gliders in Conway's Game of Life. Gliders are particularly fascinating in two- and three-dimensional spaces (see ► Gliders in Cellular Automata).

Historically, an orthogonal lattice was the main substrate for cellular automata implementation. In the last decade the limit was lifted and nowadays you can find cellular automata on non-orthogonal lattices and tilings (see ► Cellular Automata in Triangular, Pentagonal and Hexagonal Tessellations), non-Euclidean geometries, such as hyperbolic spaces (see ► Cellular Automata in Hyperbolic Spaces) and various topological spaces (see ► Dynamics of Cellular Automata in Non-compact Spaces).

Typically, a cell neighborhood is fixed during cellular automaton development, and a cell updates its state depending on current states of its neighbors. But even in this very basic setup, the space-time dynamics of cellular automata are incredibly complex, as can be observed from analysis of the simplest one-dimensional automata where a transition rule applied to the sum of two states is equal to the sum of its actions on the two states separately (see ► Additive Cellular Automata). The automata dynamics becomes much richer if we allow the topology of the cell neighborhood to be updated dynamically during automaton development (see ► Structurally Dynamic Cellular Automata) or also allow a cell's state to become dependent on the cells' previous states (see ► Cellular Automata with Memory). Talking about non-standard cell-transition rules, we must mention cellular automata with injective global functions, where every configuration has exactly one preceding configuration (see ► Reversible Cellular Automata), and also cellular automata with quantum-bit cell-states and cell-transition functions incited by principles of quantum mechanics (see ► Quantum Cellular Automata).

The reader's initial excursion into the theory of cellular automata *themselves* can conclude in reading about decision problems of cellular automata expressed in terms of filling the plane using tiles with colored edges (see ► Tiling Problem and Undecidability in Cellular Automata) and about algebraic properties of cellular automata transformations, such as group representation of the Garden of Eden theorem and matrix representation of cellular automata (see ► Cellular Automata and Groups).

Firing squad synchronization is the oldest problem of cellular automaton computation: all cells of a one-dimensional cellular automaton are quiescent apart from one cell in the firing state; we wish to design minimal cell-state transition rules enabling all other cells to assume the firing state at the same time (see ► Firing Squad Synchronization Problem in Cellular Automata). Universality of cellular automata is another classical issue. Two kinds of universality are of most importance: computation universality, that is, an ability to compute any computable function or implement a functionally complete logical system, and intrinsic, or simulation universality, such as an ability to simulate any cellular automaton (see ► Cellular Automata, Universality of).

Readers can familiarize themselves with basics of space and time complexity cellular-automata parallel computing in (see ► Cellular Automata as Models of Parallel Computation). The knowledge will then be extended by measures of complexity, parallels between cellular automata and dynamics systems, Kolmogorov complexity of cellular automata (see ► Algorithmic Complexity and Cellular Automata) and studies of cellular automata as acceptors of

formal languages (see ▶ Cellular Automata and Language Theory). As demonstrated in (see ▶ Evolving Cellular Automata) cellular automata can be evolved to perform difficult computational tasks.

Cellular automata models of natural systems such as cell differentiation, road traffic, reaction-diffusion, and excitable media (see ▶ Cellular Automata Modeling of Physical Systems) are ideal candidates for studying all important phenomena of pattern growth (see ▶ Growth Phenomena in Cellular Automata); for studying transformation of a system's state from one phase to another (see ▶ Phase Transitions in Cellular Automata), and for studying the ability of a system to be attracted to the states where boundary between the system's phases is indistinguishable (see ▶ Self-organized Criticality and Cellular Automata). Cellular automata models can be designed, in principle, by reconstructing cell-state transition rules of cellular automata from snapshots of space-time dynamics of the system we wish to simulate (see ▶ Identification of Cellular Automata).

Maximum Principle in Optimal Control

Velimir Jurdjevic
University of Toronto, Toronto, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
The Calculus of Variations and the Maximum Principle
Variational Problems with Constraints
Maximum Principle on Manifolds
Abnormal Extrema and Singular Problems
Future Directions
Bibliography

Glossary

Manifolds Manifolds M are topological spaces that are covered by a set of compatible local charts. A local chart is a pair (U, ϕ) where U is an open set on M and ϕ is a homeomorphism from U onto an open set in $\mathbb{R}^n$. If ϕ is written as $\varphi(x) = (x_1, \ldots, x_n)$ then $x_1, \ldots, x_n$ are called coordinates of a point x in U. Charts are said to be compatible if for any two charts (U, ϕ) and (V, ψ) the mapping $\psi\varphi^{-1}\colon \phi(U \cap V) \to \psi(U \cap V)$ is smooth, i. e. admits derivatives of all orders. Such manifolds are also called smooth. Manifolds on which the above mappings $\psi\varphi^{-1}$ are analytic are called analytic.

Tangent and cotangent spaces The vector space of tangent vectors at a point x in a manifold M will be denoted by T_xM. The tangent bundle TM of M is the union $\cup\{T_xM : x \in M\}$. The cotangent space at x, consisting of all linear functions on T_xM will be denoted by T_x^*M. The cotangent bundle of M, denoted by T^*M is equal to the union $\cup\{T_x^*M : x \in M\}$.

Lie groups Lie groups G are analytic manifolds on which the group operations are compatible with the manifold structure, in the sense that both $(g, h) \to gh$ from $G \times G$ onto G and $g \to g^{-1}$ from G onto G are analytic. The set of all $n \times n$ non-singular matrices is an n^2 dimensional Lie group under matrix multiplication, and so is any closed subgroup of it.

Absolutely continuous curves A parametrized curve $x\colon [0, T] \to \mathbb{R}^n$ is said to be absolutely continuous if each derivative $(\mathrm{d}x_i)/(\mathrm{d}t)(t)$ exists almost everywhere in $[0, T]$ $\int_0^T |(\mathrm{d}x_i)/(\mathrm{d}t)(t)|\mathrm{d}t < \infty$, and $x(t_2) - x(t_1) = \int_{t_1}^{t_2} (\mathrm{d}x)/(\mathrm{d}t)(t)\mathrm{d}t$ for almost all points t_1 and t_2 in $[0, T]$. This notion extends to curves on manifolds by requiring that the above condition holds in any system of coordinates.

Control systems and reachable sets A control system is any system of differential equations in $\mathbb{R}^n$ or a manifold M of the form $\frac{\mathrm{d}x}{\mathrm{d}t}(t) = F(x(t), u(t))$, where $u(t) = (u_1(t), \ldots, u_m(t))$ is a function, called control function. Control functions are assumed to be in some specified class of functions $\mathcal{U}$, called the class of admissible controls. For purposes of optimization $\mathcal{U}$ is usually assumed to consist of functions measurable and bounded on compact intervals $[t_1, t_2]$ that take values in some a priori specified set U in $\mathbb{R}^m$.

Under the usual existence and uniqueness assumptions on the vector fields F, each control $u(t)\colon \mathbb{R} \to \mathbb{R}^m$ and each initial condition x_0 determine a unique solution $x(x_0, u, t)$ that passes through x_0 at $t = 0$. These solutions are also called trajectories. A point y is said to be reachable from x_0 at time T if there exists a control $u(t)$ defined on the interval $[0, T]$ such that $x(x_0, u, T) = y$. The set of points reachable from x_0 at time T is called the reachable set at time T. The set of points reachable from x_0 at any terminal time T is called the reachable set from x_0.

Riemannian manifolds A manifold M together with a positive definite quadratic form $\langle\, , \rangle\colon T_xM \times T_xM \to \mathbb{R}$ that varies smoothly with base point x is called Riemannian. If $x(t)$, $t \in [0, T]$ is a curve on M then its Riemannian length is given by

$\int_0^T \sqrt{\int_0^T \langle (\mathrm{d}x)/(\mathrm{d}t), (\mathrm{d}x)/(\mathrm{d}t)\rangle}\,\mathrm{d}t$. Riemannian metric is a generalization of the Euclidean metric in $\mathbb{R}^n$ given by $\langle x, y\rangle = \sum_{i=1}^n x_i y_i$.

Definition of the Subject

The Maximum Principle provides necessary conditions that the terminal point of a trajectory of a control system belongs to the boundary of its reachable set or to the boundary of its reachable set at a fixed time T. In practice, however, its utility lies in problems of optimization in which a given cost functional $\int_{t_0}^{t_1} f(x(t), u(t))\mathrm{d}t$ is to be minimized over the solutions $(x(t), u(t))$ of a control system $\frac{\mathrm{d}x}{\mathrm{d}t} = F(x(t), u(t))$ that conform to some specified boundary conditions at the end points of the interval $[t_0, t_1]$. For then, the extended optimal trajectory $(\bar{x}_0(t), \bar{x}(t), \bar{u}(t))$ with $\bar{x}_0(t) = \int_{t_0}^t f(\bar{x}(t), \bar{u}(t))\mathrm{d}t$ must be on the boundary of the reachable set associated with extended control system

$$\frac{\mathrm{d}x_0}{\mathrm{d}t} = f(x(t), u(t))\,, \quad \frac{\mathrm{d}x}{\mathrm{d}t} = F(x(t), u(t))\,. \tag{1}$$

In the original publication [24] the Maximum Principle is stated for control systems in $\mathbb{R}^n$ in which the controls $u(t) = (u_1(t), \ldots, u_m(t))$ take values in an arbitrary subset U of $\mathbb{R}^m$ and are measurable and bounded on each interval $[0, T]$, under the assumptions that the cost functional f and each coordinate F^i of the vector field F in Eq. (1) together with the derivatives $\frac{\partial F^i}{\partial x^j}$ are continuous on $\mathbb{R}^n \times \bar{U}$, where $\bar{U}$ denotes the topological closure of U.

A control function $\bar{u}(t)$ and the corresponding trajectory $\bar{x}(t)$ of $\frac{\mathrm{d}x}{\mathrm{d}t} = F(x(t), \bar{u}(t))$ each defined on an interval $[0, T]$ are said to be *optimal* relative to the given boundary submanifolds S_0 and S_1 of $\mathbb{R}^n$ if $\bar{x}(0) \in S_0, \bar{x}(T) \in S_1$, and

$$\int_0^T f(\bar{x}(t), \bar{u}(t))\mathrm{d}t \le \int_0^S f(x(t), u(t))\mathrm{d}t \tag{2}$$

for any other solution $x(t)$ of $\frac{\mathrm{d}x}{\mathrm{d}t} = F(x(t), u(t))$ that satisfies $x(0) \in S_0$ and $x(S) \in S_1$. Then

Proposition 1 (The Maximum Principle (MP)) *Suppose that $(\bar{u}(t), \bar{x}(t))$ is an optimal pair on the interval $[0, T]$. Then there exist an absolutely continuous curve $p(t) = (p_1(t), \ldots, p_n(t))$ on the interval $[0, T]$ and a multiplier $p_0 \le 0$ with the following properties:*

1. *$p_0^2 + \sum_{i=1}^n p_i^2(t) > 0$, for all $t \in [0, T]$,*
2. *$\frac{dp_i}{dt}(t) = -\frac{\partial \mathcal{H}_{p_0}}{\partial x_i}(\bar{x}(t), p(t), \bar{u}(t))$, $i = 1, \ldots, n$, a. e. in $[0, T]$ where $\mathcal{H}_{p_0}(x, p, u) = p_0 f(x, u) + \sum_{i=1}^n p_i F^i(x, u)$.*
3. *$\mathcal{H}_{p_0}(\bar{x}(t), p(t), \bar{u}(t)) = \mathcal{M}(t)$ a. e. in $[0, T]$ where $\mathcal{M}(t)$ denotes the maximum value of $\mathcal{H}_{p_0}(\bar{x}(t), p(t), u)$ relative to the controls $u \in U$. Furthermore,*
4. *$\mathcal{M}(t) = 0$ for all $t \in [0, T]$.*
5. *$\langle p(0), v\rangle = 0$, $v \in T_{\bar{x}(0)}S_0$ and $\langle p(T), v\rangle = 0$, $v \in T_{\bar{x}(T)}S_1$. These conditions, known as the transversality conditions, become void when the manifolds S_0 and S_1 reduce to single points x_0 and x_1.*

The maximum principle is also valid for optimal problems in which the length of the interval $[0, T]$ is fixed, in the sense that the optimal pair $(\bar{x}(t), \bar{u}(t))$ satisfies $\int_0^T f(\bar{x}(t), \bar{u}(t))\mathrm{d}t \le \int_0^T f(x(t), u(t))\mathrm{d}t$ for any other trajectory $(x(t), u(t))$ with $x(0) \in S_0$ and $x(T) \in S_1$. In this context the maximum principle asserts the existence of a curve $p(t)$ and the multiplier p_0 subject to the same conditions as stated above except that the maximal function $\mathcal{M}(t)$ must be constant in the interval $[0, T]$ and need not be necessarily equal to zero. These two versions of the (**MP**) are equivalent in the sense that each implies the other [1].

Pairs $(x(t), p(t))$ of curves that satisfy the conditions of the maximum principle are called *extremal*. The extremal curves that correspond to the multiplier $p_0 \neq 0$ are called *normal* while the ones that correspond to $p_0 = 0$ are called *abnormal*. In the normal case it is customary to reduce the multiplier p_0 to $p_0 = -1$.

Since the original publication, however, the maximum principle has been adapted to control problems on arbitrary manifolds [12] and has also been extended to more general situations in which the vector fields that define the control system are locally Lipschitz rather than continuously differentiable [9,27]. On this level of generality the Maximum Principle stands out as a fundamental principle in differential topology that not only merges classical calculus of variations with mechanics, differential geometry and optimal control, but also reorients the classical knowledge in two major ways:

1. It shows that there is a natural "energy" Hamiltonian for arbitrary variational problems and not just for problems of mathematical physics. The passage to the appropriate Hamiltonians is direct and bypasses the Euler–Lagrange equation. The merits of this observation are not only limited to problems with inequality constraints for which the Euler-equation is not applicable; they also extend to the integration procedure of the extremal equations obtained through the integrals of motion.
2. The Hamiltonian formalism associated with (**MP**) further enriched with geometric control theory makes direct contact with the theory of Hamiltonian systems

and symplectic geometry. In this larger context, the maximum principle brings fresh insights to these classical fields and also makes their theory available for problems of optimal control.

Introduction

The Maximum Principle of Pontryagin and his collaborators [1,12,24] is a generalization of C. Weierstrass' necessary conditions for strong minima [29] and is based on the topological fact that an optimal solution must terminate on the boundary of the extended reachable set formed by the competing curves and their integral costs. An important novelty of Pontryagin's approach to the calculus of variations consists of liberating the variations along an optimal trajectory of the constricting condition that they must terminate at the given boundary data. Control theoretic context induces a natural class of variations that generates a cone of directions locally tangent to the reachable set at the terminal point defined by the optimal trajectory. As a consequence of optimality, the direction of the decreasing cost cannot be in the interior of this cone. This observation leads to the separation theorem, a generalization of the classic Legendre transform in the calculus of variations, that ultimately produces the appropriate Hamiltonian function. The maximum principle asserts that the Hamiltonian that corresponds to the optimal trajectory must be maximal relative to the completing directions, and it also asserts that each optimal trajectory is the projection of an integral curve of the corresponding Hamiltonian field.

The methodology used in the original publication extends the maximum principle to optimal control problems on arbitrary manifolds where, combined with Lie theoretic criteria for reachable sets, it stands out as a most important tool of optimal control available for problems of mathematical physics and differential geometry. Much of this article, particularly the selection of the illustrating examples is devoted to justifying this claim.

The exposition begins with comparisons between (**MP**) and the classical theory of the calculus of variations in the absence of constraints. Then it proceeds to optimal control problems in $\mathbb{R}^n$ with constraints amenable by the (**MP**) stated in Proposition 1. This section, illustrated by two famous problems of classical theory, the geodesic problem on the ellipsoid of C.J.G. Jacobi and the mechanical problem of C. Newman–J. Moser is deliberately treated by control theoretic means, partly to illustrate the effectiveness of (**MP**), but mostly to motivate extensions to arbitrary manifolds and to signal important connections to the theory of integrable systems.

The exposition then shifts to the geometric version of the maximum principle for control problems on arbitrary manifolds, with a brief discussion of the symplectic structure of the cotangent bundle. The maximum principle is first stated for extremal trajectories (that terminate on the boundary of the reachable sets) and then specialized to optimal control problems. The passage from the first to the second clarifies the role of the multiplier. There is brief discussion of canonical coordinates as a bridge that connects geometric version to the original formulation in $\mathbb{R}^n$ and also leads to left invariant adaptations of the maximum principle for problems on Lie groups.

Left invariant variational control problems on Lie groups make contact with completely integrable Hamiltonian systems, Lax pairs and the existence of spectral parameters. For that reason there is a section on the Poisson manifolds and the symplectic structure of the coadjoint orbits of a Lie group G which is an essential ingredient of the theory of integrable systems.

The exposition ends with a brief discussion of the abnormal and singular extremals.

The Calculus of Variations and the Maximum Principle

The simplest problems in the calculus of variations can be formulated as optimal control problems of the form $U = \mathbb{R}^n$, $\frac{\mathrm{d}x}{\mathrm{d}t}(t) = u(t)$, $S_0 = \{x_0\}$ and $S_1 = \{x_1\}$, with one subtle qualification connected with the basic terminology. In the literature on the calculus of variations one usually assumes that there is a curve $\bar{x}(t)$ defined on an interval $[0, T]$ that provides a "local minimum" for the integral $\int_0^T f(x(t), \frac{\mathrm{d}x}{\mathrm{d}t}(t))\mathrm{d}t$ in the sense that there is a "neighborhood" N in the space of curves on $[0, T]$ such that $\int_0^T f(\bar{x}(t), \frac{\mathrm{d}\bar{x}}{\mathrm{d}t}(t))\mathrm{d}t \leq \int_0^T f(x(t), \frac{\mathrm{d}x}{\mathrm{d}t}(t))\mathrm{d}t$ for any other curve $x(t) \in N$ that satisfies the same boundary conditions as $\bar{x}(t)$.

There are two distinctive topologies, strong and weak on the space of admissible curves relative to which optimality is defined. In strong topology admissible curves consist of absolutely continuous curves with bounded derivatives on $[0, T]$ in which an ϵ neighborhood N consists of all admissible curves $x(t)$ such that $\|\bar{x}(t) - x(t)\| < \epsilon$ for all $t \in [0, T]$. In this setting local minima are called strong minima. For weak minima admissible curves consist of continuously differentiable curves on $[0, T]$ with an ϵ neighborhood of $\bar{x}(t)$ defined by

$$\|\bar{x}(t) - x(t)\| + \left\| \frac{\mathrm{d}\bar{x}}{\mathrm{d}t}(t) - \frac{\mathrm{d}x}{\mathrm{d}t}(t) \right\| < \epsilon \text{ for all } t \in [0, T]\,. \tag{3}$$

Evidently, any strong local minimum that is continuously differentiable is also a weak local minimum. The converse, however, may not hold (see p. 341 in [12]).

The Maximum Principle is a necessary condition for local strong minima $\bar{x}(t)$ under suitable restriction of the state space. It is a consequence of conditions (1) and (3) in Proposition 1 that the multiplier p_0 can not be equal to 0. Then it may be normalized to $p_0 = -1$ and (**MP**) can be rephrased in terms of the excess function of Weierstrass [7,29] as

$$f(\bar{x}(t), u) - f\left(\bar{x}(t), \frac{\mathrm{d}\bar{x}}{\mathrm{d}t}(t)\right) \geq \sum_{i=1}^{n} \frac{\partial f}{\partial u_i}\left(\bar{x}(t), \frac{\mathrm{d}\bar{x}}{\mathrm{d}t}(t)\right)(\bar{u}_i - u_i), \text{ for all } u \in \mathbb{R}^n, \tag{4}$$

because the critical points of $\mathcal{H}(\bar{x}(t), p(t), u) = -f(\bar{x}(t), u) + \sum_{i=1}^{n} p_i(t)u_i$ relative to $u \in \mathbb{R}^n$ are given by $p_i(t) = \frac{\partial f}{\partial u_i}(\bar{x}(t), u)$. Since $\frac{\mathrm{d}\bar{x}}{\mathrm{d}t}(t) = \bar{u}(t)$ yields the maximum of $\mathcal{H}(\bar{x}(t), p(t), u)$ it follows that

$$p_i(t) = \frac{\partial f}{\partial u_i}(\bar{x}(t), \bar{u}(t)) . \tag{5}$$

Combining Eq. (5) with condition 2 of the maximum principle yields the Euler–Lagrange equation in integrated form $\frac{\partial f}{\partial u_i}(\bar{x}(t), \bar{u}(t)) - \int_0^t \frac{\partial f}{\mathrm{d}x_i}(\bar{x}(\tau), \bar{u}(\tau))\,\mathrm{d}\tau = c$, with c a constant, which under further differentiability assumptions can be stated in its usual form

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial f}{\partial u_i}(\bar{x}(t), \bar{u}(t))\right) - \frac{\partial f}{\mathrm{d}x_i}(\bar{x}(\tau), \bar{u}(\tau))\,\mathrm{d}\tau = 0 . \tag{6}$$

As a way of illustration consider

Example 2 (The harmonic oscillator) The problem of minimizing $\int_0^T \frac{1}{2}(mu^2 - kx^2)\mathrm{d}t$ over the trajectories of $\frac{\mathrm{d}x}{\mathrm{d}t} = u$ leads to the family of Hamiltonians $\mathcal{H}_u = -\frac{1}{2}(mu^2 - kx^2) + pu$. According to the Maximum Principle every optimal trajectory $x(t)$ is the projection of a curve $p(t)$ that satisfies $\frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{\partial \mathcal{H}_u}{\partial x} = -kx(t)$, subject to the maximality condition that

$$-\frac{1}{2}(mu(t)^2 - kx(t)^2) + p(t)u(t) \geq -\frac{1}{2}(mv^2 - kx(t)^2) + p(t)v \tag{7}$$

for any choice of v. That implies that the optimal control that generates $x(t)$ is of the form

$$u(t) = \frac{1}{m}p(t) \tag{8}$$

which then further implies that optimal solutions are the integral curves of a single Hamiltonian function $H = \frac{1}{2m}p^2 + \frac{1}{2}kx^2$. This Hamiltonian is equal to the total energy of the oscillator. The Euler–Lagrange equation $\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = \frac{\mathrm{d}u}{\mathrm{d}t} = -\frac{k}{m}x(t)$ is easily obtained by differentiating, but there is no need for it since all the information is already contained in the Hamiltonian equations.

It follows from the above that the projections of the extremal curves are given by $x(t) = A\cos\sqrt{t\frac{k}{m}} + B\sin\sqrt{t\frac{k}{m}}$ for arbitrary constants A and B. It can be shown, by a separate argument [12], that the preceding curves are optimal if and only if $\sqrt{t\frac{k}{m}} \leq \pi$. This example also illustrates that *the Principle of Least Action* in mechanics may be valid only on small time intervals $[t_0, t_1]$ (in the sense that it yields the least action).

Variational Problems with Constraints

The early applications of (**MP**) are best illustrated through time optimal problems for linear control systems $\frac{\mathrm{d}x}{\mathrm{d}t}(t) = Ax(t) + Bu(t)$ with control functions $u(t) = (u^1(t), \ldots, u^r(t))$ taking values in a compact neighborhood U of the origin in $\mathbb{R}^r$. Here A and B are matrices of appropriate dimensions such that the "controllability" matrix $[B, AB, A^2B, \ldots, A^{n-1}B]$ is of rank n. In this situation if $(x(t), u(t))$ is a time optimal pair then the corresponding Hamiltonian $\mathcal{H}_{p_0}(x, p, u) = p_0 + \langle p(t), Ax(t) + Bu(t)\rangle$ defined by the Maximum Principle is equal to 0 almost everywhere and is also maximal almost everywhere relative to the controls $u \in U$. The rank condition together with the fact that $p(t)$ is the solution of a linear differential equation $\frac{\mathrm{d}p}{\mathrm{d}t} = -A^T p(t)$ easily implies that the control $u(t)$ cannot take value in the interior of U for any convergent sequence of times $\{t_n\}$ otherwise, $\lim p(t_n) = 0$, and therefore $p_0 = 0$, which in turn contradicts condition 1 of Proposition 1. It then follows that each optimal control $u(t)$ must take values on the boundary of U for all but possibly finitely many times. This fact is known as the *Bang-Bang Principle*, since when U is a polyhedral set then optimal control "bangs" from one face of U to another.

In general, however, optimal controls may take values both in the interior and on the boundary of U, and the extremal curves could be concatenations of pieces generated by controls with values in the interior of U and the pieces generated by controls with values on the boundary. Such concatenations may exhibit dramatic oscillations at the juncture points, as in the following example.

Example 3 (Fuller's problem) Minimize $\frac{1}{2}\int_0^T x_1^2(t)\mathrm{d}t$ over the trajectories of $\frac{\mathrm{d}x_1}{\mathrm{d}t} = x_2$, $\frac{\mathrm{d}x_2}{\mathrm{d}t} = u(t)$ subject to the

constraint that $|u(t)| \leq 1$. Here T may be taken fixed and sufficiently large that admits an optimal trajectory $x(t) = (x_1(t), x_2(t))$ that transfers the initial point a to a given terminal point b in T units of time.

Evidently, the zero trajectory is generated by zero control and is optimal relative to $a = b = 0$. The (**MP**) reveals that any other optimal trajectory is a concatenation of this singular trajectory and a bang-bang trajectory. At the point of juncture, whether leaving or entering the origin, optimal control oscillates infinitely often between the boundary values ± 1 (in fact, the oscillations occur in a geometric sequence [12,19]). This behavior is known as *Fuller's phenomenon*. Since the original discovery Fuller's phenomena have been detected in many situations [18,30].

The Maximum Principle is the only tool available in the literature for dealing with variational problems exhibiting such chattering behavior. However, even for problems of geometry and mechanics which are amenable by the classical methods the (**MP**) offers certain advantages as the following examples demonstrate.

Example 4 (Ellipsoidal Geodesics) This problem, initiated and solved by C.G. Jacobi in 1839 [23] consists of finding the curves of minimal length on a general ellipsoid

$$\langle x, A^{-1}x\rangle = \frac{x_1^2}{a_1^2} + \frac{x_2^2}{a_2^2} + \cdots + \frac{x_n^2}{a_n^2} = 1\,. \tag{9}$$

Recall that the length of a curve $x(t)$ on an interval $[0, T]$ is given by $\int_0^T \|\frac{\mathrm{d}x}{\mathrm{d}t}(t)\|\mathrm{d}t$, where $\|\frac{\mathrm{d}x}{\mathrm{d}t}(t)\| = \sqrt{((\mathrm{d}x_1)/(\mathrm{d}t))^2 + \cdots + ((\mathrm{d}x_n)/(\mathrm{d}t))^2}$. This problem can be recast as an optimal control problem either as a time optimal problem when the curves are parametrized by arc length, or as the problem of minimizing the energy functional $\frac{1}{2}\int_0^T \|\frac{\mathrm{d}x}{\mathrm{d}t}(t)\|^2\mathrm{d}t$ over arbitrary curves [15]. In the latter formulation the associated optimal control problems consists of minimizing the integral $\frac{1}{2}\int_0^T \|u(t)\|^2\mathrm{d}t$ over the trajectories of $\frac{\mathrm{d}x}{\mathrm{d}t}(t) = u(t)$ that satisfy $\langle x(t), A^{-1}x(t)\rangle = 1$.

Since there are no abnormal extremals in this case, it follows that the adjoint curve $p(t)$ associated with an optimal trajectory $x(t)$ must maximize $\mathcal{H}_u = -\frac{1}{2}\|u\|^2 + \langle p(t), u\rangle$ on the cotangent bundle of the manifold $x\colon \langle x(t), A^{-1}x(t)\rangle - 1 = 0$.

The latter is naturally identified with the constrains $G_1 = G_2 = 0$, where $G_1 = \langle x(t), A^{-1}x(t)\rangle - 1$ and $G_2 = \langle p, A^{-1}x\rangle$. According to the Lagrange multiplier rule the correct maximum of $\mathcal{H}_u$ subject to these constrains is obtained by maximizing the function $G_u = -\frac{1}{2}||u||^2 + \langle p, u\rangle + \lambda_1 G_1 + \lambda_2 G_2$ relative to u. The the maximal Hamiltonian is given by $H = \frac{1}{2}||p||^2 + \lambda_1 G_1 + \lambda_2 G_2$ obtained by substituting $u = p$. The correct multipliers λ_1 and λ_2 are determined by requiring that the integral curves of the associated Hamiltonian vector field $\overrightarrow{H}$ respect the constraints $G_1 = G_2 = 0$. It follows that $\lambda_1 = \frac{\langle A^{-1}p, p\rangle}{2||A^{-1}x||}$ and $\lambda_2 = 0$. Hence the solutions are the projections of the integral curves of

$$H = \frac{1}{2}\|p\|^2 + \frac{\langle A^{-1}p, p\rangle}{2\|A^{-1}x\|^2} G_1 \text{ restricted to } G_1 = G_2 = 0\,. \tag{10}$$

The corresponding equations are

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= \frac{\partial H}{\partial p} = p\,, \quad \text{and} \\ \frac{\mathrm{d}p}{\mathrm{d}t} &= -\frac{\partial H}{\partial x} = \frac{\langle A^{-1}p, p\rangle}{\|A^{-1}x\|^2}\,. \end{aligned} \tag{11}$$

The projections of these equations on the ellipsoid that reside on the energy level $H = \frac{1}{2}$ are called *geodesics*. It is well known in differential geometry that geodesics are only locally optimal (up to the first conjugate point).

The relatively simple case, when the ellipsoid degenerates to the sphere occurs when $A = I$. Then the above Hamiltonian reduces to $H = \frac{1}{2}(\|p\|^2) + \frac{\langle p, p\rangle}{2\|x\|^2}(\|x\|^2 - 1)$ and the corresponding equations are given by

$$\frac{\mathrm{d}x}{\mathrm{d}t} = p, \quad \frac{\mathrm{d}p}{\mathrm{d}t} = \|p\|^2 x\,. \tag{12}$$

It follows by an easy calculation that the projections $x(t)$ of Eqs. (12) evolve along the great circles because $x(t) \wedge \frac{\mathrm{d}x}{\mathrm{d}t}(t) = x(0) \wedge \frac{\mathrm{d}x}{\mathrm{d}t}(0)$.

The solutions in the general case can be obtained either by the method of *separation of variables* inspired by the work of C.G.J. Jacobi (see also [2]), or by the isospectral deformation methods discovered by J. Moser [22], which are somewhat mysteriously linked to the following problem.

Example 5 (Newmann–Moser problem) This problem concerns the motion of a point mass on the unit sphere $\langle x, x\rangle = 1$ that moves in a force field with quadratic potential $V = \frac{1}{2}\langle x, Ax\rangle$ with A an arbitrary symmetric matrix. Then the Principle of Least Action applied to the Lagrangian $L = \frac{1}{2}\|\frac{\mathrm{d}x}{\mathrm{d}t}\|^2 - \frac{1}{2}\langle x, Ax\rangle$ defines an optimal control problem by maximizing $\int_0^T L(x(t), u(t))dt$ over the trajectories of $\frac{\mathrm{d}x}{\mathrm{d}t} = u(t)$, subject to the constraint $G_1 = ||x||^2 - 1 = 0$, whose extremal equations etc. In fact, $H = \frac{1}{2}(\|p\|^2 + \langle x, Ax\rangle) + \lambda_1 G_1 + \lambda_2 G_2$, where $G_2 = \langle x, p\rangle$, $\lambda_1 = -\frac{\langle p, x\rangle}{\|x\|^2}$ and $\lambda_2 = \frac{\|p\|^2}{2\|x\|^2} - \langle Ax, x\rangle$.

$$\frac{\mathrm{d}x}{\mathrm{d}t} = p, \frac{\mathrm{d}p}{\mathrm{d}t} = -Ax + (\langle Ax, x - \|p\|^2\rangle x \tag{13}$$

Equations (13) can be recast in matrix form

$$\frac{\mathrm{d}P}{\mathrm{d}t}(t) = [K(t), P(t)]\,, \quad \frac{\mathrm{d}K}{\mathrm{d}t}(t) = [P(t), A] \tag{14}$$

with $P(t) = x(t) \otimes x(t) - I$ and $K(t) = x(t) \wedge p(t) = x(t) \otimes p(t) - p(t) \otimes x(t)$, where $[M, N]$ denotes the matrix commutator $NM - MN$. Equations (14) admit a Lax pair representation in terms of scalar parameter λ.

$$\frac{\mathrm{d}L_\lambda}{\mathrm{d}t} = \left[\frac{1}{\lambda}P(t), L_\lambda(t)\right], \quad \text{where } L_\lambda(t) = P(t) - \lambda K(t) - \lambda^2 A\,, \tag{15}$$

that provides a basis for Moser's method of proving integrability of Eqs. (13). This method exploits the fact that the spectrum of $L_\lambda(t)$ is constant, and hence the functions $\phi_{k,\lambda} = \mathrm{Trace}(L_\lambda^k)$ are constants of motion for each λ and $k > 0$. Moreover, these functions are in involution with each other (to be explained in the next section). Remarkably, Eqs. (11) can be transformed to Eqs. (13) from which then can be inferred that the geodesic ellipsoidal problem is also integrable [22]. It will be shown later that this example is a particular case of a more general situation in which the same integration methods are available.

Maximum Principle on Manifolds

The formulation of the maximum principle for control systems on arbitrary manifolds requires additional geometric concepts and terminology [1,12]. Let M denote an n-dimensional smooth manifold with T_xM and T_x^*M denoting the tangent and the cotangent space at a point $x \in M$. The tangent bundle TM is equal to the union $\cup\{T_xM\colon x \in M\}$, and similarly the cotangent bundle T^*M is equal to $\{T_x^*M\colon x \in M\}$. In each of these cases there is a natural bundle projection π on the base manifold. In particular, $x \in M$ is the projection of a point $\xi \in T^*M$ if and only if $\xi \in T_x^*M$.

The cotangent bundle T^*M has a canonical symplectic form ω that turns T^*M into a symplectic manifold. This implies that for each function H on T^*M there is a vector field $\vec{H}$ on T^*M defined by $V(H) = \omega(\vec{H}, V)$ for all vector fields V on T^*M. In the symplectic context H is called *Hamiltonian* and $\vec{H}$ is called the *Hamiltonian vector field corresponding to* H.

Each vector field X on M defines a function $H_X(\xi) = \xi(X(x))$ on T^*M. The corresponding Hamiltonian vector field $\vec{H}_X$ is called the *Hamiltonian lift of* X. In particular, control systems $\frac{\mathrm{d}x}{\mathrm{d}t}(t) = F(x(t), u(t))$ lift to Hamiltonians $\mathcal{H}_u$ parametrized by controls with $\mathcal{H}_u(\xi) = \xi(X_u(x)) = \xi(F(x, u))$, $\xi \in T_x^*M$.

With these notations in place consider a control system $\frac{\mathrm{d}x}{\mathrm{d}t}(t) = F(x(t), u(t))$ on M with control functions $u(t)$ taking values in an arbitrary set U in R^m. Suppose that the system satisfies the same assumptions as in Proposition 1. Let $\mathcal{A}_{x_0}(T)$ denote the reachable set from x_0 at time T and let $\mathcal{A}_{x_0} = \cup\{\mathcal{A}_{x_0}(T)\colon T \geq 0\}$. The control u that generates trajectory $x(t)$ from x_0 to the boundary of either $\mathcal{A}_{x_0}(T)$ or $\mathcal{A}_{x_0}$ is called *extremal*. For extremal trajectories the following version of the Maximum Principle is available.

Proposition 6 (Geometric maximum principle (GMP)) *Suppose that a trajectory $x(t)$ corresponds to an extremal control $u(t)$ on an interval $[0, T]$. Then there exists an absolutely continuous curve $\xi(t)$ in T^*M in the interval $[0, T]$ that satisfies the following conditions:*

1. *$x(t)$ is the projection of $\xi(t)$ in $[0, T]$ and $\frac{\mathrm{d}\xi}{dt}(t) = \vec{\mathcal{H}}_{u(t)}(\xi(t))$, a. e. in $[0, T]$, where $\mathcal{H}_{u(t)}(\xi) = \xi(F(x, \bar{u}(t)))$, $\xi \in T_x^*M$.*
2. *$\xi(t) \neq 0$ for all $t \in [0, T]$.*
3. *$\mathcal{H}_{u(t)}(\xi(t)) = \xi(t)(F(x(t), u(t))) \geq \xi(t)(F(x(t), v))$ for all $v \in U$ a. e. in $[0, T]$.*
4. *If the extremal curve is extremal relative to the fixed terminal time then $\mathcal{H}_{u(t)}(\xi(t))$ is constant a. e. in $[0, T]$, otherwise, $\mathcal{H}_{u(t)}(\xi(t)) = 0$ a. e. in $[0, T]$.*

An absolutely continuous curve $\xi(t)$ that satisfies the conditions of the Maximum Principle is called an *extremal*.

Problems of optimization in which a cost functional $\int_0^T f(x(t), u(t))\mathrm{d}t$ is to be minimized over the trajectories of a control system in M subject to the prescribed boundary conditions with terminal time either fixed or variable are reduced to boundary problems defined above in the same manner as described in the introductory part. Then each optimal trajectory $x(t)$ is equal to the projection of an extremal for the extended system (1) on $\tilde{M} = \mathbb{R} \times M$ relative to the extended initial conditions $\tilde{x}_0 = (0, x_0)$. If $T^*\tilde{M}$ is identified with $\mathbb{R}^* \times T^*M$ and its points $\tilde{\xi}$ are written as (ξ_0, ξ) the Hamiltonian lifts of the extended control system are of the form

$$\mathcal{H}_u(\xi_0, \xi) = \xi_0 f(x, u) + \xi(F(x, \bar{u}(t))), \quad \xi_0 \in \mathbb{R}^*, \xi \in T_x^*M\,. \tag{16}$$

For each extremal curve $(\xi_0(t), \xi(t))$ associated with the extended system $\xi_0(t)$ is constant because the Hamiltonian $\mathcal{H}_u(\xi_0, \xi)$ is constant along the first factor of the extended state space. In particular, $\xi_0 \leq 0$ along the extremals that correspond to optimal trajectories. As before, such extremals are classified as normal or abnormal de-

pending whether ξ_0 is equal to zero or not, and in the normal case the variable ξ_0 is traditionally reduced to -1. The Maximum principle is then stated in terms of the reduced Hamiltonian on T^*M in which ξ_0 appear as parameters. In this context Proposition 6 can be rephrased as follows:

Proposition 7 *Let $x(t)$ denote an optimal trajectory on an interval $[0, T]$ generated by a control $u(t)$. Then there exist a number $\xi_0 \in \{0, -1\}$ and an absolutely continuous curve $\xi(t)$ in T^*M defined on the interval $[0, T]$ that projects onto $x(t)$ and satisfies:*

1. *$\xi(t) \neq 0$ whenever $\xi_0 = 0$.*
2. *$\frac{d\xi}{dt}(t) = \vec{\mathcal{H}}_{u(t)}(\xi_0, \xi(t))$ a. e. on $[0, T]$.*
3. *$\mathcal{H}_{u(t)}(\xi_0, \xi(t)) \geq \mathcal{H}_v(\xi_0, \xi(t))$ for any $v \in U$, a. e. in $[0, T]$.*
   When the initial and the terminal points are replaced by the submanifolds S_0 and S_1 there are transversality conditions:
4. *$\xi(0)(v) = 0$, $v \in T_{x(0)}S_0$ and $\xi(T)(v) = 0$, $v \in T_{x(T)}S_1$.*

The above version of (**MP**) coincides with the Euclidean version when the variables are expressed in terms of the canonical coordinates. Canonical coordinates are defined as follows.

Any choice of coordinates $(x_1, x_2, \ldots, x_n)$ on M induces coordinates $(v_1, \ldots, v_n)$ of vectors in T_xM relative to the basis $\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n}$ and it also induces coordinates $(p_1, p_2, \ldots, p_n)$ of covectors in T_x^*M relative to the dual basis $dx_1, dx_2, \ldots, dx_n$. Then $(x_1, x_2, \ldots, x_n, p_1, p_2, \ldots, p_n)$ serves as a system of coordinates for points ξ in T^*M. These coordinates in turn define coordinates for tangent vectors in $T_\xi(T^*M)$ relative to the basis $\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n}, \frac{\partial}{\partial p_1}, \ldots, \frac{\partial}{\partial p_n}$. The symplectic form ω can then be expressed in terms of vector fields $\mathcal{X} = \sum_{i=1}^n V_i \frac{\partial}{\partial x_i} + P_i \frac{\partial}{\partial p_i}$ and $\mathcal{Y} = \sum_{i=1}^n W_i \frac{\partial}{\partial x_i} + Q_i \frac{\partial}{\partial p_i}$ as

$$\omega_{(x,p)}(X, Y) = \sum_{i=1}^n Q_i V_i - P_i W_i \,. \tag{17}$$

The correspondence between functions H and their Hamiltonian fields $\vec{H}$ is given by

$$\vec{H}(x, p) = \sum_{i=1}^n \frac{\partial H}{\partial p_i}\frac{\partial}{\partial x_i} - \frac{\partial H}{\partial x_i}\frac{\partial}{\partial p_i} \,, \tag{18}$$

and the integral curves $(x(t), p(t))$ of $\vec{H}$ are given by the usual differential equations

$$\frac{dx_i}{dt} = \frac{\partial H}{\partial p_i}, \quad \frac{dp_i}{dt} = -\frac{\partial H}{\partial x_i}, \quad i = 1, \ldots, n \,. \tag{19}$$

Any choice of coordinates on T^*M that preserves Eq. (17) is called *canonical*. Canonical coordinates could be defined alternatively as the coordinates that preserve the Hamiltonian equations (19). In terms of the canonical coordinates the Maximum Principle of Proposition 7 coincides with the original version in Proposition 1.

Optimal Control Problems on Lie Groups

Canonical coordinates are not suitable for all variational problems. For instance, variational problems in geometry and mechanics often have symmetries that govern the solutions; to take advantage of these symmetries it may be necessary to use coordinates that are compatible with the symmetries and which may not necessarily be canonical. That is particularly true for optimal problems on Lie groups that are either right or left invariant.

To elaborate further, assume that G denotes a Lie group (matrix group for simplicity) and that $\mathfrak{g}$ denotes its Lie algebra. A vector field X on G is called *left-invariant* if for every $g \in G$, $X(g) = gA$ for some matrix A in $\mathfrak{g}$, i. e., X is determined by its tangent vector at the group identity. Similarly, *right-invariant* vector fields are defined as the right translations of matrices in $\mathfrak{g}$. Both the left and the right invariant vector fields form a frame on G, that is, $T_gG = \{gA\colon A \in \mathfrak{g}\} = \{Ag\colon A \in \mathfrak{g}\}$ for all $g \in G$. Therefore, the tangent bundle TG can be realized as the product $G \times \mathfrak{g}$ either via the left translations $(g, A) \to gA$, or via the right translations $(g, A) \to Ag$. Similarly, the cotangent bundle T^*G can be realized in two ways as $G \times \mathfrak{g}^*$ with $\mathfrak{g}^*$ equal to the dual of $\mathfrak{g}$. In the left-invariant realization $\xi \in T_g^*G$ is identified with $(g, l) \in G \times \mathfrak{g}^*$ via the formula $l(A) = \xi(gA)$ for all $A \in \mathfrak{g}$.

For optimal control problems which are left-invariant it is natural to identify T^*G as $G \times \mathfrak{g}^*$ via the left translations, and likewise identify T^*G as $G \times \mathfrak{g}^*$ via the right translations for right-invariant problems. In both cases the realization $T^*G = G \times \mathfrak{g}^*$ rules out canonical coordinates (assuming that G is non-abelian) and hence the Hamiltonian equations (19) take on a different form.

For concreteness sake assume that $T^*G = G \times \mathfrak{g}^*$ is realized via the left translations. Then it is natural to realize (T^*G) as $(G \times \mathfrak{g}^*) \times (\mathfrak{g} \times \mathfrak{g}^*)$ where $((g, l), (A, f)) \in (G \times \mathfrak{g}^*) \times (\mathfrak{g} \times \mathfrak{g}^*)$ denotes tangent vector (A, f) at the point (g, l). In this representation of $T(T^*G)$ the symplectic form ω is given by the following expression:

$$\omega_{(g,l)}((A_1, f_1), (A_2, f_2)) = f_2(A_1) - f_1(A_2) - l([A_1, A_2]) \,. \tag{20}$$

Functions on $G \times \mathfrak{g}^*$ that are constant over the first factor, i. e., functions on $\mathfrak{g}^*$, are called *left-invariant*

Hamiltonians. If H is left invariant then integral curves $(g(t), l(t))$ of the corresponding Hamiltonian vector field $\vec{H}$ are given by

$$\begin{aligned} \frac{\mathrm{d}g}{\mathrm{d}t}(t) &= g(t)\,\mathrm{d}H(l(t))\,, \\ \frac{\mathrm{d}l}{\mathrm{d}t}(t) &= -ad^*(\mathrm{d}H(l(t))(l(t))) \end{aligned} \tag{21}$$

where $\mathrm{d}H$ denotes the differential of H, and where $ad^*(A)\colon \mathfrak{g}^* \to \mathfrak{g}^*$ is given by $(ad^*(A)(l))(X) = l([A, X])$, $l \in \mathfrak{g}^*$, $X \in \mathfrak{g}$, [12,13].

On semi-simple Lie groups linear functions l in $\mathfrak{g}^*$ can be identified with matrices L in $\mathfrak{g}$ via an invariant quadratic form $\langle\,,\,\rangle$ so that Eqs. (21) become

$$\begin{aligned} \frac{\mathrm{d}g}{\mathrm{d}t}(t) &= g(t)\,\mathrm{d}H(l(t))\,, \\ \frac{\mathrm{d}L}{\mathrm{d}t}(t) &= [\mathrm{d}H(l(t), L(t))]\,. \end{aligned} \tag{22}$$

For instance, the problem of minimizing the integral $\frac{1}{2}\int_0^T \|u(t)\|^2 \mathrm{d}t$ over the trajectories of

$$\frac{\mathrm{d}g}{\mathrm{d}t}(t) = g(t)\left(A_0 + \sum_{i=1}^m u_i(t)A_i\right), \quad u \in \mathbb{R}^m \tag{23}$$

with $A_0, A_1, \ldots, A_m$ matrices in $\mathfrak{g}$ leads to the following Hamiltonians:

1. (Normal extrema) $H = \frac{1}{2}\sum_{i=1}^m H_i^2 + H_0$.
2. (Abnormal extrema) $H = H_0 + \sum_{i=1}^m u_i(t)H_i$, subject to $H_i = 0$, $i = 1, \ldots, m$,

with each H_i equal to the Hamiltonian lift of the left invariant vector field $g \to gA_i$. In the left invariant representation of T^*G each H_i is a linear function on $\mathfrak{g}^*$, i. e., $H_i(l) = l(A_i)$ and consequently both Hamiltonians above are left-invariant. In the abnormal case $\mathrm{d}H = A_0 + \sum_{i=1}^m u_i(t)A_i$, and

$$\begin{aligned} \frac{\mathrm{d}g}{\mathrm{d}t}(t) &= g(t)\left(A_0 + \sum_{i=1}^m u_i(t)A_i\right), \\ \frac{\mathrm{d}L}{\mathrm{d}t}(t) &= \left[A_0 + \sum_{i=1}^m u_i(t)A_i, L(t)\right], \end{aligned} \tag{24}$$

$$H_1(t) = H_2(t) = \cdots = H_m(t) = 0\,, \tag{25}$$

are the corresponding extremal equations.

In the normal case the extremal controls are given by $u_i = H_i$, $i = 1, \ldots, m$, and the corresponding Hamiltonian system is given by Eqs. (22) with $\mathrm{d}H = A_0 + \sum_{i=1}^m H_i(t)A_i$.

Left invariant Hamiltonian systems [Eqs. (21) and (22)] always admit certain functions, called integrals of motion, which are constant along their solutions. Hamiltonians which admit a "maximal" number of functionally independent integrals of motion are called integrable. For left invariant Hamiltonians on Lie groups there is a deep and beautiful theory directed to characterizing integrable systems [10,13,23]. This topic is discussed in more detail below.

Poisson Bracket, Involution and Integrability

Integrals of motion are most conveniently discussed in terms of the Poisson bracket. For that reason it becomes necessary to introduce the notion of a Poisson manifold. A manifold M that admits a bilinear and skew symmetric form $\{\,,\,\}\colon C^\infty(M) \times C^\infty(M) \to C^\infty(M)$ that satisfies the Jacobi identity $\{f, \{g, h\}\} + \{h, \{f, g\}\} + \{g, \{h, f\}\} = 0$ and is a derivation $\{fg, h\} = f\{g, h\} + g\{f, h\}$ is called *Poisson*. It is known that every symplectic manifold is Poisson, and it is also known that every Poisson manifold admits a foliation in which each leaf is symplectic. In particular, the cotangent bundle T^*M is a Poisson manifold with $\{f, h\}(\xi) = \omega_\xi(\vec{f}(\xi), \vec{h}(\xi))$ for all functions f and h. It is easy to show that F is an integral of motion for H if and only if $\{F, H\} = 0$ from which it follows that F is an integral of motion for H if and only if H is an integral of motion for F. Functions F and H for which $\{F, H\} = 0$ are also said to be in *involution*. A function H on a $2n$ dimensional symplectic manifold S is said to be *integrable or completely integrable* if there exist n functions $\varphi_1, \ldots, \varphi_n$ with $\varphi_1 = H$ which are functionally independent and further satisfy $\{\varphi_i, \varphi_j\} = 0$ for each i and j. It is known that such a system of functions is dimensionally maximal.

On Lie groups the dual $\mathfrak{g}^*$ of the Lie algebra $\mathfrak{g}$ inherits a Poisson structure from the symplectic form ω [Eq. (17)], with $\{f, h\}(l) = l([\mathrm{d}f, \mathrm{d}h])$ for any functions f and h on $\mathfrak{g}^*$. In the literature on Hamiltonian systems this structure is often called Lie–Poisson. The symplectic leaves induced by the Poisson–Lie structure coincide with the coadjoint orbits of G and the solutions of the equation $\frac{\mathrm{d}l}{\mathrm{d}t}(t) = -ad^*(\mathrm{d}H(l(t))(l(t)))$ associated with Eqs. (21) evolve on coadjoint orbits of G. Most of the literature on integrable systems is devoted to integrability properties of the above equation considered as a Hamiltonian equation on a coadjoint orbit, or to its semi-simple counterpart $\frac{\mathrm{d}L}{\mathrm{d}t}(t) = [\mathrm{d}H(L(t)), L(t)]$. In this setting integrability is relative to the Poisson–Lie structure on each orbit, which may be of different dimensions. However, integrability can also be defined relative to the entire cotangent structure in which case the system is integrable when-

ever the number of independent integrals in involution is equal to the dimension of G. Leaving these subtleties aside, left-invariant Hamiltonian systems on Lie groups (22) always admit certain integrals of motion. They fall into two classes:

1. Hamiltonian lifts of right-invariant vector fields on G Poisson commute with any left-invariant Hamiltonian because right-invariant vector fields commute with left-invariant vector fields. If $X(g) = Ag$ denote a right invariant vector field then its Hamiltonian lift F_A is equal to $F_A(L, g) = \langle L, g^{-1}Ag\rangle$. In view of the formula $\{F_A, F_B\} = F_{[A,B]}$, the maximal number of functionally independent Hamiltonian lifts of right-invariant vector fields is equal to the rank of $\mathfrak{g}$. The rank of a semi-simple Lie algebra $\mathfrak{g}$ is equal to the dimension of a maximal abelian subalgebra of $\mathfrak{g}$. Such maximal abelian subalgebras are called *Cartan subalgebras*.
2. Each eigenvalue of $L(t)$ is a constant of motion for $\frac{\mathrm{d}L}{\mathrm{d}t}(t) = [\mathrm{d}H(L(t), L(t))]$. If $\lambda(L)$ and $\mu(L)$ denote eigenvalues of L then $\{\lambda, \mu\} = 0$. Equivalently, the spectral functions $\varphi_k(L) = \mathrm{Trace}(L^k)$ are in involution and Poisson commute with H.

For three-dimensional Lie groups the above integrals are sufficient for complete integrability. For instance, every left-invariant Hamiltonian is completely integrable on SO_3 [12]. In general, it is difficult to determine when the above integrals of motion can be extended to a completely integrable system of functions for a given Hamiltonian H [10,13,25]. Affirmative answers are known only in the exceptional cases in which there are additional symmetries. For instance, integrable system (15) is a particular case of the following more general situation.

Example 8 Suppose that a semi-simple Lie group G admits an involutive automorphism $\sigma \neq I$ that splits the Lie algebra $\mathfrak{g}$ of G into a direct sum $\mathfrak{g} = \mathfrak{p} + \mathfrak{k}$ with $\mathfrak{k} = \{A\colon \sigma_*(A) = A\}$ and $\mathfrak{p} = \{A\colon \sigma_*(A) = -A\}$. Such a decomposition is known as a *Cartan decomposition* and the following Lie algebraic conditions hold

$$[\mathfrak{p}, \mathfrak{p}] = \mathfrak{k}, \quad [\mathfrak{p}, \mathfrak{k}] = \mathfrak{p}, \quad [\mathfrak{k}, \mathfrak{k}] \subseteq \mathfrak{k}\,. \tag{26}$$

Then $L \in \mathfrak{g}$ can be written as $L = K + P$ with $P \in \mathfrak{p}$ and $K \in \mathfrak{k}$. Assume that $H(L) = \frac{1}{2}\langle K, K\rangle + \langle A, P\rangle$, for some $A \in \mathfrak{p}$ where $\langle\,,\,\rangle$ denotes a scalar multiple of the Cartan–Killing form that is positive definite on $\mathfrak{k}$. This Hamiltonian describes normal extrema for the problem of minimizing the integral $\frac{1}{2}\int_0^T \|U(t)\|^2 \mathrm{d}t$ over the trajectories of

$$\frac{\mathrm{d}g}{\mathrm{d}t}(t) = g(t)(A + U(t)), \quad U(t) \in \mathfrak{k}\,. \tag{27}$$

With the aid of the above decomposition the Hamiltonian equations (22) associated with $\vec{H}$ are given by

$$\begin{aligned} \frac{\mathrm{d}g}{\mathrm{d}t} &= g(A + K)\,, \\ \frac{\mathrm{d}K}{\mathrm{d}t} &= [A, P]\,, \\ \frac{\mathrm{d}P}{\mathrm{d}t} &= [A, K] + [K, P]\,. \end{aligned} \tag{28}$$

Equations $\frac{\mathrm{d}K}{\mathrm{d}t} = [A, P]$, $\frac{\mathrm{d}P}{\mathrm{d}t} = [A, K] + [K, P]$ admit two distinct types of integrals of motion. The first type is a consequence of the spectral parameter representation

$$\frac{\mathrm{d}M_\lambda}{\mathrm{d}t} = [N_\lambda, M_\lambda], \text{ with } M_\lambda = P - \lambda K + (\lambda^2 - 1)A \text{ and } N_\lambda = \frac{1}{\lambda}(P - A)\,, \tag{29}$$

from which it follows that $\phi_{\lambda,k} = \mathrm{Trace}(M_\lambda^k)$ are constants of motion for each $\lambda \in \mathbb{R}$ and $k \in Z^+$. The second type follows from the observation that $[A, P]$ is orthogonal (relative to the Cartan–Killing form) to the subalgebra $\mathfrak{k}_0 = \{X \in \mathfrak{k}\colon [A, X] = 0\}$. Hence the projection of $K(t)$ on $\mathfrak{k}_0$ is constant. In many situations these two types of integrals of motion are sufficient for complete integrability [23].

Abnormal Extrema and Singular Problems

For simplicity of exposition the discussion will be confined to control affine systems written explicitly as

$$\frac{\mathrm{d}x}{\mathrm{d}t} = X_0(x) + \sum_{i=1}^m u_i(t)X_i(x)\,, \quad u = (u_1, \ldots, u_m) \in U\,, \tag{30}$$

with $X_0, X_1, \ldots, X_m$ smooth vector fields on a smooth manifold M. Recall that abnormal extrema are absolutely continuous curves $\xi(t) \neq 0$ in T^*M satisfying

1. $\frac{\mathrm{d}\xi}{\mathrm{d}t}(t) = \vec{H}_0(\xi(t)) + \sum_{i=1}^m u_i(t)\vec{H}_i(\xi(t))$ a. e. for some admissible control $u(t)$ where $\vec{H}_0, \ldots, \vec{H}_m$ denote the Hamiltonian vector fields associated with Hamiltonian lifts $H_i(\xi) = \xi(X_i(x))$, $\xi \in T_x^*M$, $i = 0, \ldots, m$, and
2. $H_0(\xi(t)) + \sum_{i=1}^m u_i(t)H_i(\xi(t)) \geq H_0(\xi(t)) + \sum_{i=1}^m v_iH_i(\xi(t))$ a. e. for any $v = (v_1, \ldots, v_m) \in U$.

Abnormal extremals satisfy the conditions of the Maximum Principle independently of any cost functional and can be studied in their own right. However, in practice, they are usually linked to some fixed optimization problem, such as for instance the problem of minimizing

$\frac{1}{2}\int_0^T \|u(t)\|^2 \mathrm{d}t$. In such a case there are several situations that may arise:

1. An optimal trajectory is only the projection of a normal extremal curve.
2. An optimal trajectory is the projection of both a normal and an abnormal extremal curve.
3. An optimal trajectory is only the projection of an abnormal curve (strictly abnormal case).

When U is an open subset of $\mathbb{R}$ the maximality condition (2) implies that $H_i(\xi(t)) = 0$, $i = 1, \ldots, m$. Then extremal curves which project onto optimal trajectories satisfy another set of constraints $\{H_i, H_j\}(\xi(t)) = \xi(t)([X_i, X_j](x(t))) = 0$, $1 \le i, j \le m$, known as the *Goh condition* in the literature on control theory [1]. Case (1) occurs when $X_1(x(t)), \ldots, X_m(x(t)), [X_i, X_j](x(t))$, $1 \le i, j \le m$ span $T_{x(t)}M$.

The remaining cases occur in the situations where higher order Lie brackets among $X_0, \ldots, X_m$ are required to generate the entire tangent space $T_{x(t)}M$ along an optimal trajectory $x(t)$. In the second case abnormal extrema can be ignored since every optimal trajectory is the projection of a normal extremal curve. However, that is no longer true in Case (3) as the following example shows.

Example 9 (Montgomery [21]) In this example $M = \mathbb{R}^3$ with its points parametrized by cylindrical coordinates r, θ, z. The optimal control problem consists of minimizing $\frac{1}{2}\int_0^T (u_1^2 + u_2^2)\mathrm{d}t$ over the trajectories of

$$\frac{\mathrm{d}x}{\mathrm{d}t}(t) = u_1(t)X_1(x(t)) + u_2(t)X_2(x(t)), \quad \text{where} \quad (31)$$

$$X_1 = \frac{\partial}{\partial r}, \quad X_2 = \frac{1}{r}\left(\frac{\partial}{\partial \theta} - A(r)\frac{\partial}{\partial z}\right), \quad \text{and} \quad A(r) = \frac{1}{2}r^2 - \frac{1}{4}r^4, \quad (32)$$

or more explicitly over the solutions of the following system of equations:

$$\frac{\mathrm{d}r}{\mathrm{d}t} = u_1, \quad \frac{\mathrm{d}\theta}{\mathrm{d}t} = \frac{u_2}{r}, \quad \frac{\mathrm{d}z}{\mathrm{d}t} = -\frac{u_2}{r}A(r). \quad (33)$$

Then normal extremal curves are integral curves of the Hamiltonian vector field associated to $H = \frac{1}{2}(H_1^2 + H_2^2)$, with $H_1 = p_r, H_2 = \frac{1}{r}(p_\theta - A(r)p_z)$, where p_r, p_θ, p_z denote dual coordinates of co-vectors p defined by $p = p_r\mathrm{d}r + p_\theta\,\mathrm{d}\theta + p_z\mathrm{d}z$.

An easy calculation shows that $[X_1, X_2] = -\frac{\mathrm{d}A}{\mathrm{d}r}\frac{\partial}{\partial z}$ and $[X_1, [X_1, X_2]] = -\frac{\mathrm{d}^2A}{\mathrm{d}r^2}\frac{\partial}{\partial z}$. Hence, $X_1(x), X_2(x), [X_1, X_2](x)$ spans $\mathbb{R}^3$ except at $x = (r, \theta, z)$ where $\frac{\mathrm{d}A}{\mathrm{d}r} = 0$, that is, on the cylinder $r = 1$. Since $[X_1, [X_1, X_2]] \neq 0$ on $r = 1$, it follows that $X_1, X_2, [X_1, X_2], [X_1, [X_1, X_2]]$ span $\mathbb{R}^3$ at all points $x \in \mathbb{R}^3$. The helix $r = 1, z(\theta) = A(1)\theta, \theta \in \mathbb{R}$, generated by $u_1 = 0, u_2 = 1$, is a locally optimal trajectory (shown in [21]). It is the projection of an abnormal extremal curve and not the projection of a normal extremal curve.

Trajectories of a control system that are the projections of a constrained Hamiltonian system are called *singular* [3]. For instance, the helix in the above example is singular. The terminology derives from the singularity theory of mappings, and in the control theoretic context it is associated to the end point mapping $E\colon u_{[0,T]} \to x(x_0, u, T)$, where $x(x_0, u, t)$ denotes the trajectory of $\frac{\mathrm{d}x}{\mathrm{d}t} = F(x(t), u(t))$, $x(x_0, u, 0) = x_0$, with the controls $u(t)$ in the class of locally bounded measurable with values in $\mathbb{R}^m$. It is known that the singular trajectories are the projections of the integral curves $\xi(t)$ of the constrained Hamiltonian system, obtained by the Maximum Principle:

$$\frac{\mathrm{d}\xi}{\mathrm{d}t} = \vec{H}(\xi(t), u(t)), \quad \frac{\mathrm{d}H}{\mathrm{d}u}(\xi(t), u(t)) = 0, \quad (34)$$

where $H(\xi, u) = \xi(F(x, u))$, $\xi \in T_x^*M$. For an extensive theory of singular trajectories see [3].

Future Directions

Since the original publications there has been a considerable effort to obtain the maximum principle under more general conditions and under different technical assumptions. This effort seems to be motivated by two distinct objectives: the first motivation is a quest for a high order maximum principle [4,16,17], while the second motivation is an extension of the maximum principle to differential inclusions and non-smooth problems [9,20,28]. Although there is some indication that the corresponding theoretical approaches do not lead to common theory [5], there still remains an open question how to incorporate these diverse points of view into a universal maximum principle.

Bibliography

Primary Literature

1. Agrachev AA, Sachkov YL (2005) Control theory from the geometric viewpoint. Encycl Math Sci 87. Springer, Heidelberg
2. Arnold VI (1989) Mathematical methods of classical mechanics. Graduate texts in Mathematics, vol 60. Springer, Heidelberg
3. Bonnard B, Chyba M (2003) Singular trajectories and their role in control. Springer, Heidelberg
4. Bianchini RM, Stefani G (1993) Controllability along a trajectory; a variational approach. SIAM J Control Optim 31:900–927

5. Bressan A (2007) On the intersection of a Clarke cone with a Boltyanski cone (to appear)
6. Berkovitz LD (1974) Optimal control theory. Springer, New York
7. Caratheodory C (1935) Calculus of variations. Teubner, Berlin (reprinted 1982, Chelsea, New York)
8. Clarke FH (1983) Optimization and nonsmooth analysis. Wiley Interscience, New York
9. Clarke FH (2005) Necessary conditions in dynamic optimization. Mem Amer Math Soc 816(173)
10. Fomenko AT, Trofimov VV (1988) Integrable systems on Lie algebras and symmetric spaces. Gordon and Breach
11. Gamkrelidze RV (1978) Principles of optimal control theory. Plenum Press, New York
12. Jurdjevic V (1997) Geometric control theory. Cambridge Studies in Advanced Mathematics vol 51. Cambridge University Press, Cambridge
13. Jurdjevic V (2005) Hamiltonian systems on complex Lie groups and their homogeneous spaces. Mem Amer Math Soc 178(838)
14. Lee EB, Markus L (1967) Foundations of optimal control theory. Wiley, New York
15. Liu WS, Sussmann HJ (1995) Shortest paths for SR metrics of rank 2 distributions. Mem Amer Math Soc 118(564)
16. Knobloch H (1975) High order necessary conditions in optimal control. Springer, Berlin
17. Krener AJ (1977) The high order maximum principle and its application to singular extremals. SIAM J Control Optim 17:256–293
18. Kupka IK (1990) The ubiquity of Fuller's phenomenon. In: Sussmann HJ (ed) Non-linear controllability and Optimal control. Marcel Dekker, New York, pp 313–350
19. Marchal C (1973) Chattering arcs and chattering controls. J Optim Theory App 11:441–468
20. Morduchovich B (2006) Variational analysis and generalized differentiation: I. Basic analysis, II. Applications. Grundlehren Series (Fundamental Principle of Mathematical Sciences). Springer, Berlin
21. Montgomery R (1994) Abnormal minimizers. SIAM J Control Optim 32(6):1605–1620
22. Moser J (1980) Geometry of quadrics and spectral theory. In: The Chern Symposium 1979. Proceedings of the International symposium on Differential Geometry held in honor of S.S. Chern, Berkeley, California. Springer, pp 147–188
23. Perelomov AM (1990) Integrable systems of classical mechanics and Lie algebras. Birkhauser, Basel
24. Pontryagin LS, Boltyanski VG, Gamkrelidze RV, Mischenko EF (1962) The mathematical theory of optimal processes. Wiley, New York
25. Reiman AG, Semenov Tian-Shansky MA (1994) Group-theoretic methods in the theory of finite dimensional integrable systems. In: Arnold VI, Novikov SP (eds) Encyclopedia of Mathematical Sciences. Springer, Heidelberg
26. Sussmann HJ, Willems J (2002) The brachistochrone problem and modern control theory. In: Anzaldo-Meneses A, Bonnard B, Gauthier J-P, Monroy-Perez F (eds) Contemporary trends in non-linear geometric control theory and its applications. Proceedings of the conference on Geometric Control Theory and Applications held in Mexico City on September 4–6, 2000, to celebrate the 60th anniversary of Velimir Jurdjevic. World Scientific Publishers, Singapore, pp 113–165
27. Sussmann HJ () Set separation, approximating multicones and the Lipschitz maximum principle. J Diff Equations (to appear)
28. Vinter RB (2000) Optimal control. Birkhauser, Boston
29. Young LC (1969) Lectures in the calculus of variations and optimal control. Saunders, Philadelphia
30. Zelikin MI, Borisov VF (1994) Theory of chattering control with applications to aeronautics, Robotics, Economics, and Engineering. Birkhauser, Basel

Books and Reviews

Bressan A (1985) A high-order test for optimality of bang-bang controls. SIAM J Control Optim 23:38–48
Griffiths P (1983) Exterior differential systems and the calculus of variations. Birkhauser, Boston

Measure Preserving Systems

Karl Petersen
Department of Mathematics,
University of North Carolina, Chapel Hill, USA

Article Outline

Glossary
Definition of the Subject
Introduction: The Dynamical Viewpoint
Where do Measure-Preserving Systems Come from?
Construction of Measures
Invariant Measures on Topological Dynamical Systems
Finding Finite Invariant Measures Equivalent to a Quasi-Invariant Measure
Finding σ-finite Invariant Measures Equivalent to a Quasi-Invariant Measure
Some Mathematical Background
Future Directions
Bibliography

Glossary

Dynamical system A set acted upon by an algebraic object. Elements of the set represent all possible states or configurations, and the action represents all possible changes.

Ergodic A measure-preserving system is ergodic if it is essentially indecomposable, in the sense that given any invariant measurable set, either the set or its complement has measure 0.

Lebesgue space A measure space that is isomorphic with the usual Lebesgue measure space of a subinterval of the set of real numbers, possibly together with countably or finitely many point masses.

Measure An assignment of sizes to sets. A measure that

takes values only between 0 and 1 assigns probabilities to events.

Stochastic process A family of random variables (measurable functions). Such an object represents a family of measurements whose outcomes may be subject to chance.

Subshift, shift space A closed shift-invariant subset of the set of infinite sequences with entries from a finite alphabet.

Definition of the Subject

Measure-preserving systems model processes in equilibrium by transformations on probability spaces or, more generally, measure spaces. They are the basic objects of study in ergodic theory, a central part of dynamical systems theory. These systems arise from science and technology as well as from mathematics itself, so applications are found in a wide range of areas, such as statistical physics, information theory, celestial mechanics, number theory, population dynamics, economics, and biology.

Introduction: The Dynamical Viewpoint

Sometimes introducing a dynamical viewpoint into an apparently static situation can help to make progress on apparently difficult problems. For example, equations can be solved and functions optimized by reformulating a given situation as a fixed point problem, which is then addressed by iterating an appropriate mapping. Besides practical applications, this strategy also appears in theoretical settings, for example modern proofs of the Implicit Function Theorem. Moreover, the introduction of the ideas of change and motion leads to new concepts, new methods, and even new kinds of questions. One looks at actions and orbits and instead of always seeking exact solutions begins perhaps to ask questions of a qualitative or probabilistic nature: what is the general behavior of the system, what happens for most initial conditions, what properties of systems are typical within a given class of systems, and so on. Much of the credit for introducing this viewpoint should go to Henri Poincaré [29].

Two Examples

Consider two particular examples, one simple and the other not so simple. Decimal or base 2 expansions of numbers in the unit interval raise many natural questions about frequencies of digits and blocks. Instead of regarding the base 2 expansion $x = .x_0x_1\ldots$ of a fixed $x \in [0,1]$ as being given, we can regard it as arising from a dynamical process. Define $T\colon [0,1] \to [0,1]$ by $Tx = 2x \bmod 1$ (the fractional part of $2x$) and let $\mathcal{P} = \{P_0 = [0,1/2), P_1 = [1/2,1]\}$ be a partition of $[0,1]$ into two subintervals. We code the orbit of any point $x \in [0,1]$ by 0's and 1's by letting $x_k = i$ if $T^k x \in P_i$, $k = 0,1,2,\ldots$. Then reading the expansion of x amounts to applying to the coding the shift transformation and projection onto the first coordinate. This is equivalent to following the orbit of x under T and noting which element of the partition $\mathcal{P}$ is entered at each time. Reappearances of blocks amount to recurrence to cylinder sets as x is moved by T, frequencies of blocks correspond to ergodic averages, and Borel's theorem on normal numbers is seen as a special case of the Ergodic Theorem.

Another example concerns Szemerédi's Theorem [34], which states that every subset $A \subset \mathbb{N}$ of the natural numbers which has positive upper density contains arbitrarily long arithmetic progressions: given $L \in \mathbb{N}$ there are $s, m \in \mathbb{N}$ such that $s, s+m, \ldots, s+(L-1)m \in A$. Szemerédi's proof was ingenious, direct, and long. Furstenberg [15] saw how to obtain this result as a corollary of a strengthening of Poincaré's Recurrence Theorem in ergodic theory, which he then proved. Again we have an apparently static situation: a set $A \subset \mathbb{N}$ of positive density in which we seek arbitrarily long regularly spaced subsets. Furstenberg proposed to consider the characteristic function $\mathbf{1}_A$ of A as a point in the space $\{0,1\}^{\mathbb{N}}$ of 0's and 1's and to form the orbit closure X of this point under the shift transformation σ. Because A has positive density, it is possible to find a shift-invariant measure μ on X which gives positive measure to the cylinder set $B = [1] = \{x \in X\colon x_1 = 1\}$. Furstenberg's Multiple Recurrence Theorem says that given $L \in \mathbb{N}$ there is $m \in \mathbb{N}$ such that $\mu(B \cap T^{-m}B \cap \cdots \cap T^{-(L-1)m}B) > 0$. If y is a point in this intersection, then y contains a block of L 1's, each at distance m from the next. And since y is in the orbit closure of $\mathbf{1}_A$, this block also appears in the sequence $\mathbf{1}_A \in \{0,1\}^{\mathbb{N}}$, yielding the result.

Aspects of the dynamical argument remain in new combinatorial and harmonic-analytic proofs of the Szemerédi Theorem by T. Gowers [16,17] and T. Tao [35], as well as the extension to the (density zero) set of prime numbers by B. Green and T. Tao [18,36].

A Range of Actions

Here is a sample of dynamical systems of various kinds:

1. A semigroup or group G acts on a set X. There is given a map $G \times X \to X$, $(g,x) \to gx$, and it is assumed that

$$g_1(g_2x) = (g_1g_2)x \quad \text{for all } g_1, g_2 \in G\,, \quad x \in X \quad (1)$$

$$ex = x \quad \text{for all } x \in X\,, \qquad \text{if } G \text{ has an identity element } e\,. \tag{2}$$

2. A continuous linear operator T acts on a Banach or Hilbert space V.
3. $\mathcal{B}$ is a Boolean σ-algebra (a set together with a zero element 0 and operations $\vee, \wedge, '$ which satisfy the same rules as $\emptyset, \cup, \cap, ^c$ (complementation) do for σ-algebras of sets); $\mathcal{N}$ is a σ-ideal in $\mathcal{B}$ ($N \in \mathcal{N}, B \in \mathcal{B}, B \wedge N = B$ implies $B \in \mathcal{N}$; and $N_1, N_2, \cdots \in \mathcal{N}$ implies $\vee_{n=1}^{\infty} N_n \in \mathcal{N}$); and $S\colon \mathcal{B} \to \mathcal{B}$ preserves the Boolean σ-algebra operations and $S\mathcal{N} \subset \mathcal{N}$.
4. $\mathcal{B}$ is a Boolean σ-algebra, μ is a countably additive positive (nonzero except on the zero element of $\mathcal{B}$) function on $\mathcal{B}$, and $S\colon \mathcal{B} \to \mathcal{B}$ is as above. Then $(\mathcal{B}, \mu)$ is a *measure algebra* and S is a *measure algebra endomorphism*.
5. $(X, \mathcal{B}, \mu)$ is a measure space (X is a set, $\mathcal{B}$ is a σ-algebra of subsets of X, and $\mu\colon \mathcal{B} \to [0, \infty]$ is countably additive: If $B_1, B_2, \cdots \in \mathcal{B}$ are pairwise disjoint, then $\mu(\cup_{n=1}^{\infty} B_n) = \sum_{n=1}^{\infty} \mu(B_n)$); $T\colon X \to X$ is measurable ($T^{-1}\mathcal{B} \subset \mathcal{B}$) and nonsingular ($\mu(B) = 0$ implies $\mu(T^{-1}B) = 0$ – or, more stringently, μ and μT^{-1} are equivalent in the sense of absolute continuity).
6. $(X, \mathcal{B}, \mu)$ is a measure space, $T\colon X \to X$ is a one-to-one onto map such that T and T^{-1} are both measurable (so that $T^{-1}\mathcal{B} = \mathcal{B} = T\mathcal{B}$), and $\mu(T^{-1}B) = \mu(B)$ for all $B \in \mathcal{B}$. (In practice often T is not one-to-one, or onto, or even well-defined on all of X, but only after a set of measure zero is deleted.) This is the case of most interest for us, and then we call $(X, \mathcal{B}, \mu, T)$ a *measure-preserving system*. We also allow for the possibility that T is not invertible, or that some other group (such as $\mathbb{R}$ or $\mathbb{Z}^d$) or semigroup acts on X, but the case of $\mathbb{Z}$ actions will be the main focus of this article.
7. X is a compact metric space and $T\colon X \to X$ is a homeomorphism. Then (X, T) is a *topological dynamical system*.
8. M is a compact manifold (C^k for some $k \in [1, \infty]$) and $T\colon M \to M$ is a diffeomorphism (one-to-one and onto, with T and T^{-1} both C^k). Then (M, T) is a *smooth dynamical system*. Such examples can arise from solutions of an autonomous differential equation given by a vector field on M. Recall that in $\mathbb{R}^n$, an ordinary differential equation initial-value problem $x' = f(x), x(0) = x_0$ has a unique solution $x(t)$ as long as f satisfies appropriate smoothness conditions. The existence and uniqueness theorem for differential equations then produces a flow according to $T_t x_0 = x(t)$, satisfying $T_{s+t}x_0 = T_s(T_t x_0)$. Restricting to a compact invariant set (if there is one) and taking $T = T_1$ (the time 1 map) gives us a smooth system (M, f).

Naturally there are relations and inclusions among these examples of actions. Often problems can be clarified by forgetting about some of the structure that is present or by adding desirable structure (such as topology) if it is not. There remain open problems about representation and realization; for example, taking into account necessary restrictions, which measure-preserving systems can be realized as smooth systems preserving a smooth measure? Sometimes interesting aspects of the dynamics of a smooth system can be due to the presence of a highly nonsmooth subsystem, for example a compact lower-dimensional invariant set. Thus one should be ready to deal with many kinds of dynamical systems.

Where do Measure-Preserving Systems Come from?

Systems in Equilibrium

Besides physical systems, abstract dynamical systems can also represent aspects of biological, economic, or other real-world systems. Equilibrium does not mean stasis, but rather that the changes in the system are governed by laws which are not themselves changing. The presence of an invariant measure means that the probabilities of observable events do not change with time. (But of course what happens at time 2 can still depend on what happens at time 1, or, for that matter, at time 3.)

We consider first the example of the wide and important class of Hamiltonian systems. Many systems that model physical situations, for example a large number of ideal charged particles in a container, can be studied by means of Hamilton's equations. The state of the entire system at any time is supposed to be specified by a vector $(q, p) \in \mathbb{R}^{2n}$, the *phase space*, with q listing the coordinates of the positions of all of the particles, and p listing the coordinates of their momenta. We assume that there is a time-independent *Hamiltonian function* $H(q, p)$ such that the time development of the system satisfies *Hamilton's equations*:

$$\frac{\mathrm{d}q_i}{\mathrm{d}t} = \frac{\partial H}{\partial p_i}\,, \quad \frac{\mathrm{d}p_i}{\mathrm{d}t} = -\frac{\partial H}{\partial q_i}\,, \quad i = 1, \ldots, n\,. \tag{3}$$

Often the Hamiltonian function is the sum of kinetic and potential energy:

$$H(q, p) = K(p) + U(q)\,. \tag{4}$$

The potential energy $U(q)$ may depend on interactions among the particles or with an external field, while the kinetic energy $K(p)$ depends on the velocities and masses of the particles.

As discussed above, solving these equations with initial state (q, p) for the system produces a flow $(q, p) \to$

$T_t(q,p)$ in phase space. According to *Liouville's Theorem*, this flow preserves Lebesgue measure on $\mathbb{R}^{2n}$. Calculating $\mathrm{d}H/\mathrm{d}t$ by means of the Chain Rule and using Hamilton's equations shows that H is constant on orbits of the flow, and thus each set of constant energy $X(H_0) = \{(q,p)\colon H(q,p) = H_0\}$ is an invariant set. Thus one should consider the flow restricted to the appropriate invariant set. It turns out that there are also natural invariant measures on the sets $X(H_0)$, namely the ones given by rescaling the volume element $\mathrm{d}S$ on $X(H_0)$ by the factor $1/||\nabla H||$. For details, see [25].

Systems in equilibrium can also be hiding inside systems not in equilibrium, for example if there is an attractor supporting an SRB measure (for Sinai, Ruelle, and Bowen) (for definitions of the terms used here and more explanations, see the article in this collection by A. Wilkinson). Suppose that $T\colon M \to M$ is a diffeomorphism on a compact manifold as above, and that m is a version of Lebesgue measure on M, say given by a smooth volume form. We consider m to be a "physical measure", corresponding to laboratory measurements of observable quantities, whose values can be determined to lie in certain intervals in $\mathbb{R}$. Quite possibly m is not itself invariant under T, and an experimenter might observe strange or chaotic behavior whenever the state of the system gets close to some compact invariant set X. The dynamics of T restricted to X can in fact be quite complicated – maybe a full shift, which represents completely undeterministic behavior (for example if there is a horseshoe present), or a shift of finite type, or some other complicated topological dynamical system. Possibly $m(X) = 0$, so that X is effectively invisible to the observer except through its effects. It can happen that there is a T-invariant measure μ supported on X such that

$$\frac{1}{n}\sum_{k=0}^{n-1} mT^{-k} \to \mu \quad \text{weak}^* \,, \tag{5}$$

and then the long-term equilibrium dynamics of the system is described by (X, T, μ). For a recent survey on SRB measures, see [39].

Stationary Stochastic Processes

A *stationary process* is a family $\{f_t\colon t \in T\}$ of random variables (measurable functions) on a probability space $(\Omega, \mathcal{F}, P)$. Usually T is $\mathbb{Z}, \mathbb{N}$, or $\mathbb{R}$. For the remainder of this section let us fix $T = \mathbb{Z}$ (although the following definition could make sense for T any semigroup). We say that the process $\{f_n\colon n \in \mathbb{Z}\}$ is *stationary* if its finite-dimensional distributions are translation invariant, in the sense that for each $r = 1, 2, \ldots$, each $n_1, \ldots, n_r \in \mathbb{Z}$, each choice of Borel sets $B_1, \ldots, B_r \subset \mathbb{R}$, and each $s \in \mathbb{Z}$, we have

$$\begin{aligned} &P\{\omega\colon f_{n_1}(\omega) \in B_1, \ldots, f_{n_r}(\omega) \in B_r\} \\ &\quad = P\{\omega\colon f_{n_1+s}(\omega) \in B_1, \ldots, f_{n_r+s}(\omega) \in B_r\}\,. \end{aligned} \tag{6}$$

The f_n represent measurements made at times n of some random phenomenon, and the probability that a particular finite set of measurements yield values in certain ranges is supposed to be independent of time.

Each stationary process $\{f_n\colon n \in \mathbb{Z}\}$ on $(\Omega, \mathcal{F}, P)$ corresponds to a shift-invariant probability measure μ on the set $\mathbb{R}^{\mathbb{Z}}$ (with its Borel σ-algebra) and a single observable, namely the projection π_0 onto the 0'th coordinate, as follows. Define

$$\phi\colon \Omega \to \mathbb{R}^{\mathbb{Z}} \quad \text{by} \quad \phi(\omega) = (f_n(\omega))_{-\infty}^{\infty}\,, \tag{7}$$

and for each Borel set $E \subset \mathbb{R}^{\mathbb{Z}}$, define $\mu(E) = P(\phi^{-1}E)$. Then examining the values of μ on cylinder sets – for Borel $B_1, \ldots, B_r \subset \mathbb{R}$,

$$\begin{aligned} &\mu\{x \in \mathbb{R}^{\mathbb{Z}}\colon x_{n_i} \in B_i, i = 1, \ldots, r\} \\ &\quad = P\{\omega \in \Omega\colon f_{n_i}(\omega) \in B_i, i = 1, \ldots, r\} \end{aligned} \tag{8}$$

– and using stationarity of (f_n) shows that μ is invariant under σ. Moreover, the processes (f_n) on Ω and $\pi_0 \circ \sigma^n$ on $\mathbb{R}^{\mathbb{Z}}$ have the same finite-dimensional distributions, so they are equivalent for the purposes of probability theory.

Construction of Measures

We review briefly (following [33]) the construction of measures largely due to C. Carathéodory [8], with input from M. Fréchet [13], H. Hahn [19], A. N. Kolmogorov [26], and others, then discuss the application to construction of measures on shift spaces and of stochastic processes in general.

The Carathéodory Construction

A *semialgebra* is a family S of subsets of a set X which is closed under finite intersections and such that the complement of any member of S is a finite disjoint union of members of S. Key examples are

1. the family $\mathcal{H}$ of half-open subintervals $[a, b)$ of $[0, 1)$;
2. in the space $X = A^{\mathbb{Z}}$ of doubly infinite sequences on a finite alphabet A, the family C of *cylinder sets* (determined by fixing finitely many entries)

$$\{x \in A^{\mathbb{Z}}\colon x_{n_1} = a_1, \ldots, x_{n_r} = a_r\}\,; \tag{9}$$

3. the family C_1 of *anchored cylinder sets*

$$\{x \in A^{\mathbb{N}} : x_1 = a_1, \ldots, x_r = a_r\} \tag{10}$$

in the space $X = A^{\mathbb{N}}$ of one-sided infinite sequences on a finite alphabet A.

An *algebra* is a family of subsets of a set X which is closed under finite unions, finite intersections, and complements. A *σ-algebra* is a family of subsets of a set X which is closed under countable unions, countable intersections, and complements. If S is a semialgebra of subsets of X, the *algebra $\mathcal{A}(S)$ generated* by S is the smallest algebra of subsets of X which contains S. $\mathcal{A}(S)$ is the intersection of all the subalgebras of the set 2^X of all subsets of X and consists exactly of all finite disjoint unions of elements of S. Given an algebra $\mathcal{A}$, the *σ-algebra $\mathcal{B}(\mathcal{A})$ generated by $\mathcal{A}$* is the smallest σ-algebra of subsets of X which contains $\mathcal{A}$.

A *nonnegative set function* on S is a function $\mu\colon S \to [0, \infty]$ such that $\mu(\emptyset) = 0$ if $\emptyset \in S$. We say that such a μ is

- *finitely additive* if whenever $S_1, \ldots, S_n \in S$ are pairwise disjoint and $S = \cup_{i=1}^n S_i \in S$, we have $\mu(S) = \sum_{i=1}^n \mu(S_i)$;
- *countably additive* if whenever $S_1, S_2 \cdots \in S$ are pairwise disjoint and $S = \cup_{i=1}^\infty S_i \in S$, we have $\mu(S) = \sum_{i=1}^\infty \mu(S_i)$; and
- *countably subadditive* if whenever $S_1, S_2 \cdots \in S$ and $S = \cup_{i=1}^\infty S_i \in S$, we have $\mu(S) \le \sum_{i=1}^\infty \mu(S_i)$.

A *measure* is a countably additive nonnegative set function defined on a σ-algebra.

Proposition 1 *Let S be a semialgebra and μ a nonnegative set function on S. In order that μ have an extension to a finitely additive set function on the algebra $\mathcal{A}(S)$ generated by S, it is necessary and sufficient that μ be finitely additive on S.*

Proof 1 The stated condition is obviously necessary. Conversely, given μ which is finitely additive on S, it is natural to define

$$\mu(\bigcup_{i=1}^n S_i) = \sum_{i=1}^n \mu(S_i) \tag{11}$$

whenever $A = \cup_{i=1}^n S_i$ (with the S_i pairwise disjoint) is in the algebra $\mathcal{A}(S)$ generated by S. It is necessary to verify that μ is then well defined on $\mathcal{A}(S)$, since each element of $\mathcal{A}(S)$ may have more than one representation as a finite disjoint union of members of S. But, given two such representations of a single set A, forming the common refinement and applying finite additivity on S shows that μ so defined assigns the same value to A both times. Then finite additivity on $\mathcal{A}(S)$ of the extended μ is clear. □

Proposition 2 *Let S be a semialgebra and μ a nonnegative set function on S. In order that μ have an extension to a countably additive set function on the algebra $\mathcal{A}(S)$ generated by S, it is necessary and sufficient that μ be (i) finitely additive and (ii) countably subadditive on S.*

Proof 2 Conditions (i) and (ii) are clearly necessary. If μ is finitely additive on S, then by Proposition 1 μ has an extension to a finitely additive nonnegative set function, which we will still denote by μ, on $\mathcal{A}(S)$.

Let us see that this extension μ is countably subadditive on $\mathcal{A}(S)$. Suppose that $A_1, A_2, \cdots \in \mathcal{A}(S)$ are pairwise disjoint and their union $A \in \mathcal{A}(S)$. Then A is a finite disjoint union of sets in S, as is each A_i:

$$\begin{aligned} A &= \bigcup_{i=1}^\infty A_i\ ,\ \text{each } A_i = \bigcup_{k=1}^{n_i} S_{ik}\ , \\ A &= \bigcup_{j=1}^m R_j\ ,\ \text{each } A_i \in \mathcal{A}(S)\ ,\ \text{each } S_{ik}, R_j \in S\ . \end{aligned} \tag{12}$$

Since each $R_j \in S$, by countable subadditivity of μ on S, and using $R_j = R_j \cap A$,

$$\mu(R_j) = \mu(\bigcup_{i=1}^\infty \bigcup_{k=1}^{n_i} S_{ik} \cap R_j) \le \sum_{i=1}^\infty \sum_{k=1}^{n_i} \mu(S_{ik} \cap R_j), \tag{13}$$

and hence, by finite additivity of μ on $\mathcal{A}(S)$,

$$\begin{aligned} \mu(A) &= \sum_{j=1}^m \mu(R_j) \le \sum_{i=1}^\infty \sum_{k=1}^{n_i} \sum_{j=1}^m \mu(S_{ik} \cap R_j) \\ &= \sum_{i=1}^\infty \sum_{k=1}^{n_i} \mu(S_{ik}) = \sum_{i=1}^\infty \mu(A_i)\ . \end{aligned} \tag{14}$$

Now finite additivity of μ on an algebra $\mathcal{A}$ implies that μ is *monotonic* on the algebra: if $A, B \in \mathcal{A}$ and $A \subset B$, then $\mu(A) \le \mu(B)$. Thus if $A_1, A_2, \cdots \in \mathcal{A}(S)$ are pairwise disjoint and their union $A \in \mathcal{A}(S)$, then for each n we have $\sum_{i=1}^n \mu(A_i) = \mu(\cup_{i=1}^n A_i) \le \mu(A)$, and hence $\sum_{i=1}^\infty \mu(A_i) \le \mu(A)$. □

Theorem 1 *In order that a nonnegative set function μ on an algebra $\mathcal{A}$ of subsets of a set X have an extension to a (countably additive) measure on the σ-algebra $\mathcal{B}(\mathcal{A})$ generated by $\mathcal{A}$, it is necessary and sufficient that μ be countably additive on $\mathcal{A}$.*

Here is a sketch of how the extension can be constructed. Given a countably additive nonnegative set function μ on

an algebra $\mathcal{A}$ of subsets of a set X, one defines the *outer measure* μ^* that it determines on the family 2^X of all subsets of X by

$$\mu^*(E) = \inf\{\sum_{i=1}^{\infty} \mu(A_i) \colon A_i \in \mathcal{A}, E \subset \cup_{i=1}^{\infty} A_i\} . \quad (15)$$

Then μ^* is a nonnegative, countably subadditive, monotonic set function on 2^X.

Define a set E to be μ^**-measurable* if for all $T \subset X$,

$$\mu^*(T) = \mu^*(T \cap E) + \mu^*(T \cap E^c) . \quad (16)$$

This ingenious definition can be partly motivated by noting that if μ^* is to be finitely additive on the family $\mathcal{M}$ of μ^*-measurable sets, which should contain X, then at least this condition must hold when $T = X$. It is amazing that then this definition readily, with just a little set theory and a few ε's, yields the following theorem.

Theorem 2 *Let μ be a countably additive nonnegative set function on an algebra $\mathcal{A}$ of subsets of a set X, and let μ^* be the outer measure that it determines on the family 2^X of all subsets of X as above. Then the family $\mathcal{M}$ of μ^*-measurable subsets of X is a σ-algebra containing $\mathcal{A}$ (and hence $\mathcal{B}(\mathcal{A})$) and all subsets of X which have μ^* measure 0. The restriction $\mu^*|\mathcal{M}$ is a (countably additive) measure which agrees on $\mathcal{A}$ with μ. If μ is σ-finite on $\mathcal{A}$ (so that there are $X_1, X_2, \dots \in \mathcal{A}$ with $\mu(X_i) < \infty$ for all i and $X = \cup_{i=1}^{\infty} X_i$), then μ on $\mathcal{B}(\mathcal{A})$ is the only extension of μ on $\mathcal{A}$ to $\mathcal{B}(\mathcal{A})$.*

In this way, beginning with the semialgebra $\mathcal{H}$ of half-open subintervals of $[0, 1)$ and $\mu[a, b) = b - a$, one arrives at Lebesgue measure on the σ-algebra $\mathcal{M}$ of Lebesgue measurable sets and on its sub-σ-algebra $\mathcal{B}(\mathcal{H})$ of Borel sets.

Measures on Shift Spaces

The measures that determine stochastic processes are also frequently constructed by specifying data on a semialgebra of cylinder sets. Given a finite alphabet A, denote by $\Omega(A) = A^{\mathbb{Z}}$ and $\Omega^+(A) = A^{\mathbb{N}}$ the sets of two and one-sided sequences, respectively, with entries from A. These are compact metric spaces, with $d(x, y) = 2^{-n}$ when $n = \inf\{|k| \colon x_k \neq y_k\}$. In both cases, the *shift transformation* σ defined by $(\sigma x)_n = x_{n+1}$ for all n is a homeomorphism.

Suppose (cf. [3]) that for every $k = 1, 2, \dots$ we are given a function $g_k \colon A^k \to [0, 1]$, and that these functions satisfy, for all k,

1. $g_k(B) \geq 0$ for all $B \in A^k$;
2. $\sum_{i \in A} g_{k+1}(Bi) = g_k(B)$ for all $B \in A^k$;
3. $\sum_{i \in A} g_1(i) = 1$.

Then Theorems 1 and 2 imply that there is a unique measure μ on the Borel subsets of $\Omega^+(A)$ such that for all $k = 1, 2, \dots$ and $B \in A^k$

$$\mu\{x \in \Omega^+(A) \colon x_1 \dots x_k = B\} = g_k(B) . \quad (17)$$

If in addition the g_k also satisfy

4. $\sum_{i \in A} g_{k+1}(iB) = g_k(B)$ for all $k = 1, 2, \dots$ and all $B \in A^k$, then there is a unique shift-invariant measure μ on the Borel subsets of $\Omega^+(A)$ (also $\Omega(A)$) such that for all n, all $k = 1, 2, \dots$ and $B \in A^k$

$$\mu\{x \in \Omega^+(A) \colon x_n \dots x_{n+k-1} = B\} = g_k(B) . \quad (18)$$

This follows from the Carathéodory theorem by beginning with the semialgebra C_1 of anchored cylinder sets or the semialgebra C of cylinder sets determined by finitely many consecutive coordinates, respectively.

There are two particularly important examples of this construction. First, let our finite alphabet be $A = \{0, \dots, d-1\}$, and let $p = (p_0, \dots, p_{d-1})$ be a probability vector: all $p_i \geq 0$ and $\sum_{i=0}^{d-1} p_i = 1$. For any block $B = b_1 \dots b_k \in A^k$, define

$$g_k(B) = p_{b_1} \dots p_{b_k} . \quad (19)$$

The resulting measure μ_p is the product measure on $\Omega(A) = A^{\mathbb{Z}}$ of infinitely many copies of the probability measure determined by p on the finite sample space A. The measure-preserving system $(\Omega, \mathcal{B}, \mu, \sigma)$ (with $\mathcal{B}$ the σ-algebra of Borel subsets of $\Omega(A)$, or its completion), is denoted by $\mathcal{B}(p)$ and is called the *Bernoulli system* determined by p. This system models an infinite number of independent repetitions of an experiment with finitely many outcomes, the ith of which has probability p_i on each trial.

This construction can be generalized to model stochastic processes which have some memory. Again let $A = \{0, \dots, d-1\}$, and let $p = (p_0, \dots, p_{d-1})$ be a probability vector. Let P be a $d \times d$ *stochastic matrix* with rows and columns indexed by A. This means that all entries of P are nonnegative, and the sum of the entries in each row is 1. We regard P as giving the transition probabilities between pairs of elements of A. Now we define for any block $B = b_1 \dots b_k \in A^k$

$$g_k(B) = p_{b_1} P_{b_1 b_2} P_{b_2 b_3} \dots P_{b_{k-1} b_k} . \quad (20)$$

Using the g_k to define a nonnegative set function $\mu_{p,P}$ on the semialgebra C_1 of anchored cylinder subsets of $\Omega^+(A)$, one can verify that $\mu_{p,P}$ is (vacuously) finitely additive

and countably subadditive on C_1 and therefore extends to a measure on the Borel σ-algebra of $\Omega^+(A)$, and its completion. The resulting stochastic process is a (one-step, finite-state) *Markov process*. If p and P also satisfy

$$pP = p\,, \tag{21}$$

then condition 4. above is satisfied, and the Markov process is stationary. In this case we call the (one or two-sided) measure-preserving system the *Markov shift* determined by p and P. Points in the space are conveniently pictured as infinite paths in a directed graph with vertices A and edges corresponding to the nonzero entries of P. A process with a longer memory, say of length m, can be produced by repeating the foregoing construction after recoding with a *sliding block code* to the new alphabet A^m: for each $\omega \in \Omega(A)$, let $(\phi(\omega))_n = \omega_n\omega_{n+1}\ldots\omega_{n+m-1} \in A^m$.

The Kolmogorov Consistency Theorem

There is a generalization of this method to the construction of stochastic processes indexed by any set T. (Most frequently $T = \mathbb{Z}, \mathbb{N}, \mathbb{R}, \mathbb{Z}^d$, or $\mathbb{R}^d$). We give a brief description, following [4].

Let T be an arbitrary index set. We aim to produce a $\mathbb{R}$-valued stochastic process indexed by T, that is to say, a Borel probability measure P on $\Omega = \mathbb{R}^T$, which has prespecified finite-dimensional distributions. Suppose that for every ordered k-tuple $t_1, \ldots, t_k$ of *distinct* elements of T we are given a Borel probability measure $\mu_{t_1\ldots t_k}$ on $\mathbb{R}^k$. Denoting $f \in \mathbb{R}^T$ also by $(f_t\colon t \in T)$, we want it to be the case that, for each k, each choice of distinct $t_1, \ldots t_k \in T$, and each Borel set $B \subset \mathbb{R}^k$,

$$P\{(f_t\colon t \in T)\colon (f_{t_1}, \ldots, f_{t_k}) \in B\} = \mu_{t_1\ldots t_k}(B)\,. \tag{22}$$

For consistency, we will need, for example, that

$$\mu_{t_1t_2}(B_1 \times B_2) = \mu_{t_2t_1}(B_2 \times B_1)\,, \quad \text{since} \tag{23}$$

$$P\{(f_{t_1}, f_{t_2}) \in A_1 \times A_2\} = P\{(f_{t_2}, f_{t_1}) \in A_2 \times A_1\}\,. \tag{24}$$

Thus we assume:

1. For any $k = 1, 2, \ldots$ and permutation π of $1, \ldots, k$, if $\phi_\pi : \mathbb{R}^k \to \mathbb{R}^k$ is defined by

$$\phi_\pi(x_{\pi 1}, \ldots, x_{\pi k}) = (x_1, \ldots, x_k)\,, \tag{25}$$

then for all k and all Borel $B \subset \mathbb{R}^k$

$$\mu_{t_1\ldots t_k}(B) = \mu_{t_{\pi 1}\ldots t_{\pi k}}(\phi_\pi^{-1}B)\,. \tag{26}$$

Further, since leaving the value of one of the f_{t_j} free does not change the probability in (22), we also should have

2. For any $k = 1, 2, \ldots$, distinct $t_1, \ldots, t_k, t_{k+1} \in T$, and Borel set $B \subset \mathbb{R}^k$,

$$\mu_{t_1\ldots t_k}(B) = \mu_{t_1\ldots t_k t_{k+1}}(B \times \mathbb{R})\,. \tag{27}$$

Theorem 3 (Kolmogorov Consistency Theorem [26]) *Given a system of probability measures $\mu_{t_1\ldots t_k}$ as above indexed by finite ordered subsets of a set T, in order that there exist a probability measure P on $\mathbb{R}^T$ satisfying (22) it is necessary and sufficient that the system satisfy 1. and 2. above.*

When $T = \mathbb{Z}, \mathbb{R}$, or $\mathbb{N}$, as in the example with the g_k above, the problem of consistency with regard to permutations of indices does not arise, since we tacitly use the order in T in specifying the finite-dimensional distributions.

In case T is a semigroup, by adding conditions on the given data $\mu_{t_1\ldots t_k}$ it is possible to extend this construction also to produce *stationary* processes indexed by T, in parallel with the above constructions for $T = \mathbb{Z}$ or $\mathbb{N}$.

Invariant Measures on Topological Dynamical Systems

Existence of Invariant Measures

Let X be a compact metric space and $T\colon X \to X$ a homeomorphism (although usually it is enough just that T be a continuous map). Denote by $C(X)$ the Banach space of continuous real-valued functions on X with the supremum norm and by $\mathcal{M}(X)$ the set of Borel probability measures on X. Given the weak* topology, according to which

$$\mu_n \to \mu \text{ if and only if } \int_X f_n\,\mathrm{d}\mu \to \int_X f\,\mathrm{d}\mu \quad \text{for all } f \in C(X)\,, \tag{28}$$

$\mathcal{M}(X)$ is a convex subset of the dual space $C(X)^*$ of all continuous linear functionals from $C(X)$ to $\mathbb{R}$. With the weak* topology it is metrizable and (by Alaoglu's Theorem) compact.

Denote by $\mathcal{M}_T(X)$ the set of T-invariant Borel probability measures on X. A Borel probability measure μ on X is in $\mathcal{M}(X)$ if and only if

$$\mu(T^{-1}B) = \mu(B) \;\text{ for all Borel sets } B \subset X\,, \tag{29}$$

equivalently,

$$\mu(f\,T) = \int_X f \circ T\,\mathrm{d}\mu = \int_X f\,\mathrm{d}\mu \;\text{ for all } f \in C(X)\,. \tag{30}$$

Proposition 3 *For every compact topological dynamical system (X, T) (with X not empty) there is always at least one T-invariant Borel probability measure on X.*

Proof 3 Let m be any Borel probability measure on X. For example, we could pick a point $x_0 \in X$ and let m be the point mass δ_{x_0} at x_0 defined by

$$\delta_{x_0}(f) = f(x_0) \quad \text{for all } f \in C(X) \,. \tag{31}$$

Form the averages

$$A_n m = \tfrac{1}{n} \sum_{i=0}^{n-1} m T^{-i} \,, \tag{32}$$

which are also in $\mathcal{M}(X)$. By compactness, $\{A_n m\}$ has a weak* cluster point μ, so that there is a subsequence

$$A_{n_k} m \to \mu \ \text{ weak}^* \,. \tag{33}$$

Then $\mu \in \mathcal{M}(X)$; and μ is T-invariant, because for each $f \in C(X)$

$$|\mu(fT) - \mu(f)| = \lim_{k\to\infty} \tfrac{1}{n_k} |\mu(fT^{n_k}) - \mu(f)| = 0 \,, \tag{34}$$

both terms inside the absolute value signs being bounded. □

Ergodicity and Unique Ergodicity

Among the T-invariant measures on X are the *ergodic* ones, those for which $(X, \mathcal{B}, \mu, T)$ (with $\mathcal{B}$ the σ-algebra of Borel subsets of X) forms an ergodic measure-preserving system. This means that there are no proper T-invariant measurable sets:

$$B \in \mathcal{B} \,, \mu(T^{-1}B \triangle B) = 0 \quad \text{implies } \mu(B) = 0 \text{ or } 1 \,. \tag{35}$$

Equivalently (using the Ergodic Theorem), $(X, \mathcal{B}, \mu, T)$ is ergodic if and only if for each $f \in L^1(X, \mathcal{B}, \mu)$

$$\tfrac{1}{n} \sum_{k=1}^{n-1} f(T^k x) \to \int_X f \, d\mu \quad \text{almost everywhere} \,. \tag{36}$$

It can be shown that the ergodic measures on (X, T) are exactly the *extreme points* of the compact convex set $\mathcal{M}_T(X)$, namely those $\mu \in \mathcal{M}_T(X)$ for which there do not exist $\mu_1, \mu_2 \in \mathcal{M}_T(x)$ with $\mu_1 \neq \mu_2$ and $s \in (0, 1)$ such that

$$\mu = s\mu_1 + (1 - s)\mu_2 \,. \tag{37}$$

The Krein-Milman Theorem states that in a locally convex topological vector space such as $C(X)^*$ every compact convex set is the closed convex hull of its extreme points. Thus every nonempty such set has extreme points, and so there always exist ergodic measures for (X, T). A topological dynamical system (X, T) is called *uniquely ergodic* if there is only *one* T-invariant Borel probability measure on X, in which case, by the foregoing discussion, that measure must be ergodic.

There are many examples of topological dynamical systems which are uniquely ergodic and of others which are not. For now, we just remark that translation by a generator on a compact monothetic group is always uniquely ergodic, while group endomorphisms and automorphisms tend to be not uniquely ergodic. Bernoulli and (nonatomic) Markov shifts are not uniquely ergodic, because they have many periodic orbits, each of which supports an ergodic measure.

Finding Finite Invariant Measures Equivalent to a Quasi-Invariant Measure

Let $(X, \mathcal{B}, m)$ be a σ-finite measure space, and suppose that $T\colon X \to X$ is an invertible *nonsingular* transformation. Thus we assume that T is one-to-one and onto (maybe after a set of measure 0 has been deleted), that T and T^{-1} are both measurable, so that

$$T\mathcal{B} = \mathcal{B} = T^{-1}\mathcal{B} \,, \tag{38}$$

and that T and T^{-1} preserve the σ-ideal of sets of measure 0:

$$\begin{aligned} m(B) = 0 \ &\text{ if and only if } m(T^{-1}B) = 0 \\ &\text{ if and only if } m(TB) = 0 \,. \end{aligned} \tag{39}$$

In this situation we say that m is *quasi-invariant* for T.

A nonsingular system $(X, \mathcal{B}, m, T)$ as above may model a nonequilibrium situation in which events that are impossible (measure 0) at any time are also impossible at any other time. When dealing with such a system, it can be useful to know whether there is a T-invariant measure μ that is equivalent to m (in the sense of absolute continuity – they have the same sets of measure 0–in which case we write $\mu \sim m$), for then one would have available machinery of the measure-preserving situation, such as the Ergodic Theorem and entropy in their simplest forms. Also, it is most useful if the measures are σ-finite, so that tools such as the Radon-Nikodym and Tonelli-Fubini theorems will be available.

We may assume that $m(X) = 1$. For if $X = \cup_{i=1}^{\infty} X_i$ with each $X_i \in \mathcal{B}$ and $m(X_i) < \infty$, disjointifying (replace X_i by $X_i \setminus X_{i-1}$ for $i \geq 2$) and deleting any X_i that have measure 0, we may replace m by

$$\sum_{i=1}^{\infty} \frac{m|_{X_i}}{2^i m(X_i)} \,. \tag{40}$$

Definition 1 Let $(X, \mathcal{B}, m)$ be a probability space and $T: X \to X$ a nonsingular transformation. We say that $A, B \in \mathcal{B}$ are *T-equivalent*, and write $A \sim_T B$, if there are two sequences of pairwise disjoint sets, $A_1, A_2, \ldots$ and $B_1, B_2, \ldots$ and integers $n_1, n_2, \ldots$ such that

$$A = \bigcup_{i=1}^{\infty} A_i\,, \quad B = \bigcup_{i=1}^{\infty} B_i\,, \quad \text{and } T^{n_i} A_i = B_i \text{ for all } i\,. \tag{41}$$

Definition 2 Let $(X, \mathcal{B}, m, T)$ be as above. A measurable set $A \subset X$ is called *T-nonshrinkable* if A is not T-equivalent to any proper subset: whenever $B \subset A$ and $B \sim_T A$ we have $m(A \setminus B) = 0$.

Theorem 4 (Hopf [23]) *Let $(X, \mathcal{B}, m)$ be a probability space and $T: X \to X$ a nonsingular transformation. There exists a finite invariant measure $\mu \sim m$ if and only if X is T-nonshrinkable.*

Proof 4 We present just the easy half. If $\mu \sim m$ is T-invariant and $X \sim_T B$, with corresponding decompositions $X = \cup_{i=1}^{\infty} X_i$, $B = \cup_{i=1}^{\infty} B_i$, then

$$\begin{aligned} \mu(B) &= \sum_{i=1}^{\infty} \mu(B_i) = \sum_{i=1}^{\infty} \mu(T^{n_i} X_i) \\ &= \sum_{i=1}^{\infty} \mu(X_i) = \mu(X)\,, \end{aligned} \tag{42}$$

so that $\mu(X \setminus B) = 0$ and hence $m(X \setminus B) = 0$.

For the converse, one tries to show that if X is T-nonshrinkable, then for each $A \in \mathcal{B}$ the following limit exists:

$$\lim_{n\to\infty} \tfrac{1}{n} \sum_{k=0}^{n-1} m(T^k A)\,. \tag{43}$$

The condition of T-nonshrinkability not being easy to check, subsequent authors gave various necessary and sufficient conditions for the existence of a finite equivalent invariant measure:

1. Dowker [11]. *Whenever $A \in \mathcal{B}$ and $m(A) > 0$,* $\liminf_{n\to\infty} m(T^n A) > 0$.
2. Calderón [6]. *Whenever $A \in \mathcal{B}$ and $m(A) > 0$,* $\lim inf_{n\to\infty} \frac{1}{n} \sum_{k=0}^{\infty} m(T^k A) > 0$.
3. Dowker [12]. *Whenever $A \in \mathcal{B}$ and $m(A) > 0$,* $\limsup_{n\to\infty} \frac{1}{n} \sum_{k=0}^{\infty} m(T^k A) > 0$.

Hajian and Kakutani [20] showed that the condition

$$m(A) > 0 \;\text{ implies }\; \limsup_{n\to\infty} m(T^n A) > 0 \tag{44}$$

is *not* sufficient for existence of a finite equivalent invariant measure. They also gave another necessary and sufficient condition.

Definition 3 A measurable set $W \subset X$ is called *wandering* if the sets $T^i W$, $i \in \mathbb{Z}$, are pairwise disjoint. W is called *weakly wandering* if there are infinitely many integers n_i such that $T^{n_i} W$ and $T^{n_j} W$ are disjoint whenever $n_i \neq n_j$.

Theorem 5 (Hajian-Kakutani [20]) *Let $(X, \mathcal{B}, m)$ be a probability space and $T: X \to X$ a nonsingular transformation. There exists a finite invariant measure $\mu \sim m$ if and only if there are no weakly wandering sets of positive measure.*

Finding σ-finite Invariant Measures Equivalent to a Quasi-Invariant Measure

First Necessary and Sufficient Conditions

While being able to replace a quasi-invariant measure by an equivalent finite invariant measure would be great, it may be impossible, and then finding a σ-finite equivalent measure would still be pretty good. Hopf's nonshrinkability condition was extended to the σ-finite case by Halmos:

Theorem 6 (Halmos [21]) *Let $(X, \mathcal{B}, m)$ be a probability space and $T: X \to X$ a nonsingular transformation. There exists a σ-finite invariant measure $\mu \sim m$ if and only if X is a countable union of T-nonshrinkable sets.*

Another necessary and sufficient condition is given easily in terms of solvability of a *cohomological functional equation* involving the Radon-Nikodym derivative w of mT with respect to m, defined by

$$m(TB) = \int_B w \,\mathrm{d}m \quad \text{for all } B \in \mathcal{B}\,. \tag{45}$$

Proposition 4 ([21]) *Let $(X, \mathcal{B}, m)$ be a probability space and $T: X \to X$ a nonsingular transformation. There exists a σ-finite invariant measure $\mu \sim m$ if and only if there is a measurable function $f: X \to (0, \infty)$ such that*

$$f(Tx) = w(x) f(x) \quad a.e. \tag{46}$$

Proof 5 If $\mu \sim m$ is σ-finite and T-invariant, let $f = \mathrm{d}m/\mathrm{d}\mu$ be the Radon-Nikodym derivative of m with respect to μ, so that

$$m(B) = \int_B f \,\mathrm{d}\mu \quad \text{for all } B \in \mathcal{B}\,. \tag{47}$$

Then for all $B \in \mathcal{B}$, since $\mu T = \mu$,

$$\begin{aligned} m(TB) &= \int_{TB} f \,\mathrm{d}\mu = \int_B f\, T \,\mathrm{d}\mu\,, \quad \text{while also} \\ m(TB) &= \int_B w \,\mathrm{d}m = \int_B w f \,\mathrm{d}m\,, \end{aligned} \tag{48}$$

so that $fT = wf$ a.e.

Conversely, given such an f, let

$$\mu(B) = \int_B \tfrac{1}{f}\,\mathrm{d}m \quad \text{for all } B \in \mathcal{B}\,. \tag{49}$$

Then for all $B \in \mathcal{B}$

$$\begin{aligned}\mu(TB) &= \int_{TB} \tfrac{1}{f}\,\mathrm{d}m = \int_B \tfrac{1}{fT}\,\mathrm{d}mT \\ &= \int_B \tfrac{1}{fT} w\,\mathrm{d}m = \int_B \tfrac{1}{f}\,\mathrm{d}m = \mu(B)\,.\end{aligned} \tag{50}$$

□

Conservativity and Recurrence

Definition 4 A nonsingular system $(X, \mathcal{B}, m, T)$ (with $m(X) = 1$) is called *conservative* if there are no wandering sets of positive measure. It is called *completely dissipative* if there is a wandering set W such that

$$m\left(\bigcup_{i=-\infty}^{\infty} T^i W\right) = m(X)\,. \tag{51}$$

Note that if $(X, \mathcal{B}, m, T)$ is completely dissipative, it is easy to construct a σ-finite equivalent invariant measure. With W as above, define $\mu = m$ on W and push μ along the orbit of W, letting $\mu = mT^{-n}$ on each $T^n W$. We want to claim that this allows us to restrict attention to the conservative case, which follows once we know that the system splits into a conservative and a completely dissipative part.

Theorem 7 (Hopf Decomposition [24]) *Given a nonsingular map T on a probability space $(X, \mathcal{B}, m)$, there are disjoint measurable sets C and D such that*

1. *$X = C \cup D$;*
2. *C and D are invariant: $TC = C = T^{-1}C$, $TD = D = T^{-1}D$;*
3. *$T|_C$ is conservative;*
4. *If $D \neq \emptyset$, then $T|_D$ is completely dissipative.*

Proof 6 Assume that the family $\mathcal{W}$ of wandering sets with positive measure is nonempty, since otherwise we can take $C = X$ and $D = \emptyset$. Partially order $\mathcal{W}$ by

$$W_1 \le W_2 \quad \text{if} \quad m(W_1 \setminus W_2) = 0\,. \tag{52}$$

We want to apply Zorn's Lemma to find a maximal element in $\mathcal{W}$. Let $\{W_\lambda : \lambda \in \Lambda\}$ be a chain (linearly ordered subset) in $\mathcal{W}$. Just forming $\cup_{\lambda\in\Lambda} W_\lambda$ may result in a nonmeasurable set, so we have to use the measure to form a measure-theoretic essential supremum of the chain. So let

$$s = \sup\{m(W_\lambda) : \lambda \in \Lambda\}\,, \tag{53}$$

so that $s \in (0, 1]$. If there is a λ such that $m(W_\lambda) = s$, let W be that W_λ. Otherwise, for each k choose $\lambda_k \in \Lambda$ so that

$$s_k = m(W_{\lambda_k}) \uparrow s\,, \tag{54}$$

and let

$$W = \bigcup_{k=1}^{\infty} W_{\lambda_k}\,. \tag{55}$$

We claim that in either case W is an upper bound for the chain $\{W_\lambda : \lambda \in \Lambda\}$. In both cases we have $m(W) = s$.

Note that if $\lambda, \tau \in \Lambda$ are such that $m(W_\lambda) \le m(W_\tau)$, then $W_\lambda \le W_\tau$. For if $W_\tau \le W_\lambda$, then $m(W_\tau \setminus W_\lambda) = 0$, and thus

$$\begin{aligned} m(W_\tau) &= m(W_\tau \cap W_\lambda) + m(W_\tau \setminus W_\lambda) = m(W_\tau \cap W_\lambda) \\ &\le m(W_\tau \cap W_\lambda) + m(W_\lambda \setminus W_\tau) \\ &= m(W_\lambda) \le m(W_\tau)\,, \end{aligned} \tag{56}$$

so that $m(W_\lambda \setminus W_\tau) = 0$, $W_\lambda \le W_\tau$, and hence $W_\lambda = W_\tau$.

Thus in the first case $W \in \mathcal{W}$ is an upper bound for the chain. In the second case, by discarding the measure 0 set

$$Z = \bigcup_{k=1}^{\infty} (W_{\lambda_k} \setminus W_{\lambda_{k+1}})\,, \tag{57}$$

we may assume that W is the *increasing* union of the W_{λ_k}. Then $W \ge W_{\lambda_k}$ for all k, and W is wandering: if some $T^n W \cap W \neq \emptyset$, then there must be a k such that $T^n W_{\lambda_k} \cap W_{\lambda_k} \neq \emptyset$.

Moreover, $W_\lambda \le W$ for all $\lambda \in \Lambda$. For let $\lambda \in \Lambda$ be given. Choose k with $s_k = m(W_{\lambda_k}) > m(W_\lambda)$. By the above, we have $W_{\lambda_k} \ge W_\lambda$. Since W is the increasing union of the W_{λ_k}, we have $W \ge W_{\lambda_k}$ for all k. Therefore $W \ge W_\lambda$, and W is an upper bound in $\mathcal{W}$ for the given chain.

By Zorn's Lemma, there is a maximal element W^* in $\mathcal{W}$. Then $D = \cup_{i=-\infty}^{\infty} T^i W^*$ is T-invariant, $T|_D$ is completely dissipative, and $C = X \setminus D$ cannot contain any wandering set of positive measure, by maximality of W^*, so $T|_C$ is conservative. □

Because of this decomposition, when looking for a σ-finite equivalent invariant measure we may assume that the nonsingular system $(X, \mathcal{B}, m, T)$ is conservative, for if not we can always construct one on the dissipative part.

Remark 1 If $(X, \mathcal{B}, m)$ is nonatomic and $T: X \to X$ is nonsingular, invertible, and *ergodic*, in the sense that if $A \in \mathcal{B}$ satisfies $T^{-1}A = A = TA$ then either $m(A) = 0$ or $m(A^c) = 0$, then T is conservative. For if W is a wandering set of positive measure, taking any $A \subset W$ with $0 < m(A) < m(W)$ and forming $\cup_{i=-\infty}^{\infty} T^i A$ will produce an invariant set of positive measure whose complement also has positive measure.

We want to reduce the problem of existence of a σ-finite equivalent invariant measure to that of a finite one by using first-return maps to sets of finite measure. For this purpose it will be necessary to know that every conservative nonsingular system is *recurrent*: almost every point of each set of positive measure returns at some future time to that set. This is easy to see, because for each $B \in \mathcal{B}$, the set

$$B^* = \bigcup_{i=1}^{\infty} T^{-i} B \tag{58}$$

is wandering. In fact much more is true.

Theorem 8 ([21]) *For any nonsingular system $(X, \mathcal{B}, m, T)$ the following properties are equivalent:*

1. *The system is incompressible: for each $B \in \mathcal{B}$ such that $T^{-1}B \subset B$, we have $m(B \setminus T^{-1}B) = 0$.*
2. *The system is recurrent; for each $B \in \mathcal{B}$, with B^* defined as above, $m(B \setminus B^*) = 0$.*
3. *The system is conservative: there are no wandering sets of positive measure.*
4. *The system is infinitely recurrent: for each $B \in \mathcal{B}$, almost every point of B returns to B infinitely many times, equivalently,*

$$m\left(B \setminus \bigcap_{n=0}^{\infty} \bigcup_{i=n}^{\infty} T^{-i} B\right) = m\left(B \setminus \bigcap_{n=0}^{\infty} T^{-n} B^*\right) = 0 . \tag{59}$$

There is a very slick proof by F. B. Wright [38] of this result in the even more general situation of a Boolean σ-algebra homomorphism (reproduced in [28]).

Using First-Return Maps, and Counterexamples to Existence

Now given a nonsingular conservative system $(X, \mathcal{B}, m, T)$ and a set $B \in \mathcal{B}$, for each $x \in B$ there is a smallest $n_B(x) \geq 1$ such that

$$T^{n_B}(x) \in B . \tag{60}$$

We define the *first-return map* $T_B: B \to B$ by

$$T_B(x) = T^{n_B(x)}(x) \quad \text{for all } x \in B . \tag{61}$$

Using derivative maps, it is easy to reduce the problem of existence of a σ-finite equivalent invariant measure to that of existence of finite equivalent invariant measures, in a way.

Theorem 9 *(see [14]) Let T be a conservative nonsingular transformation on a probability space $(X, \mathcal{B}, m)$. Then there is a σ-finite T-invariant measure $\mu \sim m$ if and only if there is an increasing sequence of sets $B_n \in \mathcal{B}$ with $\cup_{n=1}^{\infty} B_n = X$ such that for each n the first-return map T_{B_n} has a finite invariant measure equivalent to m restricted to B_n.*

Proof 7 Given a σ-finite equivalent invariant measure μ, let the B_n be sets of finite μ-measure that increase to X. Conversely, given such a sequence B_n with finite invariant measures μ_n for the first-return maps T_{B_n}, extend μ_1 in the obvious way to an (at least σ-finite) invariant measure on the full orbit $A_1 = \cup_{i=-\infty}^{\infty} T^i B_1$. Then replace B_2 by $B_2 \setminus A_1$, and continue. □

There are many more checkable conditions for existence of a σ-finite equivalent invariant measure in the literature. There are also examples of invertible ergodic nonsingular systems for which *there does not exist* any σ-finite equivalent invariant measure due to Ornstein [27] and subsequently Chacon [9], Brunel [5], L. Arnold [2], and others.

Invariant Measures for Maps of the Interval or Circle

Finally we mention sample theorems from a huge array of such results about existence of finite invariant measures for maps of an interval or of the circle.

Theorem 10 ("Folklore Theorem" [1]) *Let $X = (0, 1)$ and denote by m Lebesgue measure on X. Let $T: X \to X$ be a map for which there is a finite or countable partition $\alpha = \{A_i\}$ of X into half-open intervals $[a_i, b_i)$ satisfying the following conditions. Denote by A_i^0 the interior of each interval A_i. Suppose that*

1. *for each i, $T: A_i^0 \to X$ is one-to-one and onto;*
2. *T is C^2 on each A_i^0;*
3. *there is an n such that*

$$\inf_i \inf_{x \in A_i^0} |(T^n)'(x)| > 1 ; \tag{62}$$

4. *for each i,*

$$\sup_{x,y,z \in A_i^0} \left|\frac{T''(x)}{T'(y)T'(z)}\right| < \infty . \tag{63}$$

Then for each measurable set B, $\lim_{n\to\infty} m(T^{-n}B) = \mu(B)$ exists and defines the unique T-invariant ergodic probability measure on X that is equivalent to m.

Moreover, the partition α is weakly Bernoulli for T, so that the natural extension of T is isomorphic to a Bernoulli system.

A key example to which the theorem applies is that of the *Gauss map* $Tx = 1/x \bmod 1$ with the partition for which $A_i = [1/(i+1), 1/i)$ for each $i = 1, 2, \ldots$. Coding orbits to $\mathbb{N}^{\mathbb{N}}$ by letting $a(x) = a_i$ if $x \in A_i$ carries T to the shift on the continued fraction expansion $[a_1, a_2, \ldots]$ of x. It was essentially known already to Gauss that T preserves the measure whose density with respect to Lebesgue measure is $1/((1+x)\log 2)$.

Theorem 11 (see [22]) *Let $X = S^1$, the unit circle, and let $T: X \to X$ be a (noninvertible) C^2 map which is expanding, in the sense that $|T'(x)| > 1$ everywhere. Then there is a unique finite invariant measure μ equivalent to Lebesgue measure m, and in fact μ is ergodic and the Radon-Nikodym derivative $\mathrm{d}\mu/\mathrm{d}m$ has a continuous version.*

Some Mathematical Background

Lebesgue Spaces

Definition 5 Two measure spaces $(X, \mathcal{B}, \mu)$ and (Y, C, ν) are *isomorphic* (sometimes also called *isomorphic mod 0*) if there are subsets $X_0 \subset X$ and $Y_0 \subset Y$ such that $\mu(X_0) = 0 = \nu(Y_0)$ and a one-to-one onto map $\phi: X \setminus X_0 \to Y \setminus Y_0$ such that ϕ and ϕ^{-1} are measurable and $\mu(\phi^{-1}C) = \nu(C)$ for all measurable $C \subset Y \setminus Y_0$.

Definition 6 A *Lebesgue space* is a finite measure space that is isomorphic to a measure space consisting of a (possibly empty) finite subinterval of $\mathbb{R}$ with the σ-algebra of Lebesgue measurable sets and Lebesgue measure, possibly together with countably many atoms (point masses).

The *measure algebra* of a measure space $(X, \mathcal{B}, \mu)$ consists of the pair $(\hat{\mathcal{B}}, \hat{\mu})$, with $\hat{\mathcal{B}}$ the Boolean σ-algebra (see Sect. "A Range of Actions", 3.) of $\mathcal{B}$ modulo the σ-ideal of sets of measure 0, together with the operations induced by set operations in $\mathcal{B}$, and $\hat{\mu}$ is induced on $\hat{\mathcal{B}}$ by μ on $\mathcal{B}$. Every measure algebra $(\hat{\mathcal{B}}, \hat{\mu})$ is a metric space with the metric $d(A, B) = \hat{\mu}(A \triangle B)$ for all $A, B \in \hat{\mathcal{B}}$. It is *nonatomic* if whenever $A, B \in \hat{\mathcal{B}}$ and $A < B$ (which means $A \wedge B = A$), either $A = 0$ or $A = B$. A *homomorphism of measure algebras* $\psi: (\hat{C}, \hat{\nu}) \to (\hat{\mathcal{B}}, \hat{\mu})$ is a Boolean σ-algebra homomorphism such that $\hat{\mu}(\hat{C}) = \hat{\nu}(C)$ for all $\hat{C} \in \hat{C}$. The inverse of any factor map $\phi: X \to Y$ from a measure space $(X, \mathcal{B}, \mu)$ to a measure space (Y, C, ν) induces a homomorphism of measure algebras $(\hat{C}, \hat{\nu}) \to (\hat{\mathcal{B}}, \hat{\mu})$. We say that a measure algebra is *normalized* if the measure of the maximal element is 1: $\hat{\mu}(0') = 1$.

We work within the class of Lebesgue spaces because (1) they are the ones commonly encountered in the wide range of naturally arising examples; (2) they allow us to assume if we wish that we are dealing with a familiar space such as $[0, 1]$ or $\{0, 1\}^{\mathbb{N}}$; and (3) they have the following useful properties.

- (Carathéodory [7]) Every normalized and nonatomic measure algebra whose associated metric space is separable (has a countable dense set) is measure-algebra isomorphic with the measure algebra of the unit interval with Lebesgue measure.
- (von Neumann [37]) Every complete separable metric space with a Borel probability measure on the completion of the Borel sets is a Lebesgue space.
- (von Neumann [37]) Every homomorphism $\psi: (\hat{C}, \hat{\nu}) \to (\hat{\mathcal{B}}, \hat{\mu})$ of the measure algebras of two Lebesgue spaces (Y, C, ν) and $(X, \mathcal{B}, \mu)$ comes from a factor map: there are a set $X_0 \subset X$ with $\mu(X_0) = 0$ and a measurable map $\phi: X \setminus X_0 \to Y$ such that ψ coincides with the map induced by ϕ^{-1} from $\hat{C}$ to $\hat{\mathcal{B}}$.

Rokhlin Theory V. A. Rokhlin [31] provided an axiomatic, intrinsic characterization of Lebesgue spaces. The key ideas are the concept of a basis and the correspondence of factors with complete sub-σ-algebras and (not necessarily finite or countable) measurable partitions of a special kind.

Definition 7 A *basis* for a complete measure space $(X, \mathcal{B}, \mu)$ is a countable family $C = \{C_1, C_2, \ldots\}$ of measurable sets which *generates* $\mathcal{B}$: For each $B \in \mathcal{B}$ there is $C \in \mathcal{B}(C)$ (the smallest σ-algebra of subsets of X that contains C) such that $B \subset C$ and $\mu(C \setminus B) = 0$; and *separates the points of* X: For each $x, y \in X$ with $x \neq y$, there is $C_i \in C$ such that either $x \in C_i, y \notin C_i$ or else $y \in C_i, x \notin C_i$.

Coarse sub-σ-algebras of $\mathcal{B}$ may not separate points of X and thus may lead to equivalence relations, partitions, and factor maps. Partitions of the following kind deserve careful attention.

Definition 8 Let $(X, \mathcal{B}, \mu)$ be a complete measure space and ξ a partition of X, meaning that up to a set of measure 0, X is the union of the elements of ξ, which are pairwise disjoint up to sets of measure 0. We call ξ an *R-partition* if there is a countable family $D = \{D_1, D_2, \ldots\}$ of ξ-saturated sets (that is, each D_i is a union of elements of ξ) such that

$$\begin{aligned} &\text{for all distinct } E, F \in \xi\,, \text{ there is } D_i \text{ such that} \\ &\text{either } E \subseteq D_i\,,\ F \not\subseteq D_i \ \ (\text{so } F \subset D_i^c) \\ &\text{or } F \subseteq D_i\,,\ E \not\subseteq D_i \ \ (\text{so } E \subset D_i^c)\,. \end{aligned} \tag{64}$$

Any such family D is called a *basis* for ξ.

Note that each element of an R-partition is necessarily measurable: if $C \in \xi$ with basis $\{D_i\}$, then

$$C = \bigcap \{D_i : C \subseteq D_i\} . \tag{65}$$

Every countable or finite measurable partition of a complete measure space is an R-partition. The orbit partition of a measure-preserving transformation is often *not* an R-partition. (For example, if the transformation is ergodic, the corresponding factor space will be trivial, consisting of just one cell, rather than corresponding to the partition into orbits as required.)

For any set $B \subset X$, let $B^0 = B$ and $B^1 = B^c = X \setminus B$.

Definition 9 A basis $C = \{C_1, C_2, \dots\}$ for a complete measure space $(X, \mathcal{B}, \mu)$ is called *complete*, and the space is called *complete with respect to the basis*, if for every 0,1-sequence $e \in \{0,1\}^{\mathbb{N}}$,

$$\bigcap_{i=1}^{\infty} C_i^{e_i} \neq \emptyset . \tag{66}$$

C is called *complete mod 0* (and $(X, \mathcal{B}, \mu)$ is called *complete mod 0 with respect to* C, if there is a complete measure space $(X', \mathcal{B}', \mu')$ with a complete basis C' such that X is a full-measure subset of X', and $C_i = C_i' \cap X$ for all $i = 1, 2, \dots$.

From the definition of basis, each intersection in (66) contains at most one point. The space $\{0,1\}^{\mathbb{N}}$ with Bernoulli $1/2, 1/2$ measure on the completion of the Borel sets has the complete basis $C_i = \{\omega : \omega_i = 0\}$.

Proposition 5 *If a measure space is complete mod 0 with respect to one basis, then it is complete mod 0 with respect to every basis.*

Theorem 12 ([31]) *A measure space is a Lebesgue space (that is, isomorphic mod 0 with the usual Lebesgue measure space of a possibly empty subinterval of $\mathbb{R}$ possibly together with countably many atoms) if and only if it has a complete basis.*

In a Lebesgue space $(X, \mathcal{B}, \mu)$ there is a one-to-one onto correspondence between complete sub-σ-algebras of $\mathcal{B}$ (that is, those for which the restriction of the measure yields a complete measure space) and R-partitions of X:

Given an R-partition ξ, let $\mathcal{B}(\xi)$ denote the σ-algebra generated by ξ, which consists of all sets in $\mathcal{B}$ that are ξ-*saturated* – unions of members of ξ – and let $\overline{\mathcal{B}}(\xi)$ denote the completion of $\mathcal{B}(\xi)$ with respect to μ.

Conversely, given a complete sub-σ-algebra $C \subset \mathcal{B}$, define an equivalence relation on X by $x \sim y$ if for all $A \in C$, either $x, y \in A$ or else $x, y \in A^c$. The measure algebra $(\hat{C}, \hat{\mu})$ has a countable dense set $\hat{C}_0$ (take a countable dense set $\{\hat{B}_i\}$ for $(\hat{\mathcal{B}}, \hat{\mu})$ and, for each i, j for which it is possible, choose $\hat{C}_{ij}$ within distance $1/2^j$ of $\hat{B}_i$). Then representatives $C_i \in C$ of the $\hat{C}_{ij}$ will be a basis for the partition ξ corresponding to the equivalence relation $\sim$.

Given any family $\{\mathcal{B}_\lambda\}$, of complete sub-σ-algebras of $\mathcal{B}$, their *join* is the intersection of all the sub-σ-algebras that contain their union:

$$\bigvee_\lambda \mathcal{B}_\lambda = \mathcal{B}(\bigcup_\lambda \mathcal{B}_\lambda) , \tag{67}$$

and their *infimum* is just their intersection:

$$\bigwedge_\lambda \mathcal{B}_\lambda = \mathcal{B}(\bigcap_\lambda \mathcal{B}_\lambda) . \tag{68}$$

These σ-algebra operations correspond to the supremum and infimum of the corresponding families of R-partitions. We say that a partition ξ_1 is *finer* than a partition ξ_2, and write $\xi_1 \geq \xi_2$, if every element of ξ_2 is a union of elements of ξ_1. Given any family $\{\xi_\lambda\}$ of R-partitions, there is a coarsest R-partition $\bigvee_\lambda \xi_\lambda$ which refines all of them, and a finest R-partition $\bigwedge_\lambda \xi_\lambda$ which is coarser than all of them. We have

$$\bigvee_\lambda \mathcal{B}(\xi_\lambda) = \mathcal{B}(\bigvee_\lambda \xi_\lambda) , \bigwedge_\lambda \mathcal{B}(\xi_\lambda) = \mathcal{B}(\bigwedge_\lambda \xi_\lambda) . \tag{69}$$

Now we discuss the relationship among factor maps $\phi : X \to Y$ from a Lebesgue space $(X, \mathcal{B}, \mu)$ to a complete measure space (Y, C, ν), complete sub-σ-algebras of $\mathcal{B}$, and R-partitions of X. Given such a factor map ϕ, $\mathcal{B}_Y = \phi^{-1}C$ is a complete sub-σ-algebra of $\mathcal{B}$, and the equivalence relation $x_1 \sim x_2$ if $\phi(x_1) = \phi(x_2)$ determines an R-partition ξ_Y. (A basis for ξ can be formed from a countable dense set in $\hat{\mathcal{B}}_Y$ as above.)

Conversely, given a complete sub-σ-algebra $C \subset \mathcal{B}$, the identity map $(X, \mathcal{B}, \mu) \to (X, C, \mu)$ is a factor map. Alternatively, given an R-partition of X, we can form a measure space $(X/\xi, \mathcal{B}(\xi), \mu_\xi)$ and a factor map $\phi_\xi : X \to X/\xi$ as follows. The space X/ξ is just ξ itself; that is, the points of X/ξ are the members (cells, or atoms) of the partition ξ. $\mathcal{B}(\xi)$ consists of the ξ-saturated sets in $\mathcal{B}$ considered as subsets of ξ, and μ_ξ is the restriction of μ to $\mathcal{B}(\xi)$. Completeness of $(X, \mathcal{B}, \mu)$ forces completeness of $(X/\xi, \mathcal{B}(\xi), \mu_\xi)$. The map $\phi_\xi : X \to X/\xi$ is defined by letting $\phi(x) = \xi(x) =$ the element of ξ to which x belongs. Thus for a Lebesgue space $(X, \mathcal{B}, \mu)$, there is a perfect correspondence among images under factor maps, complete sub-σ-algebras of $\mathcal{B}$, and R-partitions of X.

Theorem 13 *If $(X, \mathcal{B}, \mu)$ is a Lebesgue space and (Y, C, ν) is a complete measure space that is the image of $(X, \mathcal{B}, \mu)$ under a factor map, then (Y, C, ν) is also a Lebesgue space.*

Theorem 14 *Let $(X, \mathcal{B}, \mu)$ be a Lebesgue space, (Y, C, ν) a separable measure space (that is, one with a countable basis as above, equivalently one with a countable dense set in its measure algebra), and $\phi\colon X \to Y$ a measurable map ($\phi^{-1}C \subset \mathcal{B}$). Then φ is also forward measurable: if $A \subset X$ is measurable, then $\phi(A) \subset Y$ is measurable.*

Theorem 15 *Let $(X, \mathcal{B}, \mu)$ be a Lebesgue space.*

1. *Every measurable subset of X, with the restriction of $\mathcal{B}$ and μ, is a Lebesgue space. Conversely, if a subset A of X with the restrictions of $\mathcal{B}$ and μ is a Lebesgue space, then A is measurable ($A \in \mathcal{B}$).*
2. *The product of countably many Lebesgue spaces is a Lebesgue space.*
3. *Every measure algebra isomorphism of $(\hat{\mathcal{B}}, \hat{\mu})$ (defined as above) is induced by a point isomorphism mod 0.*

Disintegration of Measures Every R-partition ξ of a Lebesgue space $(X, \mathcal{B}, \mu)$ has associated with it a *canonical system of conditional measures*: Using the notation of the preceding section, for μ_ξ-almost every $C \in X/\xi$, there are a σ-algebra $\mathcal{B}_C$ of subsets of C and a measure m_C on $\mathcal{B}_C$ such that:

1. $(C, \mathcal{B}_C, m_C)$ is a Lebesgue space;
2. for every $A \in \mathcal{B}$, $A \cap C \in \mathcal{B}_C$ for μ_ξ-almost every $C \in \xi$;
3. for every $A \in \mathcal{B}$, the map $C \to m_C(A \cap C)$ is $\mathcal{B}(\xi)$-measurable on X/ξ;
4. for every $A \in \mathcal{B}$,

$$\mu(A) = \int_{X/\xi} m_C(A \cap C)\, \mathrm{d}\mu_\xi(C)\,. \tag{70}$$

It follows that for $f \in L^1(X)$, (a version of) its conditional expectation (see the next section) with respect to the factor algebra corresponding to ξ is given by

$$\mathbb{E}(f|\mathcal{B}(\xi)) = \int_C f\, \mathrm{d}m_C \quad \text{on } \mu_\xi \text{ - a.e. } C \in \xi\,, \tag{71}$$

since the right-hand side is $\mathcal{B}(\xi)$-measurable and for each $A \in \mathcal{B}(\xi)$, its integral over any $B \in \mathcal{B}(\xi)$ is, as required, $\mu(A \cap B)$ (use the formula on $B/(\xi|B)$).

It can be shown that a canonical system of conditional measures for an R-partition of a Lebesgue space is essentially unique, in the sense that any two measures m_C and m'_C will be equal for μ_ξ-almost all $C \in \xi$. Also, any partition of a Lebesgue space that has a canonical system of conditional measures must be an R-partition.

These conditional systems of measures can be used to prove the ergodic decomposition theorem and to show that every factor situation is essentially projection of a skew product onto the base (see [32]).

Theorem 16 *Let $(X, \mathcal{B}, \mu)$ be a Lebesgue space. If ξ is an R-partition of X, $\{(C, \mathcal{B}_C, m_C)\}$ is a canonical system of conditional measures for ξ, and $A \in \mathcal{B}$, define $\mu(A|C) = m_C(A \cap C)$. Then:*

1. *for every $A \in \mathcal{B}$, $\mu(A|\xi(x))$ is a measurable function of $x \in X$;*
2. *if (ξ_n) is an increasing sequence of R-partitions of X, then for each $A \in \mathcal{B}$*

$$\mu(A|\xi_n(x)) \to \mu(A|\bigvee_n \xi_n(X)) \quad a.e.\ \mathrm{d}\mu\,; \tag{72}$$

3. *if (ξ_n) is a decreasing sequence of R-partitions of X, then for each $A \in \mathcal{B}$*

$$\mu(A|\xi_n(x)) \to \mu(A|\bigwedge_n \xi_n(X)) \quad a.e.\ \mathrm{d}\mu\,. \tag{73}$$

This is a consequence of the Martingale and Reverse Martingale Convergence Theorems. The statements hold just as well for $f \in L^1(X)$ as for $f = \mathbf{1}_A$ for some $A \in \mathcal{B}$.

Conditional Expectation

Let $(X, \mathcal{B}, \mu)$ be a σ-finite measure space, $f \in L^1(X)$, and $\mathcal{F} \subset \mathcal{B}$ a sub-σ-algebra of $\mathcal{B}$. Then

$$\nu(F) = \int_F f\, \mathrm{d}\mu \tag{74}$$

defines a finite signed measure on $\mathcal{F}$ which is absolutely continuous with respect to μ restricted to $\mathcal{F}$. So by the Radon-Nikodym Theorem there is a function $g \in L^1(X, \mathcal{F}, \mu)$ such that

$$\nu(F) = \int_F g\, \mathrm{d}\mu \quad \text{for all } F \in \mathcal{F}\,. \tag{75}$$

Any such function g, which is unique as an element of $L^1(X, \mathcal{F}, \mu)$ (and determined only up to sets of μ-measure 0) is called a version of the *conditional expectation* of f with respect to $\mathcal{F}$, and denoted by

$$g = \mathbb{E}(f|\mathcal{F})\,. \tag{76}$$

As an element of $L^1(X, \mathcal{B}, \mu)$, $\mathbb{E}(f|\mathcal{F})$ is characterized by the following two properties:

$$\mathbb{E}(f|\mathcal{F}) \quad \text{is } \mathcal{F}\text{-measurable}\,; \tag{77}$$

$$\int_F \mathbb{E}(f|\mathcal{F})\,\mathrm{d}\mu = \int_F f\,\mathrm{d}\mu \quad \text{for all } F \in \mathcal{F}\,. \tag{78}$$

We think of $\mathbb{E}(f|\mathcal{F})(x)$ as our expected value for f if we are *given the information in* $\mathcal{F}$, in the sense that for each $F \in \mathcal{F}$ we know whether or not $x \in F$. When $\mathcal{F}$ is the σ-algebra generated by a finite measurable partition α of X and f is the characteristic function of a set $A \in \mathcal{B}$, the conditional expectation gives the conditional probabilities of A with respect to all the sets in α:

$$\begin{aligned}\mathbb{E}(\mathbf{1}_A|\mathcal{F})(x) &= \mu(A|\alpha(x)) \\ &= \mu(A \cap F)/\mu(F) \;\text{ if } x \in F \in \alpha\,.\end{aligned} \tag{79}$$

We write $\mathbb{E}(f) = \mathbb{E}(f|\{\emptyset, X\}) = \int_X f\,d\mu$ for the expectation of any integrable function f. A measurable function f on X is *independent* of a sub-σ-algebra $\mathcal{F} \subset \mathcal{B}$ if for each $(a, b) \subset \mathbb{R}$ and $F \in \mathcal{F}$ we have

$$\mu(f^{-1}(a, b)) \cap F) = \mu(f^{-1}(a, b))\,\mu(F)\,. \tag{80}$$

A function $\tau : \mathbb{R} \to \mathbb{R}$ is *convex* if whenever $t_1, \ldots, t_n \geq 0$ and $\sum_{i=1}^n t_i = 1$,

$$\phi(\sum_{i=1}^n t_i x_i) \leq \sum_{i=1}^n t_i\,\phi(x_i) \quad \text{for all } x_1, \ldots, x_n \in \mathbb{R}\,. \tag{81}$$

Theorem 17 *Let $(X, \mathcal{B}, \mu)$ be a probability space and $\mathcal{F} \subset \mathcal{B}$ a sub-σ-algebra.*

1. *$\mathbb{E}(\cdot|\mathcal{F})$ is a positive contraction on $L^p(X)$ for each $p \geq 1$.*
2. *If $f \in L^1(X)$ is $\mathcal{F}$-measurable, then $\mathbb{E}(f|\mathcal{F}) = f$ a.e. If $f \in L^\infty(X)$ is $\mathcal{F}$-measurable, then $\mathbb{E}(fg|\mathcal{F}) = f\mathbb{E}(g|\mathcal{F})$ for all $g \in L^1(X)$.*
3. *If $\mathcal{F}_1 \subset \mathcal{F}_2$ are sub-σ-algebras of $\mathcal{B}$, then $\mathbb{E}(\mathbb{E}(f|\mathcal{F}_2)|\mathcal{F}_1) = \mathbb{E}(f|\mathcal{F}_1)$ a.e. for each $f \in L^1(X)$.*
4. *If $f \in L^1(X)$ is independent of the sub-σ-algebra $\mathcal{F} \subset \mathcal{B}$, then $\mathbb{E}(f|\mathcal{F}) = \mathbb{E}(f)$ a.e.*
5. *If $\phi : \mathbb{R} \to \mathbb{R}$ is convex, f and $\phi \circ f \in L^1(X)$, and $\mathcal{F} \subset \mathcal{B}$ is a sub-σ-algebra, then $\phi(\mathbb{E}(f|\mathcal{F})) \leq \mathbb{E}(\phi \circ f|\mathcal{F})$ a.e.*

The Spectral Theorem

A *separable Hilbert space* is one with a countable dense set, equivalently a countable orthonormal basis. A *normal operator* is a continuous linear operator S on a Hilbert space $\mathcal{H}$ such that $SS^* = S^*S$, S^* being the adjoint operator defined by $(Sf, g) = (f, S^*g)$ for all $f, g \in \mathcal{H}$. A continuous linear operator S is *unitary* if it is invertible and $S^* = S^{-1}$. Two operators S_1 and S_2 on Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively, are called *unitarily equivalent* if there is a unitary operator $U : \mathcal{H}_1 \to \mathcal{H}_2$ which carries S_1 to S_2: $S_2U = US_1$. The following brief account follows [10,30].

Theorem 18 *Let $S : \mathcal{H} \to \mathcal{H}$ be a normal operator on a separable Hilbert space $\mathcal{H}$. Then there are mutually singular Borel probability measures $\mu_\infty, \mu_1, \mu_2, \ldots$ such that S is unitarily equivalent to the operator M on the direct sum Hilbert space*

$$L^2(\mathbb{C}, \mu_\infty) \oplus L^2(\mathbb{C}, \mu_1) \oplus \left(\bigoplus_{k=1}^2 L^2(\mathbb{C}, \mu_2)\right) \oplus \cdots \oplus \left(\bigoplus_{k=1}^m L^2(\mathbb{C}, \mu_m)\right) \oplus \ldots \tag{82}$$

defined by

$$\begin{aligned}&M((f_{\infty,1}(z_{\infty,1}), f_{\infty,2}(z_{\infty,2}), \ldots), \\ &\qquad (f_{1,1}(z_{1,1})), (f_{2,1}(z_{2,1}), f_{2,2}(z_{2,2})), \ldots) \\ &= (z_{\infty,1} f_{\infty,1}(z_{\infty,1}), z_{\infty,2} f_{\infty,2}(z_{\infty,2}), \ldots), \\ &\qquad (z_{1,1} f_{1,1}(z_{1,1})), (z_{2,1} f_{2,1}(z_{2,1}), z_{2,2} f_{2,2}(z_{2,2})), \ldots)\,.\end{aligned} \tag{83}$$

The measures μ_i are supported on the *spectrum* $\sigma(S)$ of S, the (compact) set of all $\lambda \in \mathbb{C}$ such that $S - \lambda I$ does not have a continuous inverse. Some of the μ_i may be 0. They are uniquely determined up to absolute continuity equivalence. The smallest absolute continuity class with respect to which all the μ_i are absolutely continuous is called the *maximum spectral type of* S. A measure representing this type is $\sum_i \mu_i/2^i$. We have in mind the example for which $\mathcal{H} = L^2(X, \mathcal{B}, \mu)$ and $Sf = f \circ T$ (the "Koopman operator") for a measure-preserving system $(X, \mathcal{B}, \mu, T)$ on a Lebesgue space $(X, \mathcal{B}, \mu)$, which is *unitary*: it is linear, continuous, invertible, preserves scalar products, and has spectrum equal to the unit circle.

The proof of Theorem 18 can be accomplished by first decomposing $\mathcal{H}$ (in a careful way) into the direct sum of pairwise orthogonal *cyclic subspaces* $\mathcal{H}_n$: each $\mathcal{H}_n$ is the closed linear span of $\{S^i(S^*)^j f_n : i, j \geq 0\}$ for some $f_n \in \mathcal{H}$. This means that for each n the set $\{p(S, S^*)f_n : p \text{ is a polynomial in two variables}\}$ is dense in $\mathcal{H}_n$. Similarly, by the Stone-Weierstrass Theorem the set $\mathcal{P}_n$ of all polynomials $p(z, \overline{z})$ is dense in the set $C(\sigma(S|\mathcal{H}_n))$ of continuous complex-valued functions on $\sigma(S|\mathcal{H}_n)$. We define a bounded linear functional ϕ_n on $\mathcal{P}_n$ by

$$\phi(p) = (p(S, S^*)f_n, f_n) \tag{84}$$

and extend it by continuity to a bounded linear functional on $C(\sigma(S|\mathcal{H}_n))$. It can be proved that this functional is positive, and therefore, by the Riesz Representation Theorem, it corresponds to a positive Borel measure on $\sigma(S|\mathcal{H}_n)$.

The various L^2 spaces and multiplication operators involved in the above theorem can be amalgamated into a coherent whole, resulting in the following convenient form of the Spectral Theorem for normal operators

Theorem 19 *Let $S\colon \mathcal{H} \to \mathcal{H}$ be a normal operator on a separable Hilbert space $\mathcal{H}$. There are a finite measure space $(X, \mathcal{B}, \mu)$ and a bounded measurable function $h\colon X \to \mathbb{C}$ such that S is unitarily equivalent to the operator of multiplication by h on $L^2(X, \mathcal{B}, \mu)$.*

The form of the Spectral Theorem given in Theorem 18 is useful for discussing absolute continuity and multiplicity properties of the spectrum of a normal operator. Another form, involving spectral measures, has useful consequences such as the functional calculus.

Theorem 20 *Let $S\colon \mathcal{H} \to \mathcal{H}$ be a normal operator on a separable Hilbert space $\mathcal{H}$. There is a unique projection-valued measure E defined on the Borel subsets of the spectrum $\sigma(S)$ of S such that $E(\sigma(S)) = I$ (= the identity on $\mathcal{H}$);*

$$\left(E\left(\bigcup_{i=1}^{\infty} A_i\right)\right) f = \sum_{i=1}^{\infty} (E(A_i))f \tag{85}$$

whenever $A_1, A_2, \ldots$ are pairwise disjoint Borel subsets of $\sigma(S)$ and $f \in \mathcal{H}$, with the series converging in norm; and

$$S = \int_{\sigma(S)} \lambda \, \mathrm{d}E(\lambda) \,. \tag{86}$$

Spectral integrals such as the one in (86) can be defined by reducing to complex measures $\mu_{f,g}(A) = (E(A)f, g)$, for $f, g \in \mathcal{H}$ and $A \subset \sigma(S)$ a Borel set. Given a bounded Borel measurable function ϕ on $\sigma(S)$, the operator

$$V = \phi(S) = \int_{\sigma(S)} \phi(\lambda) \, \mathrm{d}E(\lambda) \tag{87}$$

is determined by specifying that

$$(Vf, g) = \int_{\sigma(S)} \phi(\lambda) \, \mathrm{d}\mu_{f,g} \quad \text{for all } f, g \in \mathcal{H} \,. \tag{88}$$

Then

$$S^k = \int_{\sigma(S)} \lambda^k \, \mathrm{d}E(\lambda) \quad \text{for all } k = 0, 1, \ldots \,. \tag{89}$$

These spectral integrals sometimes behave a bit strangely:

$$\begin{aligned} &\text{If } V_1 = \int_{\sigma(S)} \phi_1(\lambda) \, \mathrm{d}E(\lambda) \\ &\text{and } V_2 = \int_{\sigma(S)} \phi_2(\lambda) \, \mathrm{d}E(\lambda) \,, \\ &\text{then } V_1 V_2 = \int_{\sigma(S)} \phi_1(\lambda) \, \phi_2(\lambda) \, \mathrm{d}E(\lambda) \,. \end{aligned} \tag{90}$$

Finally, if $f \in \mathcal{H}$ and ν is a finite positive Borel measure that is absolutely continuous with respect to $\mu_{f,f}$, then there is g in the closed linear span of $\{S^i(S^*)^j f\colon i, j \geq 0\}$ such that $\nu = \mu_{g,g}$.

Theorem 20 can be proved by applying Theorem 19, which allows us to assume that $\mathcal{H} = L^2(X, \mathcal{B}, \mu)$ and S is multiplication by $h \in L^\infty(X, \mathcal{B}, \mu)$. For any Borel set $A \subset \sigma(S)$, let $E(A)$ be the projection operator given by multiplication by the 0, 1-valued function $\mathbf{1}_A \circ h$.

Future Directions

The mathematical study of dynamical systems arose in the late nineteenth and early twentieth century, along with measure theory and probability theory, so it is a young field with many interesting open problems. New questions arise continually from applications and from interactions with other parts of mathematics. Basic aspects of the problems of classification, topological or smooth realization, and systematic construction of measure-preserving systems remain open. There is much work to be done to understand the relations among systems and different types of systems (factors and relative properties, joinings and disjointness, various notions of equivalence with associated invariants). There is a continual need to determine properties of classes of systems and of particular systems arising from applications or other parts of mathematics such as probability, number theory, geometry, algebra, and harmonic analysis. Some of these questions are mentioned in more detail in the other articles in this collection.

Bibliography

Primary Literature

1. Adler RL (1973) *F*-expansions revisited. Lecture Notes in Math, vol 318. Springer, Berlin
2. Arnold LK (1968) On σ-finite invariant measures. Z Wahrscheinlichkeitstheorie Verw Geb 9:85–97
3. Billingsley P (1978) Ergodic theory and information. Robert E Krieger Publishing Co, Huntington NY, pp xiii,194; Reprint of the 1965 original
4. Billingsley P (1995) Probability and measure, 3rd edn. Wiley, New York, pp xiv,593

5. Brunel A (1966) Sur les mesures invariantes. Z Wahrscheinlichkeitstheorie Verw Geb 5:300–303
6. Calderón AP (1955) Sur les mesures invariantes. C R Acad Sci Paris 240:1960–1962
7. Carathéodory C (1939) Die Homomorphieen von Somen und die Multiplikation von Inhaltsfunktionen. Annali della R Scuola Normale Superiore di Pisa 8(2):105–130
8. Carathéodory C (1968) Vorlesungen über Reelle Funktionen, 3rd edn. Chelsea Publishing Co, New York, pp x,718
9. Chacon RV (1964) A class of linear transformations. Proc Amer Math Soc 15:560–564
10. Conway JB (1990) A Course in Functional Analysis, vol 96, 2nd edn. Springer, New York, pp xvi,399
11. Dowker YN (1955) On measurable transformations in finite measure spaces. Ann Math 62(2):504–516
12. Dowker YN (1956) Sur les applications mesurables. C R Acad Sci Paris 242:329–331
13. Frechet M (1924) Des familles et fonctions additives d'ensembles abstraits. Fund Math 5:206–251
14. Friedman NA (1970) Introduction to Ergodic Theory. Van Nostrand Reinhold Co., New York, pp v,143
15. Furstenberg H (1977) Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. J Anal Math 31:204–256
16. Gowers WT (2001) A new proof of Szemerédi's theorem. Geom Funct Anal 11(3):465–588
17. Gowers WT (2001) Erratum: *A new proof of Szemerédi's theorem*. Geom Funct Anal 11(4):869
18. Green B, Tao T (2007) The primes contain arbitrarily long arithmetic progressions. arXiv:math.NT/0404188
19. Hahn H (1933) Über die Multiplikation total-additiver Mengenfunktionen. Annali Scuola Norm Sup Pisa 2:429–452
20. Hajian AB, Kakutani S (1964) Weakly wandering sets and invariant measures. Trans Amer Math Soc 110:136–151
21. Halmos PR (1947) Invariant measures. Ann Math 48(2):735–754
22. Katok A, Hasselblatt B (1995) Introduction to the Modern Theory of Dynamical Systems. Cambridge University Press, Cambridge, pp xviii,802
23. Hopf E (1932) Theory of measure and invariant integrals. Trans Amer Math Soc 34:373–393
24. Hopf E (1937) Ergodentheorie. Ergebnisse der Mathematik und ihrer Grenzgebiete, 1st edn. Springer, Berlin, pp iv,83
25. Khinchin AI (1949) Mathematical Foundations of Statistical Mechanics. Dover Publications Inc, New York, pp viii,179
26. Kolmogorov AN (1956) Foundations of the Theory of Probability. Chelsea Publishing Co, New York, pp viii,84
27. Ornstein DS (1960) On invariant measures. Bull Amer Math Soc 66:297–300
28. Petersen K (1989) Ergodic Theory. Cambridge Studies in Advanced Mathematics, vol 2. Cambridge University Press, Cambridge, pp xii,329
29. Poincaré H (1987) Les Méthodes Nouvelles de la Mécanique Céleste. Tomes I, II,III. Les Grands Classiques Gauthier-Villars. Librairie Scientifique et Technique Albert Blanchard, Paris
30. Radjavi H, Rosenthal P (1973) Invariant subspaces, 2nd edn. Springer, Mineola, pp xii,248
31. Rohlin VA (1952) On the fundamental ideas of measure theory. Amer Math Soc Transl 1952:55
32. Rohlin VA (1960) New progress in the theory of transformations with invariant measure. Russ Math Surv 15:1–22
33. Royden HL (1988) Real Analysis, 3rd edn. Macmillan Publishing Company, New York, pp xx,444
34. Szemerédi E (1975) On sets of integers containing no k elements in arithmetic progression. Acta Arith 27:199–245
35. Tao T (2006) Szemerédi's regularity lemma revisited. Contrib Discret Math 1:8–28
36. Tao T (2006) Arithmetic progressions and the primes. Collect Math Extra:37–88
37. von Neumann J (1932) Einige Sätze über messbare Abbildungen. Ann Math 33:574–586
38. Wright FB (1961) The recurrence theorem. Amer Math Mon 68:247–248
39. Young L-S (2002) What are SRB measures, and which dynamical systems have them? J Stat Phys 108:733–754

Books and Reviews

Billingsley P (1978) Ergodic Theory and Information. Robert E. Krieger Publishing Co, Huntington, pp xiii,194

Cornfeld IP, Fomin SV, Sinaĭ YG (1982) Ergodic Theory. Fundamental Principles of Mathematical Sciences, vol 245. Springer, New York, pp x,486

Denker M, Grillenberger C, Sigmund K (1976) Ergodic Theory on Compact Spaces. Lecture Notes in Mathematics, vol 527. Springer, Berlin, pp iv,360

Friedman NA (1970) Introduction to Ergodic Theory. Van Nostrand Reinhold Co, New York, pp v,143

Glasner E (2003) Ergodic Theory via Joinings. Mathematical Surveys and Monographs, vol 101. American Mathematical Society, Providence, pp xii,384

Halmos PR (1960) Lectures on Ergodic Theory. Chelsea Publishing Co, New York, pp vii,101

Katok A, Hasselblatt B (1995) Introduction to the Modern Theory of Dynamical Systems. Encyclopedia of Mathematics and its Applications, vol 54. Cambridge University Press, Cambridge, pp xviii,802

Hopf E (1937) Ergodentheorie. Ergebnisse der Mathematik und ihrer Grenzgebiete, 1st edn. Springer, Berlin, pp iv,83

Jacobs K (1960) Neue Methoden und Ereignisse der Ergodentheorie. Jber Dtsch Math Ver 67:143–182

Petersen K (1989) Ergodic Theory. Cambridge Studies in Advanced Mathematics, vol 2. Cambridge University Press, Cambridge, pp xii,329

Royden HL (1988) Real Analysis, 3rd edn. Macmillan Publishing Company, New York, pp xx,444

Rudolph DJ (1990) Fundamentals of Measurable Dynamics. Oxford Science Publications. The Clarendon Press Oxford University Press, New York, pp x,168

Walters P (1982) An Introduction to Ergodic Theory. Graduate Texts in Mathematics, vol 79. Springer, New York, pp ix,250

Mechanical Computing: The Computational Complexity of Physical Devices

John H. Reif

Department of Computer Science, Duke University, Durham, USA

Article Outline

Glossary
Definition of the Subject
Introduction
The Computational Complexity of Motion Planning and Simulation of Mechanical Devices
Concrete Mechanical Computing Devices
Future Directions
Acknowledgments
Bibliography

Glossary

Mechanism A machine or part of a machine that performs a particular task computation: the use of a computer for calculation.

Computable Capable of being worked out by calculation, especially using a computer.

Simulation The term *simulation* will be used to denote both the modeling of a physical system by a computer as well as the modeling of the operation of a computer by a mechanical system; the difference will be clear from the context.

Definition of the Subject

Mechanical devices for computation appear to be largely displaced by the widespread use of microprocessor-based computers that are pervading almost all aspects of our lives. Nevertheless, mechanical devices for computation are of interest for at least three reasons:

(a) ***Historical***: The use of mechanical devices for computation is of central importance in the historical study of technologies, with a history dating back thousands of years and with surprising applications even in relatively recent times.

(b) ***Technical & Practical***: The use of mechanical devices for computation persists and has not yet been completely displaced by widespread use of microprocessor-based computers. Mechanical computers have found applications in various emerging technologies at the micro-scale that combine mechanical functions with computational and control functions not feasible by purely electronic processing. Mechanical computers also have been demonstrated at the molecular scale, and may also provide unique capabilities at that scale. The physical designs for these modern micro and molecular-scale mechanical computers may be based on the prior designs of the large-scale mechanical computers constructed in the past.

(c) ***Impact of Physical Assumptions on Complexity of Motion Planning, Design, and Simulation:*** The study of computation done by mechanical devices is also of central importance in providing lower bounds on the computational resources such as time and/or space required to simulate a mechanical system observing given physical laws. In particular, the problem of simulating the mechanical system can be shown to be *computationally hard* if a hard computational problem can be simulated by the mechanical system. A similar approach can be used to provide lower bounds on the computational resources required to solve various motion planning tasks that arise in the field of robotics. Typically, a robotic motion planning task is specified by a geometric description of the robot (or collection of robots) to be moved, its initial and final positions, the obstacles it is to avoid, as well as a model for the type of feasible motion and physical laws for the movement. The problem of planning, such as the robotic motion-planning task, can be shown to be computationally hard if a hard computational problem can be simulated by the robotic motion-planning task.

Introduction

Abstract Computing Machine Models

To gauge the computational power of a family of mechanical computers, we will use a widely known abstract computational model known as the Turing machine, defined in this section.

The Turing Machine

The *Turing machine* model formulated by Alan Turing [1] was the first complete mathematical model of an abstract computing machine that possessed universal computing power. The machine model has (i) a finite state transition control for logical control of the machine processing, (ii) a tape with a sequence of storage cells containing symbolic values, and (iii) a tape scanner for reading and writing values to and from the tape cells, which could be made to move (left and right) along the tape cells.

A machine model is *abstract* if the description of the machine transition mechanism or memory mechanism does not provide specification of the mechanical apparatus used to implement them in practice. Since Turing's description did not include any specification of the mechanical mechanism for executing the finite state transitions, it can't be viewed as a concrete mechanical computing machine, but instead is an abstract machine. Still it is valu-

able computational model, due to it simplicity and very widespread use in computational theory.

A *universal* Turing machine simulates any other Turing machine; it takes its input a pair consisting of a string providing a symbolic description of a Turing machine M and the input string x, and simulates M on input x. Because of its simplicity and elegance, the Turing machine has come to be the standard computing model used for most theoretical works in computer science. Informally, the Church–Turing hypothesis states that a Turing machine model can simulate a computation by any "reasonable" computational model (we will discuss some other reasonable computational models below).

Computational Problems

A *computational problem* is: given an input string specified by a string over a finite alphabet, determine the Boolean answer: 1 is the answer is YES, and otherwise 0. For simplicity, we generally will restrict the input alphabet to be the binary alphabet {0,1}. The *input size* of a computational problem is the number of input symbols; which is the number of bits of the binary specification of the input. (Note: It is more common to make these definitions in terms of language acceptance. A *language* is a set of strings over a given finite alphabet of symbols. A computational problem can be identified with the language consisting of all strings over the input alphabet where the answer is 1. For simplicity, we defined each complexity class as the corresponding class of problems.)

Recursively Computable Problems and Undecidable Problems

There is a large class of problems, known as *recursively computable* problems, that Turing machines compute in finite computations, that is, always halting in finite time with the answer. There are certain problems that are not recursively computable; these are called *undecidable* problems. The *Halting Problem* is: given a Turing Machine description and an input, output 1 if the Turing machine ever halts, and else output 0. Turing proved the halting problem is undecidable. His proof used a method known as a diagonalization method; it considered an enumeration of all Turing machines and inputs, and showed a contradiction occurs when a universal Turing machine attempts to solve the Halting problem for each Turing machine and each possible input.

Computational Complexity Classes

Computational complexity (see [2]) is the amount of computational resources required to solve a given computational problem. A *complexity class* is a family of problems, generally defined in terms of limitations on the resources of the computational model. The complexity classes of interest here will be associated with restrictions on the time (number of steps until the machine halts) and/or space (the number of tape cells used in the computation) of Turing machines. There are a number of notable complexity classes:

P is the complexity class associated with efficient computations, and is formally defined to be the set of problems solved by Turing machine computations running in time polynomial to the input size (typically, this is the number of bits of the binary specification of the input).

NP is the complexity class associated with combinatorial optimization problems which, if solved, can be easily determined to have correct solutions. It is formally defined to be the set of problems solved by Turing machine computations using nondeterministic choice running in polynomial time.

PSPACE is the complexity class defined as the set of problems solved by Turing machines running in space polynomial to the input size.

EXPTIME is the complexity class defined as the set of problems solved by Turing machine computations running in time exponential to the input size.

NP and **PSPACE** are widely considered to have instances that are not solvable in **P**, and it has been proved that **EXPTIME** has problems that are not in **P**.

Polynomial Time Reductions

A *polynomial time reduction* from a problem Q' to a problem Q is a polynomial time Turing machine computation that transforms any instance of the problem Q' into an instance of the problem Q which has an answer YES if and only if the problem Q' has an answer YES. Informally, this implies that problem Q can be used to efficiently solve the problem Q'. A problem Q is *hard* for a family F of problems if for every problem Q' in F, there is a polynomial time reduction from Q' to Q. Informally, this implies that problem Q can be used to efficiently solve *any* problem in F. A problem Q is *complete* for a family F of problems if Q is in C and also hard for F.

Hardness Proofs for Mechanical Problems

We will later consider various mechanical problems and characterize their computation power:

- *Undecidable mechanical problem*; this was typically proven by a computable reduction from the halting problem for a universal Turing machine problem to an

instance of the mechanical problem; this is equivalent to showing the mechanical problem can be viewed as a computational machine that can simulate a universal Turing machine computation.

- *Mechanical problems that are hard for* ***NP, PSPACE, or EXPTIME***; typically this was proved by a polynomial time reduction from the problems in the appropriate complexity class to an instance of the mechanical problem; again, this is equivalent to showing the mechanical problem can be viewed as a computational machine that can simulate a Turing machine computation in the appropriate complexity class.

The simulation proofs in either case often provide insight into the intrinsic computational power of the mechanical problem or mechanical machine.

Other Abstract Computing Machine Models

There are a number of abstract computing models discussed in this Chapter, that are equivalent, or nearly equivalent, to conventional deterministic Turing machines.

- **Reversible Turing machines** A computing device is *(logically) reversible* if each transition of its computation can be executed both in the forward direction as well as in the reverse direction, without loss of information. Landauer [3] showed that irreversible computations must generate heat in the computing process, and that reversible computations have the property that if executed slowly enough, can (in the limit) consume no energy in an adiabatic computation. A *reversible Turing machine* model allows the scan head to observe 3 consecutive tape symbols and to execute transitions both in the forward as well as in the reverse direction. Bennett [4] showed that any computing machine (e. g., an abstract machine such as a Turing machine) can be transformed to do only reversible computations, which implied that reversible computing devices are capable of universal computation. Bennett's reversibility construction required extra space to store information to insure reversibility, but this extra space can be reduced by increasing the time. Vitanyi [5] gave trade-offs between time and space in the resulting reversible machine. Lewis and Papadimitriou [95] showed that reversible Turing machines are equivalent in computational power to conventional Turing machines when the computations are bounded by polynomial time, and Crescenzi and Papadimitriou [6] proved a similar result when the computations are bounded by polynomial space. This implies that the definitions of the complexity classes **P** and **PSPACE** do not depend on the Turing machines being reversible or not. Reversible Turing machines are used in many of the computational complexity proofs to be mentioned involving simulations by mechanical computing machines.
- **Cellular automata** These are sets of finite state machines that are typically connected together by a grid network. There are known efficient simulations of the Turing machine by cellular automata (e. g., see Wolfram [7] for some known universal simulations). A number of the particle-based mechanical machines to be described are known to simulate cellular automata.
- **Randomized Turing machines** The machine can make random choices in its computation. While the use of randomized choice can be very useful in many efficient algorithms, there is evidence that randomization only provides limited additional computational power above conventional deterministic Turing machines (In particular, there are a variety of pseudo-random number generation methods proposed for producing long pseudo-random sequences from short truly random seeds, which are widely conjectured to be indistinguishable from truly random sequences by polynomial time Turning machines.) A number of the mechanical machines to be described using Brownian-motion have natural sources of random numbers.

There are also a number of abstract computing machine models that appear to be more powerful than conventional deterministic Turing machines.

- **Real-valued Turing machines** According to Blum et al. [8], each storage cell or register in these machines can store any real value (that may be transcendental). Operations are extended to allow infinite precision arithmetic operations on real numbers. To our knowledge, none of the analog computers that we will describe in this chapter have this power.
- **Quantum computers** A *quantum superposition* is a linear superposition of basis states; it is defined by a vector of complex amplitudes whose absolute magnitudes sum to 1. In a quantum computer, the quantum superposition of basis states is transformed in each step by a unitary transformation (this is a linear mapping that is reversible and always preserves the value of the sum of the absolute magnitudes of its inputs). The outputs of a quantum computation are read by observations that that project the quantum superposition to classical values; a given state is chosen with probability defined by the magnitude of the amplitude of that state in the quantum superposition. Feynman [9] and Benioff [10] were the first to suggest the use of quan-

tum mechanical principles for doing computation, and Deutsch [11] was the first to formulate an abstract model for quantum computing and show it was universal. Since then, there is a large body of work in quantum computing (see Gruska [12] and Nielsen [13]) and quantum information theory (see Jaeger [14] and Reif [15]). Some of the particle-based methods for mechanical computing described below make use of quantum phenomena, but generally are not considered to have the full power of quantum computers.

The Computational Complexity of Motion Planning and Simulation of Mechanical Devices

Complexity of Motion Planning for Mechanical Devices with Articulated Joints

The first known computational complexity result involving mechanical motion or robotic motion planning was in 1979 by Reif [16]. He considered a class of mechanical systems consisting of a finite set of connected polygons with articulated joints, which were required to be moved between two configurations in three dimensional space avoiding a finite set of fixed polygonal obstacles. To specify the movement problem (as well as the other movement problems described below unless otherwise stated), the object to be moved, as well as its initial and final positions, and the obstacles are all defined by linear inequalities with rational coefficients with a finite number of bits. He showed that this class of motion planning problems is hard for PSPACE. Since it is widely conjectured that PSPACE contains problems which are not solvable in polynomial time, this result provided the first evidence that these robotic motion planning problems were not solvable in time polynomial in n if the number of degrees of freedom grew with n. His proof involved simulating a reversible Turing machine with n tape cells by a mechanical device with n articulated polygonal arms that had to be maneuvered through a set of fixed polygonal obstacles similar to the channels in Swiss-cheese. These obstacles where devised to force the mechanical device to simulate transitions of the reversible Turing machine to be simulated, where the positions of the arms encoded the tape cell contents, and tape read/write operations were simulated by channels of the obstacles which forced the arms to be reconfigured appropriately. This class of movement problems can be solved by reduction to the problem of finding a path in a $O(n)$ dimensional space avoiding a fixed set of polynomial obstacle surfaces, which can be solved by a PSPACE algorithm from Canny [17]. Hence this class of movement problems are PSPACE complete. (In the case where the object to be moved consists of only one rigid polygon, the problem is known as the piano mover's problem and has a polynomial time solution by Schwartz and Sharir [18].)

Other PSPACE Completeness Results for Mechanical Devices

There were many subsequent PSPACE completeness results for mechanical devices (two of which we mention below), which generally involved *multiple degrees of freedom*:

- **The Warehouseman's Problem** In 1984 Schwartz and Sharir [19] showed that moving a set of n disconnected polygons in two dimensions from an initial position to a final position among a finite set of fixed polygonal obstacles is PSPACE hard.

There are two classes of mechanical dynamic systems, the Ballistic machines and the Browning Machines described below, that can be shown to provide simulations of polynomial space Turing machine computations.

Ballistic Collision-Based Computing Machines and PSPACE

A *ballistic computer* (see Bennett [20,21]) is a conservative dynamical system that follows a mechanical trajectory isomorphic to the desired computation. It has the following properties:

- Trajectories of distinct ballistic computers can't be merged,
- All operations of a computational must be reversible,
- Computations, when executed at constant velocity, require no consumption of energy,
- Computations must be executed without error, and need to be isolated from external noise and heat sources.

Collision-based computing [22] is computation by a set of particles, where each particle holds a finite state value, and state transformations are executed at the time of collisions between particles. Since collisions between distinct pairs of particles can be simultaneous, the model allows for parallel computation. In some cases the particles can be configured to execute cellular automata computations [23]. Most proposed methods for *Collision-based computing* are ballistic computers as defined above. Examples of concrete physical systems for collision-based computing are:

- **The billiard ball computers** Fredkin and Toffoli [24] considered a mechanical computing model, the *billiard ball computer*, consisting of spherical billiard balls with polygonal obstacles, where the billiard balls were as-

sumed to have perfect elastic collisions with no friction. They showed in 1982 that a Billiard Ball Computer, with an unbounded number of billiard balls, could simulate a reversible computing machine model that used reversible Boolean logical gates known as Toffoli gates. When restricted to a finite set of n spherical billiard balls, their construction provides a simulation of a polynomial space-reversible Turing machine.

- **Particle-like waves in excitable medium** Certain classes of excitable medium have discrete models that can exhibit particle-like waves which propagate through the media [25]. Using this phenomena, Adamatzky [26] gave a simulation of a universal Turing Machine. If restricted to n particle-waves, his simulation provides a simulation of a polynomial space Turing Machine.
- **Soliton computers** A soliton is a wave packet that maintains a self-reinforcing shape as it travels at constant speed through a nonlinear dispersive media. A soliton computer [27,28] makes use of optical solitons to hold state, and state transformations are made by colliding solitons.

Brownian Machines and PSPACE

In a mechanical system exhibiting *fully Brownian motion*, the parts move freely and independently, up to the constraints that either link the parts together or forces the parts exert on each other. In a fully Brownian motion, the movement is entirely due to heat and there is no other source of energy driving the movement of the system. An example of a mechanical system with fully Brownian motion is a set of particles exhibiting Browning motion, as with electrostatic interaction. The rate of movement a of mechanical system with fully Brownian motion is determined entirely by the drift rate in the random walk of their configurations.

In other mechanical systems, known as *driven Brownian motion*, the system's movement is only partly due to heat; in addition, there is a source of energy driving the movement of the system. Examples of driven Brownian motion systems are:

- Feynman's Ratchet and Pawl [29], which is a mechanical ratchet system that has a driving force but that can operate reversibly.
- Polymerase enzyme, which uses ATP as fuel to drive their average movement forward, but also can operate reversibly.

There is no energy consumed by fully Brownian motion devices, whereas driven Brownian motion devices require power that grows as a quadratic function of the drive rate in which operations are executed (see Bennett [21]).

Bennett [20] provides two examples of Brownian computing machines:

- An *enzymatic machine* This is a hypothetical biochemical device that simulates a Turing machine, using polymers to store symbolic values in a manner to similar to Turing machine tapes, and uses hypothetical enzymatic reactions to execute state transitions and read/write operations into the polymer memory. Shapiro [30] also describes a mechanical Turing machine whose transitions are executed by hypothetical enzymatic reactions.
- A *clockwork computer* This is a mechanism with linked articulated joints, with a Swiss-cheese like set of obstacles, which force the device to simulate a Turing machine. In the case where the mechanism of Bennett's clockwork computer is restricted to have a linear number of parts, it can be used to provide a simulation of PSPACE similar that of [16].

Hardness Results for Mechanical Devices with a Constant Number of Degrees of Freedom

There were also additional computation complexity hardness results for mechanical devices, which only involved a *constant number of degrees of freedom*. These results exploited special properties of the mechanical systems to do the simulation.

- **Motion planning with moving obstacles** Reif and Sharir [31] considered the problem of planning the motion of a rigid object (the robot) between two locations, while avoiding a set of obstacles, some of which are rotating. They showed this problem is PSPACE hard. This result was perhaps surprising, since the number of degrees of freedom of movement of the object to be moved was constant. However, the simulation used the rotational movement of obstacles to force the robot to be moved only to a position that encoded all the tape cells of M. The simulation of a Turing machine M was made by forcing the object between such locations (that encoded the entire n tape cell contents of M) at particular times, and further forced that object to move between these locations over time in a way that simulated state transitions of M.

NP Hardness Results for Path Problems in Two and Three Dimensions

Shortest path problems in fixed dimensions involve only a constant number of degrees of freedom. Nevertheless,

there are a number of NP hardness results for such problems. These results also led to proofs that certain physical simulations (in particular, simulation of multi-body molecular and celestial simulations) are NP hard, and therefore not likely efficiently computable with high precision.

- **Finding shortest paths in three dimensions** Consider the problem of finding a shortest path of a point in three dimensions (where distance is measured in the Euclidean metric) avoiding fixed polyhedral obstacles whose coordinates are described by rational numbers with a finite number of bits. This shortest path problem can be solved in PSPACE [17], but the precise complexity of the problem is an open problem. Canny and Reif [32] were the first to provide a hardness complexity result for this problem; they showed the problem is NP hard. Their proof used novel techniques called *free path encoding* that used 2^n homotopy equivalence classes of shortest paths. Using these techniques, they constructed exponentially many shortest path classes (with distinct homotopy) in single-source multiple-destination problems involving $O(n)$ polygonal obstacles. They used each of these paths to encode a possible configuration of the nondeterministic Turing machine with n binary storage cells. They also provided a technique for simulating each step of the Turing machine by the use of polygonal obstacles whose edges forced a permutation of these paths that encoded the modified configuration of the Turing machine. These encodings allowed them to prove that the single-source single-destination problem in three dimensions is NP-hard. Similar free path encoding techniques were used for a number of other complexity hardness results for the mechanical simulations described below.
- **Kinodynamic planning** *Kinodynamic planning* is the task of motion planning while subject to simultaneous kinematic and dynamic constraints. The algorithms for various classes of kinodynamic planning problems were first developed in [33]. Canny and Reif [32] also used free path encoding techniques to show that two dimensional kinodynamic motion planning with a bounded velocity is NP-hard.
- **Shortest curvature-constrained path planning in two dimensions** We now consider *curvature-constrained shortest path problems* which involve finding a shortest path by a point among polygonal obstacles, where the there is an upper bound on the path curvature. A class of curvature-constrained shortest path problems in two dimensions were shown to be NP hard by Reif and Wang [34] by devising a set of obstacles that forced the shortest curvature-constrained path to simulate a given nondeterministic Turing machine.

PSPACE Hard Physical Simulation Problems

- **Ray tracing with a rational placement and geometry** Ray tracing is defined as determining if a light ray reaches some given final position, given an optical system and the position and direction of the initial light ray. This problem of determining the path of light ray through an optical system was first formulated by Newton in his book on Optics. Ray tracing has been used for designing and analyzing optical systems. It is also used extensively in computer graphics to render scenes with complex curved objects under global illumination. Reif, Tygar, and Yoshida [35] showed the problem of ray tracing in various three dimensional optical systems, where the optical devices either consist of reflective objects defined by quadratic equations or refractive objects defined by linear equations, but in either case the coefficients are restricted to be rational. They showed that these ray tracing problems are PSPACE hard. Their proof used free path encoding techniques for simulating a nondeterministic linear space Turing machine, where the position of the ray as it enters a reflective or refractive optical object (such as a mirror or prism face) encodes the entire memory of the Turing machine to be simulated, and further steps of the Turing machine are simulated by optically inducing appropriate modifications in the position of the ray as it enters other reflective or refractive optical objects. This result implies that the apparently simple task of highly precise ray tracing through complex optical systems is not likely to be efficiently executed by a computer in polynomial time. It is another example of the use of a physical system to do powerful computations.
- **Molecular and gravitational mechanical systems** A quite surprising example of using physical systems to do computation is the work of Tate and Reif [36] on the complexity of n-body simulations, where they showed that certain n-body simulation problems are PSPACE hard, and therefore not likely to be efficiently computable with high precision. In particular, they considered multi-body systems in three dimensions with n particles and inverse polynomial force laws between each pair of particles (e. g., molecular systems with Columbic force laws or celestial simulations with gravitational force laws). It is quite surprising that such systems can be configured to do computation. Their hardness proof made use of free path encoding techniques similar to the proof of PSPACE-hardness of ray

tracing. A single particle, which we will call the *memory-encoding particle*, is distinguished. The position of a memory-encoding particle as it crosses a plane encodes the entire memory of the Turing machine to be simulated, and further steps of the Turing machine are simulated by inducing modifications in the trajectory of the memory-encoding particle. The modifications in the trajectory of the memory-encoding particle are made by the use of other particles that have trajectories which induce force fields that essentially act like force-mirrors, causing reflection-like changes in the trajectory of the memory-encoding particle. Hence, highly precise n-body molecular simulation is not likely to be efficiently executed by a polynomial time computer.

A Provably Intractable Mechanical Simulation Problem: Compliant Motion Planning with Uncertainty in Control

Next, we consider compliant motion planning with uncertainty in control. Specifically, we consider a point in 3 dimensions which is commanded to move in a straight line, but whose actual motion may differ from the commanded motion, possibly involving sliding against obstacles. Given that the point initially lies in some start region, the problem is to find a sequence of commanded velocities that is guaranteed to move the point to the goal. This problem was shown by Canny and Reif [32] to be non-deterministic EXPTIME hard, making it the first provably intractable problem in robotics. Their proof used free path encoding techniques that exploited the uncertainty of position to encode exponential number of memory bits in a Turing machine simulation.

Undecidable Mechanical Simulation Problems

- **Motion planning with friction** Consider a class of mechanical systems whose parts consist of a finite number of rigid objects defined by linear or quadratic surface patches connected by frictional contact linkages between the surfaces. (Note: this class of mechanisms is similar to the analytical engine developed by Babbage described in the next sections, except that there are smooth frictional surfaces rather than toothed gears). Reif and Sun [37] proved that an arbitrary Turing machine could be simulated by a (universal) frictional mechanical system in this class consisting of a finite number of parts. The entire memory of a universal Turing machine was encoded in the rotational position of a rod. In each step, the mechanism used a construct similar to Babbage's machine to execute a state transition. The key idea in their construction is to utilize frictional clamping to allow for setting arbitrary high gear transmission. This allowed the mechanism to make state transitions for arbitrary number of steps. Simulation of a universal Turing machine implied that the movement problem is undecidable when there are frictional linkages. (A problem is *undecidable* if there is no Turing machine that solves the problem for all inputs in finite time.) It also implied that a mechanical computer could be constructed with only a constant number of parts that has the power of an unconstrained Turing machine.
- **Ray tracing with non-rational postitioning** Consider again the problem of ray tracing in a three dimensional optical systems, where the optical devices again may either consist of reflective objects defined by quadratic equations, or refractive objects defined by linear equations. Reif et al. [35] also proved that in the case where the coefficients of the defining equations are not restricted to be rational and include at least one irrational coefficient, then the resulting ray tracing problem could simulate a universal Turing machine, and so is undecidable. This ray tracing problem for reflective objects is equivalent to the problem of tracing the trajectory of a single particle bouncing between quadratic surfaces, which is also undecidable by this same result of [35]. An independent result of Moore [38] also showed that the problem of tracing the trajectory of a single particle bouncing between quadratic surfaces is undecidable.
- **Dynamics and nonlinear mappings** Moore [39], Ditto [40] and Munakata et al. [41] have also given universal Turing machine simulations of various dynamical systems with nonlinear mappings.

Concrete Mechanical Computing Devices

Mechanical computers have a very extensive history; some surveys are given in Knott [42], Hartree [43], Engineering Research Associates [44], Chase [45], Martin [46], and Davis [47]. Norman [48] recently provided a unique overview of mechanical calculators and other historical computers, summarizing the contributions of notable manuscripts and publications on this topic.

Mechanical Devices for Storage and Sums of Numbers

Mechanical methods such as notches on stones and bones, or knots and piles of pebbles, have been used since the Neolithic period for storing and summing integer values. One example of such a device, the *abacus*, which may have been developed in Babylonia approximately 5000 years

ago, makes use of beads sliding on cylindrical rods to facilitate addition and subtraction calculations.

Analog Mechanical Computing Devices

Computing devices are considered here to be *analog* (as opposed to digital) if they don't provide a method for restoring calculated values to discrete values, whereas digital devices provide restoration of calculated values to discrete values. (Note that both analog and digital computers use some kind of physical quantity to represent values that are stored and computed, so the use of physical encoding of computational values is not necessarily the distinguishing characteristic of analog computing.) Descriptions of early analog computers are given by Horsburgh [49], Turck [50], Svoboda [51], Hartree [43], Engineering Research Associates [44], and Soroka [52]. There are a wide variety of mechanical devices used for analog computing:

- **Mechanical devices for astronomical and celestial calculation** While we do not have sufficient space in this article to fully discuss this rich history, we note that various mechanisms for predicting lunar and solar eclipses using optical illumination of configurations of stones and monoliths (for example, Stonehenge) appear to date to the Neolithic period. Mechanical mechanisms for more precisely predicting lunar and solar eclipses may have been developed in the classical period of ancient history. The most impressive and sophisticated known example of an ancient gear-based mechanical device is the Antikythera Mechanism, which is thought to have been constructed by Greeks in approximately 2200 years ago. Recent research [53] provides evidence it may have been used to predict celestial events such as lunar and solar eclipses by the analog calculation of arithmetic-progression cycles. Like many other intellectual heritages, some elements of the design of such sophisticated gear-based mechanical devices may have been preserved by the Arabs after that period, and then transmitted to the Europeans in the middle ages.
- **Planimeters** There is a considerable history of mechanical devices that integrate curves. A *planimeter* is a mechanical device that integrates the area of the region enclosed by a two dimensional closed curve, where the curve is presented as a function of the angle from some fixed interior point within the region. One of the first known planimeters was developed by J.A. Hermann in 1814 and improved (as the polar planimeter) by Hermann in 1856. This led to a wide variety of mechanical integrators known as wheel-and-disk integrators, whose input is the angular rotation of a rotating disk and whose output, provided by a tracking wheel, is the integral of a given function of that angle of rotation. More general mechanical integrators known as ball-and-disk integrators, whose input provided 2 degrees of freedom (the phase and amplitude of a complex function), were developed by James Thomson in 1886. There are also devices, such as the Integraph of Abdank Abakanoviez (1878) and C.V. Boys (1882), which integrate a one-variable real function of x presented as a curve $y = f(x)$ on the Cartesian plane. Mechanical integrators were later widely used in WWI and WWII military analog computers for solution of ballistics equations, artillery calculations and target tracking. Various other integrators are described in Morin [54].
- **Harmonic Analyzers** A *Harmonic Analyzer* is a mechanical device which calculates the coefficients of the Fourier Transform of a complex function of time, such as a sound wave. Early harmonic analyzers were developed by Thomson [55] and Henrici [56] using multiple pulleys and spheres, known as ball-and-disk integrators.
- **Harmonic Synthesizers** A *Harmonic Synthesizer* is a mechanical device that interpolates a function, given the Fourier coefficients. Thomson (then known as Lord Kelvin) [57] developed the first known Harmonic Analyzer in 1886 which used an array of James Thomson's (his brother) ball-and-disk integrators. Kelvin's Harmonic Synthesizer made use of these Fourier coefficients to reverse this process and interpolate function values, by using a wire wrapped over the wheels of the array to form a weighted sum of their angular rotations. Kelvin demonstrated the use of these analog devices predicted the tide heights of a port: first his Harmonic Analyzer calculated the amplitude and phase of the Fourier harmonics of solar and lunar tidal movements, and then his Harmonic Synthesizer formed their weighted sum to predict the tide heights over time. Many other Harmonic Analyzers were later developed, including one by Michelson and Stratton (1898) which performed Fourier analysis using an array of springs. Miller [58] gives a survey of these early Harmonic Analyzers. Fisher [59] made improvements to the tide predictor, and later Doodson and Légé increased the scale of this design to a 42-wheel version that was used up to the early 1960s.
- **Analog equation solvers** There are various mechanical devices for calculating the solution of sets of equations. Kelvin also developed one of the first known mechanical mechanisms for equation solving, involving the motion of pulleys and tilting plate that solved

sets of simultaneous linear equations specified by the physical parameters of the ropes and plates. In the 1930s, John Wilbur increased the scale of Kelvin's design to solve nine simultaneous linear algebraic equations. Leonardo Torres Quevedo constructed various rotational mechanical devices, for determining the real and complex roots of a polynomial. Svoboda [51] describes the state of art in the 1940s of mechanical analog computing devices using linkages.

- **Differential Analyzers** A *Differential Analyzer* is a mechanical analog device using linkages for solving ordinary differential equations. Vannevar Bush [60] developed the first Differential Analyzer at MIT in 1931, which used a torque amplifier to link multiple mechanical integrators. Although it was considered a general-purpose mechanical analog computer, this device required a physical reconfiguration of the mechanical connections to specify a given mechanical problem to be solved. In subsequent Differential Analyzers, the reconfiguration of the mechanical connections was made automatic by resetting electronic relay connections. In addition to the military applications already mentioned above, analog mechanical computers incorporating differential analyzers have been widely used for flight simulations and for industrial control systems.
- **Mechanical simulations of physical processes: Crystallization and packing** There are a variety of macroscopic devices used for simulations of physical processes, which can be viewed as analog devices. For example, a number of approaches have been used for mechanical simulations of crystallization and packing:
  - ***Simulation using solid macroscopic ellipsoids bodies*** Simulations of kinetic crystallization processes have been made by collections of macroscopic solid ellipsoidal objects – typically of diameter of a few millimeters – which model the molecules comprising the crystal. In these physical simulations, thermal energy is modeled by introducing vibrations; a low level of vibration is used to model freezing and increasing the level of vibrations models melting. In simple cases, the molecule of interest is a sphere, and ball bearings or similar objects are used for the molecular simulation. For example, to simulate the dense random packing of hard spheres within a crystalline solid, Bernal [61] and Finney [62] used up to 4000 ball bearings on a vibrating table. In addition, to model more general ellipsoidal molecules, orzo pasta grains as well as M&M candies (Jerry Gollub at Princeton University) have been used. Also, Cheerios have been used to simulate the liquid state packing of benzene molecules. To model more complex systems mixtures of balls of different sizes and/or composition have been used; for example a model ionic crystal formation has been made by use a mixture of balls composed of different materials that acquired opposing electrostatic charges.
  - ***Simulations using bubble rafts*** [63,64] These are the structures that assemble among equal sized bubbles floating on water. They typically they form two dimensional hexagonal arrays, and can be used for modeling the formation of close packed crystals. Defects and dislocations can also be modeled [65]; for example, the deliberate introduction of defects in the bubble rats have been used to simulate crystal dislocations, vacancies, and grain boundaries. Also, impurities in crystals (both interstitial and substitutional) have been simulated by introducing bubbles of other sizes.
- **Reaction-diffusion chemical computers** Adamatzky [66,67] described a class of analog computers where there is a chemical medium which has multiple chemical species, where the concentrations of these chemical species vary spatially and which diffuse and react in parallel. The memory values (as well as inputs and outputs) of the computer are encoded by the concentrations of these chemical species at a number of distinct locations (also known as micro-volumes). The computational operations are executed by chemical reactions whose reagents are these chemical species. Example computations [66,67] include: (i) Voronoi diagram; this is to determine the boundaries of the regions closest to a set of points on the plane, (ii) Skeleton of planar shape, and (iii) a wide variety of two dimensional patterns periodic and aperiodic in time and space.

Digital Mechanical Devices for Arithmetic Operations

Recall that we have distinguished digital mechanical devices from the analog mechanical devices described above by their use of mechanical mechanisms for insuring the values stored and computed are discrete. Such discretization mechanisms include geometry and structure (e. g., the notches of Napier's bones described below), or cogs and spokes of wheeled calculators. Surveys of the history of some these digital mechanical calculators are given by Knott [42], Turck [50], Hartree [43], Engineering Research Associates [44], Chase [45], Martin [46], Davis [47], and Norman [48].

- **Leonardo da Vinci's mechanical device and mechanical counting devices** This intriguing device, which involved a sequence of interacting wheels positioned on

a rod, which appear to provide a mechanism for digital carry operations, was illustrated in 1493 in Leonardo da Vinci's Codex Madrid [68]. A working model of its possible mechanics was constructed in 1968 by Joseph Mirabella. It's function and purpose is not decisively known, but it may have been intended for counting rotations (e. g., for measuring the distance traversed by a cart). There are a variety of apparently similar mechanical devices used to measuring distances traversed by vehicles.

- **Napier's Bones** In 1614 John Napier [69] developed a mechanical device (known as Napier's Bones) which allowed multiplication and division (as well as square and cube roots) to be done by addition and multiplication operations. It consisted of rectilinear rods, which provided a mechanical transformation to and from logarithmic values. In 1623, Wilhelm Shickard developed a six digit mechanical calculator that combined the use of Napier's Bones using columns of sliding rods, with the use of wheels used to sum up the partial products for multiplication.
- **Slide rules** Edmund Gunter devised in 1620 a method for calculation that used a single log scale with dividers along a linear scale; this anticipated key elements of the first slide rule described by William Oughtred [70] in 1632. A very large variety of slide machines were later constructed.
- **Pascaline: Pascal's wheeled calculator** Blaise Pascal [71] developed a calculator in 1642 known as the Pascaline that could calculate all four arithmetic operations (addition, subtraction, multiplication, and division) on up to eight digits. A wide variety of mechanical devices were then developed that used revolving drums or wheels (cogwheels or pinwheels) to do various arithmetical calculations.
- **Stepped drum calculators** Gottfried Wilhelm von Leibniz developed an improved calculator known as the Stepped Reckoner in 1671, which used a cylinder known as a *stepped drum* with nine teeth of different lengths that increase in equal amounts around the drum. The stepped drum mechanism allowed the use of a moving slide for specifying a number to be input to the machine, and made use of the revolving drums to do the arithmetic calculations. Charles Xavier Thomas de Colbrar developed a widely used arithmetic mechanical calculator based on the stepped drum known as the Arithometer in 1820. Other stepped drum calculating devices included Otto Shweiger's Millionaire calculator (1893) and Curt Herzstark's Curta (early 1940s).
- **Pinwheel calculators** Frank S. Baldwin and W.T. Odhner independently invented another class of calculators in the 1870s, known as pinwheel calculators; they used a pinwheel for specifying a number input to the machine and used revolving wheels to do the arithmetic calculations. Pinwheel calculators were widely used up to the 1950s, for example in William S. Burroughs' calculator/printer and the German Brunsviga.

Digital Mechanical Devices for Mathematical Tables and Functions

- **Babbage's Difference Engine** Charles Babbage [72, 73] invented a mechanical device in 1820 known as the *Difference Engine* for calculating the tables of an analytical function (such as the logarithm), which summed the change in values of the function when a small difference is made in the argument. For each table entry, the difference calculation required a small number of simple arithmetic computations. The device made use of columns of cogwheels to store digits of numerical values. Babbage planned to store 1000 variables, each with 50 digits, where each digit was stored by a unique cogwheel. It used cogwheels in registers for the required arithmetical calculations, and also made use of a rod-based control mechanism specialized for control of these arithmetic calculations. The design and operation of the mechanisms of the device were described by a symbolic scheme developed by Babbage [74]. He also conceived of a printing mechanism for the device. In 1801, Joseph-Marie Jacquard invented an automatic loom that made use of punched cards for the specification of fabric patterns woven by his loom, and Charles Babbage proposed the use of similar punched cards for providing inputs to his machines. He demonstrated over a number of years certain key portions of the mechanics of the device but never completed a complete function device.
- **Other Difference Engines** In 1832 Ludgate [75] independently designed, but did not construct, a mechanical computing machine similar but smaller in scale to Babbage's Analytical Engine. In 1853 Pehr and Edvard Scheutz [76] constructed in Sweden a cog wheel mechanical calculating device (similar to the Difference Engine originally conceived by Babbage) known as the Tabulating Machine, for computing and printing out tables of mathematical functions. This (and a later construction of Babbage's Difference Engine by Doron Swade [77] of the London Science Museum) demonstrated the feasibility of Babbage's Difference Engine.
- **Babbage's Analytical Engine** Babbage further conceived (but did not attempt to construct) a mechani-

cal computer known as the Analytical Engine to solve more general mathematical problems. Lovelace's extended description of Babbage's Analytical Engine [78] (translation of "Sketch of the Analytical Engine" by L.F. Menabrea) describes, in addition to arithmetic operations, also mechanisms for looping and memory addressing. However, the existing descriptions of Babbage's Analytical Engine appear to lack the ability to execute a full repertory of logical and/or finite state transition operations required for general computation. Babbage's background was very strong in analytic mathematics, but he (and the architects of similar cog-wheel based mechanical computing devices at that date) seemed to have lacked knowledge of sequential logic and it's Boolean logical basis, which was required for controlling the sequence of complex computations. This (and his propensity for changing designs prior to the completion of the machine construction) might have been the real reason for the lack of complete development of a universal mechanical digital computing device in the early 1800's.

- **Subsequent electromechanical digital computing devices with cog-wheels** Other electromechanical digital computing devices (see [44]) developed in the late 1940s and 1950s which contain cog-wheels included Howard Aiken's Mark 1 [79], constructed at Harvard University, and Konrad Zuse's Z series computer, constructed in Germany.

Mechanical Devices for Timing, Sequencing and Logical Control

We will use the term *mechanical automata* here to denote mechanical devices that exhibit autonomous control of their movements. These can require sophisticated mechanical mechanisms for timing, sequencing and logical control.

- **Mechanisms used for timing control** Mechanical clocks, and other mechanical devices for measuring time have a very long history, and include a very wide variety of designs, including the flow of liquids (e. g., water clocks), or sands (e. g., sand clocks), and more conventional pendulum-and-gear based clock mechanisms. A wide variety of mechanical automata and other control devices make use of mechanical timing mechanisms to control the order and duration of events automatically executed (for example, mechanical slot machines dating up to the 1970s made use of such mechanical clock mechanisms to control the sequence of operations used for payout of winnings). As a consequence, there is an interwoven history in the development of mechanical devices for measuring time and the development of devices for the control of mechanical automata.
- **Logical control of computations** A critical step in the history of computing machines was the development in the middle 1800's of Boolean logic by George Boole [80,81]. Boole's innovation was to assign values to logical propositions: 1 for true propositions and 0 for false propositions. He introduced the use of Boolean variables which are assigned these values, as well the use of Boolean connectives ("and," and "or") for expressing symbolic Booelan logic formulas. Boole's symbolic logic is the basis for the logical control used in modern computers. Shannon [82] was the first to make use of Boole's symbolic logic to analyze relay circuits (these relays were used to control an analog computer, namely MIT's Differential Equalizer).
- **The Jevons' logic piano: A mechanical logical inference machine** In 1870 William Stanley Jevons (who also significantly contributed to the development of symbolic logic) constructed a mechanical device [83,84] for the inference of logical proposition that used a piano keyboard for inputs. This *mechanical inference machine* is less widely known than it should be, since it may have had impact in the subsequent development of logical control mechanisms for machines.
- **Mechanical logical devices used to play games** Mechanical computing devices have also been constructed for executing the logical operations for playing games. For example, in 1975, a group of MIT undergraduates including Danny Hillis and Brian Silverman constructed a computing machine made of Tinkertoys that plays a perfect game of tic-tac-toe.

Mechanical Devices Used in Cryptography

- **Mechanical cipher devices using cogwheels** Mechanical computing devices that used cogwheels were also developed for a wide variety of other purposes beyond merely arithmetic. A wide variety of mechanical computing devices were developed for the encryption and decryption of secret messages. Some of these (most notably the family of German electromechanical cipher devices known as *Enigma Machines* [85] developed in the early 1920s for commercial use and refined in the late 1920s and 1930s for military use) made use of sets of cogwheels to permute the symbols of text message streams. Similar (but somewhat more advanced) electromechanical cipher devices were used by the USSR up to the 1970s.

- **Electromechanical computing devices used in breaking cyphers** In 1934 Marian Rejewski and a team including Alan Turing constructed an electrical/mechanical computing device known as the Bomb, which had an architecture similar to the abstract Turing machine described below, and which was used to decrypt ciphers made by the German Enigma cipher device mentioned above.

Mechanical and Electro-Optical Devices for Integer Factorization

- **Lehmer's number sieve computer** In 1926 Derrick Lehmer [86] constructed a mechanical device called the "number sieve computer" for various number theoretic problems including factorization of small integers and solutions of Diophantine equations. The device made use of multiple bicycle chains that rotated at distinct periods to discover solutions (such as integer factors) to these number theoretic problems.
- **Shamir's TWINKLE** Adi Shamir [87,88,89] proposed a design for a optical/electric device known as TWINKLE for factoring integers, with the goal of breaking the RSA public key cryptosystem. This was unique among mechanical computing devices in that it used time durations between optical pulses to encode possible solution values. In particular, LEDs were made to flash at certain intervals of time (where each LED is assigned a distinct period and delay) at a very high clock rate so as to execute a sieve-based integer factoring algorithm.

Mechanical Computation at the Micro Scale: MEMS Computing Devices

Mechanical computers can have advantages over electronic computation at certain scales; they are already having widespread use at the microscale. MEMS (Micro-Electro-Mechanical Systems) are manufactured by lithographic etching methods similar in nature to the processes in which microelectronics are manufactured, and have a similar microscale. A wide variety of MEMS devices [90] have been constructed for sensors and actuators, including accelerometers used in automobile safety devices and disk readers, and many of these MEMS devices execute mechanical computation do their task. Perhaps the MEMS device most similar in architecture to the mechanical calculators described above is the Recodable Locking Device [91] constructed in 1998 at Sandia Labs, which made use of microscopic gears that acted as a mechanical lock, and which was intended for mechanically locking strategic weapons.

Future Directions

Mechanical Self-Assembly Processes

Most of the mechanical devices discussed in this chapter have been assumed to be constructed top-down; that is they are designed and then assembled by other mechanisms generally of large scale. However a future direction to consider are bottom-up processes for assembly and control of devices. Self-assembly is a basic bottom-up process found in many natural processes and in particular in all living systems.

- **Domino Tiling Problems** The theoretical basis for self-assembly has its roots in Domino Tiling Problems (also known as Wang tilings) as defined by Wang [92] (also see the comprehensive text of Grunbaum et al. [93]). The input is a finite set of unit size square tiles, each of whose sides are labeled with symbols over a finite alphabet. Additional restrictions may include the initial placement of a subset of these tiles, and the dimensions of the region where tiles must be placed. Assuming an arbitrarily large supply of each tile, the problem is to place the tiles, without rotation (a criterion that cannot apply to physical tiles), to completely fill the given region so that each pair of abutting tiles have identical symbols on their contacting sides.
- **Turing-universal and NP complete self-assemblies** Domino tiling problems over an infinite domain with only a constant number of tiles were first proved by [94] to be undecidable. Lewis and Papadimitriou [95] showed the problem of tiling a given finite region is NP complete.
- **Theoretical models of tiling self-assembly processes** Domino tiling problems do not presume or require a specific process for tiling. Winfree [96] proposed kinetic models for self-assembly processes. The sides of the tiles are assumed to have some methodology for selective affinity, which we call pads. Pads function as programmable binding domains, which hold together the tiles. Each pair of pads have specified binding strengths (a real number on the range [0,1] where 0 denotes no binding and 1 denotes perfect binding). The self-assembly process is initiated by a singleton tile (the seed tile) and proceeds by tiles binding together at their pads to form aggregates known as tiling assemblies. The preferential matching of tile pads facilitates the further assembly into tiling assemblies.
- **Pad binding mechanisms** These provide a mechanism for the preferential matching of tile sides can be provided by various methods:
  - *magnetic attraction*, e. g., pads with magnetic orientations (these can be constructed by curing fer-

rite materials; e. g., PDMS polymer/ferrite composites) in the presence of strong magnet fields, and also pads with patterned strips of magnetic orientations,
  - *capillary force*, using hydrophobic/hydrophilic (capillary) effects at surface boundaries that generate lateral forces,
  - *shape matching* (also known as *shape complementarity* or *conformational affinity*), using the shape of the tile sides to hold them together.
  - (Also see the sections below *discussion of the used of molecular affinity* for pad binding.)
- **Materials for tiles** There are a variety of distinct materials for tiles, at a variety of scales: Whitesides (see [97] and http://www-chem.harvard.edu/GeorgeWhitesides.html) has developed and tested multiple technologies for meso-scale self-assembly, using capillary forces, shape complementarity, and magnetic forces. Rothemund [98] gave some of the most complex known meso-scale tiling assemblies using polymer tiles on fluid boundaries with pads that use hydrophobic/hydrophilic forces. A materials science group at the U. of Wisconsin (http://mrsec.wisc.edu/edetc/selfassembly) has also tested meso-scale self-assembly using magnetic tiles.
- **Meso-Scale Tile Assemblies** Meso-Scale Tiling Assemblies have tiles of size a few millimeters up to a few centimeters. They have been experimentally demonstrated by a number of methods, such as the placement of tiles on a liquid surface interface (e. g., at the interface between two liquids of distinct density or on the surface of an air/liquid interface), and using mechanical agitation with shakers to provide a heat source for the assembly kinetics (that is, a temperature setting is made by fixing the rate and intensity of shaker agitation).
- **Applications of Meso-scale Assemblies** The are a number of applications, including:
  - Simulation of the thermodynamics and kinetics of molecular-scale self-assemblies.
  - For placement of a variety of microelectronics and MEMS parts.

Mechanical Computation at the Molecular Scale: DNA Computing Devices

Due to the difficulty of constructing electrical circuits at the molecular scale, alternative methods for computation, and in particular mechanical methods, may provide unique opportunities for computing at the molecular scale. In particular the bottom-up self-assembly processes described above have unique applications at the molecular scale.

- **Self-assembled DNA nanostructures** Molecular-scale structures known as *DNA nanostructures* (see surveys by Seeman [99] and Reif [100]) can be made to self-assemble from individual synthetic strands of DNA. When added to a test tube with the appropriate buffer solution, and the test tube is cooled, the strands self-assemble into DNA nanostructures. This self-assembly of DNA nanostrucures can be viewed as a mechanical process, and in fact can be used to do computation. The first known example of a computation by using DNA was by Adleman [101,102] in 1994; he used the self-assembly of DNA strands to solve a small instance of a combinatorial optimization problem known as the Hamiltonian path problem.
- **DNA tiling assemblies** The Wang tiling [92] paradigm for self-assembling structures was the basis for scalable and programmable approach proposed by Winfree et al. [103] for doing molecular computation using DNA. First a number of distinct DNA nanostructures known as DNA tiles are self-assembled. End portions of the tiles, known as pads, are designed to allow the tiles to bind together a programmable manner similar to Wang tiling, but in this case uses the molecular affinity for pad binding due to the hydrogen-bonding of complementary DNA bases. This programmable control of the binding together of DNA tiles provides a capability for doing computation at the molecular scale. When the temperature of the test tube containing these tiles is further lowered, the DNA tiles bind together to form complexly patterned tiling lattices that correspond to computations.
- **Assembling patterned DNA tiling assemblies** Programmed patterning at the molecular scale can be produced by the use of strands of DNA that encode the patterns; this was first done by Yan et al. [104] in the form of bar-cord striped patterns, and more recently Rothemund [105] self-assembled complex 2D molecular patterns and shapes. Another method for the molecular patterning of DNA tiles is via computation done during the assembly.
- **Computational DNA tiling assemblies** The first experimental demonstration of computation via the self-assembly of DNA tiles was in 2000 by Mao et al. [106], and Yan et al. [107], which provided a 1 dimensional computation of a binary-carry computation (known as prefix-sum) associated with binary adders. Rothemund et al. [108] in 2004 demonstrated a 2 dimensional computational assembly of tiles displaying a pattern known as the Sierpinski triangle, which is the modulo 2 version of Pascal's triangle.
- **Other autonomous DNA devices** DNA nanostruc-

tures can also be made to make sequences of movement, and a demonstration of an autonomous moving DNA robotic device that moved without outside mediation across a DNA nanostructure was given by Yin et al. [109]. The design of an autonomous DNA device that moves under programmed control is described in [110]. Surveys of DNA autonomous devices are given in [111] and [112].

Acknowledgments

We sincerely thank Charles Bennett for his numerous suggestions and very important improvements to this survey. This work has been supported by NSF grants CCF-0432038 and CCF-0523555.

Bibliography

1. Turing A (1937) On Computable Numbers, with an Application to the Entscheidungsproblem. Proc Lond Math Soc, Second Ser, London 42:230–265. Erratum in 43:544–546
2. Lewis HR, Christos PH (1997) Elements of the Theory of Computation, 2nd edn. Prentice Hall, Upper Saddle River
3. Landauer R (1961) Irreversibility and heat generation in the computing process. IBMJ Res Dev 5:183
4. Bennett CH (1973) Logical reversibility of computation. IBM J Res Dev 17(6):525–532
5. Li M, Vitanyi P (1996) Reversibility and Adiabatic Computation: Trading Time and Space for Energy. Proc Roy Soc Lond, Series A 452:769–789. Preprint quant-ph/9703022
6. Crescenzi P, Christos PH (1995) Reversible simulation of space-bounded computations. Theor Comput Sci 143(1): 159–165
7. Wolfram S (1984) Universality and complexity in cellular automata. Phys D 10:1–35
8. Blum L, Cucker F, Shub M, Smale S (1996) Complexity and Real Computation: A Manifesto. Int J Bifurc Chaos 6(1):3–26
9. Feynman RP (1982) Simulating physics with computers. Int J Theor Phys 21(6–7):467–488
10. Benioff P (1982) Quantum mechanical models of Turing machines that dissipate no energy. Phys Rev Lett 48:1581
11. Deutsch D (1985) Quantum theory, the Church–Turing principle and the universal quantum computer. Proc Roy Soc A 400:97–117
12. Gruska J (1999) Quantum Computing. McGraw–Hill, Maidenhead
13. Nielsen M, Chuang I (2000) Quantum Computation and Quantum Information. Cambridge University Press, Cambridge
14. Jaeger G (2006) Quantum Information: An Overview. Springer, Berlin
15. Reif JH (2007) Quantum Information Processing: Algorithms, Technologies and Challenges In: Eshaghian-Wilner MM (ed) Nano-scale and Bio-inspired Computing. Wiley, Malden
16. Reif JH (1979) Complexity of the Mover's Problem and Generalizations. 20th Annual IEEE Symposium on Foundations of Computer Science, San Juan, Puerto Rico, October pp 421–427; (1987) In: Schwartz J (ed) Planning, Geometry and Complexity of Robot Motion. Ablex Pub Norwood, NJ, pp 267–281
17. Canny J (1988) Some algebraic and geometric computations in PSPACE In: Cole R (ed) Proceedings of the 20th Annual ACM Symposium on the Theory of Computing. ACM Press, Chicago, IL, pp 460–467
18. Schwartz JT, Sharir M (1983) On the piano movers problem: I the case of a two-dimensional rigid polygonal body moving amidst polygonal barriers. Comm Pure Appl Math 36:345–398
19. Hopcroft JE, Schwartz JT, Sharir M (1984) On the Complexity of Motion Planning for Multiple Independent Objects: PSPACE Hardness of the Warehouseman's Problem. Int J Robot Res 3(4):76–88
20. Bennett CH (1982) The thermodynamics of computation – a review. Int J Theor Phys 21(12):905–940. http://www.research.ibm.com/people/b/bennetc/bennettc1982666c3d53.pdf
21. Bennett CH (2003) Notes on Landauer's principle, reversible computation, and Maxwell's Demon. Stud History Philos Mod Phys 34:501–510. eprint physics/0210005: http://xxx.lanl.gov/abs/physics/0210005
22. Adamatzky A (ed) (2001) Collision-based computing. Springer, London
23. Squier R, Steiglitz K (1994) Programmable parallel arithmetic in cellular automata using a particle model. Complex Syst 8:311–323
24. Fredkin E, Toffoli T (1982) Conservative logic. Int J Theor Phys 21:219–253
25. Adamatzky AI (1996) On the particle-like waves in the discrete model of excitable medium. Neural Netw World 1:3–10
26. Adamatzky AI (1998) Universal dynamical computation in multidimensional excitable lattices. Int J Theor Phys 37:3069–3108
27. Jakubowski MH, Steiglitz K, Squier R (1998) State transformations of colliding optical solitons and possible application to computation in bulk media. Phys Rev E58:6752–6758
28. Jakubowski MH, Steiglitz K, Squier R (2001) Computing with solitons: a review and prospectus, Collision-based computing. Springer, London, pp 277–297
29. Feynman RP (1963) In: Feynman RP, Leighton RB, Sands M (eds) Ratchet and Pawl, The Feynman Lectures on Physics, vol 1, Chapter 46. Addison–Wesley, Reading
30. Shapiro E (1999) A Mechanical Turing Machine: Blueprint for a Biomolecular Computer. Fifth International Meeting on DNA-Based Computers at the Massachusetts Institute of Technology, Proc DNA Based Computers V. Cambridge, MA, pp 14–16
31. Reif JH, Sharir M (1994) Motion Planning in the Presence of Moving Obstacles. 26th Annual IEEE Symposium on Foundations of Computer Science, Portland, OR, October 1985 pp 144–154; J ACM (JACM) 41(4):764–790
32. Canny J, Reif JH (1987) New Lower Bound Techniques for Robot Motion Planning Problems. 28th Annual IEEE Symposium on Foundations of Computer Science, Los Angeles, CA, October pp 49–60
33. Canny J, Donald B, Reif JH, Xavier P (1993) On the Complexity of Kinodynamic Planning. 29th Annual IEEE Symposium on Foundations of Computer Science, White Plains, NY, October (1988) pp 306–316; Kinodynamic Motion Planning. J ACM 40(5):1048–1066
34. Reif JH, Wang H (1998) The Complexity of the Two Di-

mensional Curvature-Constrained Shortest-Path Problem. Third International Workshop on Algorithmic Foundations of Robotics (WAFR98). Peters AK Ltd, Houston, pp 49–57
35. Reif JH, Tygar D, Yoshida A (1994) The Computability and Complexity of Optical Beam Tracing. 31st Annual IEEE Symposium on Foundations of Computer Science, St Louis, MO, October (1990) pp 106–114; The Computability and Complexity of Ray Tracing. Discrete Comput Geometry 11:265–287
36. Tate SR, Reif JH (1993) The Complexity of N-body Simulation. Proceedings of the 20th Annual Colloquium on Automata, Languages and Programming (ICALP'93), Lund, pp 162–176
37. Reif JH, Sun Z (2003) The Computational Power of Frictional Mechanical Systems. Third International Workshop on Algorithmic Foundations of Robotics, (WAFR98). Peters AK Ltd, Houston, Texas, Mar 5–7 1998, pp 223–236; On Frictional Mechanical Systems and Their Computational Power. SIAM J Comput (SICOMP) 32(6):1449–1474
38. Moore C (1990) Undecidability and Unpredictability in Dynamical Systems. Phys Rev Lett 64:2354–2357
39. Moore C (1991) Generalized Shifts: Undecidability and Unpredictability in Dynamical Systems. Nonlinearity 4:199–230
40. Munakata T, Sinha S, Ditto WL (2002) Chaos Computing: Implementation of Fundamental Logical Gates by Chaotic Elements. IEEE Trans Circuits Syst-I Fundam Theory Appl 49(11):1629–1633
41. Sinha S, Ditto (1999) Computing with distributed chaos. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Top 60(1):363–77
42. Knott CG (ed) (1915) Napier tercentenary memorial volume. The Royal Society of Edinburgh, Longmans, Green, London
43. Hartree DR (1950) Calculating Instruments and Machines. Cambridge University Press, Cambridge
44. Engineering Research Associates Staff (1950) High-Speed Computing Devices. McGraw–Hill, New York
45. Chase GC (1980) History of Mechanical Computing Machinery. Ann Hist Comput 2(3):198–226
46. Martin E (1992) The Calculating Machines. The MIT Press, Cambridge, Massachusetts
47. Davis M (2000) The Universal Computer: The Road from Leibniz to Turing. Norton, New York
48. Norman JM (ed) (2002) The Origins of Cyberspace: From Gutenberg to the Internet: a sourcebook on the history of information technology. Norman Publishing, Novato
49. Horsburgh EM (ed) (1914) Modern Instruments and Methodsof Calculation: a Handbook of the Napier Tercentenary Exhibition, London, G Bell and Sons, Edinburgh, The Royal Society of Edinburgh, p 223. Reprinted 1982
50. Turck JAV (1921) Origin of Modern Calculating Machines. The Western Society of Engineers, Chicago
51. Svoboda A (1948) Computing Mechanisms and Linkages. McGraw–Hill, Columbus
52. Soroka WA (1954) Analog Methods in Computation and Simulation. McGraw–Hill
53. Freeth T, Bitsakis Y, Moussas X, Seiradakis JH, Tselikas A, Mangou H, Zafeiropoulou M, Hadland R, Bate D, Ramsey A, Allen M, Crawley A, Hockley P, Malzbender T, Gelb D, Ambrisco W, Edmunds MG (2006) Decoding the ancient Greek astronomical calculator known as the Antikythera Mechanism. Nature 444:587–591
54. de Morin H (1913) Les appariéls d'intégration: intégrateurs simples et compasés, planimètres, intégromètres, intégraphes et courbes intégrales, analyse harmonique et analyseurs. Gauthier-Villars, Paris, pp 162–171
55. Thomson W (later known as Lord Kelvin) (1878) Harmonic Analyzer. Proc Roy Soc Lond 27:371–373
56. Henrici (1894) Philos Mag 38:110
57. Lord Kelvin (1878) Harmonic analyzer and synthesizer. Proc Roy Soc 27:371
58. Miller D (1916) The Henrici harmonic analyzer and devices for extending and facilitating its use. J Franklin Inst 181:51–81; 182:285–322
59. Fisher EG (1911) Tide-predicting machine. Eng News 66: 69–73
60. Bush V (1931) The differential analyzer: A new machine for solving differential equations. J Franklin Inst 212:447
61. Bernal JD (1964) The Structure of Liquids. Proc Roy Soc Lond Ser A 280, 299
62. Finney JL (1970) Random Packings and the Structure of Simple Liquids. I The Geometry of Random Close Packing. Proc Royal Soc London, Ser A, Math Phys Sci 319(1539):479–493
63. Bragg L, Nye JF (1947) A dynamical model of a crystal structure. Proc R Soc A 190:474–481
64. Bragg L, Lomer WM (1948) A dynamical model of a crystal structure II. Proc R Soc A 196:171–181
65. Corcoran SG, Colton RJ, Lilleodden ET, Gerberich WW (1997) Phys Rev B 190:474
66. Adamatzky A, De Lacy BC, Asai T (2005) Reaction-Diffusion Computers. Elsevier, Amsterdam
67. Adamatzky AI (1994) Constructing a discrete generalized Voronoi diagram in reaction-diffusion media. Neural Netw World 6:635–643
68. da Vinci L (1493) Codex Madrid I
69. Napier J (1614) Mirifici logarithmorum cannonis descriptio (the description of the wonderful canon of logarithms). Hart, Edinburgh
70. Oughtred W (1632) Circles of Proportion and the Horizontal Instrument. William Forster, London
71. Pascal E (1645) Lettre dédicatoire à Monseigneur le Chancelier sur le sujet de la machine nouvellement inventée par le sieur BP pour faire toutes sortes d'opérations d'arithmétique par un mouvement réglé sans plume ni jetons, suivie d'un avis nécessaire à ceux qui auront curiosité de voir ladite machine et de s'en servir
72. Babbage C (1822) On Machinery for Calculating and Printing Mathematical Tables. Edinburgh Philos J VII:274–281
73. Babbage C (1822) Observations on the application of machinery to the computation of mathematical tables. Memoirs of the Astronomical Society 1:311–314
74. Babbage C (1826) On a Method of expressing by Signs the Action of Machinery. Philosophical Trans Royal Soc London 116(III):250–265
75. Ludgate P (1909) On a proposed analytical engine. Sci Proc Roy Dublin Soc 12:77–91
76. Lindgren M (1990) Glory and Failure: Difference Engines of Johann Muller, Charles Babbage and Georg and Edvard Scheutz. MIT Press, Cambridge
77. Swade D (1991) Charles Babbage and His Calculating Engines. Michigan State University Press, East Lensing
78. Lovelace A (1843) Sketch of the analytical engine invented by Charles Babbage. Translation of: Sketch of the Analytical Engine by Menabrea LF with Ada's notes and extensive commentary. Esq Sci Mem 3:666–731

79. Cohen BI, Welch GW (1999) Makin' Numbers: Howard Aiken and the Computer. MIT Press, Cambridge
80. Boole G (1847) Mathematical Analysis of Logic. Pamphlet
81. Boole G (1854) An Investigation of the Laws of Thought, on Which are Founded the Mathematical Theories of Logic and Probabilities. Macmillan, Cambridge
82. Shannon C (1938) A Symbolic Analysis of Relay and Switching Circuits. Trans Am Inst Electr Eng 57:713–719
83. Jevons SW (1870) On the Mechanical Performance of Logical Inference. Philos Trans Roy Soc, Part II 160:497–518
84. Jevons SW (1873) The Principles of Science. A Treatise on Logic and Scientific Method. Macmillan, London
85. Hamer D, Sullivan G, Weierud F (1998) Enigma Variations: an Extended Family of Machines. Cryptologia 22(3):211–229
86. Lehmer DH (1928) The mechanical combination of linear forms. Am Math Mon 35:114–121
87. Shamir A (1999) Method and apparatus for factoring large numbers with optoelectronic devices. patent 475920, awarded 08/05/2003
88. Shamir A (1999) Factoring large numbers with the TWINKLE device. Cryptographic Hardware and Embedded Systems (CHES) 1999. LNCS, vol 1717. Springer, New York, pp 2–12
89. Lenstra AK, Shamir A (2000) Analysis and optimization of the TWINKLE factoring Device. Proc Eurocrypt 2000. LNCS, vol 1807. Springer, pp 35–52
90. Madou MJ (2002) Fundamentals of Microfabrication: The Science of Miniaturization, 2nd edn. CRC Press, Boca Raton
91. Plummer D, Dalton LJ, Peter F (1999) The recodable locking device. Commun ACM 42(7):83–87
92. Wang H (1963) Dominoes and the AEA Case of the Decision Problem. In: J Fox (ed) Mathematical Theory of Automata. Polytechnic Press, Brooklyn, pp 23–55
93. Branko GS, Shepard GC (1987) Tilings and Patterns. H Freeman, Gordonsville. Chapter 11
94. Berger R (1966) The Undecidability of the Domino Problem. Mem Am Math Soc 66:1–72
95. Lewis HR, Papadimitriou CH (1981) Elements of the Theory of Computation. Prentice-Hall, Upper Saddle River, pp 296–300, 345–348
96. Winfree E (1998) Simulations of Computing by Self-Assembly. In: Proceedings of the Fourth Annual Meeting on DNA Based Computers, held at the University of Pennsylvania. pp 16–19
97. Xia Y, Whitesides GM (1998) Soft Lithography. Annu Rev Mater Sci 28:153–184
98. Rothemund PWK (2000) Using lateral capillary forces to compute by self-assembly. Proc Nat Acad Sci (USA) 97:984–989
99. Seeman NC (2004) Nanotechnology and the Double Helix. Sci Am 290(6):64–75
100. Reif JH, LaBean TH (2007) Nanostructures and Autonomous Devices Assembled from DNA In: Eshaghian-Wilner MM (ed) Nano-scale and Bio-inspired Computing. Wiley, Malden
101. Adleman LM (1994) Molecular computation of solutions to combinatorial problems. Science 266(11):1021–1024
102. Adleman L (1998) Computing with DNA. Sci Am 279(2):34–41
103. Winfree E, Liu F, Wenzler LA, Seeman NC (1998) Design and Self-Assembly of Two-Dimensional DNA Crystals. Nature 394:539–544
104. Yan H, LaBean TH, Feng L, Reif JH (2003) Directed Nucleation Assembly of Barcode Patterned DNA Lattices. PNAS 100(14):8103–8108
105. Rothemund PWK (2006) Folding DNA to create nanoscale shapes and patterns. Nature 440:297–302
106. Mao C, LaBean TH, Reif JH, Seeman (2000) Logical Computation Using Algorithmic Self-Assembly of DNA Triple-Crossover Molecules. Nature 407:493–495
107. Yan H, Feng L, LaBean TH, Reif J (2003) DNA Nanotubes, Parallel Molecular Computations of Pairwise Exclusive-Or (XOR) Using DNA String Tile Self-Assembly. J Am Chem Soc (JACS) 125(47):14246–14247
108. Rothemund PWK, Papadakis N, Winfree E (2004) Algorithmic Self-Assembly of DNA Sierpinski Triangles. PLoS Biology 2(12), e424. doi:10.1371/journal.pbio.0020424
109. Yin P, Yan H, Daniel XG, Turberfield AJ, Reif JH (2004) A Unidirectional DNA Walker Moving Autonomously Along a Linear Track. Angew Chem [Int Ed] 43(37):4906–4911
110. Reif JH, Sahu S (2008) Autonomous Programmable DNA Nanorobotic Devices Using DNAzymes, John H. Reif and Sudher Sahu. In: Garzon M, Yan H (eds) Autonomous Programmable DNA Nanorobotic Devices Using DNAzymes, 13th International Meeting on DNA Computing (DNA 13). Lecture Notes for Computer Science (LNCS). Springer, Berlin. To appear in special Journal Issue on Self-Assembly, Theoretical Computer Science (TCS)
111. Reif JH, LaBean TH (2007) Autonomous Programmable Biomolecular Devices Using Self-Assembled DNA Nanostructures. Comm ACM (CACM) 50(9):46–53. Extended version: http://www.cs.duke.edu/~reif/paper/AutonomousDNA/AutonomousDNA.pdf
112. Bath J, Turberfield AJ (2007) DNA nanomachines. Nat Nanotechnol 2:275–284

Mechanical Systems: Symmetries and Reduction

Jerrold E. Marsden[1], Tudor S. Ratiu[2]

[1] Control and Dynamical Systems, California Institute of Technology, Pasadena, USA

[2] Section de Mathématiques and Bernoulli Center, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Article Outline

Glossary
Definition of the Subject
Introduction
Symplectic Reduction
Symplectic Reduction – Further Discussion
Reduction Theory: Historical Overview
Cotangent Bundle Reduction
Future Directions
Acknowledgments
Appendix: Principal Connections
Bibliography

Glossary

Lie group action A process by which a Lie group, acting as a symmetry, moves points in a space. When points in the space that are related by a group element are identified, one obtains the quotient space.

Free action An action that moves every point under any nontrivial group element.

Proper action An action that obeys a compactness condition.

Momentum mapping A dynamically conserved quantity that is associated with the symmetry of a mechanical system. An example is angular momentum, which is associated with rotational symmetry.

Symplectic reduction A process of reducing the dimension of the phase space of a mechanical system by restricting to the level set of a momentum map and also identifying phase space points that are related by a symmetry.

Poisson reduction A process of reducing the dimension of the phase space of a mechanical system by identifying phase space points that are related by a symmetry.

Equivariance Equivariance of a momentum map is a property that reflects the consistency of the mapping with a group action on its domain and range.

Momentum cocycle A measure of the lack of equivariance of a momentum mapping.

Singular reduction A reduction process that leads to non-smooth reduced spaces. Often associated with non-free group actions.

Coadjoint orbit The orbit of an element of the dual of the Lie algebra under the natural action of the group.

KKS (Kostant-Kirillov-Souriau) form The natural symplectic form on coadjoint orbits.

Cotangent bundle A mechanical phase space that has a structure that distinguishes configurations and momenta. The momenta lie in the dual to the space of velocity vectors of configurations.

Shape space The space obtained by taking the quotient of the configuration space of a mechanical system by the symmetry group.

Principal connection A mathematical object that describes the geometry of how a configuration space is related to its shape space. Related to geometric phases through the subject of holonomy. In turn related to locomotion in mechanical systems.

Mechanical connection A special (principal) connection that is built out of the kinetic energy and momentum map of a mechanical system with symmetry.

Magnetic terms These are expressions that are built out of the curvature of a connection. They are so named because terms of this form occur in the equations of a particle moving in a magnetic field.

Definition of the Subject

Reduction theory is concerned with mechanical systems with symmetries. It constructs a lower dimensional *reduced space* in which associated conservation laws are enforced and symmetries are "factored out" and studies the relation between the dynamics of the given system with the dynamics on the reduced space. This subject is important in many areas, such as stability of relative equilibria, geometric phases and integrable systems.

Introduction

Geometric mechanics has developed in the last 30 years or so into a mature subject in its own right, and its applications to problems in Engineering, Physics and other physical sciences has been impressive. One of the important aspects of this subject has to do with symmetry; even things as apparently simple as the symmetry of a system such as the n-body problem under the group of translations and rotations in space (the Euclidean group) or a wheeled vehicle under the planar Euclidean group turns out to have profound consequences. Symmetry often gives *conservation laws* through Noether's theorem and these conservation laws can be used to *reduce* the dimension of a system.

In fact, reduction theory is an old and time-honored subject, going back to the early roots of mechanics through the works of Euler, Lagrange, Poisson, Liouville, Jacobi, Hamilton, Riemann, Routh, Noether, Poincaré, and others. These founding masters regarded reduction theory as a useful tool for simplifying and studying concrete mechanical systems, such as the use of Jacobi's *elimination of the node* in the study of the n-body problem to deal with the overall rotational symmetry of the problem. Likewise, Liouville and Routh used the elimination of cyclic variables (what we would call today an Abelian symmetry group) to simplify problems and it was in this setting that the *Routh stability method* was developed.

The modern form of symplectic reduction theory begins with the works of Arnold [7], Smale [168], Meyer [117], and Marsden and Weinstein [110]. A more detailed survey of the history of reduction theory can be found in the first sections of this article. As was the case with Routh, this theory has close connections with the stability theory of *relative equilibria*, as in Arnold [8] and Simo, Lewis, and Marsden [166]. The symplectic reduction method is, in fact, by now so well known that it is used as a standard tool, often without much mention. It has also entered many textbooks on geometric mechanics and

symplectic geometry, such as Abraham and Marsden [1], Arnold [9], Guillemin and Sternberg [57], Libermann and Marle [89] and Mcduff and Salamon [116]. Despite its relatively old age, research in reduction theory continues vigorously today.

It will be assumed that the reader is familiar with the basic concepts in [104]. For the statements of the bulk of the theorems, it is assumed that the manifolds involved are finite dimensional and are smooth unless otherwise stated. While many interesting examples are infinite-dimensional, the general theory in the infinite dimensional case is still not in ideal shape; see, for example, Chernoff and Marsden [39], Marsden and Hughes [96] and Mielke [118] and examples and discussion in [92].

Notation

To keep things reasonably systematic, we have adopted the following universal conventions for some common maps and other objects:

Configuration space of a mechanical system: Q
Phase space: P
Cotangent bundle projection: $\pi_Q : T^*Q \to Q$
Tangent bundle projection: $\tau_Q : TQ \to Q$
Quotient projection: $\pi_{P,G} : P \to P/G$
Tangent map: $T\varphi : TM \to TN$ for the tangent of a map $\varphi : M \to N$

Thus, for example, the symbol $\pi_{T^*Q,G}$ denotes the quotient projection from T^*Q to $(T^*Q)/G$.

- The Lie algebra of a Lie group G is denoted $\mathfrak{g}$.
- Actions of G on a space is denoted by concatenation. For example, the action of a group element g on a point $q \in Q$ is written as gq or $g \cdot q$
- The infinitesimal generator of a Lie algebra element $\xi \in \mathfrak{g}$ for an action of G on P is denoted ξ_P, a vector field on P
- Momentum maps are denoted $\mathbf{J} : P \to \mathfrak{g}^*$.
- Pairings between vector spaces and their duals are denoted by simple angular brackets: for example, the pairing between $\mathfrak{g}$ and $\mathfrak{g}^*$ is denoted $\langle \mu, \xi \rangle$ for $\mu \in \mathfrak{g}^*$ and $\xi \in \mathfrak{g}$
- Inner products are denoted with double angular brackets: $\langle\langle u, v \rangle\rangle$.

Symplectic Reduction

Roughly speaking, here is how symplectic reduction goes: given the symplectic action of a Lie group on a symplectic manifold that has a momentum map, one divides a level set of the momentum map by the action of a suitable subgroup to form a new symplectic manifold. Before the division step, one has a manifold (that can be singular if the points in the level set have symmetries) carrying a degenerate closed 2-form. Removing such a degeneracy by passing to a quotient space is a differential-geometric operation that was promoted by Cartan [26].

The "suitable subgroup" related to a momentum mapping was identified by Smale [168] in the special context of cotangent bundles. It was Smale's work that inspired the general symplectic construction by Meyer [117] and the version we shall use, which makes explicit use of the properties of momentum maps, by Marsden and Weinstein [110].

Momentum Maps

Let G be a Lie group, $\mathfrak{g}$ its Lie algebra, and $\mathfrak{g}^*$ be its dual. Suppose that G acts symplectically on a symplectic manifold P with symplectic form denoted by Ω. We shall denote the infinitesimal generator associated with the Lie algebra element ξ by ξ_P and we shall let the Hamiltonian vector field associated to a function $f : P \to \mathbb{R}$ be denoted X_f.

A ***momentum map*** is a map $\mathbf{J} : P \to \mathfrak{g}^*$, which is defined by the condition

$$\xi_P = X_{\langle \mathbf{J}, \xi \rangle} \tag{1}$$

for all $\xi \in \mathfrak{g}$, and where $\langle \mathbf{J}, \xi \rangle : P \to \mathbb{R}$ is defined by the natural pointwise pairing. We call such a momentum map ***equivariant*** when it is equivariant with respect to the given action on P and the coadjoint action of G on $\mathfrak{g}^*$. That is,

$$\mathbf{J}(g \cdot z) = \mathrm{Ad}^*_{g^{-1}} \mathbf{J}(z) \tag{2}$$

for every $g \in G$, $z \in P$, where $g \cdot z$ denotes the action of g on the point z, Ad denotes the adjoint action, and Ad^* the coadjoint action. Note that when we write $\mathrm{Ad}^*_{g^{-1}}$, we *literally* mean the adjoint of the linear map $\mathrm{Ad}_{g^{-1}} : \mathfrak{g} \to \mathfrak{g}$. The inverse of g is necessary for this to be a *left* action on $\mathfrak{g}^*$. Some authors let that inverse be understood in the notation. However, such a convention would be a notational disaster since we need to deal with both *left* and *right* actions, a distinction that is essential in mechanics. A quadruple $(P, \Omega, G, \mathbf{J})$, where (P, Ω) is a given symplectic manifold and $\mathbf{J} : P \to \mathfrak{g}^*$ is an equivariant momentum map for the symplectic action of a Lie group G, is sometimes called a ***Hamiltonian G-space***.

Taking the derivative of the equivariance identity (2) with respect to g at the identity yields the condition of ***infinitesimal equivariance***:

$$T_z \mathbf{J}(\xi_P(z)) = -\mathrm{ad}^*_\xi \mathbf{J}(z) \tag{3}$$

for any $\xi \in \mathfrak{g}$ and $z \in P$. Here, $\mathrm{ad}_\xi : \mathfrak{g} \to \mathfrak{g}; \eta \mapsto [\xi, \eta]$ is the adjoint map and $\mathrm{ad}^*_\xi : \mathfrak{g}^* \to \mathfrak{g}^*$ is its dual. A computation shows that (3) is equivalent to

$$\langle \mathbf{J}, [\xi, \eta] \rangle = \{ \langle \mathbf{J}, \xi \rangle, \langle \mathbf{J}, \eta \rangle \} \tag{4}$$

for any $\xi, \eta \in \mathfrak{g}$, that is, $\langle \mathbf{J}, \cdot \rangle : \mathfrak{g} \to \mathcal{F}(P)$ is a Lie algebra homomorphism, where $\mathcal{F}(P)$ denotes the Poisson algebra of smooth functions on P. The converse is also true if the Lie group is connected, that is, if G is connected then an infinitesimally equivariant action is equivariant (see §12.3 in [104]).

The idea that an action of a Lie group G with Lie algebra $\mathfrak{g}$ on a symplectic manifold P should be accompanied by such an equivariant momentum map $\mathbf{J} : P \to \mathfrak{g}^*$ and the fact that the orbits of this action are themselves symplectic manifolds both occur already in Lie [90]; the links with mechanics also rely on the work of Lagrange, Poisson, Jacobi and Noether. In modern form, the momentum map and its equivariance were rediscovered by Kostant [78] and Souriau [169,170] in the general symplectic case and by Smale [168] for the case of the lifted action from a manifold Q to its cotangent bundle $P = T^*Q$. Recall that the equivariant momentum map in this case is given explicitly by

$$\big\langle \mathbf{J}(\alpha_q), \xi \big\rangle = \big\langle \alpha_q, \xi_Q(q) \big\rangle \ , \tag{5}$$

where $\alpha_q \in T^*_q Q, \xi \in \mathfrak{g}$, and where the angular brackets denote the natural pairing on the appropriate spaces.

Smale referred to $\mathbf{J}$ as the "angular momentum" by generalization from the special case $G = \mathrm{SO}(3)$, while Souriau used the French word "moment". Marsden and Weinstein [110], followed usage emerging at that time and used the word "moment" for $\mathbf{J}$, but they were soon corrected by Richard Cushman and Hans Duistermaat, who suggested that the proper English translation of Souriau's French word was "momentum," which fit better with Smale's designation as well as standard usage in mechanics. Since 1976 or so, most people who have contact with mechanics use the term momentum map (or mapping). On the other hand, Guillemin and Sternberg popularized the continuing use of "moment" in English, and both words coexist today. It is a curious twist, as comes out in work on collective nuclear motion Guillemin and Sternberg [56] and plasma physics (Marsden and Weinstein [111] and Marsden, Weinstein, Ratiu, Schmid, and Spencer [114]), that moments of inertia and moments of probability distributions can actually be the values of momentum maps! Mikami and Weinstein [119] attempted a linguistic reconciliation between the usage of "moment" and "momentum" in the context of groupoids. See [104] for more information on the history of the momentum map and Sect. "Reduction Theory: Historical Overview" for a more systematic review of general reduction theory.

Momentum Cocycles and Nonequivariant Momentum Maps

Consider a momentum map $\mathbf{J} : P \to \mathfrak{g}^*$ that *need not be equivariant*, where P is a symplectic manifold on which a Lie group G acts symplectically. The map $\sigma : G \to \mathfrak{g}^*$ that is defined by

$$\sigma(g) := \mathbf{J}(g \cdot z) - \mathrm{Ad}^*_{g^{-1}} \mathbf{J}(z) \ , \tag{6}$$

where $g \in G$ and $z \in P$ is called a ***nonequivariance or momentum one-cocycle***. Clearly, σ *is a measure of the lack of equivariance of the momentum map*.

We shall now prove a number of facts about σ. The first claim is that σ *does not depend on the point* $z \in P$ *provided that the symplectic manifold* P *is connected* (otherwise it is constant on connected components). To prove this, we first recall the following *equivariance identity for infinitesimal generators*:

$$\begin{aligned} T_q \Phi_g \left(\xi_P(q) \right) &= (\mathrm{Ad}_g \xi)_P (g \cdot q) \ ; \\ \text{i. e.,} \quad \Phi^*_g \xi_P &= \left(\mathrm{Ad}_{g^{-1}} \xi \right)_P \ . \end{aligned} \tag{7}$$

This is an easy Lie group identity that is proved, for example, in [104]; see Lemma 9.3.7.

One shows that $\sigma(g)$ is constant by showing that its Hamiltonian vector field vanishes. Using the fact that $\sigma(g)$ is independent of z along with the basic identity $\mathrm{Ad}_{gh} = \mathrm{Ad}_g \mathrm{Ad}_h$ and its consequence $\mathrm{Ad}^*_{(gh)^{-1}} = \mathrm{Ad}^*_{g^{-1}} \mathrm{Ad}^*_{h^{-1}}$, shows that σ satisfies the ***cocycle identity***

$$\sigma(gh) = \sigma(g) + \mathrm{Ad}^*_{g^{-1}} \sigma(h) \tag{8}$$

for any $g, h \in G$. This identity shows that σ produces a new action $\Theta : G \times \mathfrak{g}^* \to \mathfrak{g}^*$ defined by

$$\Theta(g, \mu) := \mathrm{Ad}^*_{g^{-1}} \mu + \sigma(g) \tag{9}$$

with respect to which the momentum map $\mathbf{J}$ is obviously equivariant. This action Θ is not linear anymore – it is an ***affine action***.

For $\eta \in \mathfrak{g}$, let $\sigma_\eta(g) = \langle \sigma(g), \eta \rangle$. Differentiating the definition of σ, namely

$$\sigma_\eta(g) = \langle \mathbf{J}(g \cdot z), \eta \rangle - \big\langle \mathbf{J}(z), \mathrm{Ad}_{g^{-1}} \eta \big\rangle$$

with respect to g at the identity in the direction $\xi \in \mathfrak{g}$ shows that

$$T_e \sigma_\eta(\xi) = \Sigma(\xi, \eta) \ , \tag{10}$$

where $\Sigma(\xi, \eta)$, which is called the ***infinitesimal nonequivariance two-cocycle***, is defined by

$$\Sigma(\xi, \eta) = \langle \mathbf{J}, [\xi, \eta] \rangle - \{\langle \mathbf{J}, \xi \rangle, \langle \mathbf{J}, \eta \rangle\} . \tag{11}$$

Since σ does not depend on the point $z \in P$, neither does Σ. Also, it is clear from this definition that Σ *measures the lack of infinitesimal equivariance of* $\mathbf{J}$. Another way to look at this is to notice that from the derivation of Eq. (10), for $z \in P$ and $\xi \in \mathfrak{g}$, we have

$$T_z\mathbf{J}(\xi_P(z)) = -\mathrm{ad}^*_\xi \mathbf{J}(z) + \Sigma(\xi, \cdot) . \tag{12}$$

Comparison of this relation with Eq. (3) also shows the relation between Σ and the infinitesimal equivariance of $\mathbf{J}$.

The map $\Sigma : \mathfrak{g} \times \mathfrak{g} \to \mathbb{R}$ is bilinear, skew-symmetric, and, as can be readily verified, satisfies the ***two-cocycle identity***

$$\Sigma([\xi, \eta], \zeta) + \Sigma([\eta, \zeta], \xi) + \Sigma([\zeta, \xi], \eta) = 0 , \tag{13}$$

for all $\xi, \eta, \zeta \in \mathfrak{g}$.

The Symplectic Reduction Theorem

There are many precursors of symplectic reduction theory. When G is Abelian, the components of the momentum map form a system of functions in involution (i. e. the Poisson bracket of any two is zero). The use of k such functions to reduce a phase space to one having $2k$ fewer dimensions may be found already in the work of Lagrange, Poisson, Jacobi, and Routh; it is well described in, for example, Whittaker [179].

In the nonabelian case, Smale [168] noted that Jacobi's elimination of the node in SO(3) symmetric problems can be understood as division of a nonzero angular momentum level by the SO(2) subgroup which fixes the momentum value. In his setting of cotangent bundles, Smale clearly stated that the coadjoint isotropy group G_μ of $\mu \in \mathfrak{g}^*$ (defined to be the group of those $g \in G$ such that $g \cdot \mu = \mu$, where the dot indicates the coadjoint action), leaves the level set $\mathbf{J}^{-1}(\mu)$ invariant (Smale [168], Corollary 4.5). However, he only divided by G_μ after fixing the total energy as well, in order to obtain the "minimal" manifold on which to analyze the reduced dynamics. The goal of his "topology and mechanics" program was to use topology, and specifically Morse theory, to study relative equilibria, which he did with great effectiveness.

Marsden and Weinstein [110] combined Souriau's momentum map for general symplectic actions, Smale's idea of dividing the momentum level by the coadjoint isotropy group, and Cartan's idea of removing the degeneracy of a 2-form by passing to the leaf space of the form's null foliation. The key observation was that the leaves of the null foliation are precisely the (connected components of the) orbits of the coadjoint isotropy group (a fact we shall prove in the next section as the *reduction lemma*). An analogous observation was made in Meyer [117], except that Meyer worked in terms of a basis for the Lie algebra $\mathfrak{g}$ and identified the subgroup G_μ as the group which left the momentum level set $\mathbf{J}^{-1}(\mu)$ invariant. In this way, he did not need to deal with the equivariance properties of the coadjoint representation.

In the more general setting of symplectic manifolds with an equivariant momentum map for a symplectic group action, the fact that G_μ acts on $\mathbf{J}^{-1}(\mu)$ follows directly from equivariance of $\mathbf{J}$. Thus, it makes sense to form the ***symplectic reduced space*** which is defined to be the quotient space

$$P_\mu = \mathbf{J}^{-1}(\mu)/G_\mu . \tag{14}$$

Roughly speaking, the symplectic reduction theorem states that, under suitable hypotheses, P_μ is itself a symplectic manifold. To state this precisely, we need a short excursion on level sets of the momentum map and some facts about quotients.

Free and Proper Actions

The action of a Lie group G on a manifold M is called a ***free action*** if $g \cdot m = m$ for some $g \in G$ and $m \in M$ implies that $g = e$, the identity element.

An action of G on M is called ***proper*** when the map $G \times M \to M \times M; (g, m) \mapsto (g \cdot m, m)$ is a proper map – that is, inverse images of compact sets are compact. This is equivalent to the statement that if m_k is a convergent sequence in M and if $g_k \cdot m_k$ converges in M, then g_k has a convergent subsequence in G.

As is shown in, for example, [2] and Duistermaat and Kolk [48], freeness, together with properness implies that the quotient space M/G is a smooth manifold and that the projection map $\pi : M \to M/G$ is a smooth surjective submersion.

Locally Free Actions

An action of G on M is called ***infinitesimally free*** at a point $m \in M$ if $\xi_M(m) = 0$ implies that $\xi = 0$. An action of G on M is called ***locally free*** at a point $m \in M$ if there is a neighborhood U of the identity in G such that $g \in U$ and $g \cdot m = m$ implies $g = e$.

Proposition 1 *An action of a Lie group G on a manifold M is locally free at $m \in M$ if and only if it is infinitesimally free at m.*

A free action is obviously locally free. The converse is not true because the action of any discrete group is locally free, but need not be globally free. When one has an action that is locally free but not globally free, one is lead to the theory of orbifolds, as in Satake [164]. In fact, quotients of manifolds by locally free and proper group actions are orbifolds, which follows by the use of the Palais slice theorem (see Palais [144]). Orbifolds come up in a variety of interesting examples involving, for example, resonances; see, for instance, Cushman and Bates [44] and Alber, Luther, Marsden, and Robbins [3] for some specific examples.

Symmetry and Singularities

If μ is a regular value of $\mathbf{J}$ then we claim that the action is automatically locally free at the elements of the corresponding level set $\mathbf{J}^{-1}(\mu)$. In this context it is convenient to introduce the notion of the ***symmetry algebra*** at $z \in P$ defined by

$$\mathfrak{g}_z = \{\xi \in \mathfrak{g} \mid \xi_P(z) = 0\} .$$

The symmetry algebra $\mathfrak{g}_z$ is the Lie algebra of the ***isotropy subgroup*** G_z of $z \in P$ defined by

$$G_z = \{g \in G \mid g \cdot z = z\} .$$

The following result (due to Smale [168] in the special case of cotangent bundles and in general to Arms, Marsden, and Moncrief [5]), is important for the recognition of regular as well as singular points in the reduction process.

Proposition 2 *An element $\mu \in \mathfrak{g}^*$ is a regular value of $\mathbf{J}$ if and only if $\mathfrak{g}_z = 0$ for all $z \in \mathbf{J}^{-1}(\mu)$.*

In other words, *points are regular points precisely when they have trivial symmetry algebra*. In examples, this gives an easy way to recognize regular points. For example, for the double spherical pendulum (see, for example, Marsden and Scheurle [108] or [95]), one can say right away that the only singular points are those with *both* pendula pointing vertically (either straight down or straight up). This result holds whether or not $\mathbf{J}$ is equivariant.

This result, connecting the symmetry of z with the regularity of μ, suggests that *points with symmetry are bifurcation points of* $\mathbf{J}$. This observation turns out to have many important consequences, including some related key convexity theorems.

Now we are ready to state the symplectic reduction theorem. We will be making two sets of hypotheses; other variants are discussed in the next section. The following notation will be convenient in the statement of the results.

SR (P, Ω) is a symplectic manifold, G is a Lie group that acts symplectically on P and has an equivariant momentum map $\mathbf{J}\colon P \to \mathfrak{g}^*$.

SRFree G acts freely and properly on P.

SRRegular Assume that $\mu \in \mathfrak{g}^*$ is a regular value of $\mathbf{J}$ and that the action of G_μ on $\mathbf{J}^{-1}(\mu)$ is free and proper

From the previous discussion, note that condition **SRFree** implies condition **SRRegular**. The real difference is that **SRRegular** assumes local freeness of the action of G (which is equivalent to μ being a regular value, as we have seen), while **SRFree** assumes global freeness (on all of P).

Theorem 3 (Symplectic reduction theorem) *Assume that condition **SR** and that either the condition **SRFree** or the condition **SRRegular** holds. Then P_μ is a symplectic manifold, and is equipped with the **reduced symplectic form** Ω_μ that is uniquely characterized by the condition*

$$\pi_\mu^* \Omega_\mu = i_\mu^* \Omega , \tag{15}$$

where $\pi_\mu\colon \mathbf{J}^{-1}(\mu) \to P_\mu$ is the projection to the quotient space and where $i_\mu\colon \mathbf{J}^{-1}(\mu) \to P$ is the inclusion.

The above procedure is often called *point reduction* because one is fixing the value of the momentum map at a point $\mu \in \mathfrak{g}^*$. An equivalent reduction method called *orbit reduction* will be discussed shortly.

Coadjoint Orbits

A standard example (due to Marsden and Weinstein [110]) that we shall derive in detail in the next section, is the construction of the coadjoint orbits in $\mathfrak{g}^*$ of a group G by reduction of the cotangent bundle T^*G with its canonical symplectic structure and with G acting on T^*G by the cotangent lift of left (resp. right) group multiplication. In this case, one finds that $(T^*G)_\mu = \mathcal{O}_\mu$, the coadjoint orbit through $\mu \in \mathfrak{g}^*$. The reduced symplectic form is given by the ***Kostant, Kirillov, Souriau coadjoint form***, also referred to as the ***KKS form***:

$$\omega^{\mp}_{\mathcal{O}_\mu}(\nu)(\mathrm{ad}^*_\xi \nu, \mathrm{ad}^*_\eta \nu) = \mp\langle \nu, [\xi, \eta]\rangle , \tag{16}$$

where $\xi, \eta \in \mathfrak{g}$, $\nu \in \mathcal{O}_\mu$, $\mathrm{ad}_\xi\colon \mathfrak{g} \to \mathfrak{g}$ is the adjoint operator defined by $\mathrm{ad}_\xi \eta := [\xi, \eta]$ and $\mathrm{ad}^*_\xi\colon \mathfrak{g}^* \to \mathfrak{g}^*$ is its dual. In this formula, one uses the minus sign for the left action and the plus sign for the right action. We recall that coadjoint orbits, like any group orbit is always an immersed manifold. Thus, one arrives at the following result (see also Theorem 7):

Corollary 4 *Given a Lie group G with Lie algebra $\mathfrak{g}$ and any point $\mu \in \mathfrak{g}^*$, the reduced space $(T^*G)_\mu$ is the coad-*

joint orbit $\mathcal{O}_\mu$ through the point μ; it is a symplectic manifold with symplectic form given by (16).

This example, which "explains" Kostant, Kirillov and Souriau's formula for this structure, is typical of many of the ensuing applications, in which the reduction procedure is applied to a "trivial" symplectic manifold to produce something interesting.

Orbit Reduction

An important variant of the symplectic reduction theorem is called ***orbit reduction*** and, roughly speaking, it constructs $\mathbf{J}^{-1}(\mathcal{O})/G$, where $\mathcal{O}$ is a coadjoint orbit in $\mathfrak{g}^*$. In the next section – see Theorem 8 – we show that orbit reduction is equivalent to the point reduction considered above.

Cotangent Bundle Reduction

The theory of cotangent bundle reduction is a very important special case of general reduction theory. Notice that the reduction of T^*G above to give a coadjoint orbit is a special case of the more general procedure in which G is replaced by a configuration manifold Q. The theory of cotangent bundle reduction will be outlined in the historical overview in this chapter, and then treated in some detail in the following chapter.

Mathematical Physics Links

Another example in Marsden and Weinstein [110] came from general relativity, namely the reduction of the cotangent bundle of the space of Riemannian metrics on a manifold M by the action of the group of diffeomorphisms of M. In this case, restriction to the zero momentum level is the divergence constraint of general relativity, and so one is led to a construction of a symplectic structure on a space closely related to the space of solutions of the Einstein equations, a question revisited in Fischer, Marsden, and Moncrief [51] and Arms, Marsden, and Moncrief [6]. Here one sees a precursor of an idea of Atiyahf and Bott [11], which has led to some of the most spectacular applications of reduction in mathematical physics and related areas of pure mathematics, especially low-dimensional topology.

Singular Reduction

In the preceding discussion, we have been making hypotheses that ensure the momentum levels and their quotients are smooth manifolds. Of course, this is not always the case, as was already noted in Smale [168] and analyzed (even in the infinite-dimensional case) in Arms, Marsden, and Moncrief [5]. We give a review of some of the current literature and history on this singular case in Sect. "Reduction Theory: Historical Overview". For an outline of this subject, see [142] and for a complete account of the technical details, see [138].

Reduction of Dynamics

Along with the geometry of reduction, there is also a theory of *reduction of dynamics*. The main idea is that a G-invariant Hamiltonian H on P induces a Hamiltonian H_μ on each of the reduced spaces, and the corresponding Hamiltonian vector fields X_H and X_{H_μ} are π_μ-related. The reverse of reduction is reconstruction and this leads one to the theory of classical geometric phases (Hannay–Berry phases); see Marsden, Montgomery, and Ratiu [98].

Reduction theory has many interesting connections with the theory of integrable systems; we just mention some selected references Kazhdan, Kostant, and Sternberg [72]; Ratiu [154,155,156]; Bobenko, Reyman, and Semenov-Tian-Shansky [22]; Pedroni [148]; Marsden and Ratiu [103]; Vanhaecke [174]; Bloch, Crouch, Marsden, and Ratiu [19], which the reader can consult for further information.

Symplectic Reduction – Further Discussion

The symplectic reduction theorem leans on a few key lemmas that we just state. The first refers to the reflexivity of the operation of taking the symplectic orthogonal complement.

Lemma 5 *Let (V, Ω) be a finite dimensional symplectic vector space and $W \subset V$ be a subspace. Define the* ***symplectic orthogonal*** *to W by*

$$W^\Omega = \{v \in V \mid \Omega(v, w) = 0 \textit{ for all } w \in W\} .$$

Then

$$\left(W^\Omega\right)^\Omega = W . \tag{17}$$

In what follows, we denote by $G \cdot z$ and $G_\mu \cdot z$ the G and G_μ-orbits through the point $z \in P$; note that if $z \in \mathbf{J}^{-1}(\mu)$ then $G_\mu \cdot z \subset \mathbf{J}^{-1}(\mu)$.

The key lemma that is central for the symplectic reduction theorem is the following.

Lemma 6 (Reduction lemma) *Let P be a Poisson manifold and let $\mathbf{J}\colon P \to \mathfrak{g}^*$ be an equivariant momentum map of a Lie group action by Poisson maps of G on P. Let $G \cdot \mu$ denote the coadjoint orbit through a regular value $\mu \in \mathfrak{g}^*$ of $\mathbf{J}$. Then*

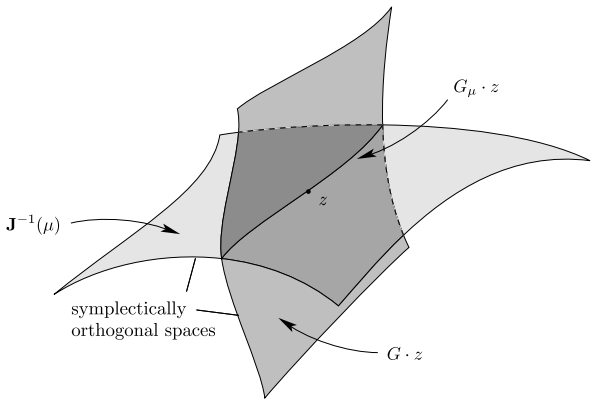

Mechanical Systems: Symmetries and Reduction, Figure 1
The geometry of the reduction lemma

(i) $\mathbf{J}^{-1}(G \cdot \mu) = G \cdot \mathbf{J}^{-1}(\mu) = \{g \cdot z \mid g \in G$ *and* $\mathbf{J}(z) = \mu\}$;
(ii) $G_\mu \cdot z = (G \cdot z) \cap \mathbf{J}^{-1}(\mu)$;
(iii) $\mathbf{J}^{-1}(\mu)$ *and* $G \cdot z$ ***intersect cleanly****, i. e.,*

$$T_z(G_\mu \cdot z) = T_z(G \cdot z) \cap T_z(\mathbf{J}^{-1}(\mu));$$

(iv) if (P, Ω) *is symplectic, then* $T_z(\mathbf{J}^{-1}(\mu)) = (T_z(G \cdot z))^\Omega$*; i. e., the sets*

$$T_z(\mathbf{J}^{-1}(\mu)) \text{ and } T_z(G \cdot z)$$

are Ω*-orthogonal complements of each other.*

Refer to Fig. 1 for one way of visualizing the geometry associated with the reduction lemma. As it suggests, the two manifolds $\mathbf{J}^{-1}(\mu)$ and $G \cdot z$ intersect in the orbit of the isotropy group $G_\mu \cdot z$ and their tangent spaces $T_z\mathbf{J}^{-1}(\mu)$ and $T_z(G \cdot z)$ are symplectically orthogonal and intersect in the space $T_z\left(G_\mu \cdot z\right)$. Notice from the statement (iv) that $T_z(\mathbf{J}^{-1}(\mu))^\Omega \subset T_z(\mathbf{J}^{-1}(\mu))$ provided that $G_\mu \cdot z = G \cdot z$. Thus, $\mathbf{J}^{-1}(\mu)$ *is* coisotropic *if* $G_\mu = G$; for example, this happens if $\mu = 0$ or if G is Abelian.

Remarks on the Reduction Theorem

1. Even if Ω is exact; say $\Omega = -\mathbf{d}\Theta$ and the action of G leaves Θ invariant, Ω_μ *need not be exact*. Perhaps the simplest example is a nontrivial coadjoint orbit of SO(3), which is a sphere with symplectic form given by the area form (by Stokes' theorem, it cannot be exact). That this is a symplectic reduced space of $T^*\mathrm{SO}(3)$ (with the canonical symplectic structure, so is exact) is shown in Theorem 7 below.
2. Continuing with the previous remark, assume that $\Omega = -\mathbf{d}\Theta$ and that the G_μ principal bundle $\mathbf{J}^{-1}(\mu) \to P_\mu := \mathbf{J}^{-1}(\mu)/G_\mu$ is trivializable; that is, it admits a *global* section $s\colon P_\mu \to \mathbf{J}^{-1}(\mu)$. Let $\Theta_\mu := s^* i_\mu^* \Theta \in \Omega^1(P_\mu)$. Then the reduced symplectic form $\Omega_\mu = -\mathbf{d}\Theta_\mu$ is exact. This statement *does not imply that the one-form* Θ *descends to the reduced space*, only that the reduced symplectic form is exact and one if its primitives is Θ_μ. In fact, if one changes the global section, another primitive of Ω_μ is found which differs from Θ_μ by a closed one-form on P_μ.
3. The assumption that μ is a regular value of $\mathbf{J}$ can be relaxed. *The only hypothesis needed is that* μ *be a clean value of* $\mathbf{J}$, i. e., $\mathbf{J}^{-1}(\mu)$ is a manifold and $T_z(\mathbf{J}^{-1}(\mu)) = \ker T_z\mathbf{J}$. This generalization applies, for instance, for zero angular momentum in the three dimensional two body problem, as was noted by Marsden and Weinstein [110] and Kazhdan, Kostant, and Sternberg [72]; see also Guillemin and Sternberg [57].
   Here are the general definitions: If $f\colon M \to N$ is a smooth map, a point $y \in N$ is called a ***clean value*** if $f^{-1}(y)$ is a submanifold and for each $x \in f^{-1}(y)$, $T_x f^{-1}(y) = \ker T_x f$. We say that f intersects a submanifold $L \subset N$ ***cleanly*** if $f^{-1}(L)$ is a submanifold of M and $T_x(f^{-1}(L)) = (T_x f)^{-1}(T_{f(x)}L)$. Note that *regular values of* f *are clean values and that if* f *intersects the submanifold* L *transversally, then it intersects it cleanly*. Also note that the definition of clean intersection of two manifolds is equivalent to the statement that the inclusion map of either one of them intersects the other cleanly. The reduction lemma is an example of this situation.
4. The freeness and properness of the G_μ action on $\mathbf{J}^{-1}(\mu)$ are used only to guarantee that P_μ is a manifold; these hypotheses can thus be replaced by the requirement that P_μ is a manifold and $\pi_\mu\colon \mathbf{J}^{-1}(\mu) \to P_\mu$ a submersion; the proof of the symplectic reduction theorem remains unchanged.
5. Even if μ is a regular value (in the sense of a regular value of the mapping $\mathbf{J}$), it need not be a ***regular point*** (also called a ***generic point***) in $\mathfrak{g}^*$; that is, a point whose coadjoint orbit is of maximal dimension. The reduction theorem *does not require that* μ *be a regular point*. For example, if G acts on itself on the left by group multiplication and if we lift this to an action on T^*G by the cotangent lift, then the action is free and so all μ are regular values, but such values (for instance, the zero element in $\mathfrak{so}(3)^*$) need not be regular. On the other hand, in many important stability considerations, a regularity assumption on the point μ is required; see, for instance, Patrick [145], Ortega and Ratiu [136] and Patrick, Roberts, and Wulff [146].

Nonequivariant Reduction

We now describe how one can carry out reduction for a *nonequivariant momentum map*.

If $\mathbf{J}\colon P \to \mathfrak{g}^*$ is a nonequivariant momentum map on the connected symplectic manifold P with nonequivariance group one-cocycle σ consider the affine action (9) and let $\widetilde{G}_\mu$ be the isotropy subgroup of $\mu \in \mathfrak{g}^*$ relative to this action. Then, under the same regularity assumptions (for example, assume that G acts freely and properly on P, or that μ is a regular value of $\mathbf{J}$ and that $\widetilde{G}_\mu$ acts freely and properly on $\mathbf{J}^{-1}(\mu)$), the quotient manifold $P_\mu := \mathbf{J}^{-1}(\mu)/\widetilde{G}_\mu$ is a symplectic manifold whose symplectic form is uniquely determined by the relation $i_\mu^*\Omega = \pi_\mu^*\Omega_\mu$. The proof of this statement is identical to the one given above with the obvious changes in the meaning of the symbols.

When using nonequivariant reduction, one has to remember that G acts on $\mathfrak{g}^*$ in an *affine* and not a linear manner. For example, while the coadjoint isotropy subgroup at the origin is equal to G; that is, $G_0 = G$, this is no longer the case for the affine action, where $\widetilde{G}_0$ in general does not equal G.

Coadjoint Orbits as Symplectic Reduced Spaces

We now examine Corollary 4 – that is, that coadjoint orbits may be realized as reduced spaces – a little more closely. Realizing them as reduced spaces shows that they are symplectic manifolds See, Chap. 14 in [104] for a "direct" or "bare hands" argument. Historically, a direct argument was found first, by Kirillov, Kostant and Souriau in the early 1960's and the (minus) coadjoint symplectic structure was found to be

$$\omega_\nu^-(\mathrm{ad}_\xi^*\nu, \mathrm{ad}_\eta^*\nu) = -\langle \nu, [\xi, \eta]\rangle \tag{18}$$

Interestingly, this is the symplectic structure on the symplectic leaves of the Lie–Poisson bracket, as is shown in, for example, [104]. (See the historical overview in Sect. "Reduction Theory: Historical Overview" below and specifically, see Eq. (21) for a quick review of the Lie–Poisson bracket).

The strategy of the reduction proof, as mentioned in the discussion in the last section, is to show that the coadjoint symplectic form on a coadjoint orbit $\mathcal{O}_\mu$ of the point μ, at a point $\nu \in \mathcal{O}$, may be obtained by symplectically reducing T^*G at the value μ. The following theorem (due to Marsden and Weinstein [110]), and which is an elaboration on the result in Corollary 4, formulates the result for left actions; of course there is a similar one for right actions, with the minus sign replaced by a plus sign.

Theorem 7 (Reduction to coadjoint orbits) *Let G be a Lie group and let G act on G (and hence on T^*G by cotangent lift) by left multiplication. Let $\mu \in \mathfrak{g}^*$ and let $\mathbf{J}_L\colon T^*G \to \mathfrak{g}^*$ be the momentum map for the left action. Then μ is a regular value of $\mathbf{J}_L$, the action of G is free and proper, the symplectic reduced space $\mathbf{J}_L^{-1}(\mu)/G_\mu$ is identified via left translation with $\mathcal{O}_\mu$, the coadjoint orbit through μ, and the reduced symplectic form coincides with ω^- given in Eq. (18).*

Remarks

1. Notice that, as in the general Symplectic Reduction Theorem 3, this result does *not* require μ to be a regular (or generic) *point* in $\mathfrak{g}^*$; that is, arbitrarily nearby coadjoint orbits may have a different dimension.
2. The form ω^- on the orbit need not be exact even though Ω is. An example that shows this is SO(3), whose coadjoint orbits are spheres and whose symplectic structure is, as shown in [104], a multiple of the area element, which is not exact by Stokes' Theorem.

Orbit Reduction

So far, we have presented what is usually called *point reduction*. There is another point of view that is called *orbit reduction*, which we now summarize. We assume the same set up as in the symplectic reduction theorem, with P connected, G acting symplectically, freely, and properly on P with an equivariant momentum map $\mathbf{J}\colon P \to \mathfrak{g}^*$.

The connected components of the point reduced spaces P_μ can be regarded as the symplectic leaves of the Poisson manifold $(P/G, \{\cdot,\cdot\}_{P/G})$ in the following way. Form a map $[i_\mu]\colon P_\mu \to P/G$ defined by selecting an equivalence class $[z]_{G_\mu}$ for $z \in \mathbf{J}^{-1}(\mu)$ and sending it to the class $[z]_G$. This map is checked to be well-defined and smooth. We then have the commutative diagram

$$\begin{array}{ccc} \mathbf{J}^{-1}(\mu) & \xrightarrow{\ i_\mu\ } & P \\ \Big\downarrow \pi_\mu & & \Big\downarrow \pi \\ P_\mu & \xrightarrow{\ [i_\mu]\ } & P/G \end{array}$$

One then checks that $[i_\mu]$ is a Poisson injective immersion. Moreover, the $[i_\mu]$-images in P/G of the connected components of the symplectic manifolds (P_μ, Ω_μ) are its symplectic leaves (see [138] and references therein

for details). As sets,

$$\left[i_\mu\right]\left(P_\mu\right) = \mathbf{J}^{-1}\left(\mathcal{O}_\mu\right)/G\,,$$

where $\mathcal{O}_\mu \subset \mathfrak{g}^*$ is the coadjoint orbit through $\mu \in \mathfrak{g}^*$. The set

$$P_{\mathcal{O}_\mu} := \mathbf{J}^{-1}\left(\mathcal{O}_\mu\right)/G$$

is called the ***orbit reduced space*** associated to the orbit $\mathcal{O}_\mu$. The smooth manifold structure (and hence the topology) on $P_{\mathcal{O}_\mu}$ is the one that makes the map $\left[i_\mu\right]: P_\mu \to P_{\mathcal{O}_\mu}$ into a diffeomorphism.

For the next theorem, which characterizes the symplectic form and the Hamiltonian dynamics on $P_{\mathcal{O}_\mu}$, recall the coadjoint orbit symplectic structure of Kirillov, Kostant and Souriau that was established in the preceding Theorem 7:

$$\omega^-_{\mathcal{O}_\mu}(\nu)(\xi_{\mathfrak{g}^*}(\nu),\ \eta_{\mathfrak{g}^*}(\nu)) = -\langle \nu,\ [\xi,\ \eta]\rangle\,, \tag{19}$$

for $\xi, \eta \in \mathfrak{g}$ and $\nu \in \mathcal{O}_\mu$.

We also recall that an injectively immersed submanifold of S of Q is called an ***initial submanifold*** of Q when for any smooth manifold P, a map $g: P \to S$ is smooth if and only if $\iota \circ g: P \to Q$ is smooth, where $\iota: S \hookrightarrow Q$ is the inclusion.

Theorem 8 (Symplectic orbit reduction theorem) *In the setup explained above:*

(i) *The momentum map $\mathbf{J}$ is transverse to the coadjoint orbit $\mathcal{O}_\mu$ and hence $\mathbf{J}^{-1}(\mathcal{O}_\mu)$ is an initial submanifold of P. Moreover, the projection $\pi_{\mathcal{O}_\mu}: \mathbf{J}^{-1}\left(\mathcal{O}_\mu\right) \to P_{\mathcal{O}_\mu}$ is a surjective submersion.*

(ii) *$P_{\mathcal{O}_\mu}$ is a symplectic manifold with the symplectic form $\Omega_{\mathcal{O}\mu}$ uniquely characterized by the relation*

$$\pi^*_{\mathcal{O}_\mu}\Omega_{\mathcal{O}\mu} = \mathbf{J}^*_{\mathcal{O}_\mu}\omega^-_{\mathcal{O}_\mu} + i^*_{\mathcal{O}_\mu}\Omega\,, \tag{20}$$

where $\mathbf{J}_{\mathcal{O}_\mu}$ is the restriction of $\mathbf{J}$ to $\mathbf{J}^{-1}\left(\mathcal{O}_\mu\right)$ and $i_{\mathcal{O}_\mu}: \mathbf{J}^{-1}\left(\mathcal{O}_\mu\right) \hookrightarrow P$ is the inclusion.

(iii) *The map $\left[i_\mu\right]: P_\mu \to P_{\mathcal{O}_\mu}$ is a symplectic diffeomorphism.*

(iv) *(Dynamics). Let H be a G-invariant function on P and define $\widetilde{H}: P/G \to \mathbb{R}$ by $H = \widetilde{H} \circ \pi$. Then the Hamiltonian vector field X_H is also G-invariant and hence induces a vector field on P/G, which coincides with the Hamiltonian vector field $X_{\widetilde{H}}$. Moreover, the flow of $X_{\widetilde{H}}$ leaves the symplectic leaves $P_{\mathcal{O}_\mu}$ of P/G invariant. This flow restricted to the symplectic leaves is again Hamiltonian relative to the symplectic form $\Omega_{\mathcal{O}\mu}$ and the Hamiltonian function $\widetilde{H}_{\mathcal{O}_\leq}$ given by*

$$\widetilde{H}_{\mathcal{O}_\mu} \circ \pi_{\mathcal{O}_\mu} = H \circ i_{\mathcal{O}_\mu}\,.$$

Note that if $\mathcal{O}_\mu$ is an embedded submanifold of $\mathfrak{g}^*$ then $\mathbf{J}$ is transverse to $\mathcal{O}_\mu$ and hence $\mathbf{J}^{-1}(\mathcal{O}_\mu)$ is automatically an embedded submanifold of P.

The proof of this theorem when $\mathcal{O}_\mu$ is an embedded submanifold of $\mathfrak{g}^*$ can be found in Marle [91], Kazhdan, Kostant, and Sternberg [72], with useful additions given in Marsden [94] and Blaom [17]. For nonfree actions and when $\mathcal{O}_\mu$ is not an embedded submanifold of $\mathfrak{g}^*$ see [138]. Further comments on the historical context of this result are given in the next section.

Remarks

1. A similar result holds for right actions.
2. Freeness and properness of the G_μ-action on $\mathbf{J}^{-1}(\mu)$ are only needed indirectly. In fact these conditions are sufficient but not necessary for P_μ to be a manifold. All that is needed is for P_μ to be a manifold and π_μ to be a submersion and the above result remains unchanged.
3. Note that the description of the symplectic structure on $\mathbf{J}^{-1}(\mathcal{O})/G$ is not as simple as it was for $\mathbf{J}^{-1}(\mu)/G$, while the Poisson bracket description is simpler on $\mathbf{J}^{-1}(\mathcal{O})/G$. Of course, the symplectic structure depends only on the orbit $\mathcal{O}$ and not on the choice of a point μ on it.

Cotangent Bundle Reduction

Perhaps the most important and basic reduction theorem in addition to those already presented is the cotangent bundle reduction theorem. We shall give an exposition of the key aspects of this theory in Sect. "Cotangent Bundle Reduction" and give a historical account of its development, along with references in the next section.

At this point, to orient the reader, we note that one of the special cases is cotangent bundle reduction at zero (see Theorem 10). This result says that if one has, again for simplicity, a free and proper action of G on Q (which is then lifted to T^*Q by the cotangent lift), then the reduced space at zero of T^*Q is given by $T^*(Q/G)$, with its canonical symplectic structure. On the other hand, reduction at a nonzero value is a bit more complicated and gives rise to modifications of the standard symplectic structure; namely, one adds to the canonical structure, the pull-back of a closed two form on Q to T^*Q. Because of their physical interpretation (discussed, for example, in [104]), such extra terms are called magnetic terms. In Sect. "Cotangent Bundle Reduction", we state the basic cotangent bundle reduction theorems along with providing some of the other important notions, such as the mechanical connection and the locked inertia tensor. Other notions that are important

in mechanics, such as the amended potential, can be found in [95].

Reduction Theory: Historical Overview

We have already given bits an pieces of the history of symplectic reduction and momentum maps. In this section we take a broader view of the subject to put things in historical and topical context.

History before 1960

In the preceding sections, reduction theory has been presented as a mathematical construction. Of course, these ideas are rooted in classical work on mechanical systems with symmetry by such masters as Euler, Lagrange, Hamilton, Jacobi, Routh, Riemann, Liouville, Lie, and Poincaré. The aim of their work was, to a large extent, to eliminate variables associated with symmetries in order to simplify calculations in concrete examples. Much of this work was done using coordinates, although the deep connection between mechanics and geometry was already evident. Whittaker [179] gives a good picture of the theory as it existed up to about 1910.

A highlight of this early theory was the work of Routh [161,163] who studied reduction of systems with cyclic variables and introduced the amended potential for the reduced system for the purpose of studying, for instance, the *stability of a uniformly rotating state* – what we would call today a *relative equilibrium*, terminology introduced later by Poincaré. Smale [168] eventually put the amended potential into a nice geometric setting. Routh's work was closely related to the reduction of systems with integrals in involution studied by Jacobi and Liouville around 1870; the Routh method corresponds to the modern theory of Lagrangian reduction for the action of Abelian groups.

The rigid body, whose equations were discovered by Euler around 1740, was a key example of reduction – what we would call today either reduction to coadjoint orbits or Lie–Poisson reduction on the Hamiltonian side, or Euler–Poincaré reduction on the Lagrangian side, depending on one's point of view. Lagrange [81] already understood reduction of the rigid body equations by a method not so far from what one would do today with the symmetry group SO(3).

Many later authors, unfortunately, relied so much on coordinates (especially Euler angles) that there is little mention of SO(3) in classical mechanics books written before 1990, which by today's standards, seems rather surprising! In addition, there seemed to be little appreciation until recently for the role of topological notions; for example, the fact that one cannot globally split off cyclic variables for the S^1 action on the configuration space of the heavy top. The Hopf fibration was patiently waiting to be discovered in the reduction theory for the classical rigid body, but it was only explicitly found later on by H. Hopf [64]. Hopf was, apparently, unaware that this example is of great mechanical interest – the gap between workers in mechanics and geometers seems to have been particularly wide at that time.

Another noteworthy instance of reduction is Jacobi's elimination of the node for reducing the gravitational (or electrostatic) n-body problem by means of the group SE(3) of Euclidean motions, around 1860 or so. This example has, of course, been a mainstay of celestial mechanics. It is related to the work done by Riemann, Jacobi, Poincaré and others on rotating fluid masses held together by gravitational forces, such as stars. Hidden in these examples is much of the beauty of modern reduction, stability and bifurcation theory for mechanical systems with symmetry.

While both symplectic and Poisson geometry have their roots in the work of Lagrange and Jacobi, it matured considerably with the work of Lie [90], who discovered many remarkably modern concepts such as the Lie–Poisson bracket on the dual of a Lie algebra. See Weinstein [176] and Marsden and Ratiu [104] for more details on the history. How Lie could have viewed his wonderful discoveries so divorced from their roots in mechanics remains a mystery. We can only guess that he was inspired by Jacobi, Lagrange and Riemann and then, as mathematicians often do, he quickly abstracted the ideas, losing valuable scientific and historical connections along the way.

As we have already hinted, it was the famous paper Poincaré [153] where we find what we call today the Euler–Poincaré equations – a generalization of the Euler equations for both fluids and the rigid body to general Lie algebras. (The Euler–Poincaré equations are treated in detail in [104]). It is curious that Poincaré did not stress either the symplectic ideas of Lie, nor the variational principles of mechanics of Lagrange and Hamilton – in fact, it is not clear to what extent he understood what we would call today Euler–Poincaré reduction. It was only with the development and physical application of the notion of a manifold, pioneered by Lie, Poincaré, Weyl, Cartan, Reeb, Synge and many others, that a more general and intrinsic view of mechanics was possible. By the late 1950's, the stage was set for an explosion in the field.

1960–1972

Beginning in the 1960's, the subject of geometric mechanics indeed did explode with the basic contributions of peo-

ple such as (alphabetically and nonexhaustively) Abraham, Arnold, Kirillov, Kostant, Mackey, MacLane, Segal, Sternberg, Smale, and Souriau. Kirillov and Kostant found deep connections between mechanics and pure mathematics in their work on the orbit method in group representations, while Arnold, Smale, and Souriau were in closer touch with mechanics.

The modern vision of geometric mechanics combines strong links to important questions in mathematics with the traditional classical mechanics of particles, rigid bodies, fields, fluids, plasmas, and elastic solids, as well as quantum and relativistic theories. Symmetries in these theories vary from obvious translational and rotational symmetries to less obvious particle relabeling symmetries in fluids and plasmas, to the "hidden" symmetries underlying integrable systems. As we have already mentioned, reduction theory concerns the removal of variables using symmetries and their associated conservation laws. Variational principles, in addition to symplectic and Poisson geometry, provide fundamental tools for this endeavor. In fact, conservation of the momentum map associated with a symmetry group action is a geometric expression of the classical Noether theorem (discovered by variational, not symplectic methods).

Arnold and Smale

The modern era of reduction theory began with the fundamental papers of Arnold [7] and Smale [168]. Arnold focused on systems whose configuration manifold is a Lie group, while Smale focused on bifurcations of relative equilibria. Both Arnold and Smale linked their theory strongly with examples. For Arnold, they were the same examples as for Poincaré, namely the rigid body and fluids, for which he went on to develop powerful stability methods, as in Arnold [8].

With hindsight, we can say that Arnold [7] was picking up on the basic work of Poincaré for both rigid body motion and fluids. In the case of fluids, G is the group of (volume preserving) diffeomorphisms of a compact manifold (possibly with boundary). In this setting, one obtains the Euler equations for (incompressible) fluids by reduction from the Lagrangian formulation of the equations of motion, an idea exploited by Arnold [7] and Ebin and Marsden [49]. This sort of description of a fluid goes back to Poincaré (using the Euler–Poincaré equations) and to the thesis of Ehrenfest (as geodesics on the diffeomorphism group), written under the direction of Boltzmann.

For Smale, the motivating example was celestial mechanics, especially the study of the number and stability of relative equilibria by a topological study of the energy-momentum mapping. He gave an intrinsic geometric account of the amended potential and in doing so, discovered what later became known as the mechanical connection. (Smale appears to not to have recognized that the interesting object he called α is, in fact, a *principal connection*; this was first observed by Kummer [79]). One of Smale's key ideas in studying relative equilibria was to link mechanics with topology via the fact that relative equilibria are critical points of the amended potential.

Besides giving a beautiful exposition of the momentum map, Smale also emphasized the connection between singularities and symmetry, observing that the symmetry group of a phase space point has positive dimension if and only if that point is not a regular point of the momentum map restricted to a fiber of the cotangent bundle (Smale [168], Proposition 6.2) – a result we have proved in Proposition 2. He went on from here to develop his topology and mechanics program and to apply it to the planar n-body problem. The topology and mechanics program definitely involved reduction ideas, as in Smale's construction of the quotients of integral manifolds, as in $I_{c,p}/S^1$ (Smale [168], page 320). He also understood Jacobi's elimination of the node in this context, although he did not attempt to give any general theory of reduction along these lines.

Smale thereby set the stage for symplectic reduction: he realized the importance of the momentum map and of quotient constructions, and he worked out explicit examples like the planar n-body problem with its S^1 symmetry group. (Interestingly, he pointed out that one should really use the nonabelian group SE(2); his feeling of unease with fixing the center of mass of an n-body system is remarkably perceptive).

Synthesis

The problem of synthesizing the Lie algebra reduction methods of Arnold [7] with the techniques of Smale [168] on the reduction of cotangent bundles by Abelian groups, led to the development of reduction theory in the general context of symplectic manifolds and equivariant momentum maps in Marsden and Weinstein [110] and Meyer [117], as we described in the last section. Both of these papers were completed by 1972.

Poisson Manifolds

Meanwhile, things were also gestating from the viewpoint of Poisson brackets and the idea of a Poisson manifold was being initiated and developed, with much duplication and rediscovery (see, Section 10.1 in [104] for additional information).

A basic example of a noncanonical Poisson bracket is the Lie–Poisson bracket on $\mathfrak{g}^*$, the dual of a Lie algebra $\mathfrak{g}$. This bracket (which comes with a plus or minus sign) is given on two smooth functions on $\mathfrak{g}^*$ by

$$\{f, g\}_{\pm}(\mu) = \pm\left\langle \mu, \left[\frac{\delta f}{\delta \mu}, \frac{\delta g}{\delta \mu}\right]\right\rangle , \tag{21}$$

where $\delta f/\delta\mu$ is the derivative of f, but thought of as an element of $\mathfrak{g}$. These Poisson structures, including the coadjoint orbits as their symplectic leaves, were known to Lie [90], although, as we mentioned previously, Lie does not seem to have recognized their importance in mechanics. It is also not clear whether or not Lie realized that the Lie Poisson bracket is the Poisson reduction of the canonical Poisson bracket on T^*G by the action of G. (See, Chap. 13 in [104] for an account of this theory). The first place we know of that has this clearly stated (but with no references, and no discussion of the context) is Bourbaki [24], Chapter III, Section 4, Exercise 6. Remarkably, this exercise also contains an interesting proof of the Duflo–Vergne theorem (with no reference to the original paper, which appeared in 1969). Again, any hint of links with mechanics is missing.

This takes us up to about 1972.

Post 1972

An important contribution was made by Marle [91], who divides the inverse image of an orbit by its characteristic foliation to obtain the product of an orbit and a reduced manifold. In particular, as we saw in Theorem 8, P_μ is symplectically diffeomorphic to an "orbit-reduced" space $P_\mu \cong J^{-1}(\mathcal{O}_\mu)/G$, where $\mathcal{O}_\mu$ is a coadjoint orbit of G. From this it follows that the P_μ are symplectic leaves in the Poisson space P/G. The related paper of Kazhdan, Kostant, and Sternberg [72] was one of the first to notice deep links between reduction and integrable systems. In particular, they found that the Calogero–Moser systems could be obtained by reducing a system that was trivially integrable; in this way, reduction provided a method of producing an interesting integrable system from a simple one. This point of view was used again by, for example, Bobenko, Reyman, and Semenov–Tian–Shansky [22] in their spectacular group theoretic explanation of the integrability of the Kowalewski top.

Noncanonical Poisson Brackets

The Hamiltonian description of many physical systems, such as rigid bodies and fluids in Eulerian variables, requires noncanonical Poisson brackets and constrained variational principles of the sort studied by Lie and Poincaré. As discussed above, a basic example of a noncanonical Poisson bracket is the Lie–Poisson bracket on the dual of a Lie algebra. From the mechanics perspective, the remarkably modern book (but which was, unfortunately, rather out of touch with the corresponding mathematical developments) by Sudarshan and Mukunda [172] showed via explicit examples how systems such as the rigid body could be written in terms of noncanonical brackets, an idea going back to Pauli [147], Martin [115] and Nambu [129]. Others in the physics community, such as Morrison and Greene [128] also discovered noncanonical bracket formalisms for fluid and magnetohydrodynamic systems. In the 1980's, many fluid and plasma systems were shown to have a noncanonical Poisson formulation. It was Marsden and Weinstein [111,112] who first applied reduction techniques to these systems.

The *reduction philosophy* concerning noncanonical brackets can be summarized by saying

> *Any mechanical system has its roots somewhere as a cotangent bundle and one can recover noncanonical brackets by the simple process of Poisson reduction. For example, in fluid mechanics, this reduction is implemented by the Lagrange-to-Euler map.*

This view ran contrary to the point of view, taken by some researchers, that one should proceed by analogy or guesswork to find Poisson structures and then to try to limit the guesses by the constraint of Jacobi's identity.

In the simplest version of the Poisson reduction process, one starts with a Poisson manifold P on which a group G acts by Poisson maps and then forms the quotient space P/G, which, if not singular, inherits a natural Poisson structure itself. Of course, the Lie–Poisson structure on $\mathfrak{g}^*$ is inherited in exactly this way from the canonical symplectic structure on T^*G. One of the attractions of this Poisson bracket formalism was its use in stability theory. This literature is now very large, but Holm, Marsden, Ratiu, and Weinstein [63] is representative.

The way in which the Poisson structure on P_μ is related to that on P/G was clarified in a generalization of Poisson reduction due to Marsden and Ratiu [103], a technique that has also proven useful in integrable systems (see, e. g., Pedroni [148] and Vanhaecke [174]).

Reduction theory for mechanical systems with symmetry has proven to be a powerful tool that has enabled key advances in stability theory (from the Arnold method to the energy-momentum method for relative equilibria) as well as in bifurcation theory of mechanical systems, geometric phases via reconstruction – the inverse of reduction – as well as uses in control theory from sta-

bilization results to a deeper understanding of *locomotion*. For a general introduction to some of these ideas and for further references, see Marsden, Montgomery, and Ratiu [98]; Simo, Lewis, and Marsden [166]; Marsden and Ostrowski [99]; Marsden and Ratiu [104]; Montgomery [122,123,124,125,126]; Blaom [16,17] and Kanso, Marsden, Rowley, and Melli-Huber [71].

Tangent and Cotangent Bundle Reduction

The simplest case of cotangent bundle reduction is the case of reduction of $P = T^*Q$ at $\mu = 0$; the answer is simply $P_0 = T^*(Q/G)$ with the canonical symplectic form. Another basic case is when G is Abelian. Here, $(T^*Q)_\mu \cong T^*(Q/G)$, but the latter has a symplectic structure modified by magnetic terms, that is, by the curvature of the mechanical connection.

An Abelian version of cotangent bundle reduction was developed by Smale [168]. Then Satzer [165] studied the relatively simple, but important case of cotangent bundle reduction at the zero value of the momentum map. The full generalization of cotangent bundle reduction for nonabelian groups at arbitrary values of the momentum map appears for the first time in Abraham and Marsden [1]. It was Kummer [79] who first interpreted this result in terms of a connection, now called the ***mechanical connection***. The geometry of this situation was used to great effect in, for example, Guichardet [54,66], Iwai [67], and Montgomery [120,123,124]. We give an account of cotangent bundle reduction theory in the following section.

The Gauge Theory Viewpoint

Tangent and cotangent bundle reduction evolved into what we now term as the "bundle picture" or the "gauge theory of mechanics". This picture was first developed by Montgomery, Marsden, and Ratiu [127] and Montgomery [120,121]. That work was motivated and influenced by the work of Sternberg [171] and Weinstein [175] on a "Yang–Mills construction" which is, in turn, motivated by Wong's equations, i. e., the equations for a particle moving in a Yang–Mills field. The main result of the bundle picture gives a structure to the quotient spaces $(T^*Q)/G$ and $(TQ)/G$ when G acts by the cotangent and tangent lifted actions. The symplectic leaves in this picture were analyzed by Zaalani [182], Cushman and Śniatycki [47] and Marsden and Perlmutter [102]. The work of Perlmutter and Ratiu [149] gives a unified study of the Poisson bracket on $(T^*Q)/G$ in both the Sternberg and Weinstein realizations of the quotient.

As mentioned earlier, we shall review some of the basics of cotangent bundle reduction theory in Sect. "Cotangent Bundle Reduction". Further information on this theory may be found in [1,95], and [92], as well as a number of the other references mentioned above.

Lagrangian Reduction

A key ingredient in Lagrangian reduction is the classical work of Poincaré [153] in which the Euler–Poincaré equations were introduced. Poincaré realized that the equations of fluids, free rigid bodies, and heavy tops could all be described in Lie algebraic terms in a beautiful way. The importance of these equations was realized by Hamel [58,59] and Chetayev [40], but to a large extent, the work of Poincaré lay dormant until it was revived in the Russian literature in the 1980's.

The more recent developments of Lagrangian reduction were motivated by attempts to understand the relation between reduction, variational principles and Clebsch variables in Cendra and Marsden [34] and Cendra, Ibort, and Marsden [33]. In Marsden and Scheurle [109] it was shown that, for matrix groups, one could view the Euler–Poincaré equations via the reduction of Hamilton's variational principle from TG to $\mathfrak{g}$. The work of Bloch, Krishnaprasad, Marsden and Ratiu [21] established the Euler–Poincaré variational structure for general Lie groups.

The paper of Marsden and Scheurle [109] also considered the case of more general configuration spaces Q on which a group G acts, which was motivated by both the Euler–Poincaré case as well as the work of Cendra and Marsden [34] and Cendra, Ibort, and Marsden [33]. The Euler–Poincaré equations correspond to the case $Q = G$. Related ideas stressing the groupoid point of view were given in Weinstein [177]. The resulting reduced equations were called the *reduced Euler–Lagrange equations*. This work is the Lagrangian analogue of Poisson reduction, in the sense that no momentum map constraint is imposed.

Lagrangian reduction proceeds in a way that is very much in the spirit of the gauge theoretic point of view of mechanical systems with symmetry. It starts with Hamilton's variational principle for a Lagrangian system on a configuration manifold Q and with a symmetry group G acting on Q. The idea is to drop this variational principle to the quotient Q/G to derive a reduced variational principle. This theory has its origins in specific examples such as fluid mechanics (see, for example, Arnold [7] and Bretherton [25]), while the systematic theory of Lagrangian reduction was begun in Marsden and Scheurle [109] and further developed in Cendra, Marsden, and Ratiu [35]. The latter reference also introduced a connection to realize the space $(TQ)/G$ as the fiber product $T(Q/G) \times \tilde{\mathfrak{g}}$ of $T(Q/G)$ with the associated bundle formed using the adjoint action

of G on $\mathfrak{g}$. The reduced equations associated to this construction are called the *Lagrange–Poincaré equations* and their geometry has been fairly well developed. Note that a G-invariant Lagrangian L on TQ induces a Lagrangian l on $(TQ)/G$.

Until recently, the Lagrangian side of the reduction story had lacked a general category that is the Lagrangian analogue of Poisson manifolds in which reduction can be repeated. One candidate is the category of Lie algebroids, as explained in Weinstein [177]. Another is that of *Lagrange–Poincaré bundles*, developed in Cendra, Marsden, and Ratiu [35]. Both have tangent bundles and Lie algebras as basic examples. The latter work also develops the Lagrangian analogue of reduction for central extensions and, as in the case of symplectic reduction by stages, cocycles and curvatures enter in a natural way.

This bundle picture and Lagrangian reduction has proven very useful in control and optimal control problems. For example, it was used in Chang, Bloch, Leonard, Marsden and Woolsey [38] to develop a Lagrangian and Hamiltonian reduction theory for controlled mechanical systems and in Koon and Marsden [76] to extend the falling cat theorem of Montgomery [123] to the case of nonholonomic systems as well as to nonzero values of the momentum map.

Finally we mention that the paper Cendra, Marsden, Pekarsky, and Ratiu [37] develops the reduction theory for Hamilton's *phase space principle* and the equations on the reduced space, along with a reduced variational principle, are developed and called *the Hamilton–Poincaré equations*. Even in the case $Q = G$, this collapses to an interesting variational principle for the Lie–Poisson equations on $\mathfrak{g}^*$.

Legendre Transformation

Of course the Lagrangian and Hamiltonian sides of the reduction story are linked by the Legendre transformation. This mapping descends at the appropriate points to give relations between the Lagrangian and the Hamiltonian sides of the theory. However, even in standard cases such as the heavy top, one must be careful with this approach, as is already explained in, for example, Holm, Marsden, and Ratiu [61]. For field theories, such as the Maxwell–Vlasov equations, this issues is also important, as explained in Cendra, Holm, Hoyle and Marsden [31] (see also Tulczyjew and Urbański [173]).

Nonabelian Routh Reduction

Routh reduction for Lagrangian systems, which goes back Routh [161,162,163] is classically associated with systems having cyclic variables (this is almost synonymous with having an Abelian symmetry group). Modern expositions of this classical theory can be found in Arnold, Koslov, and Neishtadt [10] and in [104], §8.9. Routh Reduction may be thought of as the Lagrangian analog of symplectic reduction in that a momentum map is set equal to a constant. A key feature of Routh reduction is that when one drops the Euler–Lagrange equations to the quotient space associated with the symmetry, and when the momentum map is constrained to a specified value (i. e., when the cyclic variables and their velocities are eliminated using the given value of the momentum), then the resulting equations are in Euler–Lagrange form not with respect to the Lagrangian itself, but with respect to a modified function called the *Routhian*. Routh [162] applied his method to stability theory; this was a precursor to the energy-momentum method for stability that synthesizes Arnold's and Routh's methods (see Simo, Lewis and Marsden [166]). Routh's stability method is still widely used in mechanics.

The initial work on generalizing Routh reduction to the nonabelian case was that of Marsden and Scheurle [108]. This subject was further developed in Jalnapurkar and Marsden [69] and Marsden, Ratiu and Scheurle [105]. The latter reference used this theory to give some nice formulas for geometric phases from the Lagrangian point of view.

Semidirect Product Reduction

In the simplest case of a semidirect product, one has a Lie group G that acts on a vector space V (and hence on its dual V^*) and then one forms the semidirect product $S = G \circledS V$, generalizing the semidirect product structure of the Euclidean group $\mathrm{SE}(3) = \mathrm{SO}(3) \circledS \mathbb{R}^3$.

Consider the isotropy group G_{a_0} for some $a_0 \in V^*$. The semidirect product reduction theorem states that each of the symplectic reduced spaces for the action of G_{a_0} on T^*G is symplectically diffeomorphic to a coadjoint orbit in $(\mathfrak{g} \circledS V)^*$, the dual of the Lie algebra of the semidirect product. This semidirect product theory was developed by Guillemin and Sternberg [55,56], Ratiu [154,157,158], and Marsden, Ratiu, and Weinstein [106,107].

The Lagrangian reduction analog of semidirect product theory was developed by Holm, Marsden and Ratiu [61,62]. This construction is used in applications where one has advected quantities (such as the direction of gravity in the heavy top, density in compressible fluids and the magnetic field in MHD) as well as to geophysical flows. Cendra, Holm, Hoyle and Marsden [31] applied this idea to the Maxwell–Vlasov equations of plasma physics. Cendra, Holm, Marsden, and Ratiu [32] showed how La-

grangian semidirect product theory fits into the general framework of Lagrangian reduction.

The semidirect product reduction theorem has been proved in Landsman [82], Landsman [83], Chap. 4 as an application of a stages theorem for his *special symplectic reduction method*. Even though special symplectic reduction generalizes Marsden–Weinstein reduction, the special reduction by stages theorem in Landsman [82] studies a setup that, in general, is different to the ones in the reduction by stages theorems of [92].

Singular Reduction

Singular reduction starts with the observation of Smale [168] that we have already mentioned: $z \in P$ is a regular point of a momentum map $\mathbf{J}$ if and only if z has no continuous isotropy. Motivated by this, Arms, Marsden, and Moncrief [5,6] showed that (under hypotheses which include the ellipticity of certain operators and which can be interpreted more or less, as playing the role of a properness assumption on the group action in the finite dimensional case) the level sets $\mathbf{J}^{-1}(0)$ of an equivariant momentum map $\mathbf{J}$ have quadratic singularities at points with continuous symmetry. While such a result is easy to prove for compact group actions on finite dimensional manifolds (using the equivariant Darboux theorem), the main examples of Arms, Marsden, and Moncrief [5] were, in fact, infinite dimensional – both the phase space and the group. Singular points in the level sets of the momentum map are related to convexity properties of the momentum map in that the singular points in phase space map to corresponding singular points in the the image polytope.

The paper of Otto [143] showed that if G is a Lie group acting properly on an almost Kähler manifold then the orbit space $\mathbf{J}^{-1}(\mu)/G_\mu$ decomposes into symplectic smooth manifolds constructed out of the orbit types of the G-action on P. In some related work, Huebschmann [65] has made a careful study of the singularities of moduli spaces of flat connections.

The detailed structure of $\mathbf{J}^{-1}(0)/G$ for compact Lie groups acting on finite dimensional manifolds was determined by Sjamaar and Lerman [167]; their work was extended to proper Lie group actions and to $\mathbf{J}^{-1}(\mathcal{O}_\mu)/G$ by Bates and Lerman [12], with the assumption that $\mathcal{O}_\mu$ be locally closed in $\mathfrak{g}^*$. Ortega [130] and [138] redid the entire singular reduction theory for proper Lie group actions starting with the point reduced spaces $\mathbf{J}^{-1}(\mu)/G_\mu$ and also connected it to the more algebraic approach of Arms, Cushman, and Gotay [4]. Specific examples of singular reduction, with further references, may be found in Lerman, Montgomery, and Sjamaar [84] and [44]. One of these, the "canoe" is given in detail in [92]. In fact, this is an example of singular reduction in the case of cotangent bundles, and much more can be said in this case; see Olmos and Dias [150,151]. Another approach to singular reduction based on the technique of blowing up singularities, and which was also designed for the case of singular cotangent bundle reduction, was started in Hernandez and Marsden [60] and Birtea, Puta, Ratiu, and Tudoran [14], a technique which requires further development.

Singular reduction has been extensively used in the study of the persistence, bifurcation, and stability of relative dynamical elements; see [41,42,53,85,86,132,134,135,136,139,146,159,160,180,181].

Symplectic Reduction Without Momentum Maps

The reduction theory presented so far needs the existence of a momentum map. However, more primitive versions of this procedure based on foliation theory (see Cartan [26] and Meyer [117]) do not require the existence of this object. Working in this direction, but with a mathematical program that goes beyond the reduction problem, Condevaux, Dazord, and Molino [43] introduced a concept that generalizes the momentum map. This object is defined via a connection that associates an additive holonomy group to each canonical action on a symplectic manifold. The existence of the momentum map is equivalent to the vanishing of this group. Symplectic reduction has been carried out using this generalized momentum map in Ortega and Ratiu [140,141].

Another approach to symplectic reduction that is able to avoid the possible non-existence of the momentum map is based on the optimal momentum map introduced and studied in Ortega and Ratiu [137], Ortega [131], and [92]. This distribution theoretical approach can also deal with reduction of Poisson manifolds, where the standard momentum map does not exist generically.

Reduction of Other Geometric Structures

Besides symplectic reduction, there are many other geometric structures on which one can perform similar constructions. For example, one can reduce Kähler, hyper-Kähler, Poisson, contact, Jacobi, etc. manifolds and this can be done either in the regular or singular cases. We refer to [138] for a survey of the literature for these topics.

The Method of Invariants

This method seeks to parametrize quotient spaces by group invariant functions. It has a rich history going back to Hilbert's invariant theory. It has been of great use in

bifurcation with symmetry (see Golubitsky, Stewart, and Schaeffer [52] for instance). In mechanics, the method was developed by Kummer, Cushman, Rod and coworkers in the 1980's; see, for example, Cushman and Rod [45]. We will not attempt to give a literature survey here, other than to refer to Kummer [80], Kirk, Marsden, and Silber [73], Alber, Luther, Marsden, and Robbins [3] and the book of Cushman and Bates [44] for more details and references.

Nonholonomic Systems

Nonholonomic mechanical systems (such as systems with rolling constraints) provide a very interesting class of systems where the reduction procedure has to be modified. In fact this provides a class of systems that gives rise to an almost Poisson structure, i. e. a bracket which does not necessarily satisfy the Jacobi identity. Reduction theory for nonholonomic systems has made a lot of progress, but many interesting questions still remain. In these types of systems, there is a natural notion of a momentum map, but in general it is not conserved, but rather obeys a ***momentum equation*** as was discovered by Bloch, Krishnaprasad, Marsden, and Murray [20]. This means, in particular, that point reduction in such a situation may not be appropriate. Nevertheless, Poisson reduction in the almost Poisson and almost symplectic setting is interesting and from the mathematical point of view, point reduction is also interesting, although, as remarked, one has to be cautious with how it is applied to, for example, nonholonomic systems. A few references are Koiller [75], Bates and Sniatycki [13], Bloch, Krishnaprasad, Marsden, and Murray [20], Koon and Marsden [77], Blankenstein and Van Der Schaft [15], Cushman, Śniatycki [46], Planas-Bielsa [152], and Ortega and Planas-Bielsa [133]. We refer to Cendra, Marsden, and Ratiu [36] and Bloch [18] for a more detailed historical review.

Multisymplectic Reduction

Reduction theory is by no means completed. For example, for PDE's, the multisymplectic (as opposed to symplectic) framework seems appropriate, both for relativistic and nonrelativistic systems. In fact, this approach has experienced somewhat of a revival since it has been realized that it is rather useful for numerical computation (see Marsden, Patrick, and Shkoller [100]). Only a few instances and examples of multisymplectic and multi-Poisson reduction are really well understood (see Marsden, Montgomery, Morrison, and Thompson [97]; Castrillón-López, Ratiu and Shkoller [30], Castrillón-López, Garcia Pérez and Ratiu [27], Castrillón-López and Ratiu [28], Castrillón-López and Marsden [29]), so one can expect to see more activity in this area as well.

Discrete Mechanical Systems

Another emerging area, also motivated by numerical analysis, is that of discrete mechanics. Here the idea is to replace the velocity phase space TQ by $Q \times Q$, with the role of a velocity vector played by a pair of nearby points. This has been a powerful tool for numerical analysis, reproducing standard symplectic integration algorithms and much more. See, for example, Wendlandt and Marsden [178], Kane, Marsden, Ortiz and West [70], Marsden and West [113], Lew, Marsden, Ortiz, and West [87] and references therein. This subject, too, has its own reduction theory. See Marsden, Pekarsky, and Shkoller [101], Bobenko and Suris [23] and Jalnapurkar, Leok, Marsden and West [68]. Discrete mechanics also has some intriguing links with quantization, since Feynman himself first defined path integrals through a limiting process using the sort of discretization used in the discrete action principle (see Feynman and Hibbs [50]).

Cotangent Bundle Reduction

As mentioned earlier, the cotangent bundle reduction theorems are amongst the most basic and useful of the symplectic reduction theorems. Here we only present the *regular versions* of the theorems. Cotangent bundle reduction theorems come in two forms – the *embedding cotangent bundle reduction theorem* and the *bundle cotangent bundle reduction* theorem. We start with a smooth, free, and proper, left action

$$\Phi\colon G \times Q \to Q$$

of the Lie group G on the configuration manifold Q and lift it to an action on T^*Q. This lifted action is symplectic with respect to the canonical symplectic form on T^*Q, which we denote Ω_{can}, and has an *equivariant* momentum map $\mathbf{J}\colon T^*Q \to \mathfrak{g}^*$ given by

$$\langle \mathbf{J}(\alpha_q), \xi \rangle = \langle \alpha_q, \xi_Q(q) \rangle \ ,$$

where $\xi \in \mathfrak{g}$. Letting $\mu \in \mathfrak{g}^*$, the aim of this section is to determine the structure of the symplectic reduced space $((T^*Q)_\mu, \Omega_\mu)$, which, by Theorem 3, is a symplectic manifold. We are interested in particular in the question of to what extent $((T^*Q)_\mu, \Omega_\mu)$ is a synthesis of a cotangent bundles and a coadjoint orbit.

Cotangent Bundle Reduction: Embedding Version

In this version of the theorem, we first form the quotient manifold

$$Q_\mu := Q/G_\mu \,,$$

which we call the μ-***shape space***. Since the action of G on Q is smooth, free, and proper, so is the action of the isotropy subgroup G_μ and therefore, Q_μ is a smooth manifold and the canonical projection

$$\pi_{Q,G_\mu} : Q \to Q_\mu$$

is a surjective submersion.

Consider the G_μ-action on Q and its lift to T^*Q. This lifted action is of course also symplectic with respect to the canonical symplectic form Ω_{can} and has an equivariant momentum map $\mathbf{J}^\mu : T^*Q \to \mathfrak{g}_\mu^*$ obtained by restricting $\mathbf{J}$; that is, for $\alpha_q \in T_q^*Q$,

$$\mathbf{J}^\mu(\alpha_q) = \mathbf{J}(\alpha_q)|_{\mathfrak{g}_\mu} .$$

Let $\mu' := \mu|_{\mathfrak{g}_\mu} \in \mathfrak{g}_\mu^*$ be the restriction of μ to $\mathfrak{g}_\mu$. Notice that there is a natural inclusion of submanifolds

$$\mathbf{J}^{-1}(\mu) \subset (\mathbf{J}^\mu)^{-1}(\mu') . \tag{22}$$

Since the actions are free and proper, μ and μ' are regular values, so these sets are indeed smooth manifolds. Note that, by construction, μ' is G_μ-invariant.

There will be two key assumptions relevant to the embedding version of cotangent bundle reduction. Namely,

CBR1 *In the above setting, assume there is a G_μ-invariant one-form α_μ on Q with values in $(\mathbf{J}^\mu)^{-1}(\mu')$;*

and the condition (which by (22), is a stronger condition)

CBR2 *Assume that α_μ in **CBR1** takes values in $\mathbf{J}^{-1}(\mu)$.*

For $\xi \in \mathfrak{g}_\mu$ and $q \in Q$, notice that, under the condition **CBR1**,

$$(\mathbf{i}_{\xi_Q}\alpha_\mu)(q) = \langle \mathbf{J}(\alpha_\mu(q)), \xi \rangle = \langle \mu', \xi \rangle \,,$$

and so $\mathbf{i}_{\xi_Q}\alpha_\mu$ is a constant function on Q. Therefore, for $\xi \in \mathfrak{g}_\mu$,

$$\mathbf{i}_{\xi_Q}\mathbf{d}\alpha_\mu = \pounds_{\xi_Q}\alpha_\mu - \mathbf{d}\mathbf{i}_{\xi_Q}\alpha_\mu = 0 \,, \tag{23}$$

since the Lie derivative is zero by G_μ-invariance of α_μ. It follows that

There is a unique two-form β_μ on Q_μ such that

$$\pi_{Q,G_\mu}^*\beta_\mu = \mathbf{d}\alpha_\mu \,.$$

Since π_{Q,G_μ} is a submersion, β_μ is closed (it need not be exact). Let

$$B_\mu = \pi_{Q_\mu}^*\beta_\mu \,,$$

where $\pi_{Q_\mu} : T^*Q_\mu \to Q_\mu$ is (following our general conventions for maps) the cotangent bundle projection. Also, to avoid confusion with the canonical symplectic form Ω_{can} on T^*Q, we shall denote the canonical symplectic form on T^*Q_μ, the cotangent bundle of μ-shape space, by ω_{can}.

Theorem 9 (Cotangent bundle reduction – embedding version)

(i) *If condition **CBR1** holds, then there is a symplectic embedding*

$$\varphi_\mu : ((T^*Q)_\mu, \Omega_\mu) \to (T^*Q_\mu, \omega_{\text{can}} - B_\mu) \,,$$

onto a submanifold of T^*Q_μ covering the base Q/G_μ.

(ii) *The map φ_μ in (i) gives a symplectic diffeomorphism of $((T^*Q)_\mu, \Omega_\mu)$ onto $(T^*Q_\mu, \omega_{\text{can}} - B_\mu)$ if and only if $\mathfrak{g} = \mathfrak{g}_\mu$.*

(iii) *If **CBR2** holds, then the image of φ_μ equals the vector subbundle $[T\pi_{Q,G_\mu}(V)]^\circ$ of T^*Q_μ, where $V \subset TQ$ is the vector subbundle consisting of vectors tangent to the G-orbits, that is, its fiber at $q \in Q$ equals $V_q = \{\xi_Q(q) \mid \xi \in \mathfrak{g}\}$, and $^\circ$ denotes the annihilator relative to the natural duality pairing between TQ_μ and T^*Q_μ.*

Remarks

1. A history of this result can be found in Sect. "Reduction Theory: Historical Overview".
2. As shown in the appendix on Principal Connections (see Proposition A2) the required one form α_μ may be constructed satisfying condition **CBR1** from a connection on the μ-shape space bundle $\pi_{Q,G_\mu} : Q \to Q/G_\mu$ and an α_μ satisfying **CBR2** can be constructed using a connection on the shape space bundle $\pi_{Q,G} : Q \to Q/G$.
3. Note that in the case of Abelian reduction, or, more generally, the case in which $G = G_\mu$, the reduced space is symplectically diffeomorphic to $T^*(Q/G)$ with the symplectic structure given by $\Omega_{\text{can}} - B_\mu$. In particular, if $\mu = 0$, then the symplectic form on $T^*(Q/G)$ is the canonical one, since in this case one can choose $\alpha_\mu = 0$ which yields $B_\mu = 0$.
4. The term B_μ on T^*Q is usually called a ***magnetic term***, a ***gyroscopic term***, or a ***Coriolis term***. The terminology "magnetic" comes from the Hamiltonian description of a particle of charge e moving according to the Lorentz force law in $\mathbb{R}^3$ under the influence of a magnetic field B. This motion takes

place in $T^*\mathbb{R}^3$ but with the nonstandard symplectic structure $\mathbf{d}q^i \wedge \mathbf{d}p_i - \frac{e}{c}B$, $i = 1, 2, 3$, where c is the speed of light and B is regarded as a closed two-form: $B = B_x \mathbf{d}y \wedge \mathbf{d}z - B_y \mathbf{d}x \wedge \mathbf{d}z + B_z \mathbf{d}x \wedge \mathbf{d}y$ (see §6.7 in [104] for details)

The strategy for proving this theorem is to first deal with the case of reduction at zero and then to treat the general case using a momentum shift.

Reduction at Zero

The reduced space at $\mu = 0$ is, as a set,

$$(T^*Q)_0 = \mathbf{J}^{-1}(0)/G$$

since, for $\mu = 0, G_\mu = G$. Notice that in this case, there is no distinction between orbit reduction and symplectic reduction.

Theorem 10 (Reduction at zero) *Assume that the action of G on Q is free and proper, so that the quotient Q/G is a smooth manifold. Then 0 is a regular value of $\mathbf{J}$ and there is a symplectic diffeomorphism between $(T^*Q)_0$ and $T^*(Q/G)$ with its canonical symplectic structure.*

The Case $G = G_\mu$

If one is reducing at zero, then clearly $G = G_\mu$. However, this is an important special case of the general cotangent bundle reduction theorem that, for example, includes the case of Abelian reduction. The key assumption here is that $G = G_\mu$, which indeed is always the case if G is Abelian.

Theorem 11 *Assume that the action of G on Q is free and proper, so that the quotient Q/G is a smooth manifold. Let $\mu \in \mathfrak{g}^*$, assume that $G = G_\mu$, and assume that **CBR2** holds. Then μ is a regular value of $\mathbf{J}$ and there is a symplectic diffeomorphism between $(T^*Q)_\mu$ and $T^*(Q/G)$, the latter with the symplectic form $\omega_{\text{can}} - B_\mu$; here, ω_{can} is the canonical symplectic form on $T^*(Q/G)$ and $B_\mu = \pi^*_{Q/G}\beta_\mu$, where the two form β_μ on Q/G is defined by*

$$\pi^*_{Q,G}\beta_\mu = \mathbf{d}\alpha_\mu \ .$$

Example

Consider the reduction of a general cotangent bundle T^*Q by $G = \mathrm{SO}(3)$. Here $G_\mu \cong S^1$, if $\mu \neq 0$, and so the reduced space is embedded into the cotangent bundle $T^*(Q/S^1)$. A specific example is the case of $Q = \mathrm{SO}(3)$. Then the reduced space $(T^*\mathrm{SO}(3))_\mu$ is $S^2_{\|\mu\|}$, the sphere of radius $\|\mu\|$ which is a coadjoint orbit in $\mathfrak{so}(3)^*$. In this case, $Q/G_\mu = \mathrm{SO}(3)/S^1 \cong S^2_{\|\mu\|}$ and the embedding of $S^2_{\|\mu\|}$ into $T^*S^2_{\|\mu\|}$ is the zero section.

Magnetic Terms and Curvature

Using the results of the preceding section, we will now show how one can interpret the magnetic term B_μ as the curvature of a connection on a principal bundle.

We saw in the preamble to the Cotangent Bundle Reduction Theorem 9 that $\mathbf{i}_{\xi_Q}\mathbf{d}\alpha_\mu = 0$ for any $\xi \in \mathfrak{g}_\mu$, which was used to drop $\mathbf{d}\alpha_\mu$ to the quotient. In the language of principal bundles, this may be rephrased by saying that $\mathbf{d}\alpha_\mu$ is *horizontal* and thus, once a connection is introduced, *the covariant exterior derivative of α_μ coincides with $\mathbf{d}\alpha_\mu$*.

There are two methods to construct a form α_μ with the properties in Theorem 9. We continue to work under the general assumption that G acts on Q freely and properly.

First Method

Construction of α_μ from a connection $\mathcal{A}^\mu \in \Omega^1(Q; \mathfrak{g}_\mu)$ on the principal bundle $\pi_{Q,G_\mu} : Q \to Q/G_\mu$.

To carry this out, one shows that the choice

$$\alpha_\mu := \langle \mu', \mathcal{A}^\mu \rangle \in \Omega^1(Q)$$

satisfies the condition **CBR1** in Theorem 9, where, as above, $\mu' = \mu|_{\mathfrak{g}_\mu}$. The two-form $\mathbf{d}\alpha_\mu$ may be interpreted in terms of curvature. In fact, one shows that $\mathbf{d}\alpha_\mu$ is the *μ'-component of the curvature two-form*. We summarize these results in the following statement.

Proposition 12 *If the principal bundle $\pi_{Q,G_\mu} : Q \to Q/G_\mu$ with structure group G_μ has a connection $\mathcal{A}^\mu$, then $\alpha_\mu(q)$ can be taken to equal $\mathcal{A}^\mu(q)^*\mu'$ and B_μ is induced on T^*Q_μ by $\mathbf{d}\alpha_\mu$ (a two-form on Q), which equals the μ'-component of the curvature $\mathcal{B}^\mu$ of $\mathcal{A}^\mu$.*

Second Method

Construction of α_μ from a connection $\mathcal{A} \in \Omega^1(Q; \mathfrak{g})$ on the principal bundle $\pi_{Q,G} : Q \to Q/G$. One can show that the choice (A1), that is,

$$\alpha_\mu := \langle \mu, \mathcal{A} \rangle \in \Omega^1(Q)$$

satisfies the condition **CBR2** in Theorem 9.

As with the first method, there is an interpretation of the two-form $\mathbf{d}\alpha_\mu$ in terms of curvature as follows.

Proposition 13 *If the principal bundle $\pi_{Q,G} : Q \to Q/G$ with structure group G has a connection $\mathcal{A}$, then $\alpha_\mu(q)$ can be taken to equal $\mathcal{A}(q)^*\mu$ and B_μ is the pull back to T^*Q_μ of $\mathbf{d}\alpha_\mu \in \Omega^2(Q)$, which equals the μ-component of the two form $\mathcal{B} + [\mathcal{A}, \mathcal{A}] \in \Omega^2(Q; \mathfrak{g})$, where $\mathcal{B}$ is the curvature of $\mathcal{A}$.*

Coadjoint Orbits

We now apply the Cotangent Bundle Reduction Theorem 9 to the case $Q = G$ and with the G-action given by *left* translation. The *right* Maurer–Cartan form θ^R is a flat connection associated to this action (see Theorem A13) and hence

$$\begin{aligned}\mathbf{d}\alpha_\mu(g)(u_g, v_g) &= \langle \mu, [\theta^R, \theta^R](g)(u_g, v_g)\rangle \\ &= \langle \mu, [T_g R_{g^{-1}} u_g, T_g R_{g^{-1}} v_g]\rangle \ .\end{aligned}$$

Recall from Theorem 7 that the reduced space $(T^*G)_\mu$ is the coadjoint orbit $\mathcal{O}_\mu$ endowed with the negative orbit symplectic form ω_μ^- and, according to the Cotangent Bundle Reduction Theorem, it symplectically embeds as the zero section into $(T^*\mathcal{O}_\mu, \omega_{\text{can}} - B_\mu)$, where $B_\mu = \pi^*_{\mathcal{O}_\mu}\beta_\mu$, $\pi_{\mathcal{O}_\mu} : T^*\mathcal{O}_\mu \to \mathcal{O}_\mu$ is the cotangent bundle projection, $\pi^*_{G,G_\mu}\beta_\mu = \mathbf{d}\alpha_\mu$, and $\pi_{G,G_\mu} : G \to \mathcal{O}_\mu$ is given by $\pi_{G,G_\mu}(g) = \text{Ad}^*_g\mu$. The derivative of π_{G,G_μ} is given by

$$T_g\pi_{G,G_\mu}(T_e L_g \xi) = \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \text{Ad}^*_{g\exp(t\xi)}\mu = \text{ad}^*_\xi \text{Ad}^*_g \mu$$

for any $\xi \in \mathfrak{g}$. Then a computation shows that $\beta_\mu = -\omega_\mu^-$. Thus, the embedding version of the cotangent bundle reduction theorem produces the following statement which, of course, can be easily checked directly.

Corollary 14 *The coadjoint orbit $(\mathcal{O}_\mu, \omega_\mu^-)$ symplectically embeds as the zero section into the symplectic manifold $(T^*\mathcal{O}_\mu, \omega_{\text{can}} + \pi^*_{\mathcal{O}_\mu}\omega_\mu^-)$.*

Cotangent Bundle Reduction: Bundle Version

The embedding version of the cotangent bundle reduction theorem presented in the preceding section states that $(T^*Q)_\mu$ *embeds* as a vector subbundle of $T^*(Q/G_\mu)$. The bundle version of this theorem says, roughly speaking, that $(T^*Q)_\mu$ is a *coadjoint orbit bundle* over $T^*(Q/G)$ with fiber the coadjoint orbit $\mathcal{O}$ through μ.

Again we utilize a choice of connection $\mathcal{A}$ on the shape space bundle $\pi_{Q,G} : Q \to Q/G$. A key step in the argument is to utilize orbit reduction and the identification $(T^*Q)_\mu \cong (T^*Q)_{\mathcal{O}}$.

Theorem 15 (Cotangent bundle reduction – bundle version) *The reduced space $(T^*Q)_\mu$ is a locally trivial fiber bundle over $T^*(Q/G)$ with typical fiber $\mathcal{O}$.*

This point of view is explored further and the exact nature of the coadjoint orbit bundle is identified and its symplectic structure is elaborated in [92].

Poisson Version

This same type of argument as above shows the following, which we state slightly informally.

Theorem *The Poisson reduced space $(T^*Q)/G$ is diffeomorphic to the coadjoint bundle of $\pi_{Q,G} : Q \to Q/G$. This diffeomorphism is implemented by a connection $\mathcal{A} \in \Omega^1(Q;\mathfrak{g})$. Thus the fiber of $(T^*Q)/G \to T^*(Q/G)$ is isomorphic to the Lie–Poisson space $\mathfrak{g}^*$.*

There is an interesting formula for the Poisson structure on $(T^*Q)/G$ that was originally computed in Montgomery, Marsden, and Ratiu [127], Montgomery [121]. Further developments in Cendra, Marsden, Pekarsky, and Ratiu [37] and Perlmutter and Ratiu [149] gives a unified study of the Poisson bracket on $(T^*Q)/G$ in both the Sternberg and Weinstein realizations of the quotient. Finally, we refer to, for instance, Lewis, Marsden, Montgomery and Ratiu [88] for an application of this result; in this case, the dynamics of fluid systems with free boundaries is studied.

Coadjoint Orbit Bundles

The details of the nature of the bundle and its associated symplectic structure that was sketched in Theorem 15 is due to Marsden and Perlmutter [102]; see also Zaalani [182] Cushman and Śniatycki [47], and [149]. An exposition may be found in [92].

Future Directions

One of the goals of reduction theory and geometric mechanics is to take the analysis of mechanical systems with symmetries to a deeper level of understanding. But much more needs to be done. As has already been explained, there is still a need to put many classical concepts, such as quasivelocities, into this context, with a resultant strengthening of the theory and its applications. In addition, links with Dirac structures, groupoids and algebroids is under development and should lead to further advances. Finally we mention that while much of this type of work has been applied to field theories (such as electromagnetism and gravity), greater insight is needed for many topics, stress-energy-momentum tensors being one example.

Acknowledgments

This work summarizes the contributions of many people. We are especially grateful to Alan Weinstein, Victor Guillemin and Shlomo Sternberg for their incredible insights and work over the last few decades. We also thank Hernán Cendra and Darryl Holm, our collaborators on

the Lagrangian context and Juan-Pablo Ortega, a longtime collaborator on Hamiltonian reduction and other projects; he along with Gerard Misiolek and Matt Perlmutter were our collaborators on [92], a key recent project that helped us pull many things together. We also thank many other colleagues for their input and invaluable support over the years; this includes Larry Bates, Tony Bloch, Marco Castrillón-López, Richard Cushman, Laszlo Fehér, Mark Gotay, John Harnad, Eva Kanso, Thomas Kappeler, P.S. Krishnaprasad, Naomi Leonard, Debra Lewis, James Montaldi, George Patrick, Mark Roberts, Miguel Rodríguez-Olmos, Steve Shkoller, Jędrzej Śniatycki, Leon Takhtajan, Karen Vogtmann, and Claudia Wulff.

Appendix: Principal Connections

In preparation for the next section which gives a brief exposition of the cotangent bundle reduction theorem, we now give a review and summary of facts that we shall need about principal connections. An important thing to keep in mind is that the magnetic terms in the cotangent bundle reduction theorem will appear as the curvature of a connection.

Principal Connections Defined

We consider the following basic set up. Let Q be a manifold and let G be a Lie group acting freely and properly on the left on Q. Let

$$\pi_{Q,G}\colon Q \to Q/G$$

denote the bundle projection from the configuration manifold Q to ***shape space*** $S = Q/G$. We refer to $\pi_{Q,G}\colon Q \to Q/G$ as a ***principal bundle***.

One can alternatively use right actions, which is common in the principal bundle literature, but we shall stick with the case of left actions for the main exposition.

Vectors that are infinitesimal generators, namely those of the form $\xi_Q(q)$ are called ***vertical*** since they are sent to zero by the tangent of the projection map $\pi_{Q,G}$.

Definition A1 A ***connection***, also called a ***principal connection*** on the bundle $\pi_{Q,G}\colon Q \to Q/G$ is a Lie algebra valued 1-form

$$\mathcal{A}\colon TQ \to \mathfrak{g}$$

where $\mathfrak{g}$ denotes the Lie algebra of G, with the following properties:

(i) the identity $\mathcal{A}(\xi_Q(q)) = \xi$ holds for all $\xi \in \mathfrak{g}$; that is, $\mathcal{A}$ takes infinitesimal generators of a given Lie algebra element to that same element, and

(ii) we have ***equivariance***: $\mathcal{A}(T_q\Phi_g(v)) = \mathrm{Ad}_g(\mathcal{A}(v))$

for all $v \in T_qQ$, where $\Phi_g\colon Q \to Q$ denotes the given action for $g \in G$ and where Ad_g denotes the adjoint action of G on $\mathfrak{g}$.

A remark is noteworthy at this point. The equivariance identity for infinitesimal generators noted previously (see (7)), namely,

$$T_q\Phi_g\left(\xi_Q(q)\right) = (\mathrm{Ad}_g\xi)_Q(g\cdot q)\,,$$

shows that *if the first condition for a connection holds, then the second condition holds automatically on vertical vectors*.

If the G-action on Q is a *right action*, the equivariance condition (ii) in Definition A1 needs to be changed to $\mathcal{A}(T_q\Phi_g(v)) = \mathrm{Ad}_{g^{-1}}(\mathcal{A}(v))$ for all $g \in G$ and $v \in T_qQ$.

Associated One-Forms

Since $\mathcal{A}$ is a Lie algebra valued 1-form, for each $q \in Q$, we get a linear map $\mathcal{A}(q)\colon T_qQ \to \mathfrak{g}$ and so we can form its dual $\mathcal{A}(q)^*\colon \mathfrak{g}^* \to T_q^*Q$. Evaluating this on μ produces an ordinary 1-form:

$$\alpha_\mu(q) = \mathcal{A}(q)^*(\mu)\,. \tag{A1}$$

This 1-form satisfies two important properties given in the next Proposition.

Proposition A2 *For any connection $\mathcal{A}$ and $\mu \in \mathfrak{g}^*$, the corresponding 1-form α_μ defined by (A1) takes values in $\mathbf{J}^{-1}(\mu)$ and satisfies the following G-equivariance property:*

$$\Phi_g^*\alpha_\mu = \alpha_{\mathrm{Ad}_g^*\mu}\,.$$

Notice in particular, if the group is Abelian or if μ is G-invariant, (for example, if $\mu = 0$), then α_μ is an *invariant 1-form*.

Horizontal and Vertical Spaces

Associated with any connection are vertical and horizontal spaces defined as follows.

Definition A3 Given the connection $\mathcal{A}$, its ***horizontal space*** at $q \in Q$ is defined by

$$H_q = \{v_q \in T_qQ \mid \mathcal{A}(v_q) = 0\}$$

and the ***vertical space*** at $q \in Q$ is, as above,

$$V_q = \{\xi_Q(q) \mid \xi \in \mathfrak{g}\}\,.$$

The map

$$v_q \mapsto \mathrm{ver}_q(v_q) := [\mathcal{A}(q)(v_q)]_Q(q)$$

is called the ***vertical projection***, while the map

$$v_q \mapsto \mathrm{hor}_q(v_q) := v_q - \mathrm{ver}_q(v_q)$$

is called the ***horizontal projection***.

Because connections map infinitesimal generators of a Lie algebra elements to that same Lie algebra element, the vertical projection is indeed a projection for each fixed q onto the vertical space and likewise with the horizontal projection.

By construction, we have

$$v_q = \mathrm{ver}_q(v_q) + \mathrm{hor}_q(v_q)$$

and so

$$T_qQ = H_q \oplus V_q$$

and the maps hor_q and ver_q are projections onto these subspaces.

It is sometimes convenient to *define* a connection by the specification of a space H_q declared to be the horizontal space that is complementary to V_q at each point, varies smoothly with q and respects the group action in the sense that $H_{g\cdot q} = T_q\Phi_g(H_q)$. Clearly this alternative definition of a principal connection is equivalent to the definition given above.

Given a point $q \in Q$, the tangent of the projection map $\pi_{Q,G}$ restricted to the horizontal space H_q gives an isomorphism between H_q and $T_{[q]}(Q/G)$. Its inverse $\left[T_q\pi_{Q,G}|_{H_q}\right]^{-1} : T_{\pi_{Q,G}(q)}(Q/G) \to H_q$ is called the ***horizontal lift*** to $q \in Q$.

The Mechanical Connection

As an example of defining a connection by the specification of a horizontal space, suppose that the configuration manifold Q is a Riemannian manifold. Of course, the Riemannian structure will often be that defined by the kinetic energy of a given mechanical system.

Thus, assume that Q is a Riemannian manifold, with metric denoted $\langle\langle , \rangle\rangle$ and that G acts freely and properly on Q by isometries, so $\pi_{Q,G}: Q \to Q/G$ is a principal G-bundle.

In this context we may *define the horizontal space at a point simply to be the metric orthogonal to the vertical space*. This therefore defines a connection called the **mechanical connection**.

Recall from the historical survey in the introduction that this connection was first introduced by Kummer [79] following motivation from Smale [168] and [1]. See also Guichardet [54], who applied these ideas in an interesting way to molecular dynamics. The number of references since then making use of the mechanical connection is too large to survey here.

In Proposition A5 we develop an explicit formula for the associated Lie algebra valued 1-form in terms of an inertia tensor and the momentum map. As a prelude to this formula, we show the following basic link with mechanics. In this context we write the momentum map on TQ simply as $\mathbf{J}: TQ \to \mathfrak{g}^*$.

Proposition A4 *The horizontal space of the mechanical connection at a point $q \in Q$ consists of the set of vectors $v_q \in T_qQ$ such that $\mathbf{J}(v_q) = 0$.*

For each $q \in Q$, define the ***locked inertia tensor*** $\mathbb{I}(q)$ to be the linear map $\mathbb{I}(q): \mathfrak{g} \to \mathfrak{g}^*$ defined by

$$\langle \mathbb{I}(q)\eta, \zeta\rangle = \langle\langle \eta_Q(q), \zeta_Q(q)\rangle\rangle \tag{A2}$$

for any $\eta, \zeta \in \mathfrak{g}$. Since the action is free, $\mathbb{I}(q)$ is nondegenerate, so (A2) defines an inner product. The terminology "locked inertia tensor" comes from the fact that for coupled rigid or elastic systems, $\mathbb{I}(q)$ is the classical moment of inertia tensor of the rigid body obtained by locking all the joints of the system. In coordinates,

$$I_{ab} = g_{ij}K^i_aK^j_b\,, \tag{A3}$$

where $[\xi_Q(q)]^i = K^i_a(q)\xi^a$ define the ***action functions*** K^i_a.

Define the map $\mathcal{A}: TQ \to \mathfrak{g}$ which assigns to each $v_q \in T_qQ$ the corresponding *angular velocity of the locked system*:

$$\mathcal{A}(q)(v_q) = \mathbb{I}(q)^{-1}(\mathbf{J}(v_q))\,, \tag{A4}$$

where L is the kinetic energy Lagrangian. In coordinates,

$$\mathcal{A}^a = I^{ab}g_{ij}K^i_bv^j \tag{A5}$$

since $J_a(q,p) = p_iK^i_a(q)$.

We defined the mechanical connection by declaring its horizontal space to be the metric orthogonal to the vertical space. The next proposition shows that $\mathcal{A}$ is the associated connection one-form.

Proposition A5 *The $\mathfrak{g}$-valued one-form defined by (A4) is the mechanical connection on the principal G-bundle $\pi_{Q,G}: Q \to Q/G$.*

Given a general connection $\mathcal{A}$ and an element $\mu \in \mathfrak{g}^*$, we can define the μ-component of $\mathcal{A}$ to be the ordinary one-form α_μ given by

$$\alpha_\mu(q) = \mathcal{A}(q)^*\mu \in T^*_qQ; \quad \text{i.e.,} \quad \langle\alpha_\mu(q), v_q\rangle = \langle\mu, \mathcal{A}(q)(v_q)\rangle$$

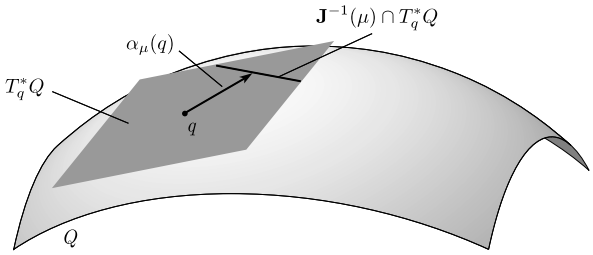

Mechanical Systems: Symmetries and Reduction, Figure 2
The extremal characterization of the mechanical connection

for all $v_q \in T_qQ$. Note that α_μ is a G_μ-invariant one-form. It takes values in $\mathbf{J}^{-1}(\mu)$ since for any $\xi \in \mathfrak{g}$, we have

$$\begin{aligned}\big\langle \mathbf{J}(\alpha_\mu(q)), \xi \big\rangle &= \big\langle \alpha_\mu(q), \xi_Q \big\rangle = \langle \mu, \mathcal{A}(q)(\xi_Q(q)) \rangle \\ &= \langle \mu, \xi \rangle \,.\end{aligned}$$

In the Riemannian context, Smale [168] constructed α_μ by a minimization process. Let $\alpha_q^\sharp \in T_qQ$ be the tangent vector that corresponds to $\alpha_q \in T_q^*Q$ via the metric $\langle\langle\, , \rangle\rangle$ on Q.

Proposition A6 *The 1-form $\alpha_\mu(q) = \mathcal{A}(q)^*\mu \in T_q^*Q$ associated with the mechanical connection $\mathcal{A}$ given by (A4) is characterized by*

$$K(\alpha_\mu(q)) = \inf\{K(\beta_q) \mid \beta_q \in \mathbf{J}^{-1}(\mu) \cap T_q^*Q\} \,, \quad \text{(A6)}$$

where $K(\beta_q) = \frac{1}{2}\|\beta_q^\sharp\|^2$ is the kinetic energy function on T^*Q. See Fig. 2.

The proof is a direct verification. We do not give here it since this proposition will not be used later in this book. The original approach of Smale [168] was to take (A6) as the definition of α_μ. To prove from here that α_μ is a smooth one-form is a nontrivial fact; see the proof in Smale [168] or of Proposition 4.4.5 in [1]. Thus, one of the merits of the previous proposition is to show easily that this variational definition of α_μ does indeed yield a smooth one-form on Q with the desired properties. Note also that $\alpha_\mu(q)$ lies in the orthogonal space to $T_q^*Q \cap \mathbf{J}^{-1}(\mu)$ in the fiber T_q^*Q relative to the bundle metric on T^*Q defined by the Riemannian metric on Q. It also follows that $\alpha_\mu(q)$ is the unique critical point of the kinetic energy of the bundle metric on T^*Q restricted to the fiber $T_q^*Q \cap \mathbf{J}^{-1}(\mu)$.

Curvature

The curvature $\mathcal{B}$ of a connection $\mathcal{A}$ is defined as follows.

Definition A7 The ***curvature*** of a connection $\mathcal{A}$ is the Lie algebra valued two-form on Q defined by

$$\mathcal{B}(q)(u_q, v_q) = \mathbf{d}\mathcal{A}(\mathrm{hor}_q(u_q), \mathrm{hor}_q(v_q)) \,, \quad \text{(A7)}$$

where $\mathbf{d}$ is the exterior derivative.

When one replaces vectors in the exterior derivative with their horizontal projections, then the result is called the ***exterior covariant derivative*** and one writes the preceding formula for $\mathcal{B}$ as

$$\mathcal{B} = \mathbf{d}^{\mathcal{A}}\mathcal{A} \,.$$

For a general Lie algebra valued k-form α on Q, the **exterior covariant derivative** is the $k+1$-form $\mathbf{d}^{\mathcal{A}}\alpha$ defined on tangent vectors $v_0, v_1, \ldots, v_k \in T_qQ$ by

$$\begin{aligned}&\mathbf{d}^{\mathcal{A}}\alpha(v_0, v_1, \ldots, v_k) \\ &\qquad = \mathbf{d}\alpha\big(\mathrm{hor}_q(v_0), \mathrm{hor}_q(v_1), \ldots, \mathrm{hor}_q(v_k)\big) \,. \quad \text{(A8)}\end{aligned}$$

Here, the symbol $\mathbf{d}^{\mathcal{A}}$ reminds us that it is like the exterior derivative but that it depends on the connection $\mathcal{A}$.

Curvature *measures the lack of integrability of the horizontal distribution* in the following sense.

Proposition A8 *On two vector fields u, v on Q one has*

$$\mathcal{B}(u, v) = -\mathcal{A}\big([\mathrm{hor}(u), \mathrm{hor}(v)]\big) \,.$$

Given a general distribution $\mathcal{D} \subset TQ$ on a manifold Q one can also define its curvature in an analogous way directly in terms of its lack of integrability. Define ***vertical vectors*** at $q \in Q$ to be the quotient space $T_qQ/\mathcal{D}_q$ and define the curvature acting on two ***horizontal vector fields*** u, v (that is, two vector fields that take their values in the distribution) to be the projection onto the quotient of their Jacobi–Lie bracket. One can check that this operation depends only on the point values of the vector fields, so indeed defines a two-form on horizontal vectors.

Cartan Structure Equations

We now derive an important formula for the curvature of a principal connection.

Theorem A9 (Cartan structure equations) *For any vector fields u, v on Q we have*

$$\mathcal{B}(u, v) = \mathbf{d}\mathcal{A}(u, v) - [\mathcal{A}(u), \mathcal{A}(v)] \,, \quad \text{(A9)}$$

where the bracket on the right hand side is the Lie bracket in $\mathfrak{g}$. We write this equation for short as

$$\mathcal{B} = \mathbf{d}\mathcal{A} - [\mathcal{A}, \mathcal{A}] \,.$$

If the G-action on Q is a *right action*, then the Cartan Structure Equations read $\mathcal{B} = \mathbf{d}\mathcal{A} + [\mathcal{A}, \mathcal{A}]$.

The following Corollary shows how the Cartan Structure Equations yield a fundamental equivariance property of the curvature.

Corollary A10 *For all $g \in G$ we have $\Phi_g^* \mathcal{B} = \mathrm{Ad}_g \circ \mathcal{B}$. If the G-action on Q is on the right, equivariance means $\Phi_g^* \mathcal{B} = \mathrm{Ad}_{g^{-1}} \circ \mathcal{B}$.*

Bianchi Identity

The ***Bianchi Identity***, which states that the exterior covariant derivative of the curvature is zero, is another important consequence of the Cartan Structure Equations.

Corollary A11 *If $\mathcal{B} = d^{\mathcal{A}}\mathcal{A} \in \Omega^2(Q; \mathfrak{g})$ is the curvature two-form of the connection $\mathcal{A}$, then the **Bianchi Identity** holds:*

$$d^{\mathcal{A}}\mathcal{B} = 0\,.$$

This form of the Bianchi identity is implied by another version, namely

$$\mathrm{d}\mathcal{B} = [\mathcal{B}, \mathcal{A}]^\wedge,$$

where the bracket on the right hand side is that of Lie algebra valued differential forms, a notion that we do not develop here; see the brief discussion at the end of §9.1 in [104]. The proof of the above form of the Bianchi identity can be found in, for example, Kobayashi and Nomizu [74].

Curvature as a Two-Form on the Base

We now show how the curvature two-form drops to a two-form on the base with values in the adjoint bundle.

The associated bundle to the given left principal bundle $\pi_{Q,G}\colon Q \to Q/G$ via the adjoint action is called the ***adjoint bundle***. It is defined in the following way. Consider the free proper action $(g, (q, \xi)) \in G \times (Q \times \mathfrak{g}) \mapsto (g{\cdot}q, \mathrm{Ad}_g\xi) \in Q{\times}\mathfrak{g}$ and form the quotient $\tilde{\mathfrak{g}} := Q{\times}_G\mathfrak{g} := (Q \times \mathfrak{g})/G$ which is easily verified to be a vector bundle $\pi_{\tilde{\mathfrak{g}}}\colon \tilde{\mathfrak{g}} \to Q/G$, where $\pi_{\tilde{\mathfrak{g}}}(g, \xi) := \pi_{Q,G}(q)$. This vector bundle has an additional structure: it is a ***Lie algebra bundle***; that is, a vector bundle whose fibers are Lie algebras. In this case the bracket is defined pointwise:

$$[\pi_{\tilde{\mathfrak{g}}}(g, \xi), \pi_{\tilde{\mathfrak{g}}}(g, \eta)] := \pi_{\tilde{\mathfrak{g}}}(g, [\xi, \eta])$$

for all $g \in G$ and $\xi, \eta \in \mathfrak{g}$. It is easy to check that this defines a Lie bracket on every fiber and that this operation is smooth as a function of $\pi_{Q,G}(q)$.

The curvature two-form $\mathcal{B} \in \Omega^2(Q; \mathfrak{g})$ (the vector space of $\mathfrak{g}$-valued two-forms on Q) naturally induces a two-form $\overline{\mathcal{B}}$ on the base Q/G with values in $\tilde{\mathfrak{g}}$ by

$$\overline{\mathcal{B}}(\pi_{Q,G}(q))\left(T_q\pi_{Q,G}(u), T_q\pi_{Q,G}(v)\right) := \pi_{\tilde{\mathfrak{g}}}\left(q, \mathcal{B}(u, v)\right) \tag{A10}$$

for all $q \in Q$ and $u, v \in T_qQ$. One can check that $\overline{\mathcal{B}}$ is well defined.

Since (A10) can be equivalently written as $\pi_{Q,G}^*\overline{\mathcal{B}} = \pi_{\tilde{\mathfrak{g}}} \circ \left(\mathrm{id}_Q \times \mathcal{B}\right)$ and $\pi_{Q,G}$ is a surjective submersion, it follows that $\overline{\mathcal{B}}$ is indeed a smooth two-form on Q/G with values in $\tilde{\mathfrak{g}}$.

Associated Two-Forms

Since $\mathcal{B}$ is a $\mathfrak{g}$-valued two-form, in analogy with (A1), for every $\mu \in \mathfrak{g}^*$ we can define the ***μ-component*** of $\mathcal{B}$, an ordinary two-form $\mathcal{B}_\mu \in \Omega^2(Q)$ on Q, by

$$\mathcal{B}_\mu(q)(u_q, v_q) := \left\langle \mu, \mathcal{B}(q)(u_q, v_q)\right\rangle \tag{A11}$$

for all $q \in Q$ and $u_q, v_q \in T_qQ$.

The adjoint bundle valued curvature two-form $\overline{\mathcal{B}}$ induces an ordinary two-form on the base Q/G. To obtain it, we consider the dual $\tilde{\mathfrak{g}}^*$ of the adjoint bundle. This is a vector bundle over Q/G which is the associated bundle relative to the coadjoint action of the structure group G of the principal (left) bundle $\pi_{Q,G}\colon Q \to Q/G$ on $\mathfrak{g}^*$. This vector bundle has additional structure: each of its fibers is a Lie–Poisson space and the associated Poisson tensors on each fiber depend smoothly on the base, that is, $\pi_{\tilde{\mathfrak{g}}^*}\colon \tilde{\mathfrak{g}}^* \to Q/G$ is a ***Lie–Poisson bundle*** over Q/G.

Given $\mu \in \mathfrak{g}^*$, define the *ordinary two-form $\overline{\mathcal{B}}_\mu$ on Q/G* by

$$\begin{aligned}
&\overline{\mathcal{B}}_\mu\left(\pi_{Q,G}(q)\right)\left(T_q\pi_{Q,G}(u_q), T_q\pi_{Q,G}(v_q)\right)\\
&:= \left\langle \pi_{\tilde{\mathfrak{g}}^*}(q, \mu), \overline{\mathcal{B}}(\pi_{Q,G}(q))\left(T_q\pi_{Q,G}(u_q), T_q\pi_{Q,G}(v_q)\right)\right\rangle\\
&= \left\langle \mu, \mathcal{B}(q)(u_q, v_q)\right\rangle = \mathcal{B}_\mu(q)(u_q, v_q)\,,
\end{aligned} \tag{A12}$$

where $q \in Q$, $u_q, v_q \in T_qQ$, and in the second equality $\langle , \rangle : \tilde{\mathfrak{g}}^* \times \tilde{\mathfrak{g}} \to \mathbb{R}$ is the duality pairing between the coadjoint and adjoint bundles. Since $\overline{\mathcal{B}}$ is well defined and smooth, so is $\overline{\mathcal{B}}_\mu$.

Proposition A12 *Let $\mathcal{A} \in \Omega^1(Q; \mathfrak{g})$ be a connection one-form on the (left) principal bundle $\pi_{Q,G}\colon Q \to Q/G$ and $\mathcal{B} \in \Omega^2(Q; \mathfrak{g})$ its curvature two-form on Q. If $\mu \in \mathfrak{g}^*$, the corresponding two-forms $\mathcal{B}_\mu \in \Omega^2(Q)$ and $\overline{\mathcal{B}}_\mu \in \Omega^2(Q/G)$ defined by (A11) and (A12), respectively,*

are related by $\pi^*_{Q,G}\overline{\mathcal{B}}_\mu = \mathcal{B}_\mu$. *In addition,* $\mathcal{B}_\mu$ *satisfies the following G-equivariance property:*

$$\Phi_g^* \mathcal{B}_\mu = \mathcal{B}_{\mathrm{Ad}_g^* \mu} .$$

Thus, if $G = G_\mu$ *then* $\mathbf{d}\alpha_\mu = \mathcal{B}_\mu = \pi^*_{Q,G}\overline{\mathcal{B}}_\mu$, *where* $\alpha_\mu(q) = \mathcal{A}(q)^*(\mu)$.

Further relations between α_μ and the μ-component of the curvature will be studied in the next section when discussing the magnetic terms appearing in cotangent bundle reduction.

The Maurer–Cartan Equations

A consequence of the structure equations relates curvature to the process of left and right trivialization and hence to momentum maps.

Theorem A13 (Maurer–cartan equations) *Let G be a Lie group and let* $\theta^R \colon TG \to \mathfrak{g}$ *be the map (called the right Maurer–Cartan form) that right translates vectors to the identity:*

$$\theta^R(v_g) = T_g R_{g^{-1}}(v_g) .$$

Then

$$\mathbf{d}\theta^R - [\theta^R, \theta^R] = 0 .$$

There is a similar result for the left trivialization θ^L, namely the identity

$$\mathbf{d}\theta^L + [\theta^L, \theta^L] = 0 .$$

Of course there is much more to this subject, such as the link with classical connection theory, Riemannian geometry, etc. We refer to [92] for further basic information and references and to Bloch [18] for applications to nonholonomic systems, and to Cendra, Marsden, and Ratiu [35] for applications to Lagrangian reduction.

Bibliography

1. Abraham R, Marsden JE (2008) Foundations of Mechanics, 2nd edn. AMS Chelsea Publ, Providence. Orginally published in 1967; second edition revised and enlarged with the assistance of Tudor Ratiu and Richard Cushman, 1978
2. Abraham R, Marsden JE, Ratiu T (1988) Manifolds, Tensor Analysis and Applications, 2nd edn. Applied Mathematical Sciences, vol 75. Springer, New York
3. Alber MS, Luther GG, Marsden JE, Robbins JM (1998) Geometric phases, reduction and Lie-Poisson structure for the resonant three-wave interaction. Physica D 123:271–290
4. Arms JM, Cushman RH, Gotay M (1991) A universal reduction procedure for Hamiltonian group actions. In: Ratiu T (ed) The Geometry of Hamiltonian systems. MSRI Series, vol 22. Springer, New York, pp 33–52
5. Arms JM, Marsden JE, Moncrief V (1981) Symmetry and bifurcations of momentum mappings. Comm Math Phys 78: 455–478
6. Arms JM, Marsden JE, Moncrief V (1982) The structure of the space solutions of Einstein's equations: II Several Killing fields and the Einstein–Yang–Mills equations. Ann Phys 144:81–106
7. Arnold VI (1966) Sur la géométrie differentielle des groupes de Lie de dimenson infinie et ses applications à l'hydrodynamique des fluids parfaits. Ann Inst Fourier Grenoble 16:319–361
8. Arnold VI (1969) On an a priori estimate in the theory of hydrodynamical stability. Am Math Soc Transl 79:267–269
9. Arnold VI (1989) Mathematical Methods of Classical Mechanics, 1st edn 1978, 2nd edn 1989. Graduate Texts in Math, vol 60. Springer, New York
10. Arnold VI, Koslov VV, Neishtadt AI (1988) Dynamical Systems III. In: Encyclopedia of Mathematics, vol 3. Springer, New York
11. Atiyahf M, Bott R (1982) The Yang-Mills equations over Riemann surfaces. Phil Trans R Soc Lond A 308:523–615
12. Bates L, Lerman E (1997) Proper group actions and symplectic stratified spaces. Pac J Math 181:201–229
13. Bates L, Sniatycki J (1993) Nonholonomic reduction. Report Math Phys 32:99–115
14. Birtea P, Puta M, Ratiu TS, Tudoran R (2005) Symmetry breaking for toral actions in simple mechanical systems. J Differ Eq 216:282–323
15. Blankenstein G, Van Der Schaft AJ (2001) Symmetry and reduction in implicit generalized Hamiltonian systems. Rep Math Phys 47:57–100
16. Blaom AD (2000) Reconstruction phases via Poisson reduction. Differ Geom Appl 12:231–252
17. Blaom AD (2001) A geometric setting for Hamiltonian perturbation theory. Mem Am Math Soc 153(727):xviii+112
18. Bloch AM (2003) Nonholonomic mechanics and control. Interdisciplinary Applied Mathematics – Systems and Control, vol 24. Springer, New York. With the collaboration of Baillieul J, Crouch P, Marsden J, with scientific input from Krishnaprasad PS, Murray RM, Zenkov D
19. Bloch AM, Crouch P, Marsden JE, Ratiu T (2002) The symmetric representation of the rigid body equations and their discretization. Nonlinearity 15:1309–1341
20. Bloch AM, Krishnaprasad PS, Marsden JE, Murray R (1996) Nonholonomic mechanical systems with symmetry. Arch Ration Mech Anal 136:21–99
21. Bloch AM, Krishnaprasad PS, Marsden JE, Ratiu T (1996) The Euler-Poincaré equations and double bracket dissipation. Commun Math Phys 175:1–42
22. Bobenko AI, Reyman AG, Semenov-Tian-Shansky MA (1989) The Kowalewski Top 99 years later: A Lax pair, generalizations and explicit solutions. Commun Math Phys 122:321–354
23. Bobenko AI, YB Suris (1999) Discrete Lagrangian reduction, discrete Euler-Poincaré equations, and semidirect products. Lett Math Phys 49:79–93
24. Bourbaki N (1998) Lie groups and Lie algebras. In: Elements of Mathematics. Springer, Berlin, Chap 1–3, No MR1728312, 2001g:17006. Translated from the French, Reprint of the 1989 English translation

25. Bretherton FP (1970) A note on Hamilton's principle for perfect fluids. J Fluid Mech 44:19–31
26. Cartan E (1922) Leçons sur les Invariants Intégraux, 1971 edn. Hermann, Paris
27. Castrillón-López M, Garcia Pérez PL, Ratiu TS (2001) Euler-Poincaré reduction on principal bundles. Lett Math Phys 58:167–180
28. Castrillón-López M, Marsden JE (2003) Some remarks on Lagrangian and Poisson reduction for field theories. J Geom Phys 48:52–83
29. Castrillón-López M, Ratiu T (2003) Reduction in principal bundles: covariant Lagrange–Poincaré equations. Comm Math Phys 236:223–250
30. Castrillón-López M, Ratiu T, Shkoller S (2000) Reduction in principal fiber bundles: Covariant Euler-Poincaré equations. Proc Amer Math Soc 128:2155–2164
31. Cendra H, Holm DD, Hoyle MJW, Marsden JE (1998) The Maxwell–Vlasov equations in Euler-Poincaré form. J Math Phys 39:3138–3157
32. Cendra H, Holm DD, Marsden JE, Ratiu T (1998) Lagrangian Reduction, the Euler–Poincaré Equations and Semidirect Products. Amer Math Soc Transl 186:1–25
33. Cendra H, Ibort A, Marsden JE (1987) Variational principal fiber bundles: a geometric theory of Clebsch potentials and Lin constraints. J Geom Phys 4:183–206
34. Cendra H, Marsden JE (1987) Lin constraints, Clebsch potentials and variational principles. Physica D 27:63–89
35. Cendra H, Marsden JE, Ratiu TS (2001) Lagrangian reduction by stages. Mem Amer Math Soc 722:1–108
36. Cendra H, Marsden JE, Ratiu T (2001) Geometric mechanics, Lagrangian reduction and nonholonomic systems. In: Enquist B, Schmid W (eds) Mathematics Unlimited-2001 and Beyond. Springer, New York, pp 221–273
37. Cendra H, Marsden JE, Pekarsky S, Ratiu TS (2003) Variational principles for Lie-Poisson and Hamilton-Poincaré equations. Mosc Math J 3:833–867
38. Chang D, Bloch AM, Leonard N, Marsden JE, Woolsey C (2002) The equivalence of controlled Lagrangian and controlled Hamiltonian systems. Control Calc Var (special issue) 8: 393–422
39. Chernoff PR, Marsden JE (1974) Properties of Infinite Dimensional Hamiltonian systems. Lecture Notes in Mathematics, vol 425. Springer, New York
40. Chetayev NG (1941) On the equations of Poincaré. J Appl Math Mech 5:253–262
41. Chossat P, Lewis D, Ortega JP, Ratiu T (2003) Bifurcation of relative equilibria in mechanical systems with symmetry. Adv Appl Math 31:10–45
42. Chossat P, Ortega JP, Ratiu T (2002) Hamiltonian Hopf bifurcation with symmetry. Arch Ration Mech Anal 163:1–33; 167:83–84
43. Condevaux M, Dazord P, Molino P (1988) Geometrie du moment. Seminaire Sud-Rhodanien, Lyon
44. Cushman R, Bates L (1997) Global Aspects of Classical Integrable Systems. Birkhäuser, Boston
45. Cushman R, Rod D (1982) Reduction of the semi-simple 1:1 resonance. Physica D 6:105–112
46. Cushman R, Śniatycki J (2002) Nonholonomic reduction for free and proper actions. Regul Chaotic Dyn 7:61–72
47. Cushman R, Śniatycki J (1999) Hamiltonian mechanics on principal bundles. C R Math Acad Sci Soc R Can 21:60–64
48. Duistermaat J, Kolk J (1999) Lie Groups. Springer, New York
49. Ebin DG, Marsden JE (1970) Groups of diffeomorphisms and the motion of an incompressible fluid. Ann Math 92:102–163
50. Feynman R, Hibbs AR (1965) Quantum Mechanics and Path Integrals. McGraw-Hill, Murray Hill
51. Fischer AE, Marsden JE, Moncrief V (1980) The structure of the space of solutions of Einstein's equations, I: One Killing field. Ann Ins H Poincaré 33:147–194
52. Golubitsky M, Stewart I, Schaeffer D (1988) Singularities and Groups in Bifurcation Theory, vol 2. Applied Mathematical Sciences, vol 69. Springer, New York
53. Grabsi F, Montaldi J, Ortega JP (2004) Bifurcation and forced symmetry breaking in Hamiltonian systems. C R Acad Sci Paris Sér I Math 338:565–570
54. Guichardet A (1984) On rotation and vibration motions of molecules. Ann Inst H Poincaré 40:329–342
55. Guillemin V, Sternberg S (1978) On the equations of motions of a classic particle in a Yang-Mills field and the principle of general covariance. Hadronic J 1:1–32
56. Guillemin V, Sternberg S (1980) The moment map and collective motion. Ann Phys 1278:220–253
57. Guillemin V, Sternberg S (1984) Symplectic Techniques in Physics. Cambridge University Press, Cambridge
58. Hamel G (1904) Die Lagrange–Eulerschen Gleichungen der Mechanik. Z Math Phys 50:1–57
59. Hamel G (1949) Theoretische Mechanik. Springer, Heidelberg
60. Hernandez A, Marsden JE (2005) Regularization of the amended potential and the bifurcation of relative equilibria. J Nonlinear Sci 15:93–132
61. Holm DD, Marsden JE, Ratiu T (1998) The Euler–Poincaré equations and semidirect products with applications to continuum theories. Adv Math 137:1–81
62. Holm DD, Marsden JE, Ratiu T (2002) The Euler–Poincaré equations in geophysical fluid dynamics. In: Norbury J, Roulstone I (eds) Large-Scale Atmosphere-Ocean Dynamics II: Geometric Methods and Models. Cambridge Univ Press, Cambridge, pp 251–300
63. Holm DD, Marsden JE, Ratiu T, Weinstein A (1985) Nonlinear stability of fluid and plasma equilibria. Phys Rep 123:1–196
64. Hopf H (1931) Über die Abbildungen der dreidimensionalen Sphäre auf die Kugelfläche. Math Ann 104:38–63
65. Huebschmann J (1998) Smooth structures on certain moduli spaces for bundles on a surface. J Pure Appl Algebra 126: 183–221
66. Iwai T (1987) A geometric setting for classical molecular dynamics. Ann Inst Henri Poincaré Phys Theor 47:199–219
67. Iwai T (1990) On the Guichardet/Berry connection. Phys Lett A 149:341–344
68. Jalnapurkar S, Leok M, Marsden JE, West M (2006) Discrete Routh reduction. J Phys A Math Gen 39:5521–5544
69. Jalnapurkar S, Marsden J (2000) Reduction of Hamilton's variational principle. Dyn Stab Syst 15:287–318
70. Kane C, Marsden JE, Ortiz M, West M (2000) Variational integrators and the Newmark algorithm for conservative and dissipative mechanical systems. Int J Num Math Eng 49: 1295–1325
71. Kanso E, Marsden JE, Rowley CW, Melli-Huber J (2005) Locomotion of articulated bodies in a perfect fluid. J Nonlinear Sci 15:255–289
72. Kazhdan D, Kostant B, Sternberg S (1978) Hamiltonian group

actions and dynamical systems of Calogero type. Comm Pure Appl Math 31:481–508
73. Kirk V, Marsden JE, Silber M (1996) Branches of stable three-tori using Hamiltonian methods in Hopf bifurcation on a rhombic lattice. Dyn Stab Syst 11:267–302
74. Kobayashi S, Nomizu K (1963) Foundations of Differential Geometry. Wiley, New York
75. Koiller J (1992) Reduction of some classical nonholonomic systems with symmetry. Arch Ration Mech Anal 118:113–148
76. Koon WS, Marsden JE (1997) Optimal control for holonomic and nonholonomic mechanical systems with symmetry and Lagrangian reduction. SIAM J Control Optim 35:901–929
77. Koon WS, Marsden JE (1998) The Poisson reduction of nonholonomic mechanical systems. Rep Math Phys 42:101–134
78. Kostant B (1966) Orbits, symplectic structures and representation theory. In: Proc US-Japan Seminar on Diff Geom, vol 77. Nippon Hyronsha, Kyoto
79. Kummer M (1981) On the construction of the reduced phase space of a Hamiltonian system with symmetry. Indiana Univ Math J 30:281–291
80. Kummer M (1990) On resonant classical Hamiltonians with *n* frequencies. J Diff Eqns 83:220–243
81. Lagrange JL (1788) Mécanique Analytique. Chez la Veuve Desaint, Paris
82. Landsman NP (1995) Rieffel induction as generalized quantum Marsden-Weinstein reduction. J Geom Phys 15:285–319. Erratum: J Geom Phys 17:298
83. Landsman NP (1998) Mathematical topics between classical and quantum mechanics. J Geom Phys 17:298
84. Lerman E, Montgomery R, Jamaar RS (1993) Examples of singular reduction. In: Symplectic Geometry. London Math Soc Lecture Note Ser, vol 192. Cambridge Univ Press, Cambridge, pp 127–155
85. Lerman E, Singer SF (1998) Stability and persistence of relative equilibria at singular values of the moment map. Nonlinearity 11:1637–1649
86. Lerman E, Tokieda T (1999) On relative normal modes. C R Acad Sci Paris Sér I Math 328:413–418
87. Lew A, Marsden JE, Ortiz M, West M (2004) Variational time integration for mechanical systems. Int J Num Meth Eng 60:153–212
88. Lewis D, Marsden JE, Montgomery R, Ratiu T (1986) The Hamiltonian structure for dynamic free boundary problems. Physica D 18:391–404
89. Libermann P, Marle CM (1987) Symplectic Geometry and Analytical Mechanics. Kluwer, Dordrecht
90. Lie S (1890) Theorie der Transformationsgruppen. Zweiter Abschnitt. Teubner, Leipzig
91. Marle CM (1976) Symplectic manifolds, dynamical groups and Hamiltonian mechanics. In: Cahen M, Flato M (eds) Differential Geometry and Relativity. Reidel, Boston, pp 249–269
92. Marsden J, Ek GM, Ortega JP, Perlmutter M, Ratiu T (2007) Hamiltonian Reduction by Stages. Springer Lecture Notes in Mathematics, vol 1913. Springer, Heidelberg
93. Marsden J, Misiolek G, Perlmutter M, Ratiu T (1998) Symplectic reduction for semidirect products and central extensions. Diff Geom Appl 9:173–212
94. Marsden JE (1981) Lectures on Geometric Methods in Mathematical Physics. SIAM, Philadelphia
95. Marsden JE (1992) Lectures on Mechanics. London Mathematical Society Lecture Notes Series, vol 174. Cambridge University Press, Cambridge
96. Marsden JE, Hughes TJR (1983) Mathematical Foundations of Elasticity. Prentice Hall, Engelwood Cliffs. Reprinted 1994 by Dover
97. Marsden JE, Montgomery R, Morrison PJ, Thompson WB (1986) Covariant Poisson brackets for classical fields. Ann Phys 169:29–48
98. Marsden JE, Montgomery R, Ratiu T (1990) Reduction, symmetry and phases in mechanics. Memoirs of the AMS, vol 436. American Mathematical Society, Providence
99. Marsden JE, Ostrowski J (1996) Symmetries in motion: Geometric foundations of motion control. Nonlinear Sci Today http://link.springer-ny.com
100. Marsden JE, Patrick GW, Shkoller S (1998) Multisymplectic geometry, variational integrators and nonlinear PDEs. Comm Math Phys 199:351–395
101. Marsden JE, Pekarsky S, Shkoller S (1999) Discrete Euler–Poincaré and Lie–Poisson equations. Nonlinearity 12: 1647–1662
102. Marsden JE, Perlmutter M (2000) The orbit bundle picture of cotangent bundle reduction. C R Math Acad Sci Soc R Can 22:33–54
103. Marsden JE, Ratiu T (1986) Reduction of Poisson manifolds. Lett Math Phys 11:161–170
104. Marsden JE, Ratiu T (1994) Introduction to Mechanics and Symmetry. Texts in Applied Mathematics, vol 17. (1999) 2nd edn. Springer, New York
105. Marsden JE, Ratiu T, Scheurle J (2000) Reduction theory and the Lagrange-Routh equations. J Math Phys 41:3379–3429
106. Marsden JE, Ratiu T, Weinstein A (1984) Semi-direct products and reduction in mechanics. Trans Amer Math Soc 281: 147–177
107. Marsden JE, Ratiu T, Weinstein A (1984) Reduction and Hamiltonian structures on duals of semidirect product Lie Algebras. Contemp Math 28:55–100
108. Marsden JE, Scheurle J (1993) Lagrangian reduction and the double spherical pendulum. ZAMP 44:17–43
109. Marsden JE, Scheurle J (1993) The reduced Euler–Lagrange equations. Fields Inst Comm 1:139–164
110. Marsden JE, Weinstein A (1974) Reduction of symplectic manifolds with symmetry. Rep Math Phys 5:121–130
111. Marsden JE, Weinstein A (1982) The Hamiltonian structure of the Maxwell–Vlasov equations. Physica D 4:394–406
112. Marsden JE, Weinstein A (1983) Coadjoint orbits, vortices and Clebsch variables for incompressible fluids. Physica D 7:305–323
113. Marsden JE, West M (2001) Discrete mechanics and variational integrators. Acta Numerica 10:357–514
114. Marsden J, Weinstein A, Ratiu T, Schmid R, Spencer R (1982) Hamiltonian systems with symmetry, coadjoint orbits and plasma physics. In: Proc. IUTAM-IS1MM Symposium on Modern Developments in Analytical Mechanics, Torino, vol 117. Atti della Acad della Sc di Torino, pp 289–340
115. Martin JL (1959) Generalized classical dynamics and the "classical analogue" of a Fermi oscillation. Proc Roy Soc A 251:536
116. Mcduff D, Salamon D (1995) Introduction to Symplectic Topology. Oxford University Press, Oxford
117. Meyer KR (1973) Symmetries and integrals in mechanics. In: Peixoto M (ed) Dynamical Systems. Academic Press, New York, pp 259–273

118. Mielke A (1991) Hamiltonian and lagrangian flows on center manifolds, with applications to elliptic variational problems. Lecture Notes in Mathematics, vol 1489. Springer, Heidelberg
119. Mikami K, Weinstein A (1988) Moments and reduction for symplectic groupoid actions. Publ RIMS Kyoto Univ 24: 121–140
120. Montgomery R (1984) Canonical formulations of a particle in a Yang–Mills field. Lett Math Phys 8:59–67
121. Montgomery R (1986) The Bundle Picture in Mechanics. Ph D thesis, University of California Berkeley
122. Montgomery R (1988) The connection whose holonomy is the classical adiabatic angles of Hannay and Berry and its generalization to the non-integrable case. Comm Math Phys 120:269–294
123. Montgomery R (1990) Isoholonomic problems and some applications. Comm Math Phys 128:565–592
124. Montgomery R (1991) Optimal control of deformable bodies and Its relation to gauge theory. In: Ratiu T (ed) The Geometry of Hamiltonian Systems. Springer, New York, pp 403–438
125. Montgomery R (1991) How much does a rigid body rotate? A Berry's phase from the 18th century. Amer J Phys 59:394–398
126. Montgomery R (1993) Gauge theory of the falling cat. Fields Inst Commun 1:193–218
127. Montgomery R, Marsden JE, Ratiu T (1984) Gauged Lie-Poisson structures. In: Fluids and plasmas: geometry and dynamics. Boulder, 1983. American Mathematical Society, Providence, pp 101–114
128. Morrison PJ, Greene JM (1980) Noncanonical Hamiltonian density formulation of hydrodynamics and ideal magnetohydrodynamics. Phys Rev Lett 45:790–794. (1982) errata 48:569
129. Nambu Y (1973) Generalized Hamiltonian dynamics. Phys Rev D 7:2405–2412
130. Ortega JP (1998) Symmetry, Reduction, and Stability in Hamiltonian Systems. Ph D thesis, University of California, Santa Cruz
131. Ortega JP (2002) The symplectic reduced spaces of a Poisson action. C R Acad Sci Paris Sér I Math 334:999–1004
132. Ortega JP (2003) Relative normal modes for nonlinear Hamiltonian systems. Proc Royal Soc Edinb Sect A 133:665–704
133. Ortega JP, Planas-Bielsa V (2004) Dynamics on Leibniz manifolds. J Geom Phys 52:1–27
134. Ortega JP, Ratiu T (1997) Persistence and smoothness of critical relative elements in Hamiltonian systems with symmetry. C R Acad Sci Paris Sér I Math 325:1107–1111
135. Ortega JP, Ratiu T (1999) Non-linear stability of singular relative periodic orbits in Hamiltonian systems with symmetry. J Geom Phys 32:160–188
136. Ortega JP, Ratiu T (1999) Stability of Hamiltonian relative equilibria. Nonlinearity 12:693–720
137. Ortega JP, Ratiu T (2002) The optimal momentum map. In: Newton P, Holmes P, Weinstein A (eds) Geometry, Mechanics and Dynamics. Springer, New York, pp 329–362
138. Ortega JP, Ratiu T (2004) Momentum maps and Hamiltonian reduction. Progress in Mathematics, vol 222. Birkhäuser, Boston, pp xxxiv+497
139. Ortega JP, Ratiu T (2004) Relative equilibria near stable and unstable Hamiltonian relative equilibria. Proc Royal Soc Lond Ser A 460:1407–1431
140. Ortega JP, Ratiu T (2006) The reduced spaces of a symplectic Lie group action. Ann Glob Analysis Geom 30:335–381
141. Ortega JP, Ratiu T (2006) The stratified spaces of a symplectic Lie group action. Rep Math Phys 58:51–75
142. Ortega JP, Ratiu T (2006) Symmetry and symplectic reduction. In: Françoise JP, Naber G, Tsun TS (eds) Encyclopedia of Mathematical Physics. Elsevier, New York, pp 190–198
143. Otto M (1987) A reduction scheme for phase spaces with almost Kähler symmetry. Regularity results for momentum level sets. J Geom Phys 4:101–118
144. Palais RS (1957) A global formulation of the Lie theory of transformation groups. Mem Am Math Soc, vol 22. American Mathematical Society, Providence, pp iii+123
145. Patrick G (1992) Relative equilibria in Hamiltonian systems: The dynamic interpretation of nonlinear stability on a reduced phase space. J Geom and Phys 9:111–119
146. Patrick G, Roberts M, Wulff C (2004) Stability of Poisson equilibria and Hamiltonian relative equilibria by energy methods. Arch Ration Mech An 174:301–344
147. Pauli W (1953) On the Hamiltonian structure of non-local field theories. Il Nuovo Cim X:648–667
148. Pedroni M (1995) Equivalence of the Drinfelćd–Sokolov reduction to a bi-Hamiltonian reduction. Lett Math Phys 35:291–302
149. Perlmutter M, Ratiu T (2005) Gauged Poisson structures. Preprint
150. Perlmutter M, Rodríguez-Olmos M, Dias MS (2006) On the geometry of reduced cotangent bundles at zero momentum. J Geom Phys 57:571–596
151. Perlmutter M, Rodríguez-Olmos M, Dias MS (2007) On the symplectic normal space for cotangent lifted actions. Diff Geom Appl 26:277–297
152. Planas-Bielsa V (2004) Point reduction in almost symplectic manifolds. Rep Math Phys 54:295–308
153. Poincaré H (1901) Sur une forme nouvelle des équations de la méchanique. C R Acad Sci 132:369–371
154. Ratiu T (1980) The Euler–Poisson equations and integrability. Ph D thesis, University of California at Berkeley
155. Ratiu T (1980) Involution theorems. In: Kaiser G, Marsden J (eds) Geometric Methods in Mathematical Physics. Lecture Notes in Mathematics, vol 775. Springer, Berlin, pp 219–257
156. Ratiu T (1980) The motion of the free *n*-dimensional rigid body. Indiana Univ Math J 29:609–629
157. Ratiu T (1981) Euler–Poisson equations on Lie algebras and the *N*-dimensional heavy rigid body. Proc Natl Acad Sci USA 78:1327–1328
158. Ratiu T (1982) Euler–Poisson equations on Lie algebras and the *N*-dimensional heavy rigid body. Am J Math 104:409–448, 1337
159. Roberts M, de Sousa Dias M (1997) Bifurcations from relative equilibria of Hamiltonian systems. Nonlinearity 10:1719–1738
160. Roberts M, Wulff C, Lamb J (2002) Hamiltonian systems near relative equilibria. J Diff Eq 179:562–604
161. Routh EJ (1860) Treatise on the Dynamics of a System of Rigid Bodies. MacMillan, London
162. Routh EJ (1877) Stability of a given state of motion. Halsted Press, New York. Reprinted (1975) In: Fuller AT (ed) Stability of Motion
163. Routh EJ (1884) Advanced Rigid Dynamics. MacMillian, London
164. Satake I (1956) On a generalization of the notion of manifold. Proc Nat Acad Sci USA 42:359–363
165. Satzer WJ (1977) Canonical reduction of mechanical systems

invariant under Abelian group actions with an application to celestial mechanics. Ind Univ Math J 26:951–976
166. Simo JC, Lewis DR, Marsden JE (1991) Stability of relative equilibria I: The reduced energy momentum method. Arch Ration Mech Anal 115:15–59
167. Sjamaar R, Lerman E (1991) Stratified symplectic spaces and reduction. Ann Math 134:375–422
168. Smale S (1970) Topology and Mechanics. Inv Math 10:305–331, 11:45–64
169. Souriau JM (1970) Structure des Systemes Dynamiques. Dunod, Paris
170. Souriau J (1966) Quantification géométrique. Comm Math Phys 1:374–398
171. Sternberg S (1977) Minimal coupling and the symplectic mechanics of a classical particle in the presence of a Yang–Mills field. Proc Nat Acad Sci 74:5253–5254
172. Sudarshan ECG, Mukunda N (1974) Classical Mechanics: A Modern Perspective. Wiley, New York. (1983) 2nd edn. Krieber, Melbourne, Fl
173. Tulczyjew WM, Urbański P (1999) A slow and careful Legendre transformation for singular Lagrangians. Acta Phys Polon B 30:2909–2978. The Infeld Centennial Meeting, Warsaw, 1998
174. Vanhaecke P (1996) Integrable Systems in the Realm of Algebraic Geometry. Lecture Notes in Mathematics, vol 1638. Springer, New York
175. Weinstein A (1978) A universal phase space for particles in Yang–Mills fields. Lett Math Phys 2:417–420
176. Weinstein A (1983) Sophus Lie and symplectic geometry. Expo Math 1:95–96
177. Weinstein A (1996) Lagrangian mechanics and groupoids. Fields Inst Commun 7:207–231
178. Wendlandt JM, Marsden JE (1997) Mechanical integrators derived from a discrete variational principle. Physica D 106: 223–246
179. Whittaker E (1937) A Treatise on the Analytical Dynamics of Particles and Rigid Bodies, 4th edn. Cambridge University Press, Cambridge. (1904) 1st edn. (1937) 5th edn. (1944) Reprinted by Dover and (1988) 4th edn, Cambridge University Press
180. Wulff C (2003) Persistence of relative equilibria in Hamiltonian systems with non-compact symmetry. Nonlinearity 16:67–91
181. Wulff C, Roberts M (2002) Hamiltonian systems near relative periodic orbits. SIAM J Appl Dyn Syst 1:1–43
182. Zaalani N (1999) Phase space reduction and Poisson structure. J Math Phys 40:3431–3438

Mechanism Design

Ron Lavi
The Technion – Israel Institute of Technology, Haifa, Israel

Article Outline

Glossary
Definition of the Subject
Introduction
Formal Model and Early Results
Quasi-Linear Utilities and the VCG Mechanism
The Importance of the Domain's Dimensionality
Budget Balancedness and Bayesian Mechanism Design
Interdependent Valuations
Future Directions
Bibliography

Glossary

A social choice function A function that determines a social choice according to players' preferences over the different possible alternatives.

A mechanism A game in incomplete information, in which player strategies are based on their private preferences. A mechanism implements a social choice function f if the equilibrium strategies yield an outcome that coincides with f.

Dominant strategies An equilibrium concept where the strategy of each player maximizes her utility, no matter what strategies the other players choose.

Bayesian–Nash equilibrium An equilibrium concept that requires the strategy of each player to maximize the expected utility of the player, where the expectation is taken over the types of the other players.

VCG mechanisms A family of mechanisms that implement in dominant strategies the social choice function that maximizes the social welfare.

Definition of the Subject

Mechanism design is a sub-field of economics and game theory that studies the construction of social mechanisms in the presence of rational but selfish individuals (players/agents). The nature of the players dictates a basic contrast between the social planner, that aims to reach a socially desirable outcome, and the players, that care only about their own private utility. The underlying question is how to incentivize the players to cooperate, in order to reach the desirable social outcomes. A *mechanism* is a game, in which each agent is required to choose one action among a set of possible actions. The social designer then chooses an *outcome*, based on the chosen actions. This outcome is typically a coupling of a physical outcome, and a payment given to each individual. Mechanism design studies how to design the mechanism such that the *equilibrium* behavior of the players will lead to the socially desired goal. The theory of mechanism design has greatly influenced several sub-fields of micro-economics, for example auction theory and contract theory, and the 2007 Nobel prize in Economics was awarded to Leonid Hur-

wicz, Eric Maskin, and Roger Myerson "for having laid the foundations of mechanism design theory".

Introduction

It will be useful to start with an example of a mechanism design setting, the well-known "public project" problem (Clarke [8]): a government is trying to decide on a certain public project (the common example is "building a bridge"). The project costs C dollars, and each player, i, will benefit from it to an amount of v_i dollars, where this number is known only to the player herself. The government desires to build the bridge if and only if $\sum_i v_i > C$. But how should this condition be checked? Clearly, every player has an interest in over-stating its own v_i, if this report is not accompanied by any payment at all, and most probably agents will understate their values, if asked to pay some proportional amount to the declared value. Clarke describes an elegant mechanism that solves this problem. His mechanism has the fantastic property that, from the point of view of every player, no matter what the other players declare, it is always *in the best interest* of the player to declare his true value. Thus, truthful reporting is a dominant-strategy equilibrium of the mechanism, and under this equilibrium, the government's goal is fully achieved. A more formal treatment of this result is given in Sect. "Quasi-Linear Utilities and the VCG Mechanism" below.

Clarke's paper, published in the early 1970s, was part of a large body of work that started to investigate mechanism design questions. Most of the early works used two different assumptions about the structure of players' utilities. Under the assumption that utilities are general, and that the influence of monetary transfers on the utility are not well predicted, the literature have produced mainly impossibilities, which are described in Sect. "Formal Model and Early Results". The assumption that utilities are quasi-linear in money was successfully used to introduce positive and impressive results, as discussed in detail in Sects. "Quasi-Linear Utilities and the VCG Mechanism" and "The Importance of the Domain's Dimensionality". These mechanisms apply the solution concept of dominant-strategy equilibrium, which is a strong solution concept that may prevent several desirable properties from being achieved. To overcome its difficulties, the weaker concept of a Bayesian–Nash equilibrium is usually employed. This concept, and one main possibility result that it provides, are described in Sect. "Budget Balancedness and Bayesian Mechanism Design". The last important model that this entry covers aims to capture settings where the players' values are not fully observed by each player separately. Rather, each player receives a *signal* that gives a partial indication to her valuation. Mechanism design for such settings is discussed in Sect. "Interdependent Valuations".

One of the most impressive applications of the general mechanism design literature is auction theory. An auction is a specific form of a mechanism, where the outcome is simply the specific allocation of the goods to the players, plus the prices they are required to pay. Vickrey [28] initiated the study of auctions in a mechanism design setting, and in fact perhaps the study of mechanisms itself. After the fundamental study of general mechanism design in the 1970s, in the 1980s the focus of the research community returned to this important application, and many models were studied.

We note that there are several other entries in this book that are strongly related to the subject of "mechanism design". In particular, the entry on ► Game Theory, Introduction to gives a broader background on the mathematical methods and tools that are used by mechanism designers, and the entry on ► Implementation Theory handles similar subjects to this entry from a different point of view.

Formal Model and Early Results

A social designer wishes to choose one possible outcome/alternative out of a set A of possible alternatives. There are n players, each has her own preference order $\succeq_i$ over A. This preference order is termed the player's "type". The set (domain) of all valid player preferences is denoted by V_i. The designer has a social choice function $f\colon V_1 \times \cdots \times V_n \to A$, that specifies the desired alternative, given any profile of individual preferences over the alternatives. The problem is that these preferences are private information of each player – the social designer does not know them, and thus cannot simply invoke f in order to determine the social alternative. Players are assumed to be strategic, and therefore we are in a game-theoretic situation.

To *implement* the social choice function, the designer constructs a "game in incomplete information", as follows. Each player is required to choose an action out of a set of possible actions $\mathcal{A}_i$, and a target function $g\colon \mathcal{A}_1 \times \cdots \times \mathcal{A}_n \to A$ specifies the chosen alternative, as a function of the players' actions. A player's choice of action may, of-course, depend on her actual preference order. Furthermore, we assume an incomplete information setting, and therefore it cannot depend on any of the other players' preferences. Thus, to play the game, player i chooses a strategy $s_i\colon V_i \to \mathcal{A}_i$.

A strategy $s_i(\cdot)$ dominates another strategy $s'_i(\cdot)$ if, for every tuple of actions a_{-i} of the other players, and for every preference $\succeq_i \in V_i$, $g(s_i(\succeq_i), a_{-i}) \succeq_i g(a_i, a_{-i}))$, for any $a_i \in \mathcal{A}_i$. In other words, no matter what the other players are doing, the player cannot improve her situation by using an action other than $s_i(\succeq_i)$.

A mechanism *implements* the social choice function f in dominant strategies if there exist dominant strategies $s_1(\cdot), \ldots, s_n(\cdot)$ such that $f(\succeq_1, \ldots, \succeq_n) = g(s_1(\succeq_1), \ldots, s_n(\succeq_n))$, for any profile of preferences $\succeq_1, \ldots, \succeq_n$. In other words, a mechanism implements the social choice function f if, given that players indeed play their equilibrium strategies (in this case the dominant strategies equilibrium), the outcome of the mechanism coincides with f's choice.

The theory of mechanism design asks: given a specific problem domain (an alternative set and a domain of preferences), and a social choice function, how can we construct a mechanism that implements it (if at all)? As we shall see below, the literature uses a variety of "solution concepts", in addition to the concept of dominant strategies equilibrium, and an impressive set of understandings have emerged.

The concept of implementing a function with a dominant-strategy mechanism seems at first too strong, as it requires each player to know exactly what action to take, regardless of the actions the others take. Indeed, as we will next describe in detail, if we do not make any further assumptions then this notion yields mainly impossibilities. Nevertheless, it is not completely empty, and it may be useful to start with a positive example, to illustrate the new notions defined above.

Consider a voting scenario, where the society needs to choose one out of two candidates. Thus, the alternative set contains two alternatives ("candidate 1" and "candidate 2"), and each player either prefers 1 over 2, 2 over 1, or is indifferent between the two. It turns out that the majority voting rule is the dominant strategy implementable, by the following mechanism: each player reports her top candidate, and the candidate that is preferred by the majority of the players is chosen. This mechanism is a "direct-revelation" mechanism, in the sense that the action space of each player is to report a preference, and g is exactly f. In a direct-revelation mechanism, the hope is that truthful reporting (i. e. $s_i(\succeq_i) = \succeq_i$) is a dominant strategy. It is not hard to verify that in this two candidates setting, this is indeed the case, and hence the mechanism implements in dominant-strategies the majority voting rule.

An elegant generalization for the case of a "single-peaked" domain is as follows. Assume that the alternatives are numbered as $A = \{a_1, \ldots, a_n\}$, and the valid preferences of a player are single-peaked, in the sense that the preference order is completely determined by the choice of a peak alternative, a_{p}. Given the peak, the preference between any two alternatives a_i, a_j is determined according to their distance from a_{p}, i. e. $a_i \succeq_i a_j$ if and only if $|j - p| \leq |i - p|$. Now consider the social choice function $f(p_1, \ldots, p_n) = \mathrm{median}(p_1, \ldots, p_n)$, i. e. the chosen alternative is the median alternative of all peak alternatives.

Theorem 1 *Suppose that the domain of preferences is single-peaked. Then the median social choice function is implementable in dominant strategies.*

Proof Consider the direct revelation mechanism in which each player reports a peak alternative, and the mechanism outputs the median of all peaks. Let us argue that reporting the true peak alternative is a dominant strategy. Suppose the other players reported p_{-i}, and that the true peak of player i is p_i. Let p_{m} be the median index. If $p_i = p_{\mathrm{m}}$ then clearly player i cannot gain by declaring a different peak. Thus, assume that $p_i < p_{\mathrm{m}}$, and let us examine a false declaration p'_i of player i. If $p'_i \leq p_{\mathrm{m}}$ then p_{m} remains the median, and the player did not gain. If $p'_i > p_{\mathrm{m}}$ then the new median is $p'_m \geq p_{\mathrm{m}}$, and since $p_i < p_{\mathrm{m}}$, this is less preferred by i. Thus, player i cannot gain by declaring a false peak alternative if the true peak alternative is smaller or equal to the median alternative. A similar argument holds for the case of $p_i > p_{\mathrm{m}}$. □

In a voting situation with two candidates, the median rule becomes the same as the majority rule, and the domain is indeed single-peaked. When we have three or more candidates, it is not hard to verify that the majority rule is different than the median rule. In addition, one can also check that the direct-revelation mechanism that uses the majority rule does not have truthfulness as a dominant strategy.

Of-course, many times one cannot order the candidates on a line, and any preference ordering over the candidates is plausible. What voting rules are implementable in such a setting? This question was asked by Gibbard [12] and Satterthwaite [27], who provided a beautiful and fundamental impossibility. A domain of player preferences is unrestricted if it contains all possible preference orderings. In our voting example, for instance, the domain is unrestricted if every ordering of the candidates is valid (in contrast to the case of a single-peaked domain). A social choice function is dictatorial if it always chooses the top alternative of a certain fixed player (the dictator).

Theorem 2 ([12,27]) *Every social choice function over an unrestricted domain of preferences, with at least three alternatives, must be dictatorial.*

The proof of this theorem, and in fact of most other impossibility theorems in mechanism design, uses as a first step the powerful direct-revelation principle. Though the examples we have seen above use a direct revelation mechanism, one can try to construct "complicated" mechanisms with "crazy" action spaces and outcome functions, and by this obtain dominant strategies. How should one reason about such vast space of possible constructions? The revelation principle says that one cannot gain extra power by such complex constructions, since if there exists an implementation to a specific function then there exists a direct-revelation mechanism that implements it.

Theorem 3 (The Direct Revelation Principle) *Any implementable social choice function can also be implemented (using the same solution concept) by a direct-revelation mechanism.*

Proof Given a mechanism M that implements f, with dominant strategies $s_i^*(\cdot)$, we construct a direct revelation mechanism M' as follows: for any tuple of preferences $\succeq = (\succeq_1, \ldots, \succeq_n)$, $g'(\succeq) = g(s^*(\succeq))$. Since $s_i^*(\cdot)$ is a dominant strategy in M, we have that for any fixed $\succeq_{-i} \in V_{-i}$ and any $\succeq_i \in V_i$, the action $a_i = s_i^*(\succeq_i)$ is dominant when i's type is $\succeq_i$. Hence declaring any other type $\tilde{\succeq}_i$ that will "produce" an action $\tilde{a}_i = s_i^*(\tilde{\succeq}_i)$, cannot increase i's utility. Therefore, the strategy $\succeq_i$ in M' is dominant. □

The proof uses the dominant-strategies solution concept, but any other equilibrium definition will also work, using the same argumentation. Though technically very simple, the revelation principle is fundamental. It states that, when checking if a certain function is implementable, it is enough to check the direct-revelation mechanism that is associated with it. If it turns out to be truthful, we still may want to implement it with an indirect mechanism that will seem more natural and "real", but if the direct-revelation mechanism is not truthful, then there is no hope of implementing the function.

The proof of the theorem of Gibbard and Satterthwaite relies on the revelation principle to focus on direct-revelation mechanisms, but this is just the beginning. The next step is to show that *any* non-dictatorial function is non-implementable. The proof achieves this by an interesting reduction to Arrow's theorem, from the field of social choice theory. This theory is concerned with the possibilities and impossibilities of social preference aggregations that will exhibit desirable properties. A social welfare function $F\colon V \to \mathcal{R}$ aggregates the individuals' preferences into a single preference order over all alternatives, where $\mathcal{R}$ is the set of all possible preference orders over A. Arrow [2] describes few desirable properties from a social welfare function, and shows that no social choice function can satisfy all:

Definition 1 (Arrow's desirable properties)

1. A social welfare function satisfies "weak Pareto" if whenever all individuals strictly prefer alternative a to alternative b then, in the social preference, a is strictly preferred to b.
2. A social welfare function is "a dictatorship" if there exists an individual for which the social preference is always identical to his own preference.
3. A social welfare function F satisfies the "Independence of Irrelevant Alternatives" property (IIA) if, for any preference orders $R, \tilde{R} \in \mathcal{R}$ and any $a, b \in A$,

$$a >_{F(R)} b \text{ and } b >_{F(\tilde{R})} a \Rightarrow \exists i\colon\ a >_{R_i} b \text{ and } b >_{\tilde{R}_i} a$$

(where $a >_{R_i} b$ iff a is preferred over b in R_i). In other words, if the social preference between a and b was flipped when the individual preferences were changed from R to $\tilde{R}$, then it must be the case that some individual flipped his own preference between a and b.

Arrow's impossibility theorem holds for the *unrestricted* domain of preferences, i. e. when all preference orders are possible:

Theorem 4 ([2]) *Assume* $|A| \geq 3$. *Any social welfare function over an unrestricted domain of preferences that satisfies both weak Pareto and Independence of Irrelevant Alternatives must be a dictatorship.*

Gibbard and Satterthwaite's proof reveals an interesting and important connection between the concept of implementation in dominant strategies, and Arrow's condition of IIA. The proof shows how to construct, from a given implementable social choice function f, a social welfare function, F, that satisfies IIA and weak Pareto. In addition, F always places the alternative chosen by f as the most preferred alternative. By Arrow's theorem, the resulting social welfare function must be dictatorial. In turn, this implies that f is dictatorial. The construction of F from f is the straight-forward one: the top alternative is f's choice to the original preferences, say a. Then a is lowered to be the least preferred alternative in all the preferences, and f's new choice is placed second, etc. The interesting exercise is to show that the implementability of f implies that F satisfies Arrow's conditions. In fact, as the proof shows that any implementable social choice function f entails a social welfare function F that "extends" f and satisfies Arrow's conditions, it actually provides a strong argument for the reasonability of Arrow's requirement – they are simply implied by the implementability requirement.

In view of these strong impossibility results, it is natural to ask whether the entire concept of a mechanism can yield positive constructions. The answer is a big yes, under the "right" set of assumptions, as discussed in the next sections.

Quasi-Linear Utilities and the VCG Mechanism

The model formalization of the previous section ignores the existence of money, or, more accurately, the fact that it has a more or less predictable effect on a player's utility. The quasi-linear utilities model takes this into account, and players are assumed to have monetary value for each alternative.

Formally, the type of a player is a valuation function $v_i : A \to \Re$ that describes the monetary value that the player will obtain from each chosen alternative (as before v_i is taken from a domain of valid valuations V_i and $V = V_1 \times \cdots \times V_n$). Note that the value of a player does not depend on the other players' values (this is termed the private value assumption). The mechanism designer can now additionally pay each player (or charge money from her), and the total utility of player i if the chosen outcome is a and in addition she pays a price P_i is $v_i(a) - P_i$. A direct mechanism for quasi-linear utilities includes an outcome function $f : V \to A$ (as before), as well as price functions $p_i : V \to \Re$ for each player i (the definition of an indirect mechanism is the natural parallel of the definition of the previous section; the revelation principle holds for quasi-linear utilities as well, and we focus here on direct mechanisms). The implicit assumption is that a player aims to maximize her resulting utility, $v_i(f(v_i, v_{-i})) - p_i(v_i, v_{-i})$, and this leads us to the definition of a truthful mechanism, that parallels that of the previous section:

Definition 2 (Truthfulness, or Incentive Compatibility, or Strategy-Proofness) A direct revelation mechanism is "truthful" (or incentive-compatible, or strategy-proof) if the dominant strategy of each player is to reveal her true type, i. e. if for every $v_{-i} \in V_{-i}$ and every $v_i, v_i' \in V_i$,

$$v_i(f(v_i, v_{-i})) - p_i(v_i, v_{-i}) \geq v_i(f(v_i', v_{-i})) - p_i(v_i', v_{-i})$$

Using this framework, we can return to the example from Sect. "Introduction" ("building a bridge"), and construct a truthful mechanism to solve it. Recall that, in this problem, a government is trying to decide on a certain public project, which costs C dollars. Each player, i, will benefit from it to an amount of v_i dollars, where this number is known only to the player herself. The government desires to build the bridge if and only if $\sum_i v_i \geq C$. Clarke [8] designed the following mechanism. Each player reports a value, $\tilde{v}_i$, and the bridge is built if and only if $\sum_i \tilde{v}_i \geq C$. If the bridge is not built, the price of each player is 0. If the bridge is built then each player, i, pays the minimal value she could have declared to maintain the positive decision. More precisely, if $\sum_{i' \neq i} \tilde{v}_{i'} \geq C$ then she still pays zero, and otherwise she pays $C - \sum_{i' \neq i} \tilde{v}_{i'}$.

Theorem 5 *Bidding the true value is a dominant strategy in the Clarke mechanism.*

Proof Consider the truthful bidding for player i, v_i, vs. another possible bid $\tilde{v}_i$ (fixing the bids of the other players to arbitrarily be $\tilde{v}_{-i}$). If with v_i the project was rejected then $v_i < C - \sum_{i' \neq i} \tilde{v}_{i'}$. In order to change the decision to an accept, the player would need to declare $\tilde{v}_i \geq C - \sum_{i' \neq i} \tilde{v}_{i'}$. In this case i's payment will be $C - \sum_{i' \neq i} \tilde{v}_{i'}$ which is smaller than v_i, as observed above. Thus, i's resulting utility will be negative, hence bidding $\tilde{v}_i$ did not improve her utility.

On the other hand, assume that with v_i the project is accepted. Therefore, the player's utility from declaring v_i is non-negative. Note that the price that the player pays in case of an accept does not depend on her bid. Thus, the only way to change i's utility (if at all) is to declare some $\tilde{v}_i$ that will cause the project to be rejected. But in this case i's utility will be zero, hence she did not gain any benefit. □

Subsequently, Groves [13] made the remarkable observation that Clarke's mechanism is in fact a special case of a much more general mechanism, that solves the welfare maximization problem on *any* domain with private values and quasi-linear utilities. For a given set of player types $v_1, \ldots, v_n$, the *welfare* obtained by an alternative $a \in A$ is $\sum_i v_i(a)$. A social choice function is termed *a welfare maximizer* if $f(v)$ is an alternative with maximal welfare, i. e. $f(v) \in \operatorname{argmax}_{a \in A}\{\sum_{i=1}^n v_i(a)\}$.

Definition 3 (VCG Mechanisms) Given a set of alternatives A, and a domain of players' types $V = V_1 \times \ldots \times V_n$, a VCG mechanism is a direct revelation mechanism such that, for any $v \in V$,

1. $f(v) \in \operatorname{argmax}_{a \in A}\{\sum_{i=1}^n v_i(a)\}$.
2. $p_i(v) = -\sum_{j \neq i} v_j(f(v)) + h_i(v_{-i})$, where $h_i : V_{-i} \to \Re$ is an arbitrary function.

Ignore for a moment the term $h_i(v_{-i})$ in the payment functions. Then the VCG mechanism has a very natural interpretation: it chooses an alternative with maximal welfare according to the reported types, and then, by making additional payments, it equates the utility of each player to that maximal welfare level.

Theorem 6 ([13]) *Any VCG mechanism truthfully implements the welfare maximizing social choice function.*

Proof We argue that $s_i(v_i) = v_i$ is a dominant strategy for i. Fix any $v_{-i} \in V_{-i}$ as the declarations (actions) of the other players, any $v'_i \neq v_i$, and assume by contradiction that $v_i(f(v_i, v_{-i})) - p_i(v_i, v_{-i}) < v_i(f(v'_i, v_{-i})) - p_i(v'_i, v_{-i})$. Replacing $p_i(\cdot)$ with the specific VCG payment function, and eliminating the term $h_i(v_{-i})$ from both sides, we get: $v_i(f(v_i, v_{-i})) + \sum_{j\neq i} v_j(f(v_i, v_{-i})) < v_i(f(v'_i, v_{-i})) + \sum_{j\neq i} v_j(f(v'_i, v_{-i}))$. Therefore, it must be that $f(v_i, v_{-i}) \neq f(v'_i, v_{-i})$. Denote $f(v_i, v_{-i}) = a$ and $f(v'_i, v_{-i}) = b$. The above equation is now $v_i(a) + \sum_{j\neq i} v_j(a) < v_i(b) + \sum_{j\neq i} v_j(b)$, or, equivalently, $\sum_{i=1}^{n} v_i(a) < \sum_{i=1}^{n} v_i(b)$, a contradiction to the fact that $f(v_i, v_{-i}) = a$, since $f(\cdot)$ is a welfare maximizer. □

Thus, we see that the welfare maximizing social choice function can always be implemented, no matter what the problem domain is, under the assumption of quasi-linear utilities. The VCG mechanism is named after Vickrey, whose seminal paper [28] on auction theory was the first to describe a special case of the above mechanism (this is the second price auction; see the entry on auction theory for more details), after Clarke, who provided the second example, and after Groves himself, that finally pinned down the general idea.

Clarke's work can be viewed, in retrospect, as a suggestion for one specific form of the function $h_i(v_{-i})$, namely $h_i(v_{-i}) = \sum_{j\neq i} v_i(f(v_{-i}))$ (this is a slight abuse of notation, as f is defined for n players, but the intention is the straight-forward one – f chooses an alternative with maximal welfare). This form for the $h_i(\cdot)$'s gives the following property: if a player does not influence the social choice, her payment is zero, and, in general, a player pays the "monetary damage" to the other players (i. e. the welfare that the others lost) as a result of i's participation. Additionally, with Clarke's payments, a truthful player is guaranteed a non-negative utility, no matter what the others declare. This last property is termed "individual rationality".

The Importance of the Domain's Dimensionality

The impressive property of the VCG mechanism is its generality with respect to the domain of preferences – it can be used for *any* domain. On the other hand, VCG is restrictive in the sense that it can be used only to implement one specific goal, namely welfare maximization. Given the possibility that VCG presents, it is natural to ask if the assumption of quasi-linear utilities and private values allows the designer to implement many other different goals. It turns out that the answer depends on the "dimensionality" of the domain, as is discussed in this section.

Single-Dimensional Domains

Consider first a domain of preferences for which the type $v_i(\cdot)$ can be completely described by a single number v_i, in the following way. For each player i, a subset of the alternatives are "losing" alternatives, and her value for all these alternatives is always 0. The other alternatives are "winning" alternatives, and the value for each "winning" alternative is the same, regardless of the specific alternative. Such a domain is "single dimensional" in the sense that one single number completely describes the entire valuation vector. As before, this single number (the value for winning), is private to the player, and here this is the *only* private information of the player. The public project domain discussed above is an example of a single-dimensional domain: the losing alternative is the rejection of the project, and the winning alternative is the acceptance of the project.

A major drawback of the VCG mechanism, in general, and with respect to the public project domain in particular, is the fact that the sum of payments is not balanced (a broader discussion on this is given in Sect. "Budget Balancedness and Bayesian Mechanism Design" below). In particular, the payments for the public project domain may not cover the entire cost of the project. Is there a different mechanism that always covers the entire cost? The positive answer that we shall soon see crucially depends on the fact that the domain is single-dimensional, and this turns out to be true for many other problem domains as well.

The following mechanism for the public project problem assumes that the designer can decide not only if the project will be built, but also which players will be allowed to use it. Thus, we now have many possible alternatives, that correspond to the different subsets of players that will be allowed to utilize the project. This is still a single-dimensional domain, as each player only cares about whether she is losing or winning, and so the alternatives, from the point of view of a specific player, can be divided to the two winning/losing subsets. The following cost-sharing mechanism was proposed by Moulin [20] in a general cost-sharing framework. The mechanism is a direct-revelation mechanism, where each player, i, first submits her winning value, v_i. The mechanism then continues in rounds, where in the first round all players are present, and in each round one or more players are declared losers and retire. Suppose that in a certain round x players remain. If all remaining players have $v_i \geq C/x$ then they are declared winners, and each one pays C/x. Otherwise, all play-

ers with $v_i < C/x$ are declared losers, and "walk out", and the process repeats. If no players remain then the project is rejected.

Clearly, the cost sharing mechanism always recovers the cost of the project, if it is indeed accepted. But is it truthful? One can analyze it directly, to show that indeed the dominant strategy of each player is to declare her true winning value. Perhaps a better way is to understand a characterization of truthfulness for the general abstract setting of a single-dimensional domain. For simplicity, we will assume that we require mechanisms to be "normalized", i. e. that a losing player will pay exactly zero to the mechanism. Now, a mechanism is said to be "value-monotone" if a winner that increases her value will always remain a winner. More formally, for all $v_i \in V_i$ and $v_{-i} \in V_{-i}$, if i is a winner in the declaration (v_i, v_{-i}) then i is a winner in the declaration (v'_i, v_{-i}), for all $v'_i \geq v_i$. Note that a value-monotone mechanism casts a "threshold value" function $v^*_i(v_{-i})$ such that, for every v_{-i}, player i wins when declaring $v_i > v^*_i(v_{-i})$, and looses when declaring $v_i < v^*_i(v_{-i})$. Quite interestingly, this structure completely characterizes incentive compatibility in single-dimensional domains:

Theorem 7 *A normalized direct-revelation mechanism for a single-dimensional domain is truthful if and only if it is value monotone and the price of a winning player is* $v^*_i(v_{-i})$.

Proof The first observation is that the price of a winner cannot depend on her declaration, v_i (only on the fact that she wins, and on the declaration of the other players). Otherwise, if it can depend on her declaration, then there are two possible bids v_i and v'_i such that i wins with both bids and pays p_i and p'_i, where $p'_i < p_i$. But then if the true value of i is v_i then bidding v'_i instead of v_i will increase i's utility, contradicting truthfulness.

We now show that a truthful mechanism must be value-monotone. Assume by contradiction that a declaration of (v_i, v_{-i}) will cause i to win, but a declaration of (v'_i, v_{-i}) will cause i to lose, for some $v'_i > v_i$. Suppose that i pays p_i for winning (when the others declare v_{-i}). Since we assume a normalized mechanism, truthfulness implies that $p_i \leq v_i$. But then when the true type of a player is v'_i, her utility from declaring the truth will be zero (she loses), and she can increase her utility by declaring v_i, which will cause her to win and to pay p_i, a contradiction.

Thus, a truthful mechanism must be value-monotone, and there exists a threshold value $v^*_i(v_{-i})$. To see that this defines p_i, let us first check the case of $p_i < v^*_i(v_{-i})$. In this case, if the type of i is v_i with $p_i < v_i < v^*_i(v_{-i})$, she will lose (by the definition of $v^*_i(v_{-i})$), and by bidding some false large enough v'_i she can win and get a positive utility of $v_i - p_i$. On the other hand, if $p_i > v^*_i(v_{-i})$ then with type v_i such that $p_i > v_i > v^*_i(v_{-i})$ a player will have negative utility of $v_i - p_i$) from declaring the truth, and she can strictly increase it by losing, again a contradiction. Therefore, it must be that $p_i = v^*_i(v_{-i})$.

To conclude, it only remains to show that a value-monotone mechanism with a price for a winner $p_i = v^*_i(v_{-i})$ is indeed truthful. Suppose first that with the truthful declaration i wins. Then $v_i > v^*_i(v_{-i}) = p_i$ and i has a positive utility. If she changes the declaration and remains a winner, her price does not change, and if she becomes a loser her utility decreases to zero. Thus, a winner cannot increase her utility. Similarly, a loser can change her utility only by becoming a winner, i. e. by declaring $v'_i > v^*_i(v_{-i}) > v_i$, but since she will then pay $v^*_i(v_{-i})$ her utility will now decrease to be negative. Thus, a loser cannot increase her utility either, and the mechanism is therefore truthful. □

This structure of truthful mechanisms is very powerful, and reduces the mechanism design problem to the algorithmic problem of designing monotone social choice functions. Another strong implication of this structure is the fact that the payments of a truthful mechanism are completely derived from the social choice rule. Consequently, if two mechanisms always choose the same set of winners and losers, then the revenues that they raise must also be equal. Myerson [21] was perhaps the first to observe that, in the context of auctions, and named this the "revenue equivalence" theorem.

As a result of this characterization, one can easily verify that the above-mentioned cost-sharing mechanism is indeed truthful. It is not hard to check that the two conditions of the theorem hold, and therefore its truthfulness is concluded. This is just one example of the usefulness of the characterization.

Multi-Dimensional Domains

In the more general case, when the domain is multi-dimensional, the simple characterization from above does not fit, but it turns out that there exists a nice generalization. We describe two properties, cyclic monotonicity (Rochet [26]) and weak monotonicity (Bikhchandani et al. [7]), which achieve that. The exposition here also relies on [14]. It will be convenient to use the abstract social choice setting described above: there is a finite set A of alternatives, and each player has a type (valuation function) $v\colon A \to \Re$ that assigns a real number to every possible alternative. $v_i(a)$ should be interpreted as i's value for

alternative a. The valuation function $v_i(\cdot)$ belongs to the domain V_i of all possible valuation functions.

Our goal is to implement in dominant strategies the social choice function $f\colon V_1 \times \cdots \times V_n \to A$. As before, it is not hard to verify that the required price function of a player i may depend on her declaration only through the choice of the alternative, i. e. that it takes the form $p_i\colon V_{-i} \times A \to \Re$, for every player i. For truthfulness, these prices should satisfy the following property. Fix any $v_{-i} \in V_{-i}$, and any $v_i, v_i' \in V_i$. Suppose that $f(v_i, v_{-i}) = a$ and $f(v_i', v_{-i}) = b$. Then it is the case that:

$$v_i(a) - p_i(a, v_{-i}) \geq v_i(b) - p_i(b, v_{-i}) \tag{1}$$

In other words, player i's utility from declaring his true v_i is no less than his utility from declaring some lie, v_i', *no matter what the other players declare*. Given a social choice function f, the underlying question is what conditions should it satisfy to guarantee the existence of such prices.

Fix a player i, and fix the declarations of the others to v_{-i}. Let us assume, without loss of generality, that f is onto A (or, alternatively, define A' to be the range of $f(\cdot, v_{-i})$, and replace A with A' for the discussion below). Since the prices of Eq. (1) now become constant, we simply seek an assignment to the variables $\{p_a\}_{a\in A}$ such that $v_i(a) - v_i(b) \geq p_a - p_b$ for every $a, b \in A$ and $v_i \in V_i$ with $f(v_i, v_{-i}) = a$. This motivates the following definition:

$$\delta_{a,b} \doteq \inf\{v_i(a) - v_i(b) \mid v_i \in V_i, f(v_i, v_{-i}) = a\} \tag{2}$$

With this we can rephrase the above assignment problem, as follows. We seek an assignment to the variables $\{p_a\}_{a\in A}$ that satisfies:

$$p_a - p_b \leq \delta_{a,b} \quad \forall a, b \in A \tag{3}$$

By adding the two inequalities $p_a - p_b \leq \delta_{a,b}$ and $p_b - p_a \leq \delta_{b,a}$ we get that a necessary condition to the existence of such prices is the inequality $\delta_{a,b} + \delta_{b,a} \geq 0$. Note that this inequality is completely determined by the social choice function. This condition is termed the non-negative 2-cycle requirement. Similarly, for any k distinct alternatives $a_1, \ldots a_k$ we have the inequalities

$$\begin{aligned} p_{a_1} - p_{a_2} &\leq \delta_{a_1,a_2} \\ &\vdots \\ p_{a_{k-1}} - p_{a_k} &\leq \delta_{a_{k-1},a_k} \\ p_{a_k} - p_{a_1} &\leq \delta_{a_k,a_1} \end{aligned}$$

and we get that any *k-cycle* must be non-negative, i. e. that $\sum_{i=1}^{k} \delta_{a_i,a_{i+1}} \geq 0$, where $a_{k+1} \equiv a_1$. It turns out that this is also a sufficient condition:

Theorem 8 *There exists a feasible assignment to (3) if and only if there are no negative-length cycles.*

One constructive way to prove this is by looking at the "allocation graph": this is a directed weighted graph $G = (V, E)$ where $V = A$ and $E = A \times A$, and an edge $a \to b$ (for any $a, b \in A$) has weight $\delta_{a,b}$. A standard basic result of graph theory states that there exists a feasible assignment to (3) if and only if the allocation graph has no negative-length cycles. Furthermore, if all cycles are non-negative, the feasible assignment is as follows: set p_a to the length of the shortest path from a to some arbitrary fixed node $a^* \in A$.

With the above theorem, we can easily state a condition for implementability:

Definition 4 (Cycle Monotonicity) A social choice function f satisfies cycle monotonicity if for every player i, $v_{-i} \in V_{-i}$, some integer $k \leq |A|$, and $v_i^1, \ldots, v_i^k \in V_i$,

$$\sum_{j=1}^{k} [v_i^j(a_j) - v_i^j(a_{j+1})] \geq 0$$

where $a_j = f(v_i^j, v_{-i})$ for $1 \leq j \leq k$, and $a_{k+1} = a_1$.

Theorem 9 *f satisfies cycle monotonicity if and only if there are no negative cycles.*

Corollary 1 *A social choice function f is dominant-strategy implementable if and only if it satisfies cycle monotonicity.*

This interesting structure implies, as another corollary, the fact that the prices are uniquely determined by the social choice function, for every connected domain (this was discussed above for the special case of single-dimensional domains). Very briefly, from the above, it follows that any two alternatives with $\delta_{ab} + \delta_{ba} = 0$ have $p_a - p_b = \delta ab = -\delta_{ba}$. Thus, determining the price of one alternative completely determines the price of the second alternative. A short argument that we omit shows that the connectedness of the domain implies that for any two alternatives a and b, there's a path $a_1, \ldots, a_k$ (with $a_1 = a$ and $a_k = b$) such that $\delta_{a_i,a_{i+1}} + \delta_{a_{i+1},a_i} = 0$ for every $1 \leq i < k$. Thus, fixing the price of one alternative completely determines the prices of all other alternatives. In particular, if there exists one alternative whose price is normalized to be (always) zero, then all other prices have also been completely determined by the δ_{ab}'s weights (who in turn are completely determined by the function f).

Cycle monotonicity satisfies our motivating goal: a condition on f that involves only the properties of f, without existential price qualifiers. However, it is quite complex. k could be large, and a "shorter" condition would have been nicer. "Weak monotonicity" (W-MON) is exactly that:

Definition 5 (Weak Monotonicity) A social choice function f satisfies W-MON if for every player i, every v_{-i}, and every $v_i, v_i' \in V_i$ with $f(v_i, v_{-i}) = a$ and $f(v_i', v_{-i}) = b$, $v_i'(b) - v_i(b) \geq v_i'(a) - v_i(a)$.

In other words, if the outcome changes from a to b when i changes her type from v_i to v_i' then i's value for b has increased at least as i's value for a in the transition v_i to v_i'. W-MON is equivalent to cycle monotonicity with $k = 2$, or, alternatively, to the requirement of no negative 2-cycles. Hence it is necessary for truthfulness. As it turns out, it is also a sufficient condition on many domains. Very recently, Monderer [19] shows that weak monotonicity must imply cycle monotonicity if and only if the closure of the domain of valuations is convex. Thus, for such domains, it is enough to look at the more simple condition of weak monotonicity.

The Implementability of Non-Welfare-Maximizing Social Goals Now that the conditions for implementability are completely understood, it should be asked what forms of social choice functions satisfy them. We already saw that the welfare-maximizer function satisfies them, for any domain, and we ask what *other* implementable functions exist? For the single-dimensional case, we saw another example of a truthful mechanism, and the literature contains many more. For the multi-dimensional case, "interesting" examples are more rare, and a beautiful result by Roberts [25] shows that when the domain has full dimensionality then only weighted welfare maximizers are implementable. In other words, weak monotonicity implies welfare maximization. More precisely, a function f is an "affine maximizer" if there exist weights $k_1, \ldots, k_n$ and $\{C_x\}_{x \in A}$ such that, for all $v \in V$,

$$f(v) \in \operatorname{argmax}_{x \in A}\{\Sigma_{i=1}^{n} k_i v_i(x) + C_x\}$$

Roberts [25] shows that, if $|A| \geq 3$ and $V_i = \Re^A$ for all i, then f is dominant-strategy implementable if and only if it is an affine maximizer.

However, most interesting domains are restricted in some meaningful way, and for this wide intermediary range of domains the current knowledge is rather scarce. One impossibility result that extends the result of Roberts to a restricted multi-dimensional case is given by Lavi et al. [18], who study multi-item auctions. In a multi-item auction, one seller (the mechanism designer) wants to allocate items to players (i. e. an alternative is an allocation of the items to the players). [18] shows that every social choice function for multi-item auctions, that additionally satisfy four other social choice properties, must be an affine maximizer.

Before concluding the discussion on dominant-strategy implementation, we demonstrate the necessity for non-welfare-maximizers by considering the following "scheduling domain". A designer wishes to assign n tasks/jobs to m workers, where worker i needs t_{ij} time units to complete job task j, and incurs a cost of t_{ij} for its processing time (one dollar per time unit). Importantly, this cost is private information of the worker, and workers are assumed to be strategic, each one selfishly trying to minimize its own cost. The load of worker i is the sum of costs of the jobs assigned to her, and the maximal load over all workers (in a given schedule) is termed the "makespan" of the schedule. The welfare maximizing social goal would put each task on the most efficient worker (for that task), which may result in a very high makespan. For example, consider a setting with two workers and n tasks. The first worker incurs a cost of 1 for every task, and the second worker incurs a cost of $1 + \epsilon$ for every task. The social welfare is the minus of the sum of the costs of the two workers, and the VCG mechanism will therefore assign all tasks to the first worker. This is a very highly unbalanced allocation, which takes twice the time that the workers optimally need in order to finish all tasks (roughly splitting the work among them).

Thus, one may wish to consider a social goal different from welfare maximization, namely *makespan* minimization. This goal aims to construct a balanced allocation, in order to minimize the completion time of the last task. Such an allocation can also be viewed as being a more "fair" allocation, in the sense of Rawls' max-min fairness criteria. Because of the strategic nature of the workers, we wish to design a truthful mechanism. While VCG is truthful, its outcome may be far from optimal, as demonstrated above. Nisan and Ronen [23], who have first studied this problem in the context of mechanism design, observed that VCG provides only an "m-approximation" to the optimal makespan, meaning that VCG may sometimes produce a makespan that is m times larger than the optimal makespan. More importantly, they have shown that *no truthful deterministic mechanism can obtain an approximation ratio better than 2*. To date, the question of closing this gap between m and 2 remains open.

Archer and Tardos [1], on the other hand, considered a natural restriction of this domain, that makes it single-di-

mensional, and showed with this they can construct many possibilities (for example, a truthful *optimal* mechanism). Thus, here too we see the contrast between single-dimensionality and multi-dimensionality. Lavi and Swamy [17] suggest a multi-dimensional special case, and give a truthful 2-approximation for the special case where the processing time of each job is known to be either "low" or "high". This special case keeps the multi-dimensionality of the domain. The construction of this result does not rely on explicit prices, but rather uses the cycle-monotonicity condition described above, to construct a monotone allocation rule.

Budget Balancedness and Bayesian Mechanism Design

The previous sections portray a concrete picture of the advantages and the disadvantages of the solution concept of truthfulness in dominant strategies. On the one hand, this is a strong and convincing concept, which admits many positive results. However, there are several problems to all these results, that cannot be solved by a truthful mechanism. Among these, the budget-imbalance problem was briefly mentioned, and this section looks again at this problem, as a motivation to the definition of the Bayesian–Nash solution concept.

To recall the budget-imbalance problem of the VCG mechanism, let us consider a specific input to the Clarke mechanism from Sect. "Quasi-Linear Utilities and the VCG Mechanism": suppose the cost of the project is \$100, and there are 102 players, each values the project by \$1. It is a simple exercise to check that the Clarke mechanism will indeed choose to perform the project, and that each player will pay a price of zero (since the project would have been conducted even if a single player is removed). Thus, the mechanism designer does not cover the project's cost. As described above, this problem, for this specific domain, can be fixed by considering the cost-sharing mechanism discussed in Sect.. "The Importance of the Domain's Dimensionality". However, this mechanism may sometimes choose not to perform the project although the society as a whole will benefit from performing it (i. e. it is not "socially efficient"), and, even more importantly, it is a solution only for the concrete domain of a public project. Is there a general mechanism (in the sense that VCG is general) that is both socially efficient and budget-balanced? In this section we describe such a mechanism, that was independently discovered by d'Aspremont and Ge'rard–Varet [10] and by Arrow [3]. Its incentive compatibility will not be in dominant strategies. Instead, it is assumed that player types are drawn i.i.d. from some fixed and known cumulative distribution function F (the assumption that the types are drawn from the same distribution is not important, and is made here only for the ease of notation; the assumption that types are not correlated is important and cannot be removed in general). The solution concept of a Bayesian–Nash equilibrium is a natural extension of the regular Nash equilibrium concept, for a setting in which the distribution F is known to all players (this is termed the "common-prior" assumption), and where players aim to maximize the expectation of their quasi-linear utility.

Definition 6 A direct mechanism $M = (f, p)$ is Bayesian incentive compatible if for every player i, and for every $v_i, v_i' \in V_i$,

$$E_{v_{-i}}[v_i(f(v_i, v_{-i})) - p_i(v_i, v_{-i})] \geq E_{v_{-i}}[v_i(f(v_i', v_{-i})) - p_i(v_i', v_{-i})]$$

In other words, Bayesian incentive compatibility requires that a player will maximize her expected utility by declaring her true type. An alternative formulation is that truthfulness in a Bayesian incentive compatible mechanism should be a "Bayesian–Nash equilibrium" (where the formal equilibrium definition naturally follows the above definition). This is an "ex-interim" equilibrium: the type of the player is already known to her, and the averaging is over the types of the others. A weaker equilibrium notion would be an "ex-ante" notion, where the player should decide on a strategy before knowing her own type, and so the averaging is done over her own types as well. A stronger notion would be an "ex-post" notion, where no-averaging is done at all, and the above inequality is required for every realization of the types of the other players. It can be shown that this stronger ex-post condition is equivalent to the requirement of dominant-strategy incentive compatibility. As a Bayesian–Nash equilibrium only considers the average over all possible realizations, it is clearly a weaker requirement than dominant-strategy implementability.

We will demonstrate the usefulness of this weaker notion by describing a general mechanism that is both ex-post socially efficient and ex-post budget balanced, and is Bayesian incentive-compatible. Define,

$$x_i(v_i) = E_{v_{-i}}\left[\sum_{j \neq i} v_j(f(v_i, v_{-i}))\right]$$

The "budget-balanced" (BB) mechanism asks the players to report their types, and then chooses the welfare-maximizing allocation according to the reported types (as VCG does). It then charges some payment $p_i(v_i, v_{-i}) =$

$-x_i(v_i)+h_i(v_{-i})$, for some function $h_i(\cdot)$ that will be chosen later on in a specific way that balances the budget. But let us first verify that the mechanism is Bayesian incentive compatible, regardless of the choice of the functions $h_i(\cdot)$. Note that, for any realization of v_{-i}, we have that,

$$v_i(f(v_i, v_{-i})) + \sum_{j\neq i} v_j(f(v_i, v_{-i})) \geq v_i(f(v_i', v_{-i})) + \sum_{j\neq i} v_j(f(v_i', v_{-i}))$$

as the mechanism chooses the maximal-welfare alternative for the given reports. Clearly, taking the expectation on both sides will maintain the inequality. Therefore we get:

$$\begin{aligned} &E_{v_{-i}}[v_i(f(v_i, v_{-i})) - p_i(v_i, v_{-i})] \\ &\quad = E_{v_{-i}}[v_i(f(v_i, v_{-i}))] + E_{v_{-i}}\left[\sum_{j\neq i} v_j(f(v_i, v_{-i}))\right] \\ &\qquad + E_{v_{-i}}[h_i(v_{-i})] \\ &\quad \geq E_{v_{-i}}[v_i(f(v_i, v_{-i}))] + E_{v_{-i}}\left[\sum_{j\neq i} v_j(f(v_i, v_{-i}))\right] \\ &\qquad + E_{v_{-i}}[h_i(v_{-i})] \\ &\quad = E_{v_{-i}}[v_i(f(v_i', v_{-i})) - p_i(v_i', v_{-i})] \end{aligned}$$

which proves Bayesian incentive compatibility. To balance the budget, consider the specific function, $h_i(v_{-i}) = 1/(n-1)\sum_{j\neq i} x_j(v_j)$. Notice that the term $x_j(v_j)$ appears $(n-1)$ times in the sum $\sum_{i=1}^n h_i(v_{-i})$ for any $j = 1, \ldots, n$. Therefore $\sum_{i=1}^n h_i(v_{-i}) = 1/(n-1)\sum_{j=1}^n (n-1)x_j(v_j) = \sum_{i=1}^n x_i(v_i)$. To conclude, we have $\sum_{i=1}^n p_i(v_i, v_{-i}) = \sum_{i=1}^n h_i(v_{-i}) - \sum_{i=1}^n x_i(v_i) = 0$, and the budget balancedness follows.

It is worth noting that such an exercise cannot be employed for the VCG mechanism, as there the "parallel" $x_i(\cdot)$ term should depend on the entire vector of declarations, not only on i's own declarations. This is the exact point where the averaging of the others' valuations is crucial.

In addition to the difference in the solution concept, one other important advantage of VCG, in comparison with the BB mechanism, is the fact that VCG (with the Clarke payments) is ex-post "individually rational": if a player declares her true valuation, it is guaranteed that she will not pay more than her value, no matter what the others will declare. Here, on the contrary, there is no reason why this should be true, in general. Can the solution concept of Bayesian incentive compatibility be used to construct a general budget-balanced and individually rational mechanism? In an important and influencing result, Myerson and Satterthwaite [22] have shown that this is impossible: there is no general mechanism that satisfies the four properties (1) Bayesian incentive compatibility, (2) budget balancedness, (3) individual rationality, and (4) social efficiency. The proof uses a simple, natural exchange setting, where two traders (one buyer and one seller) wish to exchange an item. The seller has a cost c of producing the item, and the buyer obtains a value v from receiving it. Myerson and Satterthwaite show that there is no Bayesian incentive compatible mechanism that decides to perform the exchange if and only if $v > c$, such that Bayesian incentive compatibility and individual rationality are maintained, and the price that the buyer pays exactly equals the payment that the seller gets. In particular, VCG violates this last property, while BB satisfies it, but violates individual rationality (i. e. for some realizations of the values, a buyer may pay more than her value, or the seller may get less than her cost).

Besides this disadvantage of the BB mechanism, there are also additional disadvantages that result from the underlying assumptions of the solution concept itself. In particular, Bayesian incentive compatibility entails two strong assumptions about the characteristics of the players. First, it assumes that players are risk-neutral, i. e. care only about maximizing the expectation of their profit (value minus price). Thus, when players dislike risk, for example, and prefer to decrease the variance of the outcome, even on the expense of lowering the achieved expected profit, the rational of the Bayesian–Nash equilibrium concept breaks down. Second, the assumption of a common-prior, i. e. that all players agree on the same underlying distribution, seems strong and somewhat unrealistic. Often, players have different estimations about the underlying statistical characteristics of the environment, and this concept does not handle this well. Note that the solution concept of dominant-strategies does not suffer from any of these problems, which strengthens even more its importance. Unfortunately, the classical economics literature mainly ignores these disadvantages and problems. A well-known exception is the critique known as Wilson's critique [29], who raises the above mentioned problems, and argues in favor of "detail-free" mechanisms. Recently, this critique gained more popularity, and detail-free solution concepts are re-examined. For some examples, see [5,6,11].

Interdependent Valuations

Up to now, this entry described "private value" models, i. e. models where the valuation (or the preference relation) of

a player does not depend on the types of the other players. There are many settings in which this assumption is unrealistic, and a more suitable assumption is that the valuation of a specific player is affected by the valuations of the other players. This last statement may entail two interpretations. The first is that the distribution over the valuations of a specific player is correlated with the distribution over the valuations of the other players, and, thus, knowing a player's actual valuation gives partial knowledge about the valuations of the other players. This first interpretation is still termed a private value model (but with correlated values instead of independent values), since after the player becomes aware of the actual realization of her valuation, she completely and fully knows her values for the different outcomes.

In contrast, with interdependent valuations, the actual valuation of a player depends on the actual valuations of the other players. Thus, a player does not fully know her own valuation. She only partially knows it, and can determine her full valuation only if given the others' valuations as well. A classic example is a setting where a seller sells an oil field. The oil, of-course, is not seen on the ground surface, and the only way to exactly determine how much oil is there (and, by this, determine the actual worth of the field) is to extract it. Before buying the field, though, the potential buyers are only allowed to make preliminary tests, and by this to determine an estimation of the value of the field, which is not completely accurate. If all the buyers that are interested in the field have the same technical capabilities, it seems reasonable to assume that the *true value* of the field is the average over all the estimations obtained by the different oil companies. Intuitively, a player that participates in an auction mechanism that determines who will buy the field, and at what price, has to act somehow as if she knows the value of the field, although she doesn't. Clearly, this creates different complications. Such a model is very natural in auction settings, and indeed the entry on auctions handles the subject of interdependent valuations more broadly. Since this issue is also very relevant to general mechanism design theory, we describe here one specific, rather general result for mechanisms with interdependent valuations, to exemplify the definitions and the techniques being employed.

In the formal model of interdependent valuations, player i receives a signal $s_i \in S_i$, which may be multi-dimensional. Her valuation for a specific alternative $a \in A$ is a function of the signals $s_1, \ldots, s_n$, i. e. $v_i \colon A \times S_1 \times \cdots \times S_n \to \Re$. The case where $v_i(a, s_1, \ldots, s_n) = v_j(a, s_1, \ldots, s_n)$ for all players i, j and all $a, s_1, \ldots, s_n$ is termed the "common value" case, as the actual values of all players are identical, and only their signals are different (as in the oil field example). The other extreme is when i's valuation depends only on i's signal, i. e. $v_i(a, s_1, \ldots, s_n) = v_i(a, s_i)$, which is a return to the private value case. The entire range in general is termed the case of interdependent valuations. All the results described in the previous sections fail when we move to interdependent valuations. For example, in the VCG mechanism, a player is required to report her valuation function, which is not fully known to her in the interdependent valuation case. It turns out that the straightforward modification of reporting the players' signals does not maintain the truthfulness property, and, in fact, some strong impossibilities exist (Jehiel et al. [15]). However, interdependent valuations may also enable possibilities, and the classic result of Cremer and McLean [9] will be described here to exemplify this. This result shows how to use the interdependencies in order to increase the revenue of the mechanism designer, so that the entire surplus of the players can be extracted. [9] study an auction setting where there is one item for sale. n bidders have interdependent values for the item, and it is assumed that the signal that each player receives is single-dimensional, i. e. each player receives a single real number as her signal. The valuation functions are assumed to be known to the mechanism designer, so that the only private information of the players are their signals. It is also assumed that the valuation functions are monotonically non-decreasing in the signals. For simplicity, it is assumed here that the signal space is discretisized to be $S_i = \{0, \Delta, 2\Delta, \ldots\}$. The last (and crucial) assumption is that the valuation functions satisfy the "single-crossing" property: if $v_i(s_i, s_{-i}) \geq v_j(s_i, s_{-i})$ then $v_i(s_i + \Delta, s_{-i}) \geq v_j(s_i + \Delta, s_{-i})$. This says that i's signal affects i's own value (weakly) more than it affects the value of any other player. This last assumption is strong, but in some sense necessary, as it is possible to construct interdependent valuation functions (that violate single-crossing) for which no truthful mechanism can be efficient (i. e. allocate the item to the player with the highest value).

Consider the following CM mechanism for this problem: each player reports her signal, and the player with the highest *value* (note that this may be different than the player with the highest signal) receives the object. In order to determine her payment, define the "threshold signal" $T_i(s_{-i})$ of any player i to be the minimal signal that will enable her to win (given the signals of the other players), i. e. $T_i(s_{-i}) = \min\{\tilde{s}_i \in S_i \mid v_i(\tilde{s}_i, s_{-i}) \geq \max_{j \neq i} v_j(\tilde{s}_i, s_{-i})\}$. The payment of the winner, i, is her value if her signal was $T_i(s_{-i})$, i. e. $P_i(s_{-i}) = v_i(T_i(s_{-i}), s_{-i})$. Clearly, if all players report their true signals, then the player with the highest value receives the item. Truthful reporting is also an ex-post Nash equilibrium, which means the following: if all

other players report the true signal (no matter what that is) then it is a best response for i to report her true signal as well.

To verify that truthfulness is indeed an ex-post Nash equilibrium, notice first that each player has a price for winning which does not depend on her declaration. Now, truthful reporting will ensure winning (given that the others are truthful as well) if and only if the true value of the player is higher than her price (i. e. iff winning will yield a positive utility). Thus, when a player "wants to win", truthful reporting will do that, and when a player "wants to lose", truthful reporting will do that as well, and so truthfulness will always maximize the player's utility.

The notion of an ex-post equilibrium is stronger than Bayesian–Nash equilibrium, since, here, even after the signals are revealed no player regrets her declaration (while in Bayesian–Nash equilibrium, since only the *expected* utility is maximized, there are some realizations for which a player can deviate and gain). On the other hand, ex-post equilibrium is weaker than dominant strategies, in which truthfulness is the best strategy no matter what the others choose to declare, while here truthfulness is a best response only if the others are truthful as well.

As seen above, both for the VCG mechanism as well as for the BB mechanism, adding a "constant" to the prices (i. e. setting $\tilde{P}_i(s_{-i}) = P_i + h_i(s_{-i})$) maintains the strategic properties of the mechanism, since the function $h_i(\cdot)$ does not depend on the declaration of player i. The correlation in the values can help the mechanism designer to extract more payments from the players, as follows. Consider the matrix that describes the conditional probability for a specific tuple of signals of the other players, given i's own signal. There is a row for every signal s_i of i, a column for every tuple of signals s_{-i} of the other players, and the cell (s_i, s_{-i}) contains the conditional probability $Pr(s_{-i}|s_i)$. In the private value case, the signals of the players are not correlated, hence the matrix has rank one (all rows are identical). As the correlation between the signals "increases", the rank increases, and we consider here the case when the matrix has full row rank. Let $q_i(s_i, s_{-i})$ be an indicator to the event that i is the winner when the signals are (s_i, s_{-i}). The expected surplus of player i in the CM mechanism is $U_i^*(s_i) = \sum_{s_{-i}} Pr(s_{-i}|s_i) \cdot (q_i(s_i, s_{-i}) \cdot v_i(s_i, s_{-i})) - P_i(s_{-i}))$. ($P_i(s_{-i})$ is defined to be zero whenever i is not a winner). Now find "constants" $h_i(s_{-i})$ such that, for every s_i, $\sum_{s_{-i}} h_i(s_{-i}) \cdot Pr(s_{-i}|s_i) = U_i^*(s_i)$. Note that such an $h_i(\cdot)$ function exists: we have a system of linear equations, where the variables are the function values $h_i(s_{-i})$ for all possible tuples s_{-i}, and the qualifiers are the probabilities and the expected surplus. Since the matrix of qualifies has full row rank, a solution exists. It is now not hard to verify that, with prices $\tilde{P}_i(\cdot)$, the expected utility of a truthful player is zero.

As mentioned above, truthfulness is still an ex-post equilibrium of this mechanism. It is not ex-post individually rational, though, but rather only ex-ante, since a player pays her expected surplus even if the actual signals cause her to lose. Thus, this mechanism can be considered a fair lottery. Also note that the crucial property was the correlation between the values, the interdependence assumption was not important.

Future Directions

As surveyed here, the last three decades have seen the theory of mechanism design being developed in many different directions. The common thread of all settings is the requirement to implement some social goal in the presence of incomplete information – the social designer does not know the players' preferences for the different outcomes. We have seen several alternative assumptions about the structure of players' preferences, the different equilibria solution concepts that are suitable for the different cases, and several positive examples for elegant solutions. We have also discussed some impossibilities, demonstrating that some attractive definitions may turn out almost powerless. One relatively new research direction in mechanism design is the analysis of new models for the emerging Internet economy, and the development of new alternative solution concepts that better suit this setting. A very recent example is the new model of "dynamic mechanism design", where the parameters of the problem (e. g. the number of players, or their types) vary over time. Such settings become more and more important as the economic environment becomes more dynamic, for example due to the growing importance of the electronic markets. Examples for such models include e. g. the works by Lavi and Nisan [16] in the context of computer science models, and by Athey and Segal [4] in a more classical economic context, among many other works that study such dynamic settings.

The Internet environment also strengthens the question marks posed on the solution concept of Bayesian incentive compatibility, which was the most common solution concept in mechanism design literature in the 1980s and throughout the 1990s, due to the accompanying assumption of a common prior. Such an assumption seems problematic in general, and in particular in an environment like the Internet, that brings together players from many different parts of the world. It seems that the research community agrees more and more that alternative, detail-free solution concepts should be sought. The de-

scription of more recently new solution concepts is beyond the scope of this entry, and the interested reader is referred, for example, to the papers by [5,6,11] for some recent examples.

Another aspect of mechanism design that is largely ignored in the classic research is the computational feasibility of the mechanisms being suggested. This question is not just a technicality – some classic mechanisms imply heavy computational and communicational requirements that scale exponentially as the number of players increase, making them completely infeasible for even moderate numbers of players. The computer science community has begun looking at the design of computationally efficient mechanisms, and the recent book by Nisan et al. [24] contains several surveys on the subject.

Bibliography

1. Archer A, Tardos E (2001) Truthful mechanisms for one-parameter agents. In: Proceedings of the 42st Annual Symposium on Foundations of Computer Science, FOCS'01, Las Vegas. IEEE Computer Society
2. Arrow K (1951) Social Choice and Individual Values. Wiley, New York
3. Arrow K (1979) The property rights doctrine and demand revelation under incomplete information. In: Boskin M (ed) Economies and Human Welfare. Academic Press, New York
4. Athey S, Segal I (2007) Designing dynamic mechanisms. Am Econ Rev 97(2):131–136
5. Babaioff M, Lavi R, Pavlov E (2006) Single-value combinatorial auctions and implementation in undominated strategies. In: Proceedings of the 17th Symposium on Discrete Algorithms, SODA, Miami. ACM Press
6. Bergemann D, Morris S (2005) Robust mechanism design. Econometrica 73:1771–1813
7. Bikhchandani S, Chatterjee S, Lavi R, Mu'alem A, Nisan N, Sen A (2006) Weak monotonicity characterizes deterministic dominant-strategy implementation. Econometrica 74(4): 1109–1132
8. Clarke E (1971) Multipart pricing of public goods. Public Choice 8:17–33
9. Cremer J, McLean R (1985) Optimal selling strategies under uncertainty for a discriminating monopolist when demands are interdependent. Econometrica 53:345–361
10. d'Aspremont C, Gérard-Varet L (1979) Incentives and incomplete information. J Public Econ 11:25–45
11. Dekel E, Wolinsky A (2003) Rationalizable outcomes of large private-value first-price discrete auctions. Games Econ Behav 43(2):175–188
12. Gibbard A (1973) Manipulation of voting schemes: A general result. Econometrica 41(4):587–601
13. Groves T (1973) Incentives in teams. Econometrica 41(4): 617–631
14. Gui H, Muller R, Vohra RV (2004) Characterizing dominant strategy mechanisms with multi-dimensional types. Working paper, unpublished
15. Jehiel P, Meyer ter Vehn M, Moldovanu B, Zame WR (2006) The limits of ex-post implementation. Econometrica 74(3):585–610
16. Lavi R, Nisan N (2004) Competitive analysis of incentive compatible on-line auctions. Theor Comput Sci 310:159–180
17. Lavi R, Swamy C (2007) Truthful mechanism design for multidimensional scheduling. In: The Proceedings of the 8th ACM Conference on Electronic Commerce, EC'07, San Diego. ACM Press
18. Lavi R, Mu'alem A, Nisan N (2003) Towards a characterization of truthful combinatorial auctions. In: Proceedings of the 44rd Annual Symposium on Foundations of Computer Science, FOCS'03, Cambridge. IEEE Computer Society
19. Monderer D (2007) Monotonicity and implementability. Working paper, unpublished
20. Moulin H (1999) Incremental cost sharing: Characterization by coalition strategy-proofness. Soc Choice Welf 16:279–320
21. Myerson R (1981) Optimal auction design. Math Oper Res 6: 58–73
22. Myerson R, Satterthwaite M (1983) Efficient mechanisms for bilateral trading. J Econ Theor 29:265–281
23. Nisan N, Ronen A (2001) Algorithmic mechanism design. Games Econ Behav 35:166–196
24. Nisan N, Roughgarden T, Tardos E, Vazirani VV (eds) (2007) Algorithmic Game Theory. Cambridge University Press, New York
25. Roberts K (1979) The characterization of implementable choise rules. In: Laffont JJ (ed) Aggregation and Revelation of Preferences. North-Holland, Amsterdam, pp 321–349
26. Rochet JC (1987) A necessary and sufficient condition for rationalizability in a quasilinear context. J Math Econ 16:191–200
27. Satterthwaite M (1975) Strategy-proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. J Econ Theor 10:187–217
28. Vickrey W (1961) Counterspeculations, auctions, and competitive sealed tenders. J Financ 16:8–37
29. Wilson R (1987) Game-theoretic analyses of trading processes. In: Bewley T (ed) Advances in Economic Theory: Fifth World Congress. Cambridge University Press, New York, pp 33–70

Membrane Computing

Gheorghe Păun
Institute of Mathematics of the Romanian Academy, Bucureşti, Romania

Article Outline

Glossary
Definition of the Subject
Introduction
Cell-like P Systems
Ingredients of P Systems
Spiking Neural P Systems
Computing Power
Computational Efficiency
Applications
Future Directions
Bibliography

Glossary

Active membranes P systems whose rules involve not only objects, but also membranes, are called systems with *active membranes*. By using such rules one can both specify the way the objects can be processed and moved (for instance, the rule $a[_i]_i \to [_i b]_i$ says that object a can enter membrane with label i and gets transformed into b), and the way the membrane structure itself can be transformed (membranes can be dissolved, created, divided, merged, etc.).

Evolution rule The multisets of objects evolve by means of rules corresponding to chemical reactions and to other processes taking place in a cell. The abstract counterpart of a chemical reaction is a multiset rewriting rule, of the form $u \to v$, where u and v are strings representing multisets. For instance, $aad \to abc$ is a rule with the meaning that 2 copies of object a react with one copy of object d, these copies are consumed, and as a result we produce one copy of each of a, b, and c (hence one copy of a is reproduced). Such a rule is said to be cooperative (several objects react); particular cases are those of catalytic rules, of the form $ca \to cv$, with c being a catalyst which assists object a in getting transformed in the multiset v, and of non-cooperative rules, where only one object evolves, $a \to u$ (a is an object and u a multiset).

Besides such multiset rewriting rules, there are many other types of rules corresponding to bio-chemical processes or mathematically inspired: rules for handling membranes (create, merge, divide, etc.), to move objects across membranes, and combinations of rules.

Halting Successful computations are usually defined in terms of halting: a computation stops when it reaches a configuration where no rule can be applied in the whole system. A weaker condition is to consider the computation finished when at least one compartment of the system cannot apply any of its rules (this is called local halting).

Membrane structure The cells are separated from the environment by a membrane; the internal compartments of a cell (nucleus, mitochondria, Golgi apparatus, vesicles, etc.) are also delimited by membranes. A membrane is a *separator* of a compartment, and it also has *filtering* properties (only certain substances, in certain conditions, can pass through a membrane). In a cell, the membranes are hierarchically arranged, and they delimit "protected reactors", compartments where specific chemicals evolve according to specific reactions. Such a cell-like arrangement of membranes is called *membrane structure*. In membrane computing, the complexity (number of membranes and of levels) of membrane structures is not restricted.

A cell-like (hence hierarchical) membrane structure corresponds to a tree, hence a natural representation is by means of a tree or any mathematical representation of a tree, for instance, by means of expressions of labeled parentheses (a pair of parentheses is associated with a membrane). For instance $[_1[_2\,]_2[_3\,]_3]_1$ describes the membrane structure consisting of two membranes, with labels 2 and 3, placed in a membrane with label 1. Besides cell-like membrane structures, in membrane computing one also considers tissue-like systems, hence with elementary membranes (i. e., membranes without any membrane inside) placed in the nodes of an arbitrary graph.

Multiset A *multiset* is a set with multiplicities associated with its elements. For instance, $\{(a, 2), (b, 3)\}$ is the multiset consisting of 2 copies of element a and 3 copies of element b. Mathematically, a multiset is identified with a mapping μ from a *universe set* U (an alphabet) to $\mathbf{N}$, the set of natural numbers, $\mu\colon U \longrightarrow \mathbf{N}$. In the previous case, U can be any superset of $\{a, b\}$, for instance, $U = \{a, b, c\}$, and $\mu(a) = 2, \mu(b) = 3, \mu(c) = 0$. A compact representation of a multiset is by strings: any string w over U represents a multiset, with the number of occurrences of a symbol in w being the multiplicity of that element in the multiset represented by w. Consequently, any permutation of a string represents the same multiset. In the previous case, any permutation of the string a^2b^3, for instance, ab^2ab represents the multiset μ defined above. Multisets can also be considered on infinite universe sets or even with multiplicities being negative or non-integer numbers.

In membrane computing, multisets model solutions, chemical "objects" (ions, small molecules, macromolecules, etc.) swimming in water, in the same compartment of a cell, with the multiplicity of each object relevant for the bio-chemistry taking place in that compartment.

P system The computing devices investigated in membrane computing are called *P systems*. Basically, such a system consists of a membrane structure, with multisets of objects placed in its compartments and sets of evolution rules associated with either the compartments or the membranes. The objects evolve according to the local rules (the objects can also pass across membranes, from a compartment to an adjacent compartment). Also the membrane structure can evolve. Depending on the form of the membrane structure, there

are cell-like P systems, tissue-like P systems, with neural P systems being an important subclass of the latter. Specifying a P system means to define its initial configuration (the membrane structure and the multisets of objects present in each compartment) and its evolution rules, as well as the way of using the rules (hence the transition among configurations), the successful computations, and the output of a computation. There are many possibilities from all these points of view.

Parallelism Most computer science investigations in membrane computing deal with synchronized P systems, where a global clock marks the time for the whole system and each compartment/membrane evolves in each time unit, hence in a parallel way. In many cases, the rules are also used in parallel in each compartment: (like in bio-chemistry) all rules which can be applied are applied to the existing objects. This is the so-called *maximal parallelism* (objects are assigned to rules until no further rule can be applied). There are many variants: minimal parallelism (if at least one rule can be used in a compartment of a P system, then at least one must be used), bounded parallelism (at least, at most, or exactly k rules are used, for a given k), and so on. The rules can also be used sequentially (one in each compartment), or the system can be asynchronous.

Spiking neural P system Neurons are linked in a tissue-like way, communicating along axons by means of *spikes*, electrical impulses of a constant shape and voltage, with the frequency (distance in time between consecutive spikes) carrying information. Based on these neurological facts, *spiking neural P systems* are defined: neurons are represented by membranes, where copies of the single object a, representing the spike, are placed; these membranes are associated with nodes of a graph whose edges represent the synapses; rules for processing spikes are provided in each neuron, and in this way computations are defined.

Symport/antiport An important way of selectively passing chemicals across biological membranes is the coupled transport through protein channels. The process of moving two or more objects in the same direction is called *symport*; the process of simultaneously moving two or more objects across a membrane, in opposite directions, is called *antiport*. For uniformity, the particular case of moving only one object is called *uniport*. In membrane computing, a symport rule is written in the form (u, in) or (u, out), meaning that the objects specified by (the multiset represented by) the string u can enter, or respectively exit, the membrane to which the rule is associated. An antiport rule is written in the form $(u, out; v, in)$, meaning that the objects of u exit from and, simultaneously, those of v enter in the membrane with which the rule is associated.

Tissue P system P systems consisting of one-membrane cells placed in the nodes of a graph are called *tissue-like P systems*. The channels between cells can be given in advance, fixed, or they can evolve during the computation. A class of tissue-like P systems with a dynamic structure of channels among cells is that of *population P systems* (with motivation and applications related to populations/colonies of bacteria).

Universality A computing model which is equivalent in power with Turing machines (the Standardabweichung model of algorithmic computing) is said to be computationally complete, or Turing complete. In membrane computing, many classes of P systems are *Turing complete*. Because the proofs are constructive (Turing machines or equivalent devices, such as counter/register machines, Chomsky grammars, etc., are simulated by means of P systems), the computing completeness also means *universality* in the restricted sense, of the existence of a programmable device, which can simulate any device in a given class after taking as input a "code" of the particular device.

Definition of the Subject

Membrane computing is a branch of natural computing initiated in [9] which abstracts computing models from the architecture and the functioning of living cells, as well as from the organization of cells in tissues, organs (brain included) or other higher order structures. The initial goal of membrane computing was to learn from the cell biology something possibly useful to computer science, and the area fast developed in this direction. Several classes of computing models (called *P systems*) were defined in this context, inspired from biological facts or motivated from mathematical or computer science points of view. A series of applications were reported in the last years, in biology/medicine, linguistics, computer graphics, economics, approximate optimization, cryptography, etc.

The main ingredients of a P system are (i) the membrane structure, (ii) the multisets of objects placed in the compartments of the membrane structure, and (iii) the rules for processing the objects and the membranes. Thus, membrane computing can be defined as a framework for devising cell-like or tissue-like computing models which process multisets in compartments defined by means of membranes. These models are (in general) distributed and parallel. When a P system is considered as a computing device, hence it is investigated in terms of (theoretical)

computer science, the main issues investigated concern the *computing power* (in comparison with standard models from computability theory, especially Turing machines and their restrictions) and the *computing efficiency* (the possibility of using the parallelism for solving computationally hard problems in a feasible time). Computationally and mathematically oriented ways of using the rules and of defining the result of a computation are considered in this case (e. g., maximal or minimal parallelism, halting, counting objects). When a P system is constructed as a model of a bio-chemical process, then it is examined in terms of dynamical systems, with the evolution in time being the issue of interest, not a specific output.

Membrane computing is the first systematic framework for studying cell-inspired computing models, but after a considerable theoretical development, the domain returned to the originating area, biology, proving to be a promising modeling framework.

From a theoretical point of view, P systems are both powerful (most classes are Turing complete, even when using ingredients of a reduced complexity – a small number of membranes, rules of simple forms, ways of controlling the use of rules directly inspired from biology) and efficient (many classes of P systems, especially those with an enhanced parallelism, can solve computationally hard problems – typically **NP**-complete problems, but also harder problems – in a feasible time – typically polynomial). Then, as a modeling framework, membrane computing is rather adequate for handling discrete (biological) processes, having many attractive features: easy understandability, scalability and programmability, inherent compartmentalization and non-linearity, etc. Ideas from cell biology as captured by membrane computing proved to be rather useful in handling various computer science topics – one typical example is that of membrane evolutionary algorithms, used for solving optimization problems.

Introduction

The literature of membrane computing has grown very fast (already in 2003, Thompson Institute for Scientific Information, ISI, has qualified the initial paper as "fast breaking" and the domain as "emergent research front in computer science" – see http://esi-topics.com), while the bibliography of the field counts, at the middle of 2007, almost 900 titles – see the web site from http://psystems.disco.unimib.it, with a mirror at http://bmc.hust.edu.cn/psystems. Moreover, the domain is now very diverse, as a consequence of the many motivations of introducing new variants of P systems: to be biologically oriented/realistic, mathematically elegant, computationally powerful and efficient. That is why it is possible to give here only a few basic notions and only a few (types of) results and of applications. The reader interested in details should consult the monograph [10], the volume [2], where a friendly introduction to membrane computing can be found in the first chapter, and the comprehensive bibliography from the above mentioned web page.

The field started by looking to the cell in order to learn something possibly useful to computer science, but then the research also considered cell organization in tissues (in general, populations of cells, such as colonies of bacteria), and, recently, also neurons organization in brain. Thus, at the moment there are three main types of P systems: (i) cell-like P systems, (ii) tissue-like P systems, and (iii) neural-like P systems.

The first type imitates the (eukaryotic) cell, and its basic ingredient is the *membrane structure*, a hierarchical arrangement of membranes (understood as three dimensional vesicles), delimiting compartments where multisets of symbol objects are placed; rules for evolving these multisets as well as the membranes are provided, also localized, acting in specified compartments or on specified membranes. The objects not only evolve, but they also pass through membranes (we say that they are "communicated" among compartments). The rules can have several forms, and their use can be controlled in various ways: promoters, inhibitors, priorities, etc.

In tissue-like P systems, several one-membrane cells are considered as evolving in a common environment. They contain multisets of objects, while also the environment contains objects. Certain cells can communicate directly (channels are provided between them) and all cells can communicate through the environment. The channels can be given in advance or they can be dynamically established – this latter case appears in so-called *population P systems*.

Finally, there are two types of neural-like P systems. One of them are similar to tissue-like P systems in the fact that the cells (neurons) are placed in the nodes of an arbitrary graph and they contain multisets of objects, but they also have a *state* which controls the evolution. Another promising variant was recently introduced, under the name of *spiking neural P systems*, where one uses only one type of objects, the *spike*, and the main information one works with is the distance between consecutive spikes.

The cell-like P systems were introduced first and their theory is now very well developed; tissue-like P systems have also attracted a considerable interest, while the neural-like systems, mainly under the form of spiking neural P systems, are only recently investigated.

In what follows, in order to let the reader get a flavor of membrane computing, we will discuss in some detail only cell-like P systems and spiking neural P systems, and we refer to the area literature for other classes.

Cell-like P Systems

Because in this section we only consider cell-like P systems, they will be simply called P systems.

In short, such a system consists of a hierarchical arrangement of *membranes*, which delimit *compartments*, where *multisets* (sets with multiplicities associated with their elements) of abstract *objects* are placed. These objects correspond to the chemicals from the compartments of a cell; the chemicals swim in water (many of them are bound on membranes, but we do not consider this case here), and their multiplicity matters – that is why the data structure most adequate to this situation is the multiset (a multiset can be seen as a string modulo permutation, that is why in membrane computing one usually represents the multisets by strings). In what follows the objects are supposed to be unstructured, hence we represent them by symbols from a given alphabet. There also are classes of P systems dealing with string objects.

The objects evolve according to *rules* which are also associated with the regions. The rules say both how the objects are changed and how they can be moved (*communicated*) across membranes. There also are rules which only move objects across membranes, as well as rules for evolving the membranes themselves (e. g., by destroying, creating, dividing, or merging membranes). By using these rules, we can change the *configuration* of a system (the multisets from their compartments as well as the membrane structure); we say that we get a *transition* among system configurations.

The rules can be applied in many ways. The basic mode imitates the biological way chemical reactions are performed – in parallel, with the mathematical additional restriction to have a *maximal parallelism*: one applies a bunch of rules which are maximal, no further object can evolve at the same time by any rule. Besides this mode, there were considered several others: sequential (one rule is used in each step), bounded parallelism (the number of membranes to evolve and/or the number of rules to be used in any step is bounded in advance), minimal parallelism (in each compartment where a rule *can* be used, at least one rule *must be* used). In all cases, a common feature is that the objects to evolve and the rules by which they evolve are chosen in a *non-deterministic* manner. A sequence of transitions forms a *computation* and with computations which *halt* (reach a configuration where no rule is applicable) we associate a *result*, for instance, in the form of the multiset of objects present in the halting configuration in a specified membrane.

This way of using a P system, starting from an initial configuration and computing a number, is a grammar-like (generative) one. We can also work in an automata style: an input is introduced in the system, for instance, in the form of a number represented by the multiplicity of an object placed in a specified membrane, and we start computing; the input number is accepted if and only if the computation halts. A combination of the two modes leads to a functional behavior: an input is introduced in the system (at the beginning, or symbol by symbol during the computation) and also an output is produced. In particular, we can have a decidability case, where the input encodes a decision problem and the output is one of two special objects representing the answers `yes` and `no` to the problem.

The generalization of this approach is obvious. We start from the cell, but the abstract model deals with very general notions: membranes interpreted as separators of regions with filtering capabilities, objects and rules assigned to regions; the basic data structure is the multiset. Thus, membrane computing can be interpreted as a *bio-inspired framework for distributed parallel processing of multisets*.

As briefly introduced above, the P systems are synchronous systems, and this feature is useful for theoretical investigations (e. g., for obtaining universality results or results related to the computational complexity of P systems). Also non-synchronized systems were considered, asynchronous in the standard sense or even time-free, or clock-free (e. g., generating the same output, irrespective of the duration associated with the evolution rules). Similarly, in applications to biology, specific strategies of evolution are considered. We do not enter here into details, rather we refer the reader to the bibliography given below.

Ingredients of P Systems

Let us now go into some more specific details – still remaining at an informal level.

As said above, we look to the cell structure and functioning, trying to get suggestions for an abstract computing model. The fundamental feature of a cell is its compartmentalization through membranes. Accordingly, the main ingredient of a P system is the *membrane structure*, a hierarchical arrangement of membranes, which delimit compartments. Figure 1 illustrates this notion and the related terminology.

We distinguish the external membrane (corresponding to the plasma membrane and usually called the *skin*

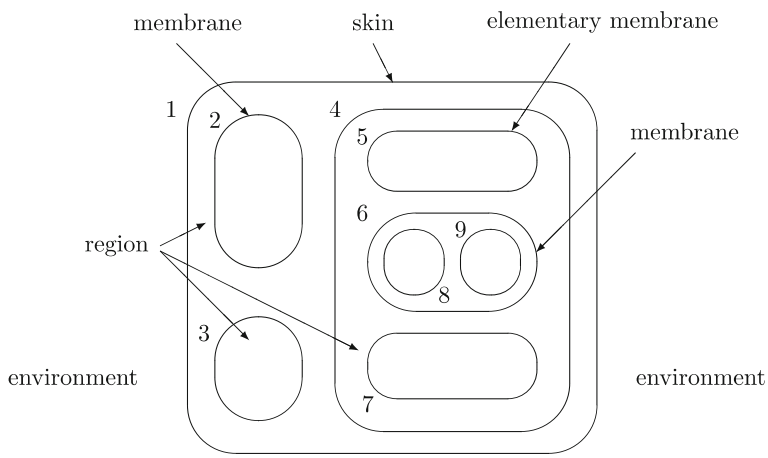

Membrane Computing, Figure 1
A membrane structure

membrane) and several internal membranes; a membrane without any other membrane inside it is said to be *elementary*. Each membrane determines a compartment, also called *region*, the space delimited from above by it and from below by the membranes placed directly inside, if any exists. The correspondence membrane–region is one-to-one, so that we identify by the same label a membrane and its associated region.

In the basic class of P systems, each region contains a multiset of symbol-objects, described by symbols from a given alphabet.

The objects evolve by means of evolution rules, which are also localized, associated with the regions of the membrane structure. The typical form of such a rule is $cd \to (a, here)(b, out)(b, in)$, with the following meaning: one copy of object c and one copy of object d react and the reaction produces one copy of a and two copies of b; the newly produced copy of a remains in the same region (indication *here*), one of the copies of b exits the compartment, going to the surrounding region (indication *out*) and the other enters one of the directly inner membranes (indication *in*). We say that the objects a, b, b are *communicated* as indicated by the commands associated with them in the right hand member of the rule. When an object exits the skin membrane, it is "lost" in the environment, it never comes back into the system. If no inner membrane exists (that is, the rule is associated with an elementary membrane), then the indication *in* cannot be followed, and the rule cannot be applied.

A membrane structure and the multisets of objects from its compartments identify a *configuration* of a P system. By a non-deterministic maximally parallel use of rules as suggested above we pass to another configuration; such a step is called a *transition*. A sequence of transitions constitutes a *computation*. A computation is successful if it halts, it reaches a configuration where no rule can be applied to the existing objects. With a halting computation we can associate a *result* in various ways. The simplest possibility is to count the objects present in the halting configuration in a specified elementary membrane; this is called *internal output*. We can also count the objects which leave the system during the computation, and this is called *external output*. In both cases the result is a number. If we distinguish among different objects, then we can have as the result a vector of natural numbers. The objects which leave the system can also be arranged in a sequence according to the moments when they exit the skin membrane, and in this case the result is a string.

Because of the non-determinism of the application of rules, starting from an initial configuration, we can get several successful computations, hence several results. Thus, a P system *computes* (one also uses to say *generates*) a set of numbers, or a set of vectors of numbers, or a language.

Many classes of P systems can be obtained by considering various possibilities for the various ingredients. We enumerate here several of these possibilities, without exhausting the list:

- Objects: symbols, strings of symbols, spikes, arrays, trees, numerical variables, other data structures, combinations.
- Data structure: multisets, sets (languages in the case of strings), fuzzy sets or fuzzy multisets.
- Place of objects: in compartments, on membranes, combined.
- Forms of rules: multiset rewriting, symport/antiport,

communication rules, boundary rules, with active membranes, combined, string rewriting, array/trees processing, spike processing.
- Controls on rules: catalysts, priority, promoters, inhibitors, activators, sequencing, energy.
- Form of membrane structure: cell-like (tree), tissue-like (arbitrary graph).
- Type of membrane structure: static, dynamical, precomputed (arbitrarily large).
- Timing: synchronized, non-synchronized, local synchronization, time-free.
- Ways of using the rules: maximal parallelism, minimal parallelism, bounded parallelism, sequential.
- Successful computations: global halting, local halting, with specified events signaling the end of a computation, non-halting.
- Modes of using a system: generative, accepting, computing an input-output function, deciding.
- Types of evolution: deterministic, non-deterministic, confluent, probabilistic.
- Ways to define the output: internal, external, traces, tree of membrane structure, spike train.
- Types of results: set of numbers, set of vectors of numbers, languages, set of arrays, `yes/no`.

We refer to the literature for details and we only add here the fact that when using P systems as models of biological systems/processes, we have to apply the rules in ways suggested by bio-chemistry, according to reaction rates or probabilities; in many cases, these rates are computed dynamically, depending on the current population of objects in the system.

In general, a (cell-like) P system is formalized as a construct $\Pi = (O, \mu, w_1, \ldots, w_m, R_1, \ldots, R_m, i_0)$, where O is the alphabet of objects, μ is the membrane structure (with m membranes), $w_1, \ldots, w_m$ are multisets of objects present in the m regions of μ at the beginning of a computation, $R_1, \ldots, R_m$ are finite sets of evolution rules associated with the regions of μ, and i_0 is the label of a membrane, used as the output membrane.

We end this section with a simple example, illustrating the architecture and the functioning of a (cell-like) P system. Figure 2 indicates the initial configuration (the rules included) of a system which computes a function, namely $n \longrightarrow n^2$, for any natural number $n \geq 1$. Besides catalytic and non-cooperating rules, the system also contains a rule with promoters, $b_2 \to b_2(e, in)|_{b_1}$: the object b_2 evolves to $b_2 e$ only if at least one copy of object b_1 is present in the same region.

In symbols, the system is given as follows:

$$\Pi = (O, C, \mu, w_1, w_2, R_1, R_2, i_0), \text{ where:}$$

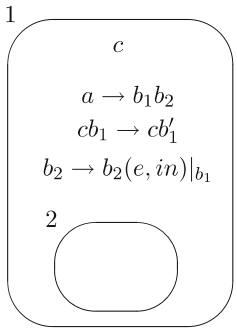

Membrane Computing, Figure 2
A P system with catalysts and promoters

$$\begin{aligned}
O &= \{a, b_1, b_1', b_2, c, e\} \text{ (the set of objects)}\\
C &= \{c\} \text{ (the set of catalysts)}\\
\mu &= [_1[_2]_2]_1 \text{ (membrane structure)}\\
w_1 &= c \text{ (initial objects in region 1)}\\
w_2 &= \lambda \text{ (initial objects in region 2)}\\
R_1 &= \{a \to b_1 b_2, cb_1 \to cb_1', b_2 \to b_2 e_{\text{in}}|_{b_1}\}\\
&\quad \text{(rules in region 1)}\\
R_2 &= \emptyset \text{ (rules in region 2)}\\
i_0 &= 2 \text{ (the output region).}
\end{aligned}$$

We start with only one object in the system, the catalyst c. If we want to compute the square of a number n, then we have to input n copies of the object a in the skin region of the system. In that moment, the system starts working, by using the rule $a \to b_1 b_2$, which has to be applied in parallel to all copies of a; hence, in one step, all objects a are replaced by n copies of b_1 and n copies of b_2. From now on, the other two rules from region 1 can be used. The catalytic rule $cb_1 \to cb_1'$ can be used only once in each step, because the catalyst is present in only one copy. This means that in each step one copy of b_1 gets primed. Simultaneously (because of the maximal parallelism), the rule $b_2 \to b_2(e, in)|_{b_1}$ should be applied as many times as possible and this means n times, because we have n copies of b_2. Note the important difference between the promoter b_1, which allows using the rule $b_2 \to b_2(e, in)|_{b_1}$, and the catalyst c: the catalyst is involved in the rule, it is counted when applying the rule, while the promoter makes possible the use of the rule, but it is not counted; the same (copy of an) object can promote any number of rules. Moreover, the promoter can evolve at the same time by means of another rule (the catalyst is never changed).

In this way, in each step we change one b_1 to b_1' and we produce n copies of e (one for each copy of b_2); the copies of e are sent to membrane 2 (the indication *in* from

the rule $b_2 \to b_2(e, in)|_{b_1}$). The computation should continue as long as there are applicable rules. This means exactly n steps: in n steps, the rule $cb_1 \to cb_1'$ will exhaust the objects b_1 and in this way neither this rule can be applied, nor $b_2 \to b_2(e, in)|_{b_1}$, because its promoter does no longer exist. Consequently, in membrane 2, considered as the output membrane, we get n^2 copies of object e.

Note that the computation is deterministic, always the next configuration of the system is unique, and that, changing the rule $b_2 \to b_2(e, in)|_{b_1}$ with $b_2 \to b_2(e, out)|_{b_1}$, the n^2 copies of e will be sent to the environment, hence we can read the result of the computation outside the system, and in this case membrane 2 is useless.

Spiking Neural P Systems

Spiking neural P systems (SN P systems) were introduced in [5] with the aim of defining P systems based on ideas specific to spiking neurons, recently much investigated in neural computing.

Very shortly, an SN P system consists of a set of *neurons* (cells, consisting of only one membrane) placed in the nodes of a directed graph and sending signals (*spikes*, denoted in what follows by the symbol a) along *synapses* (arcs of the graph). Thus, the architecture is that of a tissue-like P system, with only one kind of object present in the cells. The objects evolve by means of *spiking rules*, which are of the form $E/a^c \to a; d$, where E is a regular expression over $\{a\}$ and c, d are natural numbers, $c \geq 1, d \geq 0$. The meaning is that a neuron containing k spikes such that $a^k \in L(E), k \geq c$, can consume c spikes and produce one spike, after a delay of d steps. This spike is sent to all neurons to which a synapse exists outgoing from the neuron where the rule was applied. There also are *forgetting rules*, of the form $a^s \to \lambda$, with the meaning that $s \geq 1$ spikes are forgotten, provided that the neuron contains exactly s spikes. We say that the rules "cover" the neuron, all spikes are taken into consideration when using a rule. The system works in a synchronized manner, i. e., in each time unit, each neuron which can use a rule should do it, but the work of the system is sequential in each neuron: only (at most) one rule is used in each neuron. One of the neurons is considered to be the *output neuron*, and its spikes are also sent to the environment. The moments of time when a spike is emitted by the output neuron are marked with 1, the other moments are marked with 0. This binary sequence is called the *spike train* of the system – it might be infinite if the computation does not stop.

In the spirit of spiking neurons, the result of a computation is encoded in the distance between consecutive spikes sent into the environment by the (output neuron of the) system. For example, we can consider only the distance between the first two spikes of a spike train, or the distance between the first k spikes, the distances between all consecutive spikes, taking into account all intervals or only intervals that alternate, all computations or only halting computations, etc.

An SN P system can also be used in the accepting mode: a neuron is designated as the *input neuron* and two spikes are introduced in it, at an interval of n steps; the number n is accepted if the computation halts.

Another possibility is to consider the spike train itself as the result of a computation, and then we obtain a (binary) language generating device. We can also consider input neurons and then an SN P system can work as a transducer. Languages on arbitrary alphabets can be obtained by generalizing the form of rules: take rules of the form $E/a^c \to a^p; d$, with the meaning that, provided that the neuron is covered by E, c spikes are consumed and p spikes are produced, and sent to all connected neurons after d steps (such rules are called *extended*). Then, with a step when the system sends out i spikes, we associate a symbol b_i, and thus we get a language over an alphabet with as many symbols as the number of spikes simultaneously produced. Another natural extension is to consider several output neurons, thus producing vectors of numbers, not only single numbers.

Also for SN P systems we skip the technical details, but we consider a simple example. We give it first in a formal manner (if a rule $E/a^c \to a; d$ has $L(E) = \{a^c\}$, then we write it in the simplified form $a^c \to a; d$):

$$\begin{aligned}
\Pi_1 &= (O, \sigma_1, \sigma_2, \sigma_s, syn, out), \text{ with} \\
O &= \{a\} \text{ (alphabet, with only one object, the spike)} \\
\sigma_1 &= (2, \{a^2/a \to a; 0, a \to \lambda\}) \\
&\quad \text{(first neuron: initial spikes, rules)} \\
\sigma_2 &= (1, \{a \to a; 0,\ a \to a; 1\}) \\
&\quad \text{(second neuron: initial spikes, rules)} \\
\sigma_3 &= (3, \{a^3 \to a; 0,\ a \to a; 1,\ a^2 \to \lambda\}) \\
&\quad \text{(third neuron: initial spikes, rules)} \\
syn &= \{(1, 2), (2, 1), (1, 3), (2, 3)\} \text{ (synapses)} \\
out &= 3 \text{ (output neuron).}
\end{aligned}$$

This system is represented in a graphical form in Fig. 3 and it functions as follows. All neurons can fire in the first step, with neuron σ_2 choosing non-deterministically between its two rules. Note that neuron σ_1 can fire only if it contains two spikes; one spike is consumed, the other remains available for the next step.

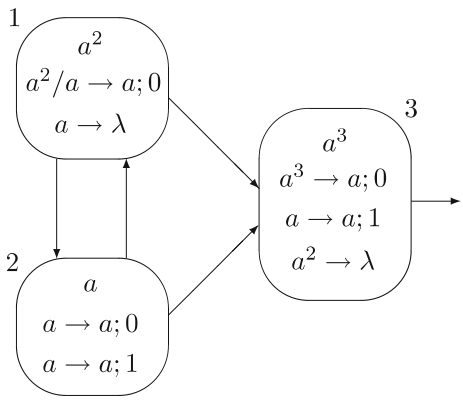

Membrane Computing, Figure 3
An SN P system generating all natural numbers greater than 1

Both neurons σ_1 and σ_2 send a spike to the output neuron, σ_3; these two spikes are forgotten in the next step. Neurons σ_1 and σ_2 also exchange their spikes; thus, as long as neuron σ_2 uses the rule $a \to a; 0$, the first neuron receives one spike, thus completing the needed two spikes for firing again.

However, at any moment, starting with the first step of the computation, neuron σ_2 can choose to use the rule $a \to a; 1$. On the one hand, this means that the spike of neuron σ_1 cannot enter neuron σ_2, it only goes to neuron σ_3; in this way, neuron σ_2 will never work again because it remains empty. On the other hand, in the next step neuron σ_1 has to use its forgetting rule $a \to \lambda$, while neuron σ_3 fires, using the rule $a \to a; 1$. Simultaneously, neuron σ_2 emits its spike, but it cannot enter neuron σ_3 (it is closed this moment); the spike enters neuron σ_1, but it is forgotten in the next step. In this way, no spike remains in the system. The computation ends with the expelling of the spike from neuron σ_3. Because of the waiting moment imposed by the rule $a \to a; 1$ from neuron σ_3, the two spikes of this neuron cannot be consecutive, but at least two steps must exist in between.

Thus, we conclude that Π computes/generates all natural numbers greater than or equal to 2.

Computing Power

As we have mentioned before, many classes of P systems, combining various ingredients (as described above or similar) are able of simulating Turing machines, hence they are *computationally complete*. Always, the proofs of results of this type are constructive, and this has an important consequence from the computability point of view: there are *universal* (hence *programmable*) P systems. In short, starting from a universal Turing machine (or an equivalent universal device), we get an equivalent universal P system. Among others, this implies that in the case of Turing complete classes of P systems, the hierarchy on the number of membranes always collapses (at most at the level of the universal P systems). Actually, the number of membranes sufficient in order to characterize the power of Turing machines by means of P systems is always rather small.

We only mention here three of the most interesting (types of) universality results for cell-like P systems:

1. P systems with symbol-objects with catalytic rules, using only two catalysts and two membranes, are universal.
2. P systems with symport/antiport rules of a restricted size (example: three membranes, symport rules of weight 2, and no antiport rules, or three membranes and minimal symport and antiport rules) are universal.
3. P systems with symport/antiport rules (of arbitrary size), using only three membranes and only three objects, are universal.

There are several other similar results, improvements or extensions of them. Many results are also known for tissue-like P systems. Details can be found, e. g., in the proceedings of the yearly Workshops of Membrane Computing mentioned in the bibliography of this chapter. For instance, universality results were obtained also in the case of P systems working in the accepting mode, and an interesting problem appears in this case, because we can then use deterministic systems. Most universality results were obtained in the deterministic case, but there also are situations where the deterministic systems are strictly less powerful than the non-deterministic ones. This is proven in [4], for the accepting catalytic P systems.

The hierarchy on the number of membranes collapses in many cases also for non-universal classes of P systems, but there also are cases when "the number of membrane matters", to cite the title of [3], where two classes of P systems were defined for which the hierarchies on the number of membranes are infinite.

Also various classes of SN P systems are computationally complete, as devices which generate or accept sets of numbers. This is true when no bound is imposed on the number of spikes present in any neuron; if such a bound exists, then the sets of numbers generated (or accepted) are semilinear.

It is worth noting that the proofs of computational completeness are based on simulating various types of grammars with restricted derivation (mainly matrix grammars with appearance checking) or on simulating register machines. In the case of SN P systems, this has an interesting consequence: starting the proofs from small univer-

sal register machines, as those produced in [6], one can find small universal SN P systems. For instance, as shown in [8], there are universal computing SN P systems with 84 neurons using standard rules and with only 49 neurons using extended rules. In the generative case, the best results are 79 and 50 neurons, respectively. Of course, these results are probably not optimal, hence it is a research topic to improve them.

Computational Efficiency

The computational power (the "competence") is only one of the important questions to be dealt with when defining a new (bio-inspired) computing model. The other fundamental question concerns the computing *efficiency*. Because P systems are parallel computing devices, it is expected that they can solve hard problems in an efficient manner – and this expectation is confirmed for systems provided with ways for producing an exponential workspace in a linear time. Three main such biologically inspired possibilities have been considered so far in the literature, and *all of them were proven to lead to polynomial solutions to* **NP**-*complete problems*.

These three ideas are *membrane division*, *membrane creation*, and *string replication*. The standard problems addressed in this framework were decidability problems, starting with SAT, the Hamiltonian Path problem, the Node Covering problem, but also other types of problems were considered, such as the problem of inverting one-way functions, or the Subset-sum and the Knapsack problems (note that the last two are numerical problems, where the answer is not of the yes/no type, as in decidability problems).

Roughly speaking, the framework for dealing with complexity matters is that of *accepting P systems with input*: a family of P systems of a given type is constructed starting from a given problem, and an instance of the problem is introduced as an input in such systems; working in a deterministic mode (or a *confluent* mode: some nondeterminism is allowed, provided that the branching converges after a while to a unique configuration, or, in the weak confluent case, all computations halt and all of them provide the same result), in a given time one of the answers yes/no is obtained, in the form of specific objects sent to the environment. The family of systems should be constructed in a uniform mode by a Turing machine, working a polynomial time.

This direction of research is very active at the present moment. More and more problems are considered, the membrane computing complexity classes are refined, characterizations of the **P**≠**NP** conjecture were obtained in this framework, several characterizations of the class **P**, even problems which are **PSPACE**-complete were proven to be solvable in polynomial time by means of membrane systems provided with membrane division or membrane creation. An important (and difficult) problem is that of finding the borderline between efficiency and non-efficiency: which ingredients should be used in order to be able to solve hard problems in a polynomial time? Many results in this respect were reported by M.J. Pérez-Jiménez and his co-workers (see the bibliography below), but still many problems remain open in this respect.

Similarly, so far, the efficiency issue was only very preliminarily investigated for SN P systems.

Applications

There are many features of membrane computing which make it attractive for applications in several disciplines, especially for biology.

First, there are several keywords which are genuinely proper to membrane computing and which are of interest for many applications: *distribution* (with the important system-part interaction, emergent behavior, non-linearly resulting from the composition of local behaviors), *programmability*, *scalability/extensibility*, *transparency* (multiset rewriting rules are nothing else than reaction equations as customarily used in chemistry and bio-chemistry), *parallelism*, *non-determinism*, *communication*, and so on and so forth.

Now, in what concerns the applications themselves reported up to now, they are developed at various levels. In many cases, what is actually used is the *language* of membrane computing, having in mind three dimensions of this aspect: (i) the long list of concepts either newly introduced, or related in a new manner in this area, (ii) the mathematical formalism of membrane computing, and (iii) the graphical language, the way to represent cell-like structures or tissue-like structures, together with the contents of the compartments and the associated evolution rules (the "evolution engine"). The next level is to use tools, techniques, results of membrane computing, and here there appears an important question: to which aim? Solving problems already stated, e. g., by biologists, in other terms and another framework, could be an impressive achievement, and this is the most natural way to proceed – but not necessarily the most efficient one, at least in the long term. New tools can suggest new problems, which either cannot be formulated in a previous framework or have no chance to be solved in the previous framework.

Applications of all these types were reported in the literature of membrane computing and, as expected, most of

them were carried out in biology. These applications are usually based on experiments using programs for simulating/implementing P systems on usual computers, and there are already several such programs, more and more elaborated (e. g., with better and better interfaces, which allow for the friendly interaction with the program). An overview of membrane computing software reported in literature (some programs are available in the web page http://psystems.disco.unimib.it) can be found in the volume [2]. Several applications are presented in detail – software included – at the web page of Sheffield membrane computing research group, http://www.dcs.shef.ac.uk/~marian/PSimulatorWeb/P_Systems_applications.htm, and at the page of Verona group, http://www.di.univr.it/dol/main?ent=arearic&id=21.

Of course, when using a P system for simulating a biological process we are no longer interested in its computing behavior (power, efficiency, etc.), but in its evolution in time; the P system is then interpreted as a dynamical system, and its trajectories are of interest, its "life". Moreover, the ingredients we use are different from those considered in theoretical investigations. For instance, in mathematical terms, we are interested in results obtained with a minimum of premises and with weak prerequisites, while the rules are used in ways inspired from automata and language theory (e. g., in a maximally or minimally parallel way), but when dealing with applications the systems are constructed in such a way to capture the features of reality (for instance, the rules are of a general form, they are applied according to probabilistic strategies, based on stoichiometric calculations, the systems are not necessarily synchronized, and so on).

The typical applications run as follows. One starts from a biological process described in general in graphical terms (chemicals are related by reactions represented in a graph-like manner, with special conventions for capturing the context-sensitivity of reactions, the existence of promoters or inhibitors, etc.) or already available in data bases in SBML (system biology mark-up language) form; these data are converted into a P system which is introduced in a simulator; the way the evolution rules (reactions) are applied is the key point in constructing this simulator (often, the classical Gilespie algorithm is used in compartments, or multi-compartmental variants of it are considered); as a result, the evolution in time of the multiplicity of certain chemicals is displayed, thus obtaining a graphical representation of the interplay in time of certain chemicals, their growth and decay, and so on. Many illustrations of this scenario can be found in the literature, many times dealing with rather complex processes.

Besides applications in biology, applications were reported in computer graphics (where the compartmentalization seems to add a significant efficiency to well-known techniques based on L systems), linguistics (both as a representation language for various concepts related to language evolution, dialog, semantics, and making use of the parallelism, in solving parsing problems in an efficient way), economics (where many bio-chemical metaphors find a natural counterpart, with the mentioning that the "reactions" which take place in economics, for instance, in market-like frameworks, are not driven only by probabilities/stoichiometric calculations, but also by psychological influences, which makes the modeling still more difficult than in biology), computer science (in devising sorting and ranking algorithms), cryptography, etc.

A very promising direction of research, namely, applying membrane computing in devising approximate algorithms for hard optimization problems, was initiated by Nishida, in [7], who proposed *membrane algorithms*, as a new class of distributed evolutionary algorithms. These algorithms can be considered as a high level (distributed and dynamically evolving their structure during the computation) evolutionary algorithms. In short, candidate solutions evolve in compartments of a (dynamical) membrane structure according to local algorithms, with better solutions migrating down in the membrane structure; after a specified halting condition is met, the current best solution is extracted as the result of the algorithm.

Nishida has checked this strategy for the traveling salesman problem, and the results were more than encouraging for a series of benchmark problems: the convergence is very fast, the number of membranes is rather influential on the quality of the solution, the method is reliable, both the average quality and the worst solutions were good enough and always better than the average and the worst solutions given by simulated annealing.

Similarly good results were obtained in a series of subsequent papers which have followed the same approach in addressing other hard optimization problems. In the bibliography below we have recalled a few recent titles, dealing mainly with the application of membrane algorithms in solving hard optimization problems. Always, benchmark problems were considered and the results were compared with those provided by other methods, existing in literature.

Future Directions

Although so much developed in the less than nine years since the investigations were initiated, membrane computing still has a large number of open problems and research topics which wait for research efforts, and new areas are

continuously appearing – the most recent one is the study of spiking neural P systems.

A general class of theoretical questions concerns the borderline between universality and non-universality or between efficiency and non-efficiency, i. e., concerning the succinctness of P systems able to compute at the level of Turing machines or to solve hard problems in polynomial time, respectively. Then, because universality implies undecidability of all non-trivial questions, an important issue is that of finding classes of P systems with decidable properties.

This is also related to the use of membrane computing as a modeling framework: if no insights can be obtained in an analytical manner, algorithmically, then what remains is to simulate the system on a computer. To this aim, better programs are still needed, maybe parallel implementations, able to handle real-life questions (for instance, in the quorum sensing area, existing applications deal with hundreds of bacteria, but biologists would need simulations at the level of thousands of bacteria in order to get convincing results).

Several research topics concern the neural inspiration for membrane computing, starting with the need of introducing a more elaborated model of neural nets. An important question in this respect is to use neuro-inspired models for efficiently solving problems, maybe borrowing ideas from the traditional neural computing (learning, solving pattern matching problems, etc.). This issue seems to also open new research directions for the traditional computability theory. Here is only one example. Efficiency is usually achieved in membrane computing by means of tools which allow producing an exponential working space in a linear time, and the standard way to do it is membrane division. However, in SN P systems we do not have such possibilities, the number of neurons remains the same and the number of spikes only increases polynomially with respect to the number of steps of a computation. How to introduce possibilities of generating an exponential workspace in a linear time remains as a research topic. Still, with inspiration from the fact that the brain consists of a huge number of neurons out of which only a small part are used, in [1] one proposes a way (illustrated for SAT) to address computationally hard problems in this framework, by assuming that an arbitrarily large SN P system is given "for free", pre-computed, with a structure as regular as possible, and without spikes inside; solving a problem starts by introducing a polynomial number of spikes in a polynomially bounded number of neurons; then, by moving spikes along synapses, the system self-activates, and a specific output provides the answer to the problem. This way of solving problems, by activating a pre-computed resource, is not at all usual in computability and is a research direction worth exploring.

Bibliography

Primary Literature

1. Chen H, Ionescu M, Ishdorj TO (2006) On the efficiency of spiking neural P systems. In: Proceedings of 8th International Conference on Electronics, Information, and Communication. Ulanbator, Mongolia, June 2006, pp 49–52
2. Ciobanu G, Păun G, Pérez-Jiménez MJ (eds) (2006) Applications of Membrane Computing. Springer, Berlin
3. Ibarra OH (2004) The number of membranes matters. In: Martín-Vide et al (eds) International Workshop, 1044 WMC2003, Tarragona, Reviced Papers LNCS, vol 2911. Springer, Berlin, pp 218–231
4. Ibarra OH, Yen HC (2006) Deterministic catalytic systems are not universal. Theor Comput Sci 363(2):149–161
5. Ionescu M, Păun G, Yokomori T (2006) Spiking neural P systems. Fundam Inform 71(2–3):279–308
6. Korec I (1996) Small universal register machines. Theor Comput Sci 168(2):267–301
7. Nishida TY (2004) An application of P systems: A new algorithm for NP-complete optimization problems. In: Callaos N et al (eds) Proceedings of the 8th World Multi-Conference on Systems, Cybernetics and Informatics, vol V. pp 109–112
8. Păun A, Păun G (2007) Small universal spiking neural P systems. BioSystems 90(1):48–60
9. Păun G (1998/2000) Computing with membranes. J Comput Syst Sci 61(1):108–143. (and Turku Center for Computer ScienceReport – TUCS 208, November (1998) http://www.tucs.fi)
10. Păun G (2002) Membrane Computing. An Introduction. Springer, Berlin

Books and Reviews

Alhazov A, Freund R, Rogozhin Y (2006) Computational power of symport/antiport: history, advances, and open problems. In: Freund R et al (eds) Springer, Berlin, pp 44–78

Bernardini F, Gheorghe M (2004) Population P systems. J Univers Comput Sci 10(5):509–539

Calude CS, Păun G, Rozenberg G, Salomaa A (eds) (2001) Multiset Processing. Mathematical, Computer Science, and Molecular Computing Points of View. LNCS, vol 2235. Springer, Berlin

Cardelli L, Păun G (2006) An universality result for a (mem)brane calculus based on mate/drip operations. Int J Found Comput Sci 17(1):49–68

Ciobanu G, Pan L, Păun G, Pérez-Jiménez MJ (2007) P systems with minimal parallelism. Theor Comput Sci 378(1):117–130

Csuhaj-Varjú E (2005) P automata. Models, results, and research topics. In: Mauri G et al (eds) Springer, Berlin, pp 1–11

Freund R, Kari L, Oswald M, Sosik P (2005) Computationally universal P systems without priorities: two catalysts are sufficient. Theor Comput Sci 330(2):251–266

Freund R, Păun G, Rozenberg G, Salomaa A (eds) (2006) Membrane Computing. In: 6th International Workshop, WMC6, Vienna, Austria, July 2005, Revised, Selected, and Invited Papers. LNCS, vol 3859. Springer, Berlin

Gutiérrez-Naranjo MA, Păun G, Pérez-Jiménez MJ (eds) (2005) Cellular Computing. Complexity Aspects. Fenix Editora, Sevilla

Hoogeboom HJ, Păun G, Rozenberg G, Salomaa A (eds) (2006) Membrane Computing. In: International Workshop, WMC7, Leiden, The Netherlands, 2006, Revised, Selected, and Invited Papers. LNCS, vol 4361. Springer, Berlin

Huang L, Wang N (2006) An optimization algorithms inspired by membrane computing. In: Jiao L et al (eds) Proc. ICNC (2006). LNCS, vol 4222. Springer, Berlin, pp 49–55

Jonoska N, Margenstern M (2004) Tree operations in P systems and λ-calculus. Fundam Inform 59(1):67–90

Kleijn J, Koutny M (2006) Synchrony and asynchrony in membrane systems. In: Hoogeboom HJ et al (eds) Springer, Berlin, pp 66–85

Leporati A, Mauri G, Zandron C (2006) Quantum sequential P systems with unit rules and energy assigned to membranes. In: Freund R et al (eds) Springer, Berlin, pp 310–325

Manca V (2006) MP systems approaches to biochemical dynamics: biological rhythms and oscillations. In: Hoogeboom HJ et al (eds) Springer, Berlin, pp 86–99

Martín-Vide C, Mauri G, Păun G, Rozenberg G, Salomaa A (eds) (2004) Membrane Computing. In: International Workshop, WMC2003, Tarragona, Spain, Revised Papers LNCS, vol 2933. Springer, Berlin

Mauri G, Păun G, Pérez-Jiménez MJ, Rozenberg G, Salomaa A (eds) (2005) Membrane Computing. In: International Workshop, WMC5, Milano, Italy, 2004, Selected Papers. LNCS, vol 3365. Springer, Berlin

Păun A, Păun G (2002) The power of communication: P systems with symport/antiport. New Gener Comput 20(3):295–306

Păun A, Pérez-Jiménez MJ, Romero-Campero FJ (2006) Modeling signal transduction using P systems. In: Hoogeboom HJ et al (eds) Springer, Berlin, pp 100–122

Păun G (2001) P systems with active membranes: Attacking NP-complete problems. J Autom, Lang Comb 6(1):75–90

Păun G, Păun R (2005) Membrane computing as a framework for modeling economic processes. In: Proceedings of Conference SYNASC, Timişoara. Press IEEE, pp 11–18

Păun G, Pazos J, Pérez-Jiménez MJ, Rodriguez-Patón A (2005) Symport/antiport P systems with three objects are universal. Fundam Inform 64(1–4):353–367

Păun G, Pérez-Jiménez MJ, Salomaa A (2007) Spiking neural P systems. An early survey. Int J Found Comput Sci 18:435–456

Păun G, Rozenberg G (2002) A guide to membrane computing. Theor Comput Sci 287(1):73–100

Păun G, Rozenberg G, Salomaa A, Zandron C (eds) (2003) Membrane Computing. International Workshop, WMC 2002, Curtea de Argeş, Romania, August 2002. Revised Papers. LNCS, vol 2597. Springer, Berlin

Pérez-Jiménez MJ (2005) An approach to computational complexity in membrane computing. In: Mauri G et al (eds) Springer, Berlin, pp 85–109

Pérez-Jiménez MJ, Romero-Jiménez A, Sancho-Caparrini F (2002) Teoria de la complejidad en modelos de computacion celular con membranas. Kronos Editorial, Sevilla

Sosik P, Rodriguez-Patón A (2007) Membrane computing and complexity theory: A characterization of International PSPACE. J Found Comput Sci 73(1):137–152

Zaharie D, Ciobanu G (2006) Distributed evolutionary algorithms inspired by membranes in solving continuous optimization problems. In: Hoogeboom HJ et al (eds) Springer, Berlin, pp 536–554

Metabolic Systems Biology

Nathan E. Lewis[1,2], Neema Jamshidi[1], Ines Thiele[2], Bernhard Ø. Palsson[1]

[1] Department of Bioengineering, University of California, San Diego, La Jolla, USA

[2] Bioinformatics program, University of California, San Diego, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Reconstructions, Knowledge Bases, and Models
Constraint-Based Modeling
Metabolic Systems Biology and Constraint-Based Modeling: Applications
Future Directions
Acknowledgments
Bibliography

Glossary

Bibliome The collection of primary literature, review literature and textbooks on a particular topic.

Biochemically, genetically and genomically (BiGG) structured reconstruction A structured genome-scale metabolic network reconstruction which incorporates knowledge about the genomic, proteomic, and biochemical components, including relationships between each component in a particular organism or cell (See Sect. "Reconstructions, Knowledge Bases, and Models").

Biomass function A pseudo-reaction representing the stoichiometric consumption of metabolites necessary for cellular growth (i. e., to produce biomass). When this pseudo-reaction is placed in a model, a flux through it represents the *in silico* growth rate of the organism or population (See Sect. "Constraint-Based Methods of Analysis").

Constraint-based reconstruction and analysis (COBRA) A set of approaches for constructing manually curated, stoichiometric network reconstructions and analyzing the resulting models by applying equality and inequality constraints and computing functional states. In general, mass conservation and thermodynamics (for directionality) are the fundamental constraints. Additional constraints reflecting experimental conditions

and other biological constraints (such as regulatory states) can be applied. The analysis approaches generally fall into two classes: biased and unbiased methods. Biased methods involve the application of various optimization approaches which require the definition of an objective function. Unbiased methods do not require an objective function (See Sect. "Constraint-Based Modeling").

Convex space A multi-dimensional space in which a straight line can be drawn from any two points, without leaving the space (see Sect. "Constraint-Based Methods of Analysis").

Extreme pathways (ExPa) analysis An approach for calculating a unique, linearly independent, but biochemically feasible reaction basis that can describe all possible steady state flux combinations in a biochemical network. ExPas are closely related to Elementary Modes (See Sect. "Constraint-Based Methods of Analysis").

Flux-balance analysis (FBA) The formalism in which a metabolic network is framed as a linear programming optimization problem. The principal constraints in FBA are those imposed by steady state mass conservation of metabolites in the system (See Sect. "Constraint-Based Methods of Analysis").

Gene-protein-reaction association (GPR) A mathematical representation of the relationships between gene loci, gene transcripts, protein sub-units, enzymes, and reactions using logical relationships (and/or) (See Sect. "Reconstructions, Knowledge Bases, and Models").

Genome-scale The characterization of a cellular function/system on its genome scale, i. e., incorporation/consideration of all known associated components encoded in the organism's genome.

Isocline A line in a phenotypic phase plane diagram, along which the ratio between the shadow prices for two metabolites is fixed (See Sect. "Constraint-Based Methods of Analysis").

Knowledge base A specific type of reconstruction which also accounts for the following information: molecular formulae, subsystem assignments, GPRs, references to primary and review literature, and additional pertinent notes (See Sect. "Reconstructions, Knowledge Bases, and Models").

Line of optimality The isocline in a phenotypic phase plane diagram that achieves the highest value of the objective in the phase plane (See Sect. "Constraint-Based Methods of Analysis").

Linear programming problem A class of optimization problems in which a linear objective function is maximized or minimized subject to linear equality and inequality constraints (See Sect. "Constraint-Based Methods of Analysis").

Metabolic network null space The set of independent vectors that satisfy the equations: $\mathbf{S} \bullet \boldsymbol{v} = 0$; i. e., a flux basis satisfying the steady state conditions, also referred to as the steady state flux solution space (See Sect. "Constraint-Based Methods of Analysis").

Network reconstruction An assembly of the components and their interconversions for an organism, based on the genome annotation and the bibliome (See Sect. "Reconstructions, Knowledge Bases, and Models").

Objective function A function which is maximized or minimized in optimization problems. In FBA, the objective function is a linear combination of fluxes. For prokaryotes and simple eukaryotes grown in the laboratory under controlled conditions, the biomass function is often used as the objective function (See Sect. "Constraint-Based Methods of Analysis").

Open reading frame (ORF) A DNA segment that has a start and stop site for translation and can encode for a protein product (see Sect. "The Human Metabolic Network Reconstruction: Characterizing the Knowledge Landscape and a Framework for Drug Target Discovery").

Phenotypic phase plan (PhPP) analysis A constraint-based method of analysis which uses FBA simulations to perform a sensitivity analysis by optimizing the objective function as two uptake fluxes are varied simultaneously. The results of generally displayed graphically. Isoclines and the line of optimality can be used to characterize different functional states in the phase plane (See Sect. "Constraint-Based Methods of Analysis").

Sensitivity analysis The analysis of how the output of a model changes as input parameters are varied (See Sect. "Constraint-Based Methods of Analysis").

Shadow price For FBA optimization problems, the (negative) change in the objective function divided by the change in the availability of a particular metabolite (i. e., the negative sensitivity of the objective function with respect to a particular metabolite) (See Sect. "Constraint-Based Methods of Analysis").

Single nucleotide polymorphism (SNP) A genetic sequence variation that involves a change or variation of a single base (See Sect. "Causal SNP Classification Using Co-sets").

Solution space The set of feasible solutions for a system under a defined set of constraints (See Sect. "Constraint-Based Methods of Analysis").

Uniform random sampling (Monte Carlo sampling) A constraint-based method of analysis that uses Monte

Carlo sampling methods to obtain a uniform distribution of random samples from the allowable flux space in order to find the range and probability distributions for reaction fluxes (See Sect. "Constraint-Based Methods of Analysis").

Definition of the Subject

Systems biology has various definitions. Common features among accepted definitions generally involve the description and analysis of interacting biomolecular components. Systems analysis of biological network is quickly demonstrating its utility as it helps to characterize biomolecular behavior that could not otherwise be produced by the individual components alone [46]. Three areas in which systems analysis has been implemented in biology include: (1) the generation and statistical analysis of high-throughput data in an effort to catalog and characterize cellular components; (2) the construction and analysis of computational models for various biological systems (e. g., metabolism, signaling, and transcriptional regulation); and (3) the integration of the knowledge of parts and computational models to predict and engineer biological systems (synthetic biology) [18,45,46]. Metabolism, as a system, has played an important role in the development of systems biology, especially in the modeling sense. This is because the network components (e. g., enzymes and metabolites) have been studied in detail for decades, and many links between components have been experimentally characterized. Metabolic systems biology, compared to systems biology in general, entails the computational analysis of these enzymes and metabolites and the metabolic pathways in which they participate. Metabolic systems biology, using genome-scale metabolic network reconstructions and their models, has helped (1) to elucidate biomolecular function [75]; (2) to predict phenotypic behavior [21]; (3) to discover new biological knowledge [19,75]; and (4) to design experiments for engineering applications [3,54]. Constraint-Based methods have played a pivotal role in the analysis of large and genome-scale metabolic networks. The structure, mathematical formulation, and analytical techniques of constraints-based methods have also paved the way for the successful modeling of other complex biological networks, such as transcriptional regulation [5,19,30,34] and signaling networks [60].

Introduction

The analysis of network capabilities, prediction of cellular phenotypes, and *in silico* hypothesis generation are among the goals in metabolic systems biology. In order to build the models that enable such analysis, a large amount of knowledge about the biological system is required. For a growing number of organisms detailed knowledge about the molecular components and their interactions is becoming available. The increasing availability of various types of high-throughput data, such as transcriptomic, proteomic, metabolomic, and interactomic (e. g., protein-protein, protein-DNA), has facilitated their identification.

Biological networks are too complex to be described by traditional mechanistic modeling approaches. This is due to the large number of components, the various physicochemical interactions, and complex hierarchical organization in space and time. Consequently, constraint-based modeling approaches have been developed which combine fundamental physicochemical constraints with mathematical methods to circumvent the need for comprehensive parametrization. These models are able to retain critical mechanistic aspects such as network structure via stoichiometry and thermodynamics. However, these constraints will not yield a uniquely determined system, but rather an underdetermined system of linear equations. Hence it is important to develop methods to characterize the functional properties of the solution spaces. There has been intense activity in developing such methods, many of which have been reviewed in Price, et al. [70] and are listed later in this chapter. The general principles underlying genome-scale modeling techniques will be further described in this chapter.

This chapter will introduce the reconstruction process and describe some constraint-based methods of analysis. This will be followed by example studies involving *E. coli* and human metabolism in which these constraint-based approaches have been successfully implemented for:

- Predicting phenotypes and outcomes of adaptive evolution in *E. coli* (Sect. "Growth Predictions of Evolved Strains");
- Discovering gene function in *E. coli* (Sect. "Discovery of Gene Function");
- Characterizing healthy and disease states in the human cardiomyocyte mitochondria (Sec. "Effects of Perturbed Mitochondrial States");
- Functionally classifying correlated reaction sets to understand disease states and potential treatment targets in human metabolism (Sect. "Causal SNP Classification Using Co-sets"):
- Expanding genome-scale modeling to human metabolism (Sect. "The Human Metabolic Network Reconstruction: Characterizing the Knowledge Landscape and a Framework for Drug Target Discovery").

Reconstructions, Knowledge Bases, and Models

Where is the Life we have lost in living?
Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?
T.S. Eliot, "The Rock", Faber & Faber 1934.

A reconstruction is an assembly of the components and their interactions for an organism, based on the genome annotation and the bibliome. A knowledge base is a very specific type of reconstruction which also accounts for the following information: molecular formulae, subsystem assignments, gene-protein-reaction associations (GPRs), references to primary and review literature, and additional notes regarding data quality and sources. Therefore, this knowledge base represents an assimilation of the current state of knowledge about biochemistry of the particular organism. A knowledge base also highlights missing information (e. g., network gaps and missing GPRs). Throughout the remainder of the chapter we will refer to knowledge bases and reconstruction interchangeably even though we have defined them distinctly.

A model is the result of converting a knowledge base or reconstruction into a computable, mathematical form by translating the networks into a matrix format and by defining system boundaries (see Sect. "From Reconstruction to Models"). It is important to note that a single reconstruction or knowledge base may yield multiple condition specific models (Fig. 1). The relationships between a genome and its derivative proteomes and phentoypes are analogous to the relationships between a knowledge base, the resulting models, and the possible functional states (Fig. 1).

Reconstructions, in a way, reverse the concern voiced above by T.S. Eliot by structuring data to provide information and processing information to find knowledge, which is then cataloged in a knowledge base. The models derived from this knowledge base can then be used for *in silico* hypothesis generation and predictive modeling, which can lead to biological discovery and provide insight into how biological systems operate.

There are two prominent approaches for network reconstructions top-down and bottom-up. Top-down reconstructions rely on high-throughput data, genome sequence, and genome annotation to computationally piece together component interaction networks based on statistical measures. Top-down reconstructions often aim to characterize the entire network. However, the resulting network links may be "soft", since they are based on statistical approaches. Top-down approaches may lead to the discovery of previously unknown components and relationships. Bottom-up reconstructions, in contrast, aim to be accurate and well-defined in their scope, as the components and interactions are experimentally verified. The bottom-up reconstruction process requires extensive manual curation and validation of its content to ensure the desired accuracy and self-consistency. This process results in a Biochemically, Genetically and Genomically (BiGG) structured knowledge base, in which genes are connected to the proteins and enzymes they encode, and each enzyme is connected to the reactions it catalyzes, also known as a gene-protein-reaction-association (GPR). Bottom-up reconstructions have been shown to be useful for many applications such as generating hypotheses and analyzing system processes.

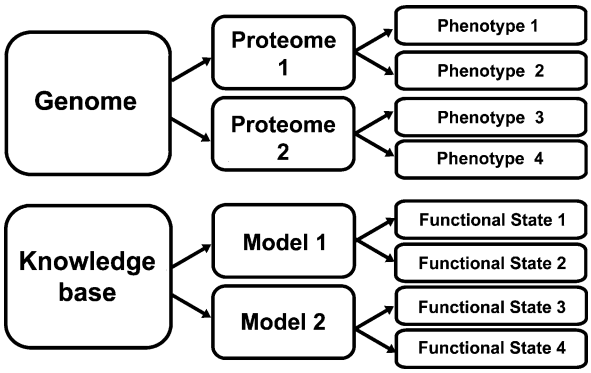

Metabolic Systems Biology, Figure 1
An analogy between genomes and knowledge bases. Regulation plays a significant role in defining phenotypes for a given genome. The regulatory program is driven by environmental cues. Similarly models derived from a knowledge base are subject to the constraints reflecting the regulatory and environmental factors (which also govern the proteome). The set of candidate functional states of different models reflects all of the possible phenotypes

The Reconstruction Process

The bottom-up reconstruction of genome-scale metabolic networks is a well established procedure that has been conducted for many organisms [73] and can be carried out in an algorithmic manner (Fig. 2). Briefly, the main phases are: (1) the generation of an initial component list based on genome annotation, (2) the manual curation of this initial list based on primary and review literature, (3) the functional validation of network capabilities using experimental data, and (4) simulation and analysis. This last step may lead to iterative improvements through reconciliation of the network with new data.

Step 1: Generation of the Initial Component List The first step in the reconstruction of a metabolic network is the selection of an organism and generation of a list of all currently known components (e. g., genes, proteins, and

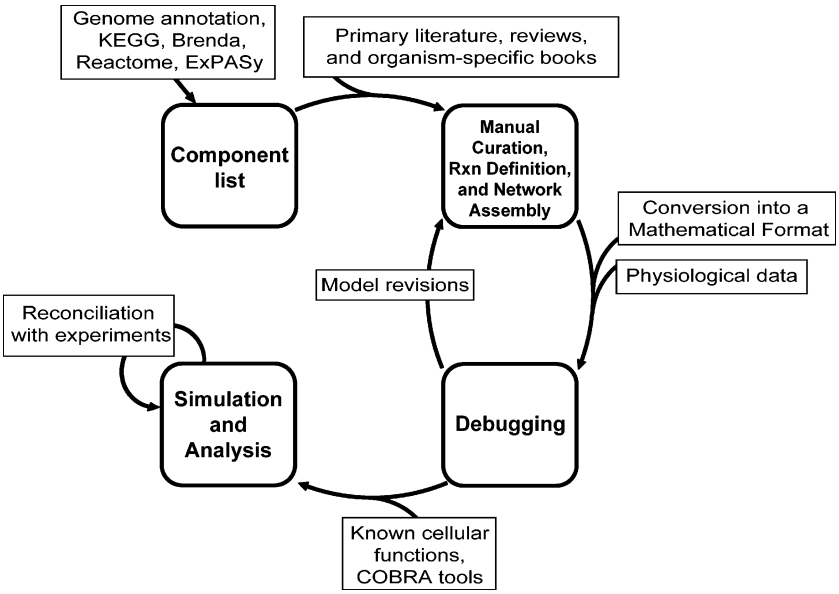

Metabolic Systems Biology, Figure 2

The reconstruction process. First, a component list is generated from the genome annotation and information in various databases. Second, the component list is manually curated using primary and review literature. Furthermore, the reactions are mass and charge balanced and assembled into pathways. Third, debugging of the reconstruction is done by computationally testing the properties and capabilities of the reconstruction to ensure that the derived models have capabilities similar to the organism. Pathway gaps may be filled if supported by experimental evidence or if required for network functionality. Fourth, simulation and analysis is conducted to reconcile the reconstruction with experimental data, which may lead to further iterations and refinements of the reconstruction

metabolites) involved in its metabolism. A sequenced and annotated genome is a prerequisite for building genome-scale networks. The quality of the reconstruction depends greatly on the quality of the annotation and the available literature describing the physiology and biochemistry of the organism. Parsing of the genome annotation for genes with metabolic functions results in the initial component list, and this list may be extended by obtaining associated reactions for these functions from databases such as KEGG [43], BRENDA [82], ExPASy [29], Reactome [99], and MetaCyc [17]. It is critical to manually curate this list, since the specific enzymes in the reconstructed organism may not act upon all of the substrates and cofactors included in the reactions in these databases.

Step 2: Manual Curation Once a component list is compiled, biochemical reactions must be manually defined, verified, assigned a confidence score, and assembled into pathways. For each reaction in the network, several properties must be defined, such as substrate specificity and their corresponding products, reaction stoichiometry, reaction directionality, subcellular localization, and chemical formulae for the metabolites with their corresponding charges. In addition, all genes and their associated gene products are connected to reactions in GPRs, using Boolean logic to describe each association. This thorough description of each reaction involves manual curation, as information is gathered from the primary literature, review articles, and organism-specific books. The completeness of this description depends heavily on how extensively the organism has been studied. Confidence scores for the reactions are a measure for the level of experimental support for the inclusion of a gene and its associated reaction(s). Generally, confidence scores have been defined on a scale from 1 to 4 and are assigned to each reaction in a reconstruction. Direct biochemical characterization of reaction activity is considered the gold standard; therefore reactions with such data receive a confidence score of 4. A score of 3 is given to reactions that are supported by genetic data, such as gene cloning. When a reaction is supported by sequence homology, physiological data, or localization data, the reaction is given a confidence score of 2. Reactions that are added only because they were needed for modeling purposes would receive a score of 1 (for a more detailed description, refer to Reed et al. [73]).

Step 3: Debugging and Functional Validation of the Reconstruction Even for well studied organisms,

a metabolic reconstruction at this stage will have a substantial number of gaps, resulting in limited network functionality. These gaps exist because the annotation of the genome is often incomplete. For example, even in *E. coli*, 20% of all genes do not have known functions, according to the latest genome annotation [77].

Gaps can be classified as knowledge or scope gaps. Knowledge gaps result from a lack of knowledge about the presence of transport and/or biochemical transformations for a particular metabolite in the target organism. Preferably, these gaps will be filled using primary literature. Alternatively, reactions may be included as hypotheses that require experimental verification and therefore are assigned a low confidence score. After converting the reconstruction into a mathematical format (see Fig. 3b and Sect. "From Reconstruction to Models"), computational algorithms can be used to assist the gap-filling process [55,78]. Using tools such as flux balance analysis (FBA) the reconstruction can be tested for the functionality of all physiologically relevant pathways. For example, if an amino acid is known to be non-essential for an organism, the complete biosynthetic pathway is needed, even if some of the required genes have not been annotated in the genome. Scope gaps involve transformations that that are outside the scope of interest in the network, such as DNA methylation reactions and tRNA charging. These gaps will not be filled, but it is important to document and classify these in the knowledge base.

Step 4: Simulation and Analysis Once the network is manually curated and debugged, its capabilities and accuracy should be tested by comparing *in silico* with experimental observations. These tests may include gene essentiality studies and evaluation of growth phenotypes under various conditions [55,73]. This step includes the simulation and analysis of secretion products and alternate nutrient sources. Additional experimental studies and newly generated data will lead to further iterations and refinement of the reconstruction content.

From Reconstruction to Models

A network reconstruction is converted into a mathematical model in two steps. First, system boundaries and the

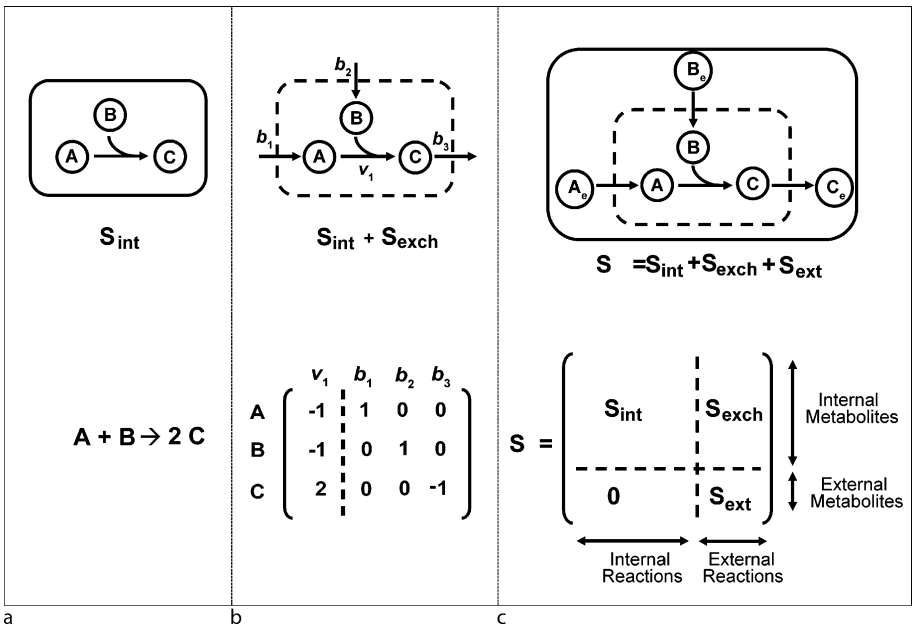

Metabolic Systems Biology, Figure 3

Converting a reconstruction into a model. The conversion of a reconstruction into a model involves the definition of system boundaries (*top*) and the conversion to a mathematical format (*bottom*). **a** Biochemical reactions can be written as conversions from reactant(s) into product(s). **b** Input and output fluxes that transport metabolites in and/or out of the system are defined (designated by b_x). The dashed line indicates an open system. **c** In a cell, for example, the cell membrane acts as a natural boundary. In a model it can be represented by an open system boundary, allowing the transfer of metabolites across the cell membrane. The complete mathematical representation of the network, containing the entire set of internal reactions with exchange (transport) reactions, is termed the stoichiometric matrix, **S**. Abbreviations: S_{int} = internal reactions, S_{exch} = exchange reactions across the open system boundaries, S_{ext} = extracellular metabolites

relevant inputs and outputs are defined. Second, the network is represented by a matrix. In this matrix, each column represents a reaction and each row represents a unique metabolite. The elements of the matrix are the stoichiometric coefficients for each metabolite in each reaction (reactants are negative and products are positive). The collection of reactions represented in this manner is called the stoichiometric matrix, **S**. At this stage condition specific constraints (e. g., measured uptake and secretion rates, or known regulatory constraints) will be applied to external and internal reactions, thus resulting in a distinct, condition-specific set. Different sets of constraints applied to the same reconstruction will result in different models.

Constraint-Based Modeling

As discussed above, constraint-based modeling has enabled the analysis of genome-scale networks in a mechanistic and predictive manner without relying on data-intensive parametrization. In addition, this technique has provided great flexibility, allowing different methods to be used while requiring few changes to the model structure. The strengths of this modeling approach are demonstrated in the size of the networks that can be modeled (e. g. the human metabolic network involves about 3,300 reactions [20]) and the ability it has to make predictions despite incomplete knowledge of the system. This section will focus on the types of constraints and some of the associated methods.

Biological Basis for Constraints

Constraints on cells can be grouped into three major classes: physicochemical, environmental, and regulatory constraints. Physicochemical constraints, the first type, are inviolable "hard" constraints on cell function. These constraints include osmotic balance, electroneutrality, the laws of thermodynamics, and mass and energy conservation. Spatial constraints, another type of physicochemical constraints, affect the function of biological systems due to mass transport limitations and molecular crowding. The second class of constraints, environmental constraints, are condition and time dependent, and include variables such as pH, nutrients, temperature, and extracellular osmolarity. Since environmental conditions and their effects on a cell can vary widely, predictive models rely heavily on well-defined experimental conditions. The third type of constraints, regulatory constraints, are self-imposed constraints in which pathway fluxes are modulated by allosteric regulation of enzymes and/or by gene expression via transcriptional control. These constraints are "soft" and can be altered through evolutionary processes [35]. These collective constraints contribute to a specific phenotype; therefore, their consideration in constraint-based modeling will assist the identification of relevant functional states.

The Mathematical Description of Constraints

Constraints can be quantitatively represented by balances and bounds, where balances are equalities and bounds are inequalities. The conservation of mass dictates that there is no net accumulation or depletion of metabolites at the steady state. This mass balance can be described mathematically as

$$\mathbf{S} \bullet \boldsymbol{v} = 0 \,, \tag{1}$$

where **S** is the $m \times n$ stoichiometric matrix (with m metabolites and n reactions), and $\boldsymbol{v}(n \times 1)$ is the flux vector, which represents the flux through each network reaction [98]. Similar steady state balance representations can be used for other physicochemical constraints, such as electroneutrality [42,57], osmotic pressure [12,41,49], and thermodynamic constraints around biochemical loops [7].

Bounds can be added as additional constraints. Environmental and regulatory constraints can be added by placing bounds on individual chemical transformations. For example, upper and lower flux rate limits

$$\boldsymbol{v}_{\min} \leqslant \boldsymbol{v}_i \leqslant \boldsymbol{v}_{\max} \tag{2}$$

can be placed on the ith reaction or transporter, to reflect experimentally measured enzyme capacity or metabolite uptake rates in a given environment. Thermodynamic constraints for each reaction can be applied by constraining the reaction directionality or by applying a set of linear thermodynamic constraints to eliminate thermodynamically infeasible fluxes [32].

Constraint-Based Methods of Analysis

A plethora of methods have been developed to analyze constraint-based models and many have been reviewed thoroughly [70]. Table 1 provides a list of methods and potential questions they can address. Constraint-Based methods can be grouped into biased and unbiased approaches.

Biased Methods Biased methods necessitate an objective function, i. e., a reaction or pseudo-reaction for which one optimizes. Some examples of commonly used objective functions include biomass production, ATP production, or the production of a byproduct of interest [80,100].

Metabolic Systems Biology, Table 1

There are numerous methods that have been developed to analyze constraint-based reconstructions of metabolic networks using experimental data to answer biological questions. Below is a list of some of these methods and questions they can help to answer

Method	Question	Examples
Alternate Optima	How many flux states can be attained by maximizing or minimizing an objective function (e. g., maximum growth or ATP production)?	[53,56,74,96,100]
Energy Balance Analysis	How can one evaluate the thermodynamic feasibility of FBA simulation results?	[7]
ExPa/ElMo	How does one define a biochemically feasible, unique set of reactions that span the steady state solution space?	[64,81,85]
FBA	What is the maximum (or minimum) of a specified cellular objective function?	[27,65,80,97]
Flux Confidence Interval	What are the confidence intervals of flux values when fluxomic data is mapped to a constraint based model?	[4]
Flux Coupling	What are the sets of network reactions that are fully coupled, partially coupled, or directionally coupled?	[14]
Flux Variability Analysis	What is the maximum and minimum flux for every reaction under a given set of constraints (i. e., what is the bounding box of the solution space)?	[56]
Gap-Fill/Gap-Find	What are the candidate reactions that can fill network gaps, thus helping improve the model and providing hypotheses for unknown pathways that can be experimentally validated?	[78]
Gene Annotation Refinement Algorithm	Which reactions are likely missing from the network, given a set of phenotypic observations? What are the candidate gene products with which corresponding reactions could fill the gap?	[75]
Gene Deletion Analysis	Which are the lethal gene deletions in an organism?	[22]
K-cone analysis	Given a set of fluxes and concentrations for a particular steady state, what is the range of allowable kinetic constants?	[25]
Metabolite Essentiality	How does metabolite essentiality contribute to cellular robustness?	[44]
Minimization of Metabolic Adjustment (MOMA)	Can suboptimal growth predictions be more consistent with experimental data in wild type and knock-out strains?	[86]
Net Analysis	Given metabolomic data, what are the allowable metabolite concentration ranges for other metabolites, and what are the likely regulated steps in the pathway based on nonequilibrium thermodynamics?	[51]
Objective function finder / ObjFind	What are the different possible cellular objectives?	[13,50,83]
Optimal Metabolic Network Identification	Given experimentally measured flux data, what is the most likely set of active reactions in the network under the given condition that will reconcile data with model predictions?	[33]
OptKnock / OptGene	How can one design a knock-out strain that is optimized for byproduct secretion coupled to cellular growth?	[15,66]
OptReg	What are the optimal reaction activations/inhibitions and eliminations to improve biochemical production?	[68]
OptStrain	Which reactions (not encoded by the genome) need to be added in order to enable a strain to produce a foreign compound?	[67]
PhPP	How does an objective function change as a function of two metabolite exchange rates?	[37,95]
rFBA	How do transcriptional regulatory rules affect the range of feasible *in silico* phenotypes?	[19]
Regulatory On/Off Minimization (ROOM)	After a gene knockout, what is the most probable flux distribution that requires a minimal change in transcriptional regulation?	[87]
Robustness analysis	How does an objective function change as a function of another network flux?	[23]
SR-FBA	To what extent do different levels of metabolic and transcriptional regulatory constraints determine metabolic behavior?	[88]
Stable Isotope Tracers	How can intracellular flux predictions be experimentally validated, and which pathways are active under the different conditions?	[79]
Thermodynamics based Metabolic Flux Analysis	How can one use thermodynamic data to generate thermodynamically feasible flux profiles?	[32]
Uniform Random Sampling	What are the distributions of network states that have not been excluded based on physicochemical constraints and/or experimental measurements? What are the completely or partially correlated reaction sets?	[1,71,93,101]

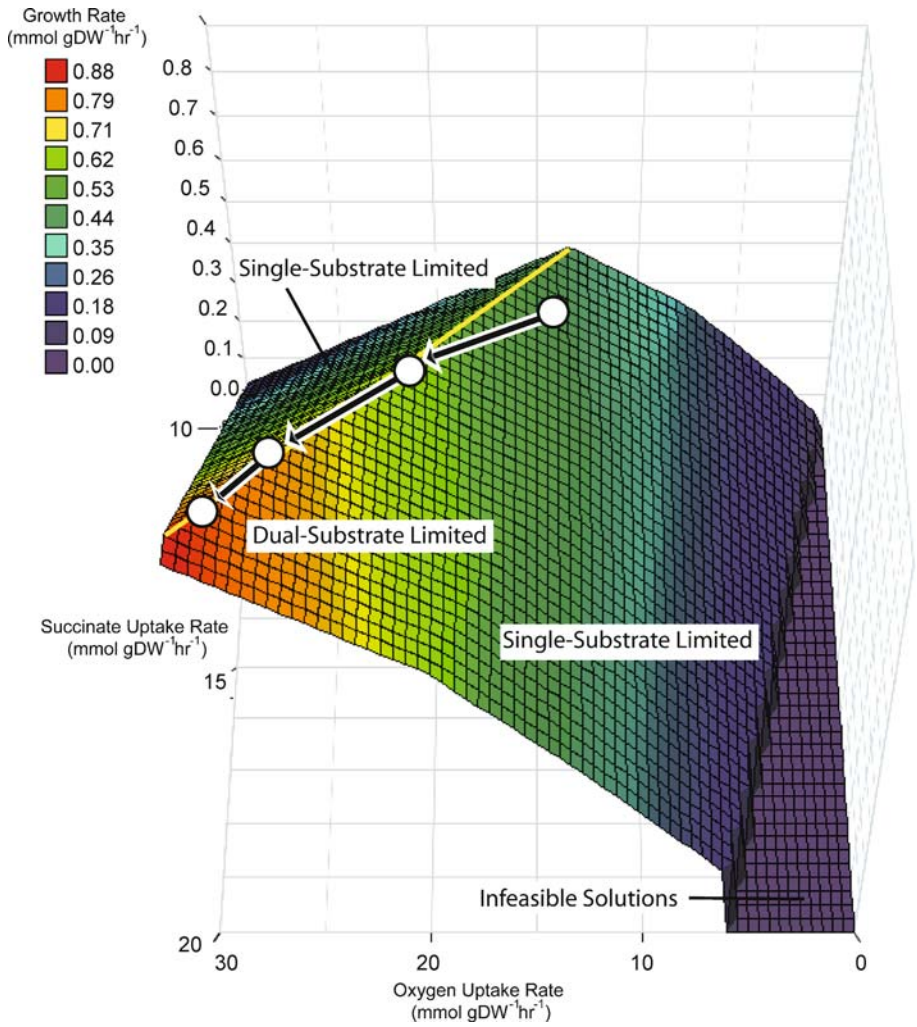

Metabolic Systems Biology, Figure 4

Phenotypic Phase Plane plot. PhPPs have been used to show that *E. coli* grown on a single carbon source does not always grow optimally (compared to *in silico* predictions). However, after growing exponentially on at least one such substrate, *E. coli* was shown to evolve to the line of optimality (*yellow line*) predicted in PhPPs [37]

For example, when bacteria are cultured in conditions selecting for growth, the cellular objective function is well-approximated by maximizing biomass production [21,86].

Flux Balance Analysis (FBA) is the formulation of linear programming problems for stoichiometric metabolic networks. Most of the biased methods employ variations or adaptations of FBA. For example, in Flux Variability Analysis (FVA) [56] every reaction in a condition-specific model is maximized and minimized. Gene Deletion Analysis [22], another FBA-based approach, involves the sequential deletion each gene in a model and optimization for growth in order to define the *in silico* knock-out phenotypes.

In **Phenotypic Phase Plane** (PhPP) analysis, a sensitivity analysis is conducted by varying two uptake or secretion network fluxes while calculating the optimal solution for the objective function (e. g., biomass production). These dependencies can be depicted graphically (Fig. 4). Each optimal solution is associated with a shadow price representing the change of the objective value when changing the availability of an external compound. The shadow prices can be used to define isoclines. An isocline is a line along which the shadow price is constant (Fig. 4). The line of optimality is a special isocline, which achieves the optimal objective [24] (yellow line in Fig. 4). The regions in PhPP plots can be classified into various

regions using isoclines: (1) futile regions occur when an increase in substrate availability leads to a decrease in the objective value; (2) single substrate limited regions occur when only one substrate increases the objective value; and (3) dual substrate limited regions occur when increases in either substrate leads to increases in the objective value [21,24]. PhPP has proven useful in designing experiments and predicting evolved microbial growth on different substrates [37].

Unbiased Methods Unbiased methods do not require the explicit definition of an objective function. These methods are of great use when the cellular objectives are not known or when a global view of all feasible *in silico* phenotypes is desired. At least three unbiased approaches have been developed which characterize the steady state solution space of the network include: ExPa analysis [62,69,81,102], Elementary Mode (ElMo) analysis [28,84,85,90], and uniform random sampling [1,71,93,101].

ExPa and ElMo Analysis have been useful over the past decade in elucidating metabolic network properties. ExPas are biochemically meaningful non-negative linear combinations of convex basis vectors of the steady state solution space [64,81]. ElMos are non-unique convex basis vectors that span the steady state solution space; hence they are a superset of ExPas. The main difference between ExPas and ElMos is in the definition of exchange reactions. In fact, when all exchange fluxes are set to be irreversible, ExPas and ElMos become identical [58]. The utility of ExPa and ElMo analysis has been demonstrated in determining network rigidity and redundancy in pathogenic microbes [61,69], proposing new media [38], and predicting regulatory points based on network structure [90]. The potential of ExPas and ElMos is limited, however, by the complexity and size of genome-scale metabolic networks, as the number of pathways increases dramatically for larger networks [48,102]. For example the core *E. coli* model (consisting of approximately 86 reactions) has approximately 20,000 ExPas in rich media growth conditions [58]. The number of ExPas in *E. coli* iJR904 [76], which consists of $\sim$ 900 reactions, has been estimated to be on the order of 10^{18} ExPas [102]. Even more impressive is the estimate of 10^{29} ExPas for the human metabolic network reconstruction [102]. These incredible pathway number estimates not only present insurmountable computational challenges but also significant difficulties in the analysis of ExPas and ElMos.

Another unbiased method is **uniform random sampling** of metabolic networks [1,101]. Uniform random sampling involves enumerating the candidate flux distributions in the steady state solution space until a statistical criterion is satisfied, e. g., a uniform set of flux distributions (Fig. 5a). There have been different approaches in implementing these procedures [71,93,101]. The distribution of random samples provides both a range of allowable fluxes and a probability distribution for flux values for the given set of constraints (Figs. 5b–d). This method not only allows for the analysis of the entire convex flux spaces of metabolic networks, but it can also be employed to study concave flux spaces and systems with non-linear constraints [72]. Thus it is apparent that uniform random sampling is a useful method that provides information about a metabolic network and a global view of all possible *in silico* phenotypes.

Co-set Analysis: Overlap between Biased and Unbiased Methods Since biased methods are a subset of unbiased methods, these two different methods can be used to calculate similar quantities, such as functionally correlated reactions, as demonstrated by **co-set analysis**. Correlated reaction sets (co-sets) are sets of reactions that are perfectly correlated ($R^2 = 1$) at the steady state (Fig. 6a-b) [63]. Reaction co-sets can be computed using different methods, such as flux coupling [14], ExPa analysis [62], or uniform random sampling [71,93]. There are pros and cons to using any of these methods. Flux coupling allows the identification of directionally coupled reactions, but requires a stated cellular objective. While ExPa analysis allows for the enumeration of co-sets without stating an objective, practical considerations such as computational time currently make ExPa analysis calculations infeasible for genome-scale models. For large networks, co-sets may be computed more rapidly using uniform random sampling. This method also allows the pairwise correlation coefficients of all reactions to be computed; therefore, partially correlated reaction sets ($R^2 < 1$) can be identified. These may be of interest when sampling the network under different environmental conditions or disease states [93].

Implementation of Constraint-Based Methods In order to make these analytical tools accessible to the scientific community, many of the methods in Table 1 have been implemented in MATLAB and released in the COBRA toolbox [8,103]. Other packages that implement various constraint-based methods include CellNetAnalyzer (FBA, ElMo analysis, and topological analysis) [47,104] and MetaFluxNet (FBA, reaction deletion analysis, and network visualization) [52]. These software packages and toolboxes are free of charge to the academic community.

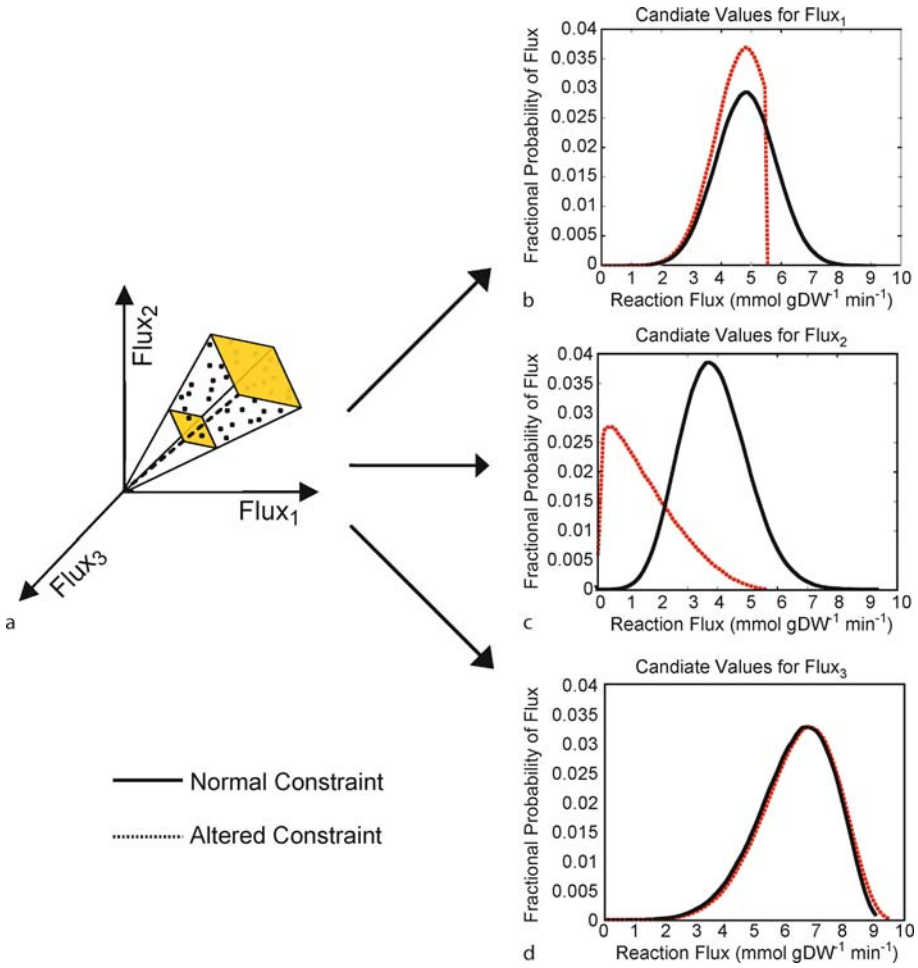

Metabolic Systems Biology, Figure 5

Uniform sampling of the solution space under normal and perturbed metabolic states. a Uniform random sampling of the solution space can be used to assign ranges of feasible fluxes and probability distributions for each reaction in the network. **b–d** The sample points can be visualized with a histogram for each network reaction (*black lines*). Measured changes in flux bounds (min/max) of network fluxes can be applied as network constraints, yielding altered flux distributions (*red lines*). These changes may reduce the maximum flux value (**b**), shift the most probable flux value (**c**), or leave the distribution unaltered (D)

Metabolic Systems Biology and Constraint-Based Modeling: Applications

As previously discussed, the past couple of decades have witnessed the development and analysis of constraint-based models. This has resulted in a wide array of analytical methods which have been employed to deepen the understanding of how biological systems function. The remainder of this chapter will discuss examples in which constraint-based models and methods were used to predict growth rates of evolved prokaryotic strains, design experiments, identify gene functions, characterize the effects of diseases and metabolic perturbations, classify genetic disorders, and propose alternative drug targets.

Growth Predictions of Evolved Strains

It has been hypothesized that incorrect log-phase growth predictions are caused by incomplete adaptation to a particular environmental given condition [37]. To test this hypothesis, *Escherichia coli* K-12 MG1655 was grown on different carbon sources (acetate, succinate, malate, glucose and glycerol) at varying concentrations and temperatures. PhPP analysis was carried out to identify the different phases and growth optimality for the different carbon sources (Fig. 4). Optimal growth in all of the substrates, except glycerol, were measured and found to lie on the calculated line of optimality (yellow line in Fig. 4). It was observed that after selecting for growth, the adap-

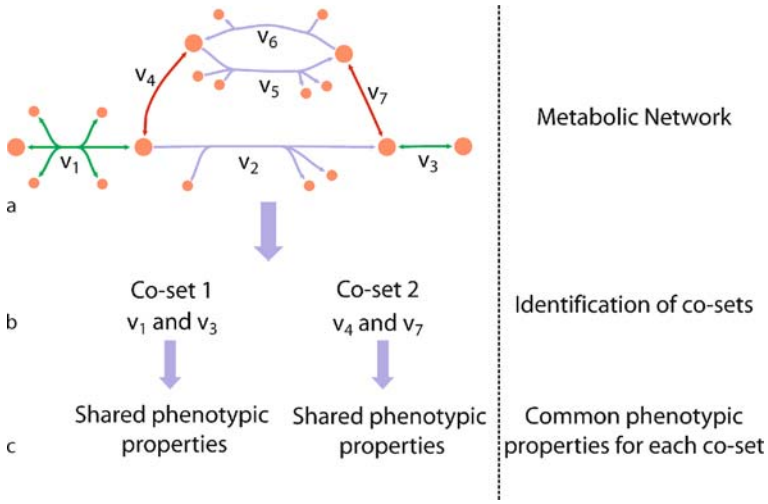

Metabolic Systems Biology, Figure 6

Mapping of SNPs onto reaction co-sets. a–c The identification of co-sets in a metabolic network leads to functionally grouped reaction sets. The reactions within each co-set are predicted to have similar disease phenotypes [39]. The same concept can be applied to the identification of alternative drug target candidates in humans to treat diseases [20] and potentially for the identification of alternative anti-microbial drug targets in human pathogens [40]

tive evolution of the parental strains (∼ 500 generations) led to increased growth rates while remaining on the line of optimality. This improved performance resulted from increased uptake of the carbon sources. The glycerol case however, showed that *E. coli* grows sub-optimally on this carbon source, i. e., the experimentally measured growth rate did not lie on the line of optimality; however, after 40 days (∼ 700 generations) the growth of the evolved strain moved to the line of optimality [37], and continued to evolve a higher growth rate while remaining on the line.

Discovery of Gene Function

When coupled with experimental data, genome-scale constraint based models can aid in hypothesis generation and can suggest functions for previously uncharacterized genes [55]. FBA was used to predict growth phenotypes of *E. coli* on a number of different carbon sources. Experimental growth phenotype data [11] was compared with the computational predictions to identify cases in which the model failed to accurately predict growth phenotypes (growth vs. non-growth) (Fig. 7a). In 54 cases, the model failed to predict experimentally measured growth phenotypes. Four failure modes were remedied using the literature, while the remaining 50 cases suggested incomplete knowledge. The computational algorithm outlined in Fig. 7b was used to predict potential reactions or transporters that could reconcile the model predictions and experimental results. Therefore, a universal database of all known metabolic reactions in living organisms [43] was queried and the minimum number of reactions needed to restore *in silico* growth of the model were computed. Solutions were found for 26 of the failure modes. A subset of the predicted solutions was chosen for experimental verification, and two sets will be discussed here.

The computational algorithm suggested the simplest solution to achieve growth on D-malate was decarboxylation of D-malate into pyruvate. A library of *E. coli* knockout mutants showed that three mutant strains demonstrated altered growth on D-malate: ΔdctA (slow growth), ΔyeaT, and ΔyeaU (both no growth). Through subsequent sequence homology analysis, gene expression measurement (with RT-PCR), and chromatin immunoprecipitation experiments, it was demonstrated that DctA is likely a transporter for D-malate, YeaU converts D-malate to pyruvate, and YeaT is a positive regulator that increases expression of yeaU.

Another example was L-galactonate. Affymetrix gene-expression data was used to identify genes involved with L-galactonate catabolism. Two candidates were found to be greatly upregulated: yjjL and yjjN. After additional experiments, these genes were annotated as follows: yjjL transports L-galactonate, yjjN is responsible for the L-galactonate oxidoreductase activity, and yjjM regulates their gene expression (Fig. 7c).

Successful *in silico* predictions can help to validate a model and unsuccessful predictions can provide opportunities to expand knowledge. These studies demonstrated

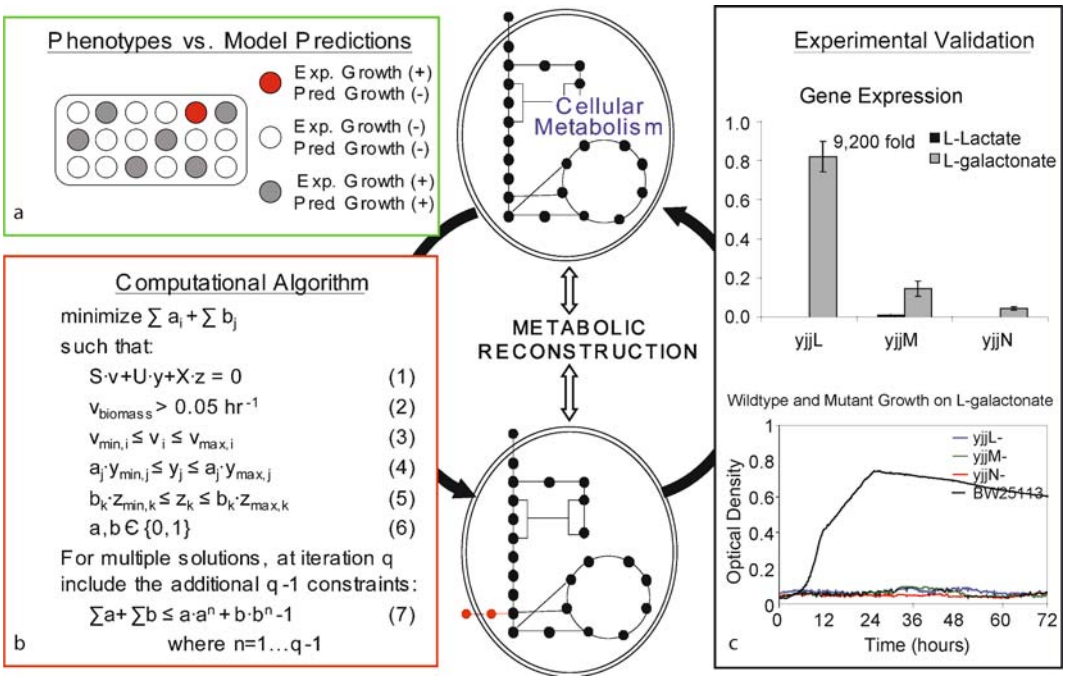

Metabolic Systems Biology, Figure 7

Refining genome annotation through computational prediction and experimental validation. a The validity of a metabolic model can be tested by comparing simulation outcomes with experimental results. In cases where the model fails to accurately predict the experimental outcome, **b** a computational algorithm can be employed that will predict the minimum number of reactions needed to reconcile the erroneous no-growth predictions from the model with the experimental data that demonstrates growth (Eq. 2). The reactions are selected from two matrices, **U** (containing all known metabolic reactions) and **X** (containing exchange reactions). The vectors **v**, **y** and **z** represent the steady state flux vectors for all of the reactions in **S**, **U**, and **X** respectively (Eq. 1, each with minimum and maximum fluxes as dictated in Eqs. 3–5. Vectors ***a*** and ***b*** are binary vectors in which an element is 1 only if the corresponding reaction in **U** or **X** is added to reconcile model with the experimental data. **c** For predicted sets of reactions, various experimental methods (e. g., growth phenotyping of gene knockout strains, measurement of gene expression levels, etc.) can be employed to validate predicted reaction sets. Accurate predictions of gene function allow for annotation of the associated genes and the corresponding reactions can be added to the network reconstruction for future use in models. Figure adapted with permission from [75]. Copyright© 2006 by The National Academy of Sciences of the United States of America, all rights reserved

that failed predictions could be used to algorithmically generate experimentally testable hypotheses and lead to refinement of the genome annotation.

Effects of Perturbed Mitochondrial States

A constraint-based network of a human cardiomyocyte mitochondrion has been used to evaluate candidate functional states in healthy and diseased individuals as well as investigate currently used therapies [93]. Uniform random sampling was used to assess all candidate metabolic flux states to characterize the effects of various metabolic perturbations, such as diabetes, ischemia, and various diets [93]. For each condition, additional constraints were applied to the network to represent the various conditions, e. g., uptake and secretion rates. It was found that the perturbations witnessed in diabetes and ischemia lead to a significant reduction of the size of the solution space, rendering the metabolic network less flexible to variations in nutrient availability (see Fig. 5).

Sampling under normal physiological conditions was found to be consistoent with experimental data, thus providing necessary network validation. Diabetic disease states were then simulated by increasing mitochondrial fatty acid uptake while decreasing cellular glucose uptake. The consequences of these constraints on the steady state solution space were found to be dramatic, meaning that for most network reactions, the range of flux values (flexibility) was significantly decreased. In particular, the oxygen requirement of the diabetic model was dramatically increased, which is consistent with the increased risk of cardiac complications seen in diabetic patients [91]. Another interesting observation was that the flux through mitochondrial pyruvate dehydrogenase was found to be

severely restricted due to network stoichiometry when fatty acid uptake was increased. Many prior studies had focused on potential inhibitory mechanisms leading to the decrease in pyruvate dehydrogenase in diabetic patients. However, the results of this study suggested that stoichiometry rather than inhibition may cause the reduced flux.

Causal SNP Classification Using Co-sets

Since reactions that are part of co-sets are either all on or off together, from the disease viewpoint, any enzyme deficiency that affects a reaction in a particular co-set would be expected to have disease phenotypes that are similar to the symptoms associated with enzymopathies of any other enzyme contained within the co-set (Fig. 6). The co-sets for the human cardiomyocyte mitochondrion were analyzed in the context of single nucleotide polymorphisms (SNPs), single base pair variations in the genes of individuals [39]. Causal SNPs result in altered phenotypes as a direct consequence of the altered genome sequence. The Online Mendelian Inheritance in Man (OMIM) database, which catalogs human genetic disorders, [31] and primary/review literature were used to map the nuclear encoded mitochondrial diseases caused by SNPs onto the co-sets using GPRs.

The resulting analysis largely confirmed the hypothesis that causal SNPs in the same reaction co-set often exhibited similar disease phenotypes. This phenotypic coherence was observed for the three different types of co-sets identified: Type A Co-sets which included sets of genes that code for sub-units of a single enzyme complex, Type B Co-sets which involved reactions in a linear pathway, and Type C Co-sets which involved non-contiguous reactions (see Fig. 6). Examples of diseases which exhibited similar phenotypes included porphyrias (Type B Co-set), fatty acid oxidation defects (Type B Co-set) and failure to thrive due to neurological problems (Type C Co-set).

It is important to recognize that these co-sets are condition dependent and have the potential to change as environmental conditions and nutrient availability vary. Furthermore it is not expected that the co-sets always have perfect agreement with clinical observations, since there are additional levels of information that are currently not accounted for in the models. This case study lent credibility to the hypothesis that co-sets can be used to understand and classify causal relationships in disease states. This concept can also be applied for proposing alternative drug targets for the treatment of disease (see Sect. "The Human Metabolic Network Reconstruction: Characterizing the Knowledge Landscape and a Framework for Drug Target Discovery") [20,40] and may serve as a rich source of hypothesis generation for alternative or new treatments for diseases.

The Human Metabolic Network Reconstruction: Characterizing the Knowledge Landscape and a Framework for Drug Target Discovery

While the human cardiac mitochondrion reconstruction has proven useful in the study of normal and diseased states, it only covers a small percentage of human metabolism. In order to account for cellular human metabolism more comprehensively, a genome-scale human reconstruction was created through a group effort resulting in the first manually curated human metabolic reconstruction, Recon 1 [20]. Recon 1 accounts for the functions of 1,496 ORFs, 2,766 metabolites, and 3,311 reactions. It accounts for the following compartments: cytoplasm, mitochondria, nucleus, endoplasmic reticulum, Golgi apparatus, lysosome, peroxisome and the extracellular environment. The network reconstruction was reconciled against 288 known metabolic functions in human.

Confidence scores were used to define the knowledge landscape of human metabolism. Of special interest are subsystems with low confidence scores, or experimental evidence, rendering them good candidates for experimental studies. For example, intracellular transport reactions and vitamin associated pathways were found to be consistently poorly characterized. Hence, the knowledge landscape provides an assessment of our current status of knowledge of human metabolism and a platform for discovery when combined with *in silico* methods (see Sect. "Discovery of Gene Function").

Another application of Recon 1 is the prediction of drug targets and consequences of metabolic perturbations using co-sets. For example, under aerobic glucose conditions, over 250 co-sets were identified. One of the largest co-sets contained the primary metabolic target, 3-Hydroxy-3-methylglutaryl-CoA Reductase, for the antilipidemic statin drugs. Following the logic of the SNPs study, the remaining members of the set are candidate drug targets for the treatment of hyperlipidemia.

Future Directions

A number of examples were discussed in this chapter demonstrating the use of constraint-based methods in biological discovery, disease characterization, and drug target prediction. Stemming from the success in these and many other applications, constraint-based genome-scale modeling will continue to be an actively growing area of research. Three general branches may include (1) the imposition of additional constraints, (2) the use of these models for bio-

logical discovery, and (3) the reconstruction of additional networks for other cellular processes.

The imposition of further constraints will reduce the size of the solution space through the elimination of biologically irrelevant flux distributions. Such constraints may include molecular crowding [9], regulation of enzymatic activity, and thermodynamic constraints [26,32]. Recently, approaches have been developed for the incorporation of metabolomic data into constraints-based models [10,16,51]. The availability of metabolomic data may also enable the application of additional physicochemical constraints, such as electroneutrality and osmotic balance [36,49]. There have also been increasing efforts to incorporate kinetic aspects into constraint-based modeling [25,89].

Furthermore, the algorithm shown in Fig. 7 demonstrated that there is still much that may be learned about metabolic networks, even in well studied organisms like *E. coli*. In addition, there are numerous ongoing research studies investigating human metabolism using Recon 1. Moreover, methods are being developed to use to constraint-based methods for engineering purposes which include strain design using bi-level optimizations [15,68] and genetic algorithms [66]. The uses of constraint-based models are also increasing to areas such as the investigation of antimicrobial agents [2,40,94].

Reconstruction of signaling networks in the constraint-based framework has been performed for the JAK-STAT pathways [59,60]. To date, there have been multiple efforts to formulate transcriptional regulatory networks based on literature and high-throughput data [5,6,19,30,34,88]. A more rigorous approach would be to describe the processes in a stoichiometric manner. This is exceedingly difficult and time-consuming, but recently progress has been made and is anticipated to be described in the near future [92]. As additional types of network reconstructions are developed and integrated with the concurrent development of new methods for analysis, the predictive capabilities and utility of these models are likely to expand.

Acknowledgments

This work was supported in part by NSF IGERT training grant DGE0504645.

Bibliography

Primary Literature

1. Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabasi AL (2004) Global organization of metabolic fluxes in the bacterium Escherichia coli. Nature 427:839–843
2. Almaas E, Oltvai ZN, Barabasi AL (2005) The activity reaction core and plasticity of metabolic networks. PLoS Comput Biol 1:e68
3. Alper H, Jin Y, Moxley JF, Stephanopoulos G (2005) Identifying gene targets for the metabolic engineering of lycopene biosynthesis in Escherichia coli. Metab Eng 7:155–164
4. Antoniewicz MR, Kelleher JK, Stephanopoulos G (2006) Determination of confidence intervals of metabolic fluxes estimated from stable isotope measurements. Metabol Eng 8:324–337
5. Barrett CL, Herring CD, Reed JL, Palsson BO (2005) The global transcriptional regulatory network for metabolism in Escherichia coli exhibits few dominant functional states. Proc Natl Acad Sci USA 102:19103–19108
6. Barrett CL, Palsson BO (2006) Iterative reconstruction of transcriptional regulatory networks: an algorithmic approach. PLoS Comput Biol 2:e52
7. Beard DA, Liang SD, Qian H (2002) Energy balance for analysis of complex metabolic networks. Biophys J 83:79–86
8. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nat Protoc 2:727–738
9. Beg QK, Vazquez A, Ernst J et al (2007) Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity. Proc Natl Acad Sci USA 104:12663–12668
10. Blank LM, Kuepfer L, Sauer U (2005) Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. Genome Biol 6:R49
11. Bochner BR, Gadzinski P, Panomitros E (2001) Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. Genome Res 11:1246–1255
12. Brumen M, Heinrich R (1984) A metabolic osmotic model of human erythrocytes. BioSystems 17:155–169
13. Burgard AP, Maranas CD (2003) Optimization-based framework for inferring and testing hypothesized metabolic objective functions. Biotechnol Bioeng 82:670–677
14. Burgard AP, Nikolaev EV, Schilling CH, Maranas CD (2004) Flux coupling analysis of genome-scale metabolic network reconstructions. Genome Res 14:301–312
15. Burgard AP, Pharkya P, Maranas CD (2003) Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol Bioeng 84:647–657
16. Cakir T, Patil KR, Onsan Z, Ulgen KO, Kirdar B, Nielsen J (2006) Integration of metabolome data with metabolic networks reveals reporter reactions. Mol Syst Biol 2:50
17. Caspi R, Foerster H, Fulcher CA et al (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 36(Database issue):D623–31
18. Church GM (2005) From systems biology to synthetic biology. Mol Syst Biol 1:2005.0032
19. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO (2004) Integrating high-throughput and computational data elucidates bacterial networks. Nature 429:92–96
20. Duarte NC, Becker SA, Jamshidi N et al (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci USA 104:1777–1782
21. Edwards JS, Ibarra RU, Palsson BO (2001) *In silico* predictions

of Escherichia coli metabolic capabilities are consistent with experimental data. Nat Biotechnol 19:125–130
22. Edwards JS, Palsson BO (2000) Metabolic flux balance analysis and the *in silico* analysis of Escherichia coli K-12 gene deletions. BMC Bioinformatics 1:1
23. Edwards JS, Palsson BO (2000) Robustness analysis of the Escherichia coli metabolic network. Biotechnol Prog 16:927–939
24. Edwards JS, Ramakrishna R, Palsson BO (2002) Characterizing the metabolic phenotype: a phenotype phase plane analysis. Biotechnol Bioeng 77:27–36
25. Famili I, Mahadevan R, Palsson BO (2005) k-Cone analysis: determining all candidate values for kinetic parameters on a network scale. Biophys J 88:1616–1625
26. Feist AM, Henry CS, Reed JL et al (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol 3:121
27. Fell DA, Small JR (1986) Fat synthesis in adipose tissue. An examination of stoichiometric constraints. Biochem J 238:781–786
28. Gagneur J, Klamt S (2004) Computation of elementary modes: a unifying framework and the new binary approach. BMC Bioinformatics 5:175
29. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 31:3784–3788
30. Gianchandani EP, Papin JA, Price ND, Joyce AR, Palsson BO (2006) Matrix formalism to describe functional states of transcriptional regulatory systems. PLoS ComputBiol 2:e101
31. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33:D514–7
32. Henry CS, Broadbelt LJ, Hatzimanikatis V (2007) Thermodynamics-based metabolic flux analysis. Biophys J 92:1792–1805
33. Herrgard MJ, Fong SS, Palsson BO (2006) Identification of genome-scale metabolic network models using experimentally measured flux profiles. PLoS Comput Biol 2:e72
34. Herrgard MJ, Lee BS, Portnoy V, Palsson BO (2006) Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae. Genome Res 16:627–635
35. Herring CD, Raghunathan A, Honisch C et al (2006) Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale. Nat Genet 38:1406–1412
36. Hohmann S, Krantz M, Nordlander B (2007) Yeast osmoregulation. Methods Enzymol 428:29–45
37. Ibarra RU, Edwards JS, Palsson BO (2002) Escherichia coli K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. Nature 420:186–189
38. Imielinski M, Belta C, Rubin H, Halasz A (2006) Systematic analysis of conservation relations in Escherichia coli genome-scale metabolic network reveals novel growth media. Biophys J 90:2659–2672
39. Jamshidi N, Palsson BO (2006) Systems biology of SNPs. Mol Syst Biol 2:38
40. Jamshidi N, Palsson BO (2007) Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the *in silico* strain iNJ661 and proposing alternative drug targets. BMC Syst Biol 1:26
41. Joshi A, Palsson BO (1989) Metabolic dynamics in the human red cell. Part I–A comprehensive kinetic model. J Theor Biol 141:515–528
42. Joshi A, Palsson BO (1989) Metabolic dynamics in the human red cell. Part II–Interactions with the environment. J Theor Biol 141:529–545
43. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32:D277–80
44. Kim PJ, Lee DY, Kim TY et al (2007) Metabolite essentiality elucidates robustness of Escherichia coli metabolism. Proc Natl Acad Sci USA 104:13638–13642
45. Kirschner MW (2005) The meaning of systems biology. Cell 121:503–504
46. Kitano H (2002) Computational systems biology. Nature 420:206–210
47. Klamt S, Saez-Rodriguez J, Gilles ED (2007) Structural and functional analysis of cellular networks with CellNetAnalyzer. BMC Syst Biol 1:2
48. Klamt S, Stelling J (2002) Combinatorial complexity of pathway analysis in metabolic networks. Mol Biol Rep 29:233–236
49. Klipp E, Nordlander B, Kruger R, Gennemark P, Hohmann S (2005) Integrative model of the response of yeast to osmotic shock. Nat Biotechnol 23:975–982
50. Knorr AL, Jain R, Srivastava R (2007) Bayesian-based selection of metabolic objective functions. Bioinformatics 23:351–357
51. Kummel A, Panke S, Heinemann M (2006) Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. Mol Syst Biol 2:2006.0034
52. Lee DY, Yun H, Park S, Lee SY (2003) MetaFluxNet: the management of metabolic reaction information and quantitative metabolic flux analysis. Bioinformatics 19:2144–2146
53. Lee S, Palakornkule C, Domach MM, Grossmann IE (2000) Recursive MILP model for finding all the alternate optima in LP models for metabolic networks. Comput Chem Eng 24:711–716
54. Lee SJ, Lee DY, Kim TY, Kim BH, Lee J, Lee SY (2005) Metabolic engineering of Escherichia coli for enhanced production of succinic acid, based on genome comparison and *in silico* gene knockout simulation. Appl Environ Microbiol 71:7880–7887
55. Leino RL, Gerhart DZ, van Bueren AM, McCall AL, Drewes LR (1997) Ultrastructural localization of GLUT 1 and GLUT 3 glucose transporters in rat brain. J Neurosci Res 49:617–626
56. Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab Eng 5:264–276
57. Marhl M, Schuster S, Brumen M, Heinrich R (1997) Modeling the interrelations between the calcium oscillations and ER membrane potential oscillations. Biophys Chem 63:221–239
58. Palsson BØ (2006) Systems biology: properties of reconstructed networks. Cambridge University Press, Cambridge, New York
59. Papin JA, Hunter T, Palsson BO, Subramaniam S (2005) Reconstruction of cellular signalling networks and analysis of their properties. Nat Rev Mol Cell Biol 6:99–111
60. Papin JA, Palsson BO (2004) The JAK-STAT signaling network in the human B-cell: an extreme signaling pathway analysis. Biophys J 87:37–46

61. Papin JA, Price ND, Edwards JS, Palsson BBO (2002) The genome-scale metabolic extreme pathway structure in Haemophilus influenzae shows significant network redundancy. J Theor Biol 215:67–82
62. Papin JA, Price ND, Palsson BO (2002) Extreme pathway lengths and reaction participation in genome-scale metabolic networks. Genome Res 12:1889–1900
63. Papin JA, Reed JL, Palsson BO (2004) Hierarchical thinking in network biology: the unbiased modularization of biochemical networks. Trends Biochem Sci 29:641–647
64. Papin JA, Stelling J, Price ND, Klamt S, Schuster S, Palsson BO (2004) Comparison of network-based pathway analysis methods. Trends Biotechnol 22:400–405
65. Papoutsakis ET (1984) Equations and calculations for fermentations of butyric acid bacteria. Biotechnol Bioeng 26:174–187
66. Patil KR, Rocha I, Forster J, Nielsen J (2005) Evolutionary programming as a platform for *in silico* metabolic engineering. BMC Bioinformatics 6:308
67. Pharkya P, Burgard AP, Maranas CD (2004) OptStrain: a computational framework for redesign of microbial production systems. Genome Res 14:2367–2376
68. Pharkya P, Maranas CD (2006) An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. Metab Eng 8:1–13
69. Price ND, Papin JA, Palsson BO (2002) Determination of redundancy and systems properties of the metabolic network of Helicobacter pylori using genome-scale extreme pathway analysis. Genome Res 12:760–769
70. Price ND, Reed JL, Palsson BO (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2:886–897
71. Price ND, Schellenberger J, Palsson BO (2004) Uniform sampling of steady-state flux spaces: means to design experiments and to interpret enzymopathies. Biophys J 87:2172–2186
72. Price ND, Thiele I, Palsson BO (2006) Candidate states of Helicobacter pylori's genome-scale metabolic network upon application of "loop law" thermodynamic constraints. Biophys J 90:3919–3928
73. Reed JL, Famili I, Thiele I, Palsson BO (2006) Towards multidimensional genome annotation. Nat Rev Genet 7:130–141
74. Reed JL, Palsson BO (2004) Genome-scale *in silico* models of E. coli have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. Genome Res 14:1797–1805
75. Reed JL, Patel TR, Chen KH et al (2006) Systems approach to refining genome annotation. Proc Natl Acad Sci USA 103:17480–17484
76. Reed JL, Vo TD, Schilling CH, Palsson BO (2003) An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). Genome Biol 4:R54
77. Riley M, Abe T, Arnaud MB et al (2006) Escherichia coli K-12: a cooperatively developed annotation snapshot–2005. Nucleic Acids Res 34:1–9
78. Satish Kumar V, Dasika MS, Maranas CD (2007) Optimization based automated curation of metabolic reconstructions. BMC Bioinformatics 8:212
79. Sauer U (2006) Metabolic networks in motion: 13C-based flux analysis. Mol Syst Biol 2:62
80. Savinell JM, Palsson BO (1992) Network analysis of intermediary metabolism using linear optimization. I. Development of mathematical formalism. J Theor Biol 154:421–454
81. Schilling CH, Letscher D, Palsson BO (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. J Theor Biol 203:229–248
82. Schomburg I, Chang A, Ebeling C et al (2004) BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res 32:D431–3
83. Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. Mol Syst Biol 3:119
84. Schuster S, Hilgetag C (1994) On Elementary Flux Modes in Biochemical Reaction Systems at Steady State. J Biol Syst 2:165–182
85. Schuster S, Hilgetag C, Woods JH, Fell DA (2002) Reaction routes in biochemical reaction systems: algebraic properties, validated calculation procedure and example from nucleotide metabolism. J Math Biol 45:153–181
86. Segre D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci USA 99:15112–15117
87. Shlomi T, Berkman O, Ruppin E (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. Proc Natl Acad Sci USA 102:7695–7700
88. Shlomi T, Eisenberg Y, Sharan R, Ruppin E (2007) A genome-scale computational study of the interplay between transcriptional regulation and metabolism. Mol Syst Biol 3:101
89. Smallbone K, Simeonidis E, Broomhead DS, Kell DB (2007) Something from nothing: bridging the gap between constraint-based and kinetic modelling. FEBS J 274:5576–5585
90. Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles ED (2002) Metabolic network structure determines key aspects of functionality and regulation. Nature 420:190–193
91. Taegtmeyer H, McNulty P, Young ME (2002) Adaptation and maladaptation of the heart in diabetes: Part I: general concepts. Circulation 105:1727–1733
92. Thiele I, Jamshidi N, Fleming RMT, Palsson BØ (2008) Genome-scale reconstruction of E. coli's transcriptional and translational machinery: A knowledge-base and its mathematical formulation. (under review)
93. Thiele I, Price ND, Vo TD, Palsson BO (2005) Candidate metabolic network states in human mitochondria. Impact of diabetes, ischemia, and diet. J Biol Chem 280:11683–11695
94. Trawick JD, Schilling CH (2006) Use of constraint-based modeling for the prediction and validation of antimicrobial targets. Biochem Pharmacol 71:1026–1035
95. Varma A, Boesch BW, Palsson BO (1993) Stoichiometric interpretation of Escherichia coli glucose catabolism under various oxygenation rates. Appl Environ Microbiol 59:2465–2473
96. Varma A, Palsson BO (1993) Metabolic Capabilities of Escherichia coli: I. Synthesis of Biosynthetic Precursors and Cofactors. J Theor Biol 165:477–502
97. Varma A, Palsson BO (1993) Metabolic Capabilities of Escherichia coli II. Optimal Growth Patterns. J Theor Biol 165:503–522
98. Varma A, Palsson BO (1994) Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. Nat Biotech 12:994–998
99. Vastrik I, D'Eustachio P, Schmidt E et al (2007) Reactome:

a knowledge base of biologic pathways and processes. Genome Biol 8:R39
100. Vo TD, Greenberg HJ, Palsson BO (2004) Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data. J BiolChem 279:39532–39540
101. Wiback SJ, Famili I, Greenberg HJ, Palsson BO (2004) Monte Carlo sampling can be used to determine the size and shape of the steady-state flux space. J Theor Biol 228:437–447
102. Yeung M, Thiele I, Palsson BO (2007) Estimation of the number of extreme pathways for metabolic networks. BMC Bioinformatics 8:363
103. http://systemsbiology.ucsd.edu/Downloads/. Accessed 4 Jul 2008
104. http://www.mpi-magdeburg.mpg.de/projects/cna/cna.html. Accessed 4 July 2008

Books and Reviews

Fiest AM, Palsson BØ (2008) The Growing Scope of Applications of Genome-scale Metabolic Reconstructions: the case of E. coli. Nat Biotechnol 26:659–667

Joyce AR, Palsson BØ (2007) Toward whole cell modeling and simulation: comprehensive functional genomics through the constraint-based approach. In: Boshoff HI, Barry III CE (eds) Systems Biological Approaches in Infectious Diseases. Birkhauser Verlag, Basel, pp 265, 267–309

Mo ML, Jamshidi N, Palsson BØ (2007) A Genome-scale, Constraint-Based Approach to Systems Biology of Human Metabolism. Mole Biosyst 3:9

Thiele I, Palsson BØ (2007) Bringing Genomes to Life: The Use of Genome-Scale *in silico* Models. In: Choi S (ed) Introduction to Systems Biology, Humana Press, Totowa

Microeconometrics

Pravin K. Trivedi
Department of Economics, Indiana University, Bloomington, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Historical Background
Two Leading Examples
Causal Modeling
New Directions in Structural Modeling
Major Insights
Bibliography

Glossary

Cowles commission approach An approach to structural econometric modeling identified with the pioneering work of the Cowles Foundation during the 1940s and 1950s.

Endogenous variable A variable whose value is determined within a specified model.

Exogenous A variable that is assumed given for the purposes of analysis because its value is determined outside the model of interest.

Reduced form models A stochastic model with relationships between endogenous variables on the one hand and all exogenous variables on the other.

Structural model A stochastic model with interdependent endogenous and exogenous variables.

Treatment effects An effect attributed to a change in the value of some policy variable analogous to a treatment in a clinical trial.

Definition of the Subject

Microeconometrics deals with model-based analysis of individual-level or grouped data on the economic behavior of individuals, households, establishments or firms. Regression methods applied to cross-section or panel (longitudinal) data constitute the core subject matter. Microeconometric methods are also broadly applicable to social and mathematical sciences that use statistical modeling. The data used in microeconometric modeling usually come from cross section and panel surveys, censuses, or social experiments. A major goal of microeconometric analysis is to inform matters of public policy. The methods of microeconometrics have also proved useful in providing model-based data summaries and prediction of hypothetical outcomes.

Introduction

Microeconometrics takes as its subject matter the regression-based modeling of economic relationships using data at the levels of individuals, households, and firms. A distinctive feature microeconometrics derives from the low level of aggregation in the data. This has immediate implications for the functional forms used to model analyze the relationships of interest. Disaggregation of data brings to the forefront heterogeneity of individuals, firms, and organizations. Modeling such heterogeneity is often essential for making valid inferences about the underlying relationships. Typically aggregation reduces noise and leads to smoothing due to averaging of movements in opposite directions whereas disaggregation leads to loss of continuity and smoothness. The range of variation in micro data is also typically greater. For example, household's average weekly consumption of (say) meat is likely to vary smoothly, while that of an individual household in a given

week may be frequently zero, and may also switch to positive values from time to time. Thus, micro data exhibit "holes, kinks and corners" [80]. The holes correspond to nonparticipation in the activity of interest, kinks correspond to the switching behavior, and corners correspond to the incidence of nonconsumption or nonparticipation at specific points of time. Consequently, discreteness and nonlinearity of response are intrinsic to microeconometrics.

Another distinctive feature of microeconometrics derives from the close integration of data and statistical modeling assumptions employed in analyzing them. Sample survey data, the raw material of microeconometrics, are subject to problems of complex survey methodology, departures from simple random sampling assumptions, and problems of sample selection, measurement errors, incomplete and/or missing data – problems that in principle impede the generalization from sample to population. Handling such issues is an essential component of microeconometric methodology.

An important application of microeconometrics is to tests predictions of microeconomic theory. Tests based on micro data are more attractive and relatively more persuasive because (a) the variables involved in such hypotheses can be measured more directly, (b) the hypotheses under test are likely to be developed from theories of individual behavior, and (c) a realistic portrayal of economic activity should accommodate a broad range of outcomes and responses that are a consequence of individual heterogeneity and that are predicted by underlying theory. In many public policy issues one is interested in the behavioral responses of a specific group of economic agents under some specified economic environment. One example is the impact of unemployment insurance on the job search behavior of young unemployed persons. To address these issues directly it is essential to use micro data.

The remainder of this article is organized as follows. In the next section I provide a historical perspective of the development of microeconometrics and sketch the topics in which important advances have occurred. In Sect. "Historical Background" we detail two models – the discrete choice model and the selection model – that are landmark developments in microeconometrics and provide important reference points for the remainder of the article. Sect. "Two Leading Examples" outlines three dominant modeling methodologies for structural modeling in microeconometrics. The final Sect. "Causal Modeling" surveys some of the major challenges in microeconometrics and the available modeling tools for dealing with these challenges. To stay within space constraints, I emphasize developments that have influenced microeconomic data analysis, and pay less attention to general theoretical analyzes.

Historical Background

Analysis of individual data has a long history. Engel [23], Allen and Bowley [2], Houthakker [43], and Prais and Houthakker [79] all made pioneering contributions to the research on consumer behavior using household budget data. Other seminal studies include Marschak and Andrews [77] in production theory, and Stone [86], and Tobin [88] in consumer demand. Nevertheless, the path-breaking econometric developments initiated by the Cowles Foundation during the 1940s and 1950s were motivated by concerns of macroeconomic modeling. The initial impact of this research was therefore largely on the development of large-scale multi-equation aggregate models of sectors and the economy. Although the Cowles Commission work was centered on the linear simultaneous equations model (SEM), while modern microeconometrics emphasizes nonlineairties and discreteness, the SEM conceptual framework has proved to be a crucial and formative influence in structural microeconometric modeling.

The early microeconometric work, with the important exception of Tobin [88], relied mainly on linear models, with little accommodation of discreteness, kinks, and corners. Daniel McFadden's [68] work on analysis of discrete choice and James Heckman's [30,31,32,33] work on models of truncation, censoring and sample selection, which combined discrete and continuous outcomes, were path-breaking developments that pioneered the development of modern microeconometrics. These developments overlapped with the availability of large micro data sets beginning in the 1950s and 1960s.

These works were a major departure from the overwhelming reliance on linear models that characterized earlier work. Subsequently, they have led to major methodological innovations in econometrics. Among the earlier textbook level treatment of this material (and more) are Maddala [76] and Amemiya [3]. As emphasized by Heckman [35], McFadden [73] and others, many of the fundamental issues that dominated earlier work based on market data remain important, especially concerning the conditions necessary for identifiability of causal economic relations. But the style of microeconometrics is sufficiently distinct to justify writing a text that is exclusively devoted to it.

Modern microeconometrics based on individual, household, and establishment level data owes a great deal to the greater availability of data from cross section and longitudinal sample surveys and census data. In the last

two decades, with the expansion of electronic recording and collection of data at the individual level, data volume has grown explosively. So too has the available computing power for analyzing large and complex data sets. In many cases event level data are available; for example, marketing science often deals with purchase data collected by electronic scanners in supermarkets, and industrial organization literature contains econometric analyzes of airline travel data collected by online booking systems. New branches of economics, such as social experimentation and experimental economics, have opened up that generate "experimental" data. These developments have created many new modeling opportunities that are absent when only aggregated market level data are available. At the same time the explosive growth in the volume and types of data has also given rise to numerous methodological issues. Processing and econometric analysis of such large micro data bases, with the objective of uncovering patterns of economic behavior, constitutes the core of microeconometrics. Econometric analysis of such data is the subject matter of this book.

Areas of Advances

Both historically and currently, microeconometrics concentrates on the so-called limited dependent variable (LDV) models. The LDV class deals with models in which the outcome of interest has a limited range of variation, in contrast to the case where variation is continuous and on the entire real line. Examples are binary valued outcomes, polychotomous outcomes, non-negative integer-valued outcomes, and truncated or censored variables where values outside a certain range are not observed. An example of censoring arises in modeling the labor supply of working women. Here the data refers to the number of hours of work of the employed women even though from an empirical perspective the economist is interested in both the decision to participate in the labor force (extensive margin) and also in the choice of hours of work (intensive margin) conditional on participation. From this perspective the sample on hours of work is censored and the analysis of hours of work of only those who participate potentially suffers from "selection bias". Analysis of transitions between states and of time spent in a state, e. g. unemployment, using the methods of hazard (survival) analysis also confronted the issue of truncation and censoring, since in many cases the spells of unemployment (durations) were only partially observed. Many economic outcomes such as choice of occupation or travel model, and event counts are inherently discrete and hence fall in the LDV class. Many others involve interdependent discrete and continuous outcomes, e. g. participation and hours of work.

- LDV topics have maintained their core status in the area. But their scope has expanded to include count data models [12] and a much wider variety of selection models. Whereas in 1975 virtually all of the models of discrete choice were static and cross sectional, now discrete choice analysis has developed in many directions, including dynamic aspects which permit dependence between past, current and future discrete choices. Dynamic discrete choice modeling is now embedded in dynamic programming models [22,83]. Individuals often state their preferences over hypothetical choices (as when they are asked to reveal preferences over goods and services not yet in the market place), and they also reveal their preferences in the market place. Modern discrete choice analysis integrates stated preferences and revealed preferences [89,90].
- In 1975 the subject of multivariate and structural estimation of discrete response models required further work in almost every respect. In modern microeconometric models generally, and discrete choice models specifically, there is greater emphasis on modeling data using flexible functional forms and allowing for heterogeneity. This often leads to mixture versions of these models. Advances in computer hardware and software technologies have made simulation-based methods of all types, including Bayesian Markov chain Monte Carlo methods, more accessible to practitioners. Varieties of LDV models that were previously outside the reach of practitioners are now widely used. Inference based on resampling methods such as bootstrap that do not require closed form expressions for asymptotic variances are now quite common in microeconometrics.
- Extensions of many, if not most, LDV models to allow for panel data are now available [44]. Random effects panel models are especially amenable to simulation-based estimation. There have been important advances in handling advanced linear panel data models (including dynamic panels) and nonlinear panel data models – especially models for binary and multinomial outcomes, censored variables, count variables, all of which are now more accessible to practitioners.
- Bayesian approaches are well-suited for analyzing complex LDV because they efficiently exploit the underlying latent variable structure. Bayesian analysis of LDV models is well-developed in the literature, but its incorporation into mainstream texts still lags [53]. Specialized monographs and texts, however, fill this gap.

- Treatment evaluation, which deals with measurement of policy impact at micro level, is now conspicuous and major new topic. The impact of the topic is broad because treatment evaluation is discussed in the context of many different LDV models, using a variety of parametric and semi- or nonparametric approaches, under a variety of different assumptions about the impact of treatment. The literature on this topic is now very extensive, see Heckman and Robb [36], Imbens and Angrist [45], Heckman and Vytlacil [39], and Lee [58] for a monograph-length treatment.
- Topic related to data structures now receive more attention. This includes the pros and cons of observational data and those from social and natural experiments. These topics arise naturally in the context of treatment evaluation. Other data related topics such as survey design and methodology, cross sectional and spatial dependence, clustered observations, and missing data also get greater attention.
- As regards estimation and inference, the classical methods of maximum likelihood, least squares and method of moments were previously dominant, with some exceptions. These methods typically make strong distributional and functional form assumptions that are often viewed with skepticism because of their potential impact on policy conclusions. By contrast, there is now a greater variety of semiparametric estimators in use, of which quantile regression is a leading example [51]. Nonparametric regression is another new topic. There is now a large literature dealing with most standard models and issues from a semi-parametric viewpoint.

Two Leading Examples

To illustrate some salient features of microeconometrics, the structure of two leading models, the first one for discrete choice and the second for sample selection, will be described and explained. Latent variables play a key role in the specification of both models, and in the specification of LDV models more generally. Distributional and structural restrictions are usually imposed through the latent variable specifications. Estimation of the models can also exploit the latent variable structure of such models.

Example 1: Random Utility Model

McFadden played a major role in the development of the random utility model (RUM) that provides the basis of discrete choice analysis; see McFadden [68,70,71,72]. Discrete choice models, firmly established in the analysis of transport mode choice, are now used extensively to model choice of occupations, purchase of consumer durables and brand choice.

The RUM framework is an extension of Thurstone [87]. In the binary RUM framework the agent chooses between alternatives 0 and 1 according to which leads to higher satisfaction or utility which is treated as a latent variable. The observed discrete variable y then takes value 1 if alternative 1 has higher utility, and takes value 0 otherwise. The additive random utility model (ARUM) specifies the utilities of alternatives 0 and 1 to be

$$\begin{aligned} U_0 &= V_0 + \varepsilon_0 \\ U_1 &= V_1 + \varepsilon_1 \,, \end{aligned} \tag{1}$$

where V_0 and V_1 are deterministic components of utility and ε_0 and ε_1 are random components of utility. The alternative with higher utility is chosen. We observe $y = 1$, say, if $U_1 > U_0$. Due to the presence of the random components of utility this is a random event with

$$\begin{aligned} \Pr\left[y = 1 | V_0, V_1\right] &= \Pr\left[U_1 > U_0\right] \\ &= \Pr\left[V_1 + \varepsilon_1 > V_0 + \varepsilon_0\right] \\ &= \Pr\left[\varepsilon_0 - \varepsilon_1 < V_1 - V_0\right] \\ &= F\left(V_1 - V_0\right) \,, \end{aligned} \tag{2}$$

where F is the c.d.f. of $(\varepsilon_0 - \varepsilon_1)$. This yields $\Pr[y = 1] = F(\mathbf{x}'\boldsymbol{\beta})$ if $V_1 - V_0 = \mathbf{x}'\boldsymbol{\beta}$. Different choices of the functional form F generate different parametric models of binary choice (outcome).

The additive RUM model has multivariate extensions. In the general m-choice multinomial model the utility of the jth choice is specified to be given by

$$U_j = V_j + \varepsilon_j \,, \qquad j = 1, 2, \ldots, m \,, \tag{3}$$

where V_j, the deterministic component of utility may be specified to be a linear index function, e. g. $V_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta}$ or $V_{ij} = \mathbf{x}_i'\boldsymbol{\beta}_j$, and ε_j denotes the random component of utility. Suppressing the individual subscript i for simplicity, using algebraic manipulations similar to those for the binary case, we obtain

$$\begin{aligned} \Pr[y = j] &= \Pr\left[U_j \geq U_k \,, \ \text{ all } \ k \neq j\right] \\ &= \Pr\left[\widetilde{\varepsilon}_{kj} \leq -\widetilde{V}_{kj} \,, \ \text{ all } \ k \neq j\right] \,, \end{aligned} \tag{4}$$

where the tilda and second subscript j denotes differencing with respect to reference alternative j.

Consider an individual choosing a mode of transport to work where the choice set consists of train, bus, or priavte car. Each mode has associated with it a deterministic utility that depends upon attributes (e. g. money cost,

time cost) of the mode and a random idiosyncratic component ("error"). Empirically the goal is to model conditional choice probabilities in terms of the mode attributes. Different multinomial models can be generated by different assumptions about the joint distribution of the error terms. These models are valid statistically, with probabilities summing to one. Additionally they are consistent with standard economic theory of rational decision-making. The idiosyncratic components of choice should exhibit correlation across choices if the alternatives are similar. For example, if the random components have independent type I extreme value distributions (a strong assumption!), then

$$\Pr[y = j] = \frac{e^{V_j}}{e^{V_1} + e^{V_2} + \cdots e^{V_m}} . \tag{5}$$

This is the conditional logit (CL) model when $V_j = \mathbf{x}_j'\boldsymbol{\beta}$, which means that attributes vary across choices only, and the multinomial legit (MNL) when $V_j = \mathbf{x}'\boldsymbol{\beta}_j$, which means that attributes are individual- but not choice-specific. Assuming that the random components have a joint multivariate normal distribution, which permits idiosyncratic components of utility to be correlated, generates the multinomial probit (MNP) model. The MNL is a special case of the Luce [59] model; it embodies an important structural restriction that the odds ratio for pair (i, j), $\Pr[y = i]/\Pr[y = j]$, is independent of all other available alternatives IIA). The MNP is the less restrictive Thurstone model, which allows for dependence between choices.

Multinomial Logit and Extensions The MNL model is much easier to compute than the MNP, but there is motivation for extending the MNL to allow for dependence in choices. One popular alternative is based on the generalized extreme value (GEV) model proposed by McFadden et al. [70], which leads to the nested logit (NL) model.

The GEV distribution is

$$F(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m) = \exp\left[-G\left(e^{-\varepsilon_1}, e^{-\varepsilon_2}, \ldots, e^{-\varepsilon_m}\right)\right]$$

where the function $G(Y_1, Y_2, \ldots, Y_m)$is chosen to satisfy several assumptions that ensure the joint distribution and resulting marginal distributions are well-defined.

If the errors are GEV distributed then an explicit solution for the probabilities in the RUM can be obtained, with

$$p_j = \Pr[y = j] = e^{V_j} \frac{G_j\left(e^{-V_1}, e^{-V_2}, \ldots, e^{-V_m}\right)}{G\left(e^{-V_1}, e^{-V_2}, \ldots, e^{-V_m}\right)} , \tag{6}$$

where $G_j(Y_1, Y_2, \ldots, Y_m) = \partial G(Y_1, Y_2, \ldots, Y_m)/\partial Y_j$, see McFadden (p. 81 in [70]). A wide range of models can be obtained by different choices of $G(Y_1, Y_2, \ldots, Y_m)$.

The nested logit model of McFadden [70] arises when the error terms ε_{jk} have the GEV joint cumulative distribution function

$$F(\boldsymbol{\varepsilon}) = \exp\left[-G\left(e^{-\varepsilon_{11}}, \ldots, e^{-\varepsilon_{1K_1}}; \ldots ; e^{-\varepsilon_{J1}}, \ldots, e^{-\varepsilon_{JK_J}}\right)\right] \tag{7}$$

for the following particular specification of the function $G(\cdot)$,

$$G(\mathbf{Y}) = G\left(Y_{11}, \ldots, Y_{1K_1}, \ldots, Y_{J1}, \ldots, Y_{JK_J}\right) = \sum_{j=1}^{J}\left(\sum_{k=1}^{K_j} Y_{jk}^{1/\rho_j}\right)^{1-\rho_j} . \tag{8}$$

The parameter ρ_j is a function of the correlation between ε_{jk} and ε_{jl} (see [13], p. 509).

The nested logit model specifies choice-making as a hierarchical process. A simple example is to consider choice of a television, where one first decides whether to buy a LCD screen or a plasma screen, and then conditional on that first choice which brand.

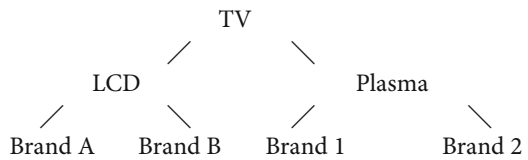

The random components in an RUM are permitted to be correlated for each option within the LCD and plasma groups, but are uncorrelated across these two groups. The GEV model can be estimated recursively by fitting a sequence of MNL models.

Multinomial Probit Another way to remove the IIA restriction is to assume that the unobserved components have a joint multivariate normal distribution. Beginning with m-choice multinomial model, with utility of the jth choice given by $U_j = V_j + \varepsilon_j$, $j = 1, 2, \ldots, m$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}[\mathbf{0}, \boldsymbol{\Sigma}]$, where the $m \times 1$ vector $\boldsymbol{\varepsilon} = [\varepsilon_1 \ldots \varepsilon_m]'$.

If the maximum likelihood equations have a unique solution for the parameters of interest, the model is said to be identified. In case that the number of equations is insufficient to yield unique estimates, restrictions on $\boldsymbol{\Sigma}$ are needed to ensure *identification*. Bunch [11] demonstrated that all but one of the parameters of the covariance matrix of the errors $\varepsilon_j - \varepsilon_1$ is identified. This can be achieved if we normalize $\varepsilon_1 = 0$, say, and then restrict one covariance element. Additional restrictions on $\boldsymbol{\Sigma}$ or $\boldsymbol{\beta}$ may be needed

for successful application, especially in models where there are no alternative-specific covariates [47]. That is, even when a MNP model is technically identified, the identification may be fragile in some circumstances, thus requiring further restrictions.

A natural estimator for this model is maximum likelihood. But, as mentioned in Sect. "Introduction", this poses a computational challenge as there is no analytical expression for the choice probabilities. For example, when $m = 3$,

$$p_1 = \Pr[y = 1] = \int_{-\infty}^{-\widetilde{V}_{31}} \int_{-\infty}^{-\widetilde{V}_{21}} f(\widetilde{\varepsilon}_{21}, \widetilde{\varepsilon}_{31})\, \mathrm{d}\widetilde{\varepsilon}_{21}\, \mathrm{d}\widetilde{\varepsilon}_{31}\,,$$

where $f(\widetilde{\varepsilon}_{21}, \widetilde{\varepsilon}_{31})$ is a bivariate normal with as many as two free covariance parameters and $\widetilde{V}_{21}$ and $\widetilde{V}_{31}$ depend on regressors and parameters $\boldsymbol{\beta}$. This bivariate normal integral can be quickly evaluated numerically, but a trivariate normal integral is the limit for numerical methods. In practice it is rare to see MNP applied when there are more than 4 choices.

Simulation methods are a potential solution for higher dimensional models [89]. For Monte Carlo integration over a region of the multivariate normal, a very popular smooth GHK simulator simulator is the GHK simulator, due to Geweke [25], Hajivassiliou et al. [29] and Keane [48]; see Train [89] for details. This discussion takes $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ as given but in practice these are estimated. The maximum simulated likelihood estimator (MSL) maximizes

$$\widehat{\mathrm{L}}_N(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \sum_{i=1}^{N} \sum_{j=1}^{m} y_{ij} \ln \widehat{p}_{ij}\,,$$

where the $\widehat{p}_{ij}$ are obtained using the GHK or other simulator. Consistency requires the number of draws in the simulator $S \to \infty$ as well as $N \to \infty$. The method is very burdensome, especially in high dimensions. This increases the appeal of alternative estimation procedures such as the method of simulated moments (MSM). The MSM estimator of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ solves the estimating equations

$$\sum_{i=1}^{N} \sum_{j=1}^{m} (y_{ij} - \widehat{p}_{ij}) \mathbf{z}_i = \mathbf{0}\,,$$

where the $\widehat{p}_{ij}$ are obtained using an unbiased simulator. Because, consistent estimation is possible even if $S = 1$, MSM is computationally less burdensome.

Finally, Bayesian methods that exploit the latent variable structure using data augmentation approach and Markov chain Monte Carlo methods have been used successfully; see Albert and Chib [1] and McCulloch and Rossi [67].

Choice probability models are of interest on their own. More usually, however, they are of interest when linked to models of other outcomes. In observational data it is common to study outcomes that are jointly determined with the choices, often through the common dependence of the two on idiosyncratic elements. Even when the main interest is in the outcome variable, modeling of the choice component is integral to the analysis. Selection models are an example of such joint models.

Example 2: Sample Selection Models

One of the most important classes of microeconometric models is the sample selection model. Goal of modeling is usually valid inference about a target population. Sample selection problem refers to the problem of making valid inference because the sample used is not representative of the target population. Observational studies are generally based on pure random samples. A sample is broadly defined to be a selected sample if, for example, it is based in part on values taken by a dependent variable. A variety of selection models arise from the many ways in which a sample may be selected, and some of these may easily go undetected.

There is a distinction between self-selection, in which the outcome of interest is determined in part by individual choice of whether or not to participate in the activity of interest, and sample-selection, in which the participants in the activity of interest are over- or under-sampled. Selection models involve modeling the participation into the activity of interest, e. g., the labor force. The outcomes of those who participate can be compared with those of non-participants, which generates the counterfactual of interest. Generating and comparing counterfactuals is a fundamental aspect of selection models. Elsewhere this topic of counterfactual analysis is called treatment evaluation. When treatment evaluation is based on observational data, issues of sample selection and self-selection almost always arise.

In the example given below, consistent estimation relies on relatively strong distributional assumptions, whereas the modern trend is to do so under weaker assumptions. The example illustrates several features of microeconometric models; specifically, the model is mixed discrete-continuous and involves truncation and latent variables.

Let y_2^* denote the outcome of interest that is observed if $y_1^* > 0$. For example, y_1^* determines whether or not to work (participation) and y_2^* determines how many hours

to work (outcome). The bivariate sample selection model has a participation equation,

$$y_1 = \begin{cases} 1 & \text{if } y_1^* > 0 \\ 0 & \text{if } y_1^* \leq 0\,, \end{cases} \tag{9}$$

and an outcome equation,

$$y_2 = \begin{cases} y_2^* & \text{if } y_1^* > 0 \\ - & \text{if } y_1^* \leq 0\,. \end{cases} \tag{10}$$

This model specifies that y_2 is observed when $y_1^* > 0$, possibly taking a negative value, while y_2 need not take on any meaningful value when $y_1^* \leq 0$.

The standard specification of the model is a linear model with additive errors for the latent variables, so

$$\begin{aligned} y_1^* &= \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1 \\ y_2^* &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2\,, \end{aligned} \tag{11}$$

where problems arise in estimating $\boldsymbol{\beta}_2$ if ε_1 and ε_2 are correlated. If β_2 were estimated using a regression of y_2 on $\mathbf{x}_2$ using only the part of the sample for which $y_2 = y_2^*$, the resulting estimates would suffer from sample selection bias. The classic early application of this model was to labor supply, where y_1^* is the unobserved desire or propensity to work, while y_2 is actual hours worked. Heckman [33] used this model to illustrate estimation given sample selection. A popular parametric specification assumes that the correlated errors are joint normally distributed and homoskedastic, with

$$\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \sim \mathcal{N}\left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right]. \tag{12}$$

which uses the normalization $\sigma_1^2 = 1$ because y_1^* is a latent variable that needs a measurement scale. Under general assumptions, and not just bivariate normality, the bivariate sample selection model therefore has likelihood function

$$L = \prod_{i=1}^{n} \left\{\Pr\left[y_{1i}^* \leq 0\right]\right\}^{1-y_{1i}} \left\{f\left(y_{2i} \mid y_{1i}^* > 0\right) \times \Pr\left[y_{1i}{}^* > 0\right]\right\}^{y_{1i}}\,, \tag{13}$$

where the first term is the contribution when $y_{1i}^* \leq 0$, since then $y_{1i} = 0$, and the second term is the contribution when $y_{1i}^* > 0$. The model is easily estimated if it is specialized to the linear models with joint normal errors, see Amemiya [3]. An important component of the identification strategy is the use of exclusion restriction(s). This refers to the restriction that some component(s) of $\mathbf{x}_1$ affects the choice variable y_1 only, and not the outcome variable. The intuition is that this provides a source of independent variation in y_1 that can robustly identify the parameters in the y_2-equation.

The maximum likelihood approach to the estimation of self-selection models can be extended to the polychotomous choice with m-alternatives by first specifying a parametric model for choice probability that takes the form of a multinomial or nested logit, or multinomial probit, and then specifying a joint distribution between the outcome of interest and the choice probabilities; see, for example, Dubin and McFadden [20]. While straight-forward in principle, this approach does pose computational challenges. This is because analytic expressions for such joint distributions are in general not available. The problem can be addressed either by using simulation-based methods or by taking a semi-parametric formulation that permits two-step estimation of the model parameters. This topic is discussed further in Sect. "Causal Modeling".

Manski [61] and Heckman [31] were early advocates of flexible semi-parametric estimation methods, of which the "two-step Heckman procedure" is a leading example. This influential modern approach seeks to avoid strong distributional and functional form assumptions and yet obtain consistent estimates with high efficiency within this class of estimators. Following in that tradition, there is a large literature, surveyed in Lee [56], that follows the semi-parametric approach. As the dependence between choices and outcomes are central to the issue, semi-parametric IV estimators are a natural choice. One strand of the literature, represented by Blundell and Powell [9], approaches this issue form a general semiparametric IV viewpoint, whereas another, represented by Lee [58] approaches this from the perspective of linear simultaneous equations viewpoint. Whereas the latent variable approach dominates discrete choice and selection models, some econometricians, e. g. Manski [62], espouse a less restrictive model that uses the basic probability formulation of the problem, with little other structure, that can still deliver informative bounds on some counterfactual outcomes. (There are also other econometric contexts in which the bounds approach can be applied; see [63].)

Causal Modeling

An important motivation for microeconometrics stems from issues of public policies that address social and economic problems of specific groups whose members react to policies in diverse ways. Then microeconometric models are used to evaluate the impact of policy. A leading example is the effect of training on jobless workers as defined in terms of their post-training wage. Accordingly, an important topic in microeconometrics is treatment evaluation. The term treatment refers to a policy and the analogy is with the model of a clinical trial with randomized as-

signment to treatment. The goal is to estimate the average effect of the treatment.

Heckman [35] has pointed out that there are two types of policy evaluation questions. The first type seeks to evaluate the effect of an existing program or policy on participants relative to an alternative program or no program at all, i. e. a treatment effect. The second formulation addresses a more difficult and ambitious task of evaluating the effect of a new program or policy for which there are no historical antecedents, or of an existing program in a new economic environment. A basic tenet of econometric modeling for policy analysis is that a structural model is required to address such policy issues.

As to how exactly to define a structural model is a difficult and unsettled issue. Indeed it is easier to say what structural models are not than to define what they are. Some modelers define structural models as those that identify parameters that are invariant with respect to policies themselves; others define structural models as those that involve mathematical-statistical relationships between jointly dependent variables, and yet others define them as relationships based on dynamic optimizing models of economic behavior that embody "fundamental" taste, technology and preference parameters.

In the next section I shall provide an overview of three major approaches to causal modeling in microeconometrics. Three dominant approaches are based on, respectively, moment conditions, the potential outcome model, and the dynamic discrete choice approach.

Structural Modeling

Broadly, structural model refers to causal rather than associative modeling. Cameron and Trivedi [13] provide a definition of a structure that is based on the distinction between exogenous variables $\mathbf{Z}$, that are taken by the modeler as given, and endogenous variables $\mathbf{Y}$, that the modeler attempts to explain within the model; this distinction derives from the classic Cowles Commission approach for the dynamic linear SEM mentioned earlier. The dynamic linear structural SEM specifies a complete model for G endogenous variables, specified to be related to K exogenous a pre-determined variables (e. g. lagged values of $\mathbf{Y}$).

Accordingly, a structure consists of

1. a set of variables $\mathbf{W}$ ("data") partitioned for convenience as $[\mathbf{Y}\sim\mathbf{Z}]$;
2. a joint probability distribution of $\mathbf{W}$, $F(\mathbf{W})$;
3. an a priori ordering of $\mathbf{W}$ according to hypothetical cause and effect relationships and specification of a priori restrictions on the hypothesized model;
4. a parametric, semiparametric or nonparametric specification of functional forms and the restrictions on the parameters of the model.

Suppose that the modeling objective is to explain the values of observable vector-valued variable $\mathbf{y}$, $\mathbf{y}' = (y_1, \ldots, y_G)$, whose elements are functions of some other elements of $\mathbf{y}$, and of explanatory variables $\mathbf{z}$ and a purely random disturbance u. Under the exogeneity assumption interdependence between elements of $\mathbf{z}$ is not modeled. The ith observation satisfies the set of implicit equations

$$\mathbf{g}(\mathbf{w}_i, \mathbf{u}_i|\boldsymbol{\theta}_0) = \mathbf{0}\,, \tag{14}$$

where $\mathbf{g}$ is a known function. By the Cameron–Trivedi definition this is a structural model, and to $\boldsymbol{\theta}_0$ is the vector of structural parameters. This corresponds to point 4 given earlier in this section. If there is a unique solution for $\mathbf{y}_i$ for every $(\mathbf{z}_i, \mathbf{u}_i)$, i. e.

$$\mathbf{y}_i = \mathbf{f}(\mathbf{z}_i, \mathbf{u}_i|\boldsymbol{\pi})\,, \tag{15}$$

then this is referred to as the reduced form of the structural model, where $\boldsymbol{\pi}$ is a vector of reduced form parameters that are functions of $\boldsymbol{\theta}$. The reduced form is obtained by solving the structural model for the endogenous variables $\mathbf{y}_i$, given $(\mathbf{z}_i, \mathbf{u}_i)$. The reduced form parameters $\boldsymbol{\pi}$ are functions of $\boldsymbol{\theta}$. If the objective of modeling is inference about elements of $\boldsymbol{\theta}$, then (14) provides a direct route. Estimation of systems of equations like (14) is referred to as structural estimation in the classic Cowles Commission approach; see Heckman [34]. When the object of modeling is conditional prediction, the reduced form model is relevant.

Moment Condition Models

The classic causal model is a moment-condition model, derived from such a framework, consists of a set of r moment conditions of the form

$$E[\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}_0)] = \mathbf{0}\,, \tag{16}$$

where $\boldsymbol{\theta}$ is a $q \times 1$ vector, $\mathbf{g}(\cdot)$ is an $r \times 1$ vector function with $r \geq q$ and $\boldsymbol{\theta}_0$ denotes the value of $\boldsymbol{\theta}$ in the data generating process (d.g.p). The vector $\mathbf{w}$ includes all observables including, where relevant, a dependent (possibly vector-valued) variable $\mathbf{y}$, potentially endogenous regressors $\mathbf{x}$ and exogenous variables $\mathbf{z}$. The expectation is with respect to all stochastic components of $\mathbf{w}$ and hence $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{z}$.

Estimation methods for moment condition models include fully parametric approaches such as maximum likelihood as well as semi-parametric methods such as the

generalized method of moments (GMM) and instrumental variables (IV).

To make valid econometric inference on $\boldsymbol{\theta}$, it must be assumed or established that this parameter is identifiable; see Heckman [34] and Manski [62]. In other words, it is assumed that there is no set of observationally equivalent moment conditions. Identification may be established using (strong) parametric restrictions or using (weaker) semiparametric restrictions. The latter approach is currently favored in theoretical work. Point identification was emphasized in the classic Cowles Foundation but partial identification in many situations may be more attainable, especially if weaker restrictions on probability distributions of data are used; see Manski [63]. However, assuming point identification and given sufficient data, in principle these moment conditions lead to a unique estimate of the parameter $\boldsymbol{\theta}$. Potentially there are many reasons for loss of identifiability. Some of these are discussed in the next section where we also consider identification strategies.

The above approach has limitations. First, the definition of structure is not absolute because the distinction between endogenous $\mathbf{Y}$ and exogenous $\mathbf{Z}$ may be arbitrary. Second, the parameters $\boldsymbol{\theta}$ need not be tied to fundamental (or "deep") parameters; indeed it includes both the policy parameters that are of intrinsic interest and others that are not. If, however, the moment conditions are derived either from a model of optimization, or from some fundamental postulates of economic behavior such as the efficient market hypothesis, then at least some subset of parameters $\boldsymbol{\theta}$ can have a "structural" interpretation that is based on preference or technology parameters. Some econometricians prefer a narrower definition of a causal parameter which focuses only on the impact of the policy on the outcome of interest; the remaining parameters are treated as non-causal. Third, the approach is often difficult to implement in a way that provides information about either of the types of policy issues mentioned at the beginning of this section.

In response to these difficulties of the conventional approach two alternative approaches have emerged. The first is the potential outcome model (POM) that can be historically traced back to Neyman and Fisher. The second (and more modern) approach is based on dynamic stochastic Markov models. The first is easier to implement and hence currently dominates the applied literature. Next I will provide a brief overview of each approach.

Treatment Effect Models

This section deals with two closely related approaches in the treatment evaluation literature which targets an important structural parameter and its variants. Treatment effect models have been used extensively to study, to give just a few examples, the effect of: schooling on earnings, the class size on scholastic performance, unions on wages, and health insurance on health care use. Although in many cases the treatment variable is dichotomous, the framework can handle polychotomous treatment variables also. Treatment need not be discrete; the framework can handle ordered as well as continuously varying treatments.

Potential Outcome Models Much econometric estimation and inference are based on observational data. Identification of and inference on causal parameters is very challenging in such a modeling environment. Great simplification in estimating causal parameters arise if one can use data from properly designed and implemented controlled social experiments. Although such experiments have been implemented in the past they are generally expensive to organize and run. Econometricians therefore seek out data generated by quasi- or natural experiments which may be thought of as settings in which some causal variable changes exogenously and independently of other explanatory variables. This is an approximation to a controlled trial.

Random assignment implies that individuals exposed to treatment are chosen randomly, and hence the treatment assignment does not depend upon the outcome and is uncorrelated with the attributes of treated subjects. The great resulting simplification in relating outcomes to policy changes is unfortunately rarely achievable because random assignment of treatment is generally not feasible in economics. Most analyzes have to depend upon observational data.

As an example, suppose one wants to study the effect of unions on wages using data from unionized and nonunionized workers. Here being a unionized worker is the treatment. For the unionized worker, being a nonunion worker is the counterfactual. The purpose of the causal model is to estimate the mean difference in wages of unionized and nonunionized workers, the difference being attributed to being in the union.

A major obstacle for causality modeling stems from the so-called fundamental problem of causal inference [40]. Accordingly, in an observational setting one can only observe an individual in either the treated or the untreated state, and not both. Hence one cannot directly observe the effect of the treatment. Consequently, nothing more can be said about causal impact without some hypothesis about the counterfactual, i. e. what value of the outcome would have been observed in the absence of the change in policy variable.

The POM, also known as the Rubin causal model (RCM), provides a solution to the problem of establishing a counterfactual for policy evaluation. Causal parameters based on counterfactuals provide statistically meaningful and operational definitions of causality. In the POM framework the term "treatment" is used interchangeably with "cause". All policy changes and changes in the policy environment are broadly covered by the term treatment. Given a group impacted by policy, and another one that is not, a measure of causal impact is the average difference in the outcomes of the treated and the nontreated groups. Examples of treatment-outcome pairs are: health insurance and health care utilization; schooling and wages; class size and scholastic performance. Of course, the fact that with observational data a treatment is often chosen, not randomly assigned, is a significant complication.

In the POM framework, assuming that every element of the target population is potentially exposed to the treatment, the variables $(y_{1i}, y_{0i}, D_i, \mathbf{x}_i)$, $i = 1, \ldots, N$, forms the basis of treatment evaluation. The categorical variable D takes the values 1 and 0, respectively when treatment is or is not received; y_{1i} measures the response for individual i receiving treatment, and y_{0i} when not receiving treatment, $\mathbf{x}_i$ is the vector of exogenous covariates. That is, $y_i = y_{1i}$ if $D_i = 1$ and $y_i = y_{0i}$ if $D_i = 0$. Receiving and not receiving treatment are mutually exclusive states so only one of the two measures is available for any given i; the unavailable measure is the counterfactual. The effect of the cause D on outcome if individual i is measured by $(y_{1i} - y_{0i})$. The average causal effect of $D_i = 1$, relative to $D_i = 0$, is measured by the average treatment effect (ATE):

$$\text{ATE} = E[y|D = 1, \mathbf{x}] - E[y|D = 0, \mathbf{x}] \,, \tag{17}$$

where expectations are with respect to the probability distribution over the target population. Unlike the conventional structural model that emphasizes marginal effects the POM framework emphasizes ATE and parameters related to it.

POM can lead to causal statements if the counterfactual can be clearly stated and made operational. In observational data, however, a clear distinction between observed and counterfactual quantities may not be possible. Then ATE will estimate a weighted function of the marginal responses of specific subpopulations. Despite these difficulties, the identifiability of the ATE parameter may be an easier research target.

Matching Methods In the POM framework a causal parameter may be unidentified because there is no suitable comparison or control group that provides the benchmark for estimation. In observational studies, by definition there are no experimental controls. Therefore, there is no direct counterpart of the ATE calculated as a mean difference between the outcomes of the treated and nontreated groups. In other words, the counterfactual is not identified.

Matching methods provide a potential solution by creating a synthetic sample which includes a comparison group that mimics the control group. Such a sample is created by matching. Potential comparison units, that are not necessarily drawn from the same population as the treated units, are those for whom the observable characteristics, $\mathbf{x}$, match those of the treated units up to some selected degree of closeness. In the context of the unionization example, one would match, as closely as possible, unionized with nonunionized workers in terms of a vector of observable characteristics. Of course, if there are significant unobserved sources of differences that cannot be controlled, then this could lead to omitted variable bias. Given a treated sample plus well matched controls, under certain assumptions it becomes possible to identify parameters related to the ATE.

Matching may produce good estimates of the average effect of the treatment on the treated, i. e. the ATET parameter if (1) we can control for a rich set of $\mathbf{x}$ variables, (2) there are many potential controls. It also requires that treatment does not indirectly affect untreated observations. The initial step of establishing the nearest matches for each observation will also clarify whether comparable control observations are available.

Suppose the treated cases are matched in terms of all observable covariates. In a restricted sense all differences between the treated and untreated groups are controlled. Given the outcomes y_{1i} and y_{0i}, for the treatment and control, respectively, the average treatment effect is

$$\begin{aligned} &E\left[y_{1i}|D_i = 1\right] - E\left[y_{0i}|D_i = 0\right] \\ &= E[y_{1i} - y_{0i}|D_i = 1] + \left\{E\left[y_{0i}|D_i = 1\right]\right. \\ &\qquad \left. - E\left[y_{0i}|D_i = 0\right]\right\} \,. \end{aligned} \tag{18}$$

The first term in the second line is the ATET, and the second bracketed term is a "bias" term which will be zero if the assignment to the treatment and control is random. The sample estimate of ATET is a simple average of the differential due to treatment.

There is an extensive literature on matching estimators covering both parametric and nonparametric matching estimators; see Lee [58] for a survey. Like the POM framework, the approach is valid for evaluating policy that is already in operation and one that does not have general equilibrium effects. An important limitation is that the ap-

proach is vague and uninformative about the mechanism through which the treatment effects occur.

Dynamic Programming Models

Dynamic programming (DP) models represent a relatively new approach to microeconometric modeling. It emphasizes structural estimation and is often contrasted with "atheoretical" models that are loosely connected to the underlying economic theory. The distinctive characteristics of this approach include: a close integration with underlying theory; adherence to the assumption of rational optimizing agents; generous use of assumptions and restrictions necessary to support that close integration; a high level of parametrization of the model; concentration on causal parameters that play a key role in policy simulation and evaluation; and an approach to estimation of model parameters that is substantially different from the standard approaches used in estimating moment condition models. The special appeal of the approach comes from the potential of this class of models to address issues relating to new policies or old policies in a new environment. Further, the models are dynamic in the sense that they can incorporate expectational factors and inter-temporal dependence between decisions.

There are many studies that follow the dynamic programming approach. Representative examples are Rust [81]; Hotz and Miller [42]; Keane and Wolpin [50]. Some key features of DP models can be exposited using a model due to Rust and Phelan [85] which provides an empirical analysis of how the incentives and constraints of the US social security and Medicare insurance system affects the labor supply of old workers. Some of the key constraints arise due to incomplete markets, while individual behavior is based in part on expectations about future income streams. Explaining transitions from work to retirement is a challenging task not only because it involves forward-looking behavior in a complex institutional environment but also because a model of retirement behavior must also capture considerable heterogeneity in individual labor supply, discontinuities in transitions from full time work to not working, and presence of part-time workers in the population, and coordination between labor supply decisions and retirement benefits decisions.

The main components of the DP model are as follows. State variable is denoted by s_t, control variable by d_t. β is the intertemporal discount factor. In implementation all continuous state variables are discretized – a step which greatly expands the dimension of the problem. Hence all continuous choices become discrete choices, d_t is a discrete choice sequence, and the choice set is finite. For example, in Rust and Phelan [85] total family income is discretized into 25 intervals, social security state into 3 states, and employment state (hours worked annually) into 3 discrete intervals, and so forth. There is a single period utility function $u_t(s, d, \theta_u)$ and $p_t(s_{t+1}|s_t, d_t, \theta_p, \alpha)$ denotes the probability density of transitions from s_t to s_{t+1}. The optimal decision sequence is denoted by $\delta = (\delta_0, \ldots, \delta_T)$ where $d_t = \delta_t(s_t)$ and is the optimal solution that maximizes the expected discounted utility:

$$V_t(s) = \max_{\delta} E_{\delta} \left\{ \sum_{j=t}^{T} \beta^{j-t} u_j(s_j, d_j, \theta_u) | s_t = s \right\} . \quad (19)$$

The model takes Social Security and Medicare policy parameters, α, as known. The structural parameters $\theta = (\beta, \theta_u, \theta_p)$ are to be estimated. To specify the stochastic structure of the model the state variables are partitioned as $s = (x, \eta)$, where x is observable and η is unobservable (for the econometrician); $\eta_t(d)$ can be thought of as the net utility or disutility impact due to factors unobserved by the econometrician at time t.

An important assumption, due to Rust [81], which restricts the role of η permits the following decomposition of the joint probability distribution of (x_{t+1}, η_{t+1}):

$$\Pr\left[x_{t+1}, \eta_{t+1} | x_t, \eta_t, d_t\right] = \Pr\left[\eta_{t+1} | x_{t+1}\right] \Pr\left[x_{t+1} | x_t, d_t\right] .$$

Note that the first term on the right-hand side implies serial independence of unobservables; the second term has a Markov structure and implies that η_t affects x_t only through d_t.

$$v_t(x_t, d_t, \theta, \alpha) = u_t(x_t, d_t, \theta_u) + \beta \int \log \left[\sum_{d_{t+1} \in D(x_{t+1})} \exp\{v_{t+1}(x_{t+1}, d_{t+1}, \theta, \alpha)\} \right] p_t(x_{t+1} | x_t, d_t, \theta_p, \alpha) , \quad (20)$$

Estimation of the model, based on panel data $\{x_t^i, d_t^i\}$, uses the likelihood function

$$L(\theta) = L(\beta, \theta_u, \theta_p) = \prod_{i=1}^{I} \prod_{t=1}^{T_i} P_t(d_t^i | x_t^i, \theta_u) p_t(x_t^i | x_{t-1}^i, d_{t-1}^i, \theta_p) . \quad (21)$$

This is a high dimensional model because a large number of state variables and associated parameters are needed to specify the future expectations. (This complexity is

highlighted to emphasize that DP models run into dimensionality problems very fast.) First, strong assumptions are needed to address the unobservable and subjective aspects of decision-making because there are a huge number of possible future contingencies to take into account. Second, restrictions are needed to estimate the belief arrays. Consistent with tenets of rational agents the model assumes rational expectations. To impose exclusion restrictions p_t is decomposed into a product of marginal and conditional densities.

As a simplification a two-stage estimation procedure is used: (1) estimate θ_p using first stage partial likelihood function involving only products of the p_t terms; (2) estimate θ_p by solving the DP problem numerically, and estimate (β, θ_u) using a second stage partial likelihood function consisting of only products of P_t. The two-stage estimation procedure is not as efficient as the full maximum likelihood estimation since the error in $\hat{\theta}_p$ contaminates the estimated covariance matrix for θ_u.

Space limitations do not permit us to provide the details of the computational procedure, for which we refer the reader to Rust [81]. In outline, at the first step the procedure estimates the transition probability parameters θ_p using the partial likelihood function and at the second stage a Nested Fixed Point (NFP) algorithm is used to estimate the remaining parameters.

New Directions in Structural Modeling

The motivation for many of the recent developments lies in the difficulties and challenges of identifying causal parameters under fewer distributional and functional form restrictions. Indeed an easily discernible trend in modern research is steady movement away from strong parametric models and towards semi-parametric models. Increasingly semiparametric identification is the stated goal of theoretic research [41]. Semiparametric identification means that unique estimates of the relevant parameters can be obtained without making assumptions about distribution of data, and some times it also means that assumptions about functional forms can also be avoided. Potentially there are numerous ways in which the identification of key model parameters can be compromised. The solution strategy in such cases is often model specific. This section provides a selective overview of recent developments in microeconometrics that address such issues.

Endogeneity and Multivariate Modeling

Structural nonlinear models involving LDVs arise commonly in microeconometrics. A leading example of a causal model involves modeling the conditional distribution (or moments) of a continuous outcome (y) which depends on variables ($\mathbf{x}$, D) where D is an endogenous binary treatment variable. For example, y is medical expenditure and D is a binary indicator of health insurance status. The causal parameter of interest is the marginal effect of D on y. More generally y could be binary, count, an ordered discrete variable, or a truncated/censored continuous variable. More generally the issue is that multivariate modeling. Currently there is no consensus on econometric methodology for handling this class of problems. Some of the currently available approaches are now summarized.

Control Functions A fully parametric ("full information") estimation strategy requires the specification of the joint distribution of (y, D), which is often difficult because such a joint distribution is rarely available. Another ("limited information") strategy is to estimate only the conditional model, quite often only the conditional mean $\mathrm{E}[y|\mathbf{x}, D]$, controlling for endogeneity of treatment. If the model is additively separable in $\mathrm{E}[y|\mathbf{x}, D]$ and the stochastic error ε which is correlated with d so that $\mathrm{E}(\varepsilon D) \neq 0$, then a two-step procedure may be used. This involves replacing D by its projection on a set of exogenous instrumental variables $\mathbf{z}$ (usually including $\mathbf{x}$), denoted $\widehat{D}(\mathbf{z})$, and estimating the conditional expectation $\mathrm{E}[y|\mathbf{x}, \widehat{D}(\mathbf{z})]$. Unfortunately, this approach does not always yield a consistent estimate of the causal parameter; for example, if the conditional mean is nonlinear in $(\mathbf{x}, D)$. Therefore this approach is somewhat ad hoc.

Another similar strategy, called the control function approach, involves replacing $\mathrm{E}[y|\mathbf{x}, D]$ by $\mathrm{E}[y|\mathbf{x}, \mathbf{w}, D]$. Here $\mathbf{w}$ is a set of additional variables in the conditional mean function such that the assumption $\mathrm{E}(\varepsilon D|\mathbf{w}) = 0$; that is D can be treated as exogenous, given the presence of $\mathbf{w}$ in the conditional mean function. Again such an approach does not in general identify the causal parameter of interest. Additional restrictions are often necessary for structural identification. In a number of cases where the approach has been shown to work some functional form and structural restrictions are invoked, such as additive separability and a triangular error structure.

Consider the following example of an additively separable model with a triangular structure. Let y_1 be the dependent variable in the outcome equation, which is written as

$$y_1 = E\left[y_1|D, \mathbf{x}\right] + u_1 + \lambda u_2 \,,$$

where $(u_1 + \lambda u_2)$ is the composite error. Let D denote the treatment indicator for which the model is

$$D = E\left[D|\mathbf{z}\right] + u_2 \,.$$

A simple assumption on the distribution of the error terms takes them to be zero-mean and mutually uncorrelated. In this case the control function approach can be used. Specifically a consistent estimate of u_2, say $\widehat{u}_2$, can be included as an additional regressor to the y_1 equation. This type of argument has been used for handling endogeneity in regression models that are specified for, instead of the conditional mean, the conditional median or conditional quantile regression; see Chesher [14], Ma and Koenker [60]. The control function approach has been adapted for treating endogeneity problems in semiparametric and nonparametric framework [9].

Latent Factor Models Another "full information" approach that simultaneously handles discrete variation and endogeneity also imposes a restriction on the structure of dependence using latent factors and resorts to simulation-assisted estimation. An example is Deb and Trivedi [18] who develop a joint model of counts, with a binary insurance plan variable (D) as a regressor, and a model for the choice of insurance plan. Endogeneity in their model arises from the presence of correlated unobserved heterogeneity in the outcome (count) equation and the binary choice equation. Their model has the following structure:

$$\Pr\left[Y_i = y_i|\mathbf{x}_i, D_i, l_{ji}\right] = f\left(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_1 D_i + \lambda l_i\right) .$$
$$\Pr\left[D_i = 1|\mathbf{z}_i, l_{ji}\right] = g\left(\mathbf{z}_i'\boldsymbol{\alpha} + \delta l_i\right) .$$

Here l_i is latent factor reflecting unobserved heterogeneity and δ is an associated factor loading. The joint distribution of selection and outcome variables, conditional on the common latent factor, can be written as

$$\Pr\left[Y_i = y_i, D_i = 1|\mathbf{x}_i, \mathbf{z}_i, l_i\right] = f\left(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_1 D_i + \lambda l_i\right) \times g\left(\mathbf{z}_i'\boldsymbol{\alpha}_j + \delta l_i\right) , \quad (22)$$

because (y, D) are conditionally independent.

The problem in estimation arises because the l_i are unknown. Although the l_i are unknown, assume that the distribution of l_i, h, is known and can therefore be integrated out of the joint density, i. e.,

$$\Pr\left[Y_i = y_i, D_i = 1|\mathbf{x}_i, \mathbf{z}_i\right] = \int \left[f\left(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_1 D_i + \lambda l_i\right) \times g\left(\mathbf{z}_i'\boldsymbol{\alpha}_j + \delta l_i\right)\right] h(l_i)\,\mathrm{d}l_i .$$

Cast in this form, the unknown parameters of the model may be estimated by maximum likelihood.

The maximum likelihood estimator maximizes the joint likelihood function $L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|y_i, D_i, \mathbf{x}_i, \mathbf{z}_i)$, where $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}, \gamma_1, \lambda)$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\alpha}, \delta)$, refer to parameters in the outcome and plan choice equations respectively, and L refers to the joint likelihood.

The main problem of estimation, given suitable specifications for f, g and h, is the fact that the integral does not have, in general, a closed form solution. The maximum simulated likelihood (MSL) estimator involves replacing the expectation by a simulated sample average, i. e.,

$$\widetilde{\Pr}[Y_i = y_i, D_i = 1|\mathbf{x}_i, \mathbf{z}_i] = \frac{1}{S}\sum_{s=1}^{S}\left[f\left(\mathbf{x}_i'\boldsymbol{\beta} + \gamma_1 D_i + \sum\nolimits_j \lambda \widetilde{l}_{is}\right) \times g\left(z_i'\boldsymbol{\alpha} + \delta \widetilde{l}_{is}\right)\right] , \quad (23)$$

where $\widetilde{l}_{is}$ is the sth draw (from a total of S draws) of a pseudo-random number from the density h and $\widetilde{\Pr}$ denotes the simulated probability.

The above approach, developed for an endogenous dummy regressor in a count regression model, can be extended to multiple dummies (e. g. several types of health insurance), and multiple outcomes, discrete or continuous (e. g. several measures of health care utilization such as number of doctor visits, prescribed medications). The limitation comes from the heavy burden of estimation compared with an IV type estimator. Further, as in any simultaneous equation model, identifiability is an issue. Applied work typically includes some nontrivial explanatory variables in the **z** vector that are excluded from the **x** vector. As an example, consider insurance premium which would be a good predictor of insurance status but will not directly affect health care use.

Instrumental Variables and Natural Experiments

If identification is jeopardized because the treatment variable is endogenous, then a standard solution is to use valid instrumental variables. To identify the treatment effect parameter we need exogenous sources of variation in the treatment. Usually this means that the model must include at least the minimum number of exogenous variables (instruments) that affect the outcome only through the treatment – an assumption usually called an exclusion or identification restriction. This requirement may be difficult to satisfy. Keane [49] gives an example where there are no possible instruments. Even this extreme possibility is discounted, agreement on valid instruments is often difficult, and when such agreement can be established the instruments may be "weak" in the sense that they do not account for substantial variation in the endogenous variables they are assumed to affect directly. The choice of the instru-

mental variable as well as the interpretation of the results obtained must be done carefully because the results may be sensitive to the choice of instruments. In practice, such instrumental variables are either hard to find, or they may generate only a limited degree of variation in the treatment by impacting only a part of the relevant population.

A natural experiment may provide a valid instrument. The idea here is simply that a policy variable may exogenously change for some subpopulation while remaining unchanged for other subpopulations. For example, minimum wage laws in one state may change while they remain unchanged in a neighboring state. Such events create natural treatment and control groups. Data on twins often provide data with both natural treatment and control, as has been argued in many studies that estimate the returns to schooling; see Angrist and Krueger [5]. If the natural experiment approximates randomized treatment assignment, then exploiting such data to estimate structural parameters can be simpler than estimation of a larger simultaneous equations model with endogenous treatment variables. However, relying on data from natural experiments is often not advisable because of such events are rare and because the results from them may not generalize to a broader population.

Limitations of the IV Approach Some limitations of the IV approach, e. g. the weak IV problem, are general but certain others are of special relevance to microeconometrics. One of these is a consequence of heterogeneity in the impact of the policy on the outcome. Consideration of this complication has led to significant refinements in the interpretation of results obtained using the IV method.

In many applications of the POM framework, the underlying assumption is that there exists a comparison group and a treatment that is homogeneous in its response to the treatment. In the heterogeneous case, the change in the participation in treatment generated by the variation in the instrument may depend both upon which instrument varies, and on the economic mechanism that links the participation to the instrument. As emphasized by Heckman and Vytlacil [38], Keane [49] and others, a mechanical application of the IV approach has a certain black box character because it fails to articulate the details of the mechanism of impact. Use of different instruments identify different policy impact parameters because they may impact differently on different members of the population. Heckman and Vytlacil [38] emphasize that the presence of unobserved heterogeneity and selection into treatment may be based on unobserved gains, a condition they call *essential heterogeneity*. The implication for the choice of IVs is that these may be independent of the idiosyncratic gains in the overall population, but conditional on those who self-select into treatment, they may no longer be independent of the idiosyncratic gains in this subgroup. Further, as a consequence of the dependence between treatment choice and IV estimates different IVs identify different parameters. In this context, an a priori specification of the choice model for treatment is necessary for the interpretation of IV estimators.

The concept of local instrumental variables is related to the local average treatment (LATE) parameter introduced by Imbens and Angrist [45]. To illustrate this we consider the following canonical linear model.

The outcome equation is a linear function of observable variables $\mathbf{x}$ and a participation indicator D:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \alpha D_i + u_i \,, \tag{24}$$

and the participation decision depends upon a single variable z, referred to as an instrument,

$$D_i^* = \gamma_0 + \gamma_1 z_i + v_i \,, \tag{25}$$

where D_i^* is a latent variable with its observable counterpart generated by

$$D_i = \begin{cases} 0 \text{ if } D_i^* \leq 0 \\ 1 \text{ if } D_i^* > 0 \end{cases} \,. \tag{26}$$

There are two assumptions: (1) There is an exclusion restriction as the variable z that appears in the equation for D that does not appear in the equation for $\mathbf{x}$. (2) Conditional on $(\mathbf{x}, z)$ $\text{Cov}\,[z, v] = \text{Cov}\,[u, z] = \text{Cov}\,[\mathbf{x}, u] = 0$, but $\text{Cov}\,[D, z] \neq 0$. It is straightforward to show that the IV estimator of the treatment effect parameter α is

$$\alpha_{IV} = \frac{E\left[y|z'\right] - E\left[y|z\right]}{\Pr\left[D\left(z'\right) = 1\right] - \Pr\left[D\left(z\right) = 1\right]} \,, \tag{27}$$

which is well-defined if $\Pr\left[D\left(z'\right) = 1\right] - \Pr\left[D\left(z\right) = 1\right] \neq 0$. The sample analog of α_{IV} is the ratio of the mean difference between the treated and the nontreated divided by the change in the proportion treated due to the change in z.

Why does this measure a "local" effect? This is because the treatment effect applies to the "compliers" only, that is those who are induced to participate in the treatment as a result of the change in z; see Angrist et al. [6]. Thus LATE depends upon the particular values of z used to evaluate the treatment and on the particular instrument chosen. Those who are impacted may not be representative of those treated, let alone the whole population. Consequently the LATE parameter may not be informative about the consequences of large policy changes brought

about by changes in instruments different from those historically observed.

If more than one instrument appears in the participation equation, as when there exist overidentifying restrictions, the LATE parameter estimated for each instrument will in general differ. However, a weighted average may be constructed.

Omitted Variables, Fixed and Random Effects

Identification may be threatened by the presence of a large number of nuisance or incidental parameter. For example, in a cross section or panel data regression model the conditional mean function may involve an individual specific effect α_i, i. e. $\mathrm{E}[y_i|\mathbf{x}_i, \alpha_i]$ or $\mathrm{E}[y_{it}|\mathbf{x}_{it}, \alpha_i]$ where $i = 1, \ldots, N$, $t = 1, \ldots, T$. The parameters α_i may be interpreted as omitted unobserved factors. Two standard statistical models for handling them are fixed and random effect formulations. In a fixed-effect (FE) model the α_i are assumed to be correlated with the observed covariates $\mathbf{x}_i$, i. e. $\mathrm{E}[|\mathbf{x}_{it}|\alpha_i] \neq \mathbf{0}$, whereas in the random effects (RE) model $\mathrm{E}[|\mathbf{x}_{it}|\alpha_i] = \mathbf{0}$ is assumed. Because the FE model is less restrictive, it has considerable appeal in microeconometrics.

FE Models In FE models this effect cannot be identified without multiple observations on each individual, i. e.. panel data. Identification is tenuous even with panel data if the panel is short, i. e. T is small; see Lancaster [54] about the incidental parameters problem. The presence of these incidental parameters in the model also hinders the identification of other parameters of direct interest. A feasible solution in the case where both N and T are large, is to introduce dummy variables for each individual and estimate all the parameters. The resulting computational problem has a large dimension but has been found to be feasible not only in the standard case of linear regression but also for some leading nonlinear regressions such as the Probit, Tobit and Poisson regressions [26,27].

If the panel is short, the $\alpha_i (i = 1, \ldots, N)$ cannot be identified and no consistent estimator is available. Then the identification strategy focuses on the remaining parameters that are estimated after eliminating α_i by a transformation of the model. Consider, as an example, the linear model with both time-varying and time-invariant exogenous regressors $(\mathbf{x}'_{it}, \mathbf{z}'_i)$

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_i\gamma + \varepsilon_{it} , \tag{28}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are common parameters, while $\alpha_1, \ldots, \alpha_N$ are incidental parameters if the panel is short as then each α_i depends on fixed T observations and there are infinitely many α_i since $N \to \infty$. Averaging over T observations yields

$$\bar{y}_i = \alpha_i + \bar{\mathbf{x}}'_i\boldsymbol{\beta} + \mathbf{z}'_i\gamma + \bar{\varepsilon}_i . \tag{29}$$

On subtracting we get the "within model"

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\boldsymbol{\beta} + (\varepsilon_{it} - \bar{\varepsilon}_i) , \quad i = 1, \ldots, N , \quad t = 1, \ldots, T , \tag{30}$$

where the α_i term and the time-invariant variables $\mathbf{z}_i$ disappear. A first difference transformation $y_{it} - y_{i,t-1}$ can also eliminate the α_i. The remaining parameters can be consistently estimated, though the disappearance of variables from the model means that prediction is no longer feasible.

Unfortunately this elimination "trick" does not generalize straight-forwardly to other models, especially nonlinear nonnormal models with fully specified distribution. There is no unified solution to the incidental parameters problem, only model-specific approaches. In some special cases the conditional likelihood approach does solve the incidental parameter problem, e. g. linear models under normality, logit models (though not probit models) for binary data, and some parametrizations of the Poisson and negative binomial models for count data. The RE model, by contrast, can be applied in more widely.

RE Panel Models If the unobservable individual effects α_i, $\alpha_i > 0$, are random variables that are distributed independently of the regressors, the model is called the random effects (RE) model. Usually the additional assumptions that both the random effects and the error term are also employed, i. e., $\alpha_i \sim [\alpha, \sigma^2_\alpha]$, and $\varepsilon_{it} \sim [0, \sigma^2_\varepsilon]$ are also employed. More accurately this is simply the *random intercept model*. As a specific example consider the Poisson individual-specific effects model which specifies

$$y_{it} \sim \mathrm{Poisson}[\alpha_i \exp(\mathbf{x}'_{it}\boldsymbol{\beta})] .$$

If we assume gamma distributed random effects distributed with mean 1, variance $1/\gamma = \eta$ and density $g(\alpha_i|\eta) = \eta^\eta \alpha_i^{\eta-1} \mathrm{e}^{-\alpha_i \eta}/\Gamma(\eta)$, there is a tractable analytical solution for the unconditional joint density for the ith observation $\int \left[\prod_{t=1}^T f(y_{it}|\mathbf{x}_{it}, \alpha_i, \boldsymbol{\beta}, \boldsymbol{\gamma})\right] g(\alpha_i|\eta)\, \mathrm{d}\alpha_i$ (see Cameron and Trivedi ([13]: chapter 23.7 for algebraic details). However, under other assumptions about the distribution (e. g. log-normal) a closed form unconditional density usually does not arise, and estimation is then based on numerical integration – an outcome that is fairly common for nonlinear random effect models.

Modeling Heterogeneity

To accommodate the diversity and complexity of responses to economic factors, it is often necessary to allow for variation in the model parameters. There are many specification strategies to accomplish such a goal. One of the most popular and well-established strategy is to model heterogeneity using some type of mixture model. Typically the specification of a mixture model involves two steps. In the first step a conditional distribution function $F(y|\mathbf{x}, \boldsymbol{\nu})$ is specified where $\mathbf{x}$ is an observed vector of covariates and ν is an unobserved heterogeneity term, referred to as frailty in biostatistics. In the second step a distribution $G(\nu)$ is specified for ν and a mixture model is derived. The distribution of ν may be continuous or discrete. Poisson-gamma mixture for count data and Weibull–gamma mixtures for survival data are two leading examples based on continuous heterogeneity assumption. The mixed multinomial logit model (MMNL) is another example [75].

The mixture class of models is very broad and includes two popular subclasses, the random coefficient approach and the latent class approach. While relatively simple in formulation, such mixture approaches often generate major identification and computational challenges [24]. Here I provide two examples that illustrate the issues associated with their use.

Latent Class Models Consider the following two-component *finite mixture* model. If the sample is a probabilistic mixture from two subpopulations with p.d.f. $f_1(y|\mu_1(\mathbf{x}))$ and $f_2(y|\mu_2(\mathbf{x}))$, then $\pi f_1(.) + (1-\pi) f_2(.)$, where $0 \leq \pi \leq 1$, defines a two-component finite mixture. That is, observations are draws from f_1 and f_2, with probabilities π and $1-\pi$ respectively. The parameters to be estimated are (π, μ_1, μ_2). The parameter π may be further parameterized.

At the simplest level we think of each subpopulation as a "type", but in many situations a more informative interpretation may be possible. There may be an a priori case for such an interpretation if there is some characteristic that partitions the sampled population in this way. An alternative interpretation is simply that the linear combination of densities is a good approximation to the observed distribution of y. Generalization to additive mixtures with three or more components is in principle straight-forward but subject to potential problems of the identifiability of the components.

Formally the marginal (mixture) distribution is

$$h\left(t_i|\mathbf{x}_i, \pi_j, \boldsymbol{\beta}\right) = \sum_{j=1}^{m} f\left(t_i|\mathbf{x}_i, \nu_j, \boldsymbol{\beta}\right) \pi_j\left(\nu_j\right) , \qquad (31)$$

where ν_j is an estimated support point and π_j is the associated probability. This representation of unobserved heterogeneity is thought of as semiparametric because it uses a discrete mass point distribution. The specification has been found to be very versatile. It has been used to model duration data where the variable of interest is the length of time spent in some state, e. g. unemployment, and individuals are thought to differ both interms of their observable and unobservable characteristics; see Heckman and Singer [37].

The estimation of the finite mixture model may be carried out either under the assumption of known or unknown number of components. More usually the proportions $\pi_j, j = 1, \ldots, m$ are unknown and the estimation involves both the π_j and the component parameters. The maximum likelihood estimator for the latter case is called Nonparametric Maximum Likelihood Estimator (NPMLE), where the nonparametric component is the number of classes. Estimation is challenging, especially if m is large because the likelihood function is generally multi-modal and gradient-based methods have to be used with care. If the number of components is unknown, as is usually the case, then some delicate issues of inference arise. In practice, one may consider model comparison criteria to select the "best" value of m. Baker and Melino [8] provide valuable practical advice for choosing this parameter using an information criterion.

LC models are very useful for generating flexible functional forms and for approximating the properties of nonparametric models. For this reason it has been used widely. Deb and Trivedi [16,17] use the approach for modeling mixtures of Poisson and negative binomial regressions. McFadden and Train [75] show that latent class multinomial logit model provides an arbitrarily good approximation to any multinomial choice model based on the RUM. This means that it provides one way of handling the IIA problem confronting the users of the MNL model. Dynamic discrete choice models also use the approach.

LC models generate a computational challenge arising from having to choose m and to estimate corresponding model parameters for a given m, and there is the model selection problem. Often there is no prior theory to guide this choice which in the end may be made largely on grounds of model goodness-of-fit. Akaike's or Bayes penalized likelihood (or information) criterion (AIC or BIC) is used in preference to the likelihood ratio test which is not appropriate because of the parameter boundary hypothesis problem. The dimension of parameters to be estimated is linear in m, the number of parameters can be quite large in many microeconometric applications that usually control for many socio-demographic factors. This

number can be decreased somewhat if some elements are restricted to be equal, for example by allowing the intercept but not the slope parameters to vary across the latent classes; see, for example, Heckman and Singer [37].

When the model is overparametrized, perhaps because the intergroup differences are small, the parameters cannot be identified. The problem is reflected in slow convergence in computation due to the presence of multiple optima, or a flat likelihood surface. The computational algorithm may converge to different points depending on the starting values.

Interpretation of the LC model can be insightful because it has the potential to capture diverse responses to different stimuli. However, a potential limitation is due to the possibility that additional components may simply reflect the presence of outliers. Though this is not necessarily a bad thing, it is useful to be able to identify the outlying observations which are responsible for one or more components.

Random Coefficient Models Random coefficient (RC) models provide another approach to modeling heterogeneity. The approach has gained increasing popularity especially in the applications of discrete choice modeling to marketing data. In this section I provide an exposition of the random coefficient logit model based on Train [89] where a more comprehensive treatment is available. The random coefficient models extend the RUM model of Sect. "Introduction" which restricts the coefficients of parameters to be constant across individuals. If individuals have different utility functions then that is a misspecification. The RC framework is one of a number of possible ways of relaxing that restriction.

The starting point is the RUM framework presented in Sect. "Introduction". Assume individual $i(i = 1, 2, \ldots, N)$ maximizes utility U_{ij} by choosing alternative j from her choice set $M_n = (0, 1)$. The utility U_{nj} has observed (systematic) part $V(\mathbf{X}_{ij}; \boldsymbol{\beta}_i)$ and random part ε_{ij};

$$U_{ij} = V(\mathbf{X}_{ij}; \boldsymbol{\beta}_i) + \varepsilon_{ij}\,, \qquad j = 0, 1\,; \quad i = 1, 2, \ldots, N\,. \tag{32}$$

Vector $\mathbf{X}_{nj}$ in $V(.,.)$ represents observed attributes of alternatives, characteristics of the individual i as well as alternative-specific constants. $\boldsymbol{\beta}_i$ is the vector of coefficients associated with $\mathbf{X}_{ij}$. Error term ε_{ij} captures unobserved individual characteristics/unobserved attributes of the alternative j and follows some distribution $\mathrm{D}(\boldsymbol{\theta}_\varepsilon)$, where $\boldsymbol{\theta}_\varepsilon$ is the unknown parameter vector to be estimated. Of course, U_{ij} is latent, so we use an indicator function, y_{ij}, such that $y_{ij} = 1$ if $U_{ij} \geq U_{ik} \forall k \neq j$ and $y_{ij} = 0$ otherwise. Probability that individual i chooses alternative j is

$$P_{ij} = P(j|\mathbf{X}_i; \boldsymbol{\beta}_i, \boldsymbol{\theta}_\varepsilon) = P(y_{ij} = 1) = P(U_{ij} \geq U_{ii} \forall i \neq j)\,,$$

and the probability that alternative j is chosen is $P(y_{ij}|\mathbf{X}_i; \boldsymbol{\beta}_i, \boldsymbol{\theta}_\varepsilon) = P_{ij}^{y_{ij}}$, which, under the independence assumption, leads to the likelihood

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}_\varepsilon) = \prod_{i=1}^{N} \prod_{j \in M_i} P_{ij}^{y_{ij}}\,. \tag{33}$$

Different assumptions on the error structure lead to different discrete choice models. The kth component of the vector $\boldsymbol{\beta}_i$, which represents the coefficient of some attribute k, can be decomposed as $\beta_{ik} = b + \boldsymbol{\delta}'\boldsymbol{\omega}_i + \sigma_k \eta_{ik}$, if the coefficient is random and simply $\beta_{nk} = b$, if the coefficient is non-random. Here b represents the average taste in the population for provider attribute k, $\boldsymbol{\omega}_i$ is a vector of choice-invariant characteristics that generates individual heterogeneity in the means of random coefficients $\boldsymbol{\beta}_i$, and $\boldsymbol{\delta}$ is the relevant parameter vector. Finally, η_{ik} is the source of random taste variation, which is be assumed to have a known distribution, e. g. normal.

If random parameters are not correlated then $\boldsymbol{\Gamma} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_K)$ is a diagonal matrix. To allow for correlated parameters, $\boldsymbol{\Gamma}$ is specified as a lower triangular matrix so that the variance-covariance matrix of the random coefficients becomes $\boldsymbol{\Gamma}\boldsymbol{\Gamma}' = \boldsymbol{\Sigma}$. Non-random parameters in the model can be easily incorporated in this formulation by specifying the corresponding rows in $\mathbf{D}$ and $\boldsymbol{\Gamma}$ to be zero. The conditional choice probability that individual i chooses alternative j, conditional on the realization of $\boldsymbol{\eta}_i$, is

$$P\left(j|\boldsymbol{\eta}_i, \boldsymbol{\theta}\right) = \frac{\exp\left(\theta_j + \boldsymbol{\beta}_i'\tilde{\mathbf{X}}_{ij}\right)}{1 + \exp\left(\theta_i + \boldsymbol{\beta}_i'\tilde{\mathbf{X}}_{ij}\right)}\,, \tag{34}$$

where $\boldsymbol{\theta} = (\mathbf{b}, \mathbf{D}, \boldsymbol{\Gamma})$ and $\boldsymbol{\eta}_i$ has distribution G with mean vector $\mathbf{0}$ and variance-covariance matrix $\mathbf{I}$.

Unconditional choice probability P_{ij} for alternative j is given by

$$P_{ij} = \int_{\boldsymbol{\eta}_i} P\left(j|\boldsymbol{\eta}_i, \boldsymbol{\theta}\right) \mathrm{d}F_{\boldsymbol{\eta}}(\boldsymbol{\eta}_i)\,, \tag{35}$$

where $F_{\boldsymbol{\eta}}(.)$ is the joint c.d.f. of $\boldsymbol{\eta}_i$. The choice probability can be interpreted as a weighted average of logit probabilities with weights given by the mixing c.d.f. $F_{\boldsymbol{\eta}}(.)$. Following (10), the log-likelihood for θ is given by:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \sum_{j=0}^{1} y_{ij} \log P_{ij}\,. \tag{36}$$

The unconditional choice probability P_{nj} involves an integral over the mixing distribution, but the log-likelihood function does not generally have a closed form. Hence one cannot differentiate the log-likelihood function with respect to the parameter vector $\boldsymbol{\theta} = (\mathbf{b}, \mathbf{D}, \boldsymbol{\Gamma})$ and solve the first order conditions in order to obtain the parameter estimates. Instead, one estimates the choice probability P_{ij} through simulation and then maximize the resulting simulated maximum likelihood (SML) with respect to the parameter vector.

Train [89] shows that this mixed logit framework leads to a tractable, unbiased and smooth simulator for the choice probability defined by:

$$\hat{P}_{ij} = \hat{P}(j|\mathbf{X}_i, \boldsymbol{\theta}) = \frac{1}{S}\sum_{s=1}^{S} P(j|\mathbf{X}_i, \boldsymbol{\beta}_i^s; \boldsymbol{\theta}) , \tag{37}$$

where $\boldsymbol{\beta}_i^s = \mathbf{b} + \mathbf{D}\boldsymbol{\omega}_i + \boldsymbol{\Gamma}\boldsymbol{\eta}_i^s$ and $\boldsymbol{\eta}_i^s$ is the sth ($s = 1, 2, \ldots, S$) draw from the joint distribution of $\boldsymbol{\eta}_i^s$, i.e., from $f(\boldsymbol{\eta}_i)$.

The log-likelihood function can be approximated by maximum simulated log-likelihood (MSL) given by

$$\begin{aligned} S\mathcal{L}(\boldsymbol{\theta}_\beta) &= \sum_{i=1}^{N}\sum_{j=0}^{1} y_{ij} \log \hat{P}_{ij} \\ &= \sum_{i=1}^{N}\sum_{j=0}^{1} y_{ij} \log\left[\frac{1}{S}\sum_{s=1}^{S} P(j|\mathbf{X}_i, \boldsymbol{\beta}_i^s; \boldsymbol{\theta})\right] . \end{aligned} \tag{38}$$

Note that although $\hat{P}_j$ is unbiased for P_j, $\ln(\hat{P}_j)$ is not unbiased for $\ln(P_j)$, therefore the simulator generates some bias. To avoid bias, the simulation approximation should be improved. That means one must choose S to be sufficiently large. How large is "sufficiently large"? There is no fixed answer. But a result due to Gourieroux and Monfort [28] states indicates that the number should increase with the sample size N. Specifically, if the number of simulations, S, increases faster than the square root of the number of observations, this bias disappears in large samples. More pragmatically, the user should check that the results do not change much if S is increased.

To simulate the choice probability P_{ij}, one generally requires a large number of pseudo-random draws from the mixing distribution so that resulting simulation errors in the parameter estimates are kept at a reasonable level. Fortunately, advances in simulation methodology, such as the use of quasi-random numbers, in place of pseudo-random numbers, makes this feasible; see Train [89].

The preceding examples illustrate the point that accommodating heterogeneity in a flexible manner comes at a considerable computational cost. In many cases they lead to simulation-assisted estimation, this being an area of microeconometrics that has developed mainly since the 1990s.

Microeconometrics, Table 1
Alternative sample stratification schemes

Stratification Scheme	Description
Simple random	One strata covers entire sample space.
Pure exogenous	Stratify on regressors only, not on dependent variable.
Pure endogenous	Stratify on dependent variable only, not on regressors.
Augmented sample	Random sample augmented by extra observations from part of the sample space.
Partitioned	Sample space split into mutually exclusive strata that fill the entire sample space.

Nonrepresentative Samples

Microeconometric methods often invoke the assumption that analysis is based on simple random samples (SRS). This assumption is hardly ever literally true for survey data. More commonly a household survey may first stratifies the population geographically into subgroups and applies differing sampling rates for different subgroups. An important strand in microeconometrics addresses issues of estimation and inference when the i.i.d. assumption no longer applies because the data are obtained from stratified and/or weighted samples. Stratified sampling methods also lead to dependence or clustering of cross section and panel observations. Clusters may have spatial, geographical, or economic dimension. In these cases the usual methods of establishing distribution of estimators based on the SRS assumption need to be adapted.

Stratified Samples For specificity it is helpful to mention some common survey stratification schemes. Table 1 based on Imbens and Lancaster [46] and Cameron and Trivedi [13], provides a summary.

Econometricians have paid special attention to endogenous stratification because this often leads to inconsistency of some standard estimation procedures such as ML; see Manski and Lerman [64], Cosslett [15], Manski and McFadden [65]. One example is choice-based sampling for binary or multinomial data where samples are chosen based on the discrete outcome y. For example, if choice is between travel to work by bus or travel by car we may over-sample bus riders if relatively few people commute by bus. A related example is count data on number of visits collected by on-site sampling of users, such as sampling at recreational sites or shopping centers or doctors offices. Then data are truncated, since those with $y = 0$ are

not sampled, and additionally high frequency visitors are over-sampled.

Endogenously stratified sampling leads to the density in the sample differing from that in the population (Cameron and Trivedi, pp. 822–827 in [13]). If the sample and population strata probabilities are known, then the standard ML and GMM estimation can be adapted to reflect the divergence. Typically this leads to weighted ML or weighted GMM estimation [46].

Clustered and Dependent Samples Survey data are usually dependent. This may reflect a feature of the survey sampling methodology, such as interviewing several households in a neighborhood. Consequently, the data may be correlated within cluster due to presence of a common unobserved cluster-specific term. Potentially, such dependence could also arise with SRS.

There are several different methods for controlling for dependence on unobservables within cluster. If the within cluster unobservables are uncorrelated with regressors then only the variances of the regression parameters need to be adjusted. This leads to cluster-correction-of-variances methods that are now well-embedded in popular software packages such as Stata. If, instead, the within cluster unobservables are correlated with regressors then the regression parameters are inconsistent and fixed effects type methods are called for. The issues and available methods closely parallel those for fixed and random effects panels models. Further, methods may also vary according to whether there are many small clusters or few large clusters. Examples and additional detail are given in Cameron and Trivedi [13].

An important new topic concerns dependence in cross section and panel data samples between independently obtained measures. Several alternative models are available to motivate such dependence. Social interactions [21] between individuals or households, and spatial dependence [7] where the observational unit is region, such as state, and observations in regions close to each other are likely to be interdependent, are examples. Models of social interaction analyze interdependence between individual choices (e. g. teenage smoking behavior) due to, for example, peer group effects. Such dependence violates the commonly deployed i.i.d. assumption, and in some cases the endogeneity assumption. Lee [57] and Andrews [4] examine the econometric implications of such dependence.

Major Insights

A major role of microeconometrics is inform public policy. But public policy issues arise not only in the context of existing policies whose effectiveness needs evaluation but also for new policies that have never been tried and old policies that are candidates for adoption in new economic environments. No single approach to microeconometrics is appropriate for all these policy settings. All policy evaluation involves comparison with counterfactuals. The complexity associated with generating counterfactuals varies according to the type of policy under consideration as well as the type of data on which models are based. A deeper understanding of this fundamental insight is a major contribution of modern microeconometrics.

A second major insight is the inherent difficulty of making causal inferences in econometrics. Many different modeling strategies are employed to overcome these challenges. At one end of the spectrum are highly structured models that make heavy use of behavioral, distributional and functional form assumptions. Such models address more detailed questions and provide, conditional on the framework, more detailed estimates of the policy impact. At the other end of the spectrum are methods that minimize on assumptions and aim to provide informative bounds for measures of policy impact. While the literature remains unsettled on the relative merits and feasibility of these approaches, the trend in microeconometrics is towards fewer and less restrictive assumptions.

There is now a greater recognition of the challenges associated with analyzes of large complex data sets generated by traditional sample surveys as well as other automated and administrative methods. In so far as such challenges are computational, advances in computer hardware and software technologies have made a major contribution to their solution.

Bibliography

Primary Literature

1. Albert JH, Chib S (1993) Bayesian Analysis of Binary and Polychotomous Response Data. J Am Stat Assoc 88:669–679
2. Allen RGD, Bowley AL (1935) Family Expenditure. PS King and Son, London
3. Amemiya T (1985) Advanced Econometrics. Harvard University Press, Cambridge (Mass)
4. Andrews DWK (2005) Cross-section Regression with Common Shocks. Econometrica 73:1551–1585
5. Angrist JD, Krueger AB (2000) Empirical Strategies in Labor Economics. In: Ashenfelter OC, Card DE (eds) Handbook of Labor Economics, vol 3A. North-Holland, Amsterdam, pp 1277–1397
6. Angrist JD, Imbens G, Rubin D (1996) Identification of Causal Effects Using Instrumental Variables. J Am Stat Assoc 91:444–455
7. Anselin L (2001) Spatial Econometrics. In: Baltagi BH (ed) A Companion to Theoretical Econometrics. Blackwell, Oxford, pp 310–330

8. Baker M, Melino A (2000) Duration Dependence and Nonparametric Heterogeneity: A Monte Carlo Study. J Econ 96:357–393
9. Blundell RW, Powell JL (2001) Endogeneity in Semiparametric Binary Response Models. Rev Econ Stud 71:581–913
10. Blundell RW, Smith RJ (1989) Estimation in a Class of Limited Dependent Variable Models. Rev Econ Stud 56:37–58
11. Bunch D (1991) Estimability in the Multinomial Probit Model. Transp Res Methodol 25B(1):1–12
12. Cameron AC, Trivedi PK (1998) Regression Analysis for Count Data. Econometric Society Monograph No. 30, Cambridge University Press, Cambridge
13. Cameron AC, Trivedi PK (2005) Microeconometrics: Methods and Applications. Cambridge University Press, Cambridge
14. Chesher A (2005) Nonparametric identification under discrete variation. Econometrica 73(5):1525–1550
15. Cosslett SR (1981) Maximum Likelihood Estimator for Choice-Based Samples. Econometrica 49:1289–1316
16. Deb P, Trivedi PK (1997) Demand for Medical Care by the Elderly: A Finite Mixture Approach. J Appl Econ 12:313–326
17. Deb P, Trivedi PK (2002) The Structure of Demand for Health Care: Latent Class versus Two-part Models. J Health Econ 21:601–625
18. Deb P, Trivedi PK (2006) Specification and Simulated Likelihood Estimation of a Non-normal Treatment-Outcome Model with Selection: Application to Health Care Utilization. Econ J 9:307–331
19. Diggle PJ, Heagerty P, Liang KY, Zeger SL (1994, 2002) Analysis of Longitudinal Data, 1st and 2nd editions. Oxford University Press, Oxford
20. Dubin J, McFadden D (1984) An Econometric Analysis of Residential Electric Appliance Holdings and Consumption. Econometrica 55:345–362
21. Durlauf S, Cohen-Cole E (2004) Social Interactions Models. In: Lempf-Leonard K (ed) Encyclopedia of Social Measurement. Academic Press, New York
22. Eckstein Z, Wolpin K (1989) The Specification and Estimation of of Dynamic Stochastic Discrete Choice Models: A Survey. J Hum Resour 24:562–598
23. Engel E (1857) Die Produktions- und Consumptionsverhältnisse des Königreichs Sachsen, Zeitschrift des Statistischen Bureaus des Königlich Sächsischen Ministerium des Innern, 22 November 1857. Reprinted in 1895 as appendix to E. Engel, Die Lebenskosten belgischer Arbeiter-Familien früher und jetzt. Bull Inst Int Stat 9:1–124
24. Frühwirth-Schnatter S (2006) Finite Mixture and Markov Switching Models. Springer, New York
25. Geweke J (1992) Evaluating the Accuracy of Sampling-based Approaches to the Calculation of Posterior Moments (with discussion). In: Bernardo J, Berger J, Dawid AP, Smith AFM, (eds) Bayesian Statistics, vol 4. Oxford University Press, Oxford, pp 169–193
26. Greene WH (2004) The Behavior of the Fixed Effects Estimator in Nonlinear Models. Econ J 7(1):98–119
27. Greene WH (2004) Fixed Effects and the Incidental Parameters Problem in the Tobit Model. Econ Rev 23(2):125–148
28. Gouriéroux C, Monfort A (1996) Simulation Based Econometrics Methods. Oxford University Press, New York
29. Hajivassiliou V, McFadden D, Ruud P (1996) Simulation of multivariate normal rectangle probabilities and their derivatives: Theoretical and computational results. J Econ 72:85–134
30. Heckman JJ (1974) Shadow Prices, Market wages, and Labor Supply. Econometrica 42:679–94
31. Heckman JJ (1976) The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for such Models. Ann Econ Soc Meas 5:475–492
32. Heckman JJ (1978) Dummy Endogenous Variables in a Simultaneous Equations System. Econometrica 46:931–960
33. Heckman JJ (1979) Sample Selection as a Specification Error. Econometrica 47:153–61
34. Heckman JJ (2000) Causal Parameters and Policy Analysis in Economics: A Twentieth Century Perspective. Quart J Econ 115:45–98
35. Heckman JJ (2001) Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture. J Political Econ 109(4):673–748
36. Heckman JJ, Robb R (1985) Alternative Methods for Evaluating the Impact of Interventions. In: Heckman JJ, Singer B (eds) Longitudinal Analysis of Labor Market Data. Cambridge University Press, Cambridge, UK
37. Heckman JJ, Singer B (1984) A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models of Duration Data. Econometrica, 52:271–320
38. Heckman JJ, Vytlacil E (2001) Local instrumental variables. In: Hsiao C, Morimue K, Powell JL (eds) Nonlinear Statistical Modeling. In: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in the Honor of Takeshi Amemiya. Cambridge University Press, New York, pp 1–46
39. Heckman JJ, Vytlacil E (2005) Structural Equations, Treatment Effects, and Econometric Policy Evaluation. Econometrica 73(3):669–738
40. Holland PW (1986) Statistics of Causal Inference. J Am Stat Assoc 81:945–60
41. Horowitz J (1998) Semiparametric Methods in Econometrics. Springer, New York
42. Hotz V, Miller R (1993) Conditional Choice Probabilities and the Estimation of Dynamic Models. Rev Econ Stud 60:497–529
43. Houthakker HS (1957) An International Comparison of Household Expenditure Patterns. Econometrica 25:532–552
44. Hsiao C (1986, 2003) Analysis of Panel Data, 1st and 2nd edn. Cambridge University Press, Cambridge
45. Imbens GW, Angrist J (1994) Identification and Estimation of Local Average Treatment Effect. Econometrica 62:467–475
46. Imbens GW, Lancaster T (1996) Efficient Estimation and Stratified Sampling. J Econ 74:289–318
47. Keane MP (1992) A Note on Identification in the Multinomial Probit Model. J Bus Econ Stat 10:193–200
48. Keane MP (1994) A Computationally Practical Simulation Estimator for Panel Data. Econometrica 62:95–116
49. Keane MP (2006) Structural vs. Atheoretical Approaches to Econometrics. Unpublished paper
50. Keane M, Wolpin K (1994) The Solutions and Estimation of Discrete Choice Dynamic Programming Models by Simulation and Interpretation: Monte Carlo Evidence. Rev Econ Stat 76:648–672
51. Koenker R (2005) Quantile Regression. Cambridge University Press, New York

52. Koenker R, Bassett G (1978) Regression Quantiles. Econometrica 46:33–50
53. Koop G, Poirier D, Tobias JL (2007) Bayesian Econometric Methods. Cambridge University Press, Cambridge
54. Lancaster T (2000) The Incidental Parameter Problem since 1948. J Econ 95:391–413
55. Lancaster T, Imbens GW (1996) Case-Control with Contaminated Controls. J Econ 71:145–160
56. Lee LF (2001) Self-selection. In: Baltagi B (ed) A Companion to Theoretical Econometrics. Blackwell Publishers, Oxford, pp 381–409
57. Lee LF (2004) Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Econometric Models. Econometrica 72:1899–1926
58. Lee MJ (2002) Micro-Econometrics for Policy, Program and Treatment Effects. Oxford University Press, Oxford
59. Luce D (1959) Individual Choice Behavior. Wiley, New York
60. Ma L, Koenker R (2006) Quantile regression methods for recursive structural equation models. J Econ 134:471–506
61. Manski CF (1975) Maximum Score Estimation of the Stochastic Utility Model of Choice. J Econ 3:205–228
62. Manski CF (1995) Identification Problems in the Social Sciences. Harvard University Press, Cambridge
63. Manski CF (2003) Partial Identification of Probability Distributions. Springer, New York
64. Manski CF, Lerman SR (1977) The Estimation of Choice Probabilities from Choice-Based Samples. Econometrica 45:1977–1988
65. Manski CF, McFadden D (1981) Alternative Estimators and Sample Design for Discrete Choice Analysis. In: Manski CF, McFadden D (eds) Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge, pp 2–50
66. Manski CF, McFadden D (eds) (1981) Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge
67. McCulloch R, Rossi P (2000) Bayesian analysis of the multinomial probit model. In: Mariano R, Schuermann T, Weeks M (eds) Simulation-Based Inference in Econometrics. Cambridge University Press, New York
68. McFadden D (1973) Conditional Logit Analysis of Qualitative Choice Behavior. In: Zarembka P (ed) Frontiers of Econometrics. Academic Press, New York
69. McFadden D (1976) Quantal Choice Analysis: A Survey. Ann Econ Soc Meas 5(4):363–390
70. McFadden D (1978) Modelling the Choice of Residential Location. In: Karlquist A et al (eds) Spatial Interaction Theory and Residential Location. North Holland, Amsterdam, New York
71. McFadden D (1981) Econometric Models of Probabilistic Choice. In: Manski CF, McFadden D (eds) Structural Analysis of Discrete Data with Economic Applications. MIT Press, Cambridge, pp 198–272
72. McFadden D (1984) Econometric Analysis of Qualitative Response Models. In: Griliches Z, Intriligator M (eds) Handbook of Econometrics, vol 2. North Holland, Amsterdam
73. McFadden D (2001) Economic Choices. Am Econ Rev 91(3):351–78
74. McFadden D, Ruud PA (1994) Estimation by Simulation. Rev Econ Stat 76:591–608
75. McFadden D, Train K (2000) Mixed MNL Models of Discrete Response. J Appl Econ 15:447–470
76. Maddala GS (1983) Limited Dependent and Qualitative Variables in Economterics. Cambridge University Press, Cambridge
77. Marschak J, Andrews WH (1944) Random Simultaneous Equations and the Theory of Production. Econometrica 12:143–205
78. Miller R (1997) Estimating Models of Dynamic Optimization with Microeconomic Data. In: Pesaran HH, Schmidt P (eds) Handbook of Applied Econometrics, vol II. Blackwell Publishers, Oxford
79. Prais SJ, Houthakker HS (1955) Analysis of Family Budgets. Cambridge University Press, Cambridge
80. Pudney S (1989) Modeling Individual Choice: The Econometrics of Corners. Basil, Kinks and Holes. Blackwell, New York
81. Rust J (1987) Optimal replacement of gmc bus engines: An empirical model of Harold Zurcher. Econometrica 55:993–1033
82. Rust J (1994) Estimation of dynamic structural models, problems and prospects: Discrete decision processes. In: Sims C (ed) Advances in Econometrics: Sixth World Congress, vol II. Cambridge University Press, New York, pp 5–33
83. Rust J (1994) Structural Estimation of Markov Decision Processes. In: Engle RF, McFadden D (eds) Handbook of Econometrics, vol 4. North-Holland, Amsterdam, pp 3081–3143
84. Rust J (1997) Using Randomization to Break the Curse of Dimensionality. Econometrica 65:487–516
85. Rust J, Phelan C (1997) How Social Security and Medicare Affect Treatment Behavior in a World of Incomplete Markets. Econometrica 65(4):781–842
86. Stone R (1953) The Measurement of Consumers' Expenditure and Behavior in the United Kingdom, vol 1. Cambridge University Press, Cambridge, pp 1920–1938
87. Thurstone L (1927) A Law of Comparative Judgment. Psychol Rev 34:273–286
88. Tobin J (1958) Estimation of Relationships for Limited Dependent Variables. Econometrica 26:24–36
89. Train KE (2003) Discrete Choice Methods with Simulation. Cambridge University Press, Cambridge
90. Walker J, Ben-Akiva M (2002) Generalized random utility model. Math Soc Sci 43:303–343

Books and Reviews

Arellano M (2003) Panel Data Econometrics. Oxford University Press, Oxord, UK

Blundell R, Powell JL (2003) Endogeneity in Nonparametric and Semiparametric Regression Models. In: Dewatripont M, Hansen LP, Turnovsky SJ (eds) Advances in Economics and Econonometrics: Theory and Applications, Eighth World Congress, vol II. Cambridge University Press, Cambridge

Deaton A (1997) The Analysis of Household Surveys: A Microeconometric Approach to Development Policy. Johns Hopkins, Baltimore

Greene WH (2007) Econometric Analysis, 6th edn. Macmillan, New York

Hensher DA, Rose J, Greene W (2005) Applied Choice Analysis: A Primer. Cambridge University Press, New York

Lancaster T (1990) The Econometric Analysis of Transitional Data. Cambridge University Press, New York

Lee MJ (2002) Panel Data Econometrics: Methods-of-Moments and Limited Dependent Variables. Academic Press, San Diego

Louviere J, Hensher D, Swait J (2000) Stated Choice Methods: Analysis and Applications. Cambridge University Press, New York
Wooldridge JM (2002) Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge
Yatchew A (2003) Semiparametric Regression for the Applied Econometrician. Cambridge University Press, Cambridge

Microfluidics

Sandip Ghosal
Department of Mechanical Engineering, Northwestern University, Evanston, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Physics of Microfluidics
Future Directions
Bibliography

Glossary

Reynolds number A characteristic dimensionless number that determines the nature of fluid flow in a given set up.

Stokes approximation A simplifying approximation often made in fluid mechanics where the terms arising due to the inertia of fluid elements is neglected. This is justified if the Reynolds number is small, a situation that arises for example in the slow flow of viscous liquids; an example is pouring honey from a jar.

Ion mobility Velocity acquired by an ion per unit applied force.

Electrophoretic mobility Velocity acquired by an ion per unit applied electric field.

Zeta-potential The electric potential at the interface of an electrolyte and substrate due to the presence of interfacial charge. Usually indicated by the Greek letter zeta (ζ).

Debye layer A thin layer of ions next to charged interfaces (predominantly of the opposite sign to the interfacial charge) due to a balance between electrostatic attraction and random thermal fluctuations.

Debye length A measure of the thickness of the Debye layer.

Debye–Hückel approximation The process of linearizing the equation for the electric potential; valid if the potential energy of ions is small compared to their average kinetic energy due to thermal motion.

Electric double layer (EDL) The Debye layer together with the set of fixed charges on the substrate constitute an EDL. It may be thought of as a parallel plate capacitor or a layer of dipoles .

Electrophoresis The motion of charged objects in a fluid due to an imposed electric field.

Dielectrophoresis The phenomenon that results in a polarizable dielectric medium experiencing a force in a non-uniform electric field.

Electroosmosis The motion of an ionic liquid relative to a fixed charged substrate due to an imposed electric field.

AC Electroosmosis or induced charge electroosmosis (ICEO) Fluid flow caused by an electric field in an ionic medium containing embedded polarizable objects (such as metal cylinders). The effect is non-linear in the electric field and can result in unidirectional flow even with an AC voltage.

Thermocapillary effect Motion of fluid or object in fluid due to spatial inhomogeneities in surface tension caused by variations in the temperature field.

Marangoni effect Same as the Thermocapillary effect except that the variation of surface tension is caused by inhomogeneities in the concentration of a dissolved chemical species such as a surfactant.

Electrowetting The phenomenon that results in changes of contact angle of a conducting liquid drop placed on a conducting plate when the electric potential of the plate is changed.

Electrowetting on dielectric (EWOD) The same phenomenon as electrowetting, except a thin dielectric coating is applied on the metal plate so that the liquid and the plate are not in electrical contact.

Definition of the Subject

Microfluidics is the science of manipulating fluids on spatial scales anywhere between one to a hundred micron. In SI units a micro-meter (μm) or micron refers to a millionth of a meter. To put this in perspective, the average width of a human hair is about 80 microns and the diameter of a hydrogen atom is 0.00005 micron. Thus, microfluidics involves engineering structures for manipulating fluids on scales that are microscopic in comparison to human dimensions but that are nevertheless much larger than atomic dimensions so that the systems can still be treated in the continuum approximation. Microfluidic devices are commonplace in the world of living things; for example, the narrowest capillaries in the human circulatory system are of the order of 5–10 microns and the diameter of a swimming bacteria maybe about 10 microns.

However, devices engineered by humans that can be considered microfluidic is of much recent vintage. The methods as well as the motivation for designing microfluidic devices entered an era of rapid development in the last ten years or so, a trend that has continued to accelerate to this date. The primary driving force behind this trend has come from Molecular Biology and the Life Sciences where progress has become dependent upon the ability to perform analytical chemistry at heretofore unprecedented speeds. As of 2006 there were about sixty companies in the US engaged in marketing microfluidic products. Revenues from such products reached 1.7 billion US dollars in 2003 and is projected to reach 2.7 billion US dollars in 2008.

Introduction

The idea of a microfluidic chip is derived from the concept of a microelectronic chip. The difference is, channels through which fluids flow take the place of conducting pathways for electrons. The transistors and other microelectronic elements are replaced by mixers, separators and other micro-fluidic elements designed to perform the basic tasks of analytical chemistry.

The Case for Microfluidics

The technology of microelectronics was invented and developed to fulfill a specific need: processing of digital information in devices that are smaller, cheaper and faster. The evolution of microfluidics is being driven by exactly the same considerations except that the function that needs to be performed is related to analytical chemistry. Applications in modern biology and medicine such as drug testing, gene sequencing and gene expression studies require the performance of extremely large numbers of repetitive steps. For example, large banks of robots were employed in the Human Genome Project which nevertheless took eleven years and over three billion dollars to complete. An analogy can be drawn with the Manhattan project where large scale numerical calculations were laboriously performed by rooms full of human 'calculators'. However, large scale automated laboratories for such functions as gene sequencing can only be a temporary measure. The ultimate enabling technology for 'super-chemistry' must be minaturization, as it was for 'super-computing'. This is because (a) as device size shrinks, the time needed to perform a specific operation (such as electrophoretic separation) becomes smaller (b) a large number of identical steps can be run in parallel on a small device (c) the amount of sample and reagents consumed becomes much smaller (d) the devices themselves can be mass produced lowering the unit cost.

A Brief History of Microfluidics

Unlike microelectronics, which by now must be considered a mature technology, microfluidics is still in its infancy. In fact microfluidics draws heavily from many of the same technological breakthroughs that led to the current sophistication in the manufacture of computer chips. It is difficult to say when microfluidics really began or what the first "microfluidic device" was. If one uses the definition of microfluidics given in the introductory paragraph as something that involves a device that manipulates fluid on a 1–100 micron scale, then perhaps the ink jet printer should qualify as one of the earliest microfluidic devices. The first inkjet printer was patented by Siemens in 1951 and subsequently went through many stages of innovations and improvement through the efforts of companies such as HP, Xerox, IBM, Cannon, Epson and others. The fluid mechanical principles involved are rich in complexity and there are excellent reviews available on the subject [1]. An early example of a microfluidic chip is due to Jacobson and Ramsey who demonstrated DNA Restriction Fragment Analysis on a single chip under computer control [2]. The first commercial "Lab on a Chip" was the AGILENT 2100 Bioanalyzer introduced in 1999. The number of patents issued per year for inventions related to some aspect of commercial microfluidic applications has grown from a handful per year in the nineties to more than 350 patents a year.

Natural Microfluidic Systems

Microfluidic systems of beautifully elegant design are commonplace in the natural world. A stunning example is the fog harvesting *Stenocara* beetle of the Kalahari desert in Namibia. The shell of this beetle is a patterned substrate of alternating hydrophillic and hydrophobic areas [3]. The beetle positions itself on top of sand dunes with its head down and facing the ocean wind that blows in the early morning fog. The surface patterning is designed to catch the water droplets, grow them to a critical size and then detach them. This size selection is important, for the droplets need to be large enough to roll down by gravity against the prevailing headwind and not be blown off the beetle's back. The African bombardier beetle (*Stenaptinus insignis*) is armed with a potent microfluidics based weapon [4]. Its abdomen contains a reaction chamber which receives a propellant (hydroquinone) and an oxidizer (hydrogen peroxide) from separate storage organs. These two chemicals combine explosively in the presence of catalysts secreted within the reaction chamber. The result is a boiling mixture of steam and corrosive liquid which the beetle is able to deliver on to the target through a steerable nozzle

Microfluidics, Figure 1
Left panel: **Chipsets for the Agilent 2100 Bioanalyzer for performing various bio-analytical procedures** ***Right panel:*** **an example of a microfluidic chip contained within the package. (Image: courtesy of Agilent Technologies)**

Microfluidics, Figure 2
Examples of Microfluidic systems in the living world (A) the ***Stenocara*** **beetle from Namibia (*left*) that harvests water droplets from fog using a textured surface with alternating hydrophobic/hydrophillic areas as seen in the microscope image (*middle*) (B) the African "Bombardier Beetle" (*Stenaptinus insignis*) discharging a boiling toxic liquid jet (*right*) through steerable micro-nozzles (Images reprinted with permission from *Nature* and *Proc. Natl. Acad. Sci. USA* – full citations in text)**

with an accompanying loud bang! Other examples include fluid flow through the proboscis of insects such as butterflies and mosquitoes, the flagellar propulsion of bacteria such as *E. Coli*, propulsion by a single waving flagellum in spermatozoa, use of cillia by numerous small animals for moving fluids along and the interesting use of surface tension forces by insects and larva that live on the air water interface in stagnant pools [5].

Understanding Microfluidics

A rich variety of microfluidic systems have arisen in the natural world through millions of years of evolution and natural selection. However, the history of microfluidics as an area of human endeavor can be measured in decades. It is therefore not surprising that practical microfluidic devices are relatively few, many ideas such as the "Lab on a Chip" exist mostly in concept and the most useful of microfluidic devices are probably yet to be conceived. In order to translate concepts into practical devices, we must learn to control fluids with great precision and speed at ultra small scales. Such control can only be achieved through a deep knowledge of the behavior of fluids on ultra small scales. It is the purpose of this article to give an overview of the mathematical laws that govern such fluid behavior in the small.

Physics of Microfluidics

The laws of fluid motion for microfluidic systems are not any different from those that govern large scale systems such as the oceanic currents on intercontinental scales. However, the relative importance of different forces and effects change dramatically as we go from macro to mi-

cro scales. Thus, for example, surface tension forces and electrostatics play no role at all in the study of ocean currents but are enormously important in the world of microfluidics.

The remainder of this section is presented in three subsections, **Stokes Flow** which deals with fluid motions that are dominated by viscous resistance, **Electrokinetics** which deals with the interaction of liquids with electrostatic forces and finally the motion of fluids driven by forces related to **Surface Tension**. The flow of gases is not considered at all in this article because microfluidic devices based on the flow of a gas is not very common. In the case of gas flow, the most important physical effect to consider is the likely break down of the classical gas dynamic equations as the mean free path approaches the channel diameter. The flow of rarefied gases have been reviewed by Muntz [6].

Stokes Flow

In microfluidics, relevant physical dimensions are sufficiently large in comparison to atomic scales that it is permissible to treat the fluid as if it were a continuum. Thus, the fluid velocity $\mathbf{u}$ and pressure p are regarded as continuous functions of position $\mathbf{x}$ and time t, and they obey the incompressible Navier–Stokes equations with a volume density of external forces $\mathbf{f}_e$

$$\rho_0(\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u}) = -\nabla p + \mu \nabla^2 \mathbf{u} + \mathbf{f}_e \,. \tag{1}$$

This is supplemented by the continuity equation which takes into account the fact that in a liquid the density changes are slight, even for large changes in pressure:

$$\nabla \cdot \mathbf{u} = 0 \,. \tag{2}$$

In the above, ρ_0 is the (constant) density of the fluid, p is the pressure, $\mathbf{u}$ is the flow velocity.

The relative size of the term on the left of Eq. (1) (due to fluid inertia) and the second term on the right (due to viscosity) is characterized by the Reynolds number

$$\mathrm{Re} = \frac{UL\rho_0}{\mu} \tag{3}$$

where U and L denote a characteristic velocity and length for the flow. In most applications of microfluidics, $\mathrm{Re} \ll 1$. In some applications, $\mathrm{Re} \sim 1$. By contrast, in large scale flows (aircraft engines, geophysical flows etc.) $\mathrm{Re} \gg 1$ is the rule. Because of the smallness of the Reynolds number, Re, in microfluidics, most well known flow instabilities leading to period doubling, chaos and finally turbulence are absent. Furthermore, the left hand side of Eq. (1) which corresponds to fluid inertia can either be neglected, or treated as a small perturbation. In the former case, we arrive at the Stokes flow equations:

$$-\nabla p + \mu \nabla^2 \mathbf{u} + \mathbf{f}_e = 0 \tag{4}$$

which is often referred to as slow, creeping or highly viscous flow. All of these terms mean the same thing, namely $\mathrm{Re} = (UL\rho_0)/\mu \ll 1$. The unknown scalar field p in Eq. (4) is determined by the constraint provided by Eq. (2).

When the flow is bounded in one or more spatial directions, boundary conditions must be specified to guarantee unique solutions to Eqs. (4) and (2). Solid boundaries are most commonly encountered where the classical "no slip" boundary condition

$$\mathbf{u}(P) = \mathbf{u}_{\mathrm{solid}}(P) \tag{5}$$

is employed: the fluid velocity at a point P on the solid fluid interface must match the velocity of the solid, $\mathbf{u}_{\mathrm{solid}}(P)$ at that interfacial point. The justification is empirical, and based on the intuitive picture that on the micro-scale any smooth surface is actually a rugged terrain of peaks and valleys which effectively randomizes the direction of the velocity vector of a molecule after it collides with the wall. On the scale of macroscopic experiments, these boundary conditions have always provided excellent agreement between theory and experiments except for describing the motion of the contact line for drops moving on a substrate. In these moving contact line problems, the no-slip boundary conditions lead to a singularity of the stress at the contact line [7]. In the flow of low density gases through microchannels, the mean free path of molecules could approach the channel diameter. In such cases deviations from the Navier-Stokes equations with the no-slip conditions have been observed [8,9]. In microfluidic systems, the validity of the no slip condition for liquid flow have been questioned both from a theoretical as well as an experimental perspective [10]. A consensus is yet to be reached, however it seems that the deviation from the no-slip condition, if there is such a deviation, is quite small. The interpretation of experimental data is often complicated by the presence of adsorbed gas microbubbles on the interface, particularly if the interface is hydrophobic in character [11].

Electrokinetics

Electrokinetics refers to mechanical effects that arise due to the motion of ions in liquids. The working fluid in microfluidic systems is normally water which contains ions

of both signs due to dissociated water molecules or other ionic components: acids, salts, and molecules with dissociable charged groups. Normally, a volume element of such a fluid considered "infinitesimal" in the continuum viewpoint still contains a sufficiently large number of ions of either sign for statistical fluctuations to be unimportant and for the fluid element to be considered charge neutral. Therefore, the net algebraic transfer of momentum due to any ambient electric field is also zero (even though a non-zero electric current may exist in the fluid due to the ordered motion of these ions). Electrokinetic effects arise when this balance of positive and negative charges is disturbed due to external factors. For example, at the silica water interface, the hydrated silica often deprotonates resulting in a net negative fixed surface charge on the silica surface. These fixed charges attract a layer of ions of the opposite sign (and repel ions of like sign) resulting in the creation of a fluid layer with a net positive charge next to the interface. Similar effects arise at the surface of large macromolecules, colloidal particles or surfactant micelles. Thus, in the macroscopic description, the fluid in these so called Debye Layers experience an electrical force with volume density $\mathbf{f}_e = -\rho_e \nabla\phi$ where ρ_e is the electric charge density and ϕ is the electric potential.

Ion Transport The electric potential ϕ is related to the charge density through the Poisson equation [12] of electrostatics:

$$\epsilon \nabla^2 \phi = -4\pi\rho_e \,, \tag{6}$$

ε being the dielectric constant of the liquid (in CGS units). If the electrolyte contains N species of ions with charges ez_k and concentration n_k ($k = 1, \ldots, N$, and e is the magnitude of the electron charge) then $\rho_e = \sum_{k=1}^{N} ez_k n_k$. Each ion species obeys a conservation equation

$$\frac{\partial n_k}{\partial t} + \nabla \cdot \mathbf{j}_k = 0 \,. \tag{7}$$

Here $\mathbf{j}_k$, the flux vector for the species k can be modeled by the Nernst–Planck equation for ion transport [13]

$$\mathbf{j}_k = -\nu_k z_k e n_k \nabla\phi - D_k \nabla n_k + n_k \mathbf{u} \,. \tag{8}$$

In Eq. (8), ν_k is the ion mobility: the velocity acquired by the ion when acted upon by a unit of external force. It is obviously related to the electrophoretic mobility: the velocity per unit of electric field as $\mu_k^{(ep)} = ez_k\nu_k$. The diffusivity of the kth species is D_k and $\mathbf{u}$ is the fluid velocity.

The boundary conditions are those of no slip at the wall for the velocity, $\mathbf{u} = 0$, and no ion flux normal to the wall $\mathbf{j}_k \cdot \hat{\mathbf{n}} = 0$ ($\hat{\mathbf{n}}$ is the unit normal directed into the fluid). In the absence of external electric fields, the chemistry at the electrolyte substrate interface leads to the establishment of a potential, $\phi = \zeta$. This so called ζ-potential at an interface depends on a number of factors including the nature of the substrate and ionic composition of the electrolyte, the presence of impurities, the temperature, the buffer pH and is even known to exhibit hysteresis effects with respect to pH [14]. Methods of determining the ζ-potential and measured values for a wide variety of surfaces used in microfluidic technology have been reviewed in [15] and [16].

Equilibrium Debye Layers The ion distribution near a planar wall at $z = 0$ with potential $\phi(z)$ is known from statistical thermodynamics: $n_k = n_k(\infty)\exp(-z_k e\phi/k_B T)$ where k_B is the Boltzmann constant and T is the absolute temperature of the solution. In order that this expression be a steady solution of Eq. (7) we must have the Einstein relation $D_k/\nu_k = k_B T$. Therefore Eq. (8) can also be written as

$$\mathbf{j}_k = -n_k \nu_k \nabla\psi_k + n_k \mathbf{u} \tag{9}$$

where $\psi_k = ez_k\phi + k_B T \ln n_k$ is called the chemical potential for the species k. To determine ϕ, one must solve a self-consistent coupled problem because the species concentrations depend upon ϕ through $n_k = n_k(\infty)\exp(-z_k e\phi/k_B T)$, but ϕ also depends on the n_k through the Poisson equation, Eq. (6).

The Gouy–Chapman model Suppose that the system is in the steady state and that there is no fluid flow or imposed electric fields. Further suppose that the geometry is such that the electrolyte-substrate interface is an iso-surface of ψ_k. Then it readily follows from Eqs. (7), (9) and the boundary condition of no flux into the wall that $\nabla\psi_k = 0$ everywhere. Therefore, $n_k = n_k^{(\infty)}\exp(-z_k e\phi/k_B T)$ where $n_k^{(\infty)}$ is the ion concentration where the potential $\phi = 0$; usually chosen as a point very far from the wall. Using the solution for n_k in the charge density ρ_e and substituting in Eq. (6), we get the non-linear Poisson–Boltzmann equation for determining the potential

$$\nabla^2\phi = -\frac{4\pi e}{\epsilon}\sum_{k=1}^{N} n_k^{(\infty)} z_k \exp\left(-z_k e\phi/k_B T\right) \,. \tag{10}$$

with the boundary condition $\phi = \zeta$ on walls.

Equation (10) was the starting point of a detailed investigation of the structure of the Electric Double Layer

(EDL) by Gouy [17] and Chapman [18]. The description of the EDL in terms of Eq. (10) is therefore known as the Gouy–Chapman model of the EDL.

The Debye–Hückel approximation Equation (10) is a nonlinear equation and solutions can only be constructed by numerical methods. Debye and Hückel linearized it by expanding the exponential terms on the right hand side in Taylor series and discarding all terms that are quadratic or of higher order in ϕ, which gives

$$\nabla^2\phi - \kappa^2\phi = 0 \tag{11}$$

where

$$\kappa = \left[\sum_{k=1}^{N} \frac{4\pi z_k^2 e^2 n_k^{(\infty)}}{\epsilon k_B T}\right]^{1/2} \tag{12}$$

is a constant determined by the ionic composition of the electrolyte. In arriving at Eq. (11) we used the condition $\sum_{k=1}^{N} n_k^{(\infty)} z_k = 0$ which expresses the fact that the bulk solution ($\phi = 0$) is free of net charge. It is easily verified that κ has units such that $\lambda_D = \kappa^{-1}$ defines a length scale that is called the Debye-Length. If a charged plane at a potential ζ is introduced in an electrolyte where the Debye–Hückel approximation is valid, then the solution to Eq. (11) may be written as $\phi = \zeta \exp(-\kappa z) = \zeta \exp(-z/\lambda_D)$ where z is distance normal to the plate. Thus, the potential due to the charged plate is shielded by the free charges in solution and the effect of the charge penetrates a distance of the order of the Debye-length λ_D; which gives a physical meaning to this very important quantity. In microfluidic applications the Debye length is typically between 1 to 10 nm.

The linearization proposed by Debye and Hückel is justified provided that $|z_k\phi| \ll k_B T/e$ uniformly in all space and for all k. At room temperature $k_B T/e \approx 30$ mV. However, for silica substrates $|\zeta| \sim 50 - 100$ mV in typical applications. Thus, the Debye–Hückel approximation is often not strictly valid. Nevertheless, it is a very useful approximation because it enormously simplifies mathematical investigations related to the Debye layer. Further, all deductions from it are usually qualitatively correct and even quantitative predictions from it outside its range of formal validity tend not to differ from the true solution by a large amount.

Thin Debye Layers and Apparent Slip The characteristic width of microfluidic channels is typically of the order of 10–100 μm, whereas the Debye length $\lambda_D \sim 1 - 10$ nm. Thus, the Debye layer is exceedingly thin compared to characteristic channel diameters. Under those circumstances, the Navier–Stokes/Poisson–Boltzmann system described in the last section may be replaced by a simpler set of equations. Indeed, the EDL then forms a very thin boundary layer at the solid fluid interface where the electrical forces are confined.

In the outer region (that is, outside the EDL) we have a fluid that is electrically conducting but charge neutral. Its motion is therefore described by the Navier-Stokes equations without the electrical force term. Inside the EDL, the problem reduces to that of an electric field parallel to a planar interface, therefore, the use of the Gouy–Chapman model is justified. From the solution (18)

$$u(z \to \infty) = -\frac{\epsilon\zeta E_0}{4\pi\mu}\,. \tag{13}$$

In the language of asymptotic theory, this is the "outer limit of the inner solution" [19] that becomes the boundary condition for the outer problem. Thus, in the limit of infinitely thin EDL, the standard no-slip boundary conditions of fluid mechanics is replaced by the following apparent slip condition:

$$\mathbf{u} - \mathbf{u}_{\text{solid}} \equiv \mathbf{u}_s = -\frac{\epsilon\mathbf{E}\zeta}{4\pi\mu}\,. \tag{14}$$

In Eq. (14), $\mathbf{u}_{\text{solid}}$ is the velocity of the solid, and $\mathbf{u}$ is the velocity of the fluid at a point on the solid fluid interface, $\mathbf{u}_s$ is the slip velocity. The electric field $\mathbf{E}$ is the field in the buffer due to external sources and can be determined by solving Laplace's equation within the electrolyte with the boundary condition $\mathbf{E}\cdot\hat{\mathbf{n}} = 0$ on the wall (since the walls are insulating and there can be no current flowing into them). Equation (14) is known as the Helmholtz–Smoluchowski (HS) slip boundary condition after the pioneering work of [20] and [21]. It is the starting point for solutions of more complex problems such as that of flow through axially inhomogeneous channels [22,23].

Electroosmosis In the presence of external fields and fluid flow the equilibrium Gouy–Chapman model is generally not applicable and one must proceed from the full electrokinetic equations presented earlier. However, if the external field and fluid velocity are both along the isosurfaces of the charge density ρ_e then the presence of the flow or the imposed field does not alter the charge density distribution which may still be obtained from the Gouy–Chapman model. Examples where such a situation holds would be

1. A planar uniformly charged substrate at $z = 0$ with an applied electric field E_0 that is tangential to the surface (the x-direction).

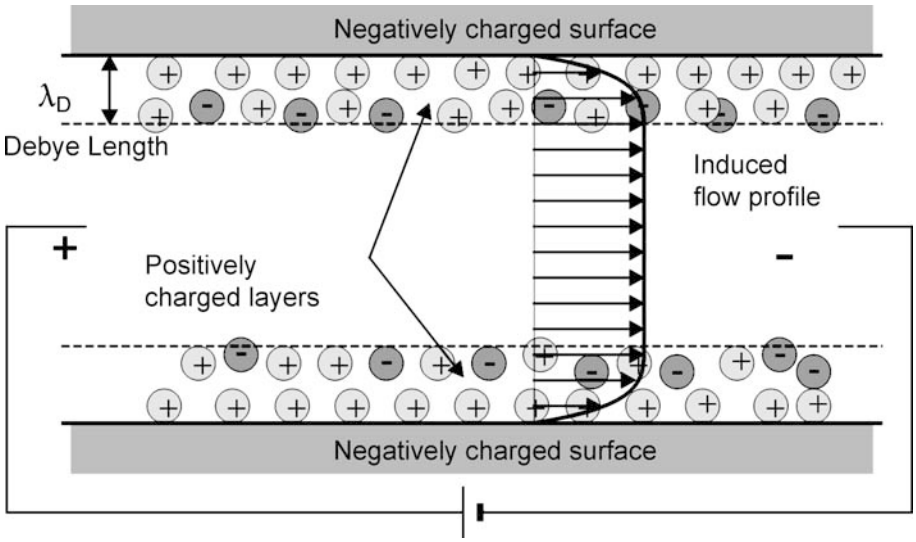

Microfluidics, Figure 3
A sketch showing the Debye layers formed at the walls of a parallel channel and the resultant electroosmotic flow in response to an applied voltage (Image: courtesy of Prof. H. Bruus of the Technical University of Denmark)

2. A uniform infinite cylindrical capillary with an imposed electric field E_0 along the axis (the x-direction).
3. A narrow slit with uniformly charged walls and an imposed constant electric field E_0 along the slit (the x-direction).

For any of the above geometries, the fluid flow equations reduce to (assuming steady state and zero imposed pressure gradient)

$$\mu\nabla^2 u + \rho_e E_0 = \mu\nabla^2 u - \frac{\epsilon E_0}{4\pi}\nabla^2\phi^{(\mathrm{EDL})} = 0 \tag{15}$$

where u is the axial velocity and $\phi^{(\mathrm{EDL})}$ is the electric potential for the equilibrium problem, that is, without the external field or flow. Therefore,

$$u = \frac{\epsilon E_0}{4\pi\mu}\phi^{(\mathrm{EDL})} + \chi \tag{16}$$

where χ satisfies

$$\nabla^2\chi = 0 \tag{17}$$

and $\chi = -(\epsilon E_0\zeta)/(4\pi\mu)$ at the boundaries. Assuming an infinitely long channel, by symmetry, solutions must be independent of the axial co-ordinate x. The only such solution is $\chi = -(\epsilon\zeta E_0)/(4\pi\mu)$. Therefore, the velocity is determined in terms of the potential distribution in the equilibrium EDL:

$$u = \frac{\epsilon E_0}{4\pi\mu}\left[\phi^{(\mathrm{EDL})} - \zeta\right] \tag{18}$$

If we adopt the Debye–Hückel approximation then

$$\phi^{(\mathrm{EDL})} = \begin{cases} \zeta\,\exp(-\kappa z) & \text{for (1) infinite plane} \\ \zeta\, I_0(\kappa r)/I_0(\kappa a) & \text{for (2) infinite cylindrical capillary} \\ \zeta\,\cosh(\kappa z)/\cosh(\kappa b) & \text{for (3) narrow slit} \end{cases} \tag{19}$$

where a is the capillary radius, r the distance from the axis and I_0 is the zero order modified Bessel function of the first kind. In the last formula, $2b$ is the channel width and z is the wall normal co-ordinate with origin on the plane (in Case 1) or origin at a point equidistant between the two walls (in Case 3). Since the fluid flow equation is linear in this limit, clearly a pressure driven flow can be added to the solution (superposition) in the event that both a pressure gradient and an electric field are simultaneously applied. The solution for an infinite capillary was first obtained by Rice and Whitehead [24]. Solutions in a narrow slit were obtained by Burgreen and Nakache [25] in the context of the Debye–Hückel approximation as well as for a 1 : 1 electrolyte (that is, in our notation $N = 2$ and $z_1 = -z_2$) directly from the full Poisson–Boltzmann equation.

Electrophoresis Electrophoresis refers to the transport of small charged objects in a fluid due to an applied electric field. Many macromolecules contain dissociable charge groups on its surface, and therefore, spontaneously acquire a charge in aqueous solution. They therefore move in

response to an applied electric field, and this motion is the basis for separating macromolecules from solution using such bioanalytical techniques as Capillary Electrophoresis (CE), Slab Gel Electrophoresis (SGE) and Isoelectric Focussing (IEF). As far as the physics is concerned, electroosmosis and electrophoresis are essentially identical, in the former case the reference frame is fixed to the substrate and in the latter case the reference frame is fixed to the fluid at infinitely large distances from the object.

Uniformly Charged Sphere The simplest 'classic' problem in electrophoresis involves determining the velocity of a uniformly charged dielectric sphere placed in an ionic fluid (usually water with dissolved salts such as KCl). An exact analytical solution to even this reduced problem is unavailable except in certain special limits. These are the limits of extremely large Debye length (corresponds to very low ionic strengths) and extremely short Debye length (corresponds to very high ionic strength).

Indeed, for low ionic strengths, the velocity v may be obtained by simply equating the electric force qE to the viscous force $6\pi\mu a v$ where E is the applied electric field, q is the charge on the sphere, μ is the dynamic viscosity of the fluid and a is the particle radius, thus,

$$\mu_{\mathrm{ep}}^{(\infty)} = \mu_{\mathrm{ep}}(\text{low salt}) = \frac{v}{E} = \frac{q}{6\pi\mu a} = \frac{\epsilon\zeta}{6\pi\mu} \tag{20}$$

where the second part of the equality follows on expressing the charge q in terms of the electrostatic potential ζ on the surface of the sphere. The ratio $v/E = \mu_{\mathrm{ep}}$ is called the electrophoretic mobility.

For high ionic strengths, we may invoke the Helmholtz Smoluchowski apparent slip boundary conditions. Then the problem reduces to solving a Stokes Flow problem for a sphere with a pseudo 'slip' boundary condition. It is easy to verify that since the electric potential ϕ is irrotational as well as solenoidal, the quantity

$$\mathbf{u} = \frac{\epsilon\zeta\nabla\phi}{4\pi\mu} \tag{21}$$

satisfies the Stokes flow equations. Further, by construction, the apparent slip boundary conditions hold on the surface of the sphere. It can be easily shown that the flow defined by Eq. (21) corresponds to zero net force and torque on the particle. Therefore, it corresponds to the actual flow field in a reference frame fixed to the particle. The flow speed at distant points gives the electrophoretic speed of the particle, thus,

$$\mu_{\mathrm{ep}}^{(0)} = \mu_{\mathrm{ep}}(\text{high salt}) = \frac{v}{E} = \frac{\epsilon\zeta}{4\pi\mu}\,. \tag{22}$$

Therefore,

$$\mu_{\mathrm{ep}}^{(\infty)} = \frac{2}{3}\mu_{\mathrm{ep}}^{(0)}\,. \tag{23}$$

Approximate solutions for intermediate Debye lengths as well as situations where the nonlinear convective term in the Navier–Stokes and species transport equations are small but not negligible have been addresses by various researchers. A careful review is provided by Saville [26].

Non-Spherical Shapes and Non-Uniform Charge For a long time, Eq. (22) had been the basis for determining experimentally the zeta potentials of particles, though the particles were not always spherical in shape. A formal proof of the validity of Eq. (22) for particles of any shape was provided by Morrison [27]. Morrison showed that in the limit $\lambda_{\mathrm{D}}/R \to 0$, where R is the smallest radius of curvature on the surface of the particle, Eq. (22) remains valid irrespective of the particle shape provided that ζ, ε and μ are uniform, the particle is non-conducting and polarization effects are neglected.

The surface charge on small objects such as macromolecules, cells and crystalline colloidal objects like particles of clay are often distributed in a non-uniform manner. The effect of non-uniformity of the ζ potential on the movement of the particle was first investigated by Anderson [28]. Anderson considered a spherical particle in the limit of infinitely thin Debye layers. The distribution of the zeta-potential on the particle surface was assumed arbitrary but known. Anderson showed that in general the particle would both translate and rotate. The translational velocity depended upon the average value of ζ (the monopole) as well as on its quadrapole moment. The angular velocity depended upon the dipole and quadrapole moments and thus vanished in the case of a uniformly charged particle.

Dielectrophoresis Particles that are uncharged but polarizable experience a force in a non-uniform electric field, the resulting motion is known as dielectrophoresis. The force density in a polarizable medium is

$$\mathbf{f}_{\mathrm{e}} = (\mathbf{P}\cdot\nabla)\mathbf{E} \tag{24}$$

where $\mathbf{P}$ is the dipole moment per unit volume and $\mathbf{E}$ is the electric field. In an isotropic and linear medium, $\mathbf{P} = \chi\mathbf{E}$ where χ is the susceptibility. Since χ is usually positive, the dielectrophoretic force is normally directed from regions of weak electric field to that of strong fields. For particles that are polarizable and also carry a net charge, electrophoretic and dielectrophoretic effects act in conjunction. The interplay of the two effects can be exploited

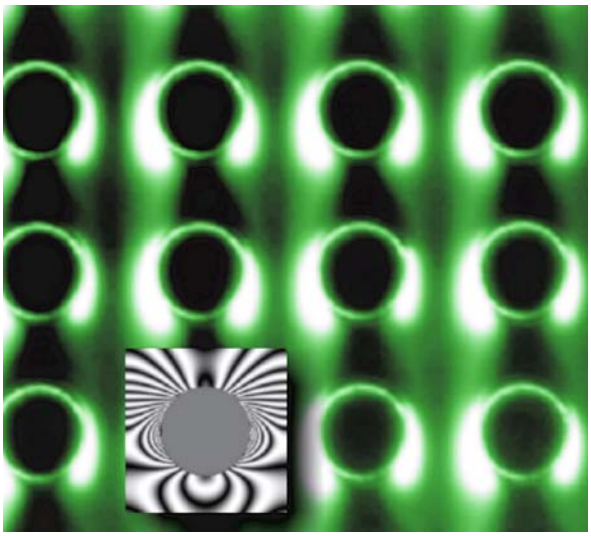

Microfluidics, Figure 4
Dielectrophoretic trapping of fluorescently labeled particles on a microchip with an etched pattern consisting of a periodic array of cylindrical obstacles. The overlayed picture shows the simulated isopotentials experienced by particles. (Reproduced with permission from [29])

for the design of elegant microfluidic elements. For example, Cummings et al. have designed a device that is able to separate live from dead bacterial cells in this way [29]. Figure 4 shows the trapping of cells labeled with a fluorescent dye in a 2D array of cylindrical posts. Another important aspect of dielectrophoresis is that unlike electrophoresis, the effect does not disappear if the constant electric field is replaced by an oscillating one. Thus, the dielectrophoretic force can be fine tuned by adjusting the frequency. Since a conducting object would develop a dipole moment in an external field, conducting regions inside a dielectric fluid (such as water droplets in oil) also experience the dielectrophoretic force.

AC Electroosmosis AC Electroosmosis [30,31,32] or more generally Induced Charge Electroosmosis (ICEO) refers to the flow of fluid due to the interaction of an applied electric field with Debye layers, except that the charges that give rise to the Debye layers are the polarization charges due to the original electric field. An example is shown in Fig. 5 which shows an insulator coated metal cylinder placed in an ionic medium. Switching on the electric field polarizes the cylinder causing the formation of the Debye layers. The tangential electric field acting on the unbalanced charges in the Debye layer drives a flow in the indicated direction. Since the polarization charges themselves are due to the applied field, ICEO is quadratic in the applied field. This nonlinearity opens up new possibilities, for example with cleverly shaped electrodes flow rectification can be achieved, that is, an oscillating field can drive a flow with a DC component. ICEO has a further advantage that electrolysis and the formation of gas bubbles, a disturbing side effect in direct current electroosmosis is eliminated if only oscillating fields are employed. Through careful design of metal on glass electrodes fluids can be effectively driven using the dielectrophoretic force by applying voltages that are much less than those required to drive flows by electroosmosis.

Surface Tension

The surface energy of a spherical drop of liquid of radius R is $E_s = 4\pi R^2\gamma$ where γ is the surface tension coefficient. If the drop rests on a surface which is taken as the zero level of potential, then the gravitational potential energy is (ρ is the fluid density) $E_g = (4/3)\pi R^3\rho g(R/2) = (2/3)\pi\rho g R^4$. Therefore, $E_g/E_s = (\rho g R^2)/(6\gamma) = \mathrm{Bo}/6$. The dimensionless number, $\mathrm{Bo} = \rho g R^2/\gamma$ is called the Bond number and it measures the relative importance of gravity and surface tension effects. The value $R = R_c = \sqrt{\gamma/\rho g}$ that makes the Bond number unity is called the capillary length. For water droplets, $R_c = 1 - 3\,\mathrm{mm}$. Thus, droplets of water smaller than one millimeter are dominated by surface tension forces, a fact that is readily apparent in the everyday world from the nearly spherical shape of small water droplets. The forces of surface tension can be used to manipulate small droplets on a surface to create a variety of useful microfluidic devices.

The Static Meniscus The shape of a liquid drop on a flat substrate may be described by some function $h(x, y)$ giving the height of the free surface above the substrate at each location (x, y). The function $h(x, y)$ has a finite support, the region D, which is the wetted area on the substrate. The equilibrium shape is the function $h(x, y)$ that minimizes the total energy of the droplet

$$E[h(x, y)] = E_{lv} + E_{ls} + E_{sv} + E_g \tag{25}$$

subject to the constraint of a fixed liquid volume, V

$$\int_D h\,\mathrm{d}S = V\,. \tag{26}$$

Here $\mathrm{d}S = \sqrt{1 + h_x^2 + h_y^2}\,\mathrm{d}x\mathrm{d}y$ is a surface element on the free surface (suffixes denote partial derivatives). The interaction energies per unit area at the liquid-vapor, liquid-solid and solid-vapor interfaces are denoted by E_{lv},

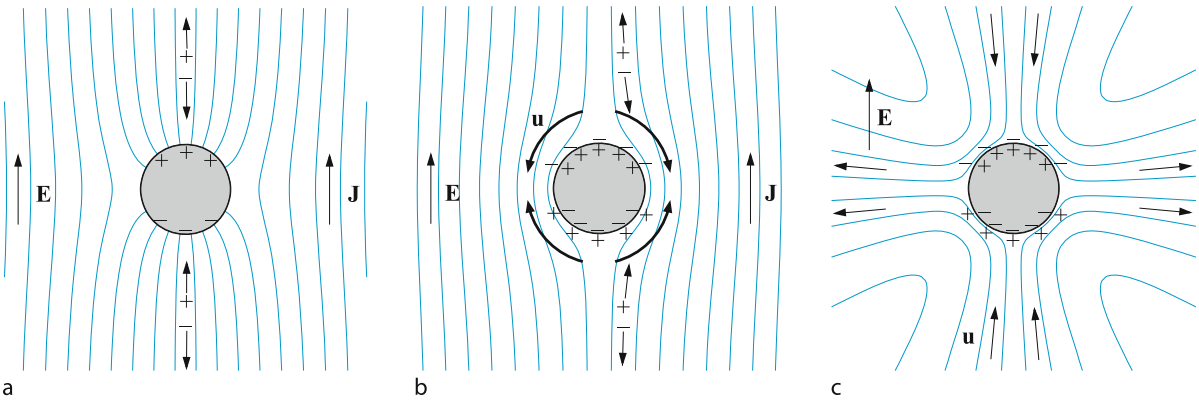

Microfluidics, Figure 5
Illustration of the principle of operation of ICEO: Lines of electric field E around a cylindrical metal wire in an electrolyte before (a) and after (b) double-layer charging in response to a suddenly applied dc field and the resulting ICEO streamlines (c). (Image: courtesy of Prof. T.M. Squires of UCSB, also see reference [30])

E_{ls} and E_{sv} and E_g is the gravitational potential energy. These quantities are functionals of the shape function $h(x, y)$ and may be written as follows:

$$E_{lv} = \int_D \gamma_{lv} dS, \qquad E_{ls} = \int_D \gamma_{ls} dx dy \tag{27}$$

$$E_{sv} = \int_{D^c} \gamma_{sv} dx dy, \qquad E_g = \frac{\rho g}{2} \int_D h dS \tag{28}$$

where D^c denotes the region outside of D. The constraint Eq. (26) is enforced by introducing the Lagrange multiplier, p, and minimizing the auxiliary functional

$$F[h(x, y)] = E[h(x, y)] - p \int_D h dS \tag{29}$$

The problem of determining the equilibrium shape then reduces to finding the shape for the domain D and the function $h(x, y)$ that minimizes $F[h(x, y)]$. The Lagrange multiplier actually corresponds to the pressure within the drop and F is the free energy. Usually, γ_{lv} is a constant, for a patterned substrate γ_{ls} is a periodic function of x and y. The Euler Lagrange equation for the variational problem formulated above can be written down but analytical solutions cannot be found easily. Numerical solutions can be readily constructed with tools such as the Surface Evolver [34]. Figure 6 shows the drop shape on a substrate with a striped pattern of γ_{ls} determined from a Lattice Boltzmann simulation [33].

An alternative to imprinting a periodic pattern of surface energy on a substrate is to change its texture by sculpting it into a series of peaks and valleys as shown in Fig. 8. The liquid drop in this case has a true minimum energy

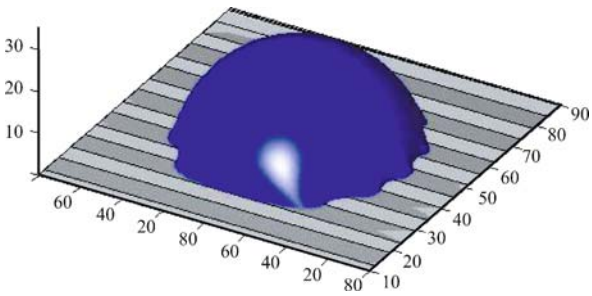

Microfluidics, Figure 6
Computed equilibrium shape of a liquid droplet on a patterned substrate consisting of alternating hydrophobic and hydrophilic stripes (Image: reproduced with permission from [33])

state corresponding to complete wetting of the region between the grooves as well as a metastable state where the drop only makes contact with the tops of the ridges (the analogy is sometimes made to a "fakir" walking on a bed of nails). In order to go from the metastable to the true energy minimum the drop will need to go through intermediate states of greatly increased surface area (energy). If an external disturbance large enough to overcome this energy barrier is not available, the drop remains in the metastable state where the "effective" contact angle is much larger than what it would be in the absence of the texture. Contact angles very close to 180 degrees can be generated in this way, and such surfaces are called "super-hydrophobic". The leaves of certain plants (such as the lotus) naturally have a micro-textured surface and exhibit super-hydrophobicity (Fig. 8). Super-hydrophobic surfaces have many applications due to its water repellent properties.

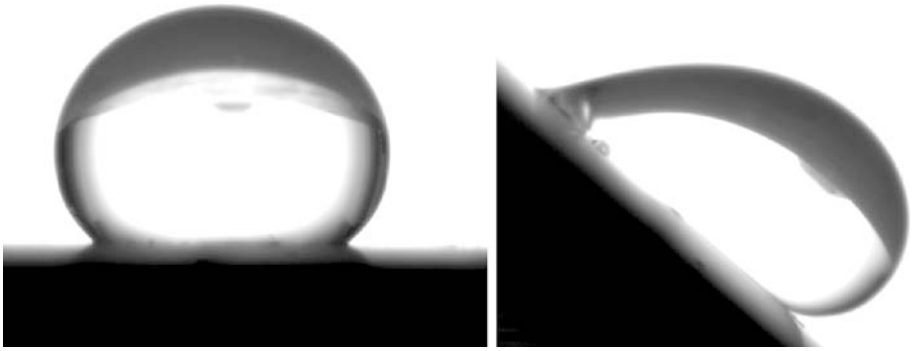

Microfluidics, Figure 7
A static drop (*left*) and a drop sliding down (*right*) an inclined plane. In the latter case the advancing contact angle (θ_a) at the front of the drop exceeds the receding contact angle (θ_r) at the rear. (Image: courtesy of Prof. A. Amirfazli of University of Alberta)

Moving Contact Lines The mathematical description of a drop moving on a substrate is considerably more complicated. Figure 7 shows a static drop on a level surface and the same drop sliding down a tilted surface. Unlike the static drop which has a symmetric shape and a contact angle θ_s given by the Young–Dupré condition

$$\gamma_{lv} \cos \theta_s = \gamma_{sv} - \gamma_{ls} \,, \tag{30}$$

the moving drop is asymmetric. The contact angle at the front of the drop, θ_a, which is called the advancing contact angle exceeds θ_s whereas the contact angle at the rear, θ_r, called the receding contact angle, is less than the equilibrium contact angle θ_s. When U is very close to zero, the contact line moves in a 'stick-slip' fashion so that smooth curves of θ_a and θ_r as a function of the speed U can no longer be drawn. This phenomenon is called contact line pinning. A consequence of contact line pinning is the phenomenon of "contact angle hysteresis". Contact angle hysteresis can be demonstrated by the following experiment with the inclined plane. If one measures the sliding speed of the drop (U) at different inclination angles (α) of the inclined plane one obtains a graph of the function $\alpha(U)$. Extrapolating to $U \to 0$ we can obtain the critical angle $\alpha_0 = \lim_{U\to 0} \alpha(U)$. One might expect then that α_0 is the angle of the incline at which the drop would first start to slide down. However, if a drop is first placed on a level surface and then the surface is gradually tilted, the drop does not start to slide when $\alpha = \alpha_0$. Instead, sliding first starts at some larger inclination $\alpha = \alpha_* > \alpha_0$. This is known as contact angle hysteresis. Contact angle hysterisis and stick slip can be eliminated for certain molecularly smooth substrates such as those with a monolayer coating. However, for most solid liquid interfaces, they are the norm.

Propulsion Mechanisms The well known phenomenon of capillary action is an example of how a liquid with a free surface can be made to flow by manipulating the contact angle. When a capillary is dipped into the horizontal free surface of a liquid pool, the interfacial energies cause the free surface to develop a curvature. Such curvature immediately causes a local pressure drop by virtue of the Young Laplace equation

$$\Delta p = \gamma \left(\frac{1}{R_1} + \frac{1}{R_2} \right) \tag{31}$$

where Δp is the pressure jump across an interfacial point and R_1 and R_2 are the principal radii of curvature at that point. This sets up a pressure gradient which then causes the liquid to flow and the contact line to move. The same principle is involved in all cases of droplet motion, the difference lies only in the physical mechanism responsible for changes in the local contact angle.

If the contact angle is shallow, very often the system admits an asymptotic reduction where the fluid motion is described in the lubrication limit. In such cases, an evolution equation for the height of the free surface $h(x, y, t)$ can be constructed:

$$\frac{\partial h}{\partial t} = \mathcal{L}[h] \tag{32}$$

where $\mathcal{L}$ is some (possibly nonlinear) differential operator in the variables x and y.

Wettability Gradients If a drop is placed on a chemically treated surface where the static contact angle θ_s varies, pressure differences are created between the fore and aft which drives a flow. An equation of the form (32) was derived by Greenspan [35], who in addition to the smallness

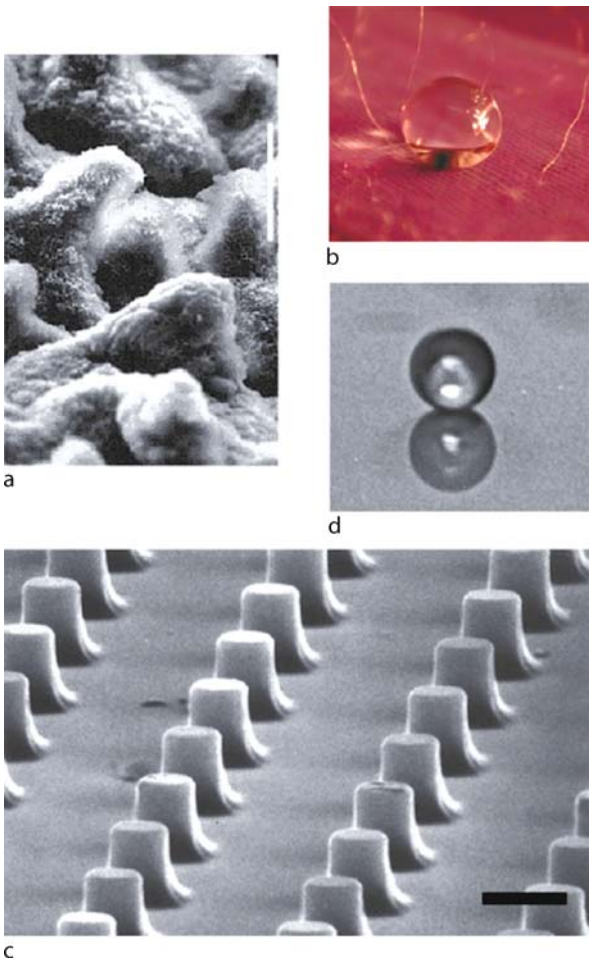

Microfluidics, Figure 8
a Lily leaf showing a rugose coating. SEM, scale bar = 3 μm. b Water droplet on the top of leaves from the South American plant Setcreasea. c Industrial rugose surface of silica. SEM, scale bar = 1 μm. d Water droplet on industrial hydrophobic coatings. Reprinted with permission from *Nature Materials* 4:277–288, 2005

of θ_s assumed that the contact line moves with a velocity $\mathbf{v}_{cl} = \kappa(\theta - \theta_s)\hat{\mathbf{n}}$ where $\hat{\mathbf{n}}$ is the unit normal to the contact line on the plane of the substrate. He further modified the no slip boundary condition as follows:

$$\left[\mathbf{v} - \frac{\alpha}{3h}\frac{\partial \mathbf{v}}{\partial z}\right]_{z=0} = 0\,. \tag{33}$$

The constant α is a dimensional constant chosen to be of the order of 10^{-10} cm^2. Equation (33) reduces to the classical non-slip boundary condition a short distance away from the contact line while reducing to a zero shear stress condition at the contact line itself. This is a way to work around the difficulty of the shear stress singularity on the contact line. Neglecting contact angle hysteresis and disjoining pressure effects Greenspan derived the following equation for $h(x, y, t)$

$$\frac{\partial h}{\partial t} + \frac{\gamma}{3\mu}\nabla \cdot \left[h(h^2 + \alpha)\nabla(\nabla^2 h)\right] = 0 \tag{34}$$

Greenspan applied Eq. (32) to some special situations with insightful results, for example, he showed that if θ_s varies slowly in the x direction, so that the fractional change over the drop length is small, then droplets with a circular wetted area of radius R_0 move with a velocity $U = -\kappa R_0(\mathrm{d}\theta_s/\mathrm{d}x)$, thus it migrates towards the more hydrophillic regions.

Thermocapillary Effects Spatial variations in temperature cause corresponding variations in the coefficient of surface tension γ, which could drive a drop on a substrate. The effect is often seen in the kitchen, heating a little oil in a non-stick pan causes a dry spot to appear at the center with the contact line moving radially out towards the cooler edges of the fry pan. An equation similar to (34) can be developed in this case as well. Since electrical heating wires can be embedded in a substrate, microfluidic devices based on controlled motion of droplets using thermocapillary effects are possible [36].

Marangoni Stresses Marangoni stresses arise when spatial variations in surface tension is set up due to the presence of a trace chemical such as a surfactant. The classic example of motion generated by Marangoni stresses is the toy called the camphor boat. A receptacle at the rear of the tiny 'boat' holds the camphor. The difference in concentration of the camphor between the fore and aft of the boat creates a surface tension gradient which propels the boat. Marangoni stresses can be exploited in microfluidics by using a chemical coating on the substrate so that its wettability can be modulated by shining UV light. Droplets can then be directed to move in a specific pattern by spatial modulation of the UV light on a substrate [36].

Electrowetting The contact angle of a liquid drop can be modified by applying an electric voltage. In early applications the drop was placed directly on top of the electrode. This however has the disadvantage of causing electrolysis and the formation of gas bubbles within the liquid drop. In modern applications, the drop is placed on a metal electrode coated with a thin layer of a dielectric material. This set up is called Electrowetting on Dielectric (EWOD). When a voltage V is applied to the plate, the equilibrium shape of the droplet will change. The new shape can be obtained by repeating the calculation sketched in Sub-

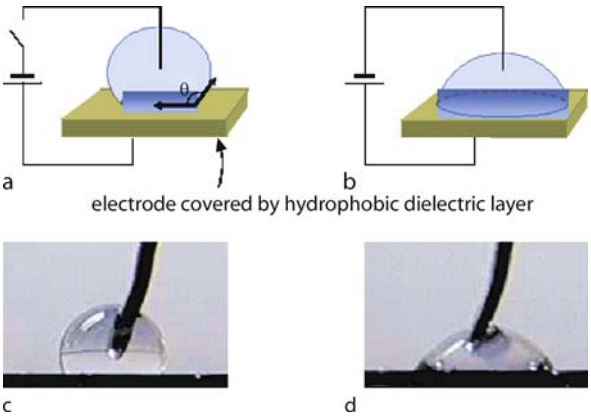

Microfluidics, Figure 9
Illustrating the physical phenomenon of Electrowetting on a dielectric or EWOD (Image: courtesy of Prof. J. Loo, UCLA)

sect. "The Static Meniscus", except that the electrostatic energy must be included in the minimization of the total energy. Such a calculation shows that the new contact angle θ can be related to the old one θ_0 through the relation

$$\cos\theta = \cos\theta_0 + \frac{\epsilon}{8\pi d\gamma_{\mathrm{lv}}} V^2 \,. \tag{35}$$

Here ε is the dielectric constant of the insulator and d is its thickness.

An interesting application of EWOD is the variable focus liquid lens [37] shown in Fig. 10. Application of a small voltage (a few volts) can alter the contact angle and thereby change the lens shape. Thus, variable focal length and focusing control can be achieved with no moving parts. If the lens is sufficiently small (Bond number much less than unity) shape changes due to changes in orientation is negligible because gravity becomes insignificant. The liquid lens is of course ubiquitous in the natural world; our eyes are made of soft materials (liquids and gels) though shape changes are brought about by pressure rather than electrostatic actuation. It is interesting to observe that one of the earliest microscopes invented by the Dutch inventor Anton van Leeuwenhoek (1623–1723) used a small water drop placed inside a hole in a brass plate as the lens.

The EWOD set up is being applied to implement early versions of the "Lab on a Chip". An array of individually addressable electrodes under a dielectric layer is used to move water drops along prescribed paths. By controlling the electrode voltages with electronics, a drop can be divided into daughter droplets or two drops (perhaps containing two different chemicals) can be merged and the contents stirred to create a microscopic reaction chamber. A great advantage of the EWOD set up, is that large voltages (and the consequent bulky power supplies) are not needed. Thus, commercial products using EWOD are likely to be more portable. One shortcoming is that evaporative loss from small drops can be significant. Figure 11 shows shape oscillations of a drop in response to an ac voltage. Such oscillations can be used to homogenize the contents of the liquid drop.

Future Directions

Since "necessity is the mother of invention" there is no doubt that the field of microfluidics will see rapid growth

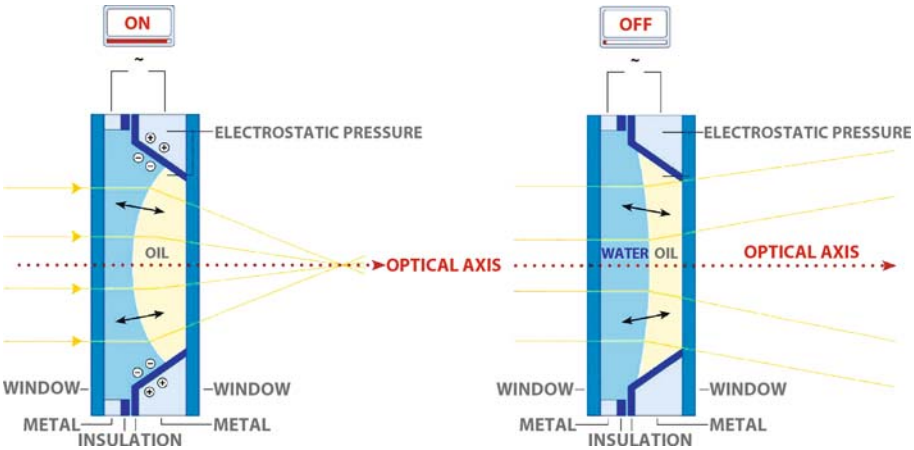

Microfluidics, Figure 10
The liquid lens based on electrowetting by *Varioptic*: a water drop is deposited on a substrate made of metal, covered by a thin insulating layer. The voltage applied to the substrate modifies the contact angle of the liquid drop. The liquid lens uses two isodensity liquids, one is an insulator while the other is a conductor. The variation of voltage leads to a change of curvature of the liquid-liquid interface, which in turn leads to a change of the focal length of the lens. (Image courtesy of Dr. B. Berge of *Varioptic*)

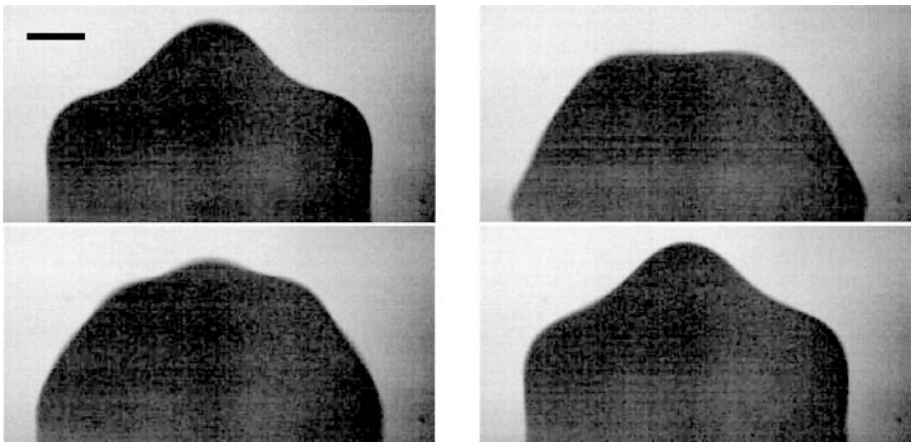

Microfluidics, Figure 11
Sequence of images showing a 7 μL KCl droplet oscillating in response to a 60 V AC forcing at 180 Hz. Images are 1 ms apart and scale bar corresponds to 0.5 mm. Such oscillations may be used to mix the contents after reagent A and B are brought together by a pair of coalescing daughter droplets (Image: courtesy of R. Miraghaie, J. D. Sterling and A. Nadim, the Keck Graduate Institute of Applied Life Sciences. See also reference [38])

in the near future fueled by the demands of the biosciences. The direction this growth might take is however much more difficult to predict.

The field of microelectronics has grown from the 1947 Bell Labs Transistor to chips that run the Blackberrys and iPods of today. Nevertheless, the basic elements of the technology such as the use of silicon wafers, thin films, photoresist and etching have not changed but have been progressively refined and improved. The situation is very different in the case of microfluidics. Let us first take the case of materials. Initially glass and silicon wafers were the only platforms being explored. But then soft materials such as PDMS came into the picture [39,40]. Soft materials have many advantages, for example chips can be mass produced by stamping from a master copy, the flexibility of the material can be exploited for pneumatic pumping [41], etc. However, they do have some drawbacks, for example the zeta potential is hard to control and electrokinetic pumping is not very efficient. It is unclear at this time whether a unique platform will emerge for microfluidics and if so, which one it is going to be.

Consider next the methods of moving fluids around. Electrokinetic methods have the advantage that the voltage needed to achieve a certain flow speed is independent of the channel diameter. By contrast, for pressure driven flows, the pressure head needed scales inversely as the cross-sectional area. Furthermore, electrokinetic flows have a uniform rather than a parabolic profile and therefore cause less Taylor dispersion. On the other hand, electrokinetic flows are very sensitive to surface contamination, a serious difficulty when dealing with proteins and peptides. Currently both pressure driven and electrokinetic approaches are being explored. One disadvantage of both of these approaches is that while the microfluidic chip itself is small and portable, the power source (compressed air for pressure driven flows or kilovolt power sources for electrokinetic ones) is usually quite bulky. Two approaches that seek to overcome this difficulty are pumping based on ICEO [42] and the approach of EWOD which only require a few volts to move droplets around on a substrate. These approaches have other drawbacks of their own, for example small droplets tend to evaporate and ICEO involves the added complexity of patterned electrodes on one or more of the substrates. It still remains to be seen whether one among these very different approaches would emerge as the winner or if they would continue to co-exist as the subject matures.

Bibliography

Primary Literature

1. Bogy D (1979) Drop formation in a circular liquid jet. Annu Rev Fluid Mech 11:207–228
2. Jacobson S, Ramsey J (1996) Integrated microdevice for DNA restriction fragment analysis. Anal Chem 68(5):720–723
3. Parker AR, Lawrence CR (2001) Water capture by a desert beetle. Nature 414:33–34
4. Eisner T, Aneshansley D (1999) Spray aiming in the bombardier beetle: Photographic evidence. Proc Natl Acad Sci, USA 97059709
5. Bush J, Hu D (2005) Walking on water: Biolocomotion at the interface. Annu Rev Fluid Mech 11:207–228
6. Muntz E (1989) Molecular gas dynamics. Annu Rev Fluid Mech 21:387–422

7. Dussan E (1979) On the spreading of liquids on solid surfaces: Static and dynamic contact lines. Annu Rev Fluid Mech 11:371–400
8. Kogan M (1973) Molecular gas dynamics. Annu Rev Fluid Mech 5:383–404
9. Sone Y (2000) Molecular gas dynamics. Annu Rev Fluid Mech 32:779–811
10. Lauga E, Brenner MP, Stone HA (2006) Microfluidics: The no-slip boundary condition. Springer, Heidelberg
11. Lauga E, Stone H (2003) Effective slip in pressure-driven stokes flow. J Fluid Mech 489:55–77
12. Feynmann R, Leighton RB, Sands M (1970) The Feynmann Lectures on Physics, vol 2. Addison-Wesley, Menlo Park
13. Landau L, Lifshitz E (2002) Course of theoretical physics, vol 10. Physical Kinetics. Butterworth-Heinenann, Oxford
14. Lambert W, Middleton D (1990) pH hysterisis effect with silica capillaries in capillary zone electrophoresis. Anal Chem 62:1585–1587
15. Kirby B, Hasselbrink E (2004) Zeta potential of microfluidic substrates: 1. theory, experimental techniques, and effects on separations. J Electrophoresis 25(2):187–202
16. Kirby B, Hasselbrink E (2004) Zeta potential of microfluidic substrates: 2. data for polymers. J Electrophoresis 25(2):203–213
17. Gouy G (1910) Sur la constitution de la électrique à la surface d'un electrolyte. J Phys Radium 9:457–468
18. Chapman D (1913) A contribution to the theory of electroencapillarity. Phil Mag 25(6):475–481
19. van Dyke M (1975) Perturbation Methods in Fluid Mechanics. The Parabolic Press, Stanford
20. Helmholtz H (1879) Studien über electrische Grenzschichten. Ann der Physik und Chemie 7:337–387
21. Smoluchowski M (1903) Contribution à la théorie de l'endosmose électrique et de quelques phenoménes corrélatifs. Bull Int de l'Academie des Sciences de Cracovie 8:182–200
22. Ghosal S (2002) Lubrication theory for electroosmotic flow in a microfluidic channel of slowly varying cross-section and wall charge. J Fluid Mech 459:103–128
23. Anderson J, Idol W (1985) Electroosmosis through pores with nonuniformly charged walls. Chem Eng Commun 38:93–106
24. Rice C, Whitehead R (1965) Electrokinetic flow in a narrow cylindrical capillary. J Phys Chem 69:4017–4024
25. Burgreen D, Nakache F (1964) Electrokinetic flow in ultrafine capillary slits. J Phys Chem 68(5):1084–1091
26. Saville D (1977) Electrokinetic effects with small particles. Annu Rev Fluid Mech 9:321–337
27. Morrison Jr F (1970) Electrophoresis of a particle of arbitrary shape. J Coll Int Sci 34(210-214):45–54
28. Anderson J (1985) Effect of nonuniform zeta potential on particle movement in electric fields. J Coll Int Sci 105(1):45–54
29. Cummings E, Singh A (2003) Dielectrophoresis in microchips containing arrays of insulating posts: Theoretical and experimental results. Anal Chem 75:4724–4731
30. Bazant M, Squires T (2004) Induced charge electrokinetic phenomena: theory and microfluidic applications. Phys Rev Lett 92(6):066101
31. Ramos A, Morgan H, Green N, Castellanos A (1998) AC electrokinetics: A review of forces in microelectrode structures. J Phys D 31:2338–2353
32. Ajdari A (2000) Pumping liquids using asymmetric electrode arrays. Phys Rev E 61:R45–48
33. Dupuis A, Yeomans J (2004) Lattice Boltzmann modelling of droplets on chemically heterogeneous surfaces. Future Generation Computer Systems 20(6):993–1001
34. Brakke K (1996) The surface evolver and the stability of liquid surfaces. Phil Trans R Soc A 354:2143–2157
35. Greenspan H (1978) On the motion of a small viscous droplet that wets a surface. J Fluid Mech 84:125–143
36. Troian S (2005) Principles of microfluidic actuation by modulation of surface stresses. Annu Rev Fluid Mech 37:425–455
37. Berge B (2005) Liquid lens technology: Principle of electrowetting based lenses and applications to imaging. In: 18th IEEE International Conference on Micro Electro Mechanical Systems Technical Digest, IEEE. Miami Beach, Florida, pp 227–230
38. Miraghaie R, Sterling JD, Nadim A (2006) Shape oscillation and internal mixing in sessile liquid drops using electrowetting-on-dielectric (ewod). In: Technical Proceedings of the 2006 NSTI Nanotechnology Conference and Trade Show, vol 2. pp 610–613
39. Anderson JR, Chiu D, Jackman R, Cherniavskaya O, McDonald J, Wu H, Whitesides S, Whitesides G (2000) Fabrication of topologically complex three-dimensional microfluidic systems in pdms by rapid prototyping. Anal Chem 72:3158–3164
40. Anderson J, McDonald J, Stone H, Whitesides G (preprint) Integrated components in microfluidic devices in pdms: A biomimetic check valve, pressure sensor & reciprocating pump
41. Whitesides G, Stroock A (2000) Flexible methods for microfluidics. Phys Today 54(6):42–48
42. Urbanski J, Thorsen T, Leviatan J, Bazant M (2006) Fast ac electro-osmotic micropumps with nonplanar electrodes. Appl Phys Lett 89:143508

Books and Reviews

Berthier J, Silberzan P (2006) Microfluidics for Biotechnology (Microelectromechanical Systems). Artech Press, Norwood

Ghiotto A (2002) Fundamentals of Microfabrication: The Science of Miniaturization. CRC Press, Boca Raton

Ghosal S (2006) Electrokinetic flow and dispersion in Capillary Electrophoresis. Annu Rev Fluid Mech 38:309–338

Nijhoff M (1983) Low Reynolds number hydrodynamics. Kluwer, Dordrecht

Karniadakis G, Beskok A, Aluru N (2005) Microflows and Nanoflows: Fundamentals and Simulation. Interdisciplinary Applied Mathematics. Springer, Heidelberg

Mugele F, Baret J (2005) Electrowetting: from basics to applications. J Phys: Condens Matter 17:R705–R774

Pozrikidis C (1992) Boundary Integral and Singularity Methods for Linearized Viscous Flow. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge

Squires TM, Quake SR (2005) Microfluidics: Fluid physics at the nanoliter scale. Rev Mod Phys 77:977–1026

Stone HA, Stroock AD, Ajdari A (2004) Engineering flows in small devices: Microfluidics toward a Lab-on-a-Chip. Annu Rev Fluid Mech 36:381–411

Tabeling P (2006) Introduction to Microfluidics. Cambridge University Press, New York (translated by Suelin Chen)

Troian S (2005) Principles of microfluidic actuation by modulation of surface stresses. Annu Rev Fluid Mech 37:425–455

Minority Games

Chi Ho Yeung[1,2,3] , Yi-Cheng Zhang[1,2,3]
[1] Department of Physics, The Hong Kong University of Science and Technology, Hong Kong, China
[2] Département de Physique, Université de Fribourg, Pérolles, Fribourg, Switzerland
[3] School of Management, University of Electronic Science and Technology of China (UESTC), Chengdu, China

Article Outline

Glossary
Definition of the Subject
Introduction
The Minority Game
The Physical Properties of the Minority Game
Variants of the Minority Game
Analytic Approaches to the Minority Game
Minority Games and Financial Markets
Future Directions
Bibliography

Glossary

Cumulated payoff It refers to the reward accumulation or the score counting for all the individual strategies being held by the agents in the game. Each strategy has its own value of cumulated payoff. When a strategy gives a winning or losing prediction for the next round of the game (no matter whether the agents follow this prediction), scores are added to or deducted from this strategy respectively. It is also known as virtual point or virtual score of the strategies. The way of rewarding or penalizing is known as the payoff function.

Attendance In the contexts of the Minority Game, attendance refers to the collective sum of all agents' actions at each round of the game. For the ordinary games, it is equal to the difference in the number of agents in choosing the two different choices or actions (in early formulation, it is equal to the number of agents in choosing one of the particular choice). The terminology "attendance" origins from the ancestor form of the Minority Game, the *El Farol* Bar problem by W.B. Arthur, in which agents choose whether to attend a bar at every round of the game.

Volatility Volatility in Minority Games is the time average variance of the attendance after each round of the game. It is an inverse measure of the efficiency of resource distribution in the game. A high volatility corresponds to large fluctuations in attendance and hence an inefficient game. A low volatility corresponds to smaller fluctuations in attendance and hence an efficient game.

Predictability It is an important macroscopic measurable in the game and also the order parameter which characterizes the major phase transition in the system. It is usually denoted by H which is a measure of non-uniform probabilistic outcome of the attendance given a certain information provided to the agents. A higher predictability refers to the case where attendance tends to be positive or negative for a certain piece of information, which makes the game outcome more predictable.

Symmetric and asymmetric phase The two important phases of the system. The symmetric phase is also known as crowded phase or the unpredictable phase. The asymmetric phase is also known as uncrowded phase, the dilute phase or the predictable phase. The system's behaviors, dynamics and characteristics are different in these two phases. The two phases are characterized by the order parameter, called the predictability H.

Endogenous and exogenous games Endogenous games refer to games which utilize the past winning history to generate signals or information for agents to make decisions in the next round. Endogenous games are also known as games with real history. Exogenous games refer to games which utilize random signals or information for agents to make decisions in the next round. Exogenous games are also known as games with random history, or external information.

On-line update and batch update On-line update refers to the evaluation of payoffs of strategies after each round of the game. Thus, the priority of strategies being employed by an agent may be altered after any round of the game. The exogenous game or random history game sometimes employ the batch update method which refers to the evaluation of payoffs on strategies only after a fixed number of rounds where all the possible signals or information have appeared. In games with ordinary batch update, all the possible signals appear once in each batch before the update of payoffs on strategies, the order of appearance of signals in each batch is thus irrelevant.

Definition of the Subject

The Minority Game (MG) refers to the simple adaptive multi-agent model of financial markets with the original formulation introduced by Challet and Zhang in 1997. In

this model of repeated games, agents choose between one of the two decisions at each round of the game, using their own simple inductive strategies. At each round, the minority group of agents win the game and rewards are given to those strategies that predict the winning side. Daily examples of minority game include drivers choosing a less crowded road or people choosing a less crowded restaurant. Unlike most economics models or theories that assume investors are deductive in nature, a trial-and-error inductive thinking approach is implicitly implemented in the process of decision making when agents choose their choices in the games. In this original formulation, the history or the information given to agents is a string of binary bits that is composed of the winning sides in the past few rounds.

While the original model is simple, many variants of the model were later introduced. In some other contexts and later literature, the term "Minority Games" is sometimes referred to as a class of multi-agent models that contains all the variants of the original Minority Game. Most of the models in this class of game share the principal features that agents are inductive in nature. Thus, strategies with accumulated virtual score are usually present in this set of models. As a result, the original formulation of the Minority Game by Challet and Zhang in 1997 is sometimes referred to as the "original Minority Game" or the "basic Minority Game".

While investigating economical dynamics, physicists found most of the economics models to be deductive in nature. Since investors have an expectation of the future, economical models conceptually differ from conventional physical models in which variables have only historical dependence. As a result, it becomes difficult for physicists to develop and analyze the traditional financial models, even with well-developed mathematical tools. The *El Farol* bar problem and the Minority Game somehow tackle the problem by assuming investors can be inductive in nature with bounded rationality, in which they predict the future by only examining the past states of the system, similar to the ordinary physics models.

In the physics community, the basic Minority Game and its variants are an interesting and newly established class of complex and disordered systems that contain a large amount of physical aspects. In addition to the modeling purpose of the financial markets, it is also a simple model where the Hamiltonian can be defined and analytic solutions are developed in some regime of the model, from which the model is viewed with a complete physical sense. It is also characterized by a clear two-phase structure with very different collective behaviors in the two phases, as in conventional physical systems. All these physical properties further raise the interest of physicists in understanding and solving the model analytically, using the techniques origin from statistical mechanics. Other than these collective behaviors, physicists are also interested in the dynamics of the games. Periodic attractors, anti-persistence and crowd-anticrowd movement of agents are also observed. In this way, the Minority Game and its variants serve as a useful tool and provide a new direction for physicists in viewing and analyzing the underlying dynamics in the financial markets, and at the same time analytical techniques from statistical physics can be widely applicable.

On the other hand, for modeling purposes, Minority Games serve as a class of simple models that are able to produce some of the macroscopic features being observed in the real financial markets. Such features are usually termed stylized facts that include the fat-tail price return distribution and volatility clustering. Crashes and bubbles are also observed in some of the variants and other models inspired by the Minority Game. The grand-canonical versions of the game suggest the conjecture of financial markets being a critical phenomenon in physics.

Because of the simplicity of the original model, a large freedom is found in modifying the models to make the models more realistic and closer to the real financial markets. Many details in the model can be fine-tuned to imitate the real markets. Minority Games setup a framework of agent-based models where predictability of financial data may be possible. Sophisticated models based on games can be setup and implemented in real trading, which show a great potential over the commonly adopted statistical techniques in financial analysis. As a result, Minority Games having caught the interest of economists, may induce some to switch to employing agent-based models in understanding the underlying mechanism behind socio-economics systems. Minority Games also shift the emphasis of some economists to investigating the formation of price pattern, rather than just performing data analysis of the price pattern.

Introduction

The basic Minority Game was formulated by Damien Challet and Yi-Cheng Zhang in 1997 [1] with their work being published in a statistical physics journal. The model was inspired by the *El Farol* Bar problem introduced by W. Brian Arthur in 1994 [2] with his work being presented and published in an economical meeting and its proceedings. This already shows the interdisciplinary nature of the Minority Game with an economical origin, in a physical perspective. The Minority Game follows the ma-

jor conceptual structure being implemented in the *El Farol* Bar problem, with some modifications on the model structure.

In the original *El Farol* Bar problem, each individual of a population choose whether to attend a bar each Thursday evening. The bar has a limited number of seats and can at most entertain x% of the population. If less than x% of the population go to the bar, the show in the bar is considered to be enjoyable and it is better to attend the bar rather than staying at home. On the other hand, if more than x% of the population go to the bar, all the people in the bar would have an unenjoyable show and staying at home is considered to be a better choice than attending the bar. In order to make decisions on whether to attend the bar, all the individuals are equipped with a certain number of strategies. These strategies provide them with predictions of the attendance in the bar next week, based on the attendance in the past few weeks. All individuals rank their strategies according to their past performance and make decisions by considering the attendance predicted by their own best strategy.

Several changes were made to the model when the Minority Game was formulated from the *El Farol* Bar problem. Instead of using the history of past attendance, a string of binary bits which records the past few winning predictions or actions are employed as information. The predictions of the strategies are the winning choices in the next round, with no prediction about the actual size of attendance. Thus, binary information and predictions are implemented, which greatly reduce the dimensional space of the system. In addition, the winning choice is determined by the minority choice (instead of the parameter x in the Bar problem) at every round of the game, hence the two choices are symmetric. Because of the minority rule, the population is restricted to be an odd integer in the original formulation.

These modifications of binary and symmetric actions make the model more accessible for the physics community. The first publication of the Minority Game led to great interest among some statistical physicists who began researching the Minority Game and formulating variants. Some physicists began to identify the study of such a class of models as within the field of econophysics. In 1999, R. Savit et al. [3] published their work on the analysis of the Minority Game, which is crucial to subsequent theoretical developments of the game. They discovered an important control parameter α, which is defined as the ratio of the total amount of possible information to the population size. It rescales the macroscopic observables of the game for different amounts of information and population size. A phase transition is observed at the critical value of α which separate the two phases, namely the symmetric phase and the asymmetric phase.

After discovery of the rescaling properties and the phase transition of the Minority Game, great efforts were put into solving the model analytically, using well-developed techniques in the field of statistical physics [4,5,6,7,8,9,10,11,12,13,14,15,16,17]. In order to solve the model, the basic Minority Game is sometimes modified to increase the feasibility of the analytical approach. In some variants of the Minority Game, the model is simplified to preserve only the major dynamical behaviors while in some other variants, features are added to the game that make the model more comparable to traditional physical models. As a result, a large number of variants of the Minority Game were produced during attempts at analytic description.

On the other hand, attempts were also made to make the models more comparable to real financial markets. Some physicists and even economists modified the basic model by adding more features from real markets [18,19,20,21,22,23,24,25,26,27,28]. Stylized facts are found in the critical regime of the grand-canonical version [23,24,25,26,27,28] of the Minority Game in which agents can choose to refrain from participation in the game. This suggests the conjecture of financial markets being in critical state and further pushes the development of the model in this direction. Some new models have also been developed to include more financial aspects. Efforts are made to have a better understanding of the market through the agent-based approach. The macroscopic observations from the models have become more realistic but at the same time, the models have become more sophisticated. Because of the analytic goal of solving the model in a physical sense and for modeling purposes, there are a vast number of variants of the Minority Game; it therefore constitutes a class of models.

In the following sections, we briefly describe the formulation of the basic model and its variants, and briefly introduce the physical properties, the analytic approaches of the model and its link with financial markets. We review the formulation of the basic Minority Game in Sect. "The Minority Game". Some major physical properties of the basic Minority Game are given in Sect. "The Physical Properties of the Minority Game", the effect of temperature is also discussed which was originally introduced in the Thermal Minority Game (TMG). In Sect. "Variants of the Minority Game", we review briefly some important variants of the Minority Game and their physical significance, these include the Evolutionary Minority Game (EMG), the TMG, the Minority Game without information and the Grand-canonical Minority Game

(GCMG) while their corresponding implications for the financial markets and some other variants will be discussed in Sect. "Minority Games and Financial Markets". In Sect. "Analytic Approaches to the Minority Game", we briefly introduce some of the analytic approaches to the Minority Game. In Sect. "Minority Games and Financial Markets", we review some of the financial features produced by Minority Games and their implications. Finally, in Sect. "Future Directions", we describe some of the possible directions for future development of the Minority Game.

Minority Games, Table 1
An example of a strategy with $M = 3$

History	Prediction
000	1
001	0
010	0
100	1
011	0
101	1
110	0
111	0

The Minority Game

The basic Minority Game [1] is defined as follows. We consider a population of N agents competing in repeated games, where N is an odd integer. At each round of the game, each agent has to choose between one of the two actions, namely "0" and "1" (in most of the subsequent literatures, "−1" and "1" instead of "0" and "1" are used as the actions, we shall keep the following discussions using the actions "−1" and "1"), which can also be interpreted as the"sell" and "buy" actions. These actions are sometimes called the bid and is denoted by $a_i(t)$, corresponding to the bid of agent i at time t. The minority choices win the game at that round and all the winning agents are rewarded.

Before the game starts, every agent draws S strategies from a strategy pool which help them to make decisions throughout the game. There is no a priori best strategy. These strategies can be visualized in the form of tables where each strategy contains a "history column" (or "signal" column) and a "prediction column". Each row of the history column is a string of M bits, which represents the *history* of the past winning actions in the previous M steps, which is also known as *signal* or *information*. The history is evolving with time and is usually denoted by $\mu(t)$. The parameter M is sometimes known as the *brain size* or the *memory* of the agents. An example of a strategy with $M = 3$ is given in Table 1. For games with memory M, the total number of possible signals is 2^M and thus the total number of possible strategies in the strategy pool is 2^{2^M}. We note that for even a relatively small M, such as $M = 5$, the total number of possible strategies is already huge.

As shown in the strategy in Table 1, a history of "110" corresponds to the case where the past three winning actions are "1", "1" and "0", and the corresponding prediction of winning choice for the next round is "0". Strategies can be conveniently represented by P-dimensional vectors that record only the P predictions, where $P = 2^M$. If the strategy gives a correct prediction on the winning choice, one point is awarded to the strategy. All the S strategies of an agent have to predict at every round of the game, and points are given to those strategies (no matter whether they are being selected by the agent to make real actions) that give correct predictions. The scores of all the strategies are accumulated which are thus known as the *virtual points*, *virtual scores* or the *cumulated payoffs* of the strategies. These scores start at zero in the basic Minority Game. At every round of the game, agents make their decisions according to the strategy with the highest virtual score at that particular moment. If there is more than one strategy with the highest score, one of these strategies is randomly employed. Agents themselves who make the winning decisions are also rewarded with points, and these are called the *real points* of the agents (to be distinguished from the *virtual points* of the strategies).

More explicitly, we define the *attendance* $A(t)$ as the collective sum of actions from all agents at time t. If we denote the prediction of strategy s of agent i under the information $\mu(t)$ to be $a_{i,s}^{\mu(t)}$ at time t, which can be either "−1" or "1", each strategy can be represented by a P-dimensional vector $\vec{a}_{i,s}$ where all the entries are either "−1" or "1". The attendance $A(t)$ can then be expressed as

$$A(t) = \sum_{i=1}^{N} a_{i,s_i(t)}^{\mu(t)} = \sum_{i=1}^{N} a_i(t) \tag{1}$$

where $s_i(t)$ denotes the best strategy of agent i at time t, i. e.,

$$s_i(t) = \arg\max_s U_{i,s}(t) \tag{2}$$

and $a_i(t)$ denotes the real actions or so-called the *bids* of the agents, i. e.,

$$a_i(t) = a_{i,s_i(t)}^{\mu(t)} \,. \tag{3}$$

With this $A(t)$, the cumulative virtual score or payoff $U_{i,s}$ of the strategy s of agent i can be updated by

$$U_{i,s}(t+1) = U_{i,s}(t) - \mathrm{sign}\left[a_{i,s}^{\mu(t)} A(t)\right] \tag{4}$$

where sign(x) is the sign function (in some literatures where "0" and "1" are employed as actions, the last term in Eq. (4) becomes $-\text{sign}[(2a_{i,s}^{\mu(t)}-1)A(t)]$). Here one point is added or deducted from the strategies which give a correct or wrong prediction respectively, and this is usually called the *step payoff* scheme. We note that the negative sign in Eq. (4) corresponds to the minority nature of the game, i. e., when $a_{i,s}^{\mu(t)}$ and $A(t)$ are of opposite signs, a point is added to the strategy. The real gain of agent i at time t is $-\text{sign}\left[a_i(t)A(t)\right]$.

Thus, every agent is considered to be adaptive, they can choose between their s strategies and the relative preference of using strategies changes with time and is adaptive to the market outcomes. They are also considered to be inductive, making their decisions according to the best choice they are aware of, with their limited number of strategies, but not the global best choice given by all possible strategies (the one with the highest virtual score among the entire strategy space). The game is also a self-contained model, in which the agents make individual actions according to the history, individual actions are then summed up to give the history for the next round which is then used by the agents to make predictions again.

As the total number of agents N in the game is an odd integer, the minority side can always be determined and the number of winners is always less than the number of losers, implying the Minority Game to be a *negative sum* game. Because of the minority nature of the game and as the two actions are in symmetric, the time average of $A(t)$ always has a value of 0 (or $\langle A(t)\rangle = N/2$ if "0" and "1" are employed as actions). Hence, instead of the average attendance, one may be more interested in the fluctuations of attendance around the average values and this turns out to be one of the most important macroscopic observables in the subsequent development. We denote σ^2 to be the variance of attendance, also known as the *volatility*, given by

$$\sigma^2 = \left\langle A^2\right\rangle - \langle A\rangle^2 \tag{5}$$

with $\langle A\rangle = 0$ for games with actions "-1" and "1". σ^2 is an inverse measure of the market efficiency in the game. We consider two extreme cases of game outcomes. For the first one, there is only one agent choosing one side while all the others choose the opposite side. There is a single winner and $N-1$ losers which is considered to be highly inefficient in the sense of resource allocation, and the supply and demand are highly unbalanced. For the second case, $(N-1)/2$ of the agents choose one side while $(N+1)/2$ of the agents choose the opposite side. There are $(N-1)/2$ winners and the supply and demand is maximally balanced. Thus, one may expect to minimize fluctuations of attendance for advantages of agents as a whole.

Some simplifications or modifications to the basic Minority Game are suggested and employed in later literature, where the major physical features of the models are preserved. A. Cavagna [29,30] observed that the variance of attendance is almost unaffected if the history string is replaced by a random *invented* string, provided that all agents receive the same string at the same time. That is, instead of their self-generated winning history, they react to a virtual random information that is completely unrelated to the previous winning groups. In this case, the signal strings are usually called *information* instead of *history*. The games which feedback the real history are known as *endogenous* games, while those games that employ random history are called *exogenous* games. In exogenous games, the total number of signals is no longer restricted to 2^M, instead it can be any integer which is usually denoted by P, and is sometimes known as the complexity of information. Every random signal or information appears with a probability of $1/P$. These random pieces of information make the dynamics of the game more stochastic, which is very advantageous for analytic approaches. In the endogenous game, $P = 2^M$.

As "-1" and "1" are employed as actions, instead of adding or deducting one virtual point to the strategies, the cumulated payoff $U_{i,s}$ can be updated by the following equation with a *linear payoff* scheme

$$U_{i,s}(t+1) = U_{i,s}(t) - a_{i,s}^{\mu(t)}A(t) \tag{6}$$

where $A(t)$ is the attendance given by Eq. (1). A factor of $1/N$, $1/\sqrt{N}$ or $1/P$ is always employed to rescale the last term. While the real gain for agent i is $-a_i(t)A(t)$, the total gain for all the agents is $\sum_i -a_i(t)A(t) = -A^2(t)$, preserving the negative sum nature of the game. This modification is important for analysis while the qualitative behaviors of the game are preserved. It also has the meaning of having a higher reward or larger penalty if a smaller minority or a larger majority group is predicted, respectively.

In the original game, the ordinary strategy space has a size of 2^{2^M}. Challet et al. [31] showed that a *reduced strategy space* (RSS) can be employed in which the qualitative behaviors of the game and the numerical values of variance are not largely affected. We first construct a set of 2^M uncorrelated strategies, in which every two strategies of the set have exactly half of the predictions different. The reduced strategy space is then formed by combining this set with the set of their anti-correlated strategies, in which every strategy in the latter set have exactly the opposite prediction to their anti-partners in the former set. Hence,

the size of the reduced strategy space is $2^M \cdot 2 = 2^{M+1}$. This virtue of the reduced strategy space simplifies the theoretical analysis in crowd-anticrowd approaches [4,5,6]. Although the dimension of the reduced strategy space is highly reduced, the ordinary strategy space is commonly employed in numerical simulations.

The Physical Properties of the Minority Game

Parameters introduced into the basic Minority Game include N, M (or P) and S corresponding to the population size, the memory of the agents (the complexity or the total amount of possible information) and the number of strategies that each agent holds. The predictions $a^{\mu}_{i,s}$ of strategies are fixed for every agent throughout the game and are considered to be quenched disorders of the system in physics. The cumulated payoffs of strategies that evolve with time are considered to be dynamic variables or annealed variables of the system. The game is also a highly frustrated model. Because of the minority nature of the model, frustration results in the fact that not all the agents can be satisfied simultaneously. We focus our discussions on the case of $S = 2$, where cases of larger S (not extensively large) will be briefly discussed and have been shown to share very similar behavior to the case of $S = 2$ [31].

Major Features: Phase Transition, Volatility and Predictability

In 1999, Robert Savit, Radu Manuca and Rick Riolo [3] found that the macroscopic behavior of the system does not depend independently on the parameters N and M, but instead depends on the ratio

$$\alpha \equiv \frac{2^M}{N} = \frac{P}{N} \tag{7}$$

(denoted by z in their original paper) which serves as the most important control parameter in the game. This scaling is also true for $P \neq 2^M$ in exogenous games. The volatility σ^2/N and the predictability H/N (which we are going to define later) for different values of N and M depend only on the ratio α. A plot of σ^2/N against the control parameter α for an endogenous game is shown in Fig. 1. We can see that the graph shows a data collapse of σ^2/N for different values of N and M. The dotted line in Fig. 1 corresponds to the coin-toss limit (random choice limit), in which agents play by making random decisions at every round of the game. This value of volatility in coin-toss limit can be obtained by simply assuming a binomial distribution of agents' actions, with probability 0.5, where $\sigma^2/N = 0.5(1-0.5)\cdot 4 = 1$. When α is small, the volatility of the game is larger than the coin-toss limit which

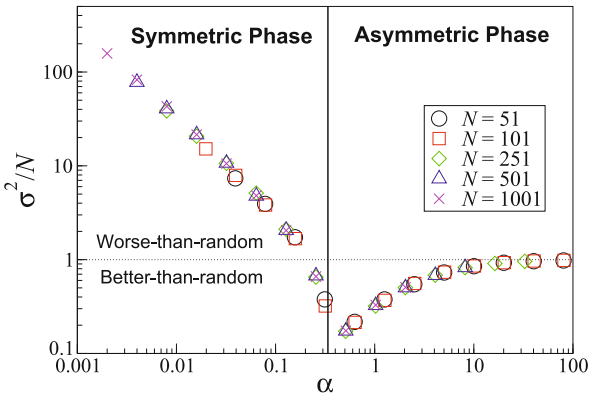

Minority Games, Figure 1
The simulation results of the volatility σ^2/N as a function of the control parameter $\alpha = 2^M/N$ for games with $S = 2$ strategies for each agent averaged over 100 samples. Endogenous information and linear payoff are adopted in these simulations. *Dotted line* shows the value of volatility in random choice limit. *Solid line* shows the critical value of $\alpha = \alpha_c \approx 0.3374$. The resolution of the curve can be improved to show σ^2/N attains minimum at $\alpha \approx \alpha_c$

implies the collective behaviors of agents are worse than the random choices. In early literature, it is known as the *worse-than-random* regime. When α increases, the volatility decreases and enters a region where agents are performing better than the random choices, which is known as the *better-than-random* regime. The volatility reaches a minimum value which is substantially smaller than the coin-toss limit. When α further increases, the volatility increases again and approaches the coin-toss limit.

These results also allow us to identify two phases in the Minority Game, as separated by the minimum of volatility in the graph. The value of α where the rescaled volatility attends its minimum is denoted by α_c, which represents the phase transition point. α_c has a value of 0.3374 ... (for $S = 2$) by analytical calculations [8,14]. Generally, for $\alpha < \alpha_c$, the volatility σ^2 and the spread of volatility for different samples of simulation are proportional to N^2. Beyond the transition point for $\alpha > \alpha_c$, the volatility σ^2 and the spread of volatility are generally proportional to N. These can be recognized by the asymptotic behavior of the graph in Fig. 1 where the slope approaches -1 for $\alpha < \alpha_c$ and approaches 0 for $\alpha > \alpha_c$.

In addition to the different scaling of volatility with N, other quantities also show different behaviors in the two phases. By examining the distributions of winning probabilities for a particular action after different history strings, R. Savit et al. [3] found that these distributions are completely different in the two phases. $P(1|\mu)$ is defined as the conditional probability of action "1" turns out to be the

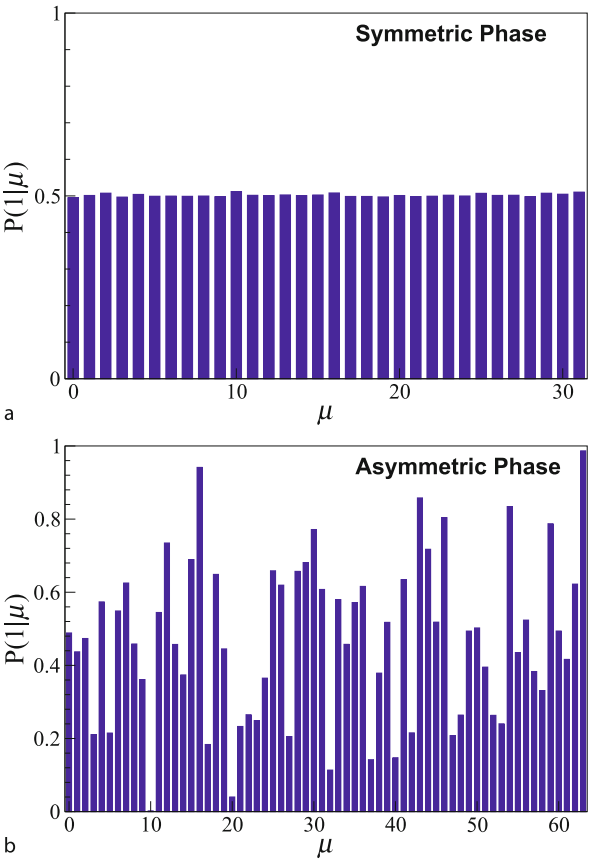

Minority Games, Figure 2

The histogram of the probabilities P(1|μ) of winning action to be "1" given information μ (plotted as the decimal representations of the binary strings of information), for games of $N = 101$ agents and $S = 2$ in a symmetric phase with $M = 5$, i.e., $\alpha \approx 0.316 < \alpha_c$ and b asymmetric phase with $M = 6$, i.e., $\alpha \approx 0.634 > \alpha_c$. Endogenous information and linear payoff are adopted. The histogram in a would have been even more uniform if step payoff was adopted, as shown in the original paper [3]

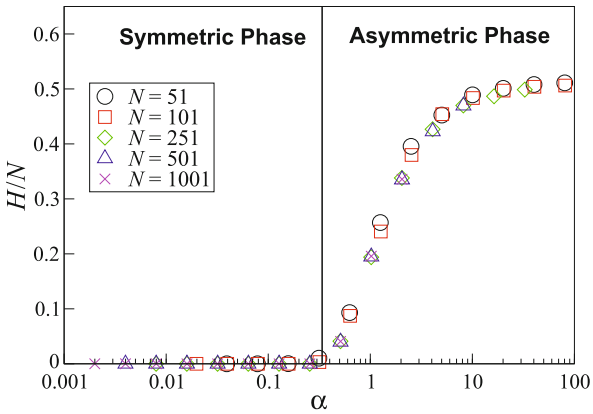

Minority Games, Figure 3

The simulation results of the predictability H as a function of the control parameter $\alpha = 2^M/N$ for games with $S = 2$ strategies for each agent averaged over 100 samples. Endogenous information and linear payoff are adopted in these simulations

minority group after the history or information μ, and the histogram for P(1|μ) is flat at 0.5 for all μ when $\alpha < \alpha_c$, as shown in Fig. 2a. For $\alpha > \alpha_c$ as shown in Fig. 2b, this histogram for P(1|μ) is not flat and uniform. This result is highly important as it implies that below α_c, there is no extractable information from the history string of length M, since the two actions have equal probability of winning (both are 0.5) for any history string. However, beyond the phase transition when $\alpha > \alpha_c$, there is an unequal winning probability of the two actions, by just looking at the past M winning actions of the game. Hence, we can call the phase for $\alpha < \alpha_c$ the *unpredictable* or the *symmetric* phase, as agents cannot predict the winning actions from the past M-bit history (the winning probabilities are symmetric). On the contrary, the phase of $\alpha > \alpha_c$ is called the *predictable* or the *asymmetric* phase, as there is bias of winning actions given the past M-bit history string (the winning probabilities are asymmetric).

Owing to the results in these histograms, a useful quantity can be defined to measure the "non-uniformity" of the winning probabilities or the information content given by the past M-bit history string. We denote H to be the *predictability* of the game which is given by the following formula,

$$H = \frac{1}{P}\sum_{\mu=1}^{P}\langle A|\mu\rangle^2 \tag{8}$$

with $P = 2^M$ again. H/N is plotted as a function of α in Fig. 3. In the symmetric phase, $\langle A|\mu\rangle = 0$ for all μ as the actions of "-1" and "1" are equally likely to appear after μ. Hence, $H = 0$ in the symmetric phase. In the asymmetric phase, $\langle A|\mu\rangle \neq 0$ for all μ as the actions of "-1" and "1" are not equally likely after μ. Hence, $H > 0$ in the asymmetric phase. H begins to increase at $\alpha = \alpha_c$ as shown in Fig. 3. Analytic approaches are developed that are based on the minimization of predictability H [7,8,9,10]. In addition to the predictability H, the fraction of *frozen* agents also increases drastically before α_c and decreases afterward. Frozen agents are agents that always use the same strategy for making decisions. In contrast, *fickle* agents are those who always switch strategies.

In exogenous games, the phase transition, the scaling by α and the properties of the two phases are preserved. The symmetric and asymmetric winning probabilities are

also preserved with μ representing the random information given to the agents in the conditional probabilities $P(1|\mu)$ (not the actual past history) [30,32]. The numerical values of volatility in the asymmetric phase are slightly deviated from that in the endogenous games. Because the winning probabilities are asymmetric, the probability of history appearance is non-uniform in endogenous games, while in exogenous games, we assume a uniform appearance probability of all the random information μ being given to the agents.

Formation of Crowds, Anticrowds and Anti-Persistence in the Symmetric Phase

In the symmetric phase with small α, the amount of available information P is small when compared to the number of agents N. Agents are able to exploit the information well and they react like a crowd which results in a large volatility. From the point of view of the strategy space, the number of independent strategies [4,5,6,31] (as discussed in RSS) is smaller than N in the symmetric phase. Many agents use identical strategies and react in the same or similar ways, forming *crowds* and *anticrowds* giving large volatility. This is sometimes known as the herd effect in the minority game. On the other hand, in the asymmetric phase with large α, the available information P is too large and complex when compared to N. Agents are not able to exploit all the information and they react in a way similar to making random decisions, resulting in a volatility approaching the coin-toss limit. In this case, the number of independent strategies is greater than N and agents are unlikely to use the same strategies, they act independently and crowds are not formed. The phase transition occurs when α_c is roughly $O(1)$, where N is roughly the same size as the available information P.

In addition to formation of crowds, anti-persistence of the winning actions exists in the symmetric phase. For small α in the symmetric phase, consecutive occurrence of the same signal leads to opposite winning actions [3,11,33], which is known as anti-persistence of the Minority Game. For example, in the case of $M = 2$, if the history "01" leads to a winning choice of "0", then the next appearance of the history "01" will lead to a winning choice of "1". This results in periodic dynamics of the game for $\alpha < \alpha_c$ with a period of $2^M \cdot 2$, where every history appears exactly twice with different winning actions for the first and second occurrence. This kind of anti-persistence disappears in the asymmetric phase. Instead, persistence is more likely [11], in which the consecutive occurrence of the same signal tends to have a higher probability in giving out the same winning actions.

Dependence on Temperature and Initial Conditions

In 1999, Cavagna et al. [34] introduced the probabilistic fashion, the *temperature*, to the decision-making process of agents in a model known as the Thermal Minority Game (TMG). This stochasticity of temperature can also be implemented in the basic game. Instead of choosing the best strategy for sure, agents employ their strategy s with the probabilities $\pi_{i,s}$ given by

$$\text{Prob}\,\{s_i(t) = s\} = \pi_{i,s} = \frac{e^{\Gamma U_{i,s}(t)}}{\sum_{s'} e^{\Gamma U_{i,s'}(t)}} \tag{9}$$

where $s_i(t)$ denotes the strategy being employed by agent i at time t. Γ is denoted by β in the original formulation, which corresponds to the *inverse temperature* (as in physical systems) of individual agents. It can also be interpreted as the *learning rate* of the system [10]. Roughly speaking, this is because the dynamics of scores take a time of approximately $1/\Gamma$ to learn a difference in the cumulated payoffs of the strategies. For small Γ, the system takes a convergence time of order N/Γ to reach the steady state [35,36], which also reveals the physical meaning of Γ as the learning rate.

In the asymmetric phase, the final state of the system and hence the volatility are independent of Γ [10]. In the symmetric phase, the final state of the system is dependent on Γ and the volatility of the system increases with increasing Γ, provided that the system has reached the steady state [36]. This property of the game is in contrast to the ordinary physical systems, where fluctuations increase with increasing temperature. In the Minority Game, fluctuations increase with increasing Γ, i. e., decreasing temperature, as Γ is implemented as an individual inverse temperature in choosing strategies. As a result, in addition to inverse temperature or learning rate, Γ can also be interpreted as *collective* or *global effective temperature* of the whole system, since global fluctuations increase with Γ. In contrast to σ^2, predictability H is independent of Γ in both symmetric and asymmetric phases [10].

In addition to dependence on Γ in the symmetry phase, the final state of the system is dependent on the initial conditions. For games with the same set of strategies among agents (identical quenched disorders), the final state of the system is dependent on the bias of initial virtual scores (heterogeneous initial conditions of annealed variables) of the strategies [10,33]. For the case of $S = 2$, the volatility of the system is smaller if larger differences are assigned to the initial virtual scores (i. e., initial bias) of the two strategies, given that the same Γ is implemented [9]. A system with a final state depending on the initial state of the annealed variable corresponds to the spin glass phase,

or the replica symmetry breaking (RSB) phenomenon in physical systems. Thus, the symmetric phase also corresponds to the behaviors of broken replica symmetry. On the other hand, in the asymmetric phase, the final state of the system and hence the values of volatility is independent of the initial conditions. Thus, the asymmetric phase also corresponds to the replica symmetry (RS) phase in physical systems.

The Cases of $S > 2$

Finally, the behaviors of the game are also dependent on S, the number of strategies that each agent holds. As shown in Fig. 4 where the volatility is plotted against $\alpha = 2^M/N$, the volatility of the system is dependent on S. Data collapse of volatility with different values of N and M is still shown by plotting volatility against α, for each value of S. While the generic shape of the curves is preserved when S increases, the points of minimum volatility shift to the right which suggests that the phase transition point is a function of S. It is also suggested in [31,37] that for the cases of $S > 2$, instead of $\alpha = 2^M/N$, the important control parameter should be $2^{M+1}/SN$. Since 2^{M+1} is the number of important strategies in the reduced strategy space and SN is the total number of strategies held by all agents, when $2^{M+1} < SN$, some agents are using identical strategies and crowds and anticrowds are formed. On the other hand, when $2^{M+1} > SN$, most agents are using independent strategies and crowds are not formed. Numerical solutions from the replica approach for different values of S show the relation of $\alpha_c(S) \approx \alpha_c(S = 2) + (S - 2)/2$ to a high degree of accuracy [9].

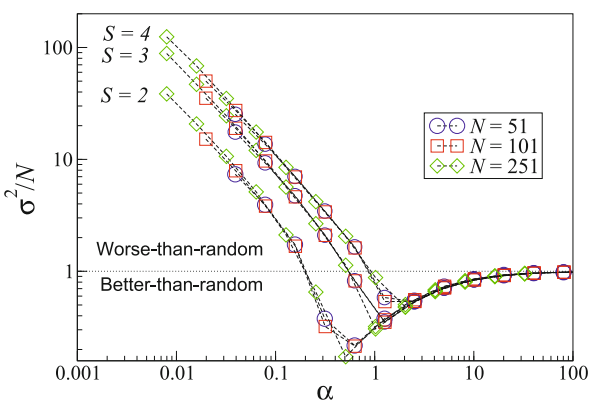

Minority Games, Figure 4
The simulation results of the volatility as a function of the control parameter $\alpha = 2^M/N$ for games with $S = 2, 3, 4$ strategies for each agent averaged over 100 samples. Endogenous information and linear payoff are adopted in these simulations. Volatility generally increases with the number of strategies S per agent. Data collapse of volatility is shown for different values of S

Variants of the Minority Game

After the first publication of Arthur's bar problem and the Minority Game, many variants of the game were established and studied by the physics community and also some economists. Some of these variant models were developed to further simplify the Minority Game or to include more features from the financial markets. Most of the modifications include the use of different kinds of strategies and payoff functions, the presence of different kinds of agents and the increased flexibility in participation of agents, evolution of agents, replacement of poorly performing agents by new agents and the individual concerns for capital. In this section, some of these variants are briefly introduced together with their physical significance to the development of the Minority Game. We leave their implications in regard to the financial markets to Sect. "Minority Games and Financial Markets".

The Evolutionary Minority Game or the Genetic Model

In 1999, N.F. Johnson et al. [20] introduced the Evolutionary Minority Game, which is usually simply described as EMG in the literature or the Genetic Model in later literature. From the name of this model, we can see that the evolution of agents is an important feature added to the game. In addition, the strategies employed by agents are also major modifications. Unlike the basic Minority Game, all agents in EMG hold only one strategy $S = 1$ and the strategy table is identical for everyone. For example in the case of $M = 3$, all agents hold one strategy as in Table 1. Instead of having a column of fixed predictions, this column records the most recent past winning action or choice for the corresponding history. Thus, this strategy table is time dependent. To make decisions, all agents are assigned a different probability p_i at the beginning, with $0 \le p_i \le 1$, which is defined as the probability that agent i acts according to the strategy table, i. e., will follow the recent winning action or the last outcome for that M-bit history. With a probability $1 - p_i$, agent i chooses the choice opposite to the past winning action for that history. This probability p_i (rather than the strategy table) acts as a role of strategy in making decisions for agents and is called "strategy" in EMG or the "gene" value in Genetic Model. Hence, the payoff or scores are rewarded or penalized subject to p_i.

To enhance the evolutionary property of the game, agents are allowed to modify the p_i if the scores fall below a threshold denoted by d, where $d < 0$, which is sometimes known as the death score. The new p_i is being drawn

with an equal probability in the range $(p_i - r/2, p_i + r/2)$ of width r, with either a periodic or reflective boundary condition at $p_i = 0$ or $p_i = 1$. This corresponds to an evolution of strategy (the probability p_i), or the mutation of the gene value p_i with mutation range r, as a result the EMG is also known as the Genetic Model. It is found that in the ordinary EMG with the winning rule being the minority group ($A(t) < N/2$), the memory M of the strategy table is not relevant in affecting the major features (including the steady state distribution $\mathrm{P}(p_i)$) of the system [38], but may be relevant with other winning rules (winning level other than $N/2$) [39]. Thus, the volatility is independent of M in the ordinary EMG, in contrast to the basic Minority Game.

In this ordinary model, one point is added to or deducted from the strategy p_i for winning or losing predictions. The possibility of having a non-unity price-to-fine ratio R was introduced by Hod et al. [38,40]. For $R > 1$, agents are in a wealthy regime since the gain is larger than the loss in one game. For $R < 1$, agents are in a tough regime. The system behaviors are dependent on R, with two thresholds $R_c^{(1)}$ and $R_c^{(2)}$, both $R_c^{(1)}$ and $R_c^{(2)}$ are less than and close to 1, with $R_c^{(1)} > R_c^{(2)}$. In the regime where $R > R_c^{(1)}$, after a sufficiently long time of evolution, the agents self-segregated into two opposing groups at $p_i = 0, 1$ and a "U" shaped" $P(p)$ distribution is found as shown in the case of $R = 1$ in Fig. 5. This implies agents tend to behave in an extreme way in a rich economy. In the regime where $R < R_c^{(2)}$, the agents cluster at $p_i = 0.5$ and an "inverse-U" shaped" distribution is found as shown in the case of $R = 0.97$ in Fig. 5, implying agents tend to be more cautious in a poor economy.

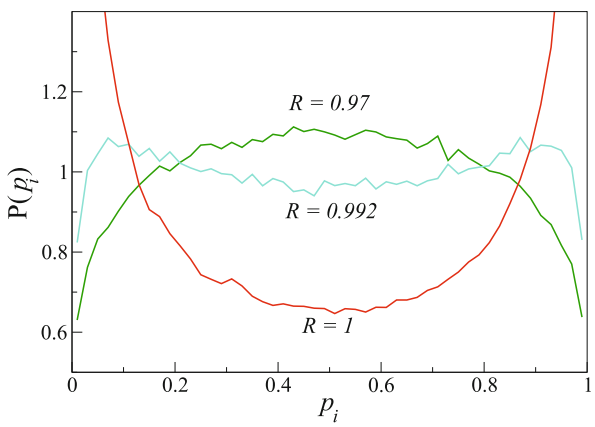

Minority Games, Figure 5
The simulation results of the distribution $\mathrm{P}(p_i)$ in a game of $N = 10001$ agents with $d = -4$ and $M = 3$. The distribution is obtained after 120 000 time steps and averaged over 50 simulations. Self-segregation and clustering of agents at different values of R are shown

The Thermal Minority Game

In addition to the stochasticity in choosing strategies as introduced by Eq. (9) discussed in Sect. "The Physical Properties of the Minority Game", there are several other modifications from the basic Minority Game in the TMG [34]. In the original formulation, the strategy is a vector in the P-dimensional real space $\boldsymbol{R}^P$ denoted by $\vec{a}_{i,s}$, with $\|\vec{a}_{i,s}\| = \sqrt{P}$. Thus, the strategy space is the surface of the P-dimensional hypersphere and the components of the strategies are continuous. This is different from the discrete strategies in the basic Minority Game. Every agent in the game draws S vector strategies before the game starts.

The information processed by the strategies is a random vector $\vec{\eta}(t)$, with a unit-length in $\boldsymbol{R}^P$. The response or the bid of the strategy is no longer integer and is given by the inner product of the strategies and the information, i. e., $\vec{a}_{i,s} \cdot \vec{\eta}(t)$. Hence, the attendance $A(t)$ is given by

$$A(t) = \sum_{i=1}^{N} \vec{a}_{i,s_i(t)} \cdot \vec{\eta}(t) \tag{10}$$

with $s_i(t)$ denoting the chosen strategy of agent i at time t. The cumulated payoff of strategy can be updated by

$$U_{i,s}(t+1) = U_{i,s}(t) - A(t)\left[\vec{a}_{i,s} \cdot \vec{\eta}(t)\right] . \tag{11}$$

TMG can be considered as a continuous formulation of the Minority Game, in which the game is no longer discrete and binary. Since the response of the strategy in TMG is defined as the inner product, i. e., a sum over all the P entries of the vectors, all components of the strategy have to predict at each round. This is a difference from the basic Minority Game where at each round, only one of the P predictions on the strategy is effective. Despite these differences and the continuous formulation, TMG reproduces the same collective behaviors as the basic MG [34,35].

The Simple Minority Game Without Information

In this simplified Minority Game, no information is given to agents and thus they have no strategy tables. Agents choose between the choice +1 and − 1 according to the following probabilities,

$$\mathrm{Prob}\{a_i(t) = \pm 1\} \equiv \frac{\mathrm{e}^{\pm U_i(t)}}{\mathrm{e}^{U_i(t)} + \mathrm{e}^{-U_i(t)}} \tag{12}$$

and $U_i(t)$ is updated by

$$U_i(t+1) = U_i(t) - \frac{\Gamma}{N} A(t) \tag{13}$$

where $U_i(t)$ can be considered as the virtual score for

agent i to make a decision of "+1", and $-U_i(t)$ correspondingly the virtual score for agent i to make a decision of "−1". If $U_i(t) > 0$, the past experience of the agents shows that it is more successful to take action $a_i(t) = +1$, and vice versa. The learning rate or the temperature Γ is implemented in this model. This model gives a very simple analytic explanation of the system's dependence on Γ. For $\Gamma < \Gamma_c$, the volatility is found to be proportional to N, i. e., $\sigma^2 \propto N$. For $\Gamma > \Gamma_c$, $\sigma^2 \propto N^2$. Γ_c is found to be dependent on the initial conditions $U_i(0)$. In addition, σ^2 decreases with increasing $U_i(0)$. Similar dependence of the volatility on Γ and initial conditions, and the dependence of Γ_c on initial conditions are also found in the basic Minority Game.

The Grand-Canonical Minority Game

The grand-canonical Minority Game (GCMG) refers to a subclass of Minority Games where the number of agents who actively participate in the market is variable. Instead of summing over all N agents, the collective action or the attendance $A(t)$ is effectively the sum of the actions of active agents at time t. Agents can be active or inactive at any time, depending on their potential profitability from the market. The term "grand-canonical" originates from the grand-canonical ensemble in statistical mechanics where the number of particles in the observing system is variable. In most of the formulations [23,26,28], when the highest virtual score of the strategies that an agent holds is below some threshold or ϵt (where ϵ is usually a positive constant being referred to as the *interest rate* and t is the number of rounds or time steps since the beginning of the game), the agent refrains from participating in that round of the game. It is equivalent to the addition of an *inactive strategy* for every agent from which agents become inactive, and the virtual score of this strategy is ϵt. Physically, it corresponds to circumstances of gaining an interest of ϵ at each time step by keeping the capital in the form of cash, so agents would only participate in the market if the gain from investments in the market is greater than the interest rate. In some other formulations, instead of the virtual score of the strategies, the real score of the agents is used to compare with the interest rate [24]. Winning probabilities of strategies within a certain time horizon are also considered [24,41]. Individual capital concerns can also be implemented to achieve the grand-canonical nature of the game [25,28], in which agents vary their investment size at each time step by considering risk, gain potential or their limited capital.

These grand-canonical modifications from the Minority Game are considered to be important and crucial in producing the stylized facts of financial markets in the Minority Game models [24,25,26,27,28], while preserving the two-phase structure of the predictable and unpredictable phase. The stylized facts being reproduced in the models include the fat-tail volatility or price return distributions and volatility clustering, when the systems are close to the critical state. Numerical tests and analytical attempts are carried out in the critical regime of the models. These models serve as a tool for physicists in understanding how macroscopic features are produced from the microscopic dynamics of individuals, which also supports the conjecture of self-organized criticality of the financial market in which the financial market is always close to or attracted to the critical state.

Analytic Approaches to the Minority Game

There are several analytic approaches to solving the Minority Game. Most of the approaches are based on the models of the basic Minority Game with little modifications or simplification. It is found that in the asymmetric phase with $\alpha > \alpha_c$, both equilibrium approaches and dynamics approaches are likely to describe the same behavior of the system, and the equilibrium approach based on the minimization of H gives an analytic solution in this phase [7,8,9,10,12]. For the symmetric phase with $\alpha < \alpha_c$, fluctuations in the dynamics have to be considered and the solution is dependent on initial conditions. The final state of the system is sensitive to initial conditions and perturbations in the dynamics [10,12,14]. In this case, a solution is available in the limit of $\Gamma \to 0$ or asymptotic behaviors can be obtained in the limit of $\alpha \to 0$.

One of the early approaches to solving the Minority Game was the crowd-anticrowd theory which provides a qualitative explanation of the volatility dependence on brain size M [4,5,6]. Consider the reduced strategy space with strategies $R = 1 \ldots 2^M$ to be the uncorrelated strategies, $\bar{R}$ to be the anti-correlated strategy of R (i. e., R and $\bar{R}$ always have opposite decisions), which constitute the 2^{M+1} strategies in the RSS. We denote n_R to be the number of agents using the strategy R, $n_{\bar{R}}$ to be the number of agents using $\bar{R}$ and $\langle \ldots \rangle$ to be time averaging. For R and R' to be uncorrelated strategies, the time average $\langle \sum_{R \neq R'} (n_R - n_{\bar{R}})(n_{R'} - n_{\bar{R}'}) \rangle = 0$ and thus the volatility σ^2 can be expressed as

$$\sigma^2 = \sum_{R=1}^{2^M} (n_R - n_{\bar{R}})^2 \tag{14}$$

which physically corresponds to the contribution to the global volatility from each crowd-anticrowd pair $(R, \bar{R})$,

as R and $\bar{R}$ are always making opposite decisions. If we consider a uniform distribution of all strategy combinations among agents at the beginning of the game, n_R and $n_{\bar{R}}$ can be determined from the ranking of virtual scores of strategies, since agents are always using the best strategies they hold. In this case, the strategy with the highest virtual points would be the most popular strategy, while its anti-correlated partner would have the lowest virtual points and becomes the least popular strategy. This happens for small M where the number of strategies is small and a large number of agents are using the best strategy. On the contrary only a small number of agents are using its anti-correlated strategy, leading to a large $|n_R - n_{\bar{R}}|$ and a large volatility. For the cases of large M, even for the best strategy, n_R is relatively small and the $n_{\bar{R}}$ may have a similar magnitude as n_R, leading to a small $|n_R - n_{\bar{R}}|$ and a small volatility. This qualitatively explains the behaviors of volatility with α based on the size of crowd-anticrowd pairs. The presence of temperature is also considered in the extended crowd-anticrowd approach [6].

In addition to this crowd-anticrowd theory, a full analytic approach can be developed. To solve the Minority Game analytically, we employ some convenient notation changes which make the tools in statistical physics more applicable [7,8,9]. For the case of $S = 2$, we denote the first strategy of an agent to be "+1" while the second one is "-1", whereas the best strategy of agent i at time t is now expressed as $s_i(t) = \pm 1$. The real bid $a_i(t)$ of agent i at time t can then be expressed as

$$a_i(t) = a^{\mu(t)}_{i,s_i(t)} = \omega^{\mu(t)}_i + s_i(t)\xi^{\mu(t)}_i \tag{15}$$

where $\omega^\mu_i = (a^\mu_{i,+} + a^\mu_{i,-})/2$ and $\xi^\mu_i = (a^\mu_{i,+} - a^\mu_{i,-})/2$. ω^μ_i and ξ^μ_i are quenched disorders and are fixed at the beginning of the game. $\omega^\mu_i, \xi^\mu_i = 0, \pm 1$ and $\omega^\mu_i \xi^\mu_i = 0$ for all μ. $s_i(t)$ is the dynamic variable and becomes explicit in the action of agents, corresponding to the Ising spins in physical systems. Thus, the attendance can be expressed as a function of spin $s_i(t)$ given by

$$A(t) = \Omega^{\mu(t)} + \sum_{i=1}^{N} \xi^{\mu(t)}_i s_i(t) \tag{16}$$

where $\Omega^\mu = \sum_i \omega^\mu_i$.

Other than the spin $s_i(t)$, the virtual scores of the strategies are also dynamic and we denote the difference of the virtual scores of the two strategies of agent i to be $Y_i(t)$ given by

$$Y_i(t) = \frac{\Gamma}{2}(U_{i,+}(t) - U_{i,-}(t)) \,. \tag{17}$$

This $Y_i(t)$ determines the relative probabilities of using the two strategies with "inverse temperature" Γ and is updated by

$$Y_i(t+1) = Y_i(t) - \frac{\Gamma}{N}\xi^{\mu(t)}_i A(t) \tag{18}$$

which is given by the update of the individual virtual scores $U_{i,+}(t)$ and $U_{i,-}(t)$ in Eq. (6) with a factor of $1/N$ in the last term. Thus, the probabilities Eq. (9) for using the strategies $s_i(t) = \pm 1$ at time t becomes

$$\text{Prob}\,\{s_i(t) = \pm 1\} = \pi_{i,\pm} = \frac{1 \pm \tanh Y_i(t)}{2} \,. \tag{19}$$

From this equation, we can calculate the time average of $s_i(t)$ at equilibrium with probabilities Eq. (19), denoted by m_i, to be

$$m_i = \langle s_i \rangle = \langle \tanh(Y_i) \rangle \,. \tag{20}$$

The system will be stationary with $\langle Y_i \rangle \sim v_i t$, corresponding to a stationary state solution of the set of m_i. From Eq. (18), v_i can be expressed as

$$v_i = -\overline{\Omega\xi_i} - \sum_{j=1}^{N} \overline{\xi_i \xi_j} m_j \,. \tag{21}$$

where $\overline{\cdots}$ denotes the average over μ

For $v_i \neq 0$, $\langle Y_i \rangle$ diverges to $\pm\infty$ and gives $m_i = \pm 1$, corresponding to the frozen agents who always use the same strategy. For $v_i = 0$, $\langle Y_i \rangle$ remains finite even after a long time and $|m_i| < 1$, corresponding to the fickle agents who always switch their active strategy even in the stationary state of the game. We can identify $\overline{\Omega\xi_i} + \sum_{j\neq i} \overline{\xi_i \xi_j} m_j$ as an external field while $\overline{\xi_i^2}$ is the self-interaction of agent i. For an agent to be frozen, the magnitude of the external field has to be greater than the self-interaction. In order to have fickle agents in the stationary state, the self-interaction term is crucial.

We note that the above equation of v_i in Eq. (21) and the corresponding conditions of frozen and fickle agents are equivalent to the minimization of predictability H, with H written in the form

$$H = \frac{1}{P}\sum_{\mu=1}^{P}\left[\Omega^\mu + \sum_{i=1}^{N} \xi^\mu_i m_i\right]^2 \,. \tag{22}$$

Since m_i's are bounded in the range $[-1, +1]$, H either attains its minimum at $\mathrm{d}H/\mathrm{d}m_i = 0$, giving $\overline{\Omega\xi_i} + \sum_j \overline{\xi_i \xi_j} m_j = 0$ (fickle agents) or at the boundary of the range $[-1, +1]$ of m_i, giving $m_i = \pm 1$ (frozen

agents). Thus, we can identify H as the Hamiltonian where the stationary state of the system is the ground state that minimizes the Hamiltonian. From Eq. (16), the volatility of the system can be expressed as

$$\sigma^2 = H + \sum_{i=1}^{N} \overline{\xi^2}(1 - m_i^2) + \sum_{i \neq j} \overline{\xi_i \xi_j} \left\langle (\tanh Y_i - m_i)(\tanh Y_j - m_j) \right\rangle . \quad (23)$$

The last term involves the fluctuations around the average behavior of the agents and is related to the dynamics of the system.

Identifying H as the Hamiltonian reduces the problem to a conventional physical problem of finding the ground state of the system by minimizing the Hamiltonian. Solving the problem involves averaging the quantity $\ln Z$ over quenched disorders corresponding to the strategies $a_{i,\pm}^{\mu}$ (now represented by ω_i^{μ} and ξ_i^{μ}) given to the agents at the beginning of the game. We note that the system is a fully connected system in which agents interact with all other agents, and can be handled by the *replica* approach as in spin glass models, under the assumption of replica symmetry. Given the fraction of frozen agents to be ϕ, which can be expressed as a function of α, $\alpha_c = 0.3374\ldots$ [8] is found to be the solution of the equation

$$\alpha = 1 - \phi(\alpha) . \quad (24)$$

In addition, this approach allows us to see that the system with different initial conditions converges to the same unique solution, corresponding to a single minima of H in the phase $\alpha > \alpha_c$ (replica symmetry). This method allows us to obtain a complete solution for the Minority Game for all Γ in the phase $\alpha > \alpha_c$. Macroscopic quantities such as σ^2 and H can be analytically calculated. For $\alpha < \alpha_c$, there are multiple minima of $H = 0$ and the system's final state is not unique (replica symmetry breaking) and depends on its initial state. In this case, dynamics has to be considered. Breaking of replica symmetry is also considered in the case with market impact [17].

We notice that in the long run, the characteristic time in the dynamics is approximately proportional to N, where all agents observe the performance of their strategies among all P states with $P = \alpha N$. This characteristic time is also inversely proportional to Γ since the dynamics of scores take a time of approximately $1/\Gamma$ to adapt a change of scores, as discussed earlier. The real time t can then be rescaled as

$$\tau = \frac{\Gamma}{N} t \quad (25)$$

in which one characteristic time step τ in the system corresponds to N/Γ real time steps t. This is the reason for the systems with small Γ having a convergence time of N/Γ. We can hence write a dynamical equation for Y_i in the rescaled time by denoting the variable $y_i(\tau) = Y_i(N\tau\Gamma)$ which gives

$$\frac{\mathrm{d}y_i}{\mathrm{d}\tau} = -\overline{\Omega \xi_i} - \sum_{j=1}^{N} \overline{\xi_i \xi_j} \tanh(y_i) + \zeta_i \quad (26)$$

where the first two terms on the right hand side represent the average behavior of agents obtained by the average frequency they play their strategies [10]. These two terms are considered to be deterministic. The last term ζ_i represents the noise or the fluctuations around the average behavior. The properties of these fluctuations are given by

$$\langle \zeta(\tau) \rangle = 0 \quad (27)$$

$$\left\langle \zeta(\tau)\zeta(\tau') \right\rangle \cong \frac{\Gamma \sigma^2}{N} \overline{\xi_i \xi_j} \delta\left(\tau - \tau'\right) . \quad (28)$$

By writing the Fokker–Planck equation for the probability distribution $P\left(\{y_i\}, t\right)$ [10], many physical implications can be obtained. We first note that the noise covariance Eq. (28) is linearly related to Γ, revealing the role of Γ as the global temperature of the system. When $\Gamma \to 0$, the noise covariance vanishes and the minimization of H gives a valid solution, even for $\alpha < \alpha_c$. It can also be deduced that in the asymmetric phase, the last term in Eq. (23) vanishes such that σ^2 is independent of Γ and initial conditions. In the symmetric phase, this last term does not vanish and σ^2 is dependent on both Γ and initial conditions.

An alternative approach to derive dynamical equations is the generating functional approach [14,15,16], which monitors the dynamics using path integrals over time. The approach was first used on the *batch* update version of the Minority Game, in which agents update their virtual scores only after a batch of P time steps and with $\Gamma \to \infty$ as in the basic Minority Game. The quenched disorder can be averaged out in the dynamical equations and in the limit of $N \to \infty$, we obtain a representative "single" agent dynamical equation with the variable $y(t)$, where $y(t)$ represents the difference in the virtual scores of the two strategies of this "single" agent after the tth batch. The dynamics are stochastic but non-Markovian in nature, and can be extended to regions inaccessible by the replica method. This method again confirms the relation of Eq. (24) and gives the same value of α_c [14]. For $\alpha > \alpha_c$, the fraction of frozen agents ϕ is obtained analytically and the volatility is calculated to a high accuracy. For $\alpha < \alpha_c$ in the limit of $\alpha \to 0$, σ is shown to diverge as $\sigma \sim \alpha^{-1/2}$ for

$y(0) < y_c$, and vanishes as $\sigma \sim \alpha^{1/2}$ for $y(0) > y_c$, with $y_c \approx 0.242$ [14]. This approach was later extended to the case of on-line update (update of virtual score after every step) and the cases of $\Gamma < \infty$ [15,16].

Minority Games and Financial Markets

The basic Minority Game model is a simple model that is used to describe the possible interaction of investors in the financial markets. Despite its simplicity, some variants of the game show certain predictive abilities for real financial data [22,24,41]. Though Minority Games are simple, they setup a framework of agent-based models from which sophisticated trading models can be built, and implementation in real trading may be possible. Although these sophisticated models are usually used for private trading and may not be open to the public, Minority Games are still a useful tool to understand the dynamics in financial markets. There are several fundamental differences between the basic game and the markets. The basic Minority Game is a negative sum game in which the sum of gain of all agents is negative. Agents are not concerned with capital and cannot refrain from participation even if they found the game unprofitable. Whether the simple payoff function in Eq. (6) correctly represents the evaluation of strategies by the real investor is questionable. We also note that the symmetric phase corresponds to a phase of information efficiency in which the game becomes unpredictable.

Although the basic Minority Game provides a very colorful collective behavior of agents from simple interactions and dynamics, some details can affect the behaviors and modifications have to be made in order to draw a more direct correspondence of Minority Games to real financial markets. Some variants of the Minority Game are modified to study a particular issue or aspect of the real markets. While with the introduction of several financial aspects, some variants of the game lead to a more realistic model of the market, at the same time they complicate the models. Among the different aspects, one of the primary issues is to draw an analogy to trading where price dynamics have to be introduced to the Minority Game. A common price dynamic used in the game is to relate the attendance $A(t)$ to the price $p(t)$, and thus the return $r(t)$ in trading is given by

$$r(t) \equiv \log[p(t+1)] - \log[p(t)] = \frac{A(t)}{\lambda} \tag{29}$$

where λ is called the *liquidity*, which is used to control the sensitivity of price on attendance. With this or similar price dynamics, the trading process can be defined in the game.

After the introduction of price dynamics, the issue of payoff function was also addressed. The mixed minority-majority game originally proposed by M. Marsili [18] is based on a simplified version of the Minority Game in which there is no strategy table and no information (as discussed in Sect. "Variants of the Minority Game"). The payoff function in this model is based on the expectations of the agents in relation to the price change in the next steps. For simplicity, we consider the expectation $E_i[A(t+1)|t]$ of agent i on the attendance $A(t+1)$ in the next step, which is expressed as

$$E_i[A(t+1)|t] = -\Phi_i A(t)\,. \tag{30}$$

For $\Phi_i > 0$, agents expect the attendance in the next step to be negatively correlated with that in the present step (i. e., the price fluctuates), revealing the minority nature of the agents and they are called *fundamentalists* or *contrarian* agents. For $\Phi_i < 0$, agents expect the attendance in the next step to be positively correlated with that in the present step (i. e., a price trend develops), revealing the majority nature of the agents and they are called *trend followers*. For both fundamentalists and trend followers, if they expect the price to go up in the next step, buying is considered to be profitable, and vice versa. Thus, the payoff function $\delta U_i(t) = U_i(t+1) - U_i(t)$ is proportional to the product of the current decision $a_i(t)$ and the expectation of a price change in the next step is given by

$$\delta U_i(t) \propto a_i(t) E_i[A(t+1)|t] = -\Phi_i a_i(t) A(t) \tag{31}$$

with $\Phi_i > 0$ and $\Phi_i < 0$ corresponding to fundamentalists and trend followers, respectively. Hence, fundamentalists are considered to be playing a minority game while trend followers play a majority game.

The two kinds of agents interact in the same game and it was found that the ratio of fundamentalists to trend followers is important in affecting the behaviors of the system. If more than half of the agents are fundamentalists, the fundamentalists prevail and the game is minority in nature with $\langle A(t+1)A(t)\rangle < 0$. On the other hand, if more than half of the agents are trend followers, the trend followers prevail and the game is majority in nature with $\langle A(t+1)A(t)\rangle > 0$. Thus, the behaviors of both minority and majority agents are found to be self-sustained, depending on the relative population of the agents.

The \$-game [19] shares some similarity to the majority nature of the trend followers in the mixed minority-majority game, but with a crucial difference of using the real attendance of the next step, not the expectations of agents, in the payoff function. In the \$-game, agents are again equipped with strategy tables and the virtual score

Minority Games, Table 2
The payoff functions of the Minority Game, the majority game and the $-game, with Φ_i set to ±1 in the mixed minority-majority game

	$\delta U_{i,s}(t)$
Minority Game	$-a^{\mu}_{i,s}(t)A(t)$
Majority Game	$a^{\mu}_{i,s}(t)A(t)$
$-game	$a^{\mu}_{i,s}(t-1)A(t)$

of strategy s is updated according to

$$U_{i,s}(t+1) = U_{i,s}(t) + a^{\mu}_{i,s}(t-1)A(t)\,. \tag{32}$$

According to this payoff scheme, the present actions $a^{\mu}_{i,s}(t)$ would only change the payoff at the next step at $t+1$. Suppose an agent buys an asset, he gains by selling the asset in the next step if the price rises, and vice versa. This payoff scheme aims to model the mode of one-step speculating in realistic markets though agents are not restricted to act oppositely in consecutive steps in the model. Bubble-like behaviors are found in the model, in which agents buy (sell) and push up (down) the price, leading to positive evaluations of the buying actions such that agents are more likely to buy (sell) again. This process continues and a persistent price trend is observed. This persistent price trend is not observed in real markets, and can be eliminated from the model if agents are concerned with their limited capital, risk or maximum holding of assets [19,21,22,28]. To summarize the different payoff schemes in the Minority game, majority game and the $-game, Table 2 shows the payoff functions of the three games.

Other than the payoff functions, we consider the stationary state of collective behaviors in the system. The stationary state of the Minority Game is not a Nash Equilibrium. There are an extremely large number of Nash Equilibria in the Minority Game [17,18] and one example is $(N+1)/2$ agents always make an action of $a_i(t) = +1$ while $(N-1)/2$ agents always make an action of $a_i(t) = -1$. In this case, $\sigma = 1$ and no individual has the incentive to change his action by himself (the majority group change if any of the losers moves). This state is not stationary in the game. The stationary state of the game is described by the minimization of predictability H, but Nash Equilibria are states of minimum volatility σ^2. Physically, instead of competing with the other $N-1$ players, agents are interacting with the total attendance $A(t)$ which includes also its own action. By subtracting their own actions from the attendance, their cumulated payoffs Eq. (13) in the simplified minority game becomes

$$U_i(t+1) = U_i(t) - \frac{\Gamma}{N}\left[A(t) - \eta a_i(t)\right] \tag{33}$$

where eta denotes the *market impact*. The Minority Game corresponds to the case of $\eta = 0$ in which Nash Equilibrium is not attained. In this simplified model, $\eta > 0$ brings heterogeneity to the behaviors of agents and Nash Equilibrium is attained [18].

In some variants of the Minority Game, the role of participants are studied. It was suggested that a symbiotic relation is present between two kinds of traders [7,25,42], namely the *producers* and the *speculators*. Producers are agents who always participate and trade with only one strategy. They have a primary interest in trading in the market for business or other reasons. Speculators are agents who speculate and have no interest in the intrinsic values of the assets traded. They can refrain from participating in the market at any time they found it unprofitable. This model corresponds to one of the grand-canonical Minority Games. In this model, it was found that the gains of producers are always negative but their losses decrease with an increasing number of speculators, since speculators provide liquidity to producers and make the market more unpredictable. On the other hand, the gain of speculators generally increases with the number of producers, since producers provide more information to speculators which make the market more predictable. As a result, producers and speculators are symbiotic.

In addition to studying the role of producers and speculators, the grand-canonical Minority Game plays a crucial role in understanding financial markets [24,26,28]. By introducing the grand-canonical nature to the model, agents can choose to refrain from participating in the market when they found the game unprofitable. While observing the market as non-trading outsiders, they can participate in the market again once they found it profitable. The predictable and unpredictable phases are usually preserved in this class of models, where fat-tail distributions of price return are found around the phase transition point. These can be fitted by power laws. While outside the critical region, the fluctuation distributions becomes Gaussian. In addition, volatility clustering, where high volatility is likely to cluster in time, is also found and can be fitted by power laws or other forms of function [25,26,28,43]. This suggests that the dynamics of agents are correlated in time.

The observation of power laws in the model coincides with the observations of fat-tail price return distributions and volatility clustering in real markets in the high frequency range [44]. While Gaussian fluctuations are not found in real markets, these properties of the model pro-

vide important implications or conjectures of financial markets being in the critical state. It also suggests the possibility of a self-organized critical system as an explanation of the behaviors of financial markets. In addition to power laws, rescaling of financial market data of different frequencies [45] also provides preliminary evidence of the property of scale invariance in time in critical systems. Although power laws in financial markets may have origins other than critical phenomenon [46], these conjectures provide a potential perspective in understanding the dynamics and behaviors of the markets. If financial markets exist as a critical phenomenon, their behaviors can be understood qualitatively from the underlying nature of interactions in similar systems within the same universality class. In this case, microscopic details of the systems are not crucial in affecting the generic behaviors for systems in the same class. As financial markets are observed to be operated close to informational efficiency, correspondingly, the grand-canonical Minority Games show stylized facts near the critical point of phase transition to the unpredictable phase, i. e., the phase of informational efficiency. In some versions of the grand-canonical minority game, rarely large fluctuations resulting from a sudden participation of a larger number of speculators are also found which draw analogy to market crashes.

Future Directions

In view of the exciting physical pictures brought along by the grand-canonical Minority Games, more analytic works can be developed to understand the dynamics of the critical regime around which the fat-tail distributions and the volatility clustering are found. The formation of power laws and anomalous fluctuations may be understood with the analytic tools. An analytic approach to the grand-canonical minority game would also provide more clues in proving or disproving the conjecture of the financial market being a self-organized critical phenomenon.

On the other hand, simple modeling that reveals the dynamics of the financial market is still possible. Development of other simple models that draw direct analogy to the financial markets, together with analytic solution is crucial in understanding how the markets work. Other than the grand-canonical games, some variants of the basic Minority Game are still simple and worth solving analytically. Although an analytic solution may not be available for complicated models that introduce more and more realistic aspects into the game, comprehensive modeling based on the inductive nature of agent-based models provides us with a new perspective in understanding the financial markets.

Other than modeling, efforts may be put in to implementing Minority-Game strategies in real trading. Although the strategies from the basic Minority Game and its variants may not accurately describe the strategies for real trading, the simplicity of the strategy does leave us a large freedom in expanding, modifying and tuning with respect to attaining profitability in real trading. Using Minority Games as a framework, a sophisticated real trading system can be built that includes a comprehensive picture of the trading mechanism. Predictive capacity may also be obtained from this kind of agent-based model, which goes beyond the standard economic assumption of deductive agents and market efficiency. All these possible future directions show great potential over the conventional statistical tools in financial analysis. These developments may be used for private trading and may not be accessible through public and academic literature. Some work [22,24,41] has already shown potential in this direction of application.

Bibliography

Primary Literature

1. Challet D, Zhang Y-C (1997) Emergence of cooperation and organization in an evolutionary game. Physica A 246:407
2. Arthur WB (1994) Inductive reasoning and bounded rationality: The El Farol problem. Am Econ Assoc Pap Proc 84:406–411
3. Savit R, Manuca R, Riolo R (1999) Adaptive competition, market efficiency, and phase transitions. Phys Rev Lett 82:2203–2206
4. Johnson NF, Hart M, Hui PM (1999) Crowd effects and volatility in a competitive market. Physica A 269:1
5. Hart ML, Jefferies PPM, Hui P, Johnson NF (2001) Crowd-anticrowd theory of multi-agent market games. Eur Phys J B 20:547–550
6. Hart ML, Jefferies P, Johnson NF, Hui PM (2000) Generalized strategies in the minority game. Phys Rev E 63:017102
7. Challet D, Marsili M, Zhang Y-C (2000) Modeling market mechanisms with minority game. Physica A 276:284
8. Challet D, Marsili M, Zecchina R (2000) Statistical mechanics of heterogeneous agents: Minority games. Phys Rev Lett 84:1824–1827
9. Marsili M, Challet D, Zecchina R (2000) Exact solution of a modified El Farol's bar problem: Efficiency and the role of market impact. Physica A 280:522
10. Marsili M, Challet D (2001) Continuum time limit and stationary states of the minority game. Phys Rev E 64:056138
11. Challet D, Marsili M (1999) Symmetry breaking and phase transition in the minority game. Phys Rev E 60:R6271
12. Garrahan JP, Moro E, Sherrington D (2000) Continuous time dynamics of the thermal minority game. Phys Rev E 62:R9
13. Sherrington D, Moro E, Garrahan JP (2002) Statistical physics of induced correlation in a simple market. Physica A 311:527–535
14. Heimel JAF, Coolen ACC (2001) Generating functional analysis of the dynamics of the batch minority game with random external information. Phys Rev E 63:056121

15. Heimel JAF, Coolen ACC, Sherrington D (2001) Dynamics of the batch minority game with inhomogeneous decision noise. Phys Rev E 65:016126
16. Coolen ACC, Heimel JAF (2001) Dynamical solution of the on-line minority game. J Phys A Math Gen 34:10783
17. De Martino AD, Marsili M (2001) Replica symmetry breaking in the minority game. J Phys A Math Gen 34:2525–2537
18. Marsili M (2001) Market mechanism and expectations in minority and majority games. Phys A 299:93–103
19. Andersen JV, Sornette D (2003) The $-game. Eur Phys J B 31:141
20. Johnson NF, Hui PM, Johnson R, Lo TS (1999) Self-organized segregation within an evolving population. Phys Rev Lett 82:3360–3363
21. Challet D (2008) Inter-pattern speculation: Beyond minority, majority and $-games. J Econ Dyn Control 32:85
22. Yeung CH, Wong KYM, Zhang Y-C (2008) Models of financial markets with extensive participation incentives. Phys Rev E 777:026107
23. Slanina F, Zhang Y-C (1999) Capital flow in a two-component dynamical system. Physica A 272:257–268
24. Jefferies P, Hart ML, Hui PM, Johnson NF (2001) From minority games to real world markets. Eur Phys J B 20:493–502
25. Challet D, Chessa A, Marsili M, Zhang Y-C (2000) From minority games to real markets. Quant Finance 1:168
26. Challet D, Marsili M (2003) Criticality and finite size effects in a simple realistic model of stock market. Phys Rev E 68:036132
27. Challet D, Marsili M, Zhang Y-C (2001) Stylized facts of financial markets in minority games. Physica A 299:228
28. Giardina I, Bouchaud J-P (2003) Bubbles, crashes and intermittency in agent based market models. Eur Phys J B 31:421
29. Cavagna A (1999) Irrelevance of memory in the minority game. Phys Rev E 59:R3783–R3786
30. Cavagna A (2000) Comment on adaptive competition, market efficiency, and phase transitions. Phys Rev Lett 84:1058
31. Challet D, Zhang Y-C (1998) On the minority game: Analytical and numerical studies. Physica A 256:514
32. Savit R (2000) Savit replies on comment on adaptive competition, market efficiency, and phase transitions. Phys Rev Lett 84:1059
33. Wong KYM, Lim SW, Gao Z (2005) Effects of diversity on multi-agent systems: Minority games. Phys Rev E 71:066103
34. Cavagna A, Garrahan JP, Giardina I, Sherrington D (1999) A thermal model for adaptive competition in a market. Phys Rev Lett 83:4429–4432
35. Challet D, Marsili M, Zecchina R (2000) Comment on thermal model for adaptive competition in a market. Phys Rev Lett 85:5008
36. Cavagna A, Garrahan JP, Giardina I, Sherrington D (2000) Reply to comment on thermal model for adaptive competition in a market. Phys Rev Lett 85:5009
37. Zhang Y-C (1998) Modeling market mechanism with evolutionary games. Europhys News 29:51
38. Burgos E, Ceva H (2000) Self organization in a minority game: The role of memory and a probabilistic approach. Physica A 284:489
39. Kay R, Johnson NF (2004) Memory and self-induced shocks in an evolutionary population competing for limited resources. Phys Rev E 70:056101
40. Hod S, Nakar E (2002) Self-segregation versus clustering in the evolutionary minority game. Phys Rev Lett 88:238702
41. Lamper D, Howison SD, Johnson NF (2001) Predictability of large future changes in a competitive evolving population. Phys Rev Lett 88:017902
42. Zhang Y-C (1999) Towards a theory of marginally efficient markets. Physica A 269:30
43. Bouchaud J-P, Giardina I, Mezard M (2001) On a universal mechanism for long-range volatility correlations. Quant Finance 1:212–216
44. Liu Y, Gopikrishnan P, Cizeau P, Meyer M, Peng CK, Stanley HE (1999) Statistical properties of the volatility of price fluctuations. Phys Rev E 60:1390
45. Mantegna R, Stanley HE (2005) Scaling behavior in the dynamics of an economic index. Nature 376:46–49
46. Gabaix X, Gopikrishnan P, Plerou V, Stanley HE (2003) A theory of power-law distributions in financial market fluctuations. Nature 423:267

Books and Reviews

Challet D, Marsili M, Zhang Y-C (2005) Minority games. Oxford University Press, Oxford
Coolen AAC (2004) The mathematical theory of minority games. Oxford University Press, Oxford
Johnson NF, Jefferies P, Hui PM (2003) Financial market complexity. Oxford University Press, Oxford
Minority Game's website: http://www.unifr.ch/econophysics/minority/

Mobile Agents

Niranjan Suri[1], Jan Vitek[2]
[1] Institute for Human and Machine Cognition, Pensacola, USA
[2] Purdue University, West Lafayette, USA

Article Outline

Definition of the Subject
Introduction
Classification of Mobile Agent Capabilities
Theoretical Foundations of Mobility
Requirements Addressed by Mobile Agents
Components of a Mobile Agent System
Security
Survey of Mobile Agent Systems
Application Areas
Future Directions
Acknowledgments
Bibliography

Definition of the Subject

Mobile agents are programs that, with varying degrees of autonomy, can move between hosts across a network. Mobile agents combine the notions of mobile code, mobile

computation, and mobile state. They are location aware and can move to new network locations through explicit mobility operations. Mobile agents realize the notion of moving the computation to the data as opposed to moving the data to the computation, which is an important paradigm for distributed computing. Mobile agents are effective in operating in networks that tend to disconnect, have low bandwidth, or high latency.

Introduction

Advances in computer communications and computing power have changed the landscape of computing: computing devices ranging from the smallest embedded sensors to the largest servers are routinely interconnected and must interoperate. Their connections often are set up over untrusted and untrustworthy networks, with limited connectivity and dynamic topologies. The computational capacities of the devices as well as the communications bandwidth between the devices are in a state of constant change and users expect computer systems to dynamically adapt to such changes. Systems should opportunistically take advantage of new resources while at the same time transparently compensate for failures of systems and communication links. Moreover, the characteristics of the applications running on those devices are quite often dynamic, with new software added to the system at run time.

While many of the characteristics of distributed systems have changed, the tools for developing distributed software have not evolved. The majority of distributed programming is still being done in languages and environments that were designed either for uni-processor hardware systems or for static software systems in which the locations and functionality of all clients and servers can be specified a priori. This entry will look at different approaches using mobile agents, a new paradigm that eases the task of developing modern distributed systems. In addition, the entry will look at programming languages and middleware designed to support mobile agents, as well as the security mechanisms required by those languages and infrastructures.

Mobile agents are software agents with the additional capability to move between computers across a network connection. By movement, we mean that the running program that constitutes an agent moves from one system to another, taking with the agent the code that constitutes the agent as well as the state information of the agent. The movement of agents may be user-directed or self-directed (i. e. autonomous). In the case of user-directed movement, agents are configured with an itinerary that dictates the movement of the agents. In the case of self-directed movement, agents may move in order to better optimize their operation. Mobility may also be a combination of user- and self-directedness.

Classification of Mobile Agent Capabilities

Mobile agents encompass three basic capabilities: mobile code, mobile computation, and mobile state. These three capabilities are shown in Fig. 1 below. Each of the capabilities is an evolution of previously developed capabilities. The following subsections describe each capability.

Mobile Computation

Mobile computation involves moving a computation from one system to another. This capability is an evolution of remote computation, which allows a system to exploit the computational resources of another system over a network connection. One of the original mechanisms for remote computation was Remote Procedure Call (RPC). Java Remote Method Invocation (RMI) is another example of remote computation as are servlets and stored procedures.

The difference between mobile and remote computation is that mobile computation supports network disconnection. In a traditional remote computation model, the system requesting the service (the client) must remain connected to the system providing the service (the server) for the duration of the remote computation operation. Additionally, depending on the interface exposed by the server, an interaction can require an arbitrary number of messages between client and server. If network con-

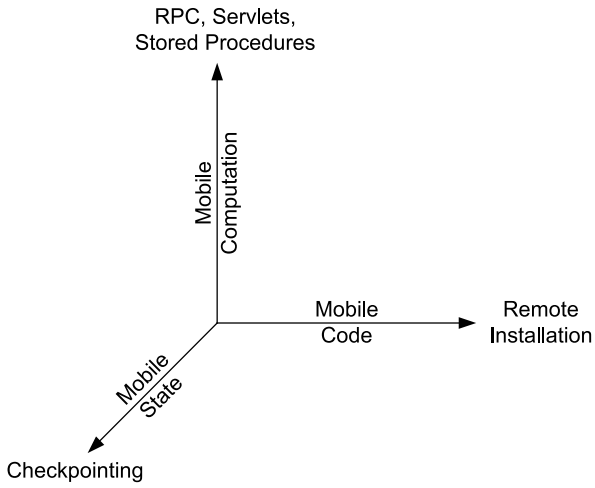

Mobile Agents, Figure 1
The Three Orthogonal Capabilities of Mobile Agents

nectivity is lost, the remote computation will become an orphaned computation that will either be terminated or whose results will be discarded. A mobile computation, on the other hand, is an autonomous entity. Once the computation moves from the first system (which may nominally be called the client) to the second system (the server), the computation continues to execute on the server even if the client becomes disconnected. The agent returns to the client with the results of the computation when (and if) the connectivity is recovered.

Mobile Code

Mobile Code is the ability to move code from one system to another. The code may be either source code that is compiled or interpreted or binary code. Binary code may further be either machine dependent or be some intermediate, machine-independent form.

Mobile code is used in other contexts besides mobile agents. For example, system administrators use mobile code in order to remotely install or upgrade software on client systems. Similarly, a web browser uses mobile code to pull an applet or script to execute as part of a web page.

Code may be mobile in two different ways: push and pull. In the push model, the system sending the code originates the code transfer operation whereas in the pull model, the system receiving the code originates the code transfer operation. An example of the pull model is a web browser downloading components such as applets or scripts. Remote installation is an example of the push model. Mobile agent systems use the push model of code mobility.

Pull mobility is often considered to be more secure and trustworthy because the host receiving the code is the one that requested the code. Usually, the origin of the request lies in some action carried out by a user of the system and hence pull mobility is superficially more secure. Push mobility on the other hand allows a system to send code to the receiving system at unexpected or unmonitored times. Hence push mobility is less trustworthy from a user's point of view. In practice the overwhelming majority of security exploits encountered in distributed systems originates in careless user actions such as executing attachments received via email.

Mobile code allows systems to be extremely flexible. New capabilities can be downloaded to systems on the fly thereby dynamically adding features or upgrading existing features. Moreover, if capabilities can be downloaded on demand, temporarily unused capabilities can also be discarded. Swapping capabilities on an as-needed basis allows systems to support small memory constrained devices. Discarding capabilities after use can also help improve system security.

Mobile State

Mobile state is an evolution of state capture, which allows the execution state of a process to be captured. State capture has been traditionally used to checkpoint systems to protect against unexpected system failure. In the event of a failure, the execution of a process can be restarted from the last checkpointed state thereby not wasting time by restarting from the very beginning. Checkpointing is thus very useful for long-running processes. Operating system research has investigated capturing entire process states, a variant of checkpointing, for load balancing purposes in the early 1980s, but that avenue of research proved to be a dead-end due to the coarse granularity of process and semantic problems due to the impossibility of capturing operating system resources such as open file descriptors.

Mobile state allows the movement of the execution state of an agent to another system for continued execution. The key advantage provided by mobile state is that the execution of the agent does not need to restart after the agent moves to a new host. Instead, the execution continues at the very next instruction in the agent.

Not all mobile agent systems provide support for state mobility [15]. The term strong mobility is used to describe systems that can capture and move execution state with the agent. Operationally, strong mobility guarantees that all variables will have identical values and the program counter will be at the same position. Weakly mobile-agent systems, on the other hand, usually support the capture of most of a program's data, but restart the program from a predefined program point and thus require some programmer involvement at each migration. The advantage of strong mobility is that the result of migrating is well defined and easier to understand, but its disadvantage is that it is much more complex to implement efficiently. The languages and systems that support strong mobility are Telescript [57], D'Agents, NOMADS, and Ara, while weak mobility is supported by a large number of mobile agent frameworks [1,3,5,16,35,39,40,45]. Results by Bettini and De Nicola suggest that strong mobility can be translated into weak mobility without affecting the application semantics [2]. The result is partial as it only works for single-threaded agents. Research is needed to be able to translate multi-threaded strongly mobile agents.

A recent advance in this area comes in the form of a modified Java-compatible VM. The IBM Jikes Research VM is designed to support pluggable Just-in-Time (JIT) compilers. Moreover, the VM is designed to allow re-

compilation of a method midstream, which requires that the state of the method be recoverable. The Mobile Jikes RVM [31] exploits this capability to provide migration of state but also provides good performance due to the Just-in-Time compiler.

The most important advantage provided by strong mobility is the ability to support external asynchronous migration requests (also known as forced mobility). This allows entities other than the agent (such as other system components, an administrator, or the owner) to request that an agent be moved. Forced mobility is useful for survivability, load-balancing, forced isolation, and replication for fault-tolerance.

Classification of Mobile Systems

The classification of mobile systems by Picco and Vigna distinguishes three broad approaches to mobility and summarizes the previous description of mobility technologies:

1. *Remote evaluation* – Remote evaluation technologies provide means for an application to invoke services on another node by specifying the code, as well as the input data, necessary to invoke the service. The code and input data are sent to the remote node, and the remote node then executes the code and sends the output data back to the client.
2. *Code on Demand* – This approach supports software components with dynamically loaded behavior. In this approach, code fragments are requested as they are needed, and dynamically compiled (if needed), verified and linked into a running system.
3. *Mobile Agents* – Mobile agents (The term 'mobile agent' is slightly misleading, as mobility is not restricted to agents, but can be used with any software component or program. Unfortunately, the literature does not differentiate between 'mobile agents' and 'mobile programs') strengthen code-on-demand with support for moving running computations. Rather than simply moving code (and possibly input data), mobile agents view a computation as a single entity and support the migration of a complete program to another node. This transfer is often seamless, so that the computation can proceed without disruption.

Remote Evaluation Remote evaluation is doubtlessly the simplest way to achieve mobility. This approach is often used for system administration tasks in which small programs written in a scripting language are submitted to hosts on a secure network. Stamos coined the name [36] to describe a technique where one computer sends another computer a request in the form of a program. The receiving computer executes the program in the request and returns the result back to the sending computer. A number of papers investigated this approach in the early 90's [34,37,38], but the only noteworthy infrastructure supporting this approach today is the SafeTCL scripting language [4,29]. The main drawback of scripting languages is that they are not suited to the development of large and reliable software systems because they often lack the basic software engineering features (e. g. encapsulation and data hiding) needed in large systems. Furthermore, the remote evaluation paradigm is confined to classical client/server settings and does not support detached operations well.

Code on Demand Code on demand is one of the main innovations of Sun Microsystems's Java programming language. This approach allows applications to be delivered piecemeal. The Java execution environment, called a Java Virtual Machine (JVM), is able to find and load any missing components at run time. These components are dynamically linked into the running system, and components that are never needed for a particular application run are never sent across the network, conserving network bandwidth. While dynamic loading is not a novel concept, the idea of allowing potentially untrusted content to be integrated into a running execution environment popularized the concept of 'safe programming languages'. The subsection on security will look in more detail at the requirement for safety. The success of Java owes as much to the safety features that were installed to ensure security as to its dynamic nature. Microsoft's .NET infrastructurez, and in particular the Common Language Runtime (CLR), provide a similar functionality, but that technology remains, as of this writing, untested and may not yet provide a comparable level of security as Java, though in the long run it is almost certainly going to play a major role. To summarize, the advantage of code-on-demand over remote evaluation is that a language such as Java is a general-purpose programming language with several mature implementations suited for building complex systems. The disadvantage of code on demand approaches is that software is not location-aware; in other words the code running in a Java system can not know where it is located nor is it able to trigger its own migration. Standards such as Java remote method invocation (RMI) extend the pure code-on-demand approach with the means to transfer data along with code under program control; they can thus be used as a basis to implement mobile agent systems but do not provide all of the functionality of a mobile agent system.

Mobile Agents Mobile software agents improve on previous approaches by bundling code and data into computational entities that can control their own mobility. Mobile agents programs can thus control their own deployment, perform load balancing, and program distributed applications. Mobile-agent infrastructures can be implemented as extensions to code-on-demand systems [1,3,5,6,16,25,35,40,45], as extensions to remote evaluation systems, or as new programming languages [39,56]. The advantage of building on an existing language such as Java is that the existing technology can be leveraged, but this comes at the price of some conceptual complexity, since the system designer must deal with inherent limitations of Java. For example, Java does not provide adequate resource management and process isolation facilities to allow untrusted computations to execute on a trusted machine. Another advantage of mobile-agent systems is that the infrastructure, not the programmer, is in charge of migrating the state and code needed by the computation. Finally, since computations are first-class entities in mobile-agent systems, they naturally can be associated with authority, access rights, and resources.

It is important to realize that, at a basic level, all of these approaches are equally powerful [30]. There is no distributed application that can be implemented with one technology and not any other [2]. Just as we now recognize that high-level programming languages increase programmer productivity, high-level mobility abstractions increase productivity even further. The goal of mobile-agent research is to find programming abstractions that are well suited to the tasks at hand, and provide well-engineered, efficient linguistic constructs to support these abstractions. In the long run, mobile-agent languages must be supplemented with automated tools for reasoning about programs and validating their properties (by static analysis, abstract interpretation or model checking). The goal of research in this arena is not only to provide the verification technology but also to design languages that are amenable to verification. Just as it is much easier to verify Java code than assembly, it is easier to verify programs that use mobility explicitly than systems in which mobility is implicit. In the long run we expect verification technologies to play an essential role to ensure correctness and provide the kind of security guarantees expected in critical information systems.

Theoretical Foundations of Mobility

A theoretical model of a system allows formal reasoning about the system. Formal reasoning can be used to establish guarantees about the behavior of the system. Understanding the semantics of mobile computation is essential for reasoning about mobile agents. Reasoning about mobility can, in turn, yield guarantees about the correctness of mission critical software. Researchers have explored a number of theoretical models based on process calculi with encouraging results. Two of the approaches that have been studied in this direction are Cardelli and Gordon's ambient calculus [11] and Vitek and Castagna's Seal calculus [50]. These models abstract both:

- *Logical mobility* (mobility of programs) and
- *Physical mobility* (mobility of nodes, such as handheld devices).

The results obtained thus far include a number of type systems for controlling agent mobility [9,10,12], and a logic for stating properties of agent programs [10]. These formalizations have wider applicability as shown by an application of the ambient calculus as a query language for XML [8]. The research on foundations of mobility is actively continuing; in the long run theoretical results can be expected to feed back into languages and infrastructures. The development of language-level abstractions will simplify the task of writing mobile software agents.

Requirements Addressed by Mobile Agents

A paradigm is as much defined by the features that are excluded as the features that are included. We now present a list of the key required features as well as some features that we explicitly do not expect mobile agent systems to support. Mobile agents, like most modern distributed systems, have to address five issues that require infrastructure support:

No global state The physical size of the network and the number of hosts that can participate in a distributed computation must scale to global networks of millions of nodes (of which tens of thousands or more might be participating in a single distributed computation). At this size, there can be no assumption of shared state in the programming model, nor can there be algorithms that require synchronization, or that rely on up to-date information.

Explicit localities Fluctuations in bandwidth, latency, and reliability are so common that they cannot be hidden. Location of resources and of the computation that accesses those resources must be explicit in the programming model.

Restricted connectivity Failures of machines and communication links can occur without warning or detection. Some of these failures may be temporary as machines may be restarted and connections reestablished,

but others may be permanent. Thus, at any time, a computation may communicate with only a subset of the network, or be fully disconnected.

Dynamic configuration The network topology, both in the physical sense and in the logical sense of available services, changes over time. New hosts and communication links appear with no advance notice, while other hosts disappear and reappear under a new name and address. Applications should be able to adapt to these changes and dynamically reconfigure themselves. Without the ability to dynamically reconfigure its components, applications will have difficulty adapting to changes in locality and connectivity.

Security Since security requirements vary from application to application, basic security mechanisms must be provided by the underlying infrastructure (type safe programming language, checked array access, access control mechanisms) and must be extended with tools for automatic validation of security properties (model checking, program analysis, proof-carrying code).

These five requirements – absence of global state, explicit localities, restricted connectivity, dynamic configuration, and security – drive many of the design decisions behind current mobile agent systems. Mobile agent architectures support the above requirements. Mobile agents do not define any notion of shared or global state. Furthermore, mobile agents do not require that hosts be connected while the agents execute. In fact mobile agents have been repeatedly advocated for disconnected operations (restricted connectivity). Finally, mobile agents, through the use of mobile code, support dynamic configuration. Security, the last issue, remains a challenge that must be addressed through infrastructural tools and services.

Components of a Mobile Agent System

Figure 2 below shows the components of a mobile agent system. This is not an exhaustive list and not all of the components shown are essential to a mobile agent system.

Every mobile agent system includes an execution environment that is responsible for receiving and hosting mobile agents. Often, the execution environment runs in the background as a daemon or service so that it is always available to receive agents. The execution environment provides a layer of separation between the mobile agent and the host platform, which is important for security.

Execution environments communicate with each other to provide mobility and messaging between agents. These protocols may be built on top of other network communication mechanisms provided by the infrastructure.

Most mobile agent systems use an interpreted or partially compiled language for the mobile agent code. In such cases, the execution environment contains an interpreter (or virtual machine, in the case of Java)

The remaining components in the execution environment are capabilities provided by the infrastructural layer. Not all of them are necessary for every mobile agent system.

Security

The main technical, and social, obstacle to approaches based on mobile software agents is security. Not only must researchers devise technical solutions, but also users and organizations must become confident enough in those solutions to permit foreign programs to migrate to and execute on their machines [14,48]. If the organization responsible for a database is not convinced of the quality of the

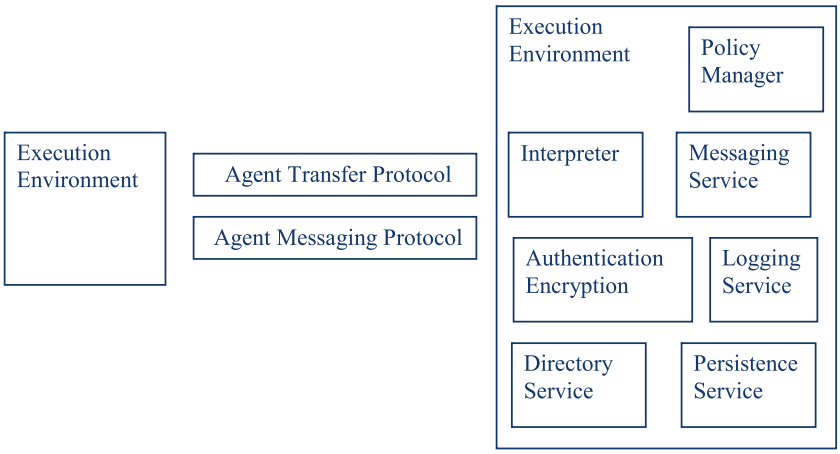

Mobile Agents, Figure 2
Components of a Typical Mobile Agent System

security mechanisms, the department will never allow mobile agents to visit the database. Although a mobile-agent application can still function (by having its agents access the database from across the network), the application will use more network bandwidth and suffer higher latencies.

Thus the security and containment of untrusted mobile code, and the objective analysis of proposed security solutions, is a critical research area. When a host receives mobile code, it ideally should evaluate the security implications of executing that particular code, but at the least, it must determine the trustworthiness of the agent's sender (and programmer). Failure to properly contain mobile code may result in serious damage to the host or in the leakage of restricted information. Such damage can be malicious (e. g. espionage or vandalism) or unintentional (programming failures or unexpected interactions with other system components). Other consequences of failing to contain mobile code include denial-of-service attacks and the infiltration of privileged networks through a downloaded Trojan horse or virus [22,28,48,52].

The symmetry of mobile-agent security concerns is remarkable as both the agent and the environment in which it executes must be protected from each other. Through purposeful engineering on the part of its developer, an agent may seek to obtain restricted data from the host on which it is running on or damage the host in some way. On the other hand, a host may seek to steal data from or corrupt the agents that migrate to it [33]. In the civilian environment, a (dishonest) company might gain an economic advantage over a competitor via a malicious agent or host, while in more critical environments such as in the military, an adversary might gain a strategic or tactical advantage during an armed conflict.

To relate the security issues to Java programming, we compare an agent to an Applet, a Java program "embedded" inside a Web page and downloaded and executed on a user's machine whenever that user browses the web page. The applet runs within an environment composed of several layers, the first layer is the Java Development Kit (JDK) and its class libraries, the second layer consists of the Java Virtual Machine, the third layer is the operating system, and the last layer is the host device itself. The distinction between these layers is important, since some layers may be easier to subvert than others. For instance, an application may trust a server that belongs to a known organization, but may not trust the libraries found on that server. In Java, this trust mismatch can occur if some of the classes against which an applet is linked have been downloaded from the network [18,54]. There are two different threats that must be considered when attempting to secure mobile-agent applications:

Exogenous threats Attacks occurring outside of the mobile-agent system. For example, if a host is running both a mobile-agent system and a Web server, an adversary might attack the host via a "standard" Web server exploit, and gain access without every attacking the mobile-agent system itself.

Endogenous threats Threats specific to a mobile-agent system.

- *Horizontal hostility* (malicious agents): Attacks between agents running on the same host in which an agent tries to disrupt the execution of other co-located agents.
- *Vertical hostility* (malicious agents & malicious hosts): Attacks against an agent by the execution environment, as well as attacks against the environment by an agent.

In the remainder, we consider only endogenous threats, as they are specific to mobile agents. There are two different viewpoints to take into account:

- For a host, it is necessary to provide protection mechanisms so that agents cannot attack each other (horizontal protection) or the host itself (vertical protection);
- For an agent, it may be necessary to protect it from attacks initiated by the host (hostile host) and other agents (horizontal).

We now consider each issue in turn.

Malicious Agents

A number of techniques have been used in the past to place protection boundaries between so-called "untrusted code" moved to a host and the remainder of the software running on that host. Traditional operating systems use virtual memory to enforce protection between processes [13]. A process cannot read or write another processes' memory and communication between processes requires traps to the kernel. By limiting the traps an untrusted process can invoke, it can be isolated to varying degrees from other processes on the host. However there is usually little point in sending a computation to a host if the computation cannot interact with other computations there, load balancing being the only exception [27]. In the context of mobile agent systems, an attractive alternative to operating system protection mechanisms is to use language-based protection mechanisms [52]. The attraction of language-based protection is twofold: precision of protection and performance. Language-based mechanisms allow access rights to be placed with more precision than traditional virtual-

memory systems, and the cost of cross-protection boundaries can often be reduced to zero, since checking is moved from runtime to the language compiler [17,23].

Safe languages The requirement for language-based security is, first and foremost, safe languages. A safe language is a language that enforces memory safety and type safety. In other words, a safe language does not permit arbitrary memory modifications, and carefully constrains how data of one type is transformed into data of another type. The objective of a safe language is to permit reasoning about security properties of programs at the source level in a compositional manner. It should be possible to check the code with automatic tools to obtain the guarantee that the mobile agent is not malicious. For this to be the true, it is necessary to produce assured implementations of safe languages, that is, implementations that do not contain hidden security vulnerabilities. Well-known examples of safe languages include Java and SafeTCL [18,29]. The Telescript agent language is a case of a safe language explicitly designed for secure mobile code [44]. Traditional languages, such as C, do not ensure memory or type safety, and thus, it is much more difficult to obtain trust in agents written in those languages. Even in safe language, there are many opportunities for security exploits. After many years, the research community is closer to producing assured implementations of the Java programming language, but much work remains. The survey by Moreau and Hartel lists several hundred papers on formalizing aspects of Java [20]. The difficulty in obtaining a clear specification of all aspects of Java underscores the need for research in semantics and formal techniques without which there can be no hope of obtaining any assurance.

Sandboxing Protection against vertical attacks is achieved by enforcing a separation between the user code and the system, a technique popularized by the well-known Java-sandbox security model; related approaches have been used in operating systems [18,19,24]. In this model user code runs with restricted access right within the same address space as the system code. Security relies on type safety, language access control mechanisms and dynamic checks. Over the years, a number of faults were discovered and fixed in this model [17,23,54,55]. The sandbox model is a basis for building more powerful security architectures that are suited to agent systems [18]. Sandboxing alone does not provide protection against horizontal attacks. For this, it is necessary to extend the protection model to include protection domains, which constrain how one agent can interact with another. Protection domains can be constructed in a safe language by providing a separate namespace for each component. Fully disjoint namespaces are not desirable as they result in disjoint applications [27]. Instead, if mobile agents must interact, some degree of sharing among namespaces is necessary. Several research systems have tried to provide better isolation of Java applications [7,53], but these attempts achieved limited success due to the constraint of working above commercial Java Virtual Machines (which allow only certain kinds of extensions to Java's basic security mechanisms). In the future, protection domains must be integrated into the Java Virtual Machine definition.

Denial of service Mobile agents can mount denial of service attacks by using an excessive amount of CPU or memory. An environment for mobile agents therefore must provide support for tracking memory and CPU usage, as well as support for termination. Termination implies stopping all threads of an agent and reclaiming its memory. Current Java systems fail to protect the Java Virtual Machine against denial of service attacks, since they support neither resource accounting nor full agent termination. Providing efficient accounting and termination support in a language-based system remains an open research problem.

Beyond safe languages Safe languages must ensure that an agent's code obeys certain well-formedness rules. In the case of Java, this assurance is obtained by verifying the bytecode of incoming agents with a complex data flow analysis algorithm [17] and by imposing some constraints on how programs may be linked [23]. A large body of research on proof-carrying code [28] is trying to broaden the set of agent languages to traditional unsafe languages such as C. Proof-carrying code associates a security proof with each program. The host need only check that the proof matches the program to determine whether the program obeys the desired security properties. Checking a proof against a program is computationally much easier than analyzing the code directly to generate the proof, making proof-carrying code an attractive approach. This direction of research is encouraging, as it may allow the expression of complex security properties, and verification of a program's compliance with those properties.

Malicious Hosts

In mobile-agent computing, an agent's owner must be able to trust that it is not subverted when visiting a series of servers, some of which may have been compromised and

made capable of malicious action against the agent [33]. Malicious servers are a particularly difficult problem, since the server must have access to all of the agent's code in order to execute it. A small body of research has attempted to solve this problem. The solutions fall into the following categories:

1. Code Signing,
2. Replication,
3. Partial Result Authentication Codes,
4. Proof Verification,
5. Code Obfuscation, and
6. Encrypted Functions.

We will assess each approach in the following paragraphs.

Code Signing can be used to protect agents from malicious hosts by attaching digital signatures to the code of the mobile agent. Code signing is being used by Sun Microsystems and Microsoft to provide guarantees of authenticity for downloaded code. The technology can be used to ensure that a server has not altered the code of an agent while in transit. Code signing does not protect the agent's data from being modified, nor does it prevent the server from accessing the information contained in the agent, but it does provide a basic level of assurance that it is essential for some applications. Furthermore, in a network in which servers can not be compromised and agents come from a single source code signing may be the best solution to security.

Replication was studied as a general method for mobile agent computation security, marrying some ideas from the fields of fault tolerance and cryptography [32,51]. The approach relies on the replication of agents and servers. The same agent computation is performed on several servers. Voting can then be used to move from one phase of a distributed computation to the next. While replication enjoys some pleasing theoretical properties, it is heavily restricted in practice. It supposes that computations are deterministic, and that several servers with the same resources are available. The connectivity assumptions also are not appropriate in scenarios involving unreliable networks.

Partial Result Authentication Codes are very similar to message authentication codes (MAC). Instead of authenticating the origins of a message, however, they authenticate the correctness of an intermediate agent state or partial result. For example, if we consider the values of selected program variables at some point during execution, we can determine whether those values could possibly have arisen from normal program execution. If not, the program has been altered in some way. PRACSs are computationally cheaper than digital signatures and have slightly different security properties (forward integrity): if an agent visits n servers and some server in m ($m < n$) is malicious, the results of servers 1 to $m - 1$ cannot be falsified. In some scenarios, mobile agents do not have intermediate results. Nevertheless, this approach can be used to ensure that results of a disconnected query are truthful [21].

Proof Verification is an approach in which a digitally signed trace of an agent's computation is returned along with the result. This trace can then be validated – a malicious host would affect the agent by changing the results and thus producing a trace that does not correspond to a valid computation [49]. Although techniques for producing compact traces have been developed, the size and complexity of the trace remains an issue.

Code Obfuscation aims at protecting a mobile agent's functioning by making it very hard to divert the agent's execution in a meaningful way. This is achieved by transforming the code such that automated reverse engineering cannot be applied. This prevents a malicious host from locating the places in the code that should be modified or that should be executed in a non-conformant way. Also, variables can be split all over the program in order to make simple read-outs impossible. Although there are many tools and also commercial products that use code obfuscation, especially in the field of digital rights management, recent theoretical results point at the impossibility of achieving perfect obfuscation [26].

Secure Coprocessors involve building a trusted execution environment for agents within a secure coprocessor [58,59]. This approach is based on tamper-proof hardware and public key infrastructures. Some experimental systems have been designed, but not validated. Secure co-processors have the potential of providing appropriate security for mobile-agent programs, but at the cost of upgrading to more expensive hardware. Secure coprocessors will be useful in some applications, but not in all (or even the majority). However, trusted coprocessors might be the only hard security anchor available today for securing mobile agent applications.

Encrypted Functions that can be executed in their encrypted form are a software-only cryptographic approach to the malicious host problem. If available, this would be the ideal way to protect any mobile agent and its payload. The approach of computing with encrypted functions was demonstrated in [33], another system was proposed in [46]. The conclusion is that for special functions it is possible to let a mobile agent

Mobile Agents, Table 1
Selected popular mobile agent systems and their key properties

Agent System	Mobility support	Base language	Host mobility	Quality of implementation	Availability	Level of support	Special Features
Telescript	Mobile agent (strong)	Telescript	No	Product	Discontinued	None	First MA system
NOMADS	Mobile agent (strong)	Java	No	Prototype	Free	Medium	Resource Control
Java	Code on demand	Java	No	Product	Commercial	High	
SafeTCL	Remote evaluation	Tcl	No	Product	Open source	High	
D'Agents	Mobile agent (strong)	Java, Tcl, Scheme	No	Prototype	Open source	Medium	Multi-language
JavaSeal	Mobile agent (weak)	Java	No	Prototype	Open source	Low	Seal Calculus
Mole	Mobile agent (weak)	Java	No	Prototype	Open source	Low	
Aglets	Mobile agent (weak)	Java	No	Product	Open source	Medium	
Lime	Mobile agent (weak)	Java	Yes	Prototype	Open source	Medium	Coordination via Tuple-spaces
Messenger	Mobile code (weak)	M0	No	Prototype	Open source	Low	

protect itself from a malicious host without having to rely on trusted hardware or on-line help from remote agents. However, the solutions proposed seem to be impractical for today's standards and no implementation has been reported so far.

Survey of Mobile Agent Systems

To summarize this entry, we now review some of the popular mobile-agent systems and classify them according to the following characteristics: the type of mobility supported by the language or system, the language(s) in which agents are written, whether host mobility (mobile devices) is supported, and the quality, availability, and current level of support of the implementation. The quality of implementation column discriminates products from research prototypes. The availability column indicates which systems can be freely used, and which require a license. Finally, the level of support column indicates whether the project is still active, and whether assistance is forthcoming.

The general conclusions that emerge are that weak mobility is by far the predominant approach. The only commercial strongly mobile system (Telescript) was discontinued several years ago, and the other two strongly mobile system (D'Agents and NOMADS) are university research prototypes. Java is the most popular agent implementation language, due to both the popularity of the language and its support for dynamic loading and advanced security features. Most available systems are research prototypes, but they have the advantage of being open source and thus can be used as a starting point for further development. A number of these projects have an active developer community, an important factor for the adoption of an infrastructure, although it is important to note that most open-source systems do not enjoy the same kind of support as commercial products.

While the above table is not an exhaustive list of existing mobile-agent systems (over 100 such systems have been implemented in the last five years), it provides a good overview of the most influential systems. The main conclusion to draw is that Java is the common thread in most current mobile-agent implementations. Another, more recent technology is Microsoft's .NET with the Common Language Runtime. Like Java, .NET is based on a virtual machine and supports the dynamic loading of programs. The suitability of .NET to mobile-agent applications has not yet been completely evaluated.

Application Areas

Information retrieval was the original application area envisioned for mobile agents. In bandwidth sensitive environments, moving the mobile agent to the data, where the agent can select a small desired subset, is more efficient than moving large amounts of data to the user. Mo-

bile agents allow bandwidth conservation and latency reduction in many information-retrieval and management applications. A mobile agent that performs a multi-step query against multiple databases can be dispatched close to the location of the databases, avoiding the transmission of intermediate results across the network. Similarly, in an unreliable network environment, the same mobile agent can continue its query task even if the network goes down temporarily.

But, mobile agents can support or improve performance in many other application areas. Mobile agents can relocate themselves during execution, an essential property for reactive and adaptive systems that must respond to changing execution environments. As soon as a change in operating conditions is detected, a mobile-agent application can reconfigure itself to relocate its computations away from a physical attack on the network or closer to a critical database after a drop in network bandwidth. Mobility also applies to the application's data itself, enabling new styles of pro-active applications where system wide actions have to be performed even before any clients are around. An example is the pre-caching of datasets to remote geographical locations to ensure instant access to essential information.

Mobile agents, with their capability to move code, allow new capabilities to be pushed dynamically to platforms. This capability is useful when existing systems need to be retasked or used for purposes not originally intended when they were deployed. Migration of capabilities is also important to accommodate changing circumstances and environmental conditions.

Mobile agents also support disconnected operations. In a mobile system, a client can send agents to a server before disconnecting. The agents can then perform their task while the client is unreachable and communicate results back whenever connectivity is regained. Similarly, a server might send a component to a handheld or other portable device to further reduce connectivity requirements.

Even if the network is stable, mobility still allows bandwidth conservation and latency reduction. For example, if a client application needs to perform a complex multi-step query, it can send the query code to the network location of the databases, avoiding the transmission of intermediate results across the network. Although the database developers could add a new database operation that performed the complex query, it is unreasonable to expect that developers can predict and address *every* client need in advance. Mobile agents allow a client application to make efficient use of network resources even when the available services expose low-level, application-independent interfaces.

In all four cases – changing network conditions, disconnected operation, bandwidth conservation, and latency reduction – the common thread is dynamic deployment and reconfiguration. Traditional programming languages constrain designers to commit to a particular system structure at build time. The choice whether a particular service is implemented on the client or server side must be made early and cannot be revisited if some of the initial assumptions about the application turn out to be invalid. Mobile agents, on the other hand, decouple system design from system deployment, and turn control over deployment to the applications themselves, allowing much more flexible design patterns. An application can deploy its components to the most attractive network locations and redeploy those components when network conditions change, leading to more efficient use of available resources and faster completion times.

Additional application areas are discussed below:

Distributed Sensor Grids

Mobile agents can migrate to key locations in a network of autonomous sensors and then filter the collected sensor data to reduce bandwidth requirements and implement local management policies. Mobile agents are particularly useful when it is not possible to pre-install stationary agents on all of the sensors. In particular, the filtering algorithm, which can be of arbitrary complexity, may depend on the phenomenon being observed and change over time. For example, different filtering agents can be used to achieve different levels of accuracy. For high-end sensors, onboard processing is possible, and agents can be sent to the sensors themselves; in other cases, agents can be sent to routers or gateway machines within and at the edge of the sensor field. The agents can be relocated on the fly, allowing the sensor application to optimize the placement of its management and filtering code with respect to current network loads.

By using automated monitoring algorithms, mobile agents can act as customized monitors that wait for phenomena of interest to be observed at distributed sensor locations and then notify human operators. Such a capability can significantly reduce the workload of human operators.

In addition to acting as on-line filters, mobile agents can also help perform data mining on off-line sensor data. In particular, mobile agents can move to several sensors and efficiently correlate information across the multiple sensors in order to classify observed phenomenon.

Mobile agent technology can be particularly useful in situations where both the sensors as well as the client systems are computationally weak and constrained by bat-

tery. This is a common occurrence in sensor networks where small sensors (such as unattended ground sensors) are being tasked by users with small PDA or cell phone client devices. Mobile agents carrying specific filtering, fusion, or other algorithms can be pushed from the PDA or cell phone to opportunistically discovered nodes in the network fabric thereby supporting the weak sensor and client platforms. Given the adaptive capabilities of mobile agents, they can also relocate themselves to other intermediate nodes when network configurations change [42,47].

Unmanned Autonomous Systems

Unmanned autonomous systems require varying degrees of remote tasking and monitoring but often operate in extreme circumstances with varying network connectivity. Mobile agents allow the dynamic configuration of the software deployed on unmanned autonomous systems. New capabilities such as new algorithms, missions, and functions that were not anticipated when the system was originally designed and deployed can be dynamically deployed while vehicles and sensors are in the field. This functionality is already used in outer space exploration projects where communication latencies and physical inaccessibility to devices require mobile-agent like approaches. Fully embracing mobile agent architectures will enable finer granularity and improve ease of use. Mobile agent technologies further allow exchange of functions between platforms, providing a very efficient and localized software update mechanism.

Certain kinds of autonomous vehicles such as undersea vehicles lose network connectivity while submerged. The capability of mobile agents to support disconnected operation is important under these circumstances.

Space exploration vehicles such as the Mars rover present a variation of the network disconnection problem – extremely long network latencies. Mobile agents, by moving themselves to the remote nodes, overcome latency problems.

Finally, mobile agents can reduce the bandwidth requirements for relaying data from unmanned vehicles by applying the necessary filtering and fusion algorithms on the platform or close to the platform.

High Availability Systems

Some large-scale distributed systems must be able to evolve and cannot tolerate downtime. Mobile agents provide a technology for dynamically upgrading such distributed systems with new procedures without interrupting their operation. Furthermore, mobile agents can provide additional fault tolerance, since an agent can be duplicated at any point of its computation and stored on secondary storage or sent to another platform.

Mobile-agent technology provides an attractive way to implement distributed monitoring tools that enforce nonlocal security policies over large-scale networks. Such nonlocal monitoring can detect distributed denial of service (DDOS) and other large-scale attacks and help pinpoint the source of those attacks. Through the use of mobile agents, this monitoring functionality can be deployed dynamically in response to previous attacks. For example, if a network is under extensive DDOS attack, the DDOS monitoring code can be dispatched on the fly to a larger number of machines. Similarly, if a network comes under a previously unknown attack, mobile agents can be implemented and deployed to detect that particular attack on the fly. The ability to dynamically distribute existing and new monitoring code makes networks more robust to cyber-attack, including attacks that are encountered for the first time.

Agile Computing

Agile computing may be defined as opportunistically (or on user demand) discovering and taking advantage of available resources in order to improve capability, performance, efficiency, fault tolerance, and survivability [43]. The term agile is used to highlight both the need to quickly react to changes in the environment as well as the ability to take advantage of transient resources only available for short periods of time. Agile computing is a promising approach to building information systems that need to operate in highly dynamic environments..

Mobile agents are one approach to building agile computing systems. Mobility of code is important in order to be opportunistic. If an idle resource is found, the likelihood that the system already has the code for a particular service or capability is small. Mobile code can be used to dynamically push the necessary services to systems.

State migration is also important for the purpose of achieving a high degree of agility. Without state migration, it would be difficult to quickly and dynamically move computations to and away from systems as resource availability changes.

Future Directions

Active research in mobile agents has petered out, with most researchers migrating to other areas. The ideas and technology heralded by mobile agents were extensions of previous concepts that were integrated together. For example, mobile computing extended remote invocation and

mobile state extended process migration. To some extent, the research area of mobile agents ended but the ideas continue to exist and continue to evolve and be used, but under different labels.

The field of mobile agents also suffered from an identity crisis caused by the use of the word agent. Researchers in the areas of Autonomous Agents and Multi Agent Systems regarded mobile agents as primarily belonging to distributed systems and networking. On the other hand, the distributed systems community did not like the connotations of the word "agent" and hence did not adopt the terminology of mobile agents, even though the concepts introduced are important and tend to get used anyway.

Much of the early research in the area of mobile agents was published as part of the ECOOP Mobile Object Systems workshop series, the Mobile Agents conferences (temporarily renamed Agent Systems and Applications / Mobile Agents), and to a lesser extent the Autonomous Agents and Multi–Agent Systems conference. The last Mobile Agents conference was held in 2002. The Agents, Interactions, Mobility, and Systems (AIMS) track at the ACM Symposium on Applied Computing was the last venue for publishing mobile agents-related papers. A collection of these papers, revised and extended, were published in a special issue entitled Mobile Software Agents in the Scalable Computing: Practice and Experience journal. A recent book on mobile agents [41] provides a review and survey of mobile agents and concepts and can serve as a reference or a textbook for a course on mobile agents.

One key unsolved research question with mobile agents continues to be protection of the mobile agent that is executing on a remote untrusted host. This continues to be a challenge with no promising solutions on the horizon.

Acknowledgments

Portions of this chapter were initially written for the Information Technology Assessment Consortium (ITAC) at the Institute for Human & Machine Cognition (IHMC) as part of a report entitled Software Agents for the Warfighter, sponsored by NASA and DARPA, and edited by Jeffrey Bradshaw at IHMC.

Bibliography

Primary Literature

1. Baumann J, Hohl F, Rothermel K, Straßer M (1998) Mole – concepts of a mobile agent system. WWW Journal, Special issue on Appl Tech Web Agents (to appear)
2. Bettini L, De Nicola R (2001) Translating strong mobility into weak mobility. Mobile Agents
3. Binder W (2001) Design and implementation of the J-SEAL2 mobile agent kernel. In: The 2001 Symposium on Applications and the Internet (SAINT-2001), San Diego, January 2001
4. Borenstein NE (1994) E-mail with a mind of its own: The SafeTcl language for enabled mail. In: Proceedings of IFIP International Conference, Barcelona, Spain, 1994
5. Bouchenak S (1999) Pickling threads state in the Java system. In: 3rd European Research Seminar on Advances in Distributed Systems (ERSADS'99), Madeira Island, Portugal, April 1999
6. Bryce C, Vitek J (1999) The javaseal mobile agent kernel. In: Milojevic D (ed) Proceedings of the 1st International Symposium on Agent Systems and Applications, Third International Symposium on Mobile Agents (ASAMA'99), Palm Springs, 9-13 May 1999. ACM Press, pp 176–189
7. Bryce C, Vitek J (2002) The JavaSeal mobile agent kernel. Auton Agents MultiAgent Syst
8. Cardelli L, Ghelli G (2001) A query language based on the ambient logic. Programming languages and systems. In: 10th European Symposium on Programming, ESOP 2001
9. Cardelli L, Gordon AD (1999) Types for mobile ambients. In: Proceedings of the 26th ACM Symposium on Principles of Programming Languages, 1999. pp 79–92
10. Cardelli L, Gordon AD (2000) Anytime, anywhere. Modal logics for mobile ambients. In: Proceedings of the 27th ACM Symposium on Principles of Programming Languages, 2000. pp 365–377
11. Cardelli L, Gordon AD (2000) Mobile ambients. TCS special issue on coordination. D Le Métayer
12. Cardelli L, Ghelli G, Gordon AD (2000) Types for the ambient calculus. In: I&C special issue on TCS'2000
13. Colusa Software (1995) Omniware: a universal substrate for mobile code. White paper, Colusa Software. http://www.colusa.com
14. Farmer WM, Guttman JD, Swarup V (1996) Security for mobile agents: Issues and requirements. In: National Information Systems Security Conference. National Institute of Standards and Technology
15. Fuggetta A, Picco GP, Vigna G (1998) Understanding code mobility. IEEE Trans Softw Eng
16. Funfrocken S (1998) Transparent migration of Java-based mobile agents: Capturing and reestablishing the state of Java programs. In: Proceedings of the Second International Workshop on Mobile Agents, September 1998. Lecture Notes in Computer Science, no 1477. Springer, Stuttgart, pp 26–37
17. Goldberg A (1998) A specification of java loading and bytecode verification. In: Proceedings of the Fifth ACM Conference on Computer and Communications Security, 1998
18. Gong L, Mueller M, Prafullchandra H, Schemers R (1997) Going beyond the sandbox: An overview of the new security architecture in the Java Development Kit 1.2. In: Proceedings of the USENIX Symposium on Internet Technologies and Systems, Monterey, California, December 1997
19. Grimm R, Bershad BN (1999) Providing policy-neutral and transparent access control in extensible systems. In: Vitek J, Jensen C (eds) Secure internet programming: Security issues for distributed and mobile objects. Lecture Notes in Computer Science, vol 1603. Springer, Berlin, pp 117–146
20. Hartel PH, Moreau LAV (2001) Formalizing the safety of Java,

the Java Virtual Machine and Java Card. ACM Comput Surv (to appear)
21. Hohl F (1997) Time limited blackbox security: protecting mobile agents from malicious hosts. Mobile Agent Security. Lecture Notes in Computer Science. Springer, Berlin
22. Jaeger T (1999) Access control in configurable systems. In: Vitek J, Jensen C (eds) Secure internet programming: Security issues for distributed and mobile objects. Lecture Notes in Computer Science, vol 1603. Springer, Berlin, pp 117–146
23. Jensen T, Le Metayer D, Thorn T (1998) Security and dynamic class loading in Java: A formalization. In: Proceedings of the 1998 IEEE International Conference on Computer Languages, May 1998, pp 4–15
24. Jones MB (1999) Interposition agents: Transparently interposing user code at the system interface. In: Vitek J, Jensen C (eds) Secure internet programming: Security issues for distributed and mobile objects. Lecture Notes in Computer Science, vol 1603. Springer, Berlin, pp 117–146
25. Lange D, Oshima M (1998) Programming and deploying Java Mobile Agents with aglets. Addison Wesley
26. Loureiro S (2001) Mobile code protection. Ph D thesis, Sophia Antipolis
27. Malkhi D, Reiter MK, Rubin AD (1998) Secure execution of Java applets using a remote playground. In: Proc of the 1998 IEEE Symp on Security and Privacy, Oakland, May 1998, pp 40–51
28. Necula GC, Lee P (1998) Safe untrusted agents using using proof-carrying code. In: Vigna G (ed) Special issue on mobile agent security, vol 1419 of Lect. Notes in Computer Science. Springer, pp 61–91
29. Ousterhout JK, Levy JY, Welch BB (1997) The Safe-Tcl Security Model. Technical report. Sun Microsystems Laboratories, Mountain View. Online at http://research.sun.com/techrep/1997/abstract-60.html
30. Puliato B, Riccobene S, Scarpa M (1999) An analytical comparison of the clientsever, Remote evaluation, and Mobile Agents paradigms. In: Proc. ASA/MA'99, pp 278–292, October 1999
31. Quitadamo R, Cabri G, Leonardi L (2006) Enabling Java mobile computing on the IBM Jikes Research Virtual Machine. In: The International Conference on the Principles and Practice of Programming in Java 2006 (PPPJ 2006). ACM Press, Mannheim
32. Roth V (1999) Mutual protection of cooperating agents. In: Vitek J, Jensen C (eds) Secure internet programming: Security issues for distributed and mobile objects. Lecture Notes in Computer Science, vol 1603. Springer, Berlin, pp 117–146
33. Sander T, Tschudin CF (1998) Protecting Mobile Agents Against Malicious Hosts. In: Vigna D (ed) Mobile Agent and Security. Lecture Notes in Computer Science, vol 1419. Springer, Berlin
34. Segal ME (1991) Extending dynamic program updating systems to support distributed systems that communicate via remote evaluation. In: Proc. International Workshop on Configurable Distributed Systems, 1991, pp 188–199
35. Sekiguchi T, Masuhara H, Yonezawa A (1999) A simple extension of Java language for controllable transparent migration and its portable implementation. In: Coordination Languages and Models. Lecture Notes in Computer Science, pp 211–226
36. Stamos JW (1986) Remote evaluation. Ph D thesis, Massachusetts Institute of Technology. Technical report MIT/LCS/TR-354
37. Stamos JW, Gifford DK (1990) Implementing remote evaluation. IEEE Trans Softw Eng 16(7)
38. Stamos JW, Gifford DK (1990) Remote evaluation. ACM Trans Program Lang Syst 12(4):537–565
39. Suri N, Bradshaw JM, Breedy MR, Groth PT, Hill GA, Jeffers R, Mitrovich TR, Pouliot BR, Smith DS (2000) NOMADS: Toward an environment for strong and safe agent mobility. In: Proceedings of Autonomous Agents '2000, Barcelona, Spain. ACM Press, New York
40. Suri N, Bradshaw JM, Breedy MR, Groth PT, Hill GA, Jeffers R (2000) Strong mobiling and fine-grained resource control in NOMADS. Agent Systems, Mobile Agents, and applications. Lecture Notes in Computer Science, vol 1882. Springer, Berlin
41. Suri N et al (2003) Applying agile computing to support efficient and policy-controlled sensor information feeds in the army future combat systems environment. In: Proceedings of the U.S. Army 2003 Annual Collaborative Technology Symposium
42. Suri N, Bradshaw J, Carvalho M, Breedy M, Cowin T, Saavedra R, Kulkarni S (2003) Applying agile computing to support efficient and policy-controlled sensor information feeds in the army future combat systems environment. In: Proceedings of the Collaborative Technologies Alliance Conference (CTA 2003), College Park
43. Suri N, Bradshaw J, Carvalho M, Cowin T, Breedy M, Groth P, Saavedra R (2003) Agile computing: Bridging the gap between Grid computing and Ad-hoc Peer-to-Peer resource sharing. In: Proceedings of the 3rd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid), 2003
44. Tardo J, Valenta L (1996) Mobile agent security and telescript. In: Proceedings of IEEE COMPCON, February 1996
45. Truyen E, Robben B, Vanhaute B, Coninx T, Joosen W, Verbaeten P (2000) Portable support for transparent thread migration in Java. In: Proceedings of the Joint Symposium on Agent Systems and Applications / Mobile Agents (ASA/MA), September 2000, pp 29–43
46. Tschudin C The messenger environment M0 – A condensed description. In: Vitek J, Tschudin C (eds) Mobile object systems. Lecture Notes in Computer Science, vol 1222. Springer, Berlin
47. Tschudin C, Lundgren H, Gulbrandsen H (2000) Active routing for Ad Hoc Networks. IEEE Commun Mag
48. Tschudin CF (1999) Mobile agent security. In: Klusch M (ed) Intelligent information agents. Springer
49. Vigna J (1998) Cryptographic traces for Mobile Agents. In: Vigna G (ed) Mobile agent security. Lecture Notes in Computer Science, vol 1419. Springer, Berlin, pp 137–153
50. Vitek J, Castagna G (1999) Seal: A framework for secure mobile computations. In: Bal HE, Belkhouche B, Cardelli L (eds) Internet programming languages. Lecture Notes in Computer Science, vol 1686. Springer
51. Vogler H, Moschgath M-L, Kunkelmann T (1997) An approach for Mobile Agent Security and fault tolerance using distributed transactions. Darmstadt Univ of Technology, ITO, Proc of ICPADS'97 (to appear)
52. Volpano D, Smith G (1998) Language issues in mobile program security. In: Vigna G (ed) Mobile agent security. Lecture Notes in Computer Science, no 1419. Springer, pp 25–43
53. Von Eicken T, Chang C-C, Czajkowski G, Hawblitzel C (1999) J-Kernel: A capability-based operating system for Java. Lect Notes Comput Sci 1603:369–394
54. Wallach DS (1999) A New approach to mobile code security. Ph D Thesis, Princeton University

55. Wallach DS, Balfanz D, Dean D, Felten EW (1997) Extensible security architectures for Java. Technical report 546–97, Department of Computer Science, Princeton University
56. White JE (1997) Telescript. In: Cockayne WR, Zyda M (eds) Mobile agents. Manning Publ, Greenwich, pp 37–57
57. White JE (1997) Mobile agents. In: Bradshaw JM (ed) Software agents. AAAI/MIT Press, Cambridge, pp 437–472
58. Wilhelm UG, Staamann S, Buttyn L (1999) Introducing trusted third parties to the mobile agent paradigm. In: Vitek J, Jensen C (eds) Secure internet programming: Security issues for distributed and mobile objects. Lecture Notes in Computer Science, vol 1603. Springer, Berlin, pp 117–146
59. Yee B (1999) A sanctuarity for mobile agents. In: Vitek J, Jensen C (eds) Secure internet programming: Security issues for distributed and mobile objects. Lecture Notes in Computer Science, vol 1603. Springer, Berlin, pp 117–146

Books and Reviews

Barak B et al (2001) On the (im)possibility of obfuscating. In: CRYPTO '01, Santa Barbara, 19–23 August. Lecture Notes in Computer Science, vol 2139. Springer, Berlin, pp 1–18.
Braun P, Rossak W (2004) Mobile Agents: Basic concepts, mobility models, and the Tracy Toolkit. Morgan Kauffman
Cardelli L, Ghelli G, Gordon AD (1999) Mobility types for mobile ambients. In: Wiedermann J, van Emde Boas P, Nielsen M (eds) Automata, languagese and programming. In: 26th International Colloquium, ICALP'99 Proceedings. Lecture Notes in Computer Science, vol 1644. Springer, Berlin
Cardelli L, Ghelli G, Gordon AD (2000) Ambient groups and mobility types. In: van Leeuwen J, Watanabe O, Hagiya M, Mosses PD, Ito T (eds) Theoretical computer science
Carriero N, Gelernter D (1989) Linda in context. Commun ACM 32(4):444–458
Carvalho M, Breedy M (2002) Supporting flexible data feeds in dynamic sensor grids through mobile agents. In: Proceedings of the 6th IEEE International Conference on Mobile Agents. Springer
Gong L, Schemers R (1998) Signing, sealing, and guarding Java Objects. In: Vigna G (ed) Mobile agent security. Lecture Notes in Computer Science, vol 1420. Springer, Berlin, pp 206–216
Gray RS (1996) Agent Tcl: A flexible and secure mobile-agent system. In: Proceedings of the 4th Annual Tcl/Tk Workshop (TCL 96), July 1996, pp 9–23
Gray R, Kotz D, Cybenko G, Rus D (1998) Security in a multiple-language mobile-agent system, In: Vigna G (ed) Lecture Notes in Computer Science: Mobile Agents and Security
McGraw G, Felten EW (1997) Java security: Hostile applets, holes, and antidotes. Wiley
Murphy AL, Picco GP, Roman G-C (2001) Lime: A middleware for physical and logical mobility. In: Proceedings of the 21 st International Conference on Distributed Computing Systems (ICDCS-21), May 2001
Picco GP, Murphy AL, Roman G-C (1999) Lime: Linda meets mobility. In: Garlan D (ed) Proceedings of the 21 st International Conference on Software Engineering, May 1999
Tschudin C (1994) An introduction to the M0 messenger system. Technical report 86 (Cahier du CUI), University of Geneva
Volpano D (1996) Provably-secure programming languages for remote evaluation. ACM Comput Surv 28A(2):electronic

Modular Self-Reconfigurable Robots

Mark Yim, Paul White, Michael Park, Jimmy Sastra
School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Modular Self-Reconfigurable Robot Review
Complexity in Robot Configurations
Control Architectures
Mechanical/Electrical/Computational Interaction
Future Directions
Bibliography

Glossary

Bonding mechanism A mechanism that allows modules to attach to other modules. Self-reconfigurable modules have the ability to selectively make and break attachments to other modules.

Configuration The connectivity arrangement of modules in a system which describes which modules is physically attached and adjacent to which.

Configuration recognition The process of automatically determining a modular robot's connectivity arrangement.

Decentralized control A control system in which the controller elements are not central in location (like the brain) but are distributed throughout the system with each component sub-system controlled by one or more controllers.

Enumeration algorithm A routine that counts and displays the number of unique, non-isomorphic configurations of a given modular robotic system.

Global bus Communication setup such that when one unit talks all other units can listen, as opposed to neighbor to neighbor communication in which communication occurs only between two units.

Isomorphic configurations Modular structures that have the same morphology but are arranged differently according to their module labels.

Morphology The form or structure of some entity, more specifically, the connectivity arrangement of modules in a system independent of module labels.

Reconfiguration algorithm A method that transforms a given robotic configuration to a desired configura-

tion via a sequence of module detachments and reattachments.

Definition of the Subject

Modular self-reconfigurable (MSR) robots are robots composed of a large number of repeated modules that can rearrange their connectedness to form a large variety of structures. An MSR system can change its shape to suit the task, whether it is climbing through a hole, rolling like a hoop, or assembling a complex structure with many arms.

These systems have three promises:

Versatility The ability to reconfigure allows a robot to disassemble and/or reassemble itself to form morphologies that are well-suited for a variety of given tasks.

Robustness Since the system is composed of many repeated parts which can be rearranged during operation, faulty parts can be discarded and replaced with an identical module on-the-fly, leading to self-repair.

Low cost MSR systems can lower module costs since mass production of identical unit modules has an economic advantage that scales favorably. Also, a range of complex machines can be made from a set of modules saving the cost versus having multiple single-function machines for doing different tasks.

Introduction

Conceptually, the best known example of an MSR robot would be the fictional T1000 liquid-metal robot from the James Cameron film, *Terminator 2: Judgment Day*. In this movie, a robot made from a futuristic liquid-like metal, (possibly many million microscopic modules) can change its shape, copy forms, or reconstitute itself to carry out sinister aims.

Real robots that change their shape, made up of many identical modules have been created and are being studied by a wide variety of groups [33]. These robots are capable of more useful contributions to society than the T1000. They promise to be versatile, low cost, and robust. While these systems do not yet behave like liquid metal, systems on the order of 100 modules have been built and promise to be useful in search and rescue or space exploration.

The concept of modular self-reconfigurable robots can be traced back to the "quick change" end effecter and automatic tool changers in computer-controlled machining centers in the 1970's. Here, special modules, each with a common connection mechanism, were automatically interchanged on the end of an electro-mechanical or robotic arm. The concept of applying a common connection mechanism to an entirely modular robot was introduced by Fukuda with the biologically-inspired CEllular roBOT (CEBOT) in the late 1980's [11]. Here each CEBOT module is 18 x 9 x 5 cm and weighs approximately 1.1 kg. These units have independent processors and motors, and can communicate with each other to approach, connect, and separate automatically.

In the early 1990's, modular reconfigurable robots were shown to have the ability to perform the task of locomotion. In 1994, Yim explored many statically stable locomotion gaits with Polypod. Polypod [28] is an MSR robot that is significantly lighter and smaller than CEBOT. A module by itself could not locomote, but through the collective behavior of the system of many modules it could move itself from place to place and achieve many different locomotion gaits [29] such as a slinky, caterpillar, or rolling track gait.

Through this work it became clear that controlling a system with a large number of modules is complex. Initial Polypod control used a gait control table to program simple gaits on a modular robot using prescribed motions. In addition to the complexity of coordinated control, the complexity of arbitrary configurations and the sequence of reconfigurations to attain those configurations quickly developed into an interesting computational problem.

Chirikjian and Murata developed lattice style configuration systems in [10,17]. As described in Sect. "Modular Self-Reconfigurable Robot Review", the lattice style robots have modules which sit on a lattice and make it easier to represent the configurations computationally. As a result this style of system quickly became popular among computational roboticists. This also presents the interesting issue of the tradeoffs between issues solved electro-mechanically versus computationally, which is developed further in Sect. "Mechanical/Electrical/Computational Interaction".

In the later 1990's Rus [14] and Shen [6], also developed hardware but their larger contributions came in the distributed programming aspects. This included seminal trends in developing provable distributed algorithms [4] and decentralized control based on local communication [23]. Two of the areas of research include configuration self-recognition and kinematic planning of the motions for rearrangement between configurations.

This paper is structured as follows: Sect. "Modular Self-Reconfigurable Robot Review" gives a classification scheme for MSR robots, potential applications, and an overview of robotic systems that are currently being developed. Section "Complexity in Robot Configurations" discusses issues regarding complexity in the configurations of MSR robots. In Sect. "Control Architectures", we

present architectures used for control in MSR systems. In Sect. "Mechanical/Electrical/Computational Interaction", we discuss the interaction between mechanical, computer, and electrical disciplines within modular robots. Lastly, in Sect. "Future Directions" we present future directions in MSR research.

Modular Self-Reconfigurable Robot Review

Categories of MSR Systems

There are several ways of categorizing MSR robotic systems. One is based on the regularity of locations for attaching; lattice vs. chain vs. mobile, and another is based on the methods of moving between those locations; stochastic vs. deterministic.

Lattice A lattice based MSR system has modules arranged nominally in a 2D or 3D grid structure. For this category, there are discrete positions that a given module can occupy. In contrast to chain-based architectures where modules are free to move in continuous space, the grid based structure of lattice systems generally simplifies the reconfiguration process. Kinematics and collision detection are comparatively simple for lattice systems. An example is shown in Fig. 1.

Modular Self-Reconfigurable Robots, Figure 1
Crystalline. The Crystalline system is a lattice style robot developed by Rus et al. [20] at Dartmouth University (then continued later at MIT) consists of modules that can expand and contract their shape in order to reconfigure and mobilize the robotic structure. Each module has three actuators: one rack-and-pinion device that allows all four sides to expand and effectively double the side length of the module and mechanical latches that allows the module to make and break bonds to its neighbors. Locomotion and shape metamorphosis was demonstrated experimentally both in simulation and with a physical implementation. The ability to self-repair a system with a malfunctioning module was demonstrated in simulation. The system was able to identify and relocate the damaged module. Both centralized and distributed planning algorithms were explored with Crystalline

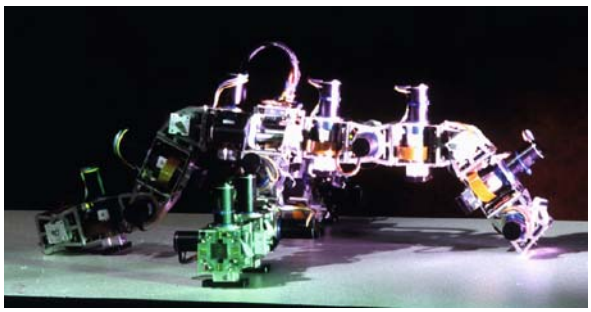

Modular Self-Reconfigurable Robots, Figure 2
PolyBot, Yim et al. developed the PolyBot chain-type MSR system at Palo Alto Research Center (PARC, formerly Xerox PARC). Each 50mm cube shaped modules is equipped with a brushless flat motor and harmonic drive which provides a single rotational DOF. Sensors provide information about neighbor proximity and contact, orientation, joint position and force torque feedback. Two hermaphroditic (electrically and mechanically) faces of the module possess redundant spring contacts to transmit power and communication and an SMA actuated mechanical latch to bond to a neighbor module. PolyBot robotic systems have shown their versatility by demonstrating locomotion as a biped, as a snake, as a rolling tread, and by climbing stairs, poles, etc. The system has also demonstrated the ability to manipulate objects and self-reconfigure

Chain A chain based MSR system consists of modules arranged in groups of connected serial chains, forming tree and loop structures. Since these modules are typically arranged in an arbitrary point in space, the coordination of a reconfiguration is complex. In particular, forward and inverse kinematics, motion planning, and collision detection are problems that do not scale well as the number of modules increases. An example is shown in Fig. 2

Mobile The mobile class of reconfiguration occurs with modules moving in the environment disconnected from other modules. When they attach, they can end up in chains or in a lattice. Examples of mobile reconfiguration devices include multiple wheeled robots that drive around and link together to form trains, modules which float in a liquid or outer space and dock with other modules.

Stochastic In a stochastic system, modules move in a 2D or 3D environment randomly and form structures by bonding to a substrate and/or other modules. Modules move in the environment in a passive state. Once a module contacts the substrate or another module, it makes a de-

cision about whether it will bond to the structure or reject a bond. The time that it takes for the system to reach a desired configuration is probabilistically bounded. The reliance on environmental forces allows the mechanical actuation to be simplified as only bonding actuation is required internal to the module. An example is shown in Fig. 3.

Deterministic In deterministic MSR systems, modules move or are manipulated directly from one position to another in the lattice or chain. The positions of each module in the system are known at all times. The amount of time it takes for a system to change from one configuration to another is determined. A module's reconfiguration mechanism requires a control structure that allows it to coordinate and perform reconfiguration sequences with its neighbors.

There are a growing number of existing physical systems that researchers are developing self-reconfigurable robots. One indication that this number is getting large is the development of a robot whose name is YaMoR (Yet another Modular Robot) [16]. Table 1 lists many of the other instantiated modular robot systems. In addition to the name, class, and author, the table lists DOF. This describes the number of actuated degrees of freedom for module motion (e. g. not latch degrees of freedom) as well as whether the system motion is planar (2D) or can move out of the plane (3D). The year is the estimated first public disclosure.

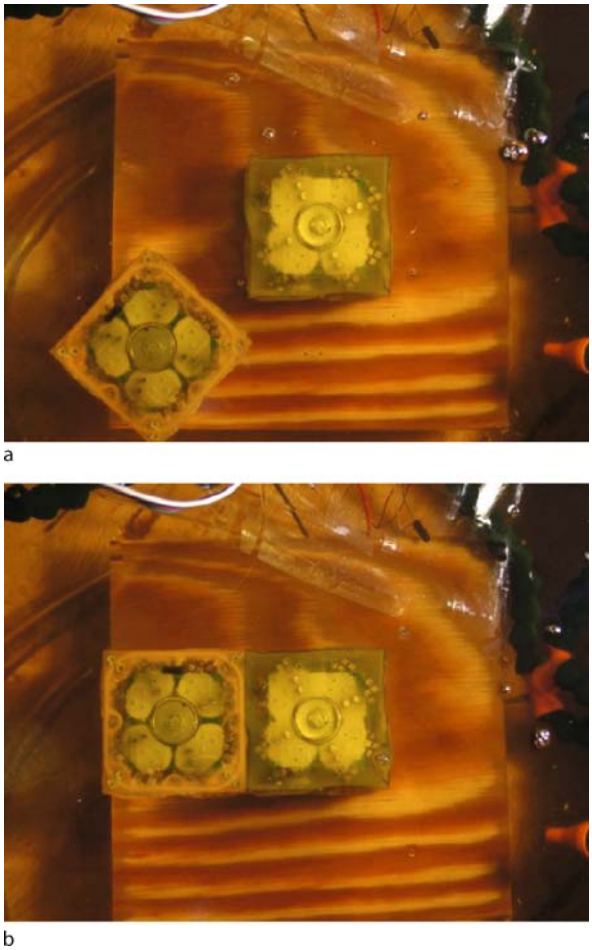

Modular Self-Reconfigurable Robots, Figure 3
Stochastic 3D. A stochastic MSR robotic system has been demonstrated in both 2D and 3D by White et al. [26] at Cornell University. The 2D system consisted of planar, square-shaped modules with electromagnets on each face that allowed the modules to selectively bond and release other modules. The modules were shuffled about randomly on an oscillating air table. The modules do not have onboard power nor do they have the capability to influence their motion. One central module acts as a powered substrate to which other modules may attach to and build desired structures. The oscillating table causes the modules to move about randomly, and when two modules collide properly, they bond to one another via the electromagnets, determine if the new configuration is desired, and release from each other if the configuration is not desired. Stochastic reconfiguration was also demonstrated in 3D where cube-shaped modules floated about in an agitated oil environment. The first generation of 3D modules (shown above) used electromagnets to provide the bonding force; a second generation used the force caused by fluid that flowed through the faces of the modules

Applications

Compared with fixed morphology robots, MSR robots are flexible in that they can adapt to a wide range of tasks and environments. However, this flexibility may compromise performance or cost. Fixed morphology systems can be optimized for a particular known task, therefore, MSR robotic systems are particularly well-suited for tasks where the operating conditions and ability requirements are not known or not well specified a priori. The following set of application examples illustrate some areas that would benefit from the development of a mature MSR system.

Space The exploration of space presents numerous challenges, including an unpredictable environment and significant limitations on the mass and volume of equipment used to study that environment. Since one set of modules can be reconfigured to perform many tasks, MSR robots can solve both the unexpected challenges while occupy little space and weight as compared to multiple devices. Graceful degradation due to failure is particularly impor-

Modular Self-Reconfigurable Robots, Table 1
List of self-reconfigurable modular systems

System name	Class	DOF	Primary author	Affiliation	Year
CEBOT	mobile	various	Fukuda et al.	Nagoya	1988
Polypod	chain	2 3D	Yim	Stanford	1993
Metamorphic	lattice	3 2D	Chirikjian	JHU	1993
Fracta	lattice	3 2D	Murata	MEL	1994
Tetrobot	chain	1 3D	Hamlin et al.	RPI	1996
3D Fracta	lattice	6 3D	Murata et al.	MEL	1998
Molecule	lattice	4 3D	Kotay and Rus	Dartmouth	1998
CONRO	chain	2 3D	Will and Shen	USC/ISI	1998
PolyBot	chain	1 3D	Yim et.al	PARC	1998
TeleCube	lattice	6 3D	Suh et.al	PARC	1998
Vertical	lattice	2D	Hosakawa et al.	Riken	1998
Crystal	lattice	4 2D	Vona and Rus	Dartmouth	1999
I-Cube	lattice	3D	Unsal	CMU	1999
Pneumatic	lattice	2D	Inoue et.al.	TiTech	2002
Uni Rover	mobile	2 2D	Hirose et al.	TiTech	2002
MTRAN II	hybrid	2 3D	Murata et al.	AIST	2002
Atron	lattice	1 3D	Stoy et al.	U.S Denmark	2003
Swarm-bot	mobile	3 2D	Mondada et al.	EPFL	2003
Stochastic 2D	stochastic	0 2D	White et al.	Cornell U.	2004
Superbot	hybrid	3 3D	Shen et al.	USC/ISI	2005
Stochastic 3D	stochastic	0 3D	White et al.	Cornell U.	2005
Catom	lattice	0 2D	Goldstein et al.	CMU	2005
Prog. parts	stochastic	0 2D	Klavins	U. Washington	2005
Molecube	chain	1 3D	Zykov et al.	Cornell U.	2005
YaMoR	chain	1 2D	Ijspeert et al.	EPFL	2005
Miche	lattice	0 3D	Rus et al.	MIT	2006

tant for robots operating in space – a component malfunction can potentially lead to mission failure. The redundant nature of MSR systems gives them the ability to discard failed modules. Modules can also be packaged in a convenient way so as to meet the volume constraints of spacecraft. Once on site, modules can be used to build structures, navigate across terrain, perform scientific studies, etc.

Search and Rescue Disaster areas such as those around collapsed buildings or other structures present another type of highly unstructured unpredictable environment where the use of an MSR robot could be beneficial. For example, the MSR system could take the form of a snake which can more easily squeeze through small void spaces to find victims. Once found, the robot could emit a locator beacon and take the form of a shelter to protect the victim until rescued.

Bucket of Stuff The term "Bucket of Stuff" is futuristic idea coined by David Duff at the Palo Alto Research Center [33]. The system would be a consumer product comprised of a container of reconfigurable modules that would reconfigure to accomplish arbitrary household tasks. This application can be seen as the most general practical goal of MSR robotics: a system that can adapt to any task in real time. A bucket of MSR modules could be used to form the desired configuration for the end user such as cleaning gutters to folding laundry.

Current Modular Robot Systems

In the previous section we presented historical examples of MSR robotic systems. In the following section we present MSR robotic systems under experimentation and development at time of this publication.

Chain The CKBot system is a reconfigurable robotic system developed by Yim et al. at the University of Pennsylvania. The CKBot system shown in Fig. 4, is a chain based system building on earlier PolyBot work at the Palo Alto Research Center. These modules utilize a servo to rotate

Modular Self-Reconfigurable Robots, Figure 4
CKbot module cluster. 4 CKbot modules and one CKbot camera module are joined together in a cluster. Several clusters can join together attaching magnetically

one portion of the module with respect to the other. In addition to statically stable locomotion gaits, Sastra et al. [22] have demonstrated a dynamic rolling gait for the CKBot system that has proven to be the fastest battery powered modular reconfigurable robot system. Global inter-module communication through CANbus as well as local neighbor-to-neighbor communication is incorporated on the modules. This system has also been used in some initial experiments in self-repair in with experiments in self-reassembly after explosion described further in Sect. "Control Architectures".

Lattice The ATRON system, shown in Fig. 5, developed by Stoy et al. [13] at the University of Southern Denmark looks to combine the reliability of reconfiguration provided by a lattice based module architecture while maintaining some of the flexibility of motion of a chain based system. Modules can distribute power via their bonding mechanisms and use a power management system for voltage regulation and battery charge maintenance. A module consists of two hemispheres where one can rotate continuously relative to the other. The bonding mechanism is extremely robust; each module has 4 female metal bars and 4 metal clasps that can be actuated to grab hold of a neighbor's bar. Reconfiguration is performed by having one module grab another and then rotate some multiple of 90 degrees to another position in the lattice structure. The ATRON system has been used to explore the value of using clusters of multiple modules to increase the manipulation, reconfiguration and locomotion abilities of the system.

The Miche system developed by Rus et al. [12] at MIT has demonstrated the ability to form desired configurations from a collection of modules. In order to self-assemble, a cluster of modules disassembles by rejecting modules that are not part of the goal configuration. Each face of the cube module has a switchable magnet and a communication interface. After the user defines the desired shape of the robotic system through the interface, a distributed algorithm determines which modules should be rejected from the system. These modules simply let go of the structure and fall due to gravity. Like many of the stochastic systems, the hardware here de-emphasizes the actuation requirements easing the ability to scale up the numbers and scale down the size.

Hybrid

The M-TRAN system developed by Murata et al. [18] at AIST/Tokyo Institute of Technology combines the positive capabilities of chain and lattice based systems to implement a highly maneuverable and reconfigurable system, Fig. 6. A module consists of one passive and one active cube that can pivot about the link that connects them and can form chains for performing tasks. However during reconfiguration, each of a module's two cubes can occupy a discrete set of positions in space when attempting to align with another module and bond for reconfiguration as in a lattice system. The current generation of M-TRAN (III) modules utilizes a mechanical latch as a bonding mechanism which is considerably faster, stronger and more reliable than the previous generation's magnetic latch. A kinematics and dynamic simulator and a GUI have been developed to aid the user in planning a reconfiguration or motion sequence of operations. This system has demonstrated the largest number of unique self-reconfiguring parallel steps in a single demonstration at 14.

The SUPERBOT system developed by Shen et al. [21] at USC/ISI is another example of a hybrid system. Building on the M-TRAN design and Shen's earlier CONRO system, one of the primary goals of this project is to develop a system robust and flexible enough to operate in harsh and uncertain environments such as space. Each module has three degrees-of-freedom (two similar to M-TRAN with an added twist degree-of-freedom) and has the capability of sharing power through its bonding mechanism and communicating via high-speed infra-red light emit-

Modular Self-Reconfigurable Robots, Figure 5
Atron system. The Atron system reconfigures in a lattice system, but can form chains as well. This image shows a four "legged" or "wheeled" configuration depending on how the modules are actuated

Modular Self-Reconfigurable Robots, Figure 6
MTRAN III four legged configuration. MTRAN modules appear similar to two cubes, one black one white. Walking occurs with chain-like motions, but reconfiguration occurs with modules at specific lattice positions. The cube half-modules checkboard space so white modules only attach to black. This eases the manufacture as one can be male and the other female. (The copyright National Institute of Advanced Industrial Science and Technology (AIST))

ting diodes (LED). A software hierarchy separates low level device specific code from high level task driven routines. The modules are controlled using hormone-inspired distributed controllers as developed for the CONRO project. Various locomotive gates have also been demonstrated in which modules traverse along carpet, sand, up a slope and across a rope, and self-reconfiguration is planned for the future.

Stochastic

Klavins et al. [3] at the University of Washington has developed a 2D stochastic MSR system named Programmable Parts. Modules are shuffled about randomly on a air hockey table by air jets. When a module collides with another module it bonds using switchable permanent magnets, communicates with the other module and decides whether or not to remain attached. The group has demonstrated that local rules can be developed that allow the system to tend toward and equilibrium of desired configurations. Using theory from statistical mechanics the group is working to develop methods for controlling stochastic MSR systems at various different scales.

Complexity in Robot Configurations

Since MSR systems are designed to be versatile, with numerous configurations for a set of modules, the problem of

recognizing and choosing useful configurations is a central area of research. The organized control of modular structures is often a complex task, involving coordinated communication between modules (each which has a processor), central controllers, and in some cases, a human user.

The computational complexity of controlling existing MSR systems varies. Factors such as processor organization (centralized or decentralized), inter-module communication schemes (i. e. global bus, local neighbor-to-neighbor, both global and local), module labeling (unique module IDs vs. unlabeled), and structural symmetry all determine a modular systems' complexity for control and coordinated computation. Ultimately, these hardware parameters determine how computationally complex the control schemes will be.

When a modular robot is controlled with a central controller it is natural to employ identifying labels so the central controller can designate explicit commands over a global bus. Since all processors and modules access a bus equally there is no indication of the relative location of modules within a configuration. Some other mechanism (e. g. the user who constructs the system, or a self-discovery mechanism) must be used to locate each module in a structure and thus map control to each module accordingly.

In the case where a system contains both a global bus and neighbor-to-neighbor communication capabilities (CEBOT, M-TRAN, CKBot, PolyBot), the system can determine a representation of the configuration (e. g., an adjacency matrix). However, for most modular systems adjacency is not enough to represent the full kinematics of the relationship between two modules as two modules maybe be attached together in different ways; for example, two cube-shaped modules may be attached face-to-face on different faces, or with different orientations on each face.

Variants of adjacency matrices [19,27] that take into account how structures are put together add essential structural information, such as inter-module port connections. While this explicit representation is not required for control, it is needed for things like simulation and any type of autonomous behavior that relies on knowing its configuration and state. For example, self-repair or any type of capability reasoning requires this explicit representation.

When doing self-discovery (automatically determining a configuration based on neighbor sensing/communication) it is often useful to see if a configuration is the same as another configuration; for example, matching a configuration to one in a library of configurations. This problem is related to finding the automorphism group of graph representations, which is known to be a hard problem with no known polynomial time algorithm [7].

The eigenvalues of the port-adjacency matrix (a generalized adjacency matrix that contains port connection numbers to designate how modules are connected) is invariant under any of the $n!$ ways a structure with unique module IDs can be rearranged or relabeled. In graph theory terminology, each relabeling is graph isomorphic to one another. Module ID mappings between isomorphic configurations can be found with a heuristic program such as *nauty* [15], a sequential search through a three-dimensional linked-list representation of the system, or with eigenvectors corresponding to the shared eigenvalues of the isomorphic structures [19].

$$\mathbf{A}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 7 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 & 6 \\ 1 & 0 & 0 & 7 & 6 & 0 & 0 \\ 0 & 1 & 7 & 0 & 4 & 0 & 0 \end{bmatrix}$$

$$\mathbf{A}_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 7 & 0 \\ 0 & 0 & 0 & 1 & 7 & 0 & 4 \\ 0 & 0 & 7 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 7 & 1 & 0 & 0 & 0 & 0 & 6 \\ 0 & 0 & 6 & 0 & 0 & 4 & 0 \end{bmatrix}$$

$$\mathrm{Det}\,(\mathbf{A}_1 - \lambda\,\mathbf{I}) = \mathrm{Det}\,(\mathbf{A}_2 - \lambda\,\mathbf{I}) = \lambda^7 - 76\,\lambda^5 + 868\,\lambda^3$$

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Consider the two isomorphic arrangements of the H-shaped modular CKBot structure and their corresponding port-adjacency matrices, $\mathbf{A}_1$ and $\mathbf{A}_2$, in Fig. 7. Note that the two configurations share the same characteristic polynomial (and hence eigenvalues), as expected since the configurations are rearrangements of the same shape. The property that swapping two columns and two rows of a square matrix does not change its determinant (row swap and column swap each change the determinant by a minus sign) corresponds to rearrangements or relabelings of module IDs. The permutation matrix $\mathbf{P}$ that maps the module IDs between the two isomorphic configurations such that $\mathbf{A}_2 = \mathbf{P}\mathbf{A}_1\mathbf{P}^{-1}$ can be determined using the methods described in [19].

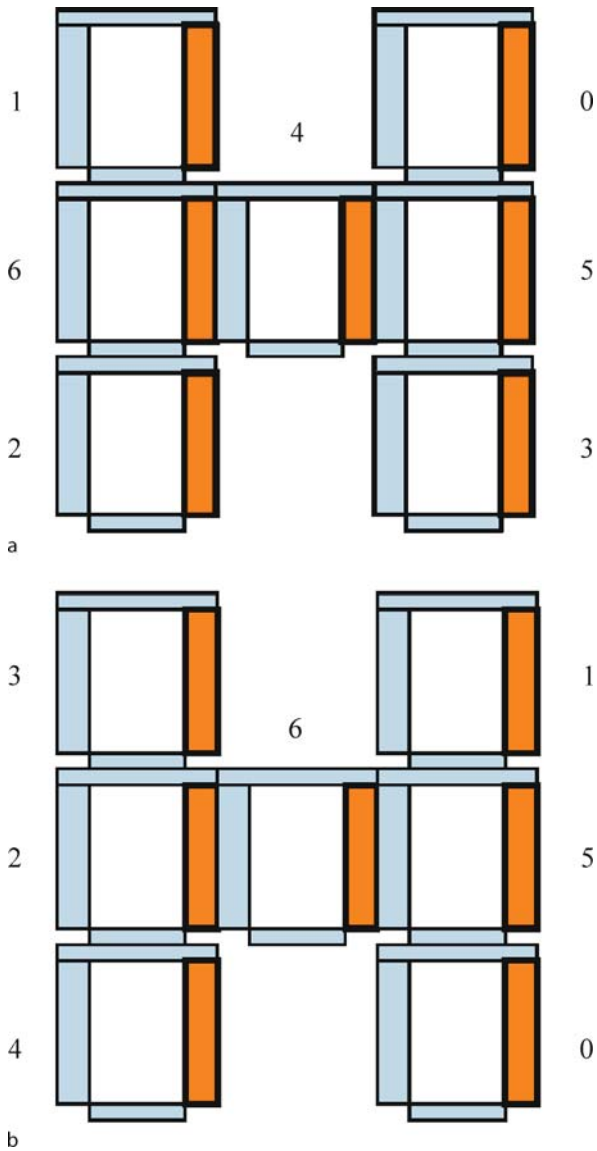

Modular Self-Reconfigurable Robots, Figure 7
Two isomorphic CKBot configurations Modules are rearranged, but the connectivity is the same (similar to nodes on a graph relabeled)

In cases where neighbor-to-neighbor communication is only present (ATRON, Conro hormone studies, Crystalline, Claytronics Atom), distributed algorithms that employ the processors of modules interacting together in parallel divide the computation required for configuration recognition and motion planning. These MSR systems typically use token-type messages where aggregate configuration information is passed from module-to-module. Complexity scaling is a critical issue for these distributed systems as the number of inter-module messages for goal configuration recognition and planning is immense for as few as 10 modules. A major benefit of the decentralized approach is that for such systems unique module IDs are not necessary since each unit can only communicate with an adjacent unit and thus the system is not limited by an address space. Decentralized approaches also promise to scale as computational resources scale with the number of modules.

For a modular robotic system composed of n homogeneous units, each with c ports, and w ways of connecting modules, an upper-bound number of structurally unique configurations is $(cw)^n$ [19]. For example, given 11 CKBot modules, each unit has 7 ports, each of which can be uniquely connected to another module in 10 ways (3 rotations for the each of the 3 top faces, and 1 orientation for the bottom connection). Therefore an upper-bound to the number of unique configurations is $(10*7)^{11} \cong 2.0*10^{20}$. This number is an upper-bound since there are inherent physical symmetries in certain structures that this approximation double counts.

An enumeration algorithm that more precisely counts the number of non-isomorphic configurations of a modular robot was developed by Chen [8]. Structural and kinematic symmetries were taken into account to find a precise number of unique configurations for a given number of modules. In this method, Polya's Enumeration Theorem is employed to count a structural state only once. For example, two cubes modules that can connect to one another on any of the six faces (each in one orientation) has 36 ways of connecting. Chen's approach takes into account the 3-fold symmetry of the cubes to determine that there is only one unique way of connecting the two modules. This assumes that all six ports on the cubes are all the same; if this symmetry is broken and one or both modules can have multiple types of ports (revolute, helical, cylindrical, etc.) then the algorithm takes these variations into account to find the number of unique ways (greater than one) that the two cubes can be connected.

Another challenge in the field of MSR robotics is the development of *reconfiguration algorithms*: a method that transforms a given robotic configuration to the desired configuration via a sequence of module reconfigurations. A naive centralized method is to perform an exhaustive search of the configuration space (all reachable configurations) beginning with the initial configuration until a path (reconfiguration sequence) to the goal configuration is found. Because it is possible for modules in an MSR system to move in parallel, the branching factor for the search tree is $O(m^n)$ with n being the number of modules free to move and m being the number of ways the module can

move. Finding an optimal path might require searching the whole space which is clearly intractable for large n.

Many groups have developed methods and tools for doing self-reconfiguration planning [9,32] that include centralized algorithms. Several groups [2,4,24,25] have presented distributed reconfiguration algorithms. In these cases, the reconfiguration algorithm is embedded on processors running on every module. Each module has an identical program with implicit or explicit knowledge of the required goal states, but only local information about the current state of neighboring modules.

One trend in developing these distributed systems is the use of "meta-modules" – groups of modules together considered to be a subset. As described further in Sect. "Mechanical/Electrical/Computational Interaction", there is a tight coupling between mechanical properties, the constraints on motion and the ease of programming such things as the reconfiguration problem. In many cases, idealized cubes that move in known and unconstrained fashions are used to develop algorithms. However, in the physical world, there are added mechanical constraints that enable the manufacture and motion of the devices. Examples of these constraints include, blocking constraints where modules may block motions under certain conditions [35], checkerboard constraints in lattice configurations where modules may only move to alternate positions (as if they were bi-partite) [18]. By grouping small numbers of modules into a meta-module, many of these constraints can be eased. For example the checkerboard constraint and the blocking constraints can removed for a group of modules moving in concert. However, the system as a whole loses resolution based on the size of the meta-module.

In [1] Abrams and Ghrist introduce the state complex as an extension of the concept of the configuration space. They present an algorithm that uses the added structure defined by the state complex to optimize (with respect to total reconfiguration time) a reconfiguration sequence generated by local planner (such as the aforementioned distributed algorithms.)

Control Architectures

Review of Existing Architectures

A design philosophy behind modular robots is that each module is very simple. In fact, one group [5] proposed the *Ensemble Axiom* "A [module] should include only enough functionality to contribute to the desired functionality of the ensemble." A module by itself cannot achieve much, but modules arranged together in a system can achieve complex tasks such as manipulation and locomotion. Similarly, the control of a single module is usually simple whereas controlling a system of many modules becomes difficult very quickly. For the overall system, different control architectures have been implemented which we will describe in more detail.

In large part, the implementation of a control architecture depends on the communication structure upon which it is built. Communication between modules can be achieved through a global bus such as CANbus (Controller Area Network, a popular automotive and more recently robotics communications protocol) and/or locally using neighbor-to-neighbor communication such as infra-red (IR) emitter/detector pairs. Many systems use both (Polybot, CKBot, M-TRAN, CONRO and Superbot). Wireless communication is also possible which is architecturally similar to a global bus. In the YaMoR system [16], Bluetooth wireless is the sole means of inter-module communication. ATRON and Crystalline modules [13,20] use only local nearest neighbor IR communication.

As mentioned earlier, control architectures can be implemented in either a centralized or decentralized fashion. In most cases it is easier to develop and analyze a centralized approach. The advantage of decentralized control architecture is that computation is shared among modules. No single unit needs to do all the heavy computation. This is also thought to be more robust and more easily lends itself to scaling to large numbers of modules. It is easier to implement centralized control has using global communications and decentralized using local and there are many examples of such. However, it is possible to implement centralized on a local bus and decentralized on a global bus.

An example of centralized control architecture is implemented on [30]. Each module has its own controller that positions its local actuator. In addition, a master controller communicates to the module controllers to set local behaviors such as setting desired joint angles under position control. In other words, a designated unit sends commands to all the individual modules and synchronizes the action of the whole system. A simple method of implementing this control is to use a *gait control table*. The gait control table is an $n \times m$ matrix where m is the number of modules and n is the number of steps of the gait. Each cell in the table holds the desired joint angle for a module. Each column of angles corresponds to the sequence of joint angles for a given module. The controller steps through this table row by row and sends these angles to the corresponding module. Typically stepping through the table occurs at a specified rate, so the vertical axis can represent time. Each module takes the next desired joint angle in the table and interpolates in joint space. The time between steps sets

the joint velocity so desired motions have C1 continuity in joint space.

Shen et al. propose a control that is based on biological hormone systems in [23]. The basic idea is that an inter-module "hormone" message is a signal that triggers different actions in different modules while leaving the low-level execution of these actions to the individual modules. The obvious biological analogy occurs when a human experiences sudden fear, and adrenaline hormones released by the brain trigger fight-or-flight behaviors in the body (i. e., the mouth opens, skin gets goose bumps, and the legs jump). Based on this principle, Shen et al. designed a control mechanism that lies somewhere between master and master-less control in that typically one or more modules need to start the hormone messages. It reduces the communication cost for locomotion controls, yet maintains some degree of global synchronization and execution monitoring.

At its root, the hormone is a local message passing system where modules can receive, act on or change messages as they are passed from module to module. An advantage of this type of control is that modules are treated identically without labels or identification numbers; instead the topology of a configuration is the differentiator and thus has a great bearing on the implementation. This lends itself well to simple locomotion control such as undulating gaits however, developing arbitrary motions can be more difficult to implement.

A fully decentralized planning system has been developed by Rus et al. In [4] an algorithm modeled after cellular automata is described. Cellular automata (CA) control uses local rules that are the same for all modules. A rule can be viewed as having a set of pre-conditions. If all those preconditions are satisfied, then a certain action is applied. For example, for a given cell, the pre-conditions could be whether a cell exists at a certain location, whether a cell does not exist at a certain location, and whether a cell is empty. If all preconditions are satisfied, the cell moves itself in a certain direction. Rather than having one master controller being in control of the whole system, modules think for themselves in a parallel distributed fashion. All modules run on the same rules and all modules are programmed with the same code. Just as the hormone method of control adds some complexity to the development of arbitrary motions, it is also difficult to do in the CA case.

Completely centralized control architecture is relatively straightforward to implement. But issues arise when dealing with millions of modules such as reaching the limits of bandwidth when using a global communication bus. On the other hand it is hard to achieve complex tasks with a completely decentralized architecture that requires only local communication because it is hard to implement behavior in a distributed fashion.

Self-Assembly After Explosion

An example of a hybrid architecture in which global as well as local communication is used is given in [34]. In this work the ability for a modular robot to repair itself is demonstrated by having the robot reassemble into one connected component after disassembly from a high energy event. As a system assembles itself, the connectivity of the robot changes many times. Having disparate disconnected pieces requires a level of decentralized control, however as the system comes together, the modules must act in a coordinated manner as well.

In [34] a demonstration is shown with 15 modules. Modules are grouped into three clusters of five modules. Clusters move as physically separate units, search for and localize each other, and crawl toward each other to connect using magnet faces and form one aggregate unit.

Within each cluster, the modules are attached using screws and an electrical header is included in between these modules to facilitate a global CANbus. The clusters connect to each other using magnet faces without an electrical header so communication is only achieved through IR communication. Thus, this hybrid architecture consisted of a global CANbus within a cluster and local IR communication in between clusters.

The hardware in SAE work is hierarchical – modules form clusters – clusters form systems – the control architecture follows that architecture as well. Each module has an onboard controller that controls the position of the local actuator. Within each cluster a controller communicates on the CANbus to all the modules in that cluster. The master cluster controller gives commands similar to a gait control table to implement behaviors such as crawling, detecting a fallen condition and self-righting, searching for other modules etc. Once clusters dock to each other magnetically, the cluster controllers can communicate to other cluster controllers using a combination of CAN and IR. For this work, one cluster controller is designated as the master whereas the other controllers are designated as slaves and follow the coordinated messages sent by the master cluster controller. For example, in the walking state with all clusters connected, the master cluster controller sends precisely timed messages to the other clusters to coordinate tasks like walking and turning.

Mechanical/Electrical/Computational Interaction

MSR systems sit at an interesting junction between mechanical, electrical, and computational interaction.

Robotics in general, is highly interdisciplinary since it requires expertise in all three of those areas. However, the configurability aspect of MSR systems adds to the intertwining of these disciplines. Enabled with electronic technologies (such as communication architecture), modular robotic structures introduce unique mechanical properties, which often require novel computational processes.

Electro-Mechanical Solutions to Computational/Information Problems

The reconfiguration planning problem consists of determining the motions of individual modules to attain a global shape under a variety of constraints. One common constraint is that the MSR system must maintain one connected component (for example, if power is shared from one module to the next if a module disconnects from the group it loses power). Determining whether a module detachment will sever the system into two or more pieces is often computationally and communication bandwidth intensive, as modules may be required to communicate with every other module for this analysis (e. g., the system may be in the shape of a large loop, in which case a disconnection motion between two modules will still leave a single connected component; however, that cannot be determined until every module has communicated at least once). If every module were to simultaneously check whether a disconnection would violate the connectedness constraint, there would be $O(n^2)$ messages and the communications system would quickly become saturated.

One electro-mechanical solution for power distributed systems (shared power between two or more modules) is to use power distribution to determine connectivity. One way this could be achieved would be to develop an intra-module connector where power between connected faces can be temporarily severed, to simulate an actual physical disconnection. If power to either module is lost, it could be concluded that that disconnection violates the connectedness constraint.

Computational Solutions to Mechanical Problems

Applying large forces or torques over some motion path is usually solved mechanically by designing stronger motors or leverage mechanisms. However, with modular robots, this problem can be moved to the computational planning domain. Robot systems with many redundant degrees-of–freedom, such as those typically found with chain style modular robots, can exploit configurations which have large mechanical advantage [31].

The idea is to utilize the very large mechanical advantage that can be obtained when a system's Jacobian is near a singularity. For example, when using a set of modules that have parallel chains, one chain can be moved to be near a singularity and have large mechanical advantage (e. g. when a human knee is straight the Jacobian representation of the leg loses rank and becomes singular, he can carry much more weight then when it is bent). Consequently, this chain then has a large mechanical advantage which can in turn apply a large force in the desired direction to move the system to a new position. Meanwhile, another parallel chain can be reoriented to be near a singularity at the new position, and then apply large forces yet again to a new position. By repeatedly switching a subset of the motors supporting the load, a ratcheting kind of action can be used to move links to desired positions while under large external forces. If the size of each ratchet motion can be made arbitrarily small, it can be arbitrarily close to the singularity with very large mechanical advantage. Thus, weak motors can be used to provide large forces. Of course, there are practical limits to this method, e. g., sensing accuracy, material strengths, joint precision etc.

The sequence of ratcheting actions that move the end points through desired trajectories is potentially a computationally intensive problem. Hence, the problem of providing sufficiently large torques has now been solved not from a mechanical viewpoint but computationally instead.

This tight integration of computational and electro-mechanical complexity can be viewed as a wider space from which to find solutions, or as a more complex problem in finding optimal solutions. This is particularly interesting in that it is likely that optimal solutions will not be found by experts in one field, but by the interaction of experts in several fields.

Future Directions

The grand challenges for MSR robotic systems were the results of a workshop where a group of researchers in the MSR robot community gathered and then presented in [33]. A proposed ultimate goal for these systems would be to one day use them in vast numbers for practical applications where un-supervised, adaptive self-organization is needed. Five grand challenges that, if overcome, would enable a next-generation of modular robots with vastly superior capabilities are summarized here:

Big systems Most systems of modular robots have been small in number, especially compared to, for example, the number of components in a living cell (which many researchers view as the best example of a self-organizing, modular system). The demonstration of

a system with at least 1000 individual units would suggest that modular robots have come of age.

Self-repairing systems A demonstration of a self-healing structure made up of many distributed, communicating parts would require rethinking algorithms for sensing and estimation of the global state, as well as truly robust hardware and algorithms for reconfiguration that work from any initial condition. A concrete example would be having a system blown up (randomly separated into many pieces) then self-assembling, or recovering from failure of a certain percentage of faulty units.

Self-sustaining systems A demonstration of a system actively running for, say 1 year, in an isolated self-sustaining robotic ecology would require new techniques in power management and energy harvesting, as well as the ability to cope with the inevitable failures.

Self-replication and self-extension While simple robotic self-replication has been demonstrated using few high-level modules, a significant challenge remains to demonstrate self-replication from elementary components and raw materials. The demonstration of a "seed" group of modular robots that can build copies of themselves from raw materials would require advancing beyond a level of complexity that Von Neumann identified as the equivalent of breaking the sound barrier for engineered systems.

Reconciliation with thermodynamics If modular robots are to be miniaturized to micro and/or nano-scale, or if the ideas discovered in this community are even to be tied to nanotechnology, the stochastic nature of nanoscale systems must be addressed. Most existing modular robot systems overcome entropy through brute force and unreasonable amounts of energy. Molecular systems, on the other hand, employ random diffusive processes and are robust to the intrinsic noise found at the nanoscale. The demonstration of a system where stochastic fluctuations are the dominant factor would represent a fundamental advance.

Bibliography

Primary Literature

1. Abrams A, Ghrist R (2004) State complexes for metamorphic robot systems. Int J Robot Res 23(7–8):809–824
2. Bhat P, Kuffner J, Goldstein S, Srinivasa S (2006) Hierarchical Motion Planning for Self-reconfigurable Modular Robots. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Beijing, October 2006. pp 886–891
3. Bishop J, Burden S, Klavins E, Kreisberg R, Malone W, Napp N, Nguyen T (2005) Self-organizing programmable parts. In: Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), August 2005, pp 3684–3691
4. Butler Z, Kotay K, Rus D, Tomita K (2002) Generic decentralized control for a class of self-reconfigurable robots. In: Proceedings of the 2002 IEEE International Conference on Robotics & Automation (ICRA), Washington DC, May 2002, pp 809–816
5. Campbell J, Pillai P, Goldstein SC (2005) The robot is the tether: active, adaptive power routing modular robots with unary inter-robot connectors. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Edmonton Alberta, August 2005, pp 4108–4115
6. Castano A, Shen W-M, Will P (2000) CONRO: Towards Deployable Robots with Inter-Robots Metamorphic Capabilities. Autonom Robot 8(3):309–324
7. Chartrand G, Lesniak L (1986) Graphs and Digraphs. Wadsworth Publ Co, Belmont
8. Chen I, Burdick J (1993) Enumerating the Non-Isomorphic Assembly Configurations of Modular Robotic Systems. In: Proceedings of the IEEE/RSJ Int Conference on Intelligent Robots and Systems (IROS), Yokohama, July 1993, pp 1985–1992
9. Chiang CJ, Chirikjian G (2001) Modular Robot Motion Planning Using Similarity Metrics. Auton Robot 10(1):91–106
10. Chirikjian G (1994) Kinematics of a Metamorphic Robotic System. In: Proceedings of the 1994 IEEE International Conference on Robotics & Automation (ICRA), San Diego 1994, pp 449–55
11. Fukuda T, Nakagawa S (1988) Dynamically reconfigurable robotic system, Robotics and Automation. In: Proceedings 1988 IEEE International Conference, Philadelphia, 24–29 Apr 1988, pp 1581–1586, vol 3
12. Gilpin K, Kotay K, Rus D (2007) Miche Modular Shape Formation by self-Disassembly. In: Proceedings of the 2007 IEEE International Conference on Robotics & Automation (ICRA). Rome, April 2007, pp 2241–2247
13. Jørgensen M, Østergaard E, Lund H (2004) Modular ATRON: modules for a self-reconfigurable robot. In: Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), October 2004, pp 2068–2073
14. Kotay K, Rus D, Vona M, McGray C (1998) The Self-reconfigurable robotic molecule. In: Proceedings of the 1998 IEEE Intl. Conf. on Robotics and Automation (ICRA), May 1994, Leuven, Belgium, May 1998, pp 424–431
15. McKay B (1981) Practical graph isomorphism. Congressus Numerantium 30:45–87
16. Moeckel R, Jaquier C, Drapel K, Dittrich E (2006) YaMoR and Bluemove – an autonomous modular robot with Bluetooth interface for exploring adaptive locomotion. Climbing and Walking. Springer, Berlin, pp 285–692
17. Murata S, Kurokawa H, Kokaji S (1994) Self-Assembling Machine. In: Proceedings of the 1994 IEEE International Conference on Robotics & Automation (ICRA), San Diego, May 1994, pp 441–448
18. Murata S, Yoshida E, Kamimura A, Kurokawa H, Tomita K, Kokaji S (2002) M-TRAN: Self-Reconfigurable Modular Robotic System. IEEE/ASME Trans Mechatron 7(4):431–41
19. Park M, Chitta S, Teichman A, Yim M (2008) Automatic Configuration Recognition in Modular Robots. Int J Robot Res 27(3–4):403–421
20. Rus D, Vona M (2000) A physical implementation of the self-reconfiguring crystallinerobot. In: Proceedings of the 2000 IEEE

International Conference on Robotics & Automation (ICRA), San Francisco, April 2000, pp 1726–1733
21. Salemi B, Moll M, Shen W-M (2006) SUPERBOT: A Deployable, Multi-Functional, and Modular Self-Reconfigurable Robotic System. In: Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Beijing, October 2006, pp 3636–3641
22. Sastra J, Chitta S, Yim M (2006) Dynamic Rolling for a modular loop robot. In: Proceedings of the International Symposium on Experimental Robotics, Rio de Janerio, July 2006
23. Shen W-M, Salemi B, Will P (2002) Hormone-inspired adaptive communication and distributed control for CONRO self-reconfigurable robots. IEEE Trans Robot Autom 18(5):700–712
24. Vassilvitskii S, Yim M, Suh J (2002) A complete, local and parallel reconfiguration algorithm for cube style modular robots. In: Proceedings of the 2002 IEEE International Conference on Robotics & Automation (ICRA), Washington DC, May 2002, pp 117–122
25. Walter J, Welch JL, Amato NM (2004) Distributed reconfiguration of metamorphic robot chains. Distrib Comput 17(2):171–189
26. White PJ, Kopanksi K, Lipson H (2004) Stochastic self-reconfigurable cellular robotics. In: Proceedings of the 2004 IEEE International Conference on Robotics & Automation (ICRA), New Orleans, April 2004, pp 2888–2893
27. Will P, Castano A (2001) Representing and Discovering the Configuration of Conro Robots. In: Proceedings of the 2001 IEEE International Conference on Robotics & Automation (ICRA), Seoul, May 2001, pp 3503–09
28. Yim M (1994) Locomotion with a Unit Modular Reconfigurable Robot. Ph D Thesis, Stanford University
29. Yim M (1994) New locomotion gaits, Robotics and Automation. In: Proceedings 1994 IEEE International Conference, San Diego, 8–13 May 1994. pp 2508–2514, vol 3
30. Yim M, Duff DG, Roufas KD (2000) PolyBot: a modular reconfigurable robot. In: Proceedings of the 2000 IEEE International Conference on Robotics & Automation (ICRA), San Francisco, April 2000, pp 514–520
31. Yim M, Duff DG, Zhang Y (2001) Closed-chain motion with large mechanical advantage. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Maui, October 2001
32. Yim M, Goldberg D, Casal A (2002) Connectivity Planning for Closed-Chain Reconfiguration. In: Proceedings of the SPIE Vol 4196, Sensor Fusion and Decentralized Control in Robotic Systems III, October 2002, pp 402–412
33. Yim M, Shen W-M, Salemi B, Rus D, Moll M, Lipson H, Klavins E, Chirikjian GS (2007) Modular Self-Reconfigurable Robot Systems: Grand Challenges of Robotics. IEEE Robot Autom Mag 14(1):43–52
34. Yim M, Shirmohammadi B, Sastra J (2007) Towards Self-reassembly After Explosion. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), October 2007, pp 2767–2772
35. Yim M, Zhang Y, Lamping J, Mao E (2001) Distributed control for 3D metamorphosis. Auton Robot 10(1):41–56

Books and Reviews

Butler Z, Fitch R, Rus D, Wang Y (2002) Distributed Goal Recognition Algorithms for Modular Robots. In: Proceedings of the 2002 IEEE International Conference on Robotics & Automation (ICRA), Washington DC, May 2002, pp 110–16

Fukuda T, Kawauchi Y (1988) Dynamically Reconfigurable Robotic System. In: Proceedings of the 1988 IEEE International Conference on Robotics & Automation (ICRA), Philadelphia, April 1988, pp 1581–86

Fukuda T, Nakagawa S, Kawauchi Y, Buss M (1989) Structure decision method for self organising robots based on cell structures-CEBOT Robotics and Automation, 1989. In: Proceedings of 1989 IEEE International Conference, Scottscale, 14–19 May 1989. vol 2. pp 695–700

Murata S, Yoshida E, Kamimura A, Kurokawa H, Tomita K, Kokaji S (2002) M-tran: self-reconfigurable modular robotic system. IEEE/ASME Trans Mechatron 7(4):431

Ostergaard EH (2004) Distributed control of the ATRON Self-Reconfigurable robot. Ph D, Univ. of Southern Denmark

White P, Zykov V, Bongard J, Lipson H (2005) Three dimensional stochastic reconfiguration of modular robots. In: Proceedings of Robotics: Science and Systems. MIT, Cambridge

Yim M, Zhang Y, Duff D (2002) Modular Reconfigurable Robots, Machines that shift their shape to suit the task at hand. IEEE Spectr Mag 39(2):30–34

Zhang Y, Roufas K, Yim M (2001) Software Architecture for Modular Self-Reconfigurable Robots. In: IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS), Hawaii, October 2001

Molecular Automata

Joanne Macdonald[1], Darko Stefanovic[2], Milan Stojanovic[1]

[1] National Chemical Bonding Center: Center for Molecular Cybernetics, Division of Experimental Therapeutics, Department of Medicine, Columbia University, New York, USA

[2] National Chemical Bonding Center: Center for Molecular Cybernetics, Department of Computer Science, University of New Mexico, Albuquerque, USA

Article Outline

Glossary
Definition
Introduction
Molecular Automata as Language Recognizers
Molecular Automata as Transducers and Controllers
Future Directions
Bibliography

Glossary

Algorithmic self-assembly of DNA tiles Spontaneous assembly of structures consisting of interlocking semi-rigid DNA molecules (the tiles). Different tiles, con-

taining different exposed ends, can be synthesized such that their affinity to interlock can be predefined. Particular sets of tiles may be constructed that result in the self-assembly of unique structures; each such set represents an algorithm for the assembly of the structure.

DNA Sierpinski triangle An algorithmically self-assembled DNA structure with a pattern approximating the Sierpinski gasket, a fractal.

DNA computing Generally, any use of the information-carrying capacity of DNA to achieve some computational or decision-making goal. The term is also used to refer specifically to the use of DNA in the manner pioneered by Adleman as a massively parallel computation substrate to solve certain combinatorial optimization problems.

DNA circuit A system consisting of multiple DNA logic gates in which DNA is used as the signal carrier between the gates; the system performs a Boolean logic function that may be more complex than the functions achievable using single logic gates.

DNA binary counter A device that uses DNA as a computational substrate to maintain its state, an integer binary numeral, and advances through successive states which represent successive integers. An algorithmic self-assembly implementation exists.

DNA logic gate A molecular device using DNA as a computational substrate to perform a Boolean logic function.

Deoxyribozyme An oligonucleotide synthesized such that its structure gives rise to enzymatic activity that can affect other oligonucleotides.

Deoxyribozyme-based automaton A molecular automaton which uses deoxyribozyme-based logic gates to achieve its function.

Deoxyribozyme-based logic gate An implementation of a DNA logic gate in which enzymatic activity of a deoxyribozyme is controlled by the presence or absence of activating or inhibiting oligonucleotides.

Ligase An enzyme, in particular a deoxyribozyme, that promotes the linking of two oligonucleotides into a single longer oligonucleotide.

Modular design of nucleic acid catalysts A method for the design of nucleic acid catalysts, in particular deoxyribozymes, wherein their nucleotide sequences are chosen by stringing together motifs (such as recognition regions, stems, and loops) which have been established to perform a desired function. With care, motifs will not interfere and will achieve a combined complex function.

Molecular automaton A device at the molecular scale which performs a predefined function; this function is seen at the macroscopic level as sampling the environment for molecular stimuli and providing

Molecular finite state automaton A device that performs a language recognition function using molecules as a computational substrate; the language symbols, the states, and the transitions between states are realized as molecules.

Molecular Mealy automaton A device using molecules as a computational substrate to achieve a general-purpose sequential logic function, viz., a finite state transducer.

Molecular logic gate A device using molecules as a computational substrate to perform a Boolean logic function.

Oligonucleotide A single-stranded DNA molecule consisting of a relatively small number of nucleotides, usually no more than a few dozen.

Phosphodiesterase An enzyme, in particular a deoxyribozyme, that promotes the cleavage of an oligonucleotide into two shorter oligonucleotides by catalyzing the hydrolysis a phosphodiester bond.

Definition

Automata are devices that autonomously perform predetermined, often complex, functions. Historically, the term referred to mechanical devices that imitated human behaviors, especially those that exhibited intelligence. In the 20th century, electronics made possible the development of various complex automata, the prime example being the computer. For its part, computer software was organized as a hierarchy of automata operating within the computer automaton. A rich formal theory of automata arose in the fields of electronics and computer science to aid in the design and use of these now ubiquitous devices.

Molecular automata employ molecular-scale phenomena of binding, dissociation, and catalytic action to achieve predetermined functions. Different chemical implementations are possible, including proteins and inorganic reactions, but the principal prototypes today use nucleic acid chemistry, and this article will focus on this modality. Nucleic acid automata may be able to interact directly with signaling molecular processes in living tissues. Thus, they provide a path towards autonomous diagnostic and therapeutic devices and even towards engineered control of cell behavior.

Introduction

Humans appear to be naturally fascinated with the construction of devices with parts that move independently, contraptions that operate autonomously, and in partic-

ular, machines with an appearance of intelligence. Each technological advance throughout the ages has coincided with the production of more and more sophisticated devices; indeed, some argue that our strong innate urge toward a mechanistic philosophy has led to the development of automata [16].

Using tools at hand, ancient cultures created a variety of automata, from the animated figures of Rhodes to the singing figures of China. As with many human endeavors, the 18th century was the era of splendor and variety in automata, particularly among the French, and sophistication continued through the 19th: from courtly displays (mechanical) to digesting ducks [33] (biomechanical) to painting devices (mechanical, but emulating creative intelligence) [4,31].

Modern automata are almost uniquely defined by the advent of computers. Just as electronics permitted building more complex systems than mechanics, software on programmable computers permitted vastly more complex systems. An entire discipline of automata theory developed to characterize the computational abilities and resource requirements of different devices; all of theoretical computer science is an outgrowth of this effort.

The formal notion of an automaton centers on a description of the configuration, or state, of the device, and of the transitions it makes from state to state. The states are typically discrete. The transitions are discrete as well and they occur in response to external stimuli formalized as the consumption of discrete input symbols. Given a current state and a current input symbol, the transition rules of an automaton define what state the automaton may enter next. A history of external stimuli is captured by an input symbol string. The automaton may produce output symbols. This has been formalized in two different but equally expressive ways: either an output symbol is associated with a specific state, or it is associated with a specific transition between two states. In either case, the successive output symbols form an output string.

Automata defined in this fashion are closely linked with formal language theory. The input and the output symbols are drawn from certain alphabets, and the input and the output strings belong to languages over these alphabets. Of particular interest are automata with a restricted form of output: a *yes* or a *no* output is associated with each state. Such automata can be viewed as language recognizers under the following interpretation: if the state in which the automaton finds itself upon consuming an input string is a *yes* output state, that string is said to belong to the language, and otherwise not. More generally, an automaton defines a correspondence between strings in the input alphabet and strings in the output alphabet; thus it can be understood as a language-translating device, or transducer. Language recognizers and transducers play a central role in traditional accounts of computational complexity, as well as in practical daily use, such as in web query processing, under the name of parsers. If the output alphabet consists of signals for mechanical actions, we have the essence of robotic control.

The number of symbols in an alphabet is generally taken to be finite. On the other hand, the number of states of an automaton might be finite or infinite. Finite state automata are readily implemented in electronics using a register (to hold an encoding of the state) and a combinational circuit (to compute the transitions). Indeed, only finite state automata can be physically implemented. Nevertheless, a beautiful theory exists for classes of automata with infinite state, this theory has led to useful programming paradigms, and such automata can be simulated using finite resources as long as we are willing to accept that they may occasionally run out of these resources and fail. Examples include pushdown automata, in which the state includes the contents of a semi-infinite push-down stack of symbols, and the Turing machine, in which the state includes the contents of an infinite tape of symbols.

Even as electronic embodiments of automata permeate the modern society, from mobile telephones to automated bank tellers to navigational satellites, they are not readily integrated into life processes: the environment of living tissues presents an obstacle for the deployment, powering, and operation of electronics. Biocompatible alternatives are needed.

With recent advances in molecular biology, research is turning toward the taming of molecules for the development of automata on the molecular scale. Within this body of research several strands can be recognized. Firstly, researchers have adapted molecular processes for explicit emulation of mathematical theory. Secondly, researchers have sought to enhance known cellular systems for alternative control of biological events. Additionally, however, researchers are also beginning to engineer molecules for completely novel purposes unrelated to their original design.

Regardless of the purpose, all of these automata fundamentally require similar mechanisms to operate autonomously within their respective environments. Molecular automata must have some ability to sense inputs (molecular or other) within their external environment, make a predetermined decision based on the sensed input, and then autonomously produce an output that affects their environment.

Here we review recent advances in the engineering of molecular automata, and outline fundamental principles

for engineering of molecular automata. Sect. "Molecular Automata as Language Recognizers" treats molecular automata created as embodiments of the mathematical concept of a language recognizer. Minimal formal background is given, and recent research exemplars are introduced, along with sufficient background in nucleic acid chemistry. We only treat exemplars that have been demonstrated in the laboratory. Sect. "Molecular Automata as Transducers and Controllers" treats molecular automata that work as transducers, and is similarly organized. In Sect. "Future Directions", we speculate on the future uses and developments of the molecular automata technology in science and medicine.

Molecular Automata as Language Recognizers

Preliminaries

Finite State Automata A finite state automaton (sometimes just "finite automaton") is a notional device that reads strings of symbols drawn from a finite alphabet and either accepts them or rejects them. Languages are sets of strings, and the language recognized by the automaton is defined to be the set of all strings which the automaton accepts.

We now consider how the automaton makes its decision. The automaton operates in a series of discrete steps, and each step consumes one symbol of the input string. The automaton exists in one of finitely many abstract states, and each step of operation causes a transition between states. A transition function describes which next states the automaton may enter upon consuming a particular symbol in a particular state. One state is designated as the start state; this is the state of the automaton before any input is read. Upon starting, the automaton reads the symbols from the input string and makes the indicated transitions until the string has been exhausted. If the final state in which it finds itself belongs to the designated subset of accepting states, we say that it has accepted the input string.

Formally, an alphabet Σ is a finite set of uninterpreted symbols. In examples, the sets $\{0, 1\}$ and $\{a, b, c, \dots\}$ are common; in computing ASCII and Unicode are used. While it may seem plausible that the nucleotides $\{A, T, G, C\}$ should be used directly as the alphabet in molecular implementations of automata, device design, such as the use of restriction enzymes with multi-nucleotide recognition sites, often dictates a less dense representation; for instance, each of a few unique short oligonucleotides may stand for a symbol of the alphabet.

A string, or word, is a finite ordered sequence of symbols. A string of no symbols is written ε. Two strings s_1 and s_2 can be concatenated to form $s_1 s_2$. Any set of strings is called a language. The notion of concatenation is extended to languages, $L_1 L_2 = \{s_1 s_2 | s_1 \in L_1 \wedge s_2 \in L_2\}$. A special operator, the Kleene star, finitely iterates concatenation: $L^\star = \varepsilon \cup L \cup LL \cup LLL \cup \dots$ If we identify single-symbol strings with symbols, the set of all strings over an alphabet Σ is $\Sigma^\star$.

Given a finite set of states Q, the transition function of an automaton is $\delta\colon Q \times \Sigma \to Q$. The start state is $S \in Q$, and the accepting (sometimes called final) states are $F \subseteq Q$. To formalize the operation of the automaton, we define the configuration of the automaton to be its current state paired with the unread input. The initial configuration for input w is (S, w). The relation $\vdash$ describes one step of operation: $(s, x) \vdash (t, y)$ if $(\exists a \in \Sigma) x = ay \wedge \delta(s, a) = t$. The relation $\vdash^\star$ is the reflexive, transitive closure of $\vdash$ and describes the complete operation. If $(S, w) \vdash^\star (s, \varepsilon)$ then the automaton accepts w if and only if $s \in F$.

Finite state automata as described are properly called deterministic. Nondeterministic finite state automata are an apparent generalization that permits transitions that read strings rather than single symbols from the input at each step of operation, which is described by a transition relation $\delta \subseteq Q \times \Sigma^\star \times Q$. With a nondeterministic automaton, we say it has accepted the input string if there is an accepting state among all states it may have reached while consuming the entire input string.

In the conventional visual representation of a finite state automaton, the transition diagram, (example shown in Fig. 1), the states are shown as vertices in a graph, and the transitions as directed edges labeled with symbols (or with strings for nondeterministic automata).

The class of languages that can be recognized by a finite state automaton is known as regular languages; the term itself derives from an alternative description of the same class using regular grammars (see below). Nondeterministic automata recognize the same class of languages but may require fewer states for the same language.

The fact that nondeterministic automata are no more powerful than deterministic ones means that a number of variations that seem to be "in between" can be freely used. For instance, a deterministic automaton with a partial transition function δ, i. e., in which the transition out of some particular state on some particular input symbol is not defined is really a nondeterministic automaton. Rather than having an explicit "stuck" state with a self-loop for all input symbols, as it would be required in a deterministic description, the automaton is stuck by virtue of not having any way out of a state. This is used in Benenson's automata description (Sect. "Benenson's Finite Automaton").

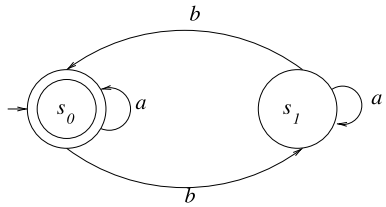

s	σ	$\delta(s,\sigma)$
s_0	a	s_0
s_0	b	s_1
s_1	a	s_1
s_1	b	s_0

a

b

current state	remaining input
s_0	*baaba*
s_1	*aaba*
s_1	*aba*
s_1	*ba*
s_0	*a*
s_0	ε

c

Molecular Automata, Figure 1
Example finite state automaton. The alphabet $\{a, b\}$, states $\{s_0, s_1\}$, start state s_0, accepting states $\{s_0\}$, and transition function given by the table in a define a deterministic finite state automaton, graphically shown in b. This automaton accepts precisely the strings that contain an even number of *b*'s. In c, the actions of the automaton in accepting the string *baaba* are shown

Finding functioning encodings for the symbols, states, and transitions in a molecular implementation is a challenging task, and therefore prototypes have dealt with small numbers of symbols and states, and it is of interest to know how many different automata can be encoded in a particular prototype. Considering deterministic automata alone, given that the transition functions δ maps $Q \times \Sigma$ to Q, the number of possible such functions is $|Q|^{|Q||\Sigma|}$; there are $|Q|$ choices for the initial state, and there are $2^{|Q|}$ choices for accepting states (any subset of the set of states). Thus, there are $|Q|2^{|Q|}|Q|^{|Q||\Sigma|}$ different automata descriptions. Of course, many of these describe automata that are identical up to a relabeling of the states; moreover, automata may be different and yet recognize the same language, so these calculations should not be taken too seriously.

More Powerful Automata Pushdown automata are a natural extension of finite state automata; these devices include a notionally infinite pushdown stack. This structure allows the device to write an unbounded number symbols to its scratch memory, but at any time only the most recently written symbol can be read back (which also erases it). These devices are capable of recognizing a larger class of languages, called the context-free languages.

Turing machines add two pushdown stacks to a finite state automaton, or, equivalently, an infinite read/write tape. These devices are capable of recognizing any language. All plausible models of what can be computed that have been put forward have been shown equivalent to Turing machines; we say that they describe universal computation.

Wang Tiles and Self Assembly One form of universal computation has been designed from a specific set of *Wang tiles* [47]. These are a mathematical model wherein square unit tiles are labeled with specific "glue" symbols on each edge. Each tile is only allowed to associate with tiles that have matching symbols, and the tiles cannot be rotated or reflected. The computational interest of Wang tiles is the question of proving whether they can be used to tile a plane. Any Turing machine can be translated into a set of Wang tiles; the Wang tiles can tile the plane if and only if the Turing machine will never halt.

Prototypes

Molecular Automata as Language Recognizers The encoding of finite automata states and transitions using DNA and restriction enzymes was first proposed by Rothemund [34]. A detailed design was also offered by Garzon [19]. Shapiro's group extended this idea to an autonomously operating cascade of cleavages of a double-stranded DNA directed by oligonucleotides [1,6,8]. In effect, this group demonstrated the first molecular equivalent of a finite automaton. What is now sometimes called the Shapiro–Benenson–Rothemund automaton consists of a mixture of two groups of DNA molecules (input and "software", i. e., transition rules) and the *FokI* restriction enzyme ("hardware"). The automaton from [6] (Fig. 3) is described in Sect. "Benenson's Finite Automaton".

DNA has also been applied in the construction of Wang tiles and general self-assembly algorithms. Erik Winfree was the first to note that planar self-assembly of DNA molecules can be applied to this form of universal computation [48]. DNA molecules analogous to Wang tiles can be constructed from double-crossover (DX) molecules [18], which consist of two side-by-side DNA duplexes that are joined by two crossovers (Fig. 2). They have a rigid stable body, and open sticky ends for attachment to other DX molecules. The sticky ends of these DNA tiles may be labeled with certain sequences, analogous to the symbols labeling the sides of Wang tiles. This labeling allows the sticky ends to bind only to the tile ends that have a complementary sequence of base pairs, which corresponds to the rule that restricts Wang tiles to only associate with tiles that have matching symbols. Triple crossover (TX) molecules, consisting of three side-by-side DNA duplexes are also good for this purpose [23]. It was

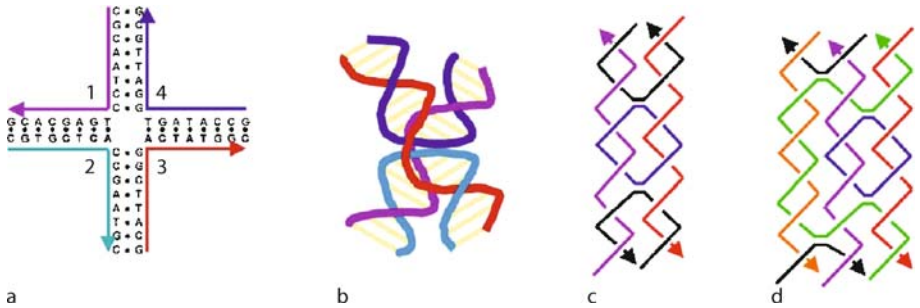

Molecular Automata, Figure 2
a A four-arm junction and **b** its three-dimensional structure; **c** a DNA DX; and **d** a DNA TX. (From [28])

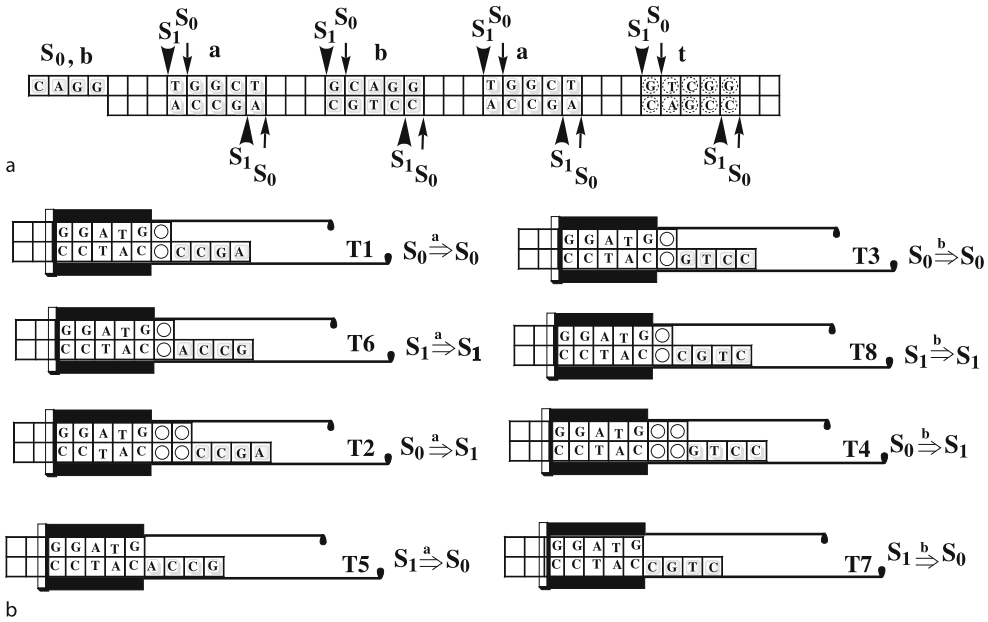

Molecular Automata, Figure 3
a The input molecule (I) is a double-stranded DNA with an overhang, and repeating motifs encoding symbols *a* (TGGCT), *b* (GCAGG) and, finally, a terminator *t* (GTCGG). The same sequences define two states (S_0 and S_1) of the finite automaton, depending on where *FokI* cleaves. Thus, S_0 in *a* is defined as a GGCG overhang and in *b* it is defined as a CAGG overhang, while S_1 is defined in *a* as a TGGC overhang and in *b* as a GCAG overhang. States are similarly defined in *t*. **b** All possible transition rules in complex with *FokI* (software-hardware complex) for a two-state, two-symbol automaton. Spacer sequences (squares with circles) between guide sequence (GGATG) and targeting overhangs define the change of states, because they serve to adjust the cleavage position within the next symbol. Overhangs define the initial state and symbol that is read by the transition rules

shown that universal computation could be accomplished by self assembling these DNA tiles [50]. An appealing interpretation is that a two-dimensional self-assembling structure can be used to represent (in one dimension) the time development of a one-dimensional (in the other dimension) cellular automaton. The first example of computing performed by DNA self-assembly [29], a four-bit cumulative XOR, is described in Sect. "DNA Self-Assembly". This is followed in Sect. "Algorithmic DNA Self-Assembly" by a detailed description of algorithmic self-assembly of DNA tiles [5,35,36] used in the construction of a binary counter and a Sierpinski triangle.

Benenson's Finite Automaton The automaton described by Benenson encodes finite automata states and transitions using DNA and restriction enzymes. The input (I) consists of repetitive sequences, with groups of five base pairs (separated by three base-pairs in [6]) which denote input symbols (Fig. 3a), and, upon enzymatic cleavage, the state of the automaton as well. Two symbols in

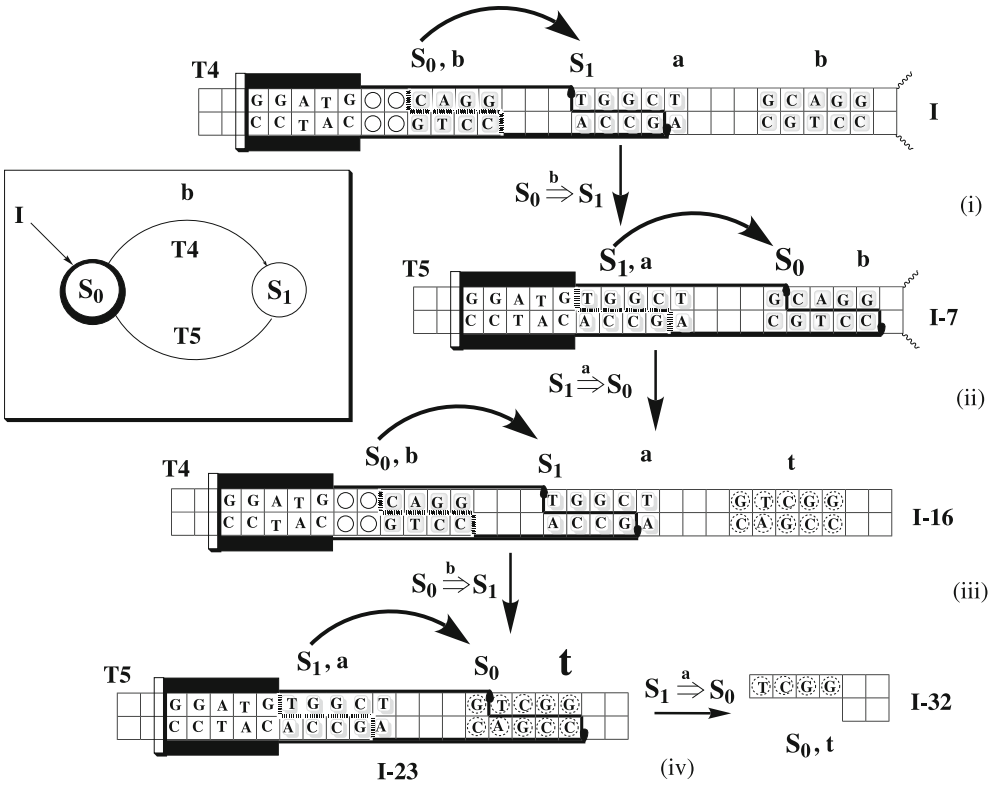

Molecular Automata, Figure 4

The effect of two transition rules, T4 and T5 on an input I from Fig. 3a is shown. The overhang at I (CAGG) defines an initial state (S_0) and a current input symbol, *b*. The input encodes the string *baba*. For clarity, only bases in overhangs, symbols, and the guide sequence are shown. (i) Complexation between T4 and I directs the cleavage of I to I-7: Overhang CAGG on I is complementary to an overhang GTCC on the T4 complex. Thus, the automaton is at the beginning of its "calculation" in the state S_0 and will "read" the next symbol *b* on the input. T4 complexes with the input and cleaves at the position 9/13 bases away within the region defining the symbol *a*. Upon this cleavage the whole complex disintegrates, leaving the shortened input (I-7) with a new overhang in place, TGGC, representing the next state S_1 and the current symbol *a*. Thus, the automaton performed according to a transition rule T4 an elementary step reading the first input symbol, *b*, and transitioned from state S_0 into state S_1, moving along the input (tape) to the second symbol, *a*. (ii) The T5 complex recognizes overhang at the I-7 and the input again at the 9/13 position, which is now within the next symbol, *b*; this leaves the new CAGG (S_0) overhang. Thus, the automaton transitioned from state S_1 into state S_0, having read the input symbol *a*. It also moved to the following input symbol *b*, on the (ever shrinking) input, now at I-16. (iii) The automaton (actually, the T4 transition rule) recognizes and cleaves I-16 and moves to the next symbol *a*, transitioning in this process to S_1 state and producing I-23. (iv) In the last step, I-23 is cleaved by T5 to I-32, producing the state S_0 in the terminator symbol

the input are *a* (TGGCT) and *b* (GCAGG). The cleavage that leaves the first four bases in an overhang is defined as a transition to an S_0 state, while the cleavage that leaves the second four bases in an overhang is defined as a transition to an S_1 state.

Interactions of an input with transition rules occur as follows: The readable input starts with an overhang of four bases, and this overhang is recognized by a complementary overhang on a transition rule (software). All transition rules are complexed with *FokI* (this is called the software-hardware complex). *FokI* recognizes the "guide" sequence in transition rules (GGATG), and then, upon interactions with input, cleaves the input at the constant position, nicking the DNA helix 9 and 13 positions away from the guide sequences (Fig. 4). In this way, the transition rule directs the cleavage of an input. As this process generates a new overhang, within the next symbol, we say that cleavage moves the automaton to read the next symbol, and that at the same time it defines the new state of the automaton. The structure of the transition rules defines whether a state of the automaton changes or not, by a distance between an overhang and the guide sequence.

This process is a cascade, because a new overhang can be recognized by another (or the same) transition rule.

With two states and two symbols, we can have a total of eight transition rules (Fig. 3b), defining all possible transitions, upon reading a symbol. (Keinan and colleagues also reported three-state, three-symbol automata [40].)

We say that this automaton (or in biochemistry a cascade) is capable of consuming input strings over a two-letter alphabet $\{a, b\}$ and transitioning between states depending on the input and in accordance with a preprogrammed control, defined through the presence of an arbitrary mixture of transition rule complexes (T1–T8). The mechanism of this cascade is best explained through an example, and we provide one in Fig. 4. For example, the transition function consisting of only transition rules T4 and T5 can move the automaton back and forth between states S_0 and S_1 endlessly, as long as it is moving along an input made of symbols a and b in alternation, $baba\cdots$, with the final result S_0. We invite the reader to try any other combination of transition rules in Fig. 3b on the same input. Some of them will stall the automaton.

Thus, through selecting a set of transition rules, or software/hardware complexes (transition rules of the form: if you find the system in state S_m and read a symbol $\sigma \in \{a, b\}$ from the input, go to state S_n and move to the next input symbol), one could write a molecular "program" able to read and degrade an input DNA molecule, which encodes a defined set of symbols. Input molecules for which transitions always succeed will be degraded up to a terminator symbol, which will read the final state of an automaton ("answer"). If there is no transition rule present in the system that matches the current state and the current symbol of the input at a certain stage of the degradation, the automaton stalls and the input molecule is not degraded to a terminator symbol. Thus, it can be said that this system indeed works as a finite-state automaton and thus recognizes the language abstractly specified by its set of transition rules.

Whilst the laboratory demonstration was of finite state automata, one could envisage pushdown automata or Turing machines; indeed Shapiro has a patent on a molecular Turing machine [38].

DNA Self-Assembly The first DNA self-assembling automaton was a four-bit cumulative XOR [29]. The Boolean function XOR evaluates to 0 if its two inputs are equal, otherwise to 1. The "cumulative" (i. e., multi-argument) XOR takes Boolean input bits $x_1, \ldots, x_n$, and computes the Boolean outputs $y_1, \ldots, y_n$, where $y_1 = x_1$, and for $i > 1$, $y_i = y_{i-1} \mathrm{XOR} x_i$. The effect of this is that y_i is equal to the even or odd parity of the first i values of x. Eight types of TX molecules were used in the implementation: two corner tiles, two input tiles, and four output tiles. The corner tiles connect a layer of input tiles to a layer of output tiles. The input tiles represent $x_i = 0$ and $x_i = 1$. There are two ways to get each of the two possible outputs of a bitwise XOR, and so four output tiles were used, leaving some interpretation to the human observer. One output tile represents the state with output bit $y_i = 0$ and input bits $x_i = 0$ and $y_{i-1} = 0$; another the state with output bit $y_i = 0$ and input bits $x_i = 1$ and $y_{i-1} = 1$. The remaining output tiles represent the two states with $y_i = 1$. The actual computation of the XOR operation is accomplished by harnessing the way the output tiles connect to the input tiles. Each output tile (y_i) attaches to a unique combination of one input tile (x_i) and one output tile (y_{i-1}), and leaves one sticky end open that encodes its own value (y_i) so that another output tile may attach to it. Thus, only the output tiles that represent the correct solution to the problem are able to attach to the input tiles.

Algorithmic DNA Self-Assembly As another type of DNA based paradigm capable of autonomous computing, we give an example of algorithmic self-assembly of DNA tiles, as first suggested, and then implemented by Winfree. Algorithmic self-assembly is an extension of various DNA nanotechnologies developed over the years by Seeman's group, and is based on a vision that one can encode in a set of tiles the growth of an aperiodic crystal. While aperiodic crystals seem on the surface irregular, the position of each component of such crystal is actually precisely encoded by a program.

The approach is based on a rigid set of DNA tiles, such as double or triple crossover tiles. Unlike standard (i. e., crossover-free) structures, such as three- and four-way junctions, these molecules are sufficiently rigid to define precise positions of other tiles interacting with them, which could lead to a regular crystal growth. This has led to many impressive periodic 2D structures. Interactions between tiles occur through complementary overhangs. For example, each of the DAO-E tiles in Fig. 7 projects four overhangs (schematically presented as different shapes), each of which can be independently defined.

We examine two computations that have been demonstrated by means of algorithmic self-assembly: a binary counter, and a Sierpinski triangle.

The binary counter [5,36] uses seven different types of tiles: two types of tiles representing 1, two types representing 0, and three types for the creation of a border (corner, bottom, and side tiles). The counter (Fig. 5) works by first setting up a tile border with the border tiles—it is convenient to think of the "side" border tiles as being on the right, as then the counter will display the digits in the customary order. Two border tiles bind together with

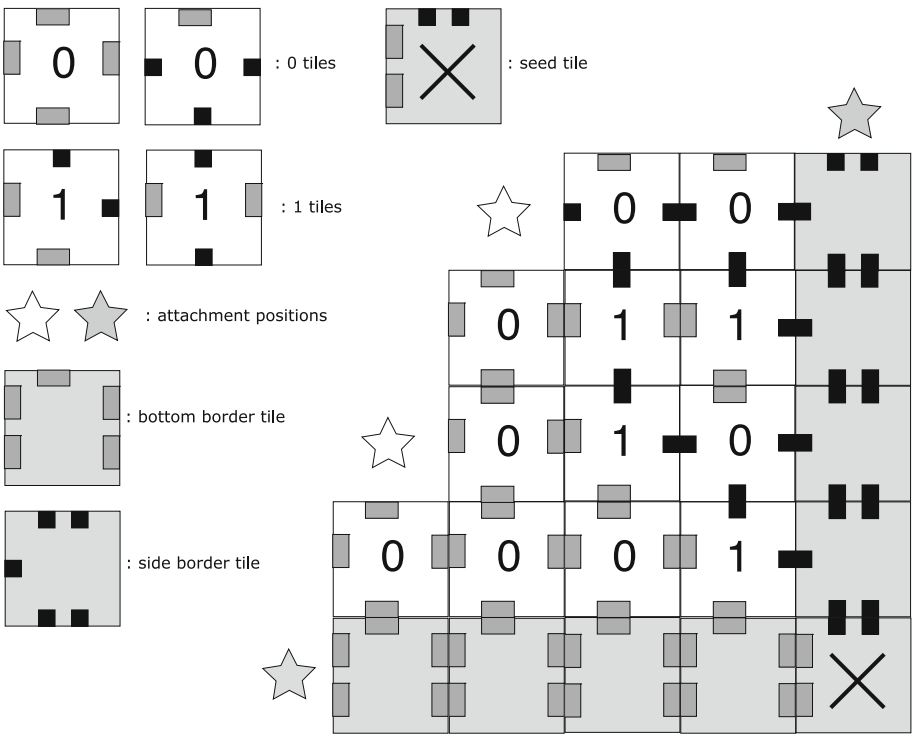

Molecular Automata, Figure 5

A binary counter in the process of self-assembly. The seed tile starts off the assembly. The right side and bottom border tiles connect to each other with double bonds, while all the other tiles connect with single bonds. A tile needs two single bonds (or one double bond) to form a stable attachment to the structure; the marked attachment positions show where a tile can form a stable attachment

a strong, double bond, while all other tiles bind to each other and to border tiles with a single bond. Since any tile except a border tile must bind to two additional tiles in order to have two bonds, but a border tile and another border tile of the correct type will form a double bond with each other, a stable border forms before other stable formations, composed of non-border tiles, are created. The bottom and side border tiles are designed such that the only tile that may bind in the border's corner (to both a side and a bottom border tile) is a specific type of 1 tile. Only one of the 0 tiles may bind to both this 1 tile and the bottom of the border, and this type of 0 tile may also bind to itself and the bottom of the border, and thus may fill out the left side of the first number in the counter with leading zeros. The only type of tile which may bind both above the 1 in the corner and to the right side of the border is the other type of 0 tile, and the only tile which may bind to the left of it is a 1 tile—we get the number 10, or two in binary. The tile binding rules are such that this can continue similarly up the structure, building numbers that always increment by one.

We now look at how an aperiodic crystal corresponding to the so-called *Sierpinski triangle* was constructed [35]. First we consider what this fascinating figure is in the abstract. The Sierpinski triangle (or *gasket*), properly speaking, is a fractal, obtained from a triangle by iteratively removing an inner triangle half its size *ad infinitum*; Fig. 6 is an approximation up to ten iterations. Any physical realization will provide only a finite number of iterations and can thus only be an approximation of the true fractal; some realizations have been fabricated in the past with interesting physical properties [20].

A peculiar connection exists with Pascal's triangle (Table 1), the table of binomial coefficients $\binom{n}{k}$. Each entry in Pascal's triangle is the sum of the two entries right above it, according to the equality $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$. If we turn an initial section consisting of 2^m rows of Pascal's triangle into an image by mapping odd numbers to black and even numbers to white (Table 2), we get an approximation to the Sierpinski triangle; various generalizations are possible, see [21,32]. Thus, we could generate the Sierpinski triangle by first generating Pascal's triangle, row by row, using the above formula. But instead of computing the entries of Pascal's triangle, which are large numbers, we can compute modulo 2, i. e., only keep track of whether the numbers are odd or even. Supposing numbers were writ-

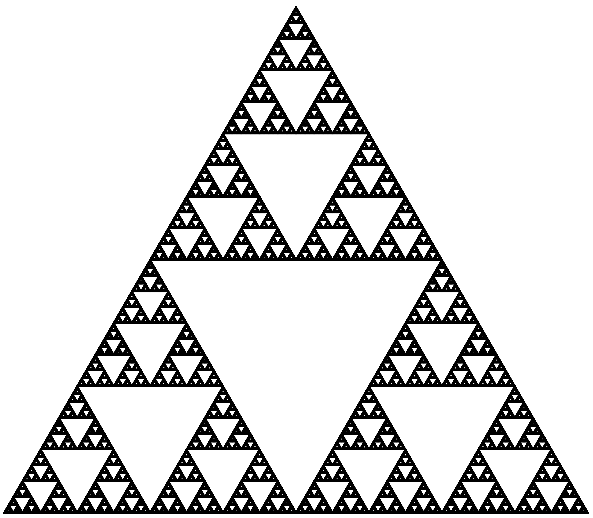

Molecular Automata, Figure 6
A Sierpinski triangle as a fractal image

Molecular Automata, Table 1
The first eight rows of Pascal's triangle

							1								
						1		1							
					1		2		1						
				1		3		3		1					
			1		4		6		4		1				
		1		5		10		10		5		1			
	1		6		15		20		15		6		1		
1		7		21		35		35		21		7		1	

Molecular Automata, Table 2
Marks indicate the odd entries in the first eight rows of Pascal's triangle

							•								
						•		•							
					•				•						
				•		•		•		•					
			•								•				
		•								•			•		
	•				•				•					•	
•			•		•		•		•		•		•		•

Molecular Automata, Table 3
The computational schema for the construction of an approximation to a Sierpinski triangle

							1								
						1	0	1							
					1	0	0	0	1						
				1	0	1	0	1	0	1					
			1	0	0	0	0	0	0	0	1				
		1	0	1	0	0	0	0	0	1	0	1			
	1	0	0	0	1	0	0	0	1	0	0	0	1		
1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	

ten in binary, this is equivalent to keeping only the least significant bit. In that case, the addition operation degenerates into an XOR. This is the chosen method for DNA self-assembly of a Sierpinski triangle (Table 3).

The set of tiles that worked the best for the synthesis of the Sierpinski triangle consists of 6 tiles, three **R** and three **S** tiles (Fig. 7). The **R** tiles interact with the **S** tiles, and vice versa. The **R** tiles in row i calculate (display their overhang) outputs to the **S** tiles in the row $i + 1$, based on inputs (interactions with complementary overhangs) from the **S** tiles in row $i - 1$ (Fig. 7c). Thus, each row is composed of either **R** or **S** tiles. There are four tiles (**S-00**, **S-11**, **R-00**, **R-11**) which are binary-encoded with values of 0 (Fig. 7a). These are incorporated in the growing crystal between two adjacent tiles in row $i - 1$ that have both the value 0 or both the value 1. There are two tiles (**R-01** and **S-01**) with values of 1 (Fig. 7b), and they have additional loops that will show as bright regions on the atomic force microscopy (AFM) characterization (cf. Fig. 9). These tiles are incorporated in between two tiles in row $i - 1$ that display different values. Due to the C2-symmetry of the displayed overhangs two tiles, there is no need to define any further tiles. In fact, the tile **R-11** is never used in this particular calculation, and is redundant.

Assembly proceeds as follows: Winfree and colleagues first constructed a long input (scaffold strand) DNA, which contained binding sites for the **S-00** tiles at regular distances. The input is necessary to ensure longer periods of growth (more than 70 individual rows were grown from the input). The input also contained a small number of randomly distributed sites to which the **S-01** tile binds. This is an initiation site, around which aperiodic growth occurs. Thus, the first row, immediately on the input strand, contains mostly **S-00** tiles, and here and there a single **S-01** tile. The second row of **R** tiles assembles on these tiles, with **R-00** between all **S-00** tiles and **R-01** on each side of **S-01** (i. e., between **S-01** and **S-00** tiles). An example of this type of crystal growth is shown in Fig. 8, while Fig. 9 shows the actual AFM results.

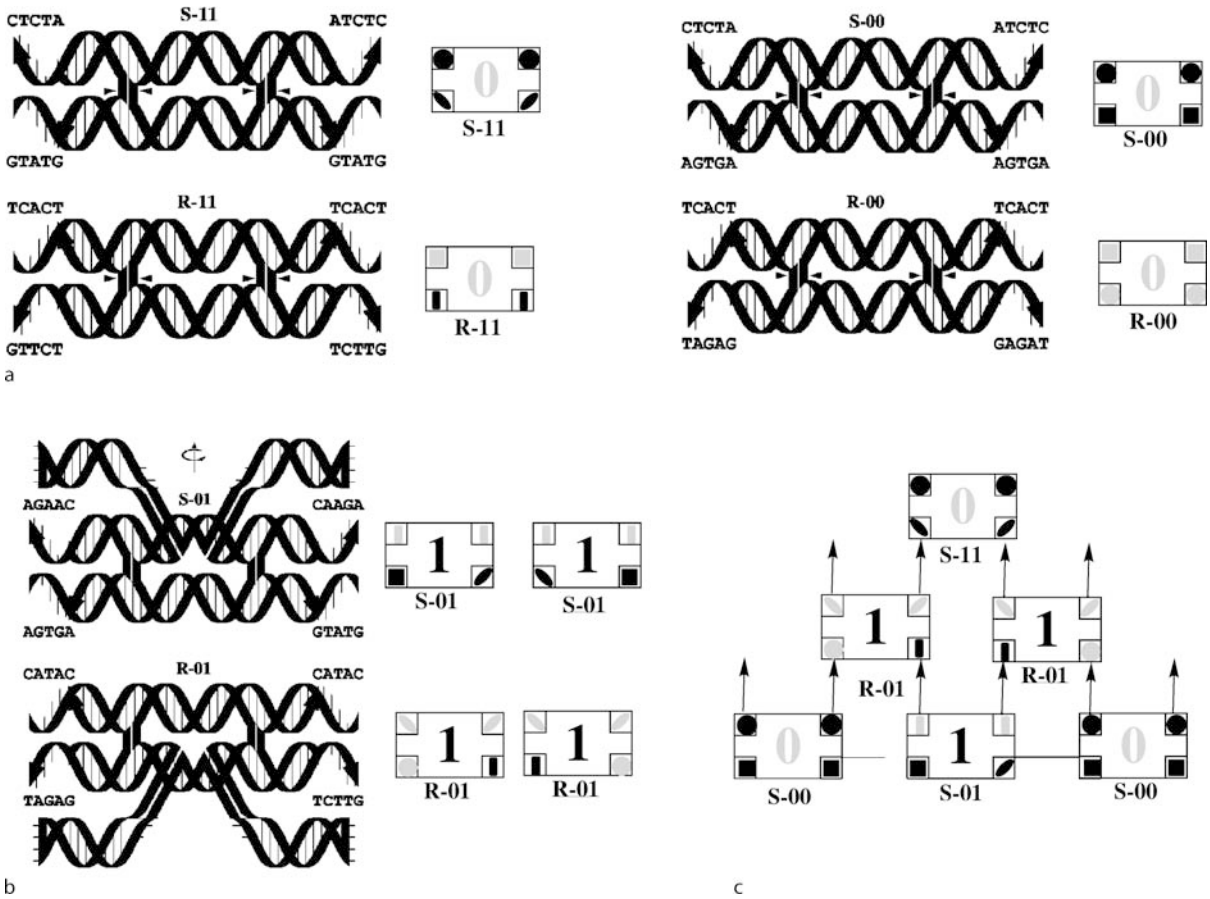

Molecular Automata, Figure 7

The DAO-E Sierpinski set of tiles, their corresponding schematic representations and the mechanism of their assembly. Each tile is assigned a schematic rectangular representation. An overhang is represented by the shapes in each corner, with complements being assigned the same shapes, but in black and gray. In c we give an example of assembly

In principle, tiling of surfaces is Turing complete. While it is certain that many interesting aperiodic crystals can be encoded using DNA tiles, the main limitation for further progress at this moment is the high error rate (1–10%) of the addition of individual tiles. Explicit error correction using some form of redundancy or "self-healing" tile sets to recover from errors is called for [49].

Molecular Automata as Transducers and Controllers

Preliminaries

Finite State Transducers A finite state transducer is a variation of the deterministic finite state automaton which, in addition to consuming one input symbol at each step of operation, produces as output a string over some alphabet. A special case is the Mealy automaton [30], which emits one output symbol at each step, so that the length of the output is the same as the length of the input; an example is shown in Fig. 10.

Molecular Automata as Language Recognizers Shapiro's group used the same type of cascades as in their finite automata to create a molecular transducer as a proof-of-concept for a diagnostic automaton [7]. This design is explained in Sect. "Therapeutic and Diagnostic Automata".

Using a conceptually different model, Stojanovic and Stefanovic created molecular transducers from deoxyribozyme-based logic gates (see Sect. "Deoxyribozyme-Based Logic Gates as Transducers"), which have been used as components for the construction of adders (Sect. "Adders and Other Elementary Arithmetic and Logic Functions")

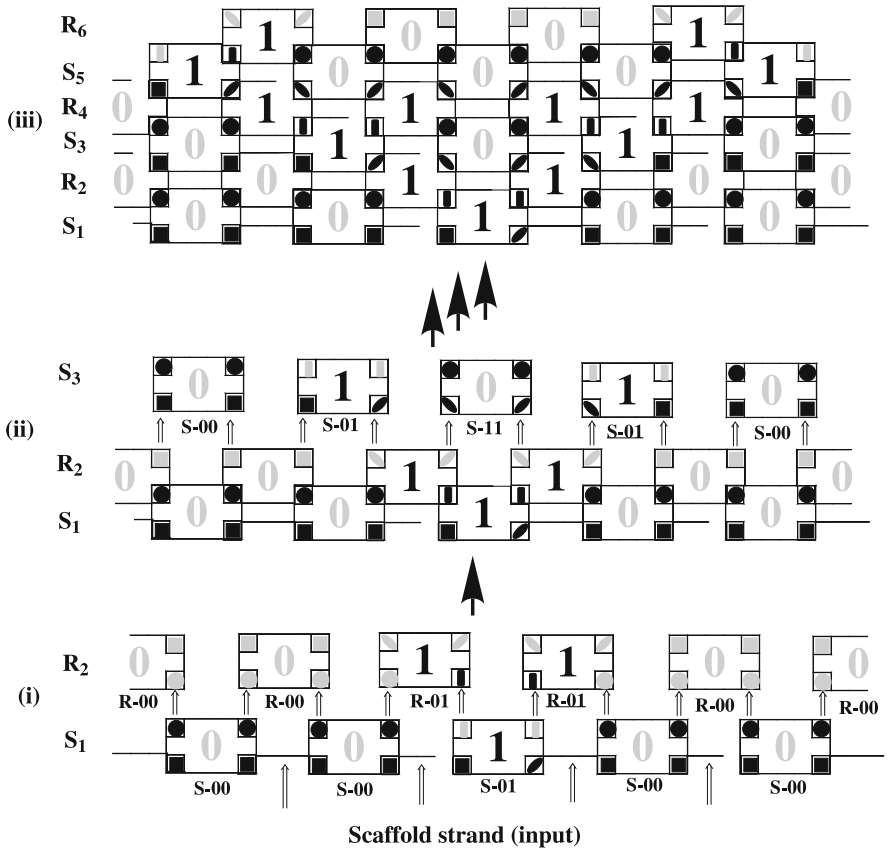

Molecular Automata, Figure 8
A schematic example of the assembly of an aperiodic 2D crystal from tiles encoding the Sierpinski triangle

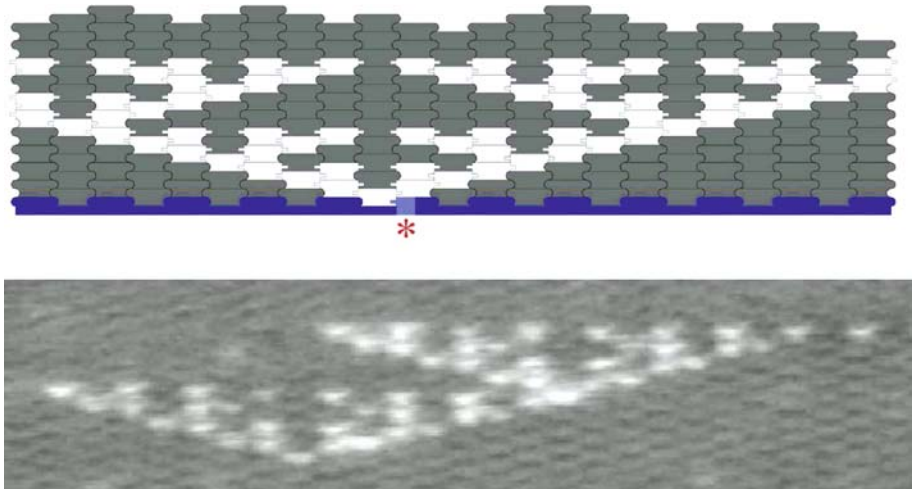

Molecular Automata, Figure 9
AFM picture (*bottom*) of a representative aperiodic crystal encoding the Sierpinski triangle in its structure (schematic, *top*). (From [35])

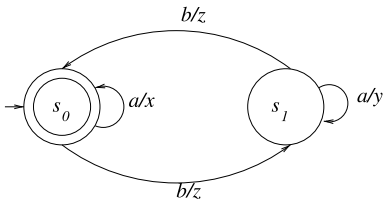

Molecular Automata, Figure 10
Example Mealy automaton. The input alphabet, the states, and the transition function are as in Fig. 1. The output alphabet is $\{x, y, z\}$. Successive runs of *a*s in the input are transduced alternately into runs of *x*s and runs of *y*s; *b*s are transduced into *z*s

and automata for games strategies (Sects. "Automata for Games of Strategy" and "Improved Automata for Games of Strategy").

Deoxyribozyme-Based Logic Gates as Transducers Simple DNA logic-gate transducers have been constructed from deoxyribozymes by Stojanovic and Stefanovic [43]. Deoxyribozymes are enzymes made of DNA that catalyze DNA reactions, such as the cleavage of a DNA strand into two separate strands, or the ligation of two strands into one. These can be modified to include allosteric regulation sites, to which specific control molecules can bind and so affect the catalytic activity.

There is a type of regulation site to which a control molecule must bind before the enzyme can complex with (i. e., bind to) the substrate, thus this control molecule promotes catalytic activity. Another type of regulation site allows the control molecule to alter the conformation of the enzyme's catalytic core, such that even if the substrate has bound to the enzyme, no cleavage occurs; thus this control molecule suppresses or inhibits catalytic activity. This allosterically regulated enzyme can be interpreted as a logic gate, the control molecules as inputs to the gate, and the cleavage products as the outputs. This basic logic gate corresponds to a conjunction, such as e. g., $a \wedge b \wedge \neg c$, here assuming two promotory sites and one inhibitory site, and using a and b as signals encoded by the promotor input molecules and c as a signal encoded by the inhibitor input molecule. Deoxyribozyme logic gates are constructed via a modular design [41,43] that combines molecular beacon stem-loops with hammerhead-type deoxyribozymes, Fig. 11. A gate is active when its catalytic core is intact (not modified by an inhibitory input) and its substrate recognition region is free (owing to the promotive inputs), allowing the substrate to bind and be cleaved.

Correct functioning of individual gates can be experimentally verified through fluorescent readouts (Fig. 21). The inputs are compatible with sensor molecules [42] that can detect cellular disease markers. Final products (outputs) can be tied to release of small molecules. All products and inputs (i. e., external signals) must be sufficiently different to minimize the error rates of imperfect oligonucleotide matching, and they must not bond to one another. The gates use oligonucleotides as both inputs and outputs, so cascading gates is possible without external interfaces. Two gates are coupled in series if the product of an "upstream" gate specifically activates a "downstream" gate. A series connection of two gates, an upstream ligase and a downstream phosphodiesterase, has been experimentally validated [44].

Prototypes

Therapeutic and Diagnostic Automata Using similar principles to their finite automata (Sect. "Benenson's Finite Automaton"), Shapiro's groups created a molecular transducer to assess the presence or absence of a series of oligonucleotides, as a proof-of-concept for a diagnostic automaton [7]. The automaton analyzes up- and down-regulated mRNA molecules that are characteristic for some pathological states (Fig. 12a). The authors introduced several new concepts in their work, primarily in order to maximize the tightness of control over a diagnostic cascade. We present here a somewhat simplified explanation that focuses on the underlining principles and the ability of these finite automata to analyze oligonucleotide inputs.

Input molecules for diagnostic automata are similar to those in Fig. 3. Each "symbol" in the input now contains recognition sites ("diagnostic" symbol), which are, upon cleavage by *FokI*, recognized by transition rules. In turn, these transition rules are regulated by the expression of genes (i. e., mRNA) characteristic for certain types of cancer.

If we are looking to diagnose an increase in certain mRNA level, there are, potentially, two types of transition rules that can cleave each of the symbols after these are exposed as overhangs through an action by *FokI*: (1) YES→YES transitions, which will lead to the next step of processing the input because a new overhang, recognizable by the next set of transition rules, will be exposed; and (2) YES→NO transitions, in which *FokI* cleaves off from the site that is recognized by the next set of transition rules, and, thus, these transitions cause the automaton to stall (i. e., no existing transition rules recognize the new overhang).

The YES→YES transition is activated by the presence of mRNA, through the removal of an inactivating (blocking) oligonucleotide from its complex with one of the strands making a transition rule (Fig. 12b). In contrast,

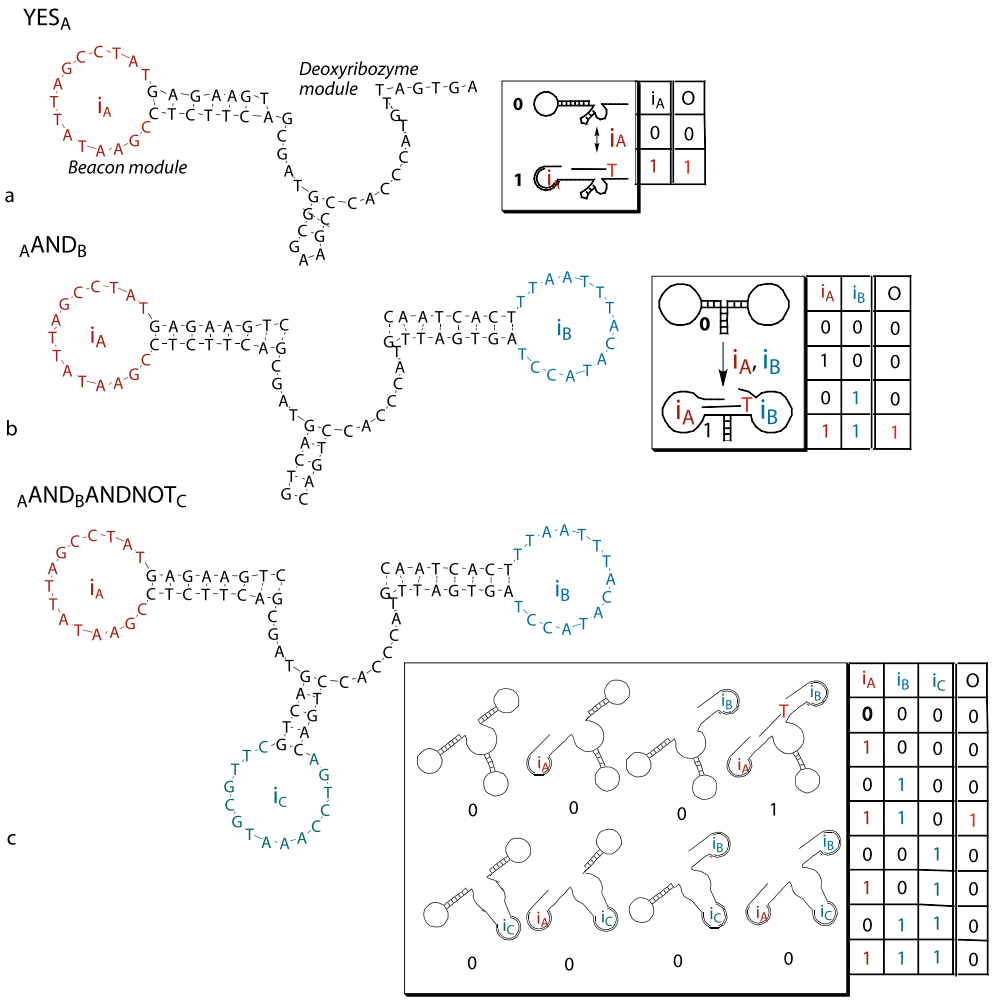

Molecular Automata, Figure 11

Examples of deoxyribozyme-based logic gates. a A YES gate showing deoxyribozyme core and molecular beacon allosteric controlling region; this gate is active in the presence of one activating input. b A two-input AND gate produces outputs only if both inputs are present. c A three-input ANDANDNOT gate produces outputs only if the two activating inputs (A and B) are present, and the inhibiting input (C) is absent

the YES→NO transitions are deactivated by an mRNA, because mRNA will compete for binding to one of the strands in the transition rule, thus, effectively breaking up the transition rule. In case we want to diagnose a downregulation of mRNA (genes 1 and 2 in Fig. 12a), the situation is reversed: The lack of characteristic mRNA leaves the YES→YES transition active, while the presence of it would break this transition rule apart. In contrast, the presence of mRNA will activate YES→NO transition by removing the inactivating oligonucleotide.

A diagnostic automaton can be turned into a "therapeutic" automaton by modifying the events at the end of the cascade. Instead of releasing a terminator symbol (cf. Figs. 3 and 4, Shapiro and colleagues described a degradation of a stem-loop, and a release of a linear oligonucleotide, which had the potential for a therapeutic (antisense) action. A further interesting detail was that a negative diagnosis could be coupled to a parallel cascade, which can release another oligonucleotide, complementary to and blocking the activity of a therapeutic oligonucleotide. The authors also described independent and parallel action of two different automata in the same solution. From the perspective of this review, it is also important that Winfree proved that these automata are capable of general Boolean computing [39].

The "therapeutic automaton" is in many respects the most challenging *ex vivo* molecular system that has ever been constructed. Although there are many reasons to

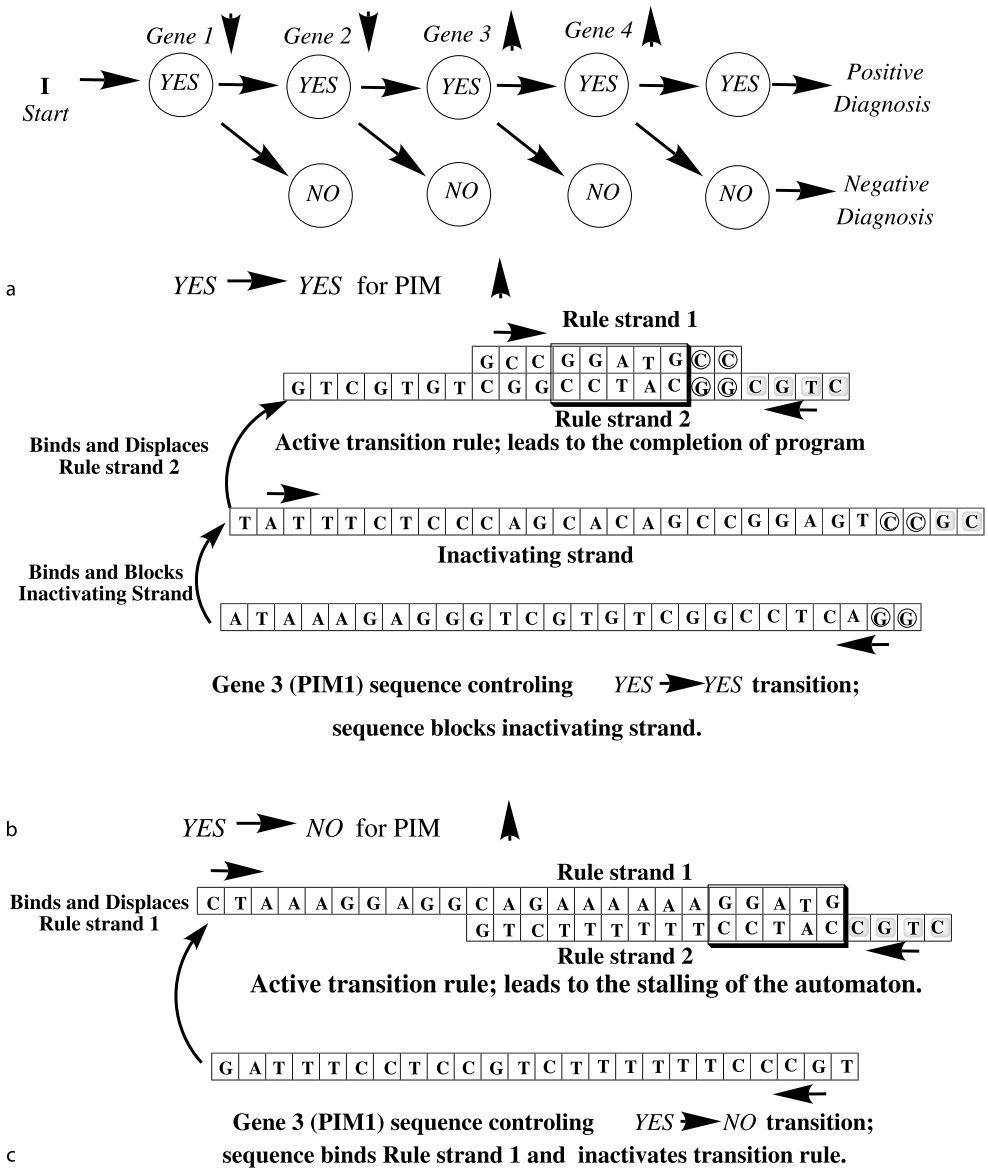

Molecular Automata, Figure 12

Therapeutic automata: a A diagnostic procedure is based on an analysis of a set of genes (e. g., Genes 1 to 4) which have the levels of expression characteristic for a particular disease; these genes regulate the transition rules that process the input molecule (see below). Input (I) molecule, similar to Fig. 3, encodes the symbols that are cleaved by these transition rules. Symbols are recognized by two types of transition rules regulated by the mRNA molecule transcribed from a gene; one type leads to continuation of the cascade, while the other leads to no further processing. b YES→YES transition rule for an upregulated gene 3 in this cascade (e. g., PIM1); in the absence of the gene this transition rule is deactivated by the inactivating strand that binds stronger to strand 2 of the rule, displacing strand 1. c YES→NO Transition rule Gene 3 expressed is formed from two software strands, by a displacement of a blocking (protector) strand. d Transition rule Gene 1 not expressed, is degraded in the presence mRNA for Gene 1, thus stalling the automaton

doubt eventual practical applications as a "doctor in cell", it will always remain an extremely impressive intellectual exercise, which will serve as an inspiration for different systems.

Adders and Other Elementary Arithmetic and Logic Functions Another type of transducing automaton is a number adder, which accepts as input two numbers expressed as strings of bits (binary digits) and produces as

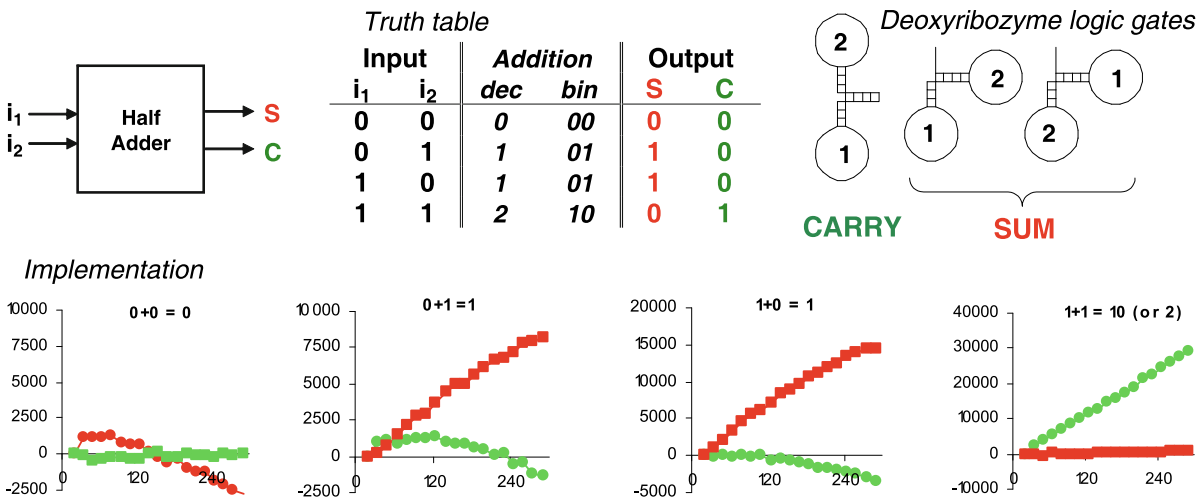

Molecular Automata, Figure 13

Components of a molecular half-adder, which adds two binary inputs to produce a sum and carry digit. *Top:* Circuit diagram, truth table of calculations and deoxyribozyme logic gates required for construction. *Bottom:* implementation of the deoxyribozyme-based half adder. The sum output is seen by an increase in red channel (tertramethylrhodamine) fluorescence ($0 + 1 = 1$ and $1 + 0 = 1$), and the carry output is seen by an increase in green channel (fluorescein) fluorescence ($1 + 1 = 2$)

output the string of bits corresponding to their sum. On the surface these may not appear to have direct relevance with diagnostic and therapeutic automata, however adders serve as extremely important computational components, and their development begins to hint at the power in automaton building through the multiple layering of logic gates.

In electronics, binary adders commonly use as a building block the full adder, which is a device that takes as input three bits (normally two digits from two numbers, and a carry-in bit) and produces a sum bit and a carry bit. The full adder is commonly realized using two half-adders; a half-adder takes as input two bits and produces a sum bit and a carry bit. A half-adder, in turn, is realized using simple logic gates. A logic gate takes some number of input bits (often just one or two) and produces one output bit as a Boolean function of the inputs, such as a negation (NOT-gate) or a conjunction (AND-gate). Adders and logic gates are simple circuits with obvious applications. Numerous molecular devices with logic gate and adder functions (as well as mixed-mode devices such as molecular-optical) have been introduced in recent years [10,11,12,13,14,15,17,22,24,51].

Logic gates built from deoxyribozymes were first described by Stojanovic and Stefanovic [43], and were subsequently used to build both a half-adder [45] and a full adder [24]. A half-adder was achieved by combining three two-input gates in solution, an AND gate for the carry bit, and an XOR, realized using two ANDNOT gates (gates of the form $x \wedge \neg y$) for the sum bit. The two substrates used are fluorogenically marked, one with red tetramethylrhodamine (T), and the other with green fluorescein (F), similar to Fig. 21, and the activity of the device can be followed by tracking the fluorescence at two distinct wavelengths. The results, in the presence of Zn^{2+} ions, are shown in Fig. 13. When both inputs are present, only the green fluorescein channel (carry bit) shows a rise in fluorescence. When only input i_1 is present or only input i_2 is present, only the red tetramethylrhodamine channel (sum bit) rises. With no inputs, neither channel rises. Thus, the two bits of output can be reliably detected and are correctly computed.

A molecular full adder was similarly constructed, requiring 7 molecular logic gates in total: three AND gates for computation of the SUM bit, and four 3-input gates for calculation of the CARRY bit (Fig. 14). The requirement for a 3-input ANDAND gate as well as several ANDNOTANDNOT gates necessitated additional deoxyribozyme logic gate design. This was achieved by precomplexing gates with a complementary oligonucleotide that would subsequently be removed by addition of inputs (Fig. 15). The length of input increased to 30 nucleotides, and served a dual function—to activate controlling elements directly and to remove precomplexed oligonucleotides when necessary. The combination of stem-loop regulation in cis and controlling elements supplied in trans is reminiscent of some classical in vivo regulatory pathways. This simple demonstration points towards the construction of even

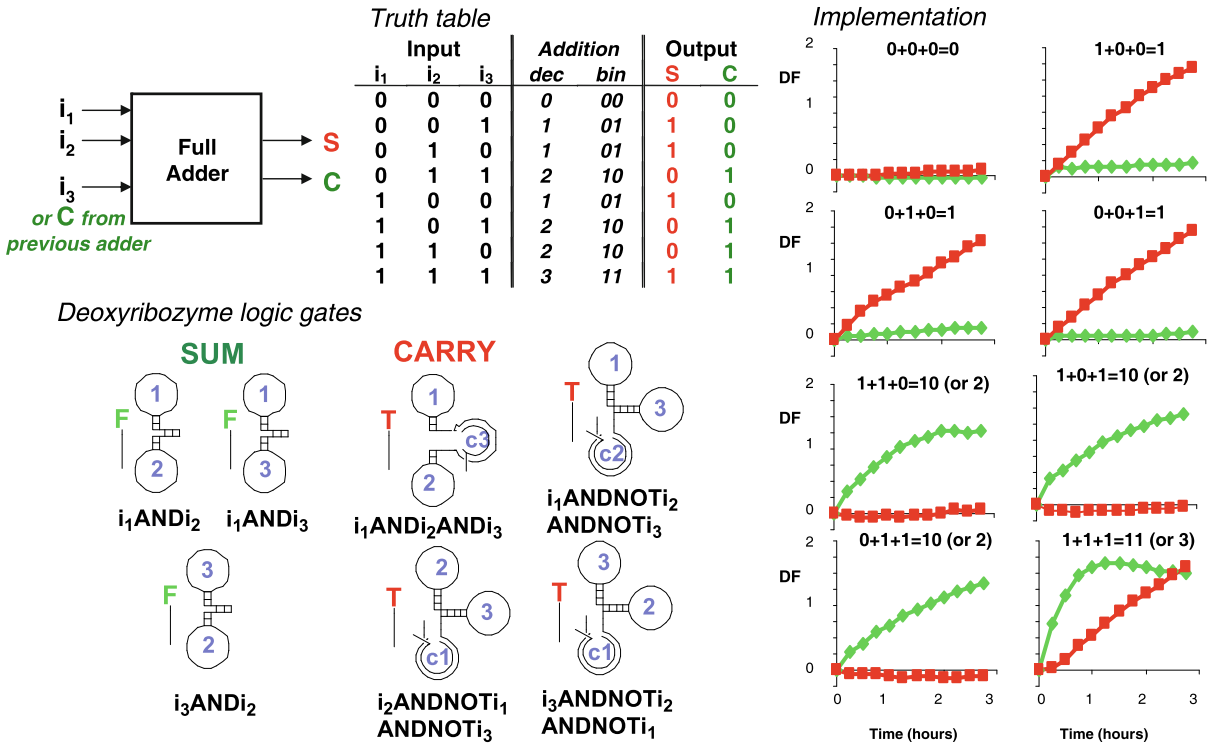

Molecular Automata, Figure 14

Components of a molecular full adder, which adds three binary inputs to produce a sum and carry digit. *Left:* circuit diagram, truth table of calculations and deoxyribozyme-based logic gates required for construction. *Right:* implementation of the deoxyribozyme-based full adder. The sum output is seen by an increase in red channel (tetramethylrhodamine) fluorescence, and the carry output is seen by an increase in green channel (fluorescein) fluorescence

more complex artificial networks, which could mimic natural systems through the use of multifunctional regulatory components.

Automata for Games of Strategy Initial construction of networks capable of simple arithmetic operations suggested the possibility that circuits of highly regulated deoxyribozymes could support complex molecular decision trees. This was subsequently demonstrated by Stojanovic and Stefanovic in the construction of the first interactive molecular transducing automaton, designed to perfectly play the game of tic-tac-toe against a human opponent [46]. To understand how this was achieved, the structure of the game will be briefly examined.

A sequential game is a game in which players take turns making decisions known as moves. A game of perfect information is a sequential game in which all the players are informed before every move of the complete state of the game. A strategy for a player in a game of perfect information is a plan that dictates what moves that player will make in every possible game state. A strategy tree is a (directed, acyclic) graph representation of a strategy (an example is shown in Fig. 16). The nodes of the graph represent reachable game states. The edges of the graph represent the opponent's moves. The target node of the edge contains the strategy's response to the move encoded on the edge. A leaf represents a final game state, and can, usually, be labelled either win, lose, or draw. Thus, a path from the root of a strategy tree to one of its leaves represents a game. In a tree, there is only one path from the root of the tree to each node. This path defines a set of moves made by the players in the game. A player's move set at any node is the set of moves made by that player up to that point in a game. For example, a strategy's move set at any node is the set of moves dictated by the strategy along the path from the root to that node. A strategy is said to be feasible if, for every pair of nodes in the decision tree for which the opponent's move sets are equal, one of the following two conditions holds: (1) the vertices encode the same decision (i. e., they dictate the same move), or (2) the strategy's move sets are equal. A feasible strategy can be successfully converted into Boolean logic implemented us-

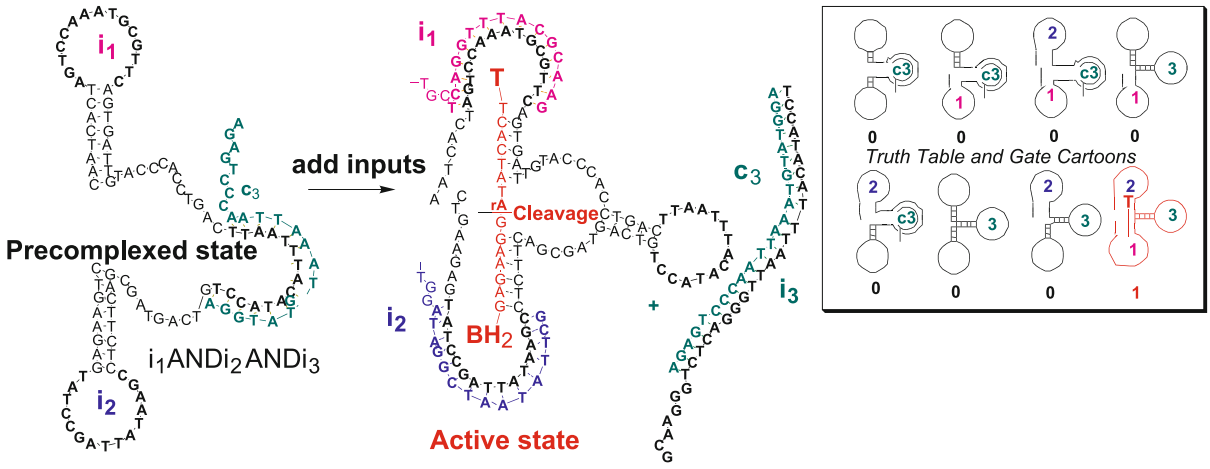

Molecular Automata, Figure 15
Three-input ANDAND gate structure

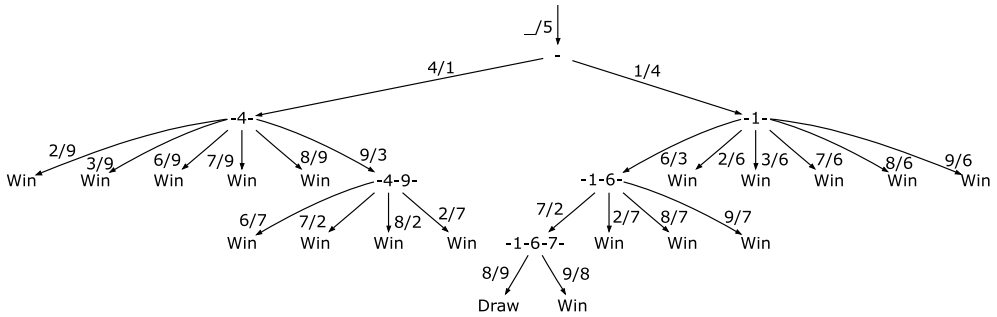

Molecular Automata, Figure 16
The chosen strategy (game tree) for the symmetry-pruned game of tic-tac-toe, drawn as the diagram of a Mealy automaton. Each state is labeled with the string of inputs seen on the path to it. Each edge is labeled *a*/*b*), where *b* is the output that is activated on input *a*

1	2	3
4	5	6
7	8	9

Molecular Automata, Figure 17
The tic-tac-toe game board with the field numbering convention

ing monotone logic gates, such as the deoxyribozyme logic gates.

In the first implementation of the tic-tac-toe automaton, the following simplifying assumptions were made to reduce the number and complexity of needed molecular species. The automaton moves first and its first move is into the center (square 5, Fig. 17). To exploit symmetry, the first move of the human, which must be either a side move or a corner move, is restricted to be either square 1 (corner) or square 4 (side). (Any other response is a matter or rotating the game board.)

The game tree in Fig. 16 represents the chosen strategy for the automaton. For example, if the human opponent moves into square 1 following the automaton's opening move into square 5, the automaton responds by moving into square 4. If the human then moves into square 6, the automaton responds by moving into square 3. If the human then moves into square 7, the automaton responds by moving into square 2. Finally, if the human then moves into square 8, the automaton responds by moving into square 9, and the game ends in a draw.

This strategy is feasible; therefore, following a conversion procedure, it is possible to reach a set of Boolean formulas that realize it, given in Table 4. (For a detailed analysis of feasibility conditions for the mapping of games of strategy to Boolean formulas, see [3].)

Molecular Automata, Table 4
Boolean formulas resulting from the tic-tac-toe game tree. The inputs are designated i_k, and the outputs are designated o_k

$$
\begin{aligned}
o_1 &= i_4 \\
o_2 &= (i_6 \wedge i_7 \wedge \neg i_2) \vee (i_7 \wedge i_9 \wedge \neg i_1) \vee (i_8 \wedge i_9 \wedge \neg i_1) \\
o_3 &= (i_1 \wedge i_6) \vee (i_4 \wedge i_9) \\
o_4 &= i_1 \\
o_5 &= 1 \\
o_6 &= (i_1 \wedge i_2 \wedge \neg i_6) \vee (i_1 \wedge i_3 \wedge \neg i_6) \vee (i_1 \wedge i_7 \wedge \neg i_6) \vee (i_1 \wedge i_8 \wedge \neg i_6) \vee (i_1 \wedge i_9 \wedge \neg i_6) \\
o_7 &= (i_2 \wedge i_6 \wedge \neg i_7) \vee (i_6 \wedge i_8 \wedge \neg i_7) \vee (i_6 \wedge i_9 \wedge \neg i_7) \vee (i_9 \wedge i_2 \wedge \neg i_1) \\
o_8 &= i_9 \wedge i_7 \wedge \neg i_4 \\
o_9 &= (i_7 \wedge i_8 \wedge \neg i_4) \vee (i_4 \wedge i_2 \wedge \neg i_9) \vee (i_4 \wedge i_3 \wedge \neg i_9) \vee (i_4 \wedge i_6 \wedge \neg i_9) \vee (i_4 \wedge i_7 \wedge \neg i_9) \vee (i_4 \wedge i_8 \wedge \neg i_9)
\end{aligned}
$$

i1	5'	TCT GCG TCT ATA AAT
i2	5'	ATC GTA TGT TGT TCA
i3	5'	GTA TAG TCT GTT TGT
i4	5'	G TAA GTG CTC AAA TGT C
i6	5'	G TCT AAT TCT CAC GGT C
i7	5'	TAG TCT GTG TGT TGT
i8	5'	TCT ATA TGA GCG TAA
i9	5'	TGT CCA TCT AAA TCC

Molecular Automata, Figure 18
Sequences of the nine oligonucleotide inputs used to indicate human move positions in the MAYA automaton

Human interaction with the automaton is achieved by the sequential addition of nine input oligonucleotides (i_1–i_9). Each oligonucleotide is 15–17 nucleotides long and has a unique sequence of A, T, C, and G's, with no more than 4 nucleotides in common between each oligonucleotide sequence (Fig. 18). The input numbering directly matches the square numbering of the tic-tac-toe game board (Fig. 17). Hence, to signify a move into square 9, the human would choose input i_9 for addition.

The game is played in a standard laboratory plate, with 9 wells representing each square of the tic-tac-toe game board. The automaton is pre-prepared by the addition of fluorescent substrate and a specific set of deoxyribozyme-based logic gates in each well. Thus the automaton moves are predetermined by the inherent logic provided in each well, and in the chosen arrangement the automaton never loses because it plays according to a perfect strategy. The arrangement of deoxyribozyme logic gates corresponding to the above formulas is given in Fig. 19. This is the initial state of the nine wells of a well-plate in which the automaton is realized in the laboratory. The automaton was named MAYA since it uses a Molecular Array of YES and AND gates to determine responses to human moves.

An example game played against MAYA is shown in Fig. 20. The play begins when Mg^{2+} ions, a required cofactor, are added to all nine wells, activating only the deoxyribozyme in well 5, i. e., prompting the automaton to play its first move into the center, since this well contains a single active deoxyribozyme which begins cleaving to produce fluorescent output. The human player detects this move by monitoring wells for increase in fluorescence emissions using a fluorescent plate reader.

After that, the game branches according to the human's inputs. In response to the automaton's first move, the human player may choose either well 1 or well 4 in this restricted game, and thus will add input oligonucleotide i_1 or i_4 to every well of the game board, so that each well of the automaton receives information on the position of the human move. This creates a chain reaction among the deoxyribozyme logic gates in each well, opening individual stem-loops complementary to the chosen input. However only one gate in one well of the board will become fully activated in response to this first human move: the YESi_1 gate placed in well 4 (if the human added i_1 to every well) or the YESi_4 gate in well 1 (if the human added i_4 to every well). The activated gate subsequently cleaves to produce fluorescent outputs, which is again detected by the human via fluorescence monitoring.

Subsequent human moves are unrestricted, so the human may choose to play in any of the remaining wells where neither player has yet moved. However at this point

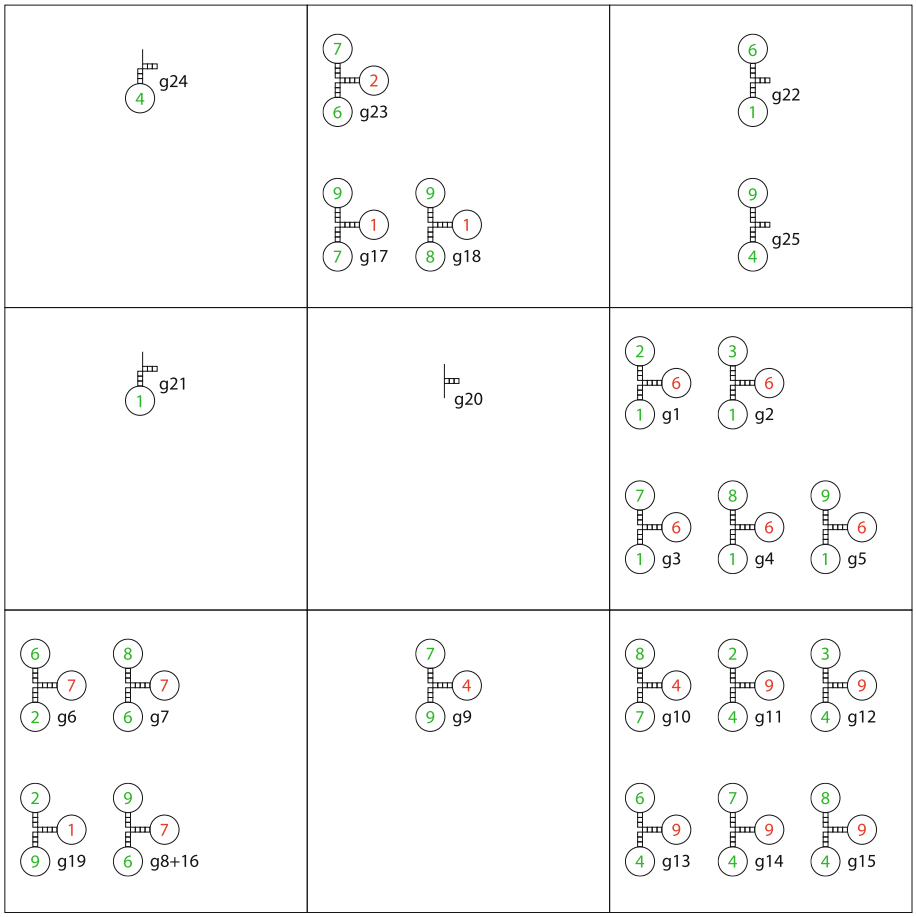

Molecular Automata, Figure 19

Realizing a tic-tac-toe automaton using deoxyribozyme logic. The center well contains a constitutively active deoxyribozyme. Each of the eight remaining wells contains a number of deoxyribozyme logic gates as indicated. In the schematic, *green* numbers are indices to the inputs that appear as positive literals in conjunctions; *red* as negative

the automaton is poised to win the game, either through a diagonal three in a row (wells 1, 5, 9; if the human chose to play in well 4) or through a horizontal three in a row (wells 4, 5, 6; if the human chose to play in well 1). Assuming perfect play, the human will choose to block an automaton win by playing into well 9 (adding input i9 to each well) or well 6 (adding input i6 to each well). This would again cause a chain reaction with complementary stem loops in every well, but again, only a single logic gate would become fully activated (either gate $i_4 \text{AND} i_9$ or $i_1 \text{AND} i_6$ in well 3, depending on the human's initial moves). The game thus continues by repeated human input addition and subsequent fluorescent monitoring until either the human makes an error leading to an automaton win, or there is a draw as illustrated in Fig. 20.

Improved Automata for Games of Strategy In order further to probe the complexity with which a molecular automaton could be rationally constructed, a second version of MAYA was constructed [27]. MAYA-II plays a non-restricted version of tic-tac-toe where the automaton still moves first in the middle square, but the human player may choose to respond in any of the remaining squares. The new automaton was also designed to be more user-friendly than the original MAYA, allowing monitoring of both the automaton and human moves using a dual-color fluorescence output system, similar to the previously constructed half and full adder [24,45] (Fig. 21).

The complexity of MAYA-II encompasses 76 possible games, of which 72 end in an automaton win and 4 end in a draw. The required arrangement of logic gates for

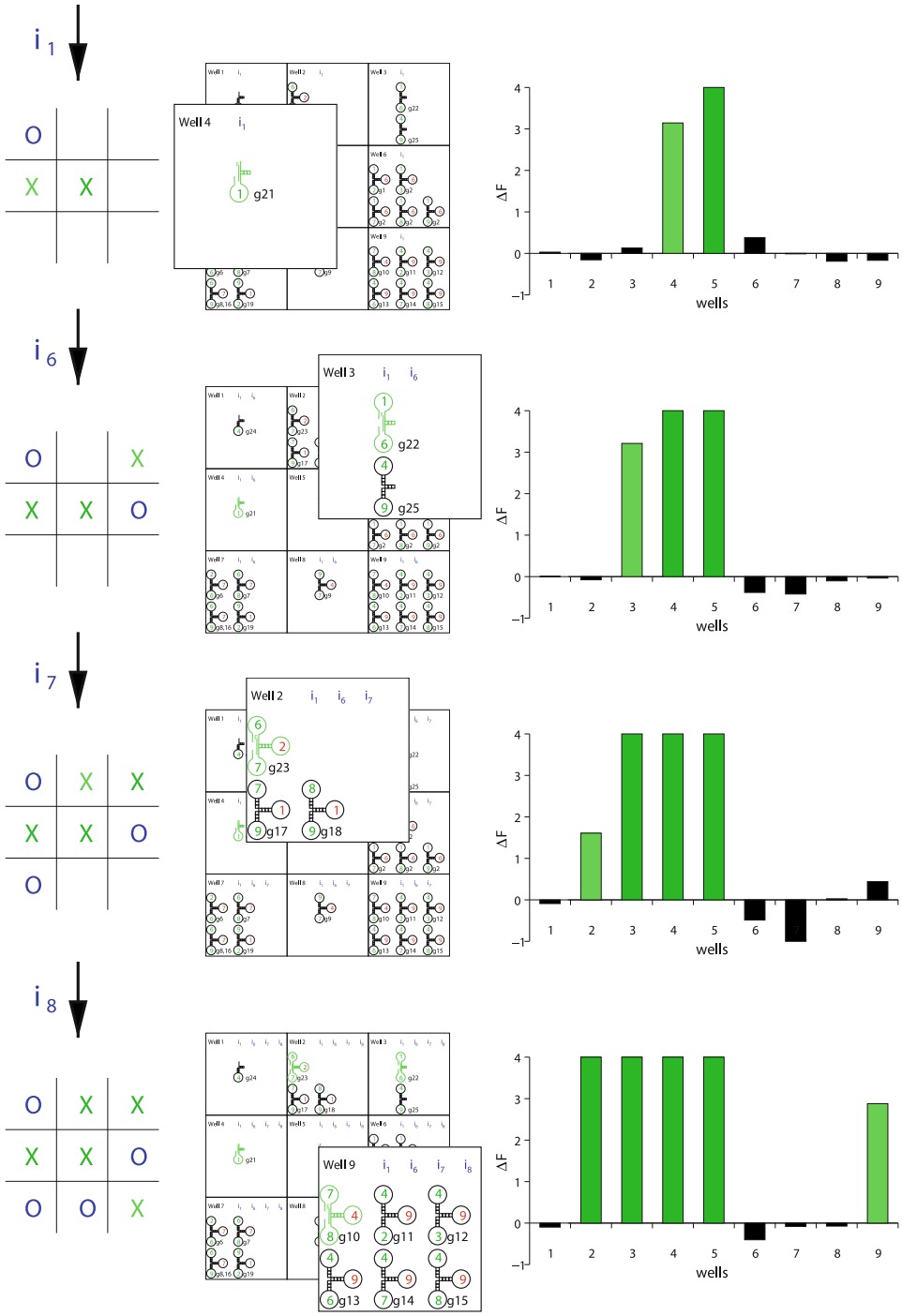

Molecular Automata, Figure 20

A game of tic-tac-toe that ends in a draw because both the automaton and its human opponent play perfectly. As the human adds input to indicate his moves, the automaton responds with its own move, activating precisely one well, which is shown enlarged. The newly activated gate is shown in *light green*. The bar chart shows the measured change in fluorescence in all the wells. Wells that are logically inactive (contain no active gates) have *black bars*, and wells that are logically active have *green bars* (the newly active well is *light green*)

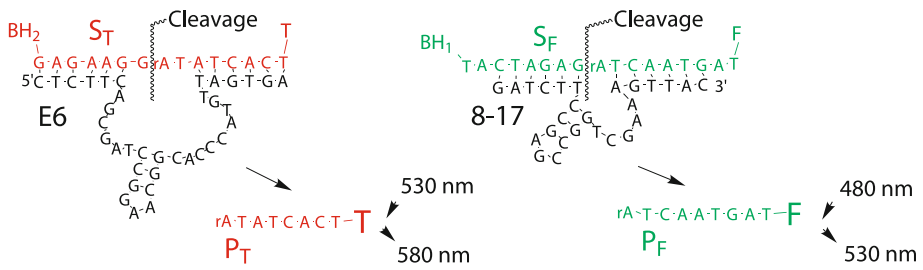

Molecular Automata, Figure 21

Dual color fluorescence output system utilized in MAYA-II. Automaton moves are determined using logic gates based on the deoxyribozyme E6 [9,24,45], which cleaves substrate S_T (dual labeled with tetramethylrhodamine and black hole quencher 2) to produce product P_T and an increase in tetramethylrhodamine fluorescence. Human moves are displayed using logic gates based on the deoxyribozyme 8.17 [24,37,45], which cleaves substrate S_F (dual labeled with fluorescein and black hole quencher 1) to produce product P_F and an increase in fluorescein fluorescence

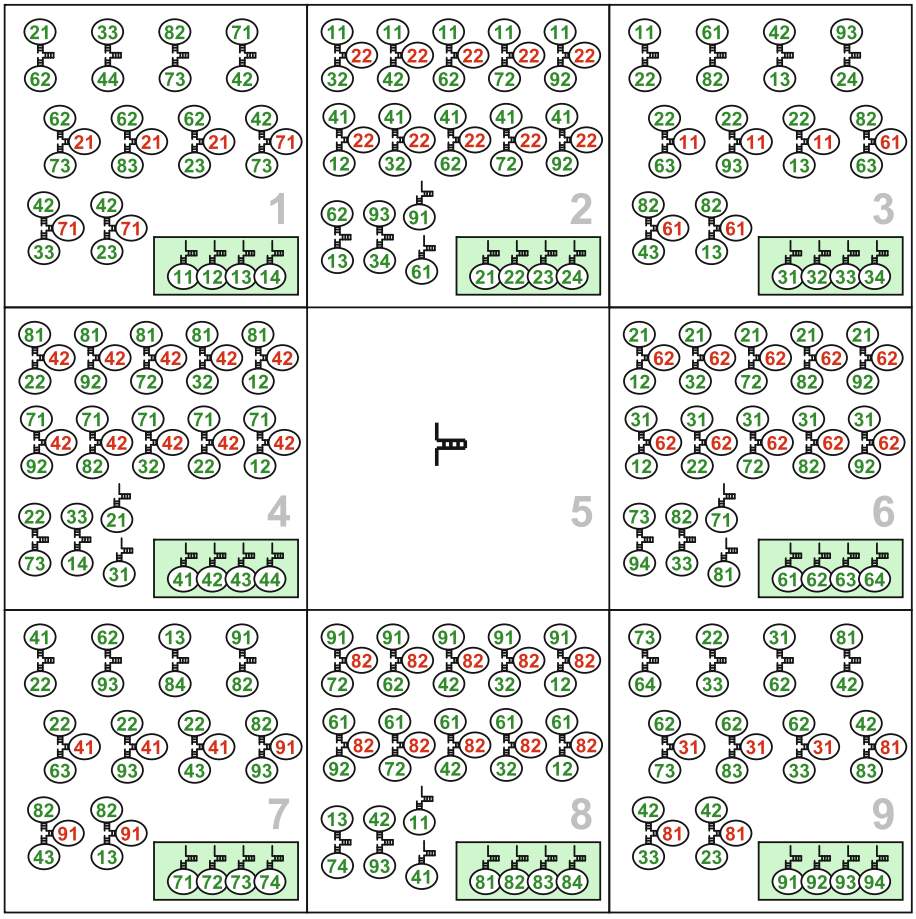

Molecular Automata, Figure 22

Realizing the MAYA-II automaton using deoxyribozyme logic. The center well contains a constitutively active deoxyribozyme. Each of the eight remaining wells contains a number of deoxyribozyme logic gates as indicated; boxed gates monitor human moves and the rest determine deoxyribozyme moves. In the schematic, *green* numbers are indices to the inputs that appear as positive literals in conjunctions; *red* as negative

a perfect automaton strategy was determined by computer modeling. However, using the gate limitations of the time, a strategy using only 9 inputs could not be determined by the program. Acceptable logic could be predicted by increasing the number of inputs to 32 different oligonucleotides. These inputs were coded i_{nm} where n refers to the position of the human move (1–4, 6–9) and m corresponds to the move order (1–4). For instance, to signify a move into well 9 on the first move, the human would add input i_{91} to all wells.

Using this input coding, the final arrangement of logic gates for MAYA-II's chosen strategy used 128 deoxyribozyme-based logic gates (Fig. 22); 32 gates monitor and display the human player's moves, and 96 gates calculate the automaton's moves based on the human-added inputs. Essentially, successive automaton moves are con-

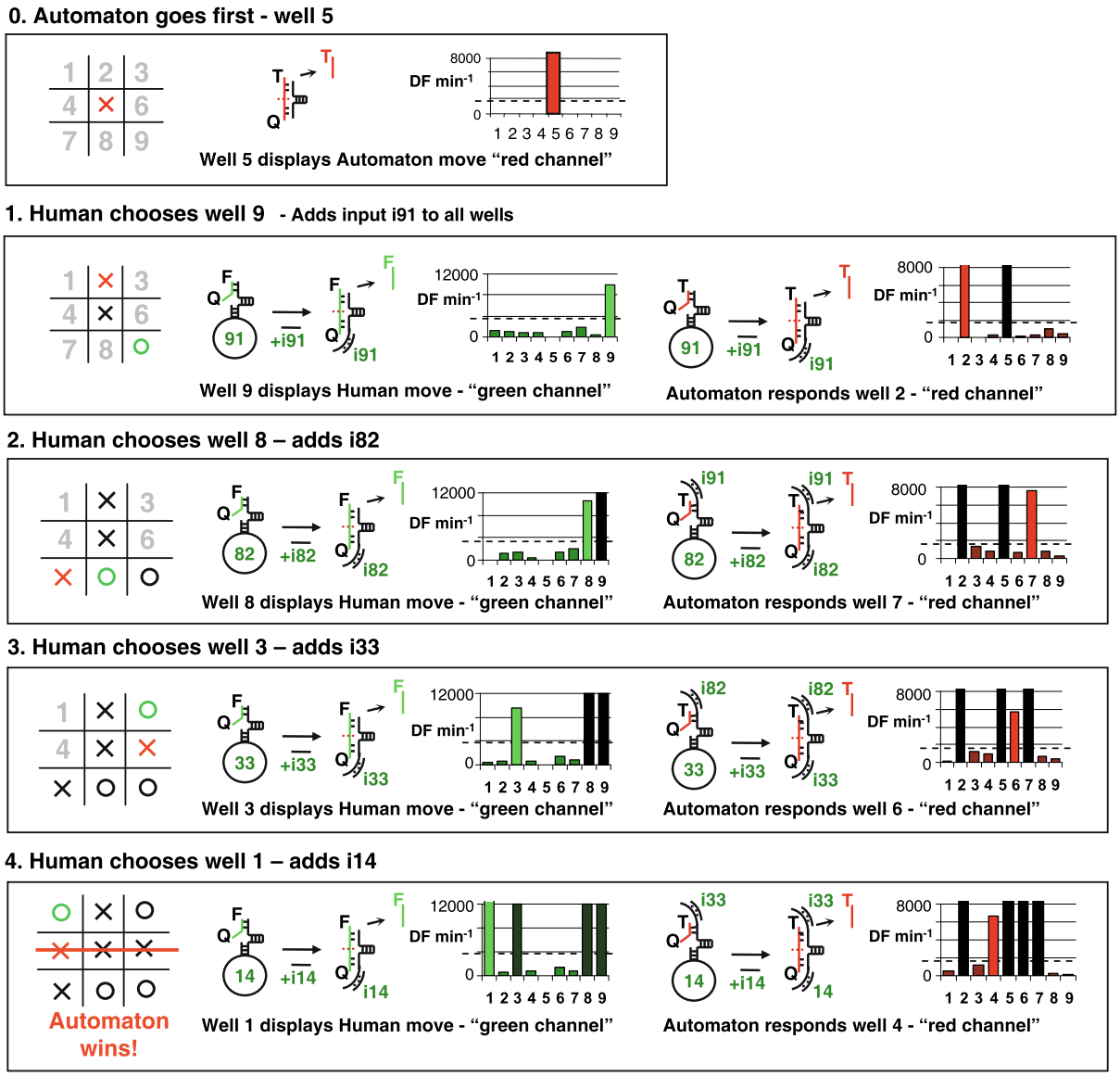

Molecular Automata, Figure 23

MAYA-II example game. A game played against MAYA-II ends in an automaton win, since the human opponent made an error in move 3. The human adds input to indicate their moves to every well, which causes a chain reaction to activate precisely one well for each fluorescent color. Human moves are displayed in the fluorescein (*green*) fluorescence channel, and automaton moves are displayed in the tetramethylrhodamine (*red*) fluorescence channel

structed as a hierarchy of AND gates, with YES gates responding to the first human move. Some NOT loops are included to prevent secondary activation in already played wells, or are redundant and included to minimize cumulative nondigital behavior in side wells over several moves. In doing this, MAYA-II is a step toward programmable and generalizable MAYAs that are trainable to play any game strategy.

Coordination of such a large number of rationally designed molecular logic gates was an unprecedented molecular engineering challenge, taking several years of construction. While spurious binding of oligonucleotides was predicted to cause serious problems, this was not observed in the building of the system. Instead, the most challenging aspect was titrating individual gates to produce similar levels of fluorescent signals, since the individual oligonucleotide input sequence could affect the catalytic activity of the molecule.

MAYA-II perfectly plays a general tic-tac-toe game by successfully signaling both human and automaton moves. An example of play against the automaton is shown in Fig. 23. It could be argued that by integrating more than 100 molecular logic gates in a single system, MAYA-II represents the first "medium-scale integrated molecular circuit" in solution. This level of rationally-designed complexity has important implications for the future of diagnostic and therapeutic molecular automata. Moreover, the increased complexity of MAYA-II enabled refinement of the deoxyribozyme logic gate model, allowing the development of design principles for optimizing digital gate behavior and the generation of a library of 32 known input sequences for future "plug and play" construction of complex automata. This library of known sequences is already being employed in the construction of other complex DNA automata [26].

Future Directions

This review mostly focused on DNA-based automata, an area of research that has its roots in Adleman's early computing experiments [2]. The early attempts to find applications for this type of DNA computing mostly focused on some kind of competition with silicon, for example, through massively parallel computing. After more then a decade of intensive research, we can safely conclude that, without some amazing new discovery, such applications are very unlikely. Instead, it seems that the most likely applications will come from the simplest systems, in which groups of molecules will have some diagnostic or therapeutic role. Further, approaches such as Winfree's can lead to completely new thinking in materials sciences. But aside from practical considerations, experiments in which mixtures of molecules perform complex functions and execute programs are of great interest for basic science. Mixtures of molecules perform complex tasks in organisms, and some of the systems we described provide us with tools to understand how such complexity may have arisen at first.

Bibliography

Primary Literature

1. Adar R, Benenson Y, Linshiz G, Rosner A, Tishby N, Shapiro E (2004) Stochastic computing with biomolecular automata. Proc Natl Acad Sci USA (PNAS) 101(27):9960–9965
2. Adleman LM (1994) Molecular computation of solutions to combinatorial problems. Science 266(5187):1021–1024
3. Andrews B (2005) Games, strategies, and boolean formula manipulation. Master's thesis, University of New Mexico
4. Bailly C (2003) Automata: The Golden Age, 1848–1914. Robert Hale, London
5. Barish RD, Rothemund PWK, Winfree E (2005) Two computational primitives for algorithmic self-assembly: Copying and counting. Nano Lett 5(12):2586–2592
6. Benenson Y, Adar R, Paz-Elizur T, Livneh Z, Shapiro E (2003) DNA molecule provides a computing machine with both data and fuel. Proc Natl Acad Sci USA (PNAS) 100(5):2191–2196
7. Benenson Y, Gil B, Ben-Dor U, Adar R, Shapiro E (2004) An autonomous molecular computer for logical control of gene expression. Nature 429:423–429
8. Benenson Y, Paz-Elizur T, Adar R, Keinan E, Livneh Z, Shapiro E (2001) Programmable and autonomous computing machine made of biomolecules. Nature 414:430–434
9. Breaker RR, Joyce GF (1995) A DNA enzyme with Mg^{2+}-dependent RNA phosphoesterase activity. Chem Biol 2:655–660
10. Collier CP, Wong EW, Belohradský M, Raymo FM, Stoddart JF, Kuekes PJ, Williams RS, Heath JR (1999) Electronically configurable molecular-based logic gates. Science 285:391–394
11. Credi A, Balzani V, Langford SJ, Stoddart JF (1997) Logic operations at the molecular level. An XOR gate based on a molecular machine. J Am Chem Soc 119(11):2679–2681
12. de Silva AP, Dixon IM, Gunaratne HQN, Gunnlaugsson T, Maxwell PRS, Rice TE (1999) Integration of logic functions and sequential operation of gates at the molecular-scale. J Am Chem Soc 121(6):1393–1394
13. de Silva AP, Gunaratne HQN, McCoy CP (1993) A molecular photoionic AND gate based on fluorescent signalling. Nature 364:42–44
14. de Silva AP, Gunaratne HQN, McCoy CP (1997) Molecular photoionic AND logic gates with bright fluorescence and "off-on" digital action. J Am Chem Soc 119(33):7891–7892
15. de Silva AP, McClenaghan ND (2000) Proof-of-principle of molecular-scale arithmetic. J Am Chem Soc 122(16):3965–3966
16. de Solla Price DJ (1964) Automata and the origins of mechanism and mechanistic philosophy. Technol Cult 5(1):9–23
17. Ellenbogen JC, Love JC (2000) Architectures for molecular electronic computers: 1. Logic structures and an adder built from molecular electronic diodes. Proc IEEE 88(3):386–426
18. Fu TJ, Seeman NC (1993) DNA double-crossover molecules. Biochemistry 32:3211–3220

19. Garzon M, Gao Y, Rose JA, Murphy RC, Deaton RJ, Franceschetti DR, Stevens SE Jr (1998) In vitro implementation of finite-state machines. In: Proceedings 2nd International Workshop on Implementing Automata WIA'97. Lecture Notes in Computer Science, vol 1436. Springer, London, pp 56–74
20. Gordon JM, Goldman AM, Maps J, Costello D, Tiberio R, Whitehead B (1986) Superconductin-normal phase boundary of a fractal network in a magnetic field. Phys Rev Lett 56(21):2280–2283
21. Holter NS, Lakhtakia A, Varadan VK, Varadan VV, Messier R (1986) On a new class of planar fractals: the Pascal–Sierpinski gaskets. J Phys A: Math Gen 19:1753–1759
22. Huang Y, Duan X, Cui Y, Lauhon LJ, Kim KH, Lieber CM (2001) Logic gates and computation from assembled nanowire building blocks. Science 294(9):1313–1317
23. LaBean TH, Yan H, Kopatsch J, Liu F, Winfree E, Reif JH, Seeman NC (2000) Construction, analysis, ligation, and self-assembly of DNA triple crossover complexes. J Am Chem Soc 122:1848–1860
24. Lederman H, Macdonald J, Stefanovic D, Stojanovic MN (2006) Deoxyribozyme-based three-input logic gates and construction of a molecular full adder. Biochemistry 45(4):1194–1199
25. Lipton RJ, Baum EB (eds) (1996) DNA Based Computers, DIMACS Workshop 1995 (Princeton University: Princeton, NJ). Series in Discrete Mathematics and Theoretical Computer Science, vol 27. American Mathematical Society, Princeton
26. Macdonald J (2007) DNA-based calculators with 7-segment displays. In: The 13th International Meeting on DNA Computing, Memphis
27. Macdonald J, Li Y, Sutovic M, Lederman H, Pendri K, Lu W, Andrews BL, Stefanovic D, Stojanovic MN (2006) Medium scale integration of molecular logic gates in an automaton. Nano Lett 6(11):2598–2603
28. Mao C (2004) The emergence of complexity: Lessons from DNA. PLoS Biology 2(12):2036–2038
29. Mao C, LaBean TH, Reif JH, Seeman NC (2000) Logical computation using algorithmic self-assembly of DNA triple-crossover molecules. Nature 407:493–496, erratum, Nature 408 (2000), p 750
30. Mealy GH (1955) A method for synthesizing sequential circuits. Bell Syst Techn J 34:1045–1079
31. Peppé R (2002) Automata and Mechanical Toys. Crowood Press, Ramsbury
32. Pickover CA (1990) On the aesthetics of sierpinski gaskets formed from large pascal's triangles. Leonardo 23(4):411–417
33. Riskin J (2003) The defecating duck, or, the ambiguous origins of artificial life. Crit Inq 29(4):599–633
34. Rothemund PWK (1996) A DNA and restriction enzyme implementation of Turing machines. In: Lipton RJ, Baum EB (eds) DNA Based Computers. American Mathematical Society, Providence, pp 75–120
35. Rothemund PWK, Papadakis N, Winfree E (2004) Algorithmic self-assembly of DNA Sierpinski triangles. PLoS Biol 2(12):2041–2053
36. Rothemund PWK, Winfree E (2000) The program-size complexity of self-assembled squares. In: STOC'00: The 32nd Annual ACM Symposium on Theory of Computing. Association for Computing Machinery, Portland, pp 459–468
37. Santoro SW, Joyce GF (1997) A general purpose RNA-cleaving DNA enzyme. Proc Natl Acad Sci USA (PNAS) 94:4262–4266
38. Shapiro E, Karunaratne KSG (2001) Method and system of computing similar to a Turing machine. US Patent 6,266,569 B1
39. Soloveichik D, Winfree E (2005) The computational power of Benenson automata. Theoret Comput Sci 344:279–297
40. Soreni M, Yogev S, Kossoy E, Shoham Y, Keinan E (2005) Parallel biomolecular computation on surfaces with advanced finite automata. J Am Chem Soc 127(11):3935–3943
41. Stojanovic MN, de Prada P, Landry DW (2001) Catalytic molecular beacons. Chem Bio Chem 2(6):411–415
42. Stojanovic MN, Kolpashchikov D (2004) Modular aptameric sensors. J Am Chem Soc 126(30):9266–9270
43. Stojanovic MN, Mitchell TE, Stefanovic D (2002) Deoxyribozyme-based logic gates. J Am Chem Soc 124(14):3555–3561
44. Stojanovic MN, Semova S, Kolpashchikov D, Morgan C, Stefanovic D (2005) Deoxyribozyme-based ligase logic gates and their initial circuits. J Am Chem Soc 127(19):6914–6915
45. Stojanovic MN, Stefanovic D (2003) Deoxyribozyme-based half adder. J Am Chem Soc 125(22):6673–6676
46. Stojanovic MN, Stefanovic D (2003) A deoxyribozyme-based molecular automaton. Nature Biotechnol 21(9):1069–1074
47. Wang H (1963) Dominoes and the AEA case of the decision problem. In: Fox J (ed) Mathematical Theory of Automata. Polytechnic Press, New York, pp 23–55
48. Winfree E (1996) On the computational power of DNA annealing and ligation. In: Lipton RJ, Baum EB (eds) (1996) DNA Based Computers. American Mathematical Society, Providence, pp 199–221
49. Winfree E (2006) Self-healing tile sets. In: Chen J, Jonoska N, Rozenberg G (eds) (2006) Natural Computing. Springer, Berlin, pp 55–78
50. Winfree E, Yang X, Seeman NC (1999) Universal computation via self-assembly of DNA: Some theory and experiments. In: Landweber LF, Baum EB (eds) DNA Based Computers II, DIMACS Workshop 1996 (Princeton University: Princeton, NJ), American Mathematical Society, Princeton. Series in Discrete Mathematics and Theoretical Computer Science, vol 44. pp 191–213; Errata: http://www.dna.caltech.edu/Papers/self-assem.errata
51. Yurke B, Mills Jr AP, Cheng SL (1999) DNA implementation of addition in which the input strands are separate from the operator strands. Bio Systems 52(1–3):165–174

Books and Reviews

Chen J, Jonoska N, Rozenberg G (eds) (2006) Natural Computing. Springer, Berlin

Lewis HR, Papadimitriou CH (1981) Elements of the Theory of Computation. Prentice-Hall, Englewood Cliffs

Seeman NC (2002) It started with Watson and Crick, but it sure didn't end there: Pitfalls and possibilities beyond the classic double helix. Nat Comput Int J 1(1):53–84

Molecular Evolution, Networks in

Andreas Wagner[1,2]

[1] Department of Biochemistry, University of Zurich, Zurich, Switzerland

[2] The Santa Fe Institute, New Mexico, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Protein Interaction Networks
Transcriptional Regulation Networks
Metabolic Networks
Summary and Outlook
Bibliography

Glossary

Genome The totality of genetic material in an organism, including all its genes.

Metabolite A small molecule that is either produced or consumed by a chemical reaction that takes place inside an organism.

Graph A mathematical object that consists of nodes or vertices. Pairs of nodes may be connected by edges.

Gene frequency or allele frequency The proportion of genes in a population that have a specific genotype (DNA sequence).

Paralogous genes Genes in the same genome that originated in a gene duplication event.

Orthologous genes Genes in the genomes of two different species that shared a common ancestor in the ancestral species.

Nonsynonymous nucleotide substitution/amino acid replacement substitution A nucleotide substitution in a gene that changes the amino acid sequence of the encoded protein.

Synonymous nucleotide substitution/amino acid replacement substitution A nucleotide substitution in a gene that does not change the amino acid sequence of the encoded protein.

Protein domain a protein region with a characteristic function and structure that often also folds autonomously.

Definition of the Subject

Molecular evolution is concerned with evolutionary change of nucleic acids and proteins. It attempts to identify the evolutionary forces that cause these molecules to change their structure over millions of years. Molecular evolution as a research field emerged in the second half of the 20th century, when information on DNA and protein sequences first became available. Although studies in the field initially focused on the evolution of genes and the proteins they encode, they increasingly concentrate on the evolution of whole genomes. This was made possible by the availability of whole genome sequences in the mid-1990s.

Recently, technological developments have made it possible to study molecular networks inside cells. These networks encompass hundreds or thousands of proteins that interact with each other, with DNA or RNA, and with small metabolites. A molecule's position in such a molecular network, as well as its interaction partners may influence the tempo and mode of the molecule's evolution. In addition, change in individual molecules of a network can influence the network's structure on an evolutionary time scale. These two topics form the core of this contribution.

Introduction

Molecular Networks

A molecular network is a highly heterogeneous assemblage of different molecules, including small metabolites, RNA, DNA, proteins, and protein complexes. Molecules in this assemblage interact with each other in a variety of ways to carry out important cellular functions. In any one cell, the structure of this network changes as a function of the cell's physiological state, and as a function of the proteins and RNA molecules that are expressed at any one time. No experimental technique is currently available that could reveal the full complexity of a molecular network, much less its temporal dynamics. However, much information is available on (sub)networks that are characterized by one kind of molecular interactions. Specifically, three kinds of such networks have been characterized extensively in different organisms. The first kind is a *protein interaction network*. It can be represented as a graph whose nodes are proteins, and where two proteins are connected by an edge if they physically interact inside a cell. The second kind of network is a *transcriptional regulation network*. Here, the nodes of the network are genes. A directed edge connects a gene A to a gene B in such a network, if A encodes a transcriptional regulator, a protein that binds to regulatory DNA near gene B, and if A activates or represses the transcription of B. The third and final class of well-characterized networks comprises *metabolic reaction networks*. They are networks of chemical reactions that sustain life by producing energy and biochemical building blocks for cell growth. Metabolic networks consist of two kinds of key parts, metabolites and metabolic enzymes. Metabolic enzymes catalyze chemical reactions that convert metabolites into other metabolites. These enzymes are encoded by genes.

Because these three kinds of networks, protein interaction networks, transcriptional regulation networks, and metabolic networks, are by far the best studied kinds of biological networks, this contribution will focus on them.

Molecular Evolution

Genes and the proteins they encode are key components of the three networks introduced above. Genomic DNA in general, and genes in particular can undergo three principal kinds of evolutionary genetic change (mutations). The first kind is a *deletion*, whereby a gene or a part of it becomes eliminated from the genome. The second kind is a *duplication*, whereby a stretch of genomic DNA becomes duplicated, such that two copies of the DNA sequence come to exist in the genome. The duplicate DNA can reside immediately adjacent or far away from the original, depending on the mechanism of duplication. If such a duplication encompasses one or more genes, one speaks of a gene duplication. Gene duplications have received considerable attention since whole genome sequences have become available, because they lead to an increase of the number of genes in a genome, and because they may facilitate the evolution of genes with new functions. The third kind of change is a *point mutation*. Here, a single nucleotide changes. If the change occurs inside of a gene, then the amino acid sequence of the encoded protein may also change, leading to a potential change of function in the protein. More complicated kinds of evolutionary change also occur, such as rearrangements of parts or domains within a protein. Their impact on network evolution is less intensely studied and has been reviewed elsewhere [1].

This characterization of evolutionary genetic change distinguishes different kinds of molecular events. In addition, one can also distinguish genetic change through its effects on fitness. Here again, there are three possible classes of change. The first class consists of *neutral mutations*. Such mutations are causing a change in genomic DNA that leaves an organism's fitness unchanged. A second class comprises *beneficial mutations*, mutations that increase an organism's fitness. Natural selection increases the frequency of genes carrying such mutations in a population. A third class consists of *deleterious mutations* which decrease the fitness of an organism, and are thus often eliminated from populations. For this reason, deleterious mutations contribute little to observed molecular variation, even though they may be the most frequent mutations. Despite 40 years of research, it is still a matter of debate whether most mutations that give rise to observed variation in a population of organisms are neutral or beneficial.

Molecular evolution as a research field emerged in the second half of the 20th century, with the availability of the first DNA and protein sequences. A key theoretical development in the field was Kimura's neutral theory of molecular evolution [2]. This theory makes specific predictions about the fate of neutral mutations. Specifically, the rate at which neutral mutations arise that will eventually go to fixation, that is, attain a frequency of one, equals the rate of neutral mutations itself, and is constant and independent of population size. The time neutral mutations take to go to fixation is proportional to the size of a population. These simple predictions do not hold for beneficial mutations whose fate also depends on the amount of fitness benefits they confer. These predictions of the neutral theory are well corroborated, but Kimura and others made additional claims that were more controversial. Specifically, they maintained that neutral mutations comprised the vast majority of mutations that give rise to genetic variation in a population, a claim that gave rise to the neutralist-selectionist controversy [3,4]. Although this debate has not been fully resolved, recent analyzes based on whole-genome data suggest that many mutations that occur in a genome have beneficial effects [5,6].

Multiple sequence characteristics can be used to determine whether the DNA sequence of a gene has been subject to mostly negative selection that eliminates deleterious mutations, to positive selection that has increased beneficial mutations in frequency, or to no selection (neutral evolution) [7]. One such characteristic, the ratio K_a/K_s of non-synonymous to synonymous nucleotide substitution is simple and widely used. In order to determine this ratio, one compares two genes and the mutations that have accumulated since their common ancestry (either since a gene duplication event for paralogous genes, or since a speciation event for orthologous genes). Specifically, one estimates the number of non-synonymous mutations, mutations that did change the amino acid sequence of the encoded protein, and the number of synonymous mutations, mutations that did not change the protein. Such mutations are possible, because the genetic code is redundant, that is, multiple codons may encode the same amino acid. More specifically still, one estimates K_s, the fraction of synonymous substitutions per synonymous nucleotide site in a gene, and K_a, the fraction of amino acid replacement substitutions per replacement site. These measures of divergence account for the fact that different genes have different length. From these estimates, one then calculates K_a/K_s. If this ratio is smaller than one, then the genes in question have tolerated fewer amino acid replacement

substitutions in their evolutionary history than synonymous substitutions. This means that the genes are under negative or purifying selection, because some amino acid substitutions have been eliminated from the evolutionary record. If the ratio is equal to one ($K_a = K_s$), then an equal number of silent and replacement substitutions have been preserved. Such genes evolve neutrally. This pattern of evolution is typical of pseudogenes, genes that have lost their function through mutations. Finally, if the ratio is greater than one, then more amino acid changing mutations have been preserved than synonymous mutations, meaning that the genes have been subject to net positive selection. For the vast majority of genes, the ratio K_a/K_s is much smaller than one, meaning that these genes are under net negative or purifying selection. For these genes, K_a/K_s is a good indicator of the evolutionary constraint a gene is subject to: Genes with small K_a/K_s are said to be more highly constrained than genes with a large K_a/K_s.

Molecular Networks and Molecular Evolution

Two principal kinds of genetic change can be distinguished in the molecular evolution of molecular networks. First, there is change that affects the number of network parts itself, either by adding network parts through duplication, or by eliminating network parts through deletion. Second, there is change that leaves the network size unaffected, but that changes existing network parts and their interactions through point mutations. A comprehensive analysis of network evolution would study both categories of change, and it would analyze how such change affects the structure of a network. Such an analysis would also study how natural selection on network function would influence the kinds of genetic change that can be tolerated on evolutionary time scales. Partly because of a lack of necessary data, no such comprehensive analysis exists for all of the molecular networks discussed here. One kind of change and its impact on a network may have been studied for one kind of network, but hardly at all for another network. The next sections highlight insights available from studies focusing on one or the other kind of change and its effects on the evolution of protein interaction networks, transcriptional regulation networks, and metabolic networks.

Protein Interaction Networks

Characterizing Protein Interaction Networks

Two prominent experimental approaches exist to characterize protein interaction networks (reviewed in [8]). These approaches illustrate the kinds of data available for evolutionary studies on such networks. The first approach is the yeast two-hybrid assay [9], a technique to identify interactions between two specific proteins A and B (not necessarily from yeast). This assay first uses recombinant DNA techniques to generate two hybrid proteins. In one of these hybrids, protein A is fused to the transcriptional activation domain of a yeast transcription factor. In the other hybrid, the transcriptional activation domain of the same transcription factor is fused to protein B. If protein A and B interact in vivo, then their interaction physically links the transcriptional activation and the DNA binding domain of the transcription factor, thus allowing transcriptional activation of a suitably chosen "reporter" gene, which can be easily detected. The two hybrid approach has been applied to detect interactions of most protein pairs A-B in a genome [10,11,12,13,14,15,16,17,18]. Even for a small genome like that of the yeast *Saccharomyces cerevisiae*, this requires screening millions of pairwise interactions.

The first genome-wide protein interaction screens that used the two-hybrid assay were carried out in the yeast proteome itself. They yielded maps of protein interactions involving some 1000 proteins [14,15]. Variations of the approach have been applied successfully to analyze protein interactions in other microbes, such as the bacterium *Helicobacter pylori* [16], and protein interactions between viral and cellular proteins [11,12]. The yeast two-hybrid approach has several commonly recognized shortcomings. One of them is the use of fusion proteins, which can lead to protein misfolding. Another problem is that the assay forces coexpression of proteins in the same compartment of a cell or an organism, although the proteins may not co-localize in vivo. These shortcomings lead to potentially high false positive and false negative error rates, i. e., to the detection of spurious interactions, and to the failure to detect actual interactions. These error rates may well exceed 50% [19,20]. This complication means that it is currently difficult to evaluate which of the (vast) differences in network composition and interactions observed among distantly related organisms is due to evolutionary divergence, and which part is due to experimental error.

Another class of techniques to characterize protein interaction networks identifies the proteins that are part of a multiprotein complex [21,22,23]. Here, the departure point of a typical experiment is some protein A of interest, and the experiment asks which protein complexes – groups of interacting proteins – this protein A is a part of. In the experiment, protein A is reversibly attached to a solid support via a chemical tag. This solid support is exposed to a protein extract from cells. As a result, proteins that can interact with A become attached to the support

via protein A. Protein A and all proteins attached to it are then released from the support, at which point the proteins can be isolated and characterized, for example through mass spectrometry. The whole approach is a variation of affinity chromatography, a chemical separation technique that takes advantage of specific binding of one molecule to another. The largest-scale approaches so far have identified more than 400 protein complexes in the yeast *Saccharomyces cerevisiae* [21,22,23,24].

The yeast two-hybrid assay and affinity chromatography based methods lead to different and complementary kinds of information. The yeast two-hybrid assay yields information about pairwise protein interactions. In contrast, affinity chromatography-based methods lead to information about the proteins that occur in a protein complex, where not all of the proteins in a complex may interact directly with each other.

Characterizing Network Structure

Perhaps the most basic and general question that one can ask about protein interaction networks (or any other molecular network) is why a network has its observed structure. To answer this question ultimately requires an evolutionary perspective, because any network's structure needs to be explained from its evolutionary history and the evolutionary forces shaping it. To answer this question, however, one has to first know what a network's structure *is*. Because molecular networks have thousands of parts, visual inspection is of little use in identifying a network's structure, and it is not always clear what features of the structure to focus on. Most existing work focuses on the simplest structural network characteristics, three of which are given below. Others are also in use, but many biologically sensible such characteristics may still await discovery.

Perhaps the simplest structural characteristic one can study is the distribution of the number d of interactions per protein, the so-called degree distribution of a network. A second characteristic are degree correlations among proteins, that is, one can ask whether highly (lowly) connected proteins preferentially connect to highly (lowly) connected other proteins. A third basic characteristic of a molecular network is the clustering coefficient C [25]. To define the clustering coefficient $C(v)$ of a node (protein) v in a graph, consider all k_v nodes adjacent to a node v, and count the number m of edges that exist among these k_v nodes (not including edges connecting them to v). The maximally possible m is $k_v(k_v - 1)/2$, in which case all k_v nodes are connected to each other. Let $C(v) := m/(k_v(k_v - 1)/2)$. $C(v)$ measures the "cliquishness" of the neighborhood of v, i. e., what fraction of the nodes adjacent to v are also adjacent to each other. The clustering coefficient C of the whole network is defined as the average of $C(v)$ over all v.

The degree distribution of protein interaction networks resembles a power law, $P(d) \sim d^{-\gamma}$, where γ is some constant [26,27], protein degrees are anticorrelated, that is, highly connected proteins preferentially interact with lowly connected proteins [28], and the clustering coefficient of protein interaction networks is much higher than that of random networks with the same number of interactions.

Protein Network Structure and Molecular Evolution

A variety of evolutionary models have attempted to ask why networks have their observed structure with respect to the above and some other simple structural features [29,30,31,32,33,34,35,36,37]. These models rely on two main ingredients, addition and deletion of network proteins (caused by gene duplications and deletions) which can change the size of a network, and "rewiring" of network interactions driven by point mutations in the genes encoding network proteins. Both processes undoubtedly play a role in network evolution. Network rewiring must occur, because individual mutations can change protein-protein interfaces necessary for interactions. Gene duplication and gene deletion must also play a role, because genomes vary in size by orders of magnitude, and so do the number of genes, encoded proteins, and protein interaction network size. In addition, some families of interacting proteins such as heterodimerizing transcription factors have arisen largely through gene duplication [38,39]. Furthermore, gene duplication plays a role in the evolution of new protein complexes in yeast [40].

Beyond these generalities, the available models differ widely in their assumptions, and about the importance they ascribe to rewiring and duplication/deletion. They include differences in assumptions about

(i) Rates of duplication, deletion, and rewiring,
(ii) Whether these processes are random with respect to network structure, or whether their rate depends on a protein's position in the network, and
(iii) Whether duplication/deletion and rewiring occur independently from one another or whether they are in some way coupled.

Most existing models constitute mathematical proofs of principle, that is, they attempt to show that a particular network feature, such as the degree distribution *could* be explained by a particular evolutionary process, whereas

a few models attempt to stay close to available molecular evolution data. However, this data is currently very limited, because no information is available about the structure of protein interaction networks in closely related organisms. That is, the available data is either derived from comparisons of protein content and/or network structure of very distantly related organisms, or from within one genome, such as from gene duplicates (whose age can estimated) and their common interaction partners [26,41]. Although such data is insufficient to validate or refute any one of the models to the exclusion of all others, a limited amount of evidence favors a preferential attachment mechanism of network evolution. In this mechanism, proteins that have arisen early during network evolution tend to be highly connected proteins, and such highly connected proteins may acquire more interactions subsequently [36,42,43,44], but see also Kunin [45].

Despite all their differences, existing network evolution models have an important unifying feature: None of them require that natural selection molds any global feature of network structure, such as the degree distribution. This observation is significant, because early work on molecular networks assumed that features of protein interaction networks, such as the power-law degree distribution reflect evolutionary optimization of some aspect of network function. For example, in protein interaction networks and other networks with power-law degree distributions, the mean distance between network nodes that can be reached from each other (via a path of edges) is very small and it increases only very little upon random removal of nodes [46]. This distance can be thought of as a measure of how compact a network is. In graphs with other degree distributions, this mean distance can increase substantially upon node removal. From this observation emerged the proposition that robustly compact networks confer some (unknown) advantages on a cell, and that the power law degree distribution reflects the action of natural selection on the degree distribution itself. The observation alone that power-law degree distributions are ubiquitous in biological and non-biological systems argues against this proposition. The models mentioned above, none of which require natural selection on the degree distribution, further speak against it. In addition, an even simpler hypothetical explanation of observed network structure has been proposed. This hypothesis explains the degree distribution and other network features by a random model of desolvation energies among interacting protein pairs [47].

One might be tempted to call network evolution in the absence of natural selection optimizing a global network feature *neutral evolution*. Doing so, however, would neglect that natural selection almost certainly influences which duplication/deletion/rewiring events are preserved in the evolutionary record. In other words, even though natural selection may not influence global network structure, it may affect the local events that change network structure in evolutionary time. Multiple lines of evidence hint at this influence of natural selection. The first comes from a study on protein complexes. In the yeast *Saccharomyces cerevisiae*, over- or underexpression of members of a protein complex may have adverse effects on fitness. The likely reason is that such expression changes affect the stoichiometric balance of the proteins in a cell which is necessary for forming complexes with the correct protein composition [48,49]. Gene duplications of proteins interacting in a complex may be harmful, because such duplications effectively change gene expression, which distorts this balance. In agreement with this observation, proteins encoded by members of large gene families, genes that have often undergone duplication, are underrepresented in protein complexes [49]. A second, similar indication of the influence of natural selection on network evolution is that the number of proteins in a complex encoded by single copy genes rises with complex size [50]. Thirdly, gene duplications seem to have been preferentially preserved in the sparsely connected regions of the yeast protein interaction network, regions that are characterized by low degree [51] or low clustering coefficients [52]. This suggests that gene duplications in densely connected network parts may have deleterious effects.

Rather than focusing on global networks structure, a limited amount of work has focused on small subgraphs of a protein interaction network. Such subgraphs comprise only few (3–5) proteins, are characterized by specific patterns of interactions, and are also known as network motifs. Proteins that occur in larger and more densely connected motifs have a greater likelihood to be preserved across distantly related species [53,54].

All work discussed thus far has focused on the evolution of the network itself. Another line of inquiry asks how a protein's position within a network constrains the protein's evolution. For example, as already discussed above, gene duplications tend to be observed preferentially for genes in sparsely connected parts of a network [51,55]. Also, proteins that have a more central role in the protein interaction network evolve more slowly [56]. In addition, early work suggested that proteins with more interaction partners are evolutionarily more constrained [57,58]. This association has become controversial, because it may be caused by bias in protein interaction data sets, and because it may be explained by differences in gene expression level among proteins with different numbers of inter-

action partners [59,60,61,62,63,64,65]. Specifically, highly expressed proteins evolve more slowly, and much of the observed variation in evolutionary rates among proteins may be due to variation in expression level [66,67], leaving only a minor role for the influence of protein-protein interactions.

Rather than just considering the numbers of interactions of a protein in a protein network when trying to explain evolutionary rate differences, it may be necessary to distinguish between different kinds of interactions. One important distinction here is that between transient and permanent interactions. Proteins that enter permanent interactions, thus forming stable complexes with other proteins, evolve at lower rates than proteins that undergo transient interactions, or proteins that are not known to interact with other proteins [55,68]. A closely related distinction is that between proteins that have multiple protein interaction interfaces, and that can thus interact with multiple proteins at the same time, and between proteins that have a single interaction interface, and that interact with multiple partners successively and transiently. Multi-interface proteins evolve more slowly, which may be readily explained by the larger fraction of their surface that is constrained [69]. Yet other distinctions among interactions may also affect evolutionary rates. For example, interactions between proteins of different cellular functions may constrain evolutionary rates particularly strongly [70].

In sum, models of protein on network evolution agree that natural selection on global network structure is not necessary to produce protein interaction networks with the global features that have been studied so far. Nonetheless, these models differ in the relative importance they ascribe to deletion and duplication on one hand, and interaction rewiring on the other hand, in network evolution. Studies focusing on the evolution of network parts within an existing network suggest that a protein's position in a network influences these constraints. Nonetheless, this influence may be minor compared to other factors, especially protein expression level.

Transcriptional Regulation Networks

Characterizing Transcriptional Regulation Networks

In a transcriptional regulation network, transcription factors bind to regulatory DNA near network genes, and activate or repress the expression of these genes. Transcription factors are proteins that are themselves encoded by genes in the network. Transcriptional regulation networks thus comprise two main kinds of genes, genes encoding transcriptional regulators, and their regulatory target genes. The two classes of genes overlap, because genes encoding transcriptional regulators may themselves be transcriptionally regulated. Even small genomes such as that of the yeast *Saccharomyces cerevisiae* contain hundreds of genes encoding transcriptional regulators.

Two principal approaches have been pursued to characterize transcriptional regulation networks. One of them is manual curation, whereby data from existing experimental literature about the targets of individual transcription factors, is assembled into a network [71,72]. The second approach is high-throughput experimental analysis of DNA binding by transcriptional regulators. This approach permits the genome-scale identification of regulatory DNA regions bound by transcription factors. It thus provides hints which genes may be regulated by which transcription factors, although transcription factor binding is only a necessary, but not a sufficient criterion for transcriptional regulation. A prominent technique used in this area is chromatin immunoprecipitation. In this technique, a transcriptional regulator is labeled with an epitope tag, a molecule that can be recognized by a specific antibody. Genomic DNA, some of which is bound by the regulator, is then isolated. This isolate is then exposed to the antibody, in order to precipitate the DNA bound by the regulator, hence the name immunoprecipitation. The precipitated DNA is then hybridized to a DNA microarray, allowing its identification and localization in the genome. In one prominent study using this technique putative candidate target genes of 106 yeast transcriptional regulators were identified [73].

These two approaches to characterize transcriptional regulation networks are complementary: Manual curation may reveal high quality information about individual transcriptional regulators, but it may capture only a limited number of regulatory interactions. The high-throughput approach, on the other hand, provides more comprehensive information at the price of greater uncertainty about the biological relevance of the observed interactions.

It is noteworthy that transcriptional regulation networks, as opposed to protein interaction networks, are directed networks. This means that interactions occur from a regulator to its target gene, but not necessarily vice versa. In a graph representation of such a network, genes are thus connected by directed edges.

Transcriptional Regulation Networks and Molecular Evolution

The molecular evolution of transcriptional regulation networks has received less attention than that of protein interaction networks. In existing work, some parallels to evolutionary patterns in protein interaction networks are ev-

ident. First, a gene's connectedness within the network may have only a weak or no impact on its rate of evolutionary change. Specifically, the number of target genes of a transcriptional regulator does not affect the regulator's evolutionary rate, as indicated by the ratio K_a/K_s. Similarly, the number of transcriptional regulators that regulate a given target gene does not strongly influence the evolutionary rate of the target gene. Second, as in protein interaction networks, gene duplications are also very important in the formation of transcriptional regulation networks [74,75,76,77]. For example, a large proportion of transcriptional regulators themselves are products of duplicate genes. The exact proportion depends on how duplicates are identified. For example, approaches that identify duplicate genes through their domain architecture may reveal that a majority of transcriptional regulators are the results of gene duplication, whereas approaches based on significant sequence similarity among transcriptional regulator genes may ascribe a lesser role to duplication. Thirdly, rewiring of transcriptional regulation interactions has also played a prominent role in transcriptional regulation networks [78,79,80,81]. Such rewiring can be accomplished in two ways. First, a mutation may change the DNA binding domain of a transcription factor, such that the factor recognizes a different spectrum of regulatory DNA motifs. However, because any one transcriptional regulator may regulate hundreds of genes, many such changes are likely to be deleterious and may not be preserved in the evolutionary record. Second and perhaps more importantly, changes in a gene's regulatory region may affect which transcription factors can bind to and regulate a gene of interest. Because the regulatory DNA sequence motifs at which a transcription factor binds are often very short, binding sites can be easily created or destroyed through mutations. For example, in a study comparing gene expression patterns between the yeasts *Candida albicans* and *Saccharomyces cerevisiae*, a strong expression correlation was found between cytoplasmatic and ribosomal proteins in *C. albicans* but not in *S. cerevisiae*. This difference was associated with a change in multiple short regulatory DNA elements that drive the expression of these genes [80].

How fast and to what extent gene functions diverge after gene duplication is a subject of considerable interest to molecular evolutionists. Gene regulation and gene expression are an important aspect of gene function. Duplicate genes in a transcriptional regulation network are thus ideal study subjects to help answer this question. For example, one can ask to what extent duplicate genes of different sequence similarity (and thus different age) share transcription factors that bind at their regulatory regions. The answer is that duplicate genes rapidly diverge in the number of shared transcription factors they share [77,82]. Specifically, duplicate genes in yeast may lose 3% of common transcription factors for every 1% of sequence divergence [82]. The process of divergence, however, does not only involve loss of common transcription factor binding sites. A gain of new sites unique to each member of a duplicate gene pair may be equally important [77,83].

One area that has received perhaps more attention than in protein interaction networks is the analysis of small and highly abundant genetic circuit motifs in transcriptional regulation networks [22,84,85]. An example for such a regulatory motif is a transcriptional feed-forward loop, where a transcriptional regulator A regulates the expression of a regulator B, which regulates the expression of some target gene C, which is also regulated by A. Multiple other classes of network motifs are known. A wide spectrum of possibilities exist for the evolutionary origin of these circuits. At the two extremes of this spectrum stand two scenarios. First, these circuits may have arisen through the duplication – and subsequent functional diversification – of one or a few ancestral circuits, that is, through the duplication of each of their constituent genes in a series of duplication events. Alternatively, most of these circuits may have arisen independently by recruitment of unrelated genes. In this case, abundant circuits would have arisen through *convergent evolution*. Convergent evolution – the independent origin of similar organismal features – is a strong indicator of optimal "design" of a feature.

Because the complete genome sequence is available for the yeast *Saccharomyces cerevisiae*, one can ask which of these scenarios better reflects the evolutionary history of transcriptional regulation motifs. The answer is that the vast majority of highly abundant transcriptional regulation motifs have not originated through gene duplication, but independently and convergently [79]. What are the favorable functional properties of such networks, the properties that would drive such convergent evolution? Answers are beginning to emerge from a mix of computational and experimental work [85,86,87]. For example, a feed-forward loop may activate the regulated ('downstream') genes only if the upstream-most regulator is persistently activated. It can thus filter intracellular gene expression noise, which is known to be ubiquitous.

Metabolic Networks

The Analysis of Genome-Scale Metabolic Networks

Complex chemical reaction networks comprising hundreds to thousands of reactions sustain all of life. In

free-living, heterotrophic organisms, these reaction networks transform food into energy and biosynthetic building blocks for growth and reproduction. Complete (or nearly so) maps of core metabolism, comprising hundreds of reactions and metabolites are available for several model organisms [88,89,90]. These maps have been assembled through painstaking analysis of decades worth of biochemical literature, aided by genome sequence analysis, which may help determine whether a genome contains a gene catalyzing a particular chemical reaction.

Some work in this area focuses on network *structure*, by characterizing one of a variety of graph representations of a metabolic network. For example, metabolic networks can be represented as graphs whose nodes are enzymes and metabolites. Two nodes are connected if they participate in the same chemical reactions. Structural network analysis, however, has one key limitation: It does not capture the flow of matter through a metabolic network, which is at the heart of metabolic network function. This function can be computationally analyzed, even though information about enzymatic reaction rates in metabolic networks is very limited. Central to any such computational analysis are approaches such as flux balance analysis that use only information about the stoichiometry and reversibility of chemical reactions [91,92]. Flux balance analysis determines the rates (fluxes) at which individual chemical reactions can proceed if fundamental constraints such as that of mass conservation have to be fulfilled. Within the limits of such constraints, flux balance analysis can determine the distribution of metabolic fluxes that will maximize some metabolic property of interest. The rate of biomass production is one of these properties. It is a proxy for cell growth-rate, itself an important component of fitness in single-cell organism. Flux balance analysis makes predictions that are often in good agreement with experimental evidence in *E. coli* and the yeast *S. cerevisiae* [90,93,94]. However, such predictions may fail if an organism has not been subject to natural selection to optimize growth in a particular environment.

Metabolic Networks and Molecular Evolution

Much as for the other two classes of networks, considerable attention has focused on the question how the structure of metabolic networks evolved [95,96,97,98,99,100,101,102,103]. With some exceptions [93,104,105], most work has not focused on metabolites, but on enzymes and their role in network evolution. The reason is that the evolution of metabolic enzymes can be better reconstructed through gene sequence and protein structure comparisons.

In the evolution of metabolic network structure, rewiring clearly has much lower importance than in the previous networks discussed here. In contrast to protein interaction networks and regulatory networks, where interactions could form between a wide variety of different proteins, in metabolic networks interactions are largely dictated by chemistry. That is, only two enzymes that share substrates or products of their reactions can be neighbors in the network.

In contrast to rewiring, gene duplications and gene deletions play an important role in network evolution. Long before information on genome-scale metabolic reaction networks became available, gene duplication already played an important role in two major hypotheses about the evolution of metabolic pathways. The first such hypothesis is that of "retrograde" evolution. According to this hypothesis, metabolic pathways evolved backwards from their (essential) end-products, through the addition of new enzymes produced by gene duplication, and in response to the depletion of substrates that are necessary for production of these end-products [106]. The second, "patchwork evolution" hypothesis postulates that enzymes originally had broad substrate specificities, and that they subsequently evolved more specialized functions through gene duplication [107]. Recent genome-scale analyzes of metabolic network evolution suggest that both processes may occur, but that retrograde evolution by gene duplication is relatively rare [96,100,108]. For example, only a small fraction of adjacent enzymes in the same pathway arose from gene duplication. In contrast, recruitment of duplicate enzymes into new pathways is very frequent [100]. Gene duplications can also produce isoenzymes, enzymes with the same catalytic function, and thus the same position in a pathway, but possibly differential regulation. The locations in a network where such duplications are most likely preserved are not random. For example, isoenzymes are most often observed for enzymes with high metabolic flux, enzymes through which a lot of matter flows per unit time [109,110].

The phenomenon of horizontal gene transfer, which can add new genes from a different organism to a network has also received some attention in the analysis of metabolic network evolution [111,112]. Genes encoding metabolic enzymes are frequently transferred horizontally among bacterial genomes. However, such transfer does not affect all classes of enzymes equally. Specifically, peripheral network reactions, which are often reactions that are involved in an organism's response to specific environmental demands, are more frequently added to a network by horizontal transfer than central reactions [112]. This does not mean, however, that central parts of metabolism

are completely invariant. Even a metabolic cycle as central as the tricarboxylic acid cycle can undergo substantial evolutionary change [113].

Gene duplication and horizontal gene transfer are both mechanisms by which metabolic networks can increase in size. Gene deletions, in contrast, reduce network size. They are especially important in the evolution of organelles or organisms with reduced genome size, such as chloroplasts [114] and endosymbiotic cells [97]. In such cells, the host cell provides most metabolites necessary for survival, and metabolic networks are thus often drastically reduced in size. Deletions of enzyme-coding genes have also been studied in the evolution of individual pathways. Examples include the vitamin B6 synthesis pathway, which has been lost multiple times independently through gene deletions during animal evolution [115].

As in the other two networks discussed above, some work has also focused on the evolution of enzyme-coding genes *within* a network and within metabolic pathways. An example regards enzymes involved in the biosynthesis of anthocyan, a plant pigment. In this pathway, upstream enzymes are subject to greater evolutionary constraints than downstream enzymes, as indicated by their lower rate at which non-synonymous substitutions accumulate [116]. Here, the upstream enzymes are located above metabolic branch points that lead to other metabolic pathways. It is thus possible that mutations in them are more likely to be deleterious, because they affect more than one pathway. This evolutionary pattern of higher conservation in upstream genes is, however, not universal [117].

More recent work has asked how the amount of metabolic flux through an enzyme might affect its evolutionary rate. Here, a negative association exists between the flux through individual enzymatic reactions in yeast, as predicted by flux balance analysis, and the ratio K_a/K_s [118]. That is, enzymes with high associated metabolic flux can tolerate fewer amino acid changes. One likely explanation is that the products of high-flux enzymes play a role in multiple metabolic pathways. Thus, mutations in such enzymes, most of which reduce their metabolic output, are more likely to have deleterious effects.

Summary and Outlook

Molecular evolution studies in molecular networks are still in their infancy, partly because genome-scale data on such networks has only become available recently. A few common patterns emerge from existing work. With the possible exception of metabolic networks, a gene's position in a network has a limited influence on its rate of evolution. In the evolution of network structure, both gene duplications and gene deletions play an important role in all three networks, and rewiring of existing interactions is important in protein interaction networks and transcriptional regulation networks. Natural selection may have only a minor role in shaping those feature of global network structure that have been studied, but many other such features remain poorly investigated. In contrast, natural selection undoubtedly influences what kinds of mutations can be tolerated during network evolution. A major future challenge is to explain the structure of biological networks in evolutionary terms through a quantitative framework that accounts for all the rates of evolutionary events that influence network structure.

Bibliography

1. Bornberg-Bauer E et al (2005) The evolution of domain arrangements in proteins and interaction networks. Cell Mol Life Sci 62:435–445
2. Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge
3. Kreitman M, Akashi H (1995) Molecular evidence for natural selection. Annu Rev Ecol Syst 26:403–422
4. Ohta T (1992) The nearly neutral theory of molecular evolution. Annu Rev Ecol Syst 23:263–286
5. Smith NGC, Eyre-Walker A (2002) Adaptive protein evolution in Drosophila. Nature 415(6875):1022–1024
6. Fay J, Wyckoff G, Wu C-I (2002) Testing the neutral theory of molecular evolution with genomic data from Drosophila. Nature 415:1024–1026
7. Kreitman M (2000) Methods to detect selection in populations with applications to the human. Annu Rev Genomics Hum Genet 1:539–559
8. Pandey A, Mann MP (2001) Proteomics to study genes and genomes. Nature 405:837–846
9. Fields S, Song OK (1989) A novel genetic system to detect protein protein interactions. Nature 340(6230):245–246
10. Fromont-Racine M, Rain JC, Legrain P (1997) Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. Nat Genet 16(3):277–282
11. Bartel PL et al (1996) A protein linkage map of Escherichia coli bacteriophage T7. Nat Genet 12(1):72–77
12. Flajolet M et al (2000) A genomic approach of the hepatitis C virus generates a protein interaction map. Gene 242(1–2):369–379
13. Ito T et al (2000) Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. In: Proceedings of the National Academy of Sciences of the United States of America 97(3):1143–1147
14. Ito T et al (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. In: Proceedings of the National Academy of Sciences of the United States of America 98(8):4569–4574
15. Uetz P et al (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403(6770):623–627

16. Rain JC et al (2001) The protein-protein interaction map of Helicobacter pylori. Nature 409(6817):211–215
17. Rual JF et al (2005) Towards a proteome-scale map of the human protein-protein interaction network. Nature 437(7062):1173
18. Li SM et al (2004) A map of the interactome network of the metazoan C-elegans. Science 303(5657):540
19. Edwards AM et al (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes. Trends Genet 18(10):529–536
20. von Mering C et al (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417(6887):399–403
21. Gavin AC et al (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Faseb J 16(4/pt.1):A523–A523
22. Ho Y et al (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415(6868):180–183
23. Krogan NJ et al (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440(7084):637
24. Gavin AC et al (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440(7084):631
25. Watts DJ, Strogatz SH (1998) Collective dynamics of small-world networks. Nature 393(6684):440–442
26. Wagner A (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol Biol Evol 18:1283–1292
27. Jeong H et al (2001) Lethality and centrality in protein networks. Nature 411:41–42
28. Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. Science 296:910–913
29. Wagner A (2003) How the global structure of protein interaction networks evolves. In: Proceedings of the Royal Society of London Series B 270:457–466
30. Sole RV et al (2002) A model of large-scale proteome evolution. Adv Complex Syst 5:43–54
31. Pastor-Satorras R, Smith E, Sole RV (2003) Evolving protein interaction networks through gene duplication. J Theoret Biol 222:199–210
32. Vazquez A et al (2001) Modelling of protein interaction networks. Complexus 1:38–44
33. Berg J, Lassig M, Wagner A (2004) Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. BMC Evol Biol 4:51
34. Przulj N, Higham DJ (2006) Modelling protein-protein interaction networks via a stickiness index. J R Soc Interface 3:711–716
35. Ispolatov I, Krapivsky PL, Yuryev A (2005) Duplication-divergence model of protein interaction network. Phys Rev E 71(6):061911
36. Middendorf M, Ziv E, Wiggins CH (2005) Inferring network mechanisms: The Drosophila melanogaster protein interaction network. In: Proceedings of the National Academy of Sciences of the United States of America 102(9):3192
37. Goh KI, Kahng B, Kim D (2005) Evolution of the protein interaction network of budding yeast: Role of the protein family compatibility constraint. J Korean Phys Soc 46(2):551
38. Amoutzias GD, Robertson DL, Bornberg-Bauer E (2004) The evolution of protein interaction networks in regulatory proteins. Comp Funct Genomics 5(1):79
39. Amoutzias GD, Weiner J, Bornberg-Bauer E (2005) Phylogenetic profiling of protein interaction networks in eukaryotic transcription factors reveals focal proteins being ancestral to hubs. Gene 347(2):247
40. Pereira-Leal J, Teichmann S (2005) Novel specificities emerge by stepwise duplication of functional modules. Genome Res 4:552–559
41. Ispolatov I et al (2005) Binding properties and evolution of homodimers in protein-protein interaction networks. Nucleic Acids Res 33(11):3629
42. Pereira-Leal JB et al (2005) An exponential core in the heart of the yeast protein interaction network. Mol Biol Evol 22(3):421
43. Wagner A (2003) How the global structure of protein interaction networks evolves. In: Proceedings of the Royal Society of London Series B-Biological Sciences 270(1514):457
44. Eisenberg E, Levanon E (2003) Preferential attachment in the protein network evolution. Phys Rev Lett 91:138701–138704
45. Kunin V, Pereira-Leal JB, Ouzounis CA (2004) Functional Evolution of the Yeast Protein Interaction Network. Mol Biol Evol 21(7):1171–1176
46. Albert R, Jeong H, Barabasi AL (2000) Error and attack tolerance of complex networks. Nature 406(6794):378–382
47. Deeds EJ, Ashenberg O, Shakhnovich EI (2006) A simple physical model for scaling in protein-protein interaction networks. In: Proceedings of the National Academy of Sciences of the United States of America 103(2):311
48. Lemos B, Meiklejohn CD, Hartl DL (2004) Regulatory evolution across the protein interaction network. Nat Genet 36(10):1059
49. Papp B, Pal C, Hurst LD (2003) Dosage sensitivity and the evolution of gene families in yeast. Nature 424(6945):194–197
50. Yang J, Lusk R, Li WH (2003) Organismal complexity, protein complexity, and gene duplicability. In: Proceedings of the National Academy of Sciences of the United States of America 100(26):15661
51. Prachumwat A, Li WH (2006) Protein function, connectivity, and duplicability in yeast. Mol Biol Evol 23(1):30
52. Li L et al (2006) Preferential duplication in the sparse part of yeast protein interaction network. Mol Biol Evol 23(12):2467
53. Wuchty S (2004) Evolution and topology in the yeast protein interaction network. Genome Res 14(7):1310
54. Wuchty S, Barabasi AL, Ferdig MT (2006) Stable evolutionary signal in a Yeast protein interaction network. BMC Evol Biol 6:8
55. Mintseris J, Weng ZP (2005) Structure, function, and evolution of transient and obligate protein-protein interactions. In: Proceedings of the National Academy of Sciences of the United States of America 102(31):10930
56. Hahn MW, Kern AD (2005) Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol 22(4):803
57. Fraser H et al (2002) Evolutionary rate in the protein interaction network. Science 296(5568):750–752
58. Fraser HB, Wall DP, Hirsh AE (2003) A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol Biol 3:11
59. Batada N, Hurst L, Tyers M (2006) Evolutionary and physiological importance of hub proteins. PloS Comput Biol 2(7):e88
60. Hahn M, Conant GC, Wagner A (2004) Molecular evolution in large genetic networks: does connectivity equal importance? J Mol Evol 58:203–211

61. Jordan IK, Wolf YI, Koonin EV (2003) No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol Biol 3:1
62. Jordan IK, Wolf YI, Koonin EV (2003) Correction: no simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors evolve slowly. BMC Evol Biol 3:5
63. Bloom JD, Adami C (2003) Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. BMC Evol Biol 3:21
64. Bloom JD, Adami C (2004) Evolutionary rate depends on number of protein-protein interactions independently of gene expression level: Response. BMC Evol Biol 4:14
65. Agrafioti I et al (2005) Comparative analysis of the Saccharomyces cerevisiae and Caenorhabditis elegans protein interaction networks. BMC Evol Biol 5:23
66. Pal C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. Genetics 158:927–931
67. Drummond DA et al (2005) Why highly expressed proteins evolve slowly. In: Proceedings of the National Academy of Sciences of the United States of America 102(40):14338
68. Teichmann SA (2002) The constraints protein-protein interactions place on sequence divergence. J Mol Evol 324(3):399
69. Kim PM et al (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. Science 314(5807):1938
70. Makino T, Gojobori T (2006) The evolutionary rate of a protein is influenced by features of the interacting partners. Mol Biol Evol 23(4):784
71. Salgado H et al (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12. Nucleic Acids Res 32:D303–D306
72. Guelzim N et al (2002) Topological and causal structure of the yeast transcriptional regulatory network. Nat Genet 31(1):60–63
73. Lee T et al (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 298(5594):799–804
74. Babu MM, Teichmann SA (2003) Evolution of transcription factors and the gene regulatory network in Escherichia coli. Nucleic Acids Res 31(4):1234
75. Teichmann SA, Babu MM (2004) Gene regulatory network growth by duplication. Nat Genet 36(5):492
76. Babu MM et al (2004) Structure and evolution of transcriptional regulatory networks. Curr Opin Struct Biol 14(3):283
77. Evangelisti A, Wagner A (2004) Molecular evolution in the transcriptional regulation network of yeast. J Exp Zool/Mol Dev Evol 302B:392–411
78. Kellis M et al (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423(6937):241–254
79. Conant GC, Wagner A (2003) Convergent evolution in gene circuits. Nat Genet 34:264–266
80. Ihmels J et al (2005) Rewiring of the yeast transcriptional network through the evolution of motif usage. Science 309(5736):938
81. Babu MM, Teichmann SA, Aravind L (2006) Evolutionary dynamics of prokaryotic transcriptional regulatory networks. J Mol Biol 358(2):614
82. Maslov S et al (2004) Upstream plasticity and downstream robustness in evolution of molecular networks. BMC Evol Biol 4:9
83. Papp B, Pal C, Hurst LD (2003) Evolution of cis-regulatory elements in duplicated genes of yeast. Trends Genet 19(8):417–422
84. Milo R et al (2002) Network motifs: Simple building blocks of complex networks. Science 298(5594):824–827
85. Shen-Orr S et al (2002) Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet 31(1): 64–68
86. Mangan S, Alon U (2003) Structure and function of the feed-forward loop network motif. In: Proceedings of the National Academy of Sciences of the United States of America 100(21):11980–11985
87. Mangan S, Zaslaver A, Alon U (2003) The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks J Mol Evol 334:197–204
88. Edwards JS, Palsson BO (1999) Systems properties of the Haemophilus influenzae Rd metabolic genotype. J Biol Chem 274(25):17410–17416
89. Edwards JS, Palsson BO (2000) The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. In: Proceedings of the National Academy of Sciences of the United States of America 97(10):5528–5533
90. Forster J et al (2003) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Res 13:244–253
91. Varma A, Palsson BO (1993) Metabolic capabilities of Escherichia coli. synthesis of biosynthetic precursors and cofactors. J Theoret Biol 165:477–502
92. Schilling CH, Edwards JS, Palsson BO (1999) Toward metabolic phenomics: Analysis of genomic data using flux balances. Biotechnol Prog 15(3):288–295
93. Segre D, Vitkup D, Church G (2002) Analysis of optimality in natural and perturbed metabolic networks. In: Proceedings of the National Academy of Sciences of the USA 99:15112–15117
94. Edwards JS, Palsson BO (2000) Robustness analysis of the Escherichia coli metabolic network. Biotechnol Prog 16(6):927–939
95. Light S, Kraulis P, Elofsson A (2005) Preferential attachment in the evolution of metabolic networks. BMC Genomics 6:159
96. Light S, Kraulis P (2004) Network analysis of metabolic enzyme evolution in Escherichia coli. BMC Bioinformatics 5:15
97. Pal C et al (2006) Chance and necessity in the evolution of minimal metabolic networks. Nature 440(7084):667
98. Sakharkar MK et al (2005) Insights to metabolic network evolution by fusion proteins. Front Biosci 10:1070
99. Ebenhoh O, Handorf T, Kahn D (2006) Evolutionary changes of metabolic networks and their biosynthetic capacities. In: IEE Proceedings Systems Biology 153(5):354
100. Teichmann SA et al (2001) The evolution and structural anatomy of the small molecule metabolic pathways in Escherichia coli. J Mol Evol 311(4):693
101. Teichmann SA et al (2001) Small-molecule metabolism: an enzyme mosaic. Trends Biotechnol 19(12):482
102. Spirin V et al (2006) A metabolic network in the evolutionary context: Multiscale structure and modularity. In: Proceedings of the National Academy of Sciences of the United States of America 103(23):8774
103. Tanaka T, Ikeo K, Gojobori T (2006) Evolution of metabolic

networks by gain and loss of enzymatic reaction in eukaryotes. Gene 365:88
104. Handorf T, Ebenhoh O, Heinrich R (2005) Expanding metabolic networks: Scopes of compounds, robustness, and evolution. J Mol Evol 61(4):498
105. Pfeiffer T, Soyer OS, Bonhoeffer S (2005) The evolution of connectivity in metabolic networks. Plos Biol 3(7):1269
106. Horowitz N (1965) The evolution of biochemical syntheses – retrospect and prospect. In: Bryson H, Vogel H (eds) Evolving genes and proteins. Academic Press, New York, pp 15–23
107. Jensen R (1976) Enzyme recruitment in evolution of new functions. Annu Rev Microbiol 30:409–425
108. Alves R, Chaleil RAG, Sternberg MJE (2002) Evolution of enzymes in metabolism: A network perspective. J Mol Biol 320(4):751
109. Vitkup D, Kharchenko P, Wagner A (2006) Influence of metabolic network structure and function on enzyme evolution. Genome Biol 7(5):R39
110. Papp B, Pal C, Hurst LD (2004) Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. Nature 429(6992):661–664
111. Ma HW, Zeng AP (2004) Phylogenetic comparison of metabolic capacities of organisms at genome level. Mol Phylogenet Evol 31(1):204
112. Pal C, Papp B, Lercher MJ (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet 37(12):1372
113. Huynen MA, Dandekar T, Bork P (1999) Variation and evolution of the citric acid cycle: a genomic perspective. Trends Microbiol 7(7):281–291
114. Wang Z et al (2006) Exploring photosynthesis evolution by comparative analysis of metabolic networks between chloroplasts and photosynthetic bacteria. BMC Genomics 7:100
115. Tanaka T, Tateno Y, Gojobori T (2005) Evolution of vitamin B-6 (Pyridoxine) metabolism by gain and loss of genes. Mol Biol Evol 22(2):243
116. Rausher MD, Miller RE, Tiffin P (1999) Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. Mol Biol Evol 16(2):266
117. Cork JM, Purugganan MD (2004) The evolution of molecular genetic pathways and networks. Bioessays 26(5):479
118. Vitkup D, Kharchenko P, Wagner A (2006) Metabolic flux and molecular evolution in a genome-scale metabolic network. Genome Biol 7(5):R39

Monte Carlo Simulations in Statistical Physics

Kurt Binder
Institut für Physik, Johannes Gutenberg Universität, Mainz, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
The Metropolis Importance Sampling Algorithm as a Tool in Classical Equilibrium Statistical Mechanics
The Dynamic Interpretation of Monte Carlo Simulation and Application to Study Dynamic Processes
Overcoming the Limitations of Finite Size
Extensions to Quantum Statistical Mechanics
Future Directions
Bibliography

Glossary

Classical statistical mechanics Statistical mechanics relates the macroscopic properties of matter to basic equations governing the motion of the (many!) constituents from which matter is built from. For classical statistical mechanics these equations are Newton's laws of classical mechanics.

Critical slowing down Divergence of the relaxation time of the model describing the dynamics of a many-particle system when one approaches a second-order phase transition (= "critical point" in the phase diagram).

Detailed balance principle Relation linking the transition probability for a move and the transition probability for the inverse move to the ratio of the probability for the occurrence of these two states connected by these moves in thermal equilibrium.

Equilibrium Statistical mechanics considers "thermal equilibrium", i. e. a many-body system in contact with a (big) heat reservoir does not take up heat from this reservoir, its macroscopic properties do not change with time, and a few global properties (like temperature, pressure, particle number) suffice to characterize the state of the system.

Ergodicity Property that ensures that ensemble averages of statistical mechanics (taken with the proper probability distribution) agree with time averages taken along the trajectory along which the system moves through its state space.

Finite-size scaling Theory that describes the rounding and shifting of singularities that thermodynamic properties exhibit when the state of a system changes from one phase to another in the "thermodynamic limit" (i. e., particle number $N \to \infty$).

Importance sampling Monte Carlo method that chooses the states that are generated according to the probability distribution that one desires to realize. For example, for statistical mechanics applications, states are chosen with weights proportional to the "Boltzmann factor" {exp[−energy of the state/temperature]}.

Master equation Rate equation describing the "time"-evolution of the probability that a state occurs as a function of a "time" coordinate labeling the sequence of states.

Molecular dynamics method Simulation method for interacting many-body systems based on the numerical solution of Newton's equations of motion of classical mechanics.

Monte Carlo step Unit of (pseudo) time in (dynamically interpreted) importance sampling where, on the average, each degree of freedom in the system gets one chance to be changed (or "updated").

Quantum statistical mechanics Statistical mechanics relates the macroscopic properties of matter to basic equations governing the motion of the (many!) constituents matter is built from. For quantum statistical mechanics, this basic equation is the Schrödinger equation for the many body wavefunction. If the eigenvalue spectrum of this equation could be obtained, the canonical formalism of statistical mechanics could be straightforwardly applied; since normally this is not possible, one has to use a reformulation of the Schroedinger equation in terms of path integrals.

Random number generator (RNG) Computer subroutine to produce pseudorandom numbers that are approximately uniformly distributed in the interval from zero to unity. Approximately the subsequently generated random numbers are uncorrelated. RNG's typically are deterministic algorithms and strictly periodic, but the period is large enough that for many applications this periodicity does not matter.

Simple sampling Monte Carlo method that chooses states uniformly and at random from the available space.

Thermodynamic variables Macroscopic pieces of matter (solids, liquids, gases) in thermal equilibrium can be characterized by a few state variables, "extensive" thermodynamic variables (proportional to the particle number, such as energy, volume) and "intensive" ones (independent of the particle number, such as temperature, pressure).

Transition probability Probability that controls the move from one state to the next one in a Markovian Monte Carlo process.

Definition of the Subject

Monte Carlo simulation in statistical physics uses powerful computers to obtain information on the collective behavior of systems of many interacting particles, based on the general framework of classical or quantum statistical mechanics. Typically these systems are too complex to allow for a reliable treatment (i. e. with errors that can be controlled) by analytical theory. Monte Carlo simulation uses (pseudo)-random numbers generated also on the computer, and hence is suitable to derive estimates of probability distributions and averages derived from them. Such probability distributions (such as the so-called "canonical" distribution characterizing the equilibrium state of matter at a given temperature and volume) are the basic objects of statistical thermodynamics. While the latter field of physics provides a convenient formal framework, it does in most cases not yield a convenient tool for explicit calculation, and such an approach is provided by Monte Carlo simulation. In fact, already the first application of the Metropolis importance sampling algorithm in 1953 addressed a problem of this kind, namely the equation of state of hard disks. Since then Monte Carlo simulation has become an extremely useful and versatile tool of statistical physics, with applications varying from many subfields of physics (from condensed matter (i. e. liquids and solids) to elementary particles) to neighboring fields (physical chemistry, theoretical biology, stochastic modeling of complex phenomena in society such as traffic flows, stock market fluctuations, etc.), where methodologies "borrowed" from physics are increasingly applied.

Introduction

One important problem of statistical physics is the explanation of the macroscopic properties of solids, liquids and gases in terms of the atomistic description: 1 cm^3 of a solid contains about $N = 10^{22}$ atoms, which in turn are composed of electrons and nuclei. The basic interactions keeping the atoms together are the Coulomb forces between particles of opposite electrical charge, and these Coulomb forces are also responsible for effective forces acting between atoms as a whole. In the present article, we are not concerned with the quantum-mechanical problem of predicting these forces, but rather assume them to be known, and deal only with the many-body problem of interacting atoms in the framework of classical statistical mechanics. Macroscopic properties (e. g. the density ρ of a fluid, the magnetization density $\boldsymbol{m}$ of a ferromagnet, etc.) will be denoted as "observables" $A(\boldsymbol{X})$ which depend on the degrees of freedom of the N particles (these degrees of freedom are formally denoted as a vector $\boldsymbol{x}$, which encompasses the configurational "phase space" of the system). Statistical thermodynamics then shows that in a thermal equilibrium state that is characterized by parameters such as temperature T, pressure p etc., averages $\langle A \rangle$ are to be

computed as

$$\langle A \rangle = \int \mathrm{d}\boldsymbol{X} P_{\mathrm{eq}}(\boldsymbol{X}) A(\boldsymbol{X}) \,, \tag{1}$$

where $P_{\mathrm{eq}}(\boldsymbol{X})$ the probability that in equilibrium the "microstate" (=point in phase space) $\boldsymbol{X}$ occurs. E. g., when the equilibrium of a system is characterized by given volume V and given T, then $P_{\mathrm{eq}}(\boldsymbol{X})$ is described by the so-called "canonic distribution"

$$P_{\mathrm{eq}}(\boldsymbol{X}) = (1/Z_N) \exp[-\mathcal{H}(\boldsymbol{X})/k_{\mathrm{B}} T] \,, \tag{2}$$

where $Z_N = \int \mathrm{d}\boldsymbol{X} \exp[-\mathcal{H}(\boldsymbol{X})/k_{\mathrm{B}} T]$ is called "partition function", k_{B} is Boltzmann's constant, and the "Hamiltonian" $\mathcal{H}(\boldsymbol{X})$ describes the interaction energies among the particles (and possible contributions to the energy due to external fields). For N particles in d-dimensional space Eq. (1) would involve a dN-dimensional integral, and in general the task posed by Eqs. (1), (2) cannot be carried out analytically. Only when the particles do not interact (i. e., an ideal gas, or related problems such as a harmonic solid where one can reduce the problem to an ideal gas of "phonons" describing the lattice vibrations, etc.), is the problem easily tractable, since the multidimensional integration factorizes. In the case of interactions, analytically soluble problems are extremely rare. E. g., the famous Ising model of ferromagnetism

$$\mathcal{H} = -\sum_{i \neq j} J_{ij} S_i S_j - H \sum_i S_i \,, \quad S_1 = \pm 1 \,, \tag{3}$$

where the first term on the right hand side describes the exchange interaction between spins at lattice sites $i\,j$ of a crystal lattice, while the second term describes the energy due to an external magnetic field, can be solved (in the case of the interaction J_{ij} being nonzero only for nearest neighbor pairs on the lattice) in $d = 1$, but there the system stays paramagnetic at all temperatures. In $d = 2$ dimensions and $H = 0$, one also can solve the problem, though this requires very tedious and subtle mathematics, but no solution is known in either $d = 2$ or $d = 3$ for any more complicated cases ($H \neq 0$, J_{ij} nonzero for next nearest or even more distant neighbors; etc.). An approximate solution, where the pairwise interaction is reduced to a coupling to an effective mean field (one puts $S_i S_j \approx S_i \langle S_j \rangle + \langle S_i \rangle S_j - \langle S_i \rangle \langle S_j \rangle$) would reduce the problem to an effective single-particle problem, similar to the problem of the ideal paramagnet (for which $Z_N = Z_1^N$). However, comparison with the exactly known cases in $d = 1$ and $d = 2$ shows that this mean field approximation is unsatisfactory, the obtained results may be even qualitatively wrong (such as the prediction of ferromagnetism for $d = 1$), and uncontrollable errors occur. In almost all cases of the statistical mechanics of many interacting degrees of freedom, no analytical tools exist to solve the problem either exactly or approximately with errors that can always be controlled (in particular near the phase transitions).

Monte Carlo simulation amounts to replacing the integration in Eq. (1) by a summation over a representative statistical sample of M points $\{\boldsymbol{X}_\nu\}$ that is suitably chosen (what this actually means, will be discussed in the next section). The choice of this sample requires random numbers, which are produced on the computer via a pseudorandom number generator (RNG). These random numbers must be uniformly distributed between zero and unity, and should be uncorrelated. Actually, all random numbers due to RNG's exhibit some residual correlations, which may cause erroneous results in Monte Carlo simulations, and hence devising "good" RNG's is a matter of longstanding concern (e. g., [11,12,14]). Of course, due to the finiteness of M there occurs the so-called "statistical errors" (as a matter of fact, for some algorithms statistical and systematic errors are not easy to disentangle, see [13]); but by making M bigger and bigger, these errors can be made smaller and smaller, and hence controlled, and techniques exist to estimate these errors reliably [1].

Of course, the application outlined above refers only to a subset of problems in statistical physics, but many other problems can be reduced to it. E. g., the problem of computing averages in quantum statistical mechanics can be reduced to Eqs. (1), (2) by path-integral Monte Carlo (PIMC) methods [7,13,20]. In this method, the quantum character of particles enters, on the one hand, by replacing each quantum particle by a chain of classical particles (the coordinate along the chain is the "imaginary time" coordinate of the path integral formalism).

While in this way the delocalization of quantum particles at low temperatures (note that Heisenberg's uncertainty principle forbids to specify both the spatial coordinates and the momentum of a quantum particle precisely) is elegantly taken into account, the statistics of the particles (for fermions the wave functions need to be strictly antisymmetric with respect to the interchange of coordinates for any pair of particles) is still a challenge for such quantum Monte Carlo methods [7].

Also seemingly unrelated problems, such as the theory of elementary particles which is a field theory of matter fields on the femtometer scale and of gauge fields respecting the basic symmetry principles of relativistic quantum field theory, can be cast into a form closely related to Eqs. (1), (2). The generating functional

$$Z = \int \mathcal{D}A \mathcal{D}\bar{\psi} \mathcal{D}\psi \exp[-S_g(A, \bar{\psi}, \psi)] \tag{4}$$

formally corresponds to the partition function in statistical mechanics. Equation (4) involves functional integration over the gauge fields (here A stands symbolically for the vector potential of electrodynamics in the four-dimensional continuum, 3 space + 1 time dimensions) and over the fermionic matter field (described symbolically by ψ and $\bar{\psi}$). The action S_g of the theory contains a coupling constant g which is related to inverse temperature, when one invokes the analogy to statistical mechanics. In fact, to remove ultraviolet divergences that would otherwise plague this quantum field theory one replaces the four-dimensional continuum by a lattice and hence the theory is called "lattice gauge theory" [15]. Monte Carlo simulations based on this formalism promise to become a powerful approach to unravel the properties of hadrons and other elementary particles, beyond the regime of parameters where analytical theories based on systematic expansions in terms of small parameters ("perturbation theory") work.

While both in the case of Eqs. (2) and (4) the explicit probability distribution of the interacting many-particle system needs to be constructed itself in the course of the Monte Carlo sampling, via Importance Sampling methods (as will be described in the next section), there exist also simpler cases where a generation of "microstates" of the N-particle system is straightforward, but the analysis of the properties of these microstates is difficult and hence requires large scale simulation. As an example, we mention the "percolation problem" [19]. Suppose a ferromagnet, described by Eq. (3), is randomly diluted with nonmagnetic atoms, such that a lattice site i carries a spin S_i with probability p and no spin with probability $1-p$, J_{ij} being nonzero only between nearest neighbor pairs of spins. Then a ferromagnetic ground state of the lattice is only possible if p exceeds the "percolation threshold" p_c, where for the first time an infinite "percolating cluster" of magnetic atoms connected by "bonds" J_{ij} and extending throughout the whole lattice occurs. Choosing random numbers η uniformly distributed between zero and unity, a chosen site is taken by a magnetic atom if $\eta < p$ and otherwise it is taken by a nonmagnetic atom. All lattice sites are occupied independently of each other and all configurations of the lattice generated in this way have equal a priori probability. While the generation of a sample of states by such a "simple random sampling" strategy hence is straightforward, the analysis of the (fractal) percolation clusters near the percolation threshold is a difficult task. Note that many other problems where simple sampling suffices exist (e. g. generation of random walks on lattices, a problem arising in the context of simulating flexible macromolecules, or of diffusion processes). However, we shall not dwell on simple sampling further but rather refer to the literature [6,13].

The Metropolis Importance Sampling Algorithm as a Tool in Classical Equilibrium Statistical Mechanics

If we would choose microstates $\boldsymbol{X}$ of a many-body system according to a simple sampling strategy to sample Eq. (2), we would fail to get useful results, except for very small values of N. In fact, the probability distribution $P_{\mathrm{eq}}(\boldsymbol{X})$ has a very sharp peak in phase space where all extensive variables (extensive variables are proportional to N for $N \to \infty$) $A(\boldsymbol{X})$ are close to their average values $\langle A \rangle$. This peak may be approximated by a Gaussian centered at $\langle A \rangle$, with a relative half-width of order $1/\sqrt{N}$. Hence, for a practically useful method, one should not sample the phase space uniformly, but the points $\boldsymbol{X}_\nu$ over which the sampling is extended must be chosen preferentially from the important region of phase space, i. e., the vicinity of the peak of this probability distribution. This goal is achieved by the importance sampling method (Metropolis et al. [16]): One generates a Markov chain of M configurations $\boldsymbol{X}_\nu$, in terms of a Markovian transition probability $W(\boldsymbol{X}_\nu \to \boldsymbol{X}'_\nu)$ that rules the "Monte Carlo moves" from an old state $\boldsymbol{X}_\nu$ to *a* new state $\boldsymbol{X}'_\nu$. Starting from some (arbitrary) initial state $\boldsymbol{X}_1$ one creates a "random walk through phase space", choosing W such that in the limit of $M \to \infty$ a point $\boldsymbol{X}_\nu$, is chosen with a probability proportional to $P_{\mathrm{eq}}(\boldsymbol{X})$, and hence the average in Eq. (1) is approximated by a simple arithmetic average,

$$\bar{A} = (M - M_0)^{-1} \sum_{\nu=M_0+1}^{M} A(\boldsymbol{X}_\nu) \,. \tag{5}$$

Here it is anticipated that in practice it is better to eliminate the residual influence of the initial state $\boldsymbol{X}_1$ by eliminating the first $M_0 \gg 1$ states from the average. A condition sufficient to ensure that the points $\boldsymbol{X}_\nu$ are actually chosen proportional to the desired probability $P_{\mathrm{eq}}(\boldsymbol{X}_\nu)$ is known as the detailed balance principle,

$$P_{\mathrm{eq}}(\boldsymbol{X}) W(\boldsymbol{X} \to \boldsymbol{X}') = P_{\mathrm{eq}}(\boldsymbol{X}') W(\boldsymbol{X}' \to \boldsymbol{X}) \,. \tag{6}$$

Detailed considerations of how the Monte Carlo moves $\boldsymbol{X} \to \boldsymbol{X}'$ need to be chosen and why the Monte Carlo method actually converges to Eq. (2) can be found in the literature (Binder and Heermann [6], Frenkel and Smit [9]). Here we only emphasize that one major disadvantage of this method is that knowledge on the normalization factor of $P_{\mathrm{eq}}(\boldsymbol{X})$, the partition function Z (Eq. (2)), is lost. This is unfortunate, since from Z one could obtain

the free energy F via $F = -k_B T \ln Z$, as well as the entropy. Finding ways of obtaining F via alternative sampling algorithms, that yield directly information on the energy density of states, is an active area of research (see "Bibliography" and Landau and Binder [13]).

It needs also to be pointed out that this Metropolis method can be used for sampling any distribution $P(\boldsymbol{X})$: one simply has to choose a transition probability $W(\boldsymbol{X} \to \boldsymbol{X}')$ that satisfies a detailed balance condition with $P(\boldsymbol{X})$ rather than $P_{eq}(\boldsymbol{X})$ as given in Eq. (2).

We continue by giving some comments on the practical implementation of Monte Carlo algorithms. One obvious question is, what is meant in practice by the transition $\boldsymbol{X} \to \boldsymbol{X}'$? However, there is no general answer to this question: the choice of the move may depend both on the model under consideration and the simulation purpose. Since Eqs. (2),(6) imply that $W(\boldsymbol{X} \to \boldsymbol{X}')/W(\boldsymbol{X}' \to \boldsymbol{X}) = \exp(-\delta\mathcal{H}/k_B T)$, $\delta\mathcal{H}$ being the energy change caused by the move from $\boldsymbol{X}$ to $\boldsymbol{X}'$, typically it is necessary to carry out small changes $\delta\boldsymbol{X} = \boldsymbol{X}' - \boldsymbol{X}$ only. This is achieved in practice by moving only one (or a few) degrees of freedom at a time. Only in rare cases ("cluster algorithms" for Ising models, "pivot algorithm" for self-avoiding walks, etc.) is it possible to find algorithms involving a large $\delta\boldsymbol{X}$. It is also useful to realize that often $W(\boldsymbol{X} \to \boldsymbol{X}')$ can be written as a product of an "attempt frequency" times an "acceptance rate". By clever choice of the attempt frequency it is sometimes possible to attempt larger moves and still have a high acceptance and thus make the computations more efficient.

The types of Monte Carlo moves can also be adjusted to the choice of statistical ensemble that one wishes to realize. E. g., for a grandcanonical ensemble the chemical potential μ (in addition to volume V and temperature T) is a given variable. Suitable moves then include particle insertions and deletions, i. e., the particle number N in the simulation box fluctuates, as well as the pressure p (which can be sampled e. g. using the virial theorem). Conversely, choosing the NpT ensemble one must include moves where the volume V of the system changes, $V \to V' = V \pm \Delta V$. Also the chemical potential then fluctuates (and can be sampled using the Widom particle insertion method). For details on all these aspects, we refer to more detailed textbooks [9,13].

Another arbitrariness concerns the order in which particles are selected for considering a move. E. g., simulating a lattice model one may go through the lattice in a typewriter fashion. Sometimes it is advisable to use sublattices, e. g. the "checker board algorithm" where white and black sublattices are updated alternatively to allow an efficient "vectorization" of the code. However, if the simulation purpose is to extract dynamic information (invoking the interpretation of the Markov chain in terms of a master equation as will be discussed below), it is better to choose lattice sites (or particle labels, respectively) at random.

We briefly mention the practical realization of the algorithm, choosing the Ising model (Eq. (3)) as a simple example. Then the move $\boldsymbol{X} \to \boldsymbol{X}'$ simply may be a single spin flip $\{S_i \to -S_i\}$, for instance. The transition probability can be chosen as

$$W(\boldsymbol{X} \to \boldsymbol{X}') = \begin{cases} \exp(-\delta\mathcal{H}/k_B T) & \text{if } \delta\mathcal{H} > 0 \\ 1 & \text{otherwise} . \end{cases} \quad (7)$$

One compares W for this trial move with a random number η, uniformly distributed in the interval $0 < \eta < 1$; if $W < \eta$, the trial move is rejected and the old state is counted once more in the average. Then another trial is made. If $W > \eta$, on the other hand, the trial move is accepted, and the new configuration thus generated is taken into account in the average, and serves as a starting point for the next step.

Since successive states $\boldsymbol{X}_\nu$ in this process differ only very little, e. g. by a single spin flip in the case of the Ising model, they are highly correlated. Therefore, it is not straightforward to estimate the error of the average. Let us assume that only every nth step is included in the average, and that after n steps these correlations have completely died out. Then the standard estimate for the statistical error $\overline{(\delta A)^2}$ of $\bar{A} = (\tilde{M}^{-1}) \sum A_k$ with $\tilde{M} = (M - M_0)/n$, $A_k = A(\boldsymbol{X}_\nu)$ with $\nu = kn$, applies

$$\overline{(\delta A)^2} = [\tilde{M}(\tilde{M} - 1)]^{-1} \sum_{k=1}^{\tilde{M}} (A_k - \bar{A})^2 , \quad \tilde{M} \gg 1 . \quad (8)$$

However, the judgment when correlations have died out is subtle (see next section), and great care is needed to derive reliable error estimates [1].

The Dynamic Interpretation of Monte Carlo Simulation and Application to Study Dynamic Processes

Often it is useful to associate a time variable t to the index ν of successively generated states and to discuss the probability $P(\boldsymbol{X}, t)$ that a state $\boldsymbol{X}$ occurs at time t. This probability evolves according to the following master equation

$$\frac{\mathrm{d}}{\mathrm{d}t} P(\boldsymbol{X}, t) = -\sum_{\boldsymbol{X}'} W(\boldsymbol{X} \to \boldsymbol{X}') P(\boldsymbol{X}, t) + \sum_{\boldsymbol{X}'} W(\boldsymbol{X}' \to \boldsymbol{X}) P(\boldsymbol{X}', t) . \quad (9)$$

Obviously the probability $P(\boldsymbol{X}, t)$ decreases due to all processes that lead away from the considered state $\boldsymbol{X}$ (first sum on the right hand side of Eq. (9)), while it increases due to all processes that lead to the considered state from other states $\boldsymbol{X}'$ (second sum on the right hand side of Eq. (9)). Thus, Monte Carlo sampling (i. e., the sequence of generated states $\boldsymbol{X}_1 \to \boldsymbol{X}_2 \to \cdots \to \boldsymbol{X}_\nu \to \boldsymbol{X}_{\nu+1} \to \cdots$) can be interpreted as a numerical realization of a stochastic time evolution described by the master equation, Eq. (9).

Of course, the units of the "time" in Eq. (9) are not in an obvious way related to the units of physical time (as it appears in the Newtonian equation of motion or in the Schrödinger equation, respectively). Thus, it is common practice to normalize the "Monte Carlo time unit" such that in a system with N particles one attempts N single particle moves per unit time. This is called "one Monte Carlo step (MCS)".

For the thermal equilibrium distribution $P(\boldsymbol{X}, \boldsymbol{t}) = P_{\mathrm{eq}}(\boldsymbol{X})$, Eq. (2) there is no change of probability with time according to Eq. (9), since the right hand side of Eq. (9) vanishes because of the detailed balance principle, Eq. (6). Thus, thermal equilibrium arises as the stationary solution of the master equation. Markov processes for which Eq. (9) holds involve a relaxation toward thermal equilibrium.

The time evolution of a physical system (whose trajectory through phase space, according to classical statistical mechanics follows from Newton's equations of motion) in general will differ from the time evolution that follows from Eq. (9), a stochastic rather than deterministic trajectory through phase space. For example, Eq. (9) never describes any propagating modes such as sound waves in a fluid, phonons in a crystal, etc.

Nevertheless, often the trajectory described by Eq. (9) does have physical significance: this is because often one does not wish to describe the full set of dynamical variables of a system, but rather a subset only: for instance, in a solid where diffusion occurs via hopping of atoms over energy barriers to the vacant lattice sites, the mean time between successive hops is orders of magnitude larger than the time scale of atomic vibrations. The latter can be well approximated as a heat bath, as far as diffusion is concerned, and – at least in principle – the transition probability in Eq. (9) describing such a hopping process in a solid can be estimated either by Molecular Dynamics methods or by approximate analytic methods such as transition state theory.

There are many examples where such a separation of time scales for different degrees of freedom occurs. As a rule of thumb, any slow relaxation phenomena may be modeled by Monte Carlo: kinetics of nucleation, spinodal decomposition in alloys, growth of ordered domains in superstructure solids or in adsorbed mono-layers on surfaces, kinetic growth of rough interfaces, etc. Of course, one must pay attention to build in the relevant conservation laws into the model properly (e. g., in a binary metallic alloy the number of atoms of both species A,B is conserved, while the number of vacancies and interstitials may not be conserved). Often (such as in surface diffusion processes) it is not a priori obvious what are the elementary steps that one needs to include into such a "kinetic Monte Carlo" study, and for a realistic description of their rates extensive electronic structure calculations may be needed, in order to be able to connect the Monte Carlo time unit to physical time.

There are also some cases where a Monte Carlo description is a borderline case, as far as the faithful modeling of relaxation phenomena is concerned. E. g., macromolecules in polymer melts undergo Brownian motion (described approximately by analytic models such as Rouse or reptation models, respectively). In such a case, the heat bath for the slow conformational transitions of the polymer is provided by the fast bond-length and bond angle vibrations. However, a Monte Carlo description is not useful for the modeling of polymer melts under flow [5].

Even if one is not interested in any dynamic properties of the model that is simulated by Monte Carlo, one should be aware of important consequences of Eq. (9). Since Eq. (9) implies that the arithmetic average Eq. (5) has the meaning of a time average ($t_M = M/N$, $t_{M_0} = M_0/N$, $A(\boldsymbol{X}_\nu) \equiv A(t)$)

$$\bar{A} = (t_M - t_{M_0})^{-1} \int_{t_{M_0}}^{t_M} A(t)\mathrm{d}t \,, \tag{10}$$

"ergodicity" is a problem here, as it is for Molecular Dynamics simulations. By "ergodicity" it is meant that time averages and ensemble averages agree, in the limit of large enough time intervals $t_M - t_{M_0}$. However, near a first order transition where (in the thermodynamic limit $N \to \infty$) several phases may coexist, each of these phases plays the role of a separate "ergodic component" from which a trajectory never escapes to another ergodic component. A consequence of this problem is the observation of hysteresis. Sometimes the considered Monte Carlo moves do not even allow one in principle to reach all the states $\boldsymbol{X}_\nu$ (a well-known example includes algorithms with local moves for self-avoiding walks on lattices, see Binder [5]). Also problems occur where the system develops a "rugged free energy landscape", e. g. spin glasses (Binder and Young [4]), where a spectrum of relaxation times develops that spans many decades of time. Similar as in experiments, one then may observe phenomena such

as "aging" in a Monte Carlo simulation, and it is difficult to judge whether or not thermal equilibrium has been reached.

Time-displaced correlation functions $\langle A(t)B(0)\rangle$ described by Eq. (9) are then estimated as

$$\overline{A(t)B(0)} = (t_M - t - t_{M_0})^{-1} \int_{t_{M_0}}^{t_M - t} A(t+t')B(t')\mathrm{d}t' . \quad (11)$$

Apart from its interest for the study of dynamical properties of the considered models, Eq. (11) is useful to interpret the error due to the correlation between subsequently generated configurations. When we do not assume that subsequent A_k's in Eq. (8) are uncorrelated, we rather obtain

$$\langle(\delta A)^2\rangle = \left\langle\left[\frac{1}{\tilde{M}}\sum_{k=1}^{\tilde{M}}(A_k - \langle A\rangle)\right]^2\right\rangle = \frac{1}{\tilde{M}}\left\{\langle A^2\rangle - \langle A\rangle^2 + 2\sum_{k=1}^{\tilde{M}}\left(1-\frac{k}{\tilde{M}}\right)[\langle A_0 A_k\rangle - \langle A\rangle^2]\right\} . \quad (12)$$

Now we remember that a time $t = k\delta t = k(n/N)$ is associated with the kth state, and transform the sum in Eq. (12) to a time integral

$$\langle(\delta A)^2\rangle = \frac{1}{\tilde{M}}[\langle A^2\rangle - \langle A\rangle^2]\left\{1 + \frac{2}{\delta t}\int_0^{t_{\tilde{M}}}(1 - t/t_{\tilde{M}})\phi_{AA}(t)\mathrm{d}t\right\} , \quad (13)$$

where a relaxation function ϕ_{AA} has been introduced,

$$\phi_{AA}(t) = [\langle A(0)A(t)\rangle - \langle A\rangle^2]/[\langle A^2\rangle - \langle A\rangle^2] . \quad (14)$$

Defining a relaxation time

$$\tau_{AA} = \int_0^\infty \phi_{AA}(t)\mathrm{d}t \quad (15)$$

one finds for $\tau_{AA} \ll \tilde{M}\delta t = t_{\mathrm{obs}}$, the "observation time " of the system during the course of the simulation, that

$$\langle(\delta A)^2\rangle = \frac{1}{\tilde{M}}[\langle A^2\rangle - \langle A\rangle^2](1 + 2\tau_{AA}/\delta t) \approx 2(\tau_{AA}/\tau_{\mathrm{obs}})[\langle A^2\rangle - \langle A\rangle^2] , \quad (16)$$

where in the last step we have assumed $\tau_{AA} \gg \delta t$. Comparing to Eq. (8), we see that the error is enhanced by a "dynamic factor" $1 + 2\tau_{AA}/\delta t$ (or, equivalently, one has to choose δt so large that $\tilde{M} = \tau_{\mathrm{obs}}/\tau_{AA}$, to avoid correlations between subsequent states). Near critical points of second-order phase transitions, τ_{AA} diverges ("critical slowing down", see [10] for a detailed discussion).

For a discussion of nonlinear relaxation processes, it is useful to consider the evolution of averages $\langle A(t)\rangle$,

$$\langle A(t)\rangle = \sum_{\boldsymbol{X}} P(\boldsymbol{X}, t)A(\boldsymbol{X}) = \sum_{\boldsymbol{X}} P(\boldsymbol{X}, 0)A(\boldsymbol{X}(t)) . \quad (17)$$

Here we used the interpretation that the ensemble average involved is an average over an ensemble of initial states (weighted by $P(\boldsymbol{X}, 0)$), which evolve according to Eq. (9). In practice, Eq. (17) means an average over a large number $n_{\mathrm{run}} \gg 1$ of statistically independent runs, $[\bar{A}(t)]_{\mathrm{av}} = n_{\mathrm{run}}^{-1}\sum_{\ell=1}^{n_{\mathrm{run}}} A(t,\ell)$, where $A(t,\ell)$ is the observable A recorded at time t in the ℓth run. Then nonlinear relaxation functions $\phi_A^{n\ell}(t)$ and relaxation times $\tau_A^{(n\ell)}$ are defined as

$$\phi_A^{(n\ell)}(t) = [\langle A(t)\rangle - \langle A(\infty)\rangle]/[\langle A(0)\rangle - \langle A(\infty)\rangle] , \quad (18)$$

$$\tau_A^{(n\ell)} = \int_0^\infty \phi_A^{(n\ell)}(t)\mathrm{d}t . \quad (19)$$

Note that the condition that enough states M_0 at the beginning of the Monte Carlo sampling were omitted to eliminate the possible influence of the starting state $\boldsymbol{X}_1$, reads $t_{M_0} \gg \tau_A^{(n\ell)}$. Again, however, care is needed when one studies second-order phase transitions: then nonlinear relaxation functions may exhibit power law behavior rather than exponential decay to equilibrium. E. g., for an Ising model in thermal equilibrium we have [13] for the magnetization $m(t)$ and susceptibility $\chi(t)$ [assuming one starts from a perfectly aligned state at T_c, and $N \to \infty$]

$$m(t) \propto t^{-\beta/z\nu} , \quad \chi(t) \propto t^{\gamma/z\nu} , \quad t \to \infty , \quad (20)$$

where β, γ are the critical exponents of the order parameter and susceptibility in thermal equilibrium (assuming a d-dimensional lattice of size L for $L \to \infty$)

$$\langle|m|\rangle \propto (1 - t/T_c)^\beta , \chi \equiv L^d[\langle m^2\rangle - \langle|m|\rangle^2] \propto |1 - T/T_c|^{-\gamma} , \quad (21)$$

while ν and z are the critical exponents of the correlation length ξ and the relaxation time τ_{mm},

$$\xi \propto |1 - T/T_c|^{-\nu} , \quad \tau_{mm} \propto \xi^z . \quad (22)$$

Also the nonlinear relaxation time diverges $\{\tau_m^{(n\ell)} \propto \xi^{z-\beta/\nu}\}$. In fact, testing for Eqs. (20) is a possible approach

to study critical phenomena by Monte Carlo methods, avoiding the need to equilibrate the system ("nonequilibrium relaxation method"). However, this approach requires the use of very large systems, since all critical divergences considered in Eqs. (21), (22) are rounded off when the correlation length $\xi(t)$ (which grows as $\xi(t) \propto t^{1/z}$, see [18]) has grown to a value comparable to L. Such finite size effects are also important to consider in the context of equilibrium Monte Carlo studies, as will be discussed in the next section. Note also that power law growth of fluctuations also occurs for $T < T_c$, when we start an Ising model in a disordered spin configuration, domains grow according to a relation $\ell(t) \propto t^{1/2}$ for their linear dimension and [18]

$$\chi(t) \propto t^{d/2} \,. \tag{23}$$

Then equilibrium is reached only when $\ell(t) \approx L$, implying that $\tau_m^{(n\ell)} \propto L^2$. This consideration shows that the choice of the initial state in Monte Carlo sampling also should be done with care: for $T < T_c$ equilibrium is reached much faster if we start from a well-ordered initial state, of course.

Overcoming the Limitations of Finite Size

While systematic analytic expansions or closed-form approximations often are useful for systems away from phase boundaries where first- or second-order phase transitions occur, such methods often are of doubtful value near phase transitions. Thus, the study of phase transitions is one of the most important fields where Monte Carlo simulations are useful and important. However, sharp phase transitions occur in the thermodynamic limit only, $N \to \infty$. Of course, this is no problem for real systems: even a small water droplet freezing into a snowflake may still contain $N = 10^{18}$ water molecules, and thus the relative shift and rounding of the transition are of order $N^{-1/3} = 10^{-6}$ and $N^{-1/2} = 10^{-9}$, respectively. But the situation is different for simulations, where an economical use of computer resources requires to study rather small systems (of order $N = 10^2$ to $N = 10^6$ are typical), and hence finite size effects need careful consideration [6].

It turns out, however, that these finite size effects are not just only a limitation, but also a valuable tool to infer properties of the infinite system from the finite size behavior. As an example, we discuss the phase transition of the Ising ferromagnet (Fig. 1), which has a second-order transition at a critical temperature T_c, with critical behavior as characterized by Eqs. (21), (22). In a finite system, of course, ξ cannot exceed L, and hence these critical singularities are smeared out. Now finite size scaling theory [8,17] implies that such finite size effects can be understood from the principle that "L scales with ξ". Hence it is plausible that the magnetization probability $P_L(m)$ can be written [2]

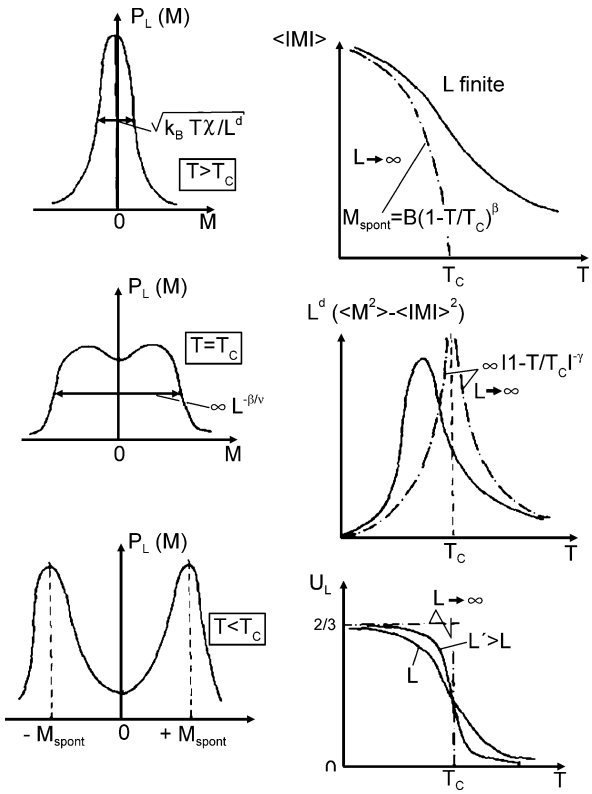

Monte Carlo Simulations in Statistical Physics, Figure 1
Schematic evolution of the order parameter probability distribution $P_L(m)$ from $T > T_c$ to $T < T_c$ (from *above* to *below*, *left part*), for an Ising ferromagnet (where M is the magnetization per spin) in a cubic box of volume $V = L^d$. The *right part* shows the corresponding temperature dependence of the mean order parameter $\langle|M|\rangle$, the susceptibility $k_B T \chi' = L^d(\langle M^2\rangle - \langle|M|\rangle^2)$, and the reduced fourth-order cumulant $U_L = 1 - \langle M^4\rangle/[3\langle M^2\rangle^2]$. *Dash-dotted curves* indicate the singular variation that results in the thermodynamic limit, $L \to \infty$

$$P_L(m) = L^x \tilde{P}(L/\xi, mL^x)\,, \quad x = \beta/\nu \,. \tag{24}$$

Here $P_L(m)$ satisfies the normalization $\int \mathrm{d}m P_L(m) = 1$, $\tilde{P}$ is a scaling function, and the result $x = \beta/\nu$ follows from the fact that

$$\langle|m|\rangle = L^{-x}\tilde{m}(L/\xi) = L^{-\beta/\nu}\tilde{m}(L/\xi) \tag{25}$$

for $L \to \infty$ must reduce to $\langle|m|\rangle \propto \xi^{-\beta/\nu}$. This is only possible when $\tilde{m}(\zeta \to \infty) \propto \zeta^x$ to cancel the power of L and when $x = \beta/\nu$. Since similarly $\langle|m|^k\rangle = L^{-k\beta}\tilde{m}_k(L/\xi)$, one derives a similar scaling relation for the

susceptibility,

$$k_B T\chi \equiv L^d(\langle m^2\rangle - \langle |m|\rangle^2) = L^{\gamma/\nu}\tilde{\chi}(L/\xi)\,, \tag{26}$$

where the hyperscaling relation $\gamma/\nu = d - 2\beta/\nu$ was invoked, and $\tilde{\chi}(\zeta) \equiv \tilde{m}_2 - \tilde{m}^2$. The fourth order cumulant U_L is a function of L/ξ only,

$$U_L = 1 - \langle m^4\rangle/[3\langle m^2\rangle^2] = \tilde{U}(L/\xi)\,. \tag{27}$$

For $T > T_c$ and large L the distribution $P_L(m)$ tends to a gaussian and hence $U_L \to 0$; for the double-gaussian distribution for $T < T_c$, however, $U_L \to 2/3$. For $T = T_c$ finite size scaling implies $U_L = \tilde{U}(0)$, independent of L. As a consequence, when one studies U_L as a function of temperature for different choices of L one should find T_c from a common intersection point. This "cumulant intersection method" has become a very widespread and useful tool for the study of critical phenomena [6].

Also the analysis of $P_L(m \approx 0)$ for $T < T_c$ is very useful [3]. The state of the system then is dominated by a two-phase configuration, i. e. (because of periodic boundary conditions) a slab-like domain with negative magnetization is separated from a domain with positive magnetization by two interfaces of area L^{d-1}. As a consequence, the deep minimum of $P_L(m \approx 0)$ in Fig. 1 is described by

$$\ln[P_L(m \approx 0)/P_L(m \approx M_{\text{spont}})] = -2L^{d-1} f_{\text{int}}/k_B T\,, \tag{28}$$

where $\pm M_{\text{spont}}$ characterizes the peak positions of $P_L(m)$ for $T < T_c$, and f_{int} is the interfacial free energy (per unit area). As a consequence, one can extract estimates for f_{int} from an analysis of $P_L(m \approx 0)$ as well. This approach has also found widespread applications for various systems (including phase separation in simple fluids and fluid mixtures and polymer solutions and blends, colloid-polymer mixtures, etc.).

A simple discussion of finite size effects at first order transitions is similarly possible. The phases coexisting at the first-order transition are again described by gaussians. In a finite system these phases coexist not only right at the transition, but over a finite parameter region. The relative weights of these states are given in terms of the free energy differences of these phases. From this phenomenological description, energy and order parameter distributions and their moments can be derived. One finds that the maximum of specific heat and susceptibility scale proportional to the volume, i. e. $k_B T\chi_{\max} \propto L^d$ (instead of $L^{\gamma/\nu}$ as at a second-order transition, Eq. (26).

Extensions to Quantum Statistical Mechanics

In quantum mechanics, an observable $A(\boldsymbol{X})$ now is represented by an quantum mechanical operator $\hat{A}$, and hence Eq. (1) is replaced by

$$\begin{aligned}\langle\hat{A}\rangle &= Z^{-1}\,Tr\exp(-\hat{\mathcal{H}}/k_B T)\hat{A}\\ &= Z^{-1}\sum_n \langle n|\exp(-\hat{\mathcal{H}}/k_B T)|n\rangle\,,\end{aligned} \tag{29}$$

where $\hat{\mathcal{H}}$ is the Hamiltonian of the system, and the states $|n\rangle$ form a complete, orthonormal set. In general, the eigenvalues E_α and eigenstates $|\alpha\rangle$ of the Hamiltonian are not known ($\mathcal{H}|\alpha\rangle = E_\alpha|\alpha\rangle$), and we wish to evaluate the trace in Eq. (29) without attempting to diagonalize the Hamiltonian. This can be achieved by path-integral Monte Carlo (PIMC); other versions of quantum Monte Carlo methods which focus on finding the ground state and its energy are outside of consideration here.

The basic idea of PIMC can already be explained for the simple case of a single particle in one dimension in a potential $V(x)$. In position representation, the Hamiltonian is

$$\hat{\mathcal{H}} = -\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + V(x) = \hat{E}_{\text{kin}} + \hat{V}\,, \tag{30}$$

where $\hbar$ is Planck's constant. The problem is the fact that the operators of kinetic energy $\hat{E}_{\text{kin}}$ and potential energy $\hat{V}$ do not commute, $[\hat{E}_{\text{kin}}, \hat{V}] \neq 0$. If $\hat{E}_{\text{kin}}$ and $\hat{V}$ commuted, we could replace $\exp[-(\hat{E}_{\text{kin}} + \hat{V})/k_B T]$ by $\exp(-\hat{E}_{\text{kin}}/k_B T)\exp(-\hat{V}/k_B T)$, and by inserting the identity $\hat{1} = \int \mathrm{d}x'|x'\rangle\langle x'|$ we would have solved the problem, since $\langle x|\exp(-\hat{E}_{\text{kin}}/k_B T)|x'\rangle$ amounts to dealing with the (known) quantum mechanical propagator of a free particle. However, by neglecting the noncommutativity of $\hat{E}_{\text{kin}}$ and $\hat{V}$ we would have reduced the problem back to the realm of classical mechanics, all quantum effects would have been lost.

But, a related recipe is provided by the exact Trotter product formula for two non-commuting operators $\hat{A}$ and $\hat{B}$, P being an integer,

$$\exp(\hat{A} + \hat{B}) \longrightarrow [\exp(\hat{A}/P)\exp(\hat{B}/P)]^P \quad \text{for } P \to \infty\,. \tag{31}$$

Using Eq. (31) the partition function can be written as

$$\begin{aligned}Z = \lim_{P\to\infty} &\int \mathrm{d}x_1 \int \mathrm{d}x_2 \cdots\\ &\int \mathrm{d}x_p \langle x_1|\exp(-\hat{E}_{\text{kin}}/k_B TP)\exp(-\hat{V}/k_B TP)|x_2\rangle\langle x_2|\\ &\cdots\langle x_p|\exp(-\hat{E}_{\text{kin}}/k_B TP)\exp(-\hat{V}/k_B TP)|x_1\rangle\,.\end{aligned} \tag{32}$$

Note the matrix elements can be worked out, with the result ($P \to \infty$)

$$Z = \left(\frac{mk_{\mathrm{B}}TP}{2\pi\hbar^2}\right)^{P/2} \int \mathrm{d}x_1 \cdots \mathrm{d}x_p \exp\left\{-\frac{1}{k_{\mathrm{B}}T}\left[\frac{\kappa}{2}\sum_{s=1}^{P}(x_s - x_{s-1})^2 + \frac{1}{P}\sum_{s=1}^{P} V(x_s)\right]\right\} , \tag{33}$$

with the boundary condition $x_{p+1} = x_1$ and the effective spring constant $\kappa = mP(k_{\mathrm{B}}T)^2/\hbar^2$. Equation (33) formally corresponds to a classical partition function of a "ring polymer" will P "monomers". Its gyration radius (which is of order $\hbar/\sqrt{mk_{\mathrm{B}}T}$, i. e. of the same order as the thermal de Broglie wavelength $\lambda_T = h/\sqrt{2\pi mk_{\mathrm{B}}T}$ of the quantum particle) represents the region over which the quantum particle typically is delocalized at temperature T. If effects of quantum statistics are ignored (which is admissible for crystals at low T, apart from solid helium) the generalization of this formalism (Eqs. (32), (33)) to N particles is straightforward. In fact, useful applications to describe low temperature properties of crystals beyond the harmonic approximation have been given [13]. Related approaches also based on the Trotter formula can be developed for various lattice problems, e. g. the Ising model on a d-dimensional lattice in a transverse field can be reduced to an ordinary Ising problem (with no transverse field) on a $(d+1)$-dimensional lattice, but with anisotropic interactions (the extra lattice direction corresponds to the "imaginary time" coordinate s in Eq. (33) along the contour of the ring polymer). But many problems of physical interest, e. g. the Hubbard Hamiltonian describing strongly correlated fermions on a lattice, cannot yet be satisfactorily simulated by such Monte Carlo methods at low enough temperatures because of the "minus sign problem": the distribution to be sampled $\{\rho(\boldsymbol{x})\}$ is not always positive, and hence cannot be interpreted as a probability density suitable for importance sampling. The brute force recipe consists of splitting $\rho(\boldsymbol{x})$ into its sign (S) and its absolute value ($\tilde{\rho} = |\rho(\boldsymbol{x})|$), so that $\rho = S_{\tilde{\rho}}$, and then the quantum average can be formally rewritten as

$$\langle A \rangle = \langle AS \rangle_{\tilde{\rho}} / \langle S \rangle_{\tilde{\rho}} , \tag{34}$$

where $\langle \cdots \rangle_{\tilde{\rho}}$ means averaging with $\tilde{\rho}$ as weight function. However, this brute force approach works only for rather small N, since the average of the sign behaves as $\langle S \rangle_{\tilde{\rho}} \propto \exp(-\text{const.} \times N)$. Alleviating this problem is still an active area of research.

Future Directions

There is still very active research going on to find more efficient algorithms, by a clever design of Monte Carlo moves adapted to specific problems, by exploiting advanced computer architecture (e. g. "parallel tempering" methods exploit parallel architectures by running n real replicas of the system at closely spaced neighboring temperatures or other control parameters in parallel and exchanging from time neighboring temperatures, to allow for a faster relaxation of the system configurations), and by techniques such as "multicanonical Monte Carlo" [1] or Wang–Landau sampling of the energy density of states [13] or related methods of so-called "umbrella sampling". The easy availability of powerful desk-top computers has also facilitated the study of rather complex model systems (unlike the early days of Monte Carlo, where the research had to focus on "toy problems" such as the Ising model, the hard disk and hard sphere fluids, the self-avoiding walk problem, percolation, etc.). While these classic problems still are useful as a testbed for new methodologies of simulation the analysis of simulation output, there is now much emphasis on applications directed towards materials sciences, soft and biological matter, and statistical mechanics far from equilibrium. In this context, also "multiscale simulation" methodology has become a very active field of research: e. g., electronic structure calculation on the subatomic scale is required to yield realistic input for potentials that can then be used for instance in kinetic Monte Carlo simulations. Monte Carlo methods developed in the context of statistical physics are very popular for applications clearly going beyond physics, such as simulations of sociological and economical processes ("econophysics"), biologically motivated models, etc. There also is a continuous and fruitful exchange of know how with the practitioners of other simulation techniques (classical and "ab initio" Molecular Dynamics, Lattice Boltzmann simulations of transport phenomena), unlike in the past where the different "communities" of simulators worked in a rather disjunct manner.

Bibliography

1. Berg BA (2004) Markov Chain Monte Carlo Simulations and Their Statistical Analysis. World Scientific, Singapore
2. Binder K (1981) Critical properties from Monte Carlo coarse graining and renormalization. Phys Rev Lett 47:693–696
3. Binder K (1982) The Monte Carlo calculation of the surface tensions for two- and three-dimensional lattice models. Phys Rev A 25:1699–1709
4. Binder K, Young D (1986) Spin glasses: Experimental facts, theoretical concepts, and open questions. Rev Mod Phys 58:801–976

5. Binder K (ed) (1995) Monte Carlo and Molecular Dynamics Simulations in Polymer Science. Oxford University Press, New York
6. Binder K, Heermann DW (2002) Monte Carlo Simulation in Statistical Physics: An Introduction, 4th edn. Springer, Berlin
7. Ceperley DM (1996) Path integral Monte Carlo methods for fermions. In: Binder K, Ciccotti G (eds) Monte Carlo and Molecular Dynamics of Condensed Matter Systems. Societa Italiana di Fisica, Bologna, pp 445–482
8. Fisher ME (1971) The theory of critical point singularities in critical phenomena. In: Green MS (ed) Proceedings of the 1970 Enrico Fermi International School of Physics, vol 51. Academic Press, New York, pp 1–99
9. Frenkel D, Smit B (2002) Understanding Molecular Simulation: From Algorithms to Applications, 2nd edn. Academic Press, San Diego
10. Hohenberg PC, Halperin BI (1977) Theory of dynamic critical phenomena. Rev Mod Phys 49:435–479
11. James F (1990) A review of pseudorandom number generators. Comp Phys Commun 60:329–344
12. Knuth D (1969) The Art of Computer Programming, vol 2. Addison-Wesley, Reading
13. Landau DP, Binder K (2005) A Guide to Monte Carlo Simulations in Statistical Physics, 2nd edn. Cambridge Univ Press, Cambridge
14. Mascagni M, Srinivasan A (2000) Algorithm 806: SPRNG: A Scalable Library for Pseudorandom Number Generation. ACM Trans Math Softw 26:436–461
15. Montvay I, Münster G (1994) Quantum Fields on the Lattice. Cambridge Univ Press, Cambridge
16. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AM, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21:1087–1092
17. Privman V (ed) (1990) Finite Size Scaling and Numerical Simulation of Statistical Systems. World Scientific, Singapore
18. Sadiq A, Binder K (1984) Dynamics of the formation of two dimensional ordered structures. J Stat Phys 35:517–585
19. Stauffer D, Aharony A (1994) Introduction to Percolation Theory. Taylor and Francis, London
20. Suzuki M (ed) (1982) Quantum Monte Carlo Methods in Condensed Matter Physics. World Scientific, Singapore

Moral Dynamics

Rainer Hegselmann
Institute of Philosophy, Bayreuth University,
Bayreuth, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Computer Tournaments: The Evolution of Cooperation
Replicator Dynamics:
The Evolution of the Social Contract
Indirect Evolutionary Approach: The Evolution of Trust
Future Directions
Bibliography

Glossary

Evolutionary game theory Evolutionary game theory was developed as a transfer of traditional game theory to biological contexts. Payoffs are no longer interpreted as representing preferences over outcomes. In evolutionary game theory payoffs represent fitness. A central idea is that strategies *replicate* according to their performance in terms of fitness. More successful strategies become more frequent, less successful become less frequent. Evolutionary game theory focuses on the success driven frequency dynamics of strategies that are played in a population. The rationality assumptions of traditional game theory are given up.

Game theory Game theory studies strategic interactions among *rational* players. They have preferences over all possible outcomes of their interaction, i. e. the game that is played. Given their preferences all players try to make the best out of the situations, knowing that all others are doing that as well. A central question is: Given the possible strategies and given the preferences, what combination of strategies, one for each player, would be a solution of a game among rational players? John von Neumann, Oskar Morgenstern and John Nash took decisive first steps in the 1940s to develop the theory.

Moral dynamics Moral dynamics refers to the processes by which moral behavior and moral attitudes emerge, evolve, spread, erode or disappear. *Moral attitudes* is broadly conceived and meant to include the whole *internal* side of morality: internalized norms, accepted values, guiding virtues, certain types of moral dispositions or morally transformed preferences, feelings like guild, regret and shame. *Moral behavior* regards the more *external* and at least partially observable side of morality.

The study of moral dynamics normally focuses on very basic problems in human interactions about which almost everybody would say that – from a moral point of view – a certain type of behavior or attitude is preferable: co-operating if there is an incentive for free riding, sharing if there is an incentive to be greedy, reciprocating if there is an incentive not to do so. The essential structure of these 'morally critical' situations can be precisely described and analyzed by means of traditional and evolutionary game theory.

Nash equilibrium The Nash equilibrium is the central

solution concept in game theory. A *Nash equilibrium* is a combination of strategies, one for each player, in which each player's strategy is a best response to the others' strategies. As a consequence, if all others play their equilibrium strategy, then no one has an incentive to deviate unilaterally from the equilibrium strategy.

Replicator dynamics The replicator dynamics is a fundamental concept in evolutionary game theory: Strategies that beat the average success, become more frequent; strategies that perform below average become less frequent. The replicator dynamics can be both, a *biological* process that involves genes, and a *cultural* process that involves imitation.

Social contract The notion of a *social contract* became famous by Hobbes' *Leviathan*: In a state of nature, in which life is nasty, brutish and short, the individuals design and sign a contract to *establish a central enforcement agency* that guarantees societal peace. Nowadays the meaning of *social contract* is often a bit different: What is meant is a set of *fundamental moral arrangements* that make societal life possible altogether – for instance, keeping promises, dividing fair, doing one's part and so on.

Definition of the Subject

Moral dynamics refers to the processes and phenomena (collective or individual) by which moral behavior and moral attitudes emerge, evolve, spread, erode or disappear. *Moral attitudes* are broadly conceived and meant to include the whole *internal* side of morality: internalized norms, accepted values, guiding virtues, certain types of moral dispositions or morally transformed preferences, feelings like guild, regret and shame. *Moral behavior* regards the more *external* and at least partially observable side of morality: Certain types of actions or omissions in certain situations in which, for instance, prescriptions or proscriptions apply.

Especially in large-scale societies with a high frequency of anonymous interactions there are lots of situations with an inherent incentive to cheat, to betray, to be unfair etc. – i. e. to act in a way that almost everybody considers to be immoral. Often external and formal inspection, control and enforcement by central authorities are not possible, too expensive or highly unattractive for different reasons. Morality is a kind of internal, decentralized and informal control. Under certain conditions this type of control may work, when external, formal and centralized control is impossible or undesired. Therefore, the construction and study of models of moral dynamics could give advice or at least some hints how to design societies and institutions.

Introduction

Questions about the origin, status and functioning of morality puzzled already the ancient Greek philosophers. Nowadays problems of morality are addressed by many disciplines, for instance sociology, psychology, economics, and biology.

In the following we will focus on influential approaches that, *firstly*, try to understand moral *dynamics* and, *secondly*, do that by means of constructing *models*, that – compared to informal dynamical drafts – allow a rigorous analysis, be it by means of computer simulations or by paper and pencil methods.

Speaking of moral dynamics requires a sufficiently clear concept of *morality*. One would expect that – after discussions of about 2500 years – moral philosophers have done the conceptual clarification. Unfortunately, they didn't. The question "What is morality?" did not find a single agreed-upon answer. Almost all fundamentals are still questioned and debated, including the conceptual requirements for an action or attitude to be called a moral one.

Trying to avoid taking sides in the philosophical debates, we will resort to the moral common sense. We focus on situations for which almost everybody would say that somehow 'morality is at stake' and that, from a moral point of view, a certain type of behavior or attitude would be preferable. The focal situations regard very basic problems in human interactions: co-operation if there is an incentive for free riding, sharing if there is an incentive to be greedy, reciprocating as promised if there is an incentive not to do so. The essential structure of these 'morally critical' situations can be precisely described and analyzed by means of *game theory*. Games that stylize situations which most of us regard as 'morally critical', are known under names as *prisoner's dilemma, stag hunt, divide the cake, dictator game, ultimatum game, game of trust* – to mention some.

Game theory is the theory of *rational* decision making in situations of mutual strategic interdependence. The decisive concept for a *rational* solution of games is the so-called *Nash equilibrium*, a combination of strategies, one for each player, in which each player's strategy is a best response to the others' strategies. Surprisingly, the rational solution of all the games, that model 'morally critical' situations, is problematic: *First*, in some of the games the rational solution is *inefficient* in the sense that there is an alternative outcome, realized by a combination of strategies *different* from the Nash-equilibrium, that all players

strictly prefer to the equilibrium outcome. *Second*, some other games have more than one Nash equilibrium. As a consequence, equilibrium selection becomes a serious problem. Both effects suggest that morality may come into play where rationality gets into trouble [103]. That conjecture gets further support by the results of thousands of laboratory experiments in which persons had to play 'morally critical' games. The test subjects *either* tend to deviate systematically from the one and only rational solution *or* – if there are many such solutions – actually realize only a small typical subset [22,27,51,60]. Again moral considerations suggest themselves for an explanation.

In the following we present and discuss *three influential approaches* to model and analyze moral dynamics. They all start with certain simple games that quintessentially describe morally critical situations. In very different ways they model and analyze processes by which 'moral solutions' may evolve, be maintained or perish.

The *first* approach regards the *evolution of cooperation* with a focus on the type of analysis pioneered by Robert Axelrod [5,6,7,8,11,12]. Core of that approach were *tournaments* of strategies for iterated prisoner's dilemmas. Despite of all the criticisms that the approach received, and regardless of the fact, that the approach is not a fully developed evolutionary approach, Axelrod, nevertheless, pioneered an evolutionary account of cooperation and other kinds of interactions of similar importance. The *second* approach is based on *evolutionary game theory*, especially *replicator dynamics*, and aims at an explanation of a set of basic moral arrangements of societal life. Evolutionary game theory – explicitly and on purpose – gives up the strong rationality assumptions that are constitutive for traditional game theory. Instead, success depending differential replication of strategies drives a biological or cultural learning process. Brian Skyrms pioneered this approach to explain the *evolution of the social contract*, understood as a set of basic moral arrangements of societal life [94,97]. The *third* approach focuses on the *evolution of trust*, while elaborating a middle ground between traditional and evolutionary game theory: The *indirect evolutionary approach* is a first step to model *explicitly* the evolution of moral attitudes (the *internal* side of morality) that accompany the emergence of moral behavior (the *external* side of morality). The internal process is conceptualized as an endogenous preference change that is driven by differential material success of the behavior based on the preferences. Thus, given their preferences, agents act rationally in the sense of traditional game theory, while their preferences undergo a dynamics in the spirit of evolutionary game theory. Werner Güth, Menachem Yaari, and Hartmut Kliemt pioneered the approach in collaboration [38,39,41].

Moral Dynamics, Table 1
Payoff matrix for a two-person prisoner's dilemma (Temptation > Reward > Punishment > Sucker's payoff)

	Cooperation	Defection
Cooperation	R R	S T
Defection	T S	P P

A concluding chapter discusses future directions and tasks ahead.

Computer Tournaments: The Evolution of Cooperation

Many social situations are a kind of *social trap:* Given the possible actions of the individuals and their incentives, the individuals tend to decide for actions that produce an inefficient, sub-optimal outcome, i. e. there is an alternative outcome that everybody would strictly prefer to the realized one.

The paradigm case for such a situation is the *prisoner's dilemma*. It is described in its normal form by the matrix in Table 1.

The standard prisoner's dilemma (abbreviation: PD) is a two-person game in which both players have to decide whether to cooperate or to defect. Both players prefer mutual cooperation to mutual defection. But they disagree about what is the best and what is the worst outcome: One-sided defection of the row player is the best outcome for the row player and at the same the worst outcome for the column player. Conversely, one-sided defection of the column player is the best for the column player and the worst for the row player. It is constitutive for the game that binding agreements are not possible (though communication may nevertheless be possible). The players decide upon their strategies simultaneously. Under these conditions every player's one and only best reply to whatever the other player might do, is defection. Therefore mutual defection is the only Nash-equilibrium of the game. At the same time, it is the one and only outcome that is *not* Pareto-optimal. Thus, the solution is a disaster. That disaster is *not* due to a lag of rationality rather than a consequence of rationality as worked out in game theory. With respect to contexts in which a PD is played repeatedly, there is often, besides the condition $T > R > P > S$, a second one: $2R > T + S$. The second condition makes sure that the aggregated or average payoffs of two players that coordinately switch between cooperation and defection are lower than the payoffs of continuously cooperating players.

A.W. Tucker was the first to characterize the structure of the PD in 1950. The game can be extended to a version with more than two persons [90]. The name of the game is motivated by the story that is often used to illustrate the PD. In that story two prisoners are suspected of having committed a crime. In the interrogation they face the decision whether to confess or to maintain silence. Already 1957 Duncan Luce and Howard Raiffa notice the considerable attention that the game received among game theorists [68]. That has never changed since then [22,27,33,51,60]. In many fields, especially in social psychology, experimental economics, some parts of sociology, political science and later biology the PD became something very close to what *escherichia coli* is in microbiology (though today the *ultimatum game* [42] receives similar attention). The reason for that attraction is clearly that the PD captures – lucidly and precisely – a situational structure that, firstly, seems to be ubiquitous and, secondly, makes functioning cooperative relationships a riddle: How is cooperation possible at all? How do human or animal beings manage to establish cooperative relationships if there is no policing central authority? Often *reciprocal altruisms* is another wording for cooperation. Then the question is: How is reciprocal altruism possible at all?

From a moral point of view, the PD describes a type of situation in which morality *often* – though *not always* – requires a cooperative choice: Mutual cooperation may have negative external effects for third parties, that are *not* players *in* the game. A price cartel is an instance for such a case, well functioning criminal gangs another one. But *without* negative externalities, the cooperative strategy is normally considered to be the moral choice: doing one's part, and not free-riding. Under that perspective the prisoner's dilemma illustrates and captures a conflict between morality and self-interested rationality. Therefore, understanding the emergence and maintenance of cooperation could *possibly* contribute to an understanding of morality.

In some articles and a book that appeared in the early1980s, Robert Axelrod worked out and made public – and that extremely successful – a new method to study problems of cooperation, namely computer based competitions of strategies for iterated prisoner's dilemmas [5,6,7,8,11,12]. The competition and the results – published in *The evolution of cooperation* [8] – were a kind of event. It made the PD and problems of cooperation well known to an audience reaching far beyond the scientific communities that had done the PD research until then.

Axelrod's 14 participants in a *first* tournament were scientists that were fairly familiar with the PD structure. Many of the participants were leading experts in the field. The participants had to submit a program that basically embodied a rule. The rule selected for each move in a finitely repeated PD that was played against the submitted rule of another participant either the cooperative or the defective strategy. The two rules that were matched for the iterated PD had access to the whole history of their interactions so far. The competition was organized as a round robin tournament, i. e. all rules were matched with each other–including their twin and RANDOM, a rule that with equal probability cooperates and defects. Every participant knew in advance that there were exactly 200 moves. The payoffs in the matrix given by Table 1 were $T = 5$, $R = 3$, $P = 1$ and $S = 0$. A very simple rule (more exactly, a *super game strategy* in the game theoretical terminology) was the winner: TIT FOR TAT (abbreviation: TFT), submitted by Anatol Rapoport. TFT starts with a cooperative move. Thereafter it simply does what the other player has done in the previous move: Thus, TFT reacts with cooperation on cooperation while defection is answered with defection. The results of the first tournament were analyzed and made public, combined with an invitation to participate in a *second* round robin tournament. Different from the first one, there was no finite number of iterations rather than a certain probability that there is no continuation after a move. The *expected* median length of the game was 200.

Much more super game strategies were submitted for the *second* tournament. But again TFT was the winner. It was also often the winner in semi-evolutionary contests in which more successful super game strategies became more and less successful less frequent ("semi-evolutionary" since there was no mutation; Axelrod refers to this type of tournament as *ecological*).

In Axelrods explanation the overwhelming success of TFT is due to *four* decisive properties. *Firstly*, TFT is *nice*: it never defects first. *Secondly*, TFT is *retaliatory*: it reacts on every defection with retaliation in the next move. *Thirdly*, TFT is *forgiving*: if the other player returns to cooperation, TFT will cooperate in the very next move – whatever the number of defection of the other player might have been. *Fourthly*, TFT is *clear*: it is easy to find out that defection will immediately be punished.

Axelrod's approach induced a lot of follow-up tournaments and computational competitions in modified and often extended or more elaborated settings [33,54]: More or less sophisticated agents in terms of memory, errors, learning, and complexity of their strategies were introduced [4,12,20,64,66,83,88]. Special initial populations were analyzed [30,53,55,66,67,68,72,76]. Payoffs were varied [30,61,77,78] or noise was added [13,14,77]. Agents got exit options [91,104].

Axelrod's book *The evolution of cooperation* [8] was a best selling book. There were 1000 quotations of his work

by 1992; for the year 2000 the *Social Science Citation Index* reports already more than 2500 quotations [54]. However, a significant fraction of them appeared in articles that cast doubt on Axelrod's theoretical claims. Ken Binmore published one of the most carping criticisms [18,19]. The latter appeared in 1998 in JASSS, the *Journal of Artificial Societies and Social Simulations*, where it is up to day – as the journal's statistics shows – among the most viewed articles ever published in that journal.

Binmore frankly states his irritation about Axelrod's success – to Binmore's mind a mere hype. The bubble was produced as the combined effect of Axelrod's scientifically unjustified championing of TFT and the additional severe misunderstandings of idolisers and science popularisers. The core of Binmore's criticism is twofold. *Firstly*, the success of TFT is not as robust as Axelrod claims. On the contrary, it is very sensitive to the special composite of other strategies with which it interacts. In certain environments the strategy TIT FOR TAT, a strategy that is *not* nice, is fairly successful [82,83,85,88]. TIT FOR TAT starts with a defection and continues to do so until the other also defects. Then it switches to cooperation and continues to cooperate until the opponent defects. At that point it returns to defection and continues the way it starts. In noisier environments strategies that are more forgiving than TIT FOR TAT may fare better [82]. FRIEDMAN is a nice strategy that never forgives. The strategy starts with cooperation and continues do so until the other defects. After the very first defection of its opponent FRIEDMAN defects forever. Despite of its unforgiveness, FRIEDMAN turns out to be very successful in many settings. *In general*: The evolutionary success of TFT crucially depends upon the environment–and that environment may change. Axelrod's tournaments do not know mutation and variation. New strategies cannot enter the stage and eliminated strategies are gone forever. The justification of far reaching evolutionary claims about niceness or forgiveness asks for much more than such a setting. Necessary are long run simulations that mimic all components of evolution: selection, variation, and mutation.

Under an evolutionary perspective it is important to note that TFT lacks a certain type of stability. It is *not* an evolutionary stable strategy – a key concept of evolutionary game theory and developed in [74]. Let $S_1, S_2, \ldots, S_i, S_j, \ldots, S_n$ be strategies. By $U(S_i, S_j)$ we denote the payoff for strategy S_i if played against S_j. Now we define:

A strategy S_i *is an evolutionary stable strategy* if and only if

(a) $U(S_i, S_i) \geqslant U(S_j, S_i)$, for all $j \neq i$ and

(b) if $U(S_i, S_i) = U(S_j, S_i)$ then $U(S_i, S_j) > U(S_j, S_j)$, for all $j \neq i$.

If (a) holds, then the strategy combination $\langle S_i, S_i \rangle$ is a Nash equilibrium in the associated two person game. Condition (b) makes sure that in a population where everybody else plays strategy S_i, natural selection works against an invading strategy S_j. TFT does not meet condition (b). It can be invaded, for instance by the unconditionally cooperating strategy ALL C. Such an invasion may then prepare the ground for other invaders, for instance, the unconditionally and always defecting strategy ALL D, which at least for a while could prey on ALL C.

A *second* line of criticism regards Axelrod's neglect of analytical insights that already exists [18,19,100,101]: It is well known that – given the probability for another iteration of the game is sufficiently high – equilibria of super game strategies exist in which the players consistently cooperate in all basic games. This insight is a corollary of a fairly general theorem, the *folk theorem*. It got its name since several people simultaneously proved it in the 1950s. – The second line of criticism is less convincing than the first one. It presupposes that Axelrod was trying to find out by simulation what *rational* players – *rational* understood in a game theoretical sense – would do. If, instead, Axelrod intended to find and study heuristics for *boundedly* rational behavior – and that seems to be the case – then the second criticism misses the target.

By his computer tournaments Axelrod pioneered an evolutionary account of cooperation. Additionally he pioneered in *The evolution of cooperation* the study of *spatially structured* problems of cooperation: Assumed is a rectangular grid, such that agents interact only within their (overlapping) local neighborhoods, i. e. the neighbors in the north, south, east, and west. Such a framework – it is a kind of cellular automaton – allows analyzing the spatial dynamics of cooperation. Questions like "How does local learning affect rise or decline of cooperation?" or "What are the chances for small clusters to grow?" can be addressed.

Lots of others took up the spatial approach to cooperation [34,56,61,81,84,86,89]. Very natural further steps are the use of *other* regular or irregular network structures [29] or the *endogenization* of the network.

An example for the latter is the study of the *evolution of support networks* [49,50]. The general idea is to model support relationships by a *two person support game*, played with one another by neighboring agents, living on a torus with a rectangular grid and lots of empty cells. Agents can move and look for attractive partners. They play the game simultaneously with all their direct neighbors.

The support game is an *asymmetric* PD with a *specific* structure: Both players become needy with certain probabilities p_1 and p_2 – equal or unequal. It is assumed that individuals belong to nine different risk classes which become needy with probabilities 0.1, 0.2, . . . , 0.9. The probabilities remain constant throughout the iterated game. The first move in the constitutive support game is a chance move. According to the probabilities p_1 and p_2, chance decides, whether both, only one player, or no one needs help. A player in need of help can't help anyone. In case the other one needs help, a player, who does not need help, has to decide whether to help or not. Helping is costly. Getting help is a benefit. In this game it is a dominant strategy *not* to help. That remains true even when the mutual support is profitable for both players in terms of expected payoffs. This inefficient solution makes the support game to a variant of the prisoners' dilemma.

The model was constructed to answer the question: What about networks of mutual support in a world:

- which is exclusively inhabited by more or less *rational egoists*,
- who are unequal in that they need help with *different* probabilities,
- must *choose* their partners themselves,
- and will choose those partners in *opportunistic ways*?

In the model all agents are assumed to know their own and others' probabilities to become needy. Costs, benefits and probabilities to get a migration option in a period are common knowledge. Given their own risk class, the agents know about feasible best and worst neighborhoods and form an aspiration level in terms of a 'networking dividend'.

All the time agents make their guesses about continuation probabilities for their actual neighboring relations. Based on that they decide rationally, whether to engage in mutual support or not: If, given the risk classes of the players, costs, benefits and assumed continuation probabilities, combinations of TFT- or FRIEDMAN-strategies are an equilibrium, then the players engage in mutual support with a neighbor, otherwise not.

Figure 2 shows the evolved network of mutual support that started with the primordial soup shown in Fig. 1. For an interpretation: Black cells are *empty*. Different colors represent different risk classes according to the legend besides the figures. Risk class 1 (dark blue) becomes needy with a probability of 0.1, risk class 2 (light blue) with a probability of 0.2 and so on. Short white lines connecting two individuals indicate *functioning support relations* between them. Round cells either do not meet their aspiration level *or* don't have any chance of finding partners for mutual support under the given conditions (costs, benefits, chance to get a migration option). Filled cells meet their aspirations.

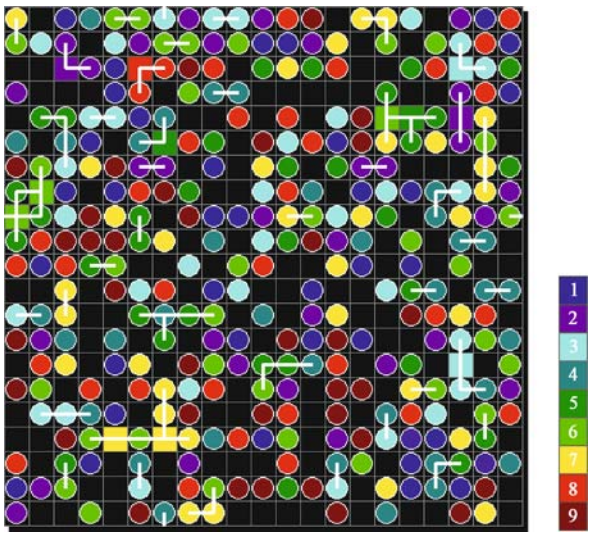

Moral Dynamics, Figure 1
Primordial soup – a random initial distribution

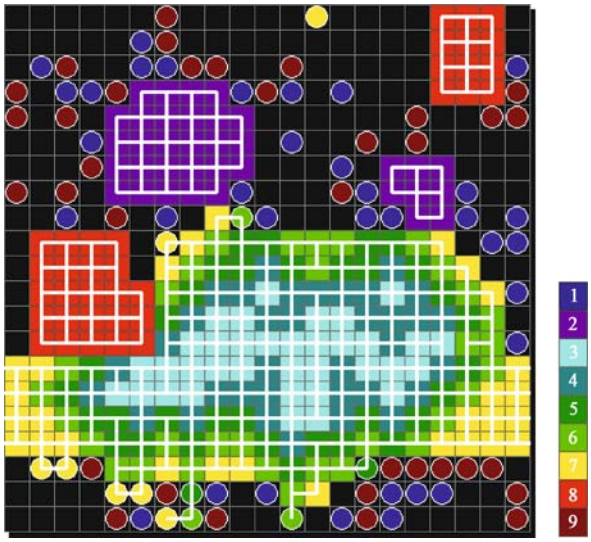

Moral Dynamics, Figure 2
Evolved network of mutual support after 2000 periods

Obviously, support networks *can* evolve even in a world of rational egoists, differently endowed by nature and choosing their partners opportunistically. However these networks will be characterized by some *class segregation*. Other results – though not visible in the two figures – are: For intermediate classes it is comparatively *easier*, to find partners for mutual support. In terms of 'network-

ing dividend' middle classes are the winners. The wider the range of classes whose members could establish functioning support relations among each other (this range increases as the probability for getting a migration option decreases) *the higher the portion of agents ending up with less than what they aspire*. With the evolution of support networks and despite the accompanying class segregation, the *equality* of the payoff distribution *increases*. One might summarize the situation a bit provocatively: rational egoism, class segregation, increasing equality and increasing wealth go hand in hand.

Looking back at all the findings about cooperation described above from a moral point of view, one might get *mixed feelings*: Lots of problems considered to be moral problems, are problems of cooperation. The PD describes the very essence of the problem. In the not repeated one shot game rationality requires defection. That changes if the game is repeated: *If* the shadow of the future [8] is threatening enough, i. e. *if* the continuation probability for the interaction is sufficiently high, *then* ongoing cooperation is possible among rational players. But is that what morality is considered to be? Or is it enlightened self-interest only?

For an answer one has to note, that among rational agents the ongoing cooperation is based on an equilibrium of *conditional* and *retaliating* super-game strategies (while a strategy combination ALL C versus ALL C is no equilibrium at all). One such conditional and retaliating super game strategy is the much-celebrated TFT. But from a moral point of view, TFT does not have the best press [65]. It looks like "An eye for an eye, a tooth for a tooth". As a principle of direct reciprocity it is not very high in Kohlberg's moral stage hierarchy [63]. Direct reciprocity is very different from the golden rule "Do to others what you would have them do to you!" – often considered as the very essence of morality. For a super game strategy like FRIEDMAN that answers the first defection with eternal punishment, the moral evaluation would be even worse.

One could try to defend TFT as a strategy of *conditional morality*: If a player cooperates in a constitutive game, then one could say that he *suspends* self-interested rationality and follows requirements of morality. A TFT-player is willing to do that, but only under the condition that others do that as well. (Obviously, other super-game strategies, for instance FRIEDMAN, could get a similar interpretation.) But then again morality seems to be based on enlightened self-interest and would not reach farther than that: Often the actual continuation probability of an interaction is too small for TFT played against TFT being an equilibrium. Is it then morally right to defect and, for instance, not to help the needy other? Or aren't the players morally obliged to cooperate nevertheless? One might answer that – by and large – *real world morality* works only *if* continuation probabilities are sufficiently high. But that would also suggest that morality sometimes asks for more than that what rational agents would do and real agents sometimes really do. If that were true, then neither the motivational nor the behavioral side of morality seems to be fully captured.

Replicator Dynamics: The Evolution of the Social Contract

Starting in the early 1990ies a series of articles and books pioneered the use of *evolutionary game theory* to explain certain phenomena in which morality seems to play a role [1,2,3,48,92,93,94,95,96,97,98,107,108]. Brian Skyrms' book *Evolution of the Social contract* [94] made the approach widely known. However, the title is a bit irritating. Most people would think of the social contract as a sort of contract that *does not evolve*. Rather, *rational* decision makers would explicitly design and sign a contract to establish a *central enforcement agency* – as in Hobbes' state of nature [57]. But Skyrms' project is different: He works out how justice and fairness *evolve* among learning agents that do *not* have the cognitive capacities and do *not* meet the strong rationality and common knowledge assumptions as assumed in standard game theory. Skyrms' social contract regards the evolution of what Hume calls *artificial virtues* [58,59]: A kind of moral dispositions, invented by humans (and in that sense 'artificial'), acquired by some sort of character transformation and maintained by the practice of mutual approval and disapproval. Thus, the *social contract*, as understood by Skyrms, is *neither* an explicit contract, *nor* is it about establishing a central enforcement agency. What is meant is a set of *fundamental moral arrangements* that make societal life possible altogether – for instance, dividing fair, doing one's part and so on.

Hume gave an evolutionary account of these fundamental moral arrangements [58,59], but it was a draft. In a still informal and qualitative manner, modern Humean philosophers and social scientists incorporated important pieces of modern science into Hume's picture [32,70,99]. Skyrms' book *Evolution of the social contract* goes an important step further. It amounts to an explanation of fundamental moral arrangements by means of *models and simulations based on evolutionary game theory* [46,47,74,75,105]. Simple games precisely characterize certain 'morally critical' situations that humans face since ever. Then evolutionary game theory is used to model and to

analyze the evolutionary dynamics of such games: Which strategies spread, which strategies go extinct? What are the patterns and structures of the dynamics? Do strategies become predominant? Do distributions and structures evolve that resemble the social contract, as we actually know it?

One of the simple games that are starting points for a rigorous evolutionary analysis is the game of *divide-the-cake* [93]. The cake in that game is a surplus or windfall. No prior and eventually unequal effort is involved. Both players demand some portion between zero and one. If the two claims total to more than one, then the cake is gone – a referee eats the cake. If the claims total to not more than one, then both players get their claims. If the claims do not add up to one, the referee gets the difference.

In experiments almost everybody facing the problem of dividing the cake will claim half the cake [22,27,51,60]. Both claiming a half *is* a strict Nash equilibrium. It looks *just* and *fair* to almost everybody. However, *all* pairs of claims that sum up to one are strict Nash equilibria. Therefore, why just the equal split? The approach under discussion tries to give an answer based on concepts and notions as developed by evolutionary game theory [46,47,74,75,105].

The fundamental idea is that strategies *replicate* according to their performance in terms of payoffs, now interpreted as fitness. Strategies that beat the average become more frequent, strategies that perform below average become less frequent. There are *two* important points to note: *Firstly*, the evolutionary mechanism sketched above allows for both, a *biological* interpretation that involves genes, and a *cultural* interpretation that involves learning by imitating more successful others [21,102,105]. *Secondly*, different from traditional game theory, the mechanism *does not* imply anything like anticipation and considerations what others might do [95].

Claims in the divide-the-cake-game are strategies. Once the intuitive ideas about the success driven dynamics of strategy frequencies are made precise, the dynamics of such strategies can rigorously be studied.

Let the strategies $S_1, S_2, \ldots, S_i, S_j, \ldots, S_n$ be n possible strategies. x_i is the frequency of strategy S_i . The composition of the population is given by the vector

$$\vec{x} = (x_1, x_2, \ldots, x_n) \quad \text{with} \sum_{i=1}^{n} x_i = 1 \,.$$

The $n \times n$ matrix $A = \left[a_{ij}\right]$ gives the payoff for strategy S_i in an encounter with strategy S_j. We assume *random* encounters and an *infinite* population. Under these conditions the average payoff for strategy S_i is the *expected* payoff

$$f_i\left(\vec{x}\right) = \sum_{j=1}^{n} x_j a_{ij} \,.$$

The average payoff of the whole population is

$$\phi\left(\vec{x}\right) = \sum_{j=1}^{n} x_j f_j\left(\vec{x}\right) \,.$$

If we now assume that the *rate of increase of the frequency* x_i *of strategy* S_i – in formal terms: the time derivative of x_i – is directly proportional to, *firstly*, the current frequency x_i and, *secondly*, the difference between the average payoff for strategy S_i and the average payoff in the population, then we are led to the *replicator equations*

$$\dot{x}_i = x_i\left[f_i\left(\vec{x}\right) - \phi\left(\vec{x}\right)\right] \,.$$

The differential equations describe what is called the *replicator dynamics*. It is a deterministic dynamics of *selection*; mutation is *not* involved. The replicator equations were introduced by Peter D. Tayler and Leo B. Jonker 1978 in [102]. The equations are analyzed in detail in [105].

Figure 3 shows the evolutionary dynamics for a simplified version of the divide-the-cake game. Only three strategies exist: S_1 = demand 1/3; S_2 = demand 2/3; S_3 = demand 1/2. The vertices of the simplex are the points where one strategy has taken over the whole population. Edges are sets of points where at least one strategy has become extinct. All points inside the simplex represent positive frequencies (which always total to one). Arrows indicate the direction of the dynamics. The dynamics has an unstable equilibrium in which half of the population

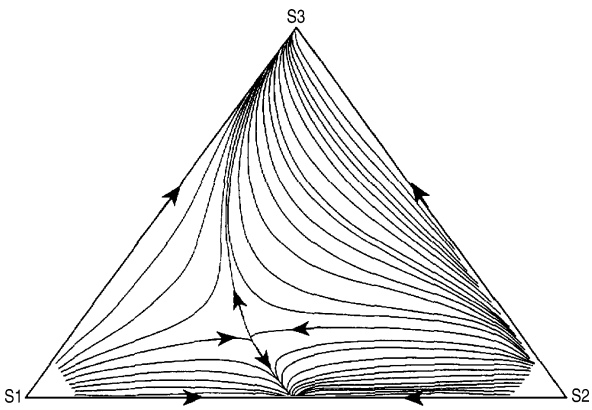

Moral Dynamics, Figure 3
Dynamics of the *divide-the-cake* game with three strategies. Reprinted from [94] with kind permission of Cambridge UP

plays S_1, one third S_2 and a sixths S_3. There is an attraction towards half the population using strategy S1 and the other half using S_2. But there is also a major basin of attraction toward universality of equal division.

The model for the dynamics of sharing can be modified and extended in a way that it covers fundamental features of social reality: The population may be *finite* or the cake may have a more or less *granular* structure (what reduces the number of possible strategies). Random encounters are probably a rare event. More often encounters are *correlated*. That can be covered by strategy dependent probabilities for the pairing of strategies. As a consequence, like-minded people may meet more frequently. Furthermore, the model can get underlying network structures of all sorts [2]. *Signaling* can be added.

The evolutionary dynamics of *other simple games* was analyzed in the same style, above all the PD, the stag-hunt game and the ultimatum game. The *stag-hunt* game (also known as *assurance game*) is again a stylized cooperation problem–though less severe than the PD. The difference is that both players now prefer mutual cooperation to their own one-sided defection. As a consequence the game has *two* Nash-equilibria, mutual cooperation *and* mutual defection. The problem is to coordinate on one of them. The *ultimatum game* [42] is about dividing a German Mark (or a dollar etc.), which is supposed to be a windfall. The two players have different roles in the game: Player one makes a proposal how to divide. Player two decides whether to accept or to reject the proposal. In case of rejection nobody gets anything at all. Given the players utilities are linear in money, there is a unique Nash equilibrium: The proposer offers the least possible amount, which player two is then glad to accept since that amount is more than nothing. However, there is overwhelming evidence [22,27,51,60] from all over the world that real players behave very different: The offers made by player one are normally below but remarkably close to 50% – and if not, then player one runs a high risk of rejection.

The evolutionary analysis of all these simple games and their extensions – especially adding correlated matching, underlying networks, and signaling – leads to the general result that there is often a profusion of possible equilibrium outcomes. Their dynamic stability properties differ. But among them are often some that resemble very much the moral arrangements that seem to be the 'operating system' of huge parts of our societies. Additionally, these equilibrium outcomes often have a remarkably huge basin of attraction – at least under some structure, correlated encounters and signaling [2,96].

The approach based on evolutionary game theory was often appreciated as an enlightening new perspective at the social contract. Nevertheless, the approach received a lot of criticisms as well. The robustness of the results was questioned since they are sensitive to the learning mechanism [28]. But it seems possible to obtain robust results if population structures as they existed among our ancestors are simulated. Another criticism stresses that coalition formation is not taken serious enough [62].

The most severe criticisms of the approach argue that evolutionary game theory is *inherently inadequate to conceptualize morality in a satisfying way*. According to D'Arms [26] the approach fails to explain morality since essential ingredients of morality are left out: Internal sanctions, i. e. negative self-directed feelings, and external punitive or emotional sanctions imposed by others. As a consequence one cannot distinguish *between* explanations of behavior *and* explanations of moral norms that require that behavior. Kitcher [62] makes a similar point when stating that at best the simulations on the divide-the-cake game reveal why we have the propensity to conform to arrangements, which we label as just. What they do not explain is the origin of our conception of justice as such. As a consequence – so the criticism goes – the explanation does not account for the distinction of situations where others do not do their part, from other situations where they do something we dislike. With a focus on the ultimatum game Bicchieri [15] criticizes that decisive ingredients are left out: As experimental results show, there is no unique norm of fairness. There are several and they are conditional upon the context. For instance, it makes a difference whether or not the proposer is thought of as a seller in a market, whether or not the proposer and the receiver are supposed to distribute the product of an exhausting joint venture. Norms, contexts and situational clues, mutual expectations, preferences for conformity given sufficiently many other conform as well, all that plays a decisive role in morality as we know it. But the evolutionary dynamics doesn't account for any of these essential details. *To sum up* this type of objections: Evolutionary game theory is conceptually too poor to conceptualize morality, as we know it. Alexander, after having written a book length elaboration of and contribution to this kind of evolutionary approach, seems to concede that the criticisms are right to a high degree [2].

For a discussion of that objection it has to be clear, that morality regards a very rich set of phenomena: The *explanandum* includes norms, values and their mental representation, more and less general principles, certain intentions and motivations (for instance, acting out of a sense of duty), all sorts of approval and disapproval to sanction behavior, intentions and motivations, special types of reasons and arguments to justify (including apologizing)

one's own or to criticize others' actions (or vice versa), then furthermore, feelings and emotions like regret, shame or bad consciousness, personal moral ideals … – a long list, but for sure not complete. For all the elements enlisted one would like to know how they function, how they work together, and how they evolved.

It is obvious that evolutionary game theory does *not* give an explanation of the *rich explanandum*. The approach explains the evolution of certain *strategies*. In the differential equations that define the replicator dynamics is nothing that directly corresponds to something like motivation, intentions, emotions, deliberations, representations, cognitive structures, discussions, approval and reproach – all that *is not explicitly* modeled.

Not being explicit about the underlying processes is, trivially, a disadvantage *if* explicitness is the only goal. But there may be other goals: Given there are good reasons to assume that a certain *overall effect* is the standard result generated by a bunch of underlying processes in a target system, then it can make a lot of sense to forget about the underlying details and *to generate directly* the overall effect in a model. Doing that may allow one *to analyze in detail other aspects* of the system, for instance dynamical structures caused by the overall effect – a scientific goal that otherwise may be out of reach.

With regard to the social contract, i. e. fundamental moral arrangements of societal life, an assumed recurrent and macroscopic overall effect is the following: As the result of complicated intra- and inter-agent processes, behavior that performs above average becomes more frequent; behavior that performs below average becomes less frequent. The replicator dynamics equations directly produce that effect *without* any deciphering of the underlying processes. All the details that produce the effect are *not* in the model. Rather, they are part of the '*story*' that *accompanies* the model. An accompanying narrative hints to underlying micro layer processes, summarily reports what is known, and accumulates evidence and arguments for the hypothesized overall effect. Therefore, macro models without an explicit micro foundation may nevertheless have a solid scientific foundation. And if so – as in the case of the replicator dynamics – then they are *not* 'just so stories' based on fantasies about causal relations. It is only fair to say that the micro layer processes involved in the emergence and maintenance of fundamental moral arrangements, are not well understood. At the same time, it is interesting to understand macro properties of moral arrangements and to analyze, for instance, stability problems.

In such an epistemic situation, there are at least *two* reasonable scientific approaches: *Firstly*, one can try to get a grasp of the micro layer, the *micro-strategy*. Once the micro layer is sufficiently well understood, one starts working on the macro phenomena. *Secondly*, one tackles macro phenomena by models that are to a large extend independent of a detailed understanding of the micro layer, the *macro-strategy*. The micro-strategy runs the risk of never to get that far. The macro-strategy, on the other side, may be inherently blind for important phenomena and inapt to distinguish where distinctions should be made – even on the macro layer.

Under such a perspective criticisms of D'Arms, Kitcher, Bicchieri, and Alexander are best understood as pointing to *inherent blind spots* if the replicator dynamics is applied to moral arrangements: For instance, many conceptions of morality – moral common sense included – make a distinction between mere behavioral conformity with moral standards ('moral conformity') and behavior for the sake of a moral standard ('moral action'). Not only in the tradition of Kant the two cases *differ from a moral point of view*. (How much that differs varies from conception to conception; for a utilitarian it differs less than for a Kantian.) The critical point is that the replicator dynamics as used by Skyrms *does not distinguish* the two cases explicitly *in the model*. Of course, it is easy to distinguish the two cases in the accompanying narrative. But that is different from explicitly modeling it. As developed *so far*, the approach focuses on strategies; motivations that bring them about are left out. Obviously the approach based on evolutionary game theory and replicator dynamics isn't the full theory. But Skyrms, who pioneered the approach, did not claim so either. His claims are modest: Perhaps the *beginning* of an explanation of our concept of justice, *not* an attempt to present the full theory of the evolution of the social contract [94].

Indirect Evolutionary Approach: The Evolution of Trust

The *indirect evolutionary approach* tries to model moral attitudes (the *internal* side of morality) that, then, motivate moral behavior (the *external* side of morality). The internal process is conceptualized as an endogenous preference change, driven by differential material success of the behavior based on the preferences [36,39,41]. A forerunner of such an approach is Hirshleifer [52] and Frank [31].

The central idea of the indirect evolutionary approach can be applied to all games that are 'morally critical'. The starting point for the actual elaboration of the approach was the *trust problem*, a two-person social dilemma: The predicament is that one player has to move first without any guarantee that the other player later reciprocates.

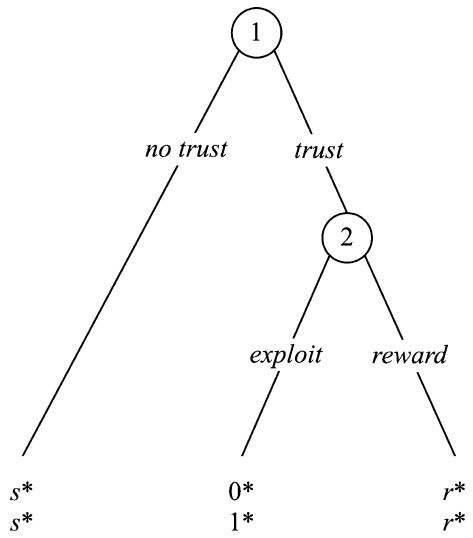

$1^* > r^* > s^* > 0^*$

Moral Dynamics, Figure 4
Trust predicament. The upper entries regard player 1, the lower entries regard player 2

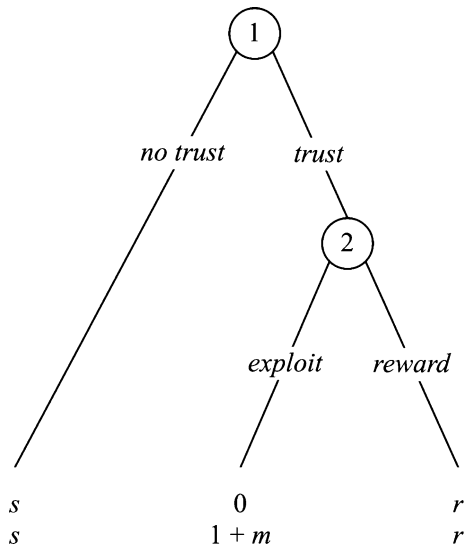

$1 > r > s > 0$

Moral Dynamics, Figure 5
The trust game. The upper entries are utilities of player 1, the lower entries are utilities of player 2. *m* is the moral and purely motivational payoff component

Such an interaction structure plays a central role already in Hume's analysis of trust and promises. In Fig. 4 a tree characterizes the situation.

If the starred entries at the ends of the branches *were* subjective utilities of rational players, the *rational* solution of the game can easily be found: At his decision node the second mover will decide to exploit since $1^* > r^*$ (the lower entries are the entries for the second mover). We assume that the tree and the payoffs are common knowledge. Therefore the first mover can foresee at his decision node that the second mover will exploit. As a consequence, the first mover's alternative is either not to trust or being exploited. Since $s^* > 0$ the first mover decides not to trust and both players end up with the payoff s^*. That result is *inefficient*: If the first mover trusts and the second mover rewards, both would receive r^* with $r^* > s^*$. But rational choice of both players does not allow getting to that efficient solution – obviously a serious predicament.

A decisive *first* step of the indirect evolutionary approach is a *different* interpretation of the payoffs in Fig. 4: The starred payoffs are *not* meant as subjective utilities rather than a sort of *objective or material payoffs that measure success or fitness*. They are linked to an evolutionary process. A function translates the objective payoffs into motivating subjective utilities. To keep it simple, it is assumed that the subjective utilities that correspond $1^*, r^*, s^*, 0^*$ are $1, r, s, 0$.

The decisive *second* step is to introduce additionally a *purely motivational* payoff component m as given in the game tree in Fig. 5. Parameter m is *endogenous*. As a component of the subjective utilities it affects choices. But – and that is important to note – m does *not* measure fitness or success. However, since choices influence the objective payoffs, m *indirectly* affects fitness or success.

In a *third* step, the indirect evolutionary approach assumes all players to be *rational*. Applying equilibrium analysis to the game specified in Fig. 5, we get as the (subgame perfect) rational solutions:

(a) For $m > (r - 1)$: The first mover does not trust; the second mover exploits.
(b) For $m < (r - 1)$: The first mover trusts; the second mover rewards.

The precise value of m is behaviorally *irrelevant*: All players for whom the same inequality between m and $(r - 1)$ applies behave *all in the same way*. Players who do not exploit the first movers' trust are referred to as the *trustworthy* $\underline{m}$-types. Players who do exploit are the untrustworthy $\overline{m}$-types.

Then, as step *four*, evolution comes into play: In a large or even infinite population individuals are randomly paired to play games of trust. The type composition of the population is given by a parameter p, the fraction

of *trustworthy* $\underline{m}$-types. After a random matching, the two roles – being either the first mover or the second mover – are assigned in each pair with a type-independent probability of one half.

In such a setting type-*detection* and type-*discrimination* is crucial: Under the assumption of *complete* type information the first mover knows the second mover's type. As a rational player the first mover will trust if the second mover is a $\underline{m}$-type; is the second mover an untrustworthy $\overline{m}$-type, then the first mover does not trust. If we calculate the objective, material success or fitness for both types, it turns out that – whatever the composition of the population may be – the *trustworthy* $\underline{m}$-types fare better than the untrustworthy $\overline{m}$-type. Without any explicit specification of the dynamics of the population composition, it is clear: Under complete type information the dynamics leads to a globally stable monomorphic population in which all are trustworthy $\underline{m}$-types. The same type of evolutionary analysis shows that if type information is *private* only, i. e. the players only know their own type, the population share p of the trustworthy $\underline{m}$-types is bound to decline towards s/r. Below that threshold there is no evolutionary pressure to drive out the trustworthy since no first mover will trust at all. As a consequence trust can't be exploited.

Both, complete type information and only private type information, are extreme cases. More interesting is the *middle ground* in between as analyzed in [37,38]: Assumed is a certain $\langle C, \mu \rangle$ technology, which reveals at cost C, $(C > 0)$ with probability μ, $(0.5 < \mu < 1)$ the true type of the second mover. An evolutionary analysis of a trust game in which such a technology is available, shows: If the costs of the technology are sufficiently low and the reliability sufficiently high, then there exists a certain *interval* of population compositions p around $p = s/r$ in which rational players in the first mover role will apply the technology. As a consequence the population share p of the trustworthy increases. To the *right* of that interval the share of the trustworthy is so high that it is not worthwhile to spend the cost to detect the untrustworthy. Therefore *all* first movers trust. As a consequence, in that region of p the untrustworthy fare better than the trustworthy and their share goes down. On the left side of the interval, there are so few trustworthy that it is not worthwhile to seek them out. *No* first mover will trust. If occasional mistakes are possible and, therefore, a first mover from time to time mistakenly trusts, then in this region of p the population share of the trustworthy will tend to decline towards zero.

The indirect evolutionary approach is very flexible and applicable to enriched and modified settings in which, for instance, institutions exist or emerge or ex-post punishment is possible [37,40,44,45].

There are *five* important characteristics of the indirect evolutionary approach:

(a) The approach does *not* give up rational choice altogether. In each single round of play the individuals are assumed to be *rational* in the usual sense. They anticipate the consequences, evaluate them and make their choices in a case-by-case manner. Different from an approach based on the replicator dynamics, the strategies are *not* hard-wired rather than result of rational choices based on the preferences that the individuals actually have.

(b) Different from standard rational choice theory, preferences are *not* given rather than subject to change. That allows modeling the emergence of morality as a preference transformation process. That idea has a longer tradition [31,52]. In the indirect evolutionary approach the moral transformation of preferences spreads if the transformation proves successful in *non*-moral terms: The population share of the 'moralized' individuals increases if and only if in terms of objective, material payoffs they fare better in their interactions than the non moralized individuals. If *past experience* proofs morality to be more successful in non-moral terms, then and only then moralized preferences and moral choices based on them can spread.

(c) As a consequence of (a) and (b) the indirect evolutionary approach *integrates a forward, backward, and sideward looking component in human decision-making*: Since the individual makes rational choices in each round of play they are forward looking. The evolutionary process, then, 'compares' past success and is thereby at the same time backward and sideward looking.

(d) The indirect evolutionary approach offers a framework in which subjective motivational factors like norms, internalization of norms, values, moral dispositions and attitudes become an *explicit* component *in* the model. It is a *simple and summary component*: A subjective preference component m is added to a certain payoff. However, that is much more than being just mentioned in an accompanying story that interprets a replicator dynamics.

(e) The approach distinguishes a *deep structure* and a *surface structure*. The deep structure regards the objective and material payoffs. The surface structure regards motivational factors. The motivational factor allows – on the surface level – to account for a kind of *intrinsic* moral motivation that *is* real and *not* a self-deception. At the same time (and very much Humean in spirit) that intrinsic moral motivation has on a deeper level

an evolutionary history: It can evolve if and only if it is worth while to have such an intrinsic motivation – measured in a *hard* currency, i. e. in terms of material and objective payoffs.

Future Directions

All approaches described and discussed above, *do not model explicitly internal cognitive processes within individuals*. Only accompanying interpretations hint to a rich set of processes that are involved: normative deliberations of all sorts, considerations about goals, norms, values, weighing pros and cons, certain emotions and emotional reactions etc. This low resolution of the cognitive structures, processes or capacities is in a sharp contrast with what is going on in certain fields of agent based modeling, agent theory and agent technology [33]. There, as a matter of fact, agents with increasingly richer and stronger cognitive structures and capacities were developed over the last two decades. Partially that is due to research programs and projects that directly aim at or necessarily need a high resolution of the cognitive structure. Such programs and projects can be found in cognitive psychology, artificial intelligence, and many fields where a technical use of agents is intended. As a consequence, agent based models with a – compared to the approaches in the chapters above – rich internal cognitive structure can now be found all over the fields of agent based modeling, multi agent systems, and social simulations in general. Ideas and requirements for agent architectures is a much-debated field. Lots of proposals for how to construct agents were made. Agent architectures, that explicitly account for norms, values, virtues, moral emotions, deliberations etc. were required, drafted and partially realized [23,24,25,79].

To avoid confusion, one should note here: *First*, agent-based modeling *as such* does not necessarily require agents with high cognitive capacities and an explicitly modeled, rich cognitive structure. For instance, a cellular automaton as applied to social processes, *is* an agent-based model. The assumed cognitive capacity may consist in not more than being able to imitate in the next period what a majority of neighbors does in the actual period. Additionally, such a capacity is normally simply assumed and *not* explicitly modeled. *Second*, low resolution with regard to the agents' minds and cognitive structures, does *not* necessarily imply that the individuals in the model are simple minded: If, for instance, agents in a rational choice-based model are assumed to play their equilibrium strategies (found out in the model by checking in just one line of code whether or not a certain inequation holds, which afore was analytically worked out by the model *builder*), then that may – on the side of the agents – amount to the solution of a cognitive task that normally is over-demanding for human beings. *Third*, a comparatively high resolution of cognitive processes and structures does *not* necessarily imply high cognitive capacities of the constructed agents. A high resolution is difficult to model. As a consequence, often the tasks that the agents have to solve are comparatively easy. *To sum up*, the two dimensions 'cognitive resolution' and 'cognitive capacities' are independent – and agent based modeling *as such* is compatible with all degrees of resolution and capacities.

One might question whether dynamical models of population shares are agent-based models at all. Howsoever, for a better understanding of the complicated interplay of intra- and inter-agent processes of moral dynamics it looks necessary and promising to *cautiously* increase the complexity of the agents' minds. Suitably constructed agents could, for instance, allow differentiating between behavioral norm conformity and actions motivated by a norm for the sake of a norm. The so far non-existing or only crude approaches to the *internal* side of morality could become much more refined. Such a refinement implies more parameters, higher complexity and endangers tractability. But it may well be the case that the external side of morality can't be understood sufficiently well without going into some details of cognitive structures and processes. At the moment there is a huge gap between macroscopic models of moral dynamics and the known variety of microscopic processes that seem to generate certain assumed overall effects. There isn't any hope to fill and close that gap in the near future – if ever. But bridges can be built. Such bridges are models that explicitly model and then study, for instance, punishment, detection, reputation, matching, norm recognition, norm internalization etc. that 'somehow' bring about macroscopic effects like changes in the population shares of certain strategies or types of players. Steps into that direction have been taken since the study of moral dynamics started in the 1980s [9,16,80]. Much more will follow.

Research into that direction is the precondition to master the probably most demanding explanatory task in the study of moral dynamics: How did homo sapiens, who used to live together in small groups, manage to learn living together in large-scale societies, in which a high proportion of interactions are no longer based on family ties or good personal acquaintanceship – and nevertheless, fairly often cooperation, fairness, trust, etc. prevail? That question has a *long scientific tradition*. Already the ancient Greeks were puzzling over that problem. In one of Plato's dialogs the sophist *Protagoras* gives a very modern answer which – after some deciphering of the myth in which it

is couched (a myth about Prometheus and Epimetheus) – amounts to saying: A high blood toll was paid to learn the lessons, but finally mankind *invented* both, *moral virtues* and *enforcement agencies* [87]. About 2000 later *David Hume* gave a very similar answer – no longer presented by telling a myth, though still a draft [58,59]. Additionally he stresses the importance of division of labor and mentions reputation mechanisms. Some first steps are done to develop models that could contribute to an answer [2,16,17,18,70,94,106]. But a well-elaborated and sufficiently precise answer is still missing.

Bibliography

Primary Literature

1. Alexander JM (2000) Evolutionary explanations of distributive justice. Philos Sci 67:490–516
2. Alexander JM (2007) The structural evolution of morality. Cambridge UP, Cambridge
3. Alexander JM, Skyrms B (1990) Bargaining with neighbors: Is justice contagious? J Philos 96:588–598
4. Andersen ES (1994) Evolutionary Economics. Pinter, London
5. Axelrod R (1980) Effective choice in the prisoner's dilemma. J Confl Resolut 24:3–25
6. Axelrod R (1980) More effective choice in the prisoner's dilemma. J Confl Resolut 24:379–403
7. Axelrod R (1981) The emergence of cooperation among egoists. Am Polit Sci Rev 75:306–318
8. Axelrod R (1984) The evolution of cooperation. Basic Books, New York
9. Axelrod R (1986) An evolutionary approach to norms. Am Polit Sci Rev 80:1095–1111
10. Axelrod R (1987) Evolution of strategies in the iterated prisoner's dilemma. In: Davis L (ed) Genetic algorithms and simulated annealing. Research notes in artificial intelligence, vol 24. Pitman, London, pp 32–41
11. Axelrod R (1997) The complexity of cooperation. Agent-based models of competition and collaboration. Princeton UP, Princeton
12. Axelrod R, Hamilton WD (1981) The evolution of cooperation. Science 211:1390–1396
13. Bendor J (1993) Uncertainty and the evolution of cooperation. J Confl Resolut 37:709–734
14. Bendor J, Kramer RM, Stout S (1991) When in doubt cooperation in a noisy prisoner's dilemma. J Confl Resolut 35:691–719
15. Bicchieri C (1999) Local fairness. Philos Phenomenolog Res 59:229–236
16. Bicchieri C (2006) The grammar of society. The nature and dynamics of social norms. Cambridge UP, New York
17. Bicchieri C, Duffy J, Tolle G (2004) Trust among strangers. Philos Sci 71:286–319
18. Binmore K (1994) Playing Fair. Game theory and the social contract I. MIT Press, Cambridge
19. Binmore K (1998) The complexity of cooperation: Agent-based models of competition and collaboration. J Artif Soc Soc Simul 1(1). http://jasss.soc.surrey.ac.uk/1/1/review1.html
20. Binmore KG, Samuelson L (1992) Evolutionary stability in repeated games played by finite automata. J Econ Theory 57:278–305
21. Björnerstedt J, Weibull J (1999) Nash equilibrium and evolution by imitation. In: Arrow K, Colombatto E, Perlman M (eds) Rationality in economics. St. Martin's Press, New York
22. Camerer CF (2003) Behavioral game theory – Experiments in strategic interaction. Princeton UP, Princeton
23. Conte R, Castelfranchi C (1995) Cognitive and social action. University College of London Press, London
24. Conte R, Castelfranchi C (1999) From conventions to prescriptions. Towards a unified theory of norms. AI Law 7:323–340
25. Conte R, Dellarocas C (2001) Social order in multiagent systems. Kluwer Academic, Dordrecht
26. D'Arms J (2000) When evolutionary game theory explains morality, what does it explain. J Conscious Stud 7:296–300
27. Davis DD, Holt CA (1993) Experimental economics. Princeton UP, Princeton
28. Ernst Z (2001) Explaining the social contract. Br J Philos Sci 52:1–24
29. Flache A, Hegselmann R (2001) Do irregular grids make a difference? Relaxing the spatial regularity assumption in cellular models of social dynamics. J Artif Soc Soc Simul 4(4). http://jasss.soc.surrey.ac.uk/4/4/6.html
30. Fogel DB (1993) Evolving behaviors in the iterated prisoner's dilemma. Evol Comput 1:77–97
31. Frank R (1988) Passions within reason. Norton, New York
32. Gibbard A (1990) Wise choices, apt feelings. Clarendon Press, Oxford
33. Gotts NM, Polhill JG, Law ANR (2003) Agent-based simulation in the study of social dilemmas. Artif Intell Rev 19:3–92
34. Grim P (1997) The undecidability of the spatialized prisoner's dilemma. Theory Decis 42:53–80
35. Güth S, Güth W, Kliemt H (2002) The dynamics of trustworthiness among the few. Jpn Econ Rev 53:369–388
36. Güth W (2001) Do banks crowd in business ethics? – An indirect evolutionary analysis. Int Rev Econ Finance 10:1–17
37. Güth W, Dufwenberg M (1999) Indirect evolution versus strategic delegation: A comparison of two approaches to explaining economic institutions. Eur J Polit Econ 15:281–295
38. Güth W, Kliemt H (1998) The indirect evolutionary approach: Bridging the gap between rationality and adaptation. Ration Soc 10:377–399
39. Güth W, Kliemt H (2000) Evolutionary stable co-operative commitments. Theory Decis 49:197–221
40. Güth W, Kliemt H (2004) The evolution of trust(worthiness) in the net. Analyse Kritik 26:203–219
41. Güth W, Yaari M (1992) Explaining reciprocal behavior in simple strategic games: An evolutionary approach. In: Witt U (ed) Explaining process and change – Approaches to evolutionary economics. The University of Michigan Press, Ann Arbor, pp 23–34
42. Güth W, Schmittberger R, Schwartze B (1982) An experimental analysis of ultimatum bargaining. J Econ Behav Org 3:367–388
43. Güth W, Kliemt H, Peleg B (2000) Co-evolution of preferences and information in a simple game of trust. Ger Econ Rev 1:83–100
44. Güth W, Kliemt H, Brennan G (2003) Trust in the shadow of the courts. J Institut Theoret Econ 159:16–36
45. Güth W, Kliemt H, Levati MV, von Wangenheim G (2007) On

the co-evolution of retribution and trustworthiness – An (indirect) evolutionary and experimental analysis. J Institut Theoret Econ 163:143–157
46. Hamilton WD (1964) Genetical evolution of social behavior I and II. J Theoret Biol 7:1–52
47. Hamilton WD (1996) Narrow roads of geneland. Freeman, San Francisco
48. Harms W (1997) Evolution and ultimatum bargaining. Theory Decis 42:147–175
49. Hegselmann R (1996) Social dilemmas in Lineland and Flatland. In: Liebrand WBG, Messick D (eds) Frontiers of social dilemmas research. Springer, Berlin, pp 337–362
50. Hegselmann R, Flache A (1998) Understanding social dynamics. A plea for cellular automata based modelling. J Artif Soc Soc Simul 1(3). http://jasss.soc.surrey.ac.uk/1/3/1.html
51. Henrich J, Boyd R, Bowles S, Camerer C, Fehr E, Gintis H (2004) Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies. Oxford UP, New York
52. Hirshleifer J (1987) On the emotions as guarantors of threats and promises. In: Dupré J (ed) The latest on the best: Essays on evolution and optimality. MIT Press, Cambridge, pp 307–326
53. Hirshleifer J, Coll JM (1988) What strategies can support the evolutionary emergence of cooperation? J Confl Resolut 32:367–398
57. Hobbes T (1996) Leviathan (ed by Gaskin JCA). Oxford UP, Oxford
54. Hoffmann R (2000) Twenty years on: The evolution of cooperation revisited. J Artif Soc Soc Simul 3(2). http://www.soc.surrey.ac.uk/JASSS/3/2/forum/1.html
55. Hoffmann R (2001) The ecology of co-operation. Theory Decis 50:101–118
56. Huberman BA, Glance NS (1993) Evolutionary games and computer simulations. Proc Nat Acad Sci USA 90:7716–7718
58. Hume D (2007) A treatise of human nature (ed by Norton DF, Norton MJ). Oxford UP, Oxford
59. Hume D (1998) An enquiry concerning the principles of morals (ed by Beauchamp TL). Oxford UP, Oxford
60. Kagel JH, Roth AE (1995) Handbook of experimental economics. Princeton UP, Princeton
61. Kirchkamp O (1995) The spatial evolution of automata in the prisoners' dilemma. In: Troitzsch KG, Mueller U, Gilbert GN, Dran JE (eds) Social science microsimulation. Springer, Berlin, pp 306–358
62. Kitcher P (1999) Games social animals play: Commentary on Brian Skyrm's 'Evolution of the social contract'. Philos Phenomenol Res 59:221–228
63. Kohlberg L (1984) Essays in moral development. In: The psychology of moral development, vol. 2. Harper Row, New York
64. Kraines D, Kraines V (1993) Learning to cooperate with Pavlov. An adaptive strategy for the iterated prisoner's dilemma with noise. Theory Decis 35:107–150
65. Krebs D (2000) Evolutionary games and morality. J Conscious Stud 7:313–321
66. Lindgren K (1992) Evolutionary phenomena in simple dynamics. In: Langton CG (ed) Artificial Life II. Addison–Wesley, Redwood City, pp 295–312
67. Linster BG (1992) Evolutionary stability in the infinitely repeated prisoner's dilemma played by two-state Moore machines. South Econ J 58:880–903
68. Lomborg B (1996) Nucleus and shield: The evolution of social structure in the iterated prisoner's dilemma. Am Sociol Rev 61:278–307
69. Luce D, Raiffa H (1957) Games and decisions. Introduction and critical survey. Wiley, New York
70. Mackie JL (1972) Ethics – Inventing right and wrong. Penguin Books, Harmondsworth
71. Macy M, Sato Y (2002) Trust, cooperation and market formation in the US and Japan. Proc Nat Acad Sci 99:7214–7220
72. Marinoff L (1992) Maximizing expected utilities in the prisoner's dilemma. J Confl Resolut 36:183–216
73. May RM (1987) More evolution of cooperation. Nature 327:15–17
75. Maynard-Smith J (1982) Evolution and the theory of games. Cambridge UP, Cambridge
74. Maynard-Smith J, Price JG (1973) The logic of animal conflict. Nature 246:15–18
76. Miller J (1996) The coevolution of automata in the repeated prisoner's dilemma. J Econ Behav Org 29:87–112
77. Mueller U (1988) Optimal retaliation for optimal cooperation. J Confl Resolut 31:692–724
78. Nachbar JH (1992) Evolution in the finitely repeated prisoner's dilemma. J Econ Behav Org 19:307–326
79. Neumann M (2008) A classification of normative architectures. Forthcoming
80. Neumann M (2008) Homo socionicus. A case study of simulation models of norms. J Artif Soc Soc Simul 11(4). http://jasss.soc.surrey.ac.uk/11/4/6.html
81. Nowak M, May RM (1993) The spatial dilemmas of evolution. Int J Bifurc Chaos 3:35–78
82. Nowak M, Sigmund K (1990) The evolution of stochastic strategies in the prisoners' dilemma. Acta Appl Math 20:247–265
83. Nowak M, Sigmund K (1992) Tit for tat in heterogeneous populations. Nature 355:250–253
84. Nowak M, Sigmund K (1992) Evolutionary games and spatial chaos. Nature 359:826–829
85. Nowak M, Sigmund K (1993) A strategy of win-shift, lose-stay that outperforms tit-for-tat in the prisoners' dilemma game. Nature 364:56–57
86. Oliphant M (1994) Evolving cooperation in the non-iterated prisoner's dilemma: The importance of spatial organization. In: Brooks R, Maes P (eds) Proceedings of the Fourth Artificial Life Workshop. MIT Press, Cambridge, pp 349–352
87. Plato (1961) Protagoras. In: Hamilton E, Huntington C (eds) The collected dialogues of Plato. Princeton UP, Princeton
88. Probst D (1996) On evolution and learning in games. Ph D thesis, University of Bonn
89. Routledge BR (1998) Economics of the prisoner's dilemma: A background. In: Danielson PA (ed) Modeling rationality, morality, and evolution. Oxford UP, Oxford, pp 92–118
90. Schelling TC (1978) Micromotives and macrobehavior. Norton, New York
91. Schüßler R (1990) Kooperation unter Egoisten. Vier Dilemmata. Oldenbourg, München
92. Skyrms B (1994) Darwin meets 'The logic of decision': Correlation in evolutionary game theory. Philos Sci 61:503–528
93. Skyrms B (1994) Sex and justice. J Philos 91:305–320
94. Skyrms B (1996) Evolution of the social contract. Cambridge UP, Cambridge
95. Skyrms B (1997) Game theory, rationality and evolution. In:

Dalla Chiara ML (ed) Structures and norms in science. Kluwer Academic, Dordrecht, pp 73–85
96. Skyrms B (1999) Stability and explanatory significance of some simple evolutionary models. Philos Sci 67:94–113
97. Skyrms B (2004) The stag hunt game and the evolution of social structure. Cambridge UP, Cambridge
98. Skyrms B, Pemantle R (2000) A dynamic model of social network formation. Proc Nat Acad Sci USA 97:9340–9346
99. Sugden R (1986) The economics of rights, co-operation and welfare. Basil Blackwell, New York
100. Taylor M (1976) Anarchy and cooperation. Wiley, London
101. Taylor M (1987) The possibility of cooperation. Cambridge UP, Cambridge
102. Taylor PD, Jonker LB (1978) Evolutionary stable strategies and game dynamics. Math Biosci 40:145–156
103. Ullmann-Margalit E (1977) The emergence of norms. Clarendon, Oxford
104. Vanberg VJ, Congleton RD (1992) Rationality, morality, and exit. Am Polit Sci Rev 86:418–431
105. Weibull JW (1995) Evolutionary game theory. MIT-Press, Cambridge
106. Will O, Hegselmann R (2008) A replication that failed. On the computational model. In: Macy MW, Sato Y: Trust, Cooperation and Market Formation in the US and Japan. Proceedings of the National Academy of Sciences, May 2002. J Artif Soc Soc Simul 11(3). http://jasss.soc.surrey.ac.uk/11/3/3.html
107. Young HP (1993) The evolution of conventions. Econometrica 61:57–84
108. Young HP (1993) An evolutionary model of bargaining. J Econ Theory 59:145–168

Books and Reviews

Gintis H (2000) Game theory evolving. A problem-centered introduction to modeling strategic interaction. Princeton UP, Princeton

Hammerstein P, Selten R (1994) Game theory and evolutionary biology. In: Aumann RJ, Hart S (eds) Handbook of game theory with economic applications. Elsevier, Amsterdam, pp 929–993

Sober E, Wilson DS (1998) Unto others: The evolution and psychology of unselfish behavior. Harvard UP, Cambridge

Motifs in Graphs

Sergi Valverde[1], Ricard V. Solé[2]
[1] Complex Systems Lab, Parc de Recerca Biomedica de Barcelona, Barcelona, Spain
[2] Santa Fe Institute, Santa Fe, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Levels of Network Complexity
Subgraph Census
Network Motifs
Dynamic Behavior of Network Motifs
Motifs as Fingerprints of Evolutionary Paths
Future Directions
Bibliography

Glossary

Graph (network) A set G of objects called points, nodes, or vertices connected by links called lines or edges.

Subgraph A subset of a graph G whose vertex and edge sets are subsets of those of G.

Modularity A network is called modular if it is subdivided into relatively autonomous, internally highly connected components.

Scale-free network A class of complex network showing a high heterogeneity in the distribution of links among its nodes. Such distribution decays as a power law.

Network motifs These are specific subgraphs that occur in different parts of a network at frequencies much higher than those found in randomized networks.

Definition of the Subject

Several nested levels of organization can be defined for any complex system, being many of such levels describable in terms of some type of network pattern. In this context, complex networks both in nature and technology have been shown to display overabundance of some characteristic, small subgraphs (so called motifs) which appear to be characteristic of the class of network considered. These tiny modules offer a powerful way of classifying networks and are the fingerprints of the rules generating network complexity.

Introduction

Complex systems can be described, on a first approximation by means of a network [1,3,5,19]. In such a network, the typical components of the system (atoms, proteins, species, computers, humans or neurons) are simply nodes with no further structure. They are linked to others by means of an edge. The presence of such a link implies that there is some type of causal relation. Such relation can be the presence of a bond among atoms or electrostatic forces among proteins. It can also be a trophic link (who eats who) or a wire. It can also be a more abstract association, such as friendship in a social network. The resulting web provides a glimpse of the topological complexity of the network. An example of one such web is illustrated in Fig. 1a, where we display part of the network of protein-protein interactions that is present inside human

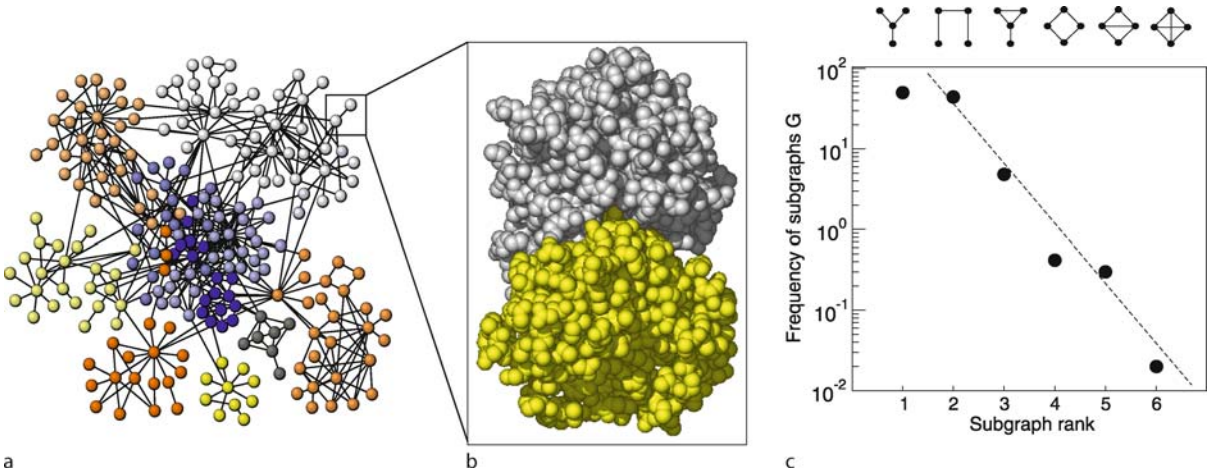

Motifs in Graphs, Figure 1

Complexity in biological systems can be seen at multiple scales. Proteins for example, typically work within pathways and assemblies and we can define a protein network or proteome as the graph of protein-protein interactions (a) where each ball corresponds to one protein. Here part of the human protein map is shown and the colors indicate different modules, i. e. groups of proteins having more connections between them than with the rest of the web. Looking closely, connections mean physical interactions among proteins, as indicated in b. Different patterns of protein interactions can be described in terms of small subgraphs. By counting the frequency of each one of these subgraphs, we obtain the so-called subgraph census, shown in c for our example (using sets of four elements)

cells. Each protein is indicated here as a sphere, although in reality it has a complex structure (Fig. 1b). Each link in this graph means that the two linked proteins interact physically (Fig. 1b).

At the small scale, the network is thus represented in terms of molecules, but as happens with all complex systems, cell behavior (largely resulting from the works of proteins) cannot be reduced to the behavior of individual units. Interactions need to be taken into account. Actually, most proteins perform their function by linking to other proteins, forming complexes. When we look at the map of protein interactions on a global scale, a complex network emerges (Fig. 1a). These assemblies then are able to interact with DNA, propagate external signals or build transport machines able to carry vesicles through the cell. Similarly, other complex networks will be describable at different levels of detail. The ultimate goal of any such description is understanding how the system works and/or how it has emerged.

An important message of complex systems approaches is the notion that most complex systems cannot be reduced to their elementary components. In particular, trying to reduce complex patterns to the properties of the underlying basic units has failed in most cases to succeed. There is quite a range of levels of analysis in any of those systems and each one involves typically some sort of emergent property. Ideally, it would be great reducing the whole complexity to the basic components. Since such a dream is not possible, we might wonder if perhaps some type of intermediate subsets of small size might be enough. Such a goal guided the proposal of searching for special types of subgraphs composed by a very small number of elements that could capture the key characteristics of complexity under the network perspective. The idea is not unreasonable, since it takes into account a subset of elements and their interactions. In other words, looking at small subgraphs (assuming their organization can be associated to a given functional trait) is a good starting point towards the analysis of higher-level complexity.

To this goal, several researchers developed a number of tools and concepts to properly capture the statistical relevance of some specific types of subgraphs. Subgraph censuses [4,7,32] and later on network motifs [13,14] were introduced in order to provide a small-scale, better characterization of complex networks. Here we will summarize their basic properties, how they are characterized and what type of relevant information they provide. The power of the approximation is illustrated with a case study on protein networks.

Levels of Network Complexity

Complex networks have been shown to exhibit a number of interesting properties. One of them is the presence of modularity. Roughly speaking, a modular system is formed by quasi-independent parts that appear inte-

grated within themselves but also exhibit a certain degree of interdependency among them. Modularity is considered a prerequisite for the adaptation of complex organisms and their evolvability [31]. It is particularly obvious in cellular networks [21], where it can be detected at the topological level. These networks include the webs of interactions among proteins, genes, enzymes and metabolites or signaling molecules. They are thus associated to the processing of energy, matter and information within and between cells. In many cases, modular structures appear to be related to functional traits: a given set of closely related proteins, for example, might be all associated to cell division or communication. All these components interact in a way or another at different times and spatial locations. How many of them need to be considered in order to capture relevant information on functional traits?

In order to explore the problem of how to define different levels of complexity, let us first provide the basic ingredients that we need. Within the context of network theory [1,3,5,19] the given system is represented as a graph $\Omega = (V, E)$ composed by a set of N nodes (say proteins) $V = \{v_i\}$ and a set of links $e_{ij} \in E$ indicating if a connection exists between nodes v_i and v_j. Each node in a graph, if connected to some other node, will have a number of links k. This is known as the node degree. Statistically, a very important description of the network structure is provided by the so-called degree distribution $P(k)$ which measures the probability of a given node having k links. Some real networks are homogeneous, meaning that their degree distribution has a well-defined average value. In these networks, most (if not all) elements will have a degree that does not strongly deviate from the average. By contrast, the majority of complex networks display a rather different architecture. This type of heterogeneous network is characterized by a probability distribution which falls off as a power law with a cut-off, i. e.

$$P(k) \sim (k + k_0)^{-\gamma} \mathrm{e}^{-k/k_c} \,. \tag{1}$$

Here k_0 is a constant and $2 < \gamma < 3$ denotes the scaling exponent (typically close to $\gamma \sim 2.5$). The cut-off k_c is a characteristic degree indicating the presence of a maximum number of links. The hubs tend to have important roles in technological but also in cellular systems [22] where the most connected nodes are often cancer-related genes and their failure typically involves some proliferative disorder. In this context, one possible attribute of an element that can give relevant information is associated to its degree. However, elements having only two nodes can be important in connecting hubs and thus degree alone is not enough.

At its smallest scale, modules are defined by means of subgraphs involving three or four elements [34]. These subgraphs have received considerable attention in relation with the so-called network motifs [13,14]. Roughly speaking, motifs are patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks. The analysis of their statistical distribution reveals that each class of natural and artificial network seems to display common patterns of motif abundances. The statistical pattern is thus interpreted as functionally meaningful. Under this view, motif abundances—as well as modularity—would be a consequence of selection forces, perhaps reflecting optimization.

Subgraph Census

Here we define the basic concepts relating small subgraphs and their abundance within complex networks. Sociologists pioneered this type of network analysis by developing the n-subgraph census, or the enumeration of all possible subgraphs with n nodes in the network. For example, in Fig. 1 the 4-subgraph census is displayed in (c). We can see that the number of subgraphs having a larger number of links (i. e. more dense ones) rapidly decays (in an exponential fashion). The way this decay occurs and the frequency distribution is characteristic for a given type of network structure. In this case, subgraphs are undirected (no arrows are taken into account) but directed links can also be taken into account, which implies much larger combinatorics (Fig. 2). Within the social sciences, the 3-subgraph census (or triad census) enables us to quantify the degree of network transitivity [4,7,32]. In a transitive social network, whenever a node i has ties with nodes j and k, there is a link between j and k and thus these nodes form a triangle (see Fig. 2). This suggests a simple transitivity test by counting the number of subgraph triangles in the network. However, any meaningful interpretation of subgraph counts requires a statistical significance test that signals important deviations from their expected value in the network (see next section).

Assuming sparse graphs ($\langle K \rangle \ll N$), the probability of a given subgraph Ω_i can be estimated. Following Itzkovitz et al., we can see how this is calculated using the subgraph example displayed in Fig. 4. Each node has a degree sequence given by the indegree list $\{K_i\}$ and the outdegree list $\{R_i\}$ (with $i = 1, \ldots, N$). The lists would be completed by the so-called mutual edges $\{M_i\}$, i. e. cases where there is a pair of edges in both directions between two nodes. For each subgraph, another degree sequence is provided by two new lists, now $\{k_j\}$ and $\{r_j\}$ for the in- and outdegrees, respectively. For our example, we have: $\{K_j\} = \{2, 1, 1, 0\}$

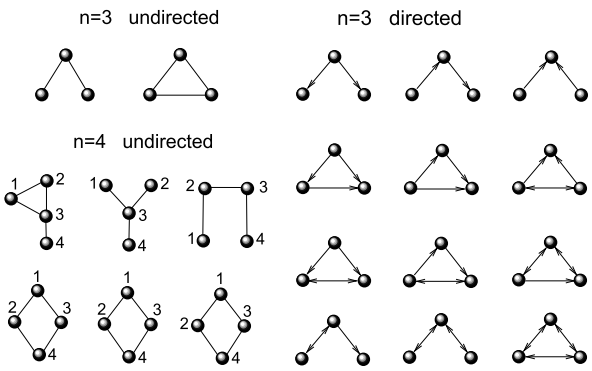

Motifs in Graphs, Figure 2
Some examples of subgraphs that can be constructed using $n = 3$ or $n = 4$ nodes. At left we show all the subgraphs that can be constructed using non-directed links (i. e. *no arrows*). A small number of graphs are obtained, but when directedness is introduced, even a very small number of nodes can generate a large number of subgraphs as shown in b where almost all combinations are shown for $n = 3$

and $\{R_j\} = \{0, 1, 1, 2\}$. The idea is to compute the different probabilities associated to each directed edge linking all pairs of nodes. For example, the probability of having a directed link from node 1 to node 2 (for $K_1 R \ll N\langle K\rangle$) is approximately:

$$P(1 \to 2) = \frac{K_1 R_2}{N\langle K\rangle} , \tag{2}$$

which can be interpreted as follows [9]: we perform K_1 attempts for the first node to connect to the target node with a probability $R_2/N\langle K\rangle$. Similarly, we would have:

$$P(1 \to 3) = \frac{(K_1 - 1)R_3}{N\langle K\rangle} \tag{3}$$

being the approach used for all edges. The average number of appearances of Ω_i is finally computed by averaging. A general derivation has been done by [9] where the power-law distribution of connectivities is taken into account. Here, we assume real networks having a scale-free in-degree distribution $P_{\mathrm{i}}(k)$ and an exponential out-degree distribution $P_{\mathrm{o}}(k)$ (the following is also valid for a network with scale-free out-degree distribution and an exponential in-degree distribution):

$$P_{\mathrm{i}}(k) = \frac{\gamma_i - 1}{k_0^{1-\gamma_i}}(k + k_0)^{-\gamma_i} , \tag{4}$$

where k_0 is the cut-off value for the distribution. The average number of appearances $\langle G\rangle$ of a given subgraph Ω_i scales with the subgraph size and the exponent of the in-degree distribution:

$$\langle G\rangle \sim N^{n-g+s-\gamma_i+1} , \tag{5}$$

where s is the maximum in-degree in the subgraph and n and g are the number of nodes and links in the subgraph, respectively. This scaling is actually valid for $2 < \gamma_i < s + 1$.

There are limitations to the above mean-field approximation. The average number of subgraphs cannot measure the high diversity of subgraph types in scale-free networks. Regular networks often display a limited repertory of connectivity patterns and they are more suitable for the mean-field analysis (see Fig. 3). The subgraph distribution provides a more adequate quantification of the impact of degree distribution in the local network structure. There is empirical evidence that subgraph distributions in scale-free networks are skewed and without a characteristic scale. In this context, [35] proposed machine-learning techniques that learn the observed subgraph distribution rather than assuming a specific type distribution. This is of uttermost importance since the choice of the model distribution strongly constrains what subgraphs are the most significant, thus introducing unwanted biases in the interpretation of the observed subgraph census.

Network Motifs

Network motifs are defined in terms of subgraphs which appear much more often than expected from pure chance. Specifically, they occur with a high frequency compared with the expected frequency from an ensemble of randomized graphs with identical degree structure [13,14]. Random networks are generated from the real network by switching links while preserving specific network properties. Random networks keep the real in/out- degree sequence but they do not have degree-degree correlations. Then, comparison between the real network and random networks signals the overabundance of sub-structures cannot be a consequence of network heterogeneity alone.

The subgraph Ω_i is a network motif when its abundance is statistically significant as indicated by the Z-score [13]:

$$Z(\Omega_i) = \frac{N_{\mathrm{real}}(\Omega_i) - \langle N_{\mathrm{rand}}(\Omega_i)\rangle}{\sigma(N_{\mathrm{rand}}(\Omega_i))} . \tag{6}$$

Here $N_{\mathrm{real}}(\Omega_i)$ is the number of times the subgraph appears in the network, whereas $\langle N_{\mathrm{rand}}(\Omega_i)\rangle$ and $\sigma(N_{\mathrm{rand}}(\Omega_i))$ refer to the mean and standard deviation of its appearances in the randomized ensemble, respectively. In order to be significant, it is required that $|Z(\Omega_i)| > 2$. When $Z(\Omega_i) > 2$ ($Z(\Omega_i) < -2$) the motif (antimotif) is

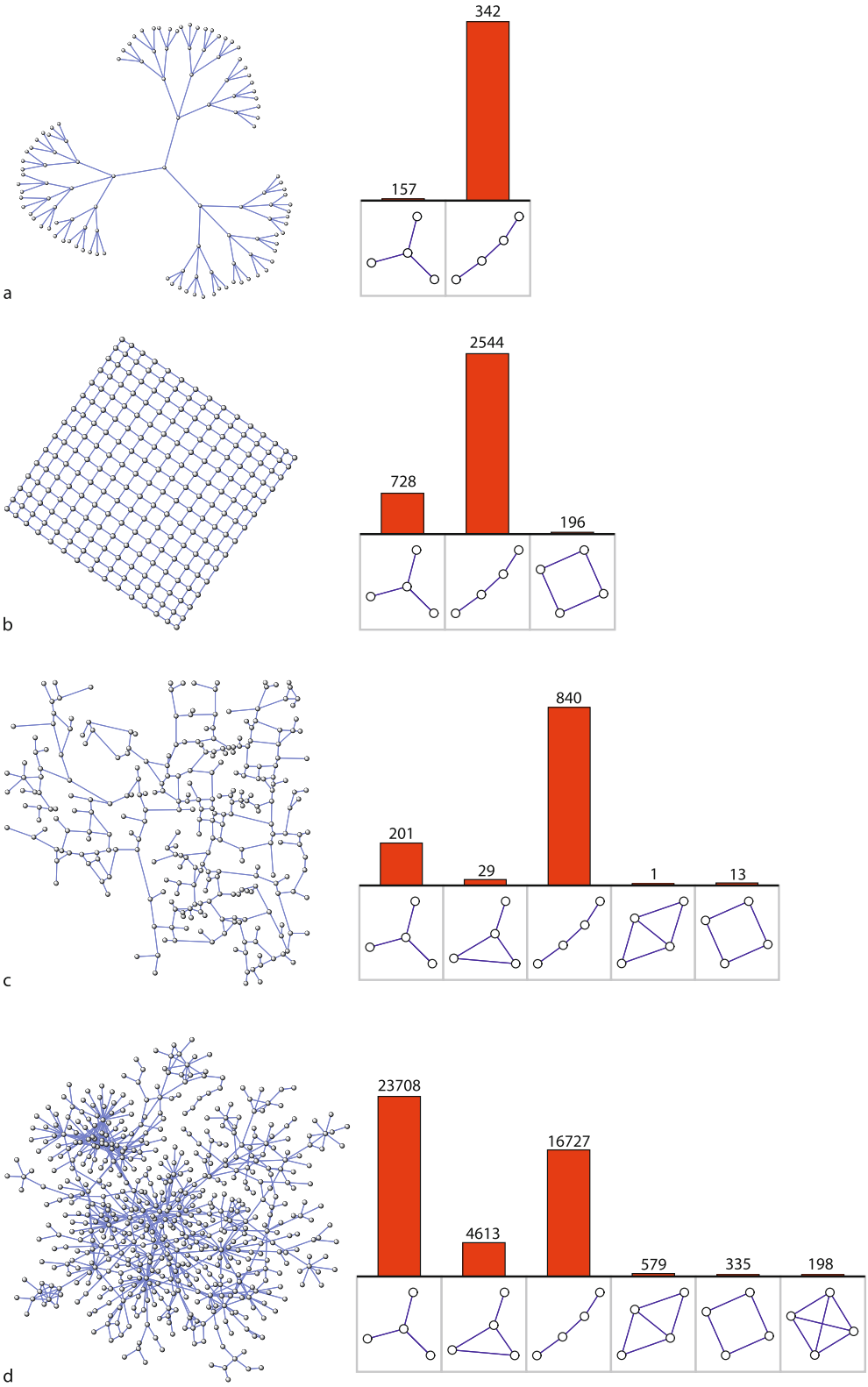
342
157
a
2544
728
196
b
840
201
29
1
13
c
23708
16727
4613
579
335
198
d

◄ **Motifs in Graphs, Figure 3**

When looking at different networks, we find that their internal structure in terms of small subgraphs differs considerably. Here four examples are shown. The first two are models, namely a a tree and b a square lattice. The other two are real networks: c a street network and d a protein interaction networks. The right column displays the observed frequencies of subgraphs of size four. As we can appreciate, different subgraphs appear to be common in some nets but not in others. Heterogeneous networks, like the protein interaction network d, display a greater variety of subgraph types than the regular networks a–c because the rich number of different node connectivity patterns

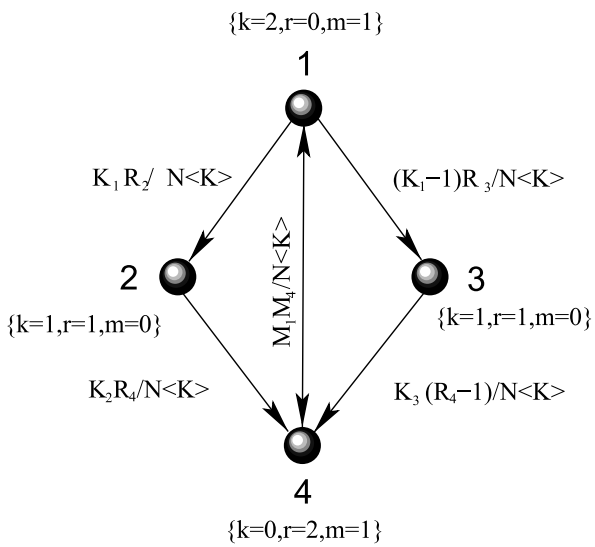

Motifs in Graphs, Figure 4

The expected frequency of a given subgraph in a random network can be computed by looking at the specific set of types of links and their numbers. Here an example of a subgraph is shown, indicating the basic quantities that can be defined in order to predict its abundance (see text)

considered to be more (less) common than expected from random.

From a practical point of view, network motif detection is a computationally intensive process. For instance, current techniques cannot detect network motifs having more than eight nodes. Motif detection requires three hard computation subtasks: (1) counting the number of subgraph instances, (2) grouping topologically equivalent subgraphs in the same class and (3) generation of random networks and comparison between observed and expected subgraph frequencies. Subgraph sampling can be used to accelerate the first task [11] but this technique is not accurate and requires a costly bias correction [33]. Much work has been done to efficiently group subgraphs having the same topology [17]. In addition, we can efficiently assess subgraph significance by comparing with theoretical estimations of the number of graphs with given degree sequence. This avoids the expensive, explicit generation of random networks during the third task [33].

A handful of motifs appear to be shared by similar or related types of networks. This is the case of Bi-parallel, Bi-fan, the feed-forward loop (FFL) and its close variants that appear both in electronic circuits and biological networks involving computations (i. e., transcriptional regulatory networks and neuronal networks). Such a common point might be easily interpreted in functional terms: similar subgraphs are abundant because they are selected or chosen to perform a given function or task. This suggests that we can classify complex networks into distinct functional families based on the set of typical motifs. As shown below, no evidence from statistical patterns supports this view.

In this context, the significance profile (SP) enables a classification of different networks according to their motif abundances [14]. The significance profile is a vector of normalized Z-scores defined as follows:

$$\mathrm{SP}(\Omega_i) = \frac{Z(\Omega_i)}{(\sum_j Z^2(\Omega_j))^{1/2}} \,. \tag{7}$$

This normalization emphasizes the relative importance of any given motif and enables a meaningful comparison across many different networks with a variable number of nodes. Figure 5 shows the SP vectors for two different types of networks using 3-size subgraphs. Although this method was proposed to obtain similar classes of networks, we must notice that networks with similar SP of 3-size subgraphs have distinct SP of 4-size subgraphs. This suggests that an effective network classification must use SP of different subgraph sizes simultaneously. Still, it is unclear what set of subgraphs is sufficient to discriminate among all possible network classes. The limitation to the maximum subgraph size handled by current motif detection algorithms may finally preclude the practical applicability of SP-based network classification. In addition, the comparison of real systems suggests the SP classification is not always reliable. Networks with different functions may have similar sets of motifs. Figure 5 shows that triad significance profiles (TSP) for software networks [28] and the transcription network in the bacteria E. coli [13] have similar SPs but they are very different systems.

We may overcome some of the above limitations by an adequate choice of the random network model used to compute the Z-score (see Eq. (6)). For example, we can assess the statistical abundance of subgraphs under different considerations besides the fixed degree sequence. The work in [8] showed that subgraph abundances of geomet-

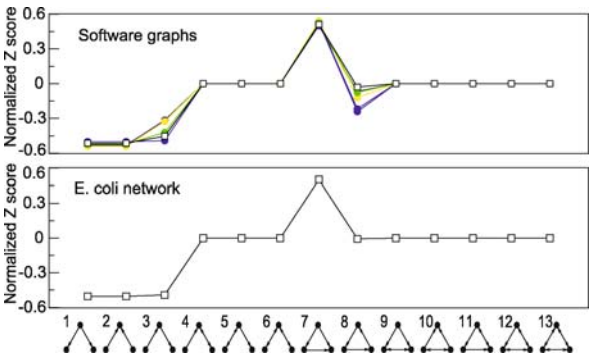

Motifs in Graphs, Figure 5
Triad significance profile (TSP) of networks representing the logical software organization in several word-processor applications (*above*) and the direct transcription network in the bacteria E. coli (*below*). Notice the remarkable similarity of TSP between these functionally unrelated systems

ric networks, where the presence of links depends on the spatial proximity of nodes, cannot be explained by spatial embedding alone. On the other hand, the large-scale network organization (i. e., described by the degree distribution together with the hierarchical structure of the network) alone might explain the natural frequencies of common motifs in biological networks [29]. Finally, standard motif detection does not handle weighted networks, where links have an associated level of activity, traffic flow or importance. The topological motif definition can be extended to take into account link weights. Here, the Z-score is replaced by the so-called motif intensity score and thus providing a dynamics perspective on subgraph importance. [20] showed that weighted motif definitions may considerably modify the conclusions drawn from pure topological statistics.

Dynamic Behavior of Network Motifs

We can further investigate the biological significance of motifs by studying their dynamical properties. In this context, a specific example of network motif, namely the feed-forward loop motif in transcriptional regulatory networks, has received much more attention than other systems. This FFL motif can be decomposed into two different kinds, i. e., coherent and incoherent, depending on the nature of individual interactions. Regulation links can be positive (activation) or negative (repression). The FFL is coherent (incoherent) if the overall sign of the indirect path from the input to the output node has the same (different) sign as the direct link. The numerical analysis of a specific instance of coherent FFL motif (i. e., with only positive links and AND-like behavior of the joint regulation of the output) shows this motif can filter out transient or fluctuating input signals [23]. This theoretical prediction was validated in experiments on motifs in real systems [15]. An extension of this work in [16] reported an exhaustive analysis of the eight possible types of coherent and incoherent FFL motifs in transcription networks. Specifically, they found that incoherent FFL motifs speed up the response time of the target gene expression following stimulus steps in one direction (e. g., off to on) but not in the other direction (on to off). This suggests the excess of FFL motifs might be explained in terms of selective pressures towards robust functioning in noisy environments.

The above emphasizes the dynamics of motifs at the local scale, where motifs are basic building blocks of functional modules. However, the same network motif could perform many different functions depending on global system requirements. For example, robust functioning can be achieved with flexible components that switch their mode of operation to replace damaged or missing components [27]. In general, the mapping between function and structure is not unique but we can expect network topology to strongly influence dynamics. For example, synthetic models of brain networks suggest they are shaped in order to maximize the number and diversity of network motifs [24]. Local structural variability allows a large space of functional states. In this context, a related question is what network features facilitate the emergence of collective synchronization (or the ability of the collective to coordinate and form a global, coherent pattern). Numerical studies suggest the pattern of motif connections determines its ability to synchronize, where denser network motifs are more prone to synchronize than less connected motifs [18].

Motifs as Fingerprints of Evolutionary Paths

An overabundance of some specific subgraphs seems a reasonable evidence for some special role. This is of course assuming that comparison is made against a random version of the same system. Moreover, it also needs assuming that changes in network structure are directly related to selective pressures. Are the abundances of network motifs evidence for such selection pressures? When looking at the network organization, we surely will recognize some forms and shapes that can easily be interpreted as resulting from selective pressures. However, it has been argued that not all patterns that we observe resulting from evolutionary change are necessarily tied to selective forces [12]. Such patterns are what evolutionary biologists Stephen Jay Gould and Richard Lewontin called *spandrels* [6]. The term spandrel was borrowed from Ar-

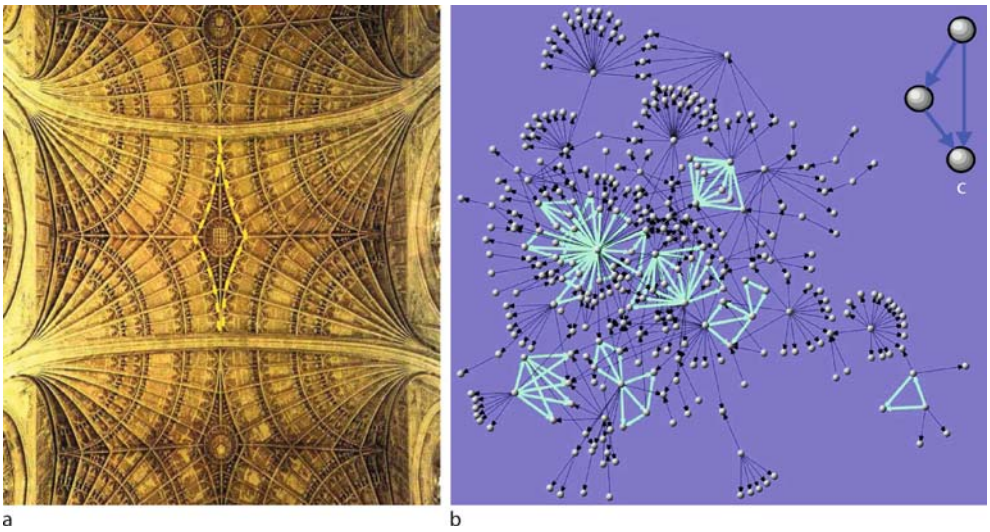

Motifs in Graphs, Figure 6

Architectural and evolved spandrels. The *left figure* shows the ceiling of King's College Chapel, in Cambridge. A spandrel is indicated in a as surrounded by a *yellow dashedline* and appears profusely decorated. The network shown in b is the gene regulatory map of *E. coli*. Here *nodes* are genes and *arrows* point from a regulatory to a regulated gene. In c a specific motif, the so-called Feed-Forward Loop (FFL) is shown. The location of this subgraph in the whole network is indicated in b by means of *lighter arrows*. Most FFL appear together, forming larger, localized structures, which are the result of duplication events (see text)

chitecture. A spandrel is the space between two arches or between an arch and a rectangular enclosure (an example is depicted with discontinuous line in Fig. 6a). In evolutionary biology, a spandrel is a phenotypic characteristic that evolved as a side effect of a true adaptation. We can summarize the features of evolutionary spandrels as follows:

1. They are the byproduct of building rules,
2. They have intrinsic, well-defined, non-random features, and
3. Their structure reveals some of the underlying rules of the system's construction.

Looking at the picture in Fig. 6a, we can see that architectural spandrels are indeed well-defined, non-random structures, arising as a side-consequence of a prior decision. Moreover, their geometric shape is fully constrained by the dominant arches.

Let us now consider cellular networks as a case study. Specifically, the patterns of protein interactions inside cells. These networks are certainly functional structures (as discussed in the Introduction) and are the result of evolution. If we look at the abundance of given motifs, we find that some of them (such as the FFL, Fig. 6c) are more common than expected from chance (for example, the FFL gives the "bump" in both top and bottom of Fig. 5). On the other hand, looking at the presence of this motif on the network, we can see that these subgraphs appear clustered together. These clustered patterns are actually characteristic of all biological motifs in cellular webs. What is the explanation for this structure? Let us forget for a moment about functionality and just consider the basic rules that make these networks grow. It is known that the genome evolves by duplication and diversification. Duplication refers to the accidental appearance of a redundant copy of a given gene and diversification to the further changes taking place in the wiring. Using a duplication model we are actually considering an essential trait of biological evolution: it takes place through tinkering [10,25]. In this context, one important source of divergence between engineering (technology) and evolution is that the engineer works according to a preconceived plan (in that he foresees the product of his efforts) and second that in order to build a new system a completely new design and units can be used without the need to resort to previous designs.

Jacob also mentions the point that the engineer will tend to approach the highest level of perfection (measured in some way) compatible with the technology available. The main point is that natural selection does not work as an engineer, but as a tinkerer, who knows what is going to be produced but is limited by the constraints present at all levels of biological organization as well as by historical circumstances. As shown below, tinkering might help explain

how special subgraphs appear more commonly than expected by chance as a result of the extensive reuse of available parts.

Changes in wiring are associated to the emergence of novelty and new functionalities. Below we consider such simple approximation based on single-gene events. One can use the simplest DD models of protein network evolution [25,30] which involves the following set of rules, to be applied a given number of times, until N nodes are present. Assuming that we have a graph of size n, we iterate the following rules:

1. Duplication: choose a node at random and duplicate it, thus generating a new node.
2. Link deletion: the new node shares a set of neighboring nodes with its predecessor. For each common pair of common links, we choose one of them and delete it with some probability δ. This rule thus removes redundant relations among proteins.
3. Link addition: a link is added among the chosen and redundant nodes with probability α. This is a small number and allows new functionalities to emerge by linking the twin proteins.

Tinkering by means of duplication and rewiring helps in understanding how some motifs might be so common and why they appear together. In Fig. 7 we show some examples of how tinkering allows us to explain why some subgraphs are expected to be more common. We can easily see that some specific arrangements (such as the graph highlighted in green) are easily obtained by a single duplication event. The *architecture* of cellular networks is not based on geometry: Topology replaces geometry and substructures need to be understood as subgraphs. Using the previous set of rules, starting from a very small graph, it is very easy to obtain a network similar to the one in Fig. 1a and with exactly the same subgraph census, as displayed in Fig. 8. In spite of the fact that the model does not consider possible functional roles, the resulting network is very close to the observed one, thus indicating that the topological patterns are a byproduct of the growth rules.

From the previous definition, motif abundances need to be understood as the spandrels of network biocomplexity [26]. Why? They follow the previous list: (a) their abundance is matched by *in silico* models lacking real functionality, and are thus a byproduct of the network building rules; (b) they exhibit highly non-random features at several scales, and these are particularly obvious when looking at the way in which motifs form clusters (Fig. 6b and c). The aggregates strongly indicate that duplication-rewiring processes, which generate the whole structure, are also responsible for their presence and specific regularities. They result from the works of the tinkerer, and their frequency and distribution within a given web have no adaptive meaning per se. The observation of a common minimal subgraph is certainly interesting, but not relevant in evolutionary terms, unless as integrated within a larger, functionally relevant set of interacting units [26].

The picture provided by network biology is a multiscale one [2] and allows one to shift from the geometry of architectural patterns to the no less fascinating universe of

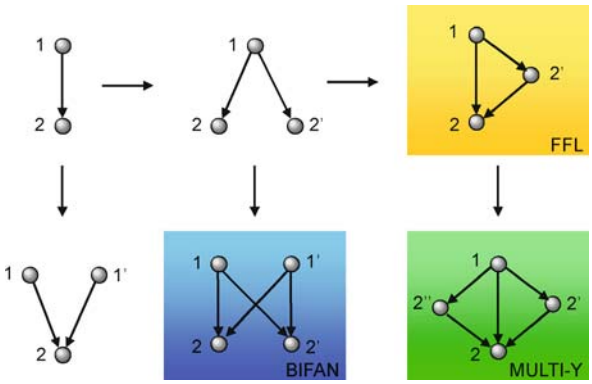

Motifs in Graphs, Figure 7
Duplication rules may explain natural motif frequencies. Here, starting from a simple two-node system we generate different subgraphs by means of duplication and rewiring rules

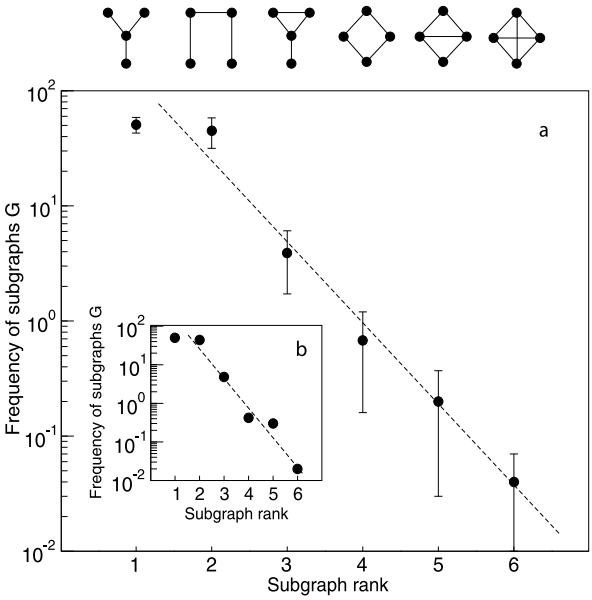

Motifs in Graphs, Figure 8
Subgraph census for the simple model of duplication and rewiring. Here we used $\delta = 0.7$ and $\alpha = 0.1$ and in a the observed frequencies of subgraphs are shown, to be compared with the ones found in human protein networks (*inset*, b)

topological patterns. Evolutionary paths proceed through extensive evolutionary tinkering, which largely influences the architectural patterns to be found. The overall topology of cellular maps includes a plethora of non-random features, some of them not being functionally selected. As the spandrels built by the architect, evolution plays the role of a tinkerer unpurposely leading to network spandrels. They include functional units, but the real relevance of them is likely to reside in their combinatorial possibilities, instead of intrinsic, individual properties.

Future Directions

The presence of subgraphs having a disproportionate number of appearances open up the possibility of detecting special properties of complex networks. They offer a good way of classifying graphs into given groups and determining the evolutionary paths that lead to such non-random abundances. The previous examples illustrate just one aspect of the possible spectrum of motif analysis, dealing with topology. But networks are weighted, the nature of the nodes is not uniform and in some cases we might have dynamical information of how things change. Although a pure topological perspective might not capture all the desired traits, an appropriate choice of the information contained in a subgraph might help in defining real functional units. Future work should test how different forms of defining motifs might overcome some of the drawbacks of using pure topology. Moreover, it remains open how motifs might somehow represent basic building blocks of complex networks or are instead another level of description of a nested hierarchy of complexity.

Bibliography

Primary Literature

1. Albert R, Barabási AL (2002) Rev Mod Phys 74:47–97
2. Bornholdt S (2005) Less is more in modeling large genetic networks. Science 310:449–450
3. Bornholdt S, Schuster G (eds) (2002) Handbook of Graphs and Networks. Wiley, Berlin
4. Davis JA, Leinhardt S (1968) The structure of positive interpersonal relations in small groups. Annual Meeting of the American Sociological Association, Boston
5. Dorogovtsev SN, Mendes JFF (2003) Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford University Press, New York
6. Gould SJ, Lewontin RC (1979) The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme. Proc Roy Soc London B 205:581–598
7. Holland PW, Leinhardt S (1970) A method for detecting structure in sociometric data. Am J Soc 70:492–513
8. Itzkovitz S, Alon U (2005) Subgraphs and Network Motifs in Geometric Networks. Phys Rev E 71:026117
9. Itzkovitz S, Milo R, Kashtan N, Ziv G, Alon U (2003) Subgraphs in random networks. Phys Rev E 68:026127
10. Jacob F (1977) Evolution as tinkering. Science 196:1161–1166
11. Kashtan N, Itzkovitz S, Milo R, Alon U (2004) Efficient Sampling Algorithm for Estimating Subgraph Concentrations and Detecting Network Motifs. Bioinformatics 20(11):1746–1758
12. Mazurie A, Bottani S, Vergassola M (2005) An evolutionary and functional assessment of regulatory network motifs. Genome Biol 6:R35
13. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network Motifs: Simple Building Blocks of Complex Networks. Science 298:824–827
14. Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S, Ayzenshtat I, Sheffer M, Alon U (2004) Superfamilies of designed and evolved networks. Science 303:1538–1542
15. Mangan S, Alon U (2003) Structure and function of the feedforward loop network motif. Proc Nat Acad Sci 100(21):11980–11985
16. Mangan S, Zaslaver A, Alon U (2003) The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. J Mol Biol 334:197–204
17. McKay BD (1981) Practical Graph Isomorphism. Congressus Numerantium 30:45–87
18. Moreno-Vega Y, Vázquez-Prada M, Pacheco A (2004) Fitness for synchronization of network motifs. Phys A 343:279–287
19. Newman MEJ (2003) SIAM Rev 45:167–256
20. Onnela J-K, Saramaki J, Kertész J, Kaski K (2005) Intensity and coherence of motifs in weighted complex networks. Phys Rev E 71:065103(R)
21. Ravasz E, Somera SL, Mongru DA, Oltvai ZN, Barabási A-L (2002) Hierarchical organization of modularity in metabolic networks. Science 297:1551–1555
22. Rodriguez-Caso C, Medina MA, Solé RV (2005) Topology, tinkering and evolution of the human transcription factor network. FEBS J 272:6423–6434
23. Shen-Orr S, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulatory network of Escherichia coli. Nat Genet 31:64–68
24. Sporns O, Kotter R (2004) Motifs in Brain Networks. PLoS Biol 2(11):e369
25. Solé RV, Ferrer I, Cancho R, Montoya JM, Valverde S (2002) Selection, Tinkering, and Emergence in Complex Networks. Complexity 8:20–33
26. Solé RV, Valverde S (2006) Are Network Motifs The Spandrels of Cellular Complexity? Trends Ecol Evol 21:419–22
27. Tononi G, Sporns O, Edelman GM (1999) Measures of degeneracy and redundancy in biological networks. Proc Natl Acad Sci USA 96:3257–3262
28. Valverde S, Solé RV (2005) Network Motifs in Computational Graphs: A Case Study in Software Architecture. Phys Rev E 72(2):026107
29. Vázquez A, Dobrin R, Sergi D, Eckmann JP, Oltvai ZN, Barabási A-L (2004) The topological relationship between the large-scale attributes and local interaction patterns of complex networks. Proc Nat Acad Sci 1001:17940–17945
30. Vázquez A, Flammini A, Maritan A, Vespignani A (2003) Modeling of Protein Interaction Networks. Complexus 1:38–44
31. Wagner G, Pavlicev M, Cheverud JM (2007) The road to modularity. Nat Rev Genet 8:921–931
32. Wasserman S, K Faust (1994) Social Network Analysis. Cambridge University Press, Cambridge

33. Wernicke S (2006) Efficient Detection of Network Motifs. IEEE/ACM Trans Comp Biol Bioinf 3(4):347–359
34. Wolf DM, Arkin AP (2003) Motifs, modules and games in bacteria. Curr Opin Microbiol 6(2):125–134
35. Ziv E, Koytcheff R, Middendorf M, Wiggins C (2005) Systematic identification of statistically significant network measures. Phys Rev E 71:016110

Books and Reviews

Banzhaf W, Kuo PD (2004) Network motifs in natural and artificial transcriptional regulatory networks. J Biol Phys Chem 4(2): 85–92

Dobrin R, Beg Q, Barabási A-L, Oltvai Z (2004) Aggregation of topological motifs in the Escherichia coli transcriptional regulatory network. BMC Bioinform 5:10

Gould SJ (2002) The Structure of Evolutionary Theory. Harvard University Press, Cambridge

Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402:C47–C52

Kuo PD, Banzhaf W, Leier A (2006) Network topology and the evolution of dynamics in an artificial regulatory network model created by whole genome duplication and divergence. Biosystems 85:177–200

Solé RV, Pastor-Satorras R, Smith E, Kepler TS (2002) A model of large-scale proteome evolution. Adv Complex Syst 5:43–54

Zhang LV et al (2005) Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network. J Biol 4(2):6

Motion Prediction for Continued Autonomy

Mario Sznaier, Octavia Camps
Electrical and Computer Engineering Department, Northeastern University, Boston, USA

Article Outline

Glossary
Introduction
Definition of the Subject
Introduction
Illustrative Examples
Summary
Future Directions
Acknowledgment
Appendix: Background Results on Linear Spaces and Robust System Identification
Bibliography

Glossary

Camshift algorithm The Continuously Adaptive Mean Shift (CAMSHIFT) algorithm is a tracking procedure based on the mean shift algorithm that was developed to cope with dynamically changing color probability distributions derived from video sequences.

Kalman filter A dynamical system (filter) that estimates the state of a linear system from measurements of its outputs corrupted by Gaussian noise.

Linear matrix inequality A matrix inequality of the form $\mathbf{A}(\mathbf{x}) \doteq \sum_i x_i \mathbf{A}_i \leq 0$, where ≤ 0 stands for negative semidefinite. An LMI of this form defines a convex constraint in the variables x_i.

Mean shift algorithm A robust non-parametric technique for climbing density gradients to find the mode (peak) of a probability density function.

Particle filter A sequential Monte Carlo method to approximate sequences of probability density functions using a large set of random samples known as particles. These particles are propagated over time using *important sampling* and resampling techniques.

Robust identification A class of deterministic identification techniques based on set descriptions of noise and allowable systems. These techniques yield both a system model compatible with the observed data and a priori assumptions, and worst-case bounds on the identification error.

Transfer matrix A (generically complex valued) matrix that relates the Z-transforms of the input $u(z)$ and the output $y(z)$ of a linear time invariant system: $y(z) = G(z)u(z)$.

Unscented Kalman filter A nonlinear estimation method where the state distribution is approximated by a Gaussian random variable chosen such that it captures the posterior mean and variance accurately up to the 3rd order of their Taylor series expansion.

Unscented particle filter A particle filter that uses an unscented Kalman filter to generate the *importance proposal distribution* for nonlinear non-Gaussian on-line estimation.

Introduction

Recent hardware developments have rendered dynamic vision – the confluence of computer vision and control– a viable option for a large number of applications, ranging from surveillance and manufacturing to assisting individuals with disabilities. This article discusses one of the critical issues currently limiting widespread use of these systems, namely their potential fragility when operating in dense, cluttered environments, and shows that robustness can be substantially enhanced by exploiting the predictive power of dynamic motion models learned from scene data. The article is organized as follows: Sect. "Definition of

the Subject" provides a brief overview of the subject. Section "Introduction" illustrates, with a simple example, the robustness challenges faced by dynamic vision methods when operating in cluttered, partially stochastic environment, and shows how to address these challenges through the use of dynamic motion models. These ideas are further developed in Sect. "Illustrative Examples", providing several examples that include inter-frame tracking, robustification of existing methods by incorporating motion models, and dynamic appearance modeling. Sections "Summary", "Future Directions" and "Background Results on Linear Spaces and Robust System Identification" summarize the issues discussed in the article, briefly covered some open issues and directions for further research, and suggest additional references, respectively. Finally, in order to make the article self-contained, we include an Appendix summarizing the key results in Linear Vector Spaces and Robust Identification used in the chapter.

Definition of the Subject

Dynamic vision and imaging – the confluence of computer vision and control – is uniquely positioned to enhance the quality of life for large segments of the general public in a cost effective way. Aware sensors endowed with moderate tracking and scene analysis capabilities can prevent crime, reduce time response to emergency scenes and allow elderly people to continue living independently. Enhanced imaging methods can substantially reduce the amount of radiation required in medical imaging procedures and in cancer therapy. Moreover, the investment required to accomplish these goals is relatively modest, since a substantial portion of the necessary hardware infrastructure already exists, since a large number of imaging sensors are already deployed and networked. For instance, the number of outdoor surveillance cameras in public spaces is already large (10,000 in Manhattan alone), and will increase exponentially with the introduction of camera cell phones capable of broadcasting and sharing live video feeds in real time. Thus, dynamic vision and imaging is arguably one of the few areas where both further advances and widespread field deployment are being held up not by the lack of a supporting infrastructure, but the lack of *supporting theory*. The challenge now is to develop a theoretical framework that allows for *robustly* processing this vast amount of information, within the constraints imposed by the need for real time operation in dynamic, partially stochastic scenarios. The goal of this chapter is to illustrate the central role that dynamic models and their associated predictions can play in developing a computationally tractable robust dynamic framework, ultimately leading to vision-based systems with enhanced autonomy, capable of operating in stochastic, cluttered environments.

Introduction

In the past few years, dynamic vision systems – i. e. systems incorporating vision as an integral part of the decision making process – have emerged as a viable option for a large number of applications, ranging from vision–based assembly [1,2,3,4,5] to vision–assisted surgery [6,7,8,9,10], assisting individuals with disabilities [11,12,13,14], and intelligent vehicle highway systems [15,16,17,18].

A requirement common to most dynamic vision applications, including the ones cited above, is the *ability to track* objects in a sequence of frames. This problem has been extensively studied in the past few years, leading to several techniques. Some of these techniques can track unknown objects [19,20,21,22,23,24,25], while others require prior knowledge of the target [26,27,28,29,30]. Orwell et al. [31] and Collins et al. [32] use color to track objects with multiple cameras. Hager and Toyama [33] track primitive features within small regions of interest (ROI) that are warped and matched against canonical configurations. Reid and Murray [34] use affine structure to track clusters of corners. Calabi et al. [19] use differential invariant signature curves to track objects. Cipolla and Blake [35] use estimates of the divergence and the deformation of closed contours to guide a robot manipulator. Blake and Isard [36] use active contours and geometrical constraints to model the likelihood of their deformations.

Correspondences between individual frames are usually integrated over time to improve robustness by exploiting the dynamical properties of the target. Kalman filter-based trackers use a model of the target dynamics and the probability distribution of the process and measurement noise to produce estimates of the future positions of the target based on (noisy) measurements of its past locations. Condensation trackers and unscented Kalman Filters [37,38,39] generalize Kalman filter-based ones by allowing more general (multimodal, nonlinear) models. In this case, analytical propagation is not longer possible and numerical methods must be used instead.

Most trackers assume a simple dynamic model such as a system moving with constant velocity. While successful in many scenarios, this approach suffers from the fact that the tracker must rely on the assumed model of the target dynamics to produce estimates of its future positions, introducing a potential source of fragility. A mismatch between this model and the actual dynamics will lead to incorrect predictions (this is the well known divergence phenomenon, see for instance [40], page 133).

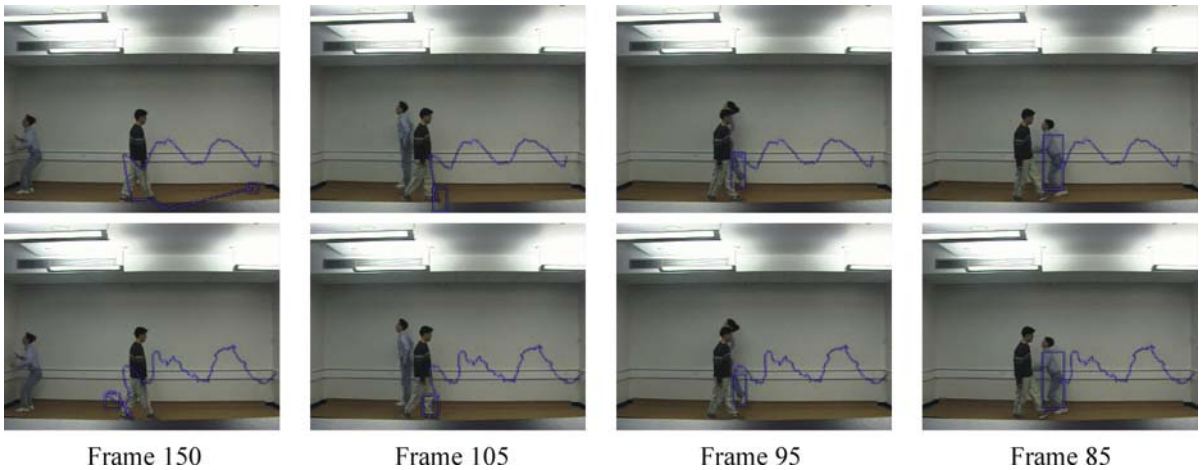

Motion Prediction for Continued Autonomy, Figure 1
Kalman (*top*) and Unscented Particle Filter (*bottom*) based tracking of a jumping individual in the presence of occlusion

This lack of robustness is shown in Fig. 1 where the effects of clutter are illustrated. As shown there, both a regular Kalman filter-based tracker and an Unscented Particle Filter (UPF), lose the target (a jumping individual) in frame 95, due to occlusion.

The objective of this chapter is to show that all of the issues noted above can be addressed by using a model of the target motion to predict a region that is guaranteed to contain it in the future (assuming that its motion modality does not change). These models can be efficiently obtained from the available data – past target positions, a priori information on motion modalities, if available, etc. – by exploiting robust identification methods developed in the control community during the past decade. As we illustrate in this chapter with several examples, this approach leads to systems capable of successfully tracking targets in the presence of substantial clutter, occlusion and even appearance changes. Thus, incorporating these algorithms in the control loop, leads to systems with increased autonomy, capable of successfully operating in dense, complex environments with minimal intervention from human operators.

Notation

In this chapter, dynamic properties of targets – such as time evolution of features or appearance, will be represented using dynamical linear models such as the one shown in Fig. 2. The model responds to some input signal u, with outputs y. The outputs are sequences of time values of measured quantities, such as the target position, size, or gray value.

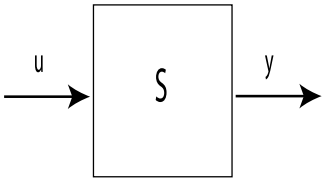

Motion Prediction for Continued Autonomy, Figure 2
Linear operator *S* with input *u* and output *y*

From an input–output viewpoint any linear model of interest S will be represented by its convolution kernel $\{s_{i,j}\}$ or by an infinite lower block triangular matrix $\mathbf{T}_S$ mapping (vector) sequences:

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} s_{0,0} & 0 & 0 & \\ s_{1,0} & s_{1,1} & 0 & \\ s_{2,0} & s_{2,1} & s_{2,2} & \\ & & & \ddots \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \end{bmatrix}. \tag{1}$$

When dealing with input–output sequences on the horizon $[0, n-1]$, we will use the finite upper left submatrix of $n \times n$, $\mathbf{T}_S^n$, obtained from the infinite matrix above.

In the sequel, we will also represent finite dimensional Linear Time Invariant (LTI) systems by using either a minimal state–space realization:

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k \\ \mathbf{y}_k &= \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k \end{aligned} \tag{2}$$

where the vectors $\mathbf{x}$, $\mathbf{u}$ and $\mathbf{y}$ represent the states, inputs and measurements, respectively, and where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$

are matrices of appropriate dimensions. Occasionally, we will also use as an alternative representation (rational) complex-valued transfer matrices:

$$S(z) \doteq \sum_{k=0}^{\infty} s_k z^k ,$$

where the coefficients s_k of the expansion are called the *Markov parameters* of $S(z)$. It is a well known fact that the coefficient s_k coincides with the impulse response of S at the kth time instant.

Finally, we summarize below some additional notation used in this chapter:

$\mathbf{x}$	real–valued (unless otherwise stated) column vector.
$\mathbf{x}_k$	kth element of a vector $\mathbf{x}$.
$\|\mathbf{x}\|_p$	p-norm of a vector: $\|\mathbf{x}\|_p \doteq \left(\sum_{k=1}^{m} \|x_k\|^p\right)^{\frac{1}{p}}$, $p \in [1, \infty)$, $\|\mathbf{x}\|_\infty \doteq \max_{k=1,\ldots,m} \|x_k\|$.
$\overline{\sigma}(\mathbf{A})$	maximum singular value of the matrix $\mathbf{A}$.
$\mathbf{A} > \mathbf{0}$	$\mathbf{A} = \mathbf{A}^{\mathrm{T}}$ is positive definite, i. e. $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} > 0$ $\forall \mathbf{x} \neq \mathbf{0}$.
$\mathcal{BX}(\gamma)$	open γ-ball in a normed space $\mathcal{X}$:
$\mathcal{H}_\infty$	Linear space of complex-valued matrix functions with bounded analytic continuation inside the unit disk, equipped with the norm: $\|G\|_\infty \doteq \sup_{\|z\|<1} \overline{\sigma}(G(z))$.

Using Predictive Models for Multiframe Tracking

In this section we briefly present an algorithm for identifying predictive motion models and using these models in the context of multiframe tracking. For the sake of clarity, the theoretical foundations of the algorithm are postponed to the Appendix. The starting point is to express the present value of a given target feature f (for instance the position of the centroid) as the output of an unknown linear system, represented by a model $\mathcal{F}$, excited by a suitable input, e. g.:

$$f(z) = \mathcal{F}(z)e(z)\,; \quad y(z) = f(z) + \eta(z) \tag{3}$$

where e represents a suitable driving signal, and f and y denote the value of the feature and its measurement, corrupted by measurement noise η, respectively. In the sequel, we will assume that the following a priori information is available:

(a) A set membership description of the measurement and process noise: $\eta_k \in \mathcal{N}$ and $e_k \in \mathcal{E}$. These sets can be used to impose correlation constraints.

(b) The model $\mathcal{F}$ admits a finite expansion of the form

$$\mathcal{F} = \overbrace{\sum_{j=1}^{N_p} p_j \mathcal{F}^j}^{\mathcal{F}_p} + \mathcal{F}_{np} .$$

Here $\mathcal{F}^j$ are N_p known, given models that contain all the information available about possible modes of motion of the target. An example of this situation is tracking moving persons where the $\mathcal{F}^j$ can be obtained off-line by training with a representative set of motions [27,42]. If this information is not available the problem reduces to purely non-parametric identification by setting $\mathcal{F}^j \equiv 0$. The second term $\mathcal{F}_{np}$ accounts for dynamics not captured by the models $\mathcal{F}_j$. Thus, it can be interpreted as the "approximation error" incurred when approximating $\mathcal{F}$ by $\mathcal{F}_p$.

(c) The magnitude of the impulse response of $\mathcal{F}_{np}$ is bounded above by $\|f_{np}(t)\|_2 \leq K\rho^t$ for some known $\rho \leq 1$ and some $K > 0$, where $f_{np}(t)$ denotes the value of the impulse response of $\mathcal{F}_{np}$ at time t.

In this context, the next value of the target feature y_k can be predicted by first identifying the system $\mathcal{F}$ and then using it to propagate its past n values. In turn, identifying the system entails finding a model $F(z) \in S \doteq \{F(z)\colon F = F_p + F_{np}\}$ such that $y - \eta = \mathcal{F}e$, precisely the class of identification problem addressed in [43] and briefly discussed in the Appendix. As shown there, such a model exists if and only if the following set of equations in the variables $\mathbf{p}$, $\mathbf{h}$ and K is feasible:

$$\mathbf{M}_R(\mathbf{h}) = \begin{bmatrix} \mathbf{R}_\rho^2 & \mathbf{T}_h^{\mathrm{T}} \\ \mathbf{T}_h & K^2\mathbf{R}_\rho^{-2} \end{bmatrix} \geq 0 \tag{4}$$

$$\mathbf{y} - \mathbf{T}_u\mathbf{P}\mathbf{p} - \mathbf{T}_u\mathbf{h} \in \mathcal{N} \tag{5}$$

where $\mathbf{T}_x$ denotes the Toeplitz matrix associated with a given sequence $\mathbf{x} = [x_1, \ldots, x_N]$, $\mathbf{R}_\rho \doteq \mathrm{diag}\,[1\rho \cdots \rho^N]$, $\mathbf{P} \doteq [f^1\ f^2\ \cdots\ f^{N_p}]$, where f^i is a column vector containing the first n Markov parameters of the ith transfer function $F^i(z)$, and where $\mathbf{h}$ contains the first n Markov parameters of $F_{np}(z)$.

In addition to providing an estimate of the next value of the target feature, this approach also has the following advantages:

(1) **Model (in)validation**: Assume that the set $\mathcal{N}$ is described by a set of Linear Matrix Inequalities (LMIs) [44] of the form:

$$\mathcal{N} \doteq \left\{ \boldsymbol{\eta} \in \Re^N\colon \mathbf{L}(\boldsymbol{\eta}) = \mathbf{L}_0 + \sum_{k=1}^{N} \mathbf{L}_k \eta_{k-1} \geq 0 \right\} \tag{6}$$

where $\mathbf{L}_i$ are given real-valued symmetric matrices. Then Eqs. (4)–(6) reduce to a set of LMIs in the variables $\mathbf{h}, \boldsymbol{\eta}$ and K^2. This allows for finding the minimum value of K^2 such that the LMIs (4)–(6) are feasible. In turn, this value can be used as a "sanity check" to assess the quality of the approximation. A large value of K indicates that the non-parametric portion of the model $\mathcal{F}_{np}$ does not provide a good description of the value of the feature, indicating that it may be necessary to re–identify the set $\{\mathcal{F}^i\}$. Infeasibility of the LMIs indicates that the experimental data is not compatible with the a priori assumptions, possibly indicating either (i) a new target activity not described by elements of the set $\{\mathcal{F}^i\}$ or (ii) the target entering a region where the noise and clutter models are no longer compatible with the description (6). Either case points to the need for re-assessing the a priori information.

(2) **Worst-case estimates of the prediction error** By construction, the operator found from the solution to the LMIs (4) is such that its response to the input $\mathbf{e}$ interpolates, within the experimental noise level η_k, the given value of the feature $f_k,\ k = 0, 2, \ldots, N-1$. However, when used to predict the *future* value of the feature, it is of interest to obtain bounds on the worst case prediction error. This can be accomplished as follows: Given a sequence $\{y_k\}_{k=0}^{N-1}$ of measurements of the value f_k of the feature, define the consistency set as:

$$\mathcal{T}(\mathbf{y}) \doteq \left\{F \in S\colon \{y_k - (F * e)_k\}_{k=0}^{N-1} \in \mathcal{N}\right\} \tag{7}$$

i. e., the set of all models consistent with both the a priori information and the experimental data. Note that the proposed method is interpolatory, that is, it always generates a candidate operator $F_{id} \in \mathcal{T}(\mathbf{y})$. Thus, since the "true" operator F_o that maps the input $\mathbf{e}$ to the feature values $\mathbf{f}$ must also belong to the consistency set (see the Appendix) it follows that, given the first N measurements $y_i,\ i = 0, \ldots, N-1$ a bound on the worst case prediction error over the horizon $[0, M-1]$, $M > N$, is given by:

$$\|\hat{\mathbf{f}} - \mathbf{f}\|_{\ell_\infty[0,M-1]} \leq \sup_{\mathbf{y}} d\left[\mathcal{T}(\mathbf{y})\right] = \mathcal{D}(\mathcal{I}) \tag{8}$$

where $d(.)$ and $\mathcal{D}(\mathcal{I})$ denote the diameter of the set $\mathcal{T}(\mathbf{y})$, in the $\ell_\infty[0, M-1]$ metric and the diameter of information, respectively. Moreover, since the a priori sets $(S, \mathcal{N})$ are convex and symmetric, with points of symmetry $F_s = 0$ and $\eta_s = 0$ respectively, it can be shown (see the Appendix) that:

$$\mathcal{D}(\mathcal{I}) \leq 2 \sup_{F \in S(\mathbf{0})} \|F\|_{\ell_\infty[0,M-1]} \tag{9}$$

where $S(\mathbf{0})$ denotes the set of operators compatible with the zero outcome: $y_k = 0,\ k = 0, 1, \ldots, N-1$. As we will illustrate in the sequel with a simple example, computing this bound reduces to a convex optimization problem.

Illustrative Examples

In this section we illustrate the use of predictive models identified from image data with several examples. In the first one we consider, for the sake of simplicity, static tracking, and indicate how to use these models to both predict future locations of the target and obtain *worst case* bounds on the prediction error. In the remainder of the examples we show how to combine these models approach with existing Kalman and UPF techniques to improve robustness. In all cases, the feature values were measured using an implementation of the Camshift algorithm in OpenCv [45]. The video sequences for these results as well as additional examples are available at http://robustsystems.ee.psu.edu.

Inter-Frame Tracking and Prediction

In this example we consider the problem of predicting the location of the centroid of the child shown in Fig. 3, from past measurements of its coordinates, (x_k, y_k), corrupted by uncorrelated noise, η. For the sake of briefness we report below only the results for the x coordinate, since those for y are similar.

The following a priori information was used:

1. $\mathcal{N} = \{\eta : \|\eta\|_\infty \leq 5.5\}$ (This value was quantified from fluctuations in the data taken when the person was at rest).
2. $\mathcal{E} = \delta(0)$, i. e. motion of the target was modeled as the impulse response of the unknown operator F (this is equivalent to lumping together the dynamics of the plant and the input signal)
3. The parametric part of the model $F_p \in \text{span}(\mathbf{G})$, $\mathbf{G}(z) \doteq [\frac{z^2}{z^2-2z+1}, \frac{z}{z^2-2z+1}]^{\mathrm{T}}$
4. The reminder, nonparametric component, which explains the unmodeled dynamics has an exponential decay with rate $\rho = 0.99$.

The experimental a posteriori data consisted of the first $N = 12$ frames of the sequence. The resulting LMI prob-

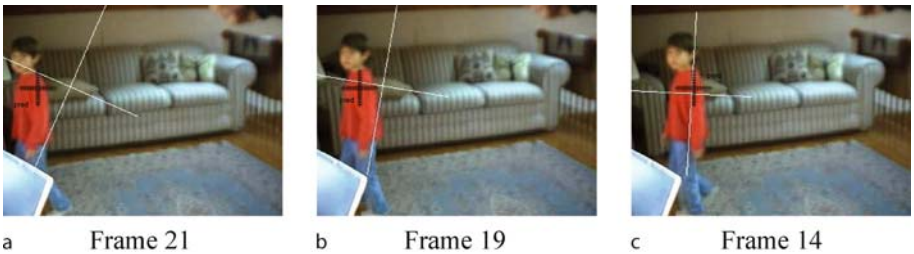

Motion Prediction for Continued Autonomy, Figure 3
Robust identification based tracking (*black cross*) versus Mean Shift (*white cross*)

Motion Prediction for Continued Autonomy, Table 1
Id error as a function of *k*. Target width is 30 pixels

Sample	13	14	15	16	17	18	19	20
Mean-Shift	25.90	35.93	41.32	45.63	54.65	57.53	65.05	64.80
Id-based	8.87	6.14	10.04	13.03	10.31	15.72	19.50	26.04
Worst case bound	13.00	15	17	19	21	23	25	27

lem was solved using MATLAB's LMI Toolbox, leading to $K_{\text{opt}} = 1.35\text{e}^{-12}$ and $\mathbf{p} = [127.7763 \;\; -135.0723]^{\text{T}}$. Note that the very low value of K indicates that indeed the parametric part F_p provides an accurate model of the dynamics of the target.

The advantage of using the predictive power of the identified models is illustrated in Fig. 3, comparing the position of the centroid predicted by the model without using new measurements (black crosses), against the results of using a Mean Shift based tracking new measurements (white crosses) implemented in Intel's Open Source Computer Vision Library [45]. Although Mean Shift is designed to improve tracking robustness by exploiting color information [46], it begins to track poorly in frame 19, and by frame 21 it has completely lost the target due to a combination of clutter and moderate occlusion. The corresponding numerical values of the error, computed as the difference between the predicted and actual values, obtained using off-line image processing, are given in Table 1. As shown there, the identified model is able to predict the location of the target, far beyond the point where the Mean Shift tracker has failed.

Finally, notice that in this case computing the worst case prediction error bound (9) reduces to a Linear Programming problem in $\mathbf{p} = [p_1 \cdots p_{N_p}]^{\text{T}}$ and $\mathbf{h} = [h_0 \cdots h_N]^{\text{T}}$. The last row in Table 1 shows the error bounds as a function of k. As expected these values increase with time, since no new data is being used beyond $k = 12$. However, they became comparable with the width of the target (30 pixels) only beyond $k = 20$.

Improving Robustness of Kalman and UPF Trackers

In this section we present several examples to illustrate how to use the identified models to improve the robustness of trackers, such as Kalman and UPF, that rely on a combination of past measurements and the dynamics of the target to estimate its future location. Proceeding as in the previous example, we used a combination of a priori information:

1. 5% noise level
2. $\mathcal{E} = \delta(0)$, i. e. motion of the target was modeled as the impulse response of the unknown operator F
3. $\mathcal{F}_p \in \text{span}[\frac{1}{z-1}, \; \frac{z}{z-a}, \; \frac{z}{(z-1)^2}, \; \frac{z^2}{(z-1)^2}, \; \frac{z^2-\cos\omega z}{z^2-2\cos\omega z+1}, \; \frac{\sin\omega z^2}{z^2-2\cos\omega z+1}]$
   where $a \in \{0.9, 1, 1.2, 1.3, 2\}$ and $\omega \in \{0.2, 0.45\}$
4. The reminder, nonparametric component, which explains the unmodeled dynamics has an exponential decay rate with $\rho = 0.99$

and the a posteriori measurements from $N = 20$ frames, where the targets were not occluded, to estimate their dynamics. These dynamics were then used in conjunction with a Kalman filter, leading to the results shown in Figs. 4 to 11.

Figure 4a shows tracking results using a Kalman filter based on the identified models for the sequence previously shown in Fig. 1. Using the predictive power of this model, the Kalman filter is now able to track the target past the casual occlusion. Figure 4b illustrates the robustness of the

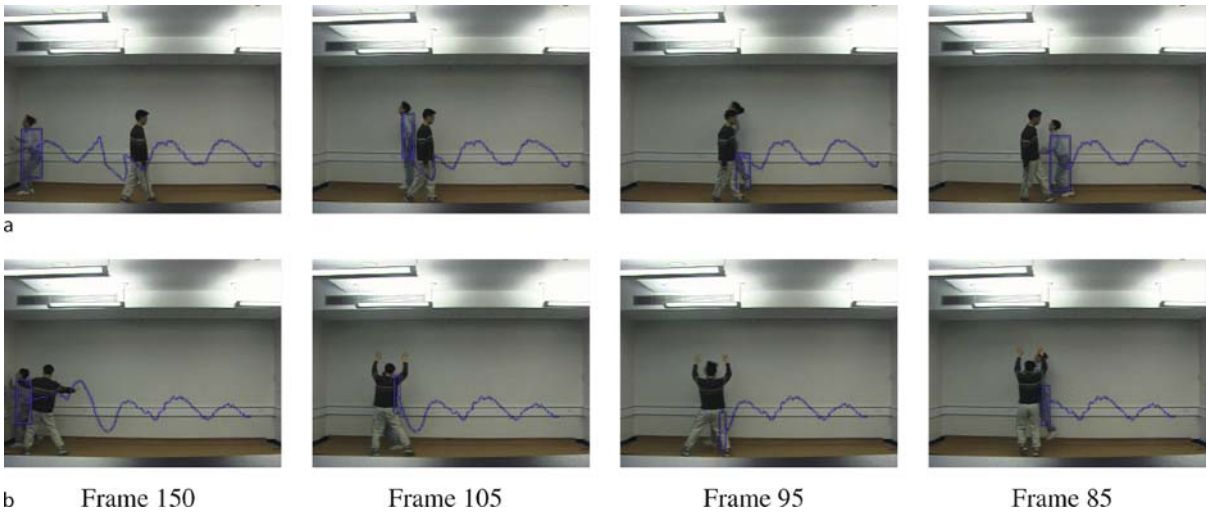

Motion Prediction for Continued Autonomy, Figure 4
Combination of Kalman and CF based tracking in the presence of occlusion a Casual occlusion. b Malicious occlusion

approach on a similar sequence but with "malicious" occlusion – i. e. the occluding person is making an effort to hide the target from the camera.

It is worth mentioning that consistent experience suggests that, for a given target and gait type, it is not necessary to re–identify the dynamics of the target for each sequence. For instance, the results at the bottom of Fig. 4 were obtained using the dynamics identified using the top sequence.

Figures 5 and 6 show examples tracking a paper plane and a bouncing ball, respectively. Figures 5a and 6a show the results of tracking these targets using a Kalman filter combined with simple constant velocity dynamics. In both cases, the tracker fails to recapture the target after it is occluded. Figures 5b and 6b show that this problem is overcome when the Kalman filter is combined with the identified models. In both examples, the blue trajectory corresponds to the frames used during the identification and the green trajectory corresponds to the trajectory "predicted" by the tracker.

Dynamic Appearance Modeling

Arguably, one of the most important challenges that needs to be solved in order to successfully track a target is to overcome changes to its appearance that might occur over time. These changes, including size, shape and color can be due to several factors such as target motion, self-occlusion, target articulations, and changes in illumination. Most tracking algorithms focus on the target position and only address changes in appearance indirectly by using probabilistic methods to compare the target to some nominal template using for example pixel statistics [45,46,47,48]. Other techniques focus on building adaptive appearance models [49] or use linear subspaces [23,26,50]. For example, Jepson et al. [49] use an online EM algorithm to adapt the appearance of a target template over time. Hager and Belhumeur [23], Black and Jepson [26] and Ho et al. [50], among others, learn target appearance models using linear subspaces. While successful in many scenarios, these approaches suffer from the fact that the obtained models tend to be too rigid and fail to capture the *dynamics* of the appearance changes.

The effects of appearance change on tracking algorithms is illustrated in Fig. 7, showing a few frames from a sequence where a blue and red football is tossed in the air. The two rows show results for tracking the ball using a combination of a Kalman Filter with motion dynamics identified using the algorithm outlined in the previous section, a combination shown there to be quite robust in the absence of appearance changes. In the first row, the measurements were obtained using mean shift [45] to compare the color distribution of the target against its initial distribution. Since initially, only the blue side of the ball is visible, this approach fails to track the ball when it becomes red. In the second row, the measurements were obtained by using a search window where the color distribution is updated using the previous frame. While this approach can cope with color change as long as the ball remains visible, it fails to recover after the ball is occluded. This is due to a combination of two facts: i) the visible color of the ball at the time when the ball comes out from below the table

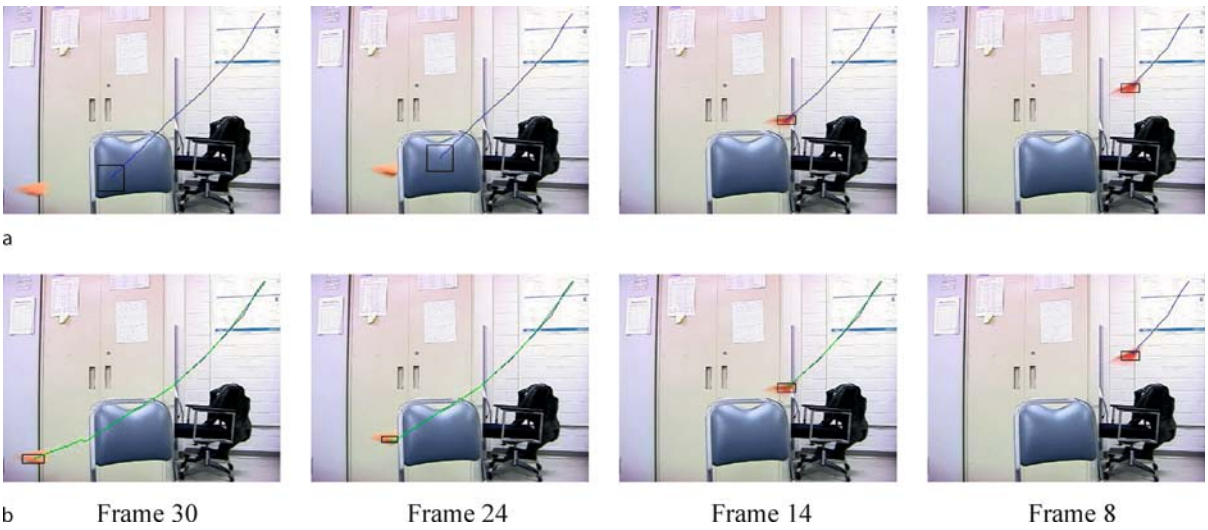

Motion Prediction for Continued Autonomy, Figure 5
Paper airplane example. a Kalman based tracking using constant velocity dynamics. b Combination of Kalman and identified models based tracking

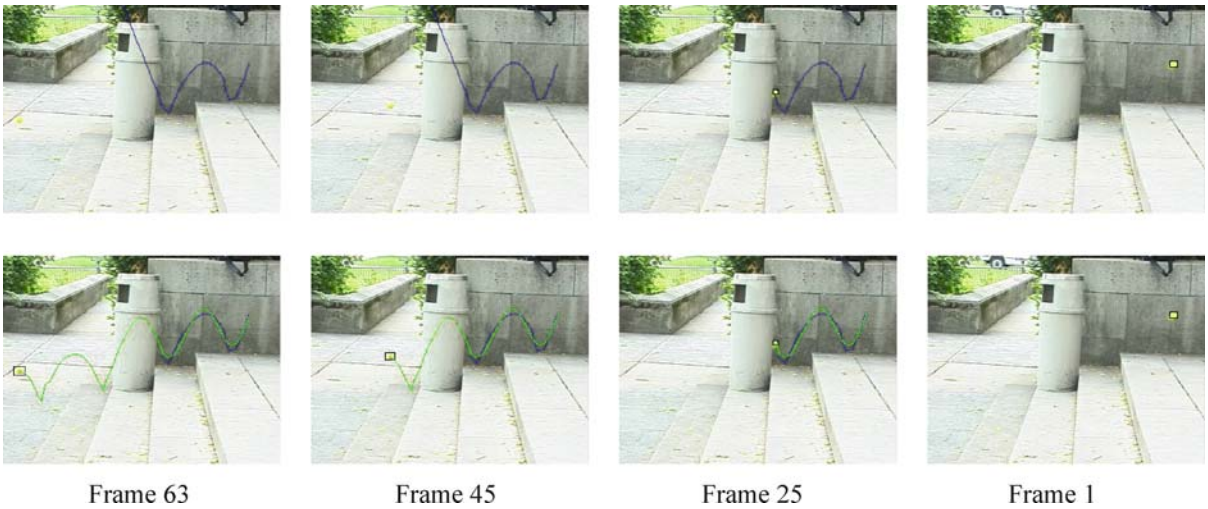

Motion Prediction for Continued Autonomy, Figure 6
Bouncing Ball Example. a Kalman based tracking using constant velocity dynamics. b Combination of Kalman and identified models based tracking

(red) is different to the visible colors just before the occlusion (blue) and ii) the binder on the table is similar in color to the blue on the ball.

The difficulties illustrated above suggest the need for more sophisticated *dynamic* appearance models that incorporate multiframe time–evolution information and have better predictive capabilities. These descriptions can be obtained by modeling the evolution of relevant appearance descriptors as the trajectories of a dynamical system. In turn, these dynamics can be obtained using the same robust identification approaches employed to identify the motion dynamics. This idea is illustrated next with several examples.

Consider first the problem of tracking the multicolored football shown in Fig. 7. In this case, the *bin counts of the hue histogram of the target* was used as appearance descriptors. These histograms were computed in small regions determined by a mean shift algorithm comparing the

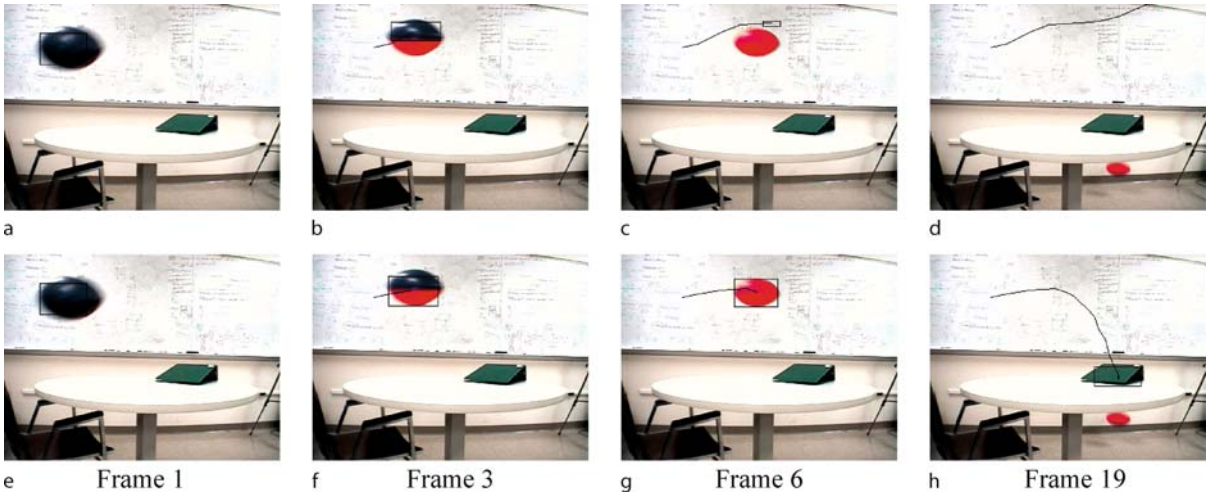

Motion Prediction for Continued Autonomy, Figure 7
Color statistics from the first frame (*top*) and from the previous frame (*bottom*) fail to track a red and blue ball

hue values of the region against the current estimates of the hue histograms. Proceeding as described in the previous section, we used a combination of a priori information:

1. 10% noise level for the appearance parameters and 3% for the location coordinates of the target.
2. $\mathcal{E} = \delta(0)$, i. e. the histogram bin counts and the location coordinates of the target were modeled as the impulse response of unknown operators F
3. The parametric parts of the operators must satisfy: $F_p \in \text{span}[\frac{1}{z+1}, \frac{z}{(z-1)^2}\frac{z^2-\cos\omega z}{z^2-2\cos\omega z+1}]$ where $\omega \in \{0.1, 0.12, 0.55\}$
4. A nonparametric component with an exponential decay rate $\rho = 0.99$

and the a posteriori measurements from $N = 21$ frames, where the target was not occluded, to estimate their dynamics. These dynamics were then used in conjunction with Kalman filters, leading to the results shown in Figs. 8 and 9, the latter showing close agreement between the actual appearance of the ball and the model prediction. Note that there are no measurements while the ball is occluded by the table, but once they become available again, there is a close match between measurements and predicted values.

The second example illustrates the use of identification methods to track a target whose size changes as it approaches the camera. As shown in Fig. 10, a conventional approach fails here due to a combination of occlusion and size change. This difficulty can be solved by using identified models that capture not only the dynamics of the motion, but also the dynamics of the change in size. The latter can be accomplished by using as appearance descriptors the *vertical and horizontal sizes* of a box around the target, found by comparing the hue histogram of the target against the initial hue histogram with a mean shift algorithm (i. e. assuming that the colors of the target are approximately constant). Figure 11 shows that the resulting tracking algorithm is able now to recover the target after the occlusion and predict its correct size.

Finally, it should be noted that a salient feature of these results is the fact that the combination Identified Models/Kalman Filter can substantially outperform the performance of a (substantially more complex) Unscented Particle Filter (UPF) acting alone. Thus the use of identified models can both improve robustness and alleviate the computational complexity of the problem.

Summary

In the past few years dynamic vision techniques have proved to be a viable option for a large number of applications, ranging from surveillance and manufacturing to assisting individuals with disabilities. Arguably, at this point one of the critical factors limiting widespread use of these techniques is the potential fragility of the resulting systems. This chapter shows that in the case of multiframe tracking this fragility can be addressed by using predictive motion models identified from the image data using robust identification tools recently developed in the control community. The advantages of this approach, and in

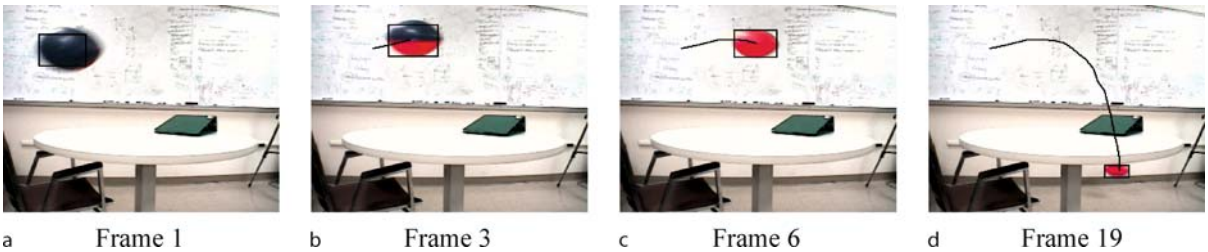

Motion Prediction for Continued Autonomy, Figure 8

Identification-based dynamical appearance model. Tracking using dynamical appearance and motion models of a red and blue ball is successful, even in the presence of occlusion. See text for details

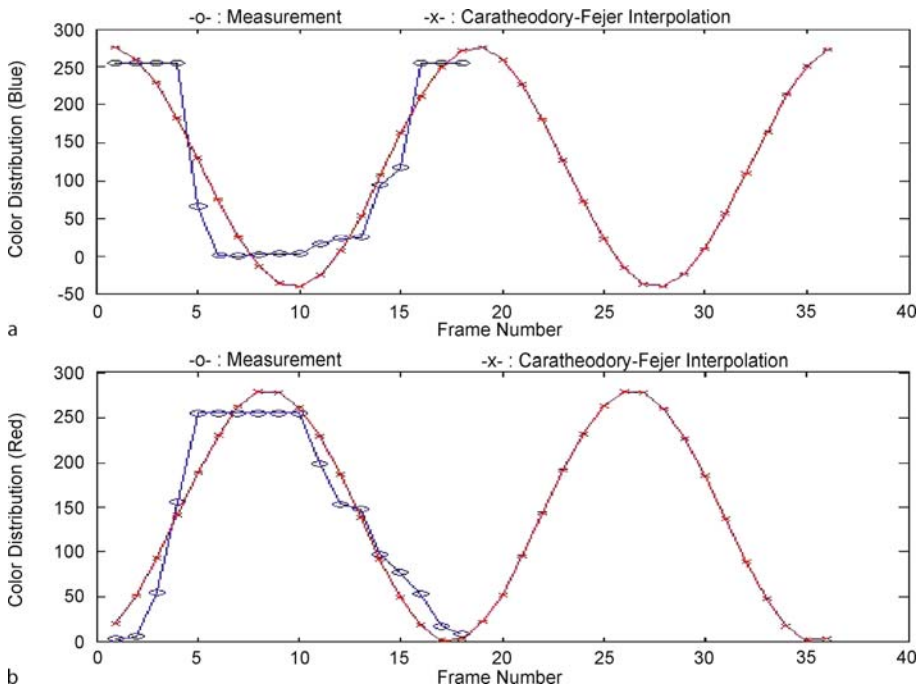

Motion Prediction for Continued Autonomy, Figure 9

Measurements and estimated values using identified models of the a blue, and b red bin counts of the hue histogram of the ball shown in Fig. 8 as it flies in front of the camera

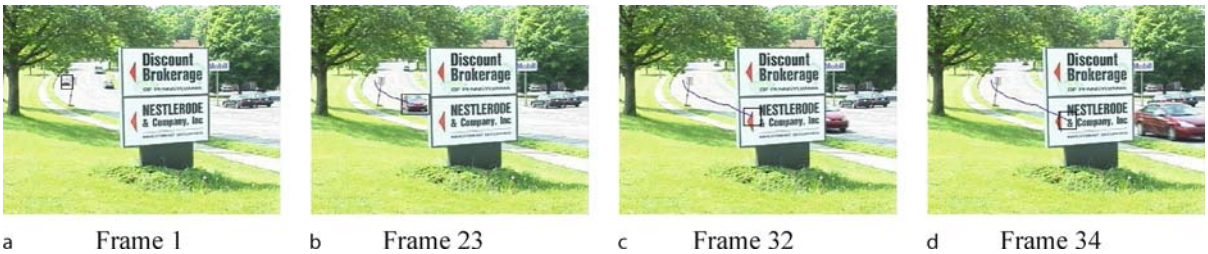

Motion Prediction for Continued Autonomy, Figure 10

Tracking an approaching car using a Kalman based tracker based constant velocity assumption

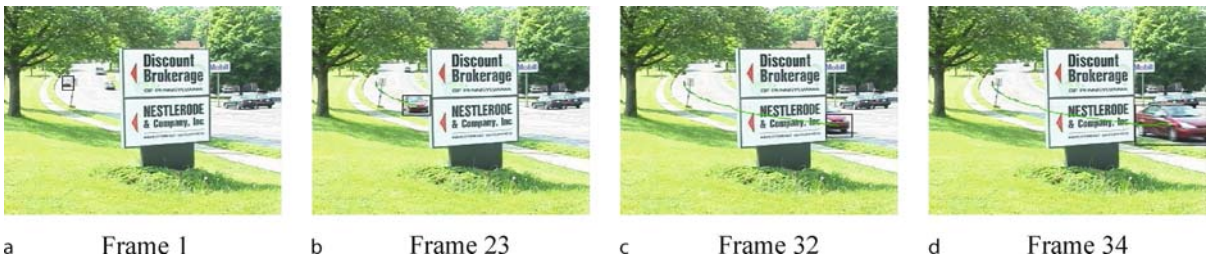

Motion Prediction for Continued Autonomy, Figure 11
Approaching car example. An increasing size car is successfully tracked in the presence of occlusion by using robust identification to obtain a predictive model for the size and the motion dynamics of the target

particular its potential to result in robust algorithms when combined with existing tracking techniques was illustrated with several experimental results.

Future Directions

As illustrated in this chapter, the use of predictive motion models can substantially enhance the robustness of dynamic vision systems, allowing for successful autonomous operation in complex scenarios. However, the computational complexity entailed in finding these models grows rapidly with the amount of data available, a situation arising for instance when attempting to track multiple targets and/or combine information from several non-overlapping sensors. Recent research (see for instance [51,52,53] suggests that a substantial reduction in computational complexity can be achieved by combining the identification techniques discussed in this chapter with machine learning and non-linear dimensionality tools to embed the problems in lower dimensional spaces. However, further research is required in order to develop a systematic procedure for finding optimal embeddings. In addition, the algorithms described here assume that the motion of the target can be adequately captured by a linear time invariant model. While this assumption holds in many cases, it will fail when the motion modality changes. An example of this situation is a person who switches from walking to running. Addressing this case requires extending the identification tools presented here to the case of *switched* (or piecewise) *linear systems*. While this problem is currently the object of considerable interest in the control community, a comprehensive solution is not yet available.

Acknowledgment

Support from NSF under grants ECS–0221562, IIS–0117387, and ITR–0312558 and AFOSR under grant FA9550–05–1–0437 is gratefully acknowledged.

Appendix: Background Results on Linear Spaces and Robust System Identification

In this appendix we summarize, for ease of reference, the background results on linear spaces and robust identification used in this chapter.

Linear Spaces

Algebraic structures are instrumental in understanding many problems arising in systems theory from an abstract point of view. In particular these tools are required to formalize and solve the optimal filtering and estimation problems arising in the context of multiframe tracking.

Field

Definition 1 A *field* $(\mathcal{F}, \&, \star)$ is an algebraic structure composed of a set $\mathcal{F}$ and two operations & and ★ with the following properties:

1. Set $\mathcal{F}$ is closed with respect to &, i. e. $a, b \in \mathcal{F} \implies (a\&b) \in \mathcal{F}$.
2. Operation & is associative, i. e. $(a\&b)\&c = a\&(b\&c) = a\&b\&c$ for $a, b, c \in \mathcal{F}$.
3. Operation & is commutative, i. e. $a\&b = b\&a$ for $a, b \in \mathcal{F}$.
4. Set $\mathcal{F}$ contains the neutral element $n_\&$ with respect to &, that is, there exists $n_\&$ such that $a\&n_\& = a$ for all $a \in \mathcal{F}$.
5. Set $\mathcal{F}$ contains the inverse element $a^I_\&$ with respect to &, that is, for all $a \in \mathcal{F}$ there exists $a^I_\& \in \mathcal{F}$, such that $a\&a^I_\& = n_\&$.
6. Set $\mathcal{F}$ is closed with respect to ★.
7. Operation ★ is associative.
8. Set $\mathcal{F}$ contains the neutral element $n_\star$ with respect to ★.
9. Set $\mathcal{F}$ contains the inverse element $a^I_\star$ with respect to ★.

10. Operation ★ is distributive with respect to &, i. e. $(a\&b) \star c = (a \star c)\&(b \star c)$ for $a, b, c \in \mathcal{F}$.

An example of a field is the set R of the real numbers, equipped with operations $(+, \times)$ as $(\&, \star)$ respectively. Here $n_\& = 0$, $n_\star = 1$, $a^I_\& = -a$ and $a^I_\star = a^{-1}$ $(a \neq 0)$.

Linear Vector Space

Definition 2 A set $\mathcal{V}$ is a *linear vector space* over the field $(\mathcal{F}, +, \times)$ if and only if the following properties are satisfied (in the sequel the elements of $\mathcal{F}$ and $\mathcal{V}$ will be called scalars and vectors respectively):

1. Set $\mathcal{V}$ is closed with respect to +.
2. Operation + is associative in $\mathcal{V}$.
3. Operation + is commutative in $\mathcal{V}$.
4. Set $\mathcal{V}$ contains the neutral element with respect to +.
5. Set $\mathcal{V}$ contains the inverse element with respect to +.
6. $\mathcal{V}$ is closed with respect to operation $\times$ between scalars and vectors.
7. Operation $\times$ among scalars and vectors is associative in the scalars, i. e. $(a \times b) \times v = a \times (b \times v) = a \times b \times v$ for $a, b \in \mathcal{F}$ and $v \in \mathcal{V}$.
8. Distributive 1: $(a + b) \times v = (a \times v) + (b \times v)$ for $a, b \in \mathcal{F}$ and $v \in \mathcal{V}$.
9. Distributive 2: $(u + v) \times a = (u \times a) + (v \times a)$ for $a \in \mathcal{F}$ and $u, v \in \mathcal{V}$.
10. Field $\mathcal{F}$ contains the neutral element of operation $\times$ between vectors and scalars, i. e. $n_\times \times v = v$ for $n_\times \in \mathcal{F}$ and $v \in \mathcal{V}$.

Examples of linear spaces over the field of real numbers are the set of matrices in $R^{n\times 1}$ and the set of sequences of real numbers, both equipped with the usual addition and scalar multiplication operations. The former is an example of a finite dimensional space, while the latter could be finite or infinite dimensional depending upon whether finite or infinite sequences are considered.

Metric, Norm and Inner Products

Definition 3 A *metric space* $(\mathcal{V}, m(\cdot,\cdot))$ is defined in terms of a linear vector space $\mathcal{V}$ and a real function (the "metric") $m(\cdot,\cdot) : \mathcal{V} \times \mathcal{V} \to R_+$, satisfying the following conditions:

1. $m(x, y) \geq 0 \quad \forall x, y \in \mathcal{V}$.
2. $m(x, y) = 0 \quad \Longleftrightarrow \quad x = y$.
3. $m(x, y) = m(y, x) \quad \forall x, y \in \mathcal{V}$.
4. $m(x, z) \leq m(x, y) + m(y, z) \quad \forall x, y, z \in \mathcal{V}$.

Here $R_+ \doteq \{x \in R,\ x \geq 0\}$.

Definition 4 A *normed space* $(\mathcal{V}, \|\cdot\|)$. is defined in terms of a linear vector space $\mathcal{V}$ and a real function $\|\cdot\| : \mathcal{V} \to R_+$ that satisfies the following conditions:

1. $\|x\| \geq 0 \quad \forall x \in \mathcal{V}$.
2. $\|x\| = 0 \quad \Longleftrightarrow \quad x = \underline{0}$.
3. $\|\alpha x\| = |\alpha| \cdot \|x\| \quad \forall x \in \mathcal{V}, \alpha \in \mathcal{F}$.
4. $\|x + y\| \leq \|x\| + \|y\| \quad \forall x, y \in \mathcal{V}$.

Here $|\cdot|$ represents the magnitude of a scalar.

The following are examples of normed spaces:

1. The linear space of n-dimensional real vectors, equipped with the norm:

$$\|x\|_p \doteq \sqrt[p]{\sum_{i=1}^{n} |x_i|^p} \quad p \geq 1 \tag{10}$$

$$\|x\|_\infty \doteq \max_{1\leq i\leq n} |x_i| \tag{11}$$

2. The linear space of real sequences, equipped with the norm:

$$\|x\|_p \doteq \sqrt[p]{\sum_{i=1}^{\infty} |x_i|^p} \quad p \geq 1 \tag{12}$$

$$\|x\|_\infty \doteq \max_{i\geq 1} |x_i| \tag{13}$$

Robust Identification

The field of system identification concerns itself with mechanisms and algorithms that process finite, partial, and corrupted data to yield abstract mathematical descriptions of real world systems.

Traditional identification approaches [55,56] assume that the data is corrupted by a stochastic process with known statistical properties and that the system to be identified has a prescribed model structure. Most of these identification procedures are based on least squares methods that estimate the parameters of the hypothesized models from the corrupted measurements. In these approaches the only source of uncertainty is the noise in the measurements while the prescribed model is assumed to be an accurate representation of the real system.

In many situations, for example when measurements are known within an accuracy range or when the available statistical information might be questionable, deterministic bounded noise descriptions are a practical and sound alternative to stochastic ones. Using this approach, the problem of system identification can be formulated as finding the sets of parameter values that are consistent

with the known noise bounds. A survey of set membership formulations for system identification can be found in [57].

Noise description is only one of the factors affecting the quality of an identified model. Perhaps a more important factor is the unrealistic presumption that a fixed model structure may fully represent the system to be identified: In practice, only partial information of the physical system is available, model parameters might change due to different operation conditions, and real systems are often too complex to be accurately modeled from first principles. These issues are addressed by *robust system identification*, which departs from traditional approaches by using a **deterministic worst-case** approach with no prior assumption about the order of the system. Instead, robust identification procedures are based on a priori assumptions on the *class* of systems and noise and on the a posteriori experimental data. Using this information robust system identification algorithms find nominal models based on the experimental data and worst-case identification error bounds over the set of models defined by the a priori information.

Information Consistency and Diameter of Information Due to the fact that the assumed a priori information is, in general, a quantification of the engineering common sense or simply a "leap of faith", there is no guarantee that it will be coherent with the a posteriori experimental data. Thus, robust identification procedures must always first test the *consistency* of both types of information.

Consistency can be better understood by considering the set of all possible models which could have produced the a posteriori data $\mathbf{y}$, in accordance with the class of systems S and the measurement noise $\eta \in \mathcal{N}$:

$$\mathcal{T}(\mathbf{y}) \doteq \{g \in S \mid \mathbf{y} = E(g, \eta), \eta \in \mathcal{N}\}$$

where $E(.,.)$ is the "experiment" operator. Intuitively, the a priori information and the a posteriori experimental data are consistent if there exists at least one element in S that could have generated the observed experimental data. This concept is formalized in the next definition:

Definition 5 The a priori information $(S, \mathcal{N})$ is *consistent* with the experimental a posteriori information $\mathbf{y}$ if and only if the set $\mathcal{T}(\mathbf{y})$ is nonempty.

Once consistency has been established, the computation of a nominal model and a valid model error bound can be attempted. There are two different types of algorithms to accomplish this. The first type of procedures [58,59] are guaranteed to converge, even when the information available is inconsistent. However, they might result on a nominal model outside the consistency set. The second type of procedures, and the type we use in the sequel, are interpolatory algorithms [60]. As we show next, these algorithms are always guaranteed to converge as the information is completed. Moreover, they are optimal to a factor of 2, in the sense that their worst-case error is never larger than twice the minimum achievable error over the set of all identification algorithms.

Worst Case Identification Error A salient feature of robust identification is its ability to provide worst-case bounds on the identification error. Given an identification algorithm $\mathcal{A}$ mapping the a priori and a posteriori information to candidate nominal model, its *local* error is defined as follows:

$$e(\mathcal{A}, \mathbf{y}) = \sup_{g \in \mathcal{T}(\mathbf{y})} m\left[g, \mathcal{A}(\mathbf{y}, S, \mathcal{N})\right] \tag{14}$$

that is, the maximum distance between the identified set and any other plant in the set $\mathcal{T}(\mathbf{y})$. Note that this error is related to the outcome of a specific experiment $\mathbf{y}$. A *global* error can be defined by considering the worst–case error over the set of all possible experimental outcomes:

Definition 6 The worst case global error of a given algorithm $\mathcal{A}(\mathbf{y}, S, \mathcal{N})$ is given by:

$$e(\mathcal{A}) = \sup_{\mathbf{y} \in \mathbf{Y}} e(\mathcal{A}, \mathbf{y}) \tag{15}$$

where $\mathbf{Y}$ is the set of all possible experimental data, consistent with sets S and $\mathcal{N}$.

Next we briefly review how to obtain mathematically tractable bounds for these errors. Recall that the set $\mathcal{T}(\mathbf{y}) \subset S$ is the smallest set of models that are indistinguishable from the view point of the input information. Therefore, roughly speaking, its size gives lower and upper bounds on the identification error defined above. In order to formalize these ideas and obtain computable bounds we need to introduce the following concepts:

Definition 7 The *radius* and *diameter* of a subset $\mathcal{A}$ of a metric space $(\mathcal{X}, m)$ are

$$r(\mathcal{A}) = \inf_{x \in \mathcal{X}} \sup_{a \in \mathcal{A}} m(x, a)$$

$$d(\mathcal{A}) = \sup_{x, a \in \mathcal{A}} m(x, a)\,.$$

The radius can be interpreted as the maximum error, measured in the metric $m(.)$, when considering the set $\mathcal{A}$ as represented by a single "central" point (which might not

belong to $\mathcal{A}$). The diameter is the maximum distance between any two points in the set. Based on these concepts of radius, we next quantify the "size" of the available information.

Definition 8 The radius and diameter of information are defined as:

$$\mathcal{R}(\mathcal{I}) \doteq \sup_{\mathbf{y}\in\mathbf{Y}} r[\mathcal{T}(\mathbf{y})]$$
$$\mathcal{D}(\mathcal{I}) \doteq \sup_{\mathbf{y}\in\mathbf{Y}} d[\mathcal{T}(\mathbf{y})]$$

where $\mathbf{Y}$ is the set of all possible experimental data consistent with the sets S and $\mathcal{N}$:

$$\mathbf{Y} \doteq \{E(g,\eta) \mid g \in S, \eta \in \mathcal{N}\}.$$

The following result gives worst–case bounds of the identification error based on these concepts:

Lemma 1 *The worst case identification error defined in (15) satisfies the following inequality:*

$$e(\mathcal{A}) \geq \mathcal{R}(\mathcal{I}) \geq \frac{1}{2}\mathcal{D}(\mathcal{I}) \tag{16}$$

for any algorithm $\mathcal{A}$. The following upper bound holds:

$$\mathcal{D}(\mathcal{I}) \geq e(\mathcal{A}_I) \tag{17}$$

for any interpolation algorithm $\mathcal{A}_I$.

The bounds above are of theoretical importance. For instance $\mathcal{R}(\mathcal{I})$ can be interpreted as an *intrinsic* error that cannot be decreased by any identification algorithm, unless extra information is added to the problem. On the other hand, these quantities are in general hard to compute. Fortunately, in practically relevant cases, they lead to mathematically tractable problems.

Definition 9 A set $\mathcal{A}$ in a linear space X is called symmetric if and only if there exists an element $c \in X$ such that for any $a \in X$ for which $c + a \in \mathcal{A}$ then $c - a \in \mathcal{A}$. The element c is called the symmetry point of set $\mathcal{A}$.

Lemma 2 *If the a priori sets S and $\mathcal{N}$ are symmetric and convex with respect to 0, and the experiment operator $E(g,\eta)$ is linear with respect to both g and η then the diameter of information satisfies:*

$$\mathcal{D}(\mathcal{I}) = \sup_{\mathbf{y}\in\mathbf{Y}} d\left[\mathcal{T}(\mathbf{y})\right] = d\left[\mathcal{T}(\mathbf{y}_0)\right], \quad \mathbf{y}_0 = E(0,0) \tag{18}$$

Furthermore,

$$d\left[\mathcal{T}(\mathbf{y}_0)\right] = 2 \sup_{g\in\mathcal{T}(\mathbf{y}_0)} m(g,0)\,. \tag{19}$$

Roughly speaking, the result above states that the experiment that yields the least amount of information is the one that results in a null outcome. Moreover, a bound on the worst case identification error is given by twice the maximum distance from any element in $\mathcal{T}(\mathbf{y_o})$ to the center of symmetry of S.

Time–Domain Based Interpolatory Identification Algorithms In this section we briefly review the properties of the specific identification algorithm, based on time–domain data, used in this paper to establish the existence of operators with the appropriate features. To this effect we need several preliminary results.

The first lemma considers the problem of the existence of a causal linear discrete-time invariant operator such that the first n terms of its transfer function are given:

Lemma 3 (Carathéodory–Fejér) *Given a matrix valued sequence $\{\mathbf{L}_i\}_{i=0}^{n-1}$, there exists a causal, discrete-time, LTI operator $L(z) \in \mathcal{BH}_\infty$ such that*

$$L(z) = \mathbf{L}_0 + \mathbf{L}_1 z + \mathbf{L}_2 z^2 + \ldots \mathbf{L}_{n-1} z^{n-1} + \ldots \tag{20}$$

if and only if

$$(\mathbf{T}_L^n)^{\mathrm{T}}\mathbf{T}_L^n \leq \mathbf{I} \tag{21}$$

where $\boldsymbol{I}$ denotes the identity matrix of compatible dimension.

Proof See for instance Chap. 1 in [61]. □

In the sequel we consider operator families of the form S:

$$S \doteq \{S(z) = H(z) + P(z)\} \tag{22}$$

where operators $S(z)$ are described in terms of a *non-parametric* component $H(z) \in \mathcal{BH}_\infty(K)$ and a *parametric component* $P(z)$. We will further assume that the parametric component $P(z)$ belongs to the following class $\mathcal{P}$ of affine operators:

$$\mathcal{P} \doteq \{P(z) = \mathbf{p}^{\mathrm{T}}\mathbf{G}_p(z),\ \mathbf{p} \in \mathcal{R}^{N_p}\}, \tag{23}$$

where the N_p components $\mathbf{G}_{p_i}(z)$ of vector $\mathbf{G}_p(z)$ are known, linearly independent, rational transfer functions.

The next lemma gives a necessary and sufficient condition for two finite vector sequences to be related by an operator in the family S.

Lemma 4 *Given a scalar K, and two vector sequences $(\mathbf{u},\mathbf{y})$, there exists an operator $S \in S$ such that $\mathbf{y} = S\mathbf{u}$ if and only if there exists a vector $\boldsymbol{h}$ satisfying:*

$$M(\mathbf{h}) \doteq \begin{bmatrix} \mathbf{I} & (\mathbf{T}_h^N)^{\mathrm{T}} \\ \mathbf{T}_h^N & \frac{1}{K^2} \end{bmatrix} \geq 0 \qquad \mathbf{y} = \mathbf{T}_u\mathbf{P}\mathbf{p} + \mathbf{T}_u\mathbf{h} \tag{24}$$

where $(\mathbf{P})_k \doteq [g_k^1\ g_k^2\ \cdots\ g_k^{N_p}]$, *with* g_k^i *denoting the kth Markov parameter of the ith transfer function* $G_{p_i}(z)$, h_k *the kth Markov parameter of the nonparametric component* $H(z)$, *respectively, and the scalar* K *is an upper bound of the* ℓ_2 *induced norm of* $H(z)$.

Moreover, in this case all such operators S *can be parametrized in terms of a free parameter* $Q(z) \in \mathcal{BH}_\infty$. *In particular, the choice* $Q(z) = 0$ *leads to the "central" model*

$$S_{\text{central}}(z) = H_o(z) + \mathbf{p}^{\mathrm{T}}\mathbf{G}_p(z)$$

where an explicit state–space realization of $H_o(z)$ *is given by:*

$$H_o(z) = \mathbf{C}_H\,(z\mathbf{I} - \mathbf{A}_H)^{-1}\,\mathbf{B}_H + \mathbf{D}_H$$

with

$$\begin{aligned}
\mathbf{A}_H &= \left\{\mathbf{A} - [\mathbf{C}_-^{\mathrm{T}}\mathbf{C}_- + (\mathbf{A}^{\mathrm{T}} - \mathbf{I})]^{-1}\mathbf{C}_-^{\mathrm{T}}\mathbf{C}_-(\mathbf{A} - \mathbf{I})\right\}^{-1}\\
\mathbf{B}_H &= [\mathbf{C}_-^{\mathrm{T}}\mathbf{C}_-(\mathbf{A}^{\mathrm{T}} - \mathbf{A} - \mathbf{I}) - (\mathbf{A}^{\mathrm{T}} - \mathbf{I})\mathbf{A}]^{-1}\mathbf{C}_-^{\mathrm{T}}\\
\mathbf{C}_H &= K\mathbf{C}_+ - K\mathbf{C}_+\left\{\mathbf{A} - [\mathbf{C}_-^{\mathrm{T}}\mathbf{C}_- + (\mathbf{A}^{\mathrm{T}} - \mathbf{I})]^{-1}\right.\\
&\qquad\qquad \left.\cdot\,\mathbf{C}_-^{\mathrm{T}}\mathbf{C}_-(\mathbf{A} - \mathbf{I})\right\}^{-1}\\
\mathbf{D}_H &= K\mathbf{C}_+\left\{[\mathbf{C}_-^{\mathrm{T}}\mathbf{C}_- + (\mathbf{A}^{\mathrm{T}} - \mathbf{I})]\right.\\
&\qquad\qquad \left.\cdot\,\mathbf{A} - \mathbf{C}_-^{\mathrm{T}}\mathbf{C}_-(\mathbf{A} - \mathbf{I})\right\}^{-1}\mathbf{C}_-^{\mathrm{T}}\,,
\end{aligned} \tag{25}$$

and

$$\mathbf{A} = \begin{bmatrix} 0 & \mathbf{I}_{N\times N} \\ 0 & 0 \end{bmatrix}, \quad \mathbf{C}_- = [\overbrace{1\quad 0\quad \ldots\quad 0}^{N+1}], \quad \mathbf{C}_+ = \frac{\mathbf{h}^{\mathrm{T}}}{K}. \tag{26}$$

Proof See Theorem 18.5.2 in [62] and [43]. □

Finally, the following corollary addresses the issue that real plants are subject to some unknown but bounded noise as represented in Fig. 12.

Corollary 1 ([43]) *Consider the problem of identifying an operator* $S \in \mathcal{S}$ *from measurements of its output* y *to a known input* u, *corrupted by additive bounded noise* η *in a given set* $\mathcal{N}$:

$$y_k = (S * u)_k + \eta_k\,, \quad k = 0, 1, \ldots, N\,. \tag{27}$$

Then there exist $S \in \mathcal{S}$ *that satisfies (27) if and only there exists a pair of vectors* $(\mathbf{h}, \mathbf{p})$ *such that* $M(\mathbf{h}) > 0$ *and* $\mathbf{y} - \mathbf{T}_u\mathbf{P}\mathbf{p} - \mathbf{T}_u\mathbf{h} \in \mathcal{N}$. *In that case, one such operator is given* $S_{central} = \mathbf{p}^{\mathrm{T}}\mathbf{G}_p + H_o$, *where* H_o *has the state–space realization (25).*

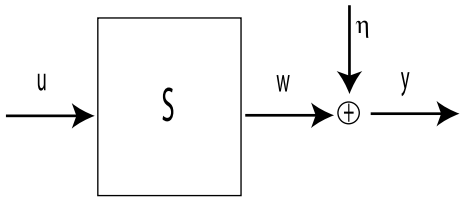

Motion Prediction for Continued Autonomy, Figure 12
Linear operator *S* with input *u* and output *y* corrupted with noise *η*

Bibliography

Primary Literature

1. Feddema J (1997) Microassembly of micro–electromechanical systems (MEMS) using visual servoing. In: Block Island Workshop on Vision and Control. Springer, Berlin, pp. 257–272
2. Ralis SJ, Vikramaditya B, Nelson BJ (2000) Micropositioning of a weakly calibrated microassembly system using coarse-to-fine visual servoing strategies. Trans IEEE Electron Packag Manuf 23(2):123–131
3. Ferreira A, Cassier C, Hirai S (2004) Automatic microassembly system assisted by vision servoing and virtual reality. IEEE Trans Mechatron 9(2):321–333
4. Xie H, Chen L, Sun L, Rong W (2005) Hybrid vision-force control for automatic assembly of miniaturized gear system. In: IEEE Int. Conf. on Robotics and Automation, Barcelona, Spain, April 2005, pp 1368–1373
5. Song Y, Li M, Sun L, Ji J (2005) Global visual servoing of miniature mobile robot inside a micro-assembly station. In: IEEE Int. Conf. on Mechatronics and Automation, Niagara Falls, Canada, July 2005, pp 1586–1591
6. Wang YF, Uecker DR, Wang Y (1996) Choreographed scope maneuvering in robotically–assisted laparoscopy with active vision guidance. In: 3rd IEEE Workshop on Applications of Computer Vision, Sarasota, December 1996
7. Krupa A, Gangloff J, Doignon C, MF de Mathelin, Morel G, Leroy J, Soler L, Marescaux J (2003) Autonomous 3-d positioning of surgical instruments in robotized laparoscopic surgery using visual servoing. IEEE Transn Robotics Autom 19(5):842–853
8. Nageotte F, Zanne P, Doignon C, de Mathelin M (2006) Visual servoing-based endoscopic path following for robot-assisted laparoscopic surgery. In: Int IEEE. Conf. on Intelligent Robots and Systems, Beijing, China, October 2006, pp 2364–2369
9. Hynes P, Dodds GI, Wilkinson AJ (2005) Uncalibrated visual-servoing of a dual-arm robot for surgical tasks. In: IEEE Int. Symposium on Computational Intelligence in Robotics and Automation, Espoo, Finland, June 2005, pp 151–156
10. Vitrani M, Morel G, Ortmaier T (2005) Automatic guidance of a surgical instrument with ultrasound based visual servoing. In: IEEE Int. Conf. on Robotics and Automation, Barcelona, Spain, April 2005, pp 508–513
11. Starner T, Pentland A (1998) Real-time american sign language recognition using desk and wearable computer based video. IEEE Trans Pattern Anal Mach Intell 20(12):1371–1375
12. Tsotsos JK, Verghese G, Dickinson S, Jenkin M, Jepson A, Milios E, Nuflo F, Stevenson S, Black M, Metaxas D, Culhane S, Ye Y, Mannn R (1998) PLAYBOT: A visually-guided robot for physically disabled children. Image Vis Comput 16(4):275–292

13. Song W, Kim J, Bien Z (2000) Visual servoing for human-robot interaction in the wheelchair-based rehabilitation robot. In: IEEE Int. Conf. on Systems, Man, and Cybernetics, October 2000, pp 1811–1816
14. Martens C, Ruchel N, Lang O, Ivlev O, Graser A (2001) A friend for assisting handicapped people. IEEE Robotics Autom Mag 8(1):57–65
15. Smith CE, Richards CA, Brandt SA, Papanikolopoulos NP (1996) Visual tracking for intelligent vehicle–highway systems. IEEE Trans Veh Tech 45(4):744–759
16. Taylor CJ, Kosecka J, Blasi R, Malik J (1999) Comparative study of vision-based lateral control strategies for autonomous highway driving. Intern J Robotics Res 18(5):442–453
17. Broggi A, Cellario M, Lombardi P, Porta M (2003) An evolutionary approach to visual sensing for vehicle navigation. IEEE Trans Ind Electron 50(1):18–29
18. Finnefrock M, Jiang X, Motai Y (2005) Visual-based assistance for electric vehicle driving. In: IEEE Intelligent Vehicles Symposium, IEEE June 2005, pp 656–661
19. Calabi E, Olver PJ, Shakiban C, Tannenbaum A, Haker S (1998) Differential and numerically invariant signature curves applied to object recognition. Intern J Comput Vis 26(2):107–135
20. Cohen LD (1991) On active contour models and balloons. Comput Vis Graph Image Process: Image Understanding 53(2):211–218
21. Coombs D, Brown C (1993) Real-time binocular smooth pursuit. Intern J Comput Vis 11(2):147–164
22. Grimson WEL, Stauffer C, Romano R, Lee L (1998) Using adaptive tracking to classify and monitor activities in a site. In: IEEE Computer Vision and Pattern Recognition, IEEE 1998, pp 22–29
23. Hager G, Belhumeur P (1997) Efficient region tracking with parametric models of geometry and illumination. IEEE Trans Pattern Anal Mach Intell 20(10):1025–1039
24. Irani M, Anandan P (1998) Unified approach to moving object detection in 2d and 3d scenes. IEEE Trans Pattern Anal Mach Intell 20(6):577–589
25. Shi J, Tomasi C (1994) Good features to track. In: IEEE Computer Vision and Pattern Recognition, IEEE 1994, pp 593–600
26. Black MJ, Jepson AD (1998) Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. Intern J Comput Vis 26(1):63–84
27. Fleet DJ, Black MJ, Yacoob Y, Jepson AD (2000) Design and use of linear models for image motion analyis. Intern J Comput Vis 36(3):171–194
28. Nayar SK, Murase H, Nene SA (1994) Learning, positioning, and tracking visual appearance. In: IEEE International Conference on Robotics and Automation, IEEE May 1994, pp 3237–3246
29. Wen CR, Azarbayejani A, Darrell T, Pentland AP (1997) Pfinder: real-time tracking of the human body. IEEE Trans Pattern Anal Mach Intell 19(7):780–785
30. Yacoob Y, Davis LS (2000) Learned models for estimation of rigid and articulated human motion from stationary or moving camera. Intern J Comput Vis 36(1):5–30
31. Orwell J, Remagnino P, Jones GA (1999) Multi-camera color tracking. In: 2nd IEEE Int. Workshop on Visual Surveillance, Fort Collins, CO, June 1999
32. Collins R, Amidi O, Kanade T (2002) An active camera system for acquiring multi-view video. In: Int. Conf. on Image Processing, vol I, IEEE pp 517–520
33. Hager G, Toyama K (2000) A new method for the nonlinear transformation of means and covariances in filters and estimators. IEEE Trans Autom Control 45(3):477–482
34. Reid ID, Murray W (1996) Active tracking of foveated feature clusters using affine structure. Intern J Comput Vis 18(1):41–60
35. Cipolla R, Blake A (1992) Surface shape from the deformation of apparent contours. Intern J Comput Vis 9(2):83–112
36. Blake A, Isard M (1998) Active Contours. Springer, Berlin
37. Isard M, Blake A (1998) CONDENSATION – conditional density propagation for visual tracking. Intern J Comput Vis 29(1):5–28
38. North B, Blake A, Isard M, Rittscher J (2000) Learning and classification of complex dynamics. IEEE Trans Pattern Anal Mach Intell 22(9):1016–1034
39. Julier S, Uhlmann J, Durrant-Whyte HF (1995) A new approach for filtering nonlinear systems. In: Proceedings of the (1995) American Control Conference, pp 1628–1632
40. Anderson BDO, Moore JB (1979) Optimal Filtering. Prentice Hall, New Jersey
41. Sánchez Peòna R, Sznaier M (1998) Robust Systems Theory and Applications. Wiley, New York
42. Bissacco A, Chiuso A, Ma Y, Soatto S (2001) Recognition of human gaits. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE December 2001
43. Parrilo PA, Sanchez Pena RS, Sznaier M (1999) A parametric extension of mixed time/frequency domain based robust identification. IEEE Trans Autom Contr 44(2):364–369
44. Boyd S, El Ghaoui L, Feron E, Balakrishnan V (1994) Linear Matrix Inequalities in System and Control Theory, vol 15. SIAM Studies in Applied Mathematics, Philadelphia
45. Bradski GR, Pisarevsky V (2000) Intel's computer vision library: applications in calibration, stereo, segmentation, tracking, gesture, face and object recognition. In: IEEE Computer Vision and Pattern Recognition, vol II. IEEE pp 796–797
46. Comaniciu D, Ramesh V, Meer P (2000) Real–time tracking of non–rigid objects using mean shift. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE June 2000, pp 142–149
47. Perez P, Hue C, Vermaak J, Gangnet M (2002) Color-based probabilistic tracking. In: 7th European Conference on Computer Vision, Copenhagen 2002, pp 661–675,
48. Zivkovic Z, Krose B (2004) An em-like algorithm for color-histogram-based object tracking. In: IEEE Computer Vision and Pattern Recognition, vol 1, IEEE June 2004, pp 798–803
49. Jepson AD, Fleet DJ, El-Maraghi TF (2003) Robust online appearance models for visual tracking. IEEE Trans Pattern Anal Mach Intell 25(10):1296–1311
50. Ho J, Lee KC, Yang MH, Kriegman D (2004) Visual tracking using learned linear subspaces. In: Int. Conference on Computer Vision and Pattern Recognition, vol 1. IEEE pp 782–789, June 2004, Washington D.C.
51. Lim H, Camps OI, Sznaier M, Morariu V (2006) Dynamic appearance modeling for human tracking. In: IEEE Computer Vision and Pattern Recognition, IEEE 2006, pp 751–757
52. Morariu V, Camps OI (2006) Modeling correspondences for multi-camera tracking using nonlinear manifold learning and target dynamics. In: IEEE Computer Vision and Pattern Recognition, IEEE pp 537–544
53. Morariu V, Camps O, Sznaier M, Lim H (2002) Robust cooperative visual tracking: A combined nonlinear dimensionality reduction/robust identification approach. In: Hirsch MJ, Pardalos PM, Murphey R, Grundel D (eds) Advances in Cooperative Control and Optimization. Springer, Berlin

54. Chen J, Gu G (2000) Control Oriented System Identification, An $\mathcal{H}_\infty$ Approach. Wiley, New York
55. Ljung L, Soderstrom T (1983) Theory and Practice of Recursive Identification. MIT Press, Cambridge
56. Ljung L (1996) Development of system identification. In: IFAC Congress, vol G, pp 141–146
57. Milanese M, Vicino A (1991) Optimal estimation theory for dynamic systems with set membership uncertainty: an overview. Automatica 27:997–1009
58. Helmicki AJ, Jacobson CA, Nett CN (1989) $\mathcal{H}_\infty$ identification of stable lsi systems: A scheme with direct application to controller design. In: American Contol Conference, Pittsburgh pp 1428–1434
59. Gu G, Khargonekar PP, Li Y (1992) Robust convergence of twostage nonlinear algorithms for system identification. Syst Control Lett 18:253–263
60. Chen J, Nett C, Fan M (1995) Worst-case system identification in $\mathcal{H}_\infty$: Validation of a Priori information, essentially optimal algorithms and error bounds. IEEE Trans Autom Control 40(7):1260–1265
61. Foias C, Frazho AE (1990) The commutant lifting approach to interpolation problems, Operator theory: Advances and Applications, vol 44. Birkhäuser, Basel
62. Ball J, Gohberg I, Rodman L (1990) Interpolation of Rational Matrix Functions, Operator Theory: Advances and Applications, vol 45. Birkhäuser, Basel

Books and Reviews

Chen J, Gu G (2000) Control Oriented System Identification, An $\mathcal{H}_\infty$ Approach. Wiley, New York. An excellent reference book with an in depth coverage of Robust Identification and the associated mathematical background in Interpolation Theory

Fisher RB, CV-Online: The Evolving, Distributed, Non-Proprietary, On-Line Compendium of Computer Vision, Online Book, 2007 Provides a very complete online, evolving hypertext summary on central topics in computer vision, including motion and tracking

Forsyth DA, Ponce J (2003) Computer Vision: A Modern Approach, Prentice Hall, Prentice Hall, Upper Saddle River. A textbook covering the fundamentals of computer vision. Chapter 4 includes an introduction to the problem of tracking using linear models

Ma Y, Soatto S, Kosecka J, J Sastry, Sastry SS (2005) An Invitation to 3-D Vision, From Images to Geometric Models. Springer, Berlin. Chapter 12 is dedicated to the topic of visual feedback including applications to autonomous navigation

Medioni G, Kang SB (2005) Emerging Topics is Computer Vision, Prentice Hall, Prentice Hall, Upper Saddle River. Chapter 11 provides a tutorial on the open source computer vision library OpenCV

Paoletti S, Juloski A, Ferrari-Trecate G, Vidal R (2007) Identification of Hybrid Systems: A Tutorial. Eur J Control 13(2–3). A survey paper covering the fundamentals of identification of piecewise affine models

Sánchez Peña R, Sznaier M (1998) Robust Systems Theory and Applications. Wiley, New York. Chapter 10 of this textbook provides a good introduction to the field of Robust Identification. In: addition, the Appendices provide a summary of several key results in Linear Systems Theory

Sznaier M, Camps O (2007) Systems Theoretic Methods in Computer Vision and Image Processing. J Soc Inst Contr Eng. (SICE) Special Issue on Control Theoretic Principles in Emerging Technologies 46:p 206, A survey paper that covers the use of systems theoretic tools to solve multiple problems arising in the context of dynamic vision and image processing, e. g. tracking, motion segmentation, structure from motion, activity recognition, texture modeling and recognition, static and dynamic inpainting, etc

Movement Coordination

Armin Fuchs[1,2], James A. S. Kelso[1]
[1] Center for Complex Systems & Brain Sciences, Florida Atlantic University, Boca Raton, USA
[2] Department of Physics, Florida Atlantic University, Boca Raton, USA

Article Outline

Glossary
Definition of the Subject
Introduction
The Basic Law of Coordination: Relative Phase
Stability: Perturbations and Fluctuations
The Oscillator Level
Breaking and Restoring Symmetries
Conclusions
Extensions of the HKB Model
Future Directions
Acknowledgment
Bibliography

Glossary

Control parameter A parameter of internal or external origin that when manipulated controls the system in a nonspecific fashion and is capable of inducing changes in the system's behavior. These changes may be a smooth function of the control parameter, or abrupt at certain critical values. The latter, also referred to as phase transitions, are of main interest here as they only occur in nonlinear systems and are accompanied by phenomena like critical slowing down and fluctuation enhancement that can be probed for experimentally.

Haken–Kelso–Bunz (HKB) model First published in 1985, the HKB model is the best known and probably most extensively tested quantitative model in human movement behavior. In its original form it describes the dynamics of the relative phase between two oscillating fingers or limbs under frequency scaling. The HKB model can be derived from coupled nonlinear os-

cillators and has been successfully extended in various ways, for instance, to situations where different limbs like an arm and a leg, a single limb and a metronome, or even two different people are involved.

Order parameter Order parameters are quantities that allow for a usually low-dimensional description of the dynamical behavior of a high-dimensional system on a macroscopic level. These quantities change their values abruptly when a system undergoes a phase transition. For example, density is an order parameter in the ice to water, or water to vapor transitions. In movement coordination the most-studied order parameter is relative phase, i. e. the difference in the phases between two or more oscillating entities.

Phase transition The best-known phase transitions are the changes from a solid to a fluid phase like ice to water, or from fluid to gas like water to vapor. These transitions are called first-order phase transitions as they involve latent heat, which means that a certain amount of energy has to be put into the system at the transition point that does not cause an increase in temperature. For the second-order phase transitions there is no latent heat involved. An example from physics is heating a magnet above its Curie temperature at which point it switches from a magnetic to a nonmagnetic state. The qualitative changes that are observed in many nonlinear dynamical systems when a parameter exceeds a certain threshold are also such second-order phase transitions.

Definition of the Subject

Movement Coordination is present all the time in daily life but tends to be taken for granted when it works. One might say it is quite an arcane subject also for science. This changes drastically when some pieces of the locomotor system are not functioning properly because of injury, disease or age. In most cases it is only then that people become aware of the complex mechanisms that must be in place to control and coordinate the hundreds of muscles and joints in the body of humans or animals to allow for maintaining balance while maneuvering through rough terrains, for example. No robot performance comes even close in such a task.

Although these issues have been around for a long time it was only during the last quarter century that scientists developed quantitative models for movement coordination based on the theory of nonlinear dynamical systems. Coordination dynamics, as the field is now called, has become arguably the most developed and best tested quantitative theory in the life sciences.

More importantly, even though this theory was originally developed for modeling of bimanual finger movements, it has turned out to be universal in the sense that it is also valid to describe the coordination patterns observed between different limbs, like an arm and a leg, different joints within a single limb, like the wrist and elbow, and even between different people that perform movements while watching each other.

Introduction

According to a dictionary definition: *Coordination* is the act of coordinating, making different people or things work together for a goal or effect.

When we think about *movement coordination* the "things" we make work together can be quite different like our legs for walking, fingers for playing the piano, mouth, tongue and lips for articulating speech, body expressions and the interplay between bodies in dancing and ballet, tactics and timing between players in team sports and so on, not to forget other advanced skill activities like skiing or golfing.

All these actions have one thing in common: they look extremely easy if performed by people who have learned and practiced these skills, and they are incredibly difficult for novices and beginners. Slight differences might exist regarding how these difficulties are perceived, for instance when asked whether they can play golf some people may say: "I don't know, let me try", and they expect to out-drive Tiger Woods right away; there are very few individuals with a similar attitude toward playing the piano.

The physics of golf as far as the ball and the club is concerned is almost trivial: hit the ball with the highest possible velocity with the club face square at impact, and it will go straight and far. The more tricky question is how to achieve this goal with a body that consists of hundreds of different muscles, tendons and joints, and, importantly, their sensory support in joint, skin and muscle receptors (proprioception), in short, hundreds of degrees of freedom. How do these individual elements work together, how are they coordinated? Notice, the question is not how do *we* coordinate them? None of the skills mentioned above can be performed by consciously controlling all the body parts involved. Conscious thinking sometimes seems to do more harm than good. So how do they/we do it? For some time many scientists sought the answer to this question in what is called *motor programs* or, more recently, *internal models*. The basic idea is straightforward: when a skill is learned it is somehow stored in the brain like a program in a computer and simply can be called and executed when needed. Additional learning or train-

ing leads to skill improvement, interpreted as refinements in the program. As intuitive as this sounds and even if one simply ignores all the unresolved issues like how such programs gain the necessary flexibility or in what form they might be stored in the first place, there are even deeper reasons and arguments suggesting that humans (or animals for that matter) don't work like that. One of the most striking of these arguments is known as motor equivalence: everybody who has learned to write with one of their hands can immediately write with the foot as well. This writing may not look too neat, but it will certainly be readable and represents the transfer of a quite complex and difficult movement from one end-effector (the hand) to another (the foot) that is controlled by a completely different set of muscles and joints. Different degrees of freedom and redundancy in the joints can still produce the same output (the letters) immediately, i. e. without any practice.

For the study of movement coordination a most important entry point is to look at situations where the movement or coordination pattern changes abruptly. An example might be the well-known gait switches from walk to trot to gallop that horses perform. It turns out, however, that switching among patterns of coordination is a ubiquitous phenomenon in human limb movements. As will be described in detail, such switching has been used to probe human movement coordination in quantitative experiments.

It is the aim of this article to describe an approach to a quantitative modeling of human movements, called coordination dynamics, that deals with quantities that are accessible from experiments and makes predictions that can and have been tested. The intent is to show that coordination dynamics represents a theory allowing for quantitative predictions of phenomena in a way that is unprecedented in the life sciences. In parallel with the rapid development of noninvasive brain imaging techniques, coordination dynamics has even pointed to new ways for the study of brain functioning.

The Basic Law of Coordination: Relative Phase

The basic experiment, introduced by one of us [27,28], that gave birth to coordination dynamics, the theory underlying the coordination of movements, is easily demonstrated and has become a classroom exercise for generations of students: if a subject is moving the two index fingers in so-called anti-phase, i. e. one finger is flexing while the other is extending, and then the movement rate is increased, there is a critical rate where the subject switches spontaneously from the anti-phase movement to in-phase, i. e. both fingers are now flexing and extending at the same time. On the other hand, if the subject starts at a high or low rate with an in-phase movement and the rate is slowed down or sped up, no such transition occurs.

These experimental findings can be translated or mapped into the language of dynamical systems theory as follows [19]:

- At low movement rates the system has two stable attractors, one representing anti-phase and one for in-phase – in short: the system is bistable;
- When the movement rate reaches a critical value, the anti-phase attractor disappears and the only possible stable movement pattern remaining is in-phase;
- There is strong hysteresis: when the system is performing in-phase and the movement rate is decreased from a high value, the anti-phase attractor may reappear but the system does not switch to it.

In order to make use of dynamical systems theory for a quantitative description of the transitions in coordinated movements, one needs to establish a measure that allows for a formulation of a dynamical system that captures these experimental observations and can serve as a phenomenological model. Essentially, the finger movements represent oscillations (as will be discussed in more detail in Subsect. "Oscillators for Limb Movements") each of which is described by an amplitude r and a phase $\varphi(t)$. For the easiest case of harmonic oscillations the amplitude r does not depend on time and the phase increases linearly with time at a constant rate ω, called the angular velocity, leading to $\varphi(t) = \omega t$. Two oscillators are said to be in the in-phase mode if the two phases are the same, or $\varphi_1(t) - \varphi_2(t) = 0$, and in anti-phase if the difference between their two phases is 180° or π radians. Therefore, the quantity that is most commonly used to model the experimental findings in movement coordination is the phase difference or *relative phase*

$$\phi(t) = \varphi_1(t) - \varphi_2(t) = \begin{cases} \phi(t) = 0 & \text{for in-phase} \\ \phi(t) = \pi & \text{for anti-phase}\,. \end{cases} \tag{1}$$

The minimal dynamical system for the relative phase that is consistent with observations is known as the Haken–Kelso–Bunz (or HKB) model and was first published in a seminal paper in 1985 [19]

$$\dot{\phi} = -a \sin\phi - 2b \sin 2\phi \quad \text{with} \quad a, b \geq 0\,. \tag{2}$$

As is the case for all one-dimensional first order differential equations, (2) can be derived from a potential function

$$\dot{\phi} = -\frac{\mathrm{d}V(\phi)}{\mathrm{d}\phi} \quad \text{with} \quad V(\phi) = -a\cos\phi - b\cos 2\phi\,. \tag{3}$$

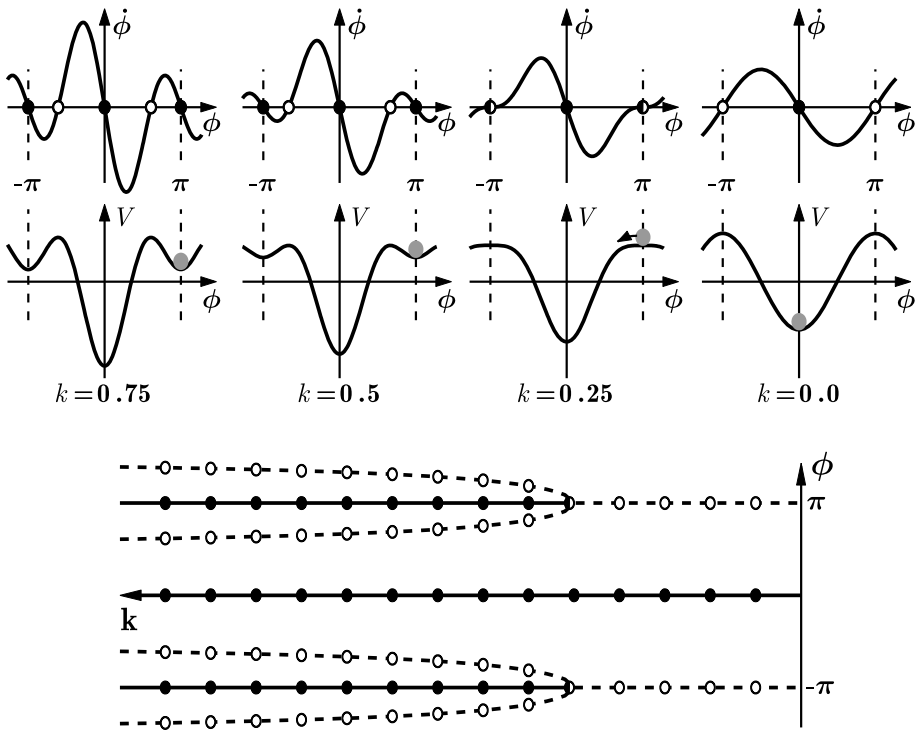

Movement Coordination, Figure 1
Dynamics of the HKB model at the coordinative, relative phase (ϕ) level as a function of the control parameter $k = \frac{b}{a}$. *Top row*: Phase space plots $\dot{\phi}$ as a function of ϕ. *Middle*: Landscapes of the potential function $V(\phi)$. *Bottom*: Bifurcation diagram, where *solid lines* with *filled circles* correspond to stable fixed points (attractors) and *dashed lines* with *open circles* denote repellers. Note that k increases from right ($k = 0$) to left ($k = 0.75$)

One of the two parameters a and b that appear in (2) and (3) can be eliminated by introducing a new time scale $\tau = \alpha t$, a procedure known as scaling and commonly used within the theory of nonlinear differential equations, leading to

$$\begin{aligned} \dot{\phi}(t) = \frac{\mathrm{d}\phi(t)}{\mathrm{d}t} \quad &\rightarrow \quad \frac{\mathrm{d}\phi\left(\frac{\tau}{\alpha}\right)}{\mathrm{d}\frac{\tau}{\alpha}} \\ &= -a \sin\phi\left(\frac{\tau}{\alpha}\right) - 2b \sin 2\phi\left(\frac{\tau}{\alpha}\right) \qquad (4) \\ \alpha \frac{\mathrm{d}\tilde{\phi}(\tau)}{\mathrm{d}\tau} &= -a \sin\tilde{\phi}(\tau) - 2b \sin 2\tilde{\phi}(\tau) \end{aligned}$$

where $\tilde{\phi}$ has the same shape as ϕ, it is just changing on a slower or faster time scale depending on whether α is bigger or smaller than 1. After dividing by α and letting the so far undetermined $\alpha = a$ (4) becomes

$$\frac{\mathrm{d}\tilde{\phi}}{\mathrm{d}\tau} = -\underbrace{\frac{a}{\alpha}}_{=1} \sin\tilde{\phi} - 2\underbrace{\frac{b}{\alpha}}_{=k} \sin 2\tilde{\phi}\,. \qquad (5)$$

Finally, by dropping the tilde ˜ (2) and (3) can be written with only one parameter $k = \frac{b}{a}$ in the form

$$\begin{aligned} \dot{\phi} &= -\sin\phi - 2k \sin 2\phi \\ &= -\frac{\mathrm{d}V(\phi)}{\mathrm{d}\phi} \quad \text{with} \quad V(\phi) = -\cos\phi - k\cos 2\phi\,. \qquad (6) \end{aligned}$$

The dynamical properties of the HKB model's *collective* or *coordinative* level of description are visualized in Fig. 1 with plots of the phase space ($\dot{\phi}$ as a function of ϕ) in the top row, the potential landscapes $V(\phi)$ in the second row and the bifurcation diagram at the bottom. The control parameter k, as shown, is the ratio between b and a, $k = \frac{b}{a}$, which is inversely related to the movement rate: a large value of k corresponds to a slow rate, whereas k close to zero indicates that the movement rate is high.

In the phase space plots (Fig. 1 top row) for $k = 0.75$ and $k = 0.5$ there exist two stable fixed points at $\phi = 0$ and $\phi = \pi$ where the function crosses the horizontal axis with a negative slope, marked by solid circles (the fixed point at $-\pi$ is the same as the point at π as the function is 2π-periodic). These attractors are separated by repellers,

zero crossings with a positive slope and marked by open circles. For the movement rates corresponding to these two values of k the model suggests that both anti-phase and in-phase movements are stable. When the rate is increased, corresponding to a decrease in the control parameter k down to the critical point at $k_c = 0.25$ the former stable fixed point at $\phi = \pi$ collides with the unstable fixed point and becomes neutrally stable indicated by a half-filled circle. Beyond k_c, i. e. for faster rates and smaller values of k the anti-phase movement is unstable and the only remaining stable coordination pattern is in-phase.

The potential functions, shown in the second row in Fig. 1, contain the same information as the phase space portraits as they are just a different representation of the dynamics. However, the strong hysteresis is more intuitive in the potential landscape than in phase space, and can best be seen through an experiment that starts out with slow movements in anti-phase (indicated by the gray ball in the minimum of the potential at $\phi = \pi$) and increasing rate. After passing the critical value $k_c = 0.25$ the slightest perturbation will put the ball on the downhill slope and initiate a switch to in-phase. If the movement is now slowed down again, going from right to left in the plots, even though the minimum at $\phi = \pi$ reappears, the ball cannot jump up and occupy it but will stay in the deep minimum at $\phi = 0$, a phenomenom known as hysteresis.

Finally, a bifurcation diagram is shown at the bottom of Fig. 1, where the locations of stable fixed points for the relative phase ϕ are plotted as solid lines with solid circles and unstable fixed points as dashed lines with open circles. Around $k_c = 0.25$ the system undergoes a subcritical pitchfork bifurcation. Note that the control parameter k in this plot increases from right to left.

Evidently, the dynamical system represented by (2) is capable of reproducing the basic experimental findings listed above. From the viewpoint of theory, this is simply one of the preliminaries for a model that have to be fulfilled. In general, any model that only reproduces what is built into it is not of much value. More important are crucial experimental tests of the consequences and additional phenomena that are predicted when the model is worked through. Several such consequences and predictions will be described in detail in the following sections. It is only after such theoretical and experimental scrutiny that the HKB model has come to qualify as an elementary law of movement coordination.

Stability: Perturbations and Fluctuations

Random fluctuations, or noise for short, exist in all systems that dissipate energy. In fact, there exists a famous theorem that goes back to Einstein, known as the dissipation-fluctuation theorem, which states that the amount of random fluctuations in a system is proportional to its dissipation of energy. There are effects from random noise on the dynamics of relative phase that can be predicted from theory both qualitatively and quantitatively, allowing for the HKB model's coordination level to be tested experimentally. Later the individual component level will be discussed.

An essential difference between the dynamical systems approach to movement coordination and the motor program or internal model hypotheses is most distinct in regions where the coordination pattern undergoes a spontaneous qualitative change as in the switch from anti-phase to in-phase in Kelso's experiment. From the latter point of view, these switches simply happen, very much like in the automatic transmission of a car: whenever certain criteria are fulfilled, the transmission switches from one gear to another. It is easy to imagine a similar mechanism to be at work and in control of the transitions in movements: as soon as a certain rate is exceeded, the anti-phase program is somehow replaced by the in-phase module, which is about all we can say regarding the mechanism of switching. On the other hand, by taking dynamic systems theory seriously, one can predict and test phenomena accompanying second-order phase transitions. Three of these phenomena, namely, critical slowing down, enhancement of fluctuations and critical fluctuations will be discussed here in detail.

For a quantitative treatment it is advantageous to expand $\dot{\phi}$ and $V(\phi)$ in (6) into Taylor series around the fixed point $\phi = \pi$ and truncate them after the linear and quadratic terms, respectively

$$\begin{aligned}
\dot{\phi} &= -\sin\phi - 2k\sin 2\phi \\
&= -\{-(\phi-\pi)+\ldots\} - 2k\{2(\phi-\pi)+\ldots\} \\
&\approx (1-4k)(\phi-\pi) \\
V(\phi) &= -\cos\phi - k\cos 2\phi \\
&= -\{-1+(\phi-\pi)^2+\ldots\} \\
&\quad - k\{1-4(\phi-\pi)^2+\ldots\} \\
&\approx 1-k-(1-4k)(\phi-\pi)^2 \,.
\end{aligned} \tag{7}$$

A typical situation that occurs when a system approaches and passes through a transition point is shown in Fig. 2. In the top row the potential function for $\phi \geq 0$ is plotted (dashed line) together with its expansion around the fixed point $\phi = \pi$ (solid). The bottom row consists of plots of time series showing how the fixed point is or is not approached when the system is initially at $\phi = \pi + \Delta$. The phenomena accompanying second-order phase tran-

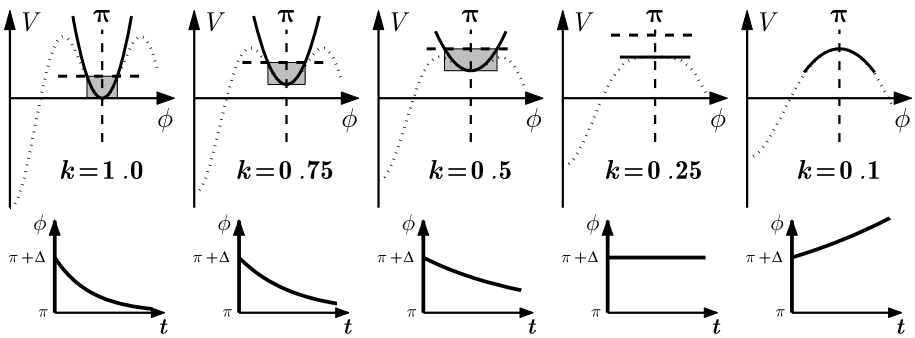

Movement Coordination, Figure 2

Hallmarks of a system that approaches a transition point: enhancement of fluctuations, indicated by the increasing size of the *shaded area*; critical slowing down shown by the time it takes for the system to recover from a perturbation (*bottom*); critical fluctuations occur where the top of the shaded area is higher than the closest maximum in the potential, initiating a switch even though the system is still stable

sitions in a system that contains random fluctuations can be best described by Fig. 2.

Critical slowing down corresponds to the time it takes the system to recover from a small perturbation Δ. In the vicinity of the fixed point the dynamics can be described by the linearization of the nonlinear equation around the fixed point (7). Such a linear equation can be readily solved leading to

$$\phi(t) = \pi + \Delta e^{(1-4k)t} .$$

As long as k is larger than its critical value $k_c = 0.25$ the exponent is negative and a perturbation will decay exponentially in time. However, as the system approaches the transition point, this decay will take longer and longer as shown in the bottom row in Fig. 2. At the critical parameter $k = 0.25$ the system will no longer return to the former stable fixed point and beyond that value it will even move away from it. In the latter parameter region the linear approximation is no longer valid. Critical slowing down can be and has been tested experimentally by perturbing a coordination state and measuring the relaxation constant as a function of movement rate prior to the transition. The experimental findings [31,44,45] are in remarkable agreement with the theoretical predictions of coordination dynamics [43].

Enhancement of fluctuations is to some extent the stochastic analog to critical slowing down. The random fluctuations that exist in all dissipative systems are a stochastic force that kicks the system away from the minimum and (on average) up to a certain elevation in the potential landscape, indicated by the shaded areas in Fig. 2. For large values of k the horizontal extent of this area is small but becomes larger and larger when the transition point is approached. Assuming that the strength of the random force does not change with the control parameter, the standard deviation of the relative phase is a direct measure of this enhancement of fluctuations and will be increasing when the control parameter is moving towards its critical value. Again experimental tests are in detailed agreement with the stochastic version of the HKB model [30,43,44].

Critical fluctuations can induce transitions even when the critical value of the control parameter has not been reached. As before, random forces will kick the system around the potential minimum and up to (on average) a certain elevation. If this height is larger than the hump it has to cross, as is the case illustrated in Fig. 2 for $k = 0.5$, a transition will occur, even though the fixed point is still classified as stable. In excellent agreement with theory, such critical fluctuations were observed in the original experiments by Kelso and colleagues [30] and have been found in a number of related experimental systems [31,42].

All these hallmarks point to the conclusion that transitions in movement coordination are not simply a switching of gears but take place in a well defined way via the instability of a former stable coordination state. Such phenomena are also observed in systems in physics and other disciplines where in situations far from thermal equilibrium macroscopic patterns emerge or change, a process termed self-organization. A general theory of self-organizing systems, called synergetics [17,18], was formulated by Hermann Haken in the early 1970s.

The Oscillator Level

The foregoing description and analysis of bimanual movement coordination takes place on the coordinative or collective level of relative phase. Looking at an actual experiment, there are two fingers moving back and forth and one may ask whether it is possible to find a model on the level of the oscillatory components from which the dynamics of the relative phase can then be derived. The challenge for such an endeavor is at least twofold: first, one needs a dynamical system that accurately describes the movements of the individual oscillatory components (the fingers). Second, one must find a coupling function for these components that leads to the correct relation for the relative phase (2).

Oscillators for Limb Movements

In terms of oscillators there is quite a variety to choose from as most second order systems of the form

$$\ddot{x} + \gamma\dot{x} + \omega^2 x + N(x, \dot{x}) = 0 \tag{8}$$

are potential candidates. Here ω is the angular frequency, γ the linear damping constant and $N(x, \dot{x})$ is a function containing nonlinear terms in x and $\dot{x}$.

Best known and most widely used are the harmonic oscillators, where $N(x, \dot{x}) = 0$, in particular for the case without damping $\gamma = 0$. In the search for a model to describe human limb movements, however, harmonic oscillators are not well suited, because they do not have stable limit cycles. The phase space portrait of an harmonic oscillator is a circle (or ellipse), but only if it is not perturbed. If such a system is slightly kicked off the trajectory it is moving on, it will not return to its original circle but continue to move on a different orbit. In contrast, it is well known that if a rhythmic human limb movement is perturbed, this perturbation decreases exponentially in time and the movement returns to its original trajectory, a stable limit cycle, which is an object that exists only for nonlinear oscillators [25,26].

Obviously, the amount of possible nonlinear terms to choose from is infinite and at first sight, the task to find the appropriate ones is like looking for a needle in a haystack. However, there are powerful arguments that can be made from both theoretical reasoning and experimental findings that restrict the nonlinearities, as we shall see, to only two. First, we assume that the function $N(x, \dot{x})$ takes the form of a polynomial in x and $\dot{x}$ and that this polynomial is of the lowest possible order. So the first choice would be to assume that N is quadratic in x and $\dot{x}$ leading to an oscillator of the form

$$\ddot{x} + \gamma\dot{x} + \omega^2 x + ax^2 + b\dot{x}^2 + cx\dot{x} = 0\,. \tag{9}$$

How do we decide whether (9) is a good model for rhythmic finger movements? If a finger is moved back and forth, that is, performs an alternation between flexion and extension, then this process is to a good approximation symmetric: flexion is the mirror image of extension. In the equations a mirror operation is carried out by substituting x by $-x$, and, in doing so, the equation of motion must not change for symmetry to be preserved. Applied to (9) this leads to

$$\begin{aligned} &-\ddot{x} + \gamma(-\dot{x}) + \omega^2(-x) + a(-x)^2 + b(-\dot{x})^2 \\ &\quad + c(-x)(-\dot{x}) = 0 \\ &-\ddot{x} - \gamma\dot{x} - \omega^2 x + ax^2 + b\dot{x}^2 + cx\dot{x} = 0 \\ &\ddot{x} + \gamma\dot{x} + \omega^2 x - ax^2 - b\dot{x}^2 - cx\dot{x} = 0 \end{aligned} \tag{10}$$

where the last equation in (10) is obtained by multiplying the second equation by -1. It is evident that this equation is not the same as (9). In fact, it is only the same if $a = b = c = 0$, which means that there must not be any quadratic terms in the oscillator equation if one wants to preserve the symmetry between flexion and extension phases of movement. The argument goes even further: $N(x, \dot{x})$ must not contain any terms of even order in x and $\dot{x}$ as all of them, like the quadratic ones, would break the required symmetry. It is easy to convince oneself that as far as the flexion-extension symmetry is concerned all odd terms in x and $\dot{x}$ are fine.

There are four possible cubic terms, namely $\dot{x}^3$, $\dot{x}x^2$, $x\dot{x}^2$ and x^3 leading to a general oscillator equation of the form

$$\ddot{x} + \gamma\dot{x} + \omega^2 x + \delta\dot{x}^3 + \epsilon\dot{x}x^2 + ax^3 + bx\dot{x}^2 = 0\,. \tag{11}$$

The effects that these nonlinear terms exert on the oscillator dynamics can be best seen by rewriting (11) as

$$\ddot{x} + \dot{x}\underbrace{\{\gamma + \epsilon x^2 + \delta\dot{x}^2\}}_{\text{damping}} + x\underbrace{\{\omega^2 + ax^2 + b\dot{x}^2\}}_{\text{frequency}} = 0 \tag{12}$$

which shows that the terms $\dot{x}^3$ and $\dot{x}x^2$ are position and velocity dependent changes to the damping constant γ, whereas the nonlinearities x^3 and $x\dot{x}^2$ mainly influence the frequency. As the nonlinear terms were introduced to obtain stable limit cycles and the main interest is in amplitude and not frequency, we will let $a = b = 0$, which reduces the candidate oscillators to

$$\ddot{x} + \dot{x}\{\gamma + \epsilon x^2 + \delta\dot{x}^2\} + \omega^2 x = 0\,. \tag{13}$$

Nonlinear oscillators with either $\delta = 0$ or $\epsilon = 0$ have been studied for a long time and have been termed in the literature as van-der-Pol and Rayleigh oscillators, respectively.

Systems of the form (13) only show sustained oscillations on a stable limit cycle within certain ranges of the parameters, as can be seen easily for the van-der-Pol oscillator, given by (13) with $\delta = 0$

$$\ddot{x} + \dot{x}\underbrace{\{\gamma + \epsilon x^2\}}_{\tilde{\gamma}} + \omega^2 x = 0\,. \tag{14}$$

The underbraced term in (14) represents the effective damping constant, $\tilde{\gamma}$, now depending on the square of the displacement, x^2, a quantity which is non-negative. For the parameters γ and ϵ one can distinguish the following four cases:

$\boldsymbol{\gamma > 0, \epsilon > 0}$ The effective damping $\tilde{\gamma}$ is always positive. The trajectories are evolving towards the origin, which is a stable fixed point.

$\boldsymbol{\gamma < 0, \epsilon < 0}$ The effective damping $\tilde{\gamma}$ is always negative. The system is unstable and the trajectories are evolving towards infinity.

$\boldsymbol{\gamma > 0, \epsilon < 0}$ For small values of the amplitude x^2 the effective damping $\tilde{\gamma}$ is positive leading to even smaller amplitudes. For large values of x^2 the effective damping $\tilde{\gamma}$ is negative leading to a further increase in amplitude. The system evolves either towards the fixed point or towards infinity depending on the initial conditions.

$\boldsymbol{\gamma < 0, \epsilon > 0}$ For small values of the amplitude x^2 the effective damping $\tilde{\gamma}$ is negative leading to an increase in amplitude. For large values of x^2 the effective damping $\tilde{\gamma}$ is positive and decreases the amplitude. The system evolves towards a stable limit cycle.

The main features for the van-der-Pol oscillator are shown in Fig. 3 with the time series (left), the phase space portrait (middle) and the power spectrum (right). The time series is not a sine function but has a fast rising increasing flank and a more shallow slope on the decreasing side. Such time series are called relaxation oscillations. The trajectory in phase space is closer to a rectangle than to a circle and the power spectrum shows pronounced peaks at the fundamental frequency ω and its odd higher harmonics ($3\omega, 5\omega, \ldots$).

In contrast to the van-der-Pol case the damping constant $\tilde{\gamma}$ for the Rayleigh oscillator, the case $\epsilon = 0$ in (13), depends on the square of the velocity $\dot{x}^2$. Arguments similar to those above lead to the conclusion that the Rayleigh oscillator shows sustained oscillations for parameters $\gamma < 0$ and $\delta > 0$.

As shown in Fig. 4 the time series and trajectories of the Rayleigh oscillator also exhibit relaxation behavior, but in this case with a slow rise and fast drop. As for the van-der-Pol, the phase space portrait is almost rectangular but the long and short axes are switched. Again the power spectrum has peaks at the fundamental frequency and contains odd higher harmonics.

Evidently, taken by themselves neither the van-der-Pol nor Rayleigh oscillators are good models for human limb movement for at least two reasons, even though they fulfill one requirement for a model: they have stable limit cycles. First, human limb movements are almost sinusoidal and their trajectories have a circular or elliptical shape. Second, it has also been found in experiments with human subjects performing rhythmic limb movements that when the movement rate is increased the amplitude of the movement decreases linearly with frequency [25]. It can be shown that for the van-der-Pol oscillator the amplitude is independent of frequency and for the Rayleigh it decreases proportional to ω^{-2}, both in disagreement with the experimental findings.

It turns out that a combination of the van-der-Pol and Rayleigh oscillator, termed the hybrid oscillator of the form (13) fulfills all the above requirements if the parameters are chosen as $\gamma < 0$ and $\epsilon \approx \delta > 0$.

As shown in Fig. 5 the time series for the hybrid oscillator is almost sinusoidal and the trajectory is elliptical. The power spectrum has a single peak at the fundamental frequency. Moreover, the relation between the amplitude and frequency is a linear decrease in amplitude when the rate is increased as shown schematically in Fig. 6. Taken together, the hybrid oscillator is a good approximation for the trajectories observed experimentally in human limb movements.

The Coupling

As pointed out already, in a second step one has to find a coupling function between two hybrid oscillators that leads to the correct dynamics for the relative phase (2). The most common realization of a coupling between two oscillators is a spring between two pendulums, leading to a force proportional to the difference in locations $f_{12} = k[x_1(t) - x_2(t)]$. It can easily be shown, that such a coupling does not lead to the required dynamics on the relative phase level. In fact, several coupling terms have been suggested that do the trick, but none of them is very intuitive. The arguably easiest form, which is one of the possible couplings presented in the original HKB model [19], is given by

$$f_{12} = (\dot{x}_1 - \dot{x}_2)\left\{\alpha + \beta(x_1 - x_2)^2\right\}\,. \tag{15}$$

Combined with two of the hybrid oscillators, the dynamical system that describes the transition from anti-phase to

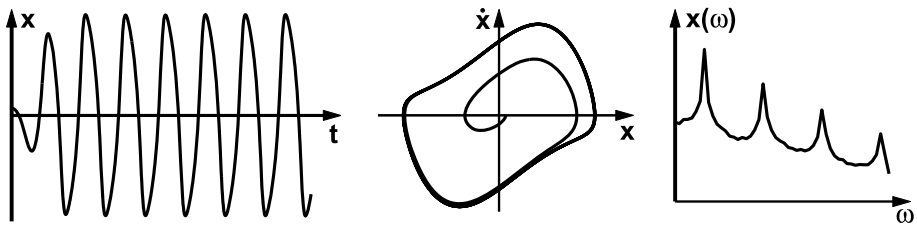

Movement Coordination, Figure 3
The van-der-Pol oscillator: time series (*left*), phase space trajectory (*middle*) and power spectrum (*right*)

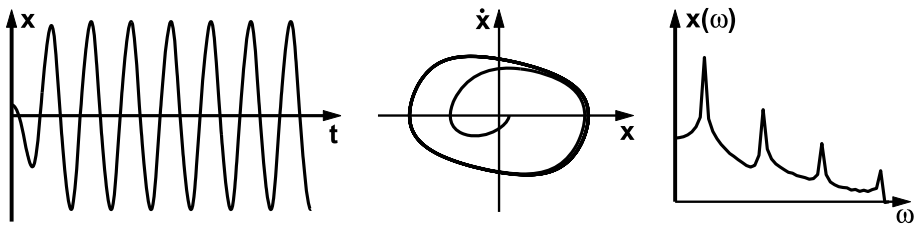

Movement Coordination, Figure 4
The Rayleigh oscillator: time series (*left*), phase space trajectory (*middle*) and power spectrum (*right*)

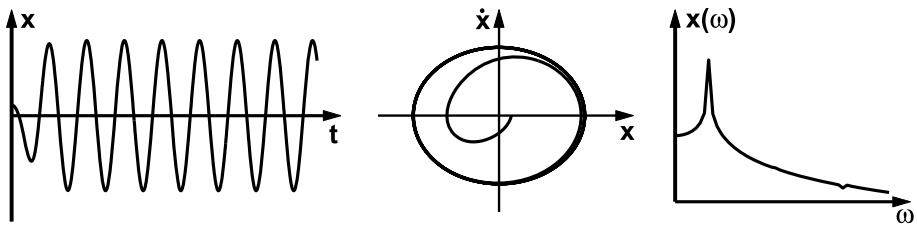

Movement Coordination, Figure 5
The hybrid oscillator: time series (*left*), phase space trajectory (*middle*) and power spectrum (*right*)

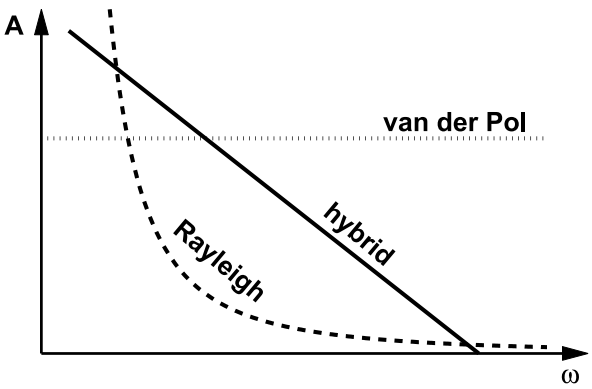

Movement Coordination, Figure 6
Amplitude-frequency relation for the van-der-Pol (*dotted*), Rayleigh ($\sim \omega^{-2}$, *dashed*) and hybrid ($\sim -\omega$, *solid*) oscillator

in-phase in bimanual finger movements takes the form

$$\begin{aligned} \ddot{x}_1 + \dot{x}_1\left(\gamma + \epsilon x_1^2 + \delta \dot{x}_1^2\right) + \omega^2 x_1 \\ = (\dot{x}_1 - \dot{x}_2)\left\{\alpha + \beta(x_1 - x_2)^2\right\} \\ \ddot{x}_2 + \dot{x}_2\left(\gamma + \epsilon x_2^2 + \delta \dot{x}_2^2\right) + \omega^2 x_2 \\ = (\dot{x}_2 - \dot{x}_1)\left\{\alpha + \beta(x_2 - x_1)^2\right\} . \end{aligned} \tag{16}$$

A numerical simulation of (16) is shown in Fig. 7. In the top row the amplitudes x_1 and x_2 are plotted as a function of time. The movement starts out in anti-phase at $\omega = 1.4$ and the frequency is continuously increased to a final value of $\omega = 1.8$. At a critical rate ω_c the anti-phase pattern becomes unstable and a transition to in-phase takes place. At the bottom a continuous estimate of the relative phase $\phi(t)$ is shown calculated as

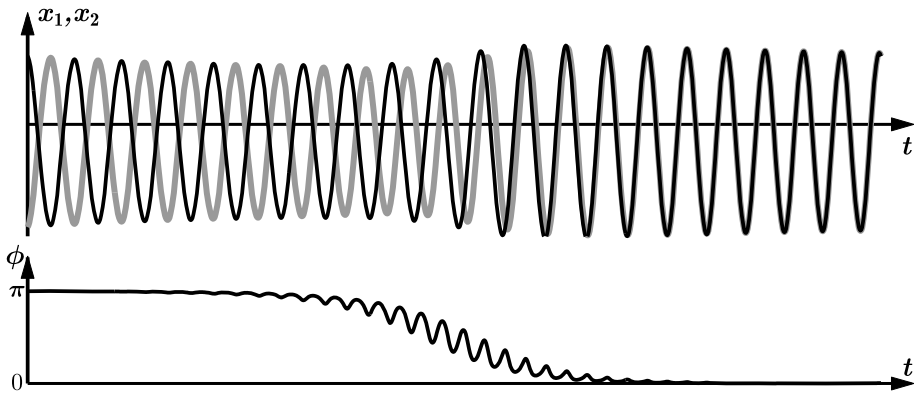

Movement Coordination, Figure 7
Simulation of (16) where the frequency ω is continuously increased from $\omega = 1.4$ on the left to $\omega = 1.8$ on the right. *Top*: time series of the amplitudes x_1 and x_2 undergoing a transition from anti-phase to in-phase when ω exceeds a critical value. *Bottom*: Continuous estimate of the relative phase ϕ changing from an initial value of π during anti-phase to 0 when the in-phase movement is established. Parameters: $\gamma = -0.7$, $\epsilon = \delta = 1$, $\alpha = -0.2$, $\beta = 0.2$, and $\omega = 1.4$ to 1.8

$$\phi(t) = \varphi_1(t) - \varphi_2(t) = \arctan\frac{\dot{x}_1}{x_1} - \arctan\frac{\dot{x}_2}{x_2}\,. \quad (17)$$

The relative phase changes from a value of π during the anti-phase movement to $\phi = 0$ when the in-phase pattern has been established.

To derive the phase relation (2) from (16) is a little lengthy but straightforward by using the ansatz (hypothesis)

$$x_k(t) = A_k(t)\mathrm{e}^{i\omega t} + A_k^*(t)\mathrm{e}^{-i\omega t} \quad (18)$$

then calculating the derivatives and inserting them into (16). Next a slowly varying amplitude approximation ($\dot{A}(t) \ll \omega$) and rotating wave approximation (neglect all frequencies $> \omega$) are applied. Finally, introducing the relative phase $\phi = \varphi_1 - \varphi_2$ after writing $A_k(t)$ in the form

$$A_k(t) = r\mathrm{e}^{i\varphi_k(t)} \quad (19)$$

leads to a relation for the relative phase ϕ of the form (2) from which the parameters a and b can be readily found in terms of the parameters that describe the oscillators and their coupling in (16)

$$a = -\alpha - 2\beta r^2\,, \quad b = \frac{1}{2}\beta r^2$$
$$\text{with} \quad r^2 = \frac{-\gamma + \alpha(1-\cos\phi)}{\epsilon + 3\delta\omega^2 - 2\beta(1-\cos\phi)^2}\,. \quad (20)$$

Breaking and Restoring Symmetries

Symmetry Breaking Through the Components

For simplicity, the original HKB model assumes on both the oscillator and the relative phase level that the two coordinating components are identical, like two index fingers. As a consequence, the coupled system (16) has a symmetry: it stays invariant if we replace x_1 by x_2 and x_2 by x_1. For the coordination between two limbs that are not the same like an arm and a leg, this symmetry no longer exists – it is said to be broken. In terms of the model, the main difference between an arm and a leg is that they have different eigenfrequencies, so the oscillator frequencies ω in (16) are no longer the same but become ω_1 and ω_2. This does not necessarily mean that during the coordination the components oscillate at different frequencies; they are still coupled, and this coupling leads to a common frequency Ω, at least as long as the eigenfrequency difference is not too big. But still, a whole variety of new phenomena originates from such a breaking of the symmetry between the components [5,22,23,29,37].

As mentioned in Subsect. "The Coupling" the dynamics for the relative phase can be derived from the level of coupled oscillators (16) for the case of the same eigenfrequencies. Performing the same calculations for two oscillators with frequencies ω_1 and ω_2 leads to an additional term in (2), which turns out to be a constant, commonly called $\delta\omega$. With this extension the equation for the relative phase reads

$$\dot{\phi} = \delta\omega - a\sin\phi - 2b\sin 2\phi$$
$$\text{with} \quad \delta\omega = \frac{\omega_1^2 - \omega_2^2}{\Omega} \approx \omega_1 - \omega_2\,. \quad (21)$$

The exact form for the term $\delta\omega$ turns out to be the difference of the squares of the eigenfrequencies divided by the rate Ω the oscillating frequency of the coupled system, which simplifies to $\omega_1 - \omega_2$ if the frequency difference is small. As before (21) can be scaled, which eliminates one

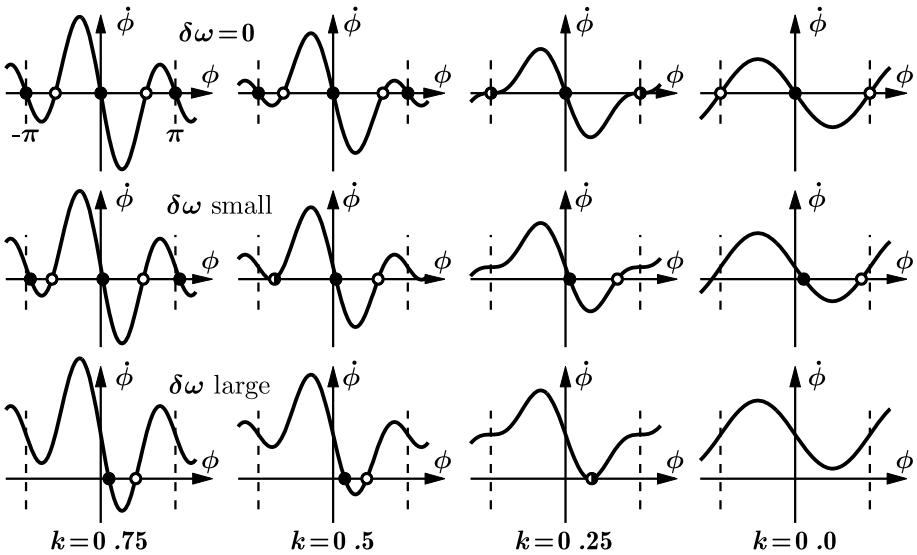

Movement Coordination, Figure 8

Phase space plots for different values of the control parameters k and $\delta\omega$. With increasing asymmetry (*top* to *bottom*) the functions are shifted more and more upwards leading to an elimination of the fixed points near $\phi = -\pi$ and $\phi = 0$ via saddle node bifurcations at $k = 0.5$ for small $\delta\omega$ and $k = 0.25$ for $\delta\omega$ large, respectively

of the parameters, and $\dot{\phi}$ can be derived from a potential function

$$\begin{aligned}\dot{\phi} &= \delta\omega - \sin\phi - 2k\sin 2\phi \\ &= -\frac{\mathrm{d}V(\phi)}{\mathrm{d}\phi} \text{ with } V(\phi) = -\delta\omega\,\phi - \cos\phi - k\cos 2\phi\,.\end{aligned} \tag{22}$$

Plots of the phase space and the potential landscape for different values of k and $\delta\omega$ are shown in Figs. 8 and 9, respectively. From these figures it is obvious that the symmetry breaking leads to a vertical shift of the curves in phase space and a tilt in the potential functions, which has several important consequences for the dynamics. First, for a nonvanishing $\delta\omega$ the stable fixed points for the relative phase are no longer located at $\phi = 0$ and $\phi = \pm\pi$ but are now shifted (see Fig. 8). The amount of this shift can be calculated for small values of $\delta\omega$ and new locations for the stable fixed points are given by

$$\phi^{(0)} = \frac{\delta\omega}{1+4k} \quad \text{and} \quad \phi^{(\pi)} = \pi - \frac{\delta\omega}{1-4k}\,. \tag{23}$$

Second, for large enough values of $\delta\omega$ not only the fixed point close to $\phi = \pi$ becomes unstable but also the in-phase pattern loses stability undergoing a saddle node bifurcation as can be seen in the bottom row in Fig. 8. Beyond this point there are no stable fixed points left and the relative phase will not settle down at a fixed value anymore but exhibits phase wrapping. However, this wrapping does not occur with a constant angular velocity, which can best be seen in the plot on the bottom right in Fig. 9. As the change in relative phase $\dot{\phi}$ is the negative derivative of the potential function, it is given by the slope. This slope is large and almost constant for negative values of ϕ, but for small positive values, where the in-phase fixed point was formerly located, the slope becomes less steep indicating that ϕ changes more slowly in this region before the dynamics picks up speed again when approaching π. So even as the fixed point has disappeared the dynamics still shows reminiscence of its former existence.

The dynamics of relative phase for the case of different eigenfrequencies from a simulation of (22) in shown in Fig. 10. Starting out at a slow movement rate on the left, the system settles into the fixed point close to $\phi = \pi$. When the movement rate is continuously increased, the fixed point drifts upwards. At a first critical point a transition to in-phase takes place, followed by another drift, this time for the fixed point representing the in-phase movement. Finally, this state also looses stability and the relative phase goes into wrapping. Reminiscence in the phase regions of the former fixed point are still visible by a flattening of the slope around $\phi \approx> 0$. With a further increase of the movement rate the function approaches a straight line.

The third consequence of this symmetry breaking is best described using the potential function for small values of $\delta\omega$ compared to the symmetric case $\delta\omega = 0$. For the lat-

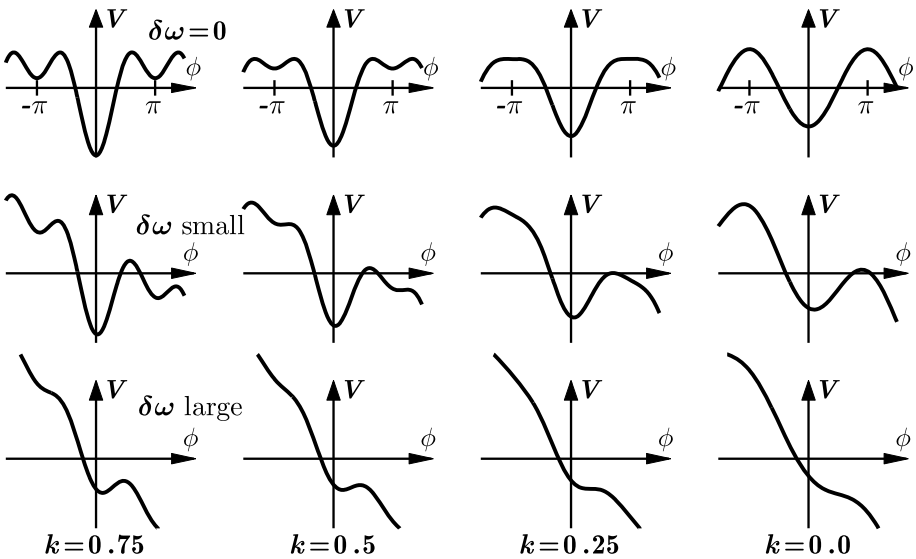

Movement Coordination, Figure 9

Potential landscape for different values of the control parameters k and $\delta\omega$. With increasing asymmetry (*top* to *bottom*) the functions get more and more tilted, destabilizing the system up to a point where there are no fixed points left on the bottom right. However, remnants of the fixed point can still be seen as changes in the curvature of the potential

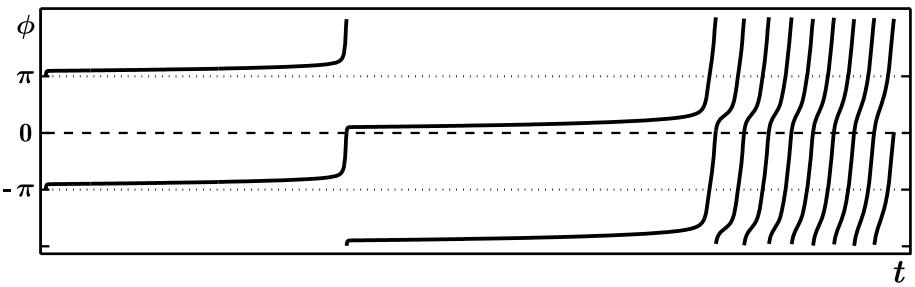

Movement Coordination, Figure 10

Relative phase ϕ as a function of time. Shown is a 4-π plot of a simulation of (22) for $\delta\omega = 1.7$ where the control parameter k is continuously decreased from $k = 2$ on the left to $k = 0$ on the right. The system settles close to anti-phase and the fixed point drifts as k is decreased (corresponding to a faster period of oscillation). At a first critical value a transition to in-phase takes place followed by another fixed point drift. Finally, the in-phase fixed point disappears and the phase starts wrapping

ter, when the system is initially in anti-phase $\phi = \pi$ and k is decreased through its critical value a switch to in-phase takes place as was shown in Fig. 1 (middle row). However, the ball there does not necessarily have to roll to the left towards $\phi = 0$ but with the same probability could roll to the right ending up in the minimum that exists at $\phi = 2\pi$ and also represents an in-phase movement. Whereas the eventual outcome is the same because due to the periodicity $\phi = 0$ and $\phi = 2\pi$ are identical, the two paths can very well be distinguished. The curve in Fig. 7 (bottom), showing the continuous estimate of the relative phase during a transition, goes from $\phi = \pi$ down to $\phi = 0$, but could, in fact with the same probability, go up towards $\phi = 2\pi$. In contrast, if the eigenfrequencies are different, also the points $-\pi$ and π, and 0 and 2π are no longer the same. If the system is in anti-phase at $\phi = \pi$ and k is decreased, it is evident from the middle row in Fig. 9 that a switch will not take place towards the left to $\phi \approx 0$, as the dynamics would have to climb over a potential hill to do so. As there are random forces acting on the dynamics a switch to $\phi \approx 0$ will still happen from time to time, but it is not equally probable to a transition to $\phi \approx 2\pi$, and it becomes even more unlikely with increasing $\delta\omega$.

These consequences, theoretically predicted to occur when the symmetry between the oscillating components is broken, can and have been tested, and have been found to

be in agreement with the experimental results [21,29] (see also [32,41]).

Asymmetry in the Mode of Coordination

Even though (16) is symmetric in the coordinating components it can only describe a transition from anti-phase to in-phase but not the other way around. Equation (16) is highly asymmetric with respect to coordination mode. This can be seen explicitly when we introduce variables that directly reflect modes of coordination

$$\psi_+ = x_1 + x_2 \quad \text{and} \quad \psi_- = x_1 - x_2 \,. \tag{24}$$

For an in-phase movement we have $x_1 = x_2$ and ψ_- vanishes, whereas for anti-phase $x_1 = -x_2$ and therefore $\psi_+ = 0$. We can now derive the dynamics in the variables ψ_+ and ψ_- by expressing the original displacements as

$$x_1 = \frac{1}{2}(\psi_+ + \psi_-) \quad \text{and} \quad x_2 = \frac{1}{2}(\psi_+ - \psi_-) \tag{25}$$

and inserting them into (16), which leads to

$$\begin{aligned} &\ddot{\psi}_+ + \epsilon\dot{\psi}_+ + \omega^2\psi_+ + \frac{\gamma}{12}\frac{\mathrm{d}}{\mathrm{d}t}\left(\psi_+^3 + 3\psi_+\psi_-^2\right) \\ &\quad + \frac{\delta}{4}\left(\dot{\psi}_+^3 + 3\dot{\psi}_+\dot{\psi}_-^2\right) = 0 \\ &\ddot{\psi}_- + \epsilon\dot{\psi}_- + \omega^2\psi_- + \frac{\gamma}{12}\frac{\mathrm{d}}{\mathrm{d}t}\left(\psi_-^3 + 3\psi_-\psi_+^2\right) \\ &\quad + \frac{\delta}{4}\left(\dot{\psi}_-^3 + 3\dot{\psi}_-\dot{\psi}_+^2\right) = 2\dot{\psi}_-\left(\alpha + \beta\psi_-^2\right) . \end{aligned} \tag{26}$$

The asymmetry between in-phase and anti-phase is evident from (26), as the right-hand side of the first equation vanishes and the equation is even independent of the coupling parameters α and β. This is the reason that the original HKB model only shows transitions from anti-phase to in-phase and not vice versa.

Transitions to Anti-phase

In 2000 Carson and colleagues [6] published results from an experiment in which subjects performed bimanual pronation-supination movements paced by a metronome of increasing rate (see also [2]). In this context an anti-phase movement corresponds to the case where one arm performs a pronation while the other arm is supinating. Correspondingly, pronation and supination with both arms at the same time represents in-phase. In their experiment Carson et al. used a manipulandum that allowed for changing the axis of rotation individually for both arms as shown in Fig. 11a. With increasing movement rate spontaneous transitions from anti-phase to in-phase, but not vice versa, were found when the subjects performed pronation-supination movements around the same axes for both arms. In trials where one arm was rotating around the axis above the hand and the other around the one below, anti-phase was found to be stable and the in-phase movement underwent a transition to anti-phase as shown for representative trials in Fig. 12.

It is evident that the HKB model in neither its original form (2) nor the mode formulation (26) is a valid model for these findings. However, Fuchs and Jirsa [11] showed that by starting from the mode description (26) it is straightforward to extend HKB such that, depending on an additional parameter σ, either the in-phase or the anti-

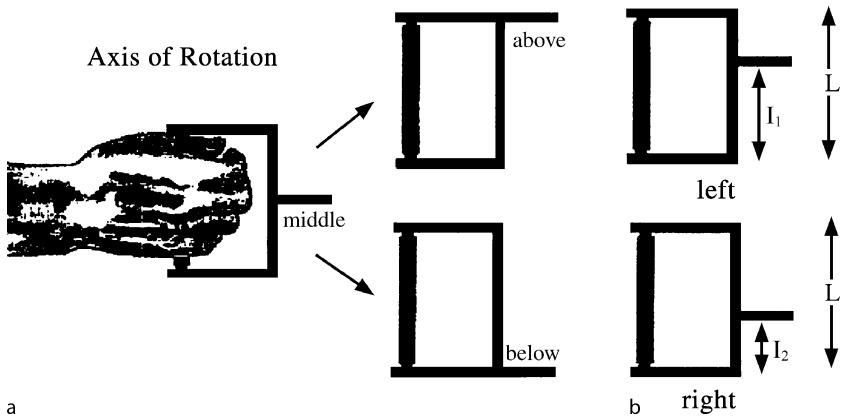

Movement Coordination, Figure 11

Manipulandum used by Carson and colleagues [6]. a The original apparatus that allowed for variation in axis of rotation above, below and in the middle of the hand. b The axis of rotation can be changed continuously, allowing us to introduce a parameter σ as a quantitative measure for the relative locations of the axes

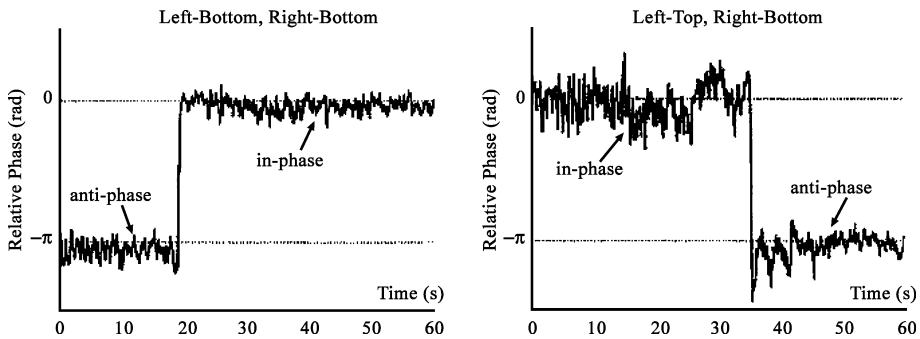

Movement Coordination, Figure 12
Relative phase over time for two representative trials from the Carson et al. experiment. *Left*: the axis of rotation is below the hand for both arms and a switch from anti-phase to in-phase occurs as the movement speeds up. *Right*: with one axis above and the other below the hand, the in-phase movement becomes unstable at higher rates leading to a transition to anti-phase

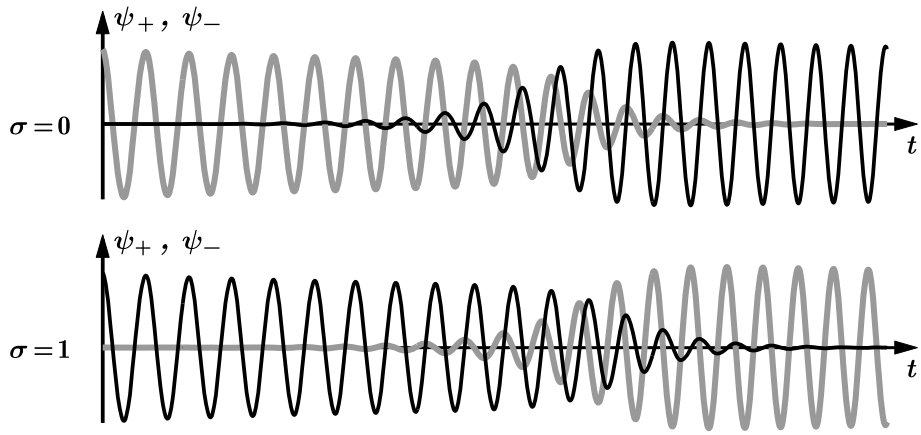

Movement Coordination, Figure 13
Simulation of (28) for $\sigma = 0$ (*top*) and $\sigma = 1$ (*bottom*) where the frequency ω is continuously increased from $\omega = 1.4$ on the left to $\omega = 1.8$ on the right. Time series of the mode amplitudes ψ_+ (black) and ψ_- (gray) undergoing transitions from anti-phase to in-phase (*top*) and from in-phase to anti-phase (*bottom*) when ω exceeds a critical value. Parameters: $\gamma = -0.7, \epsilon = \delta = 1, \alpha = -0.2, \beta = 0.2$, and $\omega = 1.4$ to 1.8

phase mode is a stable movement pattern at high rates. The additional parameter corresponds to the relative locations of the axes of rotation in the Carson et al. experiment which can be defined in its easiest form as

$$\sigma = \frac{|l_1 - l_2|}{L} \tag{27}$$

where l_1, l_2 and L are as shown in Fig. 11b. In fact, any monotonic function f with $f(0) = 0$ and $f(1) = 1$ is compatible with theory and its actual shape has to be determined experimentally.

By looking at the mode Eqs. (26) it is clear that a substitution $\psi_+ \to \psi_-$ and $\psi_- \to \psi_+$ to the left-hand side of the first equation leads to the left-hand side of the second equation and vice versa. For the terms on the right-hand side representing the coupling this is obviously not the case. Therefore, we now introduce a parameter σ and additional terms into (26) such that for $\sigma = 0$ these equations remain unchanged, whereas for $\sigma = 1$ we obtain (26) with all + and − subscripts reversed

$$\begin{aligned}
&\ddot{\psi}_+ + \epsilon\dot{\psi}_+ + \omega^2\psi_+ + \frac{\gamma}{12}\frac{\mathrm{d}}{\mathrm{d}t}\left(\psi_+^3 + 3\psi_+\psi_-^2\right) \\
&\qquad + \frac{\delta}{4}\left(\dot{\psi}_+^3 + 3\dot{\psi}_+\dot{\psi}_-^2\right) = 2\sigma\dot{\psi}_+\left(\alpha + \beta\psi_+^2\right) \\
&\ddot{\psi}_- + \epsilon\dot{\psi}_- + \omega^2\psi_- + \frac{\gamma}{12}\frac{\mathrm{d}}{\mathrm{d}t}\left(\psi_-^3 + 3\psi_-\psi_+^2\right) \\
&\qquad + \frac{\delta}{4}\left(\dot{\psi}_-^3 + 3\dot{\psi}_-\dot{\psi}_+^2\right) = 2(1-\sigma)\dot{\psi}_-\left(\alpha + \beta\psi_-^2\right).
\end{aligned} \tag{28}$$

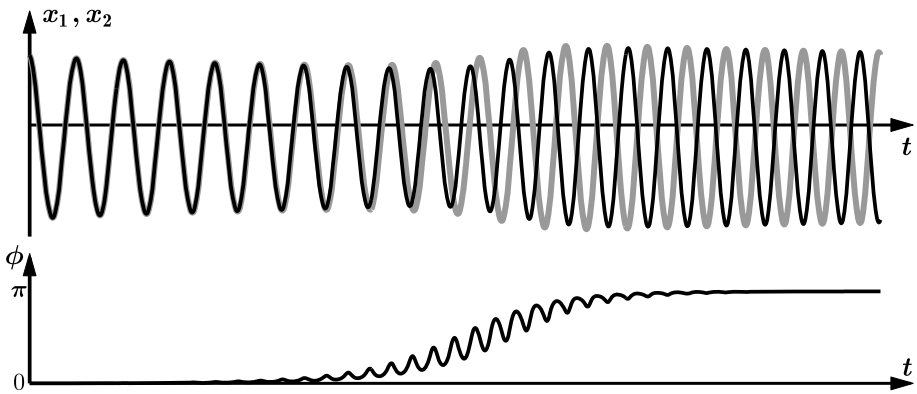

Movement Coordination, Figure 14
Simulation of (30) where the frequency ω is continuously increased from $\omega = 1.4$ on the left to $\omega = 1.8$ on the right. *Top*: time series of the amplitudes x_1 and x_2 undergoing a transition from in-phase to anti-phase when ω exceeds a critical value. *Bottom*: Continuous estimate of the relative phase ϕ changing from an initial value of 0 during the in-phase to π when the anti-phase movement is established. Parameters: $\gamma = -0.7, \epsilon = \delta = 1, \alpha = -0.2, \beta = 0.2, \sigma = 1$ and $\omega = 1.4$ to 1.8

From (28) it is straight forward to go back to the representation of the limb oscillators

$$\begin{aligned}\ddot{x}_1 + \ldots &= \frac{1}{2}\left(\ddot{\psi}_+ + \ddot{\psi}_-\right) + \ldots \\ &= \dot{\psi}_-\left(\alpha + \beta\psi_-^2\right) + \sigma\{\dot{\psi}_+\left(\alpha + \beta\psi_+^2\right) \\ &\qquad - \dot{\psi}_-\left(\alpha + \beta\psi_-^2\right)\} \\ \ddot{x}_2 + \ldots &= \frac{1}{2}\left(\ddot{\psi}_+ - \ddot{\psi}_-\right) + \ldots \\ &= -\dot{\psi}_-\left(\alpha + \beta\psi_-^2\right) + \sigma\{\dot{\psi}_+\left(\alpha + \beta\psi_+^2\right) \\ &\qquad + \dot{\psi}_-\left(\alpha + \beta\psi_-^2\right)\}\end{aligned} \tag{29}$$

where the left-hand side which represents the oscillators has been written only symbolically as all we are dealing with is the coupling on the right. Replacing the mode amplitudes ψ_+ and ψ_- in (29) using (24) one finds the generalized coupling as a function of x_1 and x_2

$$\begin{aligned}\ddot{x}_1 + \ldots &= (\dot{x}_1 - \dot{x}_2)\left\{\alpha + \beta(x_1 - x_2)^2\right\} \\ &\quad + 2\sigma\left\{\alpha\dot{x}_2 + \beta\left[\dot{x}_2\left(x_1^2 + x_2^2\right) + 2\dot{x}_1 x_1 x_2\right]\right\} \\ \ddot{x}_2 + \ldots &= (\dot{x}_2 - \dot{x}_1)\left\{\alpha + \beta(x_2 - x_1)^2\right\} \\ &\quad + 2\sigma\left\{\alpha\dot{x}_1 + \beta\left[\dot{x}_1\left(x_1^2 + x_2^2\right) + 2\dot{x}_2 x_1 x_2\right]\right\} .\end{aligned} \tag{30}$$

Like the original oscillator Eq. (16), Eq. (30) is invariant under the exchange of x_1 and x_2 but in addition allows for transitions from in-phase to anti-phase coordination if the parameter σ is chosen appropriately ($\sigma = 1$, for instance), as shown in Fig. 14.

As the final step, an equation for the dynamics of relative phase can be obtained from (30) by performing the same steps as before, which leads to a modified form of the HKB equation (2)

$$\dot{\phi} = -(1 - 2\sigma)a\sin\phi - 2b\sin 2\phi \tag{31}$$

and the corresponding potential function

$$\dot{\phi} = -\frac{\mathrm{d}V(\phi)}{\mathrm{d}\phi} \quad \text{with} \quad V(\phi) = -(1 - 2\sigma)a\cos\phi - b\cos 2\phi\,. \tag{32}$$

Both equations can be scaled again leading to

$$\begin{aligned}\dot{\phi} &= -(1 - 2\sigma)\sin\phi - 2k\sin 2\phi \\ &= -\frac{\mathrm{d}V(\phi)}{\mathrm{d}\phi} \quad \text{with} \\ V(\phi) &= -(1 - 2\sigma)\cos\phi - k\cos 2\phi\,.\end{aligned} \tag{33}$$

The landscapes of the potential for different values of the control parameters k and σ are shown in Fig. 15. The left column exhibits the original HKB case which is obtained for $\sigma = 0$. The functions in the most right column, representing the situation for $\sigma = 1$, are identical in shape to the $\sigma = 0$ case, simply shifted horizontally by a value of π. These two extreme cases are almost trivial and were the ones originally investigated in the Carson et al. experiment with the axes of rotation either on the same side or on opposite sides with respect to the hand. As the corresponding potential functions are shifted by π with respect to each other, one could assume that for an intermediate value of σ between 0 and 1 the functions are also shifted, just by a smaller amount. Such horizontal translations lead to fixed point drifts, as has been seen before for oscillation components with different eigenfrequen-

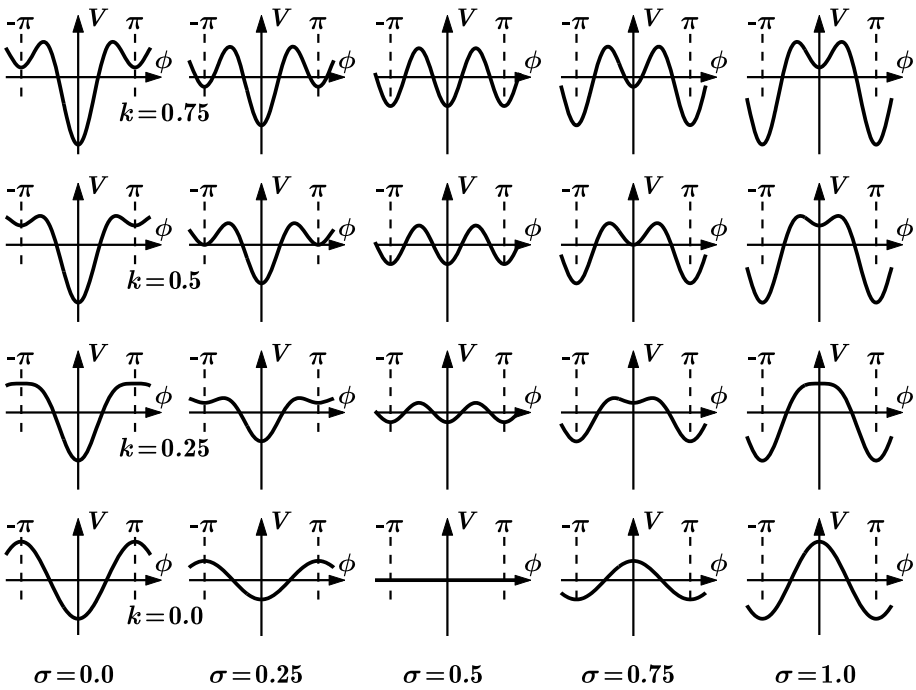

Movement Coordination, Figure 15
Potential landscape for different values of the control parameters *k* and *σ*

cies. The theory, however, predicts that this is not the case. In fact, for $\sigma = 0.5$ theory predicts that the two coordination modes in-phase and anti-phase are equally stable for all movement rates. The deep minima for slow rates indicate high stability for both movement patterns and as the rate increases both minima become more and more shallow, i. e. both movement patterns become less stable. Eventually, for high rates at $k = 0$ the potential is entirely flat, which means that there are no attractive states whatsoever. Pushed only by the stochastic forces in the system, the relative phase will now undergo a random walk. Note that this is very different from the phase wrapping encountered before where the phase was constantly increasing due to the lack of an attractive state. Here the relative phase will move back and forth in a purely random fashion, known in the theory of stochastic systems as Brownian motion. Again experimental evidence exists from the Carson group that changing the distance between the axes of rotation gradually leads to the phenomena predicted by theory.

Conclusions

The theoretical framework outlined above represents the core of the dynamical systems approach to movement coordination. Rather than going through the large variety of phenomena that coordination dynamics and the HKB model have been applied to, emphasis has been put on a detailed description of the close connection between theoretical models and experimental results. Modeling the coordination of movement as dynamical systems on both the mesoscopic level of the component oscillators and the macroscopic level of relative phase allowed for quantitative predictions and experimental tests with an accuracy that is virtually unprecedented in the life sciences, a field where most models are qualitative and descriptive.

Extensions of the HKB Model

Beyond the phenomena described above, the HKB model has been extended in various ways. Some of these extensions (by no mean exhaustive) are listed below with very brief descriptions; the interested reader is referred to the literature for details.

- The quantitative description of the influence of noise on the dynamics given in Sect. "Stability: Perturbations and Fluctuations" can be done in a quantitative fashion by adding a stochastic term to (2) [40,43] or its generalizations (21) and (31) [11] and treating them as Langevin equations within the theory of stochastic systems (see e. g. [16] for stochastic systems). In this case the system is no longer described by a single time series for the relative phase but by a probability distribu-

tion function. How such distributions evolve in time is then given by the corresponding Fokker–Planck equation and allows for a quantitative description of the stochastic phenomena such as enhancement of fluctuations and critical fluctuations. An important quantity that can be derived in this context and is also related to the critical fluctuations is the mean-first-passage time, which is the time it takes (on average) to move over a hump in the potential function.

- When subjects flex a single finger between the beats of a metronome, i. e. syncopate with the stimulus, and the metronome rate is increased, they switch spontaneously to a coordination pattern where they flex their finger on the beat, i. e. synchronize with the stimulus. This so-called syncopation-synchronization paradigm introduced by Kelso and colleagues [32] has been frequently used in brain-imaging experiments.
- A periodic patterning in the time series of the relative phase was found experimentally in the case of broken symmetry by Schmidt et al. [41] and successfully derived from the oscillator level of the HKB model [12,14].
- The metronome pacing can be explicitly included into (2) and its generalizations [24]. This so-called parametric driving allows us to explain effects in the movement trajectory known as anchoring, i. e. the variability of the movement is smaller around the metronome beat compared to other regions in phase space [10]. With parametric driving the HKB model also makes correct predictions for the stability of multi-frequency coordination, where the metronome cycle is half of the movement cycle, i. e. there is a beat at the points of maximum flexion and maximum extension [1]. There are also effects from more complicated polyrhythms that have been studied [38,39,47,48,49].
- The effect of symmetry breaking has been studied intensively in experiments where subjects were swinging pendulums with different eigenfrequencies [8,37,46].
- Transitions are also found in trajectory formation, for instance when subjects move their index finger such that they draw an "8" and this movement is sped up the pattern switches to a "0" [3,4,9].

Future Directions

One of the most exciting applications of movement coordination and its spontaneous transitions in particular is that they open a new window for probing the human brain, made possible by the rapid development of brain-imaging technologies that allow for the recording of brain activity in a noninvasive way. Electroencephalography (EEG), magnetoencephalography (MEG) and functional magnetic resonance tomography (fMRI) have been used in coordination experiments since the 1990s to study the changes in brain activations accompanying (or triggering?) the switches in movement behavior [13,33,34]. Results from MEG experiments reveal a strong frequency dependence of the dominating pattern with the contribution of the auditory system being strongest at low metronome/movement rates, whereas at high rates the signals from sensorimotor cortex dominate [15,35]. The crossover point is found at rates around 2 Hz, right where the transitions typically take place.

In two other studies the rate dependence of the auditory and sensorimotor system was investigated separately. In an MEG experiment Carver et al. [7] found a resonance-like enhancement of a brain response that occurs about 50 ms after a tone is delivered, again at a rate of about 2 Hz. In the sensorimotor system a nonlinear effect of rate was shown as well. Using a continuation paradigm, where subjects moved an index finger paced by a metronome which was turned off at a certain time while the subjects were to continue moving at the same rate, Mayville et al. [36] showed that a certain pattern of brain activation drops out when the movement rate exceeds about 1.5 Hz. Even though their contribution to behavioral transitions is far from being completely understood, it is clear that such nonlinear effects of rate exist in both the auditory and the sensorimotor system in parameter regions where behavioral transitions are observed.

Using fMRI brain areas have been identified that show a dependence of their activation level as a function of rate only, independent of coordination mode, whereas activation in other areas strongly depends on whether subjects are syncopating or synchronizing regardless of how fast they are moving [20].

Taken together, these applications of coordination dynamics to brain research have hardly scratched the surface so far but the results are already very exciting as they demonstrate that the experimental paradigms from movement coordination may be used to prepare the brain into a certain state where its responses can be studied. With further improvement of the imaging technologies and analysis procedures many more results can be expected to contribute significantly to our understanding of how the human brain works.

Acknowledgment

Work reported herein was supported by NINDS grant 48299, NIMH grants 42900 and 80838, and the Pierre de Fermat Chair to J.A.S.K.

Bibliography

Primary Literature

1. Assisi CG, Jirsa VK, Kelso JAS (2005) Dynamics of multifrequency coordination using parametric driving: Theory and Experiment. Biol Cybern 93:6–21
2. Buchanan JJ, Kelso JAS (1993) Posturally induced transitions in rhythmic multijoint limb movements. Exp Brain Res 94:131–142
3. Buchanan JJ, Kelso JAS, Fuchs A (1996) Coordination dynamics of trajectory formation. Biol Cybern 74:41–54
4. Buchanan JJ, Kelso JAS, DeGuzman GC (1997) The self-organization of trajectory formation: I. Experimental evidence. Biol Cybern 76:257–273
5. Carson RG, Goodman D, Kelso JAS, Elliott D (1995) Phase transitions and critical fluctuations in rhythmic coordination of ipsilateral hand and foot. J Mot Behav 27:211–224
6. Carson RG, Rick S, Smethrust CJ, Lison JF, Biblow WD (2000) Neuromuscular-skeletal constraints upon the dynamics of unimanual and bimanual coordination. Exp Brain Res 131:196–214
7. Carver FW, Fuchs A, Jantzen KJ, Kelso JAS (2002) Spatiotemporal analysis of neuromagnetic activity associated with rhythmic auditory stimulation. Clin Neurophysiol 113:1909–1920
8. Collins DR, Sternad D, Turvey MT (1996) An experimental note on defining frequency competition in intersegmental coordination dynamics. J Mot Behav 28:299–303
9. DeGuzman GC, Kelso JAS, Buchanan JJ (1997) The self-organization of trajectory formation: II. Theoretical model. Biol Cybern 76:275–284
10. Fink P, Kelso JAS, Jirsa VK, Foo P (2000) Local and global stabilization of coordination by sensory information. Exp Brain Res 134:9–20
11. Fuchs A, Jirsa VK (2000) The HKB model revisited: How varying the degree of symmetry controls dynamics. Hum Mov Sci 19:425–449
12. Fuchs A, Kelso JAS (1994) A theoretical note on models of Interlimb coordination. J Exp Spychol Hum Percept Perform 20:1088–1097
13. Fuchs A, Kelso JAS, Haken H (1992) Phase transitions in the human brain: Spatial mode dynamics. Int J Bifurc Chaos 2:917–939
14. Fuchs A, Jirsa VK, Haken H, Kelso JAS (1996) Extending the HKB-Model of coordinated movement to oscillators with different eigenfrequencies. Biol Cybern 74:21–30
15. Fuchs A, Mayville JM, Cheyne D, Weinberg H, Deecke L, Kelso JAS (2000) Spatiotemporal Analysis of Neuromagnetic Events Underlying the Emergence of Coordinative Instabilities. NeuroImage 12:71–84
16. Gardiner CW (1985) Handbook of stochastic Systems. Springer, Heidelberg
17. Haken H (1977) Synergetics, an introduction. Springer, Heidelberg
18. Haken H (1983) Advanced Synergetics. Springer, Heidelberg
19. Haken H, Kelso JAS, Bunz H (1985) A theoretical model of phase transition in human hand movements. Biol Cybern 51:347–356
20. Jantzen KJ, Kelso JAS (2007) Neural coordination dynamics of human sensorimotor behavior: A review. In: Jirsa VK, McIntosh AR (eds) Handbook of Brain Connectivity. Springer, Heidelberg
21. Jeka JJ, Kelso JAS (1995) Manipulating symmetry in human two-limb coordination dynamics. J Exp Psychol Hum Percept Perform 21:360–374
22. Jeka JJ, Kelso JAS, Kiemel T (1993) Pattern switching in human multilimb coordination dynamics. Bull Math Biol 55:829–845
23. Jeka JJ, Kelso JAS, Kiemel T (1993) Spontaneous transitions and symmetry: Pattern dynamics in human four limb coordination. Hum Mov Sci 12:627–651
24. Jirsa VK, Fink P, Foo P, Kelso JAS (2000) Parameteric stabilization of biological coordination: a theoretical model. J Biol Phys 26:85–112
25. Kay BA, Kelso JAS, Saltzman EL, Schöner G (1987) Space-time behavior of single and bimanual rhythmic movements: Data and limit cycle model. J Exp Psychol Hum Percept Perform 13:178–192
26. Kay BA, Saltzman EL, Kelso JAS (1991) Steady state and perturbed rhythmical movements: Dynamical modeling using a variety of analytic tools. J Exp Psychol Hum Percept Perform 17:183–197
27. Kelso JAS (1981) On the oscillatory basis of movement. Bull Psychon Soc 18:63
28. Kelso JAS (1984) Phase transitions and critical behavior in human bimanual coordination. Am J Physiol Regul Integr Comp 15:R1000–R1004
29. Kelso JAS, Jeka JJ (1992) Symmetry breaking dynamics of human multilimb coordination. J Exp Psychol Hum Percept Perform 18:645–668
30. Kelso JAS, Scholz JP, Schöner G (1986) Nonequilibrium phase transitions in coordinated biological motion: Critical fluctuations. Phys Lett A 118:279–284
31. Kelso JAS, Schöner G, Scholz JP, Haken H (1987) Phase locked modes, phase transitions and component oscillators in coordinated biological motion. Phys Scr 35:79–87
32. Kelso JAS, DelColle J, Schöner G (1990) Action-perception as a pattern forming process. In: Jannerod M (ed) Attention and performance XIII. Erlbaum, Hillsdale, pp 139–169
33. Kelso JAS, Bressler SL, Buchanan S, DeGuzman GC, Ding M, Fuchs A, Holroyd T (1992) A phase transition in human brain and behavior. Phys Lett A 169:134–144
34. Kelso JAS, Fuchs A, Holroyd T, Lancaster R, Cheyne D, Weinberg H (1998) Dynamic cortical activity in the human brain reveals motor equivalence. Nature 392:814–818
35. Mayville JM, Fuchs A, Ding M, Cheyne D, Deecke L, Kelso JAS (2001) Event-related changes in neuromagnetic activity associated with syncopation and synchronization timing tasks. Hum Brain Mapp 14:65–80
36. Mayville JM, Fuchs A, Kelso JAS (2005) Neuromagnetic motor fields accompanying self-paced rhythmic finger movements of different rates. Exp Brain Res 166:190–199
37. Park H, Turvey MT (2008) Imperfect symmetry and the elementary coordination law. In: Fuchs A, Jirsa VK (eds) Coordination: Neural, Behavioral and Social Dynamics. Springer, Heidelberg, pp 3–25
38. Peper CE, Beek PJ (1998) Distinguishing between the effects of frequency and amplitude on interlimb coupling in tapping a 2:3 polyrhythm. Exp Brain Res 118:78–92
39. Peper CE, Beek PJ, van Wieringen PC (1995) Frequency-induced phase transitions in bimanual tapping. Biol Cybern 73:303–309
40. Post AA, Peeper CE, Daffertshofer A, Beek PJ (2000) Relative phase dynamics in perturbed interlimb coordination: stability and stochasticity. Biol Cybern 83:443–459

41. Schmidt RC, Beek PJ, Treffner PJ, Turvey MT (1991) Dynamical substructure of coordinated rhythmic movements. J Exp Psychol Hum Percept Perform 17:635–651
42. Schöner G, Kelso JAS (1988) Dynamic pattern generation in behavioral and neural systems. Science 239:1513–1520
43. Schöner G, Haken H, Kelso JAS (1986) A stochastic theory of phase transitions in human hand movements. Biol Cybern 53:442–453
44. Scholz JP, Kelso JAS (1989) A quantitative approach to understanding the formation and change of coordinated movement patterns. J Mot Behav 21:122–144
45. Scholz JP, Kelso JAS, Schöner G (1987) Nonequilibrium phase transitions in coordinated biological motion: Critical slowing down and switching time. Phys Lett A 8:90–394
46. Sternad D, Collins D, Turvey MT (1995) The detuning factor in the dynamics of interlimb rhythmic coordination. Biol Cybern 73:27–35
47. Sternad D, Turvey MT, Saltzman EL (1999) Dynamics of 1:2 Coordination: Generalizing Relative Phase to n:m Rhythms. J Mot Behav 31:207–233
48. Kelso JAS, DeGuzman GC, (1988) Order in time: How the cooperation between the hands informs the design of the brain. In: Haken H (ed) Neural and Synergetic Computers. Springer, Berlin
49. DeGuzman GC, Kelso JAS (1991) Multifrequency behavioral patterns and the phase attractive circle map. Biol Cybern 64:485–495

Books and Reviews

Fuchs A, Jirsa VK (eds) (2007) Coordination: Neural, Behavioral and Social Dynamics. Springer, Heidelberg
Jirsa VK, Kelso JAS (eds) (2004) Coordination Dynamics: Issues and Trends. Springer, Heidelberg
Kelso JAS (1995) Dynamics Pattern: The Self-Organization of Brain and Behavior. MIT Press, Cambridge
Haken H (1996) Principles of Brain Functioning. Springer, Heidelberg
Tschacher W, Dauwalder JP (eds) (2003) The Dynamical Systems Approach to Cognition: Concepts and Empirical Paradigms Based on Self-Organization, Embodiment and Coordination Dynamics. World Scientific, Singapore

Multi-Granular Computing and Quotient Structure

Ling Zhang[1], Bo Zhang[2]
[1] Artificial Intelligence Institute, Anhui University, Hefei, Anhui, China
[2] Department of Computer Science, State Key Lab of Intelligent Technology and Systems, Tsinghua University, Beijing, China

Article Outline

Glossary
Definition of the Subject
Introduction
The Granularity Relation
Hierarchy
Combination
Future Directions
Acknowledgments
Bibliography

Glossary

Granularity Granularity is the relative size, scale, level of detail, or depth of penetration that characterizes an object or system.

Multi-granular computing Humans are good at viewing and solving a problem at different grain-sizes (abstraction levels) and translating from one abstraction level to the others easily. This is one of basic characteristics of human intelligence. The aim of multi-granular computing is intended to investigate the granulation problem in human cognition and endow computers with the same capability to make them more efficient in problem solving.

Quotient set Given a universe X and an equivalence relation R on X, define a set $[X] = \{[x] | x \in X\}, [x] = \{y | y \sim x, y \in X\}$. $[X]$ is called a quotient set with respect to R, or simply a quotient set.

Quotient space Given a topologic space (X, T), T is a topology on X and R is an equivalence relation on X. Define a quotient structure on $[X]$ as $[T] = \{u | p^{-1}(u) \in T, u \subset [X]\}$, where $p\colon X \to [X]$ is a natural projection from X to $[X]$. Construct a topologic space $([X], [T])$. Space $([X], [T])$ is a quotient space corresponding to R. There are authors who keep the neighborhood system structured but remove the axioms of topology [13,14].

Quotient space model The quotient space model is a mathematical model to represent a problem at different grain-sizes by using the concept of quotient space in algebra. In the model a problem (or a system) is described by a triple (X, T, f), with universe (domain) X, structure T and attribute f. If X represents the universe composed of the objects with the finest grain-size, when we view the same universe X at a coarser grain-size, we have a coarse-grained universe denoted by $[X]$. Then we have a new problem space $([X], [T], [f])$, where $[X]$ is the quotient universe of X, $[T]$ the corresponding quotient structure and $[f]$ the quotient attribute. The coarse space $([X], [T], [f])$ is called a quotient space of space (X, T, f). Therefore, a problem with different grain-sizes can be represented by a family of quotient spaces.

Definition of the Subject

On one hand, any system in the world, either natural or artificial, has a multi-granular structure. This is called structural granulation. On the other hand, man always conceptualizes the world at different granularities and handles it hierarchically. This is called cognitive granulation. We believe the granulation in cognition underlies the human power in problem solving. Unfortunately, computers are generally capable of dealing with problems in only one abstraction level so far. The motivation of the research on multi-granular computing is intended to investigate the granulation both in human cognition and real world and endow the computers with the same capacity. This will greatly reduce the computational complexity in the computerized problem solving. This strategy can be used to improve many algorithms in broad areas such as planning, search, and machine learning.

The key to the multi-granular computing is to construct a mathematical model for representing a problem at different grain-sizes. We present an algebraic model (quotient space model) of granulation that is used to analyze both human cognition and real world, especially human problem solving behaviors. The model was developed in order to represent granules and compute with them easily.

Introduction

Granularity is the relative size, scale, level of detail, or depth of penetration that characterizes an object or system. This term has been used in astronomy, physics, geography, photography, and information science and technology frequently. When a system is divided into components, it's important to get the right degree of componentization. The fine-grained components give much more details in constructing precisely the right functionality of a system. The coarse-grained components are easier to manage but may lose some important details. Performance and management considerations tend to favor the use of multi-granularity. For example, a text contains several levels such as chapter, paragraph, sentence, and word; each has a different grain-size. A video has scene, shot, and frame. A country may be split into state, city, and community. Man always conceptualizes an object (or system) at different granularities and deals with it hierarchically. This is one of man's characteristics that underlies his/her power. The aim of multi-granular computing is intended to investigate the granulation problem in human cognition and endow computers with the same capability. So multi-granular computing has widely been investigated for a long time. In the data and knowledge management research community, the formalization of the concept of time granularity and its applications were addressed. A time granularity was defined by a countable set of granules; each granule can be composed of a single instant, a set of contiguous instants (time-interval), or even a set of non-contiguous instants [3]. The paper [4] investigates the algorithm, its computational complexity and several optimization techniques to the temporal CSP (Constraint Satisfaction Problem) involving constraints in terms of different time granularities. In image processing, a multiresolution model is used extensively and it is a model which captures a wide range of levels of detail of an object and which can be used to reconstruct any one of those levels on demand. The overall goal of multiresolution modeling is to use information from objects with different spatial granularities or extract the details from complex models that are necessary for rendering a scene and to get rid of the other, unnecessary details [10,11]. In GIS applications, the representation of multi-granular spatio-temporal objects in commercially available DBSMs was addressed [2]. The research works above mostly focused on the specific application domain and lacks of a general theoretical framework. The key to the investigation is to construct a multi-granular mathematical model for a problem. Recently, there have been several models to deal with the issue such as fuzzy set [24], rough set [20], and others [1,15,18]. Each has its own characteristics. We presented a new model called quotient-space model [25,26]. In the model a problem (or a system) is described by a triple (X, T, f). Universe (or domain) X represents the whole objects of the problem that we intended to deal with. It may be a point set and its power may either be finite or infinite. T is the structure of X and represents the relationship among objects of X. Structure T may have different forms but will be described by a topology on X in the following discussion. Attribute f is a set of functions defined on universe X and $f: X \to Y$ may be multi-component such as $f = (f_1, f_2, \ldots, f_n)$, $f_i: X \to Y_i$, where Y_i is either a set of real numbers or other kinds of sets. Triple (X, T, f) is called a problem space (or simply a space). We will focus on the granularity relationships.

Suppose that X represents the universe composed of the components with the finest grain-size; each component is regarded as indecomposable. When we view the same universe X at a coarser grain-size, we have a coarse-grained universe denoted by $[X]$. Then we have a new problem space $([X], [T], [f])$, where $[X]$ is the quotient universe of X, $[T]$ is the corresponding quotient structure and $[f]$ the quotient attribute. For example, text X consists of a set of words. If "word" is regarded as an indecomposable component, then X is the finest universe. T is the text structure, i. e., the syntactic relationship among words.

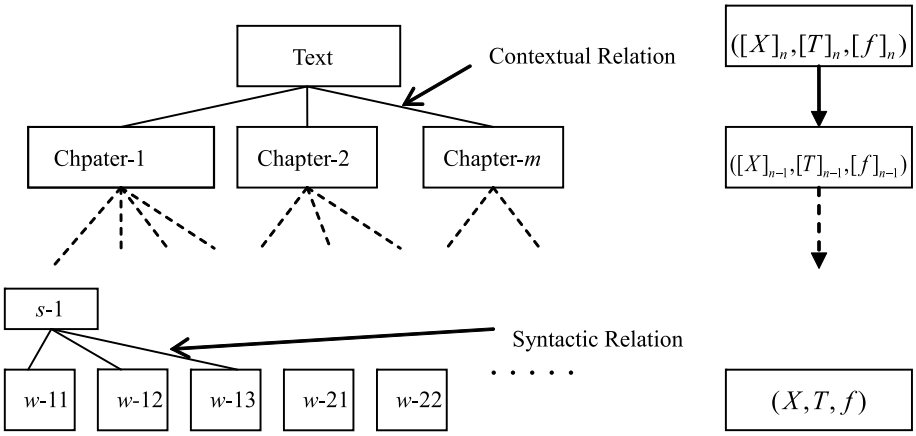

Multi-Granular Computing and Quotient Structure, Figure 1
A text represented at different grain-sizes

f is the property of words. If considering "sentence" as a new component, then we have a new (quotient) universe $[X]$ composed of a set of sentences; each sentence consists of a set of words. Then $[X]$ is a coarse grain-size universe. The quotient structure $[T]$ represents the contextual relationship among sentences. Quotient attribute $[f]$ represents the property of each sentence $[x]$. Then, space $([X], [T], [f])$ represents the same text with sentences as its components. The coarse space $([X], [T], [f])$ is called a quotient space of (X, T, f) (Fig. 1).

The coarse universe $[X]$ can be defined in many ways. Generally, universe $[X]$ is defined by an equivalence relation R on X. Then $[X]$ consists of the whole equivalence classes obtained by R; each equivalence class in X is regarded as a component in $[X]$. Universe $[X]$ can also be defined by fuzzy relation and tolerence relation. In these cases, the components may have blurry boundaries or they may overlap each other.

Assume $\Re$ is a family of equivalence relations on X. Define an order relation '<' on $\Re$ as follows. Assume $R_1, R_2 \in \Re$; then, $R_2 < R_1$ if only if for each pair $x, y \in X$, if xR_1y, then xR_2y, where xRy indicates that x and y are R-equivalent. It implies that the universe X_1 corresponding to R_1 is finer than the universe X_2 corresponding to R_2. So the family of quotient spaces defined by $\Re$ is a proper mathematical model of the granulation in cognition.

A problem is represented at different granularities in human cognition. As mentioned before, its corresponding mathematical model is a family of quotient spaces; each describes the problem at a certain abstraction level. The main characteristic of our model is that the structure T is represented at the model explicitly. Based on the 'structure' we investigate the relation and translation among different grain-size spaces. This will facilitate the multi-granular computing and reduce the computational complexity.

The Granularity Relation

The granularity relation can be represented by two basic operations on different grain-size spaces. First, the decomposition operation is the translation of a problem from a coarse level to a fine one. In the crisp-granulation issue, the universe is divided by equivalence relations, i. e., the partition. Certainly, the decomposition can be extended to fuzzy-granulation, tolerance relation, etc. In those cases, either the boundaries among granules are blurry or there is a superposition among granules. By a set of decompositions, we have a set of refined versions of the problem. Second, the projection operation is the translation from a fine level to a coarse one. By a set of projective operations; each translates a fine-grained space to a coarse one. Then we have a set of profiles of the problem, or a set of simplified versions of the problem. There are three basic kinds of projective operations, i. e., based on universe X, structure T or attribute f. By the two operations, both refined and simplified versions of the problem compose a multi-granular model.

Some relations (properties) among different grain-size spaces are shown below.

Truth Preserving Property

Assume a topological space (X, T). By projection p: $(X, T) \rightarrow ([X], [T])$, a coarse space is constructed from (X, T). Since the granule in $[X]$ is larger than that in X, the details of (X, T) are missing in space $([X], [T])$. If the properties of space (X, T) that we are interested in is still

preserved in space $([X],[T])$, then we can also solve the same problem in the simplified space. The truth preserving property among different grain-size worlds will help us to simplify the problem solving.

Assume that R is an equivalence relation on X. From R, we have a quotient set $[X]$. A quotient topology $[T]$ induced from T can be defined as follows.

$$[T] = \left\{u \,\middle|\, p^{-1}(u) \in T, u \subset [X]\right\} .$$

In addition, $p\colon X \to [X]$ is a natural projection and defined as follows.

$$p(x) = [x]$$
$$p^{-1}(u) = \{x \,|\, p(x) \in u\} .$$

From topology [9], we have the following proposition.

Proposition 1 (Truth preserving property) *If a problem $[A] \to [B]$ on $([X],[T],[f])$ has a solution path, and for $\forall [x], p^{-1}([x])$ is connected on X. Then the corresponding problem $A \to B$ has a solution path on (X,T,f) consequentially.*

Proof 1 Since $[A] \to [B]$ has a solution path on $([X],[T],[f])$, $[A]$ and $[B]$ fall on the same connected component C. Let $D = p^{-1}(C)$. We show that D is a connected set on X.

By reduction to absurdity, assume D is a union of two disjoint non-empty open closed sets D_1 and D_2. $\forall a \in C$, $p^{-1}(a)$ is connected on X. $p^{-1}(a)$ only belongs to one of D_1 and D_2. Therefore, $D_i, i = 1,2$ consists of elements of $[X]$, i. e., there exist C_1 and C_2 such that $D_1 = p^{-1}(C_1), D_2 = p^{-1}(C_2)$. Since $D_i, i = 1,2$, are open closed sets, from the definition of natural map p, C_1 and C_2 are non-empty open closed sets on $[X]$ also. Since C_1 and C_2 are the partition of C, C is not a connected set. There is a contradiction. □

In fact, in a topologic space, a problem solving (or reasoning) can be regarded as finding a connected set from the initial state (or promise) to the final state (conclusion), i. e., finding the connectivity of sets in the space. The proposition shows that if there is a solution (connected) path in the coarse-grained space $([X],[T])$, then there exists a solution path in its original space (X,T). It shows that some problems can be solved in its coarse space rather than the original one. For example, the robotic motion planning is to find the collision-free paths in a geometrical space. If a simplified topologic space is constructed properly from the geometrical one, by the truth preserving property, we know that the problem can be transformed into that of finding a connected path in the topological space, as long as we only need to know if there exists any collision-free path. Certainly, finding connected paths in a topological space is much easier than that in the geometrical one.

In order for the truth preserving property to remain between spaces $[X]$ and X, the condition that $\forall [x] \in [X], p^{-1}([x])$ is connected in X should be satisfied; that is, the connectivity of sets should be remained when the grain-size becomes finer. Sometimes, this is hard to come by, and we can obtain the following property.

Falsity Preserving Property

Proposition 2 (Falsity preserving property) *The natural projection $p\colon (X,T) \to ([X],[T])$ is a continuous mapping. If $A \subset X$ is a connected set on X, then $p(A)$ is a connected set on $[X]$.*

It means that the connectivity of sets remains unchanged when the grain-size become coarser. This is easier to obtain than the previous one in real problem solving. In fact, in the quotient space model the human problem solving (or reasoning) can be treated as finding the connectivity of sets in the problem space. The proposition shows that if there is a solution (connected) path in the original space (X,T), then there exists a solution path in its proper coarse-grained space $([X],[T])$. Conversely, in the coarse-grained space, if there does not exist a solution path, there is no solution in the original space. This is called the 'falsity preserving' property, i. e., 'no-solution' (region) property is preserved between quotient spaces. The property underlies the power of human hierarchical problem solving. If at the coarse level we do not find any solution in some regions, then there is no solution in the corresponding regions at the fine level. Therefore, the results obtained from the coarse level can guide the problem solving in the fine level effectively. In general, the coarse space is simpler than the fine one, so the computational complexity will be reduced by the hierarchical problem solving.

By using the structural (topological) relation among the quotient spaces, we have the two basic properties above. So the structure T plays an important role in our model. We cannot always have the proper properties as shown before. Sometime, we can only reach the properties to a certain degree such as in a probabilistic sense.

Hierarchy

There are two basic modes adopted by multi-granular computing, i. e., hierarchy and combination. The hierarchy means hierarchical problem solving strategy, i. e., problem solving process carries on from a coarse level to a fine one step by step. The combination means the integration of information observed from several coarse levels.

The aim of hierarchical problem solving is to reduce the computational complexity by using the information from different grain-size worlds.

The Computational Complexity Under Deterministic Model

Given a problem space (X, T, f), X is a finite set. $|X|$ is the number of components in X. If we solve the problem from (X, T, f) directly, i. e., find the goal from space (X, T, f), the computational complexity is $c(|X|)$. Assume that $c(\cdot)$ only depends on the number of components of X and is independent of its structure or other attributes. The value of $c(\cdot)$ ranges over $[0, \infty)$. Assume that X_0 is a quotient space of X. c is a complexity function of X. Suppose the number of components of X_0 that might contain the goal to be g at most. First, we consider the simplest case: the "truth preserving property" comes into existence precisely, i. e., $g \equiv 1$. Now we have a set $X_1, X_2, \ldots, X_t, X_{t+1} = X$ of quotient spaces, where X_i is a quotient space of X_{i+1}, $i = 1, 2, \ldots, t$. We estimate the asymptotic property of computational complexity denoted by $c_t(|X|)$ under the hierarchical problem solving strategy with t levels. Let $|X| = \mathrm{e}^n$.

Suppose that X_1 is a quotient space of X. We find the goal of X from X_1. At a cost of complexity $c(|X|)$, we find a unique component $a_1 \in X_1$ that contains the goal. Assume that each equivalence class has the same number of components. Now the goal is within component a_1 of X_1. If $|X_1| = \mathrm{e}^{n_1}$, from $|X| = \mathrm{e}^n$, we have $|a_1| = \mathrm{e}^{n-n_1}$. The total complexity for solving problem X by the hierarchical strategy with two levels is

$$c_1(|X|) = c(|X_1|) + c(|a_1|) = c(\mathrm{e}^{n_1}) + c(\mathrm{e}^{n-n_1}) .$$

Regarding a_1 as a set of X, if it's too large, then a_1 can further be decomposed. Assume Y_2 is a quotient space of a_1 and $|Y_2| = \mathrm{e}^{n_2}$. The total complexity for solving X with three hierarchical levels is

$$\begin{aligned} c_2(|X|) &= c(|X_1|) + c(|Y_2|) + c(|a_2|) \\ &= c(\mathrm{e}^{n_1}) + c(\mathrm{e}^{n_2}) + c(\mathrm{e}^{n-n_1-n_2}) . \end{aligned}$$

Now the goal is within the component a_2.

By induction, the total complexity for solving X using hierarchical strategy with t levels is

$$c_t(|X|) = c(\mathrm{e}^{n_1}) + c(\mathrm{e}^{n_2}) + \cdots + c(\mathrm{e}^{n_t}) \qquad \text{where } n = \sum_{i=1}^{t} n_i . \quad (1)$$

In each abstraction level, all components are classified into b equivalence classes, $b > 0$, i. e., $\forall i, \mathrm{e}^{n_i} = b$. Since $n = \sum_{i=1}^{t} n_i$, we have $\mathrm{e}^n = b^t = \mathrm{e}^{at}$, where $b = \mathrm{e}^a$, $at = n$, $t = n/a$. Substituting $t = n/a$ into (1), we have

$$c_t(|X|) = t(c(b)) = c_1 n .$$

Then we have the following proposition:

Proposition 3 *If a unique component that contains the goal can be found at each abstraction level, i. e., $g \equiv 1$, there exists a hierarchical strategy such that X can be solved in a linear time $(O(n))$ in spite of the form of the original complexity $c(|X|)$.*

Example 1 We are given N balls, one of which is known to be lighter or heavier than the rest. Using a two-pan scale, how to find the counterfeit ball such that the number of weighs is minimal. First, in each weigh, we divide the balls into three groups: on the left pan, the right pan and the table denoted by '1', '2' and '0', respectively. After n weighs, each ball is assigned a ternary number with n digits denoted by $a^i = (a_1^i, a_2^i, \ldots, a_n^i)$, $i = 1, 2, \ldots, N$. From the quotient-space model point of view, each weigh corresponds to constructing a quotient space. After n weighs, we have a tree with depth n composed by a sequence of quotient spaces as shown in Fig. 2. Then we solve the problem hierarchically. Second, in each weigh, if the weight of balls in the left pan is lighter than that in the right pan, then denote as '1'. If the weight of balls in the right pan is lighter than that in the left one, then denote as '2'. Otherwise, i. e., when the scale balances, denote as '0'. After n weighs, we have another ternary number $e = (e_1, e_2, \ldots, e_n)$ with n digits. Obviously, if the counterfeit ball is lighter, when $e_1 = 1$, the ball is annotated as '1'; when $e_1 = 2$, the ball is annotated as '2'. Similarly, when $e_1 = 0$, the counterfeit ball is annotated as '0'. Therefore, we have $a^i = e$, when the counterfeit ball is lighter.

Similarly, we have $a^i = \mathrm{e}'$, when the counterfeit ball is heavier, where e' is a conjugate of e, i. e.,

$$\mathrm{e}' = \begin{cases} 1 , & e = 2 \\ 2 , & e = 1 . \\ 0 , & e = 0 \end{cases}$$

In order to find the counterfeit ball uniquely, the tabs cannot contain their conjugates. A ternary number with n digits has 3^n entities and $(3n + 1)/2$ pairs of conjugates. When weighing the balls for the first time, in order to make two pans have the same number of balls, the number of elements having $a_1^i = 1$ and $a_1^i = 2$ in a set $\{a^i = (a_1^i, a_2^i, \ldots, a_n^i), i = 1, 2, \ldots, N\}$ of tabs should be the same. So at most there are $(3^n - 1)/2$ entities of 3^n

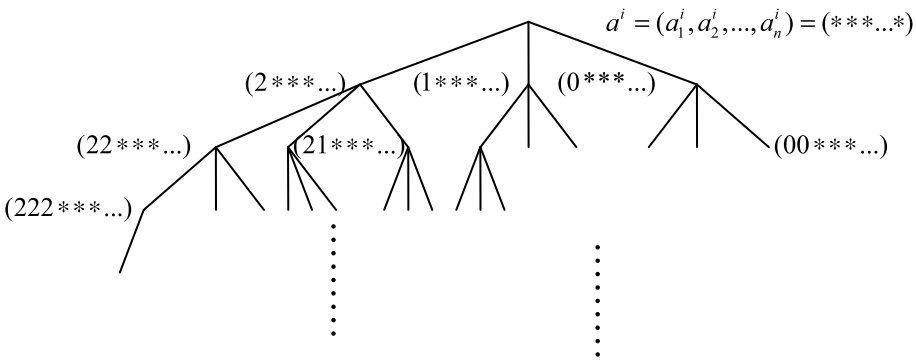

Multi-Granular Computing and Quotient Structure, Figure 2
The quotient-space model of ball weigh problem

ternary numbers with n digits that can be used as tabs. Therefore, we have the following properties.

Property 1 *A counterfeit ball can be found from* $(3^n - 1)/2$ *balls in n weighs.*

Property 2 *A set K of tabs for weighing balls can be constructed by a set of ternary numbers with n digits that satisfies the following conditions.*

(1) At most one of each pair of conjugates is in set K
(2) The numbers of elements having $a_1^i = 1$ *and* $a_1^i = 2$ *in a set* $\{a^i = (a_1^i, a_2^i, \ldots, a_n^i), i = 1, 2, \ldots, N\}$ *of tabs are the same.*

Property 3 *After n weighs, we have result e. If* $e \in K, e \neq (0, 0, \ldots, 0)$*, the tab of the counterfeit ball is e, and the ball is lighter. If* $e = (0, 0, \ldots, 0)$*, the tab of the counterfeit ball is e and we don't know whether it's lighter or heavier. Otherwise, if* $e \notin K$*, the tab of the counterfeit ball is* e' *and it's heavier.*

From Property 2, we know the whole optimal solution strategies of the problem. Any process with finite state can be represented by a set of binary numbers (or ternary, decimal numbers, etc.) i. e., a sequence of quotient spaces. Then the process can be managed hierarchically.

Finally, we have the following algorithm.

Algorithm 1 *A set K of tabs satisfies Property 2. Each ball is assigned a tab from K. Then we have the whole tabs denoted by* $a^i = (a_1^i, a_2^i, \ldots, a_n^i), i = 1, 2, \ldots, N$*. Assume that we have result* $(e_1, e_2, \ldots, e_m)$ *after m weighs. Taking out all balls assigned by* $(e_1, e_2, \ldots, e_m, a_{m+1}^i, \ldots)$ *and* $(e_1', e_2', \ldots, e_m', a_{m+1}^i, \ldots)$*, i. e., its first m components are* $(e_1, e_2, \ldots, e_m)$ *and* $(e_1', e_2', \ldots, e_m')$*, if* $a_{m+1}^i = 1$*, put the ball on the left pan; if* $a_{m+1}^i = 2$*, put the ball on the right pan; otherwise,* $a_{m+1}^i = 0$*, put it on the table. The process finishes after n weighs. If necessary, we need to put some true balls on the left and right pans such that the numbers of balls on two pans are the same.*

Combination

However, in human cognition, one usually learns things from local fragments, integrates them and forms a global picture gradually. It means inferring the fine-level representation from the information collected at coarse levels. The process is called combination (or information fusion).

For a problem space (X, T, f), given the knowledge of its two quotient spaces (X_1, T_1, f_1) and (X_2, T_2, f_2), the 'combination' operation is intended to have an overall understanding of (X, T, f) from the known knowledge.

Let (X_3, T_3, f_3) be the combination of (X_1, T_1, f_1) and (X_2, T_2, f_2), and $p_i : (X, T, f) \to (X_i, T_i, f_i), i = 1, 2$. In order to have a proper (X_3, T_3, f_3), the following three combination principles should be satisfied at least.

$$\begin{cases} p_i X_3 = X_i \\ p_i T_3 = T_i \\ p_i f_3 = f_i\,, \quad i = 1, 2 \end{cases}$$

In general, the solution satisfying the three principles is not unique. In order to have a unique result, some criteria must be added such that the solution is optimal. First, we propose the following combination rule. Let the combination universe X_3 be the least upper bound of universes X_1 and X_2. This implies that X_3 is the coarsest one among the universes that satisfy the first combination principle or the finest one that we can get from the known universes X_1 and X_2. And let the combination topology T_3 be the least upper bound of topologies T_1 and T_2. This implies that T_3 is the coarsest one among the topologies that satisfy the second combination principle or the finest one that we

can get from the known topologies T_1 and T_2. This is also the maximal amount of information that we can obtained from the known knowledge. Therefore, the proposed X_3 and T_3 are optimal in some sense.

In most cases, the optimal criteria are domain dependent. However, here we present a general criterion as follows.

$$D(f_3, f_1, f_2) = \min_f D(f, f_1, f_2) \text{ or } \max_f D(f, f_1, f_2) .$$

Where f ranges over all attribute functions on X_3 that satisfy the third combination principle.

It's noted that in the combination principle $p_i f_3 = f_i, i = 1, 2$, where p_i may be a non-deterministic mapping. We'll discuss the problem in Sect. "Combination".

To show the rationality of the above combination principles, in [25] we deduced the famous Dempster–Shafer combination rule in belief theory [21] by the principles. It shows that the D–S rule is the outcome of the combination principles under certain optimal criteria.

Now we give an example to show how to use the combination principle.

Example 2 Semantic concept detection, which is also called high-level feature extraction in TRECVID (TREC-Video Retrieval Evaluation) [19], is to automatically determine related video shots from a video dataset given some semantic concepts. This technology is very useful for automatic or semi-automatic video indexing or annotation. Usually, a small portion of data is annotated manually as training data for machine learning programs. The result of concept detection can be measured by the average precision, which summarizes the precision at all recall levels and favors algorithms that detect more relevant shots [29].

In semantic concept detection, a video shot is described by many features (attributes) with different modalities such as auto-speech recognition text, visual texture, color of segmented image regions and is related to some semantic concept. The concept detection is to determine the relevant concept of video shots from their features.

The Construction of Quotient-Space Model

A set $\{C^1, C^2, \ldots, C^i, \ldots, C^k\}$ of concepts and a video archive are given. Then we may extract a set of multi-granular and multi-modular features (speech, image and text) from each video shot. A quotient-space model for concept detection is constructed as follows. In the model, universe X represents a set of video shots. Based on some extracted feature of video shots, for example, the color of segmented image regions, X is classified into a set $\{a_1, a_2, \ldots, a_k\}$ of classes; each corresponds to a concept C^i. Letting $X_1 = \{a_i\}, i = 1, 2, \ldots, k$, X_1 is called a projection of X, the quotient space of X. Each class a_i consists of a set of video shots and corresponds to concept C^i. Based on another feature, for example, the auto-speech recognition text, the same X is classified into a new set $\{b_1, b_2, \ldots, b_k\}$ of classes. Then, we have a new space $X_2 = \{b_j\}, j = 1, 2, \ldots, k$ (Fig. 3).

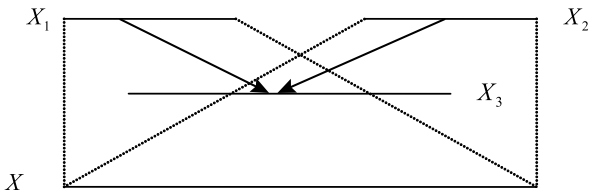

Multi-Granular Computing and Quotient Structure, Figure 3
The combination of quotient spaces

More quotient spaces can be constructed by using different features (attributes). Then, we have a set $\{(X, T, f), (X_{n-1}, T_{n-1}, f_{n-1}), \ldots, (X_0, T_0, f_0)\}$ of quotient spaces. T represents the relation among the sets of video shots. f is the description of semantic concepts of video shots that we are intended to handle. The problem is how to construct a new space X_3 based on the known information of X_1 and X_2 such that more relevant shots can be detected from X_3 when given a semantic concept. This is the combination problem in quotient-space model theory. Since a big gap is existed between a low-level feature and a semantic concept, the correspondence between them established by a small portion of training data has a prodigious uncertainty. To reduce the uncertainty, it generally uses the knowledge from different modalities.

The Combination of Quotient Spaces

Assume that (X_1, T_1, f_1) and (X_2, T_2, f_2) are the knowledge about $A = (X, T, f)$ in two different abstraction levels. The combination of (X_1, T_1, f_1) and (X_2, T_2, f_2) is defined as a new abstraction level of A denoted by (X_3, T_3, f_3), which satisfies the following three basic principles.

(1) X_1 and X_2 are quotient spaces of X_3
(2) T_1 and T_2 are quotient structures of T_3
(3) f_1 and f_2 are projections of f on X_1 and X_2, respectively (X_3, T_3, f_3) might need to satisfy some other optimal criteria.

We next discuss the combination of universe X and attribute function f.

The Combination of Universes Assume that (X_1, T_1, f_1) and (X_2, T_2, f_2) are quotient spaces of (X, T, f). R_1 and R_2 are equivalence relations with respect to X_1 and X_2, respectively.

Define the combination universe X_3 as follows.

$xR_3y \Leftrightarrow xR_1y$ and xR_2y, where R_3 is an equivalence relation with respect to X_3.

It is known that all equivalence classes on X form a semi-order lattice under the relation $<$, where $R_1 < R_2 \Leftrightarrow$ if xR_2y then xR_1y. In terms of the semi-order lattice, the combination of R_1 and R_2 can be defined as follows.

Definition 1 Assume that R_1 and R_2 are two equivalence relations on X. If R_3 is the least upper bound of R_1 and R_2 among the semi-order lattice, then R_3 is the combination of R_1 and R_2. If $X_1 = \{a_i\}$ and $X_2 = \{b_j\}$ are two partitions with respect to R_1 and R_2, respectively, then the combination of X_1 and X_2 can be represented by $X_3 = \{a_i \cap b_j \,|\, a_i \in X_1, b_j \in X_2\}$.

In the concept detection of video shots, X_1 is the partition of X by the speech feature and X_2 is the partition of X by the image feature, the combination universe X_3 is the partition of X by both speech feature and image feature. Obviously, X_3 is the finest universe which we can get from X_1 and X_2. It is also the coarsest one among the universes which satisfy the above combination principle, that is, the least upper bound.

The Combination of Attribute Functions Given (X_1, T_1, f_1) and (X_2, T_2, f_2), we find the combination space (X_3, T_3, f_3) satisfying the following conditions.

(1) $p_i f_3 = f_i, i = 1, 2$, where $p_i : (X_3, T_3, f_3) \to (X_i, T_i, f_i)$ is a natural projection
(2) $D(f, f_1, f_2)$ is a given criterion such that

$$D(f_3, f_1, f_2) = \min_f D(f, f_1, f_2) \text{ or } \max_f D(f, f_1, f_2) .$$

Where f ranges over all attribute functions on X_3 that satisfy condition (1). In fact, the solution satisfying condition (2) is not unique. The additional optimization criteria are generally needed in order to have a unique solution. There have been several kinds of combination approaches [6,12,17], sometimes called information fusion. We show one of the possible ways.

The Combination of Attribute Functions

A set $\{x^1, x^2, \ldots, x^N\}$ of video shots and a set $\{C^1, C^2, \ldots, C^i, \ldots, C^k\}$ of concepts are given. Using g different kinds of features such as speech, image and text, to classify the video shots by the annotated training samples, then we have g quotient spaces $\{X_1, X_2, \ldots, X_g\}$. For a concept C^i, there are g different classifications $\{C_1^i, C_2^i, \ldots, C_g^i\}$ on g quotient spaces, respectively; each class $C_j^i, j = 1, 2, \ldots, g$, consists of a set of video shots. Assume that each sample $x^i, i = 1, 2, \ldots, p$, of video shots is an identically independent distributed n-dimensional random variable. Then class $C_j^i, j = 1, 2, \ldots, g$, is a set of random variables; each can be described by a probability density function. We choose normal distribution function $N(x, \mu, \Sigma)$ as its probabilistic model, where mean μ may be defined as the center of the class, and $\sum$ the variance matrix. Then concept C^i can be defined as the combination of $C_j^i, j = 1, 2, \ldots, g$, as follows.

$$F_i(x) = \sum_j \alpha_j N(x, \mu_j, \Sigma_j), \quad j = 1, 2, \ldots, g .$$

$F_i(x)$ is a describer of concept C^i. From the combination principles in quotient space model, the weights α_j can be chosen by some sort of optimization techniques. Since $F_i(x)$ is a g-component finite mixture density, the maximum likelihood estimator can be used to estimate the weights. According to the iterative EM (Expectation Maximization) algorithm presented in [7], we have the following optimization procedures.

Let $K = \{(x^1, y^1), \ldots, (x^p, y^p)\}, x^i \in R^n, y^i \in \{0, 1\}^k\}$ be a set of training samples. And its corresponding classification based on the set $\{C^1, C^2, \ldots, C^i, \ldots, C^k\}$ of concepts is denoted by $C^i = \{C_1^i, C_2^i, \ldots, C_g^i\}, i = 1, 2, \ldots, k$.

Initialization Let $\alpha_j^{(0)} = d_j, d_j$ is the proportion of number of video shots in C_j^i to the total number of video shots in the ith concept.

$$\sigma_j^{(0)} = r_j, \quad r_j \text{ is the radius of } C_j^i .$$
$$\mu_j^{(0)} = a_j, \quad a_j \text{ is the center of } C_j^i .$$
$$\Sigma_j^{(0)} = (\sigma_j^{(0)})^2 I_n, \quad \text{where } I_n \text{ is a } n\text{-dimensional unit matrix} .$$

E Step The $(k+1)$-iteration, calculate the posterior probability of sample x^i from the jth component as follows.

$$\beta_{ij}^{(k)} = \beta_j\left(x^i, \Theta^{(k)}\right) = \frac{\alpha_j^{(k)} N\left(x^i, \mu_j^{(k)}, \Sigma_j^{(k)}\right)}{\sum_{j=1}^{g} \alpha_j^{(k)} N\left(x^i, \mu_j^{(k)}, \Sigma_j^{(k)}\right)}, \quad (j = 1, \ldots, g; \; i = 1, \ldots, p), \quad \text{(E-1)}$$

Multi-Granular Computing and Quotient Structure, Table 1
Comparison of uni-modal and multi-modal semantic video concept detection results

	Uni-Modal			Cross-Modal			
	ASR	**Texture**	**Region**	**A+T**	**A+R**	**T+R**	**A+T+R**
US-flag	0.0335	0.0155	0.0375	0.0359	0.0506	0.0372	0.0521
Water	0.0034	0.1143	0.0814	0.1022	0.0735	0.1333	0.1211
Mountain	0.0033	0.0693	0.1104	0.0668	0.1066	0.1176	0.1154
Sports	0.0723	0.0769	0.2156	0.1465	0.2678	0.2802	0.3050
Average	0.0281	0.0690	0.1112	0.0879	0.1246	0.1421	0.1484

$$\alpha_j^{(k+1)} = \frac{1}{p}\sum_{i=1}^{p}\beta_{ij}^{(k)}\,, \quad j = 1,\ldots,g\,. \tag{E-2}$$

M Step Find the mean and variance matrices by iteration.

$$\mu_j^{(k+1)} = \frac{\sum_{i=1}^{p}\beta_{ij}^{(k)}x^i}{\sum_{i=1}^{p}\beta_{ij}^{(k)}}\,, \quad \text{where } \beta_{ij}^{(k)} = \beta_j(x^i, \Theta^{(k)})\,,$$
$$(j = 1,\ldots,g\,;\ \ i = 1,\ldots,p)\,, \tag{M-1}$$

$$\Sigma_j^{(k+1)} = \frac{\sum_{i=1}^{p}\beta_{ij}^{(k)}(x^i - \mu j^{(k+1)})(x^i - \mu_j^{(k+1)})^T}{\sum_{i=1}^{p}\beta_{ij}^{(k)}}\,,$$
$$(j = 1,\ldots,g)\,. \tag{M-2}$$

Assume $N(x, \mu, \sum)$ is a 1-dimensional normal distribution.

$$(\sigma_j^2)^{(k+1)} = \frac{\sum_{i=1}^{p}\beta_{ij}^{(k)}\left|(x^i - \mu_j^{(k+1)})\right|^2}{\sum_{i=1}^{p}\beta_{ij}^{(k)}}\,.$$

Finally, the function $F_i(x)$ that we have is the describer (attribute) of C^i. The function $F(x) = \{F_1(x),\ldots,F_p(x)\}$ is the decision function of the set $\{C^1, C^2,\ldots,C^i,\ldots,C^n\}$ of concepts. $F(x)$ integrates all information from g quotient spaces.

In order to show the advantage of multi-granular and multi-modal computing, Our experiments were carried out on the TRECVID 2005's test dataset for the high-level feature extraction (HFE) task [19], which consists of 86.6 hours of news videos (45766 shots in 140 video clips).

Three uni-modal methods and four multi-modal methods are compared. Uni-modal results are chosen from TRECVID 2005 submissions. They include

- *ASR*, the 7th run from Fudan University [22], based on auto-speech recognition text, which is a sub-modality of the text;
- *Texture*, the 7th run from National University of Singapore [5], based on visual texture, which is a visual sub-modality; and
- *Region*, the 1st run from Tsinghua University [23], based on color of segmented image regions, which is also a visual sub-modality.

Therefore, the four multi-modal results are denoted as '$A+T$', '$A+R$', '$T+R$', and '$A+T+R$', respectively. We use the Bayes rule of Probabilistic Model Supported Rank Aggregation (*PMSRA*) [8] to integrate different modalities or sub-modalities. The distribution family adopted is the Gauss distribution.

Four concepts are chosen for comparison. They belong to different types of concept. '*US-flag*' is a concept of object. '*Water*' and '*Mountain*' are typical concepts of scenes. '*Sports*' is a kind of event.

Average precisions of the results based on all the uni-modal and multi-modal methods are shown in Table 1. On average, multi-modal strategies are significantly better than the uni-modal ones. Specifically, we can see that if there are no great disparities in performances, multi-modal methods always bring significant improvement. On the other hand, if their performances are too different, which suggests the inconsistency between modalities, the average precision based on multi-modal method may be reduced a litte.

Future Directions

The quotient space model that we discuss previously is defined by equivalence relations. The model should be extended to more general cases such as fuzzy relation, consistency relation. In these cases, how to construct a proper quotient structure from the original one, if the granularity relation such as "truth preserving" and "falsity preserving" properties still remains, more research works

should be done in the future. Unfortunately, only a few works [16,27,28] dealt the problems recently.

Acknowledgments

The work is supported by the National Natural Science Foundation of China under Grant. No. 60621062, and the National Key Foundation R&D Project under Grant. No.2003CB317007, 2007CB311003.

Bibliography

1. Bargiela A, Pedrycz W (2002) Granular computing: an introduction. Kluwer, Boston
2. Bertino E, Cuadra D, Martinez P (2005) An object-relational approach to the representation of multi-granular spatio-temporal data. In: The 17th conference on advanced information systems engineering (CAiSE'05), Porto, 13–17 June, pp 119–134
3. Bettini C, Dyreson CE, Evans WS, Snodgrass RT, Wang XS (1998) A glossary of time granularity concepts. In: Etzioni O, Jajodia S, Sripada S (eds) Temporal databases: research and practice. Lecture Notes in Computer Science, vol 1399. Springer, Berlin, pp 406–413
4. Bettini C, Wang XS, Jajodia S (2002) Solving multi-granularity temporal constraint networks. Artif Intell 140:107–152
5. Chua T-S et al (2005) TRECVID 2005 by NUS PRIS. In: TRECVID 2005. NIST. http://www-nlpir.nist.gov/projects/tvpubs/tv5.papers/nus.pdf
6. Dasarathy BV (2001) Information fusion – what, where, why, when, and how? Ed Inform Fusion 2(2):75–76
7. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data using the EM algorithm (with discussion). J R Stat Soc Ser B 39:1–38
8. Ding DY (2007) Research on information fusion technology and its application in video content analysis. Tsinghua University Dissertation. (in Chinese)
9. Eisenberg M (1974) Topology. Holt
10. Garland M (1999) Multiresolution modeling: survey and future opportunities. EUROGRAGHICS'99. State of the Art Report, September 1999
11. Iyengar SS, Prasad L (1997) Wavelet analysis with applications to image processing. CRC Press, Baca Raton
12. Jing F et al (2005) A unified framework for image retrieval using keyword and visual features. IEEE Trans Image Process 14(7):979–989
13. Lin TY (1989) Neighborhood systems and approximation in database and knowledge base systems. In: Ras ZW (ed) Proc of the fourth international symposium on methodologies of intelligent systems. Poster session. 12–15 Oct. North Holland, New York, pp 75–86
14. Lin TY (1992) Topological and fuzzy rough sets In: Slowinski R (ed) Decision support by experience – application of the rough sets theory. Kluwer, Dordrecht, pp 287–304
15. Lin TY (2001) Granular fuzzy sets: a view from rough set and probability theories. Int J Fuzzy Syst 3(2):373–381
16. Lin TY (2005) Churn-Jung Liau granular computing and rough sets. In: Maimon O, Rokach L (eds) The Data Mining and Knowledge Discovery Handbook. Spriger, pp 535–561
17. Maes F et al (1997) Multimodality image registration by maximization of mutual information. IEEE Trans Med Imaging 16(2):187–198
18. Nguyen SH, Skowron A, Stepaniuk J (2001) Granular computing: a rough set approach. Comput Intell 17:514–544
19. Over P, Kraaij W, Smeaton AF (2005) TRECVID 2005 – an introduction. In: TRECVID 2005, NIST. http://www-nlpir.nist.gov/projects/tvpubs/tv5.papers/tv5intro.pdf
20. Pawlak Z (1998) Granularity of knowledge, indicernibility, and rough sets. In: Proc of IEEE world congress on computational intelligence, vol 1. IEEE Press, New York, pp 106–110
21. Shafer G (1976) A mathematical theory of evidence. Princeton University Press, Princeton
22. Xue X et al (2005) Fudan University at TRECVID 2005. In: TRECVID 2005. NIST. http://www-nlpir.nist.gov/projects/tvpubs/tv5.papers/Fudan.pdf
23. Yuan J et al (2005) Tsinghua University at TRECVID 2005. In: TRECVID 2005. NIST. http://www-nlpir.nist.gov/projects/tvpubs/tv5.papers/tsinghua.pdf
24. Zadeh LA (1997) Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets Syst 19:111–127
25. Zhang B, Zhang L (1992) Theory and applications of problem solving. Elsevier, North-Holland
26. Zhang L, Zhang B (2004) The quotient space theory of problem solving. Fundament Inform 59(2,3):287–298
27. Zhang L, Zhang B (2005) Fuzzy reasoning model under quotient space structure. Inform Sci 173:353–364
28. Zhang L, Zhang B (2007) The theory and application of problem solving – the theory and application of quotient-space based granular computing. Tsinghua University Publishing Co. (the second version, in Chinese) Bejing
29. Zhu M (2004) Recall, precision and average precision. Department of Statistics and Actuarial Science, University of Waterloo, Waterloo

Multimillion Atom Simulations with Nemo3D

Shaikh Ahmed* [1,2], Neerav Kharche* [1], Rajib Rahman* [1], Muhammad Usman* [1], Sunhee Lee* [1], Hoon Ryu[1], Hansang Bae[1], Steve Clark[3], Benjamin Haley[1], Maxim Naumov[4], Faisal Saied[3], Marek Korkusinski[5], Rick Kennel[3], Michael McLennan[3], Timothy B. Boykin[6], Gerhard Klimeck[1,7]

[1] School of Electrical and Computer Engineering and Network for Computational Nanotechnology, Purdue University, West Lafayette, USA

[2] Electrical and Computer Engineering Department, Southern Illinois University, Carbondale, USA

[3] Rosen Center for Advanced Computing, Purdue University, West Lafayette, USA

[4] Department of Computer Science, Purdue University, West Lafayette, USA

* Authors contributed equally

[5] Institute for Microstructural Sciences, National Research Council of Canada, Ottawa, Canada
[6] Electrical and Computer Engineering Dept., The University of Alabama, Huntsville, USA
[7] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Nanoscale Device Modeling and Simulation Challenges
NEMO 3-D Simulation Package
Simulation Results
Summary and Future Directions
Acknowledgments
Bibliography

Glossary

Nanostructures Nanostructures have at least two physical dimensions of size less than 100 nm. Their size lies between atomic/molecular and microscopic structures/particles. Realistically sized nanostructures are usually composed of millions of atoms. These devices demonstrate new capabilities and functionalities where the quantum nature of charge carriers plays an important role in determining the overall device properties and performance.

Quantum dots Quantum dots (QDs) are solid-state nanostructures that provide confinement of charge carriers (electrons, holes, excitons) in all three spatial dimensions typically on the nanometer scale. This work focuses on semiconductor based quantum dots.

Atomistic simulation For device sizes in the range of tens of nanometers, the atomistic granularity of constituent materials cannot be neglected. Effects of atomistic strain, surface roughness, unintentional doping, the underlying crystal symmetries, or distortions of the crystal lattice can have a dramatic impact on the device operation and performance. In an atomistic simulation, one takes into account both the atomistic/granular and quantum properties of the underlying nanostructure.

Strain Strain is the deformation caused by the action of stress on a physical body. In nanoelectronic devices, strain typically originates from the assembly of lattice-mismatched semiconductors. Strain can be atomistically inhomogeneous and a small mechanical distortion of 2–5% can strongly modify the energy spectrum, in particular the optical bandgap, of the system by 30–100%.

Band structure Band structure of a solid originates from the wave nature of particles and depicts the allowed and forbidden energy states of electrons in the material. The knowledge of the band structure is the first and essential step towards the understanding of the device operation and reliable device design for semiconductor devices. Bandstructure is based on the assumption of an infinitely extended (bulk) material without spatial fluctuations (outside a simple repeated unit cell). For nanometer scale devices with spatial variations on the atomic scale the traditional concept of bandstructure is called into question.

Piezoelectricity A variety of advanced materials of interest, such as GaAs, InAs, GaN, are piezoelectric. Piezoelectricity arises due to charge imbalances on the bonds between atoms. Modifications of the bond angles or distances result in alterations in charge imbalance. Any spatial non-symmetric distortion/strain in nanostructures made of these materials will create piezoelectric fields, which may significantly modify the electrostatic potential landscape.

Tight binding Tight binding is an empirical model that enables calculation of single-particle energies and wave functions in a solid. The essential idea is the representation of the electronic states of the valence electrons with a local basis that contains the critical physical elements needed. The basis may contain orthogonal *s*, *p*, *d* orbitals on one atom that connect/talk to orbitals of a neighboring atom. The connection between atoms and the resulting overlapping wavefunctions form the bandstructure of a solid.

NEMO 3-D NEMO 3-D stands for NanoElectronic Modeling in three dimensions. This versatile, open source software package currently allows calculating single-particle electronic states and optical response of various semiconductor structures including bulk materials, quantum dots, impurities, quantum wires, quantum wells and nanocrystals.

nanoHUB The nanoHUB is a rich, web-based resource for research, education and collaboration in nanotechnology (www.nanoHUB.org). It was created by the NSF-funded Network for Computational Nanotechnology (NCN) with a vision to pioneer the development of nanotechnology from science to manufacturing through innovative theory, exploratory simulation, and novel cyberinfrastructure. The nanoHUB offers online nanotechnology simulation tools which one can freely access from his/her web browser.

Rappture Rappture (www.rappture.org) is a software

toolkit that supports and enables the rapid development of graphical user interfaces (GUIs) for different applications. It is developed by Network for Computational Nanotechnology at Purdue University, West Lafayette.

Definition of the Subject

The rapid progress in nanofabrication technologies has led to the emergence of new classes of nanodevices and structures which are expected to bring about fundamental and revolutionary changes in electronic, photonic, computation, information processing, biotechnology, and medical industries. At the atomic scale of novel nanostructured semiconductors the distinction between new device and new material is blurred and device physics and material science meet. The *quantum mechanical effects* in the electronic states of the device and the *granular, atomistic* representation of the underlying material become important. Modeling and simulation approaches based on a *continuum* representation of the underlying material typically used by device engineers and physicists become invalid. Typical ab initio methods used by material scientists do not represent the bandgaps and masses precisely enough for device design or they do not scale to *realistically sized devices which may contain millions of atoms*. The variety of geometries, materials, and doping configurations in semiconductor devices at the nanoscale suggests that a *general* nanoelectronic modeling tool is needed. The Nanoelectronic Modeling tool (NEMO 3-D) has been developed to address these needs. Based on the atomistic valence-force field (VFF) method and a variety of nearest-neighbor tight-binding models (s, sp^3s^*, $sp^3d^5s^*$), NEMO 3-D enables the computation of strain for over 64 million atoms and of electronic structure for over 52 million atoms, corresponding to volumes of $(110\,\text{nm})^3$ and $(101\,\text{nm})^3$, respectively. Such extreme problem sizes involve very large-scale computations, and NEMO 3-D has been designed and optimized to be scalable from single CPUs to large numbers of processors on commodity clusters and the most advanced supercomputers. Excellent scaling to 8192 cores/CPUs has been demonstrated. NEMO 3-D is continually developed by the Network for Computational Nanotechnology (NCN) under an open source license. A web-based online interactive version for educational purposes is freely available on the NCN portal http://www.nanoHUB.org. This article discusses the theoretical models, essential algorithmic and computational components, and optimization methods that have been used in the development and the deployment of NEMO 3-D. Also, successful applications of NEMO 3-D are demonstrated in the atomistic calculation of single-particle electronic states of the following realistically-sized nanostructures each consisting of multimillion atoms: (1) self-assembled quantum dots including long-range strain and piezoelectricity; (2) stacked quantum dots as used in quantum cascade lasers; (3) Phosphorus (P) impurities in Silicon used in quantum computation; (4) Si on SiGe quantum wells (QWs); and (5) SiGe nanowires. These examples demonstrate the broad NEMO 3-D capabilities and indicate the necessity of multimillion atomistic electronic structure modeling.

Introduction

Emergence of Novel Nanoscale Semiconductor Devices

The new industrial age and the new economy are driven in large measure by unprecedented advances in information technology. The electronics industry is the largest industry in the world with global sales of over one trillion dollars since 1998. If current trends continue, the sales volume of the electronics industry is predicted to reach three trillion dollars and account for about 10% of gross world product (GWP) by 2010 [93]. Basic to the electronic industry and the new information age are the *semiconductor devices* that implement all needed information processing operations. The revolution in the semiconductor industry was initiated in 1947 with the invention and fabrication of point-contact bipolar devices on *slabs* of polycrystalline germanium (Ge) used as the underlying semiconductor element [1]. Later the development of the planar process and the reliable and high-quality silicon dioxide (SiO_2) growth on silicon wafers, acting as an excellent barrier for the selective diffusion steps, led to the invention of the silicon-based bipolar integrated circuits in 1959. A metal–oxide–semiconductor field-effect transistor (MOSFET), the most critical device for today's advanced integrated circuits, was reported by Kahng and Atalla in 1960 [93]. By 1968, both complementary metal–oxide–semiconductor devices (CMOS) and polysilicon gate technology allowing self-alignment of the gate to the source/drain of the device had been developed. The industry's transition from bipolar to CMOS technology in the 1980s was mainly driven by the increased power demand for high-performance integrated circuits.

The most important factor driving the continuous device improvement has been the semiconductor industry's relentless effort to reduce the cost *per function* on a chip [96]. This is done by putting more devices on an integrated circuit chip while either reducing manufacturing costs or holding them constant. *Device scaling*, which involves reducing the transistor size while keeping the elec-

tric field constant from one generation to the next, has paved the way for a continuous and systematic increase in transistor density and improvements in system performance (described by Moore's Law [69]) for the past forty years. For example, regarding conventional/classical silicon MOSFETs, the device size is scaled in all dimensions, resulting in smaller oxide thickness, junction depth, channel length, channel width, and isolation spacing. Currently, 65 nm (with a physical gate length of 35 nm) is the state-of-the-art process technology, but even smaller dimensions are expected in the very near future.

However, recent studies by many researchers around the globe reveal the fact that the exponential growth in integrated circuit complexity as achieved through conventional scaling is finally facing its limits and will slow down in very near future. Critical dimensions, such as transistor gate length and oxide thickness, are reaching physical limitations [96]. Maintaining dimensional integrity at the limits of scaling is a challenge. Considering the manufacturing issues, photolithography becomes difficult as the feature sizes approach the wavelength of ultraviolet light. In addition, it is difficult to control the oxide thickness when the oxide is made up of just a few monolayers. Processes will be required approaching atomic-layer precision. In addition to the processing issues there are also some fundamental device issues [103]. As the silicon industry moves into the 45 nm node regime and beyond, two of the most important challenges facing us are the growing dissipation of *standby power* and the increasing variability and mismatch in device characteristics.

The Semiconductor Industry Association (SIA) forecasts [88] that the current rate of transistor performance improvement can be sustained for another 10 to 15 years, but only through the development and introduction of *new materials and transistor structures*. In addition, a major improvement in lithography will be required to continue size reduction. It is expected that these new technologies may extend MOSFETs to the 22 nm node (9 nm physical gate length) by 2016. Intrinsic device speed may exceed 1 THz and integration densities will be more than 1 billion transistors/cm^2. In many cases, the introduction of a new material requires the use of a new device structure, or vice versa. To fabricate devices beyond current scaling limits, IC companies are simultaneously pushing the planar, bulk silicon CMOS design while exploring alternative gate stack materials (high-k dielectric [108] and metal gates), band engineering methods (using strained Si [102] or SiGe [72]), and alternative transistor structures. The concept of a band-engineered transistor is to enhance the mobility of electrons and/or holes in the channel by modifying the band structure of silicon in the channel in a way such that the physical structure of the transistor remains substantially unchanged. This enhanced mobility increases the transistor transconductance (g_m) and on-drive current (I_{on}). A SiGe layer or a strained-silicon on relaxed SiGe layer is used as the enhanced-mobility channel layer. Today there is also an extensive research in double-gate (DG) structures, and FinFET transistors [23], which have better electrostatic integrity and theoretically have better transport properties than single-gated FETs. Some novel and revolutionary technology such as carbon nanotubes, silicon nanowires, or molecular transistors might be seen on the horizon, but it is not obvious, in view of the predicted future capabilities of CMOS, how competitive they will be.

A recent analysis based on fundamental quantum mechanical principles, restated by George Bourianoff of the Intel Corporation, reveals that heat/power dissipation will ultimately limit any logic device using an electronic charge [107] and operating at room temperature. This limit is about 100 watts per square centimeter for passive cooling techniques with no active or electrothermal elements. These fundamental limits have led to pessimistic predictions of the imminent end of technological progress for the semiconductor industry and simultaneously have increased interest in advanced alternative technologies that rely on something other than electronic charge—such as spin or photon fields—to store computational state. Many advocate a focus on quantum computers that make use of distinctively quantum mechanical phenomena, such as entanglement and superposition, to perform operations on data. Among a number of quantum computing proposals, the Kane scalable quantum computer is based on an array of individual phosphorus (P) donor atoms embedded in a pure silicon lattice [41]. Both the nuclear spins of the donors and the spins of the donor electrons participate in the quantum computation. The Loss–DiVincenzo quantum computer [63], also a scalable semiconductor-based quantum computer, makes use of the intrinsic spin degree of freedom of individual electrons confined to quantum dots as qubits.

Since the invention of the point-contact bipolar transistor in 1947, advanced fabrication technologies, introduction of new materials with unique properties, and broadened understanding of the underlying physical processes have resulted in tremendous growth in the number and variety of semiconductor devices and literally changed the world. To date, there are about 60 major devices, with over 100 device variations related to them. A list of most of the basic semiconductor devices (mainly based on Ref. [93]) discovered and used over the past century with the date of their introduction is shown in Table 1.

Multimillion Atom Simulations with Nemo3D, Table 1
Major semiconductor devices with the approximate date of their introduction

1874	Metal-semiconductor contact
1947	Bipolar junction transistors (BJT)
1954	Solar cell
1957	Heterojunction bipolar transistor (HBT)
1958	Tunnel diode
1959	Integrated circuits
1960	Field-effect transistors (FETs)
1962	Semiconductor lasers.
1966	Metal-semiconductor FET
1967	Nonvolatile semiconductor memory
1974	Resonant tunneling diode (RTD)
1990	Magnetoresistive Random Access Memory (MRAM)
1991	Carbon nanotubes
1994	Room-temperature single-electron memory cell (SEMC)
1994	Quantum Cascade Laser
1998	Carbon nanotube FET
1998	Proposal for Kane quantum computer
2001	15 nm MOSFET
2003	High performance Silicon nanowire FET

Need for Simulations

Simulation is playing key role in device development today. Two issues make simulation important [96]. Product cycles are getting shorter with each generation, and the demand for production wafers shadows development efforts in the factory. Consider the product cycle issue first. In order for companies to maintain their competitive edge, products have to be taken from design to production in less than 18 months. As a result, the development phase of the cycle is getting shorter. Contrast this requirement with the fact that it takes 2–3 months to run a wafer lot through a factory, depending on its complexity. The specifications for experiments run through the factory must be near the final solution. While simulations may not be completely predictive, they provide a good initial guess. This can ultimately reduce the number of iterations during the device development phase.

The second issue that reinforces the need for simulation is the production pressures that factories face. In order to meet customer demand, development factories are making way for production space. It is also expensive to run experiments through a production facility. The displaced resources could have otherwise been used to produce sellable product. Again, device simulation can be used to decrease the number of experiments run through a factory. Device simulation can be used as a tool to guide manufacturing down a more efficient path, thereby decreasing the development time and costs.

Besides offering the possibility to test hypothetical devices which have not (or could not have) yet been manufactured, device simulation offers unique insight into device behavior by allowing the observation of internal phenomena that can not be measured. Thus, a critical facet of the nanodevices development is the creation of simulation tools that can quantitatively explain or even predict experiments. In particular it would be very desirable to explore the design space before, or in conjunction with, the (typically time consuming and expensive) experiments. A general tool that is applicable over a large set of materials and geometries is highly desirable. But the tool development itself is not enough. The tool needs to be deployed to the user community so it can be made more reliable, flexible, and accurate.

Goal of this Article

The rapid progress in nanofabrication technologies has led to the development of novel devices and structures which could revolutionize many high technology industries. These devices demonstrate new capabilities and functionalities where the *quantum nature* of charge carriers plays an important role in determining the overall device properties and performance. For device sizes in the range of tens of nanometers, the *atomistic granularity* of constituent materials cannot be neglected: effects of atomistic strain, surface roughness, unintentional doping, the underlying crystal symmetries, or distortions of the crystal lattice can have a dramatic impact on the device operation and performance.

The goal of this paper is to describe the theoretical models and the essential algorithmic and computational components that have been used in the development and deployment of the Nanoelectronic Modeling tool NEMO 3-D on http://www.nanoHUB.org and to demonstrate successful applications of NEMO 3-D in the atomistic calculation of single-particle electronic states of different, realistically sized nanostructures, each consisting of multimillion atoms. We present some of the new capabilities that have been recently added to NEMO 3-D to make it one of the premier simulation tools for design and analysis of realistic nanoelectronic devices, and thus a valid tool for the computational nanotechnology community. These recent advances include algorithmic refinements, performance analysis to identify the best computational strategies, and memory saving measures. The effective scalability of NEMO 3-D code is demonstrated on the IBM BlueGene, the Cray XT3, an Intel Woodcrest cluster, and other Linux clusters. The largest electronic structure calculation, with *52 million atoms*, involved a Hamiltonian

matrix with over one *billion* complex degrees of freedom. The performance impact of storing the Hamiltonain versus recomputing the matrix, when needed, is explored. We describe the state-of-the-art algorithms that have been incorporated in the code, including very effective Lanczos, block Lanczos and Tracemin eigenvalue solvers, and present a comparison of the different solvers. While system sizes of tens of millions of atoms appear at first sight huge and wasteful, we demonstrate that some physical problems require such large scale analysis. We recently showed [44] that the analysis of valley splitting in strained Si quantum wells grown on strained SiGe required atomistic analysis of 10 million atoms to match experimental data. The insight that disorder in the SiGe buffer increases valley splitting in the Si quantum well would probably not be predictable in a continuum effective mass model. Similarly, the simulations of P impurities in silicon required multi-million atom simulations [82]. In the following, we describe NEMO 3-D capabilities in the simulation of different classes of nanodevices having carrier confinement in 3, 2, and 1 dimensions in the GaAs/InAs and SiGe materials systems.

Single and Stacked Quantum Dots (confinement in 3 dimensions) Quantum dots (QDs) are solid-state semiconducting nanostructures that provide confinement of charge carriers (electrons, holes, excitons) in all three spatial dimensions resulting in strongly localized wave functions, discrete energy eigenvalues and interesting physical and novel device properties [6,68,70,78,84,85]. Existing nanofabrication techniques tailor QDs in a variety of types, shapes and sizes. Within bottom-up approaches, QDs can be realized by colloidal synthesis at benchtop conditions. Quantum dots thus created have dimensions ranging from 2–10 nanometers, corresponding to 100–100,000 atoms.

Self-assembled quantum dots (SAQDs) grown in the coherent Stranski–Krastanov heteroepitaxial growth mode nucleate spontaneously within a lattice mismatched material system (for example, InAs grown on GaAs substrate) under the influence of strain in certain physical conditions during molecular beam epitaxy (MBE) and metalorganic vapor phase epitaxy (MOVPE) [3]. The strain produces coherently strained quantum-sized islands on top of a two-dimensional wetting-layer. The islands can be subsequently buried. Semiconducting QDs grown by self-assembly are of particular importance in quantum optics [28,67], since they can be used as detectors of infrared radiation, optical memories, and in laser applications.

The strongly peaked energy dependence of density of states and the strong overlap of spatially confined electron and hole wavefunctions provide ultra-low laser threshold current densities, high temperature stability of the threshold current, and high material and differential quantum gain/yield. Strong oscillator strength and non-linearity in the optical properties have also been observed [67]. Self-assembled quantum dots also have potential for applications in quantum cryptography as single photon sources and quantum computation [22,41]. In electronic applications QDs have been used to operate like a single-electron transistor and demonstrate a pronounced Coulomb blockade effect. Self-assembled QDs, with an average height of 1–5 nm, are typically of size (base length/diameter) 5–50 nm and consist of 5,000–2,000,000 atoms. Arrays of quantum-mechanically coupled (stacked) self-assembled quantum dots can be used as optically active regions in high-efficiency, room-temperature lasers. Typical QD stacks consist of 3–7 QDs with typical lateral extension of 10–50 nm and dot height of 1–3 nm. Such dots contain 5–50 million atoms in total, where atomistic details of interfaces are extremely important [95].

Impurities (confinement in 3 dimensions) Impurities have always played a vital role in semiconductors since the inception of the transistor. Till the end of last century, scientists and engineers had been interested in the macroscopic properties of an ensemble of dopants in a semiconductor. As technology enters the era of nanoscale electronics, devices which contain a few discrete dopants are becoming increasingly common. In recent years, there have been proposals of novel devices that operate on purely quantum mechanical principles using the quantum states of isolated or coupled donors/impurities [36,41,97]. The on-going extensive research effort on the Phosphorus (P) donor based quantum computer architecture of Kane [41] exemplifies an effort to harness the quantum nature of materials for the development of next generation electronics. As researchers strive to establish atomic scale quantum control over single impurities [19,87,91], precision modeling techniques are required to explore this new regime of device operations [25,29,65,82].

Although effective mass based approaches have been predominantly used in literature to study the physics of impurities, realistic device modeling using this technique have proved difficult in practice. Tight-binding methods [89] consider a more extensive Bloch structure for the host material, and can treat interfaces, external gates, strain, magnetic fields, and alloy disorder within a single framework. When applied to realistic nanodevices of several million atoms, this technique can prove very effective for device modeling [49]. We present a semi-empirical method for modeling impurities in Si that can be used for

a variety of applications such as *quantum computer* architecture, discretely doped FinFETs, and impurity scattering problems. Although we focus on P impurities in Si here, the method is sufficiently general to be used on other impurities and hosts.

Quantum Wires (confinement in 2 dimensions) For quite some time, nanowires have been considered a promising candidate for future building block in computers and information processing machines [17,50,64,98,106]. Nanowires are fabricated from different materials (metal, semiconductor, insulator and molecular) and assume different cross-sectional shapes, dimensions and diameters. Electrical conductivity of nanowires is greatly influenced by edge effects on the surface of the nanowire and is determined by quantum mechanical conductance. In the nanometer regime, the impact of surface roughness or alloy disorder on electronic bandstructure must be atomistically studied to further gauge the transport properties of nanowires.

Quantum Wells (confinement in 1 dimension) QW devices are already a de-facto standard technology in MOS devices and QW lasers. They continue to be examined carefully for ultra-scaled devices where interfacial details turn out to be critical. Composite channel materials with GaAs, InAs, InSb, GaSb, and Si are being considered [78,81], which effectively constitute QWs. Si QWs buffered/strained by SiGe are considered for Quantum Computing (QC) devices where valley-splitting (VS) is an important issue [27]. Si is desirable for QC due to its long spin-decoherence times, scaling potential and integrability within the present microelectronic manufacturing infrastructure. In strained Si, the 6-fold valley-degeneracy of Si is broken into lower 2-fold and raised 4-fold valley-degeneracies. The presence of 2-fold valley-degeneracy is a potential source of decoherence which leads to leakage of quantum information outside qubit Hilbert space. Therefore, it is of great interest to study the lifting of the remaining 2-fold valley degeneracy in strained Si due to sharp confinement potentials in recently proposed [27] SiGe/Si/SiGe quantum well (QW) heterostructures based quantum computing architectures.

Nanoscale Device Modeling and Simulation Challenges

The theoretical knowledge of the electronic structure of nanoscale semiconductor devices is the first and most essential step towards the interpretation and the understanding of the experimental data and reliable device design at the nanometer scale. The following is a list of the modeling and simulation challenges in the design and analysis of realistically sized engineered nanodevices.

Full Three-Dimensional Atomistic Representation

The lack of *spatial symmetry* in the overall geometry of the nanodevices usually requires explicit three-dimensional representation. For example, Stranski–Krastanov growth techniques tend to produce self-assembled InGaAs/GaAs quantum dots [68,77,84,85] with some rotational symmetry, e. g. disks, truncated cones, domes, or pyramids [6]. These structures are generally not perfect geometric objects, since they are subject to interface interdiffusion, and discretization on an atomic lattice. There is no such thing as a round disk on a crystal lattice! The underlying crystal symmetry imposes immediate restrictions on the realistic geometry and influences the quantum mechanics. Continuum methods such as effective mass [79] and $\boldsymbol{k} \cdot \boldsymbol{p}$ [35,92] typically ignore such crystal symmetry and atomistic resolution.

The required simulation domain sizes of ∼1M atoms prevent the usage of ab initio methods. Empirical methods which eliminate enough unnecessary details of core electrons, but are finely tuned to describe the atomistically dependent behavior of valence and conduction electrons, are needed. The current state-of-the-art leaves 2 choices: 1) pseudopotentials [20] and 2) Tight Binding [49]. Both methods have their advantages and disadvantages. Pseudopotentials use plane waves as a fundamental basis choice. Realistic nanostructures contain high frequency features such as alloy-disorder or hetero-interfaces. This means that the basis needs to be adjusted (by an expert) for every different device, which limit the potential impact for non-expert users. Numerical implementations of pseudopotential calculations typically require a Fourier transform between real and momentum space which demand full matrix manipulations and full transposes. This typically requires high bandwidth communication capability (i. e. extremely expensive) parallel machines, which limit the practical dissemination of the software to end users with limited compute resources. Tight-binding is a local basis representation, which naturally deals with finite device sizes, alloy-disorder and hetero-interfaces and it results in very sparse matrices. The requirements of storage and processor communication are therefore minimal compared to pseudopotentials and actual implementations perform extremely well on inexpensive clusters [49].

Tight-binding has the disadvantage that it is based on empirical fitting and some in the community continue

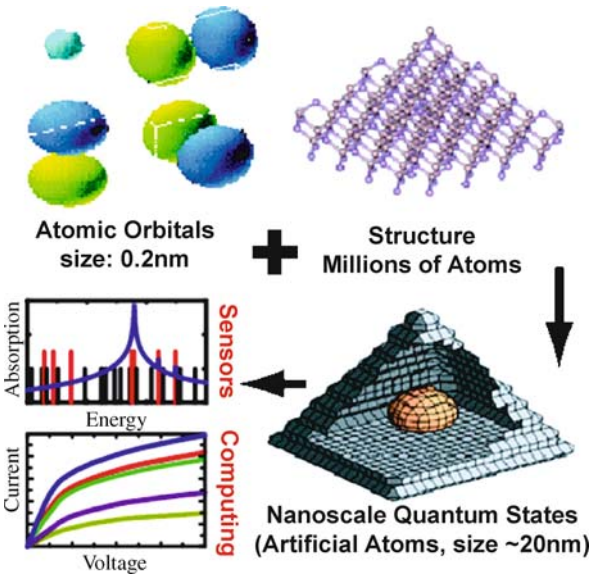

Multimillion Atom Simulations with Nemo3D, Figure 1
NEMO 3-D modeling agenda: map electronic properties of individual atoms into realistic structures containing millions of atoms, computation of nanoscale quantum dots that maps into real applications

to question the fundamental applicability of tight-binding. The NEMO team has spent a significant effort to expand and document the tight-binding capabilities with respect to handling of strain [12], electromagnetic fields [8], and Coulomb matrix elements [59] and fit them to well known and accepted bulk parameters [47,48,49]. With tight-binding the NEMO team was able early on to match experimentally verified, high-bias current-voltage curves of resonant tunneling [7,46] that could not get modeled by ether effective mass (due to the lack of physics) or pseudopotential methods (due to the lack of open boundary conditions). We continue to learn about the tight-binding method capabilities, and we are in the process of benchmarking it against more fundamental ab initio approaches and pseudopotential approaches. Our current Si/Ge parametrization is described in references [9,13]. Figure 1 depicts a range of phenomena that represent new challenges presented by new trends in nanoelectronics and lays out the NEMO 3-D modeling agenda.

Atomistic Strain

Strain that originates from the assembly of lattice-mismatched semiconductors strongly modifies the energy spectrum of the system. In the case of the InAs/GaAs quantum dots, this mismatch is around 7% and leads to a strong *long-range* strain field within the extended neighborhood (typically ~25 nm) of each quantum dot [2]. Si/Ge core/shell structured nanowires are another example of strain dominated atom arrangements [62]. Si quantum wells and SiGe quantum computing architectures rely on strain for state separation [27]. The strain can be atomistically inhomogeneous, involving not only biaxial components but also non-negligible shear components. Strain strongly influences the core and barrier material band structures, modifies the energy bandgaps, and lifts the heavy hole-light hole degeneracy at the zone center. In the nanoscale regime, the classical harmonic linear/continuum elasticity model for strain is inadequate, and device simulations must include the fundamental quantum character of charge carriers and the long-distance atomistic strain effects with proper boundary conditions on equal footing [58,101].

Piezoelectric Field

A variety of III-IV materials such as GaAs, InAs, GaN, are piezoelectric. Any spatial non-symmetric distortion in nanostructures made of these materials will create piezoelectric fields, which will modify the electrostatic potential landscape. Recent spectroscopic analyzes of self-assembled QDs demonstrate polarized transitions between confined hole and electron levels [6]. While the continuum models (effective mass or $\boldsymbol{k} \cdot \boldsymbol{p}$) can reliably predict aspects of the single-particle energy states, they fail to capture the observed non-degeneracy and optical polarization anisotropy of the excited energy states in the (001) plane. These methods fail because they use a confinement potential which is assumed to have only the *shape symmetry* of the nanostructure, and they ignore the underlying crystal symmetry. The experimentally measured symmetry is significantly lower than the assumed continuum symmetry because of (a) underlying crystalline symmetry, (b) atomistic strain relaxation and (c) piezoelectric field. For example, in the case of pyramid shaped quantum dots with square bases, continuum models treat the underlying material in $C_{4\nu}$ symmetry while the atomistic representation lowers the crystal symmetry to $C_{2\nu}$. The piezoelectric potential originating from the non-zero shear component of the strain field must be taken into account to properly model the associated symmetry breaking and the introduction of a global shift in the energy spectra of the system.

NEMO 3-D Simulation Package

Basic Features – Simulation Domains

NEMO 3-D [49,53,55,74,75] bridges the gap between the large size, classical semiconductor device models and the

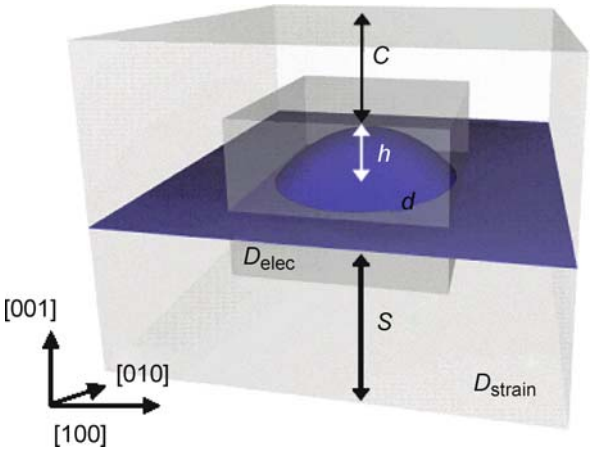

Multimillion Atom Simulations with Nemo3D, Figure 2
Simulated dome shaped InAs quantum dot buried in GaAs. Two simulation domains are shown, D_{elec}: central smaller domain for electronic structure calculation, and D_{strain}: outer larger domain for strain calculation. *In the figure*: *s* is the substrate height, *c* is the cap layer thickness, *h* is the dot height, *d* is the dot diameter

molecular level modeling. This package currently allows calculating single-particle electronic states and optical response of various semiconductor structures including bulk materials, quantum dots, quantum wires, quantum wells and nanocrystals. NEMO 3-D includes spin in its fundamental atomistic tight binding representation. Spin is therefore not added in as an afterthought into the theory, but spin-spin interactions are naturally included in the Hamiltonian. Effects of interaction with external electromagnetic fields are also included [8,32,49]. A schematic view of InAs quantum dot embedded in a GaAs barrier material the sample is presented in Fig. 2. The quantum dot is positioned on a 0.6 nm thick wetting layer (dark region). The simulation of strain is carried out in the large computational box D_{strain}, while the electronic structure computation is restricted to the smaller domain D_{elec}. Strain is long-ranged and penetrates around 25 nm into the dot substrate thus stressing the need for using large substrate thickness in the simulations. NEMO 3-D enables the computation of strain and electronic structure in an atomistic basis for over 64 and 52 million atoms, corresponding to volumes of $(110\,\text{nm})^3$ and $(101\,\text{nm})^3$, respectively. These volumes can be spread out arbitrarily over any closed geometry. For example, if a thin layer of 15 nm height is considered, the corresponding widths in the x–y plane correspond to 298 nm for strain calculations and 262 nm for electronic structure calculations. No other atomistic tool can currently handle such volumes needed for realistic device simulations. NEMO 3-D runs on serial and parallel platforms, local cluster computers as well as the NSF Teragrid.

Components and Models

The NEMO 3-D program flow consists of four main components.

Geometry Construction The first part is the geometry constructor, whose purpose is to represent the treated nanostructure in atomistic detail in the memory of the computer. Each atom is assigned three single-precision numbers representing its coordinates, stored is also its type (atomic number in short integer), information whether the atom is on the surface or in the interior of the sample (important later on in electronic calculations), what kind of computation it will take part of (strain only or strain and electronic), and what its nearest neighbor relation in a unit cell is. The arrays holding this structural information are initialized for all atoms on all CPUs, i. e., the complete information on the structure is available on each CPU. By default most of this information can be stored in short integer arrays or as single bit arrays, which does not require significant memory. This serial memory allocation of the atom positions, however, becomes significant for very large systems which must be treated in parallel.

Strain The materials making up the QD nanostructure may differ in their lattice constants; for the InAs/GaAs system this difference is of the order of 7%. This lattice mismatch leads to the appearance of strain: atoms throughout the sample are displaced from their bulk positions. Knowledge of equilibrium atomic positions is crucial for the subsequent calculation of QD's electronic properties, which makes the computation of strain a necessary step in realistic simulations of these nanostructures.

NEMO 3-D computes strain field using an atomistic valence force field (VFF) method [42] with the Keating Potential. In this approach, the total elastic energy of the sample is computed as a sum of bond-stretching and bond-bending contributions from each atom. The local strain energy at atom i is given by a phenomenological formula

$$E_i = \frac{3}{8}\sum_j \left[\frac{\alpha_{ij}}{2d_{ij}^2}\left(R_{ij}^2 - d_{ij}^2\right)^2 + \sum_{k>j}^{n} \frac{\sqrt{\beta_{ij}\beta_{ik}}}{d_{ij}d_{ik}}\left(\vec{R}_{ij}\cdot\vec{R}_{ik} - \vec{d}_{ij}\cdot\vec{d}_{ik}\right)^2 \right], \quad (1)$$

where the sum is carried out over the n nearest neighbors j of atom i, $\vec{d}_{ij}$ and $\vec{R}_{ij}$ are the bulk and actual (distorted)

distances between neighbor atoms, respectively, and α_{ij} and β_{ij} are empirical material-dependent elastic parameters. The equilibrium atomic positions are found by minimizing the total elastic energy of the system. Several other strain potentials [58,101] are also implemented in NEMO 3-D. While they modify some of the strain details they roughly have the same computational efficiency.

Electronic Structure The single-particle energies and wave functions are calculated using an empirical nearest-neighbor tight-binding model. The underlying idea of this approach is the selection of a basis consisting of atomic orbitals (such as s, p, d, and s^*) centered on each atom. These orbitals are further treated as a basis set for the Hamiltonian, which assumes the following form:

$$\hat{H} = \sum_i \varepsilon_i^{(\nu)} c_{i,\nu}^+ c_{i,\nu} + \sum_{i,\nu,\mu} t_i^{(\nu\mu)} c_{i,\nu}^+ c_{i,\mu} + \sum_{i,j,\nu,\mu} t_{\mathrm{ij}}^{(\nu\mu)} c_{i,\nu}^+ c_{j,\mu} \,, \quad (2)$$

where $c_{i,\nu}^+(c_{i,\nu})$ is the creation (annihilation) operator of an electron on the orbital ν localized on atom i. In the above equation, the first term describes the onsite orbital terms, found on the diagonal of the Hamiltonian matrix. The second term describes coupling between different orbitals localized on the same atom (only the spin-orbit coupling between p-orbitals), and the third term describes coupling between different orbitals on different atoms. The restriction in the summation of the last term is that the atoms i and j be nearest neighbors.

The characteristic parameters ε and t are treated as empirical fitting parameters for each constituent material and bond type. They are usually expressed in terms of energy constants of σ and π bonds between the atomic orbitals. For example, for a simple cubic lattice, the interaction between the s orbital localized on the atom i at origin and the orbital p_x localized on the atom j with coordinate $\vec{d}_{ij} = a\hat{x}$ with respect to the atom i would simply be expressed as $t_{ij}^{(s,p_x)} = V_{sp\sigma}$. Most of the systems under consideration, however, crystallize in the zinc-blende lattice, which means that the distance between the nearest neighbors is described by a 3-D vector $\vec{d}_{ij} = l\hat{x} + m\hat{y} + n\hat{z}$, with l, m, n being the directional cosines. These cosines rescale the interaction constants, so that the element describing the interaction of the orbitals s and p_x is $t_{ij}^{(s,p_x)} = lV_{sp\sigma}$. The parametrization of all bonds using analytical forms of directional cosines for various tight-binding models is given in [90]. NEMO 3-D provides the user with choices of the $sp^3d^5s^*$, sp^3s^*, and single s-orbital models with and without spin, in zincblende, wurzite, and simple cubic lattices.

Additional complications arise in strained structures, where the atomic positions deviate from the ideal (bulk) crystal lattice [40]. The presence of strain leads to distortions not only of bond directions, but also bond lengths. In this case, the discussed interaction constant $t_{ij}^{(s,p_x)} = l' V_{sp\sigma}\,(d/d_0)^{\eta(sp\sigma)}$, where the new directional cosine l' can be obtained analytically from the relaxed atom positions, but the bond-stretch exponent $\eta(sp\sigma)$ needs to be fitted to available data. The energy constants parametrizing the on-site interaction change as well due to bond renormalization [12,49].

The 20-band nearest-neighbor tight-binding model is thus parametrized by 34 energy constants and 33 strain parameters, which need to be established by fitting the computed electronic properties of materials to those measured experimentally. This is done by considering bulk semiconductor crystals (such as GaAs or InAs) under strain. The summation in the Hamiltonian for these systems is done over the primitive crystallographic unit cell only. The model makes it possible to compute the band structure of the semiconductor throughout the entire Brillouin zone. For the purpose of the fitting procedure, however, only the band energies and effective masses at high symmetry points and along the Δ line from Γ to C are targeted, and the tight-binding parameters are adjusted until a set of values closely reproducing these target values is found. Search for optimal parametrization is done using a genetic algorithm, described in detail in [32,49]. Once it is known for each material constituting the QD, a full atomistic calculation of the single-particle energy spectrum is carried out on samples composed of millions of atoms. No further material properties are adjusted for the nanostructure, once they are defined as basic bulk material properties.

Post Processing of Eigenstates From the single-particle eigenstates various physical properties can be calculated in NEMO 3-D such as optical matrix elements [18], Coulomb and exchange matrix elements [59], approximate single cell bandstructures from supercell bandstructure [10,11,17].

Algorithmic and Numerical Aspects

Parallel Implementation The complexity and generality of physical models in NEMO 3-D can place high demands on computational resources. For example, in the 20-band electronic calculation the discrete Hamiltonian matrix is of order 20 times the number of atoms. Thus, in a computation with 20 million atoms, the matrix is of or-

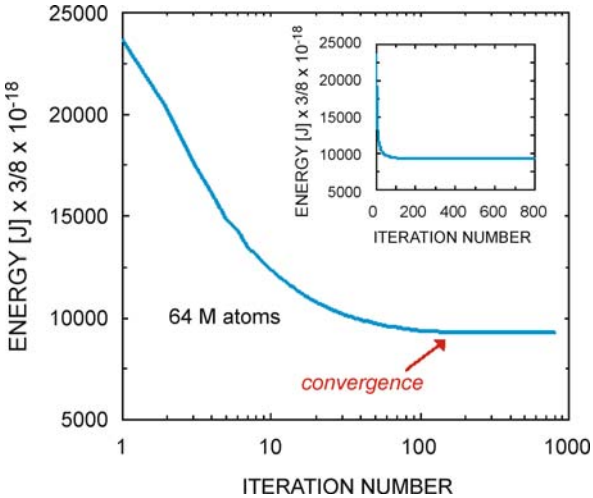

Multimillion Atom Simulations with Nemo3D, Figure 3
Elastic energy convergence profile in a typical simulation of an InAs/GaAs quantum dot with a total 64 million of atoms (*inset* – linear scale)

der 400 million. Computations of that size can be handled because of the parallelized design of the package. NEMO 3-D is implemented in ANSI C, C++ with MPI used for message-passing, which ensures its portability to all major high-performance computing platforms, and allows for an efficient use of distributed memory and parallel execution mechanisms.

Although the strain and electronic parts of the computation are algorithmically different, the key element in both is the sparse matrix-vector multiplication. This allows the use of the same memory distribution model in both phases. The computational domain is divided into slabs along one dimension. All atoms from the same slab are assigned to a single CPU, so if all nearest neighbors of an atom belong to its slab, no inter-CPU communication is necessary. The interatomic couplings are then fully contained in one of the diagonal blocks of the matrix. On the other hand, if an atom is positioned on the interface between slabs, it will couple to atoms belonging both to its own and the neighboring slab. This coupling is described by the off-diagonal blocks of the matrix. Its proper handling requires inter-CPU communication. However, due to the first-nearest-neighbor character of the strain and electronic models, the messages need to be passed only between pairs of CPUs corresponding to adjacent domains – even if the slabs are one atomic layer thick. Full duplex communication patterns are implemented such that all inter-processor communications can be performed in 2 steps [49].

Core Algorithms and Memory Requirements In the strain computation, the positions of the atoms are computed to minimize the total elastic strain energy. The total elastic energy in the VFF approach has only one, global minimum, and its functional form in atomic coordinates is quartic. The conjugate gradient minimization algorithm in this case is well-behaved and stable. Figure 3 shows the energy convergence behavior in a typical simulation of an InAs/GaAs quantum dot with a total of around 64 million of atoms. The total elastic energy operator is never stored in its matrix form, but the interatomic couplings are computed on the fly. Therefore the only data structures allocated in this phase are the vectors necessary for the conjugate gradient. The implementation used in NEMO 3-D requires six vectors, each of the total size of 3 × number of atoms (to store atomic coordinates, gradients, and intermediate data), however all those vectors are divided into slabs and distributed among CPUs as discussed above. The final atom position vectors are by default stored on all the CPU for some technical output details.

The electronic computation involves a very large eigenvector computation (matrices of order of hundreds of millions or even billion). The algorithms/solvers available in NEMO 3-D include the PARPACK library [66], a custom implementation of the Lanczos method, Block Lanczos method, the spectrum folding method [99] and the Tracemin method [86]. The research group is also exploring implementations of Lanczos with deflation method.

The Lanczos algorithm employed here is not restarted, and the Lanczos vectors are not reorthogonalized. Moreover, the spectrum of the matrix has a gap, which lies in the interior of the spectrum. Typically, a small set of eigenvalues is sought, immediately above and below the gap. The corresponding eigenstates are electron and hole wave functions, assuming effectively nonzero values only inside and in the immediate vicinity of the quantum dot. Also, in the absence of the external magnetic field the eigenvalues are repeated, which reflects the spin degeneracy of electronic states. The advantage of Lanczos algorithm is that it is fast, while the disadvantage is that it does not find the multiplicity and can potentially miss eigenvalues. Some comparisons have shown that the Lanczos method is faster by a factor of 40 for the NEMO 3-D matrix than PARPACK. Block Lanczos with block size p finds p degenerate eigenvalues relatively fast compared to PARPACK and Tracemin, however a potential instability exists as well. The Tracemin algorithm finds the correct spectrum of degenerate eigenvalues, but is slower than Lanczos. PARPACK has been found to be less reliable for this problem, taking more time than Tracemin and missing some

Multimillion Atom Simulations with Nemo3D, Table 2

Performance comparison of different eigenvalue solvers on 32 processors of Purdue University Linux cluster (Xeon x86-64 Dual Core 2.33GHz). Simulation was performed on an InAs QD structure with 268 800 atoms. Time (in hours), Relative time, Number of matrix-vector products (#MVP), Relative matrix-vector products, Memory (in GB) and number of correct eigenvalues and their multiplicity (#Eig(mul)) for Lanczos, Block Lanczos with block size 2 (BLanczos2), PARPACK, Tracemin with Quadratic mapping(QTracemin) and Tracemin with Chebyshev polynomial mapping(CTracemin)

Algorithm	Time (HRS.)	Relative Time	#MVP (×1000)	Relative MVP	Memory (GB)	#EIG.(MUL)
Lanczos	0.428	1.0	10.9	1.0	2.64	20(1)
BLanczos2	1.385	3.2	11.8	1.1	2.77	8(2)
PARPACK	18.04	42.2	59.3	5.4	2.64	8(2),4(1)
QTracemin	15.71	36.7	317.0	29.1	2.77	10(2)
CTracemin	13.70	32.1	528.8	48.5	2.64	10(2)

Multimillion Atom Simulations with Nemo3D, Table 3

List of spectrum between 1.0 ∼ 1.3 eV and the number of multiplicities obtained from different solvers. *Number* of searched eigenvalues was kept constant for these methods

Eigenvalues	Lanczos	BLanczos2	PARPACK	QTracemin	CTracemin
1.0361	1	–	–	2	2
1.0969	1	2	–	2	2
1.0976	1	2	1	2	2
1.1624	1	2	2	2	2
1.1645	1	2	2	2	2
1.1748	1	2	2	2	2
1.2304	1	2	2	2	2
1.2312	1	2	2	2	2
1.2445	1	2	2	2	2
1.2448	1	–	2	2	2
1.2975	1	–	2	–	–

of the eigenvalues and their multiplicity. Tables 2 and 3 give a comparison of Lanczos, Block Lanczos, PARPACK and Tracemin with the number of eigenvalues searched was kept constant. The majority of the memory allocated in the electronic calculation in Lanczos is taken up by the Hamiltonian matrix. This matrix is very large, but typically very sparse; this property is explicitly accounted for in the memory allocation scheme. All matrix entries are, in general, complex, and are stored in single precision. The code has an option to not store the Hamiltonian matrix, but to recompute it, each time it needs to be applied to a vector. In the Lanczos method, this is required once in each iteration. The PARPACK and Tracemin algorithms require the allocation of a significant number of vectors as a workspace, which is comparable to or larger than the Hamiltonian matrix. This additional memory need may require a matrix recompute for memory savings on memory-poor platforms like an IBM BlueGene.

Figure 4 shows the memory requirements for the dominant phase of the code (electronic structure calculations). It shows how the number of atoms that can be treated grows as a function of the number of CPUs, for a fixed amount of memory per CPU. The number of atoms can be intuitively characterized by the length of one side of a cube that would contain that many atoms. This length is shown in Fig. 4, on the vertical axis on the right side of each plot. This figure shows that the number of atoms that can be treated in NEMO 3-D continues to grow for larger CPU counts. The strain calculations have so far never been memory limited. NEMO 3-D is typically size limited in the electronic structure calculation.

Optimization in NEMO 3-D In running a scientific application that requires massive computation power, we have to consider various issues that may occur, mainly due to limited resource in a computer: Too small memory per core can limit the size of the problem and unnecessary loops in the code consumes additional time for calculation. It is crucial to design an application in a way to maximize floating operations per second and avoid inefficient loops. In NEMO-3D, several optimization ideas are implemented and those are introduced in the following sections.

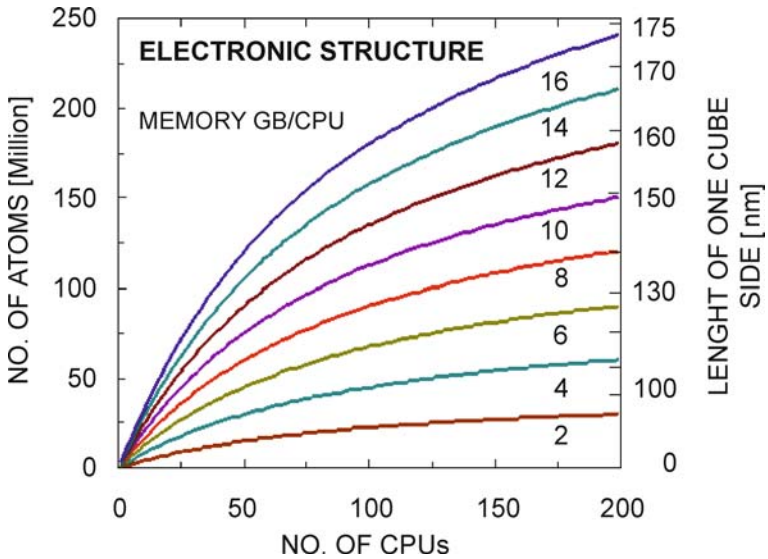

Multimillion Atom Simulations with Nemo3D, Figure 4
Number of atoms that can be treated, as a function of the number of CPUs for different amounts of memory per CPU for the electronic structure calculation. The vertical axis on the right side of each plot gives the equivalent length in nm of one side of the cube that would contain the given number of atoms

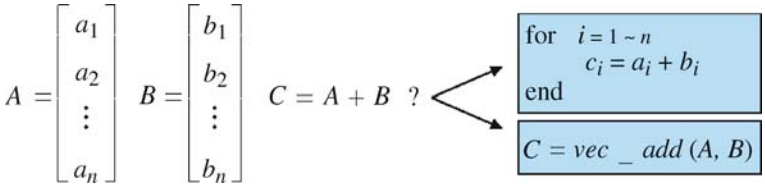

Multimillion Atom Simulations with Nemo3D, Figure 5
The conceptual diagram of vectorization. In vectorized CPU, it is capable of *n* simultaneous operations in single CPU cycle

Vectorization Vectorization is a hardware dependent optimization scheme that converts multiple single scalar operations to single vector operation. The concept is shown in Fig. 5. It is commonly used in graphic processors and supercomputers (e. g. Cray X1E machines) where massive computation load and fast processing is needed. Even recent processors in desktop computers, support similar parallel data processing scheme. The most common technique to support parallelism is Single Instruction, Multiple Data (SIMD) algorithm. It was Intel who first developed instruction sets known as Streaming SIMD Extensions, or SSE, to support in their Pentium III processors in 1999 [104]. Nowadays, AMD, Transmeta and Via also support SSE features and new enhancements are developed continuously (as of Oct. 2007, SSE5 is the latest version). A couple of single and double precision arithmetic can be carried out simultaneously resulting in fast computation. Therefore, it is possible to make use of SSE scheme in scientific applications with heavy complex number calculations. In NEMO-3D, complex multiplication and addition occurs frequently in matrix-matrix multiply routine. To this certain application, major improvement was achieved in real-complex multiplies. Figure 6 shows the speed improvement observed in NEMO-3D by replacing SSE instructions to real-complex multiplication.

Matrix–Matrix Multiplier and BLAS The Basic Linear Algebra Subprograms, or BLAS, are standardized interface for performing basic matrix-vector and matrix-matrix multiplication. The BLAS package is widely used in high-performance computing and it has been optimized to maximize the number of floating point operations for specific CPUs. For example, Intel develops its own BLAS package in the Math Kernel Library (MKL BLAS) highly optimized to their processors. Compared to native C code with double nested loops, benefits can be made from

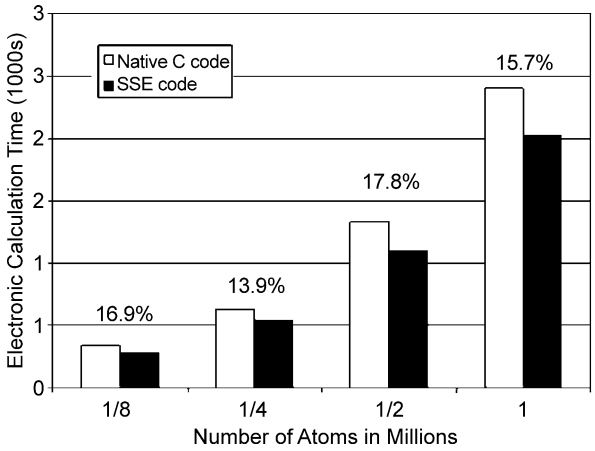

Multimillion Atom Simulations with Nemo3D, Figure 6
Comparison of electronic calculation time between SSE optimized code and native C code. Simulated on a single node of Xeon x86-64 Dual Core 2.33GHz CPU computers

BLAS, especially with matrix-matrix multiplication. From the experiment shown in Fig. 7, highly-optimized BLAS Matrix-matrix multiply instruction, or ZGEMM, is capable of utilizing the CPU to perform more floating point operations per second, reducing the total calculation time. Even for the block sizes $N = 10$, $N = 20$ corresponding to $sp^3d^5s^*$ bands significant improvement can be seen by performing block-wise operations. The data in Fig. 7 indicates an excellent incentive for the Block Lanczos and the Tracemin algorithms that perform multiple matrix-vector multiplies for the same matrix to be blocked. For example, at $N = 10$ a single vector multiply can be performed at about 1.5 GFlops while 8 multiplies can be performed at a rate of 3.6 GFlops. With the increase in relative performance for increased block size the required total CPU time increased sublinearly. Subsequent NEMO 3-D development for general 3-D spatial structures will utilize the ZGEMM multiply by arranging the data structures such that no copy is needed.

Explicit Construction of Hamiltonian in Recompute Mode
The recompute mode enables NEMO-3D to run on limited memory computers by eliminating storage of the Hamiltonian altogether and recomputing the matrix elements as they are needed. However, since the construction of the Hamiltonian consumes significant time, reducing the number of calculations in the construction of a matrix element enhances the performance. In cases where no external magnetic field is present, duplicate calculations due to the spin degeneracy can be avoided. Also, since the orbital interactions are known, unnecessary loops can be avoided, and non-zero elements may be explicitly evaluated. The doubly nested switch statements at the core of the orbital-orbital interaction loops have been replaced by customized expressions for the matrix elements for specific tight binding orbital arrangements such as sp^3s^* and $sp^3d^5s^*$. Simulation result indicates that the electronic calculation time is reduced up to 40% (Fig. 8). This customization increases computational performance but reduces the algorithmic generality.

Scaling Out of the two phases of NEMO 3-D, the strain calculation is algorithmically and computationally less challenging than the Lanczos diagonalization of the Hamiltonian matrix.

To investigate the performance of NEMO 3-D package, computation was performed in a single dome shaped InAs quantum dot nanostructure embedded in a GaAs barrier material as shown in Fig. 2. The HPC platform used in the performance studies are shown in Table 4. These include a Linux clusters at the Rosen Center for Advanced Computing (RCAC) at Purdue with Intel processors (dual core Woodcrest). The other five platforms are a BlueGene at the Rensselaer Polytechnic Institute (RPI), the Cray XT3 at the Pittsburgh Supercomputing Center (PSC), the Cray XT3/4 at ORNL, JS21 at Indiana University, and a Woodcrest machine at NCSA. Table 4 provides the relevant machine details. These platforms have proprietary interconnects, that are higher performance than Gigabit Ethernet (GigE) for the three Linux clusters at Purdue. In the following, the terms processors and cores are used interchangeably.

Figure 9 shows the performance of NEMO 3-D for each of the architectures. The wall clock times for 500 iterations of the Lanczos method for the electronic structure phase are shown as a function of the number of cores. The benchmark problem includes eight million atoms. Figure 9 shows that the PU/Woodcrest cluster is close to the performance of the Cray XT3 for lower core counts, while the XT3 performs better for higher core counts, due to its faster interconnect. The BlueGene's slower performance is consistent with its lower clock speed, while the scalability reflects its efficient interconnect.

Recomputing the Hamiltonian causes a performance reduction of about a factor of 4–6. Since the IBM BlueGene L is memory-poor, we can operate NEMO 3-D only in the Hamiltonian recomputed mode. Since the IBM BlueGene runs about a factor of 4× slower than the other HPC platforms one can see about a factor of 16 × better performance in Cray XT3/4 since it runs fast and has enough memory.

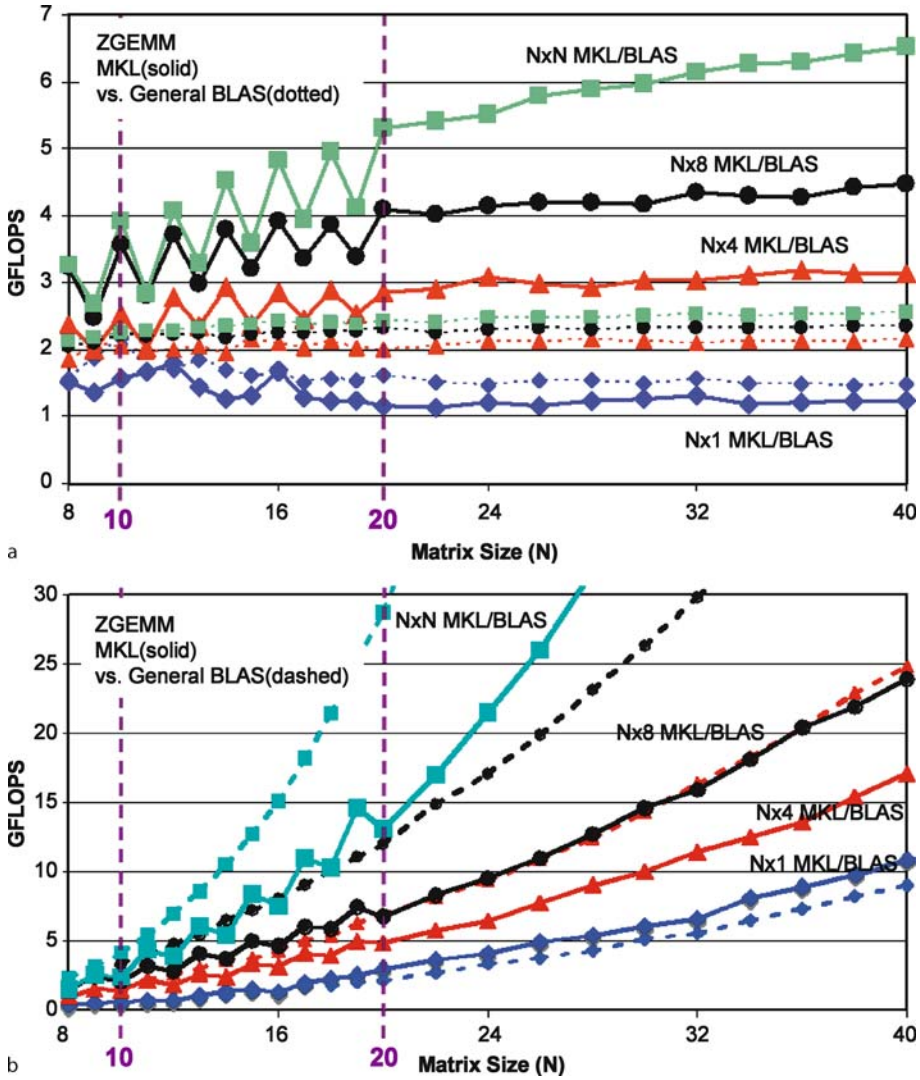

Multimillion Atom Simulations with Nemo3D, Figure 7

a Performance plots of ZGEMM (Y=AX) included in different BLAS libraries. GFLOPS (10^9 Floating Operations/second) measures of ZGEMM from MKL/BLAS (solid line) and general BLAS/LAPACK library (open markers) are plotted with varying size of $A(N\times N)$ and column size of $X(N\times M)$. Simulated on a single node of Xeon x86-64 Dual Core 2.33 GHz CPU computer. b Total compute time of data in a

Multimillion Atom Simulations with Nemo3D, Table 4

Specifications for the HPC platforms used in the performance comparisons

Platform	Type	CPU	# of Cores	Memory/ CORE	Interconnect	Top 500 June 2007	Location
ORNL/Jaguar	Cray XT3/4	Opteron x86-64 2.6GHz	23,016	2GB	Native	#2	ORNL
RPI/BGL	BlueGene/L	PowerPC 440 0.7 GHz	32,768	256MB	Native	#7	RPI
IUPU/Big Red	IBM JS21	PowerPC 970 2.5 GHz	3,072	2GB	Myrinet	#8	IUPU
PSC/XT3	Cray XT3	Opteron x86-64 2.6GHz	4,136	1GB	Native	#30	PSC
PU/Xeon D	Linux Cluster	Xeon x86-64 Dual Core 2.33GHz	672	2GB/4GB	Gigabit Ethernet	#46	RCAC Purdue

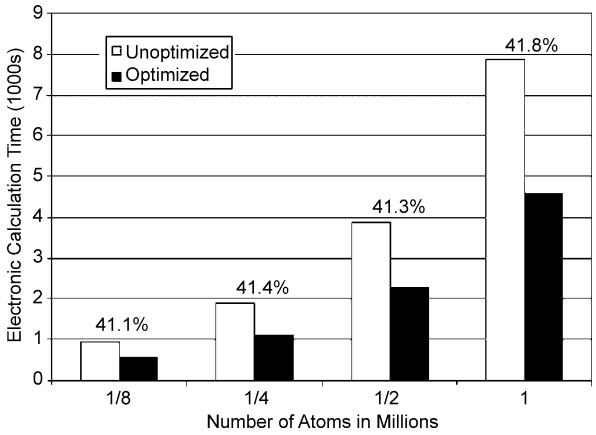

Multimillion Atom Simulations with Nemo3D, Figure 8
The electronic calculation time comparison of optimized/unoptimized Hamiltonian construction in recompute mode. Simulated on 4 nodes of Xeon x86-64 Dual Core 2.33GHz CPU computers

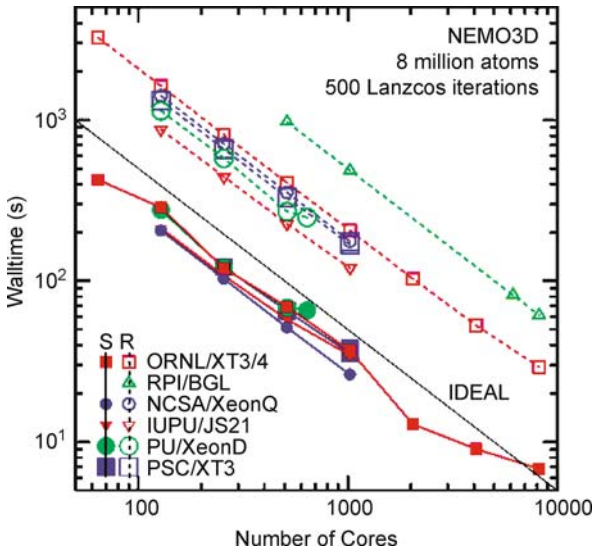

Multimillion Atom Simulations with Nemo3D, Figure 9
Strong scaling of a constant problem size (8 million atoms) on 6 different HPC platforms. *Solid/dashed lines* correspond to a stored / recomputed Hamiltonian matrix. The largest number of cores available were 8,192 on Cray XT3/4 and IBM BlueGene

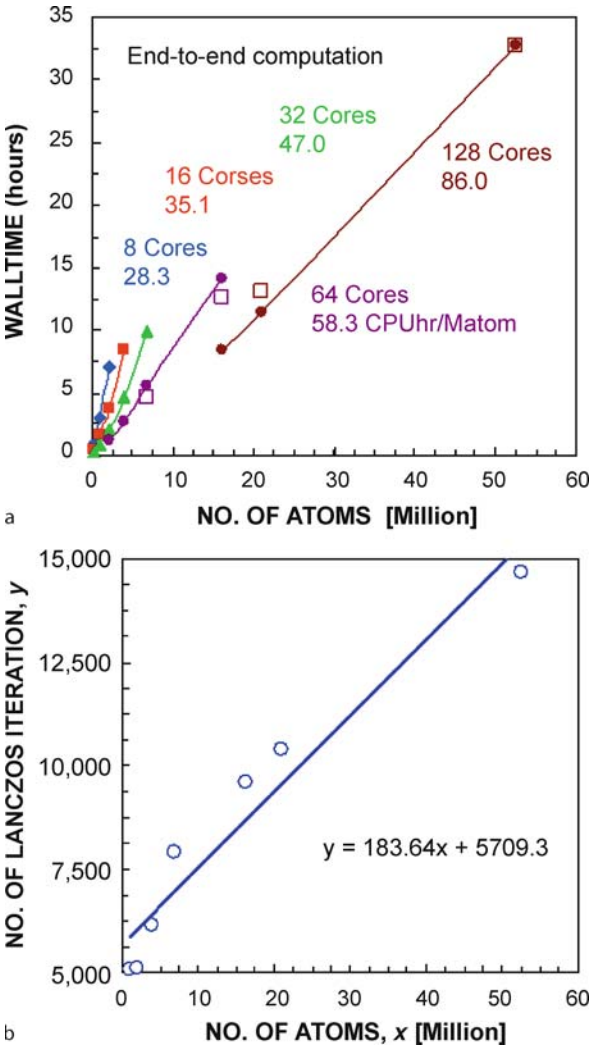

Multimillion Atom Simulations with Nemo3D, Figure 10
a Wall clock time vs. number of atoms for end-to-end computations of the electronic structure of a quantum dot, for various numbers of cores on the PU/Woodcrest cluster. Listed next to the number of cores are the CPU hours/Million of atoms needed in the simulation. b No. of Lanczos iteration vs. number of atoms for one choice of number of cores

In addition to the performance for the benchmark cases end-to-end runs on the PU/Woodcrest cluster are carried out next (Fig. 10). This involves iterating to convergence and computing the eigenstates in the desired range (4 conduction band and 4 valence band states). For each problem size, measured in millions of atoms, the end-to-end cases were run to completion, for one choice of number of cores.

The numerical experiment is designed to demonstrate NEMO 3-D's ability to extract targeted interior eigenvalues and vectors out of virtually identical systems of increasing size. A single dome shaped InAs quantum dot embedded in GaAs is considered. The GaAs buffer is increased in size to increase the dimension of the system while not affecting confined states in the QD. It is verified [4] that the eigenvectors retain the expected symmetry of the nanostructure.

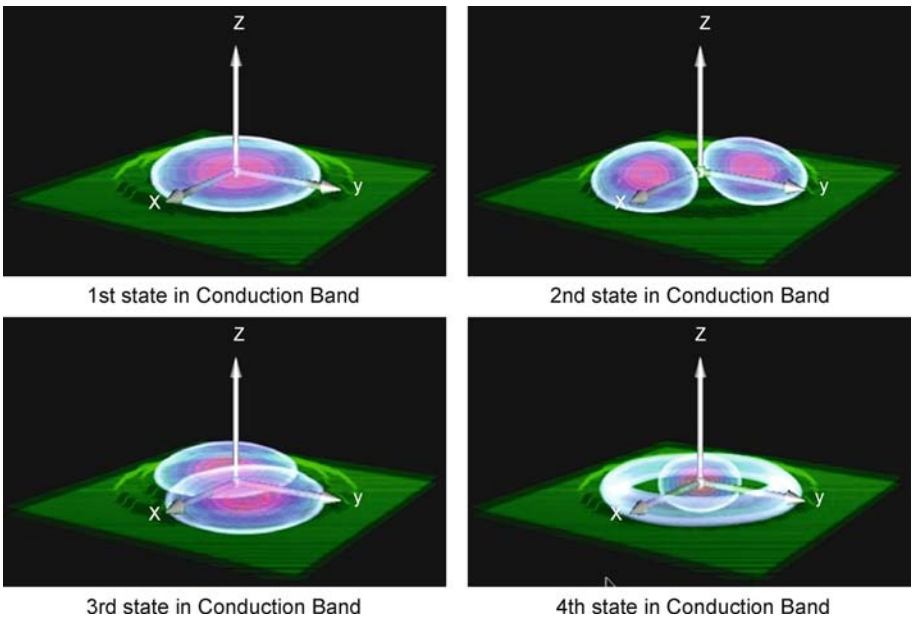

Multimillion Atom Simulations with Nemo3D, Figure 11
Wave function profiles of first 4 electron eigenstates in the conduction band. *Green color* shows active InAs region where confinement takes place

Visualization

The simulation data of NEMO 3-D contains multivariate wave functions and strain profiles of the device structure. For effective 3-D visualizations of these results, a hardware-accelerated direct volume rendering system [80] has been developed, which is combined with a graphical user interface based on *Rappture*. Rappture is a toolkit that supports and enables the rapid development of *graphical user interfaces* (GUIs) for applications, which is developed by Network for Computational Nanotechnology at Purdue University. Two approaches can be followed: (1) The legacy application is not modified at all and a *wrapper script* translates Rappture I/O to the legacy code. (2) Rappture is integrated into the source code to handle all I/O. The first step is to declare the parameters associated with one's tool by describing Rappture objects in the Extensible Markup Language (XML). Rappture reads the XML description for a tool and generates the GUI automatically. The second step is that the user interacts with the GUI, entering values, and eventually presses the Simulate button. At that point, Rappture substitutes the current value for each input parameter into the XML description, and launches the simulator with this XML description as the driver file. The third step shows that, using parser calls within the source code, the simulator gets access to these input values. Rappture has parser bindings for a variety of programming languages, including C/C++, Fortran, Python, and MATLAB. And finally, the simulator reads the inputs, computes the outputs, and sends the results through run file back to the GUI for the user to explore. The visualization system uses data set with *OPEN-DX* format that are directly generated from NEMO 3-D. *OPEN-DX* is a package of open source visualization software based on IBM's Visualization Data Explorer. Figure 11 shows the wave functions of electron on the first 4 eigenstates in conduction band of quantum dot which has 268,800 atoms in the electronic domain.

Release and Deployment of NEMO 3-D Package

NEMO 3-D was developed on Linux clusters at the Jet Propulsion Lab (JPL) and was released with an open source license in 2003. The originally released source is hosted at http://www.openchannelfoundation.org web site. As NEMO 3-D is undergoing further developments by the NCN we are planning future releases of the NEMO 3-D source through http://www.nanoHUB.org. NEMO 3-D has been ported to different high performance computing (HPC) platforms such as the NSF's TeraGrid (the Itanium2 Linux cluster at NCSA), Pittsburgh's Alpha cluster, Cary XT3, SGI Altix, IBM p690, and various Linux clusters at Purdue University and JPL.

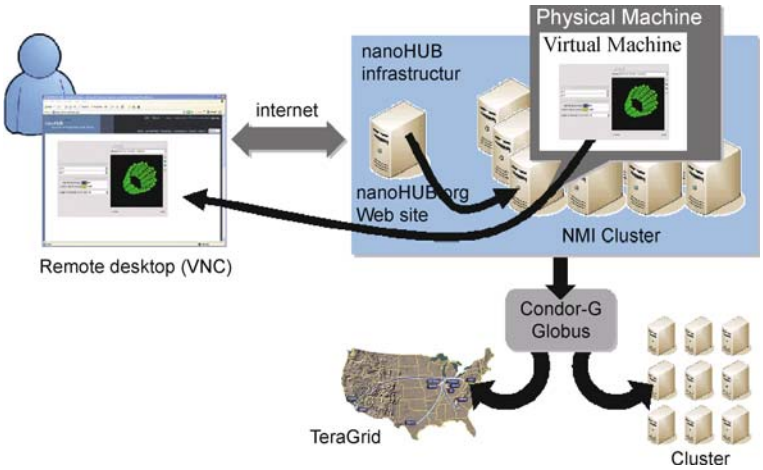

Multimillion Atom Simulations with Nemo3D, Figure 12
Deployment of the NCN nanotechnology tools on http://www.nanoHUB.org: Remote access to simulators and compute power

The NEMO 3-D project is now part of a wider initiative, the NSF Network for Computational Nanotechnology (NCN). The main goal of this initiative is to support the National Nanotechnology Initiative through research, simulation tools, and education and outreach. Deployment of these services to the science and engineering community is carried out via web-based services, accessible through the nanoHUB portal http://www.nanoHUB.org. The educational outreach of NCN is realized by enabling access to multimedia tutorials, which demonstrate state-of-the-art nanodevice modeling techniques, and by providing space for relevant debates and scientific events. The second purpose of NCN is to provide a comprehensive suite of nano simulation tools, which include electronic structure and transport simulators of molecular, biological, nanomechanical and nanoelectronic systems. Access to these tools is granted to users via the web browsers, without the necessity of any local installation by the remote users. The definition of specific sample layout and parameters is done using a dedicated Graphical User Interface (GUI) in the remote desktop (VNC) technology. The necessary computational resources are further assigned to the simulation dynamically by the web-enabled middleware, which automatically allocates the necessary amount of CPU time and memory. The end user, therefore, has access not only to the code, a user interface, and the computational resources necessary to run it but also to the scientific and engineering community responsible for its maintenance. The nanoHUB is currently considered one of the leaders in science gateways and cyber infrastructure.

The process of web-based deployment of these tools is depicted in Fig. 12. A user visits the http://www.nanoHUB.org site and finds a link to a tool. Clicking on that link will cause our middleware to create a virtual machine running on some available CPU. This virtual machine gives the user his/her own private file system. The middleware starts an application and exports its image over the Web to the user's browser. The application looks like an Applet running in the browser. The user can click and interact with the application in real time taking advantage of high-performance distributed computing power available on local clusters at Purdue University and on the NSF TeraGrid or the open science grid.

Recently, a prototype graphical user interface (GUI) based on the *Rappture* package (www.rappture.org) is incorporated within the NEMO 3-D package and a web-based online *interactive* version (Quantum Dot Lab) for educational purposes is freely available on www.nanohub.org, [38]. The currently deployed NEMO 3-D educational version is restricted to a single *s* orbital basis (single band effective mass) model and runs in seconds. Users can generate and freely rotate 3-D wavefunctions interactively powered by a remote visualization service. Quantum Dot Lab was deployed in November 2005 and has been a popular tool used by 1,541 users who ran 12,616 simulations up to August 2008. Monthly and annualized users and simulation numbers are shown in Fig. 13.

The complete NEMO 3-D package is available to selected members of the NCN community through the use of a nanoHUB workspace. A nanoHUB workspace presents a complete Linux workstation to the user within the con-

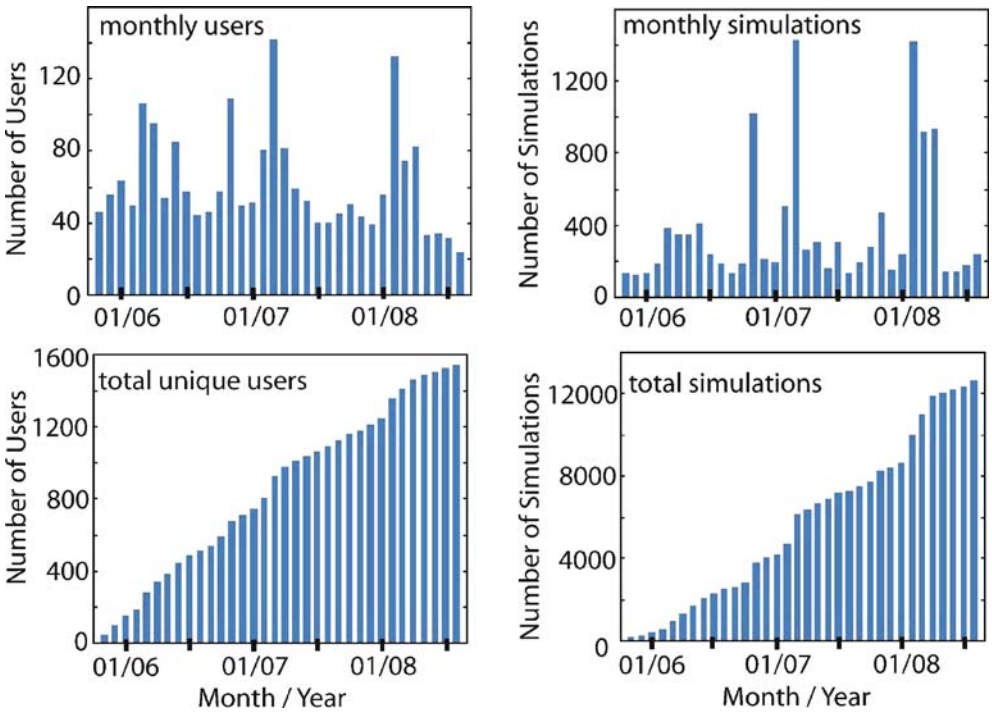

Multimillion Atom Simulations with Nemo3D, Figure 13

First row **Number of *monthly* users who have run at least one simulation and number of *monthly* simulation runs executed by nanoHUB users.** ***Second row*** **Number of *total* users who have run at least one simulation and *total* simulation runs executed by nanoHUB users**

text of a web browser. The workstation persists beyond the browser lifetime enabling to user to perform long duration simulations without requiring their constant attention. As shown in this paper the computational resources required to perform device scale simulations are considerable and beyond the reach of many researchers. With this requirement in mind NCN has joined forces with Teragrid [94] and the Open Science Grid [73] to seamlessly provide the necessary backend computational capacity to do computationally intensive computing. Computational resources necessary for large scale parallel computing are linked to nanoHUB through the Teragrid *Science Gateways* program. Access to a Teragrid allocation is provided for members of the NCN community. Development of a more comprehensive NEMO 3-D user interface continues. The more comprehensive interface will provide access to a broader audience and encourage the continued growth of the nanoHUB user base.

Simulation Results

Strain and Piezoelectricity in InAs/GaAs Single QDs

The dome shaped InAs QDs that are studied first in this work are embedded in a GaAs barrier material (schematic shown in Fig. 2) and have diameter and height of 11.3 nm and 5.65 nm respectively, and are positioned on a 0.6 nm-thick wetting layer [6,60]. The simulation of strain is carried out in the larger computational box (width D_{strain} and height H), while the electronic structure computation is usually restricted to the smaller domain (width D_{elec} and height H_{elec}). All the strain simulations in this category fix the atom positions on the bottom plane to the GaAs lattice constant, assume periodic boundary conditions in the lateral dimensions, and open boundary conditions on the top surface. The inner electronic box assumes closed boundary conditions with passivated dangling bonds [61]. The strain domain contains ~3 M atoms while the electronic structure domain contains ~0.3M atoms.

Impact of Strain Strain modifies the effective confinement volume in the device, distorts the atom bonds in length and angles, and hence modulates the local Bandstructure and the confined states. Figure 14 shows the diagonal (biaxial) components of strain distribution along the [001] direction in the quantum dot (cut through the center of the dot). There are two salient features in this plot: (a) The atomistic strain is long-ranged and penetrates deep into both the substrate and the cap layers, and (b) all

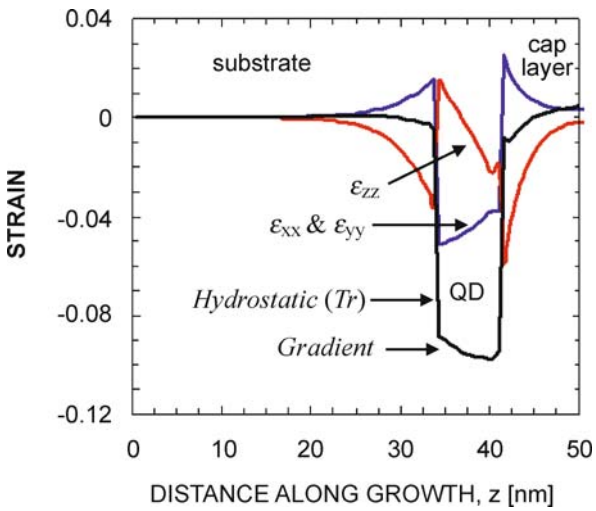

Multimillion Atom Simulations with Nemo3D, Figure 14
Atomistic *diagonal* strain profile along the [001], *z* direction. Dome shaped dot with Diameter, *d* = 11.3 nm and Height, *h* = 5.65 nm. Strain is seen to penetrate deep inside the substrate and the cap layer. Also, noticeable is the gradient in the trace of the hydrostatic strain curve (*Tr*) inside the dot region that results in optical polarization anisotropy and non-degeneracy in the electronic conduction band *P*. Atomistic strain thus lowers the symmetry of the dot

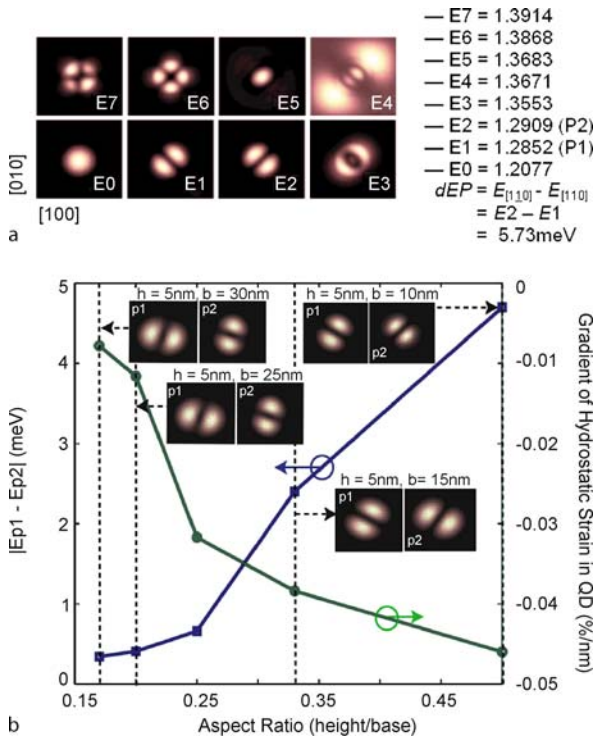

Multimillion Atom Simulations with Nemo3D, Figure 15
a Conduction band wavefunctions and spectra (eV) for first eight energy levels in the Dome shape quantum dot structure. Atomistic strain is included in the calculation. Note the optical anisotropy and non-degeneracy in the *P* energy level. The first state is oriented along [110] direction and the second state along [1$\underline{1}$0] direction. b gradient in the hydrostatic strain along the [00$\bar{1}$] direction through the center of the dot and the resulting non-degeneracy and optical anisotropy in the *P* level as a function of the dot aspect ratio

the components of biaxial stress have a non-zero slope inside the quantum dot region. The presence of the gradient in the trace of the hydrostatic strain introduces unequal stress in the zincblende lattice structure along the depth, breaks the equivalence of the [110] and [1$\underline{1}$0] directions, and finally breaks the degeneracy of the first excited electronic state (the so-called *P* level). Figure 15a shows the wavefunction distribution for the first 8 (eight) conduction band electronic states within the device region for the dot (in a 2-D projection). Note the optical anisotropy and non-degeneracy in the first excited (*P*) energy level. The first *P* state is oriented along the [110] direction and the second *P* state along the [1$\underline{1}$0] direction. The individual energy spectrum is also depicted in this figure which reveals the value of the *P* level splitting/non-degeneracy (defined as $E_{1\underline{1}0}$ – E_{110}) to be about 5.73 meV.

As explained in [6], the shape-symmetry of a quantum dot is lowered due mainly to three reasons, all originating from the fundamental atomistic nature of the underlying crystal: (1) The *interface* between the dot material (InAs) and the barrier material (GaAs), even with a common anion (As atom), is not a reflection plane and hence anisotropic with respect to the anion. The direct neighbors above the anion plane (In atoms) that align in the [$\underline{1}$10] direction are chemically different from the neighbors under the anion plane (Ga atoms) that align in the [110] direction. This creates a short-range interfacial potential. It is important to note that these atomistic interfacial potentials originating from different facets do not necessarily compensate each other in dots where the base is larger than the top (for example, pyramid, lens, truncated pyramid). (2) *Atomistic strain and relaxations* (originating from the atomic size difference between Ga and In atoms) results in a propagation of the interfacial potential further into the dot material and thus amplifies the magnitude of the asymmetry. This component is not captured if the relaxation is performed using classic harmonic continuum-elasticity approach. Noticeable is the fact that, symmetry breaking due to atomistic relaxations can even be observed in dots where the base is equal to the top (for example, box, disk); however, the effect is magnified in dots of typical shape, where the base is larger than the top (for example,

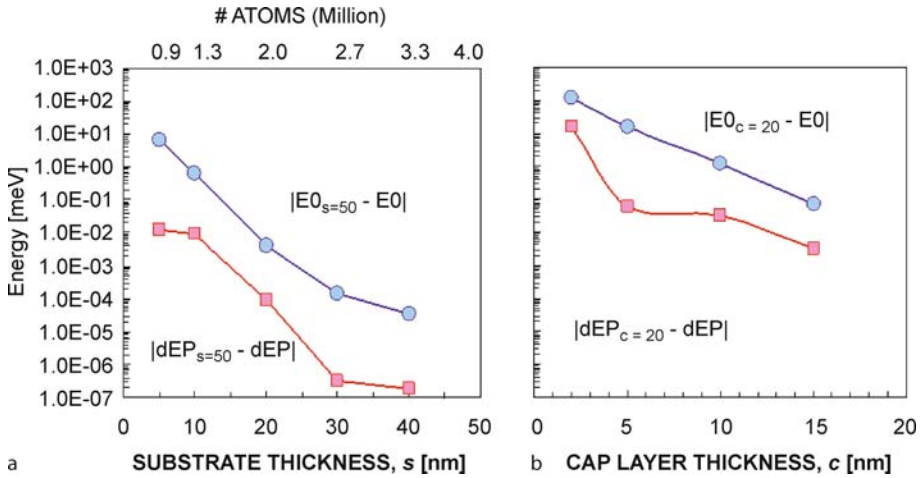

Multimillion Atom Simulations with Nemo3D, Figure 16
a Substrate layer thickness dependence of the conduction band minimum and the *P* level splitting. Other structural parameters remain constant (*h* = 5.65 nm, *d* = 11.3 nm, *c* = 10 nm, and *D* = 31.3 nm). b The impact of cap layer thickness (with substrate, *s* = 30 nm and other structural parameters remaining the same). Lanczos convergence tolerance = 1×10^{-7}

pyramid, lens, truncated pyramid) due to the presence of a gradient in the magnitude of the strain tensor between top and bottom as already explained in Fig. 15a. In order to further characterize this effect, we have simulated dome-shaped dots with varying base diameters (from 10 to 30 nm) keeping the dot height constant (at 5 nm). Figure 15b shows the gradient in the hydrostatic strain and the resulting nondegeneracy in the *P* level as a function of the dot aspect ratio (height/base). Also, shown in the insets are the wavefunctions corresponding to the split *P* levels in each of these dots. Note that the non-degeneracy and the optical anisotropy diminish as the dot aspect ratio decreases (approaching a disk shape). (3) Finally, a long-ranged *piezoelectric field* develops in these dots in response to the strain-induced displacement field, which is fundamentally anisotropic. We will discuss this effect in detail in a subsequent section.

Need for a Deep Substrate and a Realistic Cap Layer The strength of the NEMO 3-D package lies particularly in its capability of simulating device structures with realistic boundary conditions. Our simulation results based on NEMO 3-D show a significant dependence of the dot states and magnitude of level-splitting on the substrate layer thickness, *s* (underneath the dot) and the cap layer thickness, *c* (above the dot). The strain in the QD system therefore penetrates deeply into the substrate and cannot be neglected. Figure 16 shows such observed dependency where *E*0 is the ground state energy and *dEP* is the magnitude of the level splitting in the *P* electronic states due to the inclusion of atomistic strain and relaxation. The changes in both these quantities are calculated with respect to the largest *s* (50 nm) and *c* (20 nm) respectively in Figs. 16a,b. The wavefunction orientation was found to remain unchanged irrespective of the substrate depth and cap layer thickness. Figure 16a shows that it is indeed important to include enough of a substrate to capture the long-range strain, while Fig. 16b indicates opportunities to tune the eigen energy spectrum with different capping layer thicknesses.

Figure 17 reveals the reason of a strong dependency of the electronic ground state and the magnitude of nondegeneracy in *P* level on the cap layer thickness. Here the hydrostatic strain profiles for two different cap layer thicknesses (2 nm and 10 nm) are plotted. The *P* level splitting in a device with 10 nm cap layer is found to be 5.73 meV and that for a 2 nm cap layer was 20.58 meV. The reason of the reduction in the splitting in the 10 nm cap layer device can be attributed mainly to the change in the gradient of hydrostatic strain inside the device region as depicted in Fig. 17.

Impact of Piezoelectric Fields The presence of nonzero off-diagonal strain tensor elements leads to the generation of a piezoelectric field in the quantum dot structure, which is incorporated in the simulations as an external potential by solving the Poisson equation on the zincblende lattice. Figures 18a,b show the atomistic off-diagonal strain profiles in dome shaped quantum dots with heights, *h* of 2.8 nm and 5.65 nm respectively. The off-di-

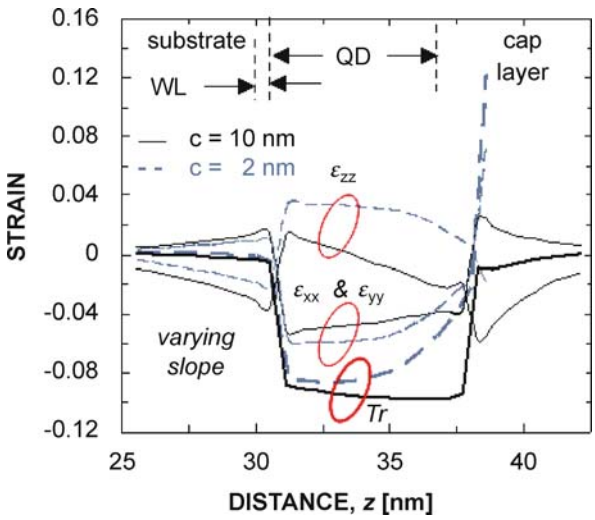

Multimillion Atom Simulations with Nemo3D, Figure 17
The impact of cap layer thickness (with deep substrate, s = 30 nm, and h = 5.65 nm, d = 11.3 nm). Shown is the significant variation of gradient/slope in the strain profile within the quantum dot region. This results in a different splitting in the conduction band P energy level for the two different thicknesses of the cap layer

agonal strain tensors are higher in the larger diameter dot. The off-diagonal strain tensors are found to be larger in the dome shaped dot. The off-diagonal strain tensors are used to calculate the first-order polarization in the underlying crystal (see [6] for the governing equations) which gives rise to a piezoelectric charge distribution throughout the device region and then used to calculate the potential by solving the Poisson equation. The relevant parameters for the piezoelectric calculation are taken from [6]. Experimentally measured polarization constants of GaAs and InAs materials (on unstrained bulk) values of -0.16 C/m^2 and -0.045 C/m^2 are used. The second order piezoelectric effect [5] is neglected here because of unavailability of reliable relevant polarization constants for an InAs/GaAs quantum dot structures.

The calculated piezoelectric charge and potential surface plots in the XY and XZ planes are shown in Figs. 19 and 20 respectively revealing a pronounced polarization effect induced in the structure. It is found that piezoelectric field alone favors the [1$\underline{1}$0] orientation of the P level. Also shown in Fig. 21 is the asymmetry in potential profile due to atomistic strain and inequivalence in the piezoelectric potential along [110] and [1$\underline{1}$0] directions at a certain height z = 1 nm from the base of the dot.

Study of Varying Sized Dots The impact of atomistic strain and piezoelectric field on the ground state energy and magnitude of the P level energy splitting in dome shaped quantum dots with varying diameter d and dot height h is shown in Figs. 22 and 23 respectively. The ground state energy for the strained system (without piezoelectricity), $E0$, decreases with an increase in both d and h because of an increase in the effective confinement volume. Figures 22a and 23a also show the change (absolute and relative to strain only) in the ground state energy due to the inclusion of piezoelectric potential in the strained system. The percentage change in the ground state energy is found to be monotonous in nature with

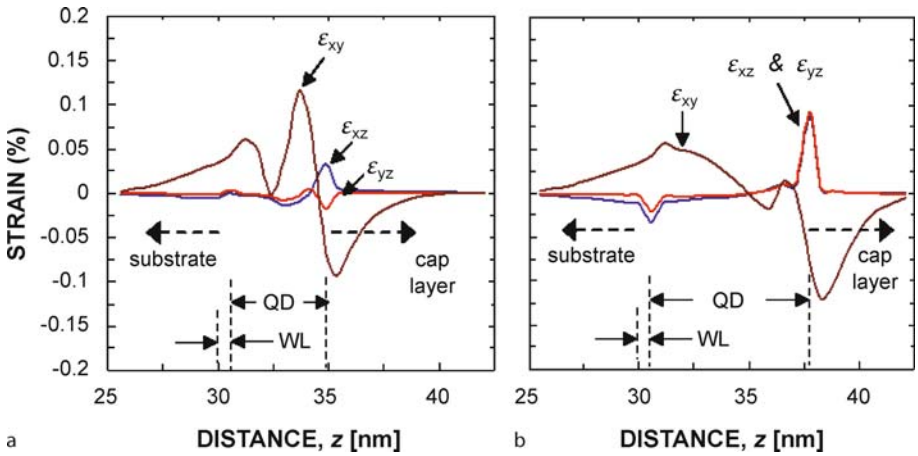

Multimillion Atom Simulations with Nemo3D, Figure 18
Atomistic *off-diagonal* strain profile along the z (vertical) direction which in effect induces polarization in the quantum dot structure. a Diameter, d = 11.3 nm and Height, h = 2.8 nm and b Diameter, d = 11.3 nm and Height, h = 5.65 nm. Note the increase in *off-diagonal* stain in b

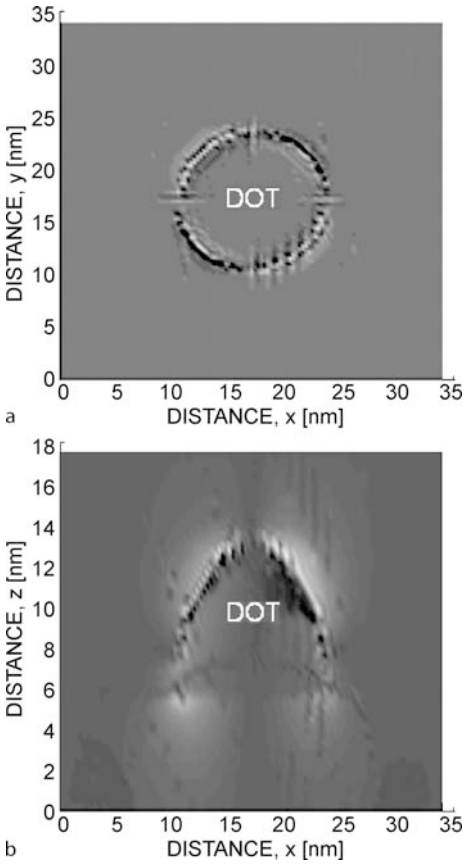

Multimillion Atom Simulations with Nemo3D, Figure 19
Charge surface plot of a dome shape quantum dot a in the *XY* plane at $z = 1$ nm from the base of the dot, and b in the *XZ* plane at $y = D_{strain}/2$. Charge is induced mainly in the vicinity of the boundary of the quantum dot. ($d = 11.3$ nm and $h = 5.65$ nm)

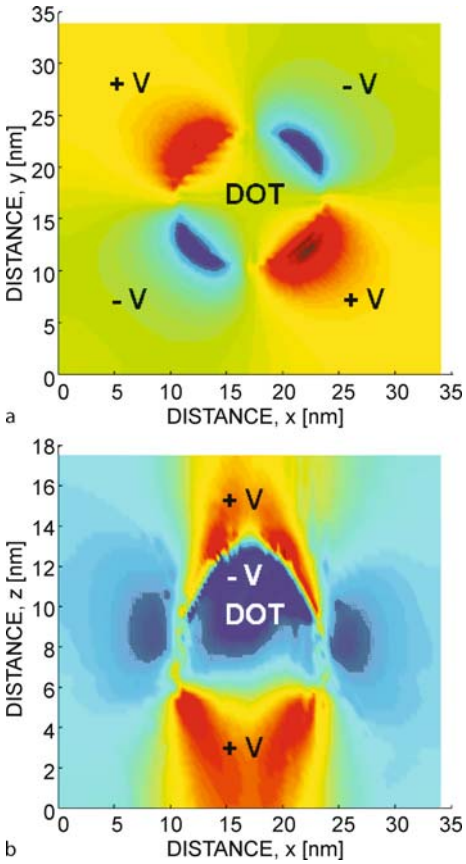

Multimillion Atom Simulations with Nemo3D, Figure 20
Piezoelectric potential surface plot of a dome shape quantum dot a in the *XY* plane at $z = 1$ nm from the base of the dot, and b in the *XZ* plane at $y = D_{strain}/2$. c Potential along [110] and [1$\underline{1}$0] directions at z = 1 nm from the base of the dot. Note the induced polarization in the potential profile and the unequal values of potential along the [110] and [1$\underline{1}$0] directions ($d = 11.3$ nm and $h = 5.65$ nm)

an increase in dot diameter while the height dependency shows saturation beyond a certain value. Figures 22b and 23b show the change of three quantities related to the first excited *P* level namely split due to strain only (circle), split due to strain combined with piezoelectricity (square) and the contribution of the piezoelectric field only (triangle), as a function of diameter *d* and dot height *h*. The piezoelectric potential introduces a global shift in the energy spectrum, and is found to be strong enough to flip the optical polarization in certain sized quantum dots. In those cases the piezoelectric contribution (triangle) dominates over that resulting from the inclusion of atomistic strain alone in the simulations (circle) as can be seen in dots (see Fig. 22b; similar trend has also been found in [6]) with diameters larger than 7 nm and (see Fig. 23b) height more than 3 nm. Figure 24 explains the reason behind this observation. Here the piezoelectric potential profiles in dots

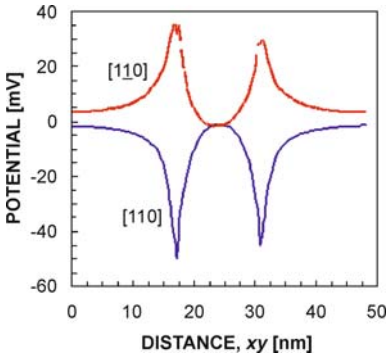

Multimillion Atom Simulations with Nemo3D, Figure 21
Potential along [110] and [1$\underline{1}$0] directions at $z = 1$ nm from the base of the dot. Note the induced polarization in the potential profile and the unequal values of potential along the [110] and [1$\underline{1}$0] directions ($d = 11.3$ nm and $h = 5.65$ nm)

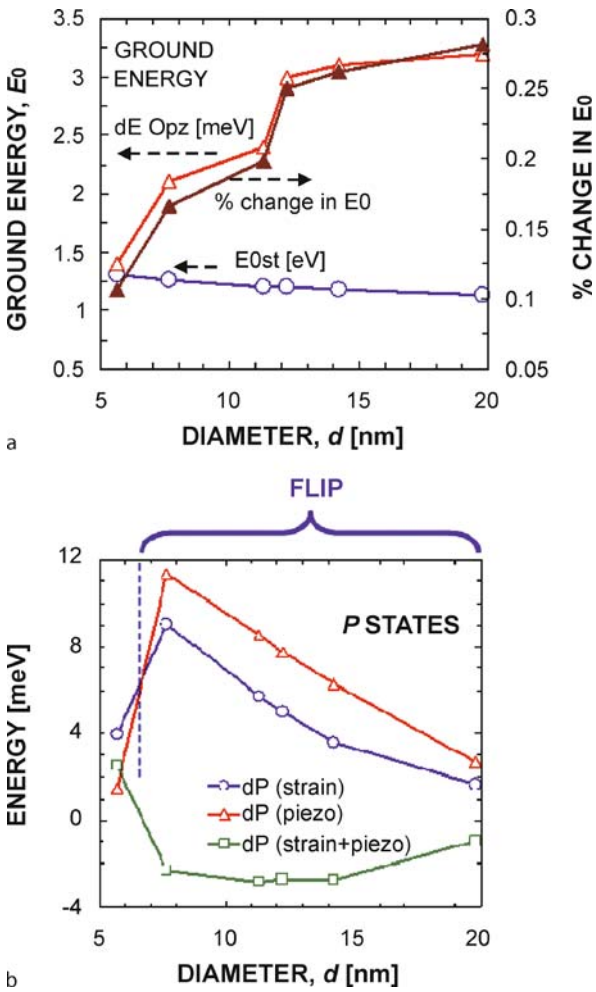

Multimillion Atom Simulations with Nemo3D, Figure 22
Study of electronic structure with the variation of *dot diameter*, *d* of the dome shaped quantum dot. a Conduction band minimum/ground state in a strained system (*circle*) and change in the conduction band minimum due to induced piezoelectricity (*triangle*). b Split in the *P* level due to strain only (*circle*), split in the *P* level due to strain and piezoelectricity (*square*), and impact of piezoelectric potential *alone* (*triangle*) in the system (dot height, *h* = 5.65 nm)

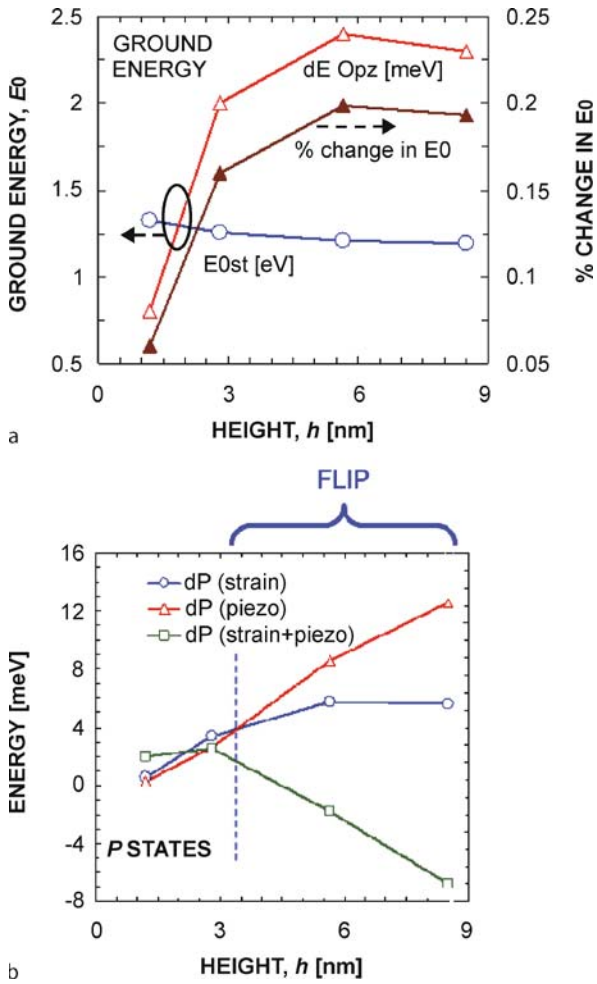

Multimillion Atom Simulations with Nemo3D, Figure 23
Study of electronic structure with the variation of *dot height*, *h* of the dome shaped quantum dot. a Conduction band minimum/ground state in a strained system (*circle*) and change in the conduction band minimum due to induced piezoelectricity (*triangle*). b Split in the *P* level due to strain only (*circle*), split in the *P* level due to strain and piezoelectricity (*square*), and impact of piezoelectric potential *alone* (*triangle*) in the system (dot diameter, *d* = 11.3 nm)

with different height h are plotted along the z direction through the dot center. Note the increase in piezoelectric potential with dot height. The stronger piezoelectric potential induced in the larger dot results in the orientational flip in the P level electronic states.

Piezoelectricity Induced Polarization Flip Figure 25 shows the conduction band wavefunctions for the ground and first three excited energy states in the quantum dot structure with diameter of 11.3 nm and height, h of 5.65 nm. In Fig. 25a strain and piezoelectricity are *not* included in the calculation. The weak anisotropy in the P level is due to the atomistic interface and material discontinuity. Material discontinuity mildly favors the [110] direction in the dot. In Fig. 25b atomistic strain and relaxation is included resulting in a 5.73 meV split in the P energy levels. Strain favors the [110] direction. In Fig. 25c piezoelectricity is included on top of strain inducing a split of −2.84 meV in the P energy level. The first P state is oriented along [1$\underline{1}$0] direction and the second state along

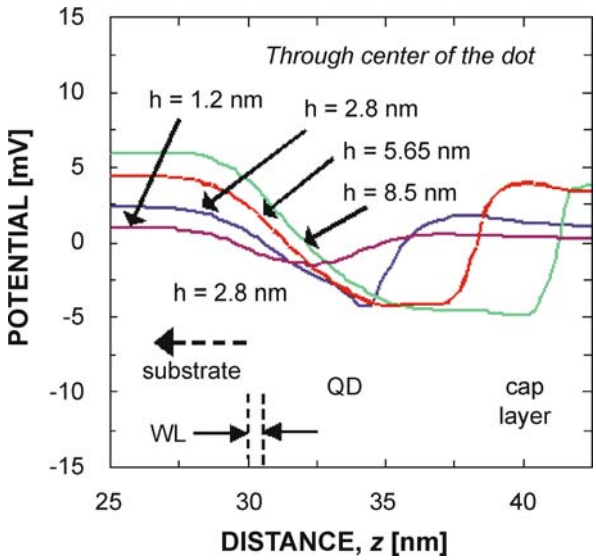

Multimillion Atom Simulations with Nemo3D, Figure 24
Piezoelectric potential in dome shaped quantum dots with h = 2.8 nm and h =5.65 nm along the z direction through the center of the dots. Noticeable is the stronger polarization in the larger dot which results in a *flip* in the P level electronic states

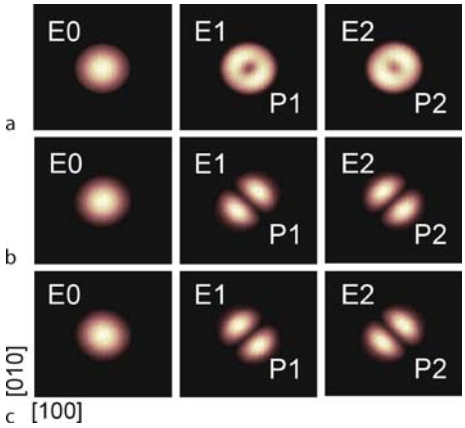

Multimillion Atom Simulations with Nemo3D, Figure 25
Conduction band wavefunctions for first three energy levels in the quantum dot structure with diameter, d = 11.3 nm and height, h = 5.65 nm a without strain and piezoelectricity, $E_{[1\bar{1}0]} - E_{[110]}$ = 1.69 meV b with atomistic strain, $E_{[1\bar{1}0]} - E_{[110]}$ = 5.73 meV and c with strain and piezoelectricity, $E_{[1\bar{1}0]} - E_{[110]}$ = −2.84 meV. Piezoelectricity *flips* the wavefunctions. An end-to-end computation involved about 4M atoms and needed CPU time of about 8 hours with 16 processors

[110] direction and piezoelectricity alone induces a potential of 8.57 meV. Piezoelectricity thereby has not only introduced a global shift in the energy spectrum but also *flipped* the orientation of the P states [6] in this case.

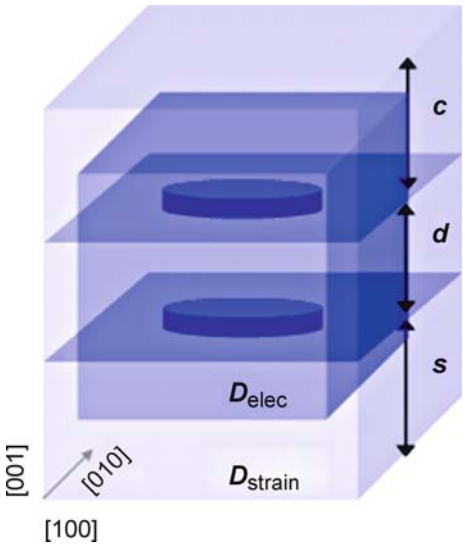

Multimillion Atom Simulations with Nemo3D, Figure 26
Simulated InAs/GaAs double quantum dots with disk/cylindrical shape. The dots are of equal size with radius r of 7 nm and height h of 1.5 nm. The separation d is varied from 0.5 nm to 8 nm. Two simulation domains have been shown. The strain domain for 8 nm spacing between the dots contained about 6 million atoms

Stacked Quantum Dot System

Self-assembled quantum dots can be grown as stacks where the QD distance can be controlled with atomic layer control. This distance determines the interaction of the artificial atomic states to form artificial molecules. The design of QD stacks becomes complicated since the structures are subject to inhomogeneous, long-range strain and growth imperfections such as non-identical dots and inter-diffused interfaces. Quantum dot stacks consisting of two QD layers are simulated next (see Fig. 26). The InAs quantum dots are disk shaped with diameter 7 nm and height 1.5 nm positioned on a 0.6 nm thick wetting layer. The substrate thickness under the first wetting layer is kept constant at 30 nm and the cap layer on top of the topmost dot is kept at 20 nm for all the simulations. The strain simulation domain (D_{strain}) contains 8–10 M atoms and the electronic structure domain (D_{elec}) contains 0.5–1.1M atoms.

Figure 27 shows both the biaxial and hydrostatic strain profiles along the z direction. As in the single dot, we see a gradient in strain profile within the dot regions which results in strain-induced asymmetry. The hydrostatic component which is responsible for conduction band well is negative within the dot and approximately zero outside the dot and the regions in-between the dots. The biaxial component of strain which have more effect on hole states

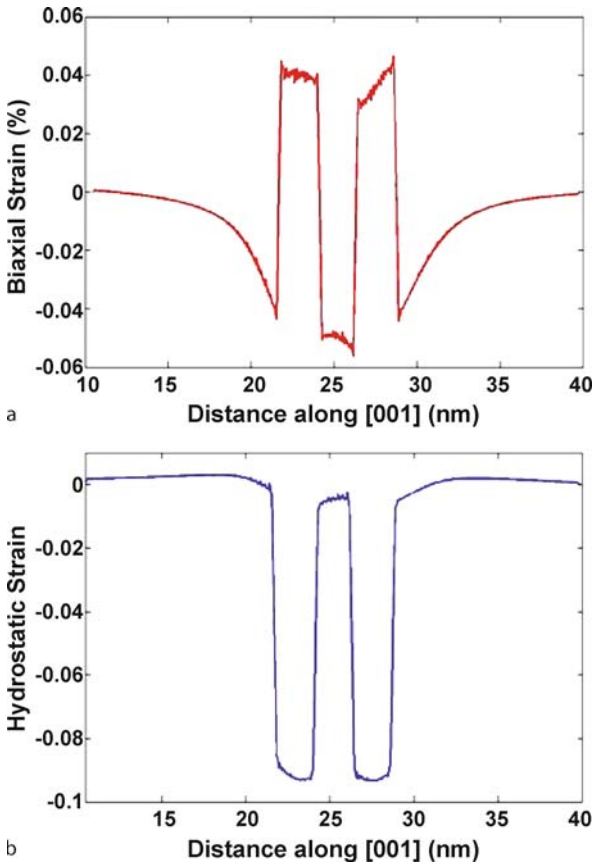

Multimillion Atom Simulations with Nemo3D, Figure 27
Atomistic a *biaxial* $\{\varepsilon_{zz} - (\varepsilon_{xx} + \varepsilon_{yy})/2\}$ and b *hydrostatic* $\{\varepsilon_{xx} + yy + \varepsilon_{zz}\}$ strain profile along the growth [001], *z* direction. Strain is seen to penetrate deep inside the substrate and the cap layer. Also, noticeable is the gradient in the trace of the hydrostatic strain curve (*Tr*) inside the dot region that results in optical polarization anisotropy and non-degeneracy in the electronic conduction band *P*. Atomistic strain thus lowers the symmetry of the dot

is positive within dots and negative in-between the dots. The magnitude for both is approximately equal. Figure 28 shows the band edge diagrams as a function of dot separations along the center of the dots in the growth direction [001]. Strain enhances the coupling between the dots. Hydrostatic component of strain makes the conduction band well shallower. Strain effects are more prominent on hole states where biaxial component of strain splits the light hole and heavy hole bands. Within the dot, heavy hole lies above the light hole edge, implying significant band mixing in the confining states. For very small separation like 0.5 nm, the well within dots is even shallower than the well in-between the dots. Figure 29 and Fig. 30 show the electron and hole state energies respectively as a function of inter-dot separation. In a system without inhomogeneous strain one would expect the identical dots to have degenerate eigenstate energies for large dot separations. Strain breaks the degeneracy even for large separations. As the dot separation is narrowed the dots interact with each other mechanically through the strain field as well as quantum mechanically through wavefunction overlaps. Wave function plots in *XZ* plane have been shown in Fig. 29 for various dot separations. Noticeably, *E*2 for 4 nm separation is a *p* like state while it is *s* like state in 6 nm separation. So there is a crossover between *p* to *s* for *E*2 as we increase separation between the dots. Also, *E*1 for 2 nm separation is confined more in the lower dot than the upper dot. This is caused by strain coupling which promotes confinement of the ground states in the lower dots in coupled quantum dot systems [53]. The electronic states and wavefunctions in a coupled QD system are thus determined through a complicated interplay of strain, QD size, and wavefunction overlap. Only a detailed simulation can reveal that interplay.

Phosphorus (P) Impurity in Silicon

Physics of P Impurity In a substitutional P impurity in Si, the 4 electrons from the outermost shell of P form bonds with the 4 neighboring Si atoms, while the 5th electron can ionize to the conduction band at moderate temperatures leaving a positively charged P atom with a coulomb potential screened by the dielectric constant of the host. At low temperatures, this potential can trap an electron, and form an Hydrogen-like system except the six fold degenerate conduction band valleys of Si give rise to a six fold degenerate 1s type ground state. In practice, this six fold degeneracy is lifted by strong coupling between the different valleys caused by deviations of the impurity potential from its coulombic nature in the vicinity of the donor nucleus. If this so called valley-orbit interaction is not taken into account, then the effective mass theory (EMT) predicts a P donor ground state binding energy of −33 meV as opposed to the experimentally measured value of −45.6 meV [87]. The influence of valley-orbit interaction is strongest for the six 1s states, and is negligible for the excited states, which are affected by the bulk properties of the host [13]. The TB model considered here also models the excited states well by its accurate representation of the Si band structure. Hence, we limit our attention here to the effect of valley-orbit interaction on the 1s states.

Study Approaches Theoretical study of donors in Si dates back to the 1950s when Kohn and Luttinger [51] employed symmetry arguments and variational envelope

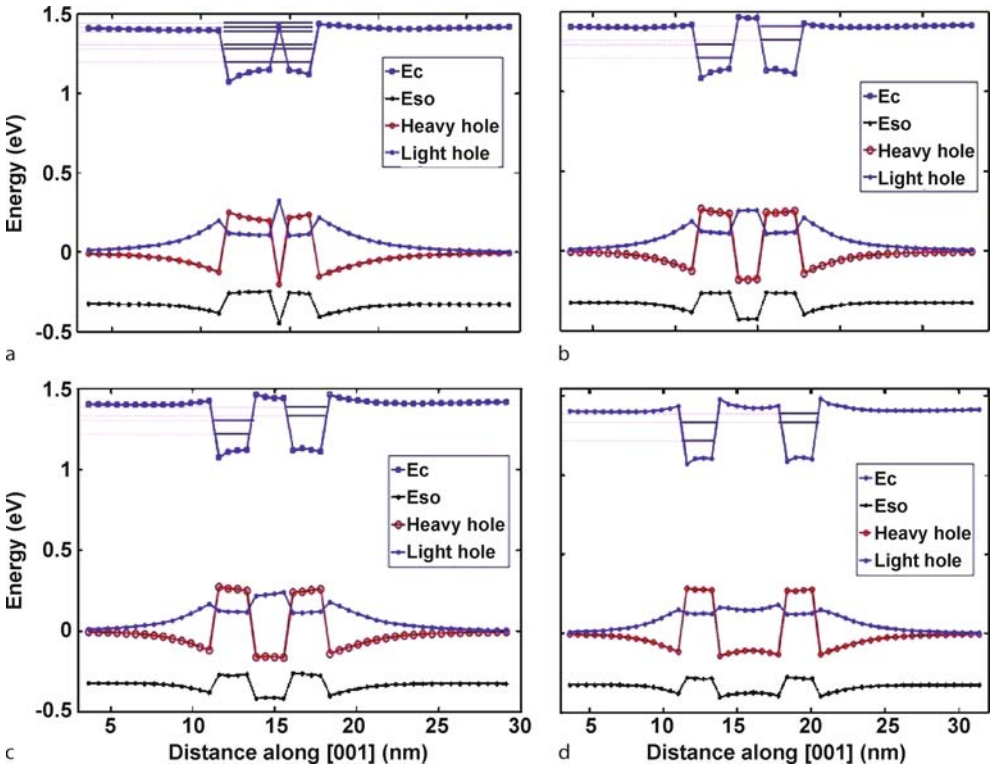

Multimillion Atom Simulations with Nemo3D, Figure 28

Band edge diagrams for double quantum dot systems for several inter-dot spacing: a 0.5 nm b 1 nm c 2 nm and d 4 nm. Strain makes InAs conduction band potential wells shallower, enhancing the coupling between the dots. Noticeable is the effect on hole wells. Strain splits the light hole and heavy hole bands. Within the dot, heavy hole lies above the light hole edge. As strain coupling decreases, heavy hole well become more and more shallower (see b and d)

functions based on EMT to predict the nature of the donor spectrum and wave functions with a fair amount of success. Although many theorists who study donor based nano devices still use the Kohn–Luttinger variational envelope functions, recent approaches [65,82,100] have highlighted the need to consider a more extended set of Bloch states than the six valley minima states and to go beyond the basic EMT assumptions for accurate modeling of impurities. For modeling high precision donor electronics, it is very important to model the basic Physics from a consistent set of assumptions, and to obtain very accurate numbers in addition to correct trends. The model presented here serves these purposes well, and can be used conveniently for large-scale device simulation.

Numerical Study of the Valley–Orbit Interaction The inset of Fig. 31 shows the lowest 6 1s type energy states of a P donor in Si. When valley-orbit coupling is ignored, the six lowest states are degenerate in energy. When Valley-orbit coupling is taken into account, the six fold degenerate states split into a ground state of symmetry $A1$, a triply degenerate state of symmetry T and a doubly degenerate state of symmetry E. Valley-orbit interaction, which arises due to the deviation of the impurity potential from its bulk-like Coulombic nature, is typically modeled by a correction term for the impurity potential in the vicinity of the donor site. The strength of this core-correcting potential determines the magnitude of the splitting of the six 1s states, and varies from impurity to impurity. Here we consider a core correcting cut-off potential U_0 at the donor site, reflecting a global shift of the orbital energies of the impurity. Figure 31 shows how the energy splitting is affected by the strength of U_0. For small U_0, the six 1s type states are degenerate in energy. As U_0 increases in magnitude, we obtain the singlet, triplet and doublet components, as mentioned earlier. Since the triplet (and doublet) states remain degenerate irrespective of U_0, we only plot one state of the T (and E) manifold.

This single core-correction term was found to reproduce the donor eigen states within a few meV, and could be adjusted to match the donor ground state binding energy within a μeV. In general however, the tight-bind-

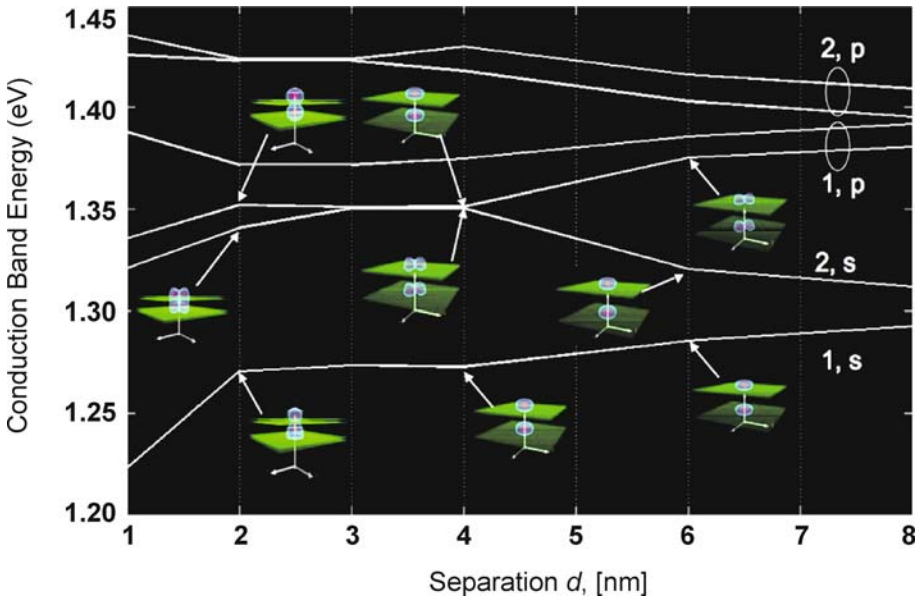

Multimillion Atom Simulations with Nemo3D, Figure 29
Dependence of six lowest electron energy levels on separation distance *d* between the dots. For electron energy levels, the state names are mentioned as *s* or *p* orbital states. Here 1 indicates bonding states whereas 2 indicates anti-bonding states. Wave function plots in *XZ* plane have been shown for some dot separations. Noticeably, e2 for 4 nm separation is a *p* like state while it is s like state in 6 nm separation. So there is a crossover between *p* to *s* for e2 as we increase separation between the dots. Also, e1 for 4 nm separation is confined in lower dot more than upper dot. This is caused by strain coupling which tries to confine ground states in the lower dots in coupled quantum dot systems

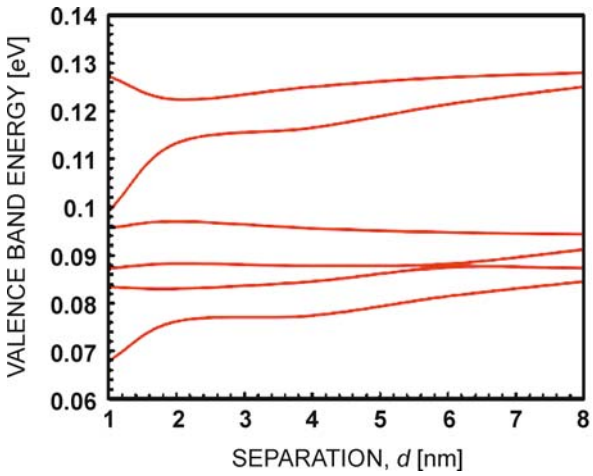

Multimillion Atom Simulations with Nemo3D, Figure 30
Dependence of six lowest hole energy levels on separation distance *d* between the dots

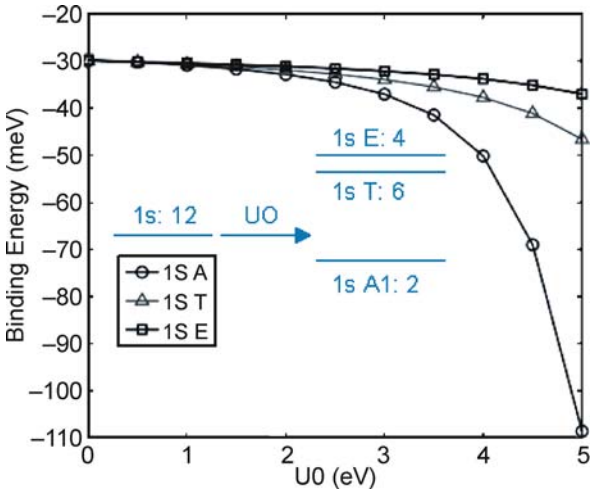

Multimillion Atom Simulations with Nemo3D, Figure 31
Effect of central cell correction U_0 on energy splitting. (*inset*) Group V donor 1s states in Si splitting into 3 components due to valley-orbit interaction

ing parameters for Si can only reproduce the full band structure within a limited accuracy. To model high precision donor electronics within a hundredth of a meV, as is needed in many quantum computing applications, additional core-correction terms are required. In semi-empirical Tight-binding, it is only natural to adjust the on-site orbital energies of the P-donor slightly from their Si counterparts to provide this additional correction. In Fig. 32, we

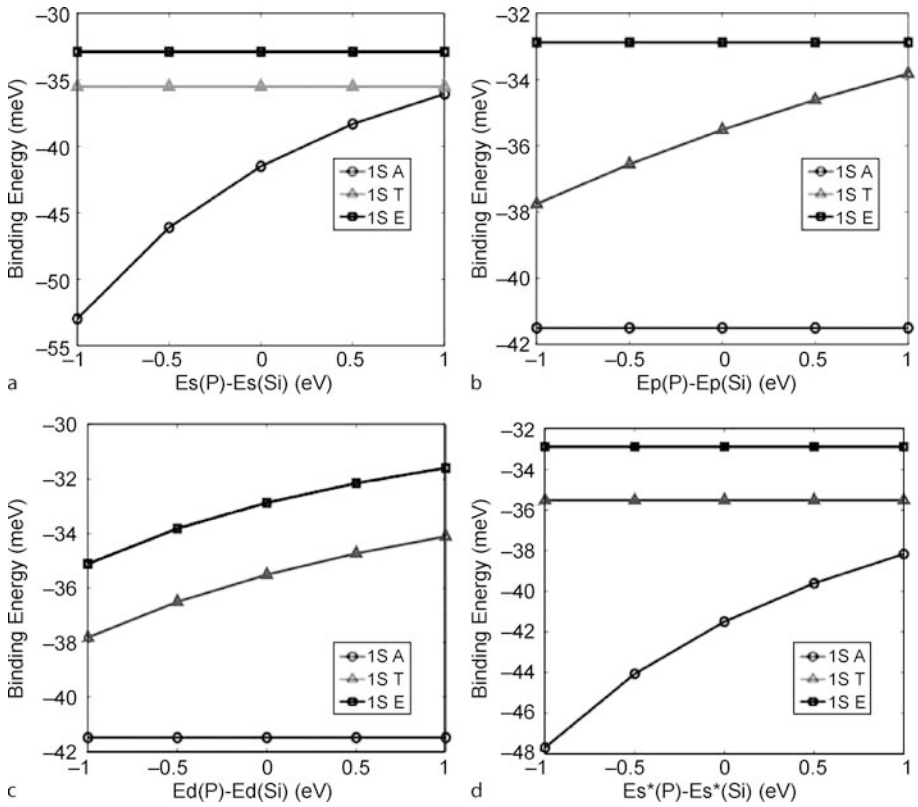

Multimillion Atom Simulations with Nemo3D, Figure 32
Variation of 1s Binding energies with on-site orbital energies. The Triplet (Doublet) states remain degenerate. Hence only 1 triplet (Doublet) is shown

show variation of the binding energy of the 1s manifold as a function of the on-site orbital energies of the donor site. The four on-site energies considered are *Es*, *Ep*, *Ed* and *Es** corresponding to the *s*, *p*, *d*, *s** orbitals respectively. The trends in the plots help us establish a recipe for optimizing the core-correction for a donor species to reflect impurity eigen states within the precision of 0.01 meV. For example, if the only donor ground state of A1 symmetry needs to have a higher binding energy, we can adjust either *Es* or *Es**, each of which will push the *A*1 state deeper in energy without affecting the excited states (Figs. 32a,d). Figure 32c shows that both the triplet and the doublet state can be adjusted in energy by *Ed* without affecting the A1 state, while Fig. 32b shows that the triplet state alone is affected by *Ep*. On the other hand, U_0 reflects a global shift of all the on-site energies, and can affect all the 1s states, as already shown in Fig. 31b. In short there are enough degrees of freedom to empirically adjust the core-correction to obtain very exact eigen values. Once a set of these parameters (U_0, *Es*, *Ep*, *Ed*, *Es**) is fixed, they can be used for a variety of applications like Stark shift, charge qubits, etc. without any additional modification. To model a generic impurity, it is recommended that U_0 be adjusted first so that the ground state binding energy is reproduced accurately. Then one can consider small deviations in a few of these on-site energies to fit the excited states accurately. In most cases, the parameters U_0, *Ep* and *Ed* can be sufficient for accurate modeling. The plots here were obtained by the tight-binding $sp^3d^5s^*$ model without spin. Clearly, this is an empirical process that does not account fully for the different nature of the impurity atom in a host lattice. Additional mapping which includes the change of the impurity to host coupling matrices could be performed possible based on an input from an ab initio method.

Solution Methods–Lanczos and Block Lanczos For a realistic simulation involving a few impurities, one needs to consider a lattice size of about 7 million atoms. In atomistic Tight-Binding with a 20 orbital nearest neighbor model, this involves solving a Hamiltonian with 140 million rows and columns. Although this matrix is considerably sparse, solving for interior eigen values occurring near

Multimillion Atom Simulations with Nemo3D, Table 5

Comparison of the single donor states relative to the conduction band minima of Si for Lanczos and Block Lanczos algorithms. The Lanczos algorithm fails to capture degenerate states, while Block Lanczos is able to resolve degeneracies at the expense of compute time. The eigenvalues were obtained by the $sp^3d^5s^*$ spin model and shows spin degenerate eigen values as well. The slight deviation of the Eigen values from the experimental values is due to the finite size (i. e. confinement effect) of the simulation domain of 30 nm^3

Experiment [5]	Lanczos	Block Lanczos (Block Size 6)	Symmetry
−45.59	−45.599	−45.599	1s (A1)
− 45.59		−45.599	1s (A1)
− 33.89	−33.932	−33.932	1s (T)
− 33.89		−33.932	1s (T)
− 33.89		−33.930	1s (T)
− 33.89		−33.930	1s (T)
− 33.89		−33.930	1s (T)
− 33.89		−33.930	1s (T)
− 32.58	−32.67	−32.670	1s (E)
− 32.58		−32.670	1s (E)
− 32.58		−32.670	1s (E)
− 32.58		−32.670	1s (E)

the conduction band poses a difficult problem. Compared to many other algorithms, the parallel Lanczos algorithm for eigen solution has proved very efficient. However, one drawback of the Lanczos algorithm is its inability to find degenerate and closely clustered eigen values with reliability. A blocked version of Lanczos resolves this problem at the cost of some additional compute time. Since there are many degenerate eigen states present in the unperturbed impurity spectrum, the block Lanczos algorithm was a suitable solution method for the problem outlined here. Table 5 shows the comparison of eigen states obtained from Lanczos, Block Lanczos, and experimentally established values for single donors in bulk Si. While Lanczos fails to capture the degeneracy of the triplet, doublet and spin states, Block Lanczos resolves all the 12 eigenvalues reliably. The computational system considered here spans a domain of 30.5 nm × 30.5 nm × 30.5 nm and contains about 1.4 million atoms. Closed boundary condition is applied in all three dimensions.

Typical Application–Donor Based Charge Qubits An impurity based charge qubit involves a single electron bound to two ionized P donors in Si. A qubit can be encoded based on the localization of the electron in either of the two impurities [36]. When the Hamiltonian of such a system is solved, a set of bonding and anti-bonding states are obtained from the set of single impurity states. An important parameter in quantum computing applications is the tunnel coupling between the two lowest eigen states. This parameter depends on the separation of the two lowest eigen states of the P_2+ problem, and is sensitive to relative donor placements and gate voltages. Figure 33 shows the tunnel coupling as a function of donor separation along [100] and [110] calculated in tight binding. The tunnel coupling tends to decay, as the impurities are located farther apart. While variation of tunnel coupling is found to be smooth along [100] direction, it is highly oscillatory along [110]. This is due to interference between Bloch parts of the impurity wave functions contributed by the Si crystal. These trends are already well established in literature [39] from effective mass theory. The impurity model in TB presented here is able to capture these effects with convenience.

Unlike EMT, the methodology developed here can consider a more extended Bloch structure of the host and incorporate many realistic device effects such as finite device sizes, interfaces under one framework, and is convenient for large scale device simulations. Treatment of such factors enables precise comparison with experimentally measured quantities, as was done in [82], where the hyperfine stark effect for a P donor was calculated in good agreement with experiment [19], and discrepancies with previous EMT [29] based calculations were resolved. Further work is under way to study CTAP [33] based architectures [37], charge qubits [36,39,52] and investigate donor-interface well hybridization in Si FinFET devices [20,57,87].

Si on SiGe Quantum Well

Many quantum dot based [30] or impurity based [41] quantum computing architectures are proposed to be fabricated in Si/SiGe heterostures. Since silicon has multi-

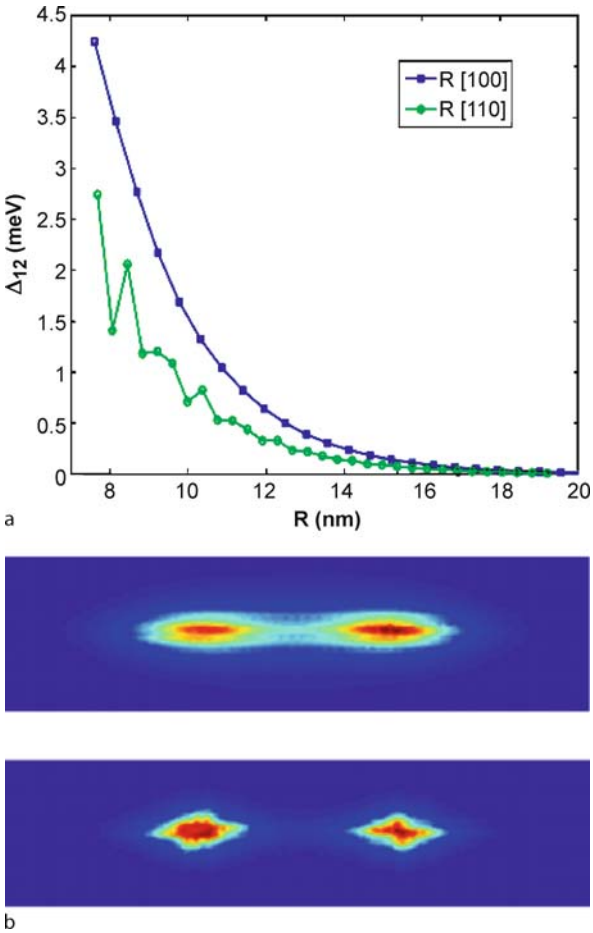

Multimillion Atom Simulations with Nemo3D, Figure 33
a Variation of tunnel coupling for a P_2+ system with impurity separation along [100] & [110]. b Formation of Molecular states for P_2+

ple degenerate values it is critical to engineer these degeneracies out of the system to avoid dephasing of qubits. Miscut substrates (Fig. 34b) as opposed to flat substrates (Fig. 34a) are often used to ensure uniform growth of Si/SiGe heterostructures. However, a miscut modifies the energy spectrum of a QW. In a flat QW the two degenerate valleys in strained Si split in energy and the valley minima occur at $\pm k_x = 0$. Valley splitting (VS) in a flat QW is a result of interaction among states in bulk z-valleys centered at $k_z = k_0$, where k_0 is position of the valley-minimum in strained Si. The energy splitting between these two lowest lying valleys is called as valley-splitting (VS). In quantum computing devices, VS is an important design parameter controlling the electron spin decoherence time [14,15,16]. In a miscut QW lowest lying valleys are degenerate with minima at $\pm k^0_{x'}$ [44]. Thus atomic scale modulation of surface topology leads to very different electronic structures in flat and miscut QWs. As a consequence of this, flat and miscut QWs respond differently to the applied electric and magnetic fields. In the presence of lateral confinement in miscut QW the two degenerate valleys in Fig. 34c interact and give rise to VS.

Traditional magnetic probe techniques such as Shubnikov de Haas oscillations are used to measure energy spectrum of QWs. Valley and Spin splittings are determined by electron-valley resonance (EVR) [31] and electron-spin resonance (ESR) [26] techniques. In these measurements in plane (lateral) confinement of the Landau-levels is provided by the magnetic field. Figure 34d shows the dependence of VS on applied magnetic field in flat and ideal miscut QWs. Ideal miscut QWs refer to the miscut QWs with no step roughness. VS in flat QW is independent of magnetic field because in these QWs VS arises from z-confinement provided by the confining SiGe buffers [44]. In miscut QWs, however, VS is the result of the combined effect of the two confinements, the z-confinement provided by the SiGe buffers and the lateral confinement provided by the applied magnetic field. The two degenerate valleys centered at $\pm k^0_{x'}$ along x' direction in the miscut QWs (Fig. 34c) interact and split in the presence of magnetic field. At low magnetic fields the dependence of VS in miscut QWs on the applied magnetic field is linear. In calculations of Fig. 34c,d QWs are assumed to be perfect. Disorders such as step roughness and alloy disorder in SiGe buffer which are inherently present is the experiments are completely ignored. As a result calculated VS is nearly an order of magnitude lower than the experimentally measured values (Fig. 35d).

Miscut substrates undergo reconstruction to reduce the surface free energy which gives rise to the step roughness [105] (Fig. 35b,c). This type of step roughness disorder is present at the Si/SiGe interface. Another type of disorder in Si/SiGe heterostructures is the random alloy disorder in SiGe buffer. These two disorders are always present in actual QW devices and thus need to be taken into account in VS computations. Schematic of an electronic structure computation domain is shown in Fig. 35a. QWs extend 15 nm along y-direction to take into account the step roughness disorder shown in Fig. 35c. x' confinement due to the magnetic field is incorporated through the Landau gauge ($\vec{A} = Bx\hat{y}$). The resulting vector potential ($\vec{A}$) is introduced into the tight-binding Hamiltonian trough the gauge invariant Peierl's substitution [8,18,32]. Closed boundary conditions are used in x and z directions while y-direction is assumed to be (quasi-)periodic. The confinement induced by closed boundary conditions in direction compete with the magnetic field confinement. The lateral extension of the strain and the electronic structure

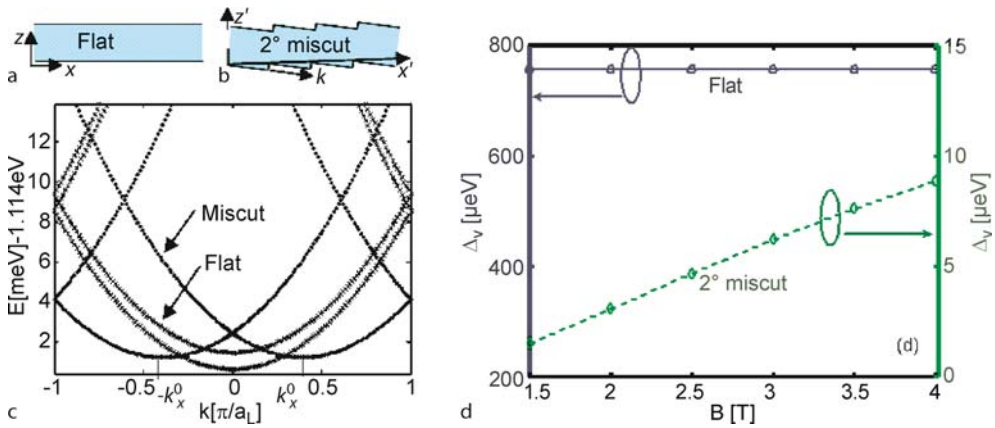

Multimillion Atom Simulations with Nemo3D, Figure 34

a Schematic of a Si QW grown on [001] substrate. The crystal symmetry directions are along x[100]and z[001]. **b** Schematic of a 2°C miscut QW unit cell. The unit cell is periodic along x' and y directions and confined in z' direction. Miscut angle is 2°C. The step height is $a/4$ which corresponds to one atomic layer, where a is lattice constant. **c** Band structure of 5.26 nm thick flat/miscut QW along x/x' direction. Flat QW shows the presence of two non-degenerate valleys separated by an energy know as VS. Miscut QW shows the presence of two degenerate valleys centered at $\pm k^0_{x'}$. **d** VS in 10 nm thick flat (001) and 2°C miscut Si QWs. Electric field in z-direction is 9 MV/m

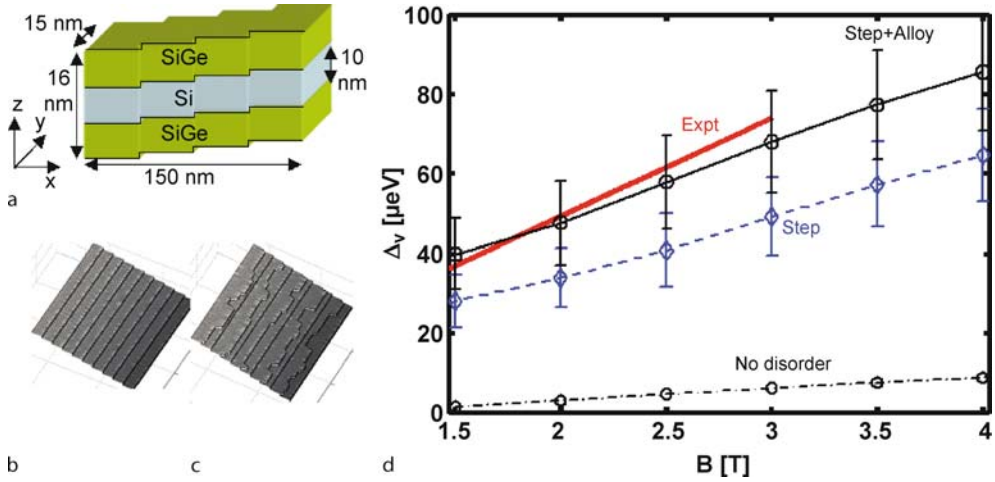

Multimillion Atom Simulations with Nemo3D, Figure 35

a Schematics of the simulation domain. **b** Ideal steps on a miscut substrate. **c** Step disorder resulting from the surface reconstruction on the miscut substrate. **d** VS of the first Landau-level in a 10 nm thick strained Si QW. The VS labeled as 'No disorder' is shown for comparison and it is same as that of in Fig. 1d. VS increases due to the step-disorder. When alloy-disorder in SiGe buffer is included along with the step disorder the computed VS matches the experimentally measured values. Error bars represent the standard deviation in VS. In the calculations of VS labeled as 'No disorder' and 'Step disorder' uniform biaxial strain of $\varepsilon|| = 0.013$ is assumed

domain is set to 150 nm, which is about 7 times larger than the maximum magnetic confinement length in a 2DEG at B = 1.5T ($\simeq$21 nm). For the magnetic field ranges of 1.5–4T confinement is dominated by the magnetic field and no lateral x-confinement effects due to the closed boundary conditions are visible in simulations. Modulation doping in Si/SiGe heterostructures induces built-in electric field. In the simulations performed here constant electric field of 9 MV/m is assumed in the QW growth direction.

Figure 35d shows the computed VS in 2°C miscut QWs. VS in ideal miscut QWs is an order of magnitude lower compared to the experimentally measured values. If the step-roughness disorder is included in the simulations, the computed VS is higher compared to that of an ideal miscut QW. In these calculations surface roughness model of [6] is used and the uniform biaxial strain of $\varepsilon|| = 0.013$ which corresponds to $Si_{0.7}Ge_{0.3}$ buffer composition is assumed. This VS, however, is slightly smaller than the ex-

perimentally measured VS. This discrepancy can be answered by adding SiGe buffers in the electronic structure simulation domain. 3 nm of SiGe buffer is included on top and bottom of the Si QW to take into account the wavefunction penetration into the finite barrier QW buffers. Strain computation domain has the same x and y dimensions as the electronic structure domain. To take into account the long range nature of strain [54] 40 nm of SiGe buffer is included on both sides of Si QW. z-dimension of the strain domain is 90 nm. Valance Force Field (VFF) model of Keating [42] is employed to calculate the relaxed geometries. The VS computed by taking both step and alloy-disorder into account is found to match closely to the experimentally measured values.

The time required to compute the 10 million atom strain calculation on 20 CPUs of an Intel x86-64 dual core linux cluster is about 9 hours. The subsequent 2 million atom electronic structure calculation requires 10 hours.

SiGe Nanowires

Semiconductor nanowires are being actively investigated as the potential candidates for the end of the semiconductor technology roadmap devices. They are also attractive for sensing applications due to their high surface-to-volume ration. Several researchers have recently demonstrated the nanowire field-effect transistors (FETs) fabricated from pure elemental or compound semiconductors like Si [24], Ge [33], and GaAs [76] as well as semiconductor alloys like SiGe [45], and their III–V counterparts. For the device design at the nanoscale, it is important to understand and to be able to predict transport properties of nanowires. Atomistic disorder such as alloy disorder, surface roughness and inhomogeneous strain strongly influence the electronic structure and the charge transport in nanoscale devices. To simulate nanodevices traditional effective mass approaches should be abandoned [98] and more accurate atomistic approaches should be adopted. Here, SiGe alloy nanowires are studied from two different perspectives. First, the electronic structure where bandstructure of a nanowire is obtained by projecting out small cell bands from a supercell eigenspectrum [10,11] and second, the transport where transmission coefficients through the nanowire are calculated using an atomistic wave function (WF) approach [17,64].

SiGe random alloy nanowires have two types of disorders: atom disorder due to random alloying and inhomogeneous strain disorder due to different Si-Si, Ge-Ge, and Si-Ge bond lengths. These disorders break the translational symmetry in semiconductor alloy nanowires. Thus one runs into the problem of choosing a unit cell for the bandstructure calculation. Disorder can be taken into account by simulating larger repeating units (supercells) containing many small cells (Fig. 34). The nanowire bandstructure obtained from the supercell calculation is folded. The one dimensional version of the zone-unfolding method [10,11] is used to project out the approximate eigenspectrum of the nanowire supercell on the small cell Brillouin-Zone. The probability sum rule [10] is used to extract the approximate bandstructure of the alloy nanowire from the projected probabilities. The small cell bandstructure obtained by this method captures the effect of SiGe alloy disorder on the electronic structure.

The nanowire geometry is specified in terms of conventional Zincblende (cubic) unit cells as $n_x \times n_y \times n_z$ where n_i is the number of cubes in direction$-i$. The wire dimensions are $40 \times 6 \times 6$ ($22.3 \times 3.3 \times 3.3$ nm) i. e. it is constructed from 40 $1 \times 6 \times 6$ slabs along [100] crystallographic direction. Figure 36a depicts a sliver cut through the center of the SiGe nanowire indicating the atomistically resolved disorder of the wire. Only the central 5 nm long portion of this 22 nm long wire is shown for visualization purpose. All electronic structure and transport calculations have been done in 20-band $sp^3d^5s^*$ tight-binding model with spin-orbit coupling. The bulk tight-binding and strain Si and Ge parameters are taken from [9,13]. Relaxed wire geometries are calculated from Valance Force Field approach.

The unfolding procedure to compute an approximate bandstructure from the large supercell calculation requires many eigenvectors. In practice these eigenstates are closely spaced in energy and Lanczos algorithm requires about 50 000 iterations to resolve 575 states in the energy range of interest. Such calculations require about 5.5 hours on 30 cores of an Intel x86-64 dual core linux cluster machine.

Figure 36b shows the conduction bandstructure of the first slab out of 40 slabs along nanowire length. Δ_4 valleys are split into four separate bands while Δ_2 valley bands are doubly degenerate. Local band-edge plots of the lowest Δ_4 and Δ_2 valley minima are shown in Fig. 36c. This so called local bandstructure of each slab is calculated assuming that this slab repeats infinitely along the nanowire. Due to fluctuations in atomic arrangements along the nanowire length one expects to see the different bandstructures for each slab as shown in Fig. 37a. Variations of band-edges along the nanowire length cause reflections which lead to the formation of the localized states and peaks in transmission plots.

The NEMO 3-D team is currently developing with Mathieu Luisier at ETH Zurich a new 3-D quantum transport simulator [64]. Here we show a comparison of a 3-D disordered system transport simulation with a NEMO 3-D

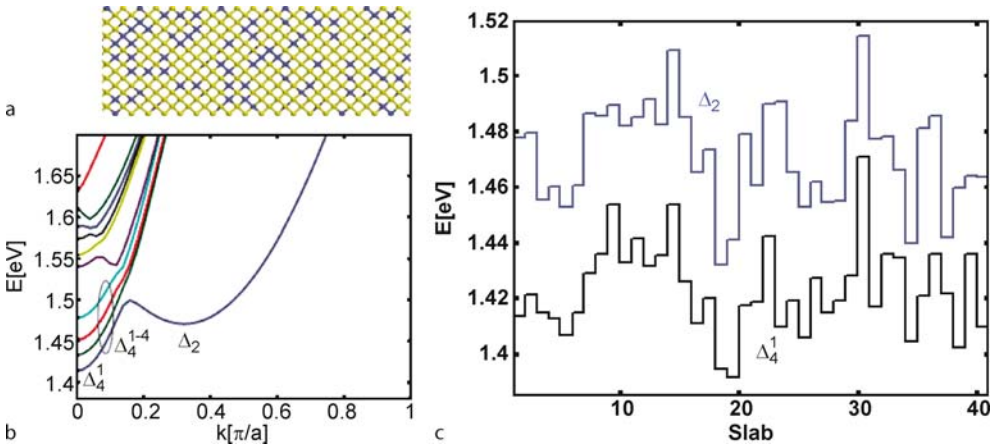

Multimillion Atom Simulations with Nemo3D, Figure 36

a Atomistically resolved disorder in the $Si_{0.8}Ge_{0.2}$ 40 × 6 × 6 nanowire. b Conduction bandstructure of the first slab assuming that the slab is repeated infinitely. Δ_4 valleys are split into four separate bands. Δ_2 valley bands are doubly degenerate. c Bandedge minima of lowest energy Δ_4 and Δ_2 valleys along length of the nanowire

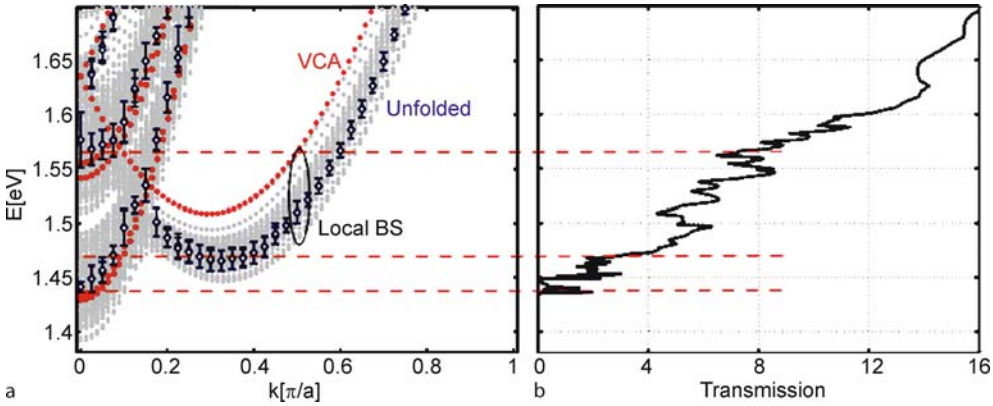

Multimillion Atom Simulations with Nemo3D, Figure 37

a Bandstructures of 40 × 6 × 6 Si0.8Ge0.2 alloy nanowire in local bandstructure (*gray*), VCA (*red*) and zone-unfolding (*blue*) formulations. b Transmission through 40 × 6 × 6 wire. Steps in transmission are identified as resulting from new bands appearing in projected bandstructure. Note that atomistic, narrow 1D wires result automatically into 1D localization

electronic structure calculation. The transmission coefficient (Fig. 37b) shows the noisy behavior due to random SiGe alloy disorder and inhomogeneous strain disorder in the wire. Steps in the transmission plot can be roughly related to the unfolded bandstructure (Fig. 35a) from supercell calculations. Four separate Δ_4 valley bands appear as a single band with a finite energy spread in the projected bandstructure. These four bands turn on near 1.44 eV which corresponds to the conduction band transmission turn on. Two Δ_2 valley bands turn on near 1.47 eV which leads to a step in the transmission. 4 more channels due to higher Δ_4 valley sub-bands turn on near 1.57eV. These transmission features can not be related to the conventional virtual crystal approximation (VCA) bandstructure shown in Fig. 35a. Peaks in the transmission plot can be related to the local density of states in the wire [43].

Projected supercell bandstructures and atomistic transport calculations are found to be complimentary and mutually supporting. Both methods provide better instight into the transport through the disordered nanowires.

Summary and Future Directions

NEMO 3-D is introduced as a versatile, open source *electronic structure* code that can handle device domains relevant for realistic large devices. Realistic devices containing millions of atoms can be computed with reasonably, easily available cluster computers. NEMO 3-D employs a VFF

Keating model for strain and the 20-band $sp^3d^5s^*$ empirical tight-binding model for the electronic structure computation. It is released under an open source license and maintained by the NCN, an organization dedicated to develop and deploy advanced nanoelectronic modeling and simulation tools. NEMO 3-D is not limited to research computing alone; A first educational version including visualization capabilities has been released on http://www.nanoHUB.org and has been used by hundreds of users for thousands of simulations. The full version of NEMO3D will soon be available for device engineers, material scientists, educators, and students through the nanoHUB, powered by the NSF Teragrid. Tool documentation, tutorials, and case studies will be posted on nanoHUB as supplemental material. We will generate and deliver tutorials on parallelization and software development through the nanoHUB.

NEMO 3-D demonstrates the capability to model a large variety of relevant, realistically sized nanoelectronic devices. The impact of atomistic strain and piezoelectricity on the electronic structure in dome shaped quantum dots is explored. Under the assumptions of realistic boundary conditions, strain is found to be long-ranged and penetrate around 25 nm into the dot substrate thus stressing the need for using large dimensions of these surrounding layers and at least 3 million atoms in the simulations. The true symmetry of the quantum dots is found to be lower than the geometrical shape symmetry because of the fundamental atomistic nature of the underlying zincblende crystal lattice. Atomistic strain is found to induce further optical polarization anisotropy favoring the [110] direction and pronounced non-degeneracy in the quantum dot excited states, magnitude (few meV) of which depends mainly on the dot size and surrounding material matrix. First order piezoelectric potential, on the other hand, favors the [1$\underline{1}$0] direction, reduces the non-degeneracy in the *P* states and is found to be strong enough to *flip* the optical polarization in certain sized quantum dots [6]. Simulations of QD stacks exemplify the complicated mechanical strain and quantum mechanical interactions on confined electronic states. Molecular states can be observed when the dots are in close proximity. Simulations of SiGe buffered Si QWs indicate the importance of band-to-band interactions that are naturally understood in the NEMO 3-D basis. Valley splitting is computed as a function of magnetic field matching experimental data. Simulations of disordered SiGe alloyed nanowires indicate the critical importance of the treatment of atomistic disorder. Typical approaches of a smoothed out material (VCA) or considerations of bandstructure in just individual slices clearly fail to represent the disordered nanowire physics. A semi-empirical tight binding model for Group V donors in Silicon is presented. The dependence of valley-orbit interaction on on-site cut-off potential and orbital energies is explored. A block based Lanczos algorithm was demonstrated as a robust and reliable method of finding eigenvalues and vectors of the resulting system. The technique outlined here enables high precision modeling of impurity based quantum electronics with relative ease and accuracy.

All these NEMO 3-D calculations underline the importance to represent explicitly the atomistically resolved physical system with a physics based local orbital representation. Such million atom systems result in system sizes of tens of millions and end-to-end 52 million atom simulations representing one *billion* degree of freedom systems were presented. The complexity of the system demands the use of well qualified, tuned, optimized algorithms and modern HPC platforms. Building and maintaining such a code is not a light undertaking and requires a significant group community effort.

Integrated circuit design faces a crisis – the 40 year process of transistor downscaling has led to atomic-scale features, making devices subject to unavoidable manufacturing irregularities at the atomic scale and to heat densities comparable to a nuclear reactor. *A new approach to design that embraces the atomistic, quantum mechanical nature of the constituent materials is necessary to develop more powerful yet energy miserly devices*. We are in the process of developing a general-purpose simulation engine. It will model not only the electronic band structure but also the out-of-equilibrium electron *transport* in realistically extended devices using fully quantum mechanical (QM) models in an atomistic material description containing millions of atoms. The research will enable discovery of new technologies for faster switching, smaller feature size, and reduced heat generation. Using this new approach, designers can directly address questions of quantization and spin, tunneling, phonon interactions, and heat generation. It is widely accepted that the Non-Equilibrium Green Function Formalism (NEGF) QM statistical mechanics theory, in conjunction with an atomistic basis, can answer these questions. It is also widely perceived that the problem is computationally hard to solve. A generalized approach to tri-level parallelism in voltage, energy, and space is highly desired. Another task addresses the bottleneck of calculating open boundary conditions (BCs) for large cross sections for realistically large structures. The BCs can be reused for each voltage point and each charge self-consistent iteration. With a view to achieving these goals, the necessary levels of parallelism to tackle the problem on 200,000+ CPUs have been designed and demonstrated to scale well. Computer scientists and HPC

experts embedded in the team will guide the implementation and explore performance, execution reliability, and alternative hardware and algorithms. The new simulation code named OMEN (with non-equilibrium Green function and 3-D atomistic representation) will be an open source project and disseminated through the nanoHUB.

Acknowledgments

The work has been supported by the Indiana 21st Century Fund, Army Research Office, Office of Naval Research, Semiconductor Research Corporation, ARDA, the National Science Foundation. The work described in this publication was carried out in part at the Jet Propulsion Laboratory, California Institute of Technology under a contract with the National Aeronautics and Space Administration. The development of the NEMO 3-D tool involved a large number of individuals at JPL and Purdue, whose work has been cited. Drs. R. Chris Bowen, Fabiano Oyafuso, and Seungwon Lee were key contributors in this large effort at JPL. The authors acknowledge an NSF Teragrid award DMR070032. Access to the Bluegene was made available through the auspices of the Computational Center for Nanotechnology Innovations (CCNI) at Rensselaer Polytechnic Institute. Access to the Oak Ridge National Lab XT3/4 was provided by the National Center for Computational Sciences project. We would also like to thank the Rosen Center for Advanced Computing at Purdue for their support. nanoHUB computational resources were used for part of this work.

Bibliography

Primary Literature

1. Agnello PD (2002) Process requirements for continued scaling of CMOS—the need and prospects for atomic-level manipulation. IBM J Res Dev 46:317–338
2. Ahmed S, Usman M, Heitzinger C, Rahman R, Schliwa A, Klimeck G (2007) Atomistic simulation of non-degeneracy and optical polarization anisotropy in zincblende quantum dots. In: 2nd Annual IEEE International Conference on Nano/Micro Engineered and Molecular Systems (IEEE-NEMS), Bangkok, Thailand
3. Arakawa Y, Sasaki H (1982) Multidimensional quantum well laser and temperature dependence of its threshold current. Appl Phys Lett 40:939
4. Bae H, Clark S, Haley B, Klimeck G, Korkusinski M, Lee S, Naumov M, Ryu H, Saied F (2007) Electronic structure computations of quantum dots with a billion degrees of freedom. Supercomputing 07, Reno, NV, USA
5. Bester G, Wu X, Vanderbilt D, Zunger A (2006) Importance of second-order piezoelectric effects in zinc-blende semiconductors. Phys Rev Lett 96:187602
6. Bester G, Zunger A (2005) Cylindrically shaped zinc-blende semiconductor quantum dots do not have cylindrical symmetry: Atomistic symmetry, atomic relaxation, and piezoelectric effects. Phys Rev B 71:045318. Also see references therein
7. Bowen R, Klimeck G, Lake R, Frensley W, Moise T (1997) Quantitative resonant tunneling diode simulation. J Appl Phys 81:207
8. Boykin T, Bowen R, Klimeck G (2001) Electromagnetic coupling and gauge invariance in the empirical tight-binding method. Phys Rev B 63:245314
9. Boykin T, Kharche N, Klimeck G (2007) Brillouin zone unfolding of perfect supercells composed of non-equivalent primitive cells. Phys Rev B 76:035310
10. Boykin T, Kharche N, Klimeck G, Korkusinski M (2007) Approximate bandstructures of semiconductor alloys from tight-binding supercell calculations. J Phys: Condens Matter 19:036203
11. Boykin T, Klimeck G (2005) Practical application of zone-folding concepts in tight-binding. Phys Rev B 71:115215
12. Boykin T, Klimeck G, Bowen R, Oyafuso F (2002) Diagonal parameter shifts due to nearest-neighbor displacements in empirical tight-binding theory. Phys Rev B 66:125207
13. Boykin T, Klimeck G, Oyafuso F (2004) Valence band effective mass expressions in the $sp^3d^5s^*$ empirical tight-binding model applied to a new Si and Ge parameterization. Phys Rev B 69:115201
14. Boykin T, Klimeck G, Eriksson M, Friesen M, Coppersmith S, Allmen P, Oyafuso F, Lee S (2004) Valley splitting in strained Si quantum wells. Appl Phys Lett 84:115
15. Boykin T, Klimeck G, Eriksson M, Friesen M, Coppersmith S, Allmen P, Oyafuso F, Lee S (2004) Valley splitting in low-density quantum-confined heterostructures studied using tight-binding models. Phys Rev B 70:165325
16. Boykin T, Klimeck G, Allmen P, Lee S, Oyafuso F (2005) Valley-splitting in V-shaped quantum wells. J Appl Phys 97:113702
17. Boykin T, Luisier M, Schenk A, Kharche N, Klimeck G (2007) The electronic structure and transmission characteristics of disordered AlGaAs nanowires. IEEE Trans Nanotechnology 6:43
18. Boykin T, Vogl P (2001) Dielectric response of molecules in empirical tight-binding theory. Phys Rev B 65:035202
19. Bradbury F et al (2006) Stark tuning of donor electron spins in silicon. Phys Rev Lett 97:176404
20. Calder'on MJ, Koiler B, Hu X, Das Sarma S (2006) Quantum control of donor electrons at the Si-SiO_2 interface. Phys Rev Lett 96:096802
21. Canning A, Wang LW, Williamson A, Zunger A (2000) Parallel empirical pseudopotential electronic structure calculations for million atom systems. J Comp Phys 160:29
22. Chen P, Piermarocchi C, Sham L (2001) Control of exciton dynamics in nanodots for quantum operations. Phys Rev Lett 87:067401
23. Colinge JP (2004) Multipole-gate SOI MOSFETs. Solid-State Elect 48:897–905
24. Cui Y, Lauhon L, Gudiksen M, Wang J, Lieber C (2001) Diameter-controlled synthesis of single-crystal silicon nanowire. Appl Phys Lett 78:2214
25. Debernardi A et al (2006) Computation of the Stark effect in P impurity states in silicon. Phys Rev B 74:035202
26. Dobers M, Klitzing K, Schneider J, Weimann G, Ploog K (1998) Electrical detection of nuclear magnetic resonance in GaAs-$Al_xGa_{1-x}As$ heterostructures. Phys Rev Lett 61:1650

27. Eriksson M, Friesen M, Coppersmith S, Joynt R, Klein L, Slinker K, Tahan C, Mooney P, Chu J, Koester S (2004) Spin-based quantum dot quantum computing in Silicon. Quantum Inf Process 3:133
28. Fafard S, Hinzer K, Raymond S, Dion M, Mccaffrey J, Feng Y, Charbonneau S (1996) Red-emitting semiconductor quantum dot lasers. Science 22:1350
29. Friesen M et al (2005) Theory of the Stark effect for P donors in Si. Phys Rev Lett 94:186403
30. Friesen M, Rugheimer P, Savage D, Lagally M, van der Weide D, Joynt R, Eriksson M (2003) Practical design and simulation of silicon-based quantum-dot qubits. Phys Rev B 67:121301
31. Goswami S, Slinker KA, Friesen M, McGuire LM, Truitt JL, Tahan C, Klein LJ, Chu JO, Mooney PM, van der Weide DW, Joynt R, Coppersmith SN, Eriksson MA (2007) Controllable valley splitting in silicon quantum devices. Nat Phys 3:41
32. Graf M, Vogl P (1995) Electromagnetic fields and dielectric response in empirical tight-binding theory. Phys Rev B 51:4940
33. Greentree A, Cole J, Hamilton A, Hollenberg L (2004) Coherent electronic transfer in quantum dot systems using adiabatic passage. Phys Rev B 70:235317
34. Greytak A, Lauhon L, Gudiksen M, Lieber C (2004) Growth and transport properties of complementary germanium nanowire field-effect transistors. Appl Phys Lett 84:4176
35. Grundmann M, Stier O, Bimberg D (1995) InAs/GaAs pyramidal quantum dots: Strain distribution, optical phonons, and electronic structure. Phys Rev B 52:11969
36. Hollenberg L et al (2004) Charge-based quantum computing using single donors in semiconductors. Phys Rev B 69:113301
37. Hollenberg L et al (2006) Two-dimensional architectures for donor-based quantum computing. Phys Rev B 74:045311
38. Klimeck G, Mannino M, McLennan M, Qiao W, Wang X (2008) https://www.nanohub.org/simulation_tools/qdot_tool_information
39. Hu X et al (2005) Charge qubits in semiconductor quantum computer architecture: Tunnel coupling and decoherence. Phys Rev B 71:235332
40. Jancu J, Scholz R, Beltram F, Bassani F (1998) Empirical spds* tight-binding calculation for cubic semiconductors: General method and material parameters. Phys Rev B 57:6493
41. Kane B (1998) A silicon-based nuclear spin quantum computer. Nature 393:133
42. Keating P (1966) Effect of invariance requirements on the elastic strain energy of crystals with application to the diamond structure. Phys Rev :145
43. Kharche N, Luisier M, Boykin T, Klimeck G (2008) Electronic structure and transmission characteristics of SiGe nanowire. J Comput Electron 7:350; Klimeck G, Ahmed S, Bae H, Kharche N, Clark S, Haley B, Lee S, Naumov M, Ryu H, Saied F, Prada M, Korkusinski M, Boykin T (2007) Atomistic simulation of realistically sized nanodevices using NEMO 3-D: Part I – Models and benchmarks. IEEE Trans Electron Devices 54:2079; Klimeck G, Ahmed S, Kharche N, Korkusinski M, Usman M, Prada M, Boykin T (2007) Atomistic simulation of realistically sized nanodevices using NEMO 3-D: Part II – Applications. IEEE Trans Electron Devices 54:2090
44. Kharche N, Prada M, Boykin T, Klimeck G (2007) Valley-splitting in strained silicon quantum wells modeled with 2 degree miscuts, step disorder, and alloy disorder. Appl Phys Lett 90:092109
45. Kim C, Yang J, Lee H, Jang H, Joa M, Park W, Kim Z, Maeng S (2007) Fabrication of $Si_{1-x}Ge_x$ alloy nanowire field-effect transistors. Appl Phys Lett 91:033104
46. Klimeck G, Boykin T, Chris R, Lake R, Blanks D, Moise T, Kao Y, Frensley W (1997) Quantitative simulation of strained InP-based resonant tunneling diodes. In: Proceedings of the 1997 55th IEEE Device Research Conference Digest:92
47. Klimeck G, Bowen R, Boykin T, Cwik T (2000) sp^3s^* tight-binding parameters for transport simulations in compound semiconductors. Superlattices Microstruct 27:519–524
48. Klimeck G, Bowen R, Boykin T, Salazar-Lazaro C, Cwik T, Stoica A (2000) Si tight-binding parameters from genetic algorithm fitting. Superlattices Microstruct 27:77–88
49. Klimeck G, Oyafuso F, Boykin T, Bowen R, Allman P (2002) Development of a nanoelectronic 3-D (NEMO 3-D) simulator for multimillion atom simulations and its application to alloyed quantum dots. Comput Model Eng Sci 3:601
50. Klimeck G, Boykin T, Luisier M, Kharche N, Schenk A (2006) A Study of alloyed nanowires from two perspectives: approximate dispersion diagrams and transmission coefficients. In: Proceedings of the 28th International Conference on the Physics of Semiconductors ICPS 2006, Vienna, Austria
51. Kohn W, Luttinger J (1995) Theory of donor states in silicon. Phys Rev 98:915
52. Koiller B, Hu X, Das Sarma S (2006) Electric-field driven donor-based charge qubits in semiconductors. Phys Rev B 73:045319
53. Korkusinski M, Klimeck G (2006) Atomistic simulations of long-range strain and spatial asymmetry molecular states of seven quantum dots. J Phys Conf Ser 38:75–78
54. Korkusinski M, Klimeck G, Xu H, Lee S, Goasguen S, Saied F (2005) Atomistic simulations in nanostructures composed of tens of millions of atoms: Importance of long-range strain effects in quantum dots. Proceedings of 2005 NSTI Conference, Anaheim, CA
55. Korkusinski M, Saied F, Xu H, Lee S, Sayeed M, Goasguen S, Klimeck G (2005) Large scale simulations in nanostructures with NEMO 3-D on linux clusters. In: Linux Cluster Institute Conference, Raleigh, NC
56. Lanczos C (1950) An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. J Res Natl Bur Stand 45
57. Lansbergen GP, Rahman R, Wellard CJ, Woo I, Caro J, Collaert N, Biesemans S, Klimeck G, Hollenberg LCL, Rogge S (2008) Gate induced quantum confinement transition of a single dopant atom in a Si FinFET. Nature Phys 4:656
58. Lazarenkova O, Allmen P, Oyafuso F, Lee S, Klimeck G (2004) Effect of anharmonicity of the strain energy on band offsets in semiconductor nanostructures. Appl Phys Lett 85:4193
59. Lee S, Kim J, Jönsson L, Wilkins J, Bryant G, Klimeck G (2002) Many-body levels of multiply charged and laser-excited InAs nanocrystals modeled by empirical tight binding. Phys Rev B 66:235307
60. Lee S, Lazarenkova O, Oyafuso F, Allmen P, Klimeck G (2004) Effect of wetting layers on the strain and electronic structure of InAs self-assembled quantum dots. Phys Rev B 70: 125307
61. Lee S, Oyafuso F, Allmen P, Klimeck G (2004) Boundary conditions for the electronic structure of finite-extent, embedded semiconductor nanostructures with empirical tight-binding model. Phys Rev B 69:045316
62. Liang G, Xiang J, Kharche N, Klimeck G, Lieber C, Lundstrom

M (2006) Performance analysis of a Ge/Si core/shell nanowire field effect transistor. cond-mat 0611226
63. Loss D, DiVincenzo DP (1998) Quantum computation with quantum dots. Phys Rev A 57:120
64. Luisier M, Schenk A, Fichtner W, Klimeck G (2006) Atomistic simulation of nanowires in the $sp^3d^5s^*$ tight-binding formalism: From boundary conditions to strain calculations. Phys Rev B 74:205323
65. Martins A et al (2004) Electric-field control and adiabatic evolution of shallow donor impurities in silicon. Phys Rev B 69:085320
66. Maschhoff K, Sorensen D (1996) A portable implementation of ARPACK for distributed memory parallel architectures. Copper Mountain Conference on Iterative Methods, Copper Mountain, 9–13 April 1996
67. Maximov M, Shernyakov Y, Tsatsul'nikov A, Lunev A, Sakharov A, Ustinov V, Egorov A, Zhukov A, Kovsch A, Kop'ev P, Asryan L, Alferov Z, Ledentsov N, Bimberg D, Kosogov A, Werner P (1998) High-power continuous-wave operation of a InGaAs/AlGaAs quantum dot laser. J Appl Phys 83:5561
68. Michler P, Kiraz A, Becher C, Schoenfeld W, Petroff P, Zhang L, Hu E, Imamoglu A (2000) A quantum dot single-photon turnstile device. Science 290:2282–2285
69. Moore G (1975) Progress in digital integrated electronics. IEDM Tech Digest, pp 11–13
70. Moreau E, Robert I, Manin L, Thierry-Mieg V, Gérard J, Abram I (2001) Quantum cascade of photons in semiconductor quantum dots. Phys Rev Lett 87:183601
71. Naumov M, Lee S, Haley B, Bae H, Clark S, Rahman R, Ryu H, Saied F, Klimeck G (2007) Eigenvalue solvers for atomistic simulations of electronic structureswith NEMO-3D. 12th International Workshop on Computational Electronics, Amherst, USA, 7–10 Oct 2007
72. Oberhuber R, Zandler G, Vogl P (1998) Subband structure and mobility of two–dimensional holes in strained Si/SiGe MOSFET's. Phys Rev B 58:9941–9948
73. Open Science Grid at http://www.opensciencegrid.org
74. Oyafuso F, Klimeck G, Allmen P, Boykin T, Bowen R (2003) Strain effects in large-scale atomistic quantum dot simulations. Phys Stat Sol (b) 239:71
75. Oyafuso F, Klimeck G, Bowen R, Boykin T, Allmen P (2003) Disorder induced broadening in multimillion atom alloyed quantum dot systems. Phys Stat Sol (c) 4:1149
76. Persson A, Larsson M, Steinström S, Ohlsson B, Samuelson L, Wallenberg L (2004) Solid phase diffusion mechanism for GaAs NW growth. Nat Mater 3:677
77. Petroff PM (2003) Single quantum dots: Fundamentals, applications, and new concepts. Springer, Berlin
78. Prada M, Kharche N, Klimeck G (2007) Electronic structure of Si/InAs composite channels. In: MRS Spring conference, Symposium G: Extending Moore's Law with Advanced Channel Materials, San Francisco, 9–13 April 2007
79. Pryor C, Kim J, Wang L, Williamson A, Zunger A (1998) Comparison of two methods for describing the strain profiles in quantum dots. J Apl Phys 83:2548
80. Qiao W, Mclennan M, Kennell R, Ebert D, Klimeck G (2006) Hub-based simulation and graphics hardware accelerated visualization for nanotechnology applications. IEEE Trans Vis Comput Graph 12:1061–1068
81. Rahman A, Klimeck G, Lundstrom M (2005) Novel channel materials for ballistic nanoscale MOSFETs bandstructure effects. In: 2005 IEEE International Electron Devices Meeting, Washington, DC 601-604
82. Rahman R et al (2007) High precision quantum control of single donor spins in silicon. Phys Rev Lett 99:036403
83. Ramdas A et al (1981) Spectroscopy of the solid-state analogues of the hydrogen atom: donors and acceptors in semiconductors. Rep Prog Phys 44
84. Reed M (1993) Quantum dots. Sci Am 268:118
85. Reed M, Randall J, Aggarwal R, Matyi R, Moore T, Wetsel A (1988) Observation of discrete electronic states in a zero-dimensional semiconductor nanostructure. Phys Rev Lett 60:535
86. Sameh A, Tong Z (2000) The trace minimization method for the symmetric generalized eigenvalue problem. J Comp Appl Math 123:155–175
87. Sellier H et al (2006) Transport spectroscopy of a single dopant in a gated silicon nanowire. Phys Rev Lett 97:206805
88. Semiconductor Industry Association (2001) International technology roadmap for semiconductors. (http://public.itrs.net/Files/2001ITRS/Home.htm)
89. Slater J, Koster G (1954) Simplified LCAO method for the periodic potential problem. Phys Rev 94:1498–1524
90. Slater J, Koster G (1954) Simplified LCAO method for the periodic potential problem. Phys Rev 94:1498
91. Stegner A et al (2006) Electrical detection of coherent P spin quantum states. Nat Phys 2:835
92. Stier O, Grundmann M, Bimberg D (1999) Electronic and optical properties of strained quantum dots modeled by 8-band $k.p$ theory. Phys Rev B 59:5688
93. Sze S, May G (2003) Fundamentals of semiconductor fabrication. John Wiley
94. TeraGrid at http://www.teragrid.org
95. Usman M, Ahmed S, Korkusinski M, Heitzinger C, Klimeck G (2006) Strain and electronic structure interactions in realistically scaled quantum dot stacks. In: Proceedings of the 28th International Conference on the Physics of Semiconductors ICPS 2006, Vienna, Austria
96. Vasileska D, Khan H, Ahmed S (2005) Quantum and Coulomb effects in nanodevices. Int J Nanosci 4:305–361
97. Vrijen R et al (2000) Electron-spin-resonance transistors for quantum computing in silicon-germanium heterostructures. Phys Rev A 62:012306
98. Wang J, Rahman A, Ghosh A, Klimeck G, Lundstrom M (2005) Performance evaluation of ballistic silicon nanowire transistors with atomic-basis dispersion relations. Appl Phys Lett 86:093113
99. Wang L, Zunger A (1994) Solving Schrödinger's equation around a desired energy: Application to silicon quantum dots. J Chem Phys 100:2394
100. Wellard C, Hollenberg L (2005) Donor electron wave functions for phosphorus in silicon: Beyond effective-mass theory. Phys Rev B 72:085202
101. Williamson A, Wang L, Zunger A (2000) Theoretical interpretation of the experimental electronic structure of lens-shaped self-assembled InAs/GaAs quantum dots. Phys Rev B 62:12963–12977
102. Welser J, Hoyt J, Gibbons J (1992) NMOS and PMOS transistors fabricated in strained silicon/relaxed silicon-germanium structures. IEDM Tech Dig, pp 1000–1002
103. Wong HS (2002) Beyond the conventional transistor. IBM J Res Dev 46:133–168

104. http://www.intel.com/cd/software/products/asmo-na/eng/307757.htm
105. Zandviet H, Elswijk H (1993) Morphology of monatomic step edges on vicinal Si(001). Phys Rev B 48:14269
106. Zheng Y, Rivas C, Lake R, Alam K, Boykin T, Klimeck G (2005) Electronic properties of Silicon nanowires. IEEE Tran Elec Dev 52:1097–1103
107. Zhirnov VV, Cavin III R K, Hutchby JA, Bourianoff GI (2003) Limits to binary logic switch–a Gedanken model. Proc IEEE 91:1934–1939
108. Zhu W, Han JP, Ma T (2004) Mobility measurement and degradation mechanisms of MOSFETs made with ultrathin high-k dielectrics. IEEE Trans Electron Dev 51:98–105

Books and Reviews

Bimberg D, Grundmann M, Ledentsov N (1999) Quantum dot heterostructures. Wiley
Datta S (2005) Quantum transport: Atom to transistor. Cambridge University Press
Harrison P (2005) Quantum wells, wires and dots: Theoretical and computational physics of semiconductor nanostructures, 2nd edn. Wiley-Interscience

Multiple Mobile Robot Teams, Path Planning and Motion Coordination in

Lynne E. Parker
Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Coupled, Centralized Approaches
Decoupled Approaches
Motion Coordination
Future Directions
Bibliography

Glossary

Autonomous robot An *autonomous robot* is a robot that can perform tasks in unstructured environments with minimal human guidance.

Planned path A *planned path* is a pre-determined, obstacle-free, trajectory that a robot can follow to reach its goal position from its starting position.

Complete path planner A *complete* path planner is an algorithm that is guaranteed to find a path, if one exists.

Deadlocked path planning A *deadlock* is a situation in path planning in which a solution cannot be found, even though one exists. Typically, this is caused by robots blocking each other's paths, and the planner being unable to find a solution in which robots move out of each other's way.

Definition of the Subject

Multi-robot path planning and motion coordination addresses the problem of how teams of autonomous mobile robots can share the same workspace while avoiding interference with each other, and/or while achieving group motion objectives. Nearly all applications of multiple autonomous mobile robots must address this issue of motion coordination, either explicitly or implicitly. Multi-robot path planning and teaming has been extensively studied since the 1980s. While many techniques have been developed to address this challenge, the general centralized multi-robot path planning problem is known to be intractable, meaning that optimal solutions cannot be found in polynomial time. Thus, alternative techniques that decouple aspects of the motion planning and coordination problem have been proposed that trade off optimality for efficiency. A wide variety of applications can benefit from teams of robots that can coordinate their motions effectively, including search and rescue, planetary exploration, mineral mining, transportation, agriculture, industrial maintenance, security and surveillance, and warehouse management.

Introduction

Many practical applications of autonomous robots require the use of multiple team members. Such teams have many potential benefits, including faster task completion time (through parallelism) and increased robustness (through redundancy). Further, teams of robots can increase the application domain of autonomous robots by providing solutions to tasks that are inherently distributed, either in time, space, or functionality. Since the 1980s, researchers have addressed many issues in multi-robot teams, such as control architectures, communication, task allocation, swarm robots, learning, and so forth [83].

A critical issue in these mobile robot teams is coordinating the motions of multiple robots interacting in the same workspace. Regardless of the mission of the robots, they must be able to effectively share the workspace to prevent interference between the team members. Solutions to the motion coordination problem are approached in a variety of ways, depending upon the underlying objectives of the robot team. In some cases, the paths of the robots are explicitly planned and coordinated in advance, as might be needed in a busy warehouse management application, for

example. In other cases, planning is relaxed and emphasis is placed on mechanisms to avoid collision, applicable for tasks such as automated hospital meal deliveries. In yet other situations, the robots could have mechanisms with little pre-planning that focus on coordinating robot motions in real-time using reactive, behavior-based, or control-theoretic approaches, such as would be used in a convoying or formation-keeping application.

The *multi-robot path planning problem* is defined as follows: given a set of m robots in k-dimensional workspace, each with an initial starting configuration (e. g., position and orientation) and a desired goal configuration, determine the path each robot should take to reach its goal, while avoiding collisions with obstacles and other robots in the workspace. More formally (adapting the notation of [58,59]), let $\mathcal{A}$ be a rigid robot in a static workspace $\mathcal{W} = \mathbb{R}^k$, where $k = 2$ or $k = 3$. The workspace is populated with obstacles. A *configuration* $\mathbf{q}$ is a complete specification of the location of every point on the robot geometry. The *configuration space* C represents the set of all the possible configurations of $\mathcal{A}$ with respect to $\mathcal{W}$. Let $\mathcal{O} \subset \mathcal{W}$ represent the region within the workspace populated by obstacles. Let the closed set $\mathcal{A}(\mathbf{q}) \subset \mathcal{W}$ denote the set of points occupied by the robot when it is in the configuration $\mathbf{q} \in C$. Then, the *C-space obstacle region*, C_{obs}, is defined as:

$$C_{\text{obs}} = \{\mathbf{q} \in C | \mathcal{A}(\mathbf{q}) \cap \mathcal{O} \neq \emptyset\} .$$

The set of configurations that avoid collision (called the *free space*) is:

$$C_{\text{free}} = C \setminus C_{\text{obs}} .$$

A *free path* between two obstacle-free configurations c_{init} and c_{goal} is a continuous map:

$$\tau[0, 1] \to C_{\text{free}}$$

such that $\tau(0) = c_{\text{init}}$ and $\tau(1) = c_{\text{goal}}$.

For a team of m robots, define a state space that considers the configurations of all the robots simultaneously:

$$X = C^1 \times C^2 \times \cdots \times C^m .$$

Note that the dimension of X is N, where $N = \sum_{i=1}^{m} \dim(C^i)$. The C-space obstacle region must now be redefined as a combination of the configurations leading to a robot-obstacle collision, together with the configurations leading to robot-robot collision. The subset of X corresponding to robot $\mathcal{A}^i$ in collision with the obstacle region, $\mathcal{O}$, is

$$X^i_{\text{obs}} = \{\mathbf{x} \in X | \mathcal{A}^i(\mathbf{q}^i) \cap \mathcal{O} \neq \emptyset\} . \quad (1)$$

The subset of X corresponding to robot $\mathcal{A}^i$ in collision with robot $\mathcal{A}^j$ is

$$X^{ij}_{\text{obs}} = \{\mathbf{x} \in X | \mathcal{A}^i(\mathbf{q}^i) \cap \mathcal{A}^j(\mathbf{q}^j) \neq \emptyset\} . \quad (2)$$

The obstacle region in X is then defined as the combination of Eqs. (1) and (2), resulting in

$$X_{\text{obs}} = \left(\bigcup_{i=1}^{m} X^i_{\text{obs}}\right) \bigcup \left(\bigcup_{ij, i \neq j} X^{ij}_{\text{obs}}\right) . \quad (3)$$

With these definitions, the planning process for multi-robot systems treats X the same as C, and X_{obs} the same as C_{obs}, where c_{init} represents the starting configurations of all the robots, and c_{goal} represents the desired goal configurations of all the robots.

Typically, optimization criteria guide the choice of a particular solution from an infinite number of possible solutions. Example criteria include minimized total path lengths, minimized time to reach goals, and minimized energy used to reach goals. Additional constraints can introduce more complexity in the planning process, such as navigational restrictions on the robots (e. g., maximum slope restrictions, inability to traverse rocky terrain, etc.), or the need for multiple robots to move in tandem with each other (e. g., a formation of robots moving over uneven terrain). Since the general optimal motion planning problem for multiple moving objects is computationally intractable (specifically, PSPACE-hard [47]), most approaches relax the requirement for global optimality, and instead seek to locally optimize portions of the path planning problem.

Planning approaches can be categorized, or taxonomized, in various ways. One taxonomy evaluates approaches in terms of completeness (i. e., whether they are guaranteed to find a solution if one exists), complexity (i. e., the computational requirements of the search process), and optimality (i. e., the quality of the resulting solution). Often, techniques that are complete and optimal are too computationally intensive to use in practice. Alternatively, techniques that achieve computational tractability typically trade off optimality and/or completeness.

Another taxonomy of multi-robot path planning techniques makes distinctions based on the amount of information used during the planning process. Approaches that use global information and plan directly in X are called *coupled, centralized* approaches. These approaches treat the robot team as a composite robot system, to which classical single-robot path planning algorithms are applied. For example, the A* algorithm [45] can generate complete and optimal solutions to the multi-robot path planning problem under a centralized and coupled approach.

However, this type of planning approach requires computation time that is exponential in the dimension of the multi-robot configuration space. Thus, these approaches can only be used in real-time for the smallest of problem sizes. Sect. "Coupled, Centralized Approaches" describes these coupled, centralized techniques.

To deal with the high-dimensionality of X, alternative approaches *decouple* the path planning problem into independent components that can find good solutions quickly, although at the cost of losing optimality and completeness. These decoupled techniques can either be centralized or decentralized. Common examples of decoupled approaches include those that separate path planning and velocity planning. Typical approaches to decoupled planning will plan individual paths for a robot or set of robots, followed by a second step to resolve any potential conflicts between the paths. Sect. "Decoupled Approaches" describes some common techniques for decoupled multi-robot path planning.

A broader problem in multi-robot teams is that of *motion coordination*. Motion coordination encompasses multi-robot path planning, but also includes other problems such as flocking, formation-keeping, multi-robot target tracking, and other similar objectives. These tasks do not necessarily require advance planning of specific paths for each robot, but do require the coordination of trajectories as the robots move, to avoid collisions with each other, or to reach other group-level objectives, such as maintaining a desired inter-robot distance. Sect. "Motion Coordination" describes some of these techniques. This chapter is concluded with Sect. "Future Directions," which offers remarks on the future directions and impact of multi-robot path planning and motion coordination.

Coupled, Centralized Approaches

In *coupled, centralized* approaches to multi-robot path planning, the robot team is considered to be a composite robot system, to which a classical single-robot path planning algorithm is applied. Motion planning algorithms for single mobile robot systems have been intensively studied for years (see [40,48,58,97]). Examples of classical single-robot path planning algorithms include sampling-based planning, potential-field techniques, and combinatorial methods. Sampling-based planners [54] avoid the explicit construction of C_{obs} by sampling different configurations to generate curves that represent collision-free paths in C_{free}. Potential field techniques (e. g., [9,10,114]) construct real-valued functions that pull the robot toward the goal, and repulse the robot away from obstacles, via a combination of force vector fields. Combinatorial methods construct roadmaps through the configuration space using techniques such as cell decomposition (e. g., [75,100]).

In an environment that contains a set of stationary obstacles, single robot path planning methods such as graph searching based on a geometric configuration of the environment are guaranteed to return optimal paths (in the sense of a performance measure such as shortest distance) in polynomial time if one exists. However, motion planning in a dynamic environment with moving obstacles is inherently harder. Even for a simple case in two dimensions, the problem is PSPACE-hard and is not solvable in polynomial time [35,47]. Motion planning in dynamic environments was originally addressed by adding the time dimension to the robot's configuration space. The approach in [29] discretizes the configuration-time space to a sequence of slices of the configuration space at successive time intervals, representing the motions of moving obstacles using the set of slices embodying space-time. In [79], moving obstacles are represented as sheared cylinders, and a methodology was proposed to provide optimal tangent paths to the goal for a dynamic robot environment.

Extending the problem still further, to multiple robot path planning, requires even more computational resources. An example centralized approach for generating complete multi-robot path solutions is the work of Parsons and Canny [85], which takes a global cell decomposition approach, incorporating obstacles and other robots in a unified configuration space representation. This algorithm first computes a decomposition of the free space into cells; it then searches through the resulting adjacency graph for a path. However, not surprisingly, the algorithm is exponential in the number of robots. Other centralized algorithms that represent the path planning problem as a cross product of the configuration spaces of the individual robots include [9,96].

Because of the high dimension of the multi-robot configuration space, centralized approaches that treat the multi-robot team as a single composite robot tend to be impractical computationally if the full search space is used. Instead, techniques that reduce the size of the search space have been shown to be practical for small-sized problems. One way to reduce the search space is to weakly constrain the allowable paths that robots can follow by limiting the motion of the robots to lie on *roadmaps* in the environment. Intuitively, roadmaps are akin to automotive highways, where robots move from their starting position to a roadmap, move along the roadmap to the proximity of the goal, and then move off the roadmap to the specific goal location. More formally, a roadmap is defined as follows [24]:

Definition 1 (Roadmap) A union of one-dimensional curves is a **roadmap** RM if for all q_{start} and q_{goal} in C_{free} that can be connected by a path, the following properties hold:

1. **Accessibility:** there exists a path from $q_{\text{start}} \in C_{\text{free}}$ to some $q'_{\text{start}} \in RM$,
2. **Departability:** there exists a path from $q'_{\text{goal}} \in RM$ to $q_{\text{goal}} \in C_{\text{free}}$, and
3. **Connectivity:** there exists a path in RM between q'_{start} and q'_{goal}.

Typically, a roadmap RM is represented as a graph $G = (V, E)$, in which the nodes V represent collision-free configurations, and the edges E represent feasible paths. (A *feasible* path is one that can be executed by robot $\mathcal{A}^i$, based on its physical motion constraints.) Various algorithms have been created that make use of the roadmap concept for motion planning, both for single robots and for multi-robot teams (e. g., [87,93,109]). The following subsections present two such approaches for multi-robot teams. The first, in work by Švestka and Overmars, is a probabilistically complete approach, meaning that the problem is solvable in finite time. Their approach creates a coordinated path for a composite robot by making use of the concept of *super-graphs*. The second, in work by Peasgood, et al., [87], is a multi-phase approach that uses a graph and spanning tree representation to create paths through the environment. This approach is shown to have linear-time complexity, and is thus scalable to much larger robot teams.

Before presenting these two approaches, it is worth noting that many other roadmapping approaches to multi-robot path planning have been proposed. For example, the work of Ryan [93] reduces the search space by decomposing the original map into subgraphs, planning paths between subgraphs, and then coordinating motions within the subgraphs. This approach has been shown to be effective for up to 10 robots. In [26], Clark, et al., introduce the concept of *dynamic networks*, which are formed between robots that are within communication range. Within this framework, only robots within the same network use a centralized planner, which is based upon probabilistic road maps [54]; otherwise, robots plan their paths using decoupled planners based on optimizing priorities (see Sect. "Decoupled Approaches"). In [95], efficiencies in the probabilistic road map are achieved by delaying collision checking along the roadmaps until necessary. The speed-up achieved by this collision-checking (on the order of a factor of 4 to 40) allows this technique to be used more practically for small-sized multi-robot teams. The authors incorporate this improved planning process into three multi-robot path planning variants: a centralized version, a decoupled planner with global coordination, and a decoupled planner with pair-wise coordination.

Super-Graph Method (Švestka and Overmars)

In [109], Švestka and Overmars present an approach for creating a composite roadmap, which represents a network of feasible motions for the composite robot. This composite roadmap is created as follows. First, a roadmap for each individual robot is constructed using the standard roadmap generation algorithm, Probabilistic Path Planner (PPP) [54]. Then, n such roadmaps are combined into a roadmap for the composite robot, which can be used to generate coordinated paths.

Specifically, the *coordinated path* for the composite robot $(\mathcal{A}^1, \ldots, \mathcal{A}^n)$ is an n-tuple of paths feasible for all robots $\mathcal{A}^i$ that, when executed simultaneously, introduce no mutual collisions between the individual robots. Formally, let $C^{[0,1]}$ represent the configuration space from time $t = 0$ to time $t = 1$, where the robot is at its starting position at time 0, and is at its goal location at time 1. Let $s_1, \ldots, s_n$ and $g_1, \ldots, g_n$ be given starting and goal configurations for the n robots, where $\forall i \in \{1, \ldots, n\}\colon s_i \in C_{\text{free}} \wedge g_i \in C_{\text{free}}$. Let P represent a free path if P is in C_{free} for all times t (i. e., $\forall t \in [0, 1]\colon P(t) \in C_{\text{free}}$). Let $A \cap B \neq \emptyset$ (i. e., A and B intersect) be represented by $A \otimes B$. Then if $P_1, \ldots, P_n \in C^{[0,1]}$ are feasible paths, such that for all $i, j \in \{1, \ldots, n\}$

- $P_i(0) = s_i \wedge P_i(1) = g_i$
- $i \neq j \Rightarrow \forall t \in [0, 1]\colon \neg\mathcal{A}(P_i(t)) \otimes \mathcal{A}(P_j(t))$

then $(P_1, \ldots P_n)$ is a coordinated path for $(\mathcal{A}^1, \ldots \mathcal{A}^n)$ solving the problem $((s_1, \ldots, s_n), (g_1, \ldots, g_n))$.

Švestka and Overmars present an approach for constructing such a coordinated path for a composite robot [109]. The basic idea is to seek paths along the roadmap, G, that allow the robots to move from their starting to their goal configurations, while disallowing simultaneous motions or motions along paths that are blocked by other robots. This type of path is called a *G-discretized coordinated path*. They introduce the concept of *super-graphs*, which represent roadmaps for the composite robots created by combining n simple robot roadmaps. Two variants of super-graphs are proposed – *flat super-graphs* and *multi-level super-graphs*.

In the flat super-graph, a node represents a feasible placement of the n simple robots at the nodes of G, and an edge represents a motion of exactly one simple robot

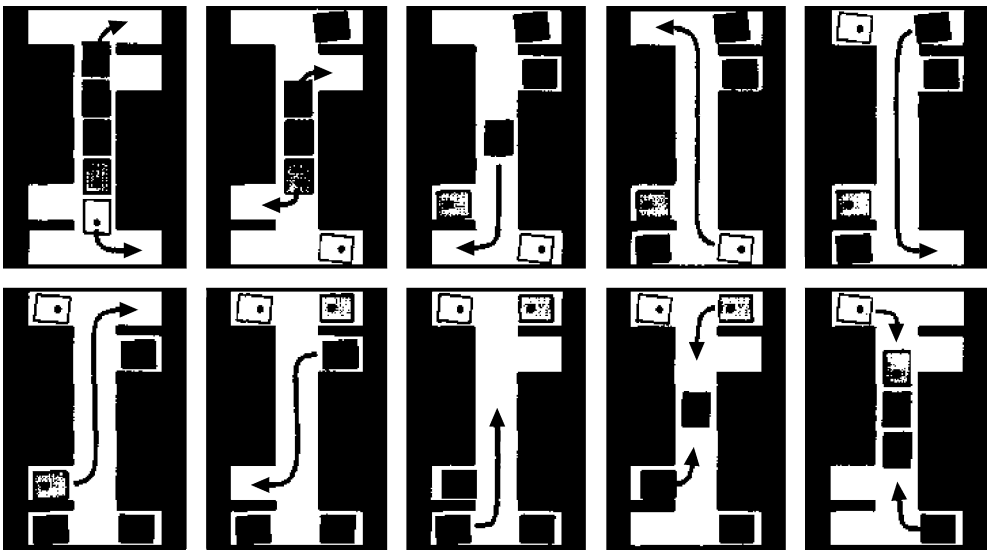

Multiple Mobile Robot Teams, Path Planning and Motion Coordination in, Figure 1
An illustration of a coordinated path generated by the super-graph approach of Švestka and Overmars, for 5 nonholonomic car-like robots (from [109])

along a non-blocked path of *G*. A disadvantage of the flat super-graph is that its size is exponential in the number of robots.

The second type of super-graph – the multi-level super-graph – reduces the size of the super-graph data structure by combining multiple nodes into a single node of the graph. This approach makes use of the concept of *subgraphs*. Whereas the nodes in a flat super-graph represent robots being located at particular nodes of *G*, the nodes in a multi-level super-graph represent robots being located in a subgraph of G. The restriction placed on node combinations is that the resultant subgraphs should not interfere with each other, meaning that the nodes in one subgraph cannot block paths in another subgraph. Experimental results have shown that the multi-level super-graphs are typically much smaller than the equivalent flat super-graphs.

Švestka and Overmars applied this approach to teams of up to 5 nonholonomic, car-like robots in simulation. An example of these results is shown in Fig. 1, illustrating the feasibility of this approach for small-sized multi-robot teams. Nevertheless, this type of approach is appropriate only for relatively small numbers of robots. For much larger sizes of robot teams, decoupled approaches are necessary (see Sect. "Decoupled Approaches").

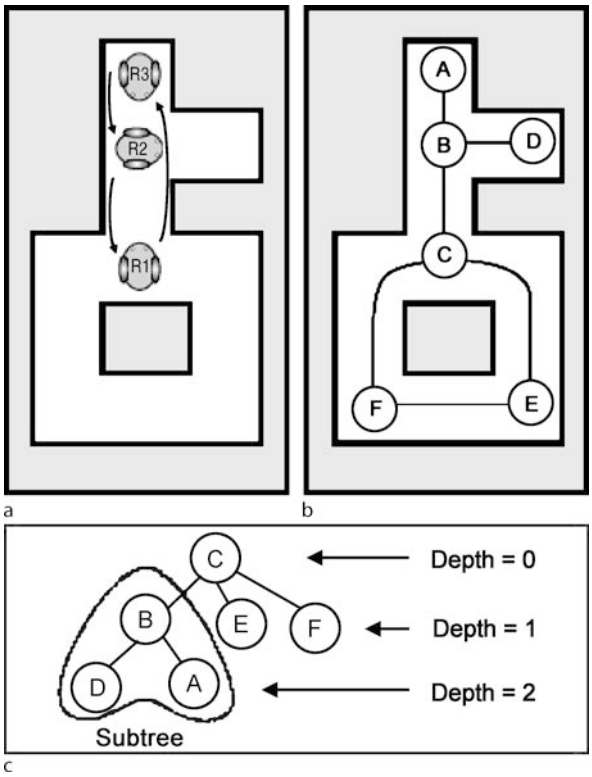

Multiple Mobile Robot Teams, Path Planning and Motion Coordination in, Figure 2
An example multi-robot path planning problem using the spanning tree method of Peasgood, et al., along with the corresponding graph and spanning tree (from [87])

Spanning Tree Method (Peasgood, et al.)

Peasgood, et al., [87] present another roadmap-based planner for multi-robot teams. This approach is a multi-

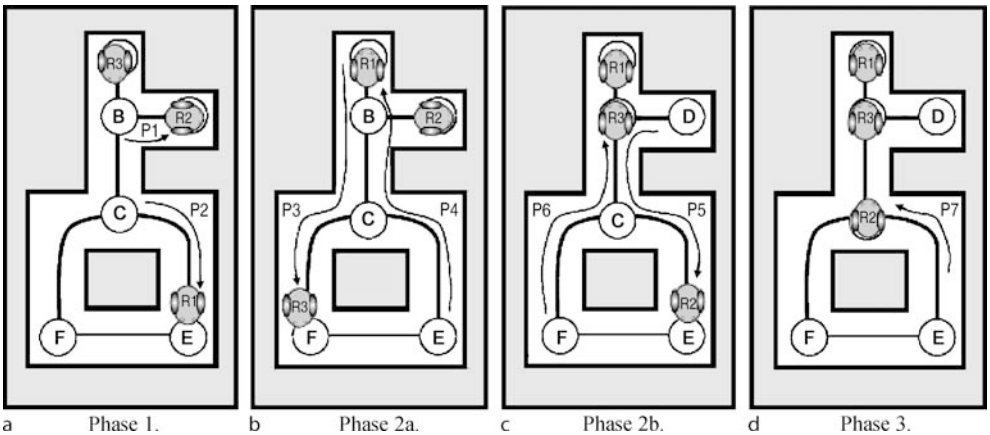

Multiple Mobile Robot Teams, Path Planning and Motion Coordination in, Figure 3
The multi-phase solution of the multi-robot path planning problem in Fig. 2, using the spanning tree method of Peasgood, et al. (from [87])

phase planner that uses a graph and spanning tree representation to create and maintain obstacle-free paths through the environment. Initially, a graph is created, in which the nodes are the robots' initial and goal positions, and the edges represent the connectivity of the node positions. An example is illustrated in Fig. 2a, in which the starting positions of the three robots (R1, R2, and R3) are (C, B, A), while the goal positions are (A, C, B). Figure 2b shows the graph-based map for this example. Then, a spanning tree of this graph is created, which is a connected subset of the original graph that includes all the nodes without cycles; Fig. 2c shows the example spanning tree. The root of this spanning tree is chosen to be the node that is closest to the geographic center of the map. Then, in the first phase of the approach, a plan is generated that moves the robots to the leaves of the spanning tree along collision-free paths, as shown in Fig. 3a. In the second phase, the robots are moved into positions where they can reach their goals without creating obstructions for other robots. This is accomplished by processing the robots in order according to the depth of their goals in the spanning tree. This is shown in Fig. 3b,c. The third phase moves robots to the remaining unfilled goal locations, as shown in Fig. 3d. These three phases result in a sequence of motions that allow only one robot to move at a time. The final phase of the process seeks to improve the quality of the concurrent plan by allowing robots to move simultaneously when doing so does not introduce any collisions.

Peasgood, et al., show that this algorithmic approach results in time complexity that is linear in the number of robots. To further improve the resulting path lengths, the authors propose a hybrid planning approach, which uses the regular multi-phase planner, but then also calls a decoupled planner (such as [15]), to attempt to find shorter path solutions. For smaller-sized robot teams (less than 20), the decoupled planner can often find better solutions. However, for larger-sized teams, the multi-phase approach is more time-efficient (increasingly so as the team size grows larger).

Decoupled Approaches

Decoupled approaches to multi-robot path planning typically trade off solution quality for efficiency by solving some aspects of the problem independently. There are many alternative ways of decomposing the planning problem. Most commonly, approaches plan individual paths for robots, followed by methods for handling collision avoidance. While decoupled approaches are typically more efficient than centralized approaches, they lose completeness. For instance, Fig. 4 shows an example of a situation that is difficult for decoupled approaches to solve. In this situation, robots must exchange positions in a narrow corridor. While a centralized approach would find a solution in which the robots first move into the open space at the end of the corridor to exchange places, a decoupled approach will have difficulties discovering this solution.

Decoupled approaches are typically divided into two broad categories [58,59]: prioritized planning and path coordination. *Prioritized planning* considers the motions of the robots one at a time, in priority order, calculating path information for the ith robot by treating the previous $i-1$ robots as moving obstacles. *Path coordination*, on the other hand, first plans independent paths for the robots separately, then seeks to plan their velocities so as

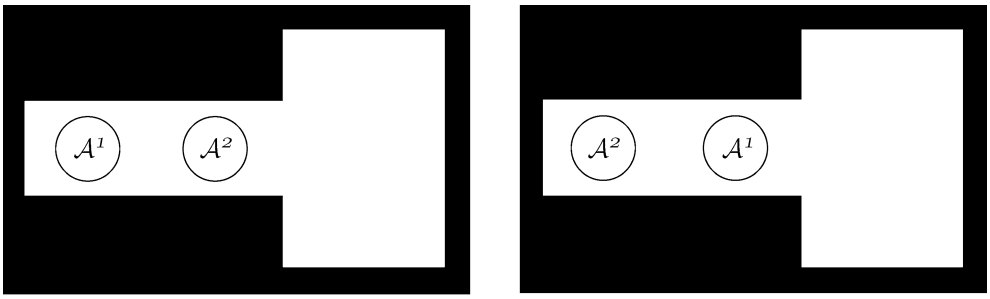

Multiple Mobile Robot Teams, Path Planning and Motion Coordination in, Figure 4
An example multi-robot path planning problem that is difficult for decoupled approaches to solve. Here, the robots must exchange positions by first moving into the open space at the end of the corridor. While a centralized approach can find this solution, most decoupled approaches would fail (recreated from [58])

to avoid collisions along those paths. The following subsections describe these approaches in more detail.

Prioritized Planning

The *prioritized planning* approach to multi-robot path planning was first proposed by Erdmann and Lozano–Peréz [29]. In this approach, priorities are assigned to each robot. These priorities could be assigned randomly, or they could be determined from motion constraints, in which more-constrained robots are given higher priority. A path is planned for the first robot using any single-robot path planning approach. The path for each successive robot, $\mathcal{A}^i$, then takes into account the plans for the previous robots $\mathcal{A}^1, \ldots, \mathcal{A}^{i-1}$, treating these higher-priority robots as moving obstacles.

More specifically, in the prioritized planning approach of [29], the configuration space is extended to account for time, since the time-varying motions of previously-planned robots must be taken into account. Configuration space-time is represented as a list of configuration space slices at particular times – specifically, those times corresponding to when a moving object changes its velocity. Motions between slices can then be interpolated via straight-line translations between these configuration space slices. The configuration space-time can be constructed in $O(m)$ time, where $m = nr$, for n edges in the environment and r time slices.

Paths through configuration space-time are computed using a visibility graph algorithm, which searches along a visibility graph consisting of the vertices of the configuration space obstacles (plus vertices for the start and goal positions), and the line-of-sight edges between the vertices. Planners using this method have time complexity $O(rn^3)$, although [29] also suggests a faster implementation. The prioritized planning approach has been demonstrated in several application domains, including the translation of multiple planar robots, as well as the motion of two-link planar articulated robot arms.

Other researchers who have studied prioritized path planning for multiple mobile robots include [16,20,32,121]. Both Ferrari, et al. [32] and Warren [121] used a fixed priority scheme for the decoupled planner. In the work of Buckley [20], a heuristic is applied to assign higher priorities to robots that can move in a straight line to their target location. Chun, et al. [25] use this priority scheme to coordinate independently-generated schedules online, as the conflicts arise. The work of Azarm and Schmidt [6] considers all possible priority assignments, although the resulting approach is computationally complex. A more tractable method for finding and optimizing priority schemes for decoupled priority-based planners is presented by Bennewitz, et al., in [16]. The proposed approach performs a centralized, randomized search with hill-climbing (i. e., the A* search algorithm [73]) to search the space of prioritization schemes to find priority schemes that minimize the overall path length. The resulting priority scheme can then be applied in decoupled priority-based planners, such as Erdmann's method described above [29].

The advantage of prioritized planning approaches is that they reduce the problem from a single planning problem in a very high-dimensional space to a sequence of planning problems in much lower dimensional space. The disadvantage, as with all decoupled approaches, is that these approaches are not complete.

Path Coordination

Path coordination techniques decouple the planning problem into path planning and velocity planning (e. g., [52]). The ideas are based on scheduling techniques for deal-

ing with limited resources, inspired by the approaches developed for concurrent access to a database by multiple users [124]. In the current context, the shared resource is space. The decomposition of path and velocity planning provides a solution through the complexity barrier caused by the additional time dimension, and also provides solutions that are relevant when robots move along fixed paths.

In the path coordination approach, the path planning step first generates individual robot paths independently, using common single-robot path planners. The second step plans a velocity profile that each robot should follow along its path so as to avoid collisions with other robots. This approach is typically called *fixed-path* coordination, since the paths planned in the first step are not altered in the second step. Instead, only the velocities taken by the robots along the paths are varied.

In more detail (using the notation of [59]), assume that the path generated for each individual robot in the first step constrains robot $\mathcal{A}^i$ to follow a path $\tau_i : [0,1] \to C^i_{\text{free}}$. Then, an m-dimensional *coordination diagram* $X = [0,1]^m$ for m robots is defined that is used to schedule the motions along their paths so that they do not collide [74]. In this diagram, the ith coordinate represents the domain, $S_i = [0,1]$, of the path of robot $\mathcal{A}^i$. At state $(0,\ldots,0) \in X$, every robot is in its initial starting configuration. At state $(1,\ldots,1) \in X$, every robot is at its goal configuration. Within the coordination diagram, obstacles form obstacle regions X_{obs} that must be avoided. Any continuous, obstacle-free path, $h\colon [0,1] \to X$, for which $h(0) = (0,\ldots,0)$ and $h(1) = (1,\ldots,1)$, is a valid path that moves the robots from their starting positions to their goals. The objective, therefore, is to find $h\colon [0,1] \to X_{\text{free}}$, in which $X_{\text{free}} = X \setminus X_{\text{obs}}$. An example coordination diagram showing a valid path for the robots is illustrated in Fig. 5.

Several authors have looked at variations of the path coordination approach. In [62], Lee and Lee use a similar idea to plan the motions of two robots. Griswold and Eem [41] take uncertainty of the moving obstacles into account while using the same principle for path planning. Pan and Luo [77] use the concept of *traversability vectors* to analyze the spatial relationship between the robot and moving obstacles, and develop a search algorithm to coordinate the robot motion. Rude [92] proposes a space-time representation for collision avoidance in preplanned individual robot paths. In [43], Guo and Parker present a decentralized path coordination approach that also incorporates optimization issues into the planning, including a global performance measurement to minimize the weighted sum of the most expensive time to reach the goals and all idle time, as well as individual optimization goals for navigation over rough terrain. In [60], LaValle and Hutchinson consider multiple robots with independent goals and performance measures, and proposes algorithms optimizing a scalarizing function that is a weighted-average of individual performance functions. Other approaches to optimal motion planning have been proposed in [17,18,23,60,88,101], sometimes in the context of robotic manipulator motion planning. In [22], one robot is randomly chosen to stop, and time delays are inserted to resolve potential collisions. Path coordination schedules, which are another form of velocity planning, are studied in [17,74,102]. A priority-based method using collision maps is presented in [78]. Extensions of the path coordination approach to coordination on roadmaps have been proposed by [39,60].

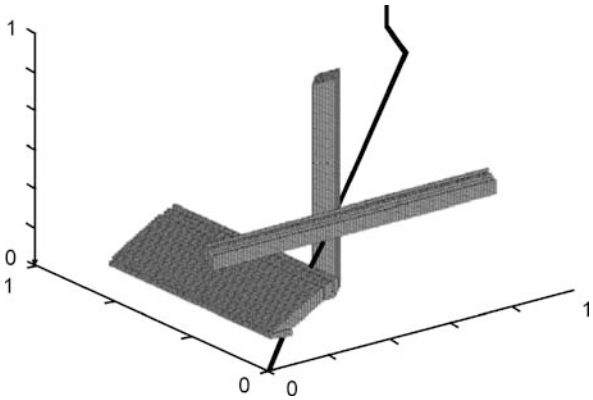

Multiple Mobile Robot Teams, Path Planning and Motion Coordination in, Figure 5
An example coordination diagram for three robots. Each axis represents the domain of an individual robot's path. The cylindrical objects are obstacles, and the path from (0, 0, 0) to (1, 1, 1) represents the coordinated velocity plans for moving all three robots to their goals without collisions (adapted from [43])

While all of these decoupled approaches typically allow good solutions to the multi-robot path planning problem, they can lead to deadlocks, in which solutions cannot be found, even though they exist. In these cases, it may be possible to make use of a centralized planner for small portions of the original problem, in order to solve the immediate deadlock problem.

Motion Coordination

Closely related to the topic of multi-robot path planning is the issue of multi-robot *motion coordination*. Unlike multi-robot *path planning* or *path coordination* approaches, which plan and/or coordinate the complete paths of all of the robots in advance, techniques for *motion*

coordination focus on decentralized, online approaches that allow robots to avoid and/or resolve conflict as the situation arises during path execution, such as through the use of *traffic control* rules. In traffic control applications, individual robots still have independent starting and goal positions, and must move so as to avoid conflict with each other. Even broader concepts of motion coordination seek to have the robots move according to some constraints on the team as a whole, such as can be seen in *formation keeping*, *flocking/swarming*, *target search/tracking*, *dispersion/aggregation*, and related topics. In these problems, the motions of individual robots are no longer independent of each other; instead, the group must move in synchrony according to pre-defined motion constraints for the entire team. The following subsections discuss some of the key research in these areas of motion coordination.

Traffic Control

Traffic control approaches to multi-robot motion coordination typically predefine traffic or control rules that robots must obey as they move through the workspace. Individual robots often move along paths to their goals that they pre-plan in advance, based only on the individual robot goals. Then, as regions involving shared resources are reached (such as the space in an intersection), robots follow the traffic or control rules to coordinate their motions with other robots who also need access to the shared resources.

An early example of traffic control is the work of Grossman [42], which addresses the motion of large numbers of Automatic Guided Vehicles (AGVs) in a factory. Grossman defines three types of control possibilities: 1) restrict the roads so that there is a unique route between all starting and goal positions; 2) allow AGVs to select their own routes autonomously; and 3) control all AGVs' paths using centralized traffic control. Grossman shows that allowing AGVs to select their routes autonomously (option 2) is preferred over the highly suboptimal restriction of roads (option 1). Of course, as previously noted, the centralized approach (option 3) has high combinatorial complexity.

The problem of the autonomous coordination of paths (option 2) is formulated as follows. A set of r AGVs are allowed to follow unconstrained paths in two dimensions, on a *grid-iron* network of roadways, with n parallel roads along each axis. Each section of roadway between intersections is called an *arc*; in this formulation, there are $2n(n-1)$ arcs in the network. Each intersection of roadways is called a *node*, representing the locations of machine tools to be serviced by the robots. It is assumed that $1 \leq r \leq n^2 - 1$, and that all vehicles move at the same speed, v. Each AGV has the task of moving from a source location (i. e., starting position) to a sink location (i. e., a goal location). Defining S to be the average number of time steps per task for each AGV, the average throughput of all the AGVs together is $W = (vr)/S$. This throughput must exactly match the throughput of all the n^2 machine tools, leading to a requirement that the AGV speed must satisfy: $v = (Sn^2)/r$. The price of r AGVs is considered negligible in comparison to the price of the machine tools. Thus, the problem is formulated as the problem of optimizing the traffic control and the value of r so as to minimize v in an $n \times n$ grid-iron floor plan. The constraints on the traffic in this environment are as follows:

- At the end of each step, at most one AGV may be at each node.
- During each step, no two AGVs may pass on the same arc.
- All AGVs have equal priority.

Different policies are investigated, including a greedy policy and a benevolent policy. Simulation results show that the benevolent policy performs the best, with a performance close to the derived lower bound. This traffic policy requires the AGVs to follow these rules:

1. From the AGVs own (i, j) location, determine in which quadrant q the goal node (i', j') lies:
   - Quadrant 1 has $i' > i$ and $j' \geq j$.
   - Quadrant 2 has $j' > j$ and $i' \leq i$.
   - Quadrant 3 has $i' < i$ and $j' \leq j$.
   - Quadrant 4 has $j' < j$ and $i' \geq i$.
2. Depending on the value of q, try to move to an adjacent node:
   - If q is 1 then $(i+1, j)$.
   - If q is 2 then $(i, j+1)$.
   - If q is 3 then $(i-1, j)$.
   - If q is 4 then $(i, j-1)$.
3. If that node is blocked, add 1 to q and try Step 2 again.
4. If that node is blocked, add 1 to q and try Step 2 again.
5. If that node is blocked, add 1 to q and try Step 2 again.
6. If all adjacent nodes are blocked, then wait at the current node.

This policy leads to an overall counterclockwise flow of traffic through the workspace. Based on analysis and simulation results, the authors conjecture that this policy is the optimal policy for AGVs without memory or task trading.

There are many variants on the traffic control and conflict resolution theme [5,53,65,117,118,125]. For example, in [53], Kato, et al., categorize the traffic rules into three types: 1) traffic rules to be applied to the current positions

of the robot (examples include *passage zone*, *stop*, *slow*); 2) traffic rules to be applied to current positions and conditions (examples include *overtaking*, *avoiding obstacles*, *crossing intersections*); and, 3) traffic rules to ensure safety in case of accidents or failures. These rules are illustrated for robot teams operating in indoor hallway-types of settings.

In [5], Asama, et al., propose two basic rules for avoiding collisions:

- "If the colliding robot is nearby to the front and approaching, then avoid from the left", and
- "If the colliding robot is nearby to the front and leaving, then stop for a while".

These rules are combined with a communication-based negotiation process that resolves conflicts by setting priorities based on the task requirements, the environmental situation, and robot performances. In the work of Yuta and Premvuti [125], robots move along pre-planned paths in network of roadways, which can involve conflicts at intersections. These deadlock situations at intersections are resolved through a "shunting" process, in which one robot, acting as a leader, devises a solution for moving robots through the intersection, and then broadcasts the instructions to the other robots for how to resolve the conflict. Another approach to conflict resolution is to use techniques from distributed computing, as illustrated in the work of Wang [117,118], in which robots use a mutual exclusion protocol to compete for the right to move along certain pathways or to resolve conflicts at intersections.

In [67], Lumelsky, et al., present a decentralized approach for motion planning that has robots plan and execute their paths "on the fly" in real time, resolving conflicts as they arise. The authors make an analogy to human cocktail parties, in which people do not plan optimal paths in advance, nor consult with others about their intended destinations; instead, they move toward their destinations while avoiding collisions as they go. Their approach is based on maze-searching techniques, and makes use of perpendicular bisectors and Voronoi diagrams [90] to allow robots to avoid collisions.

Another approach that is closely related to the decoupled path coordination research described earlier is the work of Alami, et al. [1,2]. This online *plan-merging* paradigm does not require advance planning of all robot paths in advance. Instead, robots move as needed, coordinating their paths as new goal destinations arise. In this decentralized planning approach, robots also treat segments of their paths as shared resources. However, when a robot elaborates a new plan for itself, it must validate that plan within the current multi-robot context. This is done by collecting the plans from all the other robot team members via communication, and "merging" its own plan into the existing robot plans. This merging operation is done without affecting the plans of other robots, thus allowing them to continue on with their current executions. In this approach, the environment is represented as a topological graph of areas, routes, and crossings. Routes are composed of lanes with direction, thus setting up a type of traffic pattern through the environment. The motion planning approach makes use of a graph searching technique, planning dependency graphs, and synchronization points to coordinate the motions of the robots. Figure 6 illustrates the geometrical and topological planning space for this approach in a prototypical application.

More recent work in conflict resolution for multi-robot teams is the work of Pallottino, et al., [76], which considers a more realistic kinematic model of the robot dynamics, recognizing that most robots cannot stop instantly in order to avoid collisions. This model focuses on large numbers of robots (e. g., 70) operating closely in shared, open spaces. As with other techniques discussed to this point, this approach also assumes robots have independent starting positions and goal destinations. This approach is particularly relevant for applications of aerial vehicles flying at constant altitude. This work makes use of the concept of *reserved region*, which is an area for which a robot claims exclusive ownership. The control policy is defined for a set of discrete modes of operation, including a *hold* state in which a robot is stopped, a *straight* state in which the robot is moving forward without turning, and two *roll* states – one for mild turns and a second for tight turns. Control theoretic definitions of the motions of the robots in each state are given, and the policy is shown to be *safe*, meaning that it guarantees collision avoidance. Under certain conditions, the approach is also shown to have the property of *liveness*, meaning that all the robots are guaranteed to reach their destinations in finite time.

Reactive Approaches

Reactive-style methods for coordination are useful in many applications, since they are fast, and can operate well in real-time. One common reactive method makes use of *potential fields* [55]. In the potential field approach, the robot moves through space as if it is being acted upon by a set of forces. Attractive forces pull the robot toward a goal destination, while repulsive forces push the robot away from obstacles and/or other robots. At each point in the configuration space, the robot moves along the vector representing the combined forces acting on that point in the configuration space. These concepts have been applied

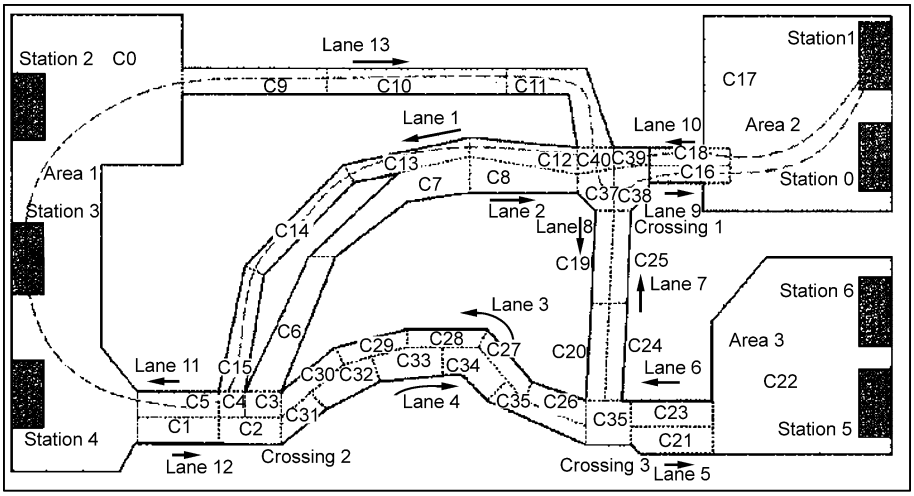

a Geometrical representation.

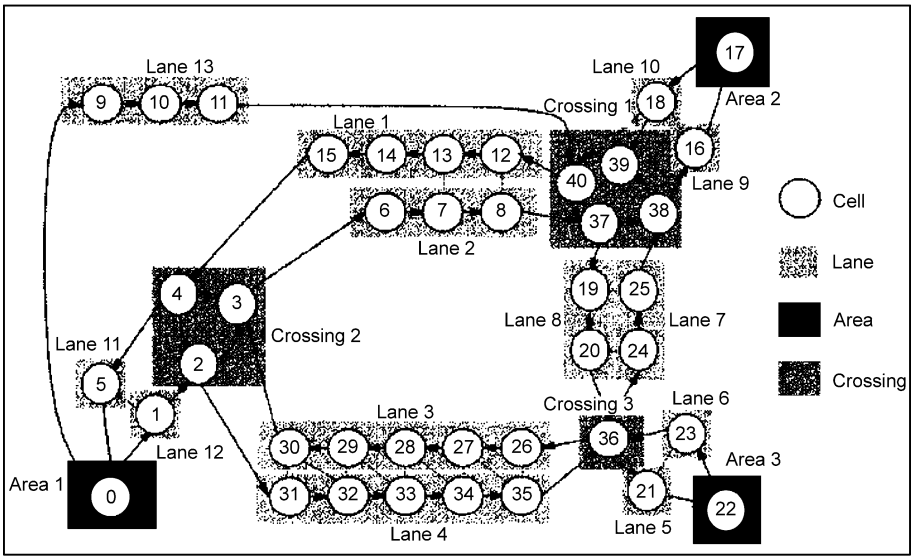

b Topological representation.

Multiple Mobile Robot Teams, Path Planning and Motion Coordination in, Figure 6
Representations used for a prototypical application of the plan-merging paradigm of Alami, et al. (from [2])

to various multi-robot applications [121], including multi-robot soccer [63]. Other potential field approaches to multi-robot coordination include [27,64,119,120]. A well-known issue in potential field methods, however, is their susceptibility to deadlock due to local minima in the potential field. Some techniques have been designed to overcome this shortcoming [11].

Other reactive approaches for collision avoidance based on local information include the work of Matarić [70], which proposes behavior-based avoidance rules in which robots either stop for a period of time or change directions. Similar rules were proposed by Arkin [4] and by Sugihara and Suzuki [107]. Shan and Hasegawa [99] present behavior-based techniques for avoiding robot collisions in narrow passages.

While all of the above techniques can work well for relatively unconstrained situations, they are not analyzed formally to provide guidance for setting the navigation

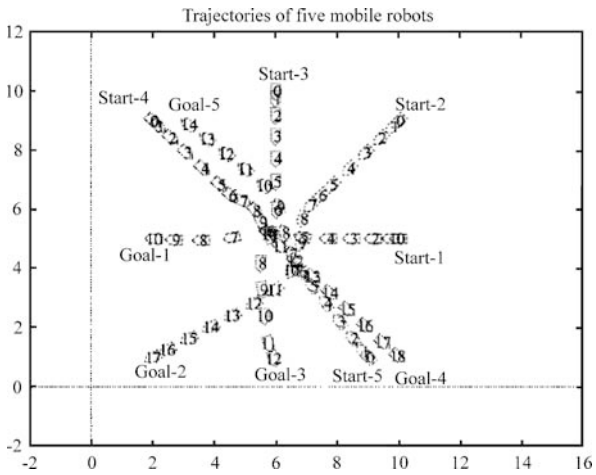

Multiple Mobile Robot Teams, Path Planning and Motion Coordination in, Figure 7
Example results for 5 robots in simulation performing adaptive navigation to avoid collisions, using the approach of Fujimori, et al. (from [34])

Multiple Mobile Robot Teams, Path Planning and Motion Coordination in, Table 1
Categories of swarm behaviors

Relative motion requirements	Swarm Behaviors
Relative to other robots	Formations [80,107], Flocking, Natural herding (as in herds of cattle), Schooling, Sorting [13], Clumping [13], Condensation, Aggregation, Dispersion
Relative to the environment	Search [36], Foraging [7], Grazing, Harvesting, Deployment, Coverage, Localization, Mapping, Exploration
Relative to external agents	Pursuit, Predator-prey, Target tracking, Forced herding/shepherding (as in shepherding sheep)
Relative to other robots and the environment	Containment, Orbiting, Surrounding, Perimeter search
Relative to other robots, external agents, and the environment	Evasion, Tactical overwatch, Soccer [19,104,115,122]

parameters. On the other hand, a more formal method for determining reactive collision avoidance parameters is given by Fujimori, et al., in [34]. These authors propose a collision avoidance method based on an *adaptive navigation technique*, in which the navigation law is given by a first-order differential equation. Navigation of the robot to the goal and obstacle avoidance are handled by switching the direction angle adaptively. Robots are assigned priorities to determine which vehicles must yield to the others. The proper value of the direction angle is calculated theoretically, based on three robot modes of operation: *navigation mode*, in which the robot is moving toward the goal without interference; *cooperative avoidance mode*, in which the robot avoids other robots; and, *final mode*, when the robot is approaching near the goal. The approach has been implemented in simulation for up to five mobile robots, and on two physical robots. Figure 7 illustrates the type of motions generated by this approach in a five-robot simulation.

Coordinated Motion of Entire Team

A significant topic of current research is the control of robot motions to achieve a group objective, such as maintaining a formation while moving to a goal position, cooperatively tracking moving targets, collective coverage tasks, and so forth. Often, these topics are studied in the context of *swarm* robot systems, involving large numbers of homogeneous robots performing the same control algorithms. A complete survey of all the work in these areas is beyond the scope of this chapter. However, this section briefly outlines the areas of active research in this domain.

Many types of swarm behaviors have been studied, such as foraging, flocking, chaining, search, herding, aggregation, and containment. The majority of these swarm behaviors deal with spatially distributed multi-robot motions, requiring robots to coordinate motions either (1) relative to other robots, (2) relative to the environment, (3) relative to external agents, (4) relative to robots and the environment, or (5) relative to all (i. e., other robots, external agents, and the environment). Table 1 categorizes swarm robot behaviors according to these groupings (see also [83]).

Much of the current research in swarm robotics is aimed at developing specific solutions to one or more of the swarm behaviors listed in Table 1. Some of these swarm behaviors have received particular attention, notably formations, flocking, search, coverage, and foraging. In general, most current work in the development of swarm behaviors is aimed at understanding the formal control theoretic principles that can predictably converge to the desired group behaviors, and remain in stable states. The following subsections outline research in some of these areas.

Flocking and Formations Coordinating the motions of robots relative to each other has been a topic of interest in multiple mobile robot systems since the inception of

the field. In particular, much attention has been paid to the flocking and formation control problems. The flocking problem can be viewed as a subcase of the formation control problem, requiring robots to move together along some path in the aggregate, but with only minimal requirements for paths taken by specific robots. Formations are more strict, requiring robots to maintain certain relative positions as they move through the environment. In these problems, robots are assumed to have only minimal sensing, computation, effector, and communications capabilities. A key question in both flocking and formation control research is determining the design of local control laws for each robot that generate the desired emergent collective behavior. Other issues include how robots cooperatively localize themselves to achieve formation control (e. g., [71,72]), and how paths can be planned for permutation-invariant multi-robot formations (e. g., [56]).

Early solutions to the flocking problem in artificial agents were generated by Reynolds [91] using a rule-based approach. Similar behavior- or rule-based approaches have been used physical robot demonstrations and studies, such as in [8,70]. These earlier solutions were based on human-generated local control rules that were demonstrated to work in practice. More recent work is based on control theoretic principles, with a focus on proving stability and convergence properties in multi-robot team behaviors. Examples of this work include [3,14,31,37,38,49,68,110,113].

Foraging and Coverage Foraging is a popular testing application for multi-robot systems, particularly for those approaches that address swarm robotics, involving very large numbers of mobile robots. In the foraging domain, objects such as pucks or simulated food pellets are distributed across the planar terrain, and robots are tasked with collecting the objects and delivering them to one or more gathering locations, such as a home base. Foraging lends itself to the study of weakly cooperative robot systems, in that the actions of individual robots do not have to be tightly synchronized with each other. This task has traditionally been of interest in multi-robot systems because of its close analogy to the biological systems that motivate swarm robotics research. However, it also has relevance to several real-world applications, such as toxic waste cleanup, search and rescue, and demining. Additionally, since foraging usually requires robots to completely explore their terrain in order to discover the objects of interest, the *coverage* domain has similar issues to the foraging application. In coverage, robots are required to visit all areas of their environment, perhaps searching for objects (such as landmines) or executing some action in all parts of the environment (e. g., for floor cleaning). The coverage application has real-world relevance to tasks such as demining, lawn care, environmental mapping, and agriculture.

In foraging and coverage applications, a fundamental question is how to enable the robots to quickly explore their environments without duplicating actions or interfering with each other. Alternative strategies can include basic stigmergy [13], forming chains [28], and making use of heterogeneous robots [7]. Other research demonstrated in the foraging and/or coverage domain includes [21,33,69,86,94,106,108,116].

Multi-Target Observation The domain of multi-target observation requires multiple robots to monitor and/or observe multiple targets moving through the environment. The objective is to maximize the amount of time, or the likelihood, that the targets remain in view by some team member. The task can be especially challenging if there are more targets than robots. This application domain can be useful for studying strongly cooperative task solutions, since robots may have to coordinate their motions or the switching of targets to follow in order to maximize their objective. In the context of multiple mobile robot applications, the planar version of this testbed was first introduced by Parker in [82] as CMOMMT (Cooperative Multi-robot Observation of Multiple Moving Targets). Similar problems have been studied by several researchers, and extended to more complex problems such as environments with complex topography or three dimensional versions for multiple aerial vehicle applications. This domain is also related to problems in other areas, such as art gallery algorithms, pursuit evasion, and sensor coverage. This domain has practical application in many security, surveillance, and reconnaissance problems. Research applied to the multi-target observation problem in multi-robot systems includes [12,51,57,61,66,111,123].

Future Directions

Many open issues in multi-robot path planning and coordination remain. Current techniques typically do not scale well to very large numbers of robots (e. g., thousands), and many still have limitations for extensions to three dimensions (e. g., aerial robots). Many approaches have difficulty in highly stochastic environments; dynamic, online replanning of paths and coordination strategies is important in these contexts. Creating provably correct interaction strategies in these domains is an ultimate goal. Developing path planning and motion coordination techniques that incorporate practical motion and sensing constraints

of physical robots is still an open issue. Integrating these techniques onto physical robots remains uncommon, due to the practical need to integrate these path planning and coordination algorithms with complete sensing, navigation, and reasoning systems, as well as the practical difficulty of experiments involving large numbers of fallible robots. Certainly, ongoing work is addressing these important issues in multi-robot path planning and coordination; it is likely that the research community will be successful in developing solutions to extend the state of the art in this domain.

Of course, understanding how to coordinate the motions of robots in a shared workspace has both practical and scientific interest. From a practical perspective, many real-world applications can potentially benefit from the use of multiple mobile robot systems. Example applications include container management in ports [2], extra-planetary exploration [105], search and rescue [50], mineral mining [98], transportation [112], industrial and household maintenance [84], construction [103], hazardous waste cleanup [81], security [30,44], agriculture [89], and warehouse management [46]. To date, relatively few real-world implementations of these multi-robot systems have occurred, primarily due to the complexities of multiple robot systems and the relative newness of the supporting technologies. Nevertheless, many proof-of-principle demonstrations of physical multi-robot systems have been achieved, and the expectation is that these systems will find their way into practical implementations as the technology continues to mature. Because of the fundamental need for motion coordination for all applications of multi-robot systems, the work described in this chapter is of critical importance.

From a scientific perspective, understanding interactions between multiple autonomous robots might lead to insights in understanding other types of complex systems, from natural interactions in biology and social systems to engineered complex systems involving multiple interacting agents. Because multi-robot systems operate in stochastic and unpredictable settings, the study of the interaction dynamics in these settings can lead to discoveries of broader impact to a wide range of complex nonlinear systems.

Bibliography

Primary Literature

1. Alami R, Robert F, Ingrand F, Suzuki S (1995) Multi-robot cooperation through incremental plan-merging. In: Proceedings of the IEEE International Conference on Robotics and Automation, Nagoya, Aichi, 21–27 May 1995, pp 2573–2578
2. Alami R, Fleury S, Herrb M, Ingrand F, Robert F (1998) Multi-robot cooperation in the MARTHA project. IEEE Robot Autom Mag 5(1):36–47
3. Antonelli G, Chiaverini S (2006) Kinematic control of platoons of autonomous vehicles. IEEE Trans Robot 22(6):1285–1292
4. Arkin (1992) Cooperation without communication: multiagent schema-based robot navigation. J Robot Syst 9:351–364
5. Asama H, Ozaki K, Itakura H, Matsumoto A, Ishida Y, Endo I (1991) Collision avoidance among multiple mobile robots based on rules and communication. In: Proceedings of IEEE/RJS International Conference on Intelligent Robots and Systems, Osaka, 3–5 Nov 1991
6. Azarm K, Schmidt G (1997) Conflict-free motion of multiple mobile robots based on decentralized motion planning and negotiation. In: Proceedings of IEEE International Conference on Robotics and Automation, 20–25 April 1997, pp 3526–3533
7. Balch T (1999) The impact of diversity on performance in robot foraging. In: Proceedings of the Third Annual Conference on Autonomous Agents, 1–5 May 1999. ACM Press, Seattle, pp 92–99
8. Balch T, Arkin R (1998) Behavior-based formation control for multi-robot teams. IEEE Trans Robot Autom 14(6):926–939
9. Barraquand J, Latombe JC (1991) Robot motion planning: A distributed representation approach. Int J Robot Res 20(6):628–649
10. Barraquand J, Langlois B, Latombe JC (1992) Numerical potential field techniques for robot motion planning. IEEE Trans Syst Man Cybern 22:224–241
11. Barraquand J, Langlois B, Latombe JC (1997) Numerical potential field techniques for robot path planning. Int J Robot Res 16(6):759–774
12. Beard RW, McLain TW, Goodrich M (2002) Coordinated target assignment and intercept for unmanned air vehicles. In: Proceedings of IEEE International Conference on Robotics and Automation, 11–15 May 2002. IEEE, Washington DC
13. Beckers R, Holland O, Deneubourg J (1994) From local actions to global tasks: Stigmergy and collective robotics. In: Brooks R, Maes P (eds) Proceedings of the 4th International Workshop on Synthesis and Simulation of Living Systems. MIT Press, Cambridge, pp 181–189
14. Belta C, Kumar V (2004) Abstraction and control for groups of robots. IEEE Trans Robot 20(5):865–875
15. Bennewitz M, Burgard W, Thrun S (2001) Optimizing schedules for prioritized path planning of multi-robot systems. In: Proceedings of IEEE International Conference on Robotics and Automation, Seoul, 21–26 May 2001, pp 271–276
16. Bennewitz M, Burgard W, Thrun S (2002) Finding and optimizing solvable priority schemes for decoupled path planning techniques for teams of mobiel robots. Robot Auton Syst 41(2):89–99
17. Bien Z, Lee J (1992) A minimum-time trajectory planning method for two robots. IEEE Trans Robot Autom 8:414–418
18. Bobrow JE (1988) Optimal robot path planning using the minimum-time criterion. IEEE Trans Robot Autom 4(4):443–450
19. Browning B, Bruce J, Bowling M, Veloso M (2005) STP: Skills, tactics and plays for multi-robot control in adversarial environments. IEEE J Control Syst Eng 219:33–52
20. Buckley SJ (1989) Fast motion planning for multiple moving robots. In: Proceedings of IEEE International Conference

on Robotics and Automation, Scottsdale, 14–19 May 1989, pp 322–326
21. Butler ZJ, Rizzi AA, Hollis RL (2000) Cooperative coverage of rectilinear environments. In: Proceedings of IEEE International Conference on Robotics and Automation, San Francisco, 24–28 April 2000. IEEE
22. Carpin S, Pagello E (2001) A distributed algorithm for multirobot motion planning. In: Proceedings of the Fourth European Workshop on Advanced Mobile Robotics, Lund 2001
23. Chang C, Chung MJ, Lee BH (1994) Collision avoidance of two general robot manipulators by minimum delay time. IEEE Trans Robot Autom 24(3):517–522
24. Choset H, Lynch K, Hutchinson S, Kantor G, Burgard W, Kavraki L, Thrun S (2005) Principles of robot motion: theory, algorithms, and implementation. MIT Press, Cambridge
25. Chun L, Zheng Z, Chang W (1999) A decentralized approach to the conflict-free motion planning for multiple mobile robots. In: Proceedings of IEEE International Conference on Robotics and Automation, Detroit, 10–15 May 1999, pp 1544–1549
26. Clark CM, Rock SM, Latombe JC (2003) Motion planning for multiple mobile robot systems using dynamic networks. In: Proceedings of IEEE International Conference on Robotics and Automation, Taipai, 14–19 Sept 2003, pp 4222–4227
27. Dimarogonas DV, Loizou SG, Kyriakopoulos KJ, Zavlanos MM (2006) A feedback stabilization adn collision avoidance scheme for multiple independent non-point agents. Automatica 42(2):229–243
28. Drogoul A, Ferber J (1992) From Tom Thumb to the Dockers: Some experiments with foraging robots. In: Proceedings of the Second International Conference on Simulation of Adaptive Behavior, Honolulu, 2–16 Dec 1992, pp 451–459
29. Erdmann M, Lozano-Perez T (1987) On multiple moving objects. Algorithmica 2:477–521
30. Everett HR, Laird RT, Carroll DM, Gilbreath GA, Heath-Pastore TA, Inderieden RS, Tran T, Grant KJ, Jaffee DM (2000) Multiple Resource Host Architecture (MRHA) for the Mobile Detection Assessment Response System (MDARS). In: SPAWAR Systems Technical Document 3026, Revision A. San Diego
31. Fax JA, Murray RM (2004) Information flow and cooperative control of vehicle formations. IEEE Trans Autom Control 49(9)
32. Ferrari C, Pagello E, Ota J, Arai T (1998) Multirobot motion coordination in space and time. Robot Auton Syst 25:219–229
33. Fontan M, Matarić M (1998) Territorial multi-robot task division. IEEE Trans Robot Autom 15(5):815–822
34. Fujimori A, Teramoto M, Nikiforuk P, Gupta M (2000) Cooperative collision avoidance between multiple mobile robots. J Robot Syst 17(7):347–363
35. Fujimura K (1991) Motion planning in dynamic environment. Computer Science Workbench. Springer, Tokyo
36. Gage D (1993) Randomized search strategies with imperfect sensors. In: Proceedings of SPIE Mobile Robots VIII. SPIE, Boston, pp 270–279
37. Gazi V (2005) Swarm aggregations using artificial potentials and sliding-mode control. IEEE Trans Robot 21(6):1208–1214
38. Ge SS, Fua CH (2005) Queues and artificial potential trenches for multirobot formations. IEEE Trans Robot 21(4):646–656
39. Ghrist R, O'Kane JM, LaValle SM (2004) Pareto optimal coordination on roadmaps. In: Proceedings of the Workshop on Algorithmic Foundations of Robotics, Utrecht, 11–13 May 2004, pp 185–200
40. Gonzalez-Banos HH, Hsu D, Latombe JC (2006) Chapter: Autonomous mobile robots: Sensing, control, decision-making, and applications. In: Motion Planning: Recent Developments. CRC, New York
41. Griswold NC, Eem J (1990) Control for mobile robots in the presence of moving objects. IEEE Trans Robot Autom 6(2):263–268
42. Grossman DD (1988) Traffic control of multiple robot vehicles. IEEE Trans Robot Autom 5(5):491–497
43. Guo Y, Parker LE (2002) A distributed and optimal motion planning approach for multiple mobile robots. In: Proceedings of IEEE International Conference on Robotics and Automation, Washington DC, 11–15 May 2002
44. Guo Y, Parker LE, Madhavan R (2004) Towards collaborative robots for infrastructure security applications. In: Proceedings of International Symposium on Collaborative Technologies and Systems, San Diego, 18–23 Jan 2004, pp 235–240
45. Hart EE, Nilsson NJ, Raphael B (1968) A formal basis for the heuristic determination of minimum cost paths. IEEE Trans Syst Sci Cybern SSC-4(2):100–107
46. Hazard C, Wurman PR, D'Andrea R (2006) Alphabet soup: A testbed for studying resource allocation in multi-vehicle systems. In: Proceedings of AAAI Workshop on Auction Mechanisms for Robot Coordination, Boston, 16–20 July 2006, pp 23–30
47. Hopcroft JE, Schwartz JT, Sharir M (1984) On the complexity of motion planning for multiple independent objects; PSPACE-Hardness of the Warehouseman's Problem. Int J Robot Res 3(4):76–88
48. Hwang Y, Ahuja N (1992) Gross motion planning – a survey. ACM Comput Surv 24(3):219–291
49. Jadbabaie A, Lin J, Morse AS (2003) Coordination of groups of mobile autonomous agents using nearest neighbor rules. IEEE Trans Autom Control 48(6):988–1001
50. Jennings JS, Whelan G, Evans WF (1997) Cooperative search and rescue with a team of mobile robots. In: Proceedings of the 8th International Conference on Advanced Robotics, Monterey, 7–9 July 1992, pp 193–200
51. Jung B, Sukhatme G (2002) Tracking targets using multiple mobile robots: The effect of environment occlusion. Auton Robot 13(3):191–205
52. Kant K, Zucker SW (1986) Toward efficient trajectory planning: the path-velocity decomposition. Int J Robot Res 5(3):72–89
53. Kato S, Nishiyama S, Takeno J (1992) Coordinating mobile robots by applying traffic rules. In: Proceedings of the 1992 IEEE/RSJ International Conference on Intelligent Robots and Systems, Raleigh, 7–17 July 1992, pp 1535–1541
54. Kavraki LE, Svestka P, Latombe JC, Overmars MH (1996) Probabilistic roadmaps for path planning in high-dimensional configuration spaces. IEEE Trans Robot Autom 12(4):566–580
55. Khatib O (1986) Real-time obstacle avoidance for manipulators and mobile robots. Int J Robot Res 5(1):90–98
56. Kloder S, Hutchinson S (2006) Path planning for permutation-invariant multirobot formations. IEEE Trans Robot 22(4):650–665
57. Kolling A, Carpin S (2006) Multirobot cooperation for surveillance of multiple moving targets – a new behavioral approach. In: Proceedings of the IEEE International Conference on Robotics and Automation, Orlando, 15–19 May 2006. IEEE, pp 1311–1316

58. Latombe JC (1991) Robot motion planning. Kluwer Academic, Boston
59. LaValle SM (2006) Planning algorithms. Cambridge University Press, Cambridge, New York
60. LaValle SM, Hutchinson SA (1998) Optimal motion planning for multiple robots having independent goals. IEEE Trans Robot Autom 14:912–925
61. LaValle SM, Gonzalez-Banos HH, Becker C, Latombe JC (1997) Motion strategies for maintaining visibility of a moving target. In: Proceedings of the 1997 IEEE International Conference on Robotics and Automation, 20–25 April 1997. IEEE, pp 731–736
62. Lee BH, Lee CS (1987) Collision-free motion planning of two robots. IEEE Trans Syst Man Cybern 17(1):21–32
63. Lee BJ, Lee SO, Park GT (1999) Trajectory generation and motion tracking for the robot soccer game. In: Proceedings of IEEE International Conference on Intelligent Robots and Systems, 17–21 Oct 1999, pp 1149–1154
64. Lee J, Bien Z (1990) Collision-free trajectory control for multiple robots based on neural optimization network. Robotica 8:185–194
65. Lin CF, Tsai WH (1991) Motion planning for multiple robots with multi-mode operations via disjunctive graphs. Robotica 9:393–408
66. Luke S, Sullivan K, Panait L, Balan G (2005) Tunably decentralized algorithms for cooperative target observation. In: Proceedings of the fourth international joint conference on Autonomous Agents and Multiagent Systems, Utrecht, 25–29 July 2005. ACM Press, pp 911–917
67. Lumelsky VJ, Harinarayan KR (1997) Decentralized motion planning for multiple mobile robots: The cocktail party model. Auton Robot 4(1):121–135
68. Marshall JA, Broucke ME, Francis BR (2004) Formations of vehicles in cyclic pursuit. IEEE Trans Autom Control 49(11):1963–1974
69. Matarić M (1997) Behavior-based control: Examples from navigation, learning, and group behavior. J Exp Theor Artif Intell 19(2–3):323–336
70. Matarić MJ (1992) Designing emergent behaviors: From local interactions to collective intelligence. In: Meyer J, Roitblat H, Wilson S (eds) Proceedings of the 2nd international conference on simulation of adaptive behavior. MIT Press, Honolulu, pp 432–441
71. Mourikis AI, Roumeliotis SI (2006) Optimal sensor scheduling for resource-constrained localization of mobile robot formations. IEEE Trans Robot 22(5):917–931
72. Mourikis AI, Roumeliotis SI (2006) Performance analysis of multirobot cooperative localization. IEEE Trans Robot 22(4):666–681
73. Nilsson N (1982) Principles of artificial intelligence. Springer, Berlin
74. O'Donnell PA, Lozano-Perez T (1989) Deadlock-free and collision-free coordination of two robot manipulators. In: Proceedings of IEEE International Conference on Robotics and Automation, Scottsdale, 14–19 May 1989, pp 484–489
75. O'Dunlaing C, Yap CK (1982) A retraction method for planning the motion of a disc. J Algorithm 6:104–111
76. Pallottino L, Scordio VG, Bicchi A, Frazzoli E (2007) Decentralized cooperative policy for conflict resolution in multivehicle systems. IEEE Trans Robot 23(6):1170–1183
77. Pan TJ, Luo RC (1990) Motion panning for mobile robots in a dynamic environment. In: Proceedings of IEEE International Conference on Robotics and Automation, 13–18 May 1990, pp 578–583
78. Park S, Lee B (2006) A new analytical representation to robot path generation with collision avoidance through the use of the collision map. Int J Control Autom Syst 4(1):77–86
79. Parker LE (1988) A robot navigation algorithm for moving obstacles. Master's thesis, The University of Tennessee
80. Parker LE (1993) Designing control laws for cooperative agent teams. In: Proceedings of the IEEE Robotics and Automation Conference, Atlanta, 2–6 May 1993. IEEE, pp 582–587
81. Parker LE (1998) Alliance: An architecture for fault-tolerant multi-robot cooperation. IEEE Trans Robot Autom 14(2):220–240
82. Parker LE (1999) Cooperative robotics for multi-target observation. Intell Autom Soft Comput 5(1):5–19
83. Parker LE (2008) Chapter 40: Multiple mobile robot systems. In: Siciliano B, Khatib O (eds) Springer handbook of robotics. Springer, New York
84. Parker LE, Draper J (1999) Robotics applications in maintenance and repair. In: Nof S (ed) Handbook of industrial robotics, 2nd edn. Wiley, New York, pp 1023–1036
85. Parsons D, Canny J (1990) A motion planner for multiple mobile robots. In: Proceedings of IEEE International Conference on Robotics and Automation, 13–18 May 1990, pp 8–13
86. Passino K (2002) Biomimicry of bacterial foraging for distributed optimization and control. IEEE Control Syst Mag 22(3):52–67
87. Peasgood M, Clark C, McPhee J (2008) A complete and scalable strategy for coordinating multiple robots within roadmaps. IEEE Trans Robot, 24(2):283–292
88. Peng J, Akella S (2005) Coordinating multiple robots with kinodynamic constraints along specified paths. Int J Robot Res 24(4):295–310
89. Pilarski T, Happold M, Pangels H, Ollis M, Fitzpatrick K, Stentz A (1999) The demeter system for automated harvesting. In: Proceedings of the 8th International Topical Meeting on Robotics and Remote Systems, Pittsburgh, 25–29 April 1999
90. Preparata F, Shamos M (1985) Computational geometry. Springer, New York
91. Reynolds CW (1987) Flocks, herds and schools: A distributed behavioral model. ACM SIGGRAPH Comput Graph 21:25–34
92. Rude M (1997) Collision avoidance by using space-time representations of motion processes. Auton Robot 4:101–119
93. Ryan MRK (2007) Graph decomposition for efficient multi-robot path planning. In: Proceedings of the International Joint Conference on Artificial Intelligence, Hyderabad, 6–12 Jan 2007, pp 2003–2008
94. Rybski P, Stoeter S, Wyman C, Gini M (1997) A cooperative multi-robot approach to the mapping and exploration of mars. In: Proceedings of AAAI/IAAI-97. AAAI, Providence
95. Sanchez G, Latombe JC (2002) On delaying collision checking in PRM planning: application to multi-robot coordination. Int J Robot Res 21(1):5–26
96. Schwartz JT, Sharir M (1983) On the 'piano movers' problem: Iii. coordinating the motion of several independent bodies: The special case of circular bodies moving amidst polygonal obstacles. Int J Robot Res 2(3):46–75
97. Schwartz JT, Sharir M (1988) A survey of motion planning and related geometric algorithms. Artif Intell J 37:157–169
98. Shaffer G, Stentz A (1992) A robotic system for underground

coal mining. In: Proceedings of IEEE International Conference on Robotics and Automation, 12–14 May 1992, pp 633–638
99. Shan L, Hasegawa T (1996) Space reasoning from action observation for motion planning of multiple robots: mutual collision avoidance in a narrow passage. J Robot Soc Jpn 14:1003–1009
100. Sharir M (2004) Algorithmic motion planning. In: Goodman JE, O'Rourke J (eds) Handbook of discrete and computational geometry, 2nd edn. Chapman Hall/CRC, New York
101. Shiller Z, Lu HH (1990) Robust computation of path constrained time optimal motions. In: Proceedings of IEEE International Conference on Robotics and Automation, 13–18 May 1990, pp 144–149
102. Simeon T, Leroy S, Laumond J (2002) Path coordination for multiple mobile robots: a resolution-complete algorithm. IEEE Trans Robot Autom 24(1):42–49
103. Simmons R, Singh S, Hershberger D, Ramos J, Smith T (2000) First results in the coordination of heterogeneous robots for large-scale assembly. In: Proc. of the ISER Seventh International Symposium on Experimental Robotics, Honolulu, 10–13 Dec 2000. Springer
104. Stone P, Veloso M (1999) Task decomposition, dynamic role assignment, and low-bandwidth communication for real-time strategic teamwork. Artif Intell 110(2):241–273
105. Stroupe A, Okon A, Robinson M, Huntsberger T, Aghazarian H, Baumgartner E (2006) Sustainable cooperative robotic technologies for human and robotic outpost infrastructure construction and maintenance. Auton Robot 20(2):113–123
106. Sugawara K, Sano M (2002) Cooperative behavior of interacting simple robots in a clockface arranged foraging field. In: Asama H, Arai T, Fukuda T, Hasegawa T (eds) Distributed Autonomous Robotic Systems. Springer, Fukjoka, pp 331–339
107. Sugihara K, Suzuki I (1996) Distributed algorithms for formation of goemetric patterns with many mobile robots. J Robot Syst 13(3):127–139
108. Sun S, Lee D, Sim K (2001) Artificial immune-based swarm behaviors of distributed autonomous robotic systems. In: Proceedings of IEEE International Conference on Robotics and Automation Seoul 21–26 May 2001. IEEE, pp 3993–3998
109. Svestka P, Overmars M (1998) Coordinated path planning for multiple robots. Robot Auton Syst 23:125–152
110. Tabuada P, Pappas G, Lima P (2005) Motion feasibility of multi-agent formations. IEEE Trans Robot 21(3):387–392
111. Tang Z, Ozguner U (2005) Motion planning for multitarget surveillance with mobile sensor agents. IEEE Trans Robot 21(5):898–908
112. Thorpe C, Jochem T, Pomerleau D (1997) The 1997 automated highway free agent demonstration. In: Proceedings of IEEE Conference on Intelligent Transportation System, Boston, 9–12 Nov 1997, pp 496–501
113. Topaz CM, Bertozzi AL (2004) Swarming patterns in two-dimensional kinematic model for biological groups. SIAM J Appl Math 65(1):152–174
114. Tournassoud P (1986) A strategy for obstacle avoidance and its application to multi-robot systems. In: Proceedings of IEEE International Conference on Robotics and Automation, San Francisco, pp 1224–1229
115. Veloso M, Stone P, Han K (1999) The CMUnited-97 robotic soccer team: Perception and multiagent control. Robot Auton Syst 29(2–3):133–143
116. Wagner I, Lindenbaum M, Bruckstein AM (2000) Mac vs. PC – determinism and randomness as complementary approaches to robotic exploration of continuous unknown domains. Int J Robot Res 19(1):12–31
117. Wang J (1991) Fully distributed traffic control strategies for many-AGV systems. In: Proceedings of the IEEE International Workshop on Intelligent Robots and Systems, Osaka 2–5 Nov 1991, pp 1199–1204
118. Wang J, Beni G (1990) Distributed computing problems in cellular robotic systems. In: Proceedings of the IEEE International Workshop on Intelligent Robots and Systems, pp 819–826
119. Wang PKC (1989) Interaction dynamics of multiple autonomous mobile robots in bounded spatial domains. Int J Control 50(6):2109–2124
120. Wang PKC (1989) Interaction dynamics of multiple mobile robots with simple navigation strategies. J Robot Syst 6(1):77–101
121. Warren CW (1990) Multiple robot path coordination using artificial potential fields. In: Proceedings of IEEE International Conference on Robotics and Automation, 13–18 May 1990, pp 500–505
122. Weigel T, Gutmann JS, Dietl M, Kleiner A, Nebel B (2002) CS Freiburg: coordinating robots for successful soccer playing. IEEE Trans Robot Autom 5(18):685–699
123. Werger BB, Matarić MJ (2000) Broadcast of local eligibility for multi-target observation. In: Parker LE, Bekey G, Barhen J (eds) Distributed autonomous robotic systems 4. Springer, New York, pp 347–356
124. Yannakakis MZ, Papadimitriou CH, Kung HT (1979) Locking policies: Safety and freedom for deadlock. In: Proceedings of the 20th Annual Symposium on Foundations of Computer Science, San Juan, 29–31 Oct 1979, pp 286–297
125. Yuta S, Premvuti S (1992) Coordinating autonomous and centralized decision making to achieve cooperative behaviors between multiple mobile robots. In: Proceedings of the 1992 IEEE/RSJ International Conference on Intelligent robots and systems, Raleigh, 7–10 July 1992, pp 1566–1574

Books and Reviews

Arai T, Ota J (1992) Motion planning of multiple mobile robots. In: Proceedings of the 1992 IEEE/RSJ International Conference on Intelligent robots and systems, Raleigh, 7–10 July 1992, pp 1761–1768

Arai T, Pagello E, Parker LE (2002) Editorial: Advances in multi-robot systems. IEEE Trans Robot Autom 18(5):655–661

Cao Y, Fukunaga A, Kahng A (1997) Cooperative mobile robotics: Antecedents and directions. Auton Robot 4:1–23

Canny J (1988) The complexity of robot motion planning. MIT Press, Cambridge

Choset H (2001) Coverage for robotics – A survey of recent results. Ann Math Artif Intell 31(1–4):113–126

Nardi D, Farinelli A, Iocchi L (2004) Multirobot systems: a classification focused on coordination. IEEE Trans Syst Man Cybern Part B 34(5):2015–2028

Parker LE (2005) Current research in multirobot teams. Artif Life Robot 7(2–3):1–5

Thrun S, Burgard W, Fox D (2005) Probabilistic robotics. MIT Press, Cambridge

Multivariate Splines and Their Applications

Ming-Jun Lai
Department of Mathematics, The University of Georgia, Athens, USA

Article Outline

Glossary
Introduction
Definition of the Subject
Various Spline Spaces
The B-form Representation of Spline Functions
Dimension of Multivariate Spline Spaces
Approximation Power of Spline Spaces
Construction of Finite Elements and Macro-Elements
Multivariate Splines for Scattered Data Fitting
Multivariate Splines for Numerical Solution of Partial Differential Equations
Multivariate Box Spline Wavelets
Open Research Problems
Bibliography

Glossary

Multivariate splines functions are smooth piecewise polynomial functions over a triangulation of a polygonal domain in the Euclidean space $\mathbb{R}^n$ for $n \geq 2$.

Box splines functions are a generalization of univariate B-spline functions over integer knot sequence. For example, $B_{\ell,m,n}$ is a box spline of degree $\ell + m + n - 2$ on three direction mesh. $B_{k,\ell,m,n}$ is a box spline of degree $k + \ell + m + n - 2$ on four direction mesh. They are bivariate piecewise polynomial functions of certain smoothness dependent on integers k, ℓ, m, n.

Finite element and macro-element Finite element and macro-element are two special spline functions which are defined over each triangle by using the same rules for all triangles so that the spline functions are smooth over the union of all triangles.

Minimal energy method is a global method to find a spline function to interpolate a set of given data values over the vertices of a triangulation and minimize the thin-plate energy functional.

Discrete least squares fitting is a global method to find a spline s to fit a given set of scattered data by minimizing the summation of the squares of the difference of s at a given location and the given value at the given location. The summation above is called the square of the discrete L_2 norm.

Penalized least squares fitting is another global method to find a spline function s to fit a given set of scattered data by minimizing the sum of the square of the discrete L_2 norm and the square of energy norm of the spline function.

Wavelets are functions whose integer translates and dilation's form an orthonormal basis for $L_2(\mathbb{R}^n)$.

Tight wavelet frame are functions whose integer translates and dilation's form a tight frame for $L_2(\mathbb{R}^n)$.

Introduction

Multivariate splines are smooth piecewise polynomial functions over a triangulation of a polygonal domain in $\mathbb{R}^n$ for $n \geq 2$. They are very efficient for evaluation and manipulation on computer and very flexible for approximating known or unknown functions or any given data sets. More than 50 years ago, Courant started using a continuous piecewise linear finite element over a triangulation of polynomial domain to approximate the solution of a partial differential equation (PDE). This seminal study generated a great deal of interest in constructing various finite elements and macro elements over the period 1965–1977 (e. g. [17,43,45]) as well as promoting the study of the dimension of multivariate spline splines in 1979 ([46]) and their approximation orders in the period 1988–1998 (cf. [10,11,34]). A new period of study of computation with multivariate splines took off at the beginning of this century (cf., e. g. [5,35,48]).

They are extremely useful for numerical solution of partial differential equations, construction of smooth surfaces to fit a given set of scattered data, etc. For example, FEM (the finite element method) is a fundamental tool for solving viscous incompressible flow equations and has direct application in the design of hydraulic turbines and rheologically complex flows which appear in many processes involving plastics or molten metals. Another example, CAGD (computer-aided geometric design) uses multivariate splines as a basic and effective tool for many engineering research fields, e. g., designing car hoods, ship hulls, and aircraft wings. They are one of the subjects of study in applied and computational mathematics such as Numerical Analysis, Approximation Theory, Computer-Aided Geometric Design, and Numerical Solution of PDE. They have found many applications in applied sciences.

In this article, we restrict our attention to multivariate splines that are defined on polygonal domains in Euclidean space $\mathbb{R}^n$ with $n \geq 2$. Spline functions, piecewise polynomial functions defined on partitions of an interval in $\mathbf{R}^1$ are called univariate splines which will not be dis-

cussed any further in this article (cf. [7] and [47] for theory and applications of univariate splines).

We shall give a definition of multivariate splines and some examples of various spline spaces. Dimension and approximation power of spline spaces will be briefly discussed. Then we will explain how to use these spline functions for data interpolation and fitting, for numerical solution of partial differential equations, for construction of multivariate wavelets and frames.

Definition of the Subject

Let us start with $\mathbb{R}^n$ with $n = 2$. Let Δ be a collection of triangles in whose union forms a polygonal domain Ω in $\mathbb{R}^2$. We say that Δ is a regular triangulation if it satisfies the following property: for any $t, t' \in \Delta$, $t \cap t'$ is either empty or a common edge of t and t' or a common vertex of t and t'.

Given two integers $d \geq 0$ and $0 \leq r < d$, let

$$S_d^r(\Delta) := \left\{ s \in C^r(\Omega) \colon s|_t \in \mathbb{P}_d, \forall\ t \in \Delta \right\}$$

be the multivariate spline space of degree d and smoothness r, where $\mathbb{P}_d$ denotes the space of all polynomials of degree $\leq d$. It is a standard spline space. Typically, $S_d^0(\Delta)$ is the continuous spline space of degree d which is a very popular finite element space. Also we consider super spline spaces. That is, let $\rho = \{\rho_v, v \in \mathcal{V}\}$ be a set of integers $\rho_v \geq 0$ associated with vertices in $\mathcal{V}$ of Δ and $\mathbf{r} = \{r_e, e \in \mathcal{E}_I\}$ be a set of integers $r_e \geq 0$ associated with interior edges $\mathcal{E}_I$ of Δ. Suppose that $\rho_v \geq r \geq 0$ for all $v \in \mathcal{V}$ and $r_e \geq r \geq 0$ for all $e \in \mathcal{E}_I$. Let

$$S_d^{\mathbf{r},\rho}(\Delta) = \left\{ s \in S_d^r(\Delta),\ s \in C^{\rho_v} \text{ at } v \in \mathcal{V} \text{ and } s \in C^{r_e} \text{ across } e \in \mathcal{E}_I \right\}$$

be the spline subspace of super smoothness ρ, smoothness $\mathbf{r}$ and degree d. If $\mathbf{r} = (r, r, \ldots, r)$ and $\rho = (r, r, \ldots, r)$, then $S_d^{\mathbf{r},\rho}(\Delta) = S_d^r(\Delta)$. When $\mathbf{r}$ and ρ have constant components, we just use $S_d^{r,\rho}(\Delta)$ to denote the super spline space. Typically, we consider $S_5^{1,2}(\Delta)$ which is the space of all spline functions that are C^1 across all interior edges of Δ and are C^2 at all vertices of Δ. When $\mathbf{r}$ is a vector of different components we call a function in $S_d^{\mathbf{r},\rho}(\Delta)$ a spline of variable smoothness.

Next let $\mathbf{d} = \{d_t, t \in \Delta\}$ be a set of integers $d_t \geq 0$ associated with triangles of Δ. Let $\mathbf{r}$ and ρ as above. Define

$$S_{\mathbf{d}}^{\mathbf{r},\rho}(\Delta) = \left\{ s \in S_d^{\mathbf{r},\rho}(\Delta),\ s|_t \in \mathbb{P}_{d_t}, t \in \Delta \right\}$$

with the spline space of variable smoothness ρ, $\mathbf{r}$ and variable degree $\mathbf{d}$ associated with the vertices, interior edges and triangles of Δ, where $\mathbb{P}_{d_t}$ stands for the space of polynomials of degree d_t. This is a user-friendly spline space allowing one to choose a spline function using polynomials of less degree in certain areas and higher degree in other areas. It is especially useful to trim off the oscillations of interpolatory surfaces.

The definition of bivariate spline spaces can be easily generalized to the multivariate setting and to the spherical setting. We refer to [36] and references therein.

Various Spline Spaces

In addition to the continuous spline space $S_d^0(\Delta)$, the following spline spaces in the bivariate setting are very popular in the literature and in applications:

1. $S_3^1(\Delta_{\mathrm{CT}})$ is the C^1 cubic spline space over Clough–Tocher refinement of triangulations. In Fig. 1., the left is a given triangulation Δ and the right is the Clough–Tocher refinement Δ_{CT} of Δ.
2. $S_2^1(\Delta_{\mathrm{PS}})$ denotes the C^1 quadratic spline space over Powell–Sabin refinement of triangulations. In Fig. 2, the left is a given triangulation Δ and the right is a Powell–Sabin refinement Δ_{PS} of Δ.
3. $S_3^1(\diamondplus)$ stands for the C^1 cubic spline space over triangulated quadrilaterals $\diamondplus$. One first decomposes a polygonal domain into strictly convex quadrilaterals. Then add two diagonals to each quadrilateral to obtain a triangulation $\diamondplus$ as shown in Fig. 3.

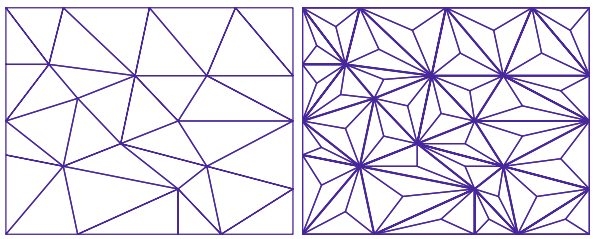

Multivariate Splines and Their Applications, Figure 1
A triangulation and its Clough–Tocher refinement

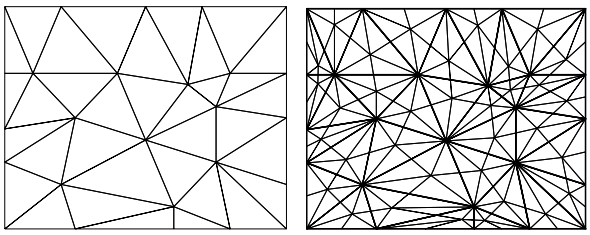

Multivariate Splines and Their Applications, Figure 2
A triangulation and its Powell–Sabin refinement

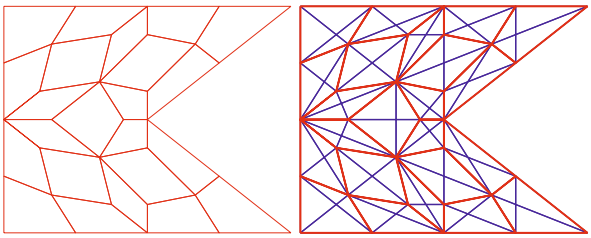

Multivariate Splines and Their Applications, Figure 3
A quadriangulation and its associated triangulation

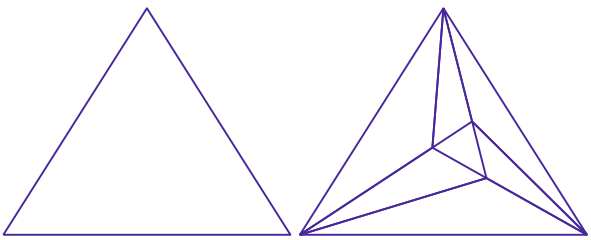

Multivariate Splines and Their Applications, Figure 4
A triangle and its Wang refinement

4. Let $S_5^2(\Delta_W)$ be the C^2 quintic spline space over Wang's refinement of triangulations. In Fig. 4. we show a triangle and its Wang's refinement.

There are many other spline spaces discussed in the literature such as Powell–Sabin's 12 split refinement method. See [36] for details. Such spline spaces have been generalized to the spherical setting and the trivariate setting except for the Wang refinement.

Another class of multivariate splines are box spline functions which are a natural generalization of uniform B-splines. They are piecewise polynomial functions defined on a uniform triangulation. To be more precise, let D be a set of non zero vectors in $\mathbb{R}^n$ (counting multiple of a same vector) which span $\mathbb{R}^n$. The box spline ϕ_D associates with the direction set D is the function whose Fourier transform is defined by

$$\hat{\phi}_D(\omega) = \prod_{\xi \in D} \frac{1 - e^{-i\xi\cdot\omega}}{i\xi \cdot \omega} .$$

It can be shown that box spline ϕ_D is a piecewise polynomial function of degree $\leq \#D - d$, where $\#D$ denotes the cardinality of D. For more properties of box splines, see [12]. In particular, for $n = 2$ and $e_1 = (1,0)^{\mathrm{T}}$, $e_2 = (0,1)^{\mathrm{T}}$, and

$$D = \{\underbrace{e_1, \ldots, e_1}_{\ell}, \underbrace{e_2, \ldots, e_2}_{m}, \underbrace{e_1 + e_2, \ldots, e_1 + e_2}_{n}\} ,$$

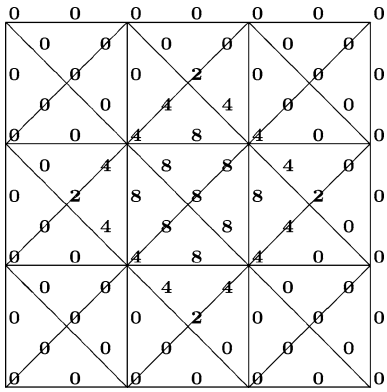

Multivariate Splines and Their Applications, Figure 5
B-coefficients of $16\phi_{1111}$

the box spline $\phi_{\ell,m,n}$ based on such a direction set D is called the 3-direction box spline whose Fourier transform is

$$\hat{\phi}_{\ell,m,n}(\xi,\eta) = \left(\frac{1 - e^{-i\xi}}{i\xi}\right)^{\ell} \left(\frac{1 - e^{-i\eta}}{i\eta}\right)^{m} \cdot \left(\frac{1 - e^{-i(\xi+\eta)}}{i(\xi+\eta)}\right)^{n} .$$

Similarly, box spline $\phi_{\ell mnk}$ based on a four direction mesh is defined in terms of Fourier transform by

$$\hat{\phi}_{\ell,m,n,k}(\xi,\eta) = \hat{\phi}_{\ell,m,n}(\xi,\eta) \left(\frac{1 - e^{-i(\xi-\eta)}}{i(\xi-\eta)}\right)^{k} .$$

Box splines on $\mathbb{R}^2$ can be shown by using their coefficients of polynomials in B-form which is explained in the next section. For example, for box spline ϕ_{1111}, its B-coefficients together with its underlying triangulation are shown in Fig. 5. For another example, the B-coefficients and the underlying triangulation of ϕ_{2111} are shown in Fig. 6.

The B-form Representation of Spline Functions

It is standard in the literature to use the B-form representation for multivariate splines over triangulations since the publication of [21] and [8].

We start with $\mathbb{R}^2$. Let $T = \langle v_1, v_2, v_3 \rangle$ be a non-degenerate triangle with $\mathbf{v}_i = (x_i, y_i)$, $i = 1, 2, 3$. It is well-known that every point $\mathbf{v} = (x, y)$ can be written uniquely in the form

$$\mathbf{v} = \lambda_1 \mathbf{v}_1 + \lambda_2 \mathbf{v}_2 + \lambda_3 \mathbf{v}_3 ,$$

with

$$\lambda_1 + \lambda_2 + \lambda_3 = 1 ,$$

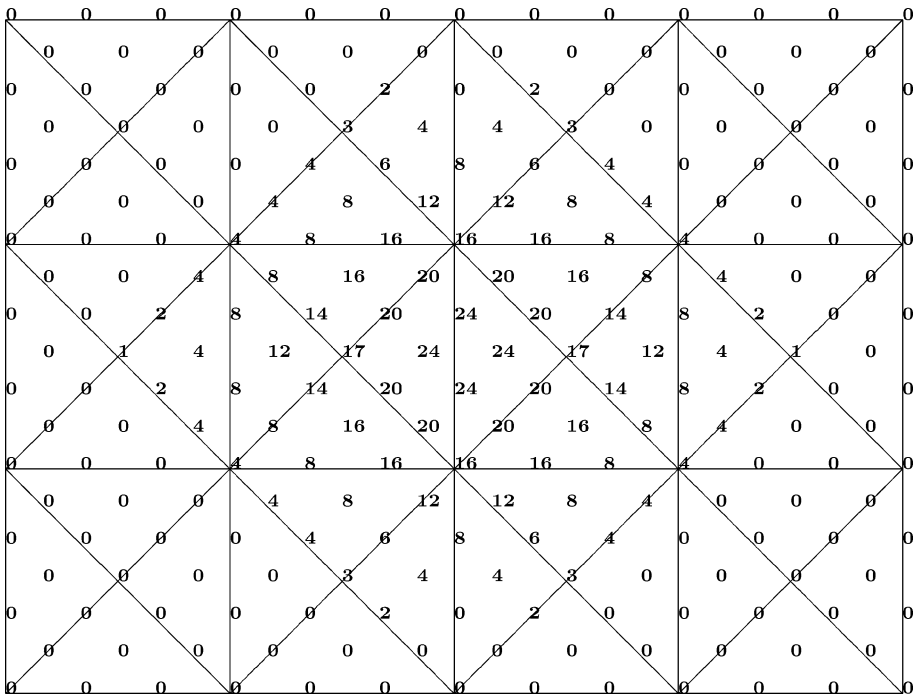

Multivariate Splines and Their Applications, Figure 6
B-coefficients of $48\phi_{2111}$

where λ_1, λ_2, and λ_3 are called the barycentric coordinates of the point $\mathbf{v} = (x, y)$ relative to the triangle T. Moreover each λ_i is a linear polynomial in x, y. Let

$$B^d_{ijk}(v) = \frac{d!}{i!j!k!}\lambda_1^i\lambda_2^j\lambda_3^k\,, \quad i+j+k=d\,.$$

They are called the Bernstein–Bézier polynomials of degree d. In fact, the set

$$\mathcal{B}^d = \left\{B^d_{ijk}(x,y,z)\,,\ i+j+k=d\right\}$$

is a basis for the space of polynomials $\mathbb{P}_d$. As a consequence any polynomial p of degree d can be written uniquely in terms of B^d_{ijk}'s, i. e.,

$$p = \sum_{i+j+k=d} c_{ijk}B^d_{ijk}\,.$$

The representation for polynomials is referred to as the B-form with respect to T. c_{ijk} are called B-coefficients of p. Let

$$\mathcal{D}_{d,T} = \left\{\xi_{ijk} = \frac{i\mathbf{v}_1 + j\mathbf{v}_2 + k\mathbf{v}_3}{d},\ i+j+k=d, T\in\Delta\right\}$$

be a set of the domain points of degree d over triangulation Δ. The polynomial p can be displayed using its B-coefficients c_{ijk} over their domain points. See Fig. 5 for the B-coefficients of spline function ϕ_{1111} over all triangles. Note that the coefficients located on an edge of two triangles belong to both triangles in the sense that ϕ_{1111} is continuous across the edge and hence, the B-coefficients over the common edge of two neighboring triangles are the same.

Let $s \in S^r_d(\Delta)$ be a spline function over triangulation Δ. Since s restricted to each triangle $T \in \Delta$ is a polynomial of degree d, we may write

$$s|_T = \sum_{i+j+k=d} c^T_{ijk}B^d_{ijk}\,, \quad T\in\Delta\,.$$

Such a representation is called the B-form representation of the spline function s. We denote by $\mathbf{c} := \{c^T_{ijk}, i+j+k=d, T\in\Delta\}$ the B-coefficient vector of s.

To evaluate a polynomial in B-form, there is the so-called de Casteljau algorithm which we now describe. For $p = \sum_{i+j+k=d} c_{ijk}B^d_{ijk}$, let us write $c_{ijk} =: c^{(0)}_{ijk}(\lambda)$ with $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ being the barycentric coordinates of $\mathbf{v} = (x, y)$ with respective to T and define for a positive integer $r \geq 1$

$$c^{(r)}_{ijk}(\lambda) = \lambda_1 c^{(r-1)}_{i+1,j,k}(\lambda) + \lambda_2 c^{(r-1)}_{i,j+1,k}(\lambda) + \lambda_3 c^{(r-1)}_{i,j,k+l}(\lambda)\,.$$

We have then

$$p = \sum_{i+j+k=d-r} c_{ijk}^{(r)}(\lambda) B_{ijk}^{d-r} , \quad 0 \le r \le d .$$

In particular, for $r = d$, we have

$$p = c_{0,0,0}^{(d)}(\lambda)$$

which is the value of p at $\mathbf{v} = (x, y)$ whose barycentric coordinates are $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ with $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

Next we discuss how to take derivatives of polynomials in B-form. We start with formulas for the directional derivatives of p in a direction defined by a vector $\mathbf{u}$.

$$D_{\mathbf{u}} p = d \sum_{i+j+k=d-1} c_{ijk}^{(1)}(\mathbf{a}) B_{ijk}^{d-1} ,$$

with $\mathbf{a} = (a_1, a_2, a_3)$ the T-coordinates of $\mathbf{u}$; that is,

$$\mathbf{u} = a_1 \mathbf{v}_1 + a_2 \mathbf{v}_2 + a_3 \mathbf{v}_3 ,$$

with

$$a_1 + a_2 + a_3 = 0 .$$

Note that if $\mathbf{u} = \mathbf{v}_1 - \mathbf{v}_2$ is the direction vector, the T-coordinates of $\mathbf{u}$ are $(1, -1, 0)$. In general, we have

$$D_{\mathbf{u}}^m p(\mathbf{v}) = \frac{d!}{(d-m)!} \sum_{i+j+k=d-m} c_{ijk}^{(m)}(\mathbf{a}) B_{ijkl}^{d-m}(\mathbf{v}) .$$

Note that for arbitrary direction vector $u = (u_1, u_2) \in \mathbb{R}^2$,

$$D_{\mathbf{u}} p = u_1 \frac{\partial}{\partial x} p + u_2 \frac{\partial}{\partial y} p .$$

For a triangle $T = \langle v_1, v_2, v_3 \rangle$, if we let $\mathbf{u} = \mathbf{v}_2 - \mathbf{v}_1 = (x_2 - x_1, y_2 - y_1)$, $\mathbf{v} = \mathbf{v}_3 - \mathbf{v}_1 = (x_3 - x_1, y_3 - y_1)$, it follows that

$$D_x p = \frac{(y_3 - y_1)}{2A_T} D_{\mathbf{u}} p - \frac{(y_2 - y_1)}{2A_T} D_{\mathbf{v}} p$$
$$D_y p = \frac{(x_2 - x_1)}{2A_T} D_{\mathbf{v}} p - \frac{(x_3 - x_1)}{2A_T} D_{\mathbf{u}} p ,$$

where A_T is the area of T.

There are precise formulas for the integrals and inner products of polynomials in B-form (cf. [14]).

Lemma 1 *Let p be a polynomial of degree d with B-coefficients c_{ijk}, $i + j + k = d$ on a triangle T. Then*

$$\int_T p(x, y) dx dy = \frac{A_T}{\binom{d+2}{2}} \sum_{i+j+k=d} c_{ijk} ,$$

where A_T = area of T.

Lemma 2 *Let q be another polynomial with B-coefficients d_{ijk}, $i + j + k = d$, the inner product of p and q over t is given by*

$$\int_t p(x, y) q(x, y) dx dy = \frac{A_T}{\binom{2d}{d}\binom{2d+2}{2}} \cdot \sum_{\substack{i+j+k=d \\ r+s+t=d}} c_{ijk} d_{rst} \binom{i+r}{i} \binom{j+s}{j} \binom{k+t}{k} .$$

We now discuss the smoothness conditions for a spline function s in $S_d^r(\Delta)$. These are well-known conditions on the coefficients of s that will assure that s has certain global smoothness properties.

Theorem 1 *Let $t = \langle v_1, v_2, v_3 \rangle$ and $t' = \langle v_1, v_2, v_4 \rangle$ be two triangles with common edge $\langle v_1, v_2 \rangle$. Then s is of class C^r on $t \cup t'$ if and only if*

$$c_{ijm}^{t'} = \sum_{\mu+\nu+\kappa=m} c_{i+\mu, j+\nu, \kappa}^{t} B_{\mu,\nu,\kappa}^{m}(v_4) ,$$
$$m = 0, \ldots, r , \quad i + j = d - m .$$

For a proof, see [21]. The geometric meaning of the smoothness condition is striking. Indeed, let us consider the coefficient c_{ijk} as a value at domain point ξ_{ijk} and connect all the coefficients by line segments as shown in Fig. 7. Then the coefficients of two polynomials located at and closest to the common interior edge form five planes across the common edge as shown in Fig. 7 when these polynomials form in C^1 fashion. For the geometric meaning of the other smoothness conditions in the bivariate setting, see [30].

This theorem guarantees the existence of a matrix H such that s is in $C^r(\Omega)$ if and only if

$$H\mathbf{c} = 0 ,$$

where $\mathbf{c}$ encodes the B-coefficients of s.

More properties on the B-form above can be found in [36]. The B-form for bivariate polynomials can be easily generalized to the multivariate setting. Indeed, let $t = \langle v^{(0)}, \ldots, v^{(n)} \rangle \in \mathbb{R}^n$ be an n-simplex with $n + 1$ distinct points $v^{(k)}$, $k = 0, 1, \ldots, n$. Suppose that the n-simplex t has nonzero volume. Then for any point $x \in \mathbb{R}^n$, $x - v^{(0)}$ can be uniquely expressed by a linear combination of $v^{(i)} - v^{(0)}$, $i = 1, \ldots, n$. That is,

$$x = v^{(0)} + \sum_{i=1}^{n} \lambda_i (v^{(i)} - v^{(0)}) .$$

Let $\lambda_0 = 1 - \sum_{i=1}^{n} \lambda_i$. Then the $(n + 1)$-tuple $(\lambda_0, \lambda_1, \ldots, \lambda_n)$ is called the barycentric coordinate of x with re-

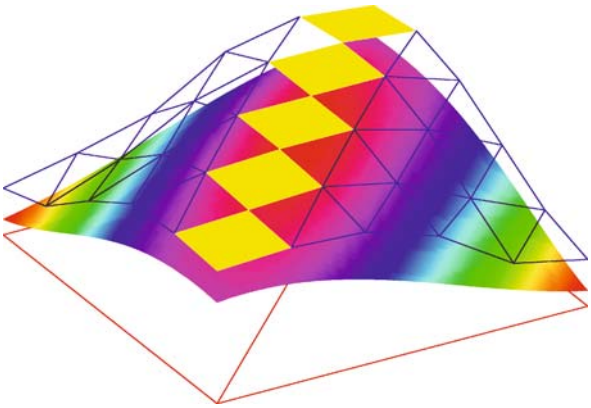

Multivariate Splines and Their Applications, Figure 7
Geometric meaning of C^1 smoothness condition

spect to t. It is easy to see that each λ_i is a linear function of x. Next let $\mathbf{Z}^{n+1}$ be the set of all multi-integers in $\mathbb{R}^{n+1}$. For a multi-integer $\alpha = (\alpha_0, \ldots, \alpha_n) \in \mathbf{Z}^{n+1}$ with $|\alpha| = \alpha_0 + \ldots + \alpha_n \geq 0$, let

$$B^t_\alpha(x) := \frac{|\alpha|!}{\alpha!}\lambda^\alpha ,$$

where $\alpha! = \alpha_0! \ldots \alpha_n!$ and

$$\lambda^\alpha = \prod_{i=0}^{n} \lambda_i^{\alpha_i} .$$

Then it is clear that $B^t_\alpha(x)$ is a polynomial of degree $|\alpha|$ in x. It can be shown that $\{B^t_\alpha(x), \alpha \in \mathbf{Z}^{n+1}, |\alpha| = d\}$ forms a basis for polynomials of degree $\leq d$ (cf. [7]). Thus, any polynomial p of total degree d may be written in terms of $B^t_\alpha(x)$'s as

$$p(x) = \sum_{|\alpha|=d} c^t_\alpha B^t_\alpha(x) \tag{1}$$

for some coefficients c^t_α's depending on t. Thus, any spline function s is given by

$$s(x) = \sum_{|\alpha|=d} c^t_\alpha B^t_\alpha(x) , \quad x \in t \in \Delta \tag{2}$$

with B-coefficient vector $\{c^t_\alpha, |\alpha| = d, t \in \Delta\}$ of length $\hat{d}T$, where T denotes the number of n-simplices in Δ and

$$\hat{d} = \binom{d+n}{n} .$$

This representation of the spline function s is called the B-form of s. (Cf. [21] and [8].)

One simple property of the B-form of polynomials is:

Lemma 3 *Let $t = \langle v^{(0)}, \ldots, v^{(n)} \rangle$ be an n-simplex in $\mathbb{R}^n$ and let $p(x)$ be a polynomial of degree d given in B-form (1) with respect to t. Then*

$$p(v^{(k)}) = c^t_{d\mathbf{e}^k} , \quad \forall \quad 0 \leq k \leq n ,$$

where $\mathbf{e}^k = (0, \ldots, 0, 1, 0, \ldots, 0)$ with 1 appearing in the $(k+1)$th place.

To evaluate $p(x)$ in B-form (1), we use the so-called de Casteljau algorithm. The derivative of $p(x)$ in B-form can be given in B-form again. The integration of a polynomial p in B-form is a sum of all coefficients of p with multiplication by an appropriate constant. See, e. g. [14] for all these properties. Another important property is the following Markov inequality:

Lemma 4 *Let $1 \leq q \leq \infty$. There exists a constant N depending only on d such that*

$$\frac{\|\{c^t_\alpha, |\alpha| = d\}\|_q}{N} \leq \|p\|_{q,t} \leq \|\{c^t_\alpha, |\alpha| = d\}\|_q ,$$

for any polynomial $p(x) = \sum_{|\alpha|=d} c^t_\alpha B^t_\alpha(x)$, where $\|p\|_{q,t}$ denotes the standard L_q norm over the n-simplex t and $\|\{c^t_\alpha, |\alpha| = d\}\|_q$ denotes the ℓ_q norm of the sequence $\{c^t_\alpha, |\alpha| = d\}$.

We refer the interested reader to [34] or [36] for a proof in the bivariate setting which can be generalized to the multivariate setting easily.

Next we look at the smoothness conditions. Let

$$t_1 = \left\langle v^{(0)}, \ldots, v^{(k)}, v^{(k+1)}, \ldots, v^{(n)} \right\rangle$$

and

$$t_2 = \left\langle v^{(0)}, \ldots, v^{(k)}, u^{(k+1)}, \ldots, u^{(n)} \right\rangle$$

be two n-simplices in $\mathbb{R}^n$ and $\tilde{t} = \langle v^{(0)}, \ldots, v^{(k)} \rangle$ the k-simplex which is a common facet of t_1 and t_2, with $0 \leq k < n$. Let F be a function defined on $t_1 \cup t_2$ by

$$F(x) = \begin{cases} p_d(x) = \sum_{|\alpha|=n} a_\alpha B^{t_1}_\alpha(x) , & \text{if } x \in t_1 \\ q_d(x) = \sum_{|\alpha|=n} b_\alpha B^{t_2}_\alpha(x) , & \text{if } x \in t_2 . \end{cases}$$

Let us assume that F is well defined on $\tilde{t}$. Writing $u^{(j)} = \sum_{i=0}^{n} c_{ji} v^{(i)}$, $j = k+1, \ldots, n$, we have the following:

Theorem 2 *Suppose that t_1 and t_2 are two n-simplices such that $\tilde{t} = t_1 \cap t_2$ is a $(n-1)$-simplex in $\mathbb{R}^n$. Let F be the function defined above. Then $F \in C^r(t_1 \cup t_2)$ if and only if the following conditions hold*

$$b_{(\alpha_0, \ldots, \alpha_{n-1}, \ell)} = \sum_{|\gamma|=\ell} a_{(\alpha_0, \ldots, \alpha_{n-1}, 0)+\gamma} B^{t_1}_\gamma(u^{(n)}) \tag{3}$$

for $0 \leq \ell \leq r$.

These are the well-known smoothness conditions (cf. [8]). Next we look at the degree reduction conditions. These conditions allow us to constrain the spline function to be of variable degree over the simplices. Let

$$\Delta_{ij} c_\alpha = c_{\alpha+e_i} - c_{\alpha+e_j}$$

be a difference operator, where $e_i = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbf{Z}^{n+1}$ with 1 in the ith entry and similar for e_j. Inductively, let

$$\Delta_{ij}^k = \Delta_{ij}(\Delta_{ij}^{k-1})$$

for $k \geq 2$. For any multi-integer $\beta = (\beta_1, \ldots, \beta_n)$, let

$$\Delta^\beta = \Delta_{10}^{\beta_1} \ldots \Delta_{n0}^{\beta_n}$$

be a difference operator of order $|\beta|$. Our degree reduction conditions are:

Theorem 3 *Let $p = \sum_{|\alpha|=d} c_\alpha B_\alpha^t$ be a polynomial of degree d in B-form with respect to t. Then p is a polynomial of degree $d_t < d$ if*

$$\Delta^\beta c_\alpha = 0\,, \quad d_t < |\beta| \leq d\,, \quad |\alpha| = d - |\beta|\,, \tag{4}$$

where $\Delta^\beta c_\alpha = \Delta^\beta c.|_\alpha$, that is, the difference operators are applied first before the evaluation at the index α.

The conditions can be verified easily and are left to the interested reader. It is easy to see that both conditions (3) and (4) are linear relations among the B-coefficients of polynomials.

Let us summarize the discussions above as follows: For each spline function in

$$S := S_{\mathbf{d}}^{\mathbf{r},\rho}(\Delta)$$

the spline space of smoothness $\mathbf{r}$, super smoothness ρ and degree $\mathbf{d}$ for three fixed sequences ρ, $\mathbf{r}$, and $\mathbf{d}$ associated with the k-simplices with $0 \leq k < n-1$, interior $n-1$ simplices, and n-simplices of Δ, we write

$$s = \sum_{t\in\Delta} \sum_{|\alpha|=d_t} c_\alpha^t \mathcal{B}_\alpha^t\,, \tag{5}$$

with $\mathbf{c} = (c_\alpha^t, |\alpha| = d_t, t \in \Delta) \in \mathbb{R}^N$, $N = \sum_{t\in\Delta} \hat{d}_t$ with $\hat{d}_t = \binom{d_t+n}{n}$ and

$$\mathcal{B}_\alpha^t(x) = \begin{cases} B_\alpha^t(x)\,, & \text{if } x \in t \\ 0\,, & x \in \Delta\backslash\{t\}\,. \end{cases}$$

In addition, $\mathbf{c}$ satisfies the constraints $H\mathbf{c} = 0$ for the smoothness conditions that S possesses and $J\mathbf{c} = 0$ for the degree reduction conditions.

Dimension of Multivariate Spline Spaces

The dimension of multivariate spline spaces is difficult to determine when spline functions are smooth in the sense that the smoothness of spline functions is greater or equal to 1. If we consider a spline space without any smoothness, then it is a piecewise polynomial function and the dimension of such a spline space is trivial. For example, in the bivariate setting,

Theorem 4 *For any triangulation Δ,*

$$\dim S_d^{-1}(\Delta) = \frac{(d+1)(d+2)}{2} N$$

where N denotes the number of triangles in Δ.

If we consider a spline space which is a space of continuous spline functions, then the dimension is also easy to determine. (See, e. g. [36] for a proof.)

Theorem 5 *For any triangulation Δ,*

$$\dim S_d^0(\Delta) = V + (d-1)E + \binom{d-1}{2} N\,,$$

where V, E, and N are the number of vertices, edges, and triangles in Δ.

When the smoothness of a spline space is bigger than or equal to 1, the dimension of $S_d^r(\Delta)$ with $r \geq 1$ is difficult to determine when d is small. If $d \geq 3r+2$, we have the following result in the bivariate setting (cf. [27]). Let V_I, V_B be the numbers of interior and boundary vertices of Δ, respectively. Similarly, let E_I, E_B be the numbers of interior and boundary edges, and let N be the number of triangles in Δ.

Theorem 6 *Suppose Δ is a regular triangulation with no holes. Then for all $d \geq 3r+2$,*

$$\dim S_d^r(\Delta) = \frac{d^2 + r^2 - r + d - 2rd}{2} V_B + (d-r)(d-2r)V_I + \frac{-2d^2 + 6rd - 3r^2 + 3r + 2}{2} + \sigma\,,$$

where σ is $\sigma = \sum_{v\in\mathcal{V}_I} \sigma_v$. Here $\mathcal{V}_I$ is the collection of all interior vertices of Δ and

$$\sigma_v := \sum_{j=1}^{d-r} \big(r + j + 1 - jm_v\big)_+ \tag{6}$$

and m_v stands for the number of edges attached to v with different slopes.

However, when $d < 3r + 2$, the dimension is extremely difficult to determine for a general triangulation Δ. For example, we still do not know the dimension of $S_3^1(\Delta)$ in the bivariate setting so far. See an open problem regarding the dimension of trivariate spline spaces in the last section. Nevertheless, we have a pretty good upper bound and lower bound as explained below.

Let $\mathcal{V}_I$ be the set of interior vertices of Δ. For each $v \in \mathcal{V}_I$, let m_v be the number of edges attached to v with different slopes. Suppose V_I and E_I are the numbers of interior vertices and edges of Δ, respectively.

Theorem 7 *For all* $0 \le r \le d$,

$$F + \sigma \le \dim S_d^r(\Delta)$$

where

$$F := \binom{d+2}{2} + \binom{d-r+1}{2} E_I - \left[\binom{d+2}{2} - \binom{r+2}{2}\right] V_I\,, \quad (7)$$

and σ *is as defined in the above theorem.*

In order to have an upper bound, we need a concept of admissible decomposition of a triangulation. Suppose Δ is a regular triangulation of a domain Ω without holes. Let $\mathcal{T}_1, \ldots, \mathcal{T}_n$ be a grouping of the triangles of Δ into disjoint subsets, and for each $i = 1, \ldots, n$, let Ω_i be the union of the triangles in $\bigcup_{j=1}^{i} \mathcal{T}_j$. Let $\Omega_0 = \emptyset$. Suppose there exist vertices $v_1, \ldots, v_n$ such that:

1) For each $1 \le i \le n$, $\mathcal{T}_i$ is the union of all triangles in $\Delta \backslash \Omega_{i-1}$ that share the vertex v_i.
2) For each $2 \le i \le n$, v_i is on the boundary of Ω_{i-1}.

Then we say that $\mathcal{T}_1, \ldots, \mathcal{T}_n$ is an *admissible decomposition of* Δ *with centers* $v_1, \ldots, v_n$.

We can construct an admissible decomposition of any regular triangulation Δ without holes by starting with $\mathcal{T}_1 := \mathrm{star}(v_1)$ for some arbitrary vertex of Δ. Then we repeatedly choose a vertex on the boundary of Ω_{i-1} and take $\mathcal{T}_i$ to be all unchosen triangles attached to v_i. Since the starting vertex can be arbitrary, it is clear that any given Δ has several different admissible decompositions. With the above concept, we have

Theorem 8 *Let* $0 \le r \le d$, *and suppose* $\mathcal{T}_1, \ldots, \mathcal{T}_n$ *is an admissible decomposition of* Δ *with centers* $v_1, \ldots, v_n$. *For each* v_i, *let* n_i *be the number of interior edges of* Δ *attached to* v_i *but not attached to any* v_j *with* $j < i$. *Let* $\tilde{w}_i$ *be the number of such edges, where we count only edges with different slopes. Finally, let*

$$\widetilde{\sigma}_i := \begin{cases} \sum_{j=1}^{d-r} \left(r + j + 1 - j\tilde{w}_i\right)_+\,, & \text{if } v_i \in \mathcal{V}_I\,, \\ 0\,, & \text{otherwise}\,. \end{cases}$$

Then

$$\dim S_d^r(\Delta) \le F + \sum_{i=1}^{n} \widetilde{\sigma}_i\,,$$

where F *is as in (7).*

For some special triangulations, we do know the dimension of spline spaces. For example, we know the dimension of $S_3^1(\Delta_{\mathrm{CT}})$ $S_2^1(\Delta_{\mathrm{PS}})$, and $S_3^1(\diamondplus)$, where $\Delta_{\mathrm{CT}}, \Delta_{\mathrm{PS}}$, and $\diamondplus$ stand for the Clough–Tocher refinement of triangulation Δ, the Powell–Sabin refinement of Δ, and triangulated quadrangulation. These spline spaces will be discussed in Sect. "Construction of Finite Elements and Macro-Elements". For more information on the dimension of $S_d^r(\Delta_{\mathrm{CT}})$, $S_d^r(\Delta_{\mathrm{PS}})$ and $S_d^r(\diamondplus)$ for some $d < 3r + 2$ dependent on r and triangulations, see [36].

Next we consider the dimension of trivariate spline spaces. Suppose Δ is a tetrahedral partition of a bounded domain $\Omega \in \mathbb{R}^3$. Given an integer $d \ge 0$, let $\mathcal{P}_d$ be the space of trivariate polynomials of degree d. Then we define the associated space of C^0 polynomial splines of degree d over Δ as

$$S_d^0(\Delta) := \left\{s \in C^0(\Omega),\ s|_T \in \mathcal{P}_d\,, \text{for all } T \in \Delta\right\}.$$

Theorem 9 *The dimension of* $S_d^0(\Delta)$ *is*

$$n := V + (d-1)E + \binom{d-1}{2} F + \binom{d-1}{3} N\,,$$

where V, E, F, T *are the number of vertices, edges, faces, and tetrahedra in* Δ, *respectively.*

Recently, Alfeld and Schumaker continued to work on the dimension of trivariate splines and found some reasonable upper and lower bounds (cf. [2]).

Approximation Power of Spline Spaces

In this section we discuss how well smooth functions can be approximated by bivariate splines. The results will be useful in deriving error bounds for various practical interpolation and approximation methods to be discussed in later sections.

Let Ω be a polygonal domain in $\mathbb{R}^2$ and recall that $\|\cdot\|_{q,\Omega}$ is the standard q-norm over Ω, for $1 \le q \le \infty$.

Given $m \geq 1$, let $W_q^m(\Omega)$ be the Sobolev space with associated seminorm $|\cdot|_{m,q,\Omega}$.

Fix $0 \leq r < d$. Let m be the largest integer such that for every polygonal domain Ω and every regular triangulation Δ of Ω with smallest angle θ, for every $f \in W_q^m(\Omega)$, there exists a spline $s \in S_d^r(\Delta)$ with

$$\|f - s\|_{q,\Omega} \leq K\,|\Delta|^m |f|_{m,q,\Omega}\,,$$

where the constant K depends only on r, d, θ, and the Lipschitz constant of the boundary of Ω. Then we say that S_d^r has *approximation power m in the q-norm*. If this holds for $m = d + 1$, we say that S_d^r has *approximation power in the q-norm*.

Let us explain the approximation power of S_d^r for various values of r and d as follows:

- If $r = 0$, then the space S_d^0 has full approximation power in all of the q-norms.
- If $r > 0$ and $d \geq 3r + 2$, then the space S_d^r has full approximation power in all of the q-norms.
- If $r > 0$ and $(3r + 2)/2 \leq d \leq 3r + 1$, then in any q-norm, the space S_d^r has approximation power at most d.
- If $r > 0$ and $d < (3r + 2)/2$, then in any q-norm, the space S_d^r has approximation power zero.

More precisely, we state the following theorems and leave their proofs and references to [36].

Theorem 10 *Suppose Δ is a regular triangulation of a polygonal domain Ω, and let $1 \leq q \leq \infty$. Then for every $f \in W_q^{d+1}(\Omega)$, there exists a spline $s \in S_d^0(\Delta)$ such that*

$$\|D_x^\alpha D_y^\beta (f - s)\|_{q,\Omega} \leq K\,|\Delta|^{d+1-\alpha-\beta}\,|f|_{d+1,q,\Omega}\,,$$

for all $0 \leq \alpha + \beta \leq d$. The constant K depends only on d, the smallest angle in Δ, and the Lipschitz constant of the boundary of Ω.

Theorem 11 *Let $d \geq 3r + 2$ with $r > 0$, and suppose Δ is a regular triangulation of Ω. Then for every $f \in W_q^{d+1}(\Omega)$, there exists a spline $s \in S_d^r(\Delta)$ such that*

$$\|D_x^\alpha D_y^\beta (f - s)\|_{q,\Omega} \leq K\,|\Delta|^{d+1-\alpha-\beta}\,|f|_{d+1,q,\Omega}\,,$$

for all $0 \leq \alpha + \beta \leq d$. If Ω is convex, then the constant K depends only on r, d, and the smallest angle in Δ. If Ω is not convex, then K also depends on the Lipschitz constant of the boundary of Ω.

To explain the approximation power of S_d^r for $r > 0$ and $d < 3r + 2$, we consider a standard unit square domain $H = [0, 1] \times [0, 1]$. Let $0 < r < d < 3r + 2$. We first show that when $d < (3r + 2)/2$, S_d^r has approximation power zero. Given a positive integer n, let

$$0 = x_0 < x_1 < \cdots < x_n < x_{n+1} = 1\,,$$
$$0 = y_0 < y_1 < \cdots < y_n < y_{n+1} = 1\,,$$

with $x_i = y_i = ih$ for $i = 0, \ldots, n + 1$, where $h := 1/(n + 1)$. We write Δ_n for the associated uniform type-I triangulation of H obtained by drawing in the northeast diagonals.

Theorem 12 *Suppose $r > 0$ and $d < (3r + 2)/2$ and $1 \leq q \leq \infty$. Then the approximation power of $S_d^r(\Delta_n)$ in the q-norm is zero.*

Next we consider the case $(3r + 2)/2 \leq d \leq 3r + 1$ which is difficult and delicate. The proof uses several different techniques and box spline theory. We again refer to [36] for a proof.

Theorem 13 *Suppose $(3r + 2)/2 \leq d \leq 3r + 1$ and let $1 \leq q \leq \infty$. Then the approximation power of the space $S_d^r(\Delta_n)$ in the q-norm is at most d.*

The approximation order of special spline spaces, e. g., $S_d^r(\Delta_{\mathrm{CT}})$, $S_d^r(\Delta_{\mathrm{PS}})$, $S_d^r(\Phi)$ are summarized in [36].

We have very little information on the approximation order for trivariate spline spaces. See the open problem in the last section.

Construction of Finite Elements and Macro-Elements

In this section, we describe some of the most useful finite elements and macro-elements in $\mathbb{R}^2$ and $\mathbb{R}^3$. They are C^1 finite element or C^1 macro-elements. Mainly we present their formulas using B-form and refer their verifications to [36] where C^r finite element, macro-elements, and locally supported functions with $r \geq 2$ can be found. Other explicit constructions of these finite elements can be found in [13] and [16].

Bivariate Finite Element and Macro-Elements

A C^1 Polynomial Macro-Element over Triangulations Let $T = \langle v_1, v_2, v_3 \rangle$ be a triangle in a triangulation Δ of a polygonal domain Ω.

Write $v_i = (x_i, y_i)$, $i = 1, 2, 3$. A C^1 quintic finite element s restricted to T can be explicitly given in terms of its

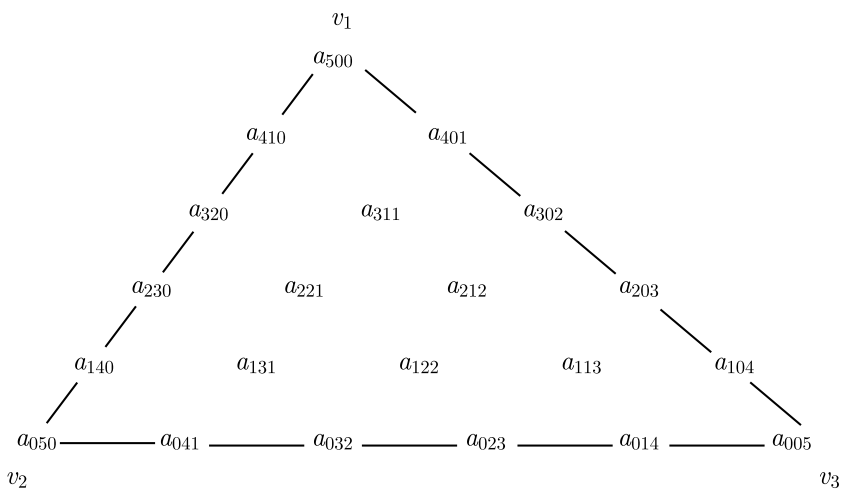

Multivariate Splines and Their Applications, Figure 8
B-coefficients of quintic polynomials on triangle *T*

B-coefficients in Fig. 8 as follows:

$$
\begin{aligned}
c^T_{500} &= s(v_1)\,,\\
c^T_{410} &= [h_2 s_x(v_1) + \hat{w}_2 s_y(v_1)]/5 + s(v_1)\,,\\
c^T_{401} &= [h_3 s_x(v_1) + \hat{w}_3 s_y(v_1)]/5 + s(v_1)\,,\\
c^T_{320} &= [h_2^2 s_{xx}(v_1) + 2h_2\hat{w}_2 s_{xy}(v_1) + \hat{w}_2^2 s_{yy}(v_1)]/20\\
&\qquad\qquad + 2c^T_{410} - s(v_1)\,,\\
c^T_{311} &= [h_2 h_3 s_{xx}(v_1) + (h_2\hat{w}_3 + h_3\hat{w}_2) s_{xy}(v_1)\\
&\qquad + \hat{w}_2\hat{w}_3 s_{yy}(v_1)]/20 + c^T_{401} + c^T_{410} - s(v_1)\,,\\
c^T_{302} &= [h_3^2 s_{xx}(v_1) + 2h_3\hat{w}_3 s_{xy}(v_1) + \hat{w}_3^2 s_{yy}(v_1)]/20\\
&\qquad\qquad + 2c^T_{401} - s(v_1)\,,
\end{aligned}
$$

where $h_i := x_i - x_1$ and $\hat{w}_i := y_i - y_1$ for $i = 2, 3$, and $s(v_1), s_x(v_1), s_y(v_1)$, etc. are any given values at v_1.

A similar formula for $c^T_{050}, c^T_{140}, c^T_{041}, c^T_{230}, c^T_{131}, c^T_{032}$ holds at v_2 as well as $c^T_{005}, c^T_{104}, c^T_{014}, c^T_{203}, c^T_{113}, c^T_{023}$ holds at v_3 respectively.

Next we need to describe the coefficients for $c^T_{122}, c^T_{212}, c^T_{221}$ as in Fig. 8. For $e = \langle v_2, v_3\rangle$, let u_e be a direction not parallel to e. Write $u_e = v_e - w_e$ and let (a_1, a_2, a_3) be the difference of the barycentric coordinates of v_e and w_e with respect to T. Then

$$
\begin{aligned}
c^T_{122} = \frac{16}{30a_1} D_{u_e} s(\eta_e) &- \frac{1}{6}\big[c^T_{140} + 4c^T_{131} + 4c^T_{113} + c^T_{104}\big]\\
&- \frac{a_2}{6a_1}\big[c^T_{050} + 4c^T_{041} + 6c^T_{032} + 4c^T_{023} + c^T_{014}\big]\\
&- \frac{a_3}{6a_1}\big[c^T_{041} + 4c^T_{032} + 6c^T_{023} + 4c^T_{014} + c^T_{005}\big]\,,
\end{aligned}
$$

where $\eta_e = (v_2 + v_3)/2$. A similar formula holds for c^T_{212} and c^T_{221}.

That is, the C^1 quintic finite element can be determined by using values and derivative values at v_i, $i = 1, 2, 3$ as well as values at the midpoint of three edges. The above explicit construction of the C^1 quintic finite element shows that for every function $f \in C^2(\Omega)$, there is a unique spline $s \in S^1_5(\Delta)$ solving the following Hermite interpolation problem

$$
\begin{aligned}
D^\alpha_x D^\beta_y s(v) &= D^\alpha_x D^\beta_y f(v), \quad \text{all } v \in \mathcal{V} \text{ and } 0 \le \alpha + \beta \le 2,\\
D_{u_e} s(\eta_e) &= D_{u_e} f(\eta_e), \quad \text{all } e \in \mathcal{E}\,,
\end{aligned}
$$

where $\mathcal{V}$ and $\mathcal{E}$ stand for the collections of vertices and edges of Δ.

A C^1 Clough–Tocher Macro-Element Let Δ be a triangulation of a polygonal domain Ω. Denote by Δ_{CT} the corresponding Clough–Tocher refinement of Δ by splitting each triangle $T \in \Delta$ at the barycenter of T.

A C^1 cubic macro-element s restricted to each triangle T can be determined by using the following formulas. That is, the B-coefficients as shown in Fig. 9 can be computed as follows:

$$
\begin{aligned}
c_1 &= s(v_1)\,,\\
c_2 &= s(v_2)\,,\\
c_3 &= s(v_3)\,,\\
c_4 &= [(x_2 - x_1)s_x(v_1) + (y_2 - y_1)s_y(v_1)]/3 + s(v_1)\,,\\
c_5 &= [(x_c - x_1)s_x(v_1) + (y_c - y_1)s_y(v_1)]/3 + s(v_1)\,,
\end{aligned}
$$

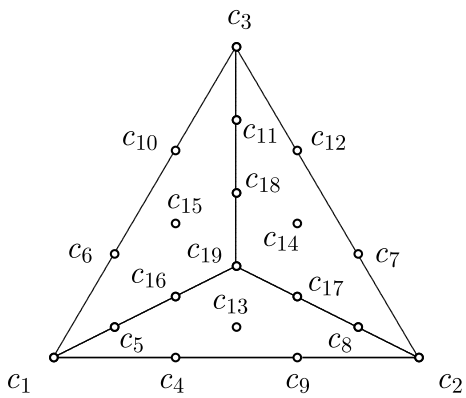

Multivariate Splines and Their Applications, Figure 9
B-coefficients of $s|_T$

$$
\begin{aligned}
c_6 &= [(x_3 - x_1)s_x(v_1) + (y_3 - y_1)s_y(v_1)]/3 + s(v_1)\,,\\
c_7 &= [(x_3 - x_2)s_x(v_2) + (y_3 - y_2)s_y(v_2)]/3 + s(v_2)\,,\\
c_8 &= [(x_c - x_2)s_x(v_2) + (y_c - y_2)s_y(v_2)]/3 + s(v_2)\,,\\
c_9 &= [(x_1 - x_2)s_x(v_2) + (y_1 - y_2)s_y(v_2)]/3 + s(v_2)\,,\\
c_{10} &= [(x_1 - x_3)s_x(v_3) + (y_1 - y_3)s_y(v_3)]/3 + s(v_3)\,,\\
c_{11} &= [(x_c - x_3)s_x(v_3) + (y_c - y_3)s_y(v_3)]/3 + s(v_3)\,,\\
c_{12} &= [(x_2 - x_3)s_x(v_3) + (y_2 - y_3)s_y(v_3)]/3 + s(v_3)\,,
\end{aligned}
$$

where $s(v_i), s_x(v_i), s_y(v_i), i = 1, 2, 3$ are given values.

To determine c_{13}, let $e := \langle v_1, v_2\rangle$, let u_e be a direction not parallel to e. Suppose (a_1, a_2, a_3) are the difference of the barycentric coordinates of $v_e - w_e = u_e$. Then

$$
c_{13} = \frac{4}{6a_3} D_{u_e} s(\eta_e) - \frac{1}{2}(c_5 + c_8) - \frac{a_1}{2a_3}(c_1 + 2c_4 + c_9) - \frac{a_2}{2a_3}(c_4 + 2c_9 + c_2)\,.
$$

A similar formula holds for c_{14} and c_{15}.

Finally, we find the remaining coefficients

$$
\begin{aligned}
c_{16} &= (c_{15} + c_5 + c_{13})/3\,,\\
c_{17} &= (c_{13} + c_8 + c_{14})/3\,,\\
c_{18} &= (c_{14} + c_{11} + c_{15})/3\,,\\
c_{19} &= (c_{18} + c_{16} + c_{17})/3\,.
\end{aligned}
$$

For each triangle T in Δ, a C^1 spline function s restricted to T is determined by the data involving evaluation at points in T. That is, the coefficients of s can be computed locally, one triangle at a time.

A C^1 Powell–Sabin Macro-Element Let Δ be a triangulation of a polygonal domain Ω. Denote by Δ_{PS} the corresponding Powell–Sabin refinement of Δ by splitting each triangle $T \in \Delta$ at the incenter of T and then connect the incenters of any two neighboring triangles.

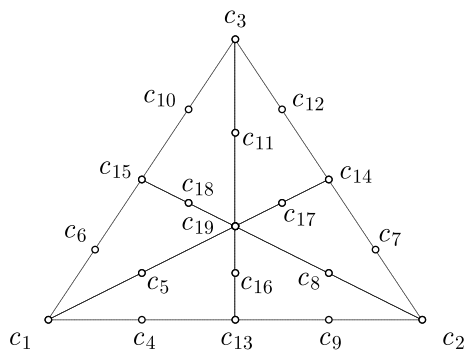

Multivariate Splines and Their Applications, Figure 10
B-coefficients of $s|_T$

Let T_{PS} be the Powell–Sabin split of a triangle $T := \langle v_1, v_2, v_3\rangle$ in Δ, and let (x_c, y_c) be the incenter of T. Then a C^1 quadratic Powell–Sabin finite element s restricted to T can be determined as follows. That is, its B-coefficients over triangle T as shown in Fig. 10 can be computed using the following formulas:

$$
\begin{aligned}
c_1 &= s(v_1)\,,\\
c_2 &= s(v_2)\,,\\
c_3 &= s(v_3)\,,\\
c_4 &= [(\hat{x}_1 - x_1)s_x(v_1) + (\hat{y}_1 - y_1)s_y(v_1)]/2 + s(v_1)\,,\\
c_5 &= [(x_c - x_1)s_x(v_1) + (y_c - y_1)s_y(v_1)]/2 + s(v_1)\,,\\
c_6 &= [(\hat{x}_3 - x_1)s_x(v_1) + (\hat{y}_3 - y_1)s_y(v_1)]/2 + s(v_1)\,,\\
c_7 &= [(\hat{x}_2 - x_2)s_x(v_2) + (\hat{y}_2 - y_2)s_y(v_2)]/2 + s(v_2)\,,\\
c_8 &= [(x_c - x_2)s_x(v_2) + (y_c - y_2)s_y(v_2)]/2 + s(v_2)\,,\\
c_9 &= [(\hat{x}_1 - x_2)s_x(v_2) + (\hat{y}_1 - y_2)s_y(v_2)]/2 + s(v_2)\,,\\
c_{10} &= [(\hat{x}_3 - x_3)s_x(v_3) + (\hat{y}_3 - y_3)s_y(v_3)]/2 + s(v_3)\,,\\
c_{11} &= [(x_c - x_3)s_x(v_3) + (y_c - y_3)s_y(v_3)]/2 + s(v_3)\,,\\
c_{12} &= [(\hat{x}_2 - x_3)s_x(v_3) + (\hat{y}_2 - y_3)s_y(v_3)]/2 + s(v_3)\,,
\end{aligned}
$$

where $w_i := (\hat{x}_i, \hat{y}_i)$ are the points on the edges of T. After the above coefficients are computed, we can find the remaining coefficients by using the following

$$
\begin{aligned}
c_{13} &= r_1 c_4 + s_1 c_9\,,\\
c_{14} &= r_2 c_7 + s_2 c_{12}\,,\\
c_{15} &= r_3 c_{10} + s_3 c_6\,,\\
c_{16} &= r_1 c_5 + s_1 c_8\,,\\
c_{17} &= r_2 c_8 + s_2 c_{11}\,,\\
c_{18} &= r_3 c_{11} + s_3 c_5\,.
\end{aligned}
$$

A C^1 Quadrilateral Macro-Element Let $\diamondsuit$ be a strictly convex quadrangulation of a polygonal domain Ω. Denote $\diamondplus$ to be the triangulation obtained from $\diamondsuit$ by drawing in the diagonals of each quadrilateral. In this section we discuss the cubic spline space $\S_3^1(\diamondplus)$. Let $\mathcal{V}$ and $\mathcal{E}$ be the sets of vertices and edges of $\diamondsuit$.

For each quadrilateral $Q := \langle v_1, v_2, v_3, v_4\rangle$, let v_Q be the intersection of the two diagonals of Q. Write $v_Q = r_1 v_1 + s_1 v_3 = r_2 v_4 + s_2 v_2$ for some r_1, s_1 and r_2, s_2. A C^1 cubic finite element s restricted to Q can be directly computed from any given values and first derivative values at each vertex as well as normal derivative values at each edge. That is, its B-coefficients over Q as shown in Fig. 11 can be computed as follows:

$$
\begin{aligned}
c_1 &= s(v_1)\,,\\
c_2 &= s(v_2)\,,\\
c_3 &= s(v_3)\,,\\
c_4 &= s(v_4)\,,\\
c_5 &= [(x_2 - x_1)s_x(v_1) + (y_2 - y_1)s_y(v_1)]/3 + s(v_1)\,,\\
c_6 &= [(x_Q - x_1)s_x(v_1) + (y_Q - y_1)s_y(v_1)]/3 + s(v_1)\,,\\
c_7 &= [(x_4 - x_1)s_x(v_1) + (y_4 - y_1)s_y(v_1)]/3 + s(v_1)\,,\\
c_8 &= [(x_3 - x_2)s_x(v_2) + (y_3 - y_2)s_y(v_2)]/3 + s(v_2)\,,\\
c_9 &= [(x_Q - x_2)s_x(v_2) + (y_Q - y_2)s_y(v_2)]/3 + s(v_2)\,,\\
c_{10} &= [(x_1 - x_2)s_x(v_2) + (y_1 - y_2)s_y(v_2)]/3 + s(v_2)\,,
\end{aligned}
$$

and

$$
\begin{aligned}
c_{11} &= [(x_4 - x_3)s_x(v_3) + (y_4 - y_3)s_y(v_3)]/3 + s(v_3)\,,\\
c_{12} &= [(x_Q - x_3)s_x(v_3) + (y_Q - y_3)s_y(v_3)]/3 + s(v_3)\,,\\
c_{13} &= [(x_2 - x_3)s_x(v_3) + (y_2 - y_3)s_y(v_3)]/3 + s(v_3)\,,\\
c_{14} &= [(x_1 - x_4)s_x(v_4) + (y_1 - y_4)s_y(v_4)]/3 + s(v_4)\,,\\
c_{15} &= [(x_Q - x_4)s_x(v_4) + (y_Q - y_4)s_y(v_4)]/3 + s(v_4)\,,\\
c_{16} &= [(x_3 - x_4)s_x(v_4) + (y_3 - y_4)s_y(v_4)]/3 + s(v_4)\,,
\end{aligned}
$$

where $(x_Q, y_Q) := v_Q$.

The coefficients $c_{17}, c_{18}, c_{19}, c_{20}$ can be computed from cross-boundary information. That is, let $e := \langle v_1, v_2\rangle$, let u_e be a direction not parallel to e. Suppose (a_1, a_2, a_3) are the difference of the barycentric coordinates of $v_e - w_e = u_e$. Then

$$
c_{17} = \frac{4}{6a_3} D_{u_e} s(\eta_e) - \frac{1}{2}(c_6 + c_9) - \frac{a_1}{2a_3}(c_1 + 2c_5 + c_{10}) - \frac{a_2}{2a_3}(c_5 + 2c_{10} + c_2)\,.
$$

A similar formula holds for c_{18}, c_{19} and c_{20}.

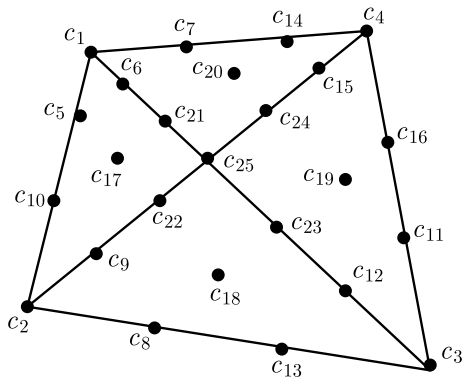

Multivariate Splines and Their Applications, Figure 11
B-coefficients of $s|_Q$

Finally, we can compute the remaining coefficients using the following formulas:

$$
\begin{aligned}
c_{21} &= r_2 c_{20} + s_2 c_{17}\,,\\
c_{22} &= r_1 c_{17} + s_1 c_{18}\,,\\
c_{23} &= r_2 c_{19} + s_2 c_{18}\,,\\
c_{24} &= r_1 c_{20} + s_1 c_{19}\,.\\
c_{25} &= r_1 r_2 c_{20} + s_1 r_2 c_{19} + r_1 s_2 c_{17} + s_1 s_2 c_{18}\,.
\end{aligned}
$$

The dimension of $S_3^1(\diamondplus)$ is equal to $3V + E$, where V and E stand for the numbers of vertices and edges of quadrangulation $\diamondsuit$.

Trivariate Finite Element and Macro-Elements

Let Δ be a tetrahedral partition of a polygonal domain Ω in $\mathbb{R}^3$. The first construction of C^1 macro-elements over tetrahedral partitions was obtained by Ženiček in [56]. See also [42]. In our notation, these finite-elements are in $S_9^{1,2,4}(\Delta)$ which is the space of all spline functions in $S_9^{-1}(\Delta)$ that are C^1 over the union of all tetrahedra in Δ, C^2 around each edge of Δ, and C^4 at each vertex of Δ. Here, a function s is said to be C^2 around an edge if s is twice differentiable at each point of e and C^4 at a vertex v if s is four times differentiable at v.

Before we describe the construction, we need some notation which will be also useful in later subsections. We denote by $\mathcal{I}_d(\Delta) := \{(\xi^t_{i,j,k,\ell} \colon i + j + k + \ell = d, t \in \Delta\}$ the collection of all domain points associated with $S_d^{-1}(\Delta)$. We shall use the concept of 3D minimal determining set. Let Γ be a proper subset of $\mathcal{I}_d(\Delta)$. Γ is a *determining set* for a spline subspace $S \subset S_d^{-1}(\Delta)$ if any spline function $s \in S$ whose B-coefficients associated with the domain points in Γ are zero is zero everywhere. Γ is a *minimal determining set* if Γ is a determining set and the cardi-

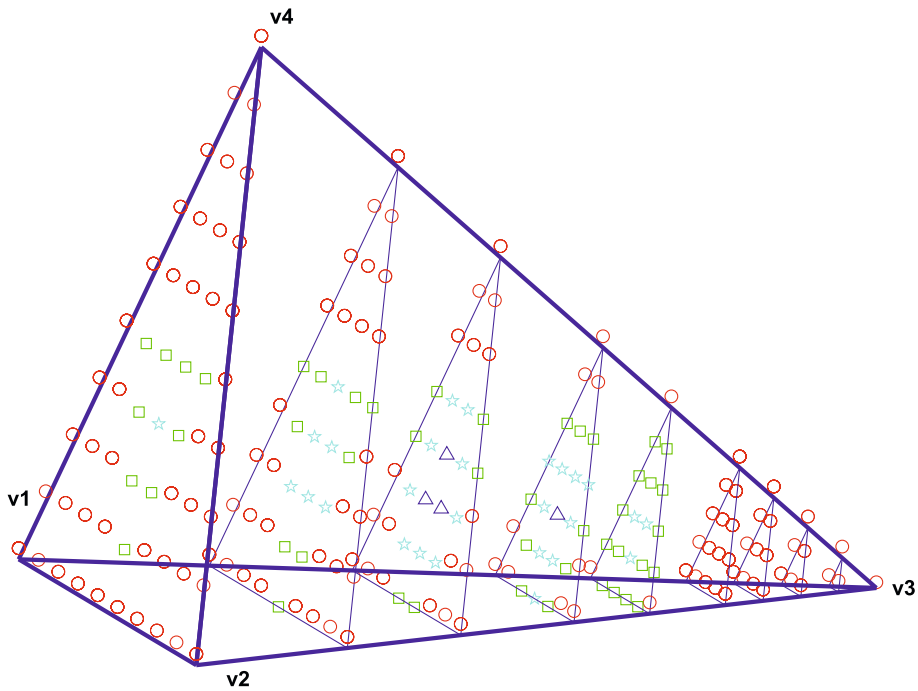

Multivariate Splines and Their Applications, Figure 12
Indication of domain point subsets on a tetrahedron

nality of Γ is the smallest possible. Next we need additional notation: Letting $e = \langle v_1, v_2\rangle$ be an edge of t and $f = \langle v_1, v_2, v_3\rangle$ be a face of t, we denote

$$\mathcal{D}_m^t(v_1) = \{\xi_{i,j,k,\ell}^t \colon i \geq d-m, i+j+k+\ell = d\}\,,$$
$$\mathcal{E}_m^t(e) = \{\xi_{i,j,k,\ell}^t \colon i+j \leq m, i+j+k+\ell = d\}\,,$$
$$\mathcal{F}_m^t(f) = \{\xi_{i,j,k,\ell}^t \colon \ell \leq m, i+j+k+\ell = d\}$$

for integer $0 \leq m \leq d$.

C^1 Splines of Degree 9 over Tetrahedral Partitions We now fix $d = 9$ and specify the following domain point sets to form a minimal determining set for $S_9^{1,2,4}(\Delta)$.

(1) For each vertex $v \in \Delta$, choose a tetrahedron t_v in Δ such that t_v contains v. Let

$$S_v := \mathcal{D}_4^{t_v}(v)\,.$$

We note that the determination of the coefficients of a spline s associated with domain points in S_v is equivalent to the assignment of all derivatives of order 4 of s at v. The domain points in $S_{v_1}, S_{v_2}, S_{v_3}, S_{v_4}$ are marked with ∘'s in Fig. 12.

(2) For each edge $e \in \Delta$, choose a tetrahedron t_e in Δ such that t_e contains e. Writing $e = \langle u, v\rangle$, we let

$$S_e := \mathcal{E}_2^{t_e}(e) \backslash \left(\mathcal{D}_4^{t_e}(u) \cup \mathcal{D}_4^{t_e}(v)\right)\,.$$

The domain points in S_e for all edges of a tetrahedron are marked with ◇'s in Fig. 12.

(3) For each face $f \in \Delta$, choose a tetrahedron t_f in Δ containing f. Writing $f = \langle u, v, w\rangle$ and $t_f = \langle u, v, w, x\rangle$, we let

$$S_f := \mathcal{F}_1^{t_f}(f) \backslash \Big(\mathcal{D}_4^{t_f}(u) \cup \mathcal{D}_4^{t_f}(v) \cup \mathcal{D}_4^{t_f}(w) \cup \mathcal{E}_2^{t_f}(\langle u, v\rangle) \cup \mathcal{E}_2^{t_f}(\langle v, w\rangle) \cup \mathcal{E}_2^{t_f}(\langle u, w\rangle)\Big)\,.$$

The domain points in S_f for all faces are marked with ★'s in Fig. 12.

(4) For each tetrahedron t, let S_t be the remaining coefficients on t, i. e.,

$$S_t = \{(t, i, j, k, \ell), i \geq 2, j \geq 2, k \geq 2, \ell \geq 2\}\,.$$

The domain points in S_t are marked with Δ's in Fig. 6.

Let

$$\Gamma := \bigcup_{v\in\Delta} S_v \cup \bigcup_{e\in\Delta} S_e \cup \bigcup_{f\in\Delta} S_f \cup \bigcup_{t\in\Delta} S_t\,.$$

Then it can be shown that Γ forms a minimal determining set for $S_9^{1,2,4}(\Delta)$. Using Γ, we can construct a basis $s_\gamma, \gamma \in \Gamma$ for $S_9^{1,2,4}(\Delta)$ as before. Thus, we have

$$\dim S_9^{1,2,4}(\Delta) = 35V + 8E + 7F + 4T\,,$$

where V, E, F, T denote the number of vertices, edges, faces, and tetrahedra of Δ. It is well-known that one can use $S_9^{1,2,4}(\Delta)$ to construct C^1 spline interpolants. We omit the details here.

The C^1 Quintic Macro-Elements over 3D Clough–Tocher Tetrahedral Partition Let Δ be a tetrahedral partition of a polygonal domain Ω in $\mathbb{R}^3$. In [1], Alfeld generalized the well-known bivariate Clough–Tocher split of triangles to the trivariate setting. He split each tetrahedron t into four subtetrahedra at the center m_t of t using the triangular planes each of which consists of an edge of t and m_t. For simplicity, we denote the refinement of a tetrahedral partition Δ by Δ_{A}. In this subsection, we present the following version of the C^1 macro-element on Alfeld's split of tetrahedra. Let

$$S_5^{1,2,3}(\Delta_{\mathrm{A}}) = \big\{s \in S_5^{-1}(\Delta_{\mathrm{A}}) \colon s \in C^1(\Omega), s \in C^2(v), \\ v \in \Delta, s \in C^3(m_t), t \in \Delta\big\}$$

be the space of all spline functions which are C^1 globally while C^2 at each vertex of Δ and C^3 at the center m_t of each tetrahedron $t \in \Delta$. Let

$$\mathcal{I}_5(\Delta_{\mathrm{A}}) = \{\xi^t_{ijk\ell}, i + j + k + \ell = 5, t \in \Delta_{\mathrm{A}}\}$$

be the collection of the domain points of $S_5^{-1}(\Delta_{\mathrm{A}})$.

(1) For each vertex v of Δ, choose a tetrahedron t_v in Δ_{A} containing v. Let

$$S_v := \mathcal{D}_2^{t_v}(v)\,.$$

The domain points in S_v for all vertices of a tetrahedron are marked with ∘'s in Figs. 13, 14, and 15.

(2) For each edge e of Δ, choose a tetrahedron t_e in Δ_{A} having an edge e. Writing $e = \langle u, v\rangle$. Let

$$S_e := \mathcal{E}_1^{t_e}(e) \backslash \left(\mathcal{D}_2^{t_e}(u) \cup \mathcal{D}_2^{t_e}(v)\right)\,.$$

The domain points in S_e for all edges of a tetrahedron are marked with ◇'s in Figs. 13, 14, and 15.

(3) For each face f of Δ, choose a tetrahedron t_f in Δ_{A} such that t_f contains f. Writing $f = \langle u, v, w\rangle$ and $t_f = \langle f, x\rangle$, let

$$S_f := \{\xi^{t_f}_{2,1,1,1}, \xi^{t_f}_{1,2,1,1}, \xi^{t_f}_{1,1,2,1}\}\,.$$

The domain points in S_f for a tetrahedron are marked with ★'s in Figs. 13, 14, and 15.

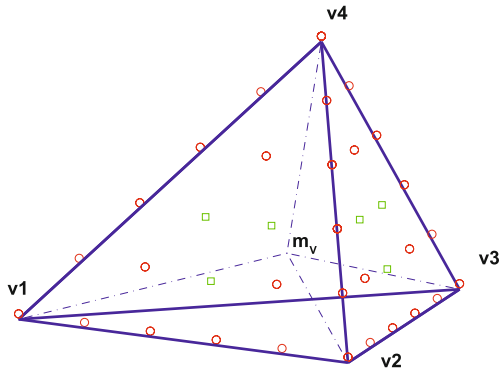

Multivariate Splines and Their Applications, Figure 13
Domain point (on the first layer) on the Alfeld split of a tetrahedron

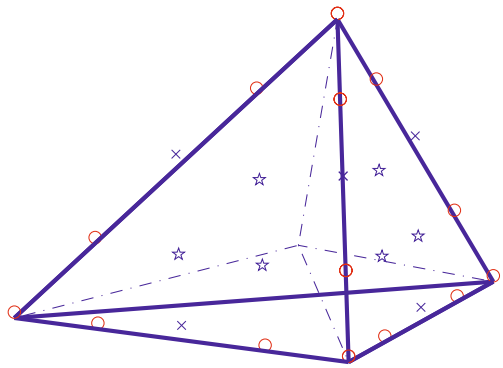

Multivariate Splines and Their Applications, Figure 14
Domain point (on the second layer) on the Alfeld split of a tetrahedron

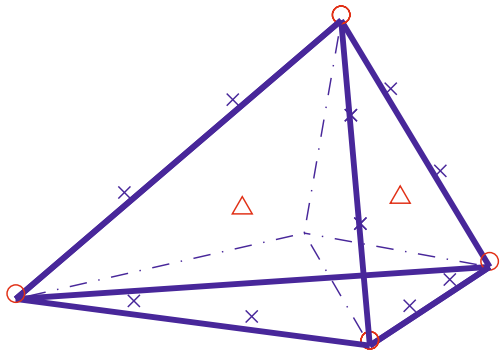

Multivariate Splines and Their Applications, Figure 15
Domain points (on the third layer) on the Alfeld split of a tetrahedron

(4) For each tetrahedron $t = \langle u, v, w, x\rangle \in \Delta$, let m_t be the center of t and $t_1 = \langle u, v, w, m_t\rangle$, $t_2 = \langle v, w, x, m_t\rangle$, $t_3 = \langle w, u, x, m_t\rangle$, and $t_4 = \langle u, v, x, m_t\rangle$ be four tetrahedra in Δ_{A} contained in t. Let

$$S_t := \{\xi^{t_1}_{1,1,1,2}, \xi^{t_2}_{1,1,1,2}, \xi^{t_3}_{1,1,1,2}, \xi^{t_4}_{1,1,1,2}\}\,.$$

Then it can be shown (cf. [33]) that

$$\Gamma := \bigcup_{v\in\Delta} S_v \cup \bigcup_{e\in\Delta} S_e \cup \bigcup_{f\in\Delta} S_f \cup \bigcup_{t\in\Delta} S_t$$

is a minimal determining set for $S_5^{1,2,3}(\Delta_{\mathrm{A}})$, and

$$\dim S_5^{1,2,3}(\Delta_{\mathrm{A}}) = 10V + 2E + 3F + 4T\,,$$

where V, E, F, T denote the number of all vertices, edges, faces, and tetrahedra of Δ. Also, there exists a locally supported basis for $S_5^{1,2,3}(\Delta_{\mathrm{A}})$.

Next we describe an interpolatory scheme using this C^1 quintic macro-element. To this end, we need the following (cf. [42] for a detailed proof).

Lemma 5 *Let $T = \langle v_1, v_2, v_3, v_4\rangle$ be a tetrahedron and v_5 be another point which does not lie on any of the four planes each of which is spanned by one of the faces of T. Given $f_{i,\alpha}, |\alpha| \le 1, i = 1, 2, 3, 4, 5$, there exists a unique cubic polynomial p satisfying the following interpolation conditions:*

$$D^\alpha p(v_i) = f_{i,\alpha}, i = 1, 2, \ldots, 5\,, \quad |\alpha| \le 1\,.$$

For each edge e, let m_e be the midpoint of e and let $e_{\perp,1}$ and $e_{\perp,2}$ be two directions which are perpendicular to e and are linearly independent to each other. For each face $f = \langle v_1, v_2, v_3\rangle$, let f_1, f_2, f_3 be the three domain points $\{(iv_1 + jv_2 + kv_3)/5, (i, j, k) = (2, 2, 1), (1, 2, 2), (2, 1, 2)\}$ on f. Let n_f be a normal unit vector of f. For each tetrahedron t, let m_t be the center point of t. Our interpolation scheme is as follows: For a function $g \in C^2(\Omega)$, we can find $S_g \in S_5^{1,2,3}(\Delta_{\mathrm{A}})$ satisfying

(1) For each vertex v of Δ,

$$D^\alpha S_g(v) = D^\alpha g(v)\,, \quad \forall |\alpha| \le 2\,;$$

(2) For each edge e of Δ,

$$D_{e_{\perp,i}} S_g(m_e) = D_{e_{\perp,i}} g(m_e)\,, \quad i = 1, 2\,;$$

Here, $D_{e_{\perp,i}}$ denotes the derivative along direction $e_{\perp,i}$, $i = 1, 2$;

(3) For each face f of Δ,

$$D_{n_f} S_g(f_j) = D_{n_f} g(f_j)\,, \quad j = 1, 2, 3\,;$$

Here, D_{n_f} denotes the derivative along direction n_f.

(4) For each tetrahedron t,

$$D^\alpha S_g(m_t) = D^\alpha g(m_t)\,, \quad \forall |\alpha| \le 1\,.$$

We use the interpolation conditions (1)–(3), C^2 smoothness conditions at the vertices of Δ, C^1 smoothness conditions around the edges and across the faces of Δ to determine the coefficients of S_g. Indeed, the interpolation conditions (1) and C^2 smoothness conditions at the vertices of Δ determine the coefficients of S_g whose domain points are in $\mathcal{D}_2^t(v)$ for each tetrahedron $t \in \Delta_{\mathrm{A}}$ which shares vertex v for all vertices of Δ. We use the interpolation conditions (2) and C^1 smoothness conditions to determine the coefficients whose domain points in $\mathcal{E}_1^t(e)$ for each tetrahedron $t \in \Delta_{\mathrm{A}}$ which shares edge e for all edges of Δ. We then use the interpolation conditions (3) and C^1 smoothness conditions to determine the coefficients whose domain points in $\mathcal{F}_1^t(f)$ for each tetrahedron $t \in \Delta_{\mathrm{A}}$ which shares face f for all faces of Δ. To determine the remaining coefficients of S_g, we consider S_g restricted on tetrahedron $t \in \Delta$. There are four interior edges inside t connecting to the center m_t of t. By the C^1 smoothness conditions around each of the four interior edges, we obtain the coefficients whose associated domain points are in the collection of

$$\{\xi_{i,j,k,2}^{t\ell}, i + j + k = 3\}\backslash\xi_{1,1,1,2}^{t\ell}, \ell = 1, 2, 3, 4$$

as well as

$$\{\xi_{2,0,0,3}^{t\ell}, \xi_{0,2,0,3}^{t\ell}, \xi_{0,0,2,3}^{t\ell}, \ell = 1, 2, 3, 4\}\,,$$

where $t_\ell, \ell = 1, 2, 3, 4$ denote the four subtetrahedra of t. Note that four of them in the first of the above two groups have already been determined by the interpolation conditions in (1). The coefficients just determined in the previous sentence can be converted to the function and first order derivative values at four vertices of t. Together with the interpolation conditions in (4), we can apply Lemma 5 to get a unique cubic polynomial p_f satisfying the interpolation conditions at the five vertices. We then find the coefficients of p_f over four subtetrahedra t_ℓ's and use these coefficients for the remaining coefficients of S_g. It follows that S_g is C^3 at m_t. Thus, we know $S_g \in S_5^{1,2,3}(\Delta_{\mathrm{A}})$. In particular, $S_g = g$ for each polynomial g of degree ≤ 5.

C^1 Cubic Splines over Worsey–Farin Refinement of Tetrahedral Partitions In [54], Worsey and Farin split each tetrahedron of Δ even further than Alfeld did. That is, they first split tetrahedron t at its center m_t of the inscribed sphere of t into four subtetrahedra by triangular planes each of which consists of one edge of t and m_t. Then for each interior face shared by two tetrahedra t and t', they connect m_t and $m_{t'}$ by a line which intersects the common face f at n_f and split the subtetrahedron of t containing f into three subsubtetrahedra using the triangular

planes each of which consists of n_f, m_t and one vertex of f. Similarly, they split the subtetrahedron of t' containing f into three subtetrahedra in the same way. Since m_t and $m_{t'}$ are the centers of the inscribed spheres, it is always true that n_f is strictly inside the face f. For a boundary face f, they split the subtetrahedra of t containing f into three subsubtetrahedra at the center n_f of f using triangular planes each of which consists of n_f, m_f, and one vertex of f. This is another 3D generalization of the well-known Clough–Tocher refinement of triangulation. Let us use Δ_{WF} to denote such a refinement.

In [54], locally supported spline functions in $S_3^1(\Delta_{\mathrm{WF}})$ were constructed. In our notation, we specify the following two domain point sets to be formed into a minimal determining set:

(1) For each vertex v of Δ, choose a tetrahedron t_v of Δ_{WF} containing v. Let

$$S_v := \mathcal{D}_1^{t_v}(v)\,.$$

(2) For each edge e of Δ, choose a tetrahedron t_e of Δ_{WF} containing e. Writing $e = \langle u, v\rangle$, let

$$S_e := \mathcal{E}_1^{t_e}(e)\Big\backslash\left(\mathcal{D}_1^{t_e}(u) \cup \mathcal{D}_1^{t_e}(v)\right)\,.$$

Then letting $\Gamma = \cup_{t\in\Delta} S_v \bigcup \cup_{e\in\Delta} S_e$, one can prove that Γ is a minimal determining set for $S_3^1(\Delta_{\mathrm{WF}})$.

Then the dimension of $S_3^1(\Delta_{\mathrm{WF}})$ is

$$\dim S_3^1(\Delta_{\mathrm{WF}}) = 4V + 2E\,.$$

There exists a locally supported basis for $S_3^1(\Delta_{\mathrm{WF}})$. Furthermore, each spline function in $S_3^1(\mathrm{WF}(\Delta))$ is in C^2 at m_t for all $t \in \Delta$.

C^1 Quadratic Splines on a Trivariate Powell–Sabin Refinement In [55], Worsey and Piper refined each tetrahedron in Δ even further than [54] to construct C^1 spline functions using piecewise quadratic polynomials. For each tetrahedron t, let v_t be a point inside t, say, the barycenter of t. For each face f of t, let v_f be a point inside f, say the barycenter of f. For each edge e of t, let v_e be a point inside e, say the middle point of e. Worsey and Piper split t into 24 subtetrahedra by first splitting t into four subtetrahedra by triangular faces spanned by v_t and two vertices of t and then splitting each subtetrahedron into six subsubtetrahedra by triangular faces spanned by v_t, v_f, v_e as well as v_t, v_f and one of the vertices of f.

The tetrahedral partition has to satisfy some stringent conditions so that their scheme will yield a C^1 spline function. The stringent conditions are (1) for two tetrahedra t, t' which share a common face f, $v_t, v_f, v_{t'}$ are on the same line; and (2) for tetrahedra $t_1, \ldots, t_m$ which share a common edge e, $v_{t_1}, \ldots, v_{t_m}$ and v_e are on the same plane. For convenience, let us call these conditions Worsey–Piper conditions.

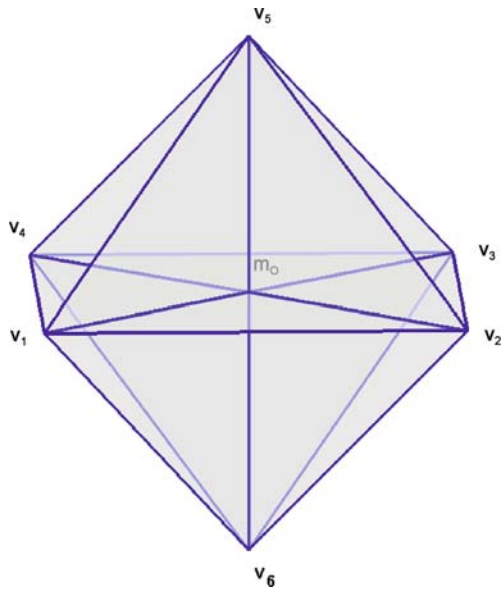

Multivariate Splines and Their Applications, Figure 16
An octahedron and its tetrahedral partition

For a general tetrahedral partition, the conditions can not be easily satisfied and hence, their construction will not be in C^1 globally. This leaves an open question how to make such stringent conditions satisfy. Recently, Schumaker, Sorokina, and Worsey overcame this difficulty by further refining the Worsey–Pipe refinement. They split a tetrahedron into more than 500 subtetrahedra so that they can construct C^1 quadratic macro-elements without using the stringent conditions (cf. [49]).

Construction of C^1 Quintic Macro-Element on Octahedral Refinement Let $O = \langle v_1, v_2, \ldots, v_6\rangle$ be an octahedron such that the three diagonals of O intersect at a common point m_O inside O as shown in Fig. 16. In this case, v_1, v_2, v_3, v_4 are coplanar. So are v_2, v_4, v_5, v_6 and v_1, v_3, v_5, v_6. In this subsection we shall restrict our attention to such tetrahedra which will be called *centrical octahedra*. It is possible to partition some common 3D solids into a collection of centrical octahedra. (See Examples in [33]). By adding the three planes $\langle v_1, v_2, v_3, v_4\rangle$, $\langle v_1, v_3, v_5, v_6\rangle$, and $\langle v_2, v_4, v_5, v_6\rangle$, we obtain eight tetrahedra in O. Let $\oplus$ be the collection of these eight tetrahedra.

We shall use piecewise trivariate polynomials of degree 5 over $\oplus$ to construct C^1 macro-elements that satisfy the following smoothness properties: It is C^1 over the union of the eight tetrahedra of $\oplus$ and C^2 at the six vertices of O. Here, a function s is said to be C^2 at a vertex v if s is twice differentiable at v. We denote by $S_5^{1,2}(\oplus)$ the space of all spline functions in $S_5^{-1}(\oplus)$ which are C^1 across each interior triangular face of $\oplus$ and C^2 at vertices of O. We specify the following subsets to be formed into a minimal determining set Γ for $S_5^{1,2}(\oplus)$:

(1) For each vertex $v \in \oplus$ except for m_O, let $t_v \in \oplus$ be a tetrahedron having v as one of its vertices. Let $S_v := \mathcal{D}_2^{t_v}(v)$. We note that in terms of traditional nodal values, the determination of the B-coefficients which are associated with domain points in S_v of any spline function s is equivalent to the assignments of $\partial^\alpha s(v)$ for all $|\alpha| \le 2$, where $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ with $|\alpha| := \alpha_1 + \alpha_2 + \alpha_3$ and

$$\partial^\alpha s(v) := \left(\frac{\partial}{\partial x}\right)^{\alpha_1} \left(\frac{\partial}{\partial y}\right)^{\alpha_2} \left(\frac{\partial}{\partial z}\right)^{\alpha_3} s(v) .$$

(2) For each boundary edge $e \in \oplus$, let $t_e \in \oplus$ be a tetrahedron containing e. Writing $e = \langle u, v\rangle$, let

$$S_e := \mathcal{E}_1^{t_e}(e) \backslash \left(\mathcal{D}_2^{t_e}(u) \cup \mathcal{D}_2^{t_e}(v)\right) .$$

We note that the determination of the B-coefficients associated with domain points in S_e of any spline function s may be replaced by the assignment of two normal derivatives of e that are perpendicular to each other at the midpoint of e if the traditional nodal values are used.

(3) For each boundary face $f \in \oplus$, let $t_f \in \oplus$ be a tetrahedron with f as one of its faces. Writing $f = \langle u_1, u_2, u_3\rangle$ and $t_f = \langle u_1, u_2, u_3, u_4\rangle$, let

$$S_f := \{\xi_{2,1,1,1}^{t_f}, \xi_{1,2,1,1}^{t_f}, \xi_{1,1,2,1}^{t_f}\} .$$

In terms of nodal values, the determination of the B-coefficients associated with the domain points in S_f of spline functions is the same as the assignment of the normal derivative of f at the three locations whose barycentric coordinates are $(2/5, 2/5, 1/5, 0)$, $(1/5, 2/5, 2/5, 0)$, $(2/5, 1/5, 2/5, 0)$ with respect to t_f. Here, for convenience, we have arranged that the last index ℓ of B-coefficients $c_{i,j,k,\ell}^t$ is associated with m_O.

(4) For m_O, let $t_{O,n}$, $n = 1, \ldots, 8$ be the 8 tetrahedra of $\oplus$. For convenience, we arrange that the last index ℓ of B-coefficients $c_{i,j,k,\ell}^{t_{O,n}}$ of any spline function is associated with m_O. Let $S_O := \{\xi_{1,1,1,2}^{t_{O,n}}, n = 1, \ldots, 8\}$.

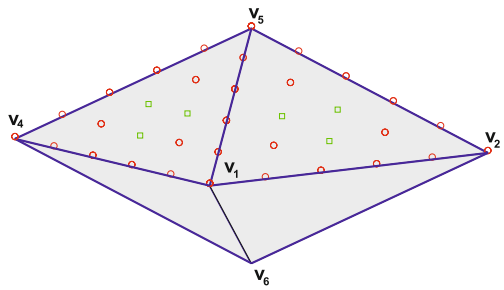

Multivariate Splines and Their Applications, Figure 17
Domain points (on the first layer) on the top half of an octahedron

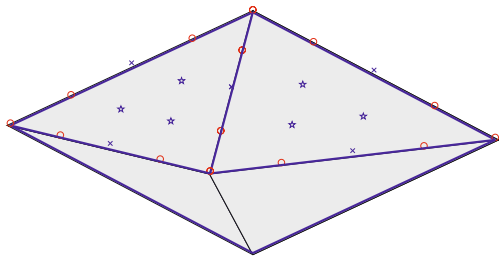

Multivariate Splines and Their Applications, Figure 18
Domain points (on the second layer) on the top half of an octahedron

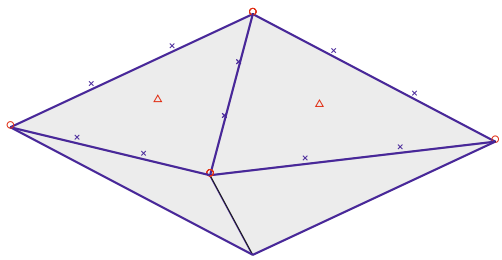

Multivariate Splines and Their Applications, Figure 19
Domain points (on the third layer) on the top half of an octahedron

We can show that the set

$$\Gamma := \bigcup_{v \in O} S_v \cup \bigcup_{e \in O} S_e \cup \bigcup_{f \in O} S_f \cup S_O$$

is a minimal determining set for $S_5^{1,2}(\oplus)$ and

$$\dim(S_5^{1,2}(\oplus)) = 10 \times 6 + 2 \times 12 + 3 \times 8 + 8 = 116 .$$

Next we let Δ be a collection of centrical octahedra O_i, $i = 1, \ldots, N$. Suppose that Δ is regular in the sense that the intersection of any two octahedra O_i and O_j is either an empty set, or their common face, or their common

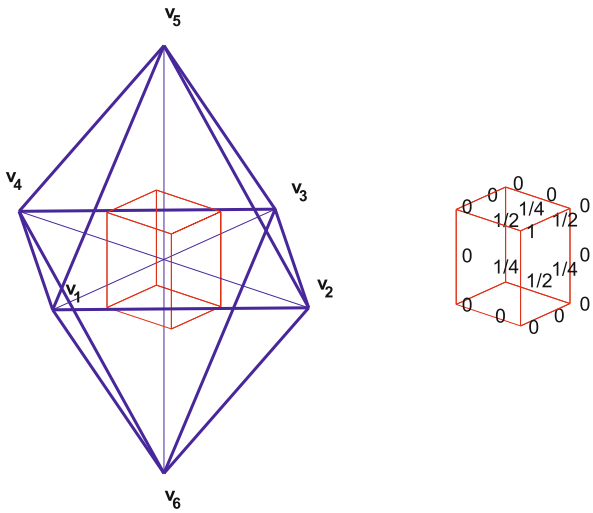

Multivariate Splines and Their Applications, Figure 20
Coefficients of spline function $s_1 \in S_5^{1,2}(\oplus)$

edge, or their common vertex. Let $\oplus_i$ be the eight tetrahedra obtained from O_i as described above, and let

$$\Delta_{LL} = \cup_{i=1}^{N} \oplus_i$$

be the corresponding special tetrahedral partition. The above construction of C^1 quintic macro-element can be applied to such special tetrahedral partitions. We get

$$\dim S_5^{1,2}(L(\Delta)) = 10V + 2E + 3F + 8N \,,$$

where V, E, F denote the numbers of all vertices, edges, and faces of $\Delta = \{O_i, i = 1, \ldots, N\}$.

Multivariate Splines for Scattered Data Fitting

Given a set of scattered data, e.g., $\{(x_i, y_i, z_i), i = 1, \ldots, N\}$, we want to find a smooth function or surface S such that

$$S(x_i, y_i) = z_i \,, \quad i = 1, \ldots, N \,,$$

if z_i are very accurate measurements or

$$S(x_i, y_i) \approx z_i \,, \quad i = 1, \ldots, N \,,$$

if z_i are subject to some random noises. Note that nowadays, N is usually very large. Three key requirements are (1) S must be a smooth surface; (2) S resembles the shape of the data; and (3) S can be efficiently computed. In this article we explain how to use multivariate splines for solving scattered data fitting problems. We restrict ourselves to the bivariate setting. For spherical setting, we refer to [6].

The following methods for fitting a given set of data are available in the literature.

- Minimal Energy Method;
- Discrete Least Squares Method;
- Penalized Least Squares Spline Method;
- L_1 Spline Method;
- Least Absolute Deviation Method;
- L_1 Smoothing Spline Method.

Let us review these methods by explaining several fundamental questions concerning each method: if a method has a solution or not (i. e., the existence and uniqueness), how to compute that solution (i. e., numerical algorithms), whether the solution surface resembles the given data (i. e., approximation properties), and what to do when the amount of data is very large.

Minimal Energy Method

Let $E(f)$ be the thin-plate energy functional

$$E(f) = \int_\Omega \left(\left(\frac{\partial^2}{\partial x^2} f \right)^2 + 2 \left(\frac{\partial^2}{\partial x \partial y} f \right)^2 + \left(\frac{\partial^2}{\partial y^2} f \right)^2 \right) \mathrm{d}x\mathrm{d}y \,.$$

Let $\Lambda(f) = \{s \in S_d^r(\Delta), s(x_i, y_i) = f_i, i = 1, \ldots, N\}$. Find $S_f \in \Lambda(f)$ such that

$$E(S_f) = \min\{E(s), \quad s \in \Lambda(f)\} \,.$$

The following result was proved in [53] and in [5] by different methods.

Theorem 14 *If $\Lambda(f)$ is not empty, there exists a unique interpolatory spline in $S_d^r(\Delta)$.*

Once we have an interpolatory surface, we would like to know how the surface resembles the given data. Let $W_\infty^2(\Omega)$ be the Sobolev space of all functions whose second derivatives are essentially bounded over Ω. $|f|_{2,\infty,\Omega}$ is the maximal norm of all second-order derivatives of f over Ω. The following results can be found in [53]

Theorem 15 *Suppose $z_i = f(x_i, y_i)$, $i = 1, \ldots, N$, for $f \in W_\infty^2(\Omega)$. Let $d \geq 3r + 2$, and let Δ be a triangulation of the data sites $\{(x_i, y_i), i = 1, \ldots, N\}$. Then*

$$\|s_f - f\|_{L_\infty(\Omega)} \leq C|\Delta|^2 |f|_{2,\infty,\Omega} \,.$$

Our next concern is how to compute interpolatory minimal energy splines using a spline space of arbitrary degree d and arbitrary smoothness r with $d \geq 3r + 2$. The following computational scheme was described in [5]. See [48] for another computational scheme.

(1) Express each $s \in S_d^{-1}(\Delta)$ in B-form, i. e.,

$$s(x,y)|_t = \sum_{i+j+k=d} c_{ijk}^t B_{ijk}^{d,t}(x,y)\,,$$

where $B_{ijk}^{d,t}$ are Bernstein–Bézier basis functions defined only on t. Let $\mathbf{c} = (c_{ijk}^t, i+j+k=d, t \in \Delta)$ be a coefficient vector for s.

(2) When $s \in S_d^r(\Delta)$, there are smoothness conditions over interior edges of Δ (cf. [21]). The smoothness conditions are linear. Put all smoothness conditions together to write

$$\mathcal{H}\mathbf{c} = 0\,,$$

for a matrix $\mathcal{H}$. I. e., $s \in S_d^r(\Delta)$ if $\mathcal{H}\mathbf{c} = 0$.

(3) Compute the energy functional $E(s) = \mathbf{c}^\mathsf{T}\mathcal{E}\mathbf{c}$ for an energy matrix $\mathcal{E}$ which is a diagonal block matrix.

(4) The interpolatory conditions can be written $\mathcal{I}\mathbf{c} = \mathbf{f}$ for a matrix $\mathcal{I}$ and a vector $\mathbf{f}$ containing all data values z_i.

(5) The minimal energy method for interpolatory splines is equivalent to finding $\mathbf{c}$ such that

$$\min\left\{\mathbf{c}^\mathsf{T}\mathcal{E}\mathbf{c},\ \text{ subject to } \mathcal{H}\mathbf{c} = 0,\ \mathcal{I}\mathbf{c} = \mathbf{f}\right\}.$$

(6) By the Lagrange multipliers method, we solve

$$\begin{bmatrix} \mathcal{E} & \mathcal{H}^\mathsf{T} & \mathcal{I}^\mathsf{T} \\ \mathcal{H} & 0 & 0 \\ \mathcal{I} & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \mathbf{f} \end{bmatrix}.$$

(7) To solve this system, we use the following iteration method introduced in [5]

$$\left(\mathcal{E} + \frac{1}{\epsilon}\begin{bmatrix}\mathcal{H}^\mathsf{T} & \mathcal{I}^\mathsf{T}\end{bmatrix}\begin{bmatrix}\mathcal{H} \\ \mathcal{I}\end{bmatrix}\right)\mathbf{c}^{(1)} = \frac{1}{\epsilon}\mathcal{I}^\mathsf{T}\mathbf{f}\,,$$

$$\left(\mathcal{E} + \frac{1}{\epsilon}\begin{bmatrix}\mathcal{H}^\mathsf{T} & \mathcal{I}^\mathsf{T}\end{bmatrix}\begin{bmatrix}\mathcal{H} \\ \mathcal{I}\end{bmatrix}\right)\mathbf{c}^{(k+1)} = \mathcal{E}\mathbf{c}^{(k)} + \frac{1}{\epsilon}\mathcal{I}^\mathsf{T}\mathbf{f}\,,$$

for $k = 1, 2, \ldots$ and $\epsilon > 0$, e. g., $\epsilon = 10^{-6}$.

We need to show that the iterative method above is convergent. To this end, recall that a matrix A is positive definite with respect to B if $\mathbf{c}^\mathsf{T}A\mathbf{c} \geq 0$ and if $A\mathbf{c} = 0$ and $B\mathbf{c} = 0$ for some $\mathbf{c}$, then $\mathbf{c} = 0$. In [3] we proved the following (cf. [5] for a similar result).

Theorem 16 *Suppose that $\mathcal{E}$ is positive definite with respect to $[\mathcal{H}, \mathcal{I}]^T$. Then the above iteration converges and*

$$\|\mathbf{c}^{(k+1)} - \mathbf{c}\| \leq C\epsilon^k\,, \quad \forall k \geq 1\,.$$

In Figs. 21 and 22 we show a given set of scattered data and the surface of a C^1 quintic spline interpolant.

In Fig. 23 we show a given set of scattered data and in Fig. 24 we show the surface of a C^1 quintic spline interpolant.

When the number of data sites is large, e. g., $N > 1000$, a computer may not be powerful enough to solve the linear system. A domain decomposition technique for computing an approximation of the minimal energy spline interpolation was proposed in [35]. The ideas of domain decomposition for scattered data fitting can be explained as follows.

Let $D_1(t)$ be the union of all triangles in Δ which share a vertex or edge with t, and $D_{k+1}(t)$ the union of all triangles sharing a vertex or edge with triangles in $D_k(t)$. For $k \geq 1$, we compute a minimal energy interpolatory spline $S_{f,t,k} \in \Lambda(f)$ such that

$$E_{D_k(t)}(S_{f,t,k}) = \min\{E_{D_k(t)}(s), s \in \Lambda(f|_t)\}\,,$$

$$E_{D_k(t)}(s) = \int_{D_k(t)} \left(\left(\frac{\partial^2}{\partial x^2}f\right)^2 + 2\left(\frac{\partial^2}{\partial x \partial y}f\right)^2 + \left(\frac{\partial^2}{\partial y^2}f\right)^2\right).$$

The following result was established in [35].

Theorem 17 *Suppose that $f \in C^2(\Omega)$. For $d \geq 3r+2$, there is a $0 < \rho < 1$ such that*

$$\|S_f - S_{f,t,k}\|_{L_\infty(t)} \leq C\rho^k|\Delta||f|_{2,\infty,\Omega}$$

for $k \geq 1$, where C is a constant dependent on d, β and the area of Ω.

This result shows that a (global) minimal energy spline interpolation S_f can be approximated by local minimal energy spline interpolations $S_{f,t,k}$ for all $t \in \Delta$. That is, for each triangle t, one can use a local minimal energy spline interpolation $S_{f,t,k}$ to replace the global one $S_f|_t$ within some tolerance.

Discrete Least Squares Fitting

The discrete least squares method is one of the classical methods for data fitting. Instead of polynomial fitting, we use multivariate splines.

Let $\ell(f) = \sum_{i=1}^N |f(x_i, y_i)|^2$. We look for $S_f \in S_d^r(\Delta)$ such that

$$\ell(S_f - f) = \min\left\{\ell(s-f)\,, s \in S_d^r(\Delta)\right\}.$$

S_f is called the discrete least squares fit of the given data $\{(x_i, y_i, f_i), i = 1, \ldots, N\}$ with $f_i = f(x_i, y_i)$.

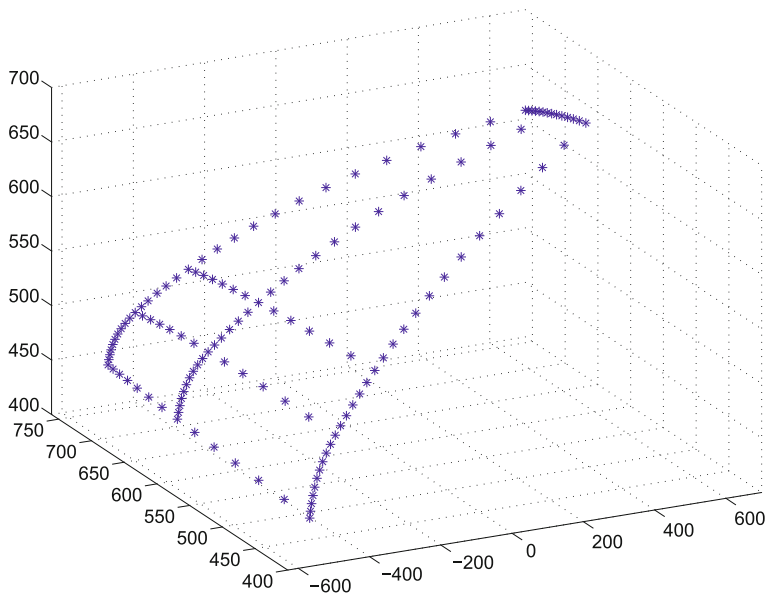

Multivariate Splines and Their Applications, Figure 21
Data points (courtesy Gerald Farin)

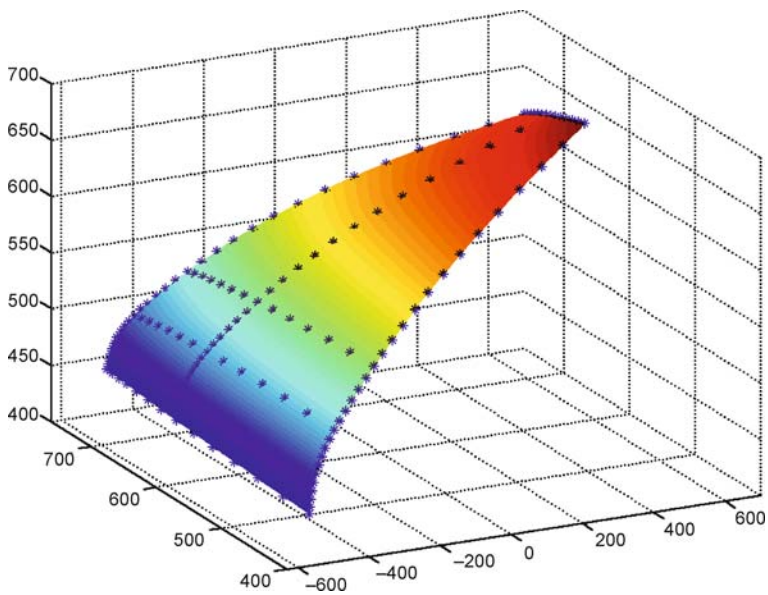

Multivariate Splines and Their Applications, Figure 22
A C^1 quintic spline fit of the data in Fig. 21

To show the existence and uniqueness of the solution S_f, we need to assume

$$A_1 \|s\|_{L_\infty(T)} \le \sqrt{\sum_{(x_i,y_i)\in T} |s(x_i, y_i)|^2}$$

for all $s \in S_d^r(\Delta)$ and all triangle $T \in \Delta$ (cf. [51]).

Theorem 18 *Suppose that the above constant A_1 is strictly positive. Then there exists a unique spline fit $S_f \in S_d^r(\Delta)$.*

Let

$$\sqrt{\sum_{(x_i,y_i)\in T} |s(x_i, y_i)|^2} \le A_2 \|s\|_{L_\infty(T)}$$

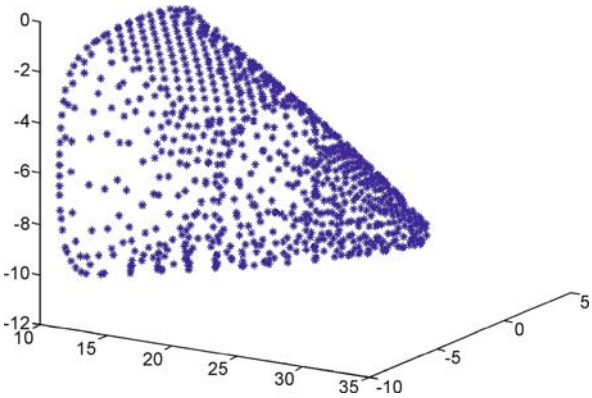

Multivariate Splines and Their Applications, Figure 23
Data points (courtesy Tom Grandine)

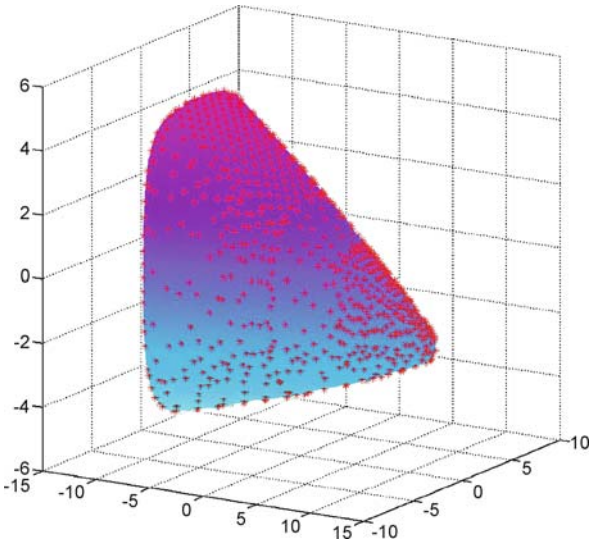

Multivariate Splines and Their Applications, Figure 24
A C^1 quintic spline fit of the data in Fig. 23

for all $T \in \Delta$ and $s \in S_d^r(\Delta)$. It is easy to see that A_2 must be less than or equal to the maximal number of points per triangle. The following result was established in [51].

Theorem 19 *Assume that $f \in W_\infty^{m+1}(\Omega)$. Then*

$$\|S_f - f\|_{L_\infty(\Omega)} \le C\frac{A_2}{A_1}|\Delta|^{m+1}|f|_{m+1,\infty,\Omega}$$

for a constant C dependent on β, d.

Furthermore, we can show the following

Corollary of Theorem 6 *Under the same assumptions above, for $|\alpha| \le m+1$,*

$$\|D^\alpha(S_f - f)\|_{L_\infty(\Omega)} \le C\frac{A_2}{A_1}|\Delta|^{m+1-|\alpha|}|f|_{m+1,\infty,\Omega}$$

for a constant C dependent only on β and d.

This can be proved by using a polynomial approximation property and Markov's inequality. Details are omitted here.

Our next question is how to compute discrete least squares fits. Recall that we write each $s \in S_d^{-1}(\Delta)$ in the B-form

$$s(x,y)|_t = \sum_{i+j+k=d} c_{ijk}^t B_{ijk}^{d,t}(x,y)$$

with coefficient vector $\mathbf{c} = (c_{ijk}^t, i+j+k=d, t \in \Delta)$.

We put all smoothness conditions of $S_d^r(\Delta)$ together as

$$\mathcal{H}\mathbf{c} = 0\,.$$

Let $\mathcal{L}$ be an observation matrix. It is easy to see

$$\ell(s-f) = \mathbf{c}^\mathsf{T}\mathcal{L}\mathcal{L}^\mathsf{T}\mathbf{c} - 2\mathbf{c}^\mathsf{T}\mathcal{L}\mathbf{f} + \mathbf{f}^\mathsf{T}\mathbf{f}\,.$$

The discrete least squares spline is the solution of

$$\min\{\mathbf{c}^\mathsf{T}\mathcal{L}\mathcal{L}^\mathsf{T}\mathbf{c} - 2\mathbf{c}^\mathsf{T}\mathcal{L}\mathbf{f},\ \text{subject to } \mathcal{H}\mathbf{c} = 0\}\,.$$

By the Lagrange multipliers method, we solve

$$\begin{bmatrix} \mathcal{L}\mathcal{L}^\mathsf{T} & \mathcal{H}^\mathsf{T} \\ \mathcal{H} & 0 \end{bmatrix}\begin{bmatrix} \mathbf{c} \\ \alpha \end{bmatrix} = \begin{bmatrix} \mathcal{L}\mathbf{f} \\ 0 \end{bmatrix}\,.$$

The ALW iteration introduced in the previous subsection can be applied to solve the above linear system. As before the iterative solutions converge the exact solution. Numerical examples are omitted here to save some space. See [48] for another approach of the computation.

When the number of data sites is large, especially when the number of triangles is large, a computer may not be powerful enough to solve the associated linear system. We again propose a domain decomposition technique for computing an approximation of the discrete least squares spline (cf. [35]). That is, for $k \ge 1$, we compute $S_{f,t,k}$ such that

$$\ell_{D_k(t)}(S_{f,t,k} - f) = \min\left\{\ell_{D_k(t)}(s-f), s \in S_d^r(\Delta)\right\}\,,$$

$$\ell_{D_k(t)}(s-f) = \sum_{(x_i,y_i)\in D_k(t)} |s(x_i,y_i) - f(x_i,y_i)|^2\,.$$

We have the following (cf. [35])

Theorem 20 *Let $S_d^r(\Delta)$ be a spline space with $d \ge 3r+2$ over a β quasi-uniform triangulation Δ. Suppose that data values are obtained from a continuously differentiable function $f \in C^{m+1}(\Omega)$ with $m \ge 0$. Suppose that $A_1 > 0$ and $A_2 < \infty$ are constants such that A_2/A_1 is independent of Δ. Then there is a positive $\rho < 1$ such that*

$$\|s_f - S_{f,k}\|_{L_\infty(t)} \le C\rho^k|\Delta|^m|f|_{m+1,\infty,\Omega}$$

for $k \ge 1$, where C is a constant dependent only on d, β, A_2/A_1, and the area of Ω.

Penalized Least Squares Spline Method

Recall that $E(f)$ denotes a thin-plate energy functional of f and $\ell(s) = \sum_{i=1}^{N}(s(x_i, y_i))^2$ as before. Fix $\lambda > 0$. Define $\mathcal{P}_\lambda(s) = \ell(s - f) + \lambda E(s)$. The PLS spline is the minimization solution $S_{f,\lambda} \in S_d^r(\Delta)$ such that

$$\mathcal{P}_\lambda(S_{f,\lambda}) = \min\left\{\mathcal{P}_\lambda(s), s \in S_d^r(\Delta)\right\}.$$

We refer to [5] for a proof of the following

Theorem 21 *Suppose that $N \geq 3$ and there exist three data sites, say, (x_i, y_i), $i = 1, 2, 3$ which are not colinear. Then there exists a unique $S_{f,\lambda}$ in $S_d^r(\Delta)$ solving the above minimization problem.*

We certainly want to know if the penalized least squares fitting surface resembles the given data or not. Since $f - S_{f,\lambda} = f - S_{f,0} + S_{f,0} - S_{f,\lambda}$ and $f - S_{f,0}$ was estimated in the previous subsection, we need to estimate $S_{f,0} - S_{f,\lambda}$. To do so, we introduce the following two quantities: (cf. [52])

$$K_1 = \sup\left\{\frac{E(s)^{1/2}}{\ell(s)^{1/2}}, s \in S_d^r(\Delta), s \neq 0\right\}$$

and

$$K_2 = \sup\left\{\frac{\|s\|_{L_\infty(\Omega)}}{\ell(s)^{1/2}}, s \in S_d^r(\Delta), s \neq 0\right\}.$$

Then in [52], von Golitschek and Schumaker proved the following

Theorem 22 *Let $S_{f,\lambda}$ be the PLS spline in $S_d^r(\Delta)$ with $d \geq 3r + 2$. Assume that K_1 and K_2 are finite. Then*

$$\|S_{f,\lambda} - S_{f,0}\|_{L_\infty(\Omega)} \leq K_2\sqrt{\lambda E(S_{f,0})}\ \min\left\{1, K_1\sqrt{\lambda}\right\}.$$

When λ is small enough, e. g., $K_1\sqrt{\lambda} < 1$, we see that the convergence is linear in λ. Let us present some numerical evidence. Consider a type-I triangulation Δ with 289 vertices and 512 triangles of $[0, 1] \times [0, 1]$. Let $f(x, y)$ be the well-known Franke function. We use domain points of degree 5 of Δ as given sample locations and $f(x, y)$ at these locations as functional values. We compute $S_{f,\lambda} \in S_5^1(\Delta)$ for $\lambda_i = 1/2^{20+i}$ and $S_{f,0} \in S_5^1(\Delta)$ and use the maximum difference of $S_{f,\lambda} - S_{f,0}$ at 100×100 equally spaced points of $[0, 1] \times [0, 1]$ to approximate $\|S_{f,\lambda} - S_{f,0}\|_\infty$. For $\lambda_i = 1/2^{10+i}$, the maximum errors of $S_{f,\lambda_i} - S_{f,0}$ are given below:

	λ_2	λ_3	λ_4	λ_5	λ_6
$S_5^1(\Delta)$	1.325(–4)	6.981(–5)	3.653(–5)	1.876(–5)	9.517(–6)
Rates		0.5267	0.5232	0.5134	0.5072

From the numerical values above, we can see that the errors decay roughly by half each time and thus, the convergence order of λ is 1. We repeat the experiment for test functions $f(x, y) = 2x^4 + 5y^4$ and $f(x, y) = \sin(\pi(x^2 + 2y^2))$. The convergence rates for both test functions are similar to the above.

We now work on estimating K_1 and K_2. It is easy to get

$$E(s) \leq \sum_{T\in\Delta} A_T\|s\|_{2,\infty,T}^2 \leq \sum_{T\in\Delta}\frac{A_T}{\rho_T^4}\|s\|_{L_\infty(T)}^2 \leq \frac{\beta^2}{(\rho_\Delta)^2}\frac{\ell(s)}{A_1^2}.$$

It follows that

$$K_1 \leq \frac{\beta}{A_1\rho_\Delta}. \tag{8}$$

Since $\|s\|_{L_\infty(\Omega)} = \|s\|_{L_\infty(T)}$ for a triangle T,

$$\|s\|_{L_\infty(\Omega)} \leq \frac{1}{A_1}\sqrt{\sum_{(x_i,y_i)\in T}|s(x_i, y_i)|^2} \leq \frac{1}{A_1}\ell(s)^{1/2}.$$

It follows that

$$K_2 \leq \frac{1}{A_1}. \tag{9}$$

With the above preparation we can prove

Theorem 23 *Suppose that $f \in W_\infty^{m+1}(\Omega)$ with $1 \leq m \leq d$. Let $S_{f,\lambda}$ be the PLS spline off in $S_d^r(\Delta)$ with $d \geq 3r + 2$. Then*

$$\|S_{f,\lambda} - f\|_{L_\infty(\Omega)} \leq C_1|\Delta|^{m+1}|f|_{m+1,\infty,\Omega} + \frac{C_2\sqrt{\lambda}}{A_1}\min\left\{1, \frac{\beta\sqrt{\lambda}}{A_1\rho_\Delta}\right\}|f|_{2,\infty,\Omega},$$

where $C_1 > 0$, $C_2 > 0$ are constants dependent on A_2/A_1, β, and the area of Ω.

As we can see the condition for the existence of penalized least squares spline fits is much weaker than that for the existence of the discrete least squares spline fits. However, the approximation result on penalized least squares spline fits is dependent on a very strong condition on the data sites, i. e., $A_1 > 0$. It is interesting to see if one can remove this condition while proving that the penalized least squares fits resemble the shape of the data.

Next we briefly explain how to compute PLS splines. Recall $\mathbf{c}$ is the coefficient vector of a spline $s \in S_d^{-1}(\Delta)$, $\mathcal{H}$ is the smoothness matrix such that $\mathcal{H}\mathbf{c} = 0$ if and only if $s \in S_d^r(\Delta)$, $\mathcal{E}$ is the energy matrix, and $\mathcal{L}$ is the observation

matrix. Then the PLS spline is the minimization solution

$$\min\{\mathbf{c}^{\mathsf{T}}\mathcal{L}\mathcal{L}^{\mathsf{T}}\mathbf{c} - 2\mathbf{c}^{\mathsf{T}}\mathcal{L}\mathbf{f} + \lambda\mathbf{c}^{\mathsf{T}}\mathcal{E}\mathbf{c},\ \text{subject to } \mathcal{H}\mathbf{c} = 0\}\,.$$

By the Lagrange multipliers method, we solve

$$\begin{bmatrix} \mathcal{L}\mathcal{L}^{\mathsf{T}} + \lambda\mathcal{E} & \mathcal{H}^{\mathsf{T}} \\ \mathcal{H} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \alpha \end{bmatrix} = \begin{bmatrix} \mathcal{L}\mathbf{f} \\ 0 \end{bmatrix}.$$

We apply the ALW iteration introduced before.

When the number of triangles is large, a computer may not be powerful enough to find the PLS splines. We use a domain decomposition technique for computing an approximation of the PLS spline (cf. [35]). For $k \geq 1$, we compute a PLS spline $S_{f,t,k}$ such that

$$P_{D_k(t)}(S_{f,t,k}) = \min\left\{P_{D_k(t)}(s)\,, s \in S_d^r(\Delta)\right\},$$

where

$$P_{D_k(t)}(s) = \sum_{(x_i,y_i)\in D_k(t)} |s(x_i,y_i) - f(x_i,y_i)|^2 + \lambda E(s|_{D_k(t)}).$$

Here $D_k(t) = \text{star}^k(t)$ for each triangle $t \in \Delta$. We have the following result (cf. [35])

Theorem 24 *Suppose that $S_d^r(\Delta)$ with $d \geq 3r+2$ over a β quasi-uniform triangulation Δ. Suppose that data values are obtained from a continuously differentiable function $f \in C^{m+1}(\Omega)$. Suppose that $A_1 > 0$ and $A_2 < \infty$ are constants such that A_2/A_1 is independent of Δ. Then there is a positive $\rho < 1$ such that*

$$\|s_f - S_{f,k}\|_{L_\infty(t)} \leq C\rho^k((k+2)^{3/2}|\Delta|^{m+1}|f|_{m+1,\infty,\Omega} + \lambda|f|_{2,\infty,\Omega})$$

for $k \geq 1$, where C is a constant dependent only on d, β and A_2/A_1.

L_1 Spline Methods

L_1 spline methods for data fitting were proposed by Lavery. He used C^1 cubic spline curves and bivariate C^1 cubic Sibson's elements for scattered data in 1D and grid data in 2D, respectively. Lai and Wenston in 2004 generalized the study to the scattered data in the bivariate setting (cf. [40]). Recall that

$$\Lambda(f) = \{s \in S_d^r(\Delta), s(x_i,y_i) = f(x_i,y_i), i = 1,\ldots,N\}.$$

Let $E_1(s)$ be the L_1 energy functional, i. e.,

$$E_1(f) = \int_\Omega \left(\left|\frac{\partial^2}{\partial x^2}f\right| + 2\left|\frac{\partial^2}{\partial x\partial y}f\right| + \left|\frac{\partial^2}{\partial y^2}f\right|\right) \mathrm{d}x\mathrm{d}y\,.$$

Find $S_f \in \Lambda(f)$ such that

$$E_1(S_f) = \min\{E_1(s), \quad s \in \Lambda(f)\}\,.$$

S_f is called the L_1 interpolatory spline of the given data $\{(x_i,y_i,f(x_i,y_i)), i = 1,\ldots,N\}$. A proof of the following theorem can be found in [40]. This can be seen from the fact that the minimization functional is convex. However, the functional is not strictly convex and hence, the solution may not be unique.

Theorem 25 *Suppose that $\Lambda(f)$ is not empty. Then there exists at least one S_f solving the above minimization problem.*

The interpolatory surfaces which minimize the L_1 energy functional are indeed quite different from the usual L_2 minimal energy splines. See Figures in [40] for detail. Mainly, the L_1 spline method reduces greatly the oscillations in the interpolatory surfaces.

It is necessary to show that L_1 interpolatory splines resemble the shape of the given data. Lai in [32] proved the following

Theorem 26 *Suppose that $f \in C^2(\Omega)$. Let S_f be the L_1 interpolatory spline of the given data locations and values $\{(x_i,y_i,f(x_i,y_i)), i = 1,\ldots,N\}$. Then*

$$\|S_f - f\|_{L_1(\Omega)} \leq C|\Delta|^2|f|_{2,\infty,\Omega}\,,$$

for a constant C dependent only on β and d.

For a given data set $\{(x_i,y_i,f(x_i,y_i)), i = 1,\ldots,N\}$, let

$$\ell_1(s) = \sum_{i=1}^N |s(x_i,y_i)|\,.$$

We find $S_f \in S_d^r(\Delta)$ such that

$$\ell_1(S_f - f) = \min\{\ell_1(s-f), s \in S_d^r(\Delta)\}\,.$$

S_f is the least absolute deviation (LAD) from the given data.

Since the minimization functional is convex, there always exist a minimizer S_f (cf. [40]). Next we would like to know how well the LAD surface resembles the given data. Let F_1 and F_2 be positive numbers such that

$$F_1\|s\|_{L_\infty(T)} \leq \sum_{(x_i,y_i)\in T} |s(x_i,y_i)| \leq F_2\|s\|_{L_\infty(T)}$$

for all $s \in S_d^r(\Delta)$ and for all $T \in \Delta$. We have the following (▶ Popular Wavelet Families and Filters and Their Use).

Theorem 27 *Suppose that two constants $F_1 > 0$ and $F_2 < \infty$ such that F_2/F_1 independent of Δ. Suppose that*

$f \in W_\infty^{m+1}(\Omega)$ for $0 \le m \le d$. Then

$$\|S_f - f\|_{L_1(\Omega)} \le C|\Delta|^{m+1}|f|_{m+1,\infty,\Omega}$$

for a positive constant C dependent on F_2/F_1, β and d.

L_1 smoothing splines are $S_f \in S_d^r(\Delta)$ which minimizes

$$\ell_1(S_f - f) + \lambda E_1(S_f) = \min\{\ell_1(s-f) + \lambda E_1(s), s \in S_d^r(\Delta)\}.$$

Since the minimization functional is convex, there exists at least one S_f solving the above minimization. We next need to show that S_f approximates f as the size of the triangulations goes to zero (► Popular Wavelet Families and Filters and Their Use).

Theorem 28 *Under the same assumptions as Theorem 27,*

$$\|S_f - f\|_{L_1(\Omega)} \le C|\Delta|^{m+1}|f|_{m+1,\infty,\Omega} + \lambda \frac{C_f}{F_1}|\Delta|^2$$

for a positive constant C dependent on F_2/F_1, β and d.

Algorithms computing these three L_1 spline methods were discussed in [40]. The main ideas are

(1) Use discontinuous piecewise polynomial functions and set the smoothness conditions as side constraints;
(2) Convert L_1 norm minimization to a linear programming problem;
(3) Use Karmarkar's algorithm to solve the linear programming problem.

Multivariate Splines for Numerical Solution of Partial Differential Equations

In this section, we will show how to solve the Poisson equation and other second-order elliptic equations by using multivariate splines of variable degree and variable smoothness. These spline functions will provide a versatile tool for numerical solution of PDE's because of the flexibility in choosing the degrees and smoothness when constructing numerical solutions. For example, it is known that the weak solution of the Poisson equation over a polygonal domain Ω is at least $H^2(V)$ for any open set $V \subset \Omega$ (cf. [20]). We should choose spline functions that are C^1 inside Ω and C^0 near the boundary of Ω to find an approximate weak solution.

Numerical Solution of Poisson Equations

Let us begin with the Poisson equation:

$$\begin{aligned} -\Delta u &= f\,, \quad \text{in } \Omega \\ u &= g\,, \quad \text{on } \partial\Omega\,, \end{aligned}$$

where Ω is a polygonal domain in $\mathbb{R}^n$, $f \in L_2(\Omega)$, and g is continuous over the boundary $\partial\Omega$ of Ω. The weak formulation of the Poisson equation is to find $u \in H^1(\Omega)$ which satisfy $u = g$ on $\partial\Omega$ and

$$a(u,v) = \langle f, v\rangle\,, \quad \forall v \in H_0^1(\Omega)\,,$$

where $a(u,v)$ is the bilinear form defined by

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x\mathrm{d}y$$

and $\langle f, v\rangle = \int_\Omega f v \mathrm{d}x\mathrm{d}y$ stands for the standard L_2 inner product of f and v. Here $H^1(\Omega)$ and $H_0^1(\Omega)$ are standard Sobolev spaces. By the standard calculus of variations, the Poisson equation is the Euler–Lagrange equation of the energy functional

$$E(w) = \int_\Omega \left(\tfrac{1}{2}\nabla w \cdot \nabla w - wf\right) \mathrm{d}x\,.$$

It is known that the weak solution of the Poisson equation is the minimizer of the energy functional $E(w)$ among the class of admissible functions

$$\mathcal{A} = \{w \in H^1(\Omega), w = g \text{ on } \partial\Omega\}\,.$$

(See §8.2.3. in [20]). That is, the weak solution u satisfies

$$E(u) = \min_{w \in \mathcal{A}} E(w)\,. \tag{10}$$

Also any minimizer satisfying (10) is the weak solution.

Next we discuss how to compute approximate weak solutions that are multivariate spline functions. For convenience, let us consider the Poisson equation in the bivariate setting first. Let Δ be a triangulation of the domain $\Omega \in \mathbb{R}^2$ and let

$$S := S_{\mathbf{d}}^{\mathbf{r},\rho}(\Delta)$$

be a spline space with fixed (global) smoothness $\mathbf{r}$, local smoothness vector ρ and degree vector $\mathbf{d}$ associated with vertices, interior edges, and triangles of Δ. Let d be the largest integer in $\mathbf{d}$. Instead of piecewise linear boundary of $\Omega' = \cup_{t\in\Delta} t$, we may use piecewise quadratic polynomials to approximate the boundary of Ω. That is, for each boundary triangle $t \in \Delta$, if the boundary edge e_t of t is not a part of boundary $\partial\Omega$ of Ω, we use a circular arc $\tilde{e}_t$ which passes through two vertices of e_t and another point on $\partial\Omega$ between the two vertices to replace e_t. Let $\tilde{t}$ be the convex hull of the vertices of t and the circular arc. All the interior triangles and new boundary triangles (with curved side) form a new domain $\widetilde{\Omega}$ which is a better approximation

of Ω than Ω'. Since each spline function $s \in S$ can be extended naturally to $\widetilde{\Omega}$, we may consider that S are defined on $\widetilde{\Omega}$.

We remark that when solving the Poisson equation with the Dirichlet boundary condition, we require spline functions to have less smoothness near the boundary while having more smoothness inside the domain according to the regularity theory of the weak solution of the Poisson and general elliptic PDE's (cf. [20]). In general, there is no spline function in S satisfying the boundary condition exactly. Let $\tilde{\mathcal{A}}$ be the subset of S satisfying the boundary condition approximately in the sense that $s_u \in S$ interpolates g at $2d+1$ distinct points at each curve edge and $d+1$ distinct points over each straight boundary edge. Here, we have assumed that the degrees of the spline functions in S are d over each boundary triangle. Otherwise, we modify the interpolation conditions appropriately. We compute the approximation $s_u \in S$ satisfying

$$E(s_u) = \min_{w \in \tilde{\mathcal{A}}} E(w) .$$

Following the same arguments in [20], the minimizer s_u is the approximate weak solution in S.

We next give an algorithm to compute such an s_u with the assumption that s_u exists and is unique. The proof of the existence and uniqueness is well-known and will be mentioned briefly later.

Let us write any spline function $s \in S$ using the B-form. That is, $s \in S$ may be expressed by

$$s(x,y)|_t = \sum_{i+j+k=d} c^t_{ijk} B^t_{ijk}(x,y) , \quad (x,y) \in t \in \Delta .$$

Let $\mathbf{c} = (c^t_{ijk}, i+j+k=d, t \in \Delta)$ be the coefficient vector associated with s. The length of the vector $\mathbf{c}$ is $\hat{d}T$ with T being the number of triangles in Δ and $\hat{d} = (d+1)(d+2)/2$. The smoothness and super smoothness conditions that s satisfies can be expressed by $H\mathbf{c} = 0$. Also, s satisfies the degree reduction conditions, i. e., $J\mathbf{c} = 0$.

Then the bilinear form $a(s, \hat{s})$ can be expressed in terms of $\mathbf{c}$ and $\hat{\mathbf{c}}$ by

$$a(s, \hat{s}) = \mathbf{c}^\mathsf{T} K \hat{\mathbf{c}}$$

where $K = \operatorname{diag}(K_t, t \in \Delta)$ with

$$K_t = \left[\int_t \nabla B^t_{ijk} \cdot \nabla B^t_{p,q,r} \mathrm{d}x\mathrm{d}y \right]_{\substack{i+j+k=d \\ p+q+r=d}} .$$

Note that the inner product $\langle f, \hat{s} \rangle$ can be approximated by $\langle s_f, \hat{s} \rangle$ where $s_f \in S_d^{-1}(\Delta)$, the space of piecewise polynomials of degree d on each triangle, interpolates f over the domain points of each triangle t. Thus,

$$\langle f, \hat{s} \rangle \approx \hat{\mathbf{c}}^\mathsf{T} M \mathbf{c}_f ,$$

where $M = \operatorname{diag}(M^t, t \in \Delta)$ is a block diagonal matrix with square blocks

$$M^t = \left[\int_t B^t_{ijk}(x,y) B^t_{p,q,r}(x,y) \mathrm{d}x\mathrm{d}y \right]_{\substack{i+j+k=d \\ p+q+r=d}}$$

and $\mathbf{c}_f$ encodes the coefficients of s_f. We need to solve the following minimization problem:

$$\begin{aligned} &\min \tfrac{1}{2}\mathbf{c}^\mathsf{T} K \mathbf{c} - \mathbf{c}^\mathsf{T} M \mathbf{c}_f \\ &\text{subject to} \\ &H\mathbf{c} = 0 , \ J\mathbf{c} = 0 , \ B\mathbf{c} = \mathbf{g} , \end{aligned}$$

where $B\mathbf{c} = \mathbf{g}$ denotes a linear system associated with the boundary conditions. Indeed, based on the de Casteljau algorithm, the evaluation of s_u at any point on a curved edge is a linear equation in terms of the unknown coefficients of s_u. As we can show that there exists a unique approximate weak solution $s_u \in S$, we know that the minimization problem has a unique solution. Since the energy functional is convex, any local minimum is the global minimum. Let us compute a local minimum by using the Lagrange multiplier method. Letting

$$\mathcal{L}(\mathbf{c}, \theta, \eta, \nu) = \tfrac{1}{2}\mathbf{c}^\mathsf{T} K \mathbf{c} - \mathbf{c}^\mathsf{T} M \mathbf{c_f} + \theta^\mathsf{T} H \mathbf{c} + \eta^\mathsf{T} J \mathbf{c} + \nu^\mathsf{T}(B\mathbf{c} - \mathbf{g}),$$

we compute

$$\begin{aligned} &\frac{\partial}{\partial \mathbf{c}} \mathcal{L}(\mathbf{c}, \theta, \eta, \nu) = 0 , \quad \frac{\partial}{\partial \theta} \mathcal{L}(\mathbf{c}, \theta, \eta, \nu) = 0 , \\ &\frac{\partial}{\partial \eta} \mathcal{L}(\mathbf{c}, \theta, \eta, \nu) = 0 , \quad \frac{\partial}{\partial \nu} \mathcal{L}(\mathbf{c}, \theta, \eta, \nu) = 0 . \end{aligned}$$

It follows that

$$\begin{aligned} &K\mathbf{c} + H^\mathsf{T}\theta + J^\mathsf{T}\eta + B^\mathsf{T}\nu = M\mathbf{c}_f \\ &H\mathbf{c} = 0, \ J\mathbf{c} = 0 , B\mathbf{c} = \mathbf{g} . \end{aligned}$$

In other words, we have

$$\begin{bmatrix} B^\mathsf{T} & J^\mathsf{T} & H^\mathsf{T} & K \\ 0 & 0 & 0 & H \\ 0 & 0 & 0 & J \\ 0 & 0 & 0 & B \end{bmatrix} \begin{bmatrix} \theta \\ \eta \\ \nu \\ \mathbf{c} \end{bmatrix} = \begin{bmatrix} M\mathbf{c}_f \\ 0 \\ 0 \\ \mathbf{g} \end{bmatrix} . \tag{11}$$

We shall apply the matrix iterative method for solving the above linear system when it is of large size. The uniqueness of the weak solution implies that K is positive definite with respect to $[B; H; J]$. Therefore, the matrix iterative method

is well-defined. We remark that assembling the matrices M and K is particularly easy and can be done without knowing the relations among the triangles in any given triangulation partition. This is also true in the multivariate setting.

This leads to a numerical method to compute approximate weak solutions for Poisson equations in $\mathbb{R}^2$. It is clear that the above arguments can be generalized to the multivariate setting.

We have implemented this method using bi- and trivariate spline spaces of any degree and any smoothness over any triangulation of any polygonal domain to solve the 2D and 3D Poisson equation. We will provide several numerical experiments near the end of this section.

Next let us briefly discuss the existence and uniqueness of the approximate weak solution s_u. The discussion is parallel to the one using finite elements. Mainly we use the well-known Lax–Milgram Theorem. Since S is a finite dimensional space, we may find a basis $\{\phi_i, i = 1, \ldots, \dim(S)\}$ which may not be locally supported. For any spline function $s \in S$, we write $s = \sum_i s_i \phi_i$ for some coefficients s_i's. Thus, for $s \in S \cap H_0^1(\Omega)$, the bilinear form can be given by

$$a(s, \hat{s}) = \mathbf{s} K' \hat{\mathbf{s}}$$

with a new stiffness matrix K'. Because $a(\cdot, \cdot)$ is coercive, it can be easily shown that K' is positive definite over $S \cap H_0^1(\Omega)$. Thus, the existence and uniqueness of the approximation weak solution s_u follows.

We remark that the Poisson equation with Neumann boundary condition

$$\begin{cases} -\Delta u = f\,, & \text{in } \Omega \\ \frac{\partial u}{\partial \mathbf{n}} = h\,, & \text{on } \partial\Omega \\ \int_\Omega u \mathrm{d}x = 0 \end{cases} \tag{12}$$

can be numerically solved in the same fashion. We leave the details to the interested reader. We have implemented the algorithm for solving (12) numerically in the bivariate and trivariate setting.

General Second-Order Elliptic Equations

We now turn our attention to general second-order elliptic equations. Consider

$$\begin{aligned} -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial}{\partial x_j}u\right) &= f\,, \quad x \in \Omega \\ u(x) &= g(x)\,, \quad x \in \partial\Omega\,, \end{aligned} \tag{13}$$

where $a_{ij}(x) = a_{ji}(x) \in L_\infty(\Omega)$ for $i, j = 1, \ldots, n$ and satisfy

$$\sum_{i,j=1}^{n} a_{ij}\lambda_i\lambda_j \geq m \sum_{i=1}^{n} \lambda_i^2\,, \quad \forall\, \lambda_i, i = 1, \ldots, n$$

for a positive constant $m > 0$. Using the results in (§8.2.3 in [20]), we can show that the weak solution of (13) is the minimizer of

$$E(w) = \int_\Omega \left(\frac{1}{2}\sum_{i,j=1}^{n} a_{ij}\frac{\partial}{\partial x_i}w\frac{\partial}{\partial x_j}w - wf\right)\mathrm{d}x$$

over the set $\mathcal{A}$ of admissible functions. Thus, to find an approximate weak solution in S, we need to solve the following minimization problem:

$$\begin{aligned} &\min \tfrac{1}{2}\mathbf{c}^\mathsf{T}\mathcal{K}\mathbf{c} - \mathbf{c}^\mathsf{T}M\mathbf{f} \\ &\text{subject to} \\ &H\mathbf{c} = 0\,, \quad J\mathbf{c} = 0\,, \quad B\mathbf{c} = \mathbf{g}\,, \end{aligned}$$

where $\mathcal{K} = \mathrm{diag}(\tilde{K}_t, t \in \Delta)$ is a block diagonal matrix with

$$\tilde{K}_t = \left[\int_t \sum_{i,j=1}^{n} a_{ij}\frac{\partial}{\partial x_i}B_\alpha^t\frac{\partial}{\partial x_j}B_{\hat{\alpha}}^t \mathrm{d}x\right]_{\substack{\alpha\in\mathbf{z}^{n+1},|\alpha|=d \\ \hat{\alpha}\in\mathbf{z}^{n+1},|\hat{\alpha}|=d}}.$$

The Lagrange multiplier method implies that we need to solve the following linear system:

$$\begin{bmatrix} B^\mathsf{T} & J^\mathsf{T} & H^\mathsf{T} & \mathcal{K} \\ 0 & 0 & 0 & H \\ 0 & 0 & 0 & J \\ 0 & 0 & 0 & B \end{bmatrix}\begin{bmatrix} \theta \\ \eta \\ \nu \\ \mathbf{c} \end{bmatrix} = \begin{bmatrix} M\mathbf{c}_f \\ 0 \\ 0 \\ \mathbf{g} \end{bmatrix}.$$

Again the uniqueness of the weak solution implies that the matrix iterative method is well defined. Furthermore, we use the maximum norm

$$\left\|\begin{bmatrix} H\mathbf{c} \\ J\mathbf{c} \\ B\mathbf{c} - \mathbf{g} \end{bmatrix}\right\| \tag{14}$$

to check if the iterative solution does satisfy the smoothness conditions, degree reduction conditions, and boundary conditions. The solution will be said to be exact if it is zero in such a norm.

Let us report on some numerical experiments for the 2D and 3D Poisson equations.

Example 1 Consider the Poisson equation with exact solution

$$u(x, y) = 10\exp(-(x^2 + y^2))$$

Multivariate Splines and Their Applications, Table 1
Approximation errors from bivariate spline spaces

	Maximum errors	CPU times
$S_3^1(\Delta)$	0.732222	0.40 s
$S_4^1(\Delta)$	0.063235	0.48 s
$S_5^1(\Delta)$	0.010793	0.78 s
$S_6^1(\Delta)$	0.001382	1.06 s
$S_7^1(\Delta)$	0.000502	1.65 s
$S_8^1(\Delta)$	0.000173	2.56 s
$S_9^1(\Delta)$	0.000013	4.03 s

over a square domain:

$$\begin{cases} -\Delta u = 40\exp(-(x^2+y^2))(1-x^2-y^2) \\ \qquad\qquad\qquad\qquad (x,y)\in[-2,2]\times[-2,2] \\ u(x,y) = 10\exp(-(x^2+y^2)) \\ \qquad\qquad\qquad\qquad (x,y)\in\partial[-2,2]\times[-2,2]\,. \end{cases}$$

The solution is relatively large inside the domain as compared to its values on the boundary. Our spline solutions can approximate it very well. We use a triangulation similar to Fig. 2 with 25 vertices and 32 triangles. We test many spline spaces and list the maximum errors of approximate weak spline solutions against the exact solution in Table 1. The maximum errors are computed based on 101×101 equally spaced points over $[-2,2]\times[-2,2]$.

Example 2 Consider the 3D Poisson equation with exact solution

$$u(x,y,z) = 10\exp(-(x^2+y^2+z^2))$$

over an octahedron $\Omega := \langle(1,0,0),(0,1,0),(-1,0,0),(0,-1,0),(0,0,1),(0,0,-1)\rangle$. We split Ω into eight tetrahedra by three coordinate planes. Let Δ denote the collection of all eight tetrahedra. We find approximate weak solutions in the 3D spline spaces $S_d^1(\Delta)$ for $d = 3,\ldots,7$. The maximum errors are computed based on $20\times 20\times 20$ equally spaced points over Ω and listed in Table 2.

Next we compare the errors by using refinements of underlying triangulations and by increasing the degrees of spline functions.

Example 3 We solve the Poisson equation with Dirichlet boundary condition over a star domain as shown in Fig. 6 with exact solution $u = \exp(x+y)$ using C^1 cubic splines over successively refined triangulations. We can only refine 3 times within the capacity of our PC and the results are listed in Table 3. In Table 4, the degrees of the spline spaces are varied.

Multivariate Splines and Their Applications, Table 2
Approximation errors from trivariate spline spaces

	Matrix size	Maximum errors	CPU times
$S_3^1(\Delta)$	160×160	0.17127	0.07 s
$S_4^1(\Delta)$	280×280	0.02737	0.17 s
$S_5^1(\Delta)$	448×448	0.00749	0.625 s
$S_6^1(\Delta)$	672×672	0.000842	1.67 s
$S_7^1(\Delta)$	960×960	0.0004601	5.18 s

Multivariate Splines and Their Applications, Table 3
Approximation from uniform refinements (Dirichlet Problem)

Refinement levels	Matrix size	Maximum errors
1	80×80	0.254346956
2	320×320	0.029554301
3	1280×1280	0.004515225
4	5120×5120	0.000535312

Multivariate Splines and Their Applications, Table 4
Approximation from degree increase (Dirichlet Problem)

Polynomial degrees	Matrix size	Maximum errors
3	80×80	0.25434695641
4	120×120	0.04251752024
5	168×168	0.00608204535
6	224×224	0.00080855135
7	288×288	0.00009770118
8	360×360	0.00001031184
9	440×440	0.00000096358
10	528×528	0.00000008441
11	624×624	0.00000000697

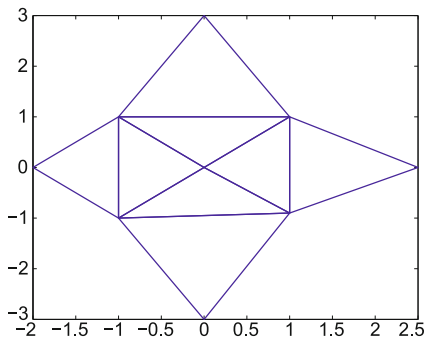

Multivariate Splines and Their Applications, Figure 25
An initial triangulation of a star-shaped domain

Example 4 Next we solve the Poisson equation with Neumann boundary condition over a star domain as shown in Fig. 6 with exact solution

$$u = 10\exp(-(x^2+y^2))$$

Multivariate Splines and Their Applications, Table 5
Approximation from uniform refinements (Neumann Problem)

Refinement levels	Matrix size	Maximum errors
1	120 × 120	2.38×10^{-2}
2	480 × 480	6.81×10^{-4}
3	1920 × 1920	3.15×10^{-5}
4	7680 × 7680	1.29×10^{-6}

Multivariate Splines and Their Applications, Table 6
Approximation from degree increase (Neumann Problem)

Polynomial degrees	Matrix size	Maximum errors
4	120 × 120	2.38×10^{-2}
5	168 × 168	5.53×10^{-3}
6	224 × 224	1.67×10^{-4}
7	288 × 288	1.24×10^{-5}
8	360 × 360	7.87×10^{-6}
9	440 × 440	2.52×10^{-6}
10	528 × 528	3.00×10^{-7}
11	624 × 634	4.08×10^{-8}

using C^1 quartic splines over successively refined triangulation. The numerical results are given in Tables 5 and 6 with refinements and degree increases.

Example 5 In this example, we show the spline approximation of a highly oscillatory solution of the Poisson equation:

$$\begin{cases} -\Delta u = f(x,y)\,, & (x,y) \in [0,1]\times[0,1] \\ u(x,y) = 10\sin(x^2+y^2) + \sin(25(x^2+y^2))\,, & \\ & (x,y) \in \partial[0,1]\times[0,1] \end{cases}$$

where $f(x,y) = 40(\cos(x^2+y^2) - (x^2+y^2)\sin(x^2+y^2)) + 50\cos(25(x^2+y^2)) - 2500\sin(25(x^2+y^2))$. The exact solution contains a high frequency part which is very hard to approximate with linear finite elements. Our spline approximation yields a good approximation of such a solution. In Table 7, we give the maximum errors of the spline approximation using degrees 5, 6, and 7 over uniformly refined triangulations.

It is possible to use spline space of variable degrees to approximate the solution of Poisson equations. We refer to [28] for an adaptive method to adjust degrees and local refinement of triangulation for numerical solution of Poisson equations.

Numerical Solution of Biharmonic Equations

In this section, we show how to solve biharmonic equations using multivariate splines of variable degree and variable smoothness. The biharmonic equation is given as follows:

Multivariate Splines and Their Applications, Table 7
Spline approximation of oscillatory solution

Levels	No. of triangles	Degree 5	Degree 6	Degree 7
1	32	6.738	10.08	4.194
2	128	1.616	0.845	0.212
3	512	0.0391	0.0086	0.0011

$$\begin{aligned} \Delta^2 u &= f\,, && \text{in } \Omega \\ u &= g\,, && \text{on } \partial\Omega \\ \frac{\partial}{\partial \mathbf{n}} u &= h\,, && \text{on } \partial\Omega\,, \end{aligned} \tag{15}$$

where Ω is a polygonal domain in $\mathbb{R}^n$, $f \in L_2(\Omega)$, g and h are in $C(\partial\Omega)$, and $\mathbf{n}$ stands for the normal direction of the boundary $\partial\Omega$. A typical biharmonic equation is the 2D Stokes equations in the stream function formulation (cf., e. g. [38]). The weak formulation for biharmonic equation is to find $u \in H^2(\Omega)$ such that u satisfies the boundary conditions in (15) and

$$a(u,v) = \langle f, v\rangle\,, \quad \forall\, v \in H_0^2(\Omega)\,,$$

where $a(u,v)$ is a bilinear form defined by

$$a(u,v) = \int_\Omega \Delta u \Delta v \mathrm{d}x$$

and $\langle f, v\rangle = \int_\Omega f v\, \mathrm{d}x$ stands for the standard L_2 inner product of f and v. Here $H^2(\Omega)$ and $H_0^2(\Omega)$ are standard Sobolev spaces. With the assumption that the boundary conditions are compatible, that is, there exists a $u_b \in H^2(\Omega)$ satisfying both boundary conditions in (15), we can show that the weak solution exists and is unique by the well-known Lax–Milgram Theorem (cf. [41]). Let

$$E_2(w) = \int_\Omega \left(\frac{1}{2}(\Delta w)^2 - wf\right) \mathrm{d}x$$

be an energy functional and

$$\mathcal{A}_2 = \left\{ w \in H^2(\Omega),\ w = g,\ \frac{\partial}{\partial \mathbf{n}} w = h,\ \text{on } \partial\Omega \right\}$$

be the class of admissible functions. By the compatibility of the boundary conditions, we know that $\mathcal{A}_2$ is not empty. As before, we shall consider the following minimization problem: Find $u \in \mathcal{A}_2$ such that

$$E_2(u) = \min\{E_2(w) \colon w \in \mathcal{A}_2\}\,.$$

On the basis of the standard calculus of variations, it is easy to prove that any minimizer u is a weak solution. Since the weak solution is unique, so is the minimizer.

To find an approximation of the weak solution u, we use multivariate splines of variable degree and variable smoothness. Let Δ be a triangulation of the domain $\Omega \subset \mathbb{R}^n$ and let

$$S := S_{\mathbf{d}}^{\mathbf{r},\rho}(\Delta)$$

be the spline space of fixed smoothness $\rho, \mathbf{r}$ and degree $\mathbf{d}$ associated with k-simplices, $0 \le k < n-1$, $(n-1)$-simplices, and n-simplices of Δ. We assume that the global smoothness $\min\{\mathbf{r}, \rho\}$ of S is bigger or equal to 1 so that $S \subset H^2(\Omega)$. Let d be the large integer in $\mathbf{d}$. The same as in the previous section, we will extend S to be defined over $\widetilde{\Omega}$. We should point out that in general, the weak solution u is smoother inside the domain Ω (cf. [23]). Thus, we should choose S such that each spline function in S is more smooth than near the boundary. Let $\tilde{\mathcal{A}}_2$ be the class of spline functions $s \in S$ satisfying the boundary conditions approximately, i. e., $s \in S$ interpolates g at $2d+1$ distinct points over each curved edge and $d+1$ distinct points over each straight edge and $\partial/(\partial\mathbf{n})s$ interpolates h at $2d-1$ distinct points over each curved edge and d distinct points over each straight edge. Our algorithm is to find $s_u \in \tilde{\mathcal{A}}_2$ such that

$$E_2(s_u) = \min\{E_2(s) \colon s \in \tilde{\mathcal{A}}_2\}\,.$$

More precisely, let us write any spline function $s \in S$ as

$$s(x)|_t = \sum_{\substack{\alpha \in \mathbf{z}^{n+1} \\ |\alpha| = d}} c_\alpha^t B_\alpha^t(x)\,, \quad x \in t \in \Delta\,,$$

where $d = \max\{d_t, t \in \Delta\}$. Let $\mathbf{c} = (c_\alpha^t, \alpha \in \mathbf{Z}^{n+1}, |\alpha| = d, t \in \Delta)$ be the coefficient vector associated with s. The smoothness and super smoothness conditions that s satisfies can be expressed by $H\mathbf{c} = 0$. Also, s satisfies the degree reduction conditions $J\mathbf{c} = 0$.

Then the bilinear form $a(s, \hat{s})$ can be expressed in terms of $\mathbf{c}$ and $\hat{\mathbf{c}}$ by

$$a(s, \hat{s}) = \mathbf{c}^{\mathsf{T}} K \hat{\mathbf{c}}$$

where $K = \operatorname{diag}(K_t, t \in \Delta)$ with

$$K_t = \left[\int_t \Delta B_\alpha^t(x) \Delta B_\gamma^t(x) \mathrm{d}x\right]_{\substack{\alpha \in \mathbf{z}^{n+1}, |\alpha| = d \\ \gamma \in \mathbf{z}^{n+1}, |\gamma| = d}}\,.$$

Note that the inner product $\langle f, \hat{s}\rangle$ can be approximated by $\langle s_f, \hat{s}\rangle$ for a spline s_f which interpolates f over the domain points of each n-simplex t. Thus

$$\langle f, \hat{s}\rangle \approx \hat{\mathbf{c}}^{\mathsf{T}} M \mathbf{c}_f$$

where $M = \operatorname{diag}(M^t, t \in \Delta)$ is a block diagonal matrix with square blocks

$$M^t = \left[\int_t B_\alpha^t(x) B_\gamma^t(x) \mathrm{d}x\right]_{\substack{|\alpha| = d \\ |\gamma| = d}}$$

and $\mathbf{c}_f$ is the coefficient vector for s_f. We need to solve the following minimization problem:

$$\begin{aligned} &\min \tfrac{1}{2}\mathbf{c}^{\mathsf{T}} K \mathbf{c} - \mathbf{c}^{\mathsf{T}} M \mathbf{c}_f \\ &\text{subject to} \\ &H\mathbf{c} = 0\,, \quad J\mathbf{c} = 0\,, \quad B\mathbf{c} = \mathbf{g}\,, \end{aligned}$$

where $B\mathbf{c} = \mathbf{g}$ denotes the linear system associated with the approximate boundary conditions. Note that the minimization problem has a unique solution. Since the energy functional is convex, any local minimum is the global minimum. Let us compute a local minimum by using the Lagrange multiplier method. Letting

$$\mathcal{L}(\mathbf{c}, \theta, \eta, \nu) = \tfrac{1}{2}\mathbf{c}^{\mathsf{T}} K \mathbf{c} - \mathbf{c}^{\mathsf{T}} M \mathbf{f} + \theta^{\mathsf{T}} H \mathbf{c} + \eta^{\mathsf{T}} J \mathbf{c} + \nu^{\mathsf{T}}(B\mathbf{c} - \mathbf{g}),$$

we compute

$$\begin{aligned} &\frac{\partial}{\partial \mathbf{c}}\mathcal{L}(\mathbf{c}, \theta, \eta, \nu) = 0\,, \quad \frac{\partial}{\partial \alpha}\mathcal{L}(\mathbf{c}, \theta, \eta, \nu) = 0\,, \\ &\frac{\partial}{\partial \eta}\mathcal{L}(\mathbf{c}, \theta, \eta, \nu) = 0\,, \quad \frac{\partial}{\partial \gamma}\mathcal{L}(\mathbf{c}, \theta, \eta, \nu) = 0\,. \end{aligned}$$

It follows that

$$\begin{aligned} &K\mathbf{c} + H^{\mathsf{T}}\theta + J^{\mathsf{T}}\eta + B^{\mathsf{T}}\nu = M\mathbf{c}_f \\ &H\mathbf{c} = 0, J\mathbf{c} = 0, B\mathbf{c} = \mathbf{g}\,. \end{aligned}$$

In other words, we need to solve the following linear system

$$\begin{bmatrix} H^{\mathsf{T}} & J^{\mathsf{T}} & B^{\mathsf{T}} & K \\ 0 & 0 & 0 & H \\ 0 & 0 & 0 & J \\ 0 & 0 & 0 & B \end{bmatrix} \begin{bmatrix} \theta \\ \eta \\ \nu \\ \mathbf{c} \end{bmatrix} = \begin{bmatrix} M\mathbf{c}_f \\ 0 \\ 0 \\ \mathbf{g} \end{bmatrix}.$$

As discussed in the previous sections, the uniqueness of the weak solution implies that the matrix K is positive definite with respect to $[H; D; B]$. Thus, the matrix iterative method is well-defined.

To make sure that the iterative solution is the weak solution of the biharmonic equation, we use the maximum norm

$$\left\| \begin{bmatrix} H\mathbf{c} \\ J\mathbf{c} \\ B\mathbf{c} - \mathbf{g} \end{bmatrix} \right\|_\infty \tag{16}$$

Multivariate Splines and Their Applications, Table 8
Numerical approximation of a biharmonic equation over a standard square domain

	Maximum errors	CPU times
$S_5^1(\Delta)$	5.7959×10^{-7}	2.8 s
$S_6^1(\Delta)$	1.1001×10^{-8}	4.5 s
$S_7^1(\Delta)$	1.3208×10^{-10}	6.2 s
$S_8^1(\Delta)$	1.1465×10^{-11}	9.8 s
$S_5^2(\Delta)$	1.1187×10^{-5}	3.5 s
$S_6^2(\Delta)$	3.2605×10^{-8}	5.2 s
$S_7^2(\Delta)$	2.7998×10^{-10}	7.9 s
$S_8^2(\Delta)$	1.1982×10^{-11}	13.2 s

to check if it does satisfy the boundary conditions, smoothness conditions, and degree reduction conditions. The solution will be said to be exact if it's zero in the above norm.

We remark that the above algorithm also gives a numerical method to determine if the boundary conditions are compatible or not. That is, if the least squares solution in the norm (16) is not close to zero as the underlying triangulations are refined or degrees of the spline functions increase, then the boundary conditions are not compatible since $S_d^1(\Delta)$ becomes dense in $H^2(\Omega)$ if d increases to ∞ and/or $|\Delta|$ decreases to 0.

We have implemented this method for 2D and 3D biharmonic equations using bivariate and trivariate spline spaces of any degree and any smoothness. That is, we are able to numerically solve biharmonic equations over any polygonal domain in the bivariate or trivariate setting. Let us present several numerical examples below.

Example 6 Consider a 2D biharmonic equation with exact solution $u(x, y) = \exp(x + y)$ over a unit square domain:

$$\begin{cases} \Delta^2 u = 4\exp(x+y)\,, & (x,y) \in [0,1]\times[0,1] \\ u(x,y) = \exp(x+y)\,, & (x,y) \in \partial[0,1]\times[0,1] \\ \frac{\partial}{\partial x}u(x,y) = \exp(x+y)\,, & (x,y) \in \partial[0,1]\times[0,1] \\ \frac{\partial}{\partial y}u(x,y) = \exp(x+y)\,, & (x,y) \in \partial[0,1]\times[0,1]\,. \end{cases}$$

We used the triangulation with 25 vertices and 32 triangles and tested many spline spaces. The maximum errors of approximate weak spline solutions against the exact solution are given in Table 8. The maximum errors are computed based on 101×101 equally spaced points over $[0, 1] \times [0, 1]$.

Example 7 Consider a 2D biharmonic equation with exact solution $u(x, y) = 10\exp(-(x^2 + y^2))$ over a unit circular domain:

$$\begin{cases} \Delta^2 u = 160\exp(-(x^2+y^2)) \\ \qquad \times(x^4+y^4+2x^2y^2+2-4x^2-4y^2)\,, \\ \qquad\qquad (x,y) \in \{(x,y), x^2+y^2<1\} \\ u(x,y) = 10\exp(-(x^2+y^2))\,, \\ \qquad\qquad (x,y) \in \{(x,y), x^2+y^2=1\} \\ \frac{\partial}{\partial x}u(x,y) = -20x\exp(-(x^2+y^2))\,, \\ \qquad\qquad (x,y) \in \{(x,y), x^2+y^2=1\} \\ \frac{\partial}{\partial y}u(x,y) = -20y\exp(-(x^2+y^2))\,, \\ \qquad\qquad (x,y) \in \{(x,y), x^2+y^2=1\}\,. \end{cases}$$

Multivariate Splines and Their Applications, Table 9
Numerical approximation of a biharmonic equation over a circular domain

Spline spaces	Matrix sizes	Maximum errors
$S_3^1(\Delta)$	1990×1990	6.6819×10^{-2}
$S_4^1(\Delta)$	2985×2985	2.0199×10^{-4}
$S_5^1(\Delta)$	4179×4179	1.3653×10^{-6}
$S_6^1(\Delta)$	5572×5572	7.6779×10^{-8}
$S_7^1(\Delta)$	7164×7164	1.7841×10^{-9}
$S_8^1(\Delta)$	8955×8955	4.4959×10^{-10}

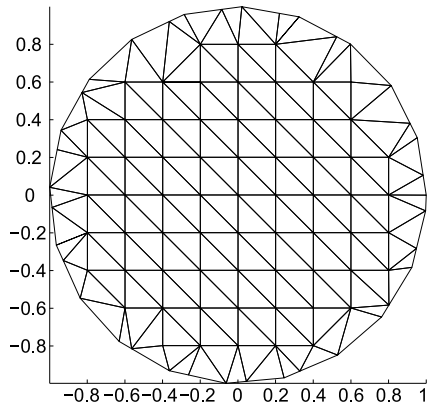

Multivariate Splines and Their Applications, Figure 26
A triangulation of the unit circular domain

We use the following triangulation and test many spline spaces. The maximum errors of approximate weak spline solutions against the exact solution are given in Table 9. The maximum errors are computed based on 101×101 equally spaced points over $[-1, 1] \times [-1, 1]$ within the circular domain.

Example 8 Consider a 3D biharmonic equation with exact solution

$$u(x, y, z) = 10\exp(-(x^2 + y^2 + z^2))$$

Multivariate Splines and Their Applications, Table 10
Approximation errors from trivariate spline spaces

Spline spaces	Matrix size	Maximum errors
$S_3^1(\Delta)$	160×160	0.248542
$S_4^1(\Delta)$	280×280	0.048342
$S_5^1(\Delta)$	448×448	0.014806
$S_6^1(\Delta)$	672×672	0.001903
$S_7^1(\Delta)$	960×960	0.000756

over an octahedron Ω as in Example 2. We use the same tetrahedral partition as above. We find approximate weak solutions from 3D spline spaces $S_d^1(\Delta)$ for $d = 3, \ldots, 7$ and Table 10 is a list of maximum errors against the exact solution evaluated at $20 \times 20 \times 20$ points over Ω.

It is possible to use spline space of variable degrees to approximate the solution of biharmonic equations. We refer to [28] for an adaptive method to automatically adjust degrees and local refinement of triangulation for numerical solution of biharmonic equations.

Bivariate Splines for Fluid Flow Simulations

In this section we use bivariate spline functions for numerical solution of 2D Navier–Stokes equations which enable us to do 2D fluid flow simulations. (See [4] for the trivariate spline approximation of 3D Navier–Stokes equations.) Our approach is like the finite element method using triangles to approximate any given 2D polygonal domains and using piecewise polynomials over triangulations to approximate the solution of the Navier–Stokes equations. The main different features are:

(1) no macro-element or locally supported spline functions are constructed;
(2) polynomials of high degrees can be easily used to get a better approximation power;
(3) smoothness can be imposed in a flexible way across the domain at places where the solution is expected to be smooth. For example, the solution of the steady state Navier–Stokes equation is H^2 inside the domain and H^1 near the boundary;
(4) the mass and stiffness matrices can be assembled easily and these processes can be done in parallel;
(5) the stream function formulation will be used and thus the spline approximation of the solution of Navier–Stokes equations satisfies the divergence-free condition exactly;
(6) the matrices that arise are singular which is an important difference from the classical finite element method.
(7) our spline method leads to a linear system of special structure. We introduce a special numerical method to solve such particularly structured linear systems.

Let us first introduce the stream function formulation. Let $\Omega \subseteq \mathbb{R}^2$ be a simply connected domain and $\mathbf{u} = (u_1, u_2)^\mathsf{T}$ be the planar velocity of a fluid flow over Ω. Also, let p be the pressure function, $\mathbf{f} = (f_1, f_2)^\mathsf{T}$ be the external body force of the fluid and $\mathbf{g} = (g_1, g_2)^\mathsf{T}$ be the velocity of the fluid flow on the boundary $\partial\Omega$. Then the steady state Navier–Stokes equations are

$$\begin{cases} -\nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p = \mathbf{f}\,, & (x,y) \in \Omega \\ \operatorname{div}\mathbf{u} = 0\,, & (x,y) \in \Omega \\ \mathbf{u} = \mathbf{g}\,, & (x,y) \in \partial\Omega\,, \end{cases} \tag{17}$$

where Δ denotes the usual Laplace operator and ∇ the gradient operator. After omitting the nonlinear terms, we have the steady state Stokes' equations:

$$\begin{cases} -\nu\Delta\mathbf{u} + \nabla p = \mathbf{f}\,, & (x,y) \in \Omega \\ \operatorname{div}\mathbf{u} = 0\,, & (x,y) \in \Omega \\ \mathbf{u} = \mathbf{g}\,, & (x,y) \in \partial\Omega\,. \end{cases} \tag{18}$$

Recall the fact that there exists a stream function φ such that $\mathbf{u} = \mathbf{curl}\,\varphi$, i. e., $u_1 = \partial\varphi/\partial y, u_2 = -\partial\varphi/\partial x$. Such φ is unique up to a constant (cf. [22]). Thus we may simplify the above Stokes and Navier–Stokes equations by canceling the pressure term. Consider the Stokes equations first. Replacing $\mathbf{u}$ by $\mathbf{curl}\,\varphi$ and then differentiating the first equation with respect to y and the second with respect to x, we subtract the first equation from the second one to obtain the following fourth-order equation

$$\nu\Delta^2\varphi = h$$

with $h = \frac{\partial f_2}{\partial x} - \frac{\partial f_1}{\partial y}$. Thus, the Stokes equations become a biharmonic equation:

$$\begin{cases} \nu\Delta^2\varphi = h\,, & \text{in } \Omega \\ \frac{\partial\varphi}{\partial x} = -g_2\,, & \text{on } \partial\Omega \\ \frac{\partial\varphi}{\partial y} = g_1\,, & \text{on } \partial\Omega \\ \varphi = b_2\,, & \text{on } \partial\Omega \end{cases} \tag{19}$$

where b_2 is an anti-derivative of the tangential derivative of φ along $\partial\Omega$ and will be examined in detail later. By a similar calculation, we easily see that the Navier–Stokes equations become the following fourth-order nonlinear

equation

$$\begin{cases} \nu\Delta^2\varphi - \frac{\partial}{\partial y}\left(\frac{\partial\varphi}{\partial y}\frac{\partial^2\varphi}{\partial x\partial y} - \frac{\partial\varphi}{\partial x}\frac{\partial^2\varphi}{\partial y^2}\right) \\ \qquad -\frac{\partial}{\partial x}\left(\frac{\partial\varphi}{\partial y}\frac{\partial^2\varphi}{\partial x^2} - \frac{\partial\varphi}{\partial x}\frac{\partial^2\varphi}{\partial x\partial y}\right) = h\,, & \text{in } \Omega \\ \frac{\partial\varphi}{\partial x} = -g_2\,, & \text{on } \partial\Omega \\ \frac{\partial\varphi}{\partial y} = g_1\,, & \text{on } \partial\Omega \\ \varphi = b_2\,, & \text{on } \partial\Omega\,. \end{cases} \tag{20}$$

Let $H^2(\Omega)$ be the usual Sobolev space and $H_0^2(\Omega)$ be the subspace of $H^2(\Omega)$ of functions whose derivatives of order less than or equal to one all vanish on the boundary $\partial\Omega$. Define the bilinear form $a_2(\varphi,\psi)$ and trilinear form $q(\theta,\varphi,\psi)$ by

$$a_2(\varphi,\psi) = \int_\Omega \Delta\varphi(x,y)\Delta\psi(x,y)\,\mathrm{d}x\mathrm{d}y$$

$$q(\theta,\varphi,\psi) = \int_\Omega \Delta\theta(x,y)\left(\frac{\partial\varphi(x,y)}{\partial x}\frac{\partial\psi(x,y)}{\partial y} - \frac{\partial\varphi(x,y)}{\partial y}\frac{\partial\psi(x,y)}{\partial x}\right)\mathrm{d}x\mathrm{d}y$$

and denote the $L_2(\Omega)$ inner product by

$$\langle h,\psi\rangle = \int_\Omega h(x,y)\psi(x,y)\,\mathrm{d}x\mathrm{d}y\,.$$

We that say $\varphi \in H^2(\Omega)$ is a weak solution of the Stokes equations (19) if φ satisfies the following

$$\begin{cases} \nu a_2(\varphi,\psi) = \langle h,\psi\rangle\,, & \forall\ \psi \in H_0^2(\Omega) \\ \frac{\partial\varphi}{\partial x} = -g_2\,, & \text{on } \partial\Omega \\ \frac{\partial\varphi}{\partial y} = g_1\,, & \text{on } \partial\Omega \\ \varphi = b_2\,, & \text{on } \partial\Omega\,. \end{cases}$$

Similarly, a function $\varphi \in H^2(\Omega)$ is a weak solution of the Navier–Stokes equations (20) if φ satisfies

$$\begin{cases} \nu a_2(\varphi,\psi) + q(\varphi,\varphi,\psi) = \langle h\,,\psi\rangle, & \forall\ \psi \in H_0^2(\Omega) \\ \frac{\partial\varphi}{\partial x} = -g_2\,, & \text{on } \partial\Omega \\ \frac{\partial\varphi}{\partial y} = g_1\,, & \text{on } \partial\Omega \\ \varphi = b_2\,, & \text{on } \partial\Omega\,. \end{cases}$$

Such weak formulations are referred to as the stream function formulation of the Stokes and Navier–Stokes equations, respectively. It is known that the weak solution for Stokes' equations exists and is unique for any $\nu > 0$. For the Navier–Stokes equations, such a weak solution exists for any $\nu > 0$, and is unique when ν is sufficiently large. (See, e. g. [22]).

There are two major advantages to using the stream function formulation over the traditional velocity-pressure formulation and vorticity-stream function formulation. Indeed, with the stream function formulation, we need to approximate only one stream function. Otherwise, we need to approximate two components of the velocity and one pressure function if the velocity-pressure formulation is used or one vorticity and one stream function if the vorticity-stream function formulation is used. In addition, the stream function formulation eliminates the pressure function which does not have an appropriate boundary condition. In the vorticity-stream function formulation, the vorticity function does not have an appropriate boundary condition. With bivariate spline functions of higher degrees, we are able to approximate stream functions very well. Thus, in this paper, we will use bivariate

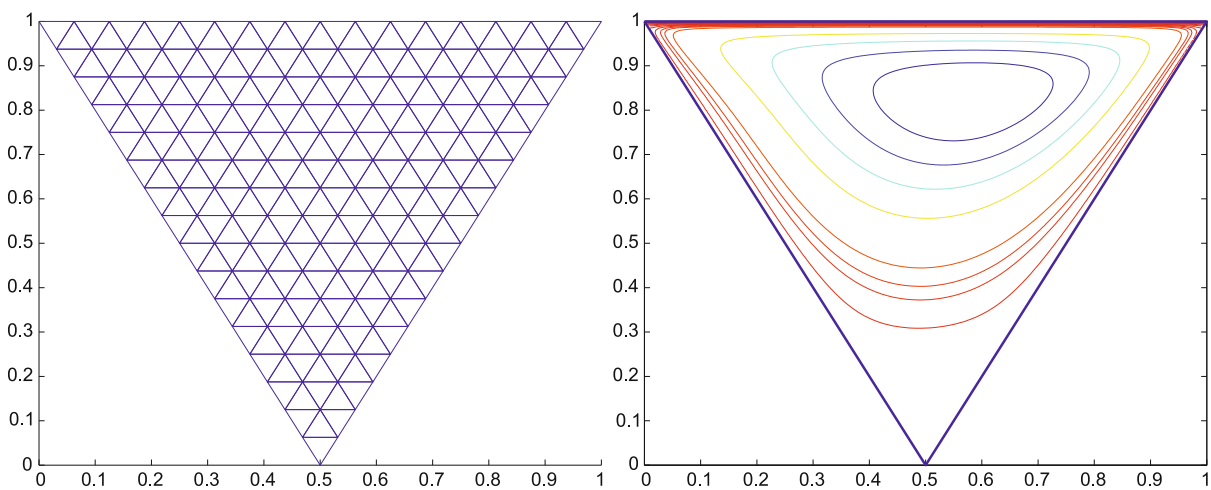

Multivariate Splines and Their Applications, Figure 27
A triangular domain and the stream lines of the triangular driven cavity flow (Reynolds number = 100)

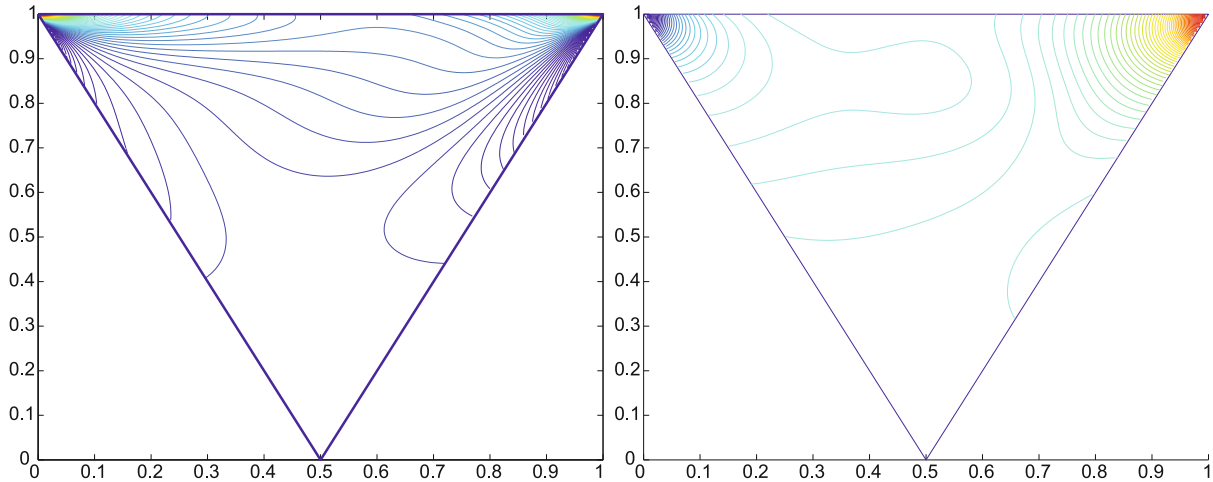

Multivariate Splines and Their Applications, Figure 28
Contour of the vorticity and pressure of the driven cavity flow (Reynolds number = 100)

splines to approximate the stream function of the Stokes and Navier–Stokes equations.

Next we discuss the computation of the pressure functions. By taking divergence, div, of the equations in Eq. (17) and Eq. (18), we can easily see that the pressure functions p of the Stokes and Navier–Stokes equations satisfy the following Poisson equations with nonhomogeneous Neumann boundary conditions involving the stream functions.

$$\begin{cases} -\Delta p = -\mathrm{div}\,(\mathbf{f})\,, & \text{in } \Omega \\ \frac{\partial p}{\partial n} = \nu\Delta(n\cdot \mathbf{curl}\,\varphi) + n\cdot\mathbf{f} & \text{on } \partial\Omega \end{cases} \tag{21}$$

for the Stokes equation and

$$\begin{cases} -\Delta p = -\mathrm{div}\,(\mathbf{f}) + \mathrm{div}\,[(\mathbf{curl}\,\varphi\cdot\nabla)\mathbf{curl}\,(\varphi)]\,, & \text{in } \Omega \\ \frac{\partial p}{\partial n} = \nu\Delta(n\cdot \mathbf{curl}\,\varphi) + n\cdot\mathbf{f} \\ \qquad\qquad + n\cdot[(\mathbf{curl}\,\varphi\cdot\nabla)(\mathbf{curl}\,\varphi)]\,, & \text{on } \partial\Omega \end{cases} \tag{22}$$

for the Navier–Stokes equation. We will use these boundary conditions to compute p after we find the spline approximations of the stream function φ.

We finally present our numerical experiments.

Example 9 Let us consider the cavity flow over a triangular domain. We uniformly refine the triangle as shown in Fig. 27 and use bivariate splines of degree 8 and smoothness 2. The boundary conditions are $\mathbf{u} = (u_1, u_2) = (0, 0)$ for all three line boundary pieces except for $u_1(x, 1) = 1$ when $0 \le x \le 1$. With Reynolds numbers of 100, 1000, and 5000, the stream lines of the cavity flow and the contours of vorticity and pressure are shown in Figs. 27–32.

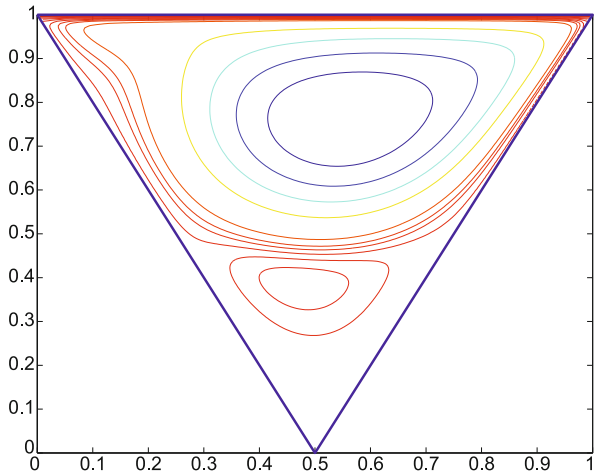

Multivariate Splines and Their Applications, Figure 29
The stream lines of the triangular driven cavity flow (Reynolds number = 1000)

More examples on benchmark flows: the driven cavity flow, the backward facing step flow, and the flow around a circular object can be found in [41].

Multivariate Box Spline Wavelets

Multivariate box splines have been used to generate a multiresolution approximation of $L_2(\mathbb{R}^n)$ and various wavelets, e. g., orthonormal wavelets, orthonormal wavelets in Sobolev spaces, biorthogonal wavelets, pre-

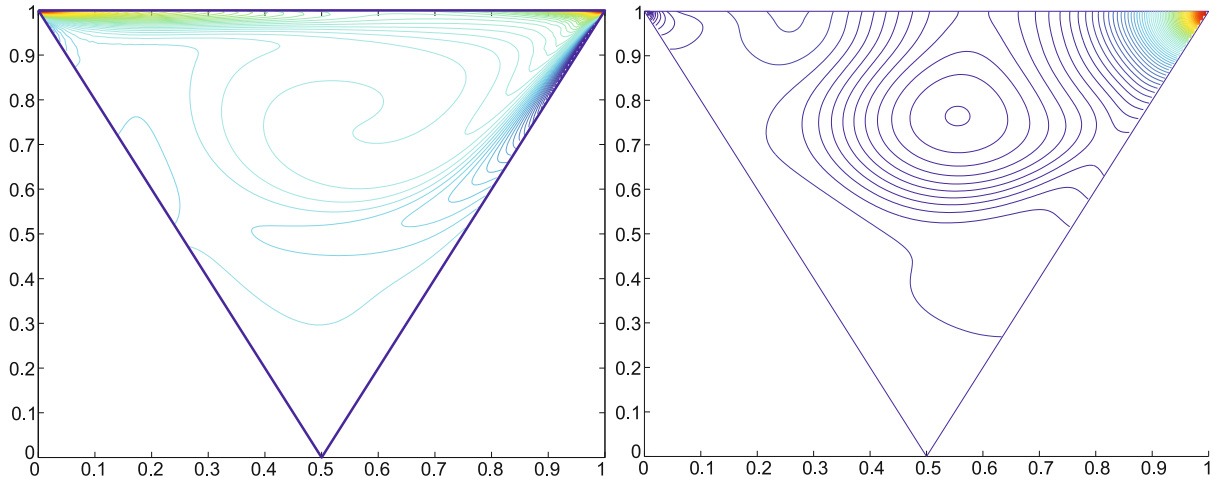

Multivariate Splines and Their Applications, Figure 30
Contour of the vorticity and pressure of the driven cavity flow (Reynolds number = 1000)

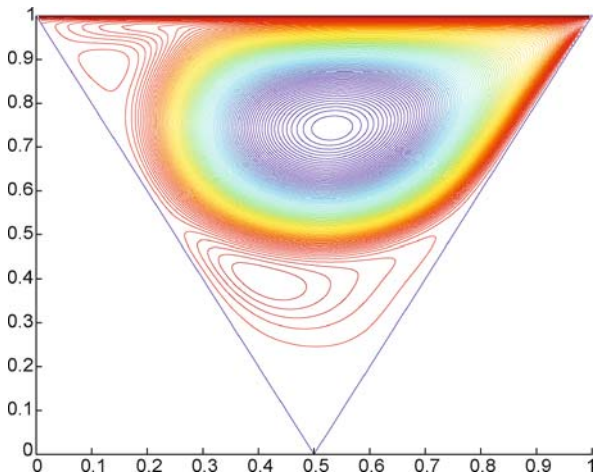

Multivariate Splines and Their Applications, Figure 31
The Stream lines of the triangular driven cavity flow (Reynolds number = 50000)

wavelets, and tight wavelet frames. In this section we just give a construction of orthonormal wavelets when $n = 2$ and $n = 3$ based on the work by Riemenschneider and Shen in [44] and leave the construction of other wavelets to the papers [24,25,26,31,37]. A short summary of these wavelets can be found in ► Popular Wavelet Families and Filters and Their Use.

Multiresolution Approximation Built by Box Splines

We begin with the definition of multiresolution approximation of $L^2(\mathbb{R}^n)$.

Definition 1 A nested sequence of subspaces

$$\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots$$

of $L^2(\mathbb{R}^d)$ form a multiresolution approximation if they satisfy

1) $\forall f \in V_{k+1}, f(\cdot/2) \in V_k$ and $\forall f \in V_k,\ f(2\cdot) \in V_{k+1}$ for all $k \in \mathbb{Z}$;
2) $\forall f \in V_k, f(\cdot - 2^{-k}\mathbf{i}) \in V_k$ for all $\mathbf{i} \in \mathbb{Z}^n$;
3) There exists $\phi \in V_0$ and two positive constants C and D such that

$$C\|\{c_\mathbf{i}\}\|_2^2 \le \|\sum_\mathbf{i} c_\mathbf{i}\phi(\cdot - \mathbf{i})\|_2^2 \le D\|\{c_\mathbf{i}\}\|_2^2;$$

and

4) $\cup_{k=-\infty}^{\infty} V_k$ is dense in $L^2(\mathbb{R}^n)$ and $\cap_{k=-\infty}^{\infty} V_k = \{0\}$.

Let ϕ be a refinable function, e. g. ϕ is a box spline function. We say that ϕ generates a multi-resolution approximation of $L^2(\mathbb{R}^d)$ if the nested sequence of subspaces defined by

$$V_0 := \text{span}\,\{\phi(\cdot - \mathbf{i}) \colon \mathbf{i} \in \mathbb{Z}^d\}$$

and

$$V_k := \{f(2^k\cdot) \colon f \in V_0\}, \quad \forall\ k \in \mathbb{Z}$$

forms a multi-resolution approximation of $L^2(\mathbb{R}^n)$.

We now consider a multiresolution approximation built by box splines. Recall that the box spline ϕ_D associates with a direction set D which is the function whose

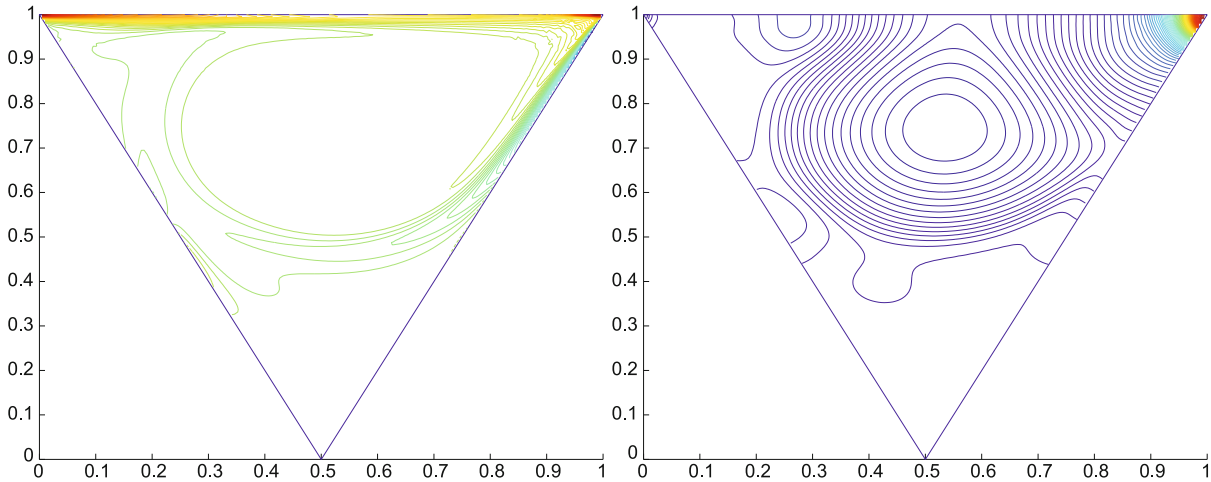

Multivariate Splines and Their Applications, Figure 32
Contour of the vorticity and pressure of the driven cavity flow (Reynolds number = 50000)

Fourier transform is defined by

$$\hat{\phi}_D(\omega) = \prod_{\xi \in D} \frac{1 - e^{-i\xi\cdot\omega}}{i\xi \cdot \omega} .$$

It is easy to see that ϕ_D is a refinable function satisfying

$$\phi_D(x) = \sum_k p_k \phi_D(2x - k)$$

for some finitely many nonzero coefficients p_k. Under some assumption on D, the box spline function ϕ_D can generate a multiresolution approximation of $L^2(\mathbb{R}^n)$.

Theorem 29 *Suppose that the direction set D satisfies*

$$|\det(Y)| = 1, \quad \forall\ Y \subset D \text{ with } |Y| = n .$$

Then there exist two positive constants A and B such that

$$A\|\{c_k\}\|_2^2 \le \| \sum_{k\in\mathbb{Z}^n} c_k \phi_D(\cdot - k)\|_2^2 \le B\|\{c_k\}\|_2^2$$

and ϕ_D generates a multi-resolution approximation of $L_2(\mathbb{R}^n)$.

The values of A and B can be computed by using Poisson's summation formula. In fact we have $B = 1$.

The multiresolution approximation of $L_2(\mathbb{R}^n)$ built by a box spline function can be used to construct orthonormal wavelets, biorthogonal wavelets, prewavelets, and tight wavelet frames. In the following, we explain how to construct orthonormal wavelets in $L_2(\mathbb{R}^n)$ with $n = 2$ and $n = 3$. The construction of orthonormal wavelets in Sobolev space, biorthogonal wavelets, prewavelets, and tight wavelet frames based on box splines can be found in ► Popular Wavelet Families and Filters and Their Use.

We begin with the definition of orthonormal wavelets. Let $\psi_k, k \in \Gamma_n\backslash\{0\}$ be wavelet functions if their integer translates and dilations form an orthonormal basis for $L_2(\mathbb{R}^n)$. That is,

$$2^{jn/2}\psi_k(2^j x - i) , \quad i \in \mathbb{Z}^n , \quad j \in \mathbb{Z} , \quad k \in \Gamma_n\backslash\{0\}$$

form an orthonormal basis for $L_2(\mathbb{R}^n)$ in the sense that they are orthonormal each and all of them span $L_2(\mathbb{R}^n)$.

To construct such functions, we first look for an orthonormal scaling function ϕ based on $V_0 = \text{span}\{\phi_D(\cdot - k), k \in \mathbb{Z}^n\}$. We begin with the following concept: ϕ is orthonormal if

$$\int_{-\infty}^{\infty} \phi(x-k)\phi(x-j)\mathrm{d}x = \begin{cases} 1 , & \text{if } k = j \\ 0 , & \text{otherwise} . \end{cases}$$

It is easy to see the following equivalent condition by using the Fourier transform and Parseval's inequality:

Lemma 6 *A function ϕ is orthonormal if and only if*

$$\sum_k |\widehat{\phi}(\omega + 2k\pi)|^2 = 1, \quad \forall\ \omega \in [0, 2\pi]^n ,$$

where $\widehat{\phi}$ stands for the Fourier transform of ϕ.

Define ϕ in terms of its Fourier transform by

$$\widehat{\phi}(\omega) = \frac{\widehat{\phi}_D(\omega)}{\sqrt{\sum_k |\widehat{\phi}_D(\omega + 2k\pi)|^2}} .$$

Then we claim that the ϕ defined above is an orthonormal scaling function generating the same multiresolution approximation built by box spline ϕ_D under the assumption that D satisfies the condition in Theorem 29.

Note that $\sum_k |\widehat{\phi_D}(\omega + 2k\pi)|^2 \geq A > 0$. Thus, its square root is well defined and hence, so is ϕ. In fact,

$$\frac{1}{\sqrt{\sum_k |\widehat{\phi_D}(\omega + 2k\pi)|^2}} = \sum_k a_k(D) e^{jk\omega}$$

with $j = \sqrt{-1}$ and the coefficients $a_k(D)$ being of exponential decay, i. e.,

$$|a_k(D)| \leq C e^{-\lambda|k|}, \quad \forall\, k \in \mathbb{Z}^n,$$

for some positive constants C and λ. Therefore,

$$\phi(x) = \sum_k a_k(D)\phi_D(x-k) \quad \in V_0 .$$

Similarly, we have

$$\sqrt{\sum_k |\widehat{\phi_D}(\omega + 2k\pi)|^2} = \sum_k b_k(D) e^{jk\omega}$$

with the coefficients $b_i(D)$ being of exponential decay, i. e.,

$$|b_k(D)| \leq C e^{-\lambda|k|}, \quad \forall\, k \in \mathbb{Z}^n$$

for some positive constants C and λ. Thus,

$$\phi_D(x) = \sum_k b_k(D)\phi(x-k) .$$

Hence, the multiresolution approximation generated by ϕ is the same one generated by ϕ_D.

It is easy to verify that ϕ defined above satisfies the orthonormal condition in Lemma 6. Thus, ϕ is an orthonormal refinable function.

Next we define the symbol of the mask associated with ϕ by

$$\widehat{\phi}(\omega) = H(\omega/2)\widehat{\phi}(\omega/2) .$$

$H(\omega)$ is the lower-pass filter associated with the refinable function ϕ. For a later application, we need the following

Theorem 30 *For $\omega \in [0, 2\pi]^s$, we have*

$$\sum_{i \in \Gamma_n} |H(e^{j(\omega + i\pi)})|^2 = 1 ,$$

where $\Gamma_n = \{0,1\}^n$.

Orthonormal Box Spline Wavelets

We are now ready to discuss the construction of orthonormal box spline wavelets. Mainly we give the construction of orthonormal box spline wavelets in $\mathbb{R}^2$ and $\mathbb{R}^3$. These two cases are more important and of more interest for applications. Suppose that $D \subset \mathbb{Z}^n$ with $n = 2$ or $n = 3$ satisfies the condition in Theorem 29, i. e.,

$$|\det(Y)| = 1, \forall\, Y \subset X_n \text{ with } \#Y = n, \ \operatorname{span}(Y) = \mathbb{R}^n .$$

For example, when $s = 2$, D consists of certain repetitions of $\mathbf{e}_1 = (1,0)$, $\mathbf{e}_2 = (0,1)$, and $\mathbf{e}_3 = (1,1)$. Then we know ϕ_D generates a multiresolution approximation of $L_2(\mathbb{R}^2)$. Let ϕ be the scaling function defined in the previous subsection which generates the same multiresolution approximation of $L^2(\mathbb{R}^2)$ as ϕ_D does.

Denote by W_0 the orthogonal complement of V_0 in V_1. Let $\Gamma_n = \{0,1\}^n$. Then we are looking for $\psi_k(t), \forall k \in \Gamma_n \backslash \{\mathbf{0}\}$ such that the integer translates of them form an orthonormal basis of W_0.

Since $\psi_k \in W_0 \subset V_1$, we have

$$\hat{\psi}_k(\omega) = H_k(\omega/2)\hat{\phi}(\omega/2)$$

for some functions $H_k(\omega)$, $k \in \Gamma_n$. For convenience, we denote $\psi_0(t) := \phi(t)$ and write $H_0(\omega) := H(e^{j\omega})$ where $H(e^{j\omega})$ is defined by

$$\widehat{\psi}_0(\omega) = \widehat{\phi}(\omega) = H_0(\omega/2)\hat{\phi}(\omega/2) .$$

Thus, for any function $f(t) = \sum_k c_k \phi(2t-k) \in V_1$, we can write it as a linear combination of the integer translates of ψ_i for $i \in \Gamma_n$. That is,

$$\sum_k c_k 2^{n/2}\phi(2x-k) = \sum_{i \in \Gamma_n} \sum_k c_{k,i}\psi_i(x-k) .$$

Because of the orthonormality of ψ_i's, we have

$$\sum_i |c_i|^2 = \sum_{k \in \Gamma_n} \sum_i |c_{k,i}|^2 .$$

In terms of its Fourier transform, we have

$$C(\omega/2)\hat{\phi}(\omega/2) = \sum_{k \in \Gamma_n} C_k(\omega)H_k(\omega/2)\hat{\phi}(\omega/2) ,$$

where

$$C(\omega) = \sum_i c_i e^{ji\omega} \text{ and } C_k(\omega) = \sum_i c_{k,i} e^{ji\omega} .$$

Thus, we have

$$C(\omega) = \sum_{k \in \Gamma_n} C_k(2\omega)H_k(\omega)$$

and

$$\begin{aligned}\sum_i |c_i|^2 &= \frac{1}{(2\pi)^n}\int_{[0,2\pi]^n} |C(\omega)|^2\,\mathrm{d}\omega \\ &= \sum_{k\in\Gamma_n}\sum_{j\in\Gamma_n}\frac{1}{(2\pi)^n}\int_{[0,2\pi]^n} C_k(2\omega)\overline{C_j(2\omega)} \\ &\qquad\cdot H_k(\omega)\overline{H_j(\omega)}\,\mathrm{d}\omega \\ &= \sum_{k\in\Gamma_n}\sum_{j\in\Gamma_n}\frac{1}{(2\pi)^n}\int_{[0,\pi]^n} C_k(2\omega)\overline{C_j(2\omega)} \\ &\qquad\cdot \sum_{i\in\Gamma_n} H_k(\omega+i\pi)\overline{H_j(\omega+i\pi)}\,\mathrm{d}\omega\end{aligned}$$

which is equal to

$$\sum_{k\in\Gamma_n}\sum_i |c_{k,i}|^2$$

for all $\{c_i\}_{i\in\mathbb{Z}^n}$ if and only if

$$\begin{aligned}&\sum_{i\in\Gamma_n} |H_k(\omega+i\pi)|^2 = 1\,, \quad \forall\ k\in\Gamma_n \\ &\sum_{i\in\Gamma_n} H_k(\omega+i\pi)\overline{H_j(\omega+i\pi)} = 0\,, \quad j\neq k\,.\end{aligned}$$

We shall call this condition the perfect reconstruction condition, which is, in fact, the condition for a perfect reconstruction of an image when using a subband coding scheme in image processing. From the experience of constructing univariate orthonormal wavelets, we would like to choose

$$H_k(\omega) = \mathrm{e}^{j\eta(k)\cdot\omega}\begin{cases} H_0(\omega+k\pi)\,, & \text{if } C(D) \text{ is an integer} \\ \overline{H_0(\omega+k\pi)}\,, & \text{otherwise}\,,\end{cases}$$

for a mapping η from Γ_2 to Γ_2 defined by

$$\begin{aligned}&\eta((0,0)) = (0,0), \quad \eta((1,0)) = (1,1) \\ &\eta((0,1)) = (0,1), \quad \eta((1,1)) = (1,0)\end{aligned}$$

and η from Γ_3 to Γ_3 defined by

$$\begin{aligned}&\eta((0,0,0)) = (0,0,0)\,, \quad \eta((1,0,0)) = (1,1,0)\,, \\ &\eta((0,1,0)) = (0,1,1)\,, \quad \eta((1,1,0)) = (1,0,0)\,, \\ &\eta((0,0,1)) = (1,0,1)\,, \quad \eta((1,0,1)) = (0,0,1)\,, \\ &\eta((0,1,1)) = (0,1,0)\,, \quad \eta((1,1,1)) = (1,1,1)\,.\end{aligned}$$

Here $C(D) = \left(\sum_{\mathbf{x}\in D}\mathbf{x}\right)/2$.

Theorem 31 *Let η be a mapping defined above. Then*

$$(\eta(k)-\eta(j))\cdot(k+j) \text{ is an odd integer } \forall\ k\neq j$$

for all $k, j\in\Gamma_n$ with $n=2$ and $n=3$.

See [44] for a proof. Then we can verify that such chosen H_k's satisfy the perfect reconstruction condition above. With the above construction, we know that ψ_k, $k\in\Gamma_n\backslash\{0\}$ whose Fourier transform is defined by

$$\widehat{\psi}_k(\omega) = H_k(\omega/2)\widehat{\phi}(\omega/2)\,,$$

are wavelets, where

$$H_k(\omega) = \mathrm{e}^{\sqrt{-1}\eta(k)\cdot\omega}\begin{cases} H_0(\omega+k\pi), & \text{if } C(D) \text{ is an integer} \\ \overline{H_0(\omega+k\pi)}, & \text{otherwise}\,.\end{cases}$$

Clearly, their integer translates form an orthonormal basis of W_0, i. e.,

$$W_0 = \operatorname{span}_{L^2}\{\psi_k(\cdot - i)\colon i\in\mathbb{Z}^n, k\in\Gamma_n\backslash\{0\}\}\,.$$

Thus,

$$V_1 = \operatorname{span}_{L^2}\psi_k(\cdot - i)\colon i\in\mathbb{Z}^n, k\in\Gamma_n\}\,.$$

Therefore, we know that

$$\left\{2^{k/2}\psi_i(2^k\cdot - j)\colon j\in\mathbb{Z}^n, i\in\Gamma_n\backslash\{0\}, k\in\mathbb{Z}\right\}$$

is an orthonormal basis of $L^2(\mathbb{R}^n)$, where $n=2$ or $n=3$.

Example 10 Consider $n=2$ and $D = \{(1,0),(0,1),(1,1)\}$. Then ϕ_D is a piecewise linear function which is 1 at $(1,1)$ and 0 at other integers. Let

$$\mathcal{B}(\omega) = \frac{1}{12}[6+2\cos\omega_1+2\cos\omega_2+2\cos(\omega_1+\omega_2)] > 0.$$

Thus, the z-transform of the lowpass filter $H(\mathrm{e}^{j\omega})$ associated with box spline scaling function ϕ is

$$\begin{aligned}H(\mathrm{e}^{j\omega}) &= \sqrt{\frac{\mathcal{B}(\omega)}{\mathcal{B}(2\omega)}}\left(\frac{1+\mathrm{e}^{j\omega_1}}{2}\right)\left(\frac{1+\mathrm{e}^{j\omega_2}}{2}\right) \\ &\qquad\cdot\left(\frac{1+\mathrm{e}^{j(\omega_1+\omega_2)}}{2}\right) \\ &= \sqrt{\frac{\mathcal{B}(\omega)}{\mathcal{B}(2\omega)}}\frac{1}{8}\Big(1+\mathrm{e}^{j\omega_1}+\mathrm{e}^{j\omega_2}+2\mathrm{e}^{j(\omega_1+\omega_2)} \\ &\qquad + \mathrm{e}^{2j\omega_1}\mathrm{e}^{j\omega_2}+\mathrm{e}^{j\omega_1}\mathrm{e}^{2j\omega_2}+\mathrm{e}^{2j(\omega_1+\omega_2)}\Big)\,.\end{aligned}$$

Then, the Fourier transform of orthonormal box spline wavelets ψ_k is given by

$$\hat{\psi}_k(2\omega) = \mathrm{e}^{j\eta(k)\cdot\omega}H((-1)^{|k|}\mathrm{e}^{j\omega})\widehat{\phi_D}(\omega)$$

for $k\in\{(1,0),(0,1),(1,1)\}$. Figures of these wavelets can be found in [12].

Box Spline Tight Wavelet Frames

We begin with the definition of tight wavelet frames based on multiresolution approximation of $L_2(\mathbb{R}^2)$. Given a function $\psi \in L_2(\mathbb{R}^2)$, we set

$$\psi_{j,k}(y) = 2^j \psi(2^j y - k) .$$

Let Ψ be a finite subset of $L_2(\mathbb{R}^2)$ and

$$\Lambda(\Psi) := \{\psi_{j,k}, \psi \in \Psi, j \in Z, k \in Z^2\} .$$

Definition 2 We say that $\Lambda(\Psi)$ is a wavelet frame if there exist two positive numbers A and B such that

$$A\|f\|^2_{L_2(\mathbb{R}^2)} \le \sum_{g\in\Lambda(\Psi)} |\langle f, g\rangle|^2 \le B\|f\|^2_{L_2(\mathbb{R}^2)}$$

for all $f \in L_2(\mathbb{R}^2)$. $\Lambda(\Psi)$ is a tight wavelet frame if it is a wavelet frame with $A = B$. In this case, after a renormalization of the g's in Ψ, we have

$$\sum_{g\in\Lambda(\Psi)} |\langle f, g\rangle|^2 = \|f\|^2_{L_2(\mathbb{R}^2)}$$

for all $f \in L_2(\mathbb{R}^2)$.

It is well known (cf. [18]) that when $\Lambda(\Psi)$ is a tight wavelet frame, any $f \in L_2(\mathbb{R}^2)$ can be recovered from $g \in \Lambda(\Psi)$, i. e.

$$f = \sum_{g\in\Lambda(\Psi)} \langle f, g\rangle g , \quad \forall\, f \in L_2(\mathbb{R}^2) .$$

Furthermore, if $f = \sum_{g\in\Lambda(\Psi)} c_g g$ for some coefficients c_g, then $\sum_{g\in\Lambda(\Psi)} |c_g|^2 \ge \sum_{g\in\Lambda(\Psi)} |\langle f, g\rangle|^2$. That is, the norm of the coefficients $\{\langle f, g\rangle, g \in \Lambda(\Psi)\}$ is the smallest among all the coefficient sequences to represent f.

Let $\phi_D \in L_2(\mathbb{R}^2)$ be a box spline based on the direction set $D \subset \mathbb{R}^2$. Then it is easy to see that

$$\hat{\phi_D}(\omega) = P(\omega/2)\hat{\phi_D}(\omega/2)$$

where $P(\omega)$ is a trigonometric polynomial in $e^{i\omega}$. P is often called the mask of refinable function ϕ_D. To construct a set of tight wavelet framelets $\psi^{(i)}$, $i = 1, \ldots, N$, we look for Q_i (trigonometric polynomial) such that

$$P(\omega)\overline{P(\omega+\ell)} + \sum_{i=0}^{N} Q_i(\omega)\overline{Q_i(\omega+\ell)} = \begin{cases} 1 , & \text{if } \ell = 0 , \\ 0 , & \ell \in \{0,1\}^2\pi\backslash\{0\} . \end{cases} \tag{23}$$

The conditions (23) are called the *Unitary Extension Principle (UEP)* in [19]. With these Q_i's we can define wavelet framelets $\psi^{(i)}$ defined in terms of the Fourier transform by

$$\hat{\psi}^{(i)}(\omega) = Q_i(\omega/2)\hat{\phi_D}(\omega/2) , \quad i = 1, \ldots, r . \tag{24}$$

Then, we know [37], $\Psi = \{\psi^{(i)}, i = 1, \ldots, r\}$ generates a tight frame, i. e., $\Lambda(\Psi)$ is a tight wavelet frame.

Furthermore, letting Q be a rectangular matrix defined by

$$\mathcal{Q} = \begin{bmatrix} Q_1(\xi,\eta) & Q_1(\xi+\pi,\eta) & Q_1(\xi,\eta+\pi) & Q_1(\xi+\pi,\eta+\pi) \\ Q_2(\xi,\eta) & Q_2(\xi+\pi,\eta) & Q_2(\xi,\eta+\pi) & Q_2(\xi+\pi,\eta+\pi) \\ Q_3(\xi,\eta) & Q_3(\xi+\pi,\eta) & Q_3(\xi,\eta+\pi) & Q_3(\xi+\pi,\eta+\pi) \\ Q_4(\xi,\eta) & Q_4(\xi+\pi,\eta) & Q_4(\xi,\eta+\pi) & Q_4(\xi+\pi,\eta+\pi) \end{bmatrix},$$

and $\mathcal{P} = (P(\xi,\eta), P(\xi+\pi,\eta), P(\xi,\eta+\pi), P(\xi+\pi,\eta+\pi))^{\mathsf{T}}$, (23) is simply

$$Q^* Q = I_{4\times 4} - \overline{\mathcal{P}}\mathcal{P}^{\mathsf{T}} . \tag{25}$$

The construction of tight wavelet frames is to find $\mathcal{Q}$ satisfying Eq. (25). Note that the mask P of box spline function ϕ_D satisfies the following *sub-QMF condition*

$$\sum_{\ell\in\{0,1\}^2\pi} |P(\omega+\ell)|^2 \le 1 . \tag{26}$$

There is a constructive method in [37] to find Q_i satisfying Eq. (25). Let us use bivariate box splines to illustrate how to construct $\psi^{(m)}$'s. Let $e_1 = (1,0)$, $e_2 = (0,1)$, $e_3 = e_1 + e_2$, $e_4 = e_1 - e_2$ and let D be a set of these vectors with some repetitions. The box spline ϕ_D associated with direction set D may be defined in terms of refinable equation by

$$\hat{\phi}_D(\omega) = P_D\left(\frac{\omega}{2}\right)\hat{\phi}_D\left(\frac{\omega}{2}\right)$$

where P_D is the mask associated with ϕ_D defined by

$$P_D(\omega) = \prod_{\xi\in D} \frac{1 + e^{-i\xi\cdot\omega}}{2} .$$

Note that the mask P_D satisfies (26). To be more precise, we present an example of tight wavelet framelets based on box spline ϕ_{2211}.

For box spline ϕ_{2211} with $D = \{e_1, e_1, e_2, e_2, e_3, e_4\}$, the graph of ϕ_{2211} is shown in Fig. 33. We have

$$P_{2211}(\omega) = \left(\frac{1+e_1}{2}\right)^2\left(\frac{1+e_2}{2}\right)^2\left(\frac{1+e_3}{2}\right)\left(\frac{1+e_4}{2}\right).$$

It is easy to check that

$$1 - \sum_{\ell\in\{0,1\}^2\pi} |P_{2211}(\omega+\ell)|^2 = \sum_{i=1}^{4} |\widetilde{P}_i(\omega)|^2 ,$$

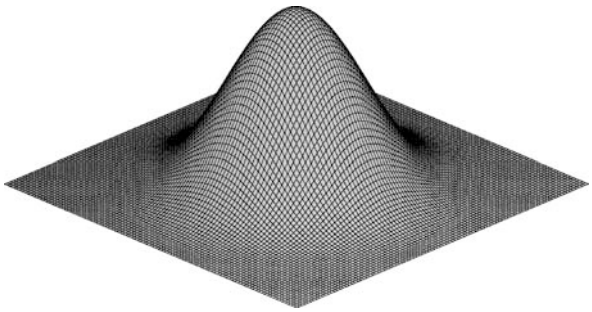

Multivariate Splines and Their Applications, Figure 33
Box spline ϕ_{2211}

where $\widetilde{P}_1(\omega) = \frac{\sqrt{1886}}{224}(1 - e^{4i\omega_1})$,

$$\widetilde{P}_2(\omega) = -\frac{3\sqrt{14}}{64} + \frac{\sqrt{40531922}}{25472} + \frac{3\sqrt{14}}{32}e^{2i\omega_2} - \left(\frac{3\sqrt{14}}{64} + \frac{\sqrt{40531922}}{25472}\right)e^{4i\omega_2}$$

$$\widetilde{P}_3(\omega) = \frac{7\sqrt{2}}{64} + \frac{7\sqrt{2}}{64}e^{4i\omega_2} - \frac{\sqrt{2}}{224}e^{i(4\omega_1+2\omega_2)} - \frac{3\sqrt{2}}{14}e^{2i(\omega_1+\omega_2)},$$

and

$$\widetilde{P}_4(\omega) = \frac{\sqrt{398}}{112} + \frac{\sqrt{398}}{112}e^{4i\omega_1} - \frac{3135\sqrt{398}}{178304}e^{2i\omega_1} - \frac{7\sqrt{398}}{25472}e^{i(2\omega_1+4\omega_2)}.$$

Hence, we will have eight tight frame generators using the constructive steps in [37]. These eight tight frames ψ_m which can be expressed in terms of Fourier transform by

$$\widehat{\psi}_m(\omega) = Q_m(\omega/2)\widehat{\phi}_{2211}(\omega/2),$$

where $Q_m, m = 1, \ldots, 8$ are given in terms of the coefficient matrix as follows: $Q_1 = \sum_{j=0}^{8}\sum_{k=0}^{6} c_{jk}e^{-ij\omega}e^{-ik\xi}$ with

$$[c_{jk}]_{\substack{0\le j\le 8\\0\le k\le 6}} = \frac{-1}{2048}\begin{bmatrix} 0 & 1 & 2 & 2 & 2 & 1 & 0\\ 1 & 4 & 7 & 8 & 7 & 4 & 1\\ 2 & 12 & 22 & 24 & 22 & 12 & 2\\ 7 & 28 & 49 & 56 & 49 & 28 & 7\\ 12 & 38 & 64 & -948 & 64 & 38 & 12\\ 7 & 28 & 49 & 56 & 49 & 28 & 7\\ 2 & 12 & 22 & 24 & 22 & 12 & 2\\ 1 & 4 & 7 & 8 & 7 & 4 & 1\\ 0 & 1 & 2 & 2 & 2 & 1 & 0 \end{bmatrix},$$

$Q_2 = \sum_{j=0}^{6}\sum_{k=0}^{6} c_{jk}e^{-ij\omega}e^{-ik\xi}$ with

$$[c_{jk}]_{\substack{0\le j\le 6\\0\le k\le 6}} = \frac{-1}{512}\begin{bmatrix} 0 & 1 & 2 & 2 & 2 & 1 & 0\\ 1 & 4 & 7 & 8 & 7 & 4 & 1\\ 2 & 7 & 12 & 14 & 12 & 7 & 2\\ 2 & 8 & 14 & -240 & 14 & 8 & 2\\ 2 & 7 & 12 & 14 & 12 & 7 & 2\\ 1 & 4 & 7 & 8 & 7 & 4 & 1\\ 0 & 1 & 2 & 2 & 2 & 1 & 0 \end{bmatrix},$$

$Q_3 = \sum_{j=0}^{8}\sum_{k=0}^{8} c_{jk}e^{-ij\omega}e^{-ik\xi}$ with

$$[c_{jk}]_{\substack{0\le j\le 8\\0\le k\le 8}} = \frac{-1}{1024}\begin{bmatrix} 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0\\ 0 & 0 & 1 & 4 & 6 & 4 & 1 & 0 & 0\\ 0 & 1 & 4 & 11 & 16 & 11 & 4 & 1 & 0\\ 1 & 4 & 11 & 24 & 32 & 24 & 11 & 4 & 1\\ 2 & 6 & 16 & 32 & -472 & 32 & 16 & 6 & 2\\ 1 & 4 & 11 & 24 & 32 & 24 & 11 & 4 & 1\\ 0 & 1 & 4 & 11 & 16 & 11 & 4 & 1 & 0\\ 0 & 0 & 1 & 4 & 6 & 4 & 1 & 0 & 0\\ 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0 \end{bmatrix},$$

$Q_4 = \sum_{j=0}^{6}\sum_{k=0}^{8} c_{jk}e^{-ij\omega}e^{-ik\xi}$ with

$$[c_{jk}]_{\substack{0\le j\le 6\\0\le k\le 8}} = \frac{-1}{2048}\begin{bmatrix} 0 & 1 & 2 & 7 & 12 & 7 & 2 & 1 & 0\\ 1 & 4 & 12 & 28 & 38 & 28 & 12 & 4 & 1\\ 2 & 7 & 22 & 49 & 64 & 49 & 22 & 7 & 2\\ 2 & 8 & 24 & 56 & -948 & 56 & 24 & 8 & 2\\ 2 & 7 & 22 & 49 & 64 & 49 & 22 & 7 & 2\\ 1 & 4 & 12 & 28 & 38 & 28 & 12 & 4 & 1\\ 0 & 1 & 2 & 7 & 12 & 7 & 2 & 1 & 0 \end{bmatrix},$$

$Q_5 = \sum_{j=0}^{8}\sum_{k=0}^{8} c_{jk}e^{-ij\omega}e^{-ik\xi}$ with

$$[c_{jk}]_{\substack{0\le j\le 8\\0\le k\le 8}} = \frac{-\sqrt{2}}{28672} \cdot \begin{bmatrix} 0 & 49 & 98 & 49 & 0 & 49 & 98 & 49 & 0\\ 49 & 196 & 294 & 196 & 98 & 196 & 294 & 196 & 49\\ 98 & 294 & 392 & 198 & 4 & 198 & 392 & 294 & 98\\ 49 & 196 & 198 & -188 & -478 & -188 & 198 & 196 & 49\\ 0 & 49 & -94 & -529 & -772 & -529 & -94 & 49 & 0\\ 0 & 0 & -98 & -392 & -588 & -392 & -98 & 0 & 0\\ 0 & 0 & -4 & -108 & -208 & -108 & -4 & 0 & 0\\ 0 & 0 & -2 & -8 & -12 & -8 & -2 & 0 & 0\\ 0 & 0 & 0 & -2 & -4 & -2 & 0 & 0 & 0 \end{bmatrix},$$

and $Q_6 = \sum_{j=0}^{8}\sum_{k=0}^{8} c_{jk}\mathrm{e}^{-ij\omega}\mathrm{e}^{-ik\xi}$ with

$$[c_{jk}]_{\substack{0\le j\le 8\\ 0\le k\le 8}} = \frac{-\sqrt{398}}{11411456} \cdot \begin{bmatrix} 0 & 1592 & 3184 & 1592 & 0 & 0 & 0 & 0 & 0 \\ 1592 & 6368 & 9552 & 6368 & 1592 & 0 & 0 & 0 & 0 \\ 3184 & 6417 & 6466 & 6417 & 3184 & -49 & -98 & -49 & 0 \\ -1543 & -6172 & -9258 & -6172 & -1592 & -196 & -294 & -196 & -49 \\ -6270 & -15626 & -18712 & -15626 & -6368 & -294 & -392 & -294 & -98 \\ -1543 & -6172 & -9258 & -6172 & -1592 & -196 & -294 & -196 & -49 \\ 3184 & 6417 & 6466 & 6417 & 3184 & -49 & -98 & -49 & 0 \\ 1592 & 6368 & 9552 & 6368 & 1592 & 0 & 0 & 0 & 0 \\ 0 & 1592 & 3184 & 1592 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Q_7 has a complicated expression which is omitted here for simplicity. Finally, we have $Q_8 = \sum_{j=0}^{8}\sum_{k=0}^{5} c_{jk}\mathrm{e}^{-ij\omega}\mathrm{e}^{-ik\xi}$ with

$$[c_{jk}]_{\substack{0\le j\le 8\\ 0\le k\le 5}} = \frac{-\sqrt{1886}}{14336}\begin{bmatrix} 0 & 1 & 2 & 1 & 0 \\ 1 & 4 & 6 & 4 & 1 \\ 2 & 6 & 8 & 6 & 2 \\ 1 & 4 & 6 & 4 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & -4 & -6 & -4 & -1 \\ -2 & -6 & -8 & -6 & -2 \\ -1 & -4 & -6 & -4 & -1 \\ 0 & -1 & -2 & -1 & 0 \end{bmatrix}.$$

These coefficient matrices are high-pass filters associated with low-pass filter P_{2211}. They satisfy (25) which is an exact reconstruction condition. They have been implemented for image denoising and edge detection. They work very well. Numerical experiment results are omitted here.

Open Research Problems

Although there has been much progress in the development of theory and applications of multivariate splines in the last 20 years, there are still many research problems left open. Some of the famous open problems are listed below:

1) **Schumaker's conjecture** Let $T = \langle v_1, v_2, v_3\rangle$ be a triangle in $\mathbb{R}^2$. For a positive integer $d > 0$, let ξ_{ijk}, $i + j + k = d$ be the domain points of degree d on T and $B_{ijk}(x, y)$ be Bézier polynomials of degree d with respect to T. That is,

$$\xi_{ijk} = \frac{1}{d}(iv_1 + jv_2 + kv_3) \quad \text{and} \quad B_{ijk} = \frac{d!}{i!j!k!}b_1^i b_2^j b_3^k$$

for $i + j + k = d$ with b_1, b_2, b_3 being the barycentric coordinates of (x, y). Then the following determinant

$$\left[B_{ijk}(\xi_{i',j',k'})\right]_{(i,j,k),(i',j',k')\in I}$$

is nonzero for any subset I of the triple index set $\{(i, j, k), i + j + k = d\}$. For small values of d, e. g., $d = 1$ and $d = 2$, one can easily verify this conjecture. Schumaker's conjecture claims that the above determinant is not zero for any subset I and for any $d \ge 1$. Similarly, we can consider a tetrahedron T and its domain points and Bézier polynomials. Then Schumaker also claims that the similar determinant is not zero.

2) **Dimension of spline spaces** In the bivariate setting, what is the dimension of spline space $S_d^r(\Delta)$ when $d < 3r + 2$? We note that the dimension of $S_2^1(\Delta_{\mathrm{MS}})$ can be different depending on whether the Morgan–Scott refinement of a triangle is symmetric or not. When $d < 3r + 2$, the dimension of spline space $S_d^r(\Delta)$ can not be equal to one value dependent only on r, d and the numbers of vertices, edges and triangles of Δ. It should be also dependent on the geometry and connectivity properties of Δ. So far we have good lower and upper bounds. The question is how to formulate the geometry and connectivity properties into a dimension formula for $S_d^r(\Delta)$. It is known that when $d < 3r + 2$, the approximation order of $S_d^r(\Delta)$ will not be full for general triangulations Δ. However, it is not known that if the dimension of $S_d^r(\Delta)$ can be determined when $d < 3r + 2$.

3) **Lai's conjecture** The Lai conjecture was formulated by Carl de Boor in [9] based on results in [15]. The conjecture says that if a spline space $S_d^r(\Delta)$ has the full approximation order for a fixed r and d over a fixed triangulation Δ, then $S_{d+k}^r(\Delta)$ has a full approximation order when $k \ge 1$. This conjecture is true when $d = 3r + 2$. It is very interesting to know that it is true for $d < 3r + 2$.

4) **Approximation order of trivariate spline spaces** We do not know the approximation order of trivariate spline spaces for general tetrahedral partitions when $d < 8r + 1$. In [29], Lai tried to establish a result that the approximation order is full for $S_d^r(\Delta)$ when $d \ge 6r + 3$. The result was not published. In addition, there is no evidence that when $d < 6r + 3$, the approximation order is not full. The approximation order of trivariate splines needs to be studied.

5) **Tetrahedral partition for Worsey–Piper's construction of macro-element** In Worsey–Piper's construc-

tion of a C^1 quadratic macro-element, they need a special tetrahedral partition which satisfies the two stringent conditions (Sect. "Construction of Finite Elements and Macro-Elements"). A standing question is how to partition an arbitrary polygonal domain into a tetrahedra such that each tetrahedron of the partition satisfies the Worsey–Piper conditions.

6) **Integration over spherical triangles** Recall that there is a nice formula for integration of bivariate polynomials over a planar triangle (cf. Lemma 1). However, when considering spherical splines, such a formula is still not available.

Bibliography

Primary Literature

1. Alfeld P (1984) A trivariate Clough–Tocher scheme for tetrahedral data. Comput Aided Geom Des 1(1984):169–181
2. Alfeld P, Schumaker LL (2008) Bounds on the dimension of trivariate spline spaces. Advances in Comp Math to appear
3. Awanou G, Lai MJ (2005) On convergence rate of the augmented Lagrangian algorithm for nonsymmetric saddle point problems. Appl Numer Math 54(2):122–134
4. Awanou G, Lai MJ (2005) Trivariate spline approximations of 3D Navier–Stokes equations. Math Comp 74:585–601
5. Awanou G, Lai MJ, Wenston P (2006) The multivariate spline method for scattered data fitting and numerical solution of partial differential equations. In: Chen G, Lai M-J (eds) Wavelets and Splines: Athens 2005. Nashboro Press, Brentwood, pp 24–74
6. Baramidze V, Lai MJ, Shum CK (2006) Spherical Splines for Data Interpolation and Fitting. SIAM J Sci Comput 28:241–259
7. de Boor C (1978) A Practical Guide to Splines. Springer, Berlin
8. de Boor C (1987) B-form basics. In: Farin G (ed) Geometric Modeling: Algorithms and New Trends. SIAM Publication, Philadelphia, pp 131–148
9. de Boor C (1993) Multivariate piecewise polynomials. Acta Numerica 1:65–109
10. de Boor C, Höllig K (1988) Approximation power of smooth bivariate pp functions. Math Z 197:343–363
11. de Boor C, Jia RQ (1993) A sharp upper bound on the approximation order of smooth bivariate pp functions. J Approx Theory 72:24–33
12. de Boor C, Höllig K, Riemenschneider SD (1993) Box Splines. Springer, Berlin
13. Brenner SC, Scott LR (1994) The mathematical theory of finite element methods. Springer, Berlin
14. Chui CK, Lai MJ (1990) Multivariate vertex splines and finite elements. J Approx Theory 60:245–343
15. Chui CK, Lai MJ (1990) On bivariate super vertex splines. Constr Approx 6:399–419
16. Ciarlet PG (1978) The Finite Element Method for Elliptic Problems. North Holland, Amsterdam
17. Clough R, Tocher J (1965) Finite element stiffness matrices for analysis of plates in bending. In: Proc of Conference on Matrix Methods in Structural Analysis, Wright–Patterson Air Force Base
18. Daubechies I (1992) Ten Lectures on Wavelets. SIAM Publications, Philadelphia
19. Daubechies I, Han B, Ron A, Shen ZW (2003) Framelets: MRA-based constructions of wavelet frames. Appl Comp Harmonic Anal 14:1–46
20. Evans L (1998) Partial Differential Equations. American Math Soc, Providence
21. Farin G (1986) Triangular Bernstein–Bézier patches. Comput Aided Geom Design 3:83–127
22. Girault V, Raviart PA (1986) Finite Element Method for Navier–Stokes Equations. Springer, Berlin
23. Grisvard (1985) Elliptic Problems in Nonsmooth Domain. Pitman Publishing Co, Boston
24. He W, Lai MJ (1998) Bivariate Box Spline Wavelets in Sobolev Spaces. In: proceedings of SPIE, vol 3458. pp 56–66
25. He W, Lai MJ (1999) Construction of bivariate compactly supported biorthogonal box spline wavelets with arbitrarily high regularities. Appl Comput Harmon Anal 6:53–74
26. He W, Lai MJ (2003) Construction of trivariate compactly supported biorthogonal box wavelets. J Approx Theory 120:1–19
27. Hong D (1991) Spaces of bivariate spline functions over triangulation. Approx Theory Appl 7:56–75
28. Hu XL, Han D, Lai MJ (2007) Bivariate splines of various degrees for numerical solution of partial differential equations. SIAM J Sci Comput 29:1338–1354
29. Lai MJ (1989) On construction of bivariate and trivariate vertex splines on mixed grid partitions, Ph D Dissertation. Texas A&M University, College Station, Texas
30. Lai MJ (1997) Geometric interpretation of smoothness conditions of triangular polynomial patches. Comput Aided Geom Des 14:191–199
31. Lai MJ (2006) Construction of multivariate compactly supported prewavelets in L_2 spaces and pre-Riesz basis in Sobolev spaces. J Approx Theory 142:83–115
32. Lai MJ (2008) Multivariate splines for data fitting and approximation. In: Neamtu M, SChumaker LL (eds) Approximation XII, San Antonio, 2007, Nashboro press, Brentwood, pp 210–228
33. Lai MJ, Le Mehaute A (2004) A new kind of trivariate C^1 finite element. Adv Comput Math 21:273–292
34. Lai MJ, Schumaker LL (1998) Approximation power of bivariate splines. Adv Comput Math 9:251–279
35. Lai MJ, Schumaker LL (2007) A domain decomposition method for computing bivariate spline fits of scattered data (appear) SIAM J Nam Anal
36. Lai MJ, Schumaker LL (2007) Spline Functions on Triangulations. Cambridge University Press, Cambridge, UK
37. Lai MJ, Stoeckler J (2006) Construction of multivariate compactly supported tight wavelet frames. Appl Comput Harmon Anal 21:324–348
38. Lai MJ, Wenston P (2000) Bivariate spline method for numerical solution of Navier–Stokes equations over polygons in stream function formulation. Numer Methods PDE 16:147–183
39. Lai MJ, Wenston P (2001) Trivariate C^1 cubic splines for numerical solution of biharmonic equation. In: Kopotun K, Lyche T, Neamtu M (eds) Trends in Approximation Theory. Vanderbilt University Press, Nashville, pp 224–234
40. Lai MJ, Wenston P (2004) L_1 spline methods for scattered data interpolation and approximation. Adv Comp Math 21:293–315
41. Lai MJ, Wenston P (2004) Bivariate splines for fluid flows. Comput Fluids 33:1047–1073
42. Le Mehauté A (1984) Interpolation et approximation par des

fonctions polynomiales the computed solutions from the domain decomposition method par morceaux dans $\mathbb{R}^n$, Ph D Thesis. Univ Rennes, Rennes
43. Powell MJD, Sabin MA (1977) Piecewise quadratic approximations on triangles. ACM Trans Math Softw 3:316–325
44. Riemenschneider S, Shen ZW (1991) Box splines, cardinal series, and wavelets. In: Chui CK (ed) Approximation Theory and Functional Analysis. Academic Press, Boston, pp 133–149
45. Sander G (1964) Bornes supérieures et inférieures dans l'analyse matricielle des plaques en flexion-torsion. Bull Soc R Sci Liège 33:456–494
46. Schumaker LL (1979) On the dimension of spaces of piecewise polynomials in two variables. In: Schempp W, Zeller K (eds) Multivariate Approximation Theory. Birkhäuser, Basel, pp 396–412
47. Schumaker LL (1981) Spline Functions: Basic Theory. Wiley, New York
48. Schumaker LL (2008) Computing bivariate splines in scattered data fitting and the finite-element method. Numer Algs 48(2008):237–260
49. Schumaker LL, Sorokina T, Worsey A (2008) A C^1 quadratic trivariate macro-element space defined over arbitrary tetrahedral partition. submitted 2008
50. Temam R (1984) Navier–Stokes Equations. Theory and Numerical Analysis. North-Holland Publishing Co, Amsterdam
51. von Golitchek M, Schumaker LL (2002) Bounds on projections onto bivariate polynomial spline spaces with stable local bases. Constr Approx 18:241–254
52. von Golitschek M, Schumaker LL (2002) Penalized least squares fitting. Serdica 18:1001–1020
53. von Golitschek M, Lai MJ, Schumaker LL (2002) Bounds for minimal energy bivariate polynomial splines. Numer Math 93:315–331
54. Worsey AJ, Farin G (1987) An n-dimensional Clough–Tocher interpolant. Constr Approx 3:99–110
55. Worsey AJ, Piper B (1988) A trivariate Powell–Sabin interpolant. Comp Aided Geom Des 5:177–186
56. Ženiček A (1973) Polynomial approximation on tetrahedrons in the finite element method. J Approx Theory 7:34–351

Books and Reviews

Chui CK (1988) Multivariate Splines. CBMS Lectures, SIAM Publication, Philadelphia
Dierckx P (1985) Curve and Surface Fitting with Splines. Oxford University Press, Oxford
Späth H (1995) Two Dimensional Spline Interpolation Algorithms. AK Peters, Wellesley

Multiwavelets

Fritz Keinert
Iowa State University, Ames, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Refinable Function Vectors
Multiresolution Approximation and Discrete Multiwavelet Transform
Moments and Approximation Order
The Discrete Multiwavelet Transform (DMWT) Algorithm
Pre- and Postprocessing and Balanced Multiwavelets
Boundary Handling
Applications
Polyphase Factorization
Lifting
Two-Scale Similarity Transform (TST)
Examples and Software
Future Directions
Bibliography

Glossary

Balanced multiwavelet A multiwavelet for which the expansion coefficients for polynomials up to a certain degree are polynomial sequences. Balanced multiwavelets do not require pre- or postprocessing.

Discrete multiwavelet transform (DMWT) The algorithm which decomposes a signal into a coarse approximation and fine detail at several levels. A direct generalization of the Discrete Wavelet Transform (DWT) for scalar wavelets.

Lifting A method for modifying an existing multiwavelet, or building one from scratch. Lifting can be used to impose approximation order, balancing, or symmetry.

Modulation matrix The modulation matrix

$$M(\xi) = \begin{pmatrix} H(\xi) & H(\xi+\pi) \\ G(\xi) & G(\xi+\pi) \end{pmatrix},$$

can be used to describe the action of the DMWT. It is also used in the construction of multiwavelets by lifting or TST.

Multiresolution, multiresolution approximation (MRA) Multiresolution is the fundamental concept underlying everything related to any kind of wavelet. A signal is decomposed into a low resolution approximation, plus fine detail at one or more levels of resolution. An MRA is a nested chain of subspaces of L^2 which describes this concept mathematically.

Multiscaling function, multiwavelet function, multiwavelet A multiwavelet of multiplicity r consists of a multiscaling function and a multiwavelet function. Both are functions from $\mathbb{R}$ to $\mathbb{C}^r$, that is, function vectors. Multiwavelets are not the same as multivariate wavelets, which are functions from $\mathbb{R}^n$ to $\mathbb{C}$.

Polyphase decomposition, polyphase matrix The polyphase decomposition separates a coefficient sequence into two sequences, by even and odd subscripts. It can be used to describe or implement the DMWT. The polyphase matrix

$$P(z) = \begin{pmatrix} H^{(0)}(z) & H^{(1)}(z) \\ G^{(0)}(z) & G^{(1)}(z) \end{pmatrix},$$

is useful in the construction of multiwavelets, especially orthogonal multiwavelets.

Pre- and postprocessing Before a DMWT can be applied to a signal, the signal needs to be written as a multiscaling function series. The DMWT works on the coefficients of the series expansion, not on the values of the signal. For scalar wavelets the distinction can usually be ignored, but not for multiwavelets. Having a multiwavelet of approximation order p does not mean that the coefficients of a polynomial up to order $p-1$ form a polynomial sequence. Preprocessing converts point samples into expansion coefficients. Postprocessing does the opposite.

Refinable function vector, refinement equation A refinable function vector of multiplicity r is a function from $\mathbb{R}$ to $\mathbb{C}^r$ which satisfies a refinement equation of the form

$$\boldsymbol{\phi}(x) = \sqrt{2} \sum_{k=k_0}^{k_1} H_k \, \boldsymbol{\phi}(2x-k)$$

with recursion coefficients H_k which are $r \times r$ matrices.

Symbol Given any sequence $\mathbf{a} = \{a_k\}$, the symbol of $\mathbf{a}$ is defined as

$$a(\xi) = \sum_k a_k \mathrm{e}^{-i\xi} \quad \text{or} \quad a(z) = \sum_k a_k z^k, \quad z = \mathrm{e}^{-i\xi},$$

possibly with a normalizing factor. The sequence may represent point samples of a signal, the recursion coefficients of a two-scale refinement equation, or other quantities. Both the trigonometric and polynomial notations are useful, depending on the setting; it is trivial to switch back and forth between them.

Two-scale similarity transform (TST) The TST is a way of moving approximation orders in a biorthogonal multiwavelet pair from one side to the other. For scalar wavelets it corresponds to moving a factor of $(1+\mathrm{e}^{-i\xi})/2$ back and forth. The TST can be used to modify an existing multiwavelet, or to build multiwavelets from scratch. It can also be used to impose or characterize symmetry, approximation order and balancing.

Definition of the Subject

Classical (scalar) wavelets have been around since the late 1980s and have become an indispensable tool in signal processing, with further applications in numerical analysis, operator theory, and other fields. Wavelets have been generalized in many ways: wavelet packets, multivariate wavelets, ridgelets, curvelets, vaguelettes, slantlets, second generation wavelets, frames, and other constructions.

One such generalization are multiwavelets, which have been around since the early 1990s. Multiwavelets use several scaling and wavelet functions. Their construction is more complicated than that of scalar wavelets, but the underlying multiresolution concepts and the decomposition and reconstruction algorithms are very similar.

Scalar wavelets are functions from $\mathbb{R}$ to $\mathbb{C}$. Multiwavelets are functions from $\mathbb{R}$ to $\mathbb{C}^r$. They include scalar wavelets as the special case $r = 1$. Multiwavelets are used to analyze one-dimensional signals, or higher-dimensional signals by using tensor products, just like scalar wavelets. They should not be confused with multivariate wavelets, which are functions from $\mathbb{R}^n$ to $\mathbb{C}$, used to analyze higher-dimensional signals.

Multiwavelets have several advantages over scalar wavelets: they can have short support coupled with high smoothness and high approximation order, and they can be both symmetric and orthogonal. They also have some disadvantages: the algorithms require preprocessing and postprocessing steps.

The applications are the same as for scalar wavelets: signal compression, signal denoising, fast operator evaluation in numerical analysis, Galerkin methods for differential and integral equations. Performance of multiwavelets is similar to that of scalar wavelets, but implementation requires a bit more effort, especially because of the need for pre- and postprocessing. Multiwavelets are best used in situations where their advantages (symmetry or short support) outweigh the extra effort.

Introduction

The fundamental concept underlying everything related to any kind of wavelet is *multiresolution*. A function (or signal or image) is decomposed into a low resolution approximation plus fine detail at one or more levels of resolution (see Fig. 3). Scalar wavelets use a scaling function for the coarse approximation, a wavelet function for the fine detail. Multiwavelets use several scaling functions and wavelet functions, combined into function vectors.

The first occurrence of multiwavelets is in the work of Alpert [1], which uses piecewise polynomial multiwavelets of high multiplicity. The Donovan–Geronimo–Hardin–

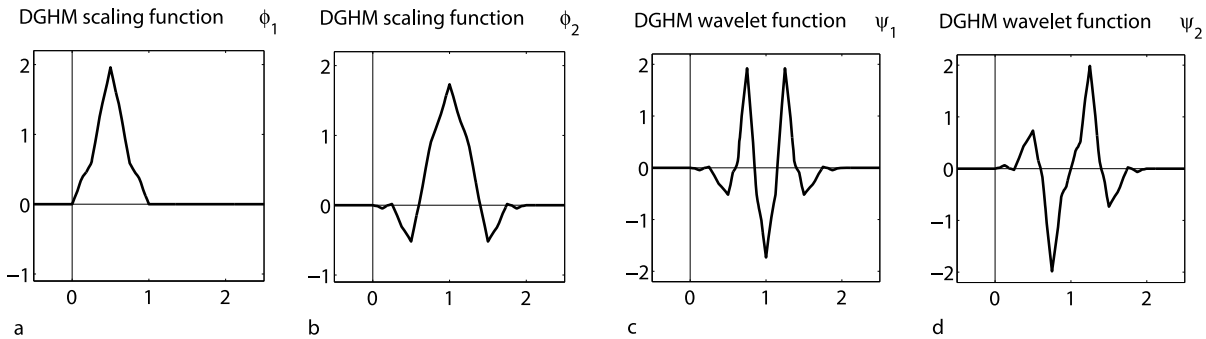

Multiwavelets, Figure 1
DGHM multiwavelet

Massopust (DGHM) multiwavelet [21] is commonly considered to be the first nontrivial example.

The DGHM multiscaling function is a vector

$$\boldsymbol{\phi}(x) = \begin{pmatrix} \phi_1(x) \\ \phi_2(x) \end{pmatrix}$$

which satisfies a recursion relation

$$\begin{aligned} \boldsymbol{\phi}(x) = \frac{1}{20}\Bigg[& \begin{pmatrix} 12 & 16\sqrt{2} \\ -\sqrt{2} & -6 \end{pmatrix} \boldsymbol{\phi}(2x) \\ & + \begin{pmatrix} 12 & 0 \\ 9\sqrt{2} & 20 \end{pmatrix} \boldsymbol{\phi}(2x-1) \\ & + \begin{pmatrix} 0 & 0 \\ 9\sqrt{2} & -6 \end{pmatrix} \boldsymbol{\phi}(2x-2) \\ & + \begin{pmatrix} 0 & 0 \\ -\sqrt{2} & 0 \end{pmatrix} \boldsymbol{\phi}(2x-3) \Bigg] . \end{aligned}$$

The first scaling function ϕ_1 is supported on $[0, 1]$ and is symmetric about $x = 1/2$. The second scaling function ϕ_2 is supported on $[0, 2]$, symmetric about $x = 1$. These functions and their integer translates are orthonormal, that is,

$$\int \phi_i(x)\phi_j(x-k)\mathrm{d}x = \delta_{ij}\delta_{0k} .$$

They have approximation order 2, which means that $f(x) = 1$ and $f(x) = x$ can be written as linear combinations of integer shifts of ϕ_1, ϕ_2. The scaling functions as well as the corresponding wavelet functions are shown in Fig. 1. The wavelet functions have support on $[0, 2]$ and are symmetric/antisymmetric about $x = 1$. They are also orthogonal to each other and to the ϕ_j.

Many other examples have been constructed since then. Much of the theory is due to Strela and Plonka, in their individual and joint papers in the late 1990s.

The most comprehensive treatment of multiwavelets in the literature is the book [34]. Useful survey articles include [48,55].

The theory of multiwavelets parallels the theory of scalar wavelets. It is highly recommended that anyone studying multiwavelets become familiar with scalar wavelets first. Good introductions to scalar wavelets can be found in [15,34], or [54].

This article assumes that the reader is familiar with scalar wavelets. It is brief on the aspects of multiwavelets which are essentially the same as for scalar wavelets, and gives more details on the areas where they differ. Similarities and differences are pointed out in many places.

The main body of this article is divided into three logical parts. The first part (Sects. "Refinable Function Vectors" through "Moments and Approximation Order") describes the basic theory and properties of the multiscaling and multiwavelet functions. The second part (Sects. "The Discrete Multiwavelet Transform (DMWT) Algorithm" through "Applications") describes the practical implementation of the Discrete Multiwavelet Transform (DMWT). The third part (Sects. "Polyphase Factorization" through "Two-Scale Similarity Transform (TST)") describes techniques for building multiwavelets, or modifying existing multiwavelets.

For conciseness, we make some simplifying assumptions:

- The dilation factor is 2, not a general $m \geq 2$.
- All functions satisfy recursion relations with finitely many coefficients and have compact support.
- All multiscaling functions satisfy the basic regularity conditions described in Sect. "Refinable Function Vectors", and lie in L^1 and L^2.

Refinable Function Vectors

A *refinable function vector* is a vector-valued function

$$\boldsymbol{\phi}(x) = \begin{pmatrix} \phi_1(x) \\ \vdots \\ \phi_r(x) \end{pmatrix}, \qquad \phi_j : \mathbb{R} \to \mathbb{C} ,$$

which satisfies a *two-scale matrix refinement equation* of the form

$$\boldsymbol{\phi}(x) = \sqrt{2} \sum_{k=k_0}^{k_1} H_k \, \boldsymbol{\phi}(2x - k) \,. \tag{1}$$

r is called the *multiplicity* of $\boldsymbol{\phi}$. The *recursion coefficients* H_k are $r \times r$ matrices. Scalar wavelets are a special case, for $r = 1$.

A pair $\boldsymbol{\phi}$, $\tilde{\boldsymbol{\phi}}$ of refinable function vectors is called *biorthogonal* if

$$\langle \boldsymbol{\phi}(x), \tilde{\boldsymbol{\phi}}(x-k) \rangle = \int \boldsymbol{\phi}(x) \tilde{\boldsymbol{\phi}}(x-k)^* \mathrm{d}x = \delta_{0k} I \,.$$

Here $\tilde{\boldsymbol{\phi}}^*$ is the complex conjugate transpose of $\tilde{\boldsymbol{\phi}}$, so this inner product produces an $r \times r$ matrix. If $\boldsymbol{\phi}$ is biorthogonal to itself, it is called *orthogonal*.

Note that the refinement Eq. (1) is similar to an eigenvalue problem: if $\boldsymbol{\phi}$ is a solution, so is any multiple of $\boldsymbol{\phi}$.

There has to be a factor of 2 or $\sqrt{2}$ in all formulas based on Eq. (1), just as there has to be a factor of 2π somewhere in every definition of the Fourier transform, but different authors put it in different places. Formulas from other sources may differ slightly from those in this article.

The support of $\boldsymbol{\phi}$ is contained in the interval $[k_0, k_1]$. It may be strictly shorter if the first or last recursion coefficient H_{k_0}, H_{k_1} is nilpotent. This happens in the case of the DGHM multiwavelet.

The *symbol* of a refinable function vector is the trigonometric matrix polynomial

$$H(\xi) = \frac{1}{\sqrt{2}} \sum_{k=k_0}^{k_1} H_k \mathrm{e}^{-ik\xi} \,. \tag{2}$$

Equivalently we could use the matrix polynomial

$$H(z) = \frac{1}{\sqrt{2}} \sum_{k=k_0}^{k_1} H_k z^k \,, \quad z = \mathrm{e}^{-i\xi}$$

The refinement Eq. (1) can only have an L^2-solution which leads to stable algorithms if it satisfies the *Basic Regularity Conditions*:

- $H(0)$ satisfies *Condition E*. That is, $H(0)$ has a simple eigenvalue of 1, and all other eigenvalues are less than 1 in magnitude.
- The coefficients H_k satisfy the *sum rules of order 1*. That is, $\mathbf{y}_0^* H(\pi) = \mathbf{0}$, where $\mathbf{y}_0^*$ is the left eigenvector of $H(0)$ to eigenvalue 1. (See Eq. (6) for general sum rules).

The three main ways to prove existence and uniqueness of $\boldsymbol{\phi}$ and obtain smoothness estimates carry over from scalar wavelets: infinite product, Cascade Algorithm, and eigenvalue problem. The second and third methods are also practical ways of obtaining point values and graphs of $\boldsymbol{\phi}$.

We define the Fourier transform as

$$\hat{f}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) \mathrm{e}^{-ix\xi} \mathrm{d}x \,.$$

The Fourier transform of refinement Eq. (1) is

$$\hat{\boldsymbol{\phi}}(\xi) = H\left(\frac{\xi}{2}\right) \hat{\boldsymbol{\phi}}\left(\frac{\xi}{2}\right) \,.$$

This leads to the formal infinite product

$$\hat{\boldsymbol{\phi}}(\xi) = \left[\prod_{k=1}^{\infty} H(2^{-k}\xi) \right] \hat{\boldsymbol{\phi}}(0) \,.$$

The infinite product can only rarely be evaluated in closed form, but its convergence can be studied. Since everything is done on the Fourier transform side, this approach can be used to investigate distribution solutions.

The *Cascade Algorithm* is fixed point iteration applied to the recursion relation. Select a suitable starting function $\boldsymbol{\phi}^{(0)}(x)$, and iterate:

$$\boldsymbol{\phi}^{(n+1)}(x) = \sqrt{2} \sum_{k=k_0}^{k_1} H_k \boldsymbol{\phi}^{(n)}(2x - k) \,.$$

The *transition operator* or *transfer operator* for the symbol $H(\xi)$ is defined by

$$TF(\xi) = H(\xi)F(\xi)H(\xi)^* + H(\xi+\pi)F(\xi+\pi)H(\xi+\pi)^* \,. \tag{3}$$

If the transition operator satisfies condition E, the cascade algorithm converges for any starting function $\boldsymbol{\phi}^{(0)}$ which satisfies

$$\mathbf{y}_0^* \sum_{k=k_0}^{k_1} \boldsymbol{\phi}^{(0)}(k) = c \neq 0 \,.$$

This condition essentially states that $\boldsymbol{\phi}^{(0)}$ must have a component in the direction of $\boldsymbol{\phi}$. Compare Eq. (5).

A third approach is the *eigenvalue method*. Write out the refinement equation at all integer points in the support:

$$\begin{aligned} \boldsymbol{\phi}(j) &= \sqrt{2} \sum_{k=k_0}^{k_1} H_k \, \boldsymbol{\phi}(2j - k) \\ &= \sqrt{2} \sum_{k=k_0}^{k_1} H_{2j-k} \, \boldsymbol{\phi}(k) \,, \quad j = k_0, \ldots, k_1 \,. \end{aligned}$$

This is an eigenvalue problem

$$\boldsymbol{\phi} = T\boldsymbol{\phi}\,,$$

where

$$\boldsymbol{\phi} = \begin{pmatrix} \boldsymbol{\phi}(k_0) \\ \boldsymbol{\phi}(k_0+1) \\ \vdots \\ \boldsymbol{\phi}(k_1) \end{pmatrix}, \quad T_{jk} = \sqrt{2}\,H_{2j-k}, \quad k_0 \le j, k \le k_1\,.$$

The basic regularity conditions guarantee that $(\mathbf{y}_0^*, \mathbf{y}_0^*, \ldots, \mathbf{y}_0^*)$ is a left eigenvector to eigenvalue 1, so a right eigenvector also exists. The right eigenvector is often unique, but not always, so this method can fail.

Unless H_{k_0} or H_{k_1} have an eigenvalue of $1/\sqrt{2}$, the values of $\boldsymbol{\phi}$ at k_0 and k_1 are zero, and we can reduce the size of $\boldsymbol{\phi}$ and T.

Once the values of $\boldsymbol{\phi}$ at the integers have been determined, we can use the refinement equation to obtain values at points of the form $k/2$, $k \in \mathbb{Z}$, then $k/2^2$, and so on to any desired resolution.

A necessary condition for biorthogonality is

$$\sum_k H_k \tilde{H}_{k-2j}^* = \delta_{0j} I\,,$$

or equivalently

$$H(\xi)\tilde{H}(\xi)^* + H(\xi+\pi)\tilde{H}(\xi+\pi)^* = I\,.$$

These conditions are sufficient if the cascade algorithm for $\boldsymbol{\phi}, \tilde{\boldsymbol{\phi}}$ converges.

Multiresolution Approximation and Discrete Multiwavelet Transform

The contents of this section parallel corresponding results for scalar wavelets.

Definition 1 A *Multiresolution Approximation (MRA)* of L^2 is a doubly infinite nested sequence of subspaces of L^2

$$\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \ldots$$

with properties

(i) $\bigcup_n V_n$ is dense in L^2
(ii) $\bigcap_n V_n = \{0\}$
(iii) $f(x) \in V_n \iff f(2x) \in V_{n+1}$ for all $n \in \mathbb{Z}$
(iv) $f(x) \in V_n \iff f(x - 2^{-n}k) \in V_n$ for all $n, k \in \mathbb{Z}$
(v) There exists a function vector $\boldsymbol{\phi} \in L^2$ so that

$$\{\phi_j(x-k) : j = 1, \ldots, r,\ k \in \mathbb{Z}\}$$

forms a stable basis of V_0.

The vector of basis functions $\boldsymbol{\phi}$ is called the *multiscaling function*. The MRA is called *orthogonal* if $\boldsymbol{\phi}$ is orthogonal.

Condition (v) means that any $f \in V_0$ can be written uniquely as

$$f(x) = \sum_{k\in\mathbb{Z}} \mathbf{f}_k^*\, \boldsymbol{\phi}(x-k)$$

with convergence in the L^2-sense; and there exist constants $0 < A \le B$, independent of f, so that

$$A \sum_k \|\mathbf{f}_k\|_2^2 \le \|f\|_2^2 \le B \sum_k \|\mathbf{f}_k\|_2^2\,.$$

Condition (iii) expresses the main property of an MRA: each V_n consists of the functions in V_0 compressed by a factor of 2^n. Thus, a stable basis of V_n is given by $\{\boldsymbol{\phi}_{n,k} : k \in \mathbb{Z}\}$, where

$$\boldsymbol{\phi}_{n,k}(x) = 2^{n/2}\boldsymbol{\phi}(2^n x - k)\,.$$

The factor $2^{n/2}$ preserves the L^2-norm.

Since $V_0 \subset V_1$, $\boldsymbol{\phi}$ can be written in terms of the basis of V_1 as

$$\boldsymbol{\phi}(x) = \sum_k H_k \boldsymbol{\phi}_{1k}(x) = \sqrt{2}\sum_k H_k \boldsymbol{\phi}(2x-k)$$

for some coefficient matrices H_k. In other words, $\boldsymbol{\phi}$ is refinable.

If the MRA is not orthogonal, further development requires the existence of a second MRA based on a *dual multiscaling function* $\tilde{\boldsymbol{\phi}}$ biorthogonal to $\boldsymbol{\phi}$.

The projection of an arbitrary function $f \in L^2$ onto V_n is given by

$$P_n f = \sum_k \langle f, \tilde{\boldsymbol{\phi}}_{n,k}\rangle \boldsymbol{\phi}_{n,k}\,.$$

The basis functions $\phi_{n,k}$ are shifted in steps of 2^{-n} as k varies, so $P_n f$ cannot represent any detail on a scale smaller than that. We say that the functions in V_n have *resolution* 2^{-n} or *scale* 2^{-n}. $P_n f$ is called an *approximation to f at resolution* 2^{-n}. An MRA provides a sequence of approximations $P_n f$ of increasing accuracy to a given function f. For $f \in L^2$, $P_n f \to f$ in L^2 as $n \to \infty$.

The true power of the multiresolution approach arises from considering the differences between approximations at different levels. The difference between the approximations at resolution 2^{-n} and 2^{-n-1} is called the *fine detail at resolution* 2^{-n}:

$$Q_n f(x) = P_{n+1} f(x) - P_n f(x)\,.$$

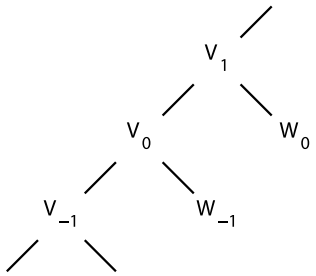

Multiwavelets, Figure 2
The spaces V_n and W_n

Q_n is also a projection (orthogonal if the MRA is orthogonal). Its range W_n satisfies

$$V_n \oplus W_n = V_{n+1} .$$

The two sequences of spaces $\{V_n\}$ and $\{W_n\}$ and their relationships can be graphically represented as in Fig. 2

The sequence of spaces $\{W_n\}$ satisfies conditions similar to conditions (i) through (v) of an MRA, except that the W_n are linearly independent (mutually orthogonal if the MRA is orthogonal) instead of nested.

A basis for W_0 is given by the integer translates of a function vector $\boldsymbol{\psi}$ called the *multiwavelet function*. The multiscaling function $\boldsymbol{\phi}$ and multiwavelet function $\boldsymbol{\psi}$ together form a *multiwavelet*.

For multiwavelets there is no simple formula for finding $\boldsymbol{\psi}$, like in the scalar case. A construction for finding $\boldsymbol{\psi}$ is given in Sect. "Polyphase Factorization".

In terms of the multiwavelet functions, the projection Q_n is given by

$$Q_n f = \sum_k \langle f, \tilde{\boldsymbol{\psi}}_{n,k} \rangle \boldsymbol{\psi}_{n,k} .$$

We now come to the main concept: the *Discrete Multiwavelet Transform (DMWT)*.

Given a function $f \in L^2$, we can represent it as

$$f = \sum_{k=-\infty}^{\infty} Q_k f$$

(complete decomposition in terms of detail at all levels), or we can start at any level N and represent f by its the approximation at resolution 2^{-N} plus all the detail at finer resolution:

$$f = P_N f + \sum_{k=N}^{\infty} Q_k f .$$

For practical applications, we need to reduce this to a finite sum. We replace f by $P_n f$ for some n. Then

$$P_n f = P_N f + \sum_{k=N}^{n-1} Q_k f .$$

This equation describes the DMWT: a high-resolution approximation $P_n f$ to the original function or signal f gets decomposed into a coarse approximation $P_N f$, and fine detail at several resolutions. See Fig. 3 for illustration. The decomposition as well as the reconstruction can be performed very efficiently on a computer. Implementation details are presented in Sect. "The Discrete Multiwavelet Transform (DMWT) Algorithm".

Moments and Approximation Order

One of the main properties of interest is the *approximation order* of a multiscaling function. A high approximation order is the basis for good performance in data compression and other applications. The results in this section are similar to corresponding results for scalar wavelets.

Definition 2 The kth *discrete moment* of $\boldsymbol{\phi}, \boldsymbol{\psi}$ is defined by

$$M_k = \frac{1}{\sqrt{2}} \sum_j j^k H_j , \quad N_k = \frac{1}{\sqrt{2}} \sum_j j^k G_j .$$

Discrete moments are $r \times r$ matrices. They are uniquely defined and easy to calculate. In particular, $M_0 = H(0)$.

Definition 3 The kth *continuous moment* of $\boldsymbol{\phi}, \boldsymbol{\psi}$ is

$$\boldsymbol{\mu}_k = \int x^k \boldsymbol{\phi}(x) \mathrm{d}x , \quad \boldsymbol{\nu}_k = \int x^k \boldsymbol{\psi}(x) \mathrm{d}x .$$

Continuous moments are r-vectors.

Continuous and discrete moments are related by

$$\begin{aligned} \boldsymbol{\mu}_k &= m^{-k} \sum_{t=0}^{k} \binom{k}{t} M_{k-t} \boldsymbol{\mu}_t , \\ \boldsymbol{\nu}_k &= m^{-k} \sum_{t=0}^{k} \binom{k}{t} N_{k-t} \boldsymbol{\mu}_t . \end{aligned} \tag{4}$$

In particular,

$$\boldsymbol{\mu}_0 = M_0 \boldsymbol{\mu}_0 = H(0) \boldsymbol{\mu}_0 .$$

$\boldsymbol{\mu}_0$ is only defined up to a constant multiple. Its scaling depends on the scaling of $\boldsymbol{\phi}$. For a biorthogonal pair $\boldsymbol{\phi}, \tilde{\boldsymbol{\phi}}$, the correct scaling is given by

$$\tilde{\boldsymbol{\mu}}_0^* \left(\sum_k \boldsymbol{\phi}(k) \right) = \tilde{\boldsymbol{\mu}}_0^* \left(\int \boldsymbol{\phi}(x) \mathrm{d}x \right) = \tilde{\boldsymbol{\mu}}_0^* \boldsymbol{\mu}_0 = 1 . \tag{5}$$

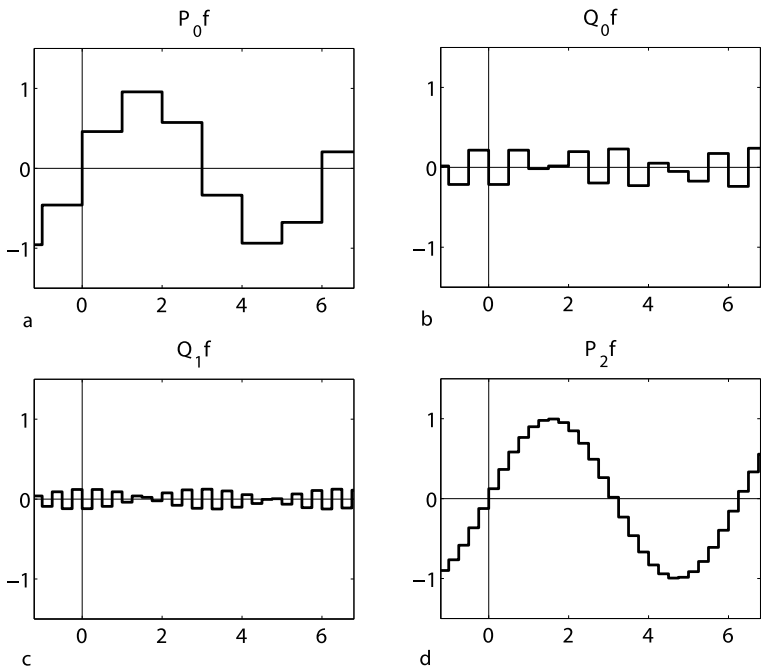

Multiwavelets, Figure 3
P_0f* (a), *Q_0f* (b), *Q_1f* (c), and *$P_2f = P_0f + Q_0f + Q_1f$* (d) for *$f(x) = \sin x$

Unlike the scalar case, the sum of point values at the integers and the integral do not have to be the same. They just have to have the same inner product with $\tilde{\boldsymbol{\mu}}_0$ (which is the same as $\mathbf{y}_0$ defined below).

For orthogonal $\boldsymbol{\phi}$,

$$\|\boldsymbol{\phi}\|^2 = \sum_{k=1}^{r} \|\phi_k\|^2 = r\,, \quad \|\boldsymbol{\mu}_0\| = 1\,.$$

For biorthogonal multiwavelets we cannot in general achieve both $\|\boldsymbol{\mu}_0\| = 1$ and $\|\tilde{\boldsymbol{\mu}}_0\| = 1$.

Once $\boldsymbol{\mu}_0$ has been chosen, all other continuous moments are uniquely defined and can be computed from Eqs. (4).

Definition 4 The multiscaling function $\boldsymbol{\phi}$ provides *approximation order p* if

$$\|f(x) - P_n f(x)\| = O(2^{-np})\,, \quad \|Q_n f\| = O(2^{-np})\,.$$

whenever f has p continuous derivatives.

Definition 5 $\boldsymbol{\phi}$ has *accuracy p* if all polynomials up to order $p-1$ can be represented as

$$x^n = \sum_k \mathbf{c}^*_{n,k} \boldsymbol{\phi}(x-k), \quad n = 0, \ldots, p-1$$

for some coefficient vectors $\mathbf{c}_{n,k}$.

Definition 6 The recursion coefficients $\{H_k\}$ satisfy the *sum rules of order p* if there exist vectors $\mathbf{y}_0, \ldots, \mathbf{y}_{p-1}$ with $\mathbf{y}_0 \neq \mathbf{0}$, which satisfy

$$\begin{aligned} \sum_{t=0}^{n} \binom{n}{t} 2^t (-i)^{n-t} \mathbf{y}_t^* D^{n-t} H(0) &= \mathbf{y}_n^*\,, \\ \sum_{t=0}^{n} \binom{n}{t} 2^t (-i)^{n-t} \mathbf{y}_t^* D^{n-t} H(\pi) &= \mathbf{0}^*\,, \end{aligned} \tag{6}$$

for $n = 0, \ldots, p-1$. D stands for the derivative operator. The vectors $\mathbf{y}_n$ are called the *approximation vectors*. Note that $\mathbf{y}_0 = \tilde{\boldsymbol{\mu}}_0$.

For scalar wavelets, the sum rules of order p reduce to "$H(\xi)$ has a zero of order p at $\xi = \pi$."

As in the scalar case, approximation order p, accuracy p, and the sum rules of order p are equivalent for sufficiently regular $\boldsymbol{\phi}$. They are also equivalent to the fact that the dual multiwavelet function has p vanishing moments, except that we need to specify vanishing *continuous* moments. In the scalar case, vanishing continuous moments are equivalent to vanishing discrete moments, but in the multiwavelet case the discrete moments are matrices. They have to annihilate certain vectors, but they do not have to be zero matrices.

For multiwavelets, accuracy p does *not* mean that the DMWT preserves polynomial sequences up to order $p-1$. See Sect. "Pre- and Postprocessing and Balanced Multiwavelets" for details.

Approximation order p is also equivalent to a certain factorization of the symbol, but not as simple as in the scalar case. This TST factorization requires a lot of machinery, and will be presented in Sect. "Two-Scale Similarity Transform (TST)".

The Discrete Multiwavelet Transform (DMWT) Algorithm

In this section we describe the implementation of the DMWT. The idea behind it was already explained in Sect. "Multiresolution Approximation and Discrete Multiwavelet Transform". We describe it in four different ways. All of them, except for the modulation formulation, can be used as the basis of a computer implementation.

We assume that the original signal is $s(x)$. The algorithm starts at some resolution level n with the coefficient sequence $\mathbf{s}_n = \{\mathbf{s}_{n,k}\}$ from

$$P_n s(x) = \sum_k \langle s, \tilde{\boldsymbol{\phi}}_{n,k}\rangle \boldsymbol{\phi}_{n,k}(x) = \sum_k \mathbf{s}^*_{n,k} \boldsymbol{\phi}_{n,k}(x) \,.$$

The decomposed signal consists of $\mathbf{s}_{n-1}$, $\mathbf{d}_{n-1}$. Note that these are sequences of *vectors*, and the recursion coefficients are matrices. We can interpret the algorithm in terms of convolutions and down- and upsampling as in the scalar case, but they are *block* convolutions and *block* down- and upsampling.

The operation count for the complete algorithm is $O(N)$, as in the scalar case. All formulations of the algorithm are straightforward generalizations of the scalar Discrete Wavelet Transform.

Direct Formulation

Decomposition:

$$\mathbf{s}_{n-1,j} = \sum_k \tilde{H}_{k-2j} \mathbf{s}_{n,k} \,,$$

$$\mathbf{d}_{n-1,j} = \sum_k \tilde{G}_{k-2j} \mathbf{s}_{n,k} \,.$$

Reconstruction:

$$\mathbf{s}_{n,k} = \sum_j H^*_{k-2j} \mathbf{s}_{n-1,j} + \sum_j G^*_{k-2j} \mathbf{d}_{n-1,j} \,.$$

Matrix Formulation

The decomposition and reconstruction steps can be interpreted as infinite matrix-vector products. The formulation becomes nicer if we interleave the **s**- and **d**-coefficients:

$$(\mathbf{sd})_{n-1,j} = \begin{pmatrix} \mathbf{s}_{n-1,j} \\ \mathbf{d}_{n-1,j} \end{pmatrix}, \qquad \tilde{L}_k = \begin{pmatrix} \tilde{H}_{2k} & \tilde{H}_{2k+1} \\ \tilde{G}_{2k} & \tilde{G}_{2k+1} \end{pmatrix} .$$

Then

$$\begin{pmatrix} \vdots \\ (\mathbf{sd})_{n-1,-1} \\ (\mathbf{sd})_{n-1,0} \\ (\mathbf{sd})_{n-1,1} \\ \vdots \end{pmatrix} = \begin{pmatrix} \cdots & \cdots & \cdots & & & & \\ \cdots & \tilde{L}_{-1} & \tilde{L}_0 & \tilde{L}_1 & \cdots & & \\ & \cdots & \tilde{L}_{-1} & \tilde{L}_0 & \tilde{L}_1 & \cdots & \\ & & \cdots & \tilde{L}_{-1} & \tilde{L}_0 & \tilde{L}_1 & \cdots \\ & & & & \cdots & \cdots & \cdots \end{pmatrix} \begin{pmatrix} \vdots \\ \mathbf{s}_{n,-1} \\ \mathbf{s}_{n,0} \\ \mathbf{s}_{n,1} \\ \vdots \end{pmatrix} , \tag{7}$$

or simply

$$(\mathbf{sd})_{n-1} = \tilde{L}\, \mathbf{s}_n \,.$$

L is an infinite banded block Toeplitz matrix with blocks of size $2r \times 2r$.

The reconstruction step can be similarly written as

$$\mathbf{s}_n = L^* (\mathbf{sd})_{n-1} \,.$$

The perfect reconstruction condition is expressed as

$$L^* \tilde{L} = I \,.$$

Modulation Formulation

This is a way of thinking about the algorithm and verifying the perfect reconstruction conditions. It is not a way to actually implement it.

We associate with the signal sequence $\mathbf{s}_n = \{\mathbf{s}_{n,k}\}$ its symbol

$$\mathbf{s}_n(\xi) = \sum_k \mathbf{s}_{n,k} \mathrm{e}^{-ik\xi} \,.$$

We can write the algorithm in matrix form as

$$\begin{pmatrix} s_{n-1}(2\xi) \\ d_{n-1}(2\xi) \end{pmatrix} = \tilde{M}(\xi) \cdot \frac{1}{\sqrt{2}} \begin{pmatrix} s_n(\xi) \\ s_n(\xi+\pi) \end{pmatrix},$$

$$\frac{1}{\sqrt{2}} \begin{pmatrix} s_n(\xi) \\ s_n(\xi+\pi) \end{pmatrix} = M(\xi)^* \begin{pmatrix} s_{n-1}(2\xi) \\ d_{n-1}(2\xi) \end{pmatrix} .$$

The matrix

$$M(\xi) = \begin{pmatrix} H(\xi) & H(\xi+\pi) \\ G(\xi) & G(\xi+\pi) \end{pmatrix}$$

is called the *modulation matrix*. The perfect reconstruc-

tion condition becomes

$$M(\xi)^* \tilde{M}(\xi) = I \,.$$

The modulation matrix of an orthogonal multiwavelet is *paraunitary*, that is,

$$M(\xi)^* M(\xi) = I \,.$$

Polyphase Formulation

Definition 7 The *phases* of the signal $\mathbf{s}_n = \{\mathbf{s}_{n,k}\}$ are defined by

$$\mathbf{s}^{(0)}_{n,k} = \mathbf{s}_{n,2k} \,, \quad \mathbf{s}^{(1)}_{n,k} = \mathbf{s}_{n,2k+1} \,.$$

The corresponding polyphase symbols are given by

$$\mathbf{s}^{(0)}_n(z) = \sum_k \mathbf{s}_{n,2k} z^k \,, \quad \mathbf{s}^{(1)}_n(z) = \sum_k \mathbf{s}_{n,2k+1} z^k \,.$$

The phases and polyphase symbols of the coefficient sequences H_k, G_k are defined similarly. Note that the symbols, defined in Eq. (2), have a factor of $1/\sqrt{2}$ in front, the polyphase symbols do not.

In matrix notation, the polyphase DMWT algorithm can be written as

$$\begin{pmatrix} \mathbf{s}_{n-1}(z) \\ \mathbf{d}_{n-1}(z) \end{pmatrix} = \tilde{P}(z) \begin{pmatrix} \mathbf{s}^{(0)}_n(z) \\ \mathbf{s}^{(1)}_n(z) \end{pmatrix} ,$$

$$\begin{pmatrix} \mathbf{s}^{(0)}_n(z) \\ \mathbf{s}^{(1)}_n(z) \end{pmatrix} = P(z)^* \begin{pmatrix} \mathbf{s}_{n-1}(z) \\ \mathbf{d}_{n-1}(z) \end{pmatrix} ,$$

where

$$P(z) = \begin{pmatrix} H^{(0)}(z) & H^{(1)}(z) \\ G^{(0)}(z) & G^{(1)}(z) \end{pmatrix}$$

is the *polyphase matrix*. The polyphase matrix has many uses in building and modifying multiwavelets.

The perfect reconstruction condition is

$$P(z)^* \tilde{P}(z) = I \,.$$

The polyphase matrix of an orthogonal multiwavelet is paraunitary.

Pre- and Postprocessing and Balanced Multiwavelets

The DMWT algorithm requires the initial expansion coefficients $\mathbf{s}_{n,k}$. Frequently the available data consists of $\boldsymbol{\sigma}_{n,k}$, which are equally spaced samples of the signal s. Converting $\boldsymbol{\sigma}_{n,k}$ to $\mathbf{s}_{n,k}$ is called *preprocessing* or *prefiltering*. After an inverse DMWT, converting $\mathbf{s}_{n,k}$ back to function values is called *postprocessing* or *postfiltering*. Postprocessing has to be the inverse of preprocessing to achieve perfect reconstruction.

For real-valued scalar wavelets,

$$2^{-n/2}\sigma_{n,k} = 2^{-n/2} s(2^{-n}k) \approx s_{n,k} = \langle s, \tilde{\boldsymbol{\phi}}_{n,k} \rangle \,,$$

and we can usually ignore the distinction. This is not true in general for multiwavelets, with some exceptions discussed below. Preprocessing and postprocessing steps are necessary.

For simplicity we assume in the remainder of this section that everything takes place at level $n = 0$, and drop the subscript 0.

For a given signal $s(x)$,

$$\mathbf{s}^*_k = \langle s(x), \tilde{\boldsymbol{\phi}}(x-k) \rangle = \int s(x) \tilde{\boldsymbol{\phi}}(x-k)^* \mathrm{d}x \,,$$

$$\boldsymbol{\sigma}^*_k = \frac{1}{\sqrt{r}} \left(s(k), s\left(k + \frac{1}{r}\right), \ldots, s\left(k + \frac{r-1}{r}\right) \right) .$$

The factor $1/\sqrt{r}$ in $\boldsymbol{\sigma}_k$ insures that for $s(x) = 1$ in the orthogonal case, $\|\mathbf{s}_k\| = \|\boldsymbol{\sigma}_k\| = 1$.

Example For the DGHM multiwavelet,

$$\boldsymbol{\mu}_0 = \frac{1}{\sqrt{3}} \begin{pmatrix} \sqrt{2} \\ 1 \end{pmatrix} .$$

This means that the function $s(x) = 1$ is represented by coefficients of the form $\mathbf{s}^*_k = 3^{-1/2}(\sqrt{2}, 1)$, while $\boldsymbol{\sigma}^*_k = 2^{-1/2}(1, 1)$.

If we use $\boldsymbol{\sigma}_k$ as the input for the DMWT, it will not be preserved during decomposition, nor will the $\mathbf{d}$ coefficients be zero. A preprocessing step needs to map $(1, 1)$ into a multiple of $(\sqrt{2}, 1)$.

A *quasi-interpolating prefilter* of order p produces correct $\boldsymbol{\sigma}_k$ for all polynomials up to degree $p - 1$. An *approximation order preserving prefilter* of order p produces $\boldsymbol{\sigma}_k$ which correspond to a polynomial of the correct degree with correct leading term, but possibly different lower-order terms. This is sufficient to achieve good results in practice.

Some common ways of constructing prefilters are listed below.

Interpolating Prefilters

Try to determine $\mathbf{s}_k$ so that the multiscaling function series matches the function values at the points $x_{k,j} = k + j/r$:

$$\sum_n \mathbf{s}^*_j \boldsymbol{\phi}(2x_{kj} - n) = s(x_{kj}) \,.$$

This may or may not be possible for a given multiwavelet. This approach preserves the approximation order but not orthogonality. See [64].

Quadrature-Based Prefilters

Approximate the integral defining $\mathbf{s}_k$ by a quadrature rule. See [31].

Quadrature-based prefilters can always be found. They preserve approximation order as long as the accuracy of the quadrature rule is at least as high as the approximation order. They do not usually preserve orthogonality.

Hardin–Roach Prefilters

These prefilters are designed to preserve orthogonality and approximation order. Preprocessing is assumed to be linear filtering

$$\mathbf{s}_k = \sum_j Q_{k-j}\boldsymbol{\sigma}_j \quad \Leftrightarrow \quad \mathbf{s}(\xi) = Q(\xi)\boldsymbol{\sigma}(\xi)\,.$$

This is an orthogonal transform if $Q(\xi)$ is paraunitary.

It is proved in [28] that such prefilters exist for arbitrarily high approximation orders. Approximation order preserving prefilters can be shorter than quasi-interpolating prefilters.

Details are described in [6,28]. Only approximation order 2 has been worked out so far.

Other Prefilters

Other approaches to prefiltering can be found in [45,56,62,65].

Balanced Multiwavelets

Balanced multiwavelets are specifically constructed to not require preprocessing. A multiwavelet is *balanced of order p* if the coefficient sequence $\mathbf{s}_k$ of any polynomial up to order $p-1$ is a polynomial sequence of the same order.

The following two characterizations of balanced multiwavelets are given in [37].

Theorem 1 *A multiwavelet is balanced of order p if and only if one of the following equivalent conditions is satisfied:*

(a) There exist constants $r_0 = 1, r_1, \ldots, r_{p-1}$ so that the symbol satisfies the sum rules of order p with approximation vectors of the form

$$\mathbf{y}_k^* = \left(\rho_k\left(\frac{0}{r}\right), \rho_k\left(\frac{1}{r}\right), \ldots, \rho_k\left(\frac{r-1}{r}\right)\right)\,,$$

where

$$\rho_k(x) = \sum_{j=0}^{k}\binom{k}{j} r_{k-j}x^j\,.$$

(b) The symbol factors as

$$H(\xi) = \frac{1}{2^p}C(2\xi)^p H_0(\xi)C(\xi)^{-p} \tag{8}$$

with

$$C(\xi) = I - \mathbf{e}_0\mathbf{e}_0^*, \quad \mathbf{e}_0^* = \frac{1}{\sqrt{r}}(1, 1, \ldots, 1)\,.$$

Part (a) are the standard sum rules (Eq. (6)) with approximation vectors of a special form. Part (b) is the TST factorization of Theorem 3 with all factors C_k equal, and of a special form.

Other conditions are derived in [50].

Any multiwavelet with approximation order 1 can be balanced of order 1, by replacing $\boldsymbol{\phi}$ with $Q\boldsymbol{\phi}(x)$ for a constant matrix Q which satisfies $\mathbf{y}_0^* Q = \mathbf{e}_0^*$. In the orthogonal case, this is the Hardin–Roach prefilter for approximation order 1.

Balancing of higher order is harder to enforce. In [10] it is shown how balancing conditions can be imposed via lifting steps. (Lifting is explained in Sect. "Lifting".)

Other examples of balanced multiwavelets are given in [10,35,37,38,50,51,52,66,67,69].

Other Multiwavelets Which Do Not Require Preprocessing

Examples are *totally interpolating* (biorthogonal) multiwavelets (see [68]), and *full rank multiwavelets* from [11].

Boundary Handling

The DMWT operates on infinite sequences of coefficients. In real life we can only work on finite sequences. We need some procedures for handling the boundary.

The finite length DMWT should be linear, so it should be of the form

$$(\mathbf{sd})_{n-1} = \tilde{L}_n\mathbf{s}_n$$

for some matrix $\tilde{L}_n$, in analogy with Eq. (7). In order to preserve the usual definition of the DMWT as much as possible, we postulate the form

$$\tilde{L}_n = \begin{pmatrix} \tilde{L}_\mathrm{b} & & \\ & \tilde{L}_{\mathrm{i},n} & \\ & & \tilde{L}_\mathrm{e} \end{pmatrix}\,,$$

where

- The interior part $\tilde{L}_{\mathrm{i},n}$ is a segment of the infinite block Toeplitz matrix $\tilde{L}$, and each row contains a complete

set of coefficients. This part will make up most of the matrix. $\tilde{L}_{\mathrm{i},n}$ approximately doubles in size when n is increased by 1.
- The matrices $\tilde{L}_{\mathrm{b}}$ at the beginning and $\tilde{L}_{\mathrm{e}}$ at the end are fairly small and remain unchanged at all levels. These matrices will handle the boundaries.
- The entire matrix $\tilde{L}_n$ has the same block structure as $\tilde{L}$ (each block row shifted by one compared to its neighbors).
- $\tilde{L}_n$ is invertible, and its inverse matrix L_n^* has an analogous structure.

In the orthogonal case, we also would like to preserve $L_n^{-1} = L_n^*$.

There are a number of ways to find suitable boundary coefficients. An excellent overview of the scalar case can be found in Sect. 8.5 in [54]. For multiwavelets, all the standard approaches that work for scalar wavelets appear to carry over in practice, but very little has actually been proved.

Data Extension Approach

This is easy to implement. We artificially extend the signal across the boundaries so that each extended coefficient is a linear combination of known coefficients.

For example, suppose the left border is at 0. We are given $\mathbf{s}_{0,k}$ for $k \geq 0$, but not for $k < 0$, and we want to compute

$$\mathbf{s}_{-1,0} = \tilde{H}_{-1}\mathbf{s}_{0,-1} + \tilde{H}_0\mathbf{s}_{0,0} + \tilde{H}_1\mathbf{s}_{0,1} + \cdots .$$

If our extension method is

$$\mathbf{s}_{0,-1} = A\mathbf{s}_{0,0} + B\mathbf{s}_{0,1} ,$$

then

$$\mathbf{s}_{-1,0} = (\tilde{H}_0 + AH_{-1})\mathbf{s}_{0,0} + (\tilde{H}_1 + BH_{-1})\mathbf{s}_{0,1} + \cdots .$$

The H- and G-coefficients that "stick out over the side" when $\tilde{L}$ is truncated to $\tilde{L}_n$ are wrapped back inside.

This gives us $\tilde{L}_n$, and its inverse will be L_n^*. There is no guarantee that $\tilde{L}_n$ will not be singular, or that $L_n = \tilde{L}_n^{-*}$ will have the correct form, but it often works in practice.

Special cases include the following:

- **Periodic Extension:** This is easy to do, and always works. It preserves orthogonality and approximation order 1. Periodic extension is not usually a good idea unless the data are truly periodic. The jump at the boundary leads to spurious large $\mathbf{d}$-coefficients.
- **Symmetric Extension:** In the scalar case there are four possibilities at each end: even or odd extension, and whole-sample or half-sample, depending on whether the boundary coefficient is repeated or not. This is analyzed in detail in [13], where it is shown that if you match the type of data extension correctly to the type of symmetry of the scaling function, the DWT will preserve the symmetry across levels.
  For multiwavelets there are many more possibilities of symmetry and symmetric extension. Some cases are treated in [63]. An ad hoc symmetric extension for the DGHM multiwavelet is used in [56].
- **Zero, Constant or Linear Extension:** These methods appear to work for multiwavelets. They preserve the corresponding approximation order (0, 1, or 2), but not symmetry.

Matrix Completion Approach

This is a linear algebra approach based on finding suitable end blocks which guarantee $\tilde{L}_n L_n^* = I$. It is described for scalar wavelets in [40,54]. The extension of these results to multiwavelets is a subject of ongoing research by the author of this article. Preliminary results indicate that this approach works in practice, but there are some new phenomena.

For example, for a multiscaling function of multiplicity 2 and support length 3 one would expect to need one boundary multiscaling function at each end, each composed of two individual scaling functions. It turns out that sometimes one has to use three individual scaling functions at one end, and one at the other.

Boundary Function Approach

This approach is the most time-consuming, but it can preserve both orthogonality and approximation order. The idea is to introduce special boundary functions at each end of the interval, and work out the resulting decomposition and reconstruction algorithms.

For scalar wavelets, this approach was pioneered in [5,17]. For multiwavelets, this has only been worked out in detail for the case of the cubic Hermite multiscaling function with one particular dual. The details are given in [18], and they are quite lengthy.

Other Approaches

There is one variation on the boundary function approach that is easy to do and has no counterpart for scalar wavelets. If we have multiwavelets with support on $[-1, 1]$, there is exactly one boundary-crossing multiscaling func-

tion at each end, and it is already orthogonal to everything inside. We can simply restrict the boundary function vector to the inside of the interval, and orthonormalize the components among themselves. Examples for this approach are given in [26,27].

This would also work for scalar wavelets, of course, except there are no wavelet pairs with support in $[-1, 1]$ except the Haar wavelet. For multiwavelets, it is possible to achieve arbitrarily high approximation order, plus symmetry, on $[-1, 1]$ by taking the multiplicity high enough.

Applications

Scalar wavelets and multiwavelets both deal with one-dimensional signals. They have the same applications: signal compression, signal denoising, fast operator evaluation in numerical analysis, Galerkin methods for differential and integral equations.

By far the largest part of wavelet applications in the literature deal with scalar wavelets. There are relatively few articles that report on implementations and performance for multiwavelets.

Some studies have compared the performance of scalar wavelets and multiwavelets in image denoising and compression, including [14,22,23,30,53,56,58,63]. The use of multiwavelets for video compression is reported in the thesis of Tham [59].

It appears that multiwavelets can do as well or better than scalar wavelets, but careful attention must be paid to preconditioning and handling of boundaries. The authors of [56] report that multiwavelet filters with short support produce fewer artifacts in the reconstruction of compressed images.

The main advantage of multiwavelets over scalar wavelets in numerical analysis lies in their short support, which makes boundaries much easier to handle.

For integral equations, multiwavelets with support $[0, 1]$ can be used. At least some of the basis functions necessarily must be discontinuous, but for integral equations that is not a problem.

Indeed, the first appearance of such multiwavelets was in the thesis and papers of Alpert (see [1,2,3,4]), before the concept of multiwavelets was invented. Multiwavelet methods for integral equations are also discussed in [16, 41,43,44,57,61].

For differential equations, multiwavelets with support $[-1, 1]$ can be used. Regularity and approximation order can be raised to arbitrary levels by taking the multiplicity high enough.

There is only one multiscaling function that crosses each boundary. It is already orthogonal to all the interior functions, so constructing the boundary multiscaling function is an easy matter: orthonormalize the truncated boundary-crossing multiscaling function. This automatically preserves approximation order. Finding the boundary multiwavelet function still takes a little effort.

If symmetric/antisymmetric multiwavelets are used, it is even possible to use only the antisymmetric components of the boundary function vector for problems with zero boundary conditions. Examples of suitable multiwavelets can be found in [18,27].

Other papers about adapting multiwavelets to the solution of differential equation include [3,8,9,39,42].

For an overview of the use of wavelets (including multiwavelets) in numerical analysis, see the collection [12].

Polyphase Factorization

Definition 8 An *orthogonal projection factor* of rank k, $1 \le k \le r-1$, is a linear paraunitary matrix of the form

$$F(z) = (I - UU^*) + UU^* z\,,$$

where U has k orthonormal columns.

Theorem 2 *Assume that $P(z)$ is the polyphase matrix of an orthogonal multiwavelet $P(z) = P_0 + P_1 z + \cdots + P_n z^n$, with $P_0 \neq 0$, $P_n \neq 0$. Then $P(z)$ can be factored in the form*

$$P(z) = QF_1(z)\ldots F_n(z)\,,$$

where Q is a constant unitary matrix, and each $F_j(z)$ is a projection factor. The number of factors equals the polynomial degree of $P(z)$.

The proof is constructive and produces the factors one by one. The factors are not necessarily unique.

The unitary matrix Q could also be put on the right, but placing it on the left has one distinct advantage: the factorization can be computed even if only the top rows of P are know. The *completion problem* (finding the multiwavelet function if the multiscaling function is known) can then be reduced to the problem of completing a constant orthogonal matrix, which is easy.

Other completion methods can be found in [24,36,55].

The factorization theorem for biorthogonal multiwavelets is considerably more complex. In addition to projection factors it also requires so-called *atoms*. Details can be found in [32,34]. In [49], atoms are called *pseudo-identity matrix pairs*.

The polyphase factorization can be used to construct orthogonal or biorthogonal multiwavelets from scratch.

Lifting

The lifting process is for constructing or modifying scalar wavelets or multiwavelets. It does not preserve orthogonality.

Assume we have a pair of matrix polynomials satisfying $P(z)^*\tilde{P}(z) = I$. We can interpret these as the polyphase matrices of a biorthogonal multiwavelet pair. The DMWT algorithm works for purely algebraic reasons, whether or not the coefficients are actually associated with multiscaling and multiwavelet functions.

Given any other pair satisfying $L(z)^*\tilde{L}(z) = I$, we obtain a new pair

$$\left[L(z)P(z)\right]^* \left[\tilde{L}(z)\tilde{P}(z)\right] = I\,.$$

The idea behind lifting is to use $L(z)$, $\tilde{L}(z)$ of a special form.

A *lifting step* is based on

$$L(z) = \begin{pmatrix} I & A(z) \\ 0 & I \end{pmatrix}, \qquad \tilde{L}(z) = \begin{pmatrix} I & 0 \\ -A(z)^* & I \end{pmatrix}$$

for arbitrary $A(z)$. The effect on the symbols is

$$\begin{aligned} H_{\text{new}}(z) &= H(z) + A(z^2)G(z) \\ G_{\text{new}}(z) &= G(z)\,, \\ \tilde{H}_{\text{new}}(z) &= \tilde{H}(z)\,, \\ \tilde{G}_{\text{new}}(z) &= \tilde{G}(z) - A(z^2)^*\tilde{H}(z)\,. \end{aligned}$$

This means that the multiscaling function $\boldsymbol{\phi}$ changes, but its dual $\tilde{\boldsymbol{\phi}}$ does not.

A *dual lifting step* is based on

$$L(z) = \begin{pmatrix} I & 0 \\ B(z) & I \end{pmatrix}, \qquad \tilde{L}(z) = \begin{pmatrix} I & -B(z)^* \\ 0 & I \end{pmatrix}.$$

It has a similar effect, with the roles of $\boldsymbol{\phi}$, $\tilde{\boldsymbol{\phi}}$ reversed. Note that some authors use a reverse definition of lifting and dual lifting steps.

For scalar wavelets, it is shown in [19] that every polyphase matrix can be factored entirely into lifting steps. An implementation of the DWT based on this factorization is faster than a direct implementation.

There is a corresponding theorem for multiwavelets (see [20]), but it requires some extra factors (unit triangular and diagonal matrices). It is not clear that this factorization is as useful as in the scalar case.

In a lifting step, conditions can be put on $A(z)$ that will create a new $\boldsymbol{\phi}$ with higher approximation order, while keeping $\tilde{\boldsymbol{\phi}}$ and its approximation order the same. Likewise, a dual lifting step can increase the approximation order of $\tilde{\boldsymbol{\phi}}$. For details see [33].

A lifting procedure which imposes symmetry conditions is described in [60]. A lifting procedure which imposes balancing conditions is described in [10]. Other papers on multiwavelet lifting include [7,8,25].

Two-Scale Similarity Transform (TST)

The two-scale similarity transform (TST) is a new, nonobvious construction for multiwavelets that has no counterpart for scalar wavelets (or rather, the concept is so trivial there that it did not need a name). Like lifting, it does not preserve orthogonality.

One application is a characterization of approximation order which is useful for both theoretical and practical purposes. It leads to the counterpart of the statement "the symbol $H(\xi)$ satisfies the sum rules of order p if and only if it contains a factor of $(1 + \mathrm{e}^{-i\xi})^p$." The TST factorization can also be used to characterize balanced multiwavelets and symmetric multiwavelets.

Assume that $\boldsymbol{\phi}$ is a refinable function vector, and let

$$\boldsymbol{\phi}_{\text{new}}(x) = \sum_k C_k \boldsymbol{\phi}(x-k)$$

for some coefficient matrices C_k. Then

$$\hat{\boldsymbol{\phi}}_{\text{new}}(\xi) = C(\xi)\hat{\boldsymbol{\phi}}(\xi)\,, \quad C(\xi) = \sum_k C_k \mathrm{e}^{-ik\xi}\,.$$

If $C(\xi)$ is nonsingular for all ξ, then

$$\begin{aligned} \hat{\boldsymbol{\phi}}_{\text{new}}(2\xi) &= C(2\xi)\hat{\boldsymbol{\phi}}(2\xi) = C(2\xi)H(\xi)\hat{\boldsymbol{\phi}}(\xi) \\ &= C(2\xi)H(\xi)C(\xi)^{-1}\hat{\boldsymbol{\phi}}_{\text{new}}(\xi)\,. \end{aligned}$$

This means that $\boldsymbol{\phi}_{\text{new}}$ is again refinable with symbol $H_{\text{new}}(\xi) = C(2\xi)H(\xi)C(\xi)^{-1}$. This is a basis change which leaves all the spaces V_n in the MRA invariant.

We can also allow singular $C(\xi)$ of a special type, and this is actually the more interesting application of this idea.

Definition 9 A *TST matrix* is a 2π-periodic, continuously differentiable matrix-valued function $C(\xi)$ which satisfies

- $C(\xi)$ is invertible for $\xi \neq 2\pi k$, $k \in \mathbb{Z}$.
- $C(0)$ has a simple eigenvalue 0 with left and right eigenvectors $\mathbf{l}$ and $\mathbf{r}$.
- This eigenvalue satisfies $\lambda'(0) \neq 0$.

The last statement requires a brief explanation: as ξ varies, the eigenvalues of $C(\xi)$ vary continuously with ξ. Simple eigenvalues vary in a differentiable manner. $\lambda(\xi)$ is the eigenvalue for which $\lambda(0) = 0$. In some neighborhood of

the origin, $\lambda(\xi)$ is uniquely defined and differentiable. This derivative must be nonzero at 0.

If $C(\xi)$ is a TST matrix, then

$$C_0(\xi) = (1 - \mathrm{e}^{-i\xi})C(\xi)^{-1}$$

is well-defined for all ξ.

The standard example is

$$\begin{aligned} C(\xi) &= I - \mathbf{r}\mathbf{l}^*\mathrm{e}^{-i\xi} = (I - \mathbf{r}\mathbf{l}^*) + \mathbf{r}\mathbf{l}^*(1 - \mathrm{e}^{-i\xi}) \,, \\ C(\xi)^{-1} &= (I - \mathbf{r}\mathbf{l}^*) + \frac{\mathbf{r}\mathbf{l}^*}{1 - \mathrm{e}^{-i\xi}} \,, \quad \xi \neq 2\pi k \,, \\ C_0(\xi) &= \mathbf{r}\mathbf{l}^* + (I - \mathbf{r}\mathbf{l}^*)(1 - \mathrm{e}^{-i\xi}) \,, \end{aligned}$$

where $\mathbf{l}$ and $\mathbf{r}$ are normalized to $\mathbf{l}^*\mathbf{r} = 1$. Here $\mathbf{r}(\xi) = \mathbf{r}$, $\lambda(\xi) = 1 - \mathrm{e}^{-i\xi}$, so $\lambda'(0) = i \neq 0$.

Main Definition

H_{new} is a *TST* of H if

$$H_{\text{new}}(\xi) = \frac{1}{2}C(2\xi)H(\xi)C(\xi)^{-1}$$

for a TST matrix $C(\xi)$ for which $C(0)$ and $H(0)$ share a common right eigenvector $\mathbf{r}$.

H_{new} is an *inverse TST* of H if

$$H_{\text{new}}(\xi) = 2C(2\xi)^{-1}H(\xi)C(\xi)$$

for a TST matrix $C(\xi)$ for which $C(0)$, $H(0)$, $H(\pi)$ share a common left eigenvector $\mathbf{l}$.

If C is a TST matrix for H, it automatically satisfies the conditions to be an inverse TST matrix for $\tilde{H}_{\text{new}}$. The eigenvector conditions ensure that $\tilde{H}_{\text{new}}$ has an approximation order one higher than before, and $\tilde{H}_{\text{new}}$ has an approximation order one lower. This is a way of moving approximation orders from one side to the other. For scalar wavelets, it corresponds to moving a factor of $(1 + \mathrm{e}^{-i\xi})/2$. The TST can be extended to cover the multiwavelet functions as well (see [34]).

Repeated application of TSTs leads to the following result.

Theorem 3 *If $H(\xi)$ has approximation order $p \geq 1$, it can be factored as*

$$\begin{aligned} H(\xi) &= H_p(\xi) \\ &= \frac{1}{2}C_p(2\xi)H_{p-1}(\xi)C_p(\xi)^{-1} = \ldots \\ &= 2^{-p}C_p(2\xi)\ldots C_1(2\xi)H_0(\xi)C_1(\xi)^{-1}\ldots C_p(\xi)^{-1} \,, \end{aligned}$$

where each $C_k(\xi)$ is a TST matrix.

As mentioned in Sect. "Pre- and Postprocessing and Balanced Multiwavelets", additional conditions on the TST matrices C_j characterize balanced multiwavelets.

TSTs were defined in [55]. A special case which required $H(0)$ to have eigenvectors of a particular structure was developed independently in [46]. The two approaches were reconciled in the joint paper [47].

Examples and Software

For the reader who wants to experiment with multiwavelets, this section lists some of the more commonly used coefficients. There is also a brief list of relevant software at the end.

In all listings, p = approximation order, s = Sobolev exponent, and α = Hölder exponent. If α is not listed, $s - 1/2$ is a lower bound for α. A common factor for all coefficients in its column is listed separately, for easier readability. All examples have multiplicity $r = 2$.

DGHM (Donovan–Geronimo–Hardin–Massopust [21]) (orthogonal)

Support $[0, 2]$, $p = 2$, $\alpha = 1$, $s = 1.5$. ϕ_1 is symmetric about $x = 1/2$, ϕ_2 is symmetric about $x = 1$. See Fig. 1 in the introduction for graphs.

	H_k	G_k
$k = 0$	$\begin{pmatrix} 12 & 16\sqrt{2} \\ -\sqrt{2} & -6 \end{pmatrix}$	$\begin{pmatrix} -\sqrt{2} & -6 \\ 2 & 6\sqrt{2} \end{pmatrix}$
1	$\begin{pmatrix} 12 & 0 \\ 9\sqrt{2} & 20 \end{pmatrix}$	$\begin{pmatrix} 9\sqrt{2} & -20 \\ -18 & 0 \end{pmatrix}$
2	$\begin{pmatrix} 0 & 0 \\ 9\sqrt{2} & -6 \end{pmatrix}$	$\begin{pmatrix} 9\sqrt{2} & -6 \\ 18 & -6\sqrt{2} \end{pmatrix}$
3	$\begin{pmatrix} 0 & 0 \\ -\sqrt{2} & 0 \end{pmatrix}$	$\begin{pmatrix} -\sqrt{2} & 0 \\ -2 & 0 \end{pmatrix}$
Factor	$1/(20\sqrt{2})$	$1/(20\sqrt{2})$

	H_k	G_k
$k = 0$	$\begin{pmatrix} 0 & 2+\sqrt{7} \\ 0 & 2-\sqrt{7} \end{pmatrix}$	$\begin{pmatrix} 0 & -2 \\ 0 & 1 \end{pmatrix}$
1	$\begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$	$\begin{pmatrix} 2 & 2 \\ -\sqrt{7} & \sqrt{7} \end{pmatrix}$
2	$\begin{pmatrix} 2-\sqrt{7} & 0 \\ 2+\sqrt{7} & 0 \end{pmatrix}$	$\begin{pmatrix} -2 & 0 \\ -1 & 0 \end{pmatrix}$
Factor	$1/(4\sqrt{2})$	$1/4$

BAT O1 (Lebrun–Vetterli [37]) (orthogonal, balanced)

Support $[0, 2]$. ϕ_2 is reflection of ϕ_1 about $x = 1$ and vice versa; wavelet functions are symmetric/antisymmetric about $x = 1$. $p = 2$, balanced of order 1, $s = 0.6406$.

The shape of ψ_1 is most likely responsible for the name BAT wavelet. There are also BAT O2 and BAT O3, which are balanced of order 2 and 3. See Fig. 4 for graphs.

Multiwavelets, Table 1
Recursion coefficients for Hermite cubic multiwavelet

	H_k	G_k	$\tilde{H}_k$	$\tilde{G}_k$
$k=-2$			$\begin{pmatrix} -2190 & -1540 \\ 13914 & 9687 \end{pmatrix}$	
-1	$\begin{pmatrix} 4 & 6 \\ -1 & -1 \end{pmatrix}$	$\begin{pmatrix} 5427 & 567 \\ -1900 & -120 \end{pmatrix}$	$\begin{pmatrix} 9720 & 3560 \\ -60588 & -21840 \end{pmatrix}$	
0	$\begin{pmatrix} 8 & 0 \\ 0 & 4 \end{pmatrix}$	$\begin{pmatrix} -19440 & -60588 \\ 7120 & 21840 \end{pmatrix}$	$\begin{pmatrix} 23820 & 0 \\ 0 & 36546 \end{pmatrix}$	$\begin{pmatrix} -2 & -1 \\ 3 & 1 \end{pmatrix}$
1	$\begin{pmatrix} 4 & -6 \\ 1 & -1 \end{pmatrix}$	$\begin{pmatrix} 28026 & 0 \\ 0 & 56160 \end{pmatrix}$	$\begin{pmatrix} 9720 & -3560 \\ 60588 & -21840 \end{pmatrix}$	$\begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}$
2		$\begin{pmatrix} -19440 & 60588 \\ -7120 & 21840 \end{pmatrix}$	$\begin{pmatrix} -2190 & 1540 \\ -13914 & 9687 \end{pmatrix}$	$\begin{pmatrix} -2 & 1 \\ -3 & 1 \end{pmatrix}$
3		$\begin{pmatrix} 5427 & -567 \\ 1900 & -120 \end{pmatrix}$		
Factor	$1/(8\sqrt{2})$	$1/(19440\sqrt{2})$	$1/(19440\sqrt{2})$	$1/(8\sqrt{2})$

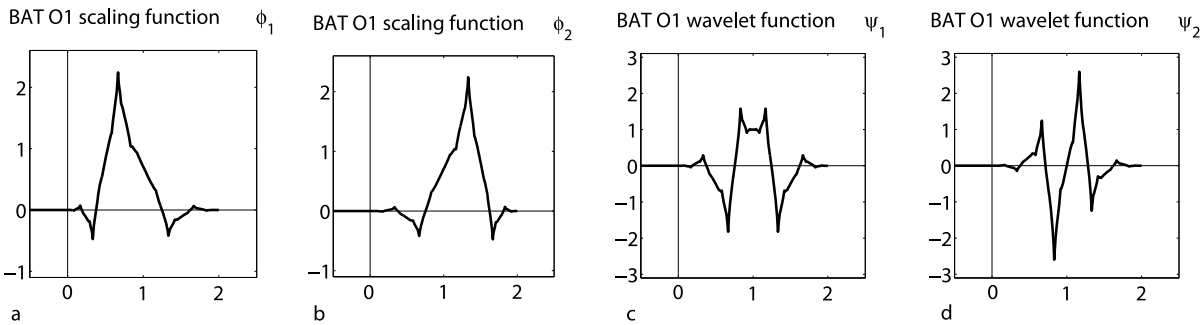

Multiwavelets, Figure 4
BAT O1 multiwavelet

HC (Hermite Cubics) (biorthogonal)

Hermite cubics are C^1 piecewise cubic polynomials on $[-1,1]$ which satisfy $\phi_1(0)=1$, $\phi_1'(0)=0$ and $\phi_2(0)=0$, $\phi_2'(0)=1$. They are not orthogonal, so there are many biorthogonal completions. Many authors also use a different scaling for ϕ_2, so not even the H_k will match what is listed here.

The completion listed here is the smoothest symmetric completion with support length 4 (see [29]).

Support of $\boldsymbol{\phi}$ is $[-1,1]$, support of $\tilde{\boldsymbol{\phi}}$ is $[-2,2]$; $\boldsymbol{\psi}$ and $\tilde{\boldsymbol{\psi}}$ have support $[-1,2]$; all functions are symmetric/antisymmetric about the center of their support. $p=4$, $\tilde{p}=2$, $\alpha=2$, $s=2.5$, $\tilde{s}=0.8279$. See Fig. 5 for graphs. See Table1 for recursion coefficients.

There are many toolboxes available for scalar wavelets. Most of them are for Matlab or are stand-alone programs, but there are tools for Mathematica, MathCAD and other systems as well. An extensive list can be found at http://www.amara.com/current/wavesoft.html. MathWorks (http://www.mathworks.com) is the maker of Matlab, so theirs is the "official" Matlab wavelet toolbox. These programs cannot usually handle multiwavelets.

There are only two software packages for multiwavelets available, both of them for Matlab.

MWMP (the Multiwavelet Matlab Package) was written by Vasily Strela; it can be found at http://www.mcs.drexel.edu/~vstrela/MWMP. A set of multiwavelet Matlab routines from the author of this entry is available through http://www.math.iastate.edu/keinert or the CRC Press download page at http://www.crcpress.com/e_products/downloads/default.asp.

Future Directions

The basic properties of multiwavelets are quite well understood by now. Many types of multiwavelets have been constructed, and new construction methods continue to be found. Nevertheless, multiwavelets have not been applied

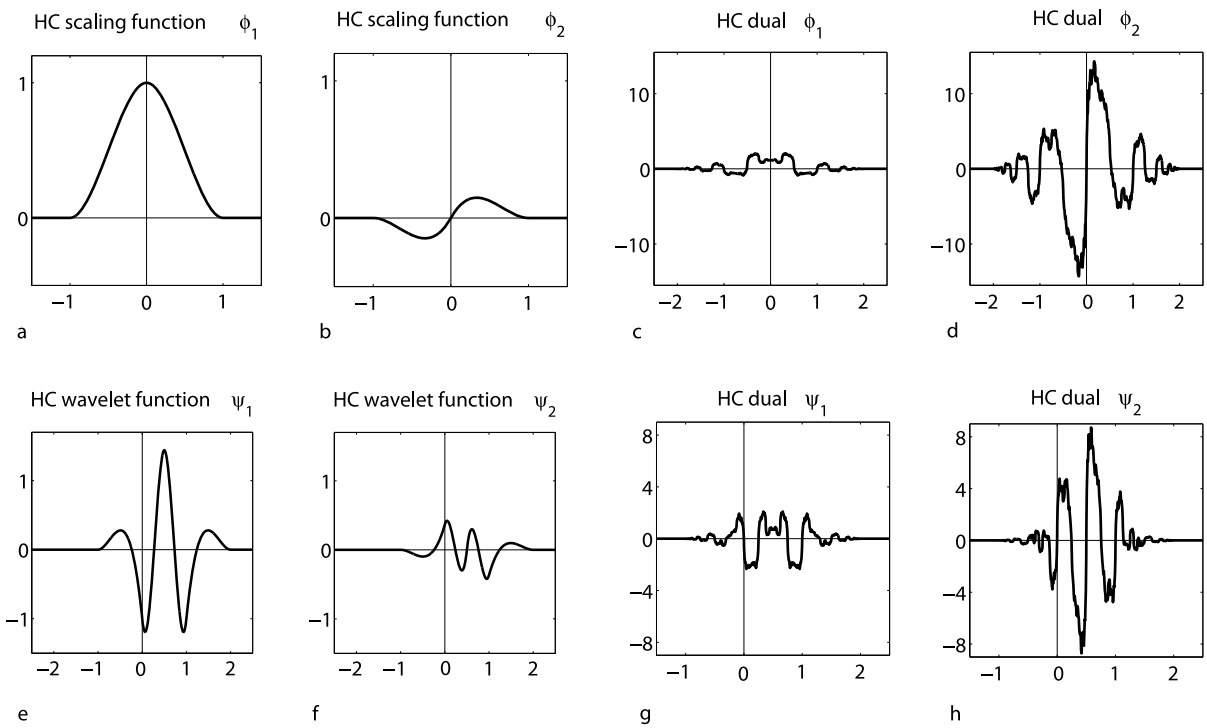

Multiwavelets, Figure 5
Hermite Cubic multiwavelet

as much as initially expected. The theory is more complicated than for scalar wavelets, and the need for pre- and postprocessing is a deterrent in applications.

For future theoretical development, the biggest need is for improved methods of preprocessing, or even better, for multiwavelets that do not require preprocessing. The study of balanced multiwavelets is a very active field at the moment.

To a lesser degree, a comprehensive theory of boundary handling is also needed. Most multiwavelet users sidestep the issue by padding the data.

In applications, the experience has been that multiwavelets can can do as well as scalar wavelets in the usual applications, and possibly better. However, they are not spectacularly better, and the extra effort is not usually warranted. The exception to this are applications that utilize one of the particular strengths of multiwavelets.

One of these strengths is that multiwavelets can be both orthogonal and symmetric. This is useful in applications where symmetry is important.

The other, more important strength is high approximation order/high smoothness coupled with short support. This is important in using multiwavelets as basis functions in the solution of differential and integral equations. Short support reduces the overlap between segments, resulting in sparser matrices, and also resolves much of the boundary problem.

For integral equations, multiwavelets with support $[0, 1]$ can be used (which are necessarily discontinuous). Differential equations require multiwavelets with support $[-1, 1]$.

Bibliography

1. Alpert BK (1990) Sparse representation of smooth linear operators. Technical Report YALEU/DCS/RR-814. Yale University, New Haven
2. Alpert BK, Beylkin G, Coifman R, Rokhlin V (1990) Wavelets for the fast solution of second-kind integral equations. Technical Report YALEU/DCS/RR-837. Yale University, New Haven
3. Alpert B, Beylkin G, Gines D, Vozovoi L (2002) Adaptive solution of partial differential equations in multiwavelet bases. J Comput Phys 182(1):149–190
4. Alpert BK (1993) A class of bases in L^2 for the sparse representation of integral operators. SIAM J Math Anal 24(1):246–262
5. Andersson L, Hall N, Jawerth B, Peters G (1994) Wavelets on closed subsets of the real line. In: Recent advances in wavelet analysis. Academic Press, Boston, pp 1–61
6. Attakitmongcol K, Hardin DP, Wilkes DM (2001) Multiwavelet prefilters II: optimal orthogonal prefilters. IEEE Trans Image Proc 10(10):1476–1487

7. Averbuch AZ, Zheludev VA (2002) Lifting scheme for biorthogonal multiwavelets originated from Hermite splines. IEEE Trans Signal Process 50(3):487–500
8. Averbuch A, Israeli M, Vozovoi L (1999) Solution of time-dependent diffusion equations with variable coefficients using multiwavelets. J Comput Phys 150(2):394–424
9. Averbuch A, Braverman E, Israeli M (2000) Parallel adaptive solution of a Poisson equation with multiwavelets. SIAM J Sci Comput 22(3):1053–1086
10. Bacchelli S, Cotronei M, Lazzaro D (2000) An algebraic construction of *k*-balanced multiwavelets via the lifting scheme. Numer Algorithms 23(4):329–356
11. Bacchelli S, Cotronei M, Sauer T (2002) Multifilters with and without prefilters. BIT 42(2):231–261
12. Bramble JH, Cohen A, Dahmen W, Canuto C (ed) (2003) Multiscale problems and methods in numerical simulations. Lecture Notes in Mathematics, vol 1825. Springer, Berlin. Lectures given at the CIME Summer School held in Martina Franca, 9–15 September, 2001
13. Brislawn C (1996) Classification of nonexpansive symmetric extension transforms for multirate filter banks. Appl Comput Harmon Anal 3(4):337–357
14. Bui TD, Chen G (1998) Translation-invariant denoising using multiwavelets. IEEE Trans Signal Process 46(12):3414–3420
15. Burrus CS, Gopinath RA, Guo H (1998) Introduction to wavelets and wavelet transforms: a primer. Prentice Hall, New York
16. Chen Z, Micchelli CA, Xu Y (1997) The Petrov–Galerkin method for second kind integral equations II: multiwavelet schemes. Adv Comput Math 7(3):199–233
17. Cohen A, Daubechies I, Vial P (1993) Wavelets on the interval and fast wavelet transforms. Appl Comput Harmon Anal 1(1):54–81
18. Dahmen W, Han B, Jia RQ, Kunoth A (2000) Biorthogonal multiwavelets on the interval: cubic Hermite splines. Constr Approx 16(2):221–259
19. Daubechies I, Sweldens W (1998) Factoring wavelet transforms into lifting steps. J Fourier Anal Appl 4(3):247–269
20. Davis GM, Strela V, and Turcajova R (2000) Multiwavelet construction via the lifting scheme. In: Wavelet analysis and multiresolution methods. Dekker, New York, pp 57–79
21. Donovan GC, Geronimo JS, Hardin DP, Massopust PR (1996) Construction of orthogonal wavelets using fractal interpolation functions. SIAM J Math Anal 27(4):1158–1192
22. Downie TR, Silverman BW (1998) The discrete multiple wavelet transform and thresholding methods. IEEE Trans Signal Process 46(9):2558–2561
23. Efromovich S (2001) Multiwavelets and signal denoising. Sankhyā Ser A 63(3):367–393
24. Goh SS, Yap VB (1998) Matrix extension and biorthogonal multiwavelet construction. Linear Algebr Appl 269:139–157
25. Goh SS, Jiang Q, Xia T (2000) Construction of biorthogonal multiwavelets using the lifting scheme. Appl Comput Harmon Anal 9(3):336–352
26. Han B, Jiang Q (2002) Multiwavelets on the interval. Appl Comput Harmon Anal 12(1):100–127
27. Hardin DP, Marasovich JA (1999) Biorthogonal multiwavelets on [−1, 1]. Appl Comput Harmon Anal 7(1):34–53
28. Hardin DP, Roach DW (1998) Multiwavelet prefilters I: orthogonal prefilters preserving approximation order $p \leq 2$. IEEE Trans Circuits Syst II. Analog Digital Signal Process 45(8): 1106–1112
29. Heil C, Strang G, Strela V (1996) Approximation by translates of refinable functions. Numer Math 73(1):75–94
30. Hsung TC, Lun DPK, Ho KC (2005) Optimizing the multiwavelet shrinkage denoising. IEEE Trans Signal Process 53(1): 240–251
31. Johnson BR (2000) Multiwavelet moments and projection prefilters. IEEE Trans Signal Process 48(11):3100–3108
32. Kautsky J, Turcajova R (1995) Discrete biorthogonal wavelet transforms as block circulant matrices. Linear Algebr Appl 223/224:393–413
33. Keinert F (2001) Raising multiwavelet approximation order through lifting. SIAM J Math Anal 32(5):1032–1049
34. Keinert F (2004) Wavelets and multiwavelets. Studies in Advanced Mathematics. Chapman & Hall/CRC, Boca Raton
35. Kessler B (2005) Balanced scaling vectors using linear combinations of existing scaling vectors. In: Approximation theory, XI. Mod Methods Math. Nashboro Press, Brentwood, pp 197–210
36. Lawton W, Lee SL, Shen Z (1996) An algorithm for matrix extension and wavelet construction. Math Comp 65(214):723–737
37. Lebrun J, Vetterli M (2001) High-order balanced multiwavelets: theory, factorization, and design. IEEE Trans Signal Process 49(9):1918–1930
38. Lian JA (2005) Armlets and balanced multiwavelets: flipping filter construction. IEEE Trans Signal Process 53(5):1754–1767
39. Lin EB, Xiao Z (2000) Multiwavelet solutions for the Dirichlet problem. In: Wavelet analysis and multiresolution methods. Lecture Notes in Pure and Appl Math, vol 212. Dekker, New York, pp 241–254
40. Madych WR (1997) Finite orthogonal transforms and multiresolution analyses on intervals. J Fourier Anal Appl 3(3):257–294
41. Maleknejad K, Yousefi M (2006) Numerical solution of the integral equation of the second kind by using wavelet bases of Hermite cubic splines. Appl Math Comput 183(1):134–141
42. Massopust PR (1998) A multiwavelet based on piecewise C^1 fractal functions and related applications to differential equations. Bol Soc Mat Mex 4(2):249–283
43. Micchelli CA, Xu Y (1994) Using the matrix refinement equation for the construction of wavelets II: smooth wavelets on [0, 1]. In: Approximation and computation. Birkhäuser, Boston, pp 435–457
44. Micchelli CA, Xu Y (1994) Using the matrix refinement equation for the construction of wavelets on invariant sets. Appl Comput Harmon Anal 1(4):391–401
45. Miller J, Li CC (1998) Adaptive multiwavelet initialization. IEEE Trans Signal Process 46(12):3282–3291
46. Plonka G (1997) Approximation order provided by refinable function vectors. Constr Approx 13(2):221–244
47. Plonka G, Strela V (1998) Construction of multiscaling functions with approximation and symmetry. SIAM J Math Anal 29(2):481–510
48. Plonka G, Strela V (1998) From wavelets to multiwavelets. In: Mathematical methods for curves and surfaces, II. Vanderbilt Univ Press, Nashville, pp 375–399
49. Resnikoff HL, Tian J, Wells RO Jr (2001) Biorthogonal wavelet space: parametrization and factorization. SIAM J Math Anal 33(1):194–215
50. Selesnick IW (1998) Multiwavelet bases with extra approximation properties. IEEE Trans Signal Process 46(11):2898–2908
51. Selesnick IW (2000) Balanced multiwavelet bases based on symmetric FIR filters. IEEE Trans Signal Process 48(1):184–191

52. Shen L, Tan HH (2001) On a family of orthonormal scalar wavelets and related balanced multiwavelets. IEEE Trans Signal Process 49(7):1447–1453
53. Shi H, Cai Y, Qiu Z (2006) On design of multiwavelet prefilters. Appl Math Comput 172(2):1175–1187
54. Strang G, Nguyen T (1996) Wavelets and filter banks. Wellesley-Cambridge Press, Wellesley
55. Strela V (1996) Multiwavelets: theory and applications. Ph D thesis, Massachusetts Institute of Technology
56. Strela V, Heller PN, Strang G, Topiwala P, Heil C (1999) The application of multiwavelet filter banks to image processing. IEEE Trans Signal Process 8:548–563
57. Tausch J (2002) Multiwavelets for geometrically complicated domains and their application to boundary element methods. In: Integral methods in science and engineering. Birkhäuser, Boston, pp 251–256
58. Tham JY, Shen L, Lee SL, Tan HH (2000) A general approach for analysis and application of discrete multiwavelet transform. IEEE Trans Signal Process 48(2):457–464
59. Tham JY (2002) Multiwavelets and scalable video compression. Ph D thesis, National University of Singapore. Available at http://www.cwaip.nus.edu.sg/thamjy
60. Turcajová R (1999) Construction of symmetric biorthogonal multiwavelets by lifting. In: Unser MA, Aldroubi A, Laine AF (eds) Wavelet applications in signal and image processing (SPIE), VII, vol 3813. SPIE, Bellingham, pp 443–454
61. von Petersdorff T, Schwab C, Schneider R (1997) Multiwavelets for second-kind integral equations. SIAM J Numer Anal 34(6):2212–2227
62. Vrhel MJ, Aldroubi A (1998) Projection based prefiltering for multiwavelet transforms. IEEE Trans Signal Process 46(11):3088
63. Xia T, Jiang Q (1999) Optimal multifilter banks: design, related symmetric extension transform, and application to image compression. IEEE Trans Signal Process 47(7):1878–1889
64. Xia XG, Geronimo JS, Hardin DP, Suter BW (1996) Design of prefilters for discrete multiwavelet transforms. IEEE Trans Signal Process 44(1):25–35
65. Xia XG (1998) A new prefilter design for discrete multiwavelet transforms. IEEE Trans Signal Process 46(6):1558–1570
66. Yang S, Peng L (2006) Construction of high order balanced multiscaling functions via PTST. Sci Chin Ser F 49(4):504–515
67. Yang S, Wang H (2006) High-order balanced multiwavelets with dilation factor *a*. Appl Math Comput 181(1):362–369
68. Zhang JK, Davidson TN, Luo ZQ, Wong KM (2001) Design of interpolating biorthogonal multiwavelet systems with compact support. Appl Comput Harmon Anal 11(3):420–438
69. Zhou S, Gao X (2006) A study of orthonormal multi-wavelets on the interval. Int J Comput Math 83(11):819–837

N

Nanocomputers

FERDINAND PEPER
National Institute of Information
and Communications Technology, Kobe, Japan

Article Outline

Glossary
Definition of the Subject
Introduction
Wires and Devices
Nanofabrication Techniques and Architectures
Heat Dissipation
Fault-Tolerance
Cellular Automaton-Based Architectures
Crossbar Array-Based Architectures
Neural Network-Based Architectures
Future Directions
Bibliography

Glossary

Adiabatic switching Switching with asymptotically zero speed with the aim of reducing power consumption in a circuit.

Asynchronous circuit Circuit that is designed to work in the absence of a clock.

Babbage engine Mechanical calculator built by Charles Babbage in the early 19th century.

Bottom-up fabrication Fabrication method employing the natural ability of physical structures (including atoms and molecules) to organize themselves into desired structures.

Brownian motion Random movement of micrometer-sized particles due to collisions with molecules. The term is also used to indicate random movement of smaller-sized particles, or the mathematical model of such movements.

Carbon nanotube Nanometer-scale tube consisting of a graphite sheet rolled up into a seamless cylinder. Carbon nanotubes intended for nanoelectronic applications are mainly single-walled.

Cellular automaton Discrete regular array of cells, each of which is in one of a finite number of states. A cell is updated in discrete time steps according to a transition rule that takes the states of the cell and its direct neighbors and determines the next state of the cell. If all cells are updated at the same time, the model is called *synchronous*, otherwise *asynchronous*.

Complementary metal-oxide-semiconductor (CMOS) Currently dominant technology used to implement digital logic circuits as well as a wide variety of analog circuits such as image sensors and data convertors.

Computational universality Ability of a computing system to compute every function in a certain class of systems.

Computer architecture Functional and structural design and operational specification of a computer system.

Construction universality Ability of a logic system to construct every arbitrary logic structure in a certain class of systems.

Crossbar array Array consisting of two layers of wires, whereby wires within one layer are parallel to each other but perpendicular to the wires in the other layer. The wires are connected to each other at their crosspoints through devices.

Coulomb blockade Increased resistance at certain voltages to the flow of electrons in a *tunnel-junction*, which is a thin insulating barrier between two conducting electrodes.

Defect-tolerant Ability of a system to remain relatively unaffected by the occurrence of permanent defects.

Delay-insensitive circuit Asynchronous circuit in which the outcomes of operations are unaffected by delays in wires and functional elements.

Device Functional element in a circuit that takes input signals and produces output signals. A transistor is a well-known example of a device.

Dielectric material Medium that is a poor conductor of electricity, but an effective supporter of electric fields. A dielectric with a high dielectric constant κ is the preferred material used as a gate-dielectric in a transistor, since it allows for a thicker insulating layer between gate and channel given a certain gate capacitance. A low-κ dielectric is the preferred material used for insulating layers between wires, since it allows for smaller wire pitches.

Entropy Measure of uncertainty associated with a certain information-theoretic or physical variable.

Error correcting code (ECC) Encoding that adds redundancy to information to increase the ability to correct or detect errors caused by noise or other impairments.

Fault-tolerant Ability of a system to remain relatively unaffected by the occurrence of errors.

Field-effect transistor (FET) Transistor in which the conductivity of the channel depends on the electric field controlled by the transistor's gate.

Fine-grained parallelism Scheme for the subdivision of tasks in a large number of small subtasks that can be simultaneously executed by a large number of simple information processing elements.

Finite automaton Logical scheme consisting of a finite number of states, including a start state and an accept state, as well as a specification of state changes under the influence of inputs.

Field-programmable gate array (FPGA) Type of chip containing logic components and programmable wires that can be programmed to perform a wide variety of combinational logic functions.

Gate A logic gate is a digital device that carries out a Boolean bit-operation, such as the AND, OR, NOT, etc. A transistor gate is the part of the transistor that controls the conductivity of the channel between the transistor's source and drain.

Hamiltonian Property corresponding to the total energy of a system's state that is determined by some sequence of physical operations. A Hamiltonian is mathematically expressed by a unitarian operator H.

Heisenberg uncertainty principle Relationship in quantum mechanics, giving a lower bound on the product of the uncertainties of two physical observables. These two observables may be position and momentum, or, alternatively, energy and time.

Lithography Microfabrication process in which a pattern is transferred to a photosensitive material to selectively remove parts of a thin substrate.

Majority gate A logic gate in which the value of the output bit is set to the logic value occurring in the majority of the input bits. A majority gate usually has an odd number of input bits.

Markov chain Discrete-time stochastic process of which the next state solely depends on the present state and not directly on previous states. In other words, the process is *memoryless*.

Molecular electronics Electronics in which the components (wires and devices) are realized in terms of molecules. These molecules are usually organic.

Moore's law Trend along which integration density of microelectronics has developed since the 1960's, the time Gordon E. Moore first observed this trend. Integration density according to this trend increases exponentially, doubling approximately every two years.

MOSFETS FETs implemented by CMOS technology.

Neural network Mathematical model based on the biological nervous system, consisting of *neurons* that receive (usually) analog values from each other through weighted interconnections. Learning in a neural network takes place through updating the weights based on the values of the neurons and the values of the input signals.

Neuron Biological neuron is a cell in the nervous system that processes and transmits information. The central part of a neuron is its *soma* (cell body), and it has an extension called *axon* to transmit information to other neurons via axon terminals called *synapses*. This information is received by a neuron through its *dendrites*. A neuron in an artificial neural network is modeled after a biological neuron. It receives inputs via weighted interconnections, which model the strengths of synapses.

NP-complete Class of decision problems for which no polynomial-time (expressed in terms of the input size) algorithms are known. Any member of the wider class of NP (Non-deterministic Polynomial time) problems can be transformed in an NP-complete problem, in which the time overhead of the translation is at most a polynomial factor.

Parallel A computation is parallel if it is divided in smaller computations that can be executed simultaneously.

Perceptron A type of neural network model invented in 1957 by Frank Rosenblatt that is used for classification. Though originally this model consisted only of a layer of input neurons and a layer of output neurons, modern uses of the term includes the possible presence of one or more hidden layers of neurons. The perceptron is a *feedforward* neural network, which means that information flows in one direction, from the input to the

output, so their are no backward connections between neurons.

Pitch Term used in integrated circuits to denote distance between elements, such as between cells in RAM memory or between the centers of two wires. Commonly used is the term *half-pitch*, indicating half this distance.

Processor The computing part of a computer also called Central Processing Unit (CPU).

Quantum dot cellular automaton Cellular automaton in which cells based on quantum dots containing electrons interact with each other through electrostatic forces.

Ratchet Device that allows a process (such as the movement of particles) to take place in only one direction.

Repeater Logic device placed on a wire to reproduce signals input to it. Usually implemented in terms of NOT-gates, repeaters speed up the propagation of signals along wires in highly integrated microelectronics.

Resonant clock Timing mechanism on synchronous chips that employs the resonance of oscillators to achieve sharply reduced power consumption and increased preciseness in timing.

Resonant tunneling device Device using quantum effects to allow very efficient transmission of electrons through a double barrier tunneling structure.

Reversible computation Computation of a function that is one-to-one.

Scanning tunneling microscope (STM) Type of electron microscope to view surfaces at the atomic level with resolutions of up to 0.1 nm lateral and 0.01 nm in depth. The STM employs a tip from which electrons tunnel to the surface, whereby the tunnel current depends on the distance between the tip and the surface as an exponential function. The STM can also be used to manipulate individual atoms and molecules.

Single electron tunneling device Device based on the tunneling of individual electrons through one or more tunneling barriers, which are thin insulating layers between electrodes.

Spintronic device Device based on the magnetic spin states of electrons.

Superposition of states Linear combination of states in a quantum system describing a situation in which a physical observable possesses two or more values simultaneously.

Synapse Receptor on a neuron's axon that connects to a dendrite of a neuron to transmit information.

Top-down fabrication Fabrication method in which structures are formed under the control of a master plan. Optical lithography is the usual top-down method for fabricating microelectronic chips.

Tunneling Quantum mechanical phenomenon in which a particle passes through an energy barrier that would, given the particle's kinetic energy, be too high for it to pass by classical physical laws.

Tunneling phase logic Logic that uses the phases of waves to conduct logic operations.

Turing machine Abstract logic model consisting of a tape, a reading and writing head, and a finite automaton to control the head. This simple machine model is used to study computation and the relations between computational models.

Very large scale integration (VLSI) Microelectronics technology with millions of devices on a chip.

Voltage encoding Encoding of the value of a signal by the level of a voltage. This is the most commonly used method to encode signals in electronics. Opposite of *charge encoding*, in which the value of a signal is determined by the presence of a small number of elementary electrical charges.

von Neumann neighborhood Set of cells in a cellular automaton that neighbors orthogonally to a cell. The von Neumann neighborhood of a cell in a 2-dimensional cellular automaton consists of the cell's northern, eastern, southern, and western neighboring cells. Often the cell itself is also included in the definition of neighborhood.

Definition of the Subject

Nanocomputers are (not-yet-realized) computers that will be based on technology employing devices and wires with feature sizes in the order of a few nanometers (10^{-9} m). If the increase in integration density of microelectronics according to Moore's law [143] continues at the same pace as it has for almost 40 years, such computers will be around in a few decades. Computational power and speed of nanocomputers will likely dwarf those of most contemporary computers if trends from the past continue. It is anticipated that silicon-based CMOS technology can be extended to up to the year 2015; beyond that, major scientific and technological breakthroughs will be required, according to the International Technology Roadmap for Semiconductors (ITRS) [94]. Many of these breakthroughs will take place on the physical level, via the development of new devices, but progress on an architectural and algorithmic level will also be indispensable. Among the issues that need to be addressed in this context are the following:

- How to build circuitry with feature sizes well below what can be potentially realized by extensions of traditional manufacturing technologies, like those based on optical lithography, even if taking into account major

advances in technology? Will manufacturing based on molecular self-assembly and self-organization be able to deliver such a feat? What will the impact of the adoption of such bottom-up manufacturing have on the architectures of computers built by it? Will it require structures to be highly regular or random, and how will this affect the way such computers are configured and programmed?

- How to cope with heat dissipation, which will be especially debilitating at the extremely high integration densities expected in nanocomputers? Cooling techniques seem to be insufficient in this context, leaving only the reduction of heat dissipation as an option, which will inevitably involve the reduction of energy consumption. Will it be sufficient to this end to develop new devices, like carbon nanotube-based devices, resonant tunneling devices, single-electron tunneling devices, spintronic devices, molecular devices, etc., or will other innovations be required? Will it be necessary to use asynchronous circuits, i. e., circuits that work without a clock? Alternatively, will it be necessary and possible to use reversible and adiabatic computing schemes (► Reversible Computing)? And how will these schemes affect a nanocomputer's architecture?
- How to cope with the unprecedented rates of errors and manufacturing defects that are expected in nanocomputers? In case of manufacturing defects, will it be possible to reconfigure a nanocomputer around such defects, and if so, how to detect the locations of such defects and how to reconfigure circuitry around them? In case of transient errors, which are associated with noise and fluctuations, and which are expected to occur at rates far beyond those experienced in CMOS technology, how to resolve such errors?

Some of above issues have attracted interest in the past, but, ironically, the success of CMOS technology has been a delaying factor in the initiation of systematic studies into them.

Introduction

In his 1959 lecture entitled *There's plenty of Room at the Bottom* [60], Feynman predicted a future in which a single bit would require a mere 100 atoms to be stored. The same lecture pointed out the possibility of computers with devices miniaturized to similar scales. Nowadays such machines are called *nanocomputers*. Feynman's proposal, though visionary, was not made in a vacuum. Molecular conduction, which is of paramount importance to molecular electronics, was investigated by Mulliken and Szent-Gyorgi in preceding decades [89]. Furthermore, as early as 1956 Arthur von Hippel advanced his view toward the engineering of materials from atoms and molecules [200], based on a firm theoretical understanding of material properties. These ideas made their way into the electronics industry through Westinghouse, which started a Molecular Systems Engineering program in 1957 [33]. The first conference on Molecular Electronics was held in 1958, jointly organized by the Air Research and Development Command, as well as by the National Security Industrial Association, and attracting approximately 300 participants [33]. The interest in military circles for this new field stemmed especially from the need to reduce size, weight, and cost of devices, as well as to increase reliability of electronics, which at the time was far worse than today. Westinghouse's endeavor in molecular electronics received funding from 1959 to 1962 from the US Air Force, but it was discontinued due to a lack of concrete results, especially with respect to manufacturing issues, though it appears to have had a positive spin-off on clean room techniques and scanning electron microscopy [33]. The effort disappeared silently from public view in successive years (see p. 181 in [30]). In retrospect, the idea was far ahead of its time.

A decade later, in the mid 1970's, molecular electronics received renewed attention, when Aviram pursued his radical vision of devices based on individual molecules together with Ratner [6]. Around the same time, Carter began work on molecular wires, switches, molecular logic elements, and molecular computers [26,27,28], thereby becoming one of the pioneers of nanocomputer architecture.

The 1980's saw an intense interest in the use of physics for the implementation of computation [12,19,65], with special focus on cellular automaton-like architectures. Drexler, around the same time, started to popularize nanotechnology and computers built by it to a general audience [55,56].

In the 1990's, interest widened to a variety of architectures and methods with potential for nanocomputing. Reliability issues of nanocomputers also started to attract attention. Representative of these efforts is the Teramac computer [84].

This trend appears to continue in the first decade of the 21st century, with unconventional computing models finding an increasing audience.

Though nanocomputers have always been overshadowed by their silicon-based counterparts, they receive attention due to CMOS technology being thought of in danger of running into serious physical limitations. Some predictions for limitations over the years are listed in Table 1 (see [97]). Though most of the expected limitations in the past have been overcome through a reexamination

Nanocomputers, Table 1

Past predicted limitations on gate lengths for downscaling of MOSFETs. The listed values are of gate lengths, which are smaller than the generation size listed in the International Technology Roadmap for Semiconductors (ITRS) [94]. For example, a 30 nm gate length corresponds to the 100 nm technology generation. The term *red brick wall* is named after the color-coded tables in the ITRS report, and indicates a limit beyond which no technology solutions are known (at the time)

Period	Expected limit	Cause
Late 1970's	1000 nm	Short -channel effects, lithography limitations
Early 1980's	500 nm	Source / Drain resistance
1980	250 nm	Tunneling leakage, dopant fluctuation [136]
Late 1980's	100 nm	Red brick wall: various
1999	30 nm	Lower limit to supply voltage [43]
2004	50 nm	Red brick wall: various
2004	10 nm	Fundamental

of the assumptions underlying them, coupled with human ingenuity, this may become difficult as the limitations increasingly assume a fundamental physical nature. Important fundamental physical limits include the following[127,138]:

Thermal limit refers to the energy required for a binary switching operation, which should at least exceed the energy of thermal fluctuations, less an error occurs. This amounts to an energy of $(\ln 2)kT$ J, whereby $k = 1.38 \cdot 10^{-23}$ J/K is the Boltzmann constant and T the temperature in Kelvin. The probability P of a thermal fluctuation of energy E_s within a response time τ of the circuit follows the Boltzmann relation [136] $P = \exp\left[-E_s/kT\right]$, which implies a rate of $(1/\tau)\exp\left[-E_s/kT\right]$ failures per second per device. The probability due to error can thus be reduced by increasing the signal energy. To be on the safe side, a limit of $100kT$ J per switching operation is usually assumed. The limit associated with this minimum level of reliability is still four orders of magnitude below the switching energy for 100 nm CMOS technology.

Quantum limit relates the signal switching energy transfer ΔE to the transition time t via the Heisenberg uncertainty principle, $\Delta E \cdot t \geq h$, where $h \approx 6.6260 \times 10^{-34}$ is Planck's constant. Resulting from the wave nature of the electron and the associated uncertainty in the energy-time (or the position-momentum) relations, this limit imposes an upper bound on device switching speed given a certain average switching energy. CMOS circuits operate far above the quantum limit.

Speed-of-light limit imposes an upper bound on the distance traveled by a signal in one clock cycle. Given that electrical signals travel half the speed of light ($c \approx 3 \cdot 10^8$ m/s) in typical materials, a distance of merely a few centimeters can be covered within one clock cycle on a 10 GHz chip.

An extensive analysis of limits, ranging from a fundamental level, materials, devices, circuits, up to systems, is given in [138].

Nanocomputers come in various flavors and are distinguished according to the underlying mechanisms involved. Montemerlo et al. [142] distinguish four types:

Electronic Nanocomputers The natural successors of electronic digital computers, these computers represent and process information through the manipulation of electrons. Though conventional transistor technology will be prevalent in the coming decade, alternatives that use tunneling of electrons or their spins have received serious attention, because of their low power potential.

Mechanical Nanocomputers These computers are miniaturized equivalents of Babbage engines that conduct their operations through moving molecular-scale rods and rotating molecular-scale wheels, as envisioned in [55,56]. When combined with electrical signals, switches in such computers have excellent On/Off ratios. The high mass of atoms, as compared to electrons, however, gives these computers a speed disadvantage to their electronic counterparts.

Chemical Nanocomputers Based on processes involving making and breaking chemical bonds to store and process information, these computers have extremely small devices, but they also will likely be slow. Most efforts in this framework find their roots in biochemistry. DNA-computers [3] (▶ DNA Computing) have attracted much interest in this context, though their much-touted use for solving NP-complete problems appears to run into practical limits [81].

Quantum Nanocomputers Quantum nanocomputers (► Quantum Computing) aim to exploit quantum effects such as superpositions of states to achieve vastly improved performance for certain algorithms (see [176] for an overview of recent progress in quantum algorithms). Though quantum effects increasingly play a role as feature sizes decrease, they have mostly be considered of significance in the context of devices, rather than in a quantum computational framework. Quantum nanocomputers seem not to be within the reach of practical applications before the year 2020 [8].

This chapter focuses mostly on electronic nanocomputers and mechanical nano-computers, these being the most likely successors of current computers. It is organized as follows. After an overview of devices and wires for nanoscale implementations in Sect. "Wires and Devices", nanofabrication techniques are discussed in Sect. "Nanofabrication Techniques and Architectures", because they have a direct impact on the architectures of nanocomputers. Top-down techniques, such as optical lithography, face serious limitations in this context. Unsurprisingly there is an increasing interest in bottom-up techniques, which can fabricate structures varying from highly regular on one hand to highly random on the other hand. Heat dissipation, and especially strategies to minimize it, will be the topic of Sect. "Heat Dissipation". Asynchronous timing and reversible logic receive particular attention, but computation schemes that aim to cope and even to exploit noise are also discussed. Section "Fault–Tolerance" discusses fault-tolerance, distinguishing faults occurring during operation – which can be corrected by employing redundancy of some resource in the architecture – and faults of a permanent character, caused during manufacturing – which tend to lend themselves more to reconfiguration techniques. Following these basic topics are three sections describing architectures with particular promise for nanocomputers, i. e., Cellular Automata in Sect. "Cellular Automaton-Based Architectures", Crossbar Arrays in Sect. "Crossbar Array-Based Architectures", and Neural Networks in Sect. "Neural Network-Based Architectures". This chapter finishes with a discussion about future directions.

Wires and Devices

Three decades of development has left the basic concepts and structures of wires and devices implemented by CMOS relatively unaffected [211]. This may change, however, as feature sizes decrease further, and technological limits are approached.

The function of an interconnect or wiring system is to distribute clock and other signals and to provide power/ground, to and among, the various circuits/systems functions on a chip [96]. The ever-increasing integration densities of chips are accompanied by a quest for increasing speeds of signal transmissions over wires. Unfortunately, the expected delay of a wire of a certain length, which is proportional to the wire's resistance as well as to its capacitance, tends to compare unfavorably to gate delays when feature sizes decrease. In practice, the increased delays of wires relative to gate delays under scaling are mostly felt by wires whose lengths cannot be downscaled, due to them running across the chip. Wires that do scale down in length are less sensitive (though not immune) to this problem [86]. Since most architectures require a substantial level of non-local interconnects, many wires will need to be sufficiently wide to offer the decreased delays associated with lower resistance. This has resulted in designs with an interconnection hierarchy on chip layouts of local thin wires at the bottom, and intermediate wires connecting functional units in the middle, to fat wires for global routing at the top of the hierarchy [192]. There is much incentive to develop new conductor and dielectric materials, to keep pace with technology requirements. Notable has been the replacement of aluminum by copper for conductors from around the late 1990's, as well as the introduction of low-κ dielectrics, which allow decreased pitch (interspacings) between wires.

Whereas transistor delay used to be the bottleneck in past technology generations, it has been superseded by wire delays. The reason for this lies in the increased needs for relatively long wires due to the increased complexities of designs [86]. The growing gap between transistor delays and wire delays can be addressed through the use in wires of *repeaters* at constant-length intervals to each other. Though this tends to result in smaller delays, it goes at the expense of increased silicon area and power consumption [86]. Ultimately the solution lies in shorter wires, and this favors architectures with local interconnection schemes [11]. Absent the use of such architectures, other radical concepts and solutions to the interconnect problem need to be contemplated. The ITRS of 2005 [96] mentions the following options:

1. Different signaling methods, like *Raised Cosine Signaling* [10], which uses raised cosine pulses instead of square pulses to cope with noise crosstalk problems. Also being investigated is *Resonant Clocking* [153], which uses on-chip inductors to effectuate resonance of clock pulses, and which leads to substantially decreased power dissipation.

2. Innovative design and package options that have optimization of the interconnects as their prime objectives.
3. Three-dimensional interconnects, to decrease the average wire length [166]. It may be a challenge to remove heat from such structures, however; it may also be necessary to use new system architectures and design tools [188].
4. Different physical principles by which to transfer signals. Among these, optical interconnects [139] promise high propagation speeds and bandwidths, high precision clock distribution, absence of crosstalk, and reduced power dissipation. Another possibility is signal transmission by Radio Frequency (RF) microwaves [151], which is accomplished though the placement of small antennas on a chip that are able to receive a global signal, like a clock signal. Finally, guided terahertz waves [104] and plasmons (e. g. [205]), which are hybrids of RF and optical signaling, use transmission frequencies around 1 THz and may result in significantly increased bandwidth.
5. Nanoelectronics-based solutions. Carbon nanotubes [106] are especially interesting due to their high conductance, and their potential to be used as semiconductors as well. Other solutions may include the use of conducting channels, such as molecular interconnects, and quantum-mechanical interactions, such as spin coupling and tunneling (see [203] for an overview).

Connecting wires to each other on nanometer-scales is nontrivial due to alignment issues. For this reason, wires are often interconnected through a perpendicular arrangement, like in crossbar arrays, which are discussed in more detail in Sect. "Crossbar Array-Based Architectures".

The function of devices is the processing of signals input to them and, in response, outputting signals according to certain specifications. The Field-Effect Transistor (FET) has long been the basis for successive generations of CMOS. The generation commercially available from the end of 2007 – processor chips based on a 45 nm process – will feature high-κ dielectrics and metal gates [22]. This combination allows for a thicker gate insulation layer given a certain gate capacitance, thus substantially reducing leakage through tunneling – the reason for this being that tunneling decreases exponentially with an increasing thickness of the insulation.

Key factors in the success of MOSFETs and circuits based on them have been noise tolerance of digital circuits and a good ability for fan-out due to the high signal gains of transistors [167], and these factors may also feature prominently in nano-electronics. Moreover, in order to be competitive with CMOS, nano-electronic devices and circuits should satisfy the following requirements [95,189]:

- Scalable by several orders of magnitude beyond CMOS.
- High information/signal processing capacity. This may be achieved through high switching speeds or a high degree of parallelism.
- Energy dissipation that is much less than in CMOS.
- Room temperature operation.

Nanoelectronic devices also need to satisfy requirements specific to their small feature sizes. Alignment, for example will be much more difficult for three-terminal devices, like transistors, than if only two terminals are involved, like in FETs based on crossed nanowires [40].

Another important requirement for nanometer-scale devices is that they constitute a *universal set of operators*, which means that every desired circuit can be constructed as a circuit based on these devices. Most designers take it for granted that any arbitrary logic circuit can be constructed from p-type and n-type transistors, but this requirement may become less trivial when alternative devices are used. What then are promising candidates to become successors of MOSFETs?

Carbon Nanotube (CNT) FETs are considered to be the foremost candidate in this context. The channel through which current is rectified by a gate in CNT FETs consists of a single-walled carbon nanotube with semiconducting transport properties (e. g. see [4]). There are still many open issues with CNTs, including the need for a better understanding of the physical mechanisms underlying their operation, problems with the synthesis of CNTs with appropriate characteristics, and fabrication issues like placement, material integration, On/Off ratios, etc. [95].

Resonant tunneling devices [64,129] include resonant tunneling diodes (RTD), which have two terminals, and resonant tunneling transistors, with three terminals. An RTD consists of a double barrier structure through which electron transmission is highly efficient about certain resonance energy levels. These devices have novel characteristics, such as negative differential resistance, and they have potentially very high switching speeds and could result in circuits with a reduced number of components and power dissipation. However, they seem to be unable to form a universal set of operators, since no designs are known for memory cells and logic gates only involving RTDs [148].

Single Electron Tunneling (SET) devices [198] are based on the controlled motion of individual electrons. These devices involve junctions, through which tunneling takes place, and Coulomb islands that temporarily store electrons and that are usually implemented as quantum

dots. SET devices have the potential for high density and power efficiency, while allowing fast switching, but they tend to suffer from low noise immunity as well as from the problem that their gain is insufficient to allow satisfactory fan-out [95].

Molecular devices [134,159] utilize single molecules – typically small organic ones – acting as electronic switches and storage elements. Extremely high densities may be achieved with these devices, and they are thought to be very suitable for fabrication by chemical synthesis. Though efficient power usage is expected, these devices may offer switching speeds that are not much higher than those possible by CMOS. Experimental devices have been built, but it has not been made sufficiently clear yet whether the observed device properties are those of the molecules, or rather influenced by those of the used electrodes [95].

Spin logic devices [210], also referred to as magnetoelectronics or spintronics, encode information by the magnetic spin orientations of electrons. Based on the manipulation of the magnetic spin of electrons by magnetic fields or by an applied voltage, these devices may allow very low power dissipation, but unfortunately their projected switching speeds may be limited.

Other devices that may play a future role in nanocomputers are Rapid Single Flux Quantum (RSFQ) devices [124], molecular-scale electromechanical devices [36, 37], and quantum interference devices [46,213].

Nanofabrication Techniques and Architectures

There is a close relationship between the technology used to manufacture computers and their architecture. The irregular, aperiodic, structures of current computers are only possible through the use of *top-down* microfabrication technology, like optical lithography (e. g. see [123,128]). Notwithstanding the physical limitations faced by lithography, its life-time has been extended over and again, through for example resolution enhancement techniques [123], which can carry feature sizes below the illumination wavelength. Alternatives to optical lithography are nano-imprinting [34], in which a master with nanometer-scale features is made using techniques like electron-beam lithography. This master is then used as a stamp and pressed onto a (soft) target surface that can then be filled in with for example metal to create wires. Expected to allow manufacturing down to 10 nm, these techniques will impose relatively few restrictions on structures, though the number of times a master can be used is limited. As top-down fabrication techniques are being stretched to smaller feature sizes, they will increasingly impose restrictions on layouts, as recent developments involving resolution enhancement techniques show [123].

There is a lack of consensus as to what is beyond the year 2015 [94] in the context of top-down technologies, though – given the investments made in them – they are unlikely to disappear for the foreseeable future. That said, building the facilities for top-down fabrication will gradually run into economic limitations, as their costs tend to increase by a factor of two for every chip generation – a trend known as *Moore's second law*. Moreover, feature sizes of less than 10 nm may be intractable for top-down mass fabrication, necessitating additional techniques based on different principles. Denoted as *bottom-up*, these techniques exploit interactions between atoms and between molecules to create useful structures. Bottom-up manufacturing is expected to come at the expense of decreased complexity in architectures, because of the lesser degree of control over the manufacturing process. Architectures possible through bottom-up techniques can be divided in three classes: regular, random, and – a combination of these two – quasi-regular.

Regular architectures are characterized by a repetitive pattern of simple features that can not be changed after fabrication. Typically, regular architectures require the use of some form of molecular *self-assembly*, which is defined as the autonomous organization of components into patterns or structures without human intervention [206]. Self-assembly is related to, but not synonymous with, self-organization, the latter being mainly concerned with pattern formation, and the former mainly involving pre-existing components (separate or distinct parts of a disordered structure) that are designed with a certain outcome of the process in mind [206]. Self-assembly as a nanofabrication method offers a number of advantages [156]:

- it is an inherently parallel process,
- it can generate structures with sub-nanometer precision, and
- it can generate 3-dimensional structures.

When self-assembly's outcome is controlled through external forces or geometrical constraints that are supplementary to the original interactions driving it, it is called *directed self-assembly*. The control allowed by directed self-assembly extends from merely tuning the interactions between individual assembling components to positioning the components at a desired location [156], like the positioning of wires through an electrical field [177]. Manufacturing by self-assembly requires sufficient control over the process to guarantee regularity. This implies the existence of some mechanism that – though in itself lacking the control of top-down fabrication – can check and

correct errors in the fabrication process. The manufacturing of structures based on the self-assembly of DNA tiles [116,160,169,209] is an example of this. Such tiles form a skeleton that can be filled in later, using techniques to transport matter towards certain locations in DNA tiles [175]. Though practical architectures based on such tiles are yet to be realized, computing on such structures in the future can be envisioned. An important question in this context is how complex each tile should be. If its complexity approaches that of a conventional processor, it is unlikely that it can be manufactured by self-assembly, because of the absence of much regularity, even in the case the processor is not very complex. Self-assembly would require a much simpler tile, with a complexity not exceeding that of a simple finite automaton. This has led various researchers to investigate *Cellular Automata*, which are based on regular (usually 2-dimensional) arrays of tiles (cells), each filled with a finite automaton (► Cellular Automata as Models of Parallel Computation). Section "Cellular Automaton-Based Architectures" describes the use of this model for nanocomputer architectures.

Another regular architecture attracting attention from researchers is the *Crossbar array*. It employs a 2-dimensional mesh of wires that have switches at their crosspoints (see Sect. "Crossbar Array-Based Architectures"). Regular architectures require post-fabrication configuration to impose a certain desired functionality on them. Configuration may take place, for example, by software initializing cell states of a cellular automaton or setting the switches of a crossbar array. To allow reuse of the architecture for computation, one usually assumes that it can be configured more than once in various configurations. In other words, the architecture should be *reconfigurable*.

At the other end of the spectrum are *random* or *unstructured* architectures. The irregularity of such architectures tends to be mainly caused by the inability of the manufacturing process to perfectly control the formation of structures. The resulting randomness is usually considered part of the architecture to begin with. The elements that are not part of an unstructured architecture will be considered defects, because they are not planned. Lacking a structure imposed on them at manufacturing time, such architectures rely on post-fabrication configuration. To configure a circuit – which may for example be as complex as a processor – on the underlying hardware, it is necessary to have detailed information in advance about the hardware's structure. This is usually accomplished by an external host computer. In other words, scanning of the underlying hardware is not self-contained in the architecture. This indicates that random structures will be mostly used in combination with one-time configuration of a circuit, since on-line scanning of the hardware by a host each time that a circuit needs to be mapped may be impractical and too time-consuming. Examples of random architectures are Neural Networks [197] and – related to them – the Nanocell [195] (see Sect. "Neural Network-Based Architectures").

Randomness is hard to avoid in bottom-up fabrication of nanometer-scale features, due to the occurrence of defects. For this reason, perfectly regular architectures may be hard to achieve in nanocomputers. When defects can be dealt with by a one-time configuration of the hardware around them, the term *quasi-regularity* is used. See Sect. "Fault-Tolerance" for a discussion of defect-tolerance and fault-tolerance.

It is unlikely that one day mankind suddenly wakes up in a world dominated by bottom-up manufacturing. Rather, it will be gradually introduced, incorporated at first in the framework of top-down manufacturing, and increasing its share over the years. This will likely result in architectures in which CMOS is combined with nanoelectronics, like in [32,122,125,181,190,219].

Heat Dissipation

The huge integration densities facilitated by nanotechnology will have far-reaching consequences for heat dissipation. Being a growing problem in VLSI chips, the heat dissipated by nanoelectronics per unit of area will reach impractical values if techniques, materials, devices, circuits, and architectures continue to be used as they have been in the past. Chips dissipate power proportional to the number of devices N, to the clock frequency f, to the probability p (typically 0.1) that a device is active during a clock cycle, and to the average energy dissipation E of an active device in one clock cycle [148,159], giving a power dissipation of $P = NfpE$. CMOS designs have been by and large governed by the scaling rules proposed in [51]. Table 2 shows some parameters in this context: the power dissipation per device decreases quadratically with downward scaling, and this is indeed what happened in the 30 years since the publication of the table. Coupled with a quadratically increasing number of devices per unit of area, this rule implies a constant power dissipation per unit area over successive generations. There is a limit in this regard, however, because the required downscaling of the power supply voltage causes an increased susceptibility to noise [102] and increased off-state drain leakage currents [137]. Ultimately the decreased switching energies associated with downscaling run into the thermal limit, discussed in Sect. "Introduction". To make matters worse, the number of devices N and the frequency f will

Nanocomputers, Table 2

Technological scaling rules for MOSFETs. From the 1970's on, each new chip generation – introduced at intervals of two to three years – has seen its minimum feature sizes reduced by a factor $K \approx 0.7$. The first item *device dimensions* applies to channel length, oxide thickness, gate width, etc. The table assumes unchanged electric fields in devices over successive generations

Device Parameter	Scaling Factor
Device dimensions	$1/K$
Voltage	$1/K$
Current	$1/K$
Gate capacitance	$1/K$
Delay time	$1/K$
Power dissipation	$1/K^2$

increase as feature sizes move into the range of nanometers. To see where this ultimately leads to, Zhirnov et al. take the physical parameters of an imaginary system to its extreme, resulting in a "limit technology", in which a chip contains the smallest binary switches, packed to maximum density and operating at the lowest possible energy per bit. This would give a power density in the range of a few MW/cm^2 [218], which dwarfs the power density of the sun's surface by three orders of magnitude. Unfortunately, the capacity to remove heat is limited to several hundred W/cm^2 when known cooling methods for two-dimensional structures are used [218]. Additional measures thus seem justified, and this section looks for them in terms of architectures and algorithms.

Goldstein [74] proposes to reduce heat dissipation problems by spreading circuits out in space, reducing the number of times each part of a circuit is used in a computation. The resulting increase in parallelism allows for higher clock rates. Since the circuitry is spread over a larger area, a lower power density results, but this goes at the cost of longer wires, which causes increased delay and power consumption.

Other approaches seek to reduce heat dissipation in logic circuits by focusing on the clock. Each time clock signals are distributed over a circuit, energy will be pumped into wires, much of which gets lost in the form of heat. A novel approach to clocking problems is to use *resonant clocks*. These designs employ clock distribution circuitry that act as transmission lines with uniformly distributed capacitors and inductors, in which electromagnetic waves propagate long distances without the need for repeaters to amplify signals [9,31,141,152,212]. Since electrical charges on parasitic capacitors can be extracted and preserved by the inductors, this approach allows for a significantly reduced power consumption, but it has problems of its own, like unpredictability due to noise and reflections, difficulties to distribute clock signals in a clock-tree due to impedance mismatches, and the inability to yield constant phase and magnitude of clock signals simultaneously [9]. To deal with some of these problems, while preserving the important advantage of low-power, Banu et al. [9] propose to use two identical side-by-side transmission lines with linear topologies in which clock signals travel in opposite directions, allowing the extraction of an absolute phase at any point of the clocking circuitry by averaging the two signals' timings. The overhead of the averaging circuitry, however, may compromise the efficiency of this solution.

A radical way to deal with clocking problems is to remove the clock all together, resulting in *asynchronous circuits*, which have a range of advantages. Apart from not having the clock's overhead of increased energy consumption and increased area, such circuits have the fortunate tendency to restrict energy consumption to only those parts that are actively switching. In other words, asynchronous circuits have an inherent mechanism to selectively shut off parts of them – even down to fine granularities of individual devices – a feature for which synchronous circuits require special *clock gating* circuitry.

Apart from the reduced power consumption and heat dissipation offered by asynchronous electronics – provided it is well-designed – there are also other advantages:

- Problems with the distribution of clock signals disappear, like *clock-skew* (timing differences at which different parts of a circuit receive the clock-signal) and *race conditions* (the failure of signals to reach their destinations within a clock cycle). These problems tend to get worse, otherwise, with increasing integration densities.
- Less noise and electromagnetic interference.
- Insensitivity to physical implementations and conditions. This implies more freedom in the timing of signals and layouts of circuits, which is useful in the reconfigurable architectures expected in nanocomputers [74].
- Average rather than worst-case performance. An asynchronous circuit operates as fast as switching times of devices allow, rather than being limited by the slowest parts of a circuit, which in synchronous systems imposes an upper limit to the global clock rate.
- Modularity. Asynchronous circuits can be designed in terms of modules that can be combined without considerations of timing restrictions. This facilitates flexible rearrangement of circuit elements into various circuits.

Asynchronous circuits come in many flavors [44,82], which are distinguished by the extent as to which assump-

tions are made about timing of signals, both of a circuit and in the interaction of a circuit with its environment. The most general class of asynchronous circuits assumes certain bounds on delays of signals [199], but unfortunately it is the least robust to unexpected behaviors, which complicates their design. Forming the second class are *self-timed circuits*, which only make assumptions on timing within circuit elements, like bounded delay etc., but not on timing between modules [174]. *Speed-independent circuits*, the third class, assume unbounded delays in circuit elements, but zero or negligible delays on wires between the elements [147]. The fourth and fifth class form the *quasi-delay-insensitive circuits* [130] and the *delay-insensitive circuits* [178]. The latter are robust to delays in both circuit elements as well as wires, but this is at the cost of significant design complexity. For this reason quasi-delay-insensitive circuits are more common. They differ from delay-insensitive circuits in their requiring a so-called *isochronic fork* [133], which is a fanout element of which all output branches have a similar delay. In practice, this class of circuits shares many similarities with speed-independent circuits.

Notwithstanding the advantages of asynchronous circuits, their use has been limited to few applications. This is partly due to the much smaller base of design and testing infrastructure when compared to synchronous technology, though asynchronous design support technology is increasingly being developed [191]. One of the key motivations to use synchronous circuits in the first place used to be ease of design, as there is – as compared to asynchronous circuits – less unexpected circuit behavior accompanying changes in the values of signals. The advantages of synchronous design have become less evident today than in the past, as timing problems have become more urgent with the increase of clock frequency, but synchrony still defines the mindset of generations of circuit designers.

The conditions surrounding nanoelectronics may change that. Not only will problems accompanying a clock become more severe with the increase of integration densities, the entire nature of systems will change. Whereas signals in digital logic have been voltage-encoded in virtually all systems, they may be represented in quite different ways on nanometer scales, to better reflect the physical environment. As signals more and more approach the scales of individual particles, a token-based representation in which signals have a discrete indivisible character may be more appropriate. Circuits based on single electron tunneling [198] are already moving in this direction [154], as do experimental circuits based on molecule cascades [58,85]. Asynchronous timing is very suitable for token-based circuits, and may thus play an important role in this context [157].

Another approach to minimize heat dissipation aims to minimize the energy required for switching, even eliminate it altogether. This is possible – at least in theory – by making computations reversible [16,113] (▶ Reversible Computing). The key to this line of thought is the observation that the information stored on a computer is ultimately represented in terms of physical entities [115]. Operations in a computer that change the entropy of this information will accordingly change the thermodynamical entropy of the physical system. The erasure of a bit by overwriting it with a new value, is such an operation [113]. Whereas the bit's entropy corresponds to two possible (unknown) states of the bit before the operation, the bit's value will be fixed into one state by the operation, which decreases the entropy. This implies a decrease of the bit's physical degrees of freedom, and – to compensate for this – there is an increase of unobservable degrees of freedom such as in the microscopic motion of molecules [127], which usually translates into the dissipation of heat. The energy loss associated with such an operation would be at least $kT \ln 2$ J, according to the thermal limit in Sect. "Introduction". While Landauer [113] initially believed that no universal computations could be performed without a change in entropy, Bennett [16] proved this belief wrong by constructing a universal Turing-machine that conducts its operations reversibly, implying that it would cause no change in entropy.

This line of thought, which started in the early 1960's, is widely known today, but it has not resulted yet in practical circuits. Arguably, it will be difficult to overcome noise and friction, which are fundamental to physical processes. Most proponents of reversible computing agree that these factors are an impediment to dissipation-less computing, but they tend to argue that, absent these factors, reversible computing would *in principle* be possible. This would make it a matter of technological progress to approach the zero average switching energy limit arbitrarily close, rather than there being a hard physical lower limit.

One way to go about in this context is to slow down switching speed. Called *Adiabatic Switching* [5], this approach decreases the dissipation caused by the charging and discharging of a gate capacitance C through an associated resistance R. The dissipated energy for charging such a gate is given by [5]:

$$E = \frac{RC}{\tau} CV^2$$

whereby τ is the time required to charge the gate, and V is the voltage supplied to the circuit. An increase in τ

thus implies a corresponding decrease in the dissipated energy, the other factors remaining equal. Schemes of adiabatic switching usually follow the requirement that, first, a switch is only closed when there is no voltage difference between its terminals and, second, a switch is only opened when no current flows through it. Adiabatic schemes of switching tend to be presented in combination with a reversible mode of computation [63]. When reversibility is left out, the resulting adiabatic scheme is usually characterized as *charge recovery* or *energy recycling* [173].

Reversible computing appears to have particular appeal for nanoelectronics, since physical interactions at nanometer scales are typically reversible. Notwithstanding the reversible computing paradigm's wide acceptance, there are still concerns about its feasibility [29,98,165]. Concerns based on problems with noise and friction have been pointed out in [165] in the early years of reversible computing. This is a recurrent theme with skeptics of reversible computation. The overwhelming responses generated by [165] are particularly educating [13,18,114,193], (see also the accompanying reporting in [168]). Other concerns [98] highlight some principal reasons why reversible computing would only work under unrealistic conditions, such as the necessity to have an infinite information-theoretic temperature of the scheme. There are also concerns [29] about possible problems when reversible schemes are combined with adiabatic schemes, the first supposedly being thermally isolated from the environment, while the second scheme is usually not. Other fundamental problems are pointed out in [29] as well with adiabatic switching: apart from thermal noise interfering with the reliability of a simple adiabatic bit-flipping operation, this noise is also responsible for a less than ideal charging curve of a circuit's capacitances, which causes an increased power consumption. Finally, a problem in [29] is claimed with the power supply of adiabatic circuits, which, in order to be the approximate ideal current source needed for adiabatic charging, would require a high internal resistance. The latter would cause much heat dissipation, not in the circuit itself, but in the power supply, so that the overall energy efficiency of the scheme would be inferior to non-adiabatic schemes. A more optimistic note about the prospects of reversible adiabatic computing is found in [63], though it is admitted that there remain significant challenges, the most fundamental of which are the requirement of switching devices that are less resistive and less leaky, and the requirement of higher quality power supplies and clocks, like the resonant clocks mentioned earlier in this section.

The problems with clocks in the framework of reversible computing have led some researchers to consider whether reversible computing can be combined with asynchronous circuits, especially asynchronous circuits that are token-based and delay-insensitive [118,120,145]. Circuits, in which only one single signal token moves around at a time, do not need a clock, so they would qualify as reversible and asynchronous. Though computational universal circuits can be constructed in this way, they have obviously a problem in terms of efficiency. The alternative is circuits in which multiple tokens move around in parallel. This may lead, however, to situations in which reversibility is hard to define [118]. A circuit element having two or more tokens as asynchronous inputs for an operation will – in order to preserve a strict one-to-one relationship between input and output as defined through a (unitary) Hamiltonian operator [59] – need to monitor the arrival times of the individual tokens, but this requires a large overhead and it is counterintuitive to how a circuit is supposed to operate, suggesting that the underlying physical model may need reconsideration.

An unconventional strategy to reduce energy consumption is to use switching operations with a reduced reliability. Palem and co-workers [105,155] assert that by accepting a reduced reliability of switching, less energy will be required to conduct a switching operation. Palem [155] shows that if a switching operation is correct with a probability $p(> 1/2)$, then its energy consumption – assuming that the operation is irreversible – is bounded from below by $kTJ\ln(2p)$. A reduced reliability of switching limits the range of possible algorithms to those that accept some degree of randomness, like probabilistic algorithms [146], image processing algorithms, and classification algorithms (e. g. [66]). Somewhat related is a scheme proposed by Kish [103] that drives computations by noise. Shown is a simple model circuit containing a comparator with error rate around 0.5. Inspired by the highly efficient way in which information is processed in the brain, this thermal noise driven computing scheme requires a minimum energy requirement of $1.1kTJ$ per bit. This scheme may require extensive error correction to make it work. The use of noise and fluctuations as a computational resource is also outlined in [42], as likely being a driving force for mechanisms in biological organisms [216]. Brownian motion provides a free search mechanism, allowing a molecule, signal, or other entity to explore its environment as it randomly moves around, thereby undergoing certain operations (like molecular binding) when it arrives at certain locations (like molecular binding sites). An early mentioning of Brownian motion for computations is made by Bennett in [17], who uses it for reversible computation in case thermal noise cannot be avoided. Bennett considers a Brownian implementation for a mechanical Turing

machine, in which signals move around randomly, searching their way through the machine's circuitry, only to be restricted by the circuit's topology. This search would, in principle, not require any energy to be provided from a power supply, the drawback being that the computation would move forward and backward evenly without any bias, and thus would take a long time to complete. Funneling the random motions into preferred directions would require the selective use of ratchets, but ratchets tend to consume energy, as shown by Feynman [61]. A mechanism somewhat similar to a Brownian Turing machine is found in nature, in the form of the Brownian tape-copying machine embodied by RNA polymerase [17]. The efficient – in terms of energy consumption – use of Brownian motion in various biological mechanisms, such as those involving Brownian molecular motors, suggest that biological organisms may provide inspiration for future nanocomputing systems in this respect.

Noise and fluctuations have also been used in the simulated annealing process of a Boltzmann machine that can be implemented by SET devices [215] (see also Sect. "Neural Network-Based Architectures"), though reductions in power consumption have not been the main purpose of this work.

Fault-Tolerance

When a system experiences a failure in one of its parts and can continue to operate, possibly at a reduced level, rather than – as most systems would do – fail completely, it is called *fault-tolerant*. A fault-tolerant system may experience a reduced throughput or an increased response time in the face of one (or more) of its components failing. Fault-tolerance is an important property for nanocomputers, since the reliability of their components will likely be low due to:

- The probabilistic nature of interactions on nanometer scales, rooted in thermodynamics, quantum mechanics, etc.
- The individual behavior of particles, molecules, etc. gaining importance over their averaged behavior due to their decreased numbers, causing the "Law of Large Numbers" to become invalid below the device level [21,48].

When faults have a permanent character – either incurred at manufacturing time or later – they are usually referred to as *defects*. A nanocomputer architecture that can cope with defects is called *defect-tolerant*.

When faults have a transient character, i. e. when they occur during operation and are temporary, they are referred to as *transient faults*, or, simply, *faults*, like the more general term. They are mainly caused by environmental conditions, such as thermal noise, quantum fluctuations, and electromagnetic perturbations. Cosmic radiation, which is a major source of transient faults in CMOS technology, may be less of a concern, on the other hand, for nanoelectronics based on materials different from semiconductor silicon [72]. The term used to describe robustness of an architecture to the occurrence of transient faults is *fault-tolerance*, like the more general term.

Faults may also be *intermittent*, meaning that they occur during a certain (extended) time interval, then disappear, and may (or may not) later reappear. Intermittent faults tend to be caused by unstable or marginal hardware, due to for example manufacturing residues and process variations. These problems tend to increase with the increase in integration density [38]. Changes in operational parameters, such as temperature, voltage, and frequency changes may trigger intermittent faults [38]. Though transient faults and intermittent faults have many characteristics in common, the latter differ in that they tend to occur in bursts, at the same location. Repairing a part affected by intermittent faults usually eliminates them, whereas transient faults can not be prevented by repairs.

Tolerance to faults and defects can be accomplished at various levels of abstraction [76]: the *physical device level*, the *architectural level*, and the *application level*. An example relating to the physical device level is the tolerance of digital devices to noise due to the digital encoding of signals. Tolerance at the architectural level compensates for malfunctioning of devices by organizing them in structures with a certain level of *redundancy*. Such a structure may consist, for example, of replicated (identical) components, operating in parallel, on which majority voting is used to extract the result of an operation. Tolerance at the application level refers to the ability of a computing application to operate correctly despite defects and faults in the underlying hardware. Research aiming to improve the tolerance at the application level may be found in the framework of probabilistic algorithms or image processing algorithms that tolerate some noise in their input. As tolerance at the physical device level tends to decrease with increasing integration density, the tolerance at the architectural and application levels need to improve to compensate for this. It also works the other way around: it may make less sense, for example, to improve tolerance at the physical device level, if the application level is able to cope with a certain level of faults or defects on its own. Most research efforts on fault-tolerance in the nanocomputing field are in practice directed to the architectural level, and this section is no exception on that.

Redundancy in an architecture is achieved by equipping it with additional resources [112]. Especially important are the following types of redundancy:

- *Information redundancy* refers to the addition of information to the original data such as to be able to recover it in case of faults. *Error Correcting Codes (ECC)* are the usual form to represent redundant information (e. g. [207]). Used extensively to achieve reliable operation in communications and memory, error correcting codes provide a way to cope with the corruption of bits by encoding messages as code words that contain redundant information. This information is used to reconstruct the original code word in case errors occur.
- *Hardware redundancy* refers to the inclusion of additional hardware to the system. This may encompass the use of extra components working in parallel to mask any faults that occur – a technique called *static redundancy*, which includes the use of ECC. When only one element of the extra components are used at a time, to be switched out and replaced by a spare in case of a fault, the term *dynamic redundancy* is used. Dynamic redundancy is usually achieved through periodic testing of circuits or self-checking of circuits, followed up by the *reconfiguration* of a circuit, or even its *self-repair*.
- *Time redundancy* involves the repetition of operations in a system. This may take place after an error is detected, in which case a recovery procedure needs to be started up. Alternatively, as a standard procedure operations may be repeated a fixed number of times, after which the consecutive results are compared to each other. Time redundancy is only suitable to cope with non-permanent faults, since repetitions of operations by a defective component will only repeat incorrect results.

The different nature of defects as opposed to faults is reflected in the different strategies usually followed to make an architecture robust to them, though there tends to be an overlap.

The permanent nature of a defect implies that it needs to somehow be isolated from non-defective parts to prevent it from affecting a system's correct functioning. Strategies to cope with defects thus tend to first detect defects and then reconfigure the system around them. This requires spare components that become activated due to the reconfiguration, while the defective components (or defective connections) are made inaccessible. Absent reconfiguration, a system's performance will be aversely affected by a defect, in terms of throughput, speed, reliability, etc., but this may be ameliorated to an acceptable level if sufficient spare resources are available.

Depending on whether reconfiguration takes place at the same time that a nanocomputer is processing data or not, the terms *on-line* or *off-line* reconfiguration, respectively, are used. On-line reconfiguration implies a continuous monitoring of a computer for defects and accordingly taking action if they are detected. Usually envisioned as being self-contained in the computer, this type of operation needs the detection of defects and the associated reconfiguration to be robust to defects and faults themselves, which complicates its realization. Off-line reconfiguration is usually done under the direction of a host-computer, which creates a defect-table and reconfigures the system accordingly. While easier to realize than on-line configuration, it may be only feasible if the size of a nanocomputer is limited, since otherwise the detection and reconfiguration operations would consume prohibitive amounts of time. A well-known example of off-line configuration of a prototype defect-tolerant nanocomputer is the Teramac [84]. Consisting of 10^6 gates operating at 1 MHz, it is based on Field Programmable Gate Arrays (FPGAs), 75 percent of which contain at least one defect. Having a total of three percent defects out of a total of 7,670,000 resources including wiring and gates, the Teramac needs a rich interconnection structure to find alternative routes around defects. To this end, it employs a crossbar array, which is described in more detail in Sect. "Crossbar Array-Based Architectures".

Fault-tolerance requires a different strategy than defect-tolerance, due to the temporary nature of faults. Such a strategy usually involves a combination of information redundancy and hardware redundancy.

Motivated by the low reliability of electronic components in his time, von Neumann studied how reliable computation could be obtained through the use of unreliable components [201]. His *multiplexing* technique employs a circuit built from faulty universal gates – all of the same type – to simulate a single gate of this type, but with increased reliability. Each (input- or output-) wire of the gate is replaced by a bundle of N wires in the circuit. The circuits investigated by von Neumann are based on the NAND-gate, but other gates could also be used in principle. The logic state 1 or 0 of a bundle of wires is determined by whether the fraction of wires in state 1 – called its *excitation level* – lies above resp. below a certain threshold. The scheme consists of three stages, of which the first – the *Executive Unit* – conducts the logic operation, and the last two – the *Restorative Units* – act as nonlinear amplifiers that restore the excitation level of the output bundle to its nominal level (see Fig. 1). The resulting circuit has

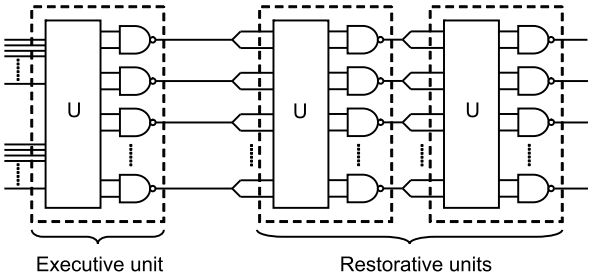

Nanocomputers, Figure 1
Von Neumann's fault-tolerant circuit that is used to replace a single faulty NAND-gate. It consists of three stages: The Executive Unit conducts the actual logic operation, whereas the two later Restorative Units reduce the fault-rate of the Executive Unit's output. The bundle of output wires of the circuit will effectively be reduced to one wire (not shown here), of which the value is determined by the excitation level of the bundle

two input bundles of wires and one output bundle of wires. Wires from the first input bundle are randomly paired by a permutation network labeled U with the wires from the second bundle; the resulting pairs of wires are then used as inputs to the faulty NAND-gates in the Executive Unit. The wires in the output bundle are split, and again randomly paired to form the inputs to the faulty NAND-gates in the first Restorative Unit. This is repeated for the second Restorative Unit. This scheme suffers from a limited reliability, while requiring a high redundancy. For example, an error rate of 0.005 per device requires a redundancy of 1000 to achieve a probability of failure of the gate of merely 0.027. A similar scheme by von Neumann that uses three-input *majority gates*, instead of NAND-gates, does not fare better.

A drawback of von Neumann's scheme is that it entirely focuses on the error rate of gates, but not on that of wires. While the latter can be modeled in terms of the former if wires have a constant length, this condition will not be valid in densely connected circuits, causing failure rates of gates to depend on the length of their input wires [163]. Von Neumann's scheme is at the root of many follow-up schemes proposed in the decades thereafter. Most of these efforts tend to minimize the number of required gates, given a certain failure probability of the components. Dobrushin and Ortyukov, for example, show in [54] that a function computed by m fault-free elements can be reliably computed by $O(m \log m)$ unreliable elements with a high probability. This is further improved in [161,162] to $O(m)$ unreliable elements. An excellent review of these models is found in [163]. This review also shows some variations on this line of research, in which it is not gates that may fail, but contacts of switches. Thought to have become irrelevant with relays becoming obsolete for computational purposes, these variations regain significance in the context of nanocomputing, given that nanoscale switching devices with mechanical contacts have attracted renewed interest [36,37].

Another method, *R-fold modular redundancy* [52], achieves tolerance to faults by employing R copies of a unit (R preferably an odd number), of which the majority establishes the output in accordance with an operation conducted by a majority gate. If $R = 3$, this technique is called *Triple Modular Redundancy* (TMR) [52]. When the outputs of three TMR units are combined by a majority gate on a second level and so on in a hierarchy of levels, the result is a model with increased reliability higher in the hierarchy: *Cascaded Triple Modular Redundancy* (CTMR) [186]. TMR, CTMR, and von Neumann's multiplexing technique lead to high redundancies, as compared to reconfiguration techniques [149], but on the other hand they are more suitable for transient faults. Von Neumann's original multiplexing technique is reconsidered in [172], with claims that its redundancy can be reduced by using output wire bundles as input bundles to other gates, rather than – as von Neumann advocated – reducing them to a single wire between the outputs bundles of a gate-equivalent unit and the inputs bundles of the next unit. Von Neumann's multiplexing technique is applied in a hierarchical way in combination with reconfiguration techniques in [78]. The claim is of a significantly reduced redundancy in this design, even when the device error rate is up to 0.01. The usefulness of Markov chains in the analysis of these models is reviewed in [79].

The constructions of von Neumann and others have some resemblances with error correcting codes, especially to so-called *repetition codes*, which are codes (but not very efficient ones at that) in which each symbol of a message is repeated a number of times to create redundancy. The use of error correcting codes in this context leads to a scheme in which computation takes place in the encoded space, whereby errors are corrected locally, and encoding and decoding is only necessary at the beginning and the end, respectively, of the computation. This line of thought is followed in [187], but then through the use of *generalized Reed–Solomon codes* (e. g. [207]), rather than repetition codes, to realize a fault-tolerant computing model with improved reliability.

Fault-tolerance in architectures based on cellular automata and on crossbar arrays are discussed in the sections describing these architectures, i. e. in Sects. "Cellular Automaton-Based Architectures" and "Crossbar Array-Based Architectures", respectively.

Cellular Automaton-Based Architectures

A Cellular Automaton (► Cellular Automata as Models of Parallel Computation) consist of cells that are organized in a regular array. Each cell contains a finite automaton that, taking as input the state of the cell and the states of its neighboring cells, changes its state according to a set of transition rules. These rules are designed such as to produce certain global behavior. When cellular automata are used as architectures for nanocomputers, this behavior is usually general-purpose computation, but application-specific behavior, like pattern recognition tasks, has also been investigated.

Originating in von Neumann's research on self-reproduction in the 1940's, cellular automata have experienced a resurgence in interest in recent years, especially in the context of nanocomputing architectures. Their simplicity – evidenced by a tiling-based structure in combination with local neighborhoods of cells with finite automata-like complexity – has much potential for bottom-up fabrication by molecular self-assembly. Moreover, their local interconnection structure implies a constant length of wires, which makes them efficiently scalable to high integration densities [11]. Cellular automata are unlikely to be used in combination with top-down fabrication methods, because of their significant overhead. Apart from the low percentage of cells that typically participate simultaneously in a computation, there may also be – depending on the design – relatively high hardware requirements for each cell, given its limited functionality as compared to conventional VLSI circuit designs. One of these factors may be improved, but this tends to be at the cost of efficiency with respect to the other factor: increasing the functionality of each cell, may increase the percentage of them that engage in a computation, since configurations of cells implementing certain tasks can be smaller, but on the other hand a cell will require more hardware resources to implement its functionality. A cellular automaton's overhead, which may exceed a factor of 10, could still be acceptable if it can be compensated for by the low-cost availability of cells in the huge quantities that can be expected with bottom-up fabrication. Even when only bottom-up fabrication methods are available, however, it is important to minimize the complexity of cells, because success in this area translates directly into more efficient physical realizations.

Though traditionally the complexity of a cellular automaton's cells is measured in terms of the number of cell states, this is inadequate for a nanocomputing framework, as it leaves unmentioned a cell's functionality in terms of the (number of) transition rules. The centralized storage of the table of transition rules, which is common in traditional cellular automata or software to simulate them, is infeasible for nanocomputers, because it violates one of the key tenets of cellular automata, i. e., locality. That leaves only storage of the rule table in each cell individually as a viable option, implying that this table should be as small as possible, less a huge overhead occurs. The alternative to storage of the rule table is finding an efficient representation through inherently available physical interactions implied by the physical structure of a cell, and this is ultimately what many researchers aim for when using cellular automata for nanocomputers. This alternative, however, imposes even stricter limitations on transition rules.

A cell's complexity is determined by the information required to represent its functionality, and this includes the cell's state as well as the transition rules or any equivalent physical structure or mechanism. There are various ways to represent this information. A representation requiring a decreased number of bits for storage of transition rules may require increased resources, both in time and hardware, to decode the information. This implies that it is important to consider the overall costs, including the required supporting circuitry, to be taken into account when evaluating a cell's complexity. Absent a concrete physical implementation, it tends to be easiest to settle for a cost measurement in terms of the number of bits to represent the cell states plus the number of bits to represent the transition rules, under the assumption that decoding of this information is done through a straightforward and simple scheme. For the cellular automaton in [117], for example, three bits are required to represent a cell's five states. The von Neumann neighborhood of a cell in this model requires each transition rule to encode in its left-hand side the states of a cell itself and the states of four neighboring cells, and in the right-hand side the new state of the cell (see Fig. 2). Given that there are 58 transition rules in this model, this results in a total of

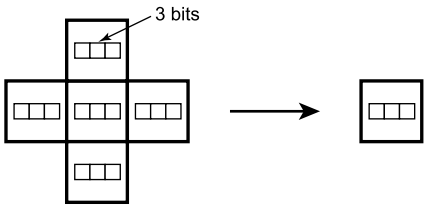

Nanocomputers, Figure 2
Bits required to encode a transition rule in a cellular automaton with 5-state cells and von Neumann neighborhood. The left-hand side of a transition rule requires three bits for the cell state itself and three bits each for the four neighboring cells' states. The right-hand side requires three bits to represent the new state of the cell after the transition

$3 + 58 \times 3 \times (1 + 4 + 1) = 957$ bits that are required to encode the functionality in a cell. A similar calculation yields 148 bits of information for the model in [157] and 100 bits of information for the model in [158]. A further reduction may be required to allow feasible physical realizations.

The use of cellular automata for nanocomputers originates in the late 1970's work of Fredkin, Toffoli, and Margolus, who aimed for computation inspired by physical models. A major issue in this research is the reduction of energy consumption as a road to high-performance high-integration computers. Signals in this paradigm have a discrete particle-like character and are conserved throughout operations. Moreover, a reversible scheme of logic (▶ Reversible Cellular Automata) is adopted.

One of the first proposals for a (2-dimensional) cellular automaton based on molecular interactions is due to Carter [28]. Written from a chemist's perspective, this proposal focuses on how particular operations of cells – including propagation in wires and switching by various automata – could be realized by means of the cascaded exchange of single and double bonds in molecules. This bond exchange – termed *soliton propagation in conjugated systems* – is relatively slow, but due to the small distances at which switches can be placed (in the order of 20 nm), switching times in the subnanosecond time scale are claimed to be feasible. Follow-up on this work is needed to show how the proposed operations could be combined into a cellular automaton capable of useful computations, but unfortunately the work has remained relatively unknown in the computer science community.

The Quantum Dot Cellular Automaton (QCA) [121, 164,194] has cells that consist of four quantum dots, two of which contain an electron each (Fig. 3a). Due to Coulomb interactions inside a cell, the electrons will settle along one of two possible polarizations, interpreted as signals 0 and 1, respectively. On a larger scale, there are also Coulomb interactions between cells through the electrons in them. These interactions can be used for the propagation of signals along cells (Fig. 3b and c). Logic operations are also possible, like the NOT-gate in Fig. 3d, or the majority gate in Fig. 3e, which covers both an AND gate and an OR gate in its functionality. The QCA model is not a cellular automaton in the true sense, since it assumes a layout of cells along the topology of the circuitry they represent. This may even include layouts in which a cell is placed with an offset of half a cell at the side of another cell [194]. QCA promise extremely low power consumption as their operation does not involve flows of electrons, but rather tunneling of electrons on a local scale, i. e., between quantum dots within a cell. The problem with QCA, however, is that they tend to be very sensitive to noise, imposing a low operating temperature on them, which is an obstacle to their efficient application. A supposedly more practical variation on QCA is the Magnetic QCA [39], in which the dots are implemented as submicron magnetic dots and which may allow operating temperatures as high as room temperature. There is considerable doubt among some researchers [24] whether QCA will ever provide a significant competitive advantage with respect to CMOS in terms of integration density, speed, power consumption, and the implementation of complex systems.

A 2-dimensional cellular automaton using optical signals to trigger transitions has been proposed by Biafore [20]. Employing physical interactions rather than explicit transition rules, this model has very simple partitioned cells that are structured such as to give control over both (a) which cells interact as neighbors and (b) when these interactions occur. Its design is based on Margolus' reversible cellular automaton [131] implementing Fredkin's billiard ball model (BBM) [65] (▶ Reversible Cellular Automata). The cells are suitable for implementation by quantum-dot devices that are organized such that electrical charge is transferred under the control of optical clock signals of different wave lengths supplied in a specific order. A similar way to control operations by electromagnetic or optical signals in a cellular automaton – though then applied to binary addition rather than universal computation – is described in [14]. Somewhat related is the cellular automaton model with less regular topologies in [15]. A 1-dimensional cellular automaton in which transitions are triggered by electromagnetic pulses of well-defined frequency and length is described in [126]. Apart from classical operations like the AND, OR, and NOT, quantum-mechanical operations are in principle possible by this scheme.

Cellular automata governed by nonlinear dynamics, so-called *Cellular Neural Networks* [35], hold another promise for nanocomputers, since the required physical implementations may be relatively simple. They are explored for this purpose in more detail in [101,217]. The logic states of such cellular automata can be expressed in terms of the electrical phases in a dynamic physical process. A possible implementation of such a model is by *tunneling phase logic* [217]. Tunneling phase logic still suffers from many problems [23], however, ranging from the difficulty to manufacture uniform tunneling diodes to the occurrence of stray background charges.

Implementations of cellular automata utilizing physical interactions between CO molecules arranged on a copper (Cu) surface are described in [58,85]. The atoms in the Cu surface form a triangular grid, on which the CO

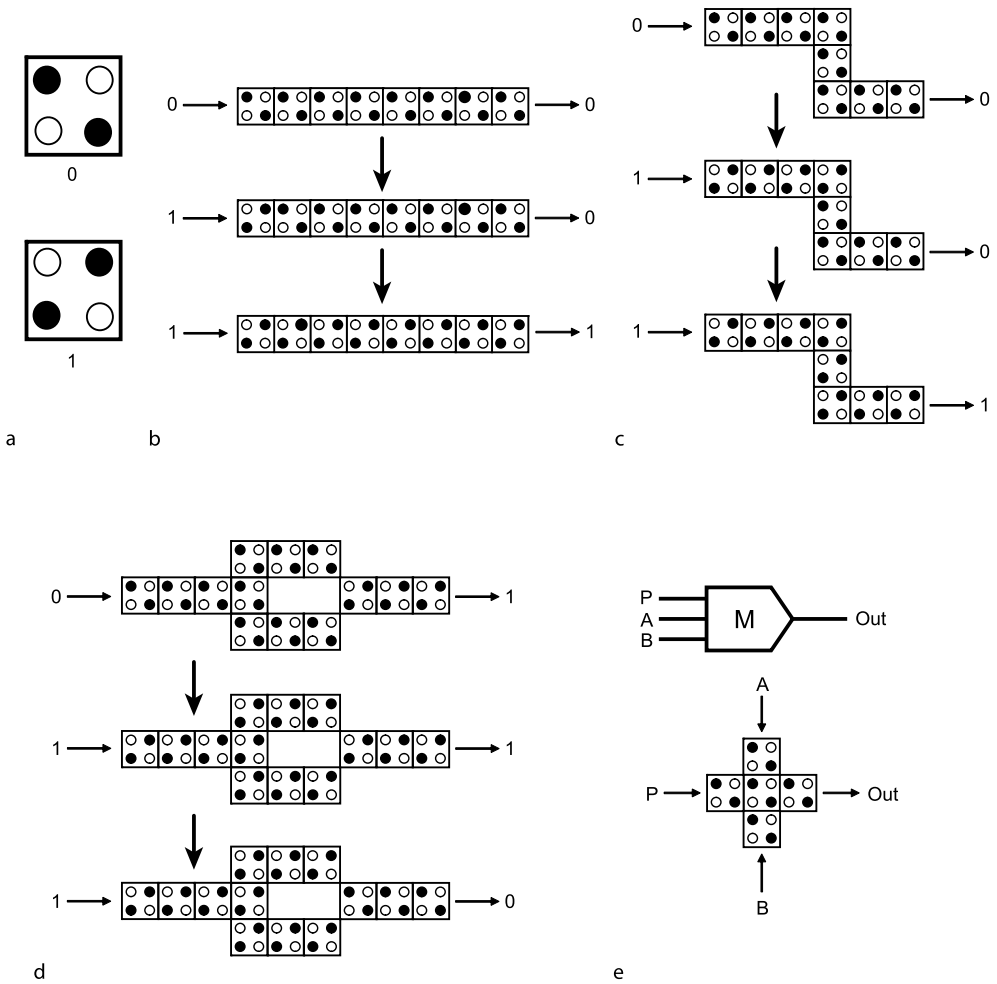

Nanocomputers, Figure 3

Quantum Dot Cellular Automaton (QCA). a A cell in the QCA contains four quantum dots storing two electrons. Since the electrons repel each other, they are likely to be in opposite corners of the cell. This allows for two possible cell configurations, which are interpreted as signals 0 and 1. b Due to the Coulomb interactions between electrons in quantum dots of neighboring cells, signals propagate along cells. A wire with cells in configuration 0 will change from the left to the right into a wire with cells in configuration 1 if the left-most cell is kept in configuration 1. c The same mechanism also works for wires taking left- and right-turns. d A NOT-gate formed from cells in the QCA. Change of input signal 0 to 1 results in the output signal changing from 1 to 0. e A majority-gate formed from cells in the QCA. From the majority-gate an AND-gate can be constructed by setting one of the inputs to the constant value 0. An OR-gate can be constructed by setting one input to 1

molecules settle, with the carbon atom of a molecule oriented toward a grid point. When two molecules are in nearest grid points, they will slightly point away from each other, due to a weak repulsive force between them. This force is sufficiently strong in some configurations of CO molecules to move certain CO molecules by one grid point, particularly in the so-called *Chevron-configuration* (Fig. 4a). Organized in suitable configurations, CO molecules will move on the grid in succession, triggering each other's hops, like domino stones falling in succession. These configurations, called *Molecule Cascades*, can be exploited to transmit signals on the grid (Fig. 4b), and to conduct Boolean AND operations on signals (Fig. 4c). More complex circuits can also be formed, and these will require configurations to cross wires, and in some cases OR-gates. As compared to the interactions based on the Chevron configuration, the interactions required for wire crossings and OR-gates are more complicated, and unfortunately less reliable. Based on a wide variety of configurations, a three-bit sorter is even possible, as shown in [85]. A formal analysis of the configurations possible by molecule cascades is given in [25]. Though molecule cas-

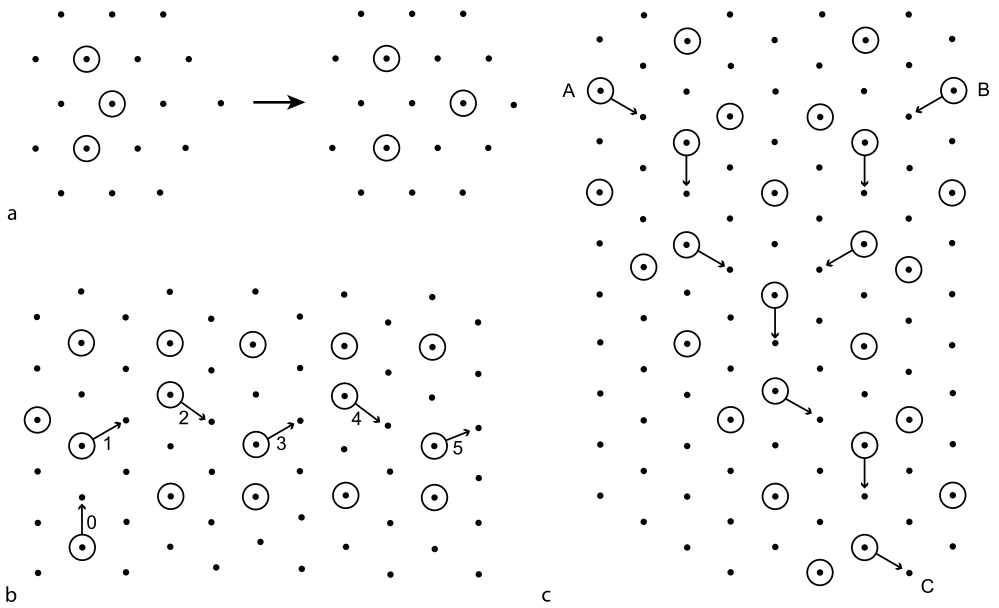

Nanocomputers, Figure 4
Molecule Cascades based on CO molecules (indicated by *circles*) placed on a grid of copper atoms (indicated by *dots*). a Chevron configuration, in which the center molecule moves one grid point away due to repulsive forces. b Wire based on the Chevron configuration. The arrows indicate the hops made by the CO molecules. c AND-gate based on the Chevron configuration ($C = A \wedge B$)

cades allow impressive integration densities – the three-bit sorter would require a area of merely 200 nm^2 – their speed of computation is very low: it takes about one hour for the three-bit sorter to complete its computation. There is no fundamental reason, though, why different systems of molecules would not be much faster. Setting up a molecular cascade system for computation involves a careful and time-consuming manipulation of individual molecules by a Scanning Tunneling Microscope (STM), so these types of systems are far from practical applications. A distinct disadvantage of the molecule cascades in [85] is that they allow only one-time computation: once a computation ends, the molecules have to be moved back to their initial positions to conduct the computation once more. An important step toward a practical system would be the development of a mechanism inherently present in the system to reinitialize molecules.

The cellular automata discussed above utilize the physical interactions inherent in their design to conduct transitions, an approach perhaps best illustrated in [132]. Though this tends to result in extremely simple cells, it may be challenging to build additional functionality into cells, such as functionality to configure cells into states necessary to initiate certain computations. This may be especially difficult if such a configuration process needs to take place on only part of the cell space. Likewise, error correction functionality is difficult to implement through simple physical interactions in a cellular space. For these reasons some researchers advocate cells that are more complex, but still sufficiently simple to potentially allow a huge number of them to be fabricated by bottom-up techniques and organized in arrays at nanometer scale integration densities.

One approach along these lines is the *Cell Matrix* [57]. Each cell in this model, consisting of a memory of less than 100 bytes and a few dozen gates, can be configured to conduct an operation like NAND, XOR, one-bit addition, etc.

The above cellular automaton models are all timed synchronously, requiring the delivery of a central clock signal to each cell. A synchronous mode of timing causes a wide array of problems as pointed out in Sect. "Heat Dissipation", like a high power consumption and heat dissipation. For this reason, cellular automata in which the principle of locality is carried one step further to the timing of the architecture, have started to attract interest in recent years. The resulting *asynchronous cellular automata* have a mode of operation in which all cells conduct their transitions at random times and independent of each other. Notwithstanding the advantage of asynchronous timing from a physical point of view, it brings up the question how to actually compute on such cellular automata in a deterministic way. The usual method to do this is to simulate a timing mechanism on an asynchronous cellular automaton to force the cells approximately into lock-step,

and then using well-established methods to compute synchronously. Unfortunately, such an approach causes a significant overhead in terms of the number of cell states and transition rules. To simulate an n-state synchronous cellular automaton on its asynchronous counterpart, one needs, depending on the method, a number of states of up to $O(n^2)$ [157]. Worse yet, to synchronize different parts of an asynchronous cellular automaton with each other – necessary, because in the end it is a synchronous cellular automaton that is being simulated – exchange of signals between these parts is required. This exchange takes the form of so-called *synchronization waves* that propagate along the cell space [157]. Since all cells need to continuously change their states to let these waves pass, there are many transitions being conducted that do not contribute to the computation, even in areas of the cell space where no signals or active configurations are present. Implemented physically, such asynchronous cellular automata would thus need to consume much energy for dummy transitions, making them hardly better candidates for nanocomputer architectures than synchronous cellular automata.

A more efficient approach to conduct computations on asynchronous cellular automata is to conduct asynchronous computations directly, using only synchronization between cells on strictly local scales. This approach is based on embedding a circuit on the cellular space and then simulating the circuit's operation through the cooperative transitions of the cells. While the circuits in synchronous models are typically standard logic circuits involving AND, OR, and NOT gates, different functionalities are required for asynchronous circuits, but the general idea of embedding circuits on the cell space remains the same. The approach in [1,117,119,157,158] is to use delay-insensitive circuits (see Sect. "Heat Dissipation"), which, due to their robustness to signal delays, combine very well with asynchronous cellular automata, since the requirement no longer holds that signals must arrive at certain locations at certain times dictated by a central clock, as in synchronous circuits. This takes away concerns about variations in the operational speed of cells and considerably simplifies the design of configurations representing circuit elements embedded in the cellular space.

To avoid the randomness in the timing of transitions interfering with proper deterministic operation of an asynchronous cellular automaton, its design involves transition rules that serialize transitions into well-controlled sequences of cell states. Some asynchronous cellular automata designed according to these principles can be found in [1,117,119]. Unfortunately, many of these models still require tens of transition rules, which may hinder

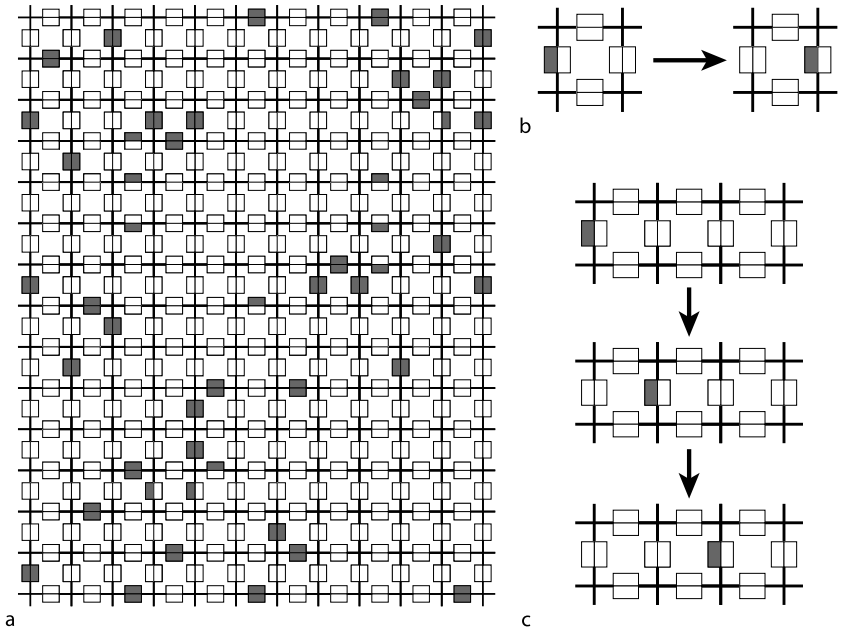

Nanocomputers, Figure 5

A Self-Timed Cellular Automaton (STCA) consists of cells with four partitions. All four partitions as well as one partition of each of the four neighboring cells are rewritten by transition rules. a Example of a small configuration of cells with partitions having values 0 (*white*) and 1 (*black*). b Transition rule for signal propagation. c Applying this rule to the configuration at the left results in a 1-partition moving to the right

efficient physical implementations. To this end, alternative models have been proposed, called *Self-Timed Cellular Automata (STCA)* [157], in which cells are subdivided in four partitions, each representing a part of a cell's state information (see Fig. 5a). This results in a substantial reduction of the number of transition rules. The model in [158] employs merely six rules. One of these rules, a rule for signal propagation is shown in Fig. 5b, and its effect on a cell configuration of a signal is shown in Fig. 5c.

In order to compute on a cellular automaton, it is necessary to configure it for the computation. One way to do this is by having a separate mechanism for (re-) configuration, through a separate layer of circuitry to address each individual cell like in a memory, and writing the required state into it. Another way is to have the reconfiguration done by the cells themselves. This requires the transition rules to contain sufficient functionality to represent the reconfiguration mechanism. Such functionality resembles self-reproduction of configurations in cellular automata. In other words, such cellular automata require *construction universality* (► Self-Replication and Cellular Automata). Unfortunately, it turns out that implementing construction universality on a cellular automaton tends to give a large overhead in terms of the number of transition rules. A self-contained mechanism for reconfiguration may thus not be practical, absent a novel ingenious way to implement construction universality, probably inspired by some biological mechanisms.

Reconfiguration in cellular automata can also be used to achieve defect-tolerance. In the approach in [90], defects are detected and isolated from non-defective cells through waves of cells in certain states propagating over the cell space, leaving defective cells, which are unable to adhere to the state changes, standing out to be detected as defective [90]. This approach is taken one step further in [93] by additionally scanning the cell-space for defect-free areas on which circuits can be configured. Being self-contained, this scanning process uses reconfiguration mechanisms resembling the self-reproduction functionality on cellular automata. An online approach to defect-tolerance is followed in [92], in which configurations called *random flies* float around randomly in the cellular space, and stick to configurations that are static. A configuration unable to compute due to defects is static in this model and will thus be isolated by a layer of random flies stuck to it. Being highly experimental, these approaches require an overhead in terms of the number of transition rules that may be too high to be of practical significance for nanocomputers.

Approaches to fault-tolerance in cellular automaton-based nanocomputers tend to focus on economic ways to implement redundancy. Early work in this line of research is reported in [150]. This model can correct at most one error in 19 cells, but to this end each cell requires a neighborhood of 49 cells in its transition rules, which is much higher than usual cellular automata. The increased complexity of cells suggests that they will be very error-prone in physical implementations, limiting the practical value of this work. The model in [80] suffers from similar problems.

Better fault-tolerance is obtained in [67,68,70] with synchronous cellular automata, and in [69,204] with asynchronous cellular automata simulating synchronous cellular automata: the idea is to organize cells in blocks that perform a fault-tolerant simulation of a second cellular automaton, which on its turn is also organized in blocks, simulating even more reliably a third cellular automaton, and so on. This results in a hierarchical structure with high reliability at the higher levels, like with the CTMR technique mentioned in Sect. "Fault-Tolerance". The cells in these models, however, are too complex to be suitable for efficient physical implementations, since they contain information regarding the hierarchical organization, such as block structure, address within a block, programs selecting transition rules, and timing information. In [91] a fault-tolerant STCA based on *BCH codes* (e. g. [207]) is proposed, but computation on this model is inefficient, since only one signal at a time is allowed.

Spielman's work [187] on implementing a computation scheme in an encoded space based on an error correcting code, mentioned in Sect. "Fault-Tolerance", has resulted in related work [158] in the context of STCAs. Each partition of an STCA's cell and each adjacent partition of a neighboring cell is stored in a memory and encoded by an error correcting code. Up to one third of the memory's bits can – if corrupted – be corrected by this method.

Cellular automata in which cells have the complexities of processing elements [62,202], rather than (much simpler) finite automata, form the other end of the spectrum of cellular nanocomputer architectures. Since the fabrication of even simple processing elements are likely to be beyond the reach of bottom-up techniques, this approach may be infeasible for the same reasons as to why alternatives to von Neumann architectures are proposed in the first place for nanocomputers.

In conclusion, the main challenges for cellular automaton-based nanocomputers are:

- the design of models requiring a very small number of transition rules i. e., in the order of less than 10,
- the design or discovery of a mechanism to implement transition rules in a physically feasible way,

- to implement reconfiguration ability, while keeping the number of transition rules small,
- to implement fault- and defect-tolerance while keeping the number of transition rules and cell states small.

Crossbar Array-Based Architectures

A crossbar consists of a set of, say, N horizontal wires that cross another set of M vertical wires. The $N \times M$ cross points of the wires form a matrix of switches, which can be individually set (programmed) in a low-resistance state (closed) or a high-resistance state (opened). The simple structure of two planes of parallel wires allows for a wide variety of bottom-up fabrication techniques, such as self-assembly by fluidics-based alignment [87], directed self-assembly based on chemical techniques in combination with an alternating current (AC) electric field [53], or fabrication including top-down techniques like nanoimprinting [99,214]. Alignment of wires in order to connect them – a major problem in nanoelectronics – is relatively straightforward for crossbar arrays, due to the wires being perpendicular. Combined with a significant flexibility in making and altering connections, this accounts for much of the attention for crossbar arrays in the context of nanocomputers. Nanowires with diameters of only a few nanometers that can be used in implementations of crossbar arrays have already been fabricated, like single-crystal nanowires [41,87,100,144] and carbon nanotube wires [184]. The crossbar in [171] employs carbon nanotubes as its wires that can be programmed by applying a voltage across a crosspoint of two wires. This voltage, which is higher than that for normal (non-programming) operation of the crossbar, results in the wires being bent toward each other to close the switch. After removing the voltage, these deformations are preserved due to van der Waals forces, so memory functionality is inherent to this mechanism. To open the switch again, an opposite voltage is applied, which drives the wires apart. Nanometer scale implementations of crossbar arrays like the above provide significant efficiency as compared to their CMOS-based counterparts, the latter requiring a separate 1-bit memory for each cross point to store the state of the corresponding switch.

Another nanoscale crossbar design places a very thin layer of rotaxane molecules between the two wire planes in the crossbar [32]. Molecules in this layer at crosspoints can be set into low-resistance or high-resistance states – again through applying a voltage of sufficient strength – and these states tend to remain unchanged throughout successive read operations.

The crossbar is used in the following contexts:

Routing When used as a interconnection array, the crossbar array allows the routing of any input from a (single) vertical wire to any output on a (single) horizontal wire, by closing the switch between these wires. The resulting flexibility in connectivity provides an efficient way to route around defects.

Logic Logic when implemented by crossbar arrays usually takes the form of wired (N)OR-planes or (N)AND-planes [47,109,182,183], whereby wires in one plane provide inputs and wires in the other plane provide outputs.

Memory The 1-bit memories at the crosspoints offer the potential of huge storage capacities when combined in large crossbar arrays [32,77,107,214]. Reading out a memory bit is usually done by applying a voltage to one of the corresponding wires, resulting in the state of the memory appearing on the other wire. Realizations as associative memories have also been considered, whereby information storage is distributed over the switches such that recovery is possible even if some switches fail [45,179].

Address Decoding The connections between the two wire planes can also be set such that a demultiplexer is obtained, which can address a single wire out of the N wires in one plane by much less than N wires in the other plane, e. g. down to $O(\sqrt{N})$ or even $O(\log N)$ wires [32,47,208,219]. Apart from use in combination with crossbar arrays for routing, logic, or memory, a demultiplexer is also an excellent interface between micro-scales on one hand and nano-scales on the other, allowing for a limited number of microwires in one plane of a crossbar array to address a much larger number of much thinner nanowires in the other plane. Such demultiplexers tend to be static, i. e., they do not require reprogramming of the switch settings. When used as micro/nano-interfaces, reprogramming may even be impossible as nanowires can not be accessed directly; prepatterned connections [47] or stochastically assembled connections [208] may be suitable options in this case.

The cross points in crossbar arrays can take various forms, the most common of which are resistors, diodes – both combined with a switch in each cross point – and Field Effect Transistors (FETs). Resistor crossbars, followed by diode crossbars, are easiest to fabricate by bottom-up techniques, since both are 2-terminal devices, which significantly simplifies alignment problems. Resistor crossbars, additionally, can be built using metal nanowires, facilitat-

ing low output impedances, but their linear device characteristics may complicate the approximation of nonlinear functionalities [180]. Though diode and FET crossbars will do better with regard to the latter point, they carry significant disadvantages with existing fabrication technology [180]: limited current density, forward-biased voltage drops, relatively high impedance of semiconductor nanowires, difficulty of configuration, etc. Resistor and diode crossbars, on the other hand, come with another disadvantage [189]: being 2-terminal devices, they are passive, thus lacking signal gain – a major disadvantage as compared to 3-terminal devices like transistors – complicating the restoration of signals. Moreover, 2-terminal devices tend to have higher power consumption than devices like transistors, due to the inherently higher currents flowing to ground [189].

The lack of signal restoration is especially problematic when many levels of logic are required through which signals pass. Minimizing this number of levels, however, causes an inefficient use of hardware: it prevents the exploitation of the logic structure of a circuit design, since each level – such as an OR-plane or an AND-plane in a crossbar array – has a homogeneous structure. This is a well-known problem in Programmable Logic Arrays (PLAs) [49]: an n-input XOR, for example, requires a number of product terms that is exponential in n when expressed by the 2-level logic of a single PLA. It is thus unsurprising that signal restoration in crossbar arrays has received significant attention.

Several proposals have been made to include signal restoration functionality in crossbar arrays. DeHon [50] realizes some of the cross points between two wires as FETs, such that the electric field of one wire controls the conductivity of the other wire. This requires doping of (parts of) a wire, but the problems related to that, such as alignment of doped parts of wires with crossing wires, may require further research. Kuekes et al. [110] propose to use a bistable-switch latch to restore signals, in which the logic value of a wire is controlled by two switches, one connecting to a logic 0 and one to a logic 1, such that only one switch is closed at a time. Goldstein [75] also uses a latch for signal restoration, though in this case the latch is composed of a wire with two Negative Differential Resistance molecules at either end. There are also schemes in which the inherent poor voltage margins of resistor-based crossbar logic is improved by including error-correcting codes in the scheme [109,111].

Within the class of crossbar arrays for memory applications, there is a pronounced difference in terms of storage capacity between crossbar arrays based on diodes and FETs on the one hand, and crossbar arrays based on resistors on the other hand. The use of diodes or FETs ensures that each cross point is independent of the other cross points, since electrical current can only flow in one direction in these designs. In resistor-based crossbar arrays, on the other hand, there will often be an indirect (parasitic) path through which a current flows when wires are connected at multiple cross points to perpendicular wires. This has a profound impact on the storage capacity of memories based on crossbar arrays. Whereas an $N \times N$ crossbar array will have an $O(N^2)$ storage capacity when based on diodes or FETs, it will have only $O(N \log N)$ storage capacity when based on resistors (for more detailed estimates of storage capacities see [185]).

Though a standalone crossbar array is insufficient as an architecture for nanocomputers, it can be applied at different places and at different levels in a hierarchy to form a more or less "complete" architecture.

Goldstein's NanoFabrics [75], for example, consist of so-called nanoBlocks and switch-blocks organized in tiles called clusters that are connected to each other through long nanowires of varying lengths (e. g. 1, 2, 4, and 8 clusters long). These wires allow signals to traverse a few clusters without the need to go through switches. The nanoBlocks themselves consist of logic based on crossbar arrays, and the switch-blocks consist of interconnections based on crossbar arrays. Defect-tolerance is obtained by testing the circuits – like in the Teramac, mentioned in Sect. "Fault-Tolerance" – but unlike the Teramac, testing is conducted in incremental stages in which an increasing part of the hardware – once it has been configured around defects – is used in the testing of the remaining parts [140]. Testing the initial part is conducted by a host.

DeHon's Nanowire-Based Programmable Architecture [49] provides designs for Wired-OR logic, interconnection, and signal restoration implemented by atomic-scale nanowire crossbar arrays.

The CMOL architecture [125,190] is based on a reconfigurable crossbar array of nanodevices that is superimposed at a small angle on a grid of CMOS cells. It is used to implement memory, FPGA logic, or neural network inspired models (see Sect. "Neural Network-Based Architectures"). Each CMOS cell consists of an invertor and two pass transistors serving two pins that connect to the crossbar array. Snider et al.'s Field-Programmable Nanowire Interconnect (FPNI) [181] improves on this proposal by adjusting the crossbar array's structure so that all CMOS-fabricated pins can be of the same height. Logic functionality in the architecture is restricted to the CMOS part and routing to the crossbar of nanowires.

Reconfigurability is one of the strong points of crossbar arrays, but it requires a time-intensive process of test-

ing for defects to guarantee defect-free routes. In effect, manufacturing yield is traded for an expensive die per die multi-step functional test and programming process [24]. A combination with error-correcting codes, like in [108], is thought to be more economical by some authors [7]. It is unclear as to what extent reconfiguration around defects can be made self-contained in crossbar-array-based architectures.

Neural Network-Based Architectures

A *Neural Network* is a collection of threshold elements, called neurons, of which the human brain contains about 10^{11}, each connected to 10^3 to 10^4 other neurons. Neurons receive inputs from each other through weighted connections. The strengths (weights) of the connections in a neural network are changed in accordance with the input to it; this process is called *learning*. Learning is usually accomplished through a *Hebbian learning rule*, which was first discovered by Hebb to be of fundamental importance in real brains. Hebbian learning can be shortly described as "neurons that fire together, wire together". In other words, two neurons tend to strengthen the connection between them when their patterns of fired pulses correlate positively with each other. In artificial neural networks this rule is often implemented in terms of the correlation of the time-averaged analog values of neurons, rather than the actual pulses. Hebbian learning lies at the basis of an impressive self-organizing ability of neural networks. The model has much flexibility with regard to the connectivity of the network between the neurons, as a result of which little detailed information is required about the underlying network structure to conduct learning processes. This tends to result in a good tolerance to defects and faults, which is an important reason for the interest in neural networks for nanocomputer architectures.

Neural networks tend to be especially useful to problems that involve pattern recognition, classification, associative memory, and control. Recognition of visual images, for example, is conducted by brains to an extent that is unsurpassed by traditional computers, in terms of speed (taking approximately 100 ms), as well as in terms of fidelity. Though individual neurons are slow – it takes about 10 ms for a neural signal to propagate from one neuron to the other – the system as a whole is very fast. Contrast that to the time it takes computers with clock speeds of many GHz to performing the same task, which tends to be in the order of minutes to hours.

Most studied applications of artificial neural networks have difficulty to go beyond the toy-problem level. A better knowledge about the brain may lead to better results, and there is also the expectation [125] that the availability of huge amounts of hardware resources may be of help. It is important to realize, however, that the large-scale organization of the brain is directed not only by relatively easy-to-characterize learning algorithms, but also by genetic factors, which involve a certain degree of "wetness". The latter is not very compatible with the usual perceptions about computers. It is questionable whether the abundance of hardware resources alone will result in a breakthrough in neural networks-based architectures.

Neuromorphic Engineering is the term used to describe the interdisciplinary field that takes inspiration from neuroscience, computer science and engineering to design artificial neural systems. Though initially associated with VLSI systems engineering – due to Mead [135] coining the term *neuromorphic* – it is increasingly attracting contributions related to nanocomputers. One of the early papers in this context is [170], in which quantum dots arranged in a two-dimensional array and deposited on a resonant tunneling diode (RTD) are connected by nanowires, to form nodal points of nonlinear bistable elements in a conductive network. The dynamics of this system bears strong similarities to the equations representing associative memories.

Another proposal related to neuromorphic nanoarchitectures is the *Nanocell*. A Nanocell is an element containing randomly ordered metallic islands that are interlinked with molecules – acting as switches – between micrometer-sized metallic input/output leads [196]. To impose certain functionalities on the nanocell, like an AND, a NOR, an Adder, etc., it is necessary for these switches to be set into certain configurations. Genetic Algorithms (GA) are used for this in [195], but this method requires a detailed knowledge of the random structure inside the nanocell, so it is impractical, resulting in proposals to use neural networks, in which – ideally – the switches can be programmed through a learning process employing electrical pulses on input and output terminals of the nanocell [88]. To create useful circuits from nanocells it is necessary to connect them – after programming – into certain circuits. Creating such inter-cell connectivity has not been researched in detail, but neuromorphic learning methods may be suitable in this context as well.

Boltzmann machine neural networks based on single electron tunneling devices are proposed in [215]. *Simulated annealing*, the processing in Boltzmann machines through which a minimum energy state is found in an optimization problem, is implemented through a digital oscillator based on the stochastic nature of the tunneling phenomenon. The architecture consists of a network in which outputs of neural elements are fed back to their in-

puts through weights in a interconnection network, which are implemented by capacitance values reflecting the particular problem to be optimized.

CrossNets [197] are neuromorphic architectures based on the CMOL architecture. Neurons in this model, being relatively sparse, are represented by CMOS logic, and the neural interconnections, being very dense, are represented by a crossbar array of nanometer scale wires, whereby a cross-connection between two wires represents a so-called *synapse*. The crossbar array offers a very flexible framework to connect neurons to each other: it allows connectivity ranging from hierarchical structures that are typically used in multi-layered perceptrons [83] to more homogeneous networks with quasi-local connections. CrossNets can be used for pattern classification in feedforward multilayered perceptrons as well as for pattern restoration in associative memories. The crossbar array, though very flexible, has also disadvantages in this context, because it limits interconnection weights to be binary, and it is hard to avoid interference between different switches when setting or updating weights. Some solutions to these problems may be reached by using multiple-wire cross connections per weight [197]. The above models employ neural signals that have analogous values, as opposed to the spiking signals that are common in real brains. Spikes offer the opportunity for decreased power consumption, and, more importantly, their timing appears to encode important information in brains, and is thought to allow the semantic binding of different brain areas with each other. Some preliminary results on spiking neuromorphic architectures based on CMOL CrossNets and comparisons to non-spiking models are given in [71]. A different neuromorphic architecture based on crossbar arrays is explored in [45] for use in associative memories.

Future Directions

In pondering the future directions nanocomputers may take, it is important to realize that CMOS technology is a formidable competitor. Detailed predictions on the pace at which this technology progresses have been made for decades, with varying degrees of success. When the emphasis is on technological limitations, the resulting predictions tend to be overly conservative, as they do not take into account the many ways often found around those limitations [86]. When on the other hand trends are simply extrapolated from the past, predictions tend to be on the optimistic side, as they ignore hard physical limitations that might still be ahead [86]. Notwithstanding the uncertainties of the life-time of CMOS technology, there seems to be a consensus that it can be extended to at least the year 2015. The exponential growth experienced over so many decades, however, may significantly slow down towards the end [143]. Beyond this, it is expected [23] that the benefits of increased integration densities will be extended for 30 more years through the use of emergent technologies. These technologies are likely to be used in combination with silicon technologies and gradually replace more of the latter.

Many of the limitations that are perceived in technology – and CMOS is no exception to this – are likely to be circumvented by reconsidering the underlying assumptions, as was observed by Feynman [60]. This extends to the framework of circuits and architectures as well. Reconsiderations of underlying assumptions has frequently resulted in paradigm shifts in computer science. A notable example form *Reduced Instruction Set Computers (RISC)* [73], which use a small set of rapidly executed instructions rather than a more expanded and slower set of instructions as was common until the 1980's in Complex Instruction Set Computers (CISC). The CISC approach was advantageous in the context of limited sizes of memories, necessitating a dense storage of instructions at the cost of increased processing in the processor itself. That changed with the increasing sizes and decreasing costs of memories, coupled with the realization that only a small percentage of the instructions available in CISC processors were actually used in practice. In a similar way, a changing context will determine the direction in which architectures of nanocomputers develop. This chapter has identified a number of factors of importance in this respect, like the growing problems faced in the fabrication of arbitrary structures, the increasing occurrence of defects in those structures, the increasing occurrence of faults during computations, and the increasing problems with heat dissipation. These problems will have to be addressed one way or the other, and architectural solutions are likely to play an important role in them; regular or random structures in combination with (re)configuration techniques seem especially attractive in this context.

A gradual move away from the traditional von Neumann architecture toward the increasing use of parallelism is likely to occur. Initially it will be conventional processors that will be placed in parallel configurations – a trend that has already started in the first decade of the 21st century. As feature sizes decrease further, fabrication technology will be more and more bottom-up, resulting in parallelism becoming increasingly fine-grained, to the point that computers consist of a huge number of elements with extremely simple logic functionality that can be configured according to specifications by a programmer. In the end,

architectures may look more like intelligent memory than like the computers known today. Programming methods, though apt to change as architectures change, may be relatively unaffected by the trend toward finely grained parallelism, witness the resemblance of conventional programming languages to languages for fine-grained hardware platforms like FPGAs, some of which have been designed with resemblance to the programming language C in mind.

Trends in technology are often highly influenced by breakthroughs, and it is entirely possible that future nanocomputers develop in different directions. The increasing research efforts into unconventional computing offers a clue to the wide range of what is possible in this regard. Computation models based on new media like reaction-diffusion systems, solitons, and liquid crystals [2] are just a few of the fascinating examples that may some day work their way into practical computer architectures.

Bibliography

1. Adachi S, Peper F, Lee J (2004) Computation by asynchronously updating cellular automata. J Stat Phys 114(1/2): 261–289
2. Adamatzky A (2002) New media for collision-based computing. In: Collision-Based Computing. Springer, London, pp 411–442
3. Adleman LM (1994) Molecular computation of solutions to combinatorial problems. Science 266(11):1021–1024
4. Appenzeller J, Joselevich E, Hönlein W (2003) Carbon nanotubes for data processing. In: Nanoelectronics and Information Technology. Wiley, Berlin, pp 473–499
5. Athas WC, Svensson LJ, Koller JG, Tzartzanis N, Chou EYC (1994) Low-power digital systems based on adiabatic-switching principles. IEEE Trans Very Large Scale Integr Syst 2(4):398–407
6. Aviram A, Ratner MA (1974) Molecular rectifiers. Chem Phys Lett 29(2):277–283
7. Bahar RI, Hammerstrom D, Harlow J, Joyner WH Jr, Lau C, Marculescu D, Orailoglu A, Pedram M (2007) Architectures for silicon nanoelectronics and beyond. Computer 40(1):25–33
8. Ball P (2006) Champing at the bits. Nature 440(7083):398–401
9. Banu M, Prodanov V (2007) Ultimate VLSI clocking using passive serial distribution. In: Future Trends in Microelectronics: Up the Nano Creek. Wiley, Hoboken, pp 259–276
10. Bashirullah R, Liu W (2002) Raised cosine approximation signalling technique for reduced simultaneous switching noise. Electron Lett 38(21):1256–1258
11. Beckett P, Jennings A (2002) Towards nanocomputer architecture. In: Lai F, Morris J (eds) Proc. 7th Asia-Pacific Computer Systems Architecture Conf. ACSAC'2002 (Conf. on Research and Practice in Information Technology), vol 6. Australian Computer Society, Darlinghurst, Australia
12. Benioff P (1980) The computer as a physical system: A microscopic quantum mechanical Hamiltonian model of computers as represented by Turing machines. J Stat Phys 22(5): 563–591
13. Benioff P (1984) Comment on: Dissipation in computation. Phys Rev Lett 53(12):1203
14. Benjamin SC, Johnson NF (1997) A possible nanometer-scale computing device based on an adding cellular automaton. Appl Phys Lett 70(17):2321–2323
15. Benjamin SC, Johnson NF (1999) Cellular structures for computation in the quantum regime. Phys Rev A 60(6):4334–4337
16. Bennett CH (1973) Logical reversibility of computation. IBM J Res Dev 17(6):525–532
17. Bennett CH (1982) The thermodynamics of computation – a review. Int J Theor Phys 21(12):905–940
18. Bennett CH (1984) Thermodynamically reversible computation. Phys Rev Lett 53(12):1202
19. Bennett CH (1988) Notes on the history of reversible computation. IBM J Res Dev 32(1):16–23
20. Biafore M (1994) Cellular automata for nanometer-scale computation. Physica D 70:415–433
21. Birge RR, Lawrence AF, Tallent JR (1991) Quantum effects, thermal statistics and reliability of nanoscale molecular and semiconductor devices. Nanotechnology 2(2):73–87
22. Bohr MT, Chau RS, Ghani T, Mistry K (2007) The high κ solution. IEEE Spectr 44(10):23–29
23. Bourianoff G (2003) The future of nanocomputing. Computer 36(8):44–53
24. Brillouët M (2007) Physical Limits of Silicon CMOS: Real Showstopper or Wrong Problem? In: Future Trends in Microelectronics; Up the Nano Creek. Wiley, Hoboken, pp 179–191
25. Carmona J, Cortadella J, Takada Y, Peper F (2006) From molecular interactions to gates: a systematic approach. In: ICCAD '06: Proceedings of the 2006 IEEE/ACM international conference on Computer-aided design, San Jose, 5–9 Nov 2008
26. Carter FL (1983) The chemistry in future molecular computers. In: Computer Applications in Chemistry, Proc. 6th Int. Conf. on Computers in Chemical Research and Education. Elsevier, Amsterdam, pp 225–262
27. Carter FL (1983) Molecular level fabrication techniques and molecular electronic devices. J Vac Sci Technolog B 1(4):959–968
28. Carter FL (1984) The molecular device computer: point of departure for large scale cellular automata. Physica D 10 (1–2):175–194
29. Cavin RK, Zhirnov VV, Hutchby JA, Bourianoff GI (2005) Energy barriers, demons, and minimum energy operation of electronic devices. In: Proc. SPIE, vol 5844, pp 1–9
30. Ceruzzi P (1998) A history of modern computing. MIT Press, Cambridge
31. Chan SC, Shepard KL, Restle PJ (2005) Uniform-phase uniform-amplitude resonant-load global clock distributions. IEEE J Solid-State Circuits 40(1):102–109
32. Chen Y, Jung GY, Ohlberg DAA, Li X, Steward DR, Jeppesen JO, Nielsen KA, Stoddard JF, Williams RS (2003) Nanoscale molecular-switch crossbar circuits. Nanotechnology 14(4):462–468
33. Choi H, Mody C (2007) Molecular electronics in the *longue durée*: the microelectronics origins of nanotechnology. In: Joint Wharton-Chemical Heritage Foundation Symposium on the Social Studies of Nanotechnology, Philadelphia, 7–8 Jun 2007
34. Chou SY, Krauss PR, Renstrom PJ (1996) Imprint lithography with 25-nanometer resolution. Science 272(5258):85–87
35. Chua LO, Yang L (1988) Cellular neural networks: theory. Circuit Syst IEEE Trans 35(10):1257–1272

36. Collier CP, Wong EW, Belohradský M, Raymo FM, Stoddart JF, Kuekes PJ, Williams RS, Heath JR (1999) Electronically configurable molecular-based logic gates. Science 285(5426):391–394
37. Collier CP, Mattersteig G, Wong EW, Luo Y, Beverly K, Sampaio J, Raymo FM, Stoddart JF, Heath JR (2000) A [2]Catenane-based solid state electronically reconfigurable switch. Science 289(5482):1172–1175
38. Constantinescu C (2007) Impact of intermittent faults on nanocomputing devices. In: Workshop on Dependable and Secure Nanocomputing. Edinburgh, 28 Jun 2007
39. Cowburn RP, Welland ME (2000) Room temperature magnetic quantum cellular automata. Science 287(5457):1466–1468
40. Cui Y, Lieber CM (2001) Functional nanoscale electronic devices assembled using silicon nanowire building blocks. Science 291(5505):851–853
41. Cui Y, Lieber C, Lauhon L, Gudiksen M, Wang J (2001) Diameter-controlled synthesis of single crystal silicon nanowires. Appl Phys Lett 78(15):2214–2216
42. Dasmahapatra S, Werner J, Zauner KP (2006) Noise as a computational resource. Int J Unconv Comput 2(4):305–319
43. Davari B (1999) CMOS technology: present and future. In: Proc. IEEE Symp. on VLSI circuits. Digest of Technical Papers, pp 5–9
44. Davis A, Nowick SM (1997) An introduction to asynchronous circuit design. Tech Rep UUCS-97-013, Computer Science Department, University of Utah
45. Davis BA, Principe JC, Fortes JAB (2004) Design and performance analysis of a novel nanoscale associative memory. In: Proceedings of 4th IEEE Conference on Nanotechnology, pp 314–316
46. Debray P, Raichev OE, Rahman M, Akis R, Mitchel WC (1999) Ballistic transport of electrons in T-shaped quantum waveguides. Appl Phys Lett 74(5):768–770
47. DeHon A (2003) Array-based architecture for FET-based nanoscale electronics. IEEE Trans Nanotechnol 2(1):23–32
48. DeHon A (2004) Law of large numbers system design. In: Nano, quantum and molecular computing: implications to high level design and validation. Kluwer, Norwell, pp 213–241
49. DeHon A (2005) Nanowire-based programmable architectures. ACM J Emerg Technol Comput Syst 1(2):109–162
50. DeHon A, Lincoln P, Savage JE (2003) Stochastic assembly of sublithographic nanoscale interfaces. IEEE Trans Nanotechnol 2(3):165–174
51. Dennard RH, Gaensslen FH, Yu HN, Rideout VL, Bassous E, LeBlanc AR (1974) Design of ion-implanted mosfets with very small physical dimensions. IEEE J Solid-State Circ 9(5):256–268
52. Depledge PG (1981) Fault-tolerant computer systems. IEE Proceedings A 128(4):257–272
53. Diehl MR, Yaliraki SN, Beckman RA, Barahona M, Heath JR (2002) Self-assembled deterministic carbon nanotube wiring networks. Angewandte Chem Int Ed 41(2):353–356
54. Dobrushin RL, Ortyukov SI (1977) Upper bound for the redundancy of self-correcting arrangements of unreliable functional elements. Probl Inform Transm 13(3):203–218
55. Drexler KE (1986) Engines of creation. Anchor Books, New York
56. Drexler KE (1992) Nanosystems: molecular machinery, manufacturing, and computation. Wiley, New York
57. Durbeck LJK, Macias NJ (2001) The cell matrix: an architecture for nanocomputing. Nanotechnology 12(3):217–230
58. Eigler DM, Lutz CP, Crommie MF, Mahoran HC, Heinrich AJ, Gupta JA (2004) Information transport and computation in nanometer-scale structures. Phil Trans R Soc Lond A 362(1819):1135–1147
59. Feynman RP (1985) Quantum mechanical computers. Optics News 11:11–20
60. Feynman RP (1992) There's plenty of room at the bottom (reprint of 1959 lecture). J Microelectromech Syst 1(1):60–66
61. Feynman RP, Leighton R, Sands M (2006) Ratchet and pawl. In: The Feynman Lectures on Physics, vol 1. Addison Wesley, San Francisco, pp 1–9
62. Fountain TJ, Duff MJB, Crawley DG, Tomlinson CD, Moffat CD (1998) The use of nanoelectronic devices in highly parallel computing systems. IEEE Trans VLSI Syst 6(1):31–38
63. Frank MP (2005) Introduction to reversible computing: motivation, progress, and challenges. In: CF '05: Proceedings of the 2nd conference on Computing frontiers. ACM Press, New York, pp 385–390
64. Frazier G, Taddiken A, Seabaugh A, Randall J (1993) Nanoelectronic circuits using resonant tunneling transistors and diodes. In: Digest of Technical Papers. IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, 24–26 Feb 1993, pp 174–175
65. Fredkin E, Toffoli T (1982) Conservative logic. Int J Theor Phys 21:219–253
66. Fukś H (2002) Nondeterministic density classification with diffusive probabilistic cellular automata. Phys Rev E 66(6):066106.
67. Gács P (1986) Reliable computation with cellular automata. J Comput Syst Sci 32(1):15–78
68. Gács P (1989) Self-correcting two-dimensional arrays. In: Micali S (ed) Randomness in Computation. Advances in Computing Research (a scientific annual), vol 5. JAI Press, Greenwich, pp 223–326
69. Gács P (1997) Reliable cellular automata with self-organization. In: IEEE Symposium on Foundations of Computer Science, pp 90–99
70. Gács P, Reif X (1988) A simple three-dimensional real-time reliable cellular array. J Comput Syst Sci 36(2):125–147
71. Gao C, Hammerstrom D (2007) Cortical models onto CMOL and CMOS – architectures and performance/price. IEEE Trans Circ Syst I: Regul Pap 54(11):2502–2515
72. Gil D, de Andrés D, Ruiz JC, Gil P (2007) Identifying fault mechanisms and models of emerging nanoelectronic devices. In: Workshop on Dependable and Secure Nanocomputing (DSN'07). Online proceedings www.laas.fr/WDSN07/WDSN07_files/Texts/WDSN07-POST-01-Gil.pdf. Accessed 5 Aug 2008
73. Gimarc CE, Milutinovic VM (1987) A survey of RISC processors and computers of the mid-1980s. Computer 20(9):59–69
74. Goldstein SC (2005) The impact of the nanoscale on computing systems. In: IEEE/ACM International Conference on Computer-Aided Design (ICCAD 2005). San Jose, CA, pp 655–661. Online Proceedings www.cs.cmu.edu/~seth/papers/goldstein-iccad05.pdf. Accessed 5 Aug 2008
75. Goldstein SC, Budiu M (2001) Nanofabrics: Spatial computing using molecular electronics. In: Proceedings of the 28th annual international symposium on Computer architecture, pp 178–191

76. Graham P, Gokhale M (2004) Nanocomputing in the presence of defects and faults: a survey. In: Nano, Quantum and Molecular Computing. Kluwer, Boston, pp 39–72
77. Green JE, Choi JW, Boukai A, Bunimovich Y, Johnston-Halperin E, Delonno E, Luo Y, Sheriff BA, Xu K, Shin YS, Tseng HR, Stoddart JF, Heath JR (2007) A 160-kilobit molecular electronic memory patterned at 10^{11} bits per square centimeter. Nature 445(7126):414–417
78. Han J, Jonker P (2003) A defect- and fault-tolerant architecture for nanocomputers. Nanotechnology 14(2):224–230
79. Han J, Gao J, Qi Y, Jonker P, Fortes JAB (2005) Toward hardware-redundant, fault-tolerant logic for nanoelectronics. IEEE Des & Test Comput 22(4):328–339
80. Harao M, Noguchi S (1975) Fault tolerant cellular automata. J Comput Syst Sci 11(2):171–185
81. Hartmanis J (1995) On the weight of computations. Bull Eur Assoc Theor Comput Sci 55:136–138
82. Hauck S (1995) Asynchronous design methodologies: an overview. Proc IEEE 83(1):69–93
83. Haykin S (1998) Neural Networks: A Comprehensive Foundation. Prentice Hall PTR, Upper Saddle River, NJ
84. Heath JR, Kuekes PJ, Snider GS, Williams RS (1998) A defect-tolerant computer architecture: Opportunities for nanotechnology. Science 280(5370):1716–1721
85. Heinrich AJ, Lutz CP, Gupta JA, Eigler DM (2002) Molecule cascades. Science 298(5597):1381–1387
86. Ho R, Mai KW, Horowitz MA (2001) The future of wires. Proc IEEE 89:490–504
87. Huang Y, Duan X, Wei Q, Lieber C (2001) Directed assembly of one-dimensional nanostructures into functional networks. Science 291(5504):630–633
88. Husband CP, Husband SM, Daniels JS, Tour JM (2003) Logic and memory with nanocell circuits. IEEE Trans Electron Dev 50(9):1865–1875
89. Hush NS (2003) An overview of the first half-century of molecular electronics. Ann N Y Acad Sci 1006:1–20
90. Isokawa T, Abo F, Peper F, Kamiura N, Matsui N (2003) Defect-tolerant computing based on an asynchronous cellular automaton. In: Proceedings of SICE Annual Conference, Fukui, Japan, pp 1746–1749
91. Isokawa T, Abo F, Peper F, Adachi S, Lee J, Matsui N, Mashiko S (2004) Fault-tolerant nanocomputers based on asynchronous cellular automata. Int J Mod Phys C 15(6):893–915
92. Isokawa T, Kowada S, Peper F, Kamiura N, Matsui N (2006) On-line marking of defective cells by random flies. In: Yacoubi SE, Chopard B, Bandini S (eds) Lecture Notes in Computer Science, vol 4173. Springer, Berlin, pp 347–356
93. Isokawa T, Kowada S, Takada Y, Peper F, Kamiura N, Matsui N (2007) Defect-tolerance in cellular nanocomputers. New Gener Comput 25(2):171–199
94. International Roadmap Commitee (2005) International Technology Roadmap for Semiconductors
95. International Roadmap Commitee (2005) International Technology Roadmap for Semiconductors, Emerging Research Devices. www.itrs.net/Links/2005ITRS/ERD2005.pdf. Accessed 5 Aug 2008
96. International Roadmap Commitee (2005) International Technology Roadmap for Semiconductors, Interconnect. www.itrs.net/Links/2005ITRS/ERD2005.pdf. Accessed 5 Aug 2008
97. Iwai H (2004) CMOS scaling for sub-90 nm to sub-10 nm. In: VLSID '04: Proceedings of the 17th International Conference on VLSI Design, IEEE Computer Society, Washington, DC, p 30
98. Jablonski DG (1990) A heat engine model of a reversible computation. Proc IEEE 78(5):817–825
99. Jung GY, Johnston-Halperin E, Wu W, Yu Z, Wang SY, Tong WM, Li Z, Green JE, Sheriff BA, Boukai A, Bunimovich Y, Heath JR, Williams RS (2006) Circuit fabrication at 17nm half-pitch by nanoimprint lithography. Nano Lett 6(3):351–354
100. Kamins TI, Williams RS, Chen Y, Chang YL, Chang YA (2000) Chemical vapor deposition of Si nanowires nucleated by TiSi2 islands on si. Appl Phys Lett 76(5):562–564
101. Kiehl RA (2006) Information processing in nanoscale arrays: DNA assembly, molecular devices, nano-array architectures. In: ICCAD '06: Proceedings of the 2006 IEEE/ACM international conference on Computer-aided design, San Jose, 5–9 Nov 2006
102. Kish LB (2002) End of Moore's law: thermal (noise) death of integration in micro and nano electronics. Phys Lett A 305 (3–4):144–149
103. Kish LB (2006) Thermal noise driven computing. Appl Phys Lett 89(14):144104
104. Knap W, Deng Y, Rumyantsev S, Lu JQ, Shur MS, Saylor CA, Brunel LC (2002) Resonant detection of subterahertz radiation by plasma waves in a submicron field-effect transistor. Appl Phys Lett 80(18):3433–3435
105. Korkmaz P, Akgul BES, Palem KV, Chakrapani LN (2006) Advocating noise as an agent for ultra-low energy computing: probabilistic complementary metal-oxide-semiconductor devices and their characteristics. Jpn J Appl Phys 45(4B):3307–3316
106. Kreup F, Graham AP, Liebau M, Duesberg GS, Seidel R, Unger E (2004) Carbon nanotubes for interconnect applications. In: Electron Devices Meeting, 2004. IEDM Technical Digest. IEEE International, pp 683–686
107. Kuekes PJ, Williams RS, Heath JR (2000) Demultiplexer for a molecular wire crossbar network. US Patent 6 128 214
108. Kuekes PJ, Robinett W, Seroussi G, Williams RS (2005) Defect-tolerant interconnect to nanoelecronic circuits: internally redundant demultiplexers based on error-correcting codes. Nanotechnology 16(6):869–881
109. Kuekes PJ, Robinett W, Williams RS (2005) Improved voltage margins using linear error-correcting codes in resistor-logic demultiplexers for nanoelectronics. Nanotechnology 16(9):1419–1432
110. Kuekes PJ, Steward DR, Williams RS (2005) The crossbar latch: Logic value storage, restoration, and inversion in crossbar circuits. J Appl Phys 97(3):034301
111. Kuekes PJ, Robinett W, Roth RM, Seroussi G, Snider GS, Williams RS (2006) Resistor-logic demultiplexers for nanoelectronics based on constant- weight codes. Nanotechnology 17(4):1052–1061
112. Lala PK (2001) Self-checking and fault-tolerant digital design. Morgan Kaufmann, San Francisco, CA
113. Landauer R (1961) Irreversibility and heat generation in the computing process. IBM J Res Dev 5(3):183–191
114. Landauer R (1984) Dissipation in computation. Phys Rev Lett 53(12):1205
115. Landauer R (1992) Information is physical. In: PhysComp'92: Workshop on Physics and Computation, Dallas, 2–4 Oct 1992, pp 1–4

116. Le J, Pinto Y, Seeman NC, Musier-Forsyth K, Taton TA, Kiehl RA (2004) DNA-templated self-assembly of metallic nanocomponent arrays on a surface. Nano Lett 4(12):2343–2347
117. Lee J, Adachi S, Peper F, Morita K (2003) Embedding universal delay-insensitive circuits in asynchronous cellular spaces. Fundamenta Informaticae 58(3/4):295–320
118. Lee J, Peper F, Adachi S, Mashiko S (2004) On reversible computation in asynchronous systems. In: Quantum Information and Complexity. World Scientific, Singapore, pp 296–320
119. Lee J, Adachi S, Peper F, Mashiko S (2005) Delay-insensitive computation in asynchronous cellular automata. J Comput Syst Sci 70:201–220
120. Lee J, Peper F, Adachi S (2006) Reversible logic elements operating in asynchronous mode. US Patent 6 987 402
121. Lent CS, Tougaw PD, Porod W, Bernstein GH (1993) Quantum cellular automata. Nanotechnology 4(1):49–57
122. Li C, Fan W, Lei B, Zhang D, Han S, Tang T, Liu X, Liu Z, Asano S, Meyyappan M, Han J, Zhou C (2004) Multilevel memory based on molecular devices. Appl Phys Lett 84(11):1949–1951
123. Liebmann LW (2003) Layout impact of resolution enhancement techniques: impediment or opportunity? In: Proc. 2003 Int. Symp. on Physical Design (ISPD'03), ACM Press, pp 110–117
124. Likharev KK, Semenov VK (1991) RSFQ logic/memory family: a new Josephson-junction technology for sub-terahertz-clock-frequency digital systems. IEEE Trans Appl Supercond 1(1):3–28
125. Likharev KK, Strukov DB (2005) Introduction to Molecular Electronics. In: Cuniberti G et al (ed) CMOL: Devices, circuits, and architectures. Springer, Berlin, pp 447–477
126. Lloyd S (1993) A potentially realizable quantum computer. Science 261(5128):1569–1571
127. Lloyd S (2000) Ultimate physical limits to computation. Nature 406(6799):1047–1054
128. Madou MJ (2002) Lithography. In: Fundamentals of Microfabrication, The Science of Miniaturization. CRC Press, Florida, pp 1–76
129. Maezawa K, Förster A (2003) Quantum transport devices based on resonant tunneling. In: Nanoelectronics and Information Technology, pp 407–424
130. Manohar R, Martin AJ (1995) Quasi-delay-insensitive circuits are turing-complete. Tech. Rep. CaltechCSTR:1995.cs-tr-95-11, California Institute of Technology, Pasadena, CA
131. Margolus NH (1984) Physics-like models of computation. Physica D 10(1/2):81–95
132. Margolus NH (1999) Crystalline computation. In: Feynman and computation: exploring the limits of computers. Perseus books, Cambridge, pp 267–305
133. Martin AJ (1990) Programming in VLSI: From communicating processes to delay-insensitive circuits. In: Hoare CAR (ed) Developments in Concurrency and Communication. Addison-Wesley, Reading, pp 1–64
134. Mayor M, Weber HB, Waser R (2003) Molecular Electronics. In: Nanoelectronics and Information Technology. Wiley, Berlin, pp 501–525
135. Mead C (1990) Neuromorphic electronic systems. Proc IEEE 78(10):1629–1636
136. Mead C, Conway L (1980) Introduction to VLSI Systems. Addison-Wesley, Boston
137. Meindl JD (1995) Low power microelectronics: retrospect and prospect. Proc IEEE 83(4):619–635
138. Meindl JD, Chen Q, Davis JA (2001) Limits on silicon nanoelectronics for terascale integration. Science 293(5537):2044–2049
139. Miller DAB (2000) Rationale and challenges for optical interconnects to electronic chips. Proc IEEE 88(6):728–749
140. Mishra M, Goldstein SC (2003) Defect tolerance at the end of the roadmap. In: Proceedings of the IEEE International Test Conference (ITC), vol 1, pp 1201–1210
141. Mizuno M, Anjo K, Surni Y, Wakabayashi H, Mogami T, Horiuchi T, Yamashina M (2000) On-chip multi-ghz clocking with transmission lines. In: 2000 IEEE International Solid-State Circuits Conference (ISSCC). Digest of Technical Papers, pp 366–367
142. Montemerlo MS, Love JC, Opiteck GJ, Goldhaber-Gordon DJ, Ellenbogen JC (1996) Technologies and designs for electronic nanocomputers. Tech. Rep. 96W0000044, MITRE
143. Moore GE (2003) No exponential is forever: but "forever" can be delayed! In: Solid-State Circuits Conference. Digest of Technical Papers. ISSCC. IEEE International Solid-State Circuits Conference (ISSCC), vol 1, pp 20–23
144. Morales A, Lieber C (2001) A laser ablation method for the synthesis of crystalline semiconductor nanowires. Science 291(5348):208–211
145. Morita K (2003) A simple universal logic element and cellular automata for reversible computing. Lect Notes Comput Sci 2055:102–113
146. Motwani R, Raghavan P (1995) Randomized Algorithms. Cambridge University Press, New York, NY
147. Muller DE, Bartky WS (1959) A theory of asynchronous circuits. In: Proceedings of an International Symposium on the Theory of Switching. Harvard University Press, pp 204–243
148. Nikolić K, Forshaw M (2003) The current status of nanoelectronic devices. Int J Nanosci 2(1/2):7–29
149. Nikolić K, Sadek A, Forshaw M (2002) Fault-tolerant techniques for nanocomputers. Nanotechnology 13(3):357–362
150. Nishio H, Kobuchi Y (1975) Fault tolerant cellular spaces. J Comput Syst Sci 11(2):150–170
151. O KK, Kim K, Floyd B, Mehta J, Yoon H, Hung CM, Bravo D, Dickson T, Guo X, Li R, Trichy N, Caserta J, Bomstad W, Branch J, Yang DJ, Bohorquez J, L Gao L, Sugavanam A, Lin JJ, Chen J, Martin F, Brewer J (2003) Wireless communications using integrated antennas. In: Proc. 2003 IEEE International Interconnect Technology Conference, San Francisco, 2–4 June 2003, pp 111–113
152. O'Mahony F, Yue CP, Horowitz MA, Wong SS (2003) A 10-GHz global clock distribution using coupled standing-wave oscillators. IEEE J Solid-State Circ 38(11):1813–1820
153. O'Mahony F, Yue CP, Horowitz M, Wong SS (2003) 10 GHz clock distribution using coupled standing-wave oscillators. In: Solid-State Circuits Conference. Digest of Technical Papers. IEEE International Solid-State Circuits Conference (ISSCC), vol 1, pp 428–504
154. Ono Y, Fujiwara A, Nishiguchi K, Inokawa H, Takahashi Y (2005) Manipulation and detection of single electrons for future information processing. J Appl Phys 97:031101
155. Palem KV (2005) Energy aware computing through probabilistic switching: a study of limits. IEEE Trans Comput 54(9):1123–1137

156. Parviz BA, Ryan D, Whitesides GM (2003) Using self-assembly for the fabrication of nano-scale electronic and photonic devices. IEEE Trans Adv Packag 26(3):233–241
157. Peper F, Lee J, Adachi S, Mashiko S (2003) Laying out circuits on asynchronous cellular arrays: a step towards feasible nanocomputers? Nanotechnology 14(4):469–485
158. Peper F, Lee J, Abo F, Isokawa T, Adachi S, Matsui N, Mashiko S (2004) Fault-tolerance in nanocomputers: a cellular array approach. IEEE Trans Nanotechnol 3(1):187–201
159. Petty M (2007) Molecular Electronics, from Principles to Practice. Wiley, West Sussex
160. Pinto YY, Le JD, Seeman NC, Musier-Forsyth K, Taton TA, Kiehl RA (2005) Sequence-encoded self-assembly of multiple-nanocomponent arrays by 2D DNA scaffolding. Nano Lett 5(12):2399–2402
161. Pippenger N (1985) On networks of noisy gates. In: 26th Annual Symposium on Foundations of Computer Science, 21–23 October 1985, Portland, Oregon, IEEE, pp 30–38
162. Pippenger N (1989) Invariance of complexity measures for networks with unreliable gates. J ACM 36(3):531–539
163. Pippenger N (1990) Developments in: "The synthesis of reliable organisms from unreliable components". In: Proc. of Symposia in Pure Mathematics, vol 50. pp 311–324
164. Porod W (1998) Quantum-dot cellular automata devices and architectures. International journal of high-speed electronics and systems 9(1):37–63
165. Porod W, Grondin RO, Ferry DK (1984) Dissipation in computation. Phys Rev Lett 52(3):232–235
166. Rahman A, Reif R (2000) System-level performance evaluation of three-dimensional integrated circuits. IEEE Trans Very Large Scale Integr Syst 8(6):671–678
167. Robert RW, Keyes W (1985) What makes a good computer device? Science 230(4722):138–144
168. Robinson AL (1984) Computing without dissipating energy. Science 223(4641):1164–1166
169. Rothemund PW, Papadakis N, Winfree E (2004) Algorithmic self-assembly of DNA sierpinski triangles. PLoS Biol 2(12): 2041–2053
170. Roychowdhury VP, Janes DB, Bandyopadhyay S, Wang X (1996) Collective computational activity in self-assembled arrays of quantum dots: a novel neuromorphic architecture for nanoelectronics. IEEE Trans Electron Dev 43(10):1688–1699
171. Rueckes T, Kim K, Joselevich E, Tseng G, Cheung C, Lieber C (2000) Carbon nanotube based nonvolatile random access memory for molecular computing. Science 289(5476):94–97
172. Sadek AS, Nikolić K, Forshaw M (2004) Parallel information and computation with restitution for noise-tolerant nanoscale logic networks. Nanotechnology 15(1):192–210
173. Sathe V, Chueh JY, Kim J, Ziesler CH, Kim S, Papaefthymiou M (2005) Fast, efficient, recovering, and irreversible. In: CF '05: Proceedings of the 2nd Conference on Computing Frontiers. ACM, New York, pp 407–413
174. Seitz CL (1980) System timing. In: Mead CA, Conway LA (eds) Introduction to VLSI Systems. Addison–Wesley, Boston
175. Sherman WB, Seeman NC (2004) A precisely controlled DNA biped walking device. Nano Lett 4(7):1203–1207
176. Shor PW (2004) Progress in quantum algorithms. Quantum Inf Process 3(1–5):5–13
177. Smith PA, Nordquist CD, Jackson TN, Mayer TS, Martin BR, Mbindyo J, Mallouk TE (2000) Electric-field assisted assembly and alignment of metallic nanowires. Appl Phys Lett 77(9):1399–1401
178. Snepscheut JvD (1985) Trace theory and VLSI design. In: Lecture Notes in Computer Science, vol 200. Springer, Berlin
179. Snider GS, Kuekes PJ (2003) Molecular-junction-nanowire-crossbar-based associative array. US Patent 6 898 098
180. Snider GS, Robinett W (2005) Crossbar demultiplexers for nanoelectronics based on n-hot codes. IEEE Trans Nanotechnol 4(2):249–254
181. Snider GS, Williams RS (2007) Nano/CMOS architectures using a field-programmable nanowire interconnect. Nanotechnology 18(3):1–11
182. Snider GS, Kuekes PJ, Williams RS (2004) CMOS-like logic in defective, nanoscale crossbars. Nanotechnology 15(8):881–891
183. Snider GS, Kuekes PJ, Hogg T, Williams RS (2005) Nanoelectronic architectures. Appl Phys A 80(6):1183–1195
184. Soh C, Quate C, Morpurgo C, Marcus C, Kong C, Dai C (1999) Integrated nanotube circuits: controlled growth and ohmic contacting of single-walled carbon nanotubes. Appl Phys Lett 75(5):627–629
185. Sotiriadis PP (2006) Information capacity of nanowire crossbar switching networks. IEEE Trans Inf Theory 52(7):3019–3032
186. Spagocci S, Fountain T (1999) Fault rates in nanochip devices. Proc Electrochem Soc 98-19:582–596
187. Spielman DA (1996) Highly fault-tolerant parallel computation. In: Proceedings of the 37th IEEE Symposium on Foundations of Computer Science (FOCS), Burlington, 14–16 Oct 1996, pp 154–163
188. Srivastava N, Banerjee K (2004) Interconnect challenges for nanoscale electronic circuits. TMS J Mater (JOM) 56(10):30–31
189. Stan MR, Franzon PD, Goldstein SC, Lach JC, Ziegler MM (2003) Molecular electronics: from devices and interconnect to circuits and architecture. Proc IEEE 91(11):1940–1957
190. Strukov DB, Likharev KK (2005) CMOL FPGA: a reconfigurable architecture for hybrid digital circuits with two-terminal nanodevices. Nanotechnology 16(6):888–900
191. Taubin A, Cortadella J, Lavagno L, Kondratyev A, Peeters A (2007) Design automation of real life asynchronous devices and systems. Found Trends Electron Des Autom 2(1):1–133
192. Theis TN (2000) The future of interconnection technology. IBM J Res Dev 44(3):379–390
193. Toffoli T (1984) Comment on: Dissipation in computation. Phys Rev Lett 53(12):1204
194. Tougaw PD, Lent CS (1994) Logical devices implemented using quantum cellular-automata. J Appl Phys 75:1818–1825
195. Tour JM, Van Zandt L, Husband CP, Husband SM, Wilson LS, Franzon PD, Nackashi DP (2002) Nanocell logic gates for molecular computing. IEEE Trans Nanotechnol 1(2):100–109
196. Tour JM, Cheng L, Nackashi DP, Yao Y, Flatt AK, St Angelo SK, Mallouk TE, Franzon PD (2003) Nanocell electronic memories. J Am Chem Soc 125(43):13279–13283
197. Türel Ö, Lee JH, Ma X, Likharev K (2005) Architectures for nanoelectronic implementation of artificial neural networks: new results. Neurocomputing 64:271–283
198. Uchida K (2003) Single-electron devices for logic applications. In: Nanoelectronics and Information Technology. Wiley, Berlin, pp 425–443
199. Unger SH (1969) Asynchronous Sequential Switching Circuits. Wiley, New York

200. von Hippel AR (1956) Molecular engineering. Science 123(3191):315–317
201. von Neumann J (1956) Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components. In: Automata Studies. Princeton University Press, Princeton, pp 43–98
202. Waingold E, Taylor M, Srikrishna D, Sarkar V, Lee W, Lee V, Kim J, Frank M, Finch P, Barua R, Babb J, Amarasinghe S, Agarwal A (1997) Baring it all to software: Raw machines. Computer 30(9):86–93
203. Wang KL, Khitun A, Flood AH (2005) Interconnects for nanoelectronics. In: Proc. 2005 IEEE International Interconnect Technology Conference, San Francisco, 6–8 June 2005, pp 231–233
204. Wang W (1990) An asynchronous two-dimensional self-correcting cellular automaton. Ph D thesis, Boston University, Boston, MA 02215, short version: In Proc. 32nd IEEE Symposium on the Foundations of Computer Science, San Juan, 1–4 Oct 1990. IEEE Press, pp 188–192, 1991
205. Weeber JC, González MU, Baudrion AL, Dereux A (2005) Surface plasmon routing along right angle bent metal strips. Appl Phys Lett 87(22):221101
206. Whitesides GM, Grzybowsky B (2002) Self-assembly at all scales. Science 295(5564):2418–2421
207. Mac Williams FJ, Sloane NJA (1978) The Theory of Error-Correcting Codes. North-Holland, Amsterdam
208. Williams RS, Kuekes PJ (2001) Demultiplexer for a molecular wire crossbar network. US Patent 6 256 767
209. Winfree E, Liu F, Wenzler LA, Seeman NC (1998) Design and self-assembly of two-dimensional DNA crystals. Nature 394(6693):539–544
210. Wolf SA, Awschalom DD, Buhrman RA, Daughton JM, von Molnar S, Roukes ML, Chtchelkanova AY, Treger DM (2001) Spintronics: a spin-based electronics vision for the future. Science 294(5546):1488–1495
211. Wong HSP, Frank DJ, Solomon PM, Wann CHJ, Wesler JJ (1999) Nanoscale CMOS. Proc IEEE 87(4):537–570
212. Wood J, Edwards TC, Lipa S (Nov 2001) Rotary traveling-wave oscillator arrays: a new clock technology. IEEE J Solid-State Circ 36(11):1654–1665
213. Worschech L, Beuscher F, Forchel A (1999) Quantized conductance in up to 20 μm long shallow etched GaAs/AlGaAs quantum wires. Appl Phys Lett 75(4):578–580
214. Wu W, Jung GY, Olynick DL, Straznicky J, Li Z, Li X, Ohlberg DAA, Chen Y, Wang SY, Liddle JA, Tong WM, Williams RS (2005) One-kilobit cross-bar molecular memory circuits at 30-nm half-pitch fabricated by nanoimprint lithography. Appl Phys A 80(6):1173–1178
215. Yamada T, Akazawa M, Asai T, Amemiya Y (2001) Boltzmann machine neural network devices using single-electron tunneling. Nanotechnology 12(1):60–67
216. Yanagida T, Ueda M, Murata T, Esaki S, Ishii Y (2007) Brownian motion, fluctuation and life. Biosystems 88(3):228–242
217. Yang T, Kiehl R, Chua L (2001) Tunneling phase logic cellular nonlinear networks. Int J Bifurc Chaos 11(12):2895–2911
218. Zhirnov VV, Cavin RK, Hutchby JA, Bourianoff GI (2003) Limits to binary logic switch scaling – a gedanken model. Proc IEEE 91(11):1934–1939
219. Zhong Z, Wang D, Cui Y, Bockrath MW, Lieber CM (2003) Nanowire crossbar arrays as address decoders for integrated nanosystems. Science 302(5649):1377–1379

Nanoscale Atomic Clusters, Complexity of

Anatoly I. Frenkel[1], Judith C. Yang[2], Duane D. Johnson[3], Ralph G. Nuzzo[4]
[1] Department of Physics, Yeshiva University, New York, USA
[2] Department of Mechanical Engineering and Materials Science, University of Pittsburgh, Pittsburgh, USA
[3] Department of Materials Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, USA
[4] Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Synthesis of Nanoparticles Through Self-Assembly
Observation and Characterization:
Synergy of Experiments and First Principle Modeling
Structural Relaxation of Nanoparticles
Future Directions
Acknowledgments
Bibliography

Glossary

Nanoparticles These, also called nanoclusters, are atomic agglomerates having the size less than 100 nm at least in one dimension.

Bulk material The material which grain sizes are in the micrometer scale or larger.

Self-assembly method A method that a system uses to develop and form particular structure using only pre-existing components.

Definition of the Subject

Freestanding and supported, nanoscale metallic clusters (i. e., nanoclusters or nanoparticles) constitute complex systems as they span an enormous space of potential structures, each having different collective properties, with only few potentially stable (or metastable) structures that exhibit technologically beneficial properties. The synthesis and design of collective properties is a critical area of current research – with technological beneficial outcomes often equivalent to finding a "needle-in-the-haystack". Yet

the synthesis and application of nanoparticles with specific properties has had, and will have, an increasing impact on technology, whether for improving toothpaste, 'green(er)' energy production (e. g., in petroleum refinement, catalysis, or batteries), biotechnology and national security issues (e. g., biosensors). From the perspective of theory, freestanding or supported nanoparticles constitute an enormous challenge in terms of property assessment as well as comparison of results, where the experimental processing routes to produce samples must be considered within theory for one-to-one comparison to measured data and understanding of the controlling physics to that observed.

Here we provide a glimpse into the complexity of the search space and a subset of the critical experimental and theoretical analysis tools that are capable of providing quantitative assessment of structure-property relations actually found, such as assessment of structure and reactivity, say, for catalytic performance. Hence, it is necessary (and desirable) to have a direct and meaningful coupling of experimental and theoretical analysis. It is our purpose to provide a brief overview of current quantitative methods and examples of such coupling of experiment and theoretical analysis.

First, to exemplify the challenges, we provide an example list of issues. Even with an established synthesis route for a single-component nanoparticle, one cannot control with certainty the size, only the distribution – a narrow distribution is great but the nanoparticles properties will still depend, at a minimum, on surface-to-volume ratio (i. e., inversely as particle radii or size, $1/R$), an example of which is melting [1]. Already with a two-component nanoparticle, such as Pb-Bi, melting is controlled again in part by $1/R$ but also by thermal and electronically-driven effects such as chemical segregation, leading, for example, to shell-like nanoparticles with differing melting temperatures for inner and outer regions [2]. In addition, for a particle of fixed number of atoms, N, there can be an enormous number of structural isomers possible, or bonding motifs. For example, in Fig. 1 are the calculated lowest-energy, 13-atom ($N = 13$) freestanding isomers of Pt, [3] similar results have been obtained for other elements. From Refs. [3,4,5] it can be garnered (amongst many other points) that (i) high-symmetry structures are not necessarily low in energy, (ii) structures can be degenerate in energy, (iii) there is no simple correlation between structure and local properties, such as coordination number or bond lengths, (iv) both 3d and 2d structures compete in energy, and a structure (and concomitantly their properties) will be strongly affected by placing the particles on a support, and (v) intuition regarding the important structures can be helpful but not reliable. Additionally, for fixed N and only two types of atoms (A and B) within the cluster $A_m B_n$, such that $N = m + n$, there are a binomial number $N!/(N-m)!m!$ of possible "alloy" configurations (i. e., homotopic groups of clusters, homomers that include stereoisomers) within a single structure (The term "homotops" has been introduced by J. Jellinek to describe such clusters that form homotopic groups. But, to be consistent with standard chemistry use of isomers and the general symmetry descriptions, such as enantiomers, diastereomers, and stereoisomers, we refer to these as homomers). There have been two recent and excellent reviews [6,7] thoroughly discussing these issues, mainly from a theoretical perspective. Finally, nanoparticles with structures that have useful properties may be only possible through self-assembly of organic stabilizers, e. g. Au nanoparticle stabilized via thiol ligands [8]. Clearly, the possible configuration space is daunting. Yet, it is critical that we can assess the structural properties of synthesized freestanding and supported nanoparticles in order to sort out many factors influencing their complex behaviors and correlate them with property design requirements. This is the principal scope of what follows, where we focus on the integrated use of several advanced analytic techniques – x-ray absorption fine structure (XAFS), and atom-counting methods of electron microscopy, and ab initio theoretical methods to achieve these goals and answer questions central for many complex systems, including nanoscale atomic clusters.

Introduction

Metallic nano-sized clusters (i. e., nanoclusters or nanoparticles) play a crucial role in modern science and technology. The application of design at the atomic scale can be found in petroleum refinement, catalytic converters, and nanofabrication. On a more academic level, metallic nanoparticle research can be found integrated in materials science, biotechnology, and organic chemistry [8]. Despite the ever-increasing interest of nanoparticles in different fields of study, catalysis remains perhaps the essential application of metallic nanoclusters, due to its potential technological impact.

The enhanced catalytic activity of these nanomaterials results from an increased number of atoms exposed to the surface compared to the corresponding bulk material. As a corollary, the number of neighbors to which a given atom is coordinated decreases, giving rise to nonbulk-like behaviors [9,10,11,12,13]. The energetic considerations of such systems lead to important synthetic pathways such as hydrogenation of olefins and silylation of polymers [14].

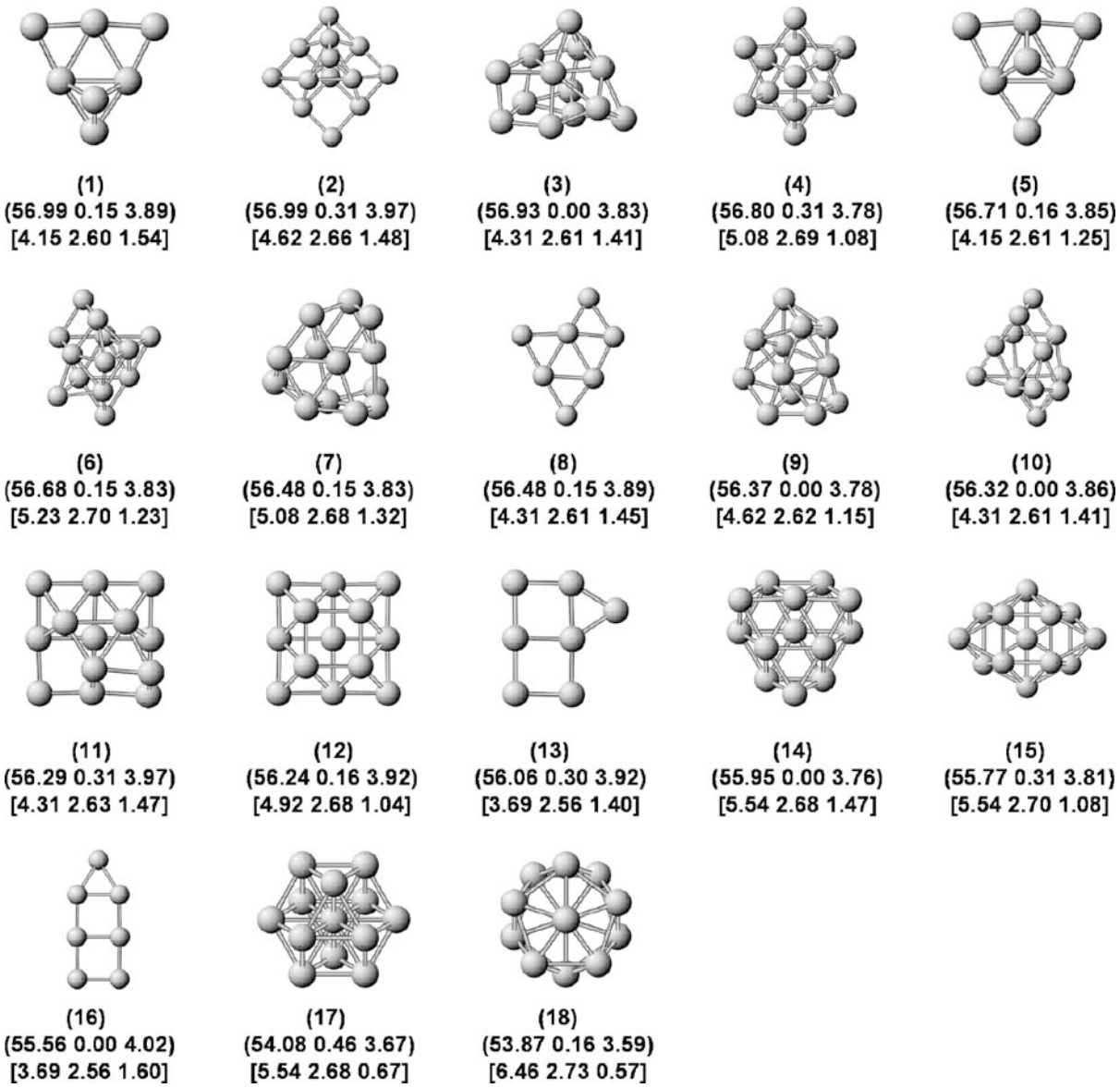

Nanoscale Atomic Clusters, Complexity of, Figure 1

Pt_{13} isomers in order of decreasing binding energy calculated via density functional theory (see text). Properties listed below the structure are (in parentheses) energy (eV), magnetic moment (μ_B/atom), mean inter-atomic distance (Å), and [in brackets] first nearest-neighbor coordination number, first NN bond length (Å) and s-d hybridization index, see Ref. [3]. Structures 17 and 18 are the compact O_h and I_h isomers, and 14 is the 2-layer cuboctahedrally stacked cluster, which forms the basis of bulk fcc Pt. Reprinted figure with permission from [3]

The syntheses of these nanoscale materials can be categorized as physical or chemical methods. Physical methods involve a "top-down" approach where starting material is sculpted down to the atomic scale in the gaseous state. Once in the gaseous state the atoms can then condense into larger structures. Examples of physical methods in the synthesis of nanoclusters are metal-vapor deposition [15,16,17] and laser ablation [15,18]. In both these methods different sources of energy are utilized to generate atoms from bulk material. Typically, the vaporized atoms are then deposited onto a substrate under vacuum, where particles form by nucleation [19]. Again, the reactive surface of these particles results in aggregation and reduction of the surface area leading to self-assem-

bly of a metallic cluster [20]. One of the drawbacks associated with this method of synthesis is the financial burden due to the expensive equipment required. Another limitation is the broad distribution of cluster sizes generated in the syntheses [15,18]. Yet despite recent advancements in physical methods for nanoclusters synthesis, [21,22,23] the field is predominantly controlled by chemistry [15].

Chemical methods take a "bottom-up" approach toward synthesis. Unlike physical methods, applying chemistry toward nanoparticle synthesis implies the assembly of the structures from chemical precursors. Traditionally, nanoparticle synthesis involves the reduction of metal salts in solution followed by the use of a capping agent to prevent aggregation of the particles. Reducing agents such as alcohols, [24] hydrogen gas, [25] and hydrides [26] can be used in the reduction of the ionic metal. Capping agents typically used are organic polymers, [24] as well as surfactants [25,26]. Coupling of the reducing and capping agent can be accomplished with molecules such as sodium citrate [27]. Chemical methods offer better control of particle size, with the assistance of capping agents, and greater flexibility because of the variety of chemical precursors that can be utilized in synthesis. In contrast to physical methods, fabrication of nanoparticles using chemistry is relatively inexpensive. Even with the improved precision intrinsic to chemical methods however, reproducibility of size, structure, and catalytic ability at the nanoscale regime (1–10 nm) is not concrete. On the other hand, developments in chemical synthesis of nanoparticles have opened doors toward other innovative methods of generating metallic nanoclusters [14].

Despite the relatively large number of available structural techniques, most of them obtain the overall, volume-average properties of nanoparticles, and give little, if any, insight into sometimes very elaborate *actual* arrangement of atoms within the particle. Two state of the art methodologies, synchrotron x-ray absorption fine-structure (XAFS) and quantitative scanning transmission electron microscopy (STEM) have been best positioned for determining the 3D structure and structural habits, both individually and as an ensemble, critical for understanding metallic nanoclusters. XAFS technique is one of the premiere tools to study both atomic and electronic structure of small ensembles due, in part, to its local structure sensitivity and excellent spatial resolution. By measuring coordination numbers, bond lengths and their disorder up to the 5th coordination shell, one can reliably determine the size of the nanoparticles, their shapes (e. g., oblate, raft-like, or truncated polyhedral), surface morphology as well as effects of surface disorder in 1–2 nm-size clusters as a function of external conditions (temperature, alloy composition, support material, etc.).

Complementary information on site-specific structure and chemistry can be obtained by STEM, which has a unique capability for providing structural and spectral information simultaneously. Supported metal nanoclusters of size of 1–100 atoms make an ideal system for examination by high-angle annular dark-field (HAADF, also known as *Z*-contrast) imaging. Correlating the absolute image intensity to the scattering cross-section has been advanced recently within the STEM-based imaging method. With this improvement, one can directly count, with accuracy of ± 2 atoms, the number of atoms in a supported nanocluster avoiding complexities associated with coherent diffraction. By utilizing state of the art electron and x-ray probe methodologies, one can explore substrate/nanoparticle interactions as a function of support and nanoparticle material, as well as by size, composition and 3-D structure of the supported nanoparticles.

Most state-of-the-art experimental work is integrated with theoretical calculations to help interpret and accelerate identification of cluster bonding motifs, possible metastable structures, and determine electronic properties and reactivity of relevant clusters. Realistic nanometer-sized, metallic clusters (i. e., those equivalent to experiment) can be reliably simulated by electronic-structure and molecular-dynamic techniques to address the issues of complex geometries (cluster size and corresponding thermodynamically– or kinetically-stabilized shapes and atomic arrangement), of shape evolution (kinetics), as well as local bonding effects that determine reactivity. As such, a correlated, self-consistent interpretation of the experimental data from methods discussed above can accelerate identification of cluster bonding motifs, possible metastable and dynamic structures, and determination of the electronic properties of relevant clusters.

In the remainder of the article will first review the methods of synthesis of nanoparticles, including those guided by self-assembly (Sect. "Synthesis of Nanoparticles Through Self-Assembly"). The modern experimental and theoretical methods of nanoparticle analysis will be reviewed in Sect. "Observation and Characterization: Synergy of Experiments and First Principle Modeling". Complexity of several typical cluster systems, mono– and bimetallic, supported and freestanding, will be illustrated by examples from recent works (Sect. "Observation and Characterization: Synergy of Experiments and First Principle Modeling"). Sect. "Structural Relaxation of Nanoparticles" discusses the possible origins of structural relaxation and surface reconstruction observed experimentally

and studied theoretically in ultra-small clusters. Future directions are outlined in the final section.

Synthesis of Nanoparticles Through Self-Assembly

Besides traditional methods of chemical synthesis, the field has expanded to include many other routes toward production of metallic nanoparticles. Among these newer techniques are photochemical synthesis, electrochemical synthesis and sonochemical synthesis. Another departure from standard chemical methods of nanoparticle synthesis is in the production of bimetallic nanoparticles [7]. The classic framework for creating nanoparticles, although efficient for some elements, requires various reagents and consequently rigorous reaction conditions [28].

Photochemically induced synthesis of nanoparticles offers the advantage of using ultraviolet light to assist in the reaction without the addition of capping or stabilizing agents [14,28,29]. The reduction reaction can be explained by the photoexcitation of a metal salt and organic molecules containing carbonyl groups. The photoexcitation of the organic compound often results in the formation of radical [28,30] or subsequent formation of a radical [29]. Reduction of the metal ion ensues by the formed organic radical. In the case of multivalent metals, reduction occurs until the atom reaches a neutral state and aggregation of a nanoparticle follows thereafter. Although the formed particles aggregate without the inclusion of a capping agent, crystal growth is mediated by the ionic liquid solution of the original reaction mixture. The surfaces are stabilized electrostatically by the concentration of ions in the solution [28,29]. Other methods of photochemical synthesis often use reducing agents which also serve as capping agents; physically adsorbing to nanoparticles and arresting further aggregation [30,31,32,33].

Another alternative to traditional methods of synthesis is sonochemistry. The inner workings of sonochemistry stem from the extreme temperatures and pressures achieved by the expansion and implosion of bubbles in solutions, a process known as cavitation [14,34,35]. The intense energy released upon collapse of the bubbles results in the decomposition of molecules and the formation of free radicals. As with photochemical synthesis, the resulting free radicals initiate the reduction reactions used to reduce the metal ions in solution [36]. Particle size is maintained by the use of capping agents [37,38,39]. The main advantages of using sonochemical methods are the fast reaction rates and the ability to generate very small sized nanoparticles. On the other hand, a broad distribution of particle sizes is also typically obtained from this type of synthesis [40].

Electrochemical synthesis of nanoparticles is another viable option capable of generating metallic nanoclusters. In this method an anode is placed in solution and oxidized in the presence of a chaperone molecules or stabilizing agents. The oxidized atoms from the anode migrate to the cathode where they are reduced. Aggregation shortly follows after the migration to the cathode where cluster growth is mediated by the stabilizing agents. Isolation of the particles is accomplished by precipitation of the product from the cathode [14,41,42,43]. There are methods where a rotating cathode [44] or a double pulse technique [45,46] is implemented to improve size distribution and yield, but ultimately the chemistry that occurs is identical to the standard electrochemical method. Comparatively, reduction using electrochemistry has many advantages over its counterparts. Once formed, the nanoparticles can be easily isolated as they begin to precipitate out of solution. The size of the clusters generated can be controlled varying the current intensity of the cell, [42,45,46] but perhaps the greatest advantage of this method is the high yields achieved [14].

Another area of synthesis important in the field of metallic nanoparticles is bimetallic nanocluster preparation. The interest in bimetallic nanoclusters arises from the changes in physical and chemical properties compared to monometallic species [47,48,49,50,51,52]. As with monometallic clusters, many routes exist toward the synthesis of bimetallic clusters. These materials can be synthesized by traditional forms of reduction, using metal ion-containing salts (precursors). Additionally, this system of reduction can be accomplished simultaneously or sequentially. In simultaneous synthesis both metallic precursors are reacted in the same solution in the presence of capping agents. The mixture can result in the spontaneous formation of a core/shell structure as well as other alloyed structures [39,53]. Alternatively, successive addition of the metallic precursors has proven to be a more efficient way to control particle structure [14,39,54,55]. In the successive method the formation of the initial metal cluster behaves as a nucleation site for the growth of the second metal, affording a core shell motif. Incidentally, this self-assembly of core/shell structures has been induced simply by the mixing of nanoclusters of two different metals [56,57]. The structural motif has been attributed to a balance between surface energies as well as binding energies where size and composition of the clusters is detrimental to structural refinement [56]. Studies have also shown that the core/shell structure can be inverted at extreme temperatures [53]. Furthermore bimetallic nanoclusters can be synthesized by using a sacrificial layer of hydrogen adsorbed onto the surface of the core metallic

cluster. Using the adsorbed hydrogen layer as a reducing agent, incoming metallic ions are reduced onto the surface of the core cluster forming a shell. Thus far this method has only been generated with specific metallic species [58]. In addition to the aforementioned methods of synthesizing bimetallic particles, synthesis can also be accomplished by sonochemical, [39] electrochemical [54] and photochemical [55] methods.

Observation and Characterization: Synergy of Experiments and First Principle Modeling

Combination of complementary structural techniques – electron microscopy and XAFS, as well as DFT/MD simulations, provide previously unavailable atomic-level understandings of the structural dynamics of the most important forms of supported metal clusters. These methods can provide understandings of the metal framework bonding present in supported nanoscale clusters and the significant impacts on them that can originate as a consequence of adsorbate bonding and (more recently revealed) electronic effects mediated by support interactions. The data from microscopy now allows one to establish precise atomic compositions in supported catalyst systems, counting atoms explicitly at the single nanoparticle level. Experimental protocols based on synchrotron x-ray absorption spectroscopy allow the elucidation of the precise structural habits adopted by supported forms of metallic nanoparticles – establishing methodologies of modeling that define the nature of the strains present in such systems; unraveling the complex nature of the atomic level bonding habits they embed; the dynamical factors that mediate transformations of these structures due to the impacts of particle size, temperature-dependent support interactions, and adsorbates; the nature of the size-dependent atomic-bond relaxations that these particles embed; and the nature of the complex electronic structures that are unique to materials of this sort [59,60,61,62,63,64].

Development and testing of these advanced methods rely on the availability of synthetic methods that have provided model systems – both in the form of discrete monolayer protected clusters and supported metal nanoparticle catalysts – with precisely defined compositions and (for the latter) extremely narrow distributions of atomic mass.

Metal nanoclusters [65] (and gold nanoclusters in particular [66]) can exhibit structures that differ significantly from that corresponding to the bulk, ones that depend strongly on cluster size. Detailed knowledge of these structures is crucial for understanding and predicting nanocluster properties, including chemical, electrical, magnetic, and optical ones. The experimental determination of atomistic structural information is a very difficult task: analyses by imaging or scattering methods are presently limited by insufficient spatial resolution or by the coherent scattering size of these techniques. As a result, most structural determinations proceed indirectly by comparing experimentally accessible properties (e. g., ion mobility, photoemission spectra, polarizability, optical absorption, etc.) with those computed theoretically for candidate structures. Synergistic approaches, however, combining several structural techniques into a self-consistent structure refinement method, make accessible the previously unknown physical picture of the nanoscale. In the following sections, we will show examples of the determination of size, shape and atomic structures of mono– and bimetallic nanoparticles by combining results of extended x-ray absorption fine-structure (EXAFS), advanced methods of electron microscopy and DFT/MD calculations.

Structure Determination by EXAFS

By using EXAFS technique, one can extract accurate information about the identities, average distances, and coordinations of the neighboring atoms to the x-ray absorbing atom in nanoclusters [67,68]. The EXAFS signal, $\chi(k)$, is the sum of all contributions, $\chi_i(k)$, from groups of neighbors at approximately equal distances from the absorbing atoms (i. e., the ith shell), which are often written as: [69]

$$\chi_i(k) = \frac{S_0^2 n_i}{k R_i^2} \left| f_i^{\mathrm{eff}}(k) \right| \cdot \sin\left[2kR_i - \frac{4}{3}\sigma_i^{(3)} k^3 + \delta_i(k) \right] \mathrm{e}^{-2\sigma_i^2 k^3} \mathrm{e}^{-2R_i/\lambda_i(k)} , \tag{1}$$

where k is the photoelectron wave number, $f_i^{\mathrm{eff}}(k)$ and $\delta_i(k)$ are the photoelectron scattering-path amplitude and phase, respectively, S_0^2 is the passive electron reduction factor, n_i is the degeneracy of the scattering path, R_i is the effective half-path-length (which is equal to the interatomic distance for single-scattering paths), σ_i^2 is the mean-square deviation in R_i, $\sigma_i^{(3)}$ is the third cumulant, and $\lambda_i(k)$ is the photoelectron mean free path. Using modern computer packages, e. g., IFEFFIT, [70] which employs a non-linear least square method to fit theoretically calculated (with the help of FEFF6 code [69]) EXAFS signal to the data, one can be obtain the best-fit values of structural parameters, together with their uncertainties. The mean-square deviation, σ^2, of the first nearest neighbor (1NN) distance can be represented to a good approximation as a superposition of static (σ_s^2) and dynamic (σ_d^2) terms:

$$\sigma^2 = \left\langle (r - \langle r \rangle)^2 \right\rangle = \sigma_s^2 + \sigma_d^2 . \tag{2}$$

To separate the temperature-independent σ_s^2 and temperature-dependent σ_d^2, one can use a simple correlated Einstein model for σ_d^2:

$$\sigma_d^2 = \frac{\hbar}{2\omega\mu}\frac{1+\exp(-\Theta_E/T)}{1-\exp(-\Theta_E/T)}\,, \tag{3}$$

where ω is a bond vibration frequency, μ is the reduced mass of the 1NN atomic pair, and $\Theta_E = \hbar\omega/k_B$ is the Einstein temperature. Thus, the total σ^2 in this approximation depends on three parameters: T, Θ_E, and σ_s^2. By replacing the total σ^2 in Eq. (2) by a sum of the dynamic and static terms (Eq. (2)), the best fit results for Θ_E and σ_s^2 can be obtained from a concurrent non-linear least square fitting of Eq. (2) to the EXAFS data taken at various temperatures. From obtaining the best fit values, with their uncertainties, of $n_i, R_i, \sigma_s^2, \sigma_d^2$ and $\sigma_i^{(3)}$, one can significantly reduce the number of possible models of cluster structure and its interaction with the environment (thermal effects, gas atmosphere effects, cluster-substrate, cluster-adsorbate interactions, etc.) EXAFS, however, only yields ensemble-averaged information; thus its application is most valuable for structure refinement of size-controlled nanoclusters.

There are several approaches that allow one to estimate cluster size from EXAFS results. The most popular method is using the first nearest neighbor (1NN) coordination number n_1 that can be directly obtained in EXAFS analysis. These data can be compared against *exact* cluster models or approximate expressions. For the first approach, calculations of Montejano–Carrizales et al. [71] are useful as they obtained geometrical characteristic of several regular polyhedral clusters (cubo-octahedral, icosahedral, body-centered cubic and simple cubic) analytically as a function of the cluster order L. Defining $L = N_E - 1$, where N_E is the number of atoms along the edge of a regular polyhedron, the following relationships can be derived for the 1NN coordination numbers in cuboctahedral (closed packed) and icosahedral (non-closed packed) clusters that have the same sequence of their magic numbers ($N = 13, 55, 147, 309, 561, 923\ldots$): [71]

$$n_1^{co} = \frac{24L\left(5L^2+3L+1\right)}{10L^3+15L^2+11L+3}\,, \tag{4}$$

$$n_1^{icos} = \frac{6L\left(20L^2+15L+7\right)}{10L^3+15L^2+11L+3}\,.$$

Calvin et al. [72] used the second approach by approximating the cluster shape as a sphere with radius R and obtaining the 1NN coordination number for a cluster with average 1NN distance r as follows:

$$n_1 \approx \left[1-\frac{3}{4}\left(\frac{r}{R}\right)+\frac{1}{16}\left(\frac{r}{R}\right)^3\right] n_1^{bulk}\,.$$

This approach is advantageous for larger clusters, and when the size distribution is relatively broad.

The method of measuring of 1NN coordination numbers for size determination in supported metal particles dates back to the end of 1970s [73,74,75,76,77,78,79,80,81,82,83,84]. An early publication detailing an analysis of the structures of supported Pt nanoparticles by EXAFS, found a strong correlation between the measured first nearest neighbor (1NN) metal coordination number and the disorder of their bond lengths; lower coordination numbers appeared to correlate strongly with increased measured disorder in the first-shell metal-metal bond lengths [73]. These effects would be most sensitively seen in particles of the smallest size. The importance of anharmonic corrections to the 1NN pair interaction potential had yet to be appreciated in this pioneering work. Neglecting this effect leads to a non-physical decrease in 1NN bond lengths measured at high temperatures. Enhanced disorder in the 1NN bond distance was later correlated with the influences that result from cluster-support interactions [79]. These studies provide an important insight into the nature of supported metal nanoparticles, namely that the bond lengths of surface atoms should exhibit enhanced structural disorder.

Another approach, to use size-dependent 1NN distance change as a measure of cluster size, was used by Frenkel et al. [61] to obtain sizes of alkanethiolate-stabilized Au nanoparticles. Such changes were first observed by Mays et al. [85] and attributed to surface tension (ST) in the framework of liquid drop model of nanoclusters. To analytically estimate particle diameter, d, the ST method utilized the equation:

$$d = \frac{4}{3} f_{rr} K \alpha\,,$$

where f_{rr} and K are the surface stress and compressibility in the bulk, and $\alpha = \Delta R/R$ is the relative lattice contraction that can be measured by EXAFS. Montano et al. [86,87,88] studied Fe, Cr, Ag and Cu nanoclusters by EXAFS and observed contractions of the average interatomic distances relative to the bulk in all cases.

An alternative approach by Jiang et al. [89] is based on the surface multiplayer relaxation phenomenon, proposed by Finnis and Heine [90]. According to their model, cluster interatomic distances can be calculated for an arbitrary polyhedral shape, using known experimental relaxation data for macroscopic crystal surfaces.

The 1st shell methods described above are based on analytical calculations of the coordination numbers or distances as a function of particle size. However, when clusters are asymmetric, these methods are not accurate since

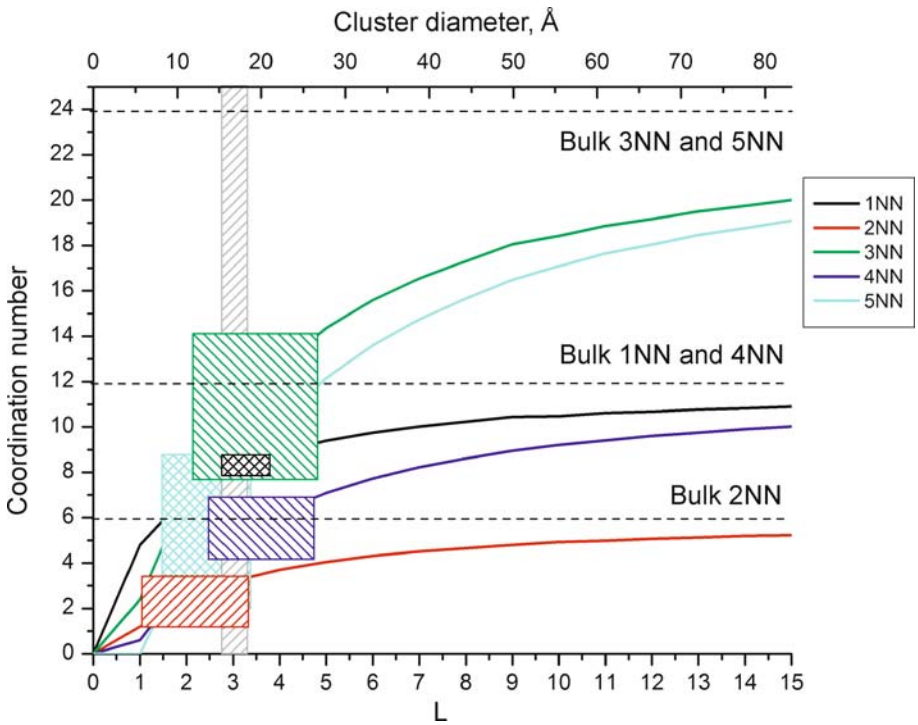

Nanoscale Atomic Clusters, Complexity of, Figure 2

Cluster size, shape and structure determination: comparison of the average coordination numbers (up to 5NN), together with their error bars, measured by EXAFS for the carbon-supported Pt nanoparticle sample and those predicted from the truncated cuboctahedron model for various cluster sizes. Reprinted with permission from [68]

model values of n_1 may be similar when not only the size, but the shape and structure of the cluster are allowed to vary. Only when coordination numbers (or path degeneracies) corresponding to the more distant shells and multiple-scattering contributions are measured, one will have a series of indices: $\{n_i\}$, $i = 1, 2, 3 \ldots$ which should be unique for any given polyhedral cluster model. Since multiple-scattering contribution to EXAFS in nanoclusters can be quantitatively analyzed, [91] not only the single-scattering coordination numbers but also multiple-scattering path degeneracies could be reliably extracted from the EXAFS data. Such additional information effectively communicates size, shape and surface orientation of nanoclusters (Fig. 2) [68].

Even though the methods above can be used to describe metal-metal coordination numbers and, therefore, size, shape and structure information in monometallic and heterometallic clusters, the latter are characterized by significantly higher level of complexity, due to their chemical heterogeneity.

In bimetallic clusters, two distinctly different types of mixing of A and B atoms are possible. They can be mixed statistically (i. e., randomly, in accordance with the overall concentration) or non-statistically. The most common example of non-statistical mixing is segregation of atoms of different elements, forming a core-shell-type particle where larger than the concentration-weighted average number of atoms of one type can be found in the core, and the other type – at the surface of the particle. In heterogeneous samples, where different clusters (A-rich and B-rich) can be formed, the situation may be further complicated [92].

For *random* alloys, the average coordination numbers n_{AA} and n_{AB} of A and B atoms relative to A atom are in the same proportion as the bulk concentrations of these elements in the sample:

$$\frac{n_{AA}}{n_{AB}} = \frac{x_A}{x_B} .$$

For alloys with nonzero short range order, the left part may be *larger* or *smaller* than the right part, indicating *positive* or *negative* tendency to clustering, respectively. In the former case, the atoms A and B segregate to different regions of the nanoalloy. In the latter, the A atoms are preferentially coordinated with B (with probability greater than x_A/x_B) and vice versa.

We can also introduce a short range order parameter, α, analogously to its definition by Cowley for bulk alloys: [93]

$$\alpha = 1 - \frac{n_{AB}/n_{AM}}{x_B} ,$$

where $n_{AM} = n_{AA} + n_{AB}$ is the coordination number of the A-metal bonds. For alloys with positive or negative tendency to clustering, α will be positive or negative, respectively. However, even after the segregation is demonstrated by examining the experimental values of n_{AA}/n_{AB} or α, more experimental information is still needed to find out whether A is predominantly in the surface or in the core, as well as for the determination of the particle size.

Such information is available by measuring EXAFS on both A and B central atoms and extracting coordination numbers n_{AA}, n_{AB} and n_{BB}. The analysis should be done concurrently, with obvious constraints imposed on the heterometallic bonds during the fits: [94]

$$n_{AB} = \frac{x_B}{x_A} n_{BA}; \quad R_{AB} = R_{BA}; \quad \sigma^2_{AB} = \sigma^2_{BA}$$

The atoms of the type A will segregate to the surface of the nanoparticle and B – to the core, if $n_{AM} < n_{BM}$, since atoms at the surface have fewer neighbors than those in the core. This criterion is useful even for alloys containing elements that are neighbors in periodic table (e. g., Fe-Ni, Pd-Ag, etc.) where only the total n_{AM}, n_{BM} numbers can be measured by EXAFS analysis of A and B absorbing atoms, respectively, due to the similarity of backscattering amplitudes of A-A and A-B pairs (as well as B-A and B-B). Another advantage of analyzing both A and B EXAFS data is for the particle geometry determination. Indeed, the average number of metal-metal neighbors per metal atom:

$$n_{MM} = x_A n_{AM} + x_B n_{BM}$$

for the first nearest neighbor shell, combined with other information (e. g., higher shell coordination numbers, transmission electron microscopy (TEM) data, etc.) allows one to estimate the particle size by methods similar to those described above for monometallic particles.

Random bimetallic alloys have a unique behavior of these coordination numbers with concentration. Assume, for simplicity, a bimetallic nanoparticle of a certain size, with random distribution of A and B atoms, where the following relationships apply:

$$n_{AM} = n_{BM} = n_{MM}; \quad n_{AA} = n_{BA} = x_A n_{MM};$$
$$n_{BB} = n_{AB} = x_B n_{MM} = (1 - x_A) n_{MM} .$$

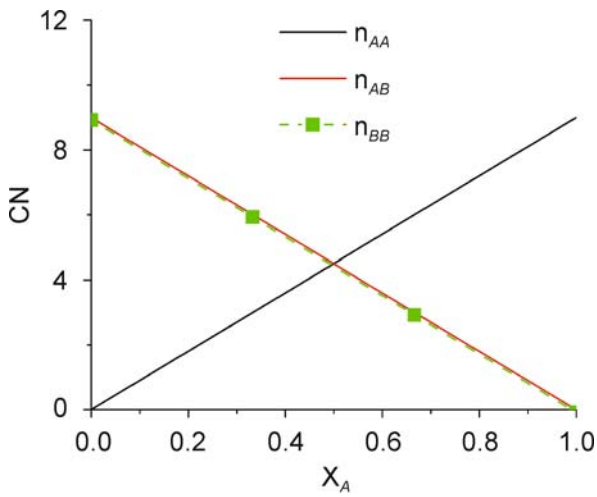

Nanoscale Atomic Clusters, Complexity of, Figure 3
Theoretical partial coordination numbers in random nanoalloys (assuming $n_{MM} = 9$) as a function of composition

Thus, partial coordination numbers should depend *linearly* on alloy concentration (Fig. 3) in *random* nanoalloys, provided that the particle size is the same at all concentrations.

Structure Determination by Advanced Methods of Transmission Electron Microscopy

As mentioned above, while multiple-scattering (MS) EXAFS allows determination of the bonding geometry in the nanoscale, it does so in the form of ensemble-average information. Advanced TEM and scanning TEM (STEM) methods, on the other hand, can probe the atomic-level structure and chemistry of individual nanoparticles and allow direct visualization of nano-scale phenomena [59]. Modern electron microscopy is a powerful tool which, if used in conjunction with EXAFS, can provide a wealth of information about the samples which otherwise could not be determined if these techniques were not used synergistically.

Using "*Z*-contrast" protocol in STEM, one can correlate the absolute image intensity to the scattering cross-section, so that the number of atoms can be counted experimentally. Furthermore, high-resolution electron microscopy (HREM) can visualize the facets and shapes of nanoparticles and the interfacial structures between the nanoparticles and their supports. In this way, a full mapping can be made of the structural landscape present in a complex heterogeneous sample.

Transmission electron microscopy has been used extensively and successfully to study nanoparticles [95]. It

has a higher level of accuracy than other methods for samples containing small number (< 100) of atoms [59]. Specifically, TEM methods have been used to examine clustering sites as well as the shapes of various clusters [95]. For example, one specific problem of interest in current heterogeneous catalysis research is determining the elemental distribution in a bimetallic catalyst particle, e. g., homogeneous distribution of elements or core-shell structures with separation of elements to the core and surface of the particles. In addressing these questions, TEM yields a wealth of information. Conventional TEM is capable of direct visualization at the resolution level of < 0.3 nm in standard instruments and < 0.1 nm in aberration-corrected microscopes. Such resolution allows the investigation of structure of catalysts at atomic level. Transmission electron microscopy uses a parallel beam of electrons, whereas analytical TEM, or STEM, methods employ a scanning attachment on the instrument that focuses the electron probe to enable energy dispersive spectroscopy (EDS) and electron energy-loss spectroscopy (EELS) detection to study elemental distribution as well as oxidation states from specific nanostructures in the sample. For example, a VG-STEM is an ultrahigh vacuum (UHV) instrument, and very fine probe (~ 0.5 nm diameter) with the ability to simultaneously image (Z-contrast) and collect EDS and EELS to acquire both structure and elemental distribution information.

One of the most common TEM techniques is HREM, which is based on phase contrast, and has been used extensively in nanoparticle studies as well. Thomas and Midgley [96] note that the two major techniques in utilizing HREM for atomic and nano-scale structures are the multislice method and basic electron diffraction studies which are analogous to x-ray crystallography. In using the multislice method one essentially matches theoretical quantum mechanical calculations of the electron diffraction patterns of the sample with the closest data set taken in order to infer information on the electronic and crystal structure of the sample [96].

The method works by essentially dividing the reciprocal space into "slices" with an "electron field/wave function" corresponding to each slice. The nth field of the nth slice is given by [97]:

$$\phi_{n+1}(b - \Delta b) = [\phi_n(b) \cdot q_n(b)] * \left(\frac{k_z}{z}\right) p(b)$$

where b is a reciprocal space vector and Δb is the shift in the origin of the wave function and the surface at each slice. $q_n(b)$ is the "phase grating" and $p(b)$ is the free space propagator. The asterisk represents the convolution of the two functions [97].

Transmission electron microscopy has an additional advantage over other techniques such as atomic force microscopy (AFM) in that the analysis of x-rays created by the electron beam/sample interaction can be used for elemental mapping [96]. The physical and chemical structure of the sample, e. g., bond structure (or electronic structure) can be obtained via analysis of the EELS. From the energy spectra of scattered or transmitted electrons, one can determine specific processes that occur in the sample, i. e., plasmon resonances and phonon excitations. These in turn offer a way to determine the electronic structures which give rise to such events in the material. The techniques include energy loss near edge structure (ELNES) which contains the chemical bond information. One can also coordinate the various states of excited atoms using the extended energy loss fine structure (EXELFS), a technique similar to EXAFS.

Determining the density of states and electronic structures of material interfaces is of particular importance in catalysis because the electronic structure at the interface governs aspects of adsorption, charge transfer and bonding. Thus EELS can be viewed as complementary to the other techniques mentioned above in allowing a detailed understanding of the chemical properties of individual nanoparticles [95] as well as their interaction with the substrate surface [98]. For example, Liu found that by using EELS analysis on a Pd-Ni bi-metallic system he could determine structural information about the thickness of Pd shell layer [98].

Perhaps the most important set of tools in nanocatalysis studies are incoherent elastic scattering techniques that are usually performed at large scattering angle. In the case of the high-angle scattering signals the incoherent scattering intensity of the signals depends on the atomic number, Z, as described by Rutherford scattering theory, which is ideal for imaging heterogeneous catalytic materials where high Z metals, e. g., Pt, are dispersed on a low Z support, e. g., alumina. Quantitative STEM which is based on high-angle annular dark-field (HAADF), or quantitative Z-contrast imaging, is the technique based on the acquisition of these very high scattering signals (> 96 mrad) [95]. HAADF is especially powerful in catalysis studies because the electronic structure and catalytic activity are directly related to the number of atoms in the particle [95]. Because many particles can be imaged simultaneously by HAADF, then this quantitative STEM method provides better statistical information, such as particle size distributions [95]. A major advantage of STEM over HREM is that the technique readily allows for EELS analysis, and x-ray emission analysis at the nano-level where scattered electrons can be simultaneously collected [96,99].

Z-contrast exploits nonlinear interference effects between the wave-packets of the scattered electrons [99]. The intensity of scattered electrons can be obtained from the expression for Rutherford scattering cross section:

$$\frac{d\sigma}{d\Omega} = \frac{1}{16}\left(\frac{Z^2 e^2}{p^2/2m}\right)\left(\frac{1}{\sin^4(\theta/2)}\right)$$

Since the differential cross-section of the scattered electrons is proportional to Z^2. A large contrast exists between the supporting structures and the nanoclusters in the image. By collecting electrons scattered to very high angles (> 96 mrad), one can suppress the in-column coherent such that the intensity of the image is simply the number of atoms in the cluster times the intensity expected for one atom of that element. Individual atoms in a particle within the sample may be counted provided that the detector has been properly calibrated. When the calibration is done, one then simply divides the total measured cross section by the cross section of an individual atom [100].

As can be seen in the Fig. 4, the contrast between the standard bright-field (BF) STEM images and the Z-contrast images can be striking. The images are those of monolayer protected gold nanoparticles. There is a large contrast in the HAADF image where one can distinguish between two elements of greatly varying atomic numbers, namely the higher atomic number gold versus carbon [101]. In that particular study Z contrast was successfully employed to determine particle diameters of the various gold clusters as well as count atoms in individual particles [101].

Since the scattered signal is directly proportional to the number of atoms in the scattering center, one can calculate particle diameters measuring both the intensity profiles of the gold and then subtracting the profile of the substrate/supporting material alone, resulting in that of the gold particles [101]. Intensity profiles across the individual gold particles can then be plotted and measured at full width at half maximum (FWHM). The relationship between the number of atoms in the particle and its diameter is characteristic of the particle's shape. By comparing the relative intensities as a function of particle size, one can infer the nucleation and growth behavior of the particles. In general, the measured intensity and particle growth modes are intimately related [101]. If the intensity is directly proportional to the diameter cubed then it can be inferred that the particles are growing uniformly in three dimensions. However, if the intensity is proportional to the diameter

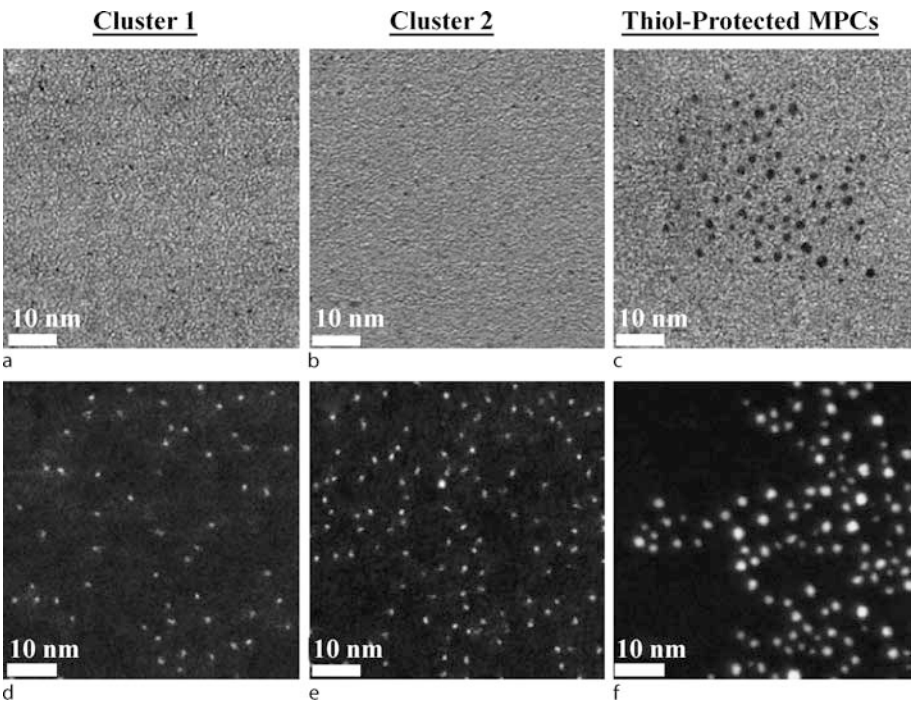

Nanoscale Atomic Clusters, Complexity of, Figure 4

Representative BF-STEM images of a mixed ligand (thiols/phosphines) Cluster 1, b mixed-ligand Cluster 2, and c thiol-protected clusters. Representative HAADF-STEM images of d Cluster 1, e Cluster 2, and f thiol-protected clusters. All images collected at 1 M× magnification. Parts a and d are images of the same area collected simultaneously. Reprinted with permission from [100]

squared, the particles are seen to be forming a two-dimensional mono-layer or bi-layer [102].

The annular dark-field detector is calibrated by tilting the beam directly onto the HAADF detector, such as by using micro-diffraction mode, with an attenuated electron beam current in order to prevent saturation of the detector. In calibrating the dark-field (DF) detector the current is reduced by a factor of 100. The image of the HAADF detector provides the angular dependence of the detector efficiency, which is taken into account when quantifying the experimental electron scattering [100]. The nanoparticles are then imaged using the very high angle annular dark-field detector such that electrons scattered above ~ 100 mrad contribute to the image intensity. The number of atoms can be quantified as follows:

$$N_0^{\text{high}} = \left(\frac{\sum_{\theta=\beta_{\min}}^{\theta=\beta_{\max}} N_0^{\text{low}}(\theta) f(\theta)^2 2\pi \sin\theta \, \mathrm{d}\theta}{\sum_{\theta=\beta_{\min}}^{\theta=\beta_{\max}} f(\theta)^2 2\pi \sin\theta \, \mathrm{d}\theta} \right) \left(\frac{I_{\text{high}}}{I_{\text{low}}} \right)$$

where $\beta_{\min}$ is the inner detector angle (130 mrad), $\beta_{\max}$ is the outer detector angle, N_0^{low} is the detector response at the attenuated current, $f(\theta)$ is the atomic electron scattering factor of the element, I_{high} is the current of the electron beam used to image the particles, and I_{low} is the attenuated beam current used to image the detector for calibration [100]. Once the calibrations have been performed the intensity of scattered electrons from the nanoparticles can be measured and the background signals can be subtracted. Once the particle intensities are determined the data can be used to count the number of atoms.

Menard et al. demonstrated the utility of these techniques in studying gold nanoparticle monolayer protected clusters on carbon supports using the quantitative HAADF method and EXAFS [59,100,102]. By utilizing these techniques in tandem, characterization of the ligand-protected Au_{13} nanoparticle size, shape, and structural disorder was accomplished. with an uncertainty of about three atoms. The combination of techniques allows for a detailed analysis of the transitioning point in terms of numbers of atoms, between classical and quantum behaviors or "bulk metallic and molecular states" as well as supporting the theoretical calculations such as those of the scattering amplitudes [59]. In these studies the HAADF imaging was performed with a field-emission Vacuum Generator HB501 STEM at 100 kV. The "atom counting" images were taken at 1-million magnification and have dimensions of 1024×1024 pixels.

Further HREM measurements were performed utilizing a field-emission JEM 2010F TEM/STEM operated at 200 kV to confirm the size and observe the crystallinity of the individual nanoparticles [59]. The importance of using the quantitative Z-contrast and HREM data is that it provides a set of constraints on the interpretation of corresponding EXAFS data as well as determines the grafting density of thiolates on the surface of the particles [59].

This summary demonstrates that high spatial resolution electron microscopies (such as quantitative Z-contrast) and HREM, are extremely useful in determining the behavior of nanomaterials, including nanoparticles, whose properties are transitional between "bulk" and "molecular" materials.

Cluster Studies by First Principle Theories and Molecular Dynamic Simulations

Theoretical methods also can greatly advance understandings of catalytic materials. First-principles, molecular dynamic (MD) simulations based on density functional theory (DFT) have been advanced to a point now where they can play a significant role in the interpretation of experimental data, revealing critical factors that determine and underlie specific metal cluster properties, see recent reviews [6,7]. (For brevity no attempt is made to include the many pertinent theoretical investigations and new methods, only a few critical papers regarding each key issues.) It immediately becomes more difficult to do this, however, as the cluster size or the chemical complexity (such as due to promoters, defects, adsorbates) of the system is increased, as noted already. In such cases, the DFT search space becomes daunting. Theory then becomes increasingly dependent on the direct experimental input and the availability of computing resources. Even with such capacities, however, the ability to identify relevant low-energy structures lying within a complex, and chemically rich, energy landscape remains a significant challenge and the insights coming from chemical "intuition" remain an important resource for developing useful analytical models.

Examples of Integrated Methods of Cluster Analysis

A recent combination of EXAFS and atom counting methods of transmission electron microscopy (TEM) on such specially-synthesized nanoclusters ($Au_{13}[PPh_3]_4[S(CH_2)_{11}CH_3]_4$) has found them to be highly monodisperse, with their overall structure possessing, on average, 13 gold atoms, with Au-Au coordination number of 6.7 ± 0.7, an average Au-Au bond length of 2.85 ± 0.02 Å, and an average Au-ligand distance of 2.324 ± 0.007 Å [59]. The presence of eight ligands per cluster was deduced from x-ray photoelectron spectroscopy data [100].

The data presented in Fig. 5a,b illustrate such integrated analyses as applied to several Au nanoclusters – the

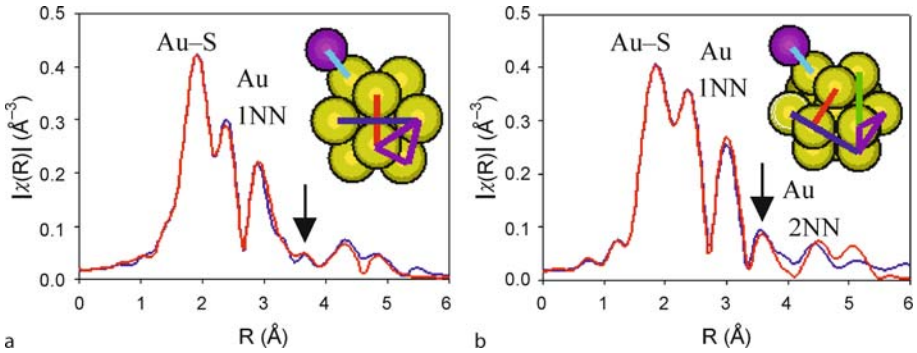

Nanoscale Atomic Clusters, Complexity of, Figure 5
Fourier transformed EXAFS spectra (*blue*) and fits (*red*) for a $Au_{13}(PPh_3)_4(SC_{12})_4$, and b fully-thiolated clusters. The *insets* show icosahedral (a) and truncated octahedral (b) units. The *arrow* indicates the fingerprint of the 2NN path in the closed packed structure (shown *green* in the *inset*) missing in the icosahedral clusters (a) but present in the truncated octahedral ones (b)

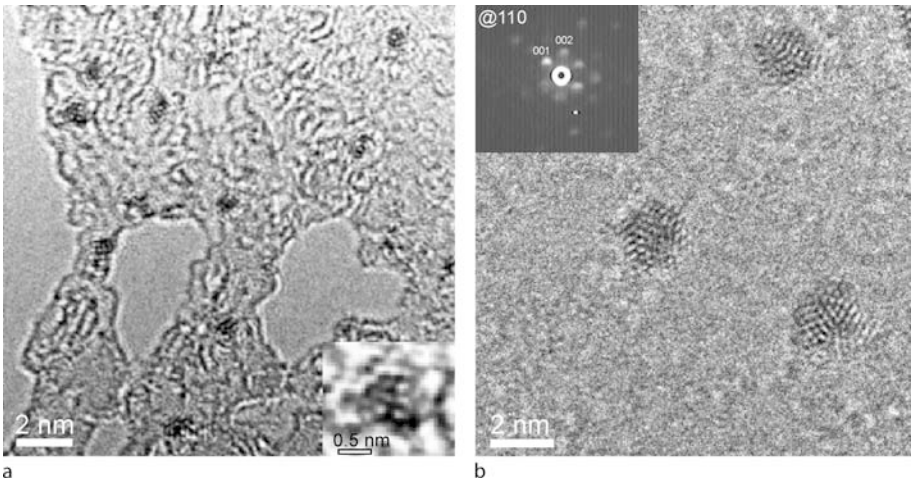

Nanoscale Atomic Clusters, Complexity of, Figure 6
a High-resolution electron micrograph of the mixed-ligand clusters. The *inset* shows a filtered image of a single cluster with icosahedral structure. b Thiol-protected MPCs with cubic packing. The *inset* shows an indexed μ-diffraction image taken from a single particle with a 2 nm diameter. Reprinted with permission from [59]

new mixed-ligand $Au_{13}[PPh_3]_4[SC_{12}]_4$ as well as a monolayer protected cluster (MPC) system comprised of larger thiol-protected particles. The differences seen in the EXAFS data (Fig. 5) are profound and can be directly related to features of the atomic-level bonding present in each case. The mixed-ligand clusters adopt quasi-spherical shapes for their 13 atom cores (Figs. 6, 7). The full analyses unambiguously establish that an icosahedral geometry is present in each case, a bonding motif that stands in marked contrast with the *fcc* truncated octahedral structures adopted by the larger, fully thiolated MPCs (Fig. 5). These interpretations were independently verified by HRTEM and electron microdiffraction measurements (Fig. 6), and *Z*-contrast technique (Fig. 7) which identified the same structural motifs, albeit with less accurate detail [59,100].

While the combination of the experimental results points towards an icosahedral shape of the Au_{13} core, theoretical verification and a detailed interpretation of such a model was lacking. In particular, two central questions – ligand placement and anomalously high Au-Au bond length disorder – were left unanswered by the experimental results. First, EXAFS is not capable of discriminating between Au-S and Au-P neighbors, treating them cumulatively as Au-L (L = S/P) pairs and obtaining the *overall* Au-L coordination number as the total number of Au-S and Au-P bonds divided by the total number of Au atoms. However, the phosphines and the thio-

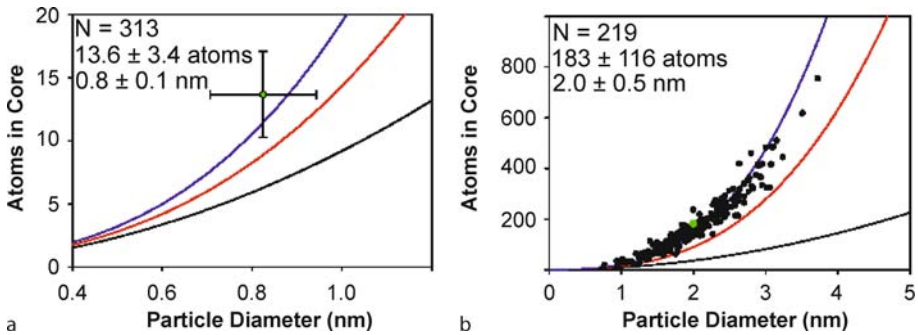

Nanoscale Atomic Clusters, Complexity of, Figure 7
Core atom counts for gold clusters measured using the quantitative HAADF-STEM technique for the mixed ligand (a) and thiolate-protected clusters (b), respectively, plotted against measured particle diameters. The *blue*, *red* and *black lines* are the spherical, hemispherical and monolayer island (111) models. The *black* and *green dots* are the data and the sample average. Reprinted with permission from [100]

lates may have distinctly different bonding motifs: on-top for phosphines [103] and bridge sites for thiolates [104]. Thus, the preferred ligand placement remains ambiguous. Second, the experimental distribution of Au-Au bond lengths was $\sigma^2 = 0.017 \pm 0.005\,\text{Å}^2$, which is much larger than in bulk gold ($0.008\,\text{Å}^2$) at the same temperature (300 K) [59]. Such enhanced σ^2 must be configurational in nature, because the temperature-dependent, dynamic component in nanoparticles has previously exhibited only weak, if any, size dependence [68]. However, neither EXAFS nor TEM provides enough information to uncover its origin.

Guliamov et al. [105] used the EXAFS and TEM results as a starting point for a theoretical analysis based on density functional theory (DFT) [106] of the mixed-ligand Au_{13} nanocluster, $Au_{13}[PPh_3]_4[S(CH_2)_{11}CH_3]_4$. For bare Au_{13} nanoclusters, many structures–including icosahedral [107], cuboctahedral [108], biplanar [109], and amorphous [110] ones-have been predicted theoretically as comprising the lowest energy configuration. Both ordered and disordered structures were predicted theoretically for ligated Au_{38} structures [111,112,113] and experimental verification was limited to scattering methods (e. g., x-ray diffraction) which are less than ideal for clusters of this size. This may reflect the existence of several energetically close isomers [113].

Calculations of different ligand structures relaxed from an initial icosahedral Au core geometry established that the symmetric bonded-ligand configuration, shown in Fig. 8, is energetically preferable. Here, the results of the calculations were found to be in good *quantitative* agreement with the EXAFS data: The mixed on-top and bridge thiol geometry was maintained, the average Au-Au bond length was 2.88 Å with a standard deviation of $\sigma^2 = 0.018\,\text{Å}^2$, the coordination number, $N_{\text{Au-L}}$, was 0.77, and

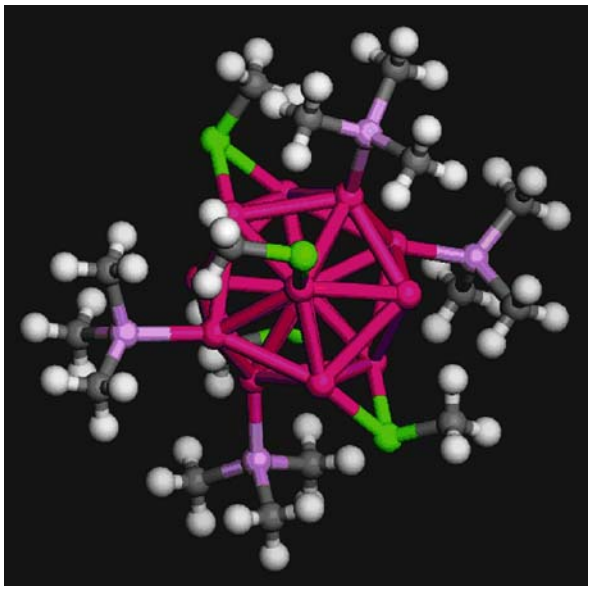

Nanoscale Atomic Clusters, Complexity of, Figure 8
Optimized structure of the icosahedral Au_{13} cluster. *Red*: Au; *Green*: S; *Purple*: P; *Gray*: C; *White*: H

the average Au-ligand distance was 2.35 Å, i. e., in good agreement with all experimental results [59].

Further analysis of the ligand-induced deviation from the ideal icosahedral symmetry of the Au core reveals that the variation in the radial distances from the central atom to the other Au atoms is very small: $R(\text{Au-Au})_{\text{rad}} = 2.78\,\text{Å}$ with $\sigma^2 = 0.005\,\text{Å}^2$. However, the in-shell tangential Au-Au bond lengths exhibit a much greater dispersion: $R(\text{Au-Au})_{\text{tan}} = 2.92\,\text{Å}$ with $\sigma^2 = 0.017\,\text{Å}^2$, with the smallest distance found between two thiol-ligated Au atoms and the largest between one thiol-ligated and one free Au atom. This is consistent with the

strong covalent interaction expected between the S and Au atoms [114,115]. For comparison, we studied the radial versus tangential bond length distribution in the relaxed, bare icosahedral Au_{13}. There, we found $R(\text{Au-Au})_{\text{rad}} = 2.73$ Å and $R(\text{Au-Au})_{\text{tan}} = 2.87$ Å.

The above results clearly reveal that the tangential strain induced by the ligands is much larger than the radial one. Both radial and tangential bond lengths in the ligated cluster are larger by $\sim 1.8\%$ with respect to the bare one, but the induced dispersion in tangential bond lengths is much larger than in the radial ones. These findings can be interpreted via a combination of the asymmetry of the effective pair potential and the non-close-packed structure of the icosahedron. The ligands do not disorder the relatively stiff radial bonds, but do disorder atoms within the shell. The results suggest that it is the weakness of the in-shell Au-Au bonds, and thus lower energy penalty compared to the strong radial Au-Au bonds, that relieves otherwise strong stresses due to the asymmetry in the nature and bonding sites of thiolate and phosphine ligands.

Experimental characterization of randomness of atomic distribution may be found by EXAFS. It is done by comparing experimental values of metal-metal coordination numbers against the model of (Fig. 3). For example, dendrimer-stabilized Pd/Au nanoalloys [116] are shown to be quasi-random or core-shell like, depending on the details of their preparation (Fig. 9).

This method is a powerful tool for quantifying short-range order in monodisperse clusters. However, a broad range of composition is required, to make such determination. If only one composition is available, the answer may be obtained from the combination of EXAFS, advanced electron microscopies (EDS, HREM, STEM) and DFT calculations. For example, EDS method revealed that the distribution of Pt:Ru was uniform, where each nanocluster contained both Pt and Ru. Using the Z-contrast method for determining the number of atoms, Yang et al. [117] discovered that the average measured scattering cross-section corresponded to four $PtRu_5$ groups or 24 atoms. The diameters of these clusters were also measured from the STEM (Fig. 10) as well as by HREM, and the average was 15.6 Å. Using an averaged number density of Ru and Pt, if the shape of the cluster is spherical, then, for a diameter of 15 Å, the particle contains 21 $PtRu_5$ groups, whereas for a hemispherical shape, the cluster contains 11 $PtRu_5$ groups. These data shows that the particles are "raft-like" on the carbon black.

To demonstrate this point, Fig. 11 is a plot of the cluster diameter vs. the number of atoms per cluster for different 3-dimensional shapes. The two theoretical fits show the number of atoms for a (a) spherical shape with diameter, d, and (b) a hemispherical shape. Clearly, since the number of atoms is considerably less than would be predicted for a hemisphere, this demonstrates that the structure of the $PtRu_5$ is oblate on the carbon black support [117]. This result confirms the *truncated* cuboctahedral model proposed by Nashner et al. [94] based on results of their EXAFS analysis.

Wang et al. [118] confirmed EXAFS and electron microscopy results using DFT calculations, including the experimentally observed enhanced Pt-Pt bond length disorder. They revealed the origin of this disorder as due to the cluster/carbon-support interactions when samples are annealed in helium, whereas samples treated in hydrogen have disorder controlled by intraparticle effects, as discussed later. They explained the EXAFS observations that supported [$PtRu_5$] metal clusters have fcc (111) cuboctahedral geometry and bulk-like metal-metal bond distances, even for small nanoparticles for which the average coordination number is much smaller than that in the

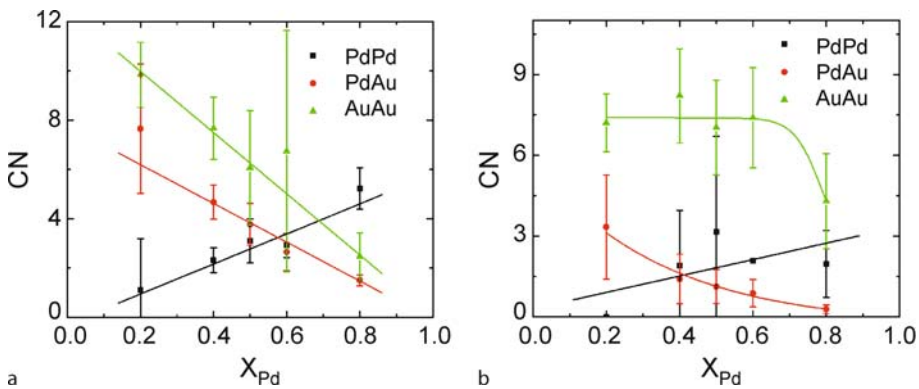

Nanoscale Atomic Clusters, Complexity of, Figure 9
Coordination numbers in nanoalloys: quasi-random dendrimer-stabilized Pd/Au alloys (a), and core-shell alloys (b). Samples courtesy of Crooks RM, U. of Texas at Austin

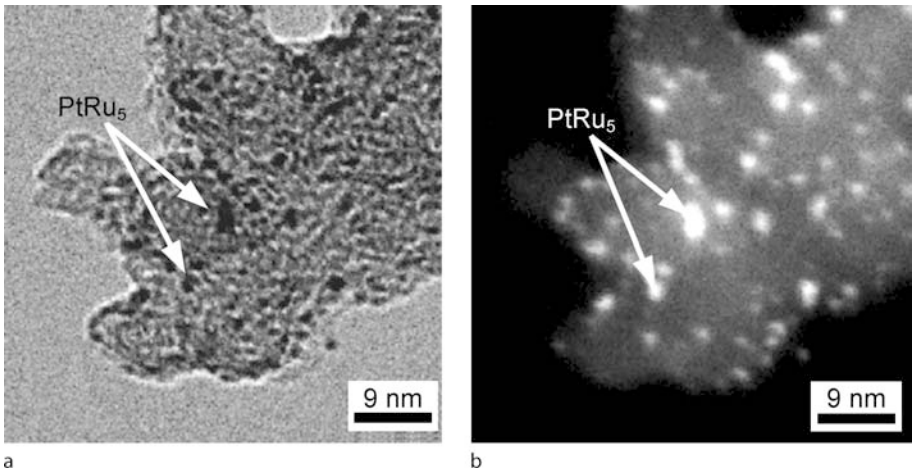

Nanoscale Atomic Clusters, Complexity of, Figure 10
BF image of $PtRu_5$ (a) and the corresponding HAADF image (b). Reprinted from [117]

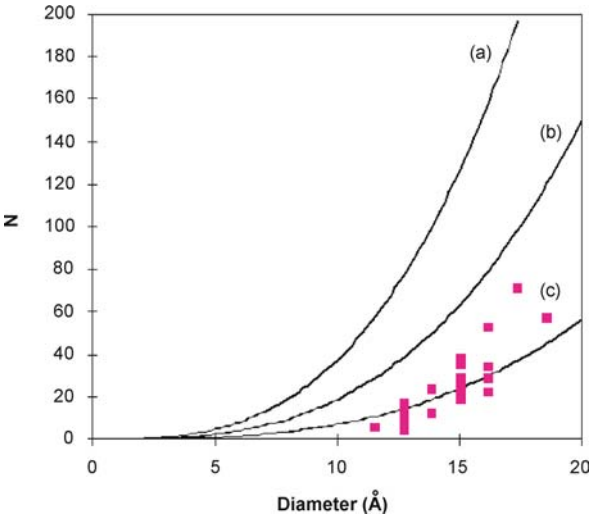

Nanoscale Atomic Clusters, Complexity of, Figure 11
Plot of diameter vs. number of atoms for different shapes: a Sphere, b Hemisphere and c are the experimental data with the best fit where the 3-D aspect ratio is kept constant. Reprinted from [117]

bulk, and that Pt in the bimetallic clusters segregates to the top (111)-layer of the nanoparticle, as hypothesized based on EXAFS results.

Structural Relaxation of Nanoparticles

The data discussed above illustrate one level of new knowledge that comes from direct measurements of atomic scale structure. It is useful, though, to look at additional levels of insight that can be developed from data of this sort. Metal nanoclusters contain many distinct types of atoms – the various habits found at the cluster surfaces, those residing at the metal support interface or cluster core, etc. [119] The metal-metal bonding, as a result of this diversity, is present as an ensemble embedding substantial configurational, static (as opposed to vibrational, dynamic) disorder.

Structural Relaxation in Freestanding and Supported Clusters

The case of the icosahedral Au clusters provides an interesting example in this regard. In this case, one moment of this effective disorder in the first shell bonding can be isolated and this in turn enables one to evaluate separately the nature of the icosahedral strain that it embeds. The latter parameter is one that directly illustrates the nature of the energetics that selects such non-bulk habits for the smallest clusters. The EXAFS analysis (Fig. 5a, inset) revealed the presence of an icosahedral strain of ca. 2.5% as compared to the 5% strain predicted by geometrically. This strain relaxation is correlated with and can be understood in terms of the ligand bonding – the simplest model of a support interaction (see below) – that terminate the cluster's surface. Obtaining such levels of structural understanding, however, is only possible if the synthetic protocols used deliver samples that are comprised of extremely monodispersed nanoparticles, a condition where it becomes possible to discriminate intra vs inter particle disorder (where the latter arises as a result of size-dependent atomic bond-length relaxation [61,85]).

It is also now well understood that more general forms of structural relaxations are common to the energy landscapes of supported metal clusters. The smallest clusters,

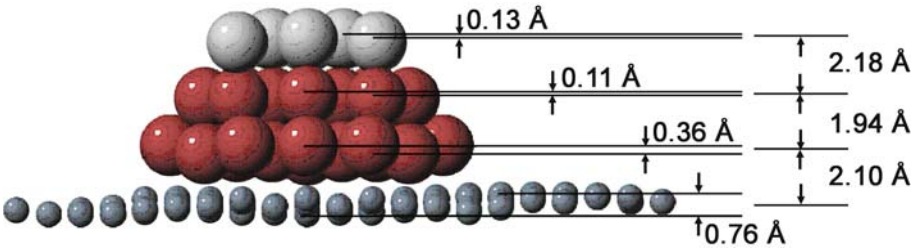

Nanoscale Atomic Clusters, Complexity of, Figure 12
Side view of the relaxed structures calculated with DFT-LDA for PtRu clusters supported on a graphite surface. The *small dark* sphere stands for C atoms, *large light* (*dark*) spheres for Pt (Ru) atoms. Both the intra-layer buckling and inter-layer distance are listed. Reprinted with permission from [118]

for example, are known to show contractions of their metal-metal bonding distances, [85] an outcome that can be understood intuitively in the context of the large number of non-bulk like atoms that would be found in a cluster that is 1nm in size – the relative fractions of surface and support-bonded atoms being large in such cases.

The data in Fig. 13 illustrate this point in greater detail, highlighting EXAFS measurements made on a series of rigorously characterized Pt catalysts supported on γ-Al_2O_3 [9]. As synthesized, the cluster distributions of atomic mass were held to exceptionally low values (a property stiffly characterized by microscopy). The data show the strong size dependent scaling of the average first-shell Pt-Pt bonding distances found in these samples. These distances are plotted here in terms of the number of first shell neighbors found experimentally by EXAFS, a parameter that explicitly correlates with both the particle's shape and diameter (as discussed for the data presented in Fig. 3, and for these samples showing by microscopy diameters spanning a range from 0.9 to 2.8 nm). The sample with the smallest Pt particles (0.9 nm) consisted of a highly homogeneous sample of Pt_{15} clusters (HAADF-STEM). The EXAFS data illustrate that size effects, in this case, elicit strong contractions of the metal-metal bonding distances – a structural relaxation that also develops a correlated and significant degree of interatomic disorder in the catalytic clusters. These relaxations are responsive to the presence of adsorbates – the data in the figure illustrate that hydrogen, which dissociates and passivates the surface bonds of the clusters lifts the relaxations (albeit not fully to bulk values for all but the largest clusters). The larger body of data implicitly suggests a model for the structural relaxations in which the bonding present at the ambient cluster surfaces/interfaces play an exceptionally important role – establishing the biases seen in the average distances and effective disorder measured by EXAFS.

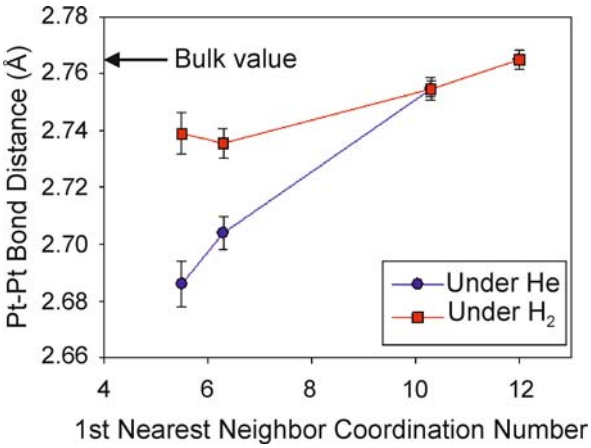

Nanoscale Atomic Clusters, Complexity of, Figure 13
The size-correlated average first shell M-M bonding distances for a series of Pt catalysts supported on γ-Al_2O_3

Advances made in the first principle theory and MD simulations allow one to investigation of structural relaxation and relative stability of various cluster morphologies. Nonetheless, DFT-based simulations must be carefully validated before use in each system. For example, Wang and Johnson [120] have shown that N-atom (for $N \leqslant 20$) cluster morphologies are highly sensitive to the DFT exchange-correlation functional, and standard generalized gradient corrected, local (spin) density approximation often yield incorrect structural properties and energetics compared to high-level quantum chemistry. In addition, with the numerous geometrical and chemical configurations associated with finding ground state and lowest-energy excited-state morphologies (e. g., as in Fig. 1), one cannot be assured with absolute certainty that a simulation has found the absolute ground state for comparison to observation because the simulation times are short (order of nanoseconds to pico-seconds) and starting geometries are few. In this regard, the most commonly employed global-optimization strategies are simulated annealing (sometimes combined, e. g., see [4] with empirical

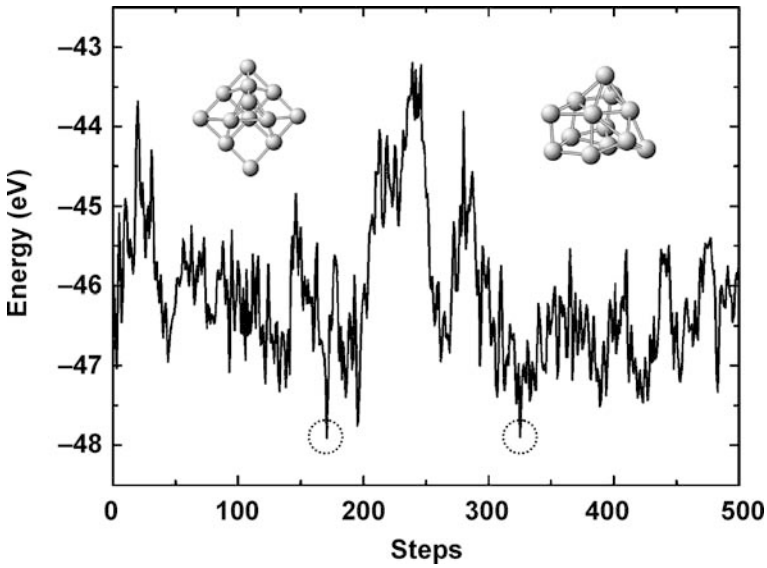

Nanoscale Atomic Clusters, Complexity of, Figure 14
First-principles MD global (potential energy surface) optimization for 500 time steps for a Pt_{13} cluster that starts with an O_h cluster. Simulation was held at 2000 K for 4 to 10 picoseconds at time steps of 20 femtoseconds and used a single k-point in a periodic box with large vacuum. Two low-energy basins are *circled*, with their corresponding structures shown in *insets* after full relaxation of the cluster. All structures in Fig. 1 were found using this approach. Reprinted figure with permission from [3]

force-fields not containing proper cluster correlation effects), genetic and evolutionary algorithms (e. g., see example review [121] and with improved GAs [122]), and variants of basin-hopping that utilize temperature as a tunable parameter to sample large, but not all, portions of solution space (e. g., Monte Carlo [123] and molecular dynamics [3,4] using very small (not converged) k-meshes). Genetic algorithms and basin hopping are comparable in efficiency, both superior to simulated annealing. An example of basin hopping within MD is shown in Fig. 14, and used to obtain results in Fig. 1.

An illustrative example of DFT studies of structural relaxation and morphological transitions in nanoparticles is shown in Fig. 15, where Wang et al. [118] were able to characterize fully the adsorbate-dependent morphological habits of a model Pt_{37} cluster supported on carbon. Importantly, for comparison to experiment, consideration was made for the annealing stages in the synthesis, including post-processing in hydrogen atmosphere before characterization. With an anneal in hydrogen, the preferred structural habit for the cluster is a truncated cuboctahedron with large Pt-Pt bond length disorder in the bond distribution (from 2.62 to 2.93 Å) that arises from structural anisotropy within a cluster. This disorder is highest for bare clusters and is reduced when hydrogen atoms are allowed to adsorb. The disorder seen in each case provides a near quantitative agreement with values deduced from experiment [118,124]. These DFT calculations further reveal that the Pt_{37} truncated cuboctahedron (TCO, shown in Fig. 15a) has a morphological instability driven by the shearing of (100) to (111) facets to lower the surface energy (see Fig. 15b) – a remnant of the electronic mechanism responsible for (100)-surface reconstruction in semi-infinite bulk Pt. [124] However, with H passivation, this shear instability is removed and the TCO is highly stable, [124] as observed in experiments [94] and which is clearly a result of the annealing in H_2.

Theory also predicts that M-M bond length contractions occur – ones that vary substantially with the presence of adsorbate bonding. They found, for example, that the DFT derived average first nearest neighbor Pt-Pt bond length for the model Pt_{37} cluster increases by 3% with H passivation (2.68 Å without H to 2.75 Å with H) [124]. These values provide a remarkable agreement with those determined experimentally (EXAFS 2.75 Å) when samples are annealed in hydrogen [68] versus helium [9]. Importantly, H-passivation effect on the Pt-Pt bond lengths are strongly dependent on the size of the cluster (both in simulations and observed); for smaller Pt nanoparticles with 1st shell coordination number (CN) from 5 to 7, an increase in Pt-Pt bond lengths is observed on various supports, whereas, for larger clusters with CN beyond 8 (i. e., approaching bulk CN of 12), no such effect is observed. The reason being that the relation between 1st shell CN and particle size is not linear, but roughly proportional to the $\sqrt{D}$, the particle diameter (D= 2R).

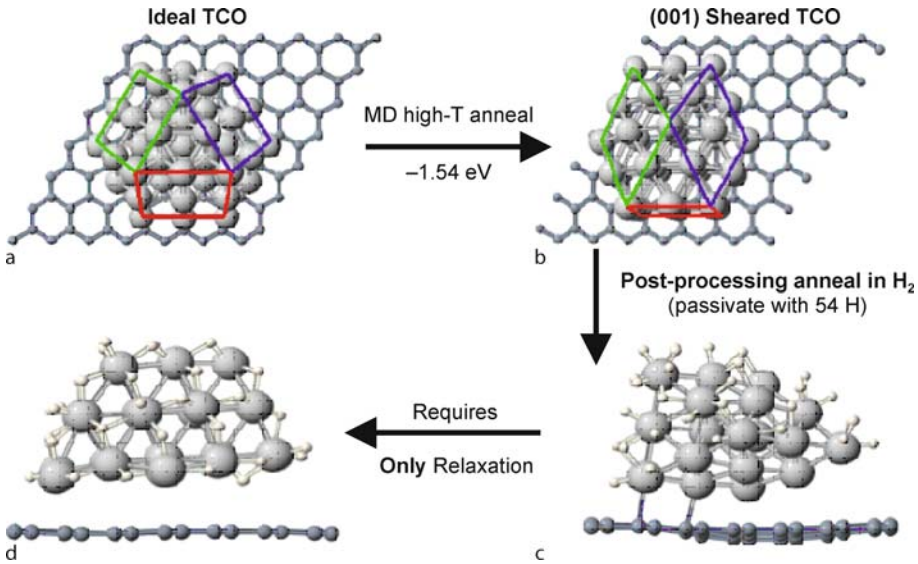

Nanoscale Atomic Clusters, Complexity of, Figure 15
Synopsis of the calculated DFT morphological changes of the structure of a Pt_{37} cluster on a carbon support during high-temperature anneal and post-processing anneal in H_2 performed during synthesis of real samples, see [124] for details. a High-T anneal of the ideal cuboctahedral (TCO) cluster, which b lowers its energy by internal shear of all (100) facets (shown by the *red*, *green* and *blue planes*) such that they resemble (111) facets, a mechanism which is a remnant of the electronically-driven shear instability for surface reconstruction in semi-infinite bulk Pt. For direct comparison to experiment, c the structure found in simulated post-processing anneal in H_2, which d structurally reverts simple by relaxation (downhill in energy) to the observed cuboctahedral structure, showing then negligible interaction with the carbon support. For structural comparison to experiment, see text. Reprinted with permission from [118]

Clearly, careful implementation and validation of theory not only shows excellent agreement with a range physical properties (from bond-length distributions, structural parameters, electronic properties, etc.) but also reveals the key physics controlling these manifested behaviors, including the important of cluster-support interactions and charge transfer that impacts the structural and concomitant catalytic behavior, so highly prized for technological purposes.

Future Directions

With respect to the atomic arrangements present in nanoparticles, the availability of a new state of the art instruments for studies by electron microscopy and x-ray absorption spectroscopy provides an exceptional opportunity to gain new levels of understandings of the structural motives present in these systems. We believe that direct imaging of the 3D atomic arrangement present in such clusters will become eventually possible. When the future synchrotron sources with exceptional brightness (equivalent to the counting statistics of 10^{12} photons/sec at the 1nm focal spot) become available, it will become possible to measure structural dynamics in individual nanoparticles by EXAFS. With the newest electron microscopes (e. g., JEOL (JEM) 2200F), high-resolution electron microscopy imaging will be possible with dramatically better resolution and, hence, improved interpretability. Since the spherical aberration causes delocalization in the images proportional to Θ^3, where Θ is the Bragg angle of the diffracted beam, the reduction of the spherical aberration will eliminate the presence of "ghost" images and/or fringes, leading to significantly improved clarity of images. The focal series reconstruction of exit wavefunction can further improve the HREM interpretability. Because of the ability to obtain spectral information atomic-by-atomic column, changes in the EELS signal across a single nanoparticle, from the surface to the interface, will be achievable. This will provide detailed information about individual arrangement of atoms within an individual particle, such as core-shell structure or surface segregation. Furthermore, with the significantly improved HAADF-STEM imaging resolution to 1Å, significantly better than the 1 nm^3 tomography resolution available on modern non-aberration corrected instruments [125]. Additional possibilities that are being engendered by methods of coherent electron diffraction further suggest that it may be possible to vastly exceed the capabilities afforded by even

the very advanced capacitiles of the JEOL 2200F instrument for electron tomography. For example, via over-sampling and phase retrieval it will be possible to invert a measured single cluster nano-diffraction pattern to obtain an atomic resolution 3D structure directly [126].

Current theoretical methods will be extended by enabling DFT-based calculations to identify low-energy structures, utilizing a rapid – optimization and selection protocols that appear to be well suited to applications involving supported nanoscale materials. An example is the DFT-based MD approach, in which the simulated nanocluster is heated to high temperature (e. g., 2000 K) and simulated rapidly using a coarse k-space mesh, is used to identify (relatively) rapidly several low-energy conformations, see Figs. 1 and 14. These conformations are then revisited with DFT-MD with a fine k-space mesh and configurations are then identified after full ionic relaxations. In the future, such protocols will be integrated with improved DFT exchange-correlation functionals to explore the nature of the energy landscapes that define the accessible and correct structures for a large variety of classes of nanoscale materials. Note that improved DFT exchange-correlation functionals [127,128,129,130,131,132,133,134,135] (such as hybrid B3LYP or PBE0, WC, and HSC) are computationally much more expensive than standard gradient corrected or local density approximation but yield much more correct results for nanoparticles, surface interactions, and some solid phases, such as ferroelectronics, which are highly sensitive to volume errors.

Complexity of nanoscale clusters can be quantitatively understood. However, due to the complexity of the characterization approaches that such understanding demands, only self-consistent, integrated, multi-disciplinary analysis methodologies appear to be heading in the direction of systematically solving their mysteries.

Acknowledgments

We are grateful to S. Sanchez, L. Li, A. Urban, L.-L. Wang, F. Dukesz and Q. Wang for help in preparing this manuscript. We would like to thank R.M. Crooks, M. Knecht and M. Meir for sharing samples used in these studies. We acknowledge support by the US Department of Energy (DOE) Grant No. DE-FG02-03ER15476.

Bibliography

Primary Literature

1. Lai SL, Guo JY, Petrova V, Ramanath G, Allen LH (1997) Size-dependent melting properties of small tin particles: nanocalorimetric measurements. Phys Rev Lett 77:99–102
2. Jesser WA, Shneck RZ, Gile WW (2004) Solid-liquid equilibria in nanoparticles of Pb-Bi alloys. Phys Rev B 69(14): 144121
3. Wang LL, Johnson DD (2007) Density functional study of structural trends for late-transition-metal 13-atom clusters. Phys Rev B 75:235405
4. Futschek T, Marsman M, Hafner J (2005) Structural and magnetic isomers of small Pd and Rh clusters: an ab initio density functional study. J Phys Condens Matter 17:5927–5963
5. Futschek T, Hafner J, Marsman M (2006) Stable structural and magnetic isomers of small transition-metal clusters from the Ni group: an ab initio density-functional study. J Phys Condens Matter 18:5703–5748
6. Ferrando R, Jellinek J, Johnston RL (2008) Nanoalloys: From theory to applications of alloy clusters and nanoparticles. Chem Rev 108(3):845–910
7. Baletto F, Ferrando R (2005) Structural properties of nanoclusters: Energetic, thermodynamic, and kinetic effects. Rev Mod Phys 77:371–423
8. Daniel MC, Astrue D (2004) Gold nanoparticles: assembly, supramolecular chemistry, quantum-size-related properties, and applications toward biology, catalysis, and nanotechnology. Chem Rev 104(1):293–346
9. Kang JH, Menard LD, Nuzzo RG, Frenkel AI (2006) Unusual non-bulk properties in nanoscale materials: thermal metal-metal bond contraction of γ-alumina-supported Pt catalysts. J Am Chem Soc 128:12068–12069
10. Kamat PV (1993) Photochemistry on nonreactive and reactive (semiconductor) surfaces. Chem Rev 93:267–300
11. Gates BC (1995) Supported metal clusters: synthesis, structure, and catalysis. Chem Rev 95:511–522
12. Shahbazyan TV, Perakis IE (1999) Size-dependent correlation effects in the ultrafast optical dynamics of metal nanoparticles. Phys Rev B 60(12):9090–9099
13. Tsunoyama H, Sakurai H, Tsukuda T (2006) Size effect on catalysis of gold clusters dispersed in water for aerobic oxidation of alcohol. Chem Phys Lett 429:528–532
14. Roucoux A, Schultz J, Patin H (2002) Reduced transition metal colloids: A novel family of reusable catalysts. Chem Rev 102:3757–3778
15. Schmid G (2006) Metal Nanoparticles, Synthesis of. In: Encyclopedia of Inorganic Chemistry, 2nd edn, vol 5. Wiley-VCH, Weinheim
16. Taneja R, Chandra R, Banerjee R, Ayyub P (2001) Structure and properties of nanocrystalline Ag and Cu_2O synthesized by high pressure sputtering. Script Mater 44:1915–1918
17. Spadavecchia J, Prete P, Lovergine N, Tapfer L, Rella R (2005) Au nanoparticles prepared by physical method on Si and sapphire substrates for biosensor applications. J Phys Chem B 109:17347–17349
18. Mafune F, Jun-ya K, Takeda Y, Kondow T (2002) Full physical preparation of size-selected gold nanoparticles in solution: laser ablation and laser-induced size control. J Phys Chem B 106(31):7575–7577
19. Ramachandra Rao CN, Kulkarni GU, Thomas PJ, Peter P Edwards (2000) Metal nanoparticles and their assemblies. Chem Soc Rev 29:27–35
20. Giesen B, Orthner HR, Kowalik A, Roth P (2004) On the interaction of coagulation and coalescence during gas-phase synthesis of Fe-agglomerates. Chem Eng Sci 59:2201–2211

21. Kamat PV, Flumiani M, Hartland GW (1998) Picosecond dynamics of silver nanoclusters. Photoejection of electrons and fragmentation. J Phys Chem B 102:3123–3128
22. Kurita H, Takami A, Koda S (1998) Size reduction of gold particles in aqueous solution by pulsed laser irradiation. Appl Phys Lett 72:789–791
23. Takami A, Kurita H, Koda S (1999) Laser-induced size reduction of noble metal particles. J Phys Chem B 103(8): 1226–1232
24. Hirai H, Nakao Y, Toshima N (1978) Colloidal rhodium in poly(vinylpyrrolidone) as hydrogenation catalyst for internal olefins. Chem Lett 7(5):545–548
25. Kiwi J, Gratzel M (1979) Protection size factors, and reaction dynamics of colloidal redox catalysis mediating light induced hydrogen evolution from water. J Am Chem Soc 101: 7214–7127
26. Chechik V, Crooks RM (2000) Dendrimer-encapsulated Pd nanoparticles as fluorous phase-soluble catalysts. J Am Chem Soc 122(6):1243–1244
27. Harriman A, Thomas JM, Millward GR (1987) Catalytic and structural properties of iridium-iridium dioxide colloids. New J Chem 11:757–762
28. Zhu J, Shen Y, Xie A, Qiu L, Zhang Q, Zhang S (2007) Photoinduced synthesis of anisotropic gold nanoparticles in room-temperature ionic liquid. J Phys Chem C 111(21):7629–7633
29. McGilvray KL, Decan MR, Wang D, Scaiano JC (2006) Facile photochemical synthesis of unprotected aqueous gold nanoparticles. J Am Chem Soc 128:15980–15981
30. Yang S, Wang Y, Qingfeng W, Zhang R, Ding B (2007) Colloids Surf A: Physicochem Eng Aspects 301:174–183
31. Mallick K, Witcomb MJ, Scurrell MS (2004) Polymer stabilized silver nanoparticles: A photochemical synthesis route. J Mat Sci 39(14):4459–4463
32. Lee J, Ryu J, Choi W (2007) Synthesis of gold and platinum nanoparticles using visible light activated Fe(III)-complex. Chem Lett 36:176–177
33. Pei L, Mori K, Adachi M (2004) Formation process of two-dimensional networked gold nanowires by citrate reduction of $AuCl_4^-$ and the shape stabilization. Langmuir 20(18):7837–7843
34. Suslick KS, Choe SB, Cichowlas AA, Grinstaff MW (1991) Sonochemical synthesis of amorphous iron. Nature 353:414–416
35. Didenko YT, McNamara WB, Suslick KS (1999) Hot spot conditions during cavitation in water. J Am Chem Soc 121: 5817–5818
36. Su CH, Wu PL, Yeh CS (2003) Sonochemical synthesis of well-dispersed gold nanoparticles at iced temperature. J Phys Chem B 107:14240–14243
37. Jiang LP, Wang AN, Zhano Y, Zhang JR, Zhu JJ (2004) A novel route for the preparation of monodisperse silver nanoparticles via a pulsed sonoelectrochemical technique. Inorg Chem Comm 7:506–509
38. Dhas NA, Gedanken A (1998) Sonochemical preparation and properties of nanostructured palladium metallic clusters. J Mater Chem 8(2):445–450
39. Kan C, Cai W, Li C, Zhang L, Hofmeister H (2003) Ultrasonic synthesis and optical properties of Au/Pd bimetallic nanoparticles in ethylene glycol. J Phys D 36:1609–1614
40. Fujimoto T, Terauchi SY, Umehara H, Kojima I, Henderson W (2001) Sonochemical preparation of single-dispersion metal nanoparticles from metal salts. Chem Mater 13:1057–1060
41. Huang S, Ma H, Zhang X, Yong F, Feng X, Pan W, Wang X, Wang Y, Chen S (2005) Electrochemical synthesis of gold nanocrystals and their 1D and 2D organization. J Phys Chem B 109(42):19823–19830
42. Rodriguez-Sanchez L, Blanco MC, Lopez-Quintela MA (2000) Electrochemical synthesis of silver nanoparticles. J Phys Chem B 104:9683–9688
43. Starowicz M, Stypula B, Banas J (2006) Electrochemical synthesis of silver nanoparticles. Electrochem Comm 8(2): 227–230
44. Zhou M, Chen S, Ren H, Wu L, Zhano S (2005) Electrochemical formation of platinum nanoparticles by a novel rotating cathode method. Physica E 27(3):341–351
45. Ueda M, Dietz H, Anders A, Kneppe H, Meixner A, Plieth W (2002) Double-pulse technique as an electrochemical tool for controlling the preparation of metallic nanoparticles. Electrochem Acta 48:377–386
46. Plieth W, Dietz H, Anders A, Sandmann G, Meixner A, Weber M, Kneppe H (2005) Electrochemical preparation of silver and gold nanoparticles: Characterization by confocal and surface enhanced Raman microscopy. Surf Sci 597:119–126
47. Toshima N, Yonezawa T, Kushihashi K (1993) Polymer-protected palladium–platinum bimetallic clusters: preparation, catalytic properties and structural considerations. J Chem Soc 89:2537–2543
48. Toshima N, Harada M, Yamazaki Y, Asakura K (1992) Catalytic activity and structural analysis of polymer-protected gold-palladium bimetallic clusters prepared by the simultaneous reduction of hydrogen tetrachloroaurate and palladium dichloride. J Phys Chem 96:9927–9933
49. Han S W, Kim Y, Kim K (1998) Dodecanethiol-derivatized Au/Ag bimetallic nanoparticles: TEM, UV/VIS, XPS, and FTIR analysis. J Colloid Interface Sci 208:272–278
50. Link S, Wang ZL, El-Sayed MA (1999) Alloy formation of gold-silver nanoparticles and the dependence of the plasmon absorption on their composition. J Phys Chem B 103: 3529–3533
51. Sun S, Murray CB, Weller D, Folks L, Moser A (2000) Monodisperse FePt nanoparticles and ferromagnetic FePt nanocrystal superlattices. Science 287:1989–1992
52. Bian B, Hirotsu Y, Sato K, Ohkubo T, Makino A (1999) Structures and magnetic properties of oriented Fe/Au and Fe/Pt nanoparticles on a-Al_2O_3. J Electron Microsc 48:753–758
53. Liu HB, Pal U, Medina A, Maldonado C, Ascencio JA (2005) Structural incoherency and structure reversal in bimetallic Au-Pd nanoclusters. Phys Rev B 71:075403
54. Dadge JW, Islam M, Dharmadhikari AK, Mahamuni SR, Aiyer RC (2006) Hyper-Rayleigh scattering in electrochemically synthesized Ag–Au coupled clusters. J Phys Condens Matter 18:5405–5413
55. Mandal M, Kundu S, Ghosh SK, Pal T (2004) Micelle-mediated UV-photoactivation route for the evolution of Pd_{core}–Au_{shell} and Pd_{core}–Ag_{shell} bimetallics from photogenerated Pd nanoparticles. J Photochem Photobiol A 167:17–22
56. Hirakawa K, Toshima N (2003) Ag/Rh bimetallic nanoparticles formed by self-assembly from Ag and Rh monometallic nanoparticles in solution. Chem Lett 32(1):78–79
57. Kanemaru M, Shiraishi Y, Koga Y, Toshima N (2005) Calorimetric study on self-assembling of two kinds of monometallic nanoparticles in solution. J Therm Analysis Calorim 81: 523–527

58. Wang Y, Toshima N (1997) Preparation of Pd-Pt bimetallic colloids with controllable core/shell structures. J Phys Chem B 101:5301–5303
59. Menard LD, Xu H, Gao S, Twesten RD, Harper AS, Song Y, Wang G, Douglas AD, Yang JC, Frenkel AI, Murray RW, Nuzzo RG (2006) Metal core bonding motifs of monodisperse icosahedral Au13 and larger Au monolayer-protected clusters as revealed by x-ray absorption spectroscopy and transmission electron microscopy. J Phys Chem B 110(30): 14564–14573
60. Wang X, Hanson JC, Frenkel AI, Kim JJ, Rodriguez JA (2004) Time-resolved studies for the mechanism of reduction of copper oxides with carbon monoxide: Complex behavior of lattice oxygen and the formation of suboxides. J Phys Chem B 108:13667
61. Frenkel AI, Nemzer S, Pister I, Soussan L, Harris T, Sun Y, Rafailovich MH (2005) Size-controlled synthesis and characterization of thiol-stabilized gold nanoparticles. J Chem Phys 123:184701–6
62. Sun Y, Frenkel AI, Isseroff R, Shonbrun C, Forman M, Shin K, Koga T, White H, Zhang L, Zhu Y, Rafailovich MH, Sokolov JC (2006) Characterization of palladium nanoparticles by using x-ray reflectivity, EXAFS, and electron microscopy. Langmuir 22(2):807–816
63. Frenkel AI, Frankel SC, Liu T (2005) Structural stability of giant polyoxomolybdate molecules as probed by EXAFS. Phys Scripta T115:721–723
64. Frenkel AI, Menard LD, Northrup P, Rodriquez JA, Zypman F, Glasner D, Gao SP, Xu H, Yang JC, Nuzzo RG (2007) Geometry and charge state of mixed-ligand Au_{13} nanoclusters. AIP Conf Proc 882:749–751
65. See, e. g., Bonačić-Koutecký V, Fantucci P, Koutecký V (1991) Quantum chemistry of small clusters of elements of groups Ia, Ib, and IIa: fundamental concepts, predictions, and interpretation of experiments. Chem Rev 91(5):1035–1108
66. Pyykkö R (2004) Theoretical chemistry of gold. Angew Chem Int Ed 43:4412–4456
67. Gurman SJ (1995) Interpretation of EXAFS data. J Synchrotron Rad 2:56–63
68. Frenkel A, Hills C, Nuzzo R (2001) A view from the inside: Complexity in the atomic scale ordering of supported metal nanoparticles. J Phys Chem B 105(51):12689–12703 (Feature Article)
69. Zabinsky SI, Rehr JJ, Ankudinov A, Albers RC, Eller MJ (1995) Multiple-scattering calculations of x-ray-absorption spectra. Phys Rev B 52:2995–3009
70. Newville M (2001) IFEFFIT: interactive XAFS analysis and FEFF fitting. J Synchrotron Rad 8:322–324
71. Montejano-Carrizale JM, Aguilera-Granja F, Moran-Lopez JL (1997) Direct enumeration of the geometrical characteristics of clusters. NanoStruct Mater 8(3):269–287
72. Calvin S, Miller MM, Goswami R, Cheng SF, Mulvaney SP, Whitman LJ, Harris VG (2003) Determination of crystallite size in a magnetic nanocomposite using extended x-ray absorption fine structure. J Appl Phys 94(1):778–783
73. Lytle FW, Via GH, Sinfelt JH (1977) New application of extended x-ray absorption fine structure (EXAFS) as a surface probe-nature of oxygen interaction with a ruthenium catalyst. J Chem Phys 67:3831–3832
74. Sinfelt JH, Via GH, Lytle FW (1977) Extended x-ray absorption fine structure (EXAFS) of supported platinum catalysts. J Chem Phys 68:2009–2010
75. Via GH, Sinfelt JH, Lytle FW (1979) Extended x-ray absorption fine structure (EXAFS) of dispersed metal catalysts. J Chem Phys 71:690–699
76. Sinfelt JH, Via GH, Lytle FW (1980) Structure of bimetallic clusters. Extended x-ray absorption fine structure (EXAFS) studies of Ru–Cu clusters. J Chem Phys 72:4832–4844
77. Sinfelt JH, Via GH, Lytle FW, Greegor RB (1981) Structure of bimetallic clusters. Extended x-ray absorption fine structure (EXAFS) studies of Os–Cu clusters. J Chem Phys 75:5527–5537
78. Sinfelt JH, Via GH, Lytle FW (1982) Structure of bimetallic clusters. Extended x-ray absorption fine structure (EXAFS) studies of Pt–Ir clusters. J Chem Phys 76:2779–2789
79. Marques EC, Sandstrom DR, Lytle FW, Greegor RB (1982) Determination of thermal amplitude of surface atoms in a supported Pt catalyst by EXAFS spectroscopy. J Chem Phys 77(2):1027–1034
80. Meitzner G, Via GH, Lytle FW, Sinfelt JH (1983) Structure of bimetallic clusters. Extended x-ray absorption fine structure (EXAFS) studies of Rh–Cu clusters. J Chem Phys 78(2):882–889
81. Meitzner G, Via GH, Lytle FW, Sinfelt JH (1983) Structure of bimetallic clusters. Extended x-ray absorption fine structure (EXAFS) studies of Ir–Rh clusters. J Chem Phys 78(5): 2533–2541
82. Meitzner G, Via GH, Lytle FW, Sinfelt JH (1983) Structure of bimetallic clusters. Extended x-ray absorption fine structure (EXAFS) studies of Re–Cu, Ir–Cu, and Pt–Cu clusters. J Chem Phys 83(1):353–360
83. Meitzner G, Via GH, Lytle FW, Sinfelt JH (1983) Structure of bimetallic clusters. Extended x-ray absorption fine structure (EXAFS) studies of Ag–Cu and Au–Cu clusters. J Chem Phys 83(9):4793–4799
84. Meitzner G, Via GH, Lytle FW, Sinfelt JH (1983) Structure of bimetallic clusters. Extended x-ray absorption fine structure (EXAFS) of Pt–Re and Pd–Re clusters. J Chem Phys 87(11):6354–6363
85. Mays CW, Vermaak JS, Kuhlmann-Wilsdorf D (1968) On surface stress and surface tension: II. Determination of the surface stress of gold. Surf Sci 12(2):134–140
86. Montano PA, Schulze W, Tesche B, Shenoy GK, Morrison TI (1984) Extended x-ray-absorption fine-structure study of Ag particles isolated in solid argon. Phys Rev B 30(2):672–677
87. Montano PA, Purdum H, Schenoy GK, Morrison TI, Schulze W (1985) X-ray absorption fine structure study of small metal clusters isolated in rare-gas solids. Surf Sci 156:228–233
88. Montano PA, Shenoy GK, Alp EE, Schulze W, Urban J (1986) Structure of copper microclusters isolated in solid argon. Phys Rev Lett 56(19):2076–2079
89. Jiang P, Jona F, Marcus PM (1987) Surface effects in metal microclusters. Phys Rev B 36(12):6336–6338
90. Finnis MW, Heine V (1974) Theory of lattice contraction at aluminium surfaces. J Phys F4(4):L37–L41
91. Frenkel AI (1999) Solving the structure of nanoparticles by multiple-scattering EXAFS analysis. J Synchrotron Rad 6: 293–295
92. Hwang BJ, Sarma LS, Chen JM, Chen CH, Shih SC, Wang GR, Liu DG, Le JF, Tang M (2005) Structural models and atomic distribution of bimetallic nanoparticles as investigated by x-ray absorption spectroscopy. J Am Chem Soc 127(31): 11140–11145

93. Cowley JM (1965) Short-range order and long-range order parameters. Phys Rev 138(5A):A1384–A1389
94. Nashner MS, Frenkel AI, Adler DL, Shapley JR, Nuzzo RG (1997) Structural characterization of carbon-supported platinum-ruthenium nanoparticles from the molecular cluster precursor $PtRu_5C(Co)_{16}$. J Am Chem Soc 119:7760–7771
95. Singhal A, Yang JC, Gibson JM (1997) STEM-based mass spectroscopy of supported Re clusters. Ultramicroscopy 67: 191–206
96. Thomas JM, Midgley PA (2004) High- resolution transmission electron microscopy. The ultimate nanoanalytical technique. Chem Commun 7:1253–1267
97. Ishizuka K (2004) FFT multislice method–the silver anniversary. Microsc Microanal 10(1):34–40
98. Liu J (2005) Scanning transmission electron microscopy and its application to the study of nanoparticles and nanoparticle systems. J Electron Microsc 54(3):251–278
99. Liu W (2007) Multi-scale catalyst design. Chem Eng Sci 62(13):3502–3512
100. Menard LD, Gao S, Xu H, Twesten RD, Harpe AS, Song Y, Wang G, Douglas AD, Yang JC, Frenke AI, Nuzzo RG, Murray RW (2006) Sub-nanometer Au monolayer-protected clusters exhibiting molecule-like electronic behavior: Quantitative high-angle annular dark-field scanning transmission electron microscopy and electrochemical characterization of clusters with precise atomic stoichiometry. J Phys Chem B 110(26):12874–12883
101. Patterson AL (1939) The Scherrer formula for x-ray particle size determination. Phyis Rev 56(10):978–982
102. Menard LD, Xu F, Nuzzo RG, Yang JC (2006) Preparation of TiO_2-supported Au nanoparticle catalysts from a Au13 cluster precursor: Ligand removal using ozone exposure versus a rapid thermal treatment. J Catal 243(1):64–73
103. Bellon P, Manassero M, Sansoni MJ (1972) Crystal and molecular structure of tri-iodoheptakis(tri-p-fluorophenylphosphine)undecagold. Dalton Trans 1481–1487
104. Bellon PL, Cariati F, Manassero M, Naldini M, Sansoni M (1971) Novel gold clusters. Preparation, properties and x-ray structure determination of salts of octakis(triarylphosphine)-enneagold. $[Au_9L_8]X_3$. J Chem Soc Chem Commun 1423–1424
105. Guliamov O, Frenkel AI, Menard LD, Nuzzo RG, Kronik L (2007) Tangential ligand-induced strain in icosahedral Au_{13} J Am Chem Soc Commun 129:10978–10979
106. Hohenberg P, Kohn W (1964) Inhomogeneous electron gas. Phys Rev 136(3B):B864–B871; Kohn W, Sham L (1965) Self-consistent equations including exchange and correlation effects. Phys Rev 140(4A):A1133–A1138
107. Michaelian K, Rendón N, Garzón IL (1999) Structure and energetics of Ni, Ag, and Au nanoclusters. Phys Rev B 60(3): 2000–2010
108. Häberlen OD, Chung S, Stener M, Rösch N (1997) From clusters to bulk: A relativistic density functional investigation on a series of gold clusters Aun, $n = 6, \ldots, 147$. J Chem Phys 106(12):5189–5201
109. Chang CM, Chou MY (2004) Alternative low-symmetry structure for 13-atom metal clusters. Phys Rev Lett 93:133401–4
110. Darby S, Mortimer-Jones TV, Johnston RL, Roberts C (2002) Theoretical study of Cu–Au nanoalloy clusters using a genetic algorithm. J Chem Phys 116(4):1536–1550
111. Häkkinen H, Barnett RN, Landman U (1999) Electronic structure of passivated $Au_{38}(SCH_3)_{24}$ nanocrystal. Phys Rev Lett 82(16):3264–3267
112. Garzon IL, Rovira C, Michaelian K, Beltran MR, Ordejon P, Junquera J, Sanchez-Portal D, Artacho E, Soler JM (2000) Do thiols merely passivate gold nanoclusters? Phys Rev Lett 85(24):5250–5251
113. Häkkinen H, Walter M, Grönbeck H (2006) Divide and protect: Capping gold nanoclusters with molecular gold-thiolate rings. J Phys Chem B 110(20):9927–9931
114. Dubbois LH, Nuzzo RG (1992) The synthesis, structure, and properties of model organic surfaces. Annu Rev Phys Chem 43:437–463
115. Garzón IL, Beltrán MR, González G, Gutíerrez-González I, Michaelian K, Reyes-Nava JA, Rodríguez-Hernández JI (2003) Chirality, defects, and disorder in gold clusters. Eur Phys J D 24:105–109
116. Knecht MR, Weir MG, Frenkel AI, Crooks RM (2008) Structural rearrangement of bimetallic alloy PdAu nanoparticles to yield core/shell configurations. Chem Mater 20:1019–1028
117. Yang JC, Bradley S, Gibson JM (2003) The oblate structure of PtRu5 supported on carbon black. Mater Charact 51:101–107
118. Wang LL, Khare SV, Johnson DD, Rockett AA, Chirita V, Frenkel AI, Mack NH, Nuzzo RG (2006) Origin of bulk-like structure and bond distributions of Pt37 and Pt6Ru31 cluster on carbon: comparison of theory and experiment. J Am Chem Soc 128(1):131–142
119. Oudenhuijzen MK, Van Bokhoven JA, Miller JT, Ramaker DE, Koningsberger DC (2005) J Am Chem Soc 127(5):1530–1540 (and references therein)
120. Wang LL, Johnson DD (2005) Removing critical errors for DFT application to transition-metal nanoclusters: correct ground-state structures of Ru clusters. J Phys Chem B Lett 109(49):23113
121. Johnston RL (2003) Evolving better nanoparticles: Genetic algorithms for optimising cluster geometries. Dalton Trans 4193–4207
122. Kumara S, Johnson DD, Goldberg DE (2007) Scalability of a hybrid extended compact genetic algorithm for ground state optimization of clusters. Mater Manuf Process 22(5):570–576
123. Wales DJ, Doye JPK (1997) Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. J Phys Chem A 101(28):5111–5116
124. Wang LL, Johnson DD (2007) Shear instabilities in metallic nanoparticles: Hydrogen-stabilized structure of Pt37 on carbon. J Am Chem Soc 129(12):3658–3664
125. Arslan I, Yates TJV, Browning ND, Midgley PA (2005) Embedded nanostructures revealed in three dimensions. Science 309(5744):2195–2198
126. Zuo JM, Vartanyants I, Gao M, Zhang R, Nagahara LA (2003) Atomic resolution imaging of a carbon nanotube from diffraction intensities. Science 300(5624):1419–1421
127. Perdew JP, Emzerhof M, Burke K (1996) Rationale for mixing exact exchange with density functional approximations. J Chem Phys 105(22):9982–9985
128. Adamo C, Barone V (1999) Toward reliable density functional methods without adjustable parameters: The PBE0 model. J Chem Phys 110(13):6158–6170
129. Becke AD (1993) A new mixing of Hartree–Fock and local density-functional theories. J Chem Phys 98(2):1372–1377

130. Becke AD (1993) Density-functional thermochemistry. III. The role of exact exchange. J Chem Phys 98(7):5648–5652
131. Bylander DM, Kleinman L (1990) Good semiconductor band gaps with a modified local-density approximation. Phys Rev B 41(11):7868–7871
132. Becke AD (1996) Density-functional thermochemistry, vol IV. A new dynamical correlation functional and implications for exact-exchange mixing. J Chem Phys 104(3):1040–1046
133. Heyd J, Scuseria GE, Ernzerhof M (2003) Hybrid functionals based on a screened Coulomb potential. J Chem Phys 118(18):8207–8215
134. Heyd J, Scuseria GE (2004) Assessment and validation of a screened Coulomb hybrid density functional. J Chem Phys 120(16):7274–7280
135. Wu Z, Cohen RE (2006) More accurate generalized gradient approximation for solids. Phys Rev B 73(23):235116–235121

Books and Reviews

Edelstein AS, Cammarata RC (eds) (1998) Nanomaterials: Synthesis, Properties and Applications. Institute of Physics Publishing, Bristol
Heiz U, Landman U (eds) (2006) Nanocatalysis. Springer, Heidelberg

Nanoscale Processes, Modeling Coupled and Transport Phenomena in Nanotechnology

Roderick Melnik[1,2]
[1] M²NeT Lab, Department of Mathematics, Wilfrid Laurier University, Waterloo, Canada
[2] Department of Physics, University of Waterloo, Waterloo, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
Hierarchy of Mathematical Models for LDSNs
Numerical Methodologies
Incorporating New Effects
Applications and Concluding Remarks
Acknowledgments
Bibliography

Glossary

Low dimensional semiconductor nanostructures Structures that have characteristic dimensions on the order of nanometers (usually between 1 and 100 nm) and such that the motion of electrons in them can be confined spatially. Quantum well heterostructures were the first low dimensional semiconductor nanostructures experimentally developed in the early 1970s. Now, a number of different techniques exist to produce a variety of different low dimensional semiconductor nanostructures, including quantum wells, wires, and dots.

Electronic structure Electrons move from one energy level to another by emission or absorption of a quantum of energy, a photon. They are placed on electronic orbitals and their configuration defines the electronic structure as the arrangement of electrons in an atom, molecule, etc. The knowledge of the electronic structure of a specific material or a structure helps us in predicting its properties.

Electromechanical effects One of the most studied examples of such effects is piezoelectricity, demonstrated for the first time by the brothers Pierre Curie and Jacques Curie in 1880. Now, it is one of the classical examples of coupled phenomena. In the heart of the piezoelectric phenomenon is a coupling mechanism between mechanical and electric fields which is a two-way interaction. In particular, electricity is produced in a piezoelectric body when stress is applied (the direct piezoeffect) and the body is stressed when an electric field is applied (the converse piezoeffect).

Multiple scales In studying complex systems we have to deal with coupled phenomena and processes at a multitude of different spatial and temporal scales. Understanding interactions in the system and its response at multiple scales is a fundamental quest of modern science.

Nanotechnology A multidisciplinary field that develops and extends our present knowledge into the nanoscale. It is the field where two main science and technology approaches, the "bottom-up" approach and the "top-down" approach, go hand in hand emphasizing the importance of systems science view.

Mathematical modeling A universal tool of modern science and technology that uses mathematical language to describe the behavior of systems, processes and phenomena in Nature and man-made.

Definition of the Subject

Low Dimensional Semiconductor Nanostructures (abbreviated, LDSNs) is a class of physical systems with characteristic dimensions on the order of 1–100 nm such that the motion of their charge carriers can be confined from one, two, or even three spatial dimensions. If we start our consideration from a three dimensional (3D) bulk crystal and create a structure where the motion of carriers is confined

from only one spatial direction, we will arrive at a simplest example of LDSNs, two dimensional (2D = 3D − 1D) nanostructures known as quantum wells. Confining the motion of carriers from two spatial directions would lead us to one dimensional (1D = 3D − 2D) nanostructures such as quantum wires or quantum rods. Finally, it is possible to confine the motion of charge carriers in LDSNs from all three spatial dimensions. Such nanostructures are often termed as 0-dimensional (0D = 3D − 3D) and known in the literature as quantum dots with other terms, such as quantum islands and artificial atoms, also used to describe them.

The history of the subject was closely interwoven first with the works on semiconductor lasers in the early 1970s, but the idea of using heterostructures for emitting purposes can easily be traced back to the early days of electronics and to works of Shockley, Kroemer, Alferov and others [1]. A heterostructure is a semiconductor structure with heterojunction, that is a junction composed of parts (or layers) of dissimilar semiconductor materials that have different electronic properties such as energy (band) gaps. Such structures have properties unmatched by the underlying bulk semiconductor materials. A structure with periodically alternating layers of several materials is called a superlattice. In semiconductor superlattices the bandwidth can be tuned by changing the width of the barriers wells. Such superlattices attracted attention in early 1960s and the early studies of the effect of a periodic modulation of the potential in one direction on the crystal band structure is due to Keldysh. Later, Esaki and Tsu and others used one dimensional models to study transport effects in semiconductor superlattices. In early 1970s, heterostructures were realized in the laboratories.

Since that time, the field of low-dimensional systems and nanostructures has grown substantially and includes now semiconductor heterostructures, quantum wells and superlattices, mesoscopic and two-dimensional electron systems, quantum wires and quantum dots. The field has experienced an unprecedented growth in terms of applications, and in addition to its early focus on semiconductor lasers, it includes now a wealth of applications in optoelectronics, quantum information processing, security and defence, health care and biotechnologies, among many other areas.

Introduction

Echoing R. Feynman's famous comments on the existence of "plenty of room at the bottom", the discussion about nanoscience and complexity continues in both popular and scientific literature [32,68,85]. In the focus of this discussion is an electron. Electrons have a wave-particle dual nature, occupying regions known as atomic orbitals. Such orbitals form a discrete set of energy levels (see Fig. 1) as they are identified with the quantum states in which electrons, surrounding an atom, may exist. These energy levels (bands) have different widths, depending on the properties of the corresponding orbitals. Mathematically, they are described by a wave function obeying the laws of quantum mechanics, and in particular the Schrödinger equation. In other words, since the electrons of all materials may only have certain allowable energies, we describe each of these allowed energies by the energy levels. From a chemistry point of view, bonding of atoms to form molecules of matter occurs through the interaction of the valence electrons of each atom and in solids atoms are brought together in such a way that the energy levels of individual atoms form bonds of energy.

Electrons tend to occupy energy states with the lowest energy possible and the energy levels corresponding to such states are the valence band's energy levels. The key in differentiating between electric properties of different materials lies with a "forbidden" band (the bandgap), the energy difference that separates the valence band from the more energetic conduction band. More precisely, as seen in Fig. 1, it is the energy difference between the top of the valence band and the bottom of the conduction band. So, from a physics point of view, electrical properties of matter are determined by three main energy bands: valence band, conduction band, and the bandgap. To move into the conduction band, the valence electrons must bridge an energy gap, which determines whether a solid acts as a conductor, a semiconductor, or an insulator (see Fig. 1). Unlike in insulators, in semiconductors enough energy exists in the valence electrons to enable them to cross the energy gap and exist as conduction (free) electrons in the conduction band. It is due to the bandgap that electrons are constrained from jumping easily from the usually more densely populated valence band to the conduction band. Since in a semiconductor bulk, we have continuous energy states, the bandgap is fixed. However, in LDSNs, and in particular in quantum dots, it can be altered to produce a range of energies between the valence and conduction bands, depending on the size, composition, and shape of these structures. Based on the band structures, which can be determined computationally by solving corresponding quantum mechanical models, we can determine optoelectronic properties of LDSNs. These properties are affected by other material properties such as thermal and mechanical, for instance, making the entire problem of determining the properties of LDSNs coupled. This leads to a situation where the systems

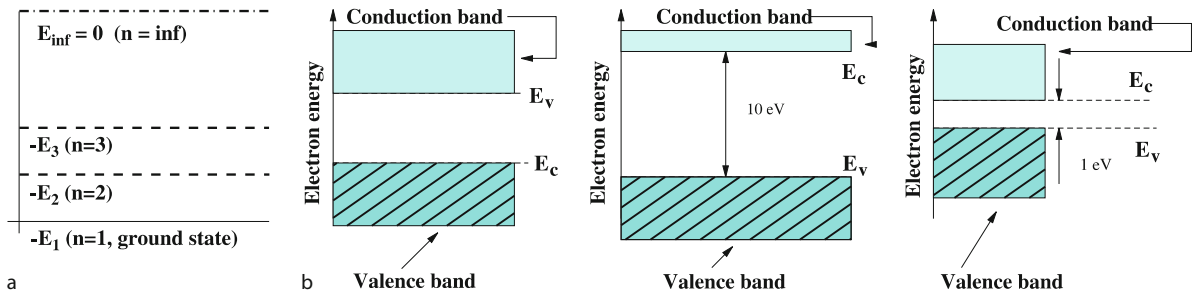

Nanoscale Processes, Modeling Coupled and Transport Phenomena in Nanotechnology, Figure 1
a A schematic electron energy level diagram; b A schematic representation of electric properties of solids, depending on the bandgap: (a) conductors, (b) insulators, (c) semiconductors

science approaches become increasingly important in this field.

Low-dimensional semiconductor nanostructures are multiscale complex systems. Parts of these systems (e. g., two bulk materials) are joined together at the atomic level via interfaces to form a new structure with properties unmatched before [6,16]. As pointed out in [70], "complexity is no longer limited to biology or human sciences: it is invading the physical sciences as deeply rooted in the laws of nature".

Today, a number of experimental techniques exist to produce both self-assembled as well as free-standing quantum dots. Typical electronic structures of such systems require calculations with $10^3 - 10^6$ atoms [86]. Already in itself, it is a task of enormous computational complexity. In addition, as applications of nanostructures and nanostructured materials continue to grow rapidly, experimental results clearly point out that there are many additional effects that may influence profoundly optoelectromechanical properties of the nanostructures. Among such effects are strain relaxation, piezoelectric and thermoelectric effects. Indeed, the formation of LDSNs, and in particular quantum dots, is a competition between the surface energy in the structure and strain energy. At the atomistic level, as two different semiconductor materials are joint together, we have a lattice mismatch in the resulting structure that leads to the properties that are substantially different compared to the underlying bulk materials. At the same time, away from the interface between these semiconductor materials, we have to deal with the effects that are pronounced at larger-than-atomistic scales, making the overall problem of determining properties of LDSNs intrinsically multiscale. The solution to this multiscale problem lies with an effective combination of methods based on bottom-up and top-down approaches at the stage of model construction and the application of efficient computational tools at the next stage. We provide further details on these two stages in the next two sections.

Hierarchy of Mathematical Models for LDSNs

The construction of a hierarchy of mathematical models for modeling LDSNs can be started from the quantum Liouville equation. If the nanostructure has n degrees of freedom, its state in the Wigner–Weyl phase space can be described by the Wigner distribution function ρ (the density operator). If we assume that the system's motion is due to a Hamiltonian function $H(\mathbf{q}, \mathbf{p}, t)$ [64], then the evolution of the state of the system can be described by the quantum Liouville equation:

$$i\hbar \frac{\partial \rho}{\partial t} = [H, \rho] = H\rho - \rho H \,, \tag{1}$$

where $\mathbf{q}$ and $\mathbf{p}$ are vectors representing the n coordinate and n momentum operators, respectively. In a number of applications a generalized form of the quantum Liouville equation has proved to be more convenient where we split the Hamiltonian operator

$$H = \sum_{i=1}^{n} \frac{p_i^2}{2m_i} + V(\mathbf{q}, t) \tag{2}$$

into two Hamiltonians describing the nuclear motion for the lower and upper electronic states

$$H_l = T + V_l, \quad H_u = T + V_u \,, \tag{3}$$

where T is the kinetic energy and V_l and V_u are the effective potentials of the neclei in the lower and upper electronic state, respectively. In this case, the generalized quantum Liouville equation can be given as [13]:

$$i\hbar \frac{\partial \rho}{\partial t} = H_u \rho - \rho H_l \,, \tag{4}$$

and it can be shown that this equation has the form of a time-dependent Schrödinger equation for a quantum system with $N = 2n$ degrees of freedom. In what follows

we mainly concern with a non-relativistic time-independent version of this equation:

$$\hat{H}\Psi = E\Psi \,, \tag{5}$$

where E is the total energy of (in this general case, many-body) Hamiltonian operator $\hat{H}$ and Ψ is the wave function.

A more conventional way to construct a model hierarchy, in particular at the device simulation level, would be to start from the classical Liouville equation for the evolution of the position-velocity probability density $f(x, v, t)$ and to derive a hierarchy of mathematical models based on the relaxation time approximations, leading to the Boltzmann, hydrodynamic type models and, in the simplest case, to the classical drift-diffusion models [55]. To account for quantum effects in such models several different avenues have been proposed in the literature and examples of such quantum-corrected models include the smooth quantum hydrodynamic approximation and the quantum drift-diffusion model [72]. The resulting models offer a substantial speed up compared to the Schrödinger models. These models start from continuum (e. g., fluid dynamics like) representations and attempt to account for the discrete nature of the problem by incorporating a contribution of quantum effects. This modelling technique is usually associated with the top-down approach. Alternatively, we can start from the Schrödinger type models and move on to incorporate additional important effects that are pronounced at larger-than-atomistic scales. This technique allows us to move from discrete to continuum representations and it is usually associated with the bottom-up approach [56]. As the characteristic dimensions of devices continue to shrink and the number of charge carriers in many applications can be small, highlighting the importance of quantum effects, models based on the mixed-state Schrödinger equation coupled to Poisson's equation for the electrostatic potential have been playing an increasingly important role. However, in the context of LDSNs, such models should be extended to account for strain and the piezoelectric effect responsible for the interaction between electric and mechanical fields, as well as for other coupled effects that may influence the overall properties of the structure.

Atomistic simulation methodologies to solving problem (5) include ab initio methods, molecular dynamics and Monte-Carlo. They have a substantial, and often prohibitive for practical applications, computational cost. Inevitably, they are based on a set of assumptions and approximations. One has to approximate the many-body Hamiltonian and the first frequently used simplification lies with the Born–Oppenheimer approximation leading to decoupling between nuclear and electronic coordinates. This decoupling procedure results in the Schrödinger equation for electrons. In deciding which methodology to apply we should balance between the accuracy of the problem solution and the computational efficiency. For example, quantum dots may contain different amount of charge, from one electrons to thousands [30,38]. If the number of electrons in the structure is small and dynamic properties are not of the major concerns, atomistic methodologies can be preferential. In most realistic cases, however, several other factors should be taken into account. Indeed, if a LDSN is part of a device or a larger structure, the resulting system usually contains a number of atoms far beyond the efficient reach of atomistic methodologies. As a result, the development of other-than-atomistic techniques has become increasingly important. Moreover, even if we are able to complete calculations of the characteristics of the structure with an atomistic methodology, understanding of the interactions between atomistic and larger-than-atomistic scales is often of utmost importance. It has been acknowledged in the literature (e. g., [27,28]) that the key conceptual difficulty in materials science in general, and in computational materials science in particular, is to learn how to deal efficiently with a large range of length and time scales. This is also important for nanostructure studies where accounting for additional phenomena may often bring unexpected results. For instance, for LDSNs nonequilibrium phenomena resulting from fluctuating material interfaces may become of great importance [31].

Even if the time-dependency is neglected in the first approximation in the modelling of LDSNs, as we did by considering model (5), we still have to face a challenge of multiple scales that is different from many other areas of computational materials science. Indeed, for a long period of time, while modelling semiconductor devices (including transport phenomena) by using tools and methods of computational solid state electronics, researchers were able successfully applied classical models where the quantum theory was used for calculating necessary parameters dependent on the microscopic atomistic properties of the structures. As the spatial regions in new devices are becoming comparable with the de Broglie wave length of charged particles, the emphasis on quantum effects is ever increasing. This should come at no surprise because it is exactly the applications of such quantum effects that lead to further advances in nanoelectronics. In the heart of what we call 'quantum devices' is a quantum phenomenon or phenomena where the application of quantum mechanics is essential for predicting properties of such devices. What is often missing in these argu-

ments is the fact that, given passive and contact regions of such devices, in most practical situations they have some "macroscopic" dimensions in a sense that we have to deal with micron or sub-micron regions of the device. In such cases, models that go beyond conventional drift-diffusion approximations and can be applied at sub-micron scales are quite useful [53,54,55]. With an increasing importance of quantum effects, microscopic and macroscopic properties of devices we study in nanotechnological applications are interconnected, they are coupled [56,58]. The study of such properties leads to multiple scale problems in computational materials science that represent non-trivial difficulties for both theory and simulation [56,61]. Relating the quantum mechanical description of the processes and phenomena that are taking place at the atomic level to the behavior of matter at the meso- or macroscopic levels is a key task in studying properties of the materials, the task that has been attempted by many scientists ever since the dawn of quantum mechanics (e. g., [26] and references therein).

Numerical Methodologies

The band structure calculation is a generic problem occurring in many fields of science and engineering. This problem is not limited to solid state physics applications as we need to face similar problems when we are dealing with wave phenomena in other fields, including acoustics, electromagnetism, and other areas [57]. In addressing this problem in the context LDSNs, the systems science approach becomes most natural. One of the reasons for that has already been emphasized in the previous section – in studying LDSNs we have to deal with intrinsically coupled systems. Another important reason lies with the fact that there is an intrinsic loss of information[1] when we construct a model for the electronic structure calculations. At the level of measurements the problem of the loss of information is well known and widely discussed in the literature (e. g. [8]). In such cases, it is essential to use prior knowledge from other methods and from other scales if the original problem was coarse-grained. Since semiconductor nanostructures are used as parts of electronic and optoelectronic devices, in their modelling a detailed microscopic scale should be combined with larger scales where such effects as strain relaxation and piezoelectric effects are pronounced. A complete description of the system in such cases can be given with a high computational efficiency within the framework of the $k \cdot p$ theory. It is true that with an increasing computational power (and indeed, we are moving steadily towards a petaflop computer [74]), several more refined methodologies, including empirical and density functional tight-binding and pseudopotential, become more approachable for the modeling of LDSNs [17]. Nevertheless, it should be realized that the problem remains its multiscale nature as long as the nanostructure under consideration is intended to be embedded as a functional element in a device, a chip, or eventually in a larger scale structure intended for the human use. In such cases, the influence of microscopic and macroscopic domains in the structure should be understood as a two-way interaction. As a result of such interactions, more refined atomistic methodologies mentioned above should be viewed only as a part of the problem solution. We need to integrate effects and phenomena that are pronounced at larger-than-atomistic scales into our mathematical models in such a way that the analysis of the entire system would still be possible within a computationally acceptable time.

Let us consider this issue in more details. It has already been emphasized that in designing new devices, it is paramount to know the properties of the materials they are made of. One of the most important steps in this direction is to find the solution to the electronic and ionic structure problem, achievable only by using efficient computational tools. As we know, under appropriate assumptions the ionic degrees of freedom can be separated from those of the electrons and then the problem of electronic structure calculations can be approached from a number of different directions which rely on a set of underlying assumptions. For example, the problem (5) can be re-formulated by using the variational principle:

$$\delta E[\Psi] = 0 \,, \tag{6}$$

so that if we look for its solutions in the subspace of the products of single particle orbitals we will arrive at the well-known Hartree method. Alternatively, if we seek the solutions to the problem in the subspace of Slater determinants (that is antisymmetrized products of single-particle orbitals), the Hartree–Fock (HF) approximation is obtained. A single Slater determinant of single-electron wave function corresponds to the spin-orbitals. In the latter case, the problem is reduced to n one-body problems which should be solved self-consistently.

An essentially different concept is used in density functional theory (DFT) approaches where instead of the wave function as in the HF method, we use the spin density (the density of electrons). In practice, by using additional assumptions such as the local density approximation or the generalized gradient approximation we can proceed with the solution of the problem. The problem is usually reduced in this case to the Kohn–Sham (KS) equations requiring a self-consistent solution. From a numerical point

[1] The definition of information can be found in [37].

of view, this can be done in several different ways, e. g. by using the plane-wave cutoff approximation or real-space methodologies [33]. In both cases, the main idea is that the occupied states of the electrons generate a charge density that corresponds to the same potential used to derive the KS equations.

In the context of LDSNs, depending on the specific application at hand, nanostructure modelling can be approached based on several different techniques. The top-down approach provides us a tool to construct a hierarchy of mathematical models of various complexity for semiconductor device modelling [55]. As the size of semiconductor devices becomes smaller, quantum effects, non-homogeneity and defects, along with surface effects, become increasingly important. As a result, an increasing attention is paid to a hierarchy of mathematical models for nanostructures that can be constructed based on the bottom-up approach, starting from the first-principle methods [83]. In a number of cases, this allows us to develop multiscale techniques based on subsequent averaging procedures (e. g., [79]). In modeling LDSNs, it is important to retain the computational efficiency while coupling the scales. Several approaches have been suggested in the literature to address the problem of scale coupling in the context of modelling nanostrucrures. Classifying multiscale methods, some authors even put the coupling methodologies into a separate group, identifying them with the "direct" coupling methodologies [76]. Note, however, that a coupling procedure is required regardless whether we follow the top-down or bottom-up approach, without formally combing them. A review of several methodologies for the atomistic to continuum coupling can be found in [14]. Other multiscale approaches are based on various forms of averaging and coarse grained techniques [18,80]. In the context of nanostuctures with interfaces such techniques have been recently discussed in [40]. Among recently proposed techniques in this direction, we also note [45], where based on the variational multiscale methodology, the authors developed an extension of the two-scale bridging technique to quantum mechanical/continuum coupling. They provided interesting examples, including the coupling procedure between a virtual atom cluster (VAC) model for the continuum modelling with tight binding calculations. In its essence, however, their approach is the top-down approach discussed above. A concurrent multiscale approach that allows one to integrate different scales "seamlessly" was proposed in [46] for metals. Finally, we note that several multiscale techniques based on the renormalization group approach have also recently been discussed [28,31]. The latter approach may prove to be potentially useful for nanostructures if the renormalization is applied to atomistic models. The underlying idea in the development of all these new methodologies is to achieve "not only higher accuracy, but also more efficient, cost-effective and if possible simpler computational methods in electronic structure calculations" [5,49].

Many applications of LDSNs, and in particular of quantum dots, make the simulations based on atomistic methodologies (ab initio, molecular dynamics or Monte Carlo) impractical. While dealing with an isolated nanostructure we have to deal with at least two scales, the atomic scale at the interfaces and the mesoscopic scale of the structure itself, any such a structure embedded into a working device would require dealing with much more scales. It is this multiscale nature of the problem that makes atomistic approaches very problematic to apply in an efficient manner in applications ranging from basic nanoelectromechanical systems (NEMS) [76] to the development of quantum dot-protein bioconjugate nanoassemblies used for imaging in systems biology [50], or in LDSN applications for RNAi technologies [12]. The field of biological and biomedical applications of LDSNs will continue to grow in the context of the development of new biosensing methodologies [2,36]. By conjugating certain proteins to quantum dots a new class of bioluminescent probes can be created [20]. The models describing these processes and systems are intrinsically multiscale. But even in relatively simple physics applications, the complexity of the problem will increase substantially if we take into account the wetting layer on which the quantum dot is grown. Furthermore, such a systems approach in considering the quantum-dot/wetting layer as a coupled structure will lead to the necessity of the formulation of correct boundary conditions and dealing with an additional disparity in spatial scales between the dot and the wetting layer. This issue of multiple scales and the formulation of correct boundary conditions in this case has been dealt with only recently [58].

In moving to more complicated cases, such as those, for example, discussed above in the context of biological and biomedical applications, an averaging over atomic scales is required. It can be carried out by a number of techniques available, and in particular by the empirical tight-binding, the pseudopotential methodology, or by using the $k \cdot p$ approximation. As we need to address the problem in a systemic way and to include a range of additional effects that are pronounced at larger-than-atomistic scales, we note that the $k \cdot p$ theory represents the electronic structure in a continuum-like manner and as such is well suited for incorporating additional effects into the model, including strain relaxation and piezoelectric effects. Within this framework, we can apply a range of pow-

erful numerical discretization procedures well established in mechanics of solids, including finite element methods (FEM). The entire problem can be solved numerically in a computationally efficient manner. It is worthwhile to note further that many methodologies described above, including the KS approach, use a representation of the wave function in a way similar to FEM:

$$\Psi(\mathbf{r}) = \sum_{i=1}^{n} \alpha_i \psi_i(\mathbf{r}) \tag{7}$$

with $\{\psi_i\}$ being a set of the basis functions and $\{\alpha_i\}$ being the set of the coefficients to solve for. For example, models based on a linear combination of atomic orbitals (LCAO) use the representation (7) as does the tight binding model.

Interface boundary conditions in nanostructure modelling deserve special attention. As mentioned, accounting for the correct boundary conditions is often a non-trivial task [58]. Furthermore, when coupling different scales in atomistic and continuum theories, one needs a procedure that should connect these two levels of modelling, atomistic and continuum. Different approaches to construct such procedures have been discussed in the literature. For example, in the context of nanoidentification problems, the authors of [35] surveyed existing so-called "handshake" approaches to coupling the localized fine-grain (e. g., atomistic) domains with their coarse-grain counterparts, where in the latter case the continuum-based modelling is acceptable. Based on the Fourier analysis of the lattice structure, they focused their attention on a particular case of this procedure that allowed them to derive multiscale boundary conditions for the specific problem they analyzed. In the remainder of this section, we show how simple averaging procedures at the fundamental atomistic level allows us to couple quantum mechanical and continuum models for LDSN bandstructure studies in an efficient computational manner. As we have seen, many interesting applications of LDSNs are connected with their optical properties. The main tool for the analysis of the optical properties of these structures is based on atomistic models of the Schrödinger–Poisson (SP) type. However, this model alone is usually not sufficient. Indeed, we recall that the formation of these structures in an industrial setting is based on a competition between the surface energy in the structure and strain energy, making mechanical effects paramount in determining and optimizing the properties of these structures. In addition, coupled effects, such as piezoelectric, could also be important, as it is the case for wurtzite materials (WZ, materials with hexagonal crystal lattice). The piezoeffect, responsible for the two-way coupling between mechanical and electric fields, is a coupled electromechanical phenomena that is convenient to describe with continuum models [51]. Moreover, as we have recently discussed in [56], nonlinear effects may also become important. As a result, in studying optoelectromechanical properties of low-dimensional semiconductor nanostructures it is essential to combine atomistic and continuum models.

Let us highlight how it can be done. We start from incorporating the lattice mismatch in the models for bandstructure calculations by defining the strain associated with that as a mismatch between two material layers. This is followed by the consideration of the Schrödinger–Poisson model where we account for the piezoelectric effect by coupling the SP model with the model for piezoelectricity. The latter procedure is carried out naturally within the $k \cdot p$ approximation of the electronic structure. This implies an averaging procedure over the atomistic scales, but as a trade-off, it allows us a straightforward coupling of the atomistic part of the model (Schrödinger–Poisson) with its continuum counterpart (elasticity with piezoeffect). The complete model is then implemented in a finite element (or finite difference) code. Recent results on the influence of electromechanical effects, including piezoelectric and strain contributions, on optoelectronic properties of the structures can be found in [56], where both linear and nonlinear strain components were considered.

The $k \cdot p$ approximation is based on the first-principle envelope function theory [10,25,39] and offers a computationally attractive tool for simulating LDSNs and the opportunity to incorporate additional effects within its framework. As all techniques arising from the effective mass theory approximations, where an appropriate fitting the model with experimental parameters is required, in the applications of $k \cdot p$ we may have to deal with the problem of spurious solutions, but several methodologies exist now to overcome this difficulty [40]. In the context of modelling LDSNs, the $k \cdot p$ framework provides a way for consistently incorporating some of the key larger-scale effects that influence the band structure computed with the SP model. Before proceeding to the description of the main implementation steps of this framework, recall that the accuracy of the $k \cdot p$ approximation is dependent on the set of basis functions that span the functional space where we seek the envelope function. A practical balance between the physics of the problem and its computational complexity for LDSNs do not usually lead to the choice of "n" in (7) higher than 8, which is the case, for example, for wurtzite semiconductors. This leads to models based on the 8×8 Hamiltonian. From a physical point of view, the model is based on 6 valence subbands and 2 conduction subbands,

accounting for spin up and down situations and has the form of (5) requiring the solution of the following PDE eigenvalue problem with respect to eigenpair (Ψ, E):

$$H\Psi = E\Psi\,, \qquad \Psi = (\psi_S^{\uparrow}, \psi_X^{\uparrow}, \psi_Y^{\uparrow}, \psi_Z^{\uparrow}, \psi_S^{\downarrow}, \psi_X^{\downarrow}, \psi_Y^{\downarrow}, \psi_Z^{\downarrow})^T\,, \tag{8}$$

where $\psi_X^{\uparrow} \equiv (|X> | \uparrow)$ denotes the wave function component that corresponds to the X Bloch function of the valence band with the spin function of the missing electron "up", the subindex "S" denotes the wave function component of the conduction band, etc, and E is the electron/hole energy, as before.

Of course, the form of the Hamiltonian in (8) depends on a particular problem at hand, but a generic representation of the Hamiltonian in the $k \cdot p$ theory can be given as

$$H \equiv H^{(\alpha,\beta)}(\vec{r}) = -\frac{\hbar^2}{2m_0}\nabla_i \mathcal{H}_{ij}^{(\alpha,\beta)}(\vec{r})\nabla_j\,, \tag{9}$$

where $\mathcal{H}$ is defined by the standard Kohn–Luttinger Hamiltonian or by its refined version based on the Burt–Foreman correction, accounting for the properties of degenerate valence states in an electric field. It represents the kinetic energy plus a nonuniform potential field and other effects contributing to the total potential energy of the system. The superindices (α, β) denote a basis for the wave function of the charge carrier.

Next, we have to incorporate into the model (8), (9) strain relaxation effects in LDSNs. In a sense, accounting for strain effects in this model provides a link between a microscopic (quasi-atomistic after averaging) description of the system with the effects that are pronounced at a larger-than-atomistic scale level as a result of interacting atoms. Indeed, in the case of self-assembled quantum dots, for example, during their growth from the crystal substrate wetting layer, atomic displacements *collectively* induce strain in our finite structure. This fact leads to a modification of the bandstructures obtainable for idealized situations without accounting for strain effects.

Fundamental works in the area by Pikus and Bir, Rashba and Sheka, as well as many others (e. g., [9]) has lead to what is now termed as the Rashba–Sheka–Pikus (RSP) Hamiltonian. Recently, such a Hamiltonian has been applied for the modelling of LDSNs, in particular those made of wurtzite materials (e. g., [23]). The question remains, however, on how to resolve adequately physical effects at edges, corners, and interfaces, including strain nonhomogeneities. Although such effects are quite important in many cases, and in particular for a wide range of quantum dot applications, conventional models for bandstructure calculations are based on the original representation of strain for the bulk materials [9], where strain can be treated on the basis of infinitesimal theory with the linear Cauchy relationships between strain and displacements. Unfortunately, geometric irregularities can make such approximate models insufficiently accurate.

Another type of nonlinearity that is worthwhile mentioning in the context of modeling LDSNs is related to material nonlinearities. For example, can we still use linear stress-strain relationships in the modelling of LDSN? Since for LDSNs, strain remains orders of magnitudes smaller than their elastic limits, the linear relationship is acceptable, at least in a first approximation. Nevertheless, material nonlinear effects may still be important. Firstly, semiconductors are piezoelectrics and higher order effects may become important at the level of device simulation. Secondly, elastic and dielectric coefficients, being functions of the structure geometry, are nonlinear, but the elasticity in this field is treated by the valence-force-field approaches.

Another very important point is related to the definition of the total potential energy when modelling LDSNs, in particular those made of wurtzite materials. For quite some time, it has been emphasized that both deformational energy and piezoelectric field functionals should be included consistently in the formulation of problems where coupled electromechanical phenomena are pronounced [51,59,66,84]. However, most results obtained so far in the context of bandstructure calculations are pertinent to minimization of elastic energy only [23]. It is true that in some cases, such an approach could be sufficient, in particular for zinc-blende materials (materilas with cubic crystal lattice) and where the piezoelectric effect is relatively small, but it does produce inaccurate results in other cases, in particular for wurtzite-based LDSNs. Furthermore, if geometric irregularities are accounted for with nonlinear strain-displacement relationships [62], the resulting coupled models based on the equation of elasticity and the Maxwell equation in dielectric approximation [51,59] become nonlinear. Even in the linear approximation, the coupling between the field of deformation and the piezoelectric field is of fundamental importance [34,84].

One of the main advantages of the $k \cdot p$ theory lies in its simplicity, but this comes at a cost of being required to fit experimental parameters into the model and generally the number of such parameters increases with the requirement of higher accuracy. For example, the model (8), (9) we described above for wurtzite-based LDSNs would contain 10 parameters. In order to reduce this number, it is important to seek any prior knowledge about the problem such as a specific geometrical shape of LDSN under consideration. Indeed, in the case of cylindrical geometry,

the above model can be reduced to a simpler one. As it has been demonstrated in [41], this can be done by applying the Sercel–Vahala (SV) approach to the Rashba–Sheka–Pikus strain Hamiltonian (9) with details for WZ materials given, e. g., in [23]. As usual, Hamiltonian entries in this case will remain PDE operators, but their representations will be simplified [41]. While the originally discussed model still need to be applied for most quantum dot structures, its simplified version is convenient to use for modelling rods, cylindrical nanowires, as well as many superlattices.

Now, we will highlight the main steps of incorporating coupled physical effects in modelling LDSNs, on the example of strain relaxation and piezoeffect. Firstly, we reduce the Maxwell equation to its dielectric approximation

$$\nabla(\epsilon\nabla\varphi) = -\rho + \nabla\cdot(P^s + P^p) \tag{10}$$

and solve it *simultaneously* with equilibrium equations by using, e. g., the finite element methodology:

$$\frac{\sigma_{xx}}{\partial x} + \frac{\partial\sigma_{xy}}{\partial y} + \frac{\partial\sigma_{xz}}{\partial z} = 0\,, \tag{11}$$

$$\frac{\sigma_{xy}}{\partial x} + \frac{\partial\sigma_{yy}}{\partial y} + \frac{\partial\sigma_{yz}}{\partial z} = 0\,, \tag{12}$$

$$\frac{\sigma_{xz}}{\partial x} + \frac{\partial\sigma_{yz}}{\partial y} + \frac{\partial\sigma_{zz}}{\partial z} = 0\,. \tag{13}$$

These equations are coupled by constitutive stress-strain relationships which differ depending on crystal configuration of the lattice [51,56,59]. In (10) P^s and P^p are spontaneous and strain-induced polarization, respectively.

Secondly, the outputs from this model allow us to define the Hamiltonian of system (9) on the same computational grid.

Thirdly, by solving the remaining eight coupled elliptic PDEs (8), we find both eigenfunctions and energies corresponding to all subbands under consideration.

This procedure can be extended to account for the carrier density and charge, in which case an additional coupling loop is necessary between the Schrödinger and Poisson parts of the model [56].

This procedure can be used for modelling most LDSNs, including dots, rods, wires, and superlattices. As a representative example in Fig. 3 we show electronic states in the conduction band in a structure depicted in Fig. 2. Different confinement patterns under coupled and uncoupled situations have important practical implications.

Indeed, LDSNs are designed with an ultimate goal to be used in functional devices where we would like to preserve their properties. For example, in designing new optoelectronic devices, we want to preserve properties of

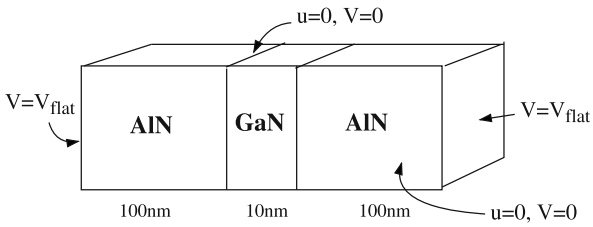

Nanoscale Processes, Modeling Coupled and Transport Phenomena in Nanotechnology, Figure 2
A schematic representation of the geometry of the three-layer AlN/GaN WZ nanostructure

LDSNs such as quantum dots. When such LDSNs are studied theoretically, atomistic (zero-dimensional) density of states in quantum dots allow us to predict some of the characteristics of such structures, e. g. optical gain. If models used for such predictions are uncoupled, a theoretically predicted optical gain can be reduced in practice which may in some cases reduce or even prohibit lasing from the ground state. The reason for that lies with the fact that strain and piezoelectric effects may lead to the elongated ground state carrier function, and hence to a smaller overlap between the electron and hole wave functions, resulting in very different properties as compared to the uncoupled case.

Incorporating New Effects

In the previous section, we discussed two important effects that can substantially influence electronic bandstructure calculations, strain relaxation and piezoeffects. It has been known for quite some time that we have to deal with an inter-dependence between confinement, strain and piezoelectric effects, but quantitative results with the coupling effects accounted for in predicting nanostructure properties have started appearing relatively recently (e. g., [23,56,66] and references therein). The effects of strain and piezo-charge on the properties of LDSNs have been analyzed within the framework of the envelope function approximation as well as within other approaches discussed in the previous section. As we have already emphasized, a systematic approach is required in incorporating such effects into the existing models. Next, we briefly review other important coupled effects that ultimately should be incorporated in the bandstructure calculation procedures.

Coupled Phenomena due to Mechanical and Electric Fields Interactions

The piezoeffect is one of the most important examples of electromechanical coupling where mechanical and elec-

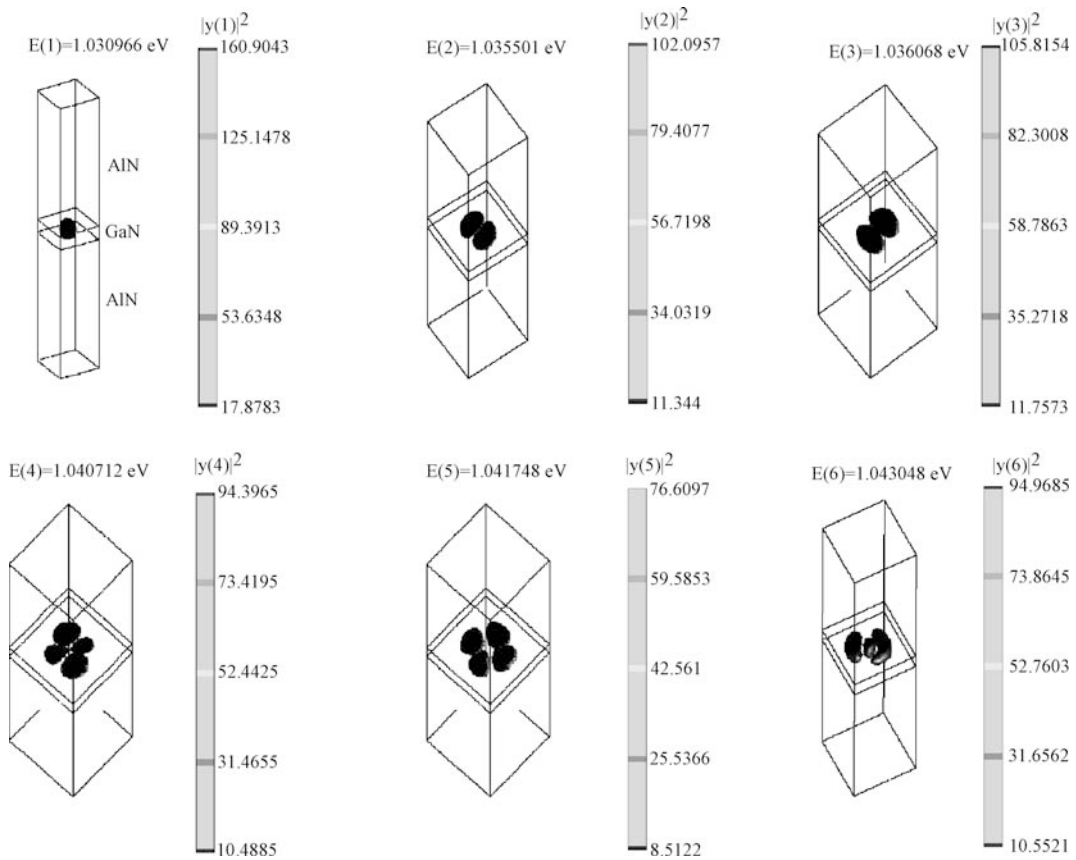

Nanoscale Processes, Modeling Coupled and Transport Phenomena in Nanotechnology, Figure 3
States of conduction band electron confinement (thanks to Roy Mahapatra for producing a figure for this example)

tric fields interact with each other in a two-way interaction manner. The models for coupled dynamic piezoelectricity were put on a rigorous mathematical basis relatively recently [51], and now it is a combination of the piezoelectric effect with semiconductor properties that has attracted attention of the engineers with expectations of a wide range of applications in coupled electronics, sensing and the development of environmentally friendly technologies [82]. Even a new word was coined to reflect these expectations: *nanopiezotronics*. Most of Group III–V compounds used for LDSNs, such as GaAs, have a cubic (zinc-blende) crystal structure and these materials are much better studied. However, it is known also that wide band gap materials such as GaN-based alloys lead to very different properties compared to their zinc-blende counterparts. Usually, they have a hexagonal (wurtzite) crystal structure and a much stronger piezoelectric effect compared to the cubic structures [3]. Anisotropic properties of the material make all the difference, resulting in different properties. The electromechanical coupling is not always linear, and nonlinear electromechanical effects such as electrostriction have recently attracted the attention of researchers working on the properties of LDSNs and materials [7,29,34,63]. The phenomenon of electromechanical coupling is also in the heart of new types of systems known as quantum electromechanical or nano-electromechanical systems (NEMS). In studying such systems quantum effects also become important, and the other level of coupling comes in this case from quantum mechanics where light is coupled to the matter. The first such coupled model was a model describing the photoelectric effect. A more complex type of coupling is observed in quantum electro-mechanical systems where mechanical motion is quantized with many interesting potential applications reported in the literature (e. g., [78] and references therein).

Coupled Effects due to Spatio-temporal Interactions in LDSNs

Although acting on very small scales, temporal effects (see p. 518 in [61]) can be important in a number of appli-

cations. These also include many types of NEMS such as quantum shuttle systems, quantum information processing and control. Furthermore, some effects in LDSNs such as Franz–Keldysh oscillations, observed originally in superlattices, require a more careful examination of spatio-temporal interactions in LDSNs. Other spatio-temporal phenomena that have recently been receiving increasing attention are related to phase transformations [60] and phase stability [83] as they have been analyzed in LDSNs [42]. Such phenomena will become even more important for LDSN-based nanocomposites [65] due to the fact that many of their properties which promise new advances in nanotechnological applications are related to interface effects and phenomena.

Coupled Phenomena due to Thermal and Electric Fields Interactions

Another important group of coupled phenomena important in LDSNs is related to thermoelectricity. Indeed, thermoelectric effects can be substantially enhanced in LDSNs compared to their bulk counterparts [43]. The analysis of such effects is important in the context of the integration of these structures into optoelectronic devices [44]. Furthermore, it is believed that thermoelectric coupling phenomena can facilitate the next generation of electronics and optoelectronics [47].

Coupled Phenomena due to Thermal, Mechanical, and Electric Fields Interactions

It is not difficult to deduce from the above discussion that a combined effect of thermal, mechanical, and electrical fields should be considered. Models of coupled thermoelectroelasticity are known [52] and await their integration into the models for electronic bandstructure calculations. There is another important reason for incorporating coupled effects due to the thermal field at the nanoscale. Indeed, at the nanoscale junctions of LDSNs, the importance of such effects as local electron heating may also become important to account for [15]. In such cases hydrodynamic types of models are needed in a way similar to what was developed in the past for sub-micron devices [53,54,55]. Such models allow us to deal with non-equilibrium and nonlocal phenomena in an efficient computational manner. Recently, in order to deal with the nonequilibrium quantum many-body problem, the time dependent local deformation approximation theory was proposed allowing to recover both Kohn–Sham and the hydrodynamic formulations [77].

Coupling due to the Magnetic Field

While it is customary for the modelling of LDSNs to apply the Maxwell equation in its dielectric approximation, there are cases where such an approximation may not be sufficient and the full electromagnetic model could be required. With the magnetic field is accounted for, there are additional nonlinear effects that in many cases would need to be incorporated into the model. For example, it is known that the quantum dot nuclear spin polarization may depend nonlinearly on applied magnetic fields [48].

Many-Body Coupled Effects

Among such effects are excitonic effects whose importance has been emphasized by a number of authors (e. g., [23] and references therein). Furthermore, in addition to electron-hole Coulomb effects responsible for excitonic transitions, electron-electron Coulomb interactions and the associated Coulomb blockade phenomena, as well as other many-particle effects, may also be important in a number of applications of LDSNs. For example, in the context of quantum dot applications, Coulomb blockade phenomena reflect the finite capacitance of the dot and is effectively an energetic discrimination against two electrons of opposite spin being on the same dot.

More on Nonlinear Coupling

It is generally true that to better understand the electronic properties of materials with strong correlations between the electrons, nonlinear properties become important. One of the examples is provided by the Kondo effect which is particularly pronounced for the low temperature transport in LDSNs [71]. In addition to nonlinear coupled effects already discussed, in the context of excitonic effects it is important to remember that Rabi oscillations of excitons [75] can show a nonlinear signature in LDSNs such as quantum dot array systems. These oscillations have interesting implications in quantum information processing applications since in the case of a single quantum dot they correspond to the one-qubit rotation (e. g., [75] and references therein). Furthermore, the optical response from LDSNs is typically nonlinear [69].

Spin Coupling

Spin is the only internal degree of freedom of an electron, and it is natural to attempt to employ the spin-orbit coupling for creating electronic and optoelectronic devices with new functionalities. One of the possibilities includes the injection of nonequilibrium spins with further manipulations of spin polarization at given locations [73].

While coupled systems is a rule rather than an exception in quantum mechanics, spin-orbit coupled systems have a special place among them as they gave birth to the entire new branch of condensed matter physics, spintronics, where we deal with devices whose functionality depends on the manipulation and control of the spin rather than the charge of electrons.

Applications and Concluding Remarks

As we have seen, LDSNs have led to a number of conceptually new ideas in many applications. Most of these ideas are related, in one way or another, to coupled phenomena being exploited and being put to work in practical applications. By now, strong coupling of light and matter at the single-photon level has led to many important advances [4]. LDSN photonics is an important player in optical information processing [67] and quantum-dot-based photonic crystal lasers represent a reality of today [19]. We are able to create sophisticated nanowires and provide a conceptual framework for nanomachines. In addition to the orbital, we can exploit the spin degree of freedom of the electron and analyze its applicability in spin-based quantum computing, e. g. with the spins of electrons confined to quantum dots [11]. Spin-quantum-NEMS is another example of such an exploitation [22]. We can use LDSNs in energy saving technologies, e. g. in solar panels, as well as in acousto-electronics, photonics, and spintronics with many exciting new developments in these fields happening over the recent years. As the range of applications of LDSNs continue to grow the complex systems science approach in modelling these structures becomes increasingly important. As we have highlighted in the previous sections, in developing new models for LDSNs a systematic combination of bottom-up and top-down techniques is necessary. Based on such a combination a hierarchy of mathematical models of various complexity can be constructed with coupled effects being incorporated in a systematic manner.

Developing this idea, we have discussed two main approaches to nanostructure modelling and highlighted the main procedures for the coupling between quantum mechanical and continuum models. Our approach to multiscale modelling of nanostructures is based on an initial averaging of the quantum mechanical models over the atomic scales and their subsequent coupling to the continuum models. This approach is sufficiently flexible to incorporate other coupled effects discussed in the previous section. We have emphasized that in assisting the design and optimization of new LDSN-based systems by developing the models for LDSNs, it is essential to account for the coupled effects that may lead to well pronounced modifications of properties of the systems being designed. The modelling framework described here is applicable to both low-density isolated LDSNs as well as dense LDSN arrays. The latter is important, given the fact that many systems we have to deal with can be viewed as complex networks. This also includes quantum networks based on LDSNs with application to quantum information processing [21] and the usage of LDSNs in biological network applications [36]. The area of applications of LDSNs in biology and biomedicine will continue to grow rapidly. Quantum dot-based bioconjugate nanoasembly promises to revolutionize the world we live in. A bridge between biology and nanoscience is being created and such complex systems as DNA-NEMS provides just one, albeit very important, example of this trend. Systems biology and systems science approaches in analyzing such coupled systems are becoming more important than ever and it opens new opportunities for a wider applications of these approaches in this exciting interdisciplinary area.

Acknowledgments

This work, conducted in the M^{2}Net Laboratory (http://www.m2netlab.wlu.ca), was supported by the NSERC CRC program, Canada and the Hans Christian Andersen Academy, Denmark. The author is greatful to many his colleagues, in particular to Profs. M. Willatzen, B. Lassen, L. Lew Yan Voon, and R. Mahapatra, for many fruitful discussions on the topics of this paper.

Bibliography

1. Alferov Z (2000) Double heterostructure lasers: early days and future perspectives. IEEE J Sel Top Quantum Electron 6(6): 832–840
2. Alivisatos P (2004) The use of nanocrystals in biological detection. Nat Biotechnol 22(1):47–52
3. Andreev AD, O'Reilly EP (2000) Theory of the electronic structure of GaN/AlN hexagonal quantum dots. Phys Rev 62(23):15851–15870
4. Aoki T et al (2006) Observation of strong coupling between one atom and a monolithic microresonator. Nature 443(12):671–674
5. Artacho E, Beck T, Hernandez E (2006) Current trends in electronic structure: Real-space, embedding and linear scaling techniques. Phys Status Solidi (b) I243(5):971–972
6. Bastard G (1988) Wave mechanics applied to semiconductor heterostructures. Halsted Press, New York (a division of John Wiley & Sons)
7. Bester G et al (2006) Effects of linear and nonlinear piezoelectricity on the electronic properties of InAs/GaAs quantum dots. Phys Rev B 74:081305(R)
8. Billinge SJ, Levin I (2007) The problem with determining atomic structure at the nanoscale. Science 316:561–565

9. Bir GL, Pikus GE (1974) Symmetry and strain-induced effects in semiconductors. Wiley, New York
10. Burt MG (1999) Fundamentals of envelope function theory for electronic states and photonic modes in nanostructures. J Phys: Condens Matter 11:R53–R83
11. Cerletti V et al (2005) Recipes for spin-based quantum computing. Nanotechnology 16:R27–R49
12. Chen AA et al (2005) Quantum dots to monitor RNAi delivery and improve gene silencing. Nucleic Acids Res 33(22):e190
13. Coalson RD, Karplus M (1983) Generalized quantum Liouville equation: its solution by wave packet dynamics. J Chem Phys 79(12):6150–6161
14. Curtin WA, Miller RE (2003) Atomistic/continuum coupling in computational materials science. Model Simul Mater Sci Eng 11(3):R33–R68
15. D'Agosta R, Sai N, Di Ventra M (2006) Local electron heating in nanoscale conductors. Nano Lett 6(12):2935–2938
16. Davies JH (1998) The physics of low-dimensional semiconductors. Cambridge University Press, Cambridge
17. Di Carlo A (2003) Microscopic theory of nanostructured semiconductor devices: beyond the envelope-function approximation. Semicond Sci Technol 18:R1–R31
18. E W et al (2007) Heterogeneous multiscale methods: a review. Commun Comput Phys 2(3):367–450
19. Ellis B et al (2007) Dynamics of quantum dot photonic crystal lasers. Appl Phys Lett 90:151102
20. Evanko D (2006) Bioluminescent quantum dots. Nat Methods 3(4):240–241
21. Faraon A et al (2007) Local quantum dot tuning on photonic crystal chips. Appl Phys Lett 90:213110
22. Fedorets D et al (2005) Spintronics of a nanoelectromechanical shuttle. Phys Rev Lett 95:057203
23. Fonoberov VA, Balandin AA (2003) Excitonic properties of strained wurtzite and zinc-blende $GaN/Al_xGa_{1-x}N$ quantum dots. J Appl Phys 94(11):7178–7186
24. Fonoberov VA, Pokatilov EP, Balandin AA (2003) Interplay of confinement, strain, and piezoelectric effects in the optical spectrum of GaN quantum dots. J Nanosci Nanotechnol 3(3):253–256
25. Foreman BA (2005) First-principles envelope-function theory for lattice-matched semiconductor heterostructures. Phys Rev B (Condens Matter Mater Phys) 72(16):165345
26. George C, Prigogine I, Rosenfeld L (1972) The macroscopic level of quantum mechanics. Nature 240:25–27
27. Ghoniem NM, Busso EP, Kioussis N, Huang H (2003) Multiscale modelling of nanomechanics and micromechanics: an overview. Philos Mag 83(31–34):3475–3528
28. Goldenfeld N, Athreya BP, Dantzig JA (2006) Renormalization group approach to multiscale modelling in materials science. J Stat Phys 125(5–6):1019–1027
29. Guy IL, Muensit S, Goldys EM (1999) Electrostriction in galium nitride. Appl Phys Lett 75(23):3641–3643
30. Harrison P (2005) Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures. Wiley, New York
31. Haselwandter CA, Vvedensky DD (2007) Multiscale theory of fluctuating interfaces: Renormalization of atomistic models. Phys Rev Lett 98(4):046102
32. Heath JR (2001) Nanometer-scale complexity. Chem Eng News 79(13):54
33. Heiskanen M et al (2001) Multigrid method for electronic structure calculations. Phys Rev B 63:245106
34. Jogai B, Albrecht JD, Pan E (2003) Effect of electromechanical coupling on the strainin AlGaN/GaN HFETs. J Appl Phys 94:3984–3989
35. Karpov E et al (2006) Multiscale boundary conditions in crystalline solids: Theory and application to nanoindentation. Int J Solids Struct 43(21):6359–6379
36. Klostranec JM, Chan WCW (2006) Quantum dots in biological and biomedical research: recent progress and present challenges. Adv Mater 18:1953–1964
37. Kolmogorov AN (1968) Three approaches to the quantitative definition of information. Int J Comput Math 2:157–168
38. Kouwenhoven L, Marcus C (1998) Quantum dots. Phys World 6:35–39
39. Lassen B et al (2004) Exact envelope-function theory versus symmetrized Hamiltonian for quantum wires: a comparison. Solid State Commun 132(3–4):141–149
40. Lassen B, Melnik RVN, Willatzen M (2007) Spurious solutions in the multiband effective mass theory applied to low dimensional nanostructures. Preprint Series of the Isaac Newton Institute, University of Cambridge
41. Lew Yan Voon L, Galeriu C, Lassen B et al (2005) Electronic structure of wurtzite quantum dots with cylindrical symmetry. Appl Phys Lett 87(4):041906
42. Liang W, Zhou M (2007) Discovery, characterization and modelling of novel shape memory behaviour of fcc metal nanowires. Philos Mag 87(14–15):2191–2220
43. Lin Y-M, Dresselhaus MS (2003) Thermoelectric properties of superlattice nanowires. Phys Rev B 68:075304
44. Liu W, Balandin AA (2005) Thermoelectric effects in wurtzite GaN and $Al_xGa_{1-x}N$ alloys. J Appl Phys 97:123705
45. Liu WK et al (2006) Bridging scale methods for nanomechanics and materials. Comp Meth Appl Mech Eng 195(13–16): 1407–1421
46. Lu G, Tadmor EB, Kaxiras E (2006) From electrons to finite elements: a concurrent multiscale approach for metals. Phys Rev B 73:024108
47. Majumdar A (2004) Thermoelectricity in semiconductor nanostructures. Science 303:777–778
48. Maletinsky P et al (2007) Nonlinear dynamics of quantum dot nuclear spins. Phys Rev B 75:035409
49. Martin RM (2004) Electronic structure: basic theory and practical methods. Cambridge University Press, Cambridge
50. Medintz IL et al (2005) Quantum dot bioconjugates for imaging, labelling and sensing. Nat Mater 4(6):435–446
51. Melnik RVN (2000) Generalised solutions, discrete models and energy estimates for a 2D problem of coupled field theory. Appl Math Comput 107(1):27–55
52. Melnik RVN (2003) Modelling coupled dynamics: Piezoelectric elements under changing temperature conditions. Int Commun Heat Mass Transf 30(1):83–92
53. Melnik RVN, He H (2000) Modelling nonlocal processes in semiconductor devices with exponential difference schemes. J Eng Math 8(3):233–263
54. Melnik RVN, He H (2000) Quasi-hydrodynamic modelling and computer simulation of coupled thermo-electrical processes in semiconductors. Math Comput Simul 52(3–4):273–287
55. Melnik RVN, He H (2000) Relaxation-time approximations of quasi-hydrodynamic type in semiconductor device modelling. Model Simul Mater Sci Eng 8(2):133–149

56. Melnik RVN, Mahapatra DR (2007) Coupled effects in quantum dot nanostructures with nonlinear strain and bridging modelling scales. Comput Struct 85(11–14):698–711
57. Melnik RVN, Povitsky A (2004) Wave phenomena in physics and engineering: new models, algorithms. Math Comput Simul 65(4–5):299–302
58. Melnik RVN, Willatzen M (2004) Bandstructures of conical quantum dots with wetting layers. Nanotechnology 15(1):1–8
59. Melnik RVN, Zotsenko KN (2004) Mixed electroelastic waves and CFL stability conditions in computational piezoelectricity. Appl Numer Math 48(1):41–62
60. Melnik RVN, Roberts AJ, Thomas KA (2000) Computing dynamics of copper-based SMA via center manifold reduction of 3D models. Comput Mater Sci 18:255–268
61. Melnik RVN et al (2003) Distance geometry algorithms in molecular modelling of polymer and composite systems. Comput Math Appl 45(1–3):515–534
62. Melnik RVN, Lassen B, Lew Yan Voon LC et al (2005) Nonlinear strain models in the analysis of quantum dot molecules. Nonlinear Anal 63(5–7):e2165–e2176
63. Morozovska AN, Eliseev EA, Glinchuk MD (2006) Ferroelectricity enhancement in confined nanorods: direct variational method. Phys Rev B 73:214106
64. Narcowich FJ (1986) A Dyson-like expansion for solutions to the quantum Liouville equation. J Math Phys 27(10): 2502–2510
65. Ovidka IA, Sheinerman AG (2006) Misfit dislocations in nanocomposites with quantum dots, nanowires and their ensembles. Adv Phys 55(7–8):627–689
66. Pan E (2002) Elastic and piezoelectric fields around a quantum dot: Fully coupled or semicoupled model? J Appl Phys 91(6):3785–3796
67. Pauzauskie PJ, Yang P (2006) Nanowire photonics. Mater Today 9(10):36–45
68. Pinto M, Prise KM, Michael BD (2005) Evidence for Complexity at the Nanometer Scale of Radiation-Induced DNA DSBs as a Determinant of Rejoining Kinetics. Radiat Res 164(1):73–85
69. Povolotskyi M et al (2004) Non-linear optical properties of InGaAs/AlGaAs nanostructures grown on (N11) surfaces. Semicond Sci Technol 19(4):S351–S353
70. Prigogine I (1987) Exploring complexity. Europ J Oper Res 30:97–103
71. Pustilnik M (2006) Kondo effect in nanostructures. Phys Status Solidi (a) 203(6):1137–1147
72. Radulovic N, Willatzen M, Melnik RVN, Voon LCLY (2006) Influence of the metal contact size on the electron dynamics and transport inside the semiconductor heterostructure nanowire. J Comput Theor Nanosci 3(4):551–559
73. Rashba EI (2006) Spin-orbit coupling and spin transport. Physica E 34:31–35
74. Sanbonmatsu KY, Tung C-S (2007) High performance computing in biology. : Multimillion atom simulations of nanoscale systems. J Struct Biol
75. Slepyan GY et al (2004) Rabi oscillations in a semiconductor quantum dot: influence of local fields. Phys Rev B 70:045320
76. Tang Z et al (2006) Finite-temperature quasicontinuum method for multiscale analysis of silicon nanostructures. Phys Rev B 74:064110
77. Tokatly IV (2007) Time-dependent deformation functional theory. Phys Rev B 75:125105
78. Twamley J et al (2006) Spin-detection in a quantum electromechanical shuttle system. New J Phys 8:1–27
79. Vvedensky DV (2004) Multiscale modelling of nanostructures. J Phys Condens Matter 16:R1537–R1576
80. Wang CY, Zhang X (2006) Multiscale modeling and related hybrid approaches. Curr Opin Solid State Mater Sci 10:2–14
81. Wang L et al (2006) Nanoparticles for multiplex diagnostics and imaging. Nanomedicine 1(4):413–426
82. Wang ZL (2007) Nanopiezotronics. Adv Mater 19:889–892
83. Wen B, Melnik RVN (2008) First principles molecular dynamics study of Cds nanostructure temperature-dependent phase stability. Appl Phys Lett 92:261911
84. Willatzen M, Lassen B, Voon LCLY et al (2006) Dynamic coupling of piezoelectric effects, spontaneous polarization, and strain in lattice-mismatched semiconductor quantum-well heterostructures. J Appl Phys 100(2):024302
85. Yan H (2004) Nucleic acid nanotechnology. Science 306(12): 2048–2049
86. Zunger A (1998) Electronic-structure theory of semiconductor quantum dots. MRS Bull 2:35–42

Navier–Stokes Equations: A Mathematical Analysis

Giovanni P. Galdi
University of Pittsburgh, Pittsburgh, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Derivation of the Navier–Stokes Equations and Preliminary Considerations
Mathematical Analysis of the Boundary Value Problem
Mathematical Analysis of the Initial-Boundary Value Problem
Future Directions
Acknowledgment
Bibliography

Glossary

Steady-State flow Flow where both velocity and pressure fields are time-independent.

Three-dimensional (or 3D) flow Flow where velocity and pressure fields depend on all three spatial variables.

Two-dimensional (or planar, or 2D) flow Flow where velocity and pressure fields depend only on two spa-

tial variables belonging to a portion of a plane, and the component of the velocity orthogonal to that plane is identically zero.

Local solution Solution where velocity and pressure fields are known to exist only for a finite interval of time.

Global solution Solution where velocity and pressure fields exist for all positive times.

Regular solution Solution where velocity and pressure fields satisfy the Navier–Stokes equations and the corresponding initial and boundary conditions in the ordinary sense of differentiation and continuity.

At times, we may interchangeably use the words "flow" and "solution".

Basic notation $\mathbb{N}$ is the set of positive integers. $\mathbb{R}$ is the field of real numbers and $\mathbb{R}^N$, $N \in \mathbb{N}$, is the set of all N-tuple $\boldsymbol{x} = (x_1, \ldots, x_N)$. The canonical base in $\mathbb{R}^N$ is denoted by $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3, \ldots, \boldsymbol{e}_N\} \equiv \{\boldsymbol{e}_i\}$. For $a, b \in \mathbb{R}$, $b > a$, we set $(a, b) = \{x \in \mathbb{R} : a < x < b\}$, $[a, b] = \{x \in \mathbb{R} : a \le x \le b\}$, $[a, b) = \{x \in \mathbb{R} : a \le x < b\}$ and $(a, b] = \{x \in \mathbb{R} : a < x \le b\}$. By $\bar{\mathcal{A}}$ we indicate the closure of the subset $\mathcal{A}$ of $\mathbb{R}^N$. A *domain* is an open connected subset of $\mathbb{R}^N$. Given a second-order tensor $\boldsymbol{A}$ and a vector $\boldsymbol{a}$, of components $\{A_{ij}\}$ and $\{a_i\}$, respectively, in the basis $\{\boldsymbol{e}_i\}$, by $\boldsymbol{a} \cdot \boldsymbol{A}$ [respectively, $\boldsymbol{A} \cdot \boldsymbol{a}$] we mean the vector with components $A_{ij}a_i$ [respectively, $A_{ij}a_j$]. (We use the Einstein summation convention over repeated indices, namely, if an index occurs twice in the same expression, the expression is implicitly summed over all possible values for that index.) Moreover, we set $|\boldsymbol{A}| = \sqrt{A_{ij}A_{ij}}$. If $\boldsymbol{h}(\boldsymbol{z}) \equiv \{h_i(\boldsymbol{z})\}$ is a vector field, by $\nabla \boldsymbol{h}$ we denote the second-order tensor field whose components $\{\nabla \boldsymbol{h}\}_{ij}$ in the given basis are given by $\{\partial h_j / \partial z_i\}$.

Function spaces notation If $\mathcal{A} \subseteq \mathbb{R}^N$ and $k \in \mathbb{N} \cup \{0\}$, by $C^k(\mathcal{A})$ [respectively, $C^k(\bar{\mathcal{A}})$] we denote the class of functions which are continuous in $\mathcal{A}$ up to their kth derivatives included [respectively, are bounded and uniformly continuous in $\mathcal{A}$ up to their kth derivatives included]. The subset of $C^k(\mathcal{A})$ of functions vanishing outside a compact subset of $\mathcal{A}$ is indicated by $C_0^k(\mathcal{A})$. If $u \in C^k(\mathcal{A})$ for all $k \in \mathbb{N} \cup \{0\}$, we shall write $u \in C^\infty(\mathcal{A})$. In an analogous way we define $C^\infty(\bar{\mathcal{A}})$ and $C_0^\infty(\mathcal{A})$. The symbols $L^q(\mathcal{A})$, $W^{m,q}(\mathcal{A})$, $m \ge 0$, $1 \le q \le \infty$, denote the usual Lebesgue and Sobolev spaces, respectively ($W^{0,q}(\mathcal{A}) = L^q(\mathcal{A})$). Norms in $L^q(\mathcal{A})$ and $W^{m,q}(\mathcal{A})$ are denoted by $\|\cdot\|_{q,\mathcal{A}}$, $\|\cdot\|_{m,q,\mathcal{A}}$. The trace space on the boundary, $\partial\mathcal{A}$, of $\mathcal{A}$ for functions from $W^{m,q}(\mathcal{A})$ will be denoted by $W^{m-1/q,q}(\partial\mathcal{A})$ and its norm by $\|\cdot\|_{m-1/q,q,\partial\mathcal{A}}$.

By $D^{k,q}(\mathcal{A})$, $k \ge 1$, $1 < q < \infty$, we indicate the homogeneous Sobolev space of order (m, q) on $\mathcal{A}$, that is, the class of functions u that are (Lebesgue) locally integrable in $\mathcal{A}$ and with $D^\beta u \in L^q(\mathcal{A})$, $|\beta| = k$, where $D^\beta = \partial^{|\beta|}/\partial x_1^{\beta_1} \partial x_2^{\beta_2} \ldots \partial x_N^{\beta_N}$, $|\beta| = \beta_1 + \beta_2 + \cdots + \beta_N$. For $u \in D^{k,q}(\mathcal{A})$, we put

$$|u|_{k,q,\mathcal{A}} = \left(\sum_{|\beta|=k} \int_{\mathcal{A}} |\, D^\beta u \,|^q \right)^{1/q} .$$

Notice that, whenever confusion does not arise, in the integrals we omit the infinitesimal volume or surface elements. Let

$$\mathcal{D}(\mathcal{A}) = \{\boldsymbol{\varphi} \in C_0^\infty(\mathcal{A}) : \operatorname{div} \boldsymbol{\varphi} = 0\} .$$

By $L_\sigma^q(\mathcal{A})$ we denote the completion of $\mathcal{D}(\mathcal{A})$ in the norm $\|\cdot\|_q$. If $\mathcal{A}$ is any domain in $\mathbb{R}^N$ we have $L^2(\mathcal{A}) = L_\sigma^2(\mathcal{A}) \oplus G(\mathcal{A})$, where $G(\mathcal{A}) = \{\boldsymbol{h} \in L^2(\mathcal{A}) : \boldsymbol{h} = \nabla p$, for some $p \in D^{1,2}(\mathcal{A})\}$; (see Sect. III.1 in [31]). We denote by P the orthogonal projection operator from $L^2(\mathcal{A})$ onto $L_\sigma^2(\mathcal{A})$. By $\mathcal{D}_0^{1,2}(\mathcal{A})$ we mean the completion of $\mathcal{D}(\mathcal{A})$ in the norm $|\cdot|_{1,2,\mathcal{A}}$. $\mathcal{D}_0^{1,2}(\mathcal{A})$ is a Hilbert space with scalar product $[\boldsymbol{v}_1, \boldsymbol{v}_2] := \int_{\mathcal{A}} (\partial \boldsymbol{v}_1/\partial x_i) \cdot (\partial \boldsymbol{v}_2/\partial x_i)$. Furthermore, $\mathcal{D}_0^{-1,2}(\mathcal{A})$ is the dual space of $\mathcal{D}_0^{1,2}(\mathcal{A})$ and $\langle \cdot, \cdot \rangle_{\mathcal{A}}$ is the associated duality pairing.

If $\boldsymbol{g} \equiv \{g_i\}$ and $\boldsymbol{h} \equiv \{h_i\}$ are vector fields on $\mathcal{A}$, we set

$$(\boldsymbol{g}, \boldsymbol{h})_{\mathcal{A}} = \int_{\mathcal{A}} g_i h_i ,$$

whenever the integrals make sense.

In all the above notation, if confusion will not arise, we shall omit the subscript $\mathcal{A}$.

Given a Banach space X, and an open interval (a, b), we denote by $L^q(a, b; X)$ the linear space of (equivalence classes of) functions $f : (a, b) \to X$ whose X-norm is in $L^q(a, b)$. Likewise, for r a nonnegative integer and I a real interval, we denote by $C^r(I; X)$ the class of continuous functions from I to X, which are differentiable in I up to the order r included. *If X denotes any space of real functions, we shall use, as a rule, the same symbol X to denote the corresponding space of vector and tensor-valued functions.*

Definition of the Subject

The Navier–Stokes equations are a mathematical model aimed at describing the motion of an incompressible viscous fluid, like many common ones as, for instance, wa-

ter, glycerin, oil and, under certain circumstances, also air. They were introduced in 1822 by the French engineer Claude Louis Marie Henri Navier and successively re-obtained, by different arguments, by a number of authors including Augustin-Louis Cauchy in 1823, Siméon Denis Poisson in 1829, Adhémar Jean Claude Barré de Saint-Venant in 1837, and, finally, George Gabriel Stokes in 1845. We refer the reader to the beautiful paper by Olivier Darrigol [17], for a detailed and thorough analysis of the history of the Navier–Stokes equations.

Even though, for quite some time, their significance in the applications was not fully recognized, the Navier–Stokes equations are, nowadays, at the foundations of many branches of applied sciences, including Meteorology, Oceanography, Geology, Oil Industry, Airplane, Ship and Car Industries, Biology, and Medicine. In each of the above areas, these equations have collected many undisputed successes, which definitely place them among the most accurate, simple and beautiful models of mathematical physics.

Notwithstanding these successes, up to the present time, a number of unresolved basic mathematical questions remain open – mostly, but not only, for the physically relevant case of three-dimensional flow.

Undoubtedly, the most celebrated is that of proving or disproving the existence of global 3D regular flow for data of arbitrary "size", no matter how smooth (*global regularity problem*). Since the beginning of the 20th century, this notorious question has challenged several generations of mathematicians who have not been able to furnish a definite answer. In fact, to date, 3D regular flows are known to exist *either* for all times but for data of "small size", *or* for data of "arbitrary size" but for a finite interval of time only. The problem of global regularity has become so intriguing and compelling that, in the year 2000, it was decided to put a generous bounty on it. In fact, properly formulated, it is listed as one of the seven $1M Millennium Prize Problems of the Clay Mathematical Institute.

However, the Navier–Stokes equations present also other fundamental open questions. For example, it is not known whether, in the 3D case, the associated initial-boundary value problem is (in an appropriate function space) well-posed in the sense of Hadamard. Stated differently, in 3D it is open the question of whether solutions to this problem exist for all times, are unique and depend continuously upon the data, without being necessarily "regular".

Another famous, unsettled challenge is whether or not the Navier–Stokes equations are able to provide a rigorous model of turbulent phenomena. These phenomena occur when the magnitude of the driving mechanism of the fluid motion becomes sufficiently "large", and, roughly speaking, it consists of flow regimes characterized by chaotic and random property changes for velocity and pressure fields throughout the fluid. They are observed in 3D as well as in two-dimensional (2D) motions (e. g., in flowing soap films). We recall that a 2D motion occurs when the relevant region of flow is contained in a portion of a plane, ϖ, and the component of the velocity field orthogonal to ϖ is negligible.

It is worth emphasizing that, in principle, the answers to the above questions may be unrelated. Actually, in the 2D case, the first two problems have long been solved in the affirmative, while the third one remains still open. Nevertheless, there is hope that proving or disproving the first two problems in 3D will require completely fresh and profound ideas that will open new avenues to the understanding of turbulence.

The list of main open problems can not be exhausted without mentioning another outstanding question pertaining to the boundary value problem describing steady-state flow. The latter is characterized by time-independent velocity and pressure fields. In such a case, if the flow region, $\mathcal{R}$, is multiply connected, it is not known (neither in 2D nor in 3D) if there exists a solution under a given velocity distribution at the boundary of $\mathcal{R}$ that *merely* satisfy the physical requirement of conservation of mass.

Introduction

Fluid mechanics is a very intricate and intriguing discipline of the applied sciences. It is, therefore, not surprising that the mathematics involved in the study of its properties can be, often, extremely complex and difficult. Complexities and difficulties, of course, may be more or less challenging depending on the mathematical model chosen to describe the physical situation.

Among the many mathematical models introduced in the study of fluid mechanics, the Navier–Stokes equations can be considered, without a doubt, the most popular one. However, this does not mean that they can correctly model any fluid under any circumstance. In fact, their range of applicability is restricted to the class of so-called *Newtonian fluids*; see Remark 3 in Sect. "Derivation of the Navier–Stokes Equations and Preliminary Considerations". This class includes several common liquids like water, most salt solutions in water, aqueous solutions, some motor oils, most mineral oils, gasoline, and kerosene.

As mentioned in the previous section, these equations were proposed in 1822 by the French engineer Claude

Navier upon the basis of a suitable molecular model. It is interesting to observe, however, that the law of interaction between the molecules postulated by Navier were shortly recognized to be totally inconsistent from the physical point of view for several materials and, in particular, for liquids. It was only more than two decades later, in 1845, that the same equations were re derived by the 26-year-old George Stokes in a quite general way, by means of the theory of continua.

The role of mathematics in the investigation of the fluid properties and, in particular, of the Navier–Stokes equations, is, as in most branches of the applied sciences, twofold and aims at the accomplishment of the following objectives. The first one, of a more fundamental nature, is the validation of the mathematical model, and consists in securing conditions under which the governing equations possess the essential requirements of well-posedness, that is, existence and uniqueness of corresponding solutions and their continuous dependence upon the data. The second one, of a more advanced character, is to prove that the model gives an adequate interpretation of the observed phenomena.

This paper is, essentially, directed toward the first objective. In fact, its goals consist of (1) formulating the primary problems, (2) describing the related known results, and (3) pointing out the remaining basic open questions, for both boundary and initial-boundary value problems. Of course, it seemed unrealistic to give a detailed presentation of such a rich and diversified subject in a few number of pages. Consequently, we had to make a choice which, we hope, will still give a fairly good idea of this fascinating and intriguing subject of applied mathematics.

The plan of this work is the following. In Sect. "Derivation of the Navier–Stokes Equations and Preliminary Considerations", we shall first give a brief derivation of the Navier–Stokes equations from continuum theory, then formulate the basic problems and, further on, discuss some basic properties. Section "Mathematical Analysis of the Boundary Value Problem" is dedicated to the boundary value problem in both bounded and exterior domains. Besides existence and uniqueness, the main topics include the study of the structure of solution set at large Reynolds number, as well as a condensed treatment of bifurcation issues. Section "Mathematical Analysis of the Initial-Boundary Value Problem" deals with the initial-boundary value problem in a bounded domain and with the related questions of well-posedness and regularity. Finally, in Sect. "Future Directions" we shall outline some future directions of research. Companion to this last section, is the list of a number of significant open questions that are mentioned throughout the paper.

Derivation of the Navier–Stokes Equations and Preliminary Considerations

In the continuum mechanics modeling of a fluid, $\mathcal{F}$, one assumes that, in the given time interval, $I \equiv [0, T]$, $T > 0$, of its motion, $\mathcal{F}$ continuously fills a region, Ω, of the three-dimensional space, $\mathbb{R}^3$. We call points, surfaces, and volumes of $\mathcal{F}$, *material points* (or *particles*), *material surfaces*, and *material volumes*, respectively. In most relevant applications, the region Ω does not depend on time. This happens, in particular, whenever the fluid is bounded by rigid walls like, for instance, in the case of a flow past a rigid obstacle, or a flow in a bounded container with fixed walls. However, there are also some significant situations where Ω depends on time as, for example, in the motion of a fluid in a pipe with elastic walls. Throughout this work, we shall consider flow of $\mathcal{F}$ where Ω *is time-independent*.

Balance Laws

In order to describe the motion of $\mathcal{F}$ it is convenient to represent the relevant physical quantities in the *Eulerian form*. Precisely, if $\boldsymbol{x} = (x_1, x_2, x_3)$ is a point in Ω and $t \in [0, T]$, we let $\rho = \rho(\boldsymbol{x}, t)$, $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{x}, t) = (v_1(\boldsymbol{x}, t), v_2(\boldsymbol{x}, t), v_3(\boldsymbol{x}, t))$ and $\boldsymbol{a} = \boldsymbol{a}(\boldsymbol{x}, t) = (a_1(\boldsymbol{x}, t), a_2(\boldsymbol{x}, t), a_3(\boldsymbol{x}, t))$ be the density, velocity, and acceleration, respectively, of that particle of $\mathcal{F}$ that, at the time t, passes through the point $\boldsymbol{x}$. Furthermore, we denote by $\boldsymbol{f} = \boldsymbol{f}(\boldsymbol{x}, t) = (f_1(\boldsymbol{x}, t), f_2(\boldsymbol{x}, t), f_3(\boldsymbol{x}, t))$ the external force per unit volume (*body force*) acting on $\mathcal{F}$.

In the continuum theory of non-polar fluids, one postulates that in every motion the following equations must hold

$$\left.\begin{aligned} &\frac{\partial \rho}{\partial t}(\boldsymbol{x}, t) + \frac{\partial}{\partial x_i}(\rho(\boldsymbol{x}, t) v_i(\boldsymbol{x}, t)) = 0\,, \\ &\rho(\boldsymbol{x}, t)\, a_i(\boldsymbol{x}, t) \\ &\quad = \frac{\partial T_{ji}}{\partial x_j} + \rho(\boldsymbol{x}, t)\, f_i(\boldsymbol{x}, t)\,, \qquad i = 1, 2, 3\,, \\ &T_{ij}(\boldsymbol{x}, t) = T_{ji}(\boldsymbol{x}, t)\,, \qquad i, j = 1, 2, 3 \end{aligned}\right\} \begin{array}{l} \text{for all} \\ (\boldsymbol{x}, t) \in \Omega \\ \times (0, T)\,. \end{array} \tag{1}$$

These equations represent the local form of the *balance laws of $\mathcal{F}$ in the Eulerian description*. Specifically, $(1)_1$ expresses the *conservation of mass*, $(1)_2$ furnishes the *balance of linear momentum*, while $(1)_3$ is equivalent to the *balance of angular momentum*. The function $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{x}, t) = \{T_{ji}(\boldsymbol{x}, t)\}$ is a second-order, symmetric tensor field, the *Cauchy stress tensor*, that takes into account the *internal forces* exerted by the fluid. More precisely, let S denote a (sufficiently smooth) fixed, closed surface in $\mathbb{R}^3$ bounding the region $\mathcal{R}_S$, let $\boldsymbol{x}$ be any point on S,

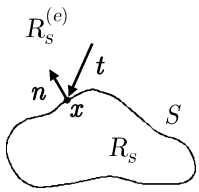

Navier–Stokes Equations: A Mathematical Analysis, Figure 1
Stress vector at the point x of the surface S

and let $\boldsymbol{n} = \boldsymbol{n}(\boldsymbol{x})$ be the outward unit normal to S at $\boldsymbol{x}$. Furthermore, let $\mathcal{R}_S^{(e)}$ be the region exterior to $\mathcal{R}_S$, with $\mathcal{R}_S^{(e)} \cap \Omega \neq \emptyset$. Then the vector $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x}, t)$ defined as

$$\boldsymbol{t}(\boldsymbol{x}, t) := \boldsymbol{n}(\boldsymbol{x}) \cdot \boldsymbol{T}(\boldsymbol{x}, t)\,, \tag{2}$$

represents the force per unit area exerted by the portion of the fluid in $\mathcal{R}_S^{(e)}$ on S at the point $\boldsymbol{x}$ and at time t; see Fig. 1.

Constitutive Equation An important kinematical quantity associated with the motion of $\mathcal{F}$ is the *stretching tensor* field, $\boldsymbol{D} = \boldsymbol{D}(\boldsymbol{x}, t)$, whose components, D_{ij}, are defined as follows:

$$D_{ij} = \frac{1}{2}\left(\frac{\partial v_j}{\partial x_i} + \frac{\partial v_i}{\partial x_j}\right), \quad i, j = 1, 2, 3\,. \tag{3}$$

The stretching tensor is, of course, symmetric and, roughly speaking, it describes the rate of deformation of parts of $\mathcal{F}$. In fact, it can be shown that a necessary and sufficient condition for a motion of $\mathcal{F}$ to be *rigid* (namely, the mutual distance between two arbitrary particles of $\mathcal{F}$ does not change in time), is that $\boldsymbol{D}(\boldsymbol{x}, t) = 0$ at each $(\boldsymbol{x}, t) \in \Omega \times I$. Moreover, $\operatorname{div} \boldsymbol{v}(\boldsymbol{x}, t) \equiv \operatorname{trace} \mathbf{D}(\boldsymbol{x}, t) \equiv \frac{\partial v_i}{\partial x_i}(\boldsymbol{x}, t) = 0$ for all $(\boldsymbol{x}, t) \in \Omega \times I$, if and only if the motion is *isochoric*, namely, every material volume does not change with time. A noteworthy class of fluids whose generic motion is isochoric is that of fluids having constant density. In fact, if ρ is a positive constant, from $(1)_1$ we find $\operatorname{div} \boldsymbol{v}(\boldsymbol{x}, t) = 0$, for all $(\boldsymbol{x}, t) \in \Omega \times I$. Fluids with constant density are called *incompressible*. Herein, incompressible fluids will be referred to as *liquids*.

During the generic motion, the internal forces will, in general, produce a deformation of parts of $\mathcal{F}$. The relation between internal forces and deformation, namely, the functional relation between $\boldsymbol{T}$ and $\boldsymbol{D}$, is called *constitutive equation* and characterizes the physical properties of the fluid. A liquid is said to be *Newtonian* if and only if the relation between $\boldsymbol{T}$ and $\boldsymbol{D}$ is linear, that is, there exist a scalar function $p = p(\boldsymbol{x}, t)$ (the *pressure*) and a constant μ (the *shear viscosity*) such that

$$\boldsymbol{T} = -p\boldsymbol{I} + 2\mu\boldsymbol{D}\,, \tag{4}$$

where $\boldsymbol{I}$ denotes the identity matrix. In a *viscous* Newtonian liquid, one assumes that the shear viscosity satisfies the restriction

$$\mu > 0\,. \tag{5}$$

We will discuss further the meaning of this assumption in Remark 2.

Navier–Stokes Equations In view of the condition $\operatorname{div} \boldsymbol{v} = 0$, it easily follows from (4) that

$$\frac{\partial T_{ji}}{\partial x_j} = -\frac{\partial p}{\partial x_i} + \mu \Delta v_i$$

where $\Delta := \frac{\partial^2}{\partial x_l \partial x_l}$ is the Laplace operator. Therefore, by (1) and (4) we deduce that the equations governing the motion of a Newtonian viscous liquid are furnished by

$$\left.\begin{aligned} &\rho a_i = -\frac{\partial p}{\partial x_i} + \mu \Delta v_i + \rho f_i\,,\ i = 1, 2, 3 \\ &\frac{\partial v_i}{\partial x_i} = 0 \end{aligned}\right\} \text{ in } \Omega \times (0, T)\,. \tag{6}$$

It is interesting to observe that *both* equations in (6) are *linear* in all kinematical variables. However, in the Eulerian description, the acceleration is a *nonlinear* functional of the velocity, and we have

$$a_i = \frac{\partial v_i}{\partial t} + v_l \frac{\partial v_i}{\partial x_l}\,,$$

or, in a vector form,

$$\boldsymbol{a} = \frac{\partial \boldsymbol{v}}{\partial t} + \boldsymbol{v} \cdot \nabla \boldsymbol{v}\,.$$

Replacing this latter expression in (6) we obtain the *Navier–Stokes equations*:

$$\left.\begin{aligned} &\rho\left(\frac{\partial \boldsymbol{v}}{\partial t} + \boldsymbol{v} \cdot \nabla \boldsymbol{v}\right) \\ &\quad = -\nabla p + \mu \Delta \boldsymbol{v} + \rho \boldsymbol{f}\,, \\ &\operatorname{div} \boldsymbol{v} = 0 \end{aligned}\right\} \quad \text{in } \Omega \times (0, T)\,. \tag{7}$$

In these equations, the (constant) density ρ, the shear viscosity μ [satisfying (5)], and the force $\boldsymbol{f}$ are given quantities, while the *unknowns* are velocity $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{x}, t)$ and pressure $p = p(\boldsymbol{x}, t)$ fields.

Some preliminary comments about the above equations are in order. Actually, we should notice that the unknowns $\boldsymbol{v}$, p do not appear in a "symmetric" way. In other words, the equation of conservation of mass $(7)_2$ does *not* involve the pressure field. This is due to the fact that, from

the mechanical point of view, the pressure plays the role of *reaction force* (Lagrange multiplier) associated with the isochoricity constraint $\operatorname{div} \boldsymbol{v} = 0$. In other words, whenever a portion of the liquid "tries to change its volume" the liquid "reacts" with a suitable distribution of pressure to keep that volume constant. Thus, the pressure field must be generally deduced in terms of the velocity field, once this latter has been determined; see Remark 1.

Initial-Boundary Value Problem In order to find solutions to the problem (7), we have to append suitable *initial*, at time $t = 0$, and *boundary* conditions. As a matter of fact, these conditions may depend on the specific physical problem we want to model. We shall assume that the region of flow, Ω, is bounded by rigid walls, $\partial\Omega$, and that the liquid does not "slip" at $\partial\Omega$. The appropriate initial and boundary conditions then become, respectively, the following ones

$$\begin{aligned} &\boldsymbol{v}(\boldsymbol{x}, 0) = \boldsymbol{v}_0(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega \\ &\boldsymbol{v}(\boldsymbol{x}, t) = \boldsymbol{v}_1(\boldsymbol{x}, t), \quad (\boldsymbol{x}, t) \in \partial\Omega \times (0, T), \end{aligned} \tag{8}$$

where $\boldsymbol{v}_0$ and $\boldsymbol{v}_1$ are *prescribed* vector fields.

Steady-State Flow and Boundary-Value Problem An important, special class of solutions to (7), called *steady-state solutions*, is that where velocity and pressure fields are independent of time. Of course, a necessary requirement for such solutions to exist is that $\boldsymbol{f}$ does not depend on time as well. From (7) we thus obtain that a generic steady-state solution, $(\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{x}), p = p(\boldsymbol{x}))$, must satisfy the following equations

$$\left.\begin{aligned} \rho\boldsymbol{v} \cdot \nabla\boldsymbol{v} &= -\nabla p + \mu\Delta\boldsymbol{v} + \rho\boldsymbol{f}, \\ \operatorname{div}\boldsymbol{v} &= 0 \end{aligned}\right\} \quad \text{in } \Omega. \tag{9}$$

Under the given assumptions on the region of flow, from $(8)_2$ it follows that the appropriate boundary conditions are

$$\boldsymbol{v}(\boldsymbol{x}) = \boldsymbol{v}_*(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial\Omega, \tag{10}$$

where $\boldsymbol{v}_*$ is a prescribed vector field.

Two-Dimensional Flow In several mathematical questions related to the unique solvability of problems (7)–(8) and (9)–(10), separate attention deserve two-dimensional solutions describing the *planar motions* of $\mathcal{F}$. For these solutions the fields $\boldsymbol{v}$ and p depend only on x_1, x_2 (say) [and t in the case (7)–(8)], and, moreover, $v_3 \equiv 0$. Consequently, the relevant (spatial) region of motion, Ω, becomes a subset of $\mathbb{R}^2$.

Remark 1 By formally operating with "div" on both sides of $(7)_1$ and by taking into account $(8)_2$, we find that, at each time $t \in (0, T)$ the pressure field $p = p(\boldsymbol{x}, t)$ must satisfy the following Neumann problem

$$\begin{aligned} &\Delta p = \rho(\boldsymbol{v} \cdot \nabla\boldsymbol{v} - \boldsymbol{f}) \quad \text{in } \Omega \\ &\frac{\partial p}{\partial n} = -[\mu\Delta\boldsymbol{v} - \rho(\boldsymbol{v}_1 \cdot \nabla\boldsymbol{v} - \boldsymbol{f})] \cdot \boldsymbol{n} \quad \text{at } \partial\Omega \end{aligned} \tag{11}$$

where $\boldsymbol{n}$ denotes the outward unit normal to $\partial\Omega$. It follows that the *prescription* of the pressure at the bounding walls or at the initial time *independently* of $\boldsymbol{v}$, could be incompatible with (8) and, therefore, could render the problem ill-posed.

Remark 2 We can give a simple qualitative explanation of the assumption (5). To this end, let us consider a steady-state flow of a viscous liquid between two parallel, rigid walls Π_1, Π_2, set at a distance d apart and parallel to the plane $x_2 = 0$ see Fig. 2. The flow is induced by a force per unit area $\boldsymbol{F} = F\boldsymbol{e}_1$, $F > 0$, applied to the wall Π_1, that moves Π_1 with a constant velocity $\boldsymbol{V} = V\,\boldsymbol{e}_1$, $V > 0$, while Π_2 is kept fixed; see Fig. 2. No external force is acting on the liquid. It is then easily checked that the velocity and pressure fields $\boldsymbol{v} = V\,(x_2/d)\boldsymbol{e}_1$, $p = p_0 = \text{const.}$ (*pure shear* flow) satisfy (9) with $\boldsymbol{f} \equiv \boldsymbol{0}$, along with the appropriate boundary conditions $\boldsymbol{v}(0) = \boldsymbol{0}$, $\boldsymbol{v}(d) = V\,\boldsymbol{e}_1$.

From (2) and (4), we obtain that the force $\boldsymbol{t}$ per unit area exerted by the fluid on Π_1 is given by

$$\boldsymbol{t}(x_1, d, x_3) = -\boldsymbol{e}_2 \cdot \mathbf{T}(x_1, d, x_3) = p_0\boldsymbol{e}_2 - \mu\frac{V}{d}\boldsymbol{e}_1,$$

that is, $\boldsymbol{t}$ has a "purely pressure" component $\boldsymbol{t}_p = p_0\,\boldsymbol{e}_2$, and a "purely shear viscous" component $\boldsymbol{t}_v = -\mu(V/d)\boldsymbol{e}_1$. As expected in a viscous liquid, $\boldsymbol{t}_v$ is directed parallel to $\boldsymbol{F}$ and, of course, if it is not zero, it should also act *against* $\boldsymbol{F}$, that is, $\boldsymbol{t}_v \cdot \boldsymbol{e}_1 < 0$. However, $(V/d) > 0$, and so we must have $\mu > 0$. Since the physical properties of the fluid are independent of the particular flow, this simple reasoning justifies the assumption made in (5).

Remark 3 As mentioned in the Introduction, the constitutive Eq. (4) and, as a consequence, the Navier–Stokes

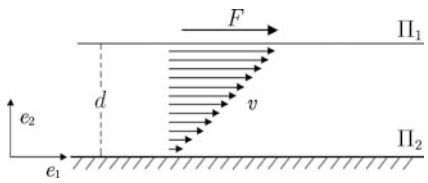

Navier–Stokes Equations: A Mathematical Analysis, Figure 2
Pure shear flow between parallel plates induced by a force F

equations, provide a satisfactory model only for a certain class of liquids, while, for others, their predictions are at odds with experimental data. These latter include, for example, biological fluids, like blood or mucus, and aqueous polymeric solutions, like common paints, or even ordinary shampoo. In fact, these liquids, which, in contrast to those modeled by (4), are called *non-Newtonian*, exhibit a number of features that the linear relation (4) is not able to predict. Most notably, in liquids such as blood, the shear viscosity is no longer a constant and, in fact, it decreases as the magnitude of shear (proportional to $\sqrt{D_{ij}D_{ij}}$) increases. Furthermore, liquids like paints or shampoos show, under a given shear rate, a distribution of stress (other than that due to the pressure) in the direction orthogonal to the direction of shear (the so-called *normal stress effect*). Modeling and corresponding mathematical analysis of a variety of non-Newtonian liquids can be found, for example, in [42].

Mathematical Analysis of the Boundary Value Problem

We begin to analyze the properties of solutions to the boundary-value problem (9)–(10). We shall divide our presentation into two parts, depending on the "geometry" of the region of flow Ω. Specifically, we shall treat the cases when Ω is either a bounded domain or an exterior domain (flow past an obstacle). For each case we shall describe methods, main results, and fundamental open questions.

Flow in Bounded Domains

In this section we shall analyze problem (9)–(10) where Ω is a bounded domain of $\mathbb{R}^3$. A similar (and simpler) analysis can be performed in the case of planar flow, with exactly the same results.

Variational Formulation and Weak Solutions To show existence of solutions, one may use, basically, two types of methods that we shall describe in some detail. The starting point of both methods is the so-called *variational formulation* of (9)–(10). Let $\boldsymbol{\varphi} \in \mathcal{D}(\Omega)$. Since $\operatorname{div}\boldsymbol{\varphi} = \operatorname{div}\boldsymbol{v} = 0$ in Ω and $\boldsymbol{\varphi}|_{\partial\Omega} = \mathbf{0}$, we have (by a formal integration by parts)

$$
\begin{aligned}
(\boldsymbol{\varphi}, \nabla p) &= \int_{\partial\Omega} p\boldsymbol{\varphi}\cdot\boldsymbol{n} = 0\,,\\
(\Delta\boldsymbol{v}, \boldsymbol{\varphi}) &= -[\boldsymbol{v}, \boldsymbol{\varphi}] + \int_{\partial\Omega} \boldsymbol{n}\cdot\nabla\boldsymbol{v}\cdot\boldsymbol{\varphi} = -[\boldsymbol{v}, \boldsymbol{\varphi}]\\
(\boldsymbol{v}\cdot\nabla\boldsymbol{v}, \boldsymbol{\varphi}) &= \int_{\partial\Omega} \boldsymbol{v}\cdot\boldsymbol{n}\boldsymbol{v}\cdot\boldsymbol{\varphi} - (\boldsymbol{v}\cdot\nabla\boldsymbol{\varphi}, \boldsymbol{v}) = -(\boldsymbol{v}\cdot\nabla\boldsymbol{\varphi}, \boldsymbol{v})
\end{aligned}
\tag{12}
$$

where $[\cdot,\cdot]$ is the scalar product in $\mathcal{D}_0^{1,2}(\Omega)$. Thus, if we dot-multiply both sides of $(9)_1$ by $\boldsymbol{\varphi} \in \mathcal{D}(\Omega)$ and integrate by parts over Ω, we (formally) obtain with $\nu := \mu/\rho$

$$
\nu\,[\boldsymbol{v}, \boldsymbol{\varphi}] - (\boldsymbol{v}\cdot\nabla\boldsymbol{\varphi}, \boldsymbol{v}) = \langle \boldsymbol{f}, \boldsymbol{\varphi}\rangle\,, \quad \text{for all } \boldsymbol{\varphi} \in \mathcal{D}(\Omega)\,, \tag{13}
$$

where we assume that $\boldsymbol{f}$ belongs to $\mathcal{D}_0^{-1,2}(\Omega)$. Equation. (13) is the *variational* (or *weak*) *form* of $(9)_1$. We observe that (13) does not contain the pressure. Moreover, every term in (13) is well-defined provided $\boldsymbol{v} \in W^{1,2}_{\mathrm{loc}}(\Omega)$.

Definition 1 A function $\boldsymbol{v} \in W^{1,2}(\Omega)$ is called a weak solution to (9)–(10) if and only if: (i) $\operatorname{div}\boldsymbol{v} = 0$ in Ω; (ii) $\boldsymbol{v}|_{\partial\Omega} = \boldsymbol{v}_*$ (in the trace sense); (iii) $\boldsymbol{v}$ satisfies (13). If, in particular, $\boldsymbol{v}_* \equiv \mathbf{0}$, we replace conditions (i), (ii) with the single requirement: (i)$'$ $\boldsymbol{v} \in \mathcal{D}_0^{1,2}(\Omega)$.

Throughout this work we shall often use the following result, consequence of the Hölder inequality and of a simple approximating procedure.

Lemma 1 *The trilinear form*

$$
(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) \in L^q(\Omega) \times W^{1,2}(\Omega) \times L^r(\Omega) \mapsto (\boldsymbol{a}\cdot\nabla\boldsymbol{b}, \boldsymbol{c}) \in \mathbb{R}\,, \quad \frac{1}{q} + \frac{1}{r} = \frac{1}{2}\,,
$$

is continuous. Moreover, $(\boldsymbol{a}\cdot\nabla\boldsymbol{b}, \boldsymbol{c}) = -(\boldsymbol{a}\cdot\nabla\boldsymbol{c}, \boldsymbol{b})$, *for any* $\boldsymbol{b}, \boldsymbol{c} \in D_0^{1,2}(\Omega)$, *and for any* $\boldsymbol{a} \in L^2_\sigma(\Omega)$. *Thus, in particular,* $(\boldsymbol{a}\cdot\nabla\boldsymbol{b}, \boldsymbol{b}) = 0$.

Regularity of Weak Solutions If $\boldsymbol{f}$ is sufficiently regular, the corresponding weak solution is regular as well and, moreover, there exists a scalar function, p, such that (9) is satisfied in the ordinary sense. Also, if $\partial\Omega$ and $\boldsymbol{v}_*$ are smooth enough, the solution $(\boldsymbol{v}, p)$ is smooth up to the boundary and (10) is satisfied in the classical sense. A key tool in the proof of these properties is the following lemma, which is a special case of a more general result due to Cattabriga [12]; (see also Lemma IV.6.2 and Theorem IV.6.1 in [31]).

Lemma 2 *Let* Ω *be a bounded domain of* $\mathbb{R}^3$, *of class* C^{m+2}, $m \geq 0$, *and let* $\boldsymbol{g} \in \boldsymbol{W}^{m,q}(\Omega)$, $\boldsymbol{u}_* \in W^{m+2-1/q,q}(\partial\Omega)$, $1 < q < \infty$, *with* $\int_{\partial\Omega} \boldsymbol{u}_*\cdot\boldsymbol{n} = 0$. *Moreover, let* $\boldsymbol{u} \in W^{1,2}(\Omega)$ *satisfy the following conditions*

(a) $\nu\,[\boldsymbol{u}, \boldsymbol{\varphi}] = (\boldsymbol{g}, \boldsymbol{\varphi})$ *for all* $\boldsymbol{\varphi} \in \mathcal{D}(\Omega)$;
(b) $\operatorname{div}\boldsymbol{u} = 0$;
(c) $\boldsymbol{u} = \boldsymbol{u}_*$ *at* $\partial\Omega$ *in the trace sense.*

Then, $\boldsymbol{u} \in \boldsymbol{W}^{m+2,q}(\Omega)$ *and there exists a unique* $\phi \in \boldsymbol{W}^{m+1,q}(\Omega)$, *with* $\int_\Omega \phi = 0$, *such that the pair* $(\boldsymbol{u}, \phi)$

satisfies the following Stokes equations

$$\left.\begin{aligned} -\nu\Delta \boldsymbol{u} &= \nabla\phi + \boldsymbol{g} \\ \operatorname{div}\boldsymbol{u} &= 0 \end{aligned}\right\} \quad \text{in } \Omega\,.$$

Furthermore, there exists a constant $C = C(\Omega, m, q) > 0$ *such that*

$$\|\boldsymbol{u}\|_{m+2,q} + \|\phi\|_{m+1,q} \le C\left(\|\boldsymbol{g}\|_{m,q} + \|\boldsymbol{u}_*\|_{m+2-1/q,q,\partial\Omega}\right)\,.$$

To give an idea of how to prove regularity of weak solutions by means of Lemma 2, we consider the case $\boldsymbol{f} \in C^\infty(\bar{\Omega})$, Ω of class C^∞ and $\boldsymbol{v}_* \in C^\infty(\partial\Omega)$. (For a general regularity theory of weak solutions, see Theorem VIII.52 in [32].) Thus, in particular, by the embedding Theorem II.2.4 in [31]

$$W^{1,2}(\Omega) \subset L^q(\Omega)\,, \quad \text{for all } q \in [1,6]\,, \tag{14}$$

and by the Hölder inequality we have that $\boldsymbol{g} := \boldsymbol{f} - \boldsymbol{v}\cdot\nabla\boldsymbol{v} \in L^{3/2}(\Omega)$. From (13) and Lemma 2, we then deduce that $\boldsymbol{v} \in W^{2,3/2}(\Omega)$ and that there exists a scalar field $p \in W^{1,3/2}$ such that $(\boldsymbol{v}, p)$ satisfy (9) a.e. in Ω. Therefore, because of the embedding $W^{2,3/2}(\Omega) \subset W^{1,3}(\Omega) \subset L^r(\Omega)$, arbitrary $r \in [1,\infty)$, Theorem II.2.4 in [31], we obtain the improved regularity property $\boldsymbol{g} \in W^{1,s}(\Omega)$, for all $s \in [1, 3/2)$. Using again Lemma 2, we then deduce $\boldsymbol{v} \in W^{3,s}(\Omega)$ and $p \in W^{2,s}(\Omega)$ which, in particular, gives further regularity for $\boldsymbol{g}$. By induction, we then prove $\boldsymbol{v}, p \in C^\infty(\bar{\Omega})$.

Existence Results. Homogeneous Boundary Conditions As we mentioned previously, there are, fundamentally, two kinds of approaches to show existence of weak solutions to (9)–(10), namely, the *finite-dimensional method* and the *function-analytic method*. We will first describe these methods in the case of homogeneous boundary conditions, $\boldsymbol{v}_* \equiv \boldsymbol{0}$, deferring the study of the non-homogeneous problem to Subsect. "Existence Results. Non-Homogeneous Boundary Conditions". In what follows, we shall refer to (9)–(10) with $\boldsymbol{v}_* = \boldsymbol{0}$ as (9)–$(10)_{\text{hom}}$.

A. The Finite-Dimensional Method This popular approach, usually called *Galerkin method*, was introduced by Fujita [28] and, independently, by Vorovich and Yudovich [96]. It consists in projecting (13) on a suitable finite dimensional space, V_N, of $\mathcal{D}_0^{1,2}(\Omega)$ and then in finding a solution, $\boldsymbol{v}_N \in V_N$, of the "projected" equation. One then passes to the limit $N \to \infty$ to show, with the help of an appropriate uniform estimate, that $\{\boldsymbol{v}_N\}$ contains at least one subsequence converging to some $\boldsymbol{v} \in \mathcal{D}_0^{1,2}(\Omega)$ that satisfies condition (13). Precisely, let $\{\boldsymbol{\psi}_k\} \subset \mathcal{D}(\Omega)$ be an orthonormal basis of $\mathcal{D}_0^{1,2}(\Omega)$, and set $\boldsymbol{v}_N = \sum_{i=1}^N c_{iN}\boldsymbol{\psi}_i$, where the coefficients c_{iN} are requested to be solutions of the following nonlinear algebraic system

$$\nu[\boldsymbol{v}_N, \boldsymbol{\psi}_k] - (\boldsymbol{v}_N\cdot\nabla\boldsymbol{\psi}_k, \boldsymbol{v}_N) = \langle \boldsymbol{f}, \boldsymbol{\psi}_k\rangle\,, \quad k = 1,\dots,N\,. \tag{15}$$

By means of the Brouwer fixed point theorem, it can be shown (see Lemma VIII.3.2 in [32]) that a solution $(c_{1N},\dots,c_{NN})$ to (15) exists, provided the following estimate holds

$$|\boldsymbol{v}_N|_{1,2} \le M \tag{16}$$

where M is a finite, positive quantity *independent of* N. Let us show that (16) indeed occurs. Multiplying through both sides of (15) by c_{kN}, summing over k from 1 to N, and observing that, by Lemma 1, $(\boldsymbol{v}_N\cdot\nabla\boldsymbol{v}_N, \boldsymbol{v}_N) = 0$, we obtain

$$\nu|\boldsymbol{v}_N|_{1,2}^2 = \langle \boldsymbol{f}, \boldsymbol{v}_N\rangle \le |\boldsymbol{f}|_{-1,2}|\boldsymbol{v}_N|_{1,2}\,,$$

which proves the desired estimate (16) with $M := |\boldsymbol{f}|_{-1,2}$. Notice that the validity of this latter estimate can be obtained by formally replacing in (13) $\boldsymbol{\varphi}$ with $\boldsymbol{v} \in \mathcal{D}_0^{1,2}(\Omega)$ and by using Lemma 1. From classical properties of Hilbert spaces and from (16), we can select a subsequence $\{\boldsymbol{v}_{N'}\}$ and find $\boldsymbol{v} \in \mathcal{D}_0^{1,2}(\Omega)$ such that

$$\lim_{N'\to\infty}[\boldsymbol{v}_{N'}, \boldsymbol{\varphi}] = [\boldsymbol{v}, \boldsymbol{\varphi}]\,, \quad \text{for all } \boldsymbol{\varphi} \in \mathcal{D}_0^{1,2}(\Omega)\,. \tag{17}$$

By Definition 1, to prove that $\boldsymbol{v}$ is a weak solution to (9)–$(10)_{\text{hom}}$ it remains to show that $\boldsymbol{v}$ satisfies (13). In view of the *Poincaré inequality* (Theorem II.4.1 in [31]):

$$\|\boldsymbol{\varphi}\|_2 \le c_P\,|\boldsymbol{\varphi}|_{1,2} \quad \boldsymbol{\varphi} \in \mathcal{D}_0^{1,2}(\Omega)\,, \quad c_P = c_P(\Omega) > 0\,, \tag{18}$$

by (16) it follows that $\{\boldsymbol{v}_N\}$ is bounded in $W_0^{1,2}(\Omega)$ and so, by Rellich compactness theorem (see Theorem II.4.2 in [31]) we can assume that $\{\boldsymbol{v}_{N'}\}$ converges to $\boldsymbol{v}$ in $L^4(\Omega)$:

$$\lim_{N'\to\infty}\|\boldsymbol{v}_{N'} - \boldsymbol{v}\|_4 = 0\,. \tag{19}$$

We now consider (15) with $N = N'$ and pass to the limit $N' \to \infty$. Clearly, by (17), we have for each fixed k

$$\lim_{N'\to\infty}[\boldsymbol{v}_{N'}, \boldsymbol{\psi}_k] = [\boldsymbol{v}, \boldsymbol{\psi}_k]\,. \tag{20}$$

Moreover, by Lemma 1 and by (19),

$$\lim_{N'\to\infty}(\boldsymbol{v}_{N'}\cdot\nabla\boldsymbol{\psi}_k, \boldsymbol{v}_{N'}) = (\boldsymbol{v}\cdot\nabla\boldsymbol{\psi}_k, \boldsymbol{v})\,,$$

and so, from this latter relation, from (20), and from (15) we conclude that

$$\nu\left[\boldsymbol{v},\boldsymbol{\psi}_k\right]-(\boldsymbol{v}\cdot\nabla\boldsymbol{\psi}_k,\boldsymbol{v})=\langle\boldsymbol{f},\boldsymbol{\psi}_k\rangle\,,\quad\text{for all } k\in\mathbb{N}\,,$$

Since $\{\boldsymbol{\psi}_k\}$ is a basis in $\mathcal{D}_0^{1,2}(\Omega)$, this latter relation, along Lemma 1, immediately implies that $\boldsymbol{v}$ satisfies (13), which completes the existence proof.

Remark 4 The Galerkin method provides existence of a weak solution corresponding to *any* given $\boldsymbol{f}\in\mathcal{D}_0^{-1,2}(\Omega)$. Moreover, it is *constructive*, in the sense that the solution can be obtained as the limit of a sequence of "approximate solutions" each of which can be, in principle, evaluated by solving the system of nonlinear algebraic Eqs. (15).

B. The Function-Analytic Method This approach, that goes back to the work of Leray [61,63] and of Ladyzhenskaya [58], consists, first, in re-writing (13) as a *nonlinear equation* in the Hilbert space $\mathcal{D}_0^{1,2}(\Omega)$, and then using an appropriate topological degree theory to prove existence of weak solutions. Though more complicated than, and not constructive like the Galerkin method, this approach has the advantage of furnishing, as a byproduct, significant information on the solutions set. By (14) and (18) we find

$$\mathcal{D}_0^{1,2}(\Omega)\subset L^4(\Omega)\,,\tag{21}$$

and so, from Lemma 1 and from the Riesz representation theorem, we have that, for each fixed $\boldsymbol{v}\in\mathcal{D}_0^{1,2}(\Omega)$, there exists $\boldsymbol{\mathcal{N}}(\boldsymbol{v})\in\mathcal{D}_0^{1,2}(\Omega)$ such that

$$-(\boldsymbol{v}\cdot\nabla\boldsymbol{\varphi},\boldsymbol{v})=[\boldsymbol{\mathcal{N}}(\boldsymbol{v}),\boldsymbol{\varphi}]\,,\quad\text{for all }\boldsymbol{\varphi}\in\mathcal{D}_0^{1,2}(\Omega)\,.\tag{22}$$

Likewise, we have $\langle\boldsymbol{f},\boldsymbol{\varphi}\rangle=[\boldsymbol{F},\boldsymbol{\varphi}]$, for some $\boldsymbol{F}\in\mathcal{D}_0^{1,2}(\Omega)$ and for all $\boldsymbol{\varphi}\in\mathcal{D}_0^{1,2}(\Omega)$. Thus, Eq. (13) can be equivalently re-written as

$$\boldsymbol{N}(\nu,\boldsymbol{v})=\boldsymbol{F}\,,\quad\text{in }\mathcal{D}_0^{1,2}(\Omega)\,,\tag{23}$$

where the map $\boldsymbol{N}$ is defined as follows

$$\boldsymbol{N}:(\nu,\boldsymbol{v})\in(0,\infty)\times\mathcal{D}_0^{1,2}(\Omega)\mapsto\nu\boldsymbol{v}+\boldsymbol{\mathcal{N}}(\boldsymbol{v})\in\mathcal{D}_0^{1,2}(\Omega)\,.\tag{24}$$

In order to show the above-mentioned property of solutions, we shall use some basic results related to Fredholm maps of index 0 [84]. This approach is preferable to that originally used by Leray – which applies to maps that are compact perturbations of homeomorphism – because it covers a larger class of problems, including flow in exterior domains [36].

Definition 2 A map $M:X\mapsto Y$, X, Y Banach spaces, is Fredholm, if and only if: (i) M is of class C^1 (in the sense of Fréchet differentiability) and denoted by $D_xM(x)$ its derivative at $x\in X$ (iii) the integers. The integers $\alpha:=\dim\{z\in X:[D_xM(x)](z)=0\}$ and $\beta:=\operatorname{codim}\{y\in Y:[D_xM(x)](z)=y\}$ for some $z\in X$ are both finite.

The integer $m:=\alpha-\beta$ is independent of the particular $x\in X$ (Sect. 5.15 in [100]), and is called the *index* of M.

Definition 3 A map $M:X\mapsto Y$, is said to be proper if $K_1:=\{x\in X:M(x)=y\,,y\in K\}$ is compact in X, whenever K is compact in Y.

By using the properties of proper Fredholm maps of index 0 [84], and of the associated *Caccioppoli–Smale degree* [10,84], one can prove the following result (see Theorem I.2.2 in [37])

Lemma 3 *Let M be a proper Fredholm map of index 0 and of class C^2, satisfying the following.*

(i) There exists $\bar{y}\in Y$ such that the equation $M(x)=\bar{y}$ has one and only one solution $\bar{x}$;
(ii) $[D_xM(\bar{x})](z)=0\Rightarrow\quad z=0$.

Then:

(a) M is surjective;
(b) There exists an open, dense set $Y_0\subset Y$ such that for any $y\in Y_0$ the solution set $\{x\in X:M(x)=y\}$ is finite and constituted by an odd number, $\kappa=\kappa(y)$, of points;
(c) The integer κ is constant on every connected component of Y_0.

The next result provides the required functional properties of the map $\boldsymbol{N}$.

Proposition 1 *The map $\boldsymbol{N}$ defined in (24) is of class C^∞. Moreover, for any $\nu>0$, $\boldsymbol{N}(\nu,\cdot):\mathcal{D}_0^{1,2}(\Omega)\mapsto\mathcal{D}_0^{1,2}(\Omega)$ is proper and Fredholm of index 0.*

Proof It is a simple exercise to prove that $\boldsymbol{N}$ is of class C^∞ (see Example I.1.6 in [37]). By the compactness of the embedding $W_0^{1,2}(\Omega)\subset L^4(\Omega)$ Theorem II.4.2 in [31], from Lemma 1 and from the definition of the map $\boldsymbol{\mathcal{N}}$ [see (22)], one can show that $\boldsymbol{\mathcal{N}}$ is compact, that is, it maps bounded sequences of $\mathcal{D}_0^{1,2}(\Omega)$ into relatively compact sequences. Therefore, $\boldsymbol{N}(\nu,\cdot)$ is a compact perturbation of a multiple of the identity operator. Moreover, by (22) and by Lemma 1 we show that $[\boldsymbol{\mathcal{N}}(\boldsymbol{v}),\boldsymbol{v}]=0$, which implies,

$$[\boldsymbol{N}(\nu,\boldsymbol{v}),\boldsymbol{v}]=\nu|\boldsymbol{v}|_{1,2}^2\,.\tag{25}$$

Using the Schwartz inequality on the left-hand side of (25), we infer that $|\boldsymbol{N}(\nu,\boldsymbol{v})|_{1,2} \geq \nu|\boldsymbol{v}|_{1,2}$ and so, for each fixed $\nu > 0$ we have $|\boldsymbol{N}(\nu,\boldsymbol{v})|_{1,2} \to \infty$ as $|\boldsymbol{v}|_{1,2} \to \infty$, that is, $\boldsymbol{N}(\nu,\cdot)$ is (weakly) *coercive*. Consequently, $\boldsymbol{N}(\nu,\cdot)$ is proper (see Theorem 2.7.2 in [7]). Finally, since the derivative of a compact map is compact (see Theorem 2.4.6 in [7]), the derivative map $D_{\boldsymbol{v}}\boldsymbol{N}(\nu,\boldsymbol{v}) \equiv \nu\boldsymbol{I} + D_{\boldsymbol{v}}\boldsymbol{N}(\boldsymbol{v})$, $\boldsymbol{I}$ identity operator, is, for each fixed $\nu > 0$, a compact perturbation of a multiple of the identity, which implies that $\boldsymbol{N}(\nu,\cdot)$ is Fredholm of index 0 (see Theorem 5.5.C in [100]). □

It is easy now to show that, for any fixed $\nu > 0$, $\boldsymbol{N}(\nu,\cdot)$ satisfies all the assumptions of Lemma 3. Actually, in view of Proposition 1, we have only to prove the validity of (i) and (ii). We take $\bar{y} \equiv \boldsymbol{0}$, and so, from (25) we obtain that the equation $\boldsymbol{N}(\nu,\boldsymbol{v}) = \boldsymbol{0}$ has only the solution $\bar{x} \equiv \boldsymbol{v} = \boldsymbol{0}$, so that (i) is satisfied. Furthermore, from (22) it follows that the equation $D_{\boldsymbol{v}}\boldsymbol{N}(\nu,0)(\boldsymbol{w}) = 0$ is equivalent to $\nu\,\boldsymbol{w} = \boldsymbol{0}$, and so also condition (ii) is satisfied for $\nu > 0$. Thus, we have proved the following result (see also [22]).

Theorem 1 *For any $\nu > 0$ and $\boldsymbol{F} \in \mathcal{D}_0^{1,2}(\Omega)$, Eq. (23) has at least one solution $\boldsymbol{v} \in \mathcal{D}_0^{1,2}(\Omega)$. Moreover, for each fixed $\nu > 0$, there exists open and dense $\mathcal{O} = \mathcal{O}(\nu) \subset \mathcal{D}_0^{1,2}(\Omega)$ with the following properties: (i) For any $\boldsymbol{F} \in \mathcal{O}$ the number of solutions to (23), $\mathfrak{n} = \mathfrak{n}(\boldsymbol{F},\nu)$ is finite and odd; (ii) the integer $\mathfrak{n}$ is constant on each connected component of $\mathcal{O}$.*

Next, for a given $\boldsymbol{F} \in \mathcal{D}_0^{1,2}(\Omega)$, consider the solution manifold

$$S(\boldsymbol{F}) = \left\{(\nu,\boldsymbol{v}) \in (0,\infty) \times \mathcal{D}_0^{1,2}(\Omega)\colon N(\nu,\boldsymbol{v}) = \boldsymbol{F}\right\}.$$

By arguments similar to those used in the proof of Theorem 1, one can show the following "generic" characterization of $S(\boldsymbol{F})$ see Example I.2.4 in [37] and also Chap. 10.3 in [93].

Theorem 2 *There exists dense $\mathcal{P} \subset \mathcal{D}_0^{1,2}(\Omega)$ such that, for every $\boldsymbol{F} \in \mathcal{P}$, the set $S(\boldsymbol{F})$ is a C^∞ 1-dimensional manifold. Moreover, there exists an open and dense subset of $(0,\infty)$, $\Lambda = \Lambda(\boldsymbol{F})$, such that for each $\nu \in \Lambda$, Eq. (23) has a finite number $\mathfrak{m} = \mathfrak{m}(\boldsymbol{F},\nu) > 0$ of solutions. Finally, the integer $\mathfrak{m}$ is constant on every open interval contained in Λ.*

In other words, Theorem 2 expresses the property that, for every $\boldsymbol{F} \in \mathcal{P}$, the set $S(\boldsymbol{F})$ is the union of smooth and *non-intersecting* curves. Furthermore, "almost all" lines $\nu = \nu_0 =$ const. intersect these curves at a *finite* number of points, $\mathfrak{m}(\nu_0,\boldsymbol{F})$, each of which is a solution to (23) corresponding to ν_0 and to $\boldsymbol{F}$. Finally, $\mathfrak{m}(\nu_0,\boldsymbol{F}) = \mathfrak{m}(\nu_1,\boldsymbol{F})$

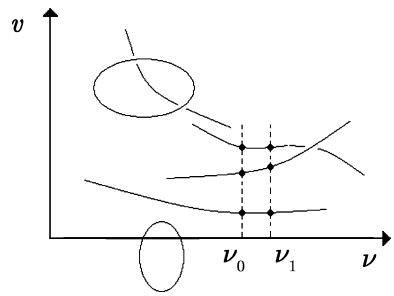

Navier–Stokes Equations: A Mathematical Analysis, Figure 3
Sketch of the manifold $S(\boldsymbol{F})$

whenever ν_0 and ν_1 belong to an open interval of a suitable dense set of $(0,\infty)$. A sketch of the manifold $S(\boldsymbol{F})$ is provided in Fig. 3.

Existence Results. Non-Homogeneous Boundary Conditions Existence of solutions to (9)–(10) when $\boldsymbol{v}_* \not\equiv \boldsymbol{0}$ leads to one of the most challenging open questions in the mathematical theory of the Navier–Stokes equations, in the case when the boundary $\partial\Omega$ is constituted by more than one connected component. In order to explain the problem, let S_i, $i = i,\ldots,K$, $K \geq 1$, denote these components. Conservation of mass $(9)_2$ along with Gauss theorem imply the following *compatibility condition* on the data $\boldsymbol{v}_*$

$$\sum_{i=1}^{K}\int_{S_i}\boldsymbol{v}_*\cdot\boldsymbol{n}_i \equiv \sum_{i=1}^{K}\Phi_i = 0\,, \tag{26}$$

where $\boldsymbol{n}_i$ denotes the outward unit normal to S_i. From the physical point of view, the quantity $\rho\,\Phi_i$ represents the mass flow-rate of the liquid through the surface S_i. Now, assuming $\boldsymbol{v}_*$ and Ω sufficiently smooth (for example, $\boldsymbol{v}_* \in W^{1/2,2}(\partial\Omega)$ and Ω locally Lipschitzian Sect. VIII.4 in [32]), we look for a weak solution to (9)–(10) in the form $\boldsymbol{v} = \boldsymbol{u} + \boldsymbol{V}$, where $\boldsymbol{V} \in W^{1,2}(\Omega)$ is an extension of $\boldsymbol{v}_*$ with $\operatorname{div}\boldsymbol{V} = 0$ in Ω. Thus, if we use, for example, the Galerkin method of Subsect. IV.1.1(A), from (15) we obtain that the "approximate solution" $\boldsymbol{u}_N = \sum_{i=1}^{N} c_{iN}\boldsymbol{\psi}_i$ must satisfy the following equations ($k = 1,\ldots,N$)

$$\begin{aligned}\nu[\boldsymbol{u}_N,\boldsymbol{\psi}_k] - (\boldsymbol{u}_N\cdot\nabla\boldsymbol{\psi}_k,\boldsymbol{u}_N) - (\boldsymbol{u}_N\cdot\nabla\boldsymbol{\psi}_k,\boldsymbol{V})\\ - (\boldsymbol{V}\cdot\nabla\boldsymbol{\psi}_k,\boldsymbol{u}_N) - (\boldsymbol{V}\cdot\nabla\boldsymbol{\psi}_k,\boldsymbol{V}) + \nu[\boldsymbol{V},\boldsymbol{\psi}_k]\\ = \langle\boldsymbol{f},\boldsymbol{\psi}_k\rangle.\end{aligned} \tag{27}$$

Therefore, existence of a weak solution will be secured provided we show that the sequence $\{\boldsymbol{v}_N := \boldsymbol{u}_N + \boldsymbol{V}\}$ satisfies the bound (16). In turn, this latter is equivalent to

showing

$$|\boldsymbol{u}_N|_{1,2} \le M_1 \,, \tag{28}$$

where M_1 is a finite, positive quantity *independent of* N. Multiplying through both sides of (27) by c_{kN}, summing over k from 1 to N, and observing that, by Lemma 1, $(\boldsymbol{u}_N \cdot \nabla \boldsymbol{u}_N, \boldsymbol{u}_N) = (\boldsymbol{V} \cdot \nabla \boldsymbol{u}_N, \boldsymbol{u}_N) = 0$, we find

$$\begin{aligned} &\nu |\boldsymbol{u}_N|_{1,2}^2 \\ &= (\boldsymbol{u}_N \cdot \nabla \boldsymbol{u}_N, \boldsymbol{V}) + (\boldsymbol{V} \cdot \nabla \boldsymbol{u}_N, \boldsymbol{V}) - [\boldsymbol{V}, \boldsymbol{u}_N] + \langle \boldsymbol{f}, \boldsymbol{u}_N \rangle \,. \end{aligned}$$

By using (14), the Hölder inequality and the Cauchy–Schwartz inequality

$$ab \le \varepsilon a^2 + (1/4\varepsilon) b^2, \quad a, b, \varepsilon > 0 \tag{29}$$

on the last three terms on the right-hand side of this latter equation, we easily find, for a suitable choice of ε,

$$\frac{\nu}{2} |\boldsymbol{u}_N|_{1,2}^2 \le (\boldsymbol{u}_N \cdot \nabla \boldsymbol{u}_N, \boldsymbol{V}) + C \,, \tag{30}$$

where $C = C(\boldsymbol{V}, \boldsymbol{f}, \Omega) > 0$. From (30) it follows that, in order to obtain the bound (28) *without restrictions on the magnitude of* ν, it suffices that $\boldsymbol{V}$ meets the following requirement:

$$\begin{aligned} &\text{Given } \varepsilon, \quad \text{there is } \boldsymbol{V} = \boldsymbol{V}(\varepsilon, \boldsymbol{x}) \in (0, \nu/4) \\ &\text{such that } (\boldsymbol{\varphi} \cdot \nabla \boldsymbol{\varphi}, \boldsymbol{V}) \le \varepsilon |\boldsymbol{\varphi}|_{1,2}^2 \quad \text{for all } \boldsymbol{\varphi} \in \mathcal{D}(\Omega) \,. \end{aligned} \tag{31}$$

As indicated by Leray pp. 28–30 in [61] and clarified by Hopf [52], if Ω is smooth enough and if $K = 1$, that is, if $\partial\Omega$ is constituted by only one connected component, S_1, it is possible to construct a family of extensions satisfying (31). Notice that, in such a case, condition (26) reduces to the single condition $\Phi_1 = 0$. If $K > 1$, the same construction is still possible but *with the limitation that* $\Phi_i = 0$, *for all* $i = 1, \ldots, K$. It should be emphasized that this condition is quite restrictive from the physical point of view, in that it does not allow for the presence, in the region of flow, of isolated "sources" and "sinks" of liquid. Nevertheless, one may wonder if, by using a different construction, it is still possible to satisfy (31). Unfortunately, as shown by Takeshita [90] by means of explicit examples, *in general, the existence of extensions satisfying* (31) *implies* $\Phi_i = 0$, *for all* $i = 1, \ldots, K$; see also § VIII.4 in [32].

We wish to emphasize that the same type of conclusion holds if, instead of the Galerkin method, we use the function-analytic approach; see [61], notes to Chap. VIII in [32].

Finally, it should be remarked that, in the special case of *two-dimensional* domains possessing suitable symmetry and of symmetric boundary data, Amick [1] and Fujita [29] have shown existence of corresponding symmetric solutions under the general assumption (26). However, we have the following.

Open Question Let Ω be a smooth domain in $\mathbb{R}^n$, $n = 2, 3$, with $\partial\Omega$ constituted by $K > 1$ connected components, S_i, $i = 1, \ldots, K$, and let $\boldsymbol{v}_*$ be any smooth field satisfying (26). *It is not known if the corresponding problem* (9)–(10) *has at least one solution.*

Uniqueness and Steady Bifurcation It is a well-established experimental fact that a steady flow of a viscous incompressible liquid is "observed", namely, it is unique and stable, if the magnitude of the driving force, usually measured through a dimensionless number $\lambda \in (0, \infty)$, say, is below a certain threshold, λ_c. However, if $\lambda > \lambda_c$, this flow becomes unstable and another, different flow is instead observed. This latter may be steady or unsteady (typically, time-periodic). In the former case, we say that a *steady bifurcation* phenomenon has occurred. From the physical point of view, bifurcation happens because the liquid finds a more "convenient" motion (than the original one) to dissipate the increasing energy pumped in by the driving force. From the mathematical point of view, bifurcation occurs when, roughly speaking, two solution-curves (parametrized with the appropriate dimensionless number) intersect. (As we know from Theorem 2, the intersection of these curves is not "generic".) It can happen that one curve exists for all values of λ, while the other only exists for $\lambda > \lambda_c$ (*supercritical bifurcation*). The point of intersection of the two curves is called the *bifurcation point*. Thus, in any neighborhood of the bifurcation point, we must have (at least) two distinct solutions and so a necessary condition for bifurcation is the occurrence of non-uniqueness. This section is dedicated to the above issues.

Uniqueness Results It is simple to show that, if $|\boldsymbol{v}|_{1,2}$ is not "too large", then $\boldsymbol{v}$ is the only weak solution to (9)–(10) corresponding to the given data.

Theorem 3 (Uniqueness) *Let* Ω *be locally Lipschitzian and let*

$$|\boldsymbol{v}|_{1,2} < \nu/\kappa \,, \tag{32}$$

with $\kappa = \kappa(\Omega) > 0$. *Then, there is only one weak solution to (9)–(10).*

Proof Let $\boldsymbol{v}, \boldsymbol{v}_1 = \boldsymbol{v} + \boldsymbol{u}$ be two different solutions. From (13) we deduce

$$\nu[\boldsymbol{u}, \boldsymbol{\varphi}] = (\boldsymbol{u} \cdot \nabla \boldsymbol{\varphi}, \boldsymbol{v}) + (\boldsymbol{v}_1 \cdot \nabla \boldsymbol{\varphi}, \boldsymbol{u}), \quad \text{for all } \boldsymbol{\varphi} \in \mathcal{D}(\Omega). \tag{33}$$

If Ω is locally Lipschitzian, then $\boldsymbol{u} \in \mathcal{D}_0^{1,2}(\Omega)$ (see Theorem II.3.2 and § II.3.5 in [31]), and since $\mathcal{D}(\Omega)$ is dense in $\mathcal{D}_0^{1,2}(\Omega)$, with the help of Lemma 1 we can replace $\boldsymbol{\varphi}$ with $\boldsymbol{u}$ in (33) to get

$$\nu |\boldsymbol{u}|_{1,2}^2 = (\boldsymbol{u} \cdot \nabla \boldsymbol{u}, \boldsymbol{v}).$$

We now use (14) and Lemma 1 on the right-hand side of this equation to find

$$(\nu - \kappa\, |\boldsymbol{v}|_{1,2}) |\boldsymbol{u}|_{1,2}^2 \le 0,$$

with $\kappa = \kappa(\Omega) > 0$, which proves $\boldsymbol{u} = 0$ in $\mathcal{D}_0^{1,2}(\Omega)$, namely, uniqueness, if $\boldsymbol{v}$ satisfies (33). □

Remark 5 As we have seen previously, if $\boldsymbol{v}_* \equiv \boldsymbol{0}$, weak solutions satisfy $\nu |\boldsymbol{v}|_{1,2} \le |\boldsymbol{f}|_{-1,2}$. Thus, (32) holds if $|\boldsymbol{f}|_{-1,2} \le \nu^2/\kappa$. If $\boldsymbol{v}_* \ne \boldsymbol{0}$, then one can show that (32) holds if $|\boldsymbol{f}|_{-1,2} + (1+\nu)\|\boldsymbol{v}_*\|_{1/2,2,\partial\Omega} + \|\boldsymbol{v}_*\|_{1/2,2,\partial\Omega}^2 \le \nu^2/\kappa_1$, where κ_1 has the same properties as κ; (see Theorem VIII.4.2 in [32]). Notice that these conditions are satisfied if a suitable non-dimensional parameter $\lambda \sim \left(|\boldsymbol{f}|_{-1,2} + \|\boldsymbol{v}_*\|_{1/2,2,\partial\Omega}\right)$ is "sufficiently small".

Remarkably enough, one can give *explicit examples of non-uniqueness*, if condition (32) is violated. More specifically, we have the following result (see Theorem VIII.2.2 in [32]).

Theorem 4 (Non-Uniqueness) *Let Ω be a bounded smooth body of revolution around an axis r, that does not include points of r. For example, Ω is a torus of arbitrary bounded smooth section. Then there are smooth fields $\boldsymbol{f}$ and $\boldsymbol{v}_*$ and a value of $\nu > 0$ such that problem (9)–(10) corresponding to these data admits at least two distinct and smooth solutions.*

Some Bifurcation Results By using the functional setting introduced in Sect. "B. The Function-Analytic Method", it is not difficult to show that steady bifurcation can be reduced to the study of a suitable nonlinear eigenvalue problem in the space $\mathcal{D}_0^{1,2}(\Omega)$. To this end, we recall certain basic definitions.

Let U be an open interval of $\mathbb{R}$ and let

$$M : (x, \mu) \in X \times U \mapsto Y. \tag{34}$$

Definition 4 The point (x_0, μ_0) is called a bifurcation point of the equation

$$M(x, \mu) = 0 \tag{35}$$

if and only if (a) $M(x_0, \mu_0) = 0$, and (b) there are (at least) two sequences of solutions, $\{(x_m, \mu_m)\}$ and $\{(x_m^*, \mu_m)\}$, to (35), with $x_m \ne x_m^*$, for all $m \in \mathbb{N}$, such that $(x_m, \mu_m) \to (x_0, \mu_0)$ and $(x_m^*, \mu_m) \to (x_0, \mu_0)$ as $m \to \infty$.

If M is suitably smooth around (x_0, μ_0), a necessary condition for (x_0, μ_0) to be a bifurcation point is that $D_x M(x_0, \mu_0)$ is not a bijection. In fact, we have the following result which is an immediate corollary of the implicit function theorem (see, e. g., Lemma III.1.1 in [37]).

Lemma 4 *Suppose that $D_x M$ exists in a neighborhood of (x_0, μ_0), and that both M and $D_x M$ are continuous at (x_0, μ_0). Then, if (x_0, μ_0) is a bifurcation point of (34), $D_x M(x_0, \mu_0)$ is not a bijection. If, in particular, $D_x M(x_0, \mu_0)$ is a Fredholm operator of index 0 (see Definition 2), then the equation $D_x M(x_0, \mu_0)x = 0$ has at least one nonzero solutions.*

Let $\boldsymbol{v}_0 = \boldsymbol{v}_0(\nu)$, $\nu \in (0, \infty)$ be a family of weak solution to (9)–(10) corresponding to given data $\boldsymbol{f}$ and $\boldsymbol{v}_*$. We assume that $\boldsymbol{f}$ and $\boldsymbol{v}_*$ are fixed. Denoting by $\boldsymbol{v}$ any other solution corresponding to the same data, by an argument completely similar to that leading to (23), we find that $\boldsymbol{u} := \boldsymbol{v} - \boldsymbol{v}_0$, satisfies the following equation in $\mathcal{D}_0^{1,2}(\Omega)$

$$\nu \boldsymbol{u} + \boldsymbol{B}(\boldsymbol{v}_0)(\boldsymbol{u}) + \boldsymbol{\mathcal{N}}(\boldsymbol{u}) = \boldsymbol{0}, \tag{36}$$

with $\boldsymbol{B}(\boldsymbol{v}_0) := D_{\boldsymbol{v}} \boldsymbol{\mathcal{N}}(\boldsymbol{v}_0)$ and $\boldsymbol{\mathcal{N}}$ defined in (22). Obviously, $(\boldsymbol{v}_0(\nu_0), \nu_0)$ is a bifurcation point for the original equation if and only if $(\boldsymbol{0}, \nu_0)$ is such for (36). Thus, in view of Lemma 4, a necessary condition for $(\boldsymbol{0}, \nu_0)$ to be a bifurcation point for (36) is that the equation

$$\nu_0 \boldsymbol{v} + \boldsymbol{B}(\boldsymbol{v}_0(\nu_0))(\boldsymbol{v}) = \boldsymbol{0} \tag{37}$$

has at least one nonzero solution $\boldsymbol{v} \in \mathcal{D}_0^{1,2}(\Omega)$.

In several significant situations it happens that, after a suitable non-dimensionalization of (36), the family of solutions $\boldsymbol{v}_0(\nu)$, $\nu \in (0, \infty)$ is independent of the parameter ν which, this time, has to be interpreted as the inverse of an appropriate dimensionless number (*Reynolds number*), like, for instance, in the Taylor–Couette problem; see the following subsection. Now, from Proposition 1, we know that $\boldsymbol{B}(\boldsymbol{u})$ is compact at each $\boldsymbol{u} \in \mathcal{D}_0^{1,2}(\Omega)$, so that $\nu \boldsymbol{I} + \boldsymbol{B}(\boldsymbol{u})$ is Fredholm of index 0, at each $\boldsymbol{u} \in \mathcal{D}_0^{1,2}(\Omega)$, for all $\nu > 0$ (see Theorem 5.5.C in [100]). Therefore, whenever $\boldsymbol{v}_0$ does not depend on ν, in a neighborhood of

ν_0, from Lemma 4 we find that a *necessary condition for* $(\boldsymbol{v}_0, \nu_0)$ *to be a bifurcation point for* (35) *is that* ν_0 *is an eigenvalue of the (compact) linear operator* $\boldsymbol{B}(\boldsymbol{v}_0)$.

The stated condition becomes also *sufficient*, provided we make the additional assumptions that ν_0 is a *simple eigenvalue* of $\boldsymbol{B}(\boldsymbol{v}_0)$. This is a consequence of the following theorem (see, e. g., Lemma III.1.2 in [37]).

Theorem 5 *Let* $X \subset Y$ *and let the operator* M *in* (34) *be of the form* $M = \mu I + T$, *where* I *is the identity in* X *and* T *is of class* C^1. *Furthermore, set* $L := D_x T(0)$. *Suppose that* $\mu_0 I + L$ *is Fredholm of index 0 (Definition 2), for some* $\mu_0 \in U$, *and that* $-\mu_0$ *is a simple eigenvalue for* L, *namely, the equation* $\mu_0 x + L(x) = 0$ *has one and only one (nonzero) independent solution,* x_1, *while the equation* $\mu_0 x + L(x) = x_1$ *has no solutions. Then,* $(0, \mu_0)$ *is a bifurcation point for the equation* $M(x, \mu) = 0$.

Bifurcation of Taylor–Couette Flow A notable application of Theorem 5 is the *bifurcation of Taylor–Couette flow*. In this case the liquid fills the space between two coaxial, infinite cylinders, C_1 and C_2, of radii R_1 and $R_2 > R_1$, respectively. C_1 rotates around the common axis, a, with constant angular velocity $\boldsymbol{\omega}$, while C_2 is at rest. Denote by (r, θ, z) a system of cylindrical coordinates with z along a and oriented as $\boldsymbol{\omega}$ and let $(\boldsymbol{e}_r, \boldsymbol{e}_\theta, \boldsymbol{e}_z)$ the associated canonical base. The components of a vector $\boldsymbol{w}$ in such a base are denoted by w_r, w_θ and w_z, respectively. If we introduce the non-dimensional quantities

$$\boldsymbol{u} = \boldsymbol{v}/(\omega r_1)\,,\ \tilde{\boldsymbol{x}} = \boldsymbol{x}/r_1\,,\ \mathsf{p} = p/\left(\rho\omega^2 r_1^2\right)\,,\ R = R_2/R_1\,,$$

with $\omega = |\boldsymbol{\omega}|$, we see at once that the following velocity and pressure fields

$$\boldsymbol{u}_0 = \left(1 - R^2\right)^{-1}\left(r - R^2/r\right)\boldsymbol{e}_\theta\,, \quad \mathsf{p}_0 = |\boldsymbol{u}_0|^2 \ln r + \text{const}\,, \tag{38}$$

solve (9), with $\boldsymbol{f} = \boldsymbol{0}$, for all values of the Reynolds number $\lambda := \rho\,\omega\, r_1^2/\mu$. Moreover, $\boldsymbol{u}_0$ satisfies the boundary conditions

$$\boldsymbol{u}_0(r) = \boldsymbol{0} \quad \text{at } r = 1\,, \quad \boldsymbol{u}_0(r) = \boldsymbol{e}_\theta \quad \text{at } r = R\,. \tag{39}$$

Experiments show that, if λ exceeds a critical value, λ_c, a flow with entirely different features than the flow (38) is observed. In fact, this new flow is dominated by large toroidal vortices, stacked one on top of the other, called *Taylor vortices*. They are periodic in the z-direction and axisymmetric (independent of θ); see Fig. 4. Therefore, we look for bifurcating solutions of the form $\boldsymbol{u}_0(r) + \boldsymbol{w}(r, z)$,

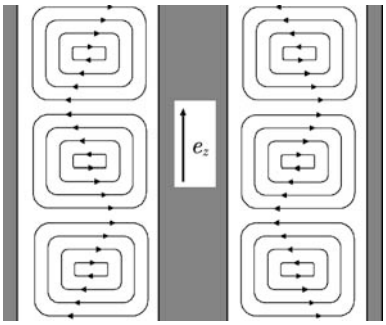

Navier–Stokes Equations: A Mathematical Analysis, Figure 4
Sketch of the streamlines of Taylor vortices, at the generic section θ=const

$\mathsf{p}_0(r) + \mathsf{p}(r, z)$, satisfying (39), and where $\boldsymbol{w}$ and p are periodic in the z-direction with period $\mathfrak{P}$, and $\boldsymbol{w}$ satisfies the following parity conditions:

$$w_r(r, z) = w_r(r, -z)\,, \quad w_\theta(r, z) = w_\theta(r, -z)\,, \quad w_z(r, z) = -w_z(r, -z)\,. \tag{40}$$

Moreover, the relevant region of flow becomes the *periodicity cell* $\Omega := (1, R) \times (0, \mathfrak{P})$. If we now introduce the *stream function* ψ

$$\frac{\partial \psi}{\partial z} = w_r\,, \quad \frac{\partial (r\psi)}{\partial r} = -r\, w_z\,; \quad \psi(r, z) = -\psi(r, -z)\,, \tag{41}$$

and the vector $\boldsymbol{u} := (\psi, w_\theta)$, it can be shown (see § 72.7 in [99]) that $\boldsymbol{u}$ satisfies an equation of the type (36), namely,

$$\boldsymbol{u} + \lambda\left(\tilde{\boldsymbol{B}}(\boldsymbol{v}_0)(\boldsymbol{u}) + \tilde{\mathcal{N}}(\boldsymbol{u})\right) = \boldsymbol{0} \quad \text{in } \mathcal{H}(\Omega) \tag{42}$$

where the operators $\tilde{\boldsymbol{B}}(\boldsymbol{v}_0)$ and $\tilde{\mathcal{N}}$ obey the same functional properties as $\boldsymbol{B}(\boldsymbol{v}_0)$ and $\mathcal{N}$, and where $\mathcal{H}(\Omega)$ is the Banach space of functions $\boldsymbol{u} := (\psi, w_\theta) \in C^{4,\alpha}(\Omega) \times C^{2,\alpha}(\Omega)$, $\alpha \in (0, 1)$, such that: (i) $\boldsymbol{u}(r, 0) = \boldsymbol{u}(r, \mathfrak{P})$ for all $r \in (1, R)$, (ii) $\boldsymbol{u}$ satisfies the parity conditions in (40), (41), and (iii) $\psi = \partial\psi/\partial r = w_\theta = 0$ at $r = 1, R$. Thus, in view of Theorem 5 and of the properties of the operators involved in (42), we will obtain that $(\boldsymbol{0}, \lambda_0)$ is a bifurcation point for (41) if the following two conditions are met

(a) $\boldsymbol{u} + \lambda_0 \tilde{\boldsymbol{B}}(\boldsymbol{v}_0)(\boldsymbol{u}) = \boldsymbol{0}$
has one and only one independent solution,
$\boldsymbol{u}_1 \in \mathcal{H}(\Omega)$; (43)

(b) the equation $\boldsymbol{u} + \lambda_0 \tilde{\boldsymbol{B}}(\boldsymbol{v}_0)(\boldsymbol{u}) = \boldsymbol{u}_1$
has no solution in $\mathcal{H}(\Omega)$.

It is in Lemma 72.14 in [99] that there exists a period $\mathfrak{P}$ for which both conditions in (43) are satisfied. In addition, for all $\lambda \in (0, \lambda_0)$ the equation $\boldsymbol{u} + \lambda \bar{\boldsymbol{B}}(\boldsymbol{v}_0)(\boldsymbol{u}) = \boldsymbol{0}$ has only the trivial solution in $\mathcal{H}(\Omega)$, which, by Lemma 4, implies that no bifurcation occurs for $\lambda \in (0, \lambda_0)$, which, in turn, means that the bifurcation is *supercritical*.

Flow in Exterior Domains

One of the most significant questions in fluid mechanics is to determine the properties of the steady flow of a liquid past a body $\mathcal{B}$ of simple symmetric shape (such as a sphere or cylinder), over the entire range of the Reynolds number $\lambda := \rho UL/\mu \in (0, \infty)$; see, e. g., [§ 4.9] in [4]. Here L is a length scale representing the linear dimension of $\mathcal{B}$, while $\boldsymbol{v}_\infty = -U\boldsymbol{e}_1$, $U = \text{const} > 0$, is the uniform velocity field of the liquid at large distances from $\mathcal{B}$. In the mathematical formulation of this problem, one assumes that the liquid fills the whole space, Ω, outside the closure of the domain $\mathcal{B} \subset \mathbb{R}^n$, $n = 2, 3$. Thus, by scaling $\boldsymbol{v}$ by U and $\boldsymbol{x}$ by L, from (9)–(10) we obtain that the steady flow past a body consists in solving the following non-dimensional *exterior boundary-value problem*

$$\left.\begin{aligned} -\Delta \boldsymbol{w} - \lambda \frac{\partial \boldsymbol{w}}{\partial x_1} + \lambda \boldsymbol{w} \cdot \nabla \boldsymbol{w} = -\nabla p + \boldsymbol{f} \\ \operatorname{div} \boldsymbol{w} = 0 \end{aligned}\right\} \quad \text{in } \Omega \tag{44}$$

$$\boldsymbol{w}(x) = \boldsymbol{e}_1\,, \quad x \in \partial\Omega\,, \quad \lim_{|x| \to \infty} \boldsymbol{w}(x) = \boldsymbol{0}\,,$$

where $\boldsymbol{w} := \boldsymbol{v} + \boldsymbol{e}_1$ and where we assume, for simplicity, that $\boldsymbol{v}_* \equiv \boldsymbol{0}$. (However, all the stated results continue to hold, more generally, if $\boldsymbol{v}_*$ belongs to a suitable trace space.)

Whereas in the three-dimensional case (e. g., *flow past a sphere*) the investigation of (44) is, to an extent, complete, in the two-dimensional case (*e. g., flow past a circular cylinder*) there are still fundamental unresolved issues. We shall, therefore, treat the two cases separately.

We need some preliminary considerations.

Extension of the Boundary Data If Ω is smooth enough (locally Lipschitzian, for example), since $\int_{\partial\Omega} \boldsymbol{e}_1 \cdot \boldsymbol{n} = 0$, by what we observed in Sect. "Existence Results. Non-Homogeneous Boundary Conditions", for any $\lambda > 0$ we find $\boldsymbol{V} = \boldsymbol{V}(\lambda, \boldsymbol{x})$, with $\operatorname{div} \boldsymbol{V} = 0$ in Ω, and satisfying (31) with $\varepsilon = 1/(2\lambda)$, say. Actually, as shown in Proposition 3.1 in [36], the extension $\boldsymbol{V}$ can be chosen, in particular, to be of class $C^\infty((0, \infty) \times \Omega)$ and such that the support of $V(\lambda, \cdot)$ is contained in a bounded set, independent of λ. The proof given in [36] is for $\Omega \subset \mathbb{R}^3$, but it can be easily extended to the case $\Omega \subset \mathbb{R}^2$.

Variational Formulation and Weak Solutions Setting $\boldsymbol{u} = \boldsymbol{w} - \mathbf{V}$ in $(44)_1$, dot-multiplying through both sides of this equation by $\boldsymbol{\varphi} \in \mathcal{D}(\Omega)$, integrating by parts and taking into account 12, we find

$$\begin{aligned} &[\boldsymbol{u}, \boldsymbol{\varphi}] - \lambda \left(\frac{\partial \boldsymbol{u}}{\partial x_1}, \boldsymbol{\varphi} \right) - \lambda\, (\boldsymbol{u} \cdot \nabla \boldsymbol{\varphi}, \boldsymbol{u}) - \lambda \{ (\boldsymbol{u} \cdot \nabla \boldsymbol{\varphi}, \mathbf{V}) \\ &+ (\mathbf{V} \cdot \nabla \boldsymbol{\varphi}, \boldsymbol{u}) \} + (\mathbf{H}, \boldsymbol{\varphi}) = \langle \boldsymbol{f}, \boldsymbol{\varphi} \rangle\,, \quad \text{for all } \boldsymbol{\varphi} \in \mathcal{D}(\Omega)\,. \end{aligned} \tag{45}$$

where

$$\mathbf{H} = \mathbf{H}(\mathbf{V}) := \Delta \mathbf{V} + \lambda \frac{\partial \mathbf{V}}{\partial x_1} - \lambda \mathbf{V} \cdot \nabla \mathbf{V}\,.$$

Definition 5 A function $\boldsymbol{w} \in D^{1,2}(\Omega)$ is called a weak solution to (44) if and only if $\boldsymbol{w} = \boldsymbol{u} + \mathbf{V}$ where $\boldsymbol{u} \in \mathcal{D}_0^{1,2}(\Omega)$ and $\boldsymbol{u}$ satisfies (45).

Regularity of Weak Solutions

(A) *Local Regularity*. The proof of differentiability properties of weak solutions can be carried out in a way similar to the case of a bounded domain. In particular, if $\boldsymbol{f}$ and Ω are of class C^∞, one can show that $\boldsymbol{w} \in C^\infty(\bar{\Omega})$ and there exists a scalar field $p \in C^\infty(\bar{\Omega})$ such that $(44)_{1,2,3}$ holds in the ordinary sense. For this and other intermediate regularity results, we refer to Theorem IX.1.1 in [32].

(B) *Regularity at Infinity*. The study of the validity of $(44)_4$ and, more particularly, of the *asymptotic structure* of weak solutions, needs a much more involved treatment, and the results depend on whether the flow is three- or two-dimensional. In the *three-dimensional case*, if $\boldsymbol{f}$ satisfies suitable summability assumptions outside a ball of large radius, by using a *local* representation of a weak solution, (namely, at all points of a ball of radius 1 and centered at $\boldsymbol{x} \in \Omega$, $\operatorname{dist}(\boldsymbol{x}, \partial\Omega) > 1$) one can prove that $\boldsymbol{w}$ and all its derivatives tend to zero at infinity uniformly pointwise and that a similar property holds for p (see Theorem X.6.1 in [32]). The starting point of this analysis is the crucial fact that

$$\begin{aligned} &\text{if } \Omega \subset \mathbb{R}^3\,, \\ &\mathcal{D}_0^{1,2}(\Omega) \text{ is continuously embedded in } L^6(\Omega)\,, \end{aligned} \tag{46}$$

which implies that $\boldsymbol{w}$ tends to zero at infinity in a suitable sense. However, the existence of the "wake" behind the body along with the *sharp* order of decay of $\boldsymbol{w}$ and p requires a more complicated analysis based on

the *global* representation of a weak solution (namely, at all points outside a ball of sufficiently large radius) by means of the *Oseen fundamental tensor*, along with maximal regularity estimates for the solution of the linearized *Oseen problem*, this latter being obtained by suppressing the nonlinear term $\boldsymbol{w}\cdot\nabla\boldsymbol{w}$ in (44) [§§ IX.6, IX.7 and IX.8] in [32]. In particular, one can prove the following result concerning the behavior of $\boldsymbol{w}$; (see Theorems IX.7.1, IX.8.1 and Remark IX.8.1 in [32]).

Theorem 6 *Let Ω be a three-dimensional exterior domain of class C^2, and let $\boldsymbol{w}$ be a weak solution to (44) corresponding to a given $\boldsymbol{f}\in L^q(\Omega)$, for all $q\in(1,q_0]$, $q_0>3$. Then*

$$\boldsymbol{w}\in L^r(\Omega)\,,\quad \textit{if and only if } r>2\,. \tag{47}$$

If, in addition, $\boldsymbol{f}$ is of bounded support, then, denoting by θ the angle made by a ray starting from the origin of coordinates (taken, without loss of generality, in $\mathbb{R}^3-\bar{\Omega}$) with the positively directed x_1-axis, we have

$$|\boldsymbol{w}(\boldsymbol{x})|\le\frac{M}{|\boldsymbol{x}|\left[1+|\boldsymbol{x}|(1+\cos\theta)\right]}\,,\quad \boldsymbol{x}\in\Omega\,, \tag{48}$$

where $M=M(\lambda,\Omega)>0$.

Remark 6 Two significant consequences of Theorem 6 are: (i) *the total kinetic energy of the flow past an obstacle* ($\equiv\frac{1}{2}\rho\|\boldsymbol{w}\|_2^2$) *is infinite*, see (47); and (ii) *the asymptotic decay of $\boldsymbol{w}$ is faster outside any semi-infinite cone with its axis coinciding with the negative x_1 -axis (existence of the "wake")*; see (48).

The study of the asymptotic properties of solutions in the two-dimensional case is deferred till Sect. "Two-Dimensional Flow. The Problem of Existence".

Three-Dimensional Flow. Existence of Solutions and Related Properties As in the case of a bounded domain, we may use two different approaches to the study of existence of weak solutions.

A. Finite-Dimensional Method Assume $\boldsymbol{f}\in\mathcal{D}_0^{-1,2}(\Omega)$. With the same notation as in Subsect. "A. The Finite-Dimensional Method", we look for "approximate solutions" to (45) of the form $\boldsymbol{u}_N=\sum_{i=1}^N c_{iN}\boldsymbol{\psi}_i$, where

$$\begin{aligned}&\left[\boldsymbol{u}_N,\boldsymbol{\psi}_k\right]-\lambda\left(\frac{\partial\boldsymbol{u}}{\partial N_{x_1}},\boldsymbol{\psi}_k\right)-\lambda\left(\boldsymbol{u}_N\cdot\nabla\boldsymbol{\psi}_k,\boldsymbol{u}_N\right)\\&-\lambda\{(\boldsymbol{u}_N\cdot\nabla\boldsymbol{\psi}_k,\boldsymbol{V})+(\boldsymbol{V}\cdot\nabla\boldsymbol{\psi}_k,\boldsymbol{u}_N)\}+(\boldsymbol{H},\boldsymbol{\psi}_k)\\&\qquad=\langle\boldsymbol{f},\boldsymbol{\psi}_k\rangle\,,\quad k=1,\dots,N\,.\end{aligned} \tag{49}$$

As in the case of a bounded domain, existence to the algebraic system (49), in the unknowns c_{iN}, will be achieved if we show the uniform estimate (28). By dot-multiplying through both sides of (49) by c_{kN}, by summing over k between 1 and N and by using Lemma 1 we find

$$|\boldsymbol{u}_N|_{1,2}^2=\lambda\left(\boldsymbol{u}_N\cdot\nabla\boldsymbol{u}_N,\boldsymbol{V}\right)-(\boldsymbol{H},\boldsymbol{u}_N)+\langle\boldsymbol{f},\boldsymbol{u}_N\rangle\,. \tag{50}$$

From the properties of the extension of the boundary data, we have that $\boldsymbol{V}$ satisfies (31) with $\varepsilon=1/(2\lambda)$ and that, moreover $-(\boldsymbol{H},\boldsymbol{u}_N)\le C|\boldsymbol{u}_N|_{1,2}$, for some $C=C(\Omega,\lambda)>0$. Thus, from this and from (50) we deduce the uniform bound (28). This latter implies the existence of a subsequence $\{\boldsymbol{u}_{N'}\}$ converging to some $\boldsymbol{u}\in\mathcal{D}_0^{1,2}(\Omega)$ weakly. Moreover, by Rellich theorem, $\boldsymbol{u}_{N'}\to\boldsymbol{u}$ in $L^4(K)$, where $K:=\Omega\cap\{|\boldsymbol{x}|>\rho\}$, all sufficiently large and finite $\rho>0$ (see Theorem II.4.2 in [31]). Consequently, recalling that $\boldsymbol{\psi}_k$ is of compact support in Ω, we proceed as in the bounded domain case and take the limit $N\equiv N'\to\infty$ in (49) to show that $\boldsymbol{u}$ satisfies (45) with $\boldsymbol{\varphi}\equiv\boldsymbol{\psi}_k$. Successively, by taking into account that every $\boldsymbol{\varphi}\in\mathcal{D}(\Omega)$ can be approximated in $L^3(\Omega)$ by linear combinations of $\boldsymbol{\psi}_k$ (see Lemma VII.2.1 in [31]) by (46) and by Lemma 1 we conclude that $\boldsymbol{u}$ satisfies (45). Notice that, as in the case of flow in bounded domains, this method furnishes existence for any $\lambda>0$ and any $\boldsymbol{f}\in\mathcal{D}_0^{-1,2}(\Omega)$.

B. The Function-Analytic Method As in the case of flow in a bounded domain, Subsect. "The Function-Analytic Method"ion IV.1.(B), our objective is to rewrite (45) as a nonlinear operator equation in an appropriate Banach space, where the relevant operator satisfies the assumptions of Lemma 3. In this way, we may draw the same conclusions of Theorem 1 also in the case of a flow past an obstacle. However, *unlike the case of flow in a bounded domain*, the map

$$\boldsymbol{\varphi}\in\mathcal{D}(\Omega)\mapsto(\boldsymbol{u}\cdot\nabla\boldsymbol{\varphi},\boldsymbol{u})\in\mathbb{R}$$

can *not* be extended to a linear, bounded functional in $\mathcal{D}_0^{1,2}(\Omega)$, if $\boldsymbol{u}$ *merely belongs to* $\mathcal{D}_0^{1,2}(\Omega)$. Analogous conclusion holds for the map $\boldsymbol{\varphi}\in\mathcal{D}(\Omega)\mapsto(\partial\boldsymbol{u}/\partial x_1,\boldsymbol{\varphi})\in\mathbb{R}$. The reason is because, in an exterior domain, the Poincaré inequality (18) and, consequently, the embedding 21 are, in general, not true. It is thus necessary to consider the above functionals for $\boldsymbol{u}$ in a space *strictly contained* in $\mathcal{D}_0^{1,2}(\Omega)$. Set

$$[\![\boldsymbol{u}]\!]:=\sup_{\boldsymbol{\varphi}\in\mathcal{D}(\Omega)}\frac{\left|\left(\frac{\partial\boldsymbol{u}}{\partial x_1},\boldsymbol{\varphi}\right)\right|}{|\boldsymbol{\varphi}|_{1,2}}$$

and let

$$X = X(\Omega) := \left\{ \boldsymbol{u} \in \mathcal{D}_0^{1,2}(\Omega) : \quad [\![\boldsymbol{u}]\!] < \infty \right\}.$$

Clearly, $X(\Omega)$ endowed with the norm $|\cdot|_{1,2} + [\![\cdot]\!]$ is a Banach space. Moreover, $X(\Omega) \subset L^4(\Omega)$, continuously; see Proposition 1.1 in [36]. We may thus conclude, by Riesz theorem, by Hölder inequality and by the properties of the extension $\boldsymbol{V}$ that, for any $\boldsymbol{u} \in X(\Omega)$, there exist $\boldsymbol{L}(\boldsymbol{u})$, $\boldsymbol{\mathcal{M}}(\boldsymbol{u})$, $\boldsymbol{\mathcal{V}}(\boldsymbol{u})$, and $\boldsymbol{\mathcal{H}}(\lambda)$ in $\mathcal{D}_0^{1,2}(\Omega)$ such that, for all $\boldsymbol{\varphi} \in \mathcal{D}(\Omega)$,

$$\begin{aligned} -\left(\frac{\partial \boldsymbol{u}}{\partial x_1}, \boldsymbol{\varphi}\right) &= [\boldsymbol{L}(\boldsymbol{u}), \boldsymbol{\varphi}]\,; \\ -(\boldsymbol{u}\cdot\nabla\boldsymbol{\varphi}, \boldsymbol{u}) &= [\boldsymbol{\mathcal{M}}(\boldsymbol{u}), \boldsymbol{\varphi}]\,; \\ -(\boldsymbol{u}\cdot\nabla\boldsymbol{\varphi}, \boldsymbol{V}) - (\boldsymbol{V}\cdot\nabla\boldsymbol{\varphi}, \boldsymbol{u}) &= [\boldsymbol{\mathcal{V}}(\boldsymbol{u}), \boldsymbol{\varphi}]\,; \\ (\boldsymbol{H}, \boldsymbol{\varphi}) &= [\boldsymbol{\mathcal{H}}(\lambda), \boldsymbol{\varphi}]\,. \end{aligned}$$

Consequently, we obtain that (45) is equivalent to the following equation

$$\boldsymbol{M}(\lambda, \boldsymbol{u}) = \boldsymbol{F} \quad \text{in } \mathcal{D}_0^{1,2}(\Omega)\,, \tag{51}$$

where

$$\begin{aligned} &\boldsymbol{M}\colon (\lambda, \boldsymbol{u}) \in (0,\infty)\times X(\Omega) \\ &\mapsto \boldsymbol{u} + \lambda\left(\boldsymbol{L}(\boldsymbol{u}) + \boldsymbol{\mathcal{V}}(\boldsymbol{u}) + \boldsymbol{\mathcal{M}}(\boldsymbol{u})\right) + \boldsymbol{\mathcal{H}}(\lambda) \in \mathcal{D}_0^{1,2}(\Omega)\,. \end{aligned}$$

A detailed study of the properties of the operator $\boldsymbol{M}$ is done in § 5 in [36], where, in particular, the following result is proved.

Lemma 5 *The operator $\boldsymbol{M}$ is of class C^∞. Moreover, for each $\lambda > 0$, $\boldsymbol{M}(\lambda, \cdot) : X(\Omega) \mapsto \mathcal{D}_0^{1,2}(\Omega)$ is proper and Fredholm of index 0, and the two equations $\boldsymbol{M}(\lambda, \boldsymbol{u}) = \boldsymbol{\mathcal{H}}(\lambda)$ and $D_{\boldsymbol{u}}\boldsymbol{M}(\lambda, \boldsymbol{0})(\boldsymbol{w}) = \boldsymbol{0}$ only have the solutions $\boldsymbol{u} = \boldsymbol{w} = \boldsymbol{0}$.*

From this lemma and with the help of Lemma 3, we obtain the following result analogous to Theorem 1.

Theorem 7 *For any $\lambda > 0$ and $\boldsymbol{F} \in \mathcal{D}_0^{1,2}(\Omega)$ the Eq. (51) has at least one solution $\boldsymbol{u} \in X(\Omega)$. Moreover, for each fixed $\lambda > 0$, there exists open and dense $\mathcal{Q} = \mathcal{Q}(\lambda) \subset \mathcal{D}_0^{1,2}(\Omega)$ with the following properties: (i) For any $\boldsymbol{F} \in \mathcal{Q}$ the number of solutions to (51), $\mathfrak{n} = \mathfrak{n}(\boldsymbol{F}, \lambda)$ is finite and odd; (ii) the integer $\mathfrak{n}$ is finite on each connected component of $\mathcal{Q}$.*

Finally, concerning the geometric structure of the set of pairs $(\lambda, \boldsymbol{u}) \in (0,\infty)\times X(\Omega)$ satisfying (51) for a fixed $\boldsymbol{F}$, a result entirely similar to Theorem 2 continues to hold; see Theorem 6.2 in [36].

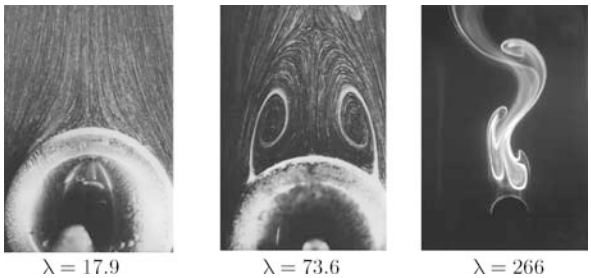

Navier–Stokes Equations: A Mathematical Analysis, Figure 5
Visualization of a flow past a sphere at increasing Reynolds numbers λ; after [91]

Three-Dimensional Flow. Uniqueness and Steady Bifurcation There is both experimental [91,97] and numerical [68,95] evidence that a steady flow past a sphere is unique (and stable) if the Reynolds number λ is sufficiently small. Moreover, experiments report that a closed recirculation zone first appears at λ around 20–25, and the flow stays steady and axisymmetric up to at least $\lambda \simeq 130$. This implies that the first bifurcation occurs through a steady motion. For higher values of λ, the wake behind the sphere becomes time-periodic, thus suggesting the occurrence of unsteady (Hopf) bifurcation; see Fig. 5.

In this section we shall collect the relevant results available for uniqueness and steady bifurcation of a flow past a three-dimensional obstacle.

Uniqueness Results Unlike the case of a flow in a bounded domain where uniqueness in the class of weak solutions is simply established (see Theorem 3), in the situation of a flow past an obstacle, the uniqueness proof requires the detailed study of the *asymptotic behavior* of a weak solution mentioned in Sect. "Flow in Exterior Domains"; see also Theorem 6. Precisely, we have the following result, for whose proof we refer to Theorem IX.5.3 in [32].

Theorem 8 *Suppose $\boldsymbol{f} \in L^{6/5}(\Omega) \cap L^{3/2}(\Omega)$, $\lambda \in (0, \bar{\lambda}]$, for some $\bar{\lambda} > 0$, and let $\boldsymbol{w}$ be the corresponding weak solution.(Notice that, under the given assumptions, $\boldsymbol{f} \in \mathcal{D}_0^{-1,2}(\Omega)$.) There exists $C = C(\Omega, \bar{\lambda}) > 0$ such that, if*

$$\|\boldsymbol{f}\|_{6/5} + \lambda < C\,,$$

then $\boldsymbol{w}$ is the only weak solution corresponding to $\boldsymbol{f}$.

Some Bifurcation Results The rigorous study of steady bifurcation of a flow past a body is a very challenging mathematical problem. However, the function-analytic framework developed in the previous section allows

us to formulate sufficient conditions for steady bifurcation, that are *formally* analogous to those discussed in Sect. "Uniqueness and Steady Bifurcation" for the case of a flow in a bounded domain; see § VII in [36]. To this end, fix $\boldsymbol{f} \in \mathcal{D}_0^{-1,2}(\Omega)$, once and for all, and let $\boldsymbol{u}_0 = \boldsymbol{u}_0(\lambda)$, λ in some open interval $I \subseteq (0, \infty)$, be a *given* curve in $X(\Omega)$, constituted by solutions to (51) corresponding to the prescribed $\boldsymbol{f}$. If $\boldsymbol{u} + \boldsymbol{u}_0$, is another solution, from (51) we easily obtain that $\boldsymbol{u}$ satisfies the following equation in $\mathcal{D}_0^{1,2}(\Omega)$

$$\boldsymbol{u} + \lambda\left(\boldsymbol{L}(\boldsymbol{u}) + \boldsymbol{\mathcal{B}}(\boldsymbol{u}_0(\lambda))(\boldsymbol{u}) + \boldsymbol{\mathcal{M}}(\boldsymbol{u})\right) = \boldsymbol{0}\,, \quad \boldsymbol{u} \in X(\Omega)\,, \tag{52}$$

where $\boldsymbol{\mathcal{B}}(\boldsymbol{u}_0) := D_{\boldsymbol{u}}\boldsymbol{\mathcal{M}}(\boldsymbol{u}_0)$. In this setting, the branch $\boldsymbol{u}_0(\lambda)$ becomes the solution $\boldsymbol{u} \equiv \boldsymbol{0}$ and the bifurcation problem thus reduces to find a nonzero branch of solutions $\boldsymbol{u} = \boldsymbol{u}(\lambda)$ to (52) in every neighborhood of some *bifurcation point* $(\boldsymbol{0}, \lambda_0)$; see Definition 4. Define the map

$$\begin{aligned} F : (\lambda, \boldsymbol{u}) &\in (0, \infty) \times X(\Omega) \\ &\mapsto \boldsymbol{u} + \lambda\left(\boldsymbol{L}(\boldsymbol{u}) + \boldsymbol{\mathcal{B}}(\boldsymbol{u}_0(\lambda))(\boldsymbol{u}) + \boldsymbol{\mathcal{M}}(\boldsymbol{u})\right) \in \mathcal{D}_0^{1,2}(\Omega)\,. \end{aligned} \tag{53}$$

In § VII in [36] the following result is shown.

Lemma 6 *The map F is of class C^∞. Moreover, the derivative*

$$D_{\boldsymbol{u}}F(\lambda, \boldsymbol{0})(\boldsymbol{w}) = \boldsymbol{w} + \lambda\left(\boldsymbol{L}(\boldsymbol{w}) + \boldsymbol{\mathcal{B}}(\boldsymbol{u}_0(\lambda))(\boldsymbol{w})\right)\,,$$

is Fredholm of index 0.

Therefore, from this lemma and from Lemma 4 we obtain that a *necessary condition* for $(\boldsymbol{0}, \lambda_0)$ to be a bifurcation point to (52) is that the linear problem

$$\boldsymbol{w}_1 + \lambda_0\left(\boldsymbol{L}(\boldsymbol{w}_1) + \boldsymbol{\mathcal{B}}(\boldsymbol{u}_0(\lambda_0))(\boldsymbol{w}_1)\right) = \boldsymbol{0}\,, \quad \boldsymbol{w}_1 \in X(\Omega)\,, \tag{54}$$

has a non-zero solution $\boldsymbol{w}_1$. Once this necessary condition is satisfied, one can formulate several sufficient conditions for the point $(\boldsymbol{0}, \lambda_0)$ to be a bifurcation point. For a review of different criteria for global and local bifurcation for Fredholm maps of index 0, we refer to Sect. 6 in [33]. Here we wish to use the criterion of Theorem 5 to present a very simple (in principle) and familiar sufficient condition in the particular case when the given curve $\boldsymbol{u}_0$ can be made (locally, in a neighborhood of λ_0) independent of λ. This may depend on the particular non-dimensionalization of the Navier–Stokes equations and on the special form of the family of solutions $\boldsymbol{u}_0$. In fact, there are several interesting problems formulated in exterior domains where this circumstance takes place, like, for example, the problem of steady bifurcation considered in the previous section and the one studied in Sect. 6 in [39]. Now, if $\boldsymbol{u}_0$ does not depend on λ, from Theorem 5 and from (54) we immediately find that a sufficient condition in order that $(\boldsymbol{0}, \lambda_0)$ be a bifurcation point is that the following problem

$$\boldsymbol{w} + \lambda_0\boldsymbol{\mathcal{L}}(\boldsymbol{w}) = \boldsymbol{w}_1\,, \quad \boldsymbol{w} \in X(\Omega)\,, \tag{55}$$

with $\boldsymbol{\mathcal{L}} := \boldsymbol{L} + \boldsymbol{\mathcal{B}}(\boldsymbol{u}_0)$ and $\boldsymbol{w}_1$ solving (54), has no solution. In different words, *a sufficient condition for* $(\boldsymbol{0}, \lambda_0)$ *to be a bifurcation point is that* $-1/\lambda_0$ *is a simple eigenvalue of the operator* $\boldsymbol{\mathcal{L}}$. It is interesting to observe that this condition is *formally* the same as the one arising in steady bifurcation problems for flow in a *bounded* domain; see Sect."Uniqueness and Steady Bifurcation". However, while in this latter case $\boldsymbol{\mathcal{L}}$ $(\equiv \boldsymbol{B}(\boldsymbol{v}_0))$ is *compact* and defined on the whole of $\mathcal{D}_0^{1,2}(\Omega)$, in the present situation $\boldsymbol{\mathcal{L}}$, with domain $D := X(\Omega) \subset \mathcal{D}_0^{1,2}(\Omega)$, is an *unbounded* operator. As such, we can not even be sure that $\boldsymbol{\mathcal{L}}$ has real *simple* eigenvalues. Nevertheless, since, by Lemma 6, $\boldsymbol{I} + \lambda\boldsymbol{\mathcal{L}}$ is Fredholm of index 0 for all $\lambda \in (0, \infty)$, and since one can show (Lemma 7.1 in [36]) that $\boldsymbol{\mathcal{L}}$ is graph-closed, if $\boldsymbol{u}_0 \in L^3(\Omega)$ (this latter condition is satisfied under suitable hypotheses on $\boldsymbol{f}$; see Theorem 6) from well-known results of spectral theory (see Theorem XVII.2.1 in [46]), it follows that the set Λ of real eigenvalues of $\mathcal{L}$ satisfies the following properties: (a) Λ is at most countable; (b) Λ is constituted by isolated points of finite algebraic and geometric multiplicities, and (c) points in Λ can only cluster at 0. Consequently, the bifurcation condition requiring the simplicity of $-1/\lambda_0$ (namely, algebraic multiplicity 1) is perfectly meaningful.

Two-Dimensional Flow. The Problem of Existence The planar motion of a viscous liquid past a cylinder is among the oldest problems to have received a systematic mathematical treatment. Actually, in 1851, it was addressed by Sir George Stokes in his work on the motion of a pendulum in a viscous liquid [89]. In the wake of his successful study of the flow past a sphere in the limit of vanishing λ (*Stokes approximation*), Stokes looked for solutions to (44), with $\boldsymbol{f} \equiv \boldsymbol{0}$ and $\lambda = 0$, in the case when Ω is the exterior of a circle. However, to his surprise, he found that this linearized problem has *no* solution, and he concluded with the following (wrong) statement [89], p. 63,

> "It appears that the supposition of steady motion is inadmissible".

Such an observation constitutes what we currently call *Stokes Paradox*.

This is definitely a very intriguing starting point for the resolution of the boundary-value problem (44), in that it suggests that, if the problem has a solution, the nonlinear terms have to play a major role. In this regard, by using, for instance, the Galerkin method and proceeding exactly as in Subsection IV.2.1(A), we can prove the existence of a weak solution to (44), *for any* $\lambda > 0$ *and* $\boldsymbol{f} \in \mathcal{D}_0^{-1,2}(\Omega)$. In addition, this solution is as smooth as allowed by the regularity of Ω and $\boldsymbol{f}$ (see Sect. "Flow in Exterior Domains") and, in such a case, it satisfies $(44)_{1,2,3}$ in the ordinary sense. However, unlike the three-dimensional case [see Eq. (46)], the space $\mathcal{D}_0^{1,2}(\Omega)$ is *not embedded in any* L^q*-space* and, therefore, we can not be sure that, even in an appropriate generalized sense, this solution vanishes at infinity, as requested by $(44)_4$. Actually, if $\Omega \subset \mathbb{R}^2$ there are functions in $\mathcal{D}_0^{1,2}(\Omega)$ becoming *unbounded* at infinity. Take, for example, $\boldsymbol{w} = (\ln|\boldsymbol{x}|)^\alpha\, \boldsymbol{e}_\theta$, $\alpha \in (0,1)$, and Ω the exterior of the unit circle. In this sense, we call these solutions *weak*, and not because of lack of local regularity (they are as smooth as allowed by the smoothness of Ω and $\boldsymbol{f}$). This problem was first pointed out by Leray (pp. 54–55 in [61]).

The above partial results leave open the worrisome possibility that a Stokes paradox could also hold for the *fully nonlinear* problem (45). If this chance turned out to be indeed true, it would cast serious doubts on the Navier–Stokes equations as a reliable fluid model, in that they would not be able to catch the physics of a very elementary phenomenon, easily reproduced experimentally.

The possibility of a nonlinear Stokes paradox was ruled out by Finn and Smith in a deep paper published in 1967 [20], where it is shown that if $\boldsymbol{f} \equiv \boldsymbol{0}$ and if Ω is sufficiently regular, then (45) has a solution, at least for "small" (but nonzero!) λ. The method used by these authors is based on the representation of solutions and careful estimates of the Green tensor of the Oseen problem. Another approach to existence for small λ, relying upon the L^q theory of the Oseen problem, was successively given by Galdi [30] where one can find the proof of the following result.

Theorem 9 *Let* Ω *be of class* C^2 *and let* $\boldsymbol{f} \in L^q(\Omega)$, *for some* $q \in (1, 6/5)$. *Then there exists* $\lambda_1 > 0$ *and* $C = C(\Omega, q, \lambda_1) > 0$ *such that if, for some* $\lambda \in (0, \lambda_1]$,

$$|\log\lambda|^{-1} + \lambda^{2(1/q-1)}\|\boldsymbol{f}\|_q < C\,,$$

problem (45) has at least one weak solution that, in addition, satisfies $(45)_4$ *uniformly pointwise.*

It can be further shown that the above solutions meet all the basic physical requirements. In particular, as in the three-dimensional case (see Sect. "Flow in Exterior Domains"), they exhibit a "wake" in the region $x_1 < 0$. Moreover, the solutions are unique in a ball of a suitable Banach space, centered at the origin and of "small" radius. For the proof of these two statements, we refer to §§ X.4 and X.5 in [32].

Though significant, these results leave open several fundamental questions. The most important is, of course, that of whether problem (44) is solvable for *all* $\lambda > 0$ and *all* $\boldsymbol{f}$ in a suitable space. As we already noticed, this solvability would be secured if we can show that the weak solution *does* satisfy $(44)_4$, even in a generalized sense. It should be emphasized that, since, as shown previously, there are functions in $\mathcal{D}_0^{1,2}(\Omega)$ that become unbounded at infinity, the proof of this asymptotic property must be restricted to functions satisfying $(44)_{1,2,3}$. This question has been taken up in a series of remarkable papers by Gilbarg and Weinberger [44], [45] and by Amick [2] in the case when $\boldsymbol{f} \equiv \boldsymbol{0}$. Some of their results lead to the following one due to Galdi; see Theorem 3.4 in [35].

Theorem 10 *Let* $\boldsymbol{w}$ *be a weak solution to (44) with* $\boldsymbol{f} \equiv \boldsymbol{0}$. *Then, there exists* $\boldsymbol{\xi} \in \mathbb{R}^2$ *such that*

$$lim_{|\boldsymbol{x}|\to\infty}\boldsymbol{w}(\boldsymbol{x}) = \boldsymbol{\xi} \quad \textit{uniformly}\,.$$

Open Question It is not known if $\boldsymbol{\xi} = \boldsymbol{0}$, and so it is not known if $\boldsymbol{w}$ satisfies $(44)_4$. Thus, *the question of the solvability of* (44) *for arbitrary* $\lambda > 0$, *even when* $\boldsymbol{f} \equiv \boldsymbol{0}$, *remains open.*

When Ω is symmetric around the direction of $\boldsymbol{e}_1$, in [33] Galdi has suggested a different approach to the solvability of (44) (with $\boldsymbol{f} \equiv \boldsymbol{0}$) for arbitrary large $\lambda > 0$. Let C be the class of vector fields, $\boldsymbol{v} = (v_1, v_2)$, and scalar fields τ such that (i) v_1 and τ are even in x_2 and v_2 is odd in x_2, and (ii) $|\boldsymbol{v}|_{1,2} < \infty$. The following result holds.

Theorem 11 *Let* Ω *be symmetric around the* x_1*-axis. Assume that the homogeneous problem:*

$$\left.\begin{aligned} \Delta\boldsymbol{u} &= \boldsymbol{u}\cdot\nabla\boldsymbol{u} + \nabla\phi \\ \operatorname{div}\boldsymbol{u} &= 0 \end{aligned}\right\} \quad \textit{in } \Omega \qquad (56)$$
$$\boldsymbol{u}|_{\partial\Omega} = \boldsymbol{0}\,, \qquad \lim_{|\boldsymbol{x}|\to\infty}\boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{0} \quad \textit{uniformly}\,,$$

has only the zero solution, $\boldsymbol{u} \equiv \boldsymbol{0}$, $p = \text{const.}$, *in the class* C. *Then, there is a set* M *with the following properties:*

(i) $M \subset (0, \infty)$;
(ii) $M \supset (0, c)$ *for some* $c = c(\Omega) > 0$;
(iii) M *is unbounded;*
(iv) For any $\lambda \in M$, *problem (44) has at least one solution in the class* C.

Open Question The difficulty with the above theorem relies in establishing the validity of its hypothesis, namely, *whether or not* (56) *has only the zero solution in the class* C. Moreover, supposing that the hypothesis holds true, the other fundamental problem is the study of the properties of the set M. For a detailed discussion of these issues, we refer to § 4.3 in [35].

Mathematical Analysis of the Initial-Boundary Value Problem

Objective of this section is to present the main results and open questions regarding the unique solvability of the initial-boundary value problem (7)–(8), and significant related properties.

Preliminary Considerations

In order to present the basic problems for (7)–(8) and the related results, we shall, for the sake of simplicity, restrict ourselves to the case when $\boldsymbol{f} \equiv \boldsymbol{v}_1 \equiv \boldsymbol{0}$. *In what follows, we shall denote by* (7)–(8)$_{\text{hom}}$ *this homogeneous problem*. The first, fundamental problem that should be naturally set for (7)–(8)$_{\text{hom}}$ is the classical one of (global) well-posedness in the sense of Hadamard.

Problem 1 Find a Banach space, X, such that for any initial data $\boldsymbol{v}_0$ in X there is a corresponding solution $(\boldsymbol{v}, p)$ to (7)-(8)$_{\text{hom}}$ satisfying the following conditions: (i) it exists for all $T > 0$, (ii) it is unique and (iii) it depends continuously on $\boldsymbol{v}_0$.

In different words, the resolution of Problem 1 will ensure that the Navier–Stokes equations furnish, at all times, a *deterministic description* of the dynamics of the liquid, provided the initial data are given in a "sufficiently rich" class. It is immediately seen that the class X should meet some necessary requirements for Problem 1 to be solvable. For instance, if we take $\boldsymbol{v}_0$ *only bounded*, we find that problem (7)–(8), with $\Omega = \mathbb{R}^n$ and $\boldsymbol{f} = 0$, admits the following two distinct solutions

$$\boldsymbol{v}_1(\boldsymbol{x}, t) = \boldsymbol{0}\,, \qquad p_1(\boldsymbol{x}, t) = 0\,;$$
$$\boldsymbol{v}_2(\boldsymbol{x}, t) = \sin t\boldsymbol{e}_1\,, \quad p_2(\boldsymbol{x}, t) = -x_1 \cos t\,;$$

corresponding to the same initial data $\boldsymbol{v}_0 = \boldsymbol{0}$.

Furthermore, we observe that the resolution of Problem 1, does *not* exclude the possibility of the formation of a "singularity", that is, the existence of points in the space-time region where the solution may become unboundedly large in certain norms. This possibility depends, of course, on the regularity of the functional class where well-posedness is established.

One is thus led to considering the next fundamental (and most popular) problem.

Problem 2 Given an initial distribution of velocity $\boldsymbol{v}_0$, no matter how smooth, with

$$\int_\Omega |\boldsymbol{v}_0(\boldsymbol{x})|^2 < \infty\,, \tag{57}$$

determine a corresponding regular solution $\boldsymbol{v}(x, t)$, $p(x, t)$ to (7)–(8)$_{\text{hom}}$ for all times $t \in (0, T)$ and all $T > 0$.

By "regular" here we mean that $\boldsymbol{v}$ and p are both of class C^∞ in the *open* cylinder $\Omega \times (0, T)$, for all $T > 0$. When $\Omega \equiv \mathbb{R}^3$, Problem 1 is, basically, the third Millennium Prize Problem posted by the Clay Mathematical Institute in May 2000.

The requirement (57) on the initial data is meaningful from the physical point of view, in that it ensures that the kinetic energy of the liquid is initially finite. Moreover, it is also necessary from a mathematical viewpoint because, if we relax (57) to the requirement that the initial distribution of velocity is, for example, *only bounded*, then Problem 2 has a simple negative answer. In fact, the following pair

$$\boldsymbol{v}(\boldsymbol{x}, t) = \frac{1}{\tau - t}\,\boldsymbol{e}_1 \quad p(x, t) = -\frac{x_1}{(\tau - t)^2}\,,$$
$$t \in [0, \tau)\,, \quad \tau > 0$$

is a solution to (7)–(8)$_{\text{hom}}$ with $\boldsymbol{f} \equiv \boldsymbol{0}$ and $\Omega = \mathbb{R}^3$, that becomes singular at $t = \tau$, for any given positive τ.

An alternative way of formulating Problem 2 in "more physical" terms is as follows.

Problem 2′ Can a spontaneous singularity arise in a finite time in a viscous liquid that is initially in an arbitrarily smooth state?

Though, perhaps, the gut answer to this question could be in the negative, one can bring very simple examples of dissipative nonlinear evolution equations where spontaneous singularities do occur, if the initial data are sufficiently large. For instance, the initial-value problem $v' + \sigma\, v = v^2$, $v(0) = v_0$, $\sigma > 0$, has the explicit solution

$$v(t) = \frac{\sigma\, v_0}{v_0 - e^{\sigma\, t}(v_0 - \sigma)}$$

which shows that, if $v_0 \le \sigma$, then v is smooth for all $t > 0$, while if $v_0 > \sigma$, then v becomes unbounded in a finite time:

$$v(t) \ge \frac{1}{\tau - t}\,, \quad t \in [0, \tau)\,, \quad \tau := \frac{1}{\sigma} \log\left(\frac{v_0}{v_0 - \sigma}\right).$$

If the occurrence of singularities for problem (7)–$(8)_{\text{hom}}$ can not be at all excluded, one can still theorize that singularities are unstable and, therefore, "undetectable". Another plausible explanation could be that singularity may appear in the Navier–Stokes equations due to the possible break-down of the continuum model at very small scales.

It turns out that, in the case of two-dimensional (2D) flow, both problems 1 and 2 are completely resolved, while they are both open in the three-dimensional (3D) case. The following section will be dedicated to these issues.

On the Solvability of Problems 1 and 2

As in the steady-state case, a basic tool for the resolution of both Problems 1 and 2 is the accomplishment of "good" a priori estimates. By "good", we mean that (i) they have to be *global*, namely, they should hold for all positive times, and (ii) they have to be valid in a sufficiently regular function class. These estimates can then be used along suitable "approximate solutions" which eventually will converge, by an appropriate limit procedure, to a solution to (7)–$(8)_{\text{hom}}$. To date, "good" estimates for 3D flow are not known.

Unless explicitly stated, throughout this section we assume that Ω is a bounded, smooth (of class C^2, for example) domain of $\mathbb{R}^n$, $n = 2, 3$.

Derivation of Some Fundamental A Priori Estimates

We recall that, for simplicity, we are assuming that the boundary data, $\boldsymbol{v}_1$, in (8) is vanishing. Thus, if we formally dot-multiply through both sides of $(7)_1$ (with $\boldsymbol{f} \equiv \boldsymbol{0}$) by $\boldsymbol{v}$, integrate by parts over Ω and take into account 12 and Lemma 1, we obtain the following equation

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{v}(t)\|_2^2 + \nu\,|\boldsymbol{v}(t)|_{1,2}^2 = 0\,. \tag{58}$$

The physical interpretation of (58) is straightforward. Actually, if we dot-multiply both sides of the identity $\operatorname{div}\boldsymbol{D}(\boldsymbol{v}) = \Delta\boldsymbol{v}$ by $\boldsymbol{v}$, where $\boldsymbol{D} = \boldsymbol{D}(\boldsymbol{v})$ is the stretching tensor (see 3), and integrate by parts over Ω, we find that $|\boldsymbol{v}|_{1,2} = \|\boldsymbol{D}(\boldsymbol{v})\|_2$. Since, as we observed in Sect. "Derivation of the Navier–Stokes Equations and Preliminary Considerations", $\boldsymbol{D}$ takes into account the deformation of the parts of the liquid, Eq. (58) simply relates the rate of decreasing of the kinetic energy to the dissipation inside the liquid, due to the combined effect of viscosity and deformation. If we integrate (58) from $s \geq 0$ to $t \geq s$, we obtain the so-called *energy equality*

$$\|\boldsymbol{v}(t)\|_2^2 + 2\nu\int_s^t |\boldsymbol{v}(\rho)|_{1,2}^2\,\mathrm{d}\rho = \|\boldsymbol{v}(s)\|_2^2 \quad 0 \leq s \leq t\,. \tag{59}$$

Notice that the nonlinear term $\boldsymbol{v}\cdot\nabla\boldsymbol{v}$ does *not* give any contribution to Eq. (58) [and, consequently, to Eq. (59)], in virtue of the fact that $(\boldsymbol{v}\cdot\nabla\boldsymbol{v}, \boldsymbol{v}) = 0$; see Lemma 1. By taking $s = 0$ in (59), we find a *bound on the kinetic energy and on the total dissipation for all times $t \geq 0$, in terms of the initial data only*, provided the latter satisfy (57). In concise words, the energy equality (59) is a *global a priori* estimate. It should be emphasized that the energy equality is, basically, *the only known global a priori estimate for 3D flow*.

From (59) it follows, in particular,

$$\boldsymbol{v} \in L^\infty(0, T; L^2(\Omega)) \cap L^2(0, T; \mathcal{D}_0^{1,2}(\Omega))\,, \quad \text{all } T > 0\,. \tag{60}$$

A second estimate can be obtained by dot-multiplying through both sides of $(7)_1$ by $P\Delta\boldsymbol{v}$ and by integrating by parts over Ω. Taking into account that

$$\left(\frac{\partial\boldsymbol{v}}{\partial t}, P\Delta\boldsymbol{v}\right) = \left(\frac{\partial\boldsymbol{v}}{\partial t}, \Delta\boldsymbol{v}\right) = -\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|\boldsymbol{v}|_{1,2}^2\,,$$
$$(\nabla p,\ P\Delta\boldsymbol{v}) = 0\,,$$

we deduce

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|\boldsymbol{v}|_{1,2}^2 + \nu\|P\Delta\boldsymbol{v}\|_2^2 = (\boldsymbol{v}\cdot\nabla\boldsymbol{v}, P\Delta\boldsymbol{v})\,. \tag{61}$$

Since the right-hand side of this equation need not be zero, we get that, unlike (58), the nonlinear term *does* contribute to (61). In addition, since the sign of this contribution is basically unknown, in order to obtain useful estimates we have to increase it appropriately. To this end, we recall the validity of the following inequalities

$$\|\boldsymbol{u}\|_\infty \leq \begin{cases} c_1\,\|\boldsymbol{u}\|_2^{\frac{1}{2}}\,\|P\Delta\boldsymbol{u}\|_2^{\frac{1}{2}} \\ \text{if } n = 2\,, \quad \text{for all } \boldsymbol{u} \in L^2_\sigma(\Omega)\,, \\ \text{with } P\Delta\boldsymbol{u} \in L^2(\Omega) \quad \text{and } \boldsymbol{u}|_{\partial\Omega} = \boldsymbol{0}\,, \\ c_2\,|\boldsymbol{u}|_{1,2}^{\frac{1}{2}}\|P\Delta\boldsymbol{u}\|_2^{\frac{1}{2}} \\ \text{if } n = 3\,, \quad \text{for all } \boldsymbol{u} \in \mathcal{D}_0^{1,2}(\Omega)\,, \\ \text{with } P\Delta\boldsymbol{u} \in L^2(\Omega)\,, \end{cases} \tag{62}$$

where $c_i = c_i(\Omega) > 0$, $i = 1, 2$. These relations follow from the Sobolev embedding theorems along with Lemma 2. We shall sketch a proof in the case $n = 3$. By the property of the projection operator P, $\boldsymbol{u}$ satisfies the assumptions of Lemma 2 with $\boldsymbol{g} := P\Delta\boldsymbol{u}$, and, consequently, we have, on the one hand, that $\boldsymbol{u} \in W^{2,2}(\Omega)$, and, on the other hand,

$$\|\boldsymbol{u}\|_{2,2} \leq c\,\|P\Delta\boldsymbol{u}\|_2\,, \tag{63}$$

with $c = c(\Omega) > 0$. We now recall that there exists an extension operator $E : \boldsymbol{u} \in W^{2,2}(\Omega) \mapsto E(\boldsymbol{u}) \in W^{2,2}(\mathbb{R}^3)$ such that

$$\|E(\boldsymbol{u})\|_{k,2} \le C_k \|\boldsymbol{u}\|_{k,2}\,, \quad k = 0, 1, 2\,; \tag{64}$$

see Chap. VI, Theorem 5 in [88]. Next, take $\boldsymbol{\varphi} \in C_0^\infty(\mathbb{R}^3)$. From the identity $\Delta(|\boldsymbol{\varphi}|^2) = 2(\boldsymbol{\varphi}\cdot\Delta\boldsymbol{\varphi} + |\nabla\boldsymbol{\varphi}|^2)$ we have the representation

$$|\boldsymbol{\varphi}(\boldsymbol{x})|^2 = -\frac{1}{2\pi}\int_{\mathbb{R}^3} \frac{\boldsymbol{\varphi}(\boldsymbol{y})\cdot\Delta\boldsymbol{\varphi}(\boldsymbol{y}) + |\nabla\boldsymbol{\varphi}(\boldsymbol{y})|^2}{|\boldsymbol{x}-\boldsymbol{y}|}\,\mathrm{d}\boldsymbol{y}\,. \tag{65}$$

Using Schwarz inequality on the right-hand side of (65) along with the classical Hardy inequality § II.5 in [31]:

$$\int_{\mathbb{R}^3} \frac{|\boldsymbol{\varphi}(\boldsymbol{y})|^2}{|\boldsymbol{x}-\boldsymbol{y}|^2}\,d\boldsymbol{y} \le 4|\boldsymbol{\varphi}|_{1,2}^2\,,$$

we recover $\|\boldsymbol{\varphi}\|_\infty \le (2/\pi)^{1/2}|\boldsymbol{\varphi}|_{1,2}^{\frac{1}{2}}|\boldsymbol{\varphi}|_{2,2}^{\frac{1}{2}}$. Since $C_0^\infty(\mathbb{R}^3)$ is dense in $W^{2,2}(\mathbb{R}^3)$, from this latter inequality we deduce, in particular,

$$\|\boldsymbol{u}\|_\infty \le (2/\pi)^{1/2}|E(\boldsymbol{u})|_{1,2}^{\frac{1}{2}}|E(\boldsymbol{u})|_{2,2}^{\frac{1}{2}}\,,$$

which, in turn, by (64) and Poincaré's inequality (18) implies that

$$\|\boldsymbol{u}\|_\infty \le c_5|\boldsymbol{u}|_{1,2}^{\frac{1}{2}}\|\boldsymbol{u}\|_{2,2}^{\frac{1}{2}}\,,$$

with $c_5 = c_5(\Omega) > 0$. Equation $(62)_2$ then follows from this latter inequality and from (63). (For the proof of (62) in more general domains, as well as in domains with less regularity, we refer to [9,66,98]. I am not aware of the validity of (62) in an arbitrary (smooth) domain.) We now employ (62) and 29 into (61) to obtain

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|\boldsymbol{v}|_{1,2}^2 + \frac{\nu}{2}\|P\Delta\boldsymbol{v}\|_2^2 \le \begin{cases} c_3\|\boldsymbol{v}\|_2^2|\boldsymbol{v}|_{1,2}^4 & \text{if } n = 2\,, \\ c_4|\boldsymbol{v}|_{1,2}^6 & \text{if } n = 3\,, \end{cases} \tag{66}$$

where $c_i = c_i(\Omega, \nu) > 0$, $i = 3, 4$. Thus, observing that, from (59), $\|\boldsymbol{v}(t)\|_2 \le \|\boldsymbol{v}_0\|_2$, if we assume, further, that $\boldsymbol{v}_0 \in \mathcal{D}_0^{1,2}(\Omega)$, from the previous differential inequality, we obtain the following uniform bound

$$|\boldsymbol{v}(t)|_{1,2} \le M(\Omega, \nu, t, \|\boldsymbol{v}_0\|_{1,2})\,,$$
$$\text{for all } t \in [0, \tau)\,, \quad \text{and some } \tau \ge 1/(K|\boldsymbol{v}_0|_{1,2}^\alpha)\,, \tag{67}$$

where M is a continuous function in t and $K = 8c_3, \alpha = 2$ if $n = 2$, while $K = 4c_4$, $\alpha = 4$ if $n = 3$. Equation (67) provides the second a priori estimate. Notice that, unlike (59), we do not know if in (67) we can take t arbitrary large, namely, we do not know if $\tau = \infty$. Integrating both sides of (66) from 0 to $t < \tau$, and taking into account (67) we find that

$$\int_0^t \|P\Delta\boldsymbol{v}(s)\|_2^2\mathrm{d}s \le M_1(\Omega, \nu, t, \|\boldsymbol{v}_0\|_{1,2})\,,$$
$$\text{for all } t \in [0, \tau)\,, \tag{68}$$

with M_1 continuous in t. From (68), (67), (60), and 63 one can then show that

$$\boldsymbol{v} \in L^\infty(0, t; W_0^{1,2}(\Omega)) \cap L^2(0, t; W^{2,2}(\Omega))\,,$$
$$t \in [0, \tau)\,. \tag{69}$$

A third a priori estimate, on the time derivative of $\boldsymbol{v}$, can be formally obtained by dot-multiplying both sides of $(7)_1$ and by integrating by parts over Ω. By using arguments similar to those leading to (61) we find

$$\frac{\nu}{2}\frac{\mathrm{d}}{\mathrm{d}t}|\boldsymbol{v}|_{1,2}^2 + \left\|\frac{\partial\boldsymbol{v}}{\partial t}\right\|_2^2 = -\left(\boldsymbol{v}\cdot\nabla\boldsymbol{v}, \frac{\partial\boldsymbol{v}}{\partial t}\right)\,, \tag{70}$$

and so, employing Hölder inequality on the right-hand side of (62) along with 29, (67) and (68) we show the following estimate

$$\int_0^t \left\|\frac{\partial\boldsymbol{v}}{\partial s}\right\|_2^2 \mathrm{d}s \le M_2(\Omega, \nu, t, \|\boldsymbol{v}_0\|_{1,2})\,,$$
$$\text{for all } t \in [0, \tau)\,, \tag{71}$$

with M_2 continuous in t. This latter inequality implies that

$$\frac{\partial\boldsymbol{v}}{\partial t} \in L^2(0, t; L^2(\Omega))\,, \quad t \in [0, \tau)\,. \tag{72}$$

Existence, Uniqueness, Continuous Dependence, and Regularity Results We shall now use estimates (59), (67), (68), (71) along a suitable approximate solution constructed by the finite-dimensional (Galerkin) method to show existence to (7)–$(8)_{\text{hom}}$ in an appropriate function class. We shall briefly sketch the argument. Similarly to the steady-state case, we look for an approximate solution to (7)–$(8)_{\text{hom}}$ of the form $\boldsymbol{v}_N(\boldsymbol{x}, t) = \sum_{i=0}^N c_{iN}(t)\boldsymbol{\psi}_i$, where $\{\boldsymbol{\psi}_i\}$ is a base of $L_\sigma^2(\Omega)$ constituted by the eigenvectors of the operator $-P\Delta$, namely,

$$-\nu\Delta\boldsymbol{\psi}_i = \lambda_i\boldsymbol{\psi}_i + \nabla\Phi_i\,, \quad \mathrm{div}\boldsymbol{\psi}_i = 0 \quad \text{in } \Omega\,,$$
$$\boldsymbol{\psi}_i|_{\partial\Omega} = \boldsymbol{0}\,, \quad i \in \mathbb{N}\,, \tag{73}$$

where λ_i are the corresponding eigenvalues. The coefficients $c_{iN}(t)$ are requested to satisfy the following system

of ordinary differential equations

$$\left(\frac{\partial \boldsymbol{v}_N}{\partial t}, \boldsymbol{\psi}_k\right) + (\boldsymbol{v}_N \cdot \nabla \boldsymbol{v}_N, \boldsymbol{\psi}_k) = \nu(\Delta \boldsymbol{v}_N, \boldsymbol{\psi}_k), \quad k = 1, \ldots, N, \tag{74}$$

with initial conditions $c_{iN}(0) = (\boldsymbol{v}_0, \boldsymbol{\psi}_i)$, $i = 1, \ldots, N$. Multiplying both sides of (74), in the order, by c_{kN}, by $\lambda_k c_{kN}$ and by dc_{kN}/dt and summing over k from 1 to N, we at once obtain, with the help of (73) and of Lemma 1, that $\boldsymbol{v}_N$ satisfies (59), (66), and (70). Consequently, $\boldsymbol{v}_N$ satisfies the uniform (in N) bounds (58) (evaluated at $s = 0$) (67), (68), and (71). Employing these bounds together with, more or less, standard limiting procedures, we can show the existence of a field $\boldsymbol{v}$ in the classes defined by (69) and (72) satisfying the relation

$$\left(\frac{\partial \boldsymbol{v}}{\partial t} + \boldsymbol{v} \cdot \nabla \boldsymbol{v} - \nu \Delta \boldsymbol{v}, \boldsymbol{\varphi}\right) = 0, \quad \text{for all } \boldsymbol{\varphi} \in \mathcal{D}(\Omega) \quad \text{and a.a. } t \in [0, \tau). \tag{75}$$

Because of (69) and (72), the function involving $\boldsymbol{v}$ in (75) belongs to $L^2(\Omega)$, a.e. in $[0, \tau)$, and therefore, in view of the orthogonal decomposition $L^2(\Omega) = L^2_\sigma(\Omega) \oplus G(\Omega)$ and of the density of $\mathcal{D}(\Omega)$ in $L^2_\sigma(\Omega)$, we find $p \in L^2(0, t; W^{1,2}(\Omega))$, $t \in [0, \tau)$, such that $(\boldsymbol{v}, p)$ satisfies $(7)_1$ for a.a. $(\boldsymbol{x}, t) \in \Omega \times [0, \tau)$. We thus find the following result, basically due to G. Prodi, to whose paper [72] we refer for all missing details; see also Chapter V.4 in [86].

Theorem 12 (Existence) *For every $\boldsymbol{v}_0 \in \mathcal{D}^{1,2}_0(\Omega)$, there exist $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{x}, t)$ and $p = p(\boldsymbol{x}, t)$ such that*

$$\begin{aligned} &\boldsymbol{v} \in L^\infty(0, \tau; L^2(\Omega)) \cap L^2(0, \tau; \mathcal{D}^{1,2}_0(\Omega)), \\ &\left.\begin{aligned} &\boldsymbol{v} \in C([0, t]; \mathcal{D}^{1,2}_0(\Omega)) \\ &\quad \cap L^2(0, t; W^{2,2}(\Omega)), \\ &\frac{\partial \boldsymbol{v}}{\partial t} \in L^2(0, t; L^2(\Omega)), \\ &p \in L^2(0, t; W^{1,2}(\Omega)), \end{aligned}\right\} \text{for all } t \in [0, \tau), \end{aligned} \tag{76}$$

with τ given in (67), satisfying (7) for a.a. $(\boldsymbol{x}, t) \in \Omega \times [0, \tau)$, and $(8)_2$ (with $\boldsymbol{v}_1 \equiv \boldsymbol{0}$) for a.a. $(\boldsymbol{x}, t) \in \partial\Omega \times [0, \tau)$. Moreover, the initial condition $(8)_1$ is attained in the following sense:

$$\lim_{t \to 0+} \|\boldsymbol{v}(t) - \boldsymbol{v}_0\|_{1,2} = 0.$$

We also have.

Theorem 13 (Uniqueness and Continuous Dependence on the Initial Data) *Let $\boldsymbol{v}_0$ be as in Theorem 12. Then the corresponding solution is unique in the class (76). Moreover, it depends continuously on $\boldsymbol{v}_0$ in the norm of $L^2(\Omega)$, in the time interval $[0, \tau)$.*

Proof Let $(\boldsymbol{v}, p)$ and $(\boldsymbol{v} + \boldsymbol{u}, p + p_1)$ be two solutions corresponding to data $\boldsymbol{v}_0$ and $\boldsymbol{v}_0 + \boldsymbol{u}_0$, respectively. From (7)–$(8)_{\text{hom}}$ we then find

$$\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{u} + \boldsymbol{v} \cdot \nabla \boldsymbol{u} + \boldsymbol{u} \cdot \nabla \boldsymbol{v} = -\nabla(p_1/\rho) + \nu \Delta \boldsymbol{u}, \quad \operatorname{div} \boldsymbol{u} = 0, \quad \text{a.e. in } \Omega \times (0, \tau). \tag{77}$$

Employing the properties of the function $\boldsymbol{v}$ and $\boldsymbol{u}$, it is not hard to show the following equation

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{u}\|_2^2 + \nu|\boldsymbol{u}|_{1,2}^2 = -(\boldsymbol{u} \cdot \nabla \boldsymbol{v}, \boldsymbol{u}), \tag{78}$$

that is formally obtained by dot-multiplying both sides of $(77)_1$ by $\boldsymbol{u}$, and by using $(77)_2$ and Lemma 1 along with the fact that $\boldsymbol{u}$ has zero trace at $\partial\Omega$. By Hölder inequality and inequalities (14), (18), and (29), we find

$$\begin{aligned} |(\boldsymbol{u} \cdot \nabla \boldsymbol{v}, \boldsymbol{u})| &\le \|\boldsymbol{u}\|_2 \|\boldsymbol{u}\|_4 |\boldsymbol{v}|_{1,4} \le c_1 \|\boldsymbol{u}\|_2 |\boldsymbol{u}|_{1,2} \|\boldsymbol{v}\|_{2,2} \\ &\le c_2 \|\boldsymbol{v}\|_{2,2}^2 \|\boldsymbol{u}\|_2^2 + \frac{\nu}{2}|\boldsymbol{u}|_{1,2}^2, \end{aligned}$$

where $c_1 = c_1(\Omega) > 0$ and $c_2 = c_2(\Omega, \nu) > 0$. If we replace this inequality back into (78) we deduce

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{u}\|_2^2 \le 2c_2 \|\boldsymbol{v}\|_{2,2}^2 \|\boldsymbol{u}\|_2^2,$$

and so, by Gronwall's lemma and by the fact that $\int_0^t \|\boldsymbol{v}(s)\|_{2,2}^2 \mathrm{d}s < \infty$, $t \in [0, \tau)$ (see Theorem 12), we prove the desired result. □

Finally, we have the following result concerning the regularity of solutions determined in Theorem 12, for whose proof we refer to [38] and Theorem 5.2 in [34].

Theorem 14 (Regularity) *Let Ω be a bounded domain of class C^∞. Then, the solution $(\boldsymbol{v}, p)$ constructed in Theorem 12 is of class $C^\infty(\bar{\Omega} \times (0, \tau))$.*

Remark 7 The results of Theorems 12–14 can be extended to *arbitrary* domains of $\mathbb{R}^n$, provided their boundary is sufficiently smooth; see [34,50]. Moreover, the continuous dependence result in Theorem 13 can be proved in the stronger norm of $\mathcal{D}^{1,2}_0(\Omega)$.

Times of Irregularity and Resolution of Problems 1 and 2 in 2D Theorems 12–14 furnish a complete and positive answer to both Problems 1 and 2, *provided we show that* $\tau = \infty$. Our next task is to give necessary and sufficient

conditions for this latter situation to occur. To this end, we give the following.

Definition 6 Let $(\boldsymbol{v}, p)$ be a solution to (7)–(8)$_{\text{hom}}$ in the class (76). We say that τ is a time of irregularity if and only if (i) $\tau < \infty$, and (ii) $(\boldsymbol{v}, p)$ can not be continued, in the class (76), to an interval $[0, \tau_1)$ with $\tau_1 > \tau$.

If τ is a time of irregularity, we expect that some norms of the solution may become infinite at $t = \tau$, while being bounded for all $t \in [0, \tau)$. In order to show this rigorously, we premise a simple but useful result.

Lemma 7 *Assume that $\boldsymbol{v}$ is a solution to (7)–(8)$_{\text{hom}}$ in the class (76) for some $\tau > 0$. Then, $|\boldsymbol{v}(t)|_{1,2} < \infty$, for all $t \in [0, \tau)$. Furthermore, for all $q \in (n, \infty]$, and all $t \in [0, \tau)$*

$$\int_0^t \|\boldsymbol{v}(s)\|_q^r \, ds < \infty \,, \quad \frac{2}{r} + \frac{n}{q} = 1 \,.$$

Proof The proof of the first statement is obvious. Moreover, by the Sobolev embedding theorem (see Theorem at p. 125 in [70]) we find, for $n = 2$,

$$\|\boldsymbol{v}\|_q \le c_1 \|\boldsymbol{v}\|_2^{2/q} |\boldsymbol{v}|_{1,2}^{1-2/q} \,, \quad q \in (2, \infty) \tag{79}$$

and

$$\|\boldsymbol{v}\|_\infty \le c_2 \|\boldsymbol{v}\|_{2,2} \,,$$

with $c_1 = c_1(q) > 0$, $c_2 = c_2(\Omega) > 0$ while, if $n = 3$,

$$\begin{aligned} &\|\boldsymbol{v}\|_q \le c_3, \|\boldsymbol{v}\|_2^{(6-q)/2q} |\boldsymbol{v}|_{1,2}^{3(q-2)/2q} \,, \quad \text{if } q \in [2, 6] \,, \\ &\|\boldsymbol{v}\|_q^{2q/(q-3)} \le c_4 \|\boldsymbol{v}\|_{2,2}^{(q-6)/(q-3)} |\boldsymbol{v}|_{1,2}^{(6+q)/(q-3)} \,, \\ &\qquad\qquad \text{if } q \in (6, \infty] \,, \end{aligned} \tag{80}$$

where $c_i = c_i(\Omega, q) > 0$, $i = 3, 4$. Since, by (76),

$$\sup_{t \in [0,\tau]} \|\boldsymbol{v}(t)\|_2 + |\boldsymbol{v}(t)|_{1,2} + \int_0^t \|\boldsymbol{v}(s)\|_{2,2}^2 \mathrm{ds} < \infty \,, \qquad t \in [0, \tau) \,,$$

the lemma follows by noting that $(q-6)/(q-3) < 2$ for $q > 3$. □

We shall now furnish some characterization of the possible times of irregularity in terms of the behavior, around them, of the norms of the solution considered in Lemma 7.

Lemma 8 (Criteria for the Existence of a Time of Irregularity) *Let $(\boldsymbol{v}, p)$ be a solution to (7)–(8)$_{\text{hom}}$ in the class (76) for some $\tau \in (0, \infty]$. Then, the following properties hold:*

(i) If τ is a time of irregularity, then

$$\lim_{t \to \tau^-} |\boldsymbol{v}(t)|_{1,2} = \infty \,. \tag{81}$$

Conversely, if $\tau < \infty$ and (81) holds, then τ is a time of irregularity. Moreover, if τ is a time of irregularity, for all $t \in (0, \tau)$ the following growth estimates hold

$$|\boldsymbol{v}(t)|_{1,2}^2 \ge \begin{cases} \dfrac{C}{\tau - t} & \text{if } n = 2 \,, \\ \dfrac{C}{\sqrt{\tau - t}} & \text{if } n = 3 \,, \end{cases} \tag{82}$$

with $C = C(\Omega, \nu) > 0$.

(ii) If τ is a time of irregularity, then, for all $q \in (n, \infty]$,

$$\int_0^\tau \|\boldsymbol{v}(s)\|_q^r \, ds = \infty \,, \quad \frac{2}{r} + \frac{n}{q} = 1 \,. \tag{83}$$

Conversely, if $\tau < \infty$ and (83) holds for some $q = \bar{q} \in (n, \infty]$, then, τ is a time of irregularity.

(iii) If $n = 3$, there exists $K = K(\Omega, \nu) > 0$ such that, if $\|\boldsymbol{v}_0\|_2 |\boldsymbol{v}_0|_{1,2} < K$, then $\tau = \infty$.

Proof

(i) Clearly, if $\tau < \infty$ and (81) holds, then τ is a time of irregularity. Conversely, suppose τ is a time of irregularity and assume, by contradiction, that there exists a sequence $\{t_k\}$ in $[0, \tau)$ and $M > 0$, independent of k, such that

$$t_k \to \tau \,, \quad |\boldsymbol{v}(t_k)|_{1,2} \le M \,.$$

Since $\boldsymbol{v}(t_k) \in \mathcal{D}_0^{1,2}(\Omega)$, by Theorem 12 we may construct a solution $(\bar{\boldsymbol{v}}, \bar{p})$ with initial data $\boldsymbol{v}(t_k)$, in a time interval $[t_k, t_k + \tau^*)$ where see (67)

$$\tau^* \ge A / |\boldsymbol{v}(t_k)|_{1,2}^\alpha \ge A M^\alpha \equiv \tau_* \,, \quad \alpha = 2(n-1) \,,$$

and A depends only on Ω and ν. By Theorem 12, $\bar{\boldsymbol{v}}$ belongs to the class (76) in the time interval $[t_k, t_k + \tau_*]$, with τ_* independent of k and, by Theorem 13, $\bar{\boldsymbol{v}}$ must coincide with $\boldsymbol{v}$ in the time interval $[t_k, \tau)$. We may now choose $t_k \equiv \tau_0$ such that $\tau_0 + \tau_* > \tau$, contradicting the assumption that τ is time of irregularity. We next show (82) when $n = 3$, the proof for $n = 2$ being completely analogous. In-

tegrating (66) (with $n = 3$) we find

$$\frac{1}{|\boldsymbol{v}(t)|_{1,2}^4} - \frac{1}{|\boldsymbol{v}(s)|_{1,2}^4} \le c_4(s-t)\,, \quad 0 < t < s < \tau.$$

Letting $s \to \tau$ and recalling (81), we prove (82).

(ii) Assume that τ is a time of irregularity. Then, $(82)_2$ holds. Now, by the Sobolev embedding theorems, one can show that (see [proof of Lemma 5.4] in [34])

$$(\boldsymbol{v}\cdot\nabla\boldsymbol{v}, P\Delta\boldsymbol{v}) \le C\,\|\boldsymbol{v}\|_q^{2q/(q-n)}|\boldsymbol{v}|_{1,2}^2 + \frac{\nu}{2}\|P\Delta\boldsymbol{v}\|_2^2\,, \quad \text{for all } q \in (n,\infty]\,,$$

where $C = C(\Omega,\nu) > 0$. If we replace this relation into (61) and integrate the resulting differential inequality from 0 to $t < \tau$, we find

$$|\boldsymbol{v}(t)|_{1,2}^2 \le |\boldsymbol{v}_0|_{1,2}^2 \exp\{2C\int_0^t \|\boldsymbol{v}(s)\|_q^r ds\}\,, \quad \text{for all } t \in [0,\tau)\,. \tag{84}$$

If condition (83) is not true for some $q \in (n,\infty]$, then (84) evaluated for that particular q, would contradict (82). Conversely, assume (83) holds for some $q = \bar{q} \in (n,\infty]$, but that, by contradiction, the solution of Theorem 12 can be extended to $[0,\tau_1)$ with $\tau_1 > \tau$. Then, by Lemma 7 we would get the invalidity of condition (83) with $q = \bar{q}$, and the proof of (ii) is completed.

(iii) By integrating the differential inequality $(66)_2$, we find

$$|\boldsymbol{v}(t)|_{1,2}^2 \le \frac{|\boldsymbol{v}_0|_{1,2}^2}{1 - 2c_4|\boldsymbol{v}_0|_{1,2}^2\int_0^t |\boldsymbol{v}(s)|_{1,2}^2 \mathrm{d}s}\,, \quad t \in [0,\tau)\,.$$

Thus, by (59) with $s = 0$ in this latter inequality we find

$$|\boldsymbol{v}(t)|_{1,2}^2 \le \frac{|\boldsymbol{v}_0|_{1,2}^2}{1 - c_5|\boldsymbol{v}_0|_{1,2}^2\|\boldsymbol{v}_0\|_2^2}\,, \quad t \in [0,\tau)\,,$$

with $c_5 = c_5(\Omega,\nu) > 0$, which shows that $(82)_2$ can not occur if the initial data satisfy the imposed "smallness" restriction.

□

A fundamental consequence of Lemma 8 is that, in the case $n = 2$, a time of irregularity can not occur. In fact, for example, (82) is incompatible with the fact that $\boldsymbol{v} \in L^2(0,\tau;\mathcal{D}_0^{1,2}(\Omega))$ (see $(77)_1$). We thus have the following theorem, which answers positively both Problems 1 and 2 in the 2D case.

Theorem 15 (Resolution of Problems 1 and 2 in 2D) *Let $\Omega \subset \mathbb{R}^2$. Then, in Theorems 12–14 we can take $\tau = \infty$.*

The conclusion of Theorem 15 also follows from Lemma 8 (ii). In fact, in the case $n = 2$, by (79) we at once find that

$$L^{2q/(q-2)}(0,T:L^q(\Omega)) \subset L^\infty(0,T;L^2(\Omega)) \cap L^2(0,T;\mathcal{D}_0^{1,2}(\Omega))\,, \text{for all } T > 0 \quad \text{and all } q \in (2,\infty)\,. \tag{85}$$

so that, from $(76)_1$, we deduce

$$\int_0^\tau \|\boldsymbol{v}(t)\|_q^{2q/(2-q)}\mathrm{d}t < \infty\,, \quad \text{for all } q \in (2,\infty)\,.$$

which, by Lemma 8 (ii), excludes the occurrence of time of irregularity.

Unfortunately, from all we know, in the case $n = 3$, we can not draw the same conclusion. Actually, in such a case, $(82)_2$ and $(77)_1$ are no longer incompatible. Moreover, from Lemma 8 (ii) it follows that a sufficient condition for τ *not* to be a time of irregularity is that

$$\int_0^\tau \|\boldsymbol{v}(t)\|_q^r\mathrm{d}t < \infty\,, \quad \frac{2}{r} + \frac{3}{q} = 1\,, \quad \text{some } q \in (3,\infty]\,. \tag{86}$$

However, from $(80)_1$ and from (76), it is immediately verified that, in the case $n = 3$, the solutions constructed in Theorem 12 satisfy the following condition

$$\int_0^\tau \|\boldsymbol{v}(t)\|_q^r\mathrm{d}t < \infty\,, \quad \frac{2}{r} + \frac{3}{q} = 1 + \frac{1}{2}\,, \quad \text{all } q \in [2,6]\,. \tag{87}$$

Therefore, in view of Lemma 8 (iii), the best conclusion we can draw is that for 3D flow Problems 1 and 2 can be positively answered *if the size of the initial data is suitably restricted.*

Remark 8 In the case $n = 3$, besides (86), one may furnish other sufficient conditions for the absence of a time of irregularity. We refer, among others, to the papers [6,8,15,18,56,57,69,80]. In particular, we would like to direct attention to the work [18,80], where the difficult borderline case $q = n = 3$ in condition (86) is worked out by completely different methods than those used here.

Open Question In the case $n = 3$, it is not known whether or not condition (86) (or any of the other conditions referred to in Remark 8) holds along solutions of Theorem 12.

Less Regular Solutions and Partial Regularity Results in 3D

As shown in the previous section, we do not know if, for 3D flow, the solutions of Theorem 12 exist in an arbitrarily large time interval, without restricting the magnitude of the initial data: they are *local solutions*. However, following the line of thought introduced by J. Leray [62], we may extend them to solutions defined for all times, for initial data of arbitrary magnitude, namely, to *global solutions*, but belonging to a functional class, C, a priori less regular than that given in (76) (*weak solutions*). Thus, if, besides existence, we could prove in C also uniqueness and continuous dependence, Problem 1 would receive a positive answer. Unfortunately, to date, the class where global solutions are proved to exist is, in principle, too large to secure the validity of these latter two properties and some extra assumptions are needed. Alternatively, in relation to Problem 2, one may investigate the "size" of the space-time regions where these generalized solutions may (possibly) become irregular. As a matter of fact, singularities, if they at all occur, have to be concentrated within "small" sets of space-time. Our objective in this section is to discuss the above issues and to present the main results.

For future purposes, we shall present some of these results also in space dimension $n = 2$, even though, as shown in Theorem 15, in this case, both Problems 1 and 2 are answered in the affirmative.

Weak Solutions and Related Properties We begin to introduce the corresponding definition of weak solutions in the sense of Leray–Hopf [54,62]. By formally dot-multiplying through both sides of (7) by $\boldsymbol{\varphi} \in \mathcal{D}(\Omega)$ and by integrating by parts over Ω, with the help of 12 we find

$$\frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{v}(t), \boldsymbol{\varphi}) + \nu(\nabla \boldsymbol{v}(t), \nabla \boldsymbol{\varphi}) + (\boldsymbol{v}(t) \cdot \nabla \boldsymbol{v}(t), \boldsymbol{\varphi}) = 0\,, \qquad \text{for all } \boldsymbol{\varphi} \in \mathcal{D}(\Omega)\,. \tag{88}$$

Definition 7 Let $\Omega \subset \mathbb{R}^n$, $n = 2, 3$. A field $\boldsymbol{v} : \Omega \times (0, \infty) \mapsto \mathbb{R}^n$ is a weak solution to (7)–$(8)_{\text{hom}}$ if and only if: (i) $\boldsymbol{v} \in L^\infty(0, T; L^2_\sigma(\Omega)) \cap L^2(0, T; \mathcal{D}^{1,2}_0(\Omega))$, for all $T > 0$; (ii) $\boldsymbol{v}$ satisfies (88) for a.a. $t \geq 0$, and (iii) $\lim_{t \to 0^+} (\boldsymbol{v}(t) - \boldsymbol{v}_0, \boldsymbol{\varphi}) = 0$, for all $\boldsymbol{\varphi} \in \mathcal{D}(\Omega)$.

A. Existence The proof of existence of weak solutions is easily carried out, for example, by the finite-dimensional (Galerkin) method indicated in Subsection V.2.2. This time, however, along the "approximate" solutions $\boldsymbol{v}_N$ to (74), we only use the estimate corresponding to the energy equality (59). We thus obtain

$$\|\boldsymbol{v}_N(t)\|_2^2 + 2\nu \int_s^t |\boldsymbol{v}_N(\rho)|_{1,2}^2 d\rho = \|\boldsymbol{v}_N(s)\|_2^2\,, \qquad 0 \leq s \leq t\,. \tag{89}$$

As we already emphasized, the important feature of this estimate is that it holds for *all* $t \geq 0$ *and all data* $v_0 \in L^2_\sigma(\Omega)$. From (89) it follows, in particular, that the sequence $\{\boldsymbol{v}_N\}$ is uniformly bounded in the class of functions specified in (i) of Definition 7. Using this fact together with classical weak and strong compactness arguments in (74), we can show the existence of at least one subsequence converging, in suitable topologies, to a weak solution $\boldsymbol{v}$. We then have the following result for whose complete proof we refer, e. g., to Theorem 3.1 in [34].

Theorem 16 *For any* $\boldsymbol{v}_0 \in L^2_\sigma(\Omega)$ *there exists at least one weak solution to (7)–*$(8)_{\text{hom}}$. *This solution verifies, in addition, the following properties.*

(i) The energy inequality:

$$\|\boldsymbol{v}(t)\|_2^2 + 2\nu \int_s^t |\boldsymbol{v}(\rho)|_{1,2}^2 \,\mathrm{d}\rho \leq \|\boldsymbol{v}(s)\|_2^2\,, \qquad \textit{for a.a. } s \in [0, \infty), 0 \textit{ included, and all } t \geq s\,. \tag{90}$$

(ii) $\lim_{t \to 0^+} \|\boldsymbol{v}(t) - \boldsymbol{v}_0\|_2 = 0$.

B. On the Energy Equality In the case $n = 3$, weak solutions in Theorem 16 only satisfy the energy *inequality* (90) instead of the energy *equality* [see (59)]. (For the case $n = 2$, see Remark 9.) This is an undesired feature that is questionable from the physical viewpoint. As a matter of fact, for fixed $s \geq 0$, in time intervals $[s, t]$ where (90) were to hold as a *strict* inequality, the kinetic energy would decrease by an amount which is not only due to the dissipation. From a strictly technical viewpoint, this happens because the convergence of (a subsequence of) the sequence $\{\boldsymbol{v}_N\}$ to the weak solution $\boldsymbol{v}$ can be proved only in the weak topology of $L^2(0, T; \mathcal{D}^{1,2}_0(\Omega))$ and this only ensures that, as $N \to \infty$, the second term on the left-hand side of (89) tends to a quantity not less than the one given by the second term on the left-hand side of (90). One may think that this circumstance is due to the special method used for constructing the weak solutions. Actually, this is not the case because, in fact, we have the following.

Open Question If $n = 3$, it is not known if there are solutions satisfying (90) with the equality sign and corresponding to initial data $\boldsymbol{v}_0 \in L^2_\sigma(\Omega)$ of unrestricted magnitude.

Remark 9 A sufficient condition for a weak solution, $\boldsymbol{v}$, to satisfy the energy equality (59) is that $\boldsymbol{v} \in L^4(0, t; L^4(\Omega))$, for all $t > 0$ (see Theorem 4.1 in [34]). Consequently, from (85) it follows that, if $n = 2$, weak solutions satisfy (59) for all $t > 0$. Moreover, from $(80)_1$, we find that the solutions of Theorem 12, for $n = 3$, satisfy the energy equality (59), at least for all $t \in [0, \tau)$.

Remark 10 For future purposes, we observe that the definition of weak solution and the results of Theorem 16 can be extended easily to the case when $\boldsymbol{f} \not\equiv \boldsymbol{0}$. In fact, it is enough to change Definition 7 by requiring that $\boldsymbol{v}$ satisfies the modification of (88) obtained by adding to its right-hand side the term $(\boldsymbol{f}, \boldsymbol{\varphi})$. Then, if $\boldsymbol{f} \in L^2(0, T; \mathcal{D}_0^{-1,2}(\Omega))$, for all $T > 0$, one can show the existence of a weak solution satisfying condition (ii) of Theorem 16 and the variant of (90) obtained by adding the term $\int_0^t (\boldsymbol{f}, \boldsymbol{v})\, ds$ on its right-hand side.

C. Uniqueness and Continuous Dependence The following result, due to Serrin (Theorem 6 in [83]) and Sather (Theorem 5.1 in [75]), is based on ideas of Leray [§ 32] in [62] and Prodi [71]. A detailed proof is given in Theorem 4.2 in [34].

Theorem 17 *Let $\boldsymbol{v}$, $\boldsymbol{u}$ be two weak solutions corresponding to data $\boldsymbol{v}_0$ and $\boldsymbol{u}_0$. Assume that $\boldsymbol{u}$ satisfies the energy inequality (90) with $s = 0$, and that*

$$\boldsymbol{v} \in L^r(0, T; L^q(\Omega)), \quad \text{for some } q \in (n, \infty] \quad \text{such that } \frac{2}{r} + \frac{n}{q} = 1. \tag{91}$$

Then,

$$\|\boldsymbol{v}(t) - \boldsymbol{u}(t)\|_2^2 \le C\|\boldsymbol{v}_0 - \boldsymbol{u}_0\|_2^2 \exp\{\int_0^t \|\boldsymbol{v}(\rho)\|_q^r d\rho\}, \quad \text{for all } t \in [0,T]\}$$

where $C = C(\Omega, \nu) > 0$. Thus, in particular, if $\boldsymbol{v}_0 = \boldsymbol{u}_0$, then $\boldsymbol{v} = \boldsymbol{u}$ a.e. in $\Omega \times [0, T]$.

Remark 11 If $n = 2$, from (85) and Remark 9 we find that every weak solution satisfies the assumptions of Theorem 17. Therefore, in such a case, every weak solution is unique in the class of weak solutions and depends continuously upon the data. Furthermore, if $n = 3$, the uniqueness result continues to hold if in condition (91), we take $q = n = 3$; see [55,87].

Open Question While the existence of weak solutions $\boldsymbol{u}$ satisfying the hypothesis Theorem 17 is secured by Theorem 16, in the case $n = 3$ it is not known if there exist weak solutions having the property stated for $\boldsymbol{v}$. [In principle, as a consequence of Definition 7(i) and $(80)_1$, $\boldsymbol{v}$ only satisfies (87).] Consequently in the case $n = 3$ uniqueness and continuous dependence in the class of weak solutions remains open, and so does the resolution of Problem 1.

Remark 12 As a matter of fact, weak solutions possess more regularity than that implied by their very definition. Actually, if $n = 2$, they are indeed smooth (see Remark 13). If $n = 3$, by means of sharp estimates for solutions to the linear problem obtained from (7)–(8) by neglecting the nonlinear term $\boldsymbol{v} \cdot \nabla \boldsymbol{v}$ (*Stokes problem*)– one can show that every corresponding weak solution satisfies the following additional properties (see Theorem 3.1 in [43], Theorem 3.4 in [87])

$$\frac{\partial \boldsymbol{v}}{\partial t} \in L^l(\delta, T; L^s(\Omega)), \quad \boldsymbol{v} \in L^l(\delta, T; W^{2,s}(\Omega)), \quad \text{for all } T > 0 \quad \text{and all } \delta \in (0, T),$$

where the numbers l, s obey the following conditions

$$\frac{2}{l} + \frac{3}{s} \ge 4, \quad l \in [1, 2], \quad s \in \left[1, \tfrac{3}{2}\right].$$

Moreover, there exists $p \in L^l(\delta, T; W^{1,s}(\Omega)) \cap L^l(\delta, T; L^{3s/(3-s)}(\Omega))$ such that the pair $(\boldsymbol{v}, p)$ satisfies (7) for a.a. $(\boldsymbol{x}, t) \in \Omega \times (0, \infty)$. If, in addition, $\boldsymbol{v}_0$ lies in a sufficiently regular subspace of $L^2_\sigma(\Omega)$, we can take $\delta = 0$. However, the above properties are still not enough to ensure the validity of condition (91). Weak solutions enjoying further regularity properties are constructed in [14,67] and Theorem 3.1 and Corollary 3.2 in [24].

D. Partial Regularity and "Suitable" Weak Solutions A problem of fundamental importance is to investigate the set of space-time where weak solutions may possibly become irregular, and to give an estimate of how "big" this set can be.

To this end, we recall that, for a given $S \subset \mathbb{R}^{d+1}$, $d \in \mathbb{N} \cup \{0\}$, and $\kappa \in (0, \infty)$, the *κ-dimensional (spherical) Hausdorff measure* $\mathcal{H}^\kappa$ of S is defined as

$$\mathcal{H}^\kappa(S) = \lim_{\delta \to 0} \mathcal{H}^\kappa_\delta(S),$$

where $\mathcal{H}^\kappa_\delta(S) = \inf \sum_i r_i^\kappa$, the infimum being taken over all at most countable coverings $\{B_i\}$ of S of closed balls $B_i \subset \mathbb{R}^d$ of radius r_i with $r_i < \delta$; see, e. g., [19]. If $d \in \mathbb{N}$, the κ-dimensional *parabolic* Hausdorff measure $\mathcal{P}^\kappa$ of S is defined as above, by replacing the ball B_i with a *parabolic*

cylinder of radius r_i:

$$Q_{r_i}(x,t) = \{(y,s) \in \mathbb{R}^d \times \mathbb{R} : |y-x| < r_i\,, |s-t| < r_i^2\}\,. \tag{92}$$

In general, it is $\mathcal{H}^\kappa(S) \le C\mathcal{P}^\kappa(S)$, $C > 0$; see § 2.10.1 in [19] for details.

The following lemma is a direct consequence of the preceding definition.

Lemma 9 *For any $S \subset \mathbb{R}^{d+1}$, $d \in \mathbb{N} \cup \{0\}$ (respectively, $d \in \mathbb{N}$), we have $\mathcal{H}^\kappa(S) = 0$ (respectively, $\mathcal{P}^\kappa(S) = 0$) if and only if, for each $\delta > 0$, S can be covered by closed balls $\{B_i\}$ (respectively, parabolic cylinders $\{Q_i\}$) of radii r_i, $i \in \mathbb{N}$, such that $\sum_{i=1}^\infty r_i^\kappa < \delta$.*

We begin to consider the collection of times where weak solutions are (possibly) not smooth and show that they constitute a very "small" region of $(0,\infty)$. Specifically, we have the following result, basically, due to Leray pp. 244–245 in [62] and completed by Scheffer [76].

Theorem 18 *Let Ω be a bounded domain of class C^∞. Assume $\boldsymbol{v}$ is a weak solution determined in Theorem 16. Then, there exists a union of disjoint and, at most, countable open time intervals $\mathcal{T} = \mathcal{T}(\boldsymbol{v}) \subset (0,\infty)$ such that:*

(i) $\boldsymbol{v}$ is of class C^∞ in $\bar\Omega \times \mathcal{T}$,
(ii) There exists $T^* \in (0,\infty)$ such that $\mathcal{T} \supset (T^*,\infty)$;
(iii) If $\boldsymbol{v}_0 \in \mathcal{D}_0^{1,2}(\Omega)$ then $\mathcal{T} \supset (0,T_1)$ for some $T_1 > 0$;
(iv) Let (s,τ) be a generic bounded interval in $\mathcal{T}(\boldsymbol{v})$ and suppose $\boldsymbol{v} \notin C^\infty(\bar\Omega \times (s,\tau_1))$, $\tau_1 > \tau$. Then, both following conditions must hold

$$|\boldsymbol{v}(t)|_{1,2}^2 \ge \frac{C}{(\tau-t)^{1/2}}\,, \quad t \in (s,\tau) \quad \text{and}$$
$$\lim_{t\to\tau^-} \int_s^t \|\boldsymbol{v}(s)\|_q^{2q/(q-3)} ds = \infty\,, \quad \text{for all } q > 3\,, \tag{93}$$

where $C = C(\Omega,\nu) > 0$;
(v) The $\frac{1}{2}$-dimensional Hausdorff measure of $\mathcal{I}(\boldsymbol{v}) := (0,\infty) - \mathcal{T}$ is zero;

Proof By (90) we may select $T^* > 0$ with the following properties: (a) $\|\boldsymbol{v}(T^*)\|_2 |\boldsymbol{v}(T^*)|_{1,2} < K$, and (b) the energy inequality (90) holds with $s = T^*$, where K is the constant introduced in Lemma 8 (iii). Let us denote by $\tilde{\boldsymbol{v}}$ the solution of Theorem 12 corresponding to the data $\boldsymbol{v}(T^*)$. By Lemma 8 (iii), $\tilde{\boldsymbol{v}}$ exists for all times $t \ge T^*$ and, by Theorem 14, it is of class C^∞ in $\Omega \times (T^*,\infty)$. By Lemma 7 and by Theorem 17 we must have $\boldsymbol{v} \equiv \tilde{\boldsymbol{v}}$ in $\Omega \times (T^*,\infty)$, and part (ii) is proved.

Next, denote by I the set of those $t \in [0,T^*)$ such that (a) $\|\boldsymbol{v}(t)\|_{1,2} < \infty$, and (b) the energy inequality (90) holds with $s \in I$. Clearly, $[0,T^*] - I$ is of zero Lebesgue measure. Moreover, for every $t_0 \in I$ we can construct in $(t_0, t_0 + T(t_0))$ a solution $\tilde{\boldsymbol{v}}$ assuming at t_0 the initial data $\boldsymbol{v}(t_0)$ $(\in \mathcal{D}_0^{1,2}(\Omega))$; see Theorem 12. From Theorem 14, Lemma 7, and Theorem 17, we know that $\tilde{\boldsymbol{v}}$ is of class C^∞ in $\Omega \times (t_0, t_0 + T(t_0))$ and that it coincides with $\boldsymbol{v}$, since this latter satisfies the energy inequality with $s = t_0$. Furthermore, if $\boldsymbol{v}_0 \in \mathcal{D}_0^{1,2}(\Omega)$, then $0 \in I$. Properties (ii)–(iv) thus follow with $\mathcal{T} \equiv \bigcup_{i\in\mathfrak{I}}(s_i,\tau_i) \cup (T^*,\infty)$, where (s_i,τ_i) are the connected components in I. Notice that

$$\begin{aligned} &(s_i,\tau_i) \subset [0,T^*]\,, \quad \text{for all } i \in \mathfrak{I};\\ &(s_i,\tau_i) \cap (s_j,\tau_j) = \emptyset\,, \quad i \ne j\,, \end{aligned} \tag{94}$$

and that, moreover, the (1-dimensional) Lebesgue measure of $\mathcal{I} := \mathcal{T} - (0,\infty)$ is 0. Finally, property (iv) is an immediate consequence of Lemma 8 and Theorem 14. It remains to show (v). From (iv) and (90) we find

$$\begin{aligned} &\sum_{i\in\mathfrak{I}} (\tau_i - s_i)^{1/2} \le 1/(2C)\\ &\sum_{i\in\mathfrak{I}} \int_{\tau_i}^{s_i} \|\nabla \boldsymbol{v}(\tau)\|_2^2 dt \le \|\boldsymbol{v}_0\|_2^2/(4C)\,. \end{aligned}$$

Thus, for every $\delta > 0$ we can find a finite part $\mathfrak{I}_\delta$ of $\mathfrak{I}$ such that

$$\sum_{i\notin\mathfrak{I}_\delta} (\tau_i - s_i) < \delta, \quad \sum_{i\notin\mathfrak{I}_\delta} (\tau_i - s_i)^{1/2} < \delta\,. \tag{95}$$

By $(94)_1$, $\cup_{i\in\mathfrak{I}}(s_i,\tau_i) \subset [0,T^*]$ and so the set $[0,T^*] - \cup_{i\in\mathfrak{I}_\delta}(s_i,\tau_i)$ consists of a finite number of disjoint closed intervals B_j, $j = 1,\dots,N$. Clearly,

$$\bigcup_{j=1}^N B_j \supset \mathcal{I}(\boldsymbol{v}). \tag{96}$$

By $(94)_2$, each interval (s_i,τ_i), $i \notin \mathfrak{I}_\delta$, is included in one and only one B_j. Denote by $\mathfrak{I}_j$ the set of all indices i satisfying $B_j \supset (s_i,\tau_i)$. We thus have

$$\begin{aligned} &\mathfrak{I} = \mathfrak{I}_\delta \cup \left(\bigcup_{j=1}^N \mathfrak{I}_j\right),\\ &B_j = \left(\bigcup_{i\in\mathfrak{I}_j} (s_i,\tau_i)\right) \cup \big(B_j \cap \mathcal{I}(\boldsymbol{v})\big)\,. \end{aligned} \tag{97}$$

Since $\mathcal{I}$ has zero Lebesgue measure, from$(97)_2$ we have $\operatorname{diam} B_j = \sum_{i\in\mathfrak{I}_j}(\tau_i - s_i)$. Thus, by (95) and $(97)_1$,

$$\operatorname{diam} B_j \le \sum_{i\notin\mathfrak{I}_\delta} (\tau_i - s_i) < \delta \tag{98}$$

and, again by (95) and $(97)_1$,

$$\sum_{j=1}^{N}(\operatorname{diam} B_j)^{1/2} \le \sum_{j=1}^{N}\left(\sum_{i\in\mathfrak{I}_j}(\tau_i - s_i)\right)^{1/2} \le \sum_{i\notin\mathfrak{I}_\delta}(\tau_i - s_i)^{1/2} < \delta\,. \tag{99}$$

Therefore, property (v) follows from (96), (98), (99), and Lemma 9. □

Remark 13 If Ω is of class C^∞, from 85, Remark 11 and from Theorem 18 (v) we at once obtain that, for $n = 2$, every weak solution is of class $C^\infty(\bar{\Omega} \times (0, t))$, for all $t > 0$.

We shall next analyze, in more detail, the set of points where weak solutions may possibly lose regularity. In view of Remark 13, *we shall restrict ourselves to consider the case n=3 only*.

Let (s, τ) be a bounded interval in $\mathcal{T}(\boldsymbol{v})$ and assume that, at $t = \tau$, the weak solution $\boldsymbol{v}$ becomes irregular. We may wish to estimate the spatial set $\Sigma = \Sigma(\tau) \subseteq \Omega$ where $\boldsymbol{v}(\tau)$ becomes irregular. By defining Σ as the set of $\boldsymbol{x} \in \Omega$ where $\boldsymbol{v}(\boldsymbol{x}, \tau)$ is not continuous, in the case $\Omega = \mathbb{R}^3$, Scheffer has shown that $\mathcal{H}^1(\Sigma) < \infty$ [77].

More generally, one may wish to estimate the "size" of the region of space-time where points of irregularity (appropriately defined, see Definition 9 below) may occur. This study, initiated by Scheffer [77,78] and continued and deepened by Caffarelli, Kohn and Nirenberg [11], can be performed in a class of solutions called *suitable weak solutions* which, in principle, due to the lack of an adequate uniqueness theory, is more restricted than that of weak solutions.

Definition 8 A pair $(\boldsymbol{v}, p)$ is called a suitable weak solution to (7)–$(8)_{\text{hom}}$ if and only if: (i) $\boldsymbol{v}$ satisfies Definition 7(i) and $p \in L^{3/2}((0, T); L^{3/2}(\Omega))$, for all $T > 0$; (ii) $(\boldsymbol{v}, p)$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{v}(t), \boldsymbol{\psi}) + \nu(\nabla\boldsymbol{v}(t), \nabla\boldsymbol{\psi}) + (\boldsymbol{v}(t)\cdot\nabla\boldsymbol{v}(t), \boldsymbol{\psi}) - (p(t), \operatorname{div}\boldsymbol{\psi}) = 0\,, \quad \text{for all } \boldsymbol{\psi} \in C_0^\infty(\Omega)\,,$$

and (iii) $(\boldsymbol{v}, p)$ obeys the following *localized energy inequality*

$$2\int_0^T\int_\Omega |\nabla\boldsymbol{v}|^2\boldsymbol{\phi}\,\mathrm{d}\boldsymbol{x}\mathrm{d}t \le \int_0^T\int_\Omega \{|\boldsymbol{v}|^2(\frac{\partial\boldsymbol{\phi}}{\partial t} + \Delta\boldsymbol{\phi}) + (|\boldsymbol{v}|^2 + 2p)\boldsymbol{v}\cdot\nabla\boldsymbol{\phi}\}\,\mathrm{d}\boldsymbol{x}\mathrm{d}t\,,$$

for all non-negative $\boldsymbol{\phi} \in C_0^\infty(\Omega \times (0, T))$.

Remark 14 By taking $l = 3/2$, $s = 9/8$ in Remark 12, it follows that *every* weak solution, corresponding to sufficiently regular initial data, matches requirements (i) and (ii) of Definition 8 (recall that Ω is bounded). However, it is not known, to date, if it satisfies also condition (iii). Moreover, it is not clear if the finite-dimensional (Galerkin) method used for Theorem 16 is appropriate to construct solutions obeying such a condition (see, [47] for a partial answer). Nevertheless, by using different methods, one can show the existence of at least one weak solution satisfying the properties stated in Theorem 16, and which, in addition, is a suitable weak solution; (see [5], Theorem A.1 in [11], and Theorem 2.2 in [64]).

Definition 9 A point $P := (\boldsymbol{x}, t) \in \Omega_T := \Omega \times (0, T)$ is called regular for a suitable weak solution $(\boldsymbol{v}, p)$, if and only if there exists a neighborhood, I, of P such that $\boldsymbol{v}$ is in $L^\infty(I)$. A point which is not regular will be called irregular.

Remark 15 The above definition of a regular point is reinforced by a result of Serrin [82], from which we deduce that, *in the neighborhood of every regular point, a suitable weak solution is, in fact, of class C^∞ in the space variables.*

The next result is crucial in assessing the "size" of the set of possible irregular points in the space-time domain. For its proof, we refer to [64], Theorem 2.2 in [60], Proposition 2 in [11]. We recall that $Q_r(\boldsymbol{x}, t)$ is defined in (92).

Lemma 10 *Let $(\boldsymbol{v}, p)$ be a suitable weak solution and let $(\boldsymbol{x}, t) \in \Omega_T$. There exists $K > 0$ such that, if*

$$\limsup_{r\to 0} r^{-1}\int_{Q_r(\boldsymbol{x},t)} |\nabla\boldsymbol{v}(\boldsymbol{y}, s)|^2\,\mathrm{d}\boldsymbol{y}ds < K\,, \tag{100}$$

then $(\boldsymbol{x}, t)$ is regular.

Now, let $S = S(\boldsymbol{v}, p) \subseteq \Omega \times (0, T)$ be the set of possible irregular points for a suitable weak solution $(\boldsymbol{v}, p)$, and let V be a neighborhood of S. By Lemma 10 we then have that, for each $\delta > 0$, there is $Q_r(\boldsymbol{x}, t) \subset V$ with $r < \delta$, such that

$$K^{-1}\int_{Q_r(\boldsymbol{x},t)} |\nabla\boldsymbol{v}(\boldsymbol{y}, s)|^2 d\boldsymbol{y}\,\mathrm{d}s > r\,. \tag{101}$$

Let $\mathcal{Q} = \{Q_r(\boldsymbol{x}, t)\}$ be the collection of all Q satisfying this property. Since $\Omega \times (0, T)$ is bounded, from Lemma 6.1 in [11] we can find an at most countable, disjoint subfamily of $\mathcal{Q}$, $\{Q_{r_i}(\boldsymbol{x}_i, t_i)\}$, such that $S \subset \cup_{i\in\mathfrak{I}} Q_{5r_i}(\boldsymbol{x}_i, t_i)$.

From (101) it follows, in particular, that

$$\sum_{i\in\mathfrak{I}} r_i \le K^{-1}\sum_{i\in\mathfrak{I}}\int_{Q_{r_i}(\boldsymbol{x}_i,t_i)} |\nabla \boldsymbol{v}(\boldsymbol{y},s)|^2\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}s \le K^{-1}\int_V |\nabla \boldsymbol{v}|^2 . \quad (102)$$

Since δ is arbitrary, 102 implies, on the one hand, that S is of zero Lebesgue measure and, on the other hand, that

$$\mathcal{P}^1(S) \le \frac{5}{K}\int_V |\nabla \boldsymbol{v}|^2 ,$$

for every neighborhood V of S. Thus, by the absolute continuity of the Lebesgue integral, from this latter inequality and from Lemma 9 we have the following Theorem B in [11].

Theorem 19 *Let $(\boldsymbol{v}, p)$ be a suitable weak solution and let $S = S(\boldsymbol{v}, p) \subseteq \Omega \times (0, T)$ be the corresponding set of possible irregular points. Then $\mathcal{P}^1(S) = 0$.*

Remark 16 Let

$$\begin{aligned}\mathcal{Z} &= \mathcal{Z}(\boldsymbol{v}, p)\\ &:= \{t \in (0, T)\colon (\boldsymbol{x}, t) \in S(\boldsymbol{v}, p)\,, \text{for some } \boldsymbol{x} \in \Omega\}\,.\end{aligned}$$

Clearly, $t \in \mathcal{Z}$ if and only if $\boldsymbol{v}$ becomes essentially unbounded around $(\boldsymbol{x}, t)$, for some $\boldsymbol{x} \in \Omega$. Namely, for any $M > 0$ there is a neighborhood of $(\boldsymbol{x}, t)$, I_M, such that $|\boldsymbol{v}(\boldsymbol{y}, s)| > M$ for a.a. $(\boldsymbol{y}, s) \in I_M$. From Theorem 18(ii) we deduce at once that $\mathcal{Z}(\boldsymbol{v}, p) \subset \mathcal{I}(\boldsymbol{v})$, where $\mathcal{I}(\boldsymbol{v})$ is the set of all possible times of irregularity; see Definition 6. Thus, from Theorem 18(i) we find $\mathcal{H}^{1/2}(\mathcal{Z}) = 0$.

Remark 17 There is a number of papers dedicated to the formulation of sufficient conditions for the absence of irregular points, $(\boldsymbol{x}, t)$, for a suitable weak solution, when $\boldsymbol{x}$ is either an interior or a boundary point. In this latter case, the definition of the regular point as well as that of the suitable weak solution must be, of course, appropriately modified. Among others, we refer to [11,48,49,64,81].

Long-Time Behavior and Existence of the Global Attractor

Suppose a viscous liquid moves in a fixed spatial bounded domain, Ω, under the action of a given time-independent driving mechanism, m, and denote by $\lambda > 0$ a non-dimensional parameter measuring the "magnitude" of m. To fix the ideas, we shall assume that the velocity field of the liquid, $\boldsymbol{v}_1$, vanishes at $\partial\Omega$, for each time $t \ge 0$. This assumption is made for the sake of simplicity. All main results can be extended to the case $\boldsymbol{v}_1 \not\equiv \boldsymbol{0}$, provided $\boldsymbol{v}_1(\boldsymbol{x}, t)\cdot\boldsymbol{n}|_{\partial\Omega} = 0$, for all $t \ge 0$, where $\boldsymbol{n}$ is the unit normal on $\partial\Omega$. We then take m to be a (non-conservative) time-independent body force, $\boldsymbol{f} \in L^2(\Omega)$ with $\lambda \sim \|\boldsymbol{f}\|_2$. We shall denote by (7)–$(8)_{\mathrm{Hom}}$ the initial-boundary value problem (7)–(8) with $\boldsymbol{v}_1 \equiv \boldsymbol{0}$. As we know from Theorem 3 and Remark 5, if λ is "sufficiently" small, less than λ_c, say, there exists one and only one (steady-state) solution, $(\bar{\boldsymbol{v}}, \bar{p})$ to the boundary-value problem (9)–(10) (with $\boldsymbol{v}_* \equiv \boldsymbol{0}$). Actually, it is easy to show that, with the above restriction on λ, every solution to the initial-boundary value problem (8)–$(8)_{\mathrm{Hom}}$, belonging to a sufficiently regular function class and corresponding to the given $\boldsymbol{f}$ and to arbitrary $\boldsymbol{v}_0 \in L^2_\sigma(\Omega)$, decays exponentially fast in time to $(\bar{\boldsymbol{v}}, \bar{p})$. In fact, by setting $\boldsymbol{u} := \boldsymbol{v} - \bar{\boldsymbol{v}}$, $P := p - \bar{p}$, from (7)–(10) we find

$$\left.\begin{aligned}\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{u} + \boldsymbol{u}\cdot\nabla\bar{\boldsymbol{v}}\\ + \bar{\boldsymbol{v}}\cdot\nabla\boldsymbol{u} = -\nabla P + \nu\Delta\boldsymbol{u}\,,\\ \operatorname{div}\boldsymbol{u} = 0\end{aligned}\right\} \quad \text{in } \Omega\times(0,T)\,;$$
$$\boldsymbol{u}(\boldsymbol{x},0) = \boldsymbol{v}_0(\boldsymbol{x}) - \bar{\boldsymbol{v}}(\boldsymbol{x})\,,\ \boldsymbol{x}\in\Omega\,;\ \boldsymbol{u} = \boldsymbol{0} \text{ at } \partial\Omega\times(0,T)\,. \quad (103)$$

If we formally dot-multiply through both sides of $(103)_1$ by $\boldsymbol{u}$, integrate by parts over Ω and take into account 12 and Lemma 1, we obtain

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{u}\|_2^2 + \nu|\boldsymbol{u}|_{1,2}^2 = -(\boldsymbol{u}\cdot\nabla\bar{\boldsymbol{v}}, \boldsymbol{u})\,. \quad (104)$$

From Lemma 1, (21), and (18) we find $|(\boldsymbol{u}\cdot\nabla\bar{\boldsymbol{v}}, \boldsymbol{u})| \le c_1\,|\boldsymbol{u}|_{1,2}^2|\bar{\boldsymbol{v}}|_{1,2}$, with $c_1 = c_1(\Omega) > 0$. Thus, by Remark 5, it follows that

$$|(\boldsymbol{u}\cdot\nabla\bar{\boldsymbol{v}}, \boldsymbol{u})| \le \frac{c_2}{\nu}\,\|\boldsymbol{f}\|_2|\boldsymbol{u}|_{1,2}^2\,,$$

with $c_2 = c_2(\Omega) > 0$, which, in turn, once replaced in (104), furnishes

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{u}\|_2^2 + \left(\nu - \frac{c_2}{\nu}\,\|\boldsymbol{f}\|_2\right)|\boldsymbol{u}|_{1,2}^2 \le 0\,.$$

Therefore, if $\gamma := \nu - (c_2/\nu)\|\boldsymbol{f}\|_2 > 0$, from (18) and from the latter displayed equation we deduce

$$\|\boldsymbol{u}(t)\|_2^2 \le \|\boldsymbol{u}(0)\|_2^2\, e^{-2(\gamma/C_P)\,t}\,, \quad (105)$$

which gives the desired result. Estimate (105) can be read, in "physical terms" in the following way: after a certain amount of time (depending on how close γ is to 0, namely, on how close λ is to λ_c), the transient motion will die out exponentially fast and the "true" dynamics of the fluid will

be described by the unique steady-state flow corresponding to the given force $\boldsymbol{f}$. From a mathematical point of view, the steady-state $(\bar{\boldsymbol{v}}, \bar{p})$ is, in a suitable function space, a one-point set which is invariant under the flow. Now, let us increase λ higher and higher beyond λ_c. Following Eberhard Hopf [53], we expect that, after a while, the transient motion will yet die out, and that the generic flow will approach a certain manifold, $\mathfrak{M} = \mathfrak{M}(\lambda)$ which need not reduce to a single point. (There are, however, explicit examples where $\mathfrak{M}(\lambda)$ remains a single point for any $\lambda > 0$; see [65].) Actually, in principle, the structure of $\mathfrak{M}$ can be very involved. Nevertheless, we envisage that $\mathfrak{M}$ is still invariant under the flow, and that it is $\mathfrak{M}$ where, eventually, the "true" dynamics of the liquid will take place. For obvious reasons, the manifold $\mathfrak{M}$ is called the *global attractor*.

The existence of a global attractor and the study of the dynamics of the liquid on it, could be of the utmost importance in the effort of formulating a mathematical theory of *turbulence*. Actually, as is well known, if the magnitude of the driving force becomes sufficiently large, the corresponding flow becomes chaotic and the velocity and pressure of the liquid exhibit large and completely random variation in space and time. According to the ideas proposed by Smale [85] and by Ruelle and Takens [74], this chaotic behavior could be explained by the existence of a very complicated global attractor, where, as mentioned before, the ultimate dynamics of the liquid occurs.

Existence of the Global Attractor for Two-Dimensional Flow, and Related Properties *Throughout this section we shall consider two-dimensional flow*, so that, in particular, $\Omega \subset \mathbb{R}^2$.

Let $\boldsymbol{f} \in L^2(\Omega)$ and $\nu > 0$ be given. Consider the one-parameter family of operators

$$S_t : \boldsymbol{a} \in L^2_\sigma(\Omega) \mapsto S_t(\boldsymbol{a}) := \boldsymbol{v}(t) \in L^2_\sigma(\Omega)\,, \qquad t \in [0, \infty)$$

where $\boldsymbol{v}(t)$ is, at each t, the weak solution to (7)–$(8)_{\mathrm{Hom}}$ corresponding to the initial data $\boldsymbol{a}$; see Remark 10. From Remark 9 and Remark 11 we deduce that the family $\{S_t\}_{t\geq 0}$ defines a (strongly) *continuous semigroup* in $L^2_\sigma(\Omega)$, namely (i) $S_{t_1}S_{t_2}(\boldsymbol{a}) = S_{t_1+t_2}(\boldsymbol{a})$ for all $\boldsymbol{a} \in L^2_\sigma(\Omega)$ and all $t_1, t_2 \in [0, \infty)$; (ii) $S_0(\boldsymbol{a}) = \boldsymbol{a}$, and (iii) the map $t \mapsto S_t(\boldsymbol{a})$ is continuous for all $a \in L^2_\sigma(\Omega)$.

Definition 10 For any given $\boldsymbol{f} \in L^2(\Omega)$, the corresponding pair $\{L^2_\sigma(\Omega), S_t\}$ is called semi-flow associated to (7)–$(8)_{\mathrm{Hom}}$

Our objective is to study the asymptotic properties (as $t \to \infty$) of the semi-flow $\{L^2_\sigma(\Omega), S_t\}$. To this end, we need to recall some basic facts.

Given $\mathcal{A}_1, \mathcal{A}_2 \subset L^2_\sigma(\Omega)$, we set

$$\mathfrak{d}(\mathcal{A}_1, \mathcal{A}_2) = \sup_{\boldsymbol{u}_1 \in \mathcal{A}_1} \inf_{\boldsymbol{u}_2 \in \mathcal{A}_2} \|\boldsymbol{u}_1 - \boldsymbol{u}_2\|_2\,.$$

Notice that $\mathfrak{d}(\mathcal{A}_1, \mathcal{A}_2) = 0 \Rightarrow \mathcal{A}_1 \subseteq \bar{\mathcal{A}}_2$. Moreover, we denote by $\mathfrak{A}$ the class of all bounded subset of $L^2_\sigma(\Omega)$.

Definition 11 $\mathcal{B} \subset L^2_\sigma(\Omega)$ is called: (i) absorbing iff for any $\mathcal{A} \in \mathfrak{A}$ there is $t_0 = t_0(\mathcal{A}) \geq 0$ such that $S_t(\mathcal{A}) \subseteq \mathcal{B}$, for all $t \geq t_0$; (ii) attracting iff $\lim_{t\to\infty} \mathfrak{d}(S_t(\mathcal{A}), \mathcal{B}) = 0$, for all $\mathcal{A} \in \mathfrak{A}$; (iii) invariant iff $S_t(\mathcal{B}) = \mathcal{B}$, for all $t \geq 0$; (iv) maximal invariant iff it is invariant and contains every invariant set in $\mathfrak{A}$; (v) global attractor iff it is compact, attracting and maximal invariant.

Clearly, if a global attractor exists, it is unique. Furthermore, roughly speaking, its existence is secured whenever the semiflow admits a bounded absorbing set on which the semiflow becomes, eventually, relatively compact; (see Theorem 1.1 in [94]). The following result holds.

Theorem 20 *For any $\boldsymbol{f} \in L^2_\sigma(\Omega)$ and $\nu > 0$, the corresponding semi-flow $\{L^2_\sigma(\Omega), S_t\}$ admits a global attractor $\mathfrak{M} = \mathfrak{M}(\boldsymbol{f}, \nu)$ which is also connected.*

Proof In view of Theorem 1.1 in [94] and of Rellich compactness theorem (Theorem II.4.2 in [31]), it suffices to show the following two properties: (a) existence of a bounded absorbing set, and (b) given $M > 0$, there is $t_0 = t_0(M, \boldsymbol{f}, \nu) > 0$ such that $\|\boldsymbol{a}\|_2 \leq M$ implies $|S_t(\boldsymbol{a})|_{1,2} \leq C$, for all $t \geq t_0$ and for some $C > 0$ independent of t. The starting point is the analog of (58) and (61) which, this time, take the form

$$\begin{aligned} &\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{v}(t)\|_2^2 + \nu|\boldsymbol{v}(t)|_{1,2}^2 = (\boldsymbol{f}, \boldsymbol{v})\,, \\ &\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|\boldsymbol{v}(t)|_{1,2}^2 + \nu\|P\boldsymbol{v}(t)\|_2^2 = (\boldsymbol{v}\cdot\nabla\boldsymbol{v}, P\Delta\boldsymbol{v}) - (\boldsymbol{f}, P\Delta\boldsymbol{v}) \end{aligned} \tag{106}$$

By using in these equations the Schwarz inequality, and inequalities (18), 29, $(62)_1$ and (63), we deduce

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{v}(t)\|_2^2 + (\nu/C_P)\|\boldsymbol{v}(t)\|_2^2 \leq F\,, \\ &\frac{\mathrm{d}}{\mathrm{d}t}|\boldsymbol{v}(t)|_{1,2}^2 - g(t)|\boldsymbol{v}(t)|_{1,2}^2 \leq F\,, \end{aligned} \tag{107}$$

where $F := \|\boldsymbol{f}\|_2^2/\nu$, $g(t) := c_1\|\boldsymbol{v}(t)\|_2^2|\boldsymbol{v}(t)|_{1,2}^2$, and $c_1 = c_1(\Omega) > 0$. By integrating $(107)_1$, we find

$$\|S_t(\boldsymbol{a})\|_2^2 := \|\boldsymbol{v}(t)\|_2^2 \leq \|\boldsymbol{a}\|_2^2 \mathrm{e}^{-\nu t/C_P} + (1 - e^{-\nu t/C_P})F/C_P\,. \tag{108}$$

Thus, setting

$$\mathfrak{B} := \{\boldsymbol{\varphi} \in L^2_\sigma(\Omega) \colon \|\boldsymbol{\varphi}\|_2 \le (2F/C_P)^{1/2} \equiv \rho\}\,, \qquad (109)$$

from (107) we deduce that, whenever $\|\boldsymbol{a}\|_2 < M$, there exists $t_1 = t_1(M, F, \nu) > 0$ such that $S_t(\boldsymbol{a}) \in \mathfrak{B}$, which shows that $\mathfrak{B}$ is absorbing. Next, again from Schwarz inequality and from $(98)_1$, we obtain

$$\begin{aligned} &\int_t^{t+1} |\boldsymbol{v}(s)|^2_{1,2}\mathrm{d}s \le \frac{1}{\nu}\|\boldsymbol{f}\|_2 \\ &\int_t^{t+1} \|\boldsymbol{v}(s)\|_2\mathrm{d}s \le \left(\frac{\rho^4 M^2 C_P}{2\nu^2}\right)^{1/2} \equiv \rho_1 \quad \text{for all } t \ge t_1\,, \end{aligned} \qquad (110)$$

which, in particular, implies

$$|\boldsymbol{v}(\bar{t})|^2_{1,2} \le \rho_1^2\,, \quad \text{for some } \bar{t} \in (t, t+1)\,, \quad \text{all } t \ge t_1\,. \qquad (111)$$

We next integrate $(107)_2$ from $\bar{t}$ to $t+1$ and use (110), to get

$$\begin{aligned} |\boldsymbol{v}(t+1)|^2_{1,2} &\le (\rho_1^2 + F)\mathrm{e}^{\int_t^{t+1} g(\zeta)\,\mathrm{d}\zeta} \\ &\le (\rho_1^2 + F)\mathrm{e}^{\rho^2 \int_t^{t+1} |\boldsymbol{v}(\zeta)|^2_{1,2}\,\mathrm{d}\zeta} \\ &\le (\rho_1^2 + F)\mathrm{e}^{\rho^2\rho_1^2}\,, \quad \text{for all } t \ge t_1\,, \end{aligned}$$

which proves also property (b). □

Remark 18 In the proof of the previous theorem the assumption of the boundedness of Ω is crucial, in order to ensure the validity of Rellich's compactness theorem. However, a different approach, due to Rosa (see Theorem 3.2 in [73]) allows us to draw the same conclusion of Theorem 20 under the more general assumption that in Ω the Poincaré inequality (18) holds. This happens whenever Ω is contained in a strip of finite width (like in a flow in an infinite channel).

We shall now list some further properties of the global attractor $\mathfrak{M}$, for whose proofs we refer to the monographs [26,59,94]

A. Smoothness The restriction of the semigroup S_t to $\mathfrak{M}$ can be extended to a group, $\tilde{S}_t$, defined for all $t \in (-\infty, \infty)$. Therefore, the pair $\{\mathfrak{M}, \tilde{S}_t\}$ constitutes a *flow* (*dynamical system*). This flow is as smooth as allowed by $\boldsymbol{f}$ and Ω. In particular, if $\boldsymbol{f}$ and Ω are of class C^∞, then the solutions to (7)–$(8)_{\mathrm{Hom}}$ belonging to $\mathfrak{M}$ are of class C^∞ in space and time as well. Further significant regularity properties can be found in [23].

B. Finite Dimensionality Let X be a bounded set of a metric space and let $N(X, \varepsilon)$ be the smallest number of balls of radius ε necessary to cover X. The non-negative (possibly infinite) number

$$d_f(X) = \limsup_{\varepsilon \to 0^+} \frac{\ln N(X, \varepsilon)}{\ln(1/\varepsilon)}\,,$$

is called the *fractal dimension* of X. If X is closed with $d_f(X) < \infty$, then there exists a Lipschitz-continuous function, $g : X \mapsto \mathbb{R}^m$, $m > 2d_f(X)$, possessing a Hölder-continuous inverse on $g(X)$ (see Theorem 1.2 in [21]).

The fundamental result states that $d_f(\mathfrak{M})$ is finite and that, moreover,

$$d_f(\mathfrak{M}) \le c\|\boldsymbol{f}\|_2/(\nu^2 C_P) := cG\,, \qquad (112)$$

where c is a positive constant depending only on the "shape" of Ω; see [92]. The quantity G is non-dimensional (often called *Grashof number*). Consequently, $\mathfrak{M}$ is (in particular) homeomorphic to a compact set of $\mathbb{R}^m$, with $m = 2c\,G + 1$. Notice that, in agreement with what conjectured by E. Hopf, (112) gives a rigorous estimate of how the dimension of $\mathfrak{M}$ is expected to increase with the magnitude of the driving force. (Recall, however, that, as remarked previously, there are examples where $d_f(\mathfrak{M}(G)) = 0$, for all $G > 0$.)

Open Question Since $\mathfrak{M}$ can be parametrized by a finite number of parameters, or, equivalently, it can be "smoothly" embedded in a finite-dimensional space, it is a natural question to ask whether or not one can construct a finite-dimensional dynamical system having a global attractor on which the dynamics is "equivalent" to the Navier–Stokes dynamics on $\mathfrak{M}$. This question, which is still *unresolved*, has led to the introduction of the idea of *inertial manifold* [27] and of the associated *approximate inertial manifold* [25], for whose definitions and detailed properties we refer to Chap. VIII in [94]; see, however, also [51].

Further Questions Related to the Existence of the Global Attractor In this section we shall address two further important aspects of the theory of attractors for the Navier–Stokes equations, namely, the three-dimensional case (in a bounded domain) and the case of a flow past an obstacle.

A. Three-Dimensional Flow in a Bounded Domain If we go through the first part of the proof of Theorem 20, we see that the assumption of planar flow has not been used. In fact, by the same token, we can still prove for three-dimensional flow the existence of an "absorbing set", in the

sense that every solution departing from any bounded set of L^2_σ, $\mathcal{A}$, will end up in the set $\mathfrak{B}$ defined in (109), after a time t_0, dependent on $\mathcal{A}$ and F. However, the difficulty in extending the results of Theorem 20 to three-dimensional flow, resides, fundamentally, in the lack of well-posedness of (7)–$(8)_{\text{Hom}}$ in the space $L^2_\sigma(\Omega)$; see Subsect. "Uniqueness and continous Dependence". In order to overcome this situation, several strategies of attack have been proposed. One way is to make an *unproved* assumption on all possible solutions, which guarantees the existence of a semiflow on $L^2_\sigma(\Omega)$; see [Chapter I] in [16]. In such a case, all the fundamental results proven for the 2D flow, continue to hold in 3D as well [16]. Another way is to weaken the definition of global attractor, by requiring the attractivity property in the weak topology of L^2; see [Chapter III.3] in [26], and a third way is to generalize the definition of semiflow in such a way that the uniqueness property is no longer required; see [3,13,79].

B. Flow Past an Obstacle In this case, the relevant initial-boundary value to be investigated is (7)–(8) with $\boldsymbol{f} \equiv \boldsymbol{v}_1 \equiv \boldsymbol{0}$, endowed with the condition at infinity $\lim_{|\boldsymbol{x}|\to\infty} \boldsymbol{v}(\boldsymbol{x}, t) = \boldsymbol{U}$, where $\boldsymbol{U} = U\boldsymbol{e}_1$ is a given, non-zero constant vector. For the reader's convenience, we shall rewrite here this set of equations and put them in a suitable non-dimensional form:

$$\begin{aligned} &\left.\begin{aligned} \frac{\partial \boldsymbol{v}}{\partial t} + \lambda \boldsymbol{v}\cdot\nabla\boldsymbol{v} &= \Delta\boldsymbol{v} - \nabla \mathsf{p} \\ \operatorname{div}\boldsymbol{v} &= 0 \end{aligned}\right\} \quad \text{in}\,\Omega\times(0,\infty) \\ &\boldsymbol{v}(x,t)|_{\partial\Omega} = \boldsymbol{0}\,,\ \lim_{|x|\to\infty} \boldsymbol{v}(x,t) = \boldsymbol{e}_1\,,\ t>0\,; \\ &\boldsymbol{v}(x,0) = \boldsymbol{v}_0(x)\,. \end{aligned} \tag{113}$$

In (113), $\lambda := |U|d/\nu$, with d a length scale, is the appropriate Reynolds number which furnishes the magnitude of the driving mechanism.

As we observed in Remark 16, the least requirement on the spatial domain for the existence of a global attractor in a two-dimensional flow, is that there holds Poincaré's inequality (18). Since this inequality is no longer valid in an exterior domain, the problem of the existence of an attractor for a flow past an obstacle is, basically, unresolved. Actually, the situation is even more complicated than what we just described. In fact, from Theorem 9 and from the considerations developed after it, we know that there is $\lambda_c > 0$ such that if $\lambda < \lambda_c$, the corresponding boundary-value problem has one and only one solution, $(\bar{\boldsymbol{v}}, \bar{p})$, in a suitable function class. Now, it is not known if this solution is attracting. More precisely, *in the two-dimensional flow, it is not known if, for sufficiently small λ, solutions to* (113), *defined in an appropriate class, tend, as $t\to\infty$, to the only corresponding steady solution.*

In the three-dimensional case, the situation is slightly better, but still, the question of the existence of an attractor is completely open. We would like to go into more detail about this point. We begin to observe that there is $\lambda_0 > 0$ such that if $\lambda < \lambda_0$, for any given $\boldsymbol{v}_0$ with $(\boldsymbol{v}_0 - \boldsymbol{e}_1) \in L^3_\sigma(\Omega)$, problem (113) has one and only one (smooth) solution that tends to the (uniquely determined) corresponding steady-state solution, $(\bar{\boldsymbol{v}}, \bar{p})$. In particular

$$\lim_{t\to\infty} \|\boldsymbol{v}(t) - \bar{\boldsymbol{v}}\|_3 = 0\,; \tag{114}$$

see [41]. The fundamental question that stays open is then that of investigating the behavior of solutions to (113) for large t, when $\lambda > \lambda_0$. As a matter of fact, *it is not known whether there exists a norm with respect to which solutions to* (113), *in a suitable class, remain bounded uniformly in time, for all* $\lambda > 0$. In this respect, it is readily seen that, unlike the bounded domain situation, solutions to (113), in general, can not be bounded in $L^2(\Omega)$, uniformly in time, *even when* $\lambda < \lambda_0$. This means that the kinetic energy associated to the motion described by (113) has to grow unbounded for large times. To see this, assume $\lambda < \lambda_0$ and that there exists $K > 0$, independent of t, such that

$$\|\boldsymbol{v}(t) - \boldsymbol{e}_1\|_2 \le K\,, \tag{115}$$

where $\boldsymbol{v}$ is a solution to (113). Then, we can find an unbounded sequence, $\{t_m\}$, and an element $\bar{\boldsymbol{w}} \in L^2_\sigma(\Omega)$ (possibly depending on the sequence) such that

$$\lim_{m\to\infty} (\boldsymbol{v}(t_m) - \boldsymbol{e}_1, \boldsymbol{\varphi}) = (\bar{\boldsymbol{w}}, \boldsymbol{\varphi})\,, \quad \text{for all}\ \boldsymbol{\varphi} \in \mathcal{D}(\Omega)\,. \tag{116}$$

By (114) and (116) we thus must have $\bar{\boldsymbol{w}} = \bar{\boldsymbol{v}} - \boldsymbol{e}_1$, which in turn implies $(\bar{\boldsymbol{v}} - \boldsymbol{e}_1) \in L^2_\sigma(\Omega)$, which, from Theorem 6, we know to be impossible. Consequently, (115) can not be true. Thus, the basic open question is whether or not there exists a function space, Y, where the solution $\boldsymbol{v}(t) - \boldsymbol{e}_1$ remains *uniformly* bounded in $t \in (0,\infty)$, for *all* $\lambda > 0$. (The bound, of course, may depend on λ.) The above considerations along with Theorem 6 suggest then that a plausible candidate for Y is $L^q_\sigma(\Omega)$, for *some* $q > 2$. However, the proof of this property for $q \ge 3$ appears to be overwhelmingly challenging because, in view of Theorem 18 and Remark 8, it would be closely related to the existence of a global, regular solution. Nevertheless, one could investigate the validity of the following weaker property

$$\|\boldsymbol{v}(t) - \boldsymbol{e}_1\|_q \le K_1\,, \quad \text{for some}\ q \in (2,3)\,, \tag{117}$$

where K_1 is independent of $t \in (0, \infty)$. Of course, the requirement is that (117) holds for all $\lambda > 0$ and for all corresponding solutions. It is worth emphasizing that the proof of (117) would be of "no harm" to the outstanding global regularity Problem 2, since, according to the available regularity criteria for weak solutions that we discussed in Sect. "Less Regular Solutions and Partial Regularity Results in 3D", the corresponding solutions, while global in time, will still be weak, even though more regular than those described in Theorem 16. However, notwithstanding its plausibility and "harmlessness", the property (117) appears to be very difficult to establish.

Future Directions

The fundamental open questions that we have pointed out throughout this work constitute as many topics for future investigation. Actually, it is commonly believed that the answer to most of these questions (in the affirmative or in the negative) will probably shed an entirely new light not only on the mathematical theory of the Navier–Stokes equations but also on other disciplines of applied mathematics.

Acknowledgment

This work was partially supported by the National Science Foundation, Grants DMS-0404834 and DMS-0707281.

Bibliography

1. Amick CJ (1984) Existence of Solutions to the Nonhomogeneous Steady Navier–Stokes Equations. Indiana Univ Math J 33:817–830
2. Amick CJ (1988) On Leray's Problem of Steady Navier–Stokes Flow Past a Body in the Plane. Acta Math 161:71–130
3. Ball JM (1997) Continuity Properties and Global Attractors of Generalized Semiflows and the Navier–Stokes Equations. J Nonlinear Sci 7:475–502
4. Batchelor GK (1981) An Introduction to Fluid Mechanics. Cambridge University Press, Cambridge
5. Beirão da Veiga H (1985) On the Construction of Suitable Weak Solutions to the Navier–Stokes Equations via a General Approximation Theorem. J Math Pures Appl 64:321–334
6. Beirão da Veiga H (1995) A New Regularity Class for the Navier–Stokes Equations in R^n. Chinese Ann Math Ser B 16:407–412
7. Berger MS (1977) Nonlinearity and Functional Analysis. Lectures on Nonlinear Problems in Mathematical Analysis. Academic Press, New York
8. Berselli LC, Galdi GP (2002) Regularity Criteria Involving the Pressure for the Weak Solutions to the Navier–Stokes Equations. Proc Amer Math Soc 130:3585–3595
9. Brown RM, Shen Z (1995) Estimates for the Stokes Operator in Lipschitz Domains. Indiana Univ Math J 44:1183–1206
10. Caccioppoli R (1936) Sulle Corrispondenze Funzionali Inverse Diramate: Teoria Generale ed Applicazioni al Problema di Plateau. Rend Accad Lincei 24:258–263; 416–421
11. Caffarelli L, Kohn R, Nirenberg L (1982) Partial Regularity of Suitable Weak Solutions of the Navier–Stokes Equations. Comm Pure Appl Math 35:771–831
12. Cattabriga L (1961) Su un Problema al Contorno Relativo al Sistema di Equazioni di Stokes. Rend Sem Mat Padova 31:308–340
13. Cheskidov A, Foiaş C (2006) On Global Attractors of the 3D Navier–Stokes Equations. J Diff Eq 231:714–754
14. Constantin P (1990) Navier–Stokes Equations and Area of Interfaces. Comm Math Phys 129:241–266
15. Constantin P, Fefferman C (1993) Direction of Vorticity and the Problem of Global Regularity for the Navier–Stokes Equations. Indiana Univ Math J 42:775–789
16. Constantin P, Foiaş C, Temam R (1985) Attractors Representing Turbulent Flows. Mem Amer Math Soc 53 no. 314 vii, pp 67
17. Darrigol O (2002) Between Hydrodynamics and Elasticity Theory: The First Five Births uof the Navier–Stokes Equation. Arch Hist Exact Sci 56:95–150
18. Escauriaza L, Seregin G, Sverák V (2003) Backward Uniqueness for Parabolic Equations. Arch Ration Mech Anal 169:147–157
19. Federer H (1969) Geometric Measure Theory. Die Grundlehren der mathematischen Wissenschaften, vol 153. Springer, New York
20. Finn R, Smith DR (1967) On the stationary solution of the Navier–Stokes equations in two dimensions. Arch Rational Mech Anal 25:26–39
21. Foiaş C, Olson EJ (1996) Finite Fractal Dimension and Hölder–Lipschitz Parametrization. Indiana Univ Math J 45:603–616
22. Foiaş C, Temam R (1977) Structure of the Set of Stationary Solutions of the Navier–Stokes Equations. Comm Pure Appl Math 30:149–164
23. Foiaş C, Temam R (1989) Gevrey Class Regularity for the Solutions of the Navier–Stokes Equations. J Func Anal 87:359–69
24. Foiaş C, Guillopé C, Temam R (1981) New A Priori Estimates for Navier–Stokes Equations in Dimension 3. Comm Partial Diff Equ 6:329–359
25. Foiaş C, Manley OP, Temam R (1988) Modelling of the Interaction of Small and Large Eddies in Two-Dimensional Turbulent Flows. RAIRO Modél Math Anal Numér 22:93–118
26. Foiaş C, Manley OP, Rosa R, Temam R (2001) Navier–Stokes Equations and Turbulence. Encyclopedia of Mathematics and its Applications, 83. Cambridge University Press, Cambridge
27. Foiaş C, Sell GR, Temam R (1988) Inertial Manifolds for Nonlinear Evolutionary Equations. J Diff Eq 73:309–353
28. Fujita H (1961) On the Existence and Regularity of the Steady–State Solutions of the Navier–Stokes Equation. J Fac Sci Univ Tokyo 9:59–102
29. Fujita H (1998) On stationary solutions to Navier–Stokes equation in symmetric plane domains under general outflow condition. In Salvi R (ed) Navier–Stokes Equations: Theory and Numerical Methods. Pitman Res Notes Math Ser 388:16–30
30. Galdi GP (1993) Existence and Uniqueness at Low Reynolds Number of Stationary Plane Flow of a Viscous Fluid in Exterior Domains. In: Galdi GP, Necas J (eds) Recent Developments in Theoretical Fluid Mechanics. Pitman Res Notes Math Ser Longman Sci Tech, Harlow 291:1–33

31. Galdi GP (1998) An Introduction to the Mathematical Theory of the Navier–Stokes Equations. vol I. Linearized Steady Problems. Springer Tracts in Natural Philosophy, 38. Springer, New York (revised Edition)
32. Galdi GP (1998) An Introduction to the Mathematical Theory of the Navier–Stokes Equations. vol II. Nonlinear Steady Problems. Springer Tracts in Natural Philosophy, 39. Springer, New York (revised Edition)
33. Galdi GP (1999) On the Existence of Symmetric Steady-State Solutions to the Plane Exterior Navier–Stokes Problem for Arbitrary Large Reynolds Number. In: Maremonti P (ed) Advances in fluid dynamics. Quad Mat Aracne Rome 4:1–25
34. Galdi GP (2000) An Introduction to the Navier–Stokes Initial-Boundary Value Problem. In: Galdi GP, Heywood JG, Rannacher R (eds) Fundamental Directions in Mathematical Fluid Mechanics. Adv Math Fluid Mech Birkhäuser, Basel 1:1–70
35. Galdi GP (2004) Stationary Navier–Stokes Problem in a Two-Dimensional Exterior Domain. In: Chipot M, Quittner P (eds) Stationary partial differential equations. Handb Diff Equ, North-Holland, Amsterdam 1:71–155
36. Galdi GP (2007) Further Properties of Steady-State Solutions to the Navier–Stokes Problem Past a Threedimensional Obstacle. J Math Phys 48:1–43
37. Galdi GP (2007) Some Mathematical Properties of the Steady-State Navier–Stokes Problem Past a Three-Dimensional Obstacle. RWTH Aachen Institut für Mathematik, Report no. 17
38. Galdi GP, Maremonti P (1988) Regularity of Weak Solutions of the Navier–Stokes System in Arbitrary Domains. Ann Univ Ferrara Sez VII 34:59–73
39. Galdi GP, Padula M (1990) A new approach to energy theory in the stability of fluid motion. Arch Rational Mech Anal 110:187–286
40. Galdi GP, Rabier PJ (1999) Functional Properties of the Navier–Stokes Operator and Bifurcation of Stationary Solutions: Planar Exterior Domains. In: Escher J, Simonett G (eds) Topics in Nonlinear Analysis. Progr Nonlinear Diff Equ Appl. Birkhäuser, Basel 35:273–303
41. Galdi GP, Heywood JG, Shibata Y (1997) On the Global Existence and Convergence to Steady State of Navier–Stokes Flow Past an Obstacle that is Started from Rest. Arch Rational Mech Anal 138:307–318
42. Galdi GP, Robertson AM, Rannacher R, Turek S (2007) Hemodynamical Flows: Modeling, Analysis and Simulation. Oberwolfach Seminar Series vol 35. Birkhäuser
43. Giga Y, Sohr H (1991) Abstract L^p Estimates for the Cauchy Problem with Applications to the Navier–Stokes Equations in Exterior Domains. J Funct Anal 102:72–94
44. Gilbarg D, Weinberger HF (1974) Asymptotic Properties of Leray's Solution of the Stationary Two-Dimensional Navier–Stokes Equations. Russian Math Surveys 29:109–123
45. Gilbarg D, Weinberger HF (1978) Asymptotic Properties of Steady Plane Solutions of the Navier–Stokes Equations with Bounded Dirichlet. Integral Ann Scuola Norm Sup Pisa 5:381–404
46. Gohberg I, Goldberg S, Kaashoek MA (1990) Classes of Linear Operators:I. Operator Theory. Advances and Applications. vol49. Birkhäuser, Basel
47. Guermond JL (2007) Faedo–Galerkin Weak Solutions of the Navier–Stokes Equations with Dirichlet Boundary Conditions are Suitable. J Math Pures Appl 88:87–106
48. Gustafson S, Kang K, Tsai TP (2006) Regularity Criteria for Suitable Weak Solutions of the Navier–Stokes Equations Near the Boundary. J Diff Equ 226:594–618
49. Gustafson S, Kang K, Tsai TP (2007) Interior Regularity Criteria for Suitable Weak Solutions of the Navier–Stokes Equations. Comm Math Phys 273:161–176
50. Heywood JG (1980) The Navier–Stokes Equations: on the Existence, Regularity and Decay of Solutions. Indiana Univ Math J 29:639–681
51. Heywood JG, Rannacher R (1993) On the Question of Turbulence Modeling by Approximate Inertial Manifolds and the Nonlinear Galerkin Method. SIAM J Numer Anal 30:1603–1621
52. Hopf E (1941) Ein Allgemeiner Endlichkeitsatz der Hydrodynamik. Math Annalen 117:764–775
53. Hopf E (1948) A Mathematical Example Displaying Features of Turbulence. Comm Pure Appl Math 1:303–322
54. Hopf E (1951) Über die Anfangswertaufgabe für die hydrodynamischen Grundgleichungen. Math Nachr 4:213–231
55. Kozono H, Sohr H (2000) Remark on Uniqueness of Weak Solutions to the Navier–Stokes Equations. Analysis 16:255–271
56. Kozono H, Taniuchi Y (2000) Bilinear Estimates in BMO and the Navier–Stokes Equations. Math Z 235:173–94
57. Kozono H, Yatsu N (2004) Extension Criterion via Two-Components of Vorticity on Strong Solutions to the 3D Navier–Stokes Equations. Math Z 246:55–68
58. Ladyzhenskaya OA (1963) The mathematical theory of viscous incompressible flow. Revised English edition. Gordon and Breach Science, New York-London
59. Ladyzhenskaya OA (1991) Attractors for Semigroups and Evolution Equations. Lezioni Lincee. Cambridge University Press, Cambridge
60. Ladyzhenskaya OA, Seregin GA (1999) On Partial Regularity of Suitable Weak Solutions to the Three-Dimensional Navier–Stokes Equations. J Math Fluid Mech 1:356–387
61. Leray J (1933) Etude de Diverses Òquations Intégrales non Linéaires et de Quelques Problèmes que Pose l' Hydrodynamique. J Math Pures Appl 12:1–82
62. Leray J (1934) Sur le Mouvement d'un Liquide Visqueux Emplissant l'Espace. Acta Math 63:193–248
63. Leray J (1936) Les Problémes non Linéaires. Enseignement Math 35:139–151
64. Lin F (1998) A New Proof of the Caffarelli–Kohn-Nirenberg Theorem. Comm. Pure Appl Math 51:241–257
65. Marchioro C (1986) An Example of Absence of Turbulence for any Reynolds Number. Comm Math Phys 105:99–106
66. Maremonti P (1998) Some Interpolation Inequalities Involving Stokes Operator and First Order Derivatives. Ann Mat Pura Appl 175:59–91
67. Málek J, Padula M, Ruzicka M (1995) A Note on Derivative Estimates for a Hopf Solution to the Navier–Stokes System in a Three-Dimensional Cube. In: Sequeira A (ed) Navier–Stokes Equations and Related Nonlinear Problems. Plenum, New York 141–146
68. Natarajan R, Acrivos A (1993) The Instability of the Steady Flow Past Spheres and Disks. J Fluid Mech 254:323–342
69. Neustupa J, Novotný A, Penel P (2002) An Interior Regularity of a Weak Solution to the Navier–Stokes Equations in Dependence on one Component of Velocity. In: Galdi GP, Rannacher R (eds) Topics in mathematical fluid mechanics. Quad Mat Aracne, Rome 10:163–183

70. Nirenberg L (1959) On Elliptic Partial Differential Equations. Ann Scuola Norm Sup Pisa 13:115–162
71. Prodi G (1959) Un Teorema di Unicità per le Equazioni di Navier–Stokes. Ann Mat Pura Appl 48:173–182
72. Prodi G (1962) Teoremi di Tipo Locale per il Sistema di Navier–Stokes e Stabilità delle Soluzioni Stazionarie. Rend Sem Mat Univ Padova 32:374–397
73. Rosa R (1998) The Global Attractor for the 2D Navier–Stokes Flow on Some Unbounded Domains. Nonlinear Anal 32:71–85
74. Ruelle D, Takens F (1971) On the Nature of Turbulence. Comm Math Phys 20:167–192
75. Sather J (1963) The Initial Boundary Value Problem for the Navier–Stokes Equations in Regions with Moving Boundaries. Ph.D. thesis, University of Minnesota
76. Scheffer V (1976) Partial Regularity of Solutions to the Navier–Stokes Equations. Pacific J Math 66:535–552
77. Scheffer V (1977) Hausdorff Measure and the Navier–Stokes Equations. Comm Math Phys 55:97–112
78. Scheffer V (1980) The Navier–Stokes Equations on a Bounded Domain. Comm Math Phys 73:1–42
79. Sell GR (1996) Global Attractors for the Three-Dimensional Navier-Stokes Equations. J Dynam Diff Eq 8:1–33
80. Seregin G, Sverák V (2002) The Navier–Stokes Equations and Backward Uniqueness. In: Birman MS, Hildebrandt S, Solonnikov VA, Uraltseva NN (eds) Nonlinear problems in mathematical physics and related topics, II. Int Math Ser (N. Y.) Kluwer/Plenum, New York 2:353–366
81. Seregin GA, Shilkin TN, Solonnikov VA (2006) Boundary Partial Regularity for the Navier–Stokes Equations. J Math Sci 132:339–358
82. Serrin JB (1962) On the Interior Regularity of Weak Solutions of the Navier–Stokes Equations. Arch Rational Mech Anal 9:187–195
83. Serrin JB (1963) The Initial Value Problem for the Navier–Stokes Equations. In: Langer RE (ed) Nonlinear Problems. University of Wisconsin Press, pp. 69–98
84. Smale S (1965) An Infinite Dimensional Version of Sard's Theorem. Amer J Math 87:861–866
85. Smale S (1967) Differentiable Dynamical Systems. Bull Amer Math Soc 73:747–817
86. Sohr H (2001) The Navier–Stokes Equations. An Elementary Functional Analytic Approach. Birkhäuser, Basel
87. Sohr H, von Wahl W (1984) On the Singular Set and the Uniqueness of Weak Solutions of the Navier–Stokes Equations. Manuscripta Math 49:27–59
88. Stein EM (1970) Singular Integrals and Differentiability Properties of Functions. Princeton University Press, Princeton
89. Stokes GG (1851) On the Effect of the Internal Friction of Fluids on the Motion of Pendulums. Trans Cambridge Phil Soc 9:8–106
90. Takeshita A (1993) A Remark on Leray's Inequality. Pacific J Math 157:151–158
91. Taneda S (1956) Experimental Investigation of the Wake Behind a Sphere at Low Reynolds Numbers. J Phys Soc Japan 11:1104–1111
92. Temam R (1986) Infinite-Dimensional Dynamical Systems in Fluid Mechanics. In: Browder F (ed) Nonlinear Functional Analysis and its Applications, Part 2. Proc. Sympos. Pure Math Amer. Math Soc. Providence, RI 45:431–445
93. Temam R (1995) Navier–Stokes Equations and Nonlinear Functional Analysis. CBMS-NSF Regional Conference Series in Applied Mathematics, 66
94. Temam R (1997) Infinite-Dimensional Dynamical Systems in Mechanics and Physics. Second edition. Applied Mathematical Sciences, 68. Springer, New York
95. Tomboulides AG, Orszag SA (2000) Numerical Investigation of Transitional and Weak Turbulent Flow Past a Sphere. J Fluid Mech 416:45–73
96. Vorovich II, Youdovic VI (1961) Stationary Flow of a Viscous Incompressible Fluid. Mat Sb 53:393–428 (in Russian)
97. Wu JS, Faeth GM (1993) Sphere Wakes in Still Surroundings at Intermediate Reynolds Numbers. AIAA J 31:1448–1460
98. Xie W (1997) Sharp Sobolev Interpolation Inequalities for the Stokes Operator. Diff Inte Equ 10:393–399
99. Zeidler E (1988) Nonlinear Functional Analysis and Applications: Application to Mathematical Physics. Springer, New York
100. Zeidler E (1995) Applied Functional Analysis: Main Principles and their Applications. Applied Math Sci., Springer, vol 109

n-Body Problem and Choreographies

Susanna Terracini
Dipartimento di Matematica e Applicazioni,
Università di Milano Bicocca, Milano, Italia

Article Outline

Glossary
Definition of the Subject
Introduction
Simple Choreographies and Relative Equilibria
Symmetry Groups and Equivariant Orbits
The 3-Body Problem
Minimizing Properties of Simple Choreographies
Generalized Orbits and Singularities
Asymptotic Estimates at Collisions
Absence of Collision for Locally Minimal Paths
Future Directions
Bibliography

Glossary

Central configurations Are the critical points of the potential constrained on the unitary moment of inertia ellipsoid. Central configurations are associated to particular solutions to the *n*-body problem: the *relative equilibrium* and the *homographic* motions defined in Definition 1.

Choreographical solution A choreographical solution of the *n*-body problem is a solution such that the particles move on the same curve, exchanging their positions after a fixed time. This property can be re-

garded as a symmetry of the trajectory. This notion finds a natural generalization in that of *G-equivariant trajectory* defined in Definition 2 for a given group of symmetries-G. The *G-equivariant minimization technique* consists in seeking action minimizing trajectories among all G-equivariant paths.

Collision and singularities When a trajectory can not be extended beyond a certain time b we say that a *singularity* occurs. Singularities can be *collisions* if the solution admits a limit configuration as $t \to b$. In such a case we term b a *collision instant*.

***n*-Body problem** The n-body problem is the system of differential equations (1) associated with suitable initial or boundary value data. A *solution* or *trajectory* is a doubly differentiable path $q(t) = (q_1(t), \ldots, q_n(t))$ satisfying (1) for all t. The weaker notion of *generalized solution* is defined in Definition 5 applies to trajectories found by variational methods.

Variational approach The variational approach to the n-body problem consists in looking at trajectories as critical points of the *action functional* defined in (4). Such critical points can be (local) *minimizers*, or *constrained minimizers* or *mountain pass*, or other type.

Definition of the Subject

The motion of n-point particles of positions $x_i(t) \in \mathbb{R}^3$ and masses $m_i > 0$, interacting in accordance with Newton's law of gravitation, satisfies the system of *differential equations*:

$$-m_i \ddot{x}_i(t) = G \sum_{j \neq i, j=1}^{n} m_i m_j \frac{x_i - x_j}{|x_i - x_j|^3}\,, \qquad i = 1, \ldots, n\,, \quad t \in \mathbb{R}\,. \tag{1}$$

The *n-body problem* consists in solving Eq. (1) associated with initial or boundary conditions.

A *simple choreography* is a periodic solution to the n-body problem Eq. (1) where the bodies lie on the same curve and exchange their mutual positions after a fixed time, namely, there exists a function $x : \mathbb{R} \to \mathbb{R}$ such that

$$x_i(t) = x(t + (i-1)\tau)\,, \quad i = 1, \ldots, n\,, \quad t \in \mathbb{R}\,, \tag{2}$$

where $\tau = 2\pi/n$.

Introduction

The *two-body problem* can be reduced, by the conservation of the linear momentum, to the *one center* Kepler problem and can be completely solved either by exploiting the conservation laws (angular momentum, energy and the Lenz vector), or by performing the Levi–Civita change of coordinates reducing the problem to that of an harmonic oscillator [58].

The three-body problem is much more complicated than the two-body and can not be solved in a simple way. A major study of the Earth–Moon–Sun system was undertaken by Delaunay in his *La Théorie du mouvement de la lune*. In the *restricted three-body problem*, the mass of one of the bodies is negligible; the circular restricted three-body problem is the special case in which two of the bodies are in circular orbits and was worked on extensively by many famous mathematicians and physicists, notably Lagrange in the 18th century, Poincaré at the end of the 19th century and Moser in the 20th century. Poincaré's work on the restricted three-body problem was the foundation of deterministic chaos theory.

A very basic – still fundamental – question concerns the real number of degrees of freedom of the n-body problem. As the motion of each point particle is represented by 3-dimensional vectors, the n-body problem has $3n$ degrees of freedom and hence it is $6n$-dimensional. First integrals are: the center of mass, the linear momentum, the angular momentum, and the energy. Hence there are 10 independent algebraic integrals. This allows the reduction of variables to $6n - 10$. It was proved in 1887 by Bruns that these are the only linearly independent integrals of the n-body problem, which are algebraic with respect to phase and time variables. This theorem was later generalized by Poincaré.

A second very natural problem is whether there exists a power series expressing – if not every – at least a large and relevant class of trajectories for the n-body problem. In 1912, after pioneering works of Mittag–Leffler and Levi–Civita, Sundman proved the existence of a series solution in powers of $t^{1/3}$ for the 3-body problem. This series is convergent for all real t, except for those initial data which correspond to *vanishing angular momentum*. These initial data have Lebesgue measure zero and therefore are not generic. An important issue in proving this result is the fact that the radius of convergence for this series is determined by the distance to the nearest singularity. Since then, the study of singularities became the main point of interest in the study of the n-body problem. Sundman's result was later generalised to the case of $n > 3$ bodies by Q. Wang in [88]. However, the rate and domain of convergence of this series are so limited to make it hardly applicable to practical and theoretical purposes.

Finally, the n-body problem can be faced from the point of view of the theory of perturbations and represents

both its starting point and its most relevant application. Delaunay and Poincaré described the spatial three-body problem as a four-dimensional Hamiltonian system and already encountered some trajectories featuring a chaotic behavior. When one of the bodies is much heavier than the other two (a system of one "star" and two "planets"), one can neglect the interaction between the small planets; hence the system can be seen as a perturbation of two decoupled two-body problems whose motions are known to be all periodic from Kepler's laws. In the planetary three-body problem, hence, two harmonic oscillators interact nonlinearly through the perturbation. Resonances between the two oscillators can be held responsible of high sensitivity with respect to initial data and other chaotic features. A natural question regards the coexistence of the irregular trajectories with regular (periodic or quasi-periodic) ones. A modern approach to the problem of stability of solutions to nearly integrable systems goes through the application of Kolmogorov–Arnold–Moser (KAM) Theorem [8,56], whose main object indeed is indeed the persistence, under perturbations, of invariant tori.

The *n*-body problem is paradigmatic of any complex system of many interacting objects, and it can neither be solved nor it can be simplified in an efficient way. A possible starting point for its analysis is to seek selected trajectories whose motion is particularly simple, in the sense that it repeats after a fixed period: the periodic solutions. Following Poincaré in his *Méthodes Nouvelles de la Mécanique Céleste, tome I, 1892*,

> "… D'ailleurs, ce qui nous rends ces solutions périodiques si précieuses, c'est qu'elles sont, pour ainsi dire, la seule brèche par où nous puissons essayer de pénetrer dans une place jusqu'ici réputée inabordable."

Indeed, just before, Poincaré conjectured that periodic trajectories are dense in the phase space:

> "… Voici un fait que je n'ai pas pu démontrer rigourousement, mais qui me parait pourtant très vraisemblable. Étant donneées des équations de la forme définie dans le n.13[1] et une solution quelconque de ces équations, on peut toujours trouver une solution périodique (dont la période peut, il est vrai, être trés longue), telle que la différence entre les deux solutions soit aussi petite qu'on le veut, pendant un temps aussi long qu'on le veut."

[1]Formula N. 13 quoted by Poincaré is Hamilton equation and covers our class od Dynamical Systems Eq. (3)

Singular Hamiltonian Systems

From an abstract point of view, the *n*-body problem is a Hamiltonian System of the form

$$m_i \ddot{x}_i = \frac{\partial U}{\partial x_i}(t, x)\,, \quad i = 1, \ldots, n\,, \tag{3}$$

where the forces $\frac{\partial U}{\partial x_i}$ are undefined on a singular set Δ, *the set of collisions between two or more particles in the n-body problem*. Such singularities play a fundamental role in the phase portrait (see, e. g. [43]) and strongly influence the global orbit structure, as they can be held responsible, among others, of the presence of *chaotic motions* (see, e. g. [39]) and of *motions becoming unbounded in a finite time* [64,90].

Two are the major steps in the analysis of the impact of the singularities in the *n*-body problem: the first consists in performing the asymptotic analysis along a single collision (total or partial) trajectory and goes back, in the classical case, to the works by Sundman ([84]), Wintner ([89]) and, in more recent years by Sperling, Pollard, Saari and other authors (see for instance [40,47,75,76,79,83]). The second step consists in blowing-up the singularity by a suitable change of coordinates introduced by McGehee in [65] and replacing it by an invariant boundary – the collision manifold – where the flow can be extended in a smooth manner. It turns out that, in many interesting applications, the flow on the collision manifold has a simple structure: it is a gradient-like, Morse–Smale flow featuring a few stationary points and heteroclinic connections (see, for instance, the surveys [39,68]). The analysis of the extended flow allows us to obtain a full picture of the behavior of solutions near the singularity, despite the flow fails to be fully regularizable (except for binary collisions).

Simple Choreographies and Relative Equilibria

A possible starting point for the study of the *n*-body problem is to find selected trajectories which are particularly simple, when regarded from some point of view. Examples of such particular solutions are the collinear periodic orbits (found in 1767 by Euler), in which three bodies of any masses move such that they oscillate along a rotation line, and the Lagrange triangular solutions, where the bodies lie at the vertices of a rotating equilateral triangle that shrinks and expands periodically, discovered in 1772. Both these trajectories are stationary in a rotating frame. A second remarkable class of trajectories – the choreographies – can be found when the masses are all equal, by exploiting the fact that particles can be interchanged without changing the structure of the system.

Among all periodic solutions of the planar 3-body problem, the relative equilibrium motions – the equilateral Lagrange and the collinear Euler–Moulton solutions – are definitely the simplest and most known. In general such simple periodic motions exist for any number of bodies.

Definition 1 A *relative equilibrium* trajectory is a solution of (1) whose configuration remains constant in a rotating frame. A *homographic* trajectory is a solution of (1) whose configuration remains constant up to homotheties.

The normalized configurations of such trajectories are named *central* and are the critical points of the potential

$$U(x) = \sum_{i<j} \frac{m_i m_j}{|x_i - x_j|} ,$$

constrained to the ellipsoid of unitary momentum of inertia:

$$I(x) = \sum_i m_i |x_i|^2 = 1 .$$

Relative equilibria feature an evident symmetry ($SO(2)$ and $O(2)$ respectively), that is, they are equivariant with respect to the symmetry group of dimension 1 acting as $SO(2)$ (resp. $O(2)$) on the time circle and on the plane, and trivially on the set of indexes $\{1, 2, 3\}$. In fact, they are minimizers of the Lagrangian action functional in the space of all loops having their same symmetry group. Therefore, generally speaking, *G-equivariant minimizers for the action functional (given a symmetry group G) can be thought as the natural generalization of relative equilibrium motions.*

This perspective has known a wide popularity in the recent literature and has produced a new boost in the study of periodic trajectories in the n-body problem; the recent proof of the existence of the Chenciner–Montgomery eight-shaped orbit is emblematic of this renewed interest (see [7,20,22,25,29,52] and the major part of our bibliographical references). In all these papers, periodic and quasi-periodic solutions of the n-body problem can be found as critical points of the Lagrangian action functional restricted to suitable spaces of symmetric paths.

Basic Definitions and Notations

Let us consider n point particles with masses $m_1, m_2, \ldots, m_n$ and positions $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$, with $d \geq 2$. We denote by $\mathcal{X}$ the space of configurations with center of mass in 0, and by $\hat{\mathcal{X}} = \mathcal{X} \backslash \Delta$ the set of collision-free configurations (collision means $x_i = x_j$ for some $i \neq j$). On the configuration space we define the homogeneous (Newton) potential of degree $-\alpha < 0$:

$$U(x) = \sum_{i<j} U_{i,j}(|x_i - x_j|), \quad U_{i,j}(|x_i - x_j|) \simeq \frac{m_i m_j}{|x_i - x_j|^\alpha} .$$

In many cases one can simply require the $U_{i,j}$ to be asymptotically homogeneous at the singularity. Furthermore, a major part of our analysis can be extended to logarithmic potentials. Relative equilibria correspond to those configurations (termed central) which are critical for the restriction of the potential U to the ellipsoid $I = \lambda$ where I denotes the momentum of inertia:

$$I(x) = \sum_i m_i |x_i|^2 .$$

On collisions the potential $U = +\infty$. We are interested in (relative) periodic (such that $\forall t\colon x(t+T) = x(t)$) solutions to the system of differential equations:

$$m_i \ddot{x}_i = \frac{\partial U}{\partial x_i} .$$

We associate with the equation the Lagrangian integrand

$$L(x, \dot{x}) = L = K + U = \sum_i \frac{1}{2} m_i |\dot{x}_i|^2 + \sum_{i<j} U_{i,j}(|x_i - x_j|)$$

and the action functional:

$$\mathcal{A}(x) = \int_0^T L(x(t), \dot{x}(t)) \mathrm{d}t . \tag{4}$$

Sometimes it will be preferable to consider the problem in a frame rotating uniformly about the vertical axis, with an angular speed ω; the corresponding action $\mathcal{A}^\omega$ then contains a gyroscopic term, associated with Coriolis force.

We shall seek periodic solutions as critical points of the action functional on the Sobolev space of T-periodic trajectories: $\Lambda = H^1(\mathbb{T}, \mathcal{X})$ or, to be more precise, of the action constrained on suitable linear subspaces $\Lambda_0 \subset \Lambda$.

Two are the major difficulties to be faced in following the variational approach; the first is due to the *lack of coercivity* (or of Palais–Smale) due to the vanishing at infinity of the force fields: indeed sequences of almost-critical points (such as minimizing sequences) may very well diverge. Furthermore, as the potential U is singular on collisions, minimizers or other critical points can a priori be *collision trajectories*. Compactness can be successfully recovered by the symmetry of the problem, as we are going to explain here below. We shall expose in the last part of this article some of the strategies which has been developed in order to overcome the problem of collisions.

Symmetry Groups and Equivariant Orbits

We can generalize the concept of relative equilibria as follows: we impose the permutation of the positions of the particles, after a given time, up to isometries of the space. This gives rise to a class of symmetric trajectories (the generalized choreoghapies). It is worthwhile noticing that these trajectories arise as a common feature of any system of interacting objects, regardless whether they come from models in Celestial mechanics or not. It applies to atoms, or molecules, galaxies or whatever system, provided the objects can be freely interchanged. Let us start by introducing some basics concepts and definitions from [52]. Let G be a finite group endowed with:

- an orthogonal representation of dimension 2 $\tau\colon G \to O(2)$ (on cyclic time $\mathbb{T} \cong S^1$),
- an orthogonal representation (on the euclidean space $\mathbb{R}^d$) $\rho\colon G \to O(d)$,
- and an homomorphism on the symmetric group on n elements ($\mathbf{n} = \{1, 2, \ldots, n\}$) $\sigma\colon G \to \Sigma_n$.

Then G acts on time (translation and reversal) $\mathbb{T}$ via τ and it acts on the configuration space $\mathcal{X}$ via ρ and σ in the following way

$$\forall i = 1 \ldots n\colon (gx)_i = \rho(g) x_{\sigma(g)^{-1}(i)} \,.$$

As a consequence we have an action on the space of trajectories:

Definition 2 A continuous function $x(t) = (x_1(t), \ldots, x_n(t))$ is *G-equivariant* if

$$\forall g \in G\colon x(gt) = (gx)(t) \,.$$

The linear subspace $\Lambda_0 = \Lambda^G \subset \Lambda$ denotes the set of periodic trajectories in Λ which are equivariant with respect to the G-action:

The Palais Principle of symmetric criticality [73] ensures that critical points of an invariant functional restricted to the space of equivariant trajectories are indeed free critical points. We stress that by the particular form of the interaction potentials, in our setting invariance is simply implied by equality of those masses which are interchanged by the action of G on the set of the indices. When the action of σ is transitive, all the masses must be equal and the associated G–equavariant trajectories give rise to *generalized choreographies*.

Cyclic and Dihedral Actions

Consider the normal subgroup $\ker\tau \triangleleft G$ and the quotient $\bar{G} = G/\ker\tau$. Since $\bar{G}$ acts effectively on $\mathbb{T}$, it is either a *cyclic* group or a *dihedral* group.

- If the group $\bar{G}$ acts trivially on the orientation of $\mathbb{T}$, then $\bar{G}$ is cyclic and we say that the action of G on Λ is of *cyclic type*.

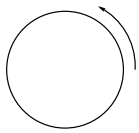

- If the group $\bar{G}$ consists of a single reflection on $\mathbb{T}$, then we say that action of G on Λ is of *brake type*.

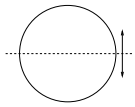

- Otherwise, we say that the action of G on Λ is of *dihedral type*.

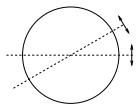

With this first classification, we can easily associate to our minimization problem some proper boundary conditions.

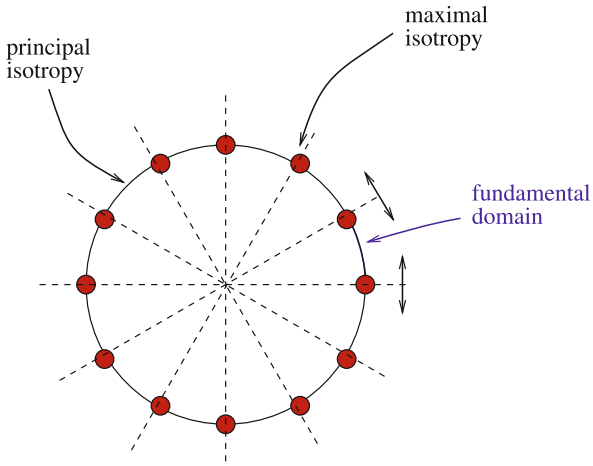

Proposition 1 *Let $\mathbb{I}$ be the fundamental domain (for a dihedral type), or any interval having as length the minimal angle of time-rotation in $\bar{G}$ (for a cyclic type). Then the G-equivariant minimization problem is equivalent to problem of minimizing the action over all paths $x\colon \mathbb{I} \to \mathcal{X}^{\ker\tau}$ subject to the boundary conditions $x(0) \in \mathcal{X}^{H_0}$ and $x(1) \in \mathcal{X}^{H_1}$, where H_0 and H_1 are the maximal isotropy subgroups of the boundary of $\mathbb{I}$.*

The Variational Approach

Let us consider the $\mathcal{A}$ restricted to the space of symmetric loops Λ^G. We recall that the action functional is said

to be coercive if $\lim_{|x|\to+\infty} \mathcal{A}(x) = +\infty$. Coercivity implies the validity of the direct method of Calculus of Variations and, consequently, the existence of a minimizer.

Proposition 2 *The action functional $\mathcal{A}$ is coercive in Λ^G if and only if $\mathcal{X}^G = 0$. Consequently, if $\mathcal{X}^G = 0$ then a minimizer of $\mathcal{A}^G$ in Λ^G exists.*

Given ρ and σ, we can compute $\dim\mathcal{X}^G$

$$\dim\mathcal{X}^G = \frac{1}{|G|}\sum_{g\in G} \mathrm{Tr}(\rho(g))\#\mathrm{Fix}(\sigma(g)) - d\,.$$

In the frame rotating with constant angular speed ω the action $\mathcal{A}^\omega$ is generally coercive, except for a (possible) discrete set values of ω.

Of course, beside seeking minimizers, one can look for other type of critical points, such as mountain pass or others. As the potential U is singular on collisions, minimizers (or other critical points) can a priori be *collision trajectories*. Many strategies were proposed in the literature in order to overcome these obstacles. The development of a suitable Critical Point theory taking into account of the contribution of fake periodic solutions (the critical points at infinity) was proposed by some authors [9,60,77] and returned a good estimate of the number of periodic trajectories satisfying an appropriate bound on the length, with the main disadvantage of requiring a strong order of infinity at the collisions (the *strong force condition*). Another strategy to recover coercivity of action functional consists in imposing a symmetry constraint on the loop space. Surprisingly enough, once coercivity is recovered, also the problem of collisions becomes much less dramatic. This fact was remarked for the first time in [34] and widely exploited in the literature, also thanks to the neat idea, due to C. Marchal, of averaging over all possible variations (generalized and exposed in Sect. "Generalized Orbits and Singularities") to avoid the occurrence of collisions for extremals of the action (Marchal idea was first exposed in [25]). This argument can be used in most of the known cases to prove that absence of collisions for minimizing trajectories and will be outlined in the last section.

The Eight Shaped Three-Body Solution

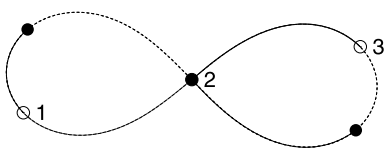

In their paper [29], Chenciner and Montomery exploited a variational argument in order to prove the existence of a periodic trajectory for the three body problem, where the three particles move one a single eight-shaped curve, interchanging their positions after a fixed time. C. Moore [71] was the first to find numerically the eight, lead by topological reasons which turned out to be insufficient to insure its existence. One of the the simplest symmetries giving rise to the eight shaped trajectory is the following. Denote

$$x(t) = (x_1(t), x_2(t), x_3(t)) \in \mathbb{R}^6\,.$$

Let $G = D_6$ be the dihedral group generated by two following reflections: the first

$$x_1(-t) = -x_1(t)\,, \quad x_2(-t) = -x_3(t)\,,$$
$$x_3(-t) = -x_2(t)\,,$$

with $g_1 = (\tau_1, \rho_1, \sigma_1)$ is $\tau_1(t) = -t$, $\rho_1(x) = -x$ and $\sigma_1(1,2,3) = (1,3,2)$. And the second

$$x_1(1-t) = -x_2(t)\,, \quad x_2(1-t) = -x_1(t)\,,$$
$$x_3(1-t) = -x_3(t)\,,$$

with $g_2 = (\tau_2, \rho_2, \sigma_2)$ is $\tau_2(t) = 1-t$, $\rho_2(x) = -x$ and $\sigma_2(1,2,3) = (2,1,3)$.

There are three possible groups yielding an eight-shaped trajectory. First, we consider the group of cyclic action type C_6 (the *cyclic eight* having order 6), which acts cyclically on $\mathbb{T}$ (i. e. by a rotation of angle $\pi/3$), by a reflection in the plane E, and by the cyclic permutation $(1,2,3)$ in the index set.

The second group, which we denote by D_{12}, is the group of order 12 obtained by extending C_6 with the element h defined as follows: $\tau(h)$ is a reflection in $\mathbb{T}$, $\rho(h)$ is the antipodal map in E (thus, the rotation of angle π), and $\sigma(h)$ is the permutation $(1,2)$. This is the symmetry group used by Chenciner and Montgomery.

The third group is the subgroup of D_{12} generated by h and the subgroup C_3 of order 3 of $C_6 \subset D_{12}$. We denote this group D_6 (since it is a dihedral group of order 6). The symmetry groups D_{12} and D_6 are of dihedral type. The choreography group C_3 is a subgroup of all the three groups, thus the action is coercive on G-equivariant loops.

The Rotating Circle Property (RCP)

As a second step of the variational approach to the n-body problem, one has to prove that the output of the minimization (or some other variational method) procedure is free of collisions. This point involves a deep analysis of the structure of the possible singularities and will be outlined in the last part of this article. After performing the analysis of collisions, we find that *if the action of G on $\mathbb{T}$ and*

$\mathcal{X}$ fulfills some conditions (computable) then (local) minimizers of the action functional $\mathcal{A}$ in $\Lambda^G \subset \Lambda$ do not have collisions.

For a group H acting orthogonally on $\mathbb{R}^d$, a circle $\mathbb{S} \subset \mathbb{R}^d$ (with center in 0) is termed *rotating under H* if $\mathbb{S}$ is *invariant* under H (that is, for every $g \in H$ $g\mathbb{S} = \mathbb{S}$) and for every $g \in H$ the restriction $g|\mathbb{S}\colon \mathbb{S} \to \mathbb{S}$ is a *rotation* (the identity is meant as a~rotation of angle 0).

Let $i \in \mathbf{n}$ be an index and $H \subset G$ a subgroup. A circle $\mathbb{S} \subset \mathbb{R}^d = V$ (with center in 0) is termed *rotating for i under H* if $\mathbb{S}$ is *rotating under H* and

$$\mathbb{S} \subset V^{H_i} \subset V = \mathbb{R}^d ,$$

where $H_i \subset H$ denotes the *isotropy subgroup* of the index i in H relative to the action of H on the index set $\mathbf{n}$ induced by restriction (that is, the isotropy $H_i = \{g \in H | gi = i\}$).

Definition 3 A group G acts with the *rotating circle property* if for every $\mathbb{T}$-isotropy subgroup $G_t \subset G$ and for at least $n-1$ indexes $i \in \mathbf{n}$ there exists in $\mathbb{R}^d$ a rotating circle $\mathbb{S}$ under G_t for i.

In most of the known examples the property is fulfilled. In [52] the following results were proved.

Theorem 1 *Consider a finite group K acting on Λ with the rotating circle property. Then a minimizer of the K-equivariant fixed-ends (Bolza) problem is free of collisions.*

Corollary 1 *For every $\alpha > 0$, minimizers of the fixed-ends (Bolza) problem are free of interior collisions.*

Corollary 2 *If the action of G on Λ is of cyclic type and $\ker\tau$ has the rotating circle property then any local minimizer of $\mathcal{A}^G$ in Λ^G is collisionless.*

Corollary 3 *If the action of G on Λ is of cyclic type and $\ker\tau = 1$ is trivial then any local minimizer of $\mathcal{A}^G$ in Λ^G is collisionless.*

Theorem 2 *Consider a finite group G acting on Λ so that every maximal $\mathbb{T}$-isotropy subgroup of G either has the rotating circle property or acts trivially on the index set n. Then any local minimizer of $\mathcal{A}^G$ yields a collision-free periodic solution of the Newton equations for the n-body problem in $\mathbb{R}^d$.*

Examples

In this section, for the sake of illustrating the power and the limitations of the approach through the G-equivariant minimization, we include a few examples fitting in our theoretical framework and we give some hints of which symmetry groups satisfy or not assumptions of Theorem 2. Well-known examples are the celebrated Chenciner–Montgomery "eight" [29], Chenciner–Venturelli "Hip–Hop" solutions [30], Chenciner "generalized Hip–Hops" solutions [26], Chen's orbit [20,22] and Terracini–Venturelli generalized Hip–Hops [85]. One word about the pictures of planar orbits: the configurations at the boundary points of the fundamental domain $\mathbb{I}$ are denoted with an empty circle (starting point $x_i(0)$) and a black disc (ending point $x_i(t)$, with t appropriate), with a label on the starting point describing the index of the particle. The trajectories of the particles with the times in $\mathbb{I}$ are painted as thicker lines (thus it is possible to recover the direction of the movement from $x_i(0)$ to $x_i(t)$). Unfortunately this feature was not possible with the three-dimensional images.

Also, in all the following examples but 4 and 5 existence of the orbits follows directly from the results of Theorem 2. The existence of the orbits described in Examples 4 and 5, which goes beyond the scope of this article, has been recently proved by Chen in [22]. Thousands of other suitable actions and the corresponding orbits have been found by a special-purpose computer program based on GAP [53].

Example 1 (Choreographies) Consider the cyclic group $G = \mathbb{Z}_n$ of order n acting trivially on V, with a cyclic permutation of order n on the index set $\mathbf{n} = \{1, \dots, n\}$ and with a rotation of angle $2\pi/n$ on the time circle $\mathbb{T}$. Since $\mathcal{X}^G = 0$, by Proposition 2 the action functional $\mathcal{A}^G$ is coercive. Moreover, since the action of G on $\mathbb{T}$ is of cyclic type and $\ker\tau = 1$, by Corollary 2 the minimum exists and it has no collisions. For several numerical results and a description of choreographies we refer the reader to [32]. An insight of the variational properties of choreographies is provided in Sect. "Minimizing Properties of Simple Choreographies".

Example 2 Let n be odd. Consider the dihedral group $G = D_{2n}$ of order $2n$, with the presentation

$$G = \langle g_1, g_2 | g_1^2 = g_2^n = (g_1 g_2)^2 = 1 \rangle .$$

Let τ be the homomorphism defined by

$$\tau(g_1) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} , \text{ and } \tau(g_2) = \begin{bmatrix} \cos\frac{2\pi}{n} & -\sin\frac{2\pi}{n} \\ \sin\frac{2\pi}{n} & \cos\frac{2\pi}{n} \end{bmatrix} .$$

Furthermore, let the homomorphism ρ be defined by

$$\rho(g_1) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \text{ and } \rho(g_2) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} .$$

Finally, let G act on $\mathbf{n}$ by the homomorphism σ defined as $\sigma(g_1) = (1, n-1)(2, n-2)\dots((n-1)/2, (n+1)/2)$,

$\sigma(g_2) = (1, 2, \ldots, n)$, where $(i_1, i_2, \ldots, i_k)$ means the standard cycle-decomposition notation for permutation groups. By the action of g_2 it is easy to show that all the loops in Λ^G are choreographies, and thus that, since $\mathcal{X}^G = 0$, the action functional is coercive. The maximal $\mathbb{T}$-isotropy subgroups are the subgroups of order 2 generated by the elements $g_1 g_2^i$ with $i = 0 \ldots n-1$. Since they are all conjugated, it is enough to show that one of them acts with the rotating circle property. Thus consider $H = \langle g_1 \rangle \subset G$. For every index $i \in \{1, 2, \ldots, n-1\}$ the isotropy $H_i \subset H$ relative to the action of H on $\mathbf{n}$ is trivial, and g_1 acts by rotation on $V = \mathbb{R}^2$. Therefore for every $i \in \{1, 2, \ldots, n-1\}$ it is possible to choose a circle rotating under H for i, since, being H_i trivial, $V^{H_i} = V$. The resulting orbits are not homographic (since all the particles pass through the origin 0 at some time of the trajectory and the configurations are centered). For $n = 3$ this is the *eight with less symmetry* of [25]. Possible trajectories are shown in Figs. 1 and 2.

Example 3 As in the previous example, let $n \geq 3$ be an odd integer. Let $G = C_{2n} \cong \mathbb{Z}_2 + \mathbb{Z}_n$ be the cyclic group of order $2n$, presented as $G = \langle g_1, g_2 | g_1^2 = g_2^n = g_1 g_2 g_1^{-1} g_2^{-1} = 1 \rangle$. The action of G on $\mathbb{T}$ is given by $\tau(g_1 g_2) = \theta_{2n}$, where θ_{2n} denotes the rotation of angle π/n (hence the action will be of cyclic type). Now, G can act on the plane $V = \mathbb{R}^2$ by the homomorphism ρ defined by

$$\rho(g_1) = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } \rho(g_2) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} .$$

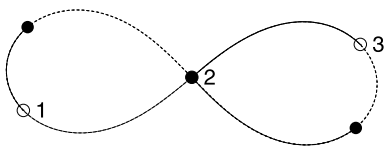

***n*-Body Problem and Choreographies, Figure 1**
The (D_6-symmetric) eight for $n = 3$

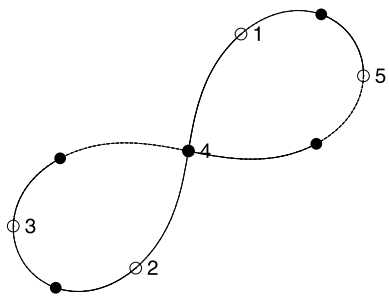

***n*-Body Problem and Choreographies, Figure 2**
The (D_{10}-symmetric) eight with $n = 5$

Finally, the action of G on $\mathbf{n} = \{1, 2, \ldots, n\}$ is given by the homomorphism $\sigma : G \to \Sigma_n$ defined by $\sigma(g_1) = ()$, $\sigma(g_2) = (1, 2, \ldots, n)$. The cyclic subgroup $H_2 = \langle g_2 \rangle \subset G$ gives the symmetry constraints of the choreographies, hence loops in Λ^G are choreographies and the functional is coercive. Furthermore, since the action is of cyclic type, by Corollary 3 the minimum of the action functional is collisionless. It is possible that such minima coincide with the minima of the previous example: this would imply that the symmetry group of the minimum contains the two groups above.

Example 4 Consider four particles with equal masses and an odd integer $q \geq 3$. Let $G = D_{4q} \times C_2$ be the direct product of the dihedral group of order $4q$ with the group C_2 of order 2. Let D_{4q} be presented by $D_{4q} = \langle g_1, g_2 | g_1^2 = g_2^{2q} = (g_1 g_2)^2 = 1 \rangle$, and let $c \in C_2$ be the non-trivial element of C_2. Now define the homomorphisms ρ, τ and σ as follows:

$$\rho(g_1) = \tau(g_1) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} ,$$

$$\rho(g_2) = \tau(g_2) = \begin{bmatrix} \cos\frac{2\pi}{2q} & -\sin\frac{2\pi}{2q} \\ \sin\frac{2\pi}{2q} & \cos\frac{2\pi}{2q} \end{bmatrix} ,$$

$$\rho(c) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} , \quad \tau(c) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} ,$$

$$\sigma(g_1) = (1, 2)(3, 4) , \quad \sigma(g_2) = (1, 3)(2, 4) ,$$

$$\sigma(c) = (1, 2)(3, 4) .$$

It is not difficult to show that $\mathcal{X}^G = 0$, and thus the action is coercive. Moreover, $\ker\tau = C_2$, which acts on $\mathbb{R}^2$ with the rotation of order 2, hence $\ker\tau$ acts with the rotating circle property. Thus, by Proposition 2 and Theorem 1 the minimizer exists and does not have interior collisions. To exclude boundary collisions we cannot invoke Theorem 2, since the maximal $\mathbb{T}$-isotropy subgroups do not act with the rotating circle property. A possible graph for such a minimum can be found in Fig. 4, for $q = 3$ (one needs to prove that the minimum is not the homographic solution – with a level estimate – and that there are no boundary collisions – with an argument similar to [20]).

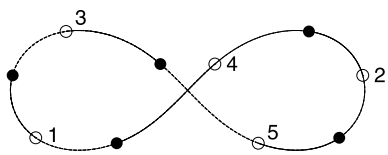

***n*-Body Problem and Choreographies, Figure 3**
Another symmetry constraint for an eight-shaped orbit ($n = 5$)

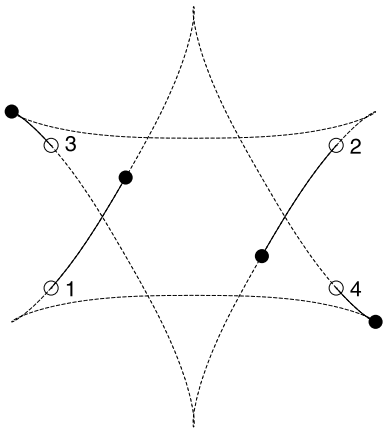

***n*-Body Problem and Choreographies, Figure 4**
The orbit of Example 4 with $q = 3$

See also [22] for an updated and much generalized treatment of such orbits.

Example 5 Consider four particles with equal masses and an even integer $q \geq 4$. Let $G = D_q \times C_2$ be the direct product of the dihedral group of order $2q$ with the group C_2 of order 2. Let D_{4q} be presented by $D_{4q} = \langle g_1, g_2 | g_1^2 = g_2^q = (g_1 g_2)^2 = 1 \rangle$, and let $c \in C_2$ be the non-trivial element of C_2. As in Example 4, define the homomorphisms ρ, τ and σ as follows.

$$\rho(g_1) = \tau(g_1) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},$$
$$\rho(g_2) = \tau(g_2) = \begin{bmatrix} \cos\frac{2\pi}{q} & -\sin\frac{2\pi}{q} \\ \sin\frac{2\pi}{q} & \cos\frac{2\pi}{q} \end{bmatrix},$$
$$\rho(c) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \tau(c) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$
$$\sigma(g_1) = (1,2)(3,4), \quad \sigma(g_2) = (1,3)(2,4),$$
$$\sigma(c) = (1,2)(3,4).$$

Again, one can show that a minimizer without interior collisions exists since $\ker\tau = C_2$ acts with the rotating circle property (a possible minimizer is shown in Fig. 5). This generalizes Chen's orbit [22]. See also [20].

Example 6 (Hip–Hops) If $G = \mathbb{Z}_2$ is the group of order 2 acting trivially on $\mathbf{n}$, acting with the antipodal map on $V = \mathbb{R}^3$ and on the time circle $\mathbb{T}$, then again $\mathcal{X}^G = 0$, so that Proposition 2 holds. Furthermore, since the action is of cyclic type Corollary 2 ensures that minimizers have no collisions. Such minimizers were called *generalized Hip–Hops* in [25]. See also [26]. A subclass of symmetric trajectories leads to a generalization of such a Hip–Hop. Let $n \geq 4$ an even integer. Consider n particles with equal masses, and the group $G = C_n \times C_2$ direct product of the cyclic group of order n (with generator g_1) and the group C_2 of order 2 (with generator g_2). Let the homomorphisms ρ, σ and τ be defined by

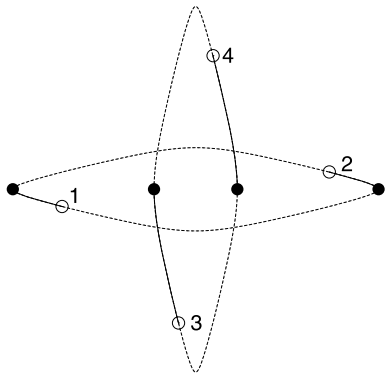

***n*-Body Problem and Choreographies, Figure 5**
A possible minimizer for Example 5

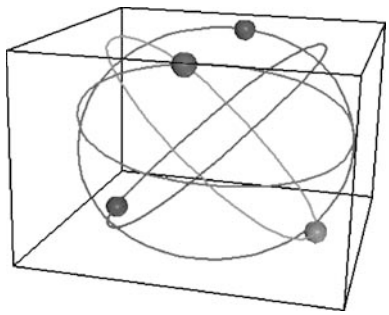

***n*-Body Problem and Choreographies, Figure 6**
The Chenciner–Venturelli Hip–Hop

$$\rho(g_1) = \begin{bmatrix} \cos\frac{2\pi}{n} & -\sin\frac{2\pi}{n} & 0 \\ \sin\frac{2\pi}{n} & \cos\frac{2\pi}{n} & 0 \\ 0 & 0 & 1 \end{bmatrix},$$
$$\rho(g_2) = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \quad \tau(g_1) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$
$$\tau(g_2) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix},$$
$$\sigma(g_1) = (1,2,3,4), \quad \sigma(g_2) = ().$$

It is easy to see that $\mathcal{X}^G = 0$, and thus a minimizer exists. Since the action is of cyclic type, it suffices to exclude interior collisions. But this follows from the fact that $\ker\tau = C_n$ has the rotating circle property. This example is the natural generalization of the Hip–Hop solution of [30] to $n \geq 4$ bodies. We can see the trajectories in Fig. 6.

Example 7 Consider the direct product $G = D_6 \times C_3$ of the dihedral group D_6 (with generators g_1 and g_2 of order 3 and 2 respectively) of order 6 and the cyclic group C_3 of order 3 generated by $c \in C_3$. Let us consider the planar n-body problem with $n = 6$ with the symmetry constraints given by the following G-action.

$$\rho(g_1) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \rho(g_2) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix},$$
$$\rho(c) = \begin{bmatrix} \cos\frac{2\pi}{3} & -\sin\frac{2\pi}{3} \\ \sin\frac{2\pi}{3} & \cos\frac{2\pi}{3} \end{bmatrix},$$
$$\tau(g_1) = \begin{bmatrix} \cos\frac{2\pi}{3} & -\sin\frac{2\pi}{3} \\ \sin\frac{2\pi}{3} & \cos\frac{2\pi}{3} \end{bmatrix},$$
$$\tau(g_2) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \tau(c) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$
$$\sigma(g_1) = (1,3,2)(4,5,6), \quad \sigma(g_2) = (1,4)(2,5)(3,6),$$
$$\sigma(c) = (1,2,3)(4,5,6).$$

By Proposition 2 one can prove that a minimizer exists, and since G acts with the rotating circle property (actually, the elements of the image of ρ are rotations) on $\mathbb{T}$-maximal isotropy subgroups, the conclusion of Theorem 2 holds. It is not difficult to see that configurations in $\mathcal{X}^{\ker\tau}$ are given by two centered equilateral triangles. Now, to guarantee that the minimizer is not a homographic solution, of course it suffices to show that there are no homographic solutions in Λ^G (like in the case of Example 2). This follows from the easy observation that at some times $t \in \mathbb{T}$ with maximal isotropy it happens that $x_1 = -x_4$, $x_2 = -x_5$ and $x_3 = -x_6$, while at some other times it happens that $x_1 = -x_5$, $x_2 = -x_6$ and $x_3 = -x_4$ or that $x_1 = -x_6$, $x_2 = -x_4$ and $x_3 = -x_5$ and this implies that there are no homographic loops in Λ^G. With no difficulties the same action can be defined for $n = 2k$, where k is any odd integer. We can see a possible trajectory in Fig. 7. Also, it is not difficult to consider a similar example in dimension 3. With $n = 6$ and the notation of D_6 and C_3 as above, consider the group $G = D_6 \times C_3 \times C_2$. Let g_1, g_2, c be as above, and let c_2 be the generator of C_2. The homomorphisms ρ, τ and σ are defined in a similar way by

$$\rho(g_1) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \rho(g_2) = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$
$$\rho(c) = \begin{bmatrix} \cos\frac{2\pi}{3} & -\sin\frac{2\pi}{3} & 0 \\ \sin\frac{2\pi}{3} & \cos\frac{2\pi}{3} & 0 \\ 0 & 0 & 1 \end{bmatrix},$$
$$\rho(c_2) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix},$$
$$\tau(g_1) = \begin{bmatrix} \cos\frac{2\pi}{3} & -\sin\frac{2\pi}{3} \\ \sin\frac{2\pi}{3} & \cos\frac{2\pi}{3} \end{bmatrix}, \quad \tau(g_2) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$
$$\tau(c) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \tau(c_2) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix},$$
$$\sigma(g_1) = (1,3,2)(4,5,6), \quad \sigma(g_2) = (1,4)(2,5)(3,6),$$
$$\sigma(c) = (1,2,3)(4,5,6), \quad \sigma(c_2) = ().$$

In the resulting collisionless minimizer (again, it follows by Proposition 2 and Theorem 2) two equilateral triangles rotate in opposite directions and have a "brake" motion on the third axis. The likely shape of the trajectories can be found in Fig. 8.

Example 8 (Marchal's P_{12}–symmetry revisited) Let $k \geq 2$ be an integer, and consider the cyclic group $G = C_{6k}$ of order $6k$ generated by the element $c \in G$. Now consider orbits for $n = 3$ bodies in the space of dimension $d = 3$. With a minimal effort and suitable changes the example can be generalized for every $n \geq 3$. We leave the details to

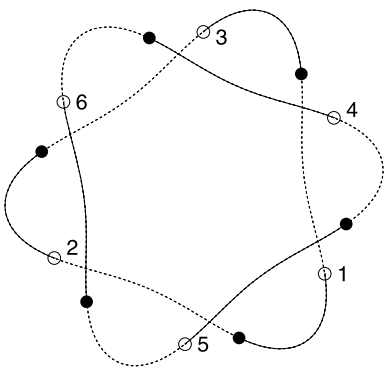

***n*-Body Problem and Choreographies, Figure 7**
The planar equivariant minimizer of Example 7

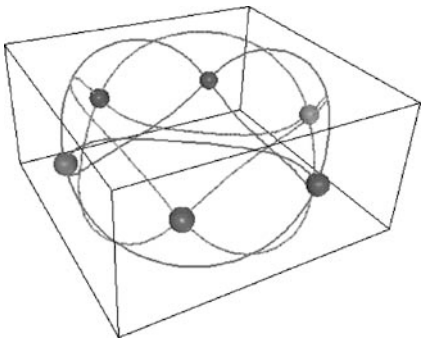

***n*-Body Problem and Choreographies, Figure 8**
The three-dimensional equivariant minimizer of Example 7

the reader. The homomorphisms ρ, τ and σ are defined by

$$\rho(c) = \begin{bmatrix} \cos\frac{\pi}{k} & -\sin\frac{\pi}{k} & 0 \\ \sin\frac{\pi}{k} & \cos\frac{\pi}{k} & 0 \\ 0 & 0 & -1 \end{bmatrix},$$

$$\tau(c) = \begin{bmatrix} \cos\frac{2\pi}{6k} & -\sin\frac{2\pi}{6k} \\ \sin\frac{2\pi}{6k} & \cos\frac{2\pi}{6k} \end{bmatrix},$$

$$\sigma(c) = (1, 2, 3)\,.$$

Straightforward calculations show that $\mathcal{X}^G = 0$ and hence Proposition 2 can be applied. Furthermore, the action is of cyclic type with $\ker\tau = 1$, and hence by Corollary 2 the minimizer does not have collisions. It is left to show that this minimum is not a homographic motion. The only homographic motion in Λ^G is a Lagrange triangle $y(t) = (y_1, y_2, y_3)(t)$, rotating with angular velocity $3-2k$ (assume that the period is 2π, i. e. that $T = |\mathbb{T}| = 2\pi$) in the plane $u_3 = 0$ (let u_1, u_2, u_3 denote the coordinates in $\mathbb{R}^3$). To be a minimum it needs to be inscribed in the horizontal circle of radius $((\alpha 3^{-\alpha/2})/(2(3-2k)^2))^{1/(2+\alpha)}$. Now, for every function $\phi(t)$ defined on $\mathbb{T}$ such that $\phi(c^3 t) = -\phi(t)$, the loop given by $v_1(t) = (0, 0, \phi(t))$, $v_2(t) = (0, 0, \phi(c^{-2}t))$ and $v_3(t) = (0, 0, \phi(c^2 t))$ is G-equivariant, and thus belongs to Λ^G. If one computes the value of Hessian of the Lagrangian action $\mathcal{A}$ in y and in the direction of the loop v one finds that

$$D_v^2\mathcal{A}\big|_y = 3\int_0^{2\pi} \dot{\phi}^2(t)\mathrm{d}t - 2(3-2k)^2 \times \int_0^{2\pi} (\phi(t) + \phi(ct))^2 \mathrm{d}t\,.$$

In particular, if we set the function $\phi(t) = \sin(kt)$, which has the desired property, elementary integration yields

$$D_v^2\mathcal{A}|_y = 3\pi(k^2 - 2(3-2k)^2)\,,$$

which does not depend on α and is negative for every $k \geq 3$. Thus for every $k \geq 3$ the minimizer is not homographic. We see a possible trajectory in Fig. 9.

Remark 1 In the previous example, if $k \not\equiv 0 \mod 3$, the cyclic group G can be written as the sum $C_3 + C_{2k}$. The generator of C_3 acts trivially on V, acts with a rotation of order 3 on $\mathbb{T}$ and with the cyclic permutation $(1, 2, 3)$ on $\{1, 2, 3\}$. This means that for all $k \not\equiv 0 \mod 3$ the orbits of Example 8 are non-planar choreographies. Furthermore, it is possible to define a cyclic action of the same kind by setting τ and σ as above and

$$\rho(c) = \begin{bmatrix} \cos\frac{p}{3k}\pi & -\sin\frac{p}{3k}\pi & 0 \\ \sin\frac{p}{3k}\pi & \cos\frac{p}{3k}\pi & 0 \\ 0 & 0 & -1 \end{bmatrix},$$

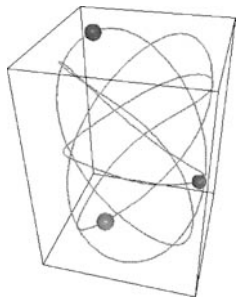

***n*-Body Problem and Choreographies, Figure 9**
The non-planar choreography of Example 8 for $k = 4$

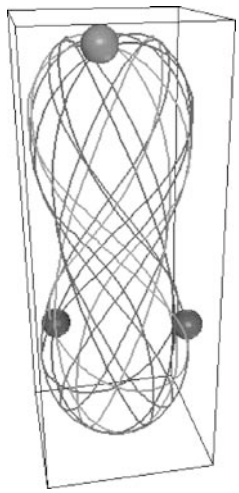

***n*-Body Problem and Choreographies, Figure 10**
A non-planar symmetric orbit of Remark 1, with $k = 3$ and $p = 1$

where p is a non-zero integer. If $p = 3$ one obtains the same action as in Example 8. One can perform similar computations and obtain that the Lagrange orbit (with angular velocity $p - 2k$, this time) is not a minimizer for all (p, k) such that $0 < p < 3k$ and $k^2 - 2(p-2k)^2 < 0$.

More Examples with non Trivial Core

Now consider the *core* of G, $\ker\tau = K \subset G$. With an abuse of notation, using the factorization of the action, we can identify K with a subgroup of $O(3)$. A very remarkable class of symmetry groups with non trivial core has been defined by Ferrario in [50], through its *Krh* decomposition:

Definition 4 Let G be a symmetry group. Then define

1. $K = \ker\tau$,
2. $[r] \in W_{O(3)}K$ as the image in the Weyl group of the generator mod K of $\ker\det\tau \subset G/K$ (corresponding

to the time-shift with minimal angle). If $\ker\det\tau = K$, then $[r] = 1$.

3. $[h] \in W_{O(3)}K$ as the image in the Weyl group of one of the time-reflections mod K in G/K, in the cases such an element exists. Otherwise it is not defined.

In short, the triple

$$(K, [r], [h])$$

is said the *Krh* data of G.

Example 9

Consider the icosahedral group Y of order 60. The group G with *Krh* data

$$\begin{pmatrix} 1 & (1,-1) \\ Y & -1 \end{pmatrix}$$

is isomorphic to the direct product $I \times Y$ of order 120, and acts on the euclidean space $\mathbb{R}^3$ as the full icosahedron group. The action on $\mathbb{T}$ is cyclic and given by the fact that $\ker\tau = 1 \times Y$. The isotropy is generated by the central inversion -1, and hence the set of bodies is $G/I \cong Y$. Thus at any time t the 60 point particles are constrained to be a Y-orbit in $\mathbb{R}^3$ (which does not mean they are vertices of a icosahedron, simply that the configuration is Y-equivariant). After half period every body is in the antipodal position: $x_i(t + T/2) = -x_i$ (in other words, the group contains the *anti-symmetry*), also known as *Italian symmetry* – see [25,26]. Of course, the group Y is just an example: one can choose also the tetrahedral group T or the octahedral O and obtain anti-symmetric orbits for 12 (tetrahedral) or 24 (octahedral) bodies, as depicted in Fig. 11. The action is by its definition transitive and coercive; local minimizers are collisionless since the maximal $\mathbb{T}$-isotropy group acts as a subgroup of $SO(3)$ (i. e. orientation–preserving).

Example 10 Let G be the group with *Krh* data

$$\begin{pmatrix} 1 & (1,-1) \\ D_k & -1 \end{pmatrix},$$

where D_k is the rotation dihedral group of order $2k$. As in the previous example, the action is such that the action functional is coercive and its local minima collisionless. At every time instant the bodies are D_k-equivariant in $\mathbb{R}^k$ and the anti-symmetry holds. Approximations of minima can be seen in Fig. 12.

Example 11 To illustrate the case of non-transitive symmetry group, consider the following (cyclic) *Krh* data:

$$\begin{pmatrix} 1 & (3,-1) \\ 1 & -1 \end{pmatrix},$$

which yields a group of order 6 acting cyclically on 3 bodies, and with the antipodal map on $\mathbb{R}^3$. Since $\ker\tau$ is trivial and the group is of cyclic type, local minima are collisionless. Now, by adding k copies of such group one obtains a symmetry group having k copies of it as its transitive components, where still local minimizers are collisionless and the restricted functional is coercive. Some possible minima can be found in Fig. 13, for $k = 3, 4$.

The 3-Body Problem

The major achievement of [14] is to give the complete description of the outcome of the equivariant minimization procedure for the planar three-body problem. First we can ensure that minimizers are always collisionless.

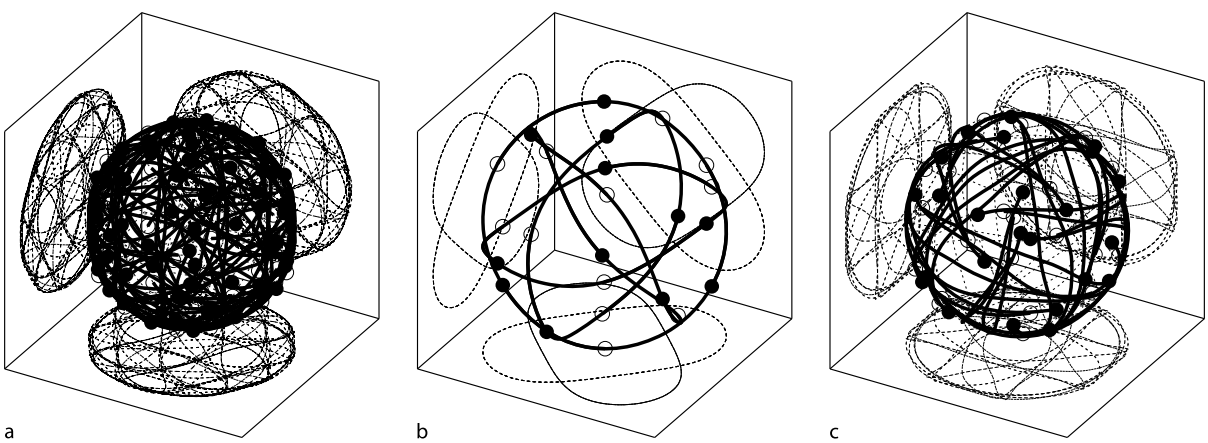

***n*-Body Problem and Choreographies, Figure 11**
60-icosahedral *Y*, 12-tetrahedral *T* and 24-octahedral *O* periodic minimizers (chiral)

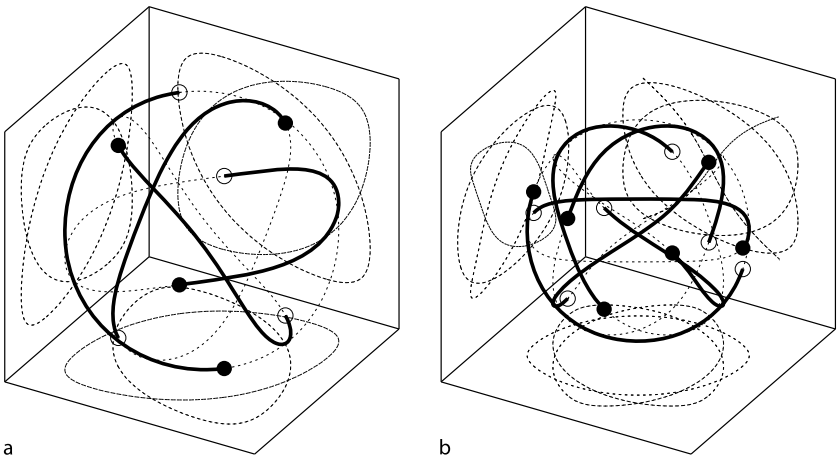

***n*-Body Problem and Choreographies, Figure 12**
4-dihedral D_2 and 6-dihedral D_6 symmetric periodic minimizers

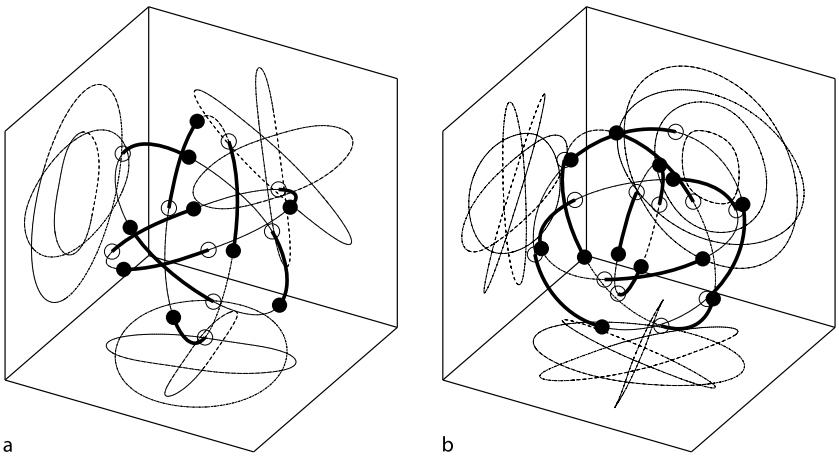

***n*-Body Problem and Choreographies, Figure 13**
9 and 12 bodies in anti-choreographic constraints grouped by 3

Theorem 3 *Let G a symmetry group of the Lagrangian in the 3-body problem (in a rotating frame or not). If G is not bound to collision (i. e. every equivariant loop has collisions), then any (possible) local minimizer is collisionless.*

A symmetry group G of the Lagrangian functional $\mathcal{A}$ is termed

- *bound to collisions* if all G-equivariant loops actually have collisions,
- *fully uncoercive* if for every possible rotation vector ω the action functional $\mathcal{A}_\omega^G$ in the frame rotating around ω with angular speed $|\omega|$ is not coercive in the space of G-equivariant loops (that is, its global minimum escapes to infinity);
- *homographic* if all G-equivariant loops are constant up to orthogonal motions and rescaling.
- The *core* of the group G is the subgroup of all the elements which do not move the time $t \in \mathbb{T}$.

If, for every angular velocity, G is a symmetry group for the Lagrangian functional in the rotating frame, then we will say that G is of type R. This is a fundamental property for symmetry groups. In fact, if G is *not* of type R, it turns out that the angular momentum of all G-equivariant trajectories vanishes.

The Classification of Planar Symmetry Groups for 3-body

Theorem 4 *Let G be a symmetry group of the Lagrangian action functional in the planar 3-body problem. Then, up to a change of rotating frame, G is either bound to collisions, fully uncoercive, homographic, or conjugated to one of the*

***n*-Body Problem and Choreographies, Table 1**
Planar symmetry groups with trivial core

Name	$\lvert G\rvert$	Type R	Action type	trans. dec.	RCP	HGM
Trivial	1	yes		$1+1+1$	yes	yes
Line	2	yes	brake	$1+1+1$	(no)	no
$2-1$-choreography	2	yes	cyclic	$2+1$	yes	no
Isosceles	2	yes	brake	$2+1$	no	yes
Hill	4	yes	dihedral	$2+1$	no	no
3-choreography	3	yes	cyclic	3	yes	yes
Lagrange	6	yes	dihedral	3	no	yes
C_6	6	no	cyclic	3	yes	no
D_6	6	no	dihedral	3	yes	no
D_{12}	12	no	dihedral	3	no	no

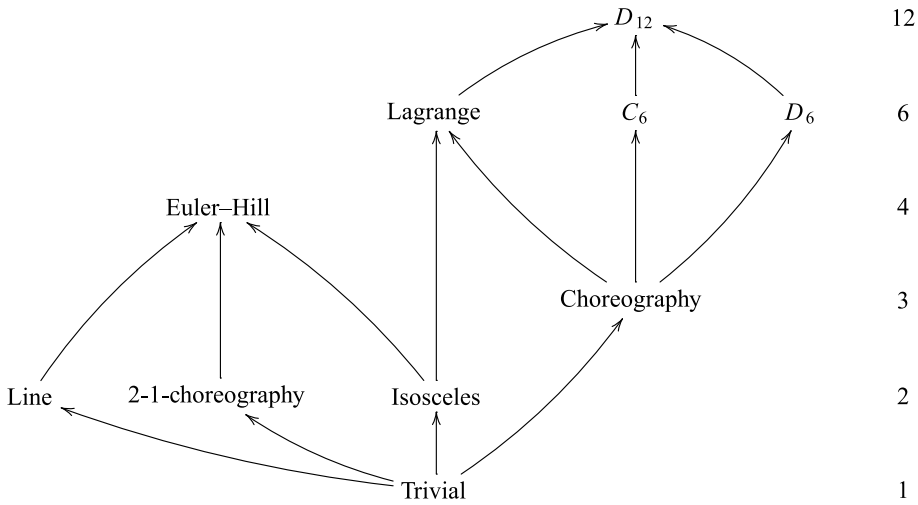

***n*-Body Problem and Choreographies, Figure 14**
The poset of symmetry groups for the planar 3-body problem

symmetry groups listed in Table 1 (RCS stands for Rotating Circle Property and HGM for Homographic Global Minimizer).

Planar Symmetry Groups

- *The trivial symmetry*
  Let G be the trivial subgroup of order 1. It is clear that it is of type R, it has the rotating circle property. It yields a coercive functional on $\Lambda^G = \Lambda$ only when ω is not an integer. If $\omega = \frac{1}{2} \mod 1$ then the minimizers are minimizers for the anti-symmetric symmetry group (also known as *Italian* symmetry) $x(at) = ax(t)$, where a is the antipodal map on $\mathbb{T}$ and E. The masses can be different.

Proposition 3 *For every $\omega \notin \mathbb{Z}$ and every choice of masses the minimum for the trivial symmetry occurs in the relative equilibrium motion associated to the Lagrange central configuration.*

- *The line symmetry*
  Another case of symmetry group that can be extended to rotating frames with arbitrary masses is the line symmetry: the group is a group of order 2 acting by a reflection on the time circle $\mathbb{T}$, by a reflection on the plane E, and trivially on the set of indexes. That means, at time 0 and π the masses are collinear, on a fixed line $l \subset E$. It is coercive only when $\omega \notin \mathbf{Z}$. In this case the Lagrangian solution cannot be a minimum, while it can the relative equilibrium associated with the Euler configuration.
- *The $2-1$-choreography symmetry*
  Consider the group of order 2 acting as follows: $\rho(g) = 1$, $\tau(g) = -1$ (that is, the translation of half-period) and $\sigma(g) = (1,2)$ (that is, $\sigma(g)(1) = 2$, $\sigma(g)(2) = 1$,

and $\sigma(g)(3) = 3$). That is, it is a half-period choreography for the bodies 1 and 2. It can be extended to rotating frames and coercive for a suitable choice of $\omega \neq 0$, $1 \mod 2$.
The Euler's orbit with $k = 1$ and the Hill's orbits with $k = \pm 1$ are equivariant for the $2-1$-choreography symmetry, while the Euler's orbit with $k = 0$ is not equivariant for this symmetry. In Figures 15 and 16 Euler 1 represents the action levels on the Euler's orbit with $k = 1$, Hill 1–2 the ones on the Hill's with $k = \pm 1$.

- *The isosceles symmetry*
  The isosceles symmetry can be obtained as follows: the group is of order 2, generated by h; $\tau(h)$ is a reflection in the time circle $\mathbb{T}$, $\rho(h)$ is a reflection along a line l in E, and $\sigma(h) = (1, 2)$ as above. The constraint is therefore that at time 0 and π the 3-body configuration is an isosceles triangle with one vertex on 1 (the third).

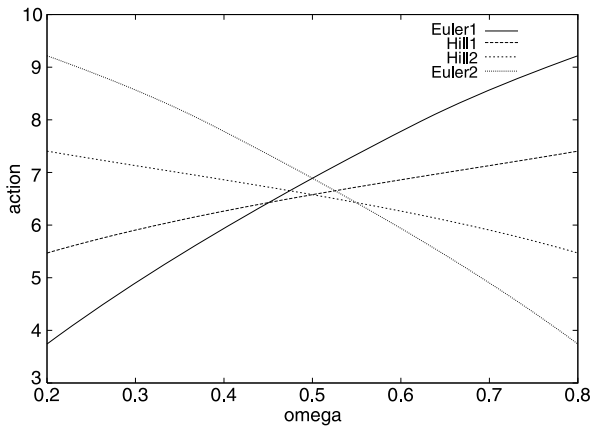

***n*-Body Problem and Choreographies, Figure 15**
Action levels for the line symmetry

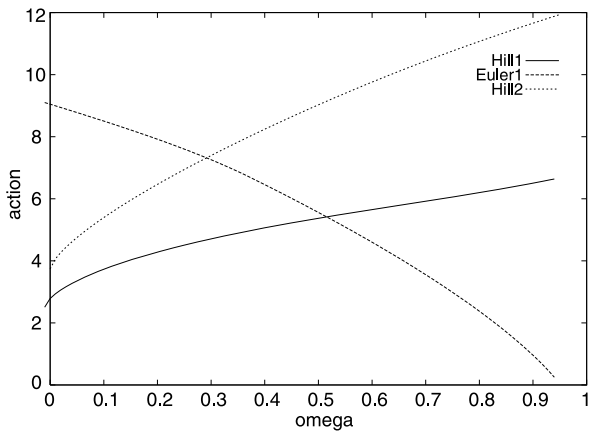

***n*-Body Problem and Choreographies, Figure 16**
Action levels for the 2 — 1-choreography symmetry

Proposition 4 *For every $\omega \notin \mathbb{Z}$ and every choice of masses the minimum for the isosceles symmetry occurs in the relative equilibrium motion associated to the Lagrange configuration.*

- *The Euler–Hill symmetry*
  Now consider the symmetry group with a cyclic generator r of order 2 (i. e. $\tau(r) = -1$) and a time reflection h (i. e. $\tau(h)$ is a reflection of $\mathbb{T}$) given by $\rho(r) = 1$, $\sigma(r) = (1, 2)$, $\rho(h)$ is a reflection and $\sigma(h) = ()$. It contains the $2-1$-choreography (as the subgroup $\ker \det(\tau)$), the isosceles symmetry (as the isotropy of $\pi/2 \in \mathbb{T}$) and the line symmetry (as the isotropy of $0 \in \mathbb{T}$) as subgroups.

Proposition 5 *The minimum of the Euler–Hill symmetry is not homographic, provided that the angular velocity ω is close to 0.5 and the values of the masses are close to 1.*

- *The choreography symmetry*
  The choreography symmetry is given by the group C_3 of order 3 acting trivially on the plane E, by a rotation of order 3 in the time circle $\mathbb{T}$ and by the cyclic permutation $(1, 2, 3)$ of the indices.

Proposition 6 *For every ω the minimal choreography of the 3 body problem is a rotating Lagrange configuration.*

- *The Lagrange symmetry*
  The Lagrange symmetry group is the extension of the choreography symmetry group by the isosceles symmetry group. Thus, it is a dihedral group of order 6, the action is of type R. Hence, the relative equilibrium motions associated to the Lagrange configuration are admissible motions for this symmetry and, again, the minimizer occurs in the relative equilibrium motion associated to the Lagrange configuration.
- *The Chenciner–Montgomery symmetry group and the eights*
  There are three symmetry groups (up to change of coordinates) that yield the Chenciner–Montgomery *figure eight* orbit: they are the only symmetry groups which do not extend to the rotating frame and we have already described them in Subsect. "The Eight Shaped Three-Body Solution". One can prove that all G-equivariant trajectories have vanishing angular momentum, whenever the group is not of type R. Moreover, we were able to partially answer to the open question (posed by

Chenciner) whether their minimizers coincide or not: for two of them (D_6 and D_{12}) the minimizer is necessarily the same.

Space Three-body Problem

Based on the classification of planar groups, by introducing a natural notion of space extension of a planar group, Ferrario gave in [51] a complete answer to the classification problem for the three-body problem in the space and at the same time to determine the resulting minimizers and describe its more relevant properties.

Theorem 5 *Symmetry groups not bound to collisions, not fully uncoercive and not homographic are, up to a change of rotating frame, either the three-dimensional extensions of planar groups (if trivial core) listed in table below or the vertical isosceles triangle (if non-trivial core).*

The next theorem is the answer to the natural questions about collisions and description of some main features of minimizers.

Theorem 6 *Let G be a symmetry group not bound to collisions and not fully uncoercive. Then*

1. *Local minima of $\mathcal{A}^\omega$ do not have collisions.*
2. *In the following cases minimizers are planar trajectories:*
   a) *If G is not of type R: $D_6^{+,-}$, $D_6^{-,+}$ and $D_{12}^{-,+}$ (and then G-equivariant minimizers are Chenciner–Montgomery eights).*
   b) *If there is a G-equivariant minimal Lagrange rotating solution: C_1^-, $H_2^{+,-}$, C_3^+, $L_6^{+,+}$, $L_6^{+,-}$ (and then the Lagrange solution is of course the minimizer).*
   c) *If the core is non-trivial and it is not the vertical isosceles (and then minimizers are homographic).*
3. *In the following cases minimizers are always non-planar:*
   a) *The groups $L_6^{-,+}$ and C_3^- for all $\omega \in (-1,1) + 6\mathbb{Z}$, $\omega \neq 0$ (the minimizers for $L_6^{-,+}$ are the elements of Marchal family P_{12}, and minimizers of C_3^- are a less-symmetric family P'_{12}[2]).*
   b) *The extensions of line and Hill–Euler type groups, for on open subset of mass distributions and angular speeds ω: $L_2^{+,-}$, $L_2^{-,+}$, $H_4^{+,-}$ and $H_4^{-,+}$ (for $L_2^{-,+}$ this happens also with equal masses).*
   c) *The vertical isosceles for suitable choices of masses and ω.*

***n*-Body Problem and Choreographies, Table 2**
Space extensions of planar symmetry groups with trivial core

Name	Extensions
Trivial	C_1^-
Line	$L_2^{+,-}, L_2^{-,+}$
Isosceles	$H_2^{+,-}, H_2^{-,+}$
Hill	$H_4^{+,-}, H_4^{-,+}$
3-choreography	C_3^+, C_3^-
Lagrange	$L_6^{+,+}, L_6^{+,-}, L_6^{-,+}$
D_6	$D_6^{+,-}, D_6^{-,+}$
D_{12}	$D_{12}^{-,+}$

Minimizing Properties of Simple Choreographies

In the space of symmetric (choregraphical) loops, the action takes the form

$$\mathcal{A}(x) = \frac{1}{2}\sum_{h=0}^{n-1}\int_0^\tau |\dot{x}(t+h\tau)|^2 \mathrm{d}t + \frac{1}{2} \times \sum_{\substack{h,l=0\\ h\neq l}}^{n-1} \int_0^\tau \frac{\mathrm{d}t}{|x(t+l\tau)-x(t+h\tau)|^\alpha}\,.$$

A natural question concerns the nature of the minimizers under the sole choregraphical constraint. Unfortunately, the bare minimization among choreographical loops gives returns only trivial motions:

Theorem 7 *For every $\alpha \in \mathbb{R}_+^*$ and $d \geq 2$, the absolute minimum of $\mathcal{A}$ on Λ is attained on a relative equilibrium motion associated to the regular n-gon.*

This theorem extends some related result for the italian symmetry by Chenciner and Desolneux [27], and the results in Sect. "The 3-Body Problem". The proof is based on a (quite involved) convexity argument together with the analysis of some spectral properties related to the choreographical constraint. Now, in order to find nontrivial minimizers, we look at the same problem in a rotating frame. In order to take into account of the Coriolis force, the new action functional has to contain a gyroscopic term:

$$\mathcal{A}(y) = \frac{1}{2}\int_0^{2\pi} |\dot{y}(t) + J\omega y(t)|^2 \mathrm{d}t + \frac{1}{2}\sum_{h=1}^{n-1}\int_0^{2\pi} \frac{\mathrm{d}t}{|y(t)-y(t+h\tau)|^\alpha}\,.$$

Consider the function $h\colon \mathbb{R}_+^* \to \mathbb{N}$, $h(\omega) = \min_{n\in\mathbb{N}^*}(\omega - n)^2/n^2$ and let $\omega^* = \frac{4}{3}$. The same technique used

[2] Highly likely they are not distinct families: this is the recurring phenomenon of "more symmetries than expected" in *n*-body problems.

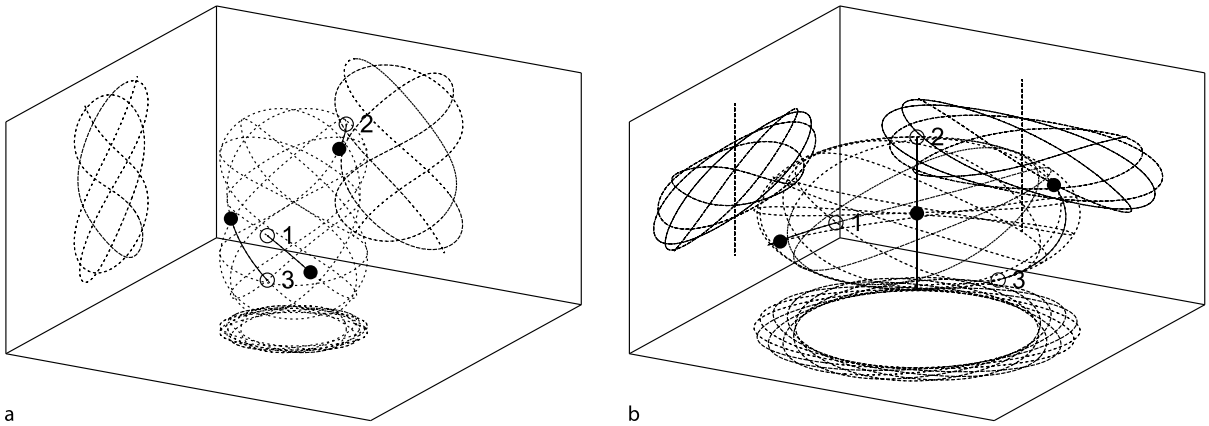

***n*-Body Problem and Choreographies, Figure 17**
Non planar minimizers for the groups $L_6^{-,+}$ and $L_2^{+,-}$, plotted in the fixed frame

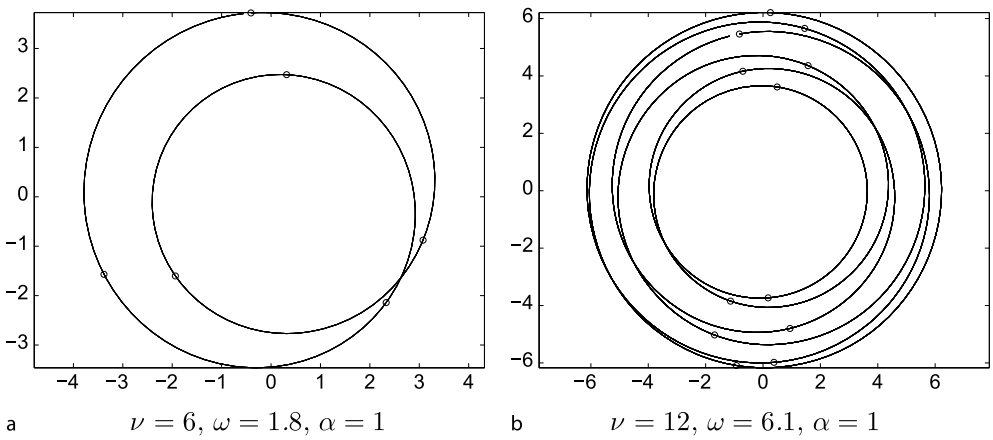

***n*-Body Problem and Choreographies, Figure 18**
Examples for Theorem 10 close to an integer that divides *n*

for the inertial system extends to rotating systems having small angular velocity; this gives the following result.

Theorem 8 *If $\omega \in (0, \omega^*)\backslash\{1\}$, then the action attains its minimum on a circle with minimal period 2π and radius depending on n, α and ω.*

When ω is Close to an Integer

The situation changes dramatically when ω is close to some integer. To understand this phenomenon, let us first check the result of the minimization procedure when ω is an integer:

Proposition 7

(1) *If $\omega = n$, then the action has a continuum of minimizers.*

(2) *If $\omega = k$, coprime with n, then the action does not achieve its infimum (escape of minimizing sequences).*

(3) *If $\omega = k$ and k divides n, then the action does not achieve its infimum (clustering of minimizing sequences).*

As a consequence, we have the following result:

Theorem 9 *Suppose that n and k are coprime. Then there exist $\epsilon = \epsilon(\alpha, n, k)$ such that if $\omega \in (k - \epsilon, k + \epsilon)$ the minimum of the action is attained on a circle with minimal period $2\pi/k$ that lies in the rotating plane with radius depending on n, α and ω.*

An interesting situation appears when the integer closest to the angular velocity is not coprime with the number of bodies. In this case we prove that the minimal orbit it is not circle anymore, as the following theorem states.

Theorem 10 *Take $k \in \mathbb{N}$ and $g.c.d.(k, n) = \tilde{k} > 1$, $\tilde{k} \neq n$. Then there exists $\epsilon = \epsilon(\alpha, n, k) > 0$ such that if $\omega \in (k - \epsilon, k + \epsilon) \backslash \{k\}$ the minimum of the action is attained on a planar 2π-periodic orbit with winding number k which is not a relative equilibrium motion.*

Also, it has be noticed that, for large number of bodies and angular velocities close to the half on an integer, the minimizer apparently is not anymore planar.

Mountain Pass Solutions for the Choreographical 3-Body Problem

The discussion carried in the previous section shows that, as the angular velocity varies, the minimizer's shape must undergoes some transitions (for example it has to pass from relative equilibrium having different winding numbers). This scenario suggests the presence of other critical points, such as local minimizers or mountain pass. This was indeed discovered numerically in [12] and then proved by a computer assisted proof in [7].

To begin with, let us look at the following figure where the values of the action functional $\mathcal{A}^{\omega}$ on the branches of circular orbits L_k^{ω} are plotted:

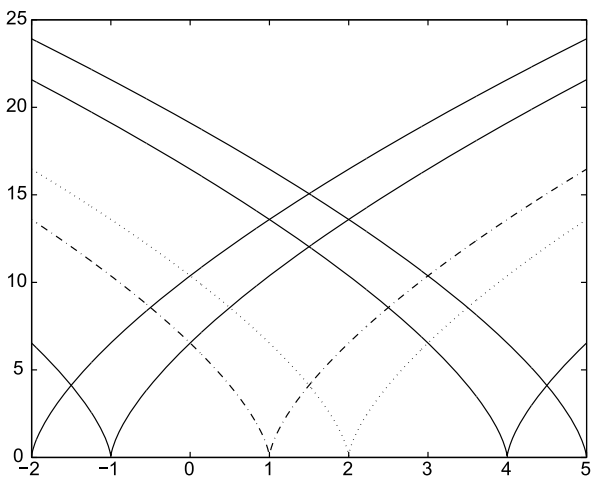

The analysis of this picture suggests the presence of critical points different from the Lagrange motions. Indeed, let us take the angular velocity $\omega = 1.5$: in this case there are two distinct global minimizers, the uniform circular motions with minimal period 2π and π, lying in the plane orthogonal to the rotation direction. This is a well known structure in Critical Point Theory, referred as the Mountain Pass geometry and gives the existence of a third critical point, provided the Palais–Smale condition is fulfilled, with an additional information on the Morse index. Next theorem follows from the application of the Mountain Pass Theorem to the action functional $\mathcal{A}^{3/2}$:

Theorem 11 *There exists a (possibly collision) critical point for the action functional $\mathcal{A}^{3/2}$ with Morse index smaller then 1 and distinct from any Lagrange motion.*

Once the existence of a Mountain Pass critical point was theoretically established, we studied its main properties in order to understand whether it belonged to some known families of periodic trajectories. To this aim we applied the bisection algorithm proposed in [13] to approximate the maximal of a locally optimal path joining the two strict global minimizers, finding in this a good numerical candidate. Of course, there could be a gap between the mountain pass solution whose existence is ensured by Theorem 11 and the numerical candidate found by applying the bisection algorithm. In order to fill this gap proved the existence of an actual solution very close to the numerical output of the Mountain Pass algorithm. The argument was based upon a fixed point principle and involved a rigorous computer assisted proof. As a consequence, we obtained the existence of a new branch of solution for the spatial 3-body problem (see Fig. 8).

Here are some relevant features of the new solution: the orbit is not planar, its winding number with respect, for instance, to the line $x = -0.2$, $y = 0$ is 2 and it does not intersect itself. A natural question is whether this solution can be continued as a function of the parameter ω. We were able to extend the numerical–rigorous argument to cover a full interval of values of the angular velocity, providing the existence of a full branch of solution.

Theorem 12 *There exists a smooth map $B(\omega)$ giving the (locally unique, up to symmetries) branch of solution of the choreographical 3-body problem for all $\omega \in [1, 2]$, starting at the Mountain Pass solutions for $\omega = 1.5$.*

A natural question is whether this mountain pass branch meets one of the known branches of choreographical periodic orbits: either one of the Lagrange or Marchal's P_{12} (described by Marchal in [62]) families. Surprisingly enough, it turns out from further numerical computation that this branch does not emanates from any of Lagrange motions L_1 or L_2; apparently, it emanates from the branch of P_{12} solutions. The details of the bifurcation diagram are depicted in Fig. 8.

Generalized Orbits and Singularities

The existence of a G-equivariant minimizer of the action is a simple consequence of the direct method of the Calculus of Variations. These trajectories, however, solve the associated differential equations only where they are far from

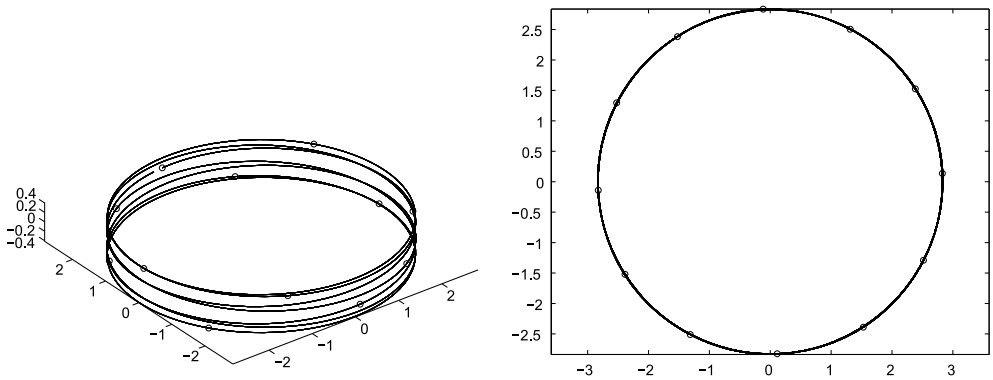

a $\nu = 12,\ \omega = 6.55,\ \alpha = 1$ b

***n*-Body Problem and Choreographies, Figure 19**

Non planar minimizers of the action with angular velocities close to the half on an integer

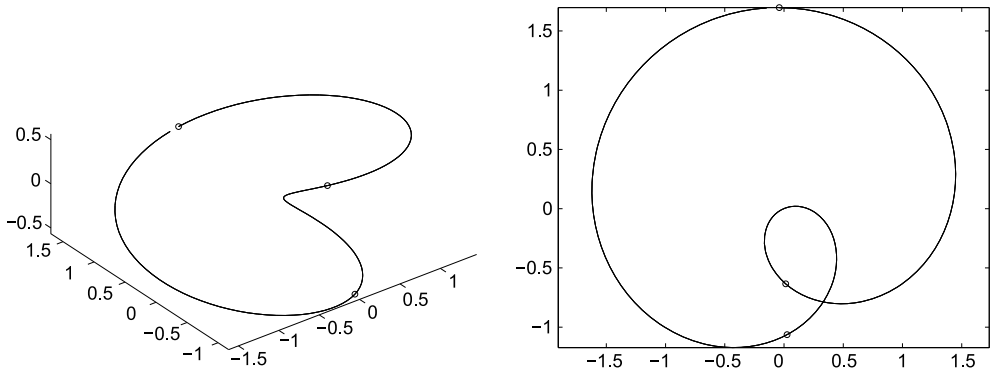

a $\nu = 3,\ \omega = 1.5,\ \alpha = 1$ b

***n*-Body Problem and Choreographies, Figure 20**

Mountain pass solution with angular velocities close to the half on an integer

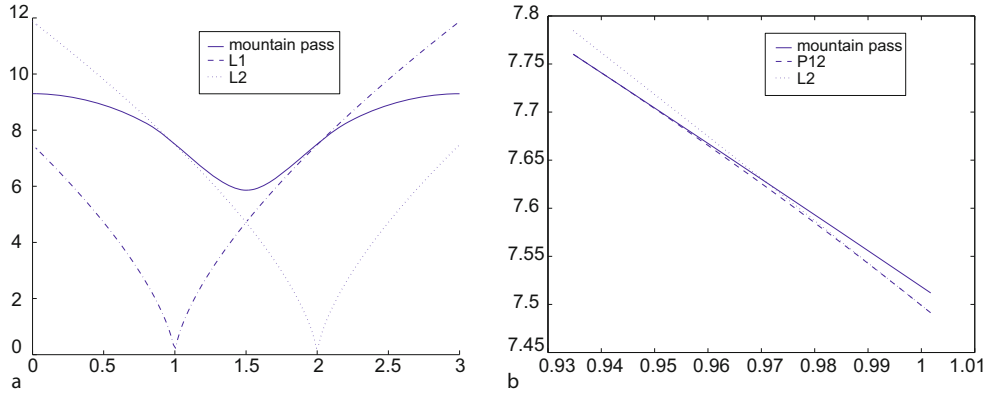

a b

***n*-Body Problem and Choreographies, Figure 21**

Action levels for the Lagrange and the mountain pass solution in the 3-body problem. On the *x*-axes the angular velocity varies in the interval [0, 3). The *left picture* focuses on the bifurcation from the P_{12} family

the singular set. Indeed, generally speaking, G-equivariant minimizers may present singularities, even though the set of collision instants must clearly be of vanishing Lebesgue measure. A first natural question concerns the number of possible collision instants. Here below we follow [15].

Definition 5 A path $x\colon (a,b) \to \mathcal{X}$ is called a *locally minimizing solution* of the N-body problem if, for every $t_0 \in (a,b)$, there exists $\delta > 0$ such that the restriction of x to $[t_0 - \delta, t_0 + \delta]$ is a local minimizer for the action with the same boundary conditions. A path which is the uniform limit of a sequence of locally minimal solutions will be called a *generalized solution*.

We remark that:

- The definition can be modified in a obvious manner to include symmetries.
- Equivariant minimizers are generalized solutions.
- If the potential is of class C^2 outside collisions then every non collision solution is a generalized solution.
- The generalized solutions possess an index: the minimal number of intervals I_j needed to cover (a,b) such that the restriction of x to I_j is a local minimizer for the action.
- There is a natural notion of maximal existence interval for generalized solutions, even without the unique extension property.

Singularities and Collisions

Generally speaking we are dealing with systems of the form

$$M\ddot{x} = \nabla U(t,x)\,, \quad t \in (a,b)\,, \quad M_{ij} = m_i \delta_{ij}\,,$$

where the potential U possesses a singular set $(a,b) \times \Delta$, in the sense that

$$[\mathrm{U0}] \quad \lim_{x \to \Delta} U(t,x) = +\infty\,, \quad \text{uniformly in } t\,.$$

The set Δ is clearly the *collision set*. We assume Δ to be a *cone* in $\mathbb{R}^{nd}$:

$$x \in \Delta \implies \lambda x \in \Delta\,.$$

Definition 6 We say that a generalized solution x on the interval (a,b), has a *singularity* at $t^* < +\infty$ if

$$\limsup_{t \to t^*} U(t, x(t)) = +\infty\,.$$

When $t^* \in (a,b)$ we will say that x has an *interior singularity* at $t = t^*$, while when $t^* = a$ or $t^* = b$ (when finite) we will talk about a *boundary singularity*.

The Theorems of Painlevé and Von Zeipel

In the usual mathematical language a classical solution x on the interval (a,b), has a singularity at $t^* < +\infty$ if it is not possible to extend x as a (classical) solution to a larger interval $(a, t^* + \delta)$. A classical results relates singularities of solutions with those of the potential.

Theorem 13 (Painlevé's Theorem) *Let $\bar{x}$ be a classical solution for the n-body dynamical system on the interval $[0, t^*)$. If $\bar{x}$ has a singularity at $t^* < +\infty$, then*

$$\lim_{t \to t^*} U(\bar{x}(t)) = +\infty\,.$$

Painlevé's Theorem does not necessarily imply that a *collision* (i. e. that is a singularity such the configuration has a definite limit) occurs when there is a singularity at a finite time (on this subject we refer to [66,75,76]). The next result has been stated by Von Zeipel in 1908 and definitely proved by Sperling in 1970.

Theorem 14 (Von Zeipel's Theorem) *If $\bar{x}$ is a classical solution for the n-body dynamical system on the interval (a, t^*) with a singularity at $t^* < +\infty$ and*

$$\lim_{t \to t^*} \|\bar{x}(t)\| < +\infty\,,$$

then $\bar{x}(t)$ has a definite limit configuration x^* a s t tends to t^*.

Von Zeipel's Theorem and the Structure of the Collision Set

To the aim of extending the Von Zeipel's Theorem to our notion of generalized solutions, we need to introduce some assumptions on the potential U and its singular set Δ.

$$\Delta = \bigcup_{\mu \in \mathcal{M}} V_\mu\,, \tag{5}$$

where the V_μ's are distinct linear subspaces of $\mathbb{R}^k$ and $\mathcal{M}$ is a finite set; observe that the set Δ is a cone, as required before. We endow the family of the V_μ's with the inclusion partial ordering and we assume the family to be closed with respect to intersection (thus we are assuming that $\mathcal{M}$ is a semilattice of linear subspaces of $\mathbb{R}^k$: it is the intersection semilattice generated by the arrangement of maximal subspaces V_μ's). With each $\xi \in \Delta$ we associate

$$\mu(\xi) = \min\{\mu\colon \xi \in V_\mu\} \quad \text{i. e.}\,, \quad V_{\mu(\xi)} = \bigcap_{\xi \in V_\mu} V_\mu\,.$$

Fixed $\mu \in \mathcal{M}$ we define the set of collision configurations satisfying

$$\Delta_\mu = \{\xi \in \Delta\colon \mu(\xi) = \mu\}$$

and we observe that this is an open subset of V_μ and its closure $\overline{\Delta_\mu}$ is V_μ. We also notice that the map $\xi \to \dim(V_{\mu(\xi)})$ is lower semicontinuous.

We denote by p_μ the orthogonal projection onto V_μ and we write

$$x = p_\mu(x) + w_\mu(x)\,,$$

where, of course, $w_\mu = \mathbb{I} - p_\mu$.

We assume that, near the collision set, the potential depends, roughly, only on the projection orthogonal to the collision set: more precisely we assume

[U5] For every $\xi \in \Delta$, there is $\varepsilon > 0$ such that

$$U(t,x) - U(t, w_{\mu(\xi)}(x)) = W(t,x) \in C^1((a,b)\times B_\varepsilon(\xi))\,,$$

where $B_\varepsilon(\xi) = \{x\colon |x-\xi| < \varepsilon\}$.

With these assumptions, we can extend then Von Zeipel's Theorem to generalized solutions:

Theorem 15 *Let $\bar{x}$ be a generalized solution on the bounded interval (a,b). If $\bar{x}$ is bounded on the whole interval (a,b) then the singularities of $\bar{x}$ are collisions.*

Asymptotic Estimates at Collisions

One Side Conditions on the Potential and Its Radial Derivative

From now on, we require our potential U to satisfy some *one-side homogeneity* conditions:

[U1] there exists $C_1 \geq 0$ such that for every $(t,x) \in ((a,b)\times\mathbb{R}^{nd}\backslash\Delta)$

$$\left|\frac{\partial U}{\partial t}(t,x)\right| \leq C_1 U(t,x)\,.$$

[U2] there exist $\alpha \in (0,2)$, $\gamma > 0$ and $C_2 \geq 0$ such that

$$\nabla U(t,x)\cdot x + \alpha U(t,x) \geq -C_2|x|^\gamma U(t,x)\,,$$

whenever $|x|$ is small .

We observe that when U is homogeneous functions of degree $-\alpha$, the equality in condition (U2) is attained with $C_2 = 0$; this assumption is satisfied also when we take into account potentials of the form $U(t,x) = U_\alpha(x) + U_\beta(x)$ where U_α is homogeneous of degree $-\alpha$, U_β is homogeneous of degree $-\beta$, $0 < \beta < \alpha$, and U_α is positive.

Isolatedness of Collisions Instants

In order to study the behavior of the solution at a collision, we firs perform a translation in such a way that the collision holds at the origin; next we introduce *polar coordinates*:

$$r = \sqrt{Mx\cdot x} = \sqrt{I}\,, \quad s = \frac{x}{r}\,,$$

where $s \in \mathcal{E} = \{x\colon I^2(x) = Mx\cdot x = 1\}$ belongs the ellipsoid of all the configurations having unitary moment of inertia.

[U3$_h$] there exists a function $\tilde{U}$ defined on $(a,b)\times(\mathcal{E}\backslash\Delta)$, such that (on compact subsets of $((a,b)\times\mathcal{E}\backslash\Delta)$):

$$\lim_{r\to 0} r^\alpha U(t,rs) = \tilde{U}(t,s)\,;$$

[U3$_l$] there exists a function $\tilde{U}$ defined on $(a,b)\times(\mathcal{E}\backslash\Delta)$, such that (on compact subsets of $((a,b)\times\mathcal{E}\backslash\Delta)$):

$$\lim_{|x|\to 0}\left[U(t,x) + M(t)\log|x|\right] = \tilde{U}(t,s)\,;$$

The following regularity theorem is proved in [15] based on a suitable variant of Sundman's inequality and the asymptotic analysis of possible collisions outlined in the sequel.

Theorem 16 *Let $x\colon (a,b)\to\mathcal{X}$ be a generalized solution of the N-body problem, with a potential U satisfying [U0–U3]. Then collision instants are isolated in (a,b). Furthermore, if (a,b) is the finite maximal extension interval of x and no escape in finite time occurs then the number of collision instants is finite.*

Some remarks are in order:

- A generalized solution does not solve the Euler–Lagrange equation in a distributional sense (the force field can not be locally integrable). Moreover its action needs not to be finite: one needs to prove it.
- a priori our solution can have a huge set of collision instants (more than countable, but null measure);
- there may be accumulation of partial collisions at a collision having more bodies;
- there is no a priori bound on the total action on the whole interval (a,b), nor on the energy;
- we have very weak assumptions on the potential and only a one side inequality on the radial component
- the theorem extends to the logarithmic potentials.

Conservation Laws

Even though they can not satisfy the differential equations in a distributional sense, generalized solutions satisfy a number of *conservation laws*.

Theorem 17 *Let x be a generalized solutions on (a,b). Then*

- *The action $\mathcal{A}(x,[a,b])$ on (a,b) is finite*
- *The energy h is bounded and belongs to the Sobolev space $W^{1,1}((a,b),\mathbb{R})$.*
- *Lagrange–Jacobi inequality holds in the sense of measures:*

$$\frac{1}{2}\ddot{I}(\bar{x}(t)) \geq 2h(t)(\bar{x}(t)) + (2-\alpha)U(t,\bar{x}(t)) - C_2|\bar{x}|^{\gamma}U(t,\bar{x}), \quad \forall t \in (a,b).$$

- *A monotonicity formula holds (extending Sundman's inequality): let us consider the angular energy*

$$\Gamma_\alpha(r,s) := r^\alpha\left[\frac{1}{2}r^2|\dot{s}|^2 - U(t,rs)\right]$$

the Γ_α is bounded variation and

$$\frac{d}{dt}\Gamma_\alpha(r,s) \geq -\frac{2-\alpha}{2}r^{1+\alpha}\dot{r}|\dot{s}|^2 - C_1 r^\alpha U(t,rs) + C_2 r^{\alpha+\gamma}\frac{\dot{r}}{r}U(t,rs).$$

Generalized Sundman–Sperling Estimates

The asymptotic analysis along a single collision (total or partial) trajectory goes back, in the classical case, to the works by Sundman ([84]), Wintner ([89]) and, in more recent years by Sperling, Pollard, Saari, Diacu and other authors (see for instance [40,47,75,76,79,83]). Now we are in a position to extend such estimate also to generalized solutions.

Theorem 18

- *The following asymptotic estimates hold:*

$$r \sim (\kappa t)^{\frac{2}{2+\alpha}} \quad (r \sim |t|\sqrt{-\log(|t|)})$$
$$K \sim U \sim \frac{1}{4-2\alpha}\ddot{I} \sim \frac{2}{(2+\alpha)^2}\kappa^2(\kappa t)^{\frac{-2\alpha}{2+\alpha}} \quad (\sim -\log|t|).$$

- *Let s be the normalized configuration of the colliding cluster $s = x/r$. Then*

$$\lim_{t\to 0} r^2|\dot{s}|^2 = 0$$
$$\lim_{t\to 0} U(s(t)) = b < +\infty$$

- *Moreover there exists the angular blow-up, that is angular scaled family $(s(\lambda t))_\lambda$ is precompact for the topology of uniform convergence on compact sets of $\mathbb{R}\backslash\{0\}$.*

As a further consequence we have the *vanishing of the total angular momentum.*

In the same setting of the Theorem, assume that the potential U verifies the further assumption:

[U4] $\lim_{r\to 0} r^{\alpha+1}\nabla_T U(t,x) = \nabla_T \tilde{U}(t,s)$.

Then we have

$$\lim_{t\to t_0}\operatorname{dist}(C^b, s(t)) = \lim_{t\to t_0}\inf_{\bar{s}\in C^b}|s(t)-\bar{s}| = 0,$$

where C^b is the set of central configurations for $\tilde{U}$ at level b.

Dissipation and McGehee Coordinates

The main tool in proving the asymptotic estimates follow is the *monotonicity formula*, which is somehow equivalent to *Sundman's inequality*. To fix our mind, let us think to a homogeneous potential U_α. A possible way to to see this dissipation is to perform the change of variables (reminiscent of the McGehee coordinates):

$$\rho = r^{\frac{2-\alpha}{4}}, \quad \rho' = \frac{2-\alpha}{4}r^{-\frac{2+\alpha}{4}}r'$$

to obtain the action functional depending on (ρ, s);

$$\mathcal{A}(\rho,s) = \int_0^{\tau^*}\frac{1}{2}\left(\frac{4}{2-\alpha}\right)^2(\rho')^2 + \rho^2\left(\frac{1}{2}|s'|^2 + U_\alpha(s)\right) - \lambda\rho^\beta\,\mathrm{d}\tau,$$

where $\beta := 2(2+\alpha)/(2-\alpha) > 2$. Here we have reparametrized the time as

$$\mathrm{d}t = r^{\frac{2+\alpha}{2}}\mathrm{d}\tau,$$

one proves that

$$\tau^* = \int_0^1 r^{-\frac{2+\alpha}{2}}\mathrm{d}t = +\infty.$$

This is the coupling between a Duffing equation and the N-body angular system. When ρ is increasing, it acts as a viscosity coefficient on the angular Lagrangian.

A classical framework for the study of collisions is given by the McGehee coordinates [65] (here and below we assume, for simplicity of notations, all the masses be equal to one):

$$r = |x| = I^{1/2}$$
$$s = \frac{x}{r}$$
$$v = r^{\alpha/2}(y\cdot s)$$
$$u = r^{\alpha/2}(y - y\cdot ss).$$

The equation of motions become (here $'$ denotes differentiation with respect to the new time variable τ):

$$\begin{aligned} r' &= rv \\ v' &= \frac{\alpha}{2}v^2 + |u|^2 - \alpha U_\alpha(s) \\ s' &= u \\ u' &= \left(\frac{\alpha}{2} - 1\right) vu - |u|^2 s + \alpha U_\alpha(s)s + \nabla U_\alpha(s)\,. \end{aligned}$$

It is worthwhile noticing that the three last equations do not depen on r and hence are defined also for $r = 0$. In this constext, the monotonicity formula means that v is a Lyapunov function for the system.

Blow-ups

For every $\lambda > 0$ let

$$x^\lambda(t) = \lambda^{-2/(2+\alpha)} x(\lambda t)\,.$$

If $\{\lambda_n\}_n$ is a sequence of positive real numbers such that $s(\lambda_n)$ converges to a normalized configuration $\bar{s}$, then

$$\forall t \in (0,1)\colon \lim_{n\to\infty} s(\lambda_n t) = \lim_{n\to\infty} s(\lambda_n) = \bar{s}\,.$$

Hence the rescaled sequence will converge uniformly to the *blow-up* of $x(t)$ relative to the colliding cluster $\mathbf{k} \subset \mathbf{n}$ (in $t = 0$).

- The blow-up $\bar{x}$ is parabolic: where a *parabolic collision trajectory* for the cluster $\mathbf{k}$ is the path

  $$\bar{x}_i(t) = |t|^{2/(2+\alpha)}\xi_i\,, \quad i \in \mathbf{k}\,, \quad t \in \mathbb{R}\,,$$

  where $\xi = (\xi_i)_{i\in\mathbf{k}}$ is a central configuration with k bodies.

Proposition 8 *The sequences x^{λ_n} and $(dx^{\lambda_n})/(dt)$ converge to the blow-up $\bar{x}$ and its derivative $\dot{\bar{x}}$ respectively, in the H^1–topology. Moreover $\bar{x}$ is a minimizing trajectory in the sense of Morse.*

$$\int_0^T [\mathcal{L}(\bar{x} + \varphi) - \mathcal{L}(\bar{q}) \geq 0]dt\,.$$

for any compactly supported variation φ.

Logarithmic Type Potentials

It is not possible to define a blow-up suitable for logarithmic type potentials. Indeed, the natural scaling should be $\bar{x}^{\lambda_n}(t) := \lambda_n^{-1}\bar{x}(\lambda_n t)$, which does not converge, since, by the asymptotics $|x(t)| \simeq |t|\sqrt{-\log|t|}$, we have:

$$\lim_{\lambda_n\to 0} |\bar{x}^{\lambda_n}(t)| = \lim_{\lambda_n\to 0} \frac{r(\lambda_n t)}{\lambda_n t\sqrt{-2M(0)\log(\lambda_n t)}} \times t\sqrt{-2M(0)\log(\lambda_n t)} = +\infty$$

for every $t > 0$.

On the other, hand, looking at the differential equation, the (right) blow-up should be:

$$\bar{q}(t) := t\bar{s}\,, \quad i \in \mathbf{k}\,,$$

where $\bar{s}$ is a central configuration for the system limit of a sequence $s(\lambda_n)$ where $(\lambda_n)_n$ is such that $\lambda_n \to 0$. This limiting function is the pointwise limit of the normalized sequence

$$\bar{x}^{\lambda_n}(t) := \frac{1}{\lambda_n\sqrt{-2M(0)\log\lambda_n}}\bar{x}(\lambda_n t)\,.$$

Unfortunately this path is not locally minimal for the limiting problem, indeed since, the sequence $(\ddot{\bar{x}}^{\lambda_n})_n$ converges to 0 as n tends to $+\infty$, and hence this blow-up minimizes only the kinetic part of the action functional.

Absence of Collision for Locally Minimal Paths

As a matter of fact, solutions to the Newtonian n-body problem which are minimals for the action are, very likely, free of any collision. This fact was observed by the construction of suitable local variation arguments for the 2 and 3-body cases by Serra and Terracini [81,82]. The 4-body case was treated afterward by Dell'Antonio ([35], with a non completely rigorous argument) and then by A. Venturelli [87]. In general, the proof goes by the sake of the contradiction and involves the construction of a suitable variation that lowers the action in presence of a collision. A recent breakthrough in this direction is due of the neat idea, due to C. Marchal, [63], of averaging over a family of variations parametrized on a sphere. The method of averaged variations for Newtonian potentials has been outlined and exposed by Chenciner [25], and then fully proved and extended to α-homogeneous potentials and various constrained minimization problems by Ferrario and Terracini in [52]. This technique can be used in many of the known cases to prove that minimizing trajectories are collisionless.

Of course, in some specific situations, other arguments can be useful, such as level estimates, on the infimum of the action on colliding paths [17,20,21,25]. However, these argument require global conditions on the potentials and

can not be applied in the present setting, where we work under local assumptions about the singularities.

We are in a position to prove the absence of collisions for locally minimal solutions when the potentials have quasi-homogeneous or logarithmic singularities. The first case is simpler, because one can take advantage the blow-up technique already exploited in [52]. On the other hand, when dealing with logarithmic potentials, the blow-up technique is no longer available and we conclude proving directly some averaging estimates that can be used to show the nonminimality of large classes of colliding motions. The following results are exposed in [15].

Quasi-Homogeneous Potentials

Let $\tilde{U}$ be the C^1 function defined on $(a,b)\times(\mathbb{R}^k\backslash\Delta)$ in the following way:

$$\tilde{U}(t,x) = |x|^{-\alpha}\lim_{r\to 0} r^{\alpha}\tilde{U}(t,rx/|x|)\,.$$

With a slight abuse of notation, we denote $\tilde{U}(x) = \tilde{U}(t^*,x)$. $\tilde{U}$ is homogeneous of degree $-\alpha$.

[U6] there is a 2-dimensional linear subspace of $V^{\perp}_{\mu(\xi)}$, say W, where $\tilde{U}$ is rotationally invariant:

$$\tilde{U}(e^{i\theta}w) = \tilde{U}(w)\,,\quad \forall w\in W\,,\quad \forall\theta\in[0,2\pi]\,;$$

[U7] for every $x\in\mathbb{R}^k$ and $\delta\in W$ there holds:

$$\tilde{U}(x+\delta)\le\tilde{U}\left(\left(\frac{\tilde{U}(\pi_W(x))}{\tilde{U}(x)}\right)^{1/\alpha}\pi_W(x) + \left(\frac{\tilde{U}(x)}{\tilde{U}(\pi_W(x))}\right)^{1/\alpha}\delta\right),$$

where π_W denotes the orthogonal projection onto W.

Theorem 19 *In addition to [U0], [U1], [U2$_h$], [U3$_h$], [U4$_h$], [U5], assume that, for all $\xi\in\Delta$ [U6] and [U7] hold. Then generalized solutions do not have collisions at the time t^*.*

Remark 2 As our potential $\tilde{U}$ is homogeneous of degree $-\alpha$ the function

$$\varphi(x) = \tilde{U}^{-1/\alpha}(x)$$

is a non negative, homogeneous of degree one function, having now Δ as zero set. In most of our applications φ will be indeed a quadratic form. Assume that φ^2 splits in the following way:

$$\varphi^2(x) = K|\pi_W(x)|^2 + \varphi^2(\pi_{W^\perp}(x))$$

for some positive constant K. Then [U6] and [U7] are satisfied. Indeed, denoting $w=\pi_W(x)$ and $z=x-w$ we have, for every $\delta\in W$,

$$\begin{aligned}\varphi^2(x+\delta) &= K|w+\delta|^2+\varphi^2(z)\\ &= K\left|\frac{\varphi(x)}{\varphi(w)}w+\frac{\varphi(w)}{\varphi(x)}\delta\right|^2 + K\frac{\varphi^2(z)}{\varphi^2(x)}|\delta|^2\\ &\ge K\left|\frac{\varphi(x)}{\varphi(w)}w+\frac{\varphi(w)}{\varphi(x)}\delta\right|^2\\ &= \varphi^2\left(\frac{\varphi(x)}{\varphi(w)}w+\frac{\varphi(w)}{\varphi(x)}\delta\right),\end{aligned}$$

which is obviously equivalent to [U7].

Proposition 9 *Assume $\tilde{U}(x)=\mathcal{Q}^{-\alpha/2}(x)$ for some non negative quadratic form $\mathcal{Q}(x)=\langle Ax,x\rangle$. Then assumptions [U6] and [U7] are satisfied whenever W is included in an eigenspace of A associated with a multiple eigenvalue.*

Given two potentials *satisfying [U6] and [U7] for a common subspace W*, their sum enjoys the same properties. On the other hand, the class of potentials satisfying [U6] and [U7] *is not stable with respect to the sum of potentials*. In order to deal with a class of potentials which is closed with respect to the sum, we introduce the following variant of the last Theorem.

Theorem 20 *In addition to [U0], [U1], [U2h], [U3h], [U4h], [U5], assume that $\tilde{U}$ has the form*

$$\tilde{U}(x) = \sum_{\nu=1}^{N}\frac{K_\nu}{(dist(x,V_\nu))^\alpha}\,,$$

where K_ν are positive constants and V_ν is a family of linear subspaces, with $codim(V_\nu)\ge 2$, for every $\nu=1,\dots,N$. Then locally minimizing trajectories do not have collisions at the time t^*.

These two theorems extend to logarithmic potentials.

Neumann Boundary Conditions and *G*-equivariant Minimizers

Our analysis allows to prove that minimizers to the fixed-ends (Bolza) problems are free of collisions: indeed all the variations of our class have compact support. However, other type of boundary conditions (generalized Neumann) can be treated in the same way. Indeed, consider a trajectory which is a (local) minimizer of the action among all paths satisfying the boundary conditions

$$x(0)\in X^0\,,\quad x(T)\in X^1\,,$$

where X^0 and X^1 are two given *linear subspaces of the configuration space*. Consider a (locally) minimizing path $\bar{x}$: of course it has not interior collisions. *In order to exclude boundary collisions we have ensure that the class of variations preserve the boundary conditions.* This can be achieved by imposing assumptions [U6] and [U7] to be fulfilled also by the restriction of the potential to the boundary subspaces X^i. The analysis of boundary conditions was a key point in the paper [52], where symmetric periodic trajectories where constructed by reflections about given subspaces. Our results can be used to prove the absence of collisions also for G-equivariant (local) minimizers, provided the group G satisfies the Rotating Circle Property defined in Definition 3. Hence, existence of G-equivariant collisionless periodic solutions can be proved for the wide class of symmetry groups described in [14,52], for a much larger class of interacting potentials, including quasi-homogeneous and logarithmic ones. On the other hand, our results can be applied to prove that G-equivariant minimals are collisionless for many relevant symmetry groups violating the rotating circle property, such as the groups of rotations recently introduced in [50,51].

The Standard Variation

In order to prove that local minimizers are free of collisions we are going to make use of the following class of variations:

Definition 7 The *standard variation* associated to $\delta \in \Xi$ and T is defined as

$$v^\delta(t) = \begin{cases} \delta & \text{if } 0 \le |t| \le T - |\delta| \\ (T-t)\frac{\delta}{|\delta|} & \text{if } T - |\delta| \le |t| \le T \\ 0 & \text{if } |t| \ge T \end{cases} .$$

Our next goal is to find a standard variation v^δ such that the trajectory $q + v^\delta$ does not have a collision at $t = 0$ and

$$\Delta\mathcal{A} := \int_{-\infty}^{+\infty} [\mathcal{L}_{\mathbf{k}}(q + v^\delta) - \mathcal{L}_{\mathbf{k}}(q)]\mathrm{d}t < 0 .$$

Let us introduce the function

$$S(\xi, \delta) = \int_0^{+\infty} \left(\frac{1}{|\xi t^{2/(2+\alpha)} - \delta|^\alpha} - \frac{1}{|\xi t^{2/(2+\alpha)}|^\alpha} \right) \mathrm{d}t ,$$

where $\xi, \delta \in \mathbb{R}^2$. We first estimate the variation of the action as $\delta \to 0$:

Theorem 21 *Let $q = \{q\}_i = \{t^{2/(2+\alpha)}\xi_i\}$, $i = 1, \ldots, k$ be a parabolic collision trajectory and v^δ a standard variation. Then, as $\delta \to 0$*

$$\Delta\mathcal{A} = 2|\delta|^{1-\alpha/2} \sum_{\substack{i<j \\ i,j\in\mathbf{k}}} m_i m_j S\left(\xi_i - \xi_j, \frac{\delta_i - \delta_j}{|\delta|}\right) + O(|\delta|).$$

We observe that

$$S(\lambda\xi, \mu\delta) = |\lambda|^{-1-\alpha/2}|\mu|^{1-\alpha/2} S(\xi, \delta)$$

and hence the sign of S depends on the angle between ξ and δ. Let

$$\Phi(\vartheta) = \int_0^{+\infty} \frac{1}{\left(t^{\frac{4}{\alpha+2}} - 2\cos\vartheta\, t^{\frac{2}{\alpha+2}} + 1\right)^{\alpha/2}} - \frac{1}{t^{\frac{2\alpha}{\alpha+2}}} \mathrm{d}t , \quad \alpha \in (0, 2) .$$

$\Phi(\theta)$ represents the *potential differential* needed for displacing the colliding particle from zero to $\mathrm{e}^{\mathrm{i}\theta}$. Expanding, we can express $\Phi(\theta)$ as a hypergeometric function:

$$\Phi(\vartheta) = \frac{\alpha(\alpha+2)}{2} \left\{ \frac{1}{\alpha - 2}\beta\left(\frac{\alpha+2}{4}, \frac{\alpha+2}{4}\right) + \frac{1}{\alpha}\sum_{k=1}^{+\infty} \binom{-\alpha/2}{k} (-1)^k 2^{k-1} (\cos\vartheta)^k \beta \times \left(\frac{\alpha}{4} - \frac{1}{2} + \frac{k}{2}, \frac{\alpha}{4} + \frac{1}{2} + \frac{k}{2}\right) \right\} .$$

Some Properties of Φ

The value of $\Phi(\theta)$ ranges from $+\infty$ to some negative value, depending on α. However, thanks to some harmonic analysis one can prove that suitable *averages* are always negative: the first inequality is particularity useful for dealing with *reflected triple collisions from the Lagrange central configuration*:

$$\Phi\left(\frac{2\pi}{3} + \gamma\right) + \Phi\left(\frac{2\pi}{3} - \gamma\right) < 0 , \quad \forall\gamma \in [0, \pi/2] .$$

A key remark was made by Christian Marchal: being the Newton potential a harmonic map averaging it on a sphere results in a truncation in the interior. In fact, is not so much a matter of harmonicity. A crucial estimate was proved in [FT] about the averages of Φ on circles:

Theorem 22 *For every $\alpha > 0$, $\xi \in \mathbb{R}^3\backslash\{0\}$ and for every circle $\mathbb{S} \subset \mathbb{R}^d$ with center in 0,*

$$\tilde{S}(\xi, \mathbb{S}) = \frac{1}{|\mathbb{S}|}\int_{\mathbb{S}} S(\xi, \delta) d\delta = |\xi|^{-1-\alpha/2}|\delta|^{1-\alpha/2} \frac{1}{2\pi}\int_0^{2\pi} \Phi(\theta) d\theta < 0 .$$

Consider $\xi = x_i - x_j$ and δ ranging in a circle. Then the above inequality implies the following principle, a generalization of Marcha's statement in [25]:

> "it is more convenient (from the point of view of the integral of the potential on the time line) to replace one of the point particles with a homogeneous circle of same mass and fixed radius which is moving keeping its center in the position of the original particle."

Future Directions

This article mainly focuses on the variational approach to the search of selected trajectories to the n-body problem. In our examples, the masses can very well be equal: hence the problem can not be regarded as a small perturbation of a simpler one and a full picture of the dynamics is out of reach. In contrast, the planetary n-body problem deals with systems where one of the bodies (a "star") is much heavier than the others (the "planets") and can be seen as a perturbation of a decoupled systems of one center problems. A small parameter ε represents the order of the ratio between the mass of the planets and that of the star. The system is then called *nearly integrable* as it can be associated with a Hamiltonian of the form:

$$H(I, \varphi) = h(I) + \varepsilon f(I, \varphi) ,$$

where $I \in \mathbb{R}^N$ and $\varphi \in \mathbb{T}^N$ (N is the number of degree of freedom) and ε is a small parameter. In the three body problem we have $N = 4$, but the integrable limit possesses motions lying on $\mathbb{T}^2$ (the product of two Keplerian orbits) so *the integrable limit depends on less action variables than the number of degrees of freedom*. For this reason the system is *properly degenerate*. The integrable limit possesses only quasi-periodic motions: in [56], the question whether these motions survive for positive values of ε is settled. KAM Theory deals with the problem of persistence of such invariant tori but can not be directly applied to the planetary n-body problem, for its degeneracy. Nevertheless, the existence of invariant tori has been recently achieved in the planar and spatial three body problem [8,16,18,57,78] and in the planetary many body problem [19,48].

The future researches will extend these results to some non perturbative settings, finding both regular motions, such as periodic, quasi–periodic or quasi–periodic trajectories and irregular, chaotic ones, through the application of suitable variational methods taking into account of collision trajectories.

Bibliography

Primary Literature

1. Albouy A (1996) The symmetric central configurations of four equal masses. Contemp Math 198:131–135
2. Albouy A (1998) Chenciner A, Le problème des *n* corps et les distances mutuelles. Invent Math 131:151–184
3. Ambrosetti A, Coti Zelati V (1993) Periodic solutions of singular Lagrangian systems. In: Progress in Nonlinear Differential Equations and their Applications, vol 10. Birkhäuser Boston Inc., Boston
4. Ambrosetti A, Coti Zelati V (1994) Non-collision periodic solutions for a class of symmetric 3-body type problems. Topol Meth Nonlin Anal 3(2):197–207
5. Ambrosetti A, Rabinowitz PH (1973) Dual variational methods in critical point theory and applications. J Funct Anal 14:349–381
6. Arioli G, Gazzola F, Terracini S (2000) Minimization properties of Hill's orbits and applications to some *N*-body problems. Ann Inst H Poincaré Anal Non Linéaire 17:(5)617–650
7. Arioli G, Barutello V, Terracini S (2006) A new branch of mountain pass solutions for the choreographical 3-body problem. Commun Math Phys 268(5):439–463
8. Arnold VI (1963) Small denominators and problems of stability of motions in classical and celestial mechanics. Uspehi Naut Nauk 18(6):91–192
9. Bahri A, Rabinowitz PH (1991) Periodic solutions of Hamiltonian systems of 3-body type. Ann Inst H Poincaré Anal Non Linéaire 8(6):561–649
10. Barutello V (2004) On the *n*-body problem, Ph. D thesis, Università di Milano–Bicocca, avaliable at http://www.matapp.unimib.it/dottorato/
11. Barutello V, Secchi S (2008) Morse index properties of colliding solutions to the *n*-body problem. Arxiv:math/0609837, Annales de l'Institut Henri Poincare (C) Non Linear Anal 25:539–565
12. Barutello V, Terracini S (2004) Action minimizing orbits in the *n*-body problem with choreography constraint. Nonlinearity 17:2015–2039
13. Barutello V, Terracini S (2007) A bisection algorithm for the numerical Mountain Pass, NoDEA 14:527–539
14. Barutello V, Ferrario DL, Terracini S (2008) Symmetry groups of the planar 3-body problem and action–minimizing trajectories. Arch Rat Mech Anal 190:189–226
15. Barutello V, Ferrario DL, Terracini S (2008) On the singularities of generalized solutions to the *N*-body problem. Int Math Res Notices 2008:rnn069–78
16. Berti M, Biasco L, Valdinoci E (2004) Periodic orbits close to invariant tori and applications to the three-body problem. Ann Scuola Norm Sup Pisa Cl Sci 5 vol III:87–138
17. Bessi U, Coti Zelati V (1991) Symmetries and noncollision closed orbits for planar *N*-body-type problems. Nonlin Anal 16(6):587–598
18. Biasco L, Chierchia L, Valdinoci E (2003) Elliptic two dimensional invariant tori for the planetary three-body problem. Arch Rat Mech Anal 170:91–135
19. Biasco L, Chierchia L, Valdinoci E (2006) *N*-dimensional elliptic invariant tori for the planetary $(N + 1)$-body problem. SIAMJ Math Anal 37:1560–1588
20. Chen K-C (2001) Action-minimizing orbits in the parallelogram

four-body problem with equal masses. Arch Rat Mech Anal 158:293–318
21. Chen K-C (2001) On Chenciner–Montgomery's orbit in the three-body problem. Discrete Contin Dyn Syst 7(1):85–90
22. Chen K-C (2003) Binary decompositions for planar *n*-body problems and symmetric periodic solutions. Arch Rat Mech Anal 170(3):247–276
23. Chen K-C (2003) Variational methods on periodic and quasi-periodic solutions for the *N*-body problem. Ergodic Theory Dyn Syst 23(6):1691–1715
24. Chenciner A (2002) Action minimizing periodic orbits in the Newtonian n-body problem Celestial Mechanics, dedicated to Don Saari. Contemp Math 292:71–90
25. Chenciner A (2002) Action minimizing solutions of the newtonian *n*-body problem: from homology to symmetry. ICM, Peking
26. Chenciner A (2002) Simple non-planar periodic solutions of the *n*-body problem. In: Proceedings of the NDDS Conference, Kyoto
27. Chenciner A, Desolneux N (1998) Minima de l'intégrale d'action et équilibre relatifs de *n* corps. C R Acad Sci Paris Sér I 326:1209–1212; Correction in: (1998) C R Acad Sci Paris Sér I 327:193
28. Chenciner A, Féjoz J (2005) L'équation aux variations verticales d'un équilibre relatif comme source de nouvelles solutions périodiques du problème des *N* corps. C R Math. Acad. Sci. Paris 340(8):593–598
29. Chenciner A, Montgomery R (2000) A remarkable periodic solution of the three body problem in the case of equal masses. Ann Math 152(3):881–901
30. Chenciner A, Venturelli A (2000) Minima de l'intégrale d'action du problème Newtonien de 4 corps de masses égales dans $\mathbb{R}^3$: orbites "hip–hop". Celest Mech 77:139–152
31. Chenciner A, Féjoz J, Montgomery R (2005) Rotating eights. I. The three Γ_i families. Nonlinearity 18(3):1407–1424
32. Chenciner A, Gerver J, Montgomery R, Simó C (2001) Simple choreographies of *N* bodies: a preliminary study. In: Geometry, Mechanics and Dynamics. Springer, New York, pp 287–308
33. Degiovanni M, Giannoni F (1988) Dynamical systems with newtonian type potentials. Ann. Scuola Norm Sup Pisa, Ser IV 15:467–494
34. Degiovanni M, Giannoni F, Marino A (1987) Dynamical systems with newtonian type potentials. Atti Accad Naz Lincei Rend Cl Sci Fis Mat Natur Ser 8(81):271–278
35. Dell'Antonio G (1998) Non-collision periodic solutions of the *N*-body system. NoDEA Nonlinear Differ Equ Appl 5:1 117–136
36. Devaney RL (1978) Collision orbits in the anisotropic Kepler problem. Invent Math 45(3):221–251
37. Devaney RL (1978) Nonregularizability of the anisotropic Kepler problem. J Differ Equ 29(2):252–268
38. Devaney RL (1980) Triple collision in the planar isosceles three-body problem. Invent Math 60(3):249–267
39. Devaney RL (1981) Singularities in classical mechanical systems. in: Ergodic theory and dynamical systems, I. College Park, Md., 1979–80, vol 10 of Progr Math. Birkhäuser, Boston, pp 211–333
40. Diacu F (1992) Regularization of partial collisions in the *N*-body problem. Differ Integral Equ 5(1):103–136
41. Diacu F (1993) Painlevé's conjecture. Math Intell 15(2):6–12
42. Diacu F (1996) Near-collision dynamics for particle systems with quasihomogeneous potentials. J Differ Equ 128(1):58–77
43. Diacu F (2002) Singularities of the *N*-body problem. in: Classical and celestial mechanics. Princeton Univ. Press, Princeton, pp 35–62, Recife, 1993/1999
44. Diacu F, Santoprete M (2004) On the global dynamics of the anisotropic Manev problem. Phys D 194(1–2):75–94
45. Diacu F, Pérez-Chavela E, Santoprete M (2006) Central configurations and total collisions for quasihomogeneous *n*-body problems. Nonlinear Anal 65(7):1425–1439
46. Diacu F, Pérez-Chavela E, Santoprete M (2005) The Kepler problem with anisotropic perturbations. J Math Phys 46(7):072701, 21
47. ElBialy MS (1990) Collision singularities in celestial mechanics. SIAM J Math Anal 21(6):1563–1593
48. Féjoz J (2004) Démonstration du "théorème d'Arnold" sur la stabilité du sistème planetaire (d'aprés Michel Herman). (french) [Proof of "Arnold's theorem" on the stability of a planetary system (following Michel Herman)], Ergodic Theory Dyn Syst 24(5):1521–1582
49. Ferrario DL (2002) Symmetric periodic orbits for the *n*-body problem: some preliminary results, Preprint of the Max-Planck-Institut für Mathematik MPI-2002-79
50. Ferrario DL (2007) Transitive decomposition of symmetry groups for the *n*-body problem. Adv Math 213:763–784
51. Ferrario DL (2006) Symmetry groups and non-planar collisionless action-minimizing solutions of the three-body problem in three-dimensional space. Arch Rat Mech Anal 179(3):389–412
52. Ferrario D, Terracini S (2004) On the existence of collisionless equivariant minimizers for the classical *n*-body problem. Invent Math 155(2)305–362
53. The GAP Group (2002) GAP – Groups, Algorithms, and Programming, Version 4.3, http://www.gap-system.org
54. Gordon WB (1975) Conservative dynamical systems involving strong forces. Trans Am Math Soc 204:113–135
55. Gordon WB (1977) A minimizing property of Keplerian orbits. Am J Math 99(5):961–971
56. Jefferys WH, Moser J (1966) Quasi-periodic solutions for the three-body problem. Celest Mech Dyn Astron J 71:508–578
57. Laskar J, Robutel P (1995) Stability of the planetary three-body problem, I: Expansion of the planetary Hamiltonian. Celest Mech Dyn Astron 62(3):193–217
58. Levi Civita T (1918) Sur la régularization du problème des trois corps. Acta Math 42:99–144
59. Majer P, Terracini S (1993) Periodic solutions to some problems of *n*-body type. Arch Rat Mech Anal 124(4):381–404
60. Majer P, Terracini S (1995) On the existence of infinitely many periodic solutions to some problems of *n*-body type. Commun Pure Appl Math 48(4):449–470
61. Majer P, Terracini S (1995) Multiple periodic solutions to some *n*-body type problems via a collision index Variational methods in nonlinear analysis (Erice, 1992). Gordon and Breach, Basel, pp 245–262
62. Marchal C (2000) The family P_{12} of the three-body problem – the simplest family of periodic orbits, with twelve symmetries per period. Celest Mech Dyn Astronom 78(1–4):279–298; (2001) New developments in the dynamics of planetary systems. Badhofgastein, 2000
63. Marchal C (2002) How the method of minimization of action avoids singularities. Celest Mech Dyn Astron 83:325–353
64. Mather JN, McGehee R (1974) Solutions of the collinear four body problem which become unbounded in finite time. In: Dynamical systems, theory and applications. Rencontres, Battelle

Res Inst, Seattle, Wash., pp 573–597. Lecture Notes in Phys., vol 38. Springer, Berlin, (1975)
65. McGehee R (1974) Triple collision in the collinear three-body problem. Invent Math 27:191–227
66. McGehee R (1986) von Zeipel's theorem on singularities in celestial mechanics. Exposition Math 4(4):335–345
67. Moeckel R (1990) On central configurations, Math Zeit 205:499–517
68. Moeckel R (1987) Some qualitative features of the three-body problem. in: Hamiltonian dynamical systems. (Boulder, 1987) vol 81 of Contemp Math, pp 1–22. Am Math Soc, Providence RI, 1988
69. Montgomery R (1998) The *N*-body problem, the braid group, and action-minimizing periodic solutions. Nonlinearity 11(2):363–376
70. Montgomery R (1999) Action spectrum and collisions in the planar three-body problem. In: Celestial Mechanics. (Evanston, 1999) vol 292 of Contemp Math Am Math Soc, Providence RI, 2002, pp 173–184
71. Moore C (1993) Braids in classical dynamics. Phys Rev Lett 70(24):3675–3679
72. Pacella F (1987) Central configurations and the equivariant Morse theory. Arch Rat Mech 97:59–74
73. Palais RS (1979) The principle of symmetric criticality. Commun Math Phys 69:19–30
74. Poincaré H (1896) Sur les solutions périodiques et le principe de moindre action. C R Acad Sci Paris, Sér I Math 123:915–918
75. Pollard H, Saari DG (1968) Singularities of the *n*-body problem I. Arch Rat Mech Anal 30:263–269
76. Pollard H, Saari DG (1970) Singularities of the *n*-body problem II. In: Inequalities II. Academic Press, New York, pp 255–259 (Proc. Second Sympos, US Air Force Acad, Colo, 1967)
77. Riahi H (1999) Study of the critical points at infinity arising from the failure of the Palais-Smale condition for *n*-body type problems. Mem Am Math Soc 138:658, viii+112
78. Robutel P (1995) Stability of the planetary three-body problem, II: KAM theory and existence of quasi-periodic motions. Celest Mech Dyn Astron 62(3):219–261
79. Saari DG (1972/73) Singularities and collisions of Newtonian gravitational systems. Arch Rat Mech Anal 49:311–320
80. Sbano L (1998) The topology of the planar three-body problem with zero total angular momentum and the existence of periodic orbits. Nonlinearity 11(3):641–658
81. Serra E, Terracini S (1992) Collisionless periodic solutions to some three-body problems. Arch Rat Mech Anal 120(4):305–325
82. Serra E, Terracini S (1994) Noncollision solutions to some singular minimization problems with Keplerian-like potentials. Nonlinear Anal 22(1):45–62
83. Sperling HJ (1970) On the real singularities of the *N*-body problem. J Reine Angew Math 245:15–40
84. Sundman KF (1913) Mémoire sur le problème des trois corps. Acta Math 36:105–179
85. Terracini S, Venturelli A (2007) Symmetric trajectories for the 2*N*-body problem with equal masses. Arch Rat Mech Anal 184(3):465–493
86. Venturelli A (2001) Une caractérisation variationnelle des solutions de Lagrange du problème plan des trois corps, C R Acad Sci Paris, Sér I Math 332(7):641–644
87. Venturelli A (2002) Application de la minimisation de l'action au Problème des *N* corps dans le plan et dans l'espace. Thesis. University Paris VII
88. Wang Q (1991) The global solution of the *n*-body problem, Celest Mech Dyn Astron 50(1):7388
89. Wintner A (1941) The analytical foundations of celestial mechanics. Princeton Mathematical Series, vol 5. Princeton University Press, Princeton
90. Xia Z (1992) The existence of non collision singularities in newtonian systems. Ann Math 135:411–468
91. Von Zeipel H (1908) Sur les singularités du problème des *n* corps. Ark Math Astr Fys 4:1–4

Books and Reviews

Arnold VI, Kozlov V, Neishtadt A (2006) Mathematical aspects of classical and celestial mechanics [Dynamical systems. III]. 3rd edn. In: Encyclopaedia of Mathematical Sciences, vol 3. Springer, Berlin, xiv and p 518, translated from the Russian original by Khukhro E

Diacu F, Holmes P (1996) Celestial encounters. The origins of chaos and stability. Princeton University Press, Princeton

Meyer, Kenneth R (1999) Periodic solutions of the *N*-body problem, Lecture Notes in Mathematics, 1719. Springer, Berlin

Moser J (1973) Stable and random motions in dynamical systems, With special emphasis on celestial mechanics, Hermann Weyl Lectures, the Institute for Advanced Study, Princeton NJ, Annals of Mathematics Studies, No. 77. Princeton University Press, Princeton

Pollard H (1976) Celestial mechanics, Carus Mathematical Monographs, 18. Mathematical Association of America, Washington

Saari D (2005) Collisions, rings, and other Newtonian *N*-body problems, CBMS Regional Conference Series in Mathematics, 104. American Mathematical Society, Providence RI, Washington, Published for the Conference Board of the Mathematical Sciences

Siegel CL, Moser JK (1995) Lectures on celestial mechanics. Classics in Mathematics. Springer, Berlin, Translated from the German by Kalme CI, Reprint of the 1971 translation

Stiefel EL, Scheifele G (1971) Linear and regular celestial mechanics. Perturbed two-body motion, numerical methods, canonical theory. Die Grundlehren der mathematischen Wissenschaften, Band 174. Springer, New York, Heidelberg

Nekhoroshev Theory

Laurent Niederman[1,2]
[1] Topologie et Dynamique – UMR 8628 du CNRS, Université Paris, Paris, France
[2] Astronomie et Systèmes Dynamiques – UMR 8028 du CNRS, IMCCE, Paris, France

Article Outline

Glossary
Definition of the Subject
Introduction

Exponential Stability of Constant Frequency Systems
Nekhoroshev Theory (Global Stability)
Applications
Future Directions
Appendix: An Example of Divergence Without Small Denominators
Bibliography

Glossary

Quasi integrable Hamiltonian A Hamiltonian is quasi integrable if it is close to another Hamiltonian whose associated system is integrable by quadrature. Here, we will consider real analytic Hamiltonians on a domain $\mathcal{D}$ which admit a holomorphic extension on a complex strip $\mathcal{D}_{\mathbb{C}}$ around $\mathcal{D}$ and the closeness to the integrable Hamiltonian is measured with the supremum norm $||.||_\infty$ over $\mathcal{D}_{\mathbb{C}}$.

Exponentially stable Hamiltonian An integrable Hamiltonian governs a system which admits a collection of first integral. We say that an integrable Hamiltonian h is exponentially stable if for any small enough Hamiltonian perturbation of h, the solutions of the perturbed system are at least defined over timescales which are exponentially long with respect to the inverse of the size of the perturbation. Moreover, the first integrals of the integrable system should remain nearly constant along the solutions of the perturbed system over the same amount of time.

Nekhoroshev theorem (1977) There exists a *generic set* of real analytic integrable Hamiltonians which are *exponential stable*.

Definition of the Subject

We only know explicitly the solutions of the Hamiltonian systems which are integrable by quadrature. Unfortunately these integrable Hamiltonians are exceptional but many physical problems can be studied by a Hamiltonian system which differs from an integrable one by a small perturbation. One of the most famous is the motions of the planets around the Sun which can be seen as a perturbation of the integrable system associated to the motion of noninteracting points around a fixed attracting center. Poincaré [70] considered the study of these quasi integrable Hamiltonian systems as the *Problème général de la dynamique*. This question was tackled with the Hamiltonian perturbation theories which were introduced at the end of the nineteenth century precisely to study the planetary motions (see ► Hamiltonian Perturbation Theory (and Transition to Chaos)).

Significant theorems concerning the mathematical justification of these methods were not put forward until the 1950s and later. A cornerstone of these results is the Kolmogorov–Arnold–Moser (KAM) theory which states that under suitable nondegeneracy and smoothness assumptions, for a small enough perturbation most of the solutions of a nearly integrable Hamiltonian system are quasi periodic, hence they are stable over infinite times. More accurately, KAM theory provides a set of large measure of invariant tori which support stable and global solutions of the perturbed system. But this set of stable quasi periodic solutions has a complicated topology: it is a Cantor set nowhere dense and without interior points. Numerically, it is extremely difficult to determine if a given solution is quasi periodic or not.

Moreover, for a n degrees of freedom Hamiltonian system (hence a $2n$-dimensional system) KAM theory provides n-dimensional tori. If $n = 2$, these two-dimensional invariant tori divide the three-dimensional energy level, therefore the solutions of the perturbed system are global and bounded over *infinite* times but an arbitrary large drift of the orbits is still possible for $n \geq 3$. Actually, Arnold [1] proposed examples of quasi integrable Hamiltonian systems where an arbitrary large instability occurs for an arbitrary small perturbation. This is known as "Arnold diffusion".

Thus, results of stability for quasi integrable Hamiltonian systems which are valid for an open set of initial condition can only be proved over *finite* times.

The first theorems of effective stability over finite but very long times were proved by Moser [58] and Littlewood [45] around an elliptic equilibrium point in an Hamiltonian system. Extending this kind of result in a general framework, Nekhoroshev [62,63] proved in 1977 a fundamental theorem of global stability which completes KAM theory and states that, for a *generic set* of integrable Hamitonian, Arnold diffusion can only occur over exponentially long times with respect to the inverse of the size of the perturbation.

These theories have been applied extensively in celestial mechanics but also to study molecular dynamics, beam dynamics, billiards, geometric numerical integrators, stability of certain solutions of nonlinear PDE (Schrodinger, wave, ...).

Introduction

We first recall that the canonical equations with a given scalar function $\mathcal{H}$ (the Hamiltonian) which is differentiable on an open set $\mathcal{D} \subset \mathbb{R}^{2n}$ equipped with the symplectic form of Liouville $\omega = \sum_i \mathrm{d}p_i \wedge \mathrm{d}q_i$ for $(q_1, \ldots, q_n) \in$

$\mathbb{R}^n$ and their conjugate coordinates $(p_1, \ldots, p_n) \in \mathbb{R}^n$ can be written:

$$\dot{p} = -\partial_q \mathcal{H}(p,q)\,; \quad \dot{q} = \partial_p \mathcal{H}(p,q)\,.$$

Actually, this kind of system can be defined over any symplectic manifold.

Moreover, a diffeomophism Φ is *symplectic* or *canonical* if it preserves Liouville form: $\Phi^*\omega = \omega$. Especially, for a Hamiltonian X, the flow Φ_X^t linked to the canonical system governed by X is a symplectic transformation over its domain of definition [54]. Such a diffeomorphism preserves the Hamilton equations: in the new variables $(p,q) = \Phi(P,Q)$ the transformed system is still canonical with the initial Hamiltonian $K(P,Q) = \mathcal{H} \circ \Phi(P,Q)$.

Integrable and Quasi-integrable Hamiltonian Systems

According to the theorem of Liouville–Arnold, under general topological conditions, a Hamiltonian system integrable by quadrature can be reduced to a system defined over $\Omega \times \mathbb{T}^n$ for an open set $\Omega \subset \mathbb{R}^n$ and the n-dimensional torus $\mathbb{T}^n$ with a Hamiltonian which does not depend on the n angles. The new variables $(I,\theta) \in \Omega \times \mathbb{T}^n$ are called *actions-angle* variables and the system is linked to a Hamiltonian $K(I)$, hence the equations take the trivial form $\dot{I} = 0, \dot{\theta} = \nabla K(I)$ where ∇K is the gradient of K and we obtain quasi-periodic motions of frequencies $\nabla K(I)$.

If a small Hamiltonian perturbation is added to an integrable system, then in action-angle variables the considered system is governed by

$$\mathcal{H}(\varepsilon, I, \theta) = h(I) + \varepsilon f(I,\theta) \quad \text{where } \mathcal{H} \in C^\omega(\Omega \times \mathbb{T}^n, \mathbb{R}) \tag{1}$$

and the equations become:

$$\dot{I} = -\varepsilon \partial_\theta f(I,\theta)\,; \quad \dot{\theta} = \nabla h(I) + \varepsilon \partial_I f(I,\theta)\,. \tag{2}$$

Hence, the variables are separated into two groups: the fast variables and the other variables which evolve slowly (the actions but also the angles with zero frequencies), over times of order $1/\varepsilon$ their evolution may be of order 1.

Averaging Principle

The *averaging principle* consists of the replacement of the initial system by its time average along the unperturbed flow Φ_h^t linked to h which means that we consider the averaged Hamiltonian:

$$\langle \mathcal{H} \rangle(I,\theta) = h(I) + \varepsilon \langle f \rangle(I,\theta)$$
$$\text{with} \langle f \rangle(I,\theta) = \lim_{t\to\infty} \left(\frac{1}{t} \int_0^t f(I, \theta + s\nabla h(I)) \mathrm{d}s \right)\,.$$

Actually, this average depends on the commensurability relations which are satisfied by the components of vector $\nabla h(I)$.

More accurately, to a submodule $\mathcal{M} \subset \mathbb{Z}^n$, we associate its *resonant zone*:

$$\mathcal{Z}_{\mathcal{M}} = \{I \in \mathbb{R}^n \text{ such that } k\,.\nabla h(I) = 0 \text{ if and only if } k \in \mathcal{M}\}\,. \tag{3}$$

Let $r \in \{0, \ldots, n-1\}$ be the rank of $\mathcal{M}$, there exists a unimodular matrix $\mathcal{R} \in \mathrm{SL}(n,\mathbb{Z})$ such that $\omega \in \mathcal{M}^\perp$ if and only if the first r lines of $\mathcal{R}\omega$ are zeros. Hence, the symplectic transformation $\phi = \mathcal{R}\theta$ and $I = {}^t\mathcal{R}^{-1}J$ defined over $\Omega \times \mathbb{T}^n$ yields the new Hamiltonian $\widetilde{h}(J) + \varepsilon \widetilde{f}(J,\phi)$ where $J = (J_1, J_2) \in \mathbb{R}^r \times \mathbb{R}^{n-r}$ and $\phi = (\phi_1, \phi_2) \in \mathbb{T}^r \times \mathbb{T}^{n-r}$ such that $\nabla \widetilde{h}(J) = (0, \omega(J))$ when ${}^t\mathcal{R}^{-1}J \in \mathcal{Z}_{\mathcal{M}}$. Moreover, the second component $\omega(J) \in \mathbb{R}^{n-r}$ does not satisfy any commensurability relation.

Equivalently, on the set $\mathcal{Z}_{\mathcal{M}}$, we can consider the torus $\mathbb{T}^n$ as the product $\mathbb{T}^r \times \mathbb{T}^{n-r}$ where the unperturbed flow is *constant* over $\mathbb{T}^r$ and *ergodic* over $\mathbb{T}^{n-r}$.

Hence, at infinity, the time average $\langle \widetilde{f} \rangle(J,\phi)$ is equal to the space average:

$$\langle \widetilde{f} \rangle(J_1, J_2, \phi_1) = \left(\frac{1}{2\pi}\right)^{n-r} \iint_{\mathbb{T}^{n-r}} \widetilde{f}(J_1, J_2, \phi_1, \phi_2)\, \mathrm{d}\phi_2\,, \tag{4}$$

and the averaged system involves only variables which evolve slowly. For instance, to determine the solutions numerically, we only need to take a step of integration of order $1/\varepsilon$ [41].

Actually, in the expression (4), we have only made a *partial averaging* by removing $n-r$ angles. Hence, the considered system is simpler but, for a generic perturbation, it is *only* for the zero module $\mathcal{M} = \{0\}$ that we obtain an integrable Hamiltonian.

It should be noticed that, for an arbitrary module $\mathcal{M}$, the solutions of the average system can be unbounded. For instance, the Hamiltonian $\mathcal{H}(I_1, I_2, \theta_1, \theta_2) = I_1 I_2 + \varepsilon \sin(\theta_2)$ is averaged with respect to the module $\mathcal{M} = \mathbb{Z}\vec{\boldsymbol{i}}$ with $\vec{\boldsymbol{i}} = (1,0)$ since θ_1 does not appear in the perturbation and it admits the unbounded solution $(I_1(t), I_2(t), \theta_1(t), \theta_2(t)) = (0, -\varepsilon t, \varepsilon t^2/2, 0)$ which starts from the origin $(0,0,0,0)$ at $t = 0$. One can also notice that we have a maximal speed of drift of the action I_2 with respect to the size of the perturbation.

On the other hand, a key observation for the proof of Nekhoroshev theorem is the fact that for an arbitrary module $\mathcal{M}$, the canonical equations ensure that the variations of the actions under the averaged vector field with respect

to $\mathcal{M}$ are located in the *subspace spanned by* $\mathcal{M}$. This point will be specified in Sect. "The Initial Statement", it will be the only place in Nekhoroshev's proof where the canonical form of the equations is used.

Hamiltonian Perturbation Theory

The averaging principle is based on the idea that the oscillating terms discarded in averaging cause only small oscillations which are superimposed to the solutions of the averaged system. In order to prove the validity of the averaging principle, one should check that any solutions of the perturbed system remain close to the solution of the averaged system with the same initial condition. Especially, this will be the case if one finds a canonical transformation ε-close to identity which transforms the perturbed Hamiltonian to its average. Hence we are reduced to a problem of normal form where one looks for a convenient system of coordinates which gives the simplest possible form to the considered system.

Here, we consider the transformation Φ_X^1 given by the time 1 flow of the Hamiltonian system governed by $X(I,\theta) = \varepsilon X_1(I,\theta)$.

With Taylor formula, the transformed Hamiltonian $\mathcal{H} \circ \Phi_X^1$ admits the following expansion with respect to ε:

$$\mathcal{H} \circ \Phi_X^1 = h + \varepsilon(f + \nabla h(I).\partial_\theta X_1(I,\theta)) + \mathcal{O}(\varepsilon^2)\,,$$

and in order to obtain $\mathcal{H} \circ \Phi_X^1 = h(I) + \varepsilon\langle f\rangle$ one must solve:

$$\nabla h(I).\partial_\theta X_1(I,\theta) = -f + \langle f\rangle\,, \tag{5}$$

which is the *homological equation*, this is the central equation of perturbation theory.

Since we are in the analytic setting, the function f admits the expansion

$$\sum_{k\in\mathbb{Z}^n} f_k(I)\exp(ik\theta)\,.$$

Hence, for an averaging with respect to a resonant module $\mathcal{M}$ around the resonant zone $\mathcal{Z}_{\mathcal{M}}$ linked to $\mathcal{M}$, the homological equation admits the formal solution:

$$X_1(I,\theta) = \sum_{k\notin\mathcal{M}} \frac{f_k(I)}{i\,(k.\nabla h(I))}\exp(ik\theta)\,, \tag{6}$$

thus one obtains a transformation which normalizes the Hamiltonian at first order in ε.

In the same way, one can *formally* eliminate the fast angles at all orders by looking at a transformation generated by a Hamiltonian $X(I,\theta) = \sum_{n\geq 1}\varepsilon^n X_n(I,\theta)$. Indeed, the same type of homological equation appears at all order to determine X_n for $n > 1$.

This is the *Linstedt method*.

With the previous construction, it is obvious that the normalizing transformation admits an expansion which is divergent if the denominators $k.\nabla h(I)$ for $k \notin \mathcal{M}$ are too close to zero on the considered domain in the action space. This is the well-known problem of the *small denominators* which was emphasized by Poincaré [70] in his celebrated theorem about nonexistence of analytic first integrals for a generic quasi integrable Hamiltonian.

It is less known that even without small denominators, the classical perturbation theory can yield divergent expansions. This is the problem of the *great multipliers* according to Poincaré terminology which come from the successive differentiations.

This problem is presented in the next subsection.

Exponential Stability of Constant Frequency Systems

The Case of a Single Frequency System

The normalization of an analytic quasi-integrable Hamiltonian system with only *one fast* phase is one of the main problems in perturbation theory. This question appears naturally to compute the time of approximate conservation of the adiabatic invariants [17,49]. It has been accurately studied in [60] and [73] and we will focus our attention on this case where the phenomenon of divergence without small denominators appears in its simplest setting.

Indeed, as in the Sect. "Hamiltonian Perturbation Theory", one can build *formally* the Hamiltonian $X(I,\theta) = \sum_{n\geq 1}\varepsilon^n X_n(I,\theta)$ which generates a normalizing symplectic transformation and eliminates the fast angle in the perturbation. But, for a generic analytic quasi-integrable Hamiltonian, it can be shown [60,73] that perturbation theory yields a *Gevrey-2* normalizing transformation over $U\times\mathbb{T}^n$ (i. e.: $\mathcal{T} \in C^\infty(U\times\mathbb{T}^n)$ with $||\partial^k\mathcal{T}||_\infty \leq CM^{2|k|}k!^2$ where C, M are positive constants and $k = (k_1,\ldots,k_{2n}) \in \mathbb{N}^{2n}$; $|k| = |k_1| + \ldots + |k_{2n}|$; $k! = k_1!\ldots k_{2n}!$) such that the initial perturbed Hamiltonian $\mathcal{H} = h + \varepsilon f$ is transformed into an integrable Hamiltonian $h_\varepsilon(\widetilde{I})$ with a one parameter family h_ε of scalar functions Gevrey-2 over U.

Hence, the normalizing transformations are usually divergent. On the other hand, by general properties of Gevrey functions (see Sect. 3.3 in [53], and the references therein) if one considers the transformation generated by the truncated expansion

$$\sum_{n\geq 1}^{N}\varepsilon^n X_n(I,\theta)$$

obtained after N steps of perturbation theory, then the transformed Hamiltonian is normalized up to a remainder of size $\varepsilon^{N+1}N!$.

In the appendix, we will study an example of a quasi-integrable Hamiltonian where this latter estimate cannot be improved and where the source of divergence of the normalizing transformation which comes from the successive differentiations in the construction can be emphasized.

We see that the remainder of size $\varepsilon^{N+1}N!$ decreases rapidly before increasing to infinity, following Poincaré [70] this is a "convergent expansion according to astronomers" and a "divergent expansion according to geometers".

Now, we can use the process of "summation at the smallest term": for a *fixed* $\varepsilon > 0$, one obtains an optimal normalization with a truncation at order N such that

$$||\partial_{\theta_2} X_{N-1}||_\infty \simeq \varepsilon ||\partial_{\theta_2} X_N||_\infty \quad \text{which yields} \quad N = E(1/\varepsilon)\,.$$

Finally, the Stirling formula yields the size of the remainder: $\sqrt{2\pi\varepsilon}\exp(-1/\varepsilon)$ which is *exponentially small* with respect to the inverse of the size of the perturbation.

More generally, Marco and Sauzin [53] have proved in the same setting that starting from a Gevrey-α Hamiltonian ($||\partial^k \mathcal{H}(I,\theta)||_\infty \leq CM^{\alpha|k|}(k!)^\alpha$), one can build a normalizing transformation which is Gevrey-α+1 and these estimates cannot be improved usually but the previous construction is still possible. Indeed, one can still make a summation at the smallest term and obtain a canonical change of coordinates which normalizes the Hamiltonian up to an exponentially small remainder with respect to the size of the perturbation.

In the case where the averaged Hamiltonian is *integrable*, according to the mean value theorem, the speed of drift of the normalized action variables is at most *exponentially slow*. Since the size of the normalizing transformation is of order ε, the initial actions admit at most a drift of size ε over an *exponentially long time*.

The Case of a Strongly Nonresonant Constant Frequency System

Systems with constant frequencies, hence $h(I) = \omega.I$ for some constant vector $\omega \in \mathbb{R}^n$, appear when we consider small nonlinear interactions of linear oscillatory systems or the action of quasi-periodic perturbations on linear oscillatory systems. In any case, the considered Hamiltonian can be written: $\mathcal{H} = \omega.I + f(I,\theta)$ with a small function $f \in C^\omega(\Omega \times \mathbb{T}^n, \mathbb{R})$ where Ω is an open set in $\mathbb{R}^n$.

Moreover, we assume here that the frequency ω is a (γ,τ)-*Diophantine vector* for some positive constants γ and τ, hence $\omega \in \Omega_{\gamma,\tau}$:

$$\Omega_{\gamma,\tau} = \left\{ \omega \in \mathbb{R}^n \ \text{ such that } \ |k.\omega| \geq \frac{\gamma}{||k||_\infty^\tau} \ \text{ for all } \ k \in \mathbb{Z}^n \backslash \{(0,\ldots,0\} \right\}. \quad (7)$$

We recall that the measure of the complementary set of $\Omega_{\gamma,\tau}$ is of order $\mathcal{O}(\gamma)$ for $\tau > n-1$.

Under these assumptions, one can prove [13,31,45,71] that for a small enough analytic perturbation, the action variables of the unperturbed problem become quasi integrals of the perturbed system over exponentially long times, more specifically:

Theorem 1 ([13,31,45,71]) *Consider a Hamiltonian $\omega.I + f(I,\theta)$ real analytic over a domain $\mathcal{U} \times \mathbb{T}^n \subset \mathbb{R}^n \times \mathbb{T}^n$ which admit a holomorphic extension on a complex strip of width $\rho > 0$ around $\mathcal{U} \times \mathbb{T}^n$ in $\mathbb{C}^{2n}$.*

The supremum norm for a holomorphic function on this complex strip is denoted $||.||_\rho$.

There exists positive constants C_1, C_2, C_3, C_4 which depend only on γ, τ, ρ, n such that if $\varepsilon = ||f||_\rho < C_1\gamma$, an arbitrary solution $(I(t), \varphi(t))$ of the perturbed system associated to $\omega.I + f(I,\theta)$ with an initial action $I(t_0) \in \mathcal{U}$ is defined at least over an exponentially long time and satisfies:

$$||I(t) - I(0)|| \leq C_2\varepsilon \quad \textit{if} \ |t| \leq C_3 \exp(C_4 \varepsilon^{-\frac{1}{1+\tau}})\,. \quad (8)$$

The proof is based on the existence of a normalizing transformation up to an exponentially small error. This is possible since we have lower bounds on the small denominators and the growth of the coefficients in the normalizing expansion is reduced to a combinatorial problem.

Finally, since the averaged Hamiltonian is integrable, the speed of drift of the action variables is at most exponentially slow.

Nekhoroshev Theory (Global Stability)

The Initial Statement

Thirty years ago, Nekhoroshev [62,63] stated a *global* result of stability which is valid for a *generic* set of integrable Hamiltonian. Especially, we don't have anymore a control on the small denominators as in the previous section but we have to handle the resonant zones. Nekhoroshev's reasonings allow one to prove a global result of stability *inde-*

pendent of the arithmetical properties of the unperturbed frequencies by taking into account the *geometry of the integrable system*. This is really a change of perspective with respect to the previous results. The key ingredient is to find a suitable property of the integrable Hamiltonian, namely the property of *steepness* introduced in the sequel, which ensures that a drift of the actions in the averaged system with respect to a module $\mathcal{M} \subset \mathbb{Z}^n$ leads to an *escape* of the resonant zone $\mathcal{Z}_{\mathcal{M}}$.

More specifically, Nekhoroshev proved global results of stability over open sets of the following type:

Definition 2 (exponential stability) Consider an open set $\Omega \subset \mathbb{R}^n$, an analytic integrable Hamiltonian $h\colon \Omega \longrightarrow \mathbb{R}$ and action-angle variables $(I, \varphi) \in \Omega \times \mathbb{T}^n$ where $\mathbb{T} = \mathbb{R}/\mathbb{Z}$.

For an arbitrary $\rho > 0$, let $\mathcal{O}_\rho$ be the space of analytic functions over a complex neighborhood $\Omega_\rho \subset \mathbb{C}^{2n}$ of size ρ around $\Omega \times \mathbb{T}^n$ equipped with the supremum norm $||.||_\rho$ over Ω_ρ.

We say that the Hamiltonian h is exponentially stable over an open set $\widetilde{\Omega} \subset \Omega$ if there exists positive constants ρ, C_1, C_2, a, b and ε_0 which depend only on h and $\widetilde{\Omega}$ such that:

i) $h \in \mathcal{O}_\rho$.
ii) For any function $\mathcal{H}(I, \varphi) \in \mathcal{O}_\rho$ such that $||\mathcal{H} - h||_\rho = \varepsilon < \varepsilon_0$, an arbitrary solution $(I(t), \varphi(t))$ of the Hamiltonian system associated to $\mathcal{H}$ with an initial action $I(t_0)$ in $\widetilde{\Omega}$ is defined over a time $\exp(C_2/\varepsilon^a)$ and satisfies:

$$||I(t) - I(t_0)|| \leq C_1 \varepsilon^b \quad \text{for } |t - t_0| \leq \exp(C_2/\varepsilon^a), \quad (9)$$

a and b are called stability exponents.

Remark Along the same lines, the previous definition can be extended to an integrable Hamiltonian in the Gevrey class (see [53]).

Hence, for a small enough perturbation, the action variables of the unperturbed problem become *quasi integrals* of the perturbed system over exponentially long times.

In order to introduce the problem, we begin by a typical example of *non-exponentially* stable integrable Hamiltonian: $h(I_1, I_2) = I_1 I_2$. Indeed, the perturbed system governed by $h(I_1, I_2) + \varepsilon \sin(\theta_2)$ admits the unbounded solution $(I_1(t), I_2(t), \theta_1(t), \theta_2(t)) = (0, \varepsilon t, \varepsilon t^2/2, 0)$ which starts from the origin $(0, 0, 0, 0)$ at $t = 0$, hence a drift of the actions $(I_1(t), I_2(t))$ on a segment of length 1 occurs over a timespan of order $1/\varepsilon$.

The important feature in this example which has to be avoided in order to ensure exponential stability is the fact that the gradient $\nabla h(I_1, 0)$ remains orthogonal to the first axis. Equivalently, the gradient of the restriction of h on this first axis is identically zero.

Nekhoroshev [61,62,63] introduced the class of *steep* functions where this problem is avoided. The property of steepness is a quantitative condition of transversality for a real valued function differentiable over an open set $\Omega \subset \mathbb{R}^n$ which involves *all* the affine subspaces which intersect Ω. Actually, steepness can be characterized by the following simple geometric criterion proved thanks to theorems of real subanalytic geometry [67]:

Theorem 3 ([67]) *A real analytic scalar function without critical points is steep if and only if its restriction to any proper affine subspace admits only isolated critical points.*

This is an extension of a previous similar result in the holomorphic case [42].

In this setting, Nekhoroshev proved the following:

Theorem 4 ([62,63]) *If the integrable Hamiltonian h is real analytic, does not admit critical points, is nondegenerate ($|\nabla^2 h(I)| \neq 0$ for any $I \in \Omega$) and steep then h is exponentially stable.*

The set of steep functions is *generic* among sufficiently smooth functions. For instance, we have seen that the function xy is not steep but it can easily be shown that $xy + x^3$ is steep. Actually, a given function can be transformed into a steep function by adding higher order terms [61,62]. It can be noticed that the (quasi-)*convex* functions are the *steepest* functions since their restrictions to any affine subspaces admit at most *one* critical point which is also nondegenerate.

The original proof of Nekhoroshev is *global*. It is based on a covering of the action space in open sets with controlled resonance properties where one can build resonant normal forms (i. e.: where only resonant harmonics are retained) up to an exponentially small remainder. The averaged Hamiltonian is not necessarily integrable but, thanks to the steepness of the integrable Hamiltonian, if a drift of the normalized actions occurs then it can only lead to a zone associated to resonances of *lower multiplicity* than the initial one (i. e.: the resonant module admits a lower dimension). Eventually, after a short distance the orbits reach a resonance-free area (i. e.: the Fourier expansion of the normalized perturbation admits only nonresonant harmonics up to an exponentially small remainder). Then, the local normal form is integrable and yields the confinement of the action variables over the desired amount of time.

Improved Versions of Nekhoroshev Theorem

The articles of Nekhoroshev remained largely unnoticed in the western countries until Benettin, Galgani and Giorgilli [12] rewrote and clarified the initial proof in the convex case.

Benettin and Gallavotti [13] proved that under an assumption of (quasi) convexity of the unperturbed Hamiltonian, the proofs of these theorems can be simplified. Indeed, after an averaging as in the steep case, the quasi convexity and the *energy conservation* ensure that the normalized Hamiltonian is an *approximate Liapunov function* over exponentially long time intervals. This allows one to confine the actions in the initial set where the considered orbit was located (we do not have to consider a drift over resonant areas of different multiplicity as in the original proof). Hence, the construction of a *single* normal form is enough to confine the actions in the convex case.

Following this idea, Lochak [46,47] has significantly simplified the proof of Nekhoroshev estimates for the convex quasi integrable Hamiltonians. His reasonings are based on normalization around the periodic orbits of the integrable Hamiltonian which represent the worst type of resonances. Using convexity, Lochak obtains open sets around the periodic orbits which hold exponential stability. Then, Dirichlet theorem about *simultaneous Diophantine approximation* ensures that these open sets recover the whole action space and yield the global result.

A remarkable feature of this proof is the fact that *improved estimates* can be obtained in the vicinity of resonances thanks to the relative abundance of periodic orbits in these areas. More specifically, periodicity corresponds to $n-1$ commensurability relations and we have already several commensurability relations at the resonances hence Dirichlet theorem can be applied on a lower dimensional space with better rates of approximation [46,47]. These improvements are important to extend Nekhoroshev estimates for large systems or infinite dimensional systems, they also fit with the speed of drift of the action variables in examples of unstable quasi integrable Hamiltonian (these points will be discussed in the sequel).

It can also be noticed that averaging along the periodic orbits of the integrable system is exactly a *one phase* averaging without small denominators.

Toward sharp estimates, Lochak–Neishtadt [50] and Pöschel [71] have independently obtained the following:

Theorem 5 ([50,71]) *If the integrable Hamiltonian h is real analytic, does not admit critical points and convex over a domain $\Omega \subset \mathbb{R}^n$ then h is exponentially stable over Ω with the global exponents $a = b = 1/2n$.*

Moreover, around the resonant zones linked to a module of rank $m < n$, the integrable Hamiltonian h is exponentially stable with the improved exponents $a = b = 1/2(n-m)$.

The proof in [50], explicitly derived in [51], relies on Lochak periodic orbits method together with a refined procedure of averaging due to Neishtadt [60]. In [71], the original scheme of Nekhoroshev is combined with a refined study of the geometry of resonances which gives an accurate partition of the action space in open sets where the action variables are confined and also Neishtadt's averaging procedure is used. Pöschel's study of the geometry of resonances should also be important in the study of Arnold diffusion.

This value of the time exponent ($a = 1/2n$) is expected to be *optimal* in the convex case according to heuristic reasonings of Chirikov [22], see also [46] on the speed of drift of Arnold diffusion. Actually, Marco–Sauzin [53] in the Gevrey cases and Marco–Lochak [52] in the analytic case have essentially proved the optimality of the improved exponent $a = 1/2(n-2)$ in the doubly resonant zones starting from an example of unstable quasi integrable Hamiltonian given by Michel Herman.

On the other hand, the previous studies except the original one of Nekhoroshev do not cover the cases of a time-dependent perturbation or a perturbed steep integrable Hamiltonian, despite their importance in physics. For instance, a time periodic perturbation of a convex Hamiltonian can be reduced to the time-independent perturbation of a quasi convex Hamiltonian in the extended phase space, but this is not the case for a general time-dependent perturbation, where energy conservation cannot be used. This problem has been studied in the light of Nekhoroshev theory by Giorgilli and Zehnder [34] in connection with the dynamics of a particle in a time-dependent potential with a high kinetic energy.

A new general study of the stability of steep integrable Hamiltonians has been carried out in [66]. This proof of stability relies on the mechanism of Nekhoroshev since we analyze the dynamics around resonances of any multiplicity and use local resonant normal forms but the original *global* construction is substituted with a *local* construction along each trajectory of the perturbed system. This construction is based on the approximation of the frequencies $\nabla h(I(t))$ at certain times by rational vectors thanks to Dirichlet theorem of simultaneous Diophantine approximation as in Lochak's proof of exponential stability in the convex case. This allows significant simplifications with respect to Nekhoroshev's original proof. Moreover, the results of Lochak and Pöschel are generalized for the steep

case since the exponents of stability derived in [66] give back the exponents $a = b = 1/2n$ in the particular case of convex integrable Hamiltonians.

In [68], Nekhoroshev estimates are proved for perturbations of integrable Hamiltonians which satisfy a *strictly weaker* condition of nondegeneracy than the initial condition of steepness. Indeed, the lack of steepness allows a drift of the action variables around the resonant zones. But, due to the exponential decay of the Fourier coefficients in the expansion of the Hamiltonian vector field, such a drift would be extremely slow if one consider resonant zones linked to a module spanned by integer vectors of large length. Thanks to this property, Morbidelli and Guzzo [57] have observed that the Hamiltonian $h(I_1, I_2) = I_1^2 - \delta I_2^2$ where δ is the square of a Diophantine number is *non steep* but nevertheless h is *exponentially stable* since its isotropic directions are the lines spanned by $(1, \pm\sqrt{\delta})$ which are "far" from the lines with rational slopes. This phenomenon has been studied numerically for the quadratic integrable Hamiltonians in [39].

Starting from this observation, a general weak condition of steepness which involves only the affine subspaces spanned by *integer* vectors has been stated in [68] with a complete proof of exponential stability in this setting.

The point in this refinement lies in the fact that it allows one to exhibit a *generic* class of real analytic integrable Hamiltonians which are exponentially stable with *fixed* exponents of stability a and b while Nekhoroshev original theory provides a generic set of exponentially stable integrable Hamiltonians but with exponents of stability which can be *arbitrarily small* [68].

More specifically, we consider genericity in a measure theoretical sense since:

Theorem 6 ([68]) *Consider an arbitrary real analytic integrable Hamiltonian h defined on a neighborhood of the closed ball $\overline{\mathrm{B}}_R^{(n)}$ of radius R centered at the origin in $\mathbb{R}^n$.*

For almost any $\Omega \in \mathbb{R}^n$, the integrable Hamiltonian $h_\Omega(x) = h(I) - \Omega.I$ is exponentially stable with the exponents:

$$a = \frac{b}{2+n^2} \quad \textit{and} \quad b = \frac{1}{2(1+2^n n)} \,.$$

Finally, all these results can be generalized for quasi integrable *symplectic mappings* thanks to a theorem of inclusion of an analytic symplectic diffeomorphism into the flow linked to a real analytic Hamiltonian up to an exponentially small accuracy ([43,75] or [37] for a direct proof).

KAM Stability, Exponential Stability, Nekhoroshev Stability

A last point which should be emphasized is the link between KAM stability, exponential stability and Nekhoroshev stability which are cornerstones in the study of stability of analytic quasi integrable Hamiltonian systems.

A first problem is the stability of the solutions in a neighborhood of a Lagrangian invariant torus over which an analytic Hamiltonian vectorfield induces a linear flow of frequency ω. Then the considered Hamiltonian can be written as $\mathcal{H}(I,\theta) = \omega.I + \mathcal{F}(I,\theta)$ in action-angle variables where the perturbation $\mathcal{F}(I,\theta)$ is analytic in a neighborhood of the origin and starts at order *two* in actions.

In the case of a KAM torus, the frequency is strongly nonresonant (Diophantine). With a suitable rescaling, the expansion of the Hamiltonian takes the form considered in Theorem 1 and the exponential estimates of stability (9) are valid.

In the general case (where the frequency can satisfy resonances of low order), the previous procedure cannot be applied. On the other hand, if one assumes that there are no resonances of order lower or equal to *four*:

$$\forall k \in \mathbb{Z}^n \backslash \{0\} \quad \text{such that} \quad |k| = |k_1| + \ldots + |k_n| \leq 4 \quad \text{then} \quad |k.\omega| \neq 0 \,,$$

it is possible to perform a Birkhoff's normalization which reduces the studied Hamiltonian to:

$$\widetilde{\mathcal{H}}\left(\widetilde{I},\widetilde{\theta}\right) = \omega.\widetilde{I} + \widetilde{Q}\left(\widetilde{I}\right) + \widetilde{\mathcal{F}}\left(\widetilde{I},\widetilde{\theta}\right) \,, \tag{10}$$

where $\widetilde{Q}$ is a quadratic form (the torsion) and $\widetilde{\mathcal{F}}(\widetilde{I},\widetilde{\theta}) = \mathcal{O}_3(\widetilde{I})$. At this point, one introduces the steepness condition required for the application of Nekhoroshev theory by imposing that the quadratic form $\widetilde{Q}$ is *sign definite* (a weaker condition is considered in [24]).

In a neighborhood of the considered torus, the problem is now reduced to the study of perturbations of a nonlinear *convex* integrable Hamiltonian and an exponential estimate of stability can be derived. These reasonings were stated in ([62], 2.2) and more specifically studied in [46,48].

A remarkable result of Morbidelli and Giorgilli [56] clearly shows that the two previous results which come respectively from Hamiltonian perturbation theory and Nekhoroshev's theorem are *independent* and can be *superimposed*. Indeed, in the case of a strongly nonresonant invariant torus (especially for a KAM torus) which admits *moreover* a sign definite torsion, one can state results

of stability over *superexponentially* long times. The proof starts with a Birkhoff's normal form like (10) but with an exponentially small remainder. Then, Nekhoroshev's theorem is applied with a perturbation which is already exponentially small, hence we obtain a superexponential time of stability. More specifically, provided that the Birkhoff normal form is quasi-convex, results of stability over times of the order of $\exp(\exp(cR^{-1/\tau}))$ can be ensured for the solutions with an initial condition in a ball of radius R small enough around the invariant torus.

Actually, the previous results were extended [46,48] for an *elliptic equilibrium* point or a lower dimensional torus in a Hamiltonian system except for an annoying problem of singularity in the action-angle transformation. This problem does not allow one to prove a stability theorem for a complete neighborhood of an elliptic equilibrium point. The corresponding theorems for *all initial data* were obtained in [28,65,72] where Nekhoroshev's estimates were established without action-angle variables.

Finally, the relationship between KAM and Nekhoroshev theory was considered in [32] and [23] where the existence of a sequence of nested domains in phase space which converge to the KAM set of invariant tori for the perturbed system and over which stability estimates are valid and growth occurs in a superexponential way was proven. Especially, on the initial domain we recover the usual statement of Nekhoroshev.

Applications

With the remaining space, we only give glimpses of application of the previous theorems in physics and astronomy. We mainly quote surveys in the sequel and these references do not form at all a complete list.

The Case of a Constant Frequency Integrable System

The question of stability of a perturbed single frequency system corresponds to the problem of preservation of the *adiabatic invariants* which appears in numerous physical problems [2,59].

Especially, for the applications in plasma physics, we can mention the beautiful survey of Northrop [69]. The problem of charge trapping by strong nonuniform magnetic fields (Van Allen belts, magnetic bottles) can also be tackled by means of the adiabatic invariant theory [15].

The same question arises in connection with the problem of *energy equipartition* in large systems of Fermi–Pasta–Ulam type and the conjecture of Boltzmann and Jeans [8,30].

The problem of existence of an approximate first integral over a long time but this time for a quasi integrable symplectic mapping appears to study the effective stability of the billiard flow near the boundary of a strictly convex domain in $\mathbb{R}^n$ [35].

For several degrees of freedom, results of stability over very long times for a quasi integrable Hamiltonian system were first obtained by Littlewood [45] about *triangular Lagrangian equilibria* in the three bodies problem. One can reduce this question to the study of a perturbed strongly nonresonant constant frequency system [31].

For effective computations, it is much more efficient to make a numerical summation at the smallest term instead of plugging the data into an abstract theorem which usually gives poor estimates. Following this scheme, Giorgilli and different coauthors have obtained effective stability results in celestial mechanics with *realistic* physical parameters (see [8,27,30,33,74]).

The same situation of a perturbed strongly nonresonant constant frequency system appears in the study of stability of symplectic numerical integrators [14,41,55].

In the realm of PDE's, results of stability around *finite dimensional nonresonant tori* can be proved ([4,11,18,44] and ► Perturbation Theory for PDEs for surveys).

The remarkable paper of Bambusi and Grebert [6] gives an extension of the previous results for *infinite dimensional* nonresonant tori.

Application of the Global Nekhoroshev Theory

Here, we look for global results of stability for perturbed nonlinear integrable Hamiltonian systems.

This situation appears in celestial mechanics where the unperturbed system is often *properly degenerate*, namely the number of constants of motion exceeds the number of degrees of freedom. This is the case for the Kepler problem which yields the integrable part of the planetary n-bodies problem (i. e.: the approximation of the n-bodies problem corresponding to the motion of the planets in the solar system). A study of this system in the light of Nekhoroshev's theory was given in [64], suggesting a modification of the original statement of Nekhoroshev by considering the proper degeneracies of this system.

The question of stability in celestial mechanics was also considered for the asteroid belt [57] where additional degeneracies and resonances appear (see also [40]).

We have seen that Nekhoroshev's theory also allows one to study stability around an elliptic equilibrium point in a Hamiltonian system with a sign definite torsion. But the most famous example of such an equilibrium in astronomy, namely the stability of an asteroid located at the top of an equilateral triangle with the Sun and Jupiter (the Lagrangian points L4, L5) cannot be tackled with the pre-

vious theorems in the convex case. It can be shown that with the actual masses of the Sun and Jupiter, the problem of stability of the Lagrangian points could be reduced to a study of a small perturbation of an integrable Hamiltonian of three degrees of freedom whose 3-jet satisfies a condition which implies steepness [10]. It allows one to prove a confinement over very long time intervals for asteroids located close enough to the Lagrangian points L4 or L5.

Other applications of Nekhoroshev's stability at an elliptic equilibrium point are the stability of the Riemann Ellipsoid [29], the fast rotation of a rigid body [9].

A Nekhoroshev-like theory has also been developed for beam dynamics [25,26].

From a numerical point of view, a new spectral formulation of the Nekhoroshev theorem has been introduced [38]. This allows one to recognize whether or not the motion in a quasi-integrable Hamiltonian system is in a Nekhoroshev regime (i. e. the action coordinates are eventually subject to an exponentially slow drift) by looking at the Fourier spectrum of the solutions.

In a geometric setting, an extension of Nekhoroshev's results for perturbations of convex, *noncommutatively* integrable Hamiltonian systems has been given in [16].

Finally, no general results like the Nekhoroshev theorem are known yet for large Hamiltonian systems or for Hamiltonian PDE's seen as infinite dimensional Hamiltonian systems. But a number of quasi-Nekhoroshev theorems for special systems of this type have been proved mostly by Bourgain and Bambusi (see [5] for large systems and [3,7,18] for PDE's).

Future Directions

The analyticity of the studied systems is only needed for the construction of the normal forms up to an exponentially high remainder. On the other hand, the steepness condition is generic for Hamiltonians of *finite* but sufficiently high smoothness [68]. It would be natural to prove the analogous stability theorems in the case of smooth functions which would give stability over *polynomially* long times.

Another question is the extension of Lochak's mechanism of stabilization around resonances for *non* quasi-convex integrable Hamiltonians. These improved exponents are at the basis of Nekhoroshev-type results which are obtained in large systems [5] or in PDE [3,7,18] hence their generalization would be important.

The global stability results considered so far are obviously valid only if one takes into account the most pessimistic estimations in the whole phase space. On the other hand, we have just seen that these results can be improved locally. It would be relevant to make a study of the *average* exponent of stability. It could be a *space* average of these exponents by considering all initial conditions or, in certain examples, the *time* average of these exponents by taking into account the variations of the speed of drift of the actions under Arnold's diffusion.

The applications of Nekhoroshev's theory in Astronomy is an active field either analytically or in a numerical way [36].

The relevance of Nekhoroshev's estimates for statistical mechanics and thermodynamics is an important question which is tackled with the problem of energy equipartition in the Fermi-Pasta-Ulam (FPU) model (see [19,20] and the references therein).

Finally, a partial generalization of KAM theory to PDEs has been carried out during the last twenty years by many high level mathematicians, but the theory developed up to now only allows one to show that *finite* dimensional invariant tori persist under perturbation. Thus, most of the initial data are outside invariant tori. It is therefore clear that it would be very important to understand the behavior of solutions starting outside the tori. This is also related to the problem of estimating the time of existence of the solutions of hyperbolic PDEs in compact domains, a problem that is one of the most important open questions in relation to hyperbolic PDEs. Up to now Nekhoroshev's theorem has been generalized to PDE's only in order to deal with small amplitude solutions but the obtention of really global results of stability would be a challenge (Kuksin has announced these kinds of theorems for the KdV equation, see ► Perturbation Theory for PDEs).

Appendix: An Example of Divergence Without Small Denominators

We would like to develop the following example of Neishtadt [60] where the phenomenons of divergence without small denominators and summation up to an exponentially small remainder appear in its simplest setting (see also [76] for another example).

Consider the quasi integrable system governed by the Hamiltonian

$$\mathcal{H}(I_1, I_2; \theta_1, \theta_2) = I_1 - \varepsilon[I_2 - \cos(\theta_1) f(\theta_2)] \quad \text{defined over } \mathbb{R}^2 \times \mathbb{T}^2 , \tag{11}$$

with

$$f(\theta) = \sum_{m \geq 1} \frac{\alpha_m}{m} \cos(m\theta) ,$$

where $\alpha_m = \mathrm{e}^{-\alpha m}$ for $0 < \alpha \leq 1$, this last choice corresponds to the exponential decay of the Fourier coefficients for a holomorphic function.

Indeed, $f'(\theta) = \mathrm{Im}(g(-\alpha + i\theta))$ where $g(z) = (\exp(z) - 1)^{-1}$ and the complex pole of f which is closest to the real axis is located at $-i\alpha$.

Conversely, all real function which admits a holomorphic extension over a complex strip of width α (i. e.: $|\mathrm{Im}(z)| \leq \alpha$) has Fourier coefficients bounded by $(C\alpha_m)_{m\in\mathbb{N}^*}$ for some constant $C > 0$.

As in the Sect. "Hamiltonian Perturbation Theory", it is possible to eliminate formally completely the fast angle θ_1 in the perturbation *without the occurrence* of any small denominators.

Indeed, one can consider a normalizing transformation generated by $X(\theta_1, \theta_2) = \sum_{n\geq 1} \varepsilon^n X_n(\theta_1, \theta_2)$ where the functions X_n satisfy the homological equations:

$$\partial_{\theta_1} X_1(\theta_1, \theta_2) = \cos(\theta_1) f(\theta_2) \text{ and}$$
$$\partial_{\theta_1} X_n(\theta_1, \theta_2) = \partial_{\theta_2} X_{n-1}(\theta_1, \theta_2)\,.$$

The solutions can be written $X_n(\theta_1, \theta_2) = \cos(\theta_1 - n\frac{\pi}{2}) f^{(n-1)}(\theta_2)$, hence:

$$X(\theta_1, \theta_2) = \sum_{n\in\mathbb{N}^*} \varepsilon^n \sum_{m\in\mathbb{N}^*} \frac{\alpha_m}{m^2} m^n \cdot \cos\left(\theta_1 - n\frac{\pi}{2}\right) \sin\left(m\theta_2 + n\frac{\pi}{2}\right)$$

and, for instance,

$$X\left(\frac{\pi}{2}, 0\right) = \sum_{m\in\mathbb{N}^*} \sum_{k\in\mathbb{N}} \frac{\mathrm{e}^{-\alpha m}}{m^2} (\varepsilon m)^{2k+1}$$

which is divergent for all $\varepsilon > 0$ since $\varepsilon \geq 1/m$ for m large enough.

We see that divergence comes from coefficients arising with successive differentiations.

On the other hand, if one considers the transformation generated by the truncated Hamiltonian $\sum_{n\geq 1}^{N} \varepsilon^n X_n(\theta_1, \theta_2)$ then the transformed Hamiltonian becomes

$$\mathcal{H}(I_1, I_2; \theta_1, \theta_2) = I_1 - \varepsilon I_2 + \varepsilon^{N+1} \partial_{\theta_2} X_N(\theta_1, \theta_2)\,,$$

where $\partial_{\theta_2} X_N(\theta_1, \theta_2) = \cos\left(\theta_1 - N\frac{\pi}{2}\right) f^{(N)}(\theta_2)$.

Especially, with $g(z) = 1/\exp(z) - 1$, we have:

$$\partial_{\theta_2} X_{2N}(\theta_1, \theta_2) = -\cos(\theta_1)\mathrm{Re}\left(g^{(2N-1)}(-\alpha + i\theta_2)\right) \quad \text{for } N > 0\,,$$

$$\partial_{\theta_2} X_{2N+1}(\theta_1, \theta_2) = \sin(\theta_1)\mathrm{Im}\left(g^{(2N)}(-\alpha + i\theta_2)\right) \quad \text{for } N \geq 0\,.$$

Now, around the real axis, the main term in the asymptotic expansion of $g^{(n)}(z)$ as n goes to infinity is given by the derivative of the polar term $1/z$ in the Laurent expansion of g at 0.

Indeed, we consider the function $h(z) = g(z) - 1/z$ which is real analytic and admits an analyticity width of size 2π around the real axis. Hence, Cauchy estimates ensure that for n large enough and z in the strip of width π around the real axis, the derivative $h^{(n)}(z)$ becomes negligible with respect to the derivative of $1/z$.

Consequently, $g^{(n)}(z)$ is *equivalent* to $(-1)^n n!/z^{n+1}$ as n goes to infinity for z in the strip of width π around the real axis.

Hence, the remainder $\partial_{\theta_2} X_N(\theta_1, \theta_2)$ has a size of order $(N-1)!/\alpha^N$ for N large. This latter estimate *cannot be improved*, since $\partial_{\theta_2} X_{2N}(\pi, 0)$ and $\partial_{\theta_2} X_{2N+1}(\pi/2, \alpha \tan(\pi/4N + 2))$ admit the same size of order $(N-1)!/\alpha^N$ for N large.

Now, we can make a "summation at the smallest term" as in Sect. "The Case of a Single Frequency System" to obtain an optimal normalization with an *exponentially small* remainder.

Bibliography

Primary Literature

1. Arnold VI (1964) Instability of dynamical systems with several degrees of freedom. Sov Math Dokl 5:581–585
2. Arnold VI, Kozlov VV, Neishtadt AI (2006) Mathematical aspects of classical and celestial mechanics, 3rd revised edn. In: Encyclopaedia of Mathematical Sciences 3. Dynamical Systems 3. Springer, New York
3. Bambusi D (1999) Nekhoroshev theorem for small amplitude solutions in nonlinear Schrödinger equations. Math Z 230(2):345–387
4. Bambusi D (1999) On long time stability in Hamiltonian perturbations of non-resonant linear PDEs. Nonlinearity 12(4):823–850
5. Bambusi D, Giorgilli A (1993) Exponential stability of states close to resonance in infinite-dimensional Hamiltonian systems. J Stat Phys 71(3–4):569–606
6. Bambusi D, Grébert B (2006) Birkhoff normal form for partial differential equations with tame modulus. Duke Math J 135(3):507–567
7. Bambusi D, Nekhoroshev NN (2002) Long time stability in perturbations of completely resonant PDE's. Acta Appl Math 70(1–3):1–22
8. Benettin G (2005) Physical applications of Nekhoroshev theorem and exponential estimates. In: Giorgilli A (ed) Cetraro (2000) Hamiltonian Dynamics, Theory and Applications. Springer, New York
9. Benettin G, Fasso F (1996) Fast rotations of the rigid body: A study by Hamiltonian perturbation theory, I. Nonlinearity 9(1):137–186

10. Benettin G, Fasso F, Guzzo M (1998) Nekhorochev stability of L4 and L5 in the spatial restricted three body problem. Regul Chaotic Dyn 3(3):56–72
11. Benettin G, Fröhlich J, Giorgilli A (1988) A Nekhoroshev-type theorem for Hamiltonian systems with infinitely many degrees of freedom. Commun Math Phys 119(1):95–108
12. Benettin G, Galgani L, Giorgilli A (1985) A proof of Nekhorochev's theorem for the stability times in nearly-integrable Hamiltonian systems. Celest Mech 37:1–25
13. Benettin G, Gallavotti G (1986) Stability of motions near resonances in quasi-integrable Hamiltonian systems. J Stat Phys 44(3–4):293–338
14. Benettin G, Giorgilli A (1994) On the Hamiltonian interpolation of near-to-the-identity symplectic mappings with application to symplectic integration algorithms. J Stat Phys 74(5–6):1117–1143
15. Benettin G, Sempio P (1994) Adiabatic invariants and trapping of a point charge in a strong non-uniform magnetic field. Nonlinearity 7(1):281–303
16. Blaom AD (2001) A geometric setting for Hamiltonian perturbation theory. Mem Am Math Soc 727(xviii):112
17. Bogolyubov NN, Mitropol'skij YA (1958) Asymptotic Methods in the Theory of Nonlinear Oscillations, 2nd edn. Nauka, Moscow. Engl. Transl. (1961) Gordon and Breach, New York
18. Bourgain J (2004) Remarks on stability and diffusion in high-dimensional Hamiltonian systems and partial differential equations. Ergod Theory Dyn Syst 24(5):1331–1357
19. Carati A, Galgani L, Giorgilli A, Ponno A (2002) The Fermi–Pasta–Ulam Problem. Nuovo Cimento B 117:1017–1026
20. Carati A, Galgani L, Giorgilli A (2006) Dynamical Systems and Thermodynamics. In: Françoise JP, Naber GL, Tsun TS (eds) Encyclopedia of mathematical physics. Elsevier, Amsterdam
21. Celletti A, Giorgilli A (1991) On the stability of the Lagrangian points in the spatial restricted problem of three bodies. Celest Mech Dyn Astron 50(1):31–58
22. Chirikov BV (1979) A universal instability in many dimensional oscillator systems. Phys Rep 52:263–279
23. Delshams A, Gutierrez P (1996) Effective Stability and KAM Theory. J Diff Eq 128(2):415–490
24. Dullin H, Fassò F (2004) An algorithm for detecting directional quasi-convexity. BIT 44(3):571–584
25. Dumas HS (1993) A Nekhoroshev-like theory of classical particle channeling in perfect crystals. In: Jones CKRT et al (ed) Dynamics reported. Expositions in dynamical systems, new series, vol 2. Springer, Berlin
26. Dumas HS (2005) Mathematical theories of classical particle channeling in perfect crystals. Nucl Inst Meth Phys Res Sect B (Beam Interactions with Materials and Atoms) 234(1–2):3–13
27. Efthymiopoulos C, Giorgilli A, Contopoulos G (2004) Nonconvergence of formal integrals. II: Improved estimates for the optimal order of truncation. J Phys A (Math Gen) 37(45):10831–10858
28. Fassò F, Guzzo M, Benettin G (1998) Nekhoroshev-stability of elliptic equilibria of Hamiltonian systems. Commun Math Phys 197(2):347–360
29. Fassò F, Lewis D (2001) Stability properties of the Riemann ellipsoids. Arch Ration Mech Anal 158(4):259–292
30. Giorgilli A (1998) On the problem of stability for near to integrable Hamiltonian systems. In: Proceedings of the International Congress of Mathematicians Berlin 1998. Documenta Mathematica III:143–152
31. Giorgilli A, Delshams A, Fontich E, Galgani L, Simó C (1989) Effective stability for a Hamiltonian system near an elliptic equilibrium point, with an application to the restricted three bodies problem. J Diff Equa 77:167–198
32. Giorgilli A, Morbidelli A (1997) Invariant KAM tori and global stability for Hamiltonian systems. Z Angew Math Phys 48(1):102–134
33. Giorgilli A, Skokos C (1997) On the stability of the Trojan asteroids. Astron Astroph 317:254–261
34. Giorgilli A, Zehnder E (1992) Exponential stability for time dependent potential. ZAMP 43:827–855
35. Gramchev T, Popov G (1995) Nekhoroshev type estimates for billiard ball maps. Ann Inst Fourier 45(3):859–895
36. Guzzo M (2003) Nekhoroshev stability of asteroids. Triennal report 2000–2003 of Commission 7-Celestial Mechanics and dynamical Astronomy of the IAU; Reports on Astronomy, 1999–2002. Transactions of the International Astronomical Union, vol XXVA, Astronomical Society of the Pacific
37. Guzzo M (2004) A direct proof of the Nekhoroshev theorem for nearly integrable symplectic maps. Ann Henri Poincaré 5(6):1013–1039
38. Guzzo M, Benettin G (2001) A spectral formulation of the Nekhoroshev theorem and its relevance for numerical and experimental data analysis. Discret Contin Dyn Syst Ser B 1(1):1–28
39. Guzzo M, Lega E, Froeschlé C (2006) Diffusion and stability in perturbed non-convex integrable systems. Nonlinearity 19(5):1049–1067
40. Guzzo M, Morbidelli A (1997) Construction of a Nekhoroshev-like result for the asteroid belt dynamical system Celest. Mech Dyn Astron 66:255–292
41. Hairer E, Lubich C, Wanner G (2006) Geometric numerical integration. Structure-preserving algorithms for ordinary differential equations, 2nd edn. Springer Series in Computational Mathematics, vol 31. Springer, New York
42. Ilyashenko IS (1986) A steepness test for analytic functions. Russ Math Surv 41:229–230
43. Kuksin S, Pöschel J (1994) On the inclusion of analytic symplectic maps in analytic Hamiltonian flows and its applications. In: Kuksin S, Lazutkin VF, Pöschel J (eds) Proceedings of the 1991 Euler Institute Conference on Dynamical Systems. Prog NonLin Diff Equ App (12) Birkhäuser, Basel
44. Kuksin SB (2006) Hamiltonian PDEs (with an appendix by Dario Bambusi). In: Hasselblatt B (ed) Handbook of dynamical systems, vol 1B. Elsevier, Amsterdam
45. Littlewood JE (1959) On the equilateral configuration in the restricted problem of the three bodies. Proc London Math Soc 9(3):343–372
46. Lochak P (1992) Canonical perturbation theory via simultaneous approximation. Russ Math Surv 47:57–133
47. Lochak P (1993) Hamiltonian perturbation theory: Periodic orbits, resonances and intermittency. Nonlinearity 6(6):885–904
48. Lochak P (1995) Stability of Hamiltonian systems over exponentially long times: the near-linear case. In: Dumas HS, Meyer K, Schmidt D (eds) Hamiltonian dynamical systems – History, theory and applications. IMA conference proceedings series, vol 63. Springer, New York, pp 221–229
49. Lochak P, Meunier C (1988) Multiphase averaging methods for Hamiltonian systems. Appl Math Sci Series, vol 72. Springer, New York

50. Lochak P, Neishtadt AI (1992) Estimates of stability time for nearly integrable systems with a quasiconvex Hamiltonian. Chaos 2(4):495–499
51. Lochak P, Neishtadt AI, Niederman L (1994) Stability of nearly integrable convex Hamiltonian systems over exponentially long times. In: Kuksin S, Lazutkin VF, Pöschel J (eds) Proceedings of the 1991 Euler Institute Conference on Dynamical Systems. Prog NonLin Diff Equ App (12). Birkhäuser, Basel
52. Marco JP, Lochak P (2005) Diffusion times and stability exponents for nearly integrable analytic systems. Central Eur J Math 3(3):342–397
53. Marco JP, Sauzin D (2003) Stability and instability for Gevrey quasi-convex near-integrable Hamiltonian Systems. Publ Math Inst Hautes Etudes Sci 96:199–275
54. Meyer KR, Hall GR (1992) Introduction to Hamiltonian dynamical systems and the N-Body problem. Applied Mathematical Sciences, vol 90. Springer, New York
55. Moan PC (2004) On the KAM and Nekhoroshev theorems for symplectic integrators and implications for error growth. Nonlinearity 17(1):67–83
56. Morbidelli A, Giorgilli A (1995) Superexponential stability of KAM tori. J Stat Phys 78:1607–1617
57. Morbidelli A, Guzzo M (1997) The Nekhoroshev theorem and the asteroid belt dynamical system. Celest Mech Dyn Astron 65(1–2):107–136
58. Moser J (1955) Stabilitätsverhalten Kanonisher Differentialgleichungssysteme. Nachr Akad Wiss Göttingen, Math Phys Kl IIa 6:87–120
59. Neishtadt AI (1981) On the accuracy of conservation of the adiabatic invariant. Prikl Mat Mekh 45:80–87. Translated in J Appl Math Mech 45:58–63
60. Neishtadt AI (1984) The separation of motions in systems with rapidly rotating phase. J Appl Math Mech 48:133–139
61. Nekhorochev NN (1973) Stable lower estimates for smooth mappings and for gradients of smooth functions. Math. USSR Sb 19(3):425–467
62. Nekhorochev NN (1977) An exponential estimate of the time of stability of nearly integrable Hamiltonian systems. Russ Math Surv 32:1–65
63. Nekhorochev NN (1979) An exponential estimate of the time of stability of nearly integrable Hamiltonian systems 2. Trudy Sem Petrovs 5:5–50. Translated In: Oleinik OA (ed) Topics in Modern Mathematics. Petrovskii Semin, vol 5. Consultant Bureau, New York
64. Niederman L (1996) Stability over exponentially long times in the planetary problem. Nonlinearity 9(6):1703–1751
65. Niederman L (1998) Nonlinear stability around an elliptic equilibrium point in an Hamiltonian system. Nonlinearity 11:1465–1479
66. Niederman L (2004) Exponential stability for small perturbations of steep integrable Hamiltonian systems. Erg Theor Dyn Syst 24(2):593–608
67. Niederman L (2006) Hamiltonian stability and subanalytic geometry. Ann Inst Fourier 56(3):795–813
68. Niederman L (2007) Prevalence of exponential stability among nearly integrable Hamiltonian systems. Erg Theor Dyn Syst 27(3):905–928
69. Northrop TG (1963) The adiabatic motion of charged particles. Interscience Publishers, New York
70. Poincaré H (1892) Méthodes Nouvelles de la Mécanique Céleste, vol 4. Blanchard, Paris
71. Pöschel J (1993) Nekhorochev estimates for quasi-convex Hamiltonian systems. Math Z 213:187–217
72. Pöschel J (1999) On Nekhoroshev's estimate at an elliptic equilibrium. Int Math Res Not 1999(4):203–215
73. Ramis JP, Schäfke R (1996) Gevrey separation of fast and slow variables. Nonlinearity 9(2):353–384
74. Steichen D, Giorgilli A (1998) Long time stability for the main problem of artificial satellites. Cel Mech 69:317–330
75. Treschev DV (1994) Continuous averaging in Hamiltonian systems. In: Kuksin S, Lazutkin VF, Pöschel J (eds) Proceedings of the 1991 Euler Institute Conference on Dynamical Systems. Prog NonLin Diff Equ App (12). Birkhäuser, Basel
76. Valdinoci E (2000) Estimates for non-resonant normal forms in Hamiltonian perturbation theory. J Stat Phys 101(3–4):905–919

Books and Reviews

Arnold VI (1983) Geometrical methods in the theory of ordinary differential equations. Transl. from the Russian by Joseph Szuecs, Mark Levi (ed) Grundlehren der Mathematischen Wissenschaften 250. Springer, New York

Arnold VI, Kozlov VV, Neishtadt AI (2006) Mathematical aspects of classical and celestial mechanics, 3rd revised edn. Encyclopaedia of Mathematical Sciences 3. Dynamical Systems 3. Springer, New York

Benettin G (2005) Physical applications of Nekhoroshev theorem and exponential estimates. In: Giorgilli A (ed) Cetraro (2000) Hamiltonian Dynamics, Theory and Applications. Springer, New York

Giorgilli A (2003) Exponential stability of Hamiltonian systems. Dynamical systems, Part I. Hamiltonian systems and celestial mechanics. Selected papers from the Research Trimester held in Pisa, Italy, February 4–April 26, 2002. Pisa: Scuola Normale Superiore. Pubblicazioni del Centro di Ricerca Matematica Ennio de Giorgi. Proceedings, 87–198

Sanders JA, Verhulst F, Murdock J (2007) Averaging methods in nonlinear dynamical systems, 2nd edn. Applied Mathematical Sciences 59. Springer, New York

Network Analysis, Longitudinal Methods of

Tom A. B. Snijders
University of Oxford, Oxford, United Kingdom

Article Outline

Glossary
Definition of the Subject
Introduction
Stochastic Models for Network Dynamics
Statistical Estimation and Testing
Example: Dynamics of Adolescent Friendship

Models for the Co-evolution of Networks and Behavior
Example: Co-evolution of Adolescent Friendship and Alcohol Use
Extensions
Future Directions
Bibliography

Glossary

Actors The social actors who are represented by the nodes of the network, and indicated by a label denoted i or j in the set $1, \ldots, n$.

Behavior An umbrella term for changing characteristics of actors, considered as components of the outcome of the stochastic system: e. g., behavioral tendencies or attitudes of human actors, performance, etc. Each behavior variable Z_h is assumed to be measured on an ordinal discrete scale with values $1, 2, \ldots, M_h$ for some $M_h \geq 2$. The value of behavior variable Z_h for actor i is denoted Z_{ih}.

Change determination process The stochastic model defining the probability distribution of changes, conditional on the event that there is an opportunity for change.

Change opportunity model The stochastic process defining the moments where tie indicators can change. This can be either *tie-based*, meaning that an ordered pair of actors (i, j) is chosen and the possibility arises that the tie variable from i to j is changed; or *actor-based*, meaning that an actor i is chosen and the possibility arises that one of the outgoing tie variables from actor i is changed.

Covariates Variables which can depend on the actors (actor covariates) or on pairs of actors (dyadic covariates), and which are considered to be deterministic, or determined outside of the 'stochastic system' under consideration.

Effects Components of the objective function.

Influence The phenomenon that change probabilities for actors' behavior depend on the network positions of the actors, usually in combination with the current behavior of the other actors.

Markov chain A stochastic process where the probability distribution of future states, given the present state, does not depend on past states.

Method of moments A general method of statistical estimation, where the parameters are estimated in such a way that expected values of a vector of selected statistics are equal to their observed values.

Network A simple directed graph representing a relation on the set of actors with binary tie indicators X_{ij} which can be regarded as a state which can change, but will normally change slowly.

Objective function Usually denoted by f_i; the informal description is that this is a measure of how attractive it is to go from an old to a new state. More formally, when there is an opportunity for change, the probability of the change is assumed to be proportional to the exponential transform of the objective function.
The objective function has a similar role as the linear predictor in generalized linear models in statistics, and is specified here as a linear combination of *effects*.

Rate function Usually denoted by λ, the expected number of opportunities for change per unit of time.

Selection The phenomenon that change probabilities for network ties depend on the behavior of one or both of the two actors involved.

Tie indicator A variable X_{ij} indicating by the value $X_{ij} = 1$ that there is a tie $i \to j$, and by the value 0 that there is no such tie. Also called tie variables.

Definition of the Subject

Social networks represent the patterns of ties between social actors. To analyze empirically the mechanisms that determine creation and termination of ties, especially if several mechanisms that may be complementary are studied simultaneously, statistical methods are needed. This chapter is aimed at the case that network panel data are available to the researcher, and treats recently developed statistical models for such data, with corresponding estimation methods. To represent the feedback processes inherent in network dynamics, it is helpful to regard such panel data as momentary observations on a continuous-time stochastic process on the space of directed graphs. Tie-oriented and actor-oriented stochastic models are presented, which can reflect endogenous network dynamics as well as effects of exogenous variables. These models can be regarded as agent-based models, and they can be implemented as computer simulation models. To estimate the parameters of the model, stochastic approximation methods can be used. Social networks are especially interesting because they are important influences on individual behavior – and in turn the network ties are influenced by individual behavior. This two-way influence can be represented by models for the co-evolution of networks and changeable actor attributes. Such models, and statistical methods to analyze panel data on networks and behavior, are also treated. An extensive example is discussed about friendship among teenagers in a school setting.

Introduction

When we think of social networks, it is quite natural to think of them as being dynamic. Ties are established, gain in strength, they can blossom and decay, and they may wither or be terminated with a bang. This applies to all kinds of relations – friendship or collaboration between humans, joint ventures between companies, bilateral agreements between countries, and so on. (Note that all these are examples of 'positive' ties; 'negative' ties are not dealt with in this chapter). It is also natural to think of such changes as being dependent on characteristics of the nodes – inclinations and abilities of humans, resources of companies, locations and capacities of countries – and on characteristics of pairs of nodes such as similarity or spatial proximity, as well as on the existing network structure – reciprocation of friendships, transitive closure of friendships (which is the case when friends of friends become friends), group formation of companies or countries. Finally, we are not surprised when such network changes have repercussions – friendships are often thought to have good or bad influences on the individuals concerned, agreements between companies and between countries will have consequences for the performance of the companies and for how the countries fare. Indeed, in the cases of the companies and countries, the links often are created with the purpose of having beneficial consequences for the companies or countries, respectively.

This sets the stage for this chapter, which is concerned with inferential data analysis for network dynamics. The focus is on social networks, and the nodes, or vertices, in the network will be referred to as (social) actors. Data analysis means that methods will be presented for analyzing empirical data on network change. Inference means that the aim is to have methods for testing hypotheses about mechanisms that may drive the network dynamics, and for estimating parameters that figure in such mechanisms. As usual in statistical inference, the 'mechanisms' will be expressed as probability models, also called stochastic models. For hypothesis testing it must be possible to play off one theory against another, which means that when we have theories, or mechanisms, T_1 and T_2, that are not contradictory but could occur together, we need models which can express both mechanisms simultaneously and reflect each mechanism in a set of parameters. The mechanism operates if some of its parameters are nonzero. Then we can test the null hypothesis that T_1 operates but not T_2, against the alternative hypothesis that both mechanisms operate; in other words, we can test for T_2 while controlling for T_1. This requires flexibility of the stochastic models being used: what was stated above implies that it must be possible to specify the models in diverse ways, e. g., with some parameters for effects of actor characteristics and others for effects of various aspects of the existing network structure, and these parameters being estimable from observed data. For data consisting of independent observations the elaboration of these principles of statistical inference in the linear regression model and its generalizations is well known. Network data have the complicating feature that they are not composed of independent observations: the occurrence, creation, or termination of one tie is highly dependent on the existence of other ties. Therefore more complex statistical models are needed, giving an adequate representation of the mutual dependence between the existence of various ties between the actors in the network. But readers who are acquainted with generalized linear statistical models will see that, although the models treated in this chapter have this greater complexity, many elements known from generalized linear models do play a role.

Social network analysis is concerned with diverse types of networks which can be represented by diverse data structures: simple graphs and their generalizations. In line with the current state of the statistical methodology for network dynamics, this chapter is restricted to data structures where the changing network is a changing simple directed graph, where the arcs represent social ties that can be regarded as *states* rather than events. This means that, although changeable, the ties have inertia, a tendency to endure. Friendship between humans and agreements between companies are examples of states; conversations and momentary transactions are events. For networks where ties are states, the dependence between ties can be represented by assuming that changes in the network are dependent upon the existing network structure. In mathematical terms, this is to say that it is reasonable to assume that – given the available 'independent' or 'explanatory' variables – the network is a *Markov chain*. A Markov chain is a stochastic process where the probability distribution of future states, given the present state, does not depend on the past states. Such models were first proposed for network dynamics by [22] and elaborated by [59]. For some sociological applications outside the realm of networks, see [2,10]. For a network of events (e. g., the network of conversations ongoing at each given moment) the assumption of a Markov chain would be untenable. For a network of states (e. g., the network of joint ventures between companies) the Markov assumption is usually not totally realistic but can be used as a first approximation, and in many cases is the best assumption one can make given the limitation of the available data. This assumption

often can be made more plausible by using relevant explanatory variables.

In the start of this introduction, ties were portrayed as possibly gaining in strength or decaying. It would be attractive to reflect this by measuring the ties on an ordinal scale. This is not considered here: we are dealing with simple and not with valued directed graphs, and the *tie indicators*, which are the variables X_{ij} indicating how actor i is tied to actor j, are constrained to having the values 0 or 1, indicating, respectively, absence and presence of the tie $i \to j$. The restriction to binary tie variables is in line with traditional network analysis, but extensions to valued ties are important and are the subject of current work. The Markov assumption will often be more reasonable for valued than for binary ties.

The assumption that we are dealing with a network of relational states must be reflected by the way in which the ties are measured. Measurement in social network analysis is a subject which tends to receive too little attention, and this chapter is no exception. In practice, ties will have to be measured in such a way that observing a tie is a good indicator for the relational state being investigated, such as friendship or collaboration. When the relation under study is a type of communication, which usually has an ephemeral nature, it will be necessary to aggregate the communication over a sufficiently long time interval, so that the resulting variable can be regarded as indicative of a relational state. For example, [25] aggregated email communication to the binary tie variable defined by at least one email being sent over a 60-day period. In other situations, shorter or longer periods may be relevant.

Stochastic Models for Network Dynamics

This section presents stochastic models for use in statistical modeling. The inferential aspects (parameter estimation and testing of hypotheses) are treated in Section "Statistical Estimation and Testing". These models were applied, e. g., to the testing of theories about dynamics of friendship networks [12,55,57], of trust networks in firms [56], of artistic prestige [13], and of ties between venture capital firms [8].

The network is represented by the node set $\{1, \ldots, n\}$ with tie variables x_{ij}, where $x_{ij} = 1$ or 0 indicates whether the tie $i \to j$ is present or absent. The tie variables are collected in the $n \times n$ adjacency matrix $x = (x_{ij})$. Self-ties are excluded, so that $x_{ii} = 0$ for all i. The concepts of network (directed graph) and matrix (its adjacency matrix) will be used interchangeably, depending on what is most convenient. In accordance with the usual notation in probability and statistics, random variables will be indicated by capitals; and observations, or other non-random variables, by small letters. The ties are assumed to be outcomes of time-dependent random variables, denoted by $X_{ij}(t)$ and collected in the time-dependent random matrix $X(t)$.

In addition to the network $X(t)$, which can be regarded as the dependent variable of the model, there can be other variables regarded as independent or explanatory variables in the sense that their values are not modeled but accepted as given, and they may influence the network. Such variables are called *covariates* and when depending on the actors they are denoted v_i, while if they depend on pairs of actors (*dyads*) the notation is w_{ij}. Examples are the age of actors (actor variable) and their spatial proximity (dyadic variable).

Basic Model Definition

The following basic assumptions are made.

1. Time, denoted by t, is a continuous variable. This does not mean that it is assumed that observations are made continuously; in most practical cases, observations are made at a number (perhaps a small number) of discrete time moments. However, it is natural and mathematically convenient to assume that there is an underlying process $X(t)$ (which may be observed only partially) which proceeds in continuous time.
2. $X(t)$ is a Markov process.
   This means that the conditional distribution of future states depends on the past only as a function of the present. In other words, to predict the future it is sufficient to know the present state of the network, and knowledge of past states will not improve predictability. This assumption was discussed above. It can be expressed by saying that the network represents a *state*, and usually goes together with inertia, i. e., the tendency of ties to remain in existence unless something special happens.
3. At any given moment t, no more than one tie variable $X_{ij}(t)$ can change.
   This assumption, first proposed by [22], means that changes of ties are not directly coordinated, and ties are mutually dependent only because tie changes will depend on the current total configuration of ties. This is an important simplifying condition and excludes, for example, partner swapping and the coordinated formation of groups. Changes of several ties are decomposed as sequences of changes of single ties.

These assumptions still allow an extremely wide array of probability models, and further specification is necessary. In network change, two aspects can be distinguished: the frequency of tie change, which may depend on the ac-

tors involved; and the network structures that tend to be formed by the tie changes – the 'direction' of change. Examples of the former are that younger individuals might change their friendship ties more frequently than older individuals, or that more central actors might change their ties more frequently than peripheral actors. Examples of the latter are tendencies toward tie reciprocation, and toward transitive network closure. These two aspects will be represented by distinct components of the model and distinct parameters, which allow the inference about the one aspect to be relatively undisturbed by inference about the other aspect. The first component is the *change opportunity* process, the second the *change determination* model.

Change Opportunity Process Two specifications of the change opportunity process are given. They use the concept of a *Poisson process*, which is a stochastic process of events occurring at a certain rate λ, which means that the probability that an event occurs in the time interval from t to $t+\epsilon$, where ϵ is a small positive number, is given (in the limit for ϵ tending to 0) by $\lambda\epsilon$. One could say that the rate is the probability of occurrence per unit of time (in short time intervals).

Two specifications of the opportunity process are given here.

1. *Tie-based change opportunities*
   For each tie variable X_{ij}, opportunities for change occur according to a Poisson process with rate λ_{ij}.
2. *Actor-based change opportunities*
   For each actor i, opportunities to establish one new outgoing tie $i \to j$, or dissolve one existing tie $i \to j$, occur according to a Poisson process with rate λ_i.

The rates λ can be constant, or depend on covariates or functions of the current state of the network; if they are not constant, they are called *rate functions*. Tie-based rates λ_{ij}, for example, could depend on the proximity between actors or on their joint embeddedness such as the current number of common friends $\sum_h X_{ih}(t)\,X_{jh}(t)$. Actor-based rates λ_i could depend on actor variables or on positional variables such as actor i's current outdegree $\sum_j X_{ij}(t)$.

Tie-based change opportunities were proposed by Robins and Pattison (personal communication), and correspond to the Gibbs sampling and Metropolis–Hastings procedures for simulating exponential random graph models, see [41]. Actor-based change opportunities were proposed by [46].

When an opportunity for change occurs, there is, in each opportunity model, a set of potential new networks that could be the result of the change. Denoting the current network by x^0, for the tie-based opportunity model this set can be denoted by $C_{ij}(x^0)$. This is the set of the two possible matrices x where all elements other than x_{ij} are equal to those in the current matrix x^0, and where x_{ij} itself can be either 0 or 1. In the actor-based opportunity model actor i, when confronted with an opportunity for change, chooses one of his outgoing tie variables and changes this into its opposite value, changing 0 to 1 (creating a new tie) or changing 1 to 0 (terminating an existing tie). Therefore the set of potential new networks here is the set composed of x^0 itself together with the $n-1$ matrices which are equal to x^0 except for exactly one non-diagonal element in line i which is replaced by its opposite, $x_{ij} = 1 - x^0_{ij}$.

The set of new possible states is denoted in shorthand applicable to either case by $C(x^0)$. Since it is allowed that the current situation is continued, it always holds that $x^0 \in C(x^0)$.

Change Determination Model The choice of the new state of the network is dependent on what is called the *objective function*, which is a function $f_i(x^0, x, v, w)$ depending on the current state of the network x^0, the potential new state x, the actor i, and the covariates summarized here as v (actor covariates) and w (dyadic covariates). The objective function can be interpreted informally as a measure of how attractive it is for actor i to change from state x^0 to state x.

> When actor i has the opportunity to change some outgoing tie variable X_{ij}, given that currently $X(t) = x^0$, the set of possible new states of the network is denoted $C(x^0)$. All $x \in C(x^0)$ differ from x^0 by at most one element x_{ij} for some j. When there is an opportunity for change from the current state x^0, the probabilities of the values of the next state $x \in C(x^0)$ are proportional to $\exp\big(f_i(x^0, x, v, w)\big)$.

The models can be summarized as follows. For the tie-oriented model, when an opportunity for change occurs, it refers to some pair (i, j); opportunities for changing X_{ij} occur at a rate λ_{ij} for each pair (i, j). When such an opportunity occurs, the probability that x^0 changes to the different state x is given by

$$\begin{aligned}
&\mathrm{P}\{X(t) \text{ changes to } x \mid (i,j) \text{ has a change opportunity at time } t, X(t) = x^0\} \\
&= p_{ij}(x^0, x, v, w) \\
&= \frac{\exp\big(f_i(x^0, x, v, w)\big)}{\exp\big(f_i(x^0, x^0, v, w)\big) + \exp\big(f_i(x^0, x, v, w)\big)},
\end{aligned} \tag{1}$$

where x and x^0 are identical except for $x_{ij} = 1 - x^0_{ij}$.

For the actor-oriented model, opportunities for change occur for actors i. Opportunities for actor i to change one of the outgoing tie variables X_{ij} ($j = 1, \ldots, n$; $j \neq i$) occur at a rate λ_i. The set of permitted new states, following on a given current state x^0, is $C(x^0)$. The probability that the new state is x, provided that x is permitted (i. e., $x \in C(x^0)$), is given by

$$\begin{aligned} &\mathrm{P}\{X(t) \text{ changes to } x \mid i \text{ has a change opportunity at} \\ &\quad \text{time } t, X(t) = x^0\} \\ &= p_i(x^0, x, v, w) \qquad (2) \\ &= \frac{\exp\left(f_i(x^0, x, v, w)\right)}{\sum_{x' \in C(x^0)} \exp\left(f_i(x^0, x', v, w)\right)} . \end{aligned}$$

The two model components can be put together by giving the transition rate matrix, also called Q-matrix, of which the elements are defined by

$$q_{x^0,x} = \lim_{\mathrm{d}t \downarrow 0} \frac{\mathrm{P}\left\{X(t + \mathrm{d}t) = x \mid X(t) = x^0\right\}}{\mathrm{d}t} \quad (x \neq x^0)$$

(see textbooks on continuous-time Markov chains, such as [34]). Note that the assumptions imply that

$$q_{x^0,x} = 0 \text{ whenever } x_{ij} \neq x^0_{ij}$$
$$\text{for more than one element } (i, j).$$

For digraphs x and x^0 which differ from each other only in the element with index (i, j), the elements of the Q-matrix are given for the tie-based opportunity process by

$$q_{x^0,x} = \lambda_{ij}(x^0, v, w)\, p_{ij}(x^0, x, v, w) \qquad (3)$$

and for the actor-based opportunity process by

$$q_{x^0,x} = \lambda_i(x^0, v, w)\, p_i(x^0, x, v, w) . \qquad (4)$$

Tie-based models with constant change rates and objective functions defined as $f(x, v, w)$ (not depending on the preceding state x^0 or on the actor i) can be regarded as Metropolis–Hastings dynamics (cf. [34]) for obtaining random draws from the digraph probability distribution with probability function

$$c \exp\left(f(x, v, w)\right)$$

where c is a normalizing constant. In statistical mechanics $f(x, v, w)$ then will be called a potential function, see [32]. When $f(x, v, w)$ is a linear combination as in (6) below, the distribution is an exponential random graph model for which the Metropolis–Hastings algorithm is treated in [41].

The actor-based opportunity model was proposed in [46] and, for a different data structure, in [45], as a stochastic actor-oriented model. In this model, the network dynamics is regarded as being driven by the social actors. Actors are assumed to control their outgoing ties, subject to inertia and the current network structure. This point of view is in accordance with the methodological approach of structural individualism [54,60], where actors are assumed to be purposeful and to behave subject to structural constraints. The purposes and constraints of the actors are summarized in the objective functions. One way to obtain the probabilities (2) is to assume that at each opportunity for change, actor i myopically optimizes the objective functions plus a random term, under the constraint that only one tie can change at a time. The myopia means that the actor only optimizes the state of the network that will be the immediate result of this change, without considering later network structures that might result in the future further ahead. The random term expresses otherwise unmodeled purposes and constraints. When the random terms have independent Gumbel distributions (the precise mathematical form of which does not matter for the present exposition), the choice probabilities are given by (2) (cf. [30,45,46]).

Simulation The Markov process defined above can be iterating by the following algorithm. The various steps in the algorithm can be derived using basic properties of Poisson processes and conditional probabilities (see [34]).

1. The process starts with a given time t and current state $X(t) = x^0$.
2. For the tie-based opportunity process define $\lambda = \sum_{ij} \lambda_{ij}$, and for the actor-based opportunity process $\lambda = \sum_i \lambda_i$.
   Let U be an independently drawn random number, uniformly distributed between 0 and 1, and let $\Delta t = -\ln(U)/\lambda$. Note that Δt has the exponential distribution with parameter λ.
   Change t into $t + \Delta t$.

3a. In the case of the tie-based opportunity process, choose a random pair (i, j) (with $i \neq j$) with probabilities λ_{ij}/λ. With probability given by (1), change $X_{ij}(t)$ into $1 - x^0_{ij}$.

3b. In the case of the tie-based opportunity process, choose a random actor i with probabilities λ_i/λ.
   To have a way for denoting the permitted new digraphs which are elements of $C_i(x^0)$, define by $x^0(i \rightsquigarrow j)$ for $j \neq i$ the digraph which is equal to x^0 except

only that $\left(x^0(i \rightsquigarrow j)\right)_{ij} = 1 - x^0_{ij}$; define $x^0(i \rightsquigarrow i) = x^0$. Then choose a random j with probabilities

$$p_i(x^0, x^0(i \rightsquigarrow j), v, w) = \frac{\exp\left(f_i(x^0, x^0(i \rightsquigarrow j), v, w)\right)}{\sum_{x' \in C(x^0)} \exp\left(f_i(x^0, x', v, w)\right)}, \quad (5)$$

which is just the same as (2). If $j \neq i$, change $X_{ij}(t)$ into $1 - x^0_{ij}$.

4. Go to step 1.

The stochastic process on the space of digraphs can be defined by this simulation algorithm just as well as by the Q-matrices (3) and (4), respectively.

Specification of the Change Determination Process

The changes can be regarded for tie-based opportunities as determinations of new values for X_{ij} according to a binary logistic regression model (see [23]); and for actor-based opportunities, according to a multinomial logistic regression model (see [29]). In the further elaboration the parallel with logistic regression is followed because the objective function f_i is specified as a linear combination

$$f_i(x^0, x, v, w) = \sum_k \beta_k \, s_{ki}(x^0, x, v, w) \quad (6)$$

where the functions s_{ki} are so-called *effects* driving the network dynamics while the weights β_k are parameters indicating the force of these effects and which can be estimated from the data.

The specification of the model will be the choice of a limited set of such effects for use in (6). A list of some effects is the following. In the formulae, replacing an index by a + means that a sum is taken over this index. Many examples of effects do not depend on x^0 or on v or w, and these arguments are then dropped from the notation.

Outdegree effect $s_{1i}(x) = x_{i+} = \sum_j x_{ij}$ This effect models the tendency to have ties at all; this tendency will also be influenced by all other effects, and therefore the interpretation of its parameter is conditional on the further selection of effects included in the model. In many models this is the only effect to which those actors j contribute who have no reciprocal link to i nor any links with any others to whom i is linked. In such models, its weight β_1 can be interpreted as the 'value' for actor i of a tie to such an other actor who is further completely isolated from i's personal network.

Reciprocity effect $s_{2i}(x) = \sum_j x_{ij} x_{ji}$ This is the number of reciprocated ties for actor i. It models the tendency toward reciprocation of choices. Thus, a higher value for its parameter β_2 will imply a higher tendency to forming reciprocated ties.

Degree Distribution The following three effects are related to modeling the dynamics of the degree distribution. Since the data are directed graphs, three distinct aspects of the degree distribution are the variability of indegrees, variability of outdegrees, and association between in- and outdegrees. Sometimes the term of preferential attachment is used [1,40] for the increased attractiveness of ties to nodes that already have high degrees. In our model these three aspects of preferential attachment can be expressed by including in the objective function terms depending on in- and out-degrees. The precise functional form for the effects is determined also by the requirement that the resulting model be amenable to statistical inference. Experience shows that using the square root of the degrees often leads to more stable estimation of the parameters than using the raw (untransformed) degrees. This suggests that, for many empirical networks, the models with squared roots of the degrees are better descriptions of reality than the models with untransformed degrees. Therefore only the models using the square roots are presented here.

Popularity effect (square root measure)
$s_{3i}(x) = \sum_j x_{ij} \sqrt{x_{+j}}$ This effect is defined by the sum of the square roots of indegrees of the others to whom i is tied. In other words, popularity of other actors is measured by the square root of their indegree. The root-popularity effect models the tendency to form ties to those actors who have high indegrees already (the Matthew effect in networks; see the chapter on ► Social Network Analysis, Graph Theoretical Approaches to). This will be reflected by the dispersion of the indegrees.

Activity effect (square root measure)
$s_{4i}(x) = \sum_j x_{ij} \sqrt{x_{j+}}$ This effect is defined by the sum of the square roots of outdegrees of the others to whom i is tied. The effect models the tendency to form ties to those actors who have high outdegrees already. This will be reflected by the association between indegrees and outdegrees.

Own outdegree, power 1.5, effect
$s_{4i}(x) = \sum_j x_{ij} \sqrt{x_{i+}} = x_{i+}^{1.5}$ The first expression show that each new tie has a 'value' equal to the actor's outdegree, comparing to a unit 'value' for the outdegree effect. This effect expresses that actors who already have many outgoing ties have a higher propensity to establish new ties. This will lead to a greater dispersion of outdegrees.

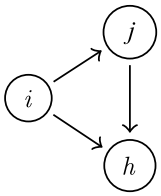

Network Analysis, Longitudinal Methods of, Figure 1
Transitive triplet

Network Closure The following two effects are related to network closure, also called transitivity or clustering: for friendship networks, this expresses that friends of friends tend to become friends. In addition to the two effects mentioned here, tendencies toward structural balance [7] and tendencies to avoid geodesic distances equal to 2 can also be implemented as effects that will lead to transitivity; also see [46,47].

Transitive triplets effect $s_{5i}(x) = \sum_{j,h} x_{ij}\, x_{ih}\, x_{jh}$ This formula represents the number of transitive patterns in i's ties as indicated in the figure below. A transitive triplet for actor i is a configuration (i, j, h) in which all three of the ties $i \to j, j \to h, i \to h$ are present (and irrespective of whether there are also other ties between these three actors).
This models the tendency toward network closure, where (for a positive parameter) formation of the tie $i \to h$ becomes increasingly likely when there are more indirect connections ('two-paths') $i \to j \to h$.

Transitive ties effect $s_{6i}(x) = \sum_h x_{ih}\ \max_j(x_{ij}\, x_{jh})]$ This is the number of actors j to whom i is directly as well as indirectly tied, i. e., for which there exists at least one h such that (i, j, h) is a transitive triplet.
The transitive triplets and transitive ties effects are two distinct ways of modeling tendencies toward network closure. For the effect on the probability of forming the tie $i \to h$, the number of two-paths $i \to j \to h$ makes no difference for the transitive ties effect, as long as there is at least one indirect connection; for the transitive triplets effect the contribution to the objective function increases linearly with the number of two-paths.

Three-cycle effect $s_{7i}(x) = \sum_{j,h} x_{ij}\, x_{jh}\, x_{hi}$ This is the number of three-cycles $i \to j \to h \to i$ in which actor i is involved. This effect models the tendency toward forming three-cycles, which is the simplest form of generalized exchange [3,27] (i gives to j, j gives to h, and h gives to i) and which is opposed to hierarchy.

Covariate Effects Actor covariates v can influence the propensities to form or terminate ties in different ways, because this propensity might be influenced by the value for the sender or the receiver of the tie, or by some combination of these two values.

v-related popularity $s_{8i}(x, v) = \sum_j x_{ij}\, v_j$ The v-related popularity effect is defined by the sum of the covariate over all actors to whom i is tied. Positive parameter values will imply that ties to actors with high v values are more attractive. This will lead to a tendency toward a correlation between v_i and the indegree of i.

v-related activity $s_{9i}(x, v) = v_i\, x_{i+}$ In words, this is v_i times i's outdegree. Positive parameter values will imply that actors with high v values tend to make more ties. This will lead to a tendency toward a correlation between v_i and the outdegree of i.

v-related similarity $s_{10,i}(x, v) = \sum_j x_{ij}\, \mathrm{sim}_v(i, j)$ Here $\mathrm{sim}_v(i, j)$ indicates the similarity between actors i and j defined by $\mathrm{sim}_v(i, j) = 1 - (|v_i - v_j|/\Delta)$, where $\Delta = \max_{h,k} |v_h - v_k|$ is the observed range of the covariate v. Thus, the effect is is the sum of similarities between i and the others to whom he is tied. Positive parameter values will imply that there is a preference for ties between actors with similar values of v_i and v_j.

v ego-alter interaction $s_{11,i}(x, v) = \sum_j x_{ij}\, v_i\, v_j$ This product interaction is an alternative to the similarity effect for expressing that the propensity for a tie to exist depends on the combined values of the v for the sending actor ('ego', i) and the receiving actor ('alter', j).

Main effect of $w s_{12,i}(x, w) = \sum_j x_{ij}\, w_{ij}$ For a dyadic covariate w, this is defined by the sum of the values of w_{ij} for all other actors j to whom i is tied.

It is clear that the list can be extended indefinitely, and that researchers have to make a limited choice reflecting theoretical and content-matter knowledge and interest. The potential complexity of network dynamics justifies to have many candidate effects that may be used to model the network dynamics. For instance, when a researcher wishes to estimate a model for testing whether there is a preference for choosing network partners with similar v values, while controlling for the tendency to have ties, and the tendencies for reciprocation and transitive closure, then effects s_1, s_2, s_5 and/or s_6, and s_{10} should be included in the model (6). Since transitive closure could be expressed by effects s_5 as well as s_6, there may be no strong prior arguments for choosing between these two 'control' effects, and empirical grounds could be used to choose either one or both. If there are grounds to suspect that the v_i values may also be associated with in- or outdegrees, the popularity and activity effects s_8 and s_9 may also be included.

Specification of the Change Opportunity Process

For the model specification it should be noted that the 'social time' which determines the speed of change of the network is not necessarily the same as the physical time elapsing between consecutive observation moments. Given the absence of an extraneous definition of this 'social time', it is not a restriction to set to 1 the total time elapsed between each pair of consecutive observations. If there are $M \geq 3$ observation moments, it is advisable to specify distinct rate parameters ρ_m governing the frequency of opportunities for change between t_m and t_{m+1}, and allow ρ_1, ρ_2, etc., to be different. If the change rate further is constant (independent of actors), ρ_m then represents the expected number of opportunities for change between t_m and t_{m+1}. This is the expected number per ordered pair (i, j) in the case of tie-based opportunities, and per actor i in the case of actor-based opportunities. The symbol ρ will denote the vector $(\rho_1, \ldots, \rho_{M-1})$.

In the more general case the rate function can be defined, e. g., for the actor-based model, by a function depending on actor covariates and positional characteristics of the actors. When, for example, a dependence on one covariate v_i and the current out-degree x^0_{i+} is considered, a logarithmic link function could be used giving a model such as

$$\lambda_i(x^0, v) = \rho_m \exp\left(\alpha_1 v_i + \alpha_2 x_{i+}\right).$$

In general the symbol α will be reserved for parameters indicating dependence of the rate function on covariates and network characteristics.

Statistical Estimation and Testing

The most usual type of longitudinal network data is panel data, where for $M \geq 2$ time points, an observation $x(t_m)$ is available of the network on the same set $\{1, \ldots, n\}$ of actors.

These models can be simulated on computers in rather straightforward ways (the algorithm is written out in [47]). Parameter estimation, however, is more complicated, because the likelihood function or explicit probabilities can be computed only for uninteresting models. This section presents the Method of Moments (MoM) estimates proposed in [46]. Maximum Likelihood (ML) estimators are presented in [48]. In the current implementation, ML estimators are more time-consuming than MoM estimators, and can be used only for relatively small data sets. For the more straightforward models (dynamics of networks only, no endowment functions), the MoM method is hardly less efficient than the ML method. For more complicated models, where it is important to squeeze every bit of information out of the data, it can be useful to employ ML methods. In the following description of the estimation method, the parameter vector (ρ, α, β) is denoted by θ.

It is undesirable to make the restrictive assumption that the distribution of the process is stationary. Instead, for each observation moment t_m $(m = 1, \ldots, M-1)$ the observed network $\mathbf{x}(t_m)$ can be used as a conditioning event for the distribution of $\mathbf{X}(t_{m+1})$. The Method of Moments requires that a vector of statistics $U_{m+1} = U\big(\mathbf{X}(t_m), \mathbf{X}(t_{m+1})\big)$ is utilized, such that the expected value

$$E_\theta\left\{U\big(\mathbf{X}(t_m), \mathbf{X}(t_{m+1})\big) \mid \mathbf{X}(t_m) = \mathbf{x}(t_m)\right\}$$

is sensitive to the parameter θ. Given the conditioning on the preceding observation, the moment equations, or estimating equations, can then be written as

$$\sum_{m=1}^{M-1} E_\theta\left\{U\big(\mathbf{X}(t_m), \mathbf{X}(t_{m+1})\big) \mid \mathbf{X}(t_m) = \mathbf{x}(t_m)\right\} = \sum_{m=1}^{M-1} U\big(\mathbf{x}(t_m), \mathbf{x}(t_{m+1})\big). \quad (7)$$

It turns out that suitable statistics are the following. The number of changed ties between consecutive observations,

$$\sum_{i,j} \left|X_{ij}(t_{m+1}) - X_{ij}(t_m)\right|,$$

is especially sensitive to the rate of change ρ_m. A vector of statistics sensitive especially to β is the sum of the individual objective functions

$$\sum_i f_i\big(\mathbf{X}(t_{m+1})\big).$$

To solve the estimating equation (7), in the absence of ways to calculate analytically the expected values, stochastic approximation methods can be used. Variants of the Robbins–Monro [9,42] algorithm have been used with good success. This is a stochastic iteration method which produces a sequence of estimates $\theta^{(N)}$ which is intended to converge to the solution of (7), and which works here as follows. For a given provisional estimate $\theta^{(N)}$, the model is simulated so that for each $m = 1, \ldots, M-1$, a simulated random draw is obtained from the conditional distribution of $\mathbf{X}(t_{m+1})$ conditional on $\mathbf{X}(t_m) = \mathbf{x}(t_m)$. This simulated network is denoted $\mathbf{X}^{(N)}(t_{m+1})$. Denote $U^{(N)}_m = U\big(\mathbf{x}(t_m), \mathbf{X}^{(N)}(t_{m+1})\big)$ and $U^{(N)} = \sum_{m=1}^{M-1} U^{(N)}_m$,

and let u^{obs} be the right-hand side of (7). Then the iteration step in the Robbins–Monro algorithm for obtaining the Method of Moments estimate is given by

$$\theta^{(N+1)} = \theta^{(N)} - a_N D^{-1} \left(U^{(N)} - u^{\text{obs}} \right) , \qquad (8)$$

where D is a suitable matrix and a_N a sequence of positive constants tending to 0. Tuning details of the algorithm, including the choices of D and a_N, are given in [46]. The experience with the convergence of this algorithm is quite good. The standard errors can be computed using the standard formulae of standard errors for the Method of Moments, based on the delta method, and applying simulation methods; such simulation methods are discussed in [44]. Bayesian estimators for these models are presented in [24] and Maximum Likelihood estimators in [48].

Example: Dynamics of Adolescent Friendship

As an example, the adolescent friendship network is considered of a year cohort at a secondary school in Glasgow (Scotland), studied in the Teenage Friends and Lifestyle Study [36,38]. This data set was collected at three measurement points t_1, t_2, t_3 in 1995–1997, at intervals of roughly one year, starting when the pupils were 12–13 years old. Here the network is studied that is formed by the 129 (out of 160) pupils who were present at all three measurement waves. Sex (boys scored as 1, girls as 2) and drinking behavior are used as actor variables, drinking (alcohol consumption) being measured on a 5-point scale ranging from 1 (not at all) to 5 (more than one a week). Both variables are centered around the mean, which is 1.43 for sex and 2.60 for drinking (averaged over t_1 and t_2). The data set is used as an illustration; more wide-ranging analyses are presented in [37,52,53].

Three models are presented. Rate parameters are assumed constant within periods between observation moments, and the duration of the periods is (arbitrarily but without loss of generality) set at 1. The first model contains only the most basic dyadic and triadic effects: outdegree, reciprocity, transitive triplets, and three-cycles. The second adds to this the three effects to model more precisely the degree distribution. The third model adds to these structural effects the effects of two covariates: gender and alcohol consumption.

Parameter estimates are approximately unbiased and normally distributed; for this assertion there is no mathematical proof yet, but it is supported by simulation studies. Therefore, effects can be tested by referring the studentized estimates (or t-ratios, i. e., estimate divided by standard error) to a standard normal distribution. When the t-ratio exceeds 2 in absolute value, the effect can be interpreted as being significant at the significance level of 5 %.

The table present first the estimated rate parameters. These indicate that the pupils had about 12 opportunities for changes in the first period (t_1 – t_2), and about 9 in the second period (t_2 – t_3).

The parameters for the objective function, which the table shows next, are more important for the interpretation. There are strong tendencies toward reciprocity and transitivity, and a tendency away from three-cycles. This is the case in all three models, although the parameter estimates are slightly different. This indicates that the structural features of transitive closure and of local hierarchy (the interpretation of the negative three-cycle effect) cannot be 'explained away' by tendencies in tie formation and dissolution that are associated to degrees, sex, or alcohol consumption.

The degree effects in Models 2 as well as 3 show that there is a positive popularity effect – with a borderline significance at the 5% level in Model 3; there is a negative activity effect; and a negative effect of own outdegree raised to the power 1.5. The positive popularity effect suggests that differential values in in-degrees tend to be self-sustaining, leading to rather strongly dispersed in-degrees. The other two, negative, degree-related effects indicate that differences in out-degrees, as well as correlations between in- and out-degrees, tend to be self-correcting, leading to relatively low dispersions of out-degrees and low in-degree – out-degree correlations.

Of the three gender-related effects, only the similarity effect is significant. In this age range, a strong preference for same-sex friendships is to be expected (the interest which perhaps exists in the other sex is not reported as friendship). For alcohol consumption, the interaction effect shows that those who drink more themselves have a higher preference for friends who also drink more; and the activity effect shows that those drinking more tend to mention less friends, for friends of average drinking habits (where the contribution of the interaction is nil). This is most clearly expressed by jointly considering the three contributions related to drinking behavior that can be made by the tie $i \to j$. In a formula, this is represented by the coefficient of the variable y_{ij} in the objective function. Recall that the actor variables are centered around the mean. Using the formulae for effects s_8, s_9, and s_{11} and filling in the coefficients in Table 1 yields

$$-0.026\,(v_j - \bar{v}) - 0.086\,(v_i - \bar{v}) + 0.107\,(v_j - \bar{v})(v_i - \bar{v}) ,$$

with $\bar{v} = 2.60$. The following picture illustrates the contributions to the objective function made by ego's drinking behavior v_i and alter's drinking behavior v_j.

Network Analysis, Longitudinal Methods of, Table 1
Parameter estimates for modeling evolution of friendship network, Glasgow school cohort. Standard errors between parentheses

	Model 1		Model 2		Model 3	
Effect	par.	(s.e.)	par.	(s.e.)	par.	(s.e.)
Rate 1	11.82	(1.00)	12.05	(1.15)	12.19	(1.13)
Rate 2	9.23	(0.83)	9.05	(0.79)	9.09	(0.77)
Outdegree	−2.697	(0.047)	−0.16	(0.34)	−0.18	(0.41)
Reciprocity	2.38	(0.10)	2.48	(0.12)	2.22	(0.12)
Transitive triplets	0.459	(0.033)	0.569	(0.036)	0.544	(0.034)
Three -cycles	−0.57	(0.10)	−0.55	(0.11)	−0.44	(0.11)
Popularity (sq. root)	–		0.223	(0.095)	0.184	(0.093)
Activity (sq. root)	–		−0.89	(0.14)	−0.92	(0.17)
Own outdegree, power 1.5	–		−0.560	(0.084)	−0.587	(0.094)
Sex (F) popularity	–		–		−0.16	(0.11)
Sex (F) activity	–		–		0.12	(0.12)
Sex similarity	–		–		0.904	(0.097)
Drinking popularity	–		–		−0.026	(0.030)
Drinking activity	–		–		−0.086	(0.038)
Drinking ego × alter	–		–		0.107	(0.024)

It can be concluded that those who do not drink alcohol ($v_i = 1$), prefer friends who drink no or little alcohol, while the reverse is true for those who drink a lot of alcohol. The range of this contribution is 0.8 for those with $v_i = 1$ and 0.9 for those with $v_i = 5$, which is comparable to the value 0.9 of the gender similarity effect. Thus, for those with the highest and the lowest values of alcohol use, the importance of the alcohol use of potential friends, when comparing potential friends with the minimum and the maximum alcohol use, is approximately as great as the importance of their gender. On the other hand, for those with medium values of alcohol use, the alcohol use of potential friends plays virtually no role at all.

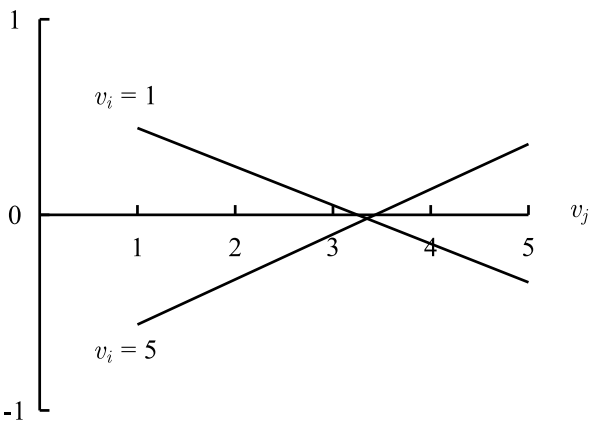

Network Analysis, Longitudinal Methods of, Figure 2
Contributions of drinking behaviors v_i, v_j to the objective function for friendship

Models for the Co-evolution of Networks and Behavior

The importance of networks derives, for an important part, from the effects of networks on the behavior and performance of the actors. This can be, for example, because of influence between friends or between collaborating partners [16], because of exchange of resources [28], or because of structural advantages [4].

A few examples drawn from the many studies in this field are concerned with influence of friends of adolescents on smoking behavior [14,37,52], effects of acquaintances on labor market outcomes [17], competitive advantage of firms [18], effects of friends on delinquency [6,19], effects of position in patent citation networks on growth rates of companies [39], and job performance [51].

This type of changing individual attributes, which might range from, e. g., behavioral tendencies and attitudes of individuals to performance of companies, will be briefly referred to as 'behavior'. When the ties in the network are influenced by the behavior and the behavior, in turn, is influenced by the network, a mutual feedback arises between the network and the behavior. The network structure of the ties between the actors together with their behavior constitute the endogenously changing environment for each of the actors [61]. Thus, this approach is well-suited to study macro-micro-macro questions of the type discussed by [11].

The vector of attributes for actor i at time t is denoted

$$z_i(t) = \left(z_{i1}(t), \ldots, z_{iH}(t)\right)$$

where $z_{ih}(t)$ denotes the hth attribute of actor i. For the n actors these are stacked in the matrix $z(t)$, which is regarded as the outcome of a random matrix $Z(t)$.

To model the co-evolution of networks and behavior, where the network and the behavior influence one another dynamically, the stochastic process $\big(X(t), Z(t)\big)$ is considered. This is treated in quite the same way as the stochastic process $X(t)$ was treated above. It is assumed that $z_i(t)$ represents an enduring, but changeable, state of actor i rather than a momentary behavior, so that it is a sensible approximation to assume that $\big(X(t), Z(t)\big)$ is a Markov process in continuous time. Now the change probabilities of $X(t)$ will depend on the current state of $X(t)$ as well as $Z(t)$. The fact that network change is co-determined by behavior is called *behavior-dependent selection*, and will often be referred to briefly as *selection*. Similarly, the change probabilities of $Z(t)$ will depend on the current state of $Z(t)$ as well as $X(t)$, and this will be called *influence*. This terminology was also used, e. g., by [14].

To remain close to the framework for network modeling, it is assumed that all of the behavior variables z_{ih} are measured on a discrete ordinal scale, with values coded as consecutive integers $\{0, 1, \ldots, M_h\}$ for some $M_h \geq 1$. The analogue of the simplifying assumption (3) made above for network change is the following. This decomposes the change between consecutive observations $\big(x(t_m), z(t_m)\big)$ and $\big(x(t_{m+1}), z(t_{m+1})\big)$ into a sequence of the smallest possible steps.

3′. At any given moment t, no more than one of all the variables $X_{ij}(t), Z_{ih}(t)$ can change. When $Z_{ih}(t)$ changes, at any given instant it can change only to an immediate neighboring value, i. e., by a decrease or increase of 1 (permitted only if this step does not take Z_{ih} outside of the permitted range from 0 to M_h).
This means that there is no direct coordination between changes in ties and changes in behavior, and the dependence between networks and behavior is brought about because both react to each other.

The model for the network is just like it was above, the rate function and objective function now being denoted by λ^X and f_i^X, and allowed to depend on the current behavior $Z(t)$, to represent behavior-dependent selection. For each of the dependent behavior variables Z_h there also is a rate function λ_i^{Zh} driving the frequency of changes and an objective function f_i^{Zh} defining the probabilities of behavior changes when there is an opportunity of change. Their dependence on the current network will represent influence.

For each actor i, opportunities to change behavior Z_{ih} occur according to a Poisson process with rate λ_i^{Zh}.

When actor i has the opportunity to change behavior Z_{ih} given that currently $(X(t), Z(t)) = (x^0, z^0)$, there are three possible new states (x^0, z), where either $z = z^0$ or the only difference between z and z^0 is that $z_{ih} = z_{ih}^0 - 1$ or $z_{ih} = z_{ih}^0 + 1$; unless one of these values is outside the range $\{0, \ldots, M_h\}$, in which case there are only the two remaining possible new states.

The probabilities of going from state (x^0, z^0) to state (x^0, z) are proportional to $\exp\big(f_i^{Zh}(z^0, z, x^0, v, w)\big)$.

The change probabilities in this model are given by

$$\begin{aligned} &\mathrm{P}\{Z(t) \text{ changes to } z \mid i \text{ has an opportunity to change} \\ &\quad Z_{ih} \text{ at time } t, X(t) = x^0, Z(t) = z^0\} \\ &\qquad = \frac{\exp\big(f_i^{Zh}(z^0, z, x^0, v, w)\big)}{\sum_{z' \in C^{Zh}(z^0)} \exp\big(f_i^{Zh}(z^0, z', x^0, v, w)\big)}\,, \end{aligned} \tag{9}$$

where $C^{Zh}(z^0)$ is the set of two or three permitted new values for Z_h.

The effects in the objective function f_i^X for network changes now are denoted by s_{ki}^X with weights β_k^X, and likewise the objective function for behavior changes is assumed to be expressed as a linear combination of effects weighted by parameters,

$$f_i^{Zh}(z^0, z, x^0, v, w) = \sum_k \beta_k^{Zh}\, s_{ki}^{Zh}(z^0, z, x^0, v, w)\,. \tag{10}$$

A set of possible effects that could be included in the objective function for behavior are the following. For simplicity of notation, it is assumed that there is only one behavior variable, so that the index h can be dropped from the notation. The first two effects are used to define the shape of the objective function as a function of z_i, and other terms depending on z_i could be added.

Linear Shape Effect $s_{i1}^Z(z) = z_i$

Squared Shape Effect $s_{i2}^Z(z) = z_i^2$ The linear and squared shape effects together define what could be regarded as a quadratic preference function on the behavior, $\beta_1^Z z_i + \beta_2^Z z_i^2$. The word 'preference function' is used with some reluctance, because it is used here only as an easy shorthand term for the combined short-term result of preferences and constraints, depending only on the actor's behavior z_i itself, net of the other terms in

the behavior objective function. If $\beta_2 < 0$ this is a unimodal function of z_i. If $\beta_2 > 0$, on the other hand, and if the minimum of the quadratic function is assumed within the range of the behavior variable, then the behavior is drawn to the extremes of the range, with actors already low on z_i being drawn to low values and actors already high on z_i being drawn to high values. This can represent, e. g., addictive behavior.

Other nonlinear functions of z_i could, of course, also be included.

Indegree effect $s^Z_{i3}(z, x) = z_i\, x_{+i}$ This represents that actors with a high indegree ('popular' actors) have a higher tendency toward high values of the behavior.

Outdegree effect $s^Z_{i4}(z, x) = z_i\, x_{i+}$ This represents that actors with a high outdegree ('active' actors) have a higher tendency toward high values of the behavior.

Total similarity effect $s^Z_{i5}(z, x) = \sum_j x_{ij}\, \mathrm{sim}_z(i, j)$ The dyadic similarity sim_z is as defined above, now applied to the dependent behavior Z. The total similarity effects adds the similarity values between i and the actors toward whom i has a tie. This is the primary representation of social influence: the preference for behavior which is close to that of one's network members.

Average alter effect $s^Z_{i6}(z, x) = z_i\, x^{-1}_{i+} \sum_j x_{ij}\, z_j$ The coefficient of z_i is the average behavior of the actors to whom i is tied (i's 'alters', j). The coefficient is defined as 0 when the outdegree x_{i+} is 0. This is another representation of social influence: the preference for behavior depends on the average behavior of one's network members.

The parameter estimation for this model is discussed in [49].

Example: Co-evolution of Adolescent Friendship and Alcohol Use

As an example for the co-evolution of networks and behavior the data of the Glasgow school cohort is used again, but now the alcohol consumption is used as a dependent, or endogenous, variable. In the treatment in Table 1 the alcohol consumption was used as an exogenous variable, i. e., its values were accepted as if determined by processes independent of the network. Now we follow a co-evolution approach where it is assumed that the dynamics in alcohol consumption can be co-determined by the network just as the network dynamics can be co-determined by the alcohol consumption.

To obtain interpretations of the numerical parameter values, it must be noted that this table refers to the centered actor variables. The raw variable ranges from 1 to 5, with averages at the three observations rising from 2.5 through 2.7 to 3.1. The overall mean, 2.8, is subtracted from the observations. We shall denote by z_i the raw value of actor i's alcohol consumption scored 1–5, by $\bar{z}$ the average equal to 2.8, and by $\tilde{z}_i$ the average of the alcohol consumption of i's friends.

Network Analysis, Longitudinal Methods of, Table 2
Parameter estimates for modeling co-evolution of friendship network and alcohol consumption, Glasgow school cohort

Effect	par.	(s.e.)
Friendship dynamics		
Rate 1	12.13	(1.26)
Rate 2	9.09	(0.97)
Outdegree	−0.22	(0.43)
Reciprocity	2.19	(0.12)
Transitive triplets	0.529	(0.036)
Three -cycles	−0.42	(0.10)
Popularity (sq. root)	0.193	(0.086)
Activity (sq. root)	−0.92	(0.15)
Own outdegree, power 1.5	−0.581	(0.099)
Sex (F) popularity	−0.16	(0.10)
Sex (F) activity	0.12	(0.13)
Sex similarity	0.90	(0.11)
Drinking popularity	−0.026	(0.043)
Drinking activity	−0.118	(0.057)
Drinking ego × alter	0.166	(0.043)
Alcohol consumption dynamics		
Rate 1	1.48	(0.25)
Rate 2	2.22	(0.37)
Linear shape	0.36	(0.42)
Squared shape	−0.34	(0.14)
Indegree	0.07	(0.12)
Outdegree	−0.08	(0.17)
Sex (F)	−0.02	(0.23)
Average alter	0.83	(0.35)

The interpretation of the friendship dynamics is quite the same as in the analysis where alcohol was treated as an exogenous variable. For the dynamics of alcohol consumption, the indegree, outdegree, and sex of the actor do not seem to have an important influence. The effect of the average drinking behavior of the friends of the focal actor does have a significant effect, however, with a t-value of $0.83/0.35 = 2.4$. When we ignore the small and nonsignificant effects of outdegree, indegree, and sex, the remaining part of the objective function for drinking behavior is

$$\begin{aligned} &0.36\,(z_i - \bar{z}) + 0.83(\tilde{z}_i - \bar{z}))(z_i - \bar{z}) - 0.34(z_i - \bar{z})^2 \\ &= \big(0.36 + 0.83(\tilde{z}_i - \bar{z})\big)(z_i - \bar{z}) - 0.34(z_i - \bar{z})^2\,. \end{aligned}$$

This is a quadratic function, unimodal, with a maximum at

$$z_i = \bar{z} + \frac{0.83}{2 \times 0.34}(\bar{z}_i - \bar{z}) = -0.62 + 1.22\,\bar{z}_i\,.$$

Taking account of the integer values of z_i, this implies that those with friends with the smallest possible average of smoking behavior $\bar{z}_i = 1$ are drawn towards the 'preferred' value of 1 themselves, while those having friends who have the highest possible average $\bar{z}_i = 5$ are themselves also drawn toward this value 5 as a 'preferred' value. It can be concluded that the data provide support for the existence, in the co-evolution of friendship and drinking tendencies, of selection (tendency to choose friends with similar behavior) as well as influence (tendency to change behavior in the direction of friends' behavior).

Extensions

The basic model specifications defined above can be extended in various ways. Above we already mentioned the possibility to let the rates of change depend on covariates or on current network structure. Another possibility is to introduce an asymmetry between the values of ties when they are formed and their values when they are lost. E.g., for friendship dynamics, there is theoretical and empirical evidence that the additional 'value' of a tie added by its being reciprocated is higher when considering a potential loss of the tie than when considering the potential new formation of the tie. This asymmetry can be modeled by endowment functions, see [49]. Technically, this means that the effects $s_{ki}(x^0, x, v, w)$ used in (6) depend not only on the new state x but, unlike the examples given above, also on the preceding state x^0. For example, the endowment effect of a tie being reciprocal is expressed by

$$s_{ki} = \sum_j x^0_{ij}\, x_{ij}\, x_{ji}\,,$$

which is sensitive to the number of i's reciprocal ties only when they are candidates to being terminated and not when they are candidates to being created.

Similarly, going upward on a behavior variable might be not the opposite of going downward, which can be modeled by endowment effects in the objective function for behavior.

Future Directions

Although the statistical modeling of network dynamics started already with [22] and [59], this area has been in rapid development only since recent years. Much work remains to be done, however, to extend these methods to other types of network data and to study their properties. The lists of effects that can be included in the objective functions illustrate the flexibility of this model and its adaptability to research questions and network data. The availability of methods for analysis of network panel data has been a stimulus also for the further collection of such data.

Plausible models and good methods for parameter estimation and testing have now been developed, as summarized in this chapter, and they are available in the SIENA (*Simulation Investigation for Empirical Network Analysis*) program. This program is available as freeware with additional material on the website http://stat.gamma.rug.nl/snijders/siena.html and has an extensive manual [50]. The examples presented here were analyzed using this program.

The tests used in the examples, based on studentized parameter estimates, can be regarded as Wald-type tests. Some limited simulation studies have supported the validity of these tests. Score-type tests associated to the Method of Moments estimators were developed in [43]. The examples in this paper underline the need for methods to assess fit of models, and also to compare non-nested models. This could be done formally based on estimated likelihoods, or informally based on the comparison of observed and expected values of relevant statistics that are not used for parameter estimation.

The open question of assessing fit also invites speculation about the robustness of the results against the use of models of which the fit is not beyond doubt. There is an inherent tension between the complexity of processes of network dynamics, and the limited amount of data that can in practice be observed concerning these processes. One issue is that the models proposed here are Markov processes. For two-wave data sets there are no clear alternatives to making such an assumption, but the assumption is certainly debatable. Including more information in the state space (by using covariates, by considering valued rather than dichotomous ties, etc.) may relax the doubts concerning such an assumption.

Another issue is the difference between the tie-oriented and actor-oriented models. Which type of model is to be preferred is a matter both of social science theory and of empirical fit. It will be important to know, supported by simulation studies and/or mathematical results, the extent to which results based on particular models for network dynamics are robust to deviations from the precise assumptions made. In addition, it will be useful to develop still other models, e. g., models accounting for actor heterogeneity (like were developed for non-longitudinal network data, e. g., by [20,35,58]) or measurement error.

Other open questions are about mathematical properties of the estimators and tests proposed. Simulation studies support the conjecture that the Method of Moments estimators have asymptotically normal distributions, but this has not been proven. It is unknown if the solution to the moment equation (7), under certain conditions, is unique. Similar questions can be asked about the Maximum Likelihood estimators. All this indicates that there is ample scope for future work on methods of statistical inference for network dynamics.

Bibliography

Primary Literature

1. Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
2. Bartholomew DJ (1982) Stochastic Models for Social Processes, 3rd edn. Wiley, New York
3. Bearman PS (1997) Generalized exchange. Am J Sociol 102:1383–1415
4. Burt RS (1992) Structural Holes. Harvard University Press, Cambridge
5. Brass DJ, Galaskiewicz J, Greve HR, Tsai W (2004) Taking stock of networks and organizations: a multilevel perspective. Acad Manag J 47:795–817
6. Burk WJ, Steglich CEG, Snijders TAB (2007) Beyond dyadic interdependence: Actor-oriented models for co-evolving social networks and individual behaviors. Int J Behav Dev 31:397–404
7. Cartwright D, Harary F (1956) Structural Balance: A Generalization of Heiders Theory. Psychol Rev 63:277–292
8. Checkley M, Steglich CEG (2007) Partners in Power: Job Mobility and Dynamic Deal-Making. Eur Manag Rev 4:161–171
9. Chen H-F (2002) Stochastic Approximation and its Applications. Kluwer, Dordrecht
10. Coleman JS (1964) Introduction to Mathematical Sociology. The Free Press of Glencoe, New York
11. Coleman JS (1990) Foundations of Social Theory. Belknap Press of Harvard University Press, Cambridge/London
12. de Federico de la Rua A (2003) La dinamica de las redes de amistad. La eleccion de amigos en el programa Erasmus. REDES 4.3, http://revista-redes.rediris.es. Accessed 10 Aug 2007
13. de Nooy W (2002) The dynamics of artistic prestige. Poetics 30:147–167
14. Ennett ST, Bauman KE (1994) The Contribution of Influence and Selection to Adolescent Peer Group Homogeneity: The Case of Adolescent Cigarette Smoking. J Personal Soc Psychol 67:653–63
15. Doreian P, Stokman FN (eds) (1997) Evolution of Social Networks. Gordon and Breach, Amsterdam
16. Friedkin NE (1998) A Structural Theory of Social Influence. Cambridge University Press, Cambridge
17. Granovetter MS (1973) The strength of weak ties. Am J Sociol 78:1360–1380
18. Gulati R, Nohria N, Zaheer A (2000) Strategic Networks. Strateg Manag J 21:203–215
19. Haynie DL (2001) Delinquent Peers Revisited: Does Network Structure Matter? Am J Sociol 106:1013–1057
20. Hoff PD, Raftery AE, Handcock MS (2002) Latent Space Approaches to Social Network Analysis. J Am Stat Assoc 97:1090–1098
21. Holland PW, Leinhardt S (1975) Local structure in social networks. Sociol Methodol – 1976 1–45
22. Holland PW, Leinhardt S (1977) A dynamic model for social networks. J Math Sociol 5:5–20
23. Hosmer D, Lemeshow S (2000) Applied logistic regression, 2nd edn. Wiley-Interscience, New York
24. Koskinen JH, Snijders TAB (2007) Bayesian inference for dynamic social network data. J Stat Plan Inference 137:3930–3938
25. Kossinets G, Watts DJ (2006) Empirical Analysis of an Evolving Social Network. Science 311:88–90
26. Leenders RTAJ (1995) Models for network dynamics: A Markovian framework. J Math Sociol 20:1–21
27. Lévi-Strauss Ce (1969; orig. 1947) The Elementary Structures of Kinship. Beacon Press, Boston
28. Lin N, Cook K, Burt RS (eds) (2001) Social Capital. Theory and Research. Aldine de Gruyter, New York
29. Long JS (1997) Regression Models for Categorical and Limited Dependent Variables. Sage Publications, Thousand Oaks
30. Maddala GS (1983) Limited-dependent and Qualitative Variables in Econometrics, 3rd edn. Cambridge University Press, Cambridge
31. Newcomb TM (1962) Student Peer-Group Influence. In: Sanford N (ed) The American College: A Psychological and Social Interpretation of the Higher Learning. Wiley, New York
32. Newman MEJ, Barkema GT (1999) Monte Carlo methods in Statistical Physics. Clarendon Press, Oxford
33. Newman MEJ, Watts DJ, Strogatz SH (2002) Random graph models of social networks. Proceedings of the National Academy of Sciences USA, 99, 2566–2572, National Academy of Sciences, Washington
34. Norris JR (1997) Markov Chains. Cambridge University Press, Cambridge
35. Nowicki K, Snijders TAB (2001) Estimation and prediction for stochastic blockstructures. J Am Stat Assoc 96:1077–1087
36. Pearson MA, Michell L (2000) Smoke Rings: Social network analysis of friendship groups, smoking and drug-taking. Drugs Educ Prev Policy 7:21–37
37. Pearson M, Steglich C, Snijders T (2006) Homophily and assimilation among sport-active adolescent substance users. Connections 27(1):47–63
38. Pearson M, West P (2003) Drifting Smoke Rings: Social Network Analysis and Markov Processes in a Longitudinal Study of Friendship Groups and Risk-Taking. Connections 25(2): 59–76
39. Podolny JM, Stuart TE, Hannan MT (1997) Networks, knowledge, and niches: competition in the worldwide semiconductor industry, 1984–1991. Am J Sociol 102:659–689
40. de Solla Price D (1976) A general theory of bibliometric and other advantage processes. J Am Soc Inf Sci 27:292–306
41. Robins GL, Woolcock J, Pattison P (2005) Small and other worlds: Global network structures from local processes. Am J Sociol 110:894–936
42. Robbins H, Monro S (1951) A stochastic approximation method. Ann Math Stat 22:400–407
43. Schweinberger M (2007) Statistical modeling of digraph panel data: Goodness-of-fit. Submitted for publication

44. Schweinberger M, Snijders TAB (2006) Markov models for digraph panel data: Monte Carlo-based derivative estimation. Comput Stat Data Anal 51:4465–4483
45. Snijders TAB (1996) Stochastic actor-oriented dynamic network analysis. J Math Sociol 21:149–172
46. Snijders TAB (2001) The statistical evaluation of social network dynamics. In: Sobel M, Becker M (eds) Sociological Methodology. Basil Blackwell, Boston and London, pp 361–395
47. Snijders TAB (2005) Models for longitudinal network data. In: Carrington PJ, Scott J, Wasserman S (eds) Models and Methods in Social Network Analysis. Cambridge University Press, New York
48. Snijders TAB, Koskinen JH, Schweinberger M (2007) Maximum Likelihood Estimation for Social Network Dynamics. Submitted for publication
49. Snijders TAB, Steglich CEG, Schweinberger M (2007) Modeling the co-evolution of networks and behavior. In: van Montfort K, Oud H, Satorra A (eds) Longitudinal models in the behavioral and related sciences. Lawrence Erlbaum, Mahwah, pp 41–71
50. Snijders TAB, Steglich CEG, Schweinberger M, Huisman M (2007) Manual for SIENA version 3. ICS, University of Groningen; Groningen, Department of Statistics, University of Oxford, Oxford http://stat.gamma.rug.nl/snijders/siena.html. Accessed 4 Dec 2007
51. Sparrowe RT, Liden RC, Wayne SJ, Kraimer ML (2001) Social Networks and the performance of individuals and groups. Acad Manag J 44:316–325
52. Steglich CEG, Snijders TAB, Pearson M (2007) Dynamic networks and behavior: Separating selection from influence. Submitted for publication
53. Steglich CEG, Snijders TAB, West P (2006) Applying SIENA: An Illustrative Analysis of the Coevolution of Adolescents' Friendship Networks, Taste in Music, and Alcohol Consumption. Methodology 2:48–56
54. Udehn L (2002) The changing face of methodological individualism. Ann Rev Sociol 8:479–507
55. van de Bunt GG, van Duijn MAJ, Snijders TAB (1999) Friendship networks through time: An actor-oriented statistical network model. Comput Math Organ Theory 5:167–192
56. van de Bunt GG, Wittek RPM, de Klepper MC (2005) The Evolution of Intra-Organizational Trust Networks; The Case of a German Paper Factory: An Empirical Test of Six Trust Mechanisms. Int Sociol 20:339–369
57. van Duijn MAJ, Zeggelink EPH, Huisman M, Stokman FN, Wasseur FW (2003) Evolution of Sociology Freshmen into a Friendship Network. J Math Sociol 27:153–191
58. van Duijn MAJ, Snijders TAB, Zijlstra BH (2004) p_2: a random effects model with covariates for directed graphs. Stat Neerlandica 58:234–254
59. Wasserman S (1980) Analyzing social networks as stochastic processes. J Am Stat Assoc 75:280–294
60. Wippler R (1978) The structural-individualistic approach in Dutch sociology. Neth J Sociol 4:135–155
61. Zeggelink EPH (1994) Dynamics of structure: an individual oriented approach. Soc Netw 16:295–333

Further Reading

For further reading, basic concepts of continuous-time Markov processes can be found in [34]. The basic definition of the model presented here and of the statistical estimation methods for network dynamics based on panel data can be studied in [46,47]. The approach to the co-evolution of networks and behavior is presented in [49,52]. Some examples of the methods presented in this chapter can be found in [6,53,55,56,57].

Networks, Flexibility and Mobility in

Michael F. Thorpe
Arizona State University, Tempe, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Flexibility
Maxwell Counting
Enumeration Methods
Molecular Framework Conjecture
Geometrical Simulation
Future Directions
Bibliography

Glossary

Atypical graph A special graph with symmetry, such as parallel lines – in contrast to a generic graph.

Covalent bond When a covalent bond exists between two atoms in a molecule, the bond length is fixed and independent of the environment. The bond angle is also usually fixed between two covalent bonds that share a common atom.

Generic graph A graph with arbitrary positions of the vertices and associated edges.

Graph A set of vertices connected by edges.

Dihedral angle rotation A rotation about a covalent bond connecting two atoms in a larger molecule.

Geometrical simulation A technique that allows the motion of the flexible parts of a network to be determined while obeying all the constraints, both equalities and inequalities.

Hyperstatic When the number of degrees of freedom is less than the number of constraints plus the number of rigid body motions.

Hypostatic When the number of degrees of freedom is greater than the number of constraints plus the number of rigid body motions.

Isostatic When the number of degrees of freedom is equal to the number of constraints plus the number of rigid body motions.

Macromolecule A large number of atoms connected by covalent bonds.

Maxwell count A global estimate of the number of degrees of freedom that assumes that every edge (constraint) is independent.

Molecular dynamics The solution of the classical equations of motion for a system of particles moving under a given potential.

Network glass Non crystalline network made up of atoms that form covalent bonds with each other.

Nuclear magnetic resonance An experimental technique that can be used to probe the different conformational states of a molecule.

Pebble game A procedure for book keeping between degrees of freedom and constraints that allows a rigid region decomposition to be performed on a graph.

Protein A large macromolecule made up of amino acids linked together to form a polypeptide chain, which folds into a compact structure that has a biological function.

Rigid region decomposition The division of a graph into rigid regions, both with and without redundant edges, and the flexible regions between them.

Redundant edge An edge that can be removed without changing the rigid regions in the graph.

Definition of the Subject

A network can be represented by a graph that can be either rigid or flexible depending on the number and distribution of the constraints. A constraint is a length that is fixed. One example of a network is a collection of points in space connected by edges. If the number of edges is large enough and homogeneous enough in distribution, then the system is a rigid body, and no motion is possible, except for translations and rotations of the whole network. If a sufficient number of edges are removed, then parts of the systems can become flexible, which means that local motion is possible while maintaining all the constraints. Another manifestation of such a system is a collection of bodies in space connected by bars. Once it is determined that there are flexible parts in the graph, the actual motion, or mobility, can be found. Such graphs were first studied by Maxwell [1] in the 1860s using ideas previously developed by Lagrange [2] in the 1780s. Exact enumeration methods were developed later, starting with Laman [3] in 1970 for two-dimensional systems and can be extended in three-dimensional graphs in some cases [4,5]. The mobility of flexible graphs can be studied using geometrical simulation [6].

This general approach, involving flexibility and mobility, has important applications in complex graphs (networks) where it is difficult to solve the classical equations of motion [7] within a potential energy landscape [8], and insight can be gained by approximating the energy landscape by a set of appropriate constraints. These constraints can be either equalities as described above, which are important in determining the flexibility of the network, or equalities augmented by inequalities which determine the subsequent mobility. Two important applications at the atomic and molecular level are network glasses and proteins.

Introduction

As well as developing the famous equations that describe electromagnetism, Maxwell also studied the stability of structures [1,9], although this work is very much less well known. He represented a structure, for example a trestle bridge, by a series of vertices and edges, and then estimated whether the structure was stable or not. In two dimensions, a triangle is rigid but a quadrilateral can be deformed as there is one internal degree of freedom. In general the number of degrees of freedom F is given by

$$F = dN - N_c$$

where there are N points embedded in a d dimensional space with N_c constraints, where a constraint is an edge. Examples are given in Fig. 1

As an introduction, we use popsicle sticks and cotter pins (also called split pins) to construct framework structures. Each popsicle stick has a small hole drilled near each end, through which the cotter pins can be placed. The pin put through two or more sticks, as shown in Fig. 1, in order to construct a two dimensional framework. A pin allows for possible hinge type motion at the joint, even if the two parts of the split pin are opened. A triangle of popsicle sticks is rigid as no internal deformations are possible, and in a similar way the two edge-sharing popsicles in the upper part of Fig. 2 are rigid.

An individual popsicle stick on a two dimensional surface has 3 degrees of freedom. This can be most easily seen as the two degrees of freedom associated with the center of mass plus a rotation about the center of mass. Thus if there are a total of N_{st} popsicle sticks, then there are initially $3N_{st}$ degrees of freedom (replacing dN the in the previous equation) before any cotter pins are added to form joints. When a cotter pin is placed through the holes in two sticks, the holes must be aligned in both the x and y directions, giving two constraints. Each additional

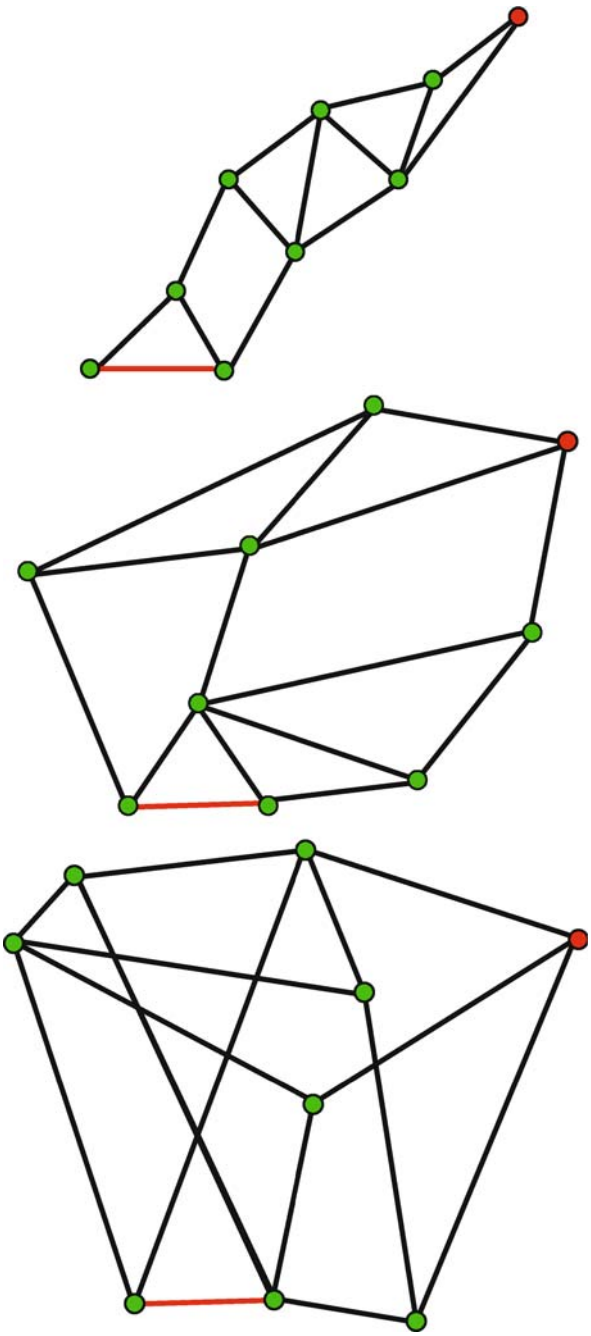

Networks, Flexibility and Mobility in, Figure 1
Showing three examples of two-dimensional isostatic networks consisting of 9 vertices and 15 edges

popsicle stick added onto a cotter pin, means two additional constraints, so that the total number of constraints is $N_c = 2\sum_r n_r(r-1)$, where there are n_r joints with r popsicle sticks coming together. Thus the number of degrees of freedom can be written as

$$F = 3N_{st} - 2\sum_r n_r(r-1)$$

where it is convenient to let the sum include dangling ends where $r = 1$ in what follows. The total number of sticks is $N_{st} = \sum_r n_r r/2$ and the total number of cotter pins is $N_{cp} = \sum_r n_r$ so that the equation above can be written in the convenient form

$$F = 2N_{cp} - N_{st}\,.$$

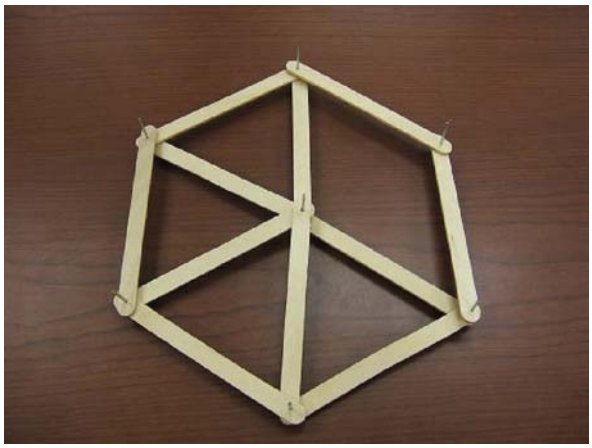

Networks, Flexibility and Mobility in, Figure 2
Showing how popsicle sticks and cotter pins can be used to construct framework structures

From Fig. 2, we have $N_{st} = 11$ and $N_{cp} = 7$, so that from the equation above, the number of floppy modes $F = 3$. These 3 are the macroscopic motions of the whole structure (two translations and one rotation) and there are no internal floppy modes. Note that the surprisingly simple formula above applies even if frameworks with unattached popsicle sticks (dangling ends with $r = 1$) are present as long as a cotter pin is placed in the hole of the dangling end. Although such a pin serves no purpose, it does help simplify the counting and leads to the simple equation above.

The approach above needs correction when redundant popsicle sticks are present – for example if the last spoke is included in Fig. 2, then there is a single redundant popsicle stick present and the count needs modifying. Note that removing any one of the 12 popsicle sticks in such a framework will remove the redundancy. The great challenge of rigidity theory is going beyond the simple counts given here to account for redundancy.

The approach taken here focused on the popsicle sticks as the fundamental objects or bodies, and the cotter pins

as constraints. However it is often convenient to reverse this and treat the cotter pins as point objects and the popsicle sticks as constraints. The meaning of the equation above now becomes clearer. Each cotter pin has 2 degrees of freedom, as it is a point in a two dimensional plane and so no rotations are involved. This is the origin of the term $2N_{cp}$and the term N_{st} represents the number of constraints (N_c). Hence in further analysis in two dimensions in this section, we will use the first equation in this section with $d = 2$ to give

$$F = 2N - N_c \,.$$

An example is given in Fig. 1, where each of the three panels has $d = 2$, $N = 9$ and $N_c = 15$, so that $F = 3$. These three degrees of freedom are the two translations and the one rotation, associated with the rigid body macroscopic motions of the whole system. This can be thought of in terms of surveying, where the conventional technique using triangulation produces a graph that is made up of a series of edge sharing triangles as shown in the upper left panel. The baseline, together with the distant point to be located with respect to the baseline, are shown in red. This produces an isostatic graph as do the other two panels, although these would probably not be used by any practicing surveyor! Thus, there are no internal motions within these three isostatic networks. This kind of single global count is referred to as the Maxwell count, and is accurate as long as all the counted constraints are independent. The problem of course is that it is hard to know whether constraints are independent or not a priori, especially in very large graphs. Thus, the Maxwell count should be thought of as an estimate or global count, which is good if the network is fairly homogeneous as in the three examples in Fig. 1.

Adding additional edges to the graphs in Fig. 1, leads to redundancy, which is of course very desirable in a building, where one does not want to have the building collapse if a single beam (edge) fails. It also leads to stability against mutations and thermal motions in proteins [10]. If the number of edges associated with each vertex varies a lot throughout the system, then the Maxwell count is incorrect and may or may not be useful as an estimate. In the examples in Fig. 1, the Maxwell count is exact, and there are just enough connections to hold the system rigid, making all three graphs isostatic. If an edge is removed in any of three graphs, there will be a single internal degree of freedom or floppy mode and the graph is called hypostatic. If an additional edge is added then there is a single redundant edge and the graph is called hyperstatic.

The study of such graphs to determine flexibility does not involve any motion, and so belongs in the realm of statics. That is the knowledge that a region is flexible, determines the potential for motion, but does not determine its amplitude. Here we are dealing with virtual displacements. To study mobility, it is necessary to make actual displacements and study the motion of the graph. Here an additional set of inequality constraints may be introduced, such that each vertex becomes a hard disc (in two dimensions) or a hard sphere in (three dimensions) where such objects are not allowed to interpenetrate. Such inequality constraints are irrelevant for the static analysis of a graph, but greatly cut down on the mobility of the flexible regions when motion is allowed.

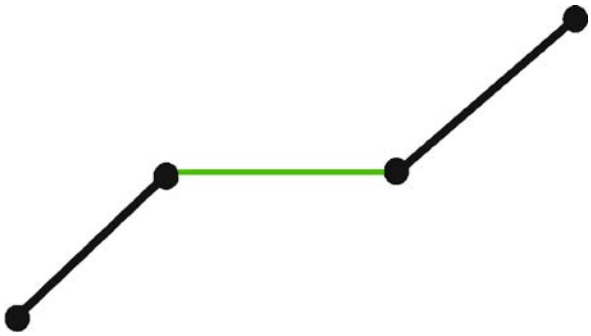

Networks, Flexibility and Mobility in, Figure 3
Showing a molecular hinge in three dimensions where dihedral angle rotations are allowed using the *green bond* as an axis

Macromolecules and disordered solids can be represented by graphs, where the atoms are the vertices and the distance between nearest neighbor atoms are the edges. This is particularly appropriate for covalent bonds [11], where the bond length is fixed to within a few percent and the bond angle between adjacent covalent bonds that share a common atom is also fixed similarly and can be represented by an edge connecting second neighbors. The system is still potentially flexible in three dimensions as dihedral angle rotations are allowed, involving a twist along a linear section with four atoms, if no other constraints are present to forbid such motion. This is a three-dimensional example of a hinge and is shown in Fig. 3.

Flexibility

A graph is defined by its vertices and edges, which can be used to represent a physical system like a glass or a protein, where the vertices are atoms. Laman's theorem [3] in two dimensions can be thought of as a Maxwell count but applied on all length scales to each subgraph. This is a topological theorem. The Maxwell count is also topological, in that the particular value of the lengths of the edges are irrelevant. Such graphs are called generic and have no special angles or lengths associated with them, and represent

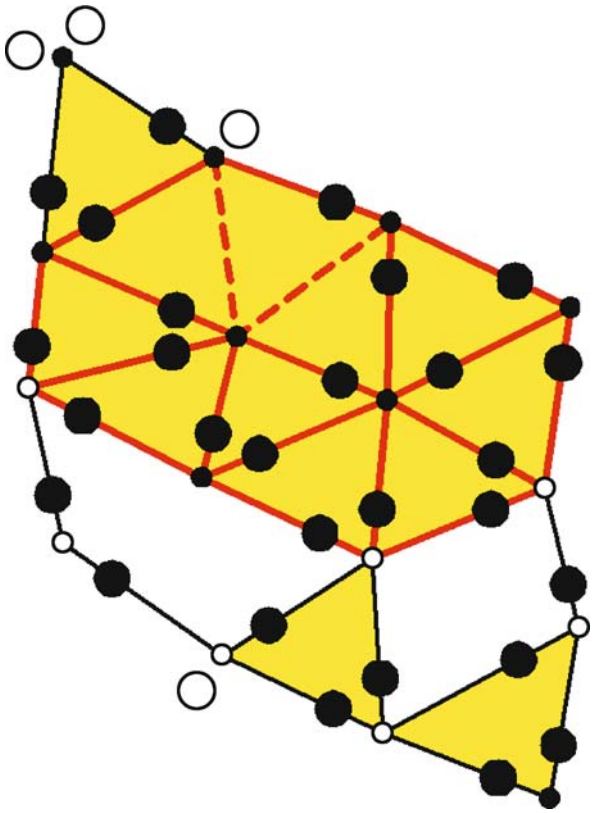

Networks, Flexibility and Mobility in, Figure 4
Showing the result of a pebble game in two dimensions, where the four free pebbles (*unshaded*) designate the three macroscopic rigid body modes (two translations and one rotation) and the single internal floppy mode, which can be visualized as a rocking of the flower part of the graph

the whole class of such graphs. Graphs that do have symmetries like parallel lines, are called atypical and there is no general theory of the flexibility of such graphs although much is known as a result of work by Dove [12,13,14,15] and Guest [16,17]. This is a rare case where the existence of symmetry is not a simplification, but rather a serious complication that can introduce additional degrees of freedom. Alternative methods using group theory appropriate for the symmetry can be used in crystalline lattices [18] and icosahedral viral capsids [16].

Laman's theorem [3] has been put into an algorithmic form by Hendrickson [19] and implemented by Jacobs and Thorpe [20,21,22] in an algorithm called the pebble game. In this algorithm, two free pebbles are associated with each vertex in the graph and can be moved onto independent edges in order to maintain the proper book keeping between degrees of freedom and constraints. An example of the output of a pebble game search for a small graph is shown in Fig. 4.

Maxwell Counting

We return now to the global count as first introduced by Maxwell. This is a very simple and useful way to estimate if a network is rigid. Let us consider a regular lattice like a triangular net in two dimensions and a face centered cubic lattice in three dimensions. These must be slightly distorted to make them generic. If we have N sites, each with z neighbors present independently with probability p, then there are dN degrees of freedom and $Nzp/2$ constraints (assumed independent) associated with the bonds between nearest neighbors. Thus, the Maxwell estimate of the number of degrees of freedom is given by

$$F = dN - \frac{Nzp}{2} = dN\left(1 - \frac{\langle r\rangle}{2d}\right)$$

where we define a mean coordination $\langle r\rangle = zp$. This means that the number of floppy modes goes to zero at

$$\langle r\rangle = zp_c = 2d$$

which gives an estimate for the critical fraction of bonds present at rigidity percolation at $p_c = 2d/z$. Of course the Maxwell count fails as the phase transition is approached, and fails catastrophically above the phase transition, where it yields a negative number of floppy modes! This estimate gives $p_c = 2/3$ for the triangular network in two dimensions and $p_c = 1/2$ for the face centered cubic lattice in three dimensions. These estimates are very close to the best current numerical estimates determined using the pebble game of 0.6602 [21] and 0.4967 [5], respectively. This shows the remarkable accuracy of the Maxwell estimates for rigidity percolation. It is sometimes said that the Maxwell count is a mean field theory, but we prefer to use the term global count [21]. Nevertheless, there are hidden long range effects in rigidity which is reminiscent of mean field theory [23]. Note that the phase transition is second order in two dimensions [21] but can be first order in three dimensions [5], but nevertheless the Maxwell estimates p_c are excellent in both cases. A typical sample of a triangular net very near percolation is shown in Fig. 5.

Enumeration Methods

Although the Maxwell count gives a good global picture in many cases, it gives no information about site to site variations. This was accomplished by Laman [3] in 1970 who showed how the Maxwell count should be applied to each and every subgraph of the network to get a complete microscopic solution as shown for example in Fig. 5. Laman's theorem was turned into an algorithm by Hendrikson [19] that has been implemented on the computer by Jacobs and

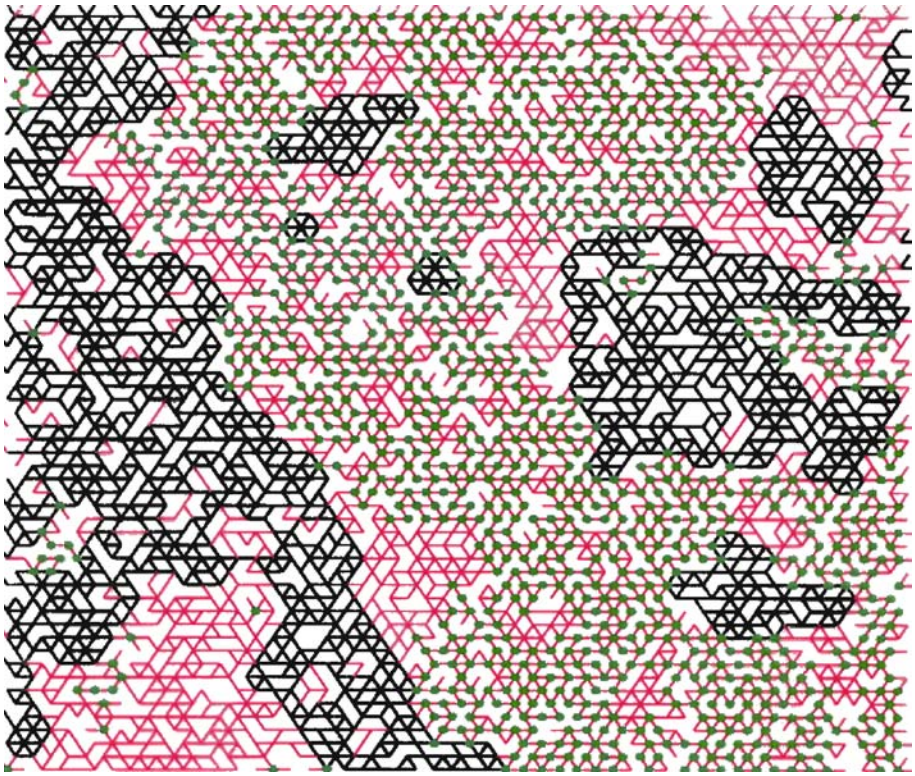

Networks, Flexibility and Mobility in, Figure 5
Showing a piece of a triangular network very near percolation, caused by bond dilution. Hinges are shown in *green*, isostatic regions in *red* and hyperstatic regions in *black*

Thorpe [20,21]. This leads to the rigid region decomposition of the kind shown in Fig. 5. This ability to examine very large networks with 100,000 atoms and more and to obtain a rather complete description of the second-order phase transition for the triangular net and the associated geometrical critical exponents [21,24] is extremely useful. Response functions, like elastic constants can also be calculated, but much less accurately as more traditional approaches like conjugate gradient methods must be used on smaller lattices [25]. The elastic constants go to zero as the phase transition is approached from the rigid side for bond dilution on both the triangular lattice and the face-centered cubic lattice [25] which is surprising as the phase transition for the face-centered cubic lattice is first order. This result requires further explanation.

Molecular Framework Conjecture

Rigidity in three dimensions is much more complicated than in two dimensions because of the existence of "banana diagrams", an example of which is shown in Fig. 6. The three rigid yellow pieces are the bananas and it

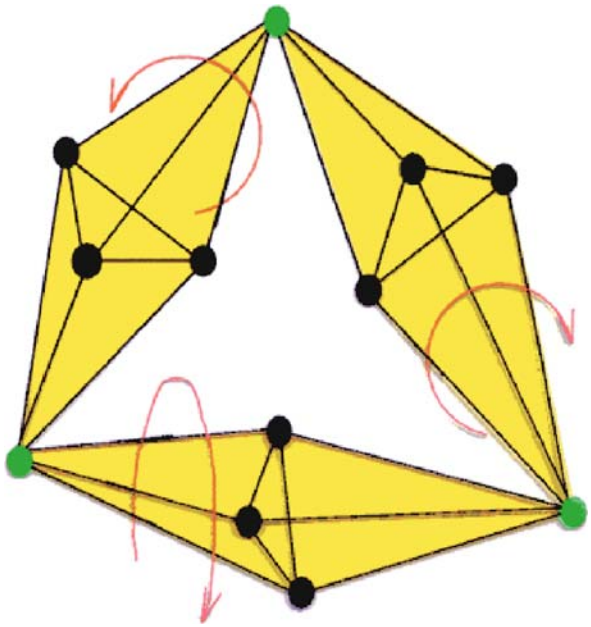

Networks, Flexibility and Mobility in, Figure 6
Showing the problematic banana diagram in three dimensions

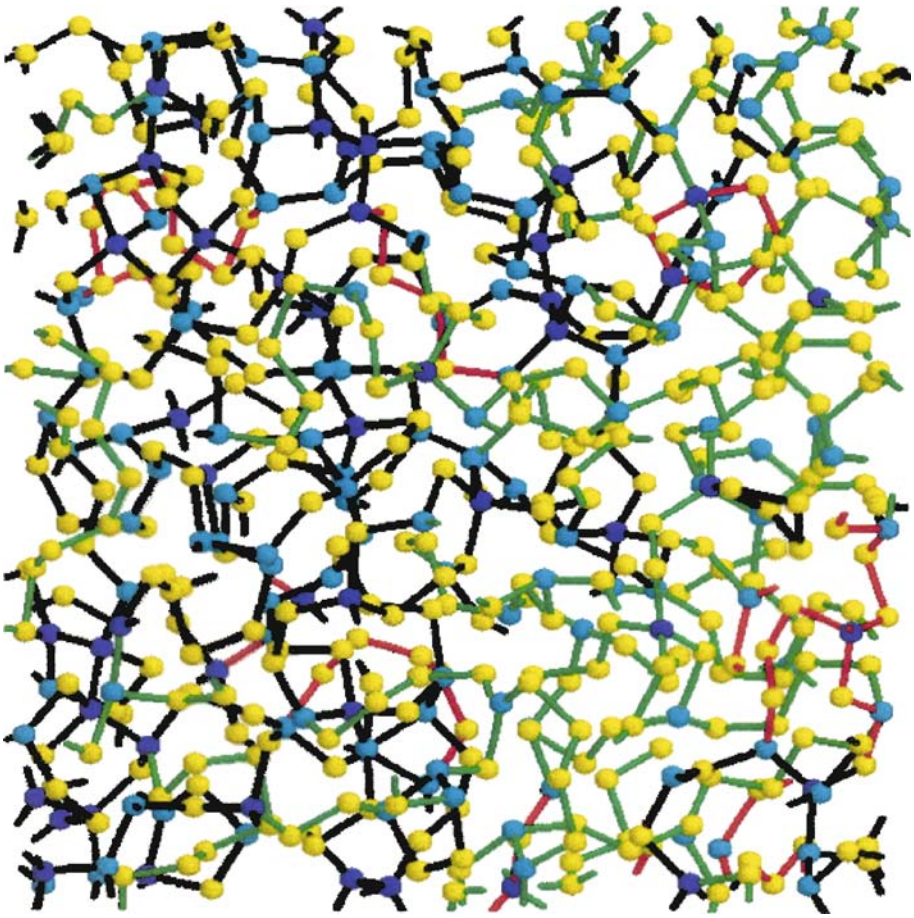

Networks, Flexibility and Mobility in, Figure 7
Showing a piece of a covalent glass near the phase transition at a mean coordination $\langle r \rangle = 2.4$, where the *green bonds* are hinges, the *red bonds* isostatic and the *black bonds* hyperstatic

can be seen that these can rotate as indicated by the arrows. The complication is that the three green sites themselves form a non-contiguous rigid cluster. This cannot happen in two dimensions (see Fig. 4 and 5), where all the individual rigid clusters are contiguous and represent a major complication. It means that there is no Laman-type theorem and hence no pebble game. Nevertheless, a very accurate, but approximate, pebble game has been developed for general three-dimensional networks [5] that for example is estimated to be accurate to 1 part in 10^9 for bond diluted face-centered cubic lattices.

There is one very important special case in three dimensions, which is the bond-bending network, where there are angular forces associated every each pair of central forces that share a common vertex. These angular forces can be thought of as second neighbor central forces. For these special networks there is a conjecture; called the *Molecular Framework Conjecture* that the appropriate pebble game is exact [4,26,27]. This is a very fortunate situation as this is the case of most practical interest. Examples of such molecular frameworks are covalent glasses like GeAsSe as shown in Fig. 7, where the coloring of the sites reflects the chemistry (dark blue for 4-coordinated Ge, light blue for 3-coordinated As and yellow for 20-coordinated Se), and the bond coloring reflects whether they are flexible (green), rigid (red) or stressed and rigid (black).

Maxwell counting can be used to estimate when these networks undergo a transition from rigid to flexible. If there are n_r sites with r-coordinated atoms, where r is 2, 3 or 4, then the total number of atoms $N = \sum_r n_r$ and the mean coordination $\langle r \rangle = \sum_r r n_r / N$. The total number of degrees of freedom is $3N$. The number of angular constraints is $r/2$ for central forces and $3r - 5$ for angular

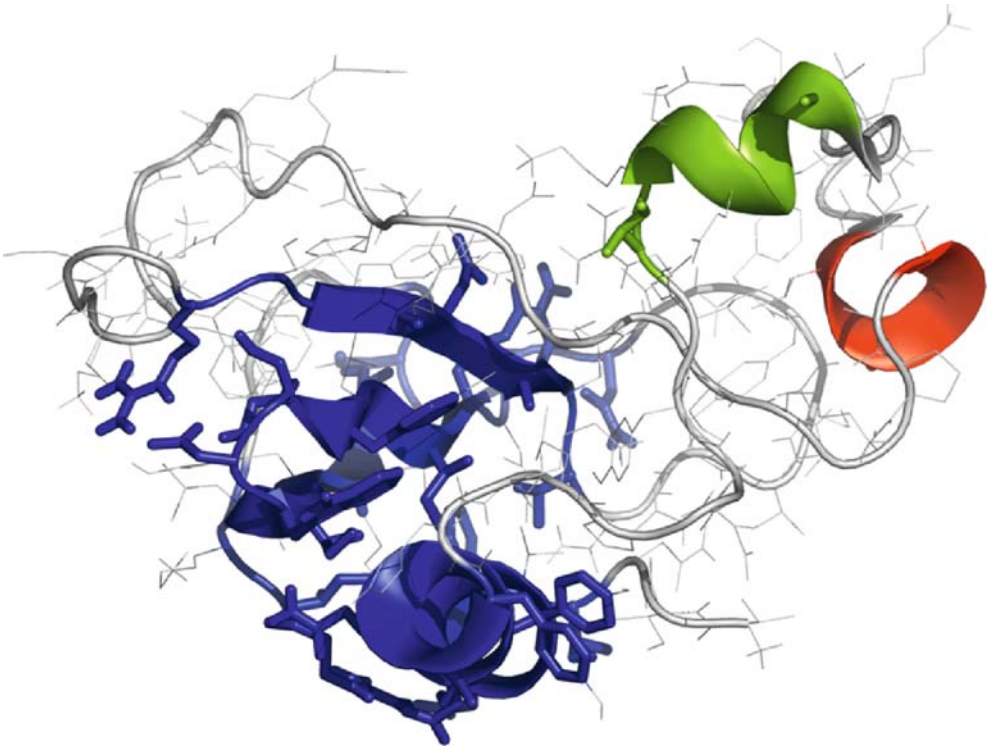

Networks, Flexibility and Mobility in, Figure 8
Showing a piece of the protein barnase with the three largest rigid regions picked out in the three different *solid colors*

forces, both associated with an r-coordinated site. Therefore, the number of floppy modes is

$$F = 3N - \sum_r n_r \left[\frac{r}{2} + (2r - 3) \right] = 6N \left[1 - 5 \frac{\langle r \rangle}{12} \right]$$

which gives a phase transition from rigid to flexible when $\langle r \rangle = 12/5 = 2.4$ [28,29,30]. Note that reference [28] which first applied these counting ideas to network glasses, contains an error as six angular constraints rather than the correct five independent angular constraints were assigned to a 4-coordinated site. Exact enumerations, using the pebble game on computer generated networks give $\langle r \rangle = 2.385$ for the location of the phase transition [31]. Thus, when the polymer chains made up of 2-coordinated atoms are lightly cross-linked, the network is flexible, but as the number of cross-links becomes quite dense and the mean coordination rises to around 2.4, the network becomes rigid [29].

The molecular framework conjecture also applies to proteins, which are large macromolecules containing many hundreds of atoms. Proteins are polypeptide chains that are crossed linked by hydrophobic tethers and hydrogen bonds to form rather compact three-dimensional structures [32]. A rigid region decomposition of a typical protein, barnase is shown in Fig. 8.

Proteins have enough rigidity to define their three-dimensional structure, while retaining sufficient flexibility to function. Thus, they exist in a narrow window around the phase transition between flexible and rigid [10]. This is shown in Fig. 9, which also shows that covalent glasses behave in a very similar way.

Geometrical Simulation

The approach described here comes from an area of mathematics called rigidity theory, which yields powerful results without much calculational or computational overhead. However, these results are limited as no actual motion is involved, and it is only the possibility of motion, or absence of motion, that is addressed. This is very valuable information.

However, for a more complete description, it is desirable to know the mobility, or the amplitudes of the motion, which do not violate any of the original constraints. In addition, the motion may involve additional constraints, the most important being inequalities that prevent objects from interpenetrating. For example, two-dimensions

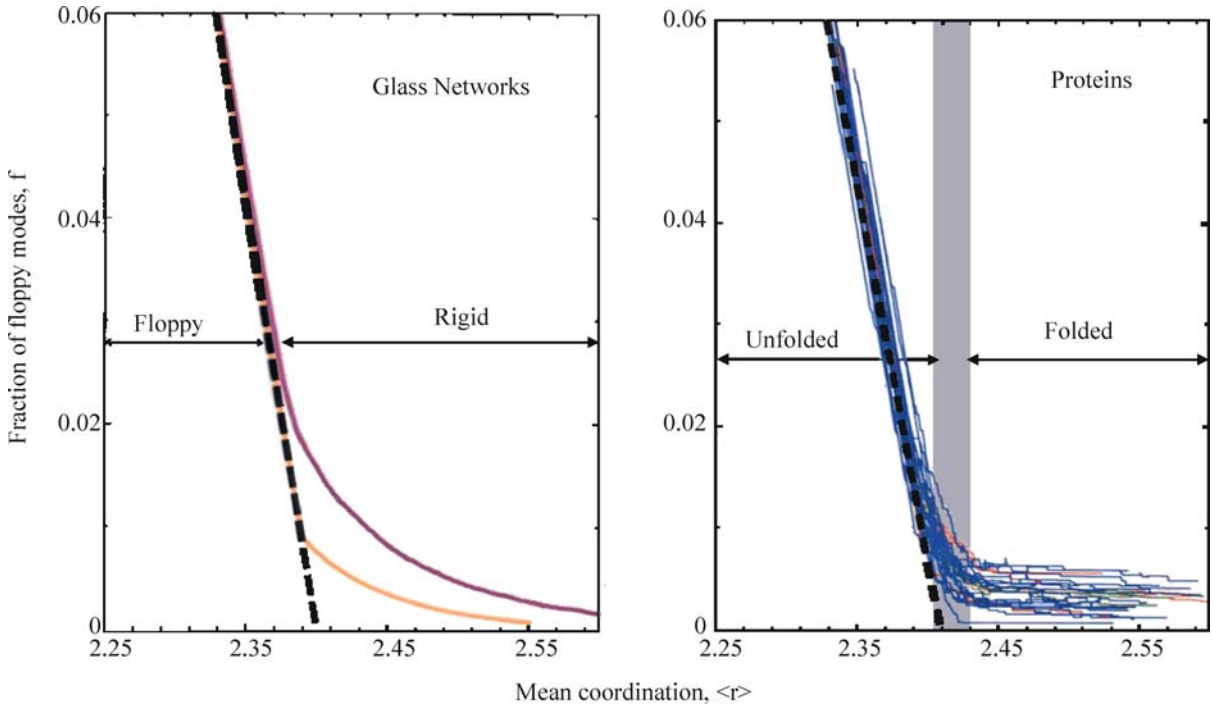

Networks, Flexibility and Mobility in, Figure 9
Showing the similar behavior of the number of floppy modes in covalent network glasses and a selection of proteins [10]. In all cases the transition from flexible to rigid occurs around a mean coordination $\langle r \rangle = 2.4$. Here $f = F/3N$ is the number of floppy modes per degree of freedom

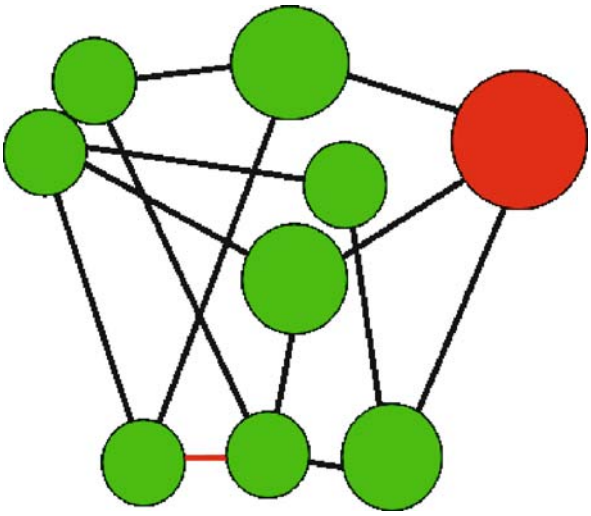

Networks, Flexibility and Mobility in, Figure 10
Showing the two-dimensional graph from the lower panel in Fig. 1, with a single edge removed, so that the graph becomes hypostatic with a single floppy mode

discs are not allowed to overlap, and in three dimensions, spheres are prevented from interpenetrating. An example is shown in Fig. 10, where the vertices from the lower panel in Fig. 1 have been replaced by discs of various radii. Pairs of discs are not allowed to overlap during the motion. As the disc radii are increased, the amplitude of the floppy mode is liable to be decreased, by the necessity to avoid collisions. Geometric simulation methods can be used to study such motion [33,34,35] and these are rather efficient. Key ingredients are the introduction of random or directed small displacements, the maintenance of constraints and collision avoidance. The reader is referred to the articles above for more details.

Geometrical simulation techniques are similar in three dimensions, and for example new conformations of molecules can be generated. An example is shown in Fig. 11 for the protein barnase, previously shown in Fig. 8. A set of 20 conformers obtained from a nuclear magnetic resonance experiment [36] are shown in the left panel. These are determined from experiments where a sufficient number of constraints are determined that the remaining indeterminacy is associated with the protein mobility. In the right panel, a single X-ray structure [37] is used as a starting point to generate many new protein conformations consistent with the constraints. The 20 simulation conformers shown are selected from a much larger set and are spaced well apart [6]. The blue regions in the left panel are rigid, and the other regions are flexible, leading to mobility. Note that distinct

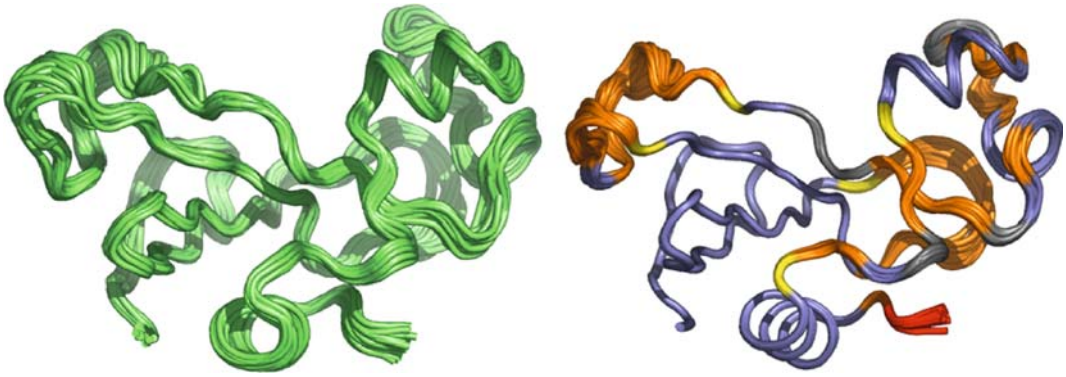

Networks, Flexibility and Mobility in, Figure 11
Showing a comparison between the NMR ensemble of conformers (*left*) for barnase and the set of conformers (*right*) generated using geometrical simulation techniques from a single X-ray crystallographic structure [6]

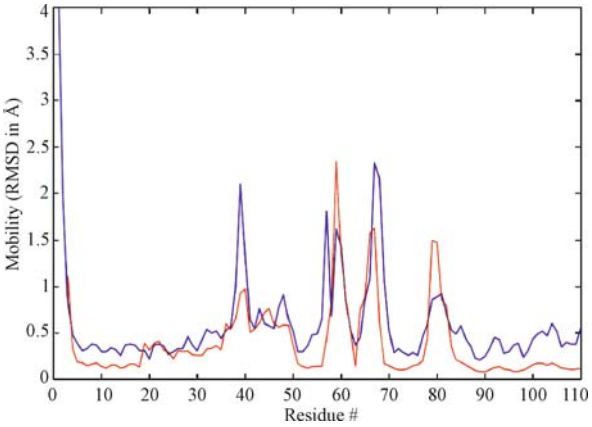

Networks, Flexibility and Mobility in, Figure 12
Plotting the mobility of each residue, as determined using experimental data (*blue*) from Fig. 11, and showing that geometric simulation (*red*) captures the main features of the mobility of the protein, when compared with nuclear magnetic resonance data [6]

rigid regions can move with respect to other rigid regions.

A quantitative comparison between the Root Mean Square Distance (RMSD) and the residue number between the nuclear magnetic resonance experiment and the geometric simulations is shown in Fig. 12. The amplitudes obtained from both the experiment and from geometrical simulation are absolute, so no scaling is required. The residue number refers to the numbering of the amino acids along the backbone of the polypeptide chain that folds to produce the three-dimensional protein structure.

Future Directions

In this brief article, we have set down the key ingredients that are needed when considering the flexibility and mobility of networks, with an emphasis on the important class of three-dimensional molecular networks. For more details of the more mathematical aspects of the theory of rigidity, the reader is referred first to the books referenced at the end. There are few review articles in this area as these approaches are quite new and are still being developed.

The static part of rigidity has led to an exact theorem [3] and fast algorithms in two dimensions [20,21,22]. There is no theorem in three dimensions for the general case, although there is the special case of the very useful molecular framework conjecture [4] that remains to be proved. Algorithms have been developed recently based on the molecular framework conjecture [31,32,38,39,40]. Although there is no theorem or conjecture for the general case in three dimensions, approximate, but rather accurate, algorithms nevertheless have been developed [5]. There is a need for a more complete description of the unique features of rigidity in three dimensions, associated with banana diagrams etc. At this time there seems no great motivation for studying rigidity in dimensions higher than three.

Geometrical simulation is still in its infancy – this is a more complex problem than the movement of robot arms [41] because many interlocking rings of constraints are involved and the environment is crowded in general. This is rather like moving groups of friends holding hands tightly in rings through a football crowd at the end of a game. Early efforts focused on ring closure [33,42], but these proved inefficient and have been largely abandoned in favor of small Monte Carlo-type moves that initially violate the constraints, followed by various guiding and shaking procedures to restore the constraints and arrive at a new conformation [6]. The inequalities associated with the van der Waals or hard sphere overlaps cause particular concern in the crowded molecular environments

found in proteins and better ways of handling this are needed. Ultimately the results of such approaches, obtained rapidly, can be used as input with phenomenological classical potentials [43] of the kind used in molecular dynamics [44,45], to be able to generate larger amplitude motions for a given amount of computer time.

Bibliography

Primary Literature

1. Maxwell JC (1864) On the calculation of the equilibrium and stiffness of frames. Philos Mag 27:294–299
2. Lagrange J-L (1788) Mécanique Analytique, 4th edn. Gauthier–Villars, Paris
3. Laman G (1970) On graphs and rigidity of plane skeletal structures. J Eng Math 4:331–340
4. Whiteley W (2005) Counting out to the flexibility of molecules. Phys Biol 2:S116–26
5. Chubynsky MV, Thorpe MF (2007) Algorithms for 3d rigidity analysis and a first order phase transition. Phys Rev E 76:041135
6. Wells SA, Menor S, Hespenheide BM, Thorpe M (2005) Constrained geometric simulation of diffusive motion in proteins. Phys Biol 2:127–136
7. Goldstein H, Poole CP, Safko JL (2002) Classical mechanics, 3rd edn. Addison Wesley, San Francisco, p 638
8. Wales D (2003) Energy landcapes. Cambridge University Press, Cambridge
9. Maxwell JC (1864) On reciprocal figures and diagrams of forces. Phil Mag 27:250–261
10. Rader AJ, Hespenheide BM, Kuhn LA, Thorpe MF (2002) Protein unfolding: Rigidity lost. Proc Natl Acad Sci USA 99:3540–5
11. Pauling L (1928) The shared-electron chemical bond. Proc Natl Acad Sci USA 14:359–62
12. Dove MT, Heine V, Hammonds KD (1995) Rigid unit modes in framework silicates. Mineral Mag 59:629–639
13. Dove MT, Hammonds KD, Heine V, Withers RL, Xiao Y, Kirkpatrick RJ (1996) Rigid unit modes in the high-temperature phase of sio2 tridymite: Calculations and electron diffraction. Phys Chem Miner 23:56–62
14. Dove MT, Keen DA, Hannon AC, Swainson IP (1997) Direct measurement of the si-o bond length and orientational disorder in the high-temperature phase of cristobalite. Phys Chem Miner 24:311–317
15. Dove MT, Trachenko KO, Tucker MG, Keen DA (2000) Rigid unit modes in framework structures: Theory, experiment and applications. Transform Process Miner 39:1–33
16. Kangwai RD, Guest S (1999) Detection of finite mechanisms in symmetric structures. Int J Solid Struct 36:5507–5527
17. Fowler PD, Guest S (2002) Symmetry and states of self-stress in toroidal frames. Int J Solid Struct 39:4385–4393
18. Ashcroft NW, Mermin ND (1976) Solid state physics. Holt Rinehart and Winston, New York, p 826
19. Hendrickson B (1992) Conditions for unique graph realizations. Siam J Comput 21:65–84
20. Jacobs DJ, Thorpe MF (1995) Generic rigidity percolation: The pebble game. Phys Rev Lett 75:4051–4054
21. Jacobs DJ, Thorpe MF (1996) Generic rigidity percolation in two dimensions. Phys Rev E 53:3682–3693
22. Jacobs D, Hendrickson B (1997) An algorithm for two-dimensional rigidity percolation: The pebble game. J Comp Phys 137:346–365
23. Kittel C, Shore H (1965) Development of a phase transition for a rigorously solvable many-body system. Phys Rev 138:A1165–A1169
24. Arbabi S, Sahimi M (1993) Mechanics of disordered solids. I Percolation on elastic networks with central forces. Phys Rev B Condens Matter 47:695–702
25. Feng S, Thorpe MF, Garboczi E (1985) Effective-medium theory of percolation on central-force elastic networks. Phys Rev B 31:276–280
26. Whiteley W (1999) Rigidity of molecular structures: Generic and geometric analysis. In: Thorpe MF, Duxbury PM (eds) Rigidity Theory and Applications. Kluwer Academic/Plenum Publishers, New York
27. Tay T-S, Whiteley W (2005) Comparison of molecular models. unpublished
28. Phillips J (1979) Topology of covalent non-crystalline solids. 1 Short-range order in chalcogenide alloys. J Non-Cryst Solid 34:153–181
29. Thorpe MF (1983) Continuous deformations in random networks. J Non-Cryst Solid 57:355–370
30. He H, Thorpe MF (1985) Elastic properties of glasses. Phys Rev Lett 54:2107–2110
31. Thorpe MF, Jacobs DJ, Chubynsky NV, Rader AJ (1999) Generic rigidity of network glasses. In: Thorpe MF, Duxbury PM (eds) Rigidity theory and applications. Kluwer Academic/Plenum Publishers, New York, pp 239–277
32. Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF (2001) Protein flexibility predictions using graph theory. Proteins 44:150–65
33. Lei M, Zavodszky MI, Kuhn LA, Thorpe MF (2004) Sampling protein conformations and pathways. J Comput Chem 25:1133–48
34. Lee A, Streinu I, Brock O (2005) A methodology for efficiently sampling the conformation space of molecular structures. Phys Biol 2:S108–S115
35. Wells S, Menor S, Hespenheide B, Thorpe MF (2005) Constrained geometric simulation of diffusive motion in proteins. Phys Biol 2:S127–36
36. Bycroft M, Ludvigsen S, Fersht AR, Poulsen FM (1991) Determination of the three-dimensional solution structure of barnase using nuclear magnetic resonance spectroscopy. Biochemistry 30:8697–701
37. Hartley RW (1989) Barnase and barstar: Two small proteins to fold and fit together. Trends Biochem Sci 14:450–4
38. Jacobs D (1998) Generic rigidity in three-dimensional bond-bending networks. J Phys A Math Gen 31:6653–6668
39. Jacobs D, Kuhn LA, Thorpe MF (1999) Flexible and rigid regions in proteins. In: Thorpe MF, Duxbury PM (eds) Rigidity Theory and Applications. Kluwer Academic/Plenum Publishers, Traverse City
40. Hespenheide BM, Jacobs DJ, Thorpe MF (2004) Structural rigidity in the capsid assembly of cowpea chlorotic mottle virus. J Phys Condens Matter 16:S5055–S5064
41. Streinu I (2000) A combinatorial approach to planar non-colliding robot arm motion planning. In: Proc 41st Annual ACM/IEEE Symposium on Foundations of Computer Science (FOCS 2000). IEEE Computer Society Washington, DC, pp 443
42. Kuhn LA, Zavodszky MI, Arora S, Lei M, Thorpe MF (2004) Modeling correlated protein main-chain and side-chain motions

in ligand docking and screening. Abstr Pap Amer Chem Soc 228:U501
43. McCammon JA, Gelin BR, Karplus M (1977) Dynamics of folded proteins. Nature 267:585–90
44. Brooks RR, Bruccoleri BE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983) Charmm: A program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 4:187–217
45. Pearlman DA, Case DA, Caldwell JW, Ross WS, Cheatham TE, Debolt S, Ferguson D, Seibel G, Kollman P (1995) Amber, a package of computer-programs for applying molecular mechanics, normal-mode analysis, molecular-dynamics and free-energy calculations to simulate the structural and energetic properties of molecules. Comput Phys Commun 91:1–41

Books and Reviews

Boolchand P (2000) Insulating and Semiconducting Glasses. Series on Directions in Condens Matter Phys 17. World Scientific, Singapore
Graver JE, Servatius B, Servatius H (1993) Combinatorial Rigidity. In: Graduate Series in Mathematics. Amer Math Soc
Thorpe MF, Duxbury PM (1999) Rigidity Theory and Applications. Kluwer Academic, Plenum Press, New York
Thorpe MF, Tichy L (2001) Properties and Applications of Amorphous Materials. In: Thorpe MF, Tichy L (eds) NATO Science Series, II Mathematics, Physics and Chemistry, vol 9

Networks and Stability

Frank H. Page Jr.[1,2], Myrna Wooders[3,4]
[1] Department of Economics, Indiana University, Bloomington, USA
[2] Centre d'Economie de la Sorbonne, Universite Paris 1, Pantheon–Sorbonne, France
[3] Department of Economics, Vanderbilt University, Nashville, USA
[4] Department of Economics, University of Warwick, Coventry, UK

Article Outline

Glossary
Definition of the Subject
Introduction
The Primitives
Abstract Games of Network Formation and Stability
Strong Stability, Pairwise Stability, Nash Stability, and Farsighted Consistency
Singleton Basins of Attraction
Future Directions
Acknowledgments
Bibliography

Glossary

Homogeneous networks A homogeneous network consists of a finite set of nodes together with a finite set of mathematical objects called links or arcs, each identifying a connection between a pair of nodes. Given finite node set N with typical element i, a homogeneous linking network G is a finite collection of sets of the form $\{i, i'\}$ called links. Link $\{i, i'\} \in G$ indicates that nodes i and i' are connected in network G. A homogeneous directed network G is a finite collection of ordered pairs (i, i') called arcs. Arc $(i, i') \in G$ indicates that nodes i and i' are connected in network G via a connection running from i to i'. In a homogeneous network (whether it be a linking network or a directed network) all connections are of the same type.

Heterogeneous networks A heterogeneous network consists of a finite set of nodes together with a finite set of mathematical objects called labeled links or labeled arcs, each identifying a particular type of connection between a pair of nodes. Given finite node set N with typical element i and given finite label set A with typical element a, a heterogeneous linking network G is a finite collection of ordered pairs of the form $(a, \{i, i'\})$ called labeled links. Labeled link $(a, \{i, i'\}) \in G$ indicates that nodes i and i' are connected in network G via a type a link. A heterogeneous directed network G is a finite collection of ordered pairs of the form $(a, (i, i'))$ called labeled arcs. Labeled arc $(a, (i, i')) \in G$ indicates that nodes i and i' are connected in network G via a type a arc running from i to i'. In a heterogeneous network (whether it be a linking network or a directed network) connections can differ and are distinguished by type.

Abstract game of network formation with respect to irreflexive dominance An abstract game of network formation with respect to irreflexive dominance consists of a feasible set of networks $\mathbb{G}$ equipped with a irreflexive dominance relation $>$. A dominance relation on $\mathbb{G}$ is a binary relation on $\mathbb{G}$ such that for all G and G' in $\mathbb{G}$, $G' > G$ (read G' dominates G) is either true or false. The dominance relation is irreflexive if $G > G$ is always false.

Abstract game of network formation with respect to path dominance An abstract game of network formation with respect to path dominance consists of a feasible set of networks $\mathbb{G}$ equipped with a path dominance relation $\geq_p$ induced by an irreflexive dominance relation $>$ on $\mathbb{G}$. Given networks G and G' in $\mathbb{G}$, $G' \geq_p G$ (read G' path dominates G) if either $G' = G$ or there

is a finite sequence of networks in $\mathbb{G}$ beginning with G and ending with G' such that each network along the sequence dominates its predecessor.

Definition of the Subject

Our subject is networks, and in particular, stable networks and the game theoretic underpinnings of stable networks.

Networks are pervasive. We routinely communicate over the internet, advance our careers by networking, travel to conferences over the transportation network and pay for the trip using the banking network. Doing this utilizes networks in our brain. The list could go on. While network models have had a long history in sociology, the natural sciences, and engineering (e. g., in modeling social organizations, brain architecture, and electrical circuits), the rise of the network paradigm in economics is relatively recent. Economists are now beginning to think of political and economic interactions as network phenomena and to model everything from terrorist activities to asset market micro structures as *games of network formation*. This trend in economics, which began with the seminal paper by Myerson [88] on graphs and cooperation and accelerated with the publication of the papers by Jackson and Wolinsky [64] and Dutta and Mutuswami [39] on stable and efficient networks, is likely to continue with the development of new algorithms, the expansion of computational capacity and the broad application of network theories to economic, political, and social phenomena.

What economists bring to the study of networks that is new is game theory. For the most part sociologists, natural scientists and engineers have used networks descriptively and have focused on the design of networks from the perspective of a single designer or on the random evolution of networks from the perspective of nature. This singularity of perspective is a consequence of the nonstrategic nature of the phenomena being explained or the problem being solved (e. g., the spread of a disease through a given population, the transmission of electrical impulses in the brain, or the optimal design of an integrated circuit). In economics the perspective is often times strategic. In particular, in many economic situations, several individuals, guided by their own self interest, behave strategically in putting into place pieces of the network of economic, political, or social interactions under their control and in so doing generate payoffs and externalities that determine the network of economic interactions that eventually emerges in equilibrium. Thus in economics, pieces of the network are the strategies and the network that ultimately prevails is the result of strategic competition rather than the design of a single individual or nature. Because large computer networks such as the internet are built, operated, and used by a many diverse individuals with competing interests, computer scientists are also beginning to use game theoretic models to analyze and understand the optimal design of secure computer networks (see for example, [105,115]). Conversely, what networks bring to the study of economics is a new way of modeling the structure of economic interactions and externalities that makes possible a game-theoretic analysis of how theses structures influence individual payoffs and the economic equilibrium that emerges from competition.

Introduction

Our main objective is to present a unified, game-theoretic development of the main concepts of stability that have appeared in the recent economics literature on strategic network formation, specifically, the notions of strong stability (Jackson and van den Nouweland [61]), pairwise stability (Jackson and Wolinsky [64]), Nash stability, and farsighted consistency (Chwe [26]). In order to accomplish this we follow the approach introduced in Page and Wooders [91,93]. The key ingredient in this approach is an abstract game model of endogenous network formation (i. e., abstract game in the sense of von Neumann–Morgenstern [122]). The model is built on four primitives: A feasible set of networks, player preferences over networks, the rules of network formation, and an irreflexive dominance relation over networks. In the remainder of this introduction we provide an overview of the four primitives of our model, a summary of results discussed, and note some important areas of research that are beyond the scope of this entry.

Feasible Sets: The feasible set may consist of networks as simple as homogenous linking networks or as complex as heterogeneous directed networks. All networks consist of a finite set of nodes (representing, for example, economic agents or players) together with a finite set of mathematical objects called links, labeled links, arcs, or labeled arcs describing the connections between nodes. Here we will focus on homogeneous linking networks, as does most of literature (see, for example, Myerson [88], Jackson and Wolinsky [64], and Jackson and van den Nouweland [61]), except in our discussion of Nash stability where we will consider homogeneous directed networks (as in [5]). What distinguishes homogeneous networks (linking or directed) from heterogeneous networks (linking or directed) is that in a homogeneous network all connections between nodes are of the same type whether represented by a link as in a linking network or by an arc as in a directed network. Thus, in a homogeneous link-

ing network all links are of the same type and in a homogeneous directed network all arcs are of the same type. While homogeneous networks are quite restrictive, they have been very important in developing our understanding of social and economic networks and have proved very useful in many economic applications (see, for example, Belleflamme and Bloch [7], Bramoulle and Kranton [17], Calvo-Armengol [19], and Furusawa and Konishi [41]). Page and Wooders [89,91,93,94] extend the existing literature on economic and social networks by introducing the notion of heterogeneous directed networks. These types of networks potentially have a rich set applications (in the natural sciences, engineering, sociology, politics, as well as economics) because connections or interactions between nodes can be distinguished by direction or intent as well as by type, intensity, or purpose.

Players' Preferences: We will assume throughout that each player's preferences are given by an irreflexive binary relation defined on the feasible set of networks. Thus, we will assume that players have strong (or strict) preferences over networks Under strong preferences, if a player prefers one network to another, then the player's preference is strict. However, we will comment where appropriate on weak preferences. Under weak preferences, if a player prefers one network to another, then the player's preference is either strict or indifferent.

Rules of Network Formation: We will focus here on three different sets of rules: Jackson–Wolinsky rules [64], Jackson–van den Nouweland rules [61], and Bala–Goyal rules [5]. In particular, in our discussions of pairwise stable homogeneous linking networks we will assume that the rules of network formation are the Jackson–Wolinsky rules. Under the Jackson–Wolinsky rules the addition of a link is bilateral (i. e., the two players that would be involved in the link must agree to adding the link), the subtraction of a link is unilateral (i. e., at least one player involved in the link must agree to subtract or delete the link), and network changes take place one link at a time (i. e., only one link can be added or subtracted at a time). In our discussion of strongly stable homogeneous linking networks, we will assume that the rules of network formation are the Jackson–van den Nouweland rules. Under the Jackson–van den Nouweland rules link addition is bilateral, link subtraction is unilateral, and in any one play of the game several links can be added and/or subtracted. Thus the Jackson–van den Nouweland rules are the Jackson–Wolinsky rules without the one-link-at-a-time restriction. Finally, in our discussion of Nash homogeneous directed networks we will assume that the rules of network formation are the Bala–Goyal rules. Under the Bala–Goyal rules an arc may be added or subtracted unilaterally by the initiating player involved in the arc and in any one play of the game only network changes brought about by an individual player are allowed. Note that all three of these sets of rules can be described as being uniform across networks. Under uniform rules the rules for changing a network are the same no matter which status quo network is being changed. Page, Wooders and Kamat [94] allow nonuniform rules and introduce a network representation of nonuniform rules.

Dominance Relations: Given players' preferences and the rules of network formation we will define a dominance relation over the feasible set of networks that incorporates both players preferences and the rules. Here we will focus on dominance relations that are either direct or indirect. Under direct dominance players are concerned with immediate consequences of their network formation strategies whereas under indirect dominance players are farsighted and consider the eventual consequences of their strategies.

General Results: A specification of the primitives induces two types of abstract games over homogeneous networks: (i) a network formation game with respect to the irreflexive dominance relation induced by preferences and rules, and (ii) a network formation game with respect to *path* dominance induced by this irreflexive dominance relation. We will begin by considering the game with respect to irreflexive dominance and present results on the existence of quasi-stable and stable networks. These results provide a network rendition of classical results from graph theory on the existence of quasi-stable sets and stable sets due to Chvatal and Lovasz [25], Berge [8], and Richardson [100]. We will also present a result on the existence and nonemptiness of the set of farsightedly consistent networks. This result is a network rendition of a result due to Chwe [26] for abstract games.

Next we will consider the game over homogeneous networks with respect to path dominance, and we will conclude that the following results hold:

1. Given preferences and the rules governing network formation, the set of homogeneous networks (linking or directed) contains a unique, finite, disjoint collection of nonempty subsets each constituting a *strategic basin of attraction*. These basins of attraction are the absorbing sets of the competitive process of network formation modeled via the game.
2. A stable set of homogeneous networks (in the sense of von Neumann–Morgenstern) with respect to path dominance consists of one network from each basin of attraction.

3. The path dominance core, defined as the set networks having the property that no network in the set is path dominated by any other homogeneous network, consists of one network from each basin of attraction containing a *single* network. Note that the path dominance core is contained in each stable set and is nonempty if and only if there is a basin of attraction containing a single network. As a corollary, we conclude that any homogeneous network contained in the path dominance core is constrained Pareto efficient.
4. From the results above it follows that if the dominance relation is transitive and irreflexive, then the path dominance core is nonempty.

These results are a special cases of results due to Page and Wooders [93].

Specific Results for Pairwise Stability, Strong Stability, Nash Stability, and Farsighted Consistency

What are the connections between our notions of stability for homogeneous networks (basins of attraction, path dominance stable sets, and path dominance core) and the notions of strong stability [39,61], pairwise stability [64], Nash stability [5], and farsighted consistency [26,94]? From the general results in [91] and [93] for heterogeneous directed networks, we will conclude for the case of homogeneous networks (linking or directed) that, depending on how we specialize the primitives of the model, the path dominance core is equal to the set of strongly stable networks, the set of pairwise stable networks, or the set of Nash networks. In particular, we will conclude that:

(a) If path dominance is induced by a direct dominance relation, then in the set of homogeneous linking networks the path dominance core is equal to the set of strongly stable networks.
(b) If, in addition, the rules of network formation are the Jackson–Wolinsky rules, then in the set of homogeneous linking networks the path dominance core is equal to the set of pairwise stable networks.
(c) If path dominance is induced by a direct dominance relation and if the rules of network formation are the Bala–Goyal rules, then in the set of homogeneous directed networks the path dominance core is equal to the set of Nash networks.

We can then conclude from (3) above that the existence of at least one basin of attraction containing a single network is, depending on how we specialize primitives, both necessary and sufficient for either (i) the existence of a strongly stable network, or (ii) a pairwise stable network, or (iii) a Nash network.

For path dominance induced by an indirect dominance relation, we can conclude from our prior results that for the case of homogeneous linking networks with Jackson–Wolinsky or Jackson–van den Nouweland rules or for the case of homogeneous directed networks with Bala–Goyal rules, each strategic basin of attraction has a nonempty intersection with the largest farsightedly consistent set of networks. This result, together with (2) above, implies that there always exists a path dominance stable set of homogeneous networks contained in the largest farsightedly consistent set. Thus, the path dominance core is contained in the largest consistent set. In light of our results on the path dominance core and stability (both strong and pairwise), we conclude that if path dominance is induced by an indirect dominance relation, then any homogeneous network contained in the path dominance core (i. e., the farsighted core) is not only farsightedly consistent but also strongly stable, as well as pairwise stable. Other papers using indirect dominance (or variations thereof) and farsighted consistency in games (not necessarily network formation games) include Li [76,77], Xue [129,130], Luo [79], Mariotti and Xue [80], Diamantoudi and Xue [36], and Mauleon and Vannetelbosch [84], Bhattacharya [9], Herrings, Mauleon and Vannetelbosch [55].

We remark that solution concepts defined using dominance relations have a long and distinguished history in the literature of game theory. First, consider the von Neuman–Morgenstern stable set (see [122] and [100]). The vN-M stable set is defined with respect to a dominance relation on a set of outcomes and consists of those outcomes that are externally and internally stable with respect to the given dominance relation. Similarly, Gillies [44] defines the core based on a given dominance relation. These solution concepts, with a few exceptions, have typically been applied to models of economies or cooperative games where the notion of dominance is based on what a coalition can achieve using only the resources owned by its members (cf., Aumann [3]) or a given set of utility vectors for each possible coalition (cf., Scarf [106]). Particularly notable exceptions are Schwartz [107], Panzer, Kalai and Schmeidler [66], Kalai and Schmeidler [67], Shenoy [109], Inarra, Kuipers, and Olaizola [58], and van Deemen [119]. Their motivations are in part similar to ours in that they take as given a set of possible choices for players (here consisting of set of networks) and a dominance relation and, based on these, describe a set of possible or likely outcomes called, by Kalai and Schmeidler, the admissible set. While their examples treat direct dominance, their general results have wider applications.

Because our objective here is to provide a unified game theoretic treatment of the main stability notions for net-

work formation games, many topics related to strategic networks are not covered here. For example, we do not discuss the conflict between stability and efficiency which is the main focus of the important papers by Dutta and Mutuswami [39] and Currarini and Morelli [30], and Mutuswami and Winter [87]. Nor do we treat the topic of network formation and cooperative games, the topic of the seminal paper by Myerson [88] and the excellent book by Slikker and van den Nouweland [113] among many other contributions, or the topic of network formation and evolution treated in Hojman and Szeidl [56]. Our game theoretic approach provides a snapshot of all possible network formation paths under dynamics which respect preferences and the rules of network formation. Our approach, however, is not explicitly dynamic. For network dynamics, we can only suggest to the reader the elegant papers by Skyrms and Pemantle [111], Watts [124], Jackson and Watts [62], Konishi and Ray [71], and Dutta, Ghosal, and Ray [38]. We do not touch on the topic of learning in networks, a topic which has been the focus of much work by Goyal [45,46], nor do we discuss random networks, introduced in economics by Kirman [68]. For random networks we refer the reader to the recent book by Vega-Redondo [121] and the references contained therein. Finally, we do not discuss the statistical mechanics of network formation. For this topic, we recommend to the reader the excellent papers by Blume [13] and Durlauf [37].

Because our focus is on foundational issues in strategic network formation, and in particular stability, we do not discuss any of the plethora of economic applications that can be found in the exploding literature on social and economic networks. For now, we can only offer the reader the following modest and incomplete list of topics and papers:

Development and insurance Bloch, Genicot, and Ray [12], Bramoulle and Kranton [18];

Employment and labor markets Rees [98], Granovetter [50], Boorman [16], Montgomery [86], Topa [118], Calvo-Armengol [19], Calvo-Armengol and Jackson [21], Calvo-Armengol and Jackson [22];

Industrial organization and R&D Demange and Henriet [32], Bloch [10], Kranton and Minehart [74], Goyal and Moraga-Gonzalez [49], Goyal and Joshi [47], Belleflamme and Bloch [7], Bloch [11], and Deroian and Gannon [35], Mauleon, Sempere-Monerris, and Vannetelbosch [83], Wang and Watts [123];

International trade Casella and Rauch [23,24] Zissimos [2], Goyal and Joshi [48], and Furusawa and Konishi [41];

Market microstructure Tesfatsion [116,117], Kirman, Herreiner and Weisbuch [69], Kranton and Minehart [75], Corominas-Bosch [28], and Even-Dar, Kearns, and Suri [40];

Public goods Bramoulle and Kranton [17];

Organizations, coordination, and communication Chwe [27], Currarini [29], Demange [34];

Marketing and advertising Galeotti and Moraga-Gonzalez [43].

The Primitives

Our abstract game of network formation rests on four primitives: The feasible set of networks, players' preferences, the rules of network formation, and a dominance relation over feasible networks. In this section, we discuss in detail these four primitives and in the next section, using these primitives we construct our abstract games of network formation.

Feasible Networks

Types of Networks Surprisingly, there is no agreed-upon definition of a network but rather several definitions depending on the application. But what all definitions have in common is a nonempty set of nodes and a precise mathematical description of how nodes are connected. What differentiates these various definitions then are the details of how nodes are connected. We begin with the most elementary notion of a network, the homogeneous linking network, and proceed to a more complex notion, the heterogeneous directed network introduced in [91].

Let N be a finite set of nodes, with typical element denoted by i, and let A be a finite set of link types or arc types, with typical element denoted by a. If the network is directed (to be defined below), we refer to the elements of A as arc types, otherwise we will refer to the elements of A as link types. For any set E, we denote by $P(E)$ the collection of *all* subsets of E. Finally, for any set E, we denote by $|E|$ the cardinality of E (note that $|\emptyset| = 0$).

Linking Networks

Definition 1 (Homogeneous Linking Networks, Myerson [88], Jackson–Wolinsky [64]) Let $P_2(N)$ denote the set of all subsets of N of size 2. A linking network, G, is a subset (possibly empty) of $P_2(N)$ and for any $G \subseteq P_2(N)$, each subset $\{i, i'\} \in G$ is called a link in G. The collection of all homogeneous linking networks is denoted by $P(P_2(N))$. □

Thus, $P_2(N)$ is the set of all possible links and a homogeneous linking network G is simply a subset of all possible links. For example, if $N = \{i_1, i_2, i_3\}$, then

$$P_2(N) = \{\{i_1, i_2\}, \{i_2, i_3\}, \{i_1, i_3\}\} \ ,$$

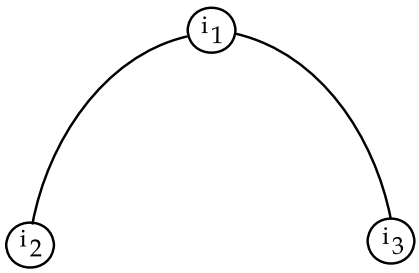

Networks and Stability, Figure 1
Homogeneous Linking Network G_1

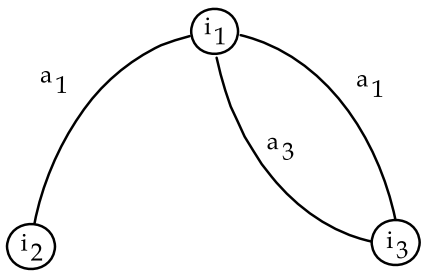

Networks and Stability, Figure 2
Heterogeneous Linking Network G_2

and the subset

$$G_1 = \{\{i_1, i_2\}, \{i_1, i_3\}\}$$

of $P_2(N)$ is a homogeneous linking network. Figure 1 depicts homogeneous linking network G_1.

Here, the link $\{i_1, i_3\} \in G_1$ denotes that nodes i_1 and i_3 are connected or linked. Note that all links are the same (i. e., links are homogeneous) and links have no orientation or direction. Also, note that in a homogeneous linking network, loops are not allowed by definition (a loop being a link between a node and itself). Finally, note that in a linking network multiple links between any pair of nodes are not allowed. However, because links are homogeneous, multiple links are unnecessary.

The following extended definition of a linking network allows for heterogeneous links. This heterogeneity is represented by a labeling of links using elements of the set A of link types.

Definition 2 (Heterogeneous Linking Networks)
A heterogeneous linking network, G, is a subset of $A \times P_2(N)$. Given any $G \subseteq A \times P_2(N)$, each ordered pair $(a, \{i, i'\}) \in G$ consisting of a link type and a link is called a labeled link in G. The collection of all heterogeneous linking networks is denoted by $P(A \times P_2(N))$. □

Thus, $A \times P_2(N)$ is the set of all possible labeled links and a heterogeneous linking network G is simply a subset of all possible labeled links. For example, given $N = \{i_1, i_2, i_3\}$ and $A = \{a_1, a_2, a_3\}$, the subset

$$G_2 = \{(a_1, \{i_1, i_2\}), (a_1, \{i_1, i_3\}), (a_3, \{i_1, i_3\})\}$$

of $A \times P_2(N)$ is a heterogeneous linking network. Figure 2 depicts heterogeneous linking network G_2.

Here, the labeled link $(a_1, \{i_1, i_3\}) \in G_2$ denotes that nodes i_1 and i_3 are linked by a type a_1 link. First, note that in network G_2 in addition to being linked by an a_1 link, nodes i_1 and i_3 are also linked by an a_3 link. Thus, in a heterogeneous linking network, links are not identical and multiple, distinct links between any given pair of nodes are possible. Second, note that in network G_2, nodes i_1 and i_2 – like nodes i_1 and i_3 – are linked by an a_1 link. Thus, in a heterogeneous linking network, link types can be used multiple times for different pairs of nodes. Finally, note that in a heterogeneous linking network, links are still without orientation or direction and loops are not possible.

Directed Networks The link orientation problem as well as the problem of loops is resolved by moving to directed networks. As is the case with linking networks, there are two categories of directed networks: Homogeneous directed networks and heterogeneous directed networks.

Definition 3 (Homogeneous Directed Networks)
A homogeneous directed network, G, is a subset of $N \times N$. Given any $G \subseteq N \times N$, each ordered pair $(i, i') \in G$ consisting of a beginning node i and an ending node i' is called an arc in G. The collection of all directed networks is denoted by $P(N \times N)$. □

Thus, $N \times N$ is the set of all possible arcs and a homogeneous directed network G is simply a subset of all possible arcs. For example, given $N = \{i_1, i_2, i_3\}$,

$$G_3 = \{(i_1, i_1), (i_1, i_2), (i_1, i_3), (i_3, i_1)\}$$

is a homogeneous directed network. Figure 3 depicts homogeneous directed network G_3.

Here, the arc $(i_1, i_3) \in G_3$ denotes that nodes i_1 and i_3 are connected by an arc *running from node i_1 to node i_3*. Note that because $(i_3, i_1) \in G_3$ there is also an arc running in the opposite direction from i_3 to i_1. Also, note that in a homogeneous directed network, while connections have direction, all connections are of the same type – that is, connections are homogeneous. Finally, note that in a directed network loops are allowed. For example, $(i_1, i_1) \in G_3$ and therefore in network G_3 there is an arc running from node i_1 to node i_1.

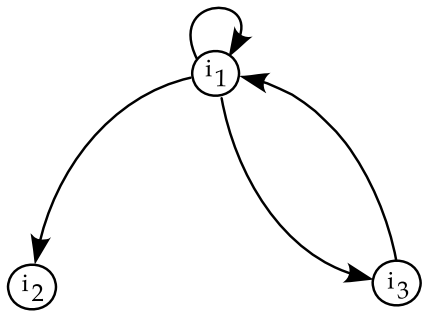

Networks and Stability, Figure 3
Homogeneous Directed Network G_3

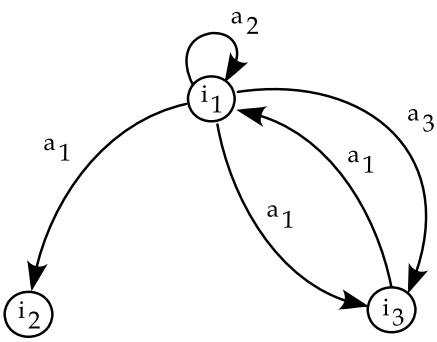

Networks and Stability, Figure 4
Heterogeneous Directed Network G_4

The following definition, from [94], allows for heterogeneous, multiple arcs by labeling arcs using the set A of arc types.

Definition 4 (Heterogeneous Directed Networks)
A heterogeneous directed network, G, is a subset of $A \times (N \times N)$. Given any $G \subseteq A \times (N \times N)$, each ordered pair $(a, (i, i')) \in G$ consisting of an arc type and an arc is called a labeled arc in G. The collection of all labeled directed networks is denoted by $P(A \times (N \times N))$. □

Thus, $A \times (N \times N)$ is the set of all possible labeled arcs and a heterogeneous directed network G is simply a subset of all possible labeled arcs. For example, given $N = \{i_1, i_2, i_3\}$ and $A = \{a_1, a_2, a_3\}$ the subset

$$G_4 = \{(a_2, (i_1, i_1)), (a_1, (i_1, i_2)), (a_1, (i_1, i_3)), (a_1, (i_3, i_1)), (a_3, (i_1, i_3))\}$$

of $A \times (N \times N)$ is a heterogeneous directed network. Figure 4 depicts heterogeneous directed network G_4.

Here, the labeled arc $(a_1, (i_1, i_3)) \in G_4$ denotes that nodes i_1 and i_3 are connected by an arc of type a_1 running from node i_1 to node i_3. Note that nodes i_1 and i_3 are also connected by an arc of type a_3 running from node i_1 to node i_3. Thus, in addition to having direction, connections are heterogeneous. Also, note that arc type a_1 is used three times in network G_4: Once in describing the connection running from i_1 to i_2, once in describing the connection running from i_1 to i_3, and once in describing the connection running from i_3 to i_1. Finally, note that loops are allowed. For example, $(a_2, (i_1, i_1)) \in G_4$ and therefore in network G_4 there is an a_2 arc running from node i_1 to node i_1.

Remarks

(1) In the terminology of graph theory (e. g., see Bollobas [15]), a homogeneous linking network is called a graph, while a homogeneous directed network is called a directed graph.

(2) The following notation is useful in describing heterogeneous directed networks. Given heterogeneous directed network $G \subseteq A \times (N \times N)$, let

$$\begin{aligned}
G(a) &:= \{(i, i') \in N \times N : (a, (i, i')) \in G\} \ , \\
G(i, i') &:= \{a \in A : (a, (i, i')) \in G\} \ , \\
G^+(i) &:= \{a \in A : (a, (i, i')) \in G \ \text{ for some } \ i' \in N\} \ , \\
G^-(i) &:= \{a \in A : (a, (i', i)) \in G \ \text{ for some } \ i' \in N\} \ , \\
G^+(a, i) &:= \{i' \in N : (a, (i, i')) \in G\} \ , \\
G^-(a, i) &:= \{i' \in N : (a, (i', i)) \in G\} \ .
\end{aligned}$$

For example, referring to heterogeneous directed network G_4 above (see Fig. 4),

$$\begin{aligned}
G_4(a_1) &:= \{(i_1, i_2), (i_1, i_3), (i_3, i_1)\} \ , \\
G_4(i_1, i_3) &:= \{a_1, a_3\} \ , \\
G_4^+(i_1) &:= \{a_1, a_2, a_3\} \ , \\
G_4^-(i_1) &:= \{a_1, a_2\} \ , \\
G_4^+(a_2, i_1) &:= \{i_1\} \ , \\
G_4^-(a_2, i_1) &:= \{i_1\} \ .
\end{aligned}$$

(3) The number $|G^+(i)|$ is the number arc types leaving node i in network G, while the number $|G^+(a, i)|$ is the out degree of node i for arc types a in network G. The number $|G^-(i)|$ is the number of arc types entering node i in network G, while the number $|G^-(a, i)|$ is the indegree of node i for arc type a in network G. For directed network G_4, we have for example,

$$\begin{aligned}
|G_4^+(i_2)| &= 0 \quad \text{and} \quad |G_4^-(i_2)| = 1 \ , \\
|G_4^+(a_1, i_1)| &= 2 \quad \text{and} \quad |G_4^-(a_1, i_1)| = 1 \ .
\end{aligned}$$

(4) Rockafellar [101], essentially defines a network G to be a nonempty subset of $A \times (N \times N)$ such that

for all $a \in A$, $G(a) \subseteq \{(i, i') \in N \times N\colon i \neq i'\}$, and $|G(a)| \leq 1$.

The Feasible Set In the abstract games of network formation we develop here, we will assume that the game is played over some feasible set of networks $\mathbb{G}$. Some examples of feasible sets are: The set of all homogeneous linking networks $P(P_2(N))$ as in Jackson–Wolinsky [64] and Jackson–van den Nouweland [61]; The set of all homogeneous directed networks $P(N \times N)$ as in Bala and Goyal [5] and an arbitrary subset of the set of heterogeneous directed networks $P(A \times (N \times N))$ as in [91] and [93]. The following example is taken from Page and Wooders [92] where the feasible set is taken to be the set of all club networks – a particular class of heterogeneous directed networks.

Example 1 (Club Networks) Let D be a finite set of players with typical element d and let C be a finite set of clubs or club locations with typical element c. As before, let A be a finite set of arc types. Finally, let $N = D \cup C$ be the set of nodes. We consider an abstract game of network formation played over the feasible set $\mathbb{G} \subset P(A \times (N \times N))$ of club networks where $G \in \mathbb{G}$ if and only if G is a nonempty subset of $A \times (D \times C)$ such that (i) for all players $d \in D$, the set

$$G(d) := \{(a, c) \in A \times C\colon (a, (d, c)) \in G\}$$

is nonempty and (ii) for all $(a, (d, c)) \in G$, $a \in A(d, c)$. Here, $A(d, c)$ is the set of actions (represented by arc types) available to player d in club c. Given club network $G \in \mathbb{G}$, $(a, (d, c)) \in G$ means that in club network G player d is a member of club c and takes action $a \in A(d, c)$ – or in the terminology of directed networks, that in club network G, there is an arc of type a running from node (player) d to node (club) c. Thus, in this example the feasible set is a set of bipartite directed networks.

We remark that the basic model of club formation underlying this example has a long history in the literature, going back to economies with essentially homogeneous agents modeled as games in characteristic function form (Shubik (bridge game) [110]) and serves as an example of several models in more recent literature on coalitional games (cf., Banerjee, Konishi and Sonmez [6], Bogomolnaia and Jackson [14], Diamantoudi and Xue [36]), and in economies with clubs (cf. Arnold and Wooders [2] and Allouch and Wooders [1]). As in Konishi, Le Breton and Weber [70] and Demange [33], for example, we allow "free entry" into clubs. □

Paths and Circuits A sequence of links $\{\{i, i'\}_k\}_k$ in $G \in \mathbb{G} \subseteq P(P_2(N))$ constitutes a path if each link $\{i, i'\}_k$ has one node in common with the preceding link $\{i, i'\}_{k-1}$ and the other node in common with the succeeding link $\{i, i'\}_{k+1}$. A circuit is a finite path $\{\{i, i'\}_k\}_{k=1}^h$ in G which begins at a node i and returns the same node. The length of a path is the number of links in the path.

A sequence of labeled links $\{(a, \{i, i'\})_k\}_k$ in $G \in \mathbb{G} \subseteq P(\mathbb{A} \times P_2(N))$ constitutes a path if each labeled link $(a, \{i, i'\})_k$ has one node in common with the preceding labeled link $(a, \{i, i'\})_{k-1}$ and the other node in common with the succeeding link $(a, \{i, i'\})_{k+1}$. A circuit is a finite path $\{(a, \{i, i'\})_k\}_{k=1}^h$ in G which begins at a node i and ends at the same node. The length of a path is the number of labeled links in the path.

A sequence of arcs $\{(i, i')_k\}_k$ in $G \in \mathbb{G} \subseteq P(N \times N)$ constitutes a path if the beginning node i of arc $(i, i')_k$ coincides with the ending node i' of preceding arc $(i, i')_{k-1}$. A circuit is a finite path $\{(i, i')_k\}_{k=1}^h$ in G such that node i of arc $(i, i')_1$ and node i' of arc $(i, i')_h$ are the same node. The length of a path is the number of arcs in the path.

Finally, a sequence of labeled arcs $\{(a, (i, i'))_k\}_k$ in $G \in \mathbb{G} \subseteq P(A \times (N \times N))$ constitutes a path if the beginning node i of labeled arc $(a, (i, i'))_k$ coincides with the ending node i' of preceding arc $(a, (i, i'))_{k-1}$. A circuit is a finite path $\{(a, (i, i'))_k\}_{k=1}^h$ in G such that node i of labeled arc $(a, (i, i'))_1$ and node i' of labeled arc $(a, (i, i'))_h$ are the same node. The length of a path is the number of labeled arcs in the path.

In Fig. 5, $\{(a_1, (i_3, i_1))_1, (a_2, (i_1, i_1))_2, (a_1, (i_1, i_2))_3\}$ is a path in G_4 of length 3, while $\{(a_1, (i_3, i_1))_1, (a_2, (i_1, i_1))_2, (a_3, (i_1, i_3))_3\}$ is a circuit in G_4 of length 3.

Players' Preferences

For the remainder of this entry we will assume that the set of players is given by the set of nodes N. Thus, henceforth the nodes represent players in the game of network formation.

Let $\Gamma(N)$ denote the collection of all coalitions of players (i. e., *nonempty* subsets of N) with typical element denoted by S.

For each player $i \in N$ let $\succ_i$ be an irreflexive binary relation on $\mathbb{G}(= P(P_2(N))$ or $P(N \times N))$ and write $G' \succ_i G$ if player $i \in N$ prefers network $G' \in \mathbb{G}$ to network $G \in \mathbb{G}$. Because $\succ_i$ is irreflexive, $G \not\succ_i G$ for all networks $G \in \mathbb{G}$. Coalition $S' \in \Gamma(N)$ *prefers network G' to network G*, written $G' \succ_{S'} G$, if $G' \succ_i G$ for all players $i \in S'$. Note that because players' preferences $\{\succ_i\}_{i \in N}$ are irreflexive, coalitional preferences, $\{\succ_S\}_{S \in \Gamma(N)}$, are also irreflexive.

A Remark on Weak Preferences Players are said to have weak preferences on $\mathbb{G}(= P(P_2(N))$ or $P(N \times N))$, denoted by $\succsim_i$ if $G' \succsim_i G$ means that player i either strongly

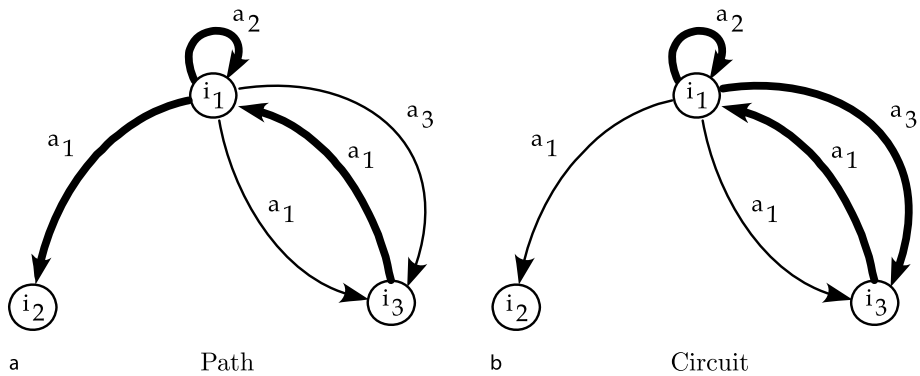

Networks and Stability, Figure 5
Path and circuit in Heterogeneous Directed Network G_4

prefers G' to G (denoted $G' \succ_i G$) or is indifferent between G' and G (denoted $G' \sim_i G$). If coalitional preferences are based on weak preference, then we say that coalition $S' \in P(N)$ *weakly prefers* network G' to network G, written $G' \succ_{wS'} G$, if for all players $i \in S'$, $G' \succsim_i G$ and if *for at least one player* $i' \in S'$, $G' \succ_{i'} G$. Note that if preferences are weak and $G' \succ_{wS'} G$ where S' consists of a single player i', so that $S' = \{i'\}$, then $G' \succ_{i'} G$. Finally, note that weak coalitional preferences $\{\succ_{wS}\}_{S \in \Gamma(N)}$ are irreflexive (i. e., $G \not\succ_{wS} G$ for all $G \in \mathbb{G}$ and $S \in \Gamma(N)$).

Network Payoff Functions In many applications, players' preferences are specified via real-valued network payoff functions, $\{v_i(\cdot)\}_{i \in N}$. If this is the case, then for each player $i \in N$ and each network $G \in \mathbb{G}$, $v_i(G)$ is the payoff to player i in network G. Note that the payoff $v_i(G)$ to player i in network G depends on the entire network. Thus, the player may be affected by connections between other players even when he himself has no direct or indirect connection with those players. Intuitively, 'widespread' network externalities are allowed.

Given payoff functions $\{v_i(\cdot)\}_{i \in N}$, player i prefers network G' to network G if $v_i(G') > v_i(G)$. Coalitional preferences can then be specified by stating that coalition $S' \in \Gamma(N)$ prefers network G' to network G if $v_i(G') > v_i(G)$ for all $i \in S'$.

Preference Supernetworks By viewing each network G in feasible set $\mathbb{G}$ as a node in a larger network, we can represent coalitional preferences as a heterogeneous directed network. To begin, let

$$\mathcal{P} := \{p_S : S \in \Gamma(N)\}$$

denote the set of arc labels for preference arcs (or p-arcs for short).

Definition 5 (Coalitional Preference Supernetworks, [94]) Given feasible set $\mathbb{G}$ $(= P(P_2(N))$ or $P(N \times N))$, a coalitional preference supernetwork $\mathbf{P}$ is a subset of $\mathcal{P} \times (\mathbb{G} \times \mathbb{G})$ such that $(p_{S'}, (G, G'))$ is contained in $\mathbf{P}$ if and only if $G' \succ_{S'} G$.

The Rules of Network Formation

The rules of network formation are specified via a collection of coalitional effectiveness relations $\{\rightarrow_S\}_{S \in \Gamma(N)}$ defined on the feasible set of networks $\mathbb{G}$ $(= P(P_2(N))$ or $P(N \times N))$. Each effectiveness relation $\rightarrow_S$ represents what a coalition S can do. Thus, if $G \rightarrow_S G'$ this means that under the rules of network formation coalition $S \in \Gamma(N)$ can change network $G \in \mathbb{G}$ to network $G' \in \mathbb{G}$ by adding, subtracting, or replacing connections in G (where, depending on the feasible set, a *connection is a link or an arc*).

Examples of Network Formation Rules

Jackson–Wolinsky Rules [64] *(Bilateral-Unilateral Rules)* Assume that the feasible set of networks $\mathbb{G}$ is equal to the set of homogeneous linking networks $P(P_2(N))$. Under the Jackson–Wolinsky rules of network formation (see [64]),

(i) adding a link from player i to player i' requires that both players i *and* i' agree to add the link (i. e., link addition is bilateral);
(ii) subtracting a link from player i to player i' requires that player i *or* player i' or both agree to subtract the link (i. e., link subtraction can be unilateral);
(iii) link addition or link subtraction takes place one link at a time.

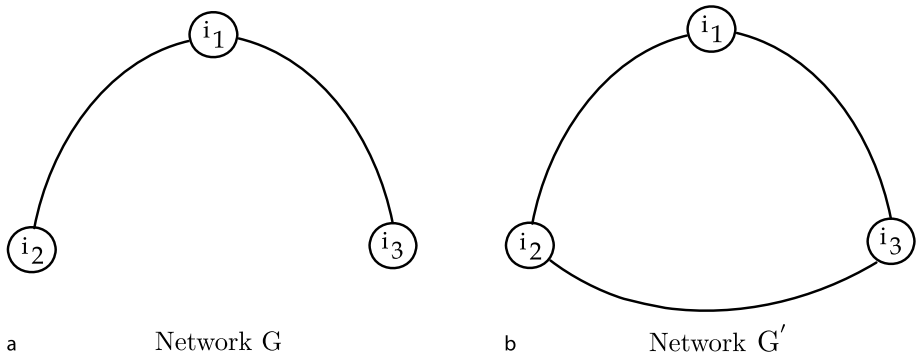

Networks and Stability, Figure 6
a Network *G*. b Network *G′*

Thus, for any pair of networks G and G' in $\mathbb{G}$, if $G \to_S G'$ and $G \neq G'$, then

$$\begin{aligned} \text{either} \quad & G' = G \cup \{i, i'\} \quad \text{for some } \{i, i'\} \in P_2(N) \\ & \quad \text{and} \quad S = \{i, i'\} \\ \text{or} \quad & G' = G \backslash \{i, i'\} \quad \text{for some } \{i, i'\} \in G \\ & \quad \text{and} \quad S = \{i\} \quad \text{or } S = \{i'\} \quad \text{or } S = \{i, i'\} . \end{aligned}$$

To illustrate, consider Fig. 6 depicting two homogeneous linking networks G and G'.

Observe that

$$G' = G \cup \{i_2, i_3\} \quad \text{and} \quad G = G' \backslash \{i_2, i_3\} .$$

Under the effectiveness relations implied by the Jackson–Wolinsky rules, for networks G and G' we have

$$G \underset{\{i_2,i_3\}}{\longrightarrow} G' , \quad G' \underset{\{i_2,i_3\}}{\longrightarrow} G , \quad G' \underset{\{i_2\}}{\longrightarrow} G , \quad G' \underset{\{i_3\}}{\longrightarrow} G .$$

Jackson–van den Nouweland Rules [61] *(Bilateral-Unilateral Rules)* Again assume that the feasible set of networks $\mathbb{G}$ is equal to the set of homogeneous linking networks $P(P_2(N))$. Under the Jackson–van den Nouweland rules of network formation,

(i) adding a link from player i to player i' requires that both players i *and* i' agree to add the link (i. e., link addition is bilateral);
(ii) subtracting a link from player i to player i' requires that player i *or* player i' or both agree to subtract the link (i. e., link subtraction can be unilateral).

Thus, the Jackson–van den Nouweland rules are the Jackson–Wolinsky rules without the one-link-at-a-time restriction. Note that if link addition is bilateral and link subtraction is unilateral (i. e., if rules (i) and (ii) hold), then $G \to_S G''$ and $G \neq G''$ implies that

(i) if $\{i, i'\} \in G''$ and $\{i, i'\} \notin G$, then $\{i, i'\} \subseteq S$;

and

(ii) if $\{i, i'\} \notin G''$ and $\{i, i'\} \in G$, then $\{i, i'\} \cap S \neq \emptyset$.

To illustrate, consider Fig. 7 depicting two homogeneous linking networks G and G''.

Observe that

$$G'' = (G \backslash \{i_1, i_3\}) \cup \{i_2, i_3\}$$

and

$$G = \big(G'' \backslash \{i_2, i_3\}\big) \cup \{i_1, i_3\} .$$

Under the effectiveness relations implied by the Jackson–van den Nouweland rules, for networks G and G'' we have

$$G \underset{\{i_1,i_2,i_3\}}{\longrightarrow} G'', \quad G \underset{\{i_2,i_3\}}{\longrightarrow} G'', \quad G'' \underset{\{i_1,i_3\}}{\longrightarrow} G, \quad G'' \underset{\{i_1,i_2,i_3\}}{\longrightarrow} G.$$

Note that under the one-link-at-a-time restriction, it is not possible under the Jackson–Wolinsky rules to move directly from network G to network G'' or directly from network G'' to network G (i. e., G and G'' are *not* related under the effectiveness relations $\{\to_S\}_{S \in \Gamma(N)}$). Instead, under the Jackson–Wolinsky rules, the change from G to G'' or from G'' to G requires two moves. For example,

$$\text{first } G \underset{\{i_2,i_3\}}{\longrightarrow} G' , \quad \text{and then,} \quad G' \underset{\{i_3\}}{\longrightarrow} G'' \quad \text{or } G' \underset{\{i_1\}}{\longrightarrow} G'' ;$$

or

$$\text{first } G'' \underset{\{i_1,i_3\}}{\longrightarrow} G' , \quad \text{and then,} \quad G' \underset{\{i_3\}}{\longrightarrow} G \quad \text{or } G' \underset{\{i_2\}}{\longrightarrow} G .$$

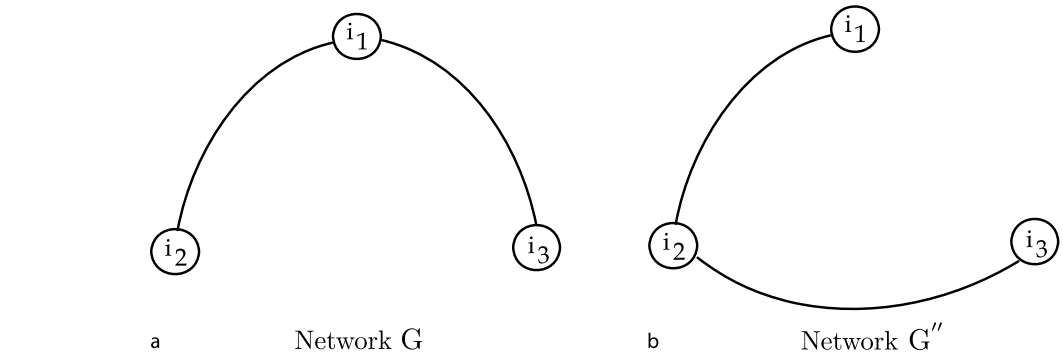

Networks and Stability, Figure 7
a Network *G*. b Network *G″*

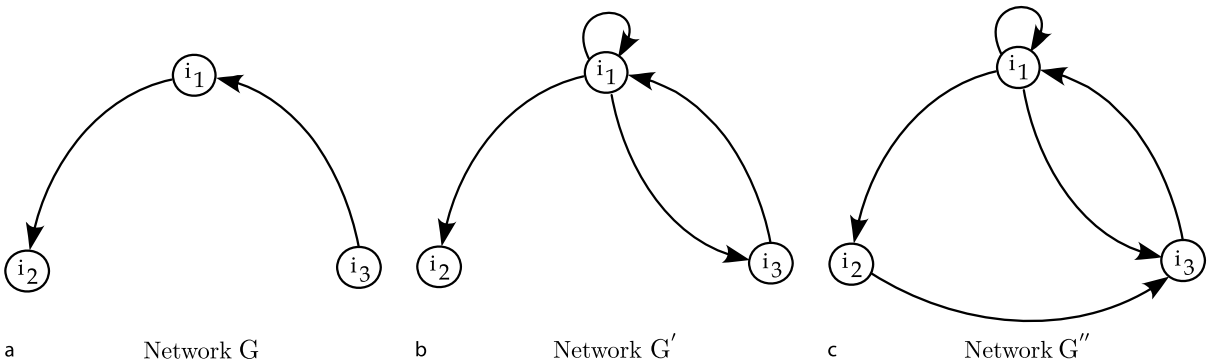

Networks and Stability, Figure 8
a Network *G*. b Network *G′*. c Network *G″*

Bala–Goyal Rules [5] *(Noncooperative Rules – Unilateral-Unilateral Rules)* Now assume that the feasible set of networks is equal to the set of homogeneous directed networks $P(N \times N)$. Translating Bala and Goyal rules into our notation and terminology,

(i) adding an arc from player i to player i' requires only that player i agree to add the arc (i. e., arc addition is unilateral and can be carried out only by the initiator, player i);
(ii) subtracting an arc from player i to player i' requires only that player i agree to subtract the arc (i. e., arc subtraction is unilateral and can be carried out only by the initiator, player i);
(iii) $G \rightarrow_S G'$ implies that $|S| = 1$ (i. e., only network changes brought about by individual players are allowed).

We shall also refer to rules (i)–(iii) as noncooperative. Note that a player i can add or subtract an arc to player i' without regard to the preferences of player i' and can add and/or subtract arcs to several players simultaneously and can do so without regard to those players' preferences. Thus in general under noncooperative rules, effectiveness relations display a type of symmetry, and in particular, if $G \underset{\{i\}}{\longrightarrow} G'$, then $G' \underset{\{i\}}{\longrightarrow} G$.

To illustrate, consider Fig. 8 depicting three homogeneous directed networks G, G', and G''.

Under the effectiveness relations implied by noncooperative rules for networks G and G' in Fig. 8 we have

$$G \underset{\{i_1\}}{\longrightarrow} G' \,, \quad G \underset{\{i_1\}}{\longrightarrow} G' \,.$$

Note that under noncooperative rules, networks G and G'' in Fig. 8 are *not* related under the effectiveness relations $\{\rightarrow_{\{i\}}\}_{i \in N}$. However, under the noncooperative rules we have, for example, the following effectiveness relations

$$G \rightarrow_{\{i_1\}} G' \,, \quad G' \rightarrow_{\{i_2\}} G''$$

and

$$G'' \rightarrow_{\{i_2\}} G' \,, \quad G' \rightarrow_{\{i_1\}} G \,.$$

Rules Supernetworks Again by viewing each network G in feasible set $\mathbb{G}$ as a node in a larger network, we can represent the rules of network formation as a heterogeneous directed network. To begin, let

$$\mathcal{M} := \{m_S : S \in \Gamma(N)\}$$

denote the set of arc labels for move arcs (or m-arcs for short).

Definition 6 (Rules Supernetworks, [94]) Given feasible set $\mathbb{G}$ ($= P(P_2(N))$ or $P(N \times N)$), a rules supernetwork $\mathbf{R}_\rho$ is a subset of $\mathcal{M}\times(\mathbb{G} \times \mathbb{G})$ such that $(m_{S'}, (G, G'))$ is contained in $\mathbf{R}_\rho$ if and only if $G \rightarrow_{S'} G'$, where ρ denotes the name of the network formation rules in force. We shall adopt the convention that $\rho = jw$ if the rules are Jackson–Wolinsky, $\rho = jn$ if the rules are Jackson–van den Nouweland, and $\rho = bg$ if the rules are Bala–Goyal or noncooperative. □

Supernetworks

Given feasible set $\mathbb{G}$ ($= P(P_2(N))$ or $P(N \times N)$), coalitional preferences $\{\succ_S\}_{S\in\Gamma(N)}$, and coalitional effectiveness relations $\{\rightarrow_S\}_{S\in\Gamma(N)}$ can be represented by a heterogeneous directed network called a supernetwork (see [89] and [94]). In particular, given preference supernetwork $\mathbf{P}$ and rules supernetwork $\mathbf{R}_\rho$, the corresponding supernetwork is given by

$$\mathbf{G}_\rho := \mathbf{P} \cup \mathbf{R}_\rho \,.$$

Letting $\mathcal{A} := \mathcal{P} \cup \mathcal{M}$ (i. e., the union of all preference arcs and move arcs), then

$$\mathbf{G}_\rho \subset \mathcal{A} \times (\mathbb{G} \times \mathbb{G}) \,.$$

Dominance Relations

Direct Dominance Given feasible set $\mathbb{G}$ ($= P(P_2(N))$ or $P(N \times N)$), coalitional preferences $\{\succ_S\}_{S\in\Gamma(N)}$, and coalitional effectiveness relations $\{\rightarrow_S\}_{S\in\Gamma(N)}$, network $G' \in \mathbb{G}$ *directly dominates* network $G \in \mathbb{G}$, written $G' \rhd G$, if for some coalition $S' \in \Gamma(N)$,

$$G \prec_{S'} G' \quad \text{and} \quad G \underset{S'}{\longrightarrow} G' \,.$$

Thus, network G' directly dominates network G if some coalition S' prefers G' to G *and* if under the rules of network formation coalition S' has the power to change G to G'.

Note that direct dominance is irreflexive but not in general transitive. Also note that if $\mathbf{G}_\rho$ is the supernetwork, then $G' \rhd G$ if and only if $(p_{S'}, (G, G')) \in \mathbf{G}_\rho$ and $(m_{S'}, (G, G')) \in \mathbf{G}_\rho$, for some coalition S'.

Indirect Dominance Given feasible set $\mathbb{G}$ ($= P(P_2(N))$ or $P(N \times N)$), coalitional preferences $\{\succ_S\}_{S\in\Gamma(N)}$, and coalitional effectiveness relations $\{\rightarrow_S\}_{S\in\Gamma(N)}$, network $G' \in \mathbb{G}$ *indirectly dominates* network $G \in \mathbb{G}$, written $G' \rhd\rhd G$, if there is a *finite* sequence of networks,

$$G_0 \,,\quad G_1 \,,\quad \ldots \,,\quad G_h \,,$$

with $G = G_0$, $G' = G_h$, and $G_k \in \mathbb{G}$ for $k = 0, 1, \ldots, h$, and a corresponding sequence of coalitions,

$$S_1 \,,\quad S_2 \,,\quad \ldots \,,\quad S_h \,,$$

such that for $k = 1, 2, \ldots, h$

$$G_{k-1} \underset{S_k}{\longrightarrow} G_k \,, \quad \text{and} \quad G_{k-1} \prec_{S_k} G_h \,.$$

Note that if network G' indirectly dominates network G (i. e., if $G' \rhd\rhd G$), then what matters to the initially deviating coalition S_1, as well as all the coalitions along the way, is that the ultimate network outcome $G' = G_h$ be preferred. Thus, for example, the initially deviating coalition S_1 will not be deterred from changing network G_0 to network G_1 even if network G_1 is not preferred to network $G = G_0$, as long as the ultimate network outcome $G' = G_h$ is preferred to G_0, that is, as long as $G_0 \prec_{S_1} G_h$. Finally, note that indirect dominance is irreflexive but not in general transitive.

In order to capture the idea of farsightedness in strategic behavior, Chwe [26] analyzed abstract games equipped with indirect dominance relations in great detail, introducing the equilibrium notions of consistency and largest consistent set. The basic idea of indirect dominance goes back to the work of Guilbaud [53] and Harsanyi [54].

Given the supernetwork representation of preferences and rules, $\mathbf{G}_\rho$, we can write, $G' \rhd\rhd G$ if there is a *finite* sequence of networks,

$$G_0 \,,\quad G_1 \,,\quad \ldots \,,\quad G_h \,,$$

with $G = G_0$, $G' = G_h$, and $G_k \in \mathbb{G}$ for $k = 0, 1, \ldots, h$, and a corresponding sequence of coalitions,

$$S_1 \,,\quad S_2 \,,\quad \ldots \,,\quad S_h \,,$$

such that for $k = 1, 2, \ldots, h$

$$(m_{S_k}, (G_{k-1}, G_k)) \in \mathbf{G}_\rho \,,$$

and

$$(p_{S_k}, (G_{k-1}, G_h)) \in \mathbf{G}_\rho \,.$$

Path Dominance Any irreflexive dominance relation $>$ on $\mathbb{G}$ ($= P(P_2(N))$ or $P(N \times N)$) – for example, direct or indirect – induces a path dominance relation on the set of networks (sometimes referred to as the transitive closure of $>$). In particular, corresponding to dominance relation $>$ on networks $\mathbb{G}$ there is a corresponding path dominance relation $\geq_p$ on $\mathbb{G}$ specified as follows: Network $G' \in \mathbb{G}$ path dominates network $G \in \mathbb{G}$ with respect to $>$ (i. e., with respect to the underlying dominance relation $>$), written $G' \geq_p G$, if $G' = G$ or if there exists a *finite* sequence of networks $\{G_k\}_{k=0}^h$ in $\mathbb{G}$ with $G_h = G'$ and $G_0 = G$ such that for $k = 1, 2, \ldots, h$

$$G_k > G_{k-1} \ .$$

We refer to such a finite sequence of networks as a *finite domination path* and we say network G' is $>$*-reachable* from network G if there exists a finite domination path from G to G'. Thus,

$$G' \geq_p G \quad \text{if and only if} \quad \begin{cases} G' & \text{is } >\text{-reachable from} \\ \quad G \ , & \text{or} \\ G' = G \ . & \end{cases}$$

Note that, even though the underlying dominance relation $>$ is irreflexive and intransitive or transitive, the induced path dominance relation $\geq_p$ on $\mathbb{G}$ is both reflexive ($G \geq_p G$) and transitive ($G' \geq_p G$ and $G'' \geq_p G'$ implies that $G'' \geq_p G$).

$>$-Supernetworks Let $>$ denote the irreflexive dominance relation on $\mathbb{G}$. It is often useful to represent $>$ as a *homogeneous* directed network, $\mathbf{D}_>$, where $\mathbf{D}_>$ is a subset of $\mathbb{G} \times \mathbb{G}$ and where $>$-arc $(G, G') \in \mathbf{D}_>$ if and only if $G' > G$ (i. e., if and only if G' $>$-dominates G). We call such a homogeneous directed network (or directed graph) a $>$-supernetwork. For example, suppose $\mathbb{G} = \{G_1, G_2, G_3, \ldots, G_7\}$ and suppose the dominance relation $>$ on $\mathbb{G}$ is a direct dominance relation and has the supernetwork representation given in Fig. 9.

Note that network G_5 is $>$-reachable through $\mathbf{D}_>$ from network G_1 by the domination path given by the $>$-arc sequence

$$\{(G_1, G_6)_1, (G_6, G_2)_2, (G_2, G_3)_3, (G_3, G_4)_4, (G_4, G_5)_5\} \ .$$

Thus, G_5 path dominates G_1. Note that network G_2 is $>$-reachable from network G_2 by the domination circuit given by the $>$-arc sequence

$$\{(G_2, G_3)_1, (G_3, G_4)_2, (G_4, G_5)_3, (G_5, G_2)_4\} \ .$$

and that network G_3 is $>$-reachable from network G_3 by two domination circuits given by the $>$-arc sequences

$$\{(G_3, G_4)_1, (G_4, G_5)_2, (G_5, G_2)_3, (G_2, G_3)_4\}$$

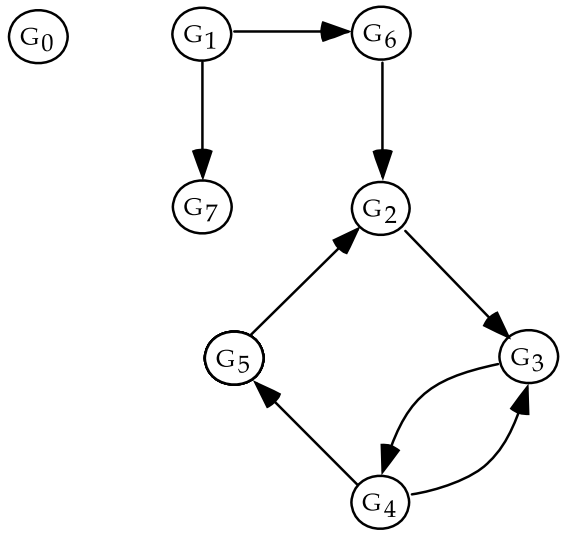

Networks and Stability, Figure 9
$>$-supernetwork $\mathbf{D}_>$

and

$$\{(G_3, G_4)_1, (G_4, G_3)_2\} \ .$$

Because networks G_2 and G_5 are on the same circuit, G_5 is $>$-reachable from G_2 *and* G_2 is $>$-reachable from G_5. Thus, G_5 path dominates G_2 (i. e., $G_5 \geq_p G_2$) *and* G_2 path dominates G_5 (i. e., $G_2 \geq_p G_5$). The same cannot be said of networks G_1 and G_5. In particular, while $G_5 \geq_p G_1$, it is not true that $G_1 \geq_p G_5$ because G_1 is not $>$-reachable from G_5. Finally, note that network G_0 is isolated in $\mathbf{D}_>$. In particular, G_0 is *not* reachable through $\mathbf{D}_>$ from any network in $\mathbb{G}$ and no network in $\mathbb{G}$ is reachable through $\mathbf{D}_>$ from G_0. In general, a network $G \in \mathbb{G}$ is isolated if there does not exist a network $G' \in \mathbb{G}$ with $G' \geq_p G$ or $G \geq_p G'$.

Note that if the direct dominance relation with $>$-supernetwork depicted in Fig. 9 has underlying coalitional preferences $\{\succ_S\}_{S \in \Gamma(N)}$ and coalitional effectiveness relations $\{\rightarrow_S\}_{S \in \Gamma(N)}$, then the $>$-arc from network G_3 to network G_4 in Fig. 9 means that for some coalition S, G_4 is preferred to G_3 and more importantly, that coalition S has the power to change network G_3 to network G_4. Thus, $G_3 \prec_{sS} G_4$ and $G_3 \rightarrow_S G_4$. But because there is a $>$-arc in the opposite direction, from network G_4 to network G_3, G_3 also directly dominates G_4. Thus for some coalition S' disjoint from coalition S ($S' \cap S = \emptyset$), $G_4 \prec_{sS'} G_3$ and $G_4 \rightarrow_{S'} G_3$. Finally, note that if coalitional preferences over networks are weak (i. e., are based on weak preferences), then the statement, 'for some coalition S' disjoint from coalition S can be weakened to for some coalition S' not equal to coalition S. With this weakening, the require-

ment that the intersection of S and S' be empty is no longer needed.

Abstract Games of Network Formation and Stability

An abstract game of network formation consists of a feasible set of networks equipped with a dominance relation. We shall consider two classes of games: (i) games where the feasible set of networks $\mathbb{G}$ ($= P(P_2(N))$ or $P(N \times N)$) is equipped with an irreflexive dominance relation $>$, either direct or indirect (i. e., $>$ is equal to $\rhd$ or $\rhd\rhd$), induced by coalitional preferences and network formation rules; and (ii) games where the feasible set of networks is equipped with a path dominance relation $\geq_p$ induced by such an irreflexive dominance relation.

Network Formation Games with Respect to Irreflexive Dominance

In this section we consider the abstract game with respect to irreflexive dominance given by the pair

$$(\mathbb{G}, >) .$$

Throughout this section we will assume that primitives are represented by supernetwork $\mathbf{G}_\rho := \mathbf{P} \cup \mathbf{R}_\rho$ (where ρ is equal to *jw*, *jn*, or *bg*) and $>$-supernetwork $\mathbf{D}_>$ (where $>$ is equal to $\rhd$ or $\rhd\rhd$), and that the feasible set of networks $\mathbb{G}$ is equal to the set of homogeneous linking networks $P(P_2(N))$ or the set of homogeneous directed networks $P(N \times N)$.

Quasi-Stability and Stability We define the $>$-*distance* from G_0 to G_1 in $\mathbf{D}_>$ to be the length of the *shortest* $>$-path from G_0 to G_1 if G_1 is $>$-*reachable* from G_0 in $\mathbf{D}_>$, and $+\infty$ if G_1 is not reachable from G_0 in $\mathbf{D}_>$. We denote the distance *from* G_0 to G_1 in $\mathbf{D}_>$ by $d_{\mathbf{D}_>}(G_0, G_1)$. Thus,

$$d_{\mathbf{D}_>}(G_0, G_1) := \begin{cases} \text{length of shortest } >\text{-path} \\ \qquad \text{from } G_0 \text{ to } G_1 \text{ in } \mathbf{D}_>, \\ \qquad \text{if } G_1 \text{ isreachable from } G_0, \\ +\infty, \text{ if } G_1 \text{ is not reachable from} \\ \qquad G_0 \text{ in } \mathbf{D}_>. \end{cases}$$

The following are network renditions of quasi-stable and stable sets.

Definition 7 (Quasi-Stable Sets and Stable Sets, [8,25])

(1) A subset $\mathbb{Q}$ of networks in $\mathbb{G}$ is said to be quasi-stable for network formation game $(\mathbb{G}, >)$ if,
   (a) $\mathbb{Q}$ is internally stable, that is, $d_{\mathbf{D}_>}(G_0, G_1) \geq 2$, whenever G_0 and G_1 are in $\mathbb{Q}$, with $G_0 \neq G_1$, and
   (b) $\mathbb{Q}$ is externally quasi-stable, that is, given any $G_0 \notin \mathbb{Q}$, there exists $G_1 \in \mathbb{Q}$ with $d_{\mathbf{D}_>}(G_0, G_1) \leq 2$.

(2) A subset $\mathbb{S}$ of networks in $\mathbb{G}$ is said to be stable for network formation game $(\mathbb{G}, >)$ if,
   (a) $\mathbb{S}$ is internally stable and
   (b) $\mathbb{S}$ is externally stable, that is, given any $G_0 \notin \mathbb{S}$, there exists $G_1 \in \mathbb{S}$ with $d_{\mathbf{D}_>}(G_0, G_1) \leq 1$. □

Thus, if $\mathbb{Q}$ is externally quasi-stable, a path of length at most 2 is required to get from any network outside of $\mathbb{Q}$ to a network in $\mathbb{Q}$, whereas, if $\mathbb{Q}$ is externally stable a path of length at most 1 is required.

Letting

$$P_>(G_0) := \{G \in \mathbb{G} : G > G_0\} ,$$

an alternative way to write part (2) of the definition above is as follows:

(2)′ A subset $\mathbb{S}$ of networks in $\mathbb{G}$ is said to be stable if

(a) (internal stability) $G \in \mathbb{S}$ implies that $P_>(G) \cap \mathbb{S} = \emptyset$ and
(b) (external stability) $G \notin \mathbb{S}$ implies that $P_>(G) \cap \mathbb{S} \neq \emptyset$.

If $\mathbb{S}$ is stable (or quasi-stable), then it is automatically nonempty. Note that a stable set $\mathbb{S}$ is simply a von Neuman–Morgenstern stable set with respect to the dominance relation $>$ defined on $\mathbb{G}$. Also, note that if $\mathbb{S}$ is stable then it is automatically quasi-stable.

We now state a remarkably simple result on the existence of quasi-stable sets. This result is a network rendition of a general result due to Chvatal and Lovasz [25] on the existence of quasi-stable sets in directed graphs (here the directed graph is the $>$-supernetwork $\mathbf{D}_>$; also see, Galeana-Sanchez and Li [42]).

Theorem 1 (Existence of quasi-stable sets for network formation games, [89]) *There exists a quasi-stable set* $\mathbb{Q}$ *for network formation game* $(\mathbb{G}, >)$.

In fact, it follows from the Theorem due to Chvatal and Lovasz [25] that any finite set $\mathbb{Z}$ equipped with an irreflexive binary relation $\prec$ has a $\prec$-quasi-stable set. Moreover, if the relation $\prec$ is transitive, then any $\prec$-quasi-stable set is $\prec$-stable.

Next we state two results on the existence of stable sets.

Theorem 2 (Existence of stable sets for network formation games, [89])

(1) If $\mathbf{D}_>$ *contains no* $>$*-circuits, then there exists a unique stable* $\mathbb{S}$ *for* $(\mathbb{G}, >)$.
(2) If $\mathbf{D}_>$ *contains no* $>$*-circuits of odd length, then there exists a stable* $\mathbb{S}$ *for* $(\mathbb{G}, >)$.

Part (1) of Theorem 2 is an immediate consequence of a 1958 result due to Berge (see Theorem 4, p. 48 in Berge [8]). Part (2) of Theorem 2 is a supernetwork version of the classical result due to Richardson [100]. A $>$-circuit in supernetwork $\mathbf{D}_>$ is said to be of odd length if there is an odd number of connections in the circuit.

Farsighted Consistency Chwe [26] in an influential paper introduced the notion of farsighted consistency in an abstract game. The following is a definition of farsighted consistency for abstract games of network formation.

Definition 8 (Farsighted Consistency, [89]) A subset $\mathbb{F}$ of networks in $\mathbb{G}$ is said to be farsightedly consistent for network formation game $(\mathbb{G}, \rhd\rhd)$ if,

$$\begin{gathered}\text{for all } G_0 \in \mathbb{F}\ , \\ (m_{S_1}, (G_0, G_1)) \in \mathbf{G}_\rho \quad \text{for some } G_1 \in \mathbb{G} \\ \text{and some coalition } S_1\ , \quad \text{implies that} \\ \text{there exists } G_2 \in \mathbb{F} \\ \text{with } G_2 = G_1 \quad \text{or } G_2 \rhd\rhd G_1 \quad \text{such that}\ , \\ (p_{S_1}, (G_0, G_2)) \notin \mathbf{G}_\rho\ .\end{gathered}$$

□

In words, a subset of directed networks $\mathbb{F}$ is said to be farsightedly consistent if given any network $G_0 \in \mathbb{F}$ and any m_{S_1}-deviation to network $G_1 \in \mathbb{G}$ by coalition S_1 (via adding, subtracting, or replacing arcs in accordance with $\mathbf{R}_\rho$) there exists further deviations leading to some network $G_2 \in \mathbb{F}$ where the initially deviating coalition S_1 is not better off – and possibly worse off. A network $G \in \mathbb{G}$ is said to be farsightedly consistent if $G \in \mathbb{F}$ where $\mathbb{F}$ is a farsightedly consistent set.

If (i) $\mathbb{G} = P(P_2(N))$, (ii) coalitional preferences are weak, denoted by $\{\succ_{wS}\}_{S\in\Gamma(N)}$, so that indirect dominance is weak (denoted by $\rhd\rhd_w$ and defined in the obvious way), and (iii) coalitional effectiveness relations are determined by Jackson–Wolinsky rules, then the notion of farsighted consistency above (essentially due to Chwe [26]) is closely related to the notion of pairwise farsighted stability introduced in Herrings, Mauleon, and Vannetelbosch [55].

For any game $(\mathbb{G}, \rhd\rhd)$, there can be many farsightedly consistent sets. We shall denote by $\mathbb{F}^*$ the *largest farsightedly consistent set*. Thus, if $\mathbb{F}$ is farsightedly consistent, then $\mathbb{F} \subseteq \mathbb{F}^*$. Unlike quasi-stable sets and stable sets where existence implies nonemptiness, in considering farsightedly consistent sets two critical questions arise: (i) does there exist a largest farsightedly consistent set of networks for $(\mathbb{G}, \rhd\rhd)$, and (ii) is it nonempty? Our next result provides a positive answer to both questions.

Theorem 3 (Existence, Uniqueness, and Nonemptiness of $\mathbb{F}^*$, [94]) *There exists a unique, nonempty largest farsightedly consistent set $\mathbb{F}^*$ for network formation game $(\mathbb{G}, \rhd\rhd)$. Moreover, $\mathbb{F}^*$ is externally stable; that is, if network G is not contained in $\mathbb{F}^*$, then there exists a network G' contained in $\mathbb{F}^*$ that indirectly dominates G (i. e., $G' \rhd\rhd G$).*

The method of proving existence and uniqueness is a straightforward, supernetwork rendition of Chwe's [26] method and is similar to the method introduced by Roth [103,104]. Page and Kamat [89] provide an alternative proof (to that of Chwe and of [94]) of the nonemptiness and external stability of the largest consistent set (with respect to indirect dominance). In particular, Page and Kamat modify the indirect dominance relation so as to make it transitive as well as irreflexive. They then show that the unique stable set with respect to path dominance induced by this new transitive indirect dominance relation is contained in the largest farsightedly consistent set – and in this way show that the largest farsightedly consistent set is nonempty and externally stable.

Network Formation Games with Respect to Path Dominance

In this section we consider the network formation game with respect to path dominance given by the pair

$$(\mathbb{G}, \geq_p)\ .$$

Throughout this section we will assume that the underlying primitives,

$$(\mathbb{G}, \{\succ_S\}, \{\rightarrow_S\}, >)_{S\in\Gamma(N)}\ ,$$

are such that $\mathbb{G}$ is equal to the set of homogeneous linking networks $P(P_2(N))$ or the set of homogeneous directed networks $P(N \times N)$, that the dominance relation $>$ on $\mathbb{G}$ is given by either direct dominance $\rhd$ or indirect dominance $\rhd\rhd$, and that primitives are represented by supernetwork $\mathbf{G}_\rho := \mathbf{P} \cup \mathbf{R}_\rho$ (with ρ equal to *jw*, *jn*, or *bg*) and $>$-supernetwork $\mathbf{D}_>$.

We will present three notions of stability introduced in [91] and [93] for abstract games of network formation with respect to path dominance over heterogeneous directed networks: (i) strategic basins of attraction, (ii) path dominance stable sets, and (iii) the path dominance core.

Preliminaries

Networks Without Descendants If $G_1 \geq_p G_0$ and $G_0 \geq_p G_1$, networks G_1 and G_0 are *equivalent*, written

$G_1 \equiv_p G_0$. If networks G_1 and G_0 are equivalent then either networks G_1 and G_0 coincide or G_1 and G_0 are on the same circuit (see Fig. 9 above for a picture of a circuit). If $G_1 \geq_p G_0$ but G_1 and G_0 are not equivalent (i. e., not $G_1 \equiv_p G_0$), then network G_1 is a *descendant* of network G_0 and we write

$$G_1 >_p G_0 \ .$$

Referring to Fig. 9, observe that network G_5 is a descendant of network G_1, that is, $G_5 >_p G_1$.

Network $G' \in \mathbb{G}$ *has no descendants in* $\mathbb{G}$ if for any network $G \in \mathbb{G}$

$$G \geq_p G' \quad \text{implies that} \quad G \equiv_p G' \ .$$

Thus, if G' has no descendants then $G \geq_p G'$ implies that G and G' coincide or lie on the same circuit. Note that any isolated network is by definition a network without descendants (e. g., network G_0 in Fig. 9).

In attempting to identify and characterize stable homogeneous networks, networks *without descendants* are of particular interest. Here is our main result concerning networks without descendants.

Theorem 4 (All path dominance network formation games have networks without descendants, [91,93]) *In network formation game $(\mathbb{G}, \geq_p)$ every network $G \in \mathbb{G}$ is path dominated by a network $G' \in \mathbb{G}$ without descendants (i. e., $G' \geq_p G$ and G' has no descendants).*

By Theorem 4, in network formation game $(\mathbb{G}, \geq_p)$, corresponding to any network $G \in \mathbb{G}$ there is a network $G' \in \mathbb{G}$ without descendants which is $>$-reachable from G. Thus, in any network formation game the set of networks without descendants is nonempty. Referring to Fig. 9, the set of networks without descendants is given by

$$\{G_0, G_2, G_3, G_4, G_5, G_7\} \ .$$

We shall denote by $\mathbb{Z}$ the set of networks without descendants.

Basins of Attraction Stated loosely, a basin of attraction is a set of *equivalent* networks to which the strategic network formation process represented by the game might tend and from which there is no escape. Formally, we have the following definition.

Definition 9 (Basin of Attraction, [91,93]) A set of networks $\mathbb{A} \subseteq \mathbb{G}$ is said to be a basin of attraction for $(\mathbb{G}, \geq_p)$ if

(a) the networks contained in $\mathbb{A}$ are equivalent (i. e., for all G' and G in $\mathbb{A}$, $G' \equiv_p G$) and for no set $\mathbb{A}'$ having $\mathbb{A}$ as a strict subset is this true that all the networks in $\mathbb{A}'$ are equivalent, and
(b) no network in $\mathbb{A}$ has descendants (i. e., there does not exist a network $G' \in \mathbb{G}$ such that $G' >_p G$ for some $G \in \mathbb{A}$).

$\mathbb{A}$ is a strict subset of $\mathbb{A}'$ if $\mathbb{A} \subset \mathbb{A}'$ and $\mathbb{A}' \backslash \mathbb{A} \neq \emptyset$.

As the following characterization result shows, there is a very close connection between networks without descendants and basins of attraction.

Theorem 5 (A characterization of basins of attraction, [91,93]) *Let $\mathbb{A}$ be a subset of networks in $\mathbb{G}$. The following statements are equivalent:*

(1) $\mathbb{A}$ is a basin of attraction for $(\mathbb{G}, \geq_p)$.
(2) There exists a network without descendants, $G \in \mathbb{Z}$, such that

$$\mathbb{A} = \{G' \in \mathbb{Z} : G' \equiv_p G\} \ .$$

In light of Theorem 5, we conclude that in any network formation game $(\mathbb{G}, \geq_p)$, $\mathbb{G}$ contains a *unique*, finite, disjoint collection of basins of attraction, say $\{\mathbb{A}_1, \mathbb{A}_2, \ldots, \mathbb{A}_m\}$, where for each $k = 1, 2, \ldots, m$ $(m \geq 1)$

$$\mathbb{A}_k = \mathbb{A}_G := \{G' \in \mathbb{Z} : G' \equiv_p G\}$$

for some network $G \in \mathbb{Z}$. Note that for networks G' and G in $\mathbb{Z}$ such that $G' \equiv_p G$, $\mathbb{A}_{G'} = \mathbb{A}_G$ (i. e. the basins of attraction $\mathbb{A}_{G'}$ and $\mathbb{A}_G$ coincide). Also, note that if network $G \in \mathbb{G}$ is isolated, then $G \in \mathbb{Z}$ and

$$\mathbb{A}_G := \{G' \in \mathbb{Z} : G' \equiv_p G\} = \{G\}$$

is, by definition, a basin of attraction – but a very uninteresting one.

Example 2 (Basins of Attraction) In Fig. 9 above the set of networks without descendants is given by

$$\mathbb{Z} = \{G_0, G_2, G_3, G_4, G_5, G_7\} \ .$$

Even though there are six networks without descendants, because networks G_2, G_3, G_4, and G_5 are equivalent, there are only three basins of attraction:

$$\mathbb{A}_1 = \{G_0\} \ , \quad \mathbb{A}_2 = \{G_2, G_3, G_4, G_5\} \ , \quad \text{and} \ \mathbb{A}_3 = \{G_7\} \ .$$

Moreover, because G_2, G_3, G_4, and G_5 are equivalent,

$$\mathbb{A}_{G_2} = \mathbb{A}_{G_3} = \mathbb{A}_{G_4} = \mathbb{A}_{G_5} = \{G_2, G_3, G_4, G_5\} \ .$$

□

Stable Sets with Respect to Path Dominance The formal definition of a $\geq_p$-stable set is as follows.

Definition 10 (Stable Sets with Respect to Path Dominance, [91,93]) A subset $\mathbb{V}$ of networks in $\mathbb{G}$ is said to be a stable set for $(\mathbb{G}, \geq_p)$ if

(a) (internal $\geq_p$-stability) whenever G_0 and G_1 are in $\mathbb{V}$, with $G_0 \neq G_1$, then neither $G_1 \geq_p G_0$ nor $G_0 \geq_p G_1$ hold, and
(b) (external $\geq_p$-stability) for any $G_0 \notin \mathbb{V}$ there exists $G_1 \in \mathbb{V}$ such that $G_1 \geq_p G_0$. □

In other words, a nonempty subset of networks $\mathbb{V}$ is a stable set for $(\mathbb{G}, \geq_p)$ if G_0 and G_1 are in $\mathbb{V}$, with $G_0 \neq G_1$, then G_1 is not $>$-reachable from G_0, nor is G_0 $>$-reachable from G_1, and if $G_0 \notin \mathbb{V}$, then there exists $G_1 \in \mathbb{V}$ reachable from G_0.

We now have our main results on the existence, construction, and cardinality of stable sets. These results can be viewed as variations on some classical results from graph theory applied to network formation games (e. g., see Berge [8], Chap. 2).

Theorem 6 (Stable sets: Existence, construction, and cardinality, [91,93]) *Without loss of generality assume that $(\mathbb{G}, \geq_p)$ has basins of attraction given by*

$$\{\mathbb{A}_1, \mathbb{A}_2, \ldots, \mathbb{A}_m\} ,$$

where basin of attraction $\mathbb{A}_k$ contains $|\mathbb{A}_k|$ many networks (i. e., $|\mathbb{A}_k|$ is the cardinality of $\mathbb{A}_k$). Then the following statements are true:

(1) $\mathbb{V} \subseteq \mathbb{G}$ is a stable set for $(\mathbb{G}, \geq_p)$ if and only if $\mathbb{V}$ is constructed by choosing one network from each basin of attraction, that is, if and only if $\mathbb{V}$ is of the form

$$\mathbb{V} = \{G_1, G_2, \ldots, G_m\} ,$$

where $G_k \in \mathbb{A}_k$ for $k = 1, 2, \ldots, m$.

(2) $(\mathbb{G}, \geq_p)$ possesses

$$|\mathbb{A}_1| \cdot |\mathbb{A}_2| \cdot \cdots \cdot |\mathbb{A}_m| := M$$

many stable sets and each stable set, $\mathbb{V}_q$, $q = 1, 2, \ldots, M$, has cardinality

$$|\mathbb{V}_q| = |\{\mathbb{A}_1, \mathbb{A}_2, \ldots, \mathbb{A}_m\}| = m .$$

Example 3 (Basins of Attraction and Stable Sets) Referring to Fig. 9, it follows from Theorem 6 that because

$$|\mathbb{A}_1| \cdot |\mathbb{A}_2| \cdot |\mathbb{A}_3| = 1 \cdot 4 \cdot 1 = 4 ,$$

the network formation game $(\mathbb{G}, \geq_p)$ has 4 stable sets, each with cardinality 3. By examining Fig. 9 in light of Theorem 6, we see that the stable sets for $(\mathbb{G}, \geq_p)$ are given by

$$\mathbb{V}_1 = \{G_0, G_2, G_7\} , \quad \mathbb{V}_2 = \{G_0, G_3, G_7\} ,$$
$$\mathbb{V}_3 = \{G_0, G_4, G_7\} , \quad \mathbb{V}_4 = \{G_0, G_5, G_7\} .$$

□

It should be noted that by equipping the abstract network formation game with the path dominance relation rather than the original dominance relation, we avoid the famous Lucas [78] example of a game with no stable set.

The Path Dominance Core

Definition 11 (The Path Dominance Core, [91,93]) A network $G \in \mathbb{G}$ is contained in the path dominance core $\mathbb{C}$ of network formation game $(\mathbb{G}, \geq_p)$ if and only if there does not exist a network $G' \in \mathbb{G}$, $G' \neq G$, such that $G' \geq_p G$.

Our next results give necessary and sufficient conditions for the path dominance core of a network formation game over homogeneous networks to be nonempty, as well as a recipe for constructing the path dominance core.

Theorem 7 (Path dominance core: Nonemptiness and construction, [91,93]) *Without loss of generality assume that $(\mathbb{G}, \geq_p)$ has basins of attraction given by*

$$\{\mathbb{A}_1, \mathbb{A}_2, \ldots, \mathbb{A}_m\} ,$$

where basin of attraction $\mathbb{A}_k$ contains $|\mathbb{A}_k|$ many networks. Then the following statements are true:

(1) $(\mathbb{G}, \geq_p)$ has a nonempty path dominance core if and only if there exists a basin of attraction containing a single network, that is, if and only if for some basin of attraction $\mathbb{A}_k$, $|\mathbb{A}_k| = 1$.

(2) Let

$$\{\mathbb{A}_{k_1}, \mathbb{A}_{k_2}, \ldots, \mathbb{A}_{k_n}\} \subseteq \{\mathbb{A}_1, \mathbb{A}_2, \ldots, \mathbb{A}_m\} ,$$

be the subset of basins of attraction containing all basins having cardinality 1. Then the path dominance core $\mathbb{C}$ of $(\mathbb{G}, \geq_p)$ is given by

$$\mathbb{C} = \{G_{k_1}, G_{k_2}, \ldots, G_{k_n}\} ,$$

where $G_{k_i} \in \mathbb{A}_{k_i}$, for $i = 1, 2, \ldots, n$.

If coalitional preferences over networks are based on weak preferences, that is, if coalitional preferences are given by $\{\succ_{wS}\}_{S \in \Gamma(N)}$, then the corresponding path dominance

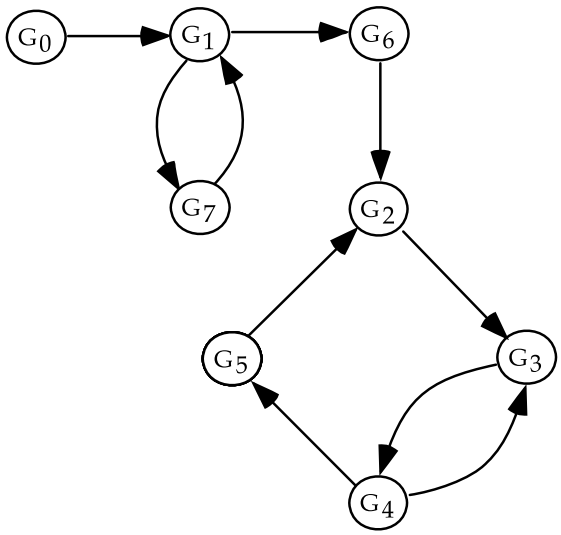

Networks and Stability, Figure 10
A different >-supernetwork $\mathbf{D}_{>}$

core – the weak path dominance core – is contained in the path dominance core based on strong preference relations.

Example 4 (Basins of Attraction and the Path Dominance Core, [91,93]*)* It follows from Theorem 7 that the path dominance core of the network formation game $(\mathbb{G}, \geq_p)$ with feasible set $\mathbb{G} = \{G_0, G_1, \ldots, G_7\}$ and path dominance relation $\geq_p$ induced by the dominance relation depicted in Fig. 9 is

$$\mathbb{C} = \{G_0, G_7\} \ .$$

Suppose that the $>$-supernetwork corresponding to the direct dominance relation $>$ on $\mathbb{G} = \{G_0, G_1, \ldots, G_7\}$ is instead depicted by Fig. 10.

Now the network formation game $(\mathbb{G}, \geq_p)$ has 3 circuits and 1 basin of attraction, $\mathbb{A} = \{G_2, G_3, G_4, G_5\}$. Because $|\mathbb{A}_1| = 4$, by Theorem 7 the path dominance core of $(\mathbb{G}, \geq_p)$ is empty. By Theorem 6, $(\mathbb{G}, \geq_p)$ has 4 stable sets each containing 1 network. These stable sets are given by

$$\mathbb{V}_1 = \{G_2\} \ , \mathbb{V}_2 = \{G_3\} \ , \mathbb{V}_3 = \{G_4\} \ , \mathbb{V}_4 = \{G_5\} \ .$$

□

The Path Dominance Core and Constrained Pareto Efficiency We say that a network $G \in \mathbb{G}$ is constrained Pareto efficient for game $(\mathbb{G}, \geq_p)$ if and only if there does not exist another network $G' \in \mathbb{G}$ such that (i) some coalition S can change network G to network G' (that is, $G \rightarrow_S G'$ for some coalition $S \in \Gamma(N)$) and (ii) G' is preferred by all players (that is, $G \prec_i G'$ for *all* players $i \in N$). Letting $\mathbb{E}$ denote the set of all constrained Pareto efficient networks, it is easy to see that the path dominance core $\mathbb{C}$ of network formation game $(\mathbb{G}, \geq_p)$ is a subset of $\mathbb{E}$, that is, $\mathbb{C} \subseteq \mathbb{E}$.

Under the classical notion of Pareto efficiency, a network G is said to be Pareto efficient if and only if there does not exists another network G' such that $G \prec_i G'$ for *all* players $i \in N$, regardless of whether or not some coalition S can change network G to network G'. Letting $\mathbb{PE}$ denote the set of all classically Pareto efficient networks, it is easy to see that $\mathbb{PE} \subseteq \mathbb{E}$. Note, however, that if under the rules of network formation, any network G can be changed to any other network G' via the actions of *some* coalition S, then the notions of constrained Pareto efficiency and classical Pareto efficiency are equivalent. Thus, if the collection of coalitional effectiveness relations $\{\rightarrow_S\}_{S \in \Gamma(N)}$ on $\mathbb{G}$ is complete, that is, if for any pair of networks G and G' in $\mathbb{G}$, $G \rightarrow_S G'$ for some coalition $S \in \Gamma(N)$, then $\mathbb{PE} = \mathbb{E}$, and we have $\mathbb{C} \subseteq \mathbb{PE} = \mathbb{E}$.

Strong Stability, Pairwise Stability, Nash Stability, and Farsighted Consistency

In this section we continue our discussion of the network formation game with respect to path dominance given by the pair

$$(\mathbb{G}, \geq_p) \ .$$

Throughout this section we will continue to assume that the underlying primitives,

$$(\mathbb{G}, \{\succ_S\}, \{\rightarrow_S\}, >)_{S \in \Gamma(N)} \ ,$$

are such that $\mathbb{G}$ is equal to the set of homogeneous linking networks $P(P_2(N))$ or the set of homogeneous directed networks $P(N \times N)$, that the dominance relation $>$ on $\mathbb{G}$ is given by either direct dominance $\rhd$ or indirect dominance $\rhd\rhd$, and that primitives are represented by supernetwork $\mathbf{G}_\rho := \mathbf{P} \cup \mathbf{R}_\rho$ (with ρ equal to jw, jn, or bg) and $>$-supernetwork $\mathbf{D}_{>}$.

We will complete our unification of stability results for network formation games by concluding that, depending on how we further specialize the primitives underlying the game $(\mathbb{G}, \geq_p)$, the path dominance core is equal to the set of pairwise stable networks [64], the set of strongly stable networks [39,61], or the set of Nash networks [5]. We also present results on the relationships between basins of attraction, the path dominance core, and the largest farsightedly consistent set [26]. All of these results follow immediately from more general results in Page and Wooders [91,93] where the notions of strong stability, pairwise

stability, Nash stability, and farsighted consistency are all extended to heterogeneous directed networks.

Strongly Stable Homogeneous Networks

We begin with a formal definition of strong stability based on that of Jackson–van den Nouweland [61].

Definition 12 (Strong Stability, [91,93]) Assume that $\mathbb{G}$ is equal to the set of all homogeneous linking networks, $P(P_2(N))$, and that the Jackson–van den Nouweland rules are in force. Network $G \in \mathbb{G}$ is said to be strongly stable in $(\mathbb{G}, \geq_p)$ if for all $G' \in \mathbb{G}$ and $S \in \Gamma(N)$, $G \rightarrow_S G'$ implies that $G \not\prec_S G'$. □

Thus, a network is strongly stable if whenever a coalition has the power to change the network to another network, the coalition will be deterred from doing so because the change is not preferred by the coalition. If coalitional preferences are strong, the change 'not being preferred' means that the change will not make *all* members of the coalition better off. If coalitional preferences are weak (i. e., based on weak preferences), the change 'not being preferred' means that the change will either make *no* members better off or will make some members better off and some members worse off. Note that under our definition of strong stability a network $G \in \mathbb{G}$ that cannot be changed to another network by any coalition is strongly stable.

If (i) coalitional preferences are weak (i. e., $\{\succ_{wS}\}_{S \in \Gamma(N)}$), and (ii) coalitional effectiveness relations are determined by Jackson–van den Nouweland rules, then the definition of strong stability above is exactly that of Jackson–van den Nouweland. As it stands, our definition is closely related to that given by Dutta–Mutuswami [39].

We now have our main result on the path dominance core and strong stability. Denote the set of strongly stable networks by $\mathbb{SS}$.

Theorem 8 (The path dominance core and strong stability, [91,93]) *Assume that $\mathbb{G}$ is equal to the set of all homogeneous linking networks, $P(P_2(N))$, and that the Jackson–van den Nouweland rules are in force.*

(1) If the path dominance core $\mathbb{C}$ of $(\mathbb{G}, \geq_p)$ is nonempty, then $\mathbb{SS}$ is nonempty and $\mathbb{C} \subseteq \mathbb{SS}$.
(2) If the dominance relation $>$ underlying $\geq_p$ is a direct dominance relation $\rhd$, then $\mathbb{C} = \mathbb{SS}$ and $\mathbb{SS}$ is nonempty if and only if there exists a basin of attraction containing a single network.

Note that the set of strongly stable homogeneous linking networks is contained in the set of constrained Pareto efficient homogeneous linking networks. Thus, $\mathbb{C} \subseteq \mathbb{SS} \subseteq \mathbb{E}$.

Pairwise Stable Networks

The following definition of pairwise stability is a translation of the Jackson–Wolinsky definition [64].

Definition 13 (Pairwise Stability, [91,93]) Assume that $\mathbb{G}$ is equal to the set of all homogeneous linking networks, $P(P_2(N))$, and that the Jackson–Wolinsky rules are in force. Network $G \in \mathbb{G}$ is said to be pairwise stable in $(\mathbb{G}, \geq_p)$ if for all $\{i, i'\} \in P_2(N)$,

(1) $G \rightarrow_{\{i,i'\}} G \cup \{i, i'\}$ implies that $G \not\prec_{\{i,i'\}} G \cup \{i, i'\}$;
(2)(a) $G \rightarrow_{\{i\}} G \setminus \{i, i'\}$ implies that $G \not\prec_{\{i\}} G \setminus \{i, i'\}$, and
(2)(b) $G \rightarrow_{\{i'\}} G \setminus \{i, i'\}$ implies that $G \not\prec_{\{i'\}} G \setminus \{i, i'\}$.

□

Thus, a homogeneous linking network is pairwise stable if there is no incentive for any pair of players to add a link to the existing network and there is no incentive for any player who is party to a link in the existing network to dissolve or remove the link. Note that under our definition of pairwise stability a network $G \in \mathbb{G}$ that cannot be changed to another network by any coalition, or can only be changed by coalitions of size greater than 2, is pairwise stable.

Let $\mathbb{PS}$ denote the set of pairwise stable networks. It follows from the definitions of strong stability and pairwise stability that $\mathbb{SS} \subseteq \mathbb{PS}$. Moreover, if the full set of Jackson–Wolinsky rules are in force, then $\mathbb{SS} = \mathbb{PS}$. Jackson–van den Nouweland [61] provide two examples of the potential for strong stability to refine pairwise stability (i. e., two examples where $\mathbb{SS}$ is a strict subset of $\mathbb{PS}$). However, *under Jackson–Wolinsky rules* because network changes can occur only one link at a time and because deviations by coalitions of more than two players are not possible such refinements are not possible driving $\mathbb{SS}$ and $\mathbb{PS}$ to equality.

We now have our main result on the path dominance core and pairwise stability.

Theorem 9 (The path dominance core and pairwise stability, [91,93]) *Assume that $\mathbb{G}$ is equal to the set of all homogeneous linking networks, $P(P_2(N))$, and that the Jackson–Wolinsky rules are in force.*

(1) If the path dominance core $\mathbb{C}$ of $(\mathbb{G}, \geq_p)$ is nonempty, then $\mathbb{PS}$ is nonempty and $\mathbb{C} \subseteq \mathbb{PS}$.

(2) *If the dominance relation* $>$ *underlying* $\geq_p$ *is a direct dominance relation* $\rhd$, *then* $\mathbb{C} = \mathbb{PS}$ *and* $\mathbb{PS}$ *is nonempty if and only if there exists a basin of attraction containing a single network.*

Theorem 9 can be viewed as an extension of a result due to Jackson and Watts [62] on the existence of pairwise stable homogeneous linking networks for network formation games induced by Jackson–Wolinsky rules. In particular, Jackson and Watts [62] show that for this particular class of Jackson–Wolinsky network formation games, if there does not exist a closed cycle of networks, then there exists a pairwise stable network. Our notion of a strategic basin of attraction containing *multiple* networks corresponds to their notion of a closed cycle of networks. Thus, stated in our terminology, Jackson and Watts show that for this class of network formation games, if there does not exist a basin of attraction containing multiple networks, then there exists a pairwise stable network. Following our approach, by part 2 of Theorem 9 the existence of *at least one* strategic basin containing a single network is both *necessary and sufficient* for the existence of a pairwise stable network.

Nash Networks

The following definition of Nash networks is a variation on the definition from Bala and Goyal [5].

Definition 14 (Nash Networks, [91,93]) Assume that $\mathbb{G}$ is equal to the set of all homogeneous directed networks, $P(N \times N)$, and that the Bala–Goyal rules are in force. Network $G \in \mathbb{G}$ is said to be a Nash network in $(\mathbb{G}, \geq_p)$ if for all $G' \in \mathbb{G}$, $G \rightarrow_{\{i'\}} G'$ implies that $G \not\prec_{\{i'\}} G'$. □

Thus, a homogeneous directed network is Nash if whenever an individual player has the power to change the network to another network, the player will have no incentive to do so. We shall denote by $\mathbb{NE}$ the set of Nash networks. Note that under our definition any network that cannot be changed to another network by a coalition of size 1 is a Nash network. Finally, note that the set of strongly stable networks $\mathbb{SS}$ is contained in the set of Nash networks $\mathbb{NE}$.

We now have our main result on the path dominance core and Nash stability.

Theorem 10 (The path dominance core and Nash networks, [91,93]) *Assume that* $\mathbb{G}$ *is equal to the set of all homogeneous directed networks,* $P(N \times N)$, *and that the Bala–Goyal rules are in force.*

(1) *If the path dominance core* $\mathbb{C}$ *of* $(\mathbb{G}, \geq_p)$ *is nonempty, then* $\mathbb{NE}$ *is nonempty and* $\mathbb{C} \subseteq \mathbb{NE}$.

(2) *If the dominance relation* $>$ *underlying* $\geq_p$ *is a direct dominance relation* $\rhd$, *then* $\mathbb{C} = \mathbb{NE}$ *and* $\mathbb{NE}$ *is nonempty if and only if there exists a basin of attraction containing a single network.*

We close this section by noting that under the Bala–Goyal rules the set of Nash networks $\mathbb{NE}$ is contained in the set of constrained Pareto efficient networks $\mathbb{E}$. If in addition the dominance relation is direct, then $\mathbb{C} = \mathbb{SS} = \mathbb{NE} \subseteq \mathbb{E}$.

Farsightedly Consistent Networks

Our final result summarizes the relationships between basins of attraction, the path dominance core, and the largest farsightedly consistent set.

Theorem 11 (Basins of attraction, the path dominance core, and the largest consistent set, [91,93]) *Assume that (i)* $\mathbb{G}$ *is equal to the set of homogeneous linking networks* $P(P_2(N))$ *and that the Jackson–Wolinsky rules or the Jackson–van den Nouweland rules are in force; or (ii) that* $\mathbb{G}$ *is equal to the set of homogeneous directed networks* $P(N \times N)$) *and that the Bala–Goyal rules are in force.*

Given network formation game $(\mathbb{G}, \geq_p)$, *where path dominance is induced by an indirect dominance relation* $\rhd\rhd$, *assume without loss of generality that* $(\mathbb{G}, \geq_p)$ *has nonempty largest consistent set given by* $\mathbb{F}^*$ *and basins of attraction given by*

$$\{\mathbb{A}_1, \mathbb{A}_2, \ldots, \mathbb{A}_m\}.$$

Then the following statements are true:

(1) *Each basin of attraction* $\mathbb{A}_k$, $k = 1, 2, \ldots, m$, *has a nonempty intersection with the largest consistent set* $\mathbb{F}^*$, *that is*

$$\mathbb{F}^* \cap \mathbb{A}_k \neq \emptyset, \quad \text{for } k = 1, 2, \ldots, m.$$

(2) *If* $(\mathbb{G}, \geq_p)$ *has a nonempty path dominance core* $\mathbb{C}$, *then*

$$\mathbb{C} \subseteq \mathbb{F}^*.$$

Singleton Basins of Attraction

In the abstract games, $(\mathbb{G}, \geq_p)$, that we have considered, the key condition guaranteeing nonemptiness of the path dominance core is the existence of basins of attraction containing a single network. Question: Are there classes of games for which this is true? *In general*, if the irreflexive dominance relation $>$ inducing path dominance $\geq_p$

is transitive, then the $>$-supernetwork $\mathbf{D}_>$ is without circuits, and therefore all basins of attraction for the game $(\mathbb{G}, \geq_p)$ contain a single network. Unfortunately, if the dominance relation is given by direct or indirect dominance, then transitivity fails to hold in general. In the next two sections we identify several classes of network formation games having singleton basins.

Network Formation Games and Potential Functions

Assume that $\mathbb{G}$ is equal to the set of homogeneous directed networks, $P(N \times N)$, and that the Bala–Goyal rules are in force, so that primitives are represented by supernetwork $\mathbf{G}_{\text{bg}} := \mathbf{P} \cup \mathbf{R}_{\text{bg}}$. In addition assume that player preferences over $P(N \times N)$ are specified via payoff functions $\{v_i(\cdot)\}_{i \in N}$ and that the dominance relation $>$ over $P(N \times N)$ is given by direct dominance $\rhd$. Thus, $G' \rhd G$ if and only if for some player $i' \in N$, $G \underset{\{i'\}}{\longrightarrow} G'$ and $v_{i'}(G') > v_{i'}(G)$.

We say that the noncooperative network formation game $(\mathbb{G}, \geq_p)$ is a potential game if there exists a function

$$P(\cdot) \colon \mathbb{G} \to R$$

such that for all G and G' with $G \to_{\{i'\}} G'$ for some player i',

$$v_{i'}(G') > v_{i'}(G) \quad \text{if and only if} \quad P(G') > P(G) \,.$$

It is easy to see that any noncooperative network formation game $(\mathbb{G}, \geq_p)$ possessing a potential function (i. e., a potential game) has no circuits, and thus possesses strategic basins of attraction each consisting of a single network. Thus, we can conclude from our Theorem 8 that any noncooperative network formation game possessing a potential function has a nonempty path dominance core. In addition, we know from our Theorem 10 that in this example the path dominance core $\mathbb{C}$ is equal to the set of Nash networks $\mathbb{NE}$.

As has been shown by Monderer and Shapley [85], potential games are closely related to congestion games introduced by Rosenthal [102]. Page and Wooders [92] introduce a club network formation game which is a variant of the noncooperative network formation game described above – but for a class of heterogenous directed networks – and using methods similar to those introduced by Hollard [57], show that this game possesses a potential function. Prior papers studying potential games in the context of linking networks include Qin [97], Slikker, Dutta, van den Nouweland, and Tijs [112] and Slikker and van den Nouweland [114]. These papers have focused on providing the strategic underpinnings of the Myerson value (Myerson [88] and Aumann and Myerson [4]).

Jackson–Wolinsky Network Formation Games

Assume that $\mathbb{G}$ is equal to the set of all homogeneous linking networks, $P(P_2(N))$, and that the Jackson–van den Nouweland rules are in force, so that rules are represented by rules supernetwork $\mathbf{R}_{\text{jn}}$. In addition assume that player preferences over $P(P_2(N))$ are weak and therefore that coalitional preferences, $\{\succ_{wS}\}_{S \in \Gamma(N)}$, are weak. Finally, assume that the dominance relation $>$ on $P(P_2(N))$ is given by direct dominance – but because coalitional preferences are weak, direct dominance is weak, denoted by $\rhd_w$.

In the Jackson–Wolinsky network formation game coalitional preferences are specified by player payoff functions, $\{v_i(\cdot)\}_{i \in N}$, and player payoff functions are in turn specified by a network value function

$$v(\cdot) \colon \mathbb{G} \to R$$

together with an allocation rule, $Y(G, v) = (Y_i(G, v))_{i \in N} \in R^{|N|}$ satisfying

$$\sum_{i \in N} Y_i(G, v) = v(G) \,.$$

Thus in the Jackson–Wolinsky game, each player's payoff function $v_i(\cdot)$ is given by $Y_i(\cdot, v)$ where $v(\cdot)$ is the network value function. The basic idea here is that given network G, $v(G)$ is the total value generated by network G and $Y_i(G, v)$ is value allocated to player i. Translating Jackson–Wolinsky into our abstract game model, if $G' \rhd_w G$ then one of the following is true:

(i) $G \to_{\{i,i'\}} G'$ where $G' = G \cup \{i, i'\}$ (a link between players i and i' is added) and $Y_i(G', v) \geq Y_i(G, v)$ and $Y_{i'}(G', v) \geq Y_{i'}(G, v)$ with strict inequality for at least on of the players;

(ii) $G \to_{\{i\}} G'$ where $G' = G \backslash \{i, i'\}$ (a link between players i and i' is subtracted) and $Y_i(G', v) > Y_i(G, v)$;

(iii) $G \to_{\{i'\}} G'$ where $G' = G \backslash \{i, i'\}$ (a link between players i and i' is subtracted) and $Y_{i'}(G', v) > Y_{i'}(G, v)$.

Each homogeneous linking network G can be partitioned into a collection of subnetworks called components as follows. Let $H \subseteq G$ be any subnetwork of G and define

$$N_H := \{i \in N \colon \exists i' \in N \quad \text{such that} \quad \{i, i'\} \in H\} \,.$$

Subnetwork $H \subseteq G$ is said to be a component if

- i and i' are in N_H, then there is a path between i and i'; and
- if $i \in N_H$, $H \subseteq G$, and $\{i, i'\} \in G$ then $\{i, i'\} \in H$.

Let $C(G)$ denote the set of all components of G.

If the allocation rule $(Y_i(G, v))_{i \in N}$ is *egalitarian* (see [64]); that is, if

$$v_i(G) = Y_i(G, v) = \frac{v(G)}{|N|} ,$$

then any network $G^* \in \arg\max_{G \in \mathbb{G}} v(G)$ is pairwise stable and hence by part (2) of Theorem 9, $(\mathbb{G}, \geq_p)$ has a basin of attraction containing a single network.

Alternatively, if the allocation rule $(Y_i(G, v))_{i \in N}$ is *componentwise egalitarian* (see [64]); that is, if

$$v_i(G) = Y_i(G, v) = \frac{v(H)}{|N_H|} , \quad \text{for } H \in C(G) \quad \text{and} \quad i \in N_H ,$$

then there exists a pairwise stable network (see Jackson [59]), and hence by part (2) of Theorem 9, $(\mathbb{G}, \geq_p)$ has a basin of attraction containing a single network.

Finally, Jackson [59] has shown that if the allocation rule is given by the Myerson value (see [88] and [4]); that is if

$$v_i(G) = Y_i(G, v) = \sum_{S \subset N \setminus \{i\}} \left(v(G_{S \cup \{i\}}) - v(G_S) \left(\frac{|S|!(|N| - |S| - 1)!}{|N|!} \right) \right),$$

where

$$G_S := \{\{i, i'\} \in G \colon i \in S \text{ and } i' \in S\} ,$$

then $(\mathbb{G}, \geq_p)$ has a basin of attraction containing a single network, and hence by part 2 of Theorem 9 $(\mathbb{G}, \geq_p)$ has at least one pairwise stable network.

Future Directions

There are many possible directions for future research on the topic of networks and stability. Here we only mention a few that, from the perspective of this entry, seem especially promising. There are a number of potential questions to be addressed concerning the path dominance core with direct or indirect dominance. For example, what is the relationship, if any, between basins of attraction and the path dominance core and partnered (or separating) collections of coalitions, as in for example Maschler and Peleg [81], Maschler, Peleg and Shapley [82], Reny and Wooders [99] and Page and Wooders [90]? Or what is relationship between basins of attraction and the path dominance core and the inner core, as in for example Qin [95,96]?

One of the most pressing issues in our view is strategic network dynamics. Future research will address the following open question: Given the rules of network formation, the preferences of individuals, the strategic behavior of coalitions, and the trembles of nature, what network dynamics are likely to emerge and persist?

Another direction is large networks, where there are many, but still a finite number of nodes or networks with a continuum of nodes. As in the framework of cooperative games (cf., Kovalenkov and Wooders [72] and Wooders, [125,127] for cores of transferable utility and nontransferable utility games) does it hold that some notion of the approximate path dominance core is nonempty if the numbers of players is sufficiently large? Do networks tend towards having some property analogous to the equal treatment property (as in, for example, Debreu and Scarf [31] and Green [52] for exchange economies, [72] or Wooders [126], for cooperative games or Gravel and Thoron [51] for local public goods economies, or Jackson and Watts [63] for a repeated game approach to a matching model). Then there is the problem of characterizing strategic behavior in large networks (as in, for example, Kalai [65] or Wooders, Cartwright and Selten [128]). Under what conditions and to what extent might Kalai's "ex-post stability" or Wooders, Cartwright and Selten's social conformity continue to hold in strategic network formation?

Acknowledgments

This paper was begun while Page and Wooders were visiting CERMSEM at the University of Paris 1 in June and October of 2007. The authors thank CERMSEM and Paris 1 for their hospitality. URLs: http://mypage.iu.edu/~fpage, http://www.myrnawooders.com.

Bibliography

1. Allouch N, Wooders M (2007) Price taking equilibrium in economies with multiple memberships in clubs and unbounded club sizes. J Econ Theory. doi:10.1016/j.jet.2007.07.06
2. Arnold T, Wooders M (2006) Club formation with coordination. University of Warwick Working Paper 640
3. Aumann RJ (1964) Markets with a continuum of traders. Econometrica 32:39–50
4. Aumann RJ, Myerson RB (1988) Endogenous formation of links between players and coalitions: An application of the

Shapley value. In: Roth A (ed) The Shapley value. Cambridge University Press, Cambridge, pp 175–191
5. Bala V, Goyal S (2000) A noncooperative model of network formation. Econometrica 68:1181–1229
6. Banerjee S, Konishi H, Sonmez T (2001) Core in a simple coalition formation game. Soc Choice Welf 18:135–158
7. Belleflamme P, Bloch F (2004) Market sharing agreements and collusive networks. Int Econ Rev 45:387–411
8. Berge C (2001) The theory of graphs. Dover, Mineola (reprint of the translated French edition published by Dunod, Paris, 1958)
9. Bhattacharya A (2005) Stable and efficient networks with farsighted players: The largest consistent set. Typescript, University of York
10. Bloch F (1995) Endogenous structures of association in oligopolies. Rand J Econ 26:537–556
11. Bloch F (2005) Group and network formation in industrial organization: A survey. In: Demange G, Wooders M (eds) Group formation in economics: Networks, clubs, and coalitions. Cambridge University Press, Cambridge, pp 335–353
12. Bloch F, Genicot G, Ray D (2008) Informal insurance in social networks. J Econ Theory. doi:10.1016/j.jet.2008.01.008
13. Blume L (1993) The statistical mechanics of strategic interaction. Games Econ Behav 5:387–424
14. Bogomolnaia A, Jackson MO (2002) The stability of hedonic coalition structures. Games Econ Behav 38:201–230
15. Bollobas B (1998) Modern graph theory. Springer, New York
16. Boorman SA (1975) A combinatorial optimization model for transmission of job information through contact networks. Bell J Econ 6:216–249
17. Bramoulle Y, Kranton R (2007) Public goods in networks. J Econ Theory 135:478–494
18. Bramoulle Y, Kranton R (2007) Risk-sharing networks. J Econ Behav Organ 64:275–294
19. Calvo-Armengol A (2004) Job contact networks. J Econ Theory 115:191–206
20. Calvo-Armengol A, Ballester C, Zenou Y (2006) Who's who in networks. Wanted: The key player. Econometrica 75:1403–1418
21. Calvo-Armengol A, Jackson MO (2004) The effects of social networks on employment and inequality. Am Econ Rev 94:426–454
22. Calvo-Armengol A, Jackson MO (2007) Social networks in labor markets: Wage and employment dynamics and inequality. J Econ Theory 132:27–46
23. Casella A, Rauch J (2002) Anonymous market and group ties in international trade. J Int Econ 58:19–47
24. Casella A, Rauch J (2003) Overcoming informational barriers in international resource allocations: Prices and ties. Econ J 113:21–42
25. Chvatal V, Lovasz L (1972) Every directed graph has a semikernel. In: Hypergraph Seminar, Lecture Notes in Mathematics, vol 411. Springer, Berlin
26. Chwe M (1994) Farsighted coalitional stability. J Econ Theory 63:299–325
27. Chwe M (2000) Communication and coordination in social networks. Rev Econ Stud 67:1–16
28. Corominas-Bosch M (2004) Bargaining in a network of buyers and sellers. J Econ Theory 115:35–77
29. Currarini S (2007) Group stability of hierarchies in games with spillovers. Math Soc Sci 54:187–202
30. Currarini S, Morelli M (2000) Network formation with sequential demands. Rev Econ Des 5:229–249
31. Debreu G, Scarf H (1963) A limit theorem on the core of an economy. Int Econ Rev 4:235–246
32. Demange G, Henreit D (1991) Sustainable oligopolies. J Econ Theory 54:417–428
33. Demange G (1994) Intermediate preferences and stable coalition structures. J Math Econ 23:45–48
34. Demange G (2004) On group stability and hierarchies in networks. J Political Econ 112:754–778
35. Deroian F, Gannon F (2005) Quality improving alliances in differentiated oligopoly. Int J Ind Organ 24:629–637
36. Diamantoudi E, Xue L (2003) Farsighted stability in hedonic games. Soc Choice Welf 21:39–61
37. Durlauf S (1997) Statistical mechanics approaches to socioeconomic behavior. In: Arthur WB, Durlauf S, Lane DA (eds) The economy as an evolving complex system II. Addison-Wesley, Reading, pp 81–104
38. Dutta B, Ghosal S, Ray D (2005) Farsighted network formation. J Econ Theory 122:143–164
39. Dutta B, Mutuswami S (1997) Stable networks. J Econ Theory 76:322–344
40. Even-Dar E, Kearns M, Suri S (2007) A network formation game for bipartite exchange economies. Computer and Information Science typescript, University of Pennsylvania
41. Furusawa T, Konishi H (2007) Free trade networks. J Int Econ 72:310–335
42. Galeana-Sanchez H, Xueliang L (1998) Semikernels and (k, l)-kernels in digraphs. SIAM J Discret Math 11:340–346
43. Galeotti A, Moraga-Gonzalez JL (2007) Segmentation, advertising and prices. Int J Ind Organ. doi:10.1016/j.ijindorg.2007.11.002
44. Gillies DB (1959) Solutions to general non-zero-sum games. In: Tucker AW, Luce RD (eds) Contributions to the Theory of Games, vol 4. Princeton University Press, Princeton, pp 47–85
45. Goyal S (2005) Learning in networks. In: Demange G, Wooders M (eds) Group formation in economics: Networks, clubs, and coalitions. Cambridge University Press, Cambridge, pp 122–167
46. Goyal S (2007) Connections: An introduction to the economics of networks. Princeton University Press, Princeton
47. Goyal S, Joshi S (2003) Networks of collaboration in oligopoly. Games Econ Behav 43:57–85
48. Goyal S, Joshi S (2006) Bilateralism and free trade. Int Econ Rev 47:749–778
49. Goyal S, Moraga-Gonzalez JL (2001) R&D networks. Rand J Econ 32:686–707
50. Granovetter M (1973) The strength of weak ties. Am J Sociol 78:1360–1380
51. Gravel N, Thoron S (2007) Does endogenous formation of jurisdictions lead to wealth stratification? J Econ Theory 132:569–583
52. Green J (1972) On the inequitable nature of core allocations. J Econ Theory 4:132–143
53. Guilbaud GT (1949) La theorie des jeux. Economie. Appliquee 2:18
54. Harsanyi JC (1974) An equilibrium-point interpretation of stable sets and a proposed alternative definition. Manag Sci 20:1472–1495
55. Herings PJ-J, Mauleon A, Vannetelbosch V (2006) Farsightedly stable networks. Meteor Research Memorandum RM/06/041

56. Hojman D, Szeidl A (2006) Endogenous networks, social games and evolution. Games Econ Behav 55:112–130
57. Hollard G (2000) On the existence of a pure strategy equilibrium in group formation games. Econ Lett 66:283–287
58. Inarra E, Kuipers J, Olaizola N (2005) Absorbing and generalized stable sets. Soc Choice Welf 24:433–437
59. Jackson MO (2003) The stability and efficiency of economic and social networks. In: Dutta B, Jackson MO (eds) Networks and groups: Models of strategic formation. Springer, Heidelberg, pp 99–141
60. Jackson MO (2005) A survey of models of network formation: Stability and efficiency. In: Demange G, Wooders M (eds) Group formation in economics: Networks, clubs, and coalitions. Cambridge University Press, Cambridge, pp 11–57
61. Jackson MO, van den Nouweland A (2005) Strongly stable networks. Games Econ Behav 51:420–444
62. Jackson MO, Watts A (2002) The evolution of social and economic networks. J Econ Theory 106:265–295
63. Jackson MO, Watts A (2008) Social games: Matching and the play of finitely repeated games. Games Econ Behav. doi:10.1016/j.geb.2008.02.004
64. Jackson MO, Wolinsky A (1996) A strategic model of social and economic networks. J Econ Theory 71:44–74
65. Kalai E (2004) Large robust games. Econometrica 72:1631–1665
66. Kalai E, Pazner A, Schmeidler D (1976) Collective choice correspondences as admissible outcomes of social bargaining processes. Econometrica 44:233–240
67. Kalai E, Schmeidler D (1977) An admissible set occurring in various bargaining situations. J Econ Theory 14:402–411
68. Kirman A (1983) Communication in markets: A suggested approach. Econ Lett 12:101–108
69. Kirman A, Herreiner D, Weisbuch G (2000) Market organization and trading relationships. Econ J 110:411–436
70. Konishi H, Le Breton M, Weber S (1998) Equilibrium in a finite local public goods economy. J Econ Theory 79:224–244
71. Konishi H, Ray D (2003) Coalition formation as a dynamic process. J Econ Theory 110:1–41
72. Kovalenkov A, Wooders M (2001) Epsilon cores of games with limited side payments: Nonemptiness and equal treatment. Games Econ Behav 36:193–218
73. Kovalenkov A, Wooders M (2003) Approximate cores of games and economies with clubs. J Econ Theory 110:87–120
74. Kranton R, Minehart D (2000) Networks versus vertical integration. RAND J Econ 31:570–601
75. Kranton R, Minehart D (2001) A theory of buyer-seller networks. Am Econ Rev 91:485–508
76. Li S (1992) Far-sighted strong equilibrium and oligopoly. Econ Lett 40:39–44
77. Li S (1993) Stability of voting games. Soc Choice Welf 10:51–56
78. Lucas WF (1968) A game with no solution. Bull Am Math Soc 74:237–239
79. Luo X (2001) General systems and φ-stable sets – a formal analysis of socioeconomic environments. J Math Econ 36:95–109
80. Mariotti M, Xue L (2002) Farsightedness in coalition formation. Typescript, University of Aarhus
81. Maschler M, Peleg B (1967) The structure of the kernel of a cooperative game. SIAM J Appl Math 15:569–604
82. Maschler M, Peleg B, Shapley LS (1971) The kernel and bargaining set for convex games. Int J Game Theory 1:73–93
83. Mauleon A, Sempere-Monerris J, Vannetelbosch V (2008) Networks of knowledge among unionized firms. Can J Econ (to appear)
84. Mauleon A, Vannetelbosch V (2004) Farsightedness and cautiousness in coalition formation games with positive spillovers. Theory Decis 56:291–324
85. Monderer D, Shapley LS (1996) Potential games. Games Econ Behav 14:124–143
86. Montgomery J (1991) Social networks and labor market outcomes: Toward an economic analysis. Am Econ Rev 81:1408–1418
87. Mutuswami S, Winter E (2002) Subscription mechanisms for network formation. J Econ Theory 106:242–264
88. Myerson RB (1977) Graphs and cooperation in games. Math Oper Res 2:225–229
89. Page FH Jr, Kamat S (2005) Farsighted stability in network formation. In: Demange G, Wooders M (eds) Group formation in economics: Networks, clubs, and coalitions. Cambridge University Press, Cambridge, pp 89–121
90. Page FH Jr, Wooders M (1996) The partnered core and the partnered competitive equilibrium. Econ Lett 52:143–152
91. Page FH Jr, Wooders M (2005) Strategic basins of attraction, the farsighted core, and network formation games. FEEM Working Paper 36.05
92. Page FH Jr, Wooders M (2007) Club networks with multiple memberships and noncooperative stability. Indiana University, Department of Economics typescript (paper presented at the Conference in Honor of Ehud Kalai, 16–18 December, 2007)
93. Page FH Jr, Wooders M (2008) Strategic basins of attraction, the path dominance core, and network formation games. Games Econ Behav. doi:10.1016/j.geb.2008.05.003
94. Page FH Jr, Wooders M, Kamat S (2005) Networks and farsighted stability. J Econ Theory 120:257–269
95. Qin C-Z (1993) A conjecture of Shapley and Shubik on competitive outcomes in the cores of NTU market games. Int J Game Theory 22:335–344
96. Qin C-Z (1994) The inner core of an N-person game. Games Econ Behav 6:431–444
97. Qin C-Z (1996) Endogenous Formations of Cooperation Structures. J Econ Theory 69:218–226
98. Rees A (1966) Information networks in labor markets. Am Econ Rev 56:218–226
99. Reny PJ, Wooders M (1996) The partnered core of a game without side payments. J Econ Theory 70:298–311
100. Richardson M (1953) Solutions of irreflexive relations. Ann Math 58:573–590
101. Rockafellar RT (1984) Network flows and monotropic optimization. Wiley, New York
102. Rosenthal RW (1973) A class of games possessing pure-strategy Nash equilibria. Int J Game Theory 2:65–67
103. Roth AE (1975) A lattice fixed-point theorem with constraints. Bull Am Math Soc 81:136–138
104. Roth AE (1977) A fixed-point approach to stability in cooperative games. In: Karamardian S (ed) Fixed points: Algorithms and applications. Academic Press, New York
105. Roughgarden T (2005) Selfish routing and the price of anarchy. MIT Press, Cambridge

106. Scarf H (1967) The core of an N-person game. Econometrica 35:50–69
107. Schwartz T (1974) Notes on the abstract theory of collective choice. Carnegie-Mellon University, School of Urban and Public Affairs typescript
108. Shapley LS, Shubik M (1969) On market games. J Econ Theory 1:9–25
109. Shenoy PP (1980) A dynamic solution concept for abstract games. J Optim Theory Appl 32:151–169
110. Shubik M (1971) The "bridge game" economy: An example of indivisibilities. J Political Econ 79:909–912
111. Skyrms B, Pemantle R (2000) A dynamic model of social network formation. Proc Nat Acad Sci 97:9340–9346
112. Slikker M, Dutta B, van den Nouweland A, Tijs S (2000) Potential maximizers and network formation. Math Soc Sci 39:55–70
113. Slikker M, van den Nouweland A (2001) Social and economic networks in cooperative game theory. Kluwer, Boston
114. Slikker M, van den Nouweland A (2002) Network formation, costs, and potential games. In: Borm P, Peters H (eds) Chapters in game theory. Kluwer, Boston, pp 223–246
115. Tardos E, Wexler T (2007) Network formation games and the potential function method. In: Nisan N, Roughgarden T, Tardos E, Vazirani V (eds) Algorithmic game theory. Cambridge University Press, Cambridge, pp 487–516
116. Tesfatsion L (1997) A trade network game with endogenous partner selection. In: Amman HM, Rustem B, Whinston AB (eds) Computational approaches to economic problems. Kluwer, Boston, pp 249–269
117. Tesfatsion L (1998) Preferential partner selection in evolutionary labor markets: A study in agent-based computational economics. In: Porto VW, Saravanan N, Waagen D, Eiben AE (eds) Evolutionary programming VII. Proceedings of the seventh annual conference on evolutionary programming. Springer, Berlin, pp 15–24
118. Topa G (2001) Social interactions, local spillovers, and unemployment. Rev Econ Stud 68:261–295
119. van Deemen AMA (1991) A note on generalized stable set. Soc Choice Welf 8:255–260
120. van den Nouweland A (2005) Models of network formation in cooperative games. In: Demange G, Wooders M (eds) Group formation in economics: Networks, clubs, and coalitions. Cambridge University Press, Cambridge, pp 58–88
121. Vega-Redondo F (2007) Complex social networks. Cambridge University Press, Cambridge
122. von Neumann J, Morgenstern O (1944) Theory of games and economic behavior. Princeton University Press, Princeton
123. Wang P, Watts A (2006) Formation of buyer-seller trade networks in a quality differentiated product market. Can J Econ 39:971–1004
124. Watts A (2001) A dynamic model of network formation. Games Econ Behav 34:331–341
125. Wooders M (1983) The epsilon core of a large replica game. J Math Econ 11:277–300
126. Wooders M (2008) Competitive markets and market games. Rev Econ Design (forthcoming)
127. Wooders M (2008) Small group effectiveness, per capita boundedness and nonemptiness of approximate cores. J Math Econ. doi:101016/j.jmateco.2007.06.006
128. Wooders M, Cartwright C, Selten R (2006) Behavioral conformity in games with many players. Games Econ Behav 57:347–360
129. Xue L (1998) Coalitional stability under perfect foresight. Econ Theory 11:603–627
130. Xue L (2000) Negotiation-proof Nash equilibrium. Int J Game Theory 29:339–357
131. Zissimos B (2005) Why are free trade agreements regional? FEEM Working Paper 67-07

Networks: Structure and Dynamics

Erzsébet Ravasz Regan
Department of Medicine, Beth Israel Deaconess Medical Center, Boston, USA

Article Outline

Glossary
Definition and Relevance
Introduction
Structural Properties of Complex Networks
Dynamics on Complex Networks
Future Directions
Bibliography

Glossary

Simple graph or network A group of N nodes (vertices) among which there exist L undirected connections (links, edges), identical in strength.

Directed graph A group of nodes among which connections are directed.

Weighted network A group of nodes among which connections are not identical in strength, but carry a weight.

Bipartite network A network with more than one type of node, in which connections only exist between different node types (the definition can be relaxed to a network were most, but not all links run between vertices of different types).

Adjacency matrix A An $N \times N$ matrix representing the network, whose elements a_{ij} are equal to 1 when there is a link from node i to j, zero otherwise.

Degree distribution $P(k)$ The probability that a node of a network, chosen uniformly at random, has degree k.

Scale-free network A network in which the tail of the degree distribution follows a power law (strictly speaking, the term scale-free implies $P(k) \sim k^{-\gamma}$, however, it is often used for networks where the tail of the distribution follows a power-law).

Degree exponent γ The power law exponent of the (tail of the) degree distribution

Scale-free model A growing network model proposed by Barabási and Albert [15]. The model builds a simple graph starting from a small connected group of nodes, to which new nodes are added one by one. These new nodes connect to m old nodes with probabilities that increase linearly with the degree of the old nodes.

Shortest path (geodesic path) The smallest collection of links that form a path through the network from one vertex to another.

Diameter D The length of the largest geodesic path in a network.

Small-world network A network in which the average shortest path length grows logarithmically (or slower) with N.

Node betweenness (betweenness centrality or load) The number of shortest paths between nodes of the network that run through a given node [62].

Edge betweennes The number of shortest paths between nodes of the network that run through a given edge.

Clustering coefficient C The fraction of connections that are realized between the neighbors of a node:

$$C_i = \frac{2\,n_i}{k_i\,(k_i - 1)}\ ,$$

where n_i denotes the number of links connecting the k_i neighbors of node i. (The average clustering coefficient is given by $\langle C\rangle = \frac{1}{N}\sum_i C_i$. An alternative global measure of clustering, also called *transitivity*, is the fraction of node triples that are linked into triangles.)

Assortativity coefficient A measure of the tendency of links to run among nodes that are similar in some respect. If the similarity is described by a scalar quantity (most often the node's degree), then the assortativity coefficient is given by

$$r = \frac{\sum_{x,y} xy\,(e_{x,y} - a_x b_y)}{\sigma_a\,\sigma_b}\ ,$$

where x (y) is the scalar at the origin (end) of a link, $e_{x,y}$ denotes the fraction of all edges in the network that go from nodes with value x to ones with value y, a_x (b_y) is the fraction of edges that start (end) at a link with values x (y), and σ_a (σ_b) is the standard deviations of the distributions of a_x (b_y) values [107].

Modularity Q The number of links between nodes within the same community minus the number expected by chance:

$$Q = \frac{1}{2L}\sum_{i=1}^{N}\sum_{j=1}^{N}(A_{ij} - P_{ij})\,\delta_{g_i,g_j}\ ,$$

where node i (j) belongs to the community g_i (g_j). P_{ij} gives the expected number of links between two nodes if the network is random with respect to communities [110]. In the simplest case, in which the null model is a random network, $P_{ij} = 2L/N^2$. A more suitable assumption is $P_{ij} = k_i\,k_j/2L$, which preserves the degree distribution of the network in question (the expected degree of node i is $\sum_j P_{ij} = k_i$) [109].

Definition and Relevance

Scientific research has had a long history of bottom-up approaches, which break the system into small or elementary constituents and map out interactions between these components. The Standard Model describing elementary particles and the four types of interactions governing our world is perhaps the most successful example. Biology has developed a very detailed description of cellular components such as the DNA molecule or the various proteins and metabolites. Furthermore, many of the interactions that govern a cells life have been investigated in great detail, but mainly in isolation: transcription of DNA, protein assembly, enzyme function, etc. Perhaps not surprisingly, the first attempts to understand complexity in physics were focused on small, simple system with complex dynamics: chaos theory. Nonetheless, large natural or social systems, like a cell, an ecosystem or the Internet are much more intuitive examples of complex systems. A meaningful description of these systems requires more than a mere account of the constituent parts: one does not understand the way the Internet works by detailing the physical characteristics of computers. Nor is the sequence of a cell's genome the final tool for understanding its behavior.

Complex systems display characteristics that are fundamentally determined by their organization, emergent phenomena created by all the interacting constituents. In many cases, if one takes a step back, avoiding the details of the interactions, a complex system as a whole is made up of an assemblage of generic elements and connections; in other words, it looks like a network. For example, a cells metabolism is maintained by a biochemical network, whose nodes are substrates and links chemical reactions [75]. Equally complex webs describe human societies, whose nodes are individuals and links represent social interactions [147], the World Wide Web (WWW) [6,28], where nodes are Web documents con-

nected by URL links, the scientific literature, whose nodes are publications and links citations [44,124,125], or language, made of words linked by various syntactic or grammatical relationships [52,61,129]. Due to the diversity and large number of the nodes and interactions, the system-level characteristics of these networks remained largely unknown and unexplored prior to the last decade. At the same time, the inability of contemporary science to address the properties of complex networks limited advances in many disciplines, including molecular biology, computer science, ecology and the social sciences. The recent availability of system-level data on the network of interactions in large numbers of systems has opened the door for interdisciplinary research in fields where the behavior of the system as a whole is a central question. Recognizing generic organizational principles and order behind diversity and apparent randomness in these different systems has certainly been a surprise along the way.

Introduction

The explosion of available data describing interaction of hundreds to millions of components in systems like the Internet or the protein interaction network is a recent development for the study of complex systems. Nonetheless, networks are not new to mathematics or the social sciences. Graph theory was born from the famous Königsberg bridge problem and its even more famous solution by Euler in 1736 [59]. The problem was simple: find a closed walk that visits each of Königsberg's seven bridges once, but only once. The trouble was, nobody could find such a walk. Euler drew a four-node graph of the pieces of land connected by the bridges and showed that graphs that have nodes with odd degrees cannot have closed self-avoiding paths. (All four pieces of land in Königsberg are connected via an odd number of bridges). Euler's emphasis on the topology of the bridge problem as key to the solution marks the starting point of two prolific subfields of mathematics: topology and graph theory.

The early 1920's witnessed the birth of social network analysis, a subfield of sociology trying to understand how social interactions organize. Data gathering methods limited the size of the networks studied, nonetheless many network measures important today were defined: degree distribution, betweenness, clustering and the small world effect.

In the 1950's two prolific hungarian mathematicians, Erdős and Rény, took the challenge of describing the structure of large social networks from social science to mathematics and formulated their famous random graph model [57,58]. Their key innovation was a statistical approach to graph theory: their theorems were proved on the set of all networks generated by the rules of their model. Their random network rules were simple: take N nodes and connect each pair with probability p. They found that above a certain threshold probability the ensemble of random graphs undergoes a percolation-type phase transition from a graph made of small disjoint subgraphs to one with a giant component comparable with the system size. Random graphs have a Poissonian degree distribution: they are homogeneous networks with a well-defined average degree. Erdős and Rényi pointed out that random networks are "small worlds": their diameter scales with the logarithm of system size (node number), quite different from the power-law relationship that holds for regular lattices.

The Erdős–Rényi model guided our thinking about complex networks until the end of the 1990's, when two seminal papers triggered a very rapid growth of the field which today is called complex networks research. The first paper, published by Watts and Strogatz in 1998 [148], showed that natural systems such as the neural network of the *C. Elegans* worm, the power grid and the network of movie actors connected by feature films have a topology somewhere between regular lattices and random graphs. These networks have large clustering coefficients, but also the small world property. The small-world model presented in the paper is the first important step away from the random world of the Erdős–Rényi graph. The second paper, published one year later by Barabási and Albert, showed that the degree distributions of the movie actor network, the WWW and the power grid are not Poissonian: they have a power-law tail, inspiring the term scale-free networks [15]. The list of power-law tailed degree distributions measured on natural and man-made systems is still growing, with degree exponents that rarely fall outside the (2,3) interval. Power law degree distributions tell us that most real networks are highly heterogeneous: the majority of nodes have very small degree, but a few hubs with degrees orders of magnitude larger than the average also exist, along with nodes with degrees of all scales in between. A good example is the US airport network connected by direct flights: Chicago and Atlanta at the high end of the degree scale, the numerous regional airports at the low end.

Naturally, the sudden explosion of data and tools to explore them brought us closer to the heart of questions fundamental to understanding complex systems: what drives their organization, what defines their emergent properties. What do these networks do, what is their function? This question is natural to biology. A metabolic network has a well-defined job in the living cell: it fuels the cell with

energy, nutrients and building blocks, and it does so adaptively and robustly. Man-made complex networks, such as the Internet or WWW, the power grid or the network of synonyms in a language also perform well-defined functions. In contrast with the delicate and well-thought-out internal structure of a computer, these systems were not designed from scratch with their function in mind: they evolved and emerged naturally. Also in contrast with the fragility of a computer (pull out a random element and it stops working), complex systems in biology and technology have a remarkable resilience to random node failure: the internet does not die when a few routers go down, in fact this probably is its normal mode of operation.

Questions about the characteristics mandated by function, common to all these systems, along with questions about the selection principles that shape these structures, are at the heart of our quest to understand complex function. These questions are typically formulated in terms of the dynamics natural to the network in question: metabolic flux driven by enzymes, packet traffic on the internet, web browsing, electric current flow on the power grid, the spread of HIV on the sexual contact network, etc. Understanding how structure affects dynamics and vice versa is in the spotlight of complex networks research today.

Structural Properties of Complex Networks

Let us look at the metabolism of a cell as an example that highlights the increasingly detailed ways one can pose the question: what is the large-scale structure of cellular metabolism? At first glance, the metabolic network is a *simple graph* in which metabolites (the nodes) are connected by chemical reactions. This representation tells us whether the network is homogeneous in degree or has hubs, whether it is a "small world", whether it has a community structure. A more detailed description of the system takes into account the direction of chemical reactions, since a large number of them are not reversible in a living cell. This leads to a *directed network* in which the in- and -out-degree distributions can be different and paths are directed. The next step is a *weighted network*: metabolites are characterized by their concentrations in the cell, edges are weighted by fluxes carried by reactions. A different way of adding complexity to the representation is by constructing a *bipartite graph*, where metabolites on one side connect to reactions on the other.

The above example suggests a natural organization for this section: presentation of simple graphs and their characteristics followed by more detailed networks, in parallel with specific real-world examples and models.

Simple Graphs

Degree Distribution The Königsberg bridge network had only four nodes; the network examples we cite nowadays have from hundreds to millions. Most of them show a degree heterogeneity best captured by the degree distribution, which carries more information than the average degree, as pointed out by Barabási and Albert [15]. Examples of scale-free networks include networks of metabolic reactions [75], genetic regulatory interactions [95,128,138], earthquake event correlations [13], word co-occurrence and synonyms [52,61,101,129], power lines [8,148], air routes [11,18], the Internet [29,60,67], the World Wide Web [1,2,6,83,88], software systems [102], Wikipedia links [31], phone calls [4], e-mails [55], co-authorship [17,104], scientific citations [124], World trade [127], innovation flow [48], sexual partnership [90] and the list goes on.

Barabási and Albert had a much shorter list of examples to work with in 1999. Nonetheless, they saw that three very different networks, the actor network, power grid and World Wide Web had similar degree distributions, different from that of a random network model. They argued that the Erdős–Rényi random network model lacks two important features present in real-world systems. First, it is a static model, while most real-world networks constantly evolve and grow. Second, the random network model is too democratic: the probability of linking any two nodes is constant. Barabási and Albert proposed that in real networks the nodes that already have a large degree are more likely to receive new links as the network grows (think of Google in the WWW network). Preferential attachment accompanied by network growth was the first mechanism ever reported to reproduce the scale-free (power-law) degree distribution seen in real networks. Interestingly, though, Barabási and Albert were not the first to report it. A sociologist named Price reported the first example of a scale-free network back in 1965, that of citations linking scientific papers [44]. Building on the "the rich get richer" idea proposed to explain wealth distributions, Price constructed a network model for citations in which the more citations a paper has, the more likely it is to acquire further citations [45].

The scale-free model, simple and analytically solvable, propelled scale-free networks and preferential attachment to the forefront of the expanding field of complex networks research. Barabási and Albert showed that both growth and preferential attachment are essential for obtaining a scale-free network. Growth along with uniform attachment leads to an exponential degree distribution, while no growth with preferential attachment leads to

a Gaussian distribution (although the system does start out with a transient power law) [16]. They also considered the effects of random rewiring and internal link formation, showing that internal link dynamics cause deviation from the power-law at low degrees, observed in real-world networks [5]. A variety of models and analytical methods were developed in the wake of these papers, addressing the effects of further changes in rules of growth or attachment on the degree distribution. The exact solution for the degree distribution of the Barabási–Albert model was worked out by graph theorists Bollobás and Riordan [27].

One of the most important generalizations of the original scale-free model was the study of nonlinear preferential attachment by Krapivsky and Redner [85,87]. They found that power-law scaling is destroyed by nonlinearity: sublinear attachment leads to a power law multiplied by a stretched exponential, while faster than linear attachment causes the network to "condense": the fraction of nodes connected to a single super-hub is finite in the thermodynamic limit. Indeed, the simple linear attachment rule of the scale-free model was later verified in real systems: citation networks, the Internet, the actor and scientific collaboration networks [76,103].

Variations on the scale-free model include linear preferential attachment offset by a constant [53], internal edge creation and removal [50,86], growing average degree [51] (seen in the WWW and co-authorship networks [17]), and edge rewiring [5,135]. One of the challenges of modeling the WWW with the original scale-free model was spotted by Adamic and Huberman: while the oldest nodes are the ones with the highest degree in the model, the WWW does not show this correlation [2]. Bianconi and Barabási proposed a multiplicative fitness model [21] in which the attachment rule is influenced by the degree as well as the "worth" of a node. This model generates both scale-free networks and "winner takes all" scenarios; the transition between the two outcomes maps beautifully onto a Bose-Einstein condensation [22].

The presence of preferential attachment can be justified in many real systems such as the WWW, citation, collaboration or airport networks through the larger visibility of high-degree nodes and the advantage nodes gain by linking to them. There are, however, scale-free networks where a different mechanism leading to preferential attachment is necessary. Biological networks offer intriguing examples: the metabolic and the protein-protein interaction networks are scale-free, even though their connections are governed by biochemistry and not choice. Protein interaction networks have inspired a class of models based on gene duplication (duplication of a node and all it's links) and subsequent mutation (addition and/or deletion of some of the copy's links) [130,145]. Preferential attachment in these models is a consequence of the evolutionary dynamics: nodes with a higher degree have more duplicating neighbors, thus receiving more links. Vertex (node) copying has been proposed as a possible mechanism for the growth of the WWW [83] and auto-catalytic networks [74].

In the spirit of the Erdős–Rényi model, Bender and Canfield proposed an ensemble model for scale-free networks. The configuration model describes the group of all networks with a prescribed degree sequence [20]. If the sequence is chosen from a power-law distribution, the resulting networks are naturally scale-free. However, the average properties of the ensemble carry no other characteristics intrinsic to evolving models (such as degree correlations or clustering). Moreover, the simple definition of the model makes it ideal for analytical approaches. Newman et al. used probability generating functions to calculate exact expressions for average path length and clustering in configuration networks. (They used a generalized definition: the ensemble of all networks with an ensemble of degree sequences drawn from the same distribution.) [112]. Many analytical results for scale-free networks were proven using the configuration model: they have asymptotically vanishing clustering coefficients [42] and they are ultra-small for the degree exponents measured on most real-world networks ($\gamma \in (2,3)$): their average path length scales as $\mathcal{O}(\log\log N)$ [34].

Some real networks are embedded in physical space and have physical connections: brain networks, the power grid, the Internet, airport networks, streets, public transportation systems, highways and rivers. Some of the above examples (brain networks, streets, rivers) are missing from the list of scale-free networks: spatial constraints can be forbidding to the formation of hubs. Several studies with both preferential attachment and bias towards shorter links show that large length costs can destroy the scale-free nature of spatial networks [19,92,151]. Systems with fixed maximum link length, such as wireless networks in which the range of a particular device is much smaller than the physical size of the system are not scale-free. Conformation networks made of all physically allowed conformations of a system (a polymer or bead chain) are also homogeneous structures naturally embedded in an n-dimensional configuration space defined by the system's degrees of freedom [123,126].

The distribution of nodes in space can also have great influence on network topology, as shown in a detailed study of the Internet by Yook et al. [152]. They found that the router density distribution is a fractal with the

same dimension $D = 1.5$ as the population density distribution. They were able to mimic the topology of the Internet using a simple evolving model in which nodes are distributed according to a fractal distribution and incoming nodes connect to old ones using preferential attachment divided by some power of their distance. Interestingly, the only model parameters able to reproduce the internet's topology were the ones actually measured for the real system: fractal dimension $D = 1.5$, linear preferential attachment and a probability of connecting two nodes that is inversely proportional to the distance between them. A similar model was also successful in describing the topology of the world-wide airport network [69].

Paths on Networks, Small Worlds and Betweenness A networks' most basic function always requires some type of communication along its edges. Thus it is natural that average shortest paths, network diameter and local betweenness measures are of great interest to networks research. Erdős and Rényi proved that the diameter of their random network model scales as the logarithm of node number. Watts and Strogatz called such networks small worlds [148], in tribute to the "small world phenomenon" in sociology: the idea that one can connect any two people on Earth by about six handshakes between mutual acquaintances. In 1967, an ingenious experiment by Milgram proved the existence of short paths of an average six hops between random people in the US, known as "six degrees of separation" [94]. The small word nature and searchability of small-world networks allow movie lovers to find short chains of movies connecting their favorite actor to Kevin Bacon. Mathematicians (and networks researchers) play the same game on the co-authorship network: one's shortest path to Pál Erdős is called the Erdős number.

The small world model introduced by Watts and Strogatz is based on the idea that real networks are in between random graphs and regular lattices: they are highly clustered (as a regular lattice), but they also have shortcuts. The model is built starting with a regular low-dimentional lattice to which one adds (or rewires) a certain number of edges, shortcuts between distant parts of the lattice [148]. Watts and Strogatz showed that increasing the number of shortcuts turns regular lattices into random networks, moreover, a small number of shortcuts is sufficient for the small-world effect without destroying the high clustering coefficient of the original lattice [111,148]. Most real networks display both characteristics of the small world model: short average path lengths and high clustering. However, the small world model has a peaked degree distribution, thus the topology of real networks with scale-free degree distribution is fundamentally different from the ones generated by the model. The Barabási–Albert scale-free model, on the other hand, fails to account for the high clustering of most real networks.

Logarithmic or slower increase of the average path length was proved for a variety of network models [25,26,32]. Bollobás and Riordan showed that the average shortest path length in scale-free networks grows no faster than $\log N/\log(\log N)$ [25]. Cohen and Havlin used the configuration model to show that random scale-free networks with $\gamma \in (2, 3)$ are ultra-small: their average shortest path length grows as $\log(\log N)$ [34].

The structure of shortest paths is crucial to any communication or flow between nodes of a network. Naturally, betweenness was found to be the relevant quantity when one deals with congestion or disruption of flow in networks. Goh et al. measured the distribution of node betweenness values for a variety of real networks as well as models, and showed that it follows a power-law with only two distinct exponents [66].

Clustering and Network Motifs Clustering is a rediscovery of "network density", a quantity widely used in sociological network analysis: it provides a measure of how well one's acquaintances know each other. High values have been observed in social networks, but also in most other real-world networks, motivating the development of clustered scale-free models. The Holme–Kim model, aimed at creating scale-free networks that also have high clustering coefficients, is a straightforward generalization of the scale-free model, with triangle-forming steps complementing preferential attachment [72]. Klemm and Eguíluz introduced a citation network model based on the idea that papers are only cited for a limited stretch of time before they are forgotten [84]. The model is built by the constant addition of nodes that connect to all "active" nodes and join their ranks. Then, one of the active nodes is deactivated with a probability inversely proportional to its in-degree (offset by a constant). This model leads to scale-free networks with high clustering coefficient, but fails to capture their small-world nature: visualized, the networks look like tubes. Gene duplication models, while motivated by evolutionary arguments, also lead to clustered scale-free networks [130,145].

An interesting generalization of the clustering coefficient was introduced by Uri Alon's group: network motifs are significantly over- or underrepresented patters of connections between n vertices (compared to the randomly rewired network). Distinct characteristic motifs were found in regulatory networks, food webs, neural networks and WWW, corresponding to local functions

performed by the network [95,128]. For example, different types of feed-forward loops (FFLs, directed motifs of 3 nodes: $A \to B$, $B \to C$ and $A \to C$) in genetic regulatory networks perform distinct signal processing roles as shown by both simulation and experiments with living cells [10]. FFLs made of activating iterations only (abundant in the regulatory network) can filter out transient changes in the concentration of the input node, while also delaying the turn-on or turn-off of the output node. On the other hand, FFLs where the $B \to C$ link is a repressing one (also abundant) act as pulse generators.

Degree Correlations and Mixing Patterns In social networks, links between people who are alike are more common; popular people are connected with popular people. Assortative mixing means that degrees at two ends of an edge are correlated, as measured by the conditional probability, $P(k'|k)$, that a node with k links is connected to another one with k' links. Measurements on the Internet and protein interaction networks show that in these system, as opposed to their social counterparts, small degrees are more likely to connect to high degree nodes [93]. The conditional probability is difficult to measure in most real networks due to poor statistics, although it is convenient in analytical work. A more compact representation of degree correlations was defined by Pastor-Satorras et al. [119]: the average degree of neighbors as a function of degree, $k_{nn}(k)$, which decreases with k for disassortative systems. One can further simplify the measure of assortativity by calculating the Pearson correlation coefficient of degrees at the ends of a network's edges: the assortativity coefficient [105,107]. Intriguingly, most social systems are assortative (actor network,company directors, coauthorship networks, phone calls, email address books), while most technological and biological systems prefer disassortative mixing (WWW, Internet, train routes, software packages, software classes, electronic circuits, peer-to-peer networks, metabolic networks, food webs, neural networks).

Degree is not the only property of a network that can show assortative or disassortative mixing. In networks where nodes can be classified in types of some kind, mixing between types can also be characterized by the assortativity coefficient (see definition). Maslov et al. showed that there are three main types of nodes in the Internet: high level providers who manage the backbone and trunk lines, Internet Service Providers who bring the network out to end-users and the end-users themselves. These three types show strong disassortativity: few end-user to end-user or ISP to ISP links in the network. In social networks, mixing by race, age or income has been observed [107].

Communities, Hierarchy and Fractality Community structure is an intuitive feature of complex networks: one expects them in social systems (circles of friends), biological networks (functional units), the WWW (websites related to a topic or organizations), co-authorship networks (scientific fields and sub-fields), etc. Communities within networks are structural features related to the function of the network as a whole, and are thus expected to have a strong influence on their dynamics. Defining and finding network communities has its history in both sociology and computer science, and has been revisited many times. The methods developed along the way paint a rich picture of the structural diversity of complex networks.

Non-Overlapping Community Structure Hierarchical clustering or cluster analysis is widely used in the study of social networks. The first step in hierarchical clustering is the construction of a similarity measure between network nodes. Next, each node is assigned to a separate cluster (the leaves of the dentogram). The two most similar clusters are joined to form junctions of the dentogram, until all clusters have been united, forming the root. Many similarity measures have been defined and used successfully. Structural equivalence, used in social network analysis, gives two nodes the highest similarity score if they have the same pattern of relationships. It can be measured using Euclidean distance or Pearson correlation coefficient between rows of the adjacency matrix, as well as topological overlap, defined as the number of overlapping neighbors divided by the smaller of the two degrees [122].

Girvan and Newman introduced a famous community detection algorithm that is similar to hierarchical clustering, but works divisively [64]. It is based on the idea that the edges most likely to run between communities are the ones that have the highest edge betweenness centrality. They first remove the edge with the highest betweenness, then recompute edge betweenness and keep removing edges until the network falls apart into non-connected nodes. As the network falls appart, they draw a dentogram where each joint is a splitting event.

Hierarchical clustering works with varying success for different systems, but it cannot determine how many communities there are in the network. Newman et al. introduced an elegant measure for the "quality" of any given partitioning into communities, called modularity, which compares a community partitioning to a null model that can be appropriately chosen for the system at hand [109]. The Girvan–Newman algorithm, together with the dentogram cut that maximizes modularity, has been successfully used for many different social and biological networks [12,24,71,110,141,150]. The main drawback of the

method is computation time ($\mathcal{O}(N^3)$ on a sparse graph): it is not feasible for networks with more than a few thousand nodes.

In a quest to develop a faster community detection method closely tied to the definition of communities, Newman used direct optimization of the modularity measure [108]. Since finding the best partitioning is an NP-hard problem, his first approach was a greedy optimization. All nodes start out as separate communities, which are then repeatedly joined such that the increase in modularity is maximal (or decrease is minimal). Newman showed that the method works well both in tests and real networks, and it works quite fast: $\mathcal{O}(N^2)$ time for sparse graphs. Still not fast enough? A collaboration with Clauset and Moore resulted in an algorithm that performed the same optimization in $\mathcal{O}(N \log^2 N)$ for sparse graphs with many levels of communities [33], solving the problem of partitioning networks with billions of nodes. (The general result for runtime is $\mathcal{O}(KD \log N)$, where K is the total number of links and D is the depth of the generated dentogram).

Interested in the theoretical foundations of community structure detection and its relationship to matrix spectra, Newman argued that community detection requires the use of the modularity matrix [109] in place of the Laplacian matrix, as done by traditional spectral methods of graph partitioning. (The Laplacian matrix of a graph, L, is a real symmetric matrix with elements $L_{ij} = k_i\, \delta_{ij} - A_{ij}$.) The modularity matrix is defined as $B_{ij} = A_{ij} - P_{ij}$ (see modularity definition). He showed that the eigenvalues and eigenvectors of the modularity matrix encode the networks's community structure. Methods based on this relationship perform as well, if not better, than previous ones, but more importantly, they are also able to detect "anti-modular", or bipartite structure [109].

Hierarchical Community Structure and Fractal Networks Ravasz et al. argued that in many real networks with interesting modular structure there is no ideal partitioning into distinct modules: the network is built of small, very cohesive communities, hierarchically embedded in larger, less cohesive ones [121,122]. Their hierarchical model is a deterministic construction of a scale-free network with hierarchically embedded modules. They showed that hubs on all scales unite communities on all scales: small nodes are part of small, very cohesive clusters (thus have large clustering coefficients), larger nodes serve as connectors of these clusters, while the largest hubs span modules at the highest level or organization (and have low clustering coefficients). Indeed, they found that the clustering coefficient $C(k)$ decreases with increasing k in a variety of real networks, such as the metabolic network, synonyms, movie actors, the WWW and the Internet represented at Autonomous System level (each node is a domain, not just a computer), and it often scales as $1/k$ [121]. This scaling was found in a variety other networks: software systems [102], the World Trade network [127], the worldwide airport network, the co-authorship network of http://arxiv.org/archive/cond-mat [18,104] and protein folding networks [120]. Not all networks are hierarchical: the power grid and the Internet at router level show no scaling of the clustering coefficient [121]. Ravasz et al. suggested that spatial embedding of both systems, where length costs are significant, works against hierarchical modularity. Both networks distribute something to a physical area. Thus, the formation of tight communities is not required for these networks to function.

Hierarchical organization in metabolic networks was found to reflect biological function on different scales of organization: the branches of the dentogram obtained by hierarchical clustering (using topological overlap as similarity measure) correspond to known functional classes of the metabolism on two different levels of organization [122].

The question of self-similarity or fractal nature has been on the mind of network researchers since the scale-free degree distribution was observed. Are networks like fractals, self-similar on all scales? Hierarchical organization strengthens this image: most networks not only have degrees in a broad range of scales, they are also made of modules of different scales embedded into each-other. The problem with complex networks as fractals was, of course, that most networks of interest to us are also small-world [148]. In a small-world network the number nodes at a distance l from any given node increases exponentially. One expects this number to grow as a power law in the case of a fractal object: measuring the "mass" within distance l from a point on the object, called the cluster growing method, is one way of measuring fractal dimension.

While some deterministic models have been constructed with the idea of fractality in mind [54,77], the breakthrough in understanding topological self-similarity in complex networks was brought forth by Song, Havlin and Makse [132,133]. They generalized the standard box counting method for measuring fractal dimension of a physical object to complex networks. How does one cover a network with boxes of different size? Divide all nodes in groups such that the shortest path between any two nodes in a group is at most l_B long: these are the boxes. Use the smallest number of groups necessary to do this

(or a decent approximation) and repeat it for $l_B \in [2, D]$. The results of this procedure were quite surprising: many scale-free real-world networks, such as the WWW, actor network and various metabolic networks show a neat fractal scaling between the number of boxes and their size. Thus these networks are self-similar, fractal structures. To prove their point further, the authors used a renormalization procedure where they collapsed each box into a node and linked these new nodes to each other if any member of the original boxes had a connection. The networks renormalized this way were also a scale-free, with the same degree distribution exponent, independent of the box size used for renormalization [132].

The question of how these networks are small world and self similar at the same time has also been resolved: the two methods of determining the fractal dimension of an object, box counting and cluster growing are not equivalent on complex networks with broad degree distribution. While box counting covers all the hubs only once (they are assigned to one box only), cluster growing finds the hubs for almost any choice of seed node, thus bringing a large part of the network into the cluster with them. This explains the exponential increase in the number of nodes within a distance l, and shows the small-world property of the system. Box counting, on the other hand, can reveal the fractal nature of the network, if present [132].

Song et al. uncover some of the requirements of fractality via proposing a network growth model based on reverse renormalization. Starting from one point, the network grows by nodes transforming into small clusters in each iteration. The conversion from a node to a small cluster mimics the renormalization process in reverse: the degree of a node grows by multiplication with a scaling factor (thus nodes of previous iterations become the hubs). The newly formed clusters have a diameter of b_B (the box size). They showed that the key feature that influences fractality in the emerging networks is how the clusters connect: if the link always runs between the hubs (leading to assortative mixing), the result is small-world network that is not a fractal. The internet at router level, found to be non-hierarchical, is a good example of such an assortative, non-fractal network. On the other hand, if the clusters only connect to each other via the non-hub (newly created) nodes, the resulting network is a fractal, but it loses it's small-world character. Many real-world networks, however, seem to be both fractal and small-world (WWW, actor network, protein interaction networks, metabolic networks). Indeed a very small number of hub-to-hub connections in this model can restore the small-world property of the network while preserving its fractal nature.

Overlapping Community Structure The community detection methods presented thus far assign each node to only one community. Palla et al. argued that many nodes in real networks belong to more than one community [114]: proteins can simultaneously be part of several complexes, people have disjoint groups of acquaintances from friends to work to extended family. They proposed a community definition that allows them to capture overlap between communities. Their method is based on "k-clique rolling": a k-clique community is the union of all k-cliques (complete subgraphs of size k) that can be connected through a series of k-cliques that share $k-1$ nodes. Increasing values of k lead to smaller, denser, but more disjoint communities. They find that k-clique communities in real networks (co-authorship, word association and protein interaction networks) have a power-law size distribution, and there are significant overlaps: the size distributions of overlap as well as node membership have fat tails.

Directed Networks

Directionality of a link is often important in real networks: scientific citations, url links, genetic regulatory interactions are only a few examples of inherently asymmetric connections. Directed links require the separate measurement of in- and out-degree distributions. In case of the WWW both are power-laws, but the degree exponents differ: $P_{\text{in}}^{\text{WWW}}(k) \sim k^{-2.1}$ and $P_{\text{out}}^{\text{WWW}}(k) \sim k^{-2.45}$ [6]. Genetic regulatory networks show an even stronger difference. Their out-degree distributions are power laws, however, the in-degree ones are scaled: genes cannot receive input from hundreds of transcription factors.

Another interesting consequence of directed links is the rich structure of directed paths. They were found to partition the WWW as well as metabolic networks into parts resembling a bow-tie [28,91]: a strongly connected component, in which there is a directed path in both directions between any pair of nodes, an IN-component, the nodes of which can reach the strongly connected component bot cannot themselves be reached, and an OUT-component that can be reached from, but has no directed paths leading into the strongly connected component. Not all directed networks have bow-tie structure: some of them are entirely acyclic, with no directed loops. Citation networks are a natural example: one can only cite papers already published. However, genetic regulatory networks are also acyclic (not considering auto-regulatory loops), even though the cause of this is now yet understood [14,138].

Gradient networks are directed graphs that generalize the concept of gradients from continuous scalar fields

to networks. They capture the backbone of a gradient-induced flow on complex networks: given a substrate graph with a scalar value associated to every node, its gradient network is formed by the collection of all directed links that lead from every node to its neighbor with the highest scalar value [139]. These directed links form collections of trees. For an independent, identically distributed association of random variables to the nodes of an Erdős–Rényi graph, the generated gradient network has been shown to be scale-free, with a connectivity exponent of $\gamma = -1$. This finding can be proved for any substrate graph with no loops shorter than 5 [140]. Moreover, the gradient networks' scale-free nature seems to be universal: it was numerically observed for a wide variety of substrate networks (including ones with short loops): regular and random trees, Erdős–Rényi and small-world networks, high dimensional regular lattices and n-tori, random geometric networks [39], the scale-free and the configuration model [123,140].

Weighted Networks

Real world networks display significant heterogeneity in the strength of their connections. The distribution of link weights has been found to have a heavy tail in metabolic networks: the steady-state flux distribution follows $\Theta(w) \sim (w_0 + w)^{-1.5} (w_0 = 3 \cdot 10^{-4})$ [9]. Moreover, the average link strength scales with the connectivity of the nodes at the two ends as $\langle w_{ij} \rangle \sim (k_i\, k_j)^{0.5}$ in both metabolic networks and the Worldwide Airport Network, indicating correlations between node degree and links weights [9,18]. Scientific collaboration networks, on the other hand, have a weight distribution that is not correlated with degree.

The existence of weighted links requires a generalization of most network measures:

- **Node strength and strength distribution.** Node strength, s_i is a natural generalization of the node degree, k_i:

$$s_i = \sum_{j=1}^{N} w_{ij}\, A_{ij} \ ,$$

the sum of weights over the links of node i. The strength distribution, $P(s)$, is typically also heavy tailed [18]. Weights are often dependent upon topology (see metabolic and airport networks), expressed by a non-linear relationship between node strength and degree: $s \sim k^{\beta}$ (in the airport network $\beta = 1.5$). Heterogeneity in the weights around a node can be measured using

$$Y_i = \sum_{j=1}^{N} \left(\frac{w_{ij}}{s_i} \right)^2 A_{ij} \ .$$

If all edges have the same weight, $Y(k)$ scales as $1/k$, while if one weigth is significantly larger than the others, $Y(k) \simeq 1$. In metabolic networks $Y(k)$ was found to scale as $k^{-0.27}$, indicating that metabolites used in a large number of reactions are more likely to have one high-flux reaction dominating their production (consumption) [9].

- **Weighted clustering coefficient.** Defined by Barrat et al. as

$$C_i^W = \frac{1}{s_i\,(k_i - 1)} \sum_{j=1}^{N} \sum_{m=1}^{N} \frac{w_{ij} + w_{im}}{2}\, A_{jm}\, A_{ij}\, A_{im} \ ,$$

the weighted clustering coefficient is often compared to the standard one: $C^W > C$ means that the triangles of the network are preferentially formed by high-weight links. This is indeed the case in the co-authorship network, for nodes with $k > 10$: more established investigators form stable, high-weight cliques (research groups) from which the main volume of their publications originate. A similar phenomenon can be seen in the world-wide airport network: larger airports form high passenger-flux triangles, so called "rich-clubs" [18].

- **Degree Correlations.** The weighted average degree of neighbors is defined as:

$$k_{nn,i}^W = \frac{1}{s_i} \sum_{j=1}^{N} w_{ij}\, k_j\, A_{ij} \ ,$$

a sensitive probe into the structure of weighted networks. The behavior of $k_{nn}(k)$ in the airport network shows a plateau for airports with more than 10 direct fights, indicating no degree preference. However, $k_{nn}^W(k)$ increases in a wider range and is in general larger $k_{nn}(k)$, showing that while large airports do not preferentially connect to large airports, the high traffic links run among members of the "rich-clubs" [18].

Dynamics on Complex Networks

The main theme of this section is: how does the structure of a network influence its dynamics? Most lessons of statistical mechanics are worth revisiting when the underlying space is a network: does an inhomogeneous topology change the dynamics? The answer, as we will see, is yes, in interesting ways.

Robustness and Vulnerability of Complex Networks

Perhaps the most intriguing feature of complex systems is their robustness. Close to 75% of the genes in *E. coli* are non-essential: the organism can survive without them (under one set of growth conditions) [63]. If the webpage of a company or university goes down, the WWW as a whole is still perfectly functional, the effect of the failure is local. On the other hand, an intentional "attack", similar to the denial-of-service attacks that crippled Yahoo, Amazon, eBay, CNN and a few other very popular websites in February of 2000, can substantially cripple the function of a complex system.

Vulnerability and robustness go hand in hand in most complex systems: bad weather in Atlanta (although usually not labeled an "attack") has very different consequences for air traffic than problems at a regional airport. Opinions expressed in the New York Times are much more likely to spread and influence events than opinions in a small town local newspaper. These trivial examples hint at interesting consequences of in-homogeneous network topology, brought to light in a 2000 Nature paper by Albert, Jeong and Barabási [7]: while scale-free networks are largely unaffected by random failure, they are very sensitive to change in their highly connected, central nodes, rendering them vulnerable (and/or responsive) to planned, targeted interventions.

Resilience One of the simplest indicators of robustness under the damage done by node or link removal is a structural one: the size of the largest connected component. Numerical studies performed by Albert et al. show that the giant connected component of Erdős–Rényi random networks falls apart at a much lower fraction of randomly removed nodes that of a scale-free network [7]. A fraction of nodes the removal of which causes a random network to falls into pieces, only slightly shrinks the giant component of a scale-free network, while small fragments of the system become isolated. An attack, on the other hand, where the most connected nodes are the first to be removed, shows a different picture. While the Erdős–Rényi random network falls appart somewhat faster but in essentially the same way, a scale-free network is blasted appart by the removal of a much lower fraction of nodes. The results hold for simulations using the actual network topologies of the Internet and WWW [7,28].

The topological aspects of random node or edge removal can be calculated by mapping random failures to percolation problems. Cohen et al. have shown that uncorrelated random networks with a diverging second moment of their degree distribution (scale-free networks with degree exponents between 2 and 3) have zero percolation thresholds [35]. Vázques and Moreno used an approach that allowed them to investigate the percolation properties of correlated networks, showing that assortative mixing is beneficial for resilience: it can push the percolation threshold down to zero, even for networks with a finite second moment [144]. The opposite effect is also true: scale-free networks with diverging second moments but dissasortative degree correlations are less resilient to node failure [89]. A general analytical approach based on generating function formalism not only reproduced the previous results, but also allowed Callaway and Newmann to investigate degree-dependent node removal scenarios, such as attacks [30]. Interestingly, assortative mixing does not help in case of an attack: Song et al. have found that non-fractal networks created via a mechanism that favors hub to hub connections on all scales are more vulnerable to intentional attack than their fractal (and also dissasortative) counterparts [133].

Cascading Failures The function of a variety of complex networks involves the transmission or flow of some conserved quantity. Removal or failure of a node in such a network has consequences that ripple through the rest of the system far beyond the effects on connectivity: the node suddenly sheds the load or flux that it carried before, thus its neighbors suddenly experience higher loads: some of these overload and fail. A cascading failure can follow, as often seen on the power grid, perhaps the most quoted example of the phenomenon. Unlike the fragmentation of a network, a cascading failure can be triggered by a relatively small number of node failures, often a single one [73,100].

Using a model of overload in which the load-bearing capacity of a node is proportional to its load (betweenness) in the full network, Motter and Lai showed that scale-free networks are more vulnerable to randomly seeded cascading failures than random networks [100]. The vulnerability becomes especially pronounced under targeted attack: overload of the largest node can cause cascades that propagate through the whole system. They find that the most dangerous targets are not the most highly connected, but the most load-baring (highest betweenness centrality) nodes. Betweenness is usually correlated with degree. The US powergrid, however, is vulnerable under attack targeting its most central nodes, but not its highest degree ones.

In a followup paper Motter introduced a fast defense strategy against cascading failures [99]. He argued that the only defense strategy which is fast enough is further removal of nodes: counterintuitive, but effective, if the nodes are carefully chosen. The reason this is possible, he argued,

is because nodes that carry small amounts of load (thus are less central to the network) actually generate much more load than they carry. Their shortest paths to the rest of the system are larger, thus all the communication (traffic, power) they receive or generate affects a larger number of intermediary nodes, increasing the total network load. Generated and handled load are anticorrelated, thus removal of nodes in ascending order of loads significantly reduces the size of cascading failures in the rest of the network, as verified by numerical simulations.

Congestion Cascading failures and congestion of traffic carried by networks are similar problems. As opposed to the node-removing effect of a failure, congestion does not disconnect a node, nonetheless, it can bring the traffic-bearing capability of a system to a halt. Most congestion models show a transition from free flow to congestion as the load is increased, regardless of the network type.

Ohira and Sawatari showed the occurrence of a jamming transition on a simple traffic model in which packets travel along the shortest paths of a two dimensional lattice between randomly chosen boundary nodes [113]. On a regular lattice shortest paths are highly degenerate, thus the authors considered two strategies: one is deterministic in picking it's route, the other is random, preferring shortest paths but occasionally picking longer than optimal ones. They found that the jamming transition occurred at larger packet creation rates if the routing was probabilistic, with an optimal randomness that does not considerably lengthen the travel paths, but relieves the network from always choosing congested paths. Guimerá et al. considered a different formulation of the congestion problem on lattices and Cayley trees where all nodes generate packets to random destinations and send them along shortest paths, but the probability for a packet to hop between two nodes along the path depends on the queues accumulated by these two nodes [70]. Thus, there is some congestion-awareness built into the model. If the processing speed of nodes decreases with an increased number of packets, the system shows a discontinuous transition from free flow to congested state: the ratio of undelivered packets (the order parameter) jumps from 0 to 1. The positive feedback between congestion and slower processing leads to the formation of congestion nuclei that spread through the network. On the other hand, no transition is observed in networks in which the processing speed of nodes increases with queue lengths, only a crossover from low-density flow to high density flow accompanied by a change in fluctuation statistics. The critical case in between is queue-independent processing, which shows a continuous transition between free flow and congestion.

Solé and Valverde proposed a generalization of the Ohira–Sawatari model [131] and showed that the system at transition point exhibits self-similar time-series dynamics with an $1/f$ power spectrum, as observed in measurements of packet transmission times on the Internet. Latency times and queue lengths also showed heavy tails with similar queue length distributions to jam size distribution in highway traffic models. Interestingly, the systems reaches its highest efficiency and information transfer regime right at the critical point, before entering the congested state [131].

Building on the observation that the time series dynamics of the model close to the critical point matches the real data, the authors introduced a self-organizing version of the model [142]. They argued that packet generation (thus user behavior) is linked to dynamics: each user tries to increase its rate until congestion of neighboring nodes is detected, at which point it starts to decrease it, dropping its rate to 0 if all it's neighbors are congested. This model generates highly heterogeneous dynamics in space and time, with a power-law congestion length distribution. Moreover, this model points to internal network dynamics being responsible for fluctuations, supported by an analysis of real fluctuations by de Menezes and Barabási. They observed a power-law scaling of flux fluctuations as a function of total flux values, with two distinct scaling exponents. $\alpha = 1/2$ scaling is generated by fluctuations internal to the system (as seen in the Internet and on a microchip), while $\alpha = 1$ corresponds to external noise (seen in the WWW, river networks and the highway system) [43]. Solé and Valverde have further extended their model to study complex topologies similar to the Internet (using the model by Yook et al. [152]), with a routing strategy that can be tuned from random routing to global shortest path routing. Each node is assumed to know its neighborhood to m hops. If the destination of a packet is within a node's search horizon it uses shortest path routing, otherwise it passes the packet to a random neighbor [143] (local routing strategies with a fixed search horizon were also investigated by Tadić et al. [136,137]). They found that the routing was optimal when the search horizon equaled the average path length of the network. Further push toward global shortest paths actually decreased the efficiency by over-specifying paths and thus exacerbating congestion. The dynamics observed with optimal search depth reproduced the 1/2 scaling between fluctuations and mean flow observed in [43], as well as the power-law exponent of the average latency time distribution measured on the Internet.

Zhao et al. investigated the effect of network topology on congestion in a model in which the packet processing

speed of nodes is determined by their degree or their betweenness. They found that if the capacity of nodes was proportional to their degree, random networks as well as scale-free ones were less prone to congestion than regular lattices or Cayley trees. However, systems in which the node capacities were proportional to their betweenness had the same critical packet generation rate regardless of topology, suggesting that selectively increasing the capacity of high-betweenness nodes is a good way of increasing the carrying capacity of the system as a whole [154].

A different approach for investigating the effect of network topology on congestion was proposed by Toroczkai and Bassler [139]. They measured the congestion factor of the Erdős–Rényi and Barabási–Albert models, defined as the average fraction of nodes that do not receive and thus process incoming traffic. Instead of an explicit choice of routing or packet generation behavior, they assumed that packets on average follow the steepest gradients towards the neighbor with the highest "potential", thus the gradient links determine the congestion factor (they assume a random distribution of node potentials). They found that the congestion factor of Erdős–Rényi random graphs increases with the size of the network, asymptotically growing to 1, while for scale-free networks it quickly reaches a value of ~ 0.7 and does not increase with system size. A followup study by Danila et al. proposed a traffic model based on the idea of congestion-gradient driven flows [41].

The recognition that congestion occurs when the average number of packets processed by the busiest node reaches 1 (time-step) highlighted the importance of betweenness in the study of congestion. Routing protocols can influence the experienced betweenness of a node: the number of packets that actually go through it on average. Several methods of reducing the largest betweenness value have been investigated: optimization of link weights such that the shortest weighted paths give rise to a small maximum betweennes [40], hub avoiding global routing schemes [134] or traffic-aware routing [56]. Sreenivasan et al. proved that for any network topology there always exist an absolute upper bound for the communication threshold, determined by network topology, above which no routing algorithm can increase the critical congestion threshold [134].

Spreading Processes and Social Dynamics

Epidemic Models Epidemiological modeling jumped to the forefront of networks research with a landmark paper by Pastor-Satorras and Vespignani which revealed the striking difference between virus spreading on scale-free networks and homogeneous systems [115,116]. They studied the "susceptible–infected–susceptible" (SIS) epidemic model, suited to describe diseases that confer no immunity, such as tuberculosis, gonorrhea as well as computer viruses on systems that do not update their virus protection software. Each susceptible node can be infected with rate ν if it is connected to one or more infected nodes, while infected nodes are cured with rate δ and become susceptible again. The SIS model was known to show a non-equilibrium phase transition at a critical spreading rate $\lambda_c(\lambda = \nu/\delta)$: if the spreading rate is higher than this threshold, the disease is endemic; a finite faction of nodes is persistently infected. Below the threshold the disease dies out completely after an initial breakout. As Pastor-Satorras and Vespignani pointed out, these results were obtained on lattices and failed to explain the very low but sustained prevalence of computer viruses. The authors report the absence of a critical point on scale-free networks with $2 < \gamma \leq 3$: there is no non-zero spreading rate for which the disease dies out. As λ approaches 0, the prevalence of the disease decreases exponentially, but only reaches 0 when $\lambda = 0$.

In a followup study together with Moreno they showed that the lack of an epidemic threshold holds for the "susceptible–infected–recovered" (SIR) model as well [96]. This model, applicable for diseases that result in immunity or death, has a radically different overall behavior from the SIS model. Recovered individuals cannot get infected again, thus the disease always dies out. Nonetheless, for homogeneous networks it shows a phase transition at a critical spreading rate λ_c, reflected in the total fraction of the population to ever get the disease. Below λ_c this fraction is vanishing as the system size goes to infinity, above λ_c there is a finite infected fraction. Moreno et al. have found that the critical spreading rate for uncorrelated complex networks is $\lambda_c = \langle k \rangle / \langle k^2 \rangle$.

Thus, networks with diverging connectivity fluctuations (such as scale-free graphs with $2 < \gamma \leq 3$) have no non-zero epidemic threshold. Although real and thus finite-size networks cannot have infinite $\langle k^2 \rangle$, the epidemic threshold is very small for large systems, much smaller than that of a similar homogeneous network, and it decreases with systems size [117]. The SIR model has been mapped onto a bond percolation problem by Grassberger [68]. This mapping was exploited by Newman, who used the generating function formalism [30,112] to obtain the exact solution of the model in the infinite time limit for simple graphs with arbitrary degree distribution. He then extended the mapping to bipartite graphs to model sexually transmitted diseases spreading between men and women [106].

The absence of an epidemic threshold has profound implications on immunization policies effective in scale-free networks. Dezsö and Barabási [47] as well as Pastor-Satorras and Vespignani [118] have shown that scale-free networks do not respond to random immunization: such a strategy cannot restore the epidemic threshold, or bring the effective spreading rate below it, even for an unrealistically high number of administered vaccines. On the other hand, immunization preferentially targeting the hubs of the network successfully reintroduces an epidemic threshold at very low vaccination rates, even if the strategy is imperfect in identifying the hubs. These results have important and controversial consequences for public policy of vaccination in the context of sexually transmitted diseases. This is partly because a policy that gives priority vaccination to prostitutes is controversial even if the public benefits are substantial, but also because it is generally difficult to identify the hubs of sexual networks. Cohen et al. proposed an immunization policy that overcomes this problem without relying on any global knowledge of the system: immunize a randomly chosen friend of a random person [36]. They showed that this strategy preferentially targets the hubs, thus restoring the epidemic threshold.

Epidemic spreading on networks with degree correlations show that degree correlations do not affect the lack of epidemic threshold in scale-free networks [23,97]. However, the epidemic incidence is smaller in these networks, while diseases are longer lived [97].

The influence of link weights on epidemic spreading has been investigated in the context of diseases such as Severe Acute Respiratory Syndrome (SARS), spreading on the world-wide airline network [37]. Colizza et al. developed a detailed multi-scale model of global epidemic spreading and found that the degree heterogeneity of the airport network accounts for most of the observed behavior of the epidemics, the heterogeneity of its weights has a much smalled influence on the spread of the disease [37]. They found that epidemics are fairly predictable, except at the very beginning of an outbreak originating from a hub, as well as the end of epidemics [37].

Information Spreading The similarity between rumor spreading and epidemic spreading was recognized as early as 1964, when Goffman and Newill used the SIR model to describe the spread of ideas instead of diseases [65]. The authors pointed out an important difference between the two processes: while one studies epidemic models with an effective immunization strategy in mind, information spread is in general favored; moreover, the process can be engineered in a way that facilitates high reliability of the spread. Later in the same year, Daley and Kendall published a spreading model that is specific to rumor (or information) spreading: instead of the random recovery rate of the SIR model (which assumes spontaneous forgetting of the news), they assumed that nodes loose their interest in spreading the news upon encounter with another node that knows it [38]. Thus the "recovery" process of the Daley–Kendall model is very different that that of SIR, leading to a significant difference in the outcome of spreading. Assuming homogeneous mixing (each infected individual, or spreader can randomly contact any node in the system), the final fraction of nodes who have heard the news is $\sim 80\%$.

A decentralized, spontaneous and robust method of spreading information was welcome by the computer science community; it allowed them to overcome the scalability problems of data disseminating protocols in large-scale distributed computing [146] and peer-to-peer networks [78]. It was also used in update management of databases duplicated at many sites [46].

In the context of complex networks, the Daley–Kendall model was first investigated on the small-world model by Zanette, who found that below a density of shortcuts the rumor dies out without reaching a finite fraction of nodes [153]. Moreover, this threshold is finite for infinite systems ($p_c = 0.2$ for $K = 2$), even though the network is turned into a small-world by a vanishingly small number of shortcuts. Moreno et al. studied the same model on homogeneous versus scale-free networks and found that as opposed to epidemic spreading, rumors spread to a higher fraction of nodes in homogeneous networks [98]. Hubs in scale-free networks have a conflicting effect on rumor spreading: while they are highly likely and quick to hear the information, they also facilitate the formation of stifler nodes who do not spread the rumor.

Searching on Complex Networks

Milgram's famous experiment showing six degrees of separation between two random people not only showed the existence of short paths on social networks, it also highlighted the fact that finding a destination node with local information is not very difficult [94]. Motivated by these findings, Kleinberg showed that the performance of a simple greedy search critically depends on the topology of the underlying network. He used a small-world type model in which nodes sit on a two-dimensional lattice (connected to their four neighbors) to which a few long-range connections are added. Unlike in the small-world model, these connections are not added entirely randomly: the probability of a shortcut is proportional to $r^{-\alpha}$, where r is the Manhattan distance between two nodes. In this model the

best-case performance of a decentralized algorithm (the expected delivery time) scales as $(\log N)^2$ for $\alpha = 2$. However, both larger and smaller values of α result in expected times that scale as polynomials in N, highlighting that not all networks with short average path lengths allow the use of efficient decentralized searches [81].

Adamic et al. proposed to exploit the heterogeneous nature of scale-free networks to significantly speed up breath-first type searches (also called burning algorithms) which typically run in $\mathcal{O}(N)$ time [3,80]. Instead of broadcasting a query to all neighbors who in turn broadcast it further, their algorithm checks whether the target is among its first neighbor, and if not, it passes the query along to the neighbor with the highest degree. This neighbor continues the search in a similar way, and if a dead end is reached, the message is recursively back-tracked until there are no neighbors left to explore. This strategy can improve the performance to $\mathcal{O}(N^{1/2})$ for $\gamma = 3$, and $\mathcal{O}(\log N)$ for $\gamma = 2$ scale-free networks. The method was tested by the authors on the GNUTELLA peer-to-peer network, as well as by Kim et al. on scale-free networks generated by the configuration model and the Barabási–Albert model [80]. (Kim, Yoon, Han and Jeong proposed the strategy independently a few months after Adamic et al. submitted their paper.)

Searchability of social networks has been revisited several times since Milgram's experiment. Killworth and Bernard found that the choice of neighbor to pass the letter on was most often based on common characteristics with the target, such as geographical location and profession [79]. An electronic version of the experiment carried out by Dodds et al. concluded that the search on the e-mail network is also guided by reliance on common "social identities", and the chosen routes do not favor social hubs [49]. Motivated by these results, Watts et al. and Kleingerg independently proposed a searchable social network model in which nodes are grouped in a nested hierarchy of social categories. A link between two nodes is exponentially less likely if their social distance (length of the path along the hierarchical tree of social groups) is large [82,149]. In a simulated Milgram–type experiment, the letter would be passed from a node to a neighbor with the smallest social distance to the target. Such a search performs well for a range of the model parameters, with it's best performance scaling as $\mathcal{O}(\log N)$ [82].

Future Directions

The short account of research advances into dynamics on complex networks presented here is by no means exhaustive. Several fairly large and lively research areas, synchronization, boolean networks, random walks on complex networks, brain networks, models of adaptive or dynamical wiring among them, have all been omitted here.

Future directions in networks research are hard to account for in a few paragraphs. It is believed that further understanding of dynamics on complex networks is the general direction of the field. Along with a shift from pure structural studies to dynamics, there has also been a shift from studies of networks in general and features that are common to most of them to more application-driven studies of increasingly narrow classes of networks. This is not to say that important lessons learned this way do not often carry over from one system to the other, but it shows that enough is known about network characteristics now to tell them apart, to look for distinguishing features as well as universal ones.

A fairly standard toolkit for the assessment of the structural properties of a network has already emerged, making the notion of networks and knowledge about them a useful toolkit in studying the architecture of complex systems. Our hope is that dynamical studies will further extend the toolkit to cover general aspects of the type of events occurring on complex networks, and perhaps lead the way in understanding how function emerges in complex systems.

Bibliography

Primary Literature

1. Adamic LA (1999) The small world web. In: Lecture Notes in Computer Science vol 1696. Springer, New York, pp 443–454
2. Adamic LA, Huberman BA (2000) Power-law distribution of the World Wide Web. Science 287:2115
3. Adamic LA, Lukose RM, Puniyani AR, Huberman BA (2001) Search in power-law networks. Phys Rev E 64(4):046 135
4. Aiello W, Chung F, Lu L (2000) A random graph model for massive graphs. In: Proceedings of the 32nd Annual ACM Symposium on Theory of Computing. ACM, New York, pp 171–180
5. Albert R, Barabási A-L (2000) Topology of evolving networks: local events and universality. Phys Rev Lett 85:5234
6. Albert R, Jeong H, Barabási A-L (1999) Diameter of the World-Wide Web. Nature 401:130–131
7. Albert R, Jeong H, Barabási A-L (2000) Attack and error tolerance of complex networks. Nature 406:378
8. Albert R, Albert I, Nakarado GL (2004) Structural vulnerability of the North American power grid. Phys Rev Lett 69:025 103
9. Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabási A-L (2004) Global organization of metabolic fluxes in the bacterium Escherichia coli. Nature 427(6977):839–843
10. Alon U (2007) Network motifs: theory and experimental approaches. Nat Rev Genet 8:450–461
11. Amaral LAN, Scala A, Barthélémy M, Stanley HE (2000) Classes of small-world networks. Proc Natl Acad Sci USA 97:11 149

12. Arenas A, Danon L, Díaz-Guilera A, Gleiser PM, Guimerà R (2004) Community analysis in social networks. Euro Phys J B 38(2):373–380
13. Baiesi M, Paczuski M (2004) Scale-free networks of earthquakes and aftershocks. Phys Rev E 69(6):066106
14. Balázsi G, Barabási A-L, Oltvai ZN (2005) Topological units of environmental signal processing in the transcriptional regulatory network of Escherichia coli. Proc Natl Acad Sci USA 102(22):7841–7846
15. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512
16. Barabási A-L, Albert R, Jeong H (1999) Mean-field theory for scale-free random networks. Physica A 272:173–187
17. Barabási A-L, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T (2002) Evolution of the social network of scientific collaborations. Physica A 311:590
18. Barrat A, Barthélemy M, Pastor-Satorras R, Vespignani A (2004) The architecture of complex weighted networks. Proc Natl Acad Sci USA 101:3747–3752
19. Barthélémy M (2003) Crossover from scale-free to spatial networks. EuroPhys Lett 63(6):915–921
20. Bender EA, R Canfield E, McKay BD (1997) The asymptotic number of labeled graphs with *n* vertices, *q* edges, and no isolated vertices. J Comb Theor: Series A 80(1):124–150
21. Bianconi G, Barabási A-L (2001) Competition and multiscaling in evolving networks. EuroPhys Lett 54:436
22. Bianconi G, Barabási A-L (2001) Bose-Einstein condensation in complex networks. Phys Rev Lett 86:5632
23. Boguñá M, Pastor-Satorras R, Vespignani A (2003) Absence of epidemic threshold in scale-free networks with degree correlations. Phys Rev Lett 90(2):028 701
24. Boguñá M, Pastor-Satorras R, Diaz-Guilera A, Arenas A (2004) Models of social networks based on social distance attachment. Phys Rev E 70(5):056122
25. Bollobás B, Riordan O (2004) The diameter of a scale-free random graph. Combinatorica 24(1):5–34
26. Bollobás B, de la Vega WF (1982) The diameter of random regular graphs. Combinatorica 2(2):125–134
27. Bollobás B, Riordan O, Spencer J, Tusnády G (2001) The degree sequence of a scale-free random process. Random Struct Algorithms 18:279–290
28. Broder A, Kumar R, Maghoul F, Raghavan P, Rajalopagan S, Stata R, Tomkins A, Wiener J (2000) Graph structure in the web. Comput Netw 33:309–320
29. Broida A, Claffy KC (2001) Internet topology: Connectivity of IP graphs. In: Fahmy S, Park K (eds) Scalability and Traffic Control in IP Networks in Proc SPIE, vol 4526. International Society for Optical Engineering, Bellingham, pp 172ñ–187
30. Callaway DS, Newman MEJ, Strogatz SH, Watts DJ (2000) Network robustness and fragility: Percolation on random graphs. Phys Rev Lett 85:5468
31. Capocci A, Servedio VDP, Colaiori F, Buriol LS, Donato D, Leonardi S, Caldarelli G (2006) Preferential attachment in the growth of social networks: the internet encyclopedia wikipedia. Phys Rev E 74:036 116
32. Chung F, Lu L (2002) The average distances in random graphs with given expected degrees. Proc Natl Acad Sci USA 99:15879–15882
33. Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70(6 Pt 2):066 111
34. Cohen R, Havlin S (2003) Scale-free networks are ultra small. Phys Rev Lett 90:058 701
35. Cohen R, Erez K, ben Avraham D, Havlin S (2000) Resilience of the Internet to random breakdowns. Phys Rev Lett 85:4626–4628
36. Cohen R, Havlin S, ben Avraham D (2003) Efficient immunization strategies for computer networks and populations. Phys Rev Lett 91(24):247 901
37. Colizza V, Barrat A, Barthelemy M, Vespignani A (2006) The role of the airline transportation network in the prediction and predictability of global epidemics. Proc Natl Acad Sci USA 103(7):2015–2020
38. Daley DJ, Kendall DG (1965) Stochastic rumours. IMA J Appl Math 1(1):42–55
39. Dall J, Christensen M (2002) Random geometric graphs. Phys Rev E 66:016 121
40. Danila B, Yu Y, Marsh JA, Bassler KE (2006) Optimal transport on complex networks. Phys Rev E 74(4):046106
41. Danila B, Yu Y, Marsh JA, Bassler KE (2007) Transport optimization on complex networks. Chaos 17(2):026102
42. Davidsen J, Ebel H, Bornholdt S (2002) Emergence of a small world from local interactions: Modeling acquaintance networks. Phys Rev Lett 88(12):128 701
43. de Menezes MA, Barabási A-L (2004) Fluctuations in network dynamics. Phys Rev Lett 92:028 701
44. de Solla Price DJ (1965) Networks of scientific papers. Science 149:510–515
45. de Solla Price DJ (1976) A general theory of bibliometric and other cumulative advantage processes. J Am Soc Inform Sci 27:292–306
46. Demers A, Greene D, Hauser C, Irish W, Larson J, Shenker S, Sturgis H, Swinehart D, Terry D (1987) Epidemic algorithms for replicated database maintenance. In: PODC '87: Proc. 6th Ann. ACM Symposium on Principles of distributed computing. ACM, New York, pp 1–12
47. Dezső Z, Barabási A-L (2002) Halting viruses in scale-free networks. Phys Rev E 65:055 103
48. Di Matteo T, Aste T, Gallegati M (2005) Innovation flow through social networks: productivity distribution in france and italy. Euro Phys J B 47(3):459–466
49. Dodds PS, Muhamad R, Watts DJ (2003) An experimental study of search in global social networks. Science 301(5634):827–829
50. Dorogovtsev SN, Mendes JFF (2000) Scaling behaviour of developing and decaying networks. Europhys Lett 52:33
51. Dorogovtsev SN, Mendes JFF (2001) Effect of the accelerating growth of communications networks on their structure. Phys Rev E 63:025 101
52. Dorogovtsev SN, Mendes JFF (2001) Language as an evolving word web. Proc R Soc London B 268:2603–2606
53. Dorogovtsev SN, Mendes JFF, Samukhin AN (2000) Structure of growing networks with preferential linking. Phys Rev Lett 85:4633–4636
54. Dorogovtsev SN, Goltsev AV, Mendes JFF (2002) Pseudofractal scale-free web. Phys Rev E 65:066 122
55. Ebel H, Mielsch LI, Bormholdt S (2002) Scale-free topology of e-mail networks. Phys Rev E 66:035 103
56. Echenique P, Gómez-Gardeñes J, Moreno Y (2004) Improved routing strategies for internet traffic delivery. Phys Rev E 70(5):056105

57. Erdős P, Rényi A (1959) On random graphs I. Publ Math (Debrecen) 6:290–297
58. Erdős P, Rényi A (1960) On the evolution of random graphs. Publ Math Inst Hung Acad Sci 5:17–61
59. Euler L (1741) Solutio problematis ad geometriam situs pertinentis. Commentarii academiae scientiarum Petropolitanae 8:128–140
60. Faloutsos M, Faloutsos P, Faloutsos C (1999) On power-law relationships of the Internet topology. Comput Commun Rev 29:251–262
61. Ferrer i Cancho R, Solé RV (2001) The small-world of human language. Proc R Soc London B 268:2261–2266
62. Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 40(1):35–41
63. Gerdes SY, Scholle MD, Campbell JW, Balázsi G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS, Bhattacharya A, Kapatral V, DíSouza M, Baev MV, Grechkin Y, Mseeh F, Fonstein MY, Overbeek R, Barabási A-L, Oltvai ZN, Osterman AL (2003) Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. J Bacteriol 185:5673–5684
64. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA 99:7821–7826
65. Goffman W, Newill VA (1964) Generalization of epidemic theory: An application to the transmission of ideas. Nature 204(4955):225–228
66. Goh K-I, Oh E, Jeong H, Kahng B, Kim D (2002) Classification of scale-free networks. Proc Natl Acad Sci USA 99(20):12583–12588
67. Govindan R, Tangmunarunkit H (2000) Heuristics for Internet map discovery. In: Proceedings of IEEE INFOCOM 2000. Nineteenth annual joint conference of the IEEE computer and communications societies, Tel Aviv, Israel, vol 3. IEEE, Piscataway, New Jersey, pp 1371–1380
68. Grassberger P (1983) Critical behavior of the general epidemic process and dynamical percolation. Math Biosci 63(2):157–172
69. Guimerà R, Amaral LAN (2004) Modeling the world-wide airport network. Euro Phys J B 38(2):381–385
70. Guimerà R, Arenas A, Díaz-Guilera A, Giralt F (2002) Dynamical properties of model communication networks. Phys Rev E 66(2):026704
71. Guimerà R, Danon L, Díaz-Guilera A, Giralt F, Arenas A (2003) Self-similar community structure in a network of human interactions. Phys Rev E 68(6):065103
72. Holme P, Kim BJ (2002) Growing scale-free networks with tunable clustering. Phys Rev E 65(2):026107
73. Holme P, Kim BJ (2002) Vertex overload breakdown in evolving networks. Phys Rev E 65(6):066109
74. Jain S, Krishna S (1998) Autocatalytic sets and the growth of complexity in an evolutionary model. Phys Rev Lett 81(25):5684–5687
75. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási A-L (2000) The large-scale organization of metabolic networks. Nature 407:651–654
76. Jeong H, Néda Z, Barabási A-L (2003) Measuring preferential attachment for evolving networks. EuroPhys Lett 61:567
77. Jung S, Kim S, Kahng B (2002) A geometric fractal growth model for scale free networks. Phys Rev E 65:056101
78. Kermarrec AM, Massoulie L, Ganesh AJ (2003) Probabilistic reliable dissemination in large-scale systems. IEEE Trans Parallel Distributed Syst 14(3):248–258
79. Killworth PD, Bernard HR (1978) The reverse small world experiment. Social Netw 1:159–192
80. Kim BJ, Yoon CN, Han SK, Jeong H (2002) Path finding strategies in scale-free networks. Phys Rev E 65(2):027103
81. Kleinberg JM (2000) Navigation in a small world. Nature 406(6798):845
82. Kleinberg JM (2002) Small-world phenomena and the dynamics of information. In: Dietterich TG, Becker S, Ghahramani Z (eds) Proceedings of the 2001 Neural Information Processing Systems Conference. MIT Press, Cambridge
83. Kleinberg JM, Kumar SR, Raghavan P, Rajagopalan S, Tomkins A (1999) The web as a graph: Measurements, models and methods. In: Proc. of the Int. Conf. on Combinatorics and Computing, COCOON'99 Berlin. Springer, Tokyo, p 1
84. Klemm K, Eguíluz VM (2002) Growing scale-free networks with small-world behavior. Phys Rev E 65:057102
85. Krapivsky PL, Redner S (2001) Organization of growing random networks. Phys Rev E 63:66–123
86. Krapivsky PL, Redner S (2002) A statistical physics perspective on web growth. Comput Netw 39:261–276
87. Krapivsky PL, Redner S, Leyvraz F (2000) Connectivity of growing random networks. Phys Rev Lett 85:4629–4632
88. Kumar R, Raghavan P, Rajagopalan S, Tomkins A (1999) Trawling the web for emerging cyber-communities. Comput Netw 31:1481–1493
89. Leone M, Vázquez A, Vespignani A, Zecchina R (2002) Ferromagnetic ordering in graphs with arbitrary degree distribution. Euro Phys J B 28:191–197
90. Liljeros F, Edling C, Amaral L, Aberg Y (2001) The web of human sexual contacts. Nature 411:907–908
91. Ma HW, Zeng AP (2003) The connectivity structure, giant strong component and centrality of metabolic networks. Bioinformatics 19(11):1423–1430
92. Manna SS, Sen P (2002) Modulated scale-free network in euclidean space. Phys Rev E 66(6):066114
93. Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. Science 296:910–913
94. Milgram S (1967) The small-world problem. Psychology Today 2:60–67
95. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298:824–827
96. Moreno Y, Pastor-Satorras R, Vespignani A (2002) Epidemic outbreaks in complex heterogeneous networks. Euro Phys J B 26(4):521–529
97. Moreno Y, Gómez JB, Pacheco AF (2003) Epidemic incidence in correlated complex networks. Phys Rev E 68(3):035103
98. Moreno Y, Nekovee M, Vespignani A (2004) Efficiency and reliability of epidemic data dissemination in complex networks. Phys Rev E 69(5):055101
99. Motter AE (2004) Cascade control and defense in complex networks. Phys Rev Lett 93(9):098701
100. Motter AE, Lai YC (2002) Cascade-based attacks on complex networks. Phys Rev E 66(6):065102
101. Motter AE, de Moura APS, Lai YC, Dasgupta P (2002) Topology of the conceptual network of language. Phys Rev E 65:065102

102. Myers CR (2003) Software systems as complex networks: Structure, function, and evolvability of software collaboration graphs. Phys Rev E 68:046116
103. Newman MEJ (2001) Clustering and preferential attachment in growing networks. Phys Rev E 64:025102(R)
104. Newman MEJ (2001) The structure of scientific collaboration networks. Proc Natl Acad Sci USA 98:404–409
105. Newman MEJ (2002) Assortative mixing in networks. Phys Rev Lett 89:208701
106. Newman MEJ (2002) Spread of epidemic disease on networks. Phys Rev E 66(1):016128
107. Newman MEJ (2003) Mixing patterns in networks. Phys Rev E 67:026126
108. Newman MEJ (2004) Fast algorithm for detecting community structure in networks. Phys Rev E 69(6 Pt 2):066133
109. Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74(3):036104
110. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2):026113
111. Newman MEJ, Watts DJ (1999) Renormalization group analysis of the small-world network model. Phys Lett A 263:341–346
112. Newman MEJ, Strogatz SH, Watts DJ (2001) Random graphs with arbitrary degree distributions and their applications. Phys Rev E 64(2):026118
113. Ohira T, Sawatari R (1998) Phase transition in a computer network traffic model. Phys Rev E 58(1):193–195
114. Palla G, Derenyi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043):814–818
115. Pastor-Satorras R, Vespignani A (2001) Epidemic dynamics and endemic states in complex networks. Phys Rev E 63(6):066117
116. Pastor-Satorras R, Vespignani A (2001) Epidemic spreading in scale-free networks. Phys Rev Lett 86:3200–3203
117. Pastor-Satorras R, Vespignani A (2002) Epidemic dynamics in finite size scale-free networks. Phys Rev E 65(3):035108
118. Pastor-Satorras R, Vespignani A (2002) Immunization of complex networks. Phys Rev Lett 65:036104
119. Pastor-Satorras R, Vázquez A, Vespignani A (2001) Dynamical and correlation properties of the Internet. Phys Rev Lett 87:258701
120. Rao F, Caflisch A (2004) The protein folding network. J Mol Biol 342:299–306
121. Ravasz E, Barabási A-L (2002) Hierarchical organization in complex networks. Phys Rev E 67:026122
122. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási A-L (2002) Hierarchical organization of modularity in metabolic networks. Science 297:1551–1555
123. Ravasz E, Gnanakaran S, Toroczkai Z (2007) Network structure of protein folding pathways. arXiv:0705.0912v1
124. Redner S (1998) How popular is your paper? An empirical study of the citation distribution. Euro Phys J B 4:131–135
125. Redner S (2004) Citation statistics from more than a century of Physical Review. arXiv:physics/0407137v2
126. Scala A, Amaral LAN, Barthélémy M (2001) Small-world networks and the conformation space of a short lattice polymer chain. Europhys Lett 55(4):594–600
127. Serrano MA, Boguñá M (2003) Topology of the world trade web. Phys Rev E 68:015101
128. Shen-Orr S, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of E. coli. Nat Genet 31:64–68
129. Sigman M, Cecchi GA (2002) Global organization of the Wordnet lexicon. Proc Natl Acad Sci USA 99(3):1742–1747
130. Solé R, Pastor-Satorras R, Smith E, Kepler T (2002) A model of large-scale proteome evolution. Adv Compl Syst 5:43–54
131. Solé RV, Valverde S (2001) Information transfer and phase transitions in a model of internet traffic. Physica A 289(3–4):595–605
132. Song C, Havlin S, Makse HA (2005) Self-similarity of complex networks. Nature 433(7024):392–395
133. Song C, Havlin S, Makse HA (2006) Origins of fractality in the growth of complex networks. Nat Phys 2(4):275–281
134. Sreenivasan S, Cohen R, Lopez E, Toroczkai Z, Stanley HE (2007) Structural bottlenecks for communication in networks. Phys Rev E 75(3):036105
135. Tadić B (2001) Dynamics of directed graphs: The world-wide web. Physica A 293:273–284
136. Tadić B, Rodgers GJ (2002) Packet Transport on Scale Free Networks. Adv Compl Syst 5:445–456
137. Tadić B, Thurner S (2004) Information super-diffusion on structured networks. Physica A 332:566–584
138. Thieffry D, Huerta AM, Perez-Rueda E, Collado-Vides J (1998) From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. Bioessays 20(5):433–440
139. Toroczkai Z, Bassler KE (2004) Network dynamics: Jamming is limited in scale-free systems. Nature 428:716
140. Toroczkai Z, Kozma B, Bassler KE, Hengartner NW, Korniss G (2004) Gradient networks. arXiv:cond-mat/0408262v1
141. Tyler JR, Wilkinson DM, Huberman BA (2003) Email as spectroscopy: Automated discovery of community structure within organizations. In: Huysman M, Wenger E, Wulf V (eds) Communities and Technologies. Proceedings of the First International Conference on Communities and Technologies. Kluwer, Norwell MA, pp 81–96
142. Valverde S, Solé RV (2002) Self-organized critical traffic in parallel computer networks. Physica A 312(3–4):636–648
143. Valverde S, Solé RV (2004) Internet's critical path horizon. Euro Phys J B 38(2):245–252
144. Vázquez A, Weigt M (2003) Computational complexity arising from degree correlations in networks. Phys Rev E 67(2):027101
145. Vázquez A, Flammini A, Maritan A, Vespignani A (2003) Modelling of protein interaction networks. ComPlexUs 1:38–44
146. Vogels W, van Renesse R, Birman K (2003) The power of epidemics: robust communication for large-scale distributed systems. SIGCOMM Comput Commun Rev 33(1):131–135
147. Wasserman S, Faust K (1994) Social Network Analysis. Cambridge University Press, Cambridge
148. Watts DJ, Strogatz SH (1998) Collective dynamics of small-world networks. Nature 393:440–442
149. Watts DJ, Dodds PS, Newman MEJ (2002) Identity and search in social networks. Science 296:130
150. Wilkinson DM, Huberman BA (2004) A method for finding communities of related genes. Proc Natl Acad Sci USA Suppl 101(1):5241–5248
151. Xulvi-Brunet R, Sokolov IM (2002) Evolving networks with disadvantaged long-range connections. Phys Rev E 66(2):026118

152. Yook SH, Jeong H, Barabási A-L (2003) Modelling the Internet's large-scale topology. Proc Natl Acad Sci USA 99:13382–13386
153. Zanette DH (2001) Critical behavior of propagation on small-world networks. Phys Rev E 64(5):050901
154. Zhao L, Lai YC, Park K, Ye N (2005) Onset of traffic congestion in complex networks. Phys Rev E 71(2):026125

Books and Reviews

Barabási A-L (2002) Linked: The New Science of Networks. Perseus Publishing, Cambridge
Pastor-Satorras R, Rubi M, Diaz-Guilera A (eds) (2003) Statistical Mechanics of Complex Networks, 625, Lecture Notes in Physics. Springer, Berlin
Mendes J, Oliveira JG, Abreu FV, Povolotsky A, Dorogovtsev SN (eds) (2005) Science of Complex Networks: From Biology to the Internet and WWW. AIP conference proceedings, CNET 2004, Aveiro, Portugal, vol 776
Newman MEJ, Barabási A-L, Watts DJ (eds) (2003) The Structure and Dynamics of Complex Networks. Princeton University Press, Princeton
Ben-Naim E, Frauenfelder H, Toroczkai Z (eds) (2005) Complex Networks, 650, Lecture Notes in Physics. Springer, Secaucus
Bornholdt S, Schuster HG (eds) (2002) Handbook of graphs and networks: from the genome to the internet. Wiley-VCH, Berlin
Bollobás B (1985) Random Graphs. Academic Press, London
Watts DJ (1999) Small Worlds: The Dynamics of Networks between Order and Randomness. Princeton University Press, Princeton
Albert R, Barabási AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74(1):47–97
Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256
Dorogovtsev SN, Mendes JFF (2002) Evolution of networks. Adv Phys 51:1079
Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex networks: Structure and dynamics. Phys Rep 424(4–5):175–308

Neuro-fuzzy Control of Autonomous Robotics

Petru Emanuel Stingu, Frank L. Lewis
Automation & Robotics Research Institute,
University of Texas at Arlington, Fort Worth, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Control Hierarchy
Map-Based Navigation
Behaviors
Future Directions
Bibliography

Glossary

Robot The word 'robot' was introduced by the Czech playwright Capek in his 1920 play Rossum's Universal Robots. The word 'robota' in Czech means simply 'work'. Although there is no definition accepted by everyone, in this chapter a robot is considered to be a human-built machine that is mobile, can sense and interact with the environment, and has the necessary intelligence in order to handle unforeseen circumstances autonomously. Most important than all, it has to do a useful task.

Autonomy Independence of control, self-sufficiency. Applied to robots, it implies the ability of the robot to find solutions by itself to the various problems that might appear while completing the assigned task.

Fuzzy logic The idea of fuzzy logic was first advanced by Dr. Lotfi Zadeh of the University of California at Berkeley in the 1960s. It came from the fact that natural language can not be easily translated in the absolute terms of 0 and 1. Fuzzy logic includes 0 and 1 as extreme cases of truth (that are representations of certainty or facts), but also includes the various states of truth in between (partial truth). As an example, using binary logic it can be said that "the target is on the left side of the robot" or "the target is not on the left side of the robot", while using fuzzy logic a more precise description can be given, like "the target is 20% on the left side of the robot".

Neural networks A neural network can be described as a system composed of many simple processing elements operating in parallel whose function is determined by network structure, connection strengths, and the processing performed at computing elements or nodes. Some of its advantages are: non-linear mappings, adaptation and learning, ease of implementation and self-organization.

Control When talking about controlling systems, control refers to the process of changing the input of the system such that its output reaches a desired value. Most of the time, the control is done in a closed loop, where the output of the system is continuously compared to the reference and the necessary control command is applied to the inputs of the system in order to reduce the error at the output.

Behavior Small independent decision-making process that fully implements a control policy for one specific sub-task. Usually multiple behaviors coexist and are enabled or disabled by an arbiter, depending on which is useful in a particular situation.

Mapping Mapping is split into two main processes that are dependent on each other: map-learning and localization. The first has to store the information from the robot sensors into a suitable internal representation (map). The latter has to estimate the position of the various objects on the map. Building the map needs localization, but in the same time localization requires a map.

Path planning The process where, given a complete description of the geometry of a robot and the static environment populated with obstacles, a collision-free path must be found such that the robot can move from an initial position and orientation to a goal position and orientation.

Definition of the Subject

Autonomous robots are robots which can perform desired tasks in unstructured environments without requiring continuous human guidance. Most of the times, the dynamics of the robot itself can be described analytically. Unfortunately, in many robotic applications, it is difficult if not impossible to obtain a precise mathematical model of the environment and its interaction with the robot through actuators and sensors. The lack of complete and precise knowledge about the environment limits the applicability of conventional control system design to the domain of autonomous robotics. Some of the requirements for a robot to successfully achieve autonomy are the possibility to acquire knowledge about the environment and itself, to reason under uncertainty and to have learning capabilities in order to adapt to the environment based on accumulated experience.

Efficient control algorithms for autonomous robots should imitate the way humans are operating manned or similar vehicles. When making decisions, humans tend to work with vague or imprecise concepts that can often be expressed linguistically. Lotfi Zadeh has proposed one way to model this decision making process by introducing the fuzzy set theory in the field of control [36]. Fuzzy logic is particularly suited for problems in which the data, the objectives and the constraints are too complex or too ill-defined to admit a precise mathematical analysis.

Neural networks were developed as an attempt to realize mathematical models of brain-like systems. The key advantage is their ability to learn from examples instead of requiring an algorithmic development from the designer. They can generalize to new situations. Once a neural network has been trained for a set of data, it can interpolate and produce answers for the cases not present in the training set.

While fuzzy logic can be used to represent knowledge in a human-readable form and to use it for reasoning, neural networks allow adaptation and learning in dynamic environments under varying conditions. Neuro-fuzzy techniques combine the advantages of both methods by having neural networks adapt the knowledge base of the fuzzy logic systems or fuzzy systems tune the weights of neural networks. These relatively simple control methods can be used to successfully implement complex intelligent autonomous robots, robust to uncertainties in their own model, in the environment and in the readings from the sensors.

Introduction

Despite the differences between neural networks and fuzzy logic systems, they can actually be unified at the level of the universal function approximator. Both of them define a nonlinear function $y = f(x)$ from inputs to outputs and if designed properly satisfy the "universal approximation property" [22,32]. Then, for any continuous function $\psi(x)$ defined on a closed and bounded set and an arbitrary number $\varepsilon > 0$, there exists a neural network or a fuzzy system $f(x)$ such that

$$\sup_x |f(x) - \psi(x)| < \varepsilon$$

This property does not say how to build the neural net or the fuzzy system. It simply shows that with the proper structure and with enough tuning it is possible to approximate a continuous function with an error that can be made as small as desired.

The multilayer perceptron should be viewed as a nonlinear network whose non-linearity can be tuned by changing the weights, biases and the parameters of the activation functions. The fuzzy logic system is also a tunable non-linearity whose shape can be tuned by changing the membership functions. In both cases, it is possible to use gradient methods for tuning. The back-propagation training method [33] is well-known from the neural networks field and inspired gradient training of fuzzy systems.

Some radial basis function neural networks are functionally equivalent to some standard fuzzy systems in the sense that given the same inputs, they will produce the same outputs. This can be shown on a normalized RBF neural net that has the same number of neurons on the hidden layer as the number of rules in the fuzzy system with product inference and centroid defuzzification. A simple example for a two-input fuzzy system is shown in Fig. 1. A RBF neural net structure that can process the information in a similar way is shown in Fig. 2.

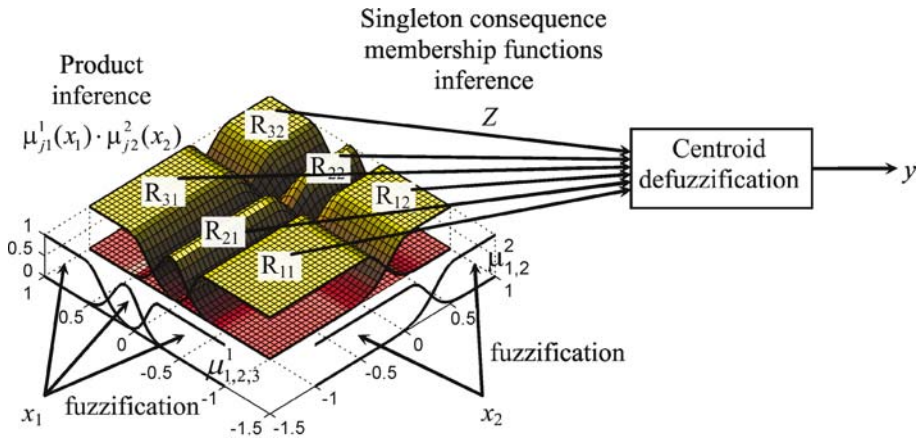

Neuro-fuzzy Control of Autonomous Robotics, Figure 1
Example of a two-variable fuzzy logic system

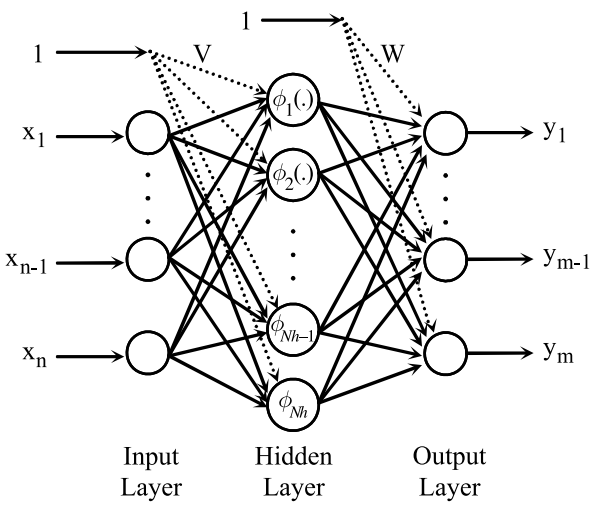

Neuro-fuzzy Control of Autonomous Robotics, Figure 2
Multi-layer neural network

The output of a fuzzy system with n input variables, possible different number of membership functions for each input variable, product inference, control representative values $z_{i,j1,j2,\ldots,jn}$ and centroid defuzzification is the following:

$$y = \frac{\sum\limits_{j1,j2,\ldots,jn} z_{i,j1,j2,\ldots,jn} \prod\limits_{k=1}^{n} \mu^k_{jk}(x_k)}{\sum\limits_{j1,j2,\ldots,jn} \prod\limits_{k=1}^{n} \mu^k_{jk}} \tag{1}$$

The kth output of an unnormalized RBF neural network with N_h neurons on the hidden layer is

$$y_k = \sum_{i=1}^{n} w_{ik} \cdot \phi_i(\mathbf{x}) \tag{2}$$

while the output of a multi-layer neural network is

$$\mathbf{y} = W^{\mathrm{T}} \boldsymbol{\varphi}(V^{\mathrm{T}}\mathbf{x}). \tag{3}$$

Because of their properties of learning, recall, function approximation, generalization, classification, association, pattern recognition and clustering, neural networks and fuzzy logic systems can be used to solve a large set of problems. They are successfully used in all types of autonomous robots. Some examples are:

- ground robots: indoor and outdoor
- aquatic robots: surface and submersible
- aerial robots: fixed-wing and rotary-wing

Control Hierarchy

We have seen that autonomous robots can borrow traits from humans. Fuzzy logic copies the way they represent knowledge and the mechanisms of logical reason. Neural networks implement a simple model for the low-level organization and the physiological mechanisms relating to information processing in the nervous system. It is a well-known fact that nature has found the optimal solutions for every imaginable problem, over thousands and thousands of years of refinements. Humans have become the most autonomous of all living beings. They can navigate all around the Earth and even in the extraterrestrial space, they gather knowledge and use it to get even further. Such a successful model should also suggest the control hierarchy to be used on autonomous robots. The organization of the human nervous system involved in movement can provide invaluable information on how to design an efficient hierarchical control architecture.

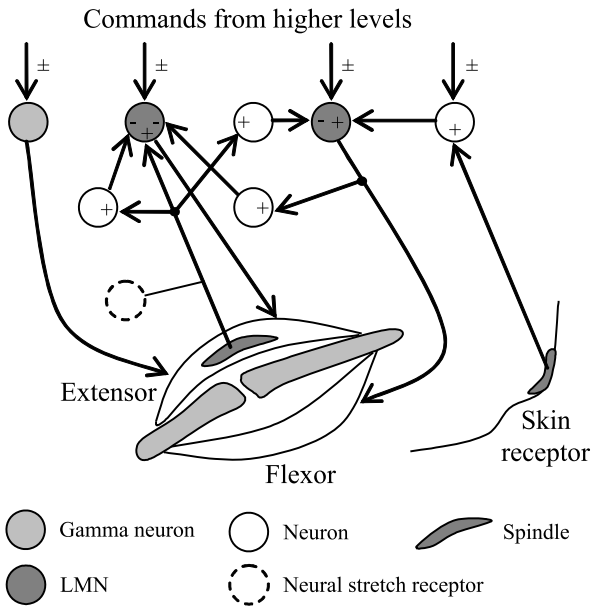

Neuro-fuzzy Control of Autonomous Robotics, Figure 3
Joint with the principal neurons involved in reflex control

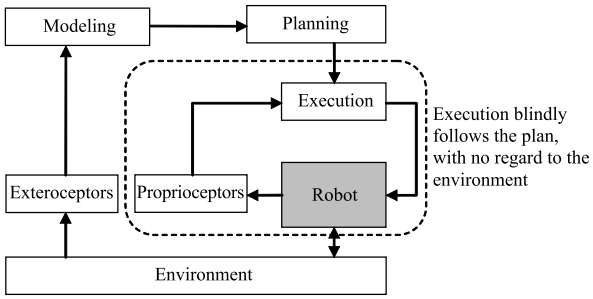

Neuro-fuzzy Control of Autonomous Robotics, Figure 4
The low-level in the hierarchical architecture blindly executes the plan

Consideration of the vast and complex system of structures and pathways involved in movement will begin at its lowest end, the spinal cord, where more basic motor control is localized. The most important human interoceptive reflex is the myotatic reflex [4], which originates from the neuro-muscular fibers. Its principal function is maintaining the joint position fixed and compensating external noise. In Fig. 3, suppose that a load is applied to the joint. This will flex the joint, causing the stretching of the extensor muscle and also stretching of the spindle. This will increase the output of the neural stretch receptor neuron, which increases the output of the local motor neuron (LMN). The resulting increase in the contraction force will compensate the load. This local feedback allows the higher system to ignore the fluctuation in contraction required to maintain a certain joint extension. On a robot, the similar function is done by the low-level controllers for the motors. For example, they receive a reference speed and they have to keep the output speed equal with the reference, independent of the torque disturbances. Neural nets and fuzzy logic speed regulators have been successfully applied for motor control, yielding better results than normally used PID controllers.

The conventional hierarchical control architecture that has been extensively used for mobile robots is shown in Fig. 4. The robot builds the model of itself using two types of sensors. The proprioceptive sensors, like shaft encoders, give information about the internal state of the robot. The exteroceptive sensors, like a sonar, provide information about the state of the environment. Using this information, a planning algorithm generates a plan that will perform the given task in the given environment. The instructions from the plan are blindly executed by the lower-level motor control layer. This layer uses proprioceptors for local feedback, but does not monitor the environment. If an obstacle is suddenly detected by the exteroceptors, it takes a long time for the robot to react. This happens because the stimulus has to pass through the higher layers first. Modeling and planning usually involve a lot of computation and take a long time. The response of the overall system to a new configuration of the environment is very slow. That is why robots implemented using this control hierarchy have to move with a very low speed in order to be able to avoid obstacles.

A solution to the above problem can be searched in the way humans react to some external stimuli, for example to pain. A simple spinal reflex can be traced in Fig. 3. When the pain receptor in the skin is excited, it fires a neuron in the LMN system, which in turn fires the LMN driving the flexor muscle. This operation removes the respective part of the body from danger, in a very fast and straightforward manner. There is no reaction expected from the higher layers of the nervous system before the motor neurons are fired. Still, the information from the skin receptor reaches the higher layers and a more intelligent measure is taken after a certain delay. The overall system now has a fast response to the environment. Reasoning about the necessary action can still be done and the movement can be corrected by the central nervous system through the direct inputs to the motor neurons. On mobile robots, the same type of reaction can be achieved by using low-level behaviors implemented in the execution layer. A simple but very important modification [25] has to be done to the structure in Fig. 4. The execution layer will not follow the planner commands blindly anymore, but will also receive

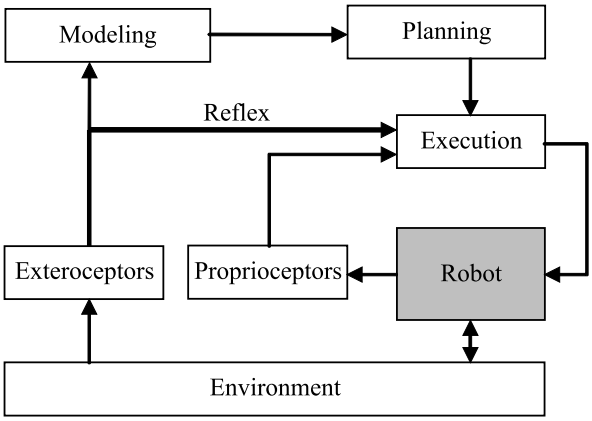

Neuro-fuzzy Control of Autonomous Robotics, Figure 5
The low-level execution layer can implement behaviors based on the sensed environment

information about the environment from the exteroceptive sensors, as shown in Fig. 5. The behaviors (that can be easily implemented using fuzzy logic) will have a fast reaction to the environment, because it is included now in the low-level feedback loop. The planner will only have to select the right behavior or to combine behaviors corresponding to the current goal and the current environment. The rapidity of decisions is not critical anymore for the higher levels.

From the overall organization of the human motor nervous system [10], shown in Fig. 6, and from the previous examples, a general principle can be extracted. First of all, (complete) sensorial data arrives to all the layers of control, starting from the lowest (the spinal cord) and up to the highest (the cerebral cortex). Each layer uses what information it needs and has certain autonomy in taking decisions, independent of the higher layer in the hierarchy. The latter will eventually correct or adjust the action initiated by the lower layers. The reaction speed is the fastest for the simple layers close to the effectors (of the order of hundreds of milliseconds) and decreases as one rises through the hierarchy and where structures become more complex.

In [3], Rodney A. Brooks has proposed a similar layered control architecture for mobile robots, organized into levels of competence. It is called the *subsumption architecture* and is based on incorporating lower levels of functionality into more general levels (Fig. 7). Each level of competence includes as a subset each earlier level of competence. Practically the higher layers impose additional restrictions on the behaviors implemented in the lower layers. The lowest-level layer is layer 0. It is implemented and tested by its own. Layer 1 is then built on top of layer 0. It can read information from layer 0 and can inject information into the internal interfaces of layer 0, suppressing the normal data flow. Level 1 of competence is achieved using layer 1 and the help of layer 0. The latter continues to run unaware of the fact that layer 1 interferes with its data paths when it wants to take control. Many layers of competence can be added one on top of the other using this mechanism. The functionality of the different layers for a mobile robot spawns across three main levels, as stated by Stephanou in [28] and by Lefebvre and Saridis in [16]: organization, coordination and execution.

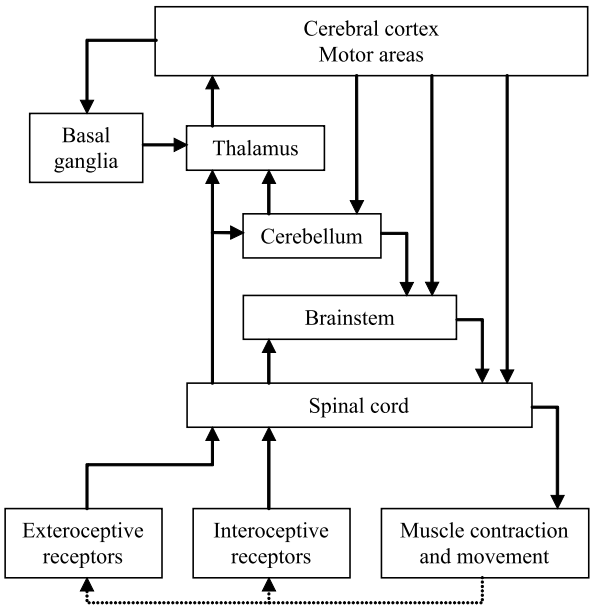

Neuro-fuzzy Control of Autonomous Robotics, Figure 6
Summary of motor control in the human nervous system

The lower levels in the hierarchy have very fast reactions to sensor readings. They don't usually do complex processing on this information. The higher levels usually do sensor fusion and apply special algorithms to extract more abstract information about the environment. The decisions are taken with longer delays and with a much lower rate. It can be observed that there is a transition between almost continuous-time control at the lower levels, with high sample rates, and relatively slow discrete events at the highest levels. This transition can be exemplified with the difference between a fast PI controller for the wheels motors, and a discrete event controller for the implementation of the main plan of action.

Map-Based Navigation

Any approach to control a dynamic system needs to use a model of the system to be controlled. In the case of a mobile robot, the system consists of the robot itself plus the

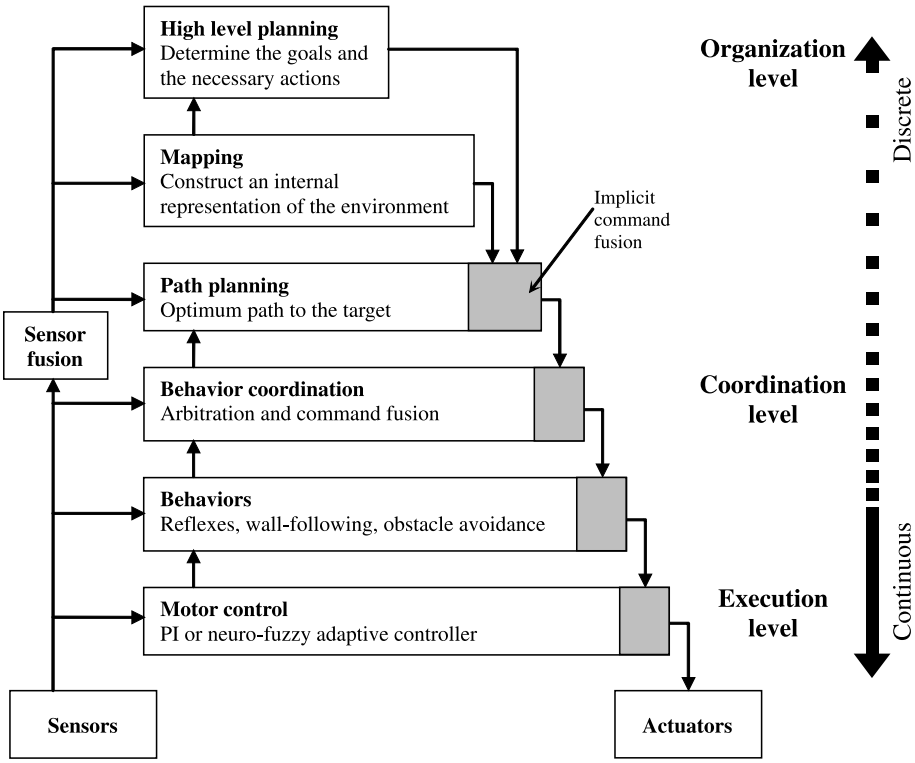

Neuro-fuzzy Control of Autonomous Robotics, Figure 7
Subsumption architecture for robot control

environment in which it operates. It is easy to obtain the model of the robot on its own, but unfortunately the situation is different if we consider the robot to be embedded in a real-world, unstructured environment, that is characterized by uncertainty which is difficult to model or to quantify. The mapping problem is generally regarded as one of the most important problems in the pursuit of building truly autonomous mobile robots.

Map-based navigation calls upon three processes:

- Map learning – the process of memorizing the data acquired by the robot during exploration in a suitable representation.
- Localization – the process of deriving the current position of the robot within the map.
- Path planning – the process of choosing a sequence of actions in order to reach a goal, starting from the current position.

Map learning and localization are interdependent. Building a map requires the position of the robot to be estimated relative to the incomplete map while localization requires that the map exists. Path planning is rather independent of the other two and takes place once the map and the robot position are already available.

Various map representations have been used in the robotics literature. They are adequate or not depending on the task and the characteristics of the robot and the environment.

Artificial Potential Fields

Artificial potential field methods for obstacle avoidance have gained increased popularity among researchers in the field of mobile robots. The idea of imaginary forces acting on a robot has been suggested by Andrews and Hogan [1] and Khatib [15]. The approach used to generate the artificial potential fields is to have obstacles exert repulsive forces onto the mobile robot, while the target applies an attractive force to it. The resultant force $\boldsymbol{F}$ determines the direction and speed of travel for the robot. The method is simple and elegant, yielding acceptable results with simple and quick implementations and without requiring many refinements. The moving robot can come in a variety of shapes and sizes. To simplify the development of the solution, the robot can be represented as a point while the obstacles and physical workspace boundaries are transformed by increasing their size with a value related to the dimension of the robot.

The potential is a scalar field whose negative gradient is a vector field of conservative forces. Any point of the force field will represent the resultant force obtained by summing the effects of the attraction force of the target and the repulsion forces of the obstacles:

$$\boldsymbol{F}(\boldsymbol{r}) = \boldsymbol{F}_{target}(\boldsymbol{r}) + \sum_{i=1}^{n} \boldsymbol{F}_{\mathrm{obs}_i}(\boldsymbol{r}) = -\nabla V(\boldsymbol{r}) \tag{4}$$

For mobile robots applications, the potential fields are usually used in a 2-dimensional space:

$$\begin{aligned} F_x(x,y) &= -\frac{\partial V(x,y)}{\partial x} \\ F_y(x,y) &= -\frac{\partial V(x,y)}{\partial y} \end{aligned} \tag{5}$$

The trajectory is generated continuously for any position (x, y) of the robot following the direction of the steepest descent. It can be considered an optimization problem that searches for the point of minimum potential.

Let's consider the model of a car-like vehicle with no mass. It is a nonholonomic system, so it will most likely deviate from the ideal trajectory resulted from the potential field. However, the potential field implicitly provides a feedback mechanism. The $\boldsymbol{F}$ force is generated for any position of the robot in such a way that it will return close to the ideal trajectory. The robot (Fig. 8) has the following dynamics:

$$\begin{aligned} \dot{x} &= v_{\mathrm{T}} \cos\phi \ \cos\theta \\ \dot{y} &= v_{\mathrm{T}} \cos\phi \ \sin\theta \\ \dot{\theta} &= \frac{v_{\mathrm{T}}}{L} \sin\phi \end{aligned} \tag{6}$$

with (x, y) the position, θ the heading angle, v_T the wheel speed, L the wheel base, and ϕ the steering angle.

A potential field corresponding to the following environment is generated:

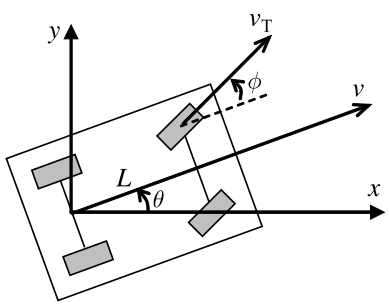

Neuro-fuzzy Control of Autonomous Robotics, Figure 8
Model of a car-like robot

- Goal: a constant force will attract the robot to the target.

$$\begin{aligned} &(x_{\mathrm{G}}, y_{\mathrm{G}}) = (10, 10) \\ &r = \sqrt{(x_{\mathrm{G}} - x)^2 + (y_{\mathrm{G}} - y)^2} \\ &F_{\mathrm{G}x}(x,y) = K_{\mathrm{G}} \frac{x_{\mathrm{G}} - x}{r}, \quad F_{\mathrm{G}y}(x,y) = K_{\mathrm{G}} \frac{y_{\mathrm{G}} - y}{r} \end{aligned}$$

- Circular obstacles: a force similar to the electrostatic force will repel the robot from the obstacles.

$$\begin{aligned} &(x_1, y_1) = (3, 3) \text{ for the first obstacle} \\ &(x_2, y_2) = (7, 2) \text{ for the second obstacle} \\ &F_{ix}(x,y) = -K_i \frac{x_i - x}{(\max\{(r - a_i), b_i\})^2}, \\ &F_{iy}(x,y) = -K_i \frac{y_i - y}{(\max\{(r - a_i), b_i\})^2}, \end{aligned}$$

where r is the distance between the robot and the goal or the obstacle, a_i is the radius of the obstacle and b_i limits the relative height of the obstacle. The total force that acts on the robot is

$$\begin{aligned} F_x &= F_{\mathrm{G}x} + F_{1x} + F_{2x} \\ F_y &= F_{\mathrm{G}y} + F_{1y} + F_{2y} \end{aligned}$$

and the angle of the force as seen from the robot is

$$\alpha = \tan^{-1}\left(\frac{F_y}{F_x}\right)$$

with corrections for the correct quadrant.

The initial position for the robot is $(x_0, y_0) = (0, 0)$ and the initial heading angle is $\phi_0 = \pi/6$. The speed v_T is often kept constant. The steering angle bay be generated by the controller based on the difference between the heading angle θ and the angle α of the resultant force acting on the robot (Fig. 9). In this example, a proportional controller is implemented:

$$\phi = K(\theta - \alpha)$$

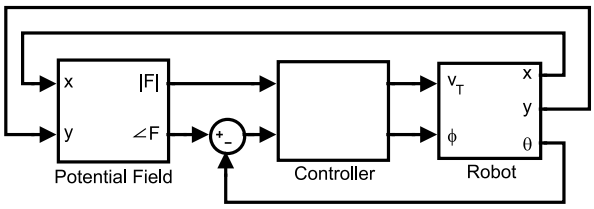

Neuro-fuzzy Control of Autonomous Robotics, Figure 9
Closed-loop control based on the potential field

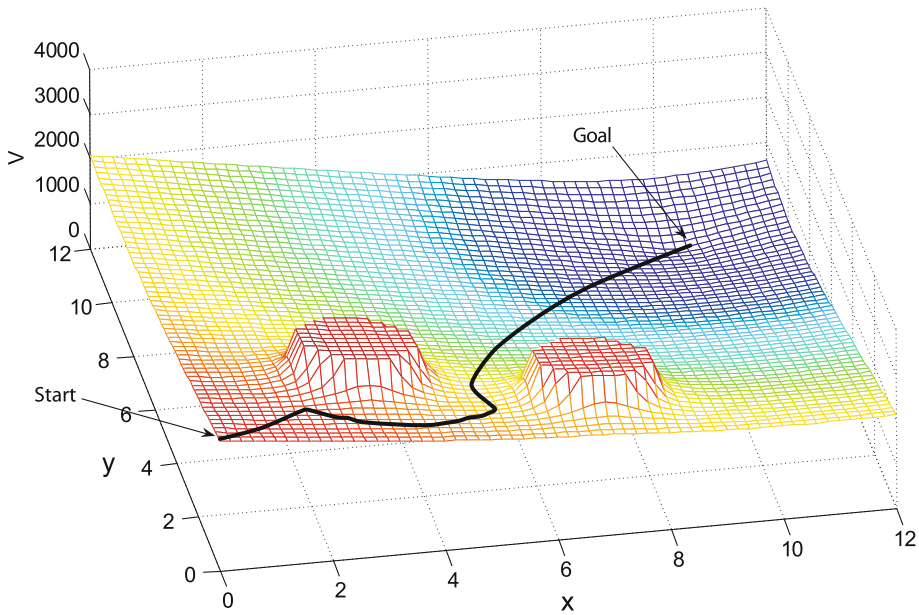

Neuro-fuzzy Control of Autonomous Robotics, Figure 10
The trajectory of the robot using the potential field

Problems Associated with Potential Field Methods

It can be seen from the potential field representation (Fig. 10) that the robot will follow the direction of the steepest descent, which is the direction of the resultant force. There are a few issues that can prevent the robot to reach the target:

- Trap situations due to local minima
- No passage between closely spaced obstacles
- Oscillations in the presence of obstacles or narrow passages

Some of the problems can be solved by modifying the height or dimension of the obstacles, by changing the type of force used to represent them or by using an intelligent controller in the structure from Fig. 9.

Fuzzy Logic and Potential Fields

A good example of combining fuzzy logic and potential fields for path planning can be found in the work of Valavanis et al. [30] that we will present in the following pages. They have implemented a two-layered fuzzy logic inference engine and a discrete type of potential field for real-time mobile robot navigation in a 2-D dynamic environment. The first layer of the fuzzy logic inference engine performs sensor fusion from sensor readings into a fuzzy variable, *collision*, providing information about possible collisions in four directions: *front*, *back*, *left* and *right*. The second layer guarantees collision avoidance with dynamic obstacles while following the trajectory generated by the Electrostatic Potential Field (EPF). The main idea is to combine planned and reactive behavior. Given a 2-D environment (with initial information from potentially existing environment a priori maps and on-line sonar sensor data), the EPF plans the initial trajectory and starts executing it. Once the object detection module (working in parallel with the EPF) detects using sensor readings a "high collision possibility", it forces the motion control module to "forget" the initial EPF path, take corrective actions in terms of robot steering and robot speed to avoid the collision, until new sensor readings dictate a "low" or "not-possible" collision possibility. Then, the motion control module takes into account the initial trajectory as computed at this time instant by the EPF planner. The EPF planner is invoked every time the environment map is updated.

The solution to the navigation problem may be compared to the flow of electric current within a sheet of conducting material. Obstacles are mapped into a discrete resistor network. The path of minimum resistance within the circuit corresponds with a path of minimum occupancy within the environment. The algorithm used to create the natural potential field follows three steps:

- Obtain an occupancy map of the environment
- Create the resistor network

- Calculate the values of the resistors in the network to obtain the potential field

Each cell in the occupancy map is replaced with a set of eight resistors (Fig. 11), each connected at a central point. The only exceptions are the cells on the outside edges and corners with five and respectively three resistors connected. The resistors values depend on the corresponding cell in the occupancy map. A minimum resistance path in the network determines a maximum potential drop. Reversing the mapping generates an optimal path in the environment corresponding to a minimum occupancy path. The path always begins at the highest potential (initial vehicle position), and ends at the lowest potential (final destination).

Consider an environment map which contains obstacles of various shapes and sizes. The initial position of the robot is q_0 and the destination point is q_f. Assume a square, bounded region centered about q_0 which includes q_f and can be divided into an $n \times n$ grid, X. The grid is discretely represented by the occupancy matrix O, where the value of each entry is the percentage of the area of the grid cell occupied by obstacles of the environment map. To determine a desired direction of travel from the EPF, a vector is associated with each cell connected to the cell containing q_0 with magnitude equal to the amount of current flowing through the specified branch. Figure 12 shows the vectors created through a solution of the resistor network.

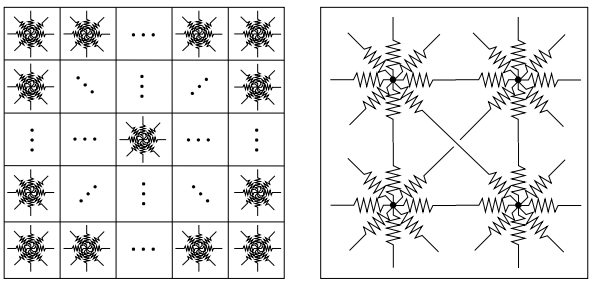

Neuro-fuzzy Control of Autonomous Robotics, Figure 11
An n by n resistor network and node detail

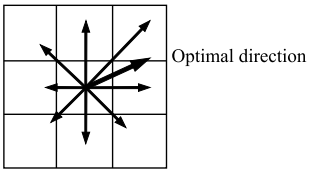

Neuro-fuzzy Control of Autonomous Robotics, Figure 12
Vector representation of the EPF solution

The magnitudes of the current along each branch are shown at the end of the associated vector. The geometric sum of these forces results in a single vector pointing in the direction of the optimal path, as shown in Fig. 12.

The control of the robot is obtained using a two-layered FL interference engine. The first layer handles obstacle detection and performs sensor data fusion from sonar readings. There are two input fuzzy variables: *sensor_direction* and *sensor_distance*. The variable *sensor_direction* has four different values that describe the sensor's membership in four cardinal relative directions: *front_collision*, *left_collision*, *back_collision*, and *right_collision*. Each of the collision values is represented as a fuzzy variable with values *not_possible*, *possible*, and *high*. The second layer has three inputs. The first receives the output of the first inference, representing the immediate collision possibility. The second input receives the output of the artificial potential field and the third receives the speed of the robot. There are two outputs that are used to control the movement of the robot: change of speed and steering.

A mobile robot, the Nomad 200, with a ring of 16 sonar sensors that cover a full 360 circle is used for experimental purposes. The sonar sensors are fixed to the body of the robot. Therefore, *sensor_direction* is a function only of the number of the sonar. Sonar S1 always belongs to *front_collision* with weight 1.0 and sonar S2 always belongs to *front_collision* with weight 0.8 and *left_collision* 0.2, as shown in Fig. 13.

To simplify the implementation, *sensor_direction* has not been used but instead each sonar has been assigned to at least one but no more than two values of the variable collision. The rule base of the obstacle detection controller contains rules of the following two types:

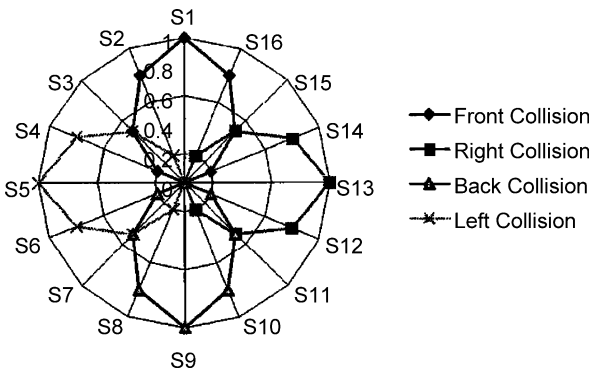

Neuro-fuzzy Control of Autonomous Robotics, Figure 13
Sonar arrangement and relative importance to collision detection [30]

R1) IF d_i is $LD^{(k)}$ THEN c_j is $LC^{(k)}$;
R2) IF d_i is $LD^{(k)}$ AND d_{i-1} is $LD^{(k)}$ AND d_{i+1} is $LD^{(k)}$ THEN c_j is $LC^{(k)}$

where

- k is the rule number,
- d_i represents the readings of sensor i,
- $LD^{(k)}$ is a linguistic value of the term set $D = \{$*Close, Near, Far*$\}$,
- c_j is the collision of type j ($j \in \{$*Front, Left, Right, Back*$\}$) and
- $LC^{(k)}$ is a linguistic value of the term set $C = \{$*Not-possible, Possible, High*$\}$.

The whole rule-base is presented in [13]. Some of the rules for the collision of type *j=front* are

IF d_1 is *Close* THEN *front* collision is *High*.
IF d_1 is *Far* AND d_2 is *Far* AND d_{16} is *Far* THEN *front* collision is *Not-Possible*.
IF d_2 is *Near* THEN *front* collision is *Possible*.
IF d_1 is *Near* AND d_2 is *High* AND d_{16} is *Far* THEN *front* collision is *Possible*.
IF d_1 is *Near* AND d_2 is *Far* AND d_{16} is *Far* THEN *front* collision is *Possible*.

The mathematical meaning of the kth single antecedent rule (type R1) is given as a fuzzy relation $R^{(k)}$ on $D \times C$, which in the membership functions domain is

$$\mu_{R*K)}(d_i, c_j) = \min\left[\mu_{LD^{(K)}}(d_i), \mu_{LC^{(k)}}(c_j)\right] \tag{7}$$

The membership function of the obstacle's position and distance from the robot is computed by the max-min composition between the sensor readings, which represent the distance from the obstacles and the fuzzy relation described by (7). The second layer fuzzy controller takes as input the variables: *collision*, *angle_error*, and *speed* and generates the control *change of speed* and *steering*. The potential field generates a heading directive based upon the sum of forces approach described in the previous section. The heading relative to the current heading of the robot is given as the *angle_error* θ. This angle represents the angle in which the vehicle should point in order to follow the desired path to the goal point. Since the desired heading given by the potential field already considers some reactive navigation, the second-layer fuzzy inference should implement the steering necessary to reduce the *angle_error* to zero, only changing the desired heading if collision in the vicinity of the desired path is possible.

Reaction-directed behaviors are controlled through analysis of the collision of the current situation. For example, IF *collision* is *front* possible AND *speed* is *normal* THEN *change of speed* is *decelerate slow*. A similar rule, which examines the *angle_error* as well as all values of *collision*, generates the control *steering*. The rules of the second fuzzy module can be described compactly as follows:

IF c_j is $LC^{(k)}$ AND θ is $L\Theta^{(k)}$ AND v is $LV^{(k)}$ THEN s is $LS^{(k)}$ AND dv is $LDV^{(k)}$

where

- k is the rule number,
- c_j is collision of type j,
- θ is the steering angle error,
- v is the speed of the robot,
- s is the steering angle correction,
- dv is the change of speed, and
- $LC^{(k)}$, $L\Theta^{(k)}$, and $LV^{(k)}$, $LS^{(k)}$, $LDV^{(k)}$ are the linguistic values of c_j, θ, v, s and dv.

In all the rules, AND is the min operator. The steering angle error θ is computed by continuously comparing the desired steering angle, θ_{desired}, i. e., the angle that guides the robot to the target point given, with the actual steering angle, θ_{actual}. The steering error is important when the robot detects no collision on the target path. In case of possible collision, the information about the error θ becomes of less importance and the rules that contain such information are firing with smaller strength than the rules which perform collision avoidance. The generic mathematical expression of the kth navigation rule is

$$\mu_{R^{(k)}}(c_j, \theta, v, s, \mathrm{d}v) = \min\left[\mu_{LC^{(k)}}(c_j), \mu_{L\Theta^{(k)}}(\theta), \mu_{LV^{(k)}}(v), \mu_{LS^{(k)}}(s), \mu_{LdV^{(k)}}(\mathrm{d}v)\right] \tag{8}$$

The overall navigation output is given by he max-min composition and in particular is

$$\mu_N^*(s, \mathrm{d}v) = \max \min_{c_j,\theta,v}\left[\mu_{AND}^*(c_j, \theta, v), \mu_R(c_j, \theta, v, s, \mathrm{d}v)\right], \tag{9}$$

where

$$\mu_R(c_j, \theta, v, s, \mathrm{d}v) = \bigcup_{k=1}^{K} \mu_{R^{(k)}}(c_j, \theta, v, s, \mathrm{d}v)$$

and $\mu_{AND}^*(c_j, \theta, v)$ is the minimum of the fuzzified sonar readings. The navigation action dictates change in robot speed and/or steering correction as it comes out from a defuzzification formula, which calculates the center of the area covered by the membership function computed from (9). Part of the rule base is shown in Table 1.

Neuro-fuzzy Control of Autonomous Robotics, Table 1
Rule base for the fuzzy inference engine implemented in [30]

Input variables						Output variables	
Collision				Speed	Steering angle error	Change of speed	Steering angle correction
Front	Left	Back	Right				
N-P	N-P	N-P	N-P	Low	RB	NC	TLF
N-P	N-P	N-P	N-P	Normal	R	NC	TL
N-P	N-P	N-P	N-P	Normal	Zero	ACC-F	NC
N-P	High	N-P	N-P	High	LS	DEC-SL	TRS
N-P	N-P	N-P	High	Low	LB	NC	TRS
High	High	N-P	N-P	Low	Zero	NC	TRF
High	N-P	High	High	Normal	RS	DEC-F	TLF
N-P	High	High	N-P	Normal	R	DEC-SL	NC
N-P	High	N-P	High	High	RB	DEC-SL	NC
High	High	High	N-P	High	Zero	DEC-F	TRF
High	N-P	High	N-P	Low	LS	NC	TR

Neuro–Fuzzy Assisted EKF for SLAM Problems

The most successful robot navigation algorithms have been derived from a probabilistic perspective [29], which takes into account vehicle motion, terrain uncertainty and sensor noise. The interest in the estimation of an autonomous robot's location and that of its surroundings, known as Simultaneous Location and Map Building (SLAM), is evident. The goal of an autonomous vehicle performing SLAM is to build a map consisting of environment features (landmarks) incrementally, by using the uncertain information extracted from its sensors, while simultaneously using that map to localize itself with respect to the reference coordinate frame. Extended Kalman Filter (EKF) [17] has been a popular choice for SLAM [8]. This approach estimates and stores the robot position and orientation and the feature positions within the map of the environment as a complete state-vector. The uncertainties in these estimates are stored as error covariance matrices. It has been shown previously that the performance of an EKF process depends largely on the accuracy of the knowledge of the process noise covariance matrix ($\mathbf{Q}$) and measurement noise covariance matrix ($\mathbf{R}$). An incorrect a priori knowledge of $\mathbf{Q}$ and $\mathbf{R}$ can lead to performance degradation or even divergence [11,20].

In real-world applications, the $\mathbf{Q}$ and $\mathbf{R}$ matrices may not be accurately known. In [5] Chatterjee and Matsuno suggest the use of fuzzy logic techniques for the adaptation of the statistical assumption of the EKF caused by unknown or possible changed sensor noise characteristics. Only the adaptive estimation of $\mathbf{R}$ is considered, which is done at each iteration. We will present the main aspects of their implementation as an example on how fuzzy logic can improve some important methods used for robot navigation, even if their intrinsic nature is completely different.

The algorithm assumes that the features do not move and are modeled as static 2D points, and that Gaussian noise is affecting the system states and the measurements (represented by the range r and the bearing θ).

The EKF is used when the process state transition is modeled by a nonlinear function or when there is a nonlinear relation between the measurements and the states of the system:

$$\boldsymbol{x}_{k+1} = \mathbf{f}(\boldsymbol{x}_k, \mathbf{u}_k) + \mathbf{q}_k \tag{10}$$

$$\mathbf{z}_{k+1} = \mathbf{h}(\boldsymbol{x}_{k+1}) + \mathbf{r}_{k+1} \tag{11}$$

where $\boldsymbol{x}_k \in \mathbf{R}^n$ is the state vector of the process with the corresponding covariance matrix $\mathbf{P}_k$, $\mathbf{z}_k \in \mathbf{R}^m$ is the measurement vector and $\mathbf{u}_k$ is the control input. The random variables $\mathbf{q}_k$ and $\mathbf{r}_k$ represent the process and measurement noise (respectively). They are assumed to be independent (of each other), white, and with normal probability distributions

$$p(\mathbf{q}) \sim N(0, \mathbf{Q})$$
$$p(\mathbf{r}) \sim N(0, \mathbf{R})$$

For the SLAM algorithm, the coordinates of the features are considered states of an augmented system. The total state vector $\boldsymbol{x}$ will contain the states of the vehicle $\boldsymbol{x}_v$ (position and orientation in the reference frame) and the states of the already observed features $\boldsymbol{x}_m$ (a vector of variable length). The mean estimate of the total state vector and the corresponding total error covariance matrix $\mathbf{P}$ are the

following:

$$\hat{\mathbf{x}} = \left[\hat{\mathbf{x}}_v^{\mathrm{T}} \quad \hat{\mathbf{x}}_m^{\mathrm{T}}\right]^{\mathrm{T}} \tag{12}$$

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_v & \mathbf{P}_{vm} \\ \mathbf{P}_{vm}^{\mathrm{T}} & \mathbf{P}_m \end{bmatrix} \tag{13}$$

$$\hat{\mathbf{x}}_v = \left[\hat{x}_v \quad \hat{y}_v \quad \hat{\varphi}_v\right]^{\mathrm{T}} \tag{14}$$

$$\hat{\mathbf{x}}_m = \left[\hat{x}_1 \quad \hat{y}_1 \quad \ldots \quad \hat{x}_n \quad \hat{y}_n\right]^{\mathrm{T}} \tag{15}$$

The state transition in (10) is modeled by the following nonlinear function that keeps the feature states constant:

$$\mathbf{f}(\boldsymbol{x}_k, \mathbf{u}_k) = \begin{bmatrix} \mathbf{f}_v(\boldsymbol{x}_{v_k}, \mathbf{u}_k) \\ \boldsymbol{x}_{m_k} \end{bmatrix} \tag{16}$$

The map is defined from the position estimates of the static features as they appear in (15). The number of features n is variable. Initially the algorithm is started with $n = 0$. They are added in subsequent iterations, when the robot starts moving and observations are carried out. The control inputs to the system are also considered to be random variables, because they are measurements (e. g. from odometers) affected by errors. They are modeled as Gaussian variations from their nominal values with normal probability distribution

$$p(\mathbf{u}) \sim N(0, \mathbf{U})$$

The following equations implement the discrete EKF. A complete description can be found in [31].

EKF Time Update ("Predict") equations:

$$\hat{\mathbf{x}}_{k+1}^{-} = \mathbf{f}(\boldsymbol{x}_k, \mathbf{u}_k) \tag{17}$$

$$\mathbf{P}_{k+1}^{-} = \nabla\mathbf{f}_{\boldsymbol{x}_k}\mathbf{P}_k\nabla\mathbf{f}_{\boldsymbol{x}_k}^{\mathrm{T}} + \nabla\mathbf{f}_{\mathbf{u}_k}\mathbf{U}_k\nabla\mathbf{f}_{\mathbf{u}_k}^{\mathrm{T}} \tag{18}$$

$$\nabla\mathbf{f}_{\boldsymbol{x}_k} = \left.\frac{\partial\mathbf{f}}{\partial\boldsymbol{x}_k}\right|_{(\hat{\mathbf{x}}_k, \hat{\mathbf{u}}_k)}$$

$$\nabla\mathbf{f}_{\mathbf{u}_k} = \left.\frac{\partial\mathbf{f}}{\partial\mathbf{u}_k}\right|_{(\hat{\mathbf{x}}_k, \hat{\mathbf{u}}_k)}$$

EKF Kalman Gain equations:

$$\nu_{i_{k+1}} = \mathbf{z}_{k+1} - \mathbf{h}_i(\hat{\mathbf{x}}_{k+1}^{-}) \tag{19}$$

$$\mathbf{S}_{i_{k+1}} = \nabla\mathbf{h}_{\boldsymbol{x}_{k+1}}\mathbf{P}_{k+1}^{-}\nabla\mathbf{h}_{\boldsymbol{x}_{k+1}}^{\mathrm{T}} + \mathbf{R}_k \tag{20}$$

$$\mathbf{K}_{i_{k+1}} = \mathbf{P}_{k+1}^{-}\nabla\mathbf{h}_{\boldsymbol{x}_{k+1}}^{\mathrm{T}}\mathbf{S}_{i_{k+1}}^{-1} \tag{21}$$

$$\nabla\mathbf{h}_{\boldsymbol{x}_{k+1}} = \left.\frac{\partial\mathbf{h}_i}{\partial\boldsymbol{x}_k}\right|_{\hat{\mathbf{x}}_{k+1}^{-}}$$

EKF Measurement Update ("Correct") equations:

$$\hat{\mathbf{x}}_{k+1}^{+} = \hat{\mathbf{x}}_{k+1}^{-} + \mathbf{K}_{i_{k+1}}\nu_{i_{k+1}} \tag{22}$$

$$\mathbf{P}_{k+1}^{+} = \mathbf{P}_{k+1}^{-} - \mathbf{K}_{i_{k+1}}\mathbf{S}_{i_{k+1}}\mathbf{K}_{i_{k+1}}^{\mathrm{T}} \tag{23}$$

The features are measured relative to the observer in terms of their range r and bearing θ. The measurement vector is given as

$$\mathbf{z} = [r \quad \theta]^{\mathrm{T}} \tag{24}$$

and the measurement noise is described by the covariance matrix

$$\mathbf{R} = \begin{bmatrix} \sigma_r^2 & 0 \\ 0 & \sigma_\theta^2 \end{bmatrix} \tag{25}$$

where it was assumed that there is no cross-correlation between the range and the bearing measurements.

In the previous equations, ν_i is the innovation of the observation corresponding to the ith landmark, $\mathbf{S}_i$ is the covariance matrix of the innovation and $\mathbf{h}_i$ is a function that maps the $(\hat{x}_i, \hat{y}_i)$ estimated coordinates of the ith landmark to range and bearing coordinates relative to the current robot position and orientation.

The approach for adapting the EKF used in [5] by Chatterjee and Matsuno is based on the innovation adaptive estimation. The method was originally proposed in [20] and combined with fuzzy logic in [19]. The basic idea behind the method is to determine innovation of the observation at every kth instant (the discrepancy between a new measurement $\mathbf{z}_k$ and its corresponding estimation $\hat{\mathbf{z}}_k$) and to use the new information to correct the estimate already made. The adaptation strategy has the objective of reducing the mismatch between the theoretical covariance of the innovation sequences ($\mathbf{S}_k$) and the corresponding actual covariance of the innovation sequences ($\hat{\mathbf{C}}_{Innk}$). $\mathbf{S}_k$ is calculated using (20) and $\hat{\mathbf{C}}_{Innk}$ is calculated as

$$\hat{\mathbf{C}}_{Innk} = \nu_k\nu_k^{\mathrm{T}} \tag{26}$$

When an observation step is carried out, there will be multiple landmarks visible simultaneously. The algorithm considers batch measurements of the form

$$\mathbf{z} = [r_1, \quad \theta_1, \quad \ldots \quad r_n, \quad \theta_n]^{\mathrm{T}} \tag{27}$$

and as a result the innovation vector is

$$\nu = \left[\nu_1^{\mathrm{T}} \quad \nu_2^{\mathrm{T}} \quad \ldots \quad \nu_n^{\mathrm{T}}\right]^{\mathrm{T}} \tag{28}$$

The matrices involved in the algorithm are also constructed for batch mode measurements and have corresponding dimensions. The EKF algorithm performs better

update steps for SLAM if the innovation vector v contains more observations simultaneously. The mismatch of the innovation sequence covariance at the kth instant is

$$\Delta\hat{\mathbf{C}}_{Innk} = \hat{\mathbf{C}}_{Innk} - \mathbf{S}_k$$

The objective is to minimize this mismatch by employing a one input – one output neuro fuzzy system for each diagonal element of the $\Delta\hat{\mathbf{C}}_{Innk}$ matrix. The fuzzy rules will adapt the sensor statistics defined by the $\mathbf{R}$ matrix so that there will be a reduction in the $\Delta\hat{\mathbf{C}}_{Innk}$ mismatch. The adaptation algorithm is run during the EKF Measurement Update step. It does the following actions:

- Determine the set of visible landmarks from the current robot position and do the range-bearing measurements for each of them to obtain $\mathbf{z}$ as in (27).
- Predict the range-bearing observations on the basis of the augmented total step vector and compute the augmented innovation sequence using (19) and (28) for batch processing.
- Compute the augmented noise covariance matrix for the measurements using the original $(2x2)$ $\mathbf{R}$ matrix, the augmented observation model $\mathbf{h}$ and the augmented $\mathbf{S}$ for batch-mode situations.
- Update the a-posteriori state estimate vector and the error covariance matrix using (22) and (23).
- Compute $\hat{\mathbf{C}}_{Innk}$, $\Delta\hat{\mathbf{C}}_{Innk}$ and normalize the diagonal entries of $\Delta\hat{\mathbf{C}}_{Innk}$ corresponding to the range mismatch or to the bearing mismatch with respect to the highest value of the respective measurement on the diagonal.
- Apply each $\Delta\hat{\mathbf{C}}_{Innk}(j,j)$ element from the diagonal as an input to the *neuro-fuzzy system* and obtain the corresponding $\Delta\mathbf{R}(j,j)$ output.
- Determine $\Delta\sigma_r^2$ and $\Delta\sigma_\theta^2$ as means of those $\Delta\mathbf{R}(j,j)$ entries which correspond to range measurements and bearing measurements, respectively.
- Adapt the original $(2x2)$ $\mathbf{R}$ matrix in the following manner:

$$\mathbf{R}_k = \mathbf{R}_{k-1} + \begin{bmatrix} \Delta\sigma_r^2 & 0 \\ 0 & \Delta\sigma_\theta^2 \end{bmatrix}. \tag{29}$$

From the algorithm, it can be seen that the neuro-fuzzy system (NFS) employs a nonlinear mapping between the input $\Delta\hat{\mathbf{C}}_{Innk}(j,j)$ and the output $\Delta\mathbf{R}(j,j)$. For each diagonal element of innovation covariance mismatch there will be a corresponding adaptation value recommended for the diagonal element of the augmented noise covariance matrix $\mathbf{R}$. Because the same NFS is applied for every diagonal element, the input to the NFS is normalized. The NFS is implemented by three IF-THEN rules:

IF $\Delta\hat{\mathbf{C}}_{Innk}(j,j)$ is N THEN $\Delta\mathbf{R}(j,j) = w_1$
IF $\Delta\hat{\mathbf{C}}_{Innk}(j,j)$ is Z THEN $\Delta\mathbf{R}(j,j) = w_2$
IF $\Delta\hat{\mathbf{C}}_{Innk}(j,j)$ is P THEN $\Delta\mathbf{R}(j,j) = w_3$

where w_1, w_2 and w_3 indicate the amount of fuzzy adaptation recommended in the form of a diagonal element of the $\Delta\mathbf{R}$ matrix that depends on the nature of the mismatch in the corresponding diagonal element of the $\Delta\hat{\mathbf{C}}_{Innk}$ matrix. The neuro-fuzzy adaptation system is represented in Fig. 14 and the fuzzy membership functions in Fig. 15.

The $\Delta\hat{\mathbf{C}}_{Innk}(j,j)$ input variable is fuzzified using three Membership Functions (MFs) μ_i: negative (N), zero (Z) and positive (P). The output of the defuzzification layer is calculated as a weighted average of all its inputs. Finally, the output layer performs a suitably scaling for the defuzzified output. The complete input-output relationship is the following:

$$\Delta\mathbf{R}_k(j,j) = K \cdot \frac{\sum_{i=1}^{3} \mu_i\left(\Delta\mathbf{C}_{Innk}(j,j)\right) \cdot w_i}{\sum_{i=1}^{3} \mu_i\left(\Delta\mathbf{C}_{Innk}(j,j)\right)} \tag{30}$$

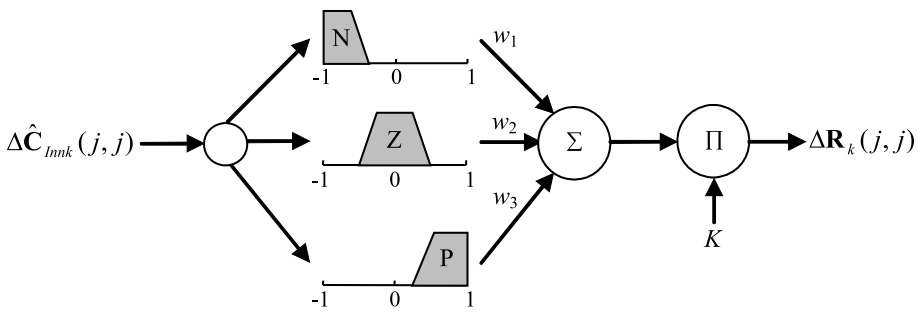

Neuro-fuzzy Control of Autonomous Robotics, Figure 14
Architecture of the neuro-fuzzy system

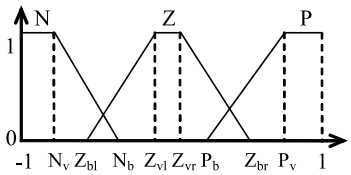

Neuro-fuzzy Control of Autonomous Robotics, Figure 15
Parametrization of the membership functions

The neuro-fuzzy model has to be trained in order to determine the suitable parameters of the membership functions, the output consequence singletons and the output gain. Unfortunately, the training can not be completed in the conventional supervised mode because the desired output for a given input is not known. A stochastic global optimization algorithm is needed for training in an unsupervised manner. The authors have chosen to use Particle Swarm Optimization (PSO), which is a relatively new algorithm based on the swarm behavior of birds and fishes. A detailed description of the algorithm can be found in [6]. The training is done as a high-dimension metaheuristic problem with the objective of minimizing a fitness function $f_{\text{fit}}(x_1, x_2, \ldots, x_n)$ on the basis of the variables $x_1, x_2, \ldots, x_n$.

The "particles" in a PSO problem are several candidate solutions of $x_1, x_2, \ldots, x_n$. At each iteration, the suitability of each solution is evaluated. For the training of the neuro-fuzzy system, the particles are formed as 12 component vectors

$$\begin{aligned} \boldsymbol{x} &= \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \ldots & \boldsymbol{x}_{12} \end{bmatrix} \\ &= \begin{bmatrix} N_v & N_b & Z_{bl} & Z_{vl} & Z_{vr} & Z_{br} \\ & & & P_b & P_v & w_1 & w_2 & w_3 & K \end{bmatrix} \end{aligned}$$

Each particle is characterized by vectors denoting its position $\boldsymbol{x}_i$ and its velocity $\mathbf{v}_i$, which are adjusted at each time step in order to minimize the fitness function $f_{\text{fit}}(x_1, x_2, \ldots, x_n)$. The traditional PSO algorithm has a quick convergence during the early phase of the training procedure, but has problems in finding the precise solution to the problem. To improve the behavior of the PSO algorithm, the variation for the speed is kept high in the initial passes and is gradually decreased as iterations progress.

The formulation of the fitness function has to be in concordance with the objective of the neuro-fuzzy assistance to the EKF. The discrepancy between the theoretical covariance and the actual covariance of the innovation sequence over the entire range of observation instants has to be minimized. For each iteration there are N_{obs} observations considered. The fitness function calculates the mean-square value of all the diagonal entries of $\Delta\hat{\mathbf{C}}_{Innk}$ for each observation instant and computes a mean of these values for the entire batch of observations:

$$\mathbf{f}_{\text{fit}} = \frac{1}{N_{\text{obs}}} \sum_{n_{\text{obs}}=1}^{N_{\text{obs}}} \left(\frac{1}{J_{Cnobs}} \sum_{j=1}^{J_{Cnobs}} \left[\Delta\mathbf{C}_{Innk}(j,j) \right]^2 \right) \quad (31)$$

where J_{Cnobs} is the number of diagonal elements of the $\Delta\hat{\mathbf{C}}_{Innk}$ matrix (2× the number of observed features) when the observation n_{obs} takes place.

The positions of each of the particles at the end of each iteration have to comply with several constrains due to the specific shapes that have to be obtained for the fuzzy membership functions (in this case trapezoidal) and also to ensure some overlapping between consecutive MFs. All the points from the universe of discourse have to be covered by at least one MF. The constraints are implemented using the following type of rules:

IF $(N_b < N_v)$ THEN $N_b = N_v$,
IF $(Z_{vl} < Z_{bl})$ THEN $Z_{vl} = Z_{bl}, \ldots$ etc.

The performance of the algorithm was compared with a conventional EKF SLAM where the noise covariance matrices $\mathbf{Q}$ and $\mathbf{R}$ are kept constant. Both algorithms start with the same values for the matrices, but the neuro-fuzzy algorithm adapts $\mathbf{R}$ continuously. The EKF-based SLAM algorithm performs very well when the sensor statistics are known. The performance degrades significantly and becomes highly unreliable when the statistics are wrongly considered, especially for a high number of landmarks. The neuro-fuzzy assistance improves the situation dramatically, helping the EKF to maintain a stable performance over several environment situations while having wrong knowledge of the sensor statistics.

The training of the neuro-fuzzy system was done in offline mode using the data gathered by the robot for a given configuration of the environment. Once the NFS parameters were determined, experiments were done for robot navigation through waypoints for several configurations of landmarks.

Behaviors

The operational characteristics of unmanned vehicles include the following:

- Perception: acquire knowledge about the environment using sensors and extracting meaningful information to be used in later tasks
- Intelligence: operate for a considerable amount of time without human intervention while accomplishing useful tasks
- Action: in general, move between certain points.

One of the most common applications of fuzzy logic for autonomous robotics is to implement individual behaviors. Considering the environment uncertainty that is difficult if not impossible to model, effective control algorithms for autonomous navigation should imitate the way humans are operating various vehicles. Fuzzy logic allows a suitable knowledge representation of inherently vague notions achieved through IF-THEN rules. These rules contain linguistic information that describes the problem in a simple and fast manner. In many applications of fuzzy logic, a mathematical model of the dynamics of the vehicle is not needed. The only requirement for the design of the inference engine is the heuristic control knowledge related to the specific problem, usually obtained from a human who knows how to handle the system. Tuning is done by trial and error methods. Because of their interpolative nature, fuzzy controllers generate smooth movement for the robots and provide robustness and a graceful degradation of performance when confronted with noise and other type of errors in the data obtained from the sensors.

Feedback Linearization Using Neural Networks

The lowest level of control on mobile robots is the execution level (Fig. 7). The higher levels generate a desired trajectory for the robot. The motor control unit has to do force control, position control or speed control. In some cases, a simple PID controller can be used with good results. But usually the exact model of the system is unknown and even worse, it is also nonlinear. A simple linear controller may not be robust or might not have a good performance. Neural networks can be used to build "model-free" controllers that offer a powerful and robust alternative to adaptive control. Mimicking the functions of human processes, these controllers learn about the systems they are controlling on-line, and thus automatically improve their performance.

Feedback linearization techniques offer a widely applicable set of design tools that are useful for broad classes of nonlinear systems. They function by basically converting the nonlinear problem into a related linear controls design problem. The performance is far exceeding that of classical linearization controllers based on Jacobian linearization techniques because no approximation is involved in feedback linearization design.

Lewis et al. [18] have done feedback linearization for a robotic manipulator using a neural network to estimate the nonlinearities of the system. The method can be applied to a large class of nonlinear systems.

The robot manipulator has dynamics

$$M(q)\ddot{q} + V_m(q,\dot{q})\dot{q} + F(\dot{q}) + G(q) + \tau_d = \tau \tag{32}$$

To make the robot manipulator follow a prescribed desired trajectory $q_d(t)$, define the tracking error $e(t)$ and filtered tracking error $r(t)$ by

$$e(t) = q_d - q \tag{33}$$

$$r = \dot{e} + \Lambda e \tag{34}$$

with $\Lambda > 0$ a positive definite design parameter matrix. The robot dynamics are expressed in terms of the filtered error as

$$M\dot{r} = -V_m r + f(x) + \tau_d - \tau \tag{35}$$

where the nonlinear robot function is defined as

$$\begin{aligned} f(x) = {} & M(q)(\ddot{q}_d + \Lambda\dot{e}) \\ & + V_m(q,\dot{q})(\dot{q}_d + \Lambda e) + F(\dot{q}) + G(q) \end{aligned} \tag{36}$$

One may define

$$x \equiv \begin{bmatrix} e^{\mathrm{T}} & \dot{e}^{\mathrm{T}} & q_d^{\mathrm{T}} & \dot{q}_d^{\mathrm{T}} & \ddot{q}_d^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} .$$

According to the universal approximation property of the neural network, there is a two-layer NN such that

$$f(x) = W^{\mathrm{T}}\sigma(V^{\mathrm{T}}x) + \varepsilon \tag{37}$$

with the approximation error ε bounded on a compact set by a known bound. W and V are unknown ideal target weights that give good approximation to $f(x)$. It is only necessary to know that they exist.

The control input is selected as

$$\tau = \hat{W}^{\mathrm{T}}\sigma(\hat{V}^{\mathrm{T}}x) + K_v r - v \tag{38}$$

with $v(t)$ a function that provides robustness:

$$v(t) = -K_z(\left\|\hat{Z}\right\|_F + Z_B)r\,, \tag{39}$$

where

$$Z \equiv \begin{bmatrix} W & 0 \\ 0 & V \end{bmatrix} \tag{40}$$

and Z_B is a known bound for the ideal NN weights such that $\|Z\|_F \leqslant Z_B$.

The NN weights and thresholds are tuned using the following algorithms:

$$\dot{\hat{W}} = F\hat{\sigma}r^{\mathrm{T}} - F\hat{\sigma}'\hat{V}^{\mathrm{T}}xr^{\mathrm{T}} - kF\,\|r\|\,\hat{W} \tag{41}$$

$$\dot{\hat{V}} = Gx(\hat{\sigma}'^{\mathrm{T}}\hat{W}r)^{\mathrm{T}} - kG\,\|r\|\,\hat{V}\,, \tag{42}$$

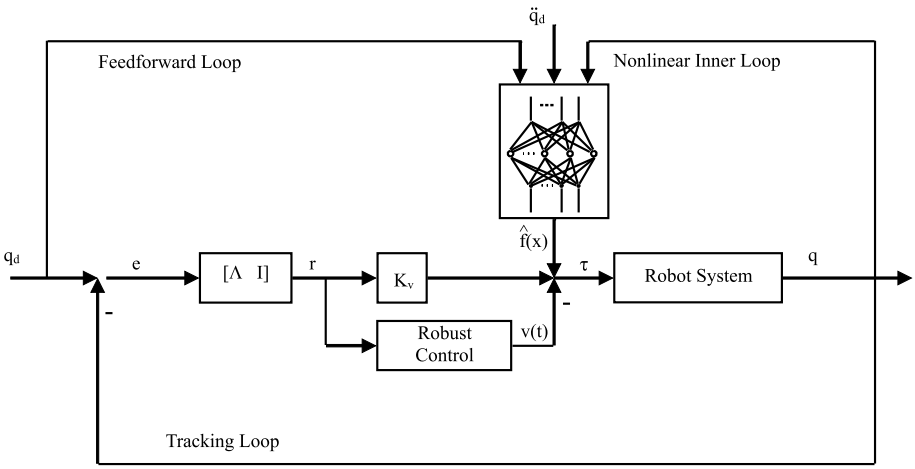

Neuro-fuzzy Control of Autonomous Robotics, Figure 16
Multilayer neural network controller structure

where the design parameters are the positive definite matrices F, G and a small scalar $k > 0$. Weight initialization is not critical and there is no preliminary off-line learning required. In fact, selecting the initial weights $\hat{W}(0)$, $\hat{V}(0)$ as zero takes the NN out of the circuit and leaves only the outer PD tracking loop. It is well known that the PD term $K_v r$ can then stabilize the plant on an interim basis until the NN begins to learn. The NN weights are tuned on-line in real time. As the NN learns, the tracking error decreases. The stability proof is given in [18]. The NN structure is shown in Fig. 16.

Combining and Coordinating Behaviors

Typically, fuzzy controllers consider a single goal. Isolated reactive behaviors are incapable of performing autonomous navigation in complex environments. However, more complex tasks can be accomplished through combination and cooperation of primitive behaviors. For the situations where there are two or more goals active simultaneously, there are two solutions:

- Build complex rules whose antecedents consider both goals in the same time.
- Write two sets of simple rules, each set specific to a single goal, and combine their output using some form of behavior arbitration or command fusion.

Behavior Arbitration

Arbitration is the process that leads to the activation of a specific behavior or that generates weights for multiple simultaneous behaviors. It decides which behaviors should influence the operation of the robot and how much.

The subsumption architecture implemented by Brooks in [3] is representative for fixed behavior arbitration. Because the robot was doing a single task, he used a fixed arbitration policy, built as a network of suppression and inhibition links. However, autonomous robots need to adapt to the environment and to perform multiple tasks. As a result, the arbitration strategies have to be dynamic, taking into consideration both the environment configuration and the current plan from the higher decisional levels. By using fuzzy logic to implement either fixed or dynamic arbitration, the transition between behaviors will be smooth and it will be possible to allow partial and concurrent activation of behaviors.

Command Fusion

The most simple and obvious method used to fuse the commands coming from multiple behaviors consists in a switching scheme (Fig. 17). The behaviors are activated only one at a time. The selected dominant behav-

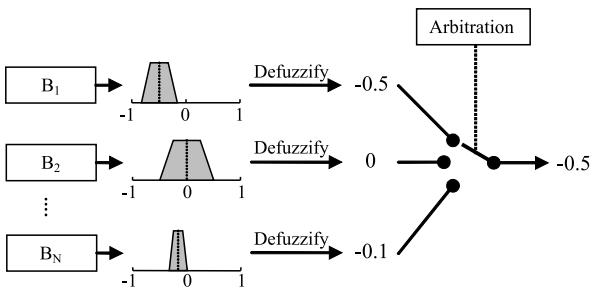

Neuro-fuzzy Control of Autonomous Robotics, Figure 17
Switching scheme for command fusion

ior solely controls the robot until the next decision cycle, whereas the motor commands of the suppressed behaviors are completely ignored. The arbitration strategy determines which behavior is activated. This switching approach achieves a poor performance in the presence of multiple goals. A good example is the potential field navigation using a reactive behavior for the avoidance of obstacles that are not represented on the map. There are two goals: the global behavior tries to reach the target, while the reactive behavior has to avoid the obstacles. When the robot approaches an obstacle, the control is taken away from the target-reaching behavior and given to the obstacle avoidance behavior. The latter can take the arbitrary decision of passing the obstacle through the left or through the right. Even if the goal of avoiding the obstacle is reached, the global solution to the navigation problem can suffer significantly if the wrong direction is chosen.

Parallel execution of multiple behaviors can overcome some of the limitations of the switching scheme. The final commands given to the robot can be calculated in two ways: by combining individual decisions (Fig. 18) or by combining individual preferences (Fig. 19).

The vector summation scheme is a good example of combining individual decisions. Considering the previous problem relating to the potential field navigation, we can identify two behaviors. Each one provides a vector characterizing the desired velocity and direction of movement. By summing the two (weighted) vectors, it is possible to obtain an intermediate speed and direction in order to reach both goals. This approach does not necessarily need fuzzy logic, although it can be easily implemented using weighted singletons as fuzzy outputs and center-of-gravity defuzzification.

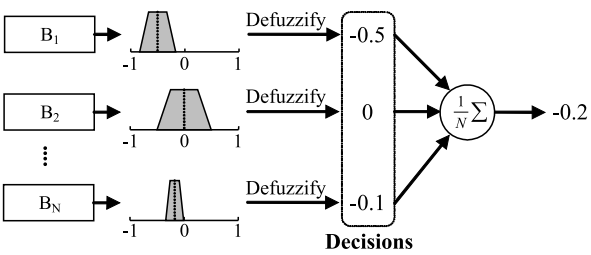

Neuro-fuzzy Control of Autonomous Robotics, Figure 18
Command fusion by combining individual decisions

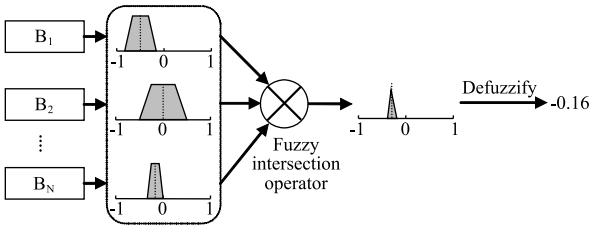

Neuro-fuzzy Control of Autonomous Robotics, Figure 19
Command fusion by combining individual preferences

The real advantage of using fuzzy logic for command fusion becomes evident when combining individual preferences. A preference can be represented as a probability density function or as a fuzzy set [23]. It provides more information than a decision, which is a single (crisp) value. It gives information about an entire range of possible values and about how desirable they are for accomplishing the required task. Fuzzy logic has many different operators to perform combination and many defuzzification functions to perform decision. If the behaviors are also implemented using fuzzy logic, it is easy to have them output a fuzzy set instead of a crisp value, by eliminating the defuzzification phase. Even behaviors implemented using a different mechanism, but that output a PDF, can be easily accommodated.

All the previous methods of command fusion suffer in some way or another when they have to handle competing behaviors that issue conflicting control commands. In this case, the resulting motor command from the compromise decision might be sub-optimal or even worse than any of the individual commands. An example could be the situation where the defuzzification results in the selection of a value that lies between two peaks of the combined fuzzy set. For a robot that has to avoid an obstacle by going to the left or to the right, this would make it go forward, straight into it. There is a need for extensions to basic fuzzy command fusion schemes capable of resolving conflicts among contradicting actions. A simple example is given in [35], where Yen proposes a different defuzzification approach, by replacing the center of gravity with centroid of largest defuzzification. This only considers the output fuzzy set with the largest area and completely ignores all the others, making the method similar to the majority voting schemes.

Context-dependent blending of behaviors is the most general type of behavior combination that can be realized using fuzzy logic. The method was suggested by Ruspini in [24] for the Flakey robot and later by Saffiotti in [26]. They reintroduce some form of behavior arbitration into fusion by having a set of higher-level supervisory fuzzy rules to activate and deactivate the individual fuzzy behaviors. As a consequence, the hierarchical behavior architecture is composed of two distinct layers. On the higher level, behavior coordination is achieved by means of supervisory fuzzy rules of the form:

IF *context* THEN *behavior*

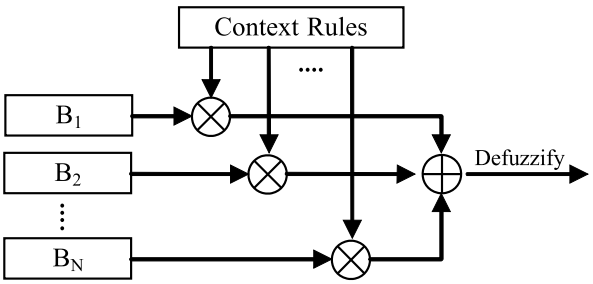

Neuro-fuzzy Control of Autonomous Robotics, Figure 20
Context-depending blending of behaviors [27]

As seen in Fig. 20, each behavior generates preferences in order to reach its goal. There is a context of activation for each behavior. It describes the applicability and desirability of that particular behavior and also reflects the needs of higher-level goals. The preferences of all behaviors, weighted by the truth value of their contexts, are combined to obtain the collective preference. The crisp value for the associated command is obtained after defuzzification.

The decisions can be event-driven or goal-driven. For example, in the potential field navigation problem with a reactive behavior for unknown obstacle avoidance, the following event-driven rules can apply:

IF *obstacle-close* THEN *avoid-obstacle*
IF *not(obstacle-close)* THEN *go-to-target*

Depending on the value of the *obstacle-close* fuzzy variable, the behaviors are activated with various strengths: for extreme values, either *avoid-obstacle* or *go-to-target* is fully enabled; for intermediate values, both behaviors are partially enabled. The goal-driven approach can be used, for example, to sequence behaviors. If the robot has to pass through a set of waypoints, these rules can be used:

IF W_1 *not reached* THEN *go-to-*W_1
IF W_1 *reached* AND W_2 *not reached* THEN *go-to-*W_2
...
IF W_{N-1} *reached* AND W_N *not reached* THEN *go-to-*W_N

The path followed by the robot will be smooth due to fuzzy interpolation of behaviors. The direction of movement and the speed will not change suddenly when a waypoint is bypassed. The event-driven and the goal-driven blending of behaviors can be combined into an arbitrarily complex set of rules in order to represent a full plan of action.

Learning of Behaviors

Until now we have discussed different applications of fuzzy logic for intelligent behaviors. Because fuzzy logic uses a rule base that works with linguistic variables, knowledge representation is very intuitive. A human operator can easily transpose his experience about controlling the robot into a set of rules. Unfortunately, the last step necessary for a functional fuzzy logic controller is not as easy. The relationship between the crisp values of the signals outside the controller and the fuzzy variables inside the controller is defined using membership functions that make possible the fuzzification and the defuzzification. A membership function must be defined for each value of a linguistic variable. A human expert can not define precisely the ideal shape for each of the membership functions. At best, he can only give fair approximations, based on his own way to associate numbers with words. Trial and error methods for tuning the membership functions and the rules could be avoided and an enormous amount of human efforts would be saved if the fuzzy system somehow possesses learning abilities.

A number of learning algorithms, such as the evolutionary algorithm [12], reinforcement learning [7,9], and supervised learning [14] have been proposed to construct the fuzzy system automatically. The evolutionary algorithm applied by itself results in a very long learning process. Although the reinforcing learning method seems quite promising because it doesn't require training data, it usually leads to a heavy learning phase as the gradient information is not provided explicitly and because the input space is too large for an efficient search. The supervised learning method has the advantages of fast convergence and is suitable for structure and parameter learning. However, it is difficult to obtain sufficient training data that doesn't contain conflicting input/output pairs. It is therefore impractical to learn behaviors using only one method for learning. Still, it is possible to use supervised learning to do a coarse tuning of the rule base and thus reduce the search domain for reinforcement learning, which will then fine-tune the rules. Ye, Yung and Wang apply this method in [34] for obstacle avoidance. The general aspects of their implementation will be shown here.

The neuro-fuzzy system in Fig. 21 receives information from the sensors as a n-dimensional input variable d'_i, $i = 1..n$ and generates a command y' to the robot. The premise membership functions are μ_{ik}. For each input variable, the associated membership functions are parametrized using a certain number of parameters x_{ri}. The fuzzy logic system has rules. The fired strength of the

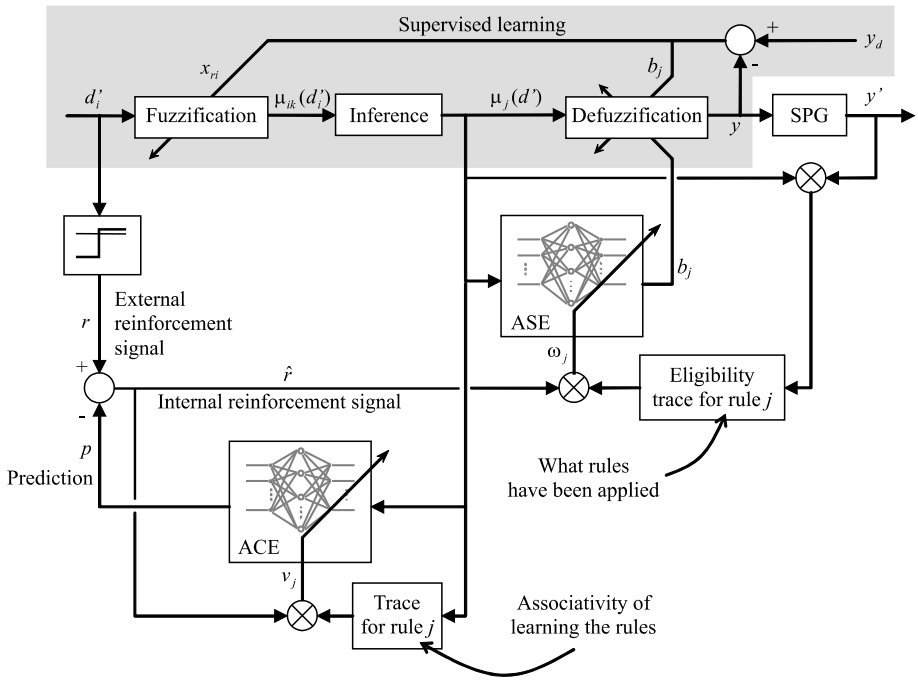

Neuro-fuzzy Control of Autonomous Robotics, Figure 21
Neuro-fuzzy system with mixed learning algorithm

jth rule is calculated as

$$\mu_j(d') = \prod_{i=1}^{n} \mu_{i,k_i}(d_i') \tag{43}$$

The output membership functions are parametrized by their centers b_j.

In the first learning phase, the supervised learning method is applied. It is shown in gray in Fig. 21. For each input state vector $\boldsymbol{d}$, the system infers an output y. The difference between the desired output y_d and the system output y is used to train the fuzzy system. A steepest descent learning algorithm is used to determine the parameters of both the input (x_{ri}) and the output (b_j) fuzzy sets. To avoid the problem of instability and in order to increase the learning speed, a simple fuzzy system (not shown on the figure) is used to adapt the learning rate and the momentum term of the algorithm.

Reinforcement learning is done using a modified model originally proposed by Sutton and Barto [2]. For learning obstacle avoidance, there is a conflict between the desire to use the rule base already learned and the desire to further explore the environment in order to improve the rule base. This phenomenon is called "the conflict between exploration and exploitation". To avoid using only exploitation and in order to maintain the efficiency of learning, a stochastic perturbation generator (SPG) was added. It generates an action $y'(t)$, which is a Gaussian random variable with mean $y(t)$ and a standard deviation that is high in case of a previous bad decision of the fuzzy system and small in case of a good decision. This way, the action of the SPG remains consistent with the fuzzy system when the previous action was a good one. During reinforcement learning, the parameters of the input fuzzy system are frozen and the method is applied to further tune the parameters of the output fuzzy sets.

For each input state d', the system infers an output y that is further adjusted by adding a stochastic perturbation and finally applied to the robot. By evaluating the new state generated by the robot movement, an internal reinforcement signal is generated and used to fine-tune the fuzzy system. The process repeats at each time step until a collision occurs. At this moment a reinforcement signal r, representing failure, is fed back to the learning network and the rules used at the previous time steps are changed in order to improve the performance of the robot. This task is accomplished by an adaptive neuron-like element, which consists of an associative search element (ASE) and an associative critic element (ACE). The ACE receives an external reinforcement signal $r(t)$ as a performance feedback and generates an internal reinforcement signal $\hat{r}(t)$ which is fed into the ASE to update its weights. The role of the

ACE is to make predictions. In order to do so, the weights v_j of the ACE are learned through the trace of the fired rules and its own output. The weights ω_j of the ASE are learned using the eligibility trace of each of the rules. These traces act like a filter and describe what rules have been used recently and what control actions have been applied.

Because the learning model adds perturbation to the control input of the robot, the resulting obstacle avoidance behavior is very robust to noisy sensor input. Moreover, from the experiments done by the authors in [34], the robot is able to perform collision-free navigation, it can achieve a path reasonably close to the shortest path and has a smooth motion.

Future Directions

Some standard algorithms used for mobile robot navigation can be improved or replaced by using neuro-fuzzy methods. While many implementations require the use of very precise sensors, full models of the systems and complicated mathematics, it can be observed in nature that the same actions can be realized by animals or humans using less precise sensorial information and without having access to or needing a mathematical model of the system or the environment. Fuzzy logic and neural networks can be used to implement a similar behavior for machines. Their full potential in this area has not been reached yet.

An example of how fuzzy logic can improve a standard method heavily used in robot navigation is the implementation of a 3D dead-reckoning system with fuzzy logic instead of Kalman filtering [21]. It uses physical reasoning about sensors' strengths and weaknesses to formulate expert system rules that fuse sensor data. Many error mechanisms can be defined more accurately by expert reasoning than by the statistics-based Kalman filter methods.

Bibliography

Primary Literature

1. Andrews JR, Hogan N (1983) Impedance Control as a Framework for Implementing Obstacle Avoidance in a Manipulator. In: Hardt DE, Book W (eds) Control of Manufacturing Processes and Robotic Systems. ASME, Boston, pp 243–251
2. Beom HR, Cho HS (1995) A sensor-based navigation for a mobile robot using fuzzy logic and reinforcement learning. IEEE Trans Syst Man Cybern 25:464–477
3. Brooks RA (1986) A Robust Layered Control System For A Mobile Robot. IEEE J Robot Autom RA-2:14–23
4. Burke RE (2001) Spinal Reflexes. In: Levitan I (ed) Encyclopedia of Life Sciences. Wiley, Chichester
5. Chatterjee A, Matsuno F (2006) Improving EKF-based solutions for SLAM problems in Mobile Robots employing Neuro–Fuzzy Supervision. 3rd International IEEE Conference Intelligent Systems, London, pp 683–689
6. Clerc M, Kennedy J (2002) The particle swarm-explosion, stability and convergence in a multidimensional complex space. IEEE Trans Evol Comput 6(1):58–73
7. Colombetti M, Dorigo M (1996) Behavior analysis and training-A methodology for behavior engineering. IEEE Trans Syst Man Cybern B 26:365–380
8. Dissanayake MWMG, Newman P, Clark S, Durrant–Whyte HF (2001) A solution to the simultaneous localization and map building (SLAM) problem. IEEE Trans Robot Autom 17(3):229–241
9. Donnart J, Meyer J (1996) Learning reactive and planning rules in a motivationally autonomous animation. IEEE Trans Syst Man Cybern B 26:381–395
10. Donoghue JP, Sanes JN (2001) Motor System Organization. In: Levitan I (ed) Encyclopedia of Life Sciences. Wiley, Chichester
11. Fitzgerald RJ (1971) Divergence of the Kalman filter. IEEE Trans Autom Control AC-16(6) 736–747
12. Homaifar A, McCormick E (1995) Simultaneous design of membership functions and rule sets for fuzzy controllers using genetic algorithms. IEEE Trans Fuzzy Syst 3:129–139
13. Hebert T (1998) Navigation of an autonomous vehicle using a combined electrostatic potential field/fuzzy inference approach. Ph.D. dissertation, University of Southwestern Louisiana, Lafayette
14. Juang CF, Lin CT (1998) An on-line self-constructing neural fuzzy inference network and its applications. IEEE Trans Fuzzy Syst 6:12–32
15. Khatib O (1985) Real-time obstacle avoidance for manipulators and mobile robots. IEEE Int Conf Robotics and Automation, St. Louis, MO, pp 500–505
16. Lefebvre DR, Saridis GN (1991) A Computer Architecture for Intelligent Machines. Intelligent Robotic Systems for Space Exploration. Rensselaer Polytechnic Institute Troy, New York, pp 31–43
17. Lewis FL (1986) Optimal Estimation: With an Introduction to Stochastic Control Theory. Wiley–Interscience, New York
18. Lewis FL, Jagannathan S, Yesildirek A (1998) Neural Network Control of Robot Manipulators and Nonlinear Systems. Taylor & Francis, Inc., Bristol
19. Loebis D, Sutton R, Chudley J, Naeem W (2004) Adaptive tuning of a Kalman filter via fuzzy logic for an intelligent AUV navigation system. Control Eng Pract 12:1531–1539
20. Mehra RK (1970) On the identification of variances and adaptive Kalman filtering. IEEE Trans Autom Control AC-15(2):175–184
21. Ojeda L, Borenstein J (2002) FLEXnav: Fuzzy Logic Expert Rule-based Position Estimation for Mobile Robots on Rugged Terrain. Proceedings of the 2002 IEEE International Conference on Robotics and Automation, Washington DC, pp 317–322
22. Park J, Sandberg IW (1991) Universal approximation using radial-basis-function networks. Neural Comput 3:246–257
23. Ruspini EH (1991) Truth as utility: a conceptual synthesis. In: Proc 7th Conf Uncertainty in Artificial Intelligence, Los Angeles, CA
24. Ruspini EH (1990) Fuzzy logic in the Flakey robot. Proc Int Conf Fuzzy Logic and Neural Networks (IIZUKA), Iizuka, Japan, pp 767–770
25. Saffiotti A (1997) The Uses of Fuzzy Logic in Autonomous Robotics: a catalogue raisonne. In: Degli Antoni G (ed) Soft Computing I(4). Springer, Berlin, pp 180–197

26. Saffiotti A, Konolige K, Ruspini EH (1995) A multivalued-logic approach to integrating planning and control. Artif Intell 76:(1–2):481–526
27. Saffiotti A, Ruspini EH, Konolige K (1993) Blending reactivity and goal-directedness in a fuzzy controller. In: Proceedings of the IEEE Int Conf Fuzzy Systems, San Francisco, California, pp 134–139
28. Saridis GN, Stephanou HE (1977) A hierarchical approach to the control of a prosthetic arm. IEEE Trans Syst Man Cybern SMC-7(6):407–420
29. Thrun S, Burgard W, Fox D (2005) Probabilistic Robotics. MIT Press, Cambridge
30. Valavanis KP, Hebert T, Kolluru R, Tsourveloudis NC (2000) Mobile robot navigation in 2-D dynamic environments using electrostatic potential fields. IEEE Trans Syst Man Cybern A 30:187–197
31. Welch G, Bishop G (2004) An introduction to the Kalman filter. Technical Report TR 95-041, University of North Carolina, Department of Computer Science, Chapel Hill
32. Wang LX (1992) Fuzzy systems are universal approximators. In: Proceedings of the 1st IEEE Conference on Fuzzy Systems, San Diego, pp 1163–1170
33. Werbos PJ (1998) Backpropagation: basics and new developments. The handbook of brain theory and neural networks. MIT Press, Cambridge, pp 134–139
34. Ye C, Yung NHC, Wang D (2003) A fuzzy controller with supervised learning assisted reinforcement learning algorithm for obstacle avoidance. IEEE Trans Syst Man Cybern B 33(1):17–27
35. Yen J, Pfluger N (1995) A Fuzzy Logic Based Extension to Payton and Rosenblatt's Command Fusion Method for Mobile Robot Navigation. IEEE Trans Syst Man Cybern 25(6):971–978
36. Zadeh LA (1973) Outline of a new approach to the analysis of complex systems and decision processes. IEEE Trans Syst Man Cybern SMC-3:28–44

Books and Reviews

Ge SS, Lewis FL (2006) Autonomous Mobile Robots: Sensing, Control, Decision Making and Applications. Autom Control Eng Ser 22:229–265

Tzafestas SG (1999) Advances in Intelligent Autonomous Systems. Microprocessor–Based and Intelligent Systems Engineering Series, vol 18. Springer, New York

Bekey GA (2005) Autonomous Robots: From Biological Inspiration to Implementation and Control. Intelligent Robotics and Autonomous Agents Series, MIT Press, Cambridge

Neuro-fuzzy Systems

Leszek Rutkowski[1], Krzysztof Cpałka[1,2], Robert Nowicki[1], Agata Pokropińska[3], Rafał Scherer[1]

[1] Departament of Computer Engineering, Częstochowa University of Technology, Częstochowa, Poland

[2] Department of Artificial Intelligence, Academy of Humanities and Economics, Lodz, Poland

[3] Institute of Mathematics and Computer Science, Jan Dlugosz University, Częstochowa, Poland

Article Outline

Glossary
Definition of the Subject
Introduction
Fuzzy Reasoning
Description of Fuzzy Inference Systems
Logical-Type Neuro-fuzzy Systems
Mamdani-Type Neuro-fuzzy Systems
Simplified Neuro-fuzzy Systems
Takagi–Sugeno Neuro-fuzzy Systems
Neuro-fuzzy Systems with Weights
Neuro-fuzzy Systems for Pattern Classification
Learning
Criteria Isolines Method
Future Directions
Acknowledgments
Bibliography

Glossary

Neuro-fuzzy systems Fusion of fuzzy logic and neural networks with the ability to automated adaptation to training data and knowledge interpretability.

Fuzzy reasoning Reasoning on the basis of fuzzy premises and fuzzy rules inferring fuzzy conclusions.

Definition of the Subject

Neuro-fuzzy systems (NFS) are part of soft computing concept. They are a synergistic fusion of fuzzy logic and neural networks with the ability to automate adaptation to training data and knowledge interpretability. Thus neuro-fuzzy systems incorporate two goals of machine learning which are difficult to meet at the same time: transparency of the rule base and high accuracy obtained thanks to learning from data. There are various types of NFS which differ in the form of fuzzy rules and inference methods. They have a very broad range of applications in life, science, economics, manufacturing, medicine etc. NFS can replace neural networks in all applications as most of them are universal approximators. Therefore they can perform well in tasks of classification, prediction, approximation and control, similarly to neural networks. Neural networks work as black boxes (usually it is not possible to tell where the knowledge is stored) and can not use prior knowledge. NFS can use almost the same learning methods and achieve the same accuracy as neural networks yet the knowledge in the form of fuzzy rules is easily interpretable for humans. On the other hand, fuzzy systems (not neuro-fuzzy) require manual tuning of their parameters and fuzzy reasoning in such systems is more trou-

blesome and computationally demanding than simplified one, used in NFS.

Introduction

Over the last decade fuzzy sets and fuzzy logic introduced in 1965 by Lotfi Zadeh [62] have been used in a wide range of problem domains, including process control, image processing, pattern recognition and classification, management, economics and decision making. Specific applications include washing-machine automation, camcorder focusing, TV color tuning, automobile transmissions and subway operations [17]. We have also been witnessing a rapid development in the area of neural networks (see e. g. [53,54,63]). Both fuzzy systems and neural networks, along with probabilistic methods [1,11,42], evolutionary algorithms [13,36], rough sets [43,44] and uncertain variables [3,4,5], constitute a consortium of soft computing techniques [1,23,26]. These techniques are often used in combination. For example, fuzzy inference systems are frequently converted into connectionist structures called neuro-fuzzy systems which exhibit advantages of neural networks and fuzzy systems. In literature various neuro-fuzzy systems have been developed (see e. g. [6,7,8,9,10,14,15,16,19,20,21,23,28,29,30,31,32,33,34,37,38,41,45,46,47,49,50,58]). They combine the natural language description of fuzzy systems and the learning properties of neural networks. Some of them are known in literature under short names such as ANFIS [18], ANNBFIS [9], DENFIS [24], FALCON [31], GARIC [2], NEFCLASS [40], NEFPROX [39,40], SANFIS [57] and others. The most popular designs of neuro-fuzzy structures fall into one of the following categories, depending on the connective between the antecedent and the consequent in fuzzy rules:

(i) Takagi–Sugeno method – consequents are functions of inputs,
(ii) Mamdani-type reasoning method – consequents and antecedents are related by the min operator or generally by a t-norm,
(iii) Logical-type reasoning method – consequents and antecedents are related by fuzzy implications, e. g. binary, Łukasiewicz, Zadeh and others.

It should be noted that most applications are dominated by the Mamdani-type fuzzy reasoning. However, it was emphasized by Yager [60,61] that "no formal reason exists for the preponderant use of the Mamdani method in fuzzy logic control as opposed to the logical method other than inertia". In this work we will outline all three types of systems. Moreover we will compare several variants of these systems in terms of efficiency for an exemplary problem.

Fuzzy Reasoning

In this section we present the idea of fuzzy reasoning with various fuzzy implications which will be useful for construction of neuro-fuzzy systems developed in the next sections. The basic rule of inference in classical logic is modus ponens. The compositional rule of inference describes a composition of a fuzzy set and a fuzzy relation. Fuzzy rule

$$\text{IF} \quad x \text{ is } A \quad \text{THEN} \quad y \text{ is } B \tag{1}$$

is represented by a fuzzy relation R. Having given input linguistic value A', we can infer an output fuzzy set B' by the composition of the fuzzy set A' and the relation R. The generalized modus ponens is the extension of the conventional modus ponens tautology, to allow partial fulfillment of the premises:

Premise	x is A'
Implication	IF x is A THEN y is B
Conclusion	y is B'

where A, B, A', B' are fuzzy sets, x and y are linguistic variables. Applying the compositional rule of inference [22], we obtain

$$B' = A' \circ R = A' \circ (A \to B) \tag{2}$$

and

$$\mu_{B'}(y) = \mu_{A' \circ R}(y) = \sup_{x \in X} \left\{ \mu_{A'}(x)^T * \mu_R(x, y) \right\} . \tag{3}$$

The problem is to determine the membership function of the fuzzy relation described by

$$\mu_R(x, y) = \mu_{A \to B}(x, y) \tag{4}$$

based on the knowledge of $\mu_A(x)$ and $\mu_B(y)$. We denote

$$\mu_{A \to B}(x, y) = I(\mu_A(x), \mu_B(y)) , \tag{5}$$

where $I(\cdot) = I_{\text{fuzzy}}(\cdot)$ is a fuzzy implication (logical approach to fuzzy reasoning, see Subsect. "Logical Approach to Fuzzy Inference Reasoning") given in Definition 1 or a t-norm (Mamdani approach to fuzzy reasoning, see Subsect. "Mamdani Approach to Fuzzy Inference Reasoning").

Neuro-fuzzy Systems, Table 1
Fuzzy implications

No	Name	Implication $I_{\text{fuzzy}}(a, b)$
1	Kleene–Dienes (binary)	$\max\{1-a, b\}$
2	Łukasiewicz	$\min\{1, 1-a+b\}$
3	Reichenbach	$1-a+a\cdot b$
4	Fodor	$\begin{cases} 1 & \text{if } a \le b \\ \max\{1-a, b\} & \text{if } a > b \end{cases}$
5	Rescher	$\begin{cases} 1 & \text{if } a \le b \\ 0 & \text{if } a > b \end{cases}$
6	Goguen	$\begin{cases} 1 & \text{if } a = 0 \\ \min\{1, \frac{b}{a}\} & \text{if } a > 0 \end{cases}$
7	Gödel	$\begin{cases} 1 & \text{if } a \le b \\ b & \text{if } a > b \end{cases}$
8	Yager	$\begin{cases} 1 & \text{if } a = 0 \\ b^a & \text{if } a > 0 \end{cases}$
9	Zadeh	$\max\{\min\{a, b\}, 1-a\}$
10	Willmott	$\min\left\{\begin{array}{c} \max\{1-a, b\}, \\ \max\{a, 1-b, \min\{1-a, b\}\} \end{array}\right\}$
11	Dubois–Prade	$\begin{cases} 1-a & \text{if } b = 0 \\ b & \text{if } a = 1 \\ 1 & \text{if otherwise} \end{cases}$

Logical Approach to Fuzzy Inference Reasoning

Definition 1 (see Fodor [12]) A fuzzy implication is a function $I = I_{\text{fuzzy}}\colon [0,1]^2 \to [0,1]$ satisfying the following conditions:

(I1) if $a_1 \le a_3$, then $I(a_1, a_2) \ge I(a_3, a_2)$, for all $a_1, a_2, a_3 \in [0,1]$,
(I2) if $a_2 \le a_3$, then $I(a_1, a_2) \le I(a_1, a_3)$, for all $a_1, a_2, a_3 \in [0,1]$,
(I3) $I(0, a_2) = 1$, for all $a_2 \in [0,1]$ (falsity implies anything),
(I4) $I(a_1, 1) = 1$, for all $a_1 \in [0,1]$ (anything implies tautology),
(I5) $I(1,0) = 0$ (Booleanity).

Selected fuzzy implications satisfying all or some of the above conditions are listed in Table 1.

In this table, Implications 1–4 are examples of an S-implication associated with a t-conorm

$$I(a,b) = S\{1-a, b\}\,. \tag{6}$$

Neuro-fuzzy systems based on fuzzy implications given in Table 1 are called logical systems. Implications (6) and (7) belong to a group of R-implications associated with the t-norm T and given by

$$I(a,b) = \sup_z\{z \mid T\{a,z\} \le b\}\,, \quad a, b \in [0,1]\,. \tag{7}$$

The Zadeh implication belongs to a group of Q-implications given by

$$I(a,b) = S\{N(a), T\{a,b\}\}\,, \quad a, b \in [0,1]\,. \tag{8}$$

It is easy to verify that S-implications and R-implications satisfy all the conditions of Definition 1. However, the Zadeh implication violates conditions I1 and I4, whereas the Willmott implication violates conditions I1, I3 and I4.

Mamdani Approach to Fuzzy Inference Reasoning

In practice we frequently use Mamdani-type operators given by

$$I(a,b) = \min\{a,b\}\,, \quad a, b \in [0,1]\,, \tag{9}$$

$$I(a,b) = a\cdot b\,, \quad a, b \in [0,1] \tag{10}$$

or generally

$$I(a,b) = T\{a,b\}\,, \quad a, b \in [0,1]\,. \tag{11}$$

It should be noted that operators (9)–(11) do not satisfy conditions of Definition 1. Operators (9)–(11) are called "correlation functions" describing strength of connections between antecedents and consequences (see [48]).

Description of Fuzzy Inference Systems

In this description, we consider multi-input-single-output fuzzy system mapping $\mathbf{X} \to \mathbf{Y}$, where $\mathbf{X} \subset \mathbf{R}^n$ and $\mathbf{Y} \subset \mathbf{R}$ (see Fig. 1). The system is composed of a fuzzifier, a fuzzy rule base, a fuzzy inference engine and a defuzzifier. The fuzzifier performs a mapping from the observed crisp input space $\mathbf{X} \subset \mathbf{R}^n$ to a fuzzy set defined in $\mathbf{X}$. The most commonly used fuzzifier is the singleton fuzzifier, which

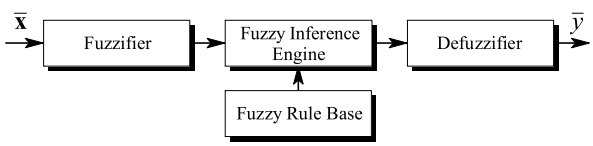

Neuro-fuzzy Systems, Figure 1
Fuzzy inference system

maps $\bar{\mathbf{x}} = [\bar{x}_1, \ldots, \bar{x}_n] \in \mathbf{X}$ into a fuzzy set $A' \subseteq \mathbf{X}$ characterized by the membership function

$$\mu_{A'}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} = \bar{\mathbf{x}} \\ 0 & \text{if } \mathbf{x} \neq \bar{\mathbf{x}} \end{cases} . \tag{12}$$

The fuzzy rule base consists of a collection of N fuzzy IF-THEN rules, aggregated by the disjunction or the conjunction, in the form

$$R^{(k)} : \begin{cases} \text{IF} & x_1 \text{ is } A_1^k \quad \text{AND} \\ & x_2 \text{ is } A_2^k \quad \text{AND} \ldots \\ & x_n \text{ is } A_n^k \\ \text{THEN} & y \text{ is } B^k \end{cases} \tag{13}$$

or

$$R^{(k)} : \text{IF} \quad \mathbf{x} \text{ is } A^k \quad \text{THEN} \quad y \text{ is } B^k , \tag{14}$$

where $\mathbf{x} = [x_1, \ldots, x_n] \in \mathbf{X}$, $y \in \mathbf{Y}$, $A^k = A_1^k \times A_2^k \times \ldots \times A_n^k$, $A_1^k, A_2^k, \ldots, A_n^k$ are fuzzy sets characterized by membership functions $\mu_{A_i^k}(x_i)$, $i = 1, \ldots, n$, $k = 1, \ldots, N$, whereas B^k are fuzzy sets characterized by membership functions $\mu_{B^k}(y)$, $k = 1, \ldots, N$. The firing strength of the kth rule, $k = 1, \ldots, N$, is defined by

$$\tau_k(\bar{\mathbf{x}}) = \mathop{\mathrm{T}}_{i=1}^{n} \{\mu_{A_i^k}(\bar{x}_i)\} = \mu_{A^k}(\bar{\mathbf{x}}) . \tag{15}$$

The fuzzy inference engine determines the mapping from the fuzzy sets in the input space $\mathbf{X}$ to the fuzzy sets in the output space $\mathbf{Y}$. Each of N rules (14) determines a fuzzy set $\bar{B}^k \subseteq \mathbf{Y}$ given by the compositional rule of inference

$$\bar{B}^k = A' \circ (A^k \to B^k) , \tag{16}$$

where $A^k = A_1^k \times A_2^k \times \cdots \times A_n^k$. Fuzzy sets $\bar{B}^k$ are characterized by membership functions expressed by the sup-star composition

$$\mu_{\bar{B}^k}(y) = \sup_{\mathbf{x} \in \mathbf{X}} \{\mu_{A'}(\mathbf{x}) * \mu_{A_1^k \times \cdots \times A_n^k \to B^k}(\mathbf{x}, y)\} , \tag{17}$$

where $*$ can be any operator in the class of t-norms. It is easily seen that for a crisp input $\bar{\mathbf{x}} = [\bar{x}_1, \ldots, \bar{x}_n] \in \mathbf{X}$, i. e., the singleton fuzzifier (12), formula (17) becomes

$$\begin{aligned} \mu_{\bar{B}^k}(y) &= \mu_{A_1^k \times \ldots \times A_n^k \to B^k}(\bar{\mathbf{x}}, y) \\ &= \mu_{A^k \to B^k}(\bar{\mathbf{x}}, y) \\ &= I(\mu_{A^k}(\bar{\mathbf{x}}), \mu_{B^k}(y)) , \end{aligned} \tag{18}$$

where $I(\cdot)$ is correlation between antecedents and consequents given by (11), in the case of Mamdani approach, or fuzzy implication $I_{\text{fuzzy}}(\cdot)$, in the case of logical approach (examples of fuzzy implications are given in Table 1). The aggregation operator, applied in order to obtain the fuzzy set B' based on fuzzy sets $\bar{B}^k$, is the t-norm or t-conorm operator, depending on the type of the fuzzy inference. As a result of the fuzzy reasoning we obtain the fuzzy set B'.

The defuzzifier performs a mapping from the fuzzy set B' to a crisp point $\bar{y}$ in $\mathbf{Y} \subset \mathbf{R}$. The COA (center of area) method is defined by the following formula:

$$\bar{y} = \frac{\int_{\mathbf{Y}} y \cdot \mu_{B'}(y) \mathrm{d}y}{\int_{\mathbf{Y}} \mu_{B'}(y) \mathrm{d}y} \tag{19}$$

or by

$$\bar{y} = \frac{\sum_{r=1}^{N} \bar{y}^r \cdot \mu_{B'}(\bar{y}^r)}{\sum_{r=1}^{N} \mu_{B'}(\bar{y}^r)} , \tag{20}$$

in the discrete form, where $\bar{y}^r$ are centers of the membership functions $\mu_{B^r}(y)$, i. e., for $r = 1, \ldots, N$

$$\mu_{B^r}(\bar{y}^r) = \max_{y \in \mathbf{Y}} \{\mu_{B^r}(y)\} . \tag{21}$$

Logical-Type Neuro-fuzzy Systems

In this approach, function $I(\cdot)$ given by 18 is a fuzzy implication (see Table 1), i. e.

$$I(\mu_{A^k}(\bar{\mathbf{x}}), \mu_{B^k}(\bar{y}^r)) = I_{\text{fuzzy}}(\mu_{A^k}(\bar{\mathbf{x}}), \mu_{B^k}(\bar{y}^r)) . \tag{22}$$

When we use the logical model, the aggregation is carried out by

$$B' = \bigcap_{k=1}^{N} \bar{B}^k . \tag{23}$$

The aggregated output fuzzy set $B' \subseteq \mathbf{Y}$ is computed by the use of a t-norm and is given by

$$\mu_{B'}(\bar{y}^r) = \mathop{\mathrm{T}}_{k=1}^{N} \{\mu_{\bar{B}^k(\bar{y}^r)}\} = \mathop{\mathrm{T}}_{k=1}^{N} \{I_{\text{fuzzy}}(\mu_{A^k}(\bar{\mathbf{x}}), \mu_{B^k}(\bar{y}^r))\} \tag{24}$$

and formula (20) becomes

$$\bar{y} = \frac{\sum_{r=1}^{N} \bar{y}^r \cdot \mathop{\mathrm{T}}_{k=1}^{N} \left\{ I_{\text{fuzzy}} \left(\mathop{\mathrm{T}}_{i=1}^{n} \{\mu_{A_i^k}(\bar{x}_i)\}, \mu_{B^k}(\bar{y}^r) \right) \right\}}{\sum_{r=1}^{N} \mathop{\mathrm{T}}_{k=1}^{N} \left\{ I_{\text{fuzzy}} \left(\mathop{\mathrm{T}}_{i=1}^{n} \{\mu_{A_i^k}(\bar{x}_i)\}, \mu_{B^k}(\bar{y}^r) \right) \right\}} . \tag{25}$$

Example 1 Using Gaussian membership functions

$$\mu_{A_i^r}(x_i) = \exp\left[-\left(\frac{x_i - \bar{x}_i^r}{\sigma_i^r}\right)^2\right], \tag{26}$$

$$\mu_{B^r}(y) = \exp\left[-\left(\frac{y - \bar{y}^r}{\sigma^r}\right)^2\right], \tag{27}$$

dependency (25), Łukasiewicz-type S-implication (see Table 1), and the min-type aggregation of conclusions, we obtain the following description of the neuro-fuzzy system

$$\bar{y} = \frac{\sum_{r=1}^{N} \bar{y}^r \cdot \min_{k=1,\ldots,N}\left\{\min\left[1, 1 - \mathop{T}_{i=1}^{n}\left(\exp\left[-\left(\frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k}\right)^2\right]\right) + \exp\left[-\left(\frac{\bar{y}^r - \bar{y}^k}{\sigma^k}\right)^2\right]\right]\right\}}{\sum_{r=1}^{N} \cdot \min_{k=1,\ldots,N}\left\{\min\left[1, 1 - \mathop{T}_{i=1}^{n}\left(\exp\left[-\left(\frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k}\right)^2\right]\right) + \exp\left[-\left(\frac{\bar{y}^r - \bar{y}^k}{\sigma^k}\right)^2\right]\right]\right\}} \tag{28}$$

Example 2 Using Gaussian membership functions (26) and (27), dependency (25), Kleene–Dienes-type S-implications (see Table 1), and the min-type aggregation of conclusions, we obtain the following description of the neuro-fuzzy system:

$$\bar{y} = \frac{\sum_{r=1}^{N} \bar{y}^r \cdot \min_{k=1,\ldots,N}\left\{\max\left[1 - \mathop{T}_{i=1}^{n}\left(\exp\left[-\left(\frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k}\right)^2\right]\right), \exp\left[-\left(\frac{\bar{y}^r - \bar{y}^k}{\sigma^k}\right)^2\right]\right]\right\}}{\sum_{r=1}^{N} \cdot \min_{k=1,\ldots,N}\left\{\max\left[1 - \mathop{T}_{i=1}^{n}\left(\exp\left[-\left(\frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k}\right)^2\right]\right), \exp\left[-\left(\frac{\bar{y}^r - \bar{y}^k}{\sigma^k}\right)^2\right]\right]\right\}} \tag{29}$$

Example 3 Using Gaussian membership functions (26) and (27), dependency (25), Reichenbach-type S-implications (see Table 1), and the min-type aggregation of conclusions, we obtain the following description of the neuro-fuzzy system:

$$\bar{y} = \frac{\sum_{r=1}^{N} \bar{y}^r \cdot \min_{k=1,\ldots,N}\left\{1 - \mathop{T}_{i=1}^{n}\left(\exp\left[-\left(\frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k}\right)^2\right]\right) \cdot \left(1 - \exp\left[-\left(\frac{\bar{y}^r - \bar{y}^k}{\sigma^k}\right)^2\right]\right)\right\}}{\sum_{r=1}^{N} \cdot \min_{k=1,\ldots,N}\left\{1 - \mathop{T}_{i=1}^{n}\left(\exp\left[-\left(\frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k}\right)^2\right]\right) \cdot \left(1 - \exp\left[-\left(\frac{\bar{y}^r - \bar{y}^k}{\sigma^k}\right)^2\right]\right)\right\}}. \tag{30}$$

Mamdani-Type Neuro-fuzzy Systems

In this approach, function $I(\cdot)$ given by (18) is a t-norm (e. g. minimum or algebraic), i. e.

$$I(\mu_{A^k}(\bar{\mathbf{x}}), \mu_{B^k}(\bar{y}^r)) = T\{\mu_{A^k}(\bar{\mathbf{x}}), \mu_{B^k}(\bar{y}^r)\}\,. \tag{31}$$

In case of the Mamdani approach, the aggregation is carried out by

$$B' = \bigcup_{k=1}^{N} \bar{B}^k\,. \tag{32}$$

The aggregated output fuzzy set $B' \subseteq \mathbf{Y}$ is computed by the use of a t-conorm and is given by

$$\mu_{B'}(\bar{y}^r) = \mathop{S}_{k=1}^{N}\{\mu_{\bar{B}^k}(\bar{y}^r)\} = \mathop{S}_{k=1}^{N}\{T\{\mu_{A^k}(\bar{\mathbf{x}}), \mu_{B^k}(\bar{y}^r)\}\}\,. \tag{33}$$

Consequently, formula (20) takes the form

$$\bar{y} = \frac{\sum_{r=1}^{N} \bar{y}^r \cdot \mathop{S}_{k=1}^{N}\left\{T\left\{\mathop{T}_{i=1}^{n}\{\mu_{A_i^k}(\bar{x}_i)\}, \mu_{B^k}(\bar{y}^r)\right\}\right\}}{\sum_{r=1}^{N} \mathop{S}_{k=1}^{N}\left\{T\left\{\mathop{T}_{i=1}^{n}\{\mu_{A_i^k}(\bar{x}_i)\}, \mu_{B^k}(\bar{y}^r)\right\}\right\}}\,. \tag{34}$$

Obviously, the t-norms used to connect the antecedents in the rules and to connect the antecedents and consequents (correlation function) do not have to be the same. Besides, they can be chosen as differentiable functions as e. g. Yager families [27].

Remark 1 Another Mamdani-type neuro-fuzzy system can be derived using so-called "height defuzzifier"

$$\bar{y} = \frac{\sum_{r=1}^{N} \bar{y}^r \cdot \mu_{\bar{B}^r}(\bar{y}^r)}{\sum_{r=1}^{N} \mu_{\bar{B}^r}(\bar{y}^r)} . \tag{35}$$

For normal fuzzy sets B^r, combining (35), (18) and (31), we obtain

$$\bar{y} = \frac{\sum_{r=1}^{N} \bar{y}^r \cdot \underset{i=1}{\overset{n}{T}} \{\mu_{A_i^r}(\bar{x}_i)\}}{\sum_{r=1}^{N} \underset{i=1}{\overset{n}{T}} \{\mu_{A_i^r}(\bar{x}_i)\}} , \tag{36}$$

and for Gaussian functions (26) and (27) and the product t-norm we get

$$\bar{y} = \frac{\sum_{r=1}^{N} \bar{y}^r \cdot \left[\prod_{i=1}^{n} \left(\exp\left[-\left(\frac{\bar{x}_i - \bar{x}_i^r}{\sigma_i^r}\right)^2\right]\right)\right]}{\sum_{r=1}^{N} \left[\prod_{i=1}^{n} \left(\exp\left[-\left(\frac{\bar{x}_i - \bar{x}_i^r}{\sigma_i^r}\right)^2\right]\right)\right]} . \tag{37}$$

It should be noted that (37) is a well-known formula derived by Wang [58] and used for tasks of modeling and control. It is interesting that formula (37) can be also derived from defuzzifier (20) if assumption (40) holds (see Sect. "Fuzzy Reasoning"). The architecture of the system (37) is depicted in Fig. 2.

Example 4 Using Gaussian membership functions (26) and (27), dependency (34), min-type Mamdani rule, and max-type aggregation of conclusions, we obtain the following description of the neuro-fuzzy system

$$\bar{y} = \frac{\sum_{r=1}^{N} \bar{y}^r \cdot \max_{k=1,\ldots,N} \left\{ \min\left(\prod_{i=1}^{n} \exp\left[-\left(\frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k}\right)^2\right] \cdot \exp\left[-\left(\frac{\bar{y}^r - \bar{y}^k}{\sigma^k}\right)^2\right]\right)\right\}}{\sum_{r=1}^{N} \max_{k=1,\ldots,N} \left\{ \min\left(\prod_{i=1}^{n} \exp\left[-\left(\frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k}\right)^2\right] \cdot \exp\left[-\left(\frac{\bar{y}^r - \bar{y}^k}{\sigma^k}\right)^2\right]\right)\right\}} . \tag{38}$$

Example 5 Using Gaussian membership functions (26) and (27), dependency (34), product-type Mamdani rule (known as the Larsen rule), and the max-type aggregation of conclusions, we obtain the following description of the neuro-fuzzy system:

$$\bar{y} = \frac{\sum_{r=1}^{N} \bar{y}^r \cdot \max_{k=1,\ldots,N} \left\{ \prod_{i=1}^{n} \exp\left[-\left(\frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k}\right)^2\right] \cdot \exp\left[-\left(\frac{\bar{y}^r - \bar{y}^k}{\sigma^k}\right)^2\right]\right\}}{\sum_{r=1}^{N} \max_{k=1,\ldots,N} \left\{ \prod_{i=1}^{n} \exp\left[-\left(\frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k}\right)^2\right] \cdot \exp\left[-\left(\frac{\bar{y}^r - \bar{y}^k}{\sigma^k}\right)^2\right]\right\}} . \tag{39}$$

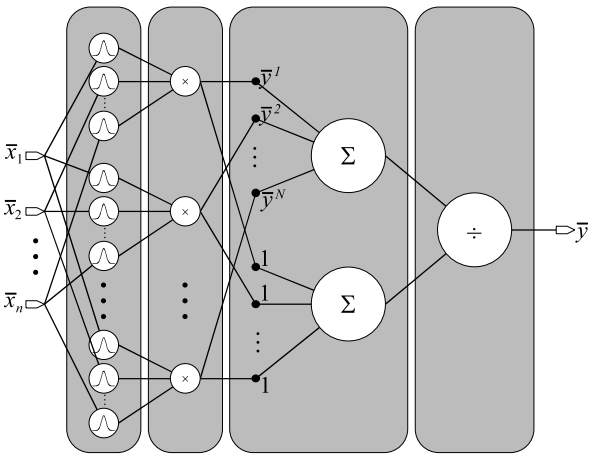

Neuro-fuzzy Systems, Figure 2
Architecture of the neuro-fuzzy system with Mamdani approach described by (37)

Simplified Neuro-fuzzy Systems

The structures presented so far are quite complex and complicated. One possibility of their simplification is to assume, that membership functions of consequent fuzzy sets are sufficiently distant from each other, so that the following assumption holds

$$\mu_{B^k}(\bar{y}^r) \approx 0 , \tag{40}$$

for $k, r = 1, \ldots, N$ and $k \neq r$. Figure 3 shows an example of such fuzzy sets.

It is easy to verify that under assumption (40) the Mamdani-type neuro-fuzzy structure given by (34), derived from defuzzifier (20), takes the simplified form

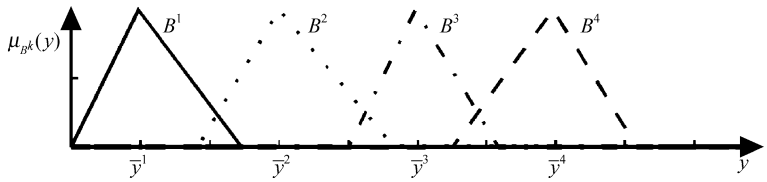

Neuro-fuzzy Systems, Figure 3
Exemplary fuzzy sets satisfying assumption (40)

which is the same as the structure (36) derived from defuzzifier (35). The logical neuro-fuzzy system (25) can also be simplified if condition (40) holds, e. g. the Lukasiewicz neuro-fuzzy system (28) takes the form

$$\bar{y} = \frac{\sum_{r=1}^{N} \bar{y}^r \mathop{T}_{\substack{k=1\\k\neq r}}^{N} \left\{1 - \mathop{T}_{i=1}^{n} \left\{\exp\left[-\left(\frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k}\right)^2\right]\right\}\right\}}{\sum_{r=1}^{N} \mathop{T}_{\substack{k=1\\k\neq r}}^{N} \left\{1 - \mathop{T}_{i=1}^{n} \left\{\exp\left[-\left(\frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k}\right)^2\right]\right\}\right\}} . \quad (41)$$

Let us observe that simplified neuro-fuzzy systems are characterized by three parameters to be found in the process of learning: $\bar{x}_i^k$, σ_i^k and $\bar{y}^k$ for $i = 1, \ldots, n$ and $k = 1, \ldots, N$. In the case of not simplified systems, see (28)–(30), we have to additionally determine parameters σ^k, $k = 1, \ldots, N$, of the output membership functions. The architecture of the system (41) is depicted in Fig. 3.

Takagi–Sugeno Neuro-fuzzy Systems

In the fuzzy Takagi–Sugeno-type model [55], the rules have fuzzy character only in the IF part, whereas in the THEN part, there are functional dependencies

$$\begin{aligned} R^{(r)}: \text{IF} \quad & \mathbf{x} \text{ is } A^r \\ \text{THEN} \quad & y_r = f^{(r)}(x_1, x_2, \ldots, x_n) . \end{aligned} \quad (42)$$

If we assume that the input of the fuzzy system is signal $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n)$, then in order to obtain the output signal $\bar{y}$ of the system, first we will determine

$$T\big(\mu_{A_1^r}(\bar{x}_1), \mu_{A_2^r}(\bar{x}_2), \ldots \mu_{A_n^r}(\bar{x}_n)\big) , \quad r = 1, \ldots, N. \quad (43)$$

The next step is to compute

$$\bar{y}_r = f^{(r)}(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n) , \quad r = 1, \ldots, N . \quad (44)$$

The output signal of the fuzzy Takagi–Sugeno system is a normalized weighted sum of particular inputs $\bar{y}_1, \ldots, \bar{y}_N$, i. e.

$$\bar{y} = \frac{\sum_{r=1}^{N} \bar{y}_r \mathop{T}_{i=1}^{n} \{\mu_{A_i^r}(\bar{x}_i)\}}{\sum_{r=1}^{N} \mathop{T}_{i=1}^{n} \{\mu_{A_i^r}(\bar{x}_i)\}} . \quad (45)$$

In the following part of this subchapter, we will consider the Takagi–Sugeno systems with linear dependencies in consequents of the base of rules, i. e.

$$\begin{aligned} R^{(r)}: \text{IF} \quad & \mathbf{x} \text{ is } A^r \\ \text{THEN} \quad & y_r = c_0^{(r)} + c_1^{(r)} x_1 + \ldots + c_n^{(r)} x_n \end{aligned} \quad (46)$$

for $r = 1, \ldots, N$. It should be noted that if $c_i^{(r)} = 0$, $i = 1, \ldots, n$, then system (45) is reduced to a simplified Mamdani system given by formula (36), and then $c_0^{(r)} = \bar{y}^r$, $r = 1, \ldots, N$.

Example 6 The Takagi–Sugeno system with Gaussian membership functions, linear model (46) and product t-norm, takes the following form

$$\bar{y} = \frac{\sum_{r=1}^{N} \left[\prod_{i=1}^{n} \left(\exp\left[-\left(\frac{\bar{x}_i - \bar{x}_i^r}{\sigma_i^r}\right)^2\right]\right)\right] \cdot \left(c_0^{(r)} + c_1^{(r)} x_1 + \ldots + c_n^{(r)} x_n\right)}{\sum_{r=1}^{N} \left[\prod_{i=1}^{n} \left(\exp\left[-\left(\frac{\bar{x}_i - \bar{x}_i^r}{\sigma_i^r}\right)^2\right]\right)\right]} . \quad (47)$$

Neuro-fuzzy Systems with Weights

We can distinguish fuzzy systems based on the following linguistic models:

(i) Systems based on linguistic model given by (14).
(ii) Systems based on linguistic model given by (14), and additionally with weights $w_k^{\text{agr}} \in [0, 1]$, $k = 1, \ldots, N$, describing the importance of each rule:

$$R^{(k)}: \begin{bmatrix} \text{IF} \quad x_1 \text{ is } A_1^k \quad \text{AND} \quad \ldots \\ \text{AND} \quad x_n \text{ is } A_n^k \\ \text{THEN} \quad y \text{ is } B^k \end{bmatrix} \left(w_k^{\text{agr}}\right) . \quad (48)$$

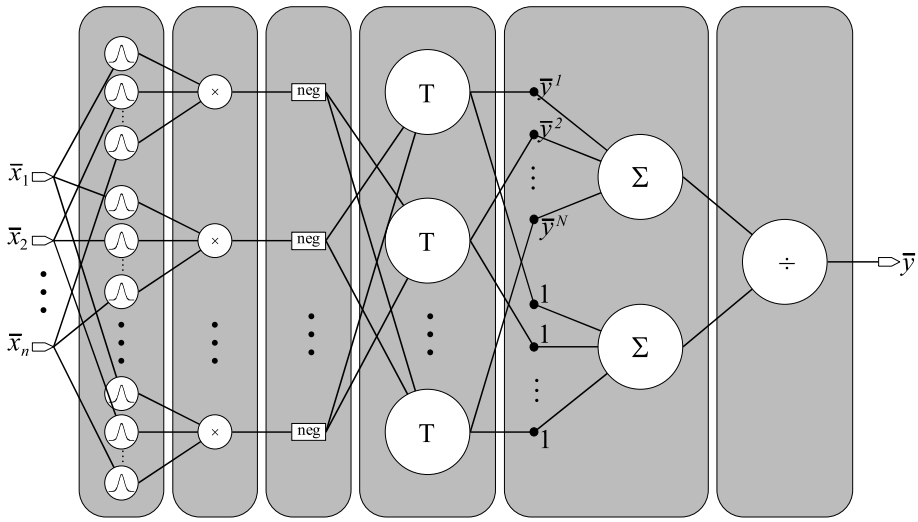

Neuro-fuzzy Systems, Figure 4
Architecture of the logical neuro-fuzzy system with Lukasiewicz implication described by (41)

(iii) Systems based on linguistic model given by (14), with weights $w_k^{\text{agr}} \in [0,1]$, $k = 1,\ldots,N$, describing the importance of each rule and weights $w_{i,k}^{\tau} \in [0,1]$, $k = 1,\ldots,N$, $i = 1,\ldots,n$, describing the importance of antecedents:

$$R^{(k)} : \begin{bmatrix} \text{IF} \quad x_1 \text{ is } A_1^k\left(w_{1,k}^{\tau}\right) \quad \text{AND} \quad \ldots \\ \text{AND} \quad x_n \text{ is } A_n^k\left(w_{n,k}^{\tau}\right) \\ \text{THEN} \quad y \text{ is } B^k \end{bmatrix} \left(w_k^{\text{agr}}\right) . \tag{49}$$

In order to incorporate the weights in models (48) and (49) into description of neuro-fuzzy systems, we can use the following formula (see [48,50]):

$$T^*\{a_1,\ldots,a_n; w_1,\ldots,w_n\} = \mathop{T}_{i=1}^{n} \{1 - w_i(1-a_i)\} , \tag{50}$$

where $T\{\cdot\}$ is any t-norm, and the weights meet the condition $w_i \in [0,1]$, $i = 1,\ldots,n$. Observe that if we assume that $w_i = 1$, $i = 1,\ldots,n$, then function $T^*\{\cdot\}$ is reduced to t-norm $T\{\cdot\}$. Analogously, we define

$$S^*\{a_1,\ldots,a_n; w_1,\ldots,w_n\} = \mathop{S}_{i=1}^{n} \{w_i a_i\} . \tag{51}$$

where S is any t co-norm.

Example 7 Simplified neuro-fuzzy system (41) with Lukasiewicz fuzzy implication and weights describing the importance of each rule takes the form

$$\bar{y} = \frac{\sum_{r=1}^{N} \bar{y}^r \mathop{T^*}_{\substack{k=1\\k\neq r}}^{N} \left\{ 1 - \mathop{T}_{i=1}^{n} \left\{ \exp\left[-\left(\frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k} \right)^2 \right] \right\} , w_k \right\}}{\sum_{r=1}^{N} \mathop{T^*}_{\substack{k=1\\k\neq r}}^{N} \left\{ 1 - \mathop{T}_{i=1}^{n} \left\{ \exp\left[-\left(\frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k} \right)^2 \right] \right\} , w_k \right\}} . \tag{52}$$

Example 8 Simplified neuro-fuzzy system (41) with Lukasiewicz fuzzy implication, weights describing the importance of each rule and weights describing the importance of antecedents takes the form

$$\bar{y} = \frac{\sum_{r=1}^{N} \bar{y}^r \mathop{T^*}_{\substack{k=1\\k\neq r}}^{N} \left\{ 1 - \mathop{T^*}_{i=1}^{n} \left\{ \exp\left[-\left(\frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k} \right)^2 \right] , w_{i,k} \right\} , w_k \right\}}{\sum_{r=1}^{N} \mathop{T^*}_{\substack{k=1\\k\neq r}}^{N} \left\{ 1 - \mathop{T^*}_{i=1}^{n} \left\{ \exp\left[-\left(\frac{\bar{x}_i - \bar{x}_i^k}{\sigma_i^k} \right)^2 \right] , w_{i,k} \right\} , w_k \right\}} . \tag{53}$$

Neuro-fuzzy Systems for Pattern Classification

We will explain how to modify linguistic model (13) to solve multi-classification problems. Let $[x_1,\ldots,x_n]$ be the vector of features of an object ν . Let $\Omega = \{\omega_1,\ldots,\omega_M\}$ be a set of classes. The knowledge is represented by a set of

N rules in the form

$$R^{(k)}: \begin{cases} \text{IF} & x_1 \text{ is } A_1^k \quad \text{AND} \\ & x_2 \text{ is } A_2^k \quad \text{AND} \ldots \\ & x_n \text{ is } A_n^k \\ \text{THEN} & \nu \in \omega_1\left(z_1^k\right), \\ & \nu \in \omega_2\left(z_2^k\right), \ldots \\ & \nu \in \omega_M\left(z_M^k\right), \end{cases}, \tag{54}$$

where z_1^k, $j = 1, \ldots, M$, $k = 1, \ldots, N$, are interpreted as "support" for class ω_j given by rule $R^{(k)}$. We will now redefine description (54). Let us introduce vector $\mathbf{z} = [z_1, \ldots, z_M]$, where z_j, $j = 1, \ldots, M$, is the "support" for class ω_j given by all M rules. We can scale the support values to the interval [0,1], so that z_j is the membership degree of an object ν to class ω_j, according to all M rules. The rules are represented by

$$R^{(k)}: \begin{cases} \text{IF} & x_1 \text{ is } A_1^k \quad \text{AND} \\ & x_2 \text{ is } A_2^k \quad \text{AND} \ldots \\ & x_n \text{ is } A_n^k \\ \text{THEN} & z_1 \text{ is } B_1^k \quad \text{AND} \\ & z_2 \text{ is } B_2^k \quad \text{AND} \ldots \\ & z_M \text{ is } B_M^k \end{cases}, \tag{55}$$

and formula (20) adopted for classification takes the form

$$\bar{z}_j = \frac{\sum_{k=1}^{N} \bar{z}_j^k \mu_{B'_j}\left(\bar{z}_j^k\right)}{\sum_{k=1}^{N} \mu_{B'_j}\left(\bar{z}_j^k\right)}, \tag{56}$$

where $\bar{z}_j^k$ are centers of fuzzy sets B_j^k, $j = 1, \ldots, M$, $k = 1, \ldots, N$.

Learning

Fuzzy and neuro-fuzzy system knowledge is expressed in the form of fuzzy rules (14). The learning process tries to pick the right number of rules and location of linguistic values (fuzzy sets defined in linguistic variables) to act like an expert or device performing given task. Traditional fuzzy systems are devised to work on the basis of knowledge formulated by an expert in the form of fuzzy rules in nearly natural language. The expert is not always able to formulate his or her knowledge or he or she is not available at all. In such cases we can rely on knowledge acquisition. The data acquired are used to supervised learning and then the learning process becomes similar to that known from neural networks, especially gradient methods. To determine all parameters of neuro-fuzzy systems using numerical data, genetic and evolutionary algorithms can be also used. As knowledge in fuzzy systems is structured in the form of rules we can also employ other methods for learning from numerical data, e. g. clustering algorithms with most often used fuzzy c-means algorithm. We can also utilize any method for generating rules from data, e. g. [59].

Having given the learning dataset of the pair $(\bar{\mathbf{x}}, d)$ where d is desired response of the system, we can use the following error measure:

$$Q(\bar{\mathbf{x}}, d) = \tfrac{1}{2}[\bar{y}(\bar{\mathbf{x}}) - d]^2 . \tag{57}$$

Our goal is to minimize the error by determining the value of parameters $\bar{x}_i^r, \sigma_i^r$ (antecedent Gaussian fuzzy sets) and $\bar{y}^r$ (consequent singleton). The parameters can be treated like weights in neural networks and in learning iteration (step) t the value of any weight can be defined by

$$p_i^k(t+1) = p_i^k(t) - \eta \frac{\partial Q(\bar{\mathbf{x}}, d; t)}{\partial p_i^k(t)} . \tag{58}$$

In the case of the simplest neuro-fuzzy system given by (37), we obtain the following formulas for gradients in recursive procedure (58)

$$\frac{\partial Q}{\partial \bar{x}_h^j} = -2\varepsilon \frac{\left(\bar{x}_h^j - \bar{x}_h\right) \prod_{i=1}^{n} \mu_{A_i^j}(\bar{x}_i) \sum_{r=1}^{N} \left(\bar{y}^j - \bar{y}^r\right) \prod_{i=1}^{n} \mu_{A_i^r}(\bar{x}_i)}{\left\{\sigma_h^j \sum_{r=1}^{N} \prod_{i=1}^{n} \mu_{A_i^r}(\bar{x}_i)\right\}^2}, \tag{59}$$

$$\frac{\partial Q}{\partial \sigma_h^j} = 2\varepsilon \frac{\left(\bar{x}_h^j - \bar{x}_h\right) \prod_{i=1}^{n} \mu_{A_i^j}(\bar{x}_i) \sum_{r=1}^{N} \left(\bar{y}^j - \bar{y}^r\right) \prod_{i=1}^{n} \mu_{A_i^r}(\bar{x}_i)}{\left\{\sigma_h^j \sum_{r=1}^{N} \prod_{Zi=1}^{n} \mu_{A_i^r}(\bar{x}_i)\right\}^2}, \tag{60}$$

$$\frac{\partial Q}{\partial \bar{y}^j} = \varepsilon \frac{\prod_{i=1}^{n} \mu_{A_i^j}(\bar{x}_i)}{\sum_{r=1}^{N} \prod_{i=1}^{n} \mu_{A_i^r}(\bar{x}_i)}, \tag{61}$$

where $\varepsilon = \bar{y} - d$ is an error calculated in the last layer of the system and $j = 1, \ldots, n$, $h = 1, \ldots, N$.

Criteria Isolines Method

In this section, we will first present the Akaike criterion, initially applied to the order estimation of autoregression processes, and next it will be adapted to evaluate the effectiveness of neuro-fuzzy systems. By the effectiveness of operation of a neuro-fuzzy system, we shall understand the precision (accuracy) of operation achieved by such a system, (expressed by mean squared error or by the number of wrongly classified samples) in the context of its size. By the system size we shall understand the number of all parameters that are subject to learning. We shall also present the concept of the so-called criteria isolines, which allow one to solve the problem of the compromise between the system accuracy and the number of parameters describing this system.

Neuro-fuzzy system should be characterized by the smallest possible error but at the same time should be as simple as possible. It should be remembered that systems with smaller number of trained parameters are characterized among others by better generalization capabilities. Here, we should mention the so-called parsimony principle [52]. This principle is very useful when determining the appropriate order of the model. It may be formulated as follows: from between two alternative and satisfactory models, we shall choose the one which contains less independent parameters. This principle remains compliant with common sense: "Do not enter any additional parameters into the process description unless they are necessary".

Estimation methods of the system order have been best developed for autoregression processes [25,35]. Time series $u(n), u(n-1), \ldots, u(n-p)$ is an autoregression process of order p, if the difference equation is satisfied

$$u(n) + \alpha_1 u(n-1) + \cdots + \alpha_p u(n-p) = e(n) \tag{62}$$

or equivalently

$$u(n) = -\sum_{k=1}^{p} \alpha_k u(n-k) + e(n)\,, \tag{63}$$

where $\alpha_1, \ldots, \alpha_p$ are process coefficient, while $e(n)$ is the white noise

$$E[e(n)e(m)] = \begin{cases} \sigma^2 & \text{dla} \quad n = m \\ 0 & \text{dla} \quad n \neq m \end{cases}. \tag{64}$$

In the autoregression theory, criteria allowing one to estimate the order of predictor p, determining first the prediction error $\hat{Q}_p$ based on the learning sequence of the length M, are well known. The most important is the Akaike information criterion (AIC), Schwarz method and the final prediction error (FPE) method [25].

The Akaike estimation order prediction method will be adapted now to the evaluation of fuzzy systems. Thanks to this, search for the desired fuzzy system based on two criteria (number of parameters and mean square error) will come down to one selected criterion, i. e. AIC. It has been adapted for the needs of evaluation of neuro-fuzzy systems in the following form:

$$\mathrm{AIC}(p, \hat{Q}_p) = M \ln \hat{Q}_p + 2p\,, \tag{65}$$

where p is the number of system parameters subject to learning (number of parameters of all membership functions and number of all weights if they occur in a given system), $\hat{Q}_p$ is the measure of error used in simulations, and M is the number of samples in a learning sequence. The product $M \cdot n$ may, therefore, be treated as a measure of size of the problem being solved. Table 2 contain the computed values of criteria for particular tested structures in case of the learning and testing sequence used in the polymerization problem [56]. Figure 5 illustrates the coordinates of the points corresponding to particular neuro-fuzzy systems tested. The coordinate p defines the number of parameters of a given system; coordinate Q defines the error with which the system realized the problem to be solved.

The criteria isolines present constant values of the AIC criterion, with different values of the error and the number of parameters. Such an approach allows one to solve the problem of the compromise between the system operation error and the number of parameters describing this system. Points located on the criteria isolines with the same values of AIC criterion characterize the neuro-fuzzy systems making up the Pareto set. In the Pareto set, none of the two values of contradictory criteria may be improved (mean square error versus system size), without worsening the other one. Points located on the criteria isolines with the smallest values of AIC criterion characterize the neuro-fuzzy systems which have been called suboptimal ones. The suboptimal neuro-fuzzy systems presented in graphs ensure the smallest value of criteria within tested structures (the terminology "optimum systems" is not used as all possible structures have not been tested). From Table 2 it follows that for both learning and testing sequences the best, in the sense of the Akaike criterion, is the structure no. 1 (Larsen simplified) followed by the structure no. 29 (Zadeh with weights of rules).

Future Directions

Mankind creates and collects large volumes of data nowadays. We deal with an enormous amount of, often multidimensional, data. We can find such data in medicine

Neuro-fuzzy Systems, Table 2
Value of the Akaike criterion for the polymerization problem for the learning and testing sequence

No.	Structure	Polymerization				
		p	Learning sequence		Testing sequence	
			Error	AIC	Error	AIC
1	Larsen simplified	42	0.0042	–299.09	0.0045	–294.26
2	Larsen simplified with weights of rules	48	0.0039	–292.27	0.0041	–288.77
3	Larsen simplified with weights of inputs and rules	66	0.0031	–272.34	0.0033	–267.97
4	Łukasiewicz simplified	42	0.0059	–275.30	0.0063	–270.70
5	Łukasiewicz simplified with weights of rules	48	0.0039	–292.27	0.0041	–288.77
6	Łukasiewicz simplified with weights of inputs and rules	66	0.0037	–259.96	0.0039	–256.27
7	Zadeh simplified	42	0.0049	–288.30	0.0053	–282.80
8	Zadeh simplified with weights of rules	48	0.0041	–288.77	0.0044	–283.83
9	Zadeh simplified with weights of inputs and rules	66	0.0038	–258.09	0.0040	–254.50
10	Binary	48	0.0063	–258.70	0.0067	–254.40
11	Binary with weights of rules	54	0.0054	–257.49	0.0057	–253.71
12	Binary with weights of inputs and rules	72	0.0036	–249.88	0.0038	–246.09
13	Larsen	48	0.0049	–276.30	0.0052	–272.14
14	Larsen with weights of rules	54	0.0043	–273.44	0.0045	–270.26
15	Larsen with weights of inputs and rules	72	0.0035	–251.85	0.0038	–246.09
16	Łukasiewicz	48	0.0065	–256.52	0.0069	–252.34
17	Łukasiewicz with weights of rules	54	0.0041	–276.77	0.0045	–270.26
18	Łukasiewicz with weights of inputs and rules	72	0.0038	–246.09	0.0041	–240.77
19	Mamdani	48	0.0041	–288.77	0.0043	–285.44
20	Mamdani with weights of rules	54	0.0039	–280.27	0.0042	–275.09
21	Mamdani with weights of inputs and rules	72	0.0034	–253.88	0.0037	–247.96
22	Reichenbach	48	0.0040	–290.50	0.0044	–283.83
23	Reichenbach with weights of rules	54	0.0037	–283.96	0.0039	–280.27
24	Reichenbach with weights of inputs and rules	72	0.0034	–253.88	0.0037	–247.96
25	Willmott	48	0.0056	–266.95	0.0060	–262.12
26	Willmott with weights of rules	54	0.0047	–267.21	0.0049	–264.30
27	Willmott with weights of inputs and rules	72	0.0039	–244.27	0.0043	–237.44
28	Zadeh	48	0.0038	–294.09	0.0043	–285.44
29	Zadeh with weights of rules	54	0.0030	–298.64	0.0033	–291.97
30	Zadeh with weights of inputs and rules	72	0.0028	–267.47	0.0031	–260.34
31	Takagi–Sugeno	60	0.0034	–277.88	0.0036	–273.88
32	Takagi–Sugeno with weights of rules	66	0.0031	–272.34	0.0034	–265.88
33	Takagi–Sugeno with weights of inputs and rules	84	0.0030	–238.64	0.0033	–231.97

and health care, business, manufacturing and finance, science and telecommunication. Neuro-fuzzy systems have to cope with these large, often multidimensional and incomplete, data sets.

In case of real world data, learning systems have to face with a problem of missing data, i. e. missing certain features (input variables). Such shortages can be replaced by the average or modal value, by a value provided by the user or by a value resulting from other records. Good solution in this case is provided by information systems based on rough sets. Such systems work very similarly to classification made by humans – having data with missing values they perform classification, increasing at the same time decision uncertainty level. In case of lack of even one feature, fuzzy systems do not work properly. Fusion of fuzzy and rough sets seems to be a promising field and in the near future rough-neuro-fuzzy systems are expected to emerge.

The development of single learning systems comes slowly to a saturation point and one of ways for further development is creating ensembles of these systems. The accuracy and generalization ability is nearly always improved comparing with a single system. Learning systems can be combined using different learning data, different systems and different aggregation methods. There are sev-

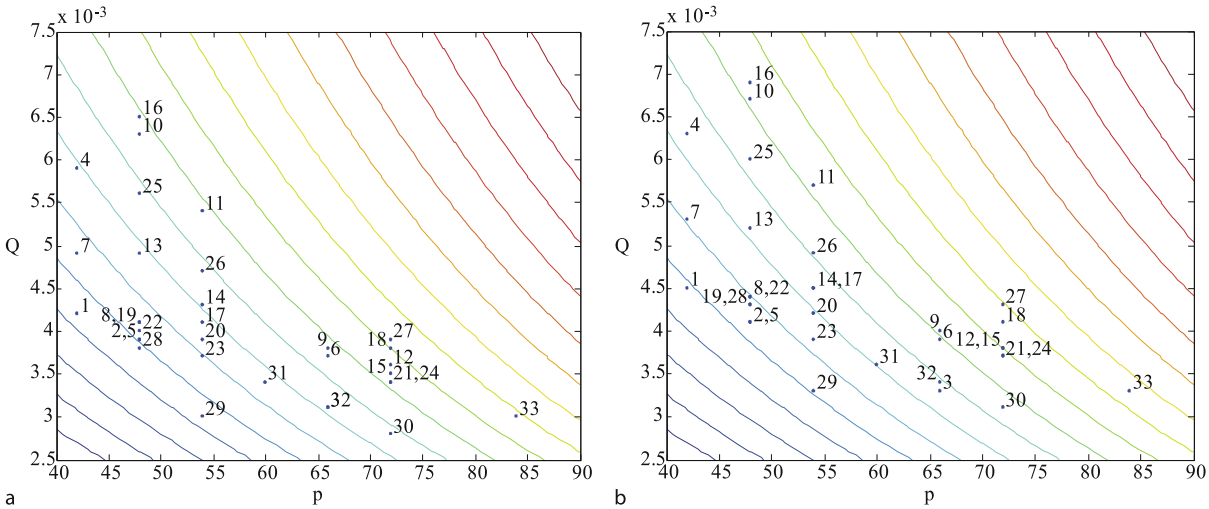

Neuro-fuzzy Systems, Figure 5

Criteria isolines: results obtained by particular systems for the Akaike criterion for the polymerization problem: a learning sequence, b testing sequence

eral methods for creating ensembles of learning systems. The most popular are bagging and boosting which are meta-algorithms allowing use of nearly any learning for ensemble members. After adding a new member to the ensemble, new learning data are created taking into account learning efficiency for every learning sample. In consecutive steps samples which are difficult to learn have more influence on learning. In the most popular AdaBoost algorithm the influence is achieved by assigning weights to learning samples according to their previous performance. As a result of learning of such an ensemble, one will obtain several fuzzy rule bases. To merge the rule bases into one knowledge base special algorithms for simplification of such large fuzzy rule bases should be developed.

The engineering community has accepted fuzzy logic and neuro-fuzzy systems for use in home appliances, automotive industry, robotics, cameras, air conditioning, image recognition, financial engineering, etc. The number and diversity of the applications is still growing and still there are many challenging problems to be solved.

Acknowledgments

This work was supported in part by the Foundation for Polish Science (Professorial Grant 2005–2008) and the Polish Ministry of Science and Higher Education (Special Research Project 2006–2009 and Polish-Singapore Research Project 2008–2010) and by science funds for 2007–2010 as research project No. N N516 1669 33 and No. N N516 1155 33.

Bibliography

1. Aliev RA, Aliev RR (2001) Soft computing and its applications. World Scientific Publishing, Singapore
2. Berenji HR, Khedkar P (1992) Learning and tuning fuzzy logic controllers through reinforcements. IEEE Trans. Neural Networks, vol 3. pp 724–740, October
3. Bubnicki Z (2001) Uncertain variables and their application to decision making. IEEE Trans. on SMC, Part A: Systems and Humans, vol 31, pp 587–596
4. Bubnicki Z (2002) Uncertain logics, variables and systems. Springer, Berlin-London-New York
5. Bubnicki Z (2002) A unified approach to descriptive and prescriptive concepts in uncertain decision systems, Systems Analysis Modeling Simulation, vol 42, issue 3. Gordon and Breach Science Publishers, Newark, pp 331–342
6. Chen MY, Linkens DA (2001) A systematic neuro-fuzzy modeling framework with application to material property prediction. IEEE Trans. on Fuzzy Systems, vol 31, pp 781–790
7. Chuang CC, Su SF, Chen SS (2001) Robust TSK fuzzy modeling for function approximation with outliers. IEEE Trans. on Fuzzy Systems, vol 9, pp 810–821, December
8. Corcoran AL, Sen S (1994) Using real-valued genetic algorithms to evolve rule sets for classification. Proc. of the 1st IEEE Conf. Evolut. Computat., Orlando FL, June pp 120–124
9. Czogała E, Łęski (2000) J fuzzy and neuro-fuzzy intelligent systems, Physica-Verlag Company, Heidelberg, New York
10. Duan JC, Chung FL (2001) Cascaded fuzzy neural network based on syllogistic fuzzy reasoning. IEEE Trans. on Fuzzy Systems, vol 9, pp 293–306
11. Eubank RL (1999) Nonparametric regression and spline smoothing. Marcel Dekker, New York
12. Fodor JC (1991) On fuzzy implication operators, Fuzzy Sets and Systems, vol 42, issue 3. Elsevier, Amsterdam, pp 293–300
13. Fogel DB (1995) Evolutionary computation: towards a new philosophy of machine intelligence. IEEE Press, New York
14. Fuller R (2000) Introduction to neuro-fuzzy systems, advances in soft computing. Physica-Verlag, New York

15. Gaweda AE, Żurada JM (2000) Fuzzy neural network with relational fuzzy rules. Proc. of the Intern. Joint Conference on Neural Networks IJCNN'2000, vol 5, pp 3–8, Como, Italy, July 23–27
16. Gaweda AE, Żurada JM (2001) Data-driven design of fuzzy system with relational input partition. Proc of the Int Conference on Fuzzy Systems FUZZ-IEEE'2001, Melbourne, Australia, December 2–5
17. Hirota K (1993) Industrial Applications of Fuzzy Technology. Springer, Tokyo, Berlin, Heidelberg, New York
18. Jang JSR (1993) ANFIS: Adaptive-network-based fuzzy inference system. IEEE Trans. Syst., Man, Cybern., vol 23, pp 665–685, June
19. Jang JSR, Sun CT (1995) Neuro-fuzzy modeling and control. Proc IEEE, vol 83, pp 378–406, March
20. Jang JS, Sun CT, Mizutani E (1997) Neuro-fuzzy and soft computing. Prentice Hall, Englewood Cliffs
21. Juang C-F, Lin C-T (1998) An on-line self-constructing neural fuzzy inference network and its applications. IEEE Trans. on Fuzzy Systems, vol 6, pp 12–32, February
22. Kacprzyk J (1997) Multistage fuzzy control. Wiley, Chichester
23. Kasabov N (1996) Foundations of neural networks, fuzzy systems and knowledge engineering. The MIT Press CA, Cambridge
24. Kasabov N (2002) DENFIS: dynamic evolving neural-fuzzy inference system and its application for time-series prediction. IEEE Trans. on Fuzzy Systems, vol 10, pp 144–154, April
25. Kay SM (1988) Modern spectral estimation. Theory and application. Prentice Hall, Englewood Cliffs
26. Kecman V (2001) Learning and soft computing. MIT Press, Cambridge
27. Klement EP, Mesiar R, Pap E (2000) Triangular norms. Kluwer, Dordrecht
28. Lee K-M, Kwak D-H, Lee-Kwang H (1994) A fuzzy neural network model for fuzzy inference and rule tuning. Int J Uncertainty, Fuzziness and Knowledge-Based Systems, vol 2, no. 3, pp 265–277
29. Lee K-M, Kwak D-H, Lee-Kwang H (1996) Fuzzy inference neural network for fuzzy model tuning. IEEE Trans on Systems, Man, and Cybernetics. Part B, vol 26, No. 4, pp 637–645
30. Lin CT (1994) Neural fuzzy control systems with structure and parameter learning. World Scientific, Singapore
31. Lin CT, Lee CSG (1991) Neural-network-based fuzzy logic control and decision system. IEEE Trans Comput, vol 40, pp 1320–1336, December
32. Lin CT, Lee GCS (1997) Neural fuzzy systems a neuro-fuzzy synergism to intelligent systems. Prentice Hall, Englewood Cliffs
33. Lin CT, Lu YC (1995) A neural fuzzy systems with linguistic teaching signals. IEEE Trans on Fuzzy Systems, vol 3, pp 169–189, May
34. Lin Y, Cunningham GA III (1995) A new approach to fuzzy-neural system modeling. IEEE Trans. on Fuzzy Systems, vol 3, pp 190–198, May
35. Marple SL Jr (1987) Digital spectral analysis with applications. Prentice Hall, Englewood Cliffs
36. Michalewicz Z (1992) Genetic algorithms + data structures = evolution programs. Springer, Berlin
37. Mouzouris GC, Mendel JM (1997) Nonsingleton fuzzy logic systems: Theory and application. IEEE Trans on Fuzzy Systems, vol 5, No. 1, pp 56–71
38. Nauck D, Kruse R (1996) Designing neuro-fuzzy systems through back-propagation. In: Pedrycz W (ed) Fuzzy modeling: paradigms and practice. Kluwer, Boston, pp 203–228
39. Nauck D, Kruse R (1999) Neuro-fuzzy systems for function approximation. Fuzzy Sets and Systems, vol 101, pp 261–271
40. Nauck D, Klawon F, Kruse R (1997) Foundations of neuro-fuzzy systems. Wiley, Chichester
41. Nie J, Linkens D (1995) Fuzzy-Neural Control. Principles, algorithms and applications. Prentice Hall, New York, London
42. Pagan A, Ullah A (1999) Nonparametric econometrics. Cambridge Univ. Press, London
43. Pawlak Z (1982) Rough sets. Int J Inform Comput Sci, vol 11, no. 341
44. Pawlak Z (1991) Rough sets. Theoretical aspects of reasoning about data. Kluwer, Dordrecht
45. Pedrycz W (1992) Fuzzy neural networks with reference neurons as pattern classifiers. IEEE Trans Neural Networks, vol 3, no. 5, pp 770–775
46. Roubos H, Setnes M (2001) Compact and transparent fuzzy models and classifiers through iterative complexity reduction, IEEE Trans. on Fuzzy Systems, vol 9, pp 516–524, August
47. Rutkowska D (2002) Neuro-fuzzy architectures and hybrid learning. Springer, Heidelberg
48. Rutkowski L (2004) Flexible neuro-fuzzy systems. Kluwer, Norwell
49. Rutkowski L, Cpałka K (2000) Flexible structures of neuro-fuzzy systems, Quo Vadis Computational Intelligence, Studies in Fuzziness and Soft Computing, vol 54. Springer, Berlin, pp 479–484
50. Rutkowski L, Cpałka K (2003) Flexible neuro-fuzzy systems. IEEE Trans. Neural Networks, vol 14, pp 554–574
51. Rutkowski L, Rafajłowicz E (1989) On global rate of convergence of some nonparametric identification procedures. IEEE Trans. on Automatic Control, vol AC-34, no.10, pp 1089–1091
52. Söderström T, Stoica P (1989) System identification. Prentice-Hall, London
53. Tadeusiewicz R (1993) Neural networks RM. Academic Publishing House, Warsaw (in Polish)
54. Tadeusiewicz R (1998) Elementary introduction to neural networks with computer programs. Academic Publishing House, Warsaw (in Polish)
55. Takagi T, Sugeno M (1985) Fuzzy identification of systems and its application to modeling and control. IEEE Trans Systems, Man, and Cybernetics, vol 15, pp 116–132
56. UCI respository of machine learning databases, Available online: http://ftp.ics.uci.edu/pub/machine-learning-databases/
57. Wang JS, Lee CSG (2002) Self-adaptive neuro-fuzzy inference systems for classification applications. IEEE Trans. on Fuzzy Systems, vol 10, pp 790–802
58. Wang LX (1994) Adaptive Fuzzy Systems and Control. PTR Prentice Hall, Englewood Cliffs
59. Wang LX, Mendel JM (1992) Generating fuzzy rules by learning from examples, IEEE Transactions on Systems. Man and Cybernetics, vol 22, no. 6, pp 1414–1427
60. Yager RR (1990) Fuzzy logic controller structures. Proc. SPIE Symp. Laser Sci. Optics Appl. 368–378
61. Yager RR (1992) A general approach to rule aggregation in fuzzy logic control. Appl Intelligence, vol 2, pp 333–351
62. Zadeh LA (1965) Fuzzy sets. Information and Control, vol 8, no. 3, pp 338–353
63. Żurada JM (1992) Introduction to artificial neural systems. West Publishing Company

Neuronal Dynamics

Nicolas Brunel[1,2], Vincent Hakim[3]
[1] Laboratoire de Neurophysique et Physiologie, Université Paris Descartes, Paris, France
[2] UMR 8119, CNRS, Paris, France
[3] Laboratoire de Physique Statistique, UMR 8550, CNRS, Paris, France

Article Outline

Glossary
Definition of the Subject
Introduction
Types of Modeling and Theoretical Tools
Intrinsic Network Dynamics
Stimulus Driven Dynamics
Future Directions
Bibliography

Glossary

Action potential or Spikes: electrical pulses of an amplitude of about 100 mV that travel along nerve fibers.

Field potential an electrical signal recorded extracellularly which arises from the synchronized activity of many cells.

Hippocampus one of the most studied area of the mammalian nervous system which is part of the limbic system and is involved in learning and memory.

Neuron the main excitable cells of nerve tissue.

Network an ensemble of synaptically connected cells.

Synapse the specialized junction between two neurons where the action potential voltage transient in the presynaptic cell is transmitted to the post-synaptic cell via neurotransmitter release (chemical synapse) or direct electrical connection (electrical synapse).

Definition of the Subject

How information is processed by nervous systems is a question of major interest, with far-reaching implications for domains as diverse as medicine and philosophy. It is however still far from understood despite much work and progress during the last sixty years. Neurons are cells which exhibit diverse dynamical behaviors. Our present partial view makes it clear that, besides physiology of single neurons and anatomy of neural systems, understanding the dynamics of coupled neuron assemblies and their collective dynamics is of utmost importance. This requires supplementing the tools of experimental biology, that themselves are making impressive progress, by modeling and theoretical analysis.

Introduction

Since the first electrical recordings, it has been noted that the mammalian brain exhibits diverse patterns of activity [14]. Oscillations at various frequencies are prominent features of human electroencephalograms that is, voltage signals recorded from the scalp [19]. Their dominant frequencies are state and task dependent shifting from slow "delta" frequencies (1–4 Hz) in certain stages of sleep, to "alpha" frequencies (8–13 Hz) in quiet wakefulness or "beta" frequencies (13–30 Hz) in attentive immobility. Intracranial recordings of field potential or neuron activity in animals have revealed higher frequency components from 40 Hz to more than 200 Hz and have also shown that neural rhythms depend on the neural structures from which they originate [3]. For instance, in the rat hippocampus, rhythms at theta (4–8 Hz) and gamma (30–100 Hz) frequencies are observed during exploratory behaviors whereas intermittent fast oscillations (100–200 Hz) are recorded during awake immobility and consummatory behaviors [34]. In the rat olfactory bulb, odorants induce strong gamma oscillations [4].

Electrophysiological recordings, visualization, pharmacological and genetic manipulations are important experimental tools for observing and assessing the mechanisms underlying this complex neural activity. However, analyzing the dynamics of a large ensemble of coupled dynamical elements is a difficult task. Modeling, theoretical analysis and computer simulations are proving increasingly helpful to interpret experimental data, as it becomes more precise and extensive, and to direct further experiments.

In the following, we first review the different types of models that are currently most useful in analyzing network dynamics. We then consider various topics to which theoretical modeling has significantly contributed. We start in Sect. "Intrinsic Network Dynamics" by discussing spontaneous neural activity, that is, activity not directly induced by stimulus-processing. Experimentally, "background" spontaneous activity appears to consist of neurons firing irregularly and at low rates and poses a first modeling challenge. The coexistence of different attractors, a question of central interest in different contexts, is considered next. We then review the different mechanisms that are thought to give rise to the various time scales of neural dynamics. In Sect. "Stimulus Driven Dynamics", we examine how stimulus presentation can modify neural dynamics and discuss various ways in which

dynamics can lead to information processing. We begin by discussing two contrasting views: classically, sensory processing has been thought to proceed by a feedforward and stimulus-driven dynamic, whereas some more recent modeling and experiments suggest that stimuli provide only a weak bias on endogenous dynamics. We then examine various other ways in which dynamics can contribute to information processing and memory. The precise role of oscillations at diverse frequencies in information processing is still ill-understood. We briefly discuss two of the better understood structures and several proposals. We finally conclude, in Sect. "Future Directions", by pointing out some approaches that appear to bear promises of future progress.

Types of Modeling and Theoretical Tools

The first task of a modeler in computational or theoretical neuroscience is to specify a particular model to be investigated. Choosing which model depends, of course, on the nature of the problem one is interested in, as well as on the availability of relevant experimental data. For example, if one is interested in understanding the origin of the irregularity of spike trains in cortex as observed in extracellular recordings in vivo, it is clear that one needs a spiking neuron model. For people working at the network level, this means specifying at least three things: the 'architecture' of the network (number of 'units', a specification of who communicates with who, and finally specification of the inputs); a model for the 'units' themselves (these 'units' can be groups or populations of functionally equivalent neurons, or single neurons); and a model for the connections between the units (synapses in the case units are single neurons). In this section, we first briefly describe popular models of units and connections, and then commonly studied network architectures. Finally, we explain briefly the theoretical methods that are used to investigate such networks.

Different Classes of Neurons Models

Broadly speaking, there are three classes of models: rate models; networks of binary neurons; and networks of spiking neurons. Though networks of binary neurons are a very useful tool to investigate associative memory properties of recurrent networks, they are not well-suited to study questions involving dynamics. We therefore restrict ourselves here to discussing rate models and networks of spiking neurons.

Rate Models In so-called 'rate models' (also called neural mass models, or neural field models), one describes the activity of a population of functionally equivalent neurons by a continuous variable r that typically obeys an ordinary differential equation. In its simplest form, the activity of a single population of neurons can be described by the equation

$$\tau \frac{\mathrm{d}r}{\mathrm{d}t} = -r + \Phi(I_{\mathrm{ext}} + Jr) \,, \tag{1}$$

where: r is the mean activity (firing rate) of the population, τ is the characteristic time constant of rate dynamics, Φ is the steady state input-output transfer function (f–I curve), that is typically taken to be sigmoidal, I_{ext} are the external inputs to the population, and J is the strength of the intra-population connections. For excitatory populations, $J > 0$, while for inhibitory populations, $J < 0$.

The function Φ that describes the 'static' transfer function of neurons in the population can take different forms. Popular transfer functions are threshold-linear,

$$\Phi(I) = \begin{cases} I - T & I > T \\ 0 & I < T \end{cases} \,, \tag{2}$$

where T is a threshold, or sigmoidal, $\Phi(I) = r_{\max}/(1 + \exp(-\beta(I - T)))$, where T is again a threshold and β measures the steepness of the f–I curve at half-maximal firing rate. Note that while the threshold-linear transfer function is unbounded, the sigmoidal one saturates at a maximal rate $r_{\max}$. In the limit $\beta \to \infty$, the sigmoidal transfer function becomes binary, $\Phi = r_{\max}$ for $I > T$, $\Phi = 0$ otherwise.

Another choice is to take Φ as the analytically computed transfer function of a specific spiking neuron model, like the integrate-and-fire model. As described below, one can compute the average firing frequency as a function of both mean and variance of inputs in some simple cases, and the resulting function can be used as a transfer function in a rate model formulation [7]. Examples of f–I curves are shown schematically in Fig. 1a.

Spiking Neurons As for most cells, ionic concentration differences and selective membrane permeability make neurons intracellularly hyperpolarized with respect to the extracellular medium. The difference between the intracellular and extracellular potentials, the cell membrane potential, is thus negative, with a typical value of about $-70\,\mathrm{mV}$. Most neurons in the nervous systems of vertebrates use action potentials[1], or spikes, which are large depolarizations of short duration ($\sim 1\,\mathrm{ms}$), as communication units. It is therefore natural to use spiking neuron

[1]Note however that there are well-documented exceptions, for instance in the retina.

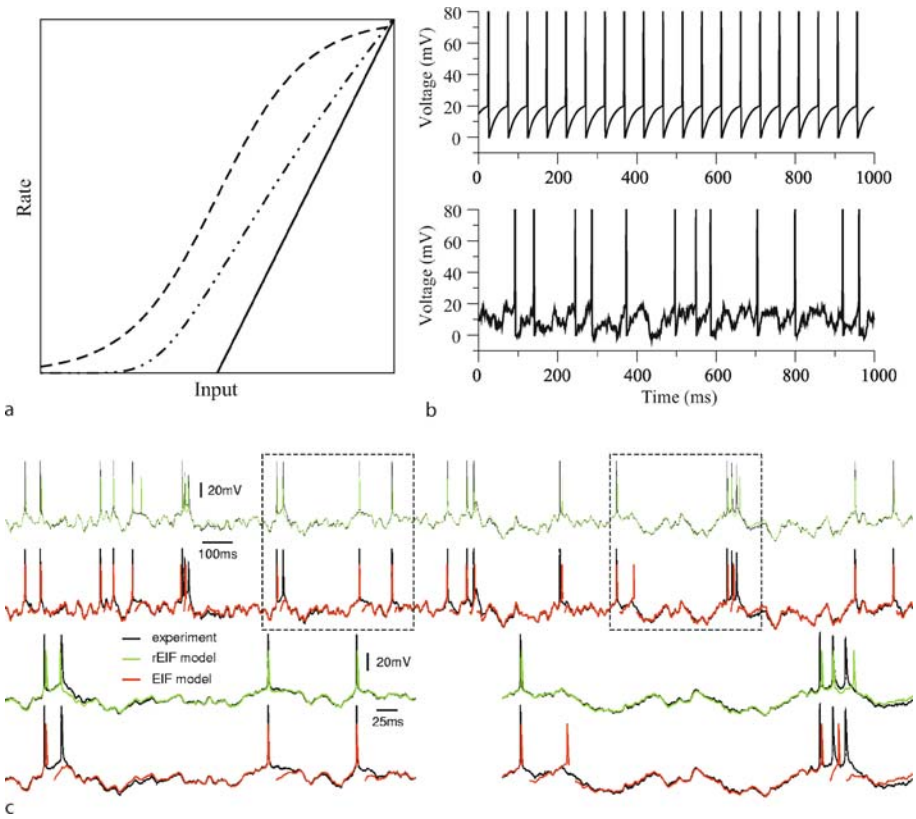

Neuronal Dynamics, Figure 1

Single unit/neuron models. a Rate model: commonly used f–I curves Φ (mean firing rate as a function of inputs). *Solid*: threshold-linear function. *Dashed*: sigmoidal function. *Dot-dashed*: f–I curve of a leaky integrate-and-fire neuron subjected to noisy inputs. b Dynamics of a leaky integrate-and-fire neuron (membrane potential) driven by constant suprathreshold input (*upper panel*) or noisy input (*lower panel*). Spikes are added by hand to better visualize the timing. c Dynamics of two versions of an exponential integrate-and-fire model (*red*: with hard reset, *green*: with adaptive threshold), together with recording of a cortical pyramidal cell. Both the artificial and the real cell receive the same injected time-dependent current (a realization of an Ornstein–Uhlenbeck process). From [8]

models in network studies. The action potential is primarily a membrane phenomenon that can be described with different degrees of realism and refinement. There is a correspondingly large diversity of single neuron spiking models in the literature, from simple, one-variable integrate-and-fire models, to complex multi-compartmental models with many voltage-gated ionic channels. The complexity of a single neuron model can be characterized along two dimensions: (i) the number of variables describing the membrane dynamics at a given location; and (ii) the number of compartments describing the spatial geometry of the neuron.

Along the first dimension, the simplest model is the integrate-and-fire model. It retains as its only variable the membrane potential [27,28,68,74,112] and simply describes the action potential by a threshold V_T in membrane potential. In the subthreshold range $V < V_T$, the membrane dynamic is simply described as that of a capacitance C (of about 1 μF/cm^2, coming from the insulating property of the lipid bilayer) in parallel with a passive leak conductance g,

$$\tau_m \frac{dV}{dt} = -(V - V_L) + I_{syn}(t) , \quad (3)$$

where $\tau_m = C/g$ is the membrane time constant (typical values are 10–20 ms), V_L is the resting potential, and $I_{syn}(t)$ are the synaptic inputs, in which we have absorbed the input resistance of the neuron (typical values are 10–100 MΩ). Thus, in Eq. (3), the inputs are in mV. In this model, a spike is emitted when the voltage hits the threshold V_T. At such times, the voltage is immediately reset to a sub-threshold voltage V_R. It is also possible to add to the model an absolute refractory period, during which the voltage is clamped to the reset, to get saturation of the

firing frequency. To choose the values of the voltage parameters, there are basically two options: take the 'realistic' values (similar to observed values in real neurons), $V_L \sim -70$ mV, $V_T \sim -50$ mV, and V_R somewhere in between; or define the origin of the voltage to be $V_L = 0$, and redefine V_T and V_R accordingly. The dynamic of the integrate-and-fire model to various types of inputs is shown in Fig. 1b.

At the opposite extreme, 'biophysically realistic' models aim to accurately represent the dynamics of the ionic channels that modify membrane permeability. They often employ the celebrated Hodgkin–Huxley formalism [55] which describes the ionic channel contribution as changes in membrane conductance via the dynamics of the activation and inactivation of various voltage-gated ionic currents. While the integrate-and-fire model, due to its very simplicity, is a model of choice for analytical studies and large-scale simulations, it lacks many features exhibited by real neurons. These features can often be accurately captured in the Hodgkin–Huxley formalism but at the price of difficulties in mathematical analysis (due to a large number of variables and pronounced non-linearities) and heavy computational cost. An intermediate class of models try to capture the best of both worlds: they contain the minimal number of variables (typically 1 or 2) and the minimal non-linearity required to capture particular phenomena of interest [21]. For example, generalized integrate-and-fire neurons contain an additional variable coupled to the voltage, which can give rise to sub-threshold resonance, as seen in many cell types [61,104]; non-linear integrate-and-fire models can accurately capture spike generation with a minimal (exponential) non-linearity in the voltage equation [8,42], see Fig. 1c; firing rate adaptation can be captured by a single adaptation variable coupled to the voltage; and so on.

Along the second (spatial) dimension, a similar diversity exists. The simplest models are composed of a single compartment. At the opposite, detailed models of cells with highly branched dendritic trees, such as the cerebellar Purkinje cell, can contain more than a thousand compartments [36]. Again, intermediate models, such as two-compartmental models, try to capture the best of both worlds: capturing non-trivial phenomena exhibited by real cells with the minimal description.

While multi-compartment models have been a valuable tool at the single neuron level, most studies of network behaviors use neurons with a single compartment. Apart from the obvious issues of mathematical tractability and computational cost, another concern limiting the use of such models is the limited present experimental knowledge on relative geometry and repartition of ionic channels. Ongoing initiatives, such as the "blue brain" project [86], may lead this to change, but we expect progress in this direction to be slow, given the huge number of parameters (especially if there are correlations in the properties of connected neurons).

Synapses

While in rate models synapses are specified by a single number, in spiking neuron models synapses are specified by dynamical variables. In the simplest case, synapses have static amplitude: their response to a particular pre-synaptic spike is the same regardless of the history of their activation. However, many, if not all, synapses in the CNS are history-dependent. This is described by models capturing short- and/or long-term plasticity properties.

Total Synaptic Inputs A typical neuron in the CNS receives inputs from something on the order of 10,000 presynaptic neurons[2]. The total synaptic inputs $I_{\text{syn}}(t)$ of Eq. (3) are typically taken as a linear sum of both individual synaptic inputs and individual spikes. This leads to a synaptic current to neuron i of the type

$$I_i(t) = \sum_j J_{ij} \sum_k s\left(t - t_j^k\right) ,$$

where the sum over j runs over pre-synaptic neurons, J_{ij} defines the amplitude of a unitary post-synaptic potential (PSP) elicited by a spike of neuron j in neuron i ($J_{ij} > 0$ if neuron j is excitatory, while $J_{ij} < 0$ if neuron j is inhibitory), the sum over k is a sum over spikes of pre-synaptic neuron j, S describes the temporal dynamics of a unitary post-synaptic current (PSC see below), and t_j^k is the timing of the kth spike emitted by neuron j.

The simplest models use the so-called "current-based" description of synaptic inputs, in which the current is independent of the voltage. A more realistic description is to take a voltage dependent $J \propto g(V - V_{\text{rev}})$ where g is the time-dependent synaptic conductance, and V_{rev} is the corresponding reversal potential – the "conductance-based" description. Typical values for PSP amplitudes are in the range 0.1–1 mV; $V_{\text{rev}} \sim 0$ mV for excitatory synapses, while $V_{\text{rev}} \sim -80$ mV for inhibitory synapses. Typical values for synaptic conductances are in the nS range.

A Simplified Semi-realistic Description for Single PSCs The time course of post-synaptic currents elicited by a sin-

[2]This is an estimate for pyramidal cells in neocortex and hippocampus. Different cell types can have widely different average number of inputs, from the 4 average inputs of a granule cell in cerebellum, to the more than 100,000 inputs of a Purkinje cell.

gle spike can be described quite accurately by a delayed difference of exponentials. When a spike is emitted at time $t = 0$, the current is zero until $t = D$, where D is the delay, or latency, of the synaptic connection. Then the current is described by

$$s(t) \propto \exp\left(-\frac{(t-D)}{\tau_d}\right) - \exp\left(-\frac{(t-D)}{\tau_r}\right), \qquad (4)$$

where $\tau_r < \tau_d$ describe the rise and decay times of the current, respectively. This can be equivalently described by the system of differential equations

$$\tau_d \frac{\mathrm{d}s}{\mathrm{d}t} = -s + x$$
$$\tau_r \frac{\mathrm{d}x}{\mathrm{d}t} = -x + \delta(t-D)\,.$$

Typical values for the time constants are: $D \sim 1\,\mathrm{ms}$; τ_r and τ_d vary widely depending on receptor type, but are $\tau_r < 1\,\mathrm{ms}$, $\tau_d \sim 2$–$10\,\mathrm{ms}$ for the fast (AMPA and GABA_A) receptor-mediated synaptic responses. See Fig. 2 for synaptic currents described by Eq. (4), as well as a few experimentally recorded synaptic currents.

Dynamic Synapses Virtually all synapses in the CNS undergo plasticity at various time scales. Short-term plasticity (on timescales of 100s of ms) is ubiquitous; Short-term depression and facilitation has been characterized experimentally and models exist that reproduce fairly accurately the characteristics of this plasticity (see [2,117]).

On longer timescales (>1 hour) various types of synapses (for example, pyramidal-to-pyramidal synapses in the neocortex; CA3 to CA1 synapses in the hippocampus; granule cell to Purkinje cell synapses in the cerebellum) exhibit various types of long-term plastic changes. There is a vast amount of material documenting such plasticity in the literature (for reviews see [16,82,83,119]), and a large number of investigators have proposed models that account for rate-based and/or spike-timing based plasticity.

In this review, we will mostly consider networks with fixed synapses; the effects of both short-term and long-term plasticity will be mentioned only briefly.

Network Architecture

The network architecture specifies how the units, or neurons, are coupled together.

Architectures in Rate Models Rate models are often used to investigate the dynamics of spatially extended networks, or systems which encode continuous stimuli. A prototypical example is given by models of the primary visual cortex, in which the activity $r(x)$ at a given location x evolves according to

$$\tau \frac{\mathrm{d}r(x)}{\mathrm{d}t} = -r(x) + \Phi\left(I_{\mathrm{ext}}(x,t) + \int \mathrm{d}y J(x,y) r(y)\right). \qquad (5)$$

The network 'architecture' is characterized by the 'synaptic footprint' $J(x, y)$ that specifies the strength of connections from y to x (the recurrent part); and the inputs $I_{\mathrm{ext}}(x, t)$, describing inputs coming from outside the modeled net-

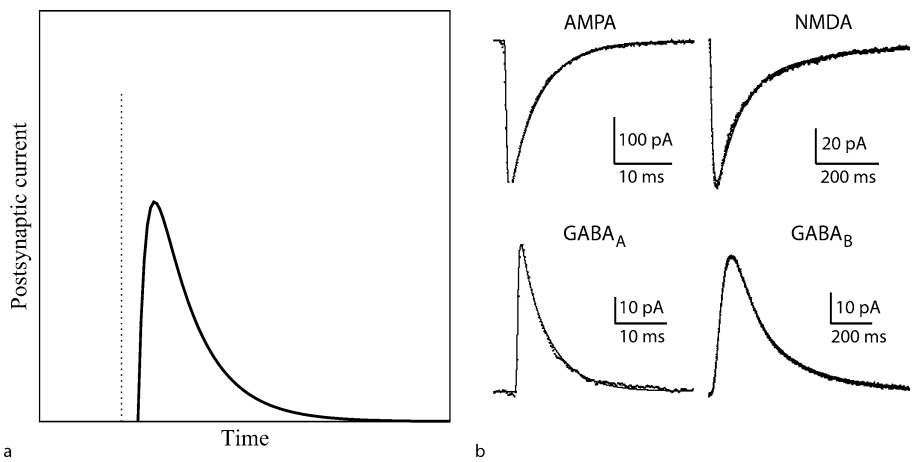

Neuronal Dynamics, Figure 2
Synaptic models. a Synaptic current elicited by a single spike of Eq. (4). Timing of the presynaptic spike is indicated schematically by a *dotted line*. b Synaptic currents mediated by various receptors, recorded in various slice preparations. From [38]

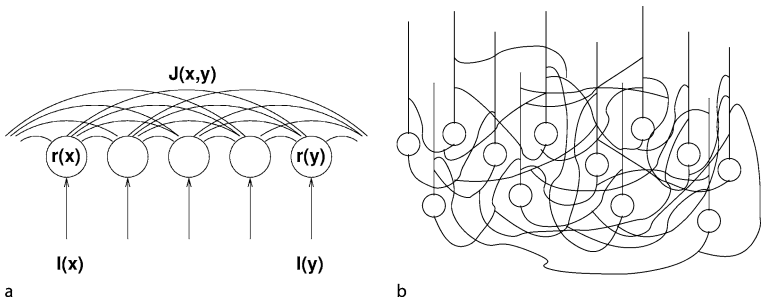

Neuronal Dynamics, Figure 3
Network architectures. a Sketch of a rate model. Each circle represents a population of cells at a given location. Lines connecting circles represent synaptic connections between populations. b A randomly connected network of individual neurons. The drawing schematically represents cell bodies (*circles*), dendrites (*vertical line*), and axons contacting dendrites of post-synaptic neurons (*curved lines*)

work (LGN and other cortical areas in the case of primary visual cortex). See Fig. 3a.

Networks of Spiking Neurons The architecture of a network of spiking neurons can be specified by answering the following list of questions:

How many classes of neurons? Early studies of networks of spiking neurons focused on one population networks (either excitatory or inhibitory networks). Networks in which there are two populations of neurons—one excitatory, and one inhibitory—have also been analyzed in detail.

How many neurons in each class? Analytical studies often use the limit in which the number of neurons goes to infinity (see theoretical tools section below). This is justified by the fact that local networks in the brain are typically composed of a very large number of neurons ($\sim 10^5$). It is also becoming possible to simulate networks of such large numbers of neurons.

What is the wiring diagram? Early studies focused on fully connected networks, which are easier to handle analytically. Another popular architecture has been a randomly connected network, with a given connection probability (see Fig. 3b). Such an architecture takes into account the relatively low connection probability of nearby neurons. Relatively few studies of networks of spiking neurons take into account spatial structure, such as a monotonically decaying connection probability.

What is the nature of the individual couplings between neurons? Theorists have investigated networks connected by chemical synapses, electrical synapses, or both. More recently, the effect of various types of synaptic plasticity phenomena (both short-term and long-term) has begun to be studied.

What are the inputs to the neurons in the network? Popular choices to study the intrinsic dynamics of networks are constant uniform inputs, or noisy inputs, which are uncorrelated from neuron to neuron. To understand the information processing capabilities of networks, one has to choose a model for the type of inputs that are processed by such network. This choice typically depends on the area that is modeled. For example, inputs to a model of a hypercolumn in visual cortex are often taken to represent (fixed or dynamic) oriented bars, and are modeled appropriately.

Theoretical Tools

Different types of analytical tools are available depending on the nature of the model. Rate models are typically specified by systems of coupled ordinary differential equations. Hence, standard tools can be borrowed from dynamical system theory. One typically starts by investigating the case of constant inputs, for which it is sometimes possible to find fixed points, and the linear stability of these fixed points. Sometimes, it is possible to analyze more complex behaviors, oscillatory states, spatially localized states, waves, etc., and their stability.

Networks of spiking neurons are substantially harder to analyze. Several types of approximations can be used, depending on the strength of coupling and the level of noise: fully connected networks in the weak coupling/weak noise scenario can be studied in the framework of the theory of coupled oscillators. When coupling and noise are strong, it is sometimes possible to approximate synaptic inputs to a neuron as a random Gaussian process, the moments of which depends on both connectivity and activity in the network. For instance, for leaky integrate-and-fire neurons, this leads to the study of the Langevin-like equa-

tion

$$\tau_m \frac{\mathrm{d}V}{\mathrm{d}t} = -(V - V_\mathrm{L}) + \bar{I}_\mathrm{syn}(t) + \sigma\sqrt{\tau_m}\xi(t)\,, \qquad (6)$$

where $\bar{I}_\mathrm{syn}$ represents the mean synaptic input and $\xi(t)$ a white noise term representing its fluctuations. In conditions in which correlations between these inputs can be neglected (as when the connection probability between cells is weak), a network of integrate-and-fire neurons can then be characterized by a Fokker–Planck equation that describes the time evolution of the instantaneous distribution of membrane potentials [1,22,105]

$$\tau_m \frac{\partial P}{\partial t} = \frac{\partial}{\partial V}\big[\big(V - \bar{I}_\mathrm{syn}(t)\big)P\big] + \frac{\sigma^2}{2}\frac{\partial^2 P}{\partial V^2}\,, \qquad (7)$$

where the mean $\bar{I}_\mathrm{syn}(t)$ is determined self-consistently from the neuron discharges.

In addition to the above mentioned analytical tools, numerical simulations are almost always indispensable, if only to check the validity of the approximations. While simulating rate models is relatively straightforward, simulations of networks of spiking neurons are, again, more demanding. In addition to standard finite-difference integration schemes [53,57,101], it is sometimes possible to use event-driven schemes, in which one jumps from one spike to the next, using an exact integration of network dynamics during inter-spike intervals (see [88]).

Intrinsic Network Dynamics

We start by a description of intrinsic network dynamics, that is, dynamics in absence of spatially or temporally structured inputs. In rate models, this means taking a constant and uniform input $I_\mathrm{ext}(x,t) = I_\mathrm{ext}$; in networks of spiking neurons, this means taking either a constant deterministic input, or a noisy stationary input. Noise is typically chosen to be a Gaussian random noise, uncorrelated from neuron to neuron, or uncorrelated Poisson processes independently activating all the neurons of the network.

The goal is then to understand the types of dynamics that can be generated by the network, independently of its inputs. Experimentally, intrinsic dynamics of a network can be observed in vitro, or in vivo, either in anesthetized animals or in awake animals in absence of specific inputs that activate the particular area under observation. Such experiments reveal many different types of activity that modelers have been trying to understand. A prominent and ubiquitous finding in in vivo recordings is irregular background activity at low rates. The issue of the origin and mechanisms of this activity can be addressed using networks of spiking neurons, and will be the topic of Subsect. "Irregular Firing at Low Rates". We will then move to the issue of multistability in networks, which can be considered in the simple setting of rate models. Finally, we will discuss the issue of synchronization and oscillations, which can be addressed both with rate models and spiking neuron models.

Irregular Firing at Low Rates

Recordings of cortical neurons in vivo show very irregular activity. The distributions of intervals between successive action potential emissions, commonly called inter spike intervals (ISI), are approximately Poissonian in a variety of cases. Understanding this irregularity, is a first theoretical challenge [15,111]. Cortical cells receive on the order of 10^4 synapses. This large number should tend to promote regular firing since fluctuations of the total input should be small compared to its mean, unless inputs are strongly correlated. Another related difficulty is that measurements of synaptic weights [40,56,87,110] (see [10] for a short survey) indicate that the detectable connections between pyramidal cells are relatively strong with a mean value of 0.5–1 mV, as measured from the peak somatic depolarization that they provoke. As a consequence, thousands of presynaptic pyramidal cells spiking at a low rate of a few Hertz should provide a strong depolarization of the post synaptic cells. It is thus not clear how a low rate of spike emission can be maintained in a recurrently coupled network. One loophole in the argument is that it does not take inhibition into account. Inhibition can potentially solve both puzzles at once when inhibitory inputs balance on average excitatory ones [6,108,120,121]. This drastically reduces the input current so that low firing rates are possible. Moreover, current fluctuations can be comparable to the mean current, allowing for irregular firing.

These studies have shown that irregular firing is obtained in randomly connected networks of spiking neurons, as soon as recurrent inhibition is strong enough. Remarkably, such an irregular firing can be obtained even when the inputs to the network are deterministic, that is, in the absence of noise [120]. The nature of this irregular firing remains the subject of intense study. Irregular firing can be associated with complex but stable trajectories in phase space [63,127], or with chaotic dynamics [127].

Recent data provide some evidence for a balance between inhibition and excitation in cortical networks [109]. Somewhat unexpectedly, a balance of excitation and inhibition has also been found in the rhythmic input onto motor neurons during scratching behavior in turtles [13]. It is proposed that the associated enhanced irregularity can be used to produce a smoother muscle excitation.

Attractors with Inhomogeneously Distributed Activity

A recurring idea in theoretical studies is that structured connectivity can lead the dynamics of a given neural network to settle in one of several coexisting attractors with mean activity differently distributed among the different neurons of the networks. It is, for instance, central to Hopfield's view of memory storage in neural networks [58], as well as to proposed explanations for persistent activity underlying short-term memory.

Many designs of this type are based on short-range recurrent excitation together with long-range inhibition. The simplest embodies the so-called "winner-take-all" mechanism. It consists of two-recurrently coupled excitatory sub-networks inhibiting each other. This can described in a rate model framework as

$$\tau \frac{\mathrm{d}r_1}{\mathrm{d}t} = -r_1 + \Phi(I_0 - Jr_2) \tag{8}$$

$$\tau \frac{\mathrm{d}r_2}{\mathrm{d}t} = -r_2 + \Phi(I_0 - Jr_1)\,, \tag{9}$$

where r_1 and r_2 describe the activity of each sub-network and $J > 0$ the magnitude of recurrent inhibition. The homogeneous steady state has $r_1 = r_2 = r_s$ with

$$r_s = \Phi(I_0 - Jr_s)\,. \tag{10}$$

However, linearization of Eqs. (8), (9) shows that homogeneously distributed activity is unstable when inhibition is sufficiently strong (the bifurcation itself can be supercritical or subcriticalm depending on the sign of the third derivative of Φ). In this case, the symmetry between the two neuronal populations is broken and there are two possible coexisting attractors with the activity enhanced in either sub-network enhanced and suppressed in the other one.

This scheme can be simply extended to several sub-networks (with global rather than reciprocal inhibition) to enlarge the possible number of coexisting attractors. Some neuronal networks, such as, for instance, the head-direction cell network, are thought to possess a line of attractors instead of a discrete set of attractors. This can again be obtained by generalizing Eqs. (8), (9) to cells labeled by a continuous parameter, which we denote by θ, with excitatory connections between cells with close values of θ and inhibition between cells with more distant values [5,39,48]. For illustrative purposes, we describe one such simple scheme proposed in [12], where for definiteness the continuous parameter is taken to correspond to an angle. This so-called "ring model" is a specific case of Eq. (5)

$$\tau \frac{\mathrm{d}}{\mathrm{d}t} r(\theta, t) = -r(\theta, t) + \Phi\left[I_{\mathrm{ext}} + \int_{-\pi}^{\pi} \frac{\mathrm{d}\theta'}{2\pi} J(\theta - \theta') r(\theta', t)\right], \quad -\pi \le \theta \le \pi\,. \tag{11}$$

It considerably simplifies the analysis [12] to restrict the synaptic coupling function to its first harmonic, $J = J_0 + J_1 \cos(\theta - \theta')$ and to take Φ to be a threshold linear f–I curve, $\Phi[I] = \beta[I]_+$ that is $\Phi[I] = \beta[I]$ for $I \ge T$ and 0 otherwise. Integration over θ reduces the dynamics to coupled differential equations for the first three moments r_0, r_{11} and r_{12} of r with

$$\begin{aligned} r_0 &= \int_{-\pi}^{\pi} \frac{\mathrm{d}\theta}{2\pi} r(\theta)\,, \\ r_{11} &= \int_{-\pi}^{\pi} \frac{\mathrm{d}\theta}{2\pi} r(\theta)\cos(\theta)\,, \\ r_{12} &= \int_{-\pi}^{\pi} \frac{\mathrm{d}\theta}{2\pi} r(\theta)\sin(\theta)\,. \end{aligned} \tag{12}$$

The homogeneous activity ($r(\theta) = r_0$) is non-zero for $I_{\mathrm{ext}} > 0$, $r_0 = I_{\mathrm{ext}}/(1 - \beta J_0)$; $\beta J_0 < 1$ is needed for the dynamics to be well-behaved that is, all-to all interactions should be inhibitory, or at least not too excitatory, to prevent activity growth without bounds (or saturated activity when saturation is included in Φ). However, homogeneous activity is unstable when interaction between cells with close values of θ is large enough that is, here for $\beta J_1 > 2$. Then, the network attractor is formed by a line of solutions in the form

$$r(\theta) = A[\cos(\theta - \theta_0) - \cos(\theta_c)]_+\,, \quad -\pi \le \theta_0 \le \pi\,. \tag{13}$$

Thus, the network activity is confined to cells with parameter θ in a interval of values of width $2\theta_c$ around an arbitrary angle θ_0. The constant θ_c, the solution selectivity, depends only on the magnitude of βJ_1. Localized states of activity and their domain of existence in parameter space are shown in Fig. 4.

The ring model is a nice example of a model with a continuous line of attractors but it also exemplifies difficulties inherent to this type of models. The main problem is that a line of attractors is a fragile structure. It is not resistant to perturbations coming, for instance, from heterogeneity in the connection strengths or neurons' properties. If this design is used in real neural systems, there should exist additional mechanisms that maintain the line of attractors and avoid its breaking to a discrete set. Some proposals exist [70,102], but the existence of a line attractor in an actual neural network remains to be demonstrated.

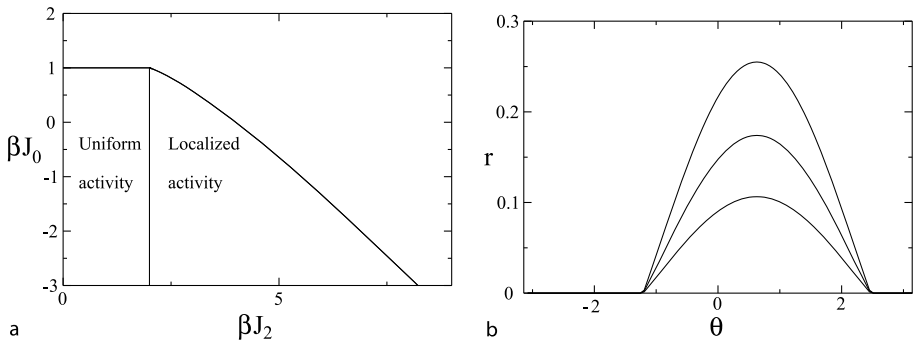

Neuronal Dynamics, Figure 4
Ring model. a Phase diagram showing the position of the uniform and localized activity for the model of Eq. (11). b Profiles of localized activity for $I_{ext} = .1, J_2 = 3.$ and $J_0 = -2, -1, . - 5$ (from bottom to top). Adapted from [12]

Dynamics at Various Time Scales

As has already been stressed, spontaneous activity in a variety of neural systems exhibits rhythmicity at time scales that range from a few millisecond to many seconds. We restrict ourselves to rhythms which are intrinsically of electrophysiological origins (unlike for instance, circadian rhythms which are accompanied by changes of neuronal excitability but which are thought to be primarily driven by changes in gene expression).

Bursts of Activity At the lower end of the spectrum, networks which produce repeated bursts of activities at a time scale of a few tenths of seconds to a few seconds are observed in different cases. They seem to be an essential component of pacemakers generating locomotion [29] or respiration [69]. Repeated bursts of activity are also observed in neuron cultures [81,93,96]. An important additional example is given by the observation of up-down states in cortical slices [107], which are proposed to be analogous to slow oscillations during slow-wave sleep.

Specific models have been proposed for different systems but they appear to be based on the same general ideas. First, recurrent excitation allows for the self-sustaining bursting state. Second, this active state terminates when a slower building process has reached sufficient magnitude to prevent self-sustainment of the active state. The main contenders for the slow process are synaptic depression (see Subsect. "Dynamic Synapses") or some intracellular mechanism such as a slow increase of an ionic concentration that opens an ion-gated channel once it has reached a threshold level. The generation of bursts of activity by synaptic depression has been modeled at a general theoretical level in [118]. In the context of respiratory rhythm generation, burst termination has been modeled as due to the slow inactivation of a Na^+ current (that is, termination of a depolarizing current) or by the activation of a calcium-gated potassium (that is, hyperpolarizing) current. For up-down states, a sodium-dependent potassium conductance has been hypothesized to mediate their termination and has been taken into account in a proposed model [32]. The mechanism is illustrated in Fig. 5. Both calcium and sodium-gated potassium conductances have been identified in spinal neurons [123]. It is worth noting however that in most cases the actual mechanism at work has not been experimentally pinpointed.

Besides burst termination, the cause of a burst should be determined to see how a repetition at about 1 Hz can

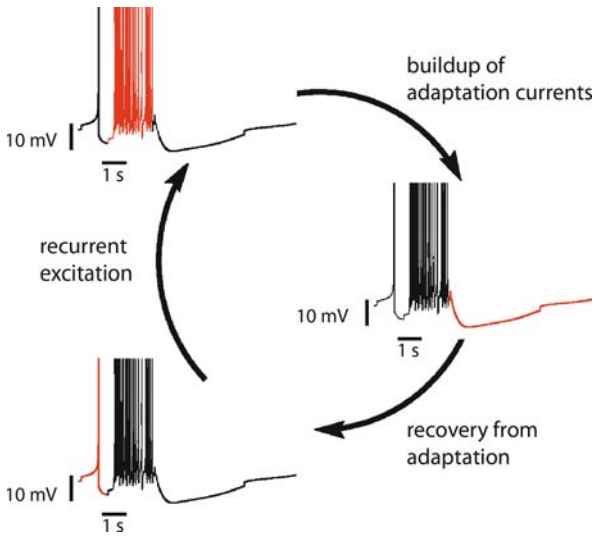

Neuronal Dynamics, Figure 5
Mechanism of slow oscillations in cortical slices. Recurrent excitation provokes a burst of activity which is terminated by a slow adaptation current. Activity resumes when the adaptation current has decayed. From [32]

come about. The general mechanisms appear again to be of two different types. Slow recovery from the termination process (for instance, return of concentration to a low level by the action of an ionic pump) can deterministically bring the networks above a threshold for excitation. Alternately, recovery can lead the network in a low activity state from which it can stochastically transit to the active state [126].

It is worth noting here that both the size distribution and duration distribution of bursts of activity in cortical slices have been examined using multi-electrode arrays [11]. Power law distributions corresponding to critical branching processes, reminiscent of a self-organized critical phenomenon [9], have been found. This was shown to emerge naturally from synaptic depression in a simplified IF neuron model without leak [77]. The applicability of this result to more physiologically realistic networks remains to be examined.

Wandering Between Different Attractors Bursts of activity can be seen as a succession of jumps between two attractors with homogeneously distributed activity—a high activity state and a low activity state—mediated by a slow process. With structured connectivity, similar mechanisms can mediate transitions between a larger number of more complex coexisting attractors, as those described in Subsect. "Attractors with Inhomogeneously Distributed Activity". A given active state then consists of one particular subnetwork of active neurons among several possible ones. Some evidence for such stochastic visits of coexisting attractors has been provided both in vitro and in vivo. Using calcium imaging, mouse neocortical slices have been found to display sparse spontaneous activity and sets of coactive neurons that repeatedly appear above chance level [33]. In vivo, neurons of cat visual cortex responding to the same preferred orientation (the so-called orientation maps) have also been reported to be spontaneously coactive, with sets belonging to different orientations appearing in time in a stochastic manner [67,115]. A model with biologically plausible connectivity and attractors consisting of orientation maps is able to reproduce these remarkable data [17]. It should be noted, however, that the experimental results seem also to be interpretable as fluctuations about a single background state driven by correlated thalamic inputs [49].

Interestingly, a similar set of ideas has been used to model slow oscillations in higher cognitive tasks. One interesting example is binocular rivalry: when the right and left eyes are presented with two different images, the perceived pattern alternates between the two images every few seconds. Moreover, several functional magnetic resonance imaging studies have shown that fluctuation in neural activity in several brain areas correlates with perception changes [80,100,114]. The dominance of the perception of one image over the other one has been modeled as a competition between two recurrently coupled excitatory networks reciprocally inhibiting each other, the "winner-take-all" design described in Subsect. "Attractors with Inhomogeneously Distributed Activity". Without further elaboration this gives rise to two attractors: either one of the two networks can be active with the other silenced. Similarly to burst or upstate termination, alternation between attractors (and percepts) can arise by supplementing this basic design with a slow process. It has been modeled either as coming in a deterministic way from synaptic depression [124] or from a stochastic jump from one of the two attractors to the other [92].

Synchronization of Periodically Firing Neurons Neural recordings shows oscillations in diverse frequency bands. They depend both on the neural structure and, when recorded in vivo, on the type of activity that the animal is performing. These rhythms, which are seen in EEG or local field potential, emerge from the coordinated activity of many neurons. Synchronization of nonlinear oscillators [99] has of course been thoroughly studied since its discovery by Huyghens. Some of the tools developed [73] have been directly applied to neurons in a regime where they emit action potentials in a periodic manner [52,122]. Many experimental and theoretical studies have focused on single neurons and coupled pairs of neurons to infer synchronization properties of large networks of weakly coupled neurons. This has acquired renewed importance with the discovery that inhibitory interneurons in the same subtype category appear to be preferentially coupled [45,47] For weak coupling, the dynamics of an oscillator can be described by the behavior of its phase, that is, its position along its limit cycle. How the spike advance/delay produced by a short current injection depends on the time of injection during the interspike interval is a crucial quantity for synchronization. This so-called phase-response-curve (PRC) has been studied in different models [52,122] with the result that generally a small and short depolarization advances the next spike, and does so to a greater extent the more closely it occurs before spike emission. This is referred to as a type I PRC. In some cases, though, (as in the original Hodgkin–Huxley model) a less intuitive type II PRC is found: a small depolarization just after the spike has a greater effect on hyperpolarizing currents than on depolarizing ones and results in a delay of the next spike. The PRC has also been experimentally measured for neocortical neurons. The results of the early experiment of [103] are shown in Fig. 6

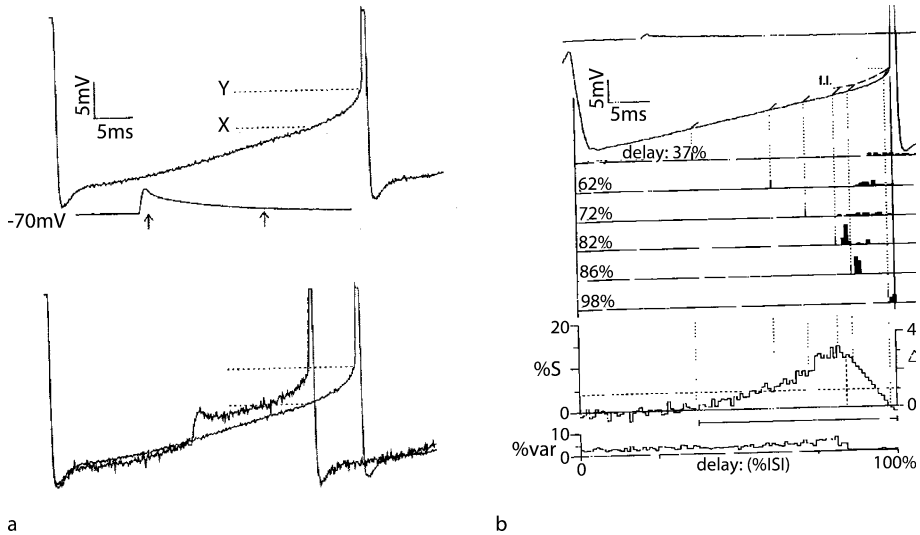

Neuronal Dynamics, Figure 6

Measurement of phase-response curve (PRC) for neocortical neurons [103]. The authors injected a short depolarizing current pulse at different times during the interspike interval. As shown on the *left panel*, this results in a timing change for the spike following the current injection. The *right panel* show results for different injection times as well as the PRC which summarizes the data by giving the spike time changes as function of the current injection time (expressed here as a percentage of the ISI). Adapted from [103]

and give a type I PRC. For instantaneous synaptic transmission, this would lead one to expect that excitatory interactions tend to make two identical neurons spike in phase while inhibitory ones would push them to antiphase spiking. It was therefore an important theoretical finding that, for more realistic synaptic currents with finite rise and decay, this is not generally the case and that inhibition can be more efficient than excitation for synchronization. For excitatory synapses, complete in-phase spiking is generally unstable for type I PRC and a finite dephasing exists even for identical neurons. On the contrary for inhibitory synapses, in-phase and anti-phase spiking are both possible at low frequency, a bistable situation. However the attraction basin of the in-phase state grows with frequency, and above a threshold frequency the anti-phase state disappears. This leads one to expect a transition from antiphase to in-phase spiking for a coupled pair of identical interneurons as their firing frequency increases. This has been verified experimentally [46] as well as other similar predictions for gap junctions or on the combined effect of inhibitory synapses and gap junctions [31,78,84].

Sparsely Synchronized Oscillations Although a substantial amount of work has been devoted to analyzing periodically firing neurons with a minimal amount of noise and heterogeneity, this seems to be a rather infrequent situation in the brain. Strong heterogeneity appears to be the norm rather than the exception. Moreover, as explained in Subsect. "Irregular Firing at Low Rates," the activity of most neurons appears quite irregular and their discharge rate is often low compared to the frequency of gamma oscillations and faster rhythms. This is also the case in vitro when pharmacologically-induced activity in hippocampal slices results in spontaneous gamma oscillations [41]. Modeling studies have shown that a mechanism distinct from the Huyghens-type synchronization of oscillators can produce a fast and robust rhythm at the network level with neurons discharging irregularly and at a lower frequency [23]. An example from a simulation of a network of IF neurons is shown in Fig. 7. Recurrent inhibition again plays an important role in these sparsely synchronized oscillations. The basic idea can be understood from a rate model description similar to Eq. (1)

$$\tau \frac{\mathrm{d}r}{\mathrm{d}t} = -r + \Phi[I_{\mathrm{ext}} + Jr(t-D)] \,, \tag{14}$$

where the activity $r(t)$ of the network at time t is influenced by its activity at a previous time $t - D$. The delay D models in a crude but effective way the latency and finite kinetics of synaptic currents (see Eq. (4)). It is not difficult to show that for $J < J_c < 0$, that is, for a sufficiently strong recurrent inhibition, the activity $r(t)$ is attracted to a limit cycle and oscillates periodically. The oscillation frequency is of order $1/D$ and increases from $f = 1/(4D)$ for delays short compared to τ to $f = 1/(2D)$ for delays long compared to τ. It is simple to see the origin of the oscillations: an increase of activity in the network at time t provokes an increase of recurrent inhibition. This results in

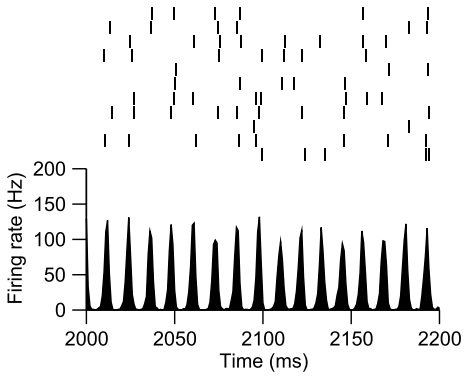

Neuronal Dynamics, Figure 7
Oscillations with sparsely firing neurons in a fully connected network of 1000 leaky integrate-and-fire neurons receiving independent white noise sources. The *top panel* shows a raster of 10 neurons, while the *bottom panel* shows the network instantaneous firing rate (computed in 1 ms bins). The network oscillates at about 90 Hz, while single cells fire at about 30 Hz, as predicted by theory [22,25] (synaptic time constants: 1 ms latency, 1 ms rise time, 6 ms decay time). Reproduced from [21]

a decrease in network activity at about $t + D$, since transmission from one neuron to the next takes a time D. This decrease of activity at $t + D$ itself again generates an increase of activity at about $t + 2D$ from which the cycle can continue. One sees that the frequency of the network oscillation, of order $1/D$, is directly linked to the kinetics of synaptic transmission but not related to the neuron discharge rate r. While the rate description of Eq. (14) simply captures some characteristics of network oscillations in the sparsely synchronized regimes, it is far from providing accurate estimates, in particular for the oscillation threshold. A more complex mathematical description based on the Fokker–Planck equation has been developed to this end [22,25]. It allows, moreover, the inclusion of various features of single neuron dynamics such as the finite rise time of the action potential [42] or of an eventual resonance in subthreshold dynamics [26,104] (see [23] for a short review).

Although many experimental recordings are suggestive of this kind of oscillations in the brain, it is not easy to eliminate the possibility that the observed rhythm is created by a population of fast-spiking unidentified cells. At present, two of the best studied examples are provided by fast oscillations in the hippocampus and the cerebellum. The "ultrafast" ripple oscillations (140–200 Hz) are the fastest among the numerous hippocampal rhythms. They occur together with "sharp waves" during awake immobility and slow wave sleep in rats. Recordings show that both the discharge rates of pyramidal cells and interneurons (average rates 8 Hz and 30 Hz respectively [34]) are much lower than the frequency of the population oscillation. In the cerebellum, fast oscillations were discovered by Adrian in one of the first intracranial recordings [3]. More recently, de Solages et al. [37] have provided strong evidence that these fast (150–250 Hz) oscillations are emerging from recurrent inhibition among Purkinje cells that themselves fire at an average of about 40 Hz.

Wave Propagation The propagation of waves of neural activity is certainly an important topic but one that has traditionally be difficult to study experimentally. It has therefore been the subject of relatively few specific theoretical works. Wave propagation in neural networks has been classically analyzed in the framework of rate models [5,39] with results for wave existence similar to those for general excitable media (for which many studies exist motivated, for instance, by applications to chemical waves in the Belousov–Zhabotinsky reaction or, in a biological context, to electrical waves in cardiac tissue). Advances in imaging techniques are now providing more data on neural activity propagation, in slices [18,72,107], in cell cultures [62] and even in vivo [98]. In parallel, theoretical interest in spiking models is developing and results have been obtained either by the analysis of simplified cases [50] or by simulations [32]. The results highlight in particular the role played by inhibition and the potential importance of long-range connections in wave propagation.

Stimulus Driven Dynamics

What happens when external stimuli are presented to a network of neurons? An external stimulus is usually modeled as an increase (or decrease) of external inputs to specific groups of neurons. For example, in the ring model, presentation of an external stimulus representing an oriented bar during an interval $[0, T]$ is described by $I_{\text{ext}}(\theta, t) = I_0 + I_1 \cos(\theta - \theta_s)\Theta(t)\Theta(T - t)$, where θ_s is the angle of the presented bar. In a spiking neuron model, an external stimulus can be represented by a transient change of the mean currents impinging on a subset of neurons, or by a transient change in the frequency of incoming spike trains. Conceptually, these transient inputs can lead to three types of phenomena: (i) During the transient input, the nature of the dynamics of the network can change qualitatively. For example, the network might switch from an asynchronous to a synchronous state or it might switch from a uniform to a spatially localized state. (ii) In multistable networks, the stimulus can switch the network from one state to another, and the network will then stay in this particular state, thereby maintaining a short-term mem-

ory of the stimulus that was presented. (iii) Presentation of a stimulus can induce synaptic plasticity between neurons that are activated/inactivated by the stimulus, leading to a remodeling of the network, and therefore potentially to a change in the repertoire of states that the network dynamics can sustain. We now briefly examine these three phenomena before focusing on specific systems.

Dynamics During Stimulus Presentation

Selective Amplification The simplest change induced by external stimuli is a change in firing rates of the neurons. This change in rate is governed by external inputs, but also by recurrent connectivity. This can be best studied in the context of rate models (see [35]). In general, excitatory networks tend to amplify external inputs, while inhibitory networks tend to attenuate them. With spatially structured connectivity, the network will tend to amplify certain inputs, at the same time attenuating another. A typical example is represented by the ring model of Subsect. "Attractors with Inhomogeneously Distributed Activity" with a 'Mexican-hat' connectivity – with this type of connectivity (excitatory at short-range, inhibitory at long-range), the network tends to amplify spatially localized inputs, while it will attenuate inputs with no spatial selectivity. Thus, the network has the property of selective amplification. When the excitatory connectivity is strong enough and one enters the marginal phase, the network generates spatially localized states even in absence of external inputs. In this case, the external stimulus selects one of the possible network states (a state in which the activity is peaked around the most strongly activated neuron) out of the repertoire of possible states (a continuum in the case of the ring model). This property is consistent with several experimental findings in various areas [67,79].

Switching from Asynchronous to Synchronous State Another effect of external stimuli can be to change the degree of synchrony in the network. For example, a stimulus can switch a network from an asynchronous state to a synchronous state (or vice versa). This is typically what happens in randomly connected networks of excitatory and inhibitory neurons (or purely inhibitory networks) when inhibition is stronger than excitation [20,22]. This is due to the fact that external inputs tend to increase the average firing rate of the network, leading to an effective increase in the strength of inhibitory feedback. This increase can potentially make the network switch from the asynchronous to the synchronous region.

This emergence of synchronous activity induced by external inputs is reminiscent of many experimental findings, from oscillations induced by odors in the olfactory system (to be discussed in more detail below), to oscillations correlated with selective visual attention in the visual system of the primate [44]. The emergence of oscillations in the presence of visual inputs has also been hypothesized to allow the system to solve the so-called 'binding problem,' though this issue remains hotly debated [59,71,106].

Multistability and Working Memory

In multistable networks, stimuli can potentially switch the network from one state to another. For example, in winner-take-all rate models, a stimulus will switch the network to an attractor state in which the population that received the largest input is active, while all others are inactive. The fact that this network configuration is an attractor means that the network maintains, in short-term memory, some information about which stimulus it was shown. This property is the hallmark of associative memory models, where many 'memory states' in which subpopulations of neurons are selectively active at higher rates, are attractors of the network dynamics, thanks to strong excitatory feedback inside such sub-populations. It is widely believed to form the basic mechanism underlying working memory, that is, the active maintenance of information in short-term memory. Such networks have been investigated extensively in the last three decades, from networks of simplified binary neurons [58] to networks of spiking neurons of increasing realism see [6,24]). The working of such a network is illustrated in Fig. 8.

Effect of Long-Term Synaptic Plasticity on Network Dynamics

Another potential effect of external stimuli is to modify the synaptic structure of the network. For example, in associative memory models, external stimuli are assumed to lead to synaptic modifications induced by increases or decreases of firing rates of single neurons. Hebbian learning rules posit that synapses connecting two neurons which are strongly activated by the same stimulus will potentiate, while synapses connecting one strongly activated neuron to an inactivated neuron will depress. This learning dynamic tends to create a synaptic structure that leads to attractors strongly correlated with the state of the network during stimulus presentation.

Most network studies have separated neuronal and synaptic dynamics, assuming much slower synaptic dynamics. Typically, one studies the network with a fixed synaptic structure that incorporates, in a Hebbian way, learning of patterns of activity presented to the network in the past. More recently, there have been stud-

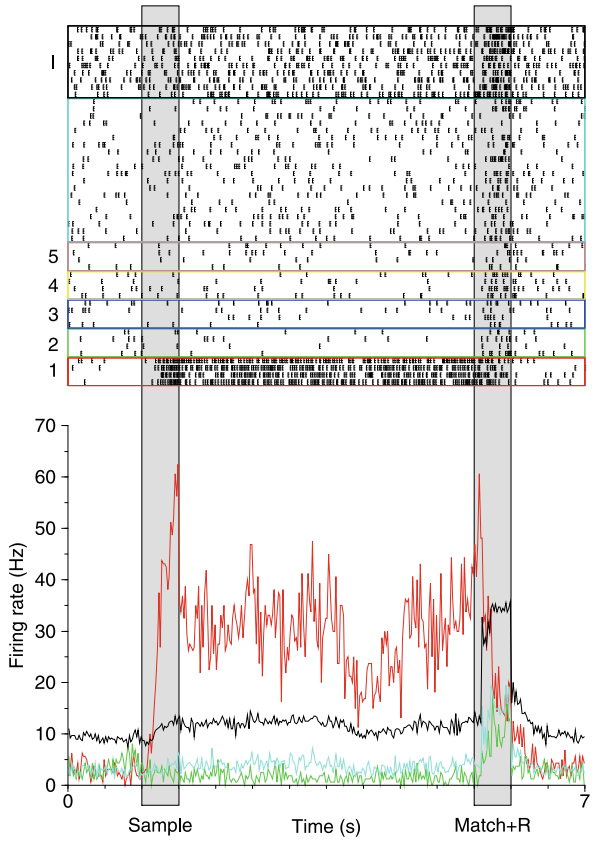

Neuronal Dynamics, Figure 8
Dynamics of a randomly connected network of excitatory and inhibitory neurons with working memory properties (*top*: raster; *bottom*: average firing rate of different populations of neurons). Inhibition is strong enough to stabilize the network in a background state (before 'sample' presentation). External stimuli have been 'stored' in the synaptic matrix, through enhanced synaptic connectivity between neurons that belong to a subpopulation that is activated by a stimulus (sub-populations 1–5 in the figure). Presentation of an external stimulus (sample 1 in the figure) increases the firing rate of the relevant neurons (see *red curve* in *bottom panel*, average firing rate of neurons in population 1). This enhanced firing rate survives removal of the stimulus because of the strong excitatory connectivity between 1 neurons. Hence, the stimulus has switched the network from the background state to the memory state corresponding to stimulus 1. Finally, a sufficiently strong non-specific stimulus switches the network back to the background state. From [24]

ies of networks with double dynamics of neurons and synapses [60,91].

Dynamics and Information Processing: Two Specific Examples

In some neural structures, the role of dynamics in information processing has been particularly scrutinized. We briefly describe here two prominent cases and some views and hypotheses that they have suggested.

Dynamics and Odor Discrimination As recalled previously, it has been known since the recordings of Adrian [4] that odors promote gamma oscillations in the olfactory bulb. The role of oscillations in odor processing has, however, remained unclear in spite of interesting proposals [43]. Olfaction is an evolutionarily ancient function and the associated neural structure are analogous in very different species. Recent recordings in insects are shedding new light on the role of dynamics in odor discrimination. The antennal lobe in insects corresponds to the olfactory bulb in mammals. It receives input from neurons in the olfactory epithelium and its principal neurons project to the next structure, the "mushroom body". In the locust, puffs of two close odors induce antennal lobe activities that are similar at start (as measured from the activity of a sample of the eight hundred projection neurons) but that diverge in a few hundred milliseconds [113]. The antennal lobe oscillations shape its output as successive volleys of projection neuron spikes emitted at a rate of about 20–30 Hz. The numerous Kenyon cells in the mushroom body then appear to function as coincidence detectors that process each projection neuron spike volley independently of the others [89,113] (reset between each volley being provided by non-specific feedforward inhibition coming from lateral horn interneurons [97]). The precise timing of the whole process appears to allow STDP plasticity at the Kenyon cells output synapses [30] and presumably memory formation in a way that remains to be related to behavioral experiments and genetic data [54,66].

Place Maps and Oscillations in the Hippocampus The hippocampus is another brain region in which oscillations are prominent and one in which their role in information processing is among the best studied. The hippocampus is an important structure for navigation. "Place" cells that fire at particular locations have been discovered thirty years ago in the rat hippocampus [94]. Different place cells discharge in different "place fields", of about 25 cm in size, and recording several cells allows the determination of the animal position in its environment. However, place cell firing is also strongly modulated at theta frequency. The phase of a place cell discharge advances as the animal enters the place cell field [95]. So, monitoring precise spike timing with respect to the theta oscillations provides supplementary information and a more accurate position determination [64,125]. Several models have been proposed for place cell formation during exploration of a new en-

vironment, some based on network dynamics [116] related to that exposed in Subsect. "Attractors with Inhomogeneously Distributed Activity", other based on temporal modulation of single cell spiking created by addition of oscillations at different frequencies [65,76,95]. Supplementary complexity and theoretical puzzle has been added by the startling discovery of "grid" cells [51,90] in the entorhinal cortex, a neural area in the hippocampal formation by which transits much of the sensory information that reaches the hippocampus.

Future Directions

We are still far from having a precise understanding of how neural systems operate, but there are daily advances at all levels from the description of detailed molecular mechanisms to that of high-level cognitive processes. We limit ourselves here to underlining some areas where further investigations of dynamics are clearly needed.

We have focused on networks with very little spatial structure but propagation and spatio-temporal dynamics most surely play important roles in neural information processing. This appears to be a very rich domain that imaging and theoretical investigations will help to explore.

A related question is that of the coordination between different neural areas. How do dynamics in one structure, for instance oscillations, serve to transmit and process information in a connected structure? As mentioned above, recent investigations of olfaction in insects provide a fascinating glimpse on this question. It will clearly be necessary to address it more generally in spite of the experimental difficulties.

Our emphasis in this article has been on electrical activity and its transmission. Another all-important aspect is neuromodulation which affects neuron properties on a slower time scale and probably a less local spatial scale. At present, the relation between these two facets of neural activity has been considered in few theoretical studies, but this needs to be thoroughly addressed in future work.

Homeostasis, or how dynamical properties of neuronal networks are maintained, is another question that needs analysis. Interesting experimental and theoretical work has been performed using the stomatogastric of the lobster as an example [75,85], but certainly this important topic requires further scrutiny.

These few questions among many others should make it clear that investigation of neuronal dynamics requires the development of powerful experimental techniques and theoretical analyses and that the subject will remain a topic of intense research in the forthcoming years.

Bibliography

Primary Literature

1. Abbott LF, van Vreeswijk C (1993) Asynchronous states in a network of pulse-coupled oscillators. Phys Rev E 48:1483–1490
2. Abbott LF, Varela JA, Sen K, Nelson SB (1997) Synaptic depression and cortical gain control. Science 275:220–224
3. Adrian ED (1934) Discharge frequencies in the cerebral and cerebellar cortex. Proc Physiol Soc 83:32–33
4. Adrian ED (1942) Olfactory reactions in the brain of the hedgehog. J Physiol 100:459–473
5. Amari S (1977) Dynamics of pattern formation in lateral-inhibition type neural fields. Biol Cybern 27:77–87
6. Amit DJ, Brunel N (1997) Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. Cereb Cortex 7:237–252
7. Amit DJ, Tsodyks MV (1991) Quantitative study of attractor neural network retrieving at low spike rates I: Substrate – spikes, rates and neuronal gain. Network 2:259–274
8. Badel L, Lefort S, Brette R, Petersen CCH, Gerstner W, Richardson MJE (2008) Dynamic I-V curves are reliable predictors of naturalistic pyramidal-neuron voltage traces. J Neurophysiol 99:656–666
9. Bak P, Tang C, Wiesenfeld K (1988) Self-organized criticality. Phys Rev A 38:364–374
10. Barbour B, Brunel N, Hakim V, Nadal J (2007) What can we learn from synaptic weight distributions? Trends Neurosci 30:622–629
11. Beggs JM, Plenz D (2003) Neuronal avalanches in neocortical circuits. J Neurosci 23:11167–77
12. Ben-Yishai R, Bar-Or RL, Sompolinsky H (1995) Theory of orientation tuning in visual cortex. Proc Natl Acad Sci USA 92:3844–3848
13. Berg RW, Alaburda A, Hounsgaard J (2007) Balanced inhibition and excitation drive spike activity in spinal half center. Science 315:390–3
14. Berger H (1929) Über das Elektroenkephalogramm des Menschen. Arch Psychiatr Nervenkrankh 87:527–570
15. Bernander O, Douglas RJ, Martin KA, Koch C (1991) Synaptic background activity determines spatio-temporal integration in single pyramidal cells. Proc Natl Acad Sci USA 88:11569–11573
16. Bliss TVP, Collingridge GL (1993) A synaptic model of memory: long-term potentiation in the hippocampus. Nature 361:31–39
17. Blumenfeld B, Bibitchkov D, Tsodyks MV (2006) Neural network model of the primary visual cortex: From functional architecture to lateral connectivity and back. J Comput Neurosci 20:219–241
18. Bolea S, Sanchez-Andres J, Huang X, Wu J (2006) Initiation and propagation of neuronal coactivation in the developing hippocampus. J Neurophysiol 95:552–561
19. Bromfield EB, Cavazos JE, Sirven JI (2006) An introduction to Epilepsy. American Epilepsy Society, Bethesda
20. Brunel N (2000) Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. J Comput Neurosci 8:183–208
21. Brunel N (2008) Modeling point neurons: from Hodgkin–Huxley to Integrate-and-Fire. In: De Schutter (ed) Computa-

tional modeling methods for neuroscientists. Mit Press, Cambridge
22. Brunel N, Hakim V (1999) Fast global oscillations in networks of integrate-and-fire neurons with low firing rates. Neural Comp 11:1621–1671
23. Brunel N, Hakim V (2008) Sparsely synchronized neuronal oscillations. Chaos 18:015113
24. Brunel N, Wang X-J (2001) Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. J Comput Neurosci 11:63–85
25. Brunel N, Wang X-J (2003) What determines the frequency of fast network oscillations with irregular neural discharges? J Neurophysiol 90:415–430
26. Brunel N, Hakim V, Richardson MJE (2003) Firing rate resonance in a generalized integrate-and-fire neuron with subthreshold resonance. Phys Rev E 67:051916
27. Burkitt AN (2006) A review of the integrate-and-fire neuron model: I. Homogeneous synaptic input. Biol Cybern 95:1–19
28. Burkitt AN (2006) A review of the integrate-and-fire neuron model: II. Inhomogeneous synaptic input and network properties. Biol Cybern 95:97–112
29. Cangiano L, Grillner S (2005) Mechanisms of rhythm generation in a spinal network deprived of crossed connections: the lamprey hemichord. J Neurosci 25:923–35
30. Cassenauer J, Laurent G (2007) Hebbian stdp in mushroom bodies facilitates the synchronous flow of olfactory information in locusts. Nature 448:709–713
31. Chow CC, Kopell N (2000) Dynamics of spiking neurons with electrical coupling. Neural Comput 12:1643–78
32. Compte A, Sanchez-Vives MV, McCormick DA, Wang X-J (2003) Cellular and network mechanisms of slow oscillatory activity (<1Hz) and wave propagations in a cortical network model. J Neurophysiol 90:2707–2725
33. Cossart R, Aronov D, Yuste R (2003) Attractor dynamics of network up states in the neocortex. Nature 423:283–288
34. Csicsvari J, Hirase H, Czurko A, Buzsáki G (1998) Reliability and state dependence of pyramidal cell-interneuron synapses in the hippocampus: an ensemble approach in the behaving rat. Neuron 21:179–189
35. Dayan P, Abbott L (2001) Theoretical neuroscience. MIT Press, Cambridge
36. De Schutter E (1999) Using realistic models to study synaptic integration in cerebellar Purkinje cells. Rev Neurosci 10:233–245
37. de Solages C, Szapiro G, Brunel N, Hakim V, Isope P, Buisseret P, Rousseau C, Barbour B, Léna C (2008) High-frequency organization and synchrony of activity in the purkinje cell layer of the cerebellum. Neuron 58:775–788
38. Destexhe A, Mainen ZF, Sejnowski TJ (1998) Kinetic models of synaptic transmission. In: Koch C, Segev I (eds) Methods in Neuronal Modeling, 2nd edn. Cambridge, MIT Press, pp 1-26
39. Ermentrout GB (1998) Neural networks as spatio-temporal pattern-forming systems. Rep Prog Phys 61:353–430
40. Feldmeyer D, Lubke J, Sakmann B (2006) Efficacy and connectivity of intracolumnar pairs of layer 2/3 pyramidal cells in the barrel cortex of juvenile rats. J Physiol 575:583–602
41. Fisahn A, Pike FG, Buhl EH, Paulsen O (1998) Cholinergic induction of network oscillations at 40 Hz in the in vitro. Nature 394:186–189
42. Fourcaud-Trocmé N, Hansel D, van Vreeswijk C, Brunel N (2003) How spike generation mechanisms determine the neuronal response to fluctuating inputs. J Neurosci 23:11628–11640
43. Freeman W (1991) The physiology of perception. Sci Am 264:78–85
44. Fries P, Reynolds J, Rorie A, Desimone R (2001) Modulation of oscillatory neuronal synchronization by selective visual attention. Science 291:1560–1563
45. Galarreta M, Hestrin S (1999) A network of fast-spiking cells in the neocortex connected by electrical synapses. Nature 402(6757):72–5
46. Gibson JR, Beierlein M, Connors BW (1999) Functional properties of electrical synapses between inhibitory interneurons of neocortical layer 4. J Neurophysiol 93:467–80
47. Gibson JR, Beierlein M, Connors BW (1999) Two networks of electrically coupled inhibitory neurons in neocortex. Nature 402(6757):75–9
48. Gierer A, Meinhardt H (1972) A theory of biological pattern formation. Kybernetik 12:30–9
49. Goldberg JA, Rokni U, Sompolinsky H (2004) Patterns of ongoing activity and the functional architecture of the primary visual cortex. Neuron 42:489–500
50. Golomb D, Ermentrout GB (2002) Slow excitation supports propagation of slow pulses in networks of excitatory and inhibitory populations. Phys Rev E 65:061911
51. Hafting T, Fyhn M, Molden S, Moser M, Moser E (2005) Microstructure of a spatial map in the entorhinal cortex. Nature 436:801–806
52. Hansel D, Mato G, Meunier C (1995) Synchrony in excitatory neural networks. Neural Computation 7:307–337
53. Hansel D, Mato G, Meunier C, Neltner L (1998) On numerical simulations of integrate-and-fire neural networks. Neural Computation 10:467–483
54. Heisenberg M (2003) Mushroom body memoir: from maps to models. Nat Rev Neurosci 4:266–275
55. Hodgkin AL, Huxley AF (1952) A quantitative description of membrane current and its application to conductance and excitation in nerve. J Physiol 117:500–544
56. Holmgren C, Harkany T, Svennenfors B, Zilberter Y (2003) Pyramidal cell communication within local networks in layer 2/3 of rat neocortex. J Physiol 551:139–153
57. Honeycutt RL (1992) Stochastic Runge–Kutta algorithms. I. White noise. Phys Rev A 45:600–603
58. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci USA 79:2554–2558
59. Issue S (1999) Special issue: The binding problem. Neuron 24:7–125
60. Izhikevich E, Gally J, Edelman G (2004) Spike-timing dynamics of neuronal groups. Cereb Cortex 14:933–944
61. Izhikevich EM (2001) Resonate-and-fire neurons. Neural Networks 14:883–894
62. Jacobi S, Moses E (2007) Variability and corresponding amplitude-velocity relation of activity propagating in one-dimensional neural cultures. J Neurophysiol 97:3597–3606
63. Jahnke S, Memmesheimer R-M, Timme M (2008) Stable irregular dynamics in spiking neural networks. Phys Rev Lett 100:048102

64. Jensen O, Lisman J (2000) Position reconstruction from an ensemble of hippocampal place cells: contribution of theta phase coding. J Neurophysiol 83:2602–2609
65. Kamondi A, Acsády L, Wang X-J, Buzsáki G (1998) Theta oscillations in somata and dendrites of hippocampal pyramidal cells in vivo: activity-dependent phase-precession of action potentials. Hippocampus 8:244–261
66. Keene A, Waddell S (2007) Drosophila olfactory memory: single genes to complex neural circuits. Nat Rev Neurosci 8:341–354
67. Kenet T, Bibitchkov D, Tsodyks MV, Grinvald A, Arieli A (2003) Spontaneously emerging cortical representations of visual attributes. Nature 425:954–956
68. Knight BW (1972) Dynamics of encoding in a population of neurons. J Gen Physiol 59:734–766
69. Koshiya N, Smith JC (1999) Neuronal pacemaker for breathing visualized in vitro. Nature 400:360–63
70. Koulakov AA, Raghavachari S, Kepecs A, Lisman JE (2002) Model for a robust neural integrator. Nat Neurosci 5:775–782
71. Kreiter AK, Singer W (1996) Stimulus dependent synchronization of neuronal responses in the visual cortex of the awake macaque monkey. J Neurosci 16:2381–2396
72. Kubota D, Colgin L, Casale M, Brucher F, Lynch G (2003) Endogenous waves in hippocampal slices. J Neurophysiol 89:81–89
73. Kuramoto Y (1984) Chemical oscillations, waves and turbulence. Springer, New York
74. Lapicque L (1907) Recherches quantitatives sur l'excitabilité électrique des nerfs traitée comme une polarisation. J Physiol Pathol Gen 9:620–635
75. LeMasson G, Marder E, Abbott L (1993) Activity-dependent regulation of conductances in model neurons. Science 259:1915–1917
76. Lengyel M, Szatmáry Z,Erdi P (2003) Dynamically detuned oscillations account for the coupled rate and temporal code of place cell firing. Hippocampus 13:700–714
77. Levina A, Hermann JM, Geisel T (2007) Dynamical synapses causing self-organized criticality in neural networks. Nature Physics 3:857–860
78. Lewis T, Rinzel J (2003) Dynamics of spiking neurons connected by both inhibitory and electrical coupling. J Comput Neurosci 14:283–309
79. Luczak A, Barthó P, Marguet S, Buzsáki G, Harris K (2007) Sequential structure of neocortical spontaneous activity in vivo. Proc Natl Acad Sci USA 104:347–352
80. Lumer ED, Friston KJ, Rees G (1998) Neural correlates of perceptual rivalry in the human brain. Science 280:1930–1934
81. Maeda E, Robinson HPC, Kawana A (1995) The mechanisms of generation and propagation of synchronized bursting in developing networks of cortical neurons. J Neurosci 15:6834–45
82. Malenka RC, Bear MF (2004) LTP and LTD: an embarrassment of riches. Neuron 44:5–21
83. Malinow R, Mainen ZF, Hayashi Y (2000) LTP mechanisms: from silence to four-lane traffic. Curr Opin Neurobiol 10:352–357
84. Mancilla JG, Lewis TJ, Pinto DJ, Rinzel J, Connors BW (2007) Synchronization of electrically coupled pairs of inhibitory interneurons in neocortex. J Neurosci 27:2058–73
85. Marder E, Goaillard J (2006) Variability, compensation and homeostasis in neuron and network function. Nat Rev Neurosci 7:563–574
86. Markram H (2006) The blue brain project. Nat Rev Neurosci 7:153–160
87. Mason A, Nicoll A, Stratford K (1991) Synaptic transmission between individual pyramidal neurons of the rat visual cortex in vitro. J Neurosci 11:72–84
88. Mattia M, Del Giudice P (2000) Efficient event-driven simulation of large networks of spiking neurons and dynamical synapses. Neural Comput 12:2305–2329
89. Mazor O, Laurent G (2005) Transient dynamics versus fixed points in odor representations by locust antennal lobe projection neurons. Neuron 48:661–673
90. McNaughton B, Battaglia F, Jensen O, Moser E, Moser M (2006) Path integration and the neural basis of the 'cognitive map'. Nat Rev Neurosci 7:663–678
91. Mongillo G, Curti E, Romani S, Amit DJ (2005) Learning in realistic networks of spiking neurons and spike-driven plastic synapses. Eur J Neurosci 21:3143–3160
92. Moreno-Bote R, Rinzel J, Rubin N (2007) Noise-induced alternations in an attractor network model of perceptual bistability. J Neurophysiol 98:1125–1139
93. Murphy TH, Blatter LA, Wier WG, Baraban JM (1992) Spontaneous synchronous synaptic calcium transients in cultured cortical neurons. J Neurosci 12:4834–45
94. O'Keefe J, Dostrovsky J (1971) The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely moving rat. Exp Brain Res 34:171–175
95. O'Keefe J, Recce M (1993) Phase relationship between hippocampal place units and the EEG theta rhythm. Hippocampus 3:317–330
96. Opitz T, Lima ADD, Voigt T (2002) Spontaneous development of synchronous oscillatory activity during maturation of cortical neurons in vitro. J Neurophysiol 88:2196–206
97. Perez-Orive J, Mazor O, Turner G, Cassenaer S, Wilson R, Laurent G (2002) Oscillations and sparsening of odor representations in the mushroom body. Science 297:359–365
98. Petersen C, Grinvald A, Sakmann B (2003) Spatiotemporal dynamics of sensory responses in layer 2/3 of rat barrel cortex measured in vivo by voltage-sensitive dye imaging combined with whole-cell voltage recordings and neuron reconstructions. J Neurosci 23:1298–1309
99. Pikovsky A, Rosenblum M, Kurth J (2001) Synchronization, a universal concept in nonlinear science. Cambridge University Press, Cambridge
100. Polonsky A, Blake R, Braun J, Heeger DJ (2000) Neuronal activity in human primary visual cortex correlates with perception during binocular rivalry. Nat Neurosci 3:1153–1159
101. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical recipes in C. Cambridge University Press, Cambridge
102. Renart A, Moreno R, Wang X-J (2003) Robust spatial working memory through homeostatic synaptic scaling in heterogenous cortical networks. Neuron 38:473–485
103. Reyes A, Fetz E (1993) Effects of transient depolarizing potentials on the firing rate of cat neocortical neurons. J Neurophysiol 69:1673–1683
104. Richardson MJE, Brunel N, Hakim V (2003) From subthreshold to firing-rate resonance. J Neurophysiol 89:2538–2554

105. Risken H (1984) The Fokker–Planck equation: methods of solution and applications. Springer, Berlin
106. Roelfsema P, Lamme V, Spekreijse H (2004) Synchrony and covariation of firing rates in the primary visual cortex during contour grouping. Nat Neurosci 7:982–991
107. Sanchez-Vives MV, McCormick DA (2000) Cellular and network mechanisms of rhythmic recurrent activity in neocortex. Nat Neurosci 3:1027–34
108. Shadlen MN, Newsome WT (1998) The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. J Neurosci 18:3870–3896
109. Shu Y, Hasenstaub A, McCormick DA (2003) Turning on and off recurrent balanced cortical activity. Nature 423:288–93
110. Sjöström PJ, Turrigiano GG, Nelson S (2001) Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. Neuron 32:1149–1164
111. Softky WR, Koch C (1993) The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. J Neurosci 13:334–350
112. Stein R (1965) A theoretical analysis of neuronal variability. Biophys J 5:173–194
113. Stopfer M, Jayaraman V, Laurent G (2003) Intensity versus identity coding in an olfactory system. Neuron 39:991–1004
114. Tong F, Nakayama K, Vaughan JT, Kanwisher N (1998) Binocular rivalry and visual awareness in human extrastriate visual cortex. Neuron 21:761–773
115. Tsodyks MV, Kenet T, Grinvald A, Arieli A (1999) Linking spontaneous activity of single cortical neurons and the underlying functional architecture. Science 286:1943–1946
116. Tsodyks MV, Skaggs W, Sejnowski T, McNaughton B (1996) Population dynamics and theta rhythm phase precession of hippocampal place cell firing: a spiking neuron model. Hippocampus 6:271–280
117. Tsodyks MV, Pawelzik, Markran H (1998) Neural networks with dynamic synapses. Neural Comp 10:821–835
118. Tsodyks MV, Uziel A, Markram H (2000) Synchrony generation in recurrent networks with frequency-dependent synapses. J Neurosci 20:RC50
119. Turrigiano GG, Nelson SB (2000) Hebb and homeostasis in neuronal plasticity. Curr Opin Neurobiol 10:358–364
120. van Vreeswijk C, Sompalinsky h (1996) Chaos in neuronal networks with balanced excitatory and inhibitory activity. Science 274:1724–1726
121. van Vreeswijk C, Sompolinsky H (1998) Chaotic balanced state in a model of cortical circuits. Neural Computation 10:1321–1371
122. van Vreeswijk C, Abbott L, Ermentrout GB (1994) When inhibition not excitation synchronizes neural firing. J Comput Neurosci 1:313–321
123. Wallén P et al (2007) Sodium-dependent potassium channels of a slack-like subtype contribute to slow afterhyperpolarization in lamprey spinal neurons. J Physiol 585:75–90
124. Wilson HR (2003) Computational evidence for a rivalry hierarchy in vision. Proc Natl Acad Sci USA 100:14499–14503
125. Wilson M, McNaughton B (1993) Dynamics of the hippocampal ensemble code for space. Science 261:1055–1058
126. Wyart C, Cocco S, Bourdieu L, Léger JF, Herr C, Chatenay D (2005) Dynamics of excitatory synaptic components in sustained firing at low rates. J Neurophysiol 96:3370–80
127. Zillmer R, Brunel N, Hansel D (2008) Irregular states in randomly diluted networks of leaky integrate-and-fire neurons. in preparation.

Books and Reviews

Buzsaki G (2006) Rhythms of the brain. Oxford University Press
Chow C, Gutkin B, Hansel D, Meunier C, Dalibard J (eds) (2004) Methods and models in neurophysics, Les Houches 2003. North-Holland
Dayan P, Abbott L (2001) Theoretical neuroscience. MIT Press, Cambridge
Gerstner W, Kistler WM (2002) Spiking neuron models: single neurons, populations, plasticity. Cambridge University Press, Cambridge
Wang X-J (2003) Neural oscillations. In: Nadel L (ed) Encyclopedia of Cognitive Science. MacMillan, London, pp 272–280

Noise and Stability in Modelocked Soliton Lasers

Brian H. Kolner[1,2]
[1] Department of Applied Science, University of California, Davis, USA
[2] Department of Electrical and Computer Engineering, University of California, Davis, USA

Article Outline

Glossary
Definition of the Subject
Introduction
List of Symbols
Basic Laser Physics
Laser Dynamics
Laser Instability, Chaos and the Lorenz Equations
Modelocking
Solitons
Modelocked Soliton Lasers
Laser Noise and Linewidth
Noise in Soliton Lasers
Large Scale Instability in Soliton Lasers
Laser Clocks
Conclusions and Outlook for the Future
Acknowledgments
Bibliography

Glossary

First and second laser thresholds Minimum conditions for producing stable and chaotic laser operation, respectively.

Lorenz equations A system of three coupled nonlinear differential equations describing convective fluid flow in cells. These equations and their solutions launched the field of chaos theory.

Modelocked laser A laser with evenly spaced modes which have their phases locked together so that the superposition of the modes creates a periodic train of very short pulses < 1 ps).

Poisson photon distribution Probability of k photons arriving in a given interval of time for which there are an average of $\bar{n}$ photons.

Relaxation oscillation Periodic fluctuation about an equilibrium point when a system, initially operating in steady-state, is subjected to a transient perturbation. The system "relaxes" back to equilibrium through a (usually) damped sinusoidal response.

Shot noise Noise produced by the random arrival of electrons in a photodetector illuminated by a laser field.

Soliton A solitary wave packet which is a solution to a nonlinear wave equation in a medium possessing both dispersion and nonlinearity. These effects balance exactly so that the wave packet maintains its shape and is stable against perturbations.

Definition of the Subject

Since its conception in 1958, the laser has become a nearly ubiquitous feature of the technological landscape finding its way into a spectacular array of devices from data storage to entertainment to fundamental science and metrology and myriad others. One of the principal features of the laser is its remarkable coherence. It is a nearly ideal source of monochromatic radiation and from this property the fields of precision spectroscopy and timekeeping have reaped huge rewards. Recently, an innovation in the design of a class of ultrashort-pulse lasers has resulted in a new type of optical clockwork with stability that rivals and even supercedes that of the best atomic clocks, including the cesium-beam clocks which, by international agreement, are used to define the second. The nearly perfect train of optical pulses emitted from these laser clocks are associated with a nearly ideal spectrum of periodic spikes, or comb-lines, in the optical frequency domain which can act as a "frequency ruler" for precision measurements. In just a few short years this new type of laser has revolutionized the fields of precision spectroscopy and timekeeping and there is much more yet to come. This raises the interesting question "how stable can these new laser clocks really be?". To begin to address this we must consider the fundamental mechanisms of noise and stability of lasers. This is the subject of the present chapter.

Introduction

Lasers are wonderfully complex systems incorporating energy storage and conversion, interaction of radiation with ensembles of atoms or molecules, classical optics, feedback and linear and nonlinear processes everywhere. Fortunately, most of the processes involved in laser action are amenable to approximations that make the system analytically tractable, although they can frequently mask very interesting and often technologically important phenomena.

The most striking feature of the laser is the ability to combine the electromagnetic radiation of, typically, 10^{15} atomic oscillators in a coherent fashion and produce a single narrow beam of light at a single frequency. The property of any radiation source that describes its power per-unit-area, per-unit-frequency interval, per-solid-angle is called the *brightness*. No other source of electromagnetic radiation in the visible portion of the spectrum comes close to the brightness of a laser.

The second most striking feature of the laser after its brightness is its degree of spatial and temporal coherence. By this we mean the stability in space and time of the harmonic nature of the electromagnetic field. Coherence is typically characterized by interfering two portions of wavefronts separated either in space or in time. High spatial coherence allows the laser field to be focused to a tight spot while high temporal coherence allows the field to accurately probe very narrow spectroscopic features of atoms or molecules or to serve as a frequency or wavelength standard. Although both of these properties influence the brightness, we will concern ourselves in this chapter exclusively with the temporal properties of the field.

To study the temporal behavior of any oscillator it will be useful to introduce the concepts of amplitude and phase instability instability. A perfect, unit-amplitude, harmonic oscillator has an output (field, voltage, current, position, etc.) given by the simple function $E_o(t) = \sin(\omega_o t)$. No physical oscillator displays perfect amplitude and phase stability so a more realistic function could be written as

$$E_n(t) = (1 + \epsilon(t)) \sin(\omega_o t + \phi(t)) \tag{1}$$

where $\epsilon(t)$ and $\phi(t)$ are random processes describing the effects of amplitude and phase noise, respectively. In "good" oscillators, these random processes do not cause dramatic deviations from the ideal behavior and if one measures the waveform in the time domain, they are seldom observable. However, if the waveform is analyzed for its frequency content, even tiny amounts of amplitude or phase noise can be readily measured. In fact, almost all oscillator characterization is done by measuring the frequency spectrum except for the case of the very most sta-

ble sources (e. g. atomic clocks and the new self-referenced laser clocks) where the statistics of the timing fluctuation are of greatest interest and the time domain yields valuable insight. In both cases very long time records are acquired. The domain of analysis is then predicated on whether the noise processes have slowly (time-domain statistics) or rapidly (frequency-domain power spectra) evolving components [1,2,3].

Figure 1a shows a simple sine-wave with a small amount of amplitude and phase noise of the form described by (1). For this example $\epsilon(t)$ and $\phi(t)$ are zero-mean Gaussian random processes with standard deviations of 0.2 and 0.02π, respectively. In addition, the spectral content of the noise was low-pass filtered to emphasize the low frequency content. Notice that the time-domain representation of the function displays no observable instability while the power spectrum $|\mathcal{E}_n(\omega)|^2$ shown in Fig. 1b contains a very strong signature of the combined amplitude and phase noise. For reference, the spectrum of the pure sinusoid $|\mathcal{E}_o(\omega)|^2$ is also shown. Its near delta-function form is only marred by the noise floor set by the limited precision of the computer. This figure demonstrates several very important ideas. First, even a very small amount of noise added to the amplitude and phase of an otherwise perfect sinusoid can dramatically alter the purity of its spectrum. Second, once the spectrum is corrupted by the noise, precise establishment of the center frequency of the oscillator is now complicated by the spreading of the energy in frequency space. This so-called "linewidth broadening" is present in all oscillators and a thorough study of the causative mechanisms is essential for good oscillator design. Finally, the spreading of energy into noise sidebands adjacent to the main peak can completely mask the presence of other desired signals in that region. For this reason, spectral purity of oscillators is a very important quality in the design and realization of receivers at any frequency from radio through microwave and optical portions of the spectrum.

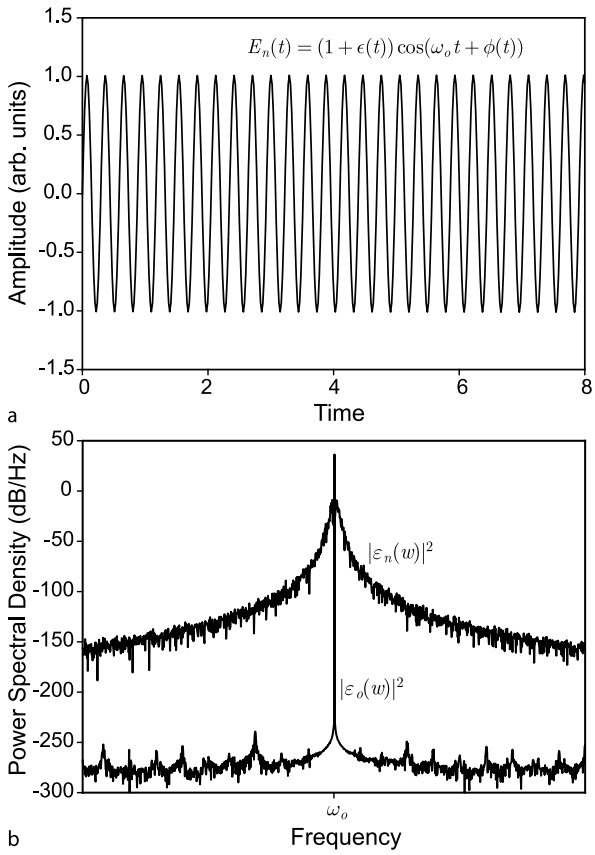

Noise and Stability in Modelocked Soliton Lasers, Figure 1
a Sinewave oscillator output with small amount of amplitude and phase noise. b Power spectrum of oscillator output $|\mathcal{E}_n(\omega)|^2$. Power spectrum of a pure sinusoid, $|\mathcal{E}_o(\omega)|^2$, (i. e. $\epsilon(t) = \phi(t) = 0$) shown for comparison

The addition of small amounts of random noise to the amplitude and phase of a laser field is but one kind of instability which we will call "small scale". The fluctuations are typically very small compared to the average quantities and the laser operates in a steady-state regime with essentially constant power flow into and out of the laser system and constant energy-level populations. Examples of processes that produce small-scale amplitude and phase noise include spontaneous emission of photons into the desired cavity mode in addition to the stimulated emission photons created by lasing action. This gives rise to a fundamental linewidth limit (i. e. the Schawlow–Townes limit). The discrete and random generation of charge carriers in a photon detector gives rise to "shot noise" when detecting laser fields. Also, amplitude noise present on the pump source will be transferred to the laser creating both amplitude and phase noise in the final output field. These are summarized in Fig. 2.

There is another type of instability which we will call "large-scale" and this is characterized by large, and sometimes chaotic, swings in the principal laser parameters such as electric field, population inversion and atomic polarization. The first instance of these phenomena always occurs when a laser first starts up as the population inversion approaches threshold for oscillation. The closely related phenomenon of relaxation oscillation occurs when a sudden change in the pumping rate or intracavity loss occurs and upsets the steady-state balance between pumping, stimulated emission and output coupling. Finally, in certain regions of parameter space, the laser can be operated in a wildly unstable manner that is quite analogous

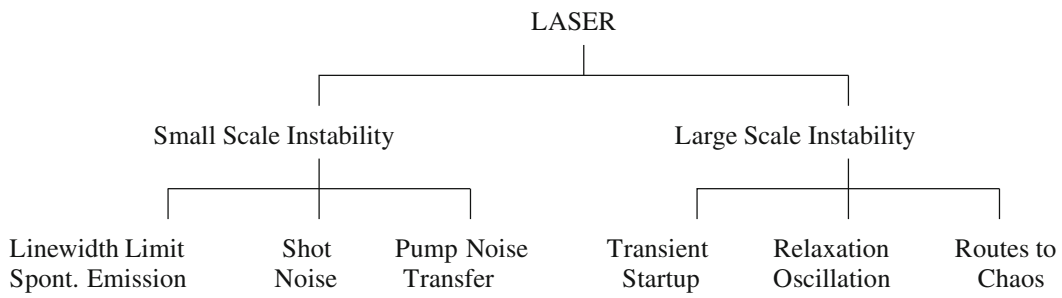

Noise and Stability in Modelocked Soliton Lasers, Figure 2
Categorization of noise and instabilities associated with laser oscillators

to nonlinear dissipative systems undergoing chaotic dynamics. All of the phenomena mentioned and shown in Fig. 2 will be discussed in this chapter. Many of the topics can be found in textbooks [4,5,6,7,8,9,10,11] and comprehensive monographs [12,13,14,15,16,17,18,19] and some, which will be presented toward the end of the chapter, are quite new.

List of Symbols

b = radius of laser rod (m).
$H_{\mathrm{AM}}(\omega_m)$ = complex AM-to-AM noise transfer function.
$H_{\mathrm{PM}}(\omega_m)$ = complex AM-to-PM noise transfer function.
k = thermal diffusivity ($\mathrm{m^2/s}$).
K = photon number/population inversion density coupling coefficient ($\mathrm{m^3\,s^{-1}}$).
L_c = cavity length without gain medium (m).
L_a = length of gain medium (m).
$n(t)$ = cavity photon number density ($\mathrm{m^{-3}}$).
$\tilde{n}_1$ = complex photon number density phasor amplitude ($\mathrm{m^{-3}}$).
$N(t)$ = population inversion density ($\mathrm{m^{-3}}$).
$\tilde{N}_1$ = complex population inversion density phasor amplitude ($\mathrm{m^{-3}}$).
p_c = round trip cavity distance (m).
p_m = round trip distance through gain medium (m).
P_{L_0} = steady-state laser output power (W).
P_{P_0} = steady-state pump power (W).
r = normalized pumping rate above threshold.
r' = radial coordinate of laser rod (m).
$R_P(t)$ = effective pumping rate density into upper laser level ($\mathrm{m^{-3}\,s^{-1}}$).
R_{P_0} = steady-state pumping rate density into upper laser level ($\mathrm{m^{-3}\,s^{-1}}$).
R_{P_1} = peak sinusoidal amplitude of pumping rate density ($\mathrm{m^{-3}\,s^{-1}}$).
R_{tot} = product of mirror reflectivities.
T = round trip cavity time (s).
T_{oc} = power transmission of output coupler.
V_c = cavity mode volume including gain medium ($\mathrm{m^3}$).
V_g = mode volume of gain medium ($\mathrm{m^3}$).
α_0 = ohmic loss coefficient ($\mathrm{m^{-1}}$).
$\beta(\omega_m)$ = peak phase deviation.
β_l = eigenvalues of the radial heat diffusion problem.
ΔP_L = peak fractional laser power deviation.
ΔP_P = peak fractional pump power deviation.
$\Delta\omega_a$ = gain bandwidth, FWHM ($\mathrm{s^{-1}}$).
γ_c = cavity decay rate = $\frac{2\alpha_0 p_c + \ln(1/R_{\mathrm{tot}})}{T}$ ($\mathrm{s^{-1}}$).
γ_{rad} = radiative decay rate out of upper level ($\mathrm{s^{-1}}$).
γ_2 = total upper level decay rate ($\mathrm{s^{-1}}$).
η = laser slope efficiency.
ω_a = resonant frequency of active gain medium ($\mathrm{s^{-1}}$).
ω_m = pump modulation frequency ($\mathrm{s^{-1}}$).
τ_{co} = unperturbed round-trip cavity time.
τ_p = modelocked pulsewidth.
3^* = polarization-field alignment factor.

Basic Laser Physics

Lasers represent a very complex system of interrelated objects and phenomena. In order to understand the origin of the systems of equations that can predict laser behavior, we should first review the basic principles of laser operation.

The fundamental mechanism of lasers is the stimulated emission of electromagnetic waves from atoms or molecules which are temporarily in an excited state. That is, a configuration corresponding to an energy level that is higher than the atom would normally find itself in thermal equilibrium. From a statistical standpoint, when a large number of atoms are gathered together and share a common temperature T, the probability of an atom being in any energy level is given by the Boltzmann distribution. This tells us that the ratio of the number of atoms, N_2, in energy level E_2, to the number of atoms, N_1 in a lower energy level, E_1 is given by

$$\frac{N_2}{N_1} = \exp\left(-\frac{E_2 - E_1}{kT}\right) = \exp\left(-\frac{\hbar\omega_a}{kT}\right) \tag{2}$$

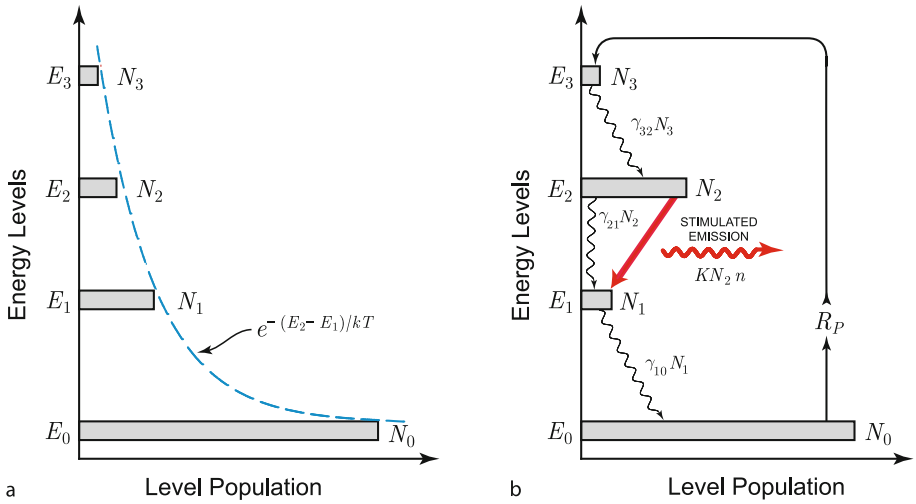

Noise and Stability in Modelocked Soliton Lasers, Figure 3

Energy levels E_i and corresponding populations N_i for the case of a thermal equilibrium with no pumping or driving signals and b pumping with rate R_p from the lowest energy level E_0 through E_3 to the upper laser level E_2. Relaxation rates $\gamma_{ij}N_i$ and stimulated emission KN_2n compete with the pumping rate to form a closed system described by rate equations

where $k = 1.381 \times 10^{-23}\,\mathrm{JK}^{-1}$ is Boltzmann's constant, $\hbar = 6.626 \times 10^{-34}$ Js is Planck's constant and $\omega_a \equiv (E_2 - E_1)/\hbar$ is the radian frequency of electromagnetic radiation corresponding to the transition between the energy levels. Any pair of energy levels must have this same exponential relationship between the corresponding populations as shown in Fig. 3a.

In order to change a medium from absorbing radiation between two energy levels to contributing energy to a radiation field (and thus exhibit amplification), there must be more atoms in a higher energy level, or state, than in a lower level. This, of course, no longer constitutes a situation of thermal equilibrium and must be created by an outside agent. The preparation of an excited state is brought about by pumping the atoms from a lower energy state to a higher energy state through some energy transfer mechanism. Figure 3b shows how energy is transferred from the ground state E_0 to the higher energy state E_3 at a rate R_P. This diagram depicts a so-called four-level laser where the uppermost level exhibits rapid relaxation, $\gamma_{32}N_3$, to the upper laser level E_2, where the atoms stay for a considerable time. If the relaxation from the lower laser level $\gamma_{10}N_1$ is also much faster than relaxation from $E_2 \rightarrow E_1$, then the population N_2 can be made to exceed N_1 creating a "population inversion". Under these circumstances, an electromagnetic wave can experience amplification by the process of stimulated emission. Stimulated emission of a quantum of electromagnetic energy occurs when the atom is driven by an electromagnetic field whose frequency is matched to the energy difference between two levels for which a population inversion exists. The remarkable feature of the stimulated emission is that it is coherent with the driving field. This is also an essential feature of laser action since any oscillator requires both gain and feedback of coherently superposed waves. The feedback, in the case of a laser, is provided by mirrors which return most of the amplified fields back to the atoms to sustain the stimulated emission process in concert with the pumping action.

As we shall see, the dynamics of laser operation are controlled by the time constants of many competing processes. Some of the these processes are microscopic and fundamental (e. g. atomic lifetimes), and many are macroscopic and environmental (e. g. Doppler effects, crystal strain broadening, cavity geometry, etc.). The multitude of different time constants, energy supply and depletion rates and the (usually) very large number of atoms required to deliver useful power output suggests that there is great potential for cooperative effects between the ensemble of atomic oscillators and the electromagnetic fields which drive them and to which they contribute.

Although the coupling of radiation to atoms is most rigorously described using quantum mechanics, when large numbers of atoms are involved a "mostly classical" treatment is satisfactory and brings out almost all of the important phenomena of laser dynamics without requiring the extra machinery of quantum physics. The one concession we make to the quantum world is reliance on the notion of defined energy levels or eigenstates of the atoms. It is the transitions between these states that characterize

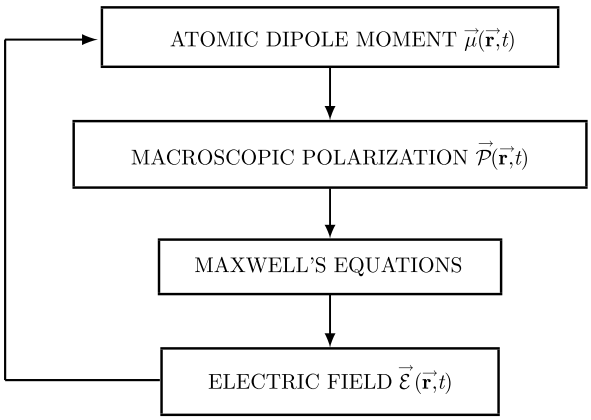

Noise and Stability in Modelocked Soliton Lasers, Figure 4
Self-consistent flow model describing coupling of atomic oscillators to electromagnetic fields

the frequencies of radiation, either absorbed or emitted, from the oscillator.

In order to gain a little insight into the physics of laser action, we can step back and take a broader view. The simplified flow model shown in Fig. 4 describes a self-consistent picture of the coupled processes that combine the radiation from individual atomic or molecular oscillators. In the classical picture of the interaction of radiation with matter, the electromagnetic wave causes a small time-varying displacement of the atom's charge cloud. The product of this displacement and the net charge displaced is called the *dipole moment*; $\vec{\mu}(\vec{\mathbf{r}}, t) = -\mathrm{e}\vec{x}(\vec{\mathbf{r}}, t)$ where e is the charge on a single electron and $\vec{x}(\vec{\mathbf{r}}, t)$ is the separation between the charge cloud and its positively charged ion located at position $\vec{\mathbf{r}}$. These microscopic dipole moments are akin to tiny bar magnets (except that these are electric dipole fields) and thus a large number of them in a macroscopic volume can have a significant effect on the local total electric field. An ensemble of microscopic dipole moments collectively makes up the polarization field $\vec{\mathcal{P}}(\vec{\mathbf{r}}, t)$, which is the sum of all the individual dipole moments divided by the volume enclosing those dipoles (therefore a "dipole moment per-unit-volume"). The polarization field combines with the electric field to yield the so-called displacement field $\vec{\mathcal{D}}(\vec{\mathbf{r}}, t)$ which is part of Maxwell's famous equations of electromagnetism [20]. In fact, one can show that the polarization field acts as a driving term for the differential wave equation for the electric field. The solution to the wave equation is the electric field that is responsible for establishing the individual dipole moments in the first place and we are back at the beginning of our flow chart.

Now we begin a little more detailed analysis to discover the basic dynamics of the material polarization which is responsible for laser action. The principles that are developed in the following section can be found in any good textbook on optics or laser physics but often the notation is inconsistent, which can cause confusion. Here we mostly follow the approach and notation of Siegman [4].

The equation of motion for the dipole moment $\vec{\mu}(\vec{\mathbf{r}}, t)$ due to a driving field $\vec{\mathcal{E}}(\vec{\mathbf{r}}, t)$ is that of the well-known damped harmonic oscillator. In the context of an electron bound to a positively-charged nucleus, this is frequently referred to as the "classical electron oscillator" or CEO model;

$$\frac{\mathrm{d}^2\mu_x(t)}{\mathrm{d}t^2} + \gamma\frac{\mathrm{d}\mu_x(t)}{\mathrm{d}t} + \omega_\mathrm{a}^2\mu_x(t) = \frac{\mathrm{e}^2}{m}\mathcal{E}_x(t) \tag{3}$$

where γ is the damping term or energy decay rate (about which more will be said shortly), ω_a is the atomic resonant frequency, and, for simplicity, we have written just one of the scalar components of the full vector system. In the absence of a forcing function, $\mathcal{E}_x(t) = 0$, the natural response for the single dipole is

$$\mu(t) = \mu_\mathrm{o}\exp\left[-\frac{\gamma}{2}t + \mathrm{i}\omega_\mathrm{a}' t\right] \tag{4}$$

where ω_a' is the actual oscillation frequency in the presence of damping

$$\omega_\mathrm{a}' \equiv \sqrt{\omega_\mathrm{a}^2 - (\gamma/2)^2} \tag{5}$$

and ω_a is the natural oscillation frequency of the atom in the CEO model. In the quantum picture $E_{21} = \hbar\omega_\mathrm{a}$ is the energy difference between the excited and ground state energy levels. If we calculate the purely radiative decay rate for a classical electron oscillator

$$\gamma_\mathrm{rad,ceo} = \frac{\mathrm{e}^2\omega_\mathrm{a}^2}{6\pi\epsilon mc^3} \tag{6}$$

we find a typical value of $\gamma_\mathrm{rad,ceo} \approx 10^8\ \mathrm{s}^{-1}$ [4]. Compared to the optical resonant frequency of $\omega_\mathrm{a} \approx 2\pi \times 10^{15}\ \mathrm{s}^{-1}$ it is reasonable to set $\omega_\mathrm{a}' = \omega_\mathrm{a}$ for all future discussions.

When we add up the dipole moments of an ensemble of atoms with density N in a small volume V, we obtain the dipole moment-per-unit volume, or, the *polarization* $\vec{\mathcal{P}}(\vec{\mathbf{r}}, t) = \sum_{i=1}^{NV}\vec{\mu}_i(\vec{\mathbf{r}}, t)$. Like the electric field, the polarization is a vector field with space and time dependence. The importance of the material polarization becomes clear when we realize that once the dipoles are set oscillating by an incident field, they become source terms for additional radiation and thus modify the total electric field. This can be summarized elegantly in the wave equation for the electric field which is derived from Maxwell's equations

$$\nabla^2\vec{\mathcal{E}}(\vec{\mathbf{r}}, t) - \mu_\mathrm{o}\epsilon_\mathrm{o}\frac{\partial^2\vec{\mathcal{E}}(\vec{\mathbf{r}}, t)}{\partial t^2} = \mu_\mathrm{o}\frac{\partial^2\vec{\mathcal{P}}(\vec{\mathbf{r}}, t)}{\partial t^2}\,. \tag{7}$$

The term on the right-hand side is the source term or forcing function for the wave equation.

Natural Lifetime and Dephasing

An important attribute of both the dipole moment and polarization is the exponential decay time associated with the natural lifetime of a single oscillator and the dephasing of an ensemble of oscillators. The CEO model of the atom includes a phenomenological damping term γ which accounts for energy lost by radiation γ_{rad} (i. e. spontaneous emission) and nonradiative effects γ_{nr} such as collisions with other atoms and vessel walls. The two combine to form a net energy decay rate γ such that the internal energy of the atom's electron cloud motion evolves according to

$$U(t) = U_{\text{o}} \exp(-\gamma t) \quad \text{where} \quad \gamma = \gamma_{\text{rad}} + \gamma_{\text{nr}} \,. \tag{8}$$

The energy decay rate γ is sometimes seen in the literature as $\gamma_{\parallel}$ and called the "longitudinal relaxation rate" [6,16] a term held over from the Bloch equations of nuclear magnetic resonance [21].

When an ensemble of simple classical electron oscillators, each with a natural response (4), radiates together we must carefully consider their superposition. If they all start oscillating in phase and there is no disruption, then it is easy to see that the macroscopic polarization field will have the same exponential decay as a single oscillator and is given by the sum

$$\mathcal{P}_x(t) = N\mu_{x\text{o}} \exp\left[-\frac{\gamma}{2}t + \mathrm{i}\omega_{\text{a}} t\right] \,. \tag{9}$$

For simplicity we have dropped the vector notation for the polarization field and just consider one Cartesian coordinate.

Unfortunately, real media do not exhibit a continuous fully coherent response as in (9) but instead undergo *dephasing events*such as collisions which tend to randomize the phases of the oscillators with respect to each other. This effect can be accounted for by introducing an additional decay rate term, $2/T_2$, which describes the dephasing time of the coherent ensemble, into the equation of motion for the polarization

$$\frac{\mathrm{d}^2\mathcal{P}_x(t)}{\mathrm{d}t^2} + \left(\gamma + \frac{2}{T_2}\right)\frac{\mathrm{d}\mathcal{P}_x(t)}{\mathrm{d}t} + \omega_{\text{a}}^2\mathcal{P}_x(t) = \frac{N\mathrm{e}^2}{m}\mathcal{E}_x(t) \,. \tag{10}$$

The factor of 2 in $2/T_2$ comes about because the phenomenon of dephasing is a field effect, not an energy decay. The natural response is

$$\mathcal{P}_x(t) = \mathcal{P}_{x\text{xo}} \exp\left[-\left(\frac{\gamma}{2} + \frac{1}{T_2}\right)t + \mathrm{i}\omega_{\text{a}} t\right] \,. \tag{11}$$

The total decay rate for the natural response polarization field is now $\gamma/2 + 1/T_2$. Equivalently, the 1/e time constant for the decay of the polarization field in $(\gamma/2 + 1/T_2)^{-1}$. The time constant T_2 has the very real physical interpretation as the average time an atom oscillates between suffering a dephasing event. The total decay rate (which we will see represents a bandwidth) or more properly, it's inverse, represents the first major time constant responsible for interesting laser dynamics. It is, however, generally dominated by the dephasing rate ($T_2 \ll 2/\gamma$). We note, also, that in the literature of nuclear magnetic resonance, the dephasing rate is often seen as $\gamma_{\perp}$ and called the "transverse" or polarization relaxation rate [6,16].

Steady-State Response and the Susceptibility

The characterization of any linear system by differential equations allows one to evaluate a natural, or homogeneous, response (11) and a forced, or inhomogeneous response. A particularly useful forcing function is the steady harmonic source $\mathrm{e}^{\mathrm{i}\omega t}$ since the system will respond in kind with a harmonic signal whose amplitude and phase depend on the amplitude and frequency of the driving function. To this end, we introduce the following forcing and response functions

$$\mathcal{E}_x(t) = \mathrm{Re}\left[\tilde{E}_x \mathrm{e}^{\mathrm{i}\omega t}\right] = \frac{1}{2}\left[\tilde{E}_x \mathrm{e}^{\mathrm{i}\omega t} + \tilde{E}_x^* \mathrm{e}^{-\mathrm{i}\omega t}\right] \tag{12}$$

$$\mathcal{P}_x(t) = \mathrm{Re}\left[\tilde{P}_x \mathrm{e}^{\mathrm{i}\omega t}\right] = \frac{1}{2}\left[\tilde{P}_x \mathrm{e}^{\mathrm{i}\omega t} + \tilde{P}_x^* \mathrm{e}^{-\mathrm{i}\omega t}\right] \tag{13}$$

into Eq. (10). The solution for the polarization amplitude is therefore dependent on the driving frequency according to

$$\tilde{P}(\omega) = \frac{N\mathrm{e}^2}{m}\frac{1}{\omega_{\text{a}}^2 - \omega^2 + \mathrm{i}\omega(\gamma + 2/T_2)}\tilde{E}(\omega) \,. \tag{14}$$

This polarization field now becomes a driving source term in Maxwell's equations through the displacement field

$$\tilde{D}(\omega) = \epsilon_0 \tilde{E}(\omega) + \tilde{P}(\omega) \,. \tag{15}$$

The atoms involved in laser action usually make up only a small fraction of the total matter density of the active medium. The remaining host atoms also have atomic resonances and contribute a polarization response but the resonant frequencies are usually far removed from the laser transitions and the contribution to the polarization is therefore rather constant throughout the range of the laser atom's response. Therefore, it is useful to separate

out the polarization due to the host medium from the laser atoms's atomic response and write

$$\tilde{P}(\omega) = \tilde{P}_{\text{host}}(\omega) + \tilde{P}_{\text{at}}(\omega) \tag{16}$$

which then results in a modified constitutive relation

$$\tilde{D}(\omega) = \epsilon_0 \tilde{E}(\omega) + \tilde{P}_{\text{host}}(\omega) + \tilde{P}_{\text{at}}(\omega) = \epsilon_{\text{host}} \tilde{E}(\omega) + \tilde{P}_{\text{at}}(\omega). \tag{17}$$

Since the atomic polarization depends linearly on the driving field, we can define a scale factor which we call the "complex atomic susceptibility" where

$$\tilde{\chi}(\omega) = \frac{\tilde{P}_{\text{at}}(\omega)}{\epsilon_{\text{host}} \tilde{E}(\omega)} = \frac{Ne^2}{\epsilon_{\text{host}} m} \frac{1}{\left[\omega_{\text{a}}^2 - \omega^2 + \text{i}\omega(\gamma + 2/T_2)\right]}. \tag{18}$$

The frequency dependence of $\tilde{\chi}(\omega)$ represents the classic "Lorentzian lineshape" function, the real part of which describes the wave speed while the imaginary part the loss or gain of the medium. The characteristic width of this lineshape is typically very small compared to the atomic transition frequency which allows us to use the resonance approximation and write (18) as

$$\tilde{\chi}(\omega) = \chi' + i\chi'' \approx -\chi_0'' \left[\frac{\Delta x}{1 + \Delta x^2} + \text{i} \frac{1}{1 + \Delta x^2} \right] \tag{19}$$

where

$$\chi_0'' \equiv \frac{Ne^2}{m\omega_{\text{a}} \epsilon \Delta\omega_{\text{a}}}, \quad \Delta x \equiv 2 \frac{\omega - \omega_{\text{a}}}{\Delta\omega_{\text{a}}} \tag{20}$$

is the normalized detuning from line center and

$$\Delta\omega_{\text{a}} \equiv \gamma + 2/T_2 \tag{21}$$

is the linewidth measured at the full-width, half maximum point (FWHM). The two components of the linewidth are the energy decay rate, γ, and the inverse of the dephasing time T_2. In most laser systems the linewidth is dominated by dephasing.

When the polarization field (16) is inserted into the wave equation (7) the solution is a monochromatic wave that propagates according to

$$\mathcal{E}(z, t) = \mathcal{E}_{\text{o}} \text{e}^{\text{i}\omega t - \Gamma(\omega) z} \tag{22}$$

in one dimension, where

$$\Gamma(\omega) \equiv \text{i}\beta \sqrt{1 + \tilde{\chi}_{\text{at}} - \text{i}\sigma/\omega\epsilon_{\text{host}}} \quad \text{and} \quad \beta \equiv \omega \sqrt{\mu_{\text{o}} \epsilon_{\text{host}}}. \tag{23}$$

Now it is straightforward to see how the complex susceptibility influences wave propagation. It is often convenient to simplify the above expression by first noting that in most materials, $|\tilde{\chi}_{\text{at}}|$ and σ are both very small compared to unity and thus we can expand the square root to first order. Also, we can group terms according to whether they are real or imaginary and label them in the usual manner of wave problems where α is used to denote gain or loss and β is used to denote phase shift. Thus,

$$\Gamma(\omega) = \text{i}(\beta + \Delta\beta_m(\omega)) - \alpha_m(\omega) + \alpha_0 \tag{24}$$

where

$$\Delta\beta_m(\omega) = \beta\chi'(\omega)/2 \quad \text{atomic contribution to phase shift} \tag{25}$$

$$\alpha_m(\omega) = \beta\chi''(\omega)/2 \quad \text{atomic contribution to gain or loss} \tag{26}$$

$$\alpha_0 = \sigma/2\epsilon_{\text{host}} c \quad \text{ohmic losses}. \tag{27}$$

Substituting these into the basic wave propagation solution (22) we have

$$\mathcal{E}(z, t) = \mathcal{E}_{\text{o}} \exp\Big([\alpha_m(\omega) - \alpha_0] z + \text{i}\omega t - \text{i}[\beta + \Delta\beta_m(\omega)] z \Big). \tag{28}$$

The net gain per-unit-distance is given by the real term in the exponent, $\alpha_m(\omega) - \alpha_0$. The phase shift per-unit-distance gives us the phase velocity, v_{ph}, in the medium through the usual ratio

$$\frac{\omega}{\beta + \Delta\beta_m(\omega)} = \frac{\omega}{\beta\,(1 + \chi'(\omega)/2)} = \frac{1}{\sqrt{\mu_{\text{o}}\epsilon_{\text{host}}}\,(1 + \chi'(\omega)/2)} = v_{\text{ph}}. \tag{29}$$

This expression can be easily separated into the free space velocity of light, c, the effects of the host medium and the effects of the laser atoms;

$$v_{\text{ph}} = \frac{1}{\sqrt{\mu_{\text{o}}\epsilon_0 \frac{\epsilon_{\text{host}}}{\epsilon_0}} \left(1 + \frac{\chi'(\omega)}{2}\right)} = \frac{c}{\sqrt{\frac{\epsilon_{\text{host}}}{\epsilon_0}} \left(1 + \frac{\chi'(\omega)}{2}\right)} = \frac{c}{n(\omega)}. \tag{30}$$

The index of refraction is thus

$$n(\omega) = \sqrt{\frac{\epsilon_{\text{host}}}{\epsilon_0}} \left(1 + \frac{\chi'(\omega)}{2}\right). \tag{31}$$

Everything derived to this point has been from a strictly classical point of view and does not yet admit the possibility of gain. To correct this we simply make a few important substitutions [4]. First, comparing the expressions for the midband susceptibility, χ_0'' (20), and the radiative decay rate for the classical electron oscillator, $\gamma_{\text{rad,ceo}}$ (6), we can write

$$\chi_0'' = \frac{3}{4\pi^2}\frac{N\lambda^3\gamma_{\text{rad,ceo}}}{\Delta\omega_{\text{a}}} . \tag{32}$$

The advantage of this form is that it uses only physically measurable quantities. Next, and most important, is to convert this classical expression to a quantum-mechanically correct expression by substituting for the oscillator density N, the *population difference density* $\Delta N \equiv N_1 - N_2$, where N_1 is the population density of the lower energy state and N_2 is the population density of the upper energy state (see Fig. 3). Therefore, (32) becomes

$$\chi_0'' = \frac{3}{4\pi^2}\frac{\Delta N\lambda^3\gamma_{\text{rad,ceo}}}{\Delta\omega_{\text{a}}} . \tag{33}$$

This form of the on-resonance susceptibility demonstrates how gain enters the classical picture of the interaction of radiation with matter. When a mechanism produces an inverted population, ($N_2 > N_1$), the sign of χ_0'' changes and thus so does the imaginary part of the index of refraction (31).

Feedback and Cavity Modes

Any oscillator, be it mechanical, electrical, optical, etc., relies on two conditions to operate; feedback and gain. For a laser, the feedback is accomplished with mirrors, in between which is placed the gain medium. The mirrors constitute an electromagnetic resonator for the optical fields and only certain frequencies, called *eigenmodes*, can exist stably within this resonator. For general resonators enclosing electromagnetic fields, the exact frequencies of the eigenmodes depend on the wavelength of the oscillating field and the geometry and dimensions of the enclosure.

For a laser, in which the sides of the enclosure are open, the exact frequency at which oscillation occurs is determined by satisfying the condition that one round trip of the electromagnetic field inside the resonator reproduces itself exactly, both in phase and amplitude. This requires that the round trip gain exactly cancels the losses and that there are an integer number, n, of cycles of the field encompassing one round trip. Equivalently, the round-trip phase shift is exactly $n2\pi$. Typically, the length of the cavity is such that n is very large at optical frequencies, of the order 10^6. Every integer n that satisfies the $n2\pi$ roundtrip phase condition results in what is called an "axial" or "longitudinal" mode. The frequency spacing between the axial modes is determined by adding (or subtracting) one more cycle of phase to the round trip so that $n \to n \pm 1$. It is easy to show that, for a resonator with two mirrors separated by a distance L, the axial mode spacing is given by

$$\Delta\nu_{\text{ax}} = \frac{c}{2L} . \tag{34}$$

For example, a 1.5 meter-long cavity has an axial mode spacing of 100 MHz. Now, in most cases in the visible and near-infrared portion of the spectrum, the axial mode spacing is smaller than the gain bandwidth of the lasing medium which means that for inhomogeneous media, several modes may oscillate at once. This is typically undesirable since the instantaneous power in the beam fluctuates rapidly due to the beating of the modes. In homogeneously broadened lasers, usually only one mode can oscillate since gain saturation pulls down the gain of all of the atoms uniformly. However, modes may compete for oscillation and this "mode competition" will appear as a fluctuation or instability in the laser output power as different axial modes within the gain curve rapidly rise and fall in power. Many techniques have evolved over the years to force lasers to operate in a single axial mode such as using unidirectional ring cavities, multiple intra-cavity etalons, etc. [9].

All resonators store energy in the cycling fields or mechanical vibrations which they contain. Because of unavoidable loss mechanisms and the need for taking energy out of the resonator, the fields in the resonator have a finite lifetime. The lifetime of the stored energy compared to the period of oscillation defines the quality factor or Q of the resonator;

$$Q \equiv 2\pi \times \frac{\text{ENERGY STORED}}{\text{ENERGY DISSIPATED PER CYCLE}} . \tag{35}$$

There is an intimate relationship between the Q of a resonator and the linewidth of an oscillator utilizing the resonator. What we will see is that the gain element in an oscillator must exactly balance any losses in the resonator so that the Q of the combined resonator plus gain will tend toward infinity. Evidently, this suggests an infinitely narrow linewidth, but, in reality this is prevented by thermal noise in electrical oscillators and quantum noise (spontaneous emission) in optical oscillators.

To gain an appreciation for the elementary properties of a laser resonator and how it contributes to the linewidth-limiting nature of an oscillator below threshold, consider the simple two-mirror laser cavity with gain region shown in Fig. 5. This configuration is called a *regenerative amplifier*. A signal input to the cavity from the

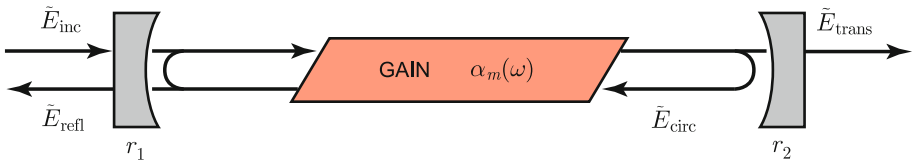

Noise and Stability in Modelocked Soliton Lasers, Figure 5
Schematic diagram of a regenerative laser amplifier

left hand side, E_{inc}, passes through a partially transmitting mirror with amplitude reflectivity r_1. Mirror reflectivity is defined in terms of a field E_{i} incident upon the mirror and the subsequent field E_{r} reflected by the mirror, $E_{\text{r}} = r * E_{\text{i}}$. The mirror amplitude transmission, t, is defined in terms of the transmitted field E_{t} and an incident field according to $E_{\text{t}} = \text{i}t * E_{\text{i}}$ [4]. For lossless mirrors, power conservation requires $r^2 + t^2 = 1$ and we define the power transmission and reflection coefficients as $T = t^2$ and $R = r^2$, respectively.

The wave experiences amplification and phase shift in the gain medium along with losses due to scattering, absorption and transmission through the mirrors. We can follow the wave around one complete round trip of the cavity (length $= p$) including a round trip through the gain (length $= p_m$) and find that

$$E_1 = E_0 \cdot r_1 r_2 \cdot \exp\left[\alpha_m(\omega)p_m - \alpha_0 p - \text{i}\omega p/c - \text{i}\Delta\beta_m(\omega)p_m\right] \tag{36}$$

where E_0 is the field just after passing through the first mirror and E_1 is the field after one round trip. $\alpha_m(\omega)$ is the frequency-dependent atomic gain (26), α_0 represents the losses due to scattering and background absorption (27) and $\Delta\beta_m$ is the phase shift due to the atomic medium (25).

From (36) we can define a complex round trip gain

$$\tilde{g}_{\text{rt}}(\omega) = \frac{E_1}{E_0} = r_1 r_2 \cdot \exp\left[\alpha_m(\omega)p_m - \alpha_0 p - \text{i}\omega p/c - \text{i}\Delta\beta_m(\omega)p_m\right] . \tag{37}$$

At the second mirror, some of the field leaks out of the cavity and we can consider the overall or net gain of the system as $\tilde{E}_{\text{trans}}/\tilde{E}_{\text{inc}}$. However, since some of the signal is reflected back into the cavity and gain medium, a circulating field is built up which can (and usually does) increase far beyond the amplitude of the incident field. This is the notion of *regenerative gain* and the remarkable phenomenon of line-narrowing occurs in this scenario. It is easy to understand qualitatively if we re-examine the expression for Q (35). The losses within the cavity, including mirror transmission, degrade the Q and therefore result in a broad resonance curve or transmission peak. As the atomic gain, $\alpha_m(\omega)$, is turned up it begins to compensate for the losses and this improves the Q and narrows the lineshape or transmission resonance of the composite system. If we follow a wave around the cavity and demand self-consistent fields at all points (i. e. steady-state conditions), we find that the overall amplitude gain of the regenerative amplifier is given by

$$\frac{\tilde{E}_{\text{trans}}}{\tilde{E}_{\text{inc}}} = -\frac{t_1 t_2}{\sqrt{r_1 r_2}}\left(\frac{\sqrt{\tilde{g}_{\text{rt}}(\omega)}}{1 - \tilde{g}_{\text{rt}}(\omega)}\right) . \tag{38}$$

Both the overall gain and the line-narrowing really occur due to the denominator in (38). As the round-trip gain approaches unity at a cavity resonance ($2n\pi$ round-trip phase condition), the denominator is quickly driven toward zero and thus the gain increases dramatically. The closer the magnitude of $\tilde{g}_{\text{rt}}(\omega)$ is to unity the more quickly this happens as we tune to a cavity resonance, thus producing a narrower response. We can see this effect in Fig. 6 where the overall regenerative power gain, $|\tilde{E}_{\text{trans}}/\tilde{E}_{\text{inc}}|^2$, is plotted as a function of frequency for two cases of mirror reflectivies and different values of midband atomic gain $2\alpha_{\text{m}_\text{o}} p_m$. Details of the parameters and quantitative features of the curves are presented in Table 1.

Figure 6a shows the regenerative power gain for mirror reflectivities $R_1 = R_2 = 30\%$. The axial mode spacing is chosen, arbitrarily, to be 1/10th of the Lorentzian linewidth $\Delta\omega_{\text{a}}$. Notice the qualitative changes in the transmission peaks as the midband atomic gain is varied from $2\alpha_{\text{m}_\text{o}} p_m = 0$ to $2\alpha_{\text{m}_\text{o}} p_m = 0.95$. There is a clear trend toward line narrowing. The regenerative amplifier results in an amplifying filter which suppresses noise or unwanted radiation except at the axial mode frequencies. As the gain is increased, even by a small amount, the whole system becomes ever more sensitive to detuning from the axial mode peaks. If we change the mirrors from a reflectivity of $R = 30\%$ to $R = 60\%$, a dramatic line-narrowing effect takes place, as shown in Fig. 6b. Not only the filtering effect improved but the overall regenerative gain has increased

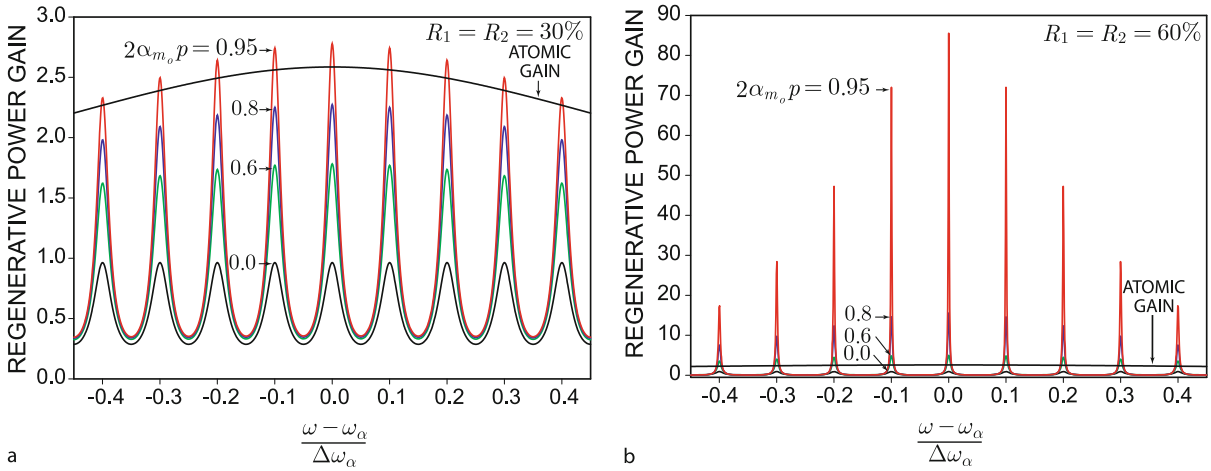

Noise and Stability in Modelocked Soliton Lasers, Figure 6

Power gain of a regenerative laser amplifier as a function of detuning from the atomic line center. a Mirror reflectivities $R_1 = R_2 =$ 30%. Regenerative gain shown for 4 values of midband atomic gain. Broad "ATOMIC GAIN" curve corresponds to midband gain $2\alpha_{m_o}p = 0.95$. Internal round-trip loss coefficient $2\alpha_o p = 0.4$. b Mirror reflectivities $R_1 = R_2 =$ 60%. Other parameters identical to a. (Adapted from [4])

Noise and Stability in Modelocked Soliton Lasers, Table 1

Parameters for regenerative laser amplifier with power gains plotted in Fig. 6. Additional assumptions, $\lambda = 800$ nm, cavity length $L = 1.5$ m, axial mode spacing $\Delta\nu_{ax} = 100$ MHz

MIDBAND GAIN COEFF. $2\alpha_{m_o}p_m$	LOSS COEFF. $2\alpha_o p$	MIDBAND PWR GAIN $\exp\left[2\alpha_{m_o}p_m\right]$	$R_{1,2}$	ROUND TRIP POWER GAIN $\lvert\tilde{g}_{rt}\rvert^2$	REGEN. POWER GAIN	CAVITY LINE-WIDTH $\Delta\nu_{cav}$	FINESSE $\mathcal{F}$	Q
0	0.04	1	30%	0.087	0.964	45.09 MHz	2.413	8.311×10^6
			60%	0.346	0.925	17.30 MHz	5.849	2.167×10^7
0.6	0.04	1.822	30%	0.158	1.783	31.75 MHz	3.282	1.180×10^7
			60%	0.630	4.983	7.375 MHz	13.58	5.081×10^7
0.8	0.04	2.226	30%	0.192	2.274	27.84 MHz	3.707	1.346×10^7
			60%	0.770	15.560	4.167 MHz	24.00	8.992×10^7
0.95	0.04	2.586	30%	0.224	2.779	25.03 MHz	4.098	1.498×10^7
			60%	0.894	85.54	1.776 MHz	56.27	2.110×10^8

from a peak value of 2.8 to a peak of 86 (note the change in the vertical scale).

There are two conventional metrics for describing the relative linewidth of a resonator. In the case of a laser resonator where the wavelength is usually much smaller than the physical cavity dimensions, the *finesse*, $\mathcal{F}$, is the ratio of the axial mode spacing to the cavity linewidth

$$\mathcal{F} \equiv \frac{\Delta\omega_{ax}}{\Delta\omega_{cav}} = \frac{\Delta\nu_{ax}}{\Delta\nu_{cav}} . \tag{39}$$

This is the commonly used figure of merit for Fabry–Perot etalons [22]. The general concept of Q described earlier (35) can also be used to relate the cavity linewidth to the *center frequency* [4] according to

$$Q \equiv \frac{\omega_a}{\Delta\omega_{cav}} = \frac{\nu_a}{\Delta\nu_{cav}} . \tag{40}$$

From Table 1 we see that both the finesse and the Q improve by a factor of ≈ 14 when the mirror reflectivity is increased from 30% to 60% for a constant round-trip power gain of $\exp\left[2\alpha_{m_o}p_m\right] = 2.586$. This dramatic and nonlinear relationship between round-trip gain/loss and overall amplifier behavior hints at what happens as the round-trip gain approaches and just compensates for all of the round-trip losses. At that point, the regenerative gain goes to infinity, the linewidth goes to zero (in this model) and no light input is required. In fact, this is just the condition for the onset of laser oscillation. With no input applied to the regenerative amplifier, a single spontaneously emitted photon which happens to propagate exactly down the resonator axis is reflected (with probability equal to the mirror reflectivity) back to the gain medium, is amplified and returns again via the other mirror. This sets off the

avalanche of field buildup until the gain medium "saturates" its gain down to the point where the net round-trip gain exactly balances the round-trip loss. (Note: the saturation phenomenon occurs when the the optical field begins to deplete the population inversion faster than it can be maintained by the pumping rate). This is an important and fascinating component of lasers (indeed of all oscillator physics) but will be left out of the present discussion in the interest of brevity. The reader may consult any of the standard textbooks on laser physics or optical spectroscopy for a comprehensive treatment, e. g. [4,9,10].

Laser Dynamics

Now that we have developed the basic concepts behind laser gain and cavity resonance, which together set the stage for optical amplification and finally oscillation, we must address the time dependence of the various coupled quantities. This is the domain of *laser dynamics* and in it we find answers to questions such as "how does a laser evolve from a cold cavity to oscillation upon being switched on?" and "what happens to an otherwise stable laser oscillator if it is subject to a sudden perturbation?" As pointed out in Sect. "Introduction", these phenomena fall under the general heading of "large scale instabilities" and the reader can find more complete treatments in, e. g., [5,6,12,13,14,16,23]. For our purposes here, we will gather up the results of the analysis and apply it to a few of the myriad interesting (in)stability problems.

Reduction of the Full Laser Equations

The major physical variables of a laser oscillator are the electric field $E(t)$, the atomic polarization (i. e. dipole moment per-unit-volume) $P(t)$ and the energy level populations $N_n(t)$. Specifically, it is the *population difference* $\Delta N = N_1(t) - N_2(t)$ that results in the gain or loss in the laser. The physics that couples the three variables is rather involved [4,5,6,12,16] and we will present only the main results here.

Upon solving Maxwell's and Schrödinger's equations for the motion of the electromagnetic fields and the atomic states of the atoms in the laser cavity, one obtains the following system of coupled nonlinear differential equations in "neoclassical" form [4],

$$\frac{\mathrm{d}^2E(t)}{\mathrm{d}t^2} + \gamma_\mathrm{c}\frac{\mathrm{d}E(t)}{\mathrm{d}t} + \omega_\mathrm{c}^2E(t) = -\frac{1}{\epsilon}\frac{\mathrm{d}^2P(t)}{\mathrm{d}t^2} \tag{41}$$

$$\frac{\mathrm{d}^2P(t)}{\mathrm{d}t^2} + \Delta\omega_\mathrm{a}\frac{\mathrm{d}P(t)}{\mathrm{d}t}\omega_\mathrm{a}^2P(t) = \kappa\Delta\mathcal{N}(t)E(t) \tag{42}$$

$$\frac{\mathrm{d}\Delta\mathcal{N}(t)}{\mathrm{d}t} + \frac{\Delta\mathcal{N}(t) - \Delta\mathcal{N}_0}{T_1} = -\frac{2^*}{\hbar\omega}E(t)\frac{\mathrm{d}P(t)}{\mathrm{d}t} \tag{43}$$

where $\Delta\mathcal{N}(t)$ is the atomic density difference averaged over the cavity volume V_c

$$\Delta\mathcal{N}(t) \equiv \frac{1}{V_\mathrm{c}}\iiint \Delta N(\vec{\mathbf{r}}, t)\,\mathrm{d}\vec{\mathbf{r}}\,, \tag{44}$$

$\Delta\mathcal{N}_0$ is the equilibrium population density difference in the absence of driving fields, γ_c is the cavity field energy decay rate which takes into account losses as well as mirror output coupling, etc., ω_c is the cavity resonant frequency, T_1 is the population-recovery or energy-relaxation time from level $E_2 \to E_1$ (see p. 205 of [4] for an excellent discussion of this concept), 2^* is the "bottlenecking factor" (that is, if the atoms get stuck emptying out of the lower energy level, $2^* = 2$, if not, $2^* = 1$, which constitutes a "good" laser) and, finally,

$$\kappa \equiv \frac{3^*\epsilon\lambda^3\omega_\mathrm{a}\gamma_\mathrm{rad}}{4\pi^2}\eta \tag{45}$$

is the coupling constant linking the polarization to the product of the population difference and the electric field (i. e. the stimulated emission). Here η is a "filling factor" which expresses the fraction of the cavity mode volume filled by the active laser atoms [4] and the 3^* describes the tensor coupling between the fields and the polarization ($0 \le 3^* \le 3$).

This rather formidable-looking set of coupled Eqs. (41)–(43) can describe almost all dynamic laser phenomena but certain approximations can be made which allows us to simplify them without compromising rigor. The first approximation is based on the fact that both of the field amplitudes of interest will change slowly on a time scale compared to the optical carrier frequency. The field amplitudes in "phasor form" separate the slowly-varying complex envelope from the time-harmonic carrier,

$$E(t) = \frac{1}{2}\left[\tilde{E}(t)\mathrm{e}^{\mathrm{i}\omega t} + \tilde{E}^*(t)\mathrm{e}^{-\mathrm{i}\omega t}\right] \tag{46}$$

$$P(t) = \frac{1}{2}\left[\tilde{P}(t)\mathrm{e}^{\mathrm{i}\omega t} + \tilde{P}^*(t)\mathrm{e}^{-\mathrm{i}\omega t}\right]\,. \tag{47}$$

If we substitute these expressions into (41)–(43) and assume that the time variations of the phasor amplitudes along with the cavity decay rate γ_c and atomic linewidth $\Delta\omega_\mathrm{a}$ are all slow compared to the carrier frequency ω, the atomic line center frequency ω_a and the cavity resonance frequency ω_c, the general Eqs. (41)–(43) become

$$\frac{\mathrm{d}\tilde{E}(t)}{\mathrm{d}t} + \left[\gamma_\mathrm{c}/2 + \mathrm{i}(\omega - \omega_\mathrm{c})\right]\tilde{E}(t) = -\mathrm{i}\frac{\omega}{2\epsilon}\tilde{P}(t) \tag{48}$$

$$\frac{\mathrm{d}\tilde{P}(t)}{\mathrm{d}t} + \left[\Delta\omega_\mathrm{a}/2 + \mathrm{i}(\omega - \omega_\mathrm{a})\right]\tilde{P}(t) = -\mathrm{i}\frac{\kappa}{2\omega}\Delta\mathcal{N}(t)\tilde{E}(t) \tag{49}$$

$$\frac{\mathrm{d}\Delta\mathcal{N}(t)}{\mathrm{d}t} + \frac{\Delta\mathcal{N}(t) - \Delta\mathcal{N}_0}{T_1} = \mathrm{i}\frac{2^*}{4\hbar}\left[\tilde{E}(t)\tilde{P}^*(t) - \tilde{E}^*(t)\tilde{P}(t)\right] . \tag{50}$$

The steps leading to the above equations are commonly known as the *slowly-varying envelope approximation* (SVEA) since we have replaced equations describing the dynamics of instantaneous quantities with equations describing the evolution of the much more slowly-varying envelope functions.

Coupled Cavity-Atom Rate Equations

At this point, we could integrate the coupled Eqs. (48)–(50), but there are several other simplifications we can take that help lend considerable physical insight into laser dynamics and yet preserve the essential physics.

(Note: An early description of the dynamics of the photon and population densities in a laser was given by Dunsmuir [24]. The description is in terms of rate equations which can be derived from the full set of coupled equations above and understood from elementary energy-balance considerations.)

Electric Field in Terms of Photons The first step is to write the electric field in terms of quanta of electromagnetic energy, or photons. The number of photons in the cavity is given by the total electromagnetic energy divided by $\hbar\omega$, the energy per photon, or

$$n(t) = \frac{\epsilon V_\mathrm{c}}{2\hbar\omega}\left|\tilde{E}(t)\right|^2 = \frac{\epsilon V_\mathrm{c}}{2\hbar\omega}\tilde{E}(t)\tilde{E}^*(t) . \tag{51}$$

Differentiating this and substituting from (48) for $\mathrm{d}\tilde{E}(t)/\mathrm{d}t$ and $\mathrm{d}\tilde{E}^*(t)/\mathrm{d}t$ results in

$$\frac{\mathrm{d}n(t)}{\mathrm{d}t} + \gamma_\mathrm{c} n(t) = \mathrm{i}\frac{V_\mathrm{c}}{4\hbar}\left[\tilde{E}(t)\tilde{P}^*(t) - \tilde{E}^*(t)\tilde{P}(t)\right] . \tag{52}$$

Population Density to Total Population In order to make the cavity-averaged population density equation have a similar driving term as the photon equation, we multiply the distributed density by the cavity mode volume to get the total population difference; $\Delta N(t) = V_\mathrm{c}\Delta\mathcal{N}(t)$. Applying this to (50) we have

$$\frac{\mathrm{d}\Delta N(t)}{\mathrm{d}t} + \frac{\Delta N(t) - \Delta N_0}{T_1} = -\mathrm{i}\frac{2^* V_\mathrm{c}}{4\hbar}\left[\tilde{E}(t)\tilde{P}^*(t) - \tilde{E}^*(t)\tilde{P}(t)\right] . \tag{53}$$

Adiabatic Elimination of the Polarization Field In many cases of interest, the dephasing of the polarization field $P(t)$ happens on a time scale that is much faster than the atomic energy decay time γ^{-1} and, therefore, the polarization field's "phase memory" is continuously interrupted making it possible for the envelope to follow fluctuations in the driving source term $\Delta\mathcal{N}(t)\tilde{E}(t)$ in (49). This allows us to establish a linear relationship between the phasor amplitudes of the electric field and the polarization with essentially no time delay. We consider the polarization to adiabatically follow the electric field and thus can solve the SVEA equation (49) in the steady state ($\mathrm{d}\tilde{P}(t)/\mathrm{d}t = 0$) to obtain

$$\tilde{P}(t) \approx \frac{-\mathrm{i}\kappa}{\omega\Delta\omega_\mathrm{a}}\frac{1}{1 + 2\mathrm{i}(\omega - \omega_\mathrm{a})/\Delta\omega_\mathrm{a}}\Delta\mathcal{N}(t)\tilde{E}(t) . \tag{54}$$

This is frequently called the "linear susceptibility" or "rate equation" approximation. We can now use this result on the right-hand sides of (52) and (53),

$$\mathrm{i}\frac{V_\mathrm{c}}{4\hbar}\left[\tilde{E}(t)\tilde{P}^*(t) - \tilde{E}^*(t)\tilde{P}(t)\right] = -\frac{\kappa}{2\hbar\omega\Delta\omega_\mathrm{a}} \times \frac{\Delta N(t)}{1 + \left[2(\omega - \omega_\mathrm{a})/\Delta\omega_\mathrm{a}\right]^2}\left|\tilde{E}(t)\right|^2 . \tag{55}$$

Substituting in for the coupling constant κ (45) and the electric field in terms of the photon number (52), we can write

$$\mathrm{i}\frac{V_\mathrm{c}}{4\hbar}\left[\tilde{E}(t)\tilde{P}^*(t) - \tilde{E}^*(t)\tilde{P}(t)\right] = -\frac{3^*\omega_\mathrm{a}\gamma_\mathrm{rad}\eta\lambda^3}{4\pi^2\Delta\omega_\mathrm{a}V_\mathrm{c}} \times \frac{1}{1 + \left[2(\omega - \omega_\mathrm{a})/\Delta\omega_\mathrm{a}\right]^2}\Delta N(t)n(t) \tag{56}$$

$$= -K\Delta N(t)n(t) \tag{57}$$

where we have defined a new "rate-equation coupling constant" K. Substituting (57) into (52) and (53), we have

$$\frac{\mathrm{d}n(t)}{\mathrm{d}t} + \gamma_\mathrm{c} n(t) = -K\Delta N(t)n(t) \tag{58}$$

and

$$\frac{\mathrm{d}\Delta N(t)}{\mathrm{d}t} + \frac{\Delta N(t) - \Delta N_0}{T_1} = -2^* K\Delta N(t)n(t) . \tag{59}$$

These equations describe the balance between population levels and photon number within the cavity for a two-level system. In order to predict dynamic laser phenomena, especially large transients, we need to add a couple of

more terms. First of all, most lasers rely on a four-level energy scheme (see Fig. 3). The lowest level, E_0 acts as a reservoir of atoms while the top level, E_3 is a momentary stop on the way to the metastable level E_2. In a typical 4-level "good" laser, pumping from the ground state E_0 to level E_3 quickly relaxes into level E_2 which is long-lived and allows for population inversion with respect to level E_1. Level E_1 will also empty quickly into the ground-state level and be ready for pumping back up into E_2 via E_3. With this in mind, it is a good approximation to set the level populations $N_3 = N_1 = 0$.

When a laser is actually oscillating, there must be a population inversion so that $\Delta N(t) = N_1(t) - N_2(t) < 0$. We can recast our equations in terms of this inversion by simply defining

$$N(t) \equiv N_2(t) - N_1(t) = -\Delta N(t) \,. \tag{60}$$

Next, to account for the effect of pumping atoms from the lower laser level to the upper level to create the inversion, we will also add a term $R_P(t)$ to (59). Also, for lasers operating in the visible and near-infrared portion of the spectrum the Boltzmann distribution governing the population levels in thermal equilibrium (i. e. with no electromagnetic signals driving the transitions) gives

$$\frac{N_2}{N_1} = \exp\left(-\frac{\hbar\omega_a}{kT}\right) \approx 0 \tag{61}$$

and thus $\Delta N_0 = N_{10} - N_{20} \approx N_{10}$. In a "good" four-level laser, there will be very fast relaxation out of the lower laser level always and therefore, to a good approximation $N_{10} = 0$ and so $\Delta N_0 = 0$. Finally, since the lower level is assumed to empty very fast, there is no bottlenecking and $2^* = 1$. We can now rewrite (58) and (59) in simplified form,

$$\frac{dn(t)}{dt} = KN(t)\left[n(t) + n_{sp}\right] - \gamma_c n(t) \tag{62}$$

$$\frac{dN(t)}{dt} = R_P(t) - \gamma_2 N(t) - KN(t)n(t) \,. \tag{63}$$

Notice that we have included one more term, n_{sp}, to account for photons spontaneously emitted into the desired cavity mode. This term is essential for getting the laser "started", but once started, the total number of photons in the cavity mode, $n(t)$, far exceeds n_{sp}, as we shall see.

The coupled rate equations (62), (63) form an intuitive and accurate platform for understanding the dynamics between the pump source, the population inversion and the photons in a laser in the approximation that these processes occur on a time scale that is long compared to a cycle of the optical field. This is a reasonable assumption in almost all cases of interest.

A physical picture that is described by (62), (63) can be constructed by realizing that these equations form an energy storage and dissipation system. Energy is pumped into the atoms via R_p and released into the electromagnetic field by stimulated emission (KNn). As the field builds up in the resonator, more energy is pulled out of the atoms and released through the output coupling mirror ($-\gamma_c n$). Soon the atoms are somewhat depleted ($-KNn$) and must "recharge" from the pump source R_p. The energy stored in the fields is reduced while the pump source increases the energy stored in the atoms. This "dance" between the energy stored in the atoms and the energy stored in the field is at the heart of the elementary "relaxation oscillations" and is also the first important dynamic laser phenomenon. Unfortunately, since the coupled rate equations (62), (63) are nonlinear, they cannot be solved in closed form. So, to get a more quantitative feeling for how the relevant quantities behave, let us first look at a couple of limiting cases.

Steady-State Response For a constant pumping rate $R_P(t) = R_{P_o}$, the laser will achieve a constant or steady-state operating point such that

$$\frac{dn(t)}{dt} = \frac{dN(t)}{dt} = 0 \,. \tag{64}$$

Applying this to (62), (63) results in the following pair of equations

$$0 = KNn - \gamma_c n \tag{65}$$

$$0 = R_{P_o} - KNn - \gamma_2 N \,. \tag{66}$$

Notice that we have removed the time dependence and also the spontaneous emission term. Its role only becomes significant during the startup of dynamic processes.

There are two solutions to the system above. We see by inspection of (65) that either $n = 0$ or $N = \gamma_c/K$. Let's investigate both cases.

Case 1: $n = 0$ Applying this condition to Eq. (66) yields

$$N = R_{P_o}/\gamma_2 \,. \tag{67}$$

The population inversion varies directly with the pumping rate. Since there are no photons generated, we call this the "below threshold" state.

Case 2: $N = \gamma_c/K$ Again we apply this condition to Eq. (66) and find

$$n = \frac{R_{P_o}}{\gamma_c} - \frac{\gamma_2}{K} = \left(\frac{R_{P_o}K}{\gamma_2\gamma_c} - 1\right)\frac{\gamma_2}{K} \,. \tag{68}$$

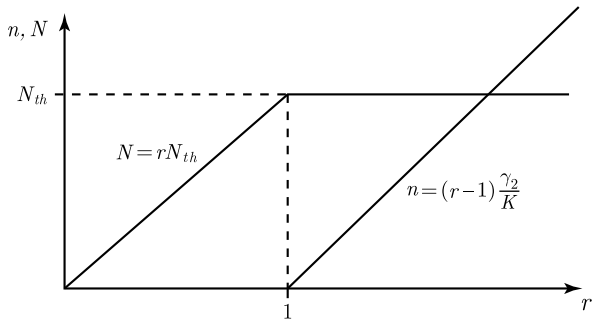

Noise and Stability in Modelocked Soliton Lasers, Figure 7
Photon number n and population inversion N as a function of the normalized pumping rate $r = R_{P_o}/\gamma_2 N_{th}$. At the first threshold, $r = 1$, ($R_{P_o} = \gamma_2 N_{th}$), the population inversion is clamped at $N = N_{th}$ and the photon number starts to rise linearly

We have deliberately written the result in this form to make a point. In *Case 2*, the population inversion is constant and is said to be "clamped" at the threshold level for lasing. We will call this the "threshold inversion density"

$$N(t) = N_{th} = \frac{\gamma_c}{K} . \tag{69}$$

The pumping rate which just achieves threshold inversion, and for which the photon number n begins to rise from zero, will be called the "first threshold" pumping rate; $R_{P_{01}} = \gamma_2 N_{th}$. Above this point, the photon density is linearly proportional to the pumping rate

$$n = n_{ss} = \left(\frac{R_{P_0}}{\gamma_2 N_{th}} - 1\right)\frac{\gamma_2}{K} = (r-1)\frac{\gamma_2}{K} . \tag{70}$$

This is the steady-state photon density and we have defined a "normalized pumping rate",

$$r \equiv R_{P_o}/\gamma_2 N_{th} , \tag{71}$$

valid above the first threshold. Inserting typical numbers for a Ti:sapphire laser; $\gamma_c = 1.280 \times 10^7\ s^{-1}$, $\gamma_2 = 3.125 \times 10^5\ s^{-1}$ and $K = 6.860 \times 10^{-17} m^3 s^{-1}$, we find

$$N_{th} = 1.866 \times 10^{23}\ m^{-3}, \quad n_{ss} = 4.555 \times 10^{21}\ m^{-3} \tag{72}$$

when pumped at twice threshold ($r = 2$).

Step Response with no Spontaneous Emission It is interesting to study the response of the system assuming no spontaneous emission into the cavity mode ($n_{sp} = 0$). In this case, there are no photons to seed the stimulated emission process and therefore $n(t) = 0$ for all time. The population inversion, however, is immediately driven by a step function pumping rate $R_P(t) = R_{P_o} H(t)$ where $H(t)$ is the Heaviside step function, defined according to

$$H(t) = \begin{cases} 0, & t \le 0 \\ 1, & t > 0 . \end{cases} \tag{73}$$

With no photons to build up and stimulate downward transitions, Eq. (63) becomes

$$\frac{dN_o(t)}{dt} = R_P(t) - \gamma_2 N_o(t) \tag{74}$$

with solution

$$N_o(t) = \frac{R_{P_o}}{\gamma_2}\left[1 - e^{-\gamma_2 t}\right], \quad t > 0 . \tag{75}$$

This tells us that the population inversion density rockets upward with a time constant γ_2. In the asymptotic limit, we find $N_o(t \to \infty) = R_{P_o}/\gamma_2$. Substituting from (70) we find that (75) can be written in the especially simple form

$$N_o(t) = rN_{th}\left[1 - e^{-\gamma_2 t}\right], \quad t > 0 . \tag{76}$$

This says that in the absence of spontaneous emission photons to seed the stimulated growth process, the population inversion will grow smoothly to r times the threshold inversion obtained under actual lasing conditions.

Full Step Response with Stimulated Emission Consider now the solutions to the full coupled nonlinear rate equations (62), (63). These have been computed numerically for a typical 1.5 meter-long Ti:sapphire laser cavity

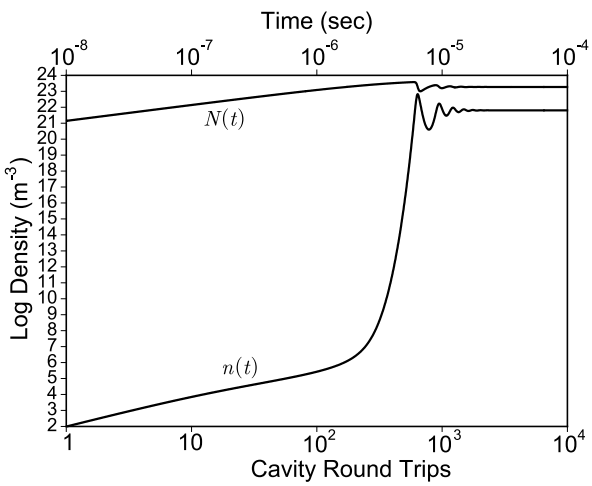

Noise and Stability in Modelocked Soliton Lasers, Figure 8
Laser photon density $n(t)$ and population inversion $N(t)$ for a step function pump $R_P(t)$. The normalized pumping rate $r = 2.0$

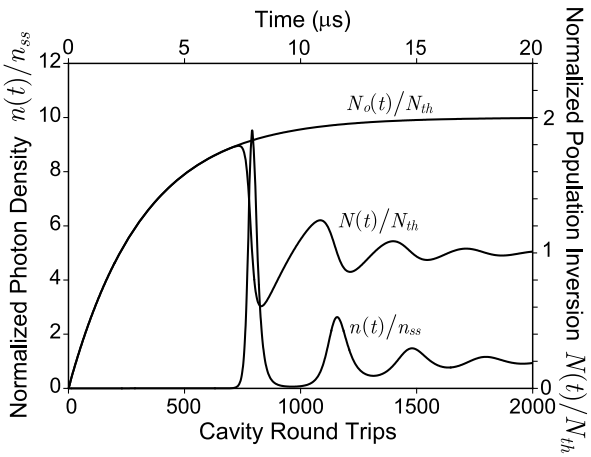

Noise and Stability in Modelocked Soliton Lasers, Figure 9
Laser photon density $n(t)$ and population inversion $N(t)$ normalized to steady-state values for a step function pump $R_p(t)$. The normalized pumping rate $r = 2.0$

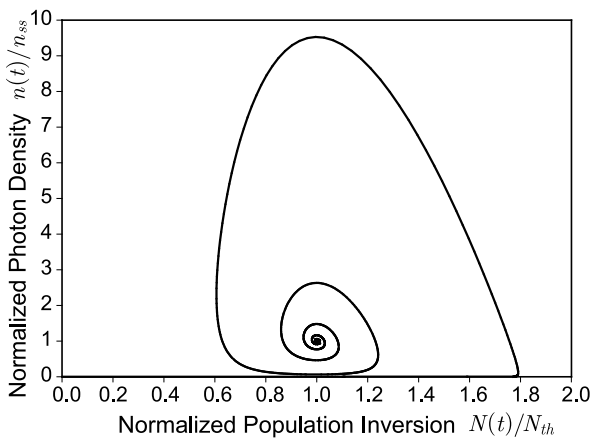

Noise and Stability in Modelocked Soliton Lasers, Figure 10
Laser photon density $n(t)$ and population inversion $N(t)$ for a step function pump $R_p(t)$ displayed in phase space

and appear in Fig. 8 on a log-log scale. The spontaneous emission term, n_{sp}, of one photon-per-cavity-mode volume, stimulates growth in the cavity photon density, $n(t)$ which climbs at a rate of about an order of magnitude per decade of round trips until $n \approx n_{sp}$ and the photon density rapidly climbs by 15 orders of magnitude until the inversion density starts to get "eaten up" by the large number of photons. This can be seen more clearly if we look at a plot of the dynamics on a linear scale as in Fig. 9. At about 700 cavity round trips, the photon density starts to eat up the inversion and thus the gain. The inversion rapidly falls below threshold, N_{th} so that there is more net loss than gain in the system. This causes the photon number to fall rapidly below the steady-state value n_{ss} which in turn allows the inversion to climb back up. When the inversion rises above the threshold value, net gain is restored to the system and the photon number starts to climb again. This behavior, including the roughly 90° phase lag between photon density and population inversion, is characteristic of the standard "predator-prey" relations of the Lotka–Volterra model [25]. The difference here is that the oscillations of the dependent variables die away with a time constant $\tau \approx \gamma_2^{-1}$.

The top curve in Fig. 9 shows the response of the population inversion density, $N_o(t)$, in the absence of the spontaneous emission factor n_{sp}. This is the solution presented in (76). There is no accompanying photon growth to eat away at the population inversion and therefore there are no oscillations and the population grows asymptotically to rN_{th}. In the case depicted here, $r = 2.0$.

A common format for viewing the dynamic behavior of two variables coupled by nonlinear differential equations with an independent variable (or parameter) such as time, is the phase plane [26]. In this format, one variable is plotted against the other with time as a parameter. Thus curves or trajectories in the phase plane represent the evolution of the system. The solution to the laser startup problem shown in Fig. 9 is plotted in the phase plane portrait of Fig. 10. Both the photon density and the population eventually settle into their steady-state values which appears as a trajectory circling around and finally settling into a "focal point". The phase plane is a highly useful device for exhibiting a wide range of nonlinear behavior and can often reveal subtle features of a nonlinear system not easily observed when plotting either of the dynamical variables alone as a function of time.

Laser Instability, Chaos and the Lorenz Equations

The general linear trend of output power versus pumping rate shown in Fig. 7 and even the transient responses depicted in Figs. 8–10 represent solutions to well-behaved "stable" equations. Gentle changes in any of the system parameters will result in gentle changes in output power once the system's transient response has died out. In the laboratory, however, wild and chaotic behavior occasionally occurs in a laser system and although rather undesirable, analysis to try and explain those aberrant phenomena was slow in developing. One of the early papers on the subject was by Kaplan and Zier [27] who derived a set of equations for the steady-state behavior of a 3-level "optical maser" and compared linearized solutions against data from Bostick and O'Connor [28]. Tang questioned the validity of the rate equation approach that was commonly used at the time for optical masers and discovered that paying close

attention to the relevant time constants (for energy decay, polarization dephasing and cavity lifetime), could admit terms into more complete equations that would give rise to undamped transient solutions [29]. In general, these treatments only revealed broad trends of instability without producing a very coherent picture of the whole landscape of the phenomena. It would not be until years after the birth of the laser that results from a completely unrelated field were recognized as being important for laser physics in the context of shedding new and more comprehensive light on the subject.

In 1963, Edward N. Lorenz published a seminal paper on the topic of hydrodynamic flow in which coupled nonlinear differential equations were developed to predict the flow in a forced dissipative system [30]. He discovered that the system would admit nonperiodic solutions which had very sensitive dependence on their initial conditions and often would yield highly irregular results. This work set the seeds for the explosive growth of the field of chaos science years later [31,32].

The equations formulated by Lorenz describe the flow of fluid in a Bénard cell which is a rectangular tank heated on the bottom [32]. At low temperature differences between the top and bottom, a steady-state gradient is produced since the liquid viscosity is high enough to prevent convection. Above a critical temperature, however, the fluid starts to move in counter-rotating rolls (convection) and at still higher temperatures turbulence and chaos sets in. Lorenz's equations, in normalized form, are

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \sigma(y-x)\,, \quad \frac{\mathrm{d}y}{\mathrm{d}t} = x(r-z)-y\,, \quad \frac{\mathrm{d}z}{\mathrm{d}t} = xy-bz\,. \tag{77}$$

For the fluid problem, in these equations x is proportional to the intensity of the convective motion, y is proportional to the temperature difference between the ascending and descending currents and z is proportional to the degree to which the temperature profile deviates from linearity in the vertical direction. The other terms are defined in Table 2. The Prandtl number is the ratio of viscous to thermal diffusivity. R is the Reynolds number and R_c is the Reynolds number for which the flow changes from conduction to convection. k_1 is a dimensionless wavenumber.

Noise and Stability in Modelocked Soliton Lasers, Table 2
Relationships between coefficients of Lorenz equations for hydrodynamic flow and the laser equations in the slowly-varying envelope approximation

Coefficient	Bénard convection	Laser
σ	Prandtl number	$\gamma_c T_2/2 = T_2/2\tau_c = \gamma_c/2\gamma_\perp$
r	R/R_c	$-\frac{\kappa T_2 \tau_c}{2\epsilon}\Delta\mathcal{N}_0$
b	$\frac{4\pi^2}{\pi^2+k_1^2}$	$\gamma T_2 = T_2/T_1 = \gamma_\parallel/\gamma_\perp$

In 1975, Haken [33] pointed out an analogy between the the equations Lorenz had developed and those governing the evolution of the laser electric field, polarization and population inversion in the slowly-varying envelope approximation. He suggested that, based on the close analogy to the Lorenz equations, new instabilities in lasers should be found when operated in special regions of parameter space. This gave birth to new analytical and theoretical studies of laser instabilities, some of which which were previously know, but not well understood. Casperson [34] was also intrigued by the problem of spontaneous pulsations and instabilities in lasers and independently developed a quantum mechanically-based set of equations which predicted this type of behavior. Shortly thereafter, many papers began to appear in the literature based on various version of the coupled equations (48)–(50) and the field of chaotic laser behavior was off and running [35,36,37,38,39,40,41,42,43,44,45,46,47,48].

The analogy between the coupled field-atom SVEA equations (48)–(50) and the Lorenz equations (77) is intuitively clear upon visual comparison. However, in order to make the Lorenz equations analytically useful, one must specify the exact transformations between the envelope and population variables and the variables of the hydrodynamic system. This will depend, of course, on the exact form of the laser equations and these will vary depending on the laser configuration and the author's approach to the analysis. Thus, when studying the literature, one may come across a myriad of differing definitions for the Lorenz variables in terms of the laser variables. For our problem in the present context, we make the following associations between the dependent variables in the two systems;

$$x = \sigma\tau_c\sqrt{\frac{2^*\kappa}{\hbar\omega}}\tilde{E}\,, \quad y = \sigma\tau_c\sqrt{\frac{2^*\kappa}{\hbar\omega}}\left(-\mathrm{i}\frac{\omega\tau_c}{\epsilon}\right)\tilde{P}\,, \quad z = \frac{\kappa\sigma\tau_c^2}{\epsilon}\left[\Delta\mathcal{N}-\Delta\mathcal{N}_0\right] \tag{78}$$

along with rescaling the time coordinate $t' \to t/T_2$. When these expressions, and those for the coefficients α, r and β (Table 2) are substituted into the Lorenz equations (77), and the derivatives are changed accordingly ($\mathrm{d}/\mathrm{d}t \to (1/T_2)\,\mathrm{d}/\mathrm{d}t'$), the full set of laser equations (48)–(50) are recovered for the "on resonance" condition, $\omega = \omega_a = \omega_c$. Within this context, we may make the following asso-

ciations;

$$x = \text{FIELD AMPLITUDE} \tag{79}$$

$$y = \text{POLARIZATION} \tag{80}$$

$$r - z = \text{POPULATION INVERSION}\,. \tag{81}$$

We may also associate with r an effective pumping rate since it sets the population levels in the absence of driving fields. The laser equations in the SVEA written in the form of the Lorenz equations are sometimes called the "Lorenz–Haken" equations or "semi-classical Maxwell–Bloch" equations.

What can we learn about laser dynamics from the Lorenz equations? A good deal, as it turns out. Theoretical predictions and experimental verifications in a large number of laser systems have been made over the years and several excellent books covering many aspects of general laser dynamics, including this analysis, have appeared [5,6,12,16,49]. However, in the limited scope of this chapter, we will only have space to point out the highlights of the varied phenomena described by (77).

First off, we remark that within the simple set of Eqs. (77), there are only three dependent variables, one independent variable and three parameters. These means that the time evolution of $x(t)$, $y(t)$ and $z(t)$ depends critically on the values of the three parameters r, σ and b. Let us now undertake a brief study of the general behavior of the Lorenz equations for some specific values of the parameters. We begin by examining the stationary or steady-state solutions of the Lorenz equations, whence we obtain

$$\frac{\mathrm{d}x}{\mathrm{d}t} = 0 = -\sigma(x - y) \tag{82}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = 0 = (r - z)x - y \tag{83}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = 0 = xy - bz \tag{84}$$

where, for simplicity, we have dropped the prime notation on the time variable ($t' \to t$). By inspection, Eq. (82) implies $x = y$ and when this is substituted into (83) we have $z = r - 1$. Finally, substituting this result into (84), we find that both x and y become $(x, y) = \pm\sqrt{bz} = \pm\sqrt{b(r-1)}$. Summarizing, our steady-state solutions can be written

$$C_\pm = \begin{bmatrix} x_\pm \\ y_\pm \\ z \end{bmatrix} = \begin{bmatrix} \pm\sqrt{b(r-1)} \\ \pm\sqrt{b(r-1)} \\ r-1 \end{bmatrix}\,. \tag{85}$$

Let us examine the behavior of solutions as a function of the parameter r [16]. For $0 \le r \le 1$, the only physically meaningful solutions require that $x = y = z = 0$. This corresponds to the laser being pumped below threshold. There is not enough gain in the system to overcome the losses and consequently there is no buildup of the cavity fields or polarization. However, as we saw in the simplified model in Subsect. "Coupled Cavity-Atom Rate Equations", the below-threshold operation results in a linear increase in the population inversion (see Fig. 7).

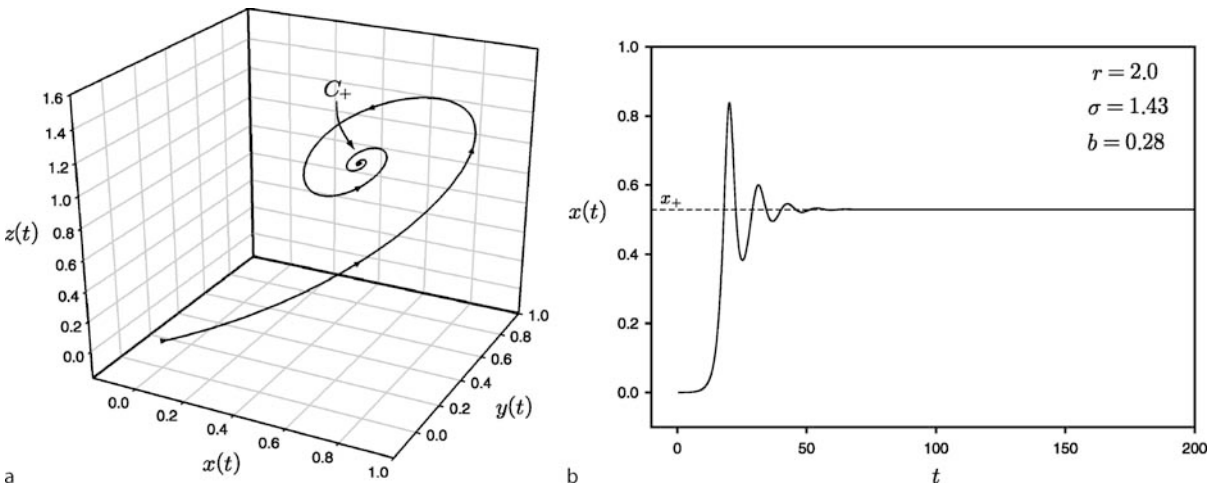

Noise and Stability in Modelocked Soliton Lasers, Figure 11
a Parametric plot of the solutions $x(t)$, $y(t)$, $z(t)$ of the Lorenz equations for $r = 2.0$, $\sigma = 1.43$ and $b = 0.28$. Coordinate of focus C_+ given by (85). b $x(t)$ alone (electric field) showing damped response relaxing to the steady state solution $x_+ = +0.53$

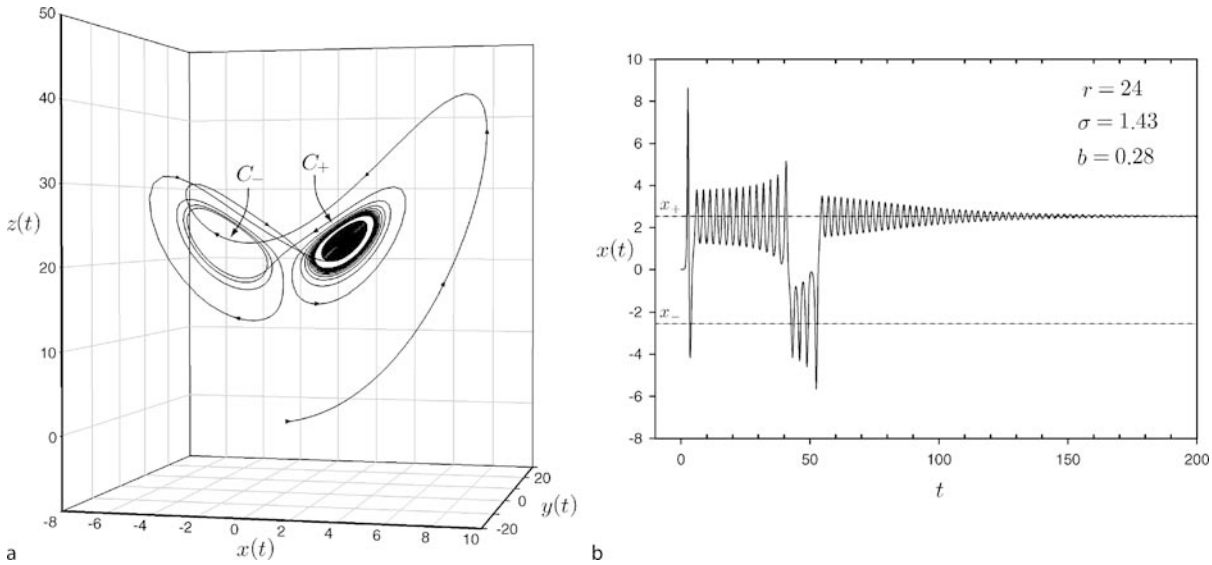

Noise and Stability in Modelocked Soliton Lasers, Figure 12
a Parametric plot of the solutions $x(t)$, $y(t)$, $z(t)$ of the Lorenz equations for $r = 24.0$, $\sigma = 1.43$ and $b = 0.28$. Coordinates of foci $C_\pm$ given by (85). b $x(t)$ alone (electric field) showing initially erratic response finally relaxing to damped steady state solution $x_+ = +2.54$

For $1 \leq r \leq 35.850$ the fields and population inversion rise to one of the two stable solutions in a uniform manner as shown in Fig. 11. A visually interesting way of plotting the results in shown in Fig. 11a where the three dependent variables, $x(t)$, $y(t)$ and $z(t)$ are plotted parametrically as a function of time. For this figure and the following, we have solved the Lorenz equations using $\sigma = 1.43$ and $b = 0.28$. We see that the solutions asymptotically spiral into the stable point C_+. Figure 11b shows just the function $x(t)$ (electric field) which has the traditional damped relaxation oscillation behavior as it approaches the stable value $x_+ = +0.53$.

As we turn up the value of r, we no longer find that the solutions uniformly converge onto one of the stable points, $C_\pm$. Rather, they tend to circle around the stable points in phase space and eventually land on one of the two. An example of this phenomenon is shown in Fig. 12 which was computed for $r = 24$. Notice that the solution trajectory (Fig. 12a) initially makes one quick pass around C_+ followed by a quick loop around C_- and then back to C_+ where it circles another 13 times, loops back to C_- and then finally goes into a long decaying orbit around C_+. Figure 12b again shows just the $x(t)$ solution and we can easily see how the electric field bounces back and forth between the two stable values $x_\pm = \pm 2.54$ until it settles down to x_+. This seemingly bizarre behavior is characteristic of the Lorenz equations and depends critically on the value of r. At special critical values (e. g. $r = 1, 4.9, 8,$ etc.), there are pronounced changes in the behavior of the orbits in phase space and these points are referred to as *bifurcations*.

When r is further increased beyond the value $r = 35.83$, a curious thing happens. The trajectory in phase space no longer settles into one of the two stable solutions, $C_\pm$, but instead circles about them indefinitely, never repeating any orbit and thus never becoming a "stable" solution. Figure 13 shows an example of this behavior for $r = 44$. In the parametric plot of Fig. 13a the initial rise from the origin has been excluded. In Fig. 13b we see the aperiodic, erratic behavior which constitutes "chaos". The region of phase space surrounding the so-called "steady-state" solutions, $C_\pm$, about which the trajectory propagates is called an *attractor*, presumably in the sense that a gravitational potential holds an orbiting body within its realm. When the dimensionality of the attractor is non-integer (i. e. fractal), then it is called a *strange* attractor [6,16].

If we continue turning up r, we find that the laser output becomes unstable at the value [16]

$$r = r_{\mathrm{H}} = \sigma(\sigma + b + 3)/(\sigma - b - 1) \tag{86}$$

which for the examples shown above occurs at $r = 44.90$. Physically, the potential well at the foci of the strange attractors has gone to zero depth. This point is referred to as the "second laser threshold". A necessary condition for

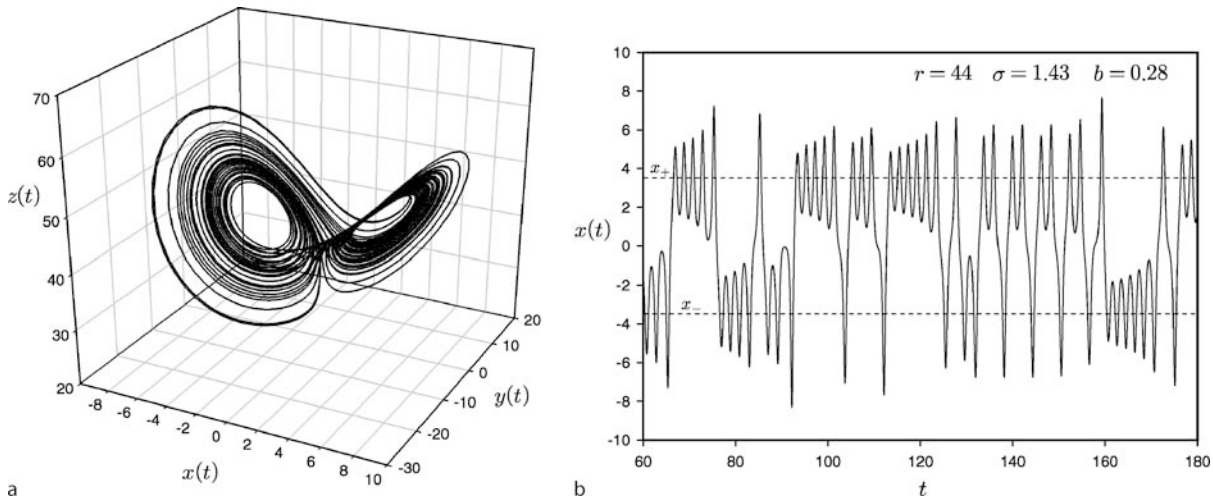

Noise and Stability in Modelocked Soliton Lasers, Figure 13
a Parametric plot of the solutions $x(t), y(t), z(t)$ of the Lorenz equations for $r = 44.0$, $\sigma = 1.43$ and $b = 0.28$. Solution propagates around foci ("strange attractors"). Coordinates of foci $C_\pm$ given by (85). b $x(t)$ alone (electric field) showing chaotic behavior with no repeating patterns

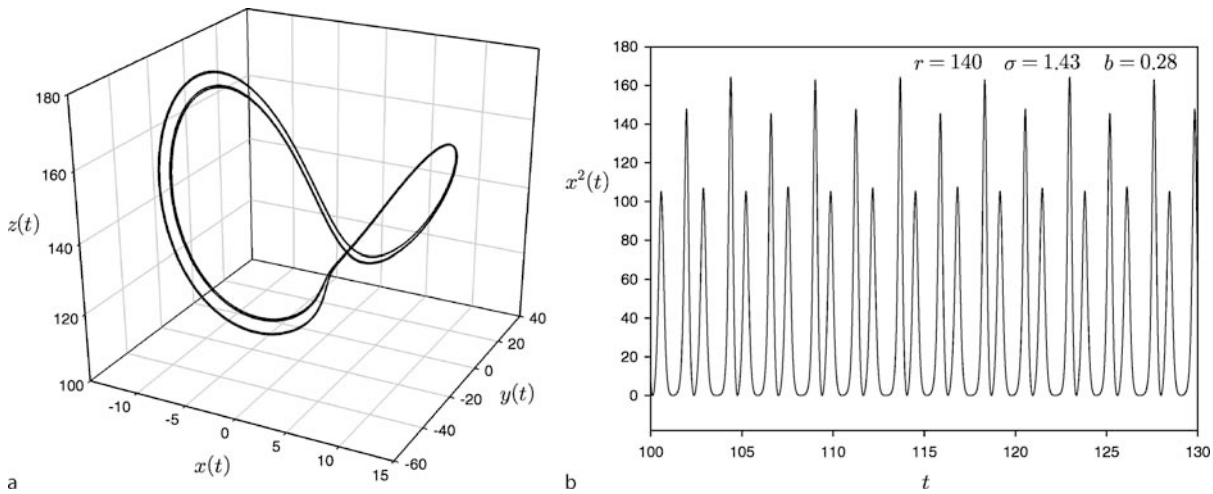

Noise and Stability in Modelocked Soliton Lasers, Figure 14
a Parametric plot of the solutions $x(t), y(t), z(t)$ of the Lorenz equations for $r = 140.0$, $\sigma = 1.43$ and $b = 0.28$. Notice that the orbit only close upon itself after 4 traverses around the original stable points $C_\pm$, thus demonstrating a four-fold period multiplication. b $x^2(t)$ alone (proportional to power in electric field)

this to occur can be seen from (86),

$$\sigma > b + 1\,, \quad \text{or} \quad \rightarrow \quad \tau_c < \frac{1}{T_1^{-1} + T_2^{-1}}\,. \tag{87}$$

This means that the cavity decay rate must exceed the sum of the energy decay plus the polarization decay rates. This is known as the *bad cavity condition* and in general, it is hard to achieve, especially in the visible portion of the spectrum. However, the spontaneous emission rate (which is, effectively, T_1^{-1}) is proportional to ω_a^2 (6) which favors long wavelength lasers. Also, for gas lasers the polarization decay rate is dominated by collisions which is proportional to pressure. These two facts favor long-wavelength gas lasers as a proving ground for investigating many of the predictions of the Lorenz equations and, indeed, most of the early work in the field was conducted with, e. g., laser-pumped ammonia (NH_3) ring lasers ($\lambda = 81\,\mu$m, homogeneously broadened) [16] and discharge-pumped He-Xe lasers ($\lambda = 3.5\,\mu$m, inhomogeneously broadened) [34].

Noise and Stability in Modelocked Soliton Lasers, Table 3
Range of values of pump parameter *r* and related laser behavior

r	CHARACTERISTIC FEATURES
$0 \le r \le 1$	Below threshold. Spontaneous emission.
$r = 1$	First laser threshold.
$1 \le r \le r_A$	Asymptotically approaches stable solutions $C_{\pm}$.
$r = r_A$	$r_A = 35.83$. Upper limit of asymptotic relaxation to $C_{\pm}$.
$r > r_A$	No periodicity. Irregular and chaotic. Strange attractors.
r_H	Second laser threshold. $r_H = \sigma(\sigma + b + 3)/(\sigma - b - 1) = 44.9$
$r > r_H$	Unstable.
$r = r_B$	$r_B \approx 138$. Boundary for strange attractor. Orbits start to close.
$r > r_C$	$r_C \approx 140$. Closed orbits. Period multiplication.

Continuing to increase r demonstrates another interesting phenomenon, that of period multiplication. In Fig. 14, we have plotted the solutions for $r = 140$ and see that the trajectory in phase space is almost a closed curve but, in fact, cycles 4 times before repeating. This is a "period-quadrupling" ($T/4$) bifurcation and occurs throughout the region for various values of $r \gtrsim 140$. Finally, as we reach far beyond $r = 150$, we start to see solid closed orbits in phase space (Fig. 15). A short summary of the important ranges of the pumping parameter r and the related laser characteristics is given in Table 3 where we have borrowed some nomenclature from the excellent text by Weiss and Vilaseca [16].

Now, we have only given a rather brief account of these interesting phenomena within the context of one specific set of laser parameters. There are many other laser systems that exhibit unusual chaotic behavior, such as semiconductor diode lasers. The reader interested in a broader perspective is encouraged to see, for example, [23,50].

Modelocking

All of the principles discussed so far regarding lasers have related to continuous-wave (CW) operation. Lasers can also be configured to send out pulses with a tremendous amount of energy and short duration. The future (and potential) of inertial confinement fusion rests on this very fact. There are predominantly two different techniques for producing short pulses in lasers. In the first, called Q-switching, the atoms in the laser gain medium are pumped up to an inverted condition far beyond that normally required to achieve CW oscillation. This can happen if an intracavity loss mechanism is introduced to prevent the feedback which initiates and sustains lasing. Thus we say that the "Q" of the cavity is temporarily spoiled. When most of the atoms are inverted, the cavity Q is restored by switching off the loss. Within a few round trips, the electromagnetic fields in the cavity build up to an extremely high level and almost all of the energy stored in the inverted atomic population is released in a giant pulse. The dynamics of this process are important. The speed with which the entire population can be inverted depends on the pumping rate and the relaxation rates of the pertinent energy levels. This can typically take 10's to 100's of round

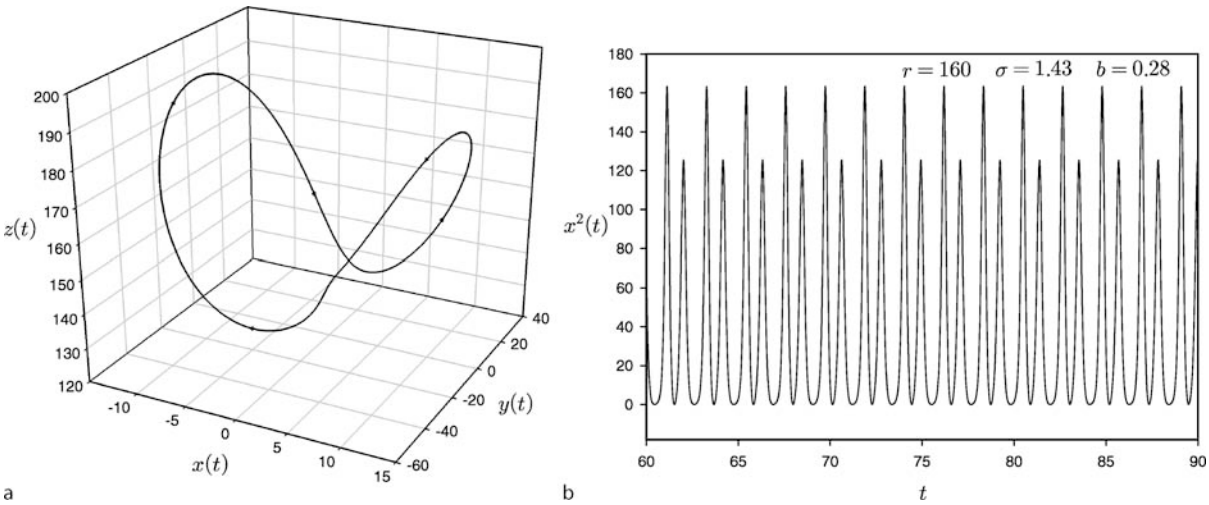

Noise and Stability in Modelocked Soliton Lasers, Figure 15
a Parametric plot of the solutions $x(t)$, $y(t)$, $z(t)$ of the Lorenz equations for $r = 160.0$, $\sigma = 1.43$ and $b = 0.28$. Now the orbit closes upon itself after one simple traversal. b $x^2(t)$ alone (proportional to power in electric field)

trip cavity-times. The final pulse width is typically on the order of a few to 10's of nanoseconds and the repetition rate can range from < 10 Hz to > 10 KHz.

Another mechanism for producing short pulses with high energy is called "modelocking". This technique produces far shorter pulses at a much higher repetition rate than Q-switching, albeit at lower energies per pulse. It is the technique of choice for producing the shortest optical pulses, and today, pulses of just a few cycles duration ($\tau \approx 5$ femtoseconds) in the visible portion of the spectrum are routinely generated in laboratories and soft x-ray pulses have even crossed the 1 femtosecond barrier ($1\,\text{fs} = 10^{-15}$ s) [51]. The term "modelocked" as applied to lasers arises when one considers the cooperative effect of many thousands, or hundreds of thousands, of spectral comb lines which constitute a train of ultrashort light pulses. Although there are a number of approaches to achieving this phenomenon, we will discuss only the one that is currently responsible for the revolution in ultrashort pulse laser technology and the explosive growth of ultrafast science as well [52].

Consider the elementary laser system shown in Fig. 16. There are four essential components in this model. All oscillators must have gain and feedback, but here we added nonlinearity and dispersion. Let us assume we have a pulse somewhere in the cavity and suppose we follow this pulse around on complete loop. Beginning with the entrance to the gain medium it is easy to understand that the pulse is simply amplified, provided that the gain medium is linear. Next, the pulse enters a nonlinear medium which preferentially passes the highest amplitude portion of the pulse and this causes a narrowing. Such an object is sometimes called a "saturable absorber" but actual energy absorption is not necessary to cause pulse narrowing (refractive effects such as self-focusing produce similar effects). This pulse narrowing is almost always accompanied by self-phase modulation. In this model some of the narrowed pulse energy escapes the cavity through a partially-transmitting output coupler. Most of the pulse energy stays within the cavity and next passes through a region of dispersion which has the effect of broadening and chirping the pulse. Finally, the broadened pulse is amplified again and we have completed one circuit of the laser. This highly simplified model neglects many subtle, but important, effects such as self-focusing, diffraction losses, bandwidth limiting and dispersion of elements like mirror coatings, etc. However, the basic four elements in this model can create a self-consistent situation that fairly accurately describes the modern generation of "Kerr-lens modelocked soliton lasers". The Kerr-lensing effect is brought about by the high peak powers in the circulating pulse causing a self-focusing effect in the laser gain medium which is designed to match the other focusing elements in the cavity. This creates a sort of unstable equilibrium in which only a high-power pulse can be a self-consistent solution since the self-focusing is *required* in order to produce a stable cavity as far as the transverse properties of the laser field are concerned. The appellation "soliton" arises from the similarity between the periodic application of nonlinearity and dispersion in the laser to the distributed effects of nonlinearity and dispersion which are required in systems that admit the solitary waves called "solitons" as solutions [53].

The self-consistent picture of a pulse evolving every round-trip within the cavity also implies that the laser produces a train of pulses with spacing, T, equal to the round-trip cavity time (see Fig. 16). Simple Fourier transformation of such a periodic sequence gives a spectrum of periodically-spaced comb-lines with interline spacing $\delta\omega = 2\pi/T$ as shown in the figure. The width of the spectrum, and hence the number of comb-lines contained therein, is inversely proportional to the pulsewidth. For example, a 5 fs pulse has a bandwidth $\Delta\nu \approx 100$ THz which therefore contains about 10^6 comb lines if the laser has a 100 MHz repetition rate. The remarkable, and es-

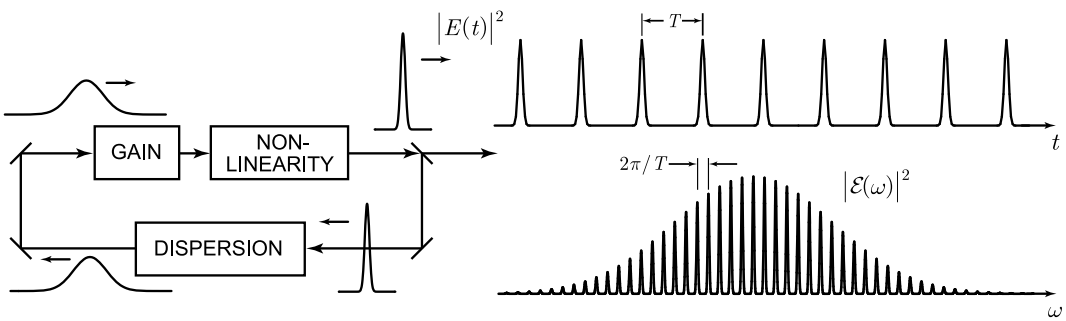

Noise and Stability in Modelocked Soliton Lasers, Figure 16
Principle of operation of the modelocked soliton laser

sential, fact about these spectral lines is that they must all have a well-defined relative phase with respect to each other. Furthermore, these phases must remain constant with time, or be "locked" together. The spectral lines of the pulse train are also, in fact, the allowed eigenmodes or axial modes of the laser cavity and thus we say that the axial modes are locked, or the laser is "modelocked".

Solitons

One of the principal features of linear systems is superposition of solutions. That is, if $u_1(x,t)$ and $u_2(x,t)$ are separate independent solution to a particular problem in a linear system, then the sum of the two, $u_1(x,t)+u_2(x,t)$, is also a solution. One could say that the two solutions are "unaware" of each other's presence in the linear system. They do not interact or depend upon each other in any way. The same cannot be said for a nonlinear system. The presence of two disturbances in a nonlinear system generally results in interaction between the two or even self-interaction within a single solution.

Except for electromagnetic waves in a vacuum, almost all wave propagation problems are nonlinear in nature. This is because the waves depend upon a medium to transfer their energy. Perturbations about equilibrium of the constituents of the medium eventually drive them into a nonlinear region of response given a large enough driving force and thus the collective wave phenomena must be described by a nonlinear wave equation. An example of such an equation is that which describes low-amplitude shallow water waves. Derived in 1895 by Korteweg and de Vries [54], this has come be known as the Korteweg–de Vries or KdV equation;

$$\frac{\partial u}{\partial t} - 6u\frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3} = 0\,. \tag{88}$$

This equation was originally developed to describe the long solitary waves that were first observed in a canal by John Scott Russell in 1834 [55].

Assuming a traveling-wave solution of the form $u(x,t) = u(x - vt)$ then it can be shown that the solution to the KdV equation is [53,56,57]

$$u(x,t) = -\frac{v}{2}\mathrm{sech}^2\left[\frac{\sqrt{v}}{2}(x - vt - x_0)\right]\,. \tag{89}$$

This function represents an isolated pulse traveling with velocity v. x_0 is the position of the center of the pulse at time $t = 0$. Notice that as the velocity increases, so does the amplitude while the width of the pulse decreases. This coupling between the amplitude, velocity and pulsewidth is quite typical of nonlinear waves.

The KdV equation was studied for many years and in 1965, in a seminal paper, Zabusky and Kruskal [58] discovered that the solitary wave solutions (89) to the KdV equation could interact *elastically*. That is, pulses could collide, exchange energy and emerge with the same profile and velocity and only a slight phase shift. This resurrection of pulse form following interaction in a nonlinear medium was unexpected and, owing to the similarity to fundamental particle scattering, the authors coined the name "soliton" for these very special types of solitary waves. The initial paper of Zabusky and Kruskal launched quite a revolution in applied mathematics and soon solitons were predicted and observed in many systems [53,59].

One of the many nonlinear wave equations that admits solitons as solutions is the nonlinear Schrödinger equation (NLS), also called the cubic Schrödinger equation. This is particularly relevant to the world of ultrafast optics since solitons were predicted by Hasegawa and Tappert as solutions to the wave equation in optical fibers under certain conditions [60]. Simply put, the required conditions are that the intrinsic anomalous dispersion of the optical fiber should exactly balance the nonlinearity due to the quadratic electro-optic, or Kerr, effect. This meant that short optical pulses (and hence high data rates) could be transmitted over very long distances without suffering the usual broadening associated with dispersion provided that the peak pulse power was high enough to drive a nonlinear reaction that exactly compensated for the linear dispersion. In an elegant series of experiments, Mollenauer et al., demonstrated this phenomenon and set the stage for all future optical fiber telecommunications technology [61].

The nonlinear Schrödinger equation, in normalized units, is

$$\mathrm{i}\frac{\partial u}{\partial t} + \frac{\partial^2 u}{\partial x^2} + 2|u|^2 u = 0\,. \tag{90}$$

We assume solutions of the general form

$$u(x,t) = \phi(x,t)\mathrm{e}^{\mathrm{i}\theta(x,t)} \tag{91}$$

where both $\phi(x,t)$ and $\theta(x,t)$ are assumed to be real. Substituting these into the NLS equation (90) above, we find that the traveling-wave forms

$$\phi(x,t) = \tilde{\phi}(x - v_\mathrm{e}t) \tag{92}$$

$$\theta(x,t) = \tilde{\theta}(x - v_\mathrm{c}t) \tag{93}$$

will work as solutions [53], where v_e and v_c are the envelope and carrier velocities, respectively. Specifically, we find that

$$\tilde{\phi} = a\,\mathrm{sech}\left[a(x - v_\mathrm{e}t)\right] \tag{94}$$

provided that

$$4a^2 = v_e^2 - 2v_e v_c \tag{95}$$

is also satisfied. The final form of the solution can then be shown to be

$$u(x,t) = a \exp\left[\mathrm{i}\frac{v_e}{2}x + \mathrm{i}\left(a^2 - \frac{v_e^2}{4}\right)\right] \operatorname{sech}\left[a(x - v_e t - x_0)\right] . \tag{96}$$

Like the solution to the KdV equation, this wave function is also a soliton and variations on it will appear again and again in modelocked laser physics and nonlinear optical fiber analysis.

Modelocked Soliton Lasers

To see how solitons come into the theory of certain types of modelocked lasers, we develop a self-consistent model of a modelocked laser following a packet of photons around a unidirectional ring cavity (Fig. 17). This will generate the so-called "master-equation" as originally developed by Haus [62] to analyze modelocking with a fast saturable absorber. It was later refined to include nonlinear phase shifts due to self-modulation [63] and the Kerr-lens effect [64] and serves as the starting point for noise and stability analyses [65,66,67].

Consider the schematic of the ring laser shown below (Fig. 17). The physical phenomena related to the passage of a pulse around the ring are grouped into four separate boxes. They account for linear loss and phase shift, gain (or amplification), group velocity dispersion (GVD) and self phase modulation (SPM) and saturable absorption. The key assumptions in the following analysis are [62] 1) The relaxation time of the saturable absorber is very short compared with the optical pulse width, 2) The saturable absorption is describable using a rate equation approach and the power dependent absorption coefficient is expanded to first order in the power, 3) The relaxation time of the laser gain medium is long compared to the optical pulse width so that the gain is essentially constant, 4) The modelocked pulse is changed by only a small amount upon each circuit of the ring, and 5) The spectrum of the modelocked pulse is small compared with the laser gain bandwidth.

The slowly-varying envelope approximation will allow us to be primarily concerned with the behavior of the envelope of the optical pulse, which we shall denote as $a(t)$ such that, in suitable units, the electric field is given by $E(t) = a(t)\exp\left[\mathrm{i}(\omega_0 t - k_0 z)\right]$. The total loss due to diffraction, dielectric loss (absorption), and finite mirror reflectivities, as well as the net linear phase shift will produce a complex change in the amplitude

$$\Delta a = -(l + \mathrm{i}x)a \tag{97}$$

per pass. For the case of low gain, the change in amplitude per pass is

$$\Delta a = g\left(1 + \frac{1}{\Omega_g^2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\right)a \tag{98}$$

where Ω_g is the bandwidth of the gain curve in the parabolic approximation. The group velocity dispersion (GVD) produces a change that is purely imaginary

$$\Delta a = \mathrm{i}D\frac{\mathrm{d}^2}{\mathrm{d}t^2}a . \tag{99}$$

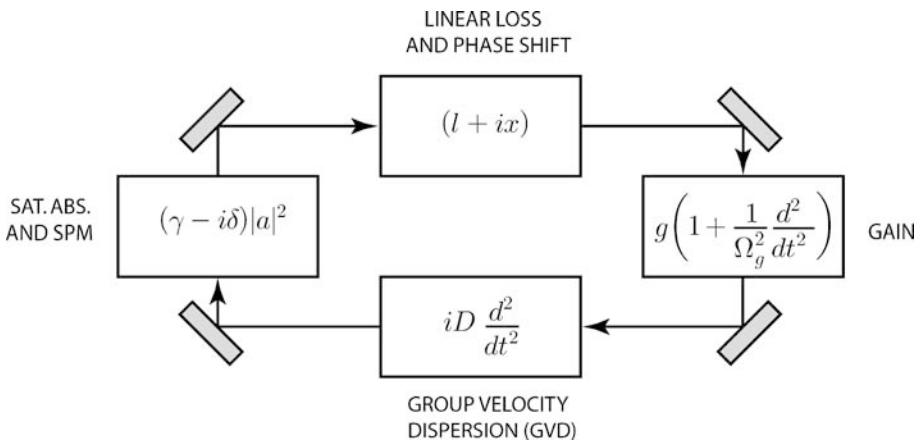

Noise and Stability in Modelocked Soliton Lasers, Figure 17
Unidirectional ring laser cavity for analysis of the modelocked soliton laser

Here D is the dispersion parameter defined as

$$D = \frac{1}{2}\frac{\mathrm{d}^2 k}{\mathrm{d}\omega^2}\, l_{\mathrm{eff}}\,, \tag{100}$$

where k is the propagation constant of the net dispersive media in the cavity with an effective length l_{eff} (Note: here again we combine all of the dispersion that can occur due to dielectrics, mirror coatings, etc., into one term for convenience).

Self phase modulation (SPM) appears as a term proportional to the peak amplitude

$$\Delta a = -\mathrm{i}\delta |a|^2 a \tag{101}$$

where the nonlinear coefficient takes the form

$$\delta = \frac{\omega_0}{c}\frac{n_2 d_{\mathrm{eff}}}{\mathcal{A}_{\mathrm{eff}}}\,. \tag{102}$$

The nonlinear index n_2 is due to the Kerr effect. It contributes to the index of refraction according to

$$n = n_{\mathrm{o}} + n_2 I \tag{103}$$

where I is the intensity (W/m^2) of the optical field. $\mathcal{A}_{\mathrm{eff}}$ is the effective field mode area and d_{eff} is the effective length of the nonlinear medium.

We also must include the effect of a fast saturable absorber,

$$\Delta a = \gamma |a|^2 a \tag{104}$$

where γ is the saturation parameter and is inversely proportional to the saturation intensity I_{sat}.

Finally, we add a term that allows for small carrier frequency changes away from the cavity mode resonances (Fabry–Perot modes). A shift in frequency of $\Delta\omega_0$ corresponds to a phase shift per pass of

$$\psi = \frac{\Delta\omega_0}{c} L_{\mathrm{eff}} \tag{105}$$

where L_{eff} is the effect round-trip cavity path length. The change in field amplitude is therefore

$$\Delta a = \exp(-\mathrm{i}\psi)a - a \approx -\mathrm{i}\psi a\,. \tag{106}$$

We now combine all of these effects by assuming that on a per-pass basis, they are all small and therefore are additive. In the steady state, then, they must all sum to zero

$$\left[-\mathrm{i}\psi - (l + \mathrm{i}x) + g\left(1 + \frac{1}{\Omega_{\mathrm{g}}^2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\right) + \mathrm{i}D\frac{\mathrm{d}^2}{\mathrm{d}t^2} + (\gamma - \mathrm{i}\delta)|a|^2\right] a = 0\,. \tag{107}$$

This is the "master equation" of steady-state modelocking [63]. A solution to this equation can be found by introducing the ansatz

$$a(t) = A_{\mathrm{o}}\,\mathrm{sech}\left(\frac{t}{\tau}\right)\exp\left[\mathrm{i}\beta\, ln\,\mathrm{sech}\left(\frac{t}{\tau}\right)\right] \tag{108}$$

where β is a so-called "chirp parameter" and τ is the pulsewidth. (Note that this form of sech pulse is remarkably similar to the soliton solutions of the KdV and NLS equations, and this version version allows for chirped pulses; that is, pulses with a sweep in frequency). If we substitute the ansatz into the master equation (107), and match coefficients common to the sech and sech^2 terms, we generate two complex equations in four unknowns,

$$-\mathrm{i}\psi + g - l - \mathrm{i}x + \frac{(1+\mathrm{i}\beta)^2}{\tau^2}\left(\frac{g}{\Omega_{\mathrm{g}}^2} + \mathrm{i}D\right) = 0 \tag{109}$$

$$\frac{2 + 3\mathrm{i}\beta - \beta^2}{\tau^2}\left(\frac{g}{\Omega_{\mathrm{g}}^2} + \mathrm{i}D\right) = (\gamma - \mathrm{i}\delta)A_{\mathrm{o}}^2\,. \tag{110}$$

The four unknowns determined by these equations are the pulsewidth, τ, the chirp parameter, β, the phase shift, ψ, and the gain, g. The pulse amplitude A is assumed to be such that the gain saturates down to a level that is approximately equal to the loss ($g \approx l$) where

$$g = \frac{g_0}{1 + 2A^2\tau/P_{\mathrm{s}}T_{\mathrm{R}}} = \frac{g_0}{1 + W/P_{\mathrm{s}}T_{\mathrm{R}}} \tag{111}$$

where the pulse *energy* is given by $W = 2A^2\tau$, P_{s} is the effective saturation power and T_{R} is the round-trip cavity time.

The complete solutions for τ, β, ψ and g and the conditions for their realization need not be elaborated here since they are covered in detail elsewhere [63,64]. Suffice it to say that steady-state solutions of a soliton nature do exist and form the basis for most modern ultrashort-pulse solid-state lasers [52]. Pulses of < 10 fs duration with peak powers of several hundred kilowatts and repetition rates of 50–1000 MHz are now routinely available in laboratories worldwide.

We next move on to learn about the fundamental processes that govern and generate noise in both CW and modelocked lasers and finally we will return the specific

case of the modelocked soliton laser to see how the important parameters, derived above, vary when subject to perturbations, for this is at the heart of the issue of instability.

Laser Noise and Linewidth

Noise accompanies all signals and from an elementary Fourier transform point of view, noise necessarily broadens the spectral linewidth of any source. One can appreciate this causal relationship better using the concepts of modulation which will be developed in Subsect. "Analytical Description of Envelope Noise". We can distinguish between two sources of noise in a laser field. The first, sometimes called "technical noise" arises from environmental perturbations to the oscillator which creates the field. These are controllable to a greater or lesser degree and are not fundamental in the sense that they are not necessary parts of the amplification or oscillation process. The second source of noise is the spontaneous emission of photons and arises due to the large number of inverted atoms (or molecules) required to produce gain through stimulated emission. Spontaneous emission occurs with equal probability in all directions but some of these photons will radiate directly into the desired laser cavity mode.

Since the generation of the spontaneously emitted photons is a random process, they are not coherent with the desired mode and thus constitute noise. If it were not for the presence of these noise photons, the laser electromagnetic field would be perfectly coherent and measurement of the spectrum of the field with an ideal spectrometer would yield a single narrow line at the oscillation frequency. The width of this line, $\Delta\nu$, would be determined solely by the inverse of the time spent in its measurement: $\Delta\nu \approx 1/\Delta t_{\text{meas}}$. However, the addition of spontaneous emission photons to the ideal coherent field produces a small perturbation to the amplitude and phase of the field which, in turn, will broaden the width of the spectrum. This noise mechanism sets a fundamental quantum limit to the minimum achievable laser linewidth and is therefore an essential concept in laser physics. We now proceed to calculate the nature of this noise and how it affects the linewidth of the desired field. (This treatment follows that of Yariv [10].)

Oscillator Spontaneous Emission Noise

In order to incorporate the presence of spontaneous emission into the coherent field of a laser oscillator, we write the field as

$$\mathcal{E}(t) = \text{Re}\left[E(t)e^{i(\omega_0 t+\theta(t))}\right] \tag{112}$$

where $E(t)$ and $\theta(t)$ are random processes whose fluctuations are very small about their mean values. Spontaneous emission noise can be modeled as adding to the otherwise perfectly coherent field a photon at the frequency of the transition, $(E_2 - E_1)/h$. The *phasor picture* of the field with the addition of this randomly oriented spontaneous photon provides a most satisfactory view of the process (see Fig. 18). We will refer to the unperturbed electromagnetic field as the *carrier*, the magnitude of which is proportional to the square root of the mean number of photons in the field mode; $|E| \propto \sqrt{\bar{n}}$. The addition of one spontaneous emission photon to the carrier is accomplished by adding to the electric field phasor a unit-amplitude phasor ($n = 1$) with random phase which we will call ϕ and measure it arbitrarily with respect to the normal to the carrier phasor following the addition of the noise photon. The phase angle ϕ is uniformly distributed on the interval $[0, 2\pi]$ and thus the probability density function $p(\phi) = 1/2\pi$. In general, the number of carrier photons will be quite large ($\bar{n} \gg 1$) and thus the phase deviation of the carrier phasor due to a single spontaneous emission photon, $\Delta\theta(1)$, can be found from the projection

$$1 \cdot \cos\phi = \sqrt{\bar{n}}\,\sin\Delta\theta(1) \approx \sqrt{\bar{n}}\,\Delta\theta(1) \tag{113}$$

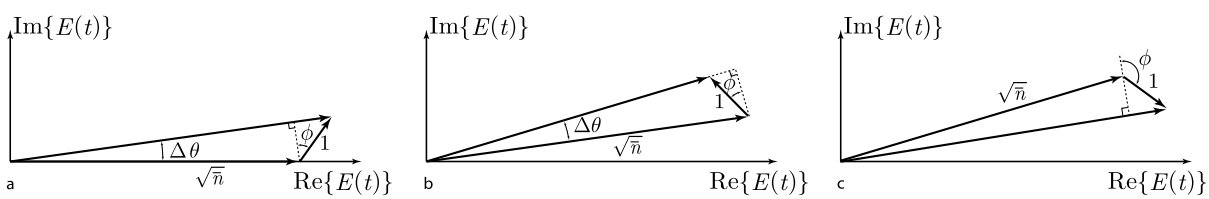

Noise and Stability in Modelocked Soliton Lasers, Figure 18
Phasor picture of the addition of a spontaneous emission photon (amplitude = 1) to the electric filed optical carrier *E*(*t*) (amplitude = $\sqrt{\bar{n}}$ where $\bar{n}$ is the mean photon number in the carrier field). The phase angle ϕ of the spontaneous emission photon is a random variable and thus its projection on the carrier phasor, cos ϕ, introduces phase ($\Delta\theta$) and amplitude perturbations to the electric field. The sequence a–c depicts three spontaneous emission events adding to the carrier field. Since the mean photon number is typically very large, $\bar{n} \gg 1$, the phasor lengths here are highly exaggerated

or

$$\Delta\theta(1) = \frac{1}{\sqrt{\bar{n}}}\cos\phi\,. \tag{114}$$

The carrier phase deviation $\Delta\theta$ is a random variable and its statistics are what determine the lineshape. Since ϕ is uniformly distributed on the interval $[0, 2\pi]$, the mean value of the carrier phase deviation after a large number N_{sp} of spontaneous emission events is obviously zero. This can be shown simply;

$$\begin{aligned}\langle\Delta\theta(N_{sp})\rangle &= \int_0^{2\pi}\Delta\theta(1)p(\phi)\,\mathrm{d}\phi \\ &= \frac{1}{2\pi\sqrt{\bar{n}}}\int_0^{2\pi}\cos\phi\,\mathrm{d}\phi = \frac{1}{2\pi\sqrt{\bar{n}}}\cdot 0 = 0\,.\end{aligned} \tag{115}$$

The *mean-squared* value of the phase deviation, on the other hand, is non-zero and is given from the results of the random walk process [68],

$$\langle[\Delta\theta(N_{sp})]^2\rangle = \langle[\Delta\theta(1)]^2\rangle N_{sp} = \frac{1}{\bar{n}}\langle\cos^2\phi\rangle N_{sp} = \frac{N_{sp}}{2\bar{n}}. \tag{116}$$

This is the mean-squared phase deviation after N_{sp} spontaneous emission photons have been emitted and combined with the carrier field. The next task is to convert this expression to an r.m.s.phase deviation in a given time interval τ. We must first calculate the average number of spontaneous emission events in a time τ. The total number of spontaneously emitted photons per second into *all modes* (i. e. the spontaneous emission rate) is given by

$$\tilde{N}_{sp}(s^{-1}) = \frac{N_2}{\tau_{rad}} = N_2\gamma_{rad}\,. \tag{117}$$

The tilde on N_{sp} serves to indicate that this is a *rate* as opposed to a simple number. Since all of the spontaneous photons are emitted into all possible modes (4π steradians), we need to normalize this to the total number of modes, p, within the atomic linewidth

$$\tilde{\mathcal{N}}_{sp}(s^{-1}\mathrm{mode}^{-1}) = \frac{N_2}{\tau_{rad}\,p} = N_2\frac{\gamma_{rad}}{p}\,. \tag{118}$$

Now, recall that the population inversion at threshold is given by

$$\Delta N_{th} \equiv (N_2 - N_1)_{th} = \frac{\gamma_c}{\gamma_{rad}}\,p \tag{119}$$

where γ_c is the energy decay rate of the cavity. We can now use this result to eliminate the mode number p from (118),

$$\begin{aligned}\tilde{\mathcal{N}}_{sp} &= N_2\frac{\gamma_{rad}}{p}\left(\frac{\Delta N_{th}}{\Delta N_{th}}\right) = \frac{N_2}{\Delta N_{th}}\frac{\gamma_{rad}}{p}\Delta N_{th} \\ &= \frac{N_2}{\Delta N_{th}}\frac{\gamma_{rad}}{p}\frac{\gamma_c}{\gamma_{rad}}\,p \qquad (120)\\ &= \frac{N_2}{\Delta N_{th}}\gamma_c = \frac{N_2}{\Delta N_{th}}\frac{1}{\tau_c}\,. \qquad (121)\end{aligned}$$

This is the spontaneous emission rate into a single mode. The total number of spontaneous photons into this mode in a time τ is therefore

$$N_{sp}(\tau) = \tilde{\mathcal{N}}_{sp}\tau = \frac{N_2}{\Delta N_{th}}\frac{\tau}{\tau_c}\,. \tag{122}$$

We can now use this result in (116) to obtain the mean-squared phase deviation in a time τ,

$$\langle[\Delta\theta(N_{sp}(\tau))]^2\rangle = \frac{\tau}{2\bar{n}\tau_c}\frac{N_2}{\Delta N_{th}} \tag{123}$$

and thus the r.m.s. phase deviation is

$$\Delta\theta_{rms}(\tau) \equiv \sqrt{\langle[\Delta\theta(N_{sp}(\tau))]^2\rangle} = \sqrt{\frac{\tau}{2\bar{n}\tau_c}\frac{N_2}{\Delta N_{th}}}\,. \tag{124}$$

Because the amplitude and phase of the electric field (112) incorporate random processes, the Fourier transform does not exist, but the power spectrum $S_{\mathcal{E}}(\omega)$ does and is related to the Fourier transform of the autocorrelation function of the electric field $R_{\mathcal{E}}(\tau)$ as established by the Wiener–Khinchine theorem [69]

$$S_{\mathcal{E}}(\omega) = \int_{-\infty}^{\infty} R_{\mathcal{E}}(\tau)\mathrm{e}^{-\mathrm{i}\omega\tau}\,\mathrm{d}\tau = \mathcal{F}\{R_{\mathcal{E}}(\tau)\}\,. \tag{125}$$

The autocorrelation function is defined by

$$R_{\mathcal{E}}(\tau) \equiv \langle\mathcal{E}(t)\mathcal{E}(t+\tau)\rangle = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}\mathcal{E}(t)\mathcal{E}(t+\tau)\mathrm{d}t. \tag{126}$$

Now, because the autocorrelation function contains expectation values of the electric field, we need to provide probability density functions for the random variables. At this point, we note that for an oscillator, the gain is, by definition, saturated, and thus there is not much amplitude fluctuation so we let $E(t) = E$. The majority of the spontaneous emission contributes to phase fluctuations. Since $\Delta\theta$ is the result of a large number of statistically independent events, we can apply the central limit theorem which

says that all such independent random processes approach a Gaussian distribution in the limit of a large number of samples [68,69]. Thus, a normalized Gaussian distribution with mean-squared value $\langle(\Delta\theta)^2\rangle$ is given by

$$p(\Delta\theta) = \frac{1}{\sqrt{2\pi\,\langle(\Delta\theta)^2\rangle}}\,\mathrm{e}^{-(\Delta\theta)^2/2\langle(\Delta\theta)^2\rangle}\,. \tag{127}$$

Since we have already calculated the mean-squared value of the phase fluctuation for the spontaneous emission (123) the autocorrelation function can be found and then we can proceed to calculating the power spectral density (the details of which will be omitted for brevity but can be found in [10])

$$S_{\mathcal{E}}(\omega) = \int_{-\infty}^{\infty} R_{\mathcal{E}}(\tau)\mathrm{e}^{-\mathrm{i}\omega\tau}\,\mathrm{d}\tau \tag{128}$$

$$= \frac{\langle E^2\rangle}{2}\frac{N_2}{4\hbar\tau_c\Delta N_{\mathrm{th}}}\left[\frac{1}{\left(\frac{N_2}{4\hbar\tau_c\Delta N_{\mathrm{th}}}\right)^2 + (\omega+\omega_o)^2} + \frac{1}{\left(\frac{N_2}{4\hbar\tau_c\Delta N_{\mathrm{th}}}\right)^2 + (\omega-\omega_o)^2}\right]. \tag{129}$$

We see that the spectral density is composed of two Lorentzian lineshapes at $\omega = \pm\omega_o$. The frequencies for which this Lorentzian is reduced to half of its value on resonance are

$$\omega_{\pm 1/2} = \omega_o \pm \frac{N_2}{4\hbar\tau_c\Delta N_{\mathrm{th}}}\,. \tag{130}$$

The full linewidth at half-maximum (FWHM) is therefore

$$\Delta\omega_{\mathrm{OSC}} = \omega_{1/2} - \omega_{-1/2} = \frac{N_2}{2\hbar\tau_c\Delta N_{\mathrm{th}}} \tag{131}$$

or, from (123), we have

$$\Delta\omega_{\mathrm{OSC}} = \frac{\langle[\Delta\theta(N_{\mathrm{sp}}(\tau))]^2\rangle}{\tau}\,. \tag{132}$$

The Lorentzian linewidth is just the mean-squared phase deviation per unit time interval τ.

We can rewrite (131) in a slightly different, but useful, form by noting that the total power emitted by the atoms into the oscillating mode is $P_{\mathrm{OSC}} = \hbar\hbar\omega_o/\tau_c$. We also note that the cold-cavity linewidth (FWHM) of the resonator is given by $\Delta\nu_c = (2\pi\tau_c)^{-1}$. Using these in (131) above we have

$$\Delta\omega_{\mathrm{OSC}} = \frac{\hbar\omega_o N_2}{2P_{\mathrm{OSC}}\tau_c^2\Delta N_{\mathrm{th}}} = \frac{\hbar\omega_o N_2(2\pi\Delta\nu_c)^2}{2P_{\mathrm{OSC}}\Delta N_{\mathrm{th}}} = \frac{2\pi^2\hbar\omega_o N_2\Delta\nu_c^2}{P_{\mathrm{OSC}}\Delta N_{\mathrm{th}}} \tag{133}$$

or

$$\Delta\nu_{\mathrm{OSC}} = \frac{\Delta\omega_{\mathrm{OSC}}}{2\pi} = \frac{\pi h\nu_o}{P_{\mathrm{OSC}}}\frac{N_2}{\Delta N_{\mathrm{th}}}\Delta\nu_c^2\,. \tag{134}$$

This fundamental result is known as the "Schawlow–Townes limit" for the linewidth of a laser oscillator [70]. Due to many other processes (described previously as "technical noise"), this ultimate linewidth is rarely achieved in practice but serves as an important benchmark for precision spectroscopy and metrology [8,9]. To get a feeling for the narrowness of the limit imposed by (134) let's assume a helium-neon laser ($\lambda = 632.8$ nm) with a 30 cm mirror spacing and an output coupler reflectivity of $R = 98\%$. We further assume that the lower energy level is always empty so that $\Delta N_{\mathrm{th}} \approx N_2$. Thus, the cold cavity linewidth is

$$\Delta\nu_c = \frac{1-R}{2\pi}\Delta\nu_{\mathrm{ax}} = \frac{(1-R)c}{4\pi L} \approx 1.6\,\mathrm{MHz} \tag{135}$$

and for 1 mW output power, the laser linewidth is

$$\Delta\nu_{\mathrm{OSC}} = \pi\frac{hc\Delta\nu_c^2}{\lambda_o P_{\mathrm{OSC}}} = 2.5\times 10^{-3}\,\mathrm{Hz}\,. \tag{136}$$

This theoretical linewidth is always overwhelmed by acoustic, thermal and other perturbations on the laser cavity and typically the actual measured linewidth is of order 10^4–10^6 Hz [9].

When lasers are used to probe narrow atomic or molecular transitions, it is often the laser linewidth that sets the ultimate resolution in the frequency domain. In practice, the transitions of greatest interest often challenge the performance of the probing lasers and this has been responsible for many advances in narrow-linewidth lasers. For example, by stabilizing the cavity length of a laser against a separate high-Q reference cavity, linewidths at the ≈ 1 Hz level have been achieved [71,72,73].

Shot Noise

The calculation of linewidth in the preceding section gives an idea of the limitations of a laser field when considering its use as a frequency reference or measurement tool. There is another type of uncertainty that enters into optical metrology that occurs when we actually measure the power (or intensity; power-per-unit area) in the field. When measuring the power an optical field we are, in effect, counting photons. This is because optical detectors rely on the conversion of field energy into quantized charge units (e. g. electrons). The photogeneration of electrons is therefore a necessarily quantized process since energy is removed from the radiation field in quantized units

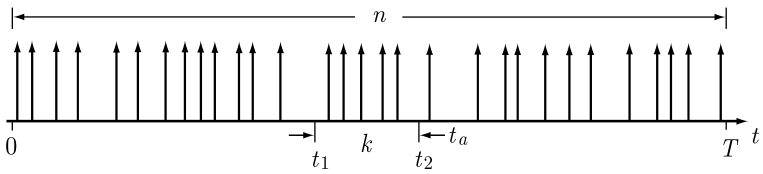

Noise and Stability in Modelocked Soliton Lasers, Figure 19

Random sequence of photoelectrons generated in a detector exposed to a laser field. If it is known that *n* events are randomly distributed over the interval *T*, the probability of finding *k* events in the time interval $t_a = t_2 - t_1$ is given by the Poisson distribution. Note that the average *rate* of events is $\bar{N} \equiv n/T$

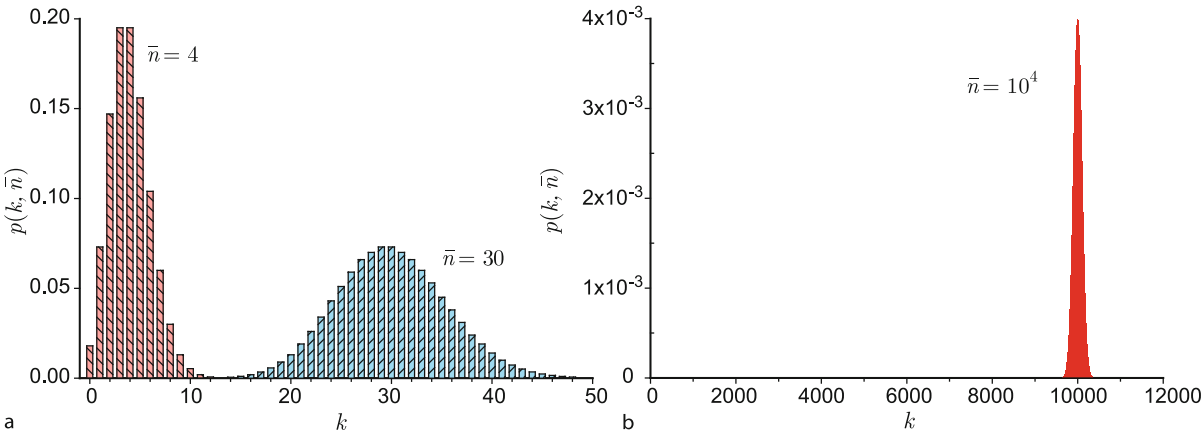

Noise and Stability in Modelocked Soliton Lasers, Figure 20

Poisson probability distributions for several cases of mean photon numbers. a *Left*; $\bar{n} = 4$ and *Right*; $\bar{n} = 30$. Notice the tendency toward Gaussian shape as $\bar{n}$ increases. b For even higher values of $\bar{n}$, (e. g. $\bar{n} = 10^4$) width of the probability distribution becomes a decreasing fraction of the mean value indicating an improvement in signal-to-noise ratio with $\bar{n}$

sufficient to liberate the charge carrier. Furthermore, the generation of a charge carrier is completely uncorrelated with the generation of any previous charge carriers. That is, the events occur independently of each other. Figure 19 depicts the generation of electrons due to photons hitting a detector. If we make many measurements during a time interval $t_a = t_2 - t_1$ we will find, on average, $\bar{n}$ photoelectrons but each individual measurement will yield a number k that will deviate from $\bar{n}$ by an amount given by the statistics of the process. The statistics of the generation of photoelectrons is that of a Poisson process [68,74,75] and we will derive several features of it now.

Suppose that n independent random events occur in a time interval T, as shown in Fig. 19. We are interested in finding the probability that k such events occur in a subinterval $t_a = t_2 - t_1$. This is a problem in Bernoulli trials and combinatorics [68,69] with the result that

$$p_n(k) = \text{Probability}\{k \text{ events in interval } (t_1, t_2)\} = \frac{n!}{k!\,(n-k)!} p^k q^{n-k} = \binom{n}{k} p^k q^{n-k} \tag{137}$$

where

$$p \equiv \frac{t_2 - t_1}{T} \tag{138}$$

is the probability of a single event in the interval $t_2 - t_1$ if only one event occurs during the longer time T, and $q = 1 - p$.

The value of k that gives the maximum probability of $p_n(k)$ is called the "most likely" value of k and is found to be about np [68]. Noting that n events occur in time T, we remark that the average rate of occurrences in time T is $n/T \equiv \bar{N}$ events/second and therefore we may interpret

$$np = \frac{n}{T}(t_2 - t_1) = \bar{N}(t_2 - t_1) \equiv \bar{n} \tag{139}$$

i. e. the *average* number of events in the interval $t_2 - t_1$.

Asymptotic Forms: Limit of Few Events If the most likely value of $k = np$ is of order $np \approx 1$, then a useful approximation to (137) is given by Poisson's theorem [76]

$$\binom{n}{k} p^k q^{n-k} \simeq e^{-np} \frac{(np)^k}{k!} \tag{140}$$

and since we have identified np as the average number of events in the interval $(t_2 - t_1)$, we may write this in the simpler form

$$p(k, \bar{n}) = \mathrm{e}^{-\bar{n}} \frac{\bar{n}^k}{k!} . \tag{141}$$

This is the *Poisson Probability Distribution* and is show in Fig. 20 for several interesting cases. In Fig. 20a, we see two sets of distributions for mean photon numbers $\bar{n} = 4$ and $\bar{n} = 30$, respectively. Clearly as the mean photon number increases, the width of the distribution also increase while the peak probability is reduced. This is consistent with the notion that the sum of probabilities over all outcomes must be unity;

$$\sum_{k=0}^{\infty} \mathrm{e}^{-\bar{n}} \frac{\bar{n}^k}{k!} = \mathrm{e}^{-\bar{n}} \sum_{k=0}^{\infty} \frac{\bar{n}^k}{k!} = \mathrm{e}^{-\bar{n}} \mathrm{e}^{\bar{n}} = 1 \tag{142}$$

or, physically, the probability that k will take on some value between 0 and ∞ is 1.

What is not quite as obvious from the two distributions in Fig. 20a is that the width of the distribution *relative to its mean* is going down as the mean increases. This is shown dramatically in Fig. 20b for the case of $\bar{n} = 10^4$. Physically, as we detect a larger number of photons, the relative uncertainty is decreasing. Or, the signal-to-noise ratio improves with the average flux of photons. We will investigate this in a little more detail shortly.

One final remark about the Poisson distribution is the apparent approach to a Gaussian shape as the mean photon number increases. This is not merely a trick of the eye but has its roots in a theorem quoted in the next section.

Asymptotic Forms: Limit of Many Events If the total number of events n in the interval T is large and so too is the target interval $(t_2 - t_1)$ such that $npq \gg 1$, then in the vicinity of the most likely value of the distribution where $|k - np| \leq \sqrt{npq}$ the theorem of DeMoivre and Laplace shows that the following approximation is valid [76]

$$p_n(k) = \binom{n}{k} p^k q^{n-k} \simeq \frac{1}{\sqrt{2\pi npq}} \mathrm{e}^{-(k-np)^2/2npq} . \tag{143}$$

That is, the Poisson distribution tends toward a Gaussian with mean value np and standard deviation $\sigma = \sqrt{npq}$.

The Moments The important statistical properties of a random process are given by their moments. The nth moment of the random variable k is defined as the expectation value of k^n; that is $E\{k^n\}$. The first moment is the "mean" and the second moment is the "mean-square" and so on. The central moments are similar except that we subtract the mean value before finding the expectation. The nth central moment is given by $E\{(k - \bar{k})^n\}$ and clearly equals zero for $n = 1$ but the case $n = 2$ is so important and useful it is given the special name "variance". In effect, the variance $\sigma^2 \equiv E\{(k - \bar{k})^2\}$ is a measure of the mean-square fluctuation of the random variable and its square root, σ, is called the "standard deviation". These important quantities are straightforward to calculate and are tabulated in Table 4.

Noise and Stability in Modelocked Soliton Lasers, Table 4
Important moments for the Poisson probability distribution function $p(k, \bar{n})$

PROBABILITY	MEAN	MEAN-SQUARE	VARIANCE	STD. DEVIATION
$p(k, \bar{n})$	$\bar{k}$	$\overline{k^2}$	$\sigma^2 = \overline{(k - \bar{n})^2}$	$\sigma = \sqrt{\overline{(k - \bar{n})^2}}$
$= \mathrm{e}^{-\bar{n}} \frac{\bar{n}^k}{k!}$	$\bar{n}$	$\bar{n}^2 + \bar{n}$	$\bar{n}$	$\sqrt{\bar{n}}$

From the moments we can immediately say something about measurement uncertainty. If we make a large number of measurements of the number of photons k in an interval t_a, then the standard deviation is a measure of the width of the distribution of the measured values. The ratio of this width to the average is a measure of the uncertainty in the measurement

$$\text{Uncertainty in } k = \frac{\sigma}{\bar{k}} = \frac{\sqrt{\bar{n}}}{\bar{n}} = \frac{1}{\sqrt{\bar{n}}} . \tag{144}$$

The uncertainty in the measurement improves (is reduced) as the square root of the average number of detected photons.

The Power Spectrum Since the photodetection process produces events that are randomly distributed in time, it does not possess a valid transform. However, a photodetector will produce a physical current when driven by the process and thus power will be transferred to a load. There are many cases when a small, coherent signal will be embedded in the optical field which will have a narrow spectral feature which can be discriminated against the random noise background. It thus is reasonable to search for an expression that describes the spectral distribution of the power, since there are fluctuations, and this is what is contained in the *power spectral density* function, which we calculate now.

The Wiener–Khinchine theorem [69] states that the power spectrum of a random process is equal to the Fourier transform of its autocorrelation. Thus, we need to

find the autocorrelation function of the process describing the photon detection. If t_i are the random points in time when a photon is detected, then the process of detected photons is described by

$$\mathbf{z}(t) = \sum_j \delta(t - t_j) \tag{145}$$

where $\delta(t)$ is the Dirac delta function defined so that

$$\int_{-\infty}^{\infty} \delta(t)\,\mathrm{d}t = 1 \tag{146}$$

and the points t_j follow the Poisson distribution (141). The autocorrelation function for this process is [68]

$$R_{\mathbf{z}}(\tau) = E\{\mathbf{z}(t)\mathbf{z}(t+\tau)\} = \bar{N}^2 + \bar{N}\,\delta(\tau)\,. \tag{147}$$

We now take the Fourier transform to find the power spectrum

$$S_{\mathbf{z}}(\nu) = \mathcal{F}\{R_{\mathbf{z}}(\tau)\} \tag{148}$$

$$= \int_{-\infty}^{\infty} R_{\mathbf{z}}(\tau)\mathrm{e}^{-\mathrm{i}2\pi\nu\tau}\,\mathrm{d}\tau \tag{149}$$

$$= \int_{-\infty}^{\infty} \left[\bar{N}^2 + \bar{N}\,\delta(\tau)\right]\mathrm{e}^{-\mathrm{i}2\pi\nu\tau}\,\mathrm{d}\tau \tag{150}$$

$$= \bar{N}^2\delta(\nu) + \bar{N} \quad (-\infty \le \omega \le \infty)\,. \tag{151}$$

In the last line we emphasize that this is a *double-sided* power spectrum. Therefore, to find the total power passing through a rectangular filter of bandwidth B, centered at ν_o, away from $\nu = 0$, we would multiply the spectral density by 2. That is, $P(\nu_\mathrm{o}) = 2S_{\mathbf{z}}(\nu_\mathrm{o})B = 2B\bar{n}$.

We remark here about a subtle issue regarding the frequency domain variable $\omega \equiv 2\pi\nu$ with units of radians/second. Almost all instrumentation and discussions of frequency-domain phenomena use the more common variable ν with units of Hertz or, the historical predecessor "cycles/second". In any case, we can define a spectral density function using either variable but the integrals over the relevant domains must yield the same power. Therefore, since $\mathrm{d}\omega = 2\pi\,\mathrm{d}\nu$, we have

$$\int_{-\infty}^{\infty} S(\omega)\,\mathrm{d}\omega = \int_{-\infty}^{\infty} S(\omega)\,2\pi\,\mathrm{d}\nu = \int_{-\infty}^{\infty} S(\nu)\,\mathrm{d}\nu \tag{152}$$

and thus

$$S(\nu) = 2\pi S(\omega)\,. \tag{153}$$

Photocurrent Suppose the photodetection process yields a single electron for every incident photon. Then we can describe the current from the photodetector in terms of the random process (145) as

$$i(t) = q\sum_j \delta(t - t_j)\,. \tag{154}$$

The average value of the current measured over a time interval T is simply

$$\overline{i(t)} = \frac{q}{T}\int_{-T/2}^{T/2} \delta(t - t_j)\,\mathrm{d}t = q\,\frac{\bar{n}}{T} = q\bar{N} \equiv i_\mathrm{avg} \tag{155}$$

since, by definition, there are on average, $\bar{n}$ events detected in a time interval T.

Using our previous results for the autocorrelation function and the power spectrum of the random process described by Poisson statistics, we have for the photocurrent

$$R_i(\tau) = E\{i(t)i(t+\tau)\} = q^2\left[\bar{N}^2 + \bar{N}\,\delta(\tau)\right] \tag{156}$$

and

$$|\mathcal{I}(\nu)|^2 = \mathcal{F}\{R_i(\tau)\} = q^2\left[\bar{N}^2\delta(\nu) + \bar{N}\right] \quad \mathrm{A^2/Hz} \tag{157}$$

valid for $(-\infty \le \nu \le \infty)$. Thus, we see that the power spectrum is comprised of two principal features. The delta function at DC ($\nu = 0$) is the power due to the average photocurrent. The second term, with uniform (two-sided) spectral spectral density, describes the power density due to the random process, or the noise. If the photocurrent drives a load resistor of value R_L, we can identify the average (or DC) power and the noise power within a bandwidth B as

$$P_\mathrm{DC} = q^2\bar{N}^2R_\mathrm{L} = i_\mathrm{avg}^2R_\mathrm{L} \quad \mathrm{W} \tag{158}$$

$$P_\mathrm{NOISE} = 2q^2\bar{N}BR_\mathrm{L} = 2q\,i_\mathrm{avg}BR_\mathrm{L} \quad \mathrm{W}\,. \tag{159}$$

Now we have expressions from which we can determine a signal-to-noise ratio for the detected photoelectrons. If we consider the DC photocurrent to be the desired measured quantity, then the signal-to-noise ratio (SNR) for shot-noise-limited detection is given by

$$\mathrm{SNR} = \frac{P_\mathrm{DC}}{P_\mathrm{NOISE}} = \frac{i_\mathrm{avg}}{2qB} \tag{160}$$

where B is the bandwidth of the measuring device. If the desired signal is a modulation of the average photocurrent,

then it will show up in the spectrum as a coherent spike with power proportional to P_{DC} and the SNR will scale according to this proportionality. The bandwidth B is still that of the measuring instrument and often it is a narrowband filter centered at the frequency of the modulation in order to optimize the SNR.

Since it is often the case that the load resistance and bandwidth of the following circuitry are flexible parameters, it is convenient to note that the single-sided ($\nu \geq 0$) *mean-squared noise current spectral density* is simply

$$I_{\mathrm{NOISE}}(\nu) \equiv 2q\, i_{\mathrm{avg}} \quad \mathrm{A}^2/\mathrm{Hz}\,. \tag{161}$$

Then, for purposes of circuit modeling and analysis, we can include a signal generator with rms current equal to $i_{\mathrm{rms}} = \sqrt{2qi_{\mathrm{avg}}}\ \mathrm{A}/\sqrt{\mathrm{Hz}}$.

Analytical Description of Envelope Noise

The noise that we will be considering in this section is that pertaining to the amplitude and phase of the laser *intensity envelope*. Instability of the optical carrier as described in Sect. "Laser Noise and Linewidth" is, indeed, an important phenomenon but it is assumed that the envelope of the electric field (and thus the intensity) carries much greater relevance in the case of modelocked lasers since we normally detect power. Furthermore, modelocking imparts a certain degree of phase coherence, or cooperation, among the many longitudinal cavity modes that make up an ultrashort pulse and it becomes rather nebulous as to what is meant by "carrier" amplitude and phase noise. So, we focus our attention on the stability of the modelocked pulse train; both amplitude and phase.

Suppose we model the laser pulse train as a periodic sequence of Gaussian pulses

$$I(t) = \sum_n I_{\mathrm{o}} \exp\left[-4\ln 2\left(\frac{t-nT}{\tau}\right)^2\right] \tag{162}$$

where I_{o} is the peak power, T is the period and the scaling factor $4\ln 2$ ensures that the pulsewidth τ represents the full-width at half-maximum (FWHM) of the pulse intensity. The form shown in (162) represents an ideal train of pulses, free of amplitude or timing fluctuations. In order to see the effects of random fluctuations in pulse amplitude and period, we take the point of view that any wideband noise process can be decomposed into narrowband slices which are quasi-coherent on a time scale that is roughly the inverse of the bandwidth of the spectral slice under study [15]. We can then introduce these effects into the model as pure sinusoids by writing

$$I(t) = \sum_n I_{\mathrm{o}}\,(1 + m\cos\omega_{\mathrm{am}}t) \cdot \exp\left[-4\ln 2\left(\frac{t-[n+(\beta/2\pi)\cos(\omega_{\mathrm{pm}}t)]T}{\tau}\right)^2\right]. \tag{163}$$

Here we see that the period is being modulated at a rate ω_{pm} and the amplitude at a rate ω_{am}. The coefficients m and β are the amplitude- and phase-modulation indices, respectively. β is also frequently referred to as the "peak phase deviation". When $\beta = 2\pi$ the pulse train has undergone one complete period of timing modulation while $m = 1$ causes 100% amplitude modulation. These two effects are shown graphically in Fig. 21. For purposes of il-

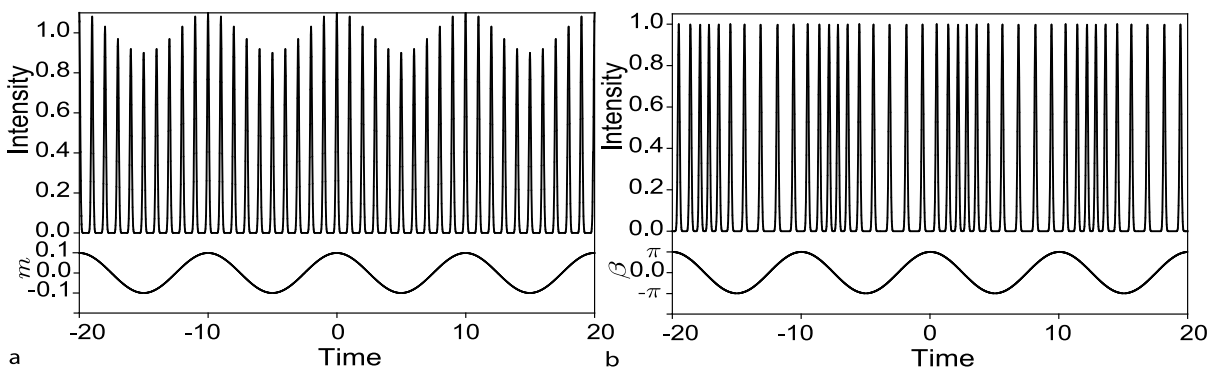

Noise and Stability in Modelocked Soliton Lasers, Figure 21

a Train of optical pulses from a modelocked laser undergoing pure sinusoidal AM modulation with a modulation index of 10% ($m = 0.1$). b Optical pulse train undergoing pure PM modulation with a peak phase deviation of π radians ($\beta = \pi$). Modulation waveform shown in bottom of each figure

lustration, values of $m = 0.1$ and $\beta = \pi$ were chosen and the modulation frequencies are 10% of the pulse repetition rate. For real sources, typically $m \ll 1$ and $\beta \ll 2\pi$. Also, the modulation rates ω_{am} and ω_{pm} are almost always much less than the laser pulse repetition rate. That is, $\omega_{\mathrm{am}}, \omega_{\mathrm{pm}} \ll 2\pi/T$.

Fourier Analysis of Sinusoidal Amplitude Modulation

The effects of amplitude and phase modulation on an infinite pulse train produce markedly different effects, yet, it turns out that if we simply measure the power spectrum associated with the two cases, the spectra will often be indistinguishable, especially for small modulation depths. To see why this is so and to understand the necessity for experimental tools that can differentiate between them, we analyze the modulation effects on each of the Fourier components in (163) using the phasor representation. Consider the effect of AM modulation first. We expand the sequence of Gaussian pulses in a Fourier series assuming that the pulsewidth is much shorter than the period, $\tau \ll T$ (as is always the case for modern modelocked lasers)

$$I_{\mathrm{AM}}(t) = I_{\mathrm{o}}\,(1 + m\cos\omega_{\mathrm{am}}t) \cdot \sum_n \exp\left[-4\ln 2\left(\frac{t-nT}{\tau}\right)^2\right] \tag{164}$$

$$= I_{\mathrm{o}}\,(1 + m\cos\omega_{\mathrm{am}}t) \cdot \frac{\tau}{T}\sqrt{\frac{\pi}{4\ln 2}}\left[\frac{1}{2} + \sum_n \cos\left(2\pi nt/T\right)\right] \tag{165}$$

$$= I_{\mathrm{o}}'\left\{\frac{1}{2} + \frac{m}{2}\cos(2\pi f_{\mathrm{am}}t) + \sum_{n=1}^{\infty}\cos(2\pi nt/T) + \frac{m}{2}\left[\cos 2\pi\left(\frac{n}{T} - f_{\mathrm{am}}\right) + \cos 2\pi\left(\frac{n}{T} + f_{\mathrm{am}}\right)\right]\right\}. \tag{166}$$

This form illuminates the important components produced by amplitude modulation. The first term in (166) is the average power carried by the pulse train and the second term is the modulation of that average power. The last three terms comprising the infinite series are harmonics of the fundamental repetition rate (n/T) along with sidebands adjacent to the harmonics due to the modulation. In the phasor representation of (166), every time-harmonic term is taken as the real part of the corresponding complex function which allows us to separate out the modulation sidebands, e. g.

$$\cos(2\pi nt/T) = \mathrm{Re}\left\{\mathrm{e}^{\mathrm{i}n\omega_{\mathrm{o}}t}\right\} \quad \text{CARRIER} \tag{167}$$

$$\cos 2\pi\left(\frac{n}{T} - f_{\mathrm{am}}\right) = \mathrm{Re}\left\{\mathrm{e}^{\mathrm{i}(n\omega_{\mathrm{o}}+\omega_{\mathrm{am}})t}\right\} \quad \text{UPPER AM SIDEBAND} \tag{168}$$

$$\cos 2\pi\left(\frac{n}{T} + f_{\mathrm{am}}\right) = \mathrm{Re}\left\{\mathrm{e}^{\mathrm{i}(n\omega_{\mathrm{o}}-\omega_{\mathrm{am}})t}\right\} \quad \text{LOWER AM SIDEBAND}\,. \tag{169}$$

The *phasor amplitudes* are the complex coefficients on the harmonic term $\exp(\mathrm{i}\omega_{\mathrm{o}}t)$. That is, the upper sideband has a phasor amplitude $\mathrm{e}^{+\mathrm{i}\omega_{\mathrm{am}}t}$ and the lower sideband has amplitude $\mathrm{e}^{-\mathrm{i}\omega_{\mathrm{am}}t}$. The phasor picture provides a very useful way of viewing the modulation processes. If we plot the phasor amplitudes in the complex plane in the rotating reference frame of the carrier, we see immediately how the addition of the upper and lower AM sidebands to the carrier effect amplitude modulation. This is shown in Fig. 22a.

Fourier Analysis of Sinusoidal Phase Modulation

We can now also analyze the phase modulation effect in a similar way. Let

$$I(t) = \sum_n I_{\mathrm{o}} \cdot \exp\left[-4\ln 2\left(\frac{t - [n + (\beta/2\pi)\cos(\omega_{\mathrm{pm}}t)]T}{\tau}\right)^2\right]. \tag{170}$$

We wish to write a Fourier series for this representation but the actual period of the pulses is based on the modulation frequency, ω_{pm} and not the repetition rate $1/T$. This makes the result cumbersome and difficult to work with. In almost all cases of interest, however, the modulation frequencies of the noise processes contributing to laser timing fluctuations are very slow compared with the pulse repetition rate, so we can approximate the time shift $(\beta/2\pi)\cos(\omega_{\mathrm{pm}}t)T$ as being static for the purposes of calculating the Fourier series. If we let $\omega_{\mathrm{pm}}t \equiv \Phi$ we may express $I(t)$ in a complex Fourier series

$$I(t) = \sum_{n=-\infty}^{\infty} c_n \mathrm{e}^{\mathrm{i}2\pi(n/T)t} \tag{171}$$

where

$$c_n = \frac{1}{T}\int_{-T/2}^{T/2} I(t)\mathrm{e}^{-\mathrm{i}2\pi(n/T)t}\,\mathrm{d}t \tag{172}$$

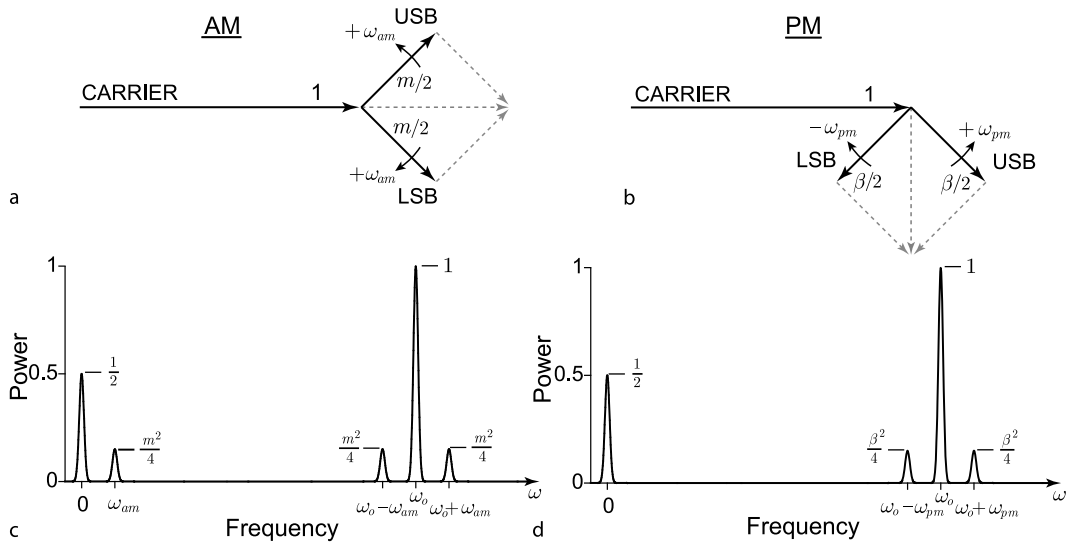

Noise and Stability in Modelocked Soliton Lasers, Figure 22

Relationships between sidebands and power spectrum for amplitude (AM) and phase (PM) modulated waveforms. a In amplitude modulation the upper (USB) and lower (LSB) modulation sidebands spinning at a rate $\pm\omega_{am}$ align in phase with the carrier and their superposition adds and subtracts from the carrier amplitude. b In phase modulation the upper (USB) and lower (LSB) modulation sidebands spinning at a rate $\pm\omega_{pm}$ align in quadrature with the carrier and their superposition periodically advances and retards the phase of the carrier. c Power spectrum of a pulse sequence undergoing AM only showing the baseband and fundamental spectral components. d Power spectrum of a pulse sequence undergoing PM only showing the baseband and fundamental spectral components. Note that there are no modulation spectral components at baseband and thus measurements of PM must be made about the fundamental ω_o (or any harmonic)

$$= \frac{I_o}{T}\int_{-T/2}^{T/2} \exp\left[-4\ln 2\left(\frac{t-(\beta T/2\pi)\cos\Phi}{\tau}\right)^2\right] \cdot e^{-i2\pi(n/T)t}\,dt \quad (173)$$

$$= I_o\frac{\tau}{T}\sqrt{\frac{\pi}{4\ln 2}}\exp\left(-\frac{(\pi(n/T)\tau)^2}{4\ln 2}\right)e^{-in\beta\cos\Phi} \quad (174)$$

$$\equiv I_o' e^{-in\beta\cos\Phi}\,. \quad (175)$$

Notice that the Gaussian term in (174) does not differ appreciably from unity until $n \approx T/\tau$. Since we seldom investigate harmonics of this high order, we can ignore the dependence on n and treat it as a constant. Substituting this result back into (171) and returning the value of $\Phi = \omega_{pm}t$ yields

$$I(t) = I_o'\sum_{n=-\infty}^{\infty} e^{i(2\pi(n/T)t-n\beta\cos\Phi)}$$
$$= I_o'\sum_{n=-\infty}^{\infty} e^{i2\pi(n/T)t}\,e^{-in\beta\cos\omega_{pm}t}\,. \quad (176)$$

The second complex exponential can be rewritten in terms of the Bessel functions $J_n(x)$ [77] as

$$e^{-in\beta\cos\omega_{pm}t} = \sum_{k=-\infty}^{\infty}(-i)^k J_k(n\beta)e^{ik\omega_{pm}t}\,. \quad (177)$$

The Bessel functions are tabulated in many books [78] and have the following symmetry properties

$$J_k(\beta) = J_{-k}(\beta), \quad J_k(-\beta) = J_k(\beta); \quad n \text{ even} \quad (178)$$

$$J_k(\beta) = -J_{-k}(\beta), \quad J_k(-\beta) = -J_k(\beta); \quad n \text{ odd}. \quad (179)$$

Thus

$$I(t) = I_o'\sum_{n=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}(-i)^k J_k(n\beta)e^{i(2\pi(n/T)+k\omega_{pm})t} \quad (180)$$

$$= I_o'\sum_{n=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}(-i)^k J_k(n\beta)e^{i(n\omega_o+k\omega_{pm})t} \quad (181)$$

where we have replaced $2\pi/T$ with ω_o, the radian fundamental frequency. This double series is actually easier to

interpret than it looks. The indices n and k set the oscillation frequencies of the time-harmonic function $e^{i\omega(n,k)}$. Since $1/T$ is the nominal repetition rate of the pulses, then n/T (or $n\omega_o$) is the frequency of the nth harmonic of the main Fourier series. The slow phase modulation rate ω_{pm} produces sidebands adjacent to the harmonics at frequencies $n\omega_o \pm k\omega_{pm}$. Let's write out a few of the important terms in the series;

$$\begin{aligned}
I(t) = I_o' \Big\{ & \cdots \\
& + e^{-i2\omega_o t}\Big[\cdots - J_{-2}(2\beta)e^{-i2\omega_{pm}t} - iJ_{-1}(2\beta)e^{-i\omega_{pm}t} \\
& \quad + J_0(2\beta) + iJ_1(2\beta)e^{i\omega_{pm}t} - J_2(2\beta)e^{i2\omega_{pm}t} + \cdots\Big] \\
& + e^{-i\omega_o t}\Big[\cdots - J_{-2}(\beta)e^{-i2\omega_{pm}t} - iJ_{-1}(\beta)e^{-i\omega_{pm}t} \\
& \quad + J_0(\beta) + iJ_1(\beta)e^{i\omega_{pm}t} - J_2(\beta)e^{i2\omega_{pm}t} + \cdots\Big] + 1 \\
& + e^{i\omega_o t}\Big[\cdots - J_{-2}(\beta)e^{-i2\omega_{pm}t} + iJ_{-1}(\beta)e^{-i\omega_{pm}t} \\
& \quad + J_0(\beta) - iJ_1(\beta)e^{i\omega_{pm}t} - J_2(\beta)e^{i2\omega_{pm}t} - \cdots\Big] \\
& + e^{i2\omega_o t}\Big[\cdots - J_{-2}(2\beta)e^{-i2\omega_{pm}t} + iJ_{-1}(2\beta)e^{-i\omega_{pm}t} \\
& \quad + J_0(2\beta) - iJ_1(2\beta)e^{i\omega_{pm}t} - J_2(2\beta)e^{i2\omega_{pm}t} - \cdots\Big] \\
& + \cdots \Big\} .
\end{aligned} \tag{182}$$

We see that all of the phase modulation information is reproduced about each harmonic, but from a practical standpoint, the terms of greatest interest in the series are those associated with the DC or baseband ($n = 0$) and the fundamental ($n = \pm 1$). Notice that at baseband there is no modulation information. That is, the series returns a constant (+1) for $n = 0$. There is no spectral energy corresponding to the timing jitter near 0 Hz, but, the power in that DC term is proportional to the average laser power and thus is still useful for normalization purposes. In contrast, the fundamental and all harmonics have modulation sidebands and spectrum analyzers tuned to these components can determine the degree of timing jitter. Let's look a little more closely at the fundamental and its associated sidebands. Using the symmetry properties (178), (179) and the small-argument approximations

$$J_0(x) \approx 1, \quad J_1(x) \approx x/2, \quad J_n(x) \approx 0, \quad |n| > 1 \tag{183}$$

we have

$$I_1(t) = I_o' \, e^{i\omega_o t}\Big[iJ_{-1}(\beta)e^{-i\omega_{pm}t} + J_0(\beta) - iJ_1(\beta)e^{i\omega_{pm}t}\Big] \tag{184}$$

$$= I_o' \, e^{i\omega_o t}\Big[1 - i\frac{\beta}{2}\left(e^{i\omega_{pm}t} + e^{-i\omega_{pm}t}\right)\Big] . \tag{185}$$

Notice that, unlike the case of the AM sidebands in (168), (169), the PM sidebands are in phase-quadrature with the carrier, owing to the presence of the imaginary term i– preceding the sidebands. We can absorb this into the sideband phase and write

$$I_1(t) = I_o'\Bigg[\underbrace{e^{i\omega_o t}}_{\text{CARRIER}} + \underbrace{\frac{\beta}{2}e^{i(\omega_{pm}t-\pi/2)}}_{\text{UPPER PM SIDEBAND}} + \underbrace{\frac{\beta}{2}e^{-i(\omega_{pm}t+\pi/2)}}_{\text{LOWER AM SIDEBAND}}\Bigg] . \tag{186}$$

A pictorial representation of the carrier and sideband phasors is shown in Fig. 22b. Notice that in the limit of very small modulation index ($\beta \ll 2\pi$) the superposition of the upper and lower-sideband phasors only affect the phase angle of the carrier and not the amplitude since they are in phase-quadrature with respect to the carrier.

Modulation Sideband Power

From a measurement standpoint, the most important attribute of the series expansions for AM and PM are the powers associated with the modulation sidebands. Because noise is a random process, we will be able to profitably discuss the power spectral density adjacent to each harmonic component in the series and we would like to be able to relate this to the relative magnitude of the noise modulation. This is most easily done by quantifying the ratio of the power in a single sideband to that in the carrier (usually the fundamental). Figure 22c shows the theoretical power contained in a purely AM-modulated periodic waveform of the type in Fig. 21a normalized to the fundamental carrier power. Figure 22d shows the same spectrum for the case of a purely PM-modulated periodic waveform of the type in Fig. 21b again normalized to the fundamental carrier power.

We can summarize the important relationships between the modulation types and their effects as in Table 5.

Noise and Stability in Modelocked Soliton Lasers, Table 5

Principal results of AM and PM modulation applied to a periodic sequence of short laser pulses. The Fourier analysis of the pulse train is applied to the laser pulse *intensity* rather than the electric field since a photodetector produces a current that is linearly proportional to the laser power. The indicated power ratios are therefore those that would be observed on an RF spectrum analyzer connected to a photodiode which is illuminated by the optical pulse train

MODULATION TYPE	MODULATION INDEX	FREQUENCY	PEAK FLUCTUATION	POWER RATIOS	
				BASEBAND	CARRIER
AMPLITUDE (AM)	m	ω_{am}	$\Delta I_{PK} = mI_o$	$\frac{P_{mod}}{P_{avg}} = \frac{m^2}{2}$	$\frac{P_{SSB-AM}}{P_c} = \frac{m^2}{4}$
PHASE (PM)	β	ω_{pm}	$\Delta t_{PK} = \frac{\beta}{2\pi}T$	N/A	$\frac{P_{SSB-PM}}{P_c} = \frac{\beta^2}{4}$

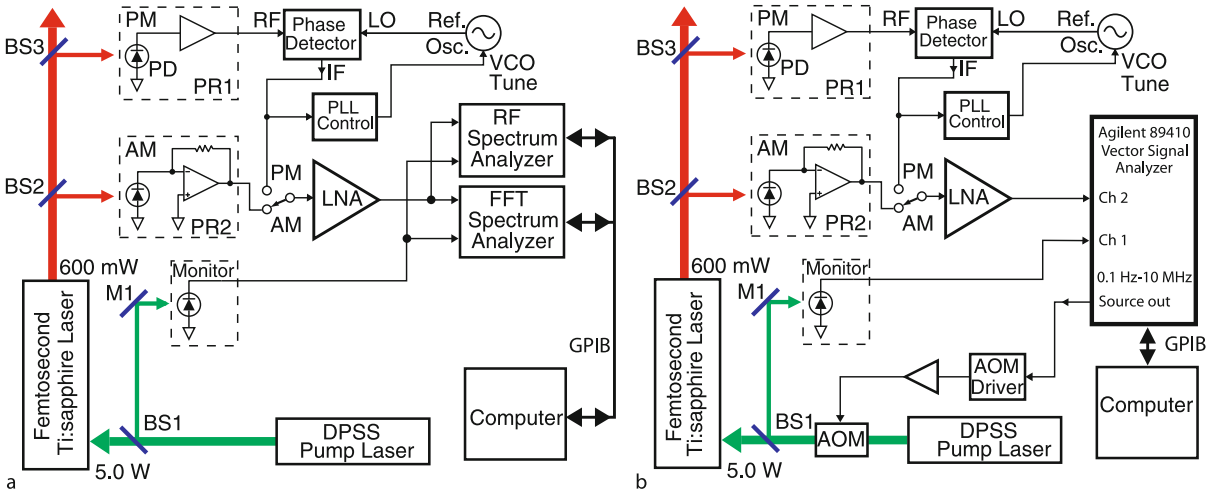

Noise and Stability in Modelocked Soliton Lasers, Figure 23

a System for measuring the amplitude and envelope phase noise of a modelocked Ti:sapphire laser and the amplitude noise of the pump laser. b Measurement system for characterizing the *complex* noise transfer function using a vector network analyzer in place of the spectrum analyzers. Legend: PR1; photoreceiver for phase noise, PR2; photoreceiver for amplitude noise, RF; radiofrequency input, LO; local oscillator or reference input, IF' intermediate (difference) frequency output, LNA; low noise amplifier, BS1, BS2, BS3; beamsplitters, M1; mirror, DPSS; diode-pumped solid-state pump laser, GPIB; general purpose interface bus, PLL; phase-locked loop, VCO; voltage-controlled oscillator, AOM; acousto-optic modulator

Measurement of Laser Amplitude and Phase Noise

Making accurate, high dynamic range noise measurements of lasers can be challenging [79]. To accurately quantify the noise associated with amplitude and timing fluctuations of a periodic pulse train we rely on the use of Fast Fourier Transform (FFT) and radiofrequency (RF) spectrum analyzers applied to the detected photocurrent produced by a photodiode illuminated with the laser under study. These instruments are, essentially, tunable narrowband frequency filters that reveal the power spectral density contained in a narrow filter bandwidth at a specific frequency. One might be tempted to argue that using a photodiode and oscilloscope would be adequate but a quick calculation indicates that in order to see a wide-enough spectrum of noise on the oscilloscope, we are limited by thermal noise to observing a noise floor of several millivolts and hence the dynamic range is of order 60 dB (for a 1 volt peak signal). this amounts to an optical power fluctuation of approximately 0.1% which is enormous for a "well-behaved" laser. In fact, as we shall see, we need a dynamic range approaching 160 dB, or better, to learn anything meaningful about laser noise.

Figure 23a shows a schematic diagram of a system for measuring the amplitude and phase noise of the continuous train of pulses from a modelocked Ti:sapphire laser [79]. This laser is pumped from a diode-pumped solid-state laser (DPSS). In this system both the pump amplitude (AM) noise and the Ti:sapphire amplitude and phase (PM) noise are measured. The amplitude noise of either laser is manifested as fluctuations in the average power so only a relatively slow photodiode and receiver

are required. The amplitude noise is measured directly in the frequency domain by the spectrum analyzers. The FFT analyzer covers the frequency range from 0.1 Hz to 25 KHz and the RF spectrum analyzer from 10 Hz to 40 MHz. The frequency overlap allows accurate calibration and matching of noise spectra. To measure the phase noise of the pulse train envelope, we must compare its phase stability against a reference source. This is done using a high-quality reference oscillator in a phase-locked loop circuit. The photodiode and amplifier associated with this portion of the circuit (PR1) are tuned to the fundamental repetition rate of the laser (e. g. 100 MHz). The resulting sinewave is mixed against the reference oscillator sinewave and the output of the mixer ($\approx$ 0 Hz) is fed back to the tuning control of the reference oscillator. This forces the oscillator to stay phaselocked to the laser's repetition rate. For timing fluctuations that occur at a rate faster than the loop can respond, the voltage at the output of the mixer is linearly proportional to the phase offset and the power spectrum of these phase fluctuations are measured by the spectrum analyzers. In short, the phase detector circuit converts phase noise into amplitude noise which is characterized by the spectrum analyzers.

Figure 24 shows a comparison of phase noise between several conventional RF sources and a free-running modelocked Ti:sapphire laser. The displayed phase noise is displayed to represent the noise power spectral density of phase fluctuations on one side of the carrier referenced to the carrier power in decibels per-unit-bandwidth or dBc/Hz. The HP 8640B is a cavity-tuned signal generator whose frequency is steered via a varactor diode (so-called "DCFM"). The HP 8662A is an RF frequency synthesizer employing many complicated analog phase-locked loops to provide 1 Hz resolution through the range 10 KHz to 1300 MHz. The 10 MHz crystal oscillator is a high-performance, temperature compensated unit that is frequency-multiplied to 80 MHz. The Ti:sapphire laser used in this measurement was running at 100 MHz and thus will display a slightly higher phase noise for the same timing fluctuation as it would if it were running at 80 MHz, the difference scaling as $(100/80)^2 = 1.56 = 1.94$ dB. From the data shown in a phase noise plot such as this, we can discern the average phase fluctuations at a given offset frequency of any source and compare them side-by-side. Modulation of the source phase, though a noise process, will show up as phase noise sidebands with magnitude given by the modulation index derived earlier in Subsect. "Modulation Sideband Power".

The format of the display of noise information is important and we should make a few remarks regarding this. Noise is a fundamentally statistical process and the noise that accompanies the laser electric field interferes with our ability to characterize it. By this we mean that given knowledge of the instantaneous electric field at any given mo-

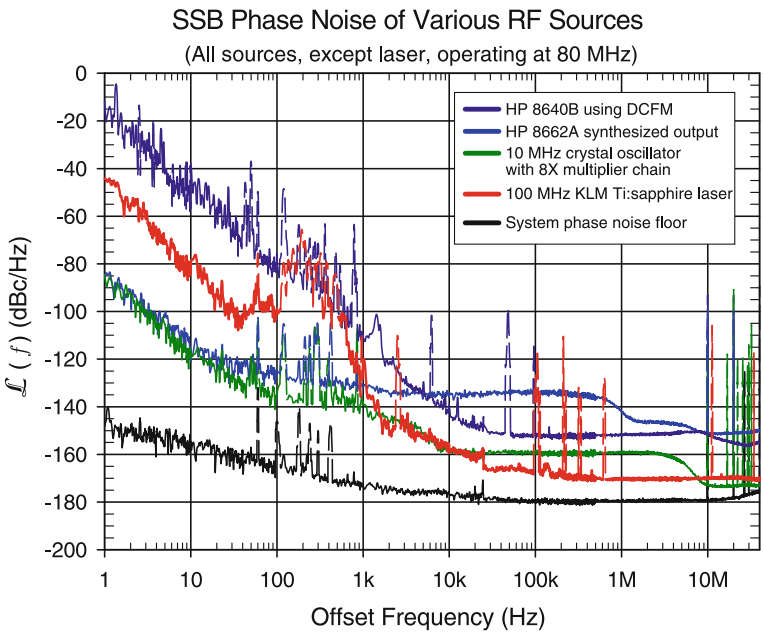

Noise and Stability in Modelocked Soliton Lasers, Figure 24
Comparison of phase noise performance of various sources

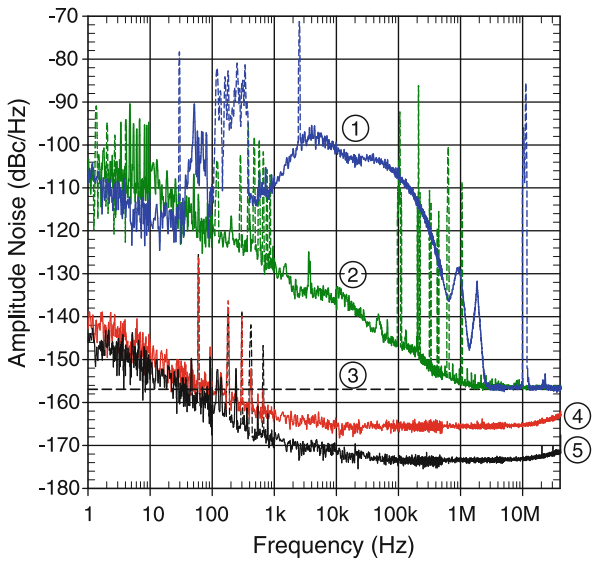

Noise and Stability in Modelocked Soliton Lasers, Figure 25
Power spectral density of amplitude noise for two different pump lasers. *1* Argon-ion laser, *2* Diode-pumped neodymium-doped yttrium orthovanadate laser (Nd:YVO_4). Both lasers producing 2.5 mA average photocurrent, *3* Calculated shot noise floor for $i_{avg} = 2.5$ mA, *4* Photoreceiver noise floor ($i_{avg} = 0$ mA), *5* Spectrum analyzer system noise floor. (From [79])

ment, we cannot say with certainly what the field will be at some specified time later. However, provided that the statistical properties of the noise are time-invariant, it can be shown that the *power spectrum* of the electric field will be stationary (unchanging). That is yet another reason why the spectrum analyzer is the appropriate tool for studying noise. It is, in effect, a tunable narrowband power meter, and thus the relevant characteristic of the noise is the power spectral density with units of Watts/Hertz. Relating the noise power-spectral-density to the fluctuating laser pulse train can best be understood using a deterministic model, which is valid on the time scale of the inverse bandwidth of the device measuring the fluctuation (in this case, the spectrum analyzer).

Figure 25 shows measurements of typical pump laser amplitude noise. The top curve (#1) is the noise spectrum of an argon-ion laser. The second curve (#2) is the noise spectrum of a diode-pumped, solid-state Nd:YVO_4 laser which is currently the preferred choice as a replacement pump source for pump Ti:sapphire lasers. Apart from the dramatic improvement in efficiency, this laser evidently has vastly superior noise performance. The photoreceiver and measurement system noise floors are shown in curves (#4) and (#5) respectively while the shot noise limit for the indicated photocurrent is shown as the straight line (#3).

Pump-Induced Noise and the Noise Transfer Function

The amplitude stability of the pump laser plays a very important role in the amplitude and timing stability of the modelocked laser which it pumps [80,81,82,83,84,85]. Figure 26 shows the effect pictorially. Here a Kerr-lens modelocked Ti:sapphire laser is pumped by a diode-pumped solid-state laser (DPSS) with a highly exaggerated amount of amplitude noise. The Ti:sapphire laser emits a stream of femtosecond-duration pulses which have both amplitude noise and timing instability (also shown highly exaggerated).

We can understand how there might be a direct relationship between pump noise and the pumped-laser noise by considering several of the possible mechanisms that couple pump amplitude to the amplitude and repetition rate of a modelocked laser. For example, the pump power level (and hence pumping rate) establishes the gain of the laser and once above threshold, the power output is directly tied to the pumping rate r (70). Thus, amplitude noise on the pump beam will induce amplitude noise on the modelocked laser beam. We call this process AM-to-AM noise transfer [85]. The mechanism for this type of noise transfer can be easily understood for slow fluctuations. Figure 27 shows a simplified plot of the laser output power P_L versus pump power P_P. Above threshold the output power is linearly related to the pump power as derived in Sect. "Laser Dynamics" (Eq. (70)). For slow or static changes in pump power, the operating point just shifts along the straight line above threshold. We typically define a "slope efficiency", η, of a laser above threshold as the slope of P_L vs. P_P at the operating point P_{P_0}. For the straight line shown in Fig. 27, we have

$$\eta = \left.\frac{dP_L}{dP_P}\right|_{P_P=P_{P_0}} = \frac{\Delta P_L}{\Delta P_P} . \tag{187}$$

If the pump modulation varies sinusoidally at a rate ω_m low enough such that the laser dynamics follow perfectly, we can define a modulation index for both the pump and the laser in terms of the peak fractional deviations in the powers,

$$m_P \equiv \frac{\Delta P_P}{P_{P_0}}, \quad m_L \equiv \frac{\Delta P_L}{P_{L_0}} . \tag{188}$$

Let us further define the AM noise transfer function, $H_{AM}(\omega_m)$ as the ratio of the induced modulation to the causative modulation. In the steady state ($\omega_m = 0$), we see the direct relationship to the slope efficiency by

$$H_{AM}(0) \equiv \frac{m_L}{m_P} = \eta \frac{P_{P_0}}{P_{L_0}} . \tag{189}$$

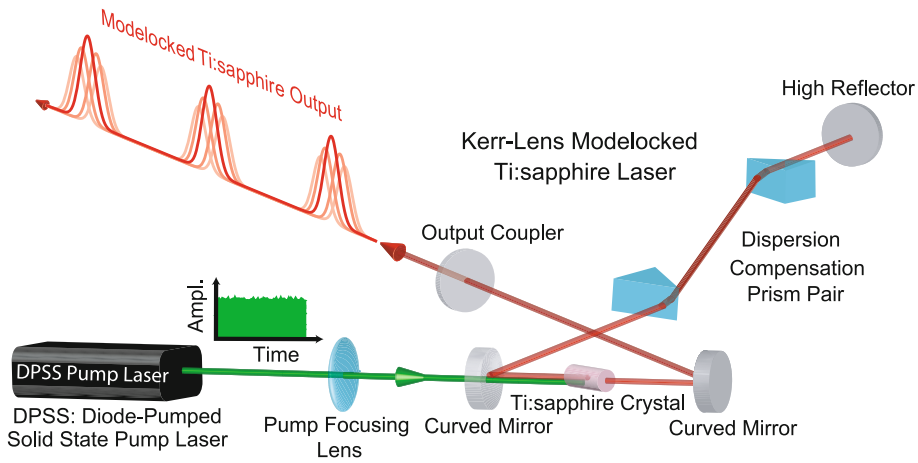

Noise and Stability in Modelocked Soliton Lasers, Figure 26
Pump laser noise contributing to amplitude noise and timing instability in a modelocked laser

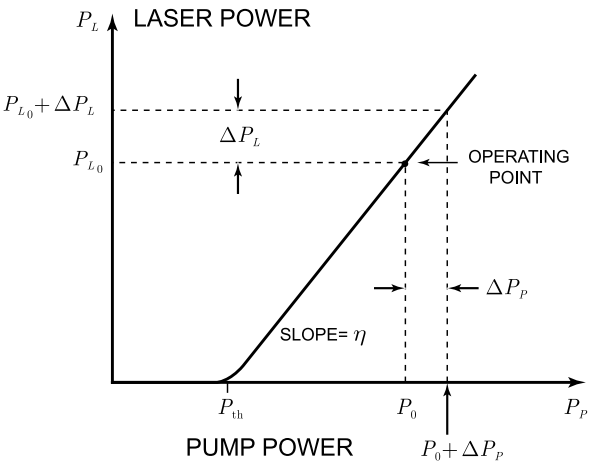

Noise and Stability in Modelocked Soliton Lasers, Figure 27
Laser output power P_L as a function of pump power P_P above threshold P_{th}

This relationship will serve as the model for the AM-to-PM noise transfer function as well. That is, both noise transfer functions relate the induced effect (amplitude or timing fluctuation) to the pump modulation. In addition, (189) will also serve as a calibration or verification point for measurements of $H_{AM}(\omega)$, over a wide range of modulation frequencies, when the pump is intentionally modulated to simulate a noise component.

In addition to inducing amplitude noise in the modelocked laser, pump noise will also cause timing jitter or envelope phase noise in the modelocked pulse train. The relationship between the two is characterized by the phase noise transfer function $H_{PM}(\omega)$, which will be developed shortly. The sensitivity of the repetition-rate of the modelocked laser to pump fluctuations has many potential contributing factors. Below threshold the population levels change in direct proportion to the pumping rate (67) but above threshold, when the laser is actually oscillating, the population inversion is clamped (69) and therefore the index of refraction is fixed. However, this is only true in the steady-state. As we have seen, if the pump is changed abruptly, the population levels cannot follow and take a while to settle down (Fig. 9). This means that for pump fluctuation frequencies that are high enough, but not substantially beyond the relaxation oscillation frequency, there can be modulation of the population levels and thus the index of refraction will also be modulated. This modulation of the index through the population inversion will cause timing jitter since the round-trip cavity time which establishes the pulse repetition rate depends on the optical path length and hence the index of refraction everywhere along the path.

Consider next the thermal loading of the laser crystal. Not every photon of absorbed pump power goes into inverting an atom. And, certainly, there is loss of energy in so-called non-radiative transitions as the atom cascades down from the highest pumped energy level to the ground state. In any case, there is a substantial amount of heat deposited into the laser crystal and often this necessitates water cooling. The equilibrium temperature achieved under steady-state pumping conditions will be many degrees above ambient room temperature and, therefore, fluctuations in the pumping power will cause fluctuations in the crystal temperature. The index of refraction of the laser crystal depends directly on the temperature through the thermo-optic coefficient which for Ti:sapphire has the value $\partial n/\partial T = 1.2 \times 10^{-6}\ \mathrm{K}^{-1}$ [86]. Once again, a change

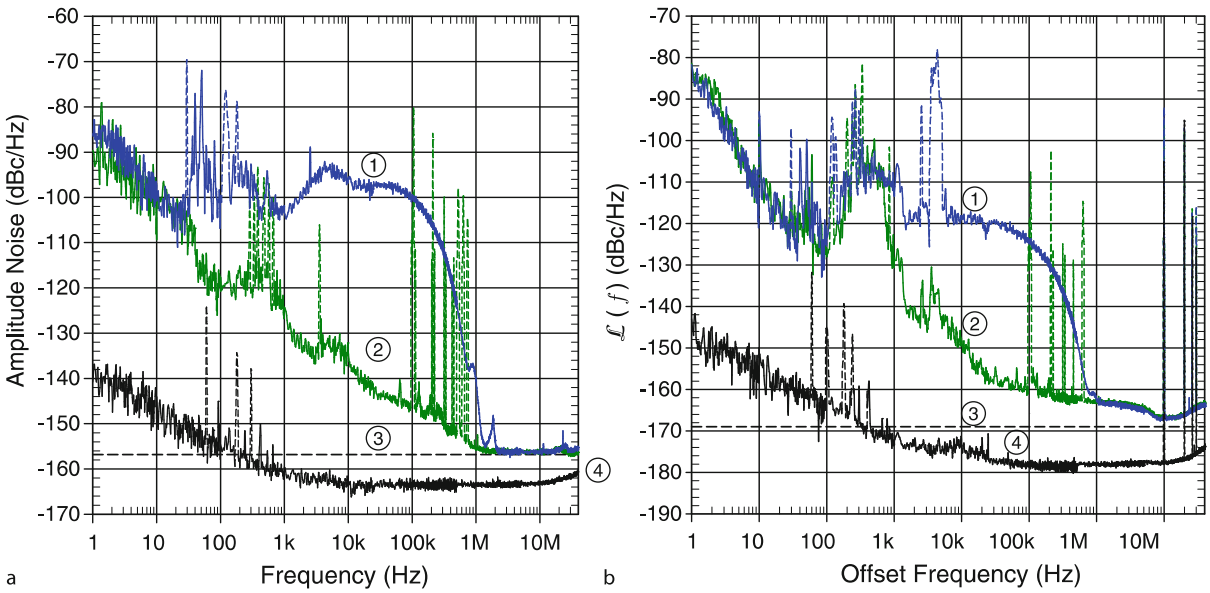

Noise and Stability in Modelocked Soliton Lasers, Figure 28

Evidence of noise transfer from pump laser. a Amplitude noise of Ti: sapphire laser pumped with *1* argon ion laser and *2* diode-pumped solid state laser. *3* Shot noise floor. *4* Photoreceiver noise floor. All spectra referenced to the average photocurrent. b Absolute single-sideband phase noise of the same Ti:sapphire laser with numbers corresponding to plots of AM spectra, left. All curves referenced to carrier amplitude. (From [79])

in the index of refraction will cause a change in the round-trip time delay for the pulses and manifest itself as timing jitter. Along with the direct thermo-optic effect, we have a thermally-driven expansion of the laser crystal at a differential rate of $5.8 \times 10^{-6}\,\mathrm{K}^{-1}$ which also increases the cavity round trip time.

Yet another contribution to timing jitter comes about in modelocked lasers running high peak-power pulses. The index of refraction has a nonlinear contribution due to the high peak power $I(t)$ such that $n = n_o + n_2 I(t)$ where $n_2 = 3.45 \times 10^{-16}\,\mathrm{cm}^2/\mathrm{W}$ [11]. In addition, when the index of refraction is modulated, the angle of refraction at the crystal surfaces due to Snell's law changes and the beam path through the dispersion compensation prism pair is modified slightly (see Fig. 26). The differential variation in time delay for a typical Kerr-lens modelocked Ti:sapphire laser due to an index change is $\mathrm{d}\tau/\mathrm{d}n \approx 4 \times 10^{-11}$ s.

Figure 28 shows directly the effect of pump noise inducing AM and PM noise on the pulse train of modelocked Ti:sapphire laser for two different types of pump laser. On the left is the amplitude noise of the Ti:sapphire laser measured at baseband and on the right is the single-sideband phase noise measured about the carrier at the fundamental repetition rate of 80 MHz. Trace #1 is the noise for the case of an argon-ion pump laser and trace #2 is for a diode-pumped solid-state (DPSS) single-frequency laser. Notice the dramatic improvement in noise reduction using the DPSS pump laser. Notice, also, the direct correlation between the noise spectra of the pump lasers, as shown in Fig. 25, and the noise spectra of the laser that is being pumped (Fig. 28). The relationship between the noise spectra of the pump lasers and the pumped laser can be characterized by the noise transfer functions $H_{\mathrm{AM}}(\omega)$ and $H_{\mathrm{PM}}(\omega)$.

Analytical Treatment of Amplitude-to-Amplitude Noise Transfer In contrast to the large-scale fluctuations characteristic of turn-on transients and routes to chaos, pump-power noise is expected to be very small in amplitude and thus a perturbation approach applied to the coupled cavity-atom rate equations is quite appropriate. We assume steady-state operation above threshold and therefore rewrite those equations without the spontaneous emission term

$$\frac{\mathrm{d}n(t)}{\mathrm{d}t} = KN(t)n(t) - \gamma_c n(t) \tag{190}$$

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = R_P(t) - \gamma_2 N(t) - KN(t)n(t)\,. \tag{191}$$

As we saw previously, the population inversion is clamped at threshold, $N_{\mathrm{th}} = \gamma_c/K$ (69), and the steady-

state photon density is given by (70)

$$n = n_{ss} = \left(\frac{R_{P_0}}{\gamma_2 N_{th}} - 1\right)\frac{\gamma_2}{K} = (r-1)\frac{\gamma_2}{K}\,. \tag{192}$$

The goal now is to find out how the population inversion and photon density respond to a small sinusoidal variation in the pumping rate. Before we apply this small-signal perturbation to the pump, we first find the natural response of the system (190), (191) to perturbations about the steady-state. We assume that $n(t)$ and $N(t)$ take the forms

$$\begin{aligned} n(t) &= n_{ss} + n_1(t)\,, \qquad & n_1(t) \ll n_{ss} \\ N(t) &= N_{th} + N_1(t)\,, \qquad & N_1(t) \ll N_{th} \end{aligned} \tag{193}$$

and substitute them into (190), (191) to obtain the following linearized small-signal rate equations

$$\frac{dn_1(t)}{dt} = (r-1)\gamma_2 N_1(t) \tag{194}$$

$$\frac{dN_1(t)}{dt} = -\gamma_c n_1(t) - r\gamma_2 N_1(t)\,. \tag{195}$$

In the usual way, we assume solutions that vary as e^{st} and obtain an equation for s with roots

$$s = s_1, s_2 = -\frac{r\gamma_2}{2} \pm \sqrt{\left(\frac{r\gamma_2}{2}\right)^2 - (r-1)\gamma_2\gamma_c}\,. \tag{196}$$

Now, in the case of solid-state lasers where $\gamma_2 \ll \gamma_c$, $(r-1)\gamma_2\gamma_c > (r\gamma_2/2)^2$, and the roots form a complex conjugate pair

$$s_1, s_2 = -\frac{r\gamma_2}{2} \pm i\sqrt{(r-1)\gamma_2\gamma_c - \left(\frac{r\gamma_2}{2}\right)^2}\,. \tag{197}$$

When we combine both complex solutions for $n_1(t)$ corresponding to s_1, s_2 with arbitrary coefficients, we obtain the real, damped, oscillatory behavior for the cavity photon number

$$n_1(t) = n_{o_1} e^{s_1 t} + n_{o_2} e^{s_2 t} = n_1 e^{-\gamma_{sp} t}\cos\omega'_{sp} t \tag{198}$$

where

$$\begin{aligned} \gamma_{sp} &\equiv \frac{r\gamma_2}{2}\,, \quad \omega'_{sp} \equiv \sqrt{\omega_{sp}^2 - \gamma_{sp}^2}\,, \\ &\omega_{sp} \equiv \sqrt{(r-1)\gamma_2\gamma_c}\,. \end{aligned} \tag{199}$$

The phase shift corresponding to initial conditions has been arbitrarily set equal to zero. The term ω_{sp} is often referred to as the "spiking" or "relaxation oscillation" frequency as is readily apparent in the plots of the turn-on transient solutions shown in Fig. 9. Thus, the solution to the first of (193) has the form

$$n(t) = n_{ss} + n_1 e^{-\gamma_{sp} t}\cos\omega'_{sp} t\,, \tag{200}$$

Next, to solve for $N_1(t)$ we substitute (198) for $n_1(t)$ back into (194) and (195) and find

$$N_1(t) = -2n_1\frac{\gamma_c}{\omega_{sp}} e^{-\gamma_{sp} t}\sin\left(\omega'_{sp} t + \phi\right) \tag{201}$$

where

$$\phi = \tan^{-1}\left[\frac{\gamma_{sp}}{\omega'_{sp}}\right]\,. \tag{202}$$

Finally, the complete solution for the population inversion is

$$N(t) = N_{th} - 2n_1\frac{\gamma_c}{\omega_{sp}} e^{-\gamma_{sp} t}\sin\left(\omega'_{sp} t + \phi\right)\,. \tag{203}$$

Note that since $\omega'_{sp} \gg \gamma_{sp}$, the population inversion *leads* the cavity photon number, as expected intuitively.

Now we can apply a small sinusoidal perturbation (modulation) to the pumping rate and solve for the effects on the photon and population inversion densities. Let

$$R_P(t) = R_{P_0} + \mathrm{Re}\left\{R_{P_1} e^{i\omega_m t}\right\} \quad s^{-1}m^{-3} \tag{204}$$

where $R_{P_1} \ll R_{P_0}$. Thinking ahead to a link with the noise transfer function, we can also write this in terms of a small-signal pump modulation index, m_p,

$$R_P(t) = R_{P_0}\left[1 + \mathrm{Re}\left\{m_p e^{i\omega_m t}\right\}\right] \tag{205}$$

where

$$m_p \equiv \frac{R_{P_1}}{R_{P_0}}\,. \tag{206}$$

When the pump is modulated sinusoidally, we expect that our linearized system will yield sinusoidally varying responses. Therefore, we seek solutions of the form

$$n(t) = n_{ss} + \mathrm{Re}\left\{\tilde{n}_1 e^{i\omega_m t}\right\}, \quad |\tilde{n}_1| \ll n_{ss} \tag{207}$$

$$N(t) = N_{th} + \mathrm{Re}\left\{\tilde{N}_1 e^{i\omega_m t}\right\}, \quad |\tilde{N}_1| \ll N_{th} \tag{208}$$

where $\tilde{n}_1$ and $\tilde{N}_1$ are complex phasor amplitudes. When these expressions are substituted back into (190) and (191) we obtain complex phasor amplitudes for the cavity photon number density

$$\tilde{n}_1 = \frac{\omega_{sp}^2 R_{P_1}/\gamma_c}{\omega_{sp}^2 - \omega_m^2 + i2\gamma_{sp}\omega_m} \tag{209}$$

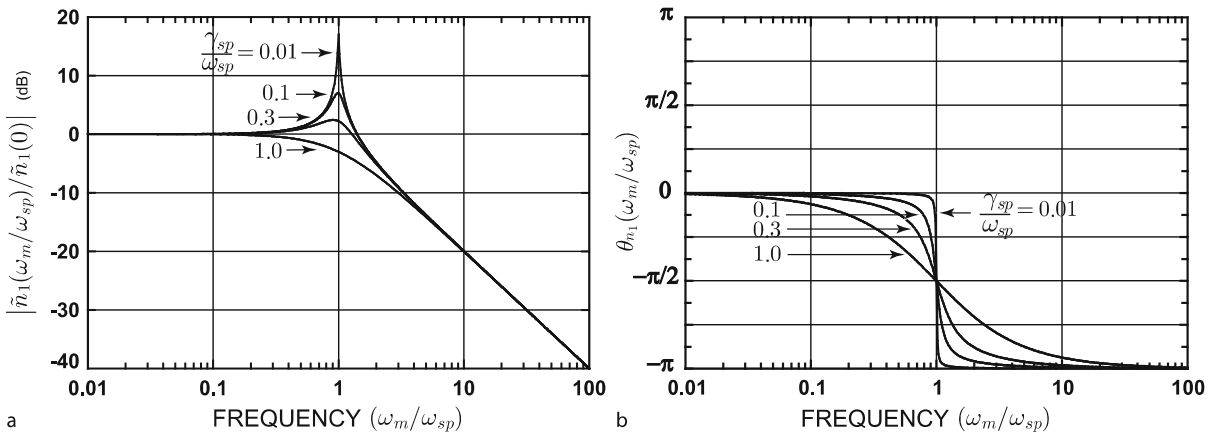

Noise and Stability in Modelocked Soliton Lasers, Figure 29
Normalized magnitude (a) and phase (b) of the cavity photon density response for several values of γ_{sp}/ω_{sp}

and the population inversion density, respectively

$$\tilde{N}_1 = \frac{i\omega_m R_{P_1}}{\omega_{sp}^2 - \omega_m^2 + i2\gamma_{sp}\omega_m} . \tag{210}$$

It is very useful to consider plots of these expressions as a function of the normalized frequency $\omega = \omega_m/\omega_{sp}$. It is also helpful to separate them into magnitude and phase. After a little algebra we find

$$|\tilde{n}_1| = \frac{R_{P_1}}{\gamma_c} \frac{1}{\sqrt{\left(1 - \left(\frac{\omega_m}{\omega_{sp}}\right)^2\right)^2 + \left(2\frac{\gamma_{sp}}{\omega_{sp}}\frac{\omega_m}{\omega_{sp}}\right)^2}} \tag{211}$$

$$\theta_{n_1} = \tan^{-1}\left[-\frac{2\frac{\gamma_{sp}}{\omega_{sp}}\frac{\omega_m}{\omega_{sp}}}{1 - \left(\frac{\omega_m}{\omega_{sp}}\right)^2}\right] = \tan^{-1}\left[-2\frac{\gamma_{sp}\omega_m}{\omega_{sp}^2 - \omega_m^2}\right] \tag{212}$$

$$|\tilde{N}_1| = \frac{R_{P_1}}{\omega_{sp}} \frac{\omega_m/\omega_{sp}}{\sqrt{\left(1 - \left(\frac{\omega_m}{\omega_{sp}}\right)^2\right)^2 + \left(2\frac{\gamma_{sp}}{\omega_{sp}}\frac{\omega_m}{\omega_{sp}}\right)^2}} \tag{213}$$

$$\theta_{N_1} = \tan^{-1}\left[\frac{1 - \left(\frac{\omega_m}{\omega_{sp}}\right)^2}{2\frac{\gamma_{sp}}{\omega_{sp}}\frac{\omega_m}{\omega_{sp}}}\right] = \tan^{-1}\left[\frac{\omega_{sp}^2 - \omega_m^2}{2\gamma_{sp}\omega_m}\right] . \tag{214}$$

These functions, in normalized from, are shown in Figs. 29 and 30 for various values of the normalized damping term γ_{sp}/ω_{sp}. They form the basis for understanding much of the behavior of the noise transfer functions that will be derived in following sections.

Now that we have expressions for the phasor amplitudes of the photon and population inversion densities, we can use them in the definitions of the noise transfer functions. We define the complex AM-to-AM noise transfer function at modulation frequency ω_m as the ratio of the induced laser amplitude modulation index to the pump modulation index

$$H_{AM}(\omega_m) \equiv \frac{\tilde{m}_L(\omega_m)}{m_P} = \frac{\tilde{n}_1(\omega_m)/n_{ss}}{R_{P_1}/R_{P_0}} \tag{215}$$

$$= \left(\frac{R_{P_0}}{n_{ss}R_{P_1}}\right)|\tilde{n}_1|e^{i\theta_{n_1}} \tag{216}$$

$$= |H_{AM}(\omega_m)|\, e^{i\theta_{AM}} \tag{217}$$

where

$$|H_{AM}(\omega_m)| \equiv \frac{R_{P_0}/n_{ss}\gamma_c}{\sqrt{\left(1 - \left(\frac{\omega_m}{\omega_{sp}}\right)^2\right)^2 + \left(2\frac{\gamma_{sp}}{\omega_{sp}}\frac{\omega_m}{\omega_{sp}}\right)^2}} \tag{218}$$

and

$$\theta_{AM} = \theta_{n_1} = \tan^{-1}\left[-2\frac{\gamma_{sp}\omega_m}{\omega_{sp}^2 - \omega_m^2}\right] . \tag{219}$$

Apart from the DC response, we see that $H_{AM}(\omega_m)$ has exactly the same form as the photon density plotted in Fig. 29.

The low frequency asymptotic form, or $H_{AM}(0)$, can be cast in a particularly revealing form by using (69) and (70);

$$H_{AM}(0) = \frac{R_{P_0}}{n_{ss}\gamma_c} = \frac{R_{P_0}K}{\gamma_c\gamma_2(r-1)} = \frac{R_{P_0}/N_{th}\gamma_2}{r-1} = \frac{r}{r-1} . \tag{220}$$

This expression tells us that, since the pumping rate is, by definition, $r > 1$ above threshold, the induced AM noise

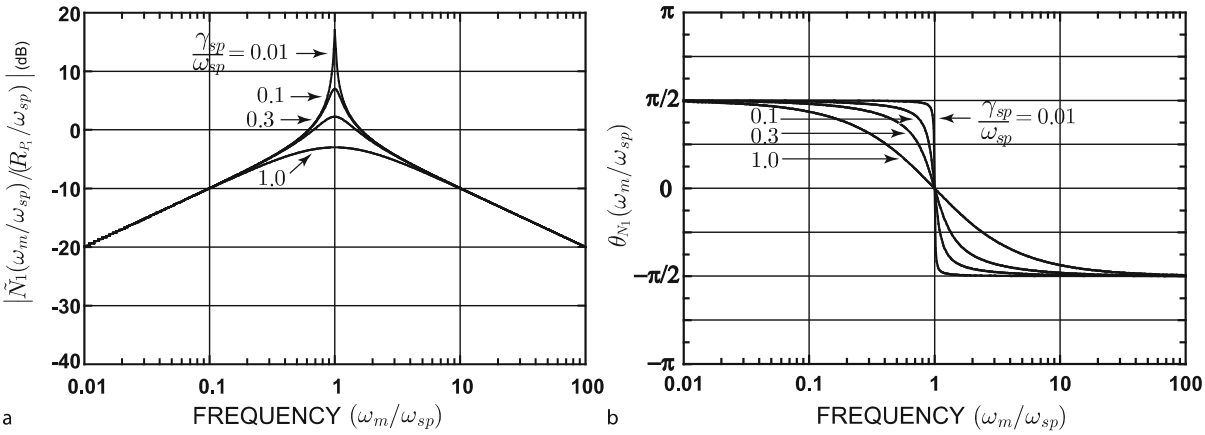

Noise and Stability in Modelocked Soliton Lasers, Figure 30
Normalized magnitude (a) and phase (b) of the population inversion density response for several values of γ_{sp}/ω_{sp}

in the pumped laser at low frequencies *always exceeds the noise of the pump!*

The other two important attributes of the spectral dependence of $H_{AM}(\omega_m)$ are the peaking at the relaxation oscillation (or spiking) frequency, ω_{sp}, and the slope for $\omega_m > \omega_{sp}$. In the classic fashion of the damped harmonic oscillator, the cavity photon number, and hence $H_{AM}(\omega_m)$ have a peak in the response at the resonance frequency with the magnitude of the peaking dependent upon the degree of damping, γ_{sp}/ω_{sp}. Above the resonance, the system cannot follow the rapid oscillations in the pumping rate and $H_{AM}(\omega_m)$ decreases at a rate of 20 dB/decade which is characteristic of the "second-order pole" in (211).

Analytical Treatment of Amplitude-to-Phase Noise Transfer As alluded to previously, the sources and mechanisms of amplitude-to-phase noise transfer are diverse and complicated. A full treatment is beyond the scope of the present discussion but we can get a glimpse of how to include all of the physical effects by beginning with a heuristic model and then establishing links between the model elements and the physical mechanisms.

Suppose we follow a packet of photons around the laser cavity and designate the round trip time as τ_c. Then, if τ_c is constant, it is reasonable to describe the train of optical pulses exiting the laser by

$$P(t) = P_o \sum_n A(t - n\tau_c) \tag{221}$$

where P_o is the peak optical power and A is the shape of an individual pulse. Since the noise processes that disturb the round trip cavity, and thus lead to timing fluctuations, are random, we will study the effects in the frequency domain using the phase noise measurement apparatus described earlier. We can analyze one particular spectral component of the power spectral density of the phase fluctuations by including in (221) a sinusoidal variation at frequency ω_m so that

$$P(t) = P_o \sum_n A\left(t - \tau_{co}\left[n + \frac{\beta(\omega_m)}{2\pi}\cos\big(\omega_m t + \phi(\omega_m)\big)\right]\right) \tag{222}$$

where τ_{co} is the unperturbed round-trip cavity time, $\beta(\omega_m)$ is the peak phase deviation, and $\phi(\omega_m)$ is a frequency-dependent phase offset between the output pulse train and whatever mechanism is causing the fluctuation. When expressed in this fashion, a peak deviation of $\beta = 2\pi$ radians corresponds to one period of the pulse train, or one round-trip cavity time. Comparing (221) and (222) we see that the instantaneous round trip cavity time with the sinusoidal fluctuation can be written

$$\tau_c(t) = \tau_{co} + \tau_{co}\frac{\beta(\omega_m)}{2\pi}\cos(\omega_m t + \phi(\omega_m)) \,. \tag{223}$$

This is a phenomenological description since it just represents the *effect* of the perturbation and not the cause. The cause of the perturbations can be included in a *physical* description in terms of propagation lengths and pulse group velocities;

$$\tau_c = 2\left[\frac{L_{ca}}{v_{air}} + \frac{L_{cp}}{v_{pr}} + \frac{L_a}{v_{ga}}\right] . \tag{224}$$

The first term accounts for the propagation time through air between the optical components in the cavity. L_{ca} is the length of the path and v_{air} is the velocity. The second term

accounts for the group delay through the prisms, if intracavity prisms are used for dispersion compensation. Finally, the third terms accounts for the group delay through the active atomic medium. L_a is the length of the gain medium and v_{ga} is the group velocity through the medium. Variations in the pump power will cause variations in all quantities except the velocity in air, v_{air}. The variations in the remaining five are caused by the pump modulating the photon density, the population inversion and the laser rod temperature. We have analyzed the link between pump modulation and the first two earlier. We can build on this by once again assuming a sinusoidal pump fluctuation

$$R_P(t) = R_{P_0} + R_{P_1} \cos \omega_m t \,. \tag{225}$$

Once again we seek solutions of the form

$$n(t) = n_{ss} + \mathrm{Re}\left\{\tilde{n}_1 e^{i\omega_m t}\right\} \tag{226}$$

$$N(t) = N_{th} + \mathrm{Re}\left\{\tilde{N}_1 e^{i\omega_m t}\right\} \tag{227}$$

$$T(t) = T_o + \mathrm{Re}\left\{\tilde{T}_1 e^{i\omega_m t}\right\} \,. \tag{228}$$

The phasor amplitudes for $\tilde{n}_1$ and $\tilde{N}_1$ have been found (209), (210) and the laser rod temperature can be shown to respond as [87]

$$\tilde{T}_1(r') = \frac{2}{k}\frac{R_{P_1}}{R_{P_0}} \sum_{l=1}^{\infty} \frac{J_0(\beta_l r'/b)}{\beta_l^2 J_1^2(\beta_l)} \frac{1}{1 + i\omega_m \frac{b^2}{k\beta_l^2}} \tag{229}$$

$$= \frac{2}{k}\frac{R_{P_1}}{R_{P_0}} \sum_{l=1}^{\infty} \frac{J_0(\beta_l r'/b)}{\beta_l^2 J_1^2(\beta_l)} \frac{1}{\sqrt{1 + \left(\omega_m \frac{b^2}{k\beta_l^2}\right)^2}} e^{i\theta_T} , \tag{230}$$

$$\theta_T \equiv \tan^{-1}\left[-\omega_m \frac{b^2}{k\beta_l^2}\right] \tag{231}$$

where k is the diffusivity (m^2/s), β_l are the eigenvalues of the radial heat diffusion problem, r' is the radial coordinate in the laser rod and b is the radius of the rod.

By carefully considering the contributions of each of the fluctuating quantities $\tilde{n}_1$, $\tilde{N}_1$ and $\tilde{T}_1$ to the total round-trip cavity time, we can gather the physical coupling factors into coefficients on the phasor amplitudes by writing

$$\tau_c(t) = A + B\,\mathrm{Re}\left\{\tilde{n}_1 e^{i\omega_m t}\right\} + C\,\mathrm{Re}\left\{\tilde{N}_1 e^{i\omega_m t}\right\} + D\,\mathrm{Re}\left\{\tilde{T}_1 e^{i\omega_m t}\right\} \,. \tag{232}$$

The derivation of the coupling coefficients is rather involved and will be presented elsewhere [88] but we note the results here

$$A \equiv \tau_{co}(n_{go}) + \frac{2L_{a0}}{c}\left[n_{host}(\omega_a) + n_2 I_{pk} + \omega_a \frac{dn_{host}}{d\omega} + \frac{e^2}{n_{host} m \epsilon_0 \Delta\omega_a^2} N_{th}\right] \tag{233}$$

$$B \equiv \frac{n_2 I_{pk}}{n_{ss}}\left[\frac{d\tau_p}{dn} + 2\frac{L_{a0}}{c}\right] \tag{234}$$

$$C \equiv \frac{2L_{a0}}{c}\frac{e^2}{n_{host} m \epsilon_0 \Delta\omega_a^2} \tag{235}$$

$$D \equiv \frac{d\tau_p}{dn}\frac{\partial n}{\partial T} + \alpha_T \frac{2L_{a0}}{c}\left[n_{host}(\omega_a) + n_2 I_{pk} + \omega_a \frac{dn_{host}}{d\omega} + \frac{e^2}{n_{host} m \epsilon_0 \Delta\omega_a^2} N_{th} + \frac{1}{\alpha_T}\frac{\partial n}{\partial T}\right] . \tag{236}$$

The terms in these expressions are defined in the beginning of this chapter (Nomenclature) with the following exceptions

$$\frac{d\tau_p}{dn} = \text{slope of the group delay through prisms} \,,$$
$$\frac{\partial n}{\partial T} = \text{thermo-optic coefficient} \,.$$

The coupling mechanisms between the fluctuating quantities $\tilde{n}_1$, $\tilde{N}_1$ and $\tilde{T}_1$ and the physical characteristics of the laser were discussed in the beginning of this section and are summarized in Table 6 below.

We now define the phase noise transfer function as the ratio of the peak induced complex phase deviation to the pump modulation index

$$H_{PM}(\omega_m) \equiv \frac{\tilde{\beta}(\omega_m)}{m_P} \tag{237}$$

Noise and Stability in Modelocked Soliton Lasers, Table 6
Coupling between the modulated laser parameters (photon density, population inversion density and laser rod temperature) and the physical characteristics that determine the round-trip cavity time τ_c

	PHOTON DENSITY $\tilde{n}_1$	POPULATION INVERSION DENSITY $\tilde{N}_1$	LASER ROD TEMPERATURE $\tilde{T}_1$
GAIN MEDIUM (LASER ROD) LENGTH			*D*
GAIN MEDIUM GROUP VELOCITY	*B*	*C*	*D*
CAVITY LENGTH (BEAM STEERING)	*B*		*D*

where the complex phase deviation $\tilde{\beta}(\omega_m) \equiv \beta(\omega_m) e^{i\phi(\omega_m)}$. Comparing (223) with (232) and using $m_p = R_{P_1}/R_{P_0}$, we have

$$H_{PM}(\omega_m) = \frac{2\pi}{\tau_{co}} \frac{R_{P_0}}{R_{P_1}} \left(B\tilde{n}_1(\omega_m) + C\tilde{N}_1(\omega_m) + D\tilde{T}_1(\omega_m) \right) \tag{238}$$

where we have explicitly emphasized the frequency dependence of the phasor amplitudes.

Measurement of the Complex Noise Transfer Function

A system for measuring the complex noise transfer function (NTF) is conceptually straightforward and quite similar to that used for laser noise measurements. Figure 23a shows a schematic diagram of such a system. The principal differences between the noise measurement system (Fig. 23b) and the NTF characterization system is the introduction of an acousto-optic amplitude modulator (AOM) into the pump beam path to simulate noise fluctuations and a vector signal analyzer to provide a stimulus and measure the induced response. The vector signal analyzer provides a swept sinusoidal signal from 0.1 Hz to 10 MHz and records the depth of pump modulation on Channel 1 and the depth of Ti:apphire laser modulation on Channel 2. By switching between two specialized receivers, either the AM or PM response can be measured both in magnitude and phase compared with the pump signal. It is the complex ratio of the induced AM or PM from the Ti:sapphire laser to that of the pump that defines the AM- or PM-NTF, respectively.

Figures 31 and 32 show the data for measurements of $H_{AM}(\omega_m)$ and $H_{PM}(\omega_m)$, along with theoretical plots based on (215) and (238) [85]. Since the AM noise transfer function, $H_{AM}(\omega_m)$, depends only on the photon density, it has exactly the same shape as the plot of $\tilde{n}_1(\omega_m)$ shown in Fig. 29. There is a small amount of peaking at the relaxation oscillation frequency ($\approx$ 500 KHz) and sharp rolloff (−20 dB/decade) thereafter. In the plots of $H_{PM}(\omega_m)$ (Fig. 32) we see the composite behavior predicted by expression (238), that is, a superposition of the frequency-domain behavior of modulations of n, N and T. At low frequencies, where thermal effects dominate, there is a −10 dB/decade rolloff consistent with the first-order pole of (229). In the middle and upper frequency ranges (1 KHz $< f <$ 10 MHz) the behavior mimics that of the photon fluctuation $\tilde{n}_1(\omega_m)$. The population inversion density, $\tilde{N}_1(\omega_m)$, depicted in Fig. 30, with its characteristic ±10 dB/decade slopes on either side of the relaxation oscillation frequency, does not appear because the magnitude of the effect of the coupling to the group velocity in the gain medium is swamped by the effects of the photon density.

Predicting Amplitude and Phase Noise from the Pump Noise Spectrum and the Noise Transfer Function

Since the noise transfer function represents the response of a linear system to the noise spectrum of a pump laser, we should be able to use it to predict the noise properties of the pumped laser once we have characterized the noise spectrum of the pump [89]. Thus, if $S_P(\omega_m)$ is the noise power spectral density (PSD) of the pump laser (as measured in the photodetected current spectrum), the amplitude noise power spectral density is given by

$$S_{AM}(\omega_m) = |H_{AM}(\omega_m)|^2 S_P(\omega_m) \tag{239}$$

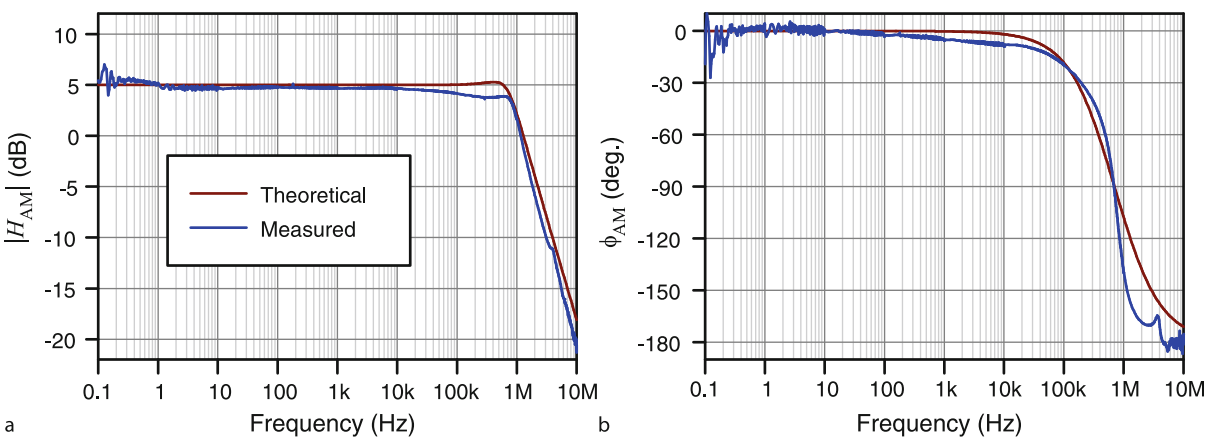

Noise and Stability in Modelocked Soliton Lasers, Figure 31

a Magnitude and b phase of the measured and theoretical complex AM NTF response. Continuous frequency measurements made with vector signal analyzer in setup shown in Fig. 23b

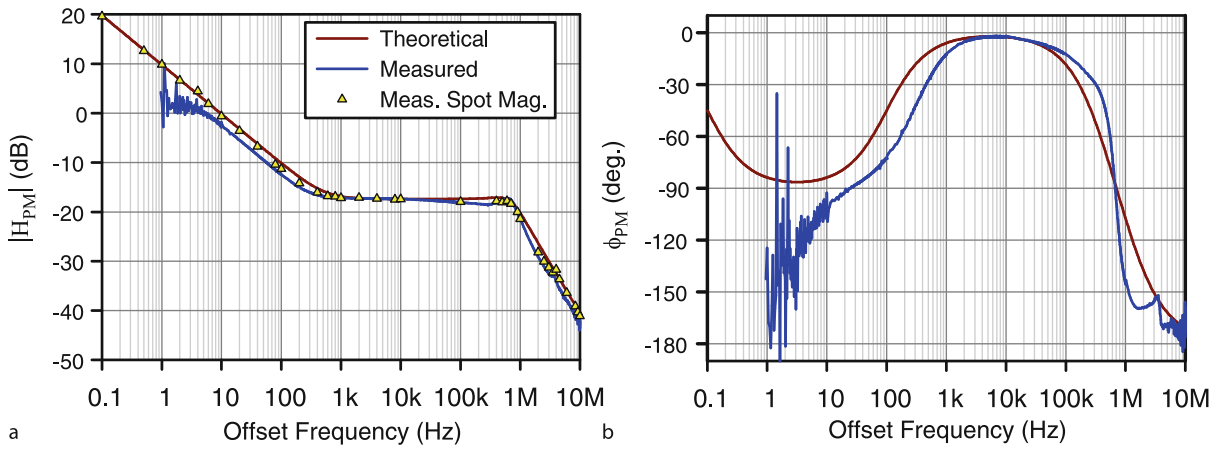

Noise and Stability in Modelocked Soliton Lasers, Figure 32

a Magnitude and b phase of the measured and theoretical complex PM NTF response. Continuous frequency measurements made with vector signal analyzer in setup shown in Fig. 23b. *Yellow triangles* are data taken at single frequency points (spot measurements) for verification of swept frequency approach and for better signal-to-noise below 5 Hz

and the single-sideband phase noise PSD is similarly given by

$$S_{\mathrm{PM-SSB}}(\omega_m) = \frac{|H_{\mathrm{PM}}(\omega_m)|^2}{2} S_{\mathrm{P}}(\omega_m) \tag{240}$$

where the factor of two comes about from the phase noise being specified as the single-sideband portion about the carrier.

Figure 33a shows the noise spectra of two types of pump lasers that can be used for pumping modelocked Ti:sapphire lasers. The multi-wavelength argon ion laser (Coherent Innova 300C) clearly has much more noise than the single longitudinal-mode diode-pumped solid-state laser (DPSS). The DPSS laser in this case was a frequency-doubled Nd:YVO_4 laser (Coherent Verdi). Figure 33b shows the absolute value of noise transfer function of a 100 MHz Ti:sapphire laser measured using both of these pump sources. The very slight difference in the data is due to the different effective pumping powers and how that affects the relaxation oscillation. When the noise spectra of the pump lasers is multiplied by the modulus-squared NTF (239), (240), one obtains the predicted Ti:sapphire laser noise as is shown in Figs. 34 and 35. The correlation between the measured and the predicted spectra is very good and validates the concept of the noise transfer function method.

Finally, it is interesting to do a comparison between the Ti:sapphire noise spectra when pumped by both lasers and this is shown in Fig. 36. Notice the considerable improvement in both amplitude and phase noise when using the DPSS pump laser.

Noise in Soliton Lasers

Much of the early work on instability in soliton lasers was developed by Haus and coworkers at M.I.T. In the following section we follow closely the pioneering approach presented by Haus and Mecozzi in their now-classic paper [65].

To study the issue of stability in soliton lasers, we return to the master equation (107) with a slight modification to allow for the fact that the pulse amplitude is not necessarily stable upon a cavity round trip [65]. Thus,

$$T_{\mathrm{R}} \frac{\partial a(T,t)}{\partial T} = \left[-l + g\left(1 - \frac{1}{\Omega_{\mathrm{g}}}\frac{\partial}{\partial t} + \frac{1}{\Omega_{\mathrm{g}}^2}\frac{\partial^2}{\partial t^2}\right) + \mathrm{i}D\frac{\partial^2}{\partial t^2} + (\gamma - \mathrm{i}\delta)|a|^2 \right] a(T,t) + T_{\mathrm{R}} S(t,T) \tag{241}$$

where $a(T,t)$ is the field amplitude which is now both a function of the local short-time variable t, and the long-time variable T, on the order of many cavity round-trip times, T_{R}. $a(T,t)$ is normalized such that $|a|^2$ is equal to the instantaneous envelope power. The other parameters are the same as described in Sect. "Modelocked Soliton Lasers" except for the new term, $S(t,T)$, which will account for noise and about which more will be said shortly. Once again we assume that the gain has a very long recovery time and thus that it saturates to a steady-state level due to the passage of many ultrashort pulses according to

$$g = \frac{g_{\mathrm{o}}}{1 + \frac{1}{P_{\mathrm{s}} T_{\mathrm{r}}} \int |a|^2 \, \mathrm{d}t} \tag{242}$$

where P_{s} is the saturation power.

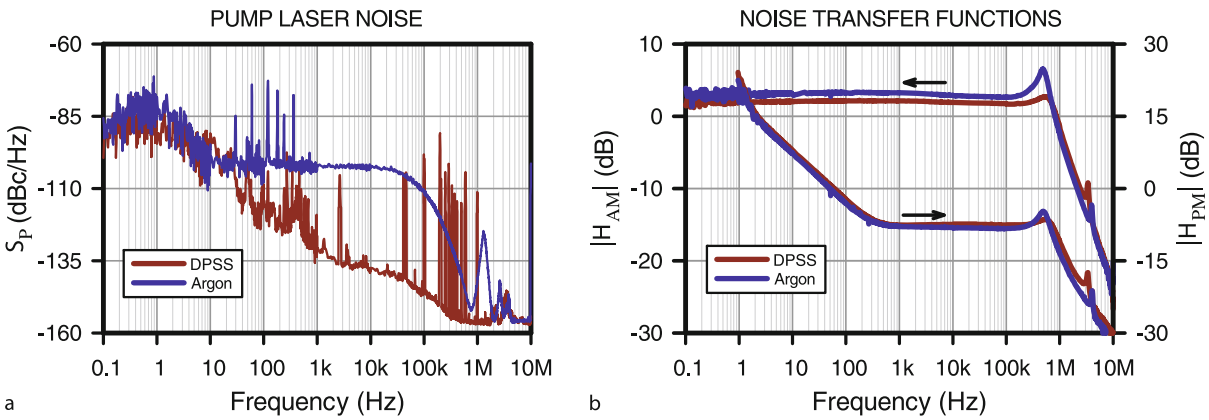

Noise and Stability in Modelocked Soliton Lasers, Figure 33

a Pump noise spectra of a single longitudinal-mode diode-pumped solid-state (DPSS) laser and an argon-ion laser. **b** Measured noise transfer functions of KLM Ti:sapphire laser using both pump lasers

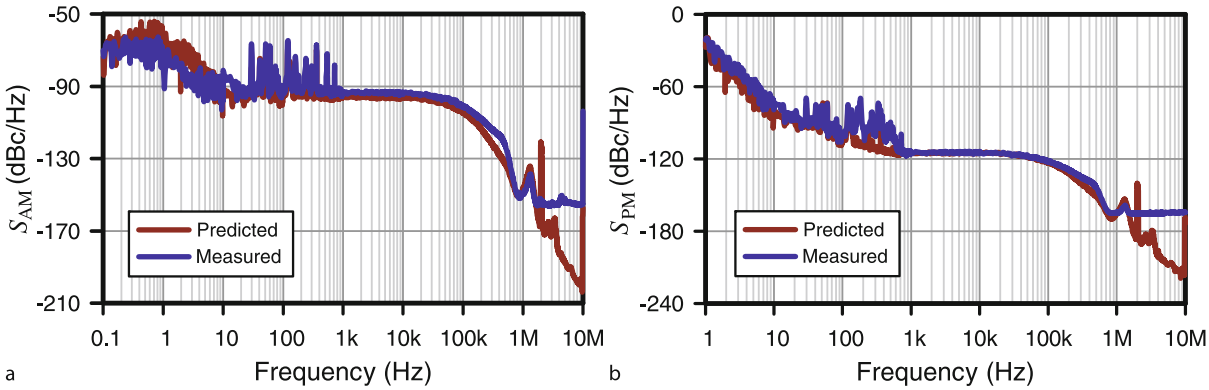

Noise and Stability in Modelocked Soliton Lasers, Figure 34

a Amplitude noise power spectral density of a modelocked Ti:sapphire laser when pumped by an argon-ion laser running all lines. Compare with the argon-ion laser noise spectrum in Fig. 33a. **b** Single sideband phase noise power spectral density of the same laser

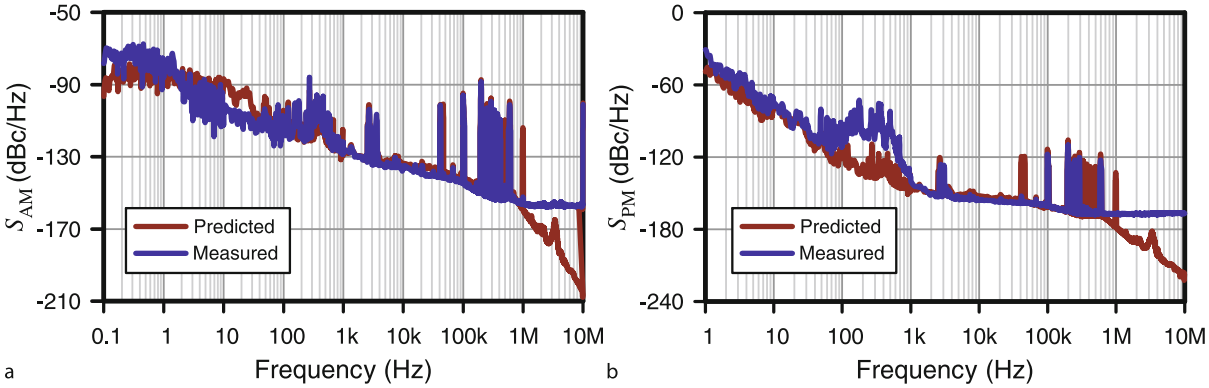

Noise and Stability in Modelocked Soliton Lasers, Figure 35

a Amplitude noise power spectral density of a modelocked Ti:sapphire laser when pumped by a DPSS laser. Compare with the argon-ion laser noise spectrum in Fig. 33a. **b** Single sideband phase noise power spectral density of the same laser

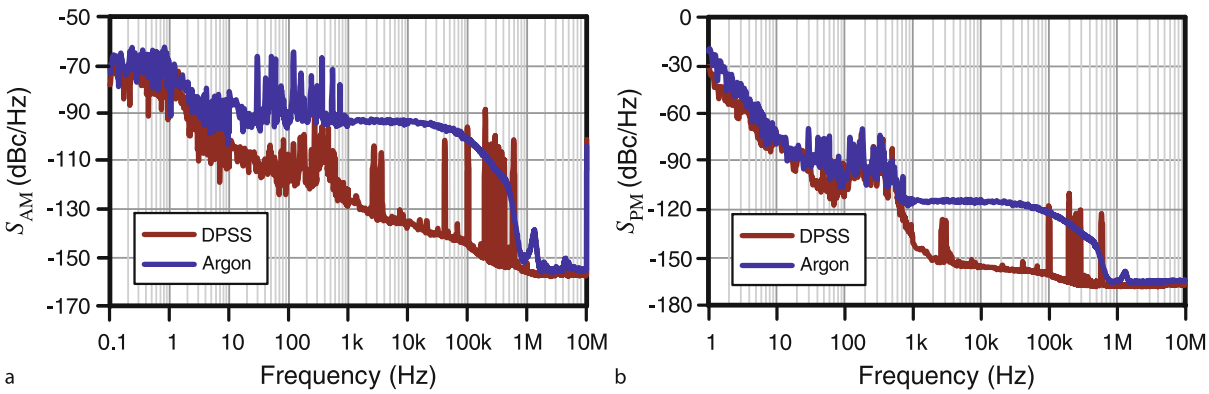

Noise and Stability in Modelocked Soliton Lasers, Figure 36

a Amplitude noise power spectral density of a modelocked Ti:sapphire laser when pumped by an argon-ion laser running all lines (*blue*) compared with the same laser pumped by a DPSS pump laser (*red*). b Single sideband phase noise under same conditions as a

Once again we introduce a solution ansatz. Let

$$a(t,T) = A_o \operatorname{sech}\left[\frac{1}{\tau}(t-t_o)\right]^{(1+i\beta)} \exp\left(i\psi\frac{T}{T_R} + i\theta\right) \tag{243}$$

which generates the pair of equations

$$-i\psi + g - l + \frac{(1+i\beta)^2}{\tau^2}\left(\frac{g}{\Omega_g^2} + iD\right) = 0 \tag{244}$$

$$\frac{2+3i\beta-\beta^2}{\tau^2}\left(\frac{g}{\Omega_g^2} + iD\right) = (\gamma - i\delta)A_o^2 \tag{245}$$

and where

$$t_o = \frac{gT}{\Omega_g T_R}\,. \tag{246}$$

From these we can find stable solutions for the physical parameters τ (pulsewidth), β (chirp parameter), ψ (phase shift), and g, the gain. Of the possible stable solutions, pulses that contain no chirp ($\beta = 0$) are to be preferred since they minimize the time-bandwidth product and thus minimize the pulsewidth for a given gain-bandwidth. With this in mind we set $\beta = 0$ and find, from (244) that the following relation obtains;

$$\frac{g/\Omega_g^2}{-D} = \frac{\gamma}{\delta} \equiv \mu\,. \tag{247}$$

The term g/Ω_g^2 represents the "gain dispersion", and its ratio to the dispersion parameter ($-D$) is given by the ratio of the saturable absorption to SPM coefficients. Applying $\beta = 0$ and (247) above to (244) and (245) results in the following,

$$A_o^2\tau^2 = \frac{2|D|}{\delta} \tag{248}$$

$$\psi = -\frac{|D|}{\tau^2} = -\frac{\delta}{2}A_o^2 \tag{249}$$

$$g - l = -\frac{1}{\tau^2}\frac{g}{\Omega_g^2} = -\frac{\gamma}{2}A_o^2\,. \tag{250}$$

If these relations are now substituted into the general form of the solution, (243), we have the following steady-state unchirped solution

$$a_o(t,T) = A_o \operatorname{sech}\left(\frac{t-t_o}{\tau}\right)\exp\left[-i\left(\frac{\delta}{2}A_o^2\frac{T}{T_R} + \theta\right)\right] \tag{251}$$

where the pulsewidth

$$\tau = \frac{1}{A_o}\sqrt{\frac{2|D|}{\delta}}\,. \tag{252}$$

Now, to study the effects of noise on the stability of the pulse solution, we can employ perturbation theory [90]. In this technique, when applied to a general differential equation such as $\mathcal{L}\{y(x)\}$ where $\mathcal{L}$ is an arbitrary differential operator, we assume that a small adjustable perturbation, ξ, can be inserted strategically into the problem such that when $\xi = 0$, the problem is exactly solvable. This gives the zero-order solution $y_0(x)$. We can then construct a perturbation solution in terms of powers of ξ,

$$y(x) = y_0(x) + \xi y_1(x) + \xi^2 y_2(x) + \ldots \tag{253}$$

Each of the $y_n(x)$ are computed from the prior terms $y_0(x), y_1(x), \ldots, y_{n-1}(x)$.

Soliton perturbation theory takes a slightly different approach [91]. To allow for the possibility of fluctuations

in the pulse parameters owing to noise, we make the new ansatz

$$a(t,T) = A_o\left[a_s(t-t_o) + \Delta a(t-t_o,T)\right]\exp\left(-\mathrm{i}\frac{\delta}{a}A_o^2\frac{T}{T_R}\right) \tag{254}$$

where

$$a_s(t) = A_o\,\mathrm{sech}\left(\frac{t}{\tau}\right). \tag{255}$$

$a_s(t)$ is the perturbation. Substituting the ansatz (254) into the master equation (241) and retaining terms to first order in Δa leads to a new equation for Δa;

$$\begin{aligned} T_R\frac{\partial}{\partial T}\Delta a(t-t_o,T) &= (\mu-\mathrm{i})\left\{-\frac{\delta}{2}A_o^2 + |D|\frac{\partial^2}{\partial t^2} + 2\delta a_s^2(t-t_o)\right\}\delta a \\ &\quad + (\mu-\mathrm{i})\delta a_s^2(t-t_o)\Delta a^* \\ &\quad - g_s\left(1-\frac{1}{\Omega_g}\frac{\partial}{\partial t}\right)\frac{a_s(t)}{2\tau A_o^2}\int \mathrm{d}t\, a_s(t)(\Delta a + \Delta a^*) \\ &\quad + T_R S(t,T) \end{aligned} \tag{256}$$

where

$$g_s \equiv \frac{g_o}{P_s T_R}\frac{2\tau A_o^2}{\left(1+\frac{2\tau A_o^2}{P_s T_R}\right)^2}. \tag{257}$$

The goal now is to expand the perturbation in a series of terms which independently account for fluctuations in the relevant physical parameters. Let these be called the "perturbation amplitudes" $\Delta w(T)$ (pulse energy), $\Delta\theta(T)$ (phase), $\Delta t(T)$ (timing) and $\Delta p(T)$ (frequency) . We now expand the perturbation in terms of these amplitudes and derivatives of the steady-state solution

$$\begin{aligned} \Delta a &= f_w(t-t_o)\Delta w(T) + f_\theta(t-t_o)\Delta\theta(T) \\ &\quad + f_t(t-t_o)\Delta t(T) + f_p(t-t_o)\Delta p(T) + \Delta a_c(t-t_o,T) \end{aligned} \tag{258}$$

where $\Delta a_c(t-t_o,T)$ accounts for continuum radiation which is not part of the pulse. The sensitivities to the perturbation amplitudes are given by derivatives of the steady-state solution a_o,

$$\text{ENERGY}\quad f_w(t) = \frac{\partial a_o}{\partial w_o} = \frac{1}{w_o}\left[1-\frac{t}{\tau}\tanh\left(\frac{t}{\tau}\right)\right]a_s(t) \tag{259}$$

$$\text{PHASE}\quad f_\theta(t) = \frac{\partial a_o}{\partial \theta_o} = \mathrm{i}a_s(t) \tag{260}$$

$$\text{TIMING}\quad f_t(t) = \frac{\partial a_o}{\partial t_o} = \frac{1}{\tau}\tanh\left(\frac{t}{\tau}\right)a_s(t) \tag{261}$$

$$\text{FREQUENCY}\quad f_p(t) \equiv \mathrm{i}\frac{2}{w_o}ta_s(t). \tag{262}$$

The next step involves defining a set of orthogonal adjoint functions to those (259)–(262) above and then projecting upon them the coefficients (amplitudes) of the perturbation expansion [91].

$$\underline{f}_w(t) = 2a_s(t) \tag{263}$$

$$\underline{f}_\theta(t) = \mathrm{i}\frac{2}{w_o}\left[1-\frac{t}{\tau}\tanh\left(\frac{t}{\tau}\right)\right]a_s(t) \tag{264}$$

$$\underline{f}_t(t) = \frac{2}{w_o}ta_s(t) \tag{265}$$

$$\underline{f}_p(t) = \mathrm{i}\left[\frac{2}{w_o\tau}\tanh\left(\frac{t}{\tau}\right)\right]a_s(t). \tag{266}$$

These functions are chosen to be orthonormal so that

$$\mathrm{Re}\left\{\int \underline{f}_i^*(t)f_j(t)\mathrm{d}t\right\} = \delta_{ij}. \tag{267}$$

The projections of the perturbation amplitudes are then

$$\Delta w(T) = \frac{1}{2}\int\left[\underline{f}_w^*(t)\Delta a(t) + \underline{f}_w(t)\Delta a^*(t)\right]\mathrm{d}t \tag{268}$$

$$\Delta\theta(T) = \frac{1}{2}\int\left[\underline{f}_\theta^*(t)\Delta a(t) + \underline{f}_\theta(t)\Delta a^*(t)\right]\mathrm{d}t \tag{269}$$

$$\Delta p(T) = \frac{1}{2}\int\left[\underline{f}_p^*(t)\Delta a(t) + \underline{f}_p(t)\Delta a^*(t)\right]\mathrm{d}t \tag{270}$$

$$\Delta t(T) = \frac{1}{2}\int\left[\underline{f}_t^*(t)\Delta a(t) + \underline{f}_t(t)\Delta a^*(t)\right]\mathrm{d}t. \tag{271}$$

When the perturbation expansion (258) using the projections (268)–(271) is substituted back into the differential equation for the original perturbation ansatz (254), we obtain the equations of motion for the perturbation amplitudes

$$T_R\frac{\partial}{\partial T}\Delta w = \left(-2g_s + 2\gamma A_o^2\right)\Delta w + T_r S_w(T) \tag{272}$$

$$T_R\frac{\partial}{\partial T}\Delta\theta = -\delta A_o^2\frac{\Delta w}{w_o} + T_r S_\theta(T) \tag{273}$$

$$T_R\frac{\partial}{\partial T}\Delta p = -\frac{4}{3}\frac{g}{\Omega_g^2\tau^2}\Delta p + T_r S_p(T) \tag{274}$$

$$T_R\frac{\partial}{\partial T}\Delta t = -2|D|\Delta p - \frac{g}{\Omega_g}\frac{\Delta w}{w_o} + T_r S_t(T) \tag{275}$$

where the $S_j(T)$ are the noise sources defined in terms of their projections

$$S_j(T) = \frac{1}{2}\int \left[\underline{f}^*_t(t)S(t,T) + \underline{f}_t(t)S^*(t,T)\right] \mathrm{d}t\,. \quad (276)$$

The physical origins of the $S(t,T)$ will be described shortly.

The equations of motion (272)–(275) are all linear first-order inhomogeneous equations driven by noise terms. Equations (272) and (274) are simple relaxation equations while the other two have no natural relaxation times and are driven by other perturbation terms as well as the noise. What can we learn about the general nature of these equations? The general form of a driven relaxation equation is

$$\frac{\mathrm{d}y(t)}{\mathrm{d}t} + \alpha y(t) = F(t) \quad (277)$$

where $F(t)$ is a forcing function. In the context of an initial-value problem, the solution is found by using an integrating factor [92],

$$y(t) = y(0)\mathrm{e}^{-\alpha t} + \int_0^t \mathrm{e}^{-\alpha(t-t')}F(t')\mathrm{d}t'\,. \quad (278)$$

After complete relaxation of the initial condition ($t \gg 1/\alpha$), the response is given by the convolution of the forcing function with the decay term. Comparing (277) with (272)–(275), we see that the noise terms represent the forcing function. Equation (272) says that if the saturable absorption action is strong enough (large $\gamma_o A_o^2$), the pulse energy will grow exponentially unless the gain saturation (g_s) is strong enough to suppress it. Equation (273) indicates that pulse energy fluctuations induce phase shift due to the Kerr effect (δ). Equation (274) says that if the center frequency of the pulse is tuned off line center, it will be pushed back ($\Delta p \to 0$) because the gain is highest at line center. In the last equation, (275), we see that both frequency change and pulse energy change cause timing shift. The effect of the first term because of group-velocity dispersion ($|D|$) and the effect of the second term because of gain changes producing index of refraction change due to the Kramers–Kronig relations.

Physical Sources of Noise

There are several principal sources of noise that will enter into the equations of motion (272)–(275) via the projections (276). The first is due to gain fluctuation. This will come about most simply bu virtue of a noisy pump mechanism. From the master equation (241), we see that gain fluctuations that occur on a relatively long time scale can be described by the equivalent noise source

$$S_{\Delta g}(t,T) = \frac{\Delta g(T)}{T_\mathrm{R}}\left(1 - \frac{1}{\Omega_\mathrm{g}}\frac{\partial}{\partial t}\right)a_\mathrm{s}(t,T)\,. \quad (279)$$

There is another, slightly subtle, effect that occurs due to gain fluctuation. If the gain and loss curves due not share the exact same line center, then as the gain goes up and down, the *net gain* (that is, the gain-loss) will necessarily shift, and this will cause a frequency-pulling term

$$S_{\Delta g}(t,T) = \mathrm{i}\frac{\Delta g(T)}{4T_\mathrm{R}(\omega_\mathrm{o} - \omega_\mathrm{oo})}\frac{\mathrm{d}}{\mathrm{d}t}a_\mathrm{s}(t,T) \quad (280)$$

where ω_o is the peak frequency of the net gain and ω_oo is the peak frequency of the gain medium.

The next physical effect is that of cavity length variation and the related effect of index fluctuations. The effective noise source is given by

$$S_{\Delta L}(t,T) = -\frac{\Delta L(T)}{T_\mathrm{R}}\left(\mathrm{i}\frac{\omega_\mathrm{o}}{c}n + \frac{1}{v_\mathrm{g}}\frac{\mathrm{d}}{\mathrm{d}t}\right)a_\mathrm{s}(t,T)\,. \quad (281)$$

The first term describes frequency pulling while the second term describes timing jitter. If we also add the effect of index variations, Δn then we simply replace ΔL by $L(\Delta n/n)$.

The final physical effect we take into account is that of spontaneous emission from the gain medium into the cavity. Spontaneous emission has a white spectral density with a correlation function

$$\left\langle S_\mathrm{qn}(T,t)S^*_\mathrm{qn}(T',t')\right\rangle = \xi\frac{2g}{T_\mathrm{R}}h\nu\delta(T-T')\delta(t-t') \quad (282)$$

where ξ is an enhancement factor that accounts for incomplete inversion of the atomic gain medium.

With the physical sources $S_{\Delta i}$ defined (279)–(282), we can project them onto the adjoint functions using the definition (276) to obtain the noise sources

$$S_\mathrm{w}(T) = 2\frac{\Delta g(T)}{T_\mathrm{R}}w_\mathrm{o} + S_{w,qn}(T) \quad (283)$$

$$S_\theta(T) = \frac{\omega_\mathrm{o}}{v_\mathrm{g}T_\mathrm{R}}\left[\Delta L(T) + L\frac{\Delta n(T)}{n}\right] + S_{\theta,qn}(T) \quad (284)$$

$$S_\mathrm{p}(T) = \frac{\Delta g(T)}{3T_\mathrm{R}(\omega_\mathrm{o} - \omega_\mathrm{oo})\tau^2} + S_{p,qn}(T) \quad (285)$$

$$S_\mathrm{t}(T) = \frac{\Delta g(T)}{T_\mathrm{R}\Omega_\mathrm{g}}\frac{1}{T_\mathrm{R}v_\mathrm{g}}\left[\Delta L(t) + L\frac{\Delta n(T)}{n}\right] + S_{t,qn}(T) \quad (286)$$

where the $S_{i,qn}(T)$ are the quantum noise sources with the correlation functions

$$\left\langle S_{i,qn}(T)S_{h,qn}(T')\right\rangle = D_{i,qn}\delta_{i,h}\delta(T-T') \quad (287)$$

and the following related diffusion coefficients

$$D_{w,qn} = 4w_o\xi\frac{2g}{T_R}h\nu \tag{288}$$

$$D_{\theta,qn} = \frac{4}{3w_o}\left(1+\frac{\pi^2}{12}\right)\xi\frac{2g}{T_R}h\nu \tag{289}$$

$$D_{p,qn} = \frac{8}{3w_o^2\tau^2}\xi\frac{2g}{T_R}h\nu \tag{290}$$

$$D_{t,qn} = \frac{\pi^2\tau^2}{3w_o}\xi\frac{2g}{T_R}h\nu\,. \tag{291}$$

We can now include (283)–(286) as the noise source terms in the equations of motion of the perturbation expansion (272)–(275).

At this point, (272)–(275) constitute a set of deterministic equations of motion for the coefficients, Δw, $\Delta\theta$, Δp and Δt of the perturbation expansion for Δa (258). If the noise sources were deterministic functions, we could solve these directly but, noise sources are random processes, and thus we must resort to statistical analyses using correlation and spectral density functions. This would take us down the correct path but would be a rather complicated addition to our already long story. Suffice it to say that Haus and Mecozzi have considered the implications of this by making reasonable assumptions about the spectral densities of the noise sources and have been able to predict the shapes of the spectra of the physical observables [65].

Large Scale Instability in Soliton Lasers

The soliton concept as applied to passively-modelocked lasers came about rather late in the general development of modelocking but was crucial to moving forward the understanding and engineering of this very important technology. Before discussing large-scale instability in soliton-like lasers, it is useful to take a brief look at instability in the other class of modelocking known as "active modelocking" [93,94]. In this approach, the net gain of the laser is modulated at the exact round-trip transit time within the cavity. This preferentially amplifies a section of the circulating fields which evolve into a pulse with pulsewidth on the order of (or better than) the width of the pumping pulse (within the limits set by the bandwidth of the gain medium). The gain itself can be directly modulated by, for example, a time varying pump, or a variable loss element such as an electro- or acousto-optic modulator inserted in the cavity. Another approach is to place an FM modulator in the cavity. A short pulse becomes synchronized with the peak of the modulation since the sides of the pulse receive strong FM modulation and the energy is shed outside of the bandwidth of the gain medium. From the standard theory of active modelocking [93], the pulsewidth is proportional to

$$\tau_p \propto \left(\frac{1}{\Delta_m}\right)^{1/4}\left(\frac{1}{f_m}\right)^{1/2} \tag{292}$$

where Δ_m is the modulation index (see Subsects. "Analytical Description of Envelope Noise"–"Fourier Analysis of Sinusoidal Phase Modulation") and f_m is the modulation frequency. The modulation index varies with drive power P_m as

$$\Delta_m \propto \begin{cases} P_m, & \text{acousto-optic AM modelockers}\,, \\ P_m^{1/2} & \text{electro-optic FM modelockers}\,. \end{cases} \tag{293}$$

Substituting these dependencies into (292) we find

$$\tau_p \propto \begin{cases} P_m^{-1/4} f_m^{-1/2} & \text{acousto-optic AM modelockers}\,, \\ P_m^{-1/8} f_m^{-1/2} & \text{electro-optic FM modelockers}\,. \end{cases} \tag{294}$$

We see that there is a greater advantage to operating at higher frequencies but the frequency is related to the round-trip cavity time $2L/c$ so they must be considered together. On the other hand, one can operate the modelocker at harmonics of the fundamental cavity rate and this succeeds at producing shorter pulses but with lower peak power since they also take on the same repetition rate as the modelocker frequency.

Early reports of instability in actively modelocked lasers included theoretical studies of pump laser and modelocked laser cavity length mismatch [95] and experimental studies studies showing period multiplication and division in a synchronously-pumped modelocked dye laser [96]. A synchronously-pumped modelocked laser is one which is pumped by another modelocked laser which has its cavity round-trip time (nominally) equal to the that of the pumped laser. This is a particular example of the technique of active modelocking described above. The authors varied both the pump laser intensity and the ratio of the pump laser cavity length to the modelocked laser cavity length. The interesting result from this investigation was the following; if the ratio of the pumping laser cavity length L to the length of the dye laser cavity L was a rational number, m/n, then the dye laser would produce a stable train of pulses at a rate of m/p times the pumping rate, where p is an integer which depends on the magnitude of the pump power. At high pump power, $p = 1$, and at low pump power, $p \to n$. This result is intuitively clear in the sense that the pump pulse defines the maximum gain point in time. If you plot the pulses with the period ratio as

a rational number, then the temporal overlap of the pulses gives the basic timing of the dye laser while if the gain is high enough, pulses in between will also form. In addition to frequency multiplication, period multiplication was also see to occur. The transitions between stable operating regimes could be considered as bifurcations and they often were accompanied by strong chaotic behavior just at the boundaries. Also, by making small ($\Delta L/L \approx 2 \times 10^{-4}$) adjustments in the dye laser cavity length, they found that the modelocking became highly irregular and chaotic.

MacFarlane and Casperson also found cavity length-dependent instability in an acousto-optically modelocked argon-ion laser [97]. Here, an acousto-optic amplitude modulator is placed within the cavity to act as a time-varying loss mechanism. An acoustic standing-wave field is established at half the round-trip cavity frequency so that every half cycle the fields collapse to zero and allow maximum optical transmission. This establishes the minimum loss point in the temporal dynamics of the laser system. They found that by detuning the acousto-optic modelocker frequency by as little as one part in 10^4, instability set in and further increasing the detuning resulted in strong spectral components at one fourth of the normal operating repetition rate (period quadrupling). Detuning by as much as 1.6×10^{-4} caused a new feature: very low frequency modulation (600 KHz) of the 81 MHz pulse train. Curiously, the trends in instability were not symmetric with respect to detuning. The strong period-four component was only present for negative relative detuning, whereas positive detuning caused an immediate change to chaotic pulsing and overall degradation of laser output. Scavennec, in a theoretical study [95], also found asymmetries with detuning. It is worth noting that the relative frequency detuning is of the same order as the relative cavity mismatch that led to chaos in [96]. Further interesting measurements in the time and radiofrequency domain were also carried out later by MacFarlane et al., [98].

Another type of instability can be seen when multiple modelockers are placed within an optical cavity to achieve shorter modelocked pulses than would otherwise be produced by one alone. Here we have an additional degree of freedom which is the relative phase of the signals driving the modulators. Scott et al., [99], placed an AM and FM modulator within a Nd:YAG laser cavity and drove the AM (acousto-optic) modulator at 40 MHz to produce fundamental modelocking at 80 MHz and the FM (electro-optic) modulator at 1.76 GHz to produce a chirp on the pulse. This combination of fundamental and high-harmonic FM modelocking achieved a pulsewidth reduction of 5.3 times over operating with the AM modelocker alone. When the relative phase between the RF drive signals to the two modelockers was adjusted, the output pulse was seen to move within the peak of the AM transmission window up to the point where there was insufficient gain to compensate for the intracavity losses. At this point, instability occurred as a new pulse tried to emerge at the maximum gain point, but with opposite chirp. The oppositely chirped pulse was less stable and had lower peak power. Further phase shift brought this pulse to the boundary of the gain > loss region and an unstable transition to a new pulse at the beginning of the gain > loss region occurred.

A large shift in the ultrafast laser community occurred around 1985 with the development of the colliding pulse modelocked dye laser incorporating intracavity dispersion compensation and adjustable saturable absorption [100]. It was recognized that there was an important interplay between the group velocity dispersion, the saturable gain and the saturable absorption and that by controlling them, optimally short pulses (< 27 fs) could be obtained. Shortly thereafter, issues of large scale instability began to receive attention and Avramopoulos et al., [101] presented theoretical and experimental results showing a causal link between certain values of intracavity dispersion and self-phase modulation and overall laser instability. The large scale amplitude fluctuations that were predicted and seen had modulation periods on the order of 100 μs which is many hundreds of round-trip cavity times.

The refinement of the twin jet colliding-pulse modelocked dye laser really set the stage for the solid state replacement a few years later where the gain medium was the Ti:sapphire crystal (with its extraordinary bandwidth) and the nonlinearity responsible for self-phase modulation and effective saturable absorption was the Kerr effect in the gain crystal itself [102]. Thus began the era of the all solid-state modelocked soliton laser. Analysis of how these lasers work proceeded rapidly and many approaches were developed, such as the master equation of Haus [64]. Later, workers also expanded the analysis to include spatial effects within the cavity [103,104,105,106]. We will not go into these techniques here since we wish to concentrate of issues of large scale instabilities.

The analytical development of the preceding Sect. "Noise in Soliton Lasers" focused on the impact of small-scale perturbations to the equations of motion for the soliton laser. To study the larger scale phenomena of instability, even more sophisticated techniques have been developed. For example, Tartwijk and Agrawal developed a modal analysis which expanded the electric field and polarization in a fiber laser in a series of the longitudinal modes and then applied the expansion to the Maxwell–Bloch equations [23]. This results in a family of Lorenz-type equations for each of the longitudinal mode expan-

sion coefficients and includes the effects of self-phase modulation, cross-phase modulation and four-wave mixing. Calculations were carried out for single mode operation but, unfortunately, they did not study the case of many modes required to simulate modelocked operation. (We conjecture that the computational resources required to accurately simulate hundreds or thousands of modes over long time scales would be formidable).

Sergeev et al., [107], studied soliton-like solutions of the Ginzburg–Landau equation (GLE, similar in form to the nonlinear Schrödinger equation (90)) which includes saturable absorption, group-velocity dispersion, finite gain-bandwidth and a Kerr nonlinearity. This equation has a hyperbolic-secant solution very similar to the master equation ansatz of Haus and Mecozzi (243). The proceeded to apply the "aberrationless approximation" which involves substituting the sech solution back into the GLE, assume the parameters for pulse amplitude, duration and chirp are functions of the propagation distance (cavity round trips, z) and then expand these expressions in a Taylor series to obtain new first-order differential equations for the parameters. Stationary equilibria are sought as a function of the saturation energy of the gain medium. By studying the solutions in the phase plane, they were able to determine an analytical criterion for stable soliton solutions in the weak absorber saturation approximation.

In 1993, Liu and Prucnal [108] discovered a new source of instability in the modelocked soliton laser. A key feature of the early generation of these lasers was the incorporation of an intracavity aperture, or slit, which preferentially passed a tightly focused beam and therefore was lossy for larger diameter beans. This forces the laser to self-modelock by virtue of the Kerr-lens effect which only produces lensing when a high peak-power pulse passes through the gain crystal. They found that by closing down the slit about 10% smaller than that for optimum pulsewidth, large amplitude modulation was obtained with depths approaching 100% and a modulation period of several μs. The process can be understood qualitatively as follows: when the slit is closed down past the size that produces the best pulses, the intracavity loss is increased because of the aperturing effect on the beam. Thus, the intracavity circulating power goes down and there is less stimulated emission in the gain medium. Less stimulated emission allows the population inversion to increase beyond its normal steady-state value which means the gain increases faster than the cavity loss rate and overshoots the steady-state value which would normally balance the static cavity losses. Now the fields start building exponentially fast which increases the Kerr effect and the self-focusing further reducing the cavity losses since the beam is now focused tighter through the narrowed aperture. At this point, two things happen. First, in overshooting the steady-state modelocked pulse amplitude, it generates a focus that is actually too small for the aperture in the sense that the divergence (convergence) angle increases and the focus shifts, which again starts to increase loss. Second, the higher field amplitude starts to eat up the population inversion by stimulated emission which also contributes to the downward trend of the field amplitude and thus the cycle repeats. The fluctuation period (or frequency) depends on the usual time constants (cavity and energy decay times) as well as the pumping rate, r. Just as in the relaxation oscillation analysis (Subsect. "Pump-Induced Noise and the Noise Transfer Function") the prime frequency of the induced modulation is very nearly the relaxation oscillation frequency (199).

If one couples an external cavity with a nonlinear element such as a single-mode fiber to the main laser cavity, the pulses in each will combine interferometrically and shorten owing to the nonlinear phase shift in the wings of the external cavity pulse. This is called "additive pulse modelocking" and was quite popular for a number of years. Because of the presence of the nonlinearity and the time-delayed feedback into the main laser cavity, this system is a perfect candidate for complex dynamics. Indeed, Sucha et al., [109] observed period doubling and quasi-periodic modelocking in a NaCl F-center laser. They also numerically simulated the laser system and found that as the fiber nonlinearity was increased from zero, the laser starts out in cw operation and then proceeds to stable modelocking. Beyond this point a bifurcation in the pulse energy takes place and, thus, period doubling. Further increases in the nonlinearity results in more period doubling and then an abrupt transition to chaos. In the period doubling region they found that the main cavity and external fiber cavity exchanged pulses on alternate round trips and these pulse had differing temporal and spectral character.

As ideas from chaos theory migrated into the laser community in the early days with studies on CW lasers, we see that they eventually found their way into analyses on modelocked lasers. The bifurcation and period doubling phenomena described above are good examples as well as the analysis by Sánchez and Hnilo who used Poincaré (iterative) maps in the complex plane to describe Kerr-lens modelocked soliton lasers [110]. We wish, now, to conclude this chapter with a brief discussion of what is happening in this field in recent years, because the emphasis has shifted back to issues of small-scale perturbation. In particular, the advent of the carrier-envelope stabilized, octave-spanning modelocked soliton laser has opened a new door onto the world of precision timekeeping and metrology and the fundamental mechanisms gov-

erning timing stability and laser linewidth are once again extremely important topics.

Laser Clocks

Researchers in the ultrafast optics community have worked tirelessly to make their laser pulses ever-shorter and evermore stable since the invention of the modelocking technique. Short pulses with a high degree of pulse-to-pulse stability are a powerful tool for time-domain studies of ultrafast phenomena in physics, chemistry, biology and electronics. But we can also appreciate this from a frequency-domain viewpoint. An ideal train of ultrashort laser pulses has an associated frequency spectrum composed of narrow, equally-spaced "comb-lines" with spacing equal to the inverse of the time between the pulses (see Fig. 16). If the exact frequency of one of the comb lines is known, and the repetition rate (interline spacing) is accurately know, then we have in effect a "ruler" in the frequency domain which can serve as a standard for measuring other optical sources or spectroscopic features. Recently, a revolution occurred in the field of laser physics when two teams reported a method for stabilizing the repetition rate of a modelocked laser using the "self-referencing" principle based on an $f-2f$ interferometer and an octave-spanning modelocked laser source [111,112]. This invention has heralded a new era in precision time and frequency metrology and new laser-based clockworks are now being created all over the world. Indeed, two out of three recipients of the 2005 Nobel Prize in Physics were recognized for their contributions to this invention and the significance of the achievement [113]. These laser systems have now demonstrated more than three orders of magnitude better timing stability than the hyperfine transition of the cesium atom which forms the basis of the currently accepted definition of the second. They are also demonstrating one to two orders of magnitude better stability than some of the other more exotic atomic standards such as the hydrogen maser and some trapped ions. In essence, the frontiers of time and frequency metrology have just advanced by more than two orders of magnitude in the short span of several years. (For good reviews on the evolution of this laser technique and precision spectroscopy in general, see [19,111,114] and references therein).

Modelocked Laser Comb Lines and Frequency Metrology

A key concept in understanding the aforementioned revolution in modelocked lasers is the "carrier-envelope offsetj" in the pulses that exit the laser cavity. Optical pulses can be decomposed into their carrier and their envelopes. The carrier propagates at the phase velocity in a medium determined by the index of refraction, $v_\mathrm{p} = c/n(\omega)$. The envelope propagates at the group velocity $v_\mathrm{g} = \mathrm{d}\omega/\mathrm{d}\beta$. Thus, for a small propagation distance δz, there will be a phase slip between the center of mass of the envelope and the carrier given by

$$\delta\phi_\mathrm{CEO} = \left(\frac{1}{v_\mathrm{g}} - \frac{1}{v_\mathrm{p}}\right)\omega_\mathrm{c}\delta z \tag{295}$$

where the subscript "CEO" stands for "carrier-envelope-offset". If we integrate this function along the whole path of the laser cavity, we arrive at a total round-trip carrier-envelope phase offset

$$\Delta\phi_\mathrm{CEO} = \left(\frac{1}{v_{\mathrm{g}_\mathrm{avg}}} - \frac{1}{v_{\mathrm{p}_\mathrm{avg}}}\right)\omega_\mathrm{c}p_\mathrm{c} \quad \text{modulo } 2\pi \tag{296}$$

where v_{g_avg} and v_{p_avg} are the path-averaged group and phase velocities and p_c is the total round-trip cavity path length. We include in our definition the modulo 2π since the total phase shift can be several hundred optical cycles but the electric field is 2π periodic and thus the net number of cycles is not seen [115]. The frequency of the recurring carrier-envelope phase offset is given by

$$f_\mathrm{CEO} = \frac{1}{2\pi}\frac{\mathrm{d}\phi_\mathrm{CEO}}{\mathrm{d}t} = \frac{1}{2\pi}\frac{\Delta\phi_\mathrm{CEO}}{T_\mathrm{c}} = \frac{\Delta\phi_\mathrm{CEO}}{2\pi}f_\mathrm{REP} \tag{297}$$

where f_REP is the repetition rate of the envelope of the modelocked pulse train. The relationship between the carrier and the envelope timing is shown in Fig. 37a. In Fig. 37b we see how the frequencies are related to each other. The dashed lines show harmonics of the fundamental repetition rate while the solid lines are the actual comb spectrum of the laser. The frequency of any comb line is given by $\nu_n = nf_\mathrm{REP} + f_\mathrm{CEO}$.

There are two requirements for using this frequency comb as a high precision basis for optical frequency measurements: 1. the comb has to be stabilized against fluctuations in f_REP and 2. f_CEO must be determined. Then by using a course frequency measuring instrument (such as a spectrometer or wavemeter) one can correctly find the integer n and all that needs to be done is to accurately determine f_CEO to have a very precise optical frequency measuring instrument.

The method of stabilizing f_REP by using an octave-spanning spectrum and cavity feedback control by self-referencing is exceptionally clever [111]. First, we must prepare a spectrum of comb lines that spans an octave. This was originally done using a highly nonlinear fiber as

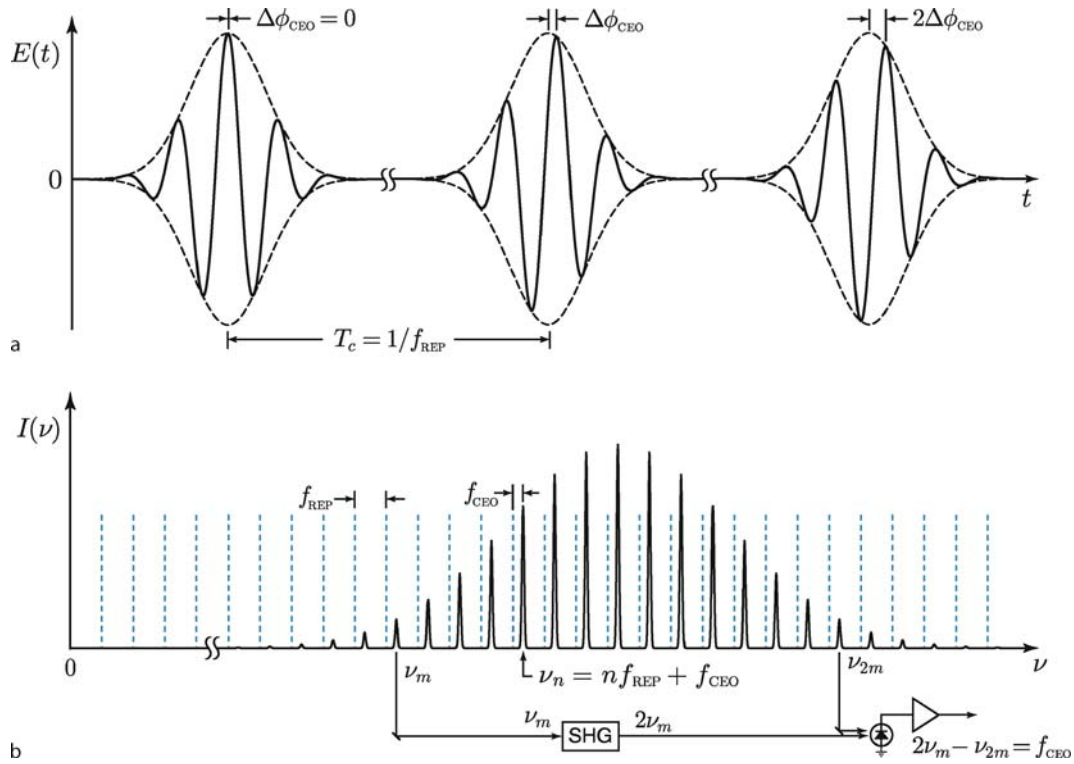

Noise and Stability in Modelocked Soliton Lasers, Figure 37

a Time domain representation of the laser pulse electric field showing the carrier-envelope phase offset $\Delta\phi_{CEO}$. b Laser power spectrum showing comb lines separated by the cavity repetition frequency f_{REP}. The absolute frequencies of the comb lines are shifted from integer multiples of f_{REP} by the carrier-envelope offset frequency f_{CEO}. Also shown is the $f - 2f$ interferometer for absolute frequency stabilization which is based on the premise that the spectrum spans at least one octave so that a low frequency component ν_m can be frequency doubled and beat against a nearby comb line at ν_{2m}

which caused dramatic spectral broadening due to self-phase modulation. In fact, the development of specialty microstructure hollow-core or "photonic crystal" fiber was the key to the birth of the new technique (see [116,117] and references therein). Suppose we spread out the spectrum with a diffraction grating, pick off a comb line at the low frequency (infrared) side, ν_m and frequency double it with a second harmonic generation crystal (SHG). Simultaneously, we pick a comb line at the high frequency side of the spectrum, which is exactly twice the index number, $2m$, of the low frequency comb. We then combine the light and shine it on a photodiode. The current that results is a sinusoid at the beat or difference frequency between the two. That is,

$$\begin{aligned}\Delta f &= 2\nu_m - \nu_{2m} \\ &= 2(mf_{REP} - f_{CEO}) - (2mf_{REP} - f_{CEO}) = f_{CEO}\,.\end{aligned} \tag{298}$$

This provides us with a direct measure of the carrier offset frequency and thus knowledge of the actual frequency of any comb line. More importantly, that difference frequency, f_{CEO} can be used as an error signal to control the cavity length and the net difference in group and phase velocities and thus drive $f_{CEO} \to 0$. This is the "self-referencing" that has revolutionized the stability of the modelocked lasers and transformed the field of time and frequency metrology [17,18,19,111,114]. As can be seen in the explosion of literature on the subject in recent years, there are many new applications for this technique, both for science and engineering, and thus questions about ultimate stability, linewidth, noise, etc., are currently a hot area of research. We will now briefly touch on a few of the most recent results.

It was recognized fairly early on that noise accompanying the pump source could be a significant source of noise in the Kerr-lens modelocked lasers (as well as any other laser) [81,82,83,84,85]. As discussed in Subsect. "Pump-Induced Noise and the Noise Transfer Function" of this chapter, the principal mechanisms for coupling pump amplitude fluctuations to modelocked laser noise are population inversion (gain) modulation for am-

plitude noise and 1. direct index modulation, 2. thermal modulation, 3. nonlinear effects, and 4. beam-pointing instability due to thermal and direct index fluctuations for the phase noise. Scott et al., [85], have been primarily concerned with the effect on the modelocked laser envelope while Xu et al., [118] Helbing et al., [81] and Matos et al., [119] studied the effect on the carrier-envelope offset frequency. The former is important for free-running (non-locked) lasers while the latter is important for the self-referenced and locked Kerr-lens lasers. The frequency-dependence of the coupling mechanisms is described in detail in Subsect. "Pump-Induced Noise and the Noise Transfer Function". We also note here that, since the low frequency phase sensitivity increases with pump power as the frequency is lowered (to DC), the pump power can be used to effectively control the modelocking frequency [120].

A detailed study of the variation in absolute and relative comb frequencies was carried out by Holman et al., [121]. They derived expressions for the sensitivity of f_{REP} and f_{CEO} to the peak circulating pulse intensity, I by differentiating

$$\frac{\mathrm{d}f_{\mathrm{REP}}}{\mathrm{d}I} = \frac{1}{L_{\mathrm{c}}}\frac{\mathrm{d}v_{\mathrm{g}}}{\mathrm{d}I} \tag{299}$$

$$\frac{\mathrm{d}f_{\mathrm{CEO}}}{\mathrm{d}I} = \frac{1}{2\pi}\frac{\partial\omega_{\mathrm{c}}}{\partial I}\left(1-\frac{v_{\mathrm{g}}}{v_{\mathrm{p}}}\right) + \frac{\omega_{\mathrm{c}}}{2\pi}\frac{v_{\mathrm{g}}}{v_{\mathrm{p}}}\left(\frac{1}{v_{\mathrm{p}}}\frac{\mathrm{d}v_{\mathrm{p}}}{\mathrm{d}I}-\frac{1}{v_{\mathrm{g}}}\frac{\mathrm{d}v_{\mathrm{g}}}{\mathrm{d}I}\right) \tag{300}$$

where ω_{c} is the spectrally-weighted center frequency and thus $\partial\omega_{\mathrm{c}}/\partial I$ is the intensity-driven shift in the optical spectrum. The authors treat only the effect of the variation of the indices of refraction with pulse intensity by letting $\bar{n} \equiv \bar{n}_0 + \bar{n}_2 I$ be the average refractive index within the laser cavity such that in one round trip, $\bar{n}L_{\mathrm{c}}$ gives the true optical path length of air plus Ti:sapphire crystal. With this definition, we can now define an effective phase and group velocity,

$$v_{\mathrm{p}} = c/\bar{n}\,, \quad v_{\mathrm{g}} = c/\left[\bar{n} + \omega_{\mathrm{c}}(\mathrm{d}\bar{n}/\mathrm{d}\omega)_{\mathrm{c}}\right]\,. \tag{301}$$

Now, carrying out the indicated differentiations in (299) and (300), we find

$$\frac{\mathrm{d}f_{\mathrm{REP}}}{\mathrm{d}I} = -\frac{1}{L_{\mathrm{c}}}\frac{v_{\mathrm{g}}^2}{c}\left[\bar{n}_2 + \omega_{\mathrm{c}}\left(\frac{\mathrm{d}\bar{n}_2}{\mathrm{d}\omega}\right)_{\omega_{\mathrm{c}}} + c\frac{\partial\omega_{\mathrm{c}}}{\partial I}\frac{\partial}{\partial\omega_{\mathrm{c}}}\left(\frac{1}{v_{\mathrm{g}}}\right)\right] \tag{302}$$

$$\frac{\mathrm{d}f_{\mathrm{CEO}}}{\mathrm{d}I} = \frac{\omega_{\mathrm{c}}^2}{2\pi}\frac{v_{\mathrm{g}}^2}{c^2}\left[\bar{n}_0\left(\frac{\mathrm{d}\bar{n}_2}{\mathrm{d}\omega}\right)_{\omega_{\mathrm{c}}} - \bar{n}_2\left(\frac{\mathrm{d}\bar{n}_0}{\mathrm{d}\omega}\right)_{\omega_{\mathrm{c}}}\right] + \frac{1}{2\pi}\frac{\partial\omega_{\mathrm{c}}}{\partial I}\left[\left(1-\frac{v_{\mathrm{g}}}{v_{\mathrm{p}}}\right) + \frac{\omega_{\mathrm{c}}v_{\mathrm{g}}^2}{v_{\mathrm{p}}}\frac{\partial}{\partial\omega_{\mathrm{c}}}\left(\frac{1}{v_{\mathrm{g}}}\right) - \frac{\omega_{\mathrm{c}}v_{\mathrm{g}}}{c}\frac{\partial\bar{n}}{\partial\omega_{\mathrm{c}}}\right]\,. \tag{303}$$

All of the terms in (302) and (303) are available in the literature except $\partial\omega_{\mathrm{c}}/\partial I$ and $\partial v_{\mathrm{g}}^{-1}/\partial\omega_{\mathrm{c}}$. However, two sets of experimental data can uniquely determine these quantities. In their experiments, the authors placed an acousto-optic amplitude modulator (AOM) between the pump laser and a pair of KLM Ti:sapphire lasers; one configured for 100 MHz and one configured for 750 MHz. For slow pump variations, standard frequency counters were used to monitor f_{REP} and f_{CEO}. For higher frequency sinusoidal pump modulations, lockin receivers were used (i. e. homodyne receivers with local oscillators derived from the AOM drive signal). For static changes in the pump power, a local minimum in the center frequency, ω_{c} of the pulse spectrum was measured. That is, $\partial\omega_{\mathrm{c}}/\partial I$ undergoes a sign change with I. This would imply a possible related zero in $\mathrm{d}f_{\mathrm{CEO}}/\mathrm{d}I$ (see Eq. 303) above) and indeed, this was confirmed in the data. Possible explanations for the shift in the spectrum with pump power include a possible offset between the gain and the loss bands within the cavity. As the pump power is changed, the gain goes up and down but the losses stay fixed, The equilibrium point for lasing thus becomes frequency-dependent, if the two curves do not exactly overlap. However, this does not provide an explanation for the local minimum. It is interesting, though, and the authors rightly point out that there are strong implications for choosing design parameters in the Ti:sapphire laser to try and minimize the sensitivity to pump fluctuations.

The authors followed the DC characterizations of Δf_{REP} and Δf_{CEO} with dynamic measurements by driving the AOM from 10 Hz to 400 KHz. Unlike the noise transfer function measurements presented in Subsect. "Pump-Induced Noise and the Noise Transfer Function", there was not a significant variation with frequency until the very lowest frequency portion of the spectrum (approximately 1 KHz), which they attribute to thermal effects. Once again, there is a definite sign dependence on the average pump power (and hence the circulating peak power). For average output powers below 750 mW, the slope $\Delta f_{\mathrm{CEO}}/\Delta I > 0$ whereas for output powers above 750 mW, $\Delta f_{\mathrm{CEO}}/\Delta I < 0$, which was consistent with the DC data.

Another very interesting result from this work is a study of the nature of the lineshape of the CEO signal. The authors studied this with a radiofrequency spectrum analyzer and found that the linewidth was narrowest at the zero-crossing point of $\Delta f_{\mathrm{CEO}}/\Delta I$ when the AOM was driven at 10 KHz. The correlation between the minimum linewidth and a zero-crossing for 10 KHz modulation is simply explained by noting that the zero-crossing implies minimum sensitivity to pump fluctuations for frequencies above the thermal response. The paper concludes with a study of the CEO signal linewidth as measured on an RF spectrum analyzer as a function of laser bandwidth. Unfortunately, this type of measurement is not conclusive as the signal-to-noise ratio and the simple technique did not afford unambiguous results. A better method would have been to separate the AM from the PM noise, measuring the AM at baseband and the PM about the f_{CEO} carrier using a quadrature mixing technique [79].

Conclusions and Outlook for the Future

The preceding sections of the chapter have really only glossed over the many topics related to noise and instability in lasers, and to modelocked soliton lasers, specifically. The author felt that a deeper appreciation could be gained if the reader were exposed to some of the details of fundamental noise sources like shot noise, the Schawlow–Townes linewidth, the master modelocking equation, as well as the Lorenz–Haken equations. These form the basis for most of the serious studies of noise and instability today. While already long, this chapter has largely neglected topics having to do with fiber lasers [122,123], issues of polarization properties of vector solitons in modelocked fiber lasers and instabilities called "exploding solitons" [124], spatial instabilities [16,23] and instabilities in semiconductor lasers [23]. These omissions were made not because they lacked merit, but because space was fundamentally limited. To those scientists in their respective disciplines, I apologize.

In just the past few years, the field has seen a reinvigoration with new ideas and a quest for determining theoretically and, hopefully, demonstrating experimentally, the ultimate limits in modelocked laser stability. New results on pump power fluctuations causing relaxation oscillations and frequency pulling can be found in [67]. The real "holy grail", of course, would be a quantum mechanically correct expression for the frequency stability or, equivalently, the modal linewidth of a modelocked soliton laser, and which would be readily accessible to experimental verification.

Acknowledgments

The author is greatly indebted to many colleagues, friends and teachers who have influenced me over the years. I wish to thank Dr. Steven T. Cundiff and Professor Curtis R. Menyuk for their tremendous contributions and stimulating conversations regarding their recent efforts in understanding this vast field. To Dr. John L. Hall I owe a great and hearty thanks for all of his pioneering work, for his patience as a sounding board and his unique perspective on physics. Thanks also to Professors Anthony E. Siegman, Stephen E. Harris, Robert L. Byer, and especially, Alwyn C. Scott, from whom I learned much. Finally, on behalf of my group at the University of California, I wish to thank Robert Temple, Tom Faulkner and Roger Muat of Agilent Technologies (formerly Hewlett-Packard Company) for numerous donations and countless hours of tutoring. This work was supported in part by the David and Lucile Packard Foundation and the National Science Foundation under grant ECS-0622235.

Bibliography

1. Barnes JA, Chi AR, Cutler LS, Healey DJ, Leeson DB, McGunigal TE, Mullen JA Jr, Smith WL, Sydnor RL, Vessot RFC, Winkler GMR (1971) Characterization of frequency stability. IEEE Trans Instrum Meas IM-20:105–120
2. Kroupa VF (ed) (1983) Frequency stability: Fundamentals and measurement. IEEE Press, New York
3. Sullivan DB, Allan DW, Howe DA, Walls FL (eds) (1990) Characterization of clocks and oscillators: NIST Technical Note 1337. United States Government Printing Office, Washington
4. Siegman AE (1986) Lasers. University Science Books, Mill Valley
5. Milonni PW, Shih M, Ackerhalt JR (1987) Chaos in Laser-Matter Interactions. World Scientific, Singapore
6. Narducci LM, Abraham NB (1988) Laser Physics and Laser Instabilities. World Scientific, Singapore
7. Saleh BEA, Teich MC (1991) Fundamentals of Photonics. Wiley, New York
8. Svanberg S (1992) Atomic and Molecular Spectroscopy, 2nd edn. Springer, Berlin
9. Demtröder W (1996) Laser Spetroscopy, 2nd edn. Springer, Berlin
10. Yariv A (1997) Optical Electronics in Modern Communications, 5th edn. Oxford University Press, Oxford
11. Koechner W (1996) Solid State Laser Engineering, 4th edn. Springer, Berlin
12. Haken H (1985) Light. In: Laser Dynamics, vol 2. North-Holland, Amsterdam
13. Boyd RW, Raymer MG, Narducci LM (eds) (1986) Optical Instabilities. In: Proceedings of the International Meeting on Instabilities and Dynamics of Lasers and Nonlinear Optical Systems, University of Rochester, 18–21 June, 1985. Cambridge University Press, Cambridge
14. Abraham NB, Mandel P, Narducci LM (1988) Dynamical instabilities and pulsations in lasers. In: Wolf E (ed) Progress in Optics, vol XXV, ch. I. North-Holland, Amsterdam, pp 1–190

15. Robins WP (1982) Phase Noise in Signal Sources. Peter Peregrinus, London
16. Weiss CO, Vilaseca R (1991) Dynamics of Lasers. Weinheim, New York
17. Riehle F (2004) Frequency Standards. Wiley-VCH, Weinheim
18. Kärtner FX, Morgner U, Schibli T, Ell R, Haus HA, Fujimoto JG, Ippen EP (2004) Few-Cycle Pulses Directly from a Laser. In: Topics in Applied Physics, vol 95. Springer, Berlin, pp 73–136
19. Ye J, Cundiff ST (eds) (2005) Femtosecond Optical Frequency Comb Technology: Principle, Operation and Application. Springer, New York
20. Jackson JD (1999) Classical Electrodynamics, 3rd edn. Wiley, New York
21. Slichter CP (1963) Principles of Magnetic Resonance. Harper and Row, New York
22. Hecht E (2002) Optics, 4th edn. Addison-Wesley, San Francisco
23. van Tartwijk GHM, Agrawal GP (1998) Laser instabilities: a modern perspective. Prog Quant Elect 22:43–122
24. Dunsmuir R (1961) Theory of relaxation oscillations in optical masers. J Electron Contr 10:453–458
25. Begon M, Harper JL, Townsend CR (1986) Ecology: Individuals, Populations, and Communities. Sinauer Assoc, Sunderland
26. Davis HT (1962) Introduction to Nonlinear Differential and Integral Equations. Dover, New York
27. Kaplan JI, Zier R (1962) Model for transient oscillations in a three-level optical maser. J Appl Phys 33:2372–2375
28. Bostick HA, O'Connor JR (1962) Infrared oscillations from CaF_2:U^{+3} and BaF_2:U^{+3} masers. Proc IRE 50:219–220
29. Tang CL (1963) On maser rate equations and transient oscillations. J Appl Phys 34:2935–2940
30. Lorenz EN (1963) Deterministic nonperiodic flow. J Atmos Sci 20:130–141
31. Gleick J (1987) Chaos. Penguin, New York
32. Schuster HG, Just W (2005) Deterministic Chaos, 4th edn. Wiley-VCH, Weinheim
33. Haken H (1975) Analogy between higher instabilities in fluids and lasers. Phys Lett 53A:77–78
34. Casperson LW (1978) Spontaneous coherent pulsations in laser oscillators. IEEE J Quantum Electron QE-14:756–761
35. Mayr M, Risken H, Vollmer HD (1981) Periodic and chaotic breathing of pulses in a ringlaser. Optics Comm 36:480–482
36. Weiss CO, Klische W (1984) On observability of Lorenz instabilities in lasers. Optics Comm 51:47–48
37. Lugiato LA, Narducci LM, Bandy DK, Pennise CA (1983) Breathing, spiking and chaos in a laser with injected signal. Opt Commun 46:64–68
38. Weiss CO, Godone A, Olafsson A (1983) Routes to chaotic emission in a cw He-Ne laser. Phys Rev A 28:892–895
39. Arrechi FT, Lippi GL, Puccioni GP, Tredicce JR (1984) Deterministic chaos in laser with injected signal. Optics Comm 51:308–314
40. Shih M, Milonni PW, Ackerhalt JR (1985) Modeling laser instabilities and chaos. J Opt Soc Amer B 2:130–135
41. Narducci LM, Sadiky H, Lugiato LA, Abraham NB (1985) Experimentally accessible periodic pulsations of a single-mode homogeneously broadened laser (the Lorenz model). Optics Comm 55:370
42. Weiss CO, Brock J (1986) Evidence for Lorenz-type chaos in a laser. Phys Rev Lett 57:2804–2806
43. Pujol J, Laguarta F, Vilaseca R, Corbalán R (1988) Influence of pump coherence on the dynamic behavior of a laser. J Opt Soc Amer B 5:1004–1010
44. Lauterborn W, Steinhoff R (1988) Bifurcation structure of a laser with pump modulation. J Opt Soc Amer B 5:1097–1104
45. Brunner W, Fischer R, Paul H (1988) Time evolution of the total electric-field strength in multimode lasers. J Opt Soc Amer B 5:1139–1143
46. Khanin YI (1988) Mechanisms of nonstationary behavior of solid-state lasers. J Opt Soc Amer B 5:889–898
47. Abraham NB, Allen UA, Peterson E, Vicens A, Vilaseca R, Espinosa V, Lippi GL (1995) Structural similarities and differences among attractors and their intensity maps in the Laser-Lorenz model. Optics Comm 117:367–384
48. Smith CP, Dykstra R (1996) Observation in the two-level spatial Maxwell-Bloch model of the anomalously large first peak as seen in experimental Lorenz-like spiral chaos from the $^{15}NH_3$ laser. Optics Comm 129:69–74
49. Lauterborn W, Kurz T (2003) Coherent Optics, 2nd edn. Springer, Berlin
50. Mørk J, Mark J, Tromborg B (1990) Route to chaos and competition between relaxation oscillations for a semiconductor laser with optical feedback. Phys Rev Lett 65:1999–2002
51. Hentschel M, Kienberger R, Spielmann C, Reider GA, Milosevic N, Brabec T, Corkum P, Heinzmann U, Drescher M, Krausz F (2001) Attosecond metrology. Nature 414:509–513
52. Diels J-C, Rudolf W (1996) Ultrashort Laser Pulse Phenomena. Academic Press, San Diego
53. Scott A (2003) Nonlinear Science; Emergence and Dynamics of Coherent Structures, 2nd edn. Oxford University Press, Oxford
54. Korteweg DJ, deVries G (1895) On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves. Philos Mag Ser 5 39:422–443
55. Russell JS (1838) Report of the committee on waves. Report of the 7th Meeting of the British Association for the Advancement of Science, pp 417–496
56. Lamb GL Jr (1980) Elements of Soliton Theory. Wiley, New York
57. Ablowitz MJ, Segur H (1981) Solitons and the Inverse Scattering Transform. Society for Industrial and Applied Mathematics, Philadelphia
58. Zabusky NJ, Kruskal MD (1965) Interaction of "solitons in a collisionless plasma and the recurrence of initial states". Phys Rev Lett 15:240–243
59. Scott AC, Chu FYF, McLaughlin DW (1973) The soliton – a new concept applied science. Proc IEEE 61:1443–1483
60. Hasegawa A, Tappert F (1973) Transmission of stationary nonlinear optical pulses in dispersive dielectric fibers. I. Anonalous dispersion. Appl Phys Lett 23:142–144
61. Mollenauer LF, Stolen RH, Gordon JP (1980) Experimental observation of picosecond pulse narrowing and solitons in optical fibers. Phys Rev Lett 45:1095–1098
62. Haus HA (1975) Theory of mode locking with a fast saturable absorber. J Appl Phys 46:3049–3058
63. Haus HA, Fujimoto JG, Ippen EP (1991) Structures for additive pulse modelocking. J Opt Soc Amer B 8:2068–2076
64. Haus HA, Fujimoto JG, Ippen EP (1992) Analytic theory of additive pulse and Kerr lens mode locking. IEEE J Quantum Electron 28:2086–2096

65. Haus HA, Mecozzi A (1993) Noise of mode-locked lasers. IEEE J Quantum Electron 29:983–996
66. Kapitula T, Kutz JN, Sandstede B (2002) Stability of pulses in the master mode-locking equation. J Opt Soc Amer B 19:740–746
67. Menyuk CR, Wahlstrand JK, Willits J, Smith RP, Schibli TR, Cundiff ST (2007) Pulse dynamics in mode-locked lasers: relaxation oscillations and frequency pulling. Opt Express 15:6677–6689
68. Papoulis A (1965) Probability, Random Variables, and Stochastic Processes. McGraw-Hill, New York
69. Cooper GR, McGillem CD (1971) Probabilistic Methods of Signal and System Analysis, 2nd edn. Holt, Rinehart and Winston, Fort Worth
70. Schawlow AL, Townes CH (1958) Infrared and optical masers. Phys Rev 112:1940–1949
71. Salomon C, Hils D, Hall JL (1988) Laser stabilization at the millihertz level. J Opt Soc Amer B 5:1576–1587
72. Eichenseer M, von Zanthier J, Walther H (2005) Common-mode-free frequency comparison of lasers with relative frequency stability at the millihertz level. Opt Lett 30:1662–1664
73. Notcutt M, Ma L-S, Hall JL (2005) Simple and compact 1 Hz laser system via an improved mounting configuration of a reference cavity. Opt Lett 30:1815–1817
74. Loudon R (1973) The Quantum Theory of Light. Oxford University Press, Oxford
75. Kingston RH (1995) Optical Sources, Detectors, and Systems. Academic Press, San Diego
76. Feller W (1957) An Introduction to Probability Theory and Its Applications. Wiley, New York, p 168
77. Gradshteyn IS, Ryzhik IM (1980) Table of Integrals, Series, and Products, corrected and enlarged 4th edn. Academic Press, New York
78. Abramowitz M, Stegun IA (eds) (1965) Handbook of Mathematical Functions, 55. US Department of Commerce, National Bureau of Standards, Washington
79. Scott RP, Langrock C, Kolner BH (2001) High dynamic range laser amplitude and phase noise measurement techniques. IEEE J Select Topics Quant Electron 7:641–655
80. Kluge J, Wiechert D, Linde DV (1984) Fluctuations in synchronously modelocked dye lasers. Opt Commun 51:271–277
81. Helbing FW, Steinmeyer G, Keller U, Windeler RS, Stenger J, Telle HR (2002) Carrier-envelope offset dynamics of mode-locked lasers. Opt Lett 27:194–196
82. Scott RP, Kolner BH, Langrock C, Byer RL, Fejer MM (2003) Ti:sapphire laser pump-noise transfer function. In: Proceedings of the Conference on Lasers and Electro-optics, Paper CFB2, Baltimore
83. Kolner BH, Scott RP, Langrock C (2003) Laser phase noise degradation from thermal effects due to pump power fluctuations. In: Proceedings of the 2003 IEEE/LEOS Summer Topical Meeting on Photonic Time/Frequency Measurement and Control, Paper TuC3.3, Vancouver, Institute of Electrical and Electronics Engineers, 14–16 July
84. Mulder TD, Scott RP, Baker KA, Kolner BH (2007) Characterization of the complex noise transfer function of a modelocked Ti:sapphire laser. In: Conference on Lasers and Electro-Optics (CLEO 2007), Baltimore, Paper JThD38
85. Scott RP, Mulder TD, Baker KA, Kolner BH (2007) Amplitude and phase noise sensitivity of modelocked Ti:sapphire lasers in terms of a complex noise transfer function. Opt Express 15:9090–9095
86. Weber R, Neuenschwander B, Donald MM, Roos MB, Weber HP (1998) Cooling schemes for longitudinally diode laser-pumped Nd:YAG rods. IEEE J Quantum Electron 34:1046–1053
87. Kolner BH (2008) Dynamic temperature distribution in cylindrical laser rods with time-varying pump sources. (in preparation)
88. Kolner BH, Mulder TD, Scott RP (2008) Laser noise modulation transfer functions. (in preparation)
89. Mulder TD, Scott RP, Kolner BH (2008) Amplitude and envelope phase noise of a modelocked laser predicted from its noise transfer function and the pump noise power spectrum. Opt Express 16(18):14186–14191
90. Bender CM, Orszag SA (1978) Advanced Mathematical Methods for Scientists and Engineers. McGraw-Hill, New York
91. Haus HA, Lai Y (1990) Quantum theory of soliton squeezing: a linearized approach. J Opt Soc Amer B 7:386–392
92. Haberman R (1987) Elementary Applied Partial Differential Equations with Fourier Series and Boundary Value Problems, 2nd edn. Prentice-Hall, New Jersey
93. Kuizenga DJ, Siegman AE (1970) FM and AM mode locking of the homogeneous laser – Part I: Theory. IEEE J Quantum Electron QE-6:694–708
94. Kuizenga DJ, Siegman AE (1970) FM and AM mode locking of the homogeneous laser – Part II: Experimental results in a Nd:YAG laser with internal FM modulation. IEEE J Quantum Electron QE-6:709–715
95. Scavennec A (1976) Mismatch effects in synchronous pumping of the continuously operated mode-locked laser. Optics Comm 17:14–17
96. Zheng JP, Sen U, Benenson DM, Kwok HS (1986) Observation of periodicity multiplication in a synchronously pumped dye laser. Opt Lett 11:632–634
97. MacFarlane DL, Casperson LW (1987) Pulse-train instabilities in a mode-locked argon laser: Experimental studies. J Opt Soc Amer B 4:1777–1780
98. MacFarlane DL, Casperson LW, Tovar AA (1988) Spectral behavior and pulse train instabilities of a synchronously pumped mode-locked dye laser. J Opt Soc Amer B 5:1144–1152
99. Scott RP, Bennett CV, Kolner BH (1997) AM and high-harmonic FM modelocking. Appl Opt 36:5908–5912
100. Valdmanis JA, Fork RL, Gordon JP (1985) Generation of optical pulses as short as 27 femtoseconds directly from a laser balancing self-phase modulation, group-velocity dispersion, saturable absorption, and saturable gain. Opt Lett 10:131–133
101. Avramopoulos H, French PMW, Williams JAR, New GHC, Taylor JR (1988) Experimental and theoretical studies of complex pulse evolutions in a passively mode-locked ring dye laser. IEEE J Quantum Electron 24:1884–1892
102. Spence DE, Kean PN, Sibbett W (1991) 60-fsec pulse generation from a self-modelocked Ti-sapphire laser. Opt Lett 16:42–44
103. Christov IP, Kapteyn HC, Murnane MM, Huang C-P, Zhou J (1995) Space-time focusing of femtosecond pulses in a Ti:sapphire laser. Opt Lett 20:309–311
104. Kalosha VP, Müller M, Herrmann J, Gatz S (1998) Spatiotemporal model of femtosecond pulse generation in Kerr-lens mode-locked solid-state lasers. J Opt Soc Amer B 15:535–550

105. Christov IP, Stoev VD (1998) Kerr-lens mode-locked laser model: role of space time effects. J Opt Soc Amer B 15:1960–1966
106. Jirauschek C, Kärtner FX, Morgner U (2003) Spatiotemporal Gaussian pulse dynamics in Kerr-lens mode-locked lasers. J Opt Soc Amer B 20:1356–1368
107. Sergeev AM, Vanin EV, Wise FW (1997) Stability of passively modelocked lasers with fast saturable absorbers. Optics Comm 140:61–64
108. Liu Y-M, Prucnal PR (1993) Slow amplitude modulation in the pulse train of a self-modelocked Ti:sapphire laser. IEEE J Quantum Electron 29:2663–2669
109. Sucha G, Bolton SR, Weiss S, Chemla DS (1995) Period doubling and quasi-periodicity in additive-pulse mode-locked lasers. Opt Lett 20:1794–1796
110. Sánchez LM, Hnilo AA (2001) Description of Kerr lens mode-locked lasers with Poincaré maps in the complex plane. Optics Comm 199:189–199
111. Hall JL, Ye J, Diddams SA, Ma L-S, Cundiff ST, Jones DJ (2001) Ultrasensitive spectroscopy, the ultrastable lasers, the ultrafast lasers, and the seriously nonlinear fiber: a new alliance for physics and metrology. IEEE J Quantum Electron 37:1482–1492
112. Holzwarth R, Zimmermann M, Udem T, Hänsch TW (2001) Optical clockworks and the measurement of laser frequencies with a mode-locked frequency comb. IEEE J Quantum Electron 37:1493–1501
113. http://nobelprize.org/physics/laureates/2005/
114. Cundiff ST, Ye J (2003) Femtosecond optical frequency combs. Rev Modern Phys 75:325–342
115. Helbing FW, Steinmeyer G, Stenger J, Telle HR, Keller U (2002) Carrier-envelope-offset dynamics and stabilization of femtosecond pulses. Appl Phys B 74:S34–S42
116. Ranka J, Windeler R, Stentz A (2000) Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. Opt Lett 25:25–27
117. Ranka JK, Windeler RS, Stentz AJ (2000) Optical properties of high-delta air silica microstructure optical fibers. Opt Lett 25:796–798
118. Xu L, Spielmann C, Poppe A, Brabec T, Krausz F, Hänsch TW (1996) Route to phase control of ultrashort light pulses. Opt Lett 21:2008–2010
119. Matos L, Mücke OD, Jian C, Kärtner FX (2006) Carrier-envelope phase dynamics and noise analysis in octave-spanning Ti:sapphire lasers. Opt Express 14:2497–2511
120. Stenger J, Talle HR (2000) Intensity-induced mode shift in a femtosecond laser by a change in the nonlinear index of refraction. Opt Lett 25:1553–1555
121. Holman KW, Jones RJ, Marian A, Cundiff ST, Ye J (2003) Detailed studies and control of intensity related dynamics of femtosecond frequency combs from mode-locked Ti-sapphire lasers. IEEE J Select Topics Quantum Electron 9:1018
122. Kutz JN (1998) Modelocking pulse dynamics in fiber lasers. In: SPIE Conference on Physics and Simulation of Optoelectronics Devices VI, vol 3283, (San Jose, CA). SPIE, pp 639–651
123. Washburn BR, Swan WC, Newberry NR (2005) Response dynamics of the frequency comb output from a femtosecond fiber laser. Opt Lett 13:10622–10633
124. Cundiff ST (2005) Soliton dynamics in mode-locked lasers. Lecture notes in physics, vol 661. Springer, Berlin, pp 183–206

Non-linear Dynamics, Symmetry and Perturbation Theory in

Giuseppe Gaeta
Dipartimento di Matematica, Università di Milano, Milan, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Symmetry of Dynamical Systems
Perturbation Theory: Normal Forms
Perturbative Determination of Symmetries
Symmetry Characterization of Normal Forms
Symmetries and Transformation to Normal Form
Generalizations
Symmetry for Systems in Normal Form
Linearization of a Dynamical System
Further Normalization and Symmetry
Symmetry Reduction of Symmetric Normal Forms
Conclusions
Future Developments
Additional Notes
Bibliography

Glossary

Perturbation theory A theory aiming at studying solutions of a differential equation (or system thereof), possibly depending on external parameters, near a known solution and/or for values of external parameters near to those for which solutions are known.

Dynamical system A system of first order differential equations $\mathrm{d}x^i/\mathrm{d}t = f^i(x, t)$, where $x \in M$, $t \in \mathbf{R}$. The space M is the phase space for the dynamical system, and $\widetilde{M} = M \times \mathbf{R}$ is the extended phase space. When f is smooth we say the dynamical system is smooth, and for f independent of t, we speak of an autonomous dynamical system.

Symmetry An invertible transformation of $\widetilde{M}$ mapping solutions into solutions. If the dynamical system is smooth, smoothness will also be required on symmetry transformations; if it is autonomous, it will be natural to consider transformations of M rather than of $\widetilde{M}$.

Symmetry reduction A method to reduce the equations under study to simpler ones (e. g. with less dependent variables, or of lower degree) by exploiting their symmetry properties.

Normal form A convenient form to which the system of differential equations under study can be brought by

means of a sequence of change of coordinates. The latter are in general well defined only in a subset of M, possibly near a known solution for the differential equations.

Further normalization A procedure to further simplify the normal form for a dynamical system, in general making use of certain degeneracies in the equations to be solved in the course of the normalization procedure.

Definition of the Subject

Given a differential equation or system of differential equations Δ with independent variables $\xi^a \in \Xi \subseteq \mathbf{R}^q$ and dependent variables $x^i \in M \subseteq \mathbf{R}^p$, a symmetry of Δ is an invertible transformation of the extended phase space $\widetilde{M} = \Xi \times M$ into itself which maps solutions of Δ into (generally, different) solutions of Δ.

The presence of symmetries is a non-generic feature; correspondingly, equations with symmetry have some special features. These can be used to obtain information about the equation and its solutions, and sometimes allow one to obtain explicit solutions.

The same applies when we consider a perturbative approach to the equations: taking into account the presence of symmetries guarantees the perturbative expansion has certain specific features (e. g. some terms are not allowed) and hence allows one to deal with simplified expansions and equations; thus this approach can be of great help in providing explicit solutions.

As mentioned above, symmetry is a non-generic feature: if we take a "generic" equation or system, it will not have any symmetry property. What makes the symmetry approach useful and widely applicable is a remarkable fact: many of the equations encountered in applications, and especially in physical and related ones (mechanical, electronic, etc.) *are* symmetric; this in turn descends from the fact that the fundamental equations of physics have a high degree of symmetry.

Thus, symmetry-based methods are at the same time "non-generic" in a mathematical sense, and "general" in a physical, or more generally real-world, sense.

Introduction

Symmetry has been a major ingredient in the development of quantum perturbation theory, and is a fundamental ingredient of the theory of integrable (Hamiltonian and non-Hamiltonian) systems; yet, the use of symmetry in the context of general perturbation theory is rather recent.

From the point of view of nonlinear dynamics, the use of symmetry has become widespread only through equivariant bifurcation theory; even in this case, attention has been mostly confined to linear symmetries.

Also, in recent years the theory and practice of symmetry methods for differential equations became increasingly popular and has been applied to a variety of problems (to a large extent, following the appearance of the book by Olver [151]). This theory is deeply geometrical and deals with symmetries of general nature (provided that they are described by smooth vector fields), i. e. in this context there is no reason to limit attention to linear symmetries.

In this article we look at the basic tools of perturbation theory, i. e. *normal forms* (first introduced by Poincaré more than a century ago for general dynamical systems; the Hamiltonian case being studied in its special features by Birkhoff several decades ago) and study their interaction with symmetries, with no limitation to linear ones. See the articles ► Normal Forms in Perturbation Theory, ► Hamiltonian Perturbation Theory (and Transition to Chaos) for an introduction to Normal Forms.

We focus on the most basic setting, i. e. systems having a fixed point (at the origin) and perturbative expansions around this; thus our theory is entirely local. We also limit to the discussion of general vector fields, i. e. we will not discuss the formulation one would obtain for the special case of Hamiltonian vector fields ► Hamiltonian Perturbation Theory (and Transition to Chaos), [111] (in which case one can deal with the Hamiltonian function rather than with the vector field it generates), referring the reader to [51] for this as well as for other extensions and for several proofs.

We start by recalling basic notions about the symmetry of differential equations, and in particular of dynamical systems; we will then discuss normal forms in the presence of symmetries, and the problem of taking into normal form the dynamical vector field and the symmetry vector field(s) at the same time.

The presence of symmetry causes several peculiar phenomena in the dynamics, and hence also in perturbative expansions. This has been explained in very effective terms by Ian Stewart [175]:

> *Symmetries abound in nature, in technology, and – especially – in the simplified mathematical models we study so assiduously. Symmetries complicate things and simplify them. They complicate them by introducing exceptional types of behavior, increasing the number of variables involved, and making vanish things that usually do not vanish. They simplify them by introducing exceptional types of behavior, increasing the number of variables involved, and making*

vanish things that usually do not vanish. They violate all the hypotheses of our favorite theorems, yet lead to natural generalizations of those theorems. It is now standard to study the "generic" behavior of dynamical systems. Symmetry is not generic. The answer is to work within the world of symmetric systems and to examine a suitably restricted idea of genericity.

Here we deal with dynamical systems, and more specially autonomous ones, i. e. systems of equations of the form $\mathrm{d}x^i/\mathrm{d}t = f^i(x)$. Now we have a single independent variable, the time $t \in \mathbf{R}$, and in view of its distinguished role we will mainly focus attention on transformations leaving it unchanged.

It is appropriate to point out here connections to several topics which we will *not* illustrate in this article.

First of all, we stress that we will work at the formal level, i. e. without considering the problem of *convergence* of the power series entering in the theory. This convergence is studied in the articles ► Perturbation Theory, ► Perturbative Expansions, Convergence of, to which the interested reader is referred in the first instance.

As hinted above, perturbation theory for symmetric systems has many points of contact with the topic of *Equivariant Bifurcation Theory*, which we will not touch upon here. The interested reader is referred to [104,107,118,160,179] for Bifurcation Theory in general, and then for the equivariant setting to the books [42,105,118]. More compact introductions are provided by the review papers [55,79].

Many facets of the interplay of symmetry and perturbation theory are also discussed in the SPT conference proceedings volumes [1,14,17,57,96,98].

Our discussion is based on the treatment in [51], with integrations and updates where appropriate.

Some considerations and remarks are given in additional notes collected in the last section; these are called for by marks$^{(xx)}$ with *xx* consecutive numbers.

Symmetry of Dynamical Systems

Symmetry of differential equations – and its use to solve or reduce the differential equations themselves – is a classical and venerable subject, being the very motivation to Sophus Lie when he created what is nowadays known as the theory of Lie groups [121]. The subject is now dealt with in a number of textbooks (see e. g. [3,19,26,27,37,80,115,125,151,152,174]) and review papers (see e. g. [116,184,185,191,192]); we will thus refer the reader to these for the general theory, and briefly recall here the special formulation one obtains when dealing with symmetries of smooth dynamical systems in $\mathbf{R}^n$.

Consider a (possibly non-autonomous) system

$$\dot{x}^i = f^i(x;t) \quad i = 1,\dots,n\,; \tag{1}$$

we assume $x \in M = \mathbf{R}^n$; M is also called the *phase space*, and $\widetilde{M} = M \times \mathbf{R}$ (the second factor representing of course time t) is the *extended phase space*.(1)

We consider now vector fields in $\widetilde{M}$; these can be written in coordinates as

$$S = \tau(x,t)\,\frac{\partial}{\partial t} + \sum_{i=1}^{n} \varphi^i(x,t)\frac{\partial}{\partial x^i}\,. \tag{2}$$

Note that (1) is identified with the vector field

$$X_f := \sum_{i=1}^{n} f^i(x,t)\,\frac{\partial}{\partial x^i}\,. \tag{3}$$

A (vector) function $x\colon \mathbf{R} \to M$ is naturally identified with the subset σ_x of $\widetilde{M}$ (corresponding to its graph) defined by

$$\sigma_x = \{(y,t) \in M \times \mathbf{R}\colon y^i = x^i(t)\} \subset \widetilde{M}\,. \tag{4}$$

The vector field S acts infinitesimally in $\widetilde{M}$ by mapping points (y,t) to points $(\widehat{y},\widehat{t})$ given by

$$\widehat{t} = t + \varepsilon\tau(y,t)\,, \quad \widehat{y}^i = y^i + \varepsilon\varphi^i(y,t)\,; \tag{5}$$

as ε is small these relations can be inverted, yielding at first order in ε

$$t = \widehat{t} - \varepsilon\tau(\widehat{y},\widehat{t})\,, \quad y^i = \widehat{y}^i - \varepsilon\varphi^i(\widehat{y},\widehat{t})\,. \tag{6}$$

Using these relations, it is easy to check that the subset $\sigma = \sigma_x$ is mapped by S to a (generally) different subset $\widehat{\sigma}$, corresponding to $y = \widehat{x}(t)$, with

$$\widehat{x}^i(t) = x^i(t) + \varepsilon\left[\varphi^i(x(t),t) - \dot{x}^i(t)\,\tau(x(t),t)\right]\,. \tag{7}$$

We say that S is a **symmetry** for the dynamical system (1) if it maps solutions into (generally, different) solutions. The condition for this to happen turns out to be [51]

$$\frac{\partial\varphi^i}{\partial t} - \frac{\partial\tau}{\partial t}\,f^i = \varphi^j\,\frac{\partial f^i}{\partial x^j} - f^j\,\frac{\partial\varphi^i}{\partial x^j}\,. \tag{8}$$

This can be more compactly expressed by introducing the Lie–Poisson bracket

$$\{f,g\} := (f\cdot\nabla)\,g - (g\cdot\nabla)\,f \tag{9}$$

between vector functions on $\widetilde{M}$. Then (8) reads

$$(\partial\varphi/\partial t) - (\partial\tau/\partial t)\,f = \{\varphi, f\}\,. \tag{10}$$

In the following we will consider *autonomous* dynamical systems; in this case it is rather natural to consider only transformations which leave t invariant, i. e. with $\tau = 0$. In this case (10) reduces to

$$(\partial\varphi/\partial t) + \{f, \varphi\} = 0 \,. \tag{11}$$

A further reduction is obtained if we only consider transformations for which the action on M is also independent of time, so that $\partial\varphi/\partial t = 0$ and the symmetry condition is

$$\{f, \varphi\} = 0 \,; \tag{12}$$

in this case one speaks of *Lie-point time-independent (LPTI) symmetries*.

The Eqs. (8) (or its reductions) will be referred to as the *determining equations* for the symmetries of the dynamical system (1).[(2)]

It should be stressed that (8) are linear in φ and τ; it is thus obvious that the solutions will span a linear space. It is also easy to check (the proof of this fact follows from the bilinearity of (9) and the Jacobi identity) that if S_1 and S_2 are solutions to (8), so is their Lie–Poisson bracket $\{S_1, S_2\}$. The set $\mathcal{G}_{X_f}$ of vector fields X_φ with φ solutions to (8) is thus a Lie algebra; it is the *symmetry algebra* for the dynamical system (1).

The symmetry algebra of a dynamical system is infinite dimensional, but has moreover an additional structure. That is, it is a *module* over the algebra $\mathcal{I}_{X_f}$ of first integrals for f (that is, scalar functions $\alpha\colon M \to \mathbf{R}$ such that $X_f(\alpha) \equiv (f \cdot \nabla)\alpha = 0$). Albeit $\mathcal{G}_{X_f}$ is infinite dimensional as a Lie algebra, it is not so as a Lie module. We have, indeed [186]:

Theorem 1 (Walcher) *The set $\mathcal{G}_{X_f}$ is a finitely generated module over $\mathcal{I}_{X_f}$.*

Perturbation Theory: Normal Forms

In this section we recall some basic facts about perturbation theory for general dynamical systems, referring to ▶ Normal Forms in Perturbation Theory, ▶ Hamiltonian Perturbation Theory (and Transition to Chaos), ▶ Perturbation Theory for details. For the sake of simplicity, we discuss perturbations around an equilibrium point; see e. g. [6,7,8,111,160,164,165] for more general settings.

As is well known – and discussed, e. g. in ▶ Normal Forms in Perturbation Theory, ▶ Hamiltonian Perturbation Theory (and Transition to Chaos) – a central objective of perturbation theory is to set (1) in normal form, i. e. to eliminate as many nonlinear terms as possible, so that the difference with respect to the linearized equation is as small as possible (see again ▶ Normal Forms in Perturbation Theory, ▶ Hamiltonian Perturbation Theory (and Transition to Chaos), or ▶ Perturbation Theory, for a precise meaning to this statement; a lengthier discussion is given e. g. in [5,7,99,104,179]).[(3)]

We briefly recall how this goes, also in order to fix notation. We consider a C^∞ dynamical system $\dot{x} = f(x)$ in $\mathbf{R}^n$, admitting $x = 0$ as an equilibrium point – that is with $f(0) = 0$. By Taylor-expanding $f(x)$ around $x = 0$ we will write this in the form

$$\dot{x} = f(x) = \sum_{k=0}^{\infty} F_k(x) \tag{13}$$

where $F_k(x)$ is homogeneous of degree $(k+1)$ in x (this seemingly odd notation will come out handy in the following). We denote the linear space of vector function $f\colon \mathbf{R}^n \to \mathbf{R}^n$ homogeneous of degree $(k+1)$ by $\mathcal{V}_k$.

Poincaré–Dulac Normal Forms

Let us consider a change of coordinates of the form

$$\xi^i = x^i + h^i_k(x) \,, \tag{14}$$

where $h_k \in \mathcal{V}_k (k \geq 1)$; we write $\Gamma^i_j = (\partial h^i/\partial x^j)$. A change of coordinates of the form (14) is called a *Poincaré transformation* (or P transformation for short), and the function h_k is also called the *generator* of the P transformation. In the following we will freely drop the subscript k when this cannot generate any confusion.

The transformation (14) is, for small x, a near-identity transformation; thus it is surely invertible in a small enough neighborhood of the origin. We apply $\Lambda := (I + \Gamma)^{-1}$ on (13), and get the P transformed dynamical system in the form

$$\dot{x} = \widetilde{f}(x) \equiv \Lambda \sum_{m=0}^{\infty} F_m\left(x + h_k(x)\right) = \sum_{m=0}^{\infty} \widetilde{F}_m(x) \,. \tag{15}$$

In order to identify the $\widetilde{F}_m$ we should consider power series expansions for Λ and for $F_m(x + h(x))$. With standard computations (we refer again to ▶ Normal Forms in Perturbation Theory, or to [7,51], for details), we obtain that the $\widetilde{F}_m$ are given (with $[q]$ the integer part of q) by

$$\widetilde{F}_m = F_m + \sum_{p=1}^{[m/k]} \left[\sum_{s=0}^{p} (-1)^s \Gamma^s \Phi^{p-s}_{h_k} \right] F_{m-kp} \,. \tag{16}$$

The Φ^r_h appearing in (16) are defined as follows. With a multi-index notation, write $J = (j_1, \ldots, j_n)$, $|J| = \sum_i j_i$; set then $\partial_J := \partial_1^{j_1} \ldots \partial_n^{j_n}$, and similarly $h^J_k :=$

$(h_k^1)^{j_1} \dots (h_k^n)^{j_n}$. The operators Φ_h^r (representing all the partial derivatives of order $|J|$) are defined as $\Phi_h^r = (1/r!) \sum_{|J|=r} (h^J \cdot \partial_J)$.

Some special cases following from this general formula should be noted. As well known, the terms of degree smaller than k are not changed at all, i. e. $\widetilde{F}_m = F_m$ for $m < k$, and the term of degree k is changed (writing $h \equiv h_k$) according to

$$\widetilde{F}_k = F_k + [\Phi_h - \Gamma] F_0 \,. \tag{17}$$

(Similarly, for $0 < \nu < k$, the term of degree $k + \nu$ is changed into $\widetilde{F}_{k+\nu} = F_{k+\nu} + [\Phi_h - \Gamma] F_\nu$.) Define now, recalling (9), the operators

$$L_k = \{F_k, .\} \,; \tag{18}$$

note $L_k \colon \mathcal{V}_m \to \mathcal{V}_{m+k}$.

The operator $\mathcal{A} = L_0$, associated with the linear part $A = (Df)(0)$ of f (that is, $F_0(x) = Ax^{(4)}$) is called the *homological operator*; it leaves the spaces $\mathcal{V}_k$ invariant, and hence it admits the decomposition $\mathcal{A} = \bigoplus_{m=0}^{\infty} \mathcal{A}^{(m)}$, where $\mathcal{A}^{(m)}$ is just the restriction of $\mathcal{A}$ to $\mathcal{V}_m$.

In the following, we will need to consider the adjoint $\mathcal{A}^+$ of the operator $\mathcal{A}$. For this we need to introduce a scalar product in the space $\mathcal{V} = \bigcup \mathcal{V}_k$. Actually, we can introduce a scalar product in each of the spaces $\mathcal{V}_k$ into which $\mathcal{V}$ decomposes.

A convenient scalar product was introduced in [60] (following [18]); we will only use this one (the reader should be warned that different definitions are also considered in the literature [5,51]). We denote by $\xi_{\mu,i}$ the vector function whose components are all zero but the ith one, given by $x^\mu := x_1^{\mu_1} \dots x_n^{\mu_n}$; with this notation, we define[(5)]

$$\left(\xi_{\mu,i}, \xi_{\nu,j}\right) = \sum_{i=1}^{n} \langle \xi_{\mu,i}, \xi_{\nu,i} \rangle \,, \tag{19}$$

where $\langle x^\mu, x^\nu \rangle \equiv \partial^\mu x^\nu$, and $\partial^\mu = \partial_1^{\mu_1} \dots \partial_n^{\mu_n}$. (When in the following we consider adjoint operators, these will be understood in terms of this.)

With this scalar product, one has the following lemma (a proof is given, e. g. in [118]).

Lemma 1 *If $\mathcal{A}$ is the homological operator associated with the matrix A, $\mathcal{A} = \{Ax, .\}$, then its adjoint $\mathcal{A}^+$ is the homological operator associated with the adjoint matrix A^+, i. e., $\mathcal{A}^+ = \{A^+ x, .\}$.*

We will also consider the projection onto the range of $\mathcal{A}^{(k)}$, denoted by π_k. The *general homological equations* (the one for $k = 0$ corresponds to the standard homological equation) are then

$$\mathcal{A}^{(k)}(h_k) = \pi_k F_k \,. \tag{20}$$

These are equations for $h_k \in \mathcal{V}_k$, and always admit a solution (thanks to the presence of the projection operator π_k). The Eq. (20) maps into a set of algebraic equations once we introduce a basis in $\mathcal{V}_k$.

The $h_k \in \mathcal{V}_k$ solving to (20) will be of the form $h_k = h_k^* + \ell_k$, where $h_k^* = \mathcal{A}^*(\pi_k F_k) \in \mathrm{Ran}[(\mathcal{A}^{(k)})^*]$ (here $\mathcal{A}^*$ is the pseudo-inverse to $\mathcal{A}$; note $\mathrm{Ker}(\mathcal{A}^*) = \mathrm{Ker}(\mathcal{A}^+)$) is unique, and ℓ_k is any function in $\mathrm{Ker}[\mathcal{A}^{(k)}]$.

Remark 1 It should be stressed that, while adding a nonzero ℓ_k to h_k^* does not change the resulting $\widetilde{F}_k$, it could – and in general will – affect the terms of higher order.

One can then normalize X_f in the standard recursive way, based on solving homological equations; this is described, e. g. in ► Normal Forms in Perturbation Theory. In this way, we are reduced to considering only systems with

$$F_k \in \left[\mathrm{Ran}\left(\mathcal{A}^{(k)}\right)\right]^\perp = \mathrm{Ker}\left[\left(\mathcal{A}^{(k)}\right)^+\right] . \tag{21}$$

Such terms are also called *resonant*.[(6)]

The presence of resonant terms is related to the existence of *resonance relations* among eigenvalues λ_i of the matrix A describing the linear part of the system; these are relations of the form $(m \cdot \lambda) \equiv \sum_i m_i \lambda_i = \lambda_s$ where the m_i are non-negative integers, with $|m| = \sum m_i > 1$ (the restriction $|m| > 1$ is to avoid trivial cases); the integer $|m|$ is also called the *order* of the resonance (► Normal Forms in Perturbation Theory, [7,8]).

If the system $\dot{x} = f(x)$ has only resonant nonlinear terms, we say that it is in *Poincaré–Dulac normal form* (► Normal Forms in Perturbation Theory, [7,8,35,44]). If all the nonlinear terms of order up to q are resonant, we say that the system is in normal form up to order q.

Theorem 2 (Poincaré–Dulac) *Any analytic vector field X_f with $f(0) = 0$ can be formally taken to normal form by a sequence of Poincaré transformations.*

Remark 2 If we do not have an exact resonance, but $(m \cdot \lambda) - \lambda_s \simeq 0$, we have a *small denominator*, and correspondingly a very large coefficient in h_k, where $k = |m|$. Such small denominators are responsible for divergencies in the normalizing series [7,8,9,32,33,35,172].

Lie Transforms

In discussing Poincaré normal forms, we have considered near-identity diffeomorphisms of M; these can be expressed as time-one maps for the flow under some vector fields X_h. In a number of cases, it is more convenient to deal directly with such vector fields. We are going to briefly discuss this approach, and its relation with the one discussed above; for further detail, see e. g. (► Normal Forms in Perturbation Theory, [25,28,29,30,51,61,84,101,145]).

Let the vector field $H \equiv X_h$ be given, in the x coordinates, by $H = h^i(x)(\partial/\partial x^i)$. We denote by $\Psi(s;x)$ the local flow under H starting at x, so that $(\mathrm{d}/\mathrm{d}s)\Psi(s;x) = H(\Psi(s;x))$. We also use exponential notation: $\Psi(s;x) = \mathrm{e}^{sH}x$.

We will denote $\Psi(1;\xi)$ as x; the direct and inverse changes of coordinates will be defined as

$$\begin{aligned} x &= \Psi(1;\xi) = \left[\mathrm{e}^{sH}\xi\right]_{s=1} , \\ \xi &= \Psi(-1;x) = \left[\mathrm{e}^{-sH}x\right]_{s=1} . \end{aligned} \tag{22}$$

Now, consider another vector field X on M, describing the dynamical system we are interested in. If we study the dynamical system $\dot{\xi}^i = f^i(\xi)$, we consider the vector field given in the x coordinates by

$$X = X_f = f^i(x)(\partial/\partial x^i) . \tag{23}$$

This also generates a (local) flow, i. e. for any $\xi_0 \in M$ we have a one-parameter family $\xi(t) \equiv \Gamma(t;\xi_0) \in M$ such that $(\mathrm{d}\xi(t)/\mathrm{d}t) = X(\xi)$. By means of (22), this also defines a one-parameter family $x(t) \in M$, which will satisfy $(\mathrm{d}x(t)/\mathrm{d}t) = \widetilde{X}(x)$ for some vector field $\widetilde{X}$ on M; this will be the transformed vector field under (22), and is given by[(7)]

$$\widetilde{X} = \left[\mathrm{e}^{sH}X\mathrm{e}^{-sH}\right]_{(s=1)} . \tag{24}$$

We call this transformation the **Lie-Poincaré transformation** generated by h. Notice that this yields, up to order one in s, (and therefore if $h \in \mathcal{V}_k$, up to terms in $\mathcal{V}_k$), just the same result as the Poincaré transformation with the same generator h.

The $\widetilde{X}$ can be given (for arbitrary s), in terms of the Baker–Campbell–Haussdorf formula, as

$$\widetilde{X} = \sum_{k=0}^{\infty} \frac{(-1)^k s^k}{k!} X^{(k)} , \tag{25}$$

where $X^{(0)} = X$, and $X^{(k+1)}$ are defined recursively by $X^{(k+1)} = [X^{(k)}, H]$.

Perturbative Determination of Symmetries

Let us now consider the problem of determining the symmetries of a given dynamical system. Writing the latter in the form

$$\dot{x} = f(x) = \sum_{k=0}^{\infty} F_k(x) \quad (F_k \in \mathcal{V}_k) , \tag{26}$$

it is quite natural to also look for symmetries in terms of a (possibly only formal) power series; this will be our approach here.

Consider a generic smooth vector field (by smooth we mean, here and elsewhere, C^∞; however, we will soon go on to actually consider vector fields represented by power series) in $M \times \mathbf{R}$,

$$Y = \tau(x,t)\,\partial_t + \sum_{i=1}^{n} \varphi^i(x,t)\partial_i . \tag{27}$$

Here and below, $\partial_i \equiv (\partial/\partial x^i)$. As remarked in Sect. "Symmetry of Dynamical Systems", (27) is too general for our purposes; we are primarily interested in vector fields acting on M alone, and mainly [48,49,51,80] in time-independent vector fields on M (note in this way the dynamical and the symmetry vector fields are on the same footing). Thus we will just consider (we will consistently use X for the vector field X_f defined by (26), and Y for the symmetry vector field X_s, in order to simplify the notation)

$$Y = \sum_{i=1}^{n} s^i(x)\,\partial_i . \tag{28}$$

Note that at this stage we are **not** assuming the dynamical system described by X has been taken into normal form; we will see later on the specific features of this case.

Determining Equations

Now, let us consider the dynamical system identified by X, and look for the determining equation identifying its symmetries Y as in (28). Condition (8) yields

$$f^j(x)\frac{\partial s^i(x)}{\partial x^j} - s^j(x)\frac{\partial f^i(x)}{\partial x^j} \equiv (f\cdot\nabla)s - (s\cdot\nabla)f = 0 . \tag{29}$$

As discussed above, using (9) this is also written as $\{f,s\} = 0$, which just means (as had to be expected)

$$[X,Y] = 0 . \tag{30}$$

We now denote the set of Y satisfying (30) by $\mathcal{G}_X$. It is obvious that $\mathcal{G}_X$, equipped with the usual commutator of

vector fields, is a Lie algebra. It is also easy to see that, as $f(0) = 0$, $Y \in \mathcal{G}_X$ implies $s(0) = 0$.

We can now expand Y, i. e. $s(x)$, in a perturbative series around $x_0 = 0$ in the same way as we did for X. We write

$$s(x) = \sum_{k=0}^{\infty} S_k(x) \quad (S_k \in \mathcal{V}_k) \ . \tag{31}$$

Plugging this into the determining Eqs. (30) we get, after rearranging the terms,

$$\sum_{k=0}^{\infty} \sum_{m=0}^{k} \{F_m, S_{k-m}\} = 0 \ . \tag{32}$$

For this to hold, the different homogeneous terms of degree k must vanish separately. Thus, we have a hierarchy (in a sense to be explained in a moment) of equations

$$\sum_{m=0}^{k} \{F_m, S_{k-m}\} = 0 \qquad k = 0, 1, 2, 3, \ldots \tag{33}$$

It is convenient to isolate the terms containing linear factors, i. e. to rewrite (33) – for $k \geq 1$ – in the form

$$\{F_0, S_k\} - \{S_0, F_k\} = -\sum_{m=1}^{k-1} \{F_m, S_{k-m}\} \equiv \chi_k \ , \tag{34}$$

where we have used the antisymmetry of $\{.,.\}$, and $\chi_0 = \chi_1 = 0$.

Recursive Solution of the Determining Equations

Let us now consider the problem of concretely solving the determining Eq. (32). As the perturbative series expansion suggests, we can proceed order by order, i. e. start with consideration of the equation for $k = 0$, then tackle $k = 1$, and so on.

Proceeding in this way we are always reduced to consider equations of the form

$$F_0^j \left(\frac{\partial S_k^i}{\partial x^j} \right) = S_k^j \left(\frac{\partial F_0^i}{\partial x^j} \right) + \Psi_k^j(x) \tag{35}$$

with the Ψ_k^j known functions of x (as they depend on the known F_k and on the S_j with $j < k$, determined at previous stages).

Notice also that F_0 is just the (known) linear part of X. If we write it in matrix form as $F_0^i(x) = A_{ij}x^j$ (we will write similarly $S_0^i(x) = B_{ij}x^j$ for S_0), then (35) reads simply

$$\left[A_{ij}x^j\right] \left(\frac{\partial S_k^i}{\partial x^j} \right) = A_{ij}S_k^j + \Psi_k^j(x) \ . \tag{36}$$

Solving the determining equations in such a recursive way only requires one to solve at each stage a system of (inhomogeneous) linear PDEs for the S_k.

For further reference, we introduce the notation X_M for the linear vector field associated with the matrix M, i. e. $X_M = (M_{ij}x^j)(\partial/\partial x^i)$. We also write L_M for the homological operator associated with a matrix M, $L_M(.) = \{Mx, .\}$ (see also note 4). In this notation, for any two matrices A, B we have $[X_A, X_B] = -X_{[A,B]}$; similarly we have, as a consequence of the Jacobi identity:

Lemma 2 *For any matrices A, B, the commutator of the associated homological operators is given by $[L_A, L_B] = L_{[B,A]} = -L_{[A,B]}$.*

Let us follow explicitly the iterative procedure for solving (32), (33) for the first steps. For $k = 0$, we require that $\{F_0, S_0\} = 0$. With our matrix notation,

$$\{F_0, S_0\} \equiv \{Ax, Bx\} = -[A, B]_{ij}\, x^j \tag{37}$$

and therefore at this stage we only have to determine the matrices B commuting with a given matrix A.

For $k = 1$, we just get $\{F_0, S_1\} + \{F_1, S_0\} = 0$; in the matrix notation this just reads $\{Ax, S_1\} = \{Bx, F_1\}$ or, using the homological operators notation,

$$\mathcal{A}(S_1) = \mathcal{B}(F_1) := \Psi_1(x) \ . \tag{38}$$

For $k = 2$ we get in the same way

$$\mathcal{A}(S_2) = \mathcal{B}(F_2) - \{F_1, S_1\} := \Psi_2(x) \ , \tag{39}$$

and so on for any k.

Remark 3 The fact that we can proceed recursively (and only deal with linear PDEs) does not mean that we are guaranteed to find solutions at any stage k. At $k = 0$ we always have at least the solutions given by $B = I$ and by $B = A^q (q = 1, 2, \ldots)$. For $k \geq 1$ we do not, in general, have solutions to the determining equations apart from $S_j = cF_j$ for all $q = 0, \ldots, k$ (i. e., as k is generic, $Y = cX$). This corresponds to the fact that symmetry is not generic.

Remark 4 As the relevant equations are not homogeneous, $S_k = 0$ is not, in general, an acceptable solution at stage k. This is quite natural if one thinks that the choice $B = I$ is always acceptable at $k = 0$. Choosing $S_k = 0$ at all the following stages would leave us with the dilation vector field $Y = x^i \partial_i$, which is a symmetry only for X linear [80,151,174].

Approximate Symmetries

Note that it could happen where we are only able to determine a commuting vector field for X up to some finite order k (either for a full symmetry does not exist or for our limited capacities, computational or otherwise).

If in this case we consider a neighborhood of the origin of small size, say a ball $\mathcal{B}_\varepsilon$ of radius $\varepsilon \ll 1$, in this we have $[X, Y] = O(\varepsilon^k)$; thus, Y represents an *approximate symmetry* for X.

Approximate symmetries are interesting and useful in a number of contexts. In particular, in some cases – notably, for Hamiltonian vector fields – there is a connection between symmetries of dynamical systems and conserved quantities (i. e. constants of motion) for it; in this case, approximate symmetries will correspond to approximate constants of motion, i. e. to quantities which are not exactly conserved, but are approximately so.

More precisely, an approximate symmetry will correspond to a quantity J whose evolution under the dynamics described by X is slow of order k, i. e. $\mathrm{d}J/\mathrm{d}t \approx \varepsilon^k$ for some finite k. It is rather clear that these can be quite useful in applications, where we are often concerned with study of the dynamics over finite times; see [51] and especially, in the Hamiltonian case, [100].

Symmetry Characterization of Normal Forms

Let us now consider the case where the dynamical system (13) has already been taken into Poincaré–Dulac normal form. We start by recalling some notions of linear algebra of use here.

Linear Algebra

A real matrix T is *semisimple* if its complexification T_C can be diagonalized, and is *normal* if it commutes with its adjoint, $[T, T^+] = 0$. A diagonal matrix is normal. For a normal matrix, $T\colon \mathrm{Ker}(T^+) \to \mathrm{Ker}(T^+)$. If T is normal we actually have $\mathrm{Ker}(T) = \mathrm{Ker}(T^+)$.

Any semisimple matrix can be transformed into a normal one by a linear transformation. If two semisimple matrices A, B commute, then they can be simultaneously diagonalized (by a linear, in general non-orthogonal, transformation), and so taken simultaneously to be normal. Thus, when considering such a pair of matrices, we can with no loss of generality assume them to be diagonal or, a fortiori, normal.

If we want to transform T into a real normal matrix, we just have to consider the transformation of T into a block diagonal matrix, the blocks corresponding to (complex conjugate) eigenvalues. It is easy to see that in this way we still get a (real) normal matrix.[8]

In the following, we will at several points restrict, for ease of discussion, to normal matrices; our statements for normal matrices will be easily extended to semisimple ones up to the appropriate linear transformation.

Normal Forms

We now note that $F \in \mathrm{Ker}(\mathcal{A})$ means that the vector field associated with F (which we denote by X_F) commutes with the linear vector field X_A associated with A. That is,

$$F \in \mathrm{Ker}(\mathcal{A}) \Leftrightarrow \{Ax, F(x)\} = 0 \Leftrightarrow [X_A, X_F] = 0\,. \quad (40)$$

Thus we have the following characterization for vector fields in normal form (note this uses Lemma 1 and hence the scalar product defined in Sect. "Perturbation Theory: Normal Forms").

Lemma 3 *A vector field $X = (A_{ij}x^j + F^i(x))\partial_i$, where the F are nonlinear functions, is in normal form if and only if its nonlinear part $X_F = F^i(x)\partial_i$ commutes with the vector field $X_{A^+} = [A^+_{ij}x^j]\partial_i$ associated with the adjoint of its linear part, i. e. if and only if $F \in \mathrm{Ker}(\mathcal{A}^+)$.*

For the sake of simplicity, we will only consider the case where the matrix A, corresponding to the linear part of X (in both the original and the normal form coordinates), commutes with its adjoint, i. e. we make the following[9] **assumption:** *The matrix A is normal:* $[A, A^+] = 0$.

If A is normal, then it follows $[L_A, L_{A^+}] = 0$ due to Lemma 2; this implies in particular that:

Lemma 4 *If A is normal, then* $\mathrm{Ker}(\mathcal{A}) = \mathrm{Ker}(\mathcal{A}^+)$.

It is important to recall that with the standard scalar product, we have $(L_A)^+ = L_{A^+}$. It is also important, although trivial, to note that if A is normal, then $\mathcal{A}$ is a normal operator (under the standard scalar product), and $\mathrm{Ker}(\mathcal{A}) \cap \mathrm{Ran}(\mathcal{A}) = \{0\}$, but $f \in \mathrm{Ker}(\mathcal{A}^2) \Rightarrow (\mathcal{A}(f)) \in \mathrm{Ker}(\mathcal{A})$ and therefore $\mathcal{A}(f) = 0$. Hence

Lemma 5 *If A is normal,* $\mathrm{Ker}(L_A^2) = \mathrm{Ker}(L_A)$.

This discussion leads to a natural characterization of Poincaré–Dulac normal forms in terms of symmetry properties.[10]

Lemma 6 *If A is a normal matrix, then $X = (Ax + F)^i(\partial/\partial x^i)$ is in normal form if and only if X_A is a symmetry of X, i. e. $F \in \mathrm{Ker}(\mathcal{A})$.*

The General Case

In the general case, i. e. when A is not normal, the resonant terms will be those in $\mathrm{K}er(\mathcal{A}^+)$; similarly, in this case we could characterize systems in normal form by $\{F(x), A^+x\} = 0$.

However, for a symmetry characterization it is better to proceed in a slightly different way. That is, we recall that any matrix A can be uniquely decomposed as $A = A_\mathrm{s} + A_\mathrm{n}$ where A_s is semisimple and A_n is nilpotent, with $[A_\mathrm{s}, A_\mathrm{n}] = 0$. Resonance properties involve eigenvalues of A_s, and resonant terms will satisfy $\{F(x), A_\mathrm{s}^+x\} = 0$.

This is a more convenient characterization, in that it shows that the full vector field (in normal form) X will commute with the linear vector field $S = (A_\mathrm{s})^i_j x^j \partial_i$ corresponding to the semisimple part of its linear part.

Symmetries and Transformation to Normal Form

We want to consider the case where the dynamical system (1), or equivalently the vector field X, admits a symmetry Y (the case of an n-dimensional symmetry algebra will be considered later on); we want to discuss how the presence of the symmetry affects the normalization procedure. Moreover, as the dynamical and symmetry vector fields are on equal footing, it will be natural to investigate if they can be both put into normal form; or even if some kind of **joint normal form** is possible (as is the case).

We will use the notation introduced above for the expression of X, Y in the x coordinates, and denote by y the normal form coordinates. Correspondingly, the bracket $\{.,.\}$ (which is defined in coordinates) will be denoted as $\{.,.\}_{(x)}$ or $\{.,.\}_{(y)}$ when confusion is possible. We have, therefore,

$$\begin{aligned} X &= f^i(x)(\partial/\partial x^i) = g^i(y)(\partial/\partial y^i)\,, \\ Y &= s^i(x)(\partial/\partial x^i) = r^i(y)(\partial/\partial y^i)\,; \end{aligned} \tag{41}$$

and similarly for the power series expansions of f, g, s, r in terms homogeneous of degree $(k+1) = 1, 2, \ldots$. We will denote the matrices associated with the linear parts of X, Y by, respectively, A and B: $(Df)(0) = (Dg)(0) = A$, $(Ds)(0) = (Dr)(0) = B$. The corresponding homological operators will be denoted by $\mathcal{A} = L_A$ and by $\mathcal{B} = L_B$. We assume that both A and B are normal matrices.

Nonlinear Symmetries (The General Case)

The key, albeit trivial, observation is that the geometric relation $[X, Y] = 0$ does not depend on the coordinate system we are using. Therefore, if $\{f, s\}_{(x)} = 0$, we must also have $\{g, r\}_{(y)} = 0$.

Another important, and again quite trivial, observation is that when we consider a P-transformation $x = y + h_k(y)$, the term F_k of order k in X changes according to $\mathcal{A}(h_k)$, but the term S_k of order k in Y changes according to $\mathcal{B}(h_k)$. The same applies when we consider a Lie-Poincaré transformation.

Thus, although when we choose $h_k + \ell_k$ [with $\ell_k \in \mathrm{Ker}(\mathcal{A})$] to generate the Poincaré transformation, we get the same transformation as that generated by h_k on the F_k (see Remark 1), the transformation on S_k can be different. This means that the freedom left by the Poincaré prescription for construction of the normalizing transformation could, in principle, be used to take the symmetry vector field Y into some convenient form. This is indeed the case, as will be shown below.

Two vector fields X, Y on M, as in (35), with A and B semisimple, are in **Joint Normal Form** if both G_k and R_k are in $\mathrm{Ker}(\mathcal{A}) \cap \mathrm{Ker}(\mathcal{B})$ for all $k \geq 1$.

Theorem 3 *Let the vector fields $X = f^i(x)\partial_i$ and $Y = s^i(x)\partial_i$ have a fixed point in $x_0 = 0$. Let them commute, $[X, Y] = 0$, and have normal semisimple linear parts $A = (Df)(0)$ and $B = (Ds)(0)$. Then, by means of a sequence of Poincaré transformations, they can be brought to Joint Normal Form.*

In this theorem (a proof of this is given in [49,51]; see also [106] for a different approach to a related problem) we did not really use the interpretation of one of the vector fields as describing the dynamics of the system and the other describing a symmetry, but only their commutation relation. From this point of view, it is natural to also consider arbitrary (possibly non-Abelian) algebras of vector fields.

Linear Symmetries

A special case of symmetries is given by *linear* symmetries, i. e. by the case where

$$Y = X_B = \left(B_{ij}x^j\right)(\partial/\partial x^i)\,. \tag{42}$$

In this case, if X, Y are in Joint Normal Form we have in particular that $f \in \mathrm{Ker}(\mathcal{B})$. We have the following corollaries to Theorem 3:

Corollary 1 *If the linear vector field $Y = X_B$ is a symmetry for $X = f^i(x)\partial_i$, it is possible to normalize f by passing to y coordinates so that $X = g^i(y)(\partial/\partial y^i)$, $Y = (By)^i(\partial/\partial y^i)$, and $g \in \mathrm{Ker}(\mathcal{A}) \cap \mathrm{Ker}(\mathcal{B})$.*

Corollary 2 *Let $s(x) = Bx + S(x)$, with S the nonlinear part of s, and let $Y = s^i(x)\partial_i$ be a symmetry of*

$X = f^i(x)\partial_i$. Then when X, Y are put in Joint Normal Form, X_B is a linear symmetry of X.

When we perform the Poincaré transformations needed to transform X into its normal form, there seems to be no reason, a priori, why the Y should keep its linear form. It is actually possible to prove the following result (the proof of this relies on a similar result by Ruelle [159,160] dealing with the center manifold mapping, and is given e. g. in [60,118]; see also [20,21,22,23]) that we quote from [118]:

Theorem 4 *If X commutes with a linear vector field $Y = X_B = (B_{ij}x^j)\partial_i$, then it is possible to find a normalizing series of Poincaré transformations with generators $h_k \in \mathrm{Ker}(\mathcal{B})$, so that in the new coordinates y, X is taken into normal form and Y is left unchanged, i. e. $Y = (B_{ij}y^j)\partial_i$.*

Note that, for resonant B, Theorem 4 is *not* a special case of Theorem 3 and the above Corollaries.[11]

Remark 5 One should avoid confusions between linear symmetries of the dynamical system and symmetries of its linearization, which do not extend in general to symmetries of the full system.

Generalizations

In this section we are going to discuss some generalizations of the results illustrated in the previous one: we will deal with the problem of transformation into normal form of a (possibly non Abelian) Lie algebra with more than two generators, not necessarily commuting.[12]

Abelian Lie Algebra

It is actually convenient to drop the distinction between the vector field defining the dynamical system and the symmetry vector fields. Thus, we simply consider an algebra $\mathcal{G}$ of vector fields X_i.[13]

First of all, we consider the case of an *Abelian* Lie algebra of vector fields. In this case we have the following result (see [51] for a proof).

Theorem 5 *Let $\{X_1, \ldots, X_r\}$ commute, and assume that the matrices $A_{(i)}$ identifying the linear parts of the X_j are normal. Then $\{X_1, \ldots, X_r\}$ can be put in Joint Normal Form by a sequence of Poincaré or Lie-Poincaré transformations.*

Note that if $\mathcal{A}_i$ are the homological operators corresponding to the $A_{(i)}$, and we write $\mathcal{K} = \bigcap_{i=0}^{k} \mathrm{Ker}(\mathcal{A}_i)$, the Theorem states there are coordinates y^i such that $X_j = (\widetilde{f}_{(j)}(y))^i(\partial/\partial y^i)$ with $\widetilde{f}_{(j)} \in \mathcal{K}$ for all j.

Nilpotent Lie Algebra

For generic Lie algebras one cannot expect results as general as in the Abelian case [4,48,51]. A significant exception to this is met in the case of *nilpotent* algebras[14] (see also [54] for a group-theoretical approach), in which we can recover essentially the same results obtained in the Abelian case (see [112] for the case of semisimple Lie algebras).

Actually, an extension of Theorem 5 to the nilpotent case should be considered with some care, as the only nilpotent algebras of nontrivial semisimple matrices are Abelian. On the other hand, we could have a non-Abelian nilpotent algebra of vector fields with linear parts given by semisimple matrices, provided that some of these vanish.

Indeed, although $[X_i, X_j] = c_{ij}^k X_k$ necessarily implies that $[A_i, A_j] = c_{ij}^k A_k$, it could happen that all the A_k for which there exists a nonzero c_{ij}^k do vanish, so that the algebra of vector fields is "Abelian at the linear level". (As a concrete example, consider the algebra spanned by $X = -x^2 \mathrm{d}/\mathrm{d}x$ and $Y = x\mathrm{d}/\mathrm{d}x$.)

Note that in this case $\mathrm{Ker}(\mathcal{A}_k)$ would be just the whole space; needless to say, we should consider the full vector fields X_k, which will produce (by assumption) a closed Lie algebra.

With this remark (and in view of the fact that the proof of Theorem 5 is based on properties of the algebra of the A_i's and of the corresponding homological operators, see [51]) it is to be expected that the result for the nilpotent case will not substantially differ from the one holding for the Abelian case (as usual, the key to this extension will be to proceed to normalization of vector fields in an order which respects the structure of the Lie algebra). This is indeed what happens, and one has the following result [48,51]:

Theorem 6 *Let the vector fields $\{X_1, \ldots, X_r\}$ form a nilpotent Lie algebra $\mathcal{G}$ under $[.,.]$; assume that the matrices $A_{(i)}$ identifying the linear parts of the X_j are normal. Then $\{X_1, \ldots, X_r\}$ can be put in Joint Normal Form by a sequence of Poincaré or Lie-Poincaré transformations.*

Corollary 3 *If the general Lie algebra $\mathcal{G}$ of vector fields $\{X_1, \ldots, X_n\}$ contains a nilpotent subalgebra $\mathcal{G}^*$, then the set of vector fields X_i spanning this $\mathcal{G}^*$ can be put in Joint Normal Form.*

General Lie Algebra

Some "partial" Joint Normal Form can be obtained, even for non-nilpotent algebras, under some special assumptions. We will just quote a result in this direction, referring as usual to [51] for details. A description of normal

forms for systems with symmetry corresponding to simple compact Lie groups is given in [86]

Theorem 7 *Let $\mathcal{G}$ be a d-dimensional algebra spanned by X_{F_a} with $F_a = A_a x + F_a (a = 1, \ldots, d)$, and let $\mathcal{G}$ admit a non-trivial center $C(\mathcal{G})$. Let the center of $\mathcal{G}$ be spanned by X_{w_b} with $w_b = C_b x + W_b (b = 1, \ldots, d_C$, where $d_C \leq d$), and assume that the semisimple parts $C_{b,s}$ are normal matrices. Denote by $C_{b,s}$ the associated homological operators, and write $\mathcal{K}_s = \cap_{b=1}^{d_C} \mathrm{Ker}(C_{b,s})$.*

Then, by means of a sequence of Poincaré transformations, all the F_a can be taken into $\widehat{F}_a \in \mathcal{K}_s$. The same holds for $\widehat{\mathcal{G}} = \mathcal{G} \oplus \mathcal{N}$, with $\mathcal{N}$ a nilpotent subalgebra.

Symmetry for Systems in Normal Form

No definite relation exists, in general, between symmetries $\mathcal{G}_X$ of a vector field X and symmetries $\mathcal{G}_A$ of its linear part, or between constants of motion $\mathcal{I}_X$ for X and constants of motion $\mathcal{I}_A$ for its linear part, but if X is in normal form, one has some interesting results [48,186]:

Lemma 7 *If X is in normal form, any constant of motion of X must also be a constant of motion of its linearization A, i. e. $\mathcal{I}_X \subseteq \mathcal{I}_A$.*

In general $\mathcal{I}_X \neq \mathcal{I}_A$, even if X is in normal form. Also, if $[B, A] = 0$ in general $\mu(x)Bx$ does not belong to $\mathcal{G}_X$, even for $\mu \in \mathcal{I}_X$ (unless $X_B \in \mathcal{L}_X$ as well, where $\mathcal{L}_X$ are the linear symmetries of X), nor to $\mathcal{G}_A$ (unless $\mu \in \mathcal{I}_A$).

Lemma 8 *If X is in normal form, then $\mathcal{G}_X \subseteq \mathcal{G}_A$.*

This result allows for the restricting our search for $Y \in \mathcal{G}_X$ to $\mathcal{G}_A$ rather than considering the full set of vector fields on M. Similarly, Lemmas 7 and 8 can be useful in the determination of the sets $\mathcal{L}_X, \mathcal{I}_X$, in that we can first solve the easier problem of determining $\mathcal{L}_A, \mathcal{I}_A$, and then look for $\mathcal{L}_X, \mathcal{I}_X$ in the class of vector fields $\mathcal{L}_A$ and of the functions in $\mathcal{I}_A$, rather than in the full set of linear vector fields on M and, respectively, in the full set of scalar functions on M.

Moreover, $\mathcal{G}_A$ can be determined in a relatively simple way, by solving the system of quasi-linear non-homogeneous first order PDEs $\{Ax, g\} = 0$, which are written explicitly as

$$\left(A_{ij} x^j\right) (\partial g^k / \partial x^i) = A_{kj} g^j \,. \tag{43}$$

By considering this, and introducing the set $\mathcal{I}_A^*$ of the meromorphic (i. e., quotients of formal power series) constants of motion of the linear problem $\dot{x} = Ax$, one can obtain [48,60,186] the following result:

Lemma 9 *$\mathcal{G}_A$ is the set of all formal power series in $\mathcal{I}_A^* \otimes \mathcal{L}_A$.*

In a more explicit form, as $\mathcal{G}_A \equiv \mathrm{Ker}(\mathcal{A})$, the resonant terms $F(x) \in \mathrm{Ker}(\mathcal{A})$ are power series of the form $F(x) = K(\rho(x)) \cdot x$, where K is a matrix commuting with A when written in terms of its real entries K_{ij}, and where $K(\rho(x))$ is the same matrix in which the entries K_{ij} are replaced by functions of the constants of motion $\rho = \rho(x) \in \mathcal{I}_A^*$. The set of the vector fields in $\mathcal{G}_A$ is of course a Lie algebra.

We summarize our discussion for dynamical systems in normal form in the following proposition:

Theorem 8 *Let X be a vector field in normal form, and let A be the normal matrix corresponding to its linear part. Then, $\mathcal{L}_X \subseteq \mathcal{G}_X \subseteq \mathcal{G}_A$; and $\mathcal{L}_X \subseteq \mathcal{L}_A \subseteq \mathcal{G}_A$.*

Remark 6 It should be mentioned that Kodama considered the problem of determining $\mathcal{G}_A$ from a more algebraic standpoint [122]. In the same work, Kodama also observed that $\mathcal{G}_A$, considered as an algebra, is not only infinite-dimensional, but has the natural structure of a graded Virasoro algebra.

Remark 7 We emphasize once again that the above results were given for X in normal form. They can obviously be no longer true if X is not in normal form.(15)

Linearization of a Dynamical System

An interesting application of the Joint Normal Form deals with the case of *linearizable* dynamical systems. Clearly, if $\mathrm{Ker}(\mathcal{A}) = \{0\}$, the dynamical system is linearizable by means of a formal Poincaré transformation. But, whatever the matrix A, the linear vector field $X_A = (Ax \cdot \nabla)$ commutes with the vector field $S = \sum_i x_i (\partial / \partial x_i) = ((Ix) \cdot \nabla)$, which generates the dilation's in $\mathbf{R}^n$. It is easy to see that, conversely, the only vector fields commuting with S are the linear ones. It is also clear that the identity does not admit resonances. Thus [16,92]:

Lemma 10 *A vector field X_f (or a dynamical system $\dot{x} = f(x)$) can be linearized if and only if it admits a (possibly formal) symmetry X_g such that $B = (Dg)(0) = I$.*

Proceeding in a similar way – but using Joint Normal Forms – we have, more generally:

Theorem 9 *The vector field X_f with $A = (Df)(0)$ can be linearized if and only if it admits a (possibly formal) symmetry X_g with $B = (Dg)(0)$ such that A and B do not admit common resonances, i. e. $\mathrm{Ker}(\mathcal{A}) \cap \mathrm{Ker}(\mathcal{B}) = \{0\}$.*

This result can be easily extended not only to the case of more than one symmetry, as an obvious consequence of Theorem 4, but also to the non-semisimple case [48].

Another interesting result related to linear dynamical systems, is the following [48,51]:

Theorem 10 *If a dynamical system in* $\mathbf{R}^n$ *can be linearized, then it admits n independent commuting symmetries, that can be simultaneously linearized.*

Further Normalization and Symmetry

As repeatedly noted above (see in particular Remark 1), when the linear part of the dynamical vector field is resonant the resulting degeneracy in the solution to the homological equation is not a real degeneracy for what concerns effects on higher order terms in the normal form.

These higher order terms could – and in general will – generate resonant terms, which cannot be eliminated by the standard algorithm. On the other hand, it is clear that this could be seen as a bonus rather than as a problem: in fact, it is conceivable that by carefully choosing the component of h_k lying in $\mathrm{Ker}(\mathcal{A})$, one could generate resonant terms which exactly cancel those already present in the vector field.

Several algorithms have been designed to take advantage, in one way or the other, of this possibility; some review of different approaches is provided in the paper [39]. Here we are concerned with those based on symmetry properties, and discuss two different approaches, developed respectively by the present author [81,83] and by Palacian and Yanguas [155].

It should be stressed that once the presence of additional symmetries – and in the Hamiltonian context, additional constants of motion – has been determined for the normal form truncated at some finite order N, one should investigate if the set of symmetries (or constants of motion) persist under small perturbations; in particular, when considering terms of higher order as well. A general tool to investigate this kind of question is provided by the Nekhoroshev generalization of the Poincaré-Lyapounov theorem [148,149,150]; see also the discussion in [15,87,88,91].

Further Normalization and Resonant Further Symmetry

We will assume again, for ease of discussion, that the matrix A associated with the linear part of the dynamical vector field X is normal. In this case, as discussed above, the normal form is written as $X = g^i(x)\partial_i$ where $g(x) = \sum_{k=0}^{\infty} G_k(x)$, with $G_k \in \mathcal{V}_k$ and all G_k being resonant. We can correspondingly write

$$X = \sum_{k=0}^{\infty} X_k \,, X_k = G_k^i \partial_i \,. \tag{44}$$

As X is in normal forms, we are guaranteed to have

$$[X_0, X_k] = 0 \quad \forall k = 0, 1, 2 \ldots \,; \tag{45}$$

this corresponds to the characterization of normal forms in terms of symmetry as discussed in Sect. "Symmetry Characterization of Normal Forms". In other words, $G_k \in \mathrm{Ker}(\mathcal{A})$ and hence, defining $\mathcal{G}_0$ as the Lie algebra of vector fields commuting with X_0, $X_k \in \mathcal{G}_0$. On the other hand, in general it will be $[X_j, X_k] \neq 0$ for $j \neq k$ and both j, k greater than zero; thus $\mathcal{G}_0$ is not, in general, an Abelian Lie algebra: we can only state that X_0 belongs to the center of $\mathcal{G}_0$, $X_0 \in Z(\mathcal{G}_0)$.

Suppose we want to operate further Lie-Poincaré transformations generated by functions which are symmetric under X_0 (i. e. are in the kernel of $\mathcal{A}$). It follows from the formulas obtained in Sect. "Perturbation Theory: Normal Forms" that these will map $\mathcal{G}_0$ into itself, i. e. $\mathcal{G}_0$ is globally invariant under this restricted set of Lie-Poincaré transformations.

As for the individual vector fields, it follows from the general formula (24) that each of them is invariant under such a transformation *at first order in* h_k, but not at higher orders. That is, making use again of Remark 1, we can still in this way generate new resonant terms in the normal form, including maybe terms which cancel some of those present in (44).

By looking at the explicit formulas (25), it is rather easy to analyze in detail the higher order terms generated in the concerned Lie-Poincaré transformation.

Note that we did not take into account problems connected with the convergence of the further normalizing transformations; it has to be expected that each step will reduce the radius of convergence, so that the further normalized forms will be actually (and not just formally) conjugated to the original dynamic in smaller and smaller neighborhoods of the fixed point; we refer to [90] for an illustration of this point by explicit examples and numerical computations; and to ▶ Perturbative Expansions, Convergence of for a general discussion on the convergence of normalizing transformations.

In order to state exactly the result obtained by this construction, we need to introduce some function spaces, which require abstract definitions. With[(16)] $\mathcal{L}_k(.) := [X_k, .]$, we set

$$H^{(p)} := \mathrm{Ker}(\mathcal{L}_0) \cap \ldots \cap \mathrm{Ker}(\mathcal{L}_{p-1}) \tag{46}$$

(note $H^{(q)} \subseteq H^{(p)}$ for $q > p$), and denote by $\mathcal{M}_p$ the restriction of $\mathcal{L}_p$ to $H^{(p)}$; thus $\mathrm{Ker}(\mathcal{M}_p) = H^{(p+1)}$. We define spaces $F^{(p)}$ (with $F^{(0)} = \mathcal{V}$) as

$$F^{(p)} := \mathrm{Ker}(\mathcal{M}_0^+) \cap \ldots \cap \mathrm{Ker}(\mathcal{M}_{p-1}^+)\,; \tag{47}$$

the adjoint should be meant in the sense of the scalar product introduced in Sect. "Perturbation Theory: Normal Forms". We also write $F_k^{(p)} = F^{(p)} \cap \mathcal{V}_k$.

We will say that X is in **Poincaré renormalized form** up to order N if $G_k \in F_k^{(k)}$ for all $k \leq N$.

Theorem 11 *The vector field X can be formally taken into Poincaré renormalized form up to (any finite) order N by means of a (finite) sequence of Lie-Poincaré transformations.*

For a proof of this theorem, and a detailed description of the renormalizing procedure, see [51,81,83,85]; see [51,81] for the case where additional symmetries are present; an improved procedure, taking full advantage of the Lie algebraic structure of $\mathcal{G}_0$, is described in [84].

Further Normalization and External Symmetry

A different approach to further reduction (simplification) of vector fields in normal form has been developed by Palacian and Yanguas [155] (and applied mainly in the context of Hamiltonian dynamics [154,156,157]), making use of a result by Meyer [144].

As discussed in the previous subsection we have $[X_0, X_k] = 0$ for all k. Suppose now there is a linear (in the coordinates used for the decomposition (44)) vector field Y such that $[X_0, Y] = 0$. Then the Jacobi identity guarantees that

$$[X_0, [X_k, Y]] = 0 \quad \forall k\,. \tag{48}$$

We assume that Y also corresponds to a normal matrix B, so that the homological operator $\mathcal{B}$ associated to it is also normal.

We can then proceed to further normalization as above, being guaranteed that – provided we choose $h_k \in \mathrm{Ker}(\mathcal{L}_0)$ – the resulting vector fields will not only still be in $\mathcal{G}_0$, but also still satisfy (48). One can use freedom in the choice of the generator $h_k \in \mathrm{Ker}(\mathcal{L}_0)$ for the further normalization in a different way than discussed above: that is, we can choose it so that $[Y, X_k] = 0$; in other words, we will get $\widehat{X} \in \mathrm{Ker}(\mathcal{A}) \cap \mathrm{Ker}(\mathcal{B})$.

Note the advantage of this: we do not have to worry about complicated matters related to relevant homological operators acting between different spaces, as we only make use of the homological operator $\mathcal{B}$ associated with the "external" symmetry linear vector field and thus mapping each $\mathcal{V}_k$ into the same $\mathcal{V}_k$.

The result one can obtain in this way is the following (which we quote in a simplified setting for the sake of brevity; in particular the normality assumption can be relaxed) [155].

Theorem 12 *Let X be in normal form; assume moreover $X_0 = Ax$ and there is a normal matrix B such that $[A, B] = 0$; denote by Y the associated vector field, $Y = (Bx)^i \partial_i$. Assume moreover for each resonant vector field R_k there is Q_k satisfying $[Y, Q_k] = 0$ and such that $\widehat{R}_k = R_k + [X_0, Q_k]$ commutes with Y. Then X can be taken to a (different) normal form $\widehat{X}$ such that $[Y, \widehat{X}] = 0$.*

Applications of this theorem, and more generally of this approach, are discussed e. g. in [154,155,156,157]. We stress that albeit the assumptions of this theorem are rather strong(17), it points out to the fact that *there are* cases in which a symmetry – and in the Hamiltonian framework, an integral of motion – of the linear part can be extended to a symmetry of the full normal form.

Symmetry Reduction of Symmetric Normal Forms

Symmetry reduction is a general – and powerful – approach to the study of nonlinear dynamical systems. (In the Hamiltonian case, this is also known as (Marsden–Weinstein) moment map [6,134,135].)

A general theory based on the geometry of group action has been developed by Louis Michel; this was originally motivated by the study of spontaneous symmetry breaking in high-energy physics [140,141] (see [136,137,138,139] for the simpler case where only stationary solutions are considered, and [142] for the full theory and applications; see also [2,82,85,89,93,166,167]).

A description of this would lead us too far away from the scope of this paper, but as this theory also applies to vector fields in normal form, we will briefly describe the results that can be obtained in this way; we will mainly follow [94] (see [95,97] for further detail, extensions, and a more abstract mathematical formulation).

As mentioned in Sect. "Symmetry of Dynamical Systems", the Lie algebra of vector fields in normal form is infinite dimensional, but also has the structure of a Lie module over the algebra of constants of motion for the linear part X_0 of the vector field (which remains the same under all the considered transformations).

Let us recall that vector monomial $\mathbf{v}_{\mu,\alpha} := x^{\mu}\mathbf{e}_{\alpha}$ is *resonant* with A if

$$(\mu \cdot \lambda) := \sum_{i=1}^{n} \mu_i \lambda_i = \lambda_\alpha \quad \text{with} \mu_i \geq 0 \,, \qquad |\mu| := \sum_{i=1}^{n} \mu_i \geq 1 \,; \tag{49}$$

here λ_i are the eigenvalues of A, which we suppose to be semisimple, for the sake of simplicity (in the general case one would consider A_s rather than A).

As mentioned in Sect. "Perturbation Theory: Normal Forms", the relation $(\mu \cdot \lambda) = \lambda_\alpha$ is said to be a *resonance relation* related to the eigenvalue λ_α, and the integer $|\mu|$ is said to be the *order* of the resonance. In the present context it is useful to include order one resonances in the definition (albeit the *trivial* order one resonances given by $\lambda_\alpha = \lambda_\alpha$ are obviously of little interest).

Let us consider again the resonance Eq. (49). It is clear that if there are non-negative integers σ_i (some of them nonzero) such that

$$\sum_{i=1}^{n} \sigma_i \lambda_i = 0 \,, \tag{50}$$

then we always have infinitely many resonances. In this case the monomial $\varphi = x^{\sigma}$ will be called a *resonant scalar monomial*. It is an invariant of X_0, and any multi-index μ with $\mu_i = k\sigma_i + \delta_{i\alpha}$ provides a resonance relation $(\mu \cdot \lambda) = \lambda_\alpha$ related to the eigenvalue λ_α; in other words, any monomial $x^{k\sigma}x^{\alpha} = \varphi^k x^{\alpha}$ is resonant, and so is any vector $\mathbf{v}_{k\sigma+e_\alpha,\alpha}$.

Therefore, we say that (50) identifies a *invariance relation*. The presence of invariance relations is the only way to have infinitely many resonances in a finite dimensional system (see [186]).

Any nontrivial resonance (49) which does not originate in an invariance relation, is said to be a *sporadic resonance*. Sporadic resonances are always in finite number (if any) in a finite dimensional system [186].

Any invariance relation (50) such that there is no ν with $\nu_i \leq \sigma_i$ (and of course $\nu \neq \sigma$) providing another invariance relation, is said to be an *elementary invariance relation*. Every invariance relation is a linear combination (with nonnegative integer coefficients) of elementary ones. Elementary invariance relations are always in finite number (if any) in a finite dimensional system [186]. If there are m independent elementary invariance relations, each of them of the form (50), we associate to these monomials $\beta_j = x^{\sigma} = \prod_{i=1}^{n} x_i^{\sigma_i} (j = 1, \ldots, m)$.

Similarly, if there are r sporadic resonances (49), we associate resonant monomials $\alpha^j(x) = x^{\mu} = \prod_{i=1}^{n} x_i^{\mu_i} (j = 1, \ldots, r)$ and resonant vectors $\mathbf{v}^{(j)}_{\mu,\alpha}$ to sporadic resonances. We then introduce two set of new coordinates: these will be the coordinates $w^1, \ldots, w^r$ in correspondence with sporadic resonances, and other new coordinates $\varphi^1, \ldots, \varphi^m$ in correspondence with elementary invariance relations. The evolution equations for the x^i can be written in simplified form using these (note that some ambiguity is present here, in that we can write these in different ways in terms of the x, w, φ), but we should also assign evolution equations for them; these will be given in agreement with the dynamics itself. That is, we set

$$\frac{\mathrm{d}w^j}{\mathrm{d}t} = \frac{\partial w^j}{\partial x^i}\frac{\mathrm{d}x^i}{\mathrm{d}t} := h^j(x, w, \varphi) \,; \tag{51}$$

$$\frac{\mathrm{d}\varphi^a}{\mathrm{d}t} = \frac{\partial \varphi^a}{\partial x^i}\frac{\mathrm{d}x^i}{\mathrm{d}t} := z^a(x, w, \varphi) \,. \tag{52}$$

We are thus led to consider the enlarged space $W = \mathbf{R}^{n+r+m}$ of the (x, w, φ), and in this the vector field

$$Y = f^i(x, w, \varphi)\,(\partial/\partial x^i) + h^j(x, w, \varphi)\,(\partial/\partial w^j) + z^a(x, w, \varphi)\,(\partial/\partial \varphi^a) \,. \tag{53}$$

The vector field Y is uniquely defined on the manifold identified by $\psi^j := w^j - \alpha^j(x) = 0, \varphi^a - \beta^a(x) = 0$. It is obvious (by construction) that the $(n+m)$-dimensional manifold $M \subset W$ identified by $\psi^i := w^i - \alpha^{(i)} = 0$ is invariant under the flow of Y, see (51). It is also easy to show that the functions z^a defined in (52) can be written in terms of the φ variables alone, i. e. $\partial z^a/\partial x^i = \partial z^a/\partial w^j = 0$. This implies[18]

Lemma 11 *The evolution of the φ variables is described by a (generally, nonlinear) equation involving the φ variables alone.*

Note that the equations for x and w depend on φ and are therefore non-autonomous. We have the following result (we refer to [94,95] for a proof; see [97] for extensions).

Theorem 13 *The analytic functions f^i and h^j defined above can be written as linear in the x and w variables, the coefficients being functions of the φ variables. Hence the evolution of the x and w variables is described by non-autonomous linear equations, obtained by inserting the solution $\varphi = \varphi(t)$ of the equations for φ in the equations $\dot{x} = f(x, w, \varphi)$, $\dot{w} = h(x, w, \varphi)$.*

Note that if no invariance relations are present, hence no φ variables are introduced, then the system describing the time evolution of the x, w variables is linear; in this case we can interpret normal forms as projections of a linear

system to an invariant manifold (without symmetry reduction).

If there are no sporadic resonances of order greater than one, then upon solving the reduced equation for the φ variables one obtains a non-autonomous linear system. Moreover, if all eigenvalues are distinct then we have a product system of one-dimensional equations.

If $\varphi(t)$ converges to some constant φ_0, the asymptotic evolution of the system is governed by a linear autonomous equation for x and w. Similarly, if there is a periodic solution $\bar{\varphi}(t)$ and $\varphi(t)$ converges to $\bar{\varphi}(t)$ for large t, the asymptotic evolution of the system is governed by a linear equation with t-periodic coefficients for x and w.

Conclusions

We have reviewed the basic notions concerning symmetry of dynamical systems and its determination, in particular in a perturbative setting. We have subsequently considered various situations where the interplay of perturbation theory and symmetry properties produce nontrivial results, either in that the perturbative expansion turns out to be simplified (with respect to the general case) due to symmetry properties, or in that computations of symmetry is simplified by dealing with a system in normal form; we then considered the problem of jointly normalizing an algebra of vector fields (with possibly but not necessarily one of these defining a dynamical system, the other being symmetry vector fields). We also discussed how normal forms can be characterized in terms of symmetry, and how this is extended to a characterization of "renormalized forms". Finally we considered symmetry reduction applied to systems in normal form.

The discussion conducted here illustrates some of the powerful conceptual and computational simplifications arising for systems with symmetry, also in the realm of perturbation theory.

As remarked in the Introduction, symmetry is a non-generic property; on the other hand, it is often the case that equations arising from physical (mechanical, electronic, etc.) systems enjoy some degree of symmetry, as a consequence of the symmetry of the fundamental equations of physics. Disregarding the symmetry properties in these cases would mean renouncing the use of what is often the only handle to grab the behavior of non-linear systems; and correspondingly a symmetry analysis can often on the one hand lead to identifying several relevant properties of the system even without a complete solution, and on the other hand being instrumental in obtaining exact (or approximate, as we are here dealing with perturbation theory) solutions.

Here we discussed some of the consequences of symmetry for the specific case of dynamical systems, such as those met in analyzing the behavior of nonlinear systems near a known (e. g. trivial) solution.

For a more general discussion, as well as for concrete applications, the reader is referred on the one hand to texts discussing symmetry for differential equations [3,19,26,27,37,80,115,125,151,152,174], on the other to texts and papers specifically dealing with the interplay of symmetry and perturbation theory, quoted in the main text and listed in the ample Bibliography below.

Future Developments

First and foremost, future developments should be expected to concern further application of the general theory in concrete cases, both in nonlinear theoretical mechanics and in specific subfields, ranging from the more applied (e. g., ship dynamics and stability [13,176]; or handling of complex electrical networks [133,183]) to the more theoretical (e. g., galactic dynamics [24,59]).

On the other hand, the theory developed so far is in many cases purely formal, in that consideration of convergence properties – and estimation of the convergence region in phase and/or parameter space – of the resulting (perturbative) series is often left aside, with the understanding that in any concrete application one will have explicit series and be ready to analyze their convergence. The story of general (i. e. non-symmetric) perturbation theory shows however that the theoretical analysis of convergence properties can be precious – not only for the conceptual understanding but also in view of concrete applications – and it should thus be expected that future developments will deal with convergence properties in the symmetric case (see e. g. ► Perturbative Expansions, Convergence of, [52] for a review of existing results).

The same issue of convergence, and estimation of the convergence region, arises in connection with further normalization (under any approach), and has so far been given little consideration.

A different kind of generalization is called for when dealing with symmetry reduction of normal forms: in fact, on the one hand it is natural to try applying the same approach (based on quite general geometrical properties) to more general systems than initially considered; on the other hand the method discussed in Sect. "Symmetry Reduction of Symmetric Normal Forms" is algorithmic and could be implemented by symbolic manipulations packages – such as those already existing for computations of symmetry of differential equations – which would be of

help to anybody having to deal with perturbation of concrete symmetric systems.

Finally, another field of future developments can be called for: here we discussed the interplay of perturbation theory and "standard" symmetries. Or, the notion of symmetry of differential equations has been generalized in various directions, producing in some cases a significant advantage in application to concrete systems. It should thus be expected that the interplay between these generalized notions of symmetry and perturbation theory will be investigated in the near future, and most probably will produce interesting and readily applicable results.

Additional Notes

We collect here the additional notes called throughout the manuscript.

(1) Note that $\widetilde{M}$ is naturally a *fiber bundle* [40,119,146,147] over t: that is, it can be decomposed as the union of copies of M, each one in correspondence to a value of t. A *section* of this bundle is simply the graph of a function $\sigma : t \to M$. The set σ_x considered in a moment is a section of this fiber bundle.

(2) For general differential equations, one would go along the same lines. A relevant difference is however present: for (systems of) first order equations there is no algorithmic way to find the general solutions to determining equations, as opposed to any other case [151,174].

(3) In the case of Hamiltonian systems, one can work directly on the Hamiltonian $H(p, q)$ rather than on the Hamilton equations of motion ► Hamiltonian Perturbation Theory (and Transition to Chaos), [111]. We will however, for the sake of brevity, not discuss the specific features of the Hamiltonian case.

(4) Later on, in particular when we deal with different homological operators, it will be convenient to also denote this $\mathcal{A}$ as L_A, with reference to the matrix A appearing in $F_0 = Ax$.

(5) This is equivalent to defining $(\xi_{\mu,i}, \xi_{\nu,j}) = \delta_{\mu,\nu}\delta_{i,j}(\mu!)$, where for the multi-index μ we have defined $\mu! = (\mu_1!)\dots(\mu_n!)$.

(6) The name "resonant" is due to the relation existing between eigenvectors of $\mathcal{A}$ and resonance relations among eigenvalues of the matrix $A = (Df)(0)$ describing the linear part $F_0(x) = Ax$ of the system.

(7) In order to determine $\widetilde{X} = \widetilde{f}^i(x)\partial/\partial x^i$, we write (using the exponential notation) $\xi(t + \delta t) = e^{\delta t X}\xi(t)$. Therefore, $x(t + \delta t) = [e^{sH}e^{(\delta t)X}\xi(t)]_{s=1}$. Using $\xi(t) = e^{-H}x(t)$, we have $[x(t + \delta t) - x(t)] = [e^{sH}(e^{(\delta t)X} - I)e^{-sH}x(t)]_{(s=1)}$, and therefore (24).

(8) Note, however, that when the diagonalizing matrix M is not unitary, this transformation changes implicitly the scalar product.

(9) The general case is discussed e. g. in [51]; see also, for a more general discussion on normal forms with non-normal and non-semisimple normal form, the article ► Perturbation of Systems with Nilpotent Real Part.

(10) Note that the terms in $\mathrm{Ker}(\mathcal{A}) = \mathrm{Ker}(\mathcal{A}^+)$ are just the resonant ones. Thus, the present characterization of normal forms is equivalent to the one given earlier on.

(11) We also note that in the framework of Lie–Poincaré transformations, i. e. when considering the time one action of a vector field Φ [25,28,145] (see above), Theorem 4 shows that Φ can be chosen to admit Y as a symmetry; see [51].

(12) Generalizations can also be obtained in the direction of relaxing the normality assumption for the matrices identifying the linear part of vector fields. We will not discuss this case, referring the reader instead to ► Perturbation of Systems with Nilpotent Real Part, [51].

(13) If this is the symmetry algebra of one of the vector fields, say X_0, we have that $[X_0, X_i] = 0$ for all the i, i. e. X_0 belongs to the center of the algebra $\mathcal{G}$.

(14) We stress that we refer here to the case of a nilpotent Lie algebra, not to the case where the relevant matrices are nilpotent!

(15) This is readily shown by the following example. Consider the two-dimensional system, which is *not* in normal form, $\dot{x}_1 = -x_1x_2$, $\dot{x}_2 = x_2$. It is easily seen that $\mu = x_1 \exp(x_2)$ is a constant of motion, but not of the linearized problem, i. e. $\mu \in \mathcal{I}_X$, but $\mu \notin \mathcal{I}_A$. Similarly, we have that $Y = x_1^2 \exp(x_2)(\partial/\partial x_1) \in \mathcal{G}_X$ but $Y \notin \mathcal{G}_A$.

(16) It should be noted that X_k, and hence $\mathcal{L}_k$, change under further normalization transformations; however, they stabilize after a finite number of steps, and in particular will not change anymore after the further normalization reaches their order.

(17) Note that the $\widehat{X}_k$ in Theorem 12 satisfies $\mathcal{B}(\widehat{X}_k) := [Y, \widehat{X}_k] = [Y, X_k] + [Y, [X_0, Q_k]]$; using the Jacobi identity, and assuming $[Y, Q_k] = 0$, this reads $\mathcal{B}(X_k) - \mathcal{A}[\mathcal{B}(Q_k)]$. On the other hand, $[A, B] = 0$ guarantees $[\mathcal{A}, \mathcal{B}] = 0$, hence we get $\mathcal{B}(\widehat{X}_k) = \mathcal{B}[X_k - \mathcal{A}(Q_k)]$. Thus the assumption of Theorem 12 could be rephrased in terms of kernels of the operators $\mathcal{A}^+$ and $\mathcal{B}$ as follows: for each $X_k \in \mathrm{Ker}(\mathcal{A}^+) \cap \mathcal{V}_k$, there exists $Q_k \in Ker(\mathcal{B})$ such that $[X_k - \mathcal{A}(Q_k)] \in \mathrm{Ker}(\mathcal{B})$.

(18) As the φ identify group orbits for the group G generated by the Lie algebra, we interpret $\dot{\varphi} = z(\varphi)$ as an equation in orbit space, and the equation for (x, w) as an equation on the Lie group G. Methods for the solution of the latter are discussed in [190], see also [38].

Bibliography

1. Abenda S, Gaeta G, Walcher S (eds) (2003) Symmetry and Perturbation Theory – SPT2002. In: Proceedings of Cala Gonone workshop, 19–26 May 2002. World Scientific, Singapore
2. Abud M, Sartori G (1983) The geometry of spontaneous symmetry breaking. Ann Phys 150:307–372
3. Aleekseevskij DV, Vinogradov AM, Lychagin VV (1991) Basic ideas and concepts of differential geometry. In: Gamkrelidze RV (ed) Encyclopaedia of Mathematical Sciences vol 28 – Geometry I. Springer, Berlin
4. Arnal D, Ben Ammar M, Pinczon G (1984) The Poincaré–Dulac theorem for nonlinear representations of nilpotent Lie algebras. Lett Math Phys 8:467–476
5. Arnold VI (1974) Equations differentielles ordinaires. MIR, Moscow, 2nd edn 1990. Arnold VI (1992) Ordinary Differential Equations. Springer, Berlin
6. Arnold V (1976) Les méthodes mathématiques de la mecanique classique. MIR, Moscow. Arnold VI (1983, 1989) Mathematical methods of classical Mechanics. Springer, Berlin
7. Arnold V (1980) Chapitres supplementaires de la théorie des equations differentielles ordinaires. MIR, Moscow. Arnold VI (1983) Geometrical methods in the theory of ordinary differential equations. Springer, Berlin
8. Arnold VI, Il'yashenko YS (1988) Ordinary differential equations. In: Anosov DV, Arnold VI (eds) Encyclopaedia of Mathematical Sciences vol 1 – Dynamical Systems I, pp 1–148. Springer, Berlin
9. Arnold VI, Kozlov VV, Neishtadt AI (1993) Mathematical aspects of classical and celestial mechanics. In: Arnold VI (ed) Encyclopaedia of Mathematical Sciences vol 3 – Dynamical Systems III, 2nd edn, pp 1–291. Springer, Berlin
10. Baider A (1989) Unique normal form for vector fields and Hamiltonians. J Diff Eqs 78:33–52
11. Baider A, Churchill RC (1988) Uniqueness and non-uniqueness of normal forms for vector fields. Proc R Soc Edinburgh A 108:27–33
12. Baider A, Sanders J (1992) Further reduction of the Takens-Bogdanov normal form. J Diff Eqs 99:205–244
13. Bakri T, Nabergoj R, Tondl A, Verhulst F (2004) Parametric excitation in non-linear dynamics. Int J Nonlin Mech 39:311–329
14. Bambusi D, Gaeta G (eds) (1997) Symmetry and Perturbation Theory. In: Proceedings of Torino Workshop, ISI, December 1996. GNFM–CNR, Roma
15. Bambusi D, Gaeta G (2002) On persistence of invariant tori and a theorem by Nekhoroshev. Math Phys El J 8:1–13
16. Bambusi D, Cicogna G, Gaeta G, Marmo G (1998) Normal forms, symmetry, and linearization of dynamical systems. J Phys A Math Gen 31:5065–5082
17. Bambusi D, Gaeta G, Cadoni M (eds) (2001) Symmetry and Perturbation Theory – SPT2001. In: Proceedings of the international conference SPT2001, Cala Gonone, 6-13 May 2001. World Scientific, Singapore
18. Bargmann V (1961) On a Hilbert space of analytic functions and an associated integral transform. Comm Pure Appl Math 14:187–214
19. Baumann G (2000) Symmetry analysis of differential equations with Mathematica. Springer, New York
20. Belitskii GR (1978) Equivalence and normal forms of germs of smooth mappings. Russ Math Surveys 33(1):107–177
21. Belitskii GR (1981) Normal forms relative to the filtering action of a group. Trans Moscow Math Soc 40(2):1–39
22. Belitskii GR (1987) Smooth equivalence of germs of vector fields with a single eigenvalue or a pair of purely imaginary eigenvalues. Funct Anal Appl 20:253–259
23. Belitskii GR (2002) C^∞-Normal forms of local vector fields. Acta Appl Math 70:23–41
24. Belmonte C, Boccaletti D, Pucacco G (2006) Stability of axial orbits in galactic potentials. Cel Mech Dyn Astr 95:101–116
25. Benettin G, Galgani L, Giorgilli A (1984) A proof of the Kolmogorov theorem on invariant tori using canonical transformations defined by the Lie method. Nuovo Cimento B 79:201–223
26. Bluman GW, Anco SC (2002) Simmetry and integration methods for differential equations. Springer, Berlin
27. Bluman GW, Kumei S (1989) Symmetries and differential equations. Springer, Berlin
28. Bogoliubov NN, Mitropolsky VA (1961) Asymptotic methods in the theory of nonlinear oscillations. Hindustan, New Delhi. (1962) Méthodes asymptothiques dans la théorie des oscillations non-linéaires. Gauthier-Villars, Paris
29. Broer HW (1979) Bifurcations of singularities in volume preserving vector fields. Ph D Thesis, Groningen
30. Broer HW (1981) Formal normal form theorems for vector fields and some consequences for bifurcations in the volume preserving case. In: Rand DA, Young LS (eds) Dynamical systems and turbulence. it Lect Notes Math 898. Springer, Berlin
31. Broer HW, Takens F (1989) Formally symmetric normal forms and genericity. Dyn Rep 2:39–59
32. Bryuno AD (1971) Analytical form of differential equations I. Trans Moscow Math Soc 25:131–288
33. Bryuno AD (1971) Analytical form of differential equations II. Trans Moscow Math Soc 26:199–239
34. Bryuno AD (1988) The normal form of a Hamiltonian system. Russ Math Sur 43(1):25–66
35. Bryuno AD (1989) Local Methods in the Theory of Differential Equations. Springer, Berlin
36. Bryuno AD, Walcher S (1994) Symmetries and convergence of normalizing transformations. J Math Anal Appl 183:571–576
37. Cantwell BJ (2002) Introduction to Symmetry Analysis. Cambridge University Press, Cambridge
38. Carinena JF, Grabowski J, Marmo G (2000) Lie-Scheffers systems: a geometric approach. Bibliopolis, Napoli
39. Chen G, Della Dora J (2000) Further reductions of normal forms for dynamical systems. J Diff Eqs 166:79–106
40. Chern SS, Chen WH, Lam KS (1999) Lectures on differential geometry. World Scientific, Singapore
41. Chossat P (2002) The reduction of equivariant dynamics to teh orbit space for compact group actions. Acta Appl Math 70:71–94
42. Chossat P, Lauterbach R (1999) Methods in equivariant bifurcations and dynamical systems with applications. World Scientific, Singapore

43. Chow SN, Hale JK (1982) Methods of bifurcation theory. Springer, Berlin
44. Chow SN, Li C, Wang D (1994) Normal forms and bifurcations of planar vector fields. Cambridge University Press, Cambridge
45. Chua LO, Kokubu H (1988) Normal forms for nonlinear vector fields Part I: theory. IEEE Trans Circ Syst 35:863–880
46. Chua LO, Kokubu H (1989) Normal forms for nonlinear vector fields Part II: applications. IEEE Trans Circ Syst 36:851–870
47. Churchill RC, Kummer M, Rod DL (1983) On averaging, reduction and symmetry in Hamiltonian systems. J Diff Eqs 49:359–414
48. Cicogna G, Gaeta G (1994) Normal forms and nonlinear symmetries. J Phys A 27:7115–7124
49. Cicogna G, Gaeta G (1994) Poincaré normal forms and Lie point symmetries. J Phys A 27:461–476
50. Cicogna G, Gaeta G (1994) Symmetry invariance and center manifolds in dynamical systems. Nuovo Cim B 109:59–76
51. Cicogna G, Gaeta G (1999) Symmetry and perturbation theory in nonlinear dynamics. Springer, Berlin
52. Cicogna G, Walcher S (2002) Convergence of normal form transformations: the role of symmetries. Acta Appl Math 70:95–111
53. Courant R, Hilbert D (1962) Methods of Mathematical Physics. Wiley, New York; (1989)
54. Cushman R, Sanders JA (1986) Nilpotent normal forms and representation theory of $sl2, R$. In: Golubitsky M, Guckenheimer J (eds) Multi-parameter bifurcation theory. Contemp Math 56, AMS, Providence,
55. Crawford JD (1991) Introduction to bifurcation theory. Rev Mod Phys 63:991–1037
56. Crawford JD, Knobloch E (1991) Symmetry and symmetry-breaking bifurcations in fluid dynamics. Ann Rev Fluid Mech 23:341–387
57. Degasperis A, Gaeta G (eds) (1999) Symmetry and Perturbation Theory II – SPT98. In: Proceedings of Roma Workshop, Universitá La Sapienza, December 1998. World Scientific, Singapore
58. Deprit A (1969) Canonical transformation depending on a small parameter. Celest Mech 1:12–30
59. de Zeeuw T, Merritt D (1983) Stellar orbits in a triaxial galaxy I Orbits in the plane of rotation. Astrophys J 267:571–595
60. Elphick C, Tirapegui E, Brachet ME, Coullet P, Iooss G (1987) A simple global characterization for normal forms of singular vector fields. Physica D 29:95–127. (1988) Addendum. Physica D 32:488
61. Fassò F (1990) Lie series method for vector fields and Hamiltonian perturbation theory. ZAMP 41:843–864
62. Fassò F, Guzzo M, Benettin G (1998) Nekhoroshev stability of elliptic equilibria of Hamiltonian systems. Comm Math Phys 197:347–360
63. Field MJ (1989) Equivariant bifurcation theory and symmetry breaking. J Dyn Dif Eqs 1:369–421
64. Field MJ (1996) Lectures on bifurcations, dynamics and symmetry. Res Notes Math 356. Pitman, Boston
65. Field MJ (1996) Symmetry breaking for compact Lie groups. Mem AMS 574:1–170
66. Field MJ, Richardson RW (1989) Symmetry breaking and the maximal isotropy subgroup conjecture for reflection groups. Arch Rat Mech Anal 105:61–94
67. Field MJ, Richardson RW (1990) Symmetry breaking in equivariant bifurcation problems. Bull Am Math Soc 22:79–84
68. Field MJ, Richardson RW (1992) Symmetry breaking and branching patterns in equivariant bifurcation theory I. Arch Rat Mech Anal 118:297–348
69. Field MJ, Richardson RW (1992) Symmetry breaking and branching patterns in equivariant bifurcation theory II. Arch Rat Mech Anal 120:147–190
70. Fokas AS (1979) Generalized symmetries and constants of motion of evolution equations. Lett Math Phys 3:467–473
71. Fokas AS (1979) Group theoretical aspects of constants of motion and separable solutions in classical mechanics. J Math Anal Appl 68:347–370
72. Fokas AS (1980) A symmetry approach to exactly solvable evolution equations. J Math Phys 21:1318–1326
73. Fokas AS (1987) Symmetries and integrability. Stud Appl Math 77:253–299
74. Fokas AS, Gelfand IM (1996) Surfaces on Lie groups, Lie algebras, and their integrability. Comm Math Phys 177:203–220
75. Fontich E, Gelfreich VG (1997) On analytical properties of normal forms. Nonlinearity 10:467–477
76. Forest E, Murray D (1994) Freedom in minimal normal forms. Physica D 74:181–196
77. Fushchich WI, Nikitin AG (1987) Symmetries of Maxwell equations. Reidel, Dordrecht
78. Fushchich WI, Shtelen WM, Slavutsky SL (1989) Symmetry analysis and exact solutions of nonlinear equations of mathematical physics. Naukova Dumka, Kiev
79. Gaeta G (1990) Bifurcation and symmetry breaking. Phys Rep 189:1–87
80. Gaeta G (1994) Nonlinear symmetries and nonlinear equations. Kluwer, Dordrecht
81. Gaeta G (1997) Reduction of Poincaré normal forms. Lett Math Phys 42:103–114 & 235
82. Gaeta G (1999) An equivariant branching lemma for relative equilibria. Nuovo Cim B 114:973–982
83. Gaeta G (1999) Poincaré renormalized forms. Ann IHP Phys Theor 70:461–514
84. Gaeta G (2001) Algorithmic reduction of Poincaré-Dulac normal forms and Lie algebraic structure. Lett Math Phys 57:41–60
85. Gaeta G (2002) Poincaré normal and renormalized forms. Acta Appl Math 70:113–131
86. Gaeta G (2002) Poincaré normal forms and simple compact Lie groups. Int J Mod Phys A 17:3571–3587
87. Gaeta G (2002) The Poincaré–Lyapounov–Nekhoroshev theorem. Ann Phys 297:157–173
88. Gaeta G (2003) The Poincaré-Nekhoroshev map. J Nonlin Math Phys 10:51–64
89. Gaeta G (2006) Finite group symmetry breaking. In: Francoise JP, Naber G, Tsou ST (eds) Encyclopedia of Mathematical Physics. Kluwer, Dordrecht
90. Gaeta G (2006) Non-quadratic additional conserved quantities in Birkhoff normal forms. Cel Mech Dyn Astr 96:63–81
91. Gaeta G (2006) The Poincaré–Lyapounov–Nekhoroshev theorem for involutory systems of vector fields. Ann Phys NY 321:1277–1295
92. Gaeta G, Marmo G (1996) Nonperturbative linearization of dynamical systems. J Phys A 29:5035–5048
93. Gaeta G, Morando P (1997) Michel theory of symmetry breaking and gauge theories. Ann Phys NY 260:149–170

94. Gaeta G, Walcher S (2005) Dimension increase and splitting for Poincaré-Dulac normal forms. J Nonlin Math Phys 12:S1327-S1342
95. Gaeta G, Walcher S (2006) Embedding and splitting ordinary differential equations in normal form. J Diff Eqs 224:98–119
96. Gaeta G, Prinari B, Rauch S, Terracini S (eds) (2005) Symmetry and Perturbation Theory – SPT2004. In: Proceedings of Cala Gonone workshop, 30 May – 6 June 2004. World Scientific, Singapore
97. Gaeta G, Grosshans FD, Scheurle J, Walcher S (2008) Reduction and reconstruction for symmetric ordinary differential equations. J Diff Eqs 244:1810–1839
98. Gaeta G, Vitolo R, Walcher S (eds) (2007) Symmetry and Perturbation Theory – SPT2007. In: Proceedings of Otranto workshop, 2–9 June 2007. World Scientific, Singapore
99. Gallavotti G (1983) The elements of mechanics. Springer, Berlin
100. Giorgilli A (1988) Rigorous results on the power expansions for the integrals of a Hamiltonian system near an elliptic equilibrium point. Ann IHP Phys Theor 48:423–439
101. Giorgilli A, Locatelli U (1997) Kolmogorov theorem and classical perturbation theory. ZAMP 48:220–261
102. Giorgilli A, Morbidelli A (1997) Invariant KAM tori and global stability for Hamiltonian systems. ZAMP 48:102–134
103. Giorgilli A, Zehnder E (1992) Exponential stability for time dependent potentials. ZAMP 43:827–855
104. Glendinning P (1994) Stability, instability and chaos: an introduction to the theory of nonlinear differential equations. Cambridge University Press, Cambridge
105. Golubitsky M, Stewart I, Schaeffer D (1988) Singularity and groups in bifurcation theory – vol II. Springer, Berlin
106. Gramchev T, Yoshino M (1999) Rapidly convergent iteration methods for simultaneous normal forms of commuting maps. Math Z 231:745–770
107. Guckenheimer J, Holmes P (1983) Nonlinear oscillations, dynamical systems, and bifurcation of vector fields. Springer, Berlin
108. Gustavson FG (1964) On constructing formal integrals of a Hamiltonian system near an equilibrium point Astron J 71:670–686
109. Guzzo M, Fassò F, Benettin G (1998) On the stability of elliptic equilibria. Math Phys El J 4(1):16
110. Hamermesh M (1962) Group theory. Addison-Wesley, Reading; reprinted by Dover, New York (1991)
111. Hanssmann H (2007) Local and semi-local bifurcations in Hamiltonian dynamical systems Results and examples. Springer, Berlin
112. Hermann R (1968) The formal linearization of a semisimple Lie algebra of vector fields about a singular point. Trans AMS 130:105–109
113. Hoveijn I (1996) Versal deformations and normal forms for reversible and Hamiltonian linear systems. J Diff Eq 126:408–442
114. Hoveijn I, Verhulst F (1990) Chaos in the 1:2:3 Hamiltonian normal form. Physica D 44:397–406
115. Hydon PE (2000) Symmetry methods for differential equations. Cambridge UP, Cambridge
116. Ibragimov N (1992) Group analysis of ordinary differential equations and the invariance principle in Mathematical Physics. Russ Math Surv 47(4):89–156
117. Il'yashenko YS, Yakovenko SY (1991) Finitely smooth normal forms of local families of diffeomorphisms and vector fields. Russ Math Surv 46(1):1–43
118. Iooss G, Adelmeyer M (1992) Topics in bifurcation theory and applications. World Scientific, Singapore
119. Isham CJ (1999) Modern differential geometry for physicists. World Scientific, Singapore
120. Kinyon M, Walcher S (1997) On ordinary differential equations admitting a finite linear group of symmetries. J Math Analysis Appl 216:180–196
121. Kirillov AA (1976, 1984) Elements of the Theory of Representations. Springer, Berlin
122. Kodama Y (1994) Normal forms, symmetry and infinite dimensional Lie algebra for systems of ODE's. Phys Lett A 191:223–228
123. Kokubu H, Oka H, Wang D (1996) Linear grading function and further reduction of normal forms. J Diff Eq 132:293–318
124. Krasil'shchik IS, Vinogradov AM (1984) Nonlocal symmetries and the theory of coverings. Acta Appl Math 2:79–96
125. Krasil'shchik IS, Vinogradov AM (1999) Symmetries and conservation laws for differential equations of mathematical physics. AMS, Providence
126. Kummer M (1971) How to avoid secular terms in classical and quantum mechanics. Nuovo Cimento B 1:123–148
127. Kummer M (1976) On resonant nonlinearly coupled oscillators with two equal frequencies. Comm Math Phys 48:53–79
128. Lamb J (1996) Local bifurcations in *k*-symmetric dynamical systems. Nonlinearity 9:537–557
129. Lamb J (1998) *k*-symmetry and return maps of spacetime symmetric flows. Nonlinearity 11:601–630
130. Lamb J, Melbourne I (2007) Normal form theory for relative equilibria and relative periodic solutions. Trans AMS 359:4537–4556
131. Lamb J, Roberts J (1998) Time reversal symmetry in dynamical systems: a survey. Physica D 112:1–39
132. Levi D, Winternitz P (1989) Non-classical symmetry reduction: example of the Boussinesq equation. J Phys A 22:2915–2924
133. Lin CM, Vittal V, Kliemann W, Fouad AA (1996) Investigation of modal interaction and its effect on control performance in stressed power systems using normal forms of vector fields. IEEE Trans Power Syst 11:781–787
134. Marsden JE (1992) Lectures on Mechanics. Cambridge University Press, Cambridge
135. Marsden JE, Ratiu T (1994) Introduction to mechanics and symmetry. Springer, Berlin
136. Michel L (1971) Points critiques de fonctions invariantes sur une G-variété. Comptes Rendus Acad Sci Paris 272-A:433–436
137. Michel L (1971) Nonlinear group action Smooth action of compact Lie groups on manifolds. In: Sen RN, Weil C (eds) Statistical Mechanics and Field Theory. Israel University Press, Jerusalem
138. Michel L (1975) Les brisure spontanées de symétrie en physique. J Phys Paris 36-C7:41–51
139. Michel L (1980) Symmetry defects and broken symmetry Configurations Hidden symmetry. Rev Mod Phys 52:617–651
140. Michel L, Radicati L (1971) Properties of the breaking of hadronic internal symmetry. Ann Phys NY 66:758–783
141. Michel L, Radicati L (1973) The geometry of the octet. Ann IHP 18:185–214
142. Michel L, Zhilinskii BI (2001) Symmetry, invariants, topology Basic tools. Phys Rep 341:11–84
143. Mikhailov AV, Shabat AB, Yamilov RI (1987) The symmetry ap-

proach to the classification of non-linear equations Complete list of integrable systems. Russ Math Surv 42(4):1–63
144. Meyer KR, Hall GR (1992) Introduction to Hamiltonian dynamical systems and the N-body problem. Springer, New York
145. Mitropolsky YA, Lopatin AK (1995) Nonlinear mechanics, groups and symmetry. Kluwer, Dordrecht
146. Nakahara M (1990) Geometry, Topology and Physics. IOP, Bristol
147. Nash C, Sen S (1983) Topology and geometry for physicists. Academic Press, London
148. Nekhoroshev NN (1994) The Poincaré–Lyapunov–Liouville-Arnol'd theorem. Funct Anal Appl 28:128–129
149. Nekhoroshev NN (2002) Generalizations of Gordon theorem. Regul Chaotic Dyn 7:239–247
150. Nekhoroshev NN (2005) Types of integrability on a submanifold and generalizations of Gordons theorem. Trans Moscow Math Soc 66:169–241
151. Olver PJ (1986) Applications of Lie groups to differential equations. Springer, Berlin
152. Olver PJ (1995) Equivalence, Invariants, and Symmetry. Cambridge University Press, Cambridge
153. Ovsjiannikov LV (1982) Group analysis of differential equations. Academic Press, London
154. Palacián J, Yanguas P (2000) Reduction of polynomial Hamiltonians by the construction of formal integrals. Nonlinearity 13:1021–1054
155. Palacián J, Yanguas P (2001) Generalized normal forms for polynomial vector fields. J Math Pures Appl 80:445–469
156. Palacián J, Yanguas P (2003) Equivariant N-DOF Hamiltonians via generalized normal forms. Comm Cont Math 5:449–480
157. Palacián J, Yanguas P (2005) A universal procedure for normalizing *n*-degree-of-freedom polynomial Hamiltonian systems. SIAM J Appl Math 65:1130–1152
158. Pucci E, Saccomandi G (1992) On the weak symmetry group of partial differential equations. J Math Anal Appl 163:588–598
159. Ruelle D (1973) Bifurcation in the presence of a symmetry group. Arch Rat Mech Anal 51:136–152
160. Ruelle D (1989) Elements of Differentiable Dynamics and Bifurcation Theory. Academic Press, London
161. Sadovskii DA, Delos JB (1996) Bifurcation of the periodic orbits of Hamiltonian systems – an analysis using normal form theory. Phys Rev A 54:2033–2070
162. Sanders JA (2003) Normal form theory and spectral sequences. J Diff Eqs 192:536–552
163. Sanders JA (2005) Normal forms in filtered Lie algebra representations. Acta Appl Math 87:165–189
164. Sanders JA, Verhulst F (1985) Averaging methods in nonlinear dynamical systems. Springer, Berlin
165. Sanders JA, Verhulst F, Murdock J (2007) Averaging methods in nonlinear dynamical systems. Springer, Berlin
166. Sartori G (1991) Geometric invariant theory A model-independent approach to spontaneous symmetry and/or supersymmetry breaking. Riv N Cim 14–11:1–120
167. Sartori G (2002) Geometric invariant theory in a model-independent analysis of spontaneous symmetry and supersymmetry breaking. Acta Appl Math 70:183–207
168. Sartori G, Valente G (2005) Constructive axiomatic approach to the determination of the orbit spaces of coregular compact linear groups. Acta Appl Math 87:191–228
169. Sattinger DH (1979) Group theoretic methods in bifurcation theory. Lecture Notes in Mathematics 762. Springer, Berlin
170. Sattinger DH (1983) Branching in the presence of symmetry. SIAM, Philadelphia
171. Sattinger DH, Weaver O (1986) Lie groups and algebras. Springer, Berlin
172. Siegel K, Moser JK (1971) Lectures on Celestial Mechanics. Springer, Berlin; reprinted in Classics in Mathematics. Springer, Berlin (1995)
173. Sokolov VV (1988) On the symmetries of evolutions equations. Russ Mah Surv 43(5):165–204
174. Stephani H (1989) Differential equations Their solution using symmetries. Cambridge University Press, Cambridge
175. Stewart I (1988) Bifurcation with symmetry. In: Bedford T, Swift J (eds) New directions in dynamical systems. Cambridge University Press, Cambridge
176. Tondl A, Ruijgrok T, Verhulst F, Nabergoj R (2000) Autoparametric resonance in mechanical systems. Cambridge University Press, Cambridge,
177. Ushiki S (1984) Normal forms for singulatrities of vector fields. Jpn J Appl Math 1:1–34
178. Vanderbauwhede A (1982) Local bifurcation and symmetry. Pitman, Boston
179. Verhulst F (1989) Nonlinear differential equations and dynamical systems. Springer, Berlin; (1996)
180. Verhulst F (1998) Symmetry and integrability in Hamiltonian normal form. In: Bambusi D, Gaeta G (eds) Symmetry and perturbation theory. CNR, Roma
181. Verhulst F (1999) On averaging methods for partial differential equations. In: Degasperis A, Gaeta G (eds) Symmetry and perturbation theory II. World Scientific, Singapore
182. Vinogradov AM (1984) Local symmetries and conservation laws. Acta Appl Math 2:21–78
183. Vittal V, Kliemann W, Ni YX, Chapman DG, Silk AD, Sobajic DJ (1998) Determination of generator groupings for an islanding scheme in the Manitoba hydro system using the method of normal forms. IEEE Trans Power Syst 13:1346–1351
184. Vorob'ev EM (1986) Partial symmetries of systems of differential equations. Soviet Math Dokl 33:408–411
185. Vorob'ev EM (1991) Reduction and quotient equations for differential equations with symmetries. Acta Appl Math 23:1–24
186. Walcher S (1991) On differential equations in normal form. Math Ann 291:293–314
187. Walcher S (1993) On transformation into normal form. J Math Anal Appl 180:617–632
188. Walcher S (1999) Orbital symmetries of first order ODEs. In: Degasperis A, Gaeta G (eds) Symmetry and perturbation theory II. World Scientific, Singapore
189. Walcher S (2000) On convergent normal form transformations in the presence of symmetry. J Math Anal Appl 244:17–26
190. Wei J, Norman E (1963) Lie algebraic solution of linear differential equations. J Math Phys 4:575–581
191. Winternitz P (1987) What is new in the study of differential equations by group theoretical methods? In: Gilmore R (ed) Group Theoretical Methods in Physics proceedings of the XV ICGTMP. World Scientific, Singapore
192. Winternitz P (1993) Lie groups and solutions of nonlinear PDEs. In: Ibort LA, Rodriguez MA (eds) Integrable systems, quantum groups, and quantum field theory NATO ASI 9009. Kluwer, Dordrecht

Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence

Tom H. Solomon
Department of Physics & Astronomy,
Bucknell University, Lewisburg, USA

Article Outline

Glossary
Definition of the Subject
Introduction and Theoretical Background
Specific Examples of Patterns in Fluid Flows
Transport and Mixing
Pattern Formation in Reaction-Diffusion and Advection-Reaction-Diffusion Systems
Other Examples of Pattern Forming Systems
Future Directions
Bibliography

Glossary

Bifurcation A change from one kind of behavior to another; e. g., a bifurcation from steady flow to time periodic flow.

Defect A location where a pattern abruptly changes; e. g., where a line ends or two lines meet at a point.

Deterministic chaos Erratic behavior arising from a low-dimensional, deterministic system; behavior that shows sensitive dependence on initial conditions where a small error in initial conditions grows roughly exponentially in time.

Free energy functional An effective energy that depends on the function describing a pattern.

Fully-developed turbulence Turbulent flow for very large Reynolds number with no spatial or temporal order.

Grain boundary Line or curve in a pattern where two different zones meet.

Kolmogorov–Arnold–Moser (KAM) invariant surface Curve or surface that separates region of ordered trajectories from region of chaotic trajectories. Tracers cannot cross between these two regions; consequently, a KAM invariant surface acts as a transport barrier.

Laminar flow Smooth, well-ordered flow, achieved with small Reynolds number.

Lévy flights Trajectories characterized by jumps between distant parts of the system. If it is a Lévy flight, the jumps in a trajectory have a probability distribution $P(L) \sim L^{-\mu}$, with $2 < \mu < 3$.

Linear stability theory Approach used to determine critical parameters (e. g., Re or Ra) at which a flow becomes unstable. Linear stability theory also determines the wavenumbers of the disturbances that are unstable.

Poincaré section A stroboscopic plot for a time-periodic flow showing the positions of one or more tracers once each period of the flow.

Rayleigh–Bénard convection A flow that develops from an instability due to unstable stratification when a fluid is heated from below.

Rayleigh number Dimensionless parameter that characterizes the magnitude of the temperature difference in a Rayleigh–Bénard flow. The Rayleigh number can also be thought of as a ratio of thermal buoyancy to diffusive damping.

Reaction-diffusion system A spatially-extended system that produces patterns or fronts as a result of the interaction between some sort of reaction (e. g., chemical reaction, biological interaction, phase transition) and molecular diffusion.

Reynolds number Dimensionless parameter that characterizes the strength of a flow. The Reynolds number can also be thought of as the ratio between fluid inertia and viscous damping.

Stratification Density variations in a fluid system.

Strong turbulence High Reynolds number turbulence.

Subdiffusion Transport where the variance σ^2 grows more slowly than normal diffusion, e. g., $\sigma^2(t) \sim t^\gamma$, with $\gamma < 1$. Associated with regions in the flow where the tracers temporarily stick.

Superdiffusion Transport where the variance σ^2 grows more rapidly than normal diffusion, e. g., $\sigma^2(t) \sim t^\gamma$, with $\gamma > 1$. Associated with Lévy flight trajectories.

Taylor–Couette flow Flow between two coaxial cylinders, the inner of which (at least) is rotating.

Weak (defect-mediated) turbulence Flows with intermediate values of the Reynolds number that still produce short-range patterns but evolve both spatially and temporally in a complicated manner. Can be characterized by the presence of numerous defects in the pattern (more defects for more turbulent flows) that move around the system.

Definition of the Subject

"Nonlinear dynamics" is the study of systems described by equations of motion that depend nonlinearly on their variables. This is of significant scientific and mathematical interest because nonlinear systems – unlike simpler

linear ones – typically show quite complicated behavior, including deterministic chaos, spontaneous formation of patterns, spatio-temporal complexity and turbulence. The equations describing fluid flows are nonlinear; consequently, there is a wide variety of types of flows that can be achieved, spanning a wide range of levels of complexity from simple, well-ordered, *laminar* flows up through turbulent flows with significant complexity both in space and time. The equations that describe the mixing of impurities in a flow are also often nonlinear (depending on the fluid flow); consequently, mixing can also be very complicated and even *chaotic*.

Introduction and Theoretical Background

The dynamics of a fluid flow are governed by the Navier–Stokes equation:

$$\frac{\partial \vec{u}}{\partial t} + \vec{u} \cdot \vec{\nabla} \vec{u} = -\frac{1}{\rho} \vec{\nabla} p + \nu \nabla^2 \vec{u} + \frac{1}{\rho} \vec{F} \,.$$

This is the fluid continuum version of Newton's second law. The first term on the left represents the change in momentum of a small parcel of a fluid. Several factors can result in a change in momentum: forces acting on the fluid element – pressure differences, viscosity and external forces (first, second and third terms on the right), and advection of higher or lower velocity fluid elements into the region of interest (the second term on the left). The Navier–Stokes equation can be non-dimensionalized by scaling the velocity by a characteristic velocity scale U, lengths by a typical length scale L, and time by an advective time scale L/U:

$$\frac{\partial \vec{u}'}{\partial t} + \vec{u}' \cdot \vec{\nabla}' \vec{u}' = -\vec{\nabla}(\Delta p)' + \frac{1}{\mathrm{Re}} \nabla'^2 \vec{u}' \,,$$

where we have neglected external forcing and where $\vec{u}' \equiv u/U$ is the non-dimensional velocity. The non-dimensional parameter $\mathrm{Re} = UL/\nu$ is the Reynolds number. The Reynolds number characterizes the ratio of inertia to viscosity in a fluid flow.

The advection term on the left is often referred to as the *inertial* term, and is the source of nonlinearity in the Navier–Stokes equation. Nonlinearity goes hand-in-hand with complexity; consequently, the Reynolds number is the fundamental parameter used to describe the transition in fluid flows from simple, laminar (smooth), well-ordered flows to more and more complicated flows. If $\mathrm{Re} \ll 1$, the inertial term is negligible and for steady flows, viscosity at every location is balanced by pressure gradients. In this limit, the Navier–Stokes equations can be rewritten (in dimensional form) as:

$$\frac{1}{\rho} \vec{\nabla} p = \nu \nabla^2 \vec{u} \,.$$

Flows in this limit are called *viscous*, *creeping* or *Stokes* flows. If $\mathrm{Re} \gg 1$, viscosity is negligible and the NS Equation can be written as:

$$\frac{\partial \vec{u}}{\partial t} + \vec{u} \cdot \vec{\nabla} \vec{u} = -\frac{1}{\rho} \vec{\nabla} p \,.$$

Flows in this limit are called *inviscid* or *Euler* flows and are dominated by inertial effects. Flows with very large Re are typically turbulent.

There has been significant research into the transition between well-ordered, laminar, viscous flows for $\mathrm{Re} \ll 1$ and fully-developed turbulent flows for $\mathrm{Re} \gg 1$. Landau [1] predicted that an increase in Re would be accompanied by a infinite series of bifurcations as more and more frequencies of fluid oscillation (and corresponding spatial complexity) are added to the flow. According to Landau, measurement of the flow velocity at a point will reveal a steady, time-independent signal for weak flows (low Re), then for stronger flows, the velocity will oscillate periodically at that point, then at higher Re, the velocity time series at that point will be a superposition of two periodic oscillations with different frequencies. As Re increases, the velocity time series will be a superposition of more and more sinusoidal functions with different frequencies. Ultimately, according to Landau, a turbulent flow is simply one with so many superposed frequencies that it *seems* complicated and unpredictable.

Landau's prediction was tested in the early 1970s by Harry Swinney and Jerry Gollub in a *Taylor–Couette* flow [2] (see Subsect. "Taylor–Couette Flow" for more detail about Taylor–Couette flow). However, they did not find an infinite sequence of transitions with more and more frequencies (and more and more complexity). Instead, they found a sudden transition from an ordered flow with two superposed frequencies directly to broad spectrum, aperiodic time dependence, a state that we now know to be *chaotic*. This direct transition in fluid flows to chaotic time dependence was also predicted theoretically by Ruelle and Takens in 1971 [3].

Deterministic chaos is denoted by two main features: (1) complicated, aperiodic time dependence from a simple, deterministic system with a small number of degrees of freedom; and (2) sensitive dependence on initial conditions, whereby small uncertainties in measurements of the initial conditions grow roughly exponentially in time, destroying long-term predictability. Despite the name, however, chaotic systems are usually quite well-ordered spa-

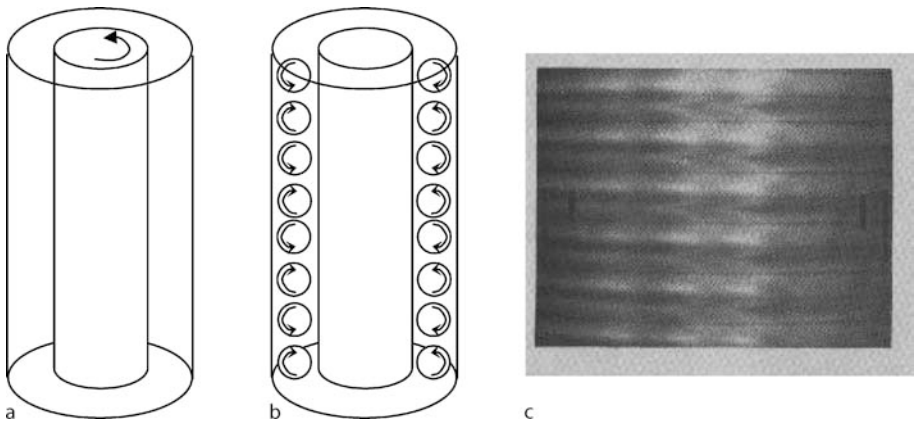

Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 1
Taylor–Couette flow. a Diagram of configuration. The fluid is contained in the gap between two coaxial cylinders, the inner of which (at least) rotates, driving a flow between the two cylinders. b Cross-section sketch of Taylor–Couette vortices. c Experimental image showing Taylor–Couette vortices; flow is visualized using Kalliroscope. (From [6])

tially; it is the time-dependence that is complicated. Further increases in Re are needed to break up the spatial patterns that persist for chaotic flows. A regime displaying disorder in both space and time is referred to as *spatio-temporal chaos* or *weak turbulence*. Weak turbulent regimes are still characterized by well-defined structures (e. g., vortices), but the structures are arranged in a disordered pattern. These structures break up if Re is increased beyond that for the weak turbulent regime; ultimately, for large enough Re, the flow is described as *fully-developed turbulence* with a broad spectrum of spatial length scales. Long-lived vortex structures may survive for surprisingly large Re, however, especially in planetary and atmospheric flows in rotating systems [4] or in plasma flows with large uniform magnetic fields.

In this article, we present several different examples of pattern formation, mixing and chaos in fluid systems. Section "Specific Examples of Patterns in Fluid Flows" presents specific examples of the kinds of transitions discussed above in fluid flows, showing the patterns that accompany these transitions. We emphasize Taylor–Couette and convective flows, mainly due to their historical importance in this field. In Sect. "Transport and Mixing", we discuss transport and mixing, highlighting in particular the relatively recent discovery that fluid mixing in *laminar*, well-ordered flows is often chaotic. In Sect. "Pattern Formation in Reaction-Diffusion and Advection-Reaction-Diffusion Systems", we discuss chemical pattern formation in both stagnant and flowing systems. Section"Other Examples of Pattern Forming Systems" contains a brief listing of some other types of pattern-forming systems. We finish in Sect. "Future Directions" with a brief overview of future directions in the field.

Specific Examples of Patterns in Fluid Flows

There are a number of different flows whose spatial patterns have been studied. Similar patterns are found in a very wide range of flows.

Taylor–Couette Flow

Taylor–Couette flow is the flow between two coaxial cylinders (Fig. 1a). In its simplest form, the outer cylinder remains stationary while the inner cylinder rotates. No-slip boundary conditions requires fluid at the outer edge to remain stationary while fluid at the inner edge rotates with the inner cylinder. The Reynolds number for this flow is defined using the gap width L as the length scale and the maximum flow velocity U as the velocity scale: $\mathrm{Re} = UL/\nu$. Taylor–Couette flows are often visualized with Kalliroscope [5], a suspension of aluminum flakes that align with shear zones in the flow. Reflection of light off the flakes depends on their orientation; consequently, kalliroscope quite effectively visualizes the flow field. This technique was originally developed by the artist Paul Matisse, but has been adopted by experimentalists who study fluid instabilities.

For small Re in the Taylor–Couette system, flow between the cylinders is laminar and purely in the azimuthal direction, with the velocity magnitude dropping smoothly from the inner cylinder velocity to zero at the outer cylinder. Beyond a critical Re this smooth flow becomes unstable due to centrifugal effects – the faster moving fluid near the inner cylinder moves outward while the slower moving fluid near the outside moves inward. The result is a pattern of stacked tori (donut-shaped structures) called *Taylor vortices* [6] (Fig. 1b and c).

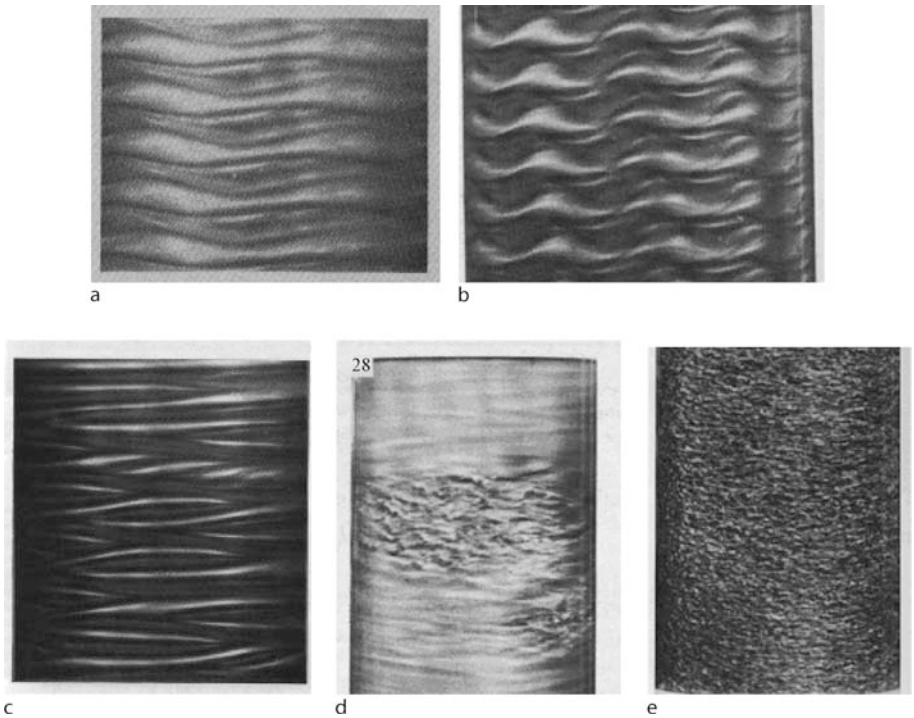

Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 2
Some examples of Taylor–Couette patterns; all visualized with Kalliroscope flakes. a Wavy-vortex flow (from [6]); b Modulated wavy vortex flow (from [7]); c Braided vortex flow (from [8]); d Intermittent turbulent spots (from [7]); and e Turbulent flow (from [7])

A fluid element within a Taylor vortex still moves azimuthally around the cylinders, but also follows a closed (vortex) path in r–z coordinates. A radial cross-section of the flow reveals a chain of counter-rotating vortices.

Further increases in Re result in a bifurcation to *wavy-vortex* flow [7], where the tori defining the Taylor vortices develop a snake-like undulation that propagates around the annulus (Fig. 2a). This spatial pattern corresponds to the time-periodic state discussed in transition to chaos in Sect. "Introduction and Theoretical Background". Further increases in Re result in transitions to quasi-periodic and chaotic states. Ultimately, for large enough Re, the flow is turbulent (Fig. 2e).

If the outer cylinder rotates as well, a wide variety of spatial patterns are found [7,8]. Some of these states are shown in Fig. 2. Taylor–Couette flows have been studied extensively because of the wide variety of different states, including states in which the flow experiences intermittent (momentary) bursts of turbulence. Various investigators have mapped out *parameter space diagrams* that show the conditions needed to obtain various different states in Taylor–Couette flow. An example of such a parameter space diagram [7] is shown in Fig. 3.

Rayleigh–Bénard Convection

When a fluid is heated from below or from the side, the warmer fluid near the bottom expands and becomes less dense than the cooler fluid above. This results in unstable stratification with heavier, more dense fluid above lighter, less dense fluid. If the stratification is large enough, the denser fluid falls and the lighter fluid rises, producing a *convective* flow. Thermal convection of this nature is quite common: it occurs when heating up water or soup on a stove; it is a dominant mechanism by which a heating element heats up the air in an oven or by which heat from a radiator warms up a room; it is the dominant mechanism by which atmospheric flows are driven; it is the mechanism by which flows within the Earth's crust are formed; and it is a major driving force for flows in stars. In all cases, thermal convection results in *significantly* faster and more efficient transport of heat than would be achieved solely with thermal diffusion.

In its simplest form, thermal convection ensues if a large enough vertical temperature gradient is applied with the bottom warmer than the top (Fig. 4). The relevant dimensionless number characterizing the magnitude of the temperature difference is the *Rayleigh Number*

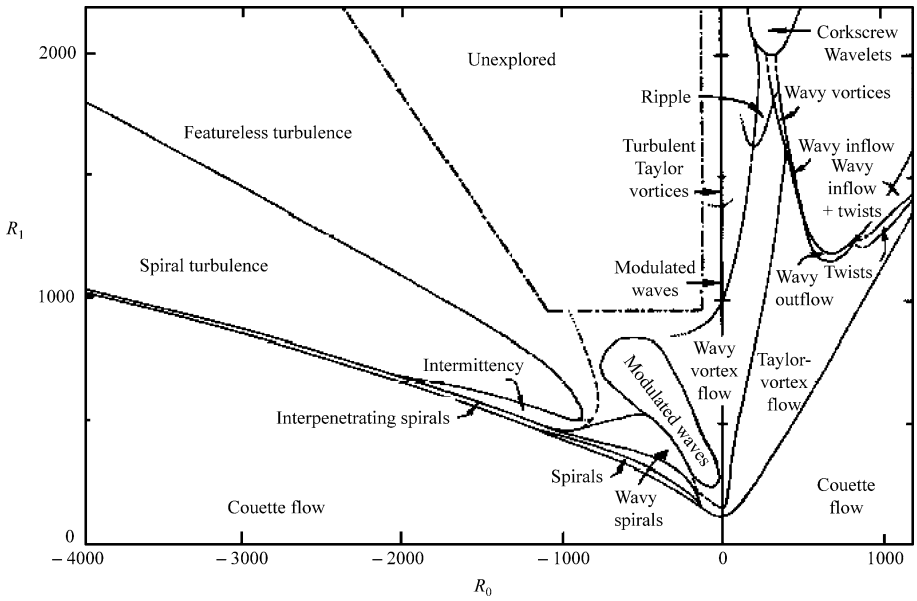

Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 3
Parameter-space diagram showing the many different regimes of pattern formation for Taylor–Couette flow between two coaxial cylinders. (From [7])

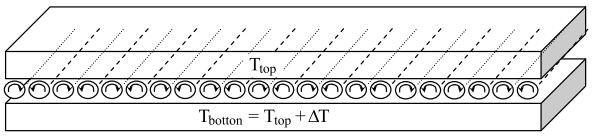

Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 4
Schematic for Rayleigh–Bénard convection. A thin layer of fluid is sandwiched between two thermally-conducting plates, with the bottom plate hotter than the top plate. For sufficiently large ΔT, convection ensues with hotter, light fluid rising from the bottom and colder, denser fluid falling from the top. In its simplest form, the convection pattern is a pattern of counter-rotating "rolls" as shown in the sketch. The *dotted* (*dashed*) lines denote "downwelling" (upwelling) regions in the flow

$\mathrm{Ra} = g\alpha\Delta T/\nu\kappa$, where $\Delta T = T_{\mathrm{bottom}} - T_{\mathrm{top}}$ is the temperature gradient, α is the coefficient describing the volume expansion of the fluid in response to a change in temperature, g is the gravitational acceleration and κ is the thermal diffusivity. Conceptually, increases in the temperature difference, gravitational acceleration, or expansion coefficient make the fluid more unstable, whereas viscosity inhibits fluid motion, and thermal diffusivity tends to diminish temperature differences.

If Ra is small, the fluid remains motionless – viscosity and thermal diffusion are sufficient to keep the fluid from becoming unstable. If Ra exceed a critical value Ra_c, the motionless state becomes unstable and the warmer, less-dense fluid near the bottom rises while the cooler, more-dense fluid near the top sinks. In a thin, rectangular box a pattern of *convection rolls* develops, with fluid circulating like rotating cigars in a box, as shown schematically in Fig. 4.

Numerous studies have been made of patterns made by Rayleigh–Benard convection. In a manner similar to that in Taylor–Couette flow, there is a series of bifurcations from well-ordered to chaotic and ultimately to turbulent flow in Rayleigh–Benard convection. The nature of the transitions depends not only on the Rayleigh number Ra but also on a second dimensionless parameter called the *Prandtl number* $\mathrm{Pr} = \nu/\kappa$ which denotes the relative magnitudes of the viscous and thermal diffusivities and is a property of the fluid used in the experiments. Large Prandtl numbers are achieved in fluids with large viscosities; even water at room temperature has a Prandtl number of around 6. Pressurized gasses have Prandtl numbers of around 1. Small Prandtl numbers are achieved in flows of liquid metals (e. g., mercury or sodium) that are very good thermal conductors and flows in gasses which have a very small kinematic viscosity.

For fluids with small Pr, the first instability from steady (time-independent) convection rolls is an oscillatory instability, similar in many respects to the wavy vortex state for Taylor–Couette flow. The convection rolls form a snake-like instability along their axes, and this undulation in the rolls propagates in a direction parallel to the axes of the

convection rolls. A cross-section of the flow in this state would reveal a pattern of counter-rotating vortices that oscillates periodically in the lateral direction.

There are also several other instabilities that appear in Rayleigh–Bénard convection including: cross-roll instabilities in which a second pattern of convection rolls develops perpendicular to the first; a skew-varicose instability in which the width of convection rolls is modulated; a zig-zag instability which is exactly what its name implies, one in which the convection rolls form a zig-zag pattern; an instability referred to as the "Eckhaus instability" where pairs of convection rolls disappear, allowing other rolls present to grow, increasing the wavelength of the pattern; and spiral convection patterns. Defects can also appear in convection patterns where either two or more convection rolls meet at a point. The motion of a defect parallel to convection rolls can result in the addition or removal of a pair of convection rolls in the flow.

One way of understanding this behavior is to consider the stability of convection rolls near onset as a function of their wavelength. A common way of analyzing stability is via a "linear stability" argument. One starts with a motionless state, and then assumes a perturbation of a periodic pattern of convection rolls on this motionless state. A growth/decay term is associated with this perturbation, which might be written in the form $A_o e^{\sigma t} f(k)$ where $f(k)$ is the perturbation with wavenumber k, and σ is the growth exponent. The growth exponent σ is typically dependent on the value of k (Fig. 5). If $\sigma < 0$ for all values of k, then all perturbations decay and the motionless state is stable and no convection rolls will develop; this is the case for $\mathrm{Ra} < \mathrm{Ra_c}$. For $\mathrm{Ra} = \mathrm{Ra_c}$, $\sigma = 0$ for one particular critical wavenumber k_c; for Ra just slightly above $\mathrm{Ra_c}$, perturbations with wavenumber k_c should be expected to grow and, for an infinite system (i. e., no boundaries), a pattern of convection rolls will appear with precisely that wavenumber. For $\mathrm{Ra} > \mathrm{Ra_c}$, there is a range of wavenumbers for which $\sigma > 0$.

The discussion above is idealistic; in realistic systems, boundary conditions have a significant effect on convection patterns. If the fluid layer is confined to a small-to-moderate rectangular box, then the rolls line up parallel to the shortest side of the box, as in Fig. 4. If the width of the box is an integer multiple of the critical wavelength $2\pi/k_c$, then linear stability correctly predicts convection with this wavelength for Ra just barely above $\mathrm{Ra_c}$. If the fluid layer is orders of magnitude wider than the critical wavelength, multiple zones of parallel convection rolls form at random orientation (Fig. 6a); the zones are separated by *grain boundaries*. Over time, some zones grow while others shrink. Ultimately, one zone may end up dominating, with a single pattern of parallel convection rolls covering the entire system.

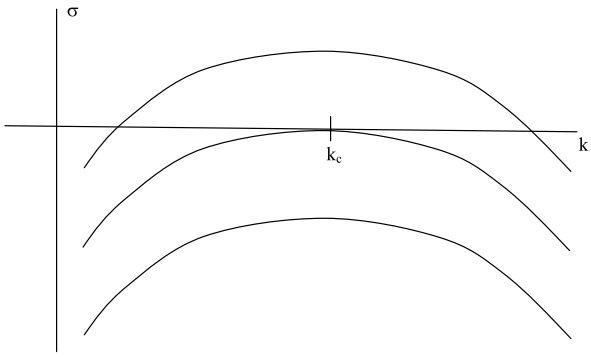

Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 5
Example of a plot of growth exponent σ as a function of wavenumber k, typical of results of linear stability analysis. The *bottom curve* corresponds to a Reynolds or Rayleigh number below the critical value for instability; σ is negative for all wavenumbers, so all disturbances decay. The *middle curve* corresponds to the onset of instability. For Ra or Re just above this value, σ is positive for wavenumbers very close to the critical value k_c. For larger Re or Ra (*top curve*), disturbances over a range wavenumbers are unstable

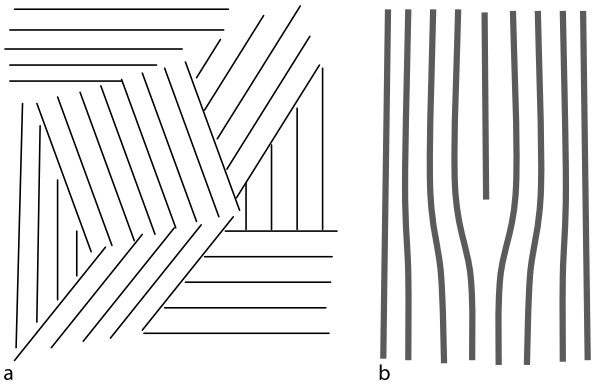

Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 6
a Sketch of zones of parallel convection rolls, separated by grain boundaries. The *lines* represent boundaries between adjacent rolls, as viewed from above (similar to dashed and dotted lines in Fig. 4). b Sketch of a defect in a convection pattern

If contained in a circular container, a pattern of parallel convection rolls cannot be achieved [9], because convection rolls tend to line up perpendicular to the side walls. A theory by Swift and Hohenberg [10] defines a free-energy functional – an effective energy that depends on the functional form of the convection pattern; the pattern evolves to minimize this free energy. The free energy is

increased by convection rolls that are parallel to the side walls, by bends in convection rolls, and by defects and grain boundaries in the patterns. Near onset (i. e., for Ra just slightly over Ra_c), a complicated convection pattern forms to minimize this free energy functional (Fig. 7a). The patterns that form have local wavenumbers that can not all be at the critical value k_c; consequently, the pattern may continually evolve in time, in contradiction to the typical case in which convection with Ra just above Ra_c is time-independent.

For large enough Ra, convective flows become turbulent. Two types of turbulent flows are often discussed. At lower (but still large) Ra, complicated convection patterns form with numerous defects. The defects move around significantly in the pattern, resulting in continuous variation in the pattern. This regime is referred to as *weak turbulence* or *defect-mediated turbulence*. For very high Ra, the convection pattern breaks up completely and a regime of *strong turbulence* is found. A side view of convection in the strongly turbulent regime (Fig. 7b) shows that the flow is divided into three regions. Most of the temperature gradient is concentrated into thin boundary layers at the top and bottom of the fluid, while the fluid is well-mixed at roughly the average temperature in the large central region of the flow. The boundary layers are not quiescent; rather, they grow in thickness as heat flows into (for the lower boundary layer) and out of (for the upper boundary layer) them, eventually becoming unstable and forming plumes that erupt, carrying blobs of warm or cold fluid into the center region. The eruption of these plumes both acts to drive the turbulent flow and also transport heat from the bottom to the top of the system.

Bénard–Marangoni Convection

If a fluid layer with a free surface is heated from below, thermal convection can be driven by variations in surface tension at the surface. This form of convection is referred to as Bénard–Marangoni convection and is usually characterized by a pattern of hexagons. Since the fluid is heated from below, it is unstably stratified. However, the main force driving the flow isn't buoyancy; rather, the flow is driven by gradients in the surface tension at the free surface. An instability occurs in which small variations in the temperature arise on the surface. The surface tension depends on the temperature; consequently, temperature gradients result in surface tension gradients that drive flows along the surface. Regions in which the fluid converges on the surface well downward, and upwelling of fluid occurs below regions in which the fluid diverges on the surface.

Transport and Mixing

The manner in which an impurity is mixed within a fluid is an issue with significant practical importance for a wide range of scientific and engineering applications. Mixing of a *passive* impurity (one that follows fluid elements in the flow and whose distribution does not affect the flow itself) is governed by the advection-diffusion equation:

$$\frac{\partial c}{\partial t} = -\vec{u} \cdot \vec{\nabla} c + D \nabla^2 c \,,$$

where c is a scalar concentration field, $\vec{u}$ is the velocity field, and D is the molecular diffusion coefficient. This equation can be written in non-dimensional form by scaling the velocity field by a characteristic flow velocity U, distance by a characteristic length L, and time by the advection timescale L/U:

$$\frac{\partial c}{\partial t'} = -\vec{u}' \cdot \vec{\nabla}' c + \frac{1}{\mathrm{Pe}} \nabla'^2 c \,,$$

where the Peclet number $\mathrm{Pe} = UL/D$ characterizes the relative strength of advective to diffusive mixing in the flow. In the limit of no flow ($\mathrm{Pe} \to 0$), mixing is entirely via molecular diffusion, caused by Brownian motion of individual tracer particles in the impurity. Molecular diffusion is a slow mixing process; the variance $\sigma^2(t)$ of a distribution grows linearly with time: $\sigma^2(t) = 2Dt$.

The presence of a fluid flow dramatically changes the mixing. An impurity is advected along with fluid elements in the flow, but can diffuse away from these deterministic fluid trajectories. Advection of an individual tracer is determined from equations of motion derived from the velocity field $\vec{u} = (u_x, u_y, u_z)$:

$$\begin{aligned} \mathrm{d}x/\mathrm{d}t &= u_x \\ \mathrm{d}y/\mathrm{d}t &= u_y \\ \mathrm{d}z/\mathrm{d}t &= u_z \,. \end{aligned}$$

Trajectories are determined by integrating these equations of motion. If the flow is two-dimensional and time-independent ($\vec{u} = \vec{u}(x, y)$ and $u_z = 0$), then tracer trajectories are *integrable* (well-ordered). For two-dimensional, time-periodic or three-dimensional, time-independent flows, however, the equations of motion in general are not integrable, and trajectories are often chaotic, with nearby tracers separating roughly exponentially in time (sensitive dependence on initial conditions).

The most well-known flow exhibiting *chaotic advection* is the blinking vortex flow proposed by Hassan Aref [11]. The flow is shown in Fig. 8a; the fluid alternates periodically between circling around the left vortex

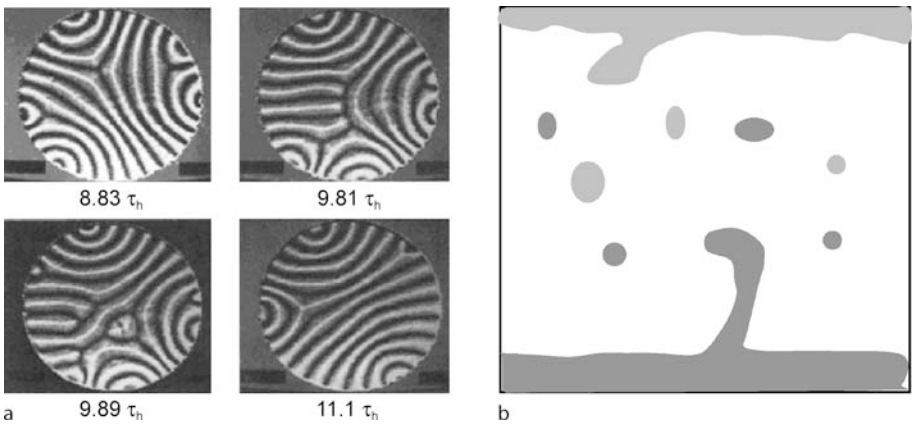

Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 7
a Images of Rayleigh–Bénard convection in a circular container near onset, viewed from above (from [9]). b Sketch of strongly-turbulent convection from side, showing the formation of thermal boundary layers at the bottom and top that erupt, sending plumes of hot and cold fluid into the middle region

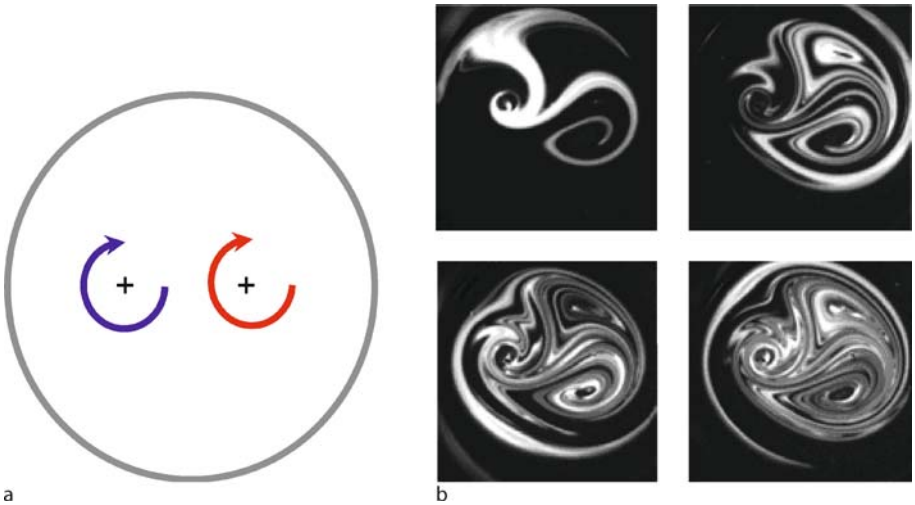

Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 8
a Sketch of blinking vortex flow. Fluid alternates between circling around left and right vortex centers. b Mixing of dye in blinking vortex flow. Images are separated in time by one period of oscillation each (from [12])

and around the right vortex. Advection of a blob of a passive dye is shown in Fig. 8b [12]. Mixing is associated with a repeated process of stretching and folding of tendrils of the impurity. Stretching and folding behavior such as this is typical of chaotic mixing.

Transport on scales larger than typical length scales of the flow can be illustrated with the alternating vortex chain [13], shown in Fig. 9. If the (two-dimensional) flow is time-independent, tracers follow ordered, closed trajectories within individual vortices. The presence of molecular diffusion allows for tracers to diffuse from one vortex to the next. The combination of advection around and diffusion between the vortices results in enhanced transport whose variance grows linearly in time: $\sigma^2(t) = 2D^*t$, where D^* is the *enhanced diffusion coefficient*.

The simplest time dependence for the alternating vortex chain is a simple lateral oscillation of the entire chain, similar to time dependence found in Taylor–Couette and Rayleigh–Benard flows for intermediate values of Re and Ra. A passive tracer moving in this flow circles around an individual vortex, but if it is near a separatrix (boundary) between vortices, it can cross between vortices. Over time, the tracer alternates between circling within and crossing between vortices in an erratic manner. The trajectories are rigorously chaotic in the sense that nearby trajectories separate roughly exponential as a function of time.

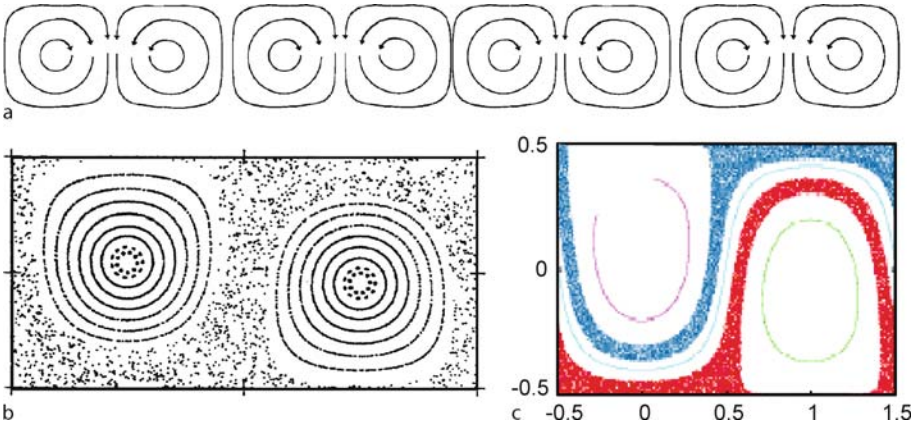

Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 9
a Sketch of alternating vortex chain. b Poincaré section for oscillating vortex chain, showing ordered regions of transport in vortex cores and chaotic region around and between vortices. c Poincaré section for oscillating and drifting vortex chain (from [14])

Not all tracers in the flow follow chaotic trajectories. Tracers near the center of a vortex remain trapped within that vortex, following a trajectory that is well-ordered and *not* sensitive to initial conditions. The separation of the system into both ordered and chaotic regions of transport can be seen easily by plotting a *Poincaré section* (Fig. 9b), a plot that shows the location of one tracer (or more) once every period of the flow. A tracer that starts in the chaotic region will eventually explore the entire chaotic region, which shows up on the Poincaré section as a pattern of dots. A chaotic trajectory can never cross into a region with ordered trajectories; consequently, the ordered regions show up as white "islands" on a Poincaré section unless tracers start within the ordered regions, revealing a pattern of closed curves, as seen in Fig 9a.

The boundaries between ordered and chaotic regions – referred to as *Kolmogorov–Arnold–Moser (KAM) invariant surfaces* – act as barriers against transport and mixing in the flow. KAM barriers are very common in laminar flows and also show up in many flows found in nature, even turbulent ones. Examples of transport barriers include the ozone hole over Antarctica, Gulf Stream rings in the Atlantic Ocean, and coherent vortices (called *meddies*) originating in the Mediterranean Sea that cross the Atlantic and maintain the salinity and biological constituency found in the Mediterranean for months and even years. Another well-known example of a transport barrier is the Great Red Spot of Jupiter, a large coherent patch of vorticity in a very turbulent atmosphere that clearly acts as a transport barrier, evidenced by the fact that it has maintained its distinct color for centuries.

Chaotic advection results in significant enhancements in long-range transport. In many cases, enhanced transport is diffusive with a variance that grows linearly in time. However, *anomalous diffusion* is also possible where the variance grows as a power law in time with a growth exponent different from 1: $\sigma^2(t) \sim t^\gamma$ with $\gamma \neq 1$. Transport with $\gamma < 1$ is called *subdiffusion* and transport with $\gamma > 1$ is called *superdiffusion*. The presence of islands of ordered trajectories is often associated with subdiffusive transport; tracers in the chaotic region stick to the outside of these islands and possibly diffuse inward due to Brownian motion. Superdiffusion is associated with *Lévy flight* trajectories where tracers jump large distances between regions in the flow. An example of a flow with superdiffusion is the alternating vortex chain with both oscillatory and drifting time dependence [14] (Fig. 9a). A Poincaré section for transport in this flow is shown in Fig. 9c. Two distinct chaotic regions are apparent, and in addition to ordered islands in the vortices, there is also a snake-like region of ordered trajectories that wind around the vortices. A tracer in the chaotic region in this flow alternately sticks to islands (remaining temporarily confined to a vortex) and to the snake region (traveling rapidly along the vortex chain).

Chaotic advection is also possible in three-dimensional flows, even those that are time-independent. An example of a such a flow is shown in Fig. 10 – a chain of alternating vortices superposed with a second chain which is shifted by half of a vortex width and is oriented with its vortex axes perpendicular to those of the first chain [15]. Trajectories of tracers in this flow are shown in Fig. 10b. A Poincaré section (Fig. 10c) again shows coexistence of both ordered and chaotic regions of transport in the flow.

There are numerous applications of chaotic mixing. Chemical engineers have already begun to develop and market devices that take advantage of chaos to enhance mixing in low-Reynolds number flows. Chaotic mixing

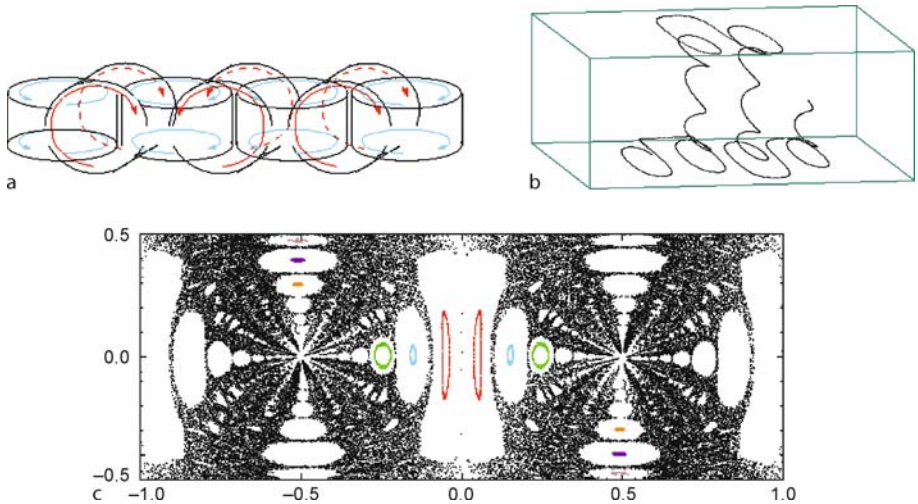

Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 10
a Sketch of time-independent, nested vortex chain flow. b Example of a chaotic trajectory for the nested vortex flow. c Poincaré section at the mid-height of the flow, showing both ordered and chaotic regions (from [15])

is particularly useful for mixing of high-viscosity fluids for which turbulent mixing would be difficult and/or tremendously expensive energetically. Chaotic mixing is also finding significant applications in the expanding field of microfluidic devices, so-called "factories-on-a-chip," again which involve very small Reynolds numbers due to the small length-scales of these systems.

Pattern Formation in Reaction-Diffusion and Advection-Reaction-Diffusion Systems

Another class of nonlinear, pattern-forming systems is the well-known reaction-diffusion (RD) problem. The reaction is a process in which one or more species somehow changes into something different. Examples include chemical reactions, biological interaction of some sort (such as the interaction of a disease with a population of people or animals or a predator-prey system), and a wide range of phase transitions. In an extended, non-flowing system, the reaction typically occurs at different times in different locations in the system, resulting in the formation of patterns. The key to the formation of these patterns is the interaction between the reaction at local regions in the system and molecular diffusion which (for a non-flowing system) is the only means of communication between different regions.

A general form of the RD equation can be written as follows:

$$\frac{\partial c}{\partial t} = f(c) + D\nabla^2 c\,,$$

where c is a concentration field of some chemical, biological or physical species, and the first term on the right denotes some sort of reaction. The details of the reaction term itself vary from system to system; however, some common behavior is found in a wide range of RD systems, independent of the details of this reaction term.

The field of reaction-diffusion studies was enhanced significantly by the discovery of the oscillatory *Belousov–Zhabotinsky* (BZ) chemical reaction. Discovered in the 1950s, this reaction is well-known to have a pH that oscillates for hours when well-mixed in a beaker or a flask. When catalyzed with a pH-indicator such as ferroin, the chemicals can be observed to alternate back and forth between a red and blue color during this period before the reaction finally runs down. The oscillation is typically nearly-periodic, but conditions have been found in which chaotic oscillations can be found in the BZ reaction.

If the chemicals for the BZ reaction are in a shallow petrie dish with no flow (or in a gel or fritted disk that inhibits a flow), spiral and/or target patterns form, as seen in Fig. 11. Patterns similar to these have been observed in a wide range of physical, chemical and biological systems. Examples include (a) spiral waves of electrical activity in the heart which act as pacemakers; breakdown of these spiral waves are associated with cardiac fibrillation; (b) spiral waves of "spreading depression" in the brain that are associated with migrane headaches; (c) spiral patterns found in developing embryos which may be partially responsible for morphogenesis in which different cells in the embryo develop different roles in the growing organism; and (d) spiral and target patterns in populations of slime molds

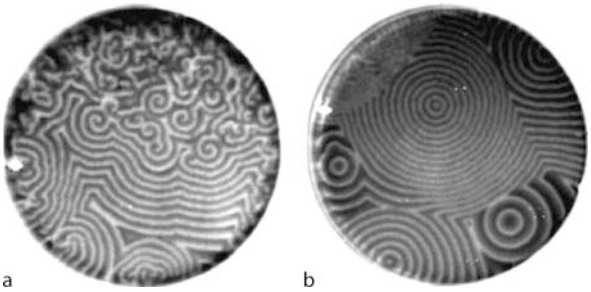

Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 11
Examples of reaction-diffusion patterns formed by the Belousov–Zhabotinsky chemical reaction in a thin layer in a petrie dish

in a stagnant system. Because of the similarity of these patterns to those found in the BZ system, there has been a significant amount of research into the properties of the BZ system, which is considered to be a paradigm of RD systems. Reaction-diffusion dynamics also explain the propagation of fronts in stagnant systems, e. g., the spreading of a fire in the absence of any winds or the growth of a solid in the absence of fluid flows.

In both natural and industrial systems, reacting systems are rarely stagnant; flows dramatically affect the mixing and therefore the communication between different parts of the system. The more general *advection-reaction-diffusion* (ARD) problem is represented by the following equation:

$$\frac{\partial c}{\partial t} = -\vec{u} \cdot \vec{\nabla} c + f(c) + D\nabla^2 c$$

where the additional term on the right denotes advection of the relevant species by the fluid flow. Advection-reaction-diffusion dynamics are relevant for pattern formation and front propagation in a wide range of systems, including marine ecological systems (e. g., algae and phytoplankton blooms), microfluidic chemical and biological processing and diagnostics, forest fires in the presence of winds, and the propagation of disease in a moving population.

If the mixing is chaotic, it has been shown [16] that ARD patterns reflect the structures that characterize chaotic mixing in the flow. An example [12] is the BZ reaction in a blinking vortex flow (Fig. 8) in which fluid circulates alternately (and periodically) between two separate vortex centers. As seen in Fig. 12, except for very weak mixing, the patterns associated with chaotic mixing end up dominating the pattern-formation process for the ARD system as well [12]. Ideas such as this have been used to explain patterns of populations of marine organisms in both the Gulf of Mexico and in the northern Atlantic Ocean.

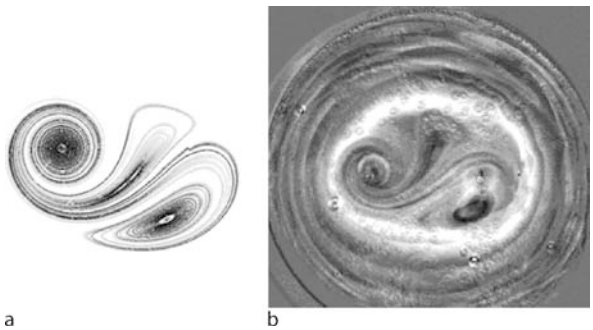

Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence, Figure 12
a Simulation of mixing field for blinking vortex flow. b Advection-reaction-diffusion pattern for Belousov–Zhabotinsky chemical reaction in the same flow (from [12])

Other Examples of Pattern Forming Systems

Patterns similar to those discussed above are found in a wide range of systems spanning all fields of science and engineering. Some examples are as follows:

- Electrohydrodynamic convection in flows of nematic liquid crystals.
- Wave patterns on the surface of a fluid oscillated periodically in the vertical direction.
- Vibrating granular systems.
- Bubble froths.
- Ferrofluids – labyrinth patterns.

This is just a small sample of the many nonlinear, pattern-forming systems that have been studied during the past few decades.

Future Directions

The study of nonlinear systems is an on-going and continually evolving field. Nonlinear dynamics span a wide range of fields of study, including all of the fields of sciences and engineering, as well as mathematics, medicine, economics and even some fields of social sciences. Not only are there numerous on-going studies of the basic mathematical and scientific behavior of nonlinear systems, but there are also many applications that are being developed, based on the principles of nonlinear systems.

Bibliography

Primary Literature

1. Landau L (1944) CR (Dokl) Acad Sci URSS 44:311
2. Gollub JP, Swinney HL (1975) Phys Rev Lett 35:927
3. Ruelle D, Takens F (1971) Commun Math Phys 20:167

4. Marcus PS (1988) Nature 331:693; Sommeria J, Meyers SD, Swinney HL (1988) Nature 331:689
5. Kalliroscope can be obtained from Kalliroscope Corporation. Groton, 978-448-6302. www.kalliroscope.com
6. Fenstermacher PR, Swinney HL, Gollub JP (1979) J Fluid Mech 94:103
7. Andereck CD, Liu SS, Swinney HL (1986) J Fluid Mech 164:155
8. Andereck CD, Dickman R, Swinney HL (1983) Phys Fluids 26:1395
9. Heutmaker MS, Fraenkel PN, Gollub JP (1985) Phys Rev Lett 54:1369
10. Swift J, Hohenberg PC (1977) Phys Rev A 15:319
11. Aref H (1984) J Fluid Mech 143:1
12. Nugent CR, Quarles WM, Solomon TH (2004) Phys Rev Lett 93:218301
13. Solomon TH, Gollub JP (1988) Phys Fluids 31:1372; Phys Rev A 38:6280; Solomon TH, Tomas S, Warner JL (1996) Phys Rev Lett 77:2682
14. Paoletti MS, Nugent CR, Solomon TH (2006) Phys Rev Lett 96:124101
15. Fogleman MA, Fawcett MJ, Solomon TH (2001) Phys Rev E 63:020101(R)
16. Tel T, de Moura A, Grebogi C, Karolyi G (2005) Chemical and biological activity in open flows: A dynamical system approach. Phys Rep 413:91

Books and Reviews

Baker GL, Gollub JP (1990) Chaotic dynamics: An introduction. Cambridge University Press, Cambridge
Ben-Avraham D, Havlin S (2000) Diffusion and reactions in fractals and disordered systems. Cambridge University Press, Cambridge
Cross MC, Hohenberg PC (1993) Pattern-Formation outside of equilibrium. Rev Mod Phys 65:851
Ott E (2002) Chaos in dynamical systems, 2nd edn. Cambridge University Press, Cambridge
Tritton DJ (1988) Physical fluid dynamics, 2nd edn. Clarendon Press, Oxford
Winfree AT (1980) The geometry of biological time. Springer, New York
Grindrod P (1996) The theory and applications of reaction-diffusion equations: patterns and waves. Clarendon Press, Oxford

Non-linear Internal Waves

Moustafa S. Abou-Dina, Mohamed A. Helal
Department of Mathematics, Faculty of Science,
Cairo University, Giza, Egypt

Article Outline

Glossary
Definition of the Subject
Introduction
Problem and Frame of Reference
Notation
Equations of Motion
The Shallow Water Theory
Free Surface and Interface Elevations of Different Modes
Secular Term
Multiple Scale Transformation of Variables
Derivation of the KdV Equation
Conclusions
Future Directions
Bibliography

Glossary

Nonlinear waves Nonlinear waves are such waves which arise as solutions of the nonlinear mathematical models simulating physical phenomena in fluids.

Shallow water Shallow water means water waves for which the ratio between the amplitude and the wave length is relatively small. The linear theory of motion is inadequate for the description of shallow water waves.

Internal waves Internal waves are gravity waves that oscillate within a fluid medium. They arise from perturbations to hydrostatic equilibrium, where balance is maintained between the force of gravity and the buoyant restoring force. A simple example is a wave propagating on the interface between two fluids of different densities, such as oil and water. Internal waves typically have much lower frequencies and higher amplitudes than surface gravity waves because the density differences (and therefore the restoring forces) within a fluid are usually much smaller than the density of the fluid itself.

Pycnocline Pycnocline is a rapid change in water density with depth. In freshwater environments, such as lakes, this density change is primarily caused by water temperature, while in seawater environments such as oceans the density change may be caused by changes in water temperature and/or salinity.

Solitary waves Solitary waves are localized traveling waves, which asymptotically tend to zero at large distances.

Solitons Solitons are waves which appear as a result of a balance between a weakly nonlinear convection and a linear dispersion. The solitons are localized highly stable waves that retain their identity (shape and speed) upon interaction, and resemble particle like behavior. In the case of a collision, solitons undergo a phase shift.

Baroclinic fluid Baroclinic fluid is such a fluid for which the density depends on both the temperature and the pressure.

In atmospheric terms, the baroclinic areas are generally found in the mid-latitude/polar regions.

Barotropic fluid Barotropic fluid is such a fluid for which the density depends only on the pressure.

In atmospheric terms, the barotropic zones of the Earth are generally found in the central latitudes, or tropics.

Definition of the Subject

The objective of the present work is to study the generation and propagation of nonlinear internal waves in the frame of the shallow water theory. These waves are generated inside a stratified fluid occupying a semi infinite channel of finite and constant depth by a wave maker situated in motion at the finite extremity of the channel. A distortion process is carried out to the variables and the nonlinear equations of the problem using a certain small parameter characterizing the motion of the wave maker and double series representations for the unknown functions is introduced. This procedure leads to a solution of the problem including a secular term, vanishing at the position of the wave maker. This inconvenient result is remedied using a multiple scale transformation of variables and it is shown that the free surface and the interface elevations satisfy the well known KdV equation. The initial conditions necessary for the solution of the KdV equations are obtained from the results of the first procedure.

Introduction

The work on internal gravity waves has been started in the middle of the twentieth century by Keulegang [21] followed by Long [25], where they applied Boussinesq's and Rayleigh's techniques, respectively, on free surface waves to the problem of internal waves.

In physical oceanography, fluid density varies with the depth due to the salinity and the temperature of the fluid [23]. Several works, dealing with variable density, have been worked out (e. g. [2,4,15,19,20,31,32], . . .).

The geophysical problem of the propagation waves in stratified fluid is very important in physical oceanography. The physical problem is simulated by a suitable model of a free-surface and interface fluid flow over a horizontal bottom or over a certain topography.

The fluid in the ocean may be considered as a stratified one according to the variation of the salinity and temperature, due to the weather, with respect to the vertical coordinate normal to the surface. The mathematical model of a two-fluid system (stratified fluid) is a good description to simulate wave motions in physical applications.

Based on results of many field observations, it is found that the fluid's density changes rapidly within the regions of pycnocline.

Nonlinear theoretical models for waves in a two-fluid system have been established with various restrictions on length scales. Among these, Benjamin [6] derived the KdV equation for thin layers in comparison with the wave length. Later, Benjamin [7], Davis and Acrivos [14] and Ono [30] constructed the BO equation under the assumptions of a thin upper layer and an infinitely deep lower layer. Kubota et al. [22] and Choi and Camassa [11,12] carried out a series of investigations to derive model equations for weakly and strongly nonlinear wave propagation in stratified fluids. The upper boundary is allowed to be either free or rigid. The assumptions they stated are as follows:

1) the wavelength of the interfacial wave is long.
2) the upper layer is thin.
3) no depth restriction is made on the lower layer.

The effect of a submerged obstacle on an incident wave inside an ideal two-laver stratified shallow water has been studied in two dimensions by Abou-Dina and Helal [3] by applying the shallow water approximation theory.

In 1995, Pinettes, Renouard and Germain presented a detailed analytical and experimental study of the oblique passing of a solitary wave over a shelf in a two-layer fluid.

Later, Barthelemy, Kabbaj and Germain [5] applied the WKBJ technique to investigate theoretically the scattering of a surface long wave by a step bottom in a two-layer fluid. This method enabled them to overcome the main inconsistency of the shallow-water theory in the presence of obstacles.

Matsuno [28] attempted to unify the KdV, BO and ILW models using the assumption of small wave steepness. Lynett and Liu [26] assumed a small density difference and derived a set of model equations.

Recently, Craig et al. [13] presented a Hamiltonian perturbation theory for the long-wave limit, and provided a uniform treatment of various long-wave models (KdV, BO and ILW models). Their formulation is shown to be very effective for perturbation calculations and represents a basis for numerical simulations. For a single-fluid system, in the last two decades, the traditional Boussinesq equations have been improved to extend their applicability from shallow water to deep water (see for example: Nwogu [29], Chen and Liu [10], Wei et al. [36], Madsen and Schaffer [27], Gobbi et al. [18]).

Most recently, Liu et al. [24] published an excellent study on the essential properties of Boussinesq equations for internal and surface waves in a two-fluid system.

Most studies applied the technique of the Padé approximate to allow a much higher-order accuracy without increasing the order of the derivatives.

Although the theoretical work for surface waves are excessive, studies of internal waves based on Boussinesq equations are still lack.

In the present work, and for a better description of the physical problem, we shall investigate, in the frame of the shallow water theory, the effect of the bounded motion of a vertical plane wave maker on the generation and propagation of waves inside a stratified fluid. The wave maker is situated at the finite end of a semi-infinite channel of constant depth occupied by two non immiscible fluid layers of different constant densities. The problem considered here, is supposed to be a two dimensional one with an irrotational motion, the fluid layers are taken ideal and the surface tension is neglected.

Double series representation for the unknown functions, in terms of a certain small parameter characterizing the motion of the wave maker is used. The first few orders of approximations obtained following this technique indicates the presence of a secular term increasing indefinitely in going away from the wave maker and vanishing at the position of this later. To overcome this physically unacceptable result, a multiple scale transformation of variables is carried out and a single series representation, for the unknown functions, is proposed. This technique leads to solutions up to the fifth order of approximations free from secular terms and shows that the generation and propagation processes of internal waves are governed by the well known KdV equation. The initial conditions needed for the solution of this nonlinear partial differential equation are obtained from the results of the preceding method at the position of the wave maker where the solution is free from secular terms. On the light of this result it is expected to obtain an internal solitonic waves.

Problem and Frame of Reference

A vertical plane wave-maker is situated at the finite extremity of a semi-infinite channel with a constant depth. The channel is occupied by two layers of immiscible and inviscid liquids of constant densities. The wave-maker is set in motion generating waves which propagate, inside both fluid layers, towards the down-stream extremity of the channel. The required is to calculate the produced waves and also to determine the motion that should be assigned to the wave maker in order to generate internal waves only in the channel.

In the mathematical model, the problem is assumed to be a two dimensional one, the flow is considered irrotational and the free surface and the interface are assumed to remain always near their positions at rest. Furthermore, the motion of the wave maker is considered slow and bounded and the problem is studied in the frame of the non-linear shallow water theory of motion.

A fixed rectangular system of reference is used for the description of the motion of the fluid. The origin O is taken in the free surface at rest with horizontal x-axis pointing along the direction of propagation of waves. The y-axis is directed vertically upwards (Fig. 1).

Notation

The following notation is used throughout this paper:

g	the acceleration of gravity
$h_1, h_2 (h = h_1 + h_2)$	the constant depths of the upper and lower layers of the fluid, respectively
$P(x, y)$	the pressure
S	the stocks parameter

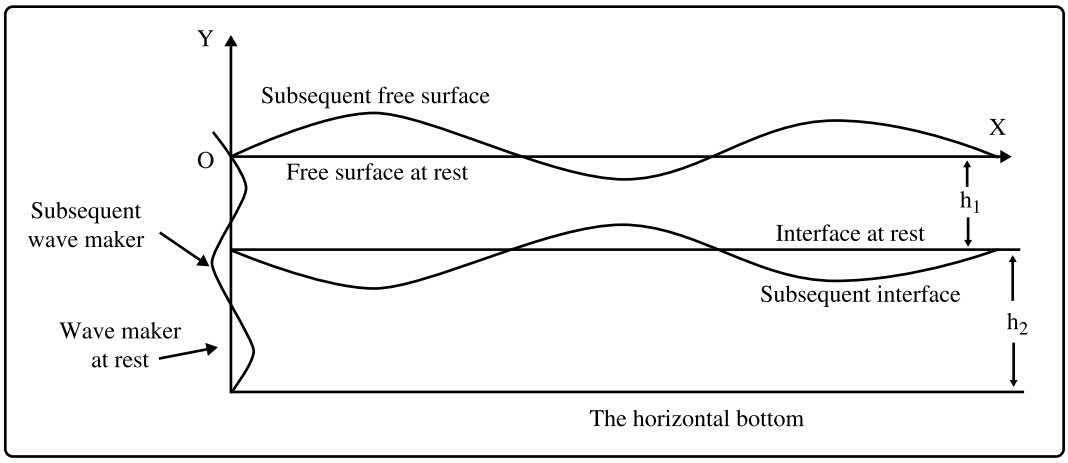

Non-linear Internal Waves, Figure 1
Problem and frame of reference

t	the time
$D_x = \varepsilon f(y, \varepsilon t)$	the horizontal displacement of the wave maker ($= \varepsilon f^{(1)}(y, \varepsilon t)$ in the upper layer and $\varepsilon f^{(2)}(y, \varepsilon t)$ in the lower layer)
$\Phi^{(j)}(x, y, t)$	the velocity potentials: $j = 1, 2$ for the upper and lower layers, respectively
$y = \eta^{(1)}(x, t)$	the equation of the free surface
$y = -h_1 + \eta^{(2)}(x, t)$	the equation of the interface
(x, y)	Cartesian coordinates of a point
$\delta_{n,m}$	the Kronecker delta $= 1$ if $n = m$; 0 if $n \neq m$
ε	a small parameter
ρ_1, ρ_2	the constant densities of the upper and lower layers, respectively ($\rho_1 < \rho_2$)
D_z	an operator of total differentiation w.r.t. z
$\partial_{x,y,\ldots z}$	an operator of partial differentiation w.r.t. $x, y, \ldots,$ and z

Superscripts

1	upper layer
2	lower layer
$'$, $''$	first and second derivatives w.r.t. the argument of the superscripted function

Subscripts

e	the barotropic (external) mode
i	the baroclinic (internal) mode

Equations of Motion

The general system of equations and simplifying conditions describing water wave motion are expressed in terms of the velocity potential in the case of irrotational motion for both homogeneous and stratified fluids (Boussinesq [9], Stoker [33], Wehausen and Laitone [35]).

As in Abou-Dina and Helal [2,3,4], we start with the set of distorted variables: $\widehat{x}, \widehat{y}, \widehat{t}$ defined in terms of x, y, t as:

$$\widehat{x} = \varepsilon x\,, \quad \widehat{y} = y\,, \quad \widehat{t} = \varepsilon t\,, \tag{1}$$

where ε is a small parameter.

Let us denote by $\widehat{\Phi}^{(j)}, \widehat{P}^{(j)}$ and $\widehat{\eta}^{(j)}$ the functions $\Phi^{(j)}(\widehat{x}/\varepsilon, \widehat{y}, \widehat{t}/\varepsilon)$, $P^{(j)}(\widehat{x}/\varepsilon, \widehat{y}, \widehat{t}/\varepsilon)$ and $\eta^{(j)}(\widehat{x}/\varepsilon, \widehat{t}/\varepsilon)$ respectively.

According to the physical and simplifying conditions adopted here, the system of equations and conditions governing the problem is written in terms of the distorted variables $\widehat{x}, \widehat{y}, \widehat{t}$ as follows (Abou-Dina and Helal [2,3,4], Abou-Dina and Hassan [1]):

(I) In the fluid layers with constant densities:

$$\varepsilon^2\, \partial_{\widehat{x}\,\widehat{x}}\, \widehat{\Phi}^{(j)} + \partial_{\widehat{y}\,\widehat{y}}\, \widehat{\Phi}^{(j)} = 0\,, \tag{2}$$

$$\widehat{P}^{(j)}(\widehat{x}, \widehat{y}, \widehat{t}) = \rho_j \Big[\varepsilon\, \partial_{\widehat{t}}\, \widehat{\Phi}^{(j)} + \frac{1}{2}\Big(\varepsilon^2 \left(\partial_{\widehat{x}}\, \widehat{\Phi}^{(j)}\right)^2 + \left(\partial_{\widehat{y}}\, \widehat{\Phi}^{(j)}\right)^2 \Big) + g\widehat{y} \Big]\,, \tag{3}$$

where j hereafter stands for both 1 and 2.

(II) On the free surface which is impermeable and isobaric:

$$\partial_{\widehat{y}}\, \widehat{\Phi}^{(1)} = \varepsilon^2\, \partial_{\widehat{x}}\, \widehat{\eta}^{(1)}\, \partial_{\widehat{x}}\, \widehat{\Phi}^{(1)} + \varepsilon\, \partial_{\widehat{t}}\, \widehat{\eta}^{(1)} \quad \text{at } \widehat{y} = \widehat{\eta}^{(1)}\left(\widehat{x}, \widehat{t}\right)\,, \tag{4}$$

$$\varepsilon\, \partial_{\widehat{t}}\, \widehat{\Phi}^{(1)} + \frac{1}{2}\left(\varepsilon^2 \left(\partial_{\widehat{x}}\, \widehat{\Phi}^{(1)}\right)^2 + \left(\partial_{\widehat{y}}\, \widehat{\Phi}^{(1)}\right)^2 \right) + g\widehat{y} = 0 \quad \text{at } \widehat{y} = \widehat{\eta}^{(1)}\left(\widehat{x}, \widehat{t}\right)\,. \tag{5}$$

(III) At the interface, the impermeability gives:

$$\partial_{\widehat{y}}\, \widehat{\Phi}^{(j)} = \varepsilon^2\, \partial_{\widehat{x}}\, \widehat{\eta}^{(2)}\, \partial_{\widehat{x}}\, \widehat{\Phi}^{(j)} + \varepsilon \partial_{\widehat{t}}\, \widehat{\eta}^{(2)} \quad \text{at } \widehat{y} = -h_1 + \widehat{\eta}^{(2)}\left(\widehat{x}, \widehat{t}\right)\,, \tag{6}$$

Also the compatibility condition for the pressure across this boundary implies:

$$\begin{aligned} \rho_1 \Big[& g\left(h_1 + \widehat{\eta}^{(2)}\right) + \varepsilon\, \partial_{\widehat{t}}\, \widehat{\Phi}^{(1)} \\ & + \frac{1}{2}\left(\varepsilon^2 \left(\partial_{\widehat{x}}\, \widehat{\Phi}^{(1)}\right)^2 + \left(\partial_{\widehat{y}}\, \widehat{\Phi}^{(1)}\right)^2 \right) \Big] \\ = \rho_2 \Big[& g\left(h_1 + \widehat{\eta}^{(2)}\right) + \varepsilon\, \partial_{\widehat{t}}\, \widehat{\Phi}^{(2)} \\ & + \frac{1}{2}\left(\varepsilon^2 \left(\partial_{\widehat{x}}\, \widehat{\Phi}^{(2)}\right)^2 + \left(\partial_{\widehat{y}}\, \widehat{\Phi}^{(2)}\right)^2 \right) \Big] \\ \text{at } \widehat{y} \quad & = -h_1 + \widehat{\eta}^{(2)}\left(\widehat{x}, \widehat{t}\right)\,. \end{aligned} \tag{7}$$

(IV) On the horizontal bottom:

$$\partial_{\widehat{y}}\, \widehat{\Phi}^{(2)} = 0 \quad \text{at } \widehat{y} = -h\,. \tag{8}$$

(V) The radiation condition implies that no wave comes from infinity.

(VI) The initial conditions: At the initial instant of time ($\widehat{t} = 0$), the fluid is at rest with horizontal free surface at $\widehat{y} = 0$ and interface at $\widehat{y} = -h_1$. This initial condition assumes that the functions $f^{(j)}(\widehat{y}, \widehat{t})$ have initial vanishing values and implies at $\widehat{t} = 0$ that

- inside the fluid layers

$$\widehat{\Phi}^{(j)}(\widehat{x}, \widehat{y}, 0) = 0 \,, \tag{9a}$$

$$\partial_{\widehat{x}} \widehat{\Phi}^{(j)}(\widehat{x}, \widehat{y}, 0) = 0 \,, \tag{9b}$$

$$\partial_{\widehat{y}} \widehat{\Phi}^{(j)}(\widehat{x}, \widehat{y}, 0) = 0 \,, \tag{9c}$$

- at the free surface

$$\widehat{\eta}^{(1)}(\widehat{x}, 0) = 0 \tag{9d}$$

and

- at the interface

$$\widehat{\eta}^{(2)}(\widehat{x}, 0) = 0 \,, \tag{9e}$$

provided that

$$f^{(j)}(\widehat{y}, 0) = 0 \quad \text{for } j = 1, 2 \,. \tag{9f}$$

(VII) On the wave maker:

$$\partial_{\widehat{x}} \widehat{\Phi}^{(j)}\left(\widehat{x}, \widehat{y}, \widehat{t}\right) = \varepsilon \frac{\partial}{\partial \widehat{t}} f^{(j)}\left(\widehat{y}, \widehat{t}\right)$$
$$\text{at } \widehat{x} = 0 \,, \quad -h + h_2 \delta_{j,1} \le \widehat{y} \le h_1 \delta_{j,2} \,. \tag{10}$$

to the system of equations and conditions governing the problem in the non-distorted (physical) space, we have to replace ε by unity and omit the hats (^) over the symbols in Eqs. (2) to (10).

The Shallow Water Theory

In the frame work of the shallow water theory, the non-linear system of Eqs. (2) to (10) will be solved using the double series representation:

$$\widehat{\Phi}^{(j)}\left(\widehat{x}, \widehat{y}, \widehat{t}\right) = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \varepsilon^n \exp\left(-\frac{m\,\pi\,\widehat{x}}{h_j\,\varepsilon}\right) \times \widehat{\Phi}^{(j)}_{n,m}\left(\widehat{x}, \widehat{y}, \widehat{t}\right) \,, \tag{11a}$$

$$\widehat{\eta}^{(j)}\left(\widehat{x}, \widehat{t}\right) = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \varepsilon^n \exp\left(-\frac{m\,\pi\,\widehat{x}}{h_j\,\varepsilon}\right) \widehat{\eta}^{(j)}_{n,m}\left(\widehat{x}, \widehat{t}\right) \,. \tag{11b}$$

The validity of using the above representation has been studied by Germain [16,17]. The small parameter ε represents the ratio of the water depth to the wave length. Hence, $S \ge 1$ and we are dealing with the non-linear acoustic analogy.

According to the theory under consideration, at each order (n,m), the above system of equations and conditions must be verified.

For simplification in the following, the hats (^) over the symbols will be omitted.

Verification of the Homogeneous Equations

The above procedure, when applied to the homogeneous Eqs. (2)–(8) along with the radiation condition (V), can lead after some manipulations to the following expression for the total velocity potential up to the second order of the small parameter ε:

$$\begin{aligned}\Phi^{(j)}(x, y, t) = {} & \varepsilon \left[A^{(j)}(x - C_1\,t) + B^{(j)}(x - C_2\,t)\right] \\ & + \varepsilon^2 \left[R^{(j)}(x - C_1\,t) + S^{(j)}(x - C_2\,t)\right] \\ & + \varepsilon^2 \sum_{m=1}^{\infty} A^{(j)}_m(t) \exp\left(-\frac{m\,\pi\,x}{h_j\,\varepsilon}\right) \\ & \times \cos\left(\frac{m\pi}{h_j}\,(y + \delta_{j2} h)\right) + O(\varepsilon^3) \,,\end{aligned} \tag{12}$$

where,

$$C_j^2 = \frac{g}{2}\left\{h - (-1)^j \sqrt{(h_1 - h_2)^2 + 4\,(\rho_1/\rho_2)\,h_1\,h_2}\right\} \,. \tag{13a}$$

The functions $A^{(j)}$, $B^{(j)}$ and $A^{(j)}_m$ are arbitrary functions to be determined, with:

$$A^{(1)}(x - C_1 t) = \gamma_1 A^{(2)}(x - C_1 t) \tag{13b}$$

$$B^{(1)}(x - C_2 t) = \gamma_2 B^{(2)}(x - C_2 t) \tag{13c}$$

and,

$$\gamma_j = \frac{g\,h_2}{C_j^2 - g\,h_1} \,. \tag{13d}$$

The functions $R^{(j)}$ and $S^{(j)}$ are arbitrary functions combined by similar relations to those in Eq. (12) in terms of $A^{(j)}$ and $B^{(j)}$ respectively.

Equations (11)–(13) indicate the existence of two different wave speeds C_1 and C_2 ($C_1 > C_2$). Accordingly, we obtain local oscillations along with progressive waves of two different modes:

(i) External (or barotropic) modes corresponding to the wave speed C_1.
(ii) Internal (or baroclinic) modes corresponding to the wave speed C_2.

The corresponding expressions of the free surface and the interface are given, to the third order, by:

$$\eta^{(1)}(x,t) = -(\varepsilon/g)\,\partial_t\,\Phi^{(1)}(x,y,t) \quad \text{at } y=0\,, \tag{14a}$$

$$\eta^{(2)}(x,t) = \left(\frac{\varepsilon}{g\,(\rho_2-\rho_1)}\right)\partial_t \times\left\{\rho_1\,\Phi^{(1)}(x,y,t) - \rho_2\,\Phi^{(2)}(x,y,t)\right\} \quad \text{at } y=-h_1. \tag{14b}$$

Complete Determination of the Solution

To complete the determination of the velocity potentials $\Phi^{(j)}(x,y,t)$, we verify the initial conditions (9) and the condition on the wave maker (10) which yield, together with expression (12) for the velocity potential functions, the following relations:

- In the upper layer:

$$A^{(1)'}(-C_1\,t) + B^{(1)'}(-C_2\,t) - \frac{\pi}{h_1}\sum_{m=1}^{\infty} m A_m^{(1)}(t) \times\cos\left(\frac{m\pi}{h_1}\,y\right) = \frac{\partial}{\partial t}\,f^{(1)}(y,t)\,, \quad -h_1 \le y \le 0\,. \tag{15a}$$

- In the lower layer:

$$A^{(2)'}(-C_1\,t) + B^{(2)'}(-C_2\,t) - \frac{\pi}{h_2}\sum_{m=1}^{\infty}(-1)^m m A_m^{(2)}(t) \times\cos\left(\frac{m\,\pi}{h_2}(y+h_1)\right) = \frac{\partial}{\partial t}\,f^{(2)}(y,t)\,, \quad -h \le y \le -h_1\,. \tag{15b}$$

Relations (15a) and (15b) together with the initial conditions (9) lead, after some manipulations, to the following expressions:

$$A^{(2)}(x-C_1\,t) = \frac{C_1}{\gamma_1-\gamma_2} \times\left[\frac{\gamma_2}{h_2}\int_{-h}^{-h_1} f^{(2)}\left(y,-\frac{x-C_1\,t}{C_1}\right)\mathrm{d}y - \frac{1}{h_1}\int_{-h_1}^{0} f^{(1)}\left(y,-\frac{x-C_1\,t}{C_1}\right)\mathrm{d}y\right] \tag{16a}$$

$$B^{(2)}(x-C_2\,t) = \frac{C_2}{\gamma_2-\gamma_1} \times\left[\frac{\gamma_1}{h_2}\int_{-h}^{-h_1} f^{(2)}\left(y,-\frac{x-C_2\,t}{C_2}\right)\mathrm{d}y - \frac{1}{h_1}\int_{-h_1}^{0} f^{(1)}\left(y,-\frac{x-C_2\,t}{C_2}\right)\mathrm{d}y\right] \tag{16b}$$

$$A_m^{(1)}(t) = \frac{-2}{m\,\pi}\int_{-h_1}^{0}\left[\frac{\partial}{\partial t}\,f^{(1)}(y,t)\right] \times\cos\left(\frac{m\,\pi}{h_1}\,y\right)\mathrm{d}y \tag{16c}$$

$$A_m^{(1)}(t) = \frac{-2}{m\,\pi}\int_{-h}^{-h_1}\left[\frac{\partial}{\partial t}\,f^{(2)}(y,t)\right] \times\cos\left(\frac{m\,\pi}{h_2}\,(y+h_1)\right)\mathrm{d}y \tag{16d}$$

The functions $A^{(1)}(x-C_1\,t)$ and $B^{(1)}(x-C_2\,t)$ are given in terms of the functions $A^{(1)}(x-C_1\,t)$ and $B^{(1)}(x-C_2\,t)$ by relations (13b) and (13c).

Free Surface and Interface Elevations of Different Modes

The contributions of the progressive waves of external and internal modes to the free surface elevation, $\eta_{\mathrm{e}}^{(1)}(x,t)$ and $\eta_{\mathrm{i}}^{(1)}(x,t)$, are given using (12) and (14a) as

$$\eta_{\mathrm{e}}^{(1)}(x,t) = \frac{-\varepsilon^2\,C_1\,\gamma_1}{g\,(\gamma_1-\gamma_2)} \times\frac{\partial}{\partial t}\left[\frac{\gamma_2}{h_2}\int_{-h}^{-h_1} f^{(2)}\left(y,-\frac{x-C_1\,t}{C_1}\right)\mathrm{d}y - \frac{1}{h_1}\int_{-h_1}^{0} f^{(1)}\left(y,-\frac{x-C_1\,t}{C_1}\right)\mathrm{d}y\right], \tag{17a}$$

$$\eta_{\mathrm{i}}^{(1)}(x,t) = \frac{-\varepsilon^2\,C_2\,\gamma_2}{g\,(\gamma_2-\gamma_1)} \times\frac{\partial}{\partial t}\left[\frac{\gamma_1}{h_2}\int_{-h}^{-h_1} f^{(2)}\left(y,-\frac{x-C_2\,t}{C_2}\right)\mathrm{d}y - \frac{1}{h_1}\int_{-h_1}^{0} f^{(1)}\left(y,-\frac{x-C_2\,t}{C_2}\right)\mathrm{d}y\right]. \tag{17b}$$

The contributions of external and internal modes to the interface elevation, $\eta_{\mathrm{e}}^{(2)}(x,t)$ and $\eta_{\mathrm{i}}^{(2)}(x,t)$, are given in

terms of $\eta_e^{(1)}(x,t)$ and $\eta_i^{(1)}(x,t)$ using (12)–(14) as

$$\eta_e^{(2)}(x,t) = \frac{\rho_2 - \gamma_1 \rho_1}{\gamma_1 (\rho_2 - \rho_1)} \eta_e^{(1)}(x,t) , \tag{18a}$$

$$\eta_i^{(2)}(x,t) = \frac{\rho_2 - \gamma_2 \rho_1}{\gamma_2 (\rho_2 - \rho_1)} \eta_i^{(1)}(x,t) . \tag{18b}$$

It is well known, in the studies dealing with stratified fluids, that the external mode is dominant in the neighborhood of the free surface and has a negligible contribution on the interface, while the internal mode is dominant in the neighborhood of the interface and has a negligible contribution on the free surface. If the motion of the wave maker is such that the function $\eta_e^{(2)}(x,t)$ given by (17a) vanishes, the major contribution of both modes is localized in the neighborhood of the interface and in such cases we say that the wave maker generates internal waves only in the channel.

Secular Term

A suitable form of the double series representation (11) was used in studying certain problems in the case of homogeneous fluids (Abou-Dina and Helal [3], Abou-Dina and Hassan [1]). It has been shown that this procedure leads to a secular term of the third order in ε in the expression of the velocity potential. This secular term, which increases indefinitely with the increase of the variable x, vanishes at $x = 0$. The same result can be shown in the case of stratified fluids, according to the analysis of Abou-Dina and Helal [2]. Hence, although expressions (12)–(18) are valid at $x = 0$, they are not adequate for the description of the propagation of waves far from the wave maker. This unacceptable result is due to certain aspects of the mathematical procedure used (see Abou-Dina and Helal [3], Abou-Dina and Hassan [1] for the justification) and needs to be remedied.

Our main interest, in the remaining part of the present paper, is to modify the mathematical procedure used above in order to describe the propagation of waves generated by the wave maker in going towards the down-stream extremity of the channel.

Multiple Scale Transformation of Variables

We use the set of variables u, v and y defined in terms of the distorted set x, y, t by (cf. Benneye [8] and Temperville [34]):

$$u = x - C\,t , \quad v = \varepsilon^2 x , \tag{19}$$

where C is a real constant to be precised.

Equations (2), (4) up to (8) can be written in terms of u, v and y as:

(i) In the fluid mass:

$$\varepsilon^6\, \partial_{vv}\, \Phi^{(j)} + 2\,\varepsilon^4\, \partial_{uv}\, \Phi^{(j)} + \varepsilon^2\, \partial_{uu}\, \Phi^{(j)} + \partial_{yy}\, \Phi^{(j)} = 0 . \tag{20}$$

(ii) On the free surface:

$$\begin{aligned} \partial_y \Phi^{(1)} &= \varepsilon^6 \partial_v\, \eta^{(1)}\, \partial_v\, \Phi^{(1)} \\ &\quad + \varepsilon^4 \left\{ \partial_u\, \eta^{(1)} \partial_v\, \Phi^{(1)} + \partial_v\, \eta^{(1)}\, \partial_u\, \Phi^{(1)} \right\} \\ &\quad + \varepsilon^2\, \partial_u\, \eta^{(1)}\, \partial_u\, \Phi^{(1)} - \varepsilon\, C\, \partial_u\, \eta^{(1)} \\ &\quad \text{at } y = \eta^{(1)}(u,v) , \end{aligned} \tag{21a}$$

$$\begin{aligned} g\,\eta^{(1)} - \varepsilon\, C\, \partial_u\, \Phi^{(1)} + \frac{1}{2}\Big[\varepsilon^2 \left\{ \partial_u\, \Phi^{(1)} + \varepsilon^2\, \partial_v\, \Phi^{(1)} \right\}^2 \\ + (\partial_y \Phi^{(1)})^2 \Big] = 0 \quad \text{at } y = \eta^{(1)}(u,v) . \end{aligned} \tag{21b}$$

(iii) On the interface:

$$\begin{aligned} \partial_y\, \Phi^{(j)} &= \varepsilon^6 \partial_v\, \eta^{(2)}\, \partial_v\, \Phi^{(j)} \\ &\quad + \varepsilon^4 \left\{ \partial_u\, \eta^{(2)}\, \partial_v\, \Phi^{(j)} + \partial_v\, \eta^{(2)}\, \partial_u\, \Phi^{(j)} \right\} \\ &\quad + \varepsilon^2\, \partial_u\, \eta^{(2)}\, \partial_u\, \Phi^{(j)} - \varepsilon\, C\, \partial_u\, \eta^{(2)} \\ &\quad \text{at } y = \eta^{(1)}(u,v) \text{ for } j = 1,2 \end{aligned} \tag{22a}$$

$$\begin{aligned} &\rho_1 \Big\{ g \left(h_1 + \eta^{(2)} \right) - \varepsilon\, C\, \partial_u\, \Phi^{(1)} \\ &\quad + \frac{1}{2} \left[\varepsilon^2 \left\{ \partial_u\, \Phi^{(1)} + \varepsilon^2\, \partial_v\, \Phi^{(1)} \right\}^2 + \left(\partial_y\, \Phi^{(1)} \right)^2 \right] \Big\} \\ &= \rho_2 \Big\{ g \left(h_1 + \eta^{(2)} \right) - \varepsilon\, C\, \partial_u\, \Phi^{(2)} \\ &\quad + \frac{1}{2} \left[\varepsilon^2 \left\{ \partial_u\, \Phi^{(2)} + \varepsilon^2\, \partial_v\, \Phi^{(2)} \right\}^2 + \left(\partial_y\, \Phi^{(2)} \right)^2 \right] \Big\} \\ &\qquad \text{at } y = -h_1 + \eta^{(2)}(u,v) . \end{aligned} \tag{22b}$$

(iv) On the bottom:

$$\partial_y\, \Phi^{(2)} = 0 \quad \text{at } y = -h . \tag{23}$$

In the regions where we are interested in far from the wave maker, the contribution of the local disturbance on the motion of the fluid layers can be neglected and we use the

following representations for the functions $\Phi^{(j)}(u,v,y)$ and $\eta^{(j)}(u,v)$ (cf. Temperville [34]):

$$\Phi^{(j)}(u,v,y) = \sum_{n=1}^{\infty} \varepsilon^{2n-1}\, \Phi^{(j)}_{2n-1}(u,v,y)\,, \tag{24a}$$

$$\eta^{(j)}(u,v) = \sum_{n=1}^{\infty} \varepsilon^{2n}\, \eta^{(j)}_{2n}(u,v)\,. \tag{24b}$$

Derivation of the KdV Equation

It can be shown from the above equations that for $n = 1$, the functions $\Phi^{(j)}_1(u,v,y)$, $j = 1,2$ are independent of the variable y and that

$$\eta^{(1)}_2(u,v) = \frac{C}{g}\,\partial_u\,\Phi^{(1)}_1(u,v,y)\,, \quad \text{at } y = 0\,, \tag{25}$$

$$\begin{aligned}\eta^{(2)}_2(u,v) = \frac{C}{g\,(\rho_2-\rho_1)}\Big[&\rho_2\,\partial_u\,\Phi^{(2)}_1(u,v,y)\\ &-\rho_1\,\partial_u\,\Phi^{(1)}_1(u,v,y)\Big]\,, \quad \text{at } y = -h_1\,.\end{aligned} \tag{26}$$

Also, for $n = 2$, the values of the constant C appearing in (19) can be obtained as

$$C = \begin{cases} C_1 & \text{for the barotropic (external) mode}\,,\\ C_2 & \text{for the baroclinic (internal) mode}\,,\end{cases} \tag{27}$$

where C_j is given by (13a) for $j = 1,2$.

It can also be shown that the results up to $n = 3$ contain no secular terms, and that the function $\eta^{(1)}_2(u,v)$, characterizing the free surface elevation, satisfies the following KdV equation:

$$L\,\partial_{uuu}\,\eta^{(1)}_2 + M\,\eta^{(1)}_2\,\partial_u\,\eta^{(1)}_2 + N\,\partial_v\,\eta^{(1)}_2 = 0\,, \tag{28}$$

where

$$L = \frac{1}{6}\left(2h^3 + h_1^3 - 3hh_1^2\right) + \frac{\gamma}{2}\left(\frac{C^2}{g} - \frac{h_1}{3}\right)\,, \tag{29a}$$

$$M = 3\gamma + \frac{2(1-\gamma)(\rho_2-\gamma\,\rho_1)}{\gamma(\rho_2-\rho_1)}\,, \tag{29b}$$

$$N = 2\,(h_2 + \gamma\,h_1) \tag{29c}$$

and

$$\gamma = \frac{g\,h_2}{C^2 - g\,h_1}\,. \tag{29d}$$

The initial condition at $v = o$ (or $x = 0$ and $u = -C\,t$), needed for the solution of the KdV Eq. (28) is obtained from (17)in the form:

- Barotropic mode:

$$\begin{aligned}\eta^{(1)}_{\mathrm{e}}(u,0) = &\frac{\varepsilon^2\,C_1^2\,\gamma_1}{g\,(\gamma_1-\gamma_2)}\\ &\times\frac{\partial}{\partial u}\left[\frac{\gamma_2}{h_2}\int_{-h}^{-h_1} f^{(2)}\left(y,-\frac{u}{C_1}\right)\mathrm{d}y\right.\\ &\left.-\frac{1}{h_1}\int_{-h_1}^{0} f^{(1)}\left(y,-\frac{u}{C_1}\right)\mathrm{d}y\right]\end{aligned} \tag{30a}$$

- Baroclinic mode:

$$\begin{aligned}\eta^{(1)}_{\mathrm{i}}(u,0) = &\frac{\varepsilon^2\,C_2^2\,\gamma_2}{g\,(\gamma_2-\gamma_1)}\\ &\times\frac{\partial}{\partial u}\left[\frac{\gamma_1}{h_2}\int_{-h}^{-h_1} f^{(2)}\left(y,-\frac{u}{C_2}\right)\mathrm{d}y\right.\\ &\left.-\frac{1}{h_1}\int_{-h_1}^{0} f^{(1)}\left(y,-\frac{u}{C_2}\right)\mathrm{d}y\right]\end{aligned} \tag{30b}$$

The following relation is also obtained between $\eta^{(2)}_2(u,v)$ and $\eta^{(1)}_2(u,v)$:

$$\eta^{(2)}_2(u,v) = \frac{\rho_2-\gamma\,\rho_1}{\gamma\,(\rho_2-\rho_1)}\,\eta^{(1)}_2(u,v)\,. \tag{31}$$

Conclusions

A multiple scale transformation of variables and a single series representations for the velocity potentials and the free surface and the interface elevations show that these elevations up to the second order satisfy the KdV equation. The initial conditions needed for the solution of this equation are obtained at the position of the wave maker using another technique depending on a double series representation of the unknown functions. The particular choice of the motion of the wave maker can minimize the elevation of the free surface and hence lead to dominant nonlinear internal waves only in the channel.

Future Directions

For a future work following the present one, it is intended to investigate the following items:

- Study the possibility of generating nonlinear internal waves by different types of the motion of the plane wave maker. This will need analytical and numerical work as well.
- Study the inverse problem, precisely the possibility of the recuperation of the water waves' energy and employing it in producing a controlled solid body motion.

- Study the possibility of generating nonlinear internal waves in model problems with different and more complicated geometry, to simulate the actual physical situations.

Bibliography

Primary Literature

1. Abou-Dina MS, Hassan FM (2006) Generation and propagation of nonlinear tsunamis in shallow water by a moving topography. Appl Math Comput 177:785–806
2. Abou-Dina MS, Helal MA (1990) The influence of submerged obstacle on an incident wave in stratified shallow water. Eur J Mech B/Fluids 9(6):545–564
3. Abou-Dina MS, Helal MA (1992) The effect of a fixed barrier on an incident progressive wave in shallow water. Il Nuovo Cimento 107B(3):331–344
4. Abou-Dina MS, Helal MA (1995) The effect of a fixed submerged obstacle on an incident wave in stratified shallow water (Mathematical Aspects). Il Nuovo Cimento B 110(8): 927–942
5. Barthelemy E, Kabbaj A, Germain JP (2000) Long surface wave scattered by a step in a two-layer fluid. Fluid Dyn Res 26: 235–255
6. Benjamin TB (1966) Internal waves of finite amplitude and permanent form. J Fluid Mech 25:241–270
7. Benjamin TB (1967) Internal waves of permanent form of great depth. J Fluid Mech 29:559–592
8. Benney DJ, LIN CC (1960) On the secondary motion induced by oscillations in a shear flow. Phys Fluids 3:656–657
9. Boussinesq MJ (1871) Théorie de l'intumescence liquide appelée onde solitaire ou de translation, se propageant dans un canal rectangulaire. Acad Sci Paris, CR Acad Sci 72:755–759
10. Chen Y, Liu PL-F (1995) Modified Boussinesq equations and associated parabolic models for water wave propagation. J Fluid Mech 288:351–381
11. Choi W, Camassa R (1996) Weakly nonlinear internal waves in a two-fluid system. J Fluid Mech 313:83–103
12. Choi W, Camassa R (1999) Fully nonlinear internal waves in a two-fluid system. J Fluid Mech 396:1–36
13. Craig W, Guyenne P, Kalisch H (2005) Hamiltonian long-wave expansions for free surfaces and interfaces. Commun Pure Appl Math 18:1587–1641
14. Davis RE, Acrivos A (1967) Solitary internal waves in deep water. J Fluid Mech 29:593–607
15. Garett C, Munk W (1979) Internal waves in the ocean. Ann Rev Fluid Mech 11:339–369
16. Germain JP (1971) Sur le caractère limite de la théorie des mouvements des liquides parfaits en eau peu profonde. CR Acad Sci Paris Série A 273:1171–1174
17. Germain JP (1972) Théorie générale d'un fluide parfait pesant en eau peu profonde de profondeur constante. CR Acad Sci Paris Série A 274:997–1000
18. Gobbi MF, Kirby JT, Wei G (2000) A fully nonlinear Boussinesq model for surface waves-Part 2. Extension to O(kh)4. J Fluid Mech 405:181–210
19. Helal MA, Moline JM (1981) Nonlinear internal waves in shallow water: A theoretical and experimental study. Tellus 33:488–504
20. Kabbaj A (1985) Contribution a l'etude du passage des ondes des gravite sur le talus continental et a la generation des ondes internes. These de doctorat d'etat, IMG Université de Grenoble
21. Keulegan GH (1953) Hydrodynamical effects of gales on Lake Erie. J Res Natl Bur Std 50:99–109
22. Kubota T, Ko DRS, Dobbs LD (1978) Propagation of weakly nonlinear internal waves in a stratified fluid of finite depth. AIAA J Hydrodyn 12:157–165
23. LeBlond PH, Mysak LA (1978) Waves in Ocean. Elsevier, Amsterdam
24. Liu C-M, Lin M-C, Kong C-H (2008) Essential properties of Boussinesq equations for internal and surface waves in a two-fluid system. Ocean Eng 35:230–246
25. Long RR (1956) Solitary waves in one- and two-fluid systems. Tellus 8:460–471
26. Lynett PJ, Liu PL-F (2002) A two-dimensional depth-integrated model for internal wave propagation over variable bathymetry. Wave Motion 36:221–240
27. Madsen PA, Schaffer HA (1998) Higher-order Boussinesq-type equations for surface gravity waves: derivation and analysis. Philos Trans R Soc Lond A 356:3123–3184
28. Matsuno Y (1993) A unified theory of nonlinear wave propagation in two-layer fluid systems. J Phys Soc Jpn 62:1902–1916
29. Nwogu O (1993) Alternative form of Boussinesq equations for nearshore wave propagation. J Waterways Port Coast Ocean Eng ASCE 119:618–638
30. Ono H (1975) Algebraic solitary waves in stratified fluids. J Phys Soc Jpn 39:1082–1091
31. Peters AS, Stoker JJ (1960) Solitary waves in liquid having nonconstant density. Comm Pure Appl Math 13:115–164
32. Robinson RM (1969) The effect of a vertical barrier on internal waves. Deep-Sea Res 16:421–429
33. Stoker JJ (1957) Water waves. Interscience, New York
34. Temperville A (1985) Contribution à la théorie des ondes de gravité en eau peu profonde. Thèse de doctorat d'état, IMG Université de Grenoble
35. Wehausen JV, Laitone EV (1960) Surface waves. In: Handbuch der Physik 9. Springer, Berlin
36. Wei G, Kirby JT, Grilli ST, Subramanya R (1995) A fully nonlinear Boussinesq model for surface waves, Part 1. Highly nonlinear, unsteady waves. J Fluid Mech 294:71–92

Books and Reviews

Germain JP, Guli L (1977) Passage d'une onde sur une barrier mince immergée en eau peu profonde. Ann Hydrog 5(746):7–11

Pinettes M-J, Renouard D, Germain J-P (1995) Analytical and experimental study of the oblique passing of a solitary wave over a shelf in a two-layer fluid. Fluid Dyn Res 16:217–235

Non-linear Ordinary Differential Equations and Dynamical Systems, Introduction to

Ferdinand Verhulst
Mathematisch Instituut, University of Utrecht, Utrecht, The Netherlands

An ordinary differential equation (ODE) is called linear if it can be written in the form $dy/dx = f(x)y + g(x)$ with x a real or complex variable and y an n-dimensional (finite) real or complex vector function. Non-linear ODEs are then n-dimensional ODEs of the form $dy/dx = F(x, y)$ that are not linear. Jean Mawhin famously compared this distinction to a division of the animal world into 'elephants' and 'non-elephants', but even admitting that the distinction between linear and non-linear ODEs is artificial and of relatively recent date, it makes a little bit more sense than it looks like at first sight. The reason is, that in many problem formulations in classical physics, linear equations are quite common. Also, when considering non-linear ODEs, but linearizing around particular solutions, a number of fundamental theorems can be used to characterize the particular solutions starting with the features of the linearized equation. This will become clear in a number of the articles that follow. On the other hand, right from the beginning of the development of classical physics, the analysis of differential equations with their fully non-linear behavior has been necessary and essential. Think of celestial mechanics that started in the 18th century and for instance the analysis of solitons in fluid mechanics.

The great scientists of the 18th century, among which Newton, Euler, Lagrange and Laplace, were all concerned with the formulation and first analysis steps of non-linear ODEs. This work was continued by mathematicians like Jacobi and Painlevé who devoted much of their attention to transformation methods and the analysis of special cases. A new stimulus came in the second half of the 19th century with the insights and fundamental ideas of Henri Poincaré. In fact his ideas are still fully alive and active today. Poincaré realized that non-linear ODEs can in general not be solved explicitly, i. e. expressed in terms of elementary functions, so that calculations should be supplemented by qualitative theory. Poincaré developed both quantitative and qualitative methods and his approach has shaped the analysis of non-linear ODEs in the period that followed up till now, becoming part of the topic of dynamical systems. This volume of the Encyclopedia reflects to a large part these new developments.

Basic questions on existence and uniqueness of solutions are discussed by Gianne Derks (see ► Existence and Uniqueness of Solutions of Initial Value Problems) with attention to recent extensions of the notion of an ODE. Most of this is classical material.

The article by Carmen Chicone (see ► Stability Theory of Ordinary Differential Equations) deals with stability theory. It is concerned with the mathematical formulation and the basic results, but also includes the stability question in the context of conservative systems and the part played by the KAM theorem. The stability of periodic orbits gets special attention and in addition the stability of the orbit structure as a whole. This is usually referred to as structural stability.

The theory of periodic solutions of non-autonomous ODEs displays a number of striking differences with the autonomous case. Jean Mawhin (see ► Periodic Solutions of Non-autonomous Ordinary Differential Equations) presents classical and new results in this field.

A more general technique to study equilibria, periodic solutions and their bifurcations is the Lyapunov–Schmidt method which is intimately connected to conditions of the implicit function theorem in its abstract form. André Vanderbauwhede (see ► Lyapunov–Schmidt Method for Dynamical Systems) explains this basic method for equilibria and for periodic solutions with extensions to infinite dimensions.

Center manifolds are manifolds associated with the critical part of the spectrum of equilibria and periodic solutions. They arise naturally in theory and applications. George Osipenko (see ► Center Manifolds) discusses the theory with special attention to the corresponding (and very effective) reduction principle.

A number of special topics have been getting a lot of attention, both in mathematical and in applied research. Relaxation oscillations are described by Johan Grasman (see ► Relaxation Oscillations), starting with the classical example of the Van der Pol-oscillator and going on to more complicated systems. This also involves the discussion of canards in geometric singular perturbation theory which continues to produce activity in dynamical systems research and applications in mathematical biology.

The dynamics of Hamiltonian systems is described by Heinz Hanßmann (see ► Dynamics of Hamiltonian Systems). Starting with the classical formulation and basic notions, the article elaborates on integrability questions and the KAM theorem, Cantor dust and tori bifurcations, thus introducing the reader to the most recent results.

Periodic solutions of Hamiltonian systems merit a special treatment in the article by Luca Sbano (see ► Periodic Orbits of Hamiltonian Systems). Natural problems are the continuation of orbits and the part played by symmetries. Variational methods have become important in this field, enabling us to consider the topology and geometry of periodic paths in sufficient generality.

Another classical topic that is fully alive is the dynamics of parametric excitation. Alan Champneys (see ► Dynamics of Parametric Excitation) introduces the necessary Floquet theory, corresponding bifurcation diagrams and a number of applications.

Non-linear ODEs has grown into the more general field of dynamical systems. Beginning with hyperbolic behavior and examples, Araújo and Viana (see ► Hyperbolic Dynamical Systems) discuss attractors and physical measures with the idea of stochastic stability. This leads naturally to the formulation of obstructions to hyperbolicity, partial and non-uniform hyperbolicity.

To handle dynamical systems in practice, one needs quantitative methods. Meijer, Dercole and Oldeman (see ► Numerical Bifurcation Analysis) consider numerical bifurcation analysis for systems depending on parameters. This involves continuation and detection of bifurcations, complications like branch switching and orbit connection and a guide to possible software environments.

The literature on non-linear ODEs grows at a fast pace. The articles are presenting a tour of the relevant literature, and even more importantly, they point out future directions of research in their respective fields.

Non-linear Partial Differential Equations, Introduction to

Italo Capuzzo Dolcetta
Dipartimento di Matematica,
Sapienza Universita' di Roma, Rome, Italy

A large number of nonlinear phenomena in fundamental sciences (physics, chemistry, biology ...), in technology (material science, control of nonlinear systems, ship and aircraft design, combustion, image processing ...) as well in economics, finance and social sciences are conveniently modeled by nonlinear partial differential equations (NLPDE, in short). Let us mention, among the most important examples for the applications and from the historical point of view, the Euler and Navier–Stokes equations in fluid dynamics and the Boltzmann equation in gas dynamics. Other fundamental models, just to mention a few of them, are reaction-diffusion, porous media, nonlinear Schrödinger, Klein–Gordon, eikonal, Burger and conservation laws, nonlinear wave Korteweg–de Vries ...

The above list is by far incomplete as one can easy realize by looking at the current scientific production in the field as documented, for example, by the American Mathematical Society database **MathSciNet**.

Despite an intense mathematical research activity going on since the second half of the XXth century, the extremely diversified landscape of the area of NLPDE is still largely unexplored and a number of difficult problems are still open.

Generally (and therefore rather vaguely) speaking, one of the main difficulties induced by the nonlinear (as opposed to linear) structure of a partial differential equation is that such equations, especially first-order ones, do not in general possess smooth solutions. The analytical approach therefore requires the adoption of appropriate generalized notions of solutions (weak solutions in Sobolev spaces, entropy solutions, viscosity solutions, re-normalized solutions ...) in order to handle the classical Hadamard well-posedness approach to the rigorous validation of the model (existence, uniqueness and stability of solutions).

Another fundamental difference with respect to the linear theory lies in the fact that no explicit representation formulas for solutions are in general available, even for simplified models. Fortunately enough however, **good theory is (almost) as useful as exact formulas**, as stated by L.C. Evans in the preface of his book *Partial Differential Equations, Graduate Studies in Mathematics* (Volume 19, American Mathematical Society, Providence Rhode Island, 1998). It is clear also, in this respect, that progress in NLPDE is necessarily strongly connected with the development of numerical methods of simulations.

A further "pathology" due to nonlinearity is that solutions of initial value problems for nonlinear evolution equations may exist only for a small lap of time (the blow-up phenomenon). These and other unavoidable features imply the need to develop new, sophisticated and often ad hoc mathematical methods for investigating different classes of NLPDE.

The nonlinear theory (rather, theories) has been widely developed for different classes of equations, mainly in the direction of existence (via monotonicity, compactness, critical points methods ...), regularity of generalized solutions, qualitative analysis of solutions, limit problems (e. g. large time asymptotic, homogenization, scaling limits ...).

Quoting P.L. Lions, **new tools are required and discovered [at this purpose] in connection with all fields of analysis (functional, real, global, complex, numerical, Fourier ...)**, see *On Some Challenging Problems in Nonlinear Partial Differential Equations* (Mathematics: Frontiers and Perspectives 2000, International Mathematical Union).

The 11 papers comprised in this section offer deep insight and up to date information on some of the various aspects mentioned above. A common feature of the papers is that their focus is on nonlinear partial differential equations arising in "real world" applications. The paper ► Hyperbolic Conservation Laws by A. Bressan reports on very recent important advances in the understanding of first-order nonlinear partial differential equations (or systems) describing the time evolution of certain basic quantities in

physical models such as mass, momentum, energy, electric charge. The main mathematical difficulty is related to the strong linearity and the absence of dissipative terms in the equations; the evolution of smooth initial data may then produce discontinuities (shocks) in finite time, thus requiring the introduction of appropriate notion of weak entropic solutions.

Many important NLPDE, e. g. Hamilton–Jacobi–Bellman and Isaacs equations arising in optimal control and game theory, mean-curvature flow and ∞-Laplace equations do not have a divergence structure. In this case, the notion of weak solution in the distribution sense is not applicable. S. Koike's ► Non-linear Partial Differential Equations, Viscosity Solution Method in provides an account of ideas and results of the theory of viscosity solutions. This generalized notion of solution, which is intimately related to the maximum principle for elliptic equations, has proved to be most appropriate to implement the well-posedness program to large classes of second-order fully NLPDE not treatable by variational methods.

First-order NLPDE of Hamilton–Jacobi-type have a central relevance both in classical and quantum mechanics as well in calculus of variations and optimal control theory. Lack of regularity occurs here at the level of the smoothness of the gradient of solutions for reasons which are somewhat similar to those indicated for hyperbolic conservation laws. The contribution by A. Siconolfi ► Hamilton–Jacobi Equations and Weak KAM Theory reports on current research on the impact of the notion of the Aubry–Mather set, imported from dynamical systems theory, in the understanding of uniqueness, extra smoothness properties and large time behavior of viscosity solutions.

In a physical model, dispersion can be understood as the decrease with time of the size of some relevant quantity such as matter or energy over a given volume. P. D'Ancona's ► Dispersion Phenomena in Partial Differential Equations examines in detail analytical techniques to evaluate, in suitable norms, the rate of dispersion in some very fundamental physical models such as the nonlinear wave and Schrödinger equations; new tools in Fourier and harmonic analysis play a key role in this respect. In connection with a general remark made above, it is worth pointing out that sharp dispersion (or Stricharz type) estimates have relevant applications to the issue of existence of solutions that are global in time.

The interaction of nonlinearity with randomness in a PDE model is the leading theme of ► Non-linear Stochastic Partial Differential Equations by G. Da Prato. This is a definitely new and rapidly developing area of investigation. The paper presents a general nonlinear semigroup approach to the study, mainly focused at global existence and well-posedness, of a large class of stochastic dynamical systems on Hilbert spaces. The large time behavior of solutions is also analyzed, relying on the notion of invariant measure. The application of general abstract results to specific models of interest, such as Ornstein–Uhlenbeck, reaction-diffusion, Burger's, 2D Navier–Stokes equations perturbed by noise is also discussed in the paper.

The Navier–Stokes equations of fluid dynamics are presently a fundamental model for simulations in several branches of applied sciences (e. g. meteorology, oceanography, biology . . .) and design in industry (airplane, car, oil . . .). The text by G.P. Galdi ► Navier–Stokes Equations: A Mathematical Analysis is an extended and fairly complete review of mathematical tools (appropriate function spaces, a priori estimates, fixed point theorems), results (concerning in particular the well-posedness of initial and boundary value problems) and open questions (including the famous one concerning the global regularity of solutions in 3D) pertaining to this important topic.

Stokes equations, which are obtained as a formal limit from Navier–Stokes equations, are central in the article ► Biological Fluid Dynamics, Non-linear Partial Differential Equations by A. De Simone, F. Alouges, and A. Lefebvre. A mathematical model for the swimming of a micro-organism at a low Reynolds number regime is proposed. Methods from sub-Riemannian geometry as well as numerical ones for the quantitative optimization of the strokes of micro-swimmers are discussed.

Control theory is concerned with the issue of influencing, by way of an external action, the evolution of a system. Many applications, both from the scientific and technological point of view, involve distributed control systems, that is systems whose evolution is governed by partial differential equations. Important examples range from control of diffusion processes to nonlinear elasticity, from control of biological systems (see above) to traffic flows on networks. The text by Alabau-Boussouira and P. Cannarsa ► Control of Non-linear Partial Differential Equations is a wide-ranging overview of the state-of-the-art in the field, covering among others the issues of controllability, stabilization and optimal control.

Traffic flows on networks are discussed in ► Vehicular Traffic: A Review of Continuum Mathematical Models by B. Piccoli and A. Tosin. The paper reviews in particular macroscopic and kinetic models. Macroscopic modeling of traffic flows stems from the Euler and Navier–Stokes equations while kinetic models rely on principles of statistical mechanics. Since macroscopic models take often the form of nonlinear hyperbolic conservation laws, special at-

tention is given in the paper to the analysis of those models by nonlinear hyperbolic equation techniques (see also the paper by Bressan in this respect).

An important issue in modeling nonlinear physical systems is the reduction of complexity. One way to realize this is to use scaling limit procedures allowing a more manageable description of the system itself through macroscopic coordinates. The paper ► Scaling Limits of Large Systems of Non-linear Partial Differential Equations contributed by D. Benedetto and M. Pulvirenti illustrates ideas and methods to treat large classical and quantum particle systems in the weak coupling regime. It is also shown that those systems can be conveniently described by macroscopic quantities whose evolution is governed by the Fokker–Planck–Landau and the Boltzmann equations for the classical and quantum case, respectively.

Non-linear Partial Differential Equations, Viscosity Solution Method in

Shigeaki Koike
Department of Mathematics, Saitama University, Saitama, Japan

Article Outline

Glossary
Definition of the Subject
Introduction
Examples
Comparison Principle
Existence Results
Boundary Value Problems
Asymptotic Analysis
Other Notions
Future Directions
Bibliography

Glossary

Weak solutions In the study of mth order partial differential equations (abbreviated, PDEs), a function is informally called a classical solution of the PDE if (a) it is m-times differentiable, and (b) the PDE holds at each point of the domain by putting its derivatives there. However, it is not easy to find such classical solutions of PDEs in general except for some special cases.

The standard strategy to find a classical solution is first to look for a candidate of solutions, which becomes a classical solution if the property (a) holds for it. Here, such a candidate is called a weak solution of the PDE.

Viscosity solutions In 1981, for first order PDEs of non-divergence type, M.G. Crandall and P.-L. Lions in [15,16] (also [17]) introduced the notion of weak solutions, which are called viscosity solutions. The definition of those is the property which the limit of approximate solutions via the vanishing viscosity method admits.

Afterwards, the notion was extended to fully nonlinear second order elliptic/parabolic PDEs.

For general theory of viscosity solutions, [3,5,7,18,20] are recommended for the interested readers.

Throughout this article, to minimize the references, [18] will be often referred to instead of the original papers except for some pioneering works or those which appeared after [18].

Ellipticity/parabolicity General second order PDEs under consideration are

$$F(x, u(x), Du(x), D^2u(x)) = 0 \quad \text{in } \Omega\,. \tag{E}$$

Here, $u\colon \Omega \to \mathbf{R}$ is the unknown function, $\Omega \subset \mathbf{R}^n$ an open set, $F\colon \Omega \times \mathbf{R} \times \mathbf{R}^n \times S^n \to \mathbf{R}$ a given (continuous) function, S^n the set of $n \times n$ symmetric matrices with the standard ordering, $Du(x) = ((\partial u)/(\partial x_1)(x), \ldots, (\partial u)/(\partial x_n)(x))$, and $D^2u(x) \in S^n$ whose (i, j)th entry is $(\partial^2 u)/(\partial x_i \partial x_j)(x)$.

According to the early literature in viscosity solution theory, (E) is called elliptic if

$$X \le Y \Longrightarrow F(x, r, p, X) \ge F(x, r, p, Y)$$

for $(x, r, p, X, Y) \in \Omega \times \mathbf{R} \times \mathbf{R}^n \times S^n \times S^n$. It should be remarked that the opposite order has been also used.

When F does not depend on the last variables (i. e. first order PDEs), it is automatically elliptic. Thus, the above notion has been called degenerate elliptic.

The evolution version of general PDEs is

$$u_t(x, t) + F(x, t, u(x, t), Du(x, t), D^2u(x, t)) = 0 \quad \text{in } Q_T := \Omega \times (0, T]\,. \tag{P}$$

Here, $u\colon Q_T \to \mathbf{R}$ is the unknown function, $F\colon \Omega \times (0, T] \times \mathbf{R} \times \mathbf{R}^n \times S^n \to \mathbf{R}$ a given function, $T > 0$, and $u_t(x, t) = (\partial u)/(\partial t)(x, t)$. In (P), the notations $Du(x, t)$ and $D^2u(x, t)$ do not contain derivatives with respect to t.

Similarly, (P) is called parabolic if $F(\cdot, t, \cdot, \cdot, \cdot)$ is elliptic for each $t \in (0, T]$.

Uniform ellipticity/uniform parabolicity Denoted by $S^n_{\lambda,\Lambda} := \{A \in S^n | \lambda I \le A \le \Lambda I\}$ for fixed $0 < \lambda \le \Lambda$, the Pucci operators $\mathcal{P}^{\pm} : S^n \to \mathbf{R}$ are defined by

$$\mathcal{P}^+(X) := \max_{A \in S^n_{\lambda,\Lambda}} \{-\mathrm{trace}(AX)\} \quad \text{and}$$
$$\mathcal{P}^-(X) := \min_{A \in S^n_{\lambda,\Lambda}} \{-\mathrm{trace}(AX)\}\,.$$

Then, the PDE (E) is called uniformly elliptic if

$$\mathcal{P}^-(X-Y) \le F(x,r,p,X) - F(x,r,p,Y) \le \mathcal{P}^+(X-Y)$$

for $(x,r,p,X,Y) \in \Omega \times \mathbf{R} \times \mathbf{R}^n \times S^n \times S^n$. This is a fully nonlinear version of the standard uniform ellipticity. For the theory of second order uniformly elliptic PDEs, [24] is the standard text book.

Similarly, (P) is called uniformly parabolic if $F(\cdot, t, \cdot, \cdot, \cdot)$ is uniformly elliptic for each $t \in (0, T]$.

Dynamic programming principle In stochastic control problems, the value function is determined by minimizing given cost functionals. The dynamic programming principle (abbreviated, DPP), which was established as the Bellman's principle of optimality, is a formula which the value function satisfies.

The DPP indicates that the value function is a viscosity solution of the associated Hamilton–Jacobi–Bellman (abbreviated, HJB) equation.

Definition of the Subject

In order to investigate phenomena in both natural and social sciences, it is important to analyze solutions of PDEs derived from certain minimization principles such as calculus of variations.

Since it is hard to find classical solutions of PDEs in general, the first strategy is to look for weak solutions of those. For PDEs of divergence type, the most celebrated notion of weak solutions is that in the distribution sense, which is formally derived through the integration by parts.

On the other hand, before viscosity solutions were introduced, there were several notions of weak solutions for PDEs of non-divergence type such as generalized solutions for first order PDEs by S.N. Kružkov.

For second order elliptic/parabolic PDEs of non-divergence type, it has turned out through much research that the notion of viscosity solutions is the most appropriate one by several reasons:

(i) Viscosity solutions admit well-posedness (i. e. existence, uniqueness, stability),
(ii) The notion of viscosity solutions is naturally derived from the vanishing viscosity method,
(iii) In deterministic/stochastic control problems, the expected solutions are the unique viscosity solutions of the associated HJB equations in principle,
(iv) The viscosity solution method can deal with first and second order elliptic/parabolic PDEs by a unified manner,
(v) The viscosity solution method can be applied to various non-divergent PDEs for which there were no good notions of weak solutions such as HJB equations, mean curvature flow equations, ∞-Laplace equation etc.

Introduction

In what follows, elliptic PDEs (E) will be mainly discussed since one can treat (P) with some modifications. For the sake of simplicity, the variables of the unknown functions and their derivatives in the PDEs will be suppressed if there is no confusion.

For $U \subset \mathbf{R}^n$, which will be Ω, $\overline{\Omega}$ or $\partial\Omega$ later, $C^1(U)$ (resp., $C^2(U)$) denotes the set of real-valued functions ϕ in U such that ϕ and $D\phi$ (resp., ϕ, $D\phi$ and $D^2\phi$) are continuous in U.

The definition of viscosity solutions presented in [17] has been widely used although the original one in [15,16] is different but equivalent.

Definition 1 $u : \Omega \to \mathbf{R}$ is called a viscosity subsolution (resp., supersolution) of (E) in Ω if

$$F(x, u(x), D\phi(x), D^2\phi(x)) \le 0 \quad (\text{resp., } \ge 0) \tag{1}$$

whenever $(u-\phi)(x) = \sup_\Omega(u-\phi)$ (resp., $\inf_\Omega(u-\phi)$) for $\phi \in C^2(\Omega)$ and $x \in \Omega$.

Finally, u is called a viscosity solution of (E) in Ω if it is a viscosity sub- and supersolution of (E) in Ω.

Remark 2 In order to study boundary value problems in the viscosity sense in Sect. "Boundary Value Problems", it is necessary to modify the above definition by replacing Ω by $\overline{\Omega}$. To this end, Ω in the semi-jets, the equivalent definition and Ishii's lemma etc. below should be also replaced by $\overline{\Omega}$.

The above definition indicates that if the so-called test function $\phi \in C^2(\Omega)$ touches u from above (resp., below) at $x \in \Omega$, by noting $0 = (u-\phi)(x) \ge$ (resp., $\le$)$(u-\phi)(y)$ for $y \in \Omega$, the one-sided inequality (1) is required, instead of the Eq. (E), with the derivatives of ϕ at $x \in \Omega$.

It is worth mentioning that if $u \in C^2(\Omega)$ is a viscosity subsolution (resp., supersolution) of (E) in Ω, then it is

a classical subsolution (resp., supersolution) of (E) in Ω;

$$F(x, u(x), Du(x), D^2u(x)) \le 0 \quad (\text{resp., } \ge 0) \quad \text{for each } x \in \Omega\,.$$

On the contrary, if $u \in C^2(\Omega)$ is a classical subsolution (resp., supersolution) of (E) in Ω, then it is a viscosity subsolution (resp., supersolution) of (E) in Ω whenever (E) is elliptic.

One can observe that if u is a viscosity subsolution (resp., supersolution) of (E) in Ω, then $-u$ is a viscosity supersolution (resp., subsolution) of

$$-F(x, -u, -Du, -D^2u) = 0 \quad \text{in } \Omega\,.$$

If F is independent of X (i. e. first order PDEs), then it is possible to replace $C^2(\Omega)$ by $C^1(\Omega)$ in the definition.

For the definition to (P), Ω and $C^2(\Omega)$, respectively, are changed by Q_T and $C^{2,1}(Q_T)$ in the above, where $Q_T = \Omega \times (0, T]$. Here, $C^{2,1}(Q_T)$ denotes the set of real-valued functions ϕ in Q_T such that ϕ, $D\phi$, $D^2\phi$ and ϕ_t are continuous in Q_T.

An equivalent definition of viscosity solutions will be utilized to establish the comparison principle for second order PDEs. To this end, it is necessary to prepare some notation: for $u\colon \Omega \to \mathbf{R}$, the semi-jets $J^{2,\pm}u(x)$ of order 2 at $x \in \Omega$ are defined by

$$J^{2,+}_{\Omega}u(x) = \left\{ (p, X) \in \mathbf{R}^n \times S^n \;\middle|\; \begin{array}{c} u(y) \le u(x) - \langle p, y - x\rangle \\ -\frac{1}{2}\langle X(y-x), y-x\rangle + o(|x-y|^2) \\ \text{for } y \in \Omega \text{ as } y \to x \end{array} \right\},$$

$$J^{2,-}_{\Omega}u(x) = \left\{ (p, X) \in \mathbf{R}^n \times S^n \;\middle|\; \begin{array}{c} u(y) \ge u(x) - \langle p, y - x\rangle \\ -\frac{1}{2}\langle X(y-x), y-x\rangle + o(|x-y|^2) \\ \text{for } y \in \Omega \text{ as } y \to x \end{array} \right\}.$$

In order to state a key lemma in Sect. "Comparison Principle", it is also needed to introduce a sort of closures of $J^{2,\pm}_{\Omega}u(x)$; for $x \in \Omega$,

$$\bar{J}^{2,\pm}_{\Omega}u(x) = \left\{ (p, X) \in \mathbf{R}^n \times S^n \;\middle|\; \begin{array}{c} \exists (p_k, X_k) \in J^{2,\pm}_{\Omega}u(x_k) \text{ for } \exists x_k \in \Omega, \\ \text{such that } (x_k, u(x_k), p_k, X_k) \\ \to (x, u(x), p, X) \text{ as } k \to \infty \end{array} \right\}.$$

Definition 3 (Equivalent definition) u is a viscosity subsolution (resp., supersolution) of (E) in Ω if and only if

$$F(x, u(x), p, X) \le 0 \quad (\text{resp., } \ge 0)$$

provided $(p, X) \in \bar{J}^{2,+}_{\Omega}u(x)$ (resp., $(p, X) \in \bar{J}^{2,-}_{\Omega}u(x)$) for $x \in \Omega$.

One can replace $\bar{J}^{2,\pm}_{\Omega}$ in the above by $J^{2,\pm}_{\Omega}$.

Examples

Since the viscosity solution method is available only for second order elliptic/parabolic PDEs up to now, all the examples below are elliptic/parabolic.

To explain how the viscosity solution method is useful, a typical example is the eikonal equation: for $\Omega_1 = (-1, 1) \subset \mathbf{R}$,

$$|Du| = 1 \quad \text{in } \Omega_1 \tag{2}$$

under the Dirichlet condition $u(\pm 1) = 0$. Since (2) is of non-divergence type, the notion of weak solutions in the distribution sense is not available to (2). It is also known that there are no classical solutions $u \in C(\overline{\Omega}_1) \cap C^1(\Omega_1)$ of (2). In fact, from an optimal control and a geometric view points, the expected solution is the distance function from the boundary $\partial\Omega_1$; $u_0(x) = \mathrm{dist}(x, \partial\Omega_1) = 1 - |x|$.

If one employs Lipschitz continuous functions satisfying (2) almost every point of Ω_1 as weak solutions of (2), then the uniqueness of such weak solutions fails. Indeed, there exist many other weak solutions in this sense; e. g.

$$u_1(x) = |x| - 1\,,$$

$$u_2(x) = \begin{cases} \min\{x+1, -x\} & (\text{for } x \in [-1, 1/2])\,, \\ x - 1 & (\text{for } x \in (1/2, 1]) \end{cases} \quad \text{etc}\,.$$

However, it can be proved that u_0 is the unique viscosity solution of (2) under $u_0(\pm 1) = 0$. It is worth mentioning that there are no test functions $\phi \in C^1(\Omega)$ touching u_0 from below at $x = 0$. Thus, no inequalities for viscosity supersolutions at $x = 0$ are required.

Moreover, if u_ε is a (classical) solution of the approximate equation via the vanishing viscosity method,

$$-\varepsilon \triangle u_\varepsilon + |Du_\varepsilon| = 1 \quad \text{in } \Omega_1 \quad (\varepsilon > 0)\,,$$

under $u_\varepsilon(\pm 1) = 0$, then one can verify that $\lim_{\varepsilon \to 0} u_\varepsilon = u_0$ in $\overline{\Omega}_1$.

The next example, to which the viscosity solution method is applicable, is a fully nonlinear PDE arising in stochastic control problems.

For a set of parameters $\mathcal{A} \neq \emptyset$, F is expressed by

$$F(x,r,p,X) = \max_{a\in\mathcal{A}}\{-\mathrm{trace}(A_a(x)X) - \langle g_a(x), p\rangle + c_a(x)r - f_a(x)\}\,,$$

where $A_a\colon \Omega \to S^n$, $g_a\colon \Omega \to \mathbf{R}^n$, $c_a, f_a\colon \Omega \to \mathbf{R}$ are given functions for each $a \in \mathcal{A}$. One can see that the mapping $(r,p,X) \to F(x,r,p,X)$ is convex in this case. The corresponding PDE is called the HJB equation:

$$\max_{a\in\mathcal{A}}\big\{-\mathrm{trace}(A_a(x)D^2u) - \langle g_a(x), Du\rangle + c_a(x)u - f_a(x)\big\} = 0\,. \quad (3)$$

If $A_a \geq 0$ in Ω for $a \in \mathcal{A}$, then (3) is elliptic.

From a stochastic control point of view, $A_a(x)$ is naturally represented by $1/2\sigma_a(x)\sigma_a^t(x)$ for an $n \times m$ matrix $\sigma_a(x)$ with some integer m (for $a \in \mathcal{A}$ and $x \in \Omega$). Thus, the property $A_a \geq 0$ in Ω holds true naturally in applications.

A fully nonlinear PDE for which the mapping $(r,p,X) \to F(x,r,p,X)$ is neither convex nor concave arises in stochastic differential games:

$$\min_{b\in\mathcal{B}}\max_{a\in\mathcal{A}}\big\{-\mathrm{trace}(A_{a,b}(x)D^2u) - \langle g_{a,b}(x), Du\rangle + c_{a,b}(x)u - f_{a,b}(x)\big\} = 0\,,$$

where $A_{a,b}\colon \Omega \to S^n$, $g_{a,b}\colon \Omega \to \mathbf{R}^n$, $c_{a,b}, f_{a,b}\colon \Omega \to \mathbf{R}$ are given for $(a,b) \in \mathcal{A} \times \mathcal{B}$. Here, $\mathcal{B}$ is another set of parameters with a different purpose.

This is called the Hamilton–Jacobi–Bellman–Isaacs (abbreviated, HJBI) equation, for which the viscosity solution method works when $A_{a,b} \geq 0$ in Ω for $(a,b) \in \mathcal{A} \times \mathcal{B}$. In principle, any fully nonlinear second order PDEs can be represented by HJBI equations with suitable choices of $A_{a,b}$, $g_{a,b}$, $c_{a,b}$, $f_{a,b}$, $\mathcal{A}$ and $\mathcal{B}$ though those HJBI equations might become complicated.

It is trivial to derive parabolic versions of HJB and HJBI equations by adding u_t together with allowing t-dependence on the given functions.

There are more PDEs, to which the viscosity solution method has been applied.

(Monge-Ampère equation) $\quad -\det(D^2u) - f(x) = 0,$

(mean curvature flow) $\quad u_t - \triangle u + \dfrac{\langle D^2uDu, Du\rangle}{|Du|^2} = 0,$

(∞-Laplace equation) $\quad -\langle D^2uDu, Du\rangle = 0,$

(Pucci equations) $\quad \mathcal{P}^{\pm}(D^2u) - f(x) = 0.$

The corresponding references are [26] for the Monge–Ampère equation, [23] for the mean curvature flow equation, [2,30] for ∞-Laplace equation, and [12,13] for Pucci equations.

Comparison Principle

The comparison principle has been one of the main issues in the study of viscosity solutions since it implies their uniqueness.

The original idea to prove the comparison principle for first order PDEs is to double the number of variables with a clever choice of a penalization term. However, for second order PDEs, this technique did not directly work. The breakthrough was achieved by Jensen in [29].

The following lemma with matrix inequalities was originally formulated by Ishii.

Lemma 4 (Ishii's lemma) *For $u, v \in C(\Omega)$ and $\phi \in C^2(\Omega \times \Omega)$, if $(x,y) \to u(x) - v(y) - \phi(x,y)$ takes its maximum at $(\hat{x},\hat{y}) \in \Omega \times \Omega$, then for each $\mu > 1$, there are $X, Y \in S^n$ such that $(D_x\phi(\hat{x},\hat{y}), X) \in \overline{J}^{2,+}_{\Omega}u(\hat{x})$, $(-D_y\phi(\hat{x},\hat{y}), Y) \in \overline{J}^{2,-}_{\Omega}v(\hat{y})$, and*

$$-(\mu + \|A\|)\begin{pmatrix} I & O \\ O & I\end{pmatrix} \leq \begin{pmatrix} X & O \\ O & -Y\end{pmatrix} \leq A + \frac{1}{\mu}A^2\,, \quad (4)$$

where $A = D^2\phi(\hat{x},\hat{y}) \in S^{2n}$.

The proof of this lemma consists of twice differentiability of convex functions by Aleksandrov, Aleksandrov–Bakelman–Pucci type maximum principle for semi-convex functions by Jensen, and the so-called magic properties of sup-convolutions.

For a typical $\phi(x,y) = \frac{\mu}{2}|x-y|^2$, the above matrix inequalities become

$$-3\mu\begin{pmatrix} I & O \\ O & I\end{pmatrix} \leq \begin{pmatrix} X & O \\ O & -Y\end{pmatrix} \leq 3\mu\begin{pmatrix} I & -I \\ -I & I\end{pmatrix}\,. \quad (5)$$

In order to establish the comparison principle, the following two hypotheses on F are sufficient. The first one is monotonicity in the second variable:

$$\text{there exists} \quad \theta > 0 \quad \text{such that} \quad r \to F(x,r,p,X) - \theta r \quad \text{is non-decreasing} \quad (6)$$

for $(x,p,X) \in \Omega \times \mathbf{R}^n \times S^n$. It should be noted that one may take $\theta = 0$ in (6) if F is uniformly elliptic.

The next one is called the structure condition on F: by setting $\mathcal{M} := \{\omega\colon [0,\infty) \to [0,\infty) | \omega$ is continuous such

that $\omega(0) = 0\}$,

$$\begin{cases} \text{there exists } \omega \in \mathcal{M} \text{ such that } F(y, r, \mu(x-y), Y) \\ \leq F(x, r, \mu(x-y), X) + \omega(\mu|x-y|^2 + |x-y|) \\ \text{for } \mu > 1 x, y \in \Omega, r \in \mathbf{R}, X, Y \in S^n \text{ satisfying (5)}. \end{cases} \tag{SC}$$

Some observations related to (SC) below indicate that the condition (SC) is sufficiently reasonable.

Sufficient/Necessary Conditions for (SC):

(i) If F satisfies (SC), then F is elliptic.

(ii) If F is uniformly elliptic, and there is $\overline{\omega} \in \mathcal{M}$ such that

$$|F(x, r, p, X) - F(y, r, p, X)| \leq \overline{\omega}(|x-y|(1+|p|+\|X\|))$$

for $(x, y, r, p, X) \in \Omega \times \Omega \times \mathbf{R} \times \mathbf{R}^n \times S^n$, then F satisfies (SC).

(iii) In the case of HJB equations (3), if there are $M_0 \geq 0$ and $\omega_0 \in \mathcal{M}$ such that

$$\begin{cases} (a) & \max\{\|\sigma_a(x)\|, |g_a(x)|, |f_a(x)|, |c_a(x)|\} \leq M_0, \\ (b) & \max\{\|\sigma_a(x) - \sigma_a(y)\|, |g_a(x) - g_a(y)|\} \\ & \qquad \leq M_0|x-y|, \\ (c) & \max\{|f_a(x) - f_a(y)|, |c_a(x) - c_a(y)|\} \\ & \qquad \leq \omega_0(|x-y|) \end{cases} \tag{7}$$

for $(x, y, a) \in \Omega \times \Omega \times \mathcal{A}$, then F satisfies (SC).

It is obvious to give the sufficient condition corresponding to (7) for HJBI equations.

Due to the comparison principle below, it follows the uniqueness of viscosity solutions of (E) in Ω such that $u = h$ on $\partial\Omega$, where $h \in C(\partial\Omega)$.

Comparison Principle: Under hypotheses (6) and (SC), if $u \in C(\overline{\Omega})$ and $v \in C(\overline{\Omega})$ are, respectively, a viscosity subsolution and a viscosity supersolution of (E) in Ω, then $u \leq v$ on $\partial\Omega$ yields $u \leq v$ in Ω.

In order to understand how Ishii's lemma is used for the proof of the comparison principle, a simple second order PDE with variable coefficients to the second derivatives will be chosen.

Doubling the Number of Variables: Lipschitz continuous functions $\sigma_j = (\sigma_{ij})\colon \Omega \to \mathbf{R}^n$ $(j = 1, \ldots, m)$ are given; $|\sigma_j(x) - \sigma_j(y)| \leq M_0|x-y|$ $(x, y \in \Omega)$. By setting $A(x) = \frac{1}{2}\sum_{k=1}^m \sigma_{ik}(x)\sigma_{jk}(x)$, the following simple linear PDE is chosen as a typical example:

$$-\mathrm{trace}(A(x)D^2u) + u = 0 \quad \text{in } \Omega\,.$$

Here Ω is a bounded domain for simplicity. It is easy to verify that this PDE satisfies (6) and (SC).

A typical proof in viscosity solution theory is by contradiction. If $\theta := \sup_{\Omega}(u - v) > 0$, and $\hat{x} \in \overline{\Omega}$ is the point such that $(u - v)(\hat{x}) = \theta$, then $\hat{x} \in \Omega$ because $u \leq v$ on $\partial\Omega$. Here, a perturbation may allow $\hat{x}$ to be the strict maximum. Now, the key idea is to consider the function $(x, y) \in \overline{\Omega} \times \overline{\Omega} \to u(x) - v(x) - \frac{\mu}{2}|x-y|^2$ for $\mu > 1$. For $(x_\mu, y_\mu) \in \overline{\Omega} \times \overline{\Omega}$ being a maximum of this function, one can first observe that

$$\lim_{\mu\to\infty}(x_\mu, y_\mu) = (\hat{x}, \hat{x})\,, \quad \text{and} \quad \lim_{\mu\to\infty} \mu|x_\mu - y_\mu|^2 = 0\,.$$

Due to Ishii's lemma, there exist X_μ and $Y_\mu \in S^n$ such that

$$\begin{pmatrix} X_\mu & O \\ O & -Y_\mu \end{pmatrix} \leq 3\mu \begin{pmatrix} I & -I \\ -I & I \end{pmatrix}\,,$$

$(\mu(x_\mu - y_\mu), X_\mu) \in \overline{J}^{2,+}_{\Omega} u(x_\mu)$ and $(\mu(x_\mu - y_\mu), Y_\mu) \in \overline{J}^{2,-}_{\Omega} v(x_\mu)$. By setting $\xi_j := (\sigma_{1j}(x_\mu), \ldots, \sigma_{nj}(x_\mu))^t$ and $\eta_j := (\sigma_{1j}(y_\mu), \ldots, \sigma_{nj}(y_\mu))^t$ $(j = 1, \ldots, m)$, the above matrix inequality implies

$$\left\langle \begin{pmatrix} X_\mu & O \\ O & -Y_\mu \end{pmatrix} \begin{pmatrix} \xi_j \\ \eta_j \end{pmatrix}, \begin{pmatrix} \xi_j \\ \eta_j \end{pmatrix} \right\rangle \leq 3\mu|\xi_j - \eta_j|^2 \leq 3\mu M_0^2|x_\mu - y_\mu|^2\,.$$

Thus, by summing these over $j = 1, \ldots, m$, it follows that

$$\mathrm{trace}(A(x_\mu)X_\mu) - \mathrm{trace}(A(y_\mu)Y_\mu) \leq \frac{3}{2}\mu m M_0^2|x_\mu - y_\mu|^2.$$

Therefore, since the mappings $x \to u(x) - v(y_\mu) - \frac{\mu}{2}|x - y_\mu|^2$ and $y \to v(y) - u(x_\mu) + \frac{\mu}{2}|y - x_\mu|^2$, respectively, take their maximum and minimum at x_μ and $y_\mu \in \Omega$ for large $\mu > 1$, the definition yields

$$u(x_\mu) - v(y_\mu) \leq \mathrm{trace}(A(x_\mu)X_\mu) - \mathrm{trace}(A(y_\mu)Y_\mu) \leq \frac{3}{2}\mu m M_0^2|x_\mu - y_\mu|^2\,,$$

which gives $\theta \leq 0$ in the limit as $\mu \to \infty$. This is a contradiction.

Existence Results

There are two basic ways to show the existence of viscosity solutions. The first one is motivated by the classical Perron's method for harmonic functions.

For Perron's method below, it is necessary to introduce some relaxations of the notion. In the definition of viscosity subsolutions (resp., supersolution), u is replaced by u^*

(resp., u_*), which is the upper (resp., lower) semicontinuous envelope of u; for $u\colon \Omega \to \mathbf{R}$ and $x \in \Omega$,

$$u^*(x) = \lim_{\varepsilon\to 0} \sup_{\Omega\cap B_\varepsilon(x)} u \left(\text{resp., } u_*(x) = \lim_{\varepsilon\to 0} \inf_{\Omega\cap B_\varepsilon(x)} u\right).$$

Here and later, $B_\varepsilon(x)$ denotes the open ball with its center $x \in \mathbf{R}^n$ and its radius $\varepsilon > 0$.

For the reader's convenience, the definition of viscosity sub- and supersolutions is restated.

Definition 5 $u\colon \Omega \to \mathbf{R}$ is called a viscosity subsolution (resp., supersolution) of (E) in Ω if

$$F(x, u^*(x), D\phi(x), D^2\phi(x)) \le 0$$
$$(\text{reps., } F(x, u_*(x), D\phi(x), D^2\phi(x)) \ge 0)$$

whenever $(u^* - \phi)(x) = \sup_\Omega(u^* - \phi)$ (resp., $(u_* - \phi)(x) = \inf_\Omega(u_* - \phi)$ for $\phi \in C^2(\Omega)$ and $x \in \Omega$.

It should be remarked that for the function u, no hypotheses are assumed in this definition.

Remark 6 Ishii's lemma holds for $u \in USC(\Omega)$ and $v \in LSC(\Omega)$, where for $U \subset \mathbf{R}^n$, $USC(U)$ and $LSC(U)$ are, respectively, the sets of upper and lower semicontinuous functions in U. With these terminologies, u and v are, respectively, replaced by u^* and v_* in the comparison principle.

It is important to note that if the comparison principle holds for (E), and a viscosity solution u is continuous on $\partial\Omega$, then u is continuous in $\overline{\Omega}$ even though u is not initially supposed to belong to $C(\Omega)$.

Ishii (cf. [27]) established Perron's method to show the existence of viscosity solutions.

Perron's Method: If there is a viscosity subsolution $\xi \in USC(\overline{\Omega})$ and a viscosity supersolution $\eta \in LSC(\overline{\Omega})$ of (E) in Ω such that $\xi \le \eta$ in $\overline{\Omega}$, by setting

$$u(x) := \sup\left\{ v(x) \;\middle|\; \begin{array}{l} v \in USC(\overline{\Omega}) \text{ is a viscosity subsolution} \\ \text{of } (E) \text{ in } \Omega \text{ such that } \xi \le v \le \eta \text{ in } \overline{\Omega} \end{array} \right\},$$

then u is a viscosity solution of (E) in Ω; u^* and u_* are, respectively, a viscosity subsolution and a viscosity supersolution of (E) in Ω.

The above u does not belong to $USC(\Omega) \cup LSC(\Omega)$ in general.

The proof of Perron's method consists of two parts. One of them should be separately stated since this property is often useful.

Proposition 7 *For a non-empty set $S \subset USC(\Omega)$ (resp., $LSC(\Omega)$) of viscosity subsolutions (resp., supersolutions) of (E) in Ω, if $u(x) := \sup\{v(x) | v \in S\}$ (resp., $:= \inf\{v(x) | v \in S\}$) $(x \in \Omega)$ is locally bounded from above (resp., from below), then u is a viscosity subsolution (resp., supersolution) of (E) in Ω.*

Remark 8 Thanks to Perron's method, for the existence of viscosity solutions, it is enough to construct a viscosity subsolution ξ and a viscosity supersolution η of (E) in Ω satisfying $\xi \le \eta$ in $\overline{\Omega}$.

For instance, concerning the Dirichlet boundary value problem, for a given $h \in C(\partial\Omega)$, if there are ξ and η satisfying $\xi \le \eta$ in Ω, and $\xi = \eta = h$ on $\partial\Omega$, then the u constructed by Perron's method is a viscosity solution of (E) in Ω under $u = h$ on $\partial\Omega$. To find such a couple (ξ, η) of viscosity sub- and supersolutions, when F is uniformly elliptic and $\partial\Omega$ satisfies the so-called uniform exterior cone condition, it is possible to build them by using a barrier function at each $x_0 \in \partial\Omega$.

For HJB/HJBI equations, there is the other approach for the existence result. Concerning HJB Eqs. (3) for instance, it is known that the expected solution is the value function associated with a stochastic control problem. To avoid the boundary condition, the domain Ω for (3) is fixed by $\mathbf{R}^n$. For a given (progressively) measurable function $\alpha\colon [0,\infty) \to \mathcal{A}$, under assumption (7), it is known that one can solve the stochastic differential equation (abbreviated, SDE)

$$dX(t) = g_{\alpha(t)}(X(t))\mathrm{d}t + \sigma_{\alpha(t)}(X(t))\mathrm{d}W(t) \text{ for } \quad t > 0\,,$$

under $X(0) = x \in \mathbf{R}^n$, where W is the m-dimensional Brownian motion in a probability space. By denoting the solution of the above by $X(\cdot; x, \alpha)$, a cost functional is given by

$$J(x,\alpha) = E\left[\int_0^\infty \mathrm{e}^{-\int_0^t c_{\alpha(s)}(X(s;x,\alpha))\mathrm{d}s} \cdot f_{\alpha(t)}(X(t;x,\alpha))\mathrm{d}t \;\middle|\; X(0;x,\alpha) = x\right],$$

where E is the expectation. Then, the value function is defined by

$$u(x) = \inf_\alpha J(x,\alpha)\,, \tag{8}$$

where the infimum is taken over all measurable controls.

The DPP is a formula which the value function satisfies when $\inf\{c_a(x) | a \in \mathcal{A}, x \in \mathbf{R}^n\} > 0$.

Dynamic Programming Principle: For $\tau > 0$,

$$u(x) = \inf_{\alpha} E\left[\int_0^\tau e^{-\int_0^t c_{\alpha(s)}(X(s;x,\alpha))\mathrm{d}s} f_{\alpha(t)}(X(t;x,\alpha))\mathrm{d}t + e^{-\int_0^\tau c_{\alpha(t)}(X(t;x,\alpha))\mathrm{d}t} u(X(\tau;x,\alpha)) \;\middle|\; X(0;x,\alpha) = x \right].$$

Due to this DPP under (7), it is shown that u is a viscosity solution of (3) in $\mathbf{R}^n$ via the Itô formula.

Boundary Value Problems

In the study of degenerate elliptic/parabolic PDEs via viscosity solution method, there is a unified way to formulate boundary value problems of

$$\begin{cases} F(x,u,Du,D^2u) = 0 & \text{in } \Omega, \\ B(x,u,Du) = 0 & \text{on } \partial\Omega\,, \end{cases} \tag{9}$$

where $B\colon \partial\Omega \times \mathbf{R} \times \mathbf{R}^n \to \mathbf{R}$ is a given (continuous) function.

Typical examples of B in the literature of PDEs are

$$\begin{cases} B(x,r,p) = r - h(x) & \text{(Dirichlet condition)}, \\ B(x,r,p) = \langle p, \gamma(x)\rangle - h(x) & \text{(oblique condition)}, \end{cases}$$

where $h \in C(\partial\Omega)$, and $\gamma\colon \partial\Omega \to \mathbf{R}^n$ with $\langle n(x), \gamma(x)\rangle > 0$ (for $x \in \partial\Omega$). Here, $n(x)$ is the outward unit normal at $x \in \partial\Omega$. When $\gamma = n$, the oblique condition is called the Neumann condition.

To formulate the boundary value problem (9) in the viscosity sense, by setting

$$G(x,r,p,X) = \begin{cases} F(x,r,p,X) & (\text{for } x \in \Omega), \\ B(x,r,p) & (\text{for } x \in \partial\Omega) \end{cases}$$

in the definition of viscosity subsolutions (resp., supersolutions), Ω and F are replaced, respectively, by $\overline{\Omega}$ and G_* (resp., G^*). Here, G^* and G_* are, respectively, an upper and lower semicontinuous envelopes of G with respect to all the variables.

When F is continuous, $G^*(x,r,p,X) = G_*(x,r,p,X) = F(x,r,p,X)$ for $(x,r,p,X) \in \Omega \times \mathbf{R} \times \mathbf{R}^n \times S^n$. Moreover, if F is continuous up to $\partial\Omega$ in the first variable, and B is continuous, then one can verify that at $x \in \partial\Omega$,

$$G_*(x,r,p,X) = \min\{F(x,r,p,X), B(x,r,p)\}\,,$$
$$G^*(x,r,p,X) = \max\{F(x,r,p,X), B(x,r,p)\}$$

for $(r,p,X) \in \mathbf{R} \times \mathbf{R}^n \times S^n$.

Under this setting, the comparison principle means that if u and v are, respectively, a viscosity subsolution and a viscosity supersolution of (9) in the above sense, then $u \le v$ in $\overline{\Omega}$. To prove the comparison principle for (9), the key tool is often a good perturbation of the penalization term to avoid the boundary condition.

If the boundary condition is imposed on the whole of $\partial\Omega$, then it would be an over-determined (i. e. unsolvable) problem when (E) is only elliptic. However, in the above setting, the given boundary condition, $B(x,u,Du) = 0$ on $\partial\Omega$, is required only on a part of $\partial\Omega$, which is not a priori known.

This formulation has been successful particularly for oblique boundary value problems. For more information on the boundary value problems, [18] and its references are suggested.

State Constraint Problems: There is a sort of boundary condition arising in optimal control problems. However, it seems that this boundary condition is not derived from physical applications.

When the infimum in (8) is taken over all controls α such that $X(t;x,\alpha) \in \overline{\Omega}$ (when $\overline{\Omega} \neq \mathbf{R}^n$) for all $t \ge 0$, it follows that the value function u is a viscosity supersolution of (3) in $\overline{\Omega}$.

Thus, to formulate the state constraint problem of (E), the supersolution property on $\partial\Omega$ is regarded as a boundary condition; $u\colon \overline{\Omega} \to \mathbf{R}$ is called a viscosity supersolution (resp., subsolution) of the state constraint problem of (E) in Ω if it is a viscosity supersolution (resp., subsolution) of

$$F(x,u,Du,D^2u) = 0 \quad \text{in } \overline{\Omega} \quad (\text{resp., } \Omega)\,.$$

The comparison principle for the state constraint problem was proved in [42] for first order PDEs.

Asymptotic Analysis

Besides the existence and uniqueness of viscosity solutions, the remaining property in the well-posedness is the stability of those, which justifies that when the PDE is perturbed a little bit, the solution may change only a little.

In order to state the stability of viscosity solutions, Barles–Perthame in [8] introduced the half relaxed limits. For functions $u_\varepsilon\colon \Omega \to \mathbf{R}$ for $\varepsilon > 0$ and $x \in \Omega$, $\overline{u}(x) = \lim_{\varepsilon\to 0} \sup^* u_\varepsilon(x)$ (resp., $\underline{u}(x) = \lim_{\varepsilon\to 0} \inf_* u_\varepsilon(x)$) are defined by

$$\lim_{\varepsilon\to 0} \sup\{u_\delta(y) | 0 < \delta < \varepsilon, y \in B_\varepsilon(x)\}$$
$$\left(\text{resp., } \lim_{\varepsilon\to 0} \inf\left\{u_\delta(y) | 0 < \delta < \varepsilon, y \in B_\varepsilon(x)\right\}\right).$$

For given $F_\varepsilon : \Omega \times \mathbf{R} \times \mathbf{R}^n \times S^n \to \mathbf{R}$ ($\varepsilon > 0$), the stability of viscosity solutions of

$$F_\varepsilon(x, u, Du, D^2u) = 0 \quad \text{in } \Omega \tag{10}$$

is stated in the following manner.

Stability: If u_ε are viscosity subsolutions (resp., supersolutions) of (10) in Ω for $\varepsilon > 0$, and $\overline{u}$ (resp., $\underline{u}$) is finite in Ω, then it is a viscosity subsolution (resp., supersolution) of

$$\underline{F}(x, u, Du, D^2u) = 0 \quad (\text{resp., } \overline{F}(x, u, Du, D^2u) = 0) \quad \text{in } \Omega\,, \tag{11}$$

where $\underline{F}(x, r, p, X)$ (resp., $\overline{F}(x, r, p, X)$) is given by

$$\liminf_{\varepsilon\to 0} \left\{ F_\delta(y, s, q, Y) \;\middle|\; \begin{array}{c} y \in B_\varepsilon(x), |r - s| < \varepsilon, q \in B_\varepsilon(p), \\ \|X - Y\| < \varepsilon, 0 < \delta < \varepsilon \end{array} \right\}$$

$$\left(\text{resp., } \limsup_{\varepsilon\to 0} \left\{ F_\delta(y, s, q, Y) \;\middle|\; \begin{array}{c} y \in B_\varepsilon(x), |r - s| < \varepsilon, q \in B_\varepsilon(p), \\ \|X - Y\| < \varepsilon, 0 < \delta < \varepsilon \end{array} \right\}\right).$$

Thus, if $\overline{F} = \underline{F}$, and the (6) and (SC) holds for $\overline{F}$, then $\overline{u} = \underline{u}$ on $\partial\Omega$ yields $\overline{u} = \underline{u}$ in Ω, which implies the convergence of u_ε to $\overline{u} = \underline{u}$.

In order to understand how this stability is powerful, it is efficient to consider the singular perturbation of (E) with the vanishing viscosity term:

$$-\varepsilon \triangle u + F(x, u, Du, D^2u) = 0 \quad \text{in } \Omega\,, \tag{12}$$

where $\varepsilon > 0$, under some boundary condition. When F is not uniformly elliptic, it is often easier to solve (12) than (E).

If u_ε is a (classical) solution of (12) in Ω for $\varepsilon > 0$, and the uniform bound of u_ε is known, then the above stability result together with the comparison principle for (E) implies the convergence of u_ε. On the other hand, if one follows the standard argument, then it is necessary to get higher a priori estimates on u_ε so that u_ε has a limit (along a subsequence if necessary).

In applications from science and engineering, there are various singular perturbation problems where $\overline{F} \neq \underline{F}$. One typical example is the following.

Homogenization: When the PDE is formulated by

$$F\left(\frac{x}{\varepsilon}, u_\varepsilon, Du_\varepsilon, D^2u_\varepsilon\right) = 0 \quad \text{in } \mathbf{R}^n\,,$$

where the mapping $y \to F(y, r, p, X)$ satisfies the periodicity

$$F(y + z, r, p, X) = F(y, r, p, X)$$

for $(y, z, r, p, X) \in \mathbf{R}^n \times \mathbf{Z}^n \times \mathbf{R} \times \mathbf{R}^n \times S^n$, this PDE corresponds to the case when it has highly oscillating coefficients in a periodic medium.

Under certain hypothesis, e. g. coercivity for first order PDEs, uniform ellipticity for second order ones, or more general cases, one may show that a subsequence of u_ε converges to u, which is a viscosity solution of the effective PDE:

$$\hat{F}(u, Du, D^2u) = 0 \quad \text{in } \Omega\,,$$

which is called the cell problem. It is important to characterize $\hat{F}$ in the homogenization theory since this new PDE may contain macroscopic information.

While there is extensive literature on homogenization for PDEs of divergence type, less was known for those of non-divergence type. The so-called perturbed test function method was proposed to study homogenization for non-divergent PDEs (see [16]). Now, even for PDEs of non-divergence type, the study of homogenization has developed in periodic/almost periodic environments.

Recently, the research of homogenization in ergodic media has started in [14,37].

There are also some singular perturbation problems in phase transitions. Section 5 in [5] is a good survey for those and their front propagations (see also [9]).

Other Notions

There are some other notions of viscosity solutions. Indeed, although the viscosity solution method can be applied to various PDEs in a unified manner, one might lose some information in each specific problem.

Semicontinuous Viscosity Solutions: In the study of first order PDEs with convex Hamiltonian with respect to Du,

$$H(x, u, Du) = 0 \quad \text{in } \Omega\,, \tag{13}$$

where $H : \Omega \times \mathbf{R} \times \mathbf{R}^n \to \mathbf{R}$ is continuous, and $p \to H(x, r, p)$ is convex for $(x, r) \in \Omega \times \mathbf{R}$, Barron and Jensen proposed the notion of semicontinuous viscosity solutions. In fact, it sometimes happens that the expected solution is discontinuous while the standard comparison principle implies its continuity. In such a problem, it is necessary to

propose a better notion of weak solutions under which one guarantees the uniqueness of discontinuous solutions.

Definition 9 $u \in USC(\Omega)$ is called a semicontinuous viscosity solution of (12) in Ω if

$$H(x, \phi(x), D\phi(x)) = 0$$

whenever $\sup_\Omega(u - \phi) = (u - \phi)(x)$ for $\phi \in C^1(\Omega)$ and $x \in \Omega$.

It should be noted that the equality holds at x instead of the inequality in the standard definition.

It is known that under general setting, the value function in deterministic optimal control problems satisfies this definition. Moreover, the uniqueness of semicontinuous viscosity solutions is proved in [6].

L^p-viscosity Solutions: In the study of (linear) uniformly elliptic/parabolic PDEs of divergence type, there are two basic regularity theories; the Schauder estimates and L^p estimates.

Initiated by Caffarelli in [11], the regularity theory of viscosity solutions for fully nonlinear uniformly elliptic/parabolic PDEs was developed. Afterwards, when F is uniformly elliptic but the mapping $x \to F(x, r, p, X)$ is merely measurable, the notion of L^p-viscosity solutions of (E) in Ω was introduced in [13], roughly speaking, by using $W^{2,p}_{\text{loc}}(\Omega)$ instead of $C^2(\Omega)$ in the definition for $p > n/2$.

Definition 10 $u \in C(\Omega)$ is an L^p-viscosity subsolution (resp., supersolution) of (E) in Ω if

$$\operatorname{ess}\liminf_{y\to x} F(y, u(y), D\phi(y), D^2\phi(y)) \le 0$$

$$\left(\text{resp., } \operatorname{ess}\limsup_{y\to x} F(y, u(y), D\phi(y), D^2\phi(y)) \ge 0\right)$$

whenever $(u - \phi)(x) = \sup_\Omega(u - \phi)$ (resp., $\inf_\Omega(u - \phi)$) for $\phi \in W^{2,p}_{\text{loc}}(\Omega)$ and $x \in \Omega$.

As usual, $u \in C(\Omega)$ is called an L^p-viscosity solution of (E) in Ω if it is an L^p-viscosity sub- and supersolution of (E) in Ω.

A remarkable fact is that there are counter-examples even for linear uniformly elliptic PDEs for which uniqueness fails when the coefficient of the second derivatives is merely bounded. Thus, one may not expect the comparison principle to hold in general.

Here, several other notions are briefly mentioned.

Though the vanishing viscosity method works well for scalar conservation law, it is not clear if the viscosity solution method can be applied to it. There is a recent trial in the study of first order PDEs of divergence/non-divergence type in [22]. The main idea for the new notion is to consider the associated level set equations, and to use test surfaces instead of test functions in the definition.

There are recent works on stochastic PDEs by introducing a stochastic version of viscosity solutions in [35,36].

Future Directions

A list of areas where more research is expected via viscosity solution method is given.

Venttsel condition: There is a boundary condition where the second derivatives are involved; B contains D^2u-variables. In the case when $X \to B(x, r, p, X)$ is elliptic, (9) with this B is called the Venttsel boundary value problem. From a view point of viscosity solution theory, this case corresponds to that of discontinuous coefficients.

More regularity theory: For fully nonlinear uniformly elliptic/parabolic PDEs, there are still many open questions on regularity.

The convexity (or concavity) assumption in the D^2u-variables on F seems necessary to show that viscosity solutions are classical ones. Indeed, some counter-examples have been discovered in [39,40] without convexity assumption. On the contrary, for a simple uniformly elliptic HJBI equation, it is shown that a viscosity solution is indeed a classical one in [10].

For degenerate elliptic but specified PDEs, there must be some criterion for regularity. For example, [41] for ∞-Laplace equation etc.

Qualitative properties: One may expect qualitative properties, which classical (or even strong) solutions possess, to hold for viscosity solutions such as Aleksandrov–Bakelman–Pucci maximum principle, Liouville theorem, Hadamard theorem, Phragmén–Lindelöf principle, symmetry etc. There have appeared several researches but there are still gaps to be fulfilled.

Concerning on the other qualitative properties, the most important one among those must be the eigenvalue problems but only a few results have appeared (e. g. [34]).

Also, the blow-up phenomenon for fully nonlinear parabolic PDEs is an attractive topic but almost nothing is known except [21].

Degenerate equations from geometry: There have been several results which treat degenerate second order PDEs arising in differential geometry; e. g. Heisenberg Laplacian in [4,38].

Dynamical systems: The weak KAM theory is an active area nowadays in viscosity solution theory. The part "Hamilton–Jacobi equations and weak KAM theory" in this section is recommended for more information.

System of PDEs/nonlocal operators: Viscosity solution method can naturally apply to monotone systems of PDEs in optimal control problems and differential games such as weakly coupled systems and switching games. Also, in applications, PDEs involving non-local operators such as impulsive control problems have been studied. There is a general treatment for PDEs with non-local operators in [28].

Recently, [1] investigated Hamilton–Jacobi equations with some special non-local operator which comes from a dislocation dynamics. One should seek new systems of PDEs, and PDEs with other non-local operators in various applications.

Applications to mathematical finance: There are lots of fully nonlinear PDEs appearing in mathematical finance. However, many of those cannot be applied directly by viscosity solution method since the PDEs contain unbounded coefficients, the domain is often unbounded, etc. Since the goal in this field is to determine (approximate) optimal policies, it is necessary to study the regularity of viscosity solutions in each problem.

There have also been several books in this field (e. g. [31]).

Differential games: Although HJBI equations appeared as examples from differential games, no details are mentioned. One interesting question in differential games is how to formulate the boundary condition for the state constraint problems. There was a trial [32] but it is not clear if this formulation covers practical problems.

Control problems in ∞-dimensional spaces: If (stochastic) PDEs govern the underlying state equation in place of (SDE) to define the value function in (8), then the domain of the resulting HJB equations becomes an ∞-dimensional space (typically, Hilbert space).

The viscosity solution theory for ∞-dimensional spaces has been studied. However, much more research should be done for specific PDEs such as the Navier–Stokes equation (e. g. in [25]). Again [18] is suggested to find more information.

To finish this section, the reader should recall the words in the "User's Guide" [18].

> "the reader should not restrict his imagination to the borders we drew above."

Bibliography

Primary Literature

1. Alvarez O, Hoch P, Le Bouar Y, Monneau R (2006) Dislocation dynamics: short time existence and uniqueness of the solution. Arch Ration Mech Anal 85:371–414
2. Aronsson G, Crandall MG, Juutinen P (2004) A tour of the theory of absolutely minimizing functions. Bull Amer Math Soc 41:439–505
3. Bardi M, Capuzzo Dolcetta I (1997) Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman Equations. Birkhäuser, Boston
4. Bardi M, Mannuci P (2006) On the Dirichlet problem for non-totally degenerate fully nonlinear elliptic equations. Comm Pure Appl Anal 5:709–731
5. Bardi M, Crandall MG, Evans LC, Soner HM, Souganidis PE (1997) Viscosity Solutions and Applications. Springer, Berlin
6. Barles G (1993) Discontinuous viscosity solutions of first-order Hamilton–Jacobi equations: A guided visit. Nonlinear Anal 20:1123–1134
7. Barles G (1994) Solutions de Viscosité des Équations de Hamilton–Jacobi. Springer, Berlin
8. Barles G, Perthame B (1987) Discontinuous solutions of deterministic optimal stopping-time problems. Model Math Anal Num 21:557–579
9. Barles G, Souganidis PE (1998) A new approach to front propagation problems: theory and applications. Arch Ration Mech Anal 141:237–296
10. Cabré X, Caffarelli LA (2003) Interior $C^{2,\alpha}$ regularity theory for a class of nonconvex fully nonlinear elliptic equations. J Math Pures Appl 83:573–612
11. Caffarelli LA (1989) Interior a priori estimates for solutions of fully non-linear equations. Ann Math 130:180–213
12. Caffarelli LP, Cabré X (1995) Fully Nonlinear Elliptic Equations. Amer Math Soc, Providence
13. Caffarelli LA, Crandall MG, Kocan M, Święch A (1996) On viscosity solutions of fully nonlinear equations with measurable ingredients. Comm Pure Appl Math 49:365–397
14. Caffarelli LA, Souganidis PE, Wang L (2005) Homogenization of fully nonlinear, uniformly elliptic and parabolic partial differential equations in stationary ergodic media. Comm Pure Appl Math 58:319–361
15. Crandall MG, Lions P-L (1981) Condition d'unicité pour les solutions generalisées des équations de Hamilton–Jacobi du premier ordre. CR Acad Sci Paris Sér I Math 292:183–186
16. Crandall MG, Lions P-L (1983) Viscosity solutions of Hamilton–Jacobi equations. Tran Amer Math Soc 277:1–42
17. Crandall MG, Evans LC, Lions P-L (1984) Some properties of viscosity solutions of Hamilton–Jacobi equations. Trans Amer Math Soc 282:487–435
18. Crandall MG, Ishii H, Lions PL (1992) User's guide to viscosity solutions of second order partial differential equations. Bull Amer Math Soc 27:1–67
19. Evans LC (1992) Periodic homogenization of certain fully nonlinear partial differential equations. Proc Roy Soc Edinb 120:245–265
20. Fleming WH, Soner HM (1993) Controlled Markov Processes and Viscosity Solutions. Springer, Berlin
21. Friedman A, Souganidis PE (1986) Blow-up of solutions of Hamilton–Jacobi equations. Comm Partial Differ Equ 11:397–443

22. Giga Y (2002) Viscosity solutions with shocks. Comm Pure Appl Math 55:431–480
23. Giga Y (2006) Surface Evolutions Equations: A Level Set Approach. Birkäuser, Basel
24. Gilbarg D, Trudinger NS (1983) Elliptic Partial Differential Equations of Second Order. Springer, New York
25. Gozzi F, Sritharan SS, Święch A (2005) Bellman equations associated to optimal control of stochastic Navier–Stokes equations. Comm Pure Appl Math 58:671–700
26. Gutiérrez CE (2001) The Monge–Ampère Equation. Birkhäuser, Boston
27. Ishii H (1987) Perron's method for Hamilton–Jacobi equations. Duke Math J 55:369–384
28. Ishii H, Koike S (1993/94) Viscosity solutions of functional-differential equations. Adv Math Sci Appl 3:191–218
29. Jensen R (1988) The maximum principle for viscosity solutions of fully nonlinear second order partial differential equations. Arch Rat Mech Anal 101:1–27
30. Jensen R (1993) Uniqueness of Lipschitz extensions: minimizing the sup norm of the gradient. Arch Ration Mech Anal 123:51–74
31. Karatzas I, Shreve SE (1998) Methods of Mathematical Finance. Springer, New York
32. Koike S (1995) On the state constraint problem for differential games. Indiana Univ Math J 44:467–487
33. Koike S, Święch A (2004) Maximum principle for fully nonlinear equations via the iterated comparison function method. Math Ann 339:461–484
34. Lions P-L (1983) Bifurcation and optimal stochastic control. Nonlinear Anal 7:177–207
35. Lions PL, Souganidis PE (1998) Fully nonlinear stochastic partial differential equations. CR Acad Sci Paris 326:1085–1092; 326:753–741
36. Lions PL, Souganidis PE (2000) Uniqueness of weak solutions of fully nonlinear stochastic partial differential equations. CR Acad Sci Paris 331:783–790
37. Lions PL, Souganidis PE (2003) Correctors for the homogenization of Hamilton–Jacobi equations in the stationary ergodic setting. Comm Pur Appl Math 56:1501–1524
38. Manfredi JJ (2002) A version of the Hopf–Lax formula in the Heisenberg group. Comm Partial Differ Equ 27:1139–1159
39. Nadirashvili N, Vlăduţ S (2007) Nonclassical solutions of fully nonlinear elliptic equations. Geom Funct Anal 17:1283–1296
40. Nadirashvili N, Vlăduţ S (2008) Singular viscosity solutions to fully nonlinear elliptic equations. J Math Pures Appl 89(9):107–113
41. Savin O (2005) C^1 regularity for infinity harmonic functions in two dimensions. Arch Ration Mech Anal 176:351–361
42. Soner MH (1986) Optimal control with state-space constraint I. SIAM J Control Optim 24:552–562

Books and Reviews

Bellman R (1957) Dynamic Programming. Princeton University Press, Princeton

Evans LC, Gangbo W (1999) Differential Equations Methods for the Monge–Kantrovich Mass Transfer Problem. Amer Math Soc, Providence

Fleming WH, Rishel R (1975) Deterministic and Stochastic Optimal Control. Springer, New York

Koike S (2004) A Beginner's Guide to the Theory of Viscosity Solutions. Math Soc Japan, Tokyo. Corrections: http://133.38.11.17/lab.jp/skoike/correction.pdf

Lions P-L (1982) Generalized Solutions of Hamilton–Jacobi Equations. Pitman, Boston

Maugeri A, Palagachev DK, Softova LG (2000) Elliptic and Parabolic Equations with Discontinuous Coefficients. Wiley, Berlin

Non-linear Stochastic Partial Differential Equations

Giuseppe Da Prato
Scuola Normale Superiore, Pisa, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
General Problems and Results
Specific Equations
Future Directions
Bibliography

Glossary

Evolution equations Let H be a Hilbert (or Banach) space, $T > 0$ and f a mapping from $[0, T] \times H$ into H. An equation of the form

$$u'(t) = f(t, u(t)) \ , \quad t \in [0, T] \ , \qquad (*)$$

is called an *abstract evolution equation*. If H is finite-dimensional we call $(*)$ an *ordinary evolution equation*, whereas if H is infinite-dimensional and f is a differential operator we call $(*)$ a *partial differential evolution equation*.

Continuous stochastic process Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A *continuous stochastic process* (with values in H) is a family of (H-valued) random variables $(X(t) = X(t, \omega))_{t \geq 0}$ $(\omega \in \Omega)$ such that $X(\cdot, \omega)$ is continuous for $\mathbb{P}$-almost all $\omega \in \Omega$.

Brownian motion A real *Brownian motion* $B = (B(t))_{t \geq 0}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ is a continuous real stochastic process in $[0, +\infty)$ such that (1) $B(0) = 0$ and for any $0 \leq s < t, B(t) - B(s)$ is a real Gaussian random variable with mean 0 and covariance $t - s$, and (2) if $0 < t_1 < \cdots < t_n$, the random variables, $B(t_1),\ B(t_2) - B(t_1), \ldots, B(t_n) - B(t_{n-1})$ are independent.

Cylindrical Wiener process A *cylindrical Wiener process* in a Hilbert space H is a process of the form

$$W(t) = \sum_{k=1}^{\infty} e_k \beta_k(t) , \quad t \geq 0 ,$$

where (e_k) is a complete orthonormal system in H and (β_k) a sequence of mutually independent standard Brownian motions in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Noise Let $W(t)$ be a cylindrical Wiener process in a Hilbert space H. A noise is an expression of the form $BW(t)$, where $B \in L(H)$. If $B = I$ (the identity operator in H) the noise is called *white*.

Stochastic dynamical system This is a system governed by a partial differential evolution equation perturbed by noise.

Definition of the Subject

Several nonlinear partial differential evolution equations can be written as abstract equations in a suitable Hilbert space H (we shall denote by $|\cdot|$ the norm and by $\langle\cdot,\cdot\rangle$ the scalar product in H) of the following form [78]:

$$\begin{cases} \dfrac{\mathrm{d}X(t)}{\mathrm{d}t} = AX(t) + b(X(t)) , \quad t \geq 0 , \\ X(0) = x \in H , \end{cases} \tag{1}$$

where $A\colon D(A) \subset H \to H$ is the infinitesimal generator of a strongly continuous semigroup e^{tA} of linear bounded operators in H and $b\colon D(b) \subset H \to H$ is a nonlinear mapping.

In order to take into account unpredictable random perturbations, one is led to add to (1) a term of the form $\sigma(X(t))dW(t)$, where $\sigma\colon D(\sigma) \subset H \to L(H)$ ($L(H)$ is the space of all linear operators from H into itself) and $W(t)$ a cylindrical Wiener process in H. Then (1) is replaced by the following stochastic partial differential equation (SPDEs)

$$\begin{cases} \mathrm{d}X(t) = (AX(t) + b(X(t)))\mathrm{d}t + \sigma(X(t))\mathrm{d}W(t) , \\ X(0) = x \in H , \end{cases} \tag{2}$$

whose precise meaning will be explained later.

The importance of the SPDEs is due to the fact that in the modelization of several phenomena, (2) is often more realistic than (1).

The study of SPDEs started in the 1970s and 1980s and interest is still growing. Among the early contributions we mention [7] for Navier–Stokes equations, [66] for SPDEs of monotone type, [79] for compactness methods, [57] for variational methods and [31].

Two approaches (essentially equivalent) are used for studying SPDEs: the first is based on the variational theory of partial differential equations and the second is based on the theory of strongly continuous semigroups of operators. In this article we are concerned with the latter [26,27]. For a recent introduction to the variational method see [71].

It is possible to consider perturbations of a partial differential equation by stochastic processes different from Brownian motion as jump processes, but we shall not consider this subject here.

Introduction

In this section we are concerned with (2) in the Hilbert space H, that is, with a partial differential equation of the form (1) perturbed by the noise $\sigma(X(t))dW(t)$, where $W(t)$ is a cylindrical Wiener process in H. We recall that $W(t)$ can be defined as follows:

$$W(t) = \sum_{k=1}^{\infty} e_k \beta_k(t) , \tag{3}$$

where e_k is a complete orthonormal system in H (often the system of eigenfunctions of A) and β_k is a sequence of mutually independent standard Brownian motions in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Intuitively $W(t)$ represents a random (white) perturbation which has the same intensity in all direction of the orthonormal basis (e_k).

Remark 1 Definition (3) is formal since

$$\mathbb{E}[|W(t)|^2] = \sum_{k=1}^{\infty} t = +\infty \quad \forall\ t > 0 .$$

However, it is possible to define rigorously $W(t)$ in a suitable space larger than H [26]. We shall see in the following how to deal with $W(t)$.

The main problems we shall consider are the following:

(i) Existence and uniqueness of solutions $X(t)$ of (2).
(ii) Asymptotic behavior of $X(t)$ as $t \to \infty$.

For the sake of brevity we shall limit ourselves to stochastic perturbations of the special form $\sigma(X(t))dW(t) = BdW(t)$, where $B \in L(H)$. In this case we say that the noise involved in (2) is *additive*. The general case (*multiplicative noise*) is important and we shall give some references later. In particular a multiplicative noise is required when for physical reasons the solution $X(t)$ must belong to some convex subset of H (for instance, $X \geq 0$) [6].

In Sect. "General Problems and Results" we shall present an overview of the main problems concerning SPDEs. Section "Specific Equations" is devoted to some significant applications. We will not give complete proofs of all assertions but only a sketch in order to give an idea of the tools which are involved. The interested reader can look at the corresponding references.

We shall concentrate on some SPDEs only. We present in the following three examples the corresponding deterministic problems.

Example 2 (reaction–diffusion equations). Consider the following partial differential equation in a bounded subset $\mathcal{O}$ of $\mathbb{R}^N$ with regular boundary $\partial\mathcal{O}$:

$$\begin{cases} \partial_t X(t,\xi) = \Delta_\xi X(t,\xi) + p(X(t,\xi)) \,, & \xi \in \mathcal{O}, t > 0 \,, \\ X(t,\xi) = 0 \,, & t > 0, \xi \in \partial\mathcal{O} \,, \\ X(0,\xi) = x(\xi) \,, & \xi \in \mathcal{O}, x \in H \,, \end{cases} \tag{4}$$

where Δ_ξ is the Laplace operator and p is a polynomial of degree $d > 1$ with negative leading coefficient.

Let us write problem (4) as an abstract equation of the form (1) in the Hilbert space $H = L^2(\mathcal{O})$ (the space of all equivalence classes of square integrable real functions on $\mathcal{O}$ with respect to the Lebesgue measure).

For this purpose, we denote by A the realization of the Laplace operator with Dirichlet boundary conditions (other boundary conditions such as Neumann or Ventzell could be considered as well),

$$\begin{cases} Ax = \Delta_\xi x \,, \quad x \in D(A) \,, \\ D(A) = H^2(\mathcal{O}) \cap H_0^1(\mathcal{O}) \,, \end{cases}$$

and define

$$b(x)(\xi) = p(x(\xi)) \,, \quad x \in D(b) = L^{2d}(\mathcal{O}) \,.$$

For $k = 1, 2$ we denote by $H^k(\mathcal{O})$ the Sobolev space consisting of all functions which belong to $L^2(\mathcal{O})$ together with their distributional derivatives of order up to k. $H_0^1(\mathcal{O})$ is the subspace of those functions in $H^1(\mathcal{O})$ which vanish at the boundary $\partial\mathcal{O}$ of $\mathcal{O}$.

Example 3 (Burgers equation) Consider the equation in $[0, 2\pi]$ with periodic boundary conditions

$$\begin{cases} \partial_t X(t,\xi) = \partial_\xi^2 X(t,\xi) + \frac{1}{2}\partial_\xi (X^2(t,\xi)), \\ \qquad\qquad \xi \in [0,2\pi], t > 0, \\ X(t,0) = X(t,2\pi) \,, \quad \partial_\xi X(t,0) = \partial_\xi X(t,2\pi), \ t > 0, \\ X(0,\xi) = x(\xi) \,, \quad \xi \in [0,2\pi] \,. \end{cases} \tag{5}$$

Problem (5) can be written in the abstract form (1) setting $H = L^2(0, 2\pi)$,

$$Ax = D_\xi^2 x \,, \quad x \in D(A) = H_\#^2(0, 2\pi) \,,$$

and

$$b(x) = \tfrac{1}{2}\, D_\xi(x^2) \,, \quad x \in D(b) = H_\#^1(0, 2\pi) \,,$$

where

$$H_\#^1(0, 2\pi) = \{x \in H^1(0, 2\pi) : \ x(0) = x(2\pi)\}$$

and

$$H_\#^2(0, 2\pi) = \\ \{x \in H^2(0, 2\pi) \colon x(0) = x(2\pi), \ \partial_\xi x(0) = \partial_\xi x(2\pi)\}.$$

Example 4 (2D Navier–Stokes equation). Consider the equation in the square $\mathcal{O} := [0, 2\pi] \times [0, 2\pi]$.

$$\begin{cases} \mathrm{d}Z = (\Delta_\xi Z - Z + D_\xi Z \cdot Z)\mathrm{d}t + \nabla p \mathrm{d}t \\ \qquad\qquad \xi \in \mathcal{O}, \ t > 0, \\ \operatorname{div} Z = 0 \quad \xi \in \mathcal{O} \,, \ t > 0 \,, \\ Z(t,\cdot) \text{ is periodic with period } 2\pi \,, \\ Z(0,\xi) = z(\xi) \quad \xi \in \mathcal{O} \,. \end{cases} \tag{6}$$

The unknown $Z = (Z_1, Z_2)$ represents the velocity and p the pressure of the fluid; $\operatorname{div} Z$ is the divergence of Z. We denote by H the space of all square integrable divergence free vectors,

$$H = \{Z = (Z_1, Z_2) \in (L^2(0, 2\pi))^2 \colon \ \operatorname{div} Z = 0\} \,,$$

and by $\mathcal{P}$ the orthogonal projector of $(L^2(0, 2\pi))^2$ onto H.

Then setting $X(t,x) = \mathcal{P}Z(t,x)$, we obtain the following problem (where the pressure p has disappeared)

$$\begin{cases} \mathrm{d}X = \mathcal{P}(\Delta X - X) + \mathcal{P}(D_\xi X \cdot X) \quad \xi \in \mathcal{O} \,, \ t > 0 \,, \\ X(t,\xi) \text{ is periodic with period } 2\pi \,, \quad t > 0 \\ X(0,\xi) = x \quad \xi \in \mathcal{O} \,. \end{cases}$$

Let us define the *Stokes* operator $A \colon D(A) \to H$ setting

$$Ax = \mathcal{P}(\Delta_\xi x - x) \,, \quad x \in D(A) = H_\#^2(\mathcal{O})$$

and the nonlinear operator b setting

$$b(x) = \mathcal{P}(D_\xi x \cdot x) \,, \quad x \in H_\#^1(\mathcal{O})$$

(b is more commonly written $b(x) = \mathcal{P}(x \cdot \nabla x)$). Here $H_\#^k(\mathcal{O})$, $k = 1, 2$ are Sobolev spaces with periodic boundary conditions. Now, problem (5) can be written in the form (1).

We have to say that many other interesting equations have been studied recently. We mention among them the following (without claiming to present an exhaustive list).

- Cahn–Hilliard equations [21,38].
- Second-order linear SPDEs and Filtering equations; see [56,74] and references therein.
- Ginzburg–Landau equations [51,52,53].
- Kortweg–de Vries equation [36,37].
- Mathematical biology and Fleming–Viot model [48, 77].
- Mathematical finance [64].
- Nonlinear Schrödinger equations [32,33,34,35].
- Porous media equations [4,30,73].
- Stochastic quantization [2,23,25,55,59,62].
- Wave equations [5,18,19,62,63,69,75].

General Problems and Results

Well-Posedness of a Stochastic Partial Differential Equation

We are here concerned with (2). The first problem is to prove the existence and uniqueness of solutions (to be defined in an appropriate way). We notice that it is not convenient to look for existence of strong solutions $X(t)$ (as for the ordinary stochastic equations) such that

$$X(t) = x + \int_0^t \Big[AX(s) + b(X(s))\Big]\mathrm{d}s + BW(t)\,, \quad t \geq 0\,.$$

This definition will require too much regularity for the solution, which does not belong to $D(A) \cap D(b)$ in general. The concept of a weaker solution has to be defined for each specific problem; see Sect. "Specific Equations".

Assume now that the existence of a unique solution $X(t,x)$ (in a suitable sense) of (2) has been proved for any $x \in H$. Then, given a real function φ in H, one is interested in the evolution in time of $\mathbb{E}[\varphi(X(t,x))]$ (the expectation of $\varphi(X(t,x))$). φ can be interpreted as an *observable*; its physical meaning could be, for instance, the mean velocity, pressure or temperature of a fluid. This leads to introduce the concept of *transition semigroup*

$$P_t\varphi(x) = \mathbb{E}[\varphi(X(t,x))]\,, \quad t \geq 0,\ x \in H,\ \varphi \in B_b(H)\,, \tag{7}$$

where $B_b(H)$ is the space of all mappings $\varphi\colon H \to \mathbb{R}$ which are bounded and Borel. It is a Banach space with the supremum norm

$$\|\varphi\|_\infty = \sup_{h\in H} |\varphi(x)|\,, \quad \varphi \in B_b(H)\,.$$

We denote by $C_b(H)$ the closed subspace of $B_b(H)$ of all uniformly continuous and bounded functions.

It is not difficult to see that P_t enjoys the semigroup property $P_{t+s} = P_tP_s$, $t,s \geq 0$, thanks to the Markovianity of the solution $X(t,x)$ to (2). However, P_t is not a strongly continuous semigroup on $C_b(H)$ in general (even when H is finite-dimensional and (2) is an ordinary differential equation).

Formally, the function $u(t,x) = P_t\varphi(x)$ is the unique solution to the following (Kolmogorov) parabolic equation:

$$\begin{cases} \dfrac{\mathrm{d}u}{\mathrm{d}t}(t,x) = K_0u(t,x)\,, \\ u(0,x) = \varphi(x)\,, \end{cases} \tag{8}$$

where K_0 is the linear differential operator

$$K_0\varphi(x) = \tfrac{1}{2}\,\mathrm{Tr}\Big[BB^*D_x^2\varphi(x)\Big] + \langle Ax + b(x), D_x\varphi(x)\rangle\,, \quad x \in D(A)\cap D(b)\,, \tag{9}$$

where

$$\mathrm{Tr}\big[BB^*D_x^2\varphi(x)\big] = \sum_{h,k=1}^{\infty} \big\langle D_x^2\varphi(x)B^*e_k\,, B^*e_k\big\rangle\,,$$

and (e_k) is a complete orthonormal system on H. Notice that $K_0\varphi$ is only defined for φ of class C^2 such that the series above is convergent.

Though the semigroup P_t is not strongly continuous one can define its infinitesimal generator K, following [72],

$$K\varphi(x) = \lim_{h\to 0}\frac{1}{h}\,(P_t\varphi(x) - \varphi(x))\,, \quad x \in H\,,$$

provided the limit exists and the following condition holds

$$\sup_{h\in(0,1]}\frac{1}{h}\,\|P_t\varphi - \varphi\|_\infty < +\infty\,.$$

Then an interesting problem consists in clarifying the relationship between the abstract operator K and the concrete differential operator K_0.

Let us list some important concepts related to P_t.

- P_t is called *Feller* if $P_t\varphi \in C_b(H)$ for any $\varphi \in C_b(H)$ and any $t \geq 0$.
- P_t is called *strong Feller* if $P_t\varphi \in C_b(H)$ for any $\varphi \in B_b(H)$ and any $t > 0$.
  The strong Feller property is related to a smoothing effect of the transition semigroup P_t and to the hypoellipticity of the Kolmogorov operator K_0 (when H is finite-dimensional) [75].

- P_t is called *irreducible* if $P_t 1_I(x) > 0$ for all $x \in H$ and all open sets I of H, where 1_I is the characteristic function of I ($1_I(x) = 1$ if $x \in I$, $1_I(x) = 0$ if $x \notin I$).

Irreducibility of P_t is related to the null-controllability of the deterministic system (that is, if for each $x \in H$ and $T > 0$ there is a control u such that $x(T) = 0$).

$$\begin{cases} \dfrac{\mathrm{d}x(t)}{\mathrm{d}t} = Ax(t) + b(x(t)) + Bu(t)\,, \quad t \geq 0\,, \\ x(0) = x \in H\,, \end{cases}$$

where u is a control; see, e. g Sects. 7.3 and 7.4 in [27].

Finally, we denote by $\pi_t(x, dy)$ the *law* of $X(t, x)$, so that the following integral representation for P_t,

$$P_t\varphi(x) = \int_H \varphi(y)\pi_t(x, \mathrm{d}y)\,, \quad \forall t > 0\,, x \in H\,, \quad \varphi \in B_b(H)\,, \tag{10}$$

holds.

Asymptotic Behavior of the Transition Semigroup

In order to study the asymptotic behavior of $P_t\varphi$, where $\varphi \in B_b(H)$, the concept of *invariant measure* is crucial. We say that a Borel probability measure ν on H is invariant for P_t if

$$\int_H P_t\varphi(x)\nu(\mathrm{d}x) = \int_H \varphi(x)\nu(\mathrm{d}x)\,, \quad \forall\, \varphi \in C_b(H)\,.$$

Notice that if ν is invariant and η is a random variable whose law is ν then the solution $X(t, \eta)$ to (2) with initial datum η is stationary.

In order to prove the existence of invariant measures an important tool is the following Krylov–Bogoliubov theorem; see, e. g., Theorem 3.1.1 in [27].

Theorem 5 *Assume that P_t is Feller and that for some $x \in H$ the family of probability measures $(\pi_t(x, dy))_{t\geq 0}$ is tight (that is, for any $\epsilon > 0$ there exists a compact set K_ϵ in H such that $\pi_t(x, K_\epsilon) > 1 - \epsilon$ for all $t \geq 0$). Then there exists an invariant measure for P_t.*

Assume now that ν is an invariant measure for P_t and let $\varphi \in C_b(H)$. Then P_t can be uniquely extended to a contraction semigroup in $L^2(H, \nu)$. In fact, by (10) and the Hölder inequality we have

$$[P_t\varphi(x)]^2 \leq P_t(\varphi^2)(x)\,, \quad t \geq 0\,, \quad x \in H\,.$$

Integrating this inequality with respect to ν over H and taking into account that $\int_H P_t(\varphi^2)(x)\nu(\mathrm{d}x) = \int_H \varphi^2(x)\nu(\mathrm{d}x)$ by the invariance of ν yields

$$\int_H [P_t\varphi(x)]^2\nu(\mathrm{d}x) \leq \int_H \varphi^2(x)\nu(\mathrm{d}x)\,.$$

Since $C_b(H)$ is dense in $L^2(H, \nu)$, this inequality can be extended to any function of $L^2(H, \nu)$. This proves that P_t can be uniquely extended to a contraction semigroup in $L^2(H, \nu)$.

Moreover, the following Von Neumann theorem holds [70].

Theorem 6 *Let ν be an invariant measure for P_t. Then for any $\varphi \in L^2(H, \nu)$ there exists the limit*

$$\lim_{T\to+\infty} \frac{1}{T}\int_0^T P_t\varphi(x)dt := \Pi\varphi\,,$$

where Π is a projection operator on

$$\Sigma := \{\varphi \in L^2(H, \nu)\colon P_t\varphi = \varphi,\ \forall\ t \geq 0\}\,.$$

If moreover

$$\lim_{T\to+\infty} \frac{1}{T}\int_0^T P_t\varphi(x)\mathrm{d}t = \int_H \varphi(y)\nu(dy) \quad \text{in } L^2(H, \nu),$$

then ν is called *ergodic*.

As in the deterministic case, ergodicity is a very important property of a stochastic dynamical system because it allows us to compute averages with respect to t in terms of averages with respect to $x \in H$ ("temporal" averages with "spatial" averages).

When the more stringent condition

$$\lim_{T\to+\infty} P_t\varphi(x)\mathrm{d}t = \int_H \varphi(y)\nu(\mathrm{d}y)\,, \quad \forall\, x \in H\,,$$

holds, we say that the measure ν is *strongly mixing*. In this direction we recall the following Doob's theorem; see e. g., Theorem 4.2.1 in [27].

Theorem 7 *Assume that P_t is irreducible and strongly Feller. Then P_t possesses at most one invariant measure which is ergodic and strongly mixing.*

Other Important Problems

Here we mention some additional problem which we cannot treat for space reasons. The first one concerns the so-called *small noise*. Often one is interested in studying stochastic perturbations of (1) of the form

$$\begin{cases} \mathrm{d}X(t) = (AX(t) + b(X(t)))\mathrm{d}t + \epsilon\mathrm{d}W(t)\,, \\ X(0) = x \in H\,, \end{cases} \tag{11}$$

where $\epsilon > 0$. In fact it can happens that (11) has a unique invariant measure μ_ε, whereas the corresponding deterministic system (1) has more. Then studying the limit points of μ_ε as $\varepsilon \to 0$ will help to select the "significant" invariant measures of (2) [39]. This problem also leads to study *large deviations* problems. Concerning Smoluchowski–Kramers approximations for parabolic SPDEs we mention [17].

Specific Equations

Ornstein–Uhlenbeck Equations

We are here concerned with the following stochastic differential equation:

$$\begin{cases} \mathrm{d}X(t) = AX(t)\mathrm{d}t + B\mathrm{d}W(t)\,, & t \geq 0 \\ X(0) = x \in H\,, \end{cases} \tag{12}$$

where $A\colon D(A) \subset H \to H$ is the infinitesimal generator of a strongly continuous semigroup e^{tA} and $B\colon H \to H$ is linear-bounded.

Formally the solution of (12) is given by the variation of constants formula

$$X(t,x) = \mathrm{e}^{tA}x + W_A(t)\,, \quad t \geq 0\,, \tag{13}$$

where the process $W_A(t)$, called *stochastic convolution*, is given by

$$W_A(t) = \int_0^t \mathrm{e}^{(t-s)A}B\mathrm{d}W(s)\,, \quad t \geq 0\,.$$

By a formal computation we see that

$$\mathbb{E}|W_A(t)|^2 = \sum_{k=1}^{\infty}\int_0^t |\mathrm{e}^{(t-s)A}Be_k|^2\mathrm{d}s = \mathrm{Tr}\,Q_t\,,$$

where

$$Q_t x = \int_0^t \mathrm{e}^{sA}C\mathrm{e}^{sA^*}x\mathrm{d}s\,, \quad x \in H,\ t \geq 0\,,$$

and $C = BB^*$. Consequently, though the definition (3) of the cylindrical Wiener process $W(t)$ is formal (see Remark 1), the stochastic convolution defined by the series

$$W_A(t) = \sum_{k=1}^{\infty}\int_0^t \mathrm{e}^{(t-s)A}Be_k\mathrm{d}\beta_k(s) \tag{14}$$

is meaningful and convergent in $L^2(\Omega, \mathcal{F}, \mathbb{P}; H)$ for all $t \geq 0$, provided $\mathrm{Tr}\,Q_t < +\infty$ for all $t > 0$. We shall assume that this hypothesis holds from now on.

It is easy to see that $W_A(t)$ is a Gaussian random variable N_{Q_t} in H. (A Borel probability measure μ in H is said to be *Gaussian* with *mean* $x \in H$ and covariance Q, where $Q \in L(H)$ is of trace class, if the Fourier transform $\hat{\mu}$ of μ is given by $\hat{\mu}(h) = \mathrm{e}^{i\langle x,h\rangle - \frac{1}{2}\langle Qh,h\rangle}$ for all $h \in H$. In this case we set $\mu = N_{x,Q}$ and $\mu = N_Q$ if $x = 0$.)

Remark 8 The limit case when $B = I$ (*white noise*) [80], is important in some application in physics, because it means that the noise acts uniformly in all directions. In this case the assumption $\mathrm{Tr}\,Q_t < +\infty$ is equivalent to

$$\int_0^t \mathrm{Tr}\left[\mathrm{e}^{sA}\mathrm{e}^{sA^*}\right]\mathrm{d}s < +\infty\,. \tag{15}$$

Assume, for instance, that A is the Laplacian in an open bounded set $\mathcal{O} \subset \mathbb{R}^d$ with Dirichlet boundary conditions. Let $Ae_k = \alpha_k e_k$, where $(e_k)_{k\in\mathbb{N}^d}$ is the complete orthonormal system of eigenfunctions of A and $(\alpha_k)_{k\in\mathbb{N}^d}$ are the corresponding eigenvalues. Then we have

$$\mathrm{Tr}\,Q_t = \sum_{k\in\mathbb{N}^d}\frac{1}{\alpha_k}\left(1 - \mathrm{e}^{-2\alpha_k t}\right)\,, \quad t > 0\,.$$

It is well known that $\alpha_k \sim |k|^2$ and so (15) is fulfilled only if $d = 1$. To deal with the case $d > 1$ one introduces the so-called *renormalization* [68,76].

Now by (13) it follows that the law of $X(t,x)$ is given by $N_{\mathrm{e}^{tA}x,Q_t}$, so the corresponding transition semigroup looks like

$$P_t\varphi(x) = \int_H \varphi(y)N_{\mathrm{e}^{tA}x,Q_t}(\mathrm{d}y)\,, \quad \varphi \in B_b(H)\,.$$

If e^{tA} is stable,

$$\|\mathrm{e}^{tA}\| \leq M\mathrm{e}^{-\omega t}\,, \quad t \geq 0\,,$$

and for some $M, \omega > 0$ it follows that

$$\lim_{t\to+\infty} P_t\varphi(x) = \int_H \varphi(y)N_{Q_\infty}(\mathrm{d}y)\,,$$

and so $\mu := N_{Q_\infty}$ is the unique invariant measure for P_t which is in addition ergodic and strongly mixing (for a necessary and sufficient condition for existence and uniqueness of an invariant measure for P_t see [26]).

Smooth Perturbations of System (12)

Let us consider the following stochastic differential equation

$$\begin{cases} \mathrm{d}X(t) = (AX(t) + b(X(t))\mathrm{d}t + B\mathrm{d}W(t)\,, & t \geq 0\,, \\ X(0) = x \in H\,, \end{cases} \tag{16}$$

where A and B are as before and $b\colon H \to H$ is of class C^1 and Lipschitz continuous. The condition on b is very strong; however, this case is important because it can be used for approximating equations with irregular coefficients.

By a *mild* solution of problem (16) on $[0, T]$ we mean a stochastic process $X \in C_W([0, T]; H)$ such that

$$X(t) = \mathrm{e}^{tA}x + \int_0^t \mathrm{e}^{(t-s)A} b(X(s))\mathrm{d}s + W_A(t)\,, \quad \mathbb{P}\text{-a.s.} \tag{17}$$

where W_A is the stochastic convolution defined by (14). By $C_W([0, T]; H)$ we mean the Banach space consisting of all continuous mappings $Y\colon [0, T] \to L^2(\Omega, \mathcal{F}, \mathbb{P}; H)$ which are adapted to W (that is, such that for all $t \in [0, T]$, $Y(t)$ is $\mathcal{F}_t$-measurable, where $\mathcal{F}_t$ is the σ-algebra generated by $\{W(s),\ s \in [0, t]\}$), endowed with the norm,

$$\|Y\|_{C_W([0,T];H))} = \left(\sup_{t\in[0,T]} \mathbb{E}\left(|Y(t)|^2\right)\right)^{1/2}.$$

It is called the space of all *mean square continuous adapted processes* on $[0, T]$ taking values on H.

Equation (17) can be solved easily by a standard fixed-point argument in the Banach space $C_W([0, T]; H)$. Thus we have the result

Proposition 9 *Equation (17) has a unique solution* $X(\cdot, x)$.

One checks easily that the corresponding transition semigroup P_t defined by (7) is Feller. When $B = I$ one can show that P_t is strongly Feller by using the *Bismut–Elworthy* formula [10,40] for the derivative of $D_x P_t\varphi(x)$ which reads as follows:

$$\langle DP_t\varphi(x), h\rangle = \frac{1}{t}\mathbb{E}\left[\varphi(X(t, x)) \int_0^t \langle D_x X(s, x)\cdot h, \mathrm{d}W(s)\rangle\right].$$

In this case one can also show that P_t is irreducible [20]. So, in view of Doob's theorem there is at most one invariant measure.

Existence of an invariant measure can be proved under the assumption

$$\langle Ax, x\rangle + \langle b(x), x\rangle \le a - b|x|^2\,, \quad \forall\, x \in D(A)\,,$$

where $a, b > 0$. In fact in this case one finds easily by Itô's formula an estimate of the second moment of $X(t, x)$ independent of t of the form

$$\mathbb{E}|X(t, x)|^2 \le c(1 + |x|^2)\,. \tag{18}$$

By (18) the tightness of $(\pi_t(x, \cdot))_{t\ge 0}$ follows for any fixed $x \in H$. We have in fact for any $R > 0$

$$\pi_t(x, B_R^c) \le \frac{1}{R^2}\int_H |y|^2 \pi_t(x, \mathrm{d}y) = \frac{1}{R^2}\,\mathbb{E}|X(t, x)|^2 \le \frac{c}{R^2}\,(1 + |x|^2)\,,$$

where B_R^c is the complement of the ball B_R of center 0 and radius R. So, the existence of an invariant measure follows from the Krylov–Bogoliubov theorem.

Reaction–Diffusion Equations Perturbed by Noise

We consider here a stochastic perturbation of problem (4) from Example 2, but for the sake of simplicity we take $\mathcal{O} = [0, 1]$ and $B = I$. So, we have $H = L^2(0, 1)$ and (4) becomes

$$\begin{cases} \mathrm{d}X(t, \xi) = \left[D_\xi^2 X(t, \xi) + p(X(t, \xi))\right]\mathrm{d}t + \mathrm{d}W(t, \xi)\,, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \xi \in [0, 1]\,, \\ X(t, 0) = X(t, 1) = 0\,, \quad t \ge 0\,, \\ X(0, \xi) = x(\xi)\,, \quad \xi \in [0, 1], \ x \in H\,, \end{cases} \tag{19}$$

where p is a polynomial of degree $d > 1$ with negative leading coefficient.

For generalizations to systems of reaction–diffusion equations in bounded subsets of $\mathbb{R}^n$ and for equations with multiplicative noise see [15,16] and references therein. For problems with reflexion see [65,82].

Now we write (19) in the mild form

$$X(t) = \mathrm{e}^{tA}x + \int_0^t \mathrm{e}^{(t-s)A} p(X(s))\mathrm{d}s + W_A(t)\,, \quad t \ge 0\,, \tag{20}$$

where $W_A(t)$ is the stochastic convolution defined by (14). Moreover,

$$Ax = D_\xi^2 x\,, \quad x \in D(A) = H^2(0, 1) \cap H_0^1(0, 1)$$

and

$$b(x)(\xi) = p(x(\xi))\,, \quad x \in D(b) = L^{2d}(0, 1)\,.$$

We notice that, since the leading coefficient of p is negative, b enjoys the basic dissipativity property

$$\langle b(x) - b(y), x - y\rangle \le c_1|x - y|^2\,, \quad \forall\, x, y \in L^{2d}(0, 1)\,,$$

where

$$c_1 = \sup_{\xi\in\mathbb{R}} p'(\xi)\,.$$

The operator A is self-adjoint and possesses a complete orthonormal system (e_k) of eigenfunctions, namely,

$$e_k(\xi) = (2/\pi)^{1/2}\ \sin(\pi\xi)\,, \quad \xi \in [0,1],\ k \in \mathbb{N}\,.$$

Moreover

$$Ae_k = -\pi^2 k^2 e_k\,, \quad k \in \mathbb{N}\,.$$

Therefore, we have $\langle Ax, x\rangle \le -\pi^2|x|^2$, for all $x \in H$.

Let us give the definition of the mild solution of (19). We say that $X \in C_W([0,T];H)$ ($C_W([0,T];H)$ was defined in Sect. "Smooth Perturbations of System (12)" is a *mild* solution of problem (19) if $X(t) \in L^{2d}(0,1)$ for all $t \ge 0$ and fulfills (20).

We notice that the condition $X(t) \in L^{2d}(0,1)$ is necessary in order for the integrand in (20) be meaningful.

The basic existence and uniqueness result is the following [3,26,41].

Theorem 10 *For any $x \in L^{2d}(\mathcal{O})$, there exists a unique mild solution $X(\cdot,x)$ of problem (19).*

Proof First we consider a regular approximating problem,

$$\begin{cases} \mathrm{d}X_\alpha(t) = (AX_\alpha(t) + b_\alpha(X_\alpha(t))\mathrm{d}t + \mathrm{d}W(t)\,, \\ X_\alpha(0) = x \in H\,, \end{cases} \tag{21}$$

where for any $\alpha > 0$, b_α is of class C^1, $b_\alpha(x) \to p(x)$ for all $x \in L^{2d}(0,1)$ as $\alpha \to 0$ and the basic dissipativity property,

$$\langle b_\alpha(x) - b_\alpha(y), x-y\rangle \le c_2|x-y|^2, \quad \forall\ x,y \in H,\ \alpha > 0,$$

is fulfilled, where c_2 is a constant independent of α. By Proposition 9 we know that problem (21) has a unique mild solution X_α, that is,

$$X_\alpha(t) = \mathrm{e}^{tA}x + \int_0^t \mathrm{e}^{(t-s)A} b_\alpha(X(s))\mathrm{d}s + W_A(t), \quad \mathbb{P}\text{-a.s.}$$

Taking advantage of the dissipativity of b_α it is possible to find suitable estimates for X_α and to show that X_α converge, as $\alpha \to 0$, to the unique solution X of (19). □

By Theorem 10 we can construct a transition semigroup P_t on $B_b(L^{2d}(0,1))$. However, it is useful to extend this semigroup to $B_b(L^2(0,1))$. For this we need a weaker notion of solution of (19) when $x \in H = L^2(0,1)$.

Let $x \in H$. We say that $X \in C_W([0,T];H)$ is a *generalized* solution of problem (19) if there exists a sequence $(x_n) \subset L^{2d}(\mathcal{O})$, such that

$$\lim_{n\to\infty} x_n = x \quad \text{in } H\,,$$

and

$$\lim_{n\to\infty} X(\cdot,x_n) = X(\cdot,x) \quad \text{in } C_W([0,T];H)\,.$$

It is easy to see, using again dissipativity of p, that this definition does not depend on the choice of the sequence (x_n) and to prove the following result.

Corollary 11 *For any $x \in H$, there exists a unique generalized solution $X(\cdot,x)$ of problem (19)*

Now we can define the transition semigroup

$$P_t\varphi(x) = \mathbb{E}[\varphi(X(t,x))]\,,$$

for all $\varphi \in B_b(H)$.

We finally discuss invariant measures of P_t assuming for simplicity that $p'(\xi) \le 0$ so that p is dissipative. In this case we can show the following result [26].

Theorem 12 *The semigroup P_t has a unique invariant measure ν, which is ergodic and strongly mixing and,*

$$\lim_{t\to+\infty} P_t\varphi(x) = \int_H \varphi(y)\nu(dy)\,, \quad x \in H\,. \tag{22}$$

Proof Let us consider the reaction–diffusion equation starting from $-s$ where $s > 0$,

$$\begin{cases} \mathrm{d}X(t,\xi) = \Big[D_\xi^2 X(t,\xi) + p(X(t,\xi))\Big]\mathrm{d}t + \mathrm{d}W(t,\xi)\,, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \xi \in [0,1]\,, \\ X(t,0) = X(t,1) = 0\,, \quad t \ge 0\,, \\ X(-s,\xi) = x(\xi)\,, \quad \xi \in [0,1]\,,\ x \in H\,. \end{cases} \tag{23}$$

Obviously we shall extend $W(t)$ to negative times setting

$$W(t) = W(-t) \quad \text{if } t \le 0\,.$$

Let us denote by $X(t,-s,x)$ the solution of (23). Then, using the dissipativity of p, one can show that there exists $\zeta \in L^2(\Omega,\mathcal{F},\mathbb{P};H)$ such that

$$\lim_{s\to+\infty} X(t,-s,x) = \zeta \quad \text{in } L^2(\Omega,\mathcal{F},\mathbb{P};H)\,, \quad x \in H\,,$$

where ζ is independent of t. Now, using the fact that the law of $X(t,-s,x)$ coincides with that of $X(t+s,x)$, it follows that the law ν of ζ is an invariant measure for P_t.

Moreover, one can show that P_t is irreducible and strongly Feller, so the uniqueness of ν as well as (22) follow from Doob's theorem. □

Burgers Equations Perturbed by Noise

We are concerned with the Burgers equation in $[0, 2\pi]$ perturbed by white noise

$$\begin{cases} \mathrm{d}X(t,\xi) = \left(\partial_\xi^2 X(t,\xi) + \frac{1}{2}\,\partial_\xi(X^2(t,\xi))\right)\mathrm{d}t \\ \qquad\qquad\qquad\qquad + \mathrm{d}W(t,\xi)\,, \xi \in [0,2\pi],\ t > 0\,, \\ X(t,0) = X(t,2\pi)\,, \quad t > 0 \\ X(0,\xi) = x(\xi)\,, \quad \xi \in [0,2\pi]\,. \end{cases} \tag{24}$$

We set $H = L^2(0, 2\pi)$ and give the following definition. A *mild solution* in $[0, T]$ of (24) is a process $X \in C_W([0, T]; H)$ such that

$$X(t) = \mathrm{e}^{tA}x + \frac{1}{2}\int_0^t \partial_\xi \mathrm{e}^{(t-s)A}(X^2(s))\mathrm{d}s + W_A(t)\,, \qquad t \geq 0\,, \tag{25}$$

where the stochastic convolution $W_A(t)$ is defined by (14).

We note that (25) is meaningful thanks to the estimate

$$\left|\partial_\xi e^{(t-s)A} y\right| \leq \kappa t^{-\frac{3}{4}}|x|_1\,, \quad x \in L^1(0, 2\pi)$$

(which can be proved by an elementary argument of harmonic analysis).

Now we state, following [29], an existence and uniqueness result.

Theorem 13 *For any $x \in H$ and $T > 0$ there exists a unique mild solution $X \in C_W([0, T]; H)$ of (24). Moreover, there exists a unique invariant measure which is ergodic and strongly mixing.*

Proof By the contraction principle it is easy to show existence and uniqueness of a local solution of (25) in a stochastic interval $[0, \tau(\omega))$ (since (25) is a semilinear equation with a locally Lipschitz nonlinearity). In order to show global existence one needs to find an a priori estimate for $\mathbb{P}$-almost all $\omega \in \Omega$. For this we first reduce (25) to a deterministic equation (more precisely to a family of deterministic equations indexed by $\omega \in \Omega$), setting

$$Y(t) = X(t) - W_A(t)\,.$$

$Y(t)$ fulfills in fact the equation

$$\begin{cases} \dfrac{\mathrm{d}Y(t)}{\mathrm{d}t} = AY(t) + b(Y(t) + W_A(t))\,, \quad t \in [0, T]\,, \\ Y(0) = x\,. \end{cases}$$

An a priori estimate for $Y(t)$ can be found by some manipulations using the basic property of b,

$$\langle b(x), x\rangle = 0\,, \quad \forall\, x \in D(b)\,.$$

For details and for the existence of an invariant measure see [29], for the uniqueness see [27]. □

Another approach for studying the Burgers equation can be found in [9].

2D Navier–Stokes Equation Perturbed by Noise

We are here concerned with the equation

$$\begin{cases} \mathrm{d}Z = (\Delta_\xi Z - Z + D_\xi Z \cdot Z)\mathrm{d}t + \nabla p\mathrm{d}t + B_1\mathrm{d}W(t) \\ \qquad\qquad\qquad\qquad \text{in } [0, +\infty) \times \mathcal{O}\,, \\ \operatorname{div} Z = 0 \quad \text{in } [0, +\infty) \times \mathcal{O}\,, \\ Z(t, \cdot) \text{ is periodic with period } 2\pi\,, \\ Z(0, \cdot) = z \quad \text{in } \mathcal{O}\,, \end{cases} \tag{26}$$

where $Z = (Z_1, Z_2)$ belongs to the space H of all square integrable divergence free vectors and $B_1 \in L(H)$.

Several papers have been devoted to stochastic 2D Navier–Stokes equations. We mention, besides the pioneering work [1,7,13,22]. Concerning 3D Navier–Stokes equations (which we do not consider here) one can look at [13,14,23,42,43,44,46,47].

We write (26) in the mild form

$$X(t) = \mathrm{e}^{tA}x + \int_0^t \mathrm{e}^{(t-s)A}b(X(s))\mathrm{d}s + W_A(t)\,, \tag{27}$$

where the operators A and b were defined in Example 4 and the stochastic convolution $W_A(t)$ is given by (14) with $B = \mathcal{P}B_1$.

In order to study (27) it is convenient to introduce the mapping

$$\Gamma(f) = \int_0^t \mathrm{e}^{(t-s)A}b(f(s))\mathrm{d}s, \quad f \in L^4((0, T)\times\mathcal{O}), \quad t \geq 0,$$

and then to write (27) as

$$X(t) = \mathrm{e}^{tA}x + \Gamma(X)(t) + W_A(t)\,.$$

In fact one can show [27] that Γ maps $L^4((0, T) \times \mathcal{O})$ into itself and for any $f, g \in L^4((0, T) \times \mathcal{O}) := L^4$ we have

$$|\Gamma(f) - \Gamma(g)|_{L^4} \leq 4(|f|_{L^4} + |g|_{L^4})\,|f - g|_{L^4}\,.$$

Theorem 14 *For any $x \in H$ there exists a unique mild solution $X(\cdot, x)$ of (26).*

Proof One reduces (27) to a family of deterministic equations as in the case of the Burgers equation. Then one uses a fixed-point argument in the space $L^4(0, T; L^4)$ [27]. □

Existence of invariant measures is not difficult. It can be obtained, as in the previous examples, by the Krylov–Bogoliubov theorem. Uniqueness is more delicate. When the noise $B\mathrm{d}W(t)$ is not too degenerate it was proved in [45]. Otherwise one has to use the *coupling* argument [54,58,60,81].

Future Directions

The number of papers devoted to nonlinear SPDEs is increasing. In fact, several models arising in the application are more realistic when a random perturbation is taken into account. Among new directions of research we mention the following.

- The direct study of the Kolmogorov equation (8) (considered as a parabolic equation with infinitely many variables). This can give important information about the properties of system (2) even in the case when the problem is not well posed. A result in this direction can be found in [28]. For a solution of (8) in the case of the 3D Navier–Stokes equation (with a failed attempt to prove uniqueness in law) see [24].
- Hamilton–Jacobi equations such as

$$\begin{cases} \dfrac{\mathrm{d}u(t,x)}{\mathrm{d}t} = K_0 u(t,x) + \mathcal{H}(u_x(t,x)) \,, \\ \qquad\qquad\qquad\qquad\qquad\qquad x \in H,\ t \geq 0 \,, \\ u(0,x) = \varphi(x) \,, \end{cases}$$

  where $\mathcal{H}$ is a Hamiltonian. This is an important subject related to stochastic optimal control problems for SPDEs. For early results see Chap. 12 in [28]. Recent results are also connected with backward equations. Here the pioneering finite-dimensional result in [67] has been extended in infinite dimensions [50].

We believe also that an extension of the following topics to infinite dimensions would be interesting:

- Homogenization and averaging. In finite dimension see, e. g., [8,49]
- Measures solutions of Kolmogorov equations. In finite dimension this is a well-known subject, see, e. g., [11,12].

Bibliography

1. Albeverio S, Cruizeiro AB (1990) Global flows with invariant (Gibbs) measures for Euler and Navier–Stokes two dimensional fluids. Commun Math Phys 129:431–444
2. Albeverio S, Röckner M (1991) Stochastic differential equations in infinite dimensions: solutions via Dirichlet forms. Probab Th Rel Fields 89:347–386
3. Bally V, Gyongy I, Pardoux E (1994) White noise driven parabolic SPDEs with measurable drift. J Funct Anal 120(2):484–510
4. Barbu V, Bogachev VI, Da Prato G, Röckner M (2006) Weak solution to the stochastic porous medium equations: the degenerate case. J Funct Anal 235(2):430–448
5. Barbu V, Da Prato G (2002) The stochastic nonlinear damped wave equation. Appl Math Optimiz 46:125–141
6. Barbu V, Da Prato G, Röckner M (2008) Existence and uniqueness of nonnegative solutions to the stochastic porous media equation. preprint S.N.S. Indiana University Math J 57(1):187–212
7. Bensoussan A, Temam R (1973) Équations stochastiques du type Navier–Stokes. J Funct Anal 13:195–222
8. Bensoussan A, Lions JL, Papanicolaou G (1978) Asymptotic analysis for periodic structures. In: Studies in mathematics and its applications, vol 5. North-Holland, Amsterdam
9. Bertini L, Cancrini N, Jona-Lasinio G (1994) The stochastic Burgers equation. Comm Math Phys 165(2):211–232
10. Bismut JM (1984) Large deviations and the Malliavin calculus. Progress in Mathematics 45. Birkhäuser, Boston
11. Bogachev VI, Da Prato G, Röckner M (2004) Existence of solutions to weak parabolic equations for measures. Proc London Math Soc 3(88):753–774
12. Bogachev VI, Krylov NV, Röckner M (2006) Elliptic equations for measures: regularity and global bounds of densities. J Math Pures Appl 9(6):743–757
13. Brzezniak Z, Capinski M, Flandoli F (1991) Stochastic partial differential equations and turbulence. Math Models Methods Appl Sci 1(1):41–59
14. Capinski M, Cutland N (1994) Statistical solutions of stochastic Navier–Stokes equations. Indiana Univ Math J 43(3):927–940
15. Cerrai S (2001) Second order PDE's in finite and infinite dimensions: A probabilistic approach. In: Lecture Notes in Mathematics, vol 1762. Springer, Berlin
16. Cerrai S (2003) Stochastic reaction-diffusion systems with multiplicative noise and non-Lipschitz reaction term. Probab Th Rel Fields 125:271–304
17. Cerrai S, Freidlin M (2006) On the Smoluchowski–Kramers approximation for a system with an infinite number of degrees of freedom. Probab Th Rel Fields 135(3):363–394
18. Crauel H, Debussche A, Flandoli F (1997) Random attractors. J Dynam Differen Equ 9:307–341
19. Dalang R, Frangos N (1998) The stochastic wave equation in two spatial dimensions. Ann Probab 26(1)187–212
20. Da Prato G (2004) Kolmogorov equations for stochastic PDEs. Birkhäuser, Basel
21. Da Prato G, Debussche A (1996) Stochastic Cahn–Hilliard equation. Nonlinear Anal 26(2):241–263
22. Da Prato G, Debussche A (2002) 2*D* Navier–Stokes equations driven by a space-time white noise. J Funct Anal 196(1):180–210
23. Da Prato G, Debussche A (2003) Strong solutions to the stochastic quantization equations. Ann Probab 31(4):1900–1916
24. Da Prato G, Debussche A (2003) Ergodicity for the 3*D* stochastic Navier–Stokes equations. J Math Pures Appl 82:877–947
25. Da Prato G, Tubaro L (2000) A new method to prove self-adjointness of some infinite dimensional Dirichlet operator. Probab Th Rel Fields 118(1):131–145

26. Da Prato G, Zabczyk J (1992) Stochastic equations in infinite dimensions. Cambridge University Press, Cambridge
27. Da Prato G, Zabczyk J (1996) Ergodicity for infinite dimensional systems. In: London Mathematical Society Lecture Notes, vol 229. Cambridge University Press, Cambridge
28. Da Prato G, Zabczyk J (2002) Second order partial differential equations in Hilbert spaces. In: London Mathematical Society Lecture Notes, vol 293. Cambridge University Press
29. Da Prato G, Debussche A, Temam R (1994) Stochastic Burgers equation. Nonlinear Differ Equ Appl 4:389–402
30. Da Prato G, Röckner M, Rozovskii BL, Wang FY (2006) Strong solutions of stochastic generalized porous media equations: Existence, uniqueness and ergodicity. Comm Partial Differ Equ 31(1–3):277–291
31. Dawson DA (1975) Stochastic evolution equations and related measures processes. J Multivariate Anal 5:1–52
32. de Bouard A, Debussche A (1999) A stochastic nonlinear Schrödinger equation with multiplicative noise. Comm Math Phys 205(1):161–181
33. de Bouard A, Debussche A (2002) On the effect of a noise on the solutions of the focusing supercritical nonlinear Schrödinger equation. Probab Th Rel Fields 123(1):76–79
34. de Bouard A, Debussche A (2003) The stochastic nonlinear Schrödinger equation in H^1. Stoch Anal Appl 21(1):97–126
35. de Bouard A, Debussche A (2005) Blow-up for the stochastic nonlinear Schrödinger equation with multiplicative noise. Ann Probab 33(3):1078–1110
36. de Bouard A, Debussche A, Tsutsumi Y (2004–05) Periodic solutions of the Korteweg–de Vries equation driven by white noise (electronic). SIAM J Math Anal 36(3):815–855
37. Debussche A, Printems J (2001) Effect of a localized random forcing term on the Korteweg–de Vries equation. J Comput Anal Appl 3(3):183–206
38. Debussche A, Zambotti L (2007) Conservative stochastic Cahn–Hilliard equation with reflection. Ann Probab 35(5):1706–1739
39. Eckmann JP, Ruelle D (1985) Rev Mod Phys 53:643–653
40. Elworthy KD (1992) Stochastic flows on Riemannian manifolds. In: Pinsky MA, Wihstutz V (eds) Diffusion processes and related problems in analysis, vol II (Charlotte, NC, 1990), Progr Probab 27. Birkhäuser, Boston, pp 37–72
41. Faris WG, Jona-Lasinio G (1982) Large fluctuations for a nonlinear heat equation with noise. J Phys A.15:3025–3055
42. Flandoli F (1994) Dissipativity and invariant measures for stochastic Navier–Stokes equations. Nonlinear Differ Equ Appl 1:403–423
43. Flandoli F (1997) Irreducibilty of the 3D stochastic Navier–Stokes equation. J Funct Anal 149:160–177
44. Flandoli F, Gątarek D (1995) Martingale and stationary solutions for stochastic Navier–Stokes equations. Probab Th Rel Fields 102:367–391
45. Flandoli F, Maslowski B (1995) Ergodicity of the 2-D Navier–Stokes equation under random perturbations. Commun Math Phys 171:119–141
46. Flandoli F, Romito M (2002) Partial regularity for the stochastic Navier–Stokes equations. Trans Am Math Soc 354(6):2207–2241
47. Flandoli F, Romito M (2006) Markov selections and their regularity for the three-dimensional stochastic Navier–Stokes equations. C R Math Acad Sci Paris 343(1):47–50
48. Fleming WH (1975) A selection-migration model in population genetics. J Math Biol 2(3):219–233
49. Freidlin M (1996) Markov processes and differential equations: asymptotic problems. In: Lectures in Mathematics ETH Zürich. Birkhäuser, Basel
50. Fuhrman M, Tessitore G (2002) Nonlinear Kolmogorov equations in infinite dimensional spaces: the backward stochastic differential equations approach and applications to optimal control. Ann Probab 30(3):1397–1465
51. Funaki T (1991) The reversible measures of multi-dimensional Ginzburg–Landau type continuum model. Osaka J Math 28(3):463–494
52. Funaki T, Olla S (2001) Fluctuations for $\nabla\phi$ interface model on a wall. Stoch Process Appl 94(1):1–27
53. Funaki T, Spohn H (1997) Motion by mean curvature from the Ginzburg–Landau $\nabla\phi$ interface model. Comm Math Phys 185(1):1–36
54. Hairer M, Mattingly JC (2006) Ergodicity of the 2D Navier–Stokes equations with degenerate stochastic forcing. Ann Math 3:993–1032
55. Jona-Lasinio G, Mitter PK (1985) On the stochastic quantization of field theory. Commun Math Phys 101(3):409–436
56. Krylov NV (1999) An analytic approach to SPDEs. In: Stochastic partial differential equations: six perspectives, Math Surveys Monogr, vol 64. Amer Math Soc, Providence, pp 185–242
57. Krylov NV, Rozovskii BL (1981) Stochastic evolution equations, Translated from Itogi Naukii Tekhniki, Seriya Sovremennye Problemy Matematiki 14:71–146. J Soviet Math 14:1233–1277
58. Kuksin S, Shirikyan A (2001) A coupling approach to randomly forced randomly forced PDE's I. Commun Math Phys 221:351–366
59. Liskevich V, Röckner M (1998) Strong uniqueness for a class of infinite dimensional Dirichlet operators and application to stochastic quantization. Ann Scuola Norm Sup Pisa Cl Sci 4(XXVII):69–91
60. Mattingly J (2002) Exponential convergence for the stochastically forced Navier–Stokes equations and other partially dissipative dynamics. Commun Math Phys 230(3):421–462
61. Mikulevicius R, Rozovskii B (1998) Martingale problems for stochastic PDE's. In: Carmona RA, Rozoskii B (eds) Stochastic partial differential equations: Six perspectives. Mathematical Surveys and Monograph, vol 64. American Mathematical Society, pp 243–325
62. Millet A, Morien PL (2001) On a nonlinear stochastic wave equation in the plane: existence and uniqueness of the solution. Ann Appl Probab 11:922–951
63. Millet A, Sanz-Solé M (2000) Approximation and support theorem for a wave equation in two space dimensions. Bernoulli 6:887–915
64. Musiela M, Rutkowski M (1997) Martingale methods in financial modelling. In: Applications of Mathematics (New York), vol 36. Springer, Berlin
65. Nualart D, Pardoux E (1992) White noise driven quasilinear SPDEs with reflection. Prob Theory Rel Fields 93:77–89
66. Pardoux E (1975) Èquations aux derivées partielles stochastiques nonlinéaires monotones. Thèse, Université Paris XI
67. Pardoux E, Peng SG (1990) Adapted solution of a backward stochastic differential equation. Syst Control Lett 14(1):55–61
68. Parisi G, Wu YS (1981) Sci Sin 24:483–490
69. Peszat S, Zabczyk J (2000) Nonlinear stochastic wave and heat equations. Probab Th Rel Fields 116(3):421–443

70. Petersen K (1983) Ergodic Theory. Cambridge, London
71. Prevot C, Röckner M (2007) A concise course on stochastic partial differential equations, Monograph 2006. In: Lecture Notes in Mathematics. Springer, Berlin
72. Priola E (1999) On a class of Markov type semigroups in spaces of uniformly continuous and bounded functions. Studia Math 136:271–295
73. Ren J, Röckner M, Wang FY (2007) Stochastic generalized porous media and fast diffusions equations. J Diff Equ 238(1):118–152
74. Rozovskii BL (2003) Linear theory and applications to nonlinear filtering (mathematics and its applications). Kluwer, Dordrecht
75. Sanz-Solé M (2005) Malliavin calculus with applications to stochastic partial differential equations. In: Fundamental sciences. EPFL Press, Lausanne; distributed by CRC Press, Boca Raton
76. Simon B (1974) The $P(\phi)_2$ Euclidean (quantum) field theory. Princeton University Press, Princeton
77. Stannat W (2000) On the validity of the log-Sobolev inequality for symmetric Fleming–Viot operators. Ann Probab 28(2):667–684
78. Temam R (1988) Infinite-dimensional dynamical systems in mechanics and physics. Springer, New York
79. Viot M (1976) Solution faibles d'équations aux dérivées partielles non linéaires. Thèse, Université Pierre et Marie Curie, Paris
80. Walsh JB (1986) An introduction to stochastic partial differential equations. In: cole d't de probabilits de Saint-Flour, XIV—1984. Lecture Notes in Math, vol 1180. Springer, Berlin, pp 265–439
81. Weinan E, Mattingly JC, Sinai YG (2001) Gibbsian dynamics and ergodicity for the stochastically forced Navier–Stokes equation. Commun Math Phys 224:83–106
82. Zambotti L (2001) A reflected stochastic heat equation as a symmetric diynamic with respect to 3-d Bessel Bridge. J Funct Anal 180(1):195–209

Non-negative Matrices and Digraphs

Abraham Berman[1], Naomi Shaked-Monderer[2]
[1] Technion – Israel Institute of Technology, Haifa, Israel
[2] Emek Yezreel College, Emek Yezreel, Israel

Article Outline

Glossary
Definition of the Subject
Introduction
Matrices, Graphs and Digraphs
Irreducible Nonnegative Matrices and Their Digraphs
Special Irreducible Matrices
Reducible Nonnegative Matrices and Their Digraphs
Examples: Ranking Problems
The Laplacian Matrix of a Directed Graph
Research Directions
Bibliography

Glossary

Nonnegative matrix A *nonnegative matrix* is a matrix whose entries are real nonnegative numbers. The matrix is *positive* if all its elements are positive.

Digraph A *digraph* (*directed graph*) Γ consists of a finite set V and a set E of ordered pairs of elements of V. The elements of V are called *vertices* and those of E are called *arcs*. It is often represented graphically by points as the vertices of Γ, and arrows between these points as the arcs.

Irreducible matrix A *reducible* matrix is a matrix which is either the 1×1 zero matrix, or a square matrix that has a zero submatrix on complementary sets of rows and columns. A square matrix which is not reducible is *irreducible*.

Strongly connected digraph A *strongly connected digraph* is a digraph in which there is a walk from every vertex to every other vertex.

(Directed) walk A *directed walk* (*walk*) in a digraph is a sequence of arcs of the digraph such that each arc starts at the end vertex of its predecessor.

Definition of the Subject

The interplay between Matrix Theory and Graph Theory is fascinating and fruitful, e. g. [4,5,8,14,18,21,25,31,32]. Here we only deal with the interplay between nonnegative matrices and digraphs.

Digraphs are of great importance in Computer Science, Social Sciences and Natural Sciences, see [1,2,55]. Nonnegative matrices have applications to a variety of disciplines, such as numerical analysis, probability, game theory, economics, optimization, dynamical systems, and data mining, see [3,6,7,39]. These two important subjects, digraphs and nonnegative matrices, are also strongly connected. (Nonnegative) matrices are a natural way to describe digraphs and conversely, digraphs describe the zero-nonzero structure of a matrix. This survey concentrates on the meeting points of the two subjects.

The theory of nonnegative matrices is a hundred years old, and digraphs have accompanied it almost since its beginning. This continues to be true today, when modern applications stimulate the study of both subjects. One of the main sources for the renewed interest is probably the possibilities offered, and the problems generated, by the use of computers. We start our survey with an example of such application.

Introduction

Google PageRank is a famous application of the theory of nonnegative matrices. Developed by Sergey Brin and

Lawrence Page [11], it ranks web pages according to their importance, under the principle that a page is important if important pages point to it. This is the idea: Suppose there are N web pages on the Web. Many of them contain links to other web pages. This can be described by a digraph: each web page is represented by a vertex, and a link on page i to page j is represented by an arc from i to j. The digraph in Fig. 1 shows a web of $N = 6$ pages. (In the real Web, the number of pages is of course huge and constantly growing, with current estimates above $N = 10^{10}$ [26]).

Let p_{ij} denote the probability that a surfer visiting page i will move to page j. Suppose there are d_i links out of page i. In a simplified version of the Google model, in the case that $d_i > 0$, $p_{ij} = \alpha \cdot \frac{1}{d_i} + (1-\alpha) \cdot \frac{1}{N}$ if there is a link from page i to page j, and $p_{ij} = (1-\alpha) \cdot \frac{1}{N}$ if there is no such link, where $\alpha = 0.85$. For a page i with no outgoing links, that is, when $d_i = 0$, $p_{ij} = \frac{1}{N}$ for every j. This represents the idea that with probability α a surfer on a page which has at least one link, randomly chooses (with equal probabilities) to follow one of the links on the page, and with probability $1-\alpha$ the surfer ignores the links, and chooses to surf to a random page on the Web (again, with equal probabilities), while if there are no links on the page, the surfer just chooses a random Web page to surf to. Note that the matrix $P = [p_{ij}]$ has positive entries, and it is row stochastic (i. e., each of its rows sums up to 1). For the web described in Fig. 1,

$$P = \begin{bmatrix} 1/40 & 7/8 & 1/40 & 1/40 & 1/40 & 1/40 \\ 1/40 & 1/40 & 19/80 & 19/80 & 19/80 & 19/80 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/40 & 1/40 & 1/40 & 1/40 & 9/20 & 9/20 \\ 1/40 & 1/40 & 1/40 & 9/20 & 1/40 & 9/20 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{bmatrix}.$$

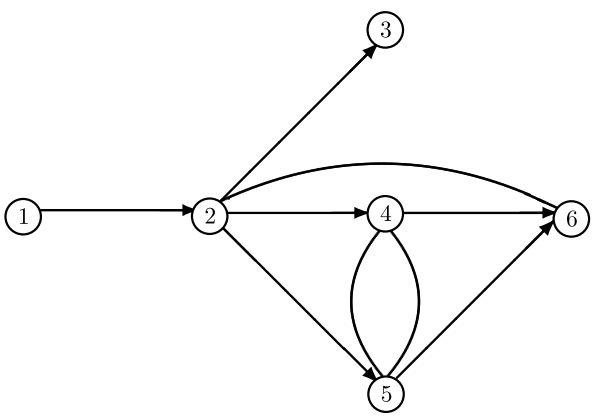

Non-negative Matrices and Digraphs, Figure 1
A 6-pages web

The PageRank of a page i, $PR(i)$, is defined as the probability that a random surfer will be in i. In this model, the larger $PR(i)$ is, the more important page i is. The relations between these PageRanks are:

$$\begin{cases} PR(1) = p_{11}PR(1) + \cdots + p_{N1}PR(N) \\ \quad\vdots \\ PR(N) = p_{1N}PR(1) + \cdots + p_{NN}PR(N) \end{cases}.$$

Let $\pi = \left[PR(1), PR(2), \ldots, PR(N)\right]^{\mathrm{T}}$ be the (column) vector of PageRanks. By the above equations, the vector π satisfies $\pi = P^{\mathrm{T}}\pi$. That is, the PageRank vector π is an eigenvector of P^{T} that corresponds to the eigenvalue 1, with nonnegative elements which sum up to 1. For the model to work we need to know that such an eigenvector exists, and that it is unique. This is guaranteed by the first major theorem on nonnegative matrices, which was proved by Perron in 1907 [43].

The Perron Theorem is a theorem on positive matrices. Its important extension to nonnegative matrices, proved by Frobenius in 1912 [23] and referred to as the Perron–Frobenius Theorem, requires the notion of irreducibility. This is the natural place to use graph theory but this was not done by Frobenius who did not care much for the, infant at his time, theory of graphs. In a very interesting article [48] Schneider describes the relations between irreducibility, the related concept of full indecomposability and the Frobenius–König Theorem as they appear in the papers of Frobenius, König and Markov (after which the highly important Markov Chains are named). He also touches the acrimonious controversy developed between Frobenius and König about their respective contributions to the topic. Without taking sides in this controversy, it is clear from historical perspective that Frobenius underestimated the relevance of digraphs to the theory of nonnegative matrices. This hopefully is demonstrated in this survey.

Matrices, Graphs and Digraphs

While we concentrate here on the interrelations between *nonnegative matrices* and *directed graphs*, we mention briefly the relations between *matrices* and *graphs*, and include the basic terminology of graphs and digraphs that will be used in this article.

A *(simple, undirected) graph* G is a pair $G = (V, E)$, where V is a finite set, and E is a set of unordered pairs of distinct elements in V. The elements of V are the *vertices* of G, and the elements of E are the *edges* of G. If $e = \{u, v\}$ is an edge of G, we say that u (or v) and e are *incident* with

each other, and that u and v are *adjacent* vertices in G. The *degree* of a vertex $v \in V$ is the number of vertices adjacent to v. The added adjective 'simple' means that in these graphs there are no loops (that is, no edge joins a vertex to itself) or multiple edges (that is, any two distinct vertices are joined by at most one edge).

Several types of graphs may be associated with matrices. Given an $m \times n$ matrix A, its *bipartite graph* $B(A) = (V, E)$ has $m + n$ vertices: $V = \{r_1, \ldots, r_m, c_1, \ldots, c_n\}$ (each vertex r_i corresponds to a row of A, and each vertex c_j corresponds to a column of A). $\{u, v\}$ is an edge of $B(A)$ if and only if $u = r_i$ and $v = c_j$ for some i and j and $a_{ij} \neq 0$.

With a *symmetric* $n \times n$ matrix A we may associate its *graph* $G(A) = (V, E)$, with vertices $V = \{1, \ldots, n\}$ and $\{i, j\} \in E$ if and only if $i \neq j$ and $a_{ij} (= a_{ji}) \neq 0$. $G(A)$ is a simple graph, and it conveys the zero-nonzero pattern of A, except for the diagonal entries of A.

We may also associate matrices with given graphs. Given a graph $G = (V, E)$ with vertices $V = \{v_1, \ldots, v_n\}$ and edges $E = \{e_1, \ldots, e_m\}$, its *incidence matrix* $C = C(G)$ is the $n \times m$ matrix with $c_{ij} = 1$ if v_i is incident with e_j in G, and $c_{ij} = 0$ otherwise. The *adjacency matrix* of G, $A = A(G)$ is the $n \times n$ $(0, 1)$-matrix with $a_{ij} = 1$ if and only if $\{v_i, v_j\} \in E$. (The diagonal of $A(G)$ is zero). Another matrix associated with G is its *(combinatorial) Laplacian*, $L = L(G) = D(G) - A(G)$, where $D = D(G)$ is the diagonal matrix of vertex degrees, i. e., d_{ii} is equal to the degree of the vertex v_i. By scaling L one obtains the *(normalized) Laplacian* $\mathcal{L} = \mathcal{L}(G)$, defined as $\mathcal{L} = D^{-1/2} L D^{-1/2}$ (for this purpose D^{-1} is actually $D^{\dagger}$, the Moore–Penrose inverse of D: it is the diagonal matrix with $\left(D^{-1}\right)_{ii} = 0$ if $d_{ii} = 0$, and $\left(D^{-1}\right)_{ii} = 1/d_{ii}$ if $d_{ii} > 0$). These matrices associated with the graph G are related: $A = CC^{\mathrm{T}} - D$, and hence $L = D - A = 2D - CC^{\mathrm{T}}$. The adjacency matrix A and the Laplacians L and $\mathcal{L}$ are symmetric matrices, and thus have real eigenvalues. The relations between these eigenvalues and the structure of the graph G were extensively studied. In particular, L (and thus also $\mathcal{L}$) is a singular positive semidefinite matrix, its smallest eigenvalue is zero ($Le = 0$, where e is the n-vector with all entries equal to 1). The second smallest eigenvalue of L was called by Fiedler the *algebraic connectivity* of G, motivated by the fact that the algebraic connectivity is nonzero if and only if the graph G is connected.

While there is a natural relation between symmetric matrices and graphs, general square matrices may be associated with digraphs. A *digraph* (*directed graph*) Γ is a pair $\Gamma = (V, E)$, where V is a finite set (we usually assume $V = \{1, \ldots, n\}$) and E is a set of ordered pairs of (not necessarily distinct) elements of V. The elements of V are called the *vertices* of Γ; The elements of E are the *arcs* of Γ. We say that $(i, j) \in E$ is an arc *from i to j*; i is its *initial vertex* or *tail*, and j is its *terminal vertex* or *head*. An arc (i, i) is called a *loop* (i is both its initial vertex and its terminal vertex). The number of arcs with i as an initial vertex is the *outdegree* of i. The number of arcs with i as terminal vertex is the *indegree* of i. We will consider only digraphs with no multiple arcs. That is, for every two distinct vertices i and j there is at most one arc from i to j. A digraph which has no loops and no multiple arcs is called *simple*. So in this article, a simple digraph is simply a digraph with no loops.

A *weighted digraph* is a digraph in which each arc (i, j) is assigned a weight ω_{ij} (usually a real or complex number). A *(directed) walk* of length m from vertex i to vertex j in a digraph Γ is a sequence of m arcs $(i_0, i_1), (i_1, i_2), \ldots, (i_{m-1}, i_m)$, with the initial vertex of each arc equal to the terminal vertex of its predecessor, and $i_0 = i$, $i_m = j$. If $i_0 = i_m$, it is a *closed walk*. If all the vertices in the walk are distinct, except, possibly, the first and last, the walk is called a *(directed) path*. When $i_m = i_0$ the path is called a *(directed) cycle*. The *distance* from vertex i to vertex j is the length of the shortest path from i to j, and the *diameter* of a digraph is largest distance between two vertices in the digraph. The *girth* of a digraph is the length of its shortest cycle.

A digraph is *strongly connected* if for every two vertices $i \neq j$ there are walks from i to j and from j to i. Every digraph on one vertex is strongly connected.

With each simple digraph Γ there is an *underlying graph*. The underlying graph G has the same vertices as Γ, and each arc (i, j) of Γ is replaced by the corresponding edge $\{i, j\}$ in G. (No multiple edges, though, so if (i, j) and (j, i) are both arcs of Γ, there is only one edge $\{i, j\}$ in G).

Let Γ be a digraph with vertices $V = \{1, \ldots, n\}$ and arcs $E = \{e_1, \ldots, e_m\}$. Its *incidence matrix* $Q = Q(\Gamma)$ is the $n \times m$ matrix with $q_{ij} = -1$ if i the initial vertex of e_j, $q_{ij} = 1$ if i the terminal vertex of e_j and $q_{ij} = 0$ otherwise. (If Γ is a simple digraph with incidence matrix Q, and G is its underlying digraph with combinatorial Laplacian L, then $QQ^{\mathrm{T}} = L$).

The *adjacency matrix* $A = A(\Gamma)$ of a digraph Γ with vertices $\{1, \ldots, n\}$ is an $n \times n$ matrix with $a_{ij} = 1$ if and only if $(i, j) \in E$, and $a_{ij} = 0$ otherwise. The adjacency matrix is a square $(0, 1)$-matrix. On the other hand, given an $n \times n$ matrix A, the *digraph of* A, $\Gamma(A)$, has vertices $\{1, \ldots, n\}$, and (i, j) is an arc if and only if $a_{ij} \neq 0$. We may also consider a *weighted digraph* associated with A by assigning the weight a_{ij} to the arc (i, j) of $\Gamma(A)$. Obviously, $\Gamma\left(A(\Gamma)\right) = \Gamma$.

Re-labeling the vertices of a digraph is equivalent to simultaneous permutation of the rows and the columns of an associated matrix. Such simultaneous permutation is achieved by pre-multiplying A by a permutation matrix P, and post-multiplying it by P^{T}. The digraph $\Gamma(PAP^{\mathrm{T}})$ is obtained from the digraph $\Gamma(A)$ by relabeling its vertices according to the permutation, and vice versa: if Γ' is a digraph obtained from $\Gamma(A)$ by re-labeling the vertices, then there exists a permutation matrix P such that $\Gamma(PAP^{\mathrm{T}}) = \Gamma'$.

A *subdigraph* of Γ is a digraph $\Gamma' = (V', E')$, where $V' \subseteq V$ and $E' \subseteq E$. Such Γ' is an *induced subdigraph* of Γ if $E' = E \cap (V' \times V')$, i. e., E' consists of all the arcs in E whose initial and terminal vertices are in V'. The adjacency matrix of the subdigraph of Γ induced by V' is a principal submatrix of the adjacency matrix $A = A(\Gamma)$ (that is, a submatrix of A lying on rows and columns with the same indices). We denote this submatrix by $A[V']$. The converse also holds. If $\Gamma = \Gamma(A)$ for some matrix A, then the digraph of a principal submatrix of A is an induced subdigraph of Γ.

In what follows, all matrices under discussion will be square, unless specifically stated otherwise.

As mentioned before, we deal here only with nonnegative matrices and digraphs. For more on the connections between matrices and graphs, see e. g., Part II of [31], and the reference there. For the many connections between the spectrum of the Laplacian of a graph to graph structure parameters see [18].

Irreducible Nonnegative Matrices and Their Digraphs

A matrix A is *nonnegative* (denoted $A \geq 0$) if all its entries are nonnegative, and *positive* (denoted $A > 0$) if all its entries are positive. If A and B are matrices of the same order, we write $A \geq B$ if $A - B \geq 0$ and $A > B$ if $A - B > 0$.

If A is an $n \times n$ matrix, then its spectrum consists of n (complex, not necessarily distinct) eigenvalues, $\lambda_1, \ldots, \lambda_n$. The *spectral radius* of A is $\rho(A) = \max\{|\lambda_i| : 1 \leq i \leq n\}$. Let $\mu(A) = \max\{|\lambda_i| : 1 \leq i \leq n, \lambda_i \neq \rho(A)\}$. With these notations, we can state Perron's Theorem [43]:

Theorem 1 (Perron Theorem) *If A is a positive square matrix, then*

a. $\rho(A) > 0$.
b. *$\rho(A)$ is a simple eigenvalue of A (called the Perron root of A).*
c. *To $\rho(A)$ there corresponds a positive eigenvector. (Such an eigenvector, normalized so that the sum of its entries is 1, is called the Perron vector of A).*
d. *$\mu(A) < \rho(A)$, that is, $\rho(A)$ is the only eigenvalue of A of maximum modulus.*
e. $\lim_{m\to\infty} \left(\frac{A}{\rho(A)}\right)^m = L$, $L = xy^{\mathrm{T}}$, *where* $Ax = \rho(A)x$, $x > 0$, $A^{\mathrm{T}}y = \rho(A)y$, $y > 0$, *and* $x^{\mathrm{T}}y = 1$.
f. *For every $\mu(A)/\rho(A) < r < 1$ there exists a constant $C = C(r, A)$ such that for every m,* $\left\|\left(\frac{A}{\rho(A)}\right)^m - L\right\|_\infty \leq Cr^m$.

The Perron Theorem may be extended to a wider class of matrices. A matrix A is *primitive* if for some natural number m, A^m is positive.

Theorem 2 *In Perron's Theorem (Theorem 1) "positive" may be replaced by "primitive".*

The theorem can be generalized to irreducible matrices. An $n \times n$ matrix A is *reducible* if $n > 1$ and there exists a permutation matrix P such that

$$PAP^{\mathrm{T}} = \begin{bmatrix} B & 0 \\ C & D \end{bmatrix},$$

with square diagonal blocks B and D, or $n = 1$ and $A = 0$. A matrix is *irreducible* if it is not reducible. Irreducibility may be described in terms of the digraph of A: A matrix A of order greater than 1 is irreducible if and only if $\Gamma(A)$ is strongly connected. Moreover,

Theorem 3 *Let A be an $n \times n$ nonnegative matrix, $n > 1$. Then the following are equivalent:*

a. *A is irreducible.*
b. *$\Gamma(A)$ is strongly connected.*
c. *There is a positive integer m such that $(I + A)^m > 0$.*
d. $(I + A)^{n-1} > 0$.

In Perron's Theorem (Theorem 1) parts *a–c*, 'positive' may be replaced by 'nonnegative and irreducible' – see parts *a–c* in Theorem 4 below. The new parts *d–f* of Theorem 4 also hold for such matrices (in particular for positive matrices). Theorem 4, the Perron–Frobenius Theorem, was first proved in [23].

Theorem 4 (Perron–Frobenius Theorem, part I) *If A is an $n \times n$ nonnegative and irreducible square matrix, then*

a. $\rho(A) > 0$.
b. *$\rho(A)$ is a simple eigenvalue of A (called the Perron root of A).*
c. *To $\rho(A)$ there corresponds a positive eigenvector. (Such an eigenvector, normalized so that the sum of its entries is 1, is called the Perron vector of A).*

In addition,

d. If $Ax = \lambda x$ and the vector x is positive, then $\lambda = \rho(A)$.
e. If $B \geq A$ and $B \neq A$, then $\rho(B) > \rho(A)$.
f. If $0 \leq B \leq A$ and $B \neq A$, then $\rho(B) < \rho(A)$.

If a nonnegative A is irreducible, so is A^{T}. The spectrum of A^{T} is equal to that of A. Hence there exists a Perron vector y of A^{T}. It satisfies: $y^{\mathrm{T}}A = \rho(A)y^{\mathrm{T}}$. The Perron vectors x and y of A and A^{T} are referred to as the *right-* and *left-* Perron vectors of A, respectively.

The following result follows from Part f of the Theorem 4.

Corollary 5 *If $B \neq A$ is a principal submatrix of a nonnegative irreducible matrix A, then $\rho(B) < \rho(A)$.*

When A is irreducible and nonnegative, but not primitive, there may be several eigenvalues of maximum modulus. The number of eigenvalues of A with modulus $\rho(A)$ is called the *index of cyclicity* of A.

Theorem 6 (Perron–Frobenius Theorem, part II) *If A is a nonnegative and irreducible square matrix with index of cyclicity $k > 1$, then*

a. The eigenvalues of A of modulus $\rho(A)$ are $\rho(A)e^{2\pi i l/k}$, $l = 0, 1, \ldots, k-1$.
b. The spectrum of A goes onto itself under rotation of the complex plane by $2\pi/k$.
c. There exists a permutation matrix P such that

$$PAP^{\mathrm{T}} = \begin{bmatrix} 0 & A_{12} & 0 & \cdots & 0 \\ 0 & 0 & A_{23} & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & A_{k-1\,k} \\ A_{k1} & 0 & 0 & \cdots & 0 \end{bmatrix},$$

where the zero diagonal blocks are square.

Corollary 7 *A nonnegative irreducible matrix A is primitive if and only if its index of cyclicity is 1.*

Using digraphs of nonnegative matrices simplifies the proof of the Perron–Frobenius Theorem. Digraphs are particularly useful in the computation of the index of cyclicity of an irreducible matrix [45]:

Theorem 8 (Romanovsky Theorem) *Let A be an $n \times n$ irreducible nonnegative matrix, and for every $1 \leq i \leq n$ let d_i be the greatest common divisor of all the lengths of closed walks through i in $\Gamma(A)$. Then all the d_i's are equal and are equal to the order of cyclicity of A.*

An actual computation of the index of cyclicity by this theorem may be very hard when A is a large matrix and $\Gamma(A)$ has many cycles, but can be done for small matrices:

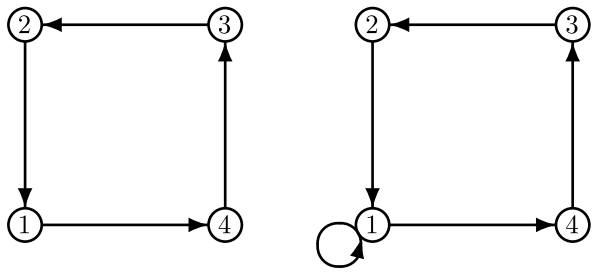

Non-negative Matrices and Digraphs, Figure 2
The digraphs of Examples 9

Example 9 Let

$$A_1 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad \text{and} \quad A_2 = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Then $\Gamma(A_1)$ is the digraph shown on the left and $\Gamma(A_2)$ is the digraph shown on the right in Fig. 2; both are strongly connected.

For $\Gamma(A_1)$ we have $d_1 = d_2 = d_3 = d_4 = 4$ and the index of cyclicity of A_1 is 4. For $\Gamma(A_2)$, the existence of the loop at vertex 1 immediately implies that $d_1 = d_2 = d_3 = d_4 = 1$ and the matrix is primitive.

Combining Romanovsky's Theorem and the Perron–Frobenius Theorem, part II, we get that

Corollary 10 *A nonnegative irreducible matrix A is primitive if* $\mathrm{trace}(A) > 0$.

Every primitive matrix is irreducible (see Theorem 3). However, the converse is not true. For example,

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

is irreducible and not primitive. Note that an $n \times n$, $n > 1$, nonnegative matrix A is irreducible if and only if its digraph is strongly connected, and it is primitive if and only if there exists m such that for every pair of (not necessarily distinct) vertices u and v of $\Gamma(A)$, there is a walk of length m from u to v. The smallest such m, that is, the smallest m for which A^m is positive, is called the *exponent* of A. For information about exponents see Sect. 3.5 in [14]. The index of cyclicity is sometimes called the *period* of the matrix. A primitive matrix is sometimes called an *aperiodic* matrix. We mention also that in a different approach, every 1×1 matrix is considered irreducible, even if it is zero. This corresponds well with the definition of every digraph on one vertex as strongly connected digraph, but requires some changes in the statement of Theorem 4.

The above results are classical, and details may be found in many books, e. g., Chap. 2 in [7], or Chap. 8 in [33].

Special Irreducible Matrices

Nearly Reducible Matrices

A strongly connected digraph Γ is *minimally connected* (or *minimally strong*) if removal of any arc of Γ results in a digraph which is not strongly connected. A matrix is *nearly reducible* if its digraph is minimally connected. That is, an $n \times n$ matrix A, with $n > 1$ is nearly reducible if it is irreducible, and replacing any one nonzero entry by a zero yields a reducible matrix. The 1×1 zero matrix is the only 1×1 nearly reducible matrix (though by our definition it is also reducible). A minimally connected digraph has no loops, and a nearly reducible matrix has zero diagonal.

Nearly reducible matrices have a specific zero-nonzero pattern, described inductively in the following theorem, due to Hartfiel [29]:

Theorem 11 *Let A be an $n \times n$ nearly reducible matrix, $n \geq 2$. Then there exists an integer $1 \leq m < n$ and a permutation matrix P such that*

$$PAP^{\mathrm{T}} = \begin{bmatrix} B & C \\ D & E \end{bmatrix},$$

where B is an $m \times m$ matrix of the form

$$B = \begin{bmatrix} 0 & * & 0 & \cdots & 0 \\ 0 & 0 & * & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & * \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix},$$

where each $*$ stands for a nonzero entry, C is an $m \times (n-m)$ matrix with a single nonzero entry, located in the last row, D is an $(n-m) \times m$ matrix with a single nonzero entry, located in the first column, and E is an $(n-m) \times (n-m)$ nearly reducible matrix.

The following theorem is a translation to nearly reducible matrices of a result of Gupta on minimally connected digraphs [27]:

Theorem 12 *Let A be an $n \times n$ nearly reducible matrix, $n > 1$. Then the number of nonzero entries in A is at least n and at most $2(n-1)$. The lower bound is attained when $\Gamma(A)$ is an n-cycle, and only in this case. The upper bound is attained only when $\Gamma(A)$ is a digraph obtained from a tree by replacing each edge $\{i, j\}$ of the tree with two oppositely directed arcs, (i, j) and (j, i).*

The results concerning nearly reducible matrices and minimally connected digraph may be found in Sect. 3.3 in [14], and Sect. 4.5 in [42]. See also Sect. 29.8 in [31].

Fully Indecomposable Matrices

A concept related to irreducibility is that of full indecomposability. We should first mention that reducible and irreducible matrices are sometimes called *decomposable* and *indecomposable*, which explains the terminology of partly- and fully- decomposable matrices introduced below.

An $n \times n$ matrix A, is *partly decomposable* if $n > 1$ and A has a zero $m \times k$ submatrix with m and k positive integers such that $m + k = n$, or $n = 1$ and $A = 0$. That is, a square matrix A of order greater than 1 is partly decomposable if and only if there exist permutation matrices P and Q such that

$$PAQ = \begin{bmatrix} B & 0 \\ C & D \end{bmatrix},$$

with square diagonal blocks B and D. A square $n \times n$ matrix A is *fully indecomposable* if and only if it is not partly decomposable.

In terms of the bipartite graph of A: A is partly decomposable if there exists a subset $S = \{r_{i_1}, \ldots, r_{i_m}\}$ of the "row vertices" $\{r_1, \ldots, r_n\}$, and a subset $T = \{c_{j_1}, \ldots, c_{j_{n-m}}\}$ of the "column vertices" $\{c_1, \ldots, c_n\}$, such that there are no edges in $B(A)$ from S to T. In terms of the digraph of A: A is partly decomposable if there exist subsets V_1 and V_2 of the vertex set $\{1, \ldots, n\}$, not necessarily disjoint, of sizes m and k such that $m + k = n$, and such that there are no arcs in $\Gamma(A)$ from V_1 into V_2. In comparison, the matrix A is reducible if there exist such subsets V_1 and V_2 which are disjoint and $V_1 \cup V_2 = \{1, \ldots, n\}$.

Clearly a reducible matrix of order greater than 1 is also partly decomposable, and a fully indecomposable matrix is irreducible. The converse does not hold: the matrix

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$$

is an example of a partly decomposable matrix ($V_1 = V_2 = \{1\}$) which is irreducible. However, irreducibility and full indecomposability are not too far apart, as the following results show. For the first of the two see Brualdi, Parter and Schneider [13] and for the second see Brualdi [12].

Theorem 13 *A square matrix A of order greater than 1 is fully indecomposable if and only if for some permutation matrix P the matrix PA is irreducible with all the entries on its main diagonal non-zero.*

Theorem 14 *A square nonnegative matrix A of order greater than 1 is irreducible if and only if $I + A$ is fully indecomposable.*

Henrik Minc opened a course on nonnegative matrices that he gave at the Technion in 1974 (which was the basis for [42]) by stating that the Frobenius–König Theorem is the fundamental result on the zero pattern of square matrices.

Theorem 15 (Frobenius–König Theorem) *A necessary and sufficient condition that every diagonal of an $n \times n$ matrix contains a zero is that the matrix contain an $m \times k$ zero submatrix with $m + k = n + 1$.*

Recall that the *permanent* of a matrix A is given by

$$\mathrm{per}(A) = \sum_{\sigma \in S_n} \prod_{i=1}^{n} a_{i\,\sigma(i)} ,$$

where S_n is the group of all permutations on $\{1, \ldots, n\}$. That is, it is the sum of all diagonal products of A.

Corollary 16 *If A is a square nonnegative matrix and $per(A) = 0$, then A is partly decomposable.*

We mention two more results, due to Marcus and Minc [40]:

Theorem 17 *A nonnegative fully indecomposable matrix is primitive.*

(The converse is not true:

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$$

is primitive but not fully indecomposable).

Theorem 18 *A fully indecomposable $n \times n$ matrix can contain at most $n(n-2)$ zeros.*

For details on fully indecomposable matrices see, e. g., Sect. II.5.4 in [41], and Sect. 4.2 in [14]. It is worth noting that the Frobenius–König Theorem played an important role in the proof of the Perron–Frobenius Theorem (see the discussion in [48]). Another important application of the Frobenius–König Theorem is Birkhoff's Theorem [9], which states that the set of all $n \times n$ doubly stochastic matrices, i. e. matrices that are both row stochastic and column stochastic, forms a convex polyhedron with the permutation matrices as vertices. A slightly more general form of the Frobenius–König Theorem is P. Hall's theorem on systems of distinct representatives [28].

Reducible Nonnegative Matrices and Their Digraphs

For reducible nonnegative matrices, there is also a strong connection between the spectral properties of the matrices and the graph theoretic properties of their digraphs. This connection has been studied extensively, and we review here a sample of the results.

If A is a reducible matrix, there is a permutation matrix P such that

$$PAP^{\mathrm{T}} = \begin{bmatrix} B & 0 \\ C & D \end{bmatrix} .$$

If any of the square diagonal blocks is reducible, we may further reduce the matrix A, until we get a lower block-triangular matrix, with irreducible and/or 1×1 zero diagonal blocks, which is permutationally equivalent to A.

This can be also described by the digraph of A. Let Γ be any digraph on vertices $V = \{1, \ldots, n\}$. We say that a vertex i *has access to* vertex j in Γ if $i = j$ or there is a walk from i to j in Γ. We say that i and j *communicate* if i has access to j and j has access to i. Communication is an equivalence relation on V, and it therefore induces a partition of V into equivalence classes $V_1, \ldots, V_k$. The induced digraphs $\Gamma_i = \Gamma(V_i)$ are strongly connected. These are the *strongly connected components* of Γ.

Let $R(\Gamma)$ be the digraph with vertices $V_1, \ldots, V_k$ and an arc from V_i to V_j, $i \neq j$, if and only if there is a walk from a vertex in V_i to a vertex in V_j (and thus from any vertex in V_i to any vertex in V_j). $R(\Gamma)$ is called the *reduced graph* of Γ.

Theorem 19 *Let Γ be a digraph. Then $R(\Gamma)$ is a simple digraph with no cycles, and its vertices may be labeled $V_1, \ldots, V_k$ so that if (V_i, V_j) is an arc of $R(\Gamma)$ then $i > j$.*

When Γ is $\Gamma(A)$, the digraph of a matrix A, the equivalence classes of the communication relation on $\{1, \ldots, n\}$ are called the *classes* of A. The reduced graph of $\Gamma(A)$ is denoted also by $R(A)$ and referred to as the *reduced graph* of A. The rows and columns of A may be permuted, according to the order of the classes described in Theorem 19. We thus get

Theorem 20 *Let A be an $n \times n$ matrix. Then there exists a permutation matrix P such that*

$$PAP^{\mathrm{T}} = \begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ A_{21} & A_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ A_{k1} & A_{k2} & \cdots & A_{kk} \end{bmatrix} , \qquad (1)$$

where each of the diagonal blocks $A_{11}, \ldots, A_{kk}$ is square and irreducible or a 1×1 zero matrix. The diagonal blocks

are uniquely determined, up to a simultaneous permutation of rows and columns, but their ordering is not necessarily unique.

The block matrix (1) is the *(lower block-triangular) Frobenius normal form* of A. The spectrum of A is the union of the spectrums of the diagonal blocks in the Frobenius normal form. In particular, the spectral radius of A, $\rho(A)$, is an eigenvalue of A, and there exists a class V_i of A such that $\rho(A[V_i]) = \rho(A)$. Such a class is called a *basic class*.

If there is a walk in $R(A)$ from V_i to V_j or $V_i = V_j$, we say that the class V_i *has access* to the class V_j, and denote it by $V_i >= V_j$. We will write $V_i > -V_j$ if V_i has access to V_j and $V_i \neq V_j$. We will also write $i >= V_j$ if i is a vertex and either $i \in V_j$ or there is a walk from i to a vertex in V_j. A class is *final* if it has no access to another class (i. e., its outdegree in $R(A)$ is 0). A class is *initial* if no other class has access to it (that is, its indegree in $R(A)$ is 0). The *level* of a basic class V_i is the maximal number of basic classes on a path in $R(A)$ that terminates at V_i. The level of an initial basic class is 1.

Example 21 Let

$$A = \begin{bmatrix} 0 & 1 & 0 & 4 & 0 & 0 & 0 \\ 3 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 5 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 7 & 0 \\ 0 & 8 & 0 & 9 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Then $\Gamma(A)$ is the digraph shown on the left in Fig. 3, and $R(A)$ is the digraph shown on the right.

The classes are $\{1,2\}$, $\{3,7\}$, $\{4\}$, $\{5\}$ and $\{6\}$. The classes $\{5\}$ and $\{3,7\}$ are initial classes, $\{4\}$ is a final class. One labeling that fits the requirements of Theorem 19 is $V_1 = \{4\}$, $V_2 = \{1,2\}$, $V_3 = \{6\}$, $V_4 = \{5\}$ and $V_5 = \{3,7\}$. Let

$$P = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

be the permutation matrix corresponding to the permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 4 & 1 & 2 & 6 & 5 & 3 & 7 \end{pmatrix},$$

then

$$PAP^{\mathrm{T}} = \left[\begin{array}{c|c|c|c|c|c|c} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 4 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 3 & 2 & 0 & 0 & 0 & 0 \\ \hline 9 & 0 & 8 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 7 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 5 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 & 3 & 0 \end{array}\right]$$

is A's Frobenius normal form. The spectra of the diagonal blocks are $\{0\}$, $\{3,-1\}$, $\{0\}$, $\{0\}$, $\{3,-3\}$. Hence V_2 and V_5 are the basic classes of A.

Finally, the basic classes V_2 and V_5 are of levels 2 and 1, respectively.

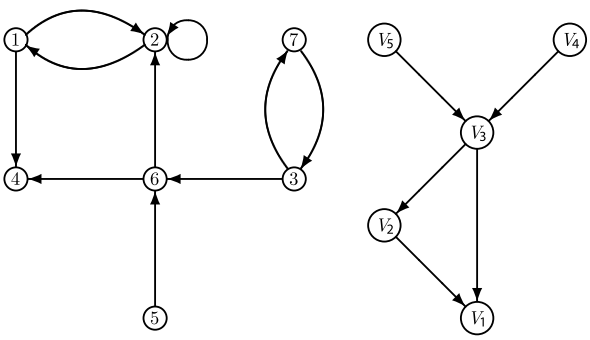

Non-negative Matrices and Digraphs, Figure 3
The digraph and reduced digraph of Example 2

The algebraic multiplicity of the Perron eigenvalue $\rho(A)$ is obviously equal to the number of basic classes of A. The size of the largest Jordan block corresponding to $\rho(A)$ can also be "read" from the reduced graph of A. Recall that the size of the largest Jordan block associated with the eigenvalue $\rho(A)$ is the multiplicity of $\rho(A)$ as a root of the minimal polynomial of A. It is called the *index* of the Perron eigenvalue $\rho(A)$ of the matrix A and we denote it by $\nu = \nu(A)$. Equivalently, the index of $\rho(A)$ is the smallest nonnegative integer ν such that $\mathcal{N}((\rho(A)I - A)^{\nu}) = \mathcal{N}((\rho(A)I - A)^{\nu+1})$, where $\mathcal{N}(B)$ denotes the nullspace of a matrix B. The following was proved in [46]:

Theorem 22 (Rothblum Index Theorem) *Let A be a square nonnegative matrix. Then the maximal level of a basic class is equal to $\nu(A)$.*

The matrix A in Example 21 thus has index $\nu(A) = 2$.

Theorem 22 implies the following (earlier) result by Schneider [47]:

Corollary 23 *Let A be a square nonnegative matrix. Then in the Jordan Canonical form of A there is only one*

block corresponding to the eigenvalue $\rho(A)$ (equivalently, $\dim\mathcal{N}((\rho(A)I - A)) = 1$) if and only if for every two basic classes in R(A), one of the two classes has access to the other.

When A is a reducible nonnegative matrix, there need not exist a positive eigenvector corresponding to $\rho(A)$. The following theorem appeared in vol. II, p. 77 of [24]:

Theorem 24 *Let A be a nonnegative matrix. There exists a positive eigenvector corresponding to the spectral radius of A if and only if the final classes of A are exactly the basic ones.*

The *algebraic eigenspace* (or the *Perron generalized eigenspace*) of A, denoted by $E(A)$, is defined to be $\mathcal{N}((\rho(A)I - A)^n) = \mathcal{N}((\rho(A)I - A)^\nu)$. The nonzero elements of the algebraic eigenspace of A are called *generalized eigenvectors* of A associated with the eigenvalue $\rho(A)$. There are several results on the existence of various nonnegative bases to the algebraic eigenspace of nonnegative matrix A. The first part of the next theorem is the *Nonnegative Basis Theorem*, due to Rothblum in [46], and the second is the *Preferred Basis Theorem* due to Richman and Schneider [44]. For an n-vector x, $\mathrm{supp}(x)$ denotes the *support* of the vector x, that is, $\mathrm{supp}(x) = \{1 \le i \le n \mid x_i \ne 0\}$.

Theorem 25 *Let A be a square nonnegative matrix with basic classes $V_{i_1}, \ldots, V_{i_m}$. Then*

a. There exists a set of vectors $\{x^1, \ldots, x^m\}$ in E(A) such that

$$x^j \ge 0 \quad \text{and}$$
$$\mathrm{supp}\left(x^j\right) = \{i \mid i >= V_{i_j}\}, \quad 1 \le j \le m\,. \tag{2}$$

Moreover, any such set of nonnegative vectors forms a basis of E(A).

b. There exists a basis for E(A) that satisfies in addition to (2) also

$$(A - \rho(A)I)x^j = \sum_{V_{i_l} \succ V_{i_j}} c_{jl}\, x^l, \quad 1 \le j \le m, \tag{3}$$

where the c_{jl} are positive.

A basis of $E(A)$ which satisfies (2) and (3) is called a *preferred-basis*. One preferred-basis for the matrix A in Example 21 is $\{x^1, x^2\}$, where

$$(x^1)^{\mathrm{T}} = [\ 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 1\ ] \quad \text{and}$$
$$(x^2)^{\mathrm{T}} = [\ 1/20 \quad 3/20 \quad 1 \quad 0 \quad 14/15 \quad 2/5 \quad 2/3\ ]\,.$$

The existence of a nonnegative Jordan basis was also studied. Recall that a *Jordan chain* of length r corresponding to the eigenvalue λ of A is a sequence of r nonzero vectors $x^0 = x, x^1 = (A - \lambda I)x, \ldots, x^{r-1} = (A - \lambda I)^{r-1}x$ such that $(A - \lambda I)^r x = 0$. A *Jordan basis* for the algebraic eigenspace $E(A)$ of a nonnegative matrix A is a basis of $E(A)$ consisting of several Jordan chains corresponding to the Perron eigenvalue $\rho(A)$. As is well-known, there always exists a Jordan basis for $E(A)$. However, there does not always exist a nonnegative Jordan basis (i. e., a Jordan basis all of whose vectors are nonnegative) for the algebraic eigenspace of a nonnegative matrix A. Necessary and sufficient conditions for the existence of a nonnegative Jordan basis were given by Richman and Schneider [44]. To state the theorem, we need some definitions and notations. Let l_i denote the number of basic classes of level i. By the Rothblum Index Theorem, there is a basic class of level ν, and no basic class of higher level. We denote by $l(A)$ the sequence of levels $(l_1, l_2, \ldots, l_\nu)$. Let $j(A)$ be the non-increasing sequence $(j_1, j_2, \ldots, j_r)$ of the sizes of the Jordan blocks corresponding to $\rho(A)$ (of course, $j_1 = \nu$). We need also the notion of a *dual sequence*. If $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_r)$ is a sequence of non-increasing positive integers, the dual sequence α^* is obtained so: Create a diagram of r left-aligned rows of dots, α_i dots in the ith row from the top. Let α^* be the (non-increasing) sequence of column-lengths of the diagram. For example, if $\alpha = (4, 3, 2, 2, 1)$, then $\alpha^* = (5, 4, 2, 1)$, as seen in the following diagram:

$$\begin{array}{cccc}
\bullet & \bullet & \bullet & \bullet \\
\bullet & \bullet & \bullet & \\
\bullet & \bullet & & \\
\bullet & \bullet & & \\
\bullet & & &
\end{array}$$

With these definitions, Richman and Schneider's theorem may now be stated.

Theorem 26 *Let A be an $n \times n$ nonnegative matrix. Then the following are equivalent:*

a. There exists a nonnegative Jordan basis for the algebraic eigenspace E(A).
b. For every $1 \le l \le \nu$ there exists a nonnegative basis for $\mathcal{N}((A - \rho(A)I)^l)$.
c. $l(A) = j(A)^*$.

For the matrix A of Example 21, there are two basic classes, one of level 2 and one of level 1. Hence $l(A) = (1, 1)$. The index of A is 2, and the algebraic multiplicity of $\rho(A) = 3$ is also 2, so $j(A) = (2)$ (see also Corollary 23. Clearly

$j(A)^* = l(A)$, so there exists a nonnegative Jordan basis for the algebraic eigenspace $E(A)$. Indeed, the preferred basis of A presented above is also a Jordan basis for $E(A)$.

A reducible nonnegative matrix may have nonnegative eigenvectors associated with eigenvalues other than the Perron eigenvalue. An eigenvalue of A is a *distinguished eigenvalue* if it has a corresponding nonnegative eigenvector. A distinguished eigenvalue is necessarily nonnegative. A class V_i in $R(A)$ is a *distinguished class* if $\rho(A_{ii}) > \rho(A_{jj})$ for every $j \neq i$ such that V_j has access to V_i. In Example 21 the basic class V_5 is distinguished, as is the (non-basic) class V_4, while the basic class V_2 and the classes V_1 and V_3 are not distinguished.

The following theorem was proved by Victory [54], and was already implicit in the original paper by Frobenius [23].

Theorem 27 (Frobenius–Victory Theorem) *Let A be an $n \times n$ nonnegative matrix. Then*

a. *A real number λ is a distinguished eigenvalue of A if and only if there exists a distinguished class V_i of A such that $\rho(A_{ii}) = \lambda$.*
b. *If V_i is a distinguished class of A and $\lambda = \rho(A_{ii})$, then there is a nonnegative eigenvector x corresponding to λ such that*

$$\operatorname{supp}(x) = \{1 \le j \le n \mid j >= V_i \text{ in } \Gamma(A)\}\,.$$

The vector x with these properties is unique up to a positive scalar multiple.

For the matrix A in Example 21, in addition to $\rho(A) = 3$ the eigenvalue 0, which is $\rho(A_{44})$, is a distinguished eigenvalue. The only vertex with access in $\Gamma(A)$ to $V_4 = \{5\}$ is 5 itself, and

$$x = [\ 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0\ ]^{\mathrm{T}}$$

is the unique nonnegative eigenvector corresponding to 0 whose support is $\{5\}$.

Some remarks about the terminology: In Rothblum's paper *level* was termed *height*. Later *level* was used (see [30]), reserving the term *height characteristic* to represent the *Weyr's characteristic*, that is, the sequence $\eta(A) = (\eta_1, \ldots, \eta_\nu)$ with $\eta_i = \dim \mathcal{N}((\rho(A)I - A)^i) - \dim \mathcal{N}((\rho(A)I - A)^{i-1})$. The sequence $l(A)$ is called the *level characteristic* of A, and the sequence $j(A)$ of Jordan blocks is known as the *Segré characteristic*, e. g. [30]. *Nonnegative bases* are referred to as *semipositive bases* in many of the works mentioned in this section, where a *semipositive* matrix or vector is a nonnegative matrix or vector which is not equal to zero.

The results here are over 20 years old – see [49], where additional references may be found. Much work was done on this subject in the years since then, which we don't cover here. In a series of papers by Tam, Tam and Wu, Tam and Schneider, the above results, and more, were extended to cone-preserving maps, providing also alternative proofs for the case of nonnegative matrices. The latest installment in this series of papers is [51], where reference can be found to the previous papers in the series, as well as to works of others. The existence of nonnegative bases, relations between characteristics of the reduced graph $R(A)$ and characteristics of the matrix A, and extensions of such results to general (not necessarily nonnegative) matrices, were further studied by Hershkowitz, Schneider and others, see [30] and the references therein. See also Sect. 9.3 in [31], and the references there.

Examples: Ranking Problems

Google PageRank Revisited

We now consider in more detail the Google PageRank Example.

In the Google model, we may think of a weighted digraph W representing the links on the Web. Each of the N pages is represented by a vertex, each link is represented by an arc, and a weight is assigned to each arc: If the outdegree of a vertex i is d_i, the weight of each of the d_i arcs is $1/d_i$. Let T be the matrix associated with this weighted digraph, that is, t_{ij} is $1/d_i$ if (i, j) is an arc in W, and 0 otherwise; Make T row stochastic by replacing each zero row by $\frac{1}{N}\, e^{\mathrm{T}}$, where e is the N-vector with all entries equal to 1. Now $Te = e$. The Google matrix is:

$$P = \alpha T + (1 - \alpha)\,\frac{1}{N}\,J\,,$$

where J is the (in this case, $N \times N$) matrix of all ones and $\alpha = 0.85$. By this representation of P it is easy to see that P is positive and that $Pe = e$ (since both $Te = \frac{1}{N}Je = e$). This implies (by Theorem 4 d) that 1 is the Perron root of P. Since a matrix and its transpose have the same spectrum, 1 is also the Perron root of P^{T}. The existence of a unique Perron vector π for P^{T} is therefore guaranteed by the Perron Theorem.

It is not known how exactly Google calculates (an approximation to) π, but probably some version of the power method is used. This is done essentially by starting with a vector x_0 and computing $\lim_{m\to\infty} x_m$, where $x_m = P^{\mathrm{T}} x_{m-1} = \left(P^{\mathrm{T}}\right)^m x_0$. By Perron's Theorem (Theorem 1 e), for every vector x_0 whose entries sum up to 1, $\left(P^{\mathrm{T}}\right)^m x_0$ converges to $\pi e^{\mathrm{T}} x_0 = \pi$. Part f of the Perron Theorem gives an idea about the rate of convergence: By

a known bound for a row stochastic matrix (Chap. 5, Theorem 5.10 in [7])

$$\mu(P) \leq 1 - \sum_{j=1}^{N} \min_i p_{ij} \,.$$

The construction of P and the Web structure imply that the quantity on the right hand side is α. Hence the choice of α determines the rate of convergence of the power method: it is roughly the rate of convergence of α^m to zero (see Theorem 1 f).

In the definition of P, the matrix T is the crucial part to the idea of Google, of ranking pages according to the hyperlink structure of the Web. It is T that contains the information on that hyperlink structure. But the matrix T is not irreducible. (The digraph representing the Web is certainly not strongly connected. Studies of the Web suggest that it has a large strongly connected component, "the core", but also three other, equally large, groups of vertices: one consisting of vertices which have access to the core, but no access from the core; another consisting of vertices which can be accessed from the core, but have no access to the core; and finally, there is a group of vertices which have no access to or from the core [15]). The addition of the matrix $\frac{1}{N}J$ may be explained as we did in Sect. "Matrices, Graphs and Digraphs", but technically it has the role of making the matrix P irreducible (in fact, positive) thus guaranteeing the existence of a unique PageRank vector. If we replace the matrix $\frac{1}{N}J$ by a matrix $E = ev^{\mathrm{T}}$, where v is a positive probability vector (i. e., $v^{\mathrm{T}}e = 1$), the new matrix P would be again positive and row stochastic. Of course, a choice of $v \neq \frac{1}{N}e$ ($E \neq \frac{1}{N}J$) means that certain pages are favored. Indeed, it is reported that Google now uses such biased perturbation matrices in the construction of P. The choice of α is the result of a balancing act: large α means that more weight is given to the hyperlink structure of the Web; small α means faster convergence to the PageRank vector.

PageRank is not the only Web information retrieval method that is based on the hyperlink structure of the Web and uses nonnegative matrices associated with the Web digraph. See [38] for an illuminating survey on three such methods, PageRank included.

Tournaments

Suppose that n players participate in a round-robin tournament. In such a tournament, every player plays against every other player once, and there is a clear winner in each of the games. The outcomes of the games can be represented by a digraph on n vertices, each representing a player, with (i, j) an arc if and only if player i beats player j. Such a digraph has no loops, and for every $i \neq j$ exactly one of (i, j) and (j, i) is an arc. It is called a *tournament*. The adjacency matrix of a tournament is called a *tournament matrix*. That is, a $(0, 1)$-matrix A is a tournament matrix if A if and only if $A + A^{\mathrm{T}} = J - I$. Figure 3 shows a tournament on 4 vertices and next to it its adjacency matrix.

The problem is: Given the outcomes of all games in a round-robin tournament, how does one rank the players? It seems that it is not enough to count the scores of the players. As in the Google PageRank example, it seems reasonable to rate players not only by their scores, but also by the scores of the players they beat.

The vector of scores of the n players is Ae. The vector of second-level scores of the players, whose ith component is the sum of the scores of the players defeated by player i, is $A(Ae) = A^2e$. The vector of kth-level scores is $A^k e$. The entries of the kth-level scores grow large with k, but we are interested in their relative magnitude, so we may normalize each kth-level scores vector. By Theorem 2, if the matrix A is primitive, then

$$\lim_{k\to\infty} \left(\frac{A}{\rho(A)}\right)^k e = s \,,$$

where s is a scalar multiple of the Perron vector of A. The vector s can then be used to rank the players in the tournament A: the larger s_i is, the higher is the rank of player i. For example, the tournament matrix in Fig. 4 is primitive, its Perron root is $\rho = 1.3953$, and $[0.6256, 0.5516, 0.3213, 0.4484]^{\mathrm{T}}$ is a corresponding eigenvector, which leads to the ranking of 1 as the top player, followed by 2, then 4, and finally 3 in the last place.

A tournament matrix need not be primitive or even irreducible; however, by the next theorem every irreducible tournament matrix of order greater than 3 is primitive.

Theorem 28 *An $n \times n$ irreducible tournament matrix is primitive if and only if $n \geq 4$.*

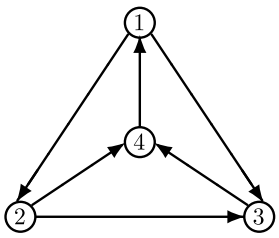

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

Non-negative Matrices and Digraphs, Figure 4
A tournament and its tournament matrix

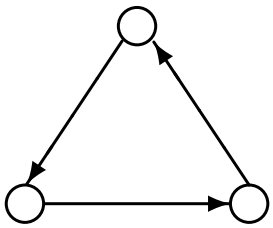

Non-negative Matrices and Digraphs, Figure 5
A 3-cycle tournament

Tournament matrices of order 1 are 2 are reducible, and the only irreducible tournament matrices on 3 vertices are the matrices whose digraph is a 3-cycle.

All these irreducible 3×3 tournament matrices are not primitive. It is quite obvious from Fig. 5 that in these tournaments all three players should have equal rank in any reasonable ranking method.

Here is how we may rank the players when the tournament matrix A is reducible. Let P be a permutation matrix such that

$$PAP^{\mathrm{T}} = \begin{bmatrix} A_{11} & 0 & \cdots & 0 \\ A_{21} & A_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ A_{k1} & A_{k2} & \cdots & A_{kk} \end{bmatrix}$$

is in Frobenius normal form, with each diagonal block square and irreducible (of size at least 3×3), or a 1×1 zero matrix. Let V_i, $i = 1, \ldots, k$, be the classes of A (V_i corresponds to A_{ii}). Since A is a tournament matrix, each of the blocks below the diagonal has all entries equal to 1. Thus in this case V_i has access to V_j if and only if $i > j$ (which means that each player in V_i beats each player in V_j for every $j < i$). Now if $|V_i| = 3$, rank all the players in V_i as equal. If $|V_i| > 3$, rank the players in V_i according to the Perron vector of A_{ii}, and for $i > j$ rank all players in V_i higher than the players in V_j. This is the Kendall–Wei ranking ([34,56]).

For the results of this section see Sect. 10.7 in [10]. See also [22] and the references therein for more on tournament matrices and ranking tournament players.

The Laplacian Matrix of a Directed Graph

Let Γ be a digraph with no loops on vertices $\{1, \ldots, n\}$. Let $A = A(\Gamma)$ be the adjacency matrix of A, and let $D = D(\Gamma)$ be the diagonal matrix of outdegrees of the vertices of Γ (that is, d_{ii} is the outdegree of the vertex i). The matrix $L(\Gamma) = D - A$ is the *Laplacian Matrix* of Γ. Observe that $L(\Gamma)e = 0$. If Γ is a weighted digraph, we may replace the matrix A with the matrix $\mathcal{A} = \mathcal{A}(\Gamma)$ whose i, j entry is the weight of the arc (i, j), and the diagonal matrix of outdegrees by the diagonal matrix $\mathcal{D} = \mathcal{D}(\Gamma)$ whose i, i entry is the sum of the weights of the arcs with i as initial vertex. The Laplacian of the weighted digraph Γ is $\mathcal{L}(\Gamma) = \mathcal{D}(\Gamma) - \mathcal{A}(\Gamma)$.

An *arborescence* is a digraph with no cycles, in which all the vertices but one vertex u have outdegree 1, and u has outdegree 0. The vertex u is called the *root* of the arborescence. That is, an arborescence is a digraph whose underlying graph is a tree, in which all the arcs are directed towards the root u. A *spanning arborescence* of a digraph Γ is a subdigraph of Γ with the same set of vertices as Γ, which is an arborescence. The Matrix-Tree Theorem below is due to Tutte [52].

Theorem 29 (The Matrix-Tree Theorem) *Let $\mathcal{L}(\Gamma)$ be the Laplacian matrix of a weighted simple digraph Γ. Let $\mathcal{L}_i(\Gamma)$ be the principal submatrix of $\mathcal{L}(\Gamma)$ obtained by deleting the ith row and ith column. Then* $\det(\mathcal{L}_i(\Gamma))$ *is the sum of weights of all spanning arborescences of Γ rooted at the vertex i, where the weight of a subdigraph is the product of the weights of its arcs.*

A non-weighted digraph Γ, may be considered as having weight 1 assigned to each arc. With these weights, $\mathcal{L}(\Gamma) = L(\Gamma)$. Thus we have:

Corollary 30 *Let $L(\Gamma)$ be the Laplacian matrix of a simple digraph Γ. Let $L_i(\Gamma)$ be the principal submatrix of $L(\Gamma)$ obtained by deleting the ith row and ith column. Then* $\det(L_i(\Gamma))$ *is the number of spanning arborescences of Γ rooted at the vertex i.*

For the results here see Sect. 9.6 in [14], and Chap. VI in [53]. They are generalizations of Kirchhoff's Theorem on the (combinatorial) Laplacian of a graph. A more general form of the Matrix-Tree Theorem is the All Minors Matrix-Tree Theorem, which relates minors of order $n - k$ of the Laplacians to sums of weights of certain spanning (directed) forests consisting of k disjoint arborescences. See e. g. [16] and the references there. See also [17] and the references there for more on such spanning forests and their relations to the Laplacian.

As mentioned in Sect. "Matrices, Graphs and Digraphs" of this article, for graphs there are many results on the connections between the spectrum of the Laplacian and the structure of the graph (see [18,21] and the references therein). There aren't many such results concerning the spectrum of the Laplacian of a digraph. Attempts in this direction have only begun in recent years, and we will touch upon these briefly in the next section, on research directions.

Research Directions

We conclude by mentioning a few directions in current research.

Relating analytic properties of a nonnegative matrix to properties of its digraph: The Romanovsky Theorem is an (old) example of such a theorem. More recent results in this spirit can be found in [35,36,37]. In the first of these, a lower bound on the second largest eigenvalue of a stochastic matrix A is given in terms of the girth (i. e., the length of the shortest cycle) of its digraph. In the second and third, directed graphs are used in measuring the sensitivity of stationary distributions (i. e., the Perron left eigenvector of the transition matrix) under perturbations of the transition matrix.

Spectral theory for digraphs: One disadvantage of the Laplacian of a directed graph, compared to that of a graph, is that it is not symmetric and need not have real eigenvalues. Fan Chung [19] and Chai Wah Wu [57] independently derived from the Laplacian $L = D - A$ of a digraph Γ a symmetric matrix whose eigenvalues may be considered instead of the eigenvalues of the original Laplacian. Following Chung, for a strongly connected digraph Γ let

$$S = \Phi - \frac{1}{2}\left(\Phi P + P^{\mathrm{T}}\Phi\right) ,$$

where $P = D^{-1}A$ and Φ is the diagonal matrix with the left Perron vector of P on its diagonal. By scaling S we get a normalized version:

$$S = \Phi^{-1/2} S \Phi^{-1/2} .$$

The eigenvalues of S (see [57]; Wu's matrix is actually a scalar multiple of S) or of S (see [19,20]) may be considered the directed graph equivalents of the eigenvalues of the combinatorial or normalized Laplacian of a graph. Both Chung and Wu examine connections between these eigenvalues (mainly the second smallest one) and some digraph parameters (such as the diameter and the Cheeger constant). This study seems to be at its beginning.

Searching the Web and Google: There is a lot of activity of analyzing the Google matrix, the ideas behind Google, possible improvements and alternative approaches. The best way to find the current state of affairs of these subjects is to google related terms (try that on http://scholar.google.com).

Bibliography

1. Alon U (2006) An Introduction to Systems Biology: Design Principles of Biological Circuites. CRC Mathematical and Computational Biology Series. Chapman, Boca Raton
2. Bang-Jensen J, Gutin G (2001) Digraphs: Theory Algorithms and Applications. Springer, London
3. Bapat RB, Raghavan TES (1997) Nonnegative Matrices and Applications. Encyclopedia of Mathematics and its Applications 64. Cambridge University Press, Cambridge
4. Beineke LW, Wilson RJ (2004) Topics in Algebraic Graph Theory. Encyclopedia of Mathematics and its Applications 102. Cambridge University Press, Cambridge
5. Berman A (2003) Graphs of matrices and matrices of graphs. Numer Math J Chin Univ (Engl Ser) 12(suppl):12–14
6. Berman A, Neumann M, Stern RJ (1989) Nonnegative Matrices in Dynamical Systems. Wiley-Interscience, New York
7. Berman A, Plemmons R (1994) Nonnegative Matrices in the Mathematical sciences. SIAM, Philadelphia
8. Berman A, Shaked-Monderer N (2003) Completely Positive Matrices. World Scientific, River Edge
9. Birkhoff G (1946) Tres obseraciones sobre el algebra lineal. Univ Nac Tucuman Rev Ser A 5:147–150
10. Bondy JA, Murty USR (1976) Graph Theory with Applications. North-Holland, New York
11. Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine. Comput Netw ISDN Syst 33(1–7):107–117
12. Brualdi RA (1979) Matrices permutation equivalent to irreducible matrices and applications. Linear Multilinear Algebra 7:1–12
13. Brualdi RA, Parter SV, Schneider H (1966) The diagonal equivalence of a nonnegative matrix to a stochastic matrix. J Math Anal Appl 16:31–50
14. Brualdi RA, Ryser HJ (1991) Combinatorial Matrix Theory. Encyclopedia of Mathematics and its Applications 39. Cambridge University Press, Cambridge
15. Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener J (2000) Comput Netw 33(1–3): 309–320
16. Chaiken S (1982) A combinatorial proof of the all-minors matrix tree theorem. SIAM J Alg Dis Meth 3:319–329
17. Chebotarev P, Agaev R (2002) Forest matrices around the Laplacian matrix. Linear Algebra Appl 356:253–274
18. Chung FRK (1997) Spectral Graph Theory. CBMS, Regional Conference Series in Mathematics 92. AMS, Providence
19. Chung F (2005) Laplacians and the Cheeger inequality for directed graphs. Ann Comb 9:1–19
20. Chung F (2006) The diameter and Laplacian eigenvalues of a directed graphs. Electron J Combin 13(4):6
21. Cvetković D, Doob M, Sachs H (1980) Spectra of Graphs: Theory and Applications. Academic Press, New York
22. Eshenbach C, Hall F, Hemansinha R, Kirkland S, Li Z, Shader B, Stuart J, Weaver J (2000) Properties of Tournaments among Well-Matched Players. Amer Math Month 107(10):881–892
23. Frobenius G (1912) Über Matrizen aus nicht negativen Elementen. Sitzungsbericht. Preussische Akademie der Wissenschaften, Berlin, pp 456–477
24. Gantmacher FR (1959) The Theory of Matrices. (Translated from Russian), Chelsea, New York
25. Godsil C, Royle G (2001) Algebraic Graph Theory. Graduate Texts in Mathematics 207. Springer, New York
26. Gulli A, Signorini A (2005) The indexable web is more than 11.5 billion pages. In: Proc. 14th WWW (Posters), pp 902–903
27. Gupta RP (1967) On basis digraphs. J Combin Theory 3: 309–311

28. Hall P (1935) On representatives of subsets. J London Math Soc 10:26–30
29. Hartfiel DJ (1970) A simplified form for nearly reducible and nearly decomposable matrices. Proc Amer Math Soc 24: 388–393
30. Hershkowitz D (1999) The combinatorial structure of generalized eigenspaces – from nonnegative matrices to general matrices. Linear Algebra Appl 302–303:173–191
31. Hogben L (2007) Handbook of Linear Algebra. Discrete Mathematics and Its Applications Series 39. CRC Press, Boca Raton
32. van der Holst H, Lovász L, Schrijver A (1999) The Colin de Verdière graph parameter. In: Graph Theory and Computational Biology. Bolyai Soc Math Stud 7:29–85
33. Horn RA, Johnson CR (1985) Matrix Analysis. Cambridge University Press, Cambridge
34. Kendall MG (1955) Further contributions to the theory of paired comparisons. Biometrics 11:43–62
35. Kirkland S (2004) Digraph-based conditioning for Markov chains. Linear Algebra Appl 385:81–93
36. Kirkland SJ (2004) A combinatorial approach to the conditioning of a single entry in the stationary distribution for a Markov chain. Electron J Linear Algebra 11:168–179
37. Kirkland SJ (2004/5) Girth and subdominant eigenvalues for stochastic matrices. Electron J Linear Algebra 12:25–41
38. Langville AN, Meyer CD (2005) A survey of eigenvector methods of Web information retrieval. SIAM Rev 47(1):135–161
39. Lee D, Seung H (2001) Algorithms for nonnegative matrix factorization. Adv Neural Process 13:556–562
40. Marcus M, Minc H (1963) Disjoint pairs of sets and incidence matrices. Illinois J Math 7:137–147
41. Marcus M, Minc H (1964) A survey of matrix theory and matrix inequalities. Allyn and Bacon, Boston
42. Minc H (1988) Nonnegative Matrices. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley, New York
43. Perron O (1907) Zur Theorie der Matrizen. Math Ann 64: 248–263
44. Richman D, Schneider H (1978) On the singular graph and the Weyr characteristic of an *M*-matrix. Aequationes Math 17: 208–234
45. Romanovsky V (1936) Recherche sur les chains de Markoff. Acta Math 66:147–251
46. Rothblum UG (1975) Algebraic eigenspaces of nonnegative matrices. Linear Algebra Appl 12:281–292
47. Schneider H (1956) The elementary divisors associated with 0 of a singular *M*-matrix. Proc Edinburgh Math Soc 10(2):108–122
48. Schneider H (1977) The concepts of irreducibility and full indecomposability of a matrix in the works of Frobenius, König and Markov. Linear Algebra Appl 18:139–162
49. Schneider H (1986) The influence of the marked reduced graph of a nonnegative matrix on the Jordan Form and on related properties: a survey. Linear Algebra Appl 84:161–189
50. Seneta E (1981) Nonnegative Matrices and Markov Chains. Springer Series in Statistics, 2nd edn. Springer, New York
51. Tam BS (2004) The Perron generalized eigenspace and the spectral cone of a cone-preserving map. Linear Algebra Appl 393:375–429
52. Tutte WT (1948) The dissection of equilateral triangles into equilateral triangles. Proc Cambridge Philos Soc 7:463–482
53. Tutte WT (1980) Graph Theory. Encyclopedia of Mathematics and its Applications 21. Cambridge University Press, Cambridge
54. Victory Jr HD (1985) On nonnegative solutions to matrix equations. SIAM J Alg Dis Meth 6:406–412
55. Wasserman S, Faust K (1994) Social network analysis: Methods and applications. Cambridge University Press, Cambridge
56. Wei TH (1952) The Algebraic Foundations of Ranking Theory. Ph D Thesis, Cambridge University
57. Wu CW (2005) On Rayleigh-Ritz rations of a generalized Laplacian matrix of directed graphs. Linear Algebra Appl 402:207–227

Nonparametric Tests for Independence

Cees Diks
University of Amsterdam, Amsterdam, The Netherlands

Article Outline

Glossary
Definition of the Subject
Introduction
Invariant Tests
Tests Based on Divergence Measures
Tests Based on Other Measures of Dependence
Bootstrap and Permutation Tests
Future Directions
Bibliography

Glossary

Hypothesis A hypothesis is a statement concerning the (joint) distribution underlying the observed data.

Nonparametric test In contrast to a parametric test, a nonparametric test does not presume a particular parametric structure concerning the data generating process.

Serial dependence Statistical dependence among time series observations.

Time series A sequence of observed values of some variable over time, such as a historical temperature record, a sequence of closing prices of a stock, etc.

Definition of the Subject

One of the central goals of data analysis is to measure and model the statistical dependence among random variables. Not surprisingly, therefore, the question whether two or more random variables are statistically independent can be encountered in a wide range of contexts. Although this article will focus on tests for independence among time series data, its relevance is not limited to the time series context only. In fact many of the dependence measures discussed could be utilized for testing independence between

random variables in other statistical settings (e. g. cross-sectional dependence in spatial statistics).

When working with time series data that are noisy by nature, such as financial returns data, testing for serial independence is often a preliminary step carried out before modeling the data generating process or implementing a prediction algorithm for future observations. A straightforward application in finance consists of testing the random walk hypothesis by checking whether increments of, for instance, log prices or exchange rates, are independent and identically distributed [8,12,80]. Another important application consists of checking for remaining dependence structure among the residuals of an estimated time series model.

Introduction

Throughout this article it will be assumed that $\{X_t\}$, $t \in \mathbb{Z}$, represents a strictly stationary time series process, and tests for serial independence are to be based on an observed finite sequence $\{X_t\}_{t=1}^n$. Unless stated otherwise, it will be assumed that the observations X_t take values in the real line $\mathbb{R}$. Admittedly, this is a limitation, since there are also time series processes that do not take values in the real line. For instance, the natural space in which wind direction data take values is the space of planar angles, which are naturally represented by the interval $[0, 2\pi]$ with the endpoints identified. However, most of the tests developed to date are designed for the real-valued case. The problem under consideration is that of testing the null hypothesis that the time series process $\{X_t\}$ consists of independent, identically distributed (i.i.d.) random variables. In practice this is tested by looking for dependence among of m consecutive observations $X_{t-m+1}, \ldots, X_t$ for a finite value $m \geq 2$.

Traditionally, tests for serial independence have focused on detecting serial dependence structure in stationary time series data by estimating the autocorrelation function (acf), $\rho_k = \mathrm{Cov}(X_{t-k}, X_t)/\mathrm{Var}(X_t)$, or the normalized spectral density, which is one-to-one related to the acf by Fourier transformation:

$$h(\omega) = (2\pi)^{-1} \sum_{k=-\infty}^{\infty} \rho_k \mathrm{e}^{-i\omega k}$$

$$\text{and} \quad \rho_k = \int_{-\pi}^{\pi} h(\omega) \mathrm{e}^{ik\omega} \mathrm{d}\omega \,.$$

Because the acf is real and symmetric ($\rho_k = \rho_{-k}$) the normalized spectral density is real and symmetric also. Since the acf and the normalized spectral density are related by an invertible transformation, they carry the same information regarding the dependence of a process. For i.i.d. processes with finite variance, $\rho_k = 0$ for $k \geq 1$ under the null hypothesis. The spectral density is flat (equal to 1 for all ω) in that case.

Tests based on the acf date back to Von Neumann [83]. Motivated by the aim to test for serial independence against the presence of trends, he introduced the ratio of the mean square first difference to the sample variance,

$$S_n := \frac{\frac{1}{n-1}\sum_{t=2}^n (X_t - X_{t-1})^2}{\frac{1}{n}\sum_{t=1}^n (X_t - \bar{X})^2} \,,$$

which may be considered a rescaled (and shifted) estimator of ρ_1. Von Neumann studied the distributional properties of this statistic in detail under the assumption of normality. Durbin and Watson [38,39] used an analogue of Von Neumann's ratio to check for first order autocorrelation among the error terms $\{\varepsilon_t\}_{t=1}^n$ in a linear regression model, based on observed residuals $\{\widehat{\varepsilon}_t\}_{t=1}^n$. As for the original statistic of Von Neumann, the null distribution (which is no longer unique in this case, but depends on the parameters of the data generating process) has been studied in detail for the normal case [40,56,101]. For the class of autoregressive integrated moving average (ARIMA) processes, acf-based tests for residual autocorrelation beyond lag 1 were proposed by Box and Jenkins [13] and Ljung and Box [79], and for autocorrelation in squared residuals by McLeod and Li [82]. Beran [9] proposed adapted tests for serial independence for processes with long-range dependence.

Although the autocovariance structure of a time series process fully characterizes the dependence structure within classes of linear Gaussian random processes, tests based solely on the acf may clearly fail to be consistent against dependence that does not show up in the acf. It is not hard to construct examples of processes for which this is the case. For instance, the bilinear process

$$X_t = aX_{t-2}\varepsilon_{t-1} + \varepsilon_t, \quad (|a| < 1)$$

where $\{\varepsilon_t\}$ is a sequence of independent standard normal random variables, clearly exhibits dependence, but has no autocorrelation structure beyond lag zero. Other examples include the ARCH(1) process [42],

$$X_t = \sqrt{h_t}\varepsilon_t, \quad h_t = c + \theta X_{t-1}^2, \quad (c > 0, 0 \leq \theta < 1)$$

and the GARCH(1,1) process [11],

$$X_t = \sqrt{h_t}\varepsilon_t, \quad h_t = c + \alpha h_{t-1} + \beta X_{t-1}^2,$$
$$(c > 0, \alpha, \beta > 0, \alpha + \beta < 1)$$

which have become popular for modeling financial returns data.

The desire to avoid specific assumptions about the process under the null hypothesis or under the possible alternatives motivates a nonparametric statistical approach to the problem of testing for serial independence. One possibility is to develop rank-based tests for serial independence against particular types of structure such autoregressive moving average (ARMA) structure. Compared to the linear Gaussian paradigm, this approach explicitly drops the assumption that the marginal distribution is normal, which in a natural way leads to tests formulated in terms of ranks. This has the advantage that the tests are distribution free (the null distributions of test statistics do not depend on the actual marginal distribution). The developments around invariant tests for serial independence are reviewed briefly in Sect. "Invariant Tests".

Another nonparametric approach consists of using nonparametric estimators of divergence measures between distributions to construct tests against unspecified alternatives. The idea is to measure the discrepancy between the joint distribution of $(X_{t-m+1}, \ldots, X_t)$ and the product of marginals with a measure of divergence between multivariate probability measures. This typically involves estimating a suitable measure of dependence, and determining the statistical significance of the observed value of the statistic for the sample at hand. In Sect. "Tests Based on Divergence Measures" several tests for serial independence based on divergence measures are reviewed. For further details on some of the earlier methods, the interested reader is referred to the overview by Tjøstheim [103].

Section "Tests Based on Other Measures of Dependence" describes tests for independence based on some other measures of serial dependence in the observed time series, such as partial sums of observations and the bispectrum.

For particular statistics of interest critical values can be obtained in different ways. The traditional way is to use critical values based on asymptotic theory, which is concerned with the large sample limiting distributions of test statistics. With the increasing computer power that became available to most researchers in the recent decades, it has become more and more popular to obtain critical values of test statistics by resampling and other computer simulation techniques. I will discuss the advantages and disadvantages of several of these numerical procedures in Sect. "Bootstrap and Permutation Tests".

Note that pairs of delay vectors such as (X_{t-1}, X_t) and $(X_{s-1}, X_s)'$ for $s \neq t$ may have elements in common, and hence are not independent even under the null; a fact which has to be taken into account when critical values of test statistics are determined. Tests for independence among m random variables, $(Y_1, \ldots, Y_m)$, say, based on a random sample therefore typically need to be adapted for applications in a time series setting. For most asymptotic results that depend on the distribution of the test statistic under the alternative (such as consistency) additional assumptions are required on the rate of decay of the dependence in the data, known as mixing conditions [14].

Notation

Let $\boldsymbol{X}_t^m$ be short-hand notation for the delay vector $(X_{t-m+1}, \ldots, X_t)$, $m \geq 3$. For the case $m = 2$, $\boldsymbol{X}_t^m$ refers to a bivariate vector (X_{t-k}, X_t), $k \geq 1$, where the value of k will be clear from the context. Under the assumption that X_t takes values in the real line $\mathbb{R}$ (or a subset thereof) one may state the null hypothesis in terms of the joint and marginal cumulative distribution functions (CDFs):

$$H_0: \quad F_m(\boldsymbol{x}) = F_1(x_1) \times \cdots \times F_1(x_m)\,,$$

where $\boldsymbol{x} = (x_1, \ldots, x_m)'$, and $F_m(\boldsymbol{x}) = P(X_1 \leq x_1, \ldots, X_m \leq x_m)$ is the joint cumulative distribution function (CDF) of $\boldsymbol{X}_t^m$, and $F_1(x) = P(X \leq x)$ the marginal CDF of $\{X_t\}$. If $\boldsymbol{X}_t^m$ is a continuous random variable, one can denote its probability density function by $f_m(\boldsymbol{x})$, and the independence of the elements of $\boldsymbol{X}_t^m$ can be written as

$$H_0: \quad f_m(\boldsymbol{x}) = f_1(x_1) \times \cdots \times f_1(x_m)\,,$$

where $f_1(x)$ is the marginal probability density function of $\{X_t\}$. For convenience I will drop the subscript m in $f_m(\boldsymbol{x})$, and introduce $g(\boldsymbol{x}) := f_1(x_1) \times \cdots \times f_1(x_m)$, so that the null hypothesis can be rephrased simply as

$$H_0: \quad f(\boldsymbol{x}) = g(\boldsymbol{x})\,.$$

Some Practical Considerations

Which of the tests described below should one choose for practical applications? The alternative against which the null hypothesis is to be tested is any deviation from the above factorizations for some $m \geq 2$. Ideally, one would like a nonparametric test to have large power against all types of dependence. However, since no uniformly most powerful test against all possible alternatives exists, among the tests proposed in the literature one typically finds that some perform better against certain alternatives and some against others, and it is often hard to identify exactly why. Although power against the alternative at hand is obviously important in applications, usually these alternatives are not known in a simple parametric form. This is precisely what motivated many of the recently developed tests tests for independence; they

are designed to have power against large classes of alternatives. When a practitioner has to choose among the large number of the omnibus tests available, besides power also some other properties can be taken into consideration. For instance, some tests are invariant (immune to invertible transformations of the data, see Sect. "Invariant Tests") while others, such as those based on true divergence measures discussed in Sect. "Tests Based on Divergence Measures", are consistent against any fixed alternative, which means that they will asymptotically (with increasing sample size) detect any given alternative with probability one.

Although invariance is a pleasant property, because it allows one to tabulate the null distribution, I would generally rank consistency against any fixed alternative as more important, since invariance can usually be achieved easily by a simple trick, such as transforming the data to ranks before applying any given independence test. At the same time I should add that if one is willing to settle for power against particular classes of alternatives only, it is sometimes possible to construct an ideal hybrid between invariance and consistency in the form of an optimal invariant test. An example is the optimal rank test of Benghabrit and Hallin [7] discussed in the next section.

A clear disadvantage of omnibus tests is that after a rejection of the null hypothesis it leaves the practitioner with the problem of having to identify the type of dependence separately. If one is confident enough to assume (or lucky enough to know) a specific parametric form for the data generating process it is arguably more efficient to rely on traditional parametric methods. However, I think that in most cases the potential efficiency gains are not worth the risk of biased test results due to a misspecification of an unknown type.

Invariant Tests

When developing nonparametric tests for serial independence it is typically assumed that the marginal distribution of the observed time series process is unknown. Because in general the distribution of the test statistic will depend on this unknown distribution, the latter plays the role of an infinite dimensional nuisance parameter. There are various ways of dealing with this problem, such as focusing on particular classes of test statistics and appealing to asymptotic theory, or using bootstrap or permutation techniques. These methods are rather common and are used in most of the tests discussed in the subsequent sections. This section is concerned with a more direct way to deal with the nuisance parameter problem. The main idea is to focus on dependence measures that are invariant under one-to-one transformations of the space in which X_t takes values (so-called static transformations $X'_t = \phi(X_t)$, where ϕ is a strictly monotonous map from $\mathbb{R}$ to itself). This naturally leads to the study of statistics based on ranks.

Rank Tests

Various analogues of the correlation coefficient have been proposed based on ranks. For the pairs $\{(X_t, Y_t)\}$, Spearman's rank correlation [97] is the sample correlation of R_t and S_t, the ranks of X_t and Y_t among the observed X's and the Y's, respectively. In a univariate time series context one can easily define a serial version of this rank correlation (e. g. the sample autocorrelation function of the sequence of ranks $\{R_t\}$ of the X's). Kendall's tau [73] for pairs $\{(X_t, Y_t)\}$ is another rank-based measure of dependence, quantifying the concordance of the signs of $X_i - X_j$ and $Y_i - Y_j$. The serial version of tau can be defined as

$$\tau_k = \binom{n-k}{2} \sum_{i=1}^{n-k} \sum_{j=1}^{i-1} \operatorname{sgn}(X_i - X_j)\operatorname{sgn}(X_{i+k} - X_{j+k}) .$$

The multivariate versions of these concordance orderings have been described by Joe [69]. Genest et al. [45] considered tests for serial independence, building on asymptotic results derived by Ferguson et al. [43] for a serialized version of Kendall's tau in a time series setting.

Many other rank-based tests for independence have been developed meanwhile. The earlier work in this direction is covered in the review paper by Dufour [36]. Later work includes that by Bartels [6], who developed a rank-based version of Von Neumann's statistic, Hallin et al. [54] who proposed rank-based tests for serial independence against ARMA structure, and Hallin and Mélard [55] who study the finite sample behavior and robustness against outliers of their proposed procedures. Kallenberg and Ledwina [72] developed a nonparametric test for the dependence of two variables by testing for dependence of the joint distribution of ranks in a parametric model for the rank dependence.

Optimal rank tests are tests based on ranks that have maximal power. Naturally such a test depends on the alternative against which the power is designed to be large. For instance, Benghabrit and Hallin [7] derived an optimal rank test for serial independence against superdiagonal bilinear dependence.

As a way to deal with the problem that the marginal distribution is unknown under the null hypothesis (the nuisance parameter problem) Genest, Ghoudi and Rémillard [48] consider rank-based versions of the BDS test

statistic (see Sect. "Correlation Integrals"), as well as several other rank-based statistics.

Empirical Copulae

As noted by Genest and Rémillard [46], a rank test for serial independence can alternatively be considered a test based on the empirical copula. The reason is that the empirical copula determines the sequence of ranks and vice versa. I therefore briefly review the notion of a copula.

If $\boldsymbol{X}_t^m$ is a continuous random variable, its copula C is defined as

$$F(x_1, \ldots, x_m) = C(F(x_1), \ldots, F(x_m)) ,$$

where $F(x)$ denotes the marginal CDF. Note that $C(u_1, \ldots, u_m)$ is defined on the unit (hyper-)cube $[0, 1]^m$ (unit square for $m = 2$), and has the properties of a CDF of some distribution on that space. This allows one to define the associated copula density on the unit cube as

$$c(u_1, \ldots, u_m) = \frac{\partial^m}{\partial u_1 \cdots \partial u_m} C(u_1, \ldots, u_m) .$$

The copula density c is obtained by taking partial derivatives with respect to each of the X_i's:

$$f(x_1, \ldots, x_m) = c(F(x_1), \ldots, F(x_m)) \times f(x_1) \times \cdots \times f(x_m).$$

The null hypothesis of serial independence states $f(\boldsymbol{x}) = g(\boldsymbol{x}) = f(x_1) \times \cdots \times f(x_m)$, which is equivalent to $c(u_1, \ldots, u_m) = 1$. This shows that the factorization of the joint distribution in the product of marginals really is a property of the copula. In this sense the copula can be viewed as containing all relevant information regarding the dependence structure of $\boldsymbol{X}_t^m$. Figure 1 shows the Gaussian copula for a bivariate distribution with correlation coefficient 0.5 and the local ARCH(1) copula (essentially a rescaled version of the copula obtained for an infinitesimal positive ARCH parameter).

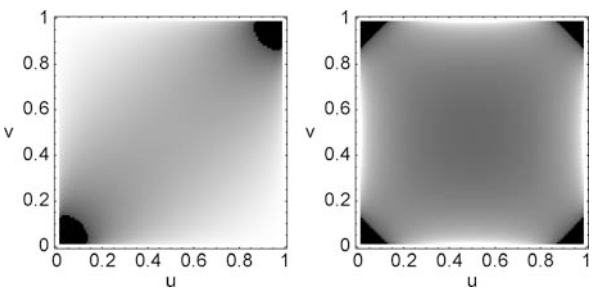

Nonparametric Tests for Independence, Figure 1
The Gaussian copula density for $\rho = 0.5$ (*left*) and local ARCH(1) copula density (*right*)

The empirical copula obtained from time series data is the empirical distribution of $(\widehat{U}_{t-m+1}, \ldots, \widehat{U}_t)$ where $\widehat{U}_t$ is the normalized rank of X_t, defined as $\widehat{U}_t = \#\{X_s \leq X_t\}/n$. Assuming that ties (identical values of X_t and X_s for $t \neq s$) are absent, each rank only occurs once, and hence the empirical copula is one-to-one connected to the sequence of ranks. This shows that test based on ranks can be considered as tests based on the empirical copula and vice versa. It also shows that the concept of an optimal rank test against a particular copula alternative is meaningful.

The connection between sequences of ranks and empirical copulae makes it rather intuitive to design tests that have high power against serial dependence described by particular (families of) copulae. Genest and Verret [47] consider rank-based tests for independence of two random variables that are locally most powerful against a number of parametric copulae. Scaillet [92] used the copula representation to test for serial independence against positive quadrant dependence between two random variables, which holds if $P[X \leq x, Y \leq Y] \geq P[X \leq x]P[Y \leq y]$, or equivalently $P[X > x, Y > y] \geq P[X > x]P[Y > y]$. In a similar vein Panchenko and I [33] derived a rank test for serial independence against the local ARCH(1) copula.

Tests Based on Divergence Measures

In this section I consider tests for serial independence based on various dependence measures. Typically the tests obtained with this approach are not invariant. However, critical values of test statistics can still be obtained using asymptotic theory or bootstrap methods (see Sect. "Bootstrap and Permutation Tests" for more details), and the tests are consistent against a wide range of alternatives.

Many popular dependence measures are based on divergences between the m-dimensional density f and its counterpart under the null hypothesis, g. Divergences are functionals of pairs of densities, which, like distances, are equal to zero whenever $f(\boldsymbol{x}) = g(\boldsymbol{x})$ and strictly positive otherwise. To qualify as a true distance notion between distributions a divergence measure must also be symmetric and satisfy the triangle inequality. Not all divergence measures discussed below are true distances in this sense. This is no problem for testing purposes, but if one is interested in comparing distances with other distances (e. g. for cluster analysis) then the triangle inequality is essential [81]. In general, a divergence measure might serve just as well as a distance as a basis for constructing a test for serial independence.

Tests for serial independence can roughly be divided into two groups: tests against specific types of dependence and omnibus tests, with power against general types of

dependence. For instance, the test of Von Neumann [83] mentioned above is sensitive to linear correlation between X_{t-1} and X_t, but is completely insensitive to some other types of dependence between these two variables. One of the great advantages of tests based on divergence measures is their omnibus nature. Typically these tests are consistent against any type of dependence among the m components of $\boldsymbol{X}_t^m$. Unfortunately, however, as noted in the Introduction, no uniformly most powerful test against serial independence exists, so different tests will be more powerful against different alternatives. Therefore, which test performs best in practice depends on the type of dependence structure present in the data.

Below I describe tests based on empirical distribution functions (empirical CDFs) as well as on densities. One can roughly state that tests based on the empirical CDFs are better suited to detecting large-scale deviations from the null distribution than small-scale deviations. In order for these tests to pick up deviations from independence there must be relatively large regions in $\mathbb{R}^m$ where the density of $\boldsymbol{X}_t^m$ is lower or higher than the hypothetical product density; the cumulative nature of the test statistics is relatively insensitive to small-scale deviations from the null distribution. If one wants to be able to detect subtle small-scale deviations between densities locally in the sample space, it seems more natural to use a test divergence measures based on density ratios or density differences, such as information theoretic divergences or correlation integrals. Note, however, that even among those tests performance may be hard to predict beforehand. For instance, in Subsect. "Information Theoretic Divergence Measures" I consider a family of closely related information theoretical divergence measures, but even within this family the relative powers of the tests depend strongly on the alternative at hand.

Empirical Distribution Functions

Empirical distribution functions have been used for studying the independence of random variables at least since Hoeffding [62], who proposed dependence measures based on the difference between the joint distribution function and the product of marginals, $F(\boldsymbol{x}) - G(\boldsymbol{x})$, where $G(\boldsymbol{x}) = \prod_{i=1}^m F_1(x_i)$ is the joint CDF under the null hypothesis.

There are various ways to define divergences in terms of distribution functions. A popular class of statistics is obtained by considering divergence measures of the type

$$d_w^2 = \int_{\mathbb{R}^m} (F(\boldsymbol{x}) - G(\boldsymbol{x}))^2 \, w(F(\boldsymbol{x}))\, \mathrm{d}F(\boldsymbol{x}) \,,$$

where $w(\cdot)$ is a positive weight function. For $w = 1$ this divergence is known as the Cramér–von Mises criterion, which has become a well-known criterion in univariate one- and two-sample problems. The Cramér–von Mises criterion suggests testing the independence of the elements of $\boldsymbol{X}_t^m$ based on its sample version

$$\tilde{n} d_n^2 = \sum_{i=m}^{n} \left(\widehat{F}(\boldsymbol{X}_i^m) - \widehat{G}(\boldsymbol{X}_i^m) \right)^2 \,,$$

where $\tilde{n} = n - m + 1$ is the number of m-dimensional delay vectors, $\widehat{F}$ is the empirical joint CDF and $\widehat{G}(\boldsymbol{x}) = \prod_{i=1}^m \widehat{F}_1(x_i)$ is the product of marginal empirical CDFs. This statistic, referred to as the Cramér–von Mises statistic, was proposed by Hoeffding [62] for testing independence in the bivariate case, based on a random sample of variables (X_i, Y_i) from a bivariate distribution (i. e. outside the time series scope). Since the statistic is invariant under one-to-one transformations of marginals, the tests based on it are automatically distribution-free. Although the null distribution for a random sample was known to converge in distribution to a mixture of scaled $\chi^2(1)$ random variables,

$$\tilde{n} d_n^2 \xrightarrow{d} \frac{1}{\pi^4} \sum_{j,k=1}^{\infty} \frac{1}{j^2 k^2} Z_{jk}^2 \,,$$

where the Z_{jk} are independent standard normal random variables, it was initially not available in a form suitable to practical applications. The distribution was tabulated eventually by Blum et al. [10], who also considered higher-variate versions of the test.

By generalizing results of Carlstein [23], Skaug and Tjøstheim [95] extended the test of Hoeffding, Blum, Kiefer and Rosenblatt to a time series context and derived the null distribution of the test statistic for first-order dependence ($m = 2$) under continuous as well as discrete marginals. In the continuous case it turns out that the first order test statistic has the same limiting null distribution as for a random sample from a bivariate distribution with independent marginals. Skaug and Tjøstheim [95] also showed that the statistic $nG_{K,n} = n \sum_{k=1}^K d_{k,n}^2$, where $d_{k,n}^2$ is the Cramér–von Mises statistic for (X_{t-k}, X_t), has a mixture of scaled $\chi^2(K)$ distributions as its limiting distribution. For high lags and moderate sample sizes Skaug and Tjøstheim [95] report that the asymptotic approximation to the finite sample null distribution is poor, and suggest the use of bootstrap methods to obtain critical values.

Delgado [29] considered the analogue test for higher-variate dependence in time series. In that case differences with the Hoeffding–Blum–Kiefer–Rosenblatt asymptotics

arise due to the presence of dependence across the delay vectors constructed from the time series. Delgado and Mora [28] investigated the test for first order independence when applied to regression residuals, and found that the test statistic in that case has the same limiting null distribution as for serially independent data.

The Kolmogorov–Smirnov statistic

$$\sqrt{n}\sup_{\boldsymbol{x}}|\widehat{F}(\boldsymbol{x})-\widehat{G}(\boldsymbol{x})|\,.$$

is another popular test statistic for comparing empirical cumulative distribution functions. Ghoudi et al. [49] developed asymptotic theory for this test statistic in rather general settings, which include the time series context. In their power simulations against several alternatives, including AR(1) and nonlinear moving average processes, the Cramér–von Mises statistic displayed a better overall performance than the Kolmogorov–Smirnov statistic.

This suggests that the Cramér–von Mises statistic might be a good choice for practical applications, provided that one wishes to compare empirical distribution functions. As noted above, for detecting subtle density variations it might be more suitable to use a dependence measure based on integrated functions of densities, described in the next subsections.

Integrated Functions of Density Differences

For the bivariate case, Rosenblatt [89] and Rosenblatt and Wahlen [90] considered a class of measures of dependence based on integrated squared differences of densities

$$d(f,g)=\int_{\mathbb{R}^2}w(\boldsymbol{x})(f(\boldsymbol{x})-g(\boldsymbol{x}))^2\mathrm{d}\boldsymbol{x}\,,$$

for some positive weight function $w(\boldsymbol{x})$. The integral can be estimated nonparametrically by plugging in kernel density estimators for the unknown densities, and performing the integration, either numerically or, if possible, analytically. Alternatively one may estimate the integral by taking sample averages of estimated densities. For instance, if $w=1$ one may estimate $\int_{\mathbb{R}^2}f^2(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=\int_{\mathbb{R}^2}f(\boldsymbol{x})\mathrm{d}F(\boldsymbol{x})=E[f(\boldsymbol{X}_t)]$ as $\tilde{n}^{-1}\sum_t\widehat{f}(\boldsymbol{X}_t)$, where $\widehat{f}(\boldsymbol{x})$ represents a consistent kernel density estimate of $f(\boldsymbol{x})$.

Chan and Tran [24] proposed a test for serial independence based on the integrated absolute difference

$$\tilde{d}(f,g)=\int_{\mathbb{R}^2}|f(\boldsymbol{x})-g(\boldsymbol{x})|\mathrm{d}\boldsymbol{x}\,,\qquad p>0\,,$$

for which they developed a histogram estimator [94]. They obtained critical values of the test statistic using a bootstrap method.

Skaug and Tjøstheim [96] explored tests for serial independence based on several dependence measures which are weighted integrals of $f(\boldsymbol{x})-g(\boldsymbol{x})$ in the bivariate case ($m=2$) including the above two measures, which they refer to as I_3 and I_2, respectively. In addition they consider the Kullback–Leibler divergence (I_1 in their notation, discussed in Subsect. "Information Theoretic Divergence Measures") and

$$I_4=\int_{\mathbb{R}^2}(f(\boldsymbol{x})-g(\boldsymbol{x}))f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\,.$$

The latter measure is not a true divergence between f and g, but if f is a bivariate normal pdf, $I_4\geq 0$ with equality if and only if $f=g$. Skaug and Tjøstheim [96] performed a simulation study in which the corresponding estimators $\widehat{I}_i$ were compared, and $\widehat{I}_4$ was found to perform well relative to the other statistics. They subsequently investigated some of the asymptotic properties of this estimator, establishing, among other results, its asymptotic normality. Despite these encouraging simulation results one should beware that there are theoretical cases with dependence where I_4 is zero, meaning that there are also processes with dependence against which the test has little or no power.

Information Theoretic Divergence Measures

By using test statistics based on true divergences, tests can be obtained that are consistent against all deviations of f from the product measure g. Although this does not guarantee high finite sample power for specific alternatives, it obviously is a desirable property if the nature of the alternative is unknown.

Joe [68] described several information theoretic divergence measures, including the Kullback–Leibler divergence between two densities f and g, defined as

$$I(f,g)=\int_{\mathbb{R}^m}f(\boldsymbol{x})\log\left(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}\right)\mathrm{d}\boldsymbol{x}\,.$$

In the case where f is a bivariate density, of X_{t-k} and X_t, say, and g is the product of marginal densities, $I(f,g)$ is also known as the mutual information between X_{t-k} and X_t.

Robinson [88] took the Kullback–Leibler divergence as a starting point for testing the equivalence of f and g. The Kullback–Leibler divergence is invariant under transformations of marginal distributions, and satisfies $I(f,g)\geq 0$ with equality if and only if $f=g$. To see why, consider the random variable $W=g(\boldsymbol{X})/f(\boldsymbol{X})$. By construction $E[W]=1$, and because $\log(W)$ is a concave function of W it follows from Jensen's inequality that $E[-\log W]\leq 0$,

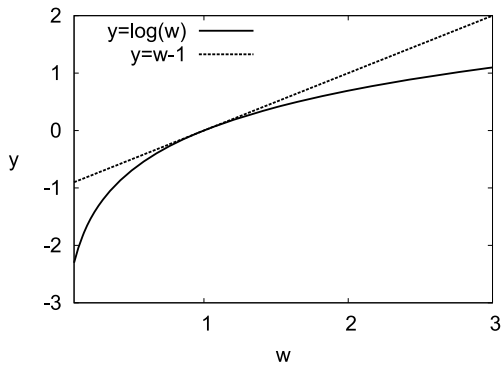

Nonparametric Tests for Independence, Figure 2
Illustration of Jensen's inequality. Since the function $y = \log w$ is concave, it is bounded from above by the tangent line at $w = 1$, given by $y = w - 1$. It follows that if $E[W] = 1$ and $Y = \log W$, $E[Y] = E[\log W] \leq E[W - 1] = E[W] - 1 = 0$ with equality if and only if $W = 1$ with probability 1

with equality if and only if $g(\boldsymbol{X}) = f(\boldsymbol{X})$ with probability one. The reason is that $\log x \leq 1 - x$ for positive x, as illustrated in Fig 2. Application of this inequality to W shows that $E[\log(W)] \leq \log E[W] = 0$ with equality if and only if $W = 1$ with probability 1.

The fact that the Kullback–Leibler divergence is positive for any difference between the true pdf f and the hypothetical pdf g makes it a suitable quantity for testing $f = g$ against unspecified alternatives. A consistent estimator for $I(f, g)$ may serve to construct a test that is consistent (i. e. asymptotically rejects with probability one) against any fixed alternative. Robinson [88] proceeded by constructing such an estimator for $I(f, g)$ using plug-in density estimates of the unknown bivariate densities $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$. For instance, one may use the Nadaraya–Watson density estimator

$$\widehat{f}(\boldsymbol{x}) = \frac{1}{\tilde{n} h^m} \sum_t K((\boldsymbol{x} - \boldsymbol{X}_t)/h) ,$$

where $K(\boldsymbol{x})$ is a probability kernel, such as the pdf of a multivariate normal random variable with independent elements, $(2\pi)^{-m/2} \exp(-\boldsymbol{x}'\boldsymbol{x}/2)$, h is a smoothing parameter, and $\tilde{n}$ the number of delay vectors $\boldsymbol{X}_t$ occurring in the summation. The resulting estimator of the Kullback–Leibler divergence is

$$\widehat{I}(\widehat{f}, \widehat{g}) = \frac{1}{|S|} \sum_{t \in S} \log \left(\frac{\widehat{f}_{X_{t-k}, X_t}(X_{t-k}, X_t)}{\widehat{f}_X(X_{t-k}) \widehat{f}_X(X_t)} \right) , \qquad (1)$$

where S is a subset of $k + 1, \ldots, n$, introduced to allow for "trimming out" some terms of the summation if desired, for instance terms in the summation for which one or more of the local density estimates are negative or zero, as may happen depending on the type of density estimators used. The number of elements of S is denoted by $|S|$. Robinson [88] showed that although the test statistic is a consistent estimator of the Kullback–Leibler divergence, no scaled version of it has a standard normal limit distribution, preventing the development of asymptotic distribution theory in a standard fashion. To overcome this problem, instead of deriving the asymptotic distribution of $\widehat{I}(\widehat{f}, \widehat{g})$, Robinson showed asymptotic normality of a modified test statistic, obtained by attaching weights to each of the terms in the sum in (1).

Hong and White [65] argued that this modification leads to the loss of asymptotic local (i. e. close to the null) power, and developed asymptotic distribution theory for the estimator $\widehat{I}(\widehat{f}, \widehat{g})$ directly. After adjusting for the asymptotic mean they found that an appropriately scaled version of the test statistic actually does have an asymptotically standard normal distribution under the null hypothesis.

Alternatively one may obtain critical values by calculating the test statistic for a large number of simulated replications of an i.i.d. process, as done by Granger and Lin [50]. Note, however, that the critical values thus obtained will depend on the marginal distribution assumed for the process. This approach was followed by Dionísio et al. [35], who used the mutual information between X_{t-k} and X_t for a range of k-values to test for serial independence in stock index returns. Critical values of the test statistic were determined by constructing a reference distribution of the test statistic under the null hypothesis by simulation, repeatedly calculating the value of the test statistic for a large number of independently generated i.i.d. normal time series. The results suggest the presence of residual dependence at several lags for log-returns on stock indices that were filtered to account for ARMA and GARCH structure.

A closely related information theoretic approach has been described by Granger et al. [51] and Racine and Maasoumi [86], who start by considering the class of divergences based on the asymmetric q-class entropy divergence measure defined as

$$I_q(f, g) = \frac{1}{1 - q} \left[1 - \int_{\mathbb{R}^m} \left(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} \right)^q g(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right] .$$

This is a generalization of the Kullback–Leibler divergence, which is recovered in the limit as $q \to 1$. The au-

thors subsequently focused on the symmetric $q = \frac{1}{2}$ case,

$$I_{\frac{1}{2}}(f,g) = 2 - 2\int_{\mathbb{R}^m} \sqrt{g(\boldsymbol{x})}\sqrt{f(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^m} \left(\sqrt{g(\boldsymbol{x})} - \sqrt{f(\boldsymbol{x})}\right)^2 \mathrm{d}\boldsymbol{x}\,,$$

known as the Hellinger distance, and used this to develop tests for various hypotheses involving the equality of two densities, including serial independence.

Fernandes and Neri [44] proposed using an estimator of the Tsallis entropy [105] to test for serial independence in a time series setting. As it turns out, the Tsallis entropy is identical to $I_q(f,g)$. In numerical simulation studies Fernandes and Neri [44] found that, depending on the time series processes under consideration and on the value of q, these tests can have higher power than the entropy-based test of Hong and White [65]. In comparison with the BDS test of Brock et al. [16] (see Sect. "Correlation Integrals") they found that the entropy-based tests perform worse in most cases, although the latter have more power for specific processes, including fractional AR(1) and threshold AR(1) processes.

Aparicio and Escribano [1] developed further tests based on information theoretic dependence measures. Their framework allows testing for short memory against long memory, as well as for the absence of cointegration against linear or nonlinear cointegration. In empirical applications they found that although the rates of the Peseta against the Yen and the US dollar do not appear to be linearly cointegrated, there is evidence supporting a nonlinear cointegrating relation between the two rates.

Characteristic Functions

Csörgő [26] noted that instead of investigating empirical distribution functions for testing independence, as Hoeffding [62] and Blum, Kiefer and Rosenblatt [10] did, a parallel approach can be based on empirical characteristic functions. Several tests for independence have been developed on the basis of this principle. I will be concerned here only with serial independence tests. As with the empirical distribution function one might consider various measures of deviations from independence, e. g. based on a maximum difference or on weighted integrals.

The test of Pinkse [84] is based on the observation that the random variables X_1 and X_2 are independent if and only if their joint characteristic function factorizes. He proposed to test the relation

$$\Psi(u,v) = E\mathrm{e}^{i(uX_1+vX_2)} - E\mathrm{e}^{iuX_1}E\mathrm{e}^{ivX_2} = 0$$

through a quantity of the form $\theta = \iint g(u)g(v)|\Psi(u,v)|^2 \mathrm{d}u\mathrm{d}v$, where $g(\cdot)$ is a positive function. In fact Pinkse introduced an estimator of a related but different functional, as detailed in Sect. "Quadratic Forms" where it is also explained why the test statistic can be estimated directly using U-statistics [93], without the need to actually perform the transformation to characteristic functions.

Hong [63,64] proposed a test for independence based on the difference between the joint characteristic function of X_{t-j} and X_t and the product of their marginal characteristic functions. The main idea is to weigh the discrepancy between F and G across all lags by considering the Fourier transforms

$$h(\omega,\boldsymbol{x}) := (2\pi)^{-1}\sum_{j=-\infty}^{\infty}\gamma_j(\boldsymbol{x})\exp(-ij\omega)$$

of $\gamma_j(\boldsymbol{x}) = F_j(\boldsymbol{x}) - G(\boldsymbol{x})$ where F_j denotes the joint CDF of X_{t-j} and X_t. An application [63] to a series of weekly Deutschmark US dollar exchange rates from 1976 until 1995 showed that although the log returns are serially uncorrelated, there is evidence of nonlinear dependence of the conditional mean return given past returns.

Correlation Integrals

Correlation integrals have been used extensively in the chaos literature, where they were introduced to characterize deterministic dynamics reconstructed from time series. The interested reader is referred to Takens [99] for details of the reconstruction theorem, and to Grassberger et al. [52,53] and the book by Tong [104] for a snapshot of the early developments around correlation integrals. Correlation integrals turn out to be very suitable also in stochastic contexts. They are well adapted to testing for serial independence against unspecified alternatives, as shown below. Moreover, since they are U-statistics asymptotic theory is readily available for them [30,31].

Brock et al. [15,16] based their test for serial independence on the correlation integral of X_t^m, defined as

$$C_m(\varepsilon) = P[|\boldsymbol{Z}_1 - \boldsymbol{Z}_2| \le \varepsilon]$$
$$\text{with } Z_i \sim \boldsymbol{X}_t^m, \text{ independent for } i = 1,2\,,$$

where $|\cdot|$ denotes the supremum norm defined by $|\boldsymbol{x}| = \sup_{i=1,\ldots,m}|x_i|$. Under the null hypothesis of serial independence the correlation integral factorizes:

$$C_m(\varepsilon) = (C_1(\varepsilon))^m\,. \tag{2}$$

This can be seen by expressing $C_m(\varepsilon)$ as a double integral

$$\begin{aligned} C_m(\varepsilon) &= \int_{\mathbb{R}^m}\int_{\mathbb{R}^m} I_{[0,\varepsilon]}(|\boldsymbol{x}-\boldsymbol{y}|)\,\mu_m(\mathrm{d}\boldsymbol{x})\,\mu_m(\mathrm{d}\boldsymbol{y}) \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} I_{[0,\varepsilon]}(|x_1-y_1|)\,\mu_1(\mathrm{d}x_1)\,\mu_1(\mathrm{d}y_1)\times\cdots \\ &\quad\times\int_{\mathbb{R}}\int_{\mathbb{R}} I_{[0,\varepsilon]}(|x_1-y_1|)\,\mu_1(\mathrm{d}x_m)\,\mu_1(\mathrm{d}y_m) \\ &= (C_1(\varepsilon))^m\ . \end{aligned}$$

In the first step the independence of $|X_i - Y_i|$ and $|X_j - Y_j|$ $(1 \le i \ne j \le m)$ was used, for two vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ drawn independently from the distribution of $\boldsymbol{X}_t^m$ under the null hypothesis (in that case all elements $X_1,\ldots,X_m,Y_1,\ldots,Y_m$ are independent and identically distributed). Note that strictly speaking $C_m(\varepsilon)-(C_1(\varepsilon))^m$ is not a divergence. Although it will typically be nonzero for most alternatives with serial dependence, it is possible to construct examples where $C_m(\varepsilon)-(C_1(\varepsilon))^m$ is zero even under serial dependence. Formally this means that testing if $C_m(\varepsilon)-(C_1(\varepsilon))^m$ is zero, amounts to testing an implication of the null hypothesis of serial independence.

For a given kernel function $K(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_k)$ that is symmetric in its arguments, the U-statistic based on a (possibly dependent) sample $\{\boldsymbol{X}_t^m\}_{t=1}^{\tilde{n}}$, consists of the sample average of the kernel function with all elements different:

$$\frac{(\tilde{n}-k)!}{\tilde{n}!}\sum_{\substack{i_1\\ i_j\text{ all different}}}\cdots\sum_{i_k} K(\boldsymbol{X}_{i_1},\ldots,\boldsymbol{X}_{i_k})\ .$$

The corresponding V-statistic is the sample average if the elements are allowed to be identical:

$$\frac{1}{\tilde{n}^k}\sum_{i_1}\cdots\sum_{i_k} K(\boldsymbol{X}_{i_1},\ldots,\boldsymbol{X}_{i_k})\ .$$

The BDS test is based on the scaled difference between the U-statistic estimators of the left- and right-hand sides of (2):

$$W_n = \sqrt{n}\,\frac{C_{m,n}(\varepsilon)-(C_{1,n}(\varepsilon))^m}{\sigma_{m,n}}\ ,$$

where the U-statistic

$$C_{m,n}(\varepsilon) = \frac{2}{(n-m+1)(n-m)} \cdot \sum_{i=2}^{n-m+1}\sum_{j=1}^{i} I_{[0,\varepsilon]}(|\boldsymbol{X}_i^m - \boldsymbol{X}_j^m|)\ , \quad (3)$$

is known as the sample correlation integral at embedding dimension m and $\sigma_{m,n}^2$ is a consistent estimator of the asymptotic variance of the scaled difference. The asymptotic distribution of the test statistic can be derived using the results for U-statistics for weakly dependent processes, described by Denker and Keller [30,31]. Under the null hypothesis of serial independence,

$$W_n \xrightarrow{d} N(0,1)\ .$$

In fact the asymptotic distribution of $C_{m,n}(\varepsilon)-(C_{1,n}(\varepsilon))^m$ is obtained from that of $(C_{m,n}(\varepsilon), C_{1,n}(\varepsilon))$. Since this is a pair of U-statistics, it follows from the results of Denker and Keller [30] that it is asymptotically bivariate normally distributed for strongly mixing stationary processes [14]. After deriving the asymptotic means and covariance matrix one can apply the functional delta method to obtain the asymptotic normal distribution of $C_{m,n}(\varepsilon)-(C_{1,n}(\varepsilon))^m$.

To apply the BDS test the user should specify a value for the bandwidth parameter ε. In numerical studies as well as applied studies, ε-values are typically taken in the range 0.5–1.5 times the sample standard deviation of the observed time series. Note that the null hypothesis tested is independence among all elements of $\boldsymbol{X}_t^m$ rather than pairwise independence of X_{t-m+1} and X_t. Because this results in a relative error of the estimated correlation integral that increases rapidly with m, for applications with moderate sample sizes ($n \approx 1000$, say) small values of m are recommendable (e. g. $m = 2$ or $m = 3$).

Brock et al. [16] derived a 'nuisance parameter theorem' for the BDS test, showing that the limiting distribution of the test statistic is asymptotically free of estimation uncertainty of an unknown parameter θ (e. g. a vector of AR(p) model parameters) provided that a root-n consistent estimator is available for θ. The nuisance parameter theorem, which covers the parameters of AR models, but not, for instance of ARCH models, states that the asymptotic distribution of the test statistic for residuals is the same as that for the true innovations. This justifies the use of residuals in place of true innovations asymptotically, which is convenient since it allows using the BDS test on residuals as a model specification test, provided that the estimated parameters are root-n consistent.

De Lima [78] formulated five conditions under which the BDS test is asymptotically nuisance parameter free (i. e. can be used as a model specification test). These involve, among others, mixing conditions and conditions ensuring the consistency of parameter estimates. Interestingly, the test is not asymptotically nuisance parameter free for GARCH residuals, but it is when applied to logarithms of the squared residuals. Caporale et al. [21] have performed a simulation study to evaluate the behavior of the test

statistic under violations of these conditions, and found the BDS test to be very robust.

Note that filtering time series data by replacing them with the (standardized) residuals of a time series model typically has the effect of whitening the data, which makes the detection of dependence more difficult. Brooks and Heravi [18] found that upon filtering data through a completely misspecified GARCH model, the frequency of rejection of the i.i.d. null hypothesis can fall dramatically. Therefore, a failure to reject the null hypothesis on the basis of GARCH residuals does not imply that a GARCH model is consistent with the data.

Wolff [106] observed that the unnormalized correlation integral, i. e. the double sum in (3) without the normalizing factor, converges to a Poisson law under some moderate assumptions regarding the marginal distribution. This motivates a nonparametric test procedure based on the correlation integral, which Wolff found to have reduced size distortion compared to the usual BDS test.

Instead of the sample correlation integral, Kočenda and Briatka [74] suggest using an estimator of the slope

$$D_m(\varepsilon) = \frac{\mathrm{d}\ln C_m(\varepsilon)}{\mathrm{d}\ln \varepsilon} ,$$

also known as the course-grained correlation dimension at embedding dimension m and distance ε, for testing the null hypothesis of serial independence. The intuition is that the theoretically $C_m(\varepsilon) \sim \varepsilon^m$ for small ε under the i.i.d. null, while $C_m(\varepsilon) \sim \varepsilon^\alpha$ for some $\alpha < m$, provided m is sufficiently large, in the case of a low-dimensional attractor with correlation dimension α. The coarse-grained correlation dimension is a measure for complexity, and deviations from the null other than chaos typically also reduce the coarse-grained correlation dimension. This makes the coarse-grained correlation dimension a promising quantity for testing the i.i.d. null hypothesis. Rather than using the slope for a single bandwidth ε, Kočenda and Briatka [74] proposed to use an estimator of the average slope across a range of ε-values, consisting of the least squares estimator of the slope parameter β_m in the regression

$$\ln(C_{m,n}(\varepsilon_i)) = \alpha_m + \beta_m \ln(\varepsilon_i) + u_i, \quad i = 1, \ldots, b ,$$

where α_m is an intercept, u_i represents an error term, and b is the number of bandwidths ε_i taken into consideration. They then determined the optimal range of ε-values by simulation. The test based on the resulting least squares estimator $\widehat{\beta}_m$ for β_m was found to have high power compared to some other tests for serial independence, and to behave well when used as a specification test.

Although it is clear that the correlation integrals from more than one bandwidth value ε_i contain more information than that from a single bandwidth, it is not clear why it would be a good idea to base a test on the estimator $\widehat{\beta}_m$. Since the correlation integral is an empirical CDF (of inter-point distances) the error terms u_i will be correlated, which typically leads to a loss of efficiency. In Subsect. "Multiple Bandwidth Permutation Tests" I discuss an alternative way to combine information from different bandwidths into a single test statistic, inspired by the rate-optimal adaptive tests of Horowitz and Spokoiny [67].

Johnson and McLelland [70,71] proposed a variation on the BDS test for testing the independence of a variable and a vector based on correlation integrals. The main idea is to test for remaining dependence between residuals and regressors, in addition to mere dependence among residuals. This might be an advisable approach in many cases, because even though theoretically a model misspecification should lead to residuals with serial dependence, it is often very hard to detect this dependence with tests on the residuals only, due to the whitening effect of the filtering.

Quadratic Forms

Quadratic forms are convenient for defining squared distances between probability distributions, which provide tests that are consistent against any type of dependence (hence including, for instance, ARCH and GARCH structure). A comparative advantage relative to the information theoretical divergences discussed in Subsect. "Information Theoretic Divergence Measures" is that they can, like correlation integrals, be estimated straightforwardly by U- and V-statistics.

The starting point for the construction of a quadratic form is a bilinear form, which may be interpreted as an inner product on the space of measures on $\mathbf{R}^m$. The quadratic forms discussed here were first applied in the context of testing for symmetries of multivariate distributions [34], and later extended to a time series context [32].

Consider, for a kernel function $K(\cdot,\cdot)$ on $\mathbb{R}^m \times \mathbb{R}^m$ the form

$$(\mu, \nu) = \int_{\mathbb{R}^m} \int_{\mathbb{R}^m} K(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\mu(\boldsymbol{x}) \mathrm{d}\nu(\boldsymbol{y})$$

for measures μ and ν. Note that this form is bilinear (linear in μ as well as ν). If this form happens to satisfy $(\mu, \mu) \geq 0$ for any (possibly signed) measure μ with $(\mu, \mu) = 0$ if and only if $\mu(A) = 0$ for all Borel subsets A of $\mathbb{R}^m$, then K is called positive definite. In the terminology introduced above, this means that $(\mu - \nu, \mu - \nu)$ is a divergence between the measures μ and ν. Note that a positive definite

form defines an inner product on the space of measures on $\mathbb{R}^m$ with the usual properties:

(i) $(\mu, \nu) = (\nu, \mu)$.
(ii) $(a\mu + b\nu, \eta) = a(\mu, \eta) + b(\nu, \eta)$ for scalars a, b.
(iii) $(\mu, \mu) \geq 0$ with equality iff $\mu(A) = 0$ for any Borel subset $A \in \mathcal{A}$.

The inner product can therefore be used to define a norm of $\mu - \nu$ as $\|\mu - \nu\| = \sqrt{(\mu - \nu, \mu - \nu)}$, which satisfies all the usual properties of a distance, such as Schwarz's inequality, the triangle inequality and the parallelogram law (See e. g. Debnath and Mikusiński [27]).

In short: any positive definite kernel K defines an inner product on the space of measures on $\mathbb{R}^m$, which in turn defines a squared distance between μ and ν, given by

$$\theta = \|\mu - \nu\|^2 = (\mu - \nu, \mu - \nu) .$$

(For simplicity the dependence of the squared distance on the kernel function K has been suppressed in the notation.)

To pinpoint some classes of kernel functions that are suitable for our purposes (i. e. that are positive definite) let us assume that the kernel function K depends on $\boldsymbol{x}$ and $\boldsymbol{y}$ only through the difference $\boldsymbol{x} - \boldsymbol{y}$, and that the kernel function factorizes, i. e.

$$K(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{m} \kappa(x_i - y_i) .$$

In that case the Fourier transform $\tilde{K}$ of the kernel function also factorizes, into $\tilde{K}(\boldsymbol{u}) = \prod_{i=1}^{m} \tilde{\kappa}(u_i)$, where $\tilde{\kappa}(u) = \int \kappa(t) \mathrm{e}^{-iut} \mathrm{d}t$, the Fourier transform of κ. The squarel distance $\theta = \|\mu - \nu\|^2$ can then be expressed directly in terms of characteristic functions $\tilde{\mu}$ and $\tilde{\nu}$ of μ and ν respectively:

$$\theta = \frac{1}{2\pi} \int \tilde{K}(\boldsymbol{u}) |\ \tilde{\mu}(\boldsymbol{u}) - \tilde{\nu}(\boldsymbol{u})|^2 \mathrm{d}\boldsymbol{u} .$$

This follows from applying Parseval's theorem to $\theta = \iint K(\boldsymbol{x}, \boldsymbol{y})(\mu - \nu)(\mathrm{d}\boldsymbol{x})(\mu - \nu)(\mathrm{d}\boldsymbol{y})$. It follows that if the kernel function is bounded and has a Fourier transform which does not vanish on any interval, its associated bilinear form is positive definite.

To illustrate this, Fig. 3 shows three kernel functions and their Fourier transforms. The Gaussian kernel (top panels) has a Gaussian as its Fourier transform, which is everywhere positive. Therefore, the Gaussian product kernel is positive definite and defines a quadratic form suitable for detecting any differences between a pair of distributions on $\mathbb{R}^m$. A similar conclusion holds for the double exponential kernel $\exp(-|x|/a)$ (middle panels). The 'naive' kernel function $I_{[-a,a]}(x)$ (bottom panels) has a Fourier transform which is negative for certain ranges of the frequency ω, and hence is not a positive definite kernel function.

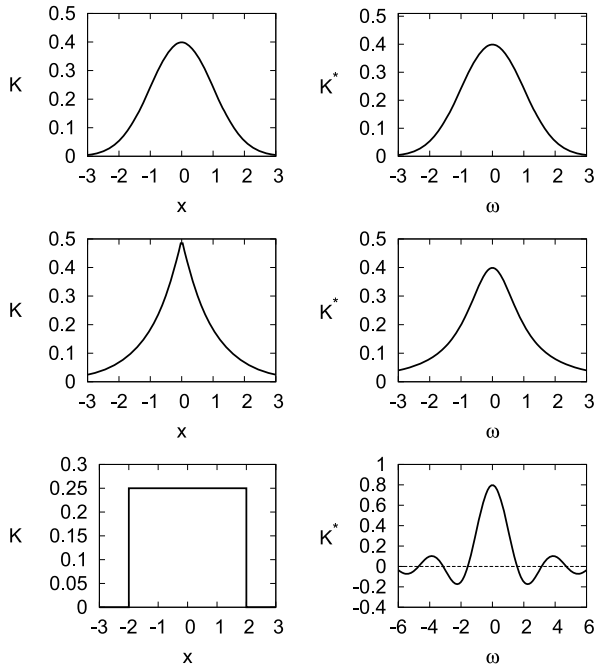

Nonparametric Tests for Independence, Figure 3
Kernel functions (*left*) and their Fourier transforms (*right*) for the Gaussian kernel (*top*), double exponential kernel (*middle*) and the naive kernel (*bottom*)

Given the kernel function K (e. g. a multivariate Gaussian product kernel) the estimation of the associated quadratic form $(\mu - \nu, \mu - \nu) = (\mu, \mu) - 2(\mu, \nu) + (\nu, \nu)$ is straightforward. Empirical versions of (μ, μ), (μ, ν) and (ν, ν) can be obtained easily as sample averages. For instance, if $\mathbf{X}_t^m$ is a sample from μ, the sample version of $(\mu, \mu) = \iint K(\boldsymbol{s}_1, \boldsymbol{s}_2) \mathrm{d}\mu(\boldsymbol{s}_1) \mathrm{d}\nu(\boldsymbol{s}_2)$ is the V-statistic

$$\widehat{(\mu, \mu)} = \frac{1}{\tilde{n}^2} \sum_i \sum_j K(\mathbf{X}_i^m, \mathbf{X}_j^m) .$$

As before, $\tilde{n} = n - m + 1$ denotes the number of m-vectors available. It follows from the results of Denker and Keller [30,31] for U- and V-statistics of dependent processes that the estimator is consistent under strong mixing conditions and asymptotically normally distributed with a variance that can be estimated consistently from the data. Note that the estimator of (μ, μ) is in fact a sample correlation integral, but with the kernel K instead of the usual naive kernel.

As shown in [32], similar consistent estimators for the other terms can be constructed easily:

$$\widehat{(\mu,\nu)} = \frac{1}{\tilde{n}}\sum_{t=1}^{\tilde{n}}\prod_{k=0}^{m-1}\widehat{C}(X_{t+k}) ,$$
$$\widehat{(\nu,\nu)} = \frac{1}{\tilde{n}^m}\prod_{k=0}^{m-1}\left(\sum_{t=1}^{n}\widehat{C}(X_{t+k})\right) ,$$

where $\widehat{C}(x) = \frac{1}{n}\sum_{i=1}^{n}\kappa(x - X_i)$ is the one-dimensional correlation integral associated with the marginal distribution. For some results concerning size and power and comparisons of those with the BDS test and the test of Granger, Maasoumi and Racine [51] see section "Multiple Bandwidth Permutation Tests".

In fact the divergence measure $\theta = \iint g(u)g(v)|\Psi(u,v)|^2 \mathrm{d}u\mathrm{d}v$, on which Pinkse [84] based his test for serial independence (see Sect. "Characteristic Functions") is also a quadratic form (for bivariate random variables). Instead of using a U-statistics estimator of θ, Pinkse used an estimator of a related quantity ϑ, which in terms of the associated inner product can be expressed as:

$$\vartheta = \left\{[(\mu,\mu)-(\mu,\nu)]^2 + [(\nu,\nu)-(\mu,\nu)]^2\right\}/2 .$$

It can be verified that also $\vartheta \geq 0$ with equality if and only if $\vartheta = 0$. Indeed, evidently $\vartheta = 0$ if $\mu = \nu$, while under any alternative one cannot have both $(\mu,\nu) = (\mu,\mu)$ and $(\mu,\nu) = (\nu,\nu)$, since in that case $(\mu-\nu,\mu-\nu) = (\mu,\mu) - 2(\mu,\nu) + (\nu,\nu) > 0$.

Tests Based on Other Measures of Dependence

Partial Sums of Data

Ashley and Patterson [2] proposed a test for independence in stock returns based on the cumulative sum $Z_t = \sum_{j=1}^{t} X_j$ where X_j represents the residuals obtained after estimating an AR(p) model on returns. The idea is that if the model is appropriate, the residuals are expected to be close to i.i.d. and Z_t corresponds to the deviation of a Brownian motion on the line after t time steps. The authors proposed to test this property using the statistic $Z^{\max} = \max\{|Z_1|,\ldots,|Z_T|\}$, assessing the statistical significance using a bootstrap method.

It was later pointed out by Corrado and Schatzberg [25] that since $\{X_t\}$ has a zero sample mean, $\{Z_t\}$ is 'tied to zero' at the endpoints ($Z_t = Z_0 = 0$), and hence the reference paths used in the bootstrap should have been constructed to satisfy the same constraints. This can, for instance, be achieved by mean adjusting the bootstrap sample, or alternatively by employing a permutation method (resampling without replacement). Moreover, Corrado and Schatzberg [25] showed that after rescaling via

$$W_t = Z_t/(\sqrt{T}\widehat{\sigma}_T)$$

where $\widehat{\sigma}_T$ is the sample standard deviation of X_j, the sample path of W_t for large T forms a Brownian bridge under the null hypothesis, which implies that the maximum absolute value has the same null distribution as the Kolmogorov–Smirnov (KS) test statistic.

Scaled versions of partial sums were also considered by Kulperger and Lockhart [75]. They focus on examining the conditional mean of $Y_j := X_{j+1}$ given X_j, by studying the dependence among successive Y-values when ordered according to the ranks of the corresponding X-values. Put simply, this replaces the ordering of pairs (X_t, Y_t) in such a way that X-values are ordered increasingly, enabling the partial sums to grasp the common dependence between Y-values on X-values, rather than on time. Motivated by this, the authors propose to study the sample path of the partial sums

$$S_i = \frac{1}{\sqrt{n}}\sum_{j=1}^{i}(Y_{(j)} - \bar{Y}) ,$$

where $Y_{(j)}$ denotes $X_{(j)+1}$ (the successor of the observation among whose rank among the original observations is j), and $\bar{Y}$ is the sample mean of $\{Y_j\}$. The authors then propose and compare various statistics to test if the realized process $\{S_i\}$ is a realization of a Brownian bridge, as predicted under the null hypothesis. Straightforward extensions can be obtained by taking $Y_{(j)} = \Phi(X_{(j)+k})$ for some fixed lag k.

The Spectral Density

Besides being able to test the strict random walk hypothesis (i.i.d. increments) for a financial time series such as a log-price, it is also of interest to be able to test the weaker hypothesis that increments have a constant conditional mean. A test based on the spectral density for this so-called martingale hypothesis was developed by Durlauf [41].

The Bispectrum

As already noted by Robinson [88], one can test for serial independence against nonlinear dependence with a test for linearity rather than independence. Here and in the next subsection I briefly discuss a few examples of linearity tests.

Extending results of Subba-Rao [98], Hinich [58] used the bispectrum to detect 'interactions between Fourier

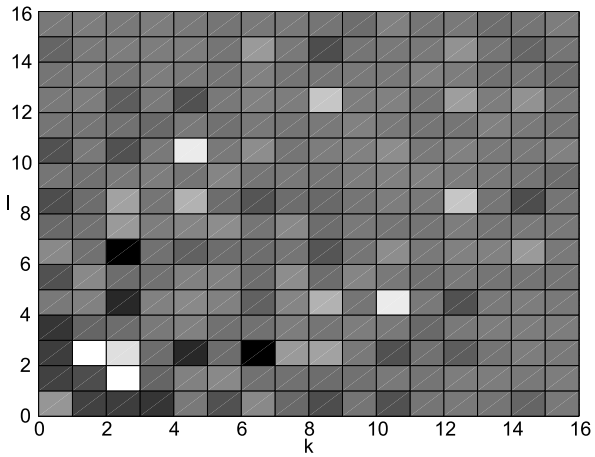

Nonparametric Tests for Independence, Figure 4
Bicovariance $c(k, \ell)$ for a bilinear process. *Lighter regions* correspond with larger values of the bicovariance. Series length $n = 4000$

components'. The motivation behind the approach is that the bicorrelation, defined as

$$c(k, \ell) = E[X_t X_{t+k} X_{t+\ell}] ,$$

should be zero for a stationary linear Gaussian random process $\{X_t\}$ with mean zero.

As an illustration of the structure that the bicovariance of a nonlinear time series may exhibit, Fig. 4 shows the bicovariance for the time series $\{X_t\}$ generated by the bilinear process

$$X_t = 0.9\varepsilon_{t-1}X_{t-2} + \varepsilon_t$$

where $\{\varepsilon_t\}$ is a sequence of independent standard normal random variables. For the simulated data I initialized the state variables at $X_{-1} = X_0 = 0$ and discarded the first 1000 iterations.

Examining the behavior of the bicorrelation $c(k, \ell)$ for many values of k and ℓ simultaneously can be achieved in various ways, for instance by examining the bispectrum

$$B(\omega_1, \omega_2) = \sum_{k=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} c(k, \ell) \exp\left[-i(\omega_1 k + \omega_2 \ell)\right] .$$

Hinich [58] introduced two functionals of the bispectrum that are suitable for testing Gaussianity and linearity, respectively. For applications to both model generated data and real data see, among others, [3,4,17,57]. Overall, these applications indicate that nonlinearity and non-Gaussianity play an important role in economic time series.

More recently Hinich [59] proposed a related test for serial independence of the innovations acting on a (unspecified) linear filter. This test is also based on the bispectrum, but the test statistic can be evaluated in the time domain since it is a function of the sample bicovariance function of the residuals. Lim et al. [77] applied the test of Hinich [59] to returns of Asian market indices. If the returns would follow a GARCH process with symmetric innovations the sequence of signs thus obtained should be i.i.d. Since the results indicate nonlinear structure in the signs of Asian stock returns, the conclusion is that GARCH models with symmetric innovations are inappropriate for the returns.

Note that although this example shows how the bispectrum can be used to detect evidence against the null of (G)ARCH, the bispectrum cannot be used to test for serial independence against (G)ARCH alternatives. The reason is that the bicovariance is not able to pick up dependence in time series processes in which the conditional mean of X_t given past observations is zero, such as ARCH and GARCH processes.

If the probabilistic structure of a stationary time series process is preserved under time-reversal (i. e. when reading the series backward in time), the time series process is called time reversible. Clearly, time reversibility should hold under serial independence. Ramsey and Rothman [87,91] developed a test for time reversibility based on a sample version of the difference of bicovariances $c(k, k) - c(0, k) = E[X_t^2 X_{t-k}] - E[X_t X_{t-k}^2]$. This test is consistent against some forms of serial dependence, but not against all. For instance, it is not consistent against ARCH and GARCH alternatives, since these have zero bicovariance at any lag k.

Terdik and Máth [100] use the information contained in the bispectrum to test whether the best predictor is linear against quadratic alternatives. The null hypothesis being tested is linearity in mean, as opposed to a linear Gaussian random processes.

Brooks and Hinich [19,20] generalized the bispectrum approach to multivariate time series settings, in order to study nonlinear lead-lag relationships. These authors indeed found evidence for nonlinear dependence structure between various exchange rates. As noted by the authors these findings have important implications for investors who try to diversify their portfolios internationally.

Nonlinearity in the Conditional Mean

Hjellvik and Tjøstheim [60] developed a nonparametric test for linearity based on the difference between the best linear predictor and a kernel-based nonparametric predictor. Hjellvik et al. [61] explored a variant of this approach where local polynomial predictors were used instead of kernel-based predictors.

Bootstrap and Permutation Tests

As shown above, in many cases it is possible to use asymptotic distribution theory for test statistics in the time series context. Notable cases are those where the test statistics are U-statistics or a function thereof, as is the case for the BDS test. In some cases, however, the resulting asymptotic approximation to the finite sample null distribution may be poor. In particular this can happen when the test statistic is a degenerate or near-degenerate U-statistic under the null hypothesis. For instance, the BDS test statistic is near-degenerate under the null hypothesis if the marginal distribution is uniform.[1]

For practical purposes, near-degeneracy means that although the test statistic is asymptotically standard normal, it may be far from asymptotic normality even for large sample sizes. Whether a particular test statistic is (near) degenerate under the null is often not known, as it depends on the marginal distribution of the data generating process. To avoid such problems with the asymptotic approximation one can use critical values obtained by simulating a fully specified model that satisfies the i.i.d. null hypothesis. However, since the distribution of the test statistic depends on the marginal distribution, it is better to reflect this in the simulated data as well. This can be done by using bootstrap or Monte Carlo methods for assessing the statistical significance of the observed value of the test statistic of interest. The Monte Carlo procedure has the additional advantage that it produces an exact randomization test.

Simulation

Although I do not recommend this procedure in practice, I include it for completeness. For simplicity I describe the approach here only for the one-sided case where large values of the test statistic provide evidence against the null hypothesis. The approach for two-sided tests is similar.

Suppose we wish to obtain critical values of a test statistic, Q_n, say, then this can be obtained by simulating a large number of i.i.d. time series of length n. The idea is to simulate time series satisfying the null hypothesis using a fully specified model. For instance one can simulate a large number B of i.i.d. normal time series data of the same length as the original time series, and then calculate the value of the test statistic for each of these artificial time series. The sample distribution of the B simulated test statistics subsequently represents the null distribution, and a p-value can be obtained as $\widehat{p} = \#\{Q_n^i \geq Q_n\}/(B+1)$. This approach is suitable if one is willing to assume a certain marginal (normality in this case) under the null, or if the distribution of the test statistic is (at least asymptotically) independent of the (unknown) marginal distribution.

Dionísio [35] implemented a test for serial independence based on the mutual information between X_t and $X_{t+\tau}$. Critical values of the test statistic were obtained for individual lags τ by simulating a large number of $N(0,1)$ i.i.d. time series of the appropriate length. This should provide a good approximation to the true critical values if the data are approximately normally distributed, or if the the (asymptotic) distribution of the test statistic is independent of the (unknown) marginal distribution. If not, this may lead to size distortions if the data are skewed or otherwise deviating from normality. A bootstrap or a permutation test may be more appropriate in those cases.

Bootstrap Tests

The bootstrap approach consists of resampling from the observed data with replacement. The idea is that under the null the best representation of the data generating process that we have is given by an i.i.d. process with the empirical distribution of the observed data. One of the motivating advantages of the bootstrap is that it yields an improved approximation to the finite sample distribution of test statistics relative to first-order asymptotic theory, provided that the statistics are asymptotically pivotal. For an overview of the use of the bootstrap in econometrics, I refer the interested reader to [66].

Although there are sophisticated bootstrap methods that are particularly designed for time series, for instance the block bootstrap [22,76,85], in the case of testing for serial independence the data are i.i.d. under the null, so under the null hypothesis we can bootstrap by simply drawing n values from the original time series independently with replacement. This procedure is often referred to as the naive bootstrap.

Hong and White [65] noted that the naive bootstrap does not produce a consistent procedure for their test statistic (essentially the estimator of the Kullback–Leibler divergence given in (1)) as it is degenerate under the null hypothesis of serial independence. They propose the use of a smoothed bootstrap procedure to overcome this. In the degenerate case also a permutation test may be used, as described next.

[1] It is exactly degenerate for i.i.d. data from the uniform distribution on the circle, i. e. the interval $[0, a]$ with the endpoints identified [16]. Theiler [102] simulated variance of $S = C_{m,n}(\varepsilon) - (C_{i,n}(\varepsilon))^m$ in the degenerate case, and found that it converges to 0 at the rate n^{-2} instead of the usual rate n^{-1}.

Permutation Tests

Under the null hypothesis of serial independence the data generating process is typically known only up to an infinite dimensional nuisance parameter (the unknown marginal distribution). This prevents one from generating time series data that have the exact same distribution as the data generating process under the null hypothesis, as Barnard [5] suggested for simple null hypotheses (i. e. null hypotheses under which the distribution of the data is fully specified). Hence the problem is that the null hypothesis of serial independence is not simple but composite, with each possible marginal distribution representing another i.i.d. process. This limitation can be overcome by considering all the possible processes under the null, conditional on an observed value of a minimal sufficient statistic for the nuisance parameter. The resulting (conditional) null distribution of the test statistic can then be shown to be free of unknown parameters. The simulated data should be drawn from the same conditional distribution as the data generating process under the null, given the sufficient statistic. This procedure can be used for constructing a randomization test procedure which is exact, i. e. the type I error rate is equal to nominal level, at least in the absence of parameter estimation uncertainty. The resulting tests are referred to as Monte Carlo tests.

Since under the null hypothesis the empirical marginal distribution is a minimal sufficient statistic for the unknown marginal distribution. The conditional distribution of the observations given their empirical marginal, assigns equal probability to each of the $n!$ possible permutations of the data. This means that every permutation of the originally observed values is equally likely. Hence an independent draw from the time series process conditional on the sufficient statistic can be obtained straightforwardly by randomly permuting the original data. The value of the test statistic for such a permuted time series is an independent draw from the conditional null distribution of the test statistic given the sufficient statistic. Although the Monte Carlo method is exact for data generated under the null hypothesis, not many investigators have studied its behavior when applied to residuals of an estimated time series model. For a general treatment of Monte Carlo testing in the presence of model parameter uncertainty, see the recent work by Dufour [37].

Multiple Bandwidth Permutation Tests

Most of the nonparametric tests described above, such as the BDS test, require a user-specified value for the bandwidth parameter. I mentioned that the BDS test is usually applied with bandwidth values in the range 0.5 to 1.5 standard deviations. Although these values appear to work reasonably well in numerical simulation studies with computer-generated data from known processes, there is no guarantee that this is an optimal range for the (usually unknown) alternative of most interest to the user (i. e. the possibly non-i.i.d. true process that generated the data at hand). In the different context of testing a parametric regression function against an unknown nonparametric alternative, Horowitz and Spokoiny [66] proposed tests with an adaptive bandwidth, that they showed to be rate-optimal. Although the present context is slightly different, and the details of their theorems most likely require some adaptation before they apply here, test statistics similar to the adaptive bandwidth statistics that they proposed can be easily implemented.

The idea is to calculate test statistics for many values of the bandwidth parameter, ε, say, and reject the null hypothesis if there is evidence against independence from one or more of the statistics calculated for the various bandwidths. To achieve this, an overall test statistic is required that will pick up evidence of dependence from any of the bandwidths. In Ref. [32] we proposed using the smallest p-value, $\widehat{p}(\varepsilon_i)$, across a set of d different bandwidths $\varepsilon_1 < \cdots < \varepsilon_d$ as an overall test statistic: $T = \inf_i \widehat{p}(\varepsilon_i)$. To establish if the value of T obtained is significant, a permutation test can be performed. I refer to this procedure as the multiple bandwidth permutation test.

Suppose that we wish to base the p-values $\widehat{p}(\varepsilon_i)$ on the permutation procedure described above, then this setup seems to require two nested permutation procedures; one global loop for replicating B values of T, T_i, $i = 1, \ldots, B$, for the B different permutations of the original data, and for each of those another loop to obtain a p-values $\widehat{p}(\varepsilon_i)$ of the observed (BDS) test statistic for each bandwidth. It turns out, however, that this can be achieved much more efficiently, in a single loop across B permutations of the original data, as follows.

Let $Q^1(\varepsilon_i)$, $i = 1, \ldots, b$, denote the value of the (BDS) test statistic for the original data at the ith bandwidth, ε_i, and $Q^k(\varepsilon_i)$, $k = 2, \ldots, B$, that of the kth randomly permuted time series, then a p-value can be obtained for each bandwidth as before: $\widehat{p}^1(\varepsilon_i) = \#\{Q^s(\varepsilon_i) \geq Q^1(\varepsilon_i)\}/B$. The superscript 1 denotes that these p-values are obtained for the original time series, $b = 1$. Subsequently one can obtain similar p-values for each of the permuted time series as $\widehat{p}^b(\varepsilon_i) = \#\{Q^s(\varepsilon_i) \geq Q^b(\varepsilon_i)\}/B$. Now we are in a position to calculate the global test statistic $T_b = \inf_i \widehat{p}^b(\varepsilon_i)$ for each of the B permuted time series, including the orig-

Nonparametric Tests for Independence, Table 1

Observed rejection rates at nominal size 0.05 of the test of Granger, Maasoumi and Racine [51] (GMR) test and the multiple bandwidth permutation procedure for the BDS test statistic [16] and the quadratic form-based test of Diks and Panchenko [32] (DP). In the model specifications $\{u_t\}$ represents a sequence of independent standard normal random variables. Embedding dimension $m = 3$, sample size 100, except for the sign process (sample size 50) and the logistic and the Hénon maps (sample size 20). Monte Carlo parameters $B = 100$ permutations and 1000 independently realized time series from each process

Model	Specification	GMR	BDS	DP
1	$X_t = u_t$	0.05	0.05	0.05
2	$X_t = u_t + 0.8u_{t-1}^2$	0.57	0.68	0.71
3	$X_t = u_t + 0.6u_{t-1}^2 + 0.6u_{t-2}^2$	0.78	0.84	0.96
4	$X_t = u_t + 0.8u_{t-1}u_{t-2}$	0.22	0.46	0.29
5	$X_t = 0.3X_{t-1} + u_t$	0.31	0.16	0.68
6	$X_t = 0.8\lvert X_{t-1}\rvert^{0.5} + u_t$	0.25	0.11	0.53
7	$X_t = \mathrm{sign}(X_{t-1}) + u_t$	0.86	0.75	0.98
8	$X_t = 0.6\varepsilon_{t-1}X_{t-2} + \varepsilon_t$	0.26	0.50	0.39
9	$X_t = \sqrt{h_t}u_t,\ h_t = 1 + 0.4X_{t-1}^2$	0.26	0.51	0.24
10	$X_t = \sqrt{h_t}u_t,\ h_t = 0.01 + 0.80h_{t-1} + 0.15X_{t-1}^2$	0.15	0.35	0.18
11	$Y_t = (-0.5 + 0.9I_{[0,\infty)}(Y_{t-1}))Y_{t-1} + \varepsilon_t$	0.34	0.06	0.87
12	$X_t = 4X_{t-1}(1 - X_{t-1}),\quad (0 < X_t < 1)$	0.95	0.71	0.90
13	$X_t = 1 + 0.3X_{t-2} - 1.4X_{t-1}^2$	0.96	0.46	0.97
14	$X_t = Z_t + \sigma u_t,\quad Z_t = 1 + 0.3Z_{t-2} - 1.4Z_{t-1}^2$	0.41	0.22	0.83

inal time series (the case $b = 1$). Finally, we can establish the significance of the test statistic T_1 obtained for the original time series by comparing it with the reference values $T_2, \ldots, T_B$. Although these values need not be independent, even under the null hypothesis, they do satisfy permutation symmetry under the null hypothesis, so that each of the possible permutations of the observed values of T_b is equally likely. By the permutation symmetry of all the time series (the original and the permuted series) under the null hypothesis, and hence of the values T_b, $b = 1, \ldots, B$, the overall p-value can still be calculated as if the values T_b were independent, i. e. $\widehat{p} = \#\{T_s \le T_1\}/B$. In other words, only the fact that all possible orderings of the values $T_1, \ldots, T_B$ are equally likely under the null hypothesis is needed, and not their independence.

So far I haven't discussed the possibility that ties may occur. In fact they occur with nonzero probability since $\widehat{p}^b(\varepsilon_i)$ is a discrete random variable for finite B. If ties are dealt with appropriately, however, then the above procedure leads to a test with a rejection rate under the null hypothesis equal to the nominal size (for details see [32]).

Table 1 shows the power of the multiple bandwidth procedure for the BDS test [16] and the test developed by Valentyn Panchenko and me [32] based on quadratic forms for various processes (referred to henceforth as the DP test). For comparison the power of the test of Granger Maasoumi and Racine [51] (GMR test) based on the Hellinger distance, discussed in Subsect. "Information Theoretic Divergence Measures", are also provided. The GMR test was performed with the R software provided by the authors, which uses a bandwidth based on cross-validation.

The processes are, in order, models of type: i.i.d. normal (1), nonlinear moving average (2–4), linear autoregressive (5), nonlinear autoregressive (6), sign autoregressive (7), bilinear (8), ARCH(1) (9), GARCH(1,1) (10), threshold autoregressive (11), logistic map (12), Hénon map (13) and the Hénon map with dynamic noise (14). The multiple bandwidth permutation test was performed with 5 bandwidth values ε_i between $\varepsilon = 0.5$ and 2.0, with a constant ratio $\varepsilon_{i+1}/\varepsilon_i$, $i = 2, \ldots, 4$ (hence the bandwidths are equally spaced on a logarithmic scale).

The table shows rejection rates for the i.i.d. process which are all close to the nominal size 0.05, hence there is no evidence for size distortion for any of the three tests. In terms of power (remaining processes) none of the tests does uniformly outperform the others, even within model classes such as the nonlinear moving average processes considered (process 2–4). This emphasizes again how hard it is to tell beforehand which test will perform best for an unknown alternative.

For applications I would have a slight preference for using a test that is consistent against any fixed alternative (such as the DP test), if only to hedge against the possibility of having no asymptotic power. However, as Table 1 shows, this does not guarantee a good finite sample performance in all cases.

Future Directions

Although permutation tests have been shown to have the advantage of providing exact tests in the ideal case of data that are truly i.i.d. under the null hypothesis, more work is required to establish the properties of these (or adapted) tests in the presence of residuals of an estimated parametric model. This requires either adaptation of the permutation procedure in that setting, or analogues of the 'nuisance parameter theorem' for the BDS test.

Another remaining challenge is the detection of dependence within observed high-variate vector-valued time series. Estimating (functionals of) probability densities in high-dimensional spaces is notoriously difficult, since the number of observations typically required grows very fast with the number of dimensions. Due to this so-called curse of dimensionality, the kernel-based methods discussed here in practice cannot be meaningfully applied to data sets with moderate sample sizes (several thousand observations) if the dimension m exceeds 5 or 6.

Additional work is also required for the development of statistical tests for time series that do not take values in the real line, but in more general manifolds. As mentioned in the introduction, an example consists of wind direction data taking values in the interval $[0, 2\pi]$ with the endpoints identified (i. e. on the circle). This touches upon the general problem of defining divergence measures between distributions on less familiar measurable spaces, and constructing and studying the statistical properties of their estimators.

Bibliography

1. Aparicio FM, Escribano A (1998) Information-theoretic analysis of serial dependence and cointegration. Stud nonlinear dyn econ 3:119–140
2. Ashley RA, Patterson DM (1986) A nonparametric, distribution-free test for serial independence in stock returns. J Financial Quant Anal 21:221–227
3. Ashley RA, Patterson DM (1989) Linear versus nonlinear macroeconomics: A statistical test. Int Econ Rev 30:165–187
4. Ashley RA, Patterson DM, Hinich MN (1986) A diagnostic check for nonlinear serial dependence in time series fitting errors. J Time Ser Anal 7:165–187
5. Barnard GA (1963) Discussion of Professor Bartlett's paper. J Royal Stat Soc Ser B 25:294
6. Bartels R (1982) The rank version of von Neumann's ratio test for randomness. J Am Stat Assoc 77:40–46
7. Benghabrit Y, Hallin M (1992) Optimal rank-based tests against 1st-order superdiagonal bilinear dependence. J Stat Plan Inference 32:45–61
8. Bera AK, Robinson PM (1989) Tests for serial dependence and other specification analysis in models of markets in equilibrium. J Bus Econ Stat 7:343–352
9. Beran J (1992) A goodness-of-fit test for time-series with long-range dependence. J Royal Stat Soc Ser B 54:749–760
10. Blum JR, Kiefer J, Rosenblatt M (1961) Distribution free tests of independence based on sample distribution functions. Ann Math Stat 32:485–498
11. Bollerslev T (1986) Generalized autoregressive heteroskedasticity. J Econometrics 31:307–327
12. Booth GG, Martikainen T (1994) Nonlinear dependence in Finnish stock returns. Eur J Oper Res 74:273–283
13. Box GEP, Pierce DA (1970) Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. J Am Stat Assoc 332:1509–1526
14. Bradley R (1986) Basic properties of strong mixing conditions. In: Eberlein E, Taqqu MS (eds) Dependence in Probability and Statistics. Birkäuser, Basel
15. Brock WA, Dechert WD, Scheinkman JA (1987) A test for independence based on the correlation dimension. Working paper 8702. University of Wisconsin, Madison
16. Brock WA, Dechert WD, Scheinkman JA, LeBaron B (1996) A test for independence based on the correlation dimension. Econometric Rev 15:197–235
17. Brockett PL, Hinich MD, Patterson D (1988) Bispectral based tests for the detection of Gaussianity and linearity in time series. J Am Stat Assoc 83:657–664
18. Brooks C, Heravi SM (1999) The effect of (mis-specified) GARCH filters on the filite sample distribution of the BDS test. Comput Econ 13:147–162
19. Brooks C, Hinich MJ (1999) Cross-correlations and cross-bicorrelations in Sterling exchange rates. J Empir Finance 6:385–404
20. Brooks C, Hinich MJ (2001) Bicorrelations and cross-bicorrelations as non-linearity tests and tools for exchange rate forecasting. J Forecast 20:181–196
21. Caporale GM, Ntantamis C, Pantelidis T, Pittis N (2005) The BDS test as a test for the adequacy of a GARCH(1,1) specification: A Monte Carlo study. J Financial Econometric 3:282–309
22. Carlstein E (1984) The use of sub-series methods for estimating the variance of a general statistic from a stationary time series. Ann Stat 14:1171–1179
23. Carlstein E (1988) Degenerate U-statistics based on non-independent observations. Calcutta Stat Assoc Bull 37:55–65
24. Chan NH, Tran LT (1992) Nonparametric tests for serial independence. J Time Ser Anal 13:19–28
25. Corrado CJ, Schatzberg J (1990) A nonparametric, distribution-free test for serial independence in stock returns: A correction. J Financial Quant Anal 25:411–415
26. Csörgő S (1985) Testing for independence by the empirical characteristic function. J Multivar Anal 16:290–299
27. Debnath L, Mikusiński P (2005) Introduction to Hilbert Spaces With Applications, 3rd edn. Elsevier Academic Press, Burlington
28. Delgado M, Mora J (2000) A nonparametric test for serial independence of regression errors. Biometrika 87:228–234
29. Delgado MA (1996) Testing serial independence using the sample distribution function. J Time Ser Anal 11:271–285
30. Denker M, Keller G (1983) On U-statistics and v. Mises' statistics for weakly dependent processes. Z Wahrscheinlichkeitstheorie verwandte Geb 64:505–522
31. Denker M, Keller G (1986) Rigorous statistical procedures for data from dynamical systems. J Stat Phys 44:67–93
32. Diks C, Panchenko V (2007) Nonparametric tests for se-

rial independence based on quadratic forms. Statistica Sin 17:81–97
33. Diks C, Panchenko V (2008) Rank-based entropy tests for serial independence. Stud Nonlinear Dyn Econom 12(1)art.2:0–19
34. Diks C, Tong H (1999) A test for symmetries of multivariate probability distributions. Biometrika 86:605–614
35. Dionísio A, Menezes R, Mendes DA (2006) Entropy-based independence test. Nonlinear Dyn 44:351–357
36. Dufour JM (1981) Rank tests for serial dependence. J Time Ser Anal 2:117–128
37. Dufour JM (2006) Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. J Econom 133:443–477
38. Durbin J, Watson GS (1950) Testing for serial correlation in least-squares regression, I. Biometrika 37:409–428
39. Durbin J, Watson GS (1951) Testing for serial correlation in least-squares regression, II. Biometrika 38:159–177
40. Durbin J, Watson GS (1971) Testing for serial correlation in least-squares regression, III. Biometrika 58:1–19
41. Durlauf S (1991) Spectral based testing of the martingale hypothesis. J Econometrics 50:355–376
42. Engle R (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50:987–1007
43. Ferguson TS, Genest C, Hallin M (2000) Kendall's tau for serial dependence. Can J Stat 28:587–604
44. Fernandes M, Neri B (2008) Nonparametric entropy-based tests of independence between stochastic processes. Econometric Reviews; Forthcoming
45. Genest C, Quessy JF, Rémillard B (2002) Tests of serial independence based on Kendall's process. Can J Stat 30:1–21
46. Genest C, Rémillard B (2004) Tests of independence and randomness based on the empirical copula process. Test 13:335–369
47. Genest C, Verret F (2005) Locally most powerful rank tests of independence for copula models. Nonparametric Stat 17:521–539
48. Genest C, Ghoudi K, Rémillard B (2007) Rank-based extensions of the Brock, Dechert, and Scheinkman test. J Am Stat Assoc 102:1363–1376
49. Ghoudi K, Kulperger RJ, Rémillard B (2001) A nonparametric test of serial independence for time series and residuals. J Multivar Anal 79:191–218
50. Granger C, Lin JL (2001) Using the mutual information coefficient to identify lags in nonlinear models. J Time Ser Anal 15:371–384
51. Granger CW, Maasoumi E, Racine J (2004) A dependence metric for possibly nonlinear processes. J Time Ser Anal 25:649–669
52. Grassberger P, Procaccia I (1983) Measuring the strangeness of strange attractors. Physica D 9:189–208
53. Grassberger P, Schreiber T, Schaffrath C (1991) Nonlinear time sequence analysis. Int J Bifurc Chaos 1:521–547
54. Hallin M, Ingenbleek J-F, Puri ML (1985) Linear serial rank tests for randomness against ARMA alternatives. Ann Stat 13:1156–1181
55. Hallin M, Mélard G (1988) Rank-based tests for randomness against first-order serial dependence. J Am Stat Assoc 83:1117–1128
56. Hannan EJ (1957) Testing for serial correlation in least squares regression. Biometrika 44:57–66
57. Hinich M, Patterson D (1985) Evidence of nonlinearity in stock returns. J Bus Econ Stat 3:69–77
58. Hinich MJ (1982) Testing for Gaussianity and linearity of a stationary time series. J Time Ser Anal 3:169–176
59. Hinich MJ (1996) Testing for dependence in the input to a linear time series model. J Nonparametric Stat 8:205–221
60. Hjellvik V, Tjøstheim D (1995) Nonparametric tests of linearity for time series. Biometrika 82:351–368
61. Hjellvik V, Yao Q, Tjøstheim D (1998) Linearity testing using polynomial approximation. J Stat Plan Inference 68:295–321
62. Hoeffding W (1948) A non-parametric test of independence. Ann Math Stat 19:546–557
63. Hong Y (1999) Hypothesis testing in time series via the empirical characteristic function: a generalized spectral density approach. J Am Stat Assoc 94:1201–1220
64. Hong Y (2000) Generalized spectral tests for serial dependence. J Royal Stat Soc Ser B 62:557–574
65. Hong Y, White H (2005) Asymptotic distribution theory for nonparametric entropy measures of serial dependence. Econometrica 73:837–901
66. Horowitz JL (2001) The bootstrap. In: Heckman JJ, Leamer EE (eds) Handbook of Econometrics, vol 5. Elsevier, Amsterdam, pp 3159–3228
67. Horowitz JL, Spokoiny VG (2001) An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. Econometrica 69:599–631
68. Joe H (1989) Relative entropy measures of multivariate dependence. J Am Stat Assoc 84:157–164
69. Joe H (1990) Multivariate concordance. J Multivar Anal 35:12–30
70. Johnson D, McLelland R (1997) Nonparametric tests for the independence of regressors and disturbances as specification tests. Rev Econ Stat 79:335–340
71. Johnson D, McLelland R (1998) A general dependence test and applications. J Appl Econometrics 13:627–644
72. Kallenberg WCM, Ledwina T (1999) Data driven rank tests for independence. J Am Stat Assoc 94:285–301
73. Kendall MG (1938) A new measure of rank correlation. Biometrika 30:81–93
74. Kočenda E, Briatka Ľ (2005) Optimal range for the IID test based on integration across the correlation integral. Econometric Rev 24:265–296
75. Kulperger RJ, Lockhart RA (1998) Tests of independence in time series. J Time Ser Anal 1998:165–185
76. Künsch HR (1989) The jackknife and the bootstrap for general stationary observations. Ann Stat 17:1217–1241
77. Lim KP, Hinich MJ, Liew VKS (2005) Statistical inadequacy of GARCH models for Asian stock markets: Evidence and implications. J Emerg Mark Finance 4:263–279
78. Lima P De (1996) Nuisance parameter free properties of correlation integral based statistics. Econometric Rev 15:237–259
79. Ljung GM, Box GEP (1978) On a measure of lack of fit in time series models. Biometrika 65:297–202
80. Lo AW (2000) Finance: A selective survey. J Am Stat Assoc 95:629–635
81. Maasoumi E (2002) Entropy and predictability of stock market returns. J Econometrics 107:291–312
82. McLeod AI, Li WK (1983) Diagnostic checking ARMA time series models using squared-residual autocorrelations. J Time

Ser Anal 4:269–273
83. Von Neumann J (1941) Distribution of the ratio of the mean square successive difference to the variance. Ann Math Stat 12:367–395
84. Pinkse J (1998) A consistent nonparametric test for serial independence. J Econometrics 84:205–231
85. Politis DN, Romano JP (1994) The stationary bootstrap. J Am Stat Assoc 89:1303–1313
86. Racine J, Maasoumi E (2007) A versatile and robust metric entropy test for time-irreversibility, and other hypotheses. J Econometrics 138:547–567
87. Ramsey JB, Rothman P (1990) Time irreversibility of stationary time series: estimators and test statistics. Unpublished manuscript, Department of Economics, New York University and University of Delaware
88. Robinson PM (1991) Consistent nonparametric entropy-based testing. Rev Econ Stud 58:437–453
89. Rosenblatt M (1975) A quadratic measure of deviation of two-dimensional density estimates and a test of independence. Ann Stat 3:1–14
90. Rosenblatt M, Wahlen BE (1992) A nonparametric measure of independence under a hypothesis of independent components. Stat Probab Lett 15:245–252
91. Rothman P (1992) The comparative power of the TR test against simple threshold models. J Appl Econometrics 7:S187–S195
92. Scaillet O (2005) A Kolmogorov–Smirnov type test for positive quadrant dependence. Can J Stat 33:415–427
93. Serfling RJ (1980) Approximation Theorems of Mathematical Statistics. Wiley, New York
94. Silverman BW (1986) Density Estimation for Statistics and Data Analysis. Chapman and Hall, New York
95. Skaug HJ, Tjøstheim D (1993a) A nonparametric test for serial independence based on the empirical distribution function. Biometrika 80:591–602
96. Skaug HJ, Tjøstheim D (1993b) Nonparametric tests of serial independence. In: Subba Rao T (ed) Developments in Time Series Analysis: the M. B. Priestley Birthday Volume. Wiley, New York
97. Spearman C (1904) The proof and measurement of association between two things. Am J Psychol 15:72–101
98. Subba Rao T, Gabr MM (1980) A test for linearity of stationary time series. J Time Ser Anal 1:145–158
99. Takens F (1981) Detecting strange attractors in turbulence. In: Rand DA, Young LS (eds) Dynamical Systems and Turbulence, Warwick 1980. (Lecture Notes in Mathematics), vol 898. Springer, Berlin, pp 366–381
100. Terdik G, Máth J (1998) A new test of linearity of time series based on the bispectrum. J Time Ser Anal 19:737–753
101. Theil H, Nagar AL (1961) Testing the independence of regression disturbances. J Am Stat Assoc 56:793–806
102. Theiler J (1990) Statistical precision of dimension estimators. Phys Rev A 41:3038–3051
103. Tjøstheim D (1996) Measures of dependence and tests of independence. Statistics 28:249–284
104. Tong H (1990) Non-linear Time Series: A Dynamical Systems Approach. Clarendon Press, Oxford
105. Tsallis C (1998) Generalized entropy-based criterion for consistent testing. Phys Rev E 58:1442–1445
106. Wolff RC (1994) Independence in time series: another look at the BDS test. Philos Trans Royal Soc Ser A 348:383–395

Nonsmooth Analysis in Systems and Control Theory

Francis Clarke
Institut universitaire de France et Université de Lyon, Lyon, France

Article Outline

Glossary
Definition of the Subject
Introduction
Elements of Nonsmooth Analysis
Necessary Conditions in Optimal Control
Verification Functions
Dynamic Programming and Viscosity Solutions
Lyapunov Functions
Stabilizing Feedback
Future Directions
Bibliography

Glossary

Generalized gradients and subgradients These terms refer to various set-valued replacements for the usual derivative which are used in developing differential calculus for functions which are not differentiable in the classical sense. The subject itself is known as *nonsmooth analysis*. One of the best-known theories of this type is that of *generalized gradients*. Another basic construct is the *subgradient*, of which there are several variants. The approach also features *generalized tangent and normal vectors* which apply to sets which are not classical manifolds. The article contains a summary of the essential definitions.

Pontryagin Maximum Principle The main theorem on necessary conditions in optimal control was developed in the 1950s by the Russian mathematician L. Pontryagin and his associates. The Maximum Principle unifies and extends to the control setting the classical necessary conditions of Euler and Weierstrass from the calculus of variations, as well as the transversality conditions. There have been numerous extensions since then, as the need to consider new types of problems continues to arise.

Verification functions In attempting to prove that a certain control is indeed the solution to a given optimal control problem, one important approach hinges upon exhibiting a function having certain properties implying the optimality of the given control. Such a function is termed a verification function. The approach

becomes widely applicable if one allows nonsmooth verification functions.

Dynamic programming A well-known technique in dynamic problems of optimization is to solve (in a discrete context) a backwards recursion for a certain value function related to the problem. This technique, which was developed notably by Bellman, can be applied in particular to optimal control problems. In the continuous setting, the recursion corresponds to the *Hamilton–Jacobi Equation*. This partial differential equation does not generally admit smooth classical solutions. The theory of *viscosity solutions* uses subgradients to define generalized solutions, and obtains their existence and uniqueness.

Lyapunov function In the classical theory of ordinary differential equations, global asymptotic stability is most often verified by exhibiting a *Lyapunov function*, a function along which trajectories decrease. In that setting, the existence of a smooth Lyapunov function is both necessary and sufficient for stability. The Lyapunov function concept can be extended to control systems, but in that case it turns out that nonsmooth functions are essential. These generalized *control Lyapunov functions* play an important role in designing optimal or stabilizing feedback.

Definition of the Subject

The term *nonsmooth analysis* refers to the body of theory which develops differential calculus for functions which are not differentiable in the usual sense, and for sets which are not classical smooth manifolds. There are several different (but related) approaches to doing this. Among the better-known constructs of the theory are the following: generalized gradients and Jacobians, proximal subgradients, subdifferentials, generalized directional (or Dini) derivates, together with various associated tangent and normal cones. Nonsmooth analysis is a subject in itself, within the larger mathematical field of differential (variational) analysis or functional analysis, but it has also played an increasingly important role in several areas of application, notably in optimization, calculus of variations, differential equations, mechanics, and control theory. Among those who have participated in its development (in addition to the author) are J. Borwein, A. D. Ioffe, B. Mordukhovich, R. T. Rockafellar, and R. B. Vinter, but many more have contributed as well.

In the case of control theory, the need for nonsmooth analysis first came to light in connection with finding proofs of necessary conditions for optimal control, notably in connection with the Pontryagin Maximum Principle. This necessity holds even for problems which are expressed entirely in terms of smooth data. Subsequently, it became clear that problems with intrinsically nonsmooth data arise naturally in a variety of optimal control settings. Generally, nonsmooth analysis enters the picture as soon as we consider problems which are truly nonlinear or nonlinearizable, whether for deriving or expressing necessary conditions, in applying sufficient conditions, or in studying the sensitivity of the problem.

The need to consider nonsmoothness in the case of stabilizing (as opposed to optimal) control has come to light more recently. It appears in particular that in the analysis of truly nonlinear control systems, the consideration of nonsmooth Lyapunov functions and discontinuous feedbacks becomes unavoidable.

Introduction

The basic object in the control theory of ordinary differential equations is the system

$$\dot{x}(t) = f(x(t), u(t)) \ \text{ a.e.}, \quad 0 \le t \le T, \tag{1}$$

where the (measurable) *control function* $u(\cdot)$ is chosen subject to the constraint

$$u(t) \in U \ \text{ a.e.}, \tag{2}$$

where U is a given set in an Euclidean space, and where the ensuing *state* $x(\cdot)$ (a function with values in $\mathbb{R}^n$) is subject to certain conditions, including most often an initial one of the form $x(0) = x_0$, and perhaps other constraints, either throughout the interval (pointwise) or at the terminal time. A control function u of this type is referred to as an *open loop control*. This indirect control of $x(\cdot)$ via the choice of $u(\cdot)$ is to be exercised for a purpose, of which there are two principal sorts:

- *positional*: $x(t)$ is to remain in a given set in $\mathbb{R}^n$, or approach that set;
- *optimal*: $x(\cdot)$, together with $u(\cdot)$, is to minimize a given functional.

The second of these criteria follows directly in the tradition of the calculus of variations, and gives rise to the subject of *optimal control*, in which the dominant issues are those of optimization: necessary conditions for optimality, sufficient conditions, regularity of the optimal control, sensitivity. We discuss below the role of nonsmooth analysis in optimal control; this was the setting of many of the earliest applications.

In contrast, a prototypical control problem of purely positional sort would be the following:

Find a control $u(\cdot)$ such that $x(\cdot)$ goes to 0 .

This rather vaguely worded goal is often more precisely expressed as that of finding a *stabilizing feedback control* $k(x)$; that is, a function k with values in U (this is often referred to as a closed-loop control) such that all solutions of the differential equation

$$\dot{x}(t) = f\big(x(t), k\big(x(t)\big)\big)\,, \quad x(0) = \alpha \tag{3}$$

converge to 0 (for all values of α) in a suitable sense.

The most common approach to designing the required stabilizing feedback uses the technique that is central to most of applied mathematics: *linearization*. In this case, one examines the linearized system

$$\dot{x}(t) = Ax(t) + Bu(t)$$

where

$$A := f_x(0,0)\,, \quad B := f_u(0,0)\,.$$

If the linearized system satisfies certain controllability properties, then classical linear systems theory provides well-known and powerful tools for designing (linear) feedbacks that stabilize the linearized system. Under further mild hypotheses, this yields a feedback that stabilizes the original nonlinear system (1) *locally*; that is, for initial values $x(0)$ sufficiently near 0.

This approach has been feasible in a large number of cases, and in fact it underlies the very successful role that control theory has played in a great variety of applications. Still, linearization does require that a certain number of conditions be met:

- The function f must be smooth (differentiable) so that the linear system can be constructed;
- The linear system must be a 'nondegenerate' approximation of the nonlinear one (that is, it must be controllable);
- The control set U must contain a neighborhood of 0, so that near 0 the choice of controls is unconstrained, and all state values near 0 must be acceptable (no state constraints);
- Both x and u must remain small so that the linear approximation remains relevant (in dealing with errors or perturbations, the feedback is operative only when they are sufficiently small).

It is not hard to envisage situations in which the last two conditions fail, and indeed such challenging problems are beginning to arise increasingly often. The first condition fails for simple problems involving electrical circuits in which a diode is present, for example (see [14]). A famous (smooth) mechanical system for which the second condition fails is the *nonholonomic integrator*, a term which refers to the following system, which is linear (separately) in the state and in the control variables:

$$\begin{cases} \dot{x}_1 = u_1 \\ \dot{x}_2 = u_2 \\ \dot{x}_3 = x_1u_2 - x_2u_1 \end{cases} \qquad \|(u_1, u_2)\| \le 1\,.$$

(Thus $n = 3$ here, and U is the closed unit ball in $\mathbb{R}^2$.) Here, the linearized system is degenerate, since its third component is $\dot{x}_3 = 0$. As discussed below, there is in fact no continuous feeback law $u = k(x)$ which will stabilize this system (even locally about the origin), but certain discontinuous stabilizing feedbacks do exist.

This illustrates the moral that when linearization is not applicable to a given nonlinear system (for whatever reason), nonsmoothness generally arises. (This has been observed in other contexts in recent decades: catastrophes, chaos, fractals.) Consider for example the issue of whether a (control) Lyapunov function exists. This refers to a pair (V, W) of positive definite functions satisfying notably the following *Infinitesimal Decrease* condition:

$$\inf_{u\in U} \langle \nabla V(x), f(x,u)\rangle \;\le\; -W(x) \quad x \ne 0\,.$$

The existence of such (smooth) functions implies that the underlying control system is *globally asymptotically controllable* (GAC), which is a necessary condition for the existence of a stabilizing feedback (and also sufficient, as recently proved [21]). In fact, exhibiting a Lyapunov function is the principal technique for proving that a given system is GAC; the function V in question then goes on to play a role in designing the stabilizing feedback.

It turns out, however, that even smooth systems that are GAC need not admit a smooth Lyapunov function. (The nonholonomic integrator is an example of this phenomenon.) But if one extends in a suitable way the concept of Lyapunov function to nonsmooth functions, then the existence of a Lyapunov function becomes a necessary and sufficient condition for a given system to be GAC. One such extension involves replacing the gradient that appears in the Infinitesimal Decrease condition above by elements of the *proximal subdifferential* (see below). How to use such extended Lyapunov functions to design a stabilizing feedback is a nontrivial topic that has only recently been successfully addressed.

In the next section we define a few basic constructs of nonsmooth analysis; knowledge of these is sufficient for interpreting the statements of the results discussed in this article.

Elements of Nonsmooth Analysis

This section summarizes some basic notions in nonsmooth analysis, in fact a minimum so that the statements of the results to come can be understood. This minimum corresponds to three types of set-valued generalized derivative (generalized gradients, proximal subgradients, limiting subgradients), together with a notion of normal vector applicable to any closed (not necessarily smooth or convex) set. (See [24] for a thorough treatment and detailed references.)

Generalized Gradients

For *smooth* real-valued functions f on $\mathbb{R}^n$ we have a well-known formula linking the usual directional derivative to the gradient:

$$f'(x;v) := \lim_{t\downarrow 0} \frac{f(x+tv)-f(x)}{t} = \langle \nabla f(x), v\rangle \ .$$

We can extend this pattern to *Lipschitz* functions: f is said to be Lipschitz on a set S if there is a constant K such that $|f(y)-f(z)| \le K\|y-z\|$ whenever y and z belong to S. A function f that is Lipschitz in a neighborhood of a point x is not necessarily differentiable at x, but we can define a *generalized directional derivative* as follows:

$$f^\circ(x;v) := \limsup_{y\to x,\ t\downarrow 0} \frac{f(y+tv)-f(y)}{t} \ .$$

Having done so, we proceed to define the *generalized gradient*:

$$\partial_C f(x) := \{\zeta \in \mathbb{R}^n : f^\circ(x;v) \ge \langle \zeta, v\rangle \ \ \forall v \in X\} \ .$$

It turns out that $\partial_C f(x)$ is a compact convex nonempty set that has a calculus reminiscent of the usual differential calculus; for example, we have $\partial(-f)(x) = -\partial f(x)$. We also have $\partial(f+g)(x) \subset \partial f(x) + \partial g(x)$; note that (as is often the case) this is an inclusion, not an equation. There is also a useful analogue of the Mean Value Theorem, and other familiar results. In addition, though, there are new formulas having no smooth counterpart, such as one for the generalized gradient of the pointwise maximum of locally Lipschitz functions ($\partial\{\max_{1\le i\le n} f_i(x)\} \subset \ldots$).

A very useful fact for actually calculating $\partial_C f(x)$ is the following Gradient Formula:

$$\partial_C f(x) = \mathrm{co}\left\{ \lim_{i\to\infty} \nabla f(x_i) : x_i \to x, x_i \notin \Gamma \right\} \ ,$$

where Γ is *any* set of measure zero containing the local points of nondifferentiability of f. Thus the generalized gradient is 'blind to sets of measure zero'.

Generalized gradients and their calculus were defined by Clarke [12] in 1973. The theory can be developed on any Banach space; the infinite-dimensional context is essential in certain control applications, but for our present purposes it suffices to limit attention to functions defined on $\mathbb{R}^n$. There is in addition a corresponding theory of tangent and normal vectors to arbitrary closed sets; we give some elements of this below.

Proximal Subgradients

We now present a different approach to developing nonsmooth calculus, one that uses the notion of *proximal subgradient*. Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be a given function (note that the value $+\infty$ is admitted here), and let x be a point where $f(x)$ is finite. A vector ζ in $\mathbb{R}^n$ is said to be a proximal subgradient of f at x provided that there exist a neighborhood Ω of x and a number $\sigma \ge 0$ such that

$$f(y) \ge f(x) + \langle \zeta, y-x\rangle - \sigma\|y-x\|^2 \ \ \forall y \in \Omega \ .$$

Thus the existence of a proximal subgradient ζ at x corresponds to the possibility of approximating f from below (thus in a *one-sided* manner) by a function whose graph is a parabola. The point $(x, f(x))$ is a contact point between the graph of f and the parabola, and ζ is the slope of the parabola at that point. Compare this with the usual derivative, in which the graph of f is approximated (in a two-sided way) by an affine function.

The set of proximal subgradients at x (which may be empty, and which is not necessarily closed, open, or bounded but which is convex) is denoted $\partial_P f(x)$, and is referred to as the *proximal subdifferential*. If f is differentiable at x, then we have $\partial_P f(x) \subset \{f'(x)\}$; equality holds if f is of class C^2 at x.

As a guide to understanding, the reader may wish to carry out the following exercise (in dimension $n = 1$): the proximal subdifferential at 0 of the function $f_1(x) := -|x|$ is empty, while that of $f_2(x) := |x|$ is the interval $[-1, 1]$.

The *proximal density theorem* asserts that $\partial_P f(x)$ is nonempty for all x in a dense subset of

$$\mathrm{dom}\, f := \{x : f(x) < \infty\} \ .$$

Although it can be empty at many points, the proximal subgradient admits a very complete calculus for the class of lower semicontinuous functions: all the usual calculus rules that the reader knows (and more) have their counterpart in terms of $\partial_P f$. Let us quote for example Ioffe's *fuzzy sum rule*: if $\zeta \in \partial_P(f+g)(x)$, then for any $\epsilon > 0$ there exist x' and x'' within ϵ of x, together with points

$\zeta' \in \partial_P f(x')$ and $\zeta'' \in \partial_P g(x'')$ such that

$$\zeta \in \zeta' + \zeta'' + \epsilon B .$$

The Limiting Subdifferential

A sequential closure operation applied to $\partial_P f$ gives rise to the *limiting subdifferential*, useful for stating results:

$$\partial_L f(x) := \{\lim \zeta_i : \zeta_i \in \partial_P f(x_i), x_i \to x, f(x_i) \to f(x)\} .$$

If f is Lipschitz near x, then $\partial_L f(x)$ is nonempty, and (for any lower semicontinuous g finite at x) we have

$$\partial_L (f+g)(x) \subset \partial_L f(x) + \partial_L g(x) .$$

When the function f is Lipschitz near x, then both the approaches given above (generalized gradients, proximal subdifferential) apply, and the corresponding constructs are related as follows:

$$\partial_C f(x) = \text{co}\ \partial_L\ f(x) .$$

Normal Vectors

Given a nonempty closed subset S of $\mathbb{R}^n$ and a point x in S, we say that $\zeta \in X$ is a *proximal normal* (vector) to S at x if there exists $\sigma \geq 0$ such that

$$\langle \zeta, x' - x \rangle \leq \sigma \left\| x' - x \right\|^2 \ \ \forall x' \in S .$$

(This is the *proximal normal inequality*.) The set (convex cone) of such ζ, which always contains 0, is denoted $N_S^P(x)$ and referred to as the proximal normal cone. We apply to $N_S^P(x)$ a sequential closure operation in order to obtain the *limiting normal cone*:

$$N_S^L(x) := \left\{\lim \zeta_i : \zeta_i \in N_S^P(x_i), x_i \to x, x_i \in S\right\} .$$

These geometric notions are consistent with the analytical ones, as illustrated by the formulas

$$\partial_P I_S(x) = N_S^P(x), \ \partial_L I_S(x) = N_S^L(x) ,$$

where I_S denotes the *indicator* of the set S: the function which equals 0 on S and $+\infty$ elsewhere. They are also consistent with the more traditional ones: When S is either a convex set, a smooth manifold, or a manifold with boundary, then both $N_S^P(x)$ and $N_S^L(x)$ coincide with the usual normal vectors (a cone, space, or half-space respectively).

Viscosity Subdifferentials

We remark that the *viscosity subdifferential* of f at x (commonly employed in pde's, but not used in this article) corresponds to the set of ζ for which we have

$$f(y) \geq f(x) + \langle \zeta, y - x \rangle + \theta(|y - x|) ,$$

where θ is a function such that $\lim_{t \downarrow 0} \theta(t)/t = 0$. This gives a potentially larger set than $\partial_P f(x)$; however, the viscosity subdifferential satisfies the same (fuzzy) calculus rules as does $\partial_P f$. In addition, the sequential closure operation described above gives rise to the *same* limiting subdifferential $\partial_L f(x)$ (see [24]). In most cases, therefore, it is equivalent to work with viscosity or proximal subgradients.

Necessary Conditions in Optimal Control

The central theorem on necessary conditions for optimal control is the Pontryagin Maximum Principle. (The literature on necessary conditions in optimal control is now very extensive; we cite some standard references in the bibliography.) Even in the somewhat special (by current standards) smooth context in which it was first proved [47], an element of generalized differentiability was required. With the subsequent increase in both the generality of the model and weakening of the hypotheses (all driven by real applications), the need for nonsmooth analysis is all the greater, even for the very statement of the result.

We give here just one example, a broadly applicable *hybrid maximum principle* taken from [11], in order to highlight the essential role played by nonsmooth analysis as well as the resulting versatility of the results obtained.

The Problem and Basic Hypotheses

We consider the minimization of the functional

$$\ell(x(a), x(b)) + \int_a^b F(t, x(t), u(t))\, \mathrm{d}t$$

subject to the boundary conditions $(x(a), x(b)) \in S$ and the standard control dynamics

$$\dot{x}(t) = f(t, x(t), u(t)) \ \ \text{a.e.}$$

The minimization takes place with respect to (absolutely continuous) state arcs x and measurable control functions $u : [a, b] \to \mathbb{R}^m$. Note that no explicit constraints are placed upon $u(t)$; if such constraints exist, they are accounted for by assigning to the extended-valued integrand F the value $+\infty$ whenever the constraints are violated. It is part of our intention here to demonstrate the

utility of indirectly taking account of constraints in this way. As basic hypotheses, we assume that F is measurable and lower semicontinuous in (x, u); ℓ is taken to be locally Lipschitz and S closed. For the theorem below, we require that f be measurable in t and continuously differentiable in (x, u) (but see the remarks for a weakening of the smoothness).

The Growth Conditions

The principal hypothesis here is that f and F satisfy the following: for every bounded subset X of $\mathbb{R}^n$, there exist a constant c and a summable function d such that, for almost every t, for every $(x, u) \in \operatorname{dom} F(t, \cdot, \cdot)$ with $x \in X$, we have

$$\|D_x f(t, x, u)\| \le c\,\{|f(t, x, u)| + F(t, x, u)\} + d(t)\,,$$

and for all (ζ, ψ) in $\partial_P F(t, x, u)$ (if any) we have

$$\frac{|\zeta|\,(1 + \|D_u f(t, x, u)\|)}{1 + |\psi|} \le c\,\{|f(t, x, u)| + F(t, x, u)\} + d(t)\,.$$

In the following, $(*)$ denotes evaluation at $(x_*(t), u_*(t))$.

Theorem 1 *Let the control function u_* give rise to an arc x_* which is a local minimum for the problem above. Then there exist an arc p on $[a, b]$ such that the following transversality condition holds*

$$(p(a), -p(b)) \in \partial_L \ell(x_*(a), x_*(b)) + N_S^L(x_*(a), x_*(b))\,,$$

and such that p satisfies the adjoint inclusion: $\dot{p}(t)$ belongs almost everywhere to the set

$$\operatorname{co}\big\{\omega\colon (\omega + D_x^* f(t, *)p(t), D_u^* f(t, *)p(t)) \in \partial_L F(t, *)\big\}\,,$$

as well as the maximum condition: for almost every t, for every u in $\operatorname{dom} F(t, x_*(t), \cdot)$, one has

$$\langle p(t), f(t, x_*(t), u)\rangle - F(t, x_*(t), u) \le \langle p(t), f(t, *)\rangle - F(t, *)\,.$$

We proceed to make a few remarks on this theorem, beginning with the fact that it can fail in the absence of the growth condition, even for smooth problems of standard type [30]. The statement of the theorem is not complete, since in general the necessary conditions may hold only in *abnormal* form; we do not discuss this technical point here for reasons of economy. The phrase 'local minimum' (which we have also not defined) can be interpreted very generally, see [11]. The 'maximum condition' above is of course what leads to the name 'maximum principle', while the exotic-looking adjoint inclusion reduces to more familiar conditions in a variety of special cases (see below). The transversality condition illustrates well the utility of couching in nonsmooth analysis terms the conclusion, since the given form encapsulates simultaneously a wide variety of conclusions obtainable in special cases. To give but one example, consider an optimal control problem in which $x(a)$ is prescribed, $x(b)$ does not appear in the cost, but is subject to an inequality constraint $g(x(b)) \le 0$ for a certain smooth scalar function g. Then the transversality condition of the theorem (interpreted via standard facts in nonsmooth analysis) asserts that $p(b)$ is of the form $\lambda \nabla g(x_*(b))$ for some nonpositive number λ (a Lagrange multiplier).

To illustrate the versatility of the theorem, we look at some special cases.

The Standard Problem

The first case we examine is that in which for each t,

$$F(t, x, u) = I_{U(t)}(u)\,,$$

the indicator function which takes the value 0 when $u \in U(t)$ and $+\infty$ otherwise. This simply corresponds to imposing the condition $u(t) \in U(t)$ on the admissible controls u; this is the *Mayer form* of the problem (no integral term in the cost, obtained by reformulation if necessary). Note that in this case the second growth condition is trivially satisfied (since $\zeta = 0$). The first growth condition is active only on $U(t)$, and certainly holds if f is smooth in (t, x, u) and $U(t)$ is uniformly bounded. The hybrid adjoint inclusion immediately implies the standard adjoint equation

$$-\dot{p}(t) = D_x^* f(t, x_*(t), u_*(t))\; p(t)\,,$$

and we recover the conclusions of the classical Maximum Principle of Pontryagin. (An extension is obtained when $U(t)$ not bounded, see [10].) When f is not assumed to be differentiable, but merely locally Lipschitz with respect to x, there is a variant of the theorem in which the adjoint inclusion is expressed as follows:

$$-\dot{p}(t) \in \partial_C\,\langle p(t), f(t, \cdot, u_*(t))\rangle\;(x_*(t))\,,$$

where the generalized gradient ∂_C (see Sect. "Elements of Nonsmooth Analysis") is taken with respect to the x variable (note the connection to the standard adjoint equation above). This is an early form of the nonsmooth maximum principle [13].

The Calculus of Variations

When we take $f(t,x,u) = u$, the problem reduces to the *problem of Bolza* in the calculus of variations. The first growth condition is trivially satisfied, and the second coincides with the generalized Tonelli–Morrey growth condition introduced in [11]; in this way we recover the state-of-the-art necessary conditions for the generalized problem of Bolza, which include as a special case the multiplier rule for problems with pointwise and/or isoperimetric constraints.

Differential Inclusions

When we specialize further by taking $F(t,\cdot)$ to be the indicator of the graph of a multifunction $M(t,\cdot)$, we obtain the principal necessary conditions for the *differential inclusion problem*. These in turn lead to necessary conditions for *generalized control systems* [11].

Mixed Constraints

Consider again the optimal control problem in Mayer form (no integral term in the cost), but now in the presence of mixed state/control pointwise constraints of the form $(x(t),u(t)) \in \Omega$ a.e. for a given closed set Ω. Obtaining general necessary conditions for such problems is a well-known challenge in the subject; see [33,34,46]. We treat this case by taking $F(t,\cdot) = I_\Omega(\cdot)$. Then the second growth condition reduces to the following geometric assumption: for every $(x,u) \in \Omega$, for every $(\zeta,\psi) \in N^P_\Omega(x,u)$, one has

$$\frac{|\zeta|\,(1 + \|D_u f(t,x,u)\|)}{|\psi|} \le c\,|f(t,x,u)| + d(t)\,.$$

By taking a suitable representation for Ω in terms of functional equalities and/or inequalities, sufficient conditions in terms of rank can be adduced which imply this property, leading to explicit multiplier rules (see [10] for details). With an appropriate 'transversal intersection' condition, we can also treat the case in which *both* the constraints $(x(t),u(t)) \in \Omega$ and $u(t) \in U(t)$ are present.

Sensitivity

The multiplier functions p that appear in necessary conditions such as the ones above play a central role in analyzing the sensitivity of optimal control problems that depend on parameters. To take but one example, consider the presence of a perturbation $\alpha(\cdot)$ in the dynamics of the problem:

$$\dot{x}(t) = f(t,x(t),u(t)) + \alpha(t) \quad \text{a.e.}$$

Clearly the minimum in the problem depends upon α; we denote it $V(\alpha)$. The function V is an example of what is referred to as a *value function*. Knowledge of the derivative of V would be highly relevant is studying the sensitivity of the problem to perturbations (errors) in the dynamics. Generally, however, value functions are not differentiable, so instead one uses nonsmooth analysis; the multipliers p give rise to estimates on the generalized gradient of V. We refer to [15,16] for examples and references.

Verification Functions

We consider now a special case of the problem considered in the preceding section: to minimize the integral cost functional

$$J(x,u) := \int_a^b F(t,x(t),u(t))\,\mathrm{d}t$$

subject to the prescribed boundary conditions

$$x(a) = A\,, \quad x(b) = B$$

and the standard control dynamics

$$\dot{x}(t) = f(t,x(t),u(t)) \text{ a.e.}, \;\; u(t) \in U(t) \;\; \text{a.e.}$$

Suppose now that a candidate (x_*,u_*) has been identified as a possible solution to this problem (perhaps through a partial analysis of the necessary conditions, or otherwise). An elementary yet powerful way (traceable to Legendre) to prove that (x_*,u_*) actually does solve the problem is to exhibit a function $\phi(t,x)$ such that:

$$F(t,x,u) \ge \phi_t(t,x) + \langle \phi_x(t,x), f(t,x,u)\rangle \qquad \forall t \in [a,b],\; (x,u) \in \mathbb{R}^n \times U(t)\,,$$

with equality along $(t,x_*(t),u_*(t))$.

The mere existence of such a function verifies that (x_*,u_*) is optimal, as we now show. For any admissible state/control pair (x,u), we have

$$\begin{aligned} F(t,x(t),u(t)) &\ge \phi_t(t,x(t)) + \langle \phi_x(t,x(t)), \dot{x}(t)\rangle \\ &= \tfrac{\mathrm{d}}{\mathrm{d}t}\{\phi(t,x(t))\} \end{aligned}$$

$$\begin{aligned} \Rightarrow J(x,u) := \int_a^b F(t,x(t),u(t))\,\mathrm{d}t &\ge \phi(t,x(t))\big|_{t=a}^{t=b} \\ &= \phi(b,B) - \phi(a,A)\,. \end{aligned}$$

But this lower bound on J holds with equality when $(x,u) = (x_*,u_*)$, which proves that (x_*,u_*) is optimal.

In this argument, we have implicitly supposed that the *verification function* φ is smooth. It is a fact that if

we limit ourselves to smooth verification functions, then there may not exist such a φ (even when the problem itself has smooth data). However, if we admit nonsmooth (locally Lipschitz) verification functions, then (under mild hypotheses on the data) the existence of a verification function φ becomes necessary and sufficient for (x_*, u_*) to be optimal.

An appropriate way to extend the smooth inequality above uses the generalized gradient $\partial_C \phi$ of φ as follows:

$$F(t, x, u) \geq \theta + \langle \zeta, f(t, x, u) \rangle \ \ \forall\, (\theta, \zeta) \in \partial_C \phi(t, x)\,,$$

together with the requirement

$$J(x_*, u_*) = \phi(b, B) - \phi(a, A)\,.$$

It is a feature of this approach to proving optimality that it extends readily to more general problems, for example those involving unilateral state constraints, boundary costs, or isoperimetric conditions. It can also be interpreted in terms of duality theory, as shown notably by Vinter. We refer to [26] for details.

The basic inequality that lies at the heart of this approach is of Hamilton–Jacobi type, and the fact that we are led to consider nonsmooth verification functions is related to the phenomenon that Hamilton–Jacobi Equations may not admit smooth solutions. This is one of the main themes of the next section.

Dynamic Programming and Viscosity Solutions

The Minimal Time Problem

By a *trajectory* of the standard control system

$$\dot{x}(t) = f(x(t), u(t)) \text{ a.e.}, \ \ u(t) \in U \ \ \text{a.e.}$$

we mean a state function $x(\cdot)$ corresponding to some choice of admissible control function $u(\cdot)$. The *minimal time problem* refers to finding a trajectory that reaches the origin as quickly as possible. Thus we seek the least $T \geq 0$ admitting a control function $u(\cdot)$ on $[0, T]$ having the property that the resulting trajectory x satisfies $x(T) = 0$. We proceed now to describe the well-known *dynamic programming* approach to solving the problem.

We begin by introducing the *minimal time function* $T(\cdot)$, defined on $\mathbb{R}^n$ as follows: $T(\alpha)$ is the least time $T \geq 0$ such that some trajectory $x(\cdot)$ satisfies

$$x(0) = \alpha, x(T) = 0\,.$$

An issue of *controllability* arises here: Is it always possible to steer α to 0 in finite time? When such is not the case, then in accord with the usual convention we set $T(\alpha) = +\infty$.

The *principle of optimality* is the dual observation that if $x(\cdot)$ is any trajectory, then we have, for $s < t$,

$$T\big(x(t)\big) - T\big(x(s)\big) \geq s - t\,.$$

Equivalently, the function

$$t \mapsto T\big(x(t)\big) + t$$

is increasing. Furthermore, if x is optimal, then the same function is constant.

Let us explain this in other terms: if $x(\cdot)$ is an optimal trajectory joining α to 0, then

$$T\big(x(t)\big) = T(\alpha) - t \quad \text{for} \quad 0 \leq t \leq T(\alpha)\,,$$

since an optimal trajectory from the point $x(t)$ is furnished by the truncation of $x(\cdot)$ to the interval $\big[t, T(\alpha)\big]$. If $x(\cdot)$ is any trajectory, then the inequality

$$T\big(x(t)\big) \geq T(\alpha) - t$$

is a reflection of the fact that in going to the point $x(t)$ from α (in time t), we may have acted optimally (in which case equality holds) or not (then inequality holds).

Since $t \mapsto T\big(x(t)\big) + t$ is increasing, we expect to have

$$\big\langle \nabla T\big(x(t)\big), \dot{x}(t) \big\rangle + 1 \geq 0\,,$$

with equality when $x(\cdot)$ is an optimal trajectory. The possible values of $\dot{x}(t)$ for a trajectory being precisely the elements of the set $f\big(x(t), U\big)$, we arrive at

$$\min_{u \in U} \big\langle \nabla T(x), f(x(t), u) \big\rangle + 1 = 0\,. \tag{4}$$

We define the (lower) *Hamiltonian function* h as follows:

$$h(x, p) := \min_{u \in U} \langle p, f(x, u) \rangle\,.$$

In terms of h, the partial differential equation obtained above reads

$$h\big(x, \nabla T(x)\big) + 1 = 0\,, \tag{5}$$

a special case of the *Hamilton–Jacobi Equation*.

Here is the first step in the dynamic programming heuristic: use the Hamilton–Jacobi Equation (5), together with the boundary condition $T(0) = 0$, to find $T(\cdot)$. How will this help us find the optimal trajectory?

To answer this question, we recall that an optimal trajectory is such that equality holds in (4). This suggests the

following procedure: for each x, let $k(x)$ be a point in U satisfying

$$\min_{u\in U}\langle \nabla T(x), f(x,u)\rangle = \langle \nabla T(x), f(x,k(x))\rangle = -1\,. \quad (6)$$

Then, if we construct $x(\cdot)$ via the initial-value problem

$$\dot{x}(t) = f\big(x(t), k\big(x(t)\big)\big)\,, \quad x(0)=\alpha\,, \quad (7)$$

we will have a trajectory that is optimal (from α)!

Here is why: Let $x(\cdot)$ satisfy (7); then $x(\cdot)$ is a trajectory, and

$$\begin{aligned}\tfrac{\mathrm{d}}{\mathrm{d}t}\,T\big(x(t)\big) &= \langle \nabla T\big(x(t)\big), \dot{x}(t)\rangle \\ &= \langle \nabla T\big(x(t)\big), f\big(x(t), k\big(x(t)\big)\big)\rangle = -1\,.\end{aligned}$$

Integrating, we find

$$T\big(x(t)\big) = T(\alpha) - t\,,$$

which implies that at $t = T(\alpha)$, we must have $x = 0$. Therefore $x(\cdot)$ is an optimal trajectory.

Let us stress the important point that $k(\cdot)$ generates the optimal trajectory from *any* initial value α (via (7)), and so constitutes what can be considered the ultimate solution for this problem: an *optimal feedback synthesis*. There can be no more satisfying answer to the problem: If you find yourself at x, just choose the control value $k(x)$ to approach the origin as fast as possible. This goes well beyond finding a single open-loop optimal control.

Unfortunately, there are serious obstacles to following the route that we have just outlined, beginning with the fact that T is nondifferentiable, as simple examples show, even when it is finite everywhere (which it generally fails to be).

We will therefore have to examine anew the argument that led to the Hamilton–Jacobi Equation (5), which, in any case, will have to be recast in some way to accommodate nonsmooth solutions. Having done so, will the generalized Hamilton–Jacobi Equation admit T as the unique solution?

The next step (after characterizing T) offers fresh difficulties of its own. Even if T were smooth, there would be in general no *continuous* function $k(\cdot)$ satisfying (6) for each x. The meaning and existence of a trajectory $x(\cdot)$ generated by $k(\cdot)$ via the differential Eq. (7), in which the right-hand side is discontinuous in the state variable, is therefore problematic in itself.

The intrinsic difficulties of this approach to the minimal-time problem have made it a historical focal point of activity in differential equations and control, and it is only recently that fully satisfying answers to all the questions raised above have been found. We begin with generalized solutions of the Hamilton–Jacobi Equation.

Subdifferentials and Viscosity Solutions

We shall say that φ is a *proximal solution* of the Hamilton–Jacobi Equation (5) provided that

$$h(x, \partial_P\phi(x)) = -1 \quad \forall x \in \mathbb{R}^n\,, \quad (8)$$

a 'multivalued equation' which means that for all x, for all $\zeta \in \partial_P\phi(x)$ (if any), we have $h(x,\zeta) = -1$. (Recall that the proximal subdifferential $\partial_P\phi$ was defined in Sect. "Elements of Nonsmooth Analysis".)

Note that the equation holds automatically at a point x for which $\partial_P\phi(x)$ is empty; such points play an important role, in fact, as we now illustrate. Consider the case in which $f(x,U)$ is equal to the unit ball for all x, in dimension $n=1$. Then $h(x,p) \equiv -|p|$ (and the equation is of *eikonal* type). Let us examine the functions $\phi_1(x) := -|x|$ and $\phi_2(x) := |x|$; they both satisfy $h(x, \nabla\phi(x)) = -1$ at all points $x \neq 0$, since (for each of these functions) the proximal subdifferential at points different from 0 reduces to the singleton consisting of the derivative. However, we have (see the exercise in the summary of nonsmooth analysis)

$$\partial_P\phi_1(0) = \emptyset,\ \ \partial_P\phi_2(0) = [-1,1]\,,$$

and it follows that ϕ_1 is (but ϕ_2 is not) a proximal solution of the Hamilton–Jacobi Equation (8).

A lesson to be drawn from this example is that in defining generalized solutions we need to look closely at the differential behavior at specific and individual points; we cannot argue in an 'almost everywhere' fashion, or by 'smearing' via integration (as is done for linear partial differential equations via distributional derivatives).

Proximal solutions are just one of the ways to define generalized solutions of certain partial differential equations, a topic of considerable interest and activity, and one which seems to have begun with the Hamilton–Jacobi Equation in every case. The first 'subdifferential type' of definition was given by the author in the 1970s, using generalized gradients and for locally Lipschitz solutions. While no uniqueness theorem holds for that solution concept, it was shown that the value function of the associated optimal control problem is a solution (hence existence holds), and is indeed a special solution: it is the maximal one. In 1980 A. I. Subbotin defined his 'minimax solutions', which are couched in terms of Dini derivates rather than subdifferentials, and which introduced the important feature of being 'two-sided'. This work featured existence and uniqueness in the class of Lipschitz functions, the solution being characterized as the value of a differential game. Subsequently, M. Crandall and P.-L. Lions incorporated both subdifferentials and two-sidedness in their *viscosity solutions*, a theory which they developed for merely

continuous functions. In the current context, and under mild hypotheses on the data, it can be shown that minimax, viscosity, and proximal solutions all coincide [19,24].

Recall that our goal (within the dynamic programming approach) is to characterize the minimal time function. This is now attained, as shown by the following (we omit the hypotheses; see [63], and also the extensive discussion in Bardi and Capuzzo-Dolcetta [4]):

Theorem 2 *There exists a unique lower semicontinuous function* $\phi : \mathbb{R}^n \to (-\infty, +\infty]$ *bounded below on* $\mathbb{R}^n$ *and satisfying the following:*
[HJ equation] $h(x, \partial_P \phi(x)) = -1 \forall\, x \neq 0;$
[Boundary condition] $\phi(0) = 0$ *and* $h(0, \partial_P \phi(0)) \geq -1.$
That unique function is $T(\cdot)$.

The proof of this theorem is based upon proximal characterizations of certain monotonicity properties of trajectories related to the inequality forms of the Hamilton–Jacobi Equation (see Sect. 4.7 of [24]). The fact that monotonicity is closely related to the solution of the minimal time problem is already evident in the following elementary assertion: a trajectory x joining α to 0 is optimal if the rate of change of the function $t \mapsto T(x(t))$ is -1 a.e.

The characterization of T given by the theorem can be applied to verify the validity of a general conjecture regarding the time-optimal trajectories from arbitrary initial values. This works as follows: we use the conjecture to calculate the supposedly optimal time $T(\alpha)$ from any initial value α, and then we see whether or not the function T constructed in this way satisfies the conditions of the theorem. If so, then (by uniqueness) T is indeed the minimal time function and the conjecture is verified (the same reasoning as in the preceding section is in fact involved here). Otherwise, the conjecture is necessarily incorrect (but the way in which it fails can provide information on how it needs to be modified; this is another story).

Another way in which the reasoning above is exploited is to discretize the underlying problem as well as the Hamilton–Jacobi Equation for T. This gives rise to the backwards recursion numerical method developed and popularized by Bellman under the name of *dynamic programming*; of course, the approach applies to problems other than minimal time.

With respect to the goal of finding an optimal feedback synthesis for our problem, we have reached the following point in our quest: given that T satisfies the proximal Hamilton–Jacobi Equation $h(x, \partial_P T(x)) = -1$, which can be written in the form

$$\min_{u \in U} \langle \zeta, f(x(t), u) \rangle = -1 \quad \forall \zeta \in \partial_P T(x), \forall\, x \neq 0\,, \quad (9)$$

how does one proceed to construct a feedback $k(x)$ having the property that any trajectory x generated by it via (7) is such that $t \mapsto T(x(t))$ decreases at a unit rate? There will also arise the issue of defining the very sense of (7) when $k(\cdot)$ is discontinuous, as it must be in general.

This is a rather complex question to answer, and it turns out to be a special case of the issue of designing stabilizing feedback (where, instead of T, we employ a *Lyapunov function*). We shall address this below, and return there to the minimal time synthesis, which will be revealed to be a special case of the procedure. We need first to examine the concept of Lyapunov function.

Lyapunov Functions

In this section we consider the standard control system

$$\dot{x}(t) = f(x(t), u(t)) \text{ a.e.}\,, \;\; u(t) \in U \text{ a.e.}\,,$$

under a mild local Lipschitz hypothesis: for every bounded set S in $\mathbb{R}^n$ there exists a constant $K = K(S)$ such that

$$|f(x, u) - f(x', u)| \leq K|x - x'| \;\; \forall x, x' \in S\,, \;\; u \in U\,.$$

We also suppose that $f(0, U)$ is bounded.

A point α is *asymptotically guidable* to the origin if there is a trajectory x satisfying $x(0) = \alpha$ and $\lim_{t\to\infty} x(t) = 0$. When every point has this property, and when additionally the origin has the familiar local stability property known as *Lyapunov stability*, it is said in the literature to be GAC: (open loop) *globally asymptotically controllable* (to 0). A well-known *sufficient* condition for this property is the existence of a smooth (C^1, say) pair of functions

$$V : \mathbb{R}^n \to \mathbb{R}\,, \quad W : \mathbb{R}^n \backslash \{0\} \to \mathbb{R}$$

satisfying the following conditions:

1. **Positive Definiteness:**

$$V(x) > 0 \text{ and } W(x) > 0 \quad \forall x \neq 0, \text{ and } V(0) \geq 0\,.$$

2. **Properness:** The sublevel sets $\{x : V(x) \leq c\}$ are bounded $\forall c$.
3. **Weak Infinitesimal Decrease:**

$$\inf_{u \in U} \langle \nabla V(x), f(x, u) \rangle \;\leq\; -W(x) \quad x \neq 0\,.$$

The last condition asserts that V decreases in *some* available direction. We refer to V as a (weak) *Lyapunov function*; it is also referred to in the literature as a *control Lyapunov function*.

It is a fact, as demonstrated by simple examples (see [17] or [57]), that the existence of a smooth function V with the above properties fails to be a necessary condition for global asymptotic controllability; that is, the familiar converse Lyapunov theorems of Massera, Barbashin and Krasovskii, and Kurzweil (in the setting of a differential equation with no control) do not extend to this weak controllability setting, at least not in smooth terms. This may be a rather general phenomenon, in view of the following result [22], which holds under the additional hypothesis that the sets $f(x, U)$ are closed and convex:

Theorem 3 *If the system admits a C^1 weak Lyapunov function, then it has the following surjectivity property: for every $\epsilon > 0$, there exists $\delta > 0$ such that $f(B(0,\epsilon), U) \supset B(0,\delta)$.*

It is not difficult to check that the nonholonomic integrator system (see Sect. "Introduction") is GAC. Since it fails to have the surjectivity property described in the theorem, it cannot admit a smooth Lyapunov function.

It is natural therefore to seek to weaken the smoothness requirement on V so as to obtain a necessary (and still sufficient) condition for a system to be GAC. This necessitates the use of some construct of nonsmooth analysis to replace the gradient of V that appears in the infinitesimal decrease condition. In this connection we use the proximal subgradient (Sect. "Elements of Nonsmooth Analysis") $\partial_P V(x)$, which requires only that the (extended-valued) function V be lower semicontinuous. In proximal terms, the Weak Infinitesimal Decrease condition becomes

$$\sup_{\zeta \in \partial_P V(x)} \inf_{u \in U} \langle \zeta, f(x,u) \rangle \ \leq \ -W(x) \quad x \neq 0 \, .$$

Note that this last condition is trivially satisfied when x is such that $\partial_P V(x)$ is empty, in particular when $V(x) = +\infty$. (The supremum over the empty set is $-\infty$.) A *general Lyapunov pair* (V, W) refers to extended-valued lower semicontinuous functions $V : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $W : \mathbb{R}^n \backslash \{0\} \to \mathbb{R} \cup \{+\infty\}$ satisfying the positive definiteness and properness conditions above, together with proximal weak infinitesimal decrease.

The following is proved in [24].

Theorem 4 *Let (V, W) be a general Lyapunov pair for the system. Then any $\alpha \in \operatorname{dom} V$ is asymptotically guidable to 0.*

It follows from the theorem that the existence of a lower semicontinuous Lyapunov pair (V, W) with V everywhere finite-valued implies the global asymptotic guidability to 0 of the system. When V is not continuous, this does not imply Lyapunov stability at the origin, however, so it cannot characterize global asymptotic controllability. An early and seminal result due to Sontag [56] considers *continuous* functions V, with the infinitesimal decrease condition expressed in terms of Dini derivates. Here is a version of that result in proximal subdifferential terms:

Theorem 5 *The system is GAC if and only if there exists a continuous Lyapunov pair (V, W).*

There is an advantage to being able to replace continuity in such a result by a stronger regularity property (particularly in connection with using V to design a stabilizing feedback, as we shall see). Of course, we cannot assert smoothness, as pointed out above. In [20] it was shown that certain locally Lipschitz *value functions* give rise to *practical* Lyapunov functions, that is, assuring stable controllability to arbitrary neighborhoods of 0. Building upon this, L. Rifford [49] was able to combine a countable family of such functions in order to construct a global *locally Lipschitz* Lyapunov function. This answered a long-standing open question in the subject. Rifford [51] also went on to show the existence of a *semiconcave* Lyapunov function, a stronger property (familiar in pde's) whose relevance to feedback construction will be seen in the next section.

The role of Lyapunov functions in characterizing various types of convergence to 0 (including guidability in finite time) is discussed in detail in [18].

Stabilizing Feedback

We now address the issue of finding a *stabilizing feedback control* $k(x)$; that is, a function k with values in U such that all solutions of the differential equation

$$\dot{x} = g(x), \ x(0) = \alpha \, , \ \text{ where } \ g(x) : = f(x, k(x)) \quad (10)$$

converge to 0 (for all values of α) in a suitable sense. Here, the origin is supposed to be an equilibrium of the system; to be precise, we take $0 \in U$ and $f(0,0) = 0$. When such a k exists, the system is termed *stabilizable*.

A necessary condition for the system to be stabilizable is that it be GAC. A central question in the subject has been whether this is sufficient as well. An early observation of Sontag and Sussmann [58] showed that the answer is negative if one requires the feedback k to be continuous (which provides the easiest classical way of interpreting the differential equation that appears in (10)). Later, Brockett showed that the nonholonomic integrator fails to admit a continuous stabilizing feedback.

One is therefore led to consider the use of discontinuous feedback, together with the attendant need to define an appropriate solution concept for a differential equation

in which the dynamics fail to be continuous in the state. The best-known solution concept in this regard is that of Filippov; it turns out, however, that the nonholonomic integrator fails to admit a (discontinuous) feedback which stabilizes it in the Filippov sense [32,55]. Clarke, Ledyaev, Sontag and Subbotin [21] gave a positive answer when the (discontinuous) feedbacks are implemented in the *closed-loop system sampling* sense (also referred to as *sample-and-hold*). We proceed now to describe the sample-and-hold implementation of a feedback.

Let $\pi = \{t_i\}_{i\geq 0}$ be a partition of $[0,\infty)$, by which we mean a countable, strictly increasing sequence t_i with $t_0 = 0$ such that $t_i \to +\infty$ as $i \to \infty$. The *diameter* of π, denoted diam (π), is defined as $sup_{i\geq 0}(t_{i+1} - t_i)$. Given an initial condition x_0, the π-trajectory $x(\cdot)$ corresponding to π and an arbitrary feedback law $k\colon \mathbb{R}^n \to U$ is defined in a step-by-step fashion as follows. Between t_0 and t_1, x is a classical solution of the differential equation

$$\dot{x}(t) = f(x(t), k(x_0)) \,, \quad x(0) = x_0 \,, \quad t_0 \leq t \leq t_1 \,.$$

(In the present context we have existence and uniqueness of x, and blow-up cannot occur.) We then set $x_1 := x(t_1)$ and restart the system at $t = t_1$ with control value $k(x_1)$:

$$\dot{x}(t) = f(x(t), k(x_1)) \,, \quad x(t_1) = x_1 \,, \quad t_1 \leq t \leq t_2 \,,$$

and so on in this fashion. The trajectory x that results from this procedure is an actual state trajectory corresponding to a piecewise constant open-loop control; thus it is a physically meaningful one. When results are couched in terms of π-trajectories, the issue of defining a solution concept for discontinuous differential equations is effectively sidestepped. Making the diameter of the partition smaller corresponds to increasing the sampling rate in the implementation.

We remark that the use of possibly discontinuous feedback has arisen in other contexts. In linear time-optimal control, one can find discontinuous feedback syntheses as far back as the classical book of Pontryagin et al. [47]; in these cases the feedback is invariably piecewise constant relative to certain partitions of state space, and solutions either follow the switching surfaces or cross them transversally, so the issue of defining the solution in other than a classical sense does not arise. Somewhat related to this is the approach that defines a multivalued feedback law [6]. In stochastic control, discontinuous feedbacks are the norm, with the solution understood in terms of stochastic differential equations. In a similar vein, in the control of certain linear partial differential equations, discontinuous feedbacks can be interpreted in a distributional sense. These cases are all unrelated to the one under discussion. We remark too that the use of discontinuous pursuit strategies in differential games [41] is well-known, together with examples to show that, in general, it is not possible to achieve the result of a discontinuous optimal strategy to within any tolerance by means of a continuous stategy (thus there can be a positive unbridgeable gap between the performance of continuous and discontinuous feedbacks).

It is natural to say that a feedback $k(x)$ (continuous or not) *stabilizes* the system in the sample-and-hold sense provided that for every initial value x_0, for all $\epsilon > 0$, there exists $\delta > 0$ and $T > 0$ such that whenever the diameter of the partition π is less than δ, then the corresponding π-trajectory x beginning at x_0 satisfies

$$\|x(t)\| \leq \epsilon \quad \forall t \geq T \,.$$

The following theorem is proven in [21].

Theorem 6 *The system is open loop globally asymptotically controllable if and only if there exists a (possibly discontinuous) feedback $k\colon \mathbb{R}^n \to U$ which stabilizes it in the sample-and-hold sense.*

The proof of the theorem used the method of *proximal aiming*, which can be viewed as a geometric version of the Lyapunov technique. We now discuss how to define stabilizing feedbacks if one has in hand a sufficiently regular Lyapunov function.

The Smooth Case

We begin with the case in which a C^1 smooth Lyapunov function exists, where a very natural approach can be used. For $x \neq 0$, we simply define $k(x)$ to be any element $u \in U$ satisfying

$$\langle \nabla V(x), f(x,u) \rangle \leq -W(x)/2 \,.$$

Note that at least one such u does exist, in light of the Infinitesimal Decrease Condition. It is then elementary to prove [18] that the pointwise feedback k described above stabilizes the system in the sample-and-hold sense.

We remark that Rifford [50] has shown that the existence of a smooth Lyapunov pair is equivalent to the existence of a locally Lipschitz one satisfying Infinitesimal Decrease in the sense of generalized gradients (that is, with $\partial_P V$ replaced by $\partial_C V$), and that this in turn is equivalent to the existence of a stabilizing feedback in the Filippov (as well as sample-and-hold) sense.

Semiconcavity

We have seen that a smooth Lyapunov function generates a stabilizing feedback in a simple and natural way.

But since a smooth Lyapunov function does not necessarily exist, we still require a way to handle the general case. It turns out that the smooth and general cases can be treated in a unified fashion through the notion of *semiconcavity*, which is a certain regularity property (not implying smoothness). Rifford has proven that any GAC system admits a semiconcave Lyapunov function; we shall see that this property permits a natural extension of the pointwise definition of a stabilizing feedback that was used in the smooth case.

A function $\phi\colon \mathbb{R}^n \to \mathbb{R}$ is said to be (globally) semiconcave provided that for every ball $B(0,r)$ there exists $\gamma = \gamma(r) \geq 0$ such that the function $x \mapsto \phi(x) - \gamma|x|^2$ is (finite and) concave on $B(0,r)$. (Hence φ is locally the sum of a concave function and a quadratic one.) Observe that any function of class C^2 is semiconcave; also, any semiconcave function is locally Lipschitz, since both concave functions and smooth functions have that property. (There is a local definition of semiconcavity that we omit for present purposes.)

Semiconcavity is an important regularity property in partial differential equations (see for example [9]). The fact that the semiconcavity of a Lyapunov function V turns out to be useful in stabilization is a new observation, and may be counterintuitive: V often has an interpretation in terms of energy, and it may seem more appropriate to seek a *convex* Lyapunov function V. We proceed now to explain why semiconcavity is a highly desirable property, and why a convex V would be of less interest (unless it were smooth, but then it would be semiconcave too).

Recall the ideal case discussed above, in which (for a smooth V) we select a function $k(x)$ such that

$$\langle \nabla V(x), f(x, k(x)) \rangle \leq -W(x)/2 \,.$$

How might this appealing idea be adapted to the case in which V is nonsmooth? We cannot use the proximal subdifferential $\partial_P V(x)$ directly, since it may be empty for some values of x. We are led to consider the limiting subdifferential $\partial_L V(x)$ (see Sect. "Elements of Nonsmooth Analysis"), which is nonempty when V is locally Lipschitz. By passing to the limit, the Weak Infinitesimal Decrease Condition for proximal subgradients implies the following:

$$\inf_{u \in U} \langle f(x,u), \zeta \rangle \leq -W(x) \quad \forall \zeta \in \partial_L V(x) \,,\ \forall x \neq 0 \,.$$

Accordingly, let us consider the following idea: for each $x \neq 0$, choose some element $\zeta \in \partial_L V(x)$, then choose $k(x) \in U$ such that

$$\langle f(x, k(x)), \zeta \rangle \leq -W(x)/2 \,.$$

Does this lead to a stabilizing feedback, when (of course) the discontinuous differential equation is interpreted in the sample-and-hold sense? When V is smooth, the answer is 'yes', as we have seen. But when V is merely locally Lipschitz, a certain 'dithering' phenomenon may arise to prevent k from being stabilizing. However, if V is semiconcave (locally on $\mathbb{R}^n \backslash \{0\}$), this does not occur, and stabilization is guaranteed. This explains the interest in finding a semiconcave Lyapunov function.

When V is a locally Lipschitz Lyapunov function with no additional regularity (neither smooth nor semiconcave), then it can still be used for defining stabilizing feedback, but less directly. It is possible to *regularize* V: to approximate it by a semiconcave function through the process of *inf-convolution* (see [24]). This leads to *practical semiglobal* stabilizing feedbacks, that is, for any $0 < r < R$, we derive a feedback which stabilizes all initial values in $B(0,R)$ to $B(0,r)$ [20].

Optimal Feedback

The strategy described above for defining stabilizing feedbacks via Lyapunov functions can also be applied to construct (nearly) optimal feedbacks as well as stabilizing ones. The key is to use an appropriate value function instead of an arbitrary Lyapunov function. We obtain in this way a unification of optimal control and feedback control, at least at the mathematical level, and as regards feedback design.

To illustrate, consider again the Minimal Time problem, at the point at which we had left it at the end of Sect. "Lyapunov Functions" (thus, unresolved as regards the time-optimal feedback synthesis). As pointed out there, T satisfies the Hamilton–Jacobi Equation: this yields Infinitesimal Decrease in subgradient terms. Consequently, if the Minimal Time function T happens to be finite everywhere as well as smooth or semiconcave, then we can use the same direct definition as above to design a feedback which (in the limiting sample-and-hold sense) produces trajectories along which T decreases at rate 1; that is, which are time-optimal. This of course yields the fastest possible stabilization (and in finite time). In general, however, T may lack such regularity, or (when the controllability to the origin is only asymptotic) not even be finite everywhere. Then it is necessary to apply an approximation (regularization) procedure in order to obtain a variant of T, and use that instead. When T is finite, we can obtain in this way an approximate time-optimal synthesis (to any given tolerance).

The whole approach described here can be carried out for a variety of optimal control contexts [26,45], and also

for finding optimal strategies in *differential games* [25]. It also carries over to problems in which *unilateral state constraints* are imposed: $x(t) \in X$, where X is a given closed set [26,28,29]. The issue of robustness, not discussed here, is particularly important in the presence of discontinuity; see [18,42,57].

Future Directions

A lesson of the past appears to be that nonsmooth analysis is likely to be required whenever linearization is not adequate or is inapplicable. It seems likely therefore to accompany the subject of control theory as it sets out to conquer new nonlinear horizons, in ways that cannot be fully anticipated. Let us nonetheless identify a few directions for future work.

The extensions of most of the results cited above to problems on manifolds, or of tracking, or with partial information (as in adaptive control) remain to be carried out to a great extent. There are a number of currently evolving contexts not discussed above in which nonsmooth analysis is highly likely to play a role, notably *hybrid control*; an example here is provided by *multiprocesses* [8,31]. Distributed control (of pde's) is another area which requires development. There is also considerable work to be done on numerical implementation; in this connection see [36,37].

Bibliography

1. Artstein Z (1983) Stabilization with relaxed controls. Nonlinear Anal TMA 7:1163–1173
2. Astolfi A (1996) Discontinuous control of nonholonomic systems. Syst Control Lett 27:37–45
3. Astolfi A (1998) Discontinuous control of the Brockett integrator. Eur J Control 4:49–63
4. Bardi M, Capuzzo-Dolcetta I (1997) Optimal control and viscosity solutions of Hamilton–Jacobi–Bellman equations. Birkhäuser, Boston
5. Bardi M, Staicu V (1993) The Bellman equation for time-optimal control of noncontrollable nonlinear systems. Acta Appl Math 31:201–223
6. Berkovitz LD (1989) Optimal feedback controls. SIAM J Control Optim 27:991–1006
7. Borwein JM, Zhu QJ (1999) A survey of subdifferential calculus with applications. Nonlinear Anal 38:687–773
8. Caines PE, Clarke FH, Liu X, Vinter RB (2006) A maximum principle for hybrid optimal control problems with pathwise state constraints. Proceedings of the 45th IEEE Conference on Decision and Control, San Diego, 13–15 Dec 2006
9. Cannarsa P, Sinestrari C (2004) Semiconcave Functions, Hamilton–Jacobi Equations, and Optimal Control. Birkhäuser, Boston
10. Clarke F (2005) The maximum principle in optimal control. J Cybern Control 34:709–722
11. Clarke F (2005) Necessary Conditions in Dynamic Optimization. Mem Amer Math Soc 173(816)
12. Clarke FH (1973) Necessary Conditions for Nonsmooth Problems in Optimal Control and the Calculus of Variations. Doctoral thesis, University of Washington
13. Clarke FH (1976) The maximum principle under minimal hypotheses. SIAM J Control Optim 14:1078–1091
14. Clarke FH (1983) Optimization and Nonsmooth Analysis. Wiley-Interscience, New York. Republished as: Classics in Applied Mathematics, vol 5. SIAM, 1990
15. Clarke FH (1986) Perturbed optimal control problems. IEEE Trans Aut Control 31:535–542
16. Clarke FH (1989) Methods of Dynamic and Nonsmooth Optimization. Regional Conference Series in Applied Mathematics, vol 57. SIAM, Philadelphia
17. Clarke FH (2001) Nonsmooth analysis in control theory: a survey. Eur J Control 7:63–78
18. Clarke FH (2004) Lyapunov functions and feedback in nonlinear control. In: de Queiroz MS, Malisoff M, Wolenski P (eds) Optimal Control, Stabilization and Nonsmooth Analysis. Lecture Notes in Control and Information Sciences, vol 301. Springer, New York, pp 267–282
19. Clarke FH, Ledyaev YS (1994) Mean value inequalities in Hilbert space. Trans Amer Math Soc 344:307–324
20. Clarke FH, Ledyaev YS, Rifford L, Stern RJ (2000) Feedback stabilization and Lyapunov functions. SIAM J Control Optim 39:25–48
21. Clarke FH, Ledyaev YS, Sontag ED, Subbotin AI (1997) Asymptotic controllability implies feedback stabilization. IEEE Trans Aut Control 42:1394–1407
22. Clarke FH, Ledyaev YS, Stern RJ (1998) Asymptotic stability and smooth Lyapunov functions. J Differ Equ 149:69–114
23. Clarke FH, Ledyaev YS, Stern RJ, Wolenski PR (1995) Qualitative properties of trajectories of control systems: a survey. J Dyn Control Syst 1:1–48
24. Clarke FH, Ledyaev YS, Stern RJ, Wolenski PR (1998) Nonsmooth Analysis and Control Theory. Graduate Texts in Mathematics, vol 178. Springer, New York
25. Clarke FH, Ledyaev YS, Subbotin AI (1997) The synthesis of universal pursuit strategies in differential games. SIAM J Control Optim 35:552–561
26. Clarke FH, Nour C (2005) Nonconvex duality in optimal control. SIAM J Control Optim 43:2036–2048
27. Clarke FH, Rifford L, Stern RJ (2002) Feedback in state constrained optimal control. ESAIM Control Optim Calc Var 7:97–133
28. Clarke FH, Stern RJ (2005) Hamilton–Jacobi characterization of the state-constrained value. Nonlinear Anal 61:725–734
29. Clarke FH, Stern RJ (2005) Lyapunov and feedback characterizations of state constrained controllability and stabilization. Syst Control Lett 54:747–752
30. Clarke FH, Vinter RB (1984) On the conditions under which the Euler equation or the maximum principle hold. Appl Math Optim 12:73–79
31. Clarke FH, Vinter RB (1989) Applications of optimal multiprocesses. SIAM J Control Optim 27:1048–1071
32. Coron J-M, Rosier L (1994) A relation between continuous time-varying and discontinuous feedback stabilization. J Math Syst Estim Control 4:67–84
33. de Pinho MR (2003) Mixed constrained control problems. J Math Anal Appl 278:293–307

34. Dmitruk AV (1993) Maximum principle for a general optimal control problem with state and regular mixed constraints. Comp Math Model 4:364–377
35. Ferreira MMA (2006) On the regularity of optimal controls for a class of problems with state constraints. Int J Syst Sci 37:495–502
36. Fontes FACC (2001) A general framework to design stabilizing nonlinear model predictive controllers. Syst Control Lett 42:127–143
37. Fontes FACC, Magni L (2003) Min-max model predictive control of nonlinear systems using discontinuous feedbacks. New directions on nonlinear control. IEEE Trans Autom Control 48:1750–1755
38. Hamzi B, Praly L (2001) Ignored input dynamics and a new characterization of control Lyapunov functions. Autom J IFAC 37:831–841
39. Ioffe AD, Tikhomirov V (1974) Theory of Extremal Problems. Nauka, Moscow. English translation: North-Holland, Amsterdam, 1979
40. Kellett CM, Teel AR (2005) On the robustness of $\mathcal{KL}$-stability for difference inclusions: smooth discrete-time Lyapunov functions. SIAM J Control Optim 44:777–800
41. Krasovskii NN, Subbotin AI (1988) Game-Theoretical Control Problems. Springer, New York
42. Ledyaev YS, Sontag ED (1999) A Lyapunov characterization of robust stabilization. Nonlinear Anal 37:813–840
43. Milyutin AA, Osmolovskii NP (1998) Calculus of Variations and Optimal Control. Amer Math Soc, Providence
44. Neustadt LW (1976) Optimization. Princeton University Press, Princeton
45. Nobakhtian S, Stern RJ (2000) Universal near-optimal feedbacks. J Optim Theory Appl 107:89–122
46. Páles Z, Zeidan V (2003) Optimal control problems with set-valued control and state constraints. SIAM J Optim 14:334–358
47. Pontryagin LS, Boltyanskii RV, Gamkrelidze RV, Mischenko EF (1962) The Mathematical Theory of Optimal Processes. Wiley-Interscience, New York
48. Prieur C, Trélat E (2005) Robust optimal stabilization of the Brockett integrator via a hybrid feedback. Math Control Signal Syst 17:201–216
49. Rifford L (2000) Existence of Lipschitz and semiconcave control-Lyapunov functions. SIAM J Control Optim 39:1043–1064
50. Rifford L (2001) On the existence of nonsmooth control-Lyapunov functions in the sense of generalized gradients. ESAIM Control Optim Calc Var 6:539–611
51. Rifford L (2002) Semiconcave control-Lyapunov functions and stabilizing feedbacks. SIAM J Control Optim 41:659–681
52. Rifford L (2003) Singularities of viscosity solutions and the stabilization problem in the plane. Indiana Univ Math J 52:1373–1396
53. Rockafellar RT, Wets R (1998) Variational Analysis. Springer, New York
54. Rodríguez H, Astolfi A, Ortega R (2006) On the construction of static stabilizers and static output trackers for dynamically linearizable systems, related results and applications. Int J Control 79:1523–1537
55. Ryan EP (1994) On Brockett's condition for smooth stabilizability and its necessity in a context of nonsmooth feedback. SIAM J Control Optim 32:1597–1604
56. Sontag ED (1983) A Lyapunov-like characterization of asymptotic controllability. SIAM J Control Optim 21:462–471
57. Sontag ED (1999) Stability and stabilization: discontinuities and the effect of disturbances. In: Clarke FH, Stern RJ (eds) Nonlinear Analysis, Differential Equations and Control, NATO ASI Montreal 1998. Kluwer, Dordrecht, pp 551–598
58. Sontag ED, Sussmann HJ (1980) Remarks on continuous feedback. In: Proc IEEE Conf Decis and Control Albuq. IEEE Publications, Piscataway, pp 916–921
59. Subbotin AI (1995) Generalized Solutions of First-Order PDEs. Birkhäuser, Boston
60. Trélat E (2006) Singular trajectories and subanalyticity in optimal control and Hamilton–Jacobi theory. Rend Semin Mat Univ Politec Torino 64:97–109
61. Vinter RB (2000) Optimal Control. Birkhäuser, Boston
62. Warga J (1972) Optimal Control of Differential and Functional Equations. Academic Press, New York
63. Wolenski PR, Zhuang Y (1998) Proximal analysis and the minimal time function. SIAM J Control Optim 36:1048–1072

Non-standard Analysis, An Invitation to

Wei-Zhe Yang
Department of Mathematics, National Taiwan University, Taipei, Taiwan

Article Outline

Glossary
Definition of the Subject
Introduction
Pre-Robinson Infinitesimals
Hyperreals
General Idea of NSA
Loeb Construction
A Future for Non-standard Analysis?
Bibliography

Glossary

Hyperreal A proper extension of the real numbers, containing infinities and infinitesimal numbers, such that a transfer principle allows first-order theorems in the reals to be extended therein.

Infinity An element in an ordered field which is larger than $1 + 1 + 1 + \cdots + 1$ for any number of 1's.

Infinitesimal An element in an ordered field whose absolute value is less than $1/n$ for any positive integer n.

Non-archimedean Infinitesimal or infinite values exist in an ordered field.

Transfer principle A rule which transforms assertions about standard sets, mappings, etc., into one about in-

ternal sets, mappings. Intuitively, it is an elementary embedding between structures.

Definition of the Subject

Generically speaking, non-standard analysis is any form of mathematics that relies on non-standard models and the transfer principle. However, it is likely safe to say that of the most common meaning for the term "non-standard analysis" is a rigorous formulation of analysis using infinitesimals, as an alternative to the analysis we all know in the ϵ-δ formulation. Abraham Robinson is usually regarded as the first person to do this coherently [12].

It is sometimes claimed that non-standard analysis can achieve more than standard models [2]. This may not be entirely true, but philosophically, non-standard analysis justifies the standard pedagogical process – infinitesimals in Leibniz notation – in which most people learned calculus, and represents the completion of the a key developmental stage of science.

Introduction

Warning 1 *What an (general) undergraduate to a student of mathematics is what a (general) mathematician to a logician.*

This article, written by a (general) mathematician, is aimed at a reader with some mathematical maturity. There are different constructions of non-standard analysis, and we have chosen to start with the "ultrapower" construction of the semantic kind, which is not as difficult to comprehend as the first presentation.

Differential = Infinitesimal Difference?

Suppose a freshman is questioned on the derivative $\frac{\mathrm{d}y}{\mathrm{d}x}$ for the cubic function $y = x^3$. Then after the difference-quotient is found,

$$\frac{\Delta y}{\Delta x} = 3x^2 + 3x(\Delta x) + (\Delta x)^2\,;$$

he just says "substituting by $\Delta x = 0$", $\frac{\mathrm{d}y}{\mathrm{d}x} = 3x^2$. And the instructors often have trouble explaining the difference between "Δx set to 0" and "$\lim_{\Delta x \to 0}$".

It is estimated that, among the mathematical-literates (i. e., those who learnt calculus), less than 10% are competent with "ϵ-δ-logy". But, of course, they are assured that what they learnt is no nonsense. The calculus we learn is from Leibniz, not Seki Kowa (even if the textbook is in Japanese), and not even Newton, who must have the priority. And the discoveries of calculus and of rigorous analysis (chiefly of Cauchy), are some 150 years apart.

In essence, the success of Leibniz is due to his elegant and mnemonic way of notations, centered around the notion of differential. When a freshman understands that "the differential of y", written as $\mathrm{d}y$, is "the infinitesimal difference of y", he deals with calculus problems just like doing routine high school algebra. Indeed calculus as envisaged by Leibniz, is a computation scheme in a number system, denoted by ${}^\star\mathbb{R}$ below. This system of "hyperreals" must include, besides the usual real numbers, infinitesimals. Leibniz conceived of the notion of infinitesimals out of necessity, because he was pre-Cauchy. He needed the infinitesimals (as differentials) in the calculus.

But from the very beginning many people, notably Bishop G. Berkeley, "discover impossibilities and contradictions". Infinitesimals cannot be, simultaneously, zero and non-zero. If they are zero, the division is illegal; if they are not, when $y = x^3$, $\frac{\mathrm{d}y}{\mathrm{d}x} = 3x^2$ is wrong because $3x\mathrm{d}x + (\mathrm{d}x)^2$ does not just evaporate away. Of course this is a contradiction that Leibniz, who was a first-rate philosopher-logician, could not fail to discover. Thus the calculus as a science was, for many people before Cauchy, is only recognized de facto and not de jure, just like the nations of Taiwan and Kosovo.

Structure of This Article

In the next Sect. "Pre-Robinson Infinitesimals", we will tell you how the people tried to handle infinitesimals before [12]. In Sect. "Hyperreals", we define the hyperreal field using the ultrapower construction. Section "General Idea of NSA" builds a small framework of non-standard analysis using hyperreals, and Sect. "Loeb Construction" introduces the Loeb construction for Wiener processes in NSA.

Pre-Robinson Infinitesimals

Herein we describe the work prior to Robinson (despite say the work of [13]), who is credited with being the creator of the field.

Recap: Naive Set Notations

We assume the concept of a set without recourse to the axioms of set theory.

Definition 1 (Binary Boolean Set-Operations) For two sets A and B, there are these Boolean combinations of them:

$$A \cup B := \{x\colon x \in A \text{ or } x \in B\},$$
$$A \cap B := \{x\colon x \in A \text{ and } x \in B\},$$
$$A \setminus B := \{x\colon x \in A \text{ and } x \notin B\}\,.$$

Non-standard Analysis, An Invitation to, Table 1
Notations of this Chapter

Notation	Meaning
iff	if and only if
$\neg, \wedge, \vee, \Rightarrow$	"negation of ", "and ", "or ", "imply ".
$\forall, \exists$	"for all ", "exists "
$:=$	"defined to be equal to "
$\mathbb{R}, \mathbb{C}, \mathbb{Z}$	set of reals, complex numbers, integers
$\mathbb{N}$	set of natural numbers (zero excluded)
$(a \ldots b), [a \ldots b]$	open and closed intervals
$(a \ldots b], [a \ldots b)$	semi-open intervals
$\{ \quad \}$	the braces holding a set
$\langle \quad \rangle$	angles holding a sequence
$\cup, \sqcup, \cap, \setminus$	set union, disjoint union, intersection, set difference
$\mathbb{P}(A)$	power set (set of all subsets including A and $\emptyset$)
$\mathbb{R}^{\mathbb{N}}_{\beta}$	set of all bounded sequences
$\mathbb{R}^{\mathbb{N}}_{c}$	set of all convergent sequences
$\mathbb{R}^{\mathbb{N}}_{0}$	set of all sequences converging to zero(=vanishing)
$\mathbb{R}^{\mathbb{N}}_{\kappa}$	set of all eventually -zero(=nullifying) sequences
$\mathbb{R}^{\times}$	$= \mathbb{R} \setminus \{0\}$, non-zero reals
$\mathbb{R}_{+}$	$[0 \ldots \infty)$, non -negative real number
$\dot{\mathbb{R}}$	$= \mathbb{R} \sqcup \{-\infty, +\infty\}$, $\mathbb{R}$ with signed infinities
$\dot{\mathbb{R}}$	$= \mathbb{R} \sqcup \{\infty\}$, $\mathbb{R}$ with unsigned infinity
${}^{\star}\mathbb{R}$	(one possible) hyperreal number system
$\mathfrak{U}$	ultra -filter, also and denote (as superscript) "ultrapower construction "
σ	superscript denotes the set of standard elements
$\pi_{\mathfrak{U}}$	denotes the projection for the ultrapower construction $\mathfrak{U}$

Definition 2 (Function) Let A and B be two non-empty sets. If for any $x \in A$, there is correspondingly assigned one $y \in B$, and written as $f(x)$, then we call this assignment f a function. Its domain is A, codomain B. The notation is $f\colon A \mapsto B$, or $f\colon x \in A \mapsto f(x) \in B$. Also, for $C \subset A$, or for $D \subset B$, we write

$$f[C] := \{f(x)\colon x \in C\}\,.$$
$$f^{-1}[D] := \{x \in A\colon f(x) \in D\}\,.$$

If $f[A] = B$, then the function f is surjective. If $f(x) = f(u)$ implies $x = u$, then the function f is injective. when the function is both injective and surjective, it is called bijective. Sometimes the codomain of a function is just understood, but suppressed, and the function is then written as an "indexed family" $\langle f_j\colon j \in I\rangle$. In this case, we write f_j instead of $f(j)$, and I, the domain of definition, here I, is called the indexing-set, and is often omitted when no confusion is likely. When the indexing set is $\mathbb{N}$ (and in a few other situations) the indexed family is called a **sequence**.

Definition 3 (Cartesian-Product, Couple, Triple, and Sequence) For two non-empty sets A, B, their Cartesian product is defined as

$$A \times B := \{\langle a, b\rangle\colon a \in A\,, \text{ and } b \in B\}\,.$$

Here "a couple", or "an ordered pair", is denoted by $\langle a, b\rangle$. An ordered n-tuple is analogously denoted like $\langle x_1, x_2, \cdots, x_n\rangle$.

Definition 4 (General Operations among a Family of Sets) If we have an indexed family of sets $(X_j\colon j \in I)$, then we define:

$$\bigcup_{j\in I} X_j := \{x\colon x \in X_j\,, \text{ for some } j \in I\}\,;$$
$$\bigcap_{j\in I} X_j := \{x\colon x \in X_j\,, \text{ for each } j \in I\}\,;$$
$$\prod_{j\in I} X_j := \{g\colon g \text{ is a function, with } g(j) \in X_j\,, \text{ for all } j \in I\}\,;$$

Definition 5 (Quotient-Set and Equivalence Relation) This is a review of the most difficult notion in naive set theory. In studying a non-empty set A, there often occur an

"equivalence relation", denoted by, say, "$\sim$". This means we write "$x \sim y$" to mean that these two elements x, y of A are "of the same kind", "similar", or "equivalent", referring to something of our concern. If we collect all similar members into "bunches", each bunch is called an "equivalence class" (of this relation $\sim$). The set of (bunches=) equivalence-classes is called the quotient set of A with respect to the relation $\sim$, and written as $\frac{A}{\sim}$. This process of forming the quotient set $\frac{A}{\sim}$ from the set A, is usually described as "identifying equivalent members". Formally speaking: a binary relation $\sim$ on A is called an equivalence relation iff it has the following three properties:

r° (Reflexivity): for any $x \in A, x \sim x$.

s° (Symmetry): when $\{x, y\} \subset A$, and $x \sim y$, then $y \sim x$.

t° (Transitivity): when $\{x, y, z\} \subset A$, and $x \sim y$, and also $y \sim z$, then $x \sim z$.

For $x \in A$, its equivalence class (modulo $\sim$), is then

$$\pi_{\sim}(x) := \{y \in A \colon x \sim y\}\,.$$

(It is a non-empty subset of A). The quotient set is then:

$$\frac{A}{\sim} := \{\pi_{\sim}(x) \colon x \in A\}\,.$$

Ordered Field Extension

The relation between the system $\mathbb{R}$ of reals and the system ${}^{\star}\mathbb{R}$ of hyper-reals may be announced thus: First of all, ${}^{\star}\mathbb{R}$ should be a proper extension of the ordered-field $\mathbb{R}$. Secondly, one should also be able to extend, from $\mathbb{R}$ to ${}^{\star}\mathbb{R}$, the usual functions exp, cos, sin, etc. This is one of the topics of this article, and we explain the meaning of "extension" in this subsection.

Definition 6 (Commutative-Semigroup) Let us consider the arithmetic operation $+$ in $\mathbb{R}$. Though it is a two-variable function, just like, e. g., $f \colon f(x, y) = \sin(x - y) * (x^3 - y^2)$, it is written "infix". (So that $x + y$ is "really" $+(x, y)$). There are the the noteworthy laws of associativity, and of commutativity:

$$\forall x\,, \forall y\,, \forall z\,,$$
$$(x + y) + z = (x + (y + z)\,; \quad x + y = y + x\,.$$

(We mathematics student read $\forall$ as "for all"). By abstraction, we call a binary operation (= two variable function) on a non-empty set A, a "commutative semigroup structure" on this set. This system, the set A together with the commutative semigroup structure, is called a commutative semigroup. We realize that: in $\mathbb{R}$, multiplication is another commutative semigroup structure.

Definition 7 (Commutative-Group) Now in $\mathbb{R}$, 0 is distinguished by the "unitality law":

$$\forall x\,, x + 0 = 0 + x = x\,.$$

There is then a negation function, mapping x into its negative $-x$, and the reciprocal law holds:

$$x + (-x) = 0\,; \quad (-x) + x = 0\,.$$

By abstraction, we call a triple of a two variable function, a distinguished element, and a one-variable function, all on a non-empty set A, satisfying all the laws mentioned above, as a "commutative group structure". And the set, together with this commutative group structure, is called a commutative group. We realize that: $(\mathbb{R}, (+, 0, -))$ is a commutative group. As to the multiplication, the distinguished unity is 1, but the reciprocation fails at precisely 0. If we all the functions are restricted to the zero-deleted set

$$\mathbb{R}^{\times} := \mathbb{R} \setminus \{0\}\,,$$

then we do have a commutative group structure on it.

Definition 8 (Field) Suppose F is a set of more than one elements. Let the triple $(+, 0, -)$ be a commutative group structure, and $(*, 1, ^{-1})$ be a commutative group structure on $F^{\times} = F \setminus \{0\}$. And suppose the distributive law holds:

$$(\forall x\,, \forall y\,, \forall z)\,, \quad x * (y + z) = x * y + x * z\,.$$

Then this couple of triple is called a "field structure" on F. (Of course, there are, altogether, six items in the structure: two distinguished elements (="null-variable functions"), two (single variable) functions, two two-variable functions). And the set F, together with this field structure, is called a field.

Definition 9 (Total-Ordering, and Ordered Field) A total-ordering on a set T is a binary relation on T, which satisfies the transitive law and the trichotomy law. For example, in $\mathbb{R}$, the relation "less than" is a total-ordering: the transitive law says that: if $a < b$, and $b < c$, then $a < c$, while the trichotomy law requires that: for any two members a, b, exactly one sentence among the following three is true: $a < b$, $a = b$, $b < a$. An ordered-field structure on a set K is a combination of a field structure, and an total-ordering $<$, with the additional requirement of compatibility law:

$$((0 < x) \wedge (0 < y)) \Rightarrow ((0 < x + y) \wedge (0 < x * y))\,.$$

$\wedge$ is read "and", and $\Rightarrow$ is read "imply". K with this structure, is then an ordered-field. Obviously the set of all rational numbers $\mathbb{Q}$, together with the natural structure, (which

is called "inherited from $\mathbb{R}$"), is also an ordered-field in this sense.

Definition 10 (Mathematical System, Inheritance and Extension) Returning to the instance of the set $\mathbb{R}$. We may consider the following substructure of five items: the three items of addition commutative-group structure, and the multiplication unital commutative semi-group structure, keeping in mind the distribution law binding them. This is a unital commutative-ring structure. A set K together with a unital commutative-ring structure on it is called a unital commutative-ring. And as is obvious from above discussion, by forgetting the two items in any field structure, we have a unital commutative-ring structure. So the latter is weakened from the former, while the former is strengthened from the latter. Essentially any one particular kind of structure is called a "category". So within a category one system may be the extension of the other, or the other way round, the latter is a sub-system of the former. For example, the rational ordered field $\mathbb{Q}$ is a sub-system of $\mathbb{R}$, and the real ordered field $\mathbb{R}$ is an extension of $\mathbb{Q}$. This extension is proper, because $\mathbb{R}$ is strictly bigger: there are irrational numbers. Now for example, the set $\mathbb{Z}$ of all integers is also a unital commutative-ring, when its structure is inherited (with the obvious meaning, or technically "restricted") from $\mathbb{Q}$. That is, $\mathbb{Z}$ is a sub-unital commutative-ring of $\mathbb{Q}$. But it is not a subsystem of $\mathbb{Q}$ as a field.

Lemma 1 (Quotient Algebra) *Let us mangle usual terminology a little and abbreviate "unital commutative-ring" as "algebra", just for convenience in this discussion. If A is an algebra, then a subset B is called an ideal, if the following condition is met: for arbitrary $x \in A$, $y \in B$, $z \in B$,*

$$y+z \in B\,, \quad x*y \in B\,, \quad (\textit{which is same as}) \quad y*x \in B\,.$$

Then we define a (modulo B) relation $\sim_B$ in A, by:

$$(x \sim_B y) \quad \text{iff} \quad (x-y) \in B\,.$$

And this is an equivalence relation. The equivalence-class of $x \in A$ is then:

$$x + B := \{x + y\colon y \in B\}\,.$$

The quotient set $\frac{A}{\sim_B}$ will be written as $A \mod B$. In it naturally we define the operations by:

$$\begin{aligned}(x+B)+(y+B) &:= (x+y)+B\,,\\ -(x+B) &:= (-x)+B\,,\\ (x+B)*(y+B) &:= (x*y)+B\,.\end{aligned}$$

(And endow it with the natural zero $0+B$, the unit $1+B$). The simple result is: these indeed form an algebra structure.

Remark 1 (Other Category) A one variable function is a particular kind of binary relation. And a two-variable function is a particular kind of trinary relation. Indeed a null-variable function (=a distinguished element), is a particular kind of unary relation. Usually structures consisting of these kind of functions, (bound by algebraic laws), are called "algebraic". In topology, we study the topological structure, and in measure theory, we need the measurable structure, before strengthening it to the measure structure. And the set $\mathbb{R}$ has on it a natural topological structure and a natural measure structure.

Classical Augmented Infinities

Definition 11 (Leibniz Program) Extend the system $\mathbb{R}$ of reals to the system ${}^\star\mathbb{R}$ of hyperreals, so as to contain infinitesimals and also infinities.

This question should be considered in the perspective of "structure of a system" explained in the last subsection. Indeed the NSA of Abraham Robinson is so powerful as to consider the super-structure which is all-encompassing. As far as the size or magnitude (=absolute-value) is concerned, hyperreals should be classified in three categories: infinitesimal, ordinary, and infinite. The reciprocal of an infinitesimal (assumed non-zero), must be infinite (=unlimited), and vice versa. While ordinary hyperreals are neither infinitesimal nor infinite. They are medium-sized, together with their reciprocal. The infinite hyperreals and the infinitesimals (zero excepted), are all non-standard real numbers. The non-zero standard real numbers are all included in the medium-sized hyperreals. Now in calculus, and especially at the beginning pages of complex variables, we sometimes talk about infinities. It turns out that they are standard infinities, and does not belong to the Leibnizian ${}^\star\mathbb{R}$ of hyperreals. Let us first try to add one or two points to $\mathbb{R}$, disregarding the infinitesimals for the present. Actually there are precisely three infinities, of two kinds: two signed infinities, and one unsigned infinity.

Definition 12 (Signed Infinities) We denote by $+\infty$ the "positive infinity", and by $-\infty$ the "negative infinity". They are added to the totally ordered set $\mathbb{R}$, so as to be, respectively, the maximum and the minimum of this order-extension of $\mathbb{R}$.

$$\ddot{\mathbb{R}} = \mathbb{R} \sqcup \{-\infty, +\infty\}\,.$$

(We use $\sqcup$ here to emphasize that it is a disjoint-union.) So there is an order-isomorphism:

$$x \in \ddot{\mathbb{R}} \mapsto \frac{2}{\pi}\arctan(x) \in [-1\ldots 1]\,, \quad (\text{a closed interval})$$

which is also an homeomorphism (=topological isomorphism). And this compact space $\ddot{\mathbb{R}}$ is therefore called the two-point compactification of $\mathbb{R}$.

Definition 13 (Supremum, Infimum) In $\ddot{\mathbb{R}}$, any (non-empty) subset A has a supremum, and also an infimum, denoted respectively as $\sup A$, $\inf A$.

We recall that "b is an upper-bound of A" means that $\forall x \in A$, $x \leq b$. And $b = \sup A$ means: b is the least among all upper-bounds of A. In measure theory, most often the values are in the positive half of $\ddot{\mathbb{R}}$, that is, the set $[0 \cdots + \infty] = \dot{\mathbb{R}}_+$. Finally there is one unsigned infinity, which is the only infinity of the one-point compactification

$$\dot{\mathbb{R}} = \mathbb{R} \sqcup \{\dot{\infty}\},$$

which is obtained by merging the two signed infinities $\pm\infty$ of $\ddot{\mathbb{R}}$ into one point. In $\dot{\mathbb{R}}$, we have, for example:

$$-\infty < -921\,, -921 < -228\,, -228 < +\infty\,;$$

If the ordering were compatible with the merging, then, we would get

$$\infty < -921\,, -921 < -228\,, -228 < \infty\,;$$

and the transitive law is violated! This excludes the possibility of ordering the set $\dot{\mathbb{R}}$. The space $\dot{\mathbb{R}}$ is homeomorphic to the unit circle, e. g. by the (one-dimensional!) stereographic projection

$$x \in \dot{\mathbb{R}} \mapsto (\cos\theta, \sin\theta)\,, \quad \text{where} \quad \theta = 2 * \arctan x - \frac{2}{\pi}\,.$$

Note: Here we use an unusual notation to distinguish $\dot{\infty} \in \dot{\mathbb{R}}$ from $+\infty \in \ddot{\mathbb{R}}$. Actually these two infinities, one unsigned, the other signed, are never coexistent, and usually they both are denoted by the same simple ∞. The deficiency of these systems is that: arithmetic operations involving the infinities are only partially defined in the system, so that there are many undefined operations, so-called "undetermined form":

$$(+\infty) + (-\infty)\,; \quad (\pm\infty) * 0\,; \quad (\pm\infty) \div (\pm\infty)\,;$$

Because of these no-definitions, the associative law is greatly compromised. So the use of these infinities is always just symbolic.

Note 1 When $\arctan(x)$ is defined for $x \in \ddot{\mathbb{R}}$, it is valued in $[\frac{-\pi}{2} \ldots \frac{\pi}{2}]$, while for $x \in \dot{\mathbb{R}}$, it is valued in $\mathbb{R} \mod (\pi)$. Also $\lim_{x\to 0}(7-2x)/x^4 = +\infty \in \ddot{\mathbb{R}}$, while $\lim_{x\to 0}(7-2x)/x^3 = \dot{\infty} \in \dot{\mathbb{R}}$. Such distinctions are sometimes useful or necessary.

Non-Archimedean Infinitesimal

Physically speaking, positive real numbers are used for measurements. The possibility of measurement is based on this simple fact: any length λ however long, may be measured against any length ϵ however short. If, for example, $\lambda = 1\,\text{m}$, and $\epsilon = 3.4\,\text{mm}$, then $0 < \lambda - n * \epsilon < \epsilon$, with $n = \text{floor}(\frac{\lambda}{\epsilon}) = 294$, so that λ is a little bit more than n units of ϵ. This basic fact is the

Axiom 1 (of Archimedes) *For any $\lambda \in \mathbb{R}_+$, and any $\epsilon \in \mathbb{R}_+$, there is an $n \in \mathbb{N}$, such that $n * \epsilon > \lambda$.*

It is precisely this axiom which denies the existence of the infinitesimals: a positive infinitesimal would not be usable as a unit for measurement.

Definition 14 (Infinitesimal) In an ordered field F, an element ϵ is called infinitesimal, if for any $n \in \mathbb{N}$, $n * |\epsilon| < 1$. In other words, no matter how many terms are there, the sum $n * \epsilon = \epsilon + \epsilon + \epsilon + \cdots + \epsilon$ (n terms) < 1.

And in an Archimedean ordered field, the only infinitesimal is zero. When people become acquainted with the concept of limit, they know that "infinitesimal" should not be interpreted statically. Rather, it is the "vanishing behavior" of, e. g. the function $(\sin^3 x)/x^2$, when $x \to 0$, that is to be identified with "infinitesimal". Along this direction there appeared several attempts to find (=construct) the system ${}^\star\mathbb{R}$ promised by Leibniz. We now describe a good and very natural try, just years before the discovery of Robinson, by Schmieden–Laugwitz.

Definition 15 (Guiding Idea) A nullifying sequence is considered as zero, a vanishing sequence is considered as an infinitesimal, and a constant sequence is identified with the number itself.

There may be better terminology, but here a sequence $\langle x_n \rangle$ is called nullifying, if after a finite number of terms, the rest of them are all zero. And a sequence $\langle x_n \rangle$ is called vanishing, if it is converging to zero.

Definition 16 (Confusing!) Limit and limiting-point recall the most confusing duo-definitions for a sophomore: "b is the limit of a sequence $\langle a_n \rangle$", and "c is a limiting point of a sequence $\langle a_n \rangle$". For any $\epsilon > 0$, there exists an $m \in \mathbb{N}$, such that, $\forall n \geq m$, $|a_n - b| < \epsilon$. For any $\epsilon > 0$, and for any $m \in \mathbb{N}$, there exists an $n \in \mathbb{N}$, such that $n > m$, and $|a_n - c| < \epsilon$. (We support Kelley's advocation, replacing "limiting" by "cluster".)

Definition 17 (Eventual Set, Frequent Set) A subset A of $\mathbb{N}$ is called "eventual", if it is "cofinite" (=complement of a finite subset). That is: $\mathbb{N} \setminus A$ is finite. It is called "frequent", if it is infinite. If a property $P(i)$ holds true for i be-

longing to an eventual set (the mention of which is unnecessary and omitted intentionally), we say $P(i)$ holds true "for eventual i". Similar meaning is applied to the phrase "for frequent i". With this kinds of abuse of the English language, we may rephrase the definitions above as: For a sequence $a := \langle a_n : n \in \mathbb{N}\rangle$, $b \in \mathbb{R}$ is a limit of a iff, for any $\epsilon > 0$, $a_i \in (b - \epsilon \ldots b + \epsilon)$ for eventual i; c is a limiting-point of a iff, for any $\epsilon > 0$, $a_i \in (c - \epsilon \ldots c + \epsilon)$ for frequent i.

Remark 2 See p. 65 in [7], for the definitions of "eventually", "frequently", for a general directed set of indices. Here the (naturally) directed set $\mathbb{N}$ is very special, because it is the least well-ordered infinite set.

Notation 1 (Sequence Spaces) *We write $x \in \mathbb{R}^{\mathbb{N}}$ to mean: $x = \langle x_n\rangle$ is a sequence of real numbers. (This is in accordance to our notating the set of mappings). Here are some subsets of $\mathbb{R}^{\mathbb{N}}$, arranged decreasingly:*

$\mathbb{R}^{\mathbb{N}}_{\boldsymbol{\beta}}$ *is the set of all bounded sequences;*

$\mathbb{R}^{\mathbb{N}}_{\boldsymbol{c}}$ *is the set of all convergent sequences;*

$\mathbb{R}^{\mathbb{N}}_{\mathbf{0}}$ *is the set of all sequences converging to zero (=vanishing);*

$\mathbb{R}^{\mathbb{N}}_{\boldsymbol{\kappa}}$ *is the set of all eventually-zero (=nullifying) sequences.*

Definition 18 (Coordinatewise Operations on Sequence Spaces)

$$\langle x_n\rangle + \langle y_n\rangle := \langle x_n + y_n\rangle\,;$$
$$\langle x_n\rangle - \langle y_n\rangle := \langle x_n - y_n\rangle\,;$$
$$\langle x_n\rangle * \langle y_n\rangle := \langle x_n * y_n\rangle\,.$$

Lemma 2 *The system $\mathbb{R}^{\mathbb{N}}$ is a commutative unital ring, a real vector space, and thence a commutative algebra (with identity). The following inclusions are subalgebra embeddings:*

$$\mathbb{R}^{\mathbb{N}}_{c} \subset \mathbb{R}^{\mathbb{N}}_{\beta} \subset \mathbb{R}^{\mathbb{N}}\,,$$

Now according to the guiding idea, a number $\lambda \in \mathbb{R}$ is identified with the sequence with all terms $= \lambda$. In this sense $\mathbb{R}^{\mathbb{N}}_{\beta}$, (and a fortiori $\mathbb{R}^{\mathbb{N}}$), is an extension of $\mathbb{R}$, considered as an algebra. But this is not field-extension, because

Example 1 Let $1_E \in \mathbb{R}^{\mathbb{N}}_{\beta}$ and $1_O \in \mathbb{R}^{\mathbb{N}}_{\beta}$ be the indicator sequences of the sets of even and odd indices, respectively. Then they are complementary idempotents:

$$1_E * 1_E = 1_E\,; \quad 1_O * 1_O = 1_O\,; \quad 1_E * 1_O = 0\,; \quad 1_E + 1_O = 1\,.$$

Since the two elements are neither 1, nor 0, of the algebra $\mathbb{R}^{\mathbb{N}}_{\beta}$, the latter cannot be a field, because even an 8th grader knows that in a field F, the only idempotent element $x = x^2$ is $x = 0$ (the zero) or $x = 1$ (the identity).

Definition 19 (Extension of Function) Consider these two functions cos and sin. For a sequence $x = \langle x_n\rangle \in \mathbb{R}^{\mathbb{N}}$, x is a "number" in a generalized sense, and we would like to define $\cos(x), \sin(x)$. If x is a constant sequence: $x_n = a, \forall n$, then $\cos(x), \sin(x)$, should be the sequences with constant terms of $\cos(a), \sin(a)$ correspondingly. So we see that a coordinatewise definition is almost forced on us: If $f : \mathbb{R} \mapsto \mathbb{R}$ is everywhere defined, then we define:

$$\overline{f}\langle x_n\rangle := \langle f(x_n)\rangle\,.$$

So far as it is defined, it works fine. As an example: the addition-formula

$$\overline{\sin}(\langle x_n\rangle + \langle y_n\rangle) = \overline{\sin}\langle x_n\rangle * \overline{\cos}\langle y_n\rangle + \overline{\sin}\langle y_n\rangle * \overline{\cos}\langle x_n\rangle$$

would simply mean that, for each index n,

$$\sin(x_n + y_n) = \sin(x_n) * \cos(y_n) + \sin(y_n) * \cos(x_n)\,.$$

Theorem 1 (A Transfer Principle: Restricted-Form) *If functions $f, g, \ldots$, are all defined on $\mathbb{R}$, and they has a formula connecting them, then this formula still holds for the extended functions on the whole sequence space $\mathbb{R}^{\mathbb{N}}$.*

We remark that coordinatewise extension is equally applicable to two-variable function, so long as it is defined everywhere. The Transfer Principle remains valid. And indeed the principle has nothing to do with the differentiability, continuity, etc., of the functions involved. This is pure naive-set theory!

Remark 3 If bounded sequences are called "moderate numbers", then only bounded functions will map "moderate number" to "moderate number".

Definition 20 (Ordering) For two elements $x = \langle x_n\rangle$, $y = \langle y_n\rangle$ of $\mathbb{R}^{\mathbb{N}}$, we define:

$$(x \le y)\,, \quad \text{iff} \quad (\forall n, x_n \le y_n)\,.$$

We write $x < y$, iff $x \le y$, and $x \ne y$. (This means the following: $x_n \le y_n$, for all $n \in \mathbb{N}$, but there is an index n, (at least one, and only one is enough!) such that $x_n < y_n$.

The binary relation $\le$ in $\mathbb{R}^{\mathbb{N}}$ is called a partial ordering, because it is transitive, and antisymmetric: If $x \le y$, and also $y \le x$, then $x = y$. This partial ordering is not a total-ordering, because some pair of members, for example 1_E

and 1_O, may be 'incomparable', as neither $1_E \geq 1_O$ nor $1_O \geq 1_E$ is true.

Lemma 3 (Nullifying and Vanishing Sequences) *The set $\mathbb{R}^{\mathbb{N}}_0$ is an ideal of $\mathbb{R}^{\mathbb{N}} \supset \mathbb{R}^{\mathbb{N}}_b \supset \mathbb{R}^{\mathbb{N}}_c$. The set $\mathbb{R}^{\mathbb{N}}_\kappa$ is an ideal of $\mathbb{R}^{\mathbb{N}}_b \supset \mathbb{R}^{\mathbb{N}}_c$. Doing modulo $\mathbb{R}^{\mathbb{N}}_0$ means that two sequences with vanishing difference are considered the same. Or, modulo $\mathbb{R}^{\mathbb{N}}_0$ is a magnifying-glass not powerful enough to distinguish a vanishing sequence from zero sequence. And you don't have non-trivial infinitesimals in the system. Also,*

Lemma 4 *The quotient algebra $\mathbb{R}^{\mathbb{N}}_c \mod \mathbb{R}^{\mathbb{N}}_0$ is (isomorphic to) $\mathbb{R}$.*

The isomorphism identifies any convergent sequence $\langle x_n \rangle$ with its limit $\lim_{n\to\infty} x_n$. The Lemma says that $\lim_{n\to\infty} x_n * y_n = \lim_{n\to\infty} x_n * \lim_{n\to\infty} y_n$, and the like. (Just the first theorems of our calculus text-book.)

Definition 21 (Quotient by Eventuality) We turn to the quotient algebra $\mathbb{R}^{\mathbb{N}}_{\mathfrak{E}} = \mathbb{R}^{\mathbb{N}} \mod \mathbb{R}^{\mathbb{N}}_\kappa$, where two sequences $x \in \mathbb{R}^{\mathbb{N}}$ and $y \in \mathbb{R}^{\mathbb{N}}$ are equivalent mod $\mathbb{R}^{\mathbb{N}}_\kappa$, iff $x_n = y_n$ for eventual n. As we denote by $\mathfrak{E}$ the set of all eventual set of indices, we will also write $x \overset{\mathfrak{E}}{\sim} y$ or $x =_{\mathfrak{E}} y$ for this equivalence, and the equivalence class of x is written $\pi\mathbb{E}(x) \in \mathbb{R}^{\mathbb{N}}_{\mathfrak{E}}$. The system $\mathbb{R}^{\mathbb{N}}_{\mathfrak{E}}$ is a gentle modification of $\mathbb{R}^{\mathbb{N}}$. It has essentially the same good or bad properties of $\mathbb{R}^{\mathbb{N}}$. Because, in essence, the idea of $\mod \mathfrak{E}$ is simply: When studying a sequence, never worry about its few exceptional terms! So we also can extend all the everywhere-defined functions for these "$\mathfrak{E}$-numbers". And the transfer principle still applies! Something better comes out from this modification $\mod \mathfrak{E}$: ordering.

Definition 22 (Ordering) For two elements $x = \langle x_n \rangle$, $y = \langle y_n \rangle$ of $\mathbb{R}^{\mathbb{N}}$, we write $x \leq_{\mathfrak{E}} y$, and also $\pi\mathbb{E}(x) \leq \pi\mathbb{E}(y)$, iff

$$x_n \leq y_n \quad \text{for eventual} \quad n\,.$$

Naturally we write $x <_{\mathfrak{E}} y$, and also $\pi_{\mathfrak{E}}(x) < \pi_{\mathfrak{E}}(y)$, iff $x \leq_{\mathfrak{E}} y$, and $x \neq_{\mathfrak{E}} y$. This means: $x_n \leq y_n$, for eventual n, but $x_n < y_n$, for frequent n.

Note in the last sentence a subtle difference with the situation in $\mathbb{R}^{\mathbb{N}}$, where "for just one n" will do. Again this partial ordering is not a total-ordering, as for example $\pi\mathbb{E}(1_E)$ and $\pi_{\mathfrak{E}}(1_O)$ are still "incomparable". But the good news is there are "infinitesimals"!

Definition 23 Let $x \in \mathbb{R}^{\mathbb{N}}$. If for all $m \in \mathbb{N}$, $\pi_{\mathfrak{E}}(x) \leq \frac{1}{m}$, then $\pi_{\mathfrak{E}}(x)$ is called an infinitesimal in $\mathbb{R}^{\mathbb{N}}_{\mathfrak{E}}$.

Lemma 5 (Infinitesimals) *For $x \in \mathbb{R}^{\mathbb{N}}$, $\pi_{\mathfrak{E}}(x)$ is an infinitesimal $\mathbb{R}^{\mathbb{N}}_{\mathfrak{E}}$, iff $x \in \mathbb{R}^{\mathbb{N}}_0$.*

If infinitesimals are just (the equivalence classes of) vanishing sequences, of course sequences "diverging to infinity" are then the infinities!

Hyperreals

The Construction of Ultimate-Reals

Let us review the system $\mathbb{R}^{\mathbb{N}}_{\mathfrak{E}}$, especially the compatibility of the equivalence relation $\overset{\mathfrak{E}}{\sim}$ with multiplication:

$$((x \overset{\mathfrak{E}}{\sim} y) \text{ and } (u \overset{\mathfrak{E}}{\sim} v)) \Rightarrow (x * u) \overset{\mathfrak{E}}{\sim} (y * v)\,.$$

The proof is simple: the conditions $x \overset{\mathfrak{E}}{\sim} y\,, u \overset{\mathfrak{E}}{\sim} v$ mean that $A := \{i\colon x_i = y_i\} \in \mathfrak{E}\,; B := \{i\colon u_i = v_i\} \in \mathfrak{E}$; while $A \cap B \subset \{i\colon x_i * u_i = y_i * v_i\}$. And the compatibility condition is guaranteed by the obvious property that:

$$(ii)\colon ((A \in \mathfrak{E})\&(B \in \mathfrak{E})) \Rightarrow ((A \cap B) \in \mathfrak{E})\,.$$

On the other hand, the appearance of the zero-divisors, (or equivalently, the nontrivial idempotents), corresponds to the fact that: $(iv^{\times})$: there is a set A of indices, such that both A and its complement B is not in $\mathfrak{E}$. E. g., A, and B are even and odd indices of $\mathbb{N}$. With these two remarks in background, we will now make a successful construction of the hyperreals. For the convenience of later possible generalizations, we will write I for $\mathbb{N}$ as the set of indices. Indeed both the well-ordering and the countably infinite cardinality of $\mathbb{N}$ are irrelevant here, and I is required to be an infinite set, just to make the construction interesting. But, for some deeper NSA, an index set of larger cardinality is really needed. So we first form the space $\mathbb{R}^I$, whose elements are real-valued functions on I. Such a function is still called a "sequence", (even when I is uncountable), and we continue to write $x(j)$ as x_j. As before, binary and unary operations on $\mathbb{R}^I$ are defined coordinatewise. And in particular, it is a commutative algebra with identity.

Definition 24 (Hyperreals) Let $\mathfrak{U}$ be a free ultra-filter on I, i. e., a set of subsets of I, satisfying some conditions to be specified presently. We call two sequences x and $y \in \mathbb{R}^I \mathfrak{U}$-related, and we write $x \overset{\mathfrak{U}}{\sim} y$, iff:

$$\{j \in I\colon x_j = y_j\} \in \mathfrak{U}\,.$$

The equivalence class of x is denoted by $\pi_{\mathfrak{U}}(x)$ or $x \mod (\overset{\mathfrak{U}}{\sim})$, and called a hyperreal (along $\mathfrak{U}$). The quotient set will be denoted simply as ${}^{\star}\mathbb{R}$, $\mathfrak{U}$ being understood. The canonical mapping from $\mathbb{R}^I$ will be denoted by $\pi_{\mathfrak{U}}$.

Definition 25 (Ultra-Filter) A subset $\mathfrak{U}$ of $\mathfrak{P}(I)$ is called a free-ultrafilter on the infinite set I, if the following conditions are fulfilled:

(i′): if $A \in \mathfrak{U}$ then A is non-empty,
(i): if $A \in \mathfrak{U}$ then A is not a finite set;
(ii): if $A \in \mathfrak{U}$ and $B \in \mathfrak{U}$, then $A \cap B \in \mathfrak{U}$;
(iii): if $A \in \mathfrak{U}$ and $B \supset A$, then $B \in \mathfrak{U}$;
(iv): if $A \subset I$, then either $A \in \mathfrak{U}$ or $(I \setminus A) \in \mathfrak{U}$.

Remark: We may say that the information about an element $x \in \mathbb{R}^I$ is indexed by I. If $A \in \mathfrak{U}$, then the information about this equivalence class $\pi_{\mathfrak{U}}(x)x$ is completely contained in $\langle x_j \colon j \in A\rangle$, the information about x on that A-indexed part, while the part off A is irrelevant for all(our) purpose. (iii) means: "more than enough (part) is enough". (ii) means: "if both A part and B part are enough, actually the intersection part is enough". In mathematical language, if I is any (infinite) set, and $\mathfrak{U} \subset \mathfrak{P}(I)$, $\mathfrak{U}$ is called a filter on I, if it satisfies the condition (i′, ii, iii). If the condition (i) also holds, the filter $\mathfrak{U}$ is called free. The dichotomy condition (iv) for an ultra-filter is very strange: "if the information about A part and the information about B part, when gathered jointly, is enough, then actually one or the other part must be enough.

Note 2 (Non-free Ultrafilter and Axiom of Choice) A non-free ultra"=filter is trivial because it is then the set of all sets of indices containing a particular index i_0. In that case, $x \overset{\mathfrak{U}}{\sim} y$ would simply mean $x(i_0) = y(i_0)$. In other words: the only relevant information is this i_0 item. The whole indexing of information would be non-sense: there is no need of votes-counting, because the policy is determined by the great-leader uniquely. The existence of a free ultrafilter is guaranteed by the axiom of choice. (Indeed we are told that this is a weaker form of the axiom choice).

From now on, a free ultra-filter $\mathfrak{U}$ is chosen. Again by abuse of English, a set $A \in \mathfrak{U}$ is called an ultimate (set of indices). And "for ultimate j" means that set of indices j validating the assertion concerned is in $\mathfrak{U}$. Also, the complement of an ultimate, or by the dichotomy equivalently a non-ultimate, is called an evanescent (or "negligible") set of indices.

Theorem 2 (Totally Ordered Field) *The quotient system $^\star\mathbb{R}$ of the algebra $\mathbb{R}^I$ by the equivalence relation $\overset{\mathfrak{U}}{\sim}$, is a totally ordered field.*

Proof of the compatibility of the operations and $\overset{\mathfrak{U}}{\sim}$: These are all tedious, but they are easy consequences of the filtering condition ii. We will take division as an example. Let $\pi_{\mathfrak{U}}(x) = \pi_{\mathfrak{U}}(u) = \xi \in {}^\star\mathbb{R}$, $\pi_{\mathfrak{U}}(y) = \pi_{\mathfrak{U}}(v) = \eta \in {}^\star\mathbb{R}$, so that $x \overset{\mathfrak{U}}{\sim} u$, $y \overset{\mathfrak{U}}{\sim} v$, and we have to show that

$$\left\{ j \in I \colon \frac{x_j}{y_j} = \frac{u_j}{v_j} \right\} \in \mathfrak{U}\,,$$

and then $\frac{\xi}{\eta} \in {}^\star\mathbb{R}$ is unambiguously defined by

$$\frac{\xi}{\eta} := \pi_{\mathfrak{U}} \left\langle \frac{x_j}{y_j} \right\rangle .$$

Of course it is assumed that (in $^\star\mathbb{R}$), $\eta \neq 0$. That is

$$A := \{ j \in I \colon y_j \neq 0 \} \in \mathfrak{U}\,.$$

Now $\pi_{\mathfrak{U}}(x) = \pi_{\mathfrak{U}}(u)$, $\pi_{\mathfrak{U}}(y) = \pi_{\mathfrak{U}}(v)$, we have:

$$B := \{ j \in I \colon x_j = u_j \} \in \mathfrak{U}\,,$$
$$C := \{ j \in I \colon y_j = v_j \} \in \mathfrak{U}\,.$$

And thus $D := A \cap B \cap C \in \mathfrak{U}$. For index $j \in D$, we have:

$$\frac{x_j}{y_j} = \frac{u_j}{v_j}\,,$$

which is what to be shown. Caution: For $j \notin A$, x_j/y_j is not defined! The obvious remedy is thus: a hyperreal $\eta \in {}^\star\mathbb{R}$ is represented by a "sequence" $y = (y_j \colon j \in A)$, which is defined by an ultimate set $A \in \mathfrak{U}$.

Proof of the total ordering: For $x \in \mathbb{R}^I$, $y \in \mathbb{R}^I$, consider the sets $A := \{ j \colon x_j \geq y_j \}$, $B := \{ j \colon y_j \geq x_j \}$. Now $A \cup B = I \in \mathfrak{U}$, so either $A \in \mathfrak{U}$, or $B \in \mathfrak{U}$, by dichotomy. Thus $\pi_{\mathfrak{U}}(x) \geq \pi_{\mathfrak{U}}(y)$ or $\pi_{\mathfrak{U}}(y) \geq \pi_{\mathfrak{U}}(x)$. Needless to say: when $A \in \mathfrak{U}$, and $B \in \mathfrak{U}$, then $A \cap B = \{ j \colon x_j = y_j \} \in \mathfrak{U}$, thus $x \overset{\mathfrak{U}}{\sim} y$. Note that in the proof above for division, we actually proved the multiplicative invertibility for $\eta \neq 0$, therefore we have shown that in

$$^\star\mathbb{R}$$

there is no non-trivial idempotent. as there is no zero-divisors.

Definition 26 (Infinitesimals and Infinities) Since $\mathbb{R}$ is embedded canonically in the totally ordered field $^\star\mathbb{R}$, a hyperreal α is called infinitesimal iff: for any $n \in \mathbb{N}$, $|\alpha| < \frac{1}{n}$. The set of all infinitesimals is denoted by $\mathbb{I}_0$, while the subset of all non-zero infinitesimals is denoted by $\mathbb{I}$.

The set $\mathbb{I}_0$ is a real linear subspace of $^\star\mathbb{R}$. Note that the only standard member therein is zero. So $\mathbb{I} \cap \mathbb{R} = \emptyset$.

Definition 27 (Monad) Any two hyperreals with infinitesimal difference are called infinitely close. This is an equivalence relation of "infinitely close", written $\approx$, whose equivalence classes are called "monads", after Leibniz.

Definition 28 (Limited and Unlimited) A hyperreal $\alpha \in {}^\star\mathbb{R}$ is called limited (="finite") if, for some $n \in \mathbb{N}$, $-n \leq \alpha \leq n$. In this case, there is uniquely one real number $a = {}^\circ\alpha$ infinitely close to α. Then $\alpha = a + (\alpha - a)$, a is

the standard part of α, and $\alpha - a$ is the infinitesimal part of α. On the other hand, a hyperreal α is called limited ("infinite") iff: for any $n \in \mathbb{N}$, $|\alpha| > n$.

An equivalent definition of this infinity is that: $\frac{1}{\alpha} \in \mathbb{I}$. The set of all limited hyperreals is a subalgebra (with identity) of ${}^\star\mathbb{R}$, with the set $\mathbb{I}_0$ of infinitesimals as an ideal. The quotient algebra is isomorphic with the real field $\mathbb{R}$, so the set of all limited hyperreals is indeed

$$\mathbb{R} + \mathbb{I}_0 = \{c + \zeta : c \in \mathbb{R}\,,\ \zeta \in \mathbb{I}_0\}\,.$$

Note: Keeping $I = \mathbb{N}$, and O, E the subsets of odd and even indices respectively, one of the two sequences $1_O, 1_E$, is (equivalent to) the unit $(=1)$ of ${}^\star\mathbb{R}$, while the other is the zero $(=0)$ of ${}^\star\mathbb{R}$. (They are idempotents complementing each other). The choice is made by the ultrafilter $\mathfrak{U}$. Or, we can choose one or the other from the two sets O, E, to be a member of $\mathfrak{U}$. The axiom of choice allow us to do this before the final set $\mathfrak{U}$ is ultimately found.

Example 2 (A monad is at least a continuum) When $I = \mathbb{N}$, the hyperreal $\varepsilon = \pi_{\mathfrak{U}}\langle \frac{1}{n} : n \in \mathbb{N}\rangle$ is infinitesimal. Then so is ε^3, and indeed for any positive real number $r > 0, \varepsilon^r$ is infinitesimal. These ε^r are inversely ordered as r. This is easily seen from the representing sequences $\langle n^{-r} : n \in \mathbb{N}\rangle$. So the monad of infinitesimals is at least a continuum, cardinally speaking.

Example 3 (A smaller infinitesimal) Suppose real numbers $r > 1, s > 1$ are fixed. Then for eventual $n \in \mathbb{N} = I$,

$$s^{-n} < \frac{1}{n^r}\,;$$

A fortiori, this is true for ultimate n; therefore

$$\nu_s = \pi_{\mathfrak{U}}\langle s^{-n} : n \in \mathbb{N}\rangle < \varepsilon^r\,.$$

To an atheist-Buddhist, an ultrafilter $\mathfrak{U}$ is a Buddha's magnifying glasses, viewing through which a tiny monad is bigger than a solar system.

Star-Extensions

Now we come to the central issue of function-extension and the intimately related transfer principle. For an everywhere-defined function f, the coordinatewise definition still works fine, and we surely have an extension $\overline{f}$ of f to the hyperreals. (The same proof works. Or rather, the assertion here is then a corollary of the former principle.) For a function like csc, which is undefined at $n\pi \in \pi * \mathbb{Z}$, $\overline{\mathrm{csc}}$ as definable at $x = \langle x_n\rangle$, is

$$\overline{\mathrm{csc}}(x) := \langle \mathrm{csc}(x_n)\rangle \in \mathbb{R}^{\mathbb{N}}\,.$$

In $\mathbb{R}^{\mathbb{N}}$, the definition breaks down, if only one term $x_n \in \pi * \mathbb{Z}$. How about the situation in $\mathbb{R}^{\mathbb{N}}_{\mathfrak{E}}$? Write temporarily the domain of definition of csc as $A := \mathrm{Dom}(\mathrm{csc}) = \mathbb{R} \setminus \pi * \mathbb{Z}$, then if for eventual n, $x_n \in B$, the definition goes through! And for such an $x \in \mathbb{R}^{\mathbb{N}}$, $\overline{\mathrm{csc}}(\pi_{\mathfrak{E}}(x)) \in \mathbb{R}^{\mathbb{N}}_{\mathfrak{E}}$ is defined. In complete analogy, for an $x \in \mathbb{R}^{\mathbb{N}}$, if only: "for ultimate n, $x_n \in A$", $\pi_{\mathfrak{U}}\langle \mathrm{csc}(x_n)\rangle \in {}^\star\mathbb{R}$ is well-defined!

Definition 29 (Star-Extension of a Set) For a non-empty set $A \subset \mathbb{R}$, the set ${}^\star A$ is the set of all $\pi_{\mathfrak{U}}(x)$, where $x = \langle x_j\rangle$ is a "sequence" such that, for ultimate j, $x_j \in A$.

Definition 30 (Extension of a Mapping) For any mapping $f : A \mapsto B$, and $x = \langle x_j\rangle \in A^I$, we have the "sequence" $y = f \circ x = \langle f(x_j)\rangle \in B^I$, whose equivalence class $\eta = \pi_{\mathfrak{U}}(y) \in {}^\star B$ is unambiguously determined by $\xi = \pi_{\mathfrak{U}}(x) \in {}^\star A$, and is independent of the representative $x \in A^I$. We then write $\eta = {}^\star f(\xi)$. This mapping ${}^\star f : {}^\star A \mapsto {}^\star B$ is called the star-extention of f. (It is surely an extension.)

It is of utmost importance to note here that these definitions of "star-extensions" are strictly notions of naive set theory. Of course A, B can be any two non-emptysets, and f any function from A to B.

Theorem 3 (Transfer Principle) *All the identities for the standard functions hold true for their star-extensions, simply because they hold true for the original functions.*

Example 4 (HyperIntegers) A hyperinteger $\pi_{\mathfrak{U}}(x) \in {}^\star\mathbb{Z}$ is the equivalence class of a sequence of integers $x \in \mathbb{Z}^{\mathbb{N}}$. If "integer" is changed to "even integer", then we get even hyperinteger $\pi_{\mathfrak{U}}(x) \in {}^\star(2 * \mathbb{Z})$, where $2 * \mathbb{Z} := \{2 * n : n \in \mathbb{Z}\}$ is the set of even integers. And there are odd hyperintegers, too. All the hyperintegers may be classified as even or odd, because of the dichotomy property of the ultrafilter $\mathfrak{U}$.

Example 5 (Cartesian Product and Two Variable Function) For two non-empty sets A, B, by canonical identification,

$${}^\star(A \times B) = {}^\star A \times {}^\star B\,.$$

So for any mapping $g : A \times B \mapsto C$, the star-extention of g may be defined: ${}^\star g : {}^\star A \times {}^\star B \mapsto {}^\star C$.

What this says is easy but tedious to explain. First, we know the identification:

$$(A \times B)^I = A^I \times B^I\,.$$

Because for $x \in A^I, u \in B^I$, the couple $\langle x, u\rangle \in A^I \times B^I$ corresponds bijectively to an element $x \otimes u$ of $(A \times B)^I$,

by:

$$x \otimes u : i \in I \mapsto \langle x_j\,,\ u_j\rangle \in A \times B\,;$$

And the equivalence class of $x \otimes u$ depends on the equivalence classes $\pi_{\mathfrak{U}}(x)$ and $\pi_{\mathfrak{U}}(u)$ only. In this way we are sure that $\pi_{\mathfrak{U}}\langle g(x_j, y_j)\rangle$ is unambiguously determined by $\pi_{\mathfrak{U}}(x)$ and $\pi_{\mathfrak{U}}(u)$ only, and is independent of the representative $x \in A^I, y \in B^I$. Actually we did give a proof for $g =$ division, $A = \mathbb{R} = C, B = \mathbb{R}^\times = \mathbb{R} \setminus \{0\}$. The proof is good for any function of any (finite) number of variables. Some examples of Transfer are:

- ${}^\star\mathrm{abs}(\alpha) = |\alpha| \in {}^\star\mathbb{R}$ is meaningful for $\alpha \in {}^\star\mathbb{R}$. And indeed $|\alpha| = -\alpha$, if $\alpha < 0$.
- $(-1)^\mu$ is meaningful for $\mu \in {}^\star\mathbb{Z}$. And indeed it is 1 for μ an even hyperinteger, it is -1 for μ an odd hyperinteger.
- The star-extensions ${}^\star\sin$, ${}^\star\cos$ of the sine, cosine functions are defined on ${}^\star\mathbb{R}$ and valued in the star-extended unit closed interval ${}^\star[0\ldots 1]$.

Note: From now on, the star-extensions of such classical functions like sin, cos, …, will not have the star displayed.

NS View of Continuous Functions

The introduction of hyperreals gives us a different standpoint to view many familiar concepts.

Remark 4 (Extended Version of a Sequence) Suppose $s = \langle s_n : n \in \mathbb{N}\rangle$ is a sequence of real numbers. This is a function from $\mathbb{N}$ to $\mathbb{R}$. Therefore, it may be star-extended to a function

$${}^\star s\colon m \in {}^\star\mathbb{N} \mapsto s_m \in {}^\star\mathbb{R}\,.$$

We call this latter the extended version of the original sequence. Indeed for any unlimited integer $\nu = \pi_{\mathfrak{U}}\langle n_j : j \in \mathbb{N}\rangle$, we may define

$$s_\nu := \pi_{\mathfrak{U}}\langle s_{n_j} : j \in \mathbb{N}\rangle \in {}^\star\mathbb{R}\,.$$

It is obvious that the convergence of the sequence s

$$\lim_{n\to\infty} s_n = b\,,$$

is equivalent to:

$$\text{for all unlimited integer } \nu \in {}^\star\mathbb{N} \setminus \mathbb{N}\,, \quad s_\nu \approx b\,.$$

As an example, when $r \in \mathbb{R}$, $y \in \mathbb{R}$, and

$$s_n := \cos^m\left(\frac{y}{\sqrt{n}}\right), \quad \text{with} \quad m = \mathrm{floor}(n * r)\,,$$

standard calculus gives us

$$\lim_{n\to\infty} s_n = \mathrm{e}^{-\frac{y^2 r}{2}}\,.$$

We will talk about a function $f\colon \mathbb{R} \mapsto \mathbb{R}$, though usually the domain of definition is only required to be an open interval.

Theorem 4 *A function f is bounded around a point $a \in \mathbb{R}$, iff ${}^\star f$ maps the monad of a into finite hyperreals.*

Proof If f maps a neighborhood $(a - e \ldots a + e)$ of a into $[-M \ldots M]$, where $e \in \mathbb{R}$ is positive, then for $\xi = \pi_{\mathfrak{U}}\langle x_j\rangle \approx a$, we have $a - e < x_j < a + e$ for ultimate j, and thence $f(x_j) \in [-M \ldots M]$. Accordingly:

$$-M^\star \le {}^\star f(\xi)^\star \le M\,.$$

On the other hand, if f is unbounded around a, then for any $j \in I = \mathbb{N}$, there is x_j (in the domain of definition of f), with $|x_j - a| < \frac{1}{j}$, and $f(x_j) > j$. Thus $\xi := \pi_{\mathfrak{U}}\langle x_j\rangle \approx a$, but $|{}^\star f(\xi)|$ is infinite. □

Theorem 5 *Let $A \subset \mathbb{R}$ be an interval, $b \in A$, and $f\colon A \mapsto\!\!\!\!\to \mathbb{R}$. Then f is continuous at b, iff: for all $\xi \in {}^\star A$,*

$$\xi \approx b \Rightarrow {}^\star f(\xi) \approx f(b)\,.$$

In other words, f is continuous at b, iff ${}^\star f$ maps the monad $a + \mathbb{I}_0$ into the monad $f(a) + \mathbb{I}_0$.

Proof By the standard definition, when f is continuous at b, and $\epsilon > 0$ is given, there is a $\delta > 0$, such that, for $|x - b| < \delta$, (and $x \in A$), $|f(x) - f(b)| < \epsilon$. Now $\xi = \pi_{\mathfrak{U}}\langle x_j\rangle$, and $\xi - b \in \mathbb{I}_0$, so for ultimate j, $|x_j - b| < \delta$. Thus $|f(x_j) - f(b)| < \epsilon$, for ultimate j. This means ${}^\star f(\xi) - f(b) \in \mathbb{I}_0$, or, ${}^\star f(\xi) \approx f(b)$. Suppose f is not continuous at b, then there is $\epsilon > 0$, such that for whatever $j \in \mathbb{N}$, there is $x_j \in A$, with $|x_j - b| < \frac{1}{j}$, while $|f(x_j) - f(b)| > \epsilon$. Then with $\xi = \pi_{\mathfrak{U}}\langle x_j\rangle \in {}^\star A$, $\xi \approx b$, while $|{}^\star f(\xi) - f(b)| > \epsilon$. □

Corollary 1 *If f is continuous at c, it is bounded around c.*

Theorem 6 *Let $A \subset \mathbb{R}$ be a compact (=closed and bounded) interval, and $f\colon A \mapsto \mathbb{R}$ is continuous. Then f is bounded.*

Proof If f is unbounded, then for $j \in \mathbb{N}$, there is $x_j \in A$, with $|f(x_j)| > j$. And we find $\xi = \pi_{\mathfrak{U}}\langle x_j\rangle$, with ${}^\star f(\xi)$ infinite. But $\xi \in {}^\star A$ is a limited hyperreal, and its standard part $b = {}^\circ\xi \in A$, $b \approx \xi$. By the continuity criterion, ${}^\star f(\xi) \approx f(b) \in \mathbb{R}$. This is a contradiction. □

Theorem 7 *Let $A \subset \mathbb{R}$ be an interval, and $f\colon A \mapsto \mathbb{R}$. Then f is uniformly continuous, iff: for all $\xi \in {}^\star A$, $\eta \in {}^\star A$,*

$$\xi \approx \eta \Rightarrow {}^\star f(\xi) \approx {}^\star f(\eta)\,.$$

Proof By the standard definition, when f is uniformly continuous on A, and $\epsilon > 0$ is given, there is a $\delta > 0$, such that, for $|x-y| < \delta$, (and $x \in A, y \in A$), $|f(x)-f(b)| < \epsilon$. Now $\xi = \pi_{\mathfrak{U}}\langle x_j\rangle \in {}^\star A, \eta = \pi_{\mathfrak{U}}\langle y_j\rangle \in {}^\star A$, and $\xi - \eta \in \mathbb{I}_0$, so for ultimately many j, $|x_j - y_j| < \delta$. Thus $|f(x_j) - f(y_j)| < \epsilon$, for ultimately many j. This means ${}^\star f(\xi) - {}^\star f(\eta) \in \mathbb{I}_0$, or, ${}^\star f(\xi) \approx {}^\star f(\eta)$. Suppose f is not uniformly continuous, then there is $\epsilon > 0$, such that for whatever $j \in \mathbb{N}$, there is a pair $x_j \in A, y_j \in A$, with $|x_j - y_j| < \frac{1}{j}$, while $|f(x_j) - f(y_j)| > \epsilon$. Then with $\xi = \pi_{\mathfrak{U}}\langle x_j : j\rangle \in {}^\star A, \eta = \pi_{\mathfrak{U}}\langle y_j : j \in I\rangle \in {}^\star A, \xi \approx \eta$, while $|{}^\star f(\xi) - {}^\star f(\eta)| > \epsilon$. Contradiction. One of the merits of NSA is that: a standard notion may be interpreted in a nonstandard way, which is pedagogically advantageous. There will be, as examples, a couple of theorems easily proved by the NS criterion just given. But first we introduce some notions very useful later on. □

Definition 31 (Hyperfinite Set) Let $T_j \subset A$ for each $j \in I$. Then the sequence defines an "internal" subset $\mathcal{T} := \pi_{\mathfrak{U}}(T)$ of ${}^\star A$:

$$\pi_{\mathfrak{U}}(T) := \{\pi_{\mathfrak{U}}(x) : x_j \in T_j \text{ for ultimate } j\} \subset {}^\star A\,.$$

If for ultimate j, the set T_j is finite: $\mathrm{card}(T_j) < \infty$, $\mathcal{T}$ is called a hyperfinite set with internal cardinal $\pi_{\mathfrak{U}}\langle \mathrm{card}(T_j)\rangle \in {}^\star\mathbb{N}$.

Example 6 (Hyperfinite Equi-Partition) Consider the hyperfinite set $\mathcal{T}$ of equi-partition points with cardinal $1 + \mathcal{N}$:

$$\begin{aligned}
\mathcal{N} &:= \pi_{\mathfrak{U}}\langle N_j\rangle \in \mathbb{N}^I; (N_j < N_{j+1} \to \infty;)\\
T_j &:= \left\{\frac{k}{N_j} : k \in \mathbb{Z}\,,\ 0 \le k \le N_j\right\};\\
\mathcal{T} &:= \pi_{\mathfrak{U}}\langle T_j\rangle\,.
\end{aligned}$$

Three convenient choices are $N_j = j!$ and $N_j = 2^j$, and $N_j = 3^j$. The corresponding partition set $\mathcal{T}$ is then called, respectively, the partition by factorial, or binary, or ternary fractions.

Lemma 6 (Density of Equi-Partition Set) *The countable set*

$$T_\infty = \cup_j T_j$$

is dense in the unit interval $\mathbb{U}$: For any number $c \in \mathbb{U}$, there is a hyperreal $\zeta \in^\star T, \zeta \approx c$.

Proof There are two convenient choices: the left- and the right-approximation:

$$\zeta = \mathrm{floor}_{\mathcal{T}}(c) = \pi_{\mathfrak{U}}\left\langle \frac{\mathrm{floor}(c * N_j)}{N_j} : j \in I\right\rangle;$$

$$\text{or}\quad \zeta = \mathrm{ceil}_{\mathcal{T}}(c) = \pi_{\mathfrak{U}}\left\langle \frac{\mathrm{ceil}(c * N_j)}{N_j} : j \in I\right\rangle.$$

□

Note that, for any $x \in \mathbb{U}$,

$$\mathrm{st}(\mathrm{floor}_{\mathcal{T}}(x)) = x\,, \quad \mathrm{st}(\mathrm{ceil}_{\mathcal{T}}(x)) = x\,.$$

Both $\mathrm{floor}_{\mathcal{T}}$ and $\mathrm{ceil}_{\mathcal{T}}$ are (Skolem?) "inverse- function" of the standard-part mapping st from $\mathbb{U}$ onto $\mathcal{T}$. Indeed for the binary-rational partition $T_j = 2^j$, or the ternary-rational partition $T_j = 3^j$, the floor-approximation gives the binary or ternary expansion respectively.

Theorem 8 *Let $A \subset \mathbb{R}$ be a compact interval, and $f : A \mapsto \mathbb{R}$ is continuous. Then f has a maximum and minimum.*

Proof Now(by simple scaling and translations) we may and will assume $A = \mathbb{U} = [0 \ldots 1]$ the unit interval. Take a hyperinteger $\mathcal{N} := \pi_{\mathfrak{U}}\langle N_j\rangle$, and then the equi-partition $\mathcal{T} = \pi_{\mathfrak{U}}(T)$ into $\mathcal{N}$ parts. Here $T_j := \{k/N_j : k \in \mathbb{Z}, 0 \le k \le N_j\}$. Now for each j, the maximum value of

$$f[T_j] := \{f(x) : x \in T_j\}$$

is unique, though the maximum point x_j may be non-unique:

$$f(x_j) \ge f(x)\,, \quad \forall x \in T_j\,;$$

We just choose (say) the least one such maximum point x_j. And define

$$\xi := \pi_{\mathfrak{U}}(x) \in {}^\star A\,; \quad x_0 := {}^\circ\xi\,; \quad x_0 \in A\,.$$

Then for any $c \in A$, and ζ as before,

$$\begin{aligned}
&{}^\star f(\zeta) \le {}^\star f(\xi)\,; \quad \zeta \approx c\,;\\
&\qquad x_0 \approx \xi;\ {}^\star f(x_0) \approx {}^\star f(\xi)\,; \quad {}^\star f(c) \approx {}^\star f(\zeta)\,;
\end{aligned}$$

We see that $f(c) \le f(x_0), \forall c \in A$, i. e., f attains its maximum-value at x_0. □

Theorem 9 (Intermediate-Value) *Let $f : A \mapsto \mathbb{R}$ be continuous, $v \in \mathbb{R}, A = [a \ldots b]$, and $(v-f(a))(v-f(b)) < 0$. Then there is $u \in (a \ldots b)$, such that $f(u) = v$.*

Proof We may and will assume $A = \mathbb{U}$, $a = 0$, $b = 1$, $f(0) < 0 < f(1)$. As before, we take a hyper-finite $\mathcal{N}$

equi-partition $\mathcal{T} = \pi_{\mathfrak{U}}\langle T_j\rangle$, $\mathcal{N} = \langle N_j\rangle$. Find the least $x_j \in T_j$, such that $f(x_j) \geq 0$. Because $0 \in T_j$, $1 \in T_j$, and $f(0) < 0 < f(1)$, we know $f(x_j) \geq 0$, $f(x_j - 1/N_j) < 0$. With $\xi := \pi_{\mathfrak{U}}\langle x_j\rangle \in {}^\star A$, $u := {}^\circ\xi$, then $\epsilon := \pi_{\mathfrak{U}}\langle 1/N_j\rangle$, we have:

$$ {}^\star f(\xi) \geq 0\,, \quad {}^\star f(\xi - \epsilon) < 0\,. $$

But $\xi \approx u$; $\xi - \epsilon \approx u$; and the continuity of f at u implies:

$$ {}^\star f(\xi) - f(u) \in \mathbb{I}_0\,, \quad {}^\star f(\xi - \epsilon) - f(u) \in \mathbb{I}_0\,. $$

Therefore:

$$ \begin{aligned} 0 \leq {}^\star f(\xi) &< {}^\star f(\xi) - {}^\star f(\xi - \epsilon) \\ &= ({}^\star f(\xi) - f(u)) - ({}^\star f(\xi - \epsilon) - f(u)) \in \mathbb{I}_0\,, \end{aligned} $$

resulting in ${}^\star f(\xi) \in \mathbb{I}_0$. Or ${}^\circ({}^\star f(\xi)) = 0$, which means $f(u) = 0$. □

Theorem 10 *Let $A \subset \mathbb{R}$ be a compact interval, and $f\colon A \mapsto \mathbb{R}$ be continuous. Then f is uniformly continuous.*

Proof Suppose $\xi \in {}^\star A, \eta \in {}^\star A$, and $\xi \approx \eta$. Now let $c = {}^\circ(\xi) \in A$. It follows that both $\xi \approx c$ and $\eta \approx c$ hold true. The continuity of f implies that both ${}^\star f(\xi) \approx f(c)$ and ${}^\star f(\eta) \approx f(c)$ hold. Therefore ${}^\star f(\xi) \approx {}^\star f(\eta)$. □

NS Calculus

According to the standard definition:

$$ f'(a) = \lim_{\Delta x \to 0} \frac{f(a + \Delta x) - f(a)}{\Delta x}\,; $$

the differentiability of f at a implies the continuity of f at a, which means that ${}^\star f$ maps the monad $a + \mathbb{I}_0$ of a into the monad $f(a) + \mathbb{I}_0$ of $f(a)$.

Definition 32 (Difference) If f is a real-valued function defined around a, then the "discern" of f at a is the (non-standard) function

$$ \Delta_a f\colon \mathrm{d}x \in \mathbb{I}_0 \mapsto {}^\star f(a + \mathrm{d}x) - f(a) \in {}^\star\mathbb{R}\,. $$

Of course we know that the continuity of f at a means that the range of $\Delta_a f$ is included in $\mathbb{I}_0$. So we are led to consider (non-standard) functions from $\mathbb{I}_0$ to $\mathbb{I}_0$. From now on, $\mathrm{d}x, \mathrm{d}y, \cdots$ will be variables whose values are infinitesimals. (So they live in $\mathbb{I}_0$.)

Definition 33 (Tangential and Tangent) If $\Phi\colon \mathbb{I}_0 \mapsto \mathbb{I}_0$, $\Psi\colon \mathbb{I}_0 \mapsto \mathbb{I}_0$, and for any nonzero infinitesimal $\epsilon \in \mathbb{I}$,

$$ \frac{\Phi(\epsilon) - \Psi(\epsilon)}{\epsilon} \approx 0\,, $$

they are called tangential to each other. For any number $c \in \mathbb{R}$, the associated tangent is the (non-standard) function:

$$ \mathrm{d}x \in \mathbb{I}_0 \mapsto c * \mathrm{d}x \in \mathbb{I}_0\,. $$

Theorem 11 (Differentiability) *A function f defined on an open interval U is differentiable at $a \in U$, and there with derivative $f'(a) = c$, iff: $\Delta_a f$ is tangential to the tangent associated with c.*

Proof If $f'(a) = c$, then for any $k \in \mathbb{N}$, there is $e_k \in \mathbb{R}$ with $e_k > 0$, such that whenever $|u| < e_k$, $|(f(a + u) - f(a))/u - c| < \frac{1}{k}$. Now fix $k \in I$. Thus to any (non-zero) $\mathrm{d}x = \pi_{\mathfrak{U}}\langle u_j\rangle \in \mathbb{I}$, for ultimate j, $0 < |u_j| < e_k$, resulting in $|(f(a + u_j) - f(a))/u_j - c| < \frac{1}{k}$. So

$$ \left|\frac{\Delta_a f(\mathrm{d}x) - c * \mathrm{d}x}{\mathrm{d}x}\right| = \left|\frac{{}^\star f(a + \mathrm{d}x) - f(a)}{\mathrm{d}x} - c\right| < \frac{1}{k}\,. $$

This being true for all $k \in \mathbb{N}$, we have: for any $\mathrm{d}x \in \mathbb{I}$,

$$ \left|\frac{\Delta_a f(\mathrm{d}x) - c * \mathrm{d}x}{\mathrm{d}x}\right| \in \mathbb{I}_0\,; $$

which means: $\Delta_a f$ is tangential to the tangent associated with c. Conversely, if f is not differentiable at a, one can show that the nonstandard condition of differentiability can not be satisfied. Indeed then the following four limits of the difference quotient $\frac{f(a+u)-f(a)}{u}$ are not all finite and coincident:

$$ \limsup_{u \downarrow 0}\,, \quad \liminf_{u \downarrow 0}\,, \quad \limsup_{u \uparrow 0}\,, \quad \liminf_{u \uparrow 0}\,. $$

For example, a case may be

$$ \begin{aligned} b := \limsup_{u \downarrow 0} \frac{f(a + u) - f(a)}{u} \\ > e := \liminf_{u \downarrow 0} \frac{f(a + u) - f(a)}{u}\,. \end{aligned} $$

In that case, there are positive sequences $u \in \mathbb{R}^{\mathbb{N}}$, $v \in \mathbb{R}^{\mathbb{N}}$, decreasing to 0, such that

$$ \begin{aligned} b &:= \lim_{j \to \infty} \frac{f(a + u_j) - f(a)}{u_j}\,, \\ e &:= \lim_{j \to \infty} \frac{f(a + v_j) - f(a)}{v_j}\,. \end{aligned} $$

Taking $\xi = \pi_{\mathfrak{U}}\langle u_j\rangle \in \mathbb{I}$, and $\eta = \pi_{\mathfrak{U}}\langle v_j\rangle \in \mathbb{I}$, we see that: the nonstandard condition of differentiability can not be satisfied simultaneously for $\mathrm{d}x = \xi$ and for $\mathrm{d}x = \eta$, whatever the choice of c. □

Theorem 12 (Chain-Rule) *If a function f is differentiable at a, and a function g is differentiable at $b = f(a)$, then the composite function $h := g \circ f$ is differentiable at a, with derivative $h'(a) = g'(b) * f'(a)$.*

Proof For any $\xi \in \mathbb{I}, \eta \in \mathbb{I}$, we have:

$$\frac{{}^\star f(a+\xi) - f(a)}{\xi} \approx f'(a)\,;$$
$$\frac{{}^\star g(b+\eta) - g(b)}{\eta} \approx g'(b)\,.$$

Let now $\eta = {}^\star f(a+\xi) - f(a)$. Then:

$$\frac{{}^\star h(a+\xi) - h(a)}{\xi} = \frac{{}^\star g({}^\star f(a+\xi)) - g(f(a))}{\xi}$$
$$= \frac{{}^\star g(b+\eta) - g(b)}{\eta} * \frac{\eta}{\xi} \approx g'(b) * f'(a)\,.$$

Here it is assumed that $\eta \neq 0$ so as to allow the division by η. But when $\eta = 0$, then necessarily $f'(a) = 0$, and the proposed identity is trivial. □

Theorem 13 (Leibniz' Product Rule) *If two functions f and g are differentiable at a, then so is their product $h, h(x) = f(x) * g(x)$. And*

$$h'(a) = f'(a) * g(a) + f(a) * g'(a)\,.$$

Proof The non-standard calculations are essentially the same as the standard ones. □

Lemma 7 *If f has a local maximum at $x = a$, then*

$$\forall dx \in \mathbb{I}\,, \quad \Delta_a f(dx) \leq 0\,.$$

Theorem 14 ("Fermat") *On an open interval, if a function f has a maximum at a, and there differentiable, then $f'(a) = 0$.*

Proof Recall the signum function $\mathrm{sign}(x) = \frac{x}{|x|}$ for $x \neq 0$. We will not put the star on its star-extension. For $\mathrm{d}x \in \mathbb{I}$,

$$\frac{\Delta_a f(\mathrm{d}x) - f'(a)\mathrm{d}x}{\mathrm{d}x} \in \mathbb{I}\,;$$

So if $f'(a) \neq 0$, we take $\mathrm{d}x \in \mathbb{I}$ to have the same signum as $f'(a)$, then there is a contradiction:

$$\mathrm{sign}(\Delta_a f(\mathrm{d}x)) = \mathrm{sign}(f'(a) * \mathrm{d}x) > 0\,.$$

□

Conventional deductions then give, as corollaries,

Theorem 15 ("Rolle") *Let a function f is continuous on a closed interval $[a \ldots b]$, and differentiable on the open interval $(a \ldots b)$. If moreover $f(a) = f(b)$, then there exists a zero of the derived function f' in the open interval. Generally, there exists ξ in the open interval $(a \ldots b)$, such that*

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}\,.$$

has a maximum at a, and there differentiable, then $f'(a) = 0$.

Remark 5 For a function f defined on a closed interval, one-sided derivative may still be defined at the endpoints naturally. Nelson mentioned that a stronger notion is in some way more convenient than differentiability. f is called derivable at a point a, iff there is a derivative $f'(a) = c$ such that:

$$(as\ x_1 \to a\,,\ x_2 \to a\,,\ x_1 \neq x_2)\,,$$
$$\lim\left(\frac{f(x_2) - f(x_1)}{x_2 - x_1} - f'(a)\right) = 0\,.$$

The NS formulation is:

$$(\epsilon_1 \in \mathbb{I}\,,\ \epsilon_2 \in \mathbb{I}\,,\ \epsilon_1 \neq \epsilon_2)$$
$$\Rightarrow \frac{{}^\star f(a+\epsilon_2) - {}^\star f(a+\epsilon_1)}{\epsilon_2 - \epsilon_1} \approx f'(a)\,.$$

Of course all the differentiation rules are valid for this stronger-sense derivative.

Theorem 16 *A function f is derivable on an interval, (i. e., at every point of the interval), iff it is differentiable with continuous derivative everywhere on the interval.*

The function $f(x) = x^2\cos(\frac{1}{x})$ (extended by $f(0) = 0$), is differentiable on $\mathbb{R}$, but $f'(x) = 2x\cos(\frac{1}{x}) + \sin(\frac{1}{x})$ (extended by $f'(0) = 0$), is not continuous at $x = 0$. Indeed it is not derivable there.

NS Integral Calculus

Let $N := (N_j\colon j \in I) \in \mathbb{N}^I$ be a strictly increasing sequence of natural numbers, and for each $j \in I$, define

$$T_j = \left\{\frac{k}{N_j}\colon k \in \mathbb{Z}\,,\ 0 \leq k \leq N_j\right\}\,.$$

The unit interval $\mathbb{U}$ is correspondingly partitioned into N_j subintervals $\left[(k-1)/N_j \ldots k/N_j\right]$, $k = 1, 2, \cdots, N_j$. Suppose f is a function on the unit interval $\mathbb{U}$. The left-Riemann sum of f, relative to this partition is

$$S_j := \frac{1}{N_j} \sum \left(f(x) \colon x \in T_j' \right) , \quad T_j' := T_j \setminus \{1\} .$$

Assuming that f is Riemann-integrable (in the standard sense), this sequence $S = \langle S_j \rangle$ of Riemann-sums converges to the Riemann integral $\int_{\mathbb{U}} f(x)\mathrm{d}x$.

Theorem 17 (Riemann Sum) *The Riemann-integral for a Riemann-integrable function on the unit interval is the standard part of the "hyperaverage" of f, over a hyperfinite equi-partition $\mathcal{T}$ as above:*

$$\int_{\mathbb{U}} f(x)dx = \mathrm{st}\left(\frac{1}{\mathcal{N}} \sum ({}^\star f(t) \colon t \in \mathcal{T}) \right) .$$

Remark 6 For the improper Riemann integral $\int_0^1 (\mathrm{d}x)/\sqrt{1-x}$, the formula is valid.

Definition 34 Let now $f \colon \mathbb{R} \mapsto \mathbb{R}$ be a function, bounded, and compactly supported, which means $f(x) = 0$ for x outside a finite interval, then we may define (the "integral over $\mathbb{R}$")

$$\int f(x)\mathrm{d}x := \mathrm{st}\left(\frac{1}{\mathcal{N}} \pi_{\mathfrak{U}} \langle S_j \rangle \right) ;$$

$$\text{where } S_j := \sum (f(x) \colon x \in T_j) ;$$

$$T_j = \left\{ \frac{k}{N_j} \colon k \in \mathbb{Z} , \ -N_j^2 \le k < N_j^2 \right\} ;$$

And then for any bounded function f defined on $\mathbb{R}$:

$$\int_a^b f(x)\mathrm{d}x := \int f(x) * 1_{[a \ldots b]}(x)\mathrm{d}x .$$

The definition is valid for any function f as long as it is bounded. (This is just like the Banach-limit for a bounded sequence. Axiom of choice is also involved in the latter!) The Riemann integrability is relevant only to prove that this "integral" is independent of the Riemann-summation scheme. We will now skip all these matter, mentioning only that, as is easily seen:

$$(a < b < c) \Rightarrow \int_a^c f(x)\mathrm{d}x = \int_a^b f(x)\mathrm{d}x + \int_b^c f(x)\mathrm{d}x .$$

This justifies the usual definition that

$$\int_a^b f(x)\mathrm{d}x := -\int_b^a f(x)\mathrm{d}x , \quad \text{for} \quad b < a .$$

Theorem 18 (Fundamental Theorem of Calculus)

(1) If f is continuous on an interval containing a point a, and $g(u) := \int_a^u f(x)dx$, then g is continuous, and for a continuity point b of f, the function g is differentiable at b, with $g'(b) = f(b)$.

(2) If f is derivable on the interval $[a \ldots b]$, so that the derivative f' is continuous, hence Riemann integrable, then

$$\int_a^b f'(x)dx = f(b) - f(a) .$$

NS-ODE

Theorem 19 (Cauchy–Peano) *If $f \colon (x, t) \in \mathbb{R} \times [0 \ldots 1] \mapsto f(x, t) \in \mathbb{R}$ is bounded and continuous, then, for any "initial-position" $a \in \mathbb{R}$, there is a "solution-path"*

$$y \colon t \in [0 \ldots 1] \mapsto y(t) \in \mathbb{R} ,$$

such that: it is continuous, and

$$y(0) = a , \quad y'(t) = f(y(t), t) , \quad \forall t \in [0 \ldots 1] .$$

Proof The polygonal approximation is the following well-known scheme

$$y = \lim_{j \to \infty} y_j \ :$$

Take a natural number N_j, and divide the time-interval into N_j equal subintervals. In the duration of kth subinterval $(k-1)/N_j \le t < k/N_j$, the velocity is approximated by the velocity-field at the latest position: $y_j(t) := f(y((k-1)/N_j), (k-1)/N_j)$, and consequently there is a recursion:

$$y_j\left(\frac{k}{N_j}\right) := y_j\left(\frac{k-1}{N_j}\right) + \frac{1}{N_j}$$
$$* f\left(y_j\left(\frac{k-1}{N_j}\right) , \quad \frac{k-1}{N_j} \right) ; \quad k = 1, 2, \cdots, N_j .$$

The standard way is to take a "limiting point" of the sequence of approximate solutions (which are polygonal = "piecewise-linear"), $y_j(t)$, $j = 1, 2, 3, \ldots$. They are considered as points of the path-space (= function space, with uniform topology). Ascoli's compactness criterion ensures the existence of such a limiting point y_∞. From the uniform convergence

$$\lim_{j \to \infty} y_j(t) = y_\infty(t) ,$$

$$\lim_{j \to \infty} f(y_j(t) , t) = f(y_\infty(t), t) ,$$

it follows easily the sought-for identity:

$$y_\infty(t) = a + \int_0^t f\left(y_\infty(s), s\right) \mathrm{d}s\,.$$

The standard proof is translated into the NS proof thus: With $N_j \uparrow \infty$, we have an internal hyper-natural number (unlimited, of course), $\mathcal{N} = \pi_{\mathfrak{U}}(N)$, $N := \langle N_j \rangle$, together with the internal set of partition-points on the NS time-interval:

$$\mathcal{T} := \left\{ \frac{k}{\mathcal{N}} : k \in \mathbb{Z}\,,\ 0 \le k \le \mathcal{N} \right\}\,.$$

For "time" $\tau = \frac{k}{\mathcal{N}}$ on this hyperfinite set, define the $\mathcal{N}$th approximate "position" $Y(\tau)$ to be

$$Y\left(\frac{k}{\mathcal{N}}\right) = a + \frac{1}{\mathcal{N}} * \sum_{i=0}^{k-1}{}^{\star} f\left(Y\left(\frac{i}{\mathcal{N}}\right), \frac{i}{\mathcal{N}}\right)\,.$$

Finally take

$$y(t) := \mathrm{st}\left(Y\left(\frac{\mathrm{floor}(t * \mathcal{N})}{\mathcal{N}}\right)\right)\,.$$

The above NS summation-multiplied-by-infinitesimal goes to Riemann integral:

$$y(t) = a + \int_0^t f(y(s), s)\mathrm{d}s\,.$$

□

It is well-known that under the very general condition of boundedness and continuity, the solution is not unique, reflecting the arbitrariness in the choice of one limiting point from a compact set. Of course this corresponds to the arbitrariness of the choice of the unlimited $\mathcal{N} \in {}^\star\mathbb{N}$, or rather, the hyper-finite partition $\mathcal{T}$.

General Idea of NSA

The Set-Terminology

Ever since G. Cantor invented the Mengenlehre, and after the Hilbertian axiomatization (r) evolution, the naive set-theory became the common language of mathematics. Everything is (to be expressed in terms of) set! As an example, the color "blue" may be defined as "the set of all blue things". (I'd like to make this as my first teaching of pedagogy.) If this sentence is hardly comprehensible, then an easier explanation is as follows. Suppose I say: "This stone ω is blue". Then what I mean is the mathematical sentence "$\omega \in B$", where B is a conveniently collected set of blue stones. Of course, B should be a subset of a convenient "total set" Ω of stones. With this primitive idea of 'membership relation', simple set-notions fit in with our daily logic. Precisely: if a is the sentence "$x \in A$", and b is "$x \in B$", then the logical combinations of assertions correspond to the Boolean combinations of sets in the well-known way:

"a or b" is "$x \in A \cup B$";
"a and b" is "$x \in A \cap B$";
"not b" is "$x \in \Omega \setminus B$".

Here a convenient total set Ω should be available for the logical analysis. From this thinking, a set $\mathcal{A}$ of subsets of Ω is called a Boolean algebra of sets, if it is closed under the three set-operations mentioned above: $\cup, \cap$, and "complementation" (set-difference by the total set). My second example of the "triumph of naive set theory" is the axiomatization of probability theory by A. N. Kolmogorov. Its publication was 10 years after N. Wiener's construction of Wiener process. Wiener wanted to talk about the "probabilistic behavior of the path W of a particle". By a path is meant a function from the time-axis to the state-space, $W(t)$ signifying "the position at time t". Here the time-axis is the half-line $[0 \cdots +\infty)$, and the state-space is the real line $\mathbb{R}$. And Wiener requires that the path is continuous, and starts from the origin: $X = 0$. In the view point of Kolmogorov, the phrase "probabilistic behavior of the path W of a particle" should be changed to "behavior of a random particle"."A random particle" just means that it is "one chosen from many"; these many form a set. (This set, the so-called probability space, is always an imagined set. It is almost always an infinite set, so never "concrete" for students of probability theory!) The picture is: Lady Luck collects a big set Ω of all particles (= chancelings). For each particle $\omega \in \Omega$, its path is $W(\cdot, \omega)$, mapping time t to its position $W(t, \omega)$ at time t. (So Wiener process W is a "two-variable function", its domain of definition is the set-product $\mathbb{R}_+ \times \Omega$. Textbooks of probability theory always suppress the variable ω which signifies the randomness.) The success of Kolmogorov's axiomatization is simply this: it makes precise what kind of "questions about probability" one can ask. Now probability comes in, only because Lady Luck chooses one chanceling ω from all conceivables. (Then this chosen one is the "reality"). We are interested (only) in the assertions of this form "$\omega \in B$". We can only ask questions of this kind: "What is the probability of the event $\omega \in B$?" And that only for some kind of set B. The answer $\mu(B)$ is the probability predestined by Her. For what kind of $B \subset \Omega$ is $\mu(B)$ meaningful is also Her designation. Then we simply call B an event. So $\mu(B)$ is now the probability of event B. (In-

stead of the "event $\omega \in B$"). Since we can do Boolean combinations of assertions, correspondingly, the set $\mathcal{A}$ of all events should be a Boolean set-algebra. But in probability theory, always the stronger condition of sigma-completeness is required: $\mathcal{A}$ is closed under countable union, thus called a σ-algebra.

Definition 35 ((Kolmogorov) Probability Space) A probability space is a nonempty set Ω, together with a measure structure $\langle \mathcal{A}, \mu \rangle$, where $\mathcal{A}$ is a σ-algebra on Ω, and $\mu : \mathcal{A} \mapsto \mathbb{U}$ satisfies the countable-additivity condition: If $\forall n, B_n \in \mathcal{A}$ and $\forall m \neq n, B_m \cap B_n = \emptyset$, then

$$\mu\left(\cup_{1 \leq n < \infty} B_n\right) = \sum_{n=1}^{\infty} \mu(B_n) .$$

Returning to Wiener's process. Lady Luck should allow assertions $\omega \in B$ of the form $W(t, \omega) \in (c_1 \ldots c_2]$, where time $t \in [0 \cdots + \infty)$ and interval $(c_1 \ldots c_2]$ are arbitrary. With this condition satisfied, the two-variable function $W : \langle t, \omega \rangle \mapsto W(t, \omega)$ is called a stochastic process on the probability space $(\Omega, \mathcal{A}, \mu)$. Actually the central questions asked by (Bachelier and) Wiener are events of the following forms:

$$B := \{\omega \in \Omega : \quad \text{for} \quad k = 1, 2, \cdots, n, W(t_k, \omega) - W(s_k, \omega) \in A_k\} ,$$

where

$$s_1 < t_1 \leq s_2 < t_2 \leq \cdots \leq s_n < t_n ; \qquad A_1, A_2, \cdots, A_n , \quad \text{are intervals}$$

Theorem 20 (Wiener) *There exists a probability space $(\Omega, \mathcal{A}, \mu)$, and a stochastic process W on it, such that: for each $\omega \in \Omega$, the sample-path $W(\cdot, \omega)$ is a continuous function on $[0 \cdots + \infty)$; and, for set B as above,*

$$\mu(B) = \prod_{1 \leq k \leq n} \frac{1}{\sqrt{2\pi(t_k - s_k)}} \int_{A_k} e^{\frac{-x^2}{2(t_k - s_k)}} dx .$$

Naive Set Operations

As we have explained above, a mathematical system is a set with structure, which is a finite set of items of functions, relations, or functions. We now explain that starting from a ground set, there is a super-structure, which will contain all conceivable structures. Naive as we are, sometimes we just have to be very careful. So we give a very careful definition of couple and relation, which will be a foundation of all that follows.

Definition 36 (Couple and Cartesian Product) For two non-empty sets A, B, the Cartesian product is defined as

$$A \times B := \{\langle a, b \rangle : a \in A , \quad \text{and} \quad b \in B\} .$$

Here "a couple", or "an ordered pair", is defined by

$$\langle a, b \rangle := \{\{a\}, \{a, b\}\} .$$

The ordered n-tuple, for $n > 2$, may be defined recursively as

$$\langle x_1 , x_2 , \ldots , x_n \rangle := \langle \langle x_1 , x_2 , \ldots , x_{n-1} \rangle , x_n \rangle .$$

Note that we could write the "1-tuple" as $\langle a \rangle := a$. Then write:

$$A^n := \{\langle x_1 , x_2 , \ldots , x_n \rangle : x_1 , x_2 , \ldots , x_n \in A\} .$$

Thus $A \times B \times C$ is defined to be $(A \times B) \times C$. (And so forth!)

The key to this definition is the observation that

$$(\langle x, y \rangle = \langle u, v \rangle) \quad \text{iff} \quad ((x = u) \quad \text{and} \quad (y = v)) .$$

Here it is to be emphasized that when forming a set, multiple enumeration of any element just counts once, but does no harm. And in the case of forming the pair $\langle a, b \rangle$, when $a = b$, the pair is reduced to $\langle a, a \rangle := \{\{a\}\}$, and this is a singleton set whose only element is the singleton set $\{a\}$.

Definition 37 (Relation) If $R \subset A \times B$, then R is a relation from A to B. And in case $A = B$, it is called a relation on A. The domain of R is the set

$$\mathrm{Dom}(R) := \{x \in A : \exists y \in B , \langle x , y \rangle \in R\} .$$

If $\langle x, y \rangle \in R$, sometimes we write (the "infix notation"!) xRy, still sometimes (the "prefix notation"!) $R(x, y)$. Suppose R and S are two relations, and $R \subset S$, then R is a restriction of S, and S an extension of R.

Definition 38 (Function and Strict-Evaluation) If f is a relation from A to B, and, for any $x \in A$, there is exactly one $y \in B$, such that $\langle x, y \rangle \in f$, then we call f a function. For a function f and an element x in its domain of definition, the 'evaluation' gives the value $f(x)$ of x under f. So this may be generalized in the following way: if s, R are fixed, and there is exactly one t, such that: $\langle s, t \rangle \in R$, we then write:

$$t := R \Rsh s .$$

Definition 39 (n-ary Relation) If $R \subset A^n$, R is an n-ary relation on A. How to define a two-variable function? Let

us take the arithmetic functions + (addition), and $*$ (multiplication), as examples. They are from $\mathbb{R} \times \mathbb{R}$ to $\mathbb{R}$, are usually infix-written, thus

$$x + y \quad \text{is precisely} \quad \text{add} \lceil\!\rightarrow \langle x, y\rangle\,;$$
$$x * y \quad \text{is precisely} \quad \text{multiply} \lceil\!\rightarrow \langle x, y\rangle\,.$$

Definition 40 (Power Set, Exponential Set) The set of all subsets of a set A, (including the empty set $\emptyset$, and the total set A), is denoted by $\mathbb{P}(A)$, and called the power-set of A. Starting with $\mathbb{P}^1(A) := \mathbb{P}(A)$, iteration of this power-set construction gives

$$\mathbb{P}^{n+1}(A) = \mathbb{P}(\mathbb{P}^n(A))\,, \quad \text{and we add} \quad \mathbb{P}^0(A) = A\,.$$

Consider the set $A = \mathbb{D}_n := \{0, 1, \ldots, n-1\}$, which has cardinality $n \in \mathbb{N}$, then $\mathbb{P}(A)$ has cardinality 2^n. We think "power-set" is an unfortunate terminology, in our opinion. A better one is the "exponential set", and a better notation is 2^A. 2 is, to be sure, the natural base for the discrete mathematics. Or, if (2 being identified with the set $\mathbb{D}_2$ above), $f \in 2^A$ is interpreted as a function from A to $\mathbb{D}_2$, it is surely the indicator function of a subset of A. On the other hand, sets like $A^2, A^3, \ldots,$ should be called the power sets of A. (A to the 2nd, 3rd, power, ...) It is generally known that in naive set theory, it is illegal to talk about "the set of all sets". A convenient rule is that any set cannot be an element of itself: $A \notin A$. For any set A, then $A \neq \{A\}$, and indeed the latter is a singleton set. Now there is no entity as a set containing everything, there are sometimes some entity which can not be a set.

Definition 41 (Ur-element) Sometimes a non-empty set S may be declared as a ground-set, and its elements declared as ur-elements. A mathematical entity x is declared as an ur-element (or individual), if a sentence $y \in x$ is never allowed true. (A mathematical entity is either an ur-element or a set).

An ur-element is just like a sodium ion which is declared decomposable in high-school or freshman Chemistry. This is the most convenient concept for the students.

Definition 42 (Super-structure) Suppose S is a set of ur-elements. We define $S_0 := S$, and then recursively:

$$S_n := S_{n-1} \cup \mathbb{P}(S_{n-1})\,.$$

The set

$$S_\infty := \cup_{j \geq 0} S_j$$

is called the super-structure over the ground-set S. Any $x \in S_\infty$ is assigned a "rank"

$$\min\{j\colon x \in S_j\}\,.$$

This rank is positive, iff x is a set.

Notation 2 *The following are the symbols for logical constructions:*

- *The symbol $\neg$ is read as "the negation of".*
- *The symbol $\wedge$ is read as "and".*
- *The symbol $\vee$ is read as "or".*
- *The symbol $\Rightarrow$ is read as "imply".*
- *The symbol $\exists$ is read as "there exists".*
- *The symbol $\forall$ is read as "for all".*

Theorem 21 *Every concept in mathematics can be explained in terms of set. In mathematical discussion, we take a suitable ground set S, build up the super-structure S_∞, then the simplest kinds of assertions are: "$a \in b$", or "$a = b$", where a and b are elements of S_∞. From these simple sentences more complicated sentences are formed with the aid of logical symbols.*

Example 7 (Group) Consider a quadruple $\langle G, e, \Lambda, *\rangle$, where G is a set, $e, \Lambda, *$ are respectively 0-, 1-, 2-variable functions with a triple $\langle e, \Lambda, *\rangle$, on G. The group axiom for this system is the sentence:

$$(\forall x \in G)(\forall y \in G)(\forall z \in G)((x * (y * z) \\ = ((x * y) * z) \wedge ((e * x) = x) \wedge (\hat{x} * x = e))\,.$$

This is too formal. But I hope by now you will have no real obstacle in understanding and accepting it. We see that if G is contained in the ground-set S_0, then $e, \Lambda, *$ are all in the superstructure S_∞ with ranks 0, 3, 5, respectively. It is important to realize that group theory should be defined as the study of the true-or-false of those sentences involving only the group structure. Suppose $\Phi\colon G \mapsto H$ is group-homomorphism. Then a sentence (about group) true in G will be true in $\Phi[G] \subset H$, after the sentence is transferred (or translated, or reinterpreted), by Φ. Commutativity is an instance:

$$(\forall x \in G)(\forall y \in G)\,, \quad (x * y = y * x)\,.$$

It is possible that G is commutative, and H is not, but the homomorphic image $\Phi[G](\subsetneqq H)$ is commutative. Any sentence (about group) valid in G is valid in $\Phi(G)$.

Theorem 22 (NSA Idea) *The star-extension from the superstructure to the hyper-universe is a "homomorphism".*

The key ideas of NSA, in the superstructure approach is the following:

- From a ground set S_0 of ur-elements, we can build up set-theoretically the super-structure

$$S_\infty := \cup_{j\geq 0} S_j,$$

the "standard universe". All the mathematical entities arising from S_0 are in it.
- Associated with S_0, there is the corresponding non-standard ground-set ${}^\star S_0 = S_0^{\mathfrak{U}}$ of ur-elements. But

$$S_\infty^{\mathfrak{U}} := \cup_{n\geq 0} S_n^{\mathfrak{U}},$$

the superstructure of $S_0^{\mathfrak{U}}$, is not interesting, because most of the entities in it are not very interesting: they are external to us. Instead of it, and inside of it, the hyper-universe (or non-standard universe)

$${}^\star S_\infty := \cup_{n\geq 0} {}^\star S_n,$$

of internal entities, is of central importance in NSA. Entities in $S_\infty^{\mathfrak{U}} \setminus {}^\star S_\infty$ are then called external.
- There is an injective mapping from S_∞ into ${}^\star S_\infty$, called the star-extension. Every standard entity $A \in S_\infty$ is star-mapped to its star-extension ${}^\star A \in {}^\star S$, which is its (non-standard or) hyper version, or star version in the hyper-universe. All these are the (hyper versions of) standard entities. Other entities of the hyper-universe ${}^\star S_\infty$ are internal but non-standard.
- There is a Transfer Principle, which says that the star-embedding is the all-encompassing "injective homomorphism": If a sentence is true in the standard interpretation in terms of the standard universe, then this sentence holds true in the non-standard interpretation in terms of the hyper-universe. By "non-standard interpretation"' we mean that all the constant entities occurring in the sentence are to be replaced by their star-extensions. (So a quantifying clause "$\forall x \in A$" is replaced by "$\forall x \in {}^\star A$").

The Ultra-Power Construction

Note 3 The notation S_0, S_n, S_∞ have the meaning as defined above. (So S_0 is a ground-set of ur-elements). There will be only another set $S_0^{\mathfrak{U}}$ of ur-elements! Any appearance like Z_n, ${}^\star S_n$, shall not be interpreted as part of superstructure construction! $\mathfrak{U}$ is a free ultra-filter on a set I of indices, to which all the phrases "ultimate" refer.

- Let Z_n be the set of "sequences" $f = \langle f_j \rangle$ such that for $j, f_j \in S_n$. The sequence Z_n is increasing with n, with union

$$Z_\infty = \cup_{n\geq 0} Z_n .$$

- For each $n \in \mathbb{N}_0$, on the set Z_n, we define an equivalence-relation $\overset{\mathfrak{U}}{\sim}$ (as before), by: ($f \overset{\mathfrak{U}}{\sim} g$) iff:

$$f_j = g_j \quad \text{for ultimate} \quad j .$$

Obviously we see that attaching a suffix n to the equivalence relation is hardly necessary: $\overset{\mathfrak{U}}{\sim}$ for Z_{n+1} is an extension of $\overset{\mathfrak{U}}{\sim}$ for Z_n.
- So we may define $S_0^{\mathfrak{U}} = \pi_{\mathfrak{U}}(Z_0)$ to be the quotient set of Z_0 under this equivalence relation $\overset{\mathfrak{U}}{\sim}$, i. e., the set of all equivalence-classes of Z_0. Now we declare that $S_0^{\mathfrak{U}}$ be a ground-set of ur-elements, so that a formula like $y \in x$ is for $x \in S_0^{\mathfrak{U}}$ is illegal. We then build up the superstructure over $S_0^{\mathfrak{U}}$ as before:

$$S_\infty^{\mathfrak{U}} := \cup_{n\geq 0} S_n^{\mathfrak{U}}; \quad S_{n+1}^{\mathfrak{U}} := S_n^{\mathfrak{U}} \cup \mathbb{P}(S_n^{\mathfrak{U}}).$$

- The canonical projection from Z_0 into $S_0^{\mathfrak{U}}$ (surjectively) being here denoted by $\pi_{\mathfrak{U}}$, we now recursively extend it to a mapping

$$\pi_{\mathfrak{U}} : f \in Z_n \mapsto \pi_{\mathfrak{U}}(f) \in S_n^{\mathfrak{U}}, \quad \forall n \in \mathbb{N}.$$

After $\pi_{\mathfrak{U}}(g) \in S_{n-1}^{\mathfrak{U}}$ are defined for every $g \in Z_{n-1}$, we then define $\pi_{\mathfrak{U}}(f)$ for $f \in Z_n \setminus Z_{n-1}$ to be

$$\pi_{\mathfrak{U}}(f) := \{\pi_{\mathfrak{U}}(g) : g \in Z_{n-1}, \\ \text{and} \quad g_j \in f_j, \quad \text{for ultimate } j\}.$$

It is easy to see that $\pi_{\mathfrak{U}}$ is well-defined on $S_\infty^{\mathfrak{U}}$, and, by this definition, all over each $S_n^{\mathfrak{U}}$, and thus over $S_\infty^{\mathfrak{U}}$, we have

$$(\pi_{\mathfrak{U}}(f) = \pi_{\mathfrak{U}}(h)) \iff \left(f \overset{\mathfrak{U}}{\sim} h\right).$$

The proof of this (and similar ones below) is tedious but easy consequence of the dichotomy condition of the ultra-filter.
- Finally we have the hyper-universe of internal entities:

$${}^\star S_\infty := \cup_{n\geq 0} {}^\star S_n; \quad {}^\star S_n := \{\pi_{\mathfrak{U}}(f) : f \in Z_n\} \subset S_n^{\mathfrak{U}}.$$

Any set $B \in S_\infty^{\mathfrak{U}} \setminus {}^\star S_\infty$ is called strictly external.
- And we see that (just as what we have done before for $I = \mathbb{N}$):

$$\pi_{\mathfrak{U}}\langle A_j \rangle \cap \pi_{\mathfrak{U}}\langle B_j \rangle = \pi_{\mathfrak{U}}\langle A_j \cap B_j \rangle;$$
$$\pi_{\mathfrak{U}}\langle A_j \rangle \cup \pi_{\mathfrak{U}}\langle B_j \rangle = \pi_{\mathfrak{U}}\langle A_j \cup B_j \rangle;$$
$$\pi_{\mathfrak{U}}\langle A_j \rangle \setminus \pi_{\mathfrak{U}}\langle B_j \rangle = \pi_{\mathfrak{U}}\langle A_j \setminus B_j \rangle;$$

And the set of all internal sets form a Boolean set-ring.

- For $\{f, g, h\} \subset Z_\infty$,

$$\pi_{\mathfrak{U}}(h) = \{\pi_{\mathfrak{U}}(f), \pi_{\mathfrak{U}}(g)\} \quad \text{iff } h_j = \{f_j, g_j\}, \quad \text{for ultimate } j;$$
$$\pi_{\mathfrak{U}}(h) = \langle \pi_{\mathfrak{U}}(f), \pi_{\mathfrak{U}}(g)\rangle \quad \text{iff } h_j = \langle f_j, g_j\rangle, \quad \text{for ultimate } j;$$
$$\pi_{\mathfrak{U}}(h) = \pi_{\mathfrak{U}}(f) \upharpoonright^{\star} \pi_{\mathfrak{U}}(g) \quad \text{iff } h_j = f_j \upharpoonright^{\star} g_j, \quad \text{for ultimate } j.$$

- For any $n \in \mathbb{N}_0$, if $A \in S_n$, it may be identified with the "sequence" with constant term A, i. e., we may consider $S_n \subset Z_n$. Then $\pi_{\mathfrak{U}}(A) \in {}^\star S_n$ is written as ${}^\star A$, and called a standard internal entity, or a star version of A, in the hyper-universe, as was mentioned before. Also note that: though $S_\infty \notin S_\infty$, and ${}^\star S_\infty \notin {}^\star S_\infty$, the star notation is still compatible.

Remark 7 If $b \in A \in S_\infty$, then the standard entity ${}^\star b \in {}^\star A$. That is

$${}^\sigma A := \{{}^\star b \colon b \in A\} \subset {}^\star A.$$

All standard members in ${}^\star A$ are gathered into the set ${}^\sigma A$. Indeed when A is a finite set, ${}^\star A = {}^\sigma A$. But when A is an infinite set, the right side above is properly contained in the left-side set. That is: the set ${}^\star A$ must contain non-standard elements. Then both ${}^\sigma A$ and ${}^\star A \setminus {}^\sigma A$ are external sets.

This is the source of all confusions!

- Internality is transitive in the sense that

$$(x \in y \quad \text{and} \quad y \in {}^\star S_\infty) \Rightarrow (x \in {}^\star S_\infty).$$

- For entities A, B, C in S_∞,

$$A = B \quad \text{iff } {}^\star A = {}^\star B;$$
$$A \in B \quad \text{iff } {}^\star A \in {}^\star B;$$
$$C = \langle A, B\rangle \quad \text{iff } {}^\star C = \langle {}^\star A, {}^\star B\rangle;$$
$$C = A \upharpoonright^{\star} B \quad \text{iff } {}^\star C = {}^\star A \upharpoonright^{\star} {}^\star B;$$

For sets A, B, in $S_\infty \setminus S_0$,

$${}^\star(A \cup B) = {}^\star A \cup {}^\star B;$$
$${}^\star(A \cap B) = {}^\star A \cap {}^\star B;$$
$${}^\star(A \setminus B) = {}^\star A \setminus {}^\star B;$$
$${}^\star(A \times B) = {}^\star A \times {}^\star B;$$

Theorem 23 (Transfer Principle) *If a sentence is true in the standard universe S_∞, then its star-version is true in the hyper-universe ${}^\star S_\infty$.*

All simple mathematical sentences are assertions about either the identity of two mathematical entities, or the membership of one against the other. By the use of "logical connectives", more complicated assertions are then constructed. So the principle means exactly the whole set of "homomorphisms" displayed above.

Theorem 24 (Standardization Principle) *Let $A \in S_\infty$ be a standard entity, and ${}^\star A \in {}^\star S_\infty$ its star-version. Let $B \in S_\infty^{\mathfrak{U}} \setminus S_0^{\mathfrak{U}}$ be any non-standard set. Then there is $C \in S_\infty$, such that: for any $x \in S_\infty$,*

$$({}^\star x \in {}^\star C) \quad \text{iff} \quad ({}^\star x \in ({}^\star A \cap B)).$$

C is called the standard part of A. The proof is simple: C is the union of all subsets $D \subset A$, such that

$$({}^\star x \in {}^\star D) \Rightarrow ({}^\star x \in ({}^\star A \cap B)).$$

Definition 43 (Concurrence Relation) A relation $R \in S_\infty \setminus S_1$ is called concurrent, if: for any finite number of elements $a_1, a_2, \cdots, a_n$ in the domain of R, there is one $b \in S_\infty$, such that

$$\langle a_i, b\rangle \in R, \quad \text{for} \quad i = 1, 2, \ldots n.$$

Example 8 An obvious example of concurrence relation is the "less than" relation $<$ in the set $\mathbb{N}$. In this case, there is an unlimited hyperinteger ω bigger than all standard integer $a \in \mathbb{N}$, that is: $\forall a \in \mathbb{N}, a < \omega$. All we need, in this particular example, is that the indicial set I has infinite cardinal.

More generally, we need an index set I with big enough cardinal. Then the following Concurrence Theorem is valid. (But the proof is bigger than our concern).

Theorem 25 (Idealization Principle) *Let $R \in S_\infty$ be a concurrent relation. Then there is $b \in {}^\star S_\infty$, such that for all $a \in \mathrm{Dom}(R)$, $\langle a, b\rangle \in R$.*

Note 4 (Internal Set Theory) There are now some axiomatic approach of NSA. In the internal set theory proposed by E. Nelson, the notion of "standardness" is considered as fundamental and undefined, and the three principles of idealization, standardization, and transfer, are taken as axioms, (in addition to the Fraenkel–Zermelo axioms with Axiom of Choice), for the NSA. This is very elegant, though not our choice in this article.

Some Consequences of Transference

Example 9 (Well-ordering in ${}^\star\mathbb{N}$) The axiom of mathematical induction for $\mathbb{N}$ is equivalent to the well-ordering

property of $\mathbb{N}$:

$$((\forall A \subset \mathbb{N}) \wedge (A \neq \emptyset)) \Rightarrow (\exists m \in A\,, (\forall n \in A, m \leq n))\,.$$

"Every non-empty subset A of $\mathbb{N}$ has a minimum element".

The transferred sentence is not:

$$((\forall A \subset {}^\star\mathbb{N}) \wedge (A \neq \emptyset)) \Rightarrow (\exists m \in A\,, (\forall n \in A\,, m \leq n))\,.$$

"Every non-empty subset A of ${}^\star\mathbb{N}$ has a minimum element". Instead, the correct "translation" of this sentence is "Every non-empty internal subset A of ${}^\star\mathbb{N}$ has a minimum element". Why? We may write the original premise as $\forall A \in \mathbb{P}(\mathbb{N})$, then, in the transference, not only the set $\mathbb{N}$ is star-extended to ${}^\star\mathbb{N}$, but also its power-set is star-extended to the standard set (in non-standard version), ${}^\star\mathbb{P}(\mathbb{N})$.

$$((\forall A\, {}^\star\!\subset {}^\star\mathbb{P}(\mathbb{N})) \wedge (A \neq \emptyset)) \Rightarrow (\exists m \in A\,, (\forall n \in A\,, m \leq n))\,.$$

Another way of thinking is to star-transform the original premise "$\forall A \subset \mathbb{N}$" into "$\forall A\, {}^\star\!\subset {}^\star\mathbb{N}$". (That is: the subset-relation gets star-extended.) In the ultra-power representation used before, $A = \pi_{\mathfrak{U}}\langle A_j\rangle$, then take $m_j = \min(A_j) \in \mathbb{N}$, which exists because of the well-ordering of $\mathbb{N}$, and $\mathbb{N} \supset A_j \neq \emptyset$. Then $m := \pi_{\mathfrak{U}}\langle m_j\rangle \in {}^\star\mathbb{N}$ is the minimum of A sought for. The wrong transference gives us the well-ordering property of ${}^\star\mathbb{N}$, which is obviously wrong, because the set of all unlimited hyper natural numbers ${}^\star\mathbb{N}\setminus\mathbb{N}$ would have a minimum element ζ, which is absurd because $\zeta - 1$ would have been still an element of this set.

Corollary 2 *The set $\mathbb{N}$ is an external subset of ${}^\star\mathbb{N}$. So is the complement ${}^\star\mathbb{N} \setminus \mathbb{N}$.*

Example 10 (Order-completeness of ${}^\star\mathbb{R}$) The order-completeness of $\mathbb{R}$ says that: "Every non-empty subset A of $\mathbb{R}$ which is bounded above has a supremum (= sup = least upper bound)".

$$\begin{aligned}&((\forall A \subset \mathbb{R}) \wedge (A \neq \emptyset) \wedge (\exists b \in \mathbb{R}\,, (\forall x \in A\,, x \leq b))\\ &\Rightarrow ((\exists m \in \mathbb{R}\,, (\forall x \in A\,, x \leq b))\\ &\quad \wedge (\forall r \in \mathbb{R}\,, ((\forall x \in A\,, x \leq r) \Rightarrow m \leq r))\,.\end{aligned}$$

The transferred sentence is not the order-completeness of the field ${}^\star\mathbb{R}$: "Every non-empty subset A of ${}^\star\mathbb{R}$ which is bounded above has a supremum". A ready counterexample is the set of infinitesimals $\mathbb{I}$. $\mathbb{I}$ is certainly bounded above by any strictly positive standard real number. But it does not have a "supremum ζ". Otherwise, ζ could only be positive. If it is a positive real number, then $\frac{\zeta}{2}$ is a smaller upper-bound. If it is a positive infinitesimal, then an infinitesimal, namely $2 * \zeta$ could not be bounded by ζ. (We may also appeal to the theorem that there is only one order -complete field, up to a canonical isomorphism.) Again the correct transference should replace "subset" by "internal subset": "In ${}^\star\mathbb{R}$, every non-empty internal subset A which is bounded above has a sup (= least upper bound)".

Corollary 3 (Overflow and Underflow) *Let A be an internal subset of ${}^\star\mathbb{R}$.*

(1): If for any $n \in \mathbb{N}$, there is a limited $x \in A$, such that $x > n$, then there is unlimited positive $\eta \in A$.
(2): If for any unlimited positive $\zeta \in A$, there is another unlimited $\xi \in A$, such that $\xi < zeta$, then there is a limited positive $\eta \in A$.
(3): Both subsets $\mathbb{R}(\subset {}^\star\mathbb{R})$, and ${}^\star\mathbb{R} \setminus \mathbb{R}$, of the standard set ${}^\star\mathbb{R}$(hyper-version) are external.

Proof

(1) We may assume that A is bounded above. But then with $\xi = \sup(A) \in {}^\star\mathbb{R}$, ξ is unlimited. Therefore by the definition of lub, there exists an $\eta \in A$, with $\frac{\xi}{2} \leq \eta \leq \xi$.
(2) Let z be the infimum (= inf = greatset lower bound) of the subset of positive elements of A, then z is limited. And by the definition of inf, there is $\eta \in A$, such that $z \leq \eta \leq z + 1$. □

Theorem 26 (Saturation Principle) *Let $\langle A_n : n \in \mathbb{N}\rangle$ be a sequence of internal sets with the "finitely-intersecting property":*

$$\forall n \in \mathbb{N}\,, \quad \cap_{1 \leq m \leq n} A_m \neq \emptyset\,.$$

Then the sequence is itself intersecting:

$$\cap_{m \in \mathbb{N}} A_m \neq \emptyset\,.$$

(For the proof, the countable indicial set $I = \mathbb{N}$ is enough.) Now for each $n \in \mathbb{N}$, $A_n = \pi_{\mathfrak{U}}\langle A_{n,j} : j \in I\rangle$. And for this fixed $n \in \mathbb{N}$,

$$\cap_{m \in \mathbb{N}} A_m = \pi_{\mathfrak{U}}\langle \cap_{1 \leq m \leq n} A_{m,j} : j \in I\rangle \neq \emptyset \in {}^\star S_\infty\,.$$

Therefore for each $n \in \mathbb{N}$, and for ultimate j, $\cap_{1 \leq m \leq n} A_{m,j} \neq \emptyset$. (In particular, we may and will just assume: for all j, $A_{1,j} \neq \emptyset$). For each $j \in I$, we define

$$k_j := \max\{n \in \mathbb{N} : \cap_{1 \leq m \leq n} A_{m,j} \neq \emptyset\,, \text{ and } n \leq j\}\,;$$

Now for each $j \in I$, choose an $x_j \in \cap_{1 \le m \le k_j} A_{m,j}$. We see that $\xi = \pi_{\mathfrak{U}} \langle x_j \rangle \in A_n$, for any $n \in \mathbb{N}$. Indeed for any $n \in \mathbb{N}$,

$$\{j \in I : x_j \in A_{n,j}\} \supset \{j : k_j \ge n\}$$
$$= \{j \in I : j \ge n\} \cap \{j \in I : \cap_{m \le n} A_{m,j} \ne \emptyset\} \in \mathfrak{U} .$$

Corollary 4 *For a sequence $\langle A_n : n \in \mathbb{N} \rangle$ of internal sets, their (countably infinite) union is internal $\cup_{n \in \mathbb{N}} A_n \in {}^\star S_\infty$, iff the union equals to a finite sub-union. I. e., iff there exists $m \in \mathbb{N}$, such that*

$$\cup_{n \in \mathbb{N}} A_n = \cup_{1 \le n \le m} A_n .$$

Proof If $B := \cup_{n \in \mathbb{N}} A_n \in {}^\star S_\infty$ (is internal), then: all $B_n := B \setminus A_n$ are internal, and $\cap_{n \in \mathbb{N}} B_n = \emptyset$, by de Morgan's law. Then the saturation principle requires that for a certain $m \in \mathbb{N}$, $\cap_{1 \le n \le m} B_n = \emptyset$. Or, $\cap_{1 \le n \le m} (B \setminus A_n) = \emptyset$, and $\cup_{1 \le n \le m} A_n = B = \cup_{n \in \mathbb{N}} A_n$. □

Loeb Construction

In this section we provide the Loeb version of constructing a non-standard version of stochastic analysis (cf. [8]).

Measure

Let a basic non-empty set Ω be fixed. We recall that a subset $\mathcal{A} \subset \mathbb{P}(\Omega)$ is called a Boolean algebra of sets, iff the following requirements are satisfied:

$$\Omega \in \mathcal{A};$$
$$(\forall x \in \mathcal{A})(\forall y \in \mathcal{A}), \quad (x \cup y \in \mathcal{A});$$
$$(\forall x \in \mathcal{A})(\forall y \in \mathcal{A}), \quad (x \cap y \in \mathcal{A});$$
$$(\forall x \in \mathcal{A})(\forall y \in \mathcal{A}), \quad (x \setminus y \in \mathcal{A}).$$

Or the last three sentences can be combined into one:

$$(\forall x \in \mathcal{A})(\forall y \in \mathcal{A}),$$
$$((x \cup y \in \mathcal{A}) \wedge (x \cap y \in \mathcal{A}) \wedge (x \setminus y \in \mathcal{A})).$$

If $\mathcal{A}$ is closed under countable operations, it is then called a σ-algebra on Ω:

$$\forall (x_n : n \in \mathbb{N}),$$
$$((\forall n \in \mathbb{N}, x_n \in \mathcal{A}) \Rightarrow (\cup_{n \in \mathbb{N}} x_n \in \mathcal{A})).$$

It is well-known that for any set $\mathcal{B}$ of subsets of Ω, there is uniquely a smallest σ-algebra $\mathcal{A} \supset \mathcal{B}$. This is the σ-algebra generated by $\mathcal{B} \in \mathbb{P}(\mathbb{P}(\Omega))$.

Definition 44 (Probability Content and Measure) Let $\mathcal{A}$ be a Boolean algebra of sets on a total set Ω. A function

$$\mu : A \in \mathcal{A} \mapsto \mu(A) \in \mathbb{R}$$

is called a content on $(\Omega, \mathcal{A})$, if the following conditions are satisfied:

$$(1) : \forall x \in \mathcal{A}, \quad \mu(x) \ge 0;$$
$$(2) : \forall x \in \mathcal{A}, \quad \forall y \in \mathcal{A}, (x \cap y = \emptyset)$$
$$\Rightarrow (\mu(x \cup y) = \mu(x) + \mu(y)).$$

If the following condition of "the continuity from above at the empty set" is satisfied, then μ is called a measure:

$$\forall (x_n : n \in \mathbb{N}), \quad ((\forall n \in \mathbb{N}, (x_n \in \mathcal{A})$$
$$\wedge (x_{n+1} \subset x_n)) \wedge (\cap_n x_n = \emptyset)) \Rightarrow \left(\lim_{n \to \infty} \mu(x_n) = 0 \right).$$

An equivalent condition is: μ is σ-additive, in the sense that: if $\langle A_n \rangle$ is a disjoint sequence in $\mathcal{A}$, with union $\cup_{n \in \mathbb{N}} A_n \in \mathcal{A}$, then

$$\mu(\cup_{n \in \mathbb{N}} A_n) = \sum_{n \in \mathbb{N}} \mu(A_n).$$

If content (or measure) μ is normalized: $\mu(\Omega) = 1$, then it is called a probability content (or measure, respectively).

Theorem 27 (Extension of a Measure) *If μ is a measure on $(\Omega, \mathcal{A})$, and $\mathcal{B}$ is the σ-algebra generated by $\mathcal{A}$, then there is a unique extension of μ to a measure $\tilde{\mu}$ on $(\Omega, \mathcal{B})$. Further, let $\overline{\mathcal{A}}$ be the collection of sets of the form $A = (B \setminus D) \cup (D \setminus B)$, where $B \in \mathcal{B}$, and $D \subset C$, with $C \in \mathcal{B}$, $\tilde{\mu}(C) = 0$, and define then $\overline{\mu}(A) = \tilde{\mu}(B)$, then the definition is valid, and $\overline{\mu}$ is a measure on $(\Omega, \overline{\mathcal{A}})$. This extension of μ is the "completion of μ". It is "complete" in the sense that whenever $A \in \overline{\mathcal{A}}$, $F \subset A$, and $\overline{\mu}(A) = 0$, then $F \in \overline{\mathcal{A}}$ with $\overline{\mu}(F) = 0$.*

Remark 8 (σ-Finiteness) The definition of a content or measure, is generally extended to allow $+\infty$ in the range. Then in the discussion the condition of σ-finiteness is is generally sufficient and necessary: there is a sequence $A_n \in \mathcal{A}$, with $\cup_{n \in \mathbb{N}} A_n = \Omega$, and $\mu(A_n)$ finite, for all n.

Loeb Measure

Definition 45 (Loeb Content) Let $\Omega \in {}^\star S_\infty$ be an internal set, $\mathcal{A}$ be an internal set of subsets of Ω, which is closed under complementation and finite union. A Loeb content on $(\Omega, \mathcal{A})$ is an internal function $\mu : \mathcal{A} \Rightarrow {}^\star\mathbb{R}$, which is

a content on $\mathcal{A}$, i. e., it is finitely-additive and non-negatively valued. (But it is valued in ${}^\star\mathbb{R}$, not real-valued.)

Theorem 28 (Loeb Measure) *For a Loeb content space $\langle \Omega, \mathcal{A}, \mu \rangle$, define the standard-part of μ as*

$$ {}^\circ\mu \colon A \in \mathcal{A} \mapsto {}^\circ\mu(A) := {}^\circ(\mu(A)) \,. $$

It is then σ-additive, and its completion is the Loeb measure $L(\mu)$ of μ, defined on the Loeb algebra $L(\mathcal{A})$ consisting of those $B \subset \Omega$ which are "μ-approximable": For any positive real ϵ, there are $A \in \mathcal{A}, C \in \mathcal{A}$, such that:

$$ A \subset B \subset C \,, \quad \mu(C) - \mu(A) < \epsilon \,. $$

Indeed there is $D \in \mathcal{A}$, (called Loeb's modification of B), such that

$$ L(\mu)(B \Delta D) = 0 \,. $$

Here Δ is the symmetric-difference:

$$ B \Delta D := (B \setminus D) \cup (D \setminus B) \,. $$

We omit the proof, but by the saturation property of internal sets, the continuity or σ-additivity of μ is trivial. It should be emphasized that $L(\mu)$ is a real-valued measure; therefore the integrands of its integration are real-valued, but defined on Ω, hence is not internal.

Definition 46 (Measurable Functions) If temporarily we denote by $\mathfrak{G}$ the set of all open sets of $\mathbb{R}$, (also called the topology of $\mathbb{R}$), and call elements of ${}^\star\mathfrak{G}$ star-open, then an internal function $F \colon \Omega \mapsto {}^\star\mathbb{R}$ is called $\mathcal{A}$-measurable, iff:

$$ (G \in {}^\star\mathfrak{G}) \Rightarrow \left(F^{-1}(G) \in \mathcal{A}\right) \,. $$

If $F := \pi_{\mathfrak{U}} \langle F_j \rangle \,, \mu = \pi_{\mathfrak{U}} \langle \mu_j \rangle$, we define naturally:

$$ {}^\star\!\int F \mathrm{d}\mu := \pi_{\mathfrak{U}} \left\langle \int F_j \mathrm{d}\mu_j \right\rangle \,. $$

This is valued in ${}^\star\mathbb{R}$.

Remark 9 (Loeb Integral) If F is real-bounded, then

$$ {}^\circ\left({}^\star\!\int F \mathrm{d}\mu\right) = \int {}^\circ F \mathrm{d}L(\mu) \,, $$

which is then called the Loeb integral. The equality here is guaranteed by the S-integrability of the internal function F:

(1): ${}^\star\!\int |F| \mathrm{d}\mu$ is limited.
(2): If $\mu(A) = 0$, then ${}^\star\!\int |F| \mathrm{d}\mu = 0$.
(3): If $F(\omega) = 0$ for all $\omega \in A$, then ${}^\star\!\int |F| \mathrm{d}\mu \approx 0$.

Now we turn to the integration of a real-valued function f on Ω, with respect to the real-valued measure $L(\mu)$.

Theorem 29 (Lifting of Function) *Any $L(\mathcal{A})$-measurable function f has a "lifting" $F \colon \Omega \mapsto {}^\star\mathbb{R}$, in the sense that: F is internal, $\mathcal{A}$-measurable function, satisfying*

$$ L(\mu)\{\omega \colon {}^\circ F(\omega) \neq f(\omega)\} = 0 \,. $$

Definition 47 (Normed Counting Measure) If $\Omega \in {}^\star S_\infty$ is a hyperfinite set, and $\mathcal{A}$ the set of all internal set, the internal probability content μ,

$$ \mu(A) := \frac{{}^\star\mathrm{card}(A)}{{}^\star\mathrm{card}(\Omega)} \,, $$

is called the normed counting measure on Ω.

Example 11 (Hyperfinite Equi-partition of the Unit-Interval) Let $\mathcal{N} \in {}^\star\mathbb{N} \setminus \mathbb{N}$ be an unlimited hyperinteger, $\mathcal{T}$ be the equi-partition hyperset of the unit interval into $\mathcal{N}$ parts, and μ be the normalized uniform counting measure. Then we get the Loeb measure $L(\mu)$ and the Loeb algebra $L(\mathcal{A})$ on $\mathcal{T} \subset {}^\star[0 \ldots 1]$. The Loeb measure space $(\mathcal{T}, L(\mathcal{A}), L(\mu))$ is simply the pull-back of the standard Lebesgue measure space $(\mathbb{U}, \overline{\mathfrak{B}(\mathbb{U})}, \lambda)$ by the standard-part map.

Here we have a dual of push-down and pull-back (or lifting). The push-down is via standard-part map $\mathrm{st} \colon \mathcal{T} \mapsto \mathbb{U}$, which is a surjection, while its inverse st^{-1} furnishes the pull-back. The push-down of $L(\mathcal{A})$, by st, is by definition,

$$ \left\{B \subset \mathbb{U} \colon \mathrm{st}^{-1}[B] \in L(\mathcal{A})\right\} \,. $$

This is precisely the Lebesgue algebra $\overline{\mathfrak{B}(\mathbb{U})}$. And for a member B in this latter algebra, then the push-down of $L(\mu)$ will assign the measure value $L(\mu)(\{x \in \mathcal{T} \colon {}^\star x \in B\})$. But we have already shown that when B is an interval, this value is simply its length $\lambda(B)$. This is enough to identify the pushed-down measure. Actually, the same kind of Loeb construction may be applied to the whole real axis $\mathbb{R}$, or the real half-axis $\mathbb{R}_+$, instead of $\mathbb{U}$ (or the analogous finite intervals). A little bit of attention should be paid to the σ-finiteness, of course.

Anderson's Wiener Process

Definition 48 (Hyperfinite Time-axis) When $\mathcal{N} \in {}^\star\mathbb{N} \setminus \mathbb{N}$ is fixed, $\mathcal{N} = \pi_{\mathfrak{U}} \langle N_j \rangle$, we have the hyperfinite time-axis

$$ \mathcal{T} := \pi_{\mathfrak{U}} \langle T_j \rangle; T_j := \left\{ \frac{k}{N_j} \colon k \in \mathbb{Z} \,,\ 0 \leq k < N_j^2 \right\} \,. $$

As before, T_j is the interval $[0 \dots N_j)$ equi-partitioned into N_j^2 parts, each with length $1/N_j$. The right-end N_j is deliberately deleted, so that $\mathrm{card}(T_j) = N_j^2$, while the augmented set is $T_j' := T_j \cup \{N_j\}$. The set T_j will be called rank k (discretized) time-axis. The left- and the right- approximations from the semi-axis $[0 \cdots +\infty)$ to $\mathcal{T}$ is well-defined:

$$\begin{cases} \mathrm{floor}_{\mathcal{T}}(c) = \pi_{\mathfrak{U}} \left\langle \dfrac{\mathrm{floor}(c * N_j)}{N_j} : j \in I \right\rangle; \\ \mathrm{ceil}_{\mathcal{T}}(c) = \pi_{\mathfrak{U}} \left\langle \dfrac{\mathrm{ceil}(c * N_j)}{N_j} : j \in I \right\rangle. \end{cases}$$

Definition 49 (Hyperfinite Coin-Tossing) With $\mathcal{N} \in {}^\star\mathbb{N} \setminus \mathbb{N}$, and $\mathcal{T} = \left\{ \frac{k}{\mathcal{N}} : k = 0, 1, 2, \cdots, \mathcal{N} - 1 \right\}$ as above, we define:

$$\begin{aligned} \mathrm{Coin} &:= \{-1, +1\}; \\ \Omega_j &:= \mathrm{Coin}^{T_j}; \\ \Omega &:= \pi_{\mathfrak{U}} \langle \Omega_j \rangle; \end{aligned}$$

Of course, +1 or −1 means "head" or "tail" of a coin toss. Any element $\omega_j \in \Omega_j$ is called a coin-particle of rank j, and an element $\omega \in \Omega$ is called a coin-particle (of hyper-finite rank). The normalized counting measure P_j on Ω_j is simple indeed, as is the one occurring most often in our high-school textbooks. (We will denote by $\mathfrak{E}_{P_j}$ the expectation with respect to P_j). The normalized counting measure P is defined on the algebra of all internal subsets of the hyper-finite internal set Ω. Therefore the Loeb construction give us the Loeb measure $L(P)$, called the hyperfinite coin-measure, on the Loeb algebra $L(\mathcal{A})$, which is called the hyper-finite coin algebra.

Definition 50 (The (Rank *j*) Random Walker) For fixed j, and any rank j coin-particle $\omega_j \in \mathrm{Coin}^{T_j}$, its path is the mapping $t \mapsto B_j(\omega_j, t) \in \mathbb{R}$, defined as follows. The domain of definition is

$$T_j' := \left\{ \frac{k}{N_j} : k = 0, 1, 2, \dots, N_j^2 \right\}.$$

And when $t = \frac{k}{N_j}$,

$$B_j(\omega_j, t) := \frac{1}{\sqrt{N_j}} \sum \left\{ \omega_j \left(\frac{m}{N_j} \right) : 0 \le m < k \right\}.$$

By definition, the walk starts from the origin $B_j(\omega_j, 0) := 0$. Then recursively, at each time $t = m/N_j$, when the (j-) particle is located at the position $B_j(\omega_j, t)$, the coin- tossing $\omega_j(t)$ will decide the plus or minus direction of the next step, occurring in the time-interval, from t to $t + \Delta_j t$. Here the rank j

$$\text{time-unit is } \Delta_j t := \frac{1}{N_j};$$
$$\text{space-unit is } \Delta_j x := \frac{1}{\sqrt{N_j}}.$$

As a coin-particle $\omega_j \in \mathrm{Coin}^{T_j}$ is randomly chosen, we may call it a random walker.

Definition 51 (Hyperfinite Random Walk) Define entities in the hyper-universe:

$$\begin{aligned} B &:= \pi_{\mathfrak{U}} \langle B_j \rangle; \\ B(\omega, t) &:= \sum \left\{ \frac{\omega(s)}{\sqrt{\mathcal{N}}} : s < t \right\}; \end{aligned}$$

B: $\langle \omega, t \rangle \in \Omega \times \mathcal{T} \mapsto B(\omega, t) \in {}^\star\mathbb{R}$ is called the hyperfinite random walk.

Lemma 8 (Independent Increment (Finite-Rank)) *For fixed j, if there are $2n$ elements of T_j', arranged according to order:*

$$s_1 < t_1 \le s_2 < t_2 \le \cdots \le s_n < t_n.$$

Then these n increments:

$$B_j(\cdot, t_1) - B_j(\cdot, s_1), B_j(\cdot, t_2) - B_j(\cdot, s_2), \cdots, B_j(\cdot, t_n) - B_j(\cdot, s_n),$$

are independent, in the sense that: for any choice of intervals (or Borel sets) $A_1, A_2, \dots, A_n$,

$$\begin{aligned} P_j\{\omega_j : B_j(\omega_j, t_k) - B_j(\omega_j, s_k) \in A_k, \forall k = 1, 2, \cdots, n, \} \\ = \prod_{1 \le k \le n} P_j\{\omega_j : B_j(\omega_j, t_k) - B_j(\omega_j, s_k) \in A_k\}. \end{aligned}$$

Proof Coin-tossings are independent, by assumption. By transference, □

Lemma 9 (Independent Increment (Hyperfinite-Rank)) *If $2n$ elements of $\mathcal{T}$ are chosen and arranged according to order:*

$$s_1 < t_1 \le s_2 < t_2 \le \cdots \le s_n < t_n.$$

Then the increments:

$$B(\cdot, t_1) - B(\cdot, s_1), \quad B(\cdot, t_2) - B(\cdot, s_2), \cdots, B(\cdot, t_n) - B(\cdot, s_n),$$

are ★-independent, in the sense that: for any choice of internal sets $\tilde{A}_1, \tilde{A}_2, \ldots, \tilde{A}_n$,

$$P\{\omega\colon B(\omega, t_k) - B(\omega, s_k) \in \tilde{A}_k\,,\ \textit{when}\ k = 1, 2, \cdots, n\} = \prod_{1 \le k \le n} P\{\omega\colon B(\omega, t_k) - B(\omega, s_k) \in \tilde{A}_k\}\,.$$

Lemma 10 (Fourier and Moments Characteristic, Finite-Rank) *For* $t \in T_j$, *and* $y \in \mathbb{R}$,

$$\mathfrak{E}_{P_j}\left(e^{\sqrt{-1}\theta * B_j(t)}\right) = \left(\cos\left(\frac{y}{\sqrt{N_j}}\right)\right)^{\frac{t}{N_j}}\,;$$

$$\mathfrak{E}_{P_j}(B_j(t)) = 0\,;$$

$$\mathfrak{E}_{P_j}(B_j(t))^2 = t\,;$$

$$\mathfrak{E}_{P_j}(B_j(t))^4 = 3t^2 - 2\frac{t}{N_j}\,.$$

Proof To calculate $\mathfrak{E}_{P_j}\left(\mathrm{e}^{\sqrt{-1}\theta * B_j(t)}\right)$, by independence, we need only calculate the case of only one-step, i. e., for $t = \frac{1}{N_j}$. Then it is

$$\frac{1}{2}\left(\mathrm{e}^{\sqrt{-1}\frac{y}{\sqrt{N_j}}} + \mathrm{e}^{\sqrt{-1}\frac{-y}{\sqrt{N_j}}}\right) = \cos\left(\frac{y}{\sqrt{N_j}}\right)\,.$$

The odd moments are zero for the even-symmetry of the distribution. Then we do, for the 2nd moment,

$$(u_1 + u_2 + u_3 + \cdots)^2 = \sum u_k^2 + 2\sum_{k<l} u_k u_l\,.$$

Also: for the 4th moment,

$$(u_1 + u_2 + \cdots)^4 = \sum u_k^4 + 6\sum_{k<l} u_k^2 u_l^2 + \text{odd terms}\,.$$

By transference, then: □

Lemma 11 (Fourier and Moments Characteristic, Hyperfinite-Rank) *For any* $t \in \mathcal{T}$, *and for any* $y \in \mathbb{R}$,

$$\mathfrak{E}_P\left(e^{\sqrt{-1}y * B(t)}\right) = \left(\cos\left(\frac{y}{\sqrt{\mathcal{N}}}\right)\right)^{\frac{t}{\mathcal{N}}}\,;$$

$$\mathfrak{E}_P(B(t)) = 0\,;$$

$$\mathfrak{E}_P(B(t))^2 = t\,;$$

$$\mathfrak{E}_P(B(t))^4 = 3t^2 - 2\frac{t}{\mathcal{N}}\,;$$

Theorem 30 (Anderson) *The function b defined below is a Wiener process.*

$$b(\omega, t) := \mathrm{st}(B(\omega, \mathrm{ceil}_{\mathcal{T}}(t)))\,; \quad b\colon \Omega \times [0 \ldots \infty) \mapsto \ddot{\mathbb{R}}\,.$$

- As to the independence of the increments:

$$b(\cdot, t_1) - b(\cdot, s_1)\,, \quad b(\cdot, t_2) - b(\cdot, s_2)\,, \cdots, \quad b(\cdot, t_n) - b(\cdot, s_n)\,,$$

  this utilizes the Loeb's modification, and the standard-part transformation between b and B.
- As to the Gaussian nature of the increment $b(\cdot, t) - b(\cdot, s)$, which is distributed just like $b(\cdot, t - s)$, calculation of its Fourier characteristic is the most convenient way of proof. Now we have (by calculus, L'Hospital or not):

$$\lim_{n\to\infty} \cos^{\mathrm{floor}(nt)}\left(\frac{y}{\sqrt{n}}\right) = \mathrm{e}^{-\frac{y^2}{2}t}\,.$$

  Therefore

$$\int \mathrm{e}^{\sqrt{-1}yb(t)}\mathrm{d}L(P) = \mathrm{e}^{\frac{-y^2}{2}t}\,.$$

- As to the continuity of sample-paths, the usual method is to calculate the moments of increment. The discussion we will omit here.

A Future for Non-standard Analysis?

The ultrapower construction above is the semantic approach. Another approach is called the *syntactic approach* to non-standard analysis requires much less logic and model theory to understand and use. This approach was developed in the mid-1970s by the mathematician Edward Nelson [9,10], using a theory called the *internal set theory* which has a unary operation "standard" that makes non-standard analysis possible. Despite the elegance exhibited in non-standard analysis, some claim that it is not very useful and can only do reinterpretations or reproofs of previous results like [4,14]. In the study of random walks and such, NSA still find some applications, such as [1].

Bibliography

1. Albeverio S, Fenstad JE, Hoegh-Krohn R, Lindstrøm T (1986) Nonstandard Methods in Stochastic Analysis and Mathematical Physics. Bull Am Math Soc 17(2):385–389
2. Bernstein A, Robinson A (eds) (1966) Solution of an invariant subspace problem of KT Smith and PR Halmos. Pacific J Math 16(3):421–431
3. Halmos P (1966) Invariant subspaces for Polynomially Compact Operators. Pacific J Math 16(3):433–437
4. Kamae T (1982) A simple proof of the ergodic theorem using nonstandard analysis. Israel J Math 42(4):284–290
5. Keisler HJ (1986) An Infinitesimal Approach to Stochastic Analysis. J Symb Logic 51(3):822–824 now included In: (1984) Mem Amer Math Soc 297

6. Keisler HJ (1986) Elementary Calculus: An Approach Using Infinitesimals, 2nd edn. Now available at http://www.math.wisc.edu/~keisler/calc.html
7. Kelley JL (1975) General Topology. Springer, New York
8. Loeb PA, Wolff MPH (2000) Nonstandard Analysis for the Working Mathematician. Math. Appl., vol. 510; Kluwer, Dordrecht
9. Nelson E (1977) Internal Set Theory A New Approach to Nonstandard Analysis. Bull Am Math Soc 83(6):1165–1198; Also see http://www.math.princeton.edu/~nelson/books/1.pdf
10. Nelson E (1989) Radically Elementary Probability Theory. Bull Am Math Soc 20(2):240–243
11. Robert A (1985) Nonstandard Analysis. Bull Am Math Soc 16(2):298–306
12. Robinson A (1966) Non-standard analysis, rev edn. North Holland, Amsterdam
13. Schmieden C, Laugwitz D (1958) Eine Erweiterung der Infinitesimalrechnung. Math Zeit 69:1–39
14. van den Dries L, Wilkie AJ (1984) Gromov's Theorem on Groups of Polynomial Growth and Elementary Logic. J Algebra 89:349–374

Normal Forms in Perturbation Theory

Henk W. Broer
University of Groningen, Groningen, The Netherlands

Article Outline

Glossary
Definition of the Subject
Introduction
Motivation
The Normal Form Procedure
Preservation of Structure
Semi-local Normalization
Non-formal Aspects
Applications
Future Directions
Bibliography

Glossary

Normal form procedure This is the stepwise 'simplification' by changes of coordinates, of the Taylor series at an equilibrium point, or of similar series at periodic or quasi-periodic solutions.

Preservation of structure The normal form procedure is set up in such a way that all coordinate changes preserve a certain appropriate structure. This applies to the class of **Hamiltonian** or **volume preserving** systems, as well as to systems that are **equivariant** or **reversible** with respect to a symmetry group. In all cases the systems may also depend on **parameters**.

Symmetry reduction The truncated normal form often exhibits a toroidal symmetry that can be factored out, thereby leading to a lower dimensional reduction.

Perturbation theory The attempt to extend properties of the (possibly reduced) normal form truncation, to the full system.

Definition of the Subject

Nonlinear dynamical systems are notoriously hard to tackle by analytic means. One of the few approaches that has been effective for the last couple of centuries, is Perturbation Theory. Here systems are studied, which in an appropriate sense, can be seen as perturbations of a given system with 'well-known' dynamical properties. Such 'well-known' systems usually are systems with a great amount of symmetry (like integrable Hamiltonian systems [1]) or very low-dimensional systems. The methods of Perturbation Theory then try to extend the 'well-known' dynamical properties to the perturbed system. Methods to do this are often based on the Implicit Function Theorem, on normal hyperbolicity [85,87] or on Kolmogorov–Arnold–Moser theory [1,10,15,18,38].

To obtain a perturbation theory set-up, normal form theory is a vital tool. In its most elementary form it amounts to 'simplifying' the Taylor series of a dynamical system at an equilibrium point by successive changes of coordinates. The lower order truncation of the series often belongs to the class of 'well-known' systems and by the Taylor formula the original system then locally can be viewed as a perturbation of this 'well-known' truncation, that thus serves as the 'unperturbed' system. More involved versions of normal form theory exist at periodic or quasi-periodic evolutions of the dynamical system.

Introduction

We review the formal theory of normal forms of dynamical systems at equilibrium points. Systems with continuous time, i. e., of vector fields, or autonomous systems of ordinary differential equations, are considered extensively. The approach is universal in the sense that it applies to many cases where a structure is preserved. In that case also the normalizing transformations preserve this structure. In particular this includes the Hamiltonian and the volume preserving case as well as cases that are equivariant or reversible with respect to a symmetry group. In all situations the systems may depend on parameters. Related topics are being dealt with concerning a vector field at a periodic solution or a quasi-periodic invariant torus as well as the case of a diffeomorphism at a fixed point. The paper

is concluded by discussing a few non-formal aspects and some applications.

Motivation

The term 'normal form' is widely used in mathematics and its meaning is very sensitive for the context. In the case of linear maps from a given vector space to itself, for example, one may consider all possible choices of a basis. Each choice gives a matrix-representation of a given linear map. A suitable choice of basis now gives the well-known Jordan canonical form. This normal form, in a simple way displays certain important properties of the linear map, concerning its spectrum, its eigenspaces, and so on.

Presently the concern is with dynamical systems, such as vector fields (i. e., systems of autonomous ordinary differential equations), or diffeomorphisms. The aim is to simplify these systems near certain equilibria and (quasi-periodic) solutions by a proper choice of coordinates. The purpose of this is to better display the local dynamical properties. This normalization is effected by a stepwise simplification of formal power series (such as Taylor series).

Reduction of Toroidal Symmetry

In a large class of cases, the simplification of the Taylor series induces a toroidal symmetry up to a certain order, and truncation of the series at that order gives a local approximation of the system at hand. From this truncation we can reduce this toroidal symmetry, thereby also reducing the dimension of the phase space. This kind of procedure is reminiscent of the classical reductions in classical mechanics related to the Noether Theorem, compare with [1], ► Dynamics of Hamiltonian Systems.

This leads us to a first perturbation problem to be considered. Indeed, by truncating and factoring out the torus symmetry we get a polynomial system on a reduced phase, and one problem is how persistent this system is with respect to the addition of higher order terms. In the case where the system depends on parameters, the persistence of a corresponding bifurcation set is of interest.

In many examples the reduced phase space is 2-dimensional where the dynamics is qualitatively determined by a polynomial and this perturbation problem can be handled by Singularity Theory. In quite a number of cases the truncated and reduced system turns out to be structurally stable [67,86].

A Global Perturbation Theory

When returning to the original phase space we consider the original system as a perturbation of the de-reduced truncation obtained so far. In the ensuing perturbation problem, several types of resonance can play a role.

A classical example [2,38] occurs when in the reduced model a Hopf bifurcation takes place and we have reduced by a 1-torus. Then in the original space generically a Hopf–Neĭmark–Sacker bifurcation occurs, where in the parameter space the dynamics is organized by resonance *tongues*. In cases where we have reduced by a torus of dimension larger than 1, we are dealing with quasi-periodic Hopf bifurcation, in which the bifurcation set gets 'Cantorized' by Diophantine conditions, compare with [6,15,21,38]. Also the theory of homo- and heteroclinic bifurcations then is of importance [68].

We use the two perturbation problems as a motivation for the Normal Form Theory to be reviewed, for details, however, we just refer to the literature. For example, compare with ► Perturbation Theory and references therein.

The Normal Form Procedure

The subject of our interest is the simplification of the Taylor series of a vector field at a certain equilibrium point. Before we explore this, however, let us first give a convenient normal form of a vector field near a non-equilibrium point.

Theorem 1 (Flow box [81]) *Let the C^∞ vector field X on $\mathbb{R}^n$ be given by $\dot{x} = f(x)$ and assume that $f(p) \neq 0$. Then there exists a neighborhood of p with local C^∞ coordinates $y = (y_1, y_2, \ldots, y_n)$, such that in these coordinates X has the form*

$$\begin{aligned} \dot{y}_1 &= 1 \\ \dot{y}_j &= 0\,, \end{aligned}$$

for $2 \leq j \leq n$.

Such a local chart usually is called a *flowbox* and the above theorem the Flowbox Theorem. A proof simply can be given using a transversal local C^∞ section that cuts the flow of the vector field transversally. For the coordinate y_1 then use the time-parametrization of the flow, while the coordinates y_j, for $2 \leq j \leq n$ come from the section, compare with, e. g., [81].

Background, Linearization

The idea of simplification near an equilibrium goes back at least to Poincaré [69], also compare with Arnold [2]. To be definite, we now let X be a C^∞ vector field on $\mathbb{R}^n$, with the origin as an equilibrium point. Suppose that X has the form $\dot{x} = Ax + f(x)$, $x \in \mathbb{R}^n$, where A is linear and

where $f(0) = 0$, $D_0 f = 0$. The first idea is to apply successive C^∞ changes of coordinates of the form $\mathrm{Id} + P$, with P a homogeneous polynomial of degree $m = 2, 3, \ldots$, 'simplifying' the Taylor series step by step.

The most 'simple' form that can be obtained in this way, is where *all* higher order terms vanish. In that case the normal form is formally linear. Such a case was treated by Poincaré, and we shall investigate this now.

We may even assume to work on $\mathbb{C}^n$, for simplicity assuming that the eigenvalues of A are distinct. A collection $\lambda = (\lambda_1, \ldots, \lambda_n)$ of points in $\mathbb{C}$ is said to be *resonant* if there exists a relation of the form

$$\lambda_s = \langle r, \lambda \rangle ,$$

for $r = (r_1, \ldots, r_n) \in Z^n$, with $r_k \geq 0$ for all k and with $\sum r_k \geq 2$. The *order* of the resonance then is the number $|r| = \sum r_k$. The Poincaré Theorem now reads

Theorem 2 (Formal linearization [2,69]) *If the (distinct) eigenvalues $\lambda_1, \ldots, \lambda_n$ of A have no resonances, there exists a formal change of variables $x = y + O(|y|^2)$, transforming the above vector field X, given by*

$$\dot{x} = Ax + f(x)$$

to

$$\dot{y} = Ay .$$

We include a proof [2], since this will provide the basis for almost all further considerations.

Proof The formal power series $x = y + O(|y|^2)$ is obtained in an inductive manner. Indeed, for $m = 2, 3, \ldots$ a polynomial transformation $x = y + P(y)$ is constructed, with P homogeneous of degree m, which removes the terms of degree m from the vector field. At the end we have to take the composition of all these polynomial transformations.

1. The basic tool for the mth step is the following. Let v be homogeneous of degree m, then if the vector fields $\dot{x} = Ax + v(x) + O(|x|^{m+1})$ and $\dot{y} = Ay$ are related by the transformation $x = y + P(y)$ with P also homogeneous of degree m, then

$$D_x PAx - AP(x) = v(x) .$$

This relation usually is called the *homological equation*, the idea being to determine P in terms of v: by this choice of P the term v can be transformed away.

The proof of this relation is straightforward. In fact,

$$\begin{aligned} \dot{x} &= (\mathrm{Id} + D_y P)Ay \\ &= (\mathrm{Id} + D_y P)A(x - P(x) + O(|x|^{m+1})) \\ &= Ax + \{D_x PAx - AP(x)\} + O(|x|^{m+1}) , \end{aligned}$$

where we used that for the inverse transformation we know $y = x - P(x) + O(|x|^{m+1})$.

2. For notational convenience we introduce the linear operator $\mathrm{ad}A$, the so-called adjoint operator, by

$$\mathrm{ad}A(P)(x) := D_x PAx - AP(x) ,$$

then the homological equation reads $\mathrm{ad}A(P) = v$. So the question is reduced to whether v is in the image of the operator $\mathrm{ad}A$. It turns out that the eigenvalues of $\mathrm{ad}A$ can be expressed in those of A. If $x_1, x_2, \ldots, x_n$ are the coordinates corresponding to the basis $e_1, e_2, \ldots, e_n$, again it is a straightforward computation to show that for $P(x) = x^r e_s$, one has

$$\mathrm{ad}A(P) = (\langle r, \lambda \rangle - \lambda_s)P .$$

Here we use the multi-index notation $x^r = x_1^{r_1} x_2^{r_2} \cdots x_n^{r_n}$. Indeed, for this choice of P one has $AP(x) = \lambda_s P(x)$, while

$$\frac{\partial x^r}{\partial x} Ax = \sum_j \frac{r_j}{x_j} x^r \lambda_j x_j = \langle r, \lambda \rangle x^r .$$

We conclude that the monomials $x^r e_s$ are eigenvectors corresponding to the eigenvalues $\langle r, \lambda \rangle - \lambda_s$. We conclude that the operator $\mathrm{ad}A$ is *semisimple*, since it has a basis of eigenvectors. Therefore, if $\ker \mathrm{ad}A = 0$, the operator is surjective. This is exactly what the non-resonance condition on the eigenvalues of A amounts to. More precisely, the homological equation $\mathrm{ad}A(P) = v$ can be solved for P for each homogeneous part v of degree m, provided that there are no resonances up to order m.

3. The induction process now runs as follows. For $m \geq 2$, given a form

$$\dot{x} = Ax + v_m(x) + O(|x|^{m+1}) ,$$

we solve the homological equation

$$\mathrm{ad}A(P_m) = v_m ,$$

then carrying out the transformation $x = y + P_m(y)$. This takes the above form to

$$\dot{y} = Ay + w_{m+1}(y) + O(|y|^{m+2}) .$$

The composition of all the polynomial transformations then gives the desired formal transformation. □

Remark

- It is well-known that the formal series usually diverge. Here we do not go into this problem, for a brief discussion see below.
- If resonances are excluded up to a finite order N, we can linearize up to that order, so obtaining a normal form.

  $$\dot{y} = Ay + O(|y|^{N+1}) .$$

  In this case the transformation can be taken as a polynomial.
- If the original problem is real, but with the matrix A having non-real eigenvalues, we still can keep all transformations real by also considering complex conjugate eigenvectors.

We conclude this introduction by discussing two further linearization theorems, one due to Sternberg and the other to Hartman–Grobman. We recall the following for a vector field $X(x) = Ax + f(x)$, $x \in \mathbb{R}^n$, with 0 as an equilibrium point, i. e., with $f(0) = 0$, $D_0 f = 0$. The equilibrium 0 is *hyperbolic* if the matrix A has no purely imaginary eigenvalues. Sternberg's Theorem reads

Theorem 3 (Smooth linearization [82]) *Let X and Y be C^∞ vector fields on $\mathbb{R}^n$, with 0 as a hyperbolic equilibrium point. Also suppose that there exists a formal transformation $(\mathbb{R}^n, 0) \to (\mathbb{R}^n, 0)$ taking the Taylor series of X at 0 to that of Y. Then there exists a local C^∞-diffeomorphism $\Phi : (\mathbb{R}^n, 0) \to (\mathbb{R}^n, 0)$, such that $\Phi_* X = Y$.*

We recall that $\Phi_* X(\Phi(x)) = D_x \Phi X(x)$. This means that X and Y are locally *conjugated* by Φ, the evolution curves of X are mapped to those of Y in a time-preserving manner. In particular Sternberg's Theorem applies when the conclusion of Poincaré's Theorem holds: for Y just take the linear part $Y(x) = Ax(\partial/\partial x)$.

Combining these two theorems we find that in the hyperbolic case, under the exclusion of all resonances, the vector field X is linearizable by a C^∞-transformation. The Hartman–Grobman Theorem moreover says that the non-resonance condition can be omitted, provided we only want a C^0-linearization.

Theorem 4 (Continuous linearization e. g., [67]) *Let X be a C^∞ vector field on $\mathbb{R}^n$, with 0 as a hyperbolic equilibrium point. Then there exists a local homeomorphism $\Phi : (\mathbb{R}^n, 0) \to (\mathbb{R}^n, 0)$, locally conjugating X to its linear part.*

Preliminaries from Differential Geometry

Before we develop a more general Normal Form Theory we recall some elements from differential geometry. One central notion used here is that of the *Lie derivative*. For simplicity all our objects will be of class C^∞. Given a vector field X we can take any tensor τ and define its Lie-derivative $\mathcal{L}_X \tau$ with respect to X as the infinitesimal transformation of τ along the flow of X. In this way $\mathcal{L}_X \tau$ becomes a tensor of the same type as τ. To be more precise, for τ a real *function* f one so defines

$$\mathcal{L}_X f(x) = X(f)(x) = \mathrm{d}f(X)(x) = \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} f(X_t(x)) ,$$

i. e., the directional derivative of f with respect to X. Here X_t denotes the flow of X over time t. For τ a *vector field* Y one similarly defines

$$\begin{aligned}\mathcal{L}_X Y(x) &= \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} (X_{-t})_* Y(x) \\ &= \lim_{t\to 0} \frac{1}{t}\{(X_{-t})_* Y(x) - Y(x)\} \\ &= \lim_{h\to 0} \frac{1}{h}\{Y(x) - (X_h)_* Y(x)\} ,\end{aligned}$$

and similarly for *differential forms*, etc. Another central notion is the *Lie-brackets* $[X, Y]$ defined for any two vector fields X and Y on $\mathbb{R}^n$ by

$$[X, Y](f) = X(Y(f)) - Y(X(f)) .$$

Here f is any real function on $\mathbb{R}^n$ while, as before, $X(f)$ denotes the directional derivative of f with respect to X.

We recall the expression of Lie-brackets in coordinates. If X is given by the system of differential equations $\dot{x}_j = X_j(x)$, with $1 \le j \le n$, then the directional derivative $X(f)$ is given by $X(f) = \sum_{j=1}^n X_j \partial f/\partial x_j$. Then, if $[X, Y]$ is given by the system $\dot{x}_j = Z_j(x)$, for $1 \le j \le n$, of differential equations, one directly shows that

$$Z_j = \sum_{k=1}^n \left(X_k \frac{\partial Y_j}{\partial x_k} - Y_k \frac{\partial X_j}{\partial x_k} \right) .$$

Here Y_j relates to Y as X_j does to X. We list some useful properties.

Proposition 5 (Properties of the Lie-derivative [81])

1. $\mathcal{L}_X(Y_1 + Y_2) = \mathcal{L}_X Y_1 + \mathcal{L}_X Y_2$ *(linearity over $\mathbb{R}$)*
2. $\mathcal{L}_X(fY) = X(f) \times Y + f \times \mathcal{L}_X Y$ *(Leibniz rule)*
3. $[Y, X] = -[X, Y]$ *(skew symmetry)*
4. $[[X, Y], Z] + [[Z, X], Y] + [[Y, Z], X] = 0$ *(Jacobi identity)*
5. $\mathcal{L}_X Y = [X, Y]$
6. $[X, Y] = 0 \Leftrightarrow X_t \circ Y_s = Y_s \circ X_t$

Proof The first four items are left to the reader.

5. The equality $\mathcal{L}_X Y = [X, Y]$ can be proven by observing the following. Both members of the equality are defined intrinsically, so it is enough to check it in any choice of (local) coordinates. Moreover, we can restrict our attention to the set $\{x | X(x) \neq 0\}$. By the Flowbox Theorem 1 we then may assume for the component functions of X that $X_1(x) = 1, X_j(x) = 0$ for $2 \leq j \leq n$. It is easy to see that both members of the equality now are equal to the vector field Z with components

$$Z_j = \frac{\partial Y_j}{\partial x_1} .$$

6. Remains the equivalence of the commuting relationships. The commuting of the flows, by a very general argument, implies that the bracket vanishes. In fact, fixing t we see that X_t conjugates the flow of Y to itself, which is equivalent to $(X_t)_* Y = Y$. By definition this implies that $\mathcal{L}_X Y = 0$.

Conversely, let $c(t) = ((X_t)_* Y)(p)$. From the fact that $\mathcal{L}_X Y = 0$ it then follows that $c(t) \equiv c(0)$. Observe that the latter assertion is sufficient for our purposes, since it implies that $(X_t)_* Y = Y$ and therefore $X_t \circ Y_s = Y_s \circ X_t$. Finally, that $c(t)$ is constant can be shown as follows.

$$\begin{aligned} c'(t) &= \lim_{h \to 0} \frac{1}{h} \{c(t+h) - c(t)\} \\ &= \lim_{h \to 0} \frac{1}{h} \{((X_{t+h})_* Y)(p) - ((X_t)_* Y)(p)\} \\ &= (X_t)_* \lim_{h \to 0} \frac{1}{h} \{((X_h)_* Y)(X_{-t}(p)) - Y(X_{-t}(p))\} \\ &= (X_t)_* \mathcal{L}_X Y(X_{-t}(p)) \\ &= (X_t)_*(0) \\ &= 0 . \end{aligned}$$

□

'Simple' in Terms of an Adjoint Action

We now return to the setting of Normal Form Theory at equilibrium points. So given is the vector field $\dot{x} = X(x)$, with $X(x) = Ax + f(x), x \in \mathbb{R}^n$, where A is linear and where $f(0) = 0, D_0 f = 0$. We recall that it is our general aim to 'simplify' the Taylor series of X at 0.

However, we have not yet said what the word 'simple' means in the present setting. In order to understand what is going on, we reintroduce the adjoint action associated to the linear part A, defined on the class of all C^∞ vector fields on $\mathbb{R}^n$. To be precise, this adjoint action ad$A$ is defined by the Lie-bracket

$$\mathrm{ad}A \colon Y \mapsto [A, Y] ,$$

where A is identified with the linear vector field $\dot{x} = Ax$. It is easily seen that this fits with the notation introduced in Theorem 2.

Let $H^m(\mathbb{R}^n)$ denote the space of polynomial vector fields, homogeneous of degree m. Then the Taylor series of X can be viewed as an element of the product $\prod_{m=1}^{\infty} H^m(\mathbb{R}^n)$. Also, it directly follows that ad$A$ induces a linear map $H^m(\mathbb{R}^n) \to H^m(\mathbb{R}^n)$, to be denoted by $\mathrm{ad}_m A$. Let

$$B^m := \mathrm{im}\, \mathrm{ad}_m A ,$$

the image of the map $\mathrm{ad}_m A$ in $H^m(\mathbb{R}^n)$. Then for *any* complement G^m of B^m in $H^m(\mathbb{R}^n)$, in the sense that

$$B^m \oplus G^m = H^m(\mathbb{R}^n) ,$$

we define the *corresponding* notion of 'simpleness' by requiring the homogeneous part of degree m to be in G^m. In the case of the Poincaré Theorem 2, since $B^m = H^m(\mathbb{R}^n)$, we have $G^m = \{0\}$.

We now quote a theorem from Sect. 7.6.1. in Dumortier et al. [16]. Although its proof is very similar to the one of Theorem 2, we include it here, also because of its format.

Theorem 6 ('Simple' in terms of G^m [74,82]) *Let X be a C^∞ vector field, defined in the neighborhood of $0 \in \mathbb{R}^n$, with $X(0) = 0$ and $D_0 X = A$. Also let $N \in \mathbb{N}$ be given and, for $m \in \mathbb{N}$, let B^m and G^m be such that $B^m \oplus G^m = H^m(\mathbb{R}^n)$. Then there exists, near $0 \in \mathbb{R}^n$, an analytic change of coordinates $\Phi \colon \mathbb{R}^n \to \mathbb{R}^n$, with $\Phi(0) = 0$, such that*

$$\Phi_\star X(y) = Ay + g_2(y) + \cdots + g_N(y) + O(|y|^{N+1}) ,$$

with $g_m \in G^m$, for all $m = 2, 3, \ldots, N$.

Proof We use induction on N. Let us assume that

$$X(x) = Ax + g_2(x) + \cdots + g_{N-1}(x) + f_N(x) + O(|x|^{N+1}) ,$$

with $g_m \in G^m$, for all $m = 2, 3, \ldots, N-1$ and with f_N homogeneous of degree N.

We consider a coordinate change $x = y + P(y)$, where P is polynomial of degree N, see above. For any such P, by substitution we get

$$(\mathrm{Id} + D_y P)\dot{y} = A(y + P(y)) + g_2(y) + \cdots + g_{N-1}(y) + f_N(y) + O(|y|^{N+1}) ,$$

or

$$\begin{aligned}\dot{y} &= (\mathrm{Id} + D_y P)^{-1}(A(y + P(y)) + g_2(y) \\ &\quad + \cdots + g_{N-1}(y) + f_N(y) + O(|y|^{N+1})) \\ &= Ay + g_2(y) + \cdots + g_{N-1}(y) + f_N(y) + AP(y) \\ &\quad - D_y PAy + O(|y|^{N+1}) ,\end{aligned}$$

using that $(\mathrm{Id} + D_y P)^{-1} = \mathrm{Id} - D_y P + O(|y|^N)$. We conclude that the terms up to order $N-1$ are unchanged by this transformation, while the Nth order term becomes

$$f_N(y) - \mathrm{ad}_N A(P)(y) .$$

Clearly, a suitable choice of P will put this term in G^N. This is the present version of the *homological equation* as introduced before. □

Remark For simplicity the formulations all are in the C^∞-context, but obvious changes can be made for the case of finite differentiability. The latter case is of importance for applications of Normal Form Theory after reduction to a center manifold [85,87].

Torus Symmetry

As a special case let the linear part A be semisimple, in the sense that it is diagonizable over the complex numbers. It then directly follows that also $\mathrm{ad}_m A$ is semisimple, which implies that

$$\mathrm{im}\,\mathrm{ad}_m A \oplus \ker \mathrm{ad}_m A = H^m(\mathbb{R}^n) .$$

The reader is invited to provide the eigenvalues of $ad_m A$ in terms of those of A, compare with the proof of Theorem 2. In the present case the obvious choice for the complementary space defining 'simpleness' is $G^m = \ker \mathrm{ad}_m A$. Moreover, the fact that the normalized, viz. simplified, terms g_m are in G^m by definition means that

$$[A, g_m] = 0 .$$

This, in turn, implies that N-jet of $\Phi_\star X$, i. e., the *normalized part* of $\Phi_\star X$, by Proposition 5 is invariant under all linear transformations

$$\exp tA, \quad t \in \mathbb{R}$$

generated by A. For further reading also compare with Sternberg [82], Takens [84], Broer [7,8] or [12].

More generally, let $A = A_s + A_n$ be the Jordan canonical splitting in the semisimple and nilpotent part. Then one directly shows that $\mathrm{ad}_m A = \mathrm{ad}_m A_s + \mathrm{ad}_m A_n$ is the Jordan canonical splitting, whence, by a general argument [89] it follows that

$$\mathrm{im}\,\mathrm{ad}_m A + \ker \mathrm{ad}_m A_s = H^m(\mathbb{R}^n) ,$$

so where the sum splitting in general no longer is direct. Now we can choose the complementary spaces G^m such that

$$G^m \subset \ker \mathrm{ad}_m A_s ,$$

ensuring equivariance of the normalized part of $\Phi_\star X$, with respect to all linear transformations

$$\exp tA_s, \quad t \in \mathbb{R} .$$

The choice of G^m can be further restricted, e. g., such that

$$G^m \subseteq \ker \mathrm{ad}_m A_s \setminus \mathrm{im}\,\mathrm{ad}_m A_n \subseteq H^m(\mathbb{R}^n) ,$$

compare with Van der Meer [88]. For further discussion on the choice of G^m, see below.

Example (Rotational symmetry [84]*)* Consider the case $n = 2$ where

$$A = A_s = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} .$$

So, the eigenvalues of A are $\pm i$ and the transformations $\exp tA_s$, $t \in \mathbb{R}$, form the rotation group $\mathrm{SO}(2, \mathbb{R})$. From this, we can arrive at once at the general format of the normal form. In fact, the normalized, Nth order part of $\Phi_\star X$ is rotationally symmetric. This implies that, if we pass to polar coordinates (r, φ), by

$$y_1 = r\cos\varphi, \quad y_2 = r\sin\varphi ,$$

the normalized truncation of $\Phi_\star X$ obtains the form

$$\begin{aligned}\dot{\varphi} &= f(r^2) \\ \dot{r} &= rg(r^2) ,\end{aligned}$$

for certain polynomials f and g, with $f(0) = 1$ and $g(0) = 0$.

Remark

- A more direct, 'computational' proof of this result can be given as follows, compare with the proof of Theorem 2 and with [16,84,87]. Indeed, in vector field notation we can write

$$\begin{aligned}A &= -x_2 \frac{\partial}{\partial x_1} + x_1 \frac{\partial}{\partial x_2} \\ &= i\left(z\frac{\partial}{\partial z} - \bar{z}\frac{\partial}{\partial \bar{z}}\right) ,\end{aligned}$$

where we complexified putting $z = x_1 + ix_2$ and use the well-known Wirtinger derivatives

$$\frac{\partial}{\partial z} = \frac{1}{2}\left(\frac{\partial}{\partial x_1} - i\frac{\partial}{\partial x_2}\right), \quad \frac{\partial}{\partial \bar{z}} = \frac{1}{2}\left(\frac{\partial}{\partial x_1} + i\frac{\partial}{\partial x_2}\right).$$

Now a basis of eigenvectors for $\mathrm{ad}_m A$ can be found directly, just computing a few Lie-brackets, compare the previous subsection. In fact, it is now given by all monomials

$$z^k \bar{z}^l \frac{\partial}{\partial z}, \quad \text{and} \quad z^k \bar{z}^l \frac{\partial}{\partial \bar{z}},$$

with $k + \ell = m$. The corresponding eigenvalues are $i(k - \ell - 1)$ viz. $i(k - \ell + 1)$. So again we see, now by a direct inspection, that $\mathrm{ad}_m A$ is semisimple and that we can take $G^m = \ker \mathrm{ad}_m A$. This space is spanned by

$$(z\bar{z})^r \left(z\frac{\partial}{\partial z} \pm \bar{z}\frac{\partial}{\partial \bar{z}}\right),$$

with $2r + 1 = m - 1$, which indeed proves that the normal form is rotationally symmetric.

- A completely similar case occurs for $n = 3$, where

$$A = A_s = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

In this case the normalized part in cylindrical coordinates (r, φ, z), given by

$$y_1 = r\cos\varphi, \quad y_2 = r\sin\varphi, \quad y_3 = z,$$

in general gets the axially symmetric form

$$\begin{aligned} \dot{\varphi} &= f(r^2, z) \\ \dot{r} &= rg(r^2, z) \\ \dot{z} &= h(r^2, z), \end{aligned}$$

for suitable polynomials f, g and h.

As a generalization of this example we state the following proposition, where the normalized part exhibits an m-torus symmetry.

Proposition 7 (Toroidal symmetry [84]) *Let X be a C^∞ vector field, defined in the neighborhood of $0 \in \mathbb{R}^n$, with $X(0) = 0$ and where $A = D_0X$ is semisimple with the eigenvalues $\pm i\omega_1, \pm i\omega_2, \ldots, \pm i\omega_m$ and 0. Here $2m \le n$. Suppose that for given $N \in \mathbb{N}$ and all integer vectors $(k_1, k_2, \ldots, k_m)$,*

$$1 \le \sum_{j=1}^{m} |k_j| \le N + 1 \Rightarrow \sum_{j=1}^{m} k_j\omega_j \ne 0, \tag{1}$$

(i. e., there are no resonances up to order $N + 1$). Then there exists, near $0 \in \mathbb{R}^n$, an analytic change of coordinates $\Phi : \mathbb{R}^n \to \mathbb{R}^n$, with $\Phi(0) = 0$, such that $\Phi_\star X$, up to terms of order N has the following form. In suitable generalized cylindrical coordinates $(\varphi_1, \ldots, \varphi_m, r_1^2, \ldots, r_m^2, z_{n-2m+1}, \ldots, z_n)$ it is given by

$$\begin{aligned} \dot{\varphi}_j &= f_j(r_1^2, \ldots, r_m^2, z_{n-2m+1}, \cdots, z_n) \\ \dot{r}_j &= r_j g_j(r_1^2, \ldots, r_m^2, z_{n-2m+1}, \cdots, z_n) \\ \dot{z}_\ell &= h_\ell(r_1^2, \ldots, r_m^2, z_{n-2m+1}, \cdots, z_n), \end{aligned}$$

where $f_j(0) = \omega_j$ and $h_\ell(0) = 0$ for $1 \le j \le m$, $n - 2m + 1 \le \ell \le n$.

Proofs can be found, e. g., in [16,84,85,87]. In fact, if one introduces a suitable complexification, it runs along the same lines as the above remark. For the fact that finitely many non-resonance conditions are needed in order to normalize up to finite order, also compare a remark following Theorem 2.

Since the truncated system of $\dot{r}_j$- and $\dot{z}_\ell$-equations is independent of the angles φ_j, this can be studied separately. A similar remark holds for the earlier examples of this section. As indicated in the introduction, this kind of 'reduction by symmetry' to lower dimension can be of great importance when studying the dynamics of X : as it enables us to consider X, viz. $\Phi_\star X$, as an N-flat perturbation of the normalized part, which is largely determined by this lower dimensional reduction. For more details see below.

On the Choices of the Complementary Space and of the Normalizing Transformation

In the previous subsection we only provided a general (symmetric) format of the normalized part. In concrete examples one has to do more. Indeed, given the original Taylor series, one has to *compute* the coefficients in the normalized expansion. This means that many choices have to be made explicit.

To begin with there is the choice of the spaces G^m, which define the notion of 'simple'. We have seen already that this choice is not unique.

Moreover observe that, even if the choice of G^m has been fixed, still P usually is not uniquely determined. In the semisimple case, for example, P is only determined modulo the kernel $G^m = \ker \mathrm{ad}_m A$.

Remark To fix thoughts, consider the former of the above examples, on $\mathbb{R}^2$, where the normalized truncation has the

rotationally symmetric form

$$\dot{\varphi} = f(r^2)$$
$$\dot{r} = rg(r^2)\,.$$

Here $g(r^2) = cr^2 + O(|r|^4)$. The coefficient c dynamically is important, just think of the case where the system is part of a family that goes through Hopf-bifurcation. The computation of (the sign of) c in a concrete model can be quite involved, as it appears from, e. g., Marsden and McCracken [56]. Machine-assisted methods largely have taken over this kind of work.

One general way to choose G^m is the following, compare with e. g., Sect. 7.6 in [16] and Sect. 2.3 in [87], :

$$G^m := \ker(\mathrm{ad}_m A^{\mathrm{T}})\,.$$

Here A^{T} is the *transpose* of A, defined by the relation $\langle A^{\mathrm{T}}x, y\rangle = \langle x, Ay\rangle$, where $\langle .,.\rangle$ is an inner product on $\mathbb{R}^n$. A suitable choice for an inner product on $H^m(\mathbb{R}^n)$ then directly gives that

$$G^m \oplus \mathrm{im}(\mathrm{ad}_m A) = H^m(\mathbb{R}^n)\,,$$

as required. Also here the normal form can be interpreted in terms of symmetry, namely with respect to the group generated by A^{T}. In the semisimple case, this choice leads to exactly the same symmetry considerations as before.

The above algorithms do not provide methods for computing the normal form yet, i. e., for actually solving the homological equation. In practice, this is an additional computation. Regarding the corresponding algorithms we give a few more references for further reading, also referring to their bibliographies. General work in this direction is Bruno [35,36], Sect. 7.6 in Dumortier et al. [16], Takens [84] or Vanderbauwhede [87], Part 2. In the latter reference also a brief description is given of the sl(2, $\mathbb{R}$)-theory of Sanders and Cushman [39], which is a powerful tool in the case where the matrix A is nilpotent [84]. For a thorough discussion on some of these methods we refer to Murdock [60]. Later on we shall come back to these aspects.

Preservation of Structure

It goes without saying that Normal Form Theory is of great interest in special cases where a given structure has to preserved. Here one may think of a symplectic or a volume form that has to be respected. Also a given symmetry group can have this rôle, e. g., think of an involution related to reversibility. Another, similar, problem is the dependence of external parameters in the system.

A natural language for preservation of such structures is that of Lie-subalgebra's of general Lie-algebra of vector fields, and the corresponding Lie-subgroup of the general Lie-group of diffeomorphisms.

The Lie-Algebra Proof

Fortunately the setting of Theorem 6 is almost completely in terms of Lie-brackets. Let us briefly reconsider its proof.

Given is a C^∞ vector field $X(x) = Ax + f(x), x \in \mathbb{R}^n$, where A is linear and where $f(0) = 0, D_0 f = 0$. We recall that in the inductive procedure a transformation

$$h = \mathrm{Id} + P\,,$$

with P a homogeneous polynomial of degree $m = 2, 3, \ldots$, is found, putting the homogeneous mth degree part of the Taylor series of X into the 'good' space G^m.

Now, nothing changes in this proof if instead of $h = \mathrm{Id} + P$, we take $h = P_1$, the flow over time 1 of the vector field P : indeed, the effect of this change is not felt until the order $2m - 1$. Here we use the following formula for $X^t := (P_t)_\star X$:

$$[X^t, P] = \mathrm{ad}_m A(P) + O(|y|^{m+1}) \quad \text{and} \quad \frac{\partial X^t}{\partial t} = [X^t, P],$$

compare with Sect. "Preliminaries from Differential Geometry". In fact, exactly this choice $h = P_1$ was taken by Roussarie [74], also see the proof of Takens [84]. Notice that $\Phi = h_N \circ h_{N-1} \circ \cdots \circ h_3 \circ h_2$. Moreover notice that if the vector field P is in a given Lie-algebra of vector fields, its time 1 map P_1 is in the corresponding Lie-group. In particular, if P is Hamiltonian, P_t is canonical or symplectic, and so on.

For the validity of this set-up for a more general Lie-subalgebra of the Lie-algebra of all C^∞ vector fields, one has to study how far the *grading*

$$\prod_{m=1}^{\infty} H^m(\mathbb{R}^n)$$

of the formal power series, as well as the *splittings*

$$B^m \oplus G^m = H^m(\mathbb{R}^n)\,,$$

are compatible with the Lie-algebra at hand. This issue was addressed by Broer [7,8] axiomatically, in terms of graded and filtered Lie-algebra's. Moreover, the methods concerning the choice of G^m briefly mentioned at the end of the previous section, all carry over to a more general Lie-algebra set-up. As a consequence there exists a version of Theorem 6 in this setting. Instead of pursuing this further, we discuss its implications in a few relevant settings.

The Volume Preserving and Symplectic Case On $\mathbb{R}^n$, resp. $\mathbb{R}^{2n}$, we consider a volume form or a symplectic form, both denoted by σ. We assume, that

$$\sigma = \mathrm{d}x_1 \wedge \cdots \wedge \mathrm{d}x_n, \quad \text{resp.} \quad \sigma = \sum_{j=1}^{n} \mathrm{d}x_j \wedge \mathrm{d}x_{j+n} .$$

In both cases, let $\mathcal{X}_\sigma$ denote the Lie-algebra of σ-preserving vector fields, i. e., vector fields X such that $\mathcal{L}_X\sigma = 0$. Here $\mathcal{L}$ again denotes the Lie-derivative, see Sect. "The Normal Form Procedure".

Indeed, one defines

$$\begin{aligned}\mathcal{L}_X\sigma(x) &= \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} (X_t)^*\sigma(x) \\ &= \lim_{t\to 0} \frac{1}{t}\{(X_t)^*\sigma(x) - \sigma(x)\} .\end{aligned}$$

Properties, similar to Proposition 5, hold here. Since in both cases σ is a closed form, one shows by 'the magic formula' [81] that

$$\begin{aligned}\mathcal{L}_X\sigma(x) &= \mathrm{d}(\iota_X\sigma) + \iota_X\,\mathrm{d}\sigma \\ &= \mathrm{d}(\iota_X\sigma) .\end{aligned}$$

Here $\iota_X\sigma$ denotes the flux-operator defined by $\iota_X\sigma(Y) = \sigma(X, Y)$. In the volume-preserving case the latter expression denotes $\mathrm{div}(X)\sigma$ and we see that preservation of σ exactly means that $\mathrm{div}(X) = 0$: the divergence of X vanishes. In the Hamiltonian case we conclude that the 1-form $\iota_X\sigma$ is closed and hence (locally) of the form $\mathrm{d}H$, for a Hamilton function H. In both cases, the fact that for a transformation h the fact that $h^*\sigma = \sigma$ implies that with X also h_*X is σ-preserving. Moreover, for a σ-preserving vector field P and $h = P_1$ one can show that indeed $h^*\sigma = \sigma$.

One other observation is, that by the *homogeneity* of the above expressions for σ, the homogeneous parts of the Taylor series of σ-preserving vector fields are again σ-preserving. This exactly means that

$$H^m(\mathbb{R}^{(2)n}) \cap \mathcal{X}_\sigma ,$$

$m = 1, 2, \ldots$, *grades* the formal power series corresponding to $\mathcal{X}_\sigma$. Here, notice that

$$H^1(\mathbb{R}^{(2)n}) \cap \mathcal{X}_\sigma = \mathrm{sl}(n, \mathbb{R}), \quad \text{resp.} \quad \mathrm{sp}(2n, \mathbb{R}) ,$$

the *special*- resp. the *symplectic* linear algebra.

In summary we conclude that both the symplectic and the volume preserving setting are covered by the axiomatic approach of [7,8] and that an appropriate version of Theorem 6 holds here. Below we shall illustrate this with a few examples.

External Parameters A C^∞ family $X = X^\lambda(x)$ of vector fields on $\mathbb{R}^n$, with a multi-parameter $\lambda \in \mathbb{R}^p$, can be regarded as one C^∞ vector field on the product space $\mathbb{R}^n \times \mathbb{R}^p$. Such a vector field is *vertical*, in the sense that it has no components in the λ-direction. In other words, if $\pi : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^p$ is the natural projection on the second component, X is tangent to the fibers of π. It is easily seen that this property defines a Lie-subalgebra of the Lie-algebra of all C^∞ vector fields on $\mathbb{R}^n \times \mathbb{R}^p$. Again, by the linearity of this projection, the gradings and splittings are compatible. The normal form transformations Φ preserve the parameter λ, i. e., $\Phi \circ \pi = \Phi$.

When studying a bifurcation problem, we often consider systems $X = X^\lambda(x)$ locally defined near $(x, \lambda) = (0, 0)$, considering series expansions both in x and in λ. Then, in the Nth order normalization $\Phi_\star X$, the normalized part consists of a polynomial in y and λ, while the remainder term is of the form $O(|y|^{N+1} + |\lambda|^{N+1})$.

As in the previous case, we shall not formulate the present analogue of Theorem 6 for this case, but illustrate its meaning in examples.

The Reversible Case In the reversible case a linear involution R is given, while for the vector fields we require $R_\star X = -X$. Let $\mathcal{X}_R$ denote the class of all such reversible vector fields. Also, let C denote the class of all X such that $R_\star X = X$. Then, both $\mathcal{X}_R$ and C are linear spaces of vector fields. Moreover, C is a Lie-subalgebra. Associated to C is the group of diffeomorphisms that commute with R, i. e., the *R-equivariant* transformations. Also it is easy to see that for each of these diffeomorphisms Φ one has $\Phi_\star(\mathcal{X}_R) \subset \mathcal{X}_R$. The above approach applies to this situation in a straightforward manner. The gradings and splittings fit, while we have to choose the infinitesimal generator P from the set C. For details compare with [29].

Remark In the case with parameters, it sometimes is possible to obtain an alternative normal form where the normalized part is polynomial in y alone, with coefficients that depend smoothly on λ. A necessary condition for this is that the origin $y = 0$ is an equilibrium for all values of λ in some neighborhood of 0. To be precise, at the Nth order we van achieve smooth dependence of the coefficients on λ for $\lambda \in \Lambda_N$, where Λ_N is a neighborhood of $\lambda = 0$, that may shrink to $\{0\}$ as $N \to \infty$. So, for $N \to \infty$ only the formal aspect remains, as is the case in the above approach. This alternative normal form can be obtained by a proper use of the Implicit Function Theorem in the spaces $H^m(\mathbb{R}^n)$; for details e. g., see Section 2.2 in Vanderbauwhede [87]. For another discussion on this topic, cf. Section 7.6.2 in Dumortier et al. [16].

In Sect. "The Normal Form Procedure" the role of symmetry was considered regarding the semisimple part of the matrix A. A question is how this discussion generalizes to Lie-subalgebra's of vector fields.

Example (Volume preserving, parameter dependent axial symmetry [7,8]) On $\mathbb{R}^3$ consider a 1-parameter family X^λ of vector fields, preserving the standard volume $\sigma = \mathrm{d}x_1 \wedge \mathrm{d}x_2 \wedge \mathrm{d}x_3$. Assume that $X^0(0) = 0$ while the spectrum of D_0X^0 consists of the eigenvalues $\pm i$ and 0. For the moment regarding λ as an extra state space coordinate, we obtain a vertical vector field on $\mathbb{R}^4$ and we apply a combination of the above considerations. The 'generic' Jordan normal form then is

$$A = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} , \tag{2}$$

with an obvious splitting in semisimple and nilpotent part. The considerations of Sect. "The Normal Form Procedure" then directly apply to this situation. For any N this yields a transformation $\Phi : \mathbb{R}^4 \to \mathbb{R}^4$, with $\Phi(0) = 0$, preserving both the projection to the 1-dimensional parameter space and the volume of the 3-dimensional phase space, such that the normalized, Nth order part of $\Phi_\star X(y, \lambda)$, in cylindrical coordinates $y_1 = r\cos\varphi$, $y_2 = r\sin\varphi$, $y_3 = z$, has the rotationally symmetric form

$$\begin{aligned} \dot\varphi &= f(r^2, z, \lambda) \\ \dot r &= rg(r^2, z, \lambda) \\ \dot z &= h(r^2, z, \lambda) , \end{aligned}$$

again, for suitable polynomials f, g and h. Note, that in cylindrical coordinates the volume has the form $\sigma = r\mathrm{d}r \wedge \mathrm{d}\varphi \wedge \mathrm{d}z$. Again the functions f, g and h have to fit with the linear part. In particular we find that $h(r^2, z, \lambda) = \lambda + az + \cdots$, observing that for $\lambda \neq 0$ the origin is no equilibrium point.

Remark If $A = A_s + A_n$ is the canonical splitting of A in $H^1(\mathbb{R}^n) = \mathrm{gl}(n, \mathbb{R})$, then automatically both A_s and A_n are in the subalgebra under consideration. In the volume preserving setting this can be seen directly. In general the same holds true as soon as the corresponding linear Lie-group is algebraic, see [7,8] and the references given there.

The Hamiltonian Case

The Normal Form Theory in the Hamiltonian case goes back at least to Poincaré [69] and Birkhoff [5]. Other references are, for instance, Gustavson [47], Arnold [1,2,3], Sanders, Verhulst and Murdock [75], Van der Meer [88], Broer, Chow and Kim [17]. In Sect. "Preservation of Structure" we already saw that the axiomatic Lie algebra approach of [7,8] applies here, especially since the symplectic group $\mathrm{SP}(2n, \mathbb{R})$ is algebraic. The canonical form here usually goes with the name Williamson, compare Galin [44], Koçak [54] and Hoveijn [49]. We discuss how the Lie algebra approach compares to the literature.

The Lie algebra of Hamiltonian vector fields can be associated to the Poisson-algebra of Hamilton functions as follows, even in an arbitrary symplectic setting. As before, the symplectic form is denoted by σ. We recall, that for any Hamiltonian H the corresponding Hamiltonian vector field X_H is given by $\mathrm{d}H = \sigma(X_H, .)$, Now, let H and K be Hamilton functions with corresponding vector fields X_H resp. X_K. Then

$$X_{\{H,K\}} = [X_H, X_K] ,$$

implying that the map $H \mapsto X_H$ is a morphism of Lie algebra's. By definition this map is surjective, while its kernel consists of the (locally) constant functions.

This implies, that the normal form procedure can be completely rephrased in terms of the Poisson-bracket. We shall now demonstrate this by an example, similar to the previous one.

Example (Symplectic parameter dependent rotational symmetry [17]) Consider $\mathbb{R}^4$ with coordinates (x_1, y_1, x_2, y_2) and the standard symplectic form $\sigma = \mathrm{d}x_1 \wedge \mathrm{d}y_1 + \mathrm{d}x_2 \wedge \mathrm{d}y_2$, considering a C^∞ family of Hamiltonian functions H^λ, where $\lambda \in \mathbb{R}$ is a parameter. The fact that $\mathrm{d}H = \sigma(X_H, .)$ in coordinates means

$$\dot x_j = \frac{\partial}{H}\partial y_j , \quad \dot y_j = -\frac{\partial}{H}\partial x_j ,$$

for $j = 1, 2$. We assume that for $\lambda = 0$ the origin of $\mathbb{R}^4$ is a singularity. Then we expand as a Taylor series in (x, y, λ)

$$H^\lambda(x, y) = H_2(x, y, \lambda) + H_3(x, y, \lambda) + \dots ,$$

where the H_m is homogeneous of degree m in (x, y, λ). It follows for the corresponding Hamiltonian vector fields that

$$X_{H_m} \in H^{m-1}(\mathbb{R}^4) ,$$

in particular $X_{H_2} \in \mathrm{sp}(4, \mathbb{R}) \subset H^1(\mathbb{R}^4)$. Let us assume that this linear part X_{H_2} at $(x, y, \lambda) = (0, 0, 0)$ has eigenvalues $\pm i$ and a double eigenvalue 0. One 'generic' Willamson's normal form then is

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} , \tag{3}$$

compare with (2), corresponding to the quadratic Hamilton function

$$H_2(x, y, \lambda) = I + \frac{1}{2} {y_2}^2 ,$$

where $I := \frac{1}{2}({x_1}^2 + {y_1}^2)$. It is straightforward to give the generic matrix for the linear part in the extended state space $\mathbb{R}^5$, see above, so we will leave this to the reader.

The semisimple part $A_s = X_I$ now can be used to obtain a rotationally symmetric normal form, as before. In fact, for any $N \in \mathbb{N}$ there exists a canonical transformation $\Phi(x, y, \lambda)$, which keeps the parameter fixed, and a polynomial $F(I, x_2, y_2, \lambda)$, such that

$$(H \circ \Phi^{-1})(x, y, \lambda) = F(I, x_2, y_2, \lambda) + O(I + x_2^2 + y_2^2 + \lambda^2)^{(N+1)/2}.$$

Instead of using the adjoint action ad_A on the spaces $H^{m-1}(\mathbb{R}^5)$, we also may use the adjoint action

$$\mathrm{ad}H_2 : f \mapsto \{H_2, f\} ,$$

where f is a polynomial function of degree m in $(x_1, y_1, x_2, y_2, \lambda)$. In the vector field language, we choose the 'good' space $G^{m-1} \subset \ker \mathrm{ad}_{m-1} A_s$, which, in the function language, translates to a 'good' subset of $\ker \mathrm{ad} I$.

Whatsoever, the normalized part of the Hamilton function $H \circ \Phi^{-1}$, viz. the vector field $\Phi_\star X_H = X_{H \circ \Phi^{-1}}$, is rotationally symmetric. The fact that the Hamilton function F Poisson-commutes with I exactly amounts to invariance under the action generated by the vector field X_I, in turn implying that I is an integral of X_F. Indeed, if we define a 2π-periodic variable φ as follows:

$$x_1 = \sqrt{2I} \sin \varphi, \quad y_1 = \sqrt{2I} \cos \varphi ,$$

then $\sigma = \mathrm{d}I \wedge \mathrm{d}\varphi + \mathrm{d}x_2 \wedge \mathrm{d}y_2$, implying that the normalized vector field X_F has the canonical form

$$\dot{I} = 0, \quad \dot{\varphi} = -\frac{\partial F}{\partial I} ,$$
$$\dot{x}_2 = \frac{\partial F}{\partial y_2}, \quad \dot{y}_2 = -\frac{\partial F}{\partial x_2} .$$

Notice that φ is a cyclic variable, making the fact that I is an integral clearly visible. Also observe that $\dot{\varphi} = -1 + \cdots$. As before, and as in, e. g., the central force problem, this enables a reduction to lower dimension. Here, the latter two equations constitute the *reduction to 1 degree of freedom*: it is a family of planar Hamiltonian vector fields, parametrized by I and λ.

Remark

- This example has many variations. First of all it can be simplified by omitting parameters and even the zero eigenvalues. The conclusion then is that a planar Hamilton function with a nondegenerate minimum or maximum has a formal rotational symmetry, up to canonical coordinate changes.
- It also can be easily made more complicated, compare with Proposition 7 in Sect. "The Normal Form Procedure". Given an equilibrium with purely imaginary eigenvalues $\pm i\omega_1, \pm i\omega_2, \ldots, \pm i\omega_m$ and with $2(n-m)$ zero eigenvalues. Provided that there are no resonances up to order $N+1$, see (1), also here we conclude that Hamiltonian truncation at the order N, by canonical transformations can be given the form

$$F(I_1, \ldots, I_m, x_{m+1}, y_{m+1} \ldots, x_n, y_n) ,$$

where $I_j := \frac{1}{2}({x_j}^2 + {y_j}^2)$. As in Proposition 7 this normal form has a toroidal symmetry. Writing $x_j = \sqrt{2I_j} \sin \varphi_j$, $y_j = \sqrt{2I_j} \cos \varphi_j$, we obtain the canonical system of equations

$$\dot{I}_j = 0, \quad \dot{\varphi}_j = -\frac{\partial F}{\partial I_j} ,$$
$$\dot{x}_\ell = \frac{\partial F}{\partial y_l}, \quad \dot{y}_\ell = -\frac{\partial F}{\partial x_l} ,$$

$1 \le j \le m, \quad m+1 \le l \le n$. Note that $\dot{\varphi}_j = -\omega_j + \cdots$. In the case $m = n$ we deal with an elliptic equilibrium. The corresponding result usually is named the Birkhoff normal form [1,5,47], ► Dynamics of Hamiltonian Systems. Then, the variables (I_j, φ_j), $1 \le j \le m$, are a set of *action-angle variables* for the truncated part of order N, [1].
- Returning to algorithmic issues, in addition to Subsect. "On the Choices of the Complementary Space and of the Normalizing Transformation", we give a few further references in cases where a structure is being preserved: Broer [7], Deprit [40,41], Meyer [57] and Ferrer, Hanßmann, Palacián and Yanguas [43]. It is to be noted that this kind of Hamiltonian result also can be obtained, where the coordinate changes are constructed using *generating functions*, compare with, e. g., [1,3,77], see also the discussion in Broer, Hoveijn, Lunter and Vegter [20] and in particular Section 4.4.2 in [23]. For another, computationally very effective normal form algorithm, see Giorgilli and Galgani [45]. Also compare with references therein.

Semi-local Normalization

This section roughly consists of two parts. To begin with, a number of subsections are devoted to related formal normal form results, near fixed points of diffeomorphisms and near periodic solutions and invariant tori of vector fields.

A Diffeomorphism Near a Fixed Point

We start by formulating a result by Takens:

Theorem 8 (Takens normal form [83]) *Let $T\colon \mathbb{R}^n \to \mathbb{R}^n$ be a C^∞ diffeomorphism with $T(0) = 0$ and with a canonical decomposition of the derivative $D_0 T = S + N$ in semisimple resp. nilpotent part. Also, let $N \in \mathbb{N}$ be given, then there exists diffeomorphism Φ and a vector field X, both of class C^∞, such that $S_\star X = X$ and*

$$\Phi^{-1} \circ T \circ \Phi = S \circ X_1 + O(|y|^{N+1}) \,.$$

Here, as before, X_1 denotes the flow over time 1 of the vector field X. Observe that the vector field X necessarily has the origin as an equilibrium point. Moreover, since $S_\star X = X$, the vector field X is invariant with respect to the group generated by S.

The proof is a bit more involved than Theorem 6 in Sect. "The Normal Form Procedure", but it has the same spirit, also compare with [37]. In fact, the Taylor series of T is modified step-by-step, using coordinate changes generated by homogeneous vector fields of the proper degree.

After a reduction to center manifolds the spectrum of S is on the complex unit circle and Theorem 8 especially is of interest in cases where this spectrum consists of roots of unity, i. e., in the case of resonance. Compare with [32,37,83].

Again, the result is completely phrased in terms of Lie algebra's and groups and therefore bears generalization to many contexts with a preserved structure, compare Sect. "Preservation of Structure". The normalizing transformations of the induction process then are generated from the corresponding Lie algebra. For a symplectic analogue see Moser [59]. Also, both in Broer, Chow and Kim [17] and in Broer and Vegter [14], symplectic cases with parameters are discussed, where $S = \mathrm{Id}$, the Identity Map, resp. $S = -\mathrm{Id}$, the latter involving a period-doubling bifurcation.

Remark Let us consider a symplectic map T of the plane, which means that T preserves both area and orientation. Assume that T is fixing the origin, while the eigenvalues of $S = D_0 T$ are on the unit circle, without being roots of unity. Then S generates the rotation group $\mathrm{SO}(2, \mathbb{R})$, so the vector field X, which has divergence zero, in this case is rotationally symmetric. Again this result often goes with the name of Birkhoff.

Near a Periodic Solution

The Normal Form Theory at a periodic solution or closed orbit has a lot of resemblance to the local theory we met before.

To fix thoughts, let us consider a C^∞ vector field of the form

$$\begin{aligned} \dot{x} &= f(x, y) \\ \dot{y} &= g(x, y) \,, \end{aligned} \tag{4}$$

with $(x, y) \in \mathbb{T}^1 \times \mathbb{R}^n$. Here $\mathbb{T}^1 = \mathbb{R}/(2\pi\mathbb{Z})$. Assuming $y = 0$ to be a closed orbit, we consider the formal Taylor series with respect to y, with x-periodic coefficients. By Floquet Theory, we can assume that the coordinates (x, y) are such that

$$f(x, y) = \omega + O(|y|), \quad g(x, y) = \Omega y + O(|y|^2) \,,$$

where $\omega \in \mathbb{R}$ is the frequency of the closed orbit and $\Omega \in \mathrm{gl}(n, \mathbb{R})$ its Floquet matrix. Again, the idea is to 'simplify' this series further. To this purpose we introduce a grading as before, letting $H^m = H^m(\mathbb{T}^1 \times \mathbb{R}^n)$ be the space of vector fields

$$Y(x, y) = L(x, y)\frac{\partial}{\partial x} + \sum_{j=1}^{n} M_j(x, y)\frac{\partial}{\partial y_j} \,,$$

with $L(x, y)$ and $M(x, y)$ homogeneous in y of degree $m - 1$ resp. m. Notice, that this space H^m is infinite-dimensional. However, this is not at all problematic for the things we are doing here. By this definition, we have that

$$A := \omega\frac{\partial}{\partial x} + \Omega y\frac{\partial}{\partial y}$$

is a member of H^1 and with this normally linear part we can define an adjoint representation $\mathrm{ad}A$ as before, together with linear maps

$$\mathrm{ad}_m A : H^m \to H^m \,.$$

Again we assume to have a decomposition

$$G^m \oplus \mathrm{Im}(\mathrm{ad}_m A) = H^m \,,$$

where the aim is to transform the terms of the series successively into the G^m, for $m = 2, 3, 4, \ldots$.

The story now runs as before. In fact, the proof of Theorem 6 in Sect. "The Normal Form Procedure", as well as

its Lie algebra versions indicated in Sect. "Preservation of Structure", can be repeated almost verbatim for this case. Moreover, if $\Omega = \Omega_s + \Omega_n$ is the canonical splitting in semisimple and nilpotent part, then

$$\omega \frac{\partial}{\partial x} + \Omega_s \, \mathrm{y} \frac{\partial}{\partial} \mathrm{y}$$

gives the semisimple part of $\mathrm{ad}_m A$, as can be checked by a direct computation. From this computation one also deduces the non-resonance conditions needed for the present torus-symmetric analogue of Proposition 7 in Sect. "The Normal Form Procedure".

There are different cases *with* resonance either between the imaginary parts of the eigenvalues of Ω (normal resonance) or between the latter and the frequency ω (normal-internal resonance). All of this extends to the various settings with preservation of structure as discussed before. In all cases direct analogues of the Theorems 6 and 8 are valid. General references in this direction are Arnold [2,3], Bruno [35,36], Chow, Li and Wang [37], Iooss [50], Murdock [60] or Sanders, Verhulst and Murdock [75].

Remark

- This approach also is important for non-autonomous systems with periodic time dependence. Here the normalization procedure includes averaging. As a special case of the above form, we obtain a system

  $$\dot{x} = \omega$$
  $$\dot{y} = g(x, y) \,,$$

  so where $x \in \mathbb{T}^1$ is proportional to the time. Apart from the general references given above, we also refer to, e. g., Broer and Vegter [14] and to Broer, Roussarie and Simó [9,19]. The latter two applications also contain parameters and deal with bifurcations. Also compare with Verhulst ► Perturbation Analysis of Parametric Resonance.
- A geometric tool that can be successfully applied in various resonant cases is a *covering space*, obtained by a Van der Pol transformation (or by passing to co-rotating coordinates). This involves equivariance with respect to the corresponding deck group. This setting (with or without preservation of structure) is completely covered by the general Lie algebra approach as described above.
  For the Poincaré map this deck group symmetry directly yields the normal form symmetry of Theorem 8. For applications in various settings, with or without preservation of structure, see [14,24,32]. This normalization technique is effective for studying bifurcation of subharmonic solutions. In the case of period doubling the covering space is just a double cover. In many cases Singularity Theory turns out to be useful.

Near a Quasi-periodic Torus

The approach of the previous subsection also applies at an invariant torus, provided that certain requirements are met. Here we refer to Braaksma, Broer and Huitema [15], Broer and Takens [12] and Bruno [35,36].

Let us consider a C^∞-system

$$\begin{aligned} \dot{x} &= f(x, y) \\ \dot{y} &= g(x, y) \end{aligned} \tag{5}$$

as before, with $(x, y) \in \mathbb{T}^m \times \mathbb{R}^n$. Here $\mathbb{T}^m = \mathbb{R}^m/(2\pi\mathbb{Z})^m$. We assume that $f(x, y) = \omega + O(|y|)$, which implies that $y = 0$ is an invariant m-torus, with on it a constant vector field with frequency-vector ω. We also assume that $g(x, y) = \Omega y + O(|y|^2)$, which is the present analogue of the Floquet form as known in the periodic case, with $\Omega \in \mathrm{gl}(n, \mathbb{R})$, independent of x. Contrarious to the situations for $m = 1$ and $n = 1$, in general reducibility to Floquet form is not possible. Compare with [10,15,18,38] and references therein. For a similar approach of a system that is not reducible to Floquet form, compare with [22].

Presently we assume this reducibility, expanding in formal series with respect to the variables y, where the coefficients are functions on $\mathbb{T}^m$. These coefficients, in turn, can be expanded in Fourier series. The aim then is, to 'simplify' this combined series by successive coordinate changes, following the above procedure. As a second requirement it then is needed that certain *Diophantine* conditions are satisfied on the pair (ω, Ω). Below we give more details on this. Instead of giving general results we again refer to [12,15,26,29,35,36]. Moreover, to fix thoughts, we present a simple example with a parameter. Here again a direct link with averaging holds.

Example (Toroidal symmetry with small divisors [38]*)* Given is a family of vector fields $\dot{x} = X(x, \lambda)$ with $(x, \lambda) \in \mathbb{T}^m \times \mathbb{R}$. We assume that $X = X(x, \lambda)$ has the form

$$X(x, \lambda) = \omega + f(x, \lambda) \,,$$

with $f(x, 0) \equiv 0$. It is assumed that the frequency vector $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ has components that satisfy the Diophantine non-resonance condition

$$|\langle \omega, k \rangle| \geq \gamma |k|^{-\tau} \,, \tag{6}$$

for all $k \in \mathbb{Z} \setminus \{0\}$. Here $\gamma > 0$ and $\tau > n-1$ are prescribed constants. We note that $\tau > n-1$ implies that this condition in $\mathbb{R}^n = \{\omega\}$ excludes a set of Lebesgue measure $O(\gamma)$ as $\gamma \downarrow 0$, compare with [38]. It follows, that by successive transformations of the form

$$h\colon (x,\lambda) \mapsto (x + P(x,\lambda), \lambda)$$

the x-dependence of X can be pushed away to ever higher order in λ, leading to a formal normal form

$$\dot{\xi} = \omega + g(\lambda)\,,$$

with $g(0) = 0$. Observe that in this case 'simple' means x- (or ξ-) independent. Therefore, in a proper formalism, x-independent systems constitute the spaces G^m. Indeed, in the induction process we get

$$X(x,\lambda) = \omega + g_2(\lambda) + \cdots + g_{N-1}(\lambda) + f_N(x,\lambda) + O(|\lambda|^{N+1})\,,$$

compare the proof of Theorem 6 in Sect. "The Normal Form Procedure". Writing $\xi = x + P(x,\lambda)$, with $P(x,\lambda) = \tilde{P}(x,\lambda)\lambda^N$, we substitute

$$\begin{aligned}\dot{\xi} &= (\mathrm{Id} + D_\xi P)\dot{x}\\ &= \omega + g_2(\lambda) + \cdots + g_{N-1}(\lambda)\\ &\quad + f_N(x,\lambda) + D_x P(x,\lambda)\,\omega + O(|\lambda|^{N+1})\,,\end{aligned}$$

Where we express the right-hand side in x. So we have to satisfy an equation

$$D_x P(x,\lambda)\omega + f_N(x,\lambda) \equiv c\lambda^N \mathrm{mod} O(|\lambda|^{N+1})\,,$$

for a suitable constant c. Writing $f_N(x,\lambda) = \tilde{f}_N(x,\lambda)\lambda^N$, this amounts to

$$D_x \tilde{P}(x,\lambda)\omega = -\tilde{f}_N(x,\lambda) + c\,,$$

which is the present form of the homological equation. If

$$\tilde{f}_N(x,\lambda) = \sum_{k\in\mathbb{Z}^n} a_k(\lambda) \mathrm{e}^{i\langle x,k\rangle}\,,$$

then $c = a_0$, i. e., the m-torus average

$$a_0(\lambda) = \frac{1}{(2\pi)^m}\int_{\mathbb{T}^m} \tilde{f}_N(x.\lambda)\mathrm{d}x\,.$$

Moreover,

$$\tilde{P}(x,\lambda) = \sum_{k\neq 0} \frac{a_k(\lambda)}{i\langle\omega,k\rangle}\mathrm{e}^{i\langle x,k\rangle}\,.$$

This procedure formally only makes sense if the frequencies $(\omega_1, \omega_2, \ldots, \omega_m)$ have no resonances, which means that for $k \neq 0$ also $\langle\omega, k\rangle \neq 0$. In other words this means that the components of the frequency vector ω are rationally independent. Even then, the denominator $i\langle\omega,k\rangle$ can become arbitrarily small, so casting doubt on the convergence. This problem of *small divisors* is resolved by the Diophantine conditions (6). For further reference, e. g., see [2,10,15,18,38]: for real analytic X, by the Paley–Wiener estimate on the exponential decay of the Fourier coefficients, the solution P again is real analytic. Also in the C^∞-case the situation is rather simple, since then the coefficients in both cases decay faster than any polynomial.

Remark

- The discussion at the end of Subsect. "Near a Periodic Solution" concerning normalization at a periodic solution, largely extends to the present case of a quasi-periodic torus. Assuming reducibility to Floquet form, we now have a normally linear part

  $$\omega\frac{\partial}{\partial x} + \Omega y\frac{\partial}{\partial y}\,,$$

  with a frequency vector $\omega \in \mathbb{R}^n$. As before $\Omega \in \mathrm{gl}(m,\mathbb{R})$. For the corresponding KAM Perturbation Theory the Diophantine conditions (6) on the frequencies are extended by including the normal frequencies, i. e., the imaginary parts $\omega_1^N, \ldots, \omega_s^N$ of the eigenvalues of Ω. To be precise, for $\gamma > 0$ and $\tau > n-1$, above the conditions (6), these extra Melnikov conditions are given by

  $$\begin{aligned}|\langle\omega,k\rangle + \omega_j^N| &\geq \gamma|k|^{-\tau}\\ |\langle\omega,k\rangle + 2\omega_j^N| &\geq \gamma|k|^{-\tau}\\ |\langle\omega,k\rangle + \omega_j^N + \omega_\ell^N| &\geq \gamma|k|^{-\tau}\,,\end{aligned} \tag{7}$$

  for all $k \in \mathbb{Z}^n \setminus \{0\}$ and for $j, \ell = 1, 2, \ldots, s$ with $\ell \neq j$. See below for the description of an application to KAM Theory (and more references) in the Hamiltonian case.
- In this setting normal resonances can occur between the ω_j^N and normal-internal resonances between the ω_j^N and ω. Certain strong normal-internal resonances occur when one of the left-hand sides of (7) vanishes. For an example see below. Apart from these now also internal resonances between the components of ω come into play. The latter generally lead to destruction of the invariant torus.
- As in the periodic case also here covering spaces turn out to be useful for studying various resonant

bifurcation scenarios often involving applications of both Singularity Theory and KAM Theory. Compare with [15,25,31,33].

Non-formal Aspects

Up to this moment (almost) all considerations have been formal, i. e., in terms of formal power series. In general, the Taylor series of a C^∞-function, say, $\mathbb{R}^n \to \mathbb{R}$ will be divergent. On the other hand, any formal power series in n variables occurs as the Taylor series of some C^∞-function. This is the content of a theorem by É. Borel, cf. Narasimhan [61].

We briefly discuss a few aspects regarding convergence or divergence of the normalizing transformation or of the normalized series. We recall that the growth rate of the formal series, including the convergent case, is described by the Gevrey index, compare with, e. g., [4,55,72,73].

Normal Form Symmetry and Genericity

For the moment we assume that all systems are of class C^∞. As we have seen, if the normalization procedure is carried out to some finite order N, the transformation Φ is a real analytic map. If we take the limit for $N \to \infty$, we only get formal power series $\hat{\Phi}$, but, by the Borel Theorem, a 'real' C^∞-map Φ exists with $\hat{\Phi}$ as its Taylor series.

Let us discuss the consequences of this, say, in the case of Proposition 7 in Sect. "The Normal Form Procedure". Assuming that there are no resonances at all between the ω_j, as a corollary, we find a C^∞-map $y = \Phi(x)$ and a C^∞ vector field $p = p(y)$, such that:

- The vector field $\Phi_\star X - p$, in corresponding generalized cylindrical coordinates has the symmetric C^∞-form

$$\begin{aligned}\dot{\varphi}_j &= f_j(r_1^2, \ldots, r_m^2, z_{n-2m+1}, \ldots, z_n)\\ \dot{r}_j &= r_j g_j(r_1^2, \ldots, r_m^2, z_{n-2m+1}, \ldots, z_n)\\ \dot{z}_\ell &= h_\ell(r_1^2, \ldots, r_m^2, z_{n-2m+1}, \ldots, z_n)\,,\end{aligned}$$

  where $f_j(0) = \omega_j$ and $h_\ell(0) = 0$ for $1 \le j \le m$, $n - 2m + 1 \le \ell \le n$.
- The Taylor series of p identically vanishes at $y = 0$.

Note, that an ∞-ly flat term p can have component functions like e^{-1/y_1^2}. We see, that the m-torus symmetry only holds up to such flat terms. Therefore, this symmetry, if present at all, also can be destroyed again by a *generic* flat 'perturbation'. We refer to Broer and Takens [12], and references therein, for further consequences of this idea. The main point is, that by a Kupka–Smale argument, which generically forbids so much symmetry, compare with [67].

Remark The Borel Theorem also can be used in the reversible, the Hamiltonian and the volume preserving setting. In the latter two cases we exploit the fact that a structure preserving vector field is generated by a function, resp. an $(n-2)$-form. Similarly the structure preserving transformations have such a generator. On these generators we then apply the Borel Theorem. Many Lie algebra's of vector fields have this 'Borel Property', saying that a formal power series of a transformation can be represented by a C^∞ map in the same structure preserving setting.

On Convergence

The above topological ideas also can be pursued in many real analytic cases, where they imply a generic *divergence* of the normalizing transformation. For an example in the case of the Hamiltonian Birkhof normal form [5,47,77] compare with Broer and Tangerman [13] and its references. As an example we now deal with the linearization of a holomorphic germ in the spirit of Sect. "The Normal Form Procedure".

Example (Holomorphic linearization [34,58,92]*)* A holomorphic case concerns the linearization of a local holomorphic map $F\colon (\mathbb{C}, 0) \to (\mathbb{C}, 0)$ of the form $F(z) = \lambda z + f(z)$ with $f(0) = f'(0) = 0$. The question is whether there exists a local biholomorphic transformation $\Phi\colon (\mathbb{C}, 0) \to (\mathbb{C}, 0)$ such that

$$\Phi \circ F = \lambda \cdot \Phi\,.$$

We say that Φ linearizes F near its fixed point 0. A formal solution as in Sect. "The Normal Form Procedure" generally exists for all $\lambda \in \mathbb{C} \setminus \{0\}$, not equal to a root of unity. The elliptic case concerns λ on the complex unit circle, so of the form $\lambda = \mathrm{e}^{2\pi i\alpha}$, where $\alpha \notin \mathbb{Q}$, and where the approximability of α by rationals is of importance, cf. Siegel [76] and Bruno [34]. Siegel introduced a sufficient Diophantine condition related to (6), which by Bruno was replaced by a sufficient condition on the continued fraction approximation of α. Later Yoccoz [92] proved that the latter condition is both necessary and sufficient. For a description and further comments also compare with [2,10,42,58].

Remark

- In certain real analytic cases Neĭshtadt [63], by truncating the (divergent) normal form series appropriately, obtains a remainder that is exponentially small in the perturbation parameter. Also compare with,

e. g., [2,3,4,78]. For an application in the context of the Bogdanov–Takens bifurcations for diffeomorphisms, see [9,19]. It follows that chaotic dynamics is confined to exponentially narrow horns in the parameter space.

- The growth of the Taylor coefficients of the usually divergent series of the normal form and of the normalizing transformation can be described using Gevrey symptotics [4,55,70,71,72,73]. Apart from its theoretical interest, this kind of approach is extremely useful for computational issues also compare with [4,46,78,79,80].

Applications

We present two main areas of application of the Normal Form Theory in Perturbation Theory. The former of these deals with more globally qualitative aspects of the dynamics given by normal form approximations. The latter class of applications concerns the Averaging Theorem, where the issue is that solutions remain close to approximating solutions given by a normal form truncation.

'Cantorized' Singularity Theory

We return to the discussion of the motivation in Sect. "Motivation", where there is a toroidal normal form symmetry up to a finite or infinite order. To begin with let us consider the quasi-periodic Hopf bifurcation [6,15], which is the analogue of the Hopf bifurcation for equilibria and the Hopf–Neĭmark–Sacker bifurcation for periodic solutions, in the case of quasi-periodic tori. For a description comparing the differences between all these cases we refer to [38]. For a description of the resonant dynamics in the resonant gaps see [30] and references therein. Apart from this, a lot of related work has been done in the Hamiltonian and reversible context as well, compare with [25,26,29,48].

To be more definite, we consider families of systems defined on $\mathbb{T}^m \times \mathbb{R}^m \times \mathbb{R}^{2p} = \{x, y, z\}$, endowed with the symplectic form $\sigma = \sum_{j=1}^m \mathrm{d}x_j \wedge \mathrm{d}y_j + \mathrm{d}z^2$. We start with 'integrable', i. e., x-independent, families of systems of the form

$$\begin{aligned} \dot{x} &= f(y, z; \lambda) \\ \dot{y} &= g(y, z; \lambda) \\ \dot{z} &= h(y, z; \lambda)\,, \end{aligned} \tag{8}$$

compare with (5), to be considered near an invariant m-torus $\mathbb{T}^m \times \{y_0\} \times \{z_0\}$, meaning that we assume $g(y_0, z_0) = 0 = h(y_0, z_0)$. The general interest is with the persistence of such a torus under nearly integrable perturbations of (8), where we include λ as a multiparameter. This problem belongs to the Parametrized KAM Theory [10,15,18] of which we sketch some background now. For y near y_0 and λ near 0 consider $\Omega(\lambda, y) = D_z h(y, z_0; \lambda)$, noting that $\Omega(\lambda, y) \in \mathrm{sp}(2p, \mathbb{R})$. Also consider the corresponding normal linear part

$$\omega(\lambda, y)\frac{\partial}{\partial x} + \Omega(\lambda, y) z \frac{\partial}{\partial z}\,.$$

As a first case assume that the matrix $\Omega(0, y_0)$ has only simple non-zero eigenvalues. Then a full neighborhood of $\Omega(0, y_0)$ in $\mathrm{sp}(2p, \mathbb{R})$ is completely parametrized by the eigenvalues of the matrices, where – in this symplectic case – we have to refrain from 'counting double'. We roughly quote a KAM Theorem as this is known in the present circumstances [15]. As a nondegeneracy condition assume that the product map

$$(\lambda, y) \mapsto (\omega(\lambda, y), \Omega(\lambda, y))$$

is a submersion near $(\lambda, y) = (0, y_0)$. Also assume all Diophantine conditions (6), (7) to hold. Then the parametrized system (8) is *quasi-periodically stable*, which implies persistence of the corresponding Diophantine tori near $\mathbb{T}^m \times \{y_0\} \times \{z_0\}$ under nearly integrable perturbations.

Remark

- A key concept in the KAM Theory is that of *Whitney smoothness* of foliations of tori over nowhere dense 'Cantor' sets. In real analytic cases even Gevrey regularity holds, and similarly when the original setting is Gevrey; compare with [71,90]. For general reference see [10,15,18] and references therein.
- The gaps in the 'Cantor sets' are centered around the various resonances. Their union forms an open and dense set of small measure, where perturbation series diverge due to small divisors. In each gap the considerations mentioned at the end of Subsects. "Near a Periodic Solution" and "Near a Quasi-Periodic Torus" apply. In [25,31,33,38] the differences between period and the quasi-periodic cases are highlighted.

Example (Quasi-periodic Hamiltonian Hopf bifurcation [26]) As a second case we take $p = 2$, considering the case of normal $1 : -1$ resonance where the eigenvalues of $\Omega(0, y_0)$ are of the form $\pm i\mu_0$, for a positive μ_0. For simplicity we only consider the non-semisimple Williamson normal form

$$\Omega(0, y_0) \sim \begin{pmatrix} 0 & -\mu_0 & 1 & 0 \\ \mu_0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -\mu_0 \\ 0 & 0 & \mu_0 & 0 \end{pmatrix},$$

where $\sim$ denotes symplectic similarity. The present format of the nondegeneracy condition regarding the product map $(\lambda, y) \mapsto (\omega(\lambda, y), \Omega(\lambda, y)$ is as before, but now it is required that the second component $(\lambda, y) \mapsto \Omega(\lambda, y)$ is a *versal unfolding* of the matrix $\Omega(0, y_0)$ in $\mathrm{sp}(4, \mathbb{R})$ with respect to the adjoint $\mathrm{SP}(4, \mathbb{R})$-action, compare with [2,26,27,29,44,49,54]. It turns out that a standard normalization along the lines of Sect. "Preservation of Structure" can be carried out in the z-direction, generically leading to a Hamiltonian Hopf bifurcation [88], characterized by its swallowtail geometry [86], which 'governs' families of invariant m-, $(m+1)$- and $(m+2)$-tori (the latter being Lagrangean). Here for the complementary spaces, see Sect. "The Normal Form Procedure", the $\mathrm{sl}(2, \mathbb{R})$-Theory is being used [39,60].

As before the question is what happens to this scenario when perturbing the system in a non-integrable way? In that case we need quasi-periodic Normal Form Theory, in the spirit of Sect. "Semi-Local Normalization". Observe that by the $1:-1$ resonance, difficulties occur with the third of the three Melnikov conditions (7). In a good set of Floquet coordinates the resonance can be written in the form $\omega_j^N + \omega_\ell^N = 0$. Nevertheless, another application of Parametrized KAM Theory [10,15,18,27] yields that the swallowtail geometry is largely preserved, when leaving out a dense union of resonance gaps of small measure. Here perturbation series diverge due to small divisors. What remains is a 'Cantorized' version of the swallowtail [26,27]. For a reversible analogue see [29,31].

Remark

- The example concerns a strong resonance and it fits in some of the larger gaps of the 'Cantor' set described in the former application of Parametrized KAM Theory. Apart from this, the previous remarks largely apply again. It turns out that in these and many other cases there is an infinite regression regarding the resonant bifurcation diagram.
- The combination of KAM Theory with Normal Form Theory generally has been very fruitful. In the example of Sect. "Applications" it implies that each KAM torus corresponds to a parameter value λ_0 that is the Lebesgue density point of such quasi-periodic tori. In the real analytic case by application of [63], it can be shown that the relative measure with non-KAM tori is exponentially small in $\lambda - \lambda_0$.
  Similar results in the real analytic Hamiltonian KAM Theory (often going by the name of exponential condensation) have been obtained by [51,52,53], also compare with [35,36], ► Diagrammatic Methods in Classical Perturbation Theory and with [3,10] and references therein.

On the Averaging Theorem

Another class of applications of normalizing-averaging is in the direction of the Averaging Theorem. There is a wealth of literature in this direction, that is not in the scope of the present paper, for further reading compare with [1,2,3,75] and references therein.

Example (A simple averaging theorem [1]) Given is a vector field

$$\dot{x} = \omega(y) + \varepsilon f(x, y)$$
$$\dot{y} = \varepsilon g(x, y)$$

with $(x, y) \in \mathbb{T}^1 \times \mathbb{R}^n$, compare with (4). Roughly following the recipe of the normalization process, a suitable near-identity transformation

$$(x, y) \mapsto (x, \eta)$$

of $\mathbb{T}^1 \times \mathbb{R}^n$ yields the following reduction, after truncating at order $O(\varepsilon^2)$:

$$\dot{\eta} = \varepsilon \bar{g}(\eta), \quad \text{where} \quad \bar{g}(\eta) = \frac{1}{2\pi} \int_0^{2\pi} g(x, \eta) \mathrm{d}x \,.$$

We now compare $y = y(t)$ and $\eta = \eta(t)$ with coinciding initial values $y(0) = \eta(0)$ as t increases. The Averaging Theorem asserts that if $\omega(\eta) > 0$ is bounded away from 0, it follows that, for a constant $c > 0$, one has

$$|y(t) - \eta(t)| < c\varepsilon, \quad \text{for all } t \quad \text{with } 0 \le t \le \frac{1}{\varepsilon} \,.$$

This theory extends to many classes of systems, for instance to Hamiltonian systems or, in various types of systems, in the immediate vicinity of a quasi-periodic torus. Further normalizing can produce sharper estimates that are polynomial in ε, while in the analytic case this even extends over exponentially long time intervals, usually known under the name of Nekhorochev estimates [64,65,66], for a description and further references also see [10], ► Nekhoroshev Theory. Another direction of generalization concerns passages through resonance, which in the example implies that the condition on $\omega(\eta)$ is no longer valid. We here mention [62], for further references and descriptions referring to [2,3] and to [75].

Future Directions

The area of research in Normal Form Theory develops in several directions, some of which are concerned with computational aspects, including the nilpotent case. Although for smaller scale projects much can be done with computer packages, for the large scale computations, e. g., needed in Celestial Mechanics, single purpose formula manipulators have to be built. For an overview of such algorithms, compare with [60,78,79,80] and references therein. Here also Gevrey asymptotics is of importance.

Another direction of development is concerned with applications in Bifurcation Theory, often in particular Singularity Theory. In many of these applications, certain coefficients in the truncated normal form are of interest and their computation is of vital importance. For an example of this in the Hamiltonian setting see [11], where the Giorgili–Galgani [45] algorithm was used to obtain certain coefficients at all arbitrary order in an efficient way. For other examples in this direction see [28,32].

Related to this is the problem how to combine the normal form algorithms as related to the present paper, with the polynomial normal forms of Singularity Theory. The latter have a universal (i. e., context independent) geometry in the product of state space and parameter space. A problem of relevance for applications is to pull-back the Singularity Theory normal form back to the original system. For early attempts in this direction see [20,23], which, among other things, involve Gröbner basis techniques.

Bibliography

Primary Literature

1. Arnold VI (1980) Mathematical methods of classical mechanics. Springer, New York (English, Russian original)
2. Arnold VI (1983) Geometrical methods in the theory of ordinary differential equations. Springer, New York (English, Russian original)
3. Arnold VI (ed) (1988) Dynamical systems III, encyclopædia of mathematical sciences, vol 3. Springer, Berlin (English, Russian original)
4. Baldomá I, Haro A (2008) One dimensional invariant manifolds of Gevrey type in real-analytic maps. DCDS series B 10(2&3):295–322
5. Birkhoff GD (1927) Dynamical systems. AMS Publications
6. Braaksma BLJ, Broer HW (1987) On a quasi-periodic Hopf bifurcation. Ann Inst Henri Poincaré, Analyse Non Linéaire 4(2):115–168
7. Broer HW (1981) Formal normal form theorems for vector fields and some consequences for bifurcations in the volume preserving case. In: Rand DA, Young LS (eds) Dynamical systems and turbulence. Warwick, 1980 LNM 898. Springer, Berlin, pp 54–74
8. Broer HW (1993) Notes on perturbation theory 1991. In: Erasmus ICP (ed) Mathematics and Fundamental Applications. Aristotle University Thessaloniki, pp 44
9. Broer HW, Roussarie R (2001) Exponential confinement of chaos in the bifurcation set of real analytic diffeomorphisms. In: Broer HW, Krauskopf B, Vegter G (eds) Global analysis of dynamical systems. Festschrift dedicated to Floris Takens for his 60th birthday Bristol Phila IOP, pp 167–210
10. Broer HW, Sevryuk MB (2008) kam Theory: quasi-periodicity in dynamical systems. In: Broer HW Hasselblatt B, Takens F (eds) Handbook of Dynamical Systems, vol 3. North-Holland (to appear 2009)
11. Broer HW, Simó C (2000) Resonance tongues in Hill's equations: a geometric approach. J Differ Equ 166:290–327
12. Broer HW, Takens F (1989) Formally symmetric normal forms and genericity. Dyn Rep 2:36–60
13. Broer HW, Tangerman FM (1986) From a differentiable to a real analytic perturbation theory, applications to the Kupka Smale theorems. Ergod Theor Dyn Syst 6:345–362
14. Broer HW, Vegter G (1992) Bifurcational aspects of parametric resonance. Dyn Rep New Ser 1:1–51
15. Broer HW, Huitema GB, Takens F, Braaksma BLJ (1990) Unfoldings and bifurcations of quasi-periodic tori. Mem AMS 83(421):i–vii; pp 1–175
16. Broer HW, Dumortier F, van Strien SJ, Takens F (1991) Structures in dynamics: Finite dimensional deterministic studies. In: van Groesen EWC, de Jager EM (eds) Studies in mathematical physics, vol 2. North-Holland, Amsterdam (Russian translation 2003)
17. Broer HW, Chow SN, Kim Y, Vegter G (1993) A normally elliptic Hamiltonian bifurcation. ZAMP 44:389–432
18. Broer HW, Huitema GB, Sevryuk MB (1996) Quasi-periodic motions in families of dynamical systems: Order amidst chaos. LNM 1645 Springer, Berlin
19. Broer HW, Roussarie R, Simó C (1996) Invariant circles in the Bogdanov-Takens bifurcation for diffeomorphisms. Ergod Theor Dyn Syst 16:1147–1172
20. Broer HW, Hoveijn I, Lunter GA, Vegter G (1998) Resonances in a Spring–Pendulum: algorithms for equivariant singularity theory. Nonlinearity 11(5):1–37
21. Broer HW, Simó C, Tatjer JC (1998) Towards global models near homoclinic tangencies of dissipative diffeomorphisms. Nonlinearity 11:667–770
22. Broer HW, Takens F, Wagener FOO (1999) Integrable and non-integrable deformations of the skew Hopf bifurcation. Reg Chaotic Dyn 4(2):17–43
23. Broer HW, Hoveijn I, Lunter GA, Vegter G (2003) Bifurcations in Hamiltonian systems: Computing singularities by Gröbner bases. LNM, vol 1806. Springer, Berlin
24. Broer HW, Golubitsky M, Vegter G (2003) The geometry of resonance tongues: A singularity theory approach. Nonlinearity 16:1511–1538
25. Broer HW, Hanßmann H, Jorba A, Villanueva J, Wagener FOO (2003) Normal-internal resonances in quasi-periodically forces oscillators: a conservative approach. Nonlinearity 16:1751–1791
26. Broer HW, Hanßmann H, Hoo J (2007) The quasi-periodic Hamiltonian Hopf bifurcations. Nonlinearity 20:417–460
27. Broer HW, Hoo J, Naudot V (2007) Normal linear stability of quasi-periodic tori. J Differ Equ 232(2):355–418

28. Broer HW, Golubitsky M, Vegter G (2007) Geometry of resonance tongues. In: Chéniot D, Dutertre N, Murolo C, Trotman D, Pichon A (eds) Proceedings of the 2005 marseille singularity school and conference, dedicated to Jean-Paul Brasselet on his 60th birthday. World Scient, pp 327–356
29. Broer HW, Ciocci MC, Hanßmann H (2007) The quasi-periodic reversible Hopf bifurcation. IJBC 17(8):2605–2623
30. Broer HW, Naudot V, Roussarie R, Saleh K, Wagener FOO (2007) Organising centres in the semi-global analysis of dynamical systems. IJAMAS 12(D07):7–36
31. Broer HW, Ciocci MC, Hanßmann H, Vanderbauwhede A (2008) Quasi-periodically stable unfoldings of normally resonant tori. Physica D (to appear)
32. Broer HW, Vegter G (2008) Generic Hopf-Neĭmark-Sacker bifurcations in feed forward systems. Nonlinearity 21:1547–1578
33. Broer HW, Hanßmann H, You J () On the destruction of resonant Lagrangean tori in Hamiltonian sytems. (in preparation)
34. Bruno AD (1965) On the convergence of transformations of differential equations to the normal form. Soviet Math Dokl 6:1536–1538 (English, Russian original)
35. Bruno AD (1989) Local methods in nonlinear differential equations. Springer, Berlin (English, Russian original)
36. Bruno AD (2000) Power geometry in algebraic and differential equations. North-Holland Mathematical Library, vol 57. North-Holland, Amsterdam (English, Russian original)
37. Chow SN, Li C, Wang D (1994) Normal forms and bifurcations of planar vector fields. Cambrigde University Press, Cambridge
38. Ciocci MC, Litvak-Hinenzon A, Broer HW (2005) Survey on dissipative KAM theory including quasi-periodic bifurcation theory based on lectures by Henk Broer. In: Montaldi J, Ratiu T (eds) Geometric mechanics and symmetry: The Peyresq lectures. LMS Lecture Notes Series, vol 306. Cambridge University Press, Cambridge, pp 303–355
39. Cushman RH, Sanders JA (1986) Nilpotent normal forms and representation theory of sl(2, $\mathbb{R}$). In: Golubitsky M, Guckenheimer J (eds) Multi-parameter bifurcation theory AMS. Providence, pp 31–51
40. Deprit A (1969) Canonical transformations depending on a small parameter. Celest Mech 1:12–30
41. Deprit A (1981) The elimination of the parallax in satellite theory. Celest Mech 24:111–153
42. Eliasson LH, Kuksin SB, Marmi S, Yoccoz JC (2002) Dynamical systems and small divisors. LNM, vol 1784. Springer, Berlin
43. Ferrer S, Hanßmann H, Palacián J, Yanguas P (2002) On perturbed oscillators in 1-1-1 resonance: the case of axially symmetric cubic potentials. J Geom Phys 40:320–369
44. Galin DM (1982) Versal deformations of linear Hamiltonian systems. Am Soc Trans Ser 2(118):1–12
45. Giorgilli A, Galgani L (1978) Formal integrals for an autonomous Hamiltonian system near an equilibrium point. Celest Mech 17:267–280
46. Giorgilli A, Delshams A, Fontich E, Galgani L, Simó C (1989) Effective stabibity for a Hamiltonian system near an elliptic equilibrium point with an application to the restricted three body problem. J Differ Equ 77:167–198
47. Gustavson FG (1966) On constructing formal integrals of a Hamiltonian system near an equilibrium point. Astron J 71:670–686
48. Hanßmann H (2007) Local and semi-local bifurcations in hamiltonian dynamical systems. LNM, vol 1893. Springer, Berlin
49. Hoveijn I (1996) Versal deformations and normal forms for reversible and Hamiltonian linear systems. J Diff Eqns 126(2)408–442
50. Iooss G (1988) Global characterization of the normal form of a vector field near a closed orbit. J Diff Eqns 76:47–76
51. Jorba À, Villanueva J (1997) On the persistence of lower dimensional invariant tori under quasi-periodic perturbations. J Nonlinear Sci 7:427–473
52. Jorba À, Villanueva J (1997) On the normal behaviour of partially elliptic lower-dimensional tori of Hamiltonian systems. Nonlinearity 10:783–822
53. Jorba À, Villanueva J (2001) The fine geometry of the Cantor families of invariant tori in Hamiltonian systems. In: Casacuberta C, Miró-Roig RM, Verdera J, Xambó-Descamps S (eds) European Congress of Mathematics, Barcelona, 2000, vol 2. Progress in Mathematics, vol 202. Birkhäuser, Basel, pp 557–564
54. Koçak H (1984) Normal forms and versal deformation of linear Hamiltonian systems. J Diff Eqns 51:359–407
55. Marco JP, Sauzin D (2002) Stability and instability for Gevrey quasi-convex near-integrable Hamiltonian systems. Publ IHES 96:199–275
56. Marsden JE, McCracken M (1976) The Hopf-bifurcation and its applications. Springer, New York
57. Meyer KR (1984) Normal forms for the general equilibrium. Funkcial Ekva 27:261–271
58. Milnor JW (2006) Dynamics in one complex variable, 3rd edn. Ann Math Stud, vol 160. Princeton Univ Press, Princeton
59. Moser JK (1968) Lectures on Hamiltonian systems. Mem Amer Math Soc 81:1–60
60. Murdock J (2003) Normal forms and unfoldings for local dynamical systems. Springer Monographs in Mathematics. Springer, New York
61. Narasimhan R (1968) Analysis on real and complex manifolds. Elsevier, North Holland
62. Neĭshtadt AI (1975) Passage through resonances in two-frequency problem. Sov Phys Dokl 20(3):189–191 (English, Russian original)
63. Neĭshtadt AI (1984) The separation of motions in systems with rapidly rotating phase. J Appl Math Mech 48(2):133–139 (English, Russian original)
64. Nekhoroshev NN (1971) On the behavior of Hamiltonian systems close to integrable ones. Funct Anal Appl 5:338–339 (English, Russian original)
65. Nekhoroshev NN (1977) An exponential estimate of the stability time of nearly integrable Hamiltonian systems. I, Russian Math Surv 32(6):1–65 (English, Russian original)
66. Nekhoroshev NN (1979) An exponential estimate of the stability time of nearly integrable Hamiltonian systems, vol II. Trudy Sem Imeni. In: Petrovskogo IG pp 5:5–50 (in Russian) English translation. In: Oleĭnik OA (ed) (1985) Topics in modern mathematics, Petrovskiĭ Seminar. Consultants Bureau, New York, pp 5:1–58
67. Palis J de Melo WC (1982) Geometric Theory of Dynamical Systems. Springer
68. Palis J, Takens F (1993) Hyperbolicity & sensitive chaotic dynamics at homoclinic bifurcations. Cambridge Studies in Advanced Mathematics, Cambridge University Press 35
69. Poincaré H (1928) Œuvres, vol I. Gauthier-Villars
70. Popov G (2000) Invariant tori, effective stability, and quasimodes with exponentially small error terms. I: Birkhoff normal forms. Ann Henri Poincaré 1(2):223–248

71. Popov G (2004) KAM theorem for Gevrey Hamiltonians. Ergod Theor Dynam Syst 24:1753–1786
72. Ramis JP (1994) Séries divergentes et théories asymptotiques. Panor Synth pp 0–74
73. Ramis JP, Schäfke R (1996) Gevrey separation of slow and fast variables. Nonlinearity 9:353–384
74. Roussarie R (1987) Weak and continuous equivalences for families of line diffeomorphisms. In: Dynamical systems and bifurcation theory, Pitman research notes in math. Series Longman 160:377–385
75. Sanders JA, Verhulst F, Murdock J (1985) Averaging methods in nonlinear dynamical systems. Revised 2nd edn, Appl Math Sciences 59, 2007. Springer
76. Siegel CL (1942) Iteration of analytic functions. Ann Math 43(2):607–612
77. Siegel CL, Moser JK (1971) Lectures on celestial mechanics. Springer, Berlin
78. Simó C (1994) Averaging under fast quasiperiodic forcing. In: Seimenis J (ed) Hamiltonian mechanics, integrability and chaotic behaviour. NATO Adv Sci Inst Ser B Phys 331:13–34, Plenum, New York
79. Simó C (1998) Effective computations in celestial mechanics and astrodynamics. In: Rumyantsev VV, Karapetyan AV (eds) Modern methods of analytical mechanics and their applications. CISM Courses Lectures, vol 387. Springer, pp 55–102
80. Simó C (2000) Analytic and numeric computations of exponentially small phenomena. In: Fiedler B, Gröger K, Sprekels J (eds) Proceedings EQUADIFF 99, Berlin. World Scientific, Singapore, pp 967–976
81. Spivak M (1970) Differential Geometry, vol I. Publish or Perish Inc
82. Sternberg S (1959) On the structure of local homeomorphisms of Euclidean *n*-space, vol II. Amer J Math 81:578–605
83. Takens F (1973) Forced oscillations and bifurcations. Applications of global analysis, vol I, Utrecht. Comm Math Inst Univ Utrecht, pp 31–59 (1974). Reprinted In: Broer HW, Krauskopf B, Vegter G (eds) (2001) Global analysis of dynamical systems. Festschrift dedicated to floris takens for his 60th birthday Leiden. Inst Phys Bristol, pp 1–61
84. Takens F (1974) Singularities of vector fields. Publ Math IHÉS 43:47–100
85. Takens F, Vanderbauwhede A (2009) Local invariant manifolds and normal forms. In: Broer HW, Hasselblatt B, Takens F (eds) Handbook of dynamical systems, vol 3. North-Holland (to appear)
86. Thom R (1989) Structural stability and morphogenesis. An outline of a general theory of models, 2nd edn. Addison-Wesley, Redwood City (English, French original)
87. Vanderbauwhede A (1989) Centre manifolds, normal forms and elementary bifurcations. Dyn Rep 2:89–170
88. van der Meer JC (1985) The hamiltonian Hopf bifurcation. LNM, vol 1160. Springer
89. Varadarajan VS (1974) Lie groups, Lie algebras and their representations. Englewood Cliffs, Prentice-Hall
90. Wagener FOO (2003) A note on Gevrey regular KAM theory and the inverse approximation lemma. Dyn Syst 18(2):159–163
91. Wagener FOO (2005) On the quasi-periodic *d*-fold degenerate bifurcation. J Differ Eqn 216:261–281
92. Yoccoz JC (1995) Théorème de Siegel, nombres de Bruno et polynômes quadratiques. Astérisque 231:3–88

Books and Reviews

Braaksma BLJ, Stolovitch L (2007) Small divisors and large multipliers (Petits diviseurs et grands multiplicateurs). Ann l'institut Fourier 57(2):603–628
Broer HW, Levi M (1995) Geometrical aspects of stability theory for Hill's equations. Arch Rat Mech An 131:225–240
Gaeta G (1999) Poincaré renormalized forms. Ann Inst Henri Poincaré 70(6):461–514
Martinet J, Ramis JP (1982) Problèmes des modules pour les èquations différentielles non linéaires du premier ordre. Publ IHES 5563–164
Martinet J Ramis JP (1983) Classification analytique des équations différentielles non linéaires résonnantes du premier ordre. Ann Sci École Norm Suprieure Sér 416(4):571–621
Vanderbauwhede A (2000) Subharmonic bifurcation at multiple resonances. In: Elaydi S, Allen F, Elkhader A, Mughrabi T, Saleh M (eds) Proceedings of the mathematics conference, Birzeit, August 1998. World Scientific, Singapore, pp 254–276

Numerical Bifurcation Analysis

Hil Meijer[1], Fabio Dercole[2], Bart Oldeman[3]
[1] Department of Electrical Engineering, Mathematics and Computer Science, University of Twente, Enschede, The Netherlands
[2] Department of Electronics and Information, Politecnico di Milano, Milano, Italy
[3] Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
Continuation and Discretization of Solutions
Normal Forms and the Center Manifold
Continuation and Detection of Bifurcations
Branch Switching
Connecting Orbits
Software Environments
Future Directions
Bibliography

Glossary

Dynamical system A rule for time evolution on a state space. The term system will be used interchangeably. Here a system is a family given by an ordinary differential equation (ODE) depending on parameters.

Equilibrium A constant solution of the system, for given parameter values.

Limit cycle An isolated periodic solution of the system, for given parameter values.

Bifurcation A qualitative change in the dynamics of a dynamical system produced by changing its parameters. Bifurcation points are the critical parameter combinations at which this happens for arbitrarily small parameter perturbations.

Normal form A simplified model system for the analysis of a certain type of bifurcation.

Codimension The minimal number of parameters needed to perturb a family of systems in a generic manner.

Defining system A set of suitable equations so that the zero set corresponds to a bifurcation of a certain type or to a particular solution of the system. Also called defining function or equation.

Continuation A numerical method suited for tracing one-dimensional manifolds, curves (here called branches) of solutions for a defining system while one or more parameters are varied.

Test function A function designed to have a regular zero at a bifurcation. During continuation a test function can be monitored to detect bifurcations.

Branch switching Several branches of different codimension can emanate from a bifurcation point. Switching from the computation of one branch to an other requires appropriate procedures.

Definition of the Subject

The theory of dynamical systems studies the behavior of solutions of systems, like nonlinear ordinary differential equations (ODEs), depending upon parameters. Using qualitative methods of bifurcation theory, the behavior of the system is characterized for various parameter combinations. In particular, the catalog of system behaviors showing qualitative differences can be identified, together with the regions in parameter space where the different behaviors occur. Bifurcations delimit such regions. Symbolic and analytical approaches are in general infeasible, but numerical bifurcation analysis is a powerful tool that aids in the understanding of a nonlinear system. When computing power became widely available, algorithms for this type of analysis matured and the first codes were developed. With the development of suitable algorithms, the advancement in the qualitative theory has found its way into several software projects evolving over time. The availability of software packages allows scientists to study and adjust their models and to draw conclusions about their dynamics.

Introduction

Nonlinear ordinary differential equations depending upon parameters are ubiquitous in science. In this article methods for *numerical bifurcation analysis* are reviewed, an approach to investigate the dynamic behavior of nonlinear dynamical systems given by

$$\dot{x} = f(x, p)\,, \quad x \in \mathbb{R}^n\,, \quad p \in \mathbb{R}^{n_p}\,, \tag{1}$$

where $f\colon \mathbb{R}^n \times \mathbb{R}^{n_p} \to \mathbb{R}^n$ is generic and sufficiently smooth. In particular, $x(t)$ represents the state of the system at time t and its components are called *state* (or *phase*) *variables*, $\dot{x}(t)$ denotes the time derivative of $x(t)$, while p denotes the *parameters* of the system, representing experimental control settings or variable inputs.

In many instances, solutions of (1), starting at an initial condition $x(0)$, appear to converge as $t \to \infty$ to *equilibria* (steady states) or *limit cycles* (periodic orbits). Bounded solutions can also converge to more complex *attractors*, like *tori* (quasi-periodic orbits) or *strange attractors* (chaotic orbits). The attractors of the system are invariant under time evolution, i. e., under the application of the time-t map Φ^t, where Φ denotes the flow induced by the system (1). Solutions attracting all nearby initial conditions, are said to be *stable*, and *unstable* if they repel some initial conditions.

Generally speaking, it is hard to obtain closed formulas for Φ^t as the system is nonlinear. In some cases, one can compute equilibria analytically, but this is often not the case for limit cycles. However, numerical simulations of (1) easily give an idea of how solutions look, although one never computes the true orbit due to numerical errors. One can verify stability conditions by linearizing the flow around equilibria and cycles. In particular, an equilibrium x_0 is stable if the eigenvalues of the linearization (*Jacobi*) matrix $A = f_x(x_0, p)$ (where the subscript denotes differentiation) all have a negative real part. Similarly, for a limit cycle $x_0(t)$ with period T, one defines the *Floquet multipliers* (or simply multipliers) as the eigenvalues of the monodromy matrix $M = \Phi_x^T(x_0(0))$. The cycle is stable if the nontrivial multipliers, there is always one equal to 1, are all within the unit circle. Equilibria and limit cycles are called *hyperbolic* if the eigenvalues and nontrivial multipliers do not have a zero real part or modulus one, respectively.

For any given parameter combination p, the state space representation of all orbits constitutes the *phase portrait* of the system. In practice, one draws a set of strategic orbits (or finite segments of them), from which all other orbits can be intuitively inferred, as illustrated in Fig. 1.

Points X_0, X_1, X_2 are equilibria, of which X_0 and X_1 are unstable and X_2 is stable. In particular, X_0 is a *repellor*, i. e., nearby orbits do not tend to remain close to X_0, while X_1 is a *saddle*, i. e., almost all nearby orbits go away from X_1 except two, which tend to X_1 and lie on the so-called *stable manifold*; the two orbits emanating from X_1

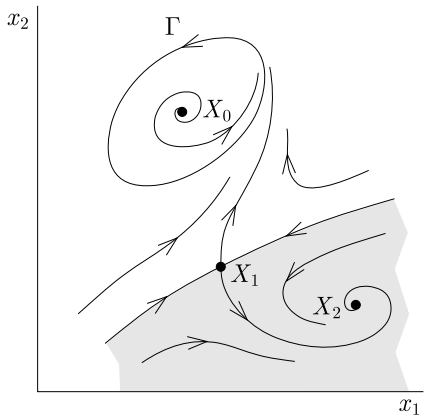

Numerical Bifurcation Analysis, Figure 1
Phase portrait of a two-dimensional system with two attractors (the equilibrium X_2 and the limit cycle Γ), a repellor (X_0), and a saddle (X_1)

compose the *unstable manifold*. There are therefore two attractors, the equilibrium X_2 and the limit cycle Γ, whose basins of attraction consist of the initial conditions in the shaded and white areas, respectively. Note that while attractors and repellors can be easily obtained through simulation, forward and backward in time, saddles can be hard to find.

The analysis of system (1) becomes even more difficult if one wants to follow the phase portrait under variation of parameters. Generically, by perturbing a parameter slightly the phase portrait changes slightly as well. Namely, if the new phase portrait is topologically equivalent to the original one, then nothing changed from a qualitative point of view, i. e., all attracting, repelling, and saddle sets are still present with unchanged stability properties, though slightly perturbed. By contrast, the critical points in parameter space where arbitrarily small parameter perturbations give rise to nonequivalent phase portraits are called *bifurcation* points, where bifurcations are said to occur. Bifurcations therefore result in a partition of parameter space into regions: parameter combinations in the same region correspond to topologically equivalent dynamics, while nonequivalent phase portraits arise for parameter combinations in neighboring regions. Most often, this partition is represented by means of a two-dimensional *bifurcation diagram*, where the regions of a parameter plane are separated by so-called *bifurcation curves*. Bifurcations are said to be *local* if they occur in an arbitrarily small neighborhood of the equilibrium or cycle; otherwise, they are said to be *global*.

Although one might hope to detect bifurcations by simulating system (1) for various parameter combinations and initial conditions, a "brute force" simulation approach is hardly effective and accurate in practice, because bifurcations of equilibria and cycles are associated with a loss of hyperbolicity, e. g., stability, so that one should dramatically increase the length of simulations while approaching the bifurcation. In particular, saddle sets are hard to find by simulation, but play a fundamental role in bifurcation analysis, since they, together with attracting and repelling sets, form the skeleton of the phase portrait. This is why numerical bifurcation analysis does not rely on simulation, but rather on *continuation*, a numerical method suited for computing (approximating through a discrete sequence of points) one-dimensional manifolds (curves, "branches" in regular) implicitly defined as the zero set of a suitable *defining function*.

The general idea is to formulate the computation of equilibria and their bifurcations as a suitable *algebraic problem* (AP) of the form

$$F(u, p) = 0 \,, \tag{2}$$

where $u \in \mathbb{R}^{n_u}$ is composed of x and possibly other variables characterizing the system, see, e. g., defining functions as in Sect. "Continuation and Detection of Bifurcations". Here, however, for simplicity of notation, u will be considered as in $\mathbb{R}^n$, but the actual dimension of u will always be clear from the context. Similarly, limit cycles and their bifurcations are formulated in the form of a *boundary-value problem* (BVP)

$$\begin{cases} \dot{u} - f(u, p) = 0 \,, \\ g(u(0), u(T), p) = 0 \,, \\ \int_0^T h(u(t), p)\mathrm{d}t = 0 \,, \end{cases} \tag{3}$$

with n_b boundary conditions, i. e., $g\colon \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^{n_p} \to \mathbb{R}^{n_b}$, n_i integral conditions, i. e., $h\colon \mathbb{R}^n \times \mathbb{R}^{n_p} \to \mathbb{R}^{n_i}$, and u in a proper function space. In other words, a list of defining functions is formulated, in the form (2) or (3), to cover all cases of interest. For example, $u = x$ and $F(x, p) = f(x, p)$ is the AP defining equilibria of (1). The commonly used cycle BVP, with the time-rescaling $t = T\tau$, is

$$\begin{cases} \dot{x} - Tf(x, p) = 0 \,, \\ x(0) - x(1) = 0 \,, \\ \int_0^1 x(\tau)^\top \dot{x}^{k-1}(\tau)\mathrm{d}\tau = 0 \,, \end{cases} \tag{4}$$

where from here on $^\top$ denotes the transpose and, for simplicity, $\dot{}$ for $\mathrm{d}/\mathrm{d}\tau$ is used. The integral condition is the so-called *phase condition*, which ensures that $x(\tau)$ is the 1 periodic solution of (1) closest to the reference solution $x^{k-1}(\tau)$ (typically known from the previous point along the continuation), among time-shifted orbits $x(\tau - \tau_0)$, $\tau_0 \in [0, 1]$.

As will be discussed in Sect. "Discretization of BVPs", a proper time-discretization of $u(\tau)$ allows one to approximate any BVP by a suitable AP. Thus, equilibria, limit cycles and their bifurcations can all be represented by an algebraic defining function like (2), and numerical continuation allows one to produce one-dimensional solution branches of (2) under the variation of strategic components of p, called *free* parameters. With this approach, equilibria and cycles can be followed without further difficulty in parameter regimes where these are unstable. Then, during continuation, the stability of equilibria and cycles is determined through linearization. Moreover, the characterization of nearby solutions of (1) can be done using *normal forms*, i. e., the simplest canonical models to which the system, close to a bifurcation, can be reduced on a lower-dimensional manifold of the state space, the so-called *center manifold*. While a branch of equilibria or cycles is followed, bifurcations can be detected as the zero of suitable test functions. Upon detection of a bifurcation, the defining function can be augmented by this test function or another appropriate function, and the new defining function can then be continued using one more free parameter.

An analytical bifurcation study is feasible for simple systems only. Numerical bifurcation analysis is one of the few but also very powerful tools to understand and describe the dynamics of systems depending on parameters. Some basic steps while performing bifurcation analysis will be outlined and software implementations of continuation and bifurcation algorithms discussed.

First a few standard and often used approaches for the computation and continuation of zeros of a defining function are reviewed in Sect. "Continuation and Discretization of Solutions". The presentation starts with the most obvious, but also naive, approaches to contrast these with the methods employed by software packages. In Sect. "Normal Forms and the Center Manifold" several possible scenarios for the loss of stability of equilibria and limit cycles are discussed. Not all bifurcations are characterized by linearization and for the detection and analysis of these bifurcations, codimension 1 normal forms are mentioned and a general method for their computation on a center manifold is presented. Then, a list of suitable test functions and defining systems for the computation of bifurcation branches is discussed in Sect. "Continuation and Detection of Bifurcations". In particular, when a system bifurcates new solution branches appear. Techniques to switch to such new branches are described in Sect. "Branch Switching". Finally, the computation and continuation of global bifurcations characterized by orbits connecting equilibria is presented, in particular *homoclinic* orbits, in Sect. "Connecting Orbits". This review concludes with an overview of existing implementations of the described algorithms in Sect. "Software Environments". Previous reviews [7,9,22,43] have similar contents. This review however, focuses more on the principles now underlying the most frequently used software packages for bifurcation analysis and the algorithms being used.

Continuation and Discretization of Solutions

The continuation of a solution u of (2) with respect to one parameter p is a fundamental application of the Implicit Function Theorem (IFT).

Generally speaking, to define one-dimensional solution manifolds (branches), the number of unknowns in (2) should be one more than the number of equations, i. e., $n_p = 1$. However, during continuation it is better not to distinguish between state variables and parameters as will become apparent in Sects. "Pseudo-Arclength Continuation" and "Moore–Penrose Continuation". Therefore, write $y = (u, p) \in Y = \mathbb{R}^{n+1}$ for the continuation variables in the continuation space Y and consider the continuation problem

$$F(y) = 0\,, \tag{5}$$

with $F\colon \mathbb{R}^{n+1} \to \mathbb{R}^n$.

Let F be at least continuous differentiable, $y_0 = (u_0, p_0)$ be a known solution point of (5), and the matrix $F_y(y_0) = [F_u(u_0, p_0)|F_p(u_0, p_0)]$ be full rank, i. e.,

$$\operatorname{rank}(F_y(y_0)) = n \Leftrightarrow \begin{cases} \text{(i)} & \operatorname{rank}(F_u(u_0, p_0)) = n\,, \text{ or} \\ \text{(ii)} & \operatorname{rank}(F_u(u_0, p_0)) = n-1 \text{ and} \\ & F_p(u_0, p_0) \notin \mathcal{R}(F_u(u_0, p_0))\,, \end{cases} \tag{6}$$

where $\mathcal{R}(F_u)$ denotes the range of F_u. Then the IFT states that there exists a unique solution branch of (5) locally to y_0. Introducing a scalar coordinate s parametrizing the branch, e. g., the *arclength* positively measured from y_0 in one of the two directions along the solution branch, then one can represent the branch by $y(s) = (u(s), p(s))$ and the IFT guarantees that $F(y(s)) = 0$ for $|s|$ is sufficiently small. Moreover, $y(s)$ is continuous differentiable and the vector $\phi(s) = y_s(s) = (u_s(s), p_s(s)) = (v(s), q(s))$, tangent to the solution branch at $y(s)$, exists and is the unique solution of

$$\begin{cases} F_y(y(s))\phi(s) = F_u(u(s)\,,\ p(s))v(s) + F_p(u(s)\,, \\ \qquad\qquad\qquad\qquad\qquad\qquad p(s))q(s) = 0\,, \\ \phi(s)^\top\phi(s) = v(s)^\top v(s) + q(s)^2 = 1\,. \end{cases} \tag{7}$$

In other words, the matrix $F_y(y_0)$ has a one-dimensional nullspace $\mathcal{N}(F_y(y_0))$ spanned by $\phi(0)$ and y_0 is said to be a *regular* point of the continuation space Y.

Below, several variants of numerical continuation will be described. The aim is to produce a sequence of points y_k, $k \geq 0$ that approximate the solution branch $y(s)$ in one direction. Starting from y_0, the general idea is to make a suitable prediction y_1^0, typically along the tangent vector, from which the Newton method is applied to find the new point y_1. The predictor-corrector procedure is then iterated. First, the simplest implementation is presented and it is shown where it might fail. Many continuation packages for bifurcation theory use an alternative implementation of which two variants are discussed. Many more advanced predictor-corrector schemes have been designed, see [1,18,46] and references therein.

Parameter Continuation

Parameter continuation assumes that the solution branch of (5) can be parameterized by the parameter $p \in \mathbb{R}$. Indeed, if F_u has full rank, i. e., case (i) in (6), then this is possible by the IFT. Starting from (u_0, p_0) and perturbing the parameter a little, with a *stepsize* h, the new parameter is $p_1 = p_0 + h$ and the most simple predictor for the state variable is given by $u_1^0 = u_0$.

Application of Newton's method to find u_1 satisfying (5) leads to

$$u_1^{j+1} = u_1^j - F_u(u_1^j, p_1)^{-1} F(u_1^j, p_1)\,, \quad j = 0, 1, 2, \ldots$$

The iterations are stopped when a certain accuracy is achieved, i. e., $\|\Delta u\| = \|u_1^{j+1} - u_1^j\| < \varepsilon_u$ and/or $\|F(u_1^j, p_1)\| < \varepsilon_F$. In practice, also the maximum number of Newton steps is bounded, in order to guarantee termination. If this maximum is reached before convergence, the computation is restarted with a smaller (typically halved) stepsize. However, in case of quick convergence, e. g., after only a few iterations, the stepsize is multiplied (1.3 is a typical factor). In any case, the stepsize is varied between two assigned limits $h_{\min}$ and $h_{\max}$, so that continuation cannot proceed when convergence is not reached even with minimum stepsize. When h is chosen too small, too much computational work may be performed, while for h that is chosen too large, little detail of the solution branch is obtained.

As a first improved predictor, note that the IFT suggests to use the tangent prediction for the state variables $u_1^0 = u_0 + hv_0$, where v_0 is obtained from (7) with $s = 0$ and as $v(0)/q(0)$. Indeed, the tangent vector can be approximated by the difference $v_k = (u_k - u_{k-1})/h_k$ or, even better, computed at negligible cost, since the numerical decomposition of the matrix $F_u(u_k, p_k)$ is known from the last Newton iteration.

These methods are illustrated in Fig. 2. Note in particular the folds in the sketch. Here parameter continuation does not work, since exactly at the fold $F_u(u, p)$ is singular such that Newton's method does not converge and beyond, for larger p, there is no local solution to (5).

Pseudo-Arclength Continuation

Near folds, the solution branch is not well parameterized by the parameter, but one can use a state variable for the parametrization. In fact, the fold is a regular point (case (ii) in (6)) at which the tangent vector $\phi = (v, q)$ has no parameter component, i. e., $q = 0$. So, without distinguishing between parameters and state variables, one takes the

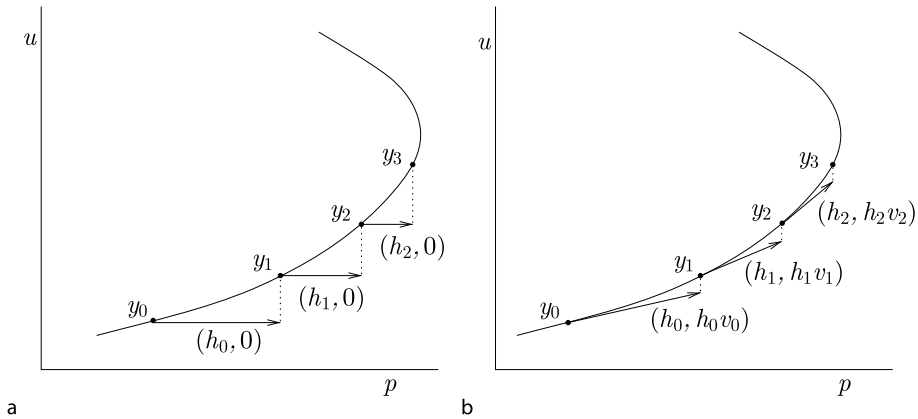

Numerical Bifurcation Analysis, Figure 2
Parameter continuation without (a) and with (b) tangent prediction. The *dotted lines* indicate subspaces where solutions are searched

tangent prediction $y_1^0 = y_0 + h\phi_0$, as long as the starting solution y_0 is a regular point. Since now both p and u are corrected, one more constraint is needed. Pseudo-arclength continuation uses the stepsize h as an approximation of the required distance, in arclength, between y_0 and the next point y_1. This leads to the so-called *pseudo-arclength equation* $\phi_0^\top(y_1 - y_0) = h$. In this way, solution branches can be followed past folds. The idea for this continuation method is due to Keller [50].

The Newton iteration, applied to

$$\begin{cases} F(y_1) = 0\,, \\ \phi_0^\top(y_1 - y_0) - h = 0\,, \end{cases}$$

is given by

$$y_1^{j+1} = y_1^j - \begin{pmatrix} cF_y(y_1^j) \\ \phi_0^\top \end{pmatrix}^{-1} \begin{pmatrix} cF(y_1^j) \\ 0 \end{pmatrix}, \quad j = 0, 1, 2, \ldots, \tag{8}$$

where $\Delta y = y_1^{j+1} - y_1^j$ is forced to lie in the hyperplane orthogonal to the tangent vector, as illustrated in Fig. 3a. Upon convergence, the new tangent vector ϕ_1 is obtained by solving (7) at y_1.

Moore–Penrose Continuation

This continuation method is based on optimization. Starting with the tangent prediction $y_1^0 = y_0 + h\phi_0$, a point y_1 with $F(y_1) = 0$ nearest to y_1^0 is searched, so the following is optimized

$$\min_{y_1}\{\|y_1 - y_1^0\| \,|\, F(y_1) = 0\}\,.$$

Each correction is therefore required to be orthogonal to the nullspace of $F_y(y_1^j)$, i. e.,

$$\begin{cases} F(y_1^{j+1}) = 0\,, \\ (\phi_1^j)^\top(y_1^{j+1} - y_1^j) = 0\,. \end{cases}$$

Starting with $\phi_1^0 = \phi_0$ the Newton iterations are given by

$$\begin{cases} y_1^{j+1} = y_1^j - \begin{pmatrix} F_y(y_1^j) \\ (\phi_1^j)^\top \end{pmatrix}^{-1} \begin{pmatrix} F(y_1^j) \\ 0 \end{pmatrix}, \\ \phi_1^{j+1} = \begin{pmatrix} F_y(y_1^{j+1}) \\ (\phi_1^j)^\top \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad j = 0, 1, 2, \ldots \end{cases} \tag{9}$$

As illustrated in Fig. 3b, the Moore–Penrose continuation can be interpreted as a variant of Keller's method in which the tangent vector is updated at every Newton step. When the new point y_1 is found, the tangent vector ϕ_1 is immediately obtained as $\phi_1^{j+1}/\|\phi_1^{j+1}\|$ from the last Newton iteration, since ϕ_1^j does not necessarily have unit length.

Finally, for both the pseudo-arclength and Moore-Penrose continuation methods one can prove that they converge (with superlinear convergence), provided that y_0 is a regular point and the stepsize is sufficiently small.

Discretization of BVPs

In this section, *orthogonal collocation* [3,15] is described, a discretization technique to approximate the solution of a generic BVP (3) by a suitable AP. Let u be at least in the space $C^1([0,1], \mathbb{R}^n)$ of continuous differentiable vector-valued functions defined on $[0, 1]$. For BVPs the rescaled

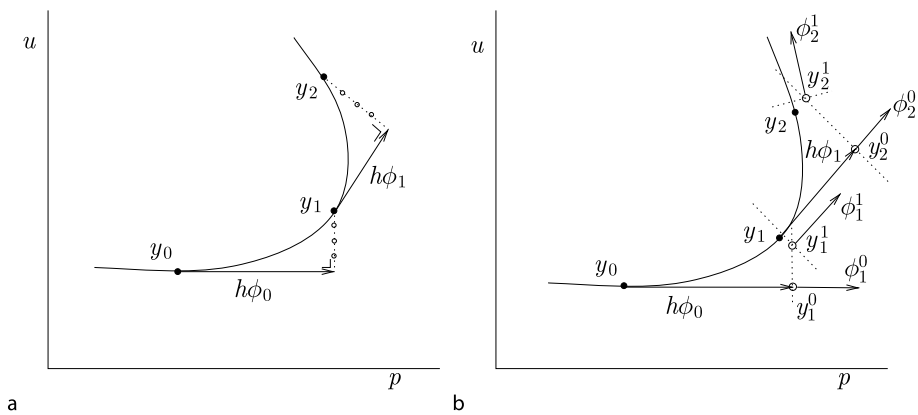

Numerical Bifurcation Analysis, Figure 3

Pseudo-arclength (a) and Moore–Penrose (b) continuation. Searching a solution in hyperplanes without (a) and with (b) updating the tangent vector. The *open dots* correspond to Newton iterations, *full dots* to points on the curve. The *dotted lines* indicate subspaces where solutions are searched

time $t = T\tau$ is used, so that the period T becomes a parameter and in the sequel T will be addressed as such. Introduce a time mesh $0 = \tau_0 < \tau_1 < \ldots < \tau_N = 1$ and, on each interval $[\tau_{j-1}, \tau_j]$, approximate the function u by a vector-valued polynomial $\wp_j$ of degree m, $j = 1, \ldots, N$. The polynomials $\wp_j$ are determined by imposing the ODE in (3) at m *collocation points* $z_{j,i}$, $i = 1, \ldots, m$, i. e.,

$$\dot{\wp}_j(z_{j,i}) = f(\wp_j(z_{j,i}), p), \quad j = 1, \ldots, N, \; i = 1, \ldots, m. \tag{10}$$

One usually chooses the so-called Gauss points as the collocation points, the roots of the mth order Legendre polynomials. Moreover, $\wp_1(0)$ and $\wp_N(1)$ must satisfy the boundary conditions and the whole piecewise polynomial must satisfy the integral conditions.

Counting the number of unknowns, the discretization of (3) leads to nN polynomials, each with $(m+1)$ degrees of freedom, plus n_p free parameters, so there are $nN(m+1) + n_p$ continuation variables. These variables are matched by nmN collocation equations from (10), $n(N-1)$ continuity conditions at the mesh points, n_b boundary conditions, and n_i integral conditions, for a total of $nN(m+1) + n_b + n_i - n$ algebraic equations. Thus, in order for these equations to compose an AP, the number of free parameters is generically $n_p = n_b + n_i - n + 1$ and, typically, one is the period T.

The collocation method yields high accuracy with superconvergence at the mesh points [15]. The mesh can also be adapted during continuation, for instance to minimize the local discretization error [68]. The equations can be solved efficiently by exploiting the particular sparsity structure of the Jacobi matrix in the Newton iteration. In particular, a few full but essentially smaller systems are solved instead of one sparse but large system. During this process one finds two nonsingular $(n \times n)$ -submatrices M_0 and M_1 such that $M_0 u(0) + M_1 u(1) = 0$, i. e., the monodromy matrix $M = -M_1^{-1} M_0$ is found as a by-product. In the case of a periodic BVP ($u(1) = u(0)$) the Floquet multipliers are therefore computed at low computational cost.

Normal Forms and the Center Manifold

Local bifurcation analysis relies on the reduction of the dynamics of system (1) to a lower-dimensional center manifold H_0 near nonhyperbolic equilibria or limit cycles, i. e., when they bifurcate at some critical parameter p_0. The existence of H_0 follows from the Center Manifold Theorem (CMT), see, e. g., [11], while the reduction principle is shown in [72]. The reduced ODE has the same dimension as H_0 given by the number n_c of critical eigenvalues or nontrivial multipliers (counting multiplicity), and is transformed to a normal form. The power of this approach is that the bifurcation scenario of the normal form is preserved in the original system. The normal form for a specific bifurcation is usually studied only up to a finite order, i. e., truncated, and many diagrams for bifurcations with higher codimension are in principle incomplete due to global phenomena, such as connecting orbits. Also, H_0 is not necessarily unique or smooth [75], but fortunately, one can still draw some useful qualitative conclusions.

The codimension (codim) of a bifurcation is the minimal number of parameters needed to encounter the bifurcation and to unfold the corresponding normal form generically. Therefore, in practice, one finds codim 1 phenomena when varying a single parameter, and continues them as curves in two-parameter planes. Codim 2 phenomena are found as isolated points along codim 1 bifurcation curves. Still, codim 2 bifurcations are important as they are the roots of codim 1 bifurcations, in particular of global phenomena. For this reason they are called organizing centers as, around these points in parameter space, one-parameter bifurcation scenarios change. For parameter-dependent systems the center manifold H_0 can be extended to a parameter-dependent invariant manifold $H(p)$, $H_0 = H(p_0)$, so that the bifurcation scenario on $H(p)$ is preserved in the original system for $\|p - p_0\|$ sufficiently small.

In the following, the normal forms for all codim 1 bifurcations of equilibria and limit cycles are presented and their bifurcation scenarios discussed. Then a general computational method for the approximation, up to a finite order, of the parameter-dependent center manifold $H(p)$ is presented. The method gives, as a by-product, explicit formulas for the coefficients of a given normal form in terms of the vector field f of system (1).

Normal Forms

Bifurcations can be defined by certain algebraic conditions. For instance, an equilibrium is nonhyperbolic if $\Re(\lambda) = 0$ holds for some eigenvalue. The simplest possibilities are $\lambda = 0$ (limit point bifurcation or branch point, though the latter is nongeneric, see Sect. "Branch Switching") and $\lambda_{1,2} = \pm i\omega_0, \omega_0 > 0$ (Hopf bifurcation). Bifurcations of limit cycles appear if some of the nontrivial multipliers cross the unit circle. The three simplest possibilities are $\mu = 1$ (limit point of cycles), $\mu = -1$ (period-doubling) or $\mu_{1,2} = \mathrm{e}^{\pm i\theta_0}, 0 < \theta_0 < \pi$ (Neimark–Sacker).

At the bifurcation ($p = p_0$), linearization of system (1) near the equilibrium x_0 or around a limit cycle, does not

result in any stability information in the center manifold. In this case, nonlinear terms are also necessary to obtain such knowledge. This is provided by the critical normal form coefficients as discussed below. The state variable in the normal form will be denoted by w and the unfolding parameter by $\alpha \in \mathbb{R}$, with $w = 0$ at $\alpha = 0$ being a nonhyperbolic equilibrium. Bifurcations are labeled in accordance with the scheme of [38].

Codimension 1 Bifurcations of Equilibria

Limit point bifurcation (LP): The equilibrium has a simple eigenvalue $\lambda = 0$ and the restriction of (1) to a one-dimensional center manifold can be transformed to the normal form

$$\dot{w} = \alpha + a_{\mathrm{LP}} w^2 + O(|w|^3)\,, \quad w \in \mathbb{R}\,, \tag{11}$$

where generically $a_{\mathrm{LP}} \neq 0$ and O denotes higher order terms in state variables depending on parameters too. When the unfolding parameter α crosses the critical value ($\alpha = 0$), two equilibria, one stable and one unstable in the center manifold, collide and disappear. This bifurcation is also called *saddle-node*, *fold* or *tangent* bifurcation. Note that this bifurcation occurs at the folds in Figs. 2 and 3.

Hopf bifurcation (H): The equilibrium has a complex pair of eigenvalues $\lambda_1 = -\lambda_2 = i\omega_0$ and the restriction of (1) to the two-dimensional center manifold is given by

$$\dot{w} = (i\omega_0 + \alpha)w + c_{\mathrm{H}} w^2 \bar{w} + O(|w|^4)\,, \quad w \in \mathbb{C}\,, \tag{12}$$

where generically the first Lyapunov coefficient $d_{\mathrm{H}} = \Re(c_{\mathrm{H}}) \neq 0$. When α crosses the critical value, a limit cycle is born. It is stable (and present for $\alpha > 0$) if $d_{\mathrm{H}} < 0$ and unstable if $d_{\mathrm{H}} > 0$ (and present for $\alpha < 0$). The case $d_{\mathrm{H}} < 0$ is called supercritical or "soft", while $d_{\mathrm{H}} > 0$ is called subcritical or "hard" as there is no (local) attractor left after the bifurcation. This bifurcation is most often called Hopf, but also Poincaré–Andronov–Hopf as this was also known to the first two.

Codimension 1 Bifurcations of Limit Cycles Bifurcations of limit cycles are theoretically very well understood using the notion of a Poincaré map. To define this map, choose a $(n-1)$-dimensional smooth cross-section Σ transversal to the cycle and introduce a local coordinate $z \in \mathbb{R}^{n-1}$ such that $z = Z(x)$ is defined on Σ and invertible. For example, one chooses a coordinate plane $x_j = 0$ such that $f_j(x)|_{x_j=0} \neq 0$. Let $x_0(t)$ be the cycle with period T, so that $z_0 = Z(x_0(0))$ is the cycle intersection with Σ, where $z = 0$ can always be assumed without loss of generality. Denote by $T(z)$ the return time to Σ defined by the flow Φ with $T(z_0) = T$. Now, the Poincaré map $P\colon \mathbb{R}^{n-1} \to \mathbb{R}^{n-1}$ maps each point close enough to $z = 0$ to the next return point on Σ, i. e., $P\colon z \mapsto Z(\Phi^{T(z)}(Z^{-1}(z)))$. Thus, bifurcations of limit cycles turn into bifurcations of fixed points of the Poincaré map which can be easily described using local bifurcation theory. Moreover, it can be shown that the $n-1$ eigenvalues of the linearization $P_z(0)$ are the nontrivial eigenvalues of the monodromy matrix $M = \Phi_x^T(x_0(0))$, which also has a trivial eigenvalue equal to 1 (the vector $f(x_0(0), p)$, tangent to the cycle at $x_0(0)$, is mapped by M to itself). The eigenvalues of $P_z(0)$ are therefore the nontrivial multipliers of the cycle. Although the Poincaré map and its linearization can also be computed numerically through suitably organized simulations (so-called *shooting* techniques [17]), it is better to handle both the cycle multipliers and normal form computations associated to nonhyperbolic cycles using BVPs [9,23,57]. Here, however, the normal forms on a Poincaré section are presented, where $w = 0$ at $\alpha = 0$ is the fixed point of the Poincaré map corresponding to a nonhyperbolic limit cycle.

Limit point of cycles (LPC): The fixed point has one simple nontrivial multiplier $\mu = 1$ on the unit circle and the restriction of P to a one-dimensional center manifold has the form

$$w \mapsto \alpha + w + a_{\mathrm{LPC}} w^2 + O(w^3)\,, \quad w \in \mathbb{R}\,,$$

where $a_{\mathrm{LPC}} \neq 0$. As for the LP bifurcation two fixed points collide and disappear when α crosses the critical value, provided $a_{\mathrm{LPC}} \neq 0$. This implies the collision of two limit cycles of the original vector field f.

Period-doubling (PD): The fixed point has one simple multiplier $\mu = -1$ on the unit circle and the restriction of P to a one-dimensional center manifold can be transformed to the normal form

$$w \mapsto -(1+\alpha)w + b_{\mathrm{PD}} w^3 + O(w^4)\,, \quad w \in \mathbb{R}\,,$$

where $b_{\mathrm{PD}} \neq 0$. When the parameter α crosses the critical value and $b_{\mathrm{PD}} \neq 0$, a cycle of period 2 for P bifurcates from the fixed point corresponding to a limit cycle of period $2T$ for the original system (1). This phenomenon is also called the *flip* bifurcation. If b_{PD} is positive [negative], the bifurcation is supercritical [subcritical] and the double period cycle is stable [unstable] (and present for $\alpha > 0[\alpha < 0]$).

Neimark–Sacker (NS): The fixed point has simple critical multipliers $\mu_{1,2} = \mathrm{e}^{\pm i\theta_0}$ and no other multipliers on the unit circle. Assume that $\mathrm{e}^{ik\theta_0} \neq 1$ for $k = 1, 2, 3, 4$, i. e., there are no strong resonances. Then, the restriction of P to a two-dimensional center manifold can be transformed to the normal form

$$w \mapsto \mathrm{e}^{i\theta(\alpha)}(1+\alpha)w + c_{\mathrm{NS}} w^2 \bar{w} + O(|w|^4)\,, \quad w \in \mathbb{C}\,,$$

where c_{NS} is a complex number and $\theta(0) = \theta_0$. Provided $d_{\mathrm{NS}} = \Re(\mathrm{e}^{-i\theta_0} c_{\mathrm{NS}}) \neq 0$, a unique closed invariant curve for P appears around the fixed point, when α crosses the critical value. In the original vector field, this corresponds to the appearance of a two-dimensional torus with (quasi-) periodic motion. This bifurcation is also called *secondary Hopf* or *torus* bifurcation. If d_{NS} is negative [positive], the bifurcation is supercritical [subcritical] and the invariant curve (torus) is stable [unstable] (and present for $\alpha > 0[\alpha < 0]$).

Center Manifolds

Generally speaking, the CMT allows one to restrict the dynamics of (1) to a suspended system

$$\dot{w} = G(w, \alpha), \quad G\colon \mathbb{R}^{n_c} \times \mathbb{R}^{n_p} \to \mathbb{R}^{n_c} , \tag{13}$$

on the center manifold H and here n_p is typically 1 or 2 depending on the codimension of the bifurcation. Although the normal forms (13) to which one can restrict the system near nonhyperbolic equilibria and cycles are known, these results are not directly applicable. Thus, efficient numerical algorithms are needed in order to verify the nondegeneracy conditions in the normal forms listed above.

Here, a powerful normalization method due to Iooss and coworkers is reviewed, see [9,14,29,37,55,59]. This method assumes very little a priori information, actually only the type of bifurcation such that the form, i. e., the nonzero coefficients of G, is known. This fits very well in a numerical bifurcation setting where one computes families of solutions and monitors and detects the occurrence of bifurcations with higher codimension during the continuation.

Without loss of generality it is assumed that $x_0 = 0$ at the bifurcation point $p_0 = 0$. Expand $f(x, p)$ in Taylor series

$$\begin{aligned} f(x, p) = {} & Ax + \tfrac{1}{2}B(x, x) + \tfrac{1}{6}C(x, x, x) + J_1 p \\ & + A_1(x, p) + \dots , \end{aligned} \tag{14}$$

parametrize, locally to $(x, p) = (0, 0)$, the parameter-dependent center manifold by

$$x = H(w, \alpha) , \quad H\colon \mathbb{R}^{n_c} \times \mathbb{R}^{n_p} \to \mathbb{R}^{n} , \tag{15}$$

and define a relation $p = V(\alpha)$ between the original and unfolding parameters. The invariance of the center manifold can be exploited by differentiating this parametrization with respect to time to obtain the so-called *homological equation*

$$f(H(w, \alpha), V(\alpha)) = H_w(w, \alpha) G(w, \alpha) . \tag{16}$$

To verify nondegeneracy conditions, only an approximation to the solution of the homological equation is required. To this end, G, H and V are expanded in Taylor series:

$$\begin{aligned} G(w, \alpha) &= \sum_{|\mu|+|\nu|\geq 1} \frac{1}{\mu!\nu!} g_{\mu\nu} w^{\mu} \alpha^{\nu} , \\ H(w, \alpha) &= \sum_{|\mu|+|\nu|\geq 1} \frac{1}{\mu!\nu!} h_{\mu\nu} w^{\mu} \alpha^{\nu} , \\ V(\alpha) &= v_{10}\alpha_1 + v_{01}\alpha_2 + O(\|\alpha\|^2) , \end{aligned} \tag{17}$$

where $g_{\mu\nu}$ are the desired normal form coefficients and μ, ν are multi-indices. For a multi-index μ one has $\mu = (\mu_1, \mu_2, \dots, \mu_n)$ for nonnegative integers μ_i, $\mu! = \mu_1!\mu_2!\dots\mu_n!$, $|\mu| = \mu_1 + \mu_2 + \dots + \mu_n$, $\tilde{\mu} \leq \mu$ if $\tilde{\mu}_i \leq \mu_i$ for all $i = 1, \dots, n$ and $w^{\mu} = w_1^{\mu_1} \dots w_n^{\mu_n}$. When dealing with just the critical coefficients, i. e., $\alpha = 0$, the index ν is omitted. Substitution of this ansatz into (16) gives a formal power series in w and α. As both sides should be equal for all w and α, the coefficients of the corresponding powers should be equal. For each vector $h_{\mu\nu}$, (16) gives linear systems of the form

$$L_{\mu\nu} h_{\mu\nu} = R_{\mu\nu} , \tag{18}$$

where $L_{\mu\nu} = A - \gamma_{\mu\nu} I_n$ ($\gamma_{\mu\nu}$ is a weighted sum of the critical eigenvalues) and $R_{\mu\nu}$ involves known quantities of G and H of order less than or equal to $|\mu| + |\nu|$. This leads to an iterative procedure, where, either system (18) is nonsingular, or the required coefficients $g_{\mu\nu}$ are obtained by imposing solvability, i. e., $R_{\mu\nu}$ lies in the range of $L_{\mu\nu}$ and is therefore orthogonal to the eigenvectors of $L_{\mu\nu}^{\top}$ associated to the zero eigenvalue. In the second case, the solution of (18) is not unique, and one typically selects the $h_{\mu\nu}$ without components in the nullspace of $L_{\mu\nu}$. However, the nonuniqueness of the center manifold does not affect qualitative conclusions. The parameter transformation, i. e., the v_{ν}, is obtained by imposing certain conditions on some normal form coefficients, leading to a solvable system.

One can perform an analogous procedure for the Poincaré map by using a Taylor expansion

$$P(z, p) = Az + \tfrac{1}{2}B(z, z) + \tfrac{1}{6}C(z, z, z) + \dots \tag{19}$$

and the homological equation for maps

$$P(H(w, \alpha), V(\alpha)) = H(G(w, \alpha), \alpha) . \tag{20}$$

The detailed derivation of the formulas for all codim 1 and 2 cases for equilibria and cycles can be found in [9,55,

Numerical Bifurcation Analysis, Table 1

Critical normal form coefficients for generic codim 1 bifurcations of equilibria and fixed points. Here, *A*, *B* and *C* refer to the expansion (14) for equilibria, while for fixed points they refer to (19)

	Eigenvectors	Critical normal form coefficients
LP	$Av = 0$ $A^\top w = 0$	$a_{\mathrm{LP}} = \frac{1}{2} w^\top B(v, v)$
H	$Av = i\omega_0 v$ $A^\top w = -i\omega_0 w$	$c_{\mathrm{H}} = \frac{1}{2}\bar{w}^\top \left[C(v, v, \bar{v}) + 2B(v, h_{11}) + B(\bar{v}, h_{20})\right]$ $h_{11} = -A^{-1}B(v, \bar{v}),\ h_{20} = (2i\omega_0 I_n - A)^{-1}B(v, v)$
LPC	$Av = v$ $A^\top w = w$	$a_{\mathrm{LPC}} = \frac{1}{2} w^\top B(v, v)$
PD	$Av = -v$ $A^\top w = -w$	$b_{\mathrm{PD}} = \frac{1}{6} w^\top \left[C(v, v, v) + 3B(v, h_2)\right]$ $h_2 = (I_{n-1} - A)^{-1}B(v, v)$
NS	$Av = \mathrm{e}^{i\theta_0} v$ $A^\top w = \mathrm{e}^{-i\theta_0} w$ $\mathrm{e}^{ik\theta_0} \neq 1, k = 1, 2, 3, 4$	$c_{\mathrm{NS}} = \frac{1}{2}\bar{w}^\top \left[C(v, v, \bar{v}) + 2B(v, h_{11}) + B(\bar{v}, h_{20})\right]$ $h_{11} = (I_{n-1} - A)^{-1}B(v, \bar{v}),\ h_{20} = (\mathrm{e}^{2i\theta_0} I_{n-1} - A)^{-1}B(v, v)$

56,59,60]. The formulas for the critical normal form coefficients for codim 1 bifurcations are presented in Table 1. Note once more that for limit cycles a numerical more appropriate method exists [57] based on periodic normal forms [47,48].

Continuation and Detection of Bifurcations

Along a solution branch one generically passes through bifurcation points of higher codimension. To detect such an event, a test function φ is defined, where the event corresponds to a regular zero. If at two consecutive points y_{k-1}, y_k along the branch the test function changes sign, i. e., $\varphi(y_k)\varphi(y_{k-1}) < 0$, then the zero can be located more precisely. Usually, a one-dimensional secant method is used to find such a point. Now, if system (1) has a bifurcation at $y_0 = (x_0, p_0)$, then there is generically a curve $y = y(s)$ where the system displays this bifurcation. In order to find this curve, one starts with a known point y_0 and formulates a defining system and then continue that solution in one extra free parameter.

Test Functions for Codimension 1 Bifurcations

An equilibrium may lose stability through a limit point, a Hopf bifurcation or in a branch point. At a limit or branch point bifurcation the Jacobi matrix $A = f_x(x_0, p_0)$ has an algebraically simple eigenvalue $\lambda = 0$ (see Sect. "Branch Switching" for branch points), while at a Hopf point there is a pair of complex conjugate eigenvalues $\lambda = \pm i\omega_0, \omega_0 \neq 0$ and only one such pair.

The simplest way of detecting the passage through a bifurcation during continuation, is to monitor the eigenvalues of the Jacobi matrix. For large systems and stiff problems this is prohibitive as it is numerically expensive and not always accurate. Instead, one can base test functions on determinants.

Test Functions for Limit Point Bifurcations Along an equilibrium curve the product of the eigenvalues changes sign at a limit point. Recall that the determinant of A is the product of its eigenvalues. Therefore, the following test function can be computed

$$\varphi_{\mathrm{LP}} = \det(f_x(x, p)) \tag{21}$$

without computing the eigenvalues explicitly.

For the LP bifurcation the pseudo-arclength or Moore–Penrose continuation methods provide an excellent test function as a by-product of the continuation. Note that while passing through the fold, the last component of the tangent vector ϕ changes sign as the continuation direction in the parameter reverses. The test function is therefore defined as

$$\varphi_{\mathrm{LP}} = \phi_{n+1}\,. \tag{22}$$

Test Functions for Hopf Bifurcations Denote the eigenvalues of A by $\lambda_i(x, p)$, $i = 1 \ldots, n$ and consider the following product

$$\varphi_{\mathrm{H}} = \prod_{i<j} (\lambda_i(x, p) + \lambda_j(x, p))\,.$$

It can be shown that this product has a regular zero at a simple Hopf point [9], but it should be checked that this

zero corresponds to an imaginary pair and not to the neutral saddle case $\lambda_i = -\lambda_j$, $\lambda_i \in \mathbb{R}$.

Also here one can compute this product without explicit computation of the eigenvalues using the bi-alternate product [34,42,45,56]. The bi-alternate product of two $(n \times n)$-matrices A and B, denoted by $A \odot B$, is a $(m \times m)$-matrix C $(m = n(n-1)/2)$ with row index (i, j) and column index (k, l) and elements

$$C_{(i,j)(k,l)} = \frac{1}{2}\left\{ \begin{vmatrix} a_{ik} a_{il} \\ b_{jk} b_{jl} \end{vmatrix} + \begin{vmatrix} b_{ik} b_{il} \\ a_{jk} a_{jl} \end{vmatrix} \right\}$$
$$\text{where} \begin{array}{l} i = 2, 3, \ldots, n, \; j = 1, 2, \ldots i-1 \,, \\ k = 2, 3, \ldots, n, \; l = 1, 2, \ldots k-1 \,. \end{array}$$

Let A be an $n \times n$-matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$, then [73]

- $A \odot A$ has eigenvalues $\lambda_i \lambda_j$,
- $2A \odot I_n$ has eigenvalues $\lambda_i + \lambda_j$.

The test function can now be expressed as

$$\varphi_{\mathrm{H}} = \det(2 f_x(x, p) \odot I_n) \,. \tag{23}$$

For higher dimensional systems, this matrix becomes very large and one should use precondition or subspace methods, see [36].

Test Functions for Codimension 1 Cycle Bifurcations Recall that the nontrivial multipliers $\mu_1, \ldots, \mu_{n-1}$ determine the stability of the cycle and can be efficiently computed as the nontrivial multipliers of the monodromy matrix M, see Sect. "Discretization of BVPs". Now the following two sets of test functions can be used to detect LPC, PD and NS bifurcations

$$\varphi_{\mathrm{LPC}} = \prod_{i=1}^{n-1} (\mu_i - 1) \,, \qquad \varphi_{\mathrm{LPC}} = \phi_p \,,$$
$$\varphi_{\mathrm{PD}} = \prod_{i=1}^{n-1} (\mu_i + 1) \,, \qquad \varphi_{\mathrm{PD}} = \det(M + I_n) \,,$$
$$\varphi_{\mathrm{NS}} = \prod_{1=i<j}^{n-1} (\mu_i \mu_j - 1) \,, \qquad \varphi_{\mathrm{NS}} = \det(M \odot M - I_{n(n-1)/2}) \,.$$

where ϕ_p denotes the parameter component of the tangent vector similar to (22). It should also be checked that a zero of φ_{NS} corresponds to nonreal multipliers $\mathrm{e}^{i\theta_0}$, similar to the test function to detect the Hopf bifurcation.

There are alternatives for these test functions. One can define bordered systems using the monodromy matrix [9,42] or a BVP formulation [23].

Defining Systems for Codimension 1 Bifurcations of Equilibria

To compute curves of codim 1 equilibria bifurcations, first a defining system of the form (2) needs to be formulated to define the bifurcation curve, and then a second parameter for the continuation must be freed, so that now $p \in \mathbb{R}^2$. This is done by adding to the equilibrium equation $f(x, p) = 0$ appropriate equations that characterize the bifurcation.

Defining systems come in two flavors, *fully* (also standard) and *minimally* augmented systems. The first computes all relevant eigenspaces, while the latter exploits the rank deficiency of the Jacobi matrix and adds only a few strategic equations to regularize the continuation problem. The evaluation of such equations requires the eigenspaces, but these can be computed separately. As the names suggest, the difference is in the dimension of the defining system leading to differently sized problems. In particular, the advantage of minimally augmented systems is that of solving several smaller linear problems, instead of a big one, which is known to be better in terms of both accuracy and computational time. For small phase dimension n there is little difference in computational effort. Both minimally and fully extended defining systems for both limit point and Hopf bifurcations are presented. The regularity of these systems is also known, e. g., see [42].

Defining Systems for Limit Point Bifurcations The first defining system is minimally extended adding the test function (21)

$$\begin{cases} f(x, p) = 0 \,, \\ \det(f_x(x, p)) = 0 \,. \end{cases} \tag{24}$$

This system consists of $n + 1$ equations in $n + 2$ unknowns (x, p). One problem is that the computation of the determinant can lose accuracy for large systems. This can be avoided in two ways, by augmenting the system with the eigenspaces or using a bordering technique.

Fully extended systems include the eigenvectors and for a LP bifurcation this leads to

$$\begin{cases} f(x, p) = 0 \,, \\ f_x(x, p) v = 0 \,, \\ v_0^{\top} v - 1 = 0 \,, \end{cases} \tag{25}$$

where v_0 is a vector not orthogonal to $\mathcal{N}(f_x(x, p))$. This system consists of $2n + 1$ equations in $2n + 2$ unknowns (x, p, v).

The bordering technique uses the bordering lemma [41]. Let $A \in \mathbb{R}^{n \times n}$ be a singular matrix and let

$B, C \in \mathbb{R}^{n\times m}$ such that the system

$$\begin{pmatrix} A & B \\ C^\top & 0_m \end{pmatrix}\begin{pmatrix} V \\ g \end{pmatrix} = \begin{pmatrix} 0_{n\times m} \\ I_m \end{pmatrix} \tag{26}$$

is nonsingular ($V \in \mathbb{R}^{n\times m}, g \in \mathbb{R}^{m\times m}$). Typically B and C are associated to the eigenspaces of $A^\top$ and A corresponding to the zero eigenvalue, respectively or, during continuation, approximated by their values computed at the previous point along the branch. It follows from the bordering lemma that A has rank deficiency m if and only if g has rank deficiency m.

With $A = f_x(x, p)$ and $m = 1$, one has $g = 0$ if and only if $\det(f_x(x, p)) = 0$. A modified and minimally extended system for limit points is thus given by

$$\begin{cases} f(x, p) = 0\,, \\ g(x, p) = 0\,, \end{cases}$$

where g is defined by (26) with $A^\top B = AC = 0$ at a previously computed point. During the continuation the derivatives of g w.r.t. to x and p are needed. They can either be approximated by finite differences, or explicitly (and efficiently) obtained from the second-derivatives of the vector field f, see [42].

Defining Systems for Hopf Bifurcations Defining systems for Hopf bifurcations are formulated analogously to the LP case. Adding the test function (23) creates a minimally extended system

$$\begin{cases} f(x, p) = 0\,, \\ \det(2f_x(x, p) \odot I_n) = 0\,, \end{cases} \tag{27}$$

while the fully extended system is given by

$$\begin{cases} f(x, p) = 0\,, \\ f_x(x, p)v_1 + \omega v_2 = 0\,, \\ f_x(x, p)v_2 - \omega v_1 = 0\,, \\ w_1^\top v_1 + w_2^\top v_2 - 1 = 0\,, \\ w_1^\top v_2 - w_2^\top v_1 = 0\,, \end{cases} \tag{28}$$

where $w = w_1 + iw_2$ is not orthogonal to the eigenvector $v = v_1 + iv_2$ corresponding to the eigenvalue $i\omega$. The vector $w = v^{k-1}$ computed at the previous point is a suitable choice during continuation. System (28) is expressed using real variables and has $3n + 2$ equations for $3n + 3$ unknowns (x, p, v_1, v_2, ω).

A reduced defining system can be obtained from (28) by noting that the matrix $f_x(x, p)^2 + \kappa I_n$ has rank deficiency two at a Hopf bifurcation point with $\kappa = \omega^2$ [67]. An alternative to (28) is now formulated as

$$\begin{cases} f(x, p) = 0\,, \\ \left[f_x(x, p)^2 + \kappa I_n\right] v = 0\,, \\ v^\top v - 1 = 0\,, \\ w^\top v = 0\,, \end{cases} \tag{29}$$

where w is not orthogonal to the two-dimensional real eigenspace of the eigenvalues $\pm i\omega$. It has $2n + 2$ equations for $2n + 3$ unknowns (x, p, v, κ). However, w needs to be updated during continuation, e. g., as the solution of $\left([f_x(x, p)^2 + \kappa I_n]^\top w,\ v^\top w\right) = (0,\ 0)$ computed at the previous continuation point.

A further reduction is obtained exploiting the rank deficiency. Consider the system

$$\begin{pmatrix} f_x(x, p)^2 + \kappa I_n & B \\ C^\top & 0_2 \end{pmatrix}\begin{pmatrix} V \\ g \end{pmatrix} = \begin{pmatrix} 0_{n\times 2} \\ I_2 \end{pmatrix}$$

and it follows from the bordering lemma that g vanishes at Hopf points and any two components of g, e. g., g_{11} and g_{22}, see [42], can be taken to augment Eq. (2) to obtain the following minimally augmented system

$$\begin{cases} f(x, p) = 0\,, \\ g_{11} = 0\,, \\ g_{22} = 0\,, \end{cases} \tag{30}$$

which has $n + 2$ equations for $n + 3$ unknowns (x, p, κ).

Defining Systems for Codimension 1 Bifurcations of Limit Cycles

In principle, to study bifurcations of limit cycles one can compute numerically the Poincaré map and study bifurcations of fixed points. If system (1) is not stiff, then the Poincaré map and its derivatives may be obtained with satisfactory accuracy. In many cases, however, continuation using BVP formulations is much more efficient.

Suppose a cycle x bifurcates at $p = p_0$, then the BVP (4) defining the limit cycle must be augmented with suitable extra functions. As for codim 1 branches of equilibria, one can define either fully extended systems by including the relevant eigenfunctions in the computation [9], or minimally extended systems using bordered BVPs [23,39]. The regularity of these defining systems is also discussed in these references. Since the discretization of the cycle leads to large APs, here the minimally extended approach can lead to faster results even though some more algebra is involved, see the comparison in [57]. Below only the equations are presented, which are added to the defining system (4) for the continuation of limit cycles.

Fully Extended Systems The following equations can be used to augment (4) and continue codim 1 bifurcations of limit cycles. The eigenfunctions v need to be discretized in a similar way to that in Sect. "Discretization of BVPs". The previous cycle x^{k-1} and eigenfunction v^{k-1} are assumed to be known.

LPC: For the *limit point of cycles* bifurcation, the BVP (4) is augmented with the equations

$$\begin{cases} \dot{v}(\tau) - Tf_x(x(\tau), p)v(\tau) - \sigma f(x(\tau), p) = 0\,, \\ v(1) - v(0) = 0\,, \\ \int_0^1 v^\top(\tau)\dot{x}^{k-1}(\tau)\mathrm{d}\tau = 0\,, \\ \int_0^1 v^\top(\tau)v(\tau)^{k-1}\mathrm{d}\tau + \sigma\sigma^{k-1} = 1\,, \end{cases} \tag{31}$$

for the variables (x, p, T, v, σ). Note that

$$\begin{cases} \dot{v}(\tau) - Tf_x(x(\tau), p)v(\tau) - Tf_p(x(\tau), p)q \\ \qquad -\sigma f(x(\tau), p) = 0\,, \\ v(1) - v(0) = 0\,, \\ \int_0^1 v^\top(\tau)\dot{x}^{k-1}(\tau)\mathrm{d}\tau = 0\,, \\ \int_0^1 v^\top(\tau)v(\tau)^{k-1}\mathrm{d}\tau + q^{k-1}q + \sigma^{k-1}\sigma = 1\,, \end{cases}$$

defines the tangent vector $\phi = (v, q, \sigma)$ to the solution branch, so that (31) simply imposes $q = 0$, i. e., the limit point. Together with (4), they compose a BVP with $2n$ ODEs, $2n$ boundary conditions, and 2 integral conditions, i. e., $n_p = 2n + 2 - 2n + 1 = 3$, namely T and two free parameters. Similar dimensional considerations hold for the PD and NS cases below.

PD: For the *period-doubling* bifurcation, the extra equations augmenting (4) are

$$\begin{cases} \dot{v}(\tau) - Tf_x(x(\tau), p)v(\tau) = 0\,, \\ v(1) + v(0) = 0\,, \\ \int_0^1 v^\top(\tau)v^{k-1}(\tau)\mathrm{d}\tau = 1\,, \end{cases} \tag{32}$$

for the variables (x, p, T, v). Here v is the eigenfunction of the linearized ODE associated with the multiplier $\mu = -1$. In fact, the second equation in (32) imposes $v(1) = Mv(0) = -v(0)$, where M is the monodromy matrix, while the third equation scales the eigenfunction against the previous continuation point.

NS: For the *Neimark–Sacker* bifurcation, the BVP (4) is augmented with the equations

$$\begin{cases} \dot{v}(\tau) - Tf_x(x(\tau), p)v(\tau) = 0\,, \\ v(1) - \mathrm{e}^{i\theta}v(0) = 0\,, \\ \int_0^1 \bar{v}^\top(\tau)v^{k-1}(\tau)\mathrm{d}\tau = 1\,, \end{cases} \tag{33}$$

for the variables (x, p, T, v, θ) with $v \in C^1([0, 1], \mathbb{C}^n)$. Here v is the eigenfunction of the linearized ODE associated with the multiplier $\mu = \mathrm{e}^{i\theta}$. Of course, the real formulation should be used in practice.

Minimally Extended Systems For limit cycle continuation the discretization of the fully extended BVP (4) with (31), (32) or (33) may lead to large APs to be solved. In [39] a minimally extended formulation is proposed to augmenting (4) with a function g with only a few components. The corresponding function g is defined using bordered systems.

LPC: For this bifurcation, one uses suitable bordering functions v_1, w_1 and vectors v_2, w_2, w_3 such that the following system linear in (v, σ, g) is regular

$$\begin{cases} \dot{v}(\tau) - Tf_x(x(\tau), p)v - f(x(\tau), p)\sigma + w_1 g = 0\,, \\ v(1) - v(0) + w_2 g = 0\,, \\ \int_0^1 f(x(\tau), p)^\top v(\tau)\mathrm{d}\tau + w_3 g = 0\,, \\ \int_0^1 v_1^\top v(\tau)\mathrm{d}\tau + v_2\sigma = 1\,. \end{cases} \tag{34}$$

The function $g = g(x, T, p)$ vanishes at a LPC point. The bordering functions v_1, w_1 and vectors v_2, w_2, w_3 can be updated to keep (34) nonsingular, in particular, $v_1 = v^{k-1}$ and $v_2 = \sigma^{k-1}$ from the previously computed point are used. It is convenient to introduce the Dirac operator $\delta_i f = f(i)$ and the integral operator $\mathrm{Int}_{v(\cdot)}f = \int_0^1 v(\tau)^\top f(\tau)\mathrm{d}\tau$ and to rewrite (34) in operator form

$$\begin{pmatrix} D - Tf_x(x(\cdot), p) & -f(x(\cdot), p) & w_1 \\ \delta_0 - \delta_1 & 0 & w_2 \\ \mathrm{Int}_{f(x(\cdot),p)} & 0 & w_3 \\ \mathrm{Int}_{v_1(\cdot)} & v_2 & 0 \end{pmatrix} \begin{pmatrix} v \\ \sigma \\ g \end{pmatrix} = \begin{pmatrix} c0 \\ 0 \\ 0 \\ 1 \end{pmatrix}. \tag{35}$$

PD: The same notation as for the minimally extended LPC defining system is used and suitable bordering functions v_1, w_1 and vector w_2 are chosen such that the following system is regular

$$\begin{pmatrix} D - Tf_x(x(\cdot), p) & w_1 \\ \delta_0 + \delta_1 & w_2 \\ \mathrm{Int}_{v_1(\cdot)} & 0 \end{pmatrix} \begin{pmatrix} v \\ g \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \tag{36}$$

At a PD bifurcation $g(x, T, p)$ defined by (36) vanishes.

NS: Let $\hat{\kappa} = \cos(\theta)$ denote the real part of the nonhyperbolic multiplier and choose bordering functions v_1, v_2, w_{11}, w_{12} and vectors w_{21}, w_{22} such that the following system is nonsingular and defines the four components of g

$$\begin{pmatrix} D - Tf_x(x(\cdot), p) & w_{11} & w_{12} \\ \delta_2 - 2\hat{\kappa}\delta_1 + \delta_0 & w_{21} & w_{22} \\ \mathrm{Int}_{v_1(\cdot)} & 0 & 0 \\ \mathrm{Int}_{v_2(\cdot)} & 0 & 0 \end{pmatrix} \begin{pmatrix} v \\ g \end{pmatrix} = \begin{pmatrix} 0_{n\times 2} \\ I_2 \end{pmatrix}. \tag{37}$$

At a NS bifurcation the four components of $g(x, T, p)$ defined by (37) vanish, and, similar to the Hopf bifurcation, the BVP (4) can be augmented with any two components of g.

Test Functions for Codimension 2 Bifurcations

During the continuation of codim 1 branches, one meets generically codim 2 bifurcations. Some of which arise through extra instabilities in the linear terms, while other codim 2 bifurcations are defined through degeneracies in the normal form coefficients. For equilibria, codim 2 bifurcations of the first type are the Bogdanov–Takens (BT, two zero eigenvalues with only one associated eigenvector), the zero-Hopf (ZH, also called Gavrilov-Guckenheimer, a simple zero eigenvalue and a simple imaginary pair), and the double Hopf (HH, two distinct imaginary pairs), while higher order degeneracies lead to cusp (CP, $a_{\mathrm{LP}} = 0$ in the normal form (11)) or generalized Hopf (GH, $d_{\mathrm{H}} = 0$ in the normal form (12), also called Bautin or degenerate Hopf). For cycles, there are strong resonances (R1–R4), fold-flip (LPPD), fold-Neimark–Sacker (LPNS), flip-Neimark–Sacker (PDNS), and double Neimark–Sacker (NSNS) among those involving linear terms, while higher order degeneracies lead to cusp (CP), degenerate flip (GPD), and Chenciner (CH) bifurcations. Naturally the normal form coefficients, see [9,56], are a suitable choice for the corresponding test functions. In Tables 2 and 3, test functions are given which are defined along the corresponding codim 1 branches of equilibrium and limit cycle bifurcations, respectively. The functions refer to the corresponding defining system and to Table 1. Upon detecting and locating a zero of a test function it may be necessary to check that a bifurcation is really involved, similar to the Hopf case where neutral saddles are excluded. For details about the dynamics and the bifurcation diagrams at codim 2 points, see [2,44,56].

Numerical Bifurcation Analysis, Table 2
Test functions along limit point and Hopf bifurcation curves. The matrix A_c for the test function of the double Hopf bifurcation can be obtained as the orthogonal complement in $\mathbb{R}^n$ of the Jacobi matrix A w.r.t. the two-dimensional eigenspace associated with the computed branch of Hopf bifurcations

	Label	LP	H
cusp	CP	a_{LP}	
generalized Hopf	GH		d_{H}
Bogdanov–Takens	BT	$w_{\mathrm{LP}}^{\top} v_{\mathrm{LP}}$	κ
zero-Hopf	ZH	φ_{H}	φ_{LP}
double Hopf	HH		$\det(2A_c \odot I_{n-2})$

Branch Switching

This section considers points in the continuation space from which several solution branches of interest, with the same codimension, emanate. At these points suitable "branch switching" procedures are required to switch from one solution branch to another. First, the transversal intersection of two solution branches of the same continuation problem is considered, which occurs at so-called Branch Points (BP) (also called "singular" or "transcritical" bifurcation points). Branch points are nongeneric, in the sense that arbitrarily small perturbations of F in (2) turn the intersection into two separated branches, which come close to the "ghost" of the (disappeared) intersec-

Numerical Bifurcation Analysis, Table 3
Test functions along LPC, PD and NS bifurcation curves. The matrix M_c along the Neimark–Sacker bifurcation curve is defined similarly as A_c in Table 2 along the Hopf bifurcation as the orthogonal complement of the monodromy matrix M w.r.t. the two-dimensional eigenspace associated with the computed branch of Neimark–Sacker bifurcations

	Label	LPC	PD	NS
cusp	CP	a_{LPC}		
degenerate flip	GPD		b_{PD}	
Chenciner	CH			d_{NS}
resonance 1:1	R1	$w_{\mathrm{LP}}^{\top} v_{\mathrm{LP}}$		$\hat{\kappa} - 1$
resonance 1:2	R2		$w_{\mathrm{PD}}^{\top} v_{\mathrm{PD}}$	$\hat{\kappa} + 1$
resonance 1:3	R3			$\hat{\kappa} + \frac{1}{2}$
resonance 1:4	R4			$\hat{\kappa}$
fold-flip	LPPD	φ_{PD}	φ_{LP}	
fold-Neimark–Sacker	LPNS	φ_{NS}		φ_{LP}
flip-Neimark–Sacker	PDNS		φ_{NS}	φ_{PD}
double Neimark–Sacker	NSNS			$\det(M_c \odot M_c - I_{(n-2)(n-3)/2})$

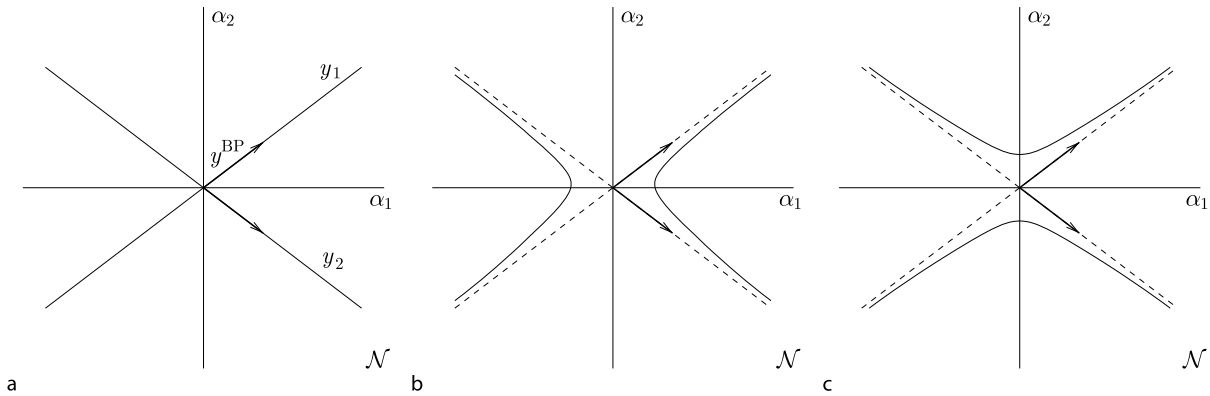

Numerical Bifurcation Analysis, Figure 4

(a) Projection of two solution branches of (2), intersecting at y^{BP}, on the null-space of $F_y(y^{\mathrm{BP}})$ ($\mathcal{N}$), close to y^{BP} (planar representation in coordinates (α_1, α_2) with respect to a given basis). The two solution branches are approximated by the *straight lines* in $\mathcal{N}$ spanned by their tangent vectors at y^{BP} (*thick vectors*). (b) and (c) Projection on $\mathcal{N}$, close to y^{BP}, of the two solution branches of the perturbed problem (see (43) for $b > 0$ and $b < 0$, respectively)

tion but then fold (LP) and leave as if they follow the other branch (see Fig. 4).

BPs, however, are very common in applications due to particular symmetries of the continuation problem, like reflections in state space, conserved quantities or the presence of trivial solutions. This is why BP detection and continuation recently received attention [16,25,28]. Then, the switch from a codim 0 solution branch to that of a different continuation problem at codim 1 bifurcations is examined. In particular, the equilibrium-to-cycle switch at a Hopf bifurcation and the period-1-to-period-2 cycle switch at a flip bifurcation are discussed. Finally, various switches between codim 1 solution branches of different continuation problems at codim 2 bifurcations are addressed.

Branch Switching at Simple Branch Points

Simple BPs are points $y^{\mathrm{BP}} = (u^{\mathrm{BP}}, p^{\mathrm{BP}})$, encountered along a solution branch of (2), at which the nullspace $\mathcal{N}(F_y(y))$ of $F_y(y)$ is two-dimensional, i. e., the nullspace is spanned by two independent vectors $\phi_1, \phi_2 \in Y$, with $\phi_i^\top \phi_i = 1,\ i = 1, 2$. Generically, two solution branches of (2) pass through y^{BP}, with transversal tangent vectors given by suitable combinations of ϕ_1 and ϕ_2. In the following, only the case of the AP (2), i. e., $Y = \mathbb{R}^{n+1}$ ($u \in \mathbb{R}^n$ and $p \in \mathbb{R}$) and $F\colon \mathbb{R}^{n+1} \to \mathbb{R}^n$ is considered. Similar considerations hold for the BVP (3) (see [16] for details), though, loosely speaking, results for APs can be applied to BVPs after time discretization.

A BP is not a regular point, since rank $(F_y(y^{\mathrm{BP}})) = n - 1$. Distinguishing between state and parameters, there are two possibilities

$$\begin{cases} \text{(i)}\ \dim\mathcal{N}(F_u(y^{\mathrm{BP}})) = 1, F_p(y^{\mathrm{BP}}) \in \mathcal{R}(F_u(y^{\mathrm{BP}})) \\ \qquad \Longrightarrow \phi_1 = (v_1, 0), \phi_2 = (v_2, q_2), \\ \text{(ii)}\ \dim\mathcal{N}(F_u(y^{\mathrm{BP}})) = 2, F_p(y^{\mathrm{BP}}) \notin \mathcal{R}(F_u(y^{\mathrm{BP}})) \\ \qquad \Longrightarrow \phi_1 = (v_1, 0), \phi_2 = (v_2, 0)\,, \end{cases} \tag{38}$$

for suitably chosen v_1, v_2, q_2. In particular, in the first case, v_1 spans the nullspace of $F_u(y^{\mathrm{BP}})$ and ϕ_2 is determined by solving $\left(F_u(y^{\mathrm{BP}})v_2 + F_p(y^{\mathrm{BP}})q_2, v_1^\top v_2, v_2^\top v_2 + q_2^2\right) = (0, 0, 1)$.

BPs can be detected by means of the following test function

$$\varphi_{\mathrm{BP}} = \det\left(\begin{bmatrix} F_y(y) \\ \phi^\top \end{bmatrix}\right)\,, \tag{39}$$

where ϕ is the tangent vector to the solution branch during continuation, which indeed vanishes when (2) admits a second independent tangent vector. Note from (38) that test function (21) also vanishes at BPs, so that test function (22) is more appropriate for LPs.

The vectors tangent to the two solution branches intersecting at y^{BP} can be computed as follows. Parametrize one of the two solution branches by a scalar coordinate s, e. g., the arclength, so that $y(s)$ and $y_s(s)$ denote the branch and its tangent vector locally to $y(0) = y^{\mathrm{BP}}$. Then, $F(y(s))$ is identically equal to zero, so taking twice

the derivative w.r.t. s one obtains $F_{yy}(y(s))[y_s(s), y_s(s)] + F_y(y(s))y_{ss}(s) = 0$, which at y^{BP} reads

$$F_{yy}(y^{BP})[y_s(0), y_s(0)] + F_y(y^{BP})y_{ss}(0) = 0\,, \tag{40}$$

with $y_s(0) = \alpha_1\phi_1 + \alpha_2\phi_2$. Let $\psi \in \mathbb{R}^n$ span the nullspace of $F_y(y^{BP})^\top$ with $\psi^\top\psi = 1$. Since the range of $F_y(y^{BP})$ is orthogonal to the nullspace of $F_y(y^{BP})^\top$, one can eliminate $y_{ss}(0)$ in (40) by left-multiplying both sides by $\psi^\top$, thus obtaining

$$\psi^\top F_{yy}(y^{BP})[\alpha_1\phi_1 + \alpha_2\phi_2, \alpha_1\phi_1 + \alpha_2\phi_2] = 0\,. \tag{41}$$

Equation (41) is called the *algebraic branching equation* [50] and is often written as

$$c_{11}\alpha_1^2 + 2c_{12}\alpha_1\alpha_2 + c_{22}\alpha_2^2 = 0\,, \tag{42}$$

with $c_{ij} = \psi^\top F_{yy}(y^{BP})[\phi_i, \phi_j]$, $i, j = 1, 2$. At BP detection, the discriminant $c_{11}c_{22} - c_{12}^2$ is generically negative (otherwise the BP would be an isolated solution point of (2)), so that two distinct pairs (α_1, α_2) and $(\tilde{\alpha}_1, \tilde{\alpha}_2)$, uniquely defined up to scaling, solve (42) and give the directions of the two emanating branches.

Once the two directions are known, one can easily perform branch switching by an initial prediction from y^{BP} along the desired direction. This, however, requires the second-order derivatives of F w.r.t. all continuation variables. Though good approximations can often be achieved by finite differences, an alternative and computationally cheap prediction can be taken in the nullspace of $F_y(y^{BP})$ along the direction orthogonal to $y_s(0)$. The vector $y_s(s)$ is in fact known at each point during the continuation of the solution branch up to BP detection, so that the cheap prediction for the other branch spans the (one-dimensional) nullspace of

$$\begin{bmatrix} F_y(y^{BP}) \\ y_s(0)^\top \end{bmatrix}.$$

Branch Point Continuation

Generic Problems Several defining systems have been proposed for BP continuation, see [16,25,28,63,64,65]. Among fully extended formulations, the most compact one characterizes BPs as points at which the range of $F_y(y)$ has rank defect 1, i. e., the nullspace of $F_y(y)^\top$ is one-dimensional. BP continuation is therefore defined by

$$\begin{cases} F(u, p) = 0\,, \\ F_u(u, p)^\top\psi = 0\,, \\ F_p(u, p)^\top\psi = 0\,, \\ \psi^\top\psi - 1 = 0\,. \end{cases}$$

Counting equations, $2n + 2$, and variables $u, \psi \in \mathbb{R}^n$, $p \in \mathbb{R}$, i. e., $2n + 1$ scalar variables, it follows that two extra parameters generically need to be freed. In other words, BPs are codim 2 bifurcations, which are not expected along generic solution branches of (2).

Non-generic Problems BP continuation can be performed in a single extra free parameter for nongeneric problems characterized by symmetries that persist for all parameter values. In such cases, the continuation problem (2) is perturbed into

$$F(y) + bu_b = 0\,, \tag{43}$$

where $b \in \mathbb{R}$ and $u_b \in \mathbb{R}^n$ are new variables of the defining system. The idea is that u_b "breaks the symmetry", in the sense that problem (43) has no BP for small $b \neq 0$, and BP continuation can be performed in two extra free parameters, one of which, b, remains zero during the continuation. The choice of u_b is not trivial. Geometrically, u_b must be such that small values of b perturb Fig. 4a into Fig. 4b, say for $b > 0$, and into Fig. 4c for $b < 0$. It turns out (see, e. g., [16]) that $u_b = \psi$ is a good choice, i. e., perturbations not in the range of $F_y(y^{BP})$ break the symmetry, since close to the BP, they must be balanced by the nonlinear terms of the expansion of F in (43), and this implies significant deviations of the perturbed solution branch y from the unperturbed $y(s)$.

BP continuation for nongeneric problems is therefore defined by

$$\begin{cases} F(x, p) + b\psi = 0\,, \\ F_x(x, p)^\top\psi = 0\,, \\ F_p(x, p)^\top\psi = 0\,, \\ \psi^\top\psi - 1 = 0\,. \end{cases} \tag{44}$$

This defining system is also useful for accurately computing BPs. In fact, the basin of convergence of the Newton iterations in (8) or (9) shrinks at BPs (recall $F_y(y)$ does not have full rank at BPs), while system (44), in the $2n + 2$ variables (u, p, b, ψ), has a unique solution $(u^{BP}, p^{BP}, 0, \psi)$ close to the BP. Thus, when the BP test function (39) changes sign along a solution branch of (2), Newton corrections can be applied to (44), starting from the best possible prediction, i. e., with $b = 0$ and ψ as the eigenvector of $F_u(u, p)^\top$ associated with the real eigenvalue closest to zero.

Minimally Extended Formulation A minimally extended defining system for BP continuation requires two scalar conditions, $g_1(u, p) = 0$, $g_2(u, p) = 0$, to be added to the unperturbed or perturbed problem (2) or (43)

for generic and nongeneric problems, respectively. These functions g_1 and g_2 are defined in [28] by solving

$$\begin{cases} F_y(y)\phi_1 + g_1\psi^{k-1} = 0\,, \\ F_y(y)\phi_2 + g_2\psi^{k-1} = 0\,, \\ (\phi_1^{k-1})^\top\phi_1 - 1 = 0\,, \\ (\phi_2^{k-1})^\top\phi_2 - 1 = 0\,, \\ (\phi_1^{k-1})^\top\phi_2 = 0\,, \\ (\phi_2^{k-1})^\top\phi_1 = 0\,, \end{cases}$$

in the unknowns ϕ_1, ϕ_2, g_1, g_2, while ψ is updated by solving

$$\begin{cases} F_y(y)^\top\psi + g_1\phi_1 + g_2\phi_2 = 0\,, \\ \psi^\top\psi - 1 = 0\,, \end{cases}$$

in the unknowns ψ, g_1, g_2, after each Newton convergence.

Branch Switching at Hopf Points

At a Hopf bifurcation point $y^{\mathrm{H}} = (x^{\mathrm{H}}, p^{\mathrm{H}})$, one typically wants to start the continuation of the emanating branch of limit cycles. For this, one might think of using the branch switching procedure described above to switch from a constant to a periodic solution branch of the limit cycle BVP (4). Unfortunately, y^{H} is not a simple BP for problem (4), since the period T is undetermined along the constant solution branch, so that, formally, an infinite number of branches emanate from y^{H}. Thus, a prediction in the proper direction, i. e., along the vector $\phi = (v, q)$ tangent to the periodic solution branch, is required.

Let $y(s)$ represent the periodic solution branch, with $y(0) = y^{\mathrm{H}}$. Then, x and v are period-1 vector-valued functions in $C^1([0,\,1], \mathbb{R}^n)$, $p \in \mathbb{R}$ and T are the free parameters, and $q = (p_s, T_s)$. The Hopf bifurcation theorem [56] ensures that $p_s = T_s = 0$ and that v is the unit-length solution of the linearized, time-independent equation $\dot{v} = T(0)f_x(x^{\mathrm{H}}, p^{\mathrm{H}})v$, i. e., $v(\tau) = \sin(2\pi\tau)w_r + \cos(2\pi\tau)w_i$, where $w = w_r + iw_i$ ($w_r^\top w_r + w_i^\top w_i = 1, w_r^\top w_i = 0$) is the complex eigenvector of $f_x(x^{\mathrm{H}}, p^{\mathrm{H}})$ associated to the eigenvalue $i\omega$, $\omega = 2\pi/T(0)$.

The periodic solution branch of the limit cycle BVP (4) can therefore be followed, provided the phase condition (see the last equation in (4)) is replaced by $\int_0^1 x^\top \dot{v}\,\mathrm{d}\tau = 0$ at the first Newton correction. Otherwise, x would be undetermined among time-shifted solutions.

Branch Switching at Flip Points

At a flip bifurcation point $y^{\mathrm{PD}} = (x^{\mathrm{PD}}, p^{\mathrm{PD}})$, where $x^{\mathrm{PD}} \in C^1([0,\,1], \mathbb{R}^n)$, $x^{\mathrm{PD}}(1) = x^{\mathrm{PD}}(0)$, one typically wants to start the continuation of the emanating branch of "period-2" limit cycles, i. e., those which close to y^{PD} have approximately the double of the period of the bifurcating cycle. For this, branch switching at simple BPs can be used. In fact, two solution branches of the limit cycle BVP (4) transversely intersect at y^{PD} if one considers T as the doubled period: the branch of interest and the branch along which the corresponding period-1 cycle is traced twice. In other words, one can see the period-doubling bifurcation in the period-1 branch as the "period-halving" bifurcation in the period-2 branch.

Alternatively, the vector $\phi = (v, q)$ tangent to the period-2 branch at y^{PD} is given by the flip theorem [56] and does not need to be computed by solving the algebraic branching Eq. (41). In particular, the initial solution of the period-2 BVP is $x(t) = x^{\mathrm{PD}}(2t)$, $p = p^{\mathrm{PD}}$, $T = 2T^{\mathrm{PD}}$, while $q = (p_s, T_s) = 0$ and

$$v(t) = \begin{cases} w(t)\,, & 0 \le t < 1\,, \\ -w(t-1)\,, & 1 \le t < 2\,, \end{cases}$$

where $w(t)$ is the unit-length eigenfunction of the linearized (time-dependent) ODE associated with the multiplier -1, i. e.,

$$\begin{cases} \dot{w} - T^{\mathrm{PD}} f_x(x^{\mathrm{PD}}, p^{\mathrm{PD}})w = 0\,, \\ w(1) + w(0) = 0\,, \\ \int_0^1 w(\tau)^\top w(\tau)\,\mathrm{d}\tau = 1\,. \end{cases}$$

Branch Switching at Codimension 2 Equilibria

Assume that $y_2 = (x_2, p_2)$, $x_2 \in \mathbb{R}^n$, $p_2 \in \mathbb{R}^2$, identifies a codim 2 equilibrium bifurcation point of system (1), either cusp (CP), Generalized Hopf (GH), Bogdanov–Takens, (BT), zero-Hopf (ZH), or double Hopf (HH). Then, according to the analysis of the corresponding normal form (13) with two parameters and critical dimension $n_c = 1, 2, 3, 4$ at CP, BT and GH, ZH, HH points, respectively, several curves of codim 1 bifurcations of equilibria and limit cycles emanate from p_2 in the parameter plane [56]. Here, the problem of how the continuation of such curves can be started from y_2 is discussed, by restricting the attention to equilibria bifurcations (an example of a cycle bifurcation is given at the end, and see [9,61]).

In general, the normal form analysis also provides an approximation of each emanating codim 1 bifurcation, in the form of a parameterized expansion

$$w = \sum_{\mu \ge 1} \frac{1}{\mu!} w_\mu \varepsilon^\mu\,, \quad \alpha = \sum_{\nu \ge 1} \frac{1}{\nu!} \alpha_\nu \varepsilon^\nu\,, \tag{45}$$

for small $\varepsilon > 0$ and up to some finite order. Then, the (parameter-dependent) center manifold (15) maps such an

approximation back to the solution branch of the original system (1), locally to $y_2 = (H(0,0), V(0))$, and allows one to compute the proper prediction from y_2 along the direction of the desired codim 1 bifurcation. However, as will be concluded in the following, such approximations are not needed to start equilibria bifurcations and are therefore not derived (see [9] for all available details).

Switching at a Cusp Point At a generic cusp point $y^{\mathrm{CP}} = (u^{\mathrm{CP}}, p^{\mathrm{CP}})$, two fold bifurcation curves terminate tangentially. Generically, the cusp point is a regular point for the fold defining systems (24) and (25), where the tangent vector has zero p-components. The cusp geometrically appears once the fold solution branch is projected in the parameter plane. Thus, the continuation of the two fold bifurcations can simply be started as forward and backward fold continuation from y^{CP}. Since the continuation direction is uniquely defined, neither a tangent prediction nor a nonlinear expansion of the desired solution branch are necessary.

Switching at a Generalized Hopf Point At a Hopf point with vanishing Lyapunov coefficient $y^{\mathrm{GH}} = (x^{\mathrm{GH}}, p^{\mathrm{GH}})$, a LPC bifurcation terminates tangentially to a Hopf bifurcation, which turns from super- to subcritical, and vice versa. Thus, y^{GH} is a regular point for the Hopf defining systems (27)–(30).

Switching at a Bogdanov–Takens Point At a generic Bogdanov–Takens point $y^{\mathrm{BT}} = (u^{\mathrm{BT}}, p^{\mathrm{BT}})$, a Hopf and a (saddle) homoclinic bifurcation terminate tangentially to a fold bifurcation, along which a real nonzero eigenvalue of the Jacobi matrix $f_x(x,p)$ changes sign at y^{BT}. Generically, y^{BT} is a regular point for the fold defining systems (24) and (25) and for the Hopf defining systems (27), (29), and (30). However, y^{BT} is a simple BP for the Hopf defining system (28). In fact, the fold and Hopf branches are both solution branches of the continuation problem (28), where $\omega = 0$ and $v_1 = v_2$, with $v_1^\top v_1 = v_2^\top v_2 = 1/2$, along the fold branch. The branch switch procedures described in this section readily apply in this case.

Switching at Zero-Hopf and Double Hopf Points Generically, zero-Hopf and double Hopf points are regular points for codim 1 equilibria bifurcations (fold and Hopf), so that the proper initial prediction is uniquely defined. For limit cycle bifurcations and connecting orbits, nonlinear expansions of the fold and Hopf branches are needed to derive initial predictions for the emanating branches. For cycles the switching procedure can be set up using the center manifold [61]. However, initial predictions for homoclinic and heteroclinic bifurcations for both zero-Hopf and double Hopf cases are not available in general, but see [35].

The double Hopf bifurcation appears when two different branches of Hopf bifurcations intersect. Several bifurcation curves are rooted at the double Hopf point. In particular, it is known that there are generically two branches, two half-lines in the parameter plane, emanating from this point along which a Neimark–Sacker bifurcation of limit cycles occurs [56]. Here it is discussed how to initialize the continuation of a Neimark–Sacker branch (using (37)) from a double Hopf point after continuation of a Hopf branch. The initialization requires approximations of the cycle x, the period T, the parameters p and the real part of the multiplier $\hat{\kappa}$. These can be obtained by reducing the dynamics of (1) to the center manifold. On the center manifold, the dynamics near a HH bifurcation point is governed by the following normal form

$$\begin{pmatrix} \dot{w}_1 \\ \dot{w}_2 \end{pmatrix} = \begin{pmatrix} (i\omega_1(\alpha) + \alpha_1)w_1 + g_{2100}w_1|w_1|^2 + g_{1011}w_1|w_2|^2 \\ (i\omega_2(\alpha) + \alpha_2)w_2 + g_{1110}w_2|w_1|^2 + g_{0021}w_2|w_2|^2 \end{pmatrix} + O(\|(w_1, w_2)\|^4)\,, \tag{46}$$

where $(w_1, w_2) \in \mathbb{C}^2$. In polar coordinates, $w_1 = \rho_1 e^{i\theta_1}$, $w_2 = \rho_2 e^{i\theta_2}$, the asymptotics from the normal form as in (45) for the nonhyperbolic cycle on one branch are given by

$$(\rho_1, \rho_2, \alpha_1, \alpha_2) = \left(\varepsilon, 0, -\mathbb{R}e(g_{2100})\varepsilon^2, -\Re(g_{1110})\varepsilon^2\right)\,, \tag{47}$$

with $\theta_1 \in [0, 2\pi]$, $\theta_2 = 0$.

Although high-dimensional, the computation of the coefficients and center manifold vectors is relatively straightforward. Introduce $Av_j = i\omega_j v_j$, $A^\top w_j = -i\omega_j w_j$ with $\bar{v}_j^\top v_j = \bar{w}_j^\top v_j = 1$ and let $v = (10), (01)$ and introduce the standard basis vectors $e_{10} = (1,0)$, $e_{01} = (0,1)$. Using the expansion (14), the cubic critical normal form coefficients and the parameter dependence are calculated from

$$\begin{aligned} g_{2100} &= \bar{w}_1^\top [C(v_1, v_1, \bar{v}_1) + B(h_{2000}, \bar{v}_1) \\ &\quad + 2B(h_{1100}, v_1)]/2\,, \\ g_{1011} &= \bar{w}_1^\top [C(v_1, v_2, \bar{v}_2) + B(h_{1010}, \bar{v}_2) \\ &\quad + B(h_{1001}, v_2) + B(h_{0011}, v_1)], \end{aligned}$$

$$
\begin{aligned}
g_{1110} &= \bar{w}_2^\top [C(v_2, v_1, \bar{v}_1) + B(h_{1100}, v_2) + B(h_{1010}, \bar{v}_1) \\
&\quad + B(\bar{h}_{1001}, v_1)] \,, \\
g_{0021} &= \bar{w}_2^\top [C(v_2, v_2, \bar{v}_2) + B(h_{0020}, \bar{v}_2) \\
&\quad + 2B(h_{0011}, v_2)]/2 \,, \\
\Gamma_{j,\nu} &= \bar{p}_j^\top [A_1(v_j, e_\nu) - B(v_j, A^{-1} J_1 e_\nu)]/2 \,,
\end{aligned} \tag{48}
$$

where $j = 1, 2$ and

$$
\begin{aligned}
h_{2000} &= (2i\omega_1 I_n - A)^{-1} B(v_1, v_1) \,, \\
h_{0020} &= (2i\omega_2 I_n - A)^{-1} B(v_2, v_2) \,, \\
h_{1100} &= -A^{-1} B(v_1, \bar{v}_1) \,, \\
h_{0011} &= -A^{-1} B(v_2, \bar{v}_2) \,, \\
h_{1010} &= (i(\omega_1 + \omega_2) I_n - A)^{-1} B(v_1, v_2) \,, \\
h_{1001} &= (i(\omega_1 - \omega_2) I_n - A)^{-1} B(v_1, \bar{v}_2) \,,
\end{aligned} \tag{49}
$$

where h_μ are the vectors in the expansion of the center manifold (17).

Now, to construct a cycle for (46), a mesh for θ_1 is defined and the asymptotics (47) are inserted into the polar coordinates. This cycle is mapped back to the original system by using (17) with (47), (48) and (49). The transformation between free system and unfolding parameters is given by $p(\varepsilon) = \mathcal{R}(\Gamma_1 \Gamma_2)^{-1} (\alpha_1, \alpha_2)^\top$ using (47). Finally, approximating formulas for the period and the real part of the multiplier are given by

$$
\begin{aligned}
T &= \frac{2\pi}{\omega_1 + \mathrm{d}\omega_1 \varepsilon^2} \,, \quad \hat{\kappa} = \cos(T(\omega_2 + \mathrm{d}\omega_2 \varepsilon^2)) \,, \\
(\mathrm{d}\omega_1, \mathrm{d}\omega_2) &= -\Im(\Gamma_1 \Gamma_2)^\top (\Re(\Gamma_1 \Gamma_2)^\top)^{-1} \Re(g_{2100}, g_{1110})^\top \\
&\quad + \Im(g_{2100}, g_{1110}) \,,
\end{aligned}
$$

where $\mathrm{d}\omega_1, \mathrm{d}\omega_2$ indicate the change in rotation in the angles θ_1, θ_2 for $\varepsilon \neq 0$. This construction is done up to second order in ε and leads to an initial approximation for the continuation of a Neimark–Sacker bifurcation curve starting from a double Hopf point. A similar set up can be defined for the other branch.

Connecting Orbits

Connecting orbits, such as homoclinic and heteroclinic orbits, can be continued using a variety of techniques. To fix some terminology: a heteroclinic orbit that connects two equilibrium points x_- and x_+ in the ODE system (1) is an orbit for which

$$
\lim_{t \to -\infty} x(t) = x_- \quad \text{and} \quad \lim_{t \to \infty} x(t) = x_+ \,.
$$

A homoclinic orbit is an orbit connecting an equilibrium point to itself, that is, if $x_+ = x_-$. Similarly there exist heteroclinic connecting orbits between equilibrium points and periodic orbits, and homoclinic and heteroclinic orbits connecting periodic orbits to periodic orbits.

Traditionally, homoclinic orbits to equilibrium points were computed indirectly using numerical shooting or by continuing a periodic solution with a large enough but fixed period, that is close enough to a homoclinic orbit [24]. More modern and robust techniques compute connecting orbits directly using *projection boundary conditions*. Computing connections with and between periodic orbits is subject to current research [26,27,54]. For a more detailed description of the methods described here, see [9].

Formulation as a BVP

A heteroclinic orbit can be expressed as a BVP in the following way:

$$
\begin{aligned}
\dot{x}(t) &= f(x(t), p) \,, \quad \lim_{t \to -\infty} x(t) = x_- \,, \\
\lim_{t \to \infty} x(t) &= x_+ \,, \quad \int_{-\infty}^{\infty} (x(t) - x_0(t))^\top \dot{x}_0(t) \mathrm{d}t = 0 \,,
\end{aligned}
$$

where the integral condition is with respect to a reference solution $x_0(t)$ and fixes the phase, similarly to the phase condition for periodic orbits. This BVP, however, operates on an infinite interval, while, numerically, one can only operate on a finite, truncated interval $[-T_-, T_+]$. In this case the problem can be reformulated as

$$
\begin{cases}
\dot{x}(t) - f(x(t), p) = 0 \,, \\
L_s(p)(x(-T_-) - x_-(p)) = 0 \,, \\
L_u(p)(x(T_+) - x_+(p)) = 0 \,, \\
\int_{-T_-}^{T_+} (x(t) - x_0(t))^\top \dot{x}_0(t) \mathrm{d}t = 0 \,,
\end{cases}
$$

where the equations involving $L_s(p)$ and $L_u(p)$ form the projection boundary conditions. Here $L_s(p)$ is an $n_s \times n$ matrix where the rows span the n_s-dimensional stable eigenspace of $A^\top(x_-)$, and similarly $L_u(p)$ is an $n_u \times n$ matrix where the rows span the n_u-dimensional unstable eigenspace of $A^\top(x_+)$, where $A(x)$ denotes the Jacobi matrix of (1) at x. The projection boundary conditions then ensure that the starting point $x(-T_-)$ lies in the unstable eigenspace of x_- and that the end point $x(T_+)$ lies in the stable eigenspace of x_+ (see Fig. 5).

In parallel, one must also continue of the equilibrium points $x_-(p)$ and $x_+(p)$, unless they are fixed. Then the eigenspaces can be determined by doing a Schur decomposition of the corresponding Jacobi matrices. These must

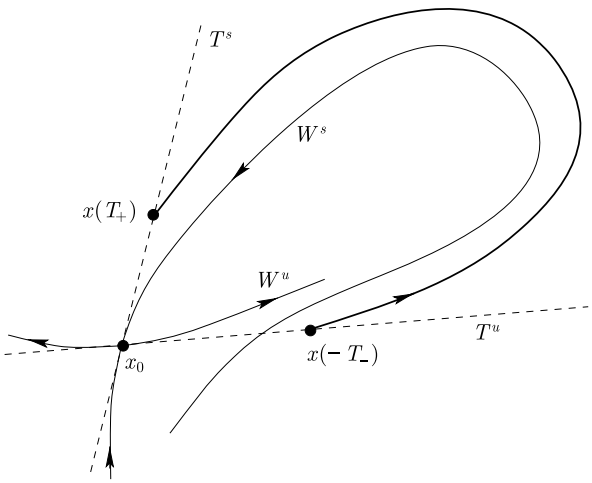

Numerical Bifurcation Analysis, Figure 5
Projection boundary conditions in two dimensions: the orbit homoclinic to x_0 is approximated by the truncated orbit from $x(-T_-)$ on the unstable eigenspace T^u to $x(T_+)$ on the stable eigenspace T^s. The unstable and stable manifolds are denoted by W^u and W^s, respectively: this figure shows that the homoclinic orbit is also approximated in parameter space, because the two manifolds do not coincide

be subsequently scaled to ensure continuity in the parameter p [6]. Alternatively, one can construct smooth projectors using an algorithm for the continuation of invariant subspaces which includes the Ricatti equation in the defining system, see [33] for this new method. All these conditions taken together give a BVP with two free parameters, since, in general, a homoclinic or heteroclinic connection is a codim 1 phenomenon.

Detecting Homoclinic Bifurcations

It is then possible to detect codim 2 bifurcations of homoclinic orbits by setting up test functions and monitoring those, detecting when they cross zero. For the continuation of these codim 2 bifurcations in three parameters such test functions can then be kept constant equal to zero, which provides an extra boundary condition. Simple test functions involve the values of leading eigenvalues of the equilibrium point, i. e., the stable and unstable eigenvalues closest to the imaginary axis. Some other bifurcations such as the inclination flip involve solving the so-called *adjoint variational equation*, which can detect whether the flow around the orbit is orientable or twisted like a Möbius strip. Homoclinics to a saddle-node, that is, where one of the eigenvalues of the equilibrium is zero, can be detected and followed similarly, by constructing appropriate test functions. For details, see [9] and the references mentioned therein.

Homoclinic Branch Switching

It is sometimes desirable to do branch switching from a homoclinic orbit to an n-homoclinic orbit, that is, a homoclinic orbit that goes through some neighborhood of the related equilibrium point $n-1$ times before finally converging to the equilibrium. Such n-homoclinic orbits arise in a number of situations involving the eigenvalues of the equilibrium and the orientation of the flow around the orbit. Suppose that the orbit is homoclinic to a saddle with complex conjugate eigenvalues (a saddle-focus). Let the *saddle-quantity* σ be the sum of the real parts of the leading stable and the leading unstable eigenvalue. If this quantity is positive, then a so-called Shil'nikov snake exists, which implies the existence of infinitely many n-periodic and n-homoclinic orbits for nearby parameters. These n-homoclinic orbits also arise from certain codim 2 bifurcations:

1. Belyakov bifurcations: Either the saddle-quantity of the saddle-focus goes through zero, or there is a transition between a saddle-focus and a real saddle (here the equilibrium has a double leading eigenvalue).
2. Homoclinic flip bifurcations: The inclination flip and the orbit flip, where the flow around the orbit changes between orientable and twisted.
3. The resonant homoclinic doubling: A homoclinic orbit for which the flow is twisted connected to a *real* saddle where the saddle quantity σ goes through zero.

Case 1. produces infinitely many n-homoclinic orbits, whereas cases 2. and 3. only produce a two-homoclinic orbit.

By breaking up a homoclinic orbit globally into two parts where the division is in a cross-section away from the equilibrium, somewhere "half-way", and then gluing pieces together, it is possible to construct n-homoclinic orbits from 1-homoclinic orbits. The gaps between to-be-glued pieces can be well-defined using Lin's method, which leaves the gaps in the direction of the adjoint vector. This vector can be found by solving the adjoint variational equation mentioned above at the "half-way" point. Combining gap distances and times taken for the flow in pieces, one can construct a well-posed BVP. This BVP can then be continued in those quantities so that if the gap sizes go to zero, one converges to an n-homoclinic orbit [66].

Lin's method was also applied in [54] to compute point-to-cycle connections by gluing a piece from the equilibrium to a cross-section to a piece from the cross section to the cycle.

Software Environments

This review has outlined necessary steps and suitable methods to perform numerical bifurcation analysis. Summarizing, the following subsequent tasks can be recognized.

Initial procedure	1. Compute the invariant solution
	2. Characterize the linearized behavior
Continuation	3. Variation of parameters
	4. Monitor dynamical indicators
Automated analysis	5. Detect special points and compute normal forms
	6. Switch branches

Ideally, these computations are automatically performed by software and indeed, many efforts have been spent on implementing the algorithms mentioned in this review and related ones. With the appearance of computers at research institutes the first codes and noninteractive packages were developed. For a recent historical overview of these and their capabilities and algorithms, see [38]. Here it is worthwhile to mention AUTO86 [24], LINBLF [51] and DSTOOL [4] as these are the predecessors to the packages AUTO-07P [21], CONTENT [58], MATCONT [19], PyDSTOOL [13] discussed here. In particular, AUTO86 is very powerful and in use to date, but not always easy to handle. Therefore, several attempts have been made to make an interactive graphical user interface. One example is XPPAUT [32] which is a standard tool in neuroscience.

The latest version of AUTO is AUTO-07P and written in Fortran and supports parallelization. It uses pseudo-arclength continuation and performs a limited bifurcation analysis of equilibria, fixed points, limit cycles and connecting orbits using HOMCONT [12]. A recent addition is the continuation of branch points [16]. It uses fully extended systems, but has specially adapted linear solvers so that it is still quite fast.

An interactive package written in C++ is CONTENT, where the Moore–Penrose continuation was first implemented. It supports bifurcation analysis of equilibria up to the continuation of codim 2 and detection of some codim 3 bifurcations. It uses a similar procedure as AUTO to continue limit cycles and detects codim 1 bifurcations of limit cycles, but does not continue these. It also handles bifurcations of cycles of maps up to the detection of codim 2 bifurcations. Normal forms for all codim 1 bifurcations are computed. Interestingly, both fully and minimally extended systems are implemented.

A new project MATCONT emerged out of CONTENT. It is written in MATLAB, a widely used software tool in modeling. In contrast to AUTO it uses Moore–Penrose continuation and minimally extended systems to compute equilibria, cycles and connecting orbits. Since MATLAB is an interpreted language, it is slower than the other packages, although a considerable speedup is obtained as the code for the Jacobi matrices for limit cycles and their bifurcations are written in C-codes and compiled. MATCONT has, however, much more functionality. It computes normal form coefficients up to codim 2 bifurcations of equilibria [20] and for codim 1 of limit cycles normal form coefficients are computed using a BVP-algorithm [28]. A recent addition is to switch branches from codim 2 bifurcation points of equilibria to codim 1 bifurcations of cycles [61].

Finally PyDSTOOL supports similar functionality for ODEs as CONTENT.

Future Directions

This review focuses on methods for equilibria and limit cycles to gain insight into the dynamics of a nonlinear ODE depending on parameters, but, generally speaking, other characteristics play a role too. For instance, a Neimark–Sacker bifurcation leads to the presence of tori. The computation and continuation of higher dimensional tori has been considered in [49,70,71]. The generalization of the methods for equilibria and cycles is, however, not straightforward. Stable and unstable manifolds are another aspect. In particular their visualization can hint at the presence of global bifurcations. A review of methods for computing such manifolds is presented in [52]. Connecting orbits can be calculated, but initializing such a continuation is a nontrivial task. This may be started from certain codim 2 bifurcation points as in [8], but good initial approximations are not available for other cases.

This review is also restricted to ODEs, but one can define other classes of dynamical systems. For instance, if the system is given by an explicitly defined map, i. e., not implicitly as a Poincaré map, the described approach can also be carried out [10,37,40,59]. Another important and related class is given by delay equations, and the algorithms for equilibria and periodic orbits in the ODE case can be applied with suitable modifications, see [5,74] and the software implementations DDE-BIFTOOL [31] and PDDE-CONT. The computation of normal forms and connecting orbits for this class is not yet thoroughly investigated or supported.

For large systems, e. g., discretizations of a partial differential equation (PDE), the algebra for some algorithms in this review becomes quite involved and numerically expensive. These problems need special treatment,

see LOCA [69], PDE-CONT [30,62] and related references. These packages focus on computing (periodic) solutions and bifurcation curves. Good algorithms for analyzing bifurcations of PDEs will be a major research topic.

Finally, there are also slow-fast (stiff) systems or systems with a particular structure such as symmetries. For these classes, many questions remain open.

While on most of the mentioned topics pioneering work has been done, the methods are far from as "complete" as for ODEs. One overview of current research topics in dynamical systems can be found in [53].

Bibliography

Primary Literature

1. Allgower EL, Georg K (2000) Numerical Continuation Methods: An Introduction. Springer, Berlin
2. Arnold VI (1983) Geometrical Methods in the Theory of Ordinary Differential Equations. Springer, Berlin
3. Ascher UC, Mattheij RMM, Russell B (1995) Numerical Solution of Boundary Value Problems for Ordinary Differenital Equations. SIAM, Philadelphia
4. Back A, Guckenheimer J, Myers M, Wicklin F, Worfolk P (1992) DsTool: Computer assisted exploration of dynamical systems. Notices Amer Math Soc 39:303–309
5. Barton DAW, Krauskopf B, Wilson RE (2007) Homoclinic bifurcations in a neutral delay model of a transmission line oscillator. Nonlinearity 20:809–829
6. Beyn WJ (1990) The numerical computation of connecting orbits in dynamical systems. IMA J Numer Anal 10:379–405
7. Beyn WJ (1991) Numerical methods for dynamical systems. In: Light W (ed) Advances in numerical analysis, Vol. I (Lancaster, 1990). Clarendon, Oxford, pp 175–236
8. Beyn WJ (1994) Numerical analysis of homoclinic orbits emanating from a Takens–Bogdanov point. IMA J Numer Anal 14:381–410
9. Beyn WJ, Champneys A, Doedel EJ, Govaerts W, Kuznetsov YA, Sandstede B (2002) Numerical continuation, and computation of normal forms. In: Fiedler B (ed) Handbook of Dynamical Systems, vol 2. Elsevier Science, Amsterdam, pp 149–219
10. Beyn WJ, Kleinkauf JM (1997) The numerical computation of homoclinic orbits for maps. Siam J Numer Anal 34:1207–1236
11. Carr J (1981) Applications of centre manifold theory. Springer, New York
12. Champneys AR, Kuznetsov YA, Sandstede B (1996) A numerical toolbox for homoclinic bifurcation analysis. Int J Bif Chaos 6:867–887
13. Clewley R, Sherwood E, LaMar D, Guckenheimer J (2005) PyDSTool, available at http://sourceforge.net/projects/pydstool. Accessed 09 Sept 2008
14. Coullet PH, Spiegel EA (1983) Amplitude equations for systems with competing instabilities. SIAM J Appl Math 43:776–821
15. de Boor C, Swartz B (1973) Collocation at Gaussian points. SIAM J Num Anal 10:582–606
16. Dercole F (2008) Bpcont: An auto driver for the continuation of branch points of algebraic and boundary-value problems. SIAM J Sci Comp 30:2405–2426
17. Deuflhard P, Bornemann F (2002) Scientific Computing With Ordinary Differential Equations. In: Marsden JE, Sirovich L, Antman S (eds) Texts in applied mathematics, vol 42. Springer, New York
18. Deuflhard P, Fiedler B, Kunkel P (1987) Efficient numerical pathfollowing beyond critical points. SIAM J Num Anal 24:912–927
19. Dhooge A, Govaerts W, Kuznetsov YA (2003) Matcont: A Matlab package for numerical bifurcation analysis of ODE's. ACM TOMS 29, pp 141–164. http://sourceforge.net/projects/matcont. Accessed 9 Sep 2008
20. Dhooge A, Govaerts W, Kuznetsov YA, Meijer HGE, Sautois B (2008) New features of the software Matcont for bifurcation analysis of dynamical systems. Math Comp Mod Dyn Syst 14:147–175
21. Doedel EJ (2007) AUTO-07P: Continuation and Bifurcation Software for Ordinary Differential Equations, User's Guide. http://cmvl.cs.concordia.ca/auto. Accessed 9 Sep 2008
22. Doedel EJ (2007) Lecture notes on numerical analysis of nonlinear equations. In: Krauskopf B, Osinga HM, Galan-Vioque J (eds) Numerical continuation methods for dynamical systems: Path following and boundary value problems. Springer-Canopus, Dordrecht, pp 1–49
23. Doedel EJ, Govaerts W, Kuznetsov YA (2003) Computation of periodic solution bifurcations in ODEs using bordered systems. SIAM J Numer Anal 41:401–435
24. Doedel EJ, Kernevez JP (1986) AUTO, Software for Continuation and Bifurcation Problems in Ordinary Differential Equations. Applied Mathematics, California Institute of Technology, Pasadena
25. Doedel EJ, Romanov VA, Paffenroth RC, Keller HB, Dichmann DJ, Vioque GJ, Vanderbauwhede A (2007) Elemental periodic orbits associated with the liberation points in the circular restricted 3-body problem. Int J Bif Chaos 17:2625–2677
26. Doedel EJ, Kooi BW, Kuznetsov YA, van Voorn GAK (2008) Continuation of connecting orbits in 3D-ODEs: (I) Point-to-cycle connections. Int J Bif Chaos 18:1889–1903
27. Doedel EJ, Kooi BW, Kuznetsov YA, van Voorn GAK (2008) Continuation of connecting orbits in 3D-ODEs: (II) Cycle-to-cycle connections. arXiv.org:0804.0179. Accessed 9 Sep 2008
28. Doedel EJ, Kuznetsov YA, Govaerts W, Dhooge A (2005) Numerical continuation of branch points of equilibria and periodic orbits. Int Bif J Chaos 15:841–860
29. Elphick C, Tirapegui E, Brachet M, Coullet P, Iooss G (1987) A simple global characterization for normal forms of singular vector fields. Phys D 32:95–127
30. Engelborghs K, Lust K, Roose D (1999) Direct computation of period doubling bifurcation points of large-scale systems of ODEs using a Newton-Picard method. IMA J Numer Anal 19:525–547
31. Engelborghs K, Luzyanina T, Roose D (2002) Numerical bifurcation analysis of delay differential equations using DDE-BIFTOOL. ACM Trans Math Softw 28:1–21
32. Ermentrout B (2002) Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students. SIAM, Philadelphia
33. Friedman M, Govaerts W, Kuznetsov YA, Sautois B (2005) Continuation of homoclinic orbits in MATLAB. In: Sunderam VS, van Albada GD, Sloot PMA, Dongarra J (eds) Proceedings ICCS 2005, Atlanta, Part I, Springer Lecture Notes in Computer Science Vol. 3514. Springer, Berlin, pp 263–270

34. Fuller AT (1968) Conditions for a matrix to have only characteristic roots with negative real parts. J Math Anal Appl 23: 71–98
35. Gaspard P (1993) Local birth of homoclinic chaos. Phys D 62:94–122
36. Govaerts W, Guckenheimer J, Khibnik A (1997) Defining functions for multiple Hopf bifurcations. SIAM J Numer Anal 34:1269–1288
37. Govaerts W, Khoshsiar Ghaziani R, Kuznetsov YA, Meijer HGE (2007) Numerical methods for two-parameter local bifurcation analysis of maps. SIAM J Sci Comput 29:2644–2667
38. Govaerts W, Kuznetsov YA (2007) Numerical continuation tools. In: Krauskopf B, Osinga HM, Galen-Vioque J (eds) Numerical continuation methods for dynamical systems: Path following and boundary value problems. Springer-Canopus, Dordrecht, pp 51–75
39. Govaerts W, Kuznetsov YA, Dhooge A (2005) Numerical continuation of bifurcations of limit cycles in matlab. SIAM J Sci Comp 27:231–252
40. Govaerts W, Kuznetsov YA, Sijnave B (1999) Bifurcations of maps in the software package content. In: Ganzha VG, Mayr EW, Vorozhtsov EV (eds) Computer algebra in scientific computing–CASC'99 (Munich). Springer, Berlin, pp 191–206
41. Govaerts W, Pryce JD (1993) Mixed block elimination for linear systems with wider borders. IMA J Num Anal 13:161–180
42. Govaerts WJF (2000) Numerical Methods for Bifurcations of Dynamical Equilibria. SIAM, Philadelphia
43. Guckenheimer J (2002) Numerical analysis of dynamical systems. In: Fiedler B (ed) Handbook of dynamical systems, Vol. 2. Elsevier Science, North-Holland, pp 346–390
44. Guckenheimer J, Holmes P (1983) Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields. Springer, New York
45. Guckenheimer J, Myers M, Sturmfels B (1997) Computing Hopf bifurcations I. SIAM J Numer Anal 34:1–21
46. Henderson Me (2007) Higher-dimensional continuation. In: Krauskopf B, Osinga HM, Galen-Vioque J (eds) Numerical continuation methods for dynamical systems: Path following and boundary value problems. Springer-Canopus, Dordrecht, pp 77–115
47. Iooss G (1988) Global characterization of the normal form for a vector field near a closed orbit. J Diff Eqs 76:47–76
48. Iooss G, Adelmeyer M (1992) Topics in Bifurcation Theory and Applications. World Scientific, Singapore
49. Jorba A (2001) Numerical computation of the normal behaviour of invariant curves of n-dimensional maps. Nonlinearity 14:943–976
50. Keller HB (1977) Numerical solution of bifurcation and nonlinear eigenvalue problems. In: Rabinowitz PH (ed) Applications of bifurcation theory. Proc Advanced Sem, Univ Wisconsin, Madison, (1976), Publ Math Res Center, No. 38. Academic Press, New York pp 359–384
51. Khibnik AI (1990) LINBLF: A program for continuation and bifurcation analysis of equilibria up to codimension three. In: Roose D, de Dier D, Spence A (eds) Continuation and bifurcations: numerical techniques and applications. Leuven, (1989), NATO Adv Sci Inst Ser C Math Phys Sci, vol 313. Kluwer, Dordrecht, pp 283–296
52. Krauskopf B, Osinga HM, Doedel EJ, Henderson ME, Guckenheimer J, Vladimirsky A, Dellnitz M, Junge O (2005) A survey of methods for computing (un)stable manifolds of vector fields. Int J Bif Chaos 15(3):763–791
53. Krauskopf B, Osinga HM, Galan-Vioque J (2007) Numerical Continuation methods for dynamical systems. Springer-Canopus, Dordrecht
54. Krauskopf B, Riess T (2008) A Lin's method approach to finding and continuing heteroclinic connections involving periodic orbits. Nonlinearity 21:1655–1690
55. Kuznetsov YA (1999) Numerical normalization techniques for all codim 2 bifurcations of equilibria in ODE's. SIAM J Numer Anal 36:1104–1124
56. Kuznetsov YA (2004) Elements of Applied Bifurcation Theory, 3rd edn. Springer, Berlin
57. Kuznetsov YA, Govaerts W, Doedel EJ, Dhooge (2005) Numerical periodic normalization for codim 1 bifurcations of limit cycles. SIAM J Numer Anal 43:1407–1435
58. Kuznetsov YA, Levitin VV (1995) Content: A multiplatform environment for analyzing dynamical systems. ftp://ftp.cwi.nl/pub/CONTENT/. Accessed 9 Sep 2008
59. Kuznetsov YA, Meijer HGE (2005) Numerical normal forms for codim 2 bifurcations of maps with at most two critical eigenvalues. SIAM J Sci Comput 26:1932–1954
60. Kuznetsov YA, Meijer HGE (2006) Remarks on interacting Neimark–Sacker bifurcations. J Diff Eqs Appl 12:1009–1035
61. Kuznetsov YA, Meijer HGE, Govaerts W, Sautois B (2008) Switching to nonhyperbolic cycles from codim 2 bifurcations of equilibria in ODEs. Physica D 237:3061–3068
62. Lust K, Roose D, Spence A, Champneys AR (1998) An adaptive Newton-Picard algorithm with subspace iteration for computing periodic solutions. SIAM J Sci Comput 19:1188–1209
63. Mei Z (1989) A numerical approximation for the simple bifurcation problems. Numer Func Anal Opt 10:383–400
64. Mei Z (2000) Numerical Bifurcation Analysis for Reaction-Diffusion Equations. Springer, Berlin
65. Moore G (1980) The numerical treatment of nontrivial bifurcation points. Numer Func Anal Opt 2:441–472
66. Oldeman BE, Champneys AR, Krauskopf B (2003) Homoclinic branch switching: a numerical implementation of Lin's method. Int J Bif Chaos 13:2977–2999
67. Roose D, Hlavaçek V (1985) A direct method for the computation of Hopf bifurcation points. SIAM J Appl Math 45:879–894
68. Russell RD, Christiansen J (1978) Adaptive mesh selection strategies for solving boundary value problems. SIAM J Num Anal 15:59–80
69. Salinger AG, Burroughs EA, Pawlowski RP, Phipps ET, Romero LA (2005) Bifurcation tracking algorithms and software for large scale applications. Int J Bif Chaos 15:1015–1032
70. Schilder F, Osinga HM, Vogt W (2005) Continuation of quasiperiodic invariant tori. SIAM J Appl Dyn Syst 4:459–488
71. Schilder F, Vogt W, Schreiber S, Osinga HM (2006) Fourier methods for quasi-periodic oscillations. Int J Num Meth Eng 67:629–671
72. Shoshitaishvili AN (1975) The bifurcation of the topological type of the singular points of vector fields that depend on parameters. Trudy Sem Petrovsk, (Vyp. 1):279–309
73. Stephanos C (1900) Sur une extension du calcul des substitutions linéaires. J Math Pures Appl 6:73–128
74. Szalai R, Stépán G, Hogan SJ (2006) Continuation of bifurcations in periodic delay-differential equations using characteristic matrices. SIAM J Sci Comput 28:1301–1317

75. van Strien SJ (1979) Center manifolds are not C^∞. Math Zeitschrift 166:143–145

Books and Reviews

Devaney RL (2003) An introduction to chaotic dynamical systems. Westview Press, Boulder

Doedel EJ, Keller HB, Kernevez JP (1991) Numerical Analysis and Control of Bifurcation Problems (I): Bifurcation in Finite Dimensions. Int J Bif Chaos 1:493–520

Doedel EJ, Keller HB, Kernevez JP (1991) Numerical Analysis and Control of Bifurcation Problems (II): Bifurcation in Infinite Dimensions. Int J Bif Chaos 1:745–772

Hirsch MW, Smale S, Devaney RL (2004) Differential equations, dynamical systems, and an introduction to chaos, 2nd edn. Pure and Applied Mathematics, vol 60 Elsevier/Academic Press, Amsterdam

Murdock J (2003) Normal Forms and Unfoldings for Local Dynamical Systems. Springer, New York

Numerical Issues When Using Wavelets

Jean-Luc Starck[1], Jalal Fadili[2]

[1] DAPNIA/SEDI-SAP, Service d'Astrophysique, CEA/Saclay, Gif sur Yvette, France

[2] GREYC CNRS UMR 6072, Image Processing Group, Ecole Nationale Supèrieure d'Ingènieurs de Caen, Caen Cedex, France

Article Outline

Notations
Glossary
Definition of the Subject
Introduction
The Continuous Wavelet Transform
The (Bi-)Orthogonal Wavelet Transform
The Lifting Scheme
The Undecimated Wavelet Transform
The 2D Isotropic Undecimated Wavelet Transform
Designing Non-orthogonal Filter Banks
Iterative Reconstruction
Future Directions
Bibliography

Notations

For a real discrete-time filter whose impulse response is $h[n], \bar{h}[n] = h[-n], n \in \mathbb{Z}$ is its time-reversed version. The hat $\hat{}$ notation will be used for the Fourier transform of square-integrable signals. For a filter h, its z-transform is written $H(z)$. The convolution product of two signals in $\ell^2(\mathbb{Z})$ will be written $*$. For the octave band wavelet representation, analysis (respectively, synthesis) filters are denoted h and g (respectively, $\tilde{h}$ and $\tilde{g}$). The scaling and wavelet functions used for the analysis (respectively, synthesis) are denoted $\phi(\phi(x/2) = \sum_k h[k]\phi(x-k), x \in \mathbb{R}$ and $k \in \mathbb{Z})$ and $\psi(\psi(x/2) = \sum_k g[k]\phi(x-k), x \in \mathbb{R}$ and $k \in \mathbb{Z})$ (respectively, $\tilde{\phi}$ and $\tilde{\psi}$). We also define the scaled dilated and translated version of ϕ at scale j and position k as $\phi_{j,k}(x) = 2^{-j}\phi(2^{-j}x - k)$, and similarly for $\psi, \tilde{\phi}$ and $\tilde{\psi}$.

Glossary

WT Wavelet Transform
CWT Continuous Wavelet Transform
DWT Discrete (decimated) Wavelet Transform
UWT Undecimated Wavelet Transform
IUWT Isotropic Undecimated Wavelet Transform

Definition of the Subject

Wavelets and related multiscale representations pervade all areas of signal processing. The recent inclusion of wavelet algorithms in JPEG 2000 – the new still-picture compression standard – testifies to this lasting and significant impact. The reason of the success of the wavelets is due to the fact that wavelet basis represents well a large class of signals, and therefore allows us to detect roughly isotropic elements occurring at all spatial scales and locations. As the noise in the physical sciences is often not Gaussian, the modeling, in the wavelet space, of many kind of noise (Poisson noise, combination of Gaussian and Poisson noise, long-memory $1/f$ noise, non-stationary noise, ...) has also been a key step for the use of wavelets in scientific, medical, or industrial applications [41]. Extensive wavelet packages exist now, commercial (see for example [29]) or non commercial (see for example [47,48]), which allows any researcher, doctor, or engineer to analyze his data using wavelets.

Introduction

Over the last two decades there has been abundant interest in wavelet methods. In many hundreds of papers published in journals throughout the scientific and engineering disciplines, a wide range of wavelet-based tools and ideas have been proposed and studied. Background texts on the wavelet transform include [11,13,26,41,44]. The most widely used wavelet transform (WT) algorithm is certainly the decimated bi-orthogonal wavelet transform (DWT) which is used in JPEG 2000. While the bi-orthogonal wavelet transform has led to successful implementa-

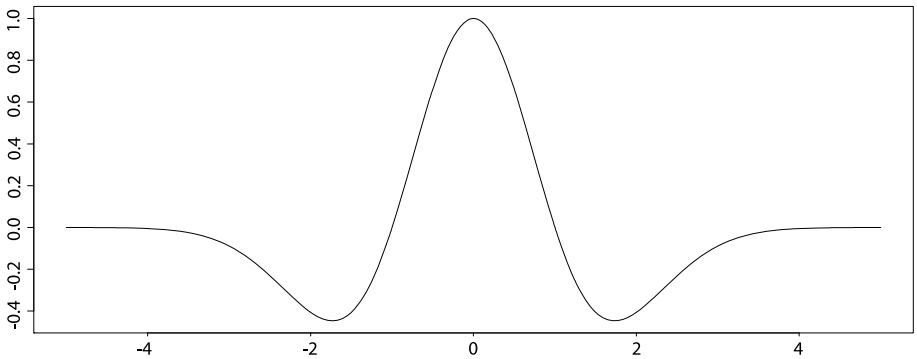

Numerical Issues When Using Wavelets, Figure 1
Mexican hat function

tion in image compression, results were far from optimal for other applications such as filtering, deconvolution, detection, or more generally, analysis of data. This is mainly due to the loss of the translation-invariance property in the DWT, leading to a large number of artifacts when an image is reconstructed after modification of its wavelet coefficients. Later efforts found that substantial improvements in perceptual quality could be obtained by translation invariant methods based on thresholding of an undecimated wavelet transform.

The Continuous Wavelet Transform

The continuous wavelet transform uses a single function $\psi(x)$ and all its dilated and shifted version to analyze signals. The Morlet–Grossmann definition [20] of the continuous wavelet transform (CWT) for a 1-dimensional real-valued function[1] $f(x) \in L^2(\mathbb{R})$, the space of all square-integrable functions, is:

$$W(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(x)\, \psi\left(\frac{x-b}{a}\right) \mathrm{d}x \tag{1}$$

where:

- $W(a,b)$ is the wavelet coefficient of the function $f(x)$,
- $\psi(x)$ is the analyzing wavelet,
- $a\,(>0)$ is the scale parameter,
- b is the position parameter.

The inverse transform is obtained by:

$$f(x) = \frac{1}{C_\psi} \int_0^{+\infty} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{a}}\, W(a,b) \times \psi\left(\frac{x-b}{a}\right) \frac{\mathrm{d}a\,\mathrm{d}b}{a^2} \tag{2}$$

[1] We only consider here real wavelets. This can be extended to complex wavelets without too much difficulty.

where:

$$C_\psi = \int_0^{+\infty} \frac{\left|\hat{\psi}\right|^2}{\nu}\, \mathrm{d}\nu = \int_{-\infty}^{0} \frac{\left|\hat{\psi}\right|^2}{\nu}\, \mathrm{d}\nu\,. \tag{3}$$

Reconstruction is only possible if C_ψ is finite (admissibility condition) which implies that $\hat{\psi}(0) = 0$, i. e. the mean of the wavelet function is 0. The wavelet is said to have a zero moment property, and appears to have a band-pass profile. A closely related relation to the inverse given in Eq. (2), is an energy conservation formula, an analogue to Plancherel's formula [26].

Figure 1 shows the Mexican hat wavelet function, which is defined by:

$$\psi(x) = \left(1 - x^2\right)\, \mathrm{e}^{-x^2/2}\,. \tag{4}$$

This is the second derivative of a Gaussian. The lower-part of Fig. 2 shows the CWT of a 1D signal (top plot of Fig. 2) computed with the Mexican Hat wavelet. This diagram is called a *scalogram*. Its y-axis represents the scale, and its x-axis represents the position parameter b.

In practice we need to discretize the scale space, and the CWT is computed for scales between $a_{\min}$ and $a_{\max}$ with a step δ_a. $a_{\min}$ must be chosen enough large to discretize properly the wavelet function, and $a_{\max}$ is limited by the number N of samples in the data. For the experiment shown in Fig. 2, $a_{\min}$ was set to 0.66 and since the dilated Mexican hat wavelet at scale a is approximately supported in $[-4a, 4a]$, we choose $a_{\max} = N/8$. The number of scales J is defined as the number of voices per octave multiplied by the number of octaves (the number of octaves is the integral part of $\log_2(a_{\max}/a_{\min})$. The number of voices per octave is generally chosen equal to 12, which guaranties a good resolution in scale and

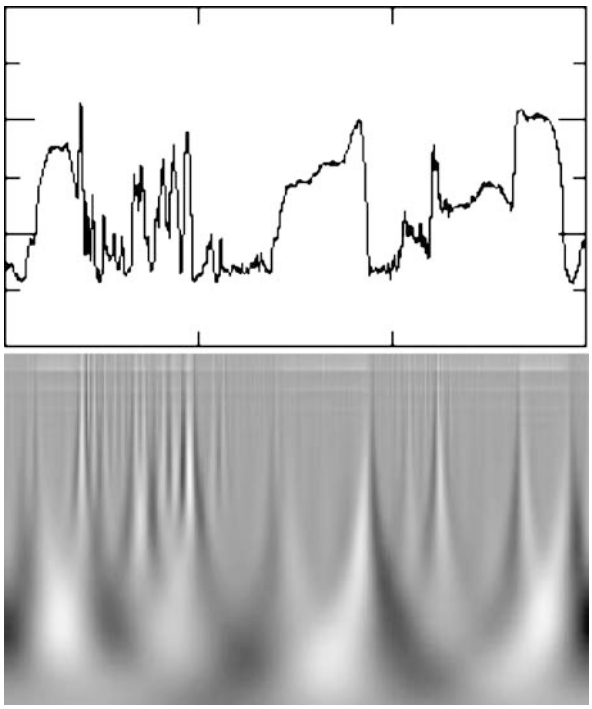

Numerical Issues When Using Wavelets, Figure 2
Top: **1D signal.** ***Bottom:*** **CWT computed with the Mexican Hat wavelet, the *y*-axis represents the scale and the *x*-axis represents the position parameter *b***

the possibility to reconstruct the signal from its wavelet coefficients. We then have $J = 12\log_2(a_{\max}/a_{\min})$, and $\delta_a = (a_{\max} - a_{\min})/(J-1)$.

The CWT algorithm is the following:

1: Set the values $a_{\min}$, $a_{\max}$, J. These values depend on both the chosen wavelet function ψ and the number of samples N.
2: Set $\delta_a = (a_{\max} - a_{\min})/(J-1)$ and $a = a_{\min}$.
3: for $a = a_{\min}$ to $a_{\max}$ with step δ_a do
- Compute $\psi_a = \psi(x/a)/\sqrt{a}$.
- Convolve the input data D with $\overline{\psi}_a$ to get $W(a,.) = 1/\sqrt{(a)}\,(\overline{\psi}_a * D)$. The convolution product can be done either in the direct space or in the Fourier space.
- $a = a + \delta_a$

4: W contains the CWT of D.

If the convolution is performed in the Fourier space (i. e. $\psi_a * D = \mathrm{IFFT}(\mathrm{FFT}(\psi_a)\mathrm{FFT}(D))$, where FFT and IFFT denote respectively the Fourier transform and its inverse), the data is assumed to be periodic. In this case, the computation of the CWT requires $O(12N(\log_2 N)^2)$ operations [26]. If the convolution is done in the direct space, we can choose other ways to deal with the borders. For instance, we may prefer to consider mirror reflexive boundary conditions (i. e. for $k = 0,\ldots,N-1$ we have $D(-k) = D(k)$ and $D(N+k) = D(N-1-k)$).

The choice of the wavelet function is let to the user. As described above, the only constraint is to have a function with a zero mean (admissibility condition). Hence, a large class of functions verifies it and we can adapt the analyzing tool, i. e. the wavelet, to the data. For oscillating data such as audio signals or seismic data, we will prefer a wavelet function which oscillates like the Morlet wavelet. While for other kind of data such as spectra, it is better to choose a wavelet function with minimum oscillation and the Mexican hat would certainly be a good choice. The wavelet function can also be complex, in which case the wavelet transform will be complex. Both the modulus and the phase will carry information about the data.

Here, we have considered only 1D data. For higher dimensional data, we can apply exactly the same approach as above. For 2D data for example, the wavelet function will be defined as a function of five parameters (position (b_x, b_y), scale in the two directions (a_x, a_y) and orientation θ) and the wavelet transform of an image will be of dimension five. But The required memory and the computation time would not be acceptable in most applications. Considering an isotropic wavelet reduces significantly to only three the dimensionality. A even more efficient approach is the (bi-)orthogonal wavelet transform algorithm.

The (Bi-)Orthogonal Wavelet Transform

Many discrete wavelet transform algorithms have been developed [26,41]. The most widely-known one is certainly the orthogonal transform, proposed by Mallat [25] and its bi-orthogonal version [13]. Here, we will introduce the bi-orthogonal through the two-channel iterated filter bank framework.

Using the bi-orthogonal wavelet transform, a signal s can be decomposed as follows:

$$s(x) = \sum_k c_J[k]\,\tilde{\phi}_{J,l}(x) + \sum_{j=1}^{J}\sum_k \tilde{\psi}_{j,k}(x)\,w_j[k] \qquad (5)$$

with $\phi_{j,l}$ and $\psi_{j,l}$ are the scaled dilated and translated version of ϕ and ψ, which are are respectively the scaling function and the wavelet function. J is the number of reso-

lution levels used in the decomposition, w_j the wavelet (or detail) coefficients at scale j, and c_J is a coarse or smooth version of the original signal s. Thus, the algorithm outputs $J+1$ subband arrays. The indexing is such that, here, $j=1$ corresponds to the finest scale (high frequencies). Coefficients $c_j[k]$ and $w_j[k]$ are obtained by means of the analysis filters h and g:

$$\begin{aligned} c_{j+1}[l] &= \sum_k h[k-2l]\, c_j[k] \\ w_{j+1}[l] &= \sum_k g[k-2l]\, c_j[k] \end{aligned} \tag{6}$$

where h and g are such that:

$$\begin{aligned} \frac{1}{2}\phi\left(\frac{x}{2}\right) &= \sum_k h[k]\,\phi(x-k) \\ \frac{1}{2}\psi\left(\frac{x}{2}\right) &= \sum_k g[k]\,\phi(x-k) \end{aligned} \tag{7}$$

and the reconstruction of the signal is performed with:

$$c_j[l] = 2\sum_k \left(\tilde{h}[k+2l]\, c_{j+1}[k] + \tilde{g}[k+2l]\, w_{j+1}[k]\right) \tag{8}$$

where the filters $\tilde{h}$ and $\tilde{g}$ must verify the conditions of dealiasing and exact reconstruction:

$$\begin{aligned} \hat{h}^*\left(\nu+\frac{1}{2}\right)\hat{\tilde{h}}(\nu) + \hat{g}^*\left(\nu+\frac{1}{2}\right)\hat{\tilde{g}}(\nu) &= 0 \\ \hat{h}^*(\nu)\,\hat{\tilde{h}}(\nu) + \hat{g}^*(\nu)\,\hat{\tilde{g}}(\nu) &= 1 \end{aligned} \tag{9}$$

or equivalently, in the z-transform domain:

$$\begin{aligned} H\left(-z^{-1}\right)\tilde{H}(z) + G\left(-z^{-1}\right)\tilde{G}(z) &= 0 \\ H\left(z^{-1}\right)\tilde{H}(z) + G\left(z^{-1}\right)\tilde{G}(z) &= 1\,. \end{aligned}$$

Note that in terms of filter banks, the bi-orthogonal wavelet transform becomes orthogonal when $h=\tilde{h}$ and $g=\tilde{g}$, in which case h is a conjugate mirror filter.

In the decomposition, c_{j+1} and w_{j+1} are computed by successively convolving a_j with the filters $\tilde{h}$ (low-pass) and g (high-pass). Each resulting channel is decimated by suppression of one sample out of two. The high-frequency channel w_{j+1} is left, and we iterate with the low-frequency part c_{j+1} (upper part of Fig. 3). In the reconstruction, we restore the sampling by inserting a 0 between each sample, then we convolve with the dual filters $\tilde{h}$ and $\tilde{g}$, we add the resulting coefficients and we multiply the result by 2. The procedure is iterated up to the smallest scale (lower part of Fig. 3).

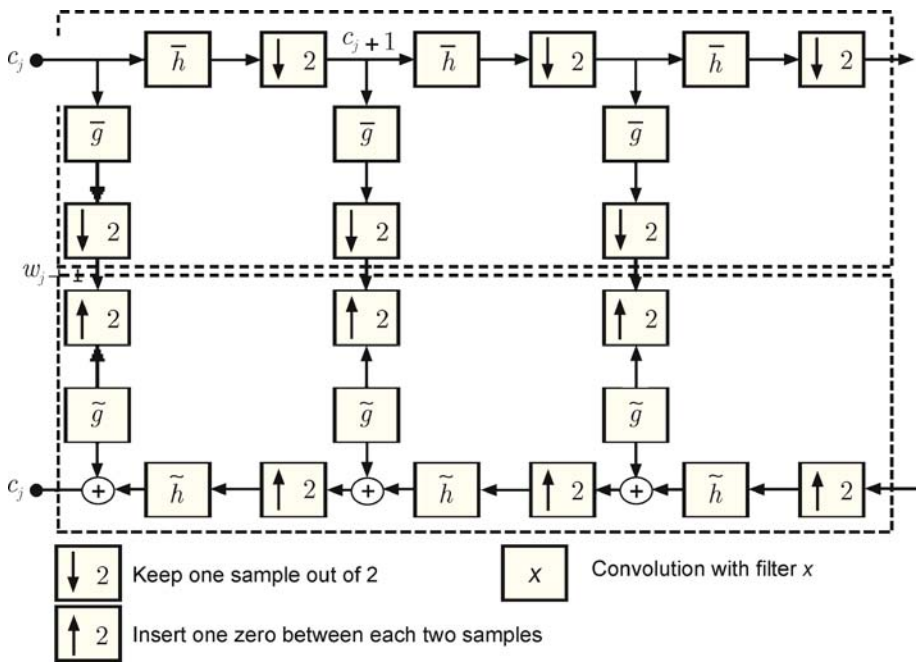

Numerical Issues When Using Wavelets, Figure 3

Fast pyramidal algorithm associated to the bi-orthogonal wavelet transform. *Top:* Fast analysis transform with a cascade of filtering with $\tilde{h}$ and $\tilde{g}$ followed by factor 2 subsampling. *Bottom:* Fast inverse transform by progressively inserting zeros and filtering with dual filters $\tilde{h}$ and $\tilde{g}$

Numerical Issues When Using Wavelets, Table 1
7/9 Filter bank (normalized to a unit mass)

h	g	$\widetilde{h}$	$\widetilde{g}$
0	0.02674875741	0.02674875741	0
−0.04563588155	0.0168641184	−0.0168641184	0.04563588155
−0.02877176311	−0.0782232665	−0.0782232665	−0.02877176311
0.295635881557	−0.26686411844	0.26686411844	−0.295635881557
0.557543526229	0.60294901823	0.60294901823	0.557543526229
0.295635881557	−0.26686411844	0.26686411844	−0.295635881557
−0.02877176311	−0.0782232665	−0.0782232665	−0.02877176311
−0.04563588155	0.0168641184	−0.0168641184	0.04563588155
0	0.02674875741	0.02674875741	0

Compared to the CWT, we have much less scales, because we consider only *dyadic scales*, i. e. scales a_j which are a power of two of the initial scale a_0 ($a_j = 2^j a_0$). Therefore, for a data set with N samples, we will typically use $J = \log(N) - 1$ scales. The algorithm is the following:

1: Set $c_0 = D$, $J = \log(N) - 1$.
2: for $j = 0$ to $J - 1$ do
- Compute $c_{j+1} = \bar{h} * c_j$, down-sample by a factor 2.
- Compute $w_{j+1} = \bar{g} * c_j$, down-sample by a factor 2.
- $j = j + 1$

3: The set $\mathcal{W} = \{w_1, \ldots, w_J, c_J\}$ represents the wavelet transform of the data.

The discrete bi-orthogonal wavelet transform (DWT) is also computationally very efficient, requiring $O(N)$ operations as compared to $O(N \log N)$ of the fast Fourier transform (N is the number of samples in data). The most used filters are certainly the 7/9 filters (by default in the JPEG 2000 norm), which are given in Table 1.

In the literature, the filter bank can be given such that it is normalized to a unit mass $\sum_k h[k] = 1$, or to a unit ℓ_2-norm $\sum_k h[k]^2 = 1$.

The above DWT algorithm can be easily extended to any dimension by *separable* (tensor) products of a scaling function ϕ and a wavelet ψ. For instance, the two-dimensional algorithm is based on separate variables leading to prioritizing of horizontal, vertical and diagonal directions. The scaling function is defined by $\phi(x, y) = \phi(x)\phi(y)$, and the passage from one resolution to the next is achieved by:

$$\begin{aligned} c_{j+1}[k, l] &= \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} h[m - 2k]\, h[n - 2l]\, c_j[m, n] \\ &= \left(\bar{h}\, \bar{h} * c_j\right)[k, l] \end{aligned} \tag{10}$$

The detail signal is obtained from three wavelets:

- Vertical wavelet: $\psi^1(x, y) = \phi(x)\,\psi(y)$
- Horizontal wavelet: $\psi^2(x, y) = \psi(x)\,\phi(y)$
- Diagonal wavelet: $\psi^3(x, y) = \psi(x)\,\psi(y)$

which leads to three wavelet subimages at each resolution level. For three dimensional data, seven wavelet subcubes are created at each resolution level, corresponding to an analysis in seven directions.

For a $N \times N$ image D, the algorithm is the following:

1: Set $c_0 = D$, $J = \log(N) - 1$.
2: for $j = 0$ to $J - 1$ do
- Compute $c_{j+1} = \bar{h}\bar{h} * c_j$, suppress one sample out of two in each dimension.
- Compute $w^1_{j+1} = \bar{g}\bar{h} * c_j$, suppress one sample out of two in each dimension.
- Compute $w^2_{j+1} = \bar{h}\bar{g} * c_j$, suppress one sample out of two in each dimension.
- Compute $w^3_{j+1} = \bar{g}\bar{g} * c_j$, suppress one sample out of two in each dimension.
- $j = j + 1$

3: The set $\mathcal{W} = \{w^1_1, w^2_1, w^3_1, \ldots, w^1_J, w^2_J, w^3_J, c_J\}$ represents the wavelet transform of the data.

The Lifting Scheme

A lifting is an elementary modification of perfect reconstruction filters, which is used to improve the wavelet properties. The lifting scheme [46] is a flexible technique that has been used in several different settings, for easy construction and implementation of traditional wavelets [46], and for the construction of wavelets on arbitrary domains such as bounded regions of $\mathbb{R}^d$ (second generation wavelets [45]) or surfaces (spherical wavelets [33]). To optimize the approximation and compression of signals and images, the lifting scheme has also been widely

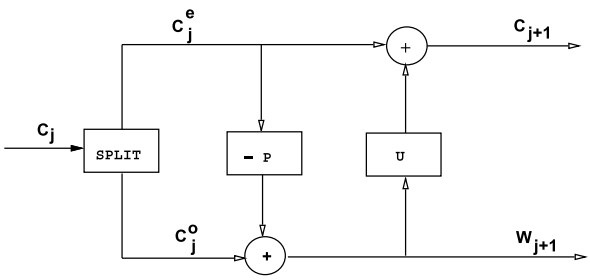

Numerical Issues When Using Wavelets, Figure 4
The lifting scheme – forward direction

used to construct adaptive wavelet bases with signal-dependent liftings. For example, short wavelets are needed in the neighborhood of singularities, but long wavelets with more vanishing moments improve the approximation of regular regions.

Its principle is to compute the difference between a true coefficient and its prediction:

$$w_{j+1}[l] = c_j[2l+1] - \mathcal{P}\big(c_j[2l-2L], \ldots, c_j[2l-2], c_j[2l], c_j[2l+2], \ldots, c_j[2l+2L]\big) . \quad (11)$$

A pixel at an odd location $2l+1$ is then predicted using pixels at even locations.

Computing the wavelet transform using lifting scheme consists of several stages. The idea is to first compute a trivial wavelet transform (the Lazy wavelet) and then improve its properties using alternating lifting and dual lifting steps. The transformation is done in three steps:

1. **Split:** This corresponds to Lazy wavelets which splits the signal into even and odd indexed samples:

$$\begin{aligned} c_j^e[l] &= c_j[2l] \\ c_j^o[l] &= c_j[2l+1] . \end{aligned} \quad (12)$$

2. **Predict:** Calculate the wavelet coefficient $w_{j+1}[l]$ as the prediction error of $c_j^o[l]$ from $c_j^e[l]$ using the prediction operator $\mathcal{P}$:

$$w_{j+1}[l] = c_j^o[l] - \mathcal{P}\big(c_j^e[l]\big) . \quad (13)$$

3. **Update:** The coarse approximation c_{j+1} of the signal is obtained by using $c_j^e[l]$ and $w_{j+1}[l]$ and the update operator $\mathcal{U}$:

$$c_{j+1}[l] = c_j^e[l] + \mathcal{U}\big(w_{j+1}[l]\big) . \quad (14)$$

The lifting steps are easily inverted by:

$$\begin{aligned} c_j[2l] &= c_j^e[l] = c_{j+1}[l] - \mathcal{U}\big(w_{j+1}[l]\big) \\ c_j[2l+1] &= c_j^o[l] = w_{j+1}[l] + \mathcal{P}\big(c_j^e[l]\big) . \end{aligned} \quad (15)$$

Some examples of wavelet transforms via the lifting scheme are:

- Haar wavelet via lifting: the Haar transform can be performed via the lifting scheme by taking the predict operator equal to the identity, and an update operator which halves the difference. The transform becomes:

$$\begin{aligned} w_{j+1}[l] &= c_j^o[l] - c_j^e[l] \\ c_{j+1}[l] &= c_j^e[l] + \frac{w_{j+1}[l]}{2} . \end{aligned}$$

  All computation can be done in-place.
- Linear wavelets via lifting: the identity predictor used before is correct when the signal is constant. In the same way, we can use a linear predictor which is correct when the signal is linear. The predictor and update operators are now:

$$\begin{aligned} \mathcal{P}\big(c_j^e[l]\big) &= \frac{1}{2}\big(c_j^e[l] + c_j^e[l+1]\big) \\ \mathcal{U}\big(w_{j+1}[l]\big) &= \frac{1}{4}\big(w_{j+1}[l-1] + w_{j+1}[l]\big) . \end{aligned}$$

  It is easy to verify that:

$$\begin{aligned} c_{j+1}[l] = &-\frac{1}{8}\, c_j[2l-2] + \frac{1}{4}\, c_j[2l-1] + \frac{3}{4}\, c_j[2l] \\ &+ \frac{1}{4}\, c_j[2l+1] - \frac{1}{8}\, c_j[2l+2] \end{aligned}$$

  which is the bi-orthogonal Cohen–Daubechies–Feauveau [12] wavelet transform.
  The lifting factorization of the popular (7/9) filter pair leads to the following implementation [14]:

$$\begin{aligned} s^{(0)}[l] &= c_j[2l] \\ d^{(0)}[l] &= c_j[2l+1] \\ d^{(1)}[l] &= d^{(0)}[l] + \alpha\big(s^{(0)}[l] + s^{(0)}\big)[l+1] \\ s^{(1)}[l] &= s^{(0)}[l] + \beta\big(d^{(1)}[l] + d^{(1)}[l-1]\big) \\ d^{(2)}[l] &= d^{(1)}[l] + \gamma\big(s^{(1)}[l] + s^{(1)}[l+1]\big) \\ s^{(2)}[l] &= s^{(1)}[l] + \delta\big(d^{(2)}[l] + d^{(2)}[l-1]\big) \\ c_{j+1}[l] &= u s^{(2)}[l] \\ w_{j+1}[l] &= \frac{d^{(2)}[l]}{u} \end{aligned} \quad (16)$$

  with

$$\begin{aligned} \alpha &= -1.586134342 \\ \beta &= -0.05298011854 \\ \gamma &= 0.8829110762 \\ \delta &= 0.4435068522 \\ u &= 1.149604398 . \end{aligned} \quad (17)$$

Every wavelet transform can be written via lifting.

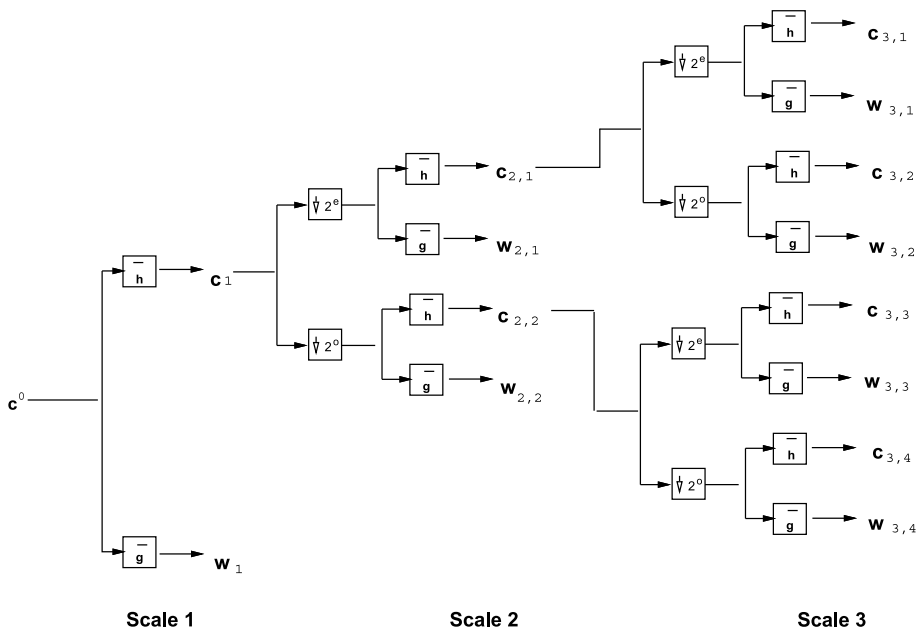

Numerical Issues When Using Wavelets, Figure 5
1D undecimated wavelet transform

Integer Wavelet Transform

When the input data consist of integer values, the wavelet transform is not necessarily integer-valued. For lossless coding and compression, it is useful to have a wavelet transform which produces integer values. We can build an integer version of every wavelet transform [7]. For instance, denoting $\lfloor x \rfloor$ as the largest integer not exceeding x, the integer Haar transform (also called "S" transform) can be calculated by:

$$\begin{aligned} w_{j+1}[l] &= c_j^{\mathrm{o}}[l] - c_j^{\mathrm{e}}[l] \\ c_{j+1}[l] &= c_j^{\mathrm{e}}[l] + \left\lfloor \frac{w_{j+1}[l]}{2} \right\rfloor \end{aligned} \tag{18}$$

while the reconstruction is

$$\begin{aligned} c_j[2l] &= c_{j+1}[l] - \left\lfloor \frac{w_{j+1}[l]}{2} \right\rfloor \\ c_j[2l+1] &= w_{j+1}[l] + c_j[2l]\,. \end{aligned} \tag{19}$$

More generally, the lifting operators for an integer version of the wavelet transform are:

$$\begin{aligned} \mathcal{P}\big(c_j^{\mathrm{e}}[l]\big) &= \left\lfloor \sum_k p[k]\, c_j^{\mathrm{e}}[l-k] + \frac{1}{2} \right\rfloor \\ \mathcal{U}\big(w_{j+1}[l]\big) &= \left\lfloor \sum_k u[k]\, w_{j+1}[l-k] + \frac{1}{2} \right\rfloor \end{aligned} \tag{20}$$

where p and u are appropriate filters associated to primal and dual lifting steps.

For instance, the linear integer wavelet transform[2] is given by

$$\begin{aligned} w_{j+1}[l] &= c_j^{\mathrm{o}}[l] - \left\lfloor \frac{1}{2}\left(c_j^{\mathrm{e}}[l] + c_j^{\mathrm{e}}[l+1]\right) + \frac{1}{2} \right\rfloor \\ c_{j+1}[l] &= c_j^{\mathrm{e}}[l] + \left\lfloor \frac{1}{4}\left(w_{j+1}[l-1] + w_{j+1}[l]\right) + \frac{1}{2} \right\rfloor . \end{aligned} \tag{21}$$

More filters can be found in [7]. In lossless compression of integer-valued digital images, even if there is no filter that consistently performs better than all the other filters on all images, it was observed that the linear integer wavelet transform performs generally better than other integer wavelet transforms using other filters [7].

The Undecimated Wavelet Transform

The undecimated wavelet transform, UWT, consists of keeping the filter bank construction which provides a fast and dyadic algorithms, but eliminating the decimation step in the orthogonal wavelet transform [17,22]: $c_1 = \bar{h} * c_0$ and $w_1 = \bar{g} * c_0$. By separating even and odd

[2]This integer wavelet transform is based upon a symmetric, bi-orthogonal wavelet transform built from the interpolating Deslauriers–Dubuc scaling function where both the high-pass filter and its dual has 2 vanishing moments moments [26].

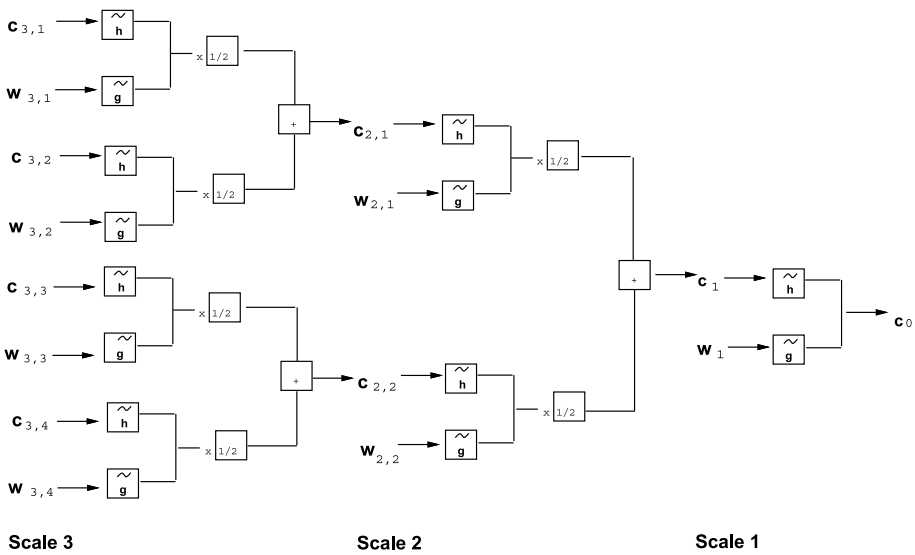

Numerical Issues When Using Wavelets, Figure 6
1D undecimated wavelet reconstruction

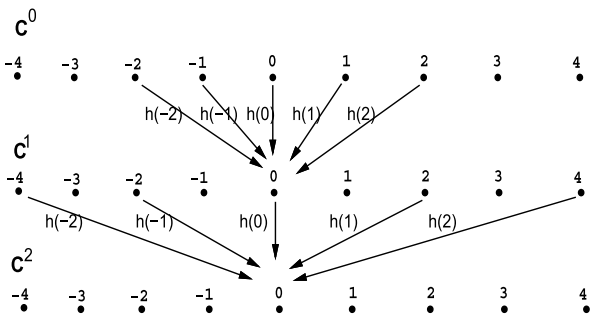

Numerical Issues When Using Wavelets, Figure 7
Passage from c_0 to c_1, and from c_1 to c_2 with the UWT à trous algorithm

pixels in c_1 and w_1, we get $(c_1^{\mathrm{E}}, w_1^{\mathrm{E}})$ and $(c_1^{\mathrm{O}}, w_1^{\mathrm{O}})$, and both parts obviously allow us to reconstruct perfectly c_0. The reconstruction can be obtained by

$$c_0 = \frac{1}{2}\left(\tilde{h} * c_1^{\mathrm{E}} + \tilde{g} * w_1^{\mathrm{E}} + \tilde{h} * c_1^{\mathrm{O}} + \tilde{g} * w_1^{\mathrm{O}}\right) . \quad (22)$$

For the passage to the next resolution, both c_1^{E} and c_1^{O} are decomposed, leading, after the splitting into even and odd pixels, to four coarse arrays associated with c_2. All of the four data sets can again be decomposed in order to obtain the third decomposition level, and so on.

Figure 5 shows the 1D UWT (UWT) decomposition. The decimation step is not applied and both w_1 and c_1 have the same size as c_0. c_1 is then split into c_1^{E} (even pixels) and c_1^{O} (odd pixels), and the same decomposition is applied to both c_1^{E} and c_1^{O}. c_1^{E} produces $c_{2,1}$ and $w_{2,1}$, while c_1^{O} produces $c_{2,2}$ and $w_{2,2}$. $w_2 = \{w_{2,1}, w_{2,1}\}$ contains the wavelet coefficients at the second scale, and is also of the same size as c_0. Figure 6 shows the 1D UWT reconstruction.

It is clear that this approach is much more complicated than the decimated bi-orthogonal wavelet transform. There exists, however, a very efficient way to implement it, called the "à trous" algorithm ("à trous", a French term, meaning *with holes*) [22,34]. $c_{j+1}[l]$ and $w_{j+1}[l]$ can be expressed as

$$\begin{aligned} c_{j+1}[l] &= \left(\tilde{h}^{(j)} * c_j\right)[l] = \sum_k h[k]\, c_j[l + 2^j k] \\ w_{j+1}[l] &= \left(\tilde{g}^{(j)} * c_j\right)[l] = \sum_k g[k]\, c_j[l + 2^j k] \,, \end{aligned} \quad (23)$$

where $h^{(j)}[l] = h[l]$ if $l/2^j$ is an integer and 0 otherwise. For example, we have

$$h^{(1)} = (\ldots, h[-2], 0, h[-1], 0, h[0], 0, h[1], 0, h[2], \ldots) .$$

The reconstruction is obtained by

$$c_j[l] = \frac{1}{2}\left[\left(\tilde{h}^{(j)} * c_{j+1}\right)[l] + \left(\tilde{g}^{(j)} * w_{j+1}\right)[l]\right] . \quad (24)$$

The filter bank $(h, g, \tilde{h}, \tilde{g})$ needs only to verify the exact reconstruction condition written in the z-transform domain:

$$H\left(z^{-1}\right) \tilde{H}(z) + G\left(z^{-1}\right) \tilde{G}(z) = 1 . \quad (25)$$

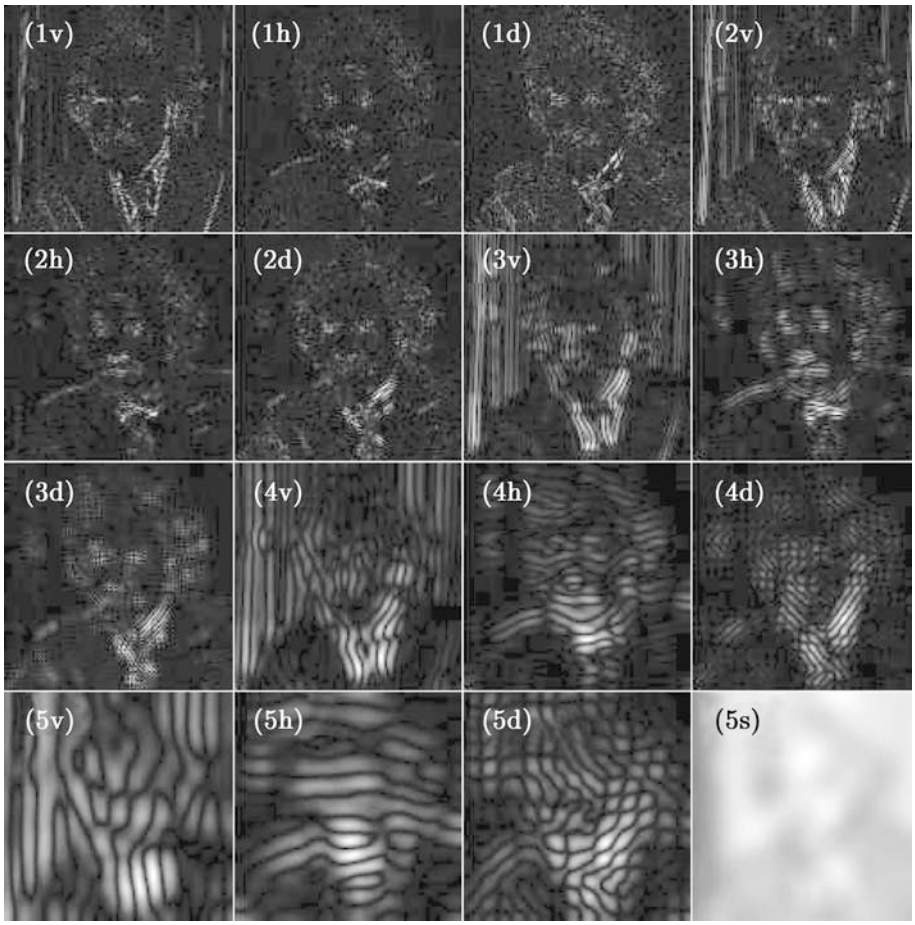

Numerical Issues When Using Wavelets, Figure 8
Undecimated wavelet transform of the *Einstein* image

This provides us with a higher degree of freedom when designing the synthesis prototype filter bank.

The à trous algorithm can be extended to 2D, by:

$$\begin{aligned} c_{j+1}[k,l] &= \left(\bar{h}^{(j)}\bar{h}^{(j)} * c_j\right)[k,l] \\ w^1_{j+1}[k,l] &= \left(\bar{g}^{(j)}\bar{h}^{(j)} * c_j\right)[k,l] \\ w^2_{j+1}[k,l] &= \left(\bar{h}^{(j)}\bar{g}^{(j)} * c_j\right)[k,l] \\ w^3_{j+1}[k,l] &= \left(\bar{g}^{(j)}\bar{g}^{(j)} * c_j\right)[k,l]\,. \end{aligned} \tag{26}$$

where $hg * c$ is the convolution of c by the separable filter hg (i. e. convolution first along the columns by h and then convolution along the rows by g). At each scale, we have three wavelet images, w^1, w^2, w^3, and each has the same size as the original image. The redundancy factor is therefore $3(J-1)+1$ [26].

The 2D Isotropic Undecimated Wavelet Transform

The Isotropic Undecimated Wavelet Transform, IUWT, algorithm is well known in the astronomical domain, because it is well adapted to astronomical data where objects are more or less isotropic in most cases [40]. Requirements for a good analysis of such data are:

- Filters must be symmetric ($\bar{h}[k] = h[k]$, and $\bar{g}[k] = g[k]$).
- In 2D or higher dimension, h, g, ψ, ϕ must be nearly isotropic.

Filters do not need to be orthogonal or bi-orthogonal and this lack of the need for orthogonality or bi-orthogonality is beneficial for design freedom. For computational reasons, we also prefer to have the separability; $h[k,l] = h[k]h[l]$. Separability is not a required condi-

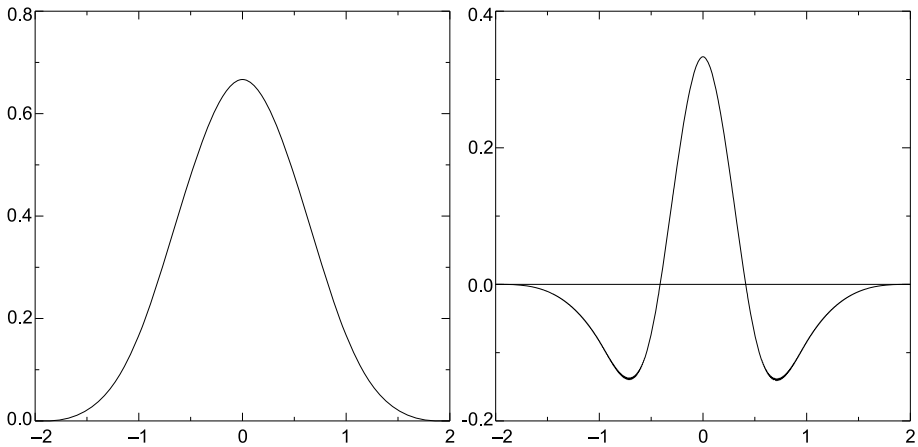

Numerical Issues When Using Wavelets, Figure 9
Left: **the cubic spline function ϕ.** ***Right:*** **the wavelet ψ. $\psi(x)$ is the difference between two resolutions**

tion, but it allows us to have a fast calculation, which is important for a large data set.

This has motivated the following choice for the analysis scaling and wavelet functions [40]:

$$\begin{aligned}\phi_1(x) &= \frac{1}{12}\big(|x-2|^3 - 4|x-1|^3 + 6|x|^3 \\ &\qquad - 4|x+1|^3 + |x+2|^3\big) \\ \phi(x,y) &= \phi_1(x)\,\phi_1(y) \\ \frac{1}{4}\,\psi\Big(\frac{x}{2},\frac{y}{2}\Big) &= \phi(x,y) - \frac{1}{4}\,\phi\Big(\frac{x}{2},\frac{y}{2}\Big)\end{aligned} \tag{27}$$

where $\phi_1(x)$ is the spline of order 3, and the wavelet function is defined as the difference between two resolutions. The related filters h and g are defined by:

$$\begin{aligned}h^{(1D)}[k] &= \frac{[1,4,6,4,1]}{16}\,, \quad k=-2,\ldots,2 \\ h[k,l] &= h^{(1D)}[k]h^{(1D)}[l] \\ g[k,l] &= \delta[k,l] - h[k,l]\end{aligned} \tag{28}$$

where δ is defined as $\delta[0,0]=1$ and $\delta[k,l]=0$ for all (k,l) different from $(0,0)$.

The following useful properties characterize any pair of even-symmetric analysis FIR (finite impulse response) filters $(h, g=\delta-h)$ such as those of Eq. (28):

Property 1 *For any pair of even symmetric filters h and g such that $g=\delta-h$, the following holds:*

(i) This FIR filter bank implements a frame decomposition, and perfect reconstruction using FIR filters is possible.
(ii) The above filters can not implement a tight frame decomposition.

See [39] for a proof.

Figure 9 shows respectively the cubic B-spline scaling function ϕ and the wavelet ψ.

From the structure of g, it is easily seen that the wavelet coefficients are obtained just by taking the difference between two resolutions:

$$w_{j+1}[k,l] = c_j[k,l] - c_{j+1}[k,l] \tag{29}$$

where $c_{j+1}[k,l] = (\bar{h}^{(j)}\bar{h}^{(j)} * c_j)[k,l]$. At each scale j, we obtain one subband $\{w_j\}$ (and not three as in the undecimated WT, denoted UWT above) which has the same number of pixels as the input image.

The reconstruction is obtained by a simple co-addition of all wavelet scales and the final smoothed array, namely

$$c_0[k,l] = c_J[k,l] + \sum_{j=1}^{J} w_j[k,l]\,. \tag{30}$$

That is, the synthesis filters are $\tilde{h}=\delta$ and $\tilde{g}=\delta$, which are indeed FIR as expected from Property 1(i). This wavelet transformation is very well adapted to the analysis of images which contain isotropic objects such as in astronomy [40] or in biology [19]. This construction has a close relation to the Laplacian pyramidal construction introduced by Burt and Adelson [6] or the FFT-based pyramidal wavelet transform [41].

Figure 10 shows the undecimated isotropic wavelet transform of the image *Einstein* using six resolution levels. This transformation contains 6 bands, each one being of the same size as the original image. The redundancy factor is therefore equal to 6. The simple addition of these six images reproduce exactly the original image.

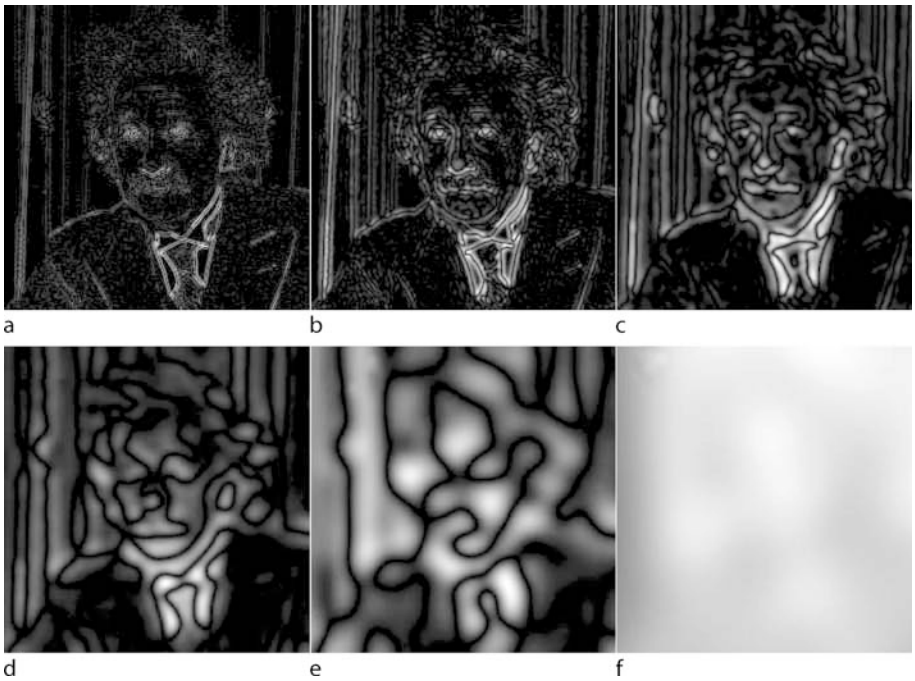

Numerical Issues When Using Wavelets, Figure 10
Undecimated isotropic wavelet transform of the *Einstein* image. The addition of these six images reproduce exactly the original image

Relation Between the UWT and the IUWT

Since the dealiasing filter bank condition is not required anymore in the UWT decomposition, we can build the standard three-directional undecimated filter bank using the non-(bi-)orthogonal "Astro" filter bank ($h^{1D} = [1, 4, 6, 4, 1]/16$, $g^{1D} = \delta - h^{1D} = [-1, -4, 10, -4, -1]/16$ and $\tilde{h} = \tilde{g} = \delta$). In two dimensions, this filter bank leads to a wavelet decomposition with three orientations w_j^1, w_j^2, w_j^3 at each scale j, but with the same property as for the IUWT, i. e. the sum of all scales reproduces the original image:

$$c_0[k, l] = c_J[k, l] + \sum_{j=1}^{J} \sum_{d=1}^{3} w_j^d[k, l] \,. \tag{31}$$

Indeed, a straightforward calculation immediately shows that [39]:

$$w_j^1 + w_j^2 + w_j^3 = c_j - c_{j+1} \,. \tag{32}$$

Therefore, the sum of the three directions reproduces the IUWT detail band at scale j. Figure 11 shows the UWT of the galaxy NGC2997. When we add the three directional wavelet bands at a given scale, we recover exactly the isotropic undecimated scale. When we add all bands, we recover exactly the original image. The relation between the two undecimated decompositions is clear.

Designing Non-orthogonal Filter Banks

A Surprising Result

Because the decomposition is non-subsampled, there are many ways to reconstruct the original image from its wavelet transform[3]. For a given filter bank (h, g), any filter bank $(\tilde{h}, \tilde{g})$ which satisfies the reconstruction condition of Eq. (25) leads to exact reconstruction. For instance, for isotropic h, if we choose $\tilde{h} = h$. (the synthesis scaling function $\tilde{\phi} = \phi$) we obtain a filter $\tilde{g}$ defined by [39]:

$$\tilde{g} = \delta + h$$

Again, as expected from Property 1, the analysis filter bank $(h, g = \delta - h)$ implements a (non-tight) frame decomposition for FIR symmetric h, where $\tilde{h} = h$ and $\tilde{g} = \delta + h$ are also FIR filters. For instance, if $h = [1, 4, 6, 4, 1]/16$, then $\tilde{g} = [1, 4, 22, 4, 1]/16$. $\tilde{g}$ *is positive* [39]. This means that $\tilde{g}$ is no longer related to a wavelet function. The synthesis scaling function related to $\tilde{g}$ is defined by:

$$\frac{1}{2}\,\tilde{\psi}\left(\frac{x}{2}\right) = \phi(x) + \frac{1}{2}\,\phi\left(\frac{x}{2}\right) \,. \tag{33}$$

[3]In frame theory parlance, we would say that the UWT frame synthesis operator is not injective.

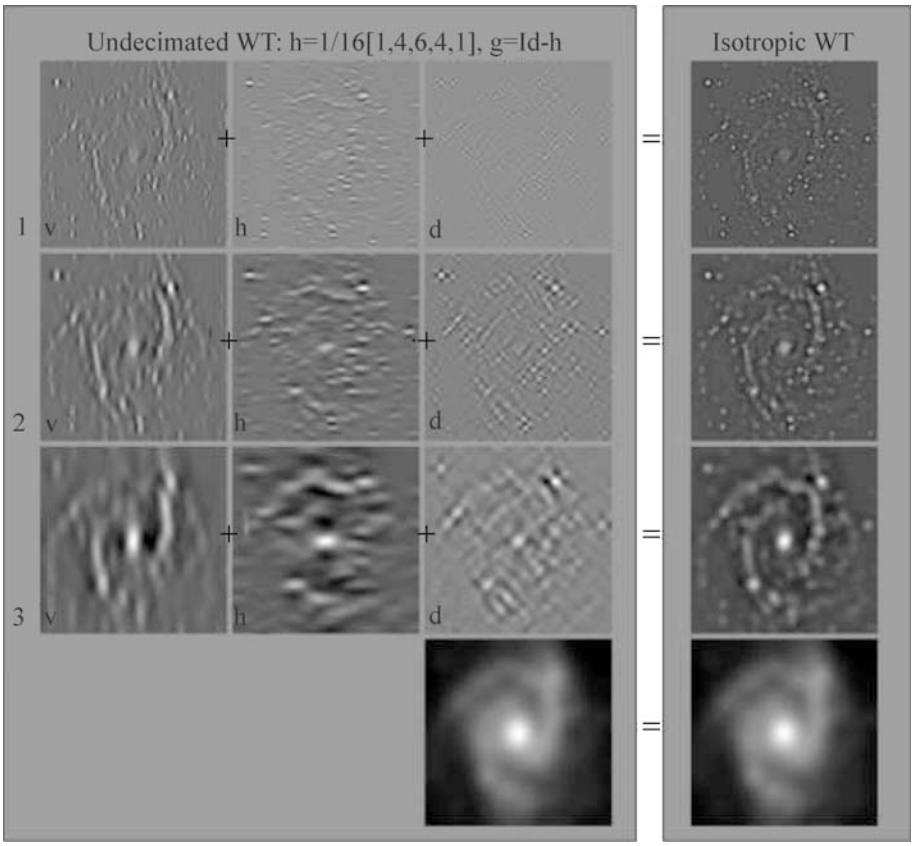

Numerical Issues When Using Wavelets, Figure 11
UWT of the galaxy NGC2997 using the Astro filter bank. The addition of three bands at a given scale is exactly the band related to the isotropic wavelet transform. Addition of all bands reproduces exactly the original image

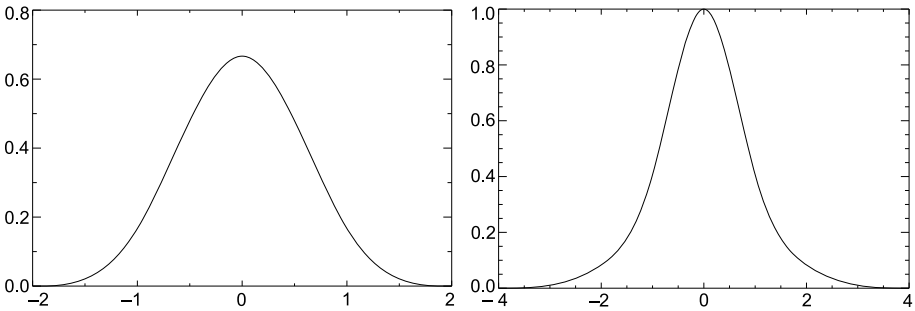

Numerical Issues When Using Wavelets, Figure 12
Left: **the $\tilde{\phi}$ synthesis scaling function.** ***Right:*** **the $\tilde{\psi}$ detail synthesis function**

Finally, note that choosing $\tilde{\phi} = \phi$, any synthesis function $\tilde{\psi}$ which satisfies

$$\hat{\tilde{\psi}}(2\nu)\,\hat{\psi}(2\nu) = \hat{\phi}^2(\nu) - \hat{\phi}^2(2\nu) \tag{34}$$

leads to an exact reconstruction [26] and $\hat{\tilde{\psi}}(0)$ can take any value. The synthesis function $\tilde{\psi}$ does not need to verify the admissibility condition (i. e. to have a zero mean).

Figure 12 shows the two scaling functions $\tilde{\phi}(x)$ $(= \phi)$ and $\tilde{\psi}(x)$ used in the reconstruction in 1D, corresponding to the synthesis filters $\tilde{h} = h$ and $\tilde{g} = \delta + h$. Figure 13 shows the backprojection of a wavelet coefficient in 2D (all wavelet coefficients are set to zero, except one), when the non-zero coefficient belongs to different bands. We can see that the reconstruction functions are positive.

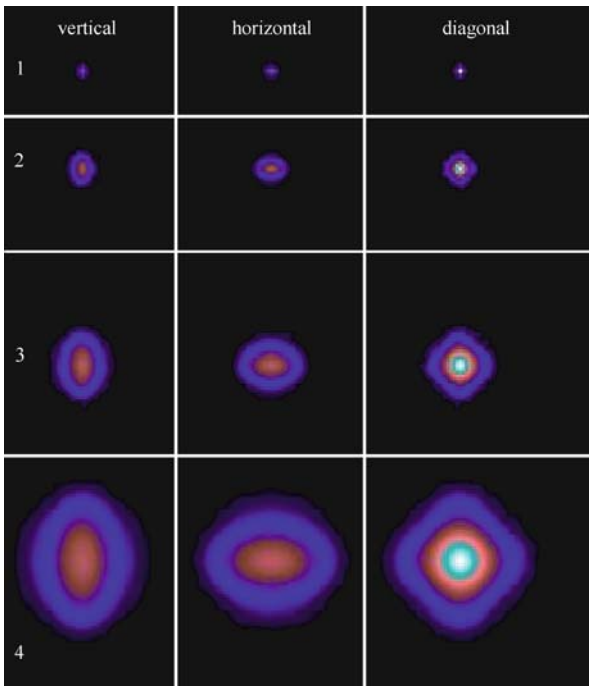

Numerical Issues When Using Wavelets, Figure 13
Back projection: Each image corresponds to the backprojection of one wavelet coefficient. All these reconstructed images are positive (no negative values). From *left* to *right*, the coefficient belongs to the vertical, horizontal and diagonal direction. From *top* to *bottom*, the scale index increases

Finally, we have an expansion of a 1D signal s,

$$s(x) = \sum_k c_J[k]\,\tilde{\phi}_{J,k}(x) + \sum_{j=1}^{J}\sum_k w_j[k]\,\tilde{\psi}_{j,k}(x) \quad (35)$$

where $\tilde{\phi}$ and $\tilde{\psi}$ are not wavelet functions (both of them have a non-zero mean and are positive), but the w_j are wavelet coefficients.

Reconstruction from the Haar Undecimated Coefficients

The Haar filters ($h = \tilde{h} = [1/2, 1/2]$, $g = \tilde{g} = [-1/2, 1/2]$) are not considered as good filters in practice because of their lack of smoothness. They are however very useful in many situations such as denoising where their simplicity allows us to derive analytical or semi-analytical detection levels even when the noise does not follow a Gaussian distribution.

Adopting the same design approach as before, we can reconstruct a signal from its Haar wavelet coefficients choosing a smooth scaling function. For instance, if $\tilde{h} = [1, 4, 6, 4, 1]/16$, it is easy to derive that the z transforms of these three filters are respectively:

$$H(z) = \frac{1 + z^{-1}}{2}\,,$$
$$G(z) = \frac{z^{-1} - 1}{2}\,,$$
$$\tilde{H}(z) = \frac{z^2 + 4z + 6 + 4z^{-1} + z^{-2}}{16}\,. \quad (36)$$

From the exact reconstruction condition in Eq. (25), we obtain:

$$\tilde{G}(z) = \frac{1 - \tilde{H}(z)H\left(z^{-1}\right)}{G\left(z^{-1}\right)}\,. \quad (37)$$

In the case of the spline filter bank, this yields after some re-arrangement (where we used simple convolution properties of splines),

$$\tilde{G}(z) = -2\,\frac{1 - z^3\left(\frac{1+z^{-1}}{2}\right)^5}{1 - z^{-1}} = z^3\,\frac{1 + 6z^{-1} + 16z^{-2} - 6z^{-3} - z^{-4}}{16} \quad (38)$$

which is the z-transform of the corresponding filter $\tilde{g} = [1, 6, 16, -6, -1]/16$.

The Haar analysis filters fulfill the following property:

Property 2 *Haar analysis filters can implement a tight frame expansion (more precisely, one scale of the Haar wavelet UWT does). Perfect reconstruction with FIR synthesis filters is possible.*

Figure 14, upper left and right, depicts the coarsest scale and a wavelet scale of the Haar transform when the input signal contains only zero values except one sample (Dirac $\delta[k]$). Figure 14, bottom left, portrays the backprojection of a Dirac at the coarsest scale (all coefficients are set to zero) and Fig. 14, bottom right, shows the backprojection of a Haar wavelet coefficient. Since the synthesis filters are regular, the backprojection of a Dirac does not produce any block staircase-like artifact. Finally, we would like to point out that other alternatives exist. For example the filter bank ($h = [1/2, -1/2]$, $g = [-1/4, 1/2, -1/4]$, $\tilde{h} = [1, 3, 3, 1]/8$ and $\tilde{g} = [1, 6, 1]/4$ leads also to an interesting solution where the synthesis filters are both positive.

Another Interesting Filter Bank

A particular case is obtained when $\hat{\tilde{\phi}} = \hat{\phi}$ and $\hat{\psi}(2\nu) = (\hat{\phi}^2(\nu) - \hat{\phi}^2(2\nu))/\hat{\phi}(\nu)$, which leads to a filter g equal to $\delta - h * h$. In this case, the synthesis function $\tilde{\psi}$

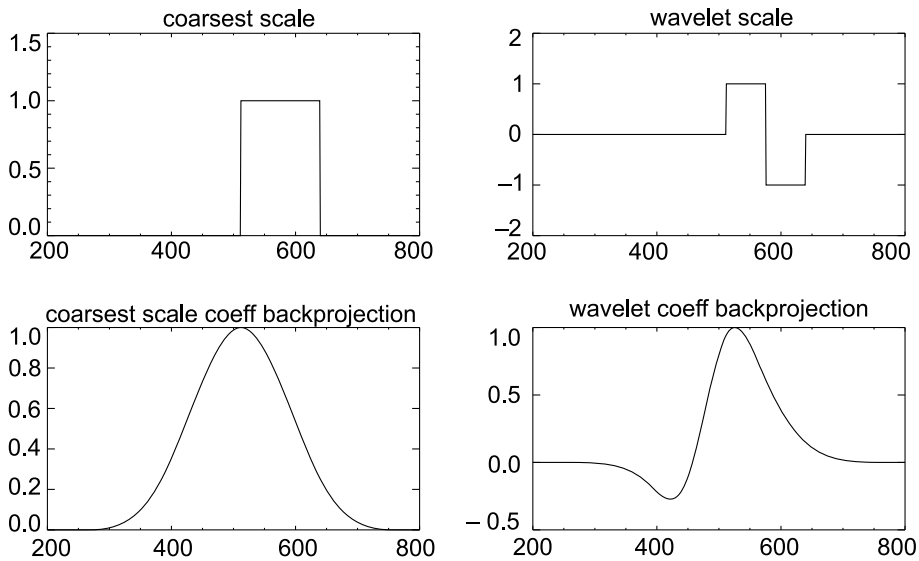

Numerical Issues When Using Wavelets, Figure 14
Haar Undecimated Transform: *Upper left:* coarsest scale when the signal is $\delta[k]$. *Upper right:* one wavelet scale of the Dirac decomposition. *Bottom left:* backprojection of a Dirac at the coarsest scale. *Bottom right:* backprojection of a Haar wavelet coefficient

is defined by $1/2\,\tilde{\psi}(x/2) = \phi(x)$ and the filter $\tilde{g} = \delta$ is the solution to Eq. (25). We end up with a synthesis scheme where only the smooth part is convolved during the reconstruction. Furthermore, for a symmetric FIR filter h, it can be easily shown that this filter bank fulfills the statements of Property 1.

Deriving h from a spline scaling function, for instance B_1 ($h_1 = [1, 2, 1]/4$) or B_3 ($h_3 = [1, 4, 6, 4, 1]/16$) (note that $h_3 = h_1 * h_1$), since h is even-symmetric (i. e. $H(z) = H(z^{-1})$), the z-transform of g is then:

$$\begin{aligned} G(z) &= 1 - H^2(z) \\ &= 1 - z^4 \left(\frac{1+z^{-1}}{2}\right)^8 \\ &= \frac{-z^4 - 8z^3 - 28z^2 - 56z + 186 - 56z^{-1}}{256} \\ &\quad - \frac{28z^{-2} - 8z^{-3} - z^{-4}}{256} \end{aligned} \tag{39}$$

which is the z-transform of the filter $g = [-1, -8, -28, -56, 186, -56, -28, -8, -1]/256$. We get the following filter bank:

$$\begin{aligned} h &= h_3 = \tilde{h} = \frac{[1, 4, 6, 4, 1]}{16} \\ g &= \delta - h * h \\ &= \frac{[-1, -8, -28, -56, 186, -56, -28, -8, -1]}{256} \end{aligned} \tag{40}$$

$$\tilde{g} = \delta \tag{41}$$

With this filter bank, there is a no convolution with the filter $\tilde{g}$ during the reconstruction. Only the low-pass synthesis filter $\tilde{h}$ is used. The reconstruction formula is:

$$c_j[l] = \left(h^{(j)} * c_{j+1}\right)[l] + w_{j+1}[l] \tag{42}$$

and denoting $L^j = h^{(0)} * \cdots * h^{(j-1)}$ and $L^0 = \delta$, we have

$$c_0[l] = \left(L^J * c_J\right)[l] + \sum_{j=1}^{J} \left(L^{j-1} * w_j\right)[l]\,. \tag{43}$$

Each wavelet scale is convolved with a low-pass filter.

Figure 15 shows the analysis scaling and wavelet functions. The synthesis functions $\tilde{\phi}$ and $\tilde{\psi}$ are the same as those in Fig. 12.

Iterative Reconstruction

Denoting $\mathcal{W}$ the undecimated wavelet transform operator and $\mathcal{R}$ the reconstruction operator, and thanks to the exact reconstruction formulae, we have the relation: $\alpha_S = \mathcal{W}\mathcal{R}\alpha_S$, where S is a signal or image and α_S its wavelet coefficients (i. e. $\alpha_S = \mathcal{W}S$). But we loose one fundamental property of the (bi-)orthogonal WT. Indeed, the relation $\alpha = \mathcal{W}\mathcal{R}\alpha$ is not true for all α sets. For example, if we set all wavelet coefficients to zero except one at a coarse scale, there is no image such that its UWT would produce a Dirac at a coarse scale. Another way

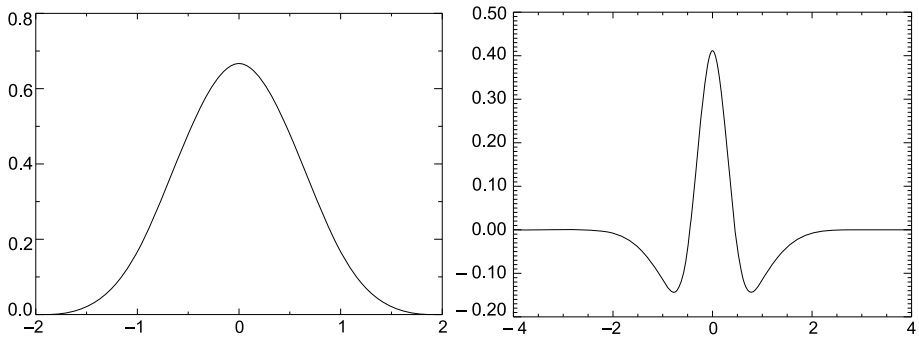

Numerical Issues When Using Wavelets, Figure 15
Left: **the ϕ analysis scaling function.** ***Right:*** **the ψ analysis wavelet function. The synthesis functions $\tilde{\phi}$ and $\tilde{\psi}$ are the same as those in Fig. 12**

to understand this point is to consider the Fourier domain of a given undecimated scale. Indeed, wavelet coefficients α_j at scale j obtained using the wavelet transform operator will contain information only localized at a given frequency band. But any modification of the coefficients at this scale, such as a thresholding ($\alpha_T = \Delta_T(\alpha)$, where Δ_T is the thresholding operator with threshold T and α_T are the thresholded coefficients), will introduce some frequency components which should not exist at this scale j, and we have $\alpha_T \neq \mathcal{W}\mathcal{R}\alpha_T$.

Reconstruction from a Subset of Coefficients

Without loss of generality, we consider hereafter the case of 1D signals. If only a subset of coefficients (for instance after thresholding) is different from zero, we would like to reconstruct an image $\tilde{S}$ such that its wavelet transform reproduces the non-zero wavelet coefficients. This can be cast as an inverse problem. We want to solve the following optimization problem $\min_{\tilde{S}} \|M(\alpha_T - \mathcal{W}\tilde{S})\|_2^2$ where $M_j[k]$ is the multiresolution support of α, i. e. $M_j[k] = 1$ if the wavelet coefficient $\alpha_j[k]$ at scale j and at position k is different from zero, and $M_j[k] = 0$ otherwise. A solution can be obtained using the Landweber iterative scheme [35,41]:

$$\tilde{S}^{n+1} = \tilde{S}^n + \mathcal{R}M\left[\alpha_T - \mathcal{W}\tilde{S}^n\right] . \tag{44}$$

If the solution is known to be positive, the positivity constraint can be introduced using the following equation:

$$\tilde{S}^{n+1} = P_+\left(\tilde{S}^n + \mathcal{R}M\left[\alpha_T - \mathcal{W}\tilde{S}^n\right]\right) \tag{45}$$

where P_+ is the projection on the cone of non-negative images. This iterative scheme can also be interpreted in terms of alternating projections onto convex sets (POCS). It has also proven very effective at many tasks such as image approximation and restoration when using the UWT [39].

Future Directions

For 2D or 3D data set, wavelet bases present some intrinsic limitations, because they are not adapted to the detection of highly anisotropic elements, such as lines or curvilinear structures in an image, or sheets in a cube. Recently, other multiscale systems like curvelets [8,9,15,36] and ridgelets [10] which are very different from wavelet-like systems have been developed. Curvelets and ridgelets take the form of basis elements which exhibit very high directional sensitivity and are highly anisotropic. A digital implementation of both the ridgelet and the curvelet transform for image denoising has been described in [36]. These new data representations, combined with wavelets, have been used in many applications such as denoising [21,32,36], deconvolution [43], contrast enhancement [42], texture analysis [2,38], detection [23], watermarking [49], component separation [37], inpainting [18] or blind source separation [4,5].

To reach higher sparsity levels, the transforms just mentioned with a fixed geometry can be replaced by adaptive representations using an optimized basis. Geometric transforms such as wedgelets [16] or bandlets [24,28] allow to define an adapted multiscale geometry. These transforms perform a non-linear search for an optimal representation. They offer geometrical adaptivity together with fast and stable algorithms. Recently, Mallat [27] proposed a more biologically inspired procedure named the grouplet transform, which defines a multiscale association field by grouping together pairs of wavelet coefficients.

Following Olshausen and Field [31], one can push one step forward the idea of adaptive sparse representation and requires that the dictionary is not fixed but rather optimized to sparsify a set of exemplar signals/images. Such a learning problem corresponds to finding a sparse matrix factorization as exposed in the K-SVD framework [1]. Ex-

plicit structural constraints such as translation invariance can also be enforced on the learned dictionary [3,30].

Bibliography

Primary Literature

1. Aharon M, Elad M, Bruckstein AM (2006) The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation. IEEE Trans Signal Process 54(11):4311–4322
2. Arivazhagan S, Ganesan L, Kumar TS (2006) Texture classification using curvelet statistical and co-occurrence features. In: Proceeding of the 18th International Conference on Pattern Recognition (ICPR 2006), vol 2. IEEE Computer Society, Los Alamitos, pp 938–941
3. Blumensath T, Davies M (2006) Sparse and shift-invariant representations of music. IEEE Trans Speech Audio Process 14(1):50–57
4. Bobin J, Moudden Y, Starck J-L, Elad M (2006) Morphological diversity and source separation. IEEE Trans Signal Process 13(7):409–412
5. Bobin J, Starck J-L, Fadili J, Moudden Y (2007) Sparsity, morphological diversity and blind source separation. IEEE Trans Image Process 16(11):2662–2674
6. Burt P, Adelson A (1983) The Laplacian pyramid as a compact image code. IEEE Trans Commun 31:532–540
7. Calderbank R, Daubechies I, Sweldens W, Yeo B-L (1998) Wavelet transforms that map integers to integers. Appl Comput Harmon Anal 5:332–369
8. Candès EJ, Demanet L, Donoho DL, Ying L (2006) Fast discrete curvelet transforms. SIAM Multiscale Model Simul 5(3): 861–899
9. Candès EJ, Donoho DL (1999) Curvelets – a surprisingly effective nonadaptive representation for objects with edges. In: Cohen A, Rabut C, Schumaker LL (eds) Curve and surface fitting: Saint-Malo 1999. Vanderbilt University Press, Nashville
10. Candès EJ, Donoho DL (1999) Ridgelets: the key to high dimensional intermittency? Philos Trans R Soc Lond A 357:2495–2509
11. Cohen A (2003) Numerical analysis of wavelet methods. Elsevier, Amsterdam
12. Cohen A, Daubechies I, Feauveau J (1992) Biorthogonal bases of compactly supported wavelets. Commun Pure Appl Math 45:485–560
13. Daubechies I (1992) Ten lectures on wavelets. Society for Industrial and Applied Mathematics Press, Philadelphia
14. Daubechies I, Sweldens W (1998) Factoring wavelet transforms into lifting steps. J Fourier Anal Appl 4:245–267
15. Do MN, Vetterli M (2003) Contourlets. In: Stoeckler J, Welland GV (eds) Beyond wavelets. Academic Press, New York
16. Donoho DL (1999) Wedgelets: nearly-minimax estimation of edges. Ann Statist 27:859–897
17. Dutilleux P (1989) An implementation of the "algorithme à trous" to compute the wavelet transform. In: Combes JM, Grossmann A, Tchamitchian P (eds) Wavelets: Time-frequency methods and phase-space. Springer, New York
18. Elad M, Starck J-L, Donoho DL, Querre P (2006) Simultaneous cartoon and texture image inpainting using Morphological Component Analysis (MCA). J Appl Comput Harmonic Anal 19:340–358
19. Genovesio A, Olivo-Marin J-C (2003) Tracking fluorescent spots in biological video microscopy. In: Conchello J-A, Cogswell CJ, Wilson T (eds) Three-dimensional and multidimensional microscopy: Image acquisition and processing X, vol 4964. SPIE Publications, pp 98–105
20. Grossmann A, Kronland-Martinet R, Morlet J (1989) Reading and understanding the continuous wavelet transform. In: Combes JM, Grossmann A, Tchamitchian P (eds) Wavelets: Time-frequency methods and phase-space. Springer, Berlin, pp 2–20
21. Hennenfent G, Herrmann FJ (2006) Seismic denoising with nonuniformly sampled curvelets. IEEE Comput Sci Eng 8(3): 16–25
22. Holschneider M, Kronland-Martinet R, Morlet J, Tchamitchian P (1989) A real-time algorithm for signal analysis with the help of the wavelet transform. In: Combes JM, Grossmann A, Tchamitchian P (eds) Wavelets: Time-frequency methods and phase-space. Springer, New York, pp 286–297
23. Jin J, Starck J-L, Donoho DL, Aghanim N, Forni O (2005) Cosmological non-Gaussian signatures detection: Comparison of statistical tests. Eurasip J 15:2470–2485
24. Le Pennec E, Mallat S (2005) Bandelet image approximation and compression. SIAM Multiscale Modeling Simul 4(3): 992–1039
25. Mallat S (1989) A theory for multiresolution signal decomposition: The wavelet representation. IEEE Trans Pattern Anal Mach Intell 11:674–693
26. Mallat S (1998) A wavelet tour of signal processing. Academic Press, New York
27. Mallat S (2006) Geometrical grouplets. Appl Comput Harmonic Anal, submitted
28. Mallat S, Peyrè G (2006) Orthogonal bandlet bases for geometric images approximation. Com Pure Appl Math, to appear
29. MR/1 (2001) Multiresolution image and data analysis software package, Version 3.0. Multi Resolutions Ltd, http://www.multiresolution.com
30. Olshausen BA (2002) Sparse coding of time-varying natural images. J Vision 2(7):130
31. Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive-field properties by learning a sparse code for natural images. Nature 381(6583):607–609
32. Saevarsson B, Sveinsson J, Benediktsson J (2006) Speckle reduction of SAR images using adaptive curvelet domain. In: Proceeding of the IEEE International Conference on Geoscience and Remote Sensing Symposium (IGARSS 2003), vol 6. IEEE Computer Society, Los Alamitos, pp 4083–4085
33. Schröder P, Sweldens W (1995) Spherical wavelets: Efficiently representing functions on the sphere. In: Computer Graphics Proceedings (SIGGRAPH 95). ACM SIGGRAPH Publications, pp 161–172
34. Shensa M (1992) Discrete wavelet transforms: Wedding the à trous and Mallat algorithms. IEEE Trans Signal Process 40:2464–2482
35. Starck J-L, Bijaoui A, Murtagh F (1995) Multiresolution support applied to image filtering and deconvolution. CVGIP: Graph Model Image Process 57:420–431
36. Starck J-L, Candès EJ, Donoho DL (2002) The curvelet transform for image denoising. IEEE Trans Image Process 11(6): 131–141
37. Starck J-L, Elad M, Donoho DL (2004) Redundant multiscale transforms and their application for morphological component analysis. Adv Imaging Electron Phys 132:288–345

38. Starck J-L, Elad M, Donoho DL (2005) Image decomposition via the combination of sparse representation and a variational approach. IEEE Trans Image Process 14(10):1570–1582
39. Starck J-L, Fadili J, Murtagh F (2007) The undecimated wavelet decomposition and its reconstruction. IEEE Trans Image Process 16:297–309
40. Starck J-L, Murtagh F (2002) Astronomical image and data analysis. Springer, Berlin
41. Starck J-L, Murtagh F, Bijaoui A (1998) Image processing and data analysis: The multiscale approach. Cambridge University Press, Cambridge
42. Starck J-L, Murtagh F, Candès EJ, Donoho DL (2003) Gray and color image contrast enhancement by the curvelet transform. IEEE Trans Image Process 12(6):706–717
43. Starck J-L, Nguyen M, Murtagh F (2003) Wavelets and curvelets for image deconvolution: A combined approach. Signal Process 83(10):2279–2283
44. Strang G, Nguyen T (1996) Wavelet and filter banks. Wellesley-Cambridge Press, Wellesley
45. Sweldens W (1997) The lifting scheme: A construction of second generation wavelets. SIAM J Math Anal 29:511–546
46. Sweldens W, Schröder P (1996) Building your own wavelets at home. In: Wavelets in Computer Graphics, ACM SIGGRAPH Course notes. ACM SIGGRAPH Publications, pp 15–87
47. Wavelab 802 (2001) Stanford University, http://www-stat.stanford.edu/~wavelab/index_wavelab802.html
48. Wavelab 805 (2005) Stanford University, http://www-stat.stanford.edu/~wavelab/
49. Zhang Z, Huang W, Zhang J, Yu H, Lu Y (2006) Digital image watermark algorithm in the curvelet domain. In: Proceeding of the International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2006). IEEE Computer Society, Los Alamitos, pp 105–108

Books and Reviews

Antoine J-P, Murenzi R, Vandergheynst P, Ali ST (2004) Two-dimensional wavelets and their relatives. Cambridge University Press, Cambridge

Chan T, Shen J (2005) Image processing and analysis: Variational, Pde, wavelet, and stochastic methods. Society for Industrial and Applied Mathematics Press, Philadelphia

Cohen A (2003) Numerical analysis of wavelet methods. Elsevier, Amsterdam

Daubechies I (1992) Ten lectures on wavelets. In: Proceedings of CBMS-NSF Regional Conference Series in Applied Mathematics Philadelphia 1992, vol 61. Society for Industrial and Applied Mathematics Press, Philadelphia

Hubbard BB (1995) The world according to wavelets: The story of a mathematical technique in the making. AK Peters, Wellesley

Jaffard S, Ryan RD, Meyer Y (2005) Wavelets: Tools for science and technology. Society for Industrial and Applied Mathematics Press, Philadelphia

Mallat S (1999) A wavelet tour of signal processing. Academic Press, New York

Meyer Y, Ryan R (1993) Wavelets: Algorithms and applications. Society for Industrial and Applied Mathematics Press, Philadelphia

Starck J-L, Murtagh F (2006) Astronomical data analysis, 2nd edn. Springer, Berlin

Starck J-L, Murtagh F, Bijaoui A (1998) Image processing and data analysis: The multiscale approach. Cambridge University Press, Cambridge

Vetterli M, Kovacevic J (1995) Wavelets and subband coding. Prentice-Hall, New Jersey

Vidakovic B (1999) Statistical modeling by wavelets. In: Wiley Series in probability and statistics. Wiley, New York

Walker JS (1999) A primer on wavelets and their scientific applications. Chapman and Hall/CRC, Boca Raton

O

Observability (Deterministic Systems) and Realization Theory

Jean-Paul André Gauthier
Department of Electrical Engineering,
University of Toulon, Toulon, France

Article Outline

Glossary
Definition of the Subject
Introduction
Preliminaries
Linear Systems
Observability of Nonlinear Systems
Realization Theory
Observers
Future Directions
Bibliography

Glossary

Observability An observed (and eventually controlled) dynamical system is observable if two distinct initial conditions can be distinguished (via the observations) by choosing the control function.

Universal inputs A universal input is a control function allowing to distinguish between all initial conditions.

Observer An observer system is a device, given in general under the guise of a differential equation (or a differences equation in the discrete case), allowing to track asymptotically the state trajectory of the system, using only the controls and the observations.

Input-output map An input-output map is a mapping (for fixed initial condition) which to "control functions" associates "output functions". It is in general assumed to be "causal" in some sense.

Realization A realization of an input-output map is a (controlled) nonlinear system realizing the given input-output map. A realization (system) is said to be minimal if it is controllable and observable.

Definition of the Subject

Observability analysis, design of nonlinear observers and realization of input-output maps are subjects of central interest in control theory and systems analysis. Related to the synthesis of observer systems is the very important question of "dynamic output stabilization": usually in practice a stabilizing feedback law is applied to the system via the estimation of the state provided by some observer device. Also, the topic is strongly connected with filtering theory, including the standard linear Kalman filter but also nonlinear filtering theory. Realization of some input-output behavior covers the practical idea of modeling systems by differential equations on the basis of input-output experiments (identification).

Introduction

In this article, we discuss the basic concepts and methods in observability, observation and realization theories. The area is so large that there are thousands of contributions. We provide a nonexhaustive limited list of references which is certainly far from complete, but corresponds to our taste: an entirely subjective selection. We focus on the continuous finite dimensional case, but there are very important developments for systems governed by PDE's, and for discrete time systems.

In this continuous, finite dimensional context, we chose the geometric setting, however there are other possibilities (algebraic setting, formal power series, Volterra series, ...).

For more details, we provide a list of books of significant interest dealing with the topics.

After setting the general definitions, we consider briefly linear systems for which the theory has been well established for a long time, the pioneers being Kalman and Luenberger.

Then we state some important results from the geometric nonlinear observability theory, the most significant contributions being undoubtedly those of Hermann and Krener [12] and Sussmann [25,26,27]. Also, contrarily

to the case of linear systems, the observability of a system depends on the control applied to it. The existence of universal controls is a very important point, clarified by Sussmann [28]. We state the main result. Concerning observability in an analytic-geometry setting, there are also interesting and important results by Bartoziewicz.

The next part of the paper is devoted to realization theory, where mostly two problems may be considered:

1. Given a nonlinear system, find a minimal realization;
2. Given some input-output mapping, find a realization of it (it will be minimal by construction).

The most important contribution in this setting is that of Jakubczyk [14,15]. In fact, it follows a basic idea of Kalman, first for finite automata and second for linear systems. We like Jakubczyk's approach since in particular, it contains very naturally the linear case. To our knowledge, this natural approach has not been used (in the nonlinear framework) for practical identification of nonlinear systems. However, it is rather clear that interesting developments are possible. Moreover, it is not so hard to show complete equivalence between this geometric approach and the formal power series approach.

The contribution of Crouch [4] about realization of finite Volterra series is also important, original and involves a lot of geometric considerations. We just refer to the original paper.

Realizing or approximating a system by a bilinear or state linear one is an important question in view of the observer synthesis problem. We state some results on the subject. In particular, there is an important geometric representation theorem (by bilinear systems) due to Fliess and Kupka [10], that we explain.

After these theoretical considerations, we go to a more practical topic: observers. Besides the linear case, there are several contributions on nonlinear observers synthesis (sliding modes, high gain, ...). Here, we focus on two natural generalizations of the linear results:

1. The output injection method (the equivalent for observability of feedback linearization) due mostly to Isidori, Krener, Respondek [19,20];
2. The use of the deterministic version of the linear Kalman's filter: it applies to bilinear systems, that are popular also by several approximation results (Fliess and Jacob in particular [13]).

Preliminaries

Surprisingly in the nonlinear case controllability plays a role in the observability properties of a system. It is the reason for the title of the next section.

Nonlinear Systems Under Consideration and Controllability

We consider nonlinear systems (Σ) of the usual form:

$$(\Sigma)\ \begin{cases} \dot{x} = f(x,u)\,, \quad u \in U\,, \\ y = h(x)\,. \end{cases} \tag{1}$$

Here, the state x lives either in $\mathbb{R}^n$ or more generally in some n-dimensional differentiable manifold X. The set U of values of control u is some arbitrary set (for simplicity, we assume a closed subset of $\mathbb{R}^l$, may be finite). The observation function h takes values in $\mathbb{R}^p$. To simplify, we will consider the analytic case only, i. e. f and h are real-analytic w.r.t. x. In the special cases where U has some analytic structure (i. e. $U = R^l$ for instance) we assume joint real analyticity w.r.t. (x,u).

If W is an open subset of X, we denote by $\Sigma|_W$ the system Σ restricted to W.

Some initial condition $x_0 \in X$ being fixed, such a system Σ defines (via Cauchy existence and uniqueness Theorem) an input-output mapping $P_\Sigma : L^\infty[U] \to AC[\mathbb{R}^p]$, $u(\cdot) \to y(\cdot)$, where $L^\infty[U]$ is the set of functions defined on semi-open intervals $[0, T_u[$ (depending on the control $u(\cdot)$). Possibly $T_u = +\infty$. Here $AC[\mathbb{R}^p]$ denotes the set of absolutely continuous functions over some interval $[0, T_y[$ possibly depending on the output function $y(\cdot)$. Moreover, $T_y = \inf\{T_u, e(u,x_0)\}$, where $e(u,x_0)$ is the explosion time of the solution of (1) associated with the initial condition x_0, and the control $u(\cdot)$.

Particular cases of systems under consideration are the usual linear systems (L), bilinear systems (B) or state-linear systems (LX):

$$\begin{cases} (L) & \dot{x} = Ax + Bu,\ y = Cx;\ X = \mathbb{R}^n,\ U = \mathbb{R}^l\,, \\ (B) & \dot{x} = Ax + Bx \otimes u,\ y = Cx;\ X = \mathbb{R}^n,\ U = \mathbb{R}^l, \\ (LX) & \dot{x} = A(u)x,\ y = Cx;\ X = \mathbb{R}^n\,. \end{cases} \tag{2}$$

In these formulas, $A, B, C, A(u)$ are linear. Of course, in the case of a linear system (L), with initial condition $x_0 = 0$, the input-output mapping P_L is a linear mapping.

Our system Σ is said to be "**controllable**" if the Lie algebra $Lie(\Sigma)$ of smooth vector fields on X generated by the vector fields f_u, $u \in U$ (where $f_u(x) = f(x,u)$) has dimension n at each point of X.

Also, we say that a system Σ is **symmetric** if $\forall\, u \in U$, $\exists\, v \in U$ s. t. $f_v = -f_u$, and Σ is **complete** if all the vector fields f_u, $u \in U$, are complete.

The following fact is standard, for analytic systems. A system is controllable iff:

1. The accessibility set $A(x_0)$ of $x_0 \in X$, i. e. the set of points that can be reached from x_0 by some trajectory of Σ, **in positive time**, has open interior in X, whatever $x_0 \in X$.
2. The orbit $O(x_0)$ of $x_0 \in X$, i. e. the set of points that can be joined to x_0 by some continuous curve which is a concatenation of trajectories of Σ **in positive or negative time**, is equal to X, whatever $x_0 \in X$.

Moreover in 1, 2 above, it is enough to restrict to piecewise constant control functions. Also, if Σ is symmetric, $O(x_0) = A(x_0)$, $\forall\, x_0 \in X$.

Definition and Characterization of Observability, Minimal Systems

Here, $C^\omega(X)$ denotes the vector space of real analytic functions over X. First, let $\Theta \subset C^\omega(X)$ denote the "**observation space of Σ**", i. e. the smallest vector subspace of $C^\omega(X)$ containing the p components $h_i(\cdot)$ of the output function h and closed under Lie derivation L_{f_u} in the direction of the vector fields f_u, $u \in U$. Then, Θ is also closed under Lie derivation in the direction of vector fields in $Lie(\Sigma)$ and Θ is generated as a real vector space by the functions $(L_{f_{u_r}})^{k_r}(L_{f_{u_{r-1}}})^{k_{r-1}} \cdots (L_{f_{u_1}})^{k_1} h_i$.

Definition 1 The observability distribution Δ of Σ is the distribution $\ker(\mathrm{d}\Theta)$ formed by the kernel of the one-forms $\mathrm{d}\theta, \theta \in \Theta$. The system Σ is said to be rank-observable if the distribution Δ is trivial. This fact is also called the "observability rank condition".

The important fact relating the observability and controllability properties is that the observability distribution Δ **has no singularities as soon as Σ is controllable**: the rank of Δ is preserved along trajectories of vector fields f_u. Moreover, it is clear that Δ is involutive, hence integrable by Frobenius's Theorem. Leaves of Δ are **levels** of Θ.

Definition 2 (Indistinguishability and weak indistinguishability relations) Let I be the binary relation over X defined by $x_0^1 I x_0^2$ if for any (piecewise constant) control $u(\cdot)$: $[0, T_u[\to U$ such that $e(u, x_0^1) = e(u, x_0^2) = T_u$, then the corresponding output functions $y_1(t), y_2(t)$ from both initial conditions x_0^1, x_0^2 are equal, $t \in [0, T_u[$. The relation I is called the indistinguishability relation for Σ. If V is an open subset of X, we denote by I_V (V-indistinguishability relation) the indistinguishability relation for the restriction $\Sigma|_V$. The weak-indistinguishability relation, denoted by I^w is the equivalence relation associated with the foliation of X generated by Δ.

The indistinguishability relation is an equivalence relation as soon as Σ is complete. It is not an equivalence relation in general. Hence in general, V-indistinguishability also is not equivalence over V.

Definition 3 The system Σ is said to be observable if the relation I is the trivial relation. It is said to be weakly observable if for all $x_0 \in X$, there is a neighborhood W of x_0 such that for each neighborhood V of x_0, $V \subset W$, $I_V(x_0) = x_0$.

Then weak observability means that **locally**, we can find inputs such that the initial conditions are distinguished by the observations, in arbitrarily short time. Observability means just that distinct initial conditions can be distinguished by observations. The system Σ being observable, analytic, this can be done in arbitrary short time.

In view of realization theory, we say that Σ is **minimal** if it is both controllable and observable. We say that it is **weakly minimal** if it is controllable and weakly observable.

Definition 4 A universal input for Σ is an input $u(\cdot)$, that distinguishes among any pair of distinct states in arbitrarily short time.

Observers

For a system Σ of the form (1) (that we assume to be observable) an observer is a system of the form:

$$\begin{cases} \dot{z} = F(z, y, u)\,, \\ \hat{x} = H(z, u)\,, \end{cases} \tag{3}$$

where $z \in Z$, some manifold. The observer system is fed by $y(t)$ and $u(t)$, the output and input of Σ. The mapping $H\colon Z \times U \to X$, and we require that, for a large set of initial condition z_0 for z, the output $\hat{x}(t)$ tracks asymptotically the state $x(t)$ of the system, i. e. at least,

$$\lim_{t \to +\infty} d(\hat{x}(t), x(t)) = 0\,, \tag{4}$$

where d is some (Riemannian) metric over X. In general, there are additional requirements on the rate of convergence to zero of the **estimation error** $\varepsilon(t) = d(\hat{x}(t), x(t))$ (such as exponential convergence, with arbitrary exponential rate).

Of course even without such additional requirements, this definition is very vague and not serious at all. It has to be made more precise, depending on the context. There are mostly two types of problems:

- This definition depends on the metric d. It may happen that $\varepsilon(t)$ goes to zero for some Riemannian metric d, although it goes to infinity for some other metric d'.

Also, the state variables z or x may explode in finite time. Therefore, in general it is reasonable to require (4) only for trajectories of Σ that remain in a given compact subset of X for all positive times. In that case, the usual convergence requirements becomes independent of the Riemannian metric d.

- One cannot expect to observe unobservable systems. Therefore, one has to require convergence for "good" inputs only.

Abstract Definition of an Input-Output Map

We define the topological group G (resp. the topological semi group S) of extended (resp. positive time) **piecewise constant** controls as follows: typical elements of G and S are words of the form:

$$\check{u}(\check{t}) = (u_k, t_k)\cdots(u_1, t_1)\,, \tag{5}$$

where $u_i \in U$ and $t_i \in \mathbb{R}$ (resp. $\mathbb{R}_+$). The operation over G and S is the concatenation of words. We consider also the neutral element ε: $\check{u}(\check{t})\varepsilon = \varepsilon\check{u}(\check{t}) = \check{u}(\check{t})$. We define the equivalence relation $\sim$ over G and S as being generated by the relations:

$$\begin{cases}(u, 0) \sim \varepsilon\,,\\ (u, s)(u, \theta) \sim (u, s+\theta)\,.\end{cases} \tag{6}$$

We consider the quotient spaces $G := G/\sim$, $S := S/\sim$. Both are embedded with the topology co-induced by the maps:

$$\check{u}(\cdot)\colon \mathbb{R}^k \to G \quad (\text{resp. } (\mathbb{R}_+)^k \to S)\,.$$

For $\theta \in \mathbb{R}_+$ and $\check{u}(\check{t}) = (u_k, t_k)\cdots(u_1, t_1) \in S$, we define

$$\begin{aligned}\theta * \check{u}(\check{t}) &= (u_{r+1}, \theta - \eta_r)(u_r, t_r)\cdots(u_1, t_1)\\ &\qquad\qquad \text{for } \theta \in [\eta_r, \eta_{r+1}[\,,\\ \eta_r &= t_1 + \cdots + t_r\,,\\ \theta * \check{u}(\check{t}) &= \check{u}(\check{t}) \quad \text{for } \theta \geq \eta_k\,.\end{aligned}$$

A real mapping: $P\colon D \subseteq G \to \mathbb{R}$ (resp. $S \to \mathbb{R}$) with open domain D is said to be analytic if, for all $\check{u}(\check{t}) \in D$, the mapping $\check{t} \to P(\check{u}(\check{t}))$ is analytic at $\check{t}$ as a mapping $\mathbb{R}^k \to \mathbb{R}$.

The domain D of P: $D \subseteq S \to \mathbb{R}$ is said to be "star-shaped" if $\theta * a \in D$ for all $\theta \in \mathbb{R}_+$ and $a \in D$.

Denote $\check{B}(\check{s}) = ((\check{b}_m(\check{s}_m), \ldots, \check{b}_1(\check{s}_1)) \in G^m$ (resp. S^m), with

$$\check{b}_i(\check{s}_i) = \left(b_{i_{n_i}}, s_{i_{n_i}}\right)\cdots\left(b_{i_1}, s_{i_1}\right), \quad b_{i_j} \in U,$$

and set:

$$\begin{aligned}\Psi^{\check{B}(\check{s})}_{\check{u}(\check{t})} &= \left(\Psi^{\check{b}_m(\check{s}_m)}_{\check{u}(\check{t})}, \ldots, \Psi^{\check{b}_1(\check{s}_1)}_{\check{u}(\check{t})}\right)\,,\\ \Psi^{\check{b}_i(\check{s}_i)}_{\check{u}(\check{t})} &= P(\check{b}_i(\check{s}_i)\check{u}(\check{t}))\,.\end{aligned}$$

The rank of P is defined as

$$\operatorname{rank}(P) = \sup_{k, \check{B}(\check{s}), \check{u}(\check{t})} \operatorname{rank} D_{\check{t}} \Psi^{\check{B}(\check{s})}_{\check{u}(\check{t})}\,,$$

where $D_{\check{t}}$ means the differential w.r.t. $\check{t} \in \mathbb{R}^k$, and all the arguments belong to the possible domain defined by the domain D of P.

Definition 5 An (abstract) input-output mapping P is an analytic mapping, from some open and star-shaped subset $D \subset S$, with finite rank.

An extension P^+ of an analytic mapping P is an analytic mapping such that $\operatorname{dom}(P) \subset \operatorname{dom}(P^+) \subset S$ and $P = P^+|_{\operatorname{dom}(P)}$ (restriction of P^+ to $\operatorname{dom}(P)$).

Remark 6 Given a pointed nonlinear system (Σ, x_0) where Σ is of the form (1) and $x_0 \in X$, it is clear that the associated input-output mapping defines an abstract input-output mapping, the rank of which is the dimension n of the state space.

Linear Systems

The simplest case for observability, design of observers and realization theory is the linear case.

Given a linear system (L) from (2) the following results are standard and more or less obvious:

- The observability property is independent of the control $u(\cdot)$ applied to the system, i.e (L) is observable iff it is observable for some fixed arbitrary control $u(\cdot)$.
- The observability distribution Δ is a field of constant planes, given by $\Delta = \cap_{i=1}^{n-1} \ker(CA^i)$. Then Σ is observable iff $\operatorname{rank}(\Delta) = 0$. This condition is known as the **observability rank condition**.
- If (L) is observable the following device (**Luenberger observer**):

$$\begin{cases}\dot{z} = (A - \Omega C)z + \Omega y + Bu\,,\\ \Omega : \mathbb{R}^n \to \mathbb{R}^p\,, \quad z \in \mathbb{R}^n\,,\end{cases} \tag{7}$$

is an **arbitrary exponential rate observer**, i. e. the matrix Ω can be chosen in such a way that the matrix $A - \Omega C$ has arbitrary spectrum, which implies:

$$\begin{aligned}\|\varepsilon(t)\| &= \|z(t) - x(t)\|\\ &\leq k(\alpha)e^{-\alpha t}\|z_0 - x_0\| = k(\alpha)e^{-\alpha t}\|\varepsilon_0\|\,,\end{aligned} \tag{8}$$

where $\alpha > 0$ is arbitrary, and k is some polynomial in α, independent of Ω.

- Any linear system restricts to a controllable one on some subspace, and is mapped to an observable one, by the canonical projection $\Pi : \mathbb{R}^n \to \mathbb{R}^n/I$ (where I is the indistinguishability relation from Definition 2).
- Let $Y(t)$ denote an "impulse response" ($Y(t) : \mathbb{R}^l \to \mathbb{R}^p, t \geq 0$). The input-output map is the causal linear mapping $P : u(\cdot) \to y(\cdot) = Y * u$, where $*$ denotes the convolution of (positive time) signals. Then assume that (as a formal power series) $Y(t) = \sum_{k=1}^{\infty} G_k \frac{t^k}{k!}$, and let $\mathcal{H}$ denote the infinite block-Hankel matrix constructed from the sequence of blocks $G_1, G_2, \ldots, G_k, \ldots$
  Then, $Y(t)$ is the impulse response of a linear system (L) iff $\mathcal{H}$ has finite rank n.

Observability of Nonlinear Systems

What is clear is that if a system is rank-observable, then it is weakly observable. This is due to a Baker–Campbell–Hausdorf like formula, valid for piecewise constant controls $\check{u}(\check{t})$:

$$y(t) = \sum (L_{f_{u_k}})^{r_k} \cdots (Lf_{u_1})^{r_1} h(x_0) \frac{t_k^{r_k} \cdots t_1^{r_1}}{r_k! \cdots r_1!} . \qquad (9)$$

Indeed by real analyticity, if $y_1(t) = y_2(t)$, all the terms $(L_{f_{u_k}})^{r_k} \cdots (Lf_{u_1})^{r_1} h(x_0^1)$ and $(L_{f_{u_k}})^{r_k} \cdots (Lf_{u_1})^{r_1} h(x_0^2)$ are equal, which contradicts the rank assumption, for x_0^1, x_0^2 close enough.

Conversely, assume that Σ is controllable and not rank-observable. Then, the observability distribution Δ is constant rank, integrable, nontrivial. Leaves of Δ are levels of Θ. By the same formula (9) points of such leaves are indistinguishable. Therefore Σ is not weakly observable. Then, the following theorem holds:

Theorem 7 *A controllable system Σ is weakly observable iff it is rank-observable.*

The other important result (Sussmann [28]) is:

Theorem 8 *If Σ is observable, there is a universal input. Moreover, the set of universal inputs is generic.*

Realization Theory

Minimal Realizations Given a Realization

We are given a realization i. e. a pointed system (Σ, x_0), $x_0 \in X$. In fact, the results follow from the Sussmann's theorem on quotient manifolds: a closed equivalence relation $\mathcal{R}$ differentiably passes to the quotient (i. e. quotient is a manifold and canonical projection is submersive) if there are enough complete vector fields that respect $\mathcal{R}$. We apply this theorem to the indistinguishability relation $\mathcal{R} = I$ in the case of a complete and controllable system. Then all vector fields of $Lie(\Sigma)$ respect I. This is exactly Sussmann's requirement, so that not only there is a quotient manifold and canonical mapping is submersive, but moreover vector fields of $Lie(\Sigma)$ pass to the quotient. Also, the elements of Θ obviously pass to the quotient.

If Σ is not controllable, then, as a first step, we can use the standard Hermann–Nagano Theorem to restrict to a (controllable) leaf of the distribution $Lie(\Sigma)$. Then, we have a similar theorem to the one of the linear case.

Theorem 9 *If Σ is complete, then we can restrict to the leaf of $Lie(\Sigma)$ containing $x_0 \in X$ to get a controllable system. Passing to the quotient manifold by the indistinguishability relation I, we get a minimal realization. Moreover, two minimal realizations are unique up to a diffeomorphism of the state spaces.*

For complete systems, there is an interesting refinement of this theorem. A realization is said to be **weakly-minimal** if it is controllable, weakly observable. It turns out that the equivalence relation I^w associated to Δ meets also Sussmann's conditions. It follows that the system goes to the quotient, and we get a weakly minimal realization $\hat{\Sigma}$ with state space $\hat{X}$. We can apply the previous Theorem 9 to $\hat{\Sigma}$ to get again the (unique) minimal realization Σ_m of Σ, with state space X_m. The following theorem is almost obvious.

Theorem 10 *$\hat{X}$ is a covering space of X_m. Moreover, any covering space of X_m determines a weakly-minimal realization of Σ, by a trivial lifting procedure.*

In particular, there is (up to diffeomorphisms) a single simply-connected weakly-minimal realization.

Note that in fact the relation I^w is the same relation as: $x_0^1 I^w x_0^2$ if there is a continuous curve $\gamma : [0, 1] \to X$ connecting x_0^1 to x_0^2 and for $r, s \in [0, 1]$, $\gamma(r) I \gamma(s)$.

Minimal Realizations Given an Abstract Input-Output Map

The set of controls U being given, we consider an abstract input-output map defined over the whole group G (domain $D = G$). Note that this is the case in particular for the input-output mappings determined by a complete symmetric system.

In that case we have the following theorem, due to Jakubczyk [14].

Theorem 11 *An abstract input-output mapping with domain G has a unique minimal realization, which is complete.*

Remark 12 The finite rank assumption for the input-output mapping is a generalization of the finite rank assumption of the Hankel matrix of the linear case. It is also the analog of certain finite rank assumptions appearing in the formal power series approach of Fliess, or in the Volterra-kernels approach.

Remark 13 There is one ugly detail in this theory: in general, we do not get a paracompact manifold as the state space X.

The idea for the proof of the theorem is very simple: we consider the subgroup H of G defined by $H = \{a \in G | P(ca) = P(c), \forall c \in G\}$. Then, the state space will just be $X = G/H$. The finite rank condition implies that X has the structure of a Hausdorff analytic manifold. The output function h is defined by $h(gH) = P(g)$. The vector-field f_u is defined via its one parameter group: $\exp(tf_u)(gH) = \Pi((u,t)g)$, where $\Pi : G \to G/H$ is the canonical projection.

A more practical result is the following: if we assume that the set U of values of the control **is a finite set**, then the following global result (containing a local one) can be proven.

Theorem 14 *Assume U is finite, then, a necessary and sufficient condition for P to have a realization (weakly-minimal) is that P has an extension P^+ with star-shaped domain D^+. The state space X of this realization is Hausdorff, paracompact.*

In the general analytic case with infinite U, existence of certain **local** realizations only, can be proved.

Bilinear or State-Linear Realization

This point will be extremely important for the problem of constructing observer systems (Sect. "Observers"). A system is said to be control affine if the vector fields f_u form an affine family w.r.t. u. The single control case ($l = 1$) is just the case $f(x,u) = f(x) + g(x)u$ where f and g are two vector fields on X. Note that a bilinear system is just a state linear system, which is moreover affine in the controls.

A state linear realization (LX, x_0) from Eq. (2) is said to be minimal if it is observable and controllable in the following sense: the orbit of x_0 is not contained in a strict subspace of $\mathbb{R}^n$ (the smallest such subspace would be automatically invariant under all the operators $A(u), u \in U$). First, it is rather simple to show that any pointed state-linear system (LX, x_0) has a minimal state-linear realization. Of course, the additional property to be bilinear is hereditary.

Note that for state-linear systems, the observation space is a (finite-dimensional) vector space of linear forms over X. It turns out that this finite dimensionality condition is in fact a necessary and sufficient condition. This is a very important result from Fliess and Kupka [10]:

Theorem 15 *Assume that Σ has a finite dimensional observation space Θ. Then, Σ is embeddable in a state-linear system. In other terms (Σ, x_0) has a state linear (minimal) realization.*

The proof is very easy. It is enough to take:

- $X = \Theta^*$ (dual space of Θ),
- For $\varphi \in \Theta^*$, $C_i\varphi = \varphi(h_i)$, $i = 1, \ldots, p$,
- $A(u) = (L_{f_u})^*$ (transpose of L_{f_u}),
- The initial state $\hat{x}_0$ meets $\hat{x}_0(\varphi) = \varphi(x_0)$ for $\varphi \in \Theta$.

Besides the fact that this result allows one to solve the observer problem for such systems, "truncating" in some manner the observation space is a way to approximate systems by state-linear ones, and to get approximate observers.

An interesting particular case where this theorem applies is the case of systems with polynomial observation h and state-linear dynamics:

$$\begin{cases} \dot{x} = A(u)x \,, \\ y = P(x) \,, \end{cases}$$

where P is some polynomial mapping. It is clear that Θ is finite-dimensional. More generally, if we start with a system with state-linear dynamics, we can uniformly approximate h on compact sets by a polynomial mapping to get a state-linear realization (and later on, an approximate observer device).

State-Linear Skew-Adjoint Realization

Here, for the sake of simplicity in the exposition we limit ourselves to the single output case $p = 1$.

This section describes some particular cases and some generalizations of the results of the previous section, in view of synthesis of observers with a method presented in Subsect. "Observers for Skew-Adjoint State Linear Systems".

For some reason that will be made clear in the Subsect. "Observers for Skew-Adjoint State Linear Systems" we would like to know when it is possible to embed a system (or to have a realization of a system) into a skew-symmetric, or more generally skew-adjoint, state-linear one.

This means that all the matrices $A(u)$ are skew-symmetric w.r.t. the usual scalar product over the state space $\mathbb{R}^n$ of the realization. By Theorem 15, a necessary condition for the nonlinear system Σ (minimal and complete) to be embeddable is that Θ be finite dimensional and hence the group of diffeomorphisms of X generated by the vector fields f_u be a Lie group G. One could think that a necessary condition is that G be a compact Lie group. This is not the case as the following example shows:

$$\begin{cases} \dot{x} = u\,, & x, u \in \mathbb{R}\,, \\ y = \cos(x) + \cos(\alpha x)\,, & \text{where } \alpha \text{ is irrational}\,, \end{cases}$$

since the group of diffeomorphisms is $\mathbb{R}$, while a skew symmetric embedding does exist.

The proper condition is given by the following theorem:

Theorem 16 *The system Σ (complete, minimal) can be embedded into a state-linear skew symmetric system iff:*

1. $\dim(\Theta) < \infty$ *(from which it follows that G is a Lie group),*
2. *The observation function $h(x)$ lifts over G into $\tilde{h}$ (in a natural way), an almost periodic function over G.*

Recall that an almost periodic function over G is a function that prolongs into a continuous function over the Bohr compactification $G^{\flat}$ of G [5]. The two conditions of Theorem 16 are equivalent to the fact that G is a Lie group and $\tilde{h}$ is a finite linear combination of coefficients[1] of unitary irreducible finite dimensional representations of G.

If G is "embeddable in a compact group", i. e. if G is the semi-direct product of a compact group by a finite dimensional real vector space then, any h can be approximated in some sense by an almost periodic one.

Actually, a special interesting case is the following: the system Σ is such that $X = G$, a compact Lie group, and the vector fields f_u are right invariant vector fields over G. We can take h as any continuous function $h\colon G \to \mathbb{R}$, and consider the abstract Fourier transform $\hat{h}$ of h. In fact, by Peter–Weyl's Theorem [3], h is a uniform limit over G of **finite** linear combinations of the form

$$h(g) = \sum_i \alpha_i \Phi_i(g)\,,$$

where $\Phi_i(g)$ is a coefficient of an irreducible (hence finite dimensional) unitary representation of G. This means that h has approximations h_m that converge uniformly to h over G, such that each system

$$(\Sigma_m) \begin{cases} \dot{g} = A(u)g\,, \\ y = h_m(g)\,, \end{cases}$$

has a state-linear minimal realization of the form:

$$\begin{cases} \dot{x} = A_m(u)x\,, \quad x \in \mathbb{C}^n\,, \quad A_m(u) \text{ is skew-adjoint}\,, \\ y = C_m x\,. \end{cases}$$

Hence **the input-output mapping of any right invariant system over a compact group can be approximated by the one of a skew-adjoint state-linear one**.

Now, let us consider again a (complete, minimal) system Σ, with finite dimensional Lie algebra, but the group G is not compact. In that case $\tilde{h}$ (a lift of h over G) can be approximated uniformly on any compact subset of G by a function h_m, which is a finite linear combination of "positive type" functions over G. This approximation result is known as the Gelfand–Raikov Theorem [5]. As a consequence we have the theorem:

Theorem 17 *The system*

$$(\Sigma_n) \begin{cases} \dot{g} = A(u)g\,, \\ y = h_m(g)\,, \end{cases}$$

has a (infinite dimensional) skew-adjoint state linear realization on a separable complex Hilbert space $\mathcal{H}$, i.e:

$$\begin{cases} \dot{\Psi} = A(u)\Psi\,, \\ y = \langle \Psi, \xi \rangle\,. \end{cases}$$

Here $\xi, \Psi \in \mathbb{H}$ and $\langle \cdot, \cdot \rangle$ is the scalar product over $\mathcal{H}$. All the operators $A(u)$ are densely defined, essentially skew-adjoint operators, infinitesimal generators of strongly continuous one parameter groups of unitary operators over $\mathcal{H}$.

With this result, in Subsect. "Observers for Skew-Adjoint State Linear Systems", we will be able to construct reasonable approximate observers for Σ.

Observers

Kalman's Observer for State-Linear Systems

This is just the deterministic version of the linear time-dependant Kalman filter. Therefore, inputs being known, it applies to state-linear systems (LX) from (2). Contrarily to linear systems, observability for those systems is not a property independent of the inputs: for some input $u(\cdot)$

[1] A coefficient of a representation is a coefficient of the matrix representing the representation operator in certain orthonormal basis.

it might be observable, for others it might not be. Clearly, if we want the observer to have some asymptotic property of convergence of the estimation error, it is reasonable to require that the input under consideration keeps a certain **minimum level of observability** when the time grows to infinity. It is natural to consider inputs living in the space $\mathcal{U} = L^{\infty}_{[0,\infty[,\mathbb{R}^p}$ of measurable U-valued bounded functions. For an input $u \in \mathcal{U}$ and for a real $a \geq 0$, set $u_a(t) = u(t+a)$. We denote by $\Phi_u(t)$ the matrix resolvent of the linear equation $\dot{\Phi}_u(t) = A(u(t))\Phi_u(t)$. Then for $T > 0$, the Gramm-observability matrix:

$$G_{u,T} = \int_0^T \Phi_u(t)^* C^* C \Phi_u(t)\, \mathrm{d}t\,, \tag{10}$$

where * stands for adjoint operator, measures observability in the following sense: the system is observable for $u\colon [0,T] \to U$ iff $G_{u,T}$ is positive definite. Hence there are several types of assumptions that are possible to express that $u\colon [0,+\infty[\to U$ keeps a certain level of observability when time passes. The most simple one is the following:

There are $\alpha, T, T_0 > 0$ such that for all $\theta \geq T_0$, $G_{u_\theta,T} \geq \alpha.Id_n$, where Id_n is the identity matrix. This condition means intuitively that, from time T_0 on, the input u has minimum observability level α on all time intervals of length T. Such an input could be called a "**persistent-excitation**" for Σ.

Then, the following theorem is just a restatement of the classical results about the deterministic version of the linear time-dependant Kalman's filter:

Theorem 18 ([18]) *The matrices Q and R being positive definite symmetric matrices with adequate dimensions, the Riccati system:*

$$\begin{cases} (1) \quad \dot{S} = -A(u(t))'S(t) - S(t)A(u(t)) + C^* R^{-1} C - SQS, \\ (2) \quad \dot{z} = A(u(t))z - S^{-1}C^*R^{-1}(Cz - y(t))\,, \end{cases} \tag{11}$$

is an asymptotic observer for persistent-excitations $u(\cdot)$. Convergence of the estimation error is exponential. The matrices $S(t)$ (as soon as the same holds for the initial condition S_0) live in the open cone of positive definite symmetric matrices.

Observers for Systems that are Injectable in a State-Linear One

Of course, the technique of the previous section applies stricto-sensu to such systems from Subsect. "Bilinear or State-Linear Realization".

The Output-Injection Idea

It turns out that both the Luenberger observer (7) for linear systems and the Kalman observer (11) for state-linear systems can be applied in more general nonlinear situations.

Assume that Σ is **linear "up to output injection"**, i. e.

$$(\Sigma) \left\{ \begin{array}{c} \dot{x} = Ax + \varphi(y,u) \\ y = Cx \end{array} \right\}, \tag{12}$$

or respectively that Σ is **state-linear (or bilinear) up to output injection**, i. e.

$$(\Sigma) \left\{ \begin{array}{c} \dot{x} = A(u)x + \varphi(y,u) \\ y = Cx \end{array} \right\}, \tag{13}$$

where φ (the output injection) is some nonlinear term depending on the output and input only. Then there are easy modifications of the Luenberger observer (resp. the Kalman's observer) that provide exactly the same results of convergence of the estimation error as for the corresponding systems without the output-injection term.

For case (12) we take the observer under the Luenberger-modified form:

$$\dot{z} = (A - \Omega C)z + \varphi(y,u) + \Omega(y - Cz)\,,$$

while for case (13) we take:

$$\begin{cases} \dot{S} = -A(u(t))'S(t) - S(t)A(u(t)) + C^*R^{-1}C - SQS \\ \dot{z} = A(u(t))z + \varphi(y,u) - S^{-1}C^*R^{-1}(Cz - y(t))\,. \end{cases}$$

To check the result it is enough to write the estimation error equation and to see that it is exactly the same as in the situation without output-injection.

For that reason, **it is important to characterize systems that can be put under the form of a linear or state-linear system up to output-injection**.

There is an industry around this question. It starts with the works of Isidori, Krener, Respondek [19,20]. The first result of this type is in the uncontrolled case. For an uncontrolled system

$$(\Sigma) \begin{cases} \dot{x} = f(x) \\ y = h(x)\,, \end{cases}$$

with single output ($p = 1$), consider the vector fields X_i defined by

$$L_{X_1}(L_f)^{i-1}h = \delta_{i,n}\,, \quad i = 1,\ldots,n\,,$$
$$\text{where } \delta \text{ is the Kronecker symbol}\,,$$
$$X_j = -[f, X_{j-1}]\,, \quad j = 2,\ldots,n\,,$$

The system Σ can be linearized up to a diffeomorphism

and an output injection iff the two following conditions are met [19]:

1. The family $\{\mathrm{d}h, \mathrm{d}L_f h, \ldots, \mathrm{d}(L_f)^{n-1}h\}$ has full rank n at all points of X.
2. $[X_k, X_m] = 0$ for $1 \leq k, m \leq n$.

Of course, this is a "local almost everywhere result" only.

There is also a lot of results on the problem of characterizing systems that are diffeomorphic to or **embeddable in state-linear systems up to output injection**. A significant result to the problem of embedding up to output injection is the one of Jouan [16].

Observers for Skew-Adjoint State Linear Systems

Again, to simplify the exposition we consider the single output case $p = 1$ only.

In this case we have a (minimal) state-linear realization which is also skew-adjoint, there is a construction of an observer which is **much simpler** than Kalman's one [no Riccati equation besides the prediction-correction Eq. (11), (2)]. Moreover this construction **extends to infinite-dimensional realizations**, a fact which allows it to treat any (complete minimal) system with finite dimensional Lie algebra.

To start, consider some skew-symmetric state linear system:

$$(LX)\begin{cases} \dot{x} = A(u)x\,, \quad A(u) \text{ skew-symmetric } \forall\, u \in U\,, \\ y = Cx\,, \end{cases} \tag{14}$$

We consider the following candidate observer system:

$$\dot{z} = A(u)z - rC^*(Cz - y)\,, \tag{15}$$

in which $r > 0$ is a parameter. The estimation error is $\varepsilon = z - x$:

$$\dot{\varepsilon} = (A(u) - rC^*C)\varepsilon\,.$$

Then it is not so hard to show that, if $u\colon [0, \infty[\to U$ is a "persistent excitation" of Σ in some sense (for instance in the sense of Subsect. "Kalman's Observer for State-Linear Systems", then we have:

$$\lim_{t\to+\infty} \|\varepsilon(t)\| = 0\,.$$

As a consequence, the systems with compact group G of diffeomorphisms (or with G semidirect product of compact group by vector space), admit also approximate observers, using the results of Subsect. "State-Linear Skew-Adjoint Realization".

It turns out that this method can be extended in a reasonable way to systems with (infinite dimensional) skew-adjoint state-linear realization. In particular, it is possible to construct approximate observers for all (complete minimal) systems with finite dimensional Lie algebra.

Consider a skew adjoint realization from Subsect. "State-Linear Skew-Adjoint Realization":

$$\begin{cases} \dot{\Psi} = A(u)\Psi\,, \\ y = \langle \Psi, \xi \rangle \end{cases}$$

on the (separable) Hilbert space $\mathcal{H}$. Then, the candidate observer device is:

$$\dot{\Lambda} = A(u)\Lambda - r\xi(\langle \Lambda, \xi \rangle - y(t))\,. \tag{16}$$

In fact, the persistency assumption cannot be of the same type as in the finite dimensional case. The reason is that the Gramm observability matrix $G_{u,T}$ is a compact operator in that case. Hence it cannot satisfy an inequality of the type $G_{u,T} \geq \alpha Id_{\mathbb{H}}$ since $\mathcal{H}$ is infinite dimensional.

Hence, the definition of a persistent excitation has to be replaced by one of the following type: there is a time $T > 0$ and a real sequence θ_n, $\theta_n \to +\infty$, with $\theta_{n+1} - \theta_n$ bounded, such that the translated inputs $u_{\theta_n}\colon [0, T] \to \mathbb{R}^l$ converge to u^* in the weak–* topology of $L^\infty_{[0,T],\mathbb{R}^l}$ (which topology is precompact over bounded sets) and u^* is a universal input for Σ on $[0, T]$.

This means also that a certain level of observability is preserved, on regularly spaced time intervals, while the time increases.

In that case, of course the result is weaker than in the finite dimensional case. We have only:

$$\text{weak–}\lim_{t\to+\infty} \varepsilon(t) = 0\,.$$

Future Directions

For observability and synthesis of observers, besides the improvement of the current methods (including sliding modes, high gain, ...) several directions have to be investigated more deeply, namely infinite dimensional systems, delay and hybrid systems.

For realization theory, and as a consequence identification theory, almost no "**practical result**" is known in the nonlinear context. However, we think interesting and consistent developments are possible, even starting from the apparently abstract theory outlined there. This is clearly the challenge for the future.

Bibliography

Primary Literature

1. Bartosiewicz Z (1995) Local observability of nonlinear systems. Syst Control Lett 25(4,1):295–298
2. Bartosiewicz Z (1998) Real analytic geometry and local observability. Proc Sympos Pure Math. Am Math Soc 64:65–72
3. Barut AO, Raczca R (1986) Theory of Group Representations and Applications. World Scientific, Singapore
4. Crouch PE (1981) Dynamical realizations of finite volterra series. SIAM J Control Opt 19:177–202
5. Dixmier J (1964) Les $\mathbb{C}^*$ algébres et leurs représentations. Cahiers Scientifiques, fasc. XXIX. Gauthier-Villars, Paris
6. Fliess M (1981) Fonctionnelles causales non linéaires et indéterminées non commutatives. Bull Soc Math France 109: 3–40
7. Fliess M (1983) Réalisation locale des systèmes non linéaires, algèbres de Lie filtrés transitives et séries génératrices non commutatives. Inventiones Mathematicae 71:521–537. Springer
8. Fliess M (1973) Sur la realization des systemes dynamiques bilineaires. CR Acad Sc Paris A 277:923–926
9. Fliess M (1980) Nonlinear realization theory and abstract transitive Lie algebras. Bull Amer Math Soc (NS) 2:444–446
10. Fliess M, Kupka I (1983) A finiteness criterion for nonlinear input-output differential systems. SIAM J Control Optim 21:721–728
11. Grizzle JW, Moraal PE (1995) Observer design for nonlinear systems with discrete-timemeasurements. IEEE Trans Autom Control 40(3):395–404
12. Hermann R, Krener AJ (1977) Nonlinear controllability and observability. IEEE Trans Autom Control 22(5):728–740
13. Jacob G, Hespel C (1991) Approximation of nonlinear dynamic systems by rational series. Theor Comp Sci Arch 79(1):151–162
14. Jakubczyk B (1986) Local realizations of nonlinear causal operators. SIAM J Control Optim 24(2):230–242
15. Jakubczyk B (1984) Réalisations locales des opérateurs causals non linéaires. Comptes rendus des séances de l'Académie des sciences. Série A 299(15):787–793
16. Jouan P (2003) Immersion of nonlinear systems into linear systems modulo output injection. SIAM J Control Optim 41(6):1756–1778
17. Kalman RE (1963) Mathematical description of linear dynamical systems. SIAM J Control Optim 1(2):152–192
18. Kalman RE, Bucy RS (1961) New results in linear filtering and prediction theory. J Basic Eng Trans ASME 83:95–108
19. Krener AJ, Isidori A (1983) Linearization by output injection and nonlinear observers. Syst Control Lett 3(1):47–52
20. Krener AJ, Respondek W (1985) Nonlinear observers with linearizable error dynamics. SIAM J Control Optim 23:197–216
21. Luenberger D (1966) Observers for multivariable systems. IEEE Trans Autom Control 11(2):190–197
22. Luenberger D (1971) An introduction to observers. IEEE Trans Autom Control 16(6):596–602
23. Sontag E (1979) On the observability of polynomial systems, I: Finite-time problems. SIAM J Control Optim 17(1):139–151
24. Sussmann HJ (1973) Orbits of families of vector fields and integrability of distributions. Trans Amer Math Soc 180:171–188
25. Sussmann HJ (1975) A generalization of the closed subgroup theorem to quotients of arbitrary manifolds. J Diff Geom 10:151–166
26. Sussmann HJ (1974) On quotients of manifolds: a generalization of the closed subgroup theorem. Bull Amer Math Soc 80:573–575
27. Sussmann HJ (1976) Existence and uniqueness of minimal realizations of nonlinear systems. J Theor Comput Syst 10(1):263–284, 371–393
28. Sussmann HJ (1978) Single-input observability of continuous-time systems. J Theor Comput Syst 12(1):371–393
29. Yamamoto Y (1981) Realization theory of infinite-dimensional linear systems, Part I. Math Syst Theor 15:55–77

Books and Reviews

Isidori A (1995) Nonlinear Control Systems. Springer Communications and Control Engineering Series. Springer, New York

Kalman RE, Falb PL, Arbib MA (1969) Topics in Mathematical System Theory. Mc Graw Hill, New York

Khalil H (2002) Nonlinear systems, 2nd edn. Prentice Hall, Upper Slade River

Nijmeijer H, van der Schaft A (1990) Nonlinear Dynamical Control Systems. Springer

Rugh WJ (1981) Nonlinear System Theory, The Volterra/Wiener Approach. The Johns Hopkins University Press, Baltimore

Sontag ED (1998) Mathematical Control Theory: Deterministic Finite Dimensional Systems, 2nd edn. Springer, New York

Utkin VI (1992) Sliding Modes in Control and Optimization. Springer

Opinion Dynamics and Sociophysics

DIETRICH STAUFFER
Institute for Theoretical Physics, Cologne University, Cologne, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Schelling Model
Opinion Dynamics
Languages, Hierarchies and Football
Future Directions
Bibliography

Glossary

Cluster Clusters are sets of neighboring sites of the same type.

Ising model Each site carries a magnetic dipole which points up or down; neighboring dipoles "want" to be parallel.

Opinion dynamics How do people change opinions? Simulations usually ignore all details of the brain and represent the opinion by one or several numbers which can be changed due to contact with others.

Schelling model People belonging to different groups may produce segregated neighborhoods just by their personal preferences, not by outside force.

Sociophysics Application of methods from (mostly statistical) physics to human relations can be traced centuries backwards.

Definition of the Subject

Sociophysics is the study of social questions by physicists using their physics methods. In contrast to biophysics, it is a field which is not yet very well established. Opinion dynamics is one of the most widespread topics of sociophysics.

Introduction

The application of concepts from the natural sciences to social sciences, partly to be reviewed here, is at least 25 centuries old. Then the Greek philosopher Empedokles stated (according to J. Mimkes) that humans are like liquids: Some mix easily like wine and water, and others like oil and water refuse to mix. We start with the Schelling model of 1971, which implemented this idea, and its criticism (see ► Social Processes, Simulation Models of). Then we will review opinion dynamics in large populations, summarizing only shortly other aspects like self-organization of hierarchies or competition between human languages.

Humans do not like to be treated like a number, and indeed the human brain is much more complex than a binary variable (called "spin" by physicists) which is either $+1$ or -1. We do not deal here with the psychological processes of an individual but with mass psychology, and this author learned half a century ago in school that mass psychology is different from individual psychology: The law of large numbers averages out over individual fluctuations and makes general trends more clearly visible. Thus what we call today statistical physics plays a useful rule, and social scientists [15,36] have applied it, without knowing then that they dealt with an Ising model of ferromagnets.

The astronomer Halley, best known through his comet, tried to establish mortality tables already three centuries ago. Of course, the time of death of one given individual is usually difficult to predict but averaged over millions of people the statistical offices of many countries prepare regularly life tables which tell us how probable it is for a newborn child to live up to x years, provided there are no changes of the mortalities in the coming decades. Insurance for automobiles is another example: We do not want to produce accidents, but we know that they happen, and take precautions against their financial consequences. Thus the whole insurance industry is based on treating humans like numbers, ignoring their individuality.

Finally, human opinions are often fluctuating and ill-defined, but nevertheless in elections people cast one choice, out of a limited number of choices. And election results belong to those social data for which we have lots of accurate numbers, based on large populations.

Thus it is not at all the merit (or ignorance) of physicists which treats humans like numbers; this method has a very long tradition and is an indispensable part of modern life.

Schelling Model

Ising Simulations

Following (but not citing) Empedokles, the later economics Nobel laureate Schelling [36] asked whether the racial segregation in American cities can emerge from intrinsic behavior of the individual people, instead of or in addition to extrinsic reasons like discrimination, rent differences, etc. In particular, can "black" ghettos in the predominantly "white" USA arise just because people prefer to have neighbors of their own group over neighbors from the other group? In many other countries we find many other types of residential segregation, based on religion, ethnicity, ... In physics, such a process is easily simulated through the two-dimensional Ising model, as shown in Figs. 1 and 2.

In this Ising model, each site on a square lattice carries a variable $S_i = \pm 1$, and each pair $\langle i, k \rangle$ of nearest neighbors produces an "energy" $-JS_iS_k$ with some proportionality constant J. The total energy E (= total unhap-

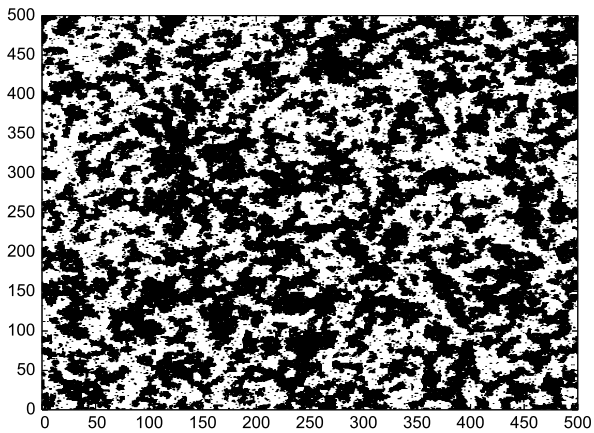

Opinion Dynamics and Sociophysics, Figure 1
Ising model after 20 Glauber kinetic steps per site on a 500 × 500 square lattice at $k_BT/J = 2$. We start from a random distribution of equally many black and white sites

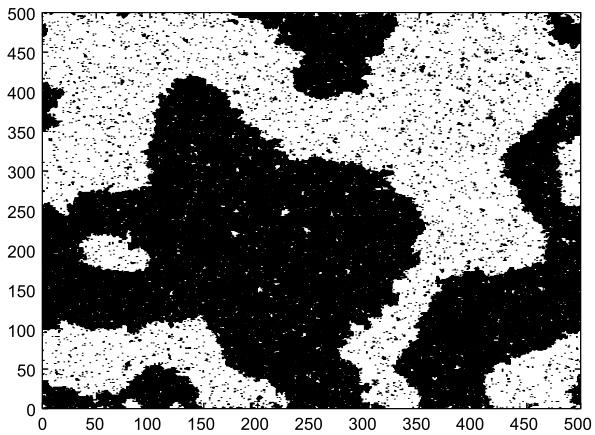

Opinion Dynamics and Sociophysics, Figure 2
As Fig. 1 but after 2000 instead of 20 iterations

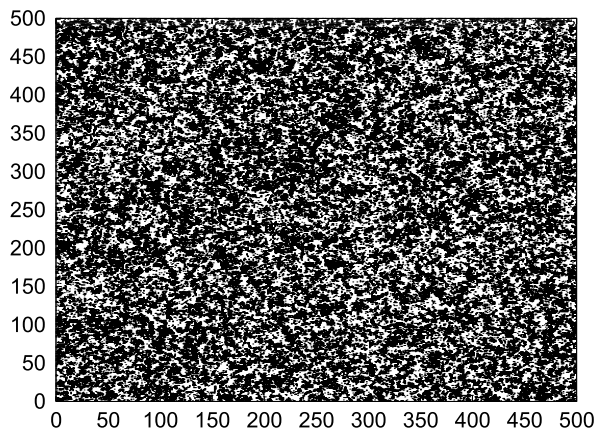

Opinion Dynamics and Sociophysics, Figure 3
As Fig. 2 but at $k_B T/J = 3$ instead of 2. Only small clusters and no large domains are formed. After 200 and 20,000 iterations the pictures look similar to this one made after 2000 iterations

piness) is the sum of this pair energy over all neighbor pairs of the lattice. In physics, different distributions of the "spins" S_i are realized with a probability proportional to $\exp(-E/k_B T)$ where T is the absolute temperature and k_B the Boltzmann constant. There is no need to worry about values for T, k_B, J since the only relevant quantity is the ratio $k_B T/J$, taken as 2 in these pictures. The "Glauber" kinetics is simulated on the computer by flipping a spin if and only if a random number between 0 and 1 is smaller than the probability $\exp(-\Delta E/k_B T)/[1+\exp(-\Delta E/k_B T)]$. The Fortran program of Algorithm 1 contains less than 40 lines and takes a few seconds.

Such models and programs are taught in courses on computational or theoretical physics all over the world; the model was published in 1925. If in the above flipping probability the denominator is omitted one gets the Metropolis kinetics. If instead of flipping one spin, we exchange two opposite spins, we get the Kawasaki dynamics. For Glauber or Metropolis, after very long times (measured by the number of sweeps through the lattice) one of the two possibilities dominates at the end, if T is not larger than the critical temperature T_c, with $2J/k_B T_c = \ln(1+\sqrt{2}) \simeq 0.88$ known since 1940. For Kawasaki dynamics the fraction of black sites remains constant, and we get two large domains. For higher temperatures above T_c only small clusters and no large domains are formed, Fig. 3.

In this Ising model, two neighboring spins have due to their interaction $-JS_i S_k$ a higher probability to belong to the same group than to belong to the two different groups. If the difference between these two probabilities is large enough, $T < T_c$, domain sizes can grow to infinity in an infinite lattice, Figs. 1 and 2, while only small clusters are formed in Fig. 3 for smaller differences in the probabilities, $T > T_c$. That these probabilities, controlled through $-J/k_B T$, lead to these different regimes, separated by a sharp phase transition at $T = T_c$, is not obvious from the definition of the interaction $JS_i S_k$, took physicists many years to find, and is typical of complex systems.

The social meaning of temperature T is not what we hear in the weather reports but an overall approximation for all the more or less random events which influence our decisions but are not explicitly included in the model. For residential segregation the model only counts how many neighbors of which group one has. But not all people of one group are alike, housing in different parts of a city costs different amounts of money, some parts are more beautiful then others, and job hunting may force us into a temporary residence of a new city which does not conform to our wishes. In this way, a positive temperature allows for rare moves which increase the energy, i. e. we move to a new residence where the neighborhood composition along makes us less happy. At zero temperature, the Ising model does not properly order into one or two "infinite" domains.

Schelling's Version and Later Improvements

Schelling [36] avoided probabilistic rules and thus counted neighbors $S_k = \pm 1$ at zero temperature. Then it does not matter if all neighbors or only a majority of them belong to the own group. Thus people are defined as happy if at least half of the neighbors belong to the own group, and as unhappy otherwise (i. e. if the majority belong to the opposite group). Unhappy people move to the nearest place

```
      parameter(L=500,Lmax=(L+2)*L)
      dimension is(Lmax),iex(9)
      byte is
      data t,max,ibm/2.00,2000,1/
      print *, L,max,ibm,t
      Lp1=L+1
      L2pL=L*L+L
      do 1 i=1,Lmax
         is(i)=-1
         ibm=ibm*16807
1         if(ibm.gt.0) is(i)=1
      do 2 ie=1,9
         ex=exp(-2*(ie-5)/t)
2         iex(ie)=(2.0*ex/(1.0+ex) - 1.0)*2147483647
      ibm=2*ibm+1
      do 3 mc=1,max
         do 4 i=Lp1,L2pL
            ie=5+is(i)*(is(i-1)+is(i+1)+is(i-L)+is(i+L))
            ibm=ibm*16807
4           if(ibm.lt.iex(ie)) is(i)=-is(i)
         mag=0
         do 6 i=Lp1,L2pL
6            mag=mag+is(i)
3        if(mc.eq.(mc/100000)*100000) print *, mc,mag
      do 5 i=Lp1,L2pL
         if(is(i).ne.1) goto 5
         iy = (i-1)/L
         ix=i-L*iy
         print *, ix, iy+1
5     continue
      stop
      end
```

Opinion Dynamics and Sociophysics, Algorithm 1
Simple Fortran program to produce pictures like Figs. 1 to 3

where they are happy. Since Schelling moved only one person (or family) at a time, and made no exchange of two people simultaneously as in Kawasaki kinetics, he introduced a large fraction of empty residences. Thus at each step, one unhappy person or family moves into the closest vacancy where life would be happy.

This model, and also many variants [16,36], fail to give large domains; only small clusters are seen. In reality, Harlem in Manhattan (New York), is not a cluster of a few houses but extends over many square kilometers. Thus the original version does not give the desired results. Large domains are formed if people also change residences if this brings no improvement [46] (hardly a realistic assumption), or at a finite temperature [41]. The latter paper also gives some alternatives to the Schelling model which also allow for large domains, and a simple example of a finite cluster where everybody is "happy" and which therefore never grows or dissolves on its own "will". More quantitative analyses of domain growth are given by [13, 27,39].

Much earlier and simpler is the zero-temperature version of Jones [15] who at each iteration removes a random fraction of the people and fills the vacancies with people who are there happy in the Schelling sense. This randomness, just like the finite temperature, leads to large domains as desired. Neither physicists nor social scientists have taken much note of [15]. The history of the Schelling model is an example how the lack of communication between disciplines has hampered progress in research, even very recently [41,46]. Only computational statistical physicists know everything. ([15] also mention a probabilistic version closer to the Ising model.)

For finite temperatures, [41] follow the above Glauber dynamics, but instead of an energy E uses a variable which is 0 or 1 depending on the happiness of the residents. Moving from one place to the other then depends exponentially

on the ratio of this variable to $k_B T$, instead of on the ratio $\Delta E/k_B T$. Many variants are possible, e. g. in the treatment of neutral cases [30,46] where the number of neighbors of both groups is exactly the same.

But we are on safer grounds and can use decades of physics research if we use the normal Ising model, or its generalization to Q different groups, the Q-state Potts model. Then Refs. [28,37] implemented a suggestion of Weidlich [47] that people slowly learn to live together with neighbors from the other group. Thus T not only takes into account the various accidents from outside the model, but also measures the tolerance: The higher T is the more are people willing to live in neighborhoods of the other group. In the limit $T = \infty$ the neighbors would not matter at all, for intermediate T, Fig. 3 showed small clusters but no large domains, and for low T the domains grow to infinite sizes on an infinite lattice. The learning suggested by Weidlich thus means that this parameter T (= temperature or tolerance) no longer is kept constant but slowly increases.

For an Ising model, [28] showed how an initial large domain dissolves if the temperature is slowly increased from below to above T_c. More realistically, for five (instead of only two) different groups in a modified five-state Potts model, [37] increased T from low to high values and showed that with a slow increase one has appreciable domain formation during intermediate times, while with a fast increase this segregation is mostly avoided, Fig. 4.

Instead of imposing a fixed temperature or tolerance T to everybody, one can also let it self-organize according to

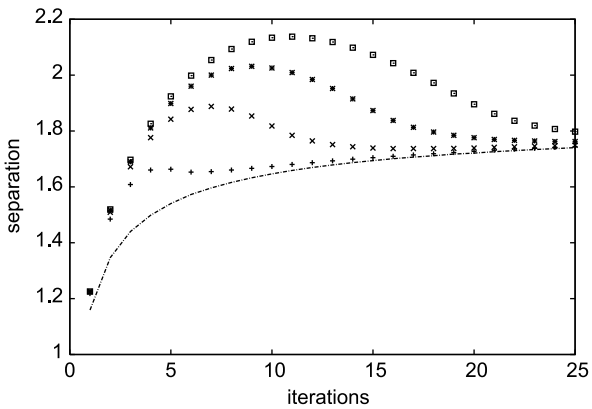

Opinion Dynamics and Sociophysics, Figure 4
Amount of neighbors of the same type in a Potts model of five groups, normalized to unity for the initial random distribution. The temperature or tolerance increases from low to high values, slowly in the *top curves*, and fast in the *lower curves*; the latter mostly avoid the segregation into different group. (The *lowest line* holds for a constant high temperature.). From [37]

the neighborhood [30]. Or one may introduce two different T, one for tolerance against people of the other group, and the other for the random noise from events outside the model, like marriages, job losses, deaths [32].

Poor people cannot afford expensive housing. If we assume one of the two groups to be poor and the other to be rich, and if we assume that each residence is randomly either expensive or cheap, then we have a random-field Ising model [43]. This "field" gives the probability for the poor to select only cheap housing and for the rich to live in expensive residences. For intermediate lattice sizes and intermediate times, the field prevents the growth of infinite domains, and the clusters are the smaller the larger the field is [43].

Opinion Dynamics

The following section describes several rules for simulated people to change opinions; each of these rules is applied again and again to these agents until some stationary or static state has been achieved.

Ising Model

Also for human opinions, one could use the Ising model of the previous section [9,19]; see also [47]. People can vote for or against the government or a new constitution, for one of two presidential candidates, or (using generalized models) for one out of $Q > 2$ different parties. Their neighbors on a lattice influence them in their vote, and in addition mass media may influence everybody in one direction. The latter effect can be modeled through an external "magnetic" field, Eq. (1b) in "Phase transitions ..." by this author in this encyclopedia. No motion of people needs to be taken into account, and the complications of Kawasaki kinetics (exchange of two people with opposing opinions) are not needed. Thus the Glauber program of the previous section still can be used, and we only refer here to the old generalization into the social impact model [24,25] and to a recent financial application [50].

Voter Model

Also quite old is the voter model [26]: Each person chooses between two opinions, by taking over the one of a randomly selected neighbor. One may rewrite this rule as stating that each person selects the opinion of the neighborhood, with a probability proportional to the number n of neighbors having that opinion. Thus in contrast to the Ising model where the probabilities depend exponentially on n, now they depend linearly on n. A final equilibrium (absorbing fixed point) is reached if everybody shares the

same opinion. The deviations from that final state can be measured by the magnetization (difference between the numbers for the two opinions) or energy (average number of neighbors having the opposite opinion). The time needed to reach the consensus increases with a power of the lattice size, and the exponent depends on the dimensionality. A nice and short review of the voter model, also on various networks, is given by the Majorca group [35].

Axelrod Model

Axelrod [1] wondered how different opinions or cultures may coexists even if people tend to become more alike in their beliefs. Looking at the above Ising Figs. 1, 2, 3, we see that due to finite time and/or finite temperature such coexistence of two opposing opinions is possible. But Axelrod generalized it not only to $Q > 2$ different possible opinions as in the Potts model of the above "Schelling Model" section, but also to F different questions. People may have one set of opinions on which political party they want to vote for, another set about what is the best football team, a third about recent cinema films, etc. This allows for Q^F different opinion sets on all F questions ("features"). Of course, one could generalize this model to the case where the number Q_f of possible choices is different for the different features f, allowing then for $\Pi_{f=1}^{F} Q_f$ instead of simply Q^F different sets of opinions.

Another aspect of the model takes into account that people prefer to talk to, or to make political coalitions with, others with whom they share many opinions. Thus the probability of one person to take over the different opinion of a neighbor is proportional to the number of features on which their opinions already agree. In the next subsection we will use a similar concept under the name of bounded confidence.

Whether a total consensus ("globalization") is reached or multiculturality persists depends on parameters: Small Q lead to consensus. Again, the Majorca group [35] reviewed the many follow-up papers on this Axelrod model.

Sznajd, Krause–Hegselmann and Deffuant Models

Much of the opinion dynamics research since 2000 centered on three different models S, KH and D, originally invented independently around that year: Sznajd [44] (S), Krause–Hegselmann [23] (KH) and Deffuant et al. [14] (D). They were also called missionaries, opportunists and negotiators by some computational physicists [42].

The S model is closest to the earlier models since it allows for Q discrete opinions, while KH and D use real opinions, e. g. between zero and one. S happens on a lattice or network while for KH and D everybody may interact with everybody. In the most widespread S version a pair of neighboring sites on a square lattice convinces its six neighbors of its opinion, if and only if the two opinions of the pair agree [29]; governments and parties usually lose support if their internal opinion differences make it to the headlines. For KH, the new opinion of a person is the arithmetic average over the opinion of the whole population. For D, each person selects randomly another person and then both move in their opinion towards each other by an amount proportional to their opinion difference.

In all three cases, "bounded confidence" applies: The KH agents average only over those people who differ from their own opinion by less than ϵ, and the D agents only select negotiation partners differing by less than ϵ from their own opinions. In both models $0 < \epsilon < 1$ is a fixed parameter. For S agents with $Q = 2$ such a rule makes no difference, but for $Q > 2$ one can modify the convincing rule such that only neighbors differing by at most ± 1 from the pair opinion adopt the pair opinion. Thus $1/Q$ for S plays the role of ϵ for D and KH. A rule similar to this bounded confidence was mentioned above for the Axelrod model [1].

In spite of the differences in their definitions, the results are quite similar for S, KH and D. For large ϵ or $Q \leq 3$ a complete consensus is usually reached; for small ϵ or $Q \geq 4$ different opinions may coexist forever. In addition to computer simulations, also analytical calculations were made [2,40] which agree with many aspects of the simulations. More results, also for opinions on more than one feature and agents sitting on scale-free networks [8], are summarized in [42].

One particular application is shown in Fig. 5: Various Brazilian election results for candidates in city councils

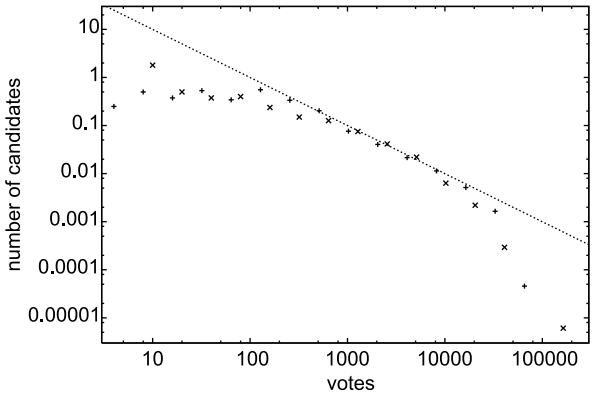

Opinion Dynamics and Sociophysics, Figure 5
Brazilian elections (x) and simulations of Sznajd model on Barabási–Albert networks (+); from [4]

showed great similarity if the number of candidates getting a given number v of votes is plotted against this v. Putting the S model with $Q = 1000$ candidates onto a scale-free network instead of a square lattice, excellent agreement of simulation and reality was found after the numbers were scaled by suitable factors. Also Indian elections were simulated this way [20], while proportional elections are different [11]. It would be nice to apply other opinion dynamics models to the same election problem. As usual in statistical physics, these studies can predict and simulate the shape of the distributions but not the winner in a specific election, just as we can predict the pressure of the air molecules around us but not where which molecule will be one minute from now.

Galam Conservatism

Galam has published since many years theoretical models which may explain why reforms are very difficult and why a minority can stay in power. Usually these models are solvable analytically and assume that the population is divided into small groups of people which to the outside are represented by one person who follows the majority wish of the group. Several of these representatives form a supergroup, and this supergroup again decides according to the majority of the representatives in it. In this way an "infinite" hierarchy of people, groups, supergroups etc can be built. In the case of equally many voting for one choice as for the opposite choice, within one unit, that unit votes for the status quo. Starting with everybody having opinion -1, a very large majority of people must switch to opinion $+1$ before the top of the hierarchy finally also changes opinion [17]. We refer to [42] for a summary of more recent Galam papers, and to [18] for a more complete review.

Languages, Hierarchies and Football

Language Competition

Darwinian survival of the fittest is established biology, but similar concepts can be applied to human languages, bridging the gap to opinion dynamics. There are now thousands of different languages, and their "size" is the number of native speakers of that language. The size distribution extends from 1 (on the verge of extinction) to 10^9 (Mandarin Chinese). The grammar of a language [22] can be characterized by F features each of which can have Q different values, just as in Axelrod's model explained above. Features can change spontaneously or be taken over from a (neighboring) language; speakers of a small language give it up and learn a widespread language (as done with physics research publications since 1945); people migrate to other places and bring their language with them. All these processes can lead to the extinction of existing languages and the creation of new ones (by the branching of one language into several sub-languages.) The present language size distribution is roughly log-normal, with an enhancement at small sizes [21]. Similar languages form families, and the size distribution of families is a power law at intermediate sizes [49] (where the size is now the number of different languages belonging to that family).

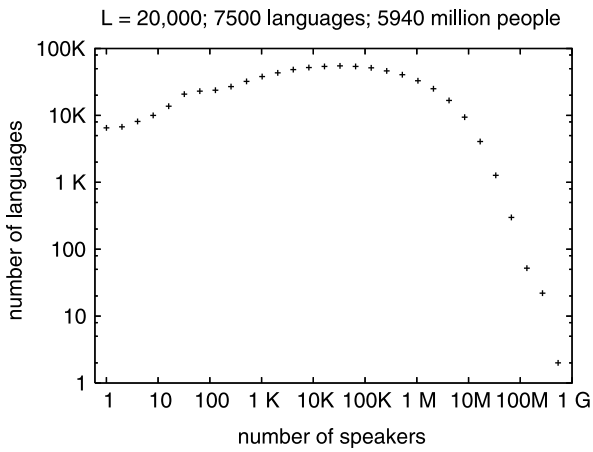

Opinion Dynamics and Sociophysics, Figure 6
Simulated language size distribution on a 20,000 × 20,000 square lattice using a modified Viviane model [33]

Various computer simulations of this language competition have been made, mostly since 2003 and reviewed recently [38]; see also [10,12]. We only mention Fig. 6 from a modified Viviane model, which agrees well with the real language size distribution. For language families, the empirical statistics is worse [49] and one model also works well [38]. Good distributions were also obtained in a model which avoids dealing with individual speakers [45].

Self-Organization of Social Hierarchies

The elites of all countries and all times always had excellent reasons why they should be on top and others on the bottom. This holds even when the United Nations criticize the school system as violating human rights. In contrast, the Bonabeau model [7] explains social hierarchies as purely accidental, without any merit. People are put on a lattice, occupying a fraction p of all lattice sites and having an initial score of zero. Then they move randomly to neighboring sites, and whenever one person wants to move into the site occupied by another person, a fight erupts. The winner takes the contested site, the loser moves into (or stays at) the other site. Also, the winner adds one point and the

loser subtracts one point in its score, and in the future the agents with a positive score have a higher probability to win, those with a negative score have a lower probability to win. Slowly the history is forgotten, by reducing the score at each time step by, say, ten percent.

With some suitable feedback between the distribution of scores and the probability to win, a phase transition was simulated such that for p above some critical concentration, the standard deviation in the scores becomes positive for long times and large populations. For $p < p_c$ it fluctuates near zero, which means that everybody has close to a 50 percent chance to win. So, just by accident at a high population density some people rise to the top, and others fall to the bottom. However, the people on top (bottom) are not always the same; only the differences between top and bottom, not the people, remain the same. [3,31,48] are some of the more recent references in this field.

Football

Football (= soccer) is the world's most popular spectator sport, though in the author's city it is more a frustration. Randomness surely plays a role and makes it attractive. Can we explain all results just by chance, in the spirit of Bonabeau hierarchies? Assuming a constant probability to make a goal within one minute, the distribution of goals and victories is more narrow than in reality. If instead we assume that this probability varies from team to team, still no good agreement is found. Good agreement with reality is obtained only if correlations are taken into account [6], in the sense that a goal makes the scoring team happy, shocks the opposing team, and thus with an enhanced probability leads to another goal for the scoring team. Thus if we lose it is not just bad luck; it is also the referee's fault.

Future Directions

The Schelling model of Sect. "Schelling Model" is not the only case of missed opportunities because of a lack of cooperation between social sciences on the one side and physics, mathematics or computer science on the other side. The two books [5,42] were written without the authors of one book knowing of the preparation of the other book. One group of authors works in physics departments; none of the other group lists physics as institutional address. Nevertheless the two books show strong overlap in fields and methods covered, but little overlap in the literature cited. More interdisciplinary cooperation would help.

Bibliography

1. Axelrod R (1997) J Confl Resolut 41:203
2. Ben-Naim E, Krapivsky P, Redner S (2003) Physica D 183:190
3. Ben-Naim E, Vazquez F, Redner S (2006) Eur Phys J B 49:531
4. Bernardes AT, Stauffer D, Kertész J (2002) Eur Phys J B 25:123
5. Billari FC, Fent T, Prskawetz A, Scheffran J (2006) Agent-based computational modelling. Physica, Heidelberg
6. Bittner E, Nussbaumer A, Janke W, Weigel M (2007) Europhys Lett 78:58002 (2007)
7. Bonabeau E, Theraulaz G, Deneubourg JL (1995) Physica A 217:373
8. Caldarelli G (2007) Scale-free networks: complex webs in nature and technology. Oxford University Press, Oxford
9. Callen E, Shapero D (1974) Phys Today 27:23
10. Cangelosi A, Parisi D (eds) (2002) Simulating the evolution of language. Springer, New York
11. Castellano C, Fortunato S, Loreto V (2008) Statistical physics of social dynamics. arXiv:0710:3256, Rev Mod Phys (preprint)
12. Culicover P, Nowak A (2003) Dynamical grammar. Oxford University Press, Oxford
13. Dall'Asta L, Castellano C, Marsili M (2008) J Stat Mech L07002
14. Deffuant G, Amblard F, Weisbuch G, Faure T (2002) J Artif Soc Soc Simul 5(4):1
15. Dethlefsen E, Moody C (1982) Byte 7:178; Jones FL (1985) Aust NZ J Sociol 21:431
16. Fossett M (2006) J Math Sociol 30:185
17. Galam S (1990) J Stat Phys 61:943
18. Galam S (2008) Int J Mod Phys C 19:409
19. Galam S, Gefen Y, Shapir Y (1982) J Math Sociol 9:1
20. González MC, Sousa AO, Herrmann HJ (2004) Int J Mod Phys C 15:45
21. Grimes BF (2000) Ethnologue: Languages of the world, 14th edn. Summer Institute of Linguistics, Dallas. Available at www.ethnologue.org
22. Haspelmath M, Dryer M, Gil D, Comrie C (eds) (2005) The world atlas of language structures. Oxford University Press, Oxford
23. Hegselmann R, Krause U (2002) J Artif Soc Soc Simul 5(3):2
24. Hołyst JA, Kacperski K, Schweitzer F (2001) Annual Reviews of Computational Physics IX. World Scientific, Singapore, p 275
25. Latané B (1981) Am Psychol 36:343
26. Liggett TM (1985) Interacting particle systems. Springer, New York
27. Lim M, Metzler R, Bar-Yam Y (2007) Science 317:1540
28. Meyer-Ortmanns H (2003) Int J Mod Phys C 14:311
29. Milgram S, Bickman L, Berkowitz L (1969) J Pers Soc Psych 13:79
30. Müller K, Schulze C, Stauffer D (2008) Int J Mod Phys C 19:385
31. Naumis GG, del Castillo-Mussot M, Perez LA, Vazquez GJ (2006) Physica A 369:789
32. Ódor G (2008) Int J Mod Phys C 19:393
33. de Oliveira PMC, Stauffer D, Lima FWS, Sousa AO, Schulze C, Moss de Oliveira S (2007) Physica A 376:609
34. de Oliveira PMC, Stauffer D, Wichmann S, Moss de Oliveira S (2008) J Linguist 44:659
35. San Miguel M, Eguíluz VM, Toral R (2005) Comput Sci Engin 7:67
36. Schelling TC (1971) J Math Sociol 1:143
37. Schulze C (2005) Int J Mod Phys C 16:351
38. Schulze S, Stauffer D, Wichmann S (2007) Comm Comput Phys 3:271
39. Singh A, Vainchtein D, Weiss H. Schelling's segregation model: Parameters, scaling, and aggregation. Preprint
40. Slanina F, Lavička H (2003) Eur Phys J B 35:279
41. Stauffer D, Solomon S (2007) Eur Phys J B 57:473

42. Stauffer D, Moss de Oliveira S, de Oliveira PMC, Sá Martins JS (2006) Biology, sociology, geology by computational physicists. Elsevier, Amsterdam
43. Sumour MA, El-Astal AH, Radwan MA, Shabat MM (2008) Int J Mod Phys C 19:637; see also Emboloni F (preprint)
44. Sznajd-Weron K, Sznajd J (2000) Int J Mod Phys C 11:1157
45. Tuncay Ç (2007) Int J Mod Phys C 18:1641
46. Vinkovic D, Kirman A (2006) Proc Natl Acad Sci USA 103:19261
47. Weidlich W (2000) Sociodynamics; A systematic approach to mathematical modelling in the social sciences. Harwood Academic Publishers, 2006 reprint, Dover, Mineola
48. Weisbuch G, Stauffer D (2007) Physica A 384:542
49. Wichmann S (2005) J Linguist 41:117
50. Zhou WX, Sornette D (2007) Eur Phys J B 55:175

Optical Computing

Thomas J. Naughton[2,3], Damien Woods[1,4]
[1] Department of Computer Science, University College Cork, Cork, Ireland
[2] Department of Computer Science, National University of Ireland, Maynooth County Kildare, Ireland
[3] Oulu Southern Institute, University of Oulu, RFMedia Laboratory, Ylivieska, Finland
[4] Department of Computer Sience and Artificial Intelligence, University if Seville, Seville, Spain

Article Outline

Glossary
Definition of the Subject
Introduction
History
Selected Elements of Optical Computing Systems
Continuous Space Machine (CSM)
Example CSM Datastructures and Algorithms
C_2-CSM
Optical Computing and Computational Complexity
Future Directions
Acknowledgments
Bibliography

Glossary

Coherent light Light of a narrow band of wavelengths (temporally coherent), and a light beam whose phase is approximately constant over its cross sectional area (spatial coherence). For example, coherent light can be produced by a laser.

Incoherent light Light which is not spatially coherent and not temporally coherent. For example, incoherent light is produced by a conventional light bulb.

Source A device for generating light.

Spatial light modulator (SLM) A device that imposes some form of spatially-varying modulation on a beam of light. An SLM may modulate the intensity, phase, or both, of the light.

Detector A device for sensing light.

Continuous space machine (CSM) A general optical model of computation that is defined in Sect. "Continuous Space Machine (CSM)".

Parallel computation thesis This thesis states that parallel time corresponds, within a polynomial, to sequential space, for reasonable parallel and sequential machines [29,52,74,98,126].

P, NP, PSPACE, NC Complexity classes, these classes are respectively defined as the set of problems solvable on polynomial time deterministic Turing machines; polynomial time nondeterministic Turing machines; polynomial space Turing machines; and parallel computers that use polylogarithmic time and polynomial hardware [97].

Definition of the Subject

An optical computer is a physical information processing device that uses photons to transport data from one memory location to another, and processes the data while it is in this form. In contrast, a conventional digital electronic computer uses electric fields (traveling along conductive paths) for this task. The optical data paths in an optical computer are effected by refraction (such as the action of a lens) or reflection (such as the action of a mirror). A principal advantage of an optical data path over an electrical data path is that optical data paths can intersect and even completely overlap without corrupting the data in either path. Optical computers make use of this property to efficiently transform the optically-encoded data from one representation to another, for example, to shuffle or reverse the order of an array of parallel paths, or to convolve the data in several arrays of parallel paths. Other advantages of optical computers include inherent parallelism and the ability to encode a two-dimensional spatial function in the cross-section of a single beam of light, higher bandwidths (in contrast to the free transmission of photons, electric fields generate noise in parallel conductors as they are pushed down their conductor), lower energy consumption (an argument deriving from the fact that optical computers in principle generate very little heat), easier circuit design, and lower latency in comparison to electrical transmission.

However, the property of non-interference of intersecting data paths means that it is not straightforward to

effect a switch or branch instruction in optics since in a vacuum the presence or absence of light in one data path cannot affect another path. In order to perform a conventional computation (e. g. solve a decision problem) optical computers invariably need to be equipped with an electronic interface, which would sense the presence or absence of light at some stage of the computation, and set the optical computer on a new course. Although this limitation has been addressed with varying levels of success through the development of nonlinear optical materials for all-optical switching and storage devices, it is generally accepted that if optical computers become mainstream, it will be through a symbiotic relationship with their extremely flexible digital electronic counterparts. Furthermore, currently there is no convincing alternative to using digital electronics for optical computer data input devices (liquid-crystal display panels, for example) and data output devices (digital cameras, for example).

Optical computing is an inherently multidisciplinary subject whose study routinely involves a spectrum of expertise that threads optical physics, materials science, optical engineering, electrical engineering, computer architecture, computer programming, and computer theory. Applying ideas from theoretical computer science, such as analysis of algorithms and computational complexity, enables us to place optical computing in a framework where we can try to answer a number of important questions. For example, which problems are optical computers suitable for solving? Also, how does the resource usage on optical computers compare with more standard (e. g. digital electronic) architectures? Furthermore, optical computing gives one an opportunity to apply computer theory on a completely new suite of machine models. In contrast to a number of other nature-inspired models of computation, optical computers have very real and immediate realization possibilities.

Traditionally, in optical information processing a distinction was made between signal/image processing through optics and numerical processing through optics, with only the latter (and often only the digital version of the latter) being called optical computing [44,73,85,142]. However, it was always difficult to clearly delineate between the two, since it was largely a question of the interpretation the programmer attached to the output optical signals. The most important argument for referring to the latter only as optical computing had to do with the fact that the perceived limits (or at least, ambitions) of the former was simply for special-purpose signal/image processing devices while the ambitions for the latter was general-purpose computation. Given recent results on the computational power of optical image processing architectures [91,131,136], it is not the case that such architectures are limited to special-purpose tasks. Furthermore, as the field become increasingly multidisciplinary, and in particular as computer scientists play a more prominent role, it is necessary to bring the definition of optical computing in line with the broad definition of computing. In particular, this facilitates analysis from the theoretical computer science point of view. The distinction between analog optical computing and digital optical computing is similarly blurred given the prevalence of digital multiplication schemes effected through analog convolution [73]. Our broad interpretation of the term optical computing has been espoused before [25].

Introduction

The three most basic hardware components of an optical information processing system are a source, a modulator, and a detector. A source generates the light, a modulator multiplies the light by a (usually, spatially varying) function, and a detector senses the resulting light. The simplest example of a modulator encoding a spatially varying function is a transparency (a sheet of clear plastic or photographic film) with an opaque pattern handwritten or printed onto it. When placed in the path of an advancing wavefront, which we define simply as being a wide beam of light, the modulator encodes its pattern onto this wavefront. The common liquid-crystal display projector is a programmable example of the same principle. Keeping this kind of system in mind, we now highlight some attributes of optical information processing systems.

Time Efficiency

Consider a light detector that converts incident light into an electrical current. Consider also an encoding scheme whereby the intensity in a beam of light represented a particular nonnegative integer. Further, assume there are no fluctuations in the light source output, that the encoding scheme is linear, and that the detector's response is linear. Then, the sum of two such nonnegative integers incident on the detector could be determined by measuring the detector's current. In fact, several nonnegative integers could be summed in this way, with a single measurement (see Fig. 1). (This concept is not unknown to designers of analog electrical ANNs. However, the important difference is that since the medium is free space, the practical fan-in limitations of Kirchoff Law summation [86] in analog electronics do not apply here.) Such an optical arrangement can find the sum of n nonnegative integers in $O(1)$ addition steps. On a model of a sequential digital electronic computer this would require $n-1$ addition

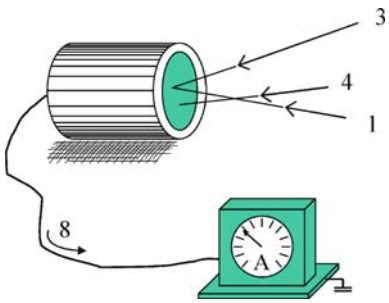

Optical Computing, Figure 1
A light detector apparatus converts incident light into an electrical current. Multiple nonnegative integers, encoded in beams of light, can be summed in unit time

operations and even a parallel digital electronic machine with n or more processors requires $O(\log n)$ timesteps. Tasks that rely on scalar summation operations (such as matrix multiplication) would benefit greatly from an optical implementation of the scalar sum operation. Similarly, $O(1)$ multiplication and $O(1)$ convolution operations can be realized optically. In Sect. "Optical Computing and Computational Complexity" we formally describe the time efficiency of a broad class of optical computers. Very recently, an optics-based digital signal processing platform has been marketed that claims digital processing speeds of tera (10^{12}) operations per second [79].

Efficiency in Interconnection Complexity

As optical pathways can cross in free space without measurable effect on the information in either channel, high interconnection densities are possible with optics [20,27]. Architectures with highly parallel many-to-many interconnections between parallel surfaces have already been proposed for common tasks such as sorting [8,37,83,115]. Currently, intra-chip, inter-chip, and inter-board connections are being investigated for manufacturing feasibility [87].

Energy Efficiency

Electrical wires suffer from induced noise and heat, which increases dramatically whenever wires are made thinner or placed closer together, or whenever the data throughput is increased [87]. As a direct consequence of their resistance-free pathways and noise-reduced environments, optical systems have the potential to generate less waste heat and so consume less energy per computation step than electronic systems [21]. This has been demonstrated experimentally with general-purpose digital optical processors [59,116,117].

Coherence

Mutually spatially coherent optical wavefronts (such as from a laser) interfere with each other just as waves in a water tank do. In the theory of physical optics, coherent wavefronts can be described by, and thus can represent, a complex-valued function (both positive and negative values in each of the real and imaginary axes). Using the language of this theory to interpret optical phenomena permits the definition of (at least) three important information processing constant-time operations: spatial modulation, Fourier transformation, and signal squaring [55]. The ability to perform such operations has resulted in many constant-time optical implementations of standard convolution-based digital image processing tasks.

While incoherent light (such as from an ordinary lightbulb) has some advantages in terms of tolerances in misalignment of optical components and less susceptibility to certain types of noise, coherent light is more general (in that its mathematics can be used to describe both coherent and incoherent wavefronts). Incoherent wavefronts can be modeled as being nonnegative everywhere, and so only admit possibilities to directly represent nonnegative spatially-varying and temporally-varying functions.

Optical Image Processing

It has long been appreciated that spatial optical signals are the most natural means of representing continuous tone 2D signals. There are many positive aspects to processing information using these (sometimes unwieldy and always inaccurate) physical signals instead of the more accurate digital electronic representations of 2D signals. These include the ability to concurrently modify all parts of an image (spatial light modulation), the capability to substitute space computational complexity for time computational complexity when performing certain transformations [22,82,91,107,131] (such as constant-time Fourier transformation with coherent light), the potential significant energy savings [21] (in both creating the signal and effecting the computation), and the ease with which analog signals can be digitized or resampled at an arbitrary frequency for subsequent digital electronic handling. The most common applications of optical image processing are pattern recognition and numerical matrix computations (see Sect. "History" for elaboration).

Pattern recognition is one of the most commonly implemented signal, image, and information processing tasks. A significant number of the algorithms for these tasks involve a convolution operation, either as the principal operation (e. g. comparing two images on a pixel-by-pixel basis using correlation) or as part of necessary pre-

processing (e. g. edge detection prior to applying a Hough transform [64]). The inherent parallel nature of optical systems can be used to facilitate low time and space complexity implementations of the convolution operation, either by multiplication in the Fourier domain or by systolic action [26].

Overview of the Chapter

In this chapter, we focus on optical image processing as it is an optical computing paradigm that makes full use of the degrees of freedom afforded by optics. Other optical computing architectures that seek to emulate perfectly in step-by-step fashion the operations of digital electronic architectures (for example, architectures built upon all-optical flip-flops [34] and all-optical network routers [39]) occupy an important place in the taxonomy of optical computing. However, in terms of computational complexity, the analyzes of these digital optical computers are in many respects identical to that of the digital electronic counterparts they emulate.

In Sect. "History", we give a brief overview of the history of optical computing, commonly referred to as optical information processing. Section "History" also includes an overview of existing optical models of computation. This is followed in Sect. "Selected Elements of Optical Computing Systems" by a summary of the most important elements of optical computing that could be used to define the functionality of an optical model of computation. In Sect. "Continuous Space Machine (CSM)", we take a detailed look at a particular model of optical computing (the CSM) that encompasses most of the functionality that coherent optical information processing has to offer. We begin by defining the CSM and a total of seven complexity measures that are inspired by real-world (optical) resources. We go on to discuss how the CSM's operations could be carried out physically. Section "Example CSM Datastructures and Algorithms" contains some example datastructures and algorithms for the CSM. In Sect. "C_2-CSM" we motivate and introduce an important restriction of the model called the C_2-CSM, and in Sect. "Optical Computing and Computational Complexity" we describe a number of C_2-CSM computational complexity results, and their implications. We conclude with some future directions in Sect. "Future Directions".

History

It could be argued that the field of optical information processing began in earnest with the realization that spatially coherent laser light could be used to conveniently Fourier transform an image, allow one to modify the complex-valued spatial frequency components, and then inverse Fourier transform back to the spatial domain. This concept is called spatial filtering [35,36,78,95,121,123,124], it is a generalization that encompasses convolution and correlation operations, and it could be performed over two-dimensional (2D) images in constant time while limited in speed only by the refresh rates of the input and output devices. It first found application in the 1950s for parallel processing and analysis of the huge amounts of radar data produced at the time. The initial special-purpose spatial filtering systems performed optical Fourier transforms, performed image processing (for example, noise reduction and edge enhancement), and recognized patterns through correlation. The fundamentals of optical spatial filtering were formulated in that decade, and built upon previous work on optimum linear filtering in communication theory. Achieving the full potential of optical spatial filtering theory requires filters that are complex-valued, and a technique to obtain such filters was first proposed by VanderLugt [123,125]. The technique allows one to physically encode a complex-valued image on an amplitude-modulating SLM such as an LCD panel.

Research continued into this form of image-based computation. Many important image processing tasks were demonstrated at that time, from character recognition [4], to real-time tracking of moving objects [48,122], to telescope/microscope image deblurring [118]. Two important strands of this research at the time were the development of sophisticated pattern recognition algorithms and numerical computation using values encoded in the complex amplitude or intensity of the light.

Optical Pattern Recognition

In pattern recognition [6,13,16,17,24,31,66,72], effort focused on achieving systems invariant to scaling, rotation, out-of-plane rotation, deformation, and signal dependent noise, while retaining the existing invariance to translating, adding noise to, and obscuring parts of the input. Effort also went into injecting nonlinearities into these inherently linear systems to achieve wider functionality [70,71]. Improvements were made to the fundamental limitations of the VanderLugt filter, most notably the joint transform correlator architecture [130].

Optical correlators that use incoherent sources of illumination (both spatially and temporally) rather than lasers are also possible [14,15,42,100,143]. The simplest incoherent correlator would have the same basic architecture as that used for matched filtering. While coherent systems in principle are more capable than incoherent systems (principally because the former naturally represents complex

functions while the latter naturally represents real functions), incoherent systems require less precise positioning when it comes to system construction and are less susceptible to noise.

A common example of an optical correlator's use in practical systems involved it being used as a front end to a generalized hybrid object recognition system. The optical processing component would quickly and efficiently identify regions of interest in a cluttered scene and pass these on to the slower but more accurate digital electronic components for false-alarm reduction, feature extraction, and classification. Today, the matched filter and the joint transform correlator are the two most widely used optical pattern recognition techniques [45,63,84,140].

Trade-offs between space and time were proposed and demonstrated. These included time integrating correlators [125] (in contrast to the space integrating correlators mentioned thus far) and systolic architectures [18, 26,50] where, for example, the propagation of an amplitude-modulated pressure wave through an acousto-optic device naturally effects the required correlation lags [108]. In addition to pattern recognition, a common application for these classes of architectures was numerical calculation.

Analog Optical Numerical Computation

An important strand of image-based optical computation involved numerical calculations: analog computation as well as multi-level discrete computation. Matrix-vector and matrix-matrix multiplication systems were proposed and demonstrated [5,44,63,73,76,85,125]. The capability to expand a beam of light and to focus many beams of light to a common point directly corresponded to high fan-out and fan-in capabilities, respectively. The limitations of encoding a number simply as an intensity value (finite dynamic range and finite intensity resolution in the modulators and detectors) could be overcome by representing the numbers in some base. Significant effort went into dealing with carry operations so that in additions, subtractions, and multiplications each digit could be processed in parallel. Algorithms based on convolution to multiply numbers in this representation were demonstrated [73], with a single post-processing step to combine the sub-calculations and deal with the carry operations. Residue arithmetic was demonstrated as a viable alternative in which carry operations did not arise at all, and for which a matrix-vector multiplier was proposed [68], but of course conversion to and from residue format is necessary.

An application that benefited greatly from the tightly-coupled parallelism afforded by optics was the solving of sets of simultaneous equations and matrix inversion [1,23]. An application that, further, was tolerant to the inherent inaccuracies and noise of analog optics was optical neural networks [27,43,65,104] including online neural learning in the presence of noise [90].

Digital Optical Computing

The next major advances came in the form of optical equivalents of digital computers [67]. The flexibility of digital systems over analog systems in general was a major factor behind the interest in this form of optical computation [109]. Specific drawbacks of the analog computing paradigm in optics that this new paradigm addressed included no perceived ability to perform general purpose computation, accumulation of noise from one computation step to another, and systematic errors introduced by imperfect analog components. The aim was to design digital optical computers that followed the same principles as conventional electronic processors but which could perform many binary operations in parallel. These systems were designed from logic gates using nonlinear optical elements: semitransparent materials whose transmitted intensity has a nonlinear dependence on the input intensity. Almost always, the coherence of the light was not used in the computation. All-optical bistable devices acting as flip-flops were demonstrated. The field drew on many decades of research into fast on-off optical switching speeds which was heralding an explosion in optical fiber communications. The difficulties that the digital optical paradigm experienced included how to fully exploit the theoretical parallelism of optics within an optical logic framework, how to efficiently manufacture very large systems of cascaded nonlinear optical elements (for which integrated optics holds promise [51]), and the more fundamental mathematical problem of how to parallelize arbitrary algorithms in the first place to exploit the parallelism afforded by digital optics.

Digital optical computing was also proposed as an application of architectures designed originally for image-based processing, for example logic effected through symbolic substitution [11,12]. At the confluence of computing and communication, optical techniques were proposed for the routing of signals in long-haul networks [44,142]. This is a promising application given that most long-haul communications already use light in optical fibers, and the conversion from optical to electronic in order to switch, and them back to optical to retransmit, can be costly. Initial implementations followed the concept of an optoelectronic crossbar switch with n inputs and n outputs [109], while latterly more effort is now going into

all-optical packet switching in a single channel configuration [34,39,49].

Optical Models of Computation

As already discussed, optical computers were designed and built to emulate conventional microprocessors (digital optical computing), and for image processing over continuous wavefronts (analog optical computing and pattern recognition). Here we are interested in the latter class: optical computers that store data as images. Numerous physical implementations exist and example applications include fast pattern recognition and matrix-vector algebra [56,125]. There have been much resources devoted to designs, implementations and algorithms for such optical information processing architectures (for example see [5,22,40,44,56,77,82,85,90,107,125,142] and their references).

However the computational complexity theory of optical computers (that is, finding lower and upper bounds on computational power in terms of known complexity classes) had received relatively little attention when compared with other nature-insired computing paradigms. Some authors have even complained about the lack of suitable models [44,82]. Many other areas of natural computing (e. g. [2,58,62,80,88,89,99,112,139]) have not suffered from this problem. Even so, we discuss some optical computation research that is close to the goals of the theoretical computer scientist.

Reif and Tyagi [107] study two optically inspired models. The first model is a 3D VLSI model augmented with a 2D discrete Fourier transform (DFT) primitive and parallel optical interconnections. The second model is a DFT circuit with operations (multiplication, addition, comparison of two inputs, DFT) that compute over an ordered ring. Parallel time complexity is defined for both models in the obvious way. For the first model, volume complexity is defined as the volume of the smallest convex box enclosing an instance of the model. For the DFT circuit, size is defined as the number of edges plus gates. Constant time, polynomial size/volume, algorithms for a number of problems are reported including 1D DFT, matrix multiplication, sorting and string matching [107].

Feitelson [44] gives a call to theoretical computer scientists to apply their knowledge and techniques to optical computing. He then goes on to generalize the concurrent read, concurrent write parallel random access machine, by augmenting it with two optically inspired operations. The first is the ability to write the same piece of data to many global memory locations at once. Secondly, if many values are concurrently written to a single memory location then a summation of those values is computed in a single timestep. Essentially Feitelson is using 'unbounded fan-in with summation' and 'unbounded fan-out'. His architecture mixes a well known discrete model with some optical capabilities.

A symbolic substitution model of computation has been proposed by Huang and Brenner, and a proof sketched of its universality [11]. This model of digital computation operates over discrete binary images and derives its efficiency by performing logical operations on each pixel in the image in parallel. It has the functionality to copy, invert, and shift laterally individual images, and OR and AND pairs of images. Suggested techniques for its optical implementation are outlined.

In computer science there are two famous classes of problems called P and NP [97]. P contains those problems that are solvable in polynomial time on a standard sequential computer, while NP is the class of problems that are solvable in polynomial time on a nondeterministic computer. NP contains P, and it is widely conjectured that they are not equal. A direct consequence of this conjecture is that there are (NP-hard) problems for which we strongly believe there is no polynomial time algorithm on a standard sequential computer.

It is known that it is possible to solve any NP (and even any PSPACE) problem in polynomial time on optical computers, albeit with exponential use of some other, space-like, resources [131,133,135]. In Sect. "C_2-CSM and Parallel Complexity Theory", we describe how parallel optical algorithms can solve such problems.

Along with these rather general results, there are a number of specific examples of algorithms with related resource usage for NP-hard problems. Shaked et al. [110,111] design an optical system for solving the NP-hard traveling salesman problem in polynomial time. Basically they use an optical matrix-vector multiplier to multiply the (exponentially large) matrix of tours by the vector of intercity weights. They give both optical experiments and simulations. Dolev and Fitoussi [38] give optical algorithms that make use of (exponentially large) masks to solve a number of NP-hard problems. Oltean [94], and Haist and Osten [60], give polynomial time algorithms for Hamiltonian path, and traveling salesman problem, respectively, via light travelling through optical cables. As is to be expected, both suffer from exponential (space-like) resource use. Nature-inspired systems that apparently solve NP-hard problems in polynomial time, while using an exponential amount of some other resource(s), have been around for many years. So the existence of massively parallel optical systems for NP-hard problems should not really surprise the reader. Nevertheless, it is interesting

to know the computational abilities, limitations, and resource trade-offs of such optical architectures, as well as to find particular (tractable or intractable) problems which are particularly suited to optical algorithms.

Reif, Tygar and Yoshida [106] examined the computational complexity of ray tracing problems. In such problems we are concerned about the geometry of the an optical system where diffraction is ignored and we wish to predict the position of light rays after passing through some system of mirrors and lenses. They gave undecidability and PSPACE hardness results, which gives an indication of the power of these systems as computational models.

Selected Elements of Optical Computing Systems

If one is designing an optical model of computation, one will incorporate the functionality of a subset of the following selected elements (devices and functionality) of optical computing systems.

Sources

Lasers are a common source of illumination because at some levels they are mathematically simpler to understand, but incoherent sources such as light-emitting diodes are also used frequently for increased tolerance to noise and when nonnegative functions are sufficient for the computation. Usually, the source is monochromatic to avoid the problem of color dispersion as the light passes through refracting optical components, unless this dispersion is itself the basis for the computation.

Spatial Light Modulators

It is possible to encode a spatial function (a 2D image) in an optical wavefront. A page of text when illuminated with sunlight, for example, does this job perfectly. This would be called an amplitude-modulating reflective SLM. Modulators can also act on phase and polarization, and can be transmissive rather than reflective. They include photographic film, and electro-optic, magneto-optic, and acousto-optic devices [5,44,56,73,85]. One class of note are the optically-addressed SLMs, in which, typically, a 2D light pattern falling on a photosensitive layer on one side of the SLM spatially varies (with an identical pattern) the reflective properties of the other side of the SLM. A beam splitter then allows one to read out this spatially-varying reflectance pattern. The liquid-crystal light valve [41,57,69,128,129] is one instance of this class. Other classes of SLMs such as liquid-crystal display panels and acousto-optic modulators allow one to dynamically alter the pattern using electronics. It is possible for a single device (such as an electronically programmed array of individual sources) to act as both source and modulator.

Detectors and Nature's Square Law

Optical signals can be regarded as having both an amplitude and phase. However, detectors will measure only the square of the amplitude of the signal (referred to as its intensity). This phenomenon is known as Nature's detector square law and applies to detectors from photographic film to digital cameras to the human eye. Detectors that obey this law are referred to as square-law detectors. This law is evident in many physical theories of light. In quantum theory, the measurement of a complex probability function is formalized as a projection onto the set of real numbers through a squaring operation. Square-law detectors need to be augmented with a interferometric or holographic arrangement to measure both amplitude and phase rather than intensity [19,47], or need to be used for multiple captures in different domains to heuristically infer the phase.

Since it squares the absolute value of a complex function, this square law can be used for some useful computation (for example, in the joint transform correlator [130]). Detectors most commonly used include high range point (single pixel) detectors such as photodiodes, highly sensitive photon detectors such as photomultiplier tubes, and 1D and 2D array detectors such as CCD- or CMOS-digital cameras. Intensity values outside the range of a detector (outside the lowest and highest intensities that the detector can record) are thresholded accordingly. The integration time of some detectors can be adjusted to sum all of the light intensity falling on them over a period of time. Other detectors can have quite large light sensitive areas and can sum all of the light intensity falling in a region of space.

Lenses

Lenses can be used to effect high fan-in and fan-out interconnections, to rescale images linearly in either one or two dimensions, and for taking Fourier transforms. In fact, a coherent optical wavefront naturally evolves into its Fresnel transform, and subsequently into its Fourier transform at infinity, and the lens simply images those frequency components at a finite fixed distance.

VanderLugt [125] has derived an expression for the coherent optical Fourier transform, given here in 1D for convenience, using the Fresnel transform of a complex-valued function $f(x)$ positioned in the front focal plane of a convex lens (plane P1 in Fig. 2), which is illuminated from the back by a plane wave of constant amplitude and phase. In terms of the physical coordinate ξ, the signal at the back

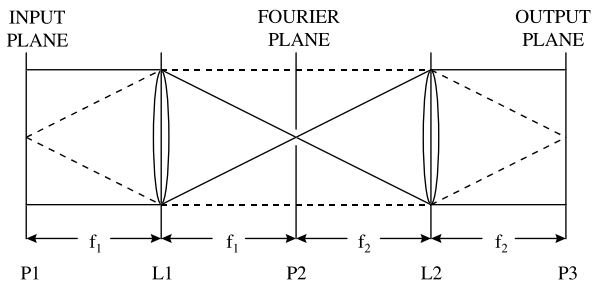

Optical Computing, Figure 2

Optical spatial frequency filtering using two convex lenses with a plane wave illuminating the input from the left. Lenses L1 and L2 have focal lengths of f_1 and f_2, respectively. If we assume that the physical dimensions of the lenses allow all diffracted orders to pass, the output is a transposed (and rescaled if $f_1 \neq f_2$) but otherwise identical version of the input. Any modification to the light field in the Fourier plane results in a spatial frequency filtered output

focal plane P2 can be written as

$$F(\xi) = \sqrt{\frac{\mathrm{i}}{\lambda L}} \int_{-\infty}^{\infty} f(x) \exp\left(\mathrm{i}2\pi x\xi/\lambda L\right) \mathrm{d}x \,, \tag{1}$$

where λ is the wavelength of the illumination and L is the focal length of the lens. Rewriting so it is a function of the spatial frequency variable α (α is a measurement of radians per unit distance) gives the common equation

$$F(\alpha) = \sqrt{\frac{\mathrm{i}}{\lambda L}} \int_{-\infty}^{\infty} f(x) \exp\left(\mathrm{i}2\pi \alpha x\right) \mathrm{d}x \,, \tag{2}$$

where $\xi = \lambda L \alpha$, and where we ignore the architecture-specific scaling constant. The Fourier transform in optics can be formed under a wide variety of conditions, and not just with a plane wave and not just in the focal plane of the lens [125]. This formalism adopts the paraxial approximation: that the distance between planes P1 and P2 is very much greater than the extent of the information in P1, thus avoiding the need for curved opposing surfaces in planes P1 and P3. When a Fourier transform is detected directly, only its square, called the power spectrum, is recorded. As mentioned, holography [19,47] overcomes this.

Interference and Complex Addition

Although photons, being bosons, do not interact with each other, coherent wavefronts can be made to interact at a detector. The addition (called superposition) of complex-valued wavefronts at a measurement is termed interference. Optically, interference is the same phenomenon as diffraction and is responsible for the formation of optical Fourier transforms. The linearity property is a useful property when analyzing coherent optical phenomena. If several images are coplanar then the optical field at their common Fourier plane is the superposition of their frequency spectra. This superposition, or interference, of complex-valued signals can be regarded as a pointwise addition of the complex amplitudes of the images.

Incoherent wavefronts are nonnegative everywhere. The addition of several incoherent wavefronts is linear in the intensity of each of those wavefronts. The addition of coherent wavefronts is linear in their complex amplitudes.

Image Copying and Combining

Images can be copied using optically-addressed SLMs or by dividing the optical energy along two paths using a beam splitter [33,125,130]. Beam splitters can also be used to combine several images to make them coplanar.

Multiplication of Images

When a signal's Fourier spectrum $F(\alpha, \beta)$ is coplanar with a transparency that encodes a second Fourier spectrum $H(\alpha, \beta)$, and if their centers are coincident, the complex signal in the plane immediately behind the transparency can be described as the pointwise multiplication of the two spectra

$$G(\alpha, \beta) = F(\alpha, \beta) H(\alpha, \beta) \,. \tag{3}$$

The signal $G(\alpha, \beta)$, which is also a frequency spectrum, could be inverse Fourier transformed to reveal a suitably correlated, convolved, or spatial frequency filtered original signal. A significant proportion of analog optics' role in the area of computation through filtering concerns convolution and those signal processing operations derived from it. Convolution filters are used extensively in the digital signal processing world to perform such tasks as deblurring, restoration, enhancement, and recognition [73]. The possibility of performing constant time convolution operations using coherent light is a promising concept.

To an approximation, optical systems can be regarded as both linear and shift-invariant. This is the basis for their convolution capabilities. Referring to Eq. (3), one can see that $H(\alpha, \beta)$ acts as a spatial frequency filter, altering the frequency content of the signal F. It could be used as a simple band-pass filter, damping the high-frequency components (noise suppression) or low frequency components (edge enhancement), or used to modulate the frequency spectrum with a more sophisticated spatial function. Mathematically, convolution with a mask A is equivalent to a frequency-domain multiplication with the Fourier transform of A.

Phase shifts can be introduced to a coherent wavefront by adding by a constant phase value. These could be constant shifts to effect numerical subtraction [44], or time-varying using a piezoelectric transducer mirror [138].

Other Elements of Optical Computation

A mirror changes the direction of the wavefront and simultaneously reflects it along some axis. A phase conjugate mirror [33] returns an image along the same path at which it approached the mirror.

In-plane rotation of an image by 180° can be accomplished using the apparatus in Fig. 2, a single lens, or even a pinhole camera. An out-of-plane tilt can be accomplished using a prism. In-plane flipping of an image (mirror image) can be accomplished using a prism, or using a beam splitter and some mirrors. Arbitrary in-plane rotation (with some tilting and translation, if required) can be achieved by combining several flip operations using a Dove prism or Pechan prism [96,119] or by combining several shearing operations [81]. Image rescaling can be accomplished using a combination of lenses (rescaling both dimensions identically), or using cylindrical lenses or an anamorphic prism (rescaling in one dimension only).

A prism or diffraction grating can be used to separate by wavelength the components of a multi-wavelength optical channel. For optical fiber communications applications, more practical (robust, economical, and scalable) alternatives exist to achieve the same effect [142].

Polarization is an important property of wavefronts, in particular in coherent optical computing, and is the basis for how liquid crystal displays work. At each point, an optical wavefront has an independent polarization value dependent on the angle, in the range $[0, 2\pi)$, of its electrical field. This can be independent of its successor (in the case of randomly polarized wavefronts), or dependent (as in the case of linear polarization), or dependent and time varying (as in the case of circular or elliptical polarization). Mathematically, a polarization state, and the transition from one polarization state to another, can be described using the Mueller calculus or the Jones calculus.

Photons can also be used for quantum computation, and quantum computers using linear optical elements (such as mirrors, polarizers, beam splitters, and phase shifters) have been proposed and demonstrated [28,75,101].

Continuous Space Machine (CSM)

For the remainder of this chapter we focus on an optical model of computation called the CSM. The model was originally proposed by Naughton [91,92]. The CSM is inspired by analog Fourier optical computing architectures, specifically pattern recognition and matrix algebra processors [56,90]. For example, these architectures have the ability to do unit time Fourier transformation using coherent (laser) light and lenses. The CSM computes in discrete timesteps over a number of two-dimensional images of fixed size and arbitrary spatial resolution. The data and program are stored as images. The (constant time) operations on images include Fourier transformation, multiplication, addition, thresholding, copying and scaling. The model is designed to capture much of the important features of optical computers, while at the same time be amenable to analysis from a computer theory point of view. Towards these goals we give an overview of how the model relates to optics as well as giving a number of computational complexity results for the model.

Section "CSM Definition" begins by defining the model. We analyze the model in terms of seven complexity measures inspired by real-world resources, these are described in Section "Complexity Measures". In Sect. "Optical Realization" we discuss possible optical implementations for the model. We then go on to give example algorithms and datastructures in Sect. "Example CSM Datastructures and Algorithms". The CSM definition is rather general, and so in Sect. "CSM Definition" we define a more restricted model called the C_2-CSM. Compared to the CSM, the C_2-CSM is somewhat closer to optical computing as it happens in the laboratory. Finally, in Sect. "Optical Computing and Computational Complexity" we show the power and limitations of optical computing, as embodied by the C_2-CSM, in terms computational complexity theory. Optical information processing is a highly parallel form of computing and we make this intuition more concrete by relating the C_2-CSM to parallel complexity theory by characterizing the parallel complexity class NC. For example, this shows the kind of worst case resource usage one would expect when applying CSM algorithms to problems that are known to be suited to parallel solutions.

CSM Definition

We begin this section by describing the CSM model in its most general setting, this brief overview is not intended to be complete and more details are to be found in [131].

A complex-valued image (or simply, image) is a function $f : [0, 1) \times [0, 1) \to \mathbb{C}$, where $[0, 1)$ is the half-open real unit interval. We let $\mathcal{I}$ denote the set of complex-valued images. Let $\mathbb{N}^+ = \{1, 2, 3, \ldots\}$, $\mathbb{N} = \mathbb{N}^+ \cup \{0\}$, and for a given CSM M let $\mathcal{N}$ be a countable set of images that encode M's addresses. An address is an element of $\mathbb{N} \times \mathbb{N}$.

$\mathrm{h}(i_1; i_2)$	: replace image at i_2 with horizontal 1D Fourier transform of i_1.
$\mathrm{v}(i_1; i_2)$	: replace image at i_2 with vertical 1D Fourier transform of image at i_1.
$*(i_1; i_2)$	: replace image at i_2 with the complex conjugate of image at i_1.
$\cdot i_1, i_2; i_3$	: pointwise multiply the two images at i_1 and i_2. Store result at i_3.
$+(i_1, i_2; i_3)$	: pointwise addition of the two images at i_1 and i_2. Store result at i_3.
$\rho(i_1, z_l, z_\mathrm{u}; i_2)$	: filter the image at i_1 by amplitude using z_l and z_u as lower and upper amplitude threshold images, respectively. Place result at i_2.
$[\xi'_1, \xi'_2, \eta'_1, \eta'_2] \leftarrow [\xi_1, \xi_2, \eta_1, \eta_2]$	: copy the rectangle of images whose bottom left-hand address is (ξ_1, η_1) and whose top right-hand address is (ξ_2, η_2) to the rectangle of images whose bottom left-hand address is (ξ_1', η_1') and whose top right-hand address is (ξ_2', η_2'). See illustration in Fig. 4.

Optical Computing, Figure 3
CSM high-level programming language instructions. In these instructions $i, z_\mathrm{l}, z_\mathrm{u} \in \mathbb{N} \times \mathbb{N}$ are image addresses and $\xi, \eta \in \mathbb{N}$. The control flow instructions are described in the main text

Additionally, for a given M there is an *address encoding function* $\mathfrak{E} : \mathbb{N} \to \mathcal{N}$ such that $\mathfrak{E}$ is Turing machine decidable, under some *reasonable* representation of images as words.

Definition 1 (CSM) A CSM is a quintuple $M = (\mathfrak{E}, L, I, P, O)$, where

- $\mathfrak{E} : \mathbb{N} \to \mathcal{N}$ is the address encoding function,
- $L = ((s_\xi, s_\eta), (a_\xi, a_\eta), (b_\xi, b_\eta))$ are the addresses: *sta*, a and b, where $a \neq b$,
- I and O are finite sets of input and output addresses, respectively,
- $P = \{(\zeta_1, p_{1_\xi}, p_{1_\eta}), \ldots, (\zeta_r, p_{r_\xi}, p_{r_\eta})\}$ are the r programming symbols ζ_j and their addresses (p_{j_ξ}, p_{j_η}) where $\zeta_j \in (\{h, v, *, \cdot, +, \rho, st, ld, br, hlt\} \cup \mathcal{N}) \subset \mathcal{I}$.

Each address is an element from $\{0, \ldots, \Xi - 1\} \times \{0, \ldots, \mathcal{Y} - 1\}$, where $\Xi, \mathcal{Y} \in \mathbb{N}^+$.

Addresses whose contents are not specified by P in a CSM definition are assumed to contain the constant image $f(x, y) = 0$. We interpret this definition to mean that M is (initially) defined on a grid of images bounded by the constants Ξ and $\mathcal{Y}$, in the horizontal and vertical directions respectively. The grid of images may grow in size as the computation progresses.

In our grid notation the first and second elements of an address tuple refer to the horizontal and vertical axes of the grid respectively, and image $(0, 0)$ is located at the lower left-hand corner of the grid. The images have the same orientation as the grid. For example the value $f(0, 0)$ is located at the lower left-hand corner of the image f.

In Definition 1 the tuple P specifies the CSM program using programming symbol images ζ_j that are from the (low-level) CSM programing language [131,136]. We refrain from giving a description of this programming language and instead describe a less cumbersome high-level language [131]. Figure 3 gives the basic instructions of this high-level language. The copy instruction is illustrated in Fig. 4. There are also **if/else** and **while** control flow instructions with conditional equality tests of the form $(f_\psi == f_\phi)$ where f_ψ and f_ϕ are *binary symbol images* (see Fig. 5a and b).

Address *sta* is the start location for the program so the programmer should write the first program instruction at *sta*. Addresses a and b define special images that are frequently used by some program instructions. The function $\mathfrak{E}$ is specified by the programmer and is used to map addresses to image pairs. This enables the programmer to choose her own address encoding scheme. We typically

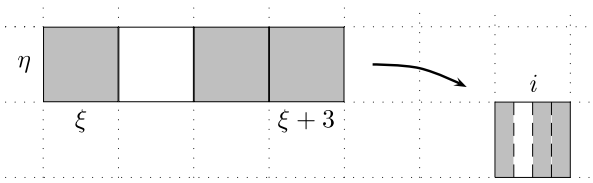

Optical Computing, Figure 4
Illustration of the instruction $i \leftarrow [\xi, \xi + 3, \eta, \eta]$ that copies four images to a single image that is denoted i

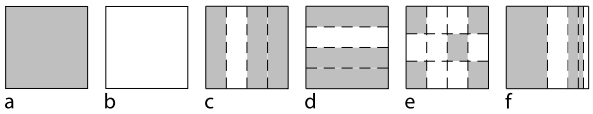

Optical Computing, Figure 5
Representing binary data. The *shaded areas* denote value 1 and the *white areas* denote value 0. a Binary symbol image representation of 1 and b of 0, c list (or row) image representation of the word 1011, d column image representation of 1011, e 3 × 4 matrix image, f binary stack image representation of 1101. *Dashed lines* are for illustration purposes only

don't want $\mathfrak{E}$ to hide complicated behavior thus the computational power of this function should be somewhat restricted. For example we put such a restriction on $\mathfrak{E}$ in Definition 7. At any given timestep, a configuration is defined in a straightforward way as a tuple $\langle c, e\rangle$ where c is an address called the control and e represents the grid contents.

Complexity Measures

In this section we define a number of CSM complexity measures. As is standard, all resource bounding functions map from $\mathbb{N}$ into $\mathbb{N}$ and are assumed to have the usual properties [7]. We begin by defining CSM TIME complexity in a manner that is standard among parallel models of computation.

Definition 2 The TIME complexity of a CSM M is the number of configurations in the computation sequence of M, beginning with the initial configuration and ending with the first final configuration.

The first of our six space-like resources is called GRID.

Definition 3 The GRID complexity of a CSM M is the minimum number of images, arranged in a rectangular grid, for M to compute correctly on all inputs.

Let $S : \mathcal{I} \times (\mathbb{N} \times \mathbb{N}) \to \mathcal{I}$, where $S(f(x, y), (\Phi, \Psi))$ is a raster image, with $\Phi\Psi$ constant-valued pixels arranged in Φ columns and Ψ rows, that approximates $f(x, y)$. If we choose a reasonable and realistic S then the details of S are not important.

Definition 4 The SPATIALRES complexity of a CSM M is the minimum $\Phi\Psi$ such that if each image $f(x, y)$ in the computation of M is replaced with $S(f(x, y), (\Phi, \Psi))$ then M computes correctly on all inputs.

One can think of SPATIALRES as a measure of the number of pixels needed during a computation. In optical image processing terms, and given the fixed size of our images, SPATIALRES corresponds to the space-bandwidth product of a detector or SLM.

Definition 5 The DYRANGE complexity of a CSM M is the ceiling of the maximum of all the amplitude values stored in all of M's images during M's computation.

In optical processing terms DYRANGE corresponds to the dynamic range of a signal.

We also use complexity measures called AMPLRES, PHASERES sand FREQ [131,136]. Roughly speaking, the AMPLRES of a CSM M is the number of discrete, evenly spaced, amplitude values per unit amplitude of the complex numbers in M's images, and so AMPLRES corresponds to the amplitude quantization of a signal. The PHASERES of M is the total number (per 2π) of discrete evenly spaced phase values in M's images, and so PHASERES corresponds to the phase quantization of a signal. Finally, the FREQ complexity of a CSM M is the minimum optical frequency necessary for M to compute correctly, this concept is explained further in [136].

Often we wish to make analogies between space on some well-known model and CSM 'space-like' resources. Thus we define the following convenient term.

Definition 6 The SPACE complexity of a CSM M is the product of all of M's complexity measures except TIME.

Optical Realization

In this section, we outline how some of the elementary operations of the CSM could be carried out physically. We do not intend to specify the definitive realization of any of the operations, but simply convince the reader that the model's operations have physical interpretations. Furthermore, although we concentrate on implementations employing visible light (optical frequencies detectable to the human eye) the CSM definition does not preclude employing other portion(s) of the electromagnetic spectrum.

A complex-valued image could be represented physically by a spatially coherent optical wavefront. Spatially coherent illumination (light of a single wavelength and emitted with the same phase angle) can be produced by a laser. SLM could be used to encode the image onto the expanded and collimated laser beam. One could write to a SLM offline (expose photographic film, or laser print or relief etch a transparency) or online (in the case of a liquid-crystal display [90,129,141] or holographic material [32,105]). The functions h and v could be effected using two convex cylindrical lenses, oriented horizontally and vertically, respectively [55,56,90,125]. As mentioned, a coherent optical wavefront will naturally evolve into its own Fourier spectrum as it propagates to infinity. What we do with a convex lens is simply image, at a finite distance, this spectrum at infinity. This finite distance is called the focal length of the lens. The constant θ used in the definitions of h and v could be effected using Fourier spectrum size reduction techniques [56,125] such as varying the focal length of the lens, varying the separation of the lens and SLM, employing cascaded Fourier transformation, increasing the dimensions/reducing the spatial resolution of the SLM, or using light with a shorter wavelength. The function $*$ could be implemented using a phase conjugate mirror [33]. The function $\cdot$ could be realized by placing a SLM encoding an image f in the path of a wavefront encoding another image g [56,123,125]. The wave-

front immediately behind the SLM would then be $\cdot(f, g)$. The function + describes the superposition of two optical wavefronts. This could be achieved using a 50:50 beam splitter [33,125,130]. The function ρ could be implemented using an electronic camera or a liquid-crystal light valve [129]. The parameters z_l and z_u would then be physical characteristics of the particular camera/light valve used. z_l corresponds to the minimum intensity value that the device responds to, known as the dark current signal, and z_u corresponds to the maximum intensity (the saturation level).

A note will be made about the possibility of automating these operations. If suitable SLMs can be prepared with the appropriate 2D pattern(s), each of the operations h, v, $*$, $\cdot$, and + could be effected autonomously and without user intervention using appropriately positioned lenses and free space propagation. The time to effect these operations would be the sum of the flight time of the image (distance divided by velocity of light) and the response time of the analog 2D detector; both of which are constants independent of the size or resolution of the images if an appropriate 2D detector is chosen. Examples of appropriate detectors would be holographic material [32,105] and a liquid-crystal light valve with a continuous (not pixellated) area [129]. Since these analog detectors are also optically-addressed SLMs, we can very easily arrange for the output of one function to act as the input to another, again in constant time independent of the size or resolution of the image. A set of angled mirrors will allow the optical image to be fed back to the first SLM in the sequence, also in constant time. It is not known, however, if ρ can be carried out completely autonomously for arbitrary parameters. Setting arbitrary parameters might fundamentally require offline user intervention (adjusting the gain of the camera, and so on), but at least for a small range of values this can be simulated online using a pair of liquid-crystal intensity filters.

We have outlined some optics principles that could be employed to implement the operations of the model. The simplicity of the implementations hides some imperfections in our suggested realizations. For example, the implementation of the + operation outlined above results in an output image that has been unnecessarily multiplied by the constant factor 0.5 due to the operation of the beam splitter. Also, in our suggested technique, the output of the ρ function is squared unnecessarily. However, all of these effects can be compensated for with a more elaborate optical setup and/or at the algorithm design stage.

A more important issue concerns the quantum nature of light. According to our current understanding, light exists as individual packets called photons. As such, in order to physically realize the CSM one would have to modify it such that images would have discrete, instead of continuous, amplitudes. The atomic operations outlined above, in particular the Fourier transform, are not affected by the restriction to quantized amplitudes, as the many experiments with electron interference patterns indicate. We would still assume, however, that in the physical world space is continuous.

A final issue concerns how a theoretically infinite Fourier spectrum could be represented by an image (or encoded by a SLM) of finite extent. This difficulty is addressed with the FREQ complexity measure [136].

Example CSM Datastructures and Algorithms

In this section we give some example data representations. We then to go on to give an example CSM algorithm that efficiently squares a binary matrix.

Representing Data as Images

There are many ways to represent data as images and interesting new algorithms sometimes depend on a new data representation. Data representations should be in some sense reasonable, for example it is unreasonable that the input to an algorithm could (non-uniformly) encode solutions to NP-hard or even undecidable problems. From Sect. "C_2-CSM", the CSM address encoding function gives the programmer room to be creative, so long as the representation is logspace computable (assuming a reasonable representation of images as words).

Here we mention some data representations that are commonly used. Figures 5a and 5b are the binary symbol image representations of 1 and 0 respectively. These images have an everywhere constant value of 1 and 0 respectively, and both have SPATIALRES of 1. The row and column image representations of the word 1011 are respectively given in Figs. 5c and 5d. These row and column images both have SPATIALRES of 4. In the matrix image representation in Fig. 5e, the first matrix element is represented at the top left corner and elements are ordered in the usual matrix way. This 3×4 matrix image has SPATIALRES of 12. Finally the binary stack image representation, which has exponential SPATIALRES of 16, is given in Fig. 5f.

Figure 4 shows how we might form a list image by copying four images to one in a single timestep. All of the above mentioned images have DYRANGE, AMPLRES and PHASERES of 1.

Another useful representation is where the value of a pixel directly encodes a number, in this case DYRANGE

becomes crucial. We can also encode values as phase values, and naturally PHASERES becomes a useful measure of the resources needed to store such values.

A Matrix Squaring Algorithm

Here we give an example CSM algorithm (taken from [133]) that makes use of the data representations described above. The algorithm squares a $n \times n$ matrix in $O(\log n)$ TIME and $O(n^3)$ SPATIALRES (number of pixels), while all other CSM resources are constant.

Lemma 1 *Let n be a power of 2 and let A be a $n \times n$ binary matrix. The matrix A^2 is computed by a C_2-CSM, using the matrix image representation, in* TIME $O(\log n)$, SPATIALRES $O(n^3)$, GRID *O(1),* DYRANGE *O(1),* AMPLRES *1 and* PHASERES *1.*

Proof In this proof the matrix and its matrix image representation (see Fig. 5e) are both denoted A. We begin with some precomputation, then one parallel pointwise multiplication step, followed by $\log n$ additions to complete the algorithm.

We generate the matrix image A_1 that consists of n vertically juxtaposed copies of A. This is computed by placing one copy of A above the other, scaling to one image, and repeating to give a total of $\log n$ iterations. The image A_1 is constructed in TIME $O(\log n)$, GRID $O(1)$ and SPATIALRES $O(n^3)$.

Next we transpose A to the column image A_2. The first n elements of A_2 are row 1 of A, the second n elements of A_2 are row 2 of A, etc. This is computed in TIME $O(\log n)$, GRID $O(1)$ and SPATIALRES $O(n^2)$ as follows.

Let $A' = A$ and $i = n$. We horizontally split A' into a left image A'_L and a right image A'_R. Then A'_L is pointwise multiplied (or masked) by the column image that represents $(10)^i$, in TIME $O(1)$. Similarly A'_R is pointwise multiplied (or masked) by the column image that represents $(01)^i$. The masked images are added. The resulting image has half the number of columns as A' and double the number of rows, and for example: row 1 consists of the first half of the elements of row 1 of A' and row 2 consists of the latter half of the elements of row 1 of A'. We call the result A' and we double the value of i. We repeat the process to give a total of $\log n$ iterations. After these iterations the resulting column image is denoted A_2.

We pointwise multiply A_1 and A_2 to give A_3 in TIME $O(1)$, GRID $O(1)$ and SPATIALRES $O(n^3)$.

To facilitate a straightforward addition we first transpose A_3 in the following way: A_3 is vertically split into a bottom and a top image, the top image is placed to the left of the bottom and the two are scaled to a single image, this splitting and scaling is repeated to give a total of $\log n$ iterations and we call the result A_4. Then to perform the addition, we vertically split A_4 into a bottom and a top image. The top image is pointwise added to the bottom image and the result is thresholded between 0 and 1. This splitting, adding and thresholding is repeated a total of $\log n$ iterations to create A_5. We 'reverse' the transposition that created A_4: image A_5 is horizontally split into a left and a right image, the left image is placed above the right and the two are scaled to a single image, this splitting and scaling is repeated a total of $\log n$ iterations to give A^2.

The algorithm highlights a few points of interest about the CSM. The CSM has quite a number of space-like resources, and it is possible to have trade-offs between them. For example in the algorithm above, if we allow GRID to increase from $O(1)$ to $O(n)$ then the SPATIALRES can be reduced from $O(n^3)$ to $O(n^2)$. In terms of optical architectures modeled by the CSM this phenomenon could be potentially very useful as certain resources may well be more economically viable than others. The algorithm is used in the proof that that polynomial TIME CSMs (and C_2-CSMs, see below) compute problems that are in the PSPACE class of languages. PSPACE includes the famous NP class. Such computational complexity results are discussed further in Sect. "Optical Computing and Computational Complexity" below.

There are a number of existing CSM algorithms, for these we point the reader to the literature [91,92,93,131,133,135,136].

C_2-CSM

In this section we define the C_2-CSM. One of the motivations for this model is the need to put reasonable upper bounds on the power of reasonable optical computers. As discussed below, it turns out that CSMs can very quickly use massive amounts of resources, and the C_2-CSM definition is an attempt to rein in this power.

Worst Case CSM Resource Usage

For the case of sequential computation it is usually obvious how the execution of a single operation will affect resource usage. In parallel models, execution of a single operation can lead to large growth in a single timestep. Characterizing resource growth is useful for proving upper bounds on power and choosing reasonable model restrictions.

We investigated the growth of complexity resources over TIME, with respect to CSM operations [131,134]. As expected, under certain operations some measures do not grow at all. Others grow at rates comparable to massively parallel models. By allowing operations like the

Optical Computing, Table 1

CSM resource usage after one timestep. For a given operation and complexity measure pair, the relevant table entry defines the worst case CSM resource usage at TIME $T+1$, in terms of the resources used at TIME T. At TIME T we have GRID $= G_T$, SPATIALRES $= R_{S,T}$, AMPLRES $= R_{A,T}$, DYRANGE $= R_{D,T}$, PHASERES $= R_{P,T}$ and FREQ $= \nu_T$

	GRID	SPATIALRES	AMPLRES	DYRANGE	PHASERES	FREQ
h	G_T	∞	∞	∞	∞	∞
v	G_T	∞	∞	∞	∞	∞
$*$	G_T	$R_{S,T}$	$R_{A,T}$	$R_{D,T}$	$R_{P,T}$	ν_T
$\cdot$	G_T	$R_{S,T}$	$(R_{A,T})^2$	$(R_{D,T})^2$	$R_{P,T}$	ν_T
$+$	G_T	$R_{S,T}$	∞	$2R_{D,T}$	∞	ν_T
ρ	unbounded	$R_{S,T}$	$R_{A,T}$	$R_{D,T}$	$R_{P,T}$	ν_T
st	unbounded	$R_{S,T}$	$R_{A,T}$	$R_{D,T}$	$R_{P,T}$	ν_T
ld	unbounded	unbounded	$R_{A,T}$	$R_{D,T}$	$R_{P,T}$	unbounded
br	G_T	$R_{S,T}$	$R_{A,T}$	$R_{D,T}$	$R_{P,T}$	ν_T
hlt	G_T	$R_{S,T}$	$R_{A,T}$	$R_{D,T}$	$R_{P,T}$	ν_T

Fourier transform we are mixing the continuous and discrete worlds, hence some measures grow to infinity in one timestep. This gave strong motivation for CSM restrictions.

Table 1 summarizes these results; the table defines the value of a complexity measure after execution of an operation (at TIME $T+1$). The complexity of a configuration at TIME $T+1$ is at least the value it was at TIME T, since complexity functions are nondecreasing. Our definition of TIME assigns unit time cost to each operation, hence we do not have a TIME column. Some entries are immediate from the complexity measure definitions, for others proofs are given in the references [131,134].

C_2-CSM

Motivated by a desire to apply standard complexity theory tools to the model, we defined [131,134] the C_2-CSM, a restricted class of CSM.

Definition 7 (C_2-CSM) A C_2-CSM is a CSM whose computation TIME is defined for $t \in \{1, 2, \ldots, T(n)\}$ and has the following restrictions:

- For all TIME t both AMPLRES and PHASERES have constant value of 2.
- For all TIME t each of GRID, SPATIALRES and DYRANGE is $2^{O(t)}$ and SPACE is redefined to be the product of all complexity measures except TIME and FREQ.
- Operations h and v compute the discrete Fourier transform in the horizontal and vertical directions respectively.
- Given some *reasonable* binary word representation of the set of addresses $\mathcal{N}$, the address encoding function $\mathfrak{E}: \mathbb{N} \to \mathcal{N}$ is decidable by a logspace Turing machine.

Let us discuss these restrictions. The restrictions on AMPLRES and PHASERES imply that C_2-CSM images are of the form $f : [0, 1) \times [0, 1) \to \{0, \pm 1/2, \pm 1, \pm 3/2, \ldots\}$. We have replaced the Fourier transform with the discrete Fourier transform [10], this essentially means that FREQ is now solely dependent on SPATIALRES; hence FREQ is not an interesting complexity measure for C_2-CSMs and we do not analyze C_2-CSMs in terms of FREQ complexity [131,134]. Restricting the growth of SPACE is not unique to our model, such restrictions are to be found elsewhere [54,98,102].

In Sect. "CSM Definition" we stated that the address encoding function $\mathfrak{E}$ should be Turing machine decidable, here we strengthen this condition. At first glance sequential logspace computability may perhaps seem like a strong restriction, but in fact it is quite weak. From an optical implementation point of view it should be the case that $\mathfrak{E}$ is not complicated, otherwise we cannot assume fast addressing. Other (sequential/parallel) models usually have a very restricted 'addressing function': in most cases it is simply the identity function on $\mathbb{N}$. Without an explicit or implicit restriction on the computational complexity of $\mathfrak{E}$, finding non-trivial upper bounds on the power of C_2-CSMs is impossible as $\mathfrak{E}$ could encode an arbitrarily complex Turing machine. As a weaker restriction we could give a specific $\mathfrak{E}$. However, this restricts the generality of the model and prohibits the programmer from developing novel, reasonable, addressing schemes.

Optical Computing and Computational Complexity

There have been a number of optical algorithms given that use the inherent parallelism of optics to speed up the solutions to certain problems. An alternative approach is to ask

the following question: How does a given optical model relate to standard sequential and parallel models? Establishing a relationship with computational complexity theory, by describing both upper and lower bounds on the model, gives immediate access to a large collection of useful algorithms and proof techniques.

The parallel computation thesis [29,52,74,98,126] states that parallel time (polynomially) corresponds to sequential space, for reasonable parallel and sequential models. An example would be the fact that the class of problems solvable in polynomial space on a number of parallel models is equivalent to PSPACE, the class of problems solvable on Turing machines that use at most polynomial space [3,9,30,46,53,54,61,113,114,120,127].

Of course the thesis can never be proved, it relates the intuitive notion of reasonable parallelism to the precise notion of a Turing machine. When results of this type were first shown researchers were suitably impressed; their parallel models truly had great power. For example if model M verifies the thesis then M decides PSPACE (including NP) languages in polynomial time. However there is another side to this coin. It is straightforward to verify that given our current best algorithms, M will use at least a superpolynomial amount of some other resource (like space or number of processors) to decide a PSPACE-complete or NP-complete language. Since the composition of polynomials is itself a polynomial, it follows that if we restrict the parallel computer to use at most polynomial time and polynomial other resources, then it can at most solve problems in P.

Nevertheless, asking if M verifies the thesis is an important question. Certain problems, such as those in the class NC, are efficiently parallelisable. NC can be defined as the class of problems that are solvable in polylogarithmic time on a parallel computer that uses a polynomial amount of hardware. So one can think of NC as those problems in P which are solved exponentially faster on parallel computation thesis models than on sequential models. If M verifies the thesis then we know it will be useful to apply M to these problems. We also know that if M verifies the thesis then there are (P-complete) problems for which it is widely believed that we will not find exponential speed up using M.

C_2-CSM and Parallel Complexity Theory

Here we summarize some characterizations of the computing power of optical computers. Such characterizations enable the algorithm designer to know what kinds of problems are solvable with resource bounded optical algorithms.

Theorem 1 below gives lower bounds on the computational power of the C_2-CSM by showing that it is at least as powerful as models that verify the parallel computation thesis.

Theorem 1 ([133,135])
$\mathrm{NSPACE}(S(n)) \subseteq C_2\text{-}CSM\text{-}\mathrm{TIME}(O(S^2(n)))$

In particular, polynomial TIME C_2-CSMs accept the PSPACE languages. PSPACE is the class of problems solvable by Turing machines that use polynomial space, which includes the famous class NP, and so NP-complete problems can be solved by C_2-CSMs in polynomial TIME. However, any C_2-CSM algorithm that we could presently write to solve PSPACE or NP problems would require exponential SPACE.

Theorem 1 is established by giving a C_2-CSM algorithm that efficiently generates, and squares, the transition matrix of a $S(n) = \Omega(\log n)$ space bounded Turing machine. This transition matrix represents all possible computations of the Turing machine and is of size $O(2^S) \times O(2^S)$. The matrix squaring part was already given as an example (Lemma 1), and the remainder of the algorithm is given in [133]. The algorithm uses SPACE that is cubic in one of the matrix dimensions. In particular the algorithm uses cubic SPATIALRES, $O(2^{3S})$, and all other space-like resources are constant. This theorem improves upon the time overhead of a previous, but similar, result [131,135] that was established via C_2-CSM simulation of the vector machines [102,103] of Pratt, Rabin, and Stockmeyer.

From the resource usage point of view, it is interesting to see that the older of these two algorithms uses GRID, DYRANGE, and SPATIALRES that are each $O(2^S)$, while the newer algorithm shows that if we allow more SPATIALRES we can in fact use only constant GRID and DYRANGE. It would be interesting to find other such resource trade-offs within the model.

Since NP is contained in PSPACE, Theorem 1 and the corresponding earlier results in [131,135], show that this optical model solves NP-complete problems in polynomial TIME. As described in Sect. "Optical Models of Computation", this has also been shown experimentally, for example Shaked et al. [110] have recently given a polynomial time, exponential space, optical algorithm to solve the NP-complete travelling salesperson problem. Their optical setup can be implemented on the CSM.

The other of the two inclusions that are necessary in order to verify the parallel computation thesis have also been shown: C_2-CSMs computing in TIME $T(n)$ are no more powerful than $T^{O(1)}(n)$ space bounded deterministic Turing machines. More precisely, we have:

Theorem 2 ([131,132])
C_2-CSM-TIME$(T(n)) \subseteq$ DSPACE$(O(T^2(n)))$

This result gives an upper bound on the power of C_2-CSMs and was established via C_2-CSM simulation by logspace uniform circuits of size and depth polynomial in SPACE and TIME respectively [132].

Via the proofs of Theorems 1 and 2 we get another (stronger) result: C_2-CSMs that simultaneously use polynomial SPACE and polylogarithmic TIME solve exactly those problems in the class NC.

Corollary 1
C_2-CSM-SPACE, TIME$(n^{O(1)}, \log^{O(1)} n) =$ NC

Problems in NC highlight the power of parallelism, as these problems can be solved exponentially faster on a polynomial amount of parallel resources than on polynomial time sequential machines. As further work in this area one could try to find alternate characterizations of NC in terms of the C_2-CSM. In particular, one could try to find further interesting trade-offs between the various space-like resources of the model. In the real world this would correspond to computing over various different CSM resources. Also, it might be interesting for optical algorithm designers to try to design optical algorithms for NC problems in an effort to find problems that are well suited to optical solutions. See [137] for details on this argument, and also fpr other CSM characterisations of complexity classes and an implementation of a fast optical search algorithm.

Future Directions

As already noted, optical computing is an inherently multidisciplinary subject whose study routinely involves a spectrum of expertise that threads optical physics, materials science, optical engineering, electrical engineering, computer architecture, computer programming, and computer theory. From the point of view of each of these fields there are various directions for future work. Also, it is generally accepted that if optical computers become mainstream, it will be through a symbiotic relationship with their extremely flexible digital electronic counterparts. At the confluence of computing and communication there is room for optical techniques such as for the routing of signals in long-haul networks via all-optical packet switching in a single channel configuration. So it seems that whether or not optical computers will be adopted in a widespread manner is both a technological and economic issue.

From the algorithmic point of view there is plenty of scope for future work. There are a number of questions related to trade-offs between resources and we believe the CSM gives a good framework to answer such questions. For example can we give useful parallel algorithms that exploit CSM resources such as PHASERES while at the same time using small SPATIALRES and GRID? In a similar vein, one can explore the computing power of restrictions and generalizations of the CSM with the goal of finding new algorithms and characterizations of complexity classes. This has immediate applications in finding new and efficient implementations of optical solutions to computational problems.

Acknowledgments

DW thanks J. Paul Gibson and Cris Moore for interesting discussions. DW acknowledges Junta de Andalucia grant TIC-581, Science Foundation Ireland grant number 04/IN3/1524, and Irish Research Council for Science Engineering and Technology grant number PD/2004/33. TN acknowledges support from the European Commission through a Marie Curie Intra-European Fellowship.

Bibliography

1. Abushagur MAG, Caulfield HJ (1987) Speed and convergence of bimodal optical computers. Opt Eng 26(1):22–27
2. Adleman LM (1994) Molecular computation of solutions to combinatorial problems. Science 266:1021–1024
3. Alhazov A, de Jesús Pérez–Jiménez M (2007) Uniform solution to QSAT using polarizationless active membranes. In: Durand–Lose J, Margenstern M (eds) Machines, Computations and Universality (MCU). LNCS, vol 4664. Springer, Orléans, pp 122–133
4. Armitage JD, Lohmann AW (1965) Character recognition by incoherent spatial filtering. Appl Opt 4(4):461–467
5. Arsenault HH, Sheng Y (1992) An Introduction to Optics in Computers. Tutorial Texts in Optical Engineering, vol TT8. SPIE Press, Bellingham, Washington
6. Arsenault HH, Hsu YN, Chalasinska–Macukow K (1984) Rotation-invariant pattern recognition. Opt Eng 23(6):705–709
7. Balcázar JL, Díaz J, Gabarró J (1988) Structural complexity, vols I and II. EATCS Monographs on Theoretical Computer Science. Springer, Berlin
8. Beyette Jr FR, Mitkas PA, Feld SA, Wilmsen CW (1994) Bitonic sorting using an optoelectronic recirculating architecture. Appl Opt 33(35):8164–8172
9. Borodin A (1977) On relating time and space to size and depth. SIAM J Comput 6(4):733–744
10. Bracewell RN (1978) The Fourier transform and its applications, 2nd edn. Electrical and electronic engineering series, McGraw–Hill, New York
11. Brenner KH, Huang A, Streibl N (1986) Digital optical computing with symbolic substitution. Appl Opt 25(18):3054–3060
12. Brenner KH, Kufner M, Kufner S (1990) Highly parallel arithmetic algorithms for a digital optical processor using symbolic substitution logic. Appl Opt 29(11):1610–1618
13. Casasent DP (1984) Unified synthetic discriminant function computational formulation. Appl Opt 23:1620–1627
14. Casasent DP, House GP (1994) Comparison of coherent and noncoherent optical correlators. In: Optical Pattern Recog-

nition V. Proceedings of SPIE, vol 2237. SPIE, Bellingham, pp 170–178
15. Casasent DP, House GP (1994) Implementation issues for a noncoherent optical correlator. In: Optical Pattern Recognition V. Proceedings of SPIE, vol 2237. SPIE, Bellingham, pp 179–188
16. Casasent DP, Psaltis D (1976) Position, rotation, and scale invariant optical correlation. Appl Opt 15(7):1795–1799
17. Casasent DP, Psaltis D (1976) Scale invariant optical transforms. Opt Eng 15(3):258–261
18. Casasent DP, Jackson J, Neuman CP (1983) Frequency-multiplexed and pipelined iterative optical systolic array processors. Appl Opt 22:115–124
19. Caulfield HJ (ed) (1979) Handbook of Optical Holography. Academic Press, New York
20. Caulfield HJ (1989) Computing with light. Byte 14:231–237
21. Caulfield HJ (1989) The energetic advantage of analog over digital computing. In: OSA Optical Computing Technical Digest Series, vol 9. Optical Society of America, Washington, pp 180–183
22. Caulfield HJ (1990) Space-time complexity in optical computing. In: Javidi B (ed) Optical information-processing systems and architectures II. SPIE, vol 1347. SPIE Press, Bellingham, pp 566–572
23. Caulfield HJ, Abushagur MAG (1986) Hybrid analog-digital algebra processors. In: Optical and Hybrid Computing II. Proceedings of SPIE, vol 634. SPIE Press, Bellingham, pp 86–95
24. Caulfield HJ, Haimes R (1980) Generalized matched filtering. Appl Opt 19(2):181–183
25. Caulfield HJ, Horvitz S, Winkle WAV (1977) Introduction to the special issue on optical computing. Proceedings of the IEEE 65(1):4–5
26. Caulfield HJ, Rhodes WT, Foster MJ, Horvitz S (1981) Optical implementation of systolic array processing. Opt Commun 40:86–90
27. Caulfield HJ, Kinser JM, Rogers SK (1989) Optical neural networks. Proc IEEE 77:1573–1582
28. Cerf NJ, Adami C, Kwiat PG (1998) Optical simulation of quantum logic. Phys Rev A 57(3):R1477–R1480
29. Chandra AK, Stockmeyer LJ (1976) Alternation. In: 17th annual symposium on Foundations of Computer Science, IEEE, Houston, Texas, pp 98–108
30. Chandra AK, Kozen DC, Stockmeyer LJ (1981) Alternation. J ACM 28(1):114–133
31. Chang S, Arsenault HH, Garcia–Martinez P, Grover CP (2000) Invariant pattern recognition based on centroids. Appl Opt 39(35):6641–6648
32. Chen FS, LaMacchia JT, Fraser DB (1968) Holographic storage in lithium niobate. Appl Phys Lett 13(7):223–225
33. Chiou AE (1999) Photorefractive phase-conjugate optics for image processing, trapping, and manipulation of microscopic objects. Proc IEEE 87(12):2074–2085
34. Clavero R, Ramos F, Martí J (2005) All-optical flip-flop based on an active Mach–Zehnder interferometer with a feedback loop. Opt Lett 30(21):2861–2863
35. Cutrona LJ, Leith EN, Palermo CJ, Porcello LJ (1960) Optical data processing and filtering systems. IRE Trans Inf Theory 6(3):386–400
36. Cutrona LJ, Leith EN, Porcello LJ, Vivian WE (1966) On the application of coherent optical processing techniques to synthetic-aperture radar. Proc IEEE 54(8):1026–1032
37. Desmulliez MPY, Wherrett BS, Waddie AJ, Snowdon JF, Dines JAB (1996) Performance analysis of self-electro-optic-effect-device-based (seed-based) smart-pixel arrays used in data sorting. Appl Opt 35(32):6397–6416
38. Dolev S, Fitoussi H (2007) The traveling beam: optical solution for bounded NP-complete problems. In: Crescenzi P, Prencipe G, Pucci G (eds) The fourth international conference on fun with algorithms (FUN). Springer, Heidelberg, pp 120–134
39. Dorren HJS, Hill MT, Liu Y, Calabretta N, Srivatsa A, Huijskens FM, de Waardt H, Khoe GD (2003) Optical packet switching and buffering by using all-optical signal processing methods. J Lightwave Technol 21(1):2–12
40. Durand–Lose J (2006) Reversible conservative rational abstract geometrical computation is Turing-universal. In: Logical Approaches to Computational Barriers, Second Conference on Computability in Europe, (CiE). Lecture Notes in Computer Science. Springer, Swansea, vol 3988. pp 163–172
41. Efron U, Grinberg J, Braatz PO, Little MJ, Reif PG, Schwartz RN (1985) The silicon liquid crystal light valve. J Appl Phys 57(4):1356+
42. Esteve–Taboada JJ, García J, Ferreira C (2000) Extended scale-invariant pattern recognition with white-light illumination. Appl Opt 39(8):1268–1271
43. Farhat NH, Psaltis D (1984) New approach to optical information processing based on the Hopfield model. J Opt Soc Am A 1:1296
44. Feitelson DG (1988) Optical Computing: A survey for computer scientists. MIT Press, Cambridge, Massachusetts
45. Feng JH, Chin GF, Wu MX, Yan SH, Yan YB (1995) Multiobject recognition in a multichannel joint-transform correlator. Opt Lett 20(1):82–84
46. Fortune S, Wyllie J (1978) Parallelism in random access machines. In: Proc. 10th Annual ACM Symposium on Theory of Computing, ACM, New York, pp 114–118
47. Gabor D (1948) A new microscopic principle. Nature 161(4098):777–778
48. Gara AD (1979) Real time tracking of moving objects by optical correlation. Appl Opt 18(2):172–174
49. Geldenhuys R, Liu Y, Calabretta N, Hill MT, Huijskens FM, Khoe GD, Dorren HJS (2004) All-optical signal processing for optical packet switching. J Opt Netw 3(12):854–865
50. Ghosh AK, Casasent DP, Neuman CP (1985) Performance of direct and iterative algorithms on an optical systolic processor. Appl Opt 24(22):3883–3892
51. Goldberg L, Lee SH (1979) Integrated optical half adder circuit. Appl Opt 18:2045–2051
52. Goldschlager LM (1977) Synchronous parallel computation. Ph D thesis, University of Toronto, Computer Science Department
53. Goldschlager LM (1978) A unified approach to models of synchronous parallel machines. In: Proc. 10th Annual ACM Symposium on Theory of Computing. ACM, New York, pp 89–94
54. Goldschlager LM (1982) A universal interconnection pattern for parallel computers. J ACM 29(4):1073–1086
55. Goodman JW (1977) Operations achievable with coherent optical information processing systems. Proc IEEE 65(1):29–38
56. Goodman JW (2005) Introduction to Fourier Optics, 3rd edn. Roberts & Company, Englewood
57. Grinberg J, Jacobson AD, Bleha WP, Miller L, Fraas L, Boswell D, Myer G (1975) A new real-time noncoherent to coherent

light image converter: The hybrid field effect liquid crystal light valve. Opt Eng 14(3):217–225
58. Grover LK (1996) A fast quantum mechanical algorithm for database search. In: Proc. 28th Annual ACM Symposium on Theory of Computing. ACM, New York, pp 212–219
59. Guilfoyle PS, Hessenbruch JM, Stone RV (1998) Free-space interconnects for high-performance optoelectronic switching. IEEE Comp 31(2):69–75
60. Haist T, Osten W (2007) An optical solution for the travelling salesman problem. Opt Expr 15(16):10473–10482
61. Hartmanis J, Simon J (1974) On the power of multiplication in random access machines. In: Proceedings of the 15th annual symposium on switching and automata theory. IEEE, The University of New Orleans, pp 13–23
62. Head T (1987) Formal language theory and DNA: an analysis of the generative capacity of specific recombinant behaviors. Bull Math Biol 49(6):737–759
63. Horner JL (ed) (1987) Optical Signal Processing. Academic Press, San Diego
64. Hough PVC (1962) Methods and measures for recognising complex patterns. U.S. Patent No. 3069654
65. Hsu KY, Li HY, Psaltis D (1990) Holographic implementation of a fully connected neural network. Proc IEEE 78(10):1637–1645
66. Hsu YN, Arsenault HH (1982) Optical pattern recognition using circular harmonic expansion. Appl Opt 21(22):4016–4019
67. Huang A (1984) Architectural considerations involved in the design of an optical digital computer. Proc IEEE 72(7):780–786
68. Huang A, Tsunoda Y, Goodman JW, Ishihara S (1979) Optical computation using residue arithmetic. Appl Opt 18(2):149–162
69. Jacobson AD, Beard TD, Bleha WP, Morgerum JD, Wong SY (1972) The liquid crystal light value, an optical-to-optical interface device. In: Proceedings of the Conference on Parallel Image Processing, document X-711-72-308, Goddard Space Flight Center. NASA, Washington, pp 288–299
70. Javidi B (1989) Nonlinear joint power spectrum based optical correlation. Appl Opt 28(12):2358–2367
71. Javidi B (1990) Generalization of the linear matched filter concept to nonlinear matched filters. Appl Opt 29(8):1215–1217
72. Javidi B, Wang J (1995) Optimum distortion-invariant filter for detecting a noisy distorted target in nonoverlapping background noise. J Opt Soc Am A 12(12):2604–2614
73. Karim MA, Awwal AAS (1992) Optical Computing: An Introduction. Wiley, New York
74. Karp RM, Ramachandran V (1990) Parallel algorithms for shared memory machines. In: van Leeuwen J (ed) Handbook of Theoretical Computer Science, vol A. Elsevier, Amsterdam, pp 869–941
75. Knill E, LaFlamme R, Milburn GJ (2001) A scheme for efficient quantum computation with linear optics. Nature 409:46–52
76. Lee JN (ed) (1995) Design Issues in Optical Processing. Cambridge Studies in Modern Optics. Cambridge University Press, Cambridge
77. Lee JN (ed) (1995) Design issues in optical processing. Cambridge studies in modern optics. Cambridge University Press, Cambridge
78. Leith EN (1977) Complex spatial filters for image deconvolution. Proc IEEE 65(1):18–28
79. Lenslet Labs (2001) Optical digital signal processing engine. white paper report, Lenslet Ltd., 12 Hachilazon St., Ramat–Gan, Israel 52522
80. Lipton RJ (1995) Using DNA to solve NP-complete problems. Science 268:542–545
81. Lohmann AW (1993) Image rotation, Wigner rotation, and the fractional Fourier transform. J Opt Soc Am A 10(10):2181–2186
82. Louri A, Post A (1992) Complexity analysis of optical-computing paradigms. Appl Opt 31(26):5568–5583
83. Louri A, Hatch Jr JA, Na J (1995) Constant-time parallel sorting algorithm and its optical implementation using smart pixels. Appl Opt 34(17):3087–3097
84. Lu XJ, Yu FTS, Gregory DA (1990) Comparison of Vander lugt and joint transform correlators. Appl Phys B 51:153–164
85. McAulay AD (1991) Optical Computer Architectures: The Application of Optical Concepts to Next Generation Computers. Wiley, New York
86. Mead C (1989) Analog VLSI and Neural Systems. Addison–Wesley, Reading
87. Miller DA (2000) Rationale and challenges for optical interconnects to electronic chips. Proc IEEE 88(6):728–749
88. Moore C (1991) Generalized shifts: undecidability and unpredictability in dynamical systems. Nonlinearity 4:199–230
89. Moore C (1997) Majority-vote cellular automata, Ising dynamics and P-completeness. J Stat Phys 88(3/4):795–805
90. Naughton T, Javadpour Z, Keating J, Klíma M, Rott J (1999) General-purpose acousto-optic connectionist processor. Opt Eng 38(7):1170–1177
91. Naughton TJ (2000) Continuous-space model of computation is Turing universal. In: Bains S, Irakliotis LJ (eds) Critical Technologies for the Future of Computing. Proc SPIE, vol 4109. San Diego, pp 121–128
92. Naughton TJ (2000) A model of computation for Fourier optical processors. In: Lessard RA, Galstian T (eds) Optics in Computing 2000. Proc SPIE, vol 4089. Quebec, pp 24–34
93. Naughton TJ, Woods D (2001) On the computational power of a continuous-space optical model of computation. In: Margenstern M, Rogozhin Y (eds) Machines, Computations and Universality: Third International Conference (MCU'01). LNCS, vol 2055. Springer, Heidelberg, pp 288–299
94. Oltean M (2006) A light-based device for solving the Hamiltonian path problem. In: Fifth International Conference on Unconventional Computation (UC'06). LNCS, vol 4135. Springer, York, pp 217–227
95. O'Neill EL (1956) Spatial filtering in optics. IRE Trans Inf Theory 2:56–65
96. Paek EG, Choe JY, Oh TK, Hong JH, Chang TY (1997) Nonmechanical image rotation with an acousto-optic dove prism. Opt Lett 22(15):1195–1197
97. Papadimitriou CH (1995) Computational complexity. Addison–Wesley, Reading
98. Parberry I (1987) Parallel complexity theory. Wiley, New York
99. Păun G (2002) Membrane computing: an introduction. Springer, Heidelberg
100. Pe'er A, Wang D, Lohmann AW, Friesem AA (1999) Optical correlation with totally incoherent light. Opt Lett 24(21):1469–1471
101. Pittman TB, Fitch MJ, Jacobs BC, Franson JD (2003) Experimental controlled-NOT logic gate for single photons in the coincidence basis. Phys Rev A 68:032316–3

102. Pratt VR, Stockmeyer LJ (1976) A characterisation of the power of vector machines. J Comput Syst Sci 12:198–221
103. Pratt VR, Rabin MO, Stockmeyer LJ (1974) A characterisation of the power of vector machines. In: Proc 6th annual ACM symposium on theory of computing. ACM, New York, pp 122–134
104. Psaltis D, Farhat NH (1985) Optical information processing based on an associative-memory model of neural nets with thresholding and feedback. Opt Lett 10(2):98–100
105. Pu A, Denkewalter RF, Psaltis D (1997) Real-time vehicle navigation using a holographic memory. Opt Eng 36(10):2737–2746
106. Reif J, Tygar D, Yoshida A (1990) The computability and complexity of optical beam tracing. In: 31st Annual IEEE Symposium on Foundations of Computer Science (FOCS). IEEE, St. Louis, pp 106–114
107. Reif JH, Tyagi A (1997) Efficient parallel algorithms for optical computing with the discrete Fourier transform (DFT) primitive. Appl Opt 36(29):7327–7340
108. Rhodes WT (1981) Acousto-optic signal processing: convolution and correlation. Proc IEEE 69(1):65–79
109. Sawchuk AA, Strand TC (1984) Digital optical computing. Proc IEEE 72(7):758–779
110. Shaked NT, Simon G, Tabib T, Mesika S, Dolev S, Rosen J (2006) Optical processor for solving the traveling salesman problem (TSP). In: Javidi B, Psaltis D, Caulfield HJ (eds) Proc SPIE, Optical Information Systems IV, vol 63110G. SPIE, Bellingham
111. Shaked NT, Messika S, Dolev S, Rosen J (2007) Optical solution for bounded NP-complete problems. Appl Opt 46(5):711–724
112. Shor P (1994) Algorithms for quantum computation: Discrete logarithms ande factoring. In: Proceedings 35th Annual Symposium on Foundations Computer Science. ACM, New York, pp 124–134
113. Sosík P (2003) The computational power of cell division in P systems: Beating down parallel computers? Nat Comput 2(3):287–298
114. Sosík P, Rodríguez–Patón A (2007) Membrane computing and complexity theory: A characterization of PSPACE. J Comput Syst Sci 73(1):137–152
115. Stirk CW, Athale RA (1988) Sorting with optical compare-and-exchange modules. Appl Opt 27(9):1721–1726
116. Stone RV (1994) Optoelectronic processor is programmable and flexible. Laser Focus World 30(8):77–79
117. Stone RV, Zeise FF, Guilfoyle PS (1991) DOC II 32-bit digital optical computer: optoelectronic hardware and software. In: Optical Enhancements to Computing Technology, Proc SPIE, vol 1563. SPIE, Bellingham, pp 267–278
118. Stroke GW, Halioua M, Thon F, Willasch DH (1974) Image improvement in high-resolution electron microscopy using holographic image deconvolution. Optik 41(3):319–343
119. Sullivan DL (1972) Alignment of rotational prisms. Appl Opt 11(9):2028–2032
120. Tromp J, van Emde Boas P (1993) Associative storage modification machines. In: Ambos–Spies K, Homer S, Schöning U (eds) Complexity theory: current research. Cambridge University Press, pp 291–313
121. Turin GL (1960) An introduction to matched filters. IRE Trans Inf Theory 6(3):311–329
122. Upatnieks J (1983) Portable real-time coherent correlator. Appl Opt 22(18):2798–2803
123. VanderLugt A (1964) Signal detection by complex spatial filtering. IEEE Trans Inf Theory 10(2):139–145
124. VanderLugt A (1974) Coherent optical processing. Proc IEEE 62(10):1300–1319
125. VanderLugt A (1992) Optical Signal Processing. Wiley, New York
126. van Emde Boas P (1990) Machine models and simulations. In: van Leeuwen J (ed) Handbook of Theoretical Computer Science, vol A. Elsevier, Amsterdam, chap 1
127. van Leeuwen J, Wiedermann J (1987) Array processing machines. BIT 27:25–43
128. Wang CH, Jenkins BK (1990) Subtracting incoherent optical neuron model – Analysis, experiment and applications. Appl Opt 29(14):2171–2186
129. Wang PY, Saffman M (1999) Selecting optical patterns with spatial phase modulation. Opt Lett 24(16):1118–1120
130. Weaver CS, Goodman JW (1966) A technique for optically convolving two functions. Appl Opt 5(7):1248–1249
131. Woods D (2005) Computational complexity of an optical model of computation. Ph D thesis, National University of Ireland, Maynooth
132. Woods D (2005) Upper bounds on the computational power of an optical model of computation. In: Deng X, Du D (eds) 16th International Symposium on Algorithms and Computation (ISAAC 2005). LNCS, vol 3827. Springer, Heidelberg, pp 777–788
133. Woods D (2006) Optical computing and computational complexity. In: Fifth International Conference on Unconventional Computation (UC'06). LNCS, vol 4135. Springer, pp 27–40
134. Woods D, Gibson JP (2005) Complexity of continuous space machine operations. In: Cooper SB, Löewe B, Torenvliet L (eds) New Computational Paradigms, First Conference on Computability in Europe (CiE 2005). LNCS, vol 3526. Springer, Amsterdam, pp 540–551
135. Woods D, Gibson JP (2005) Lower bounds on the computational power of an optical model of computation. In: Calude CS, Dinneen MJ, Păun G, Pérez–Jiménez MJ, Rozenberg G (eds) Fourth International Conference on Unconventional Computation (UC'05). LNCS, vol 3699. Springer, Heidelberg, pp 237–250
136. Woods D, Naughton TJ (2005) An optical model of computation. Theor Comput Sci 334(1–3):227–258
137. Woods D, Naughton TJ (2008) Parallel and sequential optical computing. In: International Workshop on Optical Super-Computing. LNCS. Springer, Heidelberg
138. Yamaguchi I, Zhang T (1997) Phase-shifting digital holography. Opt Lett 22(16):1268–1270
139. Yokomori T (2002) Molecular computing paradigm – toward freedom from Turing's charm. Nat Comput 1(4):333–390
140. Yu FTS (1996) Garden of joint transform correlators: an account of recent advances. In: Second International Conference on Optical Information Processing. Proc SPIE, vol 2969. SPIE, Bellingham, pp 396–401
141. Yu FTS, Lu T, Yang X, Gregory DA (1990) Optical neural network with pocket-sized liquid-crystal televisions. Opt Lett 15(15):863–865
142. Yu FTS, Jutamulia S, Yin S (eds) (2001) Introduction to information optics. Academic Press, San Diego
143. Zhai H, Mu G, Sun J, Zhu X, Liu F, Kang H, Zhan Y (1999) Color pattern recognition in white-light joint transform correlation. Appl Opt 38(35):7238–7244

Optimization Problems and Algorithms from Computer Science

Heiko Rieger
Theoretical Physics, Universität des Saarlandes, Saarbrücken, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Polymers in a Disordered Environment
Many Repulsive Elastic Lines in Random Media
Vortex Glasses and Loop Percolation
Interfaces and Elastic Manifolds
Random Field Ising Model
The Spin Glass Problem
Potts Free Energy and Submodular Functions
Future Directions
Bibliography

Glossary

Combinatorial optimization The search for an optimal configuration in terms of a cost function on a discrete set of allowed configurations.

Ground state The configuration of a model for a physical system of many interacting degrees of freedom described by a Hamiltonian or energy function that has the lowest energy. Also dented as the global minimum of the energy of the system.

Disordered system A physical system with frozen in or quenched inhomogeneities, usually modeled by an energy function containing parameters that are random numbers obeying in prescribed probability distribution.

Universal properties Properties that do not depend on microscopic details of a physical system, like the critical exponents at a continuous phase transition or fractal dimensions.

Network flows A function defined on the edges of a graph that obeys mass balance constraints at each node. A number of polynomial optimization problems relevant for disordered systems can be formulated as network flow models.

Definition of the Subject

Optimization problems in statistical physics occur whenever the ground state of a classical model for a complex condensed matter system has to be determined, which is necessary for understanding its low temperature properties. In some cases calculating the ground state is an easy task as for instance for the paradigmatic model for a ferromagnet: The configuration of all magnetic moments or spins with the lowest energy is the one, where all spins point in the same direction. But usually the situation is much more complex and the problem of calculating the state with the lowest energy is highly non-trivial. This occurs typically in systems with quenched disorder and/or frustration, which means that their Hamiltonian or energy function consists of competing terms that cannot be satisfied simultaneously. Powerful algorithms from computer science have been devised to find the optimum of complex cost-functions and in some cases this can even be achieved in polynomial time. In recent years many of these algorithms could be successfully applied to physically relevant model systems: to polymers in random media, interface problems in random ferromagnets, magnetic flux-lines in disordered environments, spin glasses, and many more.

Introduction

Solid materials which contain a substantial degree of quenched disorder, so called disordered systems, have been an experimental and a theoretical challenge for physicists for many decades. The different thermodynamic phases emerging in random magnets, the aging properties and memory effects of spin glasses, the disorder induced conductor-to-insulator transition in electronic or bosonic systems, the collective behavior of magnetic flux lines in amorphous high temperature superconductors, and the roughening transition of a disordered charge density wave systems are only a few examples for these fascinating phenomena that occur due to the presence of quenched disorder.

Analytic studies of models for these systems are usually based on perturbation theories valid for weak disorder, on phenomenological scaling pictures or on mean-field approximations. Therefore the demand for efficient numerical techniques that allow the investigation of the model Hamiltonians of disordered systems has always been high. Three facts make life difficult here: 1) The regime, where disorder effects are most clearly seen, are at low temperatures – and are even best visible at zero temperature; 2) the presence of disorder slows the dynamics of theses systems down, they become *glassy*, such that for instance conventional Monte-Carlo or molecular dynamics simulations encounter enormous equilibration problems; 3) any numerical computation of disordered systems has to incorporate an extensive disorder average.

In recent years more and more model systems with quenched disorder were found that can be investigated nu-

merically 1) at zero temperature, 2) without equilibration problems, 3) extremely fast, in polynomial time (for reviews see [1,2,3]). This *is* indeed progress, which became possible by the application of *exact* combinatorial optimization algorithms developed by mathematicians and computer scientists over the last few decades. This gift is not for free: first a mapping of the problem of finding the *exact* ground state of the model Hamiltonian under consideration onto a standard combinatorial optimization problem has to be found. If one is lucky, this problem falls into the class of P-problems, for which polynomial algorithms exist. If not, the intellectual challenge for the theoretical physicist remains to reformulate the model Hamiltonian in such a way that its universality class is not changed but a mapping on a P-problem becomes feasible.

An optimization problem can be described mathematically in the following way: let $\underline{\sigma} = (\sigma_1, \ldots, \sigma_n)$ be a vector with n elements which can take values from a domain $X^n : \sigma_i \in X$. The domain X can be either discrete, for instance $X = \{0, 1\}$ or $X = Z$ the set of all integers (in which case it is an integer optimization problem) or X can be continuous, for instance $X = R$ the real numbers. Moreover, let $\mathcal{H}$ be a real valued function, the cost function or objective, or in physics usually the Hamiltonian or the energy of the system. The *minimization problem* is then:

$$\text{Find } \underline{\sigma} \in X^n, \text{ which minimizes } \mathcal{H} \text{ !}$$

A maximization problem is defined in an analogous way. It is sufficient to consider only minimization problems, since maximizing a function H is equivalent to minimizing $-H$. Minimization problems in which the set X is *countable* are called *combinatorial* [4,5,6]. Optimization methods for real valued variables are treated mainly in mathematical literature and in books on numerical methods, see e. g. [8].

Constraints, must hold for the solution, may be expressed by additional equations or inequalities. An arbitrary value of $\underline{\sigma}$, which fulfills all constraints, is called *feasible*. Usually constraints can be expressed more conveniently without giving equations or inequalities. A famous example is the Traveling Salesman Problem (TSP) [7].

The TSP has attracted the interest of physicist several times. For an introduction, see [9]. The model is briefly presented here. Consider n cities distributed randomly in a plane. Without loss of generality the plane is considered to be the unit square. The minimization task is to find the shortest round-tour through all cities which visits each city only once. The tour stops at the city where it started. The problem is described by

$$X = \{1, 2, \ldots, n\} \tag{1}$$

$$H(\underline{\sigma}) = \sum_{i=1}^{n} d(\sigma_i, \sigma_{i+1}) \tag{2}$$

where $d(\sigma_\alpha, \sigma_\beta)$ is the distance between cities σ_α and σ_β and $\sigma_{n+1} \equiv \sigma_1$. The constraint that every city is visited only once can be realized by constraining the vector $\underline{\sigma}$ to be a permutation of the sequence $[1, 2, \ldots, n]$.

The optimum order of the cities for a TSP depends on their exact positions, i. e. on the random values of the distance matrix d. It is a feature of all problems we will encounter here that they are characterized by various random parameters. Each random realization of the parameters is called an *instance* of the problem. In general, if we have a collection of optimization problems of the same (general) type, we will call each single problem an instance of the general problem.

Because the values of the random parameters are fixed for each instance of the TSP, one speaks of *frozen* or *quenched* disorder. To obtain information about the general structure of a problem one has to average measurable quantities, like the length of the shortest tour for the TSP, over the disorder.

In this article we give an overview of *methods* how to *solve* these problems, i. e. how to find the optimum. Interestingly, there is no single way to achieve this. For some problems it is very easy while for others it is rather hard, this refers to the time you or a computer will need at least to solve the problem, it does not say anything about the elaborateness of the algorithms which are applied. Additionally, within the class of hard or within the class of easy problems, there is no universal method. Usually, even for each kind of problem there are many different ways to obtain an optimum. Once a problem becomes large, i. e. when the number of variables n is large, it is impossible to find a minimum by hand. Then computers are used to obtain a solution. Only the rapid development in the field of computer science during the last two decades has pushed forward the application of optimization methods to many problems from science and real life.

We will review some of the most fruitful applications of polynomial algorithms from the realm of combinatorial optimization to various problems in the statistical physics of disordered systems. The next section presents the application of Dijkstra's algorithm for finding shortest paths in weighted networks to the model of a non-directed polymer in a disordered environment with isotropical correlations. Then, in the 4th and 5th section, we discuss minimum cost flow problems on weighted graphs and its solution via the successive shortest path algorithm and apply it to the entanglement transition of elastic lines in a disordered environment and to the loop percolation transition

in a vortex glass model. In the 6th section we focus on the minimum cut-maximum flow problem and discuss among its many applications the roughening transition of elastic media in a disordered environment. The 7th section is devoted to the random field Ising model and how its ground states can be computed with maximum-flow-minimum-cut techniques. The spin glass problem is presented in the 8th section with a mapping onto minimum weighted matching in two dimensions and a brief outline of branch and cut methods for the higher dimensional case. The 9th section is devoted to finite temperature properties of the random bond Potts model and how its free energy can be computed in the limit of infinite Potts states. An outlook in the 10th section closes this chapter.

Polymers in a Disordered Environment

A well studied model of a single elastic line [10], like an individual polymer or a single magnetic flux line in a type-II superconductor, in a disordered environment is the following: If one excludes overhangs (and by this also self-overlaps) of the elastic lines one can parametrize its configuration by the longitudinal coordinate z. The line configuration can then be described by the transverse coordinate $\mathbf{r}(z)$ as a function of z. The presence of disorder is usually modeled by a random potential energy $V(\mathbf{r}, z)$ and the ground state configuration of the line is highly nontrivial due to the competition between the elastic energy, that tends to straighten the line, and the random energy, that tries to bend the line into positions of favorable energy:

$$\mathcal{H}_{\text{single-line}} = \mathcal{H}_{\text{elastic}} + \mathcal{H}_{\text{random}} = \int_0^H \mathrm{d}z \left\{ \frac{\gamma}{2} \left[\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}z} \right]^2 + V[\mathbf{r}(z), z] \right\} , \quad (3)$$

where H is the longitudinal length (not the proper length) of the line. The random potential energy is a Gaussian variable with prescribed mean and correlations $\langle\langle V[\mathbf{r}, z]\, V[\mathbf{r}', z']\rangle\rangle = g(\mathbf{R} - \mathbf{R}')$, where $\mathbf{R} = (\mathbf{r}, z)$ and $\langle\langle \cdots \rangle\rangle$ denotes the average over the disorder.

A lattice version of this continuum model is the *directed* polymer model: The lines correspond to directed paths on a hyper-cubic lattice that start at a specific lattice site, say $(0, 0, \ldots, 0)$ and proceed only in the $(1, 1, \ldots, 1)$ direction along the bonds. The energy contribution for a path passing bond $\mathbf{i}$ of the lattice is a *positive* random variable $e_{\mathbf{i}}$ and the total energy of a path $\mathcal{P}$ is simply

$$\mathcal{H}^{\text{lattice}}_{\text{single-line}} = \sum_{\mathbf{i} \in \mathcal{P}} e_{\mathbf{i}} = \sum_{\mathbf{i}} e_{\mathbf{i}} n_{\mathbf{i}} , \quad (4)$$

where $n_{\mathbf{i}} = 1$ if the path passes bond $\mathbf{i}$ (i. e. $\mathbf{i} \in \mathcal{P}$) and $n_{\mathbf{i}} = 0$ otherwise.

One is interested in isotropically correlated disorder and consider the problem on a *non-directed* (square) lattice (i. e. paths can pass any bond in both directions) in order not too exclude overhangs right from the beginning. In case of uncorrelated disorder overhangs were shown to be irrelevant [12], but for isotropically correlated disorder this is not clear. The latter is defined to decay algebraically with the spatial distance of the bonds

$$\langle\langle e_{\mathbf{i}} - e_{\mathbf{j}} \rangle\rangle = |\mathbf{R}_{\mathbf{i}} - \mathbf{R}_{\mathbf{j}}|^{2\alpha - 1} , \quad (5)$$

where $\mathbf{R}_{\mathbf{i}}$ spatial position of bond $\mathbf{i}$ and α is the correlation exponent: Note that one expects short-range correlations like $\langle\langle e_{\mathbf{i}} - e_{\mathbf{j}} \rangle\rangle \propto \exp(-|\mathbf{R}_{\mathbf{i}} - \mathbf{R}_{\mathbf{j}}|/\lambda)$ with a finite correlation length λ, to be irrelevant and only long-range correlations like (5) to change the universality class of the system. Increasing α imply stronger correlations, uncorrelated disorder corresponds to $\alpha \to -\infty$. The kind of correlated disorder described by (5) can be realized by generating correlated random numbers are generated using a well-established numerical procedure [11].

Exact ground states of the Hamiltonian (4) or optimal paths are calculated using Dijkstra's algorithm (note that all energies $e_{\mathbf{i}}$ are positive). This simple polynomial algorithm works as follows: Let $V = \{1, \ldots, L^d\}$ be the set of lattice sites and $A = \{(i, j) | i, j \in V \text{ nearest neighbors}\}$ the set of bonds. The algorithm increases successively a subset S of sites for which the optimal path starting at the fixed site s are known. Obviously initially $S := \{s\}$. We denote the energy of the optimal path starting at s and terminating at i with $E(i)$ and since all optimal paths can be constructed via a predecessor list, we keep track of this list, too, via an array pred(i), denoting the predecessor site of site i in a **shortest path** from s to i:

algorithm Dijkstra
begin
  $S := \{s\}$; $\overline{S} := V \backslash \{s\}$;
  $E(s) := 0$, pred$(s) := 0$;
  while $|S| < |V|$ **do**
  begin
    choose (i, j) : $E(j) := \min_{k,m}\{E(k) + e_{(k,m)} | k \in S, m \in \overline{S}, (k, m) \in A\}$;
    $\overline{S} := \overline{S} \backslash \{j\}$; $S := S \cup \{j\}$;
    pred$(j) := i$;
  end
end

In Fig. 1 we show examples of the set $\{i\}$ of lattice sites that are end-points of optimal paths starting from a fixed

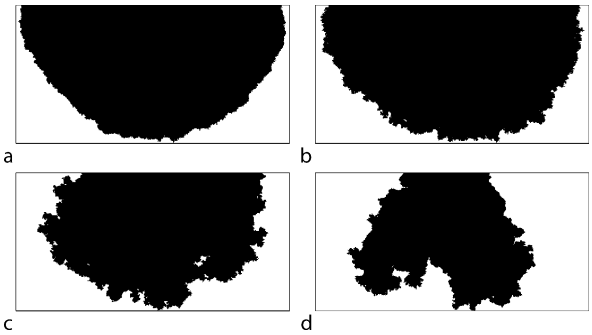

Optimization Problems and Algorithms from Computer Science, Figure 1

Example for the growth front of the non-directed polymer for uncorrelated disorder (a and b) and correlated disorder (c and d; $\alpha = 0.4$). The *black pixels* indicate the lattice sites of the (square) lattice are connected via optimal paths to the offspring (center of the top line) with energy less than a given value (from [13])

initial site and having a total energy $E(i)$ less than a given value $E_{\max}$. For uncorrelated disorder the surface of this set is roughly a semi-circle, whereas for strongly correlated disorder the surface becomes topologically more complicated.

The universal properties of the optimal paths are typically described the scaling of two characteristic quantities: The average transverse fluctuations $\langle\langle \mathbf{r}^2 \rangle\rangle$ and the average energy fluctuations $\langle\langle E^2 \rangle\rangle$. Both are expected to grow algebraically with the the longitudinal distance H between starting point and end point of the paths:

$$\langle\langle \mathbf{r}^2 \rangle\rangle \propto H^{\nu} \quad \text{and} \quad \langle\langle E^2 \rangle\rangle \propto H^{\omega} \,, \tag{6}$$

where ν is called the roughness exponent and ω the energy fluctuation exponent. For uncorrelated disorder ($\alpha \to -\infty$) one knows $\nu = 2/3$ and $\omega = 1/3$. By computing the optimal paths for several thousands of samples for a given disorder correlation exponent α and for a given longitudinal distances H and fitting the resulting data for transverse and energy fluctuations to the expected power laws we can extract the exponents ν and ω (for details see [13]). The resulting estimates in 2d show that the correlations are relevant for $\alpha > 0$ and the roughness exponent increases linearly for $\alpha > 0$ from its value for uncorrelated disorder $\nu = 2/3$. Although the number of overhangs in the optimal paths we computed in the non-directed case increased with α (i. e. increasing correlations) the fraction of bonds contributing to overhangs scaled to zero for all values of α we considered. Hence overhangs appear to be irrelevant also in the presence of correlated disorder.

Many Repulsive Elastic Lines in Random Media

When one puts interacting elastic lines together into a finite system with a given density of lines they will show interesting collective behavior. Examples are the entanglement of magnetic flux lines in high-T_c superconductors in the mixed phase [14] or the entanglement of polymers in materials like rubber [15]. The degree of entanglement of the lines usually manifests itself in various measurable properties like stiffness or shear modulus in the case of polymers and in transport or dynamical properties for magnetic flux lines in superconductors. A theoretical description of these line systems can be based on the single-line Hamiltonian (3) plus an appropriate line interaction term:

$$\mathcal{H}_{\text{many-lines}} = \sum_{i=1}^{N} \mathcal{H}^{(i)}_{\text{single-line}} + \sum_{i<j} \int_0^L \mathrm{d}z \int_0^L \mathrm{d}z' \; V_{\text{int}}[\mathbf{R}_i(z) - \mathbf{R}_j(z')] \,, \tag{7}$$

where $\mathbf{R}_i(z) = (\mathbf{r}_i(z), z)$ is the spatial position of the infinitesimal line segment $\mathrm{d}z$ of the ith line. If the interactions $V_{\text{int}}[\mathbf{R}_i(z) - \mathbf{R}_j(z')]$ are short ranged (i. e. in case of flux lines the screening length small compared to the average line distance) or just hard core repulsive, and the random, δ-correlated disorder potential $V_r[\mathbf{r}_i(z), z]$ in (3) is strong compared to the elastic energy ($\propto \gamma$) this continuum model reduces to a lattice model reminiscent of the single-line lattice model (4):

$$\mathcal{H}^{\text{lattice}}_{\text{many-lines}} = \sum_{\mathbf{i}} e_{\mathbf{i}} n_{\mathbf{i}} \,, \tag{8}$$

where $n_{\mathbf{i}} = 1$ if a line passes bond $\mathbf{i}$ and $n_{\mathbf{i}} = 0$ otherwise and the *positive* random variable $e_{\mathbf{i}}$ is the energy cost for a line segment to occupy bond $\mathbf{i}$. The hard core constraint is thus enforced on the bonds but for the sake of an easier formal description we allow the lines to touch in isolated points, the lattice sites. The lines live on the bonds of a simple cubic lattice with a lateral width L and a longitudinal height H($L \times L \times H$ sites) with free boundary conditions in all directions. Each line starts and ends at an arbitrary position on the bottom respective top planes. The number N of lines threading the sample is fixed by a prescribed density $\rho = N/L^2$. For a single line $N = 1$, one recovers the non-directed polymer model (4). The random bond energies are uniformly distributed over the interval $[0, 1]$.

Note that the allowed configurations of the bond variables $n_{\mathbf{i}}$ are only those that can be identified with lines

threading the samples (or loops inside the sample, which, however, cost energy and therefore do not occur in the ground state), which means that the number of occupied bonds connected to a lattice site that lies neither on the top nor on the bottom plane has always to be even. If we connect all sites on the top to an extra site, called the source, an all sites on the bottom to another extra site, called the target, than the latter statement remains true also for the top an bottom plane. We can now say that N lines start at the source node and terminate at the target node, or, in network flow jargon: The feasible configurations of the variables $n_{\mathbf{i}}$ constitute a flow with zero excess on all lattice sites and an excess $+N$ and $-N$ for the source and target node, respectively.

Thus the determination of the ground state configuration of the N-line problem with the Hamiltonian (8) is a **minimum-cost-flow-problem**, which can be solved with a successive shortest path algorithm [1,2,3]. In essence one starts with the zero flow $n_{\mathbf{i}} = 0$, corresponding to zero lines in the system, and sends successively one unit of flow from the source to the target, corresponding to adding one line after the other to the system. This has to happen with the minimal energy, i. e. along optimal paths, which are calculated using Dijkstra's algorithm that we encountered already in the single line problem discussed in the last section. However, when trying to add a line to a system with a number, say M, of lines already present, the existing line configuration sometimes must be changed to minimize the total energy for $M + 1$ line solution. That becomes feasible by allowing flow to be sent *backwards* on already occupied bonds. By this operation one *gains* energy (whereas occupying an empty bond $\mathbf{i}$ always costs energy $e_{\mathbf{i}} \geq 0$), which means one has to operate on a network that has to be adapted to the existing flow configuration and has negative energies on all occupied bonds. Unfortunately Dijkstra's algorithm works only for positive bond energies, and one has either to use a slower (label-correcting) algorithm to find the optimal paths in a graph with negative edge costs [3] or one has to use the concept of node potentials, by which one can make all energies in the adapted network non-negative without changing the actual shortest paths. This procedure is described in full detail in [3].

The resulting line configuration is then analyzed. One computes the winding angle of all line pairs as indicated in Fig. 2 (c.f. [16]). For each z-coordinate the vector connecting the two lines is projected onto that basal plane (*left* part of Fig. 2). $z = 0$ gives the reference line with respect to which the consecutive vectors for increasing z-coordinate have an angle $\phi(z)$. If the two lines intersect one neglects the intersection point and interpolate between the last and the next point in such a way that the global winding angle is minimized. One defines two lines to be *entangled* when $\phi(z) > 2\pi$. This choice is one that measures entanglement from the topological perspective [17], and comes from the requirement that an entangled pair of lines can not be separated by a suitable linear transformation in the basal plane (i. e. the lines almost always would cut each other, if one were shifted). The precise definition of entanglement is not of major relevance, and the one used is useful since it is the computationally easiest.

Sets or *bundles* of pairwise entangled lines are formed so that a line belongs to a bundle if it is entangled at least with one other line in the set. The topological multi-line-entanglement could be characterized by other measures

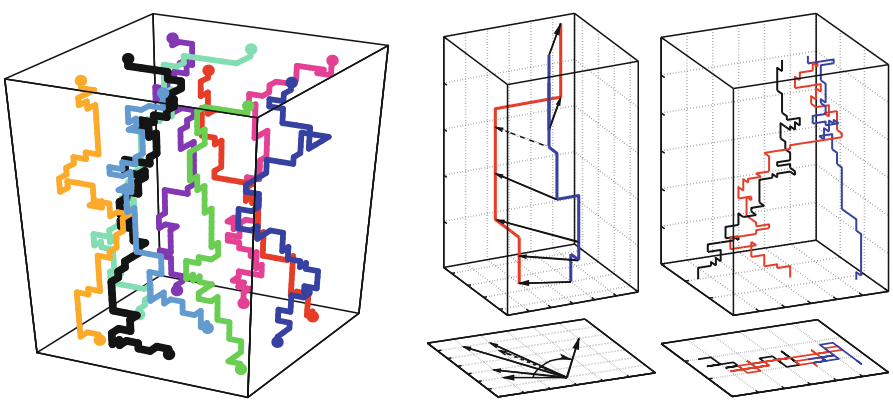

Optimization Problems and Algorithms from Computer Science, Figure 2
Left: **Ground state configuration of a N-line system with $N = 9$ defined by (8). The entry/exit points are fixed in a regular 3×3 array for better visibility.** ***Right:*** **Definition of the winding angle of two flux lines.** ***Right part, top*****: A configuration of three lines that are entangled.** ***Right part, bottom*****: The projection of the line configuration on the basal plane, defining a connected cluster**

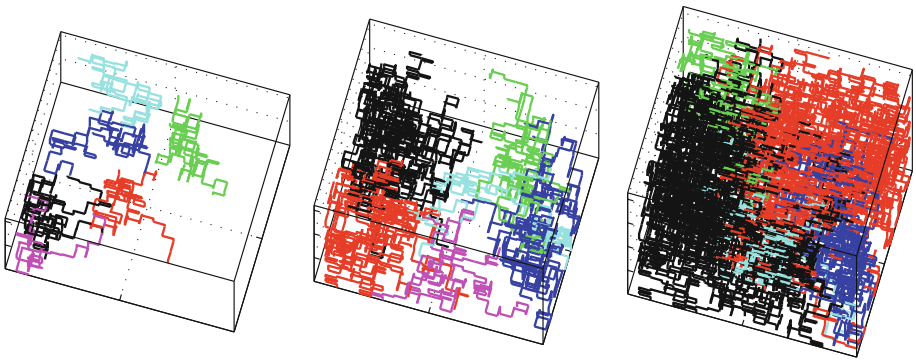

Optimization Problems and Algorithms from Computer Science, Figure 3
Line configurations for different heights H (*from left to right*: $H = 64, 96, 128$), the lateral size $L = 20$, the line density is $\rho = 0.3$. Only the largest line bundles are shown, indicated by a varying gray scale. *Black* denotes the largest cluster, which eventually percolates

as well; the universal properties of the transition will not depend on these. These line bundles are spaghetti-like – i. e. topologically complicated and knotted sets of one-dimensional objects. To study the size distribution of these objects one projects these bundles on the basal plane, as indicated in Fig. 2, where a bundle projects onto a connected cluster. The probability for two lines to be entangled increases with increasing system height. Consequently one would expect that the bundle size increases with H, and therefore also their projections, the clusters. This scenario is exemplified in Fig. 3, for the largest height the largest cluster spans from one side of the system to the other, i. e. it *percolates*.

Hence, for a given line density ρ one expect that for system heights larger than a critical value H_c an system spanning large entangled bundle occurs, containing an infinite number of lines in the limit $L \to \infty$. One calls this an *entanglement transition* occurring at a finite system height H_c. In the projection plane this appears like a percolation transition and in [18] it was shown that this transition is in the same universality class as conventional bond percolation.

Vortex Glasses and Loop Percolation

Another application of the successive shortest path algorithm for minimum-cost-flow-problems is finding the ground state of the Hamiltonian

$$H = \sum_{\mathbf{i}} (n_{\mathbf{i}} - b_{\mathbf{i}})^2$$
$$\text{with the constraint} \quad \forall k: \quad \sum_{l \text{ n.n. of } k} n_{(kl)} = 0\,, \tag{9}$$

where the integer variables $n_{\mathbf{i}}$ live on the bonds $\mathbf{i}$ of a d-dimensional hyper-cubic lattice and $b_{\mathbf{i}} \in [-2\sigma, 2\sigma]$ are real valued quenched random variables with $\sigma \geq 0$ setting the strength of the disorder. The constraint $\sum_{l \text{ n.n. of } k} n_{(kl)} = 0$ means that at all lattice sites k the incoming flow has to balance the outgoing flow, i. e. the flow $\{n_{\mathbf{i}}\}$ is divergence-less. The physical motivation of studying models these kind of models is the following:

In 2d the Hamiltonian (9) occurs for instance in the context of the solid-on-solid (SOS) model on a disordered substrate [19]. The SOS representation of a 2d surface is defined by integer height variables u_k for each lattice site k of a square lattice. The disordered substrate is modeled via random offsets $d_k \in [0, 1]$ for each lattice site, such that the total height at lattice site k is $h_k = u_k + d_k$. The the total energy of the surface is

$$\mathcal{H}_{\text{SOS}} = \sum_{(kl)} (h_k - h_l)^2 = \sum_{(\tilde{kl})} \left(n_{(\tilde{kl})} - b_{(\tilde{kl})}\right)^2 \tag{10}$$

where the first sum runs over all nearest neighbor pairs (kl) of the square lattice and the second sum runs over all bonds $(\tilde{kl})$ of the *dual* lattice (being a square lattice, too), which connect the centers of the elementary plaquettes of the original lattice. A dual bond $(\tilde{kl})$ therefore crosses perpendicularly a bond (kl) connecting neighbors k and l on the original lattice. We define $n_{(\tilde{kl})} = n_k - n_l$ and $d_{(\tilde{kl})} = d_l - d_k$ if l is either the right or the upper neighbor of k (i. e. for $k = (x, y)$ either $l = (x + 1, y)$ or $l = (x, y + 1)$ and $n_{(\tilde{kl})} = n_l - n_k$ and $d_{(\tilde{kl})} = d_k - d_l$ if l is either the left or the lower neighbor of k (i. e. for $k = (x, y)$ either $l = (x - 1, y)$ or $l = (x, y - 1)$. In this way the sum over all four dual bond variables attached to one site of the dual lattice corresponds to the sum of original height variables around an elementary plaquettes in the original lattice: $(n_{(x,y)} - n_{(x,y+1)}) + (n_{(x,y+1)} - n_{(x+1,y+1)}) + (n_{(x+1,y+1)}$

$-n_{(x+1,y)})+(n_{(x+1,y)} - n_{(x,y)}) = 0$, which implies that the flow $\{n_{(\tilde{kl})}\}$ is divergence free as inferred in (9).

In 3d the Hamiltonian (9) is the strong screening limit of the vortex glass model for disordered superconductors [20,21]

$$\mathcal{H}_{\mathrm{VG}} = \sum_{i,j}(n_i - b_i)G_\lambda(\mathbf{r}_i - \mathbf{r}_j)(n_j - b_j)\,, \tag{11}$$

where the integer vortex variables n_i live on the bonds of a simple cubic lattice and have to fulfill the constraint in (9) since they represent magnetic vortex lines that are divergence free. The real valued quenched random variables $b_i \in [-2\sigma, 2\sigma]$ are derived from the lattice curl of a random vector potential ($\sigma \geq 0$ being the strength of the disorder). The 3d vector $\mathbf{r}_i$ denotes the spatial positions of bond i in the lattice and the sum runs over all bond pairs of the lattice (not only nearest neighbors). The lattice propagator $G_\lambda(\mathbf{r})$ has the asymptotic form $G_\lambda(\mathbf{r}) \propto \exp(-|\mathbf{r}|/\lambda)/|\mathbf{r}|$, where λ is the screening length. In the strong screening limit $\lambda \to 0$ only the on-site repulsion survives [20] and gets

$$\mathcal{H}_{\mathrm{VG}}^{\lambda\to 0} = \sum_i (n_i - b_i)^2 \tag{12}$$

which is the Hamiltonian (9) in 3d that we intend to discuss here.

The ground state of (9) can again be computed within polynomial time by a successive shortest path algorithm [3]. As for the N-line problem one starts with a configuration $\{n_{\mathbf{i}}\}$ that optimizes the Hamiltonian in (9) but does not, in general, fulfill the mass balance constraint given in (9). In the N-line problem that was simply the zero-flow $n_{\mathbf{i}} = 0$, which does not fulfill the requirement that the source and the target have excess $+N$ and $-N$, respectively. Here we start with $n_{\mathbf{i}}$ the closest integer to the real number $b_{\mathbf{i}}$ for each bond $\mathbf{i}$. Since this solution violates the mass-balance constraint one successively sends flow from nodes that have an excess flow to nodes that have a deficit along optimal paths that are again found using node potentials (to make all costs non-negative) and

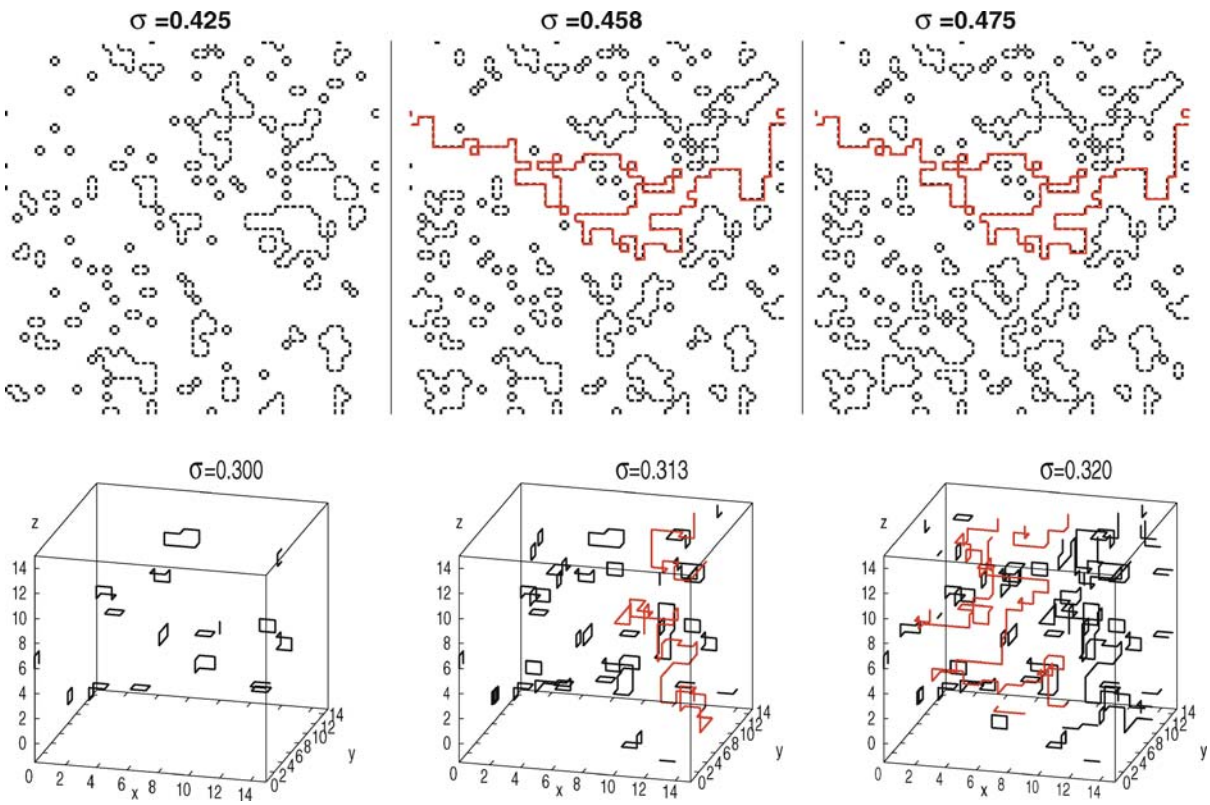

Optimization Problems and Algorithms from Computer Science, Figure 4

Examples of ground state configurations of the Hamiltonian (9) for varying disorder strengths σ (for particular disorder realizations). *Top*: 2d, $L = 50$, the critical disorder strength is $\sigma_c \approx 0.46$; *Bottom*: 3d, $L = 16$, the critical disorder strength is $\sigma_c \approx 0.31$. The occupied bonds ($n_i \neq 0$) are marked *black*, the percolating loop is marked by *light gray (red)*

Dijkstra's algorithm. The details of this algorithm can be found in [1,2,3].

Figure 4 shows three typical ground state configurations for different strength of the disorder σ in 2d and in 3d. For small σ only small isolated loops occur, whereas for larger σ one finds loops that extend through the whole system, they percolate. A finite size scaling study of the underlying percolation transition [22] yields a novel universality class with numerically estimated critical exponents that differ significantly from those for conventional bond- or site-percolation [22].

Interfaces and Elastic Manifolds

A system of strongly interacting (classical) particles or other objects, like magnetic flux lines in a type-II superconductor (as we discussed in Sect. "Many Repulsive Elastic Lines in Random Media" and for which the starting Hamiltonian would given by (7)), or a charge density wave system in a solid, will order at low temperatures into a regular arrangement a lattice (crystal lattice or flux line lattice). Fluctuations either induced by thermal noise (temperature) or by disorder (impurities, pinning centers) induce deviations of the individual particles from their equilibrium positions. As long as these fluctuations are not too strong an expansion of the potential energy around these equilibrium configuration might be appropriate. An expansion up to 2nd order is called the elastic description or elastic approximation, which in a coarse grained form (where the individual particles that undergo displacements from their equilibrium positions do not occur any more and are replaced by a continuum field $\phi(\mathbf{r})$ reads then

$$\begin{aligned}\mathcal{H}_{\text{manifold}} &= \mathcal{H}_{\text{elastic}} + \mathcal{H}_{\text{random}} \\ &= \int \mathrm{d}^d\mathbf{r}\, \left\{\frac{\gamma}{2}|\nabla\phi(\mathbf{r})|^2 + V(\phi(\mathbf{r}),\mathbf{r})\right\} .\end{aligned} \tag{13}$$

The random potential energy is a delta-correlated Gaussian variable with mean zero, $\langle\langle V(\phi,\mathbf{r})V(\phi',\mathbf{r}')\rangle\rangle = D^2\delta(\phi-\phi')\delta(\mathbf{r}-\mathbf{r}')$. The integration extends over the whole space that parameterizes the manifold, for instance $d = 1$ for an elastic line in a random potential, $d = 2$ for an interface or a surface in a disordered environment etc. Note that for $d = 1$ one recovers the single line Hamiltonian (3). The many-line Hamiltonian (7) also allows such an elastic description in the limit, in which the interactions are strong and the the random potential is weak compared to the elastic energy. In this limit the lines will only deviate moderately from a regular, translationally invariant configuration (the Abrikosov flux line lattice). This case is called an elastic periodic medium and one has to modify the φ-part of the disorder correlator such that the Hamiltonian has the correct translational symmetry [26].

Elastic Manifold

The typical example for an elastic manifold in a disordered environment are domain walls in the $d+1$ dimensional random bond ferromagnet $H = -\sum_{\langle ij\rangle} J_{ij}\sigma_i\sigma_j$ ($J_{ij} \geq 0$, random) in which we fix all spins in the

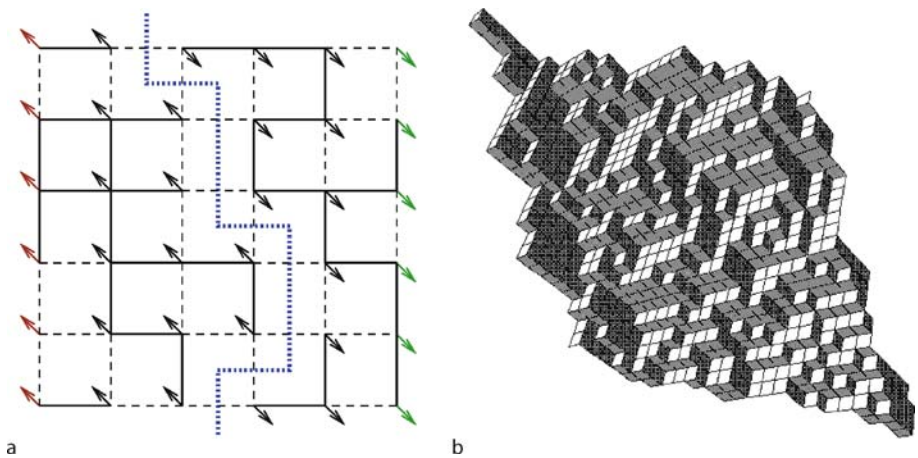

Optimization Problems and Algorithms from Computer Science, Figure 5
Left: **Sketch of a 2d (RBIM) with antiperiodic boundary conditions.** ***Broken lines*** **represent weak bonds,** ***full lines*** **strong bonds, the spin configuration with the lowest energy defines an interface, as indicated, and corresponds to the minimum cut in the corresponding network flow problem.** ***Right:*** **An optimal interface in the 111-direction of a 3d RBIM corresponding to the ground state configuration of a 2d elastic medium with scalar displacement field (from [23])**

lower (upper) plane, i. e. all σ_i with $i = (x_1 1, \ldots, x_d, y)$ and $y = 1\,(y = H)$, to be $\sigma_i = +1\,(-1)$, c.f. Fig. 5. First one maps it onto a flow problem in a capacitated network. One introduces two extra sites, a source node s, which is connected to all spins of the hyperplane $y = 1$ with bonds $J_{s,(x_1,\ldots,x_d,y=1)} = J_\infty$, and a sink node t, which is connected to all spins of the hyperplane $y = H$ with bonds $J_{s,(x_1,\ldots,x_d,y=H)} = J_\infty$. One chooses $J_\infty = 2\sum_{(ij)} J_{ij}$, i. e. strong enough that the interface cannot pass through a bond involving one of the two extra sites. Now we enforce the aforementioned boundary conditions for the spins in the upper and the lower plane by simply fixing $\sigma_s = +1$ and $\sigma_t = -1$. The graph underlying the capacitated network one has to consider is now defined by the set of vertices (or nodes) $N = \{1, \ldots, H \cdot L^d\} \cup \{s, t\}$ and the set of edges (or arcs) connecting them $A = \{(i, j)|i, j \in N,\ J_{ij} > 0\}$.

The capacities $u_i j$ of the arcs (i, j) is given by the bond strength $J_i j$. For any spin configuration $\mathbf{ff} = (\sigma_1, \ldots, \sigma_N)$ one defines $S = \{i \in N|\sigma_i = +1\}$ and $\overline{S} = \{i \in N|\sigma_i = -1\} = N \backslash S$. Obviously $\sigma_s \in S$ and $\sigma_t \in \overline{S}$. The knowledge of S is sufficient to determine the energy of any spin configuration via $H(S) = -C + 2\sum_{(i,j)\in(S,\overline{S})} J_{ij}$ where $(S, \overline{S}) = \{(i, j)|i \in S, j \in \overline{S}\}$. The constant $C = \sum_{(i,j)\in A} J_{ij}$ is irrelevant, i. e. independent of S. Note that $(S, \overline{S})$ is the set of edges (or arcs) connecting S with $\overline{S}$, this means it cuts N in two disjoint sets. Since $s \in S$ and $t \in \overline{S}$, this is a so called s-t-cut-set, abbreviated $[S, \overline{S}]$. Thus the problem of finding the ground state configuration of an interface in the random bond ferromagnet can be reformulated as a **minimum cut** problem

$$\min_{S\subset N} \{H'(S)\} = \min_{[S,\overline{S}]} \sum_{(i,j)\in(S,\overline{S})} J_{ij}\,. \tag{14}$$

in the above defined capacitated network (with $H' = (H + C)/2$). It does not come as a surprise that this minimum cut is *identical* with the interface between the $(\sigma_i = +1)$-domain and the $(\sigma_i = -1)$-domain that has the lowest energy. Actually any s-t-cut-set defines such an interface, some of them might consist of many components, which is of course energetically unfavorable.

A flow in the network G is a set of nonnegative numbers x_{ij} subject to a capacity constraint and a mass balance constraint for each arc

$$0 \le x_{ij} \le u_{ij}$$

$$\text{and} \sum_{\{j|(i,j)\in A\}} x_{ij} - \sum_{\{j|(j,i)\in A\}} x_{ji} = \begin{cases} -v & \text{for } i = s \\ +v & \text{for } i = t \\ 0 & \text{else}\,. \end{cases} \tag{15}$$

This means that at each node everything that goes in has to go out, too, with the only exception being the source and the sink. What actually flows from s to t is v, the value of the flow. The **maximum flow problem** for the capacitated network G is simply to find the flow $\mathbf{x}$ that has the maximum value v under the constraint (15).

Let $\mathbf{x}$ be a flow, v its value and $[S, \overline{S}]$ an s-t-cut. Then, by adding the mass balances for all nodes in S one has $v = \sum_{(i,j)\in(S,\overline{S})} x_{ij} - \sum_{(i,j)\in(\overline{S},S)} x_{ji}$ and since $x_{ij} \le u_{ij}$ and $x_{ji} \ge 0$ the following inequality holds: $v \le \sum_{(i,j)\in(S,\overline{S})} u_{ij} = u[S, \overline{S}]$. Thus the value of any flow $\mathbf{x}$ is less or equal to the capacity of any cut in the network. If one discovers a flow $\mathbf{x}$ whose value equals to the capacity of some cut $[S, \overline{S}]$, then $\mathbf{x}$ is a maximum flow and the cut is a minimum cut. The following implementation of the augmenting path algorithm constructs a flow whose value is equal to the capacity of a s-t-cut it defines simultaneously. Thus it will solve the maximum flow problem (and, of course, the minimum cut problem).

Given a flow $\mathbf{x}$, the residual capacity $r_i j$ of any arc $(i, j) \in A$ is the maximum additional flow that can be sent from node i to node j using the arcs (i,j) and (j,i). The residual capacity has two components: 1) $u_{ij} - x_{ij}$, the unused capacity of arc (i,j), 2) $x_j i$ the current flow on arc (j, i), which one can cancel to increase the flow from node i to j $r_{ij} = u_{ij} - x_{ij} + x_{ji}$. The residual network $G(\mathbf{x})$ with respect to the flow $\mathbf{x}$ consists of the arcs with *positive* residual capacities. An augmenting path is a directed path from the node s to the node t in the residual network. The *capacity of an augmenting path* is the minimum residual capacity of any arc in this path.

Obviously, whenever there is an augmenting path in the residual network $G(\mathbf{x})$ the flow $\mathbf{x}$ is not optimal. This motivates the following generic augmenting path algorithm:

algorithm Ford–Fulkerson
begin
  Initially set $x_{ij} := 0, x_{ji} := 0$ for all $(i, j) \in A$;
  do
    construct residual network R with capacities r_{ij};
    if there is an augmenting path from s to t in G' **then**
    begin
      Let $r_{\min}$ the minimum capacity of r along this path;
      Increase the flow in N along the path
        by a value of $r_{\min}$;
    end
  until no such path from s to t in G' is found;
end

This algorithm is polynomial in the number of lattice sites if the distribution of capacities is discrete (binary for in-

stance). In the general case it has to be improved and there are indeed more efficient algorithms to solve this problem in polynomial time. One of them is the push/relabel algorithm introduced by Goldberg and Tarjan [24]. It determines the maximal flow by successively improving a "preflow". A preflow is an edge function $f(e)$ that obeys the range constraint $0 \leq f(e) \leq w(e)$, but the conservation constraint at each node is relaxed: the sum of the $f(e)$ into or out of a node can be nonzero at internal (physical) nodes. The amount of violation of conservation at each node v give "excesses" $e(v)$. The basic operations of the algorithm, push and relabel, are used to rearrange these excesses. When the preflow can no longer be improved, it can, if desired, be converted to a maximal flow, proving the correctness of the algorithm. For details see [24,25]. It can be applied in the way sketched above to compute universal geometrical properties of elastic manifolds in 2 and 3 dimensions [23].

Periodic Medium

The presence of a periodic background potential, like a crystal potential, has a smoothening effect on the elastic manifold and tends to lock it into one of its minima. The competition between the random potential, that roughens the manifold, and such a periodic potential might lead to a roughening transition [27,28]. In 2d this is actually not the case [29], but in 3d there is as we will see. We consider a lattice version of the Hamiltonian

$$\mathcal{H} = \mathcal{H}_{\text{manifold}} + H_{\text{periodic}}$$
$$\text{with} \quad H_{\text{periodic}} = \int \mathrm{d}^d\mathbf{r}\; V_{\text{periodic}}(\phi(\mathbf{r}))\,, \tag{16}$$

where $V_{\text{periodic}}(\phi) = -\cos\phi$ represents the periodic potential.

We introduce a discrete solid-on-solid (SOS) type interface model for the elastic manifold whose continuum Hamiltonian is given in Eq. (16). Locally the EM remains flat in one of periodic potential minima at $\phi = 2\pi h$ with integer h. Due to fluctuations, some regions might shift to a different minimum with another value of h to create a step (or domain wall) separating domains. To minimize the cost of the elastic and periodic potential energy in Eq. (16), the domain-wall width must be finite, say ξ_o. Therefore, if one neglects fluctuations in length scales less than ξ_o, the continuous displacement field $\phi(\mathbf{r})$ can be replaced by the integer height variable $\{h_{\mathbf{x}}\}$ representing a $(3+1)d$ SOS interface on a simple cubic lattice with sites $\mathbf{x} \in \{1, \ldots, L\}^3$. The lattice constant is of order ξ_o and set to unity. The energy of the interface is given by the Hamiltonian

$$\mathcal{H} = \sum_{\langle \mathbf{x},\mathbf{y}\rangle} J_{(h_{\mathbf{x}},\mathbf{x});(h_{\mathbf{y}},\mathbf{y})} |h_{\mathbf{x}} - h_{\mathbf{y}}| - \sum_{\mathbf{x}} V_R(h_{\mathbf{x}}, \mathbf{x})\,, \tag{17}$$

where the first sum is over nearest neighbor site pairs. After the coarse graining, the step energy $J > 0$ as well as the random pinning potential energy V_R becomes a quenched random variable distributed independently and randomly. Note a periodic elastic medium has the same Hamiltonian as in Eq. (17) with random but periodic J and V_R in h with periodicity p [30]. In this sense, the elastic manifold emerges as in the limit $p \to \infty$ of the periodic elastic medium.

To find the ground state, one maps the 3D SOS model onto a ferromagnetic random bond Ising model in $(3+1)$d hyper-cubic lattice with anti-periodic boundary conditions in the extra dimension [23] (for the 3 space direction one uses periodic boundary conditions instead). The anti-periodic boundary conditions force a domain wall into the ground state configuration of the $(3+1)$d ferromagnet. Note that bubbles are *not* present in the ground state. A domain wall may contain an overhang which is unphysical in the interface interpretation. Fortunately, one can forbid overhangs in the Ising model representation using a technique described in [23]. If the longitudinal and transversal bond strengths are assigned with $J/2$ and $V_R/2$ occurring in Eq. (17), respectively, this domain wall of the ferromagnet becomes equivalent to the ground state configuration of (17) for the interface with the same energy. The domain wall with the lowest energy is then determined exactly by using again the max-flow/min-cost algorithm.

In elastic media described by (17) the tendency of the periodic potential to lock the displacements competes with the roughening effect of the disorder. Analytically a roughening transition was predicted in [28] and the critical exponents could be numerically estimated in three dimensions [30] with the mapping and algorithm described above.

Random Field Ising Model

The random field Ising model (RFIM, for a review see [31,32]) is defined

$$H = -\sum_{(ij)} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i \tag{18}$$

with $\sigma_i = \pm 1$ Ising spins, ferromagnetic bonds $J_{ij} \geq 0$ (random or uniform), (ij) nearest neighbor pairs on a d-dimensional lattice and at each site i a random field $h_i \in R$

that can be positive and negative. The first term prefers a ferromagnetic order, which means it tries to align all spins. The random field, however, tends to align the spins with the field which points in random directions depending on whether it is positive or negative. This is the source of competition in this model.

Let us suppose for the moment uniform interactions $J_{ij} = J$ and a symmetric distribution of the random fields with mean zero and variance h_r. It is established by now that in 3d (and higher dimensions) the RFIM shows ferromagnetic long range order at low temperatures, provided h_r is small enough. In 1d and 2d there is no ordered phase at any finite temperature. Thus in 3d one has a paramagnetic/ferromagnetic phase transition along a line $h_c(T)$ in the h_r-T-diagram.

The renormalization group picture says that the nature of the transition is the same all along the line $h_c(T)$, with the exception being the pure fixed point at $h_r = 0$ and $T_c \sim 4.51\,J$. The RG flow is dominated by a zero temperature fixed point at $h_c(T = 0)$. As a consequence, the critical exponents determining the critical behavior of the RFIM should be all identical along the phase transition line, in particular identical to those one obtains *at zero temperature* by varying h_r alone. Thus to study the universal properties of the phase transition in the RFIM one needs to calculate its ground state.

This optimization task is again equivalent to a maximum flow problem [33,34], as in the interface model discussed in the last section. Historically the RFIM was the first physical model that has been investigated with a maximum flow algorithm [36]. However, here the minimum-cut is not a geometric object within the original system.

To map the ground state problem for the RFIM onto a max-flow-min-cut problem one proceeds in the same way as in the interface problem: One adds to extra nodes s and t and attaches spins with fixed values there (see Fig. 6):

$$\sigma_s = +1 \qquad \text{and} \qquad \sigma_t = -1 \tag{19}$$

One connects all sites with positive random field to the node s and all sites with negative random field to t:

$$J_{si} = \begin{cases} h_i & \text{if } h_i \geq 0 \\ 0 & \text{if } h_i < 0 \end{cases}$$
$$J_{it} = \begin{cases} |h_i| & \text{if } h_i < 0 \\ 0 & \text{if } h_i \geq 0 \end{cases} \tag{20}$$

The a network is constructed with the set of nodes $N = \{1, \cdots, L^d\} \cup \{s, t\}$ and the set of (forward and backward) arcs $A = \{(i, j) | i, j \in N,\ J_{ij} > 0\}$. Each of them has a capacity $u_{ij} = J_{ij}$. The energy or cost function can

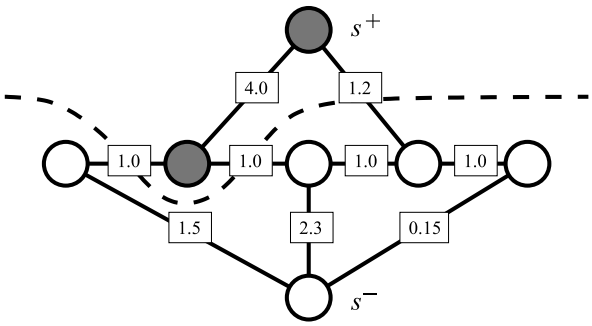

Optimization Problems and Algorithms from Computer Science, Figure 6

Representation of the ground state problem for the RFIM as an RBIM domain wall or minimum-cut problem. The physical spins are the five nodes in the single row in the figure, while the fixed external spins are s^+ and s^-. The physical RFIM coupling $J = 1.0$. A spin with $h_i > 0$ ($h_i < 0$) is connected by an auxiliary coupling of strength h_i ($-h_i$) to s^+ (s^-). The weights of each bond are indicated: the random fields are, *from left to right*, $h = -1.5, +4.0, -2.3, +1.2$, and 0.15. In the ground state, the interfacial energy between up-spin and down-spin domains is minimized, i. e., the spins are partitioned into two sets with minimal total cost for the bonds connecting the two sets. The *dashed* curve indicates the minimal weight cut. The *white (dark)* nodes indicate up (down) spins in the ground state configuration

the be written as

$$E = -\sum_{(i,j)\in A} J_{ij}\sigma_i\sigma_j \tag{21}$$

or, by denoting the set $S = \{i \in N \,|\, S_i = +1\}$ and $\overline{S} = N \backslash S$

$$E(S) = -C + 2\sum_{(i,j)\in(S,\overline{S})} J_{ij} \tag{22}$$

with $C = \sum_{(i,j)\in A} J_{ij}$. The problem is reduced to the problem of finding a minimum s-t-cut as in (14). The difference to the interface problem is that now the extra bonds connecting the two special nodes s and t with the original lattice do not have infinite capacity: they can lie *in* the cut, namely whenever it is more favorable not to break a ferromagnetic bond but to disalign a spin with its local random field. In the extended graph the s-t-cut again forms connected interface, however, in the original lattice (without the bonds leading to and from the extra nodes) the resulting structure is generally *disconnected*, a multi-component interface. Each single component surrounds a connected region in the original lattice containing spins, which all point in the same direction. In other words, they form ferromagnetically ordered domains separated by domain walls given by the subset of the s-t-cut that lies in the original lattice.

In passing we note that diluted Ising antiferromagnets in a homogeneous external field (DAFF) map straightforwardly onto a RFIM if the underlying lattice is bipartite. The 3d DAFF on a simple cubic lattice is defined by

$$H = +\sum_{(ij)} J_{ij}\varepsilon_i\varepsilon_j\sigma_i\sigma_j - \sum_i h_i\varepsilon_i\sigma_i \tag{23}$$

where $\sigma_i = \pm 1$, $J_{ij} \geq 0$, (ij) are nearest neighbor pairs on a simple cubic lattice, and $\varepsilon_i \in \{0, 1\}$ with $\varepsilon_i = 1$ with probability c, representing the concentration of spins. Because of the plus sign in front of the first term in (23) all interactions are antiferromagnetic, the model represents a diluted antiferromagnet, for which many experimental realizations exist (e. g. $Fe_cZn_{1-c}F_2$). Now that neighboring spins tend to point in opposite directions due to their antiferromagnetic interaction a uniform field competes with this ordering tendency by trying to align them all. On a bipartite lattice in zero external field the ground state would be antiferromagnetic, which means that one can define two bipartite sublattices A and B. One defines new spin and field variables via

$$\sigma_i' = \begin{cases} +\sigma_i & \text{for } i \in A \\ -\sigma_i & \text{for } i \in B \end{cases}$$

$$h_i' = \begin{cases} +\varepsilon_i h_i & \text{for } i \in A \\ -\varepsilon_i h_i & \text{for } i \in B\,. \end{cases}$$

Since $\sigma_i'\sigma_j' = -\sigma_i\sigma_j$ for all nearest neighbor pairs (ij) one can write (23) as

$$H = -\sum_{(ij)} J_{ij}'\sigma_i'\sigma_j' - \sum_i h_i'\sigma_i' \tag{24}$$

with $J_{ij}' = J_{ij}\varepsilon_i\varepsilon_j$. This is again a RFIM and ground states can be computed with the max-flow technique.

The main focus of the application of the max-flow-min-cut algorithm to the RFIM is the phase transition in the three-dimensional model occurring at a critical disorder strength h_c at zero temperature, which separates a paramagnetic phase for large disorder strength from a ferromagnetic phase. The maximum flow algorithm has first been used by Ogielski [36] to calculate the critical exponents of the RFIM via the finite size scaling. More accurate estimates were obtained more recently by Middleton and Fisher [35], where also an detailed discussion of the problems and conflicting results about the RFIM universality class is provided. For Gaussian random fields (with variance h^2) they find for the finite size scaling of magnetization $m = [S_i]\text{av}$ and specific heat $c = N^{-1}\mathrm{d}E/\mathrm{d}T$ and

$$m \sim L^{-\beta/\nu}\,,$$
$$c \sim L^{\alpha/\nu}\,, \tag{25}$$

with the magnetization exponent $x = \beta/\nu = 0.012 \pm 0.004$ the correlation length exponent $\nu = 1.37 \pm 0.09$, and the specific heat exponent $\alpha = -0.07 \pm 0.17$. Note that the magnetization exponent is very close to zero, which means that the transition is hard to discriminate from a first order transition. Also the specific heat exponents is close to zero and slightly negative, implying a lack of divergence of the specific heat at the transition.

The Spin Glass Problem

Spin glasses are the prototypes of (disordered) frustrated systems (see [37]). In the models discussed up to now, the frustration was caused by two separate terms of different physical origin (interactions and external fields or boundary conditions). Spin glasses are magnetic systems in which the magnetic moments interact ferro- or antiferromagnetically in a random way, as in the following Edwards–Anderson Hamiltonian for a short ranged Ising spin glass (SG)

$$H = -\sum_{(ij)} J_{ij}\sigma_i\sigma_j\,, \tag{26}$$

where $\sigma_i = \pm 1$, (ij) are nearest neighbor interactions on a d-dimensional lattice and the interaction strengths $J_{ij} \in R$ are unrestricted in sign. In analogy to Eq. (14) one shows that the problem of finding the ground state is again equivalent to finding a minimal cut $[S, \overline{S}]$ in a network

$$\min_{\underline{\mathbf{ff}}} \{H'(\underline{\mathbf{ff}})\} = \min_{[S,\overline{S}]} \sum_{(i,j)\in(S,\overline{S})} J_{ij}\,, \tag{27}$$

again $H' = (H + C)/2$ with $C = \sum_{(ij)} J_{ij}$. However, now the capacities $u_{ij} = J_{ij}$ of the underlying network are *not* non-negative any more, therefore it is *not* a minimum-cut problem and thus it is also not equivalent to a maximum flow problem, which we know how to handle efficiently.

It turns out that the spin glass problem is *much* harder than the questions we have discussed so far. In general (i. e. in any dimension larger than two and also for 2d in the presence of an external field) the problem of finding the SG ground state is $\mathcal{NP}$-complete [42], which means in essence that no polynomial algorithm for it is known (and also that chances to find one in the future are marginal). Nevertheless, some extremely efficient algorithms for it have been developed [38,39], which have a non-polynomial bound for their worst case running-time but which terminate (i. e. find the optimal solution) after a reasonable computing time for pretty respectable system sizes.

Two Dimensions, Planar Graph

First we discuss the only non-trivial case that can be solved with a polynomial algorithm: the two-dimensional Ising SG on a planar graph. This problem can be shown to be equivalent to finding a minimum weight perfect matching, which can be solved in polynomial time. We do not treat matching problems and the algorithms to solve them in this lecture (see [4,40,41]), however, we would like to present the idea [42]. To be concrete let us consider a square lattice with free boundary conditions. Given a spin configuration $\underline{\mathbf{ff}}$ (which is equivalent to $-\underline{\mathbf{ff}}$) we say that an edge (or arc) (i,j) is satisfied if $J_{ij}\sigma_i\sigma_j > 0$ and it is *unsatisfied* if $J_{ij}\sigma_i\sigma_j < 0$. Furthermore we say a plaquette (i. e. a unit cell of the square lattice) is *frustrated* if it is surrounded by an odd number of negative bonds (i. e. $J_{ij} \cdot J_{jk} \cdot J_{kl} \cdot J_{li} < 0$ with i, j, k and l the four corners of the plaquette)). There is a one-to-one correspondence between equivalent spin configurations ($\underline{\mathbf{ff}}$ and $-\underline{\mathbf{ff}}$) and sets of unsatisfied edges with the property that on each frustrated (unfrustrated) plaquette there is an odd (even) number of unsatisfied edges. See Fig. 7 for illustration.

Note that

$$H(\underline{\mathbf{ff}}) = -C + 2 \sum_{\text{unsatisfied edges}} |J_{ij}| \,. \tag{28}$$

which means that one has to minimize the sum over the weights of unsatisfied edges. A set of unsatisfied edges will be constituted by a set of paths (in the dual lattice) from one frustrated plaquette to another and a set of closed circles (see Fig. 7). Obviously the latter always increase the energy so that we can neglect them. The problem of finding the ground state is therefore equivalent to finding the minimum possible sum of the weights of these paths between the frustrated plaquettes. This means that we have to *match* the black dots in the Fig. 7 with one another in an optimal way. One can map this problem on a minimum weight **perfect matching** problem (a perfect matching of a graph $G = (N, A)$ is a set $M \subseteq A$ such that each node has only has only one edge of M adjacent to it). This can be solved in polynomial time (see [42] for further details).

Note that for binary couplings, i. e. $J_{ij} = \pm J$, where $J_{ij} = +J$ with probability p the weight of a matching is simply proportional to the sum of the lengths of the various paths connecting the centers of the frustrated plaquettes, which simplifies the actual implementation of the algorithm. In [43] the 2d $\pm J$ spin glass and the site disordered SG has been studied extensively with this algorithm. The site disordered spin glass is defined as follows: occupy the sites of a square lattice randomly with A (with concentration c) and B (with concentration $1 - c$) atoms. Now define the interactions J_{ij} between neighboring atoms: $J_{ij} = -J$ if on both sites are A-atoms and J_{ij} otherwise.

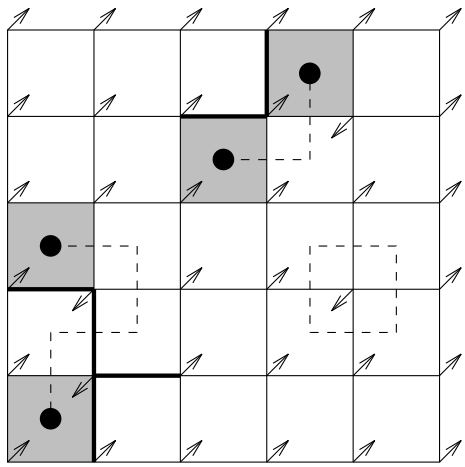

Optimization Problems and Algorithms from Computer Science, Figure 7
Two-dimensional Ising spin glass with ±-J couplings: *Thin* lines, are positive interactions, *thick* lines are negative interactions, ↗ means $\sigma_i = +1$, ↙ means $\sigma_i = -1$, *shaded* faces are frustrated plaquettes, *broken* lines cross unsatisfied edges

The main application of this algorithm is directed towards studying domain walls in spin glasses since they provide informations on the low temperature behavior and the stability of the ground state with respect to thermal fluctuations. Domain walls can be induced by applying two different boundary conditions to the system (usually periodic and anti-periodic), their energy is simply the difference between the energies of the ground states with the two different boundary conditions. The domain wall energy of the two-dimensional spin glass model with Gaussian couplings scales like

$$\Delta E \sim L^{\theta} \,, \tag{29}$$

where the stiffness exponent is $\theta = -0.282$ (see [44] for a survey). The negativity of this exponent indicates the absence of stable spin glass phase at any non-vanishing temperature in the 2d spin glass model. Recently also the fractal properties of the domain walls in 2d spin glasses with Gaussian couplings became important: They have a fractal dimension of $d_f = 1.27(1)$ and it was argued [45] that they might be a realization of a stochastic Loewner evolution (see [46] for a review) realized in disordered systems.

Three Dimensions, Non-planar Graphs

As we mentioned, in any other case except the planar lattice situation discussed above the spin glass problem is

$\mathcal{NP}$-hard. In what follows we would like to sketch the idea of an efficient but non-polynomial algorithm [39]. To avoid confusion with the minimum cut problem we discussed in connection with maximum flows one calls the problem (27) a max-cut problem (since finding the minimum of H is equivalent to finding the maximum of $-H$).

Let us consider the vector space R^A. For each cut $[S, \overline{S}]$ define $\chi^{(S,\overline{S})} \in R^A$, the incidence vector of the cut, by $\chi_e^{(S,\overline{S})} = 1$ for each edge $e = (i, j) \in (S, \overline{S})$ and $\chi_e^{(S,\overline{S})} = 0$ otherwise. Thus there is a one-to-one correspondence between cuts in G and their $\{0, 1\}$-incidence vectors in R^A. The *cut-polytope* $P_C(G)$ of G is the convex hull of all incidence vectors of cuts in G: $P_C(G) = \text{conv}\{\chi^{(S,\overline{S})} \in R^A \mid S \subseteq A\}$. Then the max-cut problem can be written as a *linear program*

$$\max \{\underline{u}^T \underline{x} \mid \underline{x} \in P_C(G)\} \tag{30}$$

since the vertices of $P_C(G)$ are cuts of G and vice versa. Linear programs usually consist of a linear cost function $\underline{u}^T \underline{x}$ that has to be maximized under the constraint of various inequalities defining a polytope in R^n (i. e. the convex hull of finite subsets of R^n) and can be solved for example by the simplex method, which proceeds from corner to corner of that polytope to find the maximum (see e. g. [40,41,48]). The crucial problem in the present case is that it is $\mathcal{NP}$-hard to write down all inequalities that represent the cut polytope $P_C(G)$.

It turns out that also *partial* systems are useful, and this is the essential idea for an efficient algorithm to solve the general spin glass problem as well as the traveling salesman problem or other so called mixed integer problems (i. e. linear programs where some of the variables x are only allowed to take on some integer values, like 0 and 1 in our case) [7,47]. One chooses a system of linear inequalities L whose solution set $P(L)$ contains $P_C(G)$ and for which $P_C(G) = \text{convex hull } \{\mathbf{x} \in P(L) | x \text{ integer}\}$. In the present case these are $0 \leq x \leq 1$, which is trivial, and the so called cycle inequalities, which are based on the observation that all cycles in G have to intersect a cut an even number of times. The most remarkable feature of this set L of inequalities is the following:

The separation problem for a set of inequalities L consists in either proving that a vector x satisfies all inequalities of this class or to find an inequality that is violated by $\mathbf{x}$. A linear program can be solved in polynomial time if and only if the separation problem is solvable in polynomial time [49]. The separation problem for the cycle inequalities can be solved in polynomial time by the *cutting plane algorithm* which, starting from some small initial set of inequalities, generates iteratively new inequalities until the optimal solution for the actual subset of inequalities is feasible. Note that one does not solve this linear program by the simplex method since the cycle inequalities are still too numerous for this to work efficiently.

Due to the insufficient knowledge of the inequalities that are necessary to describe $P_C(G)$ completely, one may end up with a non-integral solution $\mathbf{x}^*$. In this case one *branches* on some fractional variable x_e (i. e. a variable with $x_e^* \notin \{0, 1\}$), creating two subproblems in one of which x_e is set to 0 and in the other x_e is set to 1. Then one applies the cutting plane algorithm recursively for both subproblems, which is the origin of the name *branch-and-cut*. Note that in principle this algorithm is not restricted to any dimension, boundary conditions, or to the fieldless case. However, there are realizations of it that run fast (e. g. in 2d) and others that run slow (e. g. in 3d) and it is ongoing research to improve on the latter, for an overview over the current status see [47].

Potts Free Energy and Submodular Functions

The problem addressed in this chapter is not a low temperature problem but concerns the computation of the free energy of a Potts model (see [50] for a review) at *any temperature*, including some phase transition temperatures. To transform the problem of computing the free energy into an optimization problem (i. e. find a minimum in a finite set), one needs to take some limit. Usually this is a zero temperature limit as it was for all applications discussed so far in this article. Here this will be the limit of an *infinite number of states*.

Consider the q-state Potts model on a d-dimensional hyper-cubic lattice with periodic boundary conditions defined by the Hamiltonian:

$$H = -\sum_{\langle ij \rangle} J_{ij} \delta(\sigma_i, \sigma_j) \,, \tag{31}$$

where σ_i are q-state Potts variables ($\sigma_i \in \{1, \ldots, q\}$ located at lattice sites i, the sum goes over all nearest neighbor pairs $\langle ij \rangle$ of the lattice, and $J_{ij} > 0$ are ferromagnetic couplings (not that $\delta(\sigma, \sigma')$ is the Kronecker-delta, which means $\delta(\sigma, \sigma') = 1$ for $\sigma = \sigma'$ and $\delta(\sigma, \sigma') = 0$ for $\sigma \neq \sigma'$). The case $q = 2$ corresponds to the Ising model. In the random bond Potts model, which is of interest here, the couplings J_{ij} are random variables. In $d \leq 2$ dimensions the Potts model has phase transition at some critical temperature T from a paramagnetic to a ferromagnetic phase. Thermodynamic properties of the q-state Potts model are computed via its partition function

$$\mathcal{Z} = \sum_{\{\underline{\sigma}\}} \exp\left(\sum_{ij} -\beta J_{ij} \delta(\sigma_i, \sigma_j)\right) \,. \tag{32}$$

The first sum runs over all possible spin configuration, i. e. it involves q^N terms, where N is the number of spins in the system and $\beta = 1/T$ is the inverse temperature.

In the so-called random cluster representation [51] the partition sum can be written as a sum over all subsets $U \subseteq E$ of the set of edges (or bonds)

$$\begin{aligned} \mathcal{Z} &= \sum_{\{\sigma\}} \prod_{ij} \exp\left(-\beta J_{ij} \delta(\sigma_i, \sigma_j)\right) \\ &= \sum_{\{\sigma\}} \prod_{ij} \left(1 + v_{ij} \delta(\sigma_i, \sigma_j)\right) \end{aligned}$$

where $v_{ij} = \exp(\beta K_{ij}) - 1$. Note that the Kronecker-delta can only take on the values zero and one by which it is possible to identify $\exp(J\delta) = 1 + \delta(\exp(J) - 1) = 1 + v\delta$. Again one can regard the lattice as a graph $G = (V, E)$, where the sites and the bonds of the lattice are the vertices V and the edges E of the graph. Then a careful bookkeeping of the terms in the development of the above expression leads to:

$$Z = \sum_{G' \subseteq G} q^{c(G')} \prod_{e \in G'} v_e \,, \tag{33}$$

where G' denotes any subgraph of G, i. e. a graph, possibly not connected (but all vertices are kept), where some edges of G have been deleted (there are 2^m subgraphs where m is the number of edges of G). $c(G')$ is the number of connected components of the subgraph G'. For example for the empty subgraph $G' = \emptyset$ the number of connected components is the number of sites, while for $G' = G$ it is one. The product in (33) is over all the edges in G' with the convention that the product over an empty set is one. If the parameter β is small (i. e. high temperature) then the parameters v_{ij} are small and, summing in (33), only the subgraphs with few edges provides an approximation to the partition function: this is a high temperature development. Note also the way the parameter q appears in (33): it can be extended to non integer values, relating the Potts model to other problems (percolation, etc ...) [58].

Following [52] one can map the computation of the partition function Z of any ferromagnetic Potts model in the limit $q \to \infty$ onto an optimization problem by introducing another parametrization of the couplings with new variables w_e defined by

$$v_e = q^{w_e} \,.$$

Inserting this expression in (33) one gets $Z = \sum_{G' \subseteq G} q^{c(G') + \sum_{e \in G'} w_e}$, and defining $f(G) = c(G) + \sum_{e \in G} w_e$:

$$Z = \sum_{G' \subseteq G} q^{f(G')} \,.$$

In the limit $q \to \infty$ only the subgraphs $G^\star$ maximizing $f(G)$ will contribute, and computing the partition function of the Potts model in the infinite number of states limit amounts to finding the subgraphs G' of the graph G maximizing the function f, i. e. minimizing the function [52]:

$$f_P(G') = -\left(c(G') + \sum_{e \in G'} w_e\right) \,. \tag{34}$$

It turns out that this function has a property which allows to minimize it very efficiently: it is a *submodular function*.

Submodular Functions

The concept of a submodular function in discrete optimization appears to be in several respects analogous to that of a convex function in continuous optimization. In many combinatorial theorems and problems, submodularity is involved, in one form or another, and submodularity often plays an essential role in a proof or an algorithm. Moreover, analogous to the fast methods for convex function minimization, it turns out that submodular functions can also be minimized fast, i. e. in polynomial time.

Submodularity is a special property of *set functions*, which are defined as follows: Let V be a finite set and $2^V = \{X \mid X \subseteq V\}$ be the set of all the subsets of V. A function $f\colon 2^V \to \mathbb{R}$ is called a set function.

Now a set function f is **submodular** if for all subsets $A \subseteq V$ and $B \subseteq V$:

$$f(A) + f(B) \geq f(A \cap B) + f(A \cup B) \,. \tag{35}$$

It is simple to show that a function f is submodular if and only if for any subsets $S \subseteq R \subseteq V$ and for any $x \in V$:

$$f(S \cup \{x\}) - f(S) \geq f(R \cup \{x\}) - f(R) \,. \tag{36}$$

This means intuitively that adding an element to a "small" ensemble S (since $S \subseteq R$) has more effect than adding to a "large" ensemble R.

The function (34) $f_P(A) = -(c(A) + w(A))$ is submodular, because the function $-c(A)$ is submodular (and the function $w(A)$ is modular: Take two sets of edges $A \subseteq B$ and an edge e. Inspecting the three possible cases: $e \in A$, $e \notin A$ and $e \in B$, $e \notin A$ and $e \notin B$ one sees that $c(A \cup \{e\}) - c(A) \leq c(B \cup \{e\}) - c(B)$, which is the reverse of (36), so that the function $-c$ is a submodular function. Note that $c(E')$ with $E' \subseteq E$ counts the number of connected components of the graph G' that contains *all* vertices V of the complete graph but only the edges in E'. Thus adding an edge will never increase the number of components.

On the other hand it is straightforward to see that the function $w(G) = \sum_{e \in G} w_e$ verifies $w(A \cup C) + w(A \cap C) = w(A) + w(C)$. It is a so-called *modular* function. Consequently the function (34) f_P is a submodular function. In summary we are looking for the sets of edges minimizing the submodular function f_P for which a *strongly polynomial* algorithm has been recently discovered.

In passing we note that we encountered other examples of submodular functions already in the preceding sections, namely the function that defines the costs of cuts in a graph with positive edge weights, which occurs the interface problem and the random field Ising model in the last sections: Take a graph $G = (V, E)$ and define C to be a function of the subsets of the V and $C(U \subseteq V)$ is the number of edges having exactly one end in U. This function can be generalized to the case where the edges are directed and weighted, i. e. each edge carries an arrow and a positive number. The function $C(U \subseteq V)$ is then the sum of the weight of the edges having the beginning vertex in U and the ending vertex not in U. This kind of function is generally called a "cut" and is submodular.

Minimization of Submodular Function

The minimization of any submodular function can be done in polynomial time. This was first published in reference [54] in 1981. In this paper the authors utilize the so-called ellipsoid method. However this method is not a combinatorial one and is far from being efficient. In that respect this result was not quite satisfactory at least for the practical applications. Eighteen years later, Iwata–Fleischer–Fujishige [55], and independently Schrijver [56] discovered a combinatorial method which is fully satisfactory from the theoretical, as well as from the practical, point of view.

The general method uses a mathematical programming formulation. The problem is algebraically expressed as a linear program, i. e. a set of variables y_S associated to each subset $S \subset V$ is introduced, these variables are subjected to constraints and a linear function F of these variables is to be minimized. The constraints include a set of linear equations and the condition that each of the y_S is zero or one. This last condition is in general extremely difficult to realize. However, it turns out that a theorem due to Edmonds [57] indicates this condition can be simply dropped, and that automatically the set of values y_S which minimize F will all be zero or one! Actually only one variable $y_{S^\star} = 1$ will be non zero and it is precisely associated to the optimal set. Combined with the dual version of this linear program, it provides a characterization of the optimal set.

The general algorithm mentioned above can be applied to minimize (34), however, due to the specific form of the function to minimize, a more suitable method does exist. For this a property that is true for any submodular function is useful. To emphasize that the function f to minimize is defined on all the subsets of a set E we will label f with the index E as f_E. Let us now consider a subset $F \subseteq E$; one can define a set function on F by $f_F(A) = f_E(A)$ for any $A \subseteq F$. If the function f_E is submodular then its restriction f_F is also submodular. We have the following property:

Let $F \subseteq E$ and $e \in E$, if A_F is an optimal set of the set function f_F defined on F, then there will be an optimal set $A_{F \cup \{e\}}$ of the function $f_{F \cup \{e\}}$ defined on $F \cup \{e\}$ such that $A_F \subseteq A_{F \cup \{e\}}$.

To make the notation simpler we denote the function $f_{F \cup \{e\}}$ on $F \cup \{e\}$ by f_1. Let A be an optimal set of f_F on F and B an optimal set of f_1 on $F \cup \{e\}$. One has

$$f_1(A \cup B) \leq f_1(A) + f_1(B) - f_1(A \cap B) \tag{37}$$

since f_1 is submodular. But $f_1(A) = f_F(A)$ and $f_1(A \cap B) = f_F(A \cap B)$ since both A and $A \cap B$ are in A. Since A is an optimal set one has $f_F(A) \leq f_F(A \cap B)$ and consequently $f_1(A) - f_1(A \cap B) \leq 0$. Inserting this last inequality into (37) one finds that $f_1(A \cup B) \leq f_1(B)$ which proves that $A \cup B$ is one of the optimal sets (Q.E.D.).

This property has an important consequence. Indeed let us suppose that the optimal set has been found for a subset F of E. Then all the elements of E which have been selected as belonging to the optimal set of F will still belong to one optimal set of all the sets $G \supseteq F$. In other words, let us find the optimal set for $\{e_0, e_1\}$ where e_0 and e_1 are *arbitrary* elements of E; then if we find that any of these two elements belongs to the optimal set, it will belong to one optimal set for $F \subseteq E$! Such an algorithm which makes a definitive choice at each step is called a *greedy* algorithm.

Based on this observation an efficient algorithm for the minimization of (34) was developed in [59], see also [60].

Results

The algorithm based on the ideas mentioned before and presented in detail in [59,60], was applied to various two dimensional and three dimensional lattices. A realization of the disorder is chosen accordingly to a probability distribution. In practice all the weights $w(e)$ on the edge e are rational numbers with a common integer denominator q. In other words, we choose an integer $p(e)$ for each

edge and set $w(e) = p(e)/q$. To work only with integers one maximizes the product qf:

$$qf(A) = qC(A) + \sum_{e \in A} p(e) \,.$$

It is clear that if q is small compare to all the $p(e)$, then all the weights $w(e)$ will be large and the optimal set will be the set of all edges. On the contrary if q is large all the weights will be small and the optimal set will be empty. These two situations are easy to handle. Between this two limits the optimal set grows, and for a precise value q_c of q, which depends on the lattice, the optimal set percolates. This value corresponds to a phase transition. Depending on the lattice under consideration and on the distribution of the random variables $p(e)$ this transition can be first or second order.

In Fig. 8, one optimal set is shown for a lattice where each edge carries a weight 1/6 or 5/6 with probability one half (i. e. it is a critical point). The edges from the optimal set belonging to the percolation cluster are shown in black, while the others are shown in gray. The percolation cluster, which is the largest connected component in the optimal subgraph $G' \subseteq G$ is fractal with a fractal dimension $d_f = 1.809$ that is related to the critical exponent $x = \beta/\nu$ for the magnetization of the random bond $q \to \infty$ Potts model (31) in two dimensions via $x = 2 - d_f = 0.191$. Surprisingly this agrees within the error bars with the magnetization exponent $x = (3 - \sqrt{5})/4$ of the random transverse Ising chain [62], which is a one-dimensional quantum spin model. A discussion of this observation and details of the computations can be found in [61].

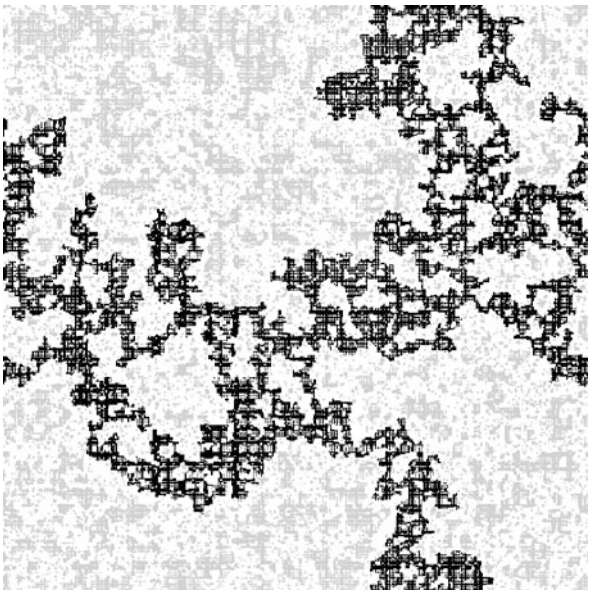

Optimization Problems and Algorithms from Computer Science, Figure 8

A 512 × 512 lattice. The edges of the optimal set belonging to the percolating cluster are shown in *black*, and the edges of the optimal set not belonging to the optimal set are in *gray* (from [60])

Future Directions

We have reviewed several applications of polynomial optimization algorithms from computer science to disordered systems in statistical physics. They were used extensively in the recent years to compute numerically universal properties like critical exponents, domain wall exponents and geometrical features like roughness and stiffness with much higher precision than with Monte-Carlo methods, which suffer notoriously from equilibration problems. A number of important issues, which were controversially debated within different analytical could be clarified, numerically, in this way – as for instance the nature of the low temperature phase of the superrough phase in the two-dimensional Bragg glass [19,63], the absence of a stable glass phase in the strongly screened vortex glass model [21] and the issue of many states in various two-dimensional glassy models [64]. Other questions still remain to be answered, as for example the phenomenon of an apparent non-universality in the three-dimensional random field Ising model [65].

NP-hard problems occurring in the statistical physics of disordered systems, still remain a challenge: Examples are the computation of ground states of spin glass models on non-planar graphs, like the three-dimensional spin glass or the random field Potts model for three or more Potts states [66]. Stochastic optimization techniques like hysteretic optimization [67] or extremal optimization [68] have reached a high level of sophistication but naturally suffer from the lack of a proof of optimality of the resulting solution. Progress in the development of exact and efficient algorithm that can handle sufficiently large system sizes to perform a reliable finite size scaling analysis is being made [47] and highly rewarding.

The cross-fertilization between computer science and statistical physics is also fruitful in the other direction: Phase transitions that occur in some combinatorial optimization problems like the satisfiability problem (SAT) were studied intensively in recent years by physicists and remarkable progress has been achieved in understanding it and inventing efficient algorithms. These developments were not covered in this article, excellent introductions can be found in [69].

Bibliography

Primary Literature

1. Rieger H (1998) Frustrated systems: Ground state properties via combinatorial optimization. In: Kertesz J, Kondor I (eds) Lect Note Phys 501:122–158
2. Alava M, Duxbury P, Moukarzel C, Rieger H (2000) Exact combinatorial algorithms: Ground states of disordered systems. In: Domb C and Lebowitz JL (eds) Phase Transit Crit Phenom 18:141–317
3. Hartmann A, Rieger H (2002) Optimization in physics. Wiley VCH, Darmstadt
4. Papadimitriou CH Steiglitz K (1998) Combinatorial Optimization. Dover Publications, Mineola (NY)
5. Cook WJ, Cunningham WH, Pulleyblank WR, Schrijver A (1998) Combinatorial Optimization. Wiley, New York
6. Korte B, Vygen J (2000) Combinatorial Optimization. Springer, Berlin
7. Lawler EL, Lenstra JK, Rinnooy Kan AHG, Shmoys DB (1990) The Travelling Salesman Problem. Wiley, Chichester
8. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1995) Numerical Recipes in C. Cambridge University Press, Cambridge
9. Kirkpatrick S, Gelatt CD Jr, Vecchi MP (1983) Optimization by simulated annealing. Science 220:671
10. Halpin-Healy T, Zhang Y-C (1995) Kinetic roughening phenomena, stochastic growth directed polymers and all that – aspects of multidisciplinary statistical-mechanics. Phys Rep 254:215
11. Peng C-K, Havlin S, Schwartz M, Stanley HE (1991) Directed-polymer and ballistic-deposition growth with correlated noise. Phys Rev A 44:2239; Pang N-N, Yu Y-K, Halpin-Healy T (1995) Interfacial kinetic roughening with correlated noise. Phys Rev E 52:3224
12. Marsili M, Zhang Y-C (1998) Overhangs in interface growth and ground-state paths. Phys Rev E 57:4814; Schwartz N, Nazaryev AL, Havlin S (1998) Optimal path in two and three dimensions. Phys Rev E 58:7642
13. Schorr R, Rieger H (2003) Universal properties of shortest paths in isotropically correlated random potentials. Europ Phys J 33:347
14. For a review see Blatter G et al (1994) Vortices in high-temperature superconductors. Rev Mod Phys 66:1125
15. Doi M, Edwards SF (1986) The Theory of Polymer Dynamics. Oxford University Press, Oxford
16. Drossel B, Kardar M (1996) Winding angle distributions for random walks and flux lines. Phys Rev E 53:5861
17. Bikbov R, Nechaev S (2001) Topological Relaxation of Entangled Flux Lattices: Single versus Collective Line Dynamics. Phys Rev Lett 87:150602
18. Petäjä V, Alava M, Rieger H (2004) Entanglement transition of elastic lines in a strongly disordered environment. Europhys Lett 66:778
19. Rieger H, Blasum U (1997) Ground state properties of solid-on-solid models with disordered substrates. Phys Rev B 55:7394R; Pfeiffer F, Rieger H (2000) Dislocations in the ground state of the solid-on-solid model on a disordered substrate. J Phys A 33:2489
20. Bokil HS, Young AP (1995) Absence of a phase transition in a three-dimensional vortex glass model with screening. Phys Rev Lett 74:3021
21. Kisker J, Rieger H (1998) Application of a minimum cost flow algorithm to the three-dimensional gauge glass model with screening. Phys Rev B 58:R8873; Pfeiffer F, Rieger H (1999) Numerical study of the strongly screened vortex glass model in an external field. Phys Rev B 60:6304
22. Pfeiffer FO, Rieger H (2002) Superconductor-to-normal phase transition in a vortex glass model: a new percolation universality glass. J Phys C 14:2361; Pfeiffer FO, Rieger H (2003) Critical properties of loop percolation models with optimization constraints. Phys Rev E 67:056113
23. Middleton AA (1995) Numerical results for the ground-state interface in a random medium. Phys Rev E 52:R3337; McNamara D, Middleton AA, Zeng C (1999) Simulation of the zero-temperature behavior of a three-dimensional elastic medium. Phys Rev B 60:10062
24. Goldberg AV, Tarjan RE (1988) A new approach to the maximum-flow problem. J Assoc Comput Mach 35:921
25. Ahuja RK, Magnati TL, Orlin JB (1993) Network Flows. Prentice Hall, London
26. Nattermann T (1990) Scaling approach to pinning: Charge density waves and giant flux creep in superconductors. Phys Rev Lett 64:2454; Giarmachi T, Le Doussal P (1994) Elastic theory of pinned flux lattices. Phys Rev Lett 72:1530; (1995) Phys Rev B 52:1242
27. Bouchaud J-P, Georges A (1992) Competition between lattice pinning and impurity pinning: Variational theory and physical realizations. Phys Rev Lett 68:3908
28. Emig T, Nattermann T (1997) A new disorder-driven roughening transition of charge-density waves and flux-line lattices. Phys Rev Lett 79:5090; (1999) Disorder driven roughening transitions of elastic manifolds and periodic elastic media. Eur J Phys B 8:525
29. Seppälä ET, Alava MJ, Duxbury PM (2001) Intermittence and roughening of periodic elastic media. Phys Rev E 63:036126
30. Noh JD, Rieger H (2001) Disorder driven critical behavior of periodic elastic media in a crystal potential. Phys Rev Lett 87:176102; (2002) Numerical study of the disorder-driven roughening transition in an elastic manifold in a periodic potential. Phys Rev E 66:036117
31. Rieger H (1995) Monte Carlo simulations of Ising spin glasses and random field systems. In: Annual Reviews of Computational Physics II. World Scientific, Singapore, pp 295–341
32. Nattermann T (1998) In: Young AP (ed) Spin Glasses and Random Fields. World Scientific, Singapore
33. Anglés d'Auriac JC, Preissman M, Rammal R (1985) The random field Ising-model - algorithmic complexity and phase-transition. J Phys (France) Lett 46:L173
34. Barahona F (1985) Finding ground-states in random-field Ising-ferromagnets. J Phys A 18:L673
35. Middleton AA, Fisher DS (2002) Three-dimensional random-field Ising magnet: Interfaces, scaling, and the nature of states. Phys Rev B 65:13411
36. Ogielski AT (1986) Integer optimization and zero-temperature fixed point in Ising random-field systems. Phys Rev Lett 57:1251
37. Kawashima N, Rieger H (2004) In: Diep HT (ed) Frustrated Spin Systems. World Scientific, Singapore
38. Grötschel M, Jünger M, Reinelt G (1985) In: van Hemmen L, Morgenstern I (eds) Heidelberg Colloqium on Glassy dynamics and Optimization. Springer, Heidelberg
39. de Simone C, Diehl M, Jünger M, Mutzel P, Reinelt G, Rinaldi G (1995) Exact ground-states of Ising spin-glasses - new exper-

imental results with a branch-and-cut algorithm. J Stat Phys 80:487
40. Lawler EL (1976) Combinatorial optimization: Networks and matroids. Holt, Rinehart and Winston, New York
41. Derigs U (1988) Programming in networks and graphs. In: Springer Series: Lecture Notes in Economics and Mathematical Systems, vol 300. Springer, Berlin
42. Barahona F (1982) On the computational-complexity of Ising spin-glass models. J Phys A 15:3241; Barahona F, Maynard R, Rammal R, Uhry JP (1982) Morphology of ground-states of two-dimensional frustration model. J Phys A 15:673
43. Kawashima N, Rieger H (1997) Finite size scaling analysis of exact ground states for $\pm J$ spin glass models. Europhys Lett 39:85
44. Hartmann AK, Young AP (2002) Large-scale low-energy excitations in the two-dimensional Ising spin glass. Phys Rev B 66:094419; Hartmann AK, Bray AJ, Carter AC, Moore MA, Young AP (2002) Stiffness exponent of two-dimensional Ising spin glasses for nonperiodic boundary conditions using aspect-ratio scaling. Phys Rev B 66:224401
45. Amoruso C, Hartmann AK, Hastings MB, Moore MA (2006) Conformal invariance and stochastic Loewner evolution processes in two-dimensional Ising spin glasses. Phys Rev Lett 97:267202; Bernard D, LeDoussal P, Middleton AA (2007) Possible description of domain walls in two-dimensional spin glasses by stochastic Loewner evolutions. Phys Rev B 76:020403(R)
46. Cardy J (2005) SLE for theoretical physicists. Ann Phys 318:81; Bauer M, Bernard D (2006) 2D growth processes: SLE and Loewner chains. Phys Rep 432:115
47. Liers F, Jünger M, Reinelt G, Rinaldi G (2004) Computing exact ground states of hard Ising spin glass problems by branch-and-cut. In: Hartmann A, Rieger H (eds) New optimization algorithms in physics. Wiley, Berlin
48. Chvátal V (1983) Linear programming. Freeman, San Francisco
49. Grötschel M, Lovász L, Schrijver A (1988) Geometric algorithms and combinatorial optimization. Springer, Berlin
50. Wu FY (1982) The Potts Model. Rev Mod Phys 54:235
51. Kasteleyn PW, Fortuin CM (1969) Phase transitions in lattice systems with random local properties. J Phys Soc Jpn 46:11
52. Juhász R, Rieger H, Iglói F (2001) The random-bond Potts model in the large-q limit. Phys Rev E 64:056122
53. Schrijver A (2003) Combinatorial Optimization – Polyhedra and Efficiency, vol B. Springer, Berlin
54. Grötschel M, Lovász L, Schrijver A (1981) The ellipsoid method and its consequences in combinatorial optimization. Comb 1:169
55. Iwata S, Fleischer L, Fujishige S (2001) A combinatorial strongly polynomial algorithm for minimizing submodular functions. J ACM 48(4):761
56. Schrijver A (2000) A combinatorial algorithm minimizing submodular functions in strongly polynomial time. J Comb Theory Ser B 80:346
57. Edmonds J (1977) In: Guy R, Hannani H, Sauer N, Schóonheim J (eds) Combinatorial Structures and Their Applications. Gordon and Breach, New York
58. Kasteleyn PW, Fortuin CM (1969) Phase transitions in lattice systems with random local properties. J Phys Soc Jpn 26:11
59. Anglés d'Auriac JC, Iglói F, Preissmann M, Sebö A (2002) Optimal cooperation and submodularity for computing Potts' partition functions with a large number of states. J Phys A 85:6973
60. Anglés d'Auriac JC (2004) Computing the Potts free energy and submodular functions. In: Hartmann A, Rieger H (eds) New optimization algorithms in physics. Wiley, Berlin
61. Anglés d'Auriac JC, Iglói F (2003) Phase transition in the 2D random Potts model in the large-q limit. Phys Rev Lett 90:190601; Mercaldo MT, Anglés d'Auriac J-C, Iglói F (2004) Disorder-induced rounding of the phase transition in the large- q -state Potts model. Phys Rev E 69:056112; Mercaldo MT, Anglés d'Auriac J-C, Iglói F (2005) Disorder-driven phase transitions of the large q-state Potts model in three dimensions. Europhys Lett 70:733
62. Fisher DS (1992) Random transverse field Ising spin chains. Phys Rev Lett 69:534; (1995) Critical behavior of random transverse-field Ising spin chains. Phys Rev B 51:6411
63. Zeng C, Middleton AA, Shapir Y (1996) Ground-state roughness of the disordered substrate and flux lines in d = 2. Phys Rev Lett 77:3204
64. Middleton AA (1999) Numerical investigation of the thermodynamic limit for ground states in models with quenched disorder. Phys Rev Lett 83:1672
65. Anglés d'Auriac J-C, Sourlas N (1997) The 3d random field Ising model at zero temperature. Europhys Lett 39:473
66. Anglés d'Auriac J-C, Preissmann M, Sebö A (1997) Optimal cuts in graphs and statistical mechanics. Math Comput Model 26:1
67. Pal KF (2004) Hysteretic opimization. In: Hartmann A, Rieger H (eds) New optimization algorithms in physics. Wiley, Berlin
68. Boettcher S (2004) Extremal optimization. In: Hartmann A, Rieger H (eds) New optimization algorithms in physics. Wiley, Berlin
69. Weigt M (2004) The random 3-satisfiability problem: From the phase transition to the efficient generation of hard, but satisfiable instances. In: Hartmann A, Rieger H (eds) New optimization algorithms in physics. Wiley, Berlin; Cocco S, Ein-Dor L, Monasson R (ibid) Analysis of backtracking procedures for random decision problems; Zecchina R (ibid) New iterative algorithms for hard combinatorial problems

Books and Reviews

Alava M, Duxbury P, Moukarzel C, Rieger H (2000) Combinatorial optimization and disordered systems. In: Domb C, Lebowitz JL (eds) Phase Transition and Critical Phenomena, vol 18. Academic Press, Cambridge

Hartmann A, Rieger H (2002) Optimization Algorithms in Physics. Wiley VCH, Berlin

Hartmann A, Rieger H (2004) New Optimization Algorithms in Physics. Wiley VCH, Berlin

Hartmann AK, Weigt M (2005) Phase Transitions in Combinatorial Optimization Problems. Wiley-VCH, Berlin

Orbital Dynamics, Chaos in

John D. Hadjidemetriou
Department of Physics, University of Thessaloniki, Thessaloniki, Greece

Article Outline

Glossary
Definition of the Subject

Introduction
Periodic Orbits in Systems with Two Degrees of Freedom
Variational Equations
Linear Stability of a Periodic Orbit
Hamiltonian Systems
Extension to Three or More Degrees of Freedom
The Poincaré Map
Poincaré Map in Hamiltonian Systems
The Gravitational Two-Body Problem
Application to the Solar System
Extrasolar Planetary Systems
Future Directions
Bibliography

Glossary

Reference frame A *reference frame* can be determined by a set of solid bodies, through which we can define a three dimensional geometric figure, for example a triedron (three non planar axes starting from a point). The surface of the Earth can be used to define a reference system. A moving car can be also used to define a reference frame, different from the first one.

Inertial frame An *inertial frame* is a special class of reference frames, in which the basic laws of motion (*Newton's laws*) are valid. According to *Galileos' Principle of Relativity*, any frame of reference moving uniformly (with constant velocity without rotation) with respect to an inertial frame is also an inertial frame. A frame of reference which is rotating with respect to an inertial frame is not inertial. The criterion for a frame to be inertial is Newton's first law to be valid. This means that in an inertial frame a free body is either at rest or moves in a straight line with constant velocity. The best approximation in nature of an inertial frame is that frame which is defined by a triedron whose origin is at the center of mass of our Solar System and its three axes are in three fixed directions in space, defined by three distant stars.

Degrees of freedom The number of independent variables that are needed to determine the position of a dynamical system is called the number of *degrees of freedom*. For example, a particle moving freely in space has three degrees of freedom, since its position is determined by its three Cartesian coordinates (x_1, x_2, x_3), which are independent.

Phase space Consider a space whose coordinates determine exactly the state of the system. This space is called the *state space* or the *phase space* of the system. Each point of the phase space determines uniquely the initial conditions of the motion. The evolution of the system in the phase space is represented by a smooth curve, which is called the *phase curve*. The phase curves do not intersect, otherwise the point of intersection would correspond to two different solutions. The set of all phase curves is called the *phase diagram* and gives important information of the stable and unstable regions of the phase space. For gravitational systems, the phase space is the space of coordinates and velocities of all the bodies. Usually, instead of the velocities, the moments are used in the definition of the phase space. In a gravitational system with n degrees of freedom, the phase space has $2n$ dimensions. For example, a body moving in the plane under the action of a force, has two degrees of freedom (coordinates x, y) and its phase space is the four-dimensional space $x, y, p_x = m\dot{x}, p_y = m\dot{y}$.

Orbit An *orbit* of a body, or a set of bodies, considered as point masses, is the path that the bodies describe in a reference frame. The orbit of the *same* body or set of bodies is different in different frames of reference.

Periodic orbit A *periodic orbit* is the orbit of one or more bodies that repeats itself after a certain time, which is called the *period* of the periodic orbit. The periodicity property is closely related to the frame of reference to which the motion is referred to. For example, an orbit may be periodic in a rotating frame, but not in the inertial frame. In this latter case, for two or more bodies, it is the relative configuration that is repeated in the inertial frame.

Poincaré map The Poincaré map is a method by which we transform the continuous flow of a dynamical system in its n-dimensional phase space, into a discrete map in a reduced phase space. The map is obtained by taking the intersections of the continuous flow in the original phase space with a *surface of section*, defined properly. This surface of section is $(n-1)$-dimensional, in general, or $(n-2)$-dimensional if an integral of motion exists, which is the energy integral in gravitational systems. These will be explained in detail in Sects. "The Poincaré Map," "Poincaré Map in Hamiltonian Systems." A periodic orbit appears as a fixed point on the Poincaré surface of section. The Poincaré map is very useful in the study of ordered and chaotic motion in a dynamical system, especially in systems with few degrees of freedom.

Stability The notion of *stability* refers to the behavior of the orbits in the vicinity of a periodic orbit. If a slight change in the initial conditions results to a new orbit, called the *perturbed orbit*, which deviates much from the periodic orbit, then the periodic orbit is called *unstable*. In the gravitational systems that we will study,

this deviation is exponential. If on the other hand, the perturbed orbit stays close to the periodic orbit, the periodic orbit is called *stable*. But there are different aspects of stability. For example, if the perturbed orbit, considered as a geometrical figure, is close to the periodic orbit, then the periodic orbit is called *orbitally stable*. However, in this latter case it may happen that two bodies, one on the periodic orbit and one on the perturbed orbit, which start very close to each other, may deviate much as each one moves on its own orbit, although the geometric figures of the two orbits are close to each other. In this aspect, the orbit is considered as unstable. A Keplerian elliptic orbit, in the two-body problem, belongs to this latter category. A different type of stability is the *asymptotic stability*. In this case any perturbed orbit, not only stays in the vicinity of the periodic orbit, but tends asymptotically to the periodic orbit. In gravitational systems asymptotic stability does not appear, unless there exists a dissipation to the system.

Ordered and chaotic motion The notion of *chaoticity* is related to the deviation of a perturbed orbit from a given orbit. It may happen that the perturbed orbit does not deviate much as time goes on. In this case we say that we are in an *ordered region*. The prediction of the evolution of the system in this case is possible. In some cases however, the perturbed orbit deviates exponentially from the original orbit. Prediction is not possible for a long time. In this latter case we are in a *chaotic region*. In general, both ordered and chaotic regions exist in the same dynamical system.

Definition of the Subject

By the term *orbital dynamics* we mean the study of the motion of one or more bodies. Motion is one of the first things that a human being noticed, since the very early stages of human life. Apart from the motion of himself, walking around, he also realized that everything around him is not still, but changes position, being it a wild animal, a dry leaf drifting in the wind, the motion of clouds in the sky, or the change of the position of the celestial bodies, most notably of the Sun and the Moon.

Evidently, motion is one of the most important aspects in everyday life. By the term *motion* we mean the change of the position of one or more bodies in space, with respect to the other bodies in that region. If only one body existed in the universe, motion could not be defined. This makes necessary the introduction of an important notion in physics, the *frame of reference*. The surface of the Earth, for example, defines a frame of reference, with respect to which we determine the position of a body and its motion, as the body changes position. But a bus moving on the road is also a frame of reference, different from that defined by the surface of the Earth, i. e., the road. And it is a different thing if the bus moves on a straight line with constant velocity, or makes rapid turns following a difficult mountain road. Among all possible frames of reference, the *inertial frames of reference* have a special status in the study of motion. It is in these frames that the basic laws of motion (Newton's laws) are valid.

If the dimensions of the body can be considered as negligible, with respect to its surroundings, we can consider it as a point mass. However, in many cases, the finite dimensions of a body cannot be ignored. In this case its motion cannot be described by the motion that a point describes, but we have to consider also the rotation of the body. Whether we consider a body as a point mass or as a body with finite dimensions, depends on the particular study. For example the Earth is considered as a point mass in the study of its motion around the Sun, but as a finite body when we study the motion of an artificial satellite. In the present study we restrict ourselves in the motion of point masses. The path that such a body describes, is called the *orbit* of the body.

The motion of the bodies takes place under the action of forces which follow definite laws. In everyday life, the dominant force is the gravitational interaction between the bodies, according to *Newton's law of gravity*. Although it is, by far, the weakest force in nature, it is the main force that we feel in everyday life, in addition to the electromagnetic force, which also manifests itself in macroscopic phenomena. These forces affect the motion of the bodies through definite laws, expressed by differential equations, which are *deterministic* equations, i. e. to a certain set of initial conditions there exists one and only one final result, which in our case is a definite orbit. In classical physics, these laws are *Newton's laws of motion*. They are expressed by differential equations of the *second order*, which implies that the initial conditions that define the motion are the *initial position* and the *initial velocity*. This is the essential property of Newton's laws. In classical physics we assume that it is possible to know exactly, *at the same time*, the position and the velocity of a body. In some other world, where the laws of motion were expressed by differential equations of the third order, the initial acceleration would be also necessary to define the motion. Alternatively, if the laws of motion were expressed as differential equations of the first order, only the initial position would be enough to determine the motion.

As we mentioned, the equations of motion are deterministic. This implies that motion would be exactly de-

fined and that we could predict the motion of one or more bodies, for example the motion of an asteroid in our Solar System, if we knew its initial conditions. This idea prevailed classical physics until the sixties of the 20th century. But what will happen if we make a small error in the initial conditions? Does it have a great effect on the final position, after a certain time (for example a few million years for the asteroid), or the final error will be of the same order as the initial error? In the latter case the small error, due for example to a not very accurate measurement of the position and the velocity of the asteroid, is not important. In many cases however, including an asteroid in certain regions in the asteroid belt close to some mean motion resonance with Jupiter, the orbit is very sensitive to a change in the initial conditions. In this latter case, after a certain time, the orbit which corresponds to the slightly changed initial conditions, differs very much from the original orbit, because it deviates exponentially. These orbits, which are very sensitive to the initial conditions, are called *chaotic orbits*. In such a case prediction of the final position of the body, after a certain time, is not possible, because a very small error in the initial conditions, beyond the accuracy of the observations for the initial conditions, will give a completely different final position, due to the exponential deviation between the two orbits. As we will see, all the physical systems are non integrable and consequently they present chaotic behavior, at least for some initial conditions, and for this reason prediction of the evolution of such a system is not possible, after a certain time interval. This time interval is different in different systems and may be two weeks for meteorological systems or some million years for the motion of an asteroid.

Among all possible orbits in a dynamical system, the *periodic orbits* play a dominant role in the study of the evolution of the system, although it is known that they form a set of measure zero. This is so because, as it will become clear in the following, the periodic orbits are the "backbone" of the topology of the phase space, because their position and their stability character (stable or unstable) determine the structure of the phase space. It is close to the unstable periodic orbits that chaotic motion appears. A special class of periodic orbits in dynamical systems that describe the motion in the Solar System are the *resonant* periodic orbits, because around the stable resonant orbits islands of stable motion exist and the system can be trapped in these regions. In addition, since in a system there exist more that one resonances, the overlap of these resonances, as a perturbation increases, will generate chaotic motion.

In the following we restrict ourselves to the study of motion under gravitational forces, focusing on our Solar System and on extrasolar planetary systems, but the theory is applicable in all cases of motion, under any force field.

Introduction

The Newtonian gravitational force is the dominant force in the N-Body systems in the universe, as for example in a planetary system, a planet with its satellites, a multiple stellar system, or a galaxy.

In many cases, there is only one massive body, whose gravitational attraction provides the dominant force, as is the case with a planetary system, where the Sun is the main attracting body, or a planet surrounded by satellites. In this case the motion of the small bodies (planets or satellites) follow Keplerian orbits, perturbed by the gravitational interaction between the small bodies. This is a *nearly integrable* dynamical system. In these systems resonances exist between the small bodies in their motion around the massive body, as will be explained in the following. These correspond to periodic motion, and this makes clear the importance of the resonances in the dynamical properties of a nearly integrable system.

The simplest model of a gravitational system is a system of two bodies moving in Keplerian orbits around their common center of mass. This is an integrable system. In such systems all motion is ordered and chaos never appears. We consider now a hierarchy of models, starting from the above mentioned integrable system and adding more bodies to the system. We have different models, which are used to study particular systems. All these systems are not integrable, although they are nearly integrable. In these latter systems both ordered and chaotic regions appear, as we will see in the following. We consider two basic models:

The restricted three-body problem: Two bodies of finite masses, called *primaries*, revolve around their common center of mass in *circular* or *elliptic* orbits and a third body with *negligible* mass moves under their gravitational attraction, but does not affect the orbits of the two primaries. In most astronomical applications the second primary has a small mass compared to the first primary (the Sun), and consequently the motion of the third, massless, body is a perturbed Keplerian orbit. This is a model for the study of an asteroid (Jupiter being the second primary) or a trans-Neptunian object (Neptune being the second primary).

The general planetary three-body problem: Three bodies with finite masses moving under their gravitational attraction. This is a model for a triple stellar system. In many astronomical applications one of the three bodies has a large mass and the other two bodies have small,

but not negligible masses. This is a model for an extrasolar planetary system, or a system of two satellites moving around a major planet. In the latter two cases the two small bodies move in perturbed Keplerian orbits.

The long term evolution of the system depends on the topology of its phase space and the existence of ordered or chaotic regions. The topology of the phase space is determined by the position and the stability character of the periodic orbits of the system (fixed points on a Poincaré map on a surface of section). Islands of stable motion exist around the stable periodic orbits. Chaotic motion appears at the unstable periodic orbits. This makes clear the importance of the periodic orbits in the study of the dynamics of such systems.

We will start with the basic elements of gravitational systems in general. Then we will focus our attention to the study of systems of two degrees of freedom, and then extend the results to three degrees of freedom. The study will be for a general dynamical system, with particular emphasis on Hamiltonian systems, because the gravitational systems are Hamiltonian.

Basic Equations and Integrals of Motion

The gravitational force between two bodies, N_i, N_j, is given by Newton's law of gravitation

$$F_{ij} = -\frac{G m_i m_j}{r_{ij}^2} ,$$

where G is the gravitational constant, m_i, m_j are the masses of the bodies N_i and N_j and r_{ij} is their distance. The minus sign indicates attraction. We have $3N$ degrees of freedom and the evolution in space is given by the system of differential equations

$$m_i \ddot{\vec{r}}_i = \vec{F}_i ,$$

where

$$\vec{F}_i = -\sum_{j=1}^{N} \frac{G m_i m_j (\vec{r}_i - \vec{r}_j)}{r_{ij}^3} = -\frac{\partial V}{\partial \vec{r}_i} ,$$

and

$$\vec{r}_i(x_i, y_i, z_i) \quad (i = 1, 2, \ldots N)$$

is the position vector of the body N_i. The system is conservative, and the potential function is

$$V(\vec{r}_m - \vec{r}_n) = -\sum_{ij} \frac{G m_i m_j}{r_{ij}} . \quad 1 \le i < j \le N .$$

The gravitational system of N bodies can be formulated in Hamiltonian dynamics, and the Hamiltonian function is

$$H = \sum \frac{\vec{p}_i^2}{2m} + V , \quad \vec{p}_i = m_i \dot{\vec{r}}_i .$$

We have the following integrals of motion:

$$\begin{aligned} \vec{r}_{\mathrm{cm}} &= \left(\sum m_i \vec{r}_i\right) && \text{Center of mass} \\ \vec{p} &= \sum m_i \vec{v}_i = \text{constant} && \text{Linear momentum} \\ \vec{L} &= \sum \vec{r}_i \times m_i \vec{v}_i = \text{constant} && \text{Angular momentum} \\ E &= T + V = \text{constant} && \text{Energy integral.} \end{aligned}$$

The total momentum of the system is equal to $\vec{p} = m\vec{v}_{\mathrm{cm}}$, $\vec{v}_{\mathrm{cm}}$ being the velocity of the center of mass. We can assume that the total momentum is equal to zero, $\vec{p} = 0$, which implies that the center of mass is at rest, $\vec{v}_{\mathrm{cm}} = 0$. Consequently, in the system where the center of mass is at rest, we have $3N - 3$ degrees of freedom. In this latter case we can take $\vec{r}_{\mathrm{cm}} = 0$.

Periodic Orbits in Systems with Two Degrees of Freedom

Periodic Orbits

The periodic orbits play an important role in understanding the dynamics of a system, because they determine critically the topology of the phase space. This will become clear in the following. For this reason it is important to know the basic families of periodic orbits in a dynamical system, because they are the "backbone" of the phase space.

Let us consider a dynamical system with two degrees of freedom, defined by the set of two second order differential equations

$$\begin{aligned} \ddot{x}_1 &= F_1(x_1, x_2, \dot{x}_1, \dot{x}_2) , \\ \ddot{x}_2 &= F_2(x_1, x_2, \dot{x}_1, \dot{x}_2) . \end{aligned} \tag{1}$$

The initial conditions that determine a solution are $(x_{10}, x_{20}, \dot{x}_{10}, \dot{x}_{20})$ and the corresponding solution has the form

$$\begin{aligned} &x_1(x_{10}, x_{20}, \dot{x}_{10}, \dot{x}_{20};\ t) , \\ &x_2(x_{10}, x_{20}, \dot{x}_{10}, \dot{x}_{20};\ t) . \end{aligned}$$

The solution is periodic, with period T, if

$$x_i(x_{10}, x_{20}, \dot{x}_{10}, \dot{x}_{20};\ t + T) = x_i(x_{10}, x_{20}, \dot{x}_{10}, \dot{x}_{20};\ t) ,$$

for every t.

Existence of Symmetric Periodic Orbits

We assume that the differential equations (1) are invariant under the transformation

$$x_1 \to x_1\,, \quad x_2 \to -x_2\,, \quad t \to -t.$$

This property appears in several models that are of astronomical interest. This means that if $x_1(t), x_2(t)$ is a solution, then $x_1(t), -x_2(-t)$ is also a solution. Note that this second solution is the symmetric of the first solution with respect to the x_1-axis. Consequently, if an orbit starts from the x_1-axis perpendicularly, $\dot{x}_{10} = 0$, and crosses again the x_1-axis perpendicularly, $\dot{x}_1 = 0$, the orbit is closed and is a *symmetric* periodic orbit with respect to the x_1-axis.

The initial conditions of a symmetric periodic orbit are $(x_{10}, x_{20} = 0, \dot{x}_{10} = 0, \dot{x}_{20})$, which means that a symmetric periodic orbit is determined only by the two nonzero initial conditions $x_{10}, \dot{x}_{20}$. From the above we see that the *periodicity conditions* are

$$x_2(x_{10}, 0, 0, \dot{x}_{20};\, T/2) = 0\,,$$
$$\dot{x}_1(x_{10}, 0, 0, \dot{x}_{20};\, T/2) = 0\,,$$

which imply that the orbit starts perpendicularly from the x_1-axis ($x_{20} = 0, \dot{x}_{10} = 0$) and crosses again perpendicularly the x_1-axis after a time interval equal to half the period T. We remark that the second *perpendicular* crossing may take place after several (non perpendicular) crossings from the x_1-axis.

The periodic orbits are not isolated, in general. They belong to families, along which the period varies. A family of symmetric periodic orbits is represented by a continuous curve in the space of initial conditions $x_{10}, \dot{x}_{20}$. This curve is called a *characteristic curve*.

Variational Equations

We study now the behavior of the system in the vicinity of a specific orbit, by considering perturbed initial conditions, i. e. initial conditions in the vicinity of the initial conditions of this orbit.

We express the system of differential equations (1) as a system of four differential equations of the first order,

$$\dot{x}_i = f_i(x_1, x_2, x_3, x_4)\,, \quad (i = 1, \ldots 4) \tag{2}$$

where $x_3 = \dot{x}_1$, $x_4 = \dot{x}_2$. Let $x_i = x_i(x_{10}, x_{20}, x_{30}, x_{40};\, t)$, $(i = 1, \ldots 4)$ be a solution of the system (2), nonperiodic in general, corresponding to the initial conditions $x_1(0), x_2(0), x_3(0), x_4(0)$. We consider new initial conditions, in the vicinity of these initial conditions, of the form $x_i(0) + \xi_i(0)$, where $\xi_i(0)$ are small. The new solution can be expressed in the form

$$x_i'(t) = x_i(t) + \xi_i(t)\,, \quad (i = 1, \ldots 4)$$

where $\xi(t)$ is the deviation vector between the initial solution $x_i(t)$ and the perturbed solution $x_i'(t)$, at the *same time* t, $\xi(t) = x_i'(t) - x_i(t)$. The behavior of the system in the vicinity of the solution $x_i(t)$ depends on the deviation vector $\xi(t)$.

We assume that the initial perturbation $\xi(0)$ is small, and consequently, for continuity reasons, the deviation $\xi(t)$ should be also small, at least for a finite time interval. For this reason we linearize the system of differential equations (2), to first order terms in the $\xi_i(t)$, by substituting the perturbed solution $x_i'(t)$ into the system (2) and keeping only the first order terms in ξ_i. We obtain the system of *variational equations*,

$$\dot{\xi}_i = \sum_{k=1}^{4} p_{ik}\xi_k\,, \quad p_{ik} = \left(\frac{\partial f_i}{\partial x_k}\right)_{x_i(t)}\,, \quad (i = 1, \ldots 4) \tag{3}$$

which describes the evolution of the system (2) in the neighborhood of the orbit $x_i(t)$, to first order terms in the deviations. The partial derivatives are computed for the solution $x_i(t)$. This is a linear system with time dependent coefficients.

The general solution of the linear system (3) is expressed as a linear combination of four linearly independent solutions. In particular, let us consider a 4×4 matrix $\Delta(t)$ whose columns are four linearly independent solutions corresponding to the initial conditions $\Delta(0) = I_4$, where I_4 is the 4×4 unit matrix. This matrix is called *fundamental matrix of solutions* and the general solution of the variational equations is expressed in the form

$$\xi(t) = \Delta(t)\xi(0)\,. \tag{4}$$

A basic property of the matrix $\Delta(t)$ is the Liouville–Jacobi formula [24,58].

$$\det \Delta(t) = \det \Delta(0) \exp \int_0^t \mathrm{trace}(P)\mathrm{d}t\,, \tag{5}$$

where P is the matrix of the coefficients of the variational equations (3), with elements p_{ij}. This relation gives the change in time of the determinant of the matrix $\Delta(t)$, which describes important properties of the evolution of the system in phase space, as we shall see in the following. Of special importance is the case where $\mathrm{trace}(P) = 0$, because in this case the determinant of the matrix $\Delta(t)$ is constant.

Another important property is that the columns of the matrix $\Delta(t)$ are the partial derivatives of the solution $x_i(x_{10}, x_{20}, x_{30}, x_{40}, t)$ with respect to the initial conditions: The solution $x_i(x_0;\ t)$ satisfies the system (2),

$$\frac{\partial x_i(x_0,t)}{\partial t} = f_i(x_1(x_0,t), x_2(x_0,t), x_3(x_0,t), x_4(x_0,t)) , \quad (i = 1, \ldots 4) .$$

If we apply to the above equations the operator $\partial/\partial x_{j0}$, $j = 1, \ldots 4$, we obtain

$$\frac{\partial}{\partial t}\left(\frac{\partial x_i}{\partial x_{j0}}\right) = \sum_k \left(\frac{\partial f_i}{\partial x_k}\right)\frac{\partial x_k}{\partial x_{j0}} , \quad (i = 1, \ldots 4) \tag{6}$$

for each x_{j0}. We note that the system (6) is the system of variational equations (3) satisfied by the vector $(\partial x_1/\partial x_{j0}, \partial x_2/\partial x_{j0}, \partial x_3/\partial x_{j0}, \partial x_4/\partial x_{j0})$. In addition, we note that $\partial x_i/\partial x_{j0} = \delta_{ij}$ for $t = 0$, which implies that these vectors, for $j = 1, \ldots 4$, are the four columns of the fundamental matrix of solutions $\Delta(t)$. This means that the fundamental matrix of solutions $\Delta(t)$ is the Jacobian of the solution $x(t)$ with respect to the initial conditions,

$$\Delta(t) = \frac{\partial(x_1, x_2, x_3, x_4)}{\partial(x_{10}, x_{20}, x_{30}, x_{40})} . \tag{7}$$

Linear Stability of a Periodic Orbit

The variational equations (3) are a system of four linear differential equations with time dependent coefficients. If the solution $x(t)$ is T-periodic, then the partial derivatives are also T-periodic. In this latter case the system of variational equations is a *linear system with periodic coefficients*. The theory related to the study of such systems is the *Floquet theory* [24] and some elements of it will be presented in the following sections.

Existence of a Periodic Solution

We shall prove that the derivative $\dot{x}_i(t)$ of the periodic solution $x_i(t)$ is a solution of the variational equations. Indeed, the solution $x_i(t)$ satisfies the system (2)

$$\dot{x}_i(t) = f_i(x_1(t), x_2(t), x_3(t), x_4(t)) , \quad (i = 1, \ldots 4)$$

and if we apply the operator $\mathrm{d}/\mathrm{d}t$ we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}(\dot{x}_i(t)) = \sum_{j=1}^{4}\left(\frac{\partial f_i}{\partial x_j}\right)_{x_i(t)} \dot{x}_j(t) .$$

This is the system of variational equations (3), for the solution $\xi_i = \dot{x}_i(t)$. So we come to the conclusion that *the variational equations that correspond to a T-periodic orbit have always a T-periodic solution, which is the derivative $\dot{x}_i(t)$ of the periodic solution.*

Mapping at Integral Multiples of the Period. The Monodromy Matrix

Let $x_i(t)$ be a periodic orbit and $x'(t)$ a perturbed orbit, which, to a linear approximation, can be expressed in the form

$$x'_i(t) = x_i(t) + \xi_i(t) ,$$

where $\xi_i(t)$ is the solution of the variational equations. This latter solution is expressed in the form

$$\xi(t) = \Delta(t)\xi(0) , \tag{8}$$

and for $t = T$,

$$\xi(T) = \Delta(T)\xi(0) . \tag{9}$$

From this expression we obtain, by induction,

$$\xi(nT) = [\Delta(T)]^n \xi(0) . \tag{10}$$

Equations (9) and (10) give the deviation, to a linear approximation, of the perturbed orbit $x'(t)$ from the periodic orbit $x(t)$ after a time interval equal to n times the period T, due to an initial deviation $\xi(0) = x'(0) - x(0)$ at $t = 0$. In fact Eq. (10) is a mapping of the initial deviation $\xi(0)$ at integral multiples of the period T (see Fig. 1a). This is a linear mapping defined by the matrix $\Delta(T)$. It is clear that the stability of the periodic orbit $x(t)$ depends on the properties of the mapping (10), i. e. on the eigenvalues of the matrix $\Delta(T)$. The matrix $\Delta(T)$ is called the *monodromy matrix*.

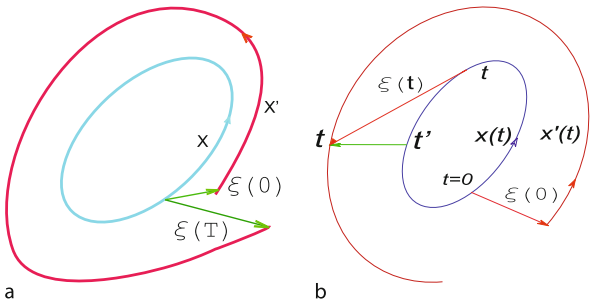

Orbital Dynamics, Chaos in, Figure 1
a Mapping at integral multiples of the period. b Orbital stability: The distance between two points *at the same time t*, $\xi(t) = x'(t) - x(t)$, is not small, but the distance between the points at two *different times*, t and t', $x'(t) - x(t')$, remains bounded

Unit Eigenvalue of the Monodromy Matrix

Existence of an Integral of Motion We shall prove that if the system of differential equations (2) has an integral of

motion,

$$G(x_1, x_2, x_3, x_4) = \text{constant}\,,$$

the system of variational equations (3) has a unit eigenvalue: Let $x_i(x_0, t)$ be a T-periodic solution. Since $G(x_1, x_2, x_3, x_4)$ is an integral, we have the relation

$$G(x_1(x_0, t), x_2(x_0, t), x_3(x_0, t), x_4(x_0, t)) = G(x_{10}, x_{20}, x_{30}, x_{40})\,.$$

We apply to the above relation the operator $\partial/\partial x_{j0}$, and we obtain

$$\sum_{k=1}^{4} \left(\frac{\partial G}{\partial x_k}\right)_t \left(\frac{\partial x_k}{\partial x_{j0}}\right)_t = \left(\frac{\partial G}{\partial x_{j0}}\right)\,.$$

We set now $t = T$ and taking into account that

$$\left(\frac{\partial G}{\partial x_k}\right)_{t=T} = \left(\frac{\partial G}{\partial x_k}\right)_{t=0}\,,$$

due to the fact that $x(t)$ is periodic, we obtain

$$\left(\Delta^{\tau}(T) - I\right) \nabla G = 0\,, \tag{11}$$

where Δ^{τ} is the transpose of Δ. From this relation we obtain that if $\nabla G \neq 0$ then $\Delta(T)^{\tau}$ has a unit eigenvalue. Thus finally, we come to the conclusion that *if the dynamical system has an integral of motion, which is not stationary along the periodic orbit, the monodromy matrix $\Delta(T)$ has a unit eigenvalue.*

Existence of a Periodic Orbit of the Variational Equations We shall also prove that if the system of variational equations has a periodic solution $\xi(t)$, the monodromy matrix has a unit eigenvalue: We have $\xi(t + T) = \xi(t)$, for any t and consequently, for $t = 0$, $\xi(T) = \xi(0)$. Due to this latter relation, Eq. (8) takes the form, for $t = T$, $\xi(0) = \Delta(T)\xi(0)$, and finally

$$(\Delta(T) - I)\,\xi(0) = 0\,. \tag{12}$$

Thus we come to the conclusion that *if the system of variational equations has a periodic solution, the monodromy matrix has a unit eigenvalue.*

We have proved above that the system of variational equations has the T-periodic solution $\xi(t) = \dot{x}_i(t)$, where $\dot{x}_i(t)$ is the periodic solution corresponding to the variational equations. Thus, the monodromy matrix $\Delta(T)$ has always a unit eigenvalue. The corresponding eigenvector is the vector $\xi(0) = \dot{x}_i(0)$, which is the tangent vector to the periodic orbit, in the phase space.

Vertical Stability of Planar Periodic Orbits

In the previous sections we studied the stability of a planar periodic orbit with respect to perturbations of the initial conditions *in the plane*. We study now the stability with respect to perturbations of the initial conditions *perpendicular* to the plane of motion. This type of stability we call *vertical stability* and completes the study of the stability of a planar periodic orbit.

Consider a dynamical system of three degrees of freedom,

$$\begin{aligned} \ddot{x}_1 &= f_1(x_1, x_2, x_3, \dot{x}_1, \dot{x}_2, \dot{x}_3)\,, \\ \ddot{x}_2 &= f_2(x_1, x_2, x_3, \dot{x}_1, \dot{x}_2, \dot{x}_3)\,, \\ \ddot{x}_3 &= x_3\, f_3(x_1, x_2, x_3, \dot{x}_1, \dot{x}_2, \dot{x}_3)\,. \end{aligned} \tag{13}$$

This is the form of the differential equations of many gravitational systems, for example of the 3-dimensional restricted three body problem, described in Sect. "Application to the Solar System" (see p. 67 in [41]). In this model, a small body with negligible mass moves under the gravitational attraction of two main bodies, which describe Keplerian orbits around their center of mass, under their mutual gravitational attraction. The plane $x_1\, x_2$ is the plane of motion of the two main bodies (in the inertial frame) and the small body moves in the three dimensional space $x_1\, x_2\, x_3$. It is intuitively clear that if the small body starts from a position in the $x_1\, x_2$ plane and its velocity is *in* this plane, then its motion is restricted in the $x_1\, x_2$ plane, since the gravitational attraction from the two main bodies is in this plane. This physical property of the motion is described by the special mathematical form of the equations of motion (13) (third equation).

It is easy to verify that the Eqs. (13) admit a planar solution, which we will assume to be periodic: $x_1(t)$, $x_2(t), x_3(t) = 0$, corresponding to the initial conditions $x_{10}, x_{20}, x_{30} = 0, \dot{x}_{10}, \dot{x}_{20}, \dot{x}_{30} = 0$. We consider now a small perturbation ϵ_3, ϵ_6 along the x_3 axis, $x_{10} + \epsilon_1$, $x_{20} + \epsilon_2$, $x_{30} = 0 + \epsilon_3$, $\dot{x}_{10} + \epsilon_4$, $\dot{x}_{20} + \epsilon_5$, $\dot{x}_{30} = 0 + \epsilon_6$, where ϵ_i are small, and we want to study the behavior of the perturbed solution. We define new variables $x_4 = \dot{x}_1, x_5 = \dot{x}_2, x_6 = \dot{x}_3$, and a simple calculation shows that the system of variational equations of the system (13), for the periodic solution $x_i(t)$, breaks into two uncoupled systems: a system in the planar displacements $\xi_1, \xi_2, \xi_4, \xi_5$, corresponding to the variational equations of the planar motion, and a system in the vertical displacements (along the x_3 axis) ξ_3, ξ_6. This latter system is

$$\begin{aligned} \dot{\xi}_3 &= \xi_6\,, \\ \dot{\xi}_6 &= f_{30}(t)\xi_3\,, \end{aligned} \tag{14}$$

where the function $f_{30}(t)$ is the T-periodic function $f_3(x_1(t), x_2(t), x_3 = 0, x_4(t), x_5(t), x_6 = 0)$, computed for the planar T-periodic solution $x_i(t)$. The system (14) is the system of variational equations for the displacements along the x_3 axis. The vertical stability depends on the eigenvalues λ_5, λ_6 of the monodromy matrix $\Delta_2(T)$ of this system.

Hamiltonian Systems

The gravitational systems are Hamiltonian. For this reason, we study in this section the special properties that a Hamiltonian system has, in addition to the general properties obtained in the previous sections. We start with systems with two degrees of freedom.

A Hamiltonian system is defined by the Hamiltonian function

$$H(x_1, x_2, x_3, x_4)\,,$$

where x_1, x_2 are the coordinates and x_3, x_4 the momenta.

The Hamiltonian equations are

$$\dot{x}_1 = \partial H/\partial x_3\,, \qquad \dot{x}_2 = \partial H/\partial x_4\,,$$
$$\dot{x}_3 = -\partial H/\partial x_1\,, \quad \dot{x}_4 = -\partial H/\partial x_2\,,$$

or

$$\dot{x} = -J\nabla H\,, \tag{15}$$

where ∇H is a column vector with elements $\partial H/\partial x_i$ and J the 4×4 symplectic matrix

$$J = \begin{pmatrix} 0 & -I_2 \\ +I_2 & 0 \end{pmatrix}\,,$$

where I_2 is the 2×2 unit matrix. Note that $J^{-1} = -J$.

Variational Equations of Hamiltonian Systems

The variational equations of a Hamiltonian system (15) have the special form given by

$$\dot{\xi} = -JA\xi\,, \tag{16}$$

where the elements a_{ij} of the 4×4 matrix A are

$$a_{ij} = \frac{\partial^2 H}{\partial x_i \partial x_j}\,. \quad (i, j = 1, \ldots 4) \tag{17}$$

Note that the matrix A is symmetric. The system (16) is called a *linear Hamiltonian system*. A complete study of such systems can be found in [58]. It is easy to see that it can be expressed in the Hamiltonian form (15) with Hamiltonian

$$H = \frac{1}{2}\xi^\tau A\xi = \frac{1}{2}\sum_{i,j=1}^{4} a_{ij}\xi_i\xi_j\,.$$

From the relations (16), (17) we can verify that the trace of the matrix of the coefficients of the linear Hamiltonian system (16) is equal to zero. Consequently, due to the general property (5), the determinant of the fundamental matrix of solutions $\Delta(t)$ is equal to unity (see also [34]), $\det \Delta(t) = \det \Delta(0) = 1$. For $t = T$ we obtain

$$\det \Delta(T) = 1\,,$$

from which we see that the determinant of the monodromy matrix is equal to one.

Using now the results of Sect. "Variational Equations," we find that

$$\det \Delta(t) = \det \left| \frac{\partial(x_1, x_2, x_3, x_4)}{\partial(x_{10}, x_{20}, x_{30}, x_{40})} \right| = 1\,. \tag{18}$$

This means that the determinant of the Jacobian of the flow in phase space is equal to one. Consequently, *the volume in phase space is conserved* (Liouville theorem).

The monodromy matrix of a Hamiltonian system is symplectic (see for example, [20])

$$\Delta^\tau(T) J \Delta(T) = J\,, \tag{19}$$

where the superscript τ means transpose. This is an important property of the monodromy matrix of a Hamiltonian system, which is called the *symplectic* property. Thus we come to the conclusion that *the monodromy matrix of a Hamiltonian system is symplectic.*

The eigenvalues of a symplectic matrix have some special properties. We express the property (19) as

$$\Delta^\tau(T) = J\Delta^{-1}(T)J^{-1}\,,$$

from which we see that the matrix $\Delta^\tau(T)$ is related to the matrix $\Delta^{-1}(T)$ by a similarity transformation. Consequently, they have the same set of eigenvalues. Thus finally, we come to the conclusion that the eigenvalues of $\Delta(T)$ are in reciprocal pairs. In addition, due to the fact that the matrix $\Delta(T)$ is real, they are also in complex conjugate pairs.

From the above we see that the four eigenvalues $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ of the monodromy matrix have the property

$$\lambda_1\lambda_2 = 1\,, \quad \lambda_3\lambda_4 = 1\,.$$

We note that the variational equations correspond to a periodic orbit $x(t)$. So, $\xi(t) = \dot{x}(t)$ is a periodic solution of

the variational equations and according to Sect. "Variational Equations," one eigenvalue is equal to one, $\lambda_1 = 1$. Using now relation $\lambda_1\lambda_2 = 1$, we come to the conclusion that *the monodromy matrix of a Hamiltonian system corresponding to a periodic orbit has a double unit eigenvalue*,

$$\lambda_1 = 1\,, \quad \lambda_2 = 1\,.$$

Stability of Hamiltonian Systems

The stability of the periodic orbit depends on the eigenvalues of the monodromy matrix, as we showed in Sect. "Linear Stability of a Periodic Orbit." Instability appears if at least one eigenvalue is outside the unit circle in the complex plane. Since two of the eigenvalues are always equal to unity, it is the other two eigenvalues, λ_3, λ_4, that determine the stability. As we proved, these eigenvalues are reciprocal and also complex conjugate, so they are either *on the unit circle*, or *on the real axis*, one inside the unit circle and the other outside. If they are real, the orbit is unstable, because one of them will be larger than $+1$ or smaller than -1. A special case is $\lambda_3 = \lambda_4 = +1$ or $\lambda_3 = \lambda_4 = -1$.

A remark is necessary at this point for the double unit eigenvalue. In Hamiltonian systems, in general, to the double unit eigenvalue there exists only one eigenvector. This introduces a secular term in the general solution of the variational equations. The two linearly independent solutions corresponding to the double unit eigenvalue are

$$\begin{aligned} \xi^1 &= f_1(t)\,, \\ \xi^2 &= f_2(t) + t f_1(t)\,, \end{aligned} \tag{20}$$

where $f_1(t), f_2(t)$ are T-periodic. This implies that the orbit is always unstable, due to the secular term $t f_1(t)$. We will show however that this secular term introduces a time shift only along the perturbed orbit, and thus we have *orbital stability*, provided that the other two eigenvalues are on the unit circle: Taking into account that $\xi^1(t) = \dot{x}(t)$, where $x(t)$ is the periodic solution corresponding to the unit eigenvalue, we note that the perturbed orbit has a term $\epsilon\xi^2$ and the corresponding part of the the solution is expressed as $x'(t) = x(t) + \epsilon t \dot{x}_i(t) + \epsilon f_2(t)$ and, to a linear approximation in ϵ,

$$x'(t) = x(t+\epsilon) + \epsilon f_2(t+\epsilon t)\,.$$

Thus, if we define a new time $t' = t + \epsilon t$, we obtain (see Fig. 1b)

$$x'(t) - x(t') = \text{bounded}\,.$$

Thus we come to the conclusion that the secular term introduces a phase shift only along the orbit. This means that the two orbits, $x(t)$ and $x'(t)$, considered as geometrical curves, are close to each other. In this case we say that we have *orbital stability*, provided that the eigenvalues λ_3, λ_4 are on the unit circle and consequently the corresponding solution is bounded.

For the other two eigenvalues λ_3, λ_4 we have the solutions

- Eigenvalues real and positive: $\xi^{3,4} = f_{3,4}(t)\,\mathrm{e}^{\pm\alpha t}$,
- Eigenvalues real and negative: $\xi^{3,4} = f_{3,4}(t)\,\mathrm{e}^{\pm\alpha t}\,\mathrm{e}^{\pm i\pi t/T}$,
- Eigenvalues complex conjugate on the unit circle $\xi^{3,4} = f_{3,4}(t)\,\mathrm{e}^{\pm i\beta t}$,

where α, β are real and the functions $f_3(t), f_4(t)$ are T-periodic. The exponent α is called the *characteristic exponent*. The general solution in the vicinity of the periodic solution is a linear combination of the above four solutions $\xi^1, \xi^2, \xi^3, \xi^4$.

The stability criteria can be obtained from the elements of the monodromy matrix as follows: The eigenvalues are the roots of the characteristic equation of $\Delta(T)$ and consequently

$$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = \text{trace}\,\Delta(T)\,,$$

$$\lambda_1\lambda_2\lambda_3\lambda_4 = \det\Delta(T) = 1\,.$$

Taking into account that $\lambda_1 = \lambda_2 = 1$ we find that the two nonzero eigenvalues λ_3, λ_4 are the roots of the quadratic equation

$$\lambda^2 - K\lambda + 1 = 0\,,$$

where

$$K = \text{trace}\,\Delta(T) - 2\,.$$

The stability depends on the value of K, which is called the *stability index*. Note that the stability index depends only on the trace of the monodromy matrix.

Asymptotic stability never appears, because it is not possible for the eigenvalues λ_3, λ_4 to be *both* inside the unit circle. This is also a consequence of the fact that the volume in phase space is conserved.

Let us assume that a periodic orbit is stable, which implies that the eigenvalues λ_3, λ_4 are on the unit circle and we assume that they are not equal to $+1$ or -1. If a parameter varies, then the eigenvalues λ_3, λ_4 are restricted to move *on the unit circle*, because they must be both inverse, $\lambda_3 = 1/\lambda_4$ and complex conjugate. Consequently, the stability is conserved. However, if λ_3, λ_4 meet at the points $+1$ or -1, then it is possible for them to go outside the unit

circle and thus generate instability. For this reason the orbits with $\lambda_3 = \lambda_4 = \pm 1$ are called *critical* as far as the stability is concerned. This is the mechanism by which instability is generated at the 3:1 resonance in the asteroid belt, where the eigenvalues λ_3, λ_4 meet at the point -1 [13].

Extension to Three or More Degrees of Freedom

All the above results concerning the eigenvalues and the stability of a periodic orbit, obtained for Hamiltonian systems with two degrees of freedom, can be easily extended to three or more degrees of freedom.

In a Hamiltonian system with n degrees of freedom the monodromy matrix is a $2n \times 2n$ symplectic matrix, and the eigenvalues are in reciprocal pairs (because of the symplectic property), and in complex conjugate pairs (because the elements of the matrix are real).

There is always a unit pair of eigenvalues, due to the existence of the energy integral $H = h =$ constant (see Sect. "Linear Stability of a Periodic Orbit"). For the other eigenvalues we have the following possibilities:

- Complex conjugate, on the unit circle, $e^{\pm i\phi}$: STABILITY.
- Real, on the real axis, in reciprocal pairs (positive or negative), $\lambda, 1/\lambda$: INSTABILITY.
- Complex, inside and outside the unit circle, in reciprocal and in complex conjugate pairs, $Re^{i\phi}$, $Re^{-i\phi}$, $R^{-1}e^{i\phi}$, $R^{-1}e^{-i\phi}$: COMPLEX INSTABILITY.

Note that in three, or more, degrees of freedom we have a new type of instability, the *complex instability*, which cannot appear in systems with two degrees of freedom.

The Poincaré Map

This a very useful method in the study of the evolution of a dynamical system. By the Poincaré map we transform the continuous flow in the n-dimensional phase space of a dynamical system to an equivalent discrete flow (map) in a phase space of $(n-1)$-dimensions (or $(n-2)$-dimensions for Hamiltonian flows).

Consider the dynamical system in $\mathcal{R}^n$: $\dot{x} = f(x)$, where $x, f(x)$: vectors in $\mathcal{R}^n$ and $\phi_t(x)$ is the flow. We consider the surface of section

$$\Sigma \subset \mathcal{R}^n : (n-1) - \text{dim}$$

and we assume that the flow is transverse: The velocity vector of the flow is not tangent to this surface (Fig. 2a): $f(x) \cdot n(x) \neq 0$, where $n(x)$ is the normal unit vector to the surface.

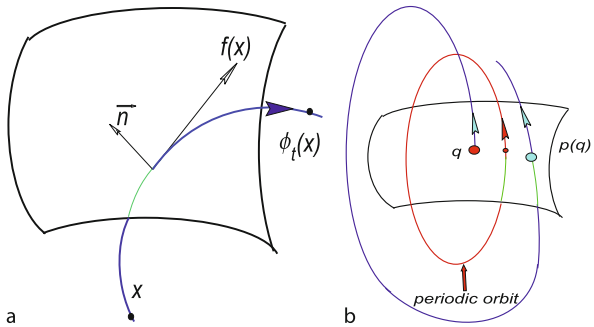

Orbital Dynamics, Chaos in, Figure 2
a The surface of section, b The Poincaré map on a surface of section

The Poincaré map is defined as:

$$q \to p(q)\,,$$
$$p(q) = \phi_\tau(q)\,,$$

where q is the position on the surface of section at a $t = 0$ and $p(q)$ is the position on this surface at the next intersection at $t = \tau$, (Fig. 2b).

The following properties apply:

- The vector $p(q)$ defines accurately the state.
- The vector $p(q)$ is a continuous function of q.
- If $\bar{x}(t)$ is a T-periodic orbit, the corresponding Poincaré map is a *fixed point*, maybe multiple (it repeats itself after several intersections) as seen in Fig. 2b (for the simple case).

Poincaré Map in Hamiltonian Systems

In this case the differential equations of motion are the canonical equations

$$\dot{q} = \partial H/\partial p\,, \quad \dot{p} = -\partial H/\partial q\,, \quad q, p \subset \mathcal{R}^n\,.$$

Let us consider the $(2n-2)$-dimensional surface of section Σ, defined as

$$H = h\,, \quad f(q,p) = 0 \quad (\text{for example} \quad q_2 = 0)\,.$$

The continuous Hamiltonian flow in the $2n$-dimensional phase space is transformed to an equivalent discrete flow (map), on a $(2n-2)$-dimensional surface of section. In addition to the general properties of the Poincaré map mentioned above, we also have the properties:

- The Poinaré map of a Hamiltonian flow is symplectic.
- The stability of the fixed points of the Poincaré map is the same as the stability of the corresponding periodic orbit. We have the same set of eigenvalues, except the double unit eigenvalue which corresponds to

the periodic orbit (the phase space now has two dimensions less). Note that this double unit eigenvalue is responsible for the phase shift along the perturbed orbit, which implies that this shift is eliminated by the Poincaré map. Thus, in the Poincaré map, the stability of the fixed point (periodic orbit) means *orbital stability*.

Hamiltonian Systems with Two Degrees of Freedom

The Poincaré map is particularly useful in systems with two degrees of freedom, where the phase space is four dimensional and the Poincaré map is in a two dimensional phase space. This makes the study very easy because we present the evolution of the system in a two dimensional space, where we can have a direct view.

We define the variables x_i as $x_1 = q_1$, $x_2 = q_2$, $x_3 = p_1$, $x_4 = p_2$. The energy integral is $H(x_1, x_2, x_3, x_4) =$ constant. We consider the surface of section

$$H(x_1, x_2, x_3, x_4) = h\,, \qquad x_2 = 0\,, \quad \text{with } x_4 > 0\,.$$

The map is in the space $x_1\, x_3$. The consecutive points of the map may lie on a smooth curve, called *invariant curve* (ordered motion), or be scattered (chaotic motion).

Let us assume that another first integral of motion exists, in addition to the energy integral $H = h =$ constant:

$$G(x_1, x_2, x_3, x_4) = c\,.$$

Then all the consecutive points of the map lie on smooth invariant curves: Let (x_1, x_3) be a point of the map on the two-dimensional surface of section. We have $x_2 = 0$ and x_4 is expressed in terms of x_1, x_3, through the energy integral $H = h$, as $x_4 = x_4(x_1, x_2 = 0, x_3)$. The points x_1, x_3 satisfy also the integral $G(x_1, x_2 = 0, x_3, x_4\,(x_1, x_2 = 0, x_3)) = c$, or

$$F(x_1, x_3) = 0\,,$$

which implies that the consecutive points (x_1, x_3) of the map lie on a smooth curve.

The Gravitational Two-Body Problem

The differential equations of the *relative motion* of two point masses m_1, m_2 are given by

$$\ddot{\vec{r}} = -\frac{GM}{r^2}\,\vec{e}_r\,,$$

where $M = m_1 + m_2$. The orbit is a conic section and in particular, for bounded motion, it is a Keplerian, elliptic orbit. The two bodies describe in the inertial frame two similar orbits around their common center of mass, whose dimensions are inversely proportional to their masses. This is one of the few integrable problems in nature. Its importance is that many real systems, as for example the asteroid problem, or the planetary systems, can be considered as perturbed two-body problems. For this reason it is important to know the basic properties of this simple two body problem and then study the evolution as a perturbation is applied.

The Two-Body Problem in a Rotating Frame

Consider a body, S, with mass m_1 and a second body, J, with mass m_2, which describe circular orbits around their common center of mass. We define a rotating frame of reference xOy, whose x-axis is the line SJ, the origin is at their center of mass and the xy plane is the orbital plane of the circular motion of the these two bodies (Fig. 3b).

Our aim is to study the motion of a massless body A in the rotating frame xOy, under the gravitational attraction of S and J. We start with a zero mass of the body J, $m_2 = 0$. In this approximation, the second body J is used only to define the rotating frame xOy, which rotates with constant angular velocity n'. Evidently, the motion of the body A is a Keplerian orbit, presented in the rotating frame. We shall give later mass to the body J, thus perturbing the Keplerian orbit of A.

The Hamiltonian function H that describes the unperturbed motion of A, in polar coordinates, r, ϕ (in the rotating frame), is

$$H_0 = \frac{p_r^2}{2} + \frac{p_\phi^2}{2r^2} - n' p_\phi - \frac{GM}{r}\,. \tag{21}$$

The moments are $p_r = \dot{r}$ and $p_\phi = r^2(\dot{\phi} + n')$. Note that the angle ϕ is an ignorable coordinate and consequently, in addition to the energy integral $H_0 = h =$ constant, we also have the angular momentum integral $p_\phi =$ constant.

The orbit of the body A (in the inertial frame) is a Keplerian orbit, which we assume to be elliptic. In terms of the elements of the orbit, the Hamiltonian (in the rotating frame) and the angular momentum are expressed as

$$H_0 = -\frac{GM}{2a} - n' p_\phi\,, \quad p_\phi = \sqrt{GMa(1-e^2)}\,.$$

Circular Orbits In the rotating frame there exist circular orbits of the body A with an arbitrary radius r_0, which correspond to the periodic solution

$$r = r_0\,, \quad p_r = 0\,, \quad \dot{\phi} = n - n'\,, \quad p_\phi = n r_0^2\,,$$

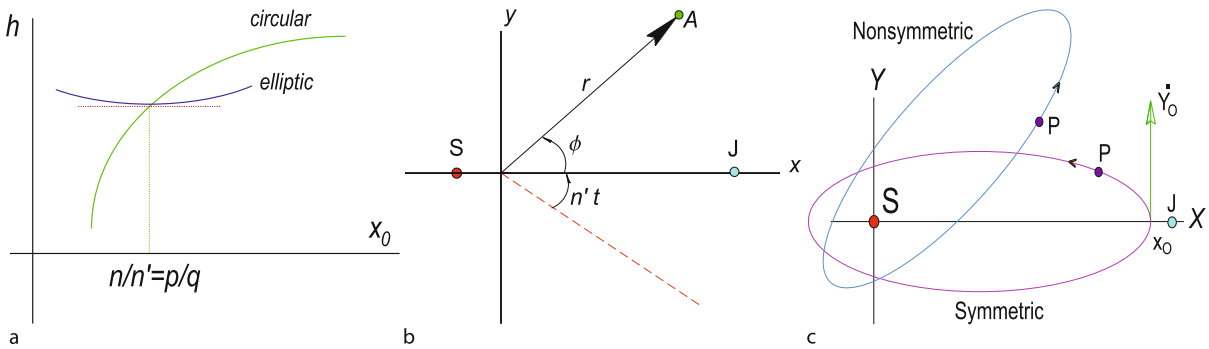

Orbital Dynamics, Chaos in, Figure 3

a The families of circular and of resonant elliptic periodic orbits in the unperturbed problem. The tangent to the elliptic family at the bifurcation point is parallel to the *x*-axis. b The rotating frame of the restricted problem. The mass, μ, of the second body, J, is equal to zero in the unperturbed problem, and is used only to define the rotating frame. In this case, the first body, S, is at the origin. c Two elliptic orbits of the small body, in the inertial frame, for $\mu = 0$. One is symmetric, corresponding to $\omega = 0^0$, and the other is asymmetric, corresponding to an arbitrary value of ω

where

$$n = p_\phi / r_0^2$$

is the angular velocity of the circular orbit (in the inertial frame). The following relations also hold:

$$\frac{p_{\phi 0}^2}{r_0^3} = \frac{GM}{r_0^2} \rightarrow \frac{GM}{r_0^3} = n^2 .$$

The period of the circular orbit in the rotating frame is

$$T = \frac{2\pi}{(n - n')} .$$

A circular orbit in xOy is a Keplerian orbit in the inertial frame, with semi major axis $a = r_0$ for any r_0. Consequently, a *family of circular periodic orbits* exists, which evidently is symmetric with respect to the x-axis. The parameter along the family is the semi major axis a (the radius), or the angular velocity (in the inertial frame) n. This family is represented by a smooth curve, in the space $h - r_0$, given by

$$-\frac{GM}{2a} - n' \sqrt{GMa} = h ,$$

obtained from the energy integral for $e = 0$ (Fig. 3a). Note that from the energy integral $H_0 = h$ we can obtain the value of $\dot{y}_0$, which together with x_0 define exactly the initial state, because $y_0 = 0$ and $\dot{x}_0 = 0$, due to the symmetry of the orbit with respect to the x-axis.

Elliptic Orbits An elliptic orbit in the inertial frame is periodic in the rotating frame only if it is resonant:

$$\frac{n}{n'} = \frac{p}{q} = \text{rational} ,$$

which means that the semi major axis must be given by

$$\frac{(GM)^{1/2} a^{-3/2}}{n'} = \frac{p}{q} .$$

Let us consider now a particular resonance p/q, which means that we keep fixed the semi major axis $a_{p/q}$. The orbit is resonant periodic for *any* eccentricity e, so a *family of elliptic periodic orbits* exists, with the eccentricity as a parameter along the family. There is however another parameter, defining the *orientation* of the elliptic orbit, which is the angle ω of the line of apsides with a fixed direction. In general, an elliptic orbit is not symmetric with respect to the rotating x-axis, contrary to the circular orbits, which are symmetric.

In the space $h - r_0$, where $r_0 = a_{p/q}(1 - e)$ is the pericenter distance ($r_0 = x_0$), an elliptic family is represented by a smooth curve (Fig. 3a), given by the energy integral

$$-\frac{GM}{2a} - n' \sqrt{GMa(1 - e^2)} = h .$$

The value of a is fixed, equal to the corresponding resonance and the eccentricity is a parameter along the family. Note that this presentation is not unique: a point on the elliptic family represents *all* the elliptic resonant orbits with the same eccentricity, but arbitrary orientation ω. An elliptic periodic orbit in the rotating frame is also periodic in the inertial frame. The resonant families of periodic orbits bifurcate from the family of circular orbits, at those points corresponding to the resonant values of the radius $a = a_{p/q}$.

Note that along the circular family the value of the semimajor axis varies, and consequently the ratio n/n'

varies and passes through resonant (rational) values. It is at these points that we have a bifurcation to an elliptic family. Evidently, all the circular and the elliptic orbits are stable, as they are Keplerian orbits.

In the following we study how the above mentioned nice picture of the families of periodic orbits change, when a perturbation is applied, and how instabilities and chaotic regions appear. We consider two cases: the restricted 3-body problem, both circular and elliptic, and the planetary problem, including our Solar System and the extrasolar planetary systems. In all these cases the Hamiltonian is expressed in the form

$$H = H_0 + \epsilon H_1 \,, \tag{22}$$

where H_0 is the integrable Hamiltonian of the two body problem.

Application to the Solar System

A Global View of the Families of Periodic Orbits

We consider the Sun and Jupiter revolving around their common center of mass in *circular orbits* or in *elliptic orbits* and a third body, with negligible mass, moving under the gravitational attraction of these two bodies. We make the approximation that the small body does not affect the motion of the two main bodies, Sun and Jupiter, which we will call *primaries*. This model is the *restricted three-body problem* and an extended study is in the books of Szebehely [46] and Roy [44]. This is a non integrable system, which is a good model to study the motion of a small body in our Solar System, for example an asteroid, a comet, or a small body in the Kuiper belt, at the edge of our Solar System (Jupiter is replaced by Neptune in this latter case).

Let us start with the study of the motion of an asteroid in the asteroid belt, a zone of small bodies between the orbits of Mars and Jupiter. For this reason we define a *rotating* frame xOy, with O the center of mass of the Sun and Jupiter and the x-axis along the line Sun–Jupiter (Fig. 3b). We start our study with the simplest case, considering that the orbits of the Sun and of Jupiter are circular (circular restricted three body problem). In this case the system xOy rotates with *constant* angular velocity n'. We start with planar motion of the asteroid and then extend the study to motion in space. Based on this model, we extend our study by assuming that the orbits of the Sun and Jupiter are elliptic.

In our study we normalize the units of length, mass and time by the relations

$$G = 1\,, \quad (m_{\text{sun}} = 1 - \mu\,,\ m_{\text{J}} = \mu)\,, \quad r_0 = 1\,, \qquad \text{which implies} \quad n' = 1\,,$$

where G is the gravitational constant, r_0 is the radius of the circular orbit of Jupiter around the Sun, and μ, the mass of Jupiter, is considered a small parameter, of the order of 10^{-3} in our case. The Hamiltonian for the motion of the small body is of the form (22), with $\epsilon = \mu$, where H_0 is the Hamiltonian (21) of the two body problem in the rotating frame.

Planar Orbits Let us start with the unperturbed problem, $\mu = 0$, for planar motion, which is the two-body problem in the rotating frame. As we mentioned in the previous section, there exists a family of circular orbits, along which the resonance n/n' varies (n is the mean motion (angular velocity) of the orbit of the asteroid) and families of resonant *elliptic* periodic orbits, which bifurcate from the circular family at all the resonant circular orbits $n/n' = p/q$, as shown schematically in Fig. 3a. Evidently, all the orbits of these families are stable, as they are Keplerian, elliptic, orbits. We study now how all these families evolve and where instabilities appear, when $\mu > 0$, i.e, when the gravitational effect of Jupiter in taken into account. A complete analysis is given in [20].

As we mentioned before, there is an infinite set of resonant periodic orbits along the circular unperturbed family. The continuation of the above mentioned circular family from $\mu = 0$ to $\mu > 0$ and the generation of instabilities depends on the resonances that appear on this family. These resonances belong to three topologically different cases, as far as the continuation to $\mu > 0$ is concerned. These are the cases (i) $n/n' = (\nu + 1)/\nu$, (ii) $n/n' = (2\nu + 1)/(2\nu - 1)$, $(\nu = 1, 2, 3, \ldots)$ and (iii) all other resonances.

(i) All the circular orbits that are not at the resonance $n/n' = 2/1, 3/2, \ldots$, are continued as nearly circular orbits in the rotating frame. The resonant circular orbits $n/n' = 2/1, 3/2, \ldots$, are not continued as periodic orbits in the rotating frame. At these resonances, a gap appears and the single unperturbed family of circular orbits breaks into a set of disconnected families of periodic orbits. From these gaps we have a bifurcation of two families of resonant elliptic periodic orbits (Fig. 4a). The stability of the circular orbits at $\mu = 0$ is preserved, except at the resonances 3/1, 5/3, ..., as we explain below. The resonant elliptic families at the 2/1, 3/2, ... resonances may be stable or unstable, depending on the phase (perihelion or aphelion at $t = 0$) and other factors (for example, close approaches).

(ii) At the circular orbits at the resonances $n/n' = 3/1$, 5/3, ... the continuation to nearly circular orbits is

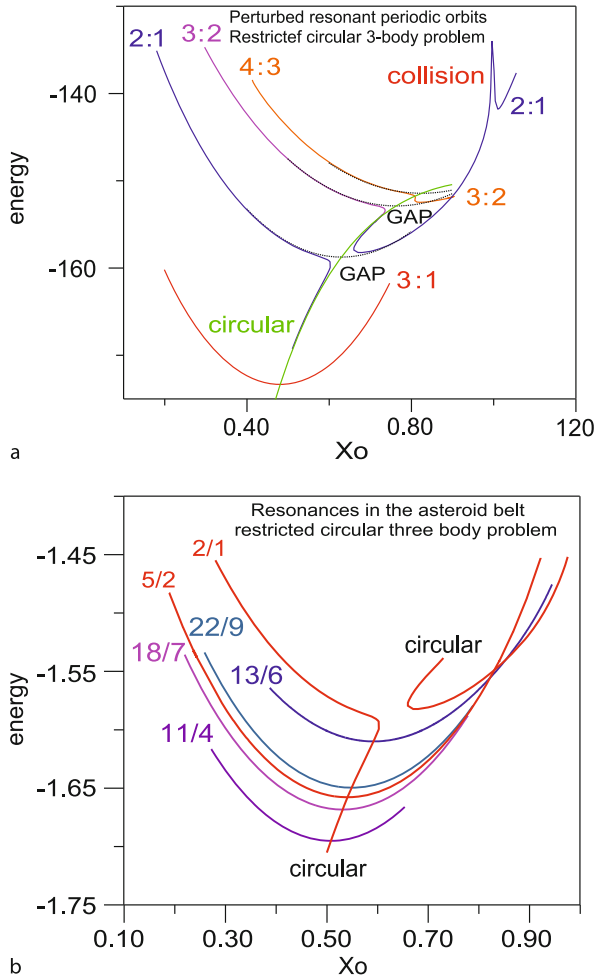

Orbital Dynamics, Chaos in, Figure 4

a The circular family and the bifurcation to resonant elliptic families at the resonances 2/1, 3/2, 4/3, . . . for $\mu = 0.000954786$. The orbits on the elliptic families are symmetric, corresponding to $\omega = 0^0$ or $\omega = 180^0$. These are the only orbits that were continued to $\mu \neq 0$. All the other unperturbed orbits, corresponding to any other value of ω (see Fig. 3c), did not survive the perturbation, as a consequence of the Poincaré–Birkhoff fixed point theorem. b A closer look at some resonant families, for different higher order resonances

possible, but the stability is destroyed. A small unstable region appears at these resonances, on the family of circular orbits. At the critical points, at the two ends of this unstable region, we have a bifurcation of two families of symmetric resonant elliptic periodic orbits, which differ in phase. One is stable and the other is unstable (Fig. 13a).

(iii) In all other resonances on the circular family, for example 5/2, 4/1, 7/3, . . . the circular orbits are continued as nearly circular orbits and in addition the stability is preserved. At these points we have a bifurcation of two families of symmetric resonant elliptic periodic orbits which differ in phase (see Fig. 4b). In this case also, one family is stable and the other unstable (but the stability may change along a family).

A remark is necessary at this point for the families of elliptic periodic orbits for $\mu > 0$. The elliptic unperturbed resonant families, shown in Fig. 3a, are two parametric, with the eccentricity e and the angle of apsides ω as the two parameters. The eccentricity increases along the elliptic family, starting from zero values, but to a fixed eccentricity there corresponds an infinity of values of ω (Fig. 3c). What happens to this two-parametric family of unperturbed periodic orbits as $\mu > 0$? It is proved [20] that, for a fixed eccentricity, out of the infinite set of periodic orbits, for different omegas, only a finite, even, number survive (usually just two), half stable and half unstable. This is a consequence of the Poincaré–Birkhoff fixed point theorem (see [1]). This theorem refers to perturbed twist mappings: In the unperturbed case there exist resonant invariant curves where *all* points are fixed points, so that on this unperturbed invariant curve there exists an infinite number of fixed points. As soon a a perturbation is applied, only a finite number of fixed points survive, half of them stable and half unstable.

Thus, all the elliptic resonant families are monoparametric families along which the eccentricity increases, starting from zero values. In most cases the orbits are symmetric with respect to the rotating x-axis ($\omega = 0$ or $\omega = \pi$) and the eccentricity can be considered as a parameter. Some of these families are stable and others are unstable. The stability depends on the phase, i. e. on whether the asteroid is at perihelion or aphelion when it crosses the x-axis, but also on other factors as, for example, to a close encounter with Jupiter. Along a family of resonant elliptic periodic orbits the resonance is almost constant. A global picture of the circular and the elliptic families is shown in Fig. 4a,b.

Three-Dimensional Orbits in the Circular Model We study now three-dimensional periodic orbits in the model of the circular restricted problem. These families bifurcate from the planar families at those points which are critical with respect to the vertical stability. It is only at these points that the vertical deviations $\xi_3(t)$ of a perturbed orbit, given by the variational equations (14), have a period equal to the period of the planar periodic orbit. We remark at this point that along a resonant family of elliptic periodic orbits, the vertical stability index is very close to crit-

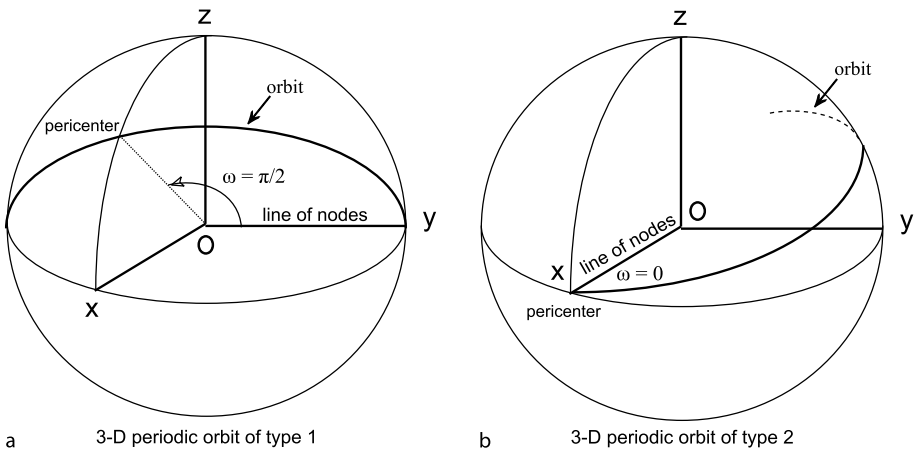

Orbital Dynamics, Chaos in, Figure 5
The two different symmetries in three-dimensional orbits: a Type 1. b Type 2

ical. (It is exactly critical on the unperturbed elliptic family). Depending on the particular resonance, such a critical point may or may not exist.

The three-dimensional periodic orbits are, in general, symmetric and there exist two types of symmetry. Their initial conditions are given below and are shown in Fig. 5:

$$\underline{\textit{Type 1}}: \quad x_{20} = \dot{x}_{10} = \dot{x}_{30}, \quad x_{10}, x_{30}, \dot{x}_{20} \neq 0 .$$

$$\underline{\textit{Type 2}}: \quad x_{20} = x_{30} = \dot{x}_{10} = 0, \quad x_{10}, \dot{x}_{20}, \dot{x}_{30} \neq 0 .$$

In Fig. 6a we present the two families of 2/1 resonant planar periodic orbits, corresponding to $\mu = 0.000954786$. The stable family, for $\omega = 0$, corresponds to the case where the asteroid is at perihelion at $t = 0$. The other family, for $\omega = \pi$, corresponds to position of asteroid at aphelion at $t = 0$ and starts as unstable up to the point where we have a collision with Jupiter. After that point the family continues, for larger eccentricities, as stable. On the stable family, corresponding to perihelion at $t = 0$, there exist two critical points, at high eccentricities, $e = 0.67$ and $e = 0.80$, as far as the vertical stability is concerned, as shown in Fig. 6a. From each one of these two points we have a bifurcation of a family of three-dimensional periodic orbits. One family, starting from $e = 0.67$, belongs to type 1 and is stable (Fig. 7a), while the other family, starting from $e = 0.80$, belongs to type 2 and is unstable (Fig. 8a). Typical three-dimensional periodic orbits on these two families are shown in Fig. 7b and Fig. 8b. Also, on the stable family in Fig. 6a, there exists a bifurcation point, at $e = 0.73$, to two families of periodic orbits of the elliptic problem (see Fig. 9), as explained in the next paragraph. Note that the bifurcation points to three dimensional periodic orbits and to the elliptic problem exist only on the stable family. On the unstable families such bifurcation points do not exist. A remark is necessary at this point: In the unperturbed case ($\mu = 0$) all points on the families of elliptic periodic orbits are critical as far as the vertical stability is concerned and also critical as far as the bifurcation to the elliptic problem is concerned (period equal to 2π). The existence or not of such critical points when $\mu \neq 0$ depends on the particular resonance. In the case we studied here, only the above critical points appeared. In other resonances the situation may be quite different.

A similar situation exists for the 3/2 resonance, as shown in Fig. 6b.

Families in the Elliptic Restricted Problem Families of resonant periodic orbits in the case where the orbits of the Sun and Jupiter are *elliptic*, with eccentricity e_J (elliptic restricted three-body problem) exist, which bifurcate from the families of the circular model, either the circular family or the elliptic families. The bifurcation can take place only at those points where the period of the periodic orbit on the families of the circular model is equal to the period of Jupiter (or a multiple of it). For a fixed value $e_J > 0$ the periodic orbits are isolated. We obtain a family by varying e_J.

Continuation from the Family of Circular Orbits Let us start from the unperturbed circular family. The period of a circular orbit ($\mu = 0$) is $T = 2\pi/(n - n')$ in the rotating frame and the period of Jupiter is $T_J = 2\pi/n'$, where n' is its mean angular velocity. We have $T = T_J/(n/n' - 1)$. At the resonance $n/n' = p/q$ we have $T = T_J q/(p - q)$ and if this orbit is described $p - q$ times, the period T^* of this orbit is an integral multiple of T_J, $T^* = qT_J$. If

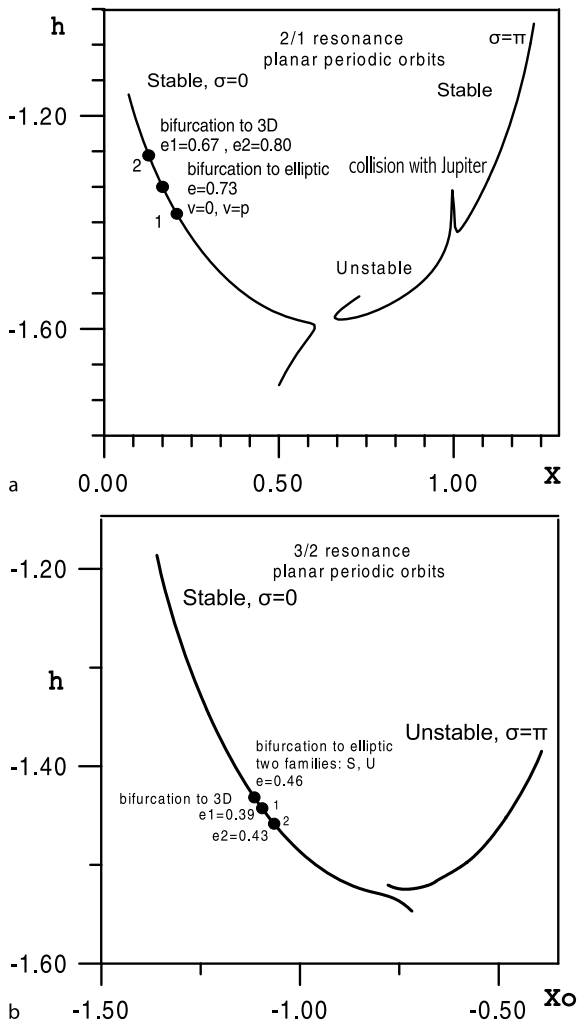

Orbital Dynamics, Chaos in, Figure 6

a The two elliptic families of resonant periodic orbits at the 2/1 resonance and the bifurcation points to three-dimensional orbits, at $e = 0.67$ and $e = 0.80$ and also to the elliptic model, at $e = 0.72$ on the stable family. b The resonant elliptic periodic family at the 3/2 resonance and the bifurcation points to three-dimensional orbits, at $e = 0.39$ and $e = 0.43$ and also to the elliptic model, at $e = 0.46$

$n/n' \neq 2/1, 3/2, \ldots$ the region around this resonant orbit is continued to $\mu > 0$, as mentioned above. On the continued circular family there exists an orbit which, if described $p - q$ times, has a period exactly equal to qT_J. This means that a bifurcation from the circular family $\mu > 0$ to a family of the elliptic problem can take place close to a resonance. This is the case with the 3/1 resonance (Fig. 13a).

Continuation from the Family of Nearly Elliptic Orbits Consider a family of $n/n' = p/q$ resonant elliptic periodic orbits of the circular planar problem, for $\mu = 0$. The period all along the family is constant, and according to the above, if the orbits of the family are described $(p - q)$ times, the period is $T^* = qT_J$. This family is continued, when $\mu > 0$, to two families of elliptic periodic orbits, differing in phase. Along each family the eccentricity increases, starting from zero values. For continuity reasons, the (multiple) period along the continued family is close to qT_J. If at a certain point, corresponding to a value e of the eccentricity, it happens to be *exactly* equal to qT_J, then a bifurcation to the elliptic problem can take place. Two families of periodic orbits exist, along which the eccentricity of Jupiter increases. For a fixed eccentricity of Jupiter, for example $e_J = 0.048$, only two *isolated* periodic orbits exist. The above mentioned two families differ in the initial phase of Jupiter on its elliptic orbit at $t = 0$, i. e. if it is at perihelion or aphelion. In general, one family is stable and the other is unstable.

The numerical computations have shown that in certain resonances, for example 2/1, 3/2, 3/1, such bifurcation points do exist, at quite large values of the eccentricity (see Fig. 6a,b for the 2/1 and 3/2 resonances and Fig. 13a for the 3/1 resonance). But in other resonances, for example 7/3, such bifurcation points do not exist. This plays an important role on the topology of the phase space close to a particular resonance, because the existence of a resonant periodic orbit/fixed point of the Poincaré map, determines the topology of the phase space. The non existence of periodic orbits in a region implies that the phase space is smooth and ordered regions exist. A systematic study along these lines has been made by Tsiganis et al. [47,48,49]. We present in Fig. 9, as an example, two resonant families of the elliptic problem, at the 2/1 resonance, one stable and one unstable. These two orbits bifurcate from the point on the stable branch of the 2/1 resonant family of the circular problem, at $e = 0.72$, as shown in Fig. 6a.

Generation of Chaos at the Unstable Periodic Orbits

Let us consider the simplest model, the circular restricted three-body problem, and study the topology of the phase space at the 2/1 and the 3/2 resonances, using as a guide the families of periodic orbits as presented in Fig. 4a and Fig. 6a,b. We compute the Poincaré map on the surface of section $y = 0$, $H = h$, for different values of the energy h. These energy levels can be visualized by considering lines parallel to the x_0 axis in Figs. 4a (or 6a,b) at different values of h. Note that these lines intersect the circular family and the resonant families, and these intersections correspond to the fixed points of the Poincaré map. For a better understanding of the physics, we mark on each map the

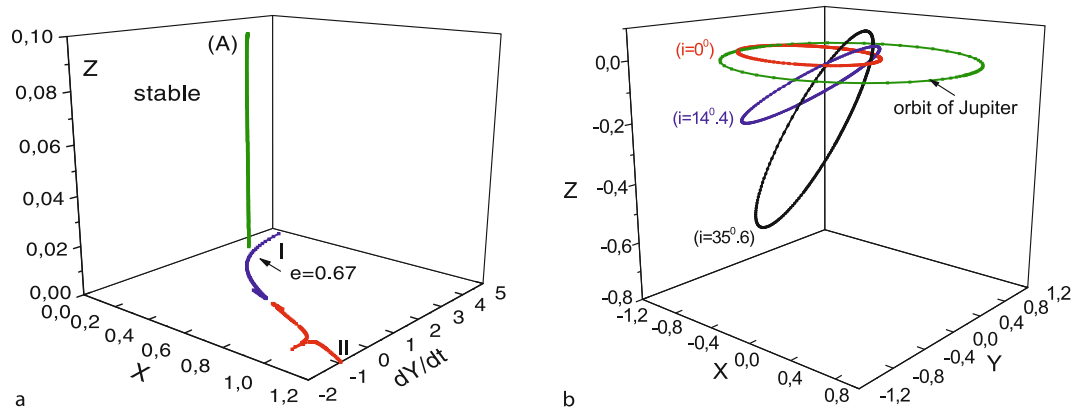

Orbital Dynamics, Chaos in, Figure 7

a The stable family of three-dimensional periodic orbits bifurcating at $e = 0.67$ b Three-dimensional periodic orbits on the stable family (type 1)

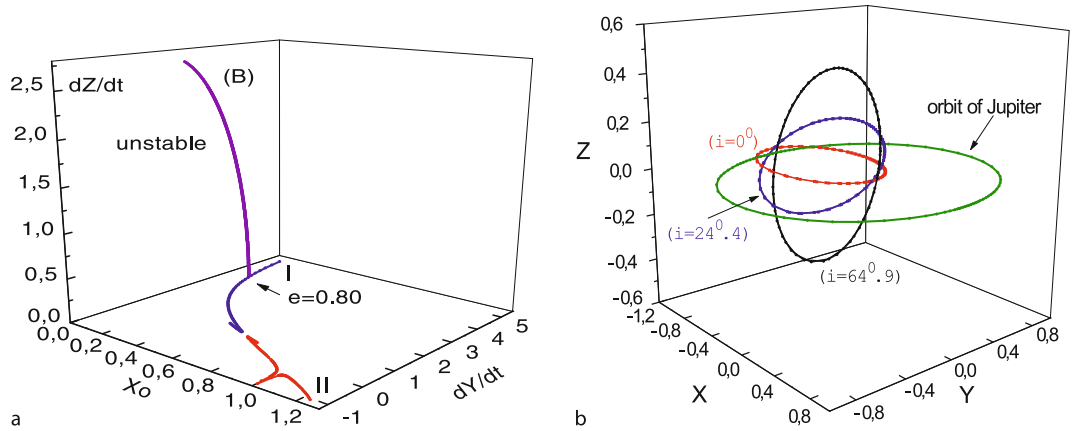

Orbital Dynamics, Chaos in, Figure 8

a The unstable family of three-dimensional periodic orbits bifurcating at $e = 0.80$. b Three-dimensional periodic orbits on the unstable family (type 2)

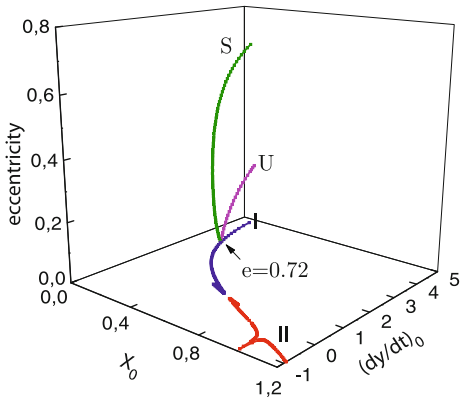

Orbital Dynamics, Chaos in, Figure 9

The two families of periodic orbit of the elliptic restricted three-body problem, bifurcating from the 2/1 resonant family at $e = 0.72$. One is stable and the other unstable

value of the eccentricity of the stable resonant fixed point, instead of the energy h, since along the family the eccentricity increases. In Fig. 10 we present several surfaces of section, at different energy levels, corresponding to different eccentricities, at the resonances 2/1 and 3/2. The fixed points corresponding to the circular periodic orbit (in the middle of the diagram) and to the stable and unstable resonant periodic orbits are clearly seen. The stable fixed points are surrounded by islands (closed invariant curves) of ordered motion, while the mapping close to the unstable fixed points is hyperbolic. Chaotic motion starts at these unstable points as the eccentricity increases as we move along the family. The chaotic orbits appear as scattered points, in contrast to the regular orbits, which are represented by smooth invariant curves. This phenomenon is stronger at the 3/2 resonance.

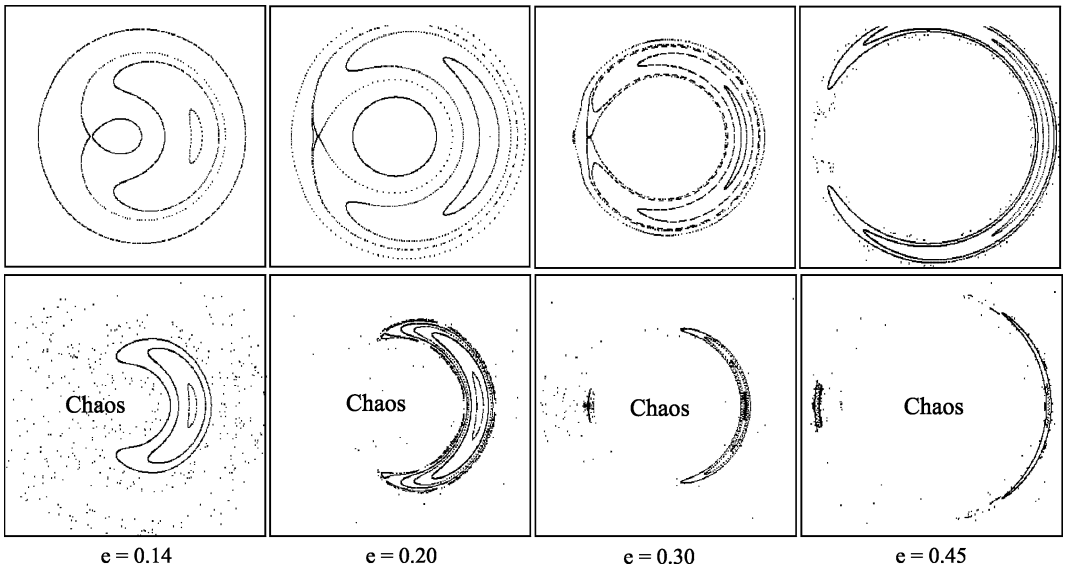

Orbital Dynamics, Chaos in, Figure 10
The Poincaré map at the 2/1 resonance (*upper row*) and the 3/2 resonance (*lower row*) for different energy levels, presented here by the corresponding eccentricity *e* of the resonant stable fixed point on the elliptic family. The stable and the unstable fixed points are clearly seen. The generation of chaotic motion at the unstable fixed points is evident. Note that the chaotic motion starts from the unstable fixed points

Note that the topology of the phase space, on the Poincaré map, is critically determined by the position of the fixed points and their stability character. This is the reason that the knowledge of the basic families of periodic orbits is so important for the study of the dynamics of the system.

Asteroid Motion Close to a Resonance

It is known that in the region between the orbits of Mars and Jupiter there exists a zone of small bodies, revolving around the Sun, called *asteroid belt*. It has been observed that the distribution of these bodies is not smooth, but gaps exist at several resonances between the mean motion of the asteroid and Jupiter, the famous Kirkwood gaps (see Fig. 12). The explanation of these gaps was an open question for many decades, and their existence was explained by realizing that the motion at the 3/1 resonance (and in many other resonances) is chaotic and consequently an asteroid could not stay in this region for a long time. The first study was made by Wisdom [53,54,55]. The study was based on the construction of a symplectic mapping model, by making use of the averaged Hamiltonian of the elliptic restricted three-body problem at the 3/1 resonance. It was shown that, due to the existence of chaos at this region, the eccentricity of an asteroid that starts its motion in a nearly circular orbit undergoes sudden jumps, after a period which may be several million years (the semi major axis remaining almost constant), and thus the orbit of the asteroid may by Mars or even Earth crossing and thus undergo additional perturbations that will eventually drive it outside the 3/1 resonance region. Several papers followed this study, for many resonances in the asteroid belt, which used mapping models based on an averaged Hamiltonian at the corresponding resonance [1,15,17,21]. For the different methods used to transform the continuous flow to a mapping model see [16]. Much work on the asteroid belt has been also made by making use of the averaging method or a combination of this method and numerical integrations [22,35,37,38,39].

The variables used in the averaged models are the Delaunay variables, transformed to resonant action angle-variables (see [41]). For example, for the 3/1 resonance, for planar motion these variables are

$$S = \sqrt{\mu_1 a}\left(1 - \sqrt{1-e^2}\right), \quad \sigma = \frac{1}{3}(3\lambda' - \lambda) - \omega,$$
$$N = \sqrt{\mu_1 a}\left(3 - \sqrt{1-e^2}\right), \quad \nu = -\frac{1}{3}(3\lambda' - \lambda) + \omega',$$

where $\mu_1 = 1 - \mu$, $e' = 0.048$, and λ, ω, a are the mean longitude, the longitude of perihelion and the semimajor axis of the asteroid and the corresponding primed quantities refer to Jupiter. The variables used to present

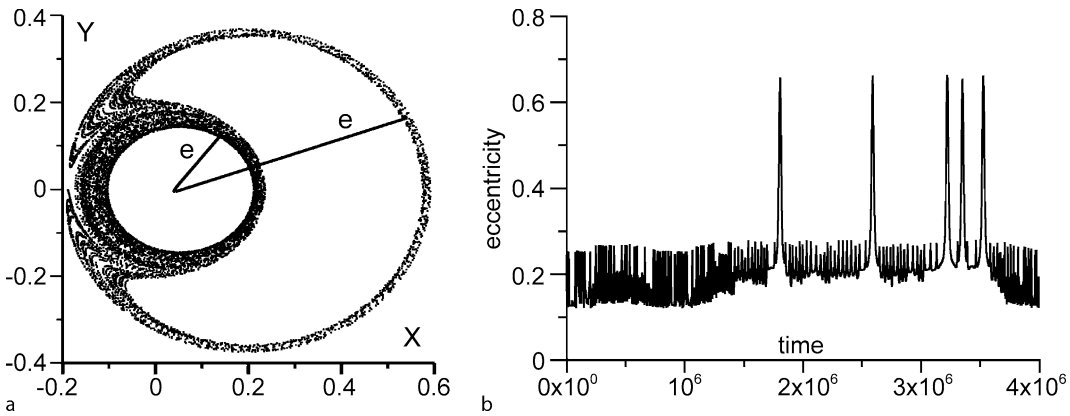

Orbital Dynamics, Chaos in, Figure 11

a The mapping, in the variables $X = e\cos(\sigma) - Y = e\sin(\sigma)$ for the motion of an asteroid at the 3/1 resonance. b The evolution of the eccentricity. Chaotic jumps of the eccentricity appear, at unpredictable times

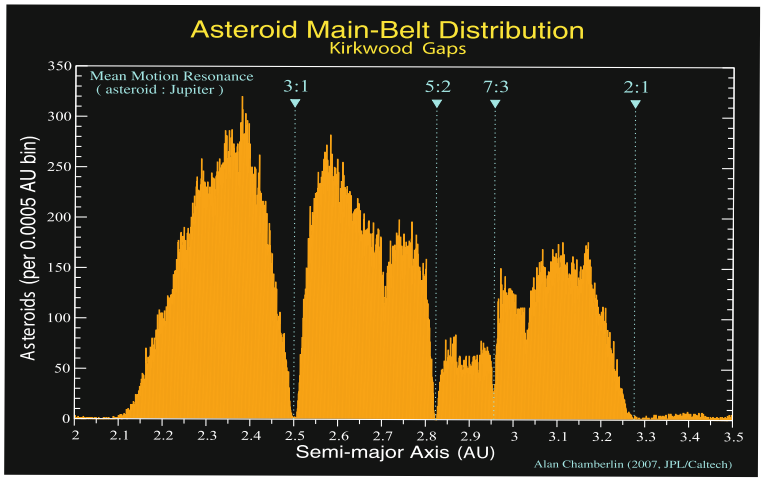

Orbital Dynamics, Chaos in, Figure 12

The distribution of the asteroids, obtained from 156 929 asteroids, as given by JPL/Caltech in 2007. The Kirkwood gaps at the 3/1, 5/2, 7/3 and 2/1 mean motion resonances are clearly seen

the mapping are the Poincaré variables $X = \sqrt{2S}\cos(\sigma)$, $Y = \sqrt{2S}\sin(\sigma)$, which are also canonical variables.

In Fig. 11a we present the mapping for an asteroid at the 3/1 resonance, from a mapping model used by Hadjidemetriou [15], equivalent to the map used by Wisdom [53]. The variables are similar to the Poincaré variables, but instead of $\sqrt{2S}$ we used the eccentricity e (note that $\sqrt{2S}$ is proportional to e, for small values of e). At the beginning the asteroid moves along the inner "diffused" circle, with small radius, corresponding to low values of the eccentricity, but eventually comes close to the chaotic region which connects the inner circle with an outer circle, with larger radius, corresponding to larger values of the eccentricity. So, through this chaotic window we have a connection between the low eccentricity regions and the high eccentricity regions. This results to a chaotic jump between small and large eccentricities, in a chaotic, unpredictable, way. This behavior is called *intermittency*. This is clearly shown in Fig. 11b.

At this point we draw the attention to an important point when we use the averaged Hamiltonian in the dynamical study. Since the averaging method is based on series expansions in a small parameter (in our case it is the eccentricity), it is not valid for high values of the parameter. In the present case, in the study of the asteroid at the 3/1 resonance, the averaged Hamiltonian used to construct the mapping which gives the evolution of the asteroid eccentricity does not contain the high eccentricity res-

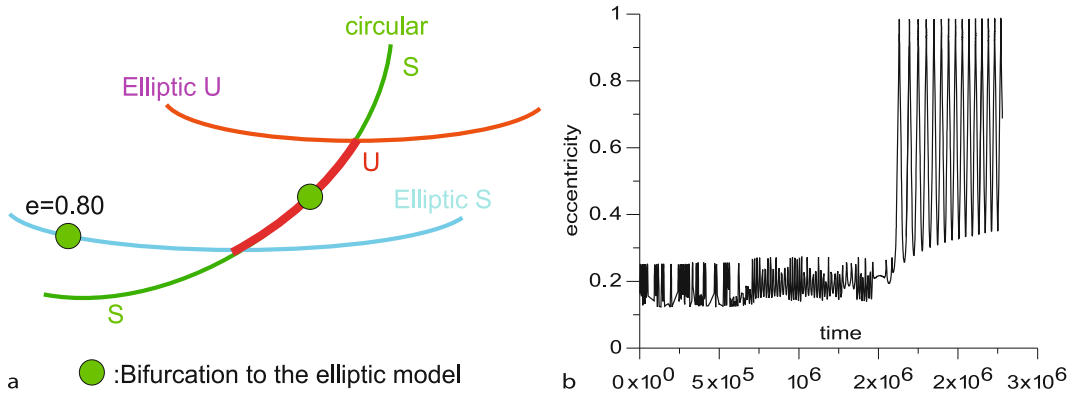

Orbital Dynamics, Chaos in, Figure 13

a The families of circular and elliptic periodic orbits of the circular model at the 3/1 resonance, and the bifurcation points to the elliptic model (schematically). S stands for stable and U for unstable. b The evolution of the eccentricity when the high eccentricity resonances are included in the model. The chaotic jumps are now up to eccentricities equal to 1

onances. This is the case with the evolution of Fig. 11. For this reason it is important, in constructing the averaged model, to know the topology of the *whole* phase space, and this can be done only if we know *all* the resonant families of periodic orbits at the 3/1 resonance (and of course in all other similar studies in other resonances). *A necessary criterion for the validity of the averaged Hamiltonian is its fixed points to coincide with the periodic orbits (fixed points of the Poincaré map) of the original model* (the elliptic restricted three-body problem in this case). This shows the importance of the periodic orbits in orbital dynamics. In Fig. 13a we show, schematically, for the model of the circular restricted problem, the family of circular orbits and the unstable region which appears at the 3/1 resonance on the circular family, and also the two families of elliptic periodic orbits that bifurcate from the critical points at the two ends of this unstable region. One family is stable and the other is unstable. It is found that on the unstable part of the circular family there exists a bifurcation point to two families of 3/1 resonant periodic orbits of the elliptic model, which start with zero eccentricities. Both of them are unstable. These are the *low eccentricity resonances* that are included in the model of Fig. 11. However, there exists one more bifurcation point, at the eccentricity $e = 0.80$ on the stable family of elliptic periodic orbits (of the circular model), from which two 3/1 resonant periodic orbits of the elliptic model appear, one stable and the other unstable, starting with high eccentricities, equal to $e = 0.80$. For a full description of the resonant structure of the restricted three-body problem at the 3/1 resonance see [14,15]. It is these *high eccentricity resonances* that are missing from the model of Fig. 11. If these high eccentricity resonances are also included in the model, the jumps in the eccentricity are higher, up to $e = 1$ and thus the asteroid not only approaches the inner planets, but may also fall on the Sun. This evolution is shown in Fig. 13b.

The study of the ordered and chaotic regions in the asteroid belt is not the only such study in our Solar System. A zone of small bodies, similar to the asteroid belt, exists at the edge of our Solar System, after the orbit of Neptune. This is the Kuiper belt, whose existence was conjectured to explain the source of low period comets. Since the last decade of the 20th century many small bodies were observed in the Kuiper belt and it was realized that ordered and chaotic regions exist in this region also, similar to those in the asteroid belt, at several resonances with Neptune. *Pluto* is one such body in the Kuiper belt, trapped at the 3/2 resonance with Neptune, together with many other smaller bodies at the same resonance, called *plutinos*. A good view of the dynamical structure in the Kuiper belt is given in [6].

All major planets, Jupiter, Saturn, Uranus, Neptune, have planetary rings, the most well known being the rings of Saturn. Although many of the properties of the ring systems can be understood by a fluid dynamics approach, several of their features are explained by resonant dynamics, as in the case of the asteroid belt or the Kuiper belt. The fine structure of the rings can be explained by resonances between the ring particles and small satellites of the planet. A description of the dynamics of the planetary rings can be found in Chap. 10 in [41].

The chaotic behavior of the Solar System, as a whole, is yet another interesting subject, and there are several numerical works on this problem, notably by Laskar [25, 26,27] and by Wisdom [56]. There are not large scale chaotic orbits of the planets, although the system is non

integrable and some chaos is expected. Especially the large planets do not show any significant change and their orbits stay, for some billion years, close to their present orbits. The inner planets, especially Mercury, have shown large deviations, but due to the chaotic nature of the Solar System and the fact that the numerical integrations are not with infinite accuracy, these results may not represent the actual evolution of the Solar System. It seems that the Solar System is stable and any chaotic motion is in a small scale and is bounded. Studies on the stability of extrasolar planetary systems have started recently, with many interesting results, as we explain in the next section.

Another interesting case of chaotic motion in the Solar System refers to the rotational motion of celestial bodies. Although the rotation of the planets is regular, there are small bodies, with irregular shape, that show chaotic rotation. Such a case is the satellite of Saturn, Hyperion, with approximate dimensions 180 km × 140 km × 112.5 km. Although its orbit is stable, due to the fact that it is at the 4:3 resonance with the more massive satellite of Saturn, Titan, its rotation is chaotic [57].

Extrasolar Planetary Systems

Some General Remarks

In the last decade of the 20th century it was discovered that our Solar System is not the only planetary system in the universe. Up to the present (May 2008) there are 281 observed extrasolar planetary systems, with 25 of them having two or more planets. In many planetary systems with two planets close to each other, the two planets are in mean motion resonance. Examples are: HD 82943 [23,33], GLIESE 876 [31,43], at the 2:1 resonance and 55Cnc at the 3:1 resonance [32]. Some of these systems have large eccentricities and are evidently stable.

There are different approaches to the study of the dynamical evolution of a planetary system and on the mechanisms that stabilize the system, or generate chaotic motion and instability: Beaugé and Michtchenko [2], Beaugé et al. [3,4,5], Ferraz-Mello et al. [8], Gozdjiewski et al. [10], Malhotra [30], Lee and Peale [29], Lee [28]. In these papers different methods have been applied, as the averaging method, direct numerical integrations of orbits, or various numerical methods which provide indicators for the exponential growth of nearby orbits. In this way the regions where stable motion exists have been detected, in the orbital elements space.

We present briefly a global view of the structure of the phase space of a planetary system with two planets, moving in the plane, as obtained from the set of the families of periodic orbits. As we have already mentioned before, the periodic orbits play a dominant role in understanding the dynamics of a system, because they determine critically the structure of the phase space. In this way, we can detect the regions where stable librations could exist. These will be the regions where a real planetary system could exist in nature. As we will see, stable regions corresponding to elliptic orbits of the two planets with relatively large eccentricities are associated with mean motion resonances. An early work on periodic orbits of the planetary type is by Hadjidemetriou [12], well before the first extrasolar planetary systems were observed. Many papers followed on these lines, after the first extrasolar planetary systems were observed [18,42,51,52]. We remark that stable motion could also exist far from resonances, if the eccentricities are small. This latter motion is close to a stable periodic orbit of the *circular family* of periodic orbits. We also remark that it is possible to have stable motion far from a periodic orbit, but in this latter case the two planets are not close to each other, so that their gravitational interaction is not very significant.

It can be proved [11] that families of periodic orbits in the planar general three body problem exist, in a *rotating* frame xOy, whose x-axis is the line $S - P_1$, with origin

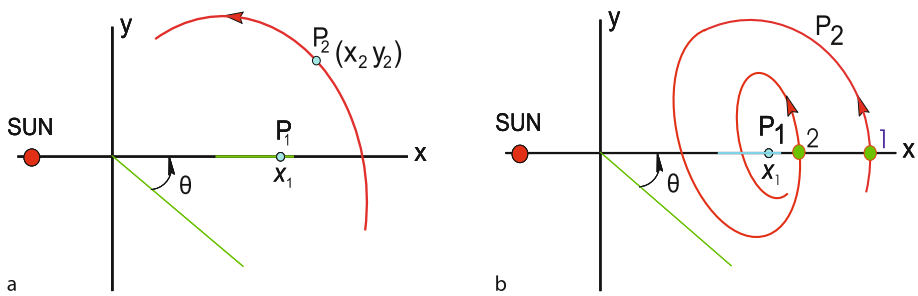

Orbital Dynamics, Chaos in, Figure 14
a The rotating frame. The planet P_1 moves on the x-axis and the planet P_2 in the xOy plane. The angle θ is an ignorable coordinate. **b** The Poincaré map at $y_2 = 0$

at the center of mass of these two bodies, where S is the Sun and P_1 the inner planet. We assume that the center of mass of the whole system is at rest with respect to an inertial frame. We have four degrees of freedom, for planar motion, with generalized variables x_1, x_2, y_2, θ (Fig. 14a). This is a non uniformly rotating frame, and the second planet P_2 moves in the plane xy. It turns out [11] that the angle θ is ignorable, so we have three degrees of freedom in the rotating frame, with variables x_1, x_2, y_2. In the planetary three body problem (one big body, the star and two small bodies, the planets) the periodic orbits are similar to the families of the restricted problem as shown in Fig. 4. There are two types of periodic orbits:

- Non resonant periodic orbits with nearly circular orbits of the two planets.
- Resonant periodic orbits with nearly elliptic orbits of the two planets.

The circular orbits are all symmetric but the elliptic orbits may be symmetric or asymmetric. There exist families of elliptic periodic orbits for every mean motion resonance. Close to a stable periodic orbit there exists a region of stable librations, and it is at these regions that a planetary system could be trapped.

Concerning the continuation of the unperturbed family of periodic orbits ($m_1 = m_2 = 0$), the situation is similar to that explained in the restricted three body problem. There are three topologically different resonant cases:

- The resonances of the form $(n+1)/n$, $(2/1, 3/2, \ldots)$ (Gaps on the circular family).
- The resonances $(2n+1)/(2n-1)$, $(3/1, 5/3, \ldots)$ (Instability on the circular family).
- All other resonances, $(5/2, 7/3, 8/3, \ldots)$ (Preservation of the stability on the circular family).

A global view of the resonant families of elliptic periodic orbits, for each one of the above resonance types can be found in [19]. There exist both *symmetric* and *asymmetric* families. The ratio of the planetary masses plays an important role on the stability and the existence of asymmetric families of periodic orbits. The sum of the masses of the planets also plays an important role on the stability and the existence of families of resonant periodic orbits. The stability of a symmetric periodic orbit depends, all other things being the same (semimajor axes, eccentricities), on the phase of the two planets, that is on whether the line of apsides are aligned or antialigned and on the position of the two planets at perihelion of aphelion at some epoch. The proper phase generates a *phase protection mechanism* so that stable planetary systems exist even for large eccentricities.

The properties of motion close to a periodic orbit are studied by considering a Poincaré map on the surface of section $y_2 = 0$, $(\dot{y}_2 > 0)$, $H = h =$ constant (Fig. 14b). The phase space of the Poincaré map is the four dimensional space $x_1, \dot{x}_1, x_2, \dot{x}_2$ ($y_2 = 0$ and $\dot{y}_2$ is obtained from $H = h, \dot{y}_2 > 0$). Close to a stable periodic orbit we have stable librations and the motion in phase space takes place on a torus. On the contrary, close to an unstable periodic orbit we have irregular, chaotic, motion and in many cases the system disrupts into a binary system (the star and one planet) and an escaping planet.

The position of some real extrasolar planetary systems is compared with the above mentioned regions of stable librations. A detailed analysis of the dynamics of extrasolar planetary systems based on the families of periodic orbits is presented in [20].

In the following, we present, as an example, the dynamics of a real extrasolar planetary system, Gliese 876, on the whole phase space, and study the stable configurations and the regions where chaotic motion appears.

A Real Extrasolar Planetary System: Gliese 876

Studies on the dynamical evolution of a planetary system, both theoretical and for real extrasolar planetary systems, have been made by different methods. One way to study the problem is to compute many orbits, for a set of initial conditions and study their behavior for a long time. A different method is to use the averaging method in order to obtain an averaged Hamiltonian, thus reducing the number of degrees of freedom. Analytic and numerical studies can then be made to find the stable regions in phase space [2,3,4,5] Ferraz-Mello et al., Gozdjiewski et al. [2], [28,29,30,45]. A systematic study of the orbital dynamics in planetary systems can be made by finding all the basic families of periodic orbits. As we mentioned before, the position and the stability character of the periodic orbits define the topology of the phase space, and in this way we find all the stable regions, close to the stable periodic orbits, where a planetary system can be trapped, and we also find the chaotic regions, close to the unstable periodic orbits, where planetary system could not exist [19].

The ordered and chaotic regions in an extrasolar planetary system, the factors that affect the stability and the mechanism of generation of chaos, will be presented here by an example from a real extrasolar planetary system, Gliese 876 [31]. This is a planetary system 15.4 light years far from our solar system. The mass of the sun in this system is equal to $m_0 = 0.32$ solar masses and the masses of the planets P_1, P_2 are $m_1 \sin i = 1.89$ MJ and $m_2 \sin i = 0.56$ MJ, where MJ stands for the mass of

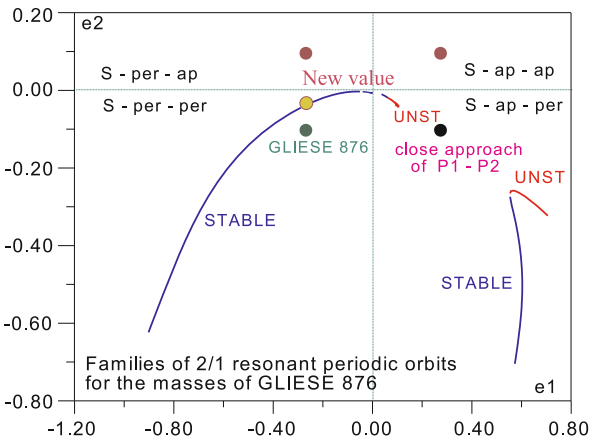

Orbital Dynamics, Chaos in, Figure 15

The families of periodic orbits at the 2/1 resonance in the space of the eccentricities. $e_i > 0$ means position at aphelion and $e_i < 0$ position at perihelion

Jupiter (i is the inclination of the orbital plane of this system with respect to the line of sight from us, and it is not known). The semimajor axes, the eccentricities and the periods of the planetary orbits are: $a_1 = 0.13$ AU, $a_2 = 0.21$ AU, $e_1 = 0.27$, $e_2 = 0.10$, $T_1 = 30.1$ days and $T_2 = 61.02$ days. The perihelia of the two planetary orbits are in the same direction. This is a system very close to the 2/1 resonance, $T_2/T_1 = 2.03$, and for this reason we study all the families of resonant periodic orbits at the 2/1 resonance, for the masses of this system (assuming $\sin i = 1$).

In Fig. 15 we present the families of resonant 2/1 periodic orbits for the masses of Gliese 876, in the space of the planetary eccentricities e_1 e_2. We used the convention that $e_i > 0$ means position of the planet at aphelion and $e_i < 0$ position at perihelion. In this way the space of the eccentricities is divided into four sections, according to the sign of the eccentricities, as shown in Fig. 15. For $e_1 < 0$, $e_2 < 0$ and $e_1 > 0$, $e_2 > 0$ the perihelia of both planets are in the same direction, while for $e_1 > 0$, $e_2 < 0$ and $e_1 < 0$, $e_2 > 0$ the perihelia are in opposite directions. We may also note that due to the 2/1 resonance, the phases where, for the

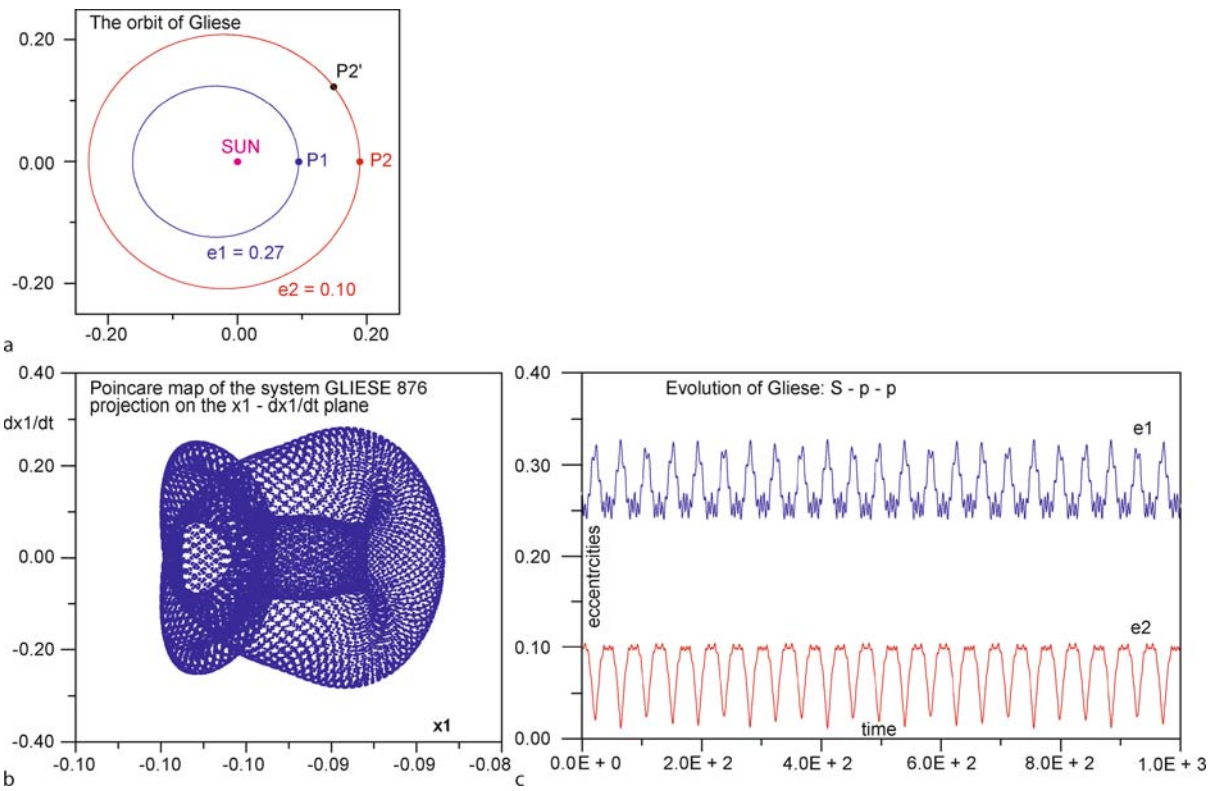

Orbital Dynamics, Chaos in, Figure 16

a The orbit, corresponding to $e_1 < 0$, $e_2 < 0$. b The Poincaré map: projection on the line $x_1\dot{x}_1$. The motion is ordered. b The evolution of the eccentricities.

same position of P_1, the position of P_2 is at perihelion or in aphelion are equivalent, corresponding to $t = 0$ and to $t = T/2$, respectively, where T is the period.

There are two families that start from the region $e_1 \approx 0$, $e_2 \approx 0$. At $e_1 = e_2 = 0$ there is a gap, similar to the gap on the family of circular orbits of the restricted three body problem, as shown in Fig. 6. The first family corresponds to $e_1 < 0$, $e_2 < 0$, along which the eccentricities of the two planets increase. In all orbits of this family the perihelia are in the same direction and at $t = 0$ both planets are at perihelia. This family is stable, even for large values of the orbital eccentricities. Another family exists, for $e_1 > 0$ and $e_2 < 0$. In this family the perihelia of the two planets are in opposite directions and at $t = 0$ the planet P_1 is at aphelion and the planet P_2 is at perihelion. This family presents a gap at the region $e_1 = -0.2$, $e_2 = 0.4$, because the two planets are close to each other and the gravitational attraction between them is so strong (for the given masses) that a resonant 2/1 orbit cannot survive. This part of the family, from zero eccentricities up to the gap, which corresponds to small eccentricities, is unstable. But after this close approach region, the family continues with large eccentricities, and this part is now stable.

In the space of the eccentricities of Fig. 15 we placed a planetary system with the same semimajor axes and eccentricities as Gliese 876, but with different phases. One of these positions, for $e_1 < 0$, $e_2 < 0$, is very close to the stable family. In Figs. 16, 17, and 18 we present the evolution of each of these systems (with the same elements a_i, e_i as Gliese 876), by making use of the Poinaré map on the surface of section defined in Fig. 14b. We note that the real system (green circle in Fig. 15) is in an ordered region (Fig. 16), with the eccentricities undergoing quasi periodic variations, and the projection of the Poincaré map on the

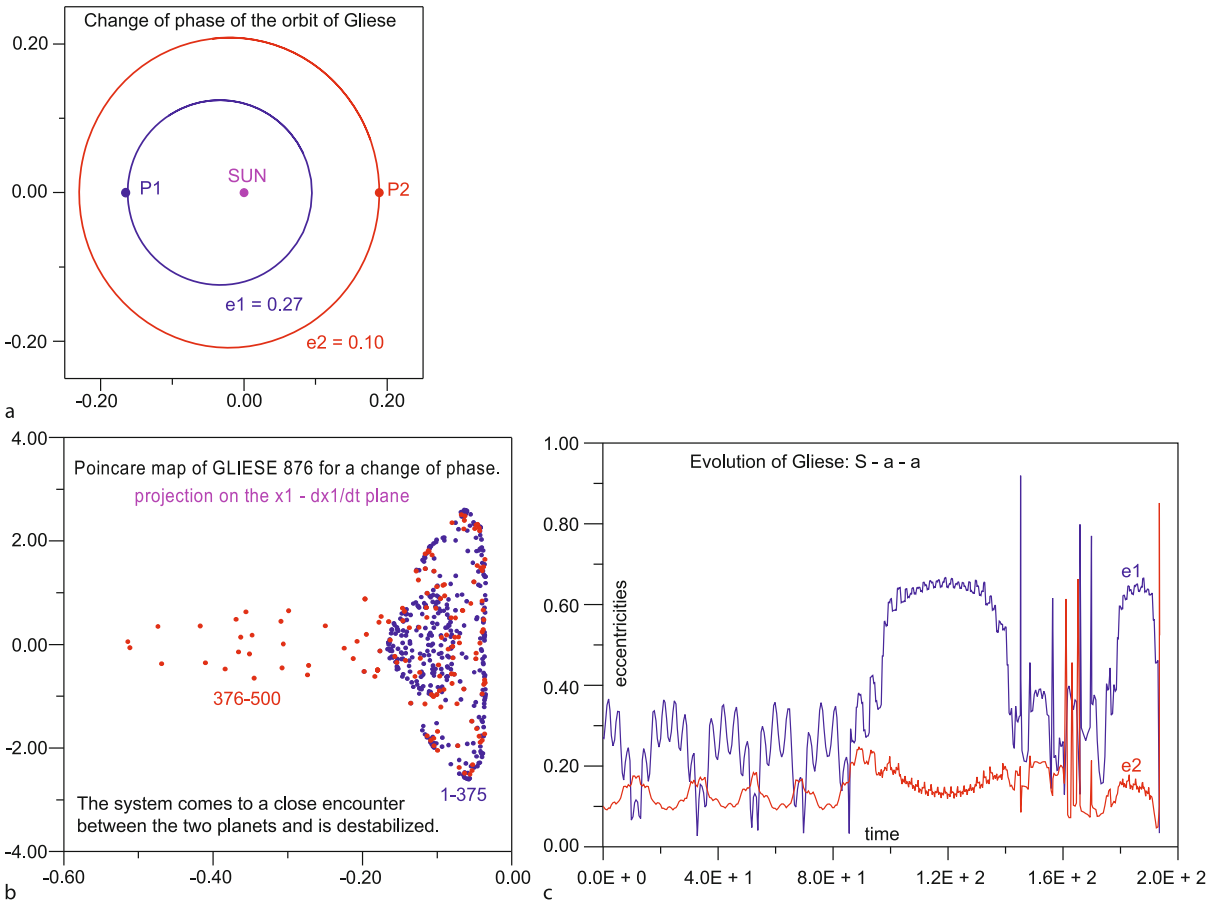

Orbital Dynamics, Chaos in, Figure 17

a The orbit, corresponding to $e_1 > 0$, $e_2 > 0$. b The Poincaré map: projection on the plane $x_1\dot{x}_1$. The motion is chaotic. c The evolution of the eccentricities

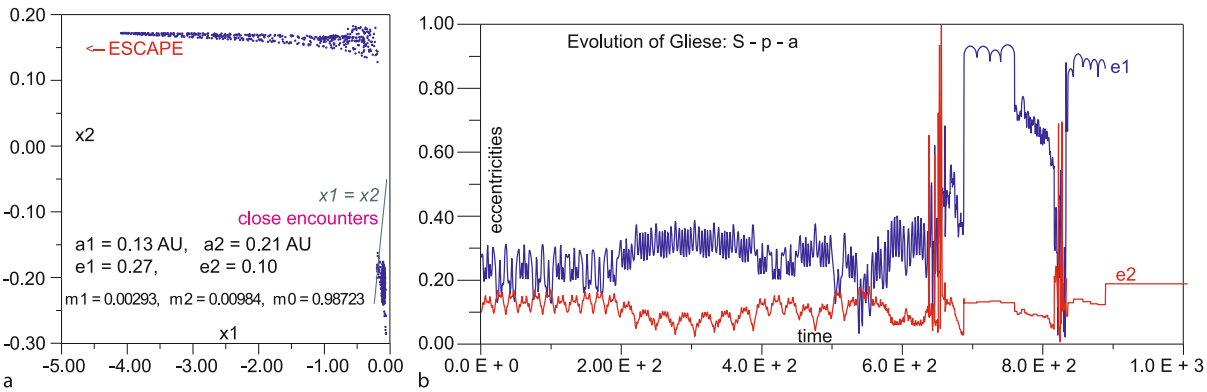

Orbital Dynamics, Chaos in, Figure 18

The orbit, corresponding to $e_1 < 0, e_2 > 0$. a The Poincaré map: projection on the plane $x_1\, x_2$. There exist points close to the $x_2 = x_1$ line, corresponding to close encounters between the two planets. The motion is chaotic. b The evolution of the eccentricities

$x_1\dot{x}_1$ plane is a nice surface (the same holds for the projection in all other planes of the four dimensional phase space of the Poincaré map). All other configurations however are unstable and present chaotic behavior, although the orbital elements are the same and only the phase differs. In Fig. 17, corresponding to $e_1 > 0$, $e_2 > 0$, chaotic behavior develops after a rather ordered motion, and the system disrupts. In Fig. 18, corresponding to $e_1 < 0$, $e_2 > 0$, also chaotic behavior develops after a long time of rather ordered motion. From the Poincaré map, which is given in its projection in the $x_1\, x_2$ space, we see that the mechanism of generation of chaos is the close encounters between the two planets, shown by the several points of intersection close to the line $x_2 = x_1$ (this is a real encounter and not just due to the projection from the four dimensional space $x_1\dot{x}_1\, x_2\dot{x}_2$ to the two dimensional plane $x_1\, x_2$, because P_1 is always on the x-axis and P_2 is also on the x-axis, which implies $y_2 = 0$, due to the definition of the map (see Fig. 14b)).

From the above we see that the phase of the two planets (perihelia in the same or in opposite directions, position of the planet at perihelion or aphelion at $t = 0$) plays a crucial role on the stability of the system. As we have seen, the stable regions are close to the stable periodic orbits, and this makes clear the importance of knowing all the families of periodic orbits. In this way we are in a position to know in what regions of the orbital elements a planetary system could exist in nature and what are the regions where a planetary system cannot exist. We note also that the orbital elements for Gliese 876 that we used in the above study were revised, as more accurate observations were taken into account. The new values correspond to a position almost *on* the stable family, as we show in Fig. 15 (yellow circle).

Future Directions

The model of the restricted three-body problem has been studied for almost a century and most of its dynamical aspects are now known. This is not so for the general, planetary, three-body problem, where several aspects of the dynamics are not yet well studied. One reason is that the phase space has more dimensions than the restricted problem. Though the motion in the plane is quite well understood, because all the basic resonant and non resonant periodic orbits (symmetric and asymmetric) are well known, the three dimensional motion is not completely studied. The main reason for this is that the observational data for the extrasolar systems are not yet accurate enough to give information on three dimensional planetary motion. The knowledge of the basic three dimensional families will give a clear picture of the topology of the phase space and of the regions where a three dimensional planetary system could be trapped.

Another problem in the study of the extrasolar planetary systems is the explanation of large planetary eccentricities. Evidently, such systems are stable, since they are observed in nature, and we know from the studies up to now that such high eccentricity planetary systems can be stable, provided we have the right phase. But how did these systems reach their present configuration? It has been proposed that they were generated as low eccentricity systems and reached the present configuration following a *migration* process. A dissipation is needed for such an evolution and several mechanisms have been proposed. It is possible that a planetary system can trapped in a stable configuration, possibly with high eccentricities, due to the migration process. It is the stable periodic orbits that correspond to

these stable configurations. Some work has been done on this problem [36], but more work is needed.

An open problem in orbital dynamics is the study of the early history of our Solar System. This study involves calculations of the N-body problem. It is believed that the orbits of the giant planets of our Solar System, from Jupiter and beyond, migrated due to the planetesimals which were left after the dispersal of the gas disk, in which the Solar System was formed. The idea is that the giant planets ejected the planetesimals and this resulted to a change of their orbits. Recent studies by Tsiganis et al. [50], Gomes et al. [9] and Morbidelli et al. [40] suggest that all outer planets started in a different configuration than the present one, with Jupiter slightly further from the Sun than its present distance, while the rest giant planets were in a distance less than 15 AU from the Sun. This is the so called *Nice model*, from the observatory of Nice where this group works. It is assumed that the planets were surrounded by a disk of planetesimals, which were ejected by the planets, and this resulted to a migration of their orbits. The assumption was made that Saturn was initially inside the 2/1 resonance with Jupiter, and as Saturn crossed this resonance, the eccentricity of the planets increased very much and the planets entered the outer planetesimal disk. This resulted to a heavy scattering of the planetesimals, which reached the inner Solar System and are responsible for the *Late Heavy Bombardment* on the surface of the Moon, which created its craters. More work is still to be done on this problem, including the effect of the giant planets on the orbits of the inner Solar System.

It has been realized recently that very small nonconservative forces, as the effect from mass loss of the sun, or the effects from theory of general relativity, must be included in the study of the past history or the long term evolution of the solar system, for billions of years. This is important for the study of the evolution of the inner planets and most notably of Mercury. Work has now started on this matter, and it is expected to give interesting results.

Bibliography

Primary Literature

1. Arnold VA, Avez A (1968) Ergodic Problems of Classical Mechanics. WA Benjamin, New York
2. Beaugé C, Michtchenko T (2003) Modelling the high-eccentricity planetary three-body problem. Application to the GJ876 planetary system. MNRAS 341:760
3. Beaugé C, Ferraz-Mello S, Michtchenko T (2003) Extrasolar Planets in Mean-Motion Resonance, Apses Alignment and Asymmetric Stationary Solutions. Ap J 593:1124
4. Beaugé C, Callegari N, Ferraz-Mello S, Michtchenko T (2005) Resonances and stability of extra-solar planetary systems. In: Knezevic Z, Milani A (eds) Dynamics of Populations of Planetary Systems. Cambridge Univ Press, p 3
5. Beaugé C, Ferraz-Mello S, Michtchenko TA (2006) Planetary Migration and Extrasolar Planets in the 2/1 Mean-motion Resonance. MNRAS 365:1160–1170
6. Celletti A, Kotoulas T, Voyatzis G, Hadjidemetriou JD (2007) The dynamical stability of a Kuiper belt-like region. MNRAS 378:1153–1164
7. Froeschlé C (1991) Modelling: An aim and a tool for the study of the chaotic behaviour of asteroidal and cometary orbits. In: Roy AE (ed) Predictability, Stability and Chaos in N-Body Dynamical Systems. Plenum Press, pp 125–155
8. Ferraz-Mello S, Beaugé C, Michtchenko T (2003) Evolution of migrating planet pairs in resonance. Cel Mech Dyn Astr 87:99–112
9. Gomes R, Levison HF, Tsiganis K, Morbidelli A (2005) Origin of the Cataclysmic Late Heavy Bombardment period of the terrestial planets. Nature 435:466
10. Gozdziewski K, Bois E, Maciejewski A (2002) Global dynamics of the Gliese 876 planetary system. MNRAS 332:839
11. Hadjidemetriou JD (1975) The continuation of periodic orbits from the restricted to the general three-body problem. Cel Mech 12:155–174
12. Hadjidemetriou JD (1976) Families of Periodic Planetary Type Orbits in the Three-Body Problem and their Stability. Astrophys Sp Sci 40:201–224
13. Hadjidemetriou JD (1982) On the Relation between Resonance and Instability in Planetary Systems. Cel Mech Dyn Astr 27:305–322
14. Hadjidemetriou JD (1992) The Elliptic Restricted Problem at the 3:1 Resonance. Celest Mech 53:151–183
15. Hadjidemetriou JD (1993) Asteroid Motion near the 3:1 Resonance. Cel Mec Dyn Astr 56:563–599
16. Hadjidemetriou JD (1998) Symplectic Maps and their use in Celestial Mechanics. In: Benest D, Froeschlé C (eds) Analysis and Modeling of Discrete Dynamical Systems, ch 9. Gordon and Breach Publ, pp 249–282
17. Hadjidemetriou JD (1999) A symplectic mapping model as a tool to understand the dynamics of the 2/1 resonant asteroid motion. Cel Mec Dyn Astr 73:65–76
18. Hadjidemetriou JD (2002) Resonant periodic motion and the stability of extrasolar planetary systems. Cel Mech Dyn Astr 83:141–154
19. Hadjidemetriou JD (2006) Symmetric and Asymmetric Librations in Extrasolar Planetary Systems: A global view. Cel Mech Dyn Astron 95:225–244
20. Hadjidemetriou JD (2006) Periodic orbits in gravitational systems. In: Steves BA et al (eds) Chaotic Worlds: From Order to Disorder in Gravitational N-Body Dynamical Systems. Springer, pp 43–79
21. Hadjidemetriou JD, Voyatzis G (2000) The 2/1 and 3/2 resonant asteroid motion: A symplectic mapping approach. Cel Mech Dyn Astron 78:137–150
22. Henrard J, Watanabe N, Moons M (1995) A bridge between Secondary and Secular Resonances inside the Hecuba Gap. Icarus 115:336–346
23. Israelian G, Santos N, Mayor M, Rebolo R (2001) Evidence for planet engulfment by the star HD82943. Nature 411:163
24. Jordan DW, Smith P (1988) Nonlinear Ordinary Differential Equations. Clarendon Press, Oxford

25. Laskar J (1988) Secular evolution of the Solar System over 10 million years. Astron Astrophys 198:341–362
26. Laskar J (1989) A numerical experiment on the chaotic behaviour of the Solar System. Nature 338:237–238
27. Laskar J (1994) Large scale chaos in the Solar System. Aston Astrophys 287:I9–12
28. Lee MH (2004) Diversity and Origin of 2:1 Orbital Resonance in Extrasolar Planetary Systems. Ap J 611:517
29. Lee MH, Peale S (2002) Dynamic and origin of the 2:1 orbital resonances of the GJ 876 planets. Ap J 567:596–609
30. Malhotra R (2002) A dynamical mechanism for establishing apsidal resonance. Ap J 575:L33–36
31. Marcy G, Butler P, Fischer D, Vogt S, Lissauer J, Rivera E (2001) Pair A of Resonant Planets Orbiting GJ 876. Ap J 556:296
32. Marcy GW, Butler RP, Fischer DA, Laughlin G, Vogt SS, Henry GW, Pourbaix D (2002) A planet at 5AU around 55Cnc. Ap J 581:1375–1388
33. Mayor M, Udry S, Naef D, Pepe F, Queloz D, Santos NC, Burnet M (2004) The CORALIE survey for southern extra-solar planets. XII Orbital solutions for 16 extra-solar planets discovered with CORALIE. A& A 415:291
34. Meyer KR, Hall GR (1992) Introduction to Hamiltonian Dynamical Systems and the N-Body Problem. Springer
35. Michtchenko TA, Ferraz-Mello S (1996) Comparative study of the asteroidal motion in the 3:2 and 2:1 resonances with Jupiter. I planar model. Astron Astrophys 310:1021–1035
36. Michtchenko TA, Beaugé C, Ferraz-Mello S (2006) Stationary Orbits in Resonant Extrasolar Planetary Systems. Cel Mech Dyn Astron 94:381–397
37. Morbidelli A (1996) On the Kirkwood Gap at the 2/1 Commensurability with Jupiter: numerical results. Astron J 111:2453–2461
38. Morbidelli A, Giorgilli A (1990) On the Dynamics in the Asteroids belt. Part I: General Theory. Celest Mech 47:145–172
39. Morbidelli A, Giorgilli A (1990) On the Dynamics in the Asteroids belt. Part II: Detailed Study of the Main Resnances. Celest Mech 47:173–1204
40. Morbidelli A, Levison HF, Tsiganis K, Gomes R (2005) Chaotic capture of Jupiter's Trojan Asteroids in the early Solar System. Nature 435:462
41. Murray CD, Dermott SF (1999) Solar System Dynamics. Cambridge University Press
42. Psychoyos D, Hadjidemetriou JD (2005) Dynamics of 2/1 resonant extrasolar systems. Application to HD82943 and Gliese 876. Cel Mech Dyn Astr 92:135–156
43. Rivera EJ, Lissauer JJ (2001) Dynamical models of the resonant pair of planets orbiting the star GJ 876. Ap J 558:392–402
44. Roy AE (1982) Orbital Motion, 2nd edn. Adam Hilder
45. Sandor ZS, Suli A, Erdi B, Pilat-Lohinger E, Dvorak R (2007) A stability catalogue of the habitable zones in extrasolar planetary systems. MNRAS 375:1495–1502
46. Szebehely V (1967) Theory of Orbits. Academic Press
47. Tsiganis K, Varvoglis H, Hadjidemetriou JD (2000) Stable chaos in the 12:7 mean motion resonance and its relation to the stickiness effect. Icarus 146:240–252
48. Tsiganis K, Varvoglis H, Hadjidemetriou JD (2002) Stable chaos in Higher-order Jovian Resonances. Icarus 155:454–474
49. Tsiganis K, Varvoglis H, Hadjidemetriou JD (2002) Stable chaos versus Kirkwood Gaps in the Asteroid Belt: A comparative Study of mean Motion Resonances. Icarus 159:284–299
50. Tsiganis K, Gomes R, Morbidelli K, Levison HF (2005) Origin of the orbital architecture of the Giant Planets of the Solar System. Nature 435:459
51. Voyatzis G, Hadjidemetriou HD (2005) Symmetric and asymmetric librations in planetary and satellite systems at the 2/1 resonance. Cel Mech Dyn Astr 93:263–294
52. Voyatzis G, Hadjidemetriou HD (2006) Symmetric and asymmetric 3:1 resonant periodic orbits. An application to the 55Cnc extra-solar system. Cel Mech Dyn Astr 95:259–271
53. Wisdom J (1982) The Origin of the Kirkwood gaps. Astron J 87:577–593
54. Wisdom J (1983) Chaotic behaviour and the Origin of the Kirkwood gaps. Icarus 56:51–74
55. Wisdom J (1985) A Perturbative Treatment of Motion Near the 3/1 Commensurability. Icarus 63:272–289
56. Wisdom J (1987) Urey Prize lecture. Chaotic Dynamics in the Solar System. Icarus 72:241–275
57. Wisdom J, Peale SJ, Mignard F (1984) The chaotic rotation of Hyperion. Icarus 58:137–152
58. Yakubovich VA, Starzhinskii VM (1975) Linear Differential Equations with Periodic Coefficients, vol 1,2. Halsted Press

Books and Reviews

Contopoulos G (2002) Order and Chaos in Dynamical Systems. Springer

Dvorak R, Freistetter F, Kurths J (eds) (2005) Chaos and Stability in Planetary Systems. Lecture Notes in Physics. Springer

Ferraz-Mello S (2007) Canonical Perturbation Theories. Degenerate Systems and resonance. Springer

Hagihara Y (1970–1972) Celestial mechanics vol 1, 2(I), 2(II). MIT Press, Cambridge

Hagihara Y (1974–1976) Celestial mechanics, vol 3(I), 3(II), 4(I), 4(II), 5(I), 5(II). Japan Society For the Promotion of Science, Tokyo

P

Partial Differential Equations that Lead to Solitons

Doğan Kaya
Department of Mathematics, Firat University, Elazig, Turkey

Article Outline

Definition of the Subject
Introduction
Some Nonlinear Models that Lead to Solitons
Future Directions
Bibliography

Definition of the Subject

In this part, we introduce the reader to a certain class of nonlinear partial differential equations which are characterized by solitary wave solutions of the classical nonlinear equations that lead to solitons. The classical nonlinear equations of interest show the existence of special types of traveling wave solutions which are either solitary waves or solitons. In this study, we will review a few solutions arising from the analytic work of the Korteweg–de Vries (KdV) equations, the generalized regularized long-wave RLW equation, Kadomtsev–Petviashvili (KP) equation, the Klein–Gordon (KG) equation, the Sine-Gordon (SG) equation, the Boussinesq equation, Pochhammer–Chree (PC) equation and the nonlinear Schrödinger (NLS) equation, the Fisher equation, Burgers equation, the Korteweg–de Vries Burgers' equation (KdVB), the two-dimensional Korteweg-deVries Burgers' (tdKdVB), the potential Kadomtsev–Petviashvili equation, the Kawahara equation, Generalized Zakharov-Kuznetsov (gZK) equation, the Sharma-Tasso-Olver equation, and the Cahn–Hilliard equation.

Introduction

Nonlinear phenomena play a crucial role in applied mathematics and physics. Calculating exact and numerical solutions, in particular, traveling wave solutions, of nonlinear PDEs in mathematical physics plays an important role in soliton theory. Moreover, these equations are mathematical models of complex physical occurrences that arise in engineering, chemistry, biology, mechanics, and physics.

In this work, we give a brief history of the above-mentioned nonlinear equations and how this type of equation has led to the soliton solutions; we then present an introduction to the theory of solitons. Soliton theory is an important branch of applied mathematics and mathematical physics. In the last decade this topic has become an active and productive area of research, and applications of the soliton equations in physical cases have been considered. These have important applications in fluid mechanics, nonlinear optics, ion plasma, classical and quantum fields' theories etc.

The best introduction to the soliton is that contained in J. Scott Russell's (a Scottish naval engineer) seminal 1844 report to the Royal Society. Scott Russell's report titled "Report on Waves" [62] was presented to the Royal Society in the 18th Century and he wrote:

> "I was observing the motion of a boat which was rapidly drawn along a narrow channel by a pair of horses, when the boat suddenly stopped—not so the mass of water in the channel which it had put in motion; it accumulated round the prow of the vessel in a state of violent agitation, then suddenly leaving it behind, rolled forward with great velocity, assuming the form of a large solitary elevation, a rounded smooth and well-defined heap of water, which continued its course along the channel apparently without change of form or diminution of speed. I followed it on horseback, and overtook it still rolling on at a rate of some eight or nine miles an hour, preserving its original figure some thirty feet long and a foot to a foot and a half in height. Its height grad-

ually diminished, and after a chase of one or two miles I lost it in the windings of the channel . . . ”

Scott Russell’s experimental observations in 1834 were followed by the theoretical scientific work of Lord Rayleigh and Joseph Boussinesq around 1870 [60] and finally, two Dutchmen, Korteweg and de Vries, developed a nonlinear partial differential equation to model the propagation of shallow water waves applicable to the situation in 1895 [47]. This work was really what Scott Russell fortuitously witnessed. This famous classical equation is known simply as the KdV equation. Korteweg and de Vries published a theory of shallow water waves which reduced Russell’s observations to its essential features. The nonlinear classical dispersive equation was formulated by Korteweg and de Vries in the form

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} + \varepsilon\frac{\partial^3 u}{\partial x^3} + \gamma\, u\frac{\partial u}{\partial x} = 0$$

where $c,\ \varepsilon,\ \gamma$ are physical parameters. This equation plays a key role in soliton theory.

In the late 1960s, Zabusky and Kruskal numerically studied and then analytically solved the KdV equation [79]. They came to the result that stable pulse-like waves could exist in a problem described by the KdV equation from their numerical study. A remarkable aspect is that the discovered solitary waves retain their shapes and speeds after collision. Zabusky and Kruskal called these waves solitons as they resemble particles in nature. In 1967, this numerical development was placed on a settled mathematical basis with Gardner et al.’s discovery of the inverse-scattering-transform method [20].

The soliton concept is related with solutions for nonlinear partial differential equations. The soliton solution of a nonlinear equation usually is used a single wave. If there are several soliton solutions, these solutions are called solitons. On the other hand, if a soliton is separated infinitely from another soliton, this soliton is called a single wave. Besides, a single wave solution can’t be sec h^2 function for equations other than nonlinear equations, such as the KdV equation. But this solution can be *sech* or $\tan^{-1}(e^{\alpha x})$.

At this stage, we could ask what the definition of the soliton solutions is? It is not easy to define the soliton concept. Wazwaz [69] describes solitons as solutions of nonlinear differential equations as follows:

(i) A long and shallow water wave should not lose its permanent form;
(ii) A long and shallow water wave of the solution is localized, meaning either the solutions decay exponentially to zero such as in the solitons admitted by the KdV equation, or approach a constant at infinity such as solitons provide by the SG equation;
(iii) A long and shallow water wave of the solution can interact with other solitons preserving its character.

There is also a more formal definition of this concept, but these definitions require substantial mathematics [17]. On the other hand, the phenomena of the solitons doesn’t always follow these three properties. For example, there is the concept of “light bullets” in the subject of nonlinear optics which are often called solitons despite losing energy during interaction. This idea can be found at an internet web page of the Simon Fraser University British Columbia, Canada [51].

Some Nonlinear Models that Lead to Solitons

In this section, we will deal with the fundamental ideas and some nonlinear partial differential equations that lead to solitons. These equations are the result of a huge amount of research work and physical implementations which focus on the most diverse and active areas of applied mathematics and mathematical physics.

Example 1 In various fields of science and engineering, nonlinear evaluation equations, as well as their analytic and numerical solutions, are of fundamental importance. One of the most attractive and surprising wave phenomena is the creation of solitary waves or solitons. It was approximately two centuries ago, that an adequate theory for solitary waves was developed, in the form of a modified wave equation known as the KdV equation [14,15,16,47,69,75]. The third-order KdV equation is a classical nonlinear partial differential equation originally formulated to model shallow water flow [47].

Herein, we are particularly interested in the original third-order KdV equation [47] given by

$$u_t + u\,u_x + \alpha\,u_{xxx} = 0\,,$$

respectively, and the generalized form

$$u_t + u^p u_x + \alpha\,u_{xxx} = 0\,,$$

where p models the dispersion and subscripts in t and x denote partial derivatives with respect to these independent variables. The study of long waves, tides, solitary waves, and related phenomena, leads to an equation referred to as the generalized KdV equation. If $p = 0$, $p = 1$, and $p = 2$, this equation becomes the linearized KdV, nonlinear KdV, and modified KdV equations, respectively [14,15,16,69,75]. The modified KdV equation has

also found applications in many areas of physical situations, for instance the equation describes nonlinear acoustic waves in an anharmonic lattice [78].

Example 2 Consider the solution $u(x, t)$ of the generalized RLW equation

$$u_t + u_x + \alpha\, u_x^{\nu} - \beta\, u_{xxt} = 0\,,$$

where ν is a positive integer, and α and β are positive constants. The generalized RLW equation was first put forward as a model for small-amplitude long waves on the surface of water in a channel by Peregrine [37,57] and later by Benjamin et al. [2]. In physical situations such as unidirectional waves propagating in a water channel, long-crested waves in near-shore zones, and many others, the generalized RLW equation serves as an alternative model to the KdV equation [4,5].

Example 3 The nonlinear KG equations in one-dimensional space

$$u_{tt} - u_{xx} + \frac{\mathrm{d}V}{\mathrm{d}u} = 0\,,$$

where $V = V(u)$ is a general nonlinear function of u, but not its derivatives. This equation is the most natural nonlinear generalization of the wave equation and first arose in a mathematical context, with $V(u) = \exp(u)$, in the theory of surfaces of constant curvature. In addition, this equation appears in many different fields of application. For instance, a polynomial nonlinearity can be used as a model field theory, while a $\cos u$ term yields the sine-Gordon equation and so on [14,15,16,69,75]. We will consider a particular case of equation KG, the so-called sine-Gordon nonlinear hyperbolic equation, which has the form

$$u_{tt} - cu_{xx} + \kappa \sin(u) = 0\,; \quad -\infty < x < \infty\,, \quad t > 0\,,$$

where c and κ are constants. The sine-Gordon equation is firstly used in the work of differential geometry and for investigation of the propagation of a 'slip' dislocation in crystals [15].

The generalized one-dimensional KG equation

$$u_{tt} - ku_{xx} + b_1 u + b_2 u^{m+1} + b_3 u^{2m+1} = 0\,,$$

is given in [80]. The KG equation represents a nonlinear model of longitudinal wave propagation of elastic rods when $m = 1$ [12].

Example 4 In this example, we consider the generalized Boussinesq-type equation [7]

$$u_{tt} - u_{xx} + \delta u_{xxxx} = -(\psi(u))_{xx}\,; \qquad -\infty < x < \infty\,, \quad t > 0\,,$$

where $\delta \geq 0$ is constant, $\psi(u) = |u|^{\alpha-1}\, u$ and $\alpha > 1$. This equation represents a generalization of the classical Boussinesq equation which arises in the modeling of nonlinear strings. The Boussinesq equation describes in the continuous limit the propagation of waves in a one-dimensional nonlinear lattice and the propagation of waves in shallow water [6,7,15,77]. A proof of the local well-posedness of the Boussinesq equation has been showed by Bona and Sachs [6]. In [13] the authors noticed that the Boussinesq equation admits solitary solutions and the existence of solitary wave solutions illustrates the perfect balance between dispersion and the nonlinearity of the Boussinesq equation. For the α integer, solitary-wave solutions have been shown to be stable under some restrictions on the wave speed by Bona and Sachs [6].

Scott Russell's study [62] of solitary water waves motivated the development of nonlinear partial differential equations for the modeling of wave phenomena in fluids, plasmas, elastic bodies, etc. The Boussinesq equation is an important model that approximately describes the propagation of long waves on shallow water like the other Boussinesq equations (with u_{xxtt}, instead of u_{xxxx}). This equation was first deduced by Boussinesq [7]. In the case $\delta > 0$ this equation is linearly stable and governs small nonlinear transverse oscillations of an elastic beam [15]. It is called the "good" Boussinesq equation, while the equation with $\delta < 0$ received the name "bad" Boussinesq equation since it possesses linear instability.

Example 5 Consider the PC equation

$$u_{tt} - \frac{1}{p}\, u_{xx}^{\ell} - u_{xx} - u_{ttxx} = 0\,,$$

where ℓ indicates different material with different values of it. The PC equation is applied as a nonlinear model of longitudinal wave propagation of elastic rods [15,16,75]. In the work of Bogolubsky [75], the author obtained exact solitary wave solutions to the PC equation $\ell = 2, 3, 5$, respectively [15].

Example 6 In this paper, we consider the Burgers equation

$$u_t + \varepsilon u u_x - \nu u_{xx} = 0$$

where ε and ν are parameters and the subscripts t and x denote differentiation. The Burgers' equation is a model of flow through a shock wave in a viscous fluid [13] and in the Burgers' model of turbulence [9].

Example 7 The Fisher equation

$$u_t - \nu u_{xx} = ku\left(1 - \frac{u}{\kappa}\right)\,,$$

is well known in population dynamics, where $\nu > 0$ is the diffusion constant, $k > 0$ is the linear growth rate, and $\kappa > 0$ is the carrying capacity of the environment. The right-hand side function $ku\left(1 - \frac{u}{\kappa}\right)$ represents a nonlinear growth rate [15]. This well-known equation was first proposed by Fisher [18] for a model of the advancement of a mutant gene in an infinite one-dimensional habitat. In recent years, the equation has been used as a basis for a wide variety of models for the spatial spread of genes in a population and for chemical wave propagation [18].

Example 8 Let us consider the $(1+1)$-dimensional NLS equation with two higher-order nonlinear terms in the form

$$iu_t + \tfrac{1}{2}u_{xx} + \alpha\,|u|^{\sigma}\,u + \beta\,|u|^{2\sigma}\,u = 0\,,$$

where σ is a positive constant, α and β are constant parameters [15,16,69]. $u = u(x,t)$ is a complex function that represents the complex amplitude of the wave form, the variable t should be interpreted as the normalized propagation distance, x the normalized transverse coordinate that represents a retarded time. The NLS equation, especially that with lower σ values, appears in various branches of contemporary physics [18,27,61,67,80,81] and has been extensively investigated for the cases of $\sigma = 1$ and $\sigma = 2$, its various solutions have been obtained. The NLS equation also arises in some form of $\alpha = 0$ and $\sigma = 1$ in some other physical systems such as nonlinear optic [3,26,38], hydro magnetic and plasma waves [30,63] and the propagation of solitary waves in piezoelectric semiconductors [55].

Example 9 In applied mathematics, the KP equation [36] is a nonlinear partial differential equation. It is also sometimes called the Kadomtsev–Petviashvili-Boussinesq equation. The KP equation is usually written as:

$$(u_t - 6uu_{xx} + u_{xxx})_x + 3\lambda^2 u_{yy} = 0\,,$$

where $\lambda = \pm 1$. This equation is implemented to describe slowly varying nonlinear waves in a dispersive medium [58]. The above written form shows that the KP equation is a generalization form of the KdV equation which is like the KdV equation; the KP equation is completely integrable. The KP equation was first discovered in 1970 by Kadomtsev and Petviashvili when they relaxed the restriction that the waves be strictly one-dimensional.

Example 10 Consider the solution $u(x,t)$ of the nonlinear partial differential equation

$$u_t + \varepsilon u u_x - \nu u_{xx} + \mu u_{xxx} = 0$$

where ε, υ and μ are positive parameters. This equation is called the Korteweg–de Vries Burgers' equation (KdVB) which is derived by Su and Gardner [65]. KdVB is a model equation for a wide class of nonlinear systems in the weak nonlinearity and long wavelength approximations because it contains both damping and dispersion. KdVB has been constructed when including electron inertia effects in the description of weak nonlinear plasma waves. The KdVB equation possesses a steady-state solution which has been demonstrated to model weak plasma shocks propagating perpendicular to a magnetic field [24]. There are some other implementation places to use the KdVB equation. Examples are its use in the study of wave propagation through a liquid-filled elastic tube [34] and for a description of shallow water waves on a viscous fluid [21,35].

Example 11 The nonlinear partial differential equation

$$(u_t + uu_x - qu_{xx} + pu_{xxx})_x + ru_{yy} = 0\,,$$

where p, q, r are real parameters. This equation is called the two-dimensional Korteweg–deVries Burgers' (tdKdVB) equation. The tdKdVB equation is a model equation for a wide class of nonlinear wave models of fluids in an elastic tube, liquids with small bubbles and turbulence [52,53].

Example 12 In this example we consider in the $(2+1)$-dimension the potential Kadomtsev–Petviashvili equation,

$$u_{xt} + \tfrac{3}{2}u_x u_{xx} + \tfrac{1}{4}u_{xxxx} + \tfrac{3}{4}u_{yy} = 0\,,$$

where the initial conditions $u(x,0,t)$ and $u_y(x,0,t)$ are given. Nonlinear phenomena play a crucial role in applied mathematics and physics. Obtaining the exact or approximate solutions of PDEs in physics and mathematics is important and it is still a hot spot to seek new methods to obtain new exact or approximate solutions [14,15,16]. Different methods have been put forward to seek various exact solutions of multifarious physical models described by nonlinear PDEs.

Example 13 We consider the numerical solution to a problem involving a nonlinear partial differential equation of the form

$$u_t + u\,u_x + u_{xxx} - u_{xxxxx} = 0\,,$$

this is called the Kawahara equation. The Kawahara equation occurs in the theory of magneto-acoustic waves in plasmas [39] and in the theory of shallow water waves with surface tension [29].

Example 14 A Generalized Zakharov–Kuznetsov (gZK) equation [14,15,16,48,75] is proposed to understand the physical and scientific mechanisms in different physical and engineering problems. The gZK equation is as follows

$$u_t + b_1 u^\rho u_x + b_2 u^{2\rho} u_x + b_3 u_{xy} + b_4 u_{xxx} + b_5 u_{xyy} = 0 ,$$

where b_1, b_2, b_3, $b_4 \neq 0$, $b_5 \neq 0$ and $\rho \neq 0$ are constant parameters [48]. When $b_1 = 6, b_2 = b_3 = 0, b_4 = b_5 = 1$ and $\rho = 1$ the equation gZK is said to be Zakharov–Kuznetsov (ZK) which can be derived for Alfvën waves in a magnetized plasma at a special, critical angle to the magnetic field by means of an asymptotic multi-scale technique [8,56]. If $b_1 = 0$, $b_2 = 1$, $b_3 = 0$, $b_4 = b_5 = 1$ and $\rho = 1$, then the equation gZK is said to be modified Zakharov–Kuznetsov (mZK) which represents an anisotropic two-dimensional generalization of the KdV equation and can be derived in a magnetized plasma for a small amplitude Alfvën wave at a critical angle to the undisturbed magnetic field. The radially symmetric positive solutions for the mZK equation have been computed [54,82].

Example 15 Let's consider the Sharma–Tasso–Olver equation (STO) with its fission and fusion [76]

$$u_t + 3\alpha u_x^2 + 3\alpha u^2 u_x + 3\alpha u u_{xx} + \alpha u_{xxx} = 0 .$$

Attention has been focused on the STO equation in [50,68] and the references therein due to its appearance in scientific applications [74].

Example 16 Consider the Cahn–Hilliard equation

$$u_t + u_{xxxx} = \left(u^3 - u\right)_{xx} + \beta\, u_x .$$

The Cahn–Hilliard equation is related to a number of interesting physical phenomena like spinodal decomposition, phase separation and phase-ordering dynamics. On the other hand this equation is very hard and difficult to solve. In this paper by considering a modified extended tanh method, we found some exact solutions of the Cahn–Hilliard equation. This equation is very crucial in materials science [10,11,25]. Many articles have focused on mathematical and numerical studies of this equation [1,19,49].

Future Directions

Nonlinear phenomena play a crucial role in applied mathematics and physics. Furthermore, when an original nonlinear equation is directly calculated, the solution will preserve the actual physical characters of solutions. Explicit solutions to the nonlinear equations are of fundamental importance. Various effective methods have been developed to understand the mechanisms of these physical models, to help physicians and engineers and to ensure knowledge for physical problems and its applications.

Many explicit exact methods have been introduced in the literature [14,15,16,31,46,69,70,71,72,73,75]. These include the Bäcklund transformation, Hopf–Cole transformation, Generalized Miura Transformation, the Inverse scattering method, Darboux transformation, Painleve method, homogeneous balance method, similarity reduction method, tanh method, Exp-function method, sine–cosine method and so on. There are also many numerical methods implemented for these equations [22,23,28,32,33,40,41,42,43,44,45,59,64,66]. These include the Finite elements method, finite difference methods and some approximate methods such as the Adomian decomposition method, Homotopy perturbation method, Variational perturbation method, Sinc-Galerkin method and so on.

There is still much work to be done by researchers in this field involving applications of nonlinear equations and exact and numerical implementations.

Bibliography

Primary Literature

1. Barrett JW, Blowey JF (1999) Finite element approximation of the Cahn–Hilliard equation with concentration dependent mobility. Math Comput 68:487–517
2. Benjamin TB, Bona JL, Mahony JJ (1972) Model Equations for Waves in Nonlinear Dispersive Systems. Phil Trans Royal Soc London 227:47–78
3. Bespalov VI, Talanov VI (1966) Filamentary structure of light beams in nonlinear liquids. JETP Lett 3:307–310
4. Bona JL, Pritchard WG, Scott LR (1981) An Evaluation for Water Waves. Phil Trans Royal Soc London A 302:457–510
5. Bona JL, Pritchard WG, Scott LR (1983) A Comparison of Solutions of two Model Equations for Long Waves. In: Lebovitz NR (ed), Fluid Dynamics in Astrophysics and Geophysics. Lectures in Applied Mathematics. Am Math Soc 20:235–267
6. Bona JL, Sachs RL (1988) Global Existence of Smooth Solutions and Stability Theory of Solitary Waves for a Generalized Boussinesq Equation. Commun Math Phys 118:15–29
7. Boussinesq J (1871) Thèorie de I'ntumescence Liquid Appelèe Onde Solitaire ou de Translation, se Propageant dans un Canal Rectangulaire. Comptes Rendus Acad Sci (Paris) 72:755–759
8. Bridges TJ, Reich S (2001) Multi-symplectic spectral discretizations for the Zakharov-Kuznetsov and Shallow water equations. Physica D 152:491–504
9. Burgers J (1948) A mathematical model illustrating the theory of turbulence, Advances in Applied Mechanics. Academic Press, New York, pp 171–199
10. Chan JW (1961) On spinodal decomposition. Acta Metall 9:795
11. Chan JW, Hilliard JE (1958) Free energy of a nonuniform system I. Interfacial free energy. J Chem Phys 28:258–267
12. Clarkson PA, LeVeque RJ, Saxton R (1986) Solitary Wave Interactions in Elastic Rods. Stud Appl Math 75:95–122

13. Cole JD (1951) On a quasilinear parabolic equation occurring in aerodynamic. Quart Appl Math 9:225–236
14. Debtnath L (1983) Nonlinear Waves. Cambrige University Press, Cambrige
15. Debtnath L (1997) Nonlinear Partial Differential Equations for Scientist and Engineers. Birkhauser, Boston
16. Drazin PG, Johnson RS (1989) Solutions: An Introduction. Cambridge University Press, Cambridge
17. Edmundson DE, Enns RH (1992) Bistable light bullets. Opt Lett 17:586
18. Fisher RA (1937) The wave of advance of advantageous genes. Ann Eugenics 7:353–369
19. Garcke H (2000) Habilitation Thesis, Bonn University, Bonn
20. Gardner CS, Greene JM, Kruskal MD, Miura RM (1967) Method for solving the Korteweg–de Vries equation. Phys Rev Lett 19:1095–1097
21. Gardner CS, Greene JM, Kruskal MD, Miura RM (1974) Korteweg–de Vries equation and generalizations. IV. Method for exact solution. Commun Pure Appl Math XXVII:97–133
22. Geyikli T, Kaya D (2005) An application for a Modified KdV equation by the decomposition method and finite element method. Appl Math Comp 169:971–981
23. Geyikli T, Kaya D (2005) Comparison of the solutions obtained by B-spline FEM and ADM of KdV equation. Appl Math Comp 169:146–156
24. Grad H, Hu PN (1967) Unified shock profile in plasma. Phys Fluids 10:2596–2601
25. Gurtin M (1996) Generalized Ginzburg-Landau and Cahn–Hilliard equations based on a microforce balance. Physica D 92:178–192
26. Hasegawa A, Tappert F (1973) Transmission of stationary nonlinear optical pulse in dispersive dielectric fibers, I: Anomalous dispersion. Appl Phys Lett 23:142–144
27. Hayata K, Koshiba M (1995) Algebraic solitary-wave solutions of a nonlinear Schrödinger equation. Phys Rev E 51:1499
28. Helal MA, Mehanna MS (2007) A comparative study between two different methods for solving the general Korteweg–de Vries equation. Chaos Solitons Fractals 33:725–739
29. Hunter JK, Scheurle J (1988) Existence of perturbed solitary wave solutions to a model equation for water-waves. Physica D 32:253–268
30. Ichikawa VH (1979) Topic on solitons in plasma. Physica Scripta 20:296–305
31. Inan IE, Kaya D (2006) Some Exact Solutions to the Potential Kadomtsev–Petviashvili Equation. Phys Lett A 355:314–318
32. Inan IE, Kaya D (2006) Some exact solutions to the potential Kadomtsev–Petviashvili equation. Phys Lett A 355:314–318
33. Inan IE, Kaya D (2007) Exact solutions of the some nonlinear partial differential equations. Physica A 381:104–115
34. Jonson RS (1970) A nonlinear equation incorporating damping and dispersion. J Phys Mech 42:49–60
35. Jonson RS (1972) Shallow water waves in a viscous fluid-the undular bore. Phys Fluids 15:1693–1699
36. Kadomtsev BB, Petviashvili VI (1970) On the Stability of Solitary Waves in Weakly Dispersive Media. Sov Phys Dokl 15:539–541
37. Kakutani T, Ona H (1969) Weak nonlinear hydomagnetic waves in a cod collision – free plasma. J Phys Soc Japan 26:1305–1319
38. Karpman VI, Krushkal EM (1969) Modulated waves in nonlinear dispersive media. Sov Phys JETP 28:277–281
39. Kawahara TJ (1972) Oscillatory Solitary Waves in Dispersive Media. Phys Soc Japan 33:260
40. Kaya D (2003) A Numerical Solution of the Sine-Gordon Equation Using the Modified Decomposition Method. Appl Math Comp 143:309–317
41. Kaya D (2006) The exact and numerical solitary-wave solutions for generalized modified boussinesq equation. Phys Lett A 348:244–250
42. Kaya D, Al-Khaled K (2007) A numerical comparison of a Kawahara equation. Phys Lett A 363:433–439
43. Kaya D, El-Sayed SM (2003) An Application of the Decomposition Method for the Generalized KdV and RLW Equations. Chaos Solitons Fractals 17:869–877
44. Kaya D, El-Sayed SM (2003) Numerical soliton-like solutions of the potential Kadomstev–Petviashvili equation by the decomposition method. Phys Lett A 320:192–199
45. Kaya D, El-Sayed SM (2003) On a Generalized Fifth Order KdV Equations. Phys Lett A 310:44–51
46. Khater AH, El-Kalaawy OH, Helal MA (1997) Two new classes of exact solutions for the KdV equation via Bäcklund transformations. Chaos Solitons Fractals 8:1901–1909
47. Korteweg DJ, de Vries H (1895) On the change of form of long waves advancing in a rectangular canal and on a new type of long stationary waves. Philosophical Magazine 39:422–443
48. Li B, Chen Y, Zhang H (2003) Exact travelling wave solutions for a generalized Zakharov-Kuznetsov equation. Appl Math Comput 146:653–666
49. Li D, Zhong C (1998) Global attractor for the Cahn–Hilliard system with fast growing nonlinearity. J Differ Equ 149(2):191
50. Lian Z, Lou SY (2005) Symmetries and exact solutions of the Sharma–Tasso–Olver equation. Nonlinear Anal 63:1167–1177
51. Edmundson D, Enns R (1996) Light Bullet Home Page. http://www.sfu.ca/~renns/lbullets.html
52. Ma WX (1993) An exact solution to two-dimensional Korteweg-deVries-Burgers equation. J Phys A 26:17–20
53. Parkas EJ (1994) Exact solutions to the two-dimensional Korteweg-deVries-Burgers equation. J Phys A 27:497–501
54. Parkes J, Munro S (1999) The derivation of a modified Zakharov-Kuznetsov equation and the stability of its solutions. J Plasma Phys 62:305–317
55. Pawlik M, Rowlands G (1975) The propagation of solitary waves in piezoelectric semiconductors. J Phys C 8:1189–1204
56. Pelinovsky DE, Grimshaw RHJ (1996) An asymptotic approach to solitary wave instability and critical collapse in long-wave KdV-type evolution equations. Physica D 98:139–155
57. Peregrine DH (1967) Long Waves on a Beach. J Fluid Mech 27:815–827
58. Peregrine DH (1996) Calculations of the Development of an Undular Bore. J Fluid Mech 25:321–330
59. Polat N, Kaya D, Tutalar HI (2006) A analytic and numerical solution to a modified Kawahara equation and a convergence analysis of the method. Appl Math Comp 179:466–472
60. Rayleigh L (1876) On Waves. The London and Edinburgh and Dublin Philosophical Magazine 5:257
61. Robert WM, Vsvolod VA, Yuri SK, John DL (1996) Optical solitons with power-law asymptotics. Phys Rev E 54:2936
62. Russell JS (1844) Report on Waves. 14th meeting of the British Association for the Advancement of Science. BAAS, London
63. Schimizu K, Ichikawa VH (1972) Auto modulation of ion oscillation modes in plasma. J Phys Soc Japan 33:789–792
64. Shawagfeh N, Kaya D (2004) Series solution to the Pochhammer–Chree equation and comparison with exact solutions. Comp Math Appl 47:1915–1920

65. Su CH, Gardner CS (1969) Derivation of the Korteweg–de Vries and Burgers' equation. J Math Phys 10:536–539
66. Ugurlu Y, Kaya D, Solution of the Cahn–Hilliard equation. Comput Math Appl (accepted for publication)
67. Wang M, Li LX, Zhang J (2007) Various exact solutions of nonlinear Schrödinger equation with two nonlinear terms. Chaos Solitons Fract 31:594–601
68. Wang S, Tang X, Lou SY (2004) Soliton fission and fusion: Burgers equation and Sharma–Tasso–Olver equation. Chaos Solitons Fractals 21:231–239
69. Wazwaz AM (2002) Partial Differential Equations: Methods and Applications. Balkema, Rotterdam
70. Wazwaz AM (2007) Analytic study for fifth-order KdV-type equations with arbitrary power nonlinearities. Comm Nonlinear Sci Num Sim 12:904–909
71. Wazwaz AM (2007) A variable separated ODE method for solving the triple sine-Gordon and the triple sinh-Gordon equations. Chaos Solitons Fractals 33:703–710
72. Wazwaz AM (2007) The extended tanh method for abundant solitary wave solutions of nonlinear wave equations. Appl Math Comp 187:1131–1142
73. Wazwaz AM, Helal MA (2004) Variants of the generalized fifth-order KdV equation with compact and noncompact structures. Chaos Solitons Fractals 21:579–589
74. Wazwaz AM (2007) New solitons and kinks solutions to the Sharma–Tasso–Olver equation. Appl Math Comp 188:1205–1213
75. Whitham GB (1974) Linear and Nonlinear Waves. Wiley, New York
76. Yan Z (2003) Integrability for two types of the $(2+1)$-dimensional generalized Sharma–Tasso–Olver integro-differential equations. MM Res 22:302–324
77. Zabusky NJ (1967) Nonlinear Partial Differential Equations. Academic Press, New York
78. Zabusky NJ (1967) A synergetic approach to problems of nonlinear dispersive wave propagations and interaction. In: Ames WF (ed) Proc. Symp. on Nonlinear Partial Differential equations. Academic Press, Boston, pp 223–258
79. Zabusky NJ, Kruskal MD (1965) Interactions of solitons in a collisionless plasma and the recurrence of initial states. Phys Rev Lett 15:240–243
80. Zhang W, Chang Q, Fan E (2003) Methods of judging shape of solitary wave and solution formulae for some evolution equations with nonlinear terms of high order. J Math Anal Appl 287:1–18
81. Zhou CT, He XT, Chen SG (1992) Basic dynamic properties of the high-order nonlinear Schrödinger equation. Phys Rev A 46:2277
82. Munro S, Parker EJ (1997) The stability of solitary-wave Solutions to a modified Zakharov–Kuznetsov equation. J Plasma Phys 64:411–426

Books and Reviews

The following, referenced by the end of the paper, is intended to give some useful for further reading.

For another obtaining of the KdV equation for water waves, see Kevorkian and Cole (1981); one can see the work of the Johnson (1972) for a different water-wave application with variable depth, for waves on arbitrary shears in the work of Freeman and Johnson (1970) and Johnson (1980) for a review of one and two-dimensional KdV equations. In addition to these; one can see the book of Drazin and Johnson (1989) for some numerical solutions of nonlinear evolution equations. In the work of the Zabusky, Kruskal and Deam (F1965) and Eilbeck (F1981), one can see the motion pictures of soliton interactions. See a comparison of the KdV equation with water wave experiments in Hammack and Segur (1974)

For further reading of the classical exact solutions of the nonlinear equations can be seen in the works: the Lax approach is described in Lax (1968); Calogero and Degasperis (1982, A.20), the Hirota's bilinear approach is developed in Matsuno (1984), the Bäckland transformations are described in Rogers and Shadwick (1982); Lamb (1980, Chap. 8), the Painleve properties is discussed by Ablowitz and Segur (1981, Sect. 3.8), In the book of Dodd, Eilbeck, Gibbon and Morris (1982, Chap. 10) can found review of the many numerical methods to solve nonlinear evolution equations and shown many of their solutions.

Patterns and Interfaces in Dissipative Dynamics

L.M. Pismen
Department of Chemical Engineering and Minerva Center for Nonlinear Physics of Complex Systems, Technion – Israel Institute of Technology, Haifa, Israel

Article Outline

Glossary
Definition of the Subject
Introduction
Stationary Patterns
Moving Interfaces
Wave Patterns
Future Directions
Bibliography

Glossary

Interface An interface separates domains where different stationary states or different patterns prevail. In the latter case, it is also called a domain wall. The interface typically has a finite thickness comparable to a characteristic intrinsic scale of the system but small compared to the overall system size.

Stationary pattern A stationary pattern is formed as a result of an instability to perturbations with a finite wavenumber. It may have any of various spatial structures (striped, square, hexagonal, or quasicrystalline in 2D, lamellar, crystalline or quasicrystalline in 3D) and may slowly evolve in time.

Wave pattern A wave pattern is formed by a combination of waves propagating in one or different directions.

Definition of the Subject

A *pattern* is an inhomogeneous state of a physical system that arises spontaneously under spatially homogeneous conditions. Spontaneous pattern formation has been first observed by Faraday [1] in vibrated liquid layers and Bénard [2] in fluids heated from below. Turing [3] envisaged pattern formation as the mechanism of morphogenesis in living Nature. Some patterns can be described as a collection of patches or domains where one of alternative homogeneous states prevails, separated by relatively narrow *interfaces*. In their turn, moving interfaces may develop corrugation patterns. Patterns can be stationary or wavelike; they can be regular, interlaced by defects, or chaotic (turbulent). In the latter part of 20th century, numerous pattern formation phenomena have been observed in chemistry, biology, fluid mechanics, granular media, nonlinear optics, and other applications, and common models describing these phenomena in physically dissimilar settings have been formulated and studied. Understanding pattern formation is important both for describing natural self-organization phenomena and for developing manufacturing processes based on self-organization.

Introduction

A typical setup of a non-equilibrium system that may undergo a symmetry-breaking transition is shown in Fig. 1. A non-equilibrium stationary state homogeneous in the "horizontal" plane is sustained by fluxes in the normal ("vertical") direction, along which an inhomogeneous "vertical structure" may be formed. This setup may be realized as a layer of fluid or granular matter; a chemically reacting system, such as an active layer or a catalytic surface; an area where different populations spread out and compete; a propagating interphase boundary, e.g a melting or crystallizing solid; a slice of nonlinear optical medium, etc. Under certain conditions, most commonly, under increased driving, this homogeneous state may be destabilized, giving way to a stationary or moving pattern with a characteristic wavelength dependent on physical properties of the system as well as on external fluxes. In chemically reacting systems, three-dimensional patterns can be also formed when a sufficient amount of reactants is stored; such patterns may exist, of course, for a limited time until the original cache is depleted. Mathematically, a pattern typically emerges as an inhomogeneous solution of a (system of) partial differential equation(s) with space-independent coefficients in the absence of lateral fluxes.

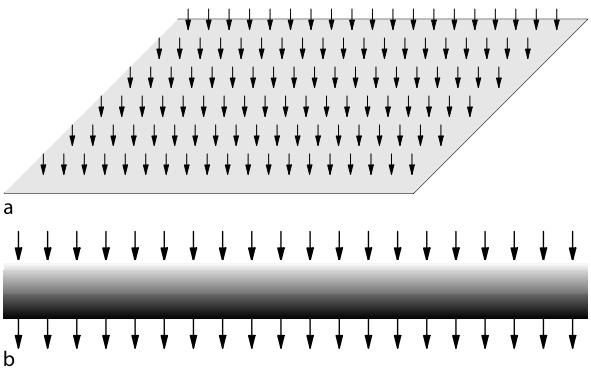

Patterns and Interfaces in Dissipative Dynamics, Figure 1
An open system isotropic in two dimensions. A truly two-dimensional system (*above*) and a cut through a system with vertical structure (*below*, shown symbolically by *varied shading*). *Arrows* indicate the direction of external fluxes

Alternative states, corresponding to different phases, may exist also in equilibrium systems. Following a fast quench past a critical point, different states, separated by domain boundaries, would be approached at spatially removed locations. Typically, these domains would consequently slowly coarsen to minimize the extent of an interphase boundary and related energetic costs. A stationary pattern with a finite wavelength may exist, however, also at equilibrium, provided it minimizes the free energy of the system. Such patterns are realized as "mesoscopic crystals" in block-copolymers consisting of two kinds of mutually repelling units [4].

In fluid mechanics, inhomogeneous states, most often disordered but still retaining a measure of regularity, are commonplace, as anybody observing wavy sea and cloud patterns could have realized long before classical 19th century experiments of Faraday and Bénard. Wave patterns generated by oscillatory chemical reactions (which long considered to be impossible due to thermodynamic misconceptions) were demonstrated in 1960s [5], while controlled experiments demonstrating persistent stationary chemical patterns in reaction-diffusion systems had to wait till early 1990s [6]. Shell growth patterns [7], striped and dotted animal skins [8], and desert vegetation patterns [9] have been always here for anybody to observe, before finding rational explanation in terms of the same nonlinear models. Corrugated interfaces were observed and described both as flame fronts in combustion theory [10] and as dendrite forms of growing crystals [11]. More recently, much attention has been drawn by nonlinear optical patterns – spontaneous images emerging in optical circuits and lasers [12].

Stationary Patterns

Symmetry-Breaking Transitions

The most direct way to formation of stationary patterns is a symmetry-breaking bifurcation. It can be demonstrated in a straightforward way taking as an example a two-component reaction-diffusion system (RDS)

$$\partial_t u = D_1 \nabla^2 u + \gamma_1^{-1} f(u, v) \,, \tag{1}$$

$$\partial_t v = D_2 \nabla^2 v + \gamma_2^{-1} g(u, v) \,, \tag{2}$$

where $f(u, v)$, $g(u, v)$ are source functions depending on the variables u and v, D_1, D_2 are diffusivities, and ∇^2 is the Laplace operator. We suppose that the system has a homogeneous stationary state (HSS) $u = u_s$, $v = v_s$ satisfying $f(u_s, v_s) = g(u_s, v_s) = 0$; the factors γ_1, γ_2 are introduced to scale the derivatives f_u, g_v computed at this HSS to unity. Stability analysis of the chosen HSS to infinitesimal perturbations $\widetilde{u}, \widetilde{v} \propto \exp(\mathrm{i}\, \boldsymbol{k} \cdot \boldsymbol{x})$ with a wave vector $\boldsymbol{k}$ shows that the most dangerous perturbations have the wavenumber

$$|\boldsymbol{k}|^2 \equiv k^2 = \frac{1}{2}\left(\frac{f_u}{\gamma_1 D_1} + \frac{g_v}{\gamma_2 D_2}\right) . \tag{3}$$

This value should be positive, which is possible only in the presence of positive feedback, or, in chemical terms, when at least one of the species is "autocatalytic", say, $f_u > 0$. Breaking of spatial symmetry preempts Hopf bifurcation, which occurs at $\gamma_1^{-1} f_u + \gamma_2^{-1} g_v = 0$ and leads to homogeneous oscillations, provided only one of the species is autocatalytic, so that $g_v < 0$, and the autocatalytic species is less diffusive. Thus, for spatial symmetry breaking in a two-component system, one needs a combination of a slowly diffusing "activator" and a rapidly diffusing "inhibitor".

The development of a pattern can be understood qualitatively in the following way. A local upsurge of the activator concentration increases also the concentration of the inhibitor, which spreads out suppressing the activator at neighboring locations. This, in turn, suppresses the inhibitor locally and, through inhibitor diffusion, enhances the activator further along the line, so that the inhomogeneous state spreads out. This scheme works with the roles of an activator and an inhibitor played, respectively, by prey and predator in population dynamics, by growing plants and seeping moisture in ecology, or, rather less directly, by buoyancy and heat conduction in natural convection.

Pattern formation may also result from nonlocal interactions. For example, a nonlocal extension of the nonlinear Schrödinger equation (NLS) for a complex field u,

$$-\mathrm{i}\, \partial_t u = \nabla^2 u - u(\boldsymbol{x}) \int U(\boldsymbol{x} - \boldsymbol{\xi}) |u(\boldsymbol{\xi})|^2 \mathrm{d}\xi \,, \tag{4}$$

generates a patterned state known as "supersolid", as compared and contrasted to superfluid solutions of the local NLS [13]. It might be possible to derive nonlocal equations from a local RDS. Thus, if in Eq. (2) $\gamma_2 \ll \gamma_1$, so that the inhibitor is fast as well as diffusive, the time derivative can be neglected; then, if the function $g(u, v)$ is linear in v, Eq. (2) can be resolved with the help of an appropriate Green's function, and substituting it in Eq. (1) yields a nonlocal activator equation.

Selection of Stationary Patterns

Symmetry breaking transitions in more than one dimension are degenerate due to spatial symmetries. In an isotropic system, an arbitrary number of differently directed modes with $k = |\boldsymbol{k}| =$ idem can be excited beyond the bifurcation point. A combination of these modes can give a variety of distinct *planforms*. Competition among the modes that determines the pattern selection is described by *amplitude equations* describing evolution of complex amplitudes a_j, which have a general form

$$\begin{aligned} \frac{\mathrm{d}a_j}{\mathrm{d}t} &= -\frac{\partial V}{\partial \overline{a}_j} \,, \\ V &= -\mu \sum |a_j|^2 + \sum \nu_{ijk} a_i a_j a_k \\ &\quad + \sum \nu_{ijkl} a_i a_j a_k a_l + \text{c.c.} \end{aligned} \tag{5}$$

Here the coefficient μ is proportional to the deviation from the bifurcation point; real coefficients ν_{ijk}, ν_{ijkl} characterize nonlinear interactions among the modes; the summation is carried out over all closed polygons formed by the wave vectors of extant modes. The product of the amplitudes $\overline{a}_j$, $\overline{a}_k$, etc. (where the overline denotes the complex conjugate) may appear in the equation for the amplitude a_i if the respective wave vectors add up to zero, $\boldsymbol{k}_i + \boldsymbol{k}_j + \boldsymbol{k}_k + \cdots = 0$. This condition ensures that the modes in question are in resonance. Otherwise, the product of these modes rapidly oscillates and is averaged out when the amplitude equation is derived using a multiscale expansion procedure. Stationary solutions, i. e. potential minima of Eq. (5) with one, two, three, or more non-vanishing modes with a symmetric star of wave vectors correspond, respectively, to a striped, square, hexagonal, or quasicrystalline pattern.

The cubic term in the potential (5) generates the lowest-order, hence, strongest nonlinear interactions. This

term vanishes in the presence of inversion symmetry $a \to -a$, which exists, in particular, in the thoroughly studied case of buoyancy-driven convection in the Boussinesq approximation. Otherwise, it is dominant near the bifurcation point, causing (in 2D) a subcritical transition to a hexagonal pattern comprising modes forming a regular triangle. These three modes are in resonance, which means that their phases are not independent but bound by a linear relationship. The sum of phases always adjusts in such a way that interactions are destabilizing. The remaining two phase degrees of freedom correspond to translational symmetry in the plane.

In 3D, the preferred patterns, or crystalline structures, comprise wave vectors forming a regular polyhedron with triangular faces – tetrahedron, octahedron or dodecahedron [14]. The former two correspond to a body-centered cubic (bcc), and the last one, to a quasicrystalline structure with fivefold symmetry. These lowest-order interactions cannot, however, stabilize the pattern at a finite amplitude, and next-order interactions generated by the quartic term in Eq. (5) are necessary to saturate the pattern. Depending on respective interaction coefficients, various structures can be chosen.

A greater variety of patterns may arise if planforms with different wavenumbers k are excited simultaneously. This can be achieved in a most natural way in two-layer systems where the wavelength of the excited pattern depends on the thickness of each layer, as in convection [15], or different diffusivities, as in a pattern-forming chemical system [16]. More possibilities arise in nonlinear optics where spatial symmetry breaking may occur on different wavelengths at rather close values of a control parameter [17]. The resulting coupled amplitude equations can generate a variety of composite planforms, which may have a form of superstructures or quasicrystals. Lowest-order interactions can generate various resonances; no rigid fitting of wavenumbers is required for this, since resonant modes can form an isosceles triangle. Dynamics of mode interactions may be complicated [18], since the gradient structure of Eq. (5) is, generally, lost.

Regular patterns may suffer various instabilities, which limit the range of admissible wavelengths or lead to a change of the planform through excitation of a noncollinear mode or decay of an extant mode. Wavelength changing instabilities, as a rule, do not saturate and lead to formation of defects.

Modulated and Distorted Patterns

Natural patterns seen both in experiment and simulations are never perfect: their amplitudes may be modulated at distances large compared to the basic wavelength, and they may have various defects: dislocations, disclinations, and domain walls. An example of an imperfect striped pattern is shown in Fig. 2. Variation of local wavelengths is possible because instability spreads out to a finite range of wavenumbers, scaled as the square root of the parametric deviation from the bifurcation point. Other imperfections are a consequence of the rotational symmetry of the system. Different orientations of stripes may be chosen at different locations, either randomly or under influence of boundary conditions or local inhomogeneities. The discrepancies of local orientations are reconciled through formation of disclinations and domain walls, while dislocations reconcile discrepancies of local wavelengths.

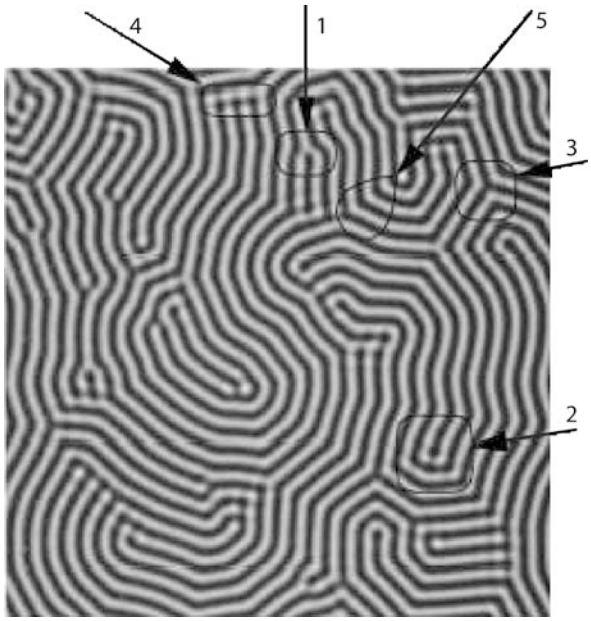

Patterns and Interfaces in Dissipative Dynamics, Figure 2
Various forms of pattern defects. *1* – dislocation, *2* – concave disclination, *3* – convex disclination, *4* – amplitude domain wall, *5* – phase domain wall ([19], reproduced with permission. Copyright by the American Physical Society)

Weak distortions, which do not contain defects, can be described by means of either space-dependent amplitude equations applicable to small-amplitude patterns near the bifurcation point, or phase dynamics applicable also to finite-amplitude patterns but restricted to long-scale distortions.

The amplitude equation must have an anisotropic form in an isotropic system, the source of anisotropy being the direction of the wave vector itself. Modulations of this amplitude along and across the direction of the wave vector $\boldsymbol{k}$ should be scaled differently, since adding a small lon-

gitudinal component, say, ϵq_x changes $k = |\boldsymbol{k}|$ by $O(\epsilon)$, while adding a transverse component of the same magnitude ϵq_y changes k by $O(\epsilon^2)$ only; thus the stripes are bent far more easily than they are compressed or extended. This leads to the Newell–Whitehead–Segel (NWS) amplitude equation [20,21], which can be written in a rescaled universal form

$$\partial_t u = \left(\partial_x - \frac{i}{2k}\partial_y^2\right)^2 u + u - |u|^2 u \,. \tag{6}$$

The mixed-order differential operator entering this equation precisely accounts for the equivalence of all structures with identical wavenumbers, independently of the direction of the wave vector.

The NWS equation is ill-suited for computations, since the orientation of the coordinate axes depends on the local phase gradient, so that the differential operator is in fact strongly nonlinear. Most model computations of striped patterns are based on the Swift–Hohenberg (SH) equation

$$\partial_t u = -(1 + \nabla^2)^2 u + u(\mu - u^2) \,. \tag{7}$$

In an *anisotropic* system where a certain direction of stripes is preferred, the situation is easier, and the amplitude equation can be reduced by rescaling to an *isotropic* real Ginzburg–Landau (RGL) equation

$$\partial_t u = \nabla^2 u + u - |u|^2 u \,. \tag{8}$$

Phase Dynamics

The idea of phase dynamics [22] is to characterize a striped pattern by means of a single variable – phase θ, which changes by 2π over the period of the pattern or, more conveniently, by a rescaled phase $\Theta = \epsilon\theta$. The derivatives of the phase are the wave vector $\boldsymbol{k} = \nabla\theta$ and frequency $\omega = -\theta_t$, which vary on an extended scale exceeding the wavelength of the underlying structure by a factor $\epsilon^{-1} \gg 1$. The general form of the phase equation in an isotropic system is determined by scaling and symmetry considerations alone:

$$\partial_T \Theta = D_1(\boldsymbol{n} \cdot \widehat{\nabla})^2 \Theta + D_2 \widehat{\nabla}^2 \Theta \,, \tag{9}$$

where ∂_T, $\widehat{\nabla}$ are derivatives with respect to slow time and extended spatial variables, $\boldsymbol{n}$ is the unit vector along $\boldsymbol{k}$, and D_1, D_2 are phase diffusivities that depend on a particular underlying problem and are, generally, functions of k. This equation can be also presented in an elegant gradient form [23].

The phase equation (9) is, in fact, strongly nonlinear due to the dependence of both the diffusivities and the direction of the unit vector $\boldsymbol{n}$ on the local phase gradient. It can be linearized, yielding an anisotropic diffusion equation, only when deviations from a prevailing wave vector $\boldsymbol{k} = \boldsymbol{k}_0$ are arbitrary small. If the X- and Y-axes are drawn, respectively, along and across $\boldsymbol{k}_0$, (9) reduces to

$$\Theta_T = D_\parallel(k_0)\Theta_{XX} + D_\perp(k_0)\Theta_{YY} \,, \tag{10}$$

where $D_\parallel = D_1 + D_2$ and $D_\perp = D_2$ are, respectively, the longitudinal and transverse phase diffusivities. The pattern with the wavenumber k_0 is stable to long-scale perturbations when both phase diffusivities are positive. Vanishing $D_\parallel$ corresponds to the *Eckhaus* instability and vanishing $D_\perp$ to the *zigzag* instability. Eckhaus instability defines the upper limit of stable wavenumbers. It never saturates, and usually leads to formation of defects effectively increasing the wavelength. Zigzag instability defines the upper limit of stable wavenumbers; it causes bending of stripes effectively decreasing the wavelength.

Dynamics of Defects

Dynamics of strongly distorted patterns is mostly governed by motion and interaction of defects. Defects are topological objects [24]: a dislocation is characterized by circulation of the phase around any enclosing contour equal to an integer multiple of 2π, and a disclination, by

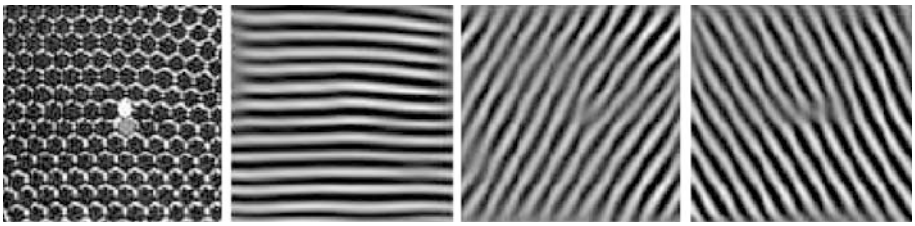

Patterns and Interfaces in Dissipative Dynamics, Figure 3
Hexagonal pattern containing a penta-hepta defect (*left*) and its three constituent modes obtained by Fourier filtering of the initial image ([32], reproduced with permission)

circulation of the direction of the wave vector equal to an integer multiple of π. A single dislocation climbing across the direction of the wave vector of a striped pattern effects a change of the wavenumber over an extended region. The force driving the dislocation is due to the deviation from the optimal wavenumber. Eckhaus instability of a striped pattern leads to the formation of a dislocation pair. It is notable that, although the far field of dislocations can be described by phase equations, their interaction is determined by the dislocation core where these equations are inapplicable [25,26].

Motion of dislocations in striped patterns is well understood and supported by experimental evidence [27] for anisotropic patterns governed by Eq. (8). The structure of dislocations in isotropic systems described by Eq. (6) is more complicated, being strongly anisotropic [28]. Disclinations pose more difficulties for the analysis, even on the topological level [24], see [29].

Paradoxically, defects enhance relaxation of the pattern to a state of minimum energy corresponding to an "optimal" wavelength. If a deviation of the control parameter from the symmetry breaking bifurcation point is of $O(\epsilon^2)$, the width of the band of excited modes is of $O(\epsilon)$, but the band width actually observed in a natural patterns containing defects is of $O(\epsilon^2)$ [19]. The band shrinks due to motion of point defects and adjustments influenced by domain walls.

The structure and interaction of dislocations in a hexagonal pattern is strongly affected by the resonant character of interactions among the constituent modes. Dislocations in any two modes of the triplet forming a hexagonal pattern, created originally at arbitrary locations, are always attracted to each other [30,31], eventually forming an immobile bound pair corresponding to a penta-hepta defect (see Fig. 3).

Equations (6), (8) are derivable from an energy functional that decreases monotonically in time until a stationary state of minimal energy is reached; this state may still contain defects necessary to satisfy boundary conditions in a confined region. In some cases, however, an additional field, besides the amplitude, is necessary to adequately describe a physical system even close to the symmetry-breaking bifurcation point. A well known example is Bénard convection in low Prandtl number fluids where the additional factor is mean flow generated by pattern distortions and advecting the entire pattern. In this case, the patterns remains weakly turbulent indefinitely long, displaying labyrinthine structures, coexisting striped and hexagonal domains [33] or spiral defect chaos [30] (see Fig. 4). Chaotic non-stationary patterns also typically appear at higher amplitudes. In reaction-diffusion systems non-stationary and chaotic patterns become more likely when the inhibitor response is slowed down.

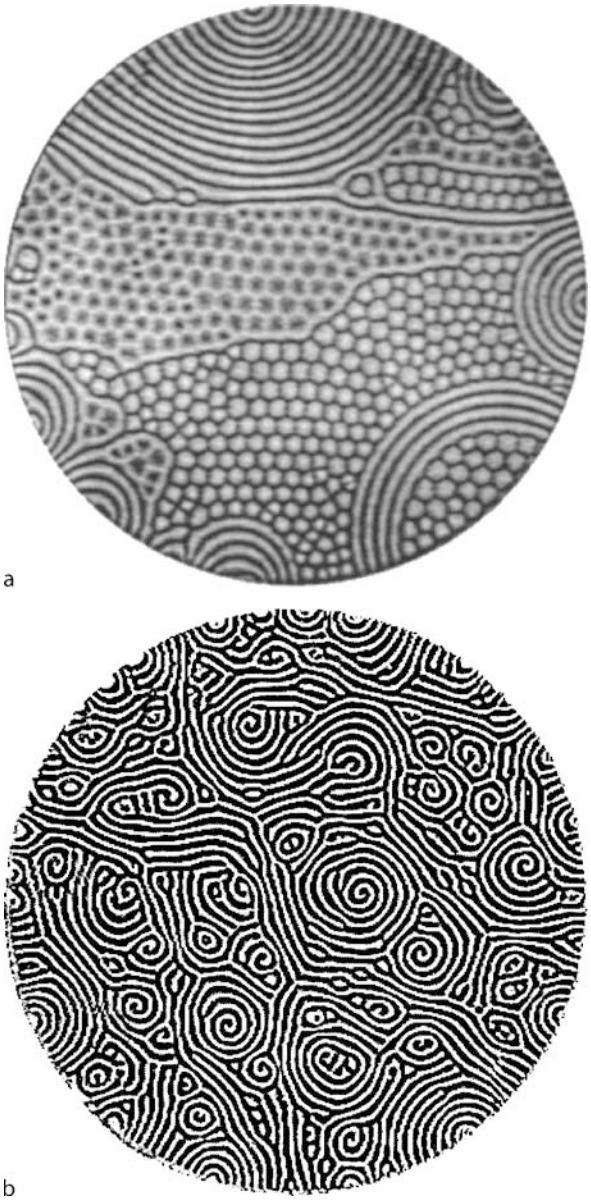

Patterns and Interfaces in Dissipative Dynamics, Figure 4
a Coexisting domains; b Spiral defect chaos [30] (reproduced with permission)

Moving Interfaces

Stationary and Propagating Fronts

Many physical systems, either at equilibrium or in a non-equilibrium steady state sustained by external fluxes, may exist in two or more alternative states. If different states are

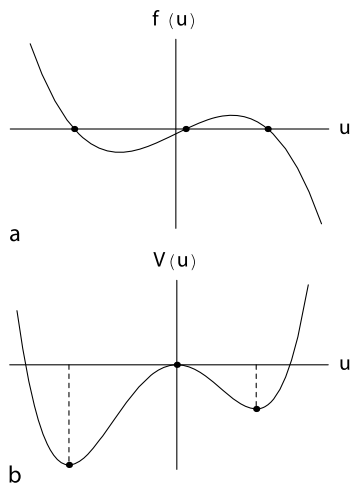

Patterns and Interfaces in Dissipative Dynamics, Figure 5
A function $f(u)$ with three zeros (a) and the respective double-well potential (b)

attained at different spatial locations, they are separated by an *interface*, carrying excess energy. The simplest model is a single "reaction–diffusion" equation

$$\partial_t u = D\nabla^2 u + f(u)\,, \tag{11}$$

where D is diffusivity and the function $f(u) = -V'(u)$ (see Fig. 5) has three zeroes that correspond to two stable (with $f'(u) < 0$) and one unstable (with $f'(u) > 0$) HSS. This equation was first used in the context of phase equilibria [34] as a model of gas–liquid interface, with u denoting density. It was later extended to the solidification problem, with u denoting a fictitious "phase field" assuming its two stable values $u = u_s^{\pm}$ in the liquid and solid phases [35]. The coefficient D is interpreted in this context as *rigidity*. The "reaction-diffusion" interpretation applies to non-equilibrium systems, such as a catalytic surface or an ecological domain, with u denoting concentration and $f(u)$, the net production rate.

A straight-line or planar interface is stationary when the potentials $V(u_s^{\pm})$ are equal. It carries then the interfacial energy

$$\sigma = D\int_{-\infty}^{\infty} u'(x)^2 \mathrm{d}x = \int_{u_s^-}^{u_s^+} \sqrt{2DV(u)}\mathrm{d}u\,, \tag{12}$$

which is identified with surface tension.

If the potentials are unequal, the front moves in the direction decreasing the total energy of the system. Assuming that the motion is stationary and directed along the x axis, (11) can be rewritten in the comoving frame propagating with the front velocity c. The steadily propagating solution depends on a single coordinate $\xi = x - ct$, and

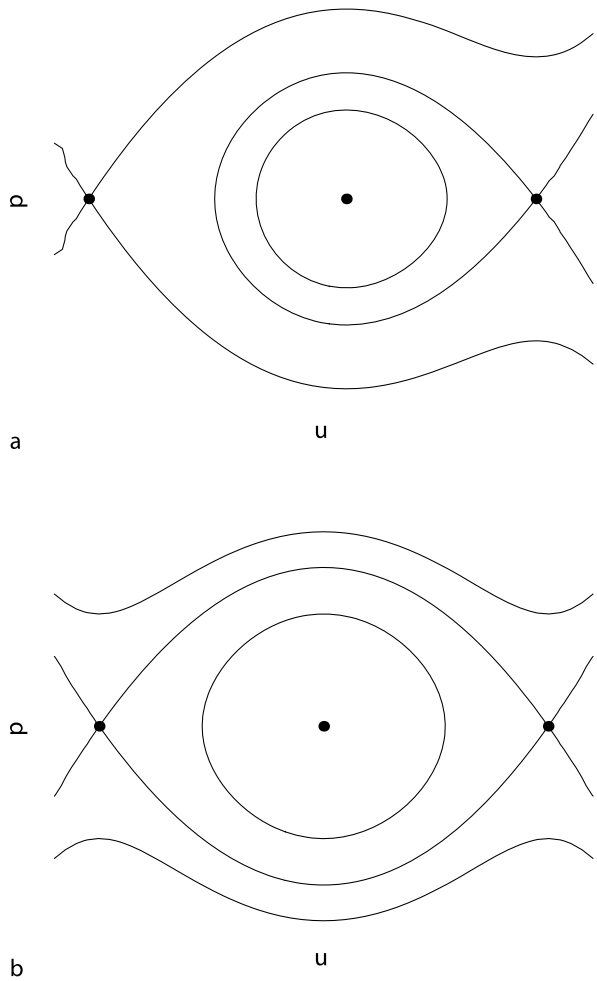

Patterns and Interfaces in Dissipative Dynamics, Figure 6
Generic trajectories in the phase plane $u, p = u'(x)$ (a) and a non-generic set of trajectories containing a heteroclinic orbit (b)

(11) reduces to an ordinary differential equation

$$cu'(\xi) + Du''(\xi) + f(u) = 0\,, \tag{13}$$

subject to the boundary conditions $u = u_s^{\pm}$ at $\xi \to \pm\infty$. When both equilibria are stable, they are saddles when viewed as equilibria of (13). The front solution corresponds to a heteroclinic trajectory connecting the equilibria $u = u_s^{\pm}$. The heteroclinic connection exists only at unique value of c (see Fig. 6); thus, the propagation speed is determined uniquely by solving a nonlinear eigenvalue problem. Its value is proportional to the difference of potentials of the two HSS:

$$c = \frac{D}{\sigma}\Delta V\,, \quad \Delta V = V(u_s^-) - V(u_s^+)\,. \tag{14}$$

The situation is different when the retreating state $u = u_s^0$ is unstable. This often happens in population dy-

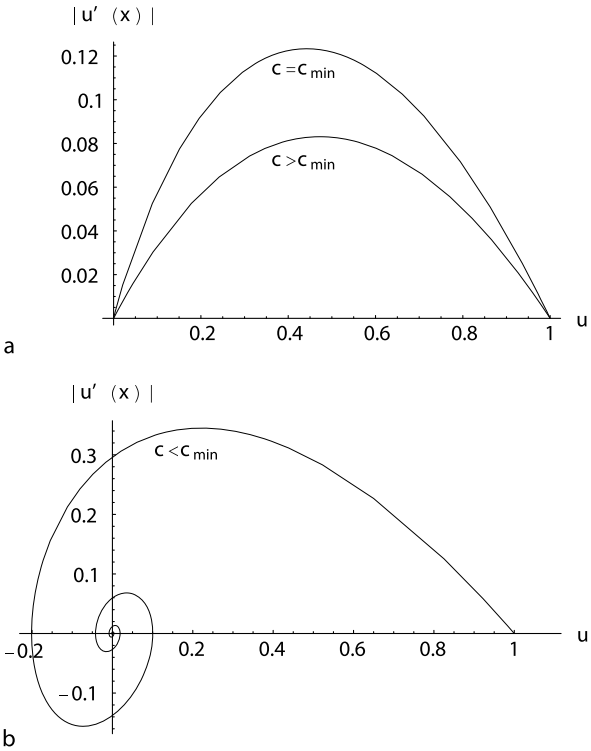

Patterns and Interfaces in Dissipative Dynamics, Figure 7
Trajectories in the phase plane connecting a stable and an unstable equilibrium

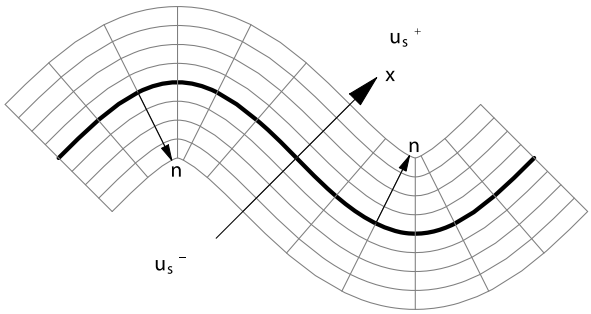

Patterns and Interfaces in Dissipative Dynamics, Figure 8
Construction of the aligned coordinate frame. The coordinate lines are shown in *gray*. *Arrows* show the local directions of the normal n and the *x*-axis. Observe a singularity developing on the concave side

namics: a state where a competitively advantageous specie is absent is formally unstable to infinitesimal perturbations but will be nevertheless preserved at any location until this specie is introduced there. An unstable state, viewed as an equilibrium point of (13), is a stable node at propagation speeds exceeding a certain threshold $c_{\min}$, Thus, a trajectory starting from the advancing stable HSS connects generically to u_s^0 at any $c > c_{\min}$ (see Fig. 7). Actual propagation speed is selected dynamically at the leading edge [37,38], and turns out to be equal to the minimum speed $c_{\min}$, which corresponds to the steepest front profile. Under certain conditions (when overshoots are allowed) a faster speed corresponding to a still steeper profile is selected nonlinearly [38]. In the former case, the front is "pulled" by perturbations growing at the leading edge and described by linearized equations, while in the latter case, it is "pushed" by nonlinear interactions favoring the advancing state.

Interfacial Instabilities

The front solution is neutrally stable to translations along the x-axis. This neutral (Goldstone) mode is weakly perturbed when the translation is weakly nonuniform, so that the front becomes curvilinear but the curvature radius still far exceeds the characteristic front thickness.

Propagation of a weakly curved front is best understood in a coordinate frame aligned with its deformed shape. The nominal front position is defined by replacing a diffuse transitional region by a planar curve C drawn along some intermediate level of the variable u. The coordinate lines $x =$ const are obtained by shifting the curve along the normal by a constant increment, as shown in Fig. 8. This shift causes the length to increase on convex, and to decrease on concave side of the curve. Eventually, a singularity develops in the latter direction, but, when the curvature radius is much larger than the characteristic front thickness, this will happen far away within the region where one of the HSS is approached.

When (13) is rewritten in the aligned frame an expanded viewing the curvature as a small parameter, the local normal propagation speed of a curved front is expressed by the eikonal equation

$$c = c_0 - D\kappa = \frac{D}{\sigma}\left(\Delta V - \sigma\kappa\right) , \tag{15}$$

where c_0 is the speed of a planar front and κ is the Gaussian curvature.

Since convex front segments propagate slower and concave segments faster, the front tends to flatten, provided c_0 is uniform everywhere. Instabilities may arise, however, when c_0 increases ahead of the front. This may happen in the presence of an externally imposed gradient, as in directional solidification [11], but most commonly is caused by an additional "control" field. The control field responsible for the Mullins–Sekerka instability of solidification fronts [11,39] is the concentration of a contaminant, which is rejected by the solid and slows down so-

lidification by lowering the melting temperature. Since the contaminant diffuses away more easily from convex segments, they tend to propagate faster, which causes instability when the driving is strong enough to overcome surface tension.

Another example is instability of a combustion front, which separates hot burnt-out and cold fuel-rich domains [10]. A thin front structure arises in this case because combustion requires both fuel and sufficient temperature for its initiation, and both fuel concentration and temperature play the role of control variables. When heat transfer is the limiting factor, convex segments cool down and propagate slower, and the front is stable. When, on the opposite, propagation is limited by fuel supply, convex segments accelerate and instability sets on, leading to corrugated fronts.

Dynamics of weak deviations $\zeta(\boldsymbol{y})$ from a stable planar front spanned by a 2-vector $\boldsymbol{y}$ is described by expanding the normal propagation speed, front curvature and the control field in powers of a small parameter scaling both the deviation ζ and its transverse derivative $\nabla_{\mathbf{y}}$, as well as time. For stable fronts, the appropriate scaling is $\zeta = O(1)$, $\nabla_{\mathbf{y}} = O(\epsilon)$, $\partial_t = O(\epsilon^2)$, leading to the Burgers equation

$$\partial_t \zeta = D\nabla_{\mathbf{y}}^2 \zeta - \tfrac{1}{2} c_0 |\nabla_{\mathbf{y}} \zeta|^2 \,. \tag{16}$$

The particular coefficients here correspond to (15), but also in other cases the same universal form can be obtained after the coefficients are removed by rescaling, provided the effective diffusivity D is positive. If the latter is negative but small, $|D| = O(\epsilon^2)$, the appropriate scaling is $\zeta = O(\epsilon)$, $\nabla_{\mathbf{y}} = O(\epsilon)$, $\partial_t = O(\epsilon^4)$, and expanding to a higher order yields, after scaling away the coefficients, the Kuramoto–Sivashinsky equation [40]

$$\partial_t \zeta + \nabla_{\mathbf{y}}^2 \zeta + (\nabla_{\mathbf{y}}^2)^2 \zeta + \tfrac{1}{2} |\nabla_{\mathbf{y}} \zeta|^2 = 0 \,. \tag{17}$$

This equation, appearing also in phase dynamics [41], is a paradigm of weak turbulence.

Front Interactions and Coarsening

Fronts of opposite polarity in a one-dimensional system attract and eventually coalesce, thereby coarsening the distribution of domains, which may have been created initially in the process of phase separation or relaxation to alternative HSS. The interaction is, however, very weak, falling off exponentially with separation. In higher dimensions, the principal cause of coarsening, or Ostwald ripening, is the curvature dependence of the propagation speed, whereby small droplets with high curvature tend to shrink and eventually disappear. This is a manifestation of the Gibbs–Thomson effect relating the equilibrium conditions with the radius of a droplet.

Coarsening most often occurs under conditions when evolution is constrained by a conservation law, so that the integral $\int u(\mathbf{x})\mathrm{d}\mathbf{x}$ expressing the total amount of material in the system remains constant. Under these conditions, fronts cannot move independently from each other. The conservation law is accounted for when (11) is replaced by the Cahn–Hilliard equation [35]

$$\partial_t u = \nabla^2 \mu \,, \quad \mu = -[D\nabla^2 u + f(u)] \,. \tag{18}$$

The eikonal equation governing the front motion retains the form (15), but the value c_0 depends on chemical potential μ. The latter shifts in the course of coarsening in such a way that the value of the critical radius $R = \kappa^{-1}$ of a droplet that neither grows or shrinks, keeps growing as smaller droplets disappear. Analytical theory [42] predicts universal asymptotic droplet size distribution at late stages of coarsening.

Structures Built up of Fronts

Coarsening can be precluded when changes in an additional control field arrest growth of large and shrinking of small domains. This leads to formation of a variety of patterns and solitary structures. The paradigmal system for exploring these phenomena is the FitzHugh–Nagumo system, which has the form (1), (2) with the function $f(u, v)$ cubic in u and linear in v and a linear function $g(u, v)$. The rescaled form suitable for the analysis of stationary structures is

$$\epsilon^2\, \partial_t u = \epsilon^2 \nabla^2 u + u - u^3 - \epsilon v \,, \tag{19}$$

$$\tau^{-1} \partial_t v = \nabla^2 v - v - v + \mu u \,, \tag{20}$$

Here $\epsilon = \sqrt{\gamma_1 D_1/\gamma_2 D_2} \ll 1$ is the ratio of the characteristic lengths associated with the activator and the inhibitor, $\tau = D_2/D_1$; the small coupling parameter ε in (19) ensures a balance between the effect of small interfacial curvature and weak symmetry breaking between the alternative HSS $u_{\mathrm{s}}^{\pm} = \pm 1 + O(\epsilon)$; the remaining parameters μ and ν regulate the coupling stress and bias.

Structures generated by the system (19), (20) are built up by assigning a region where the activator approaches one of the alternative HSS, computing the respective inhibitor distribution, and finding stationarity conditions for the fronts forming the boundaries of this region [43]. Possible stationary structures in two dimensions are a solitary band, a solitary disk, a striped pattern, or a hexagonal grid consisting of almost circular spots. The size of spots

or stripes is determined by the parameters of the system, but there is a considerable leeway in choosing the general configuration. Under certain conditions, it even might be possible to store information by creating or extinguishing spots at chosen locations [44]. In other cases, splitting of a solitary spot initiates a multiplication cascade [45], leading eventually to a hexagonal pattern filling the plane.

Instabilities of stationary structures are studied with the help of the linearized eikonal equation (15) combined with the inhibitor equation (20) where the last term is expressed through a shift of the front position. Both solitary bands and disks can suffer zigzag (leading eventually to splitting), oscillatory and traveling instabilities. The latter two become prevalent as the parameter τ decreases, so that the inhibitor response to front displacements slows down. For example, a solitary band is destabilized in the zigzag mode at $\tau > 1$, while the traveling instability comes first at smaller τ (see Fig. 9). Oscillatory instability is always preceded by traveling one in this case, but may become relevant for a solitary disk.

Traveling instability indicates transition to various propagating structures and wave patterns. A solitary spot tends to either dissolve or spread out sidewise after being immobilized; in the latter case, a spiral structure starts to develop as the ends lag behind. A traveling spot can be, however, stabilized if a second inhibitor, both fast and long-range, is added [46].

Various patterns of propagating fronts can be generated beyond this limit by the same FitzHugh–Nagumo system, which, however, should be scaled differently for this purpose. Unlike stationary or slowly evolving patterns where the characteristic length scale is set by the diffusional range of the long-scale inhibitor, the wavelength of a propagating pattern is tied to the propagation speed and remains finite even when the inhibitor is nondiffusive. The long scale should be redefined therefore on the basis of the characteristic propagation speed of the activator front $c^* = \sqrt{D_1/\gamma_1}$ and the characteristic relaxation time of the inhibitor γ_2. Using this "advective" length unit, $L^* = \gamma_2\sqrt{D_1/\gamma_1}$ brings (1), (2) to the dimensionless form

$$\gamma\partial_t u = \gamma^2\nabla^2 u + f(u,v)\,, \tag{21}$$

$$\partial_t v = \delta^2\nabla^2 v + g(u,v)\,, \tag{22}$$

where $\gamma = \gamma_1/\gamma_2$, $\delta = \gamma/\epsilon = \sqrt{\gamma D_2/D_1}$. The "inner" scale of the transitional layer, where the system switches between the two alternative activator states, $u = u_s^{\pm}$, is now set exclusively by the capacitance ratio γ, independently of diffusivities, and, provided $\gamma \ll 1$, remains small even when the inhibitor is less diffusive than the activator. The parameters can be chosen in such a way that $\delta \ll 1$, so that the inhibitor diffusion is negligible, provided $\gamma \ll D_1/D_2$. Under these conditions, the inhibitor diffusion can be neglected, reducing (22) to $\partial_t v = g(u,v)$. Although this equation contains no mechanism for healing discontinuities in v, the inhibitor field should remain smooth in the course of evolution, barring freaky initial conditions or strongly localized perturbations. This opens the easiest way of constructing various wave patterns, including such exotic objects as chaotic wave trains [47].

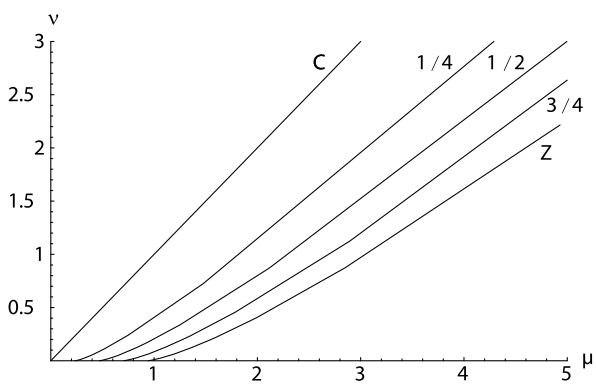

Patterns and Interfaces in Dissipative Dynamics, Figure 9
Existence boundary (C) and loci of zigzag (Z) and traveling instability for a solitary band. The loci of traveling instability are marked by respective values of τ. A stable band exists between the line C and an applicable instability locus

Interfaces of Patterns

Interfaces between different patterns or different pattern orientations (domain walls) can be described in the simplest way on the level of amplitude equations. This may give qualitatively correct results in static problems, even though changes across a domain wall in patterns generated in simulations and experiments are usually effected on a length comparable with the prevailing wavelength of the pattern. One can expect that a stationary solution exists only when the wavelengths are equal on both sides of the wall; otherwise, the wall would propagate in the direction decreasing the overall energy of the pattern. It turns out that an even stronger restriction is true, and both wavelengths should be optimal [48]. In this way, domain walls, alongside dislocations, enhance relaxation of the pattern to the optimal wavelength.

Dynamic problems are strongly influenced by detailed structure of the pattern, which is lost on the level of amplitude equations. When a pattern advances into an unstable

Patterns and Interfaces in Dissipative Dynamics, Figure 10
A scheme of depinning transitions showing crystallization (C) and melting (M) thresholds for an infinite cluster, as well as the corresponding limits for clusters of different sizes, terminating in single-cell limits 1-C, 1-M

uniform state, the wavelength selected at the leading edge is not identical to the wavelength of the full-grown pattern formed behind the front, and neither one coincides with the optimal wavelength [49].

In the case when a stable homogeneous solution coexists with a stable periodic pattern, stable stationary fronts between the two states exist within a finite parametric interval [50], rather than at a single point where the energies of both states are equal, as amplitude equations would predict. The motion of this front is affected by the discrete structure of the pattern, which causes self-induced pinning hindering the retreat of a metastable state. There are two depinning transitions, corresponding to "crystallization" or "melting" of the pattern, shown schematically by thick lines in Fig. 10. Between the two limits, various metastable stationary structures exist: a single cell ("soliton"), a finite patterned inclusion, sandwiched between semi-infinite domains occupied by a uniform state, or a semi-infinite pattern, coexisting with a uniform state. To the right of the crystallization threshold C, the pattern advances by a periodic nucleation process which creates new elementary cells at the interface [51], while to the left of the melting limit M, the pattern recedes as elementary cells at the interface are destroyed. A different, far more efficient depinning mechanism works in two dimensions [52]. It is initiated by a zigzag instability of the pattern followed by nucleation of disclinations, which further move toward the uniform state, as seen in Fig. 11. This generates stripes extending in the normal direction, turning eventually the original boundary into a domain wall separating striped patterns rotated by $\pi/2$.

Wave Patterns

Plane Waves

A simplest propagating wave pattern is a periodic solution depending on a moving coordinate $\xi = x - ct$, where $c = \omega/k$ is phase velocity, ω is frequency and k is wavenumber. A waveform $\sim \exp[\mathrm{i}(kx - \omega t)]$ may emerge directly by symmetry breaking bifurcation, but this is not the most common mechanism. It is impossible, in particular, in a two-component RDS (1), (2), where other scenarios lead to wave patterns. One of them, mentioned in the preceding Section, is traveling instability of stationary structures. Another road to wave patterns, most amenable to analytical tools, starts in the vicinity of a Hopf bifurcation, where small-amplitude oscillations weakly modulated in space are described by the complex Ginzburg–Landau (CGL) equation. Its standard rescaled form is

$$\partial_t u = (1 + \mathrm{i}\eta)\nabla^2 u + u - (1 + \mathrm{i}\nu)|u|^2 u \,. \tag{23}$$

A plane wave solution of (23) with the wave vector $\boldsymbol{k}$ is

$$\begin{aligned} u &= \rho_0 \exp[\mathrm{i}(\boldsymbol{k}\cdot\boldsymbol{x} - \omega t)] \,, \\ \rho_0 &= \sqrt{1-k^2} \,, \quad \omega = \nu + (\eta - \nu)k^2 \,. \end{aligned} \tag{24}$$

The waves are dispersive, and the group velocity is $\boldsymbol{v} = 2\boldsymbol{k}(\eta - \nu)$.

Instabilities of plane waves are studied most efficiently with the help of the phase dynamics approach, since the most dangerous perturbation modes can be viewed as long-scale distortions of neutrally stable translational modes. The longitudinal and transverse phase diffusivities are

$$D_\parallel = \frac{1 + \nu\eta - k^2(3 + \nu\eta + 2\nu^2)}{1 - k^2}, \quad D_\perp = 1 + \nu\eta. \tag{25}$$

Vanishing $D_\parallel$ marks the threshold of Eckhaus instability, which limits the range of stable wavenumbers. Vanishing $D_\perp$ signals Benjamin–Feir (self-focusing) instability, independently of the wavelength. Both instabilities arising

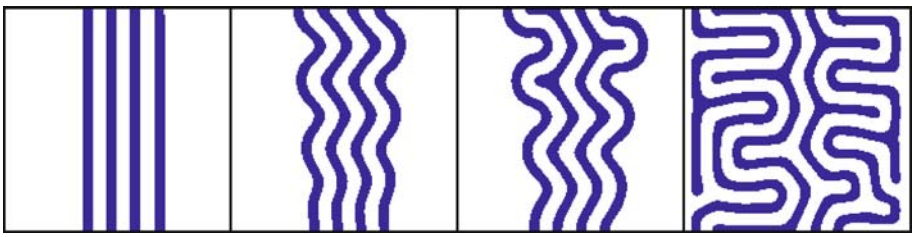

Patterns and Interfaces in Dissipative Dynamics, Figure 11
Depinning of striped pattern initiated by a zigzag instability [52]

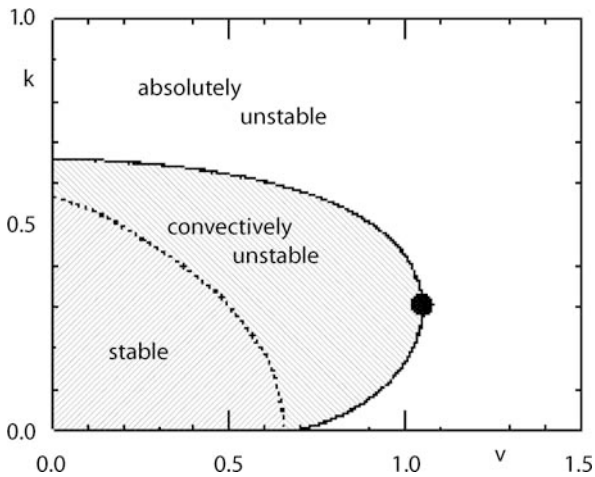

Patterns and Interfaces in Dissipative Dynamics, Figure 12
Limits of convective and absolute instabilities in the plane (ν, k) for $\eta = -3/2$. The *dot* marks the limit of convectively unstable waves ([53], reproduced with permission. Copyright by the American Physical Society)

at the respective thresholds are convective, which means that growing perturbations are washed away with the prevailing group velocity. The absolute instability condition stipulating growth of perturbation at a particular location is less restrictive (see Fig. 12). Numerical simulations [53] show that transition to turbulence occurs only when the absolute stability condition is violated, but the system is very sensitive to noise in the convectively unstable region.

Besides uniform wave trains, there is a variety of nonuniform one-dimensional solutions of the CGL equation with a constant frequency and spatially varying modulus and wavenumber, which are stationary in a frame propagating with a certain speed c and depend on the comoving coordinate $\xi = x - ct$ only. The solutions approaching asymptotically at $\xi \to \pm\infty$ either plane waves or the trivial state can be also viewed as defects separating domains where different uniform states prevail. Such solutions include pulses, approaching the trivial state at both extremes; nonlinear fronts, separating the trivial state from an invading wave train, and domain boundaries separating plane waves directed in the opposite sense and, possibly, having different wavelength [53,54]. Interactions among various defects dominate chaotic dynamics beyond the self-focusing instability limit [55].

Amplitude equations for wave patterns emerging directly from an HSS through a symmetry breaking bifurcation with $\omega \neq 0, k \neq 0$ should account for competition between waves with amplitudes $u^{\pm}$ propagating in the opposite directions, which may either suppress one another or combine to a standing wave. The normalized form of coupled equations for $u^{\pm}$ is

$$\partial_t u^{\pm} \pm c u_x^{\pm} = (1 + \mathrm{i}\eta) u_{xx}^{\pm} + u^{\pm} - (1 + \mathrm{i}\nu_+)|u^{\pm}|^2 u^{\pm} - g(1 + \mathrm{i}\nu_-)|u^{\mp}|^2 u^{\pm} , \tag{26}$$

where g is a coupling parameter. The orders of magnitude of all terms of these equations can be balanced only when the phase velocity $c = \omega/k$ is of the same $O(\epsilon)$ as $u^{\pm}$. Generically, $c = O(1)$, and the advective term $c u_x^{\pm}$ is dominant. For a single wave, it can be removed by transforming to the comoving frame. When both waves are present, each wave, viewed in its own frame $\xi_{\pm} = x \mp ct$ samples the average amplitude of its counterpart propagating in this frame with a fast speed. The appropriate amplitude equations have then the form [56]

$$\partial_t u^{\pm} = (1 + \mathrm{i}\eta) u_{\xi_{\pm}\xi_{\pm}}^{\pm} + u^{\pm} - (1 + \mathrm{i}\nu_+)|u^{\pm}|^2 u^{\pm} - g(1 + \mathrm{i}\nu_-)\langle |u^{\mp}|^2 \rangle u^{\pm} . \tag{27}$$

These equations retain only global coupling carried by the spatial averages $\langle |u^{\mp}|^2 \rangle$.

In two dimensions, the amplitude equations also involve resonant interactions of pairs of waves propagating in the opposite directions. This makes possible complex dynamics even when the amplitudes are uniform and obey space-independent equations [57]

$$\partial_t u_1^+ = u_1^+ \left[\mu - \nu_+ |u_1^+|^2 - \nu_- |u_1^-|^2 - \beta \left(|u_2^+|^2 + |u_2^-|^2\right)\right] + \gamma \overline{u_1^-} u_2^+ u_2^- . \tag{28}$$

Spiral and Scroll Waves

A ubiquitous and extensively studied waveform is a rotating spiral wave. Its specific feature is the presence of a phase singularity. An n-armed spiral wave can be constructed as a circularly symmetric vortex solution of (23) with the topological charge n, i. e. phase circulation $2\pi n$. Unlike a symmetric defect in (8), the phase must also depend on the radial coordinate, so that the vortex radiates a wave with a certain uniquely selected asymptotic wavenumber k_∞. This solution is obtained [58] in polar cooordinates r, ϕ by assuming an *ansatz*

$$u = \rho(r) \mathrm{e}^{\mathrm{i}\theta} , \quad \theta = n\phi + \psi(r) - \omega t . \tag{29}$$

Using this *ansatz* brings (23) to the form

$$\rho''(r) + r^{-1}\rho'(r) + (1 - k^2 - n^2/r^2 - \rho^2)\rho = 0 , \tag{30}$$

$$\frac{1}{r\rho^2} \frac{\mathrm{d}}{\mathrm{d}r}(rk\rho^2) = q(\rho_\infty^2 - \rho^2) , \tag{31}$$

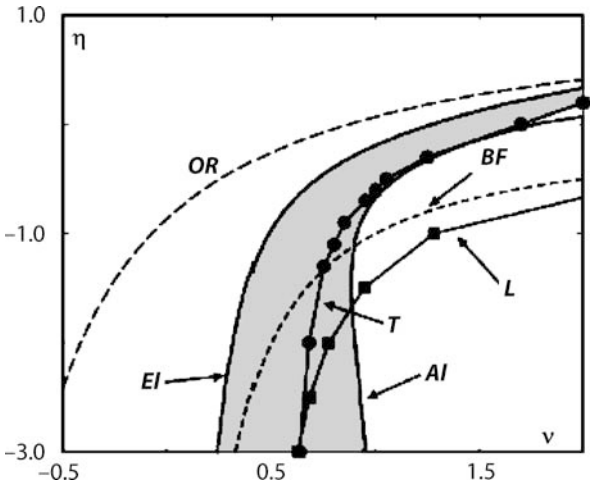

Patterns and Interfaces in Dissipative Dynamics, Figure 13
Stability limits of a spiral wave solution in the parametric plane (η, ν). The curve *EI* shows the limit of convective instability and *AI*, of absolute instability for the waves emitted by the spiral; *OR* is the boundary of the oscillatory spatial decay for the emitted waves, $q = 0.845$ (bound states exist to the right of this line). *BF* indicates the Benjamin–Feir limit $\nu\eta = -1$, *L* is the limit of phase turbulence, and *T* corresponds to the transition to defect turbulence for random initial conditions ([53], based on [59]; reproduced with permission). (Copyright by the American Physical Society)

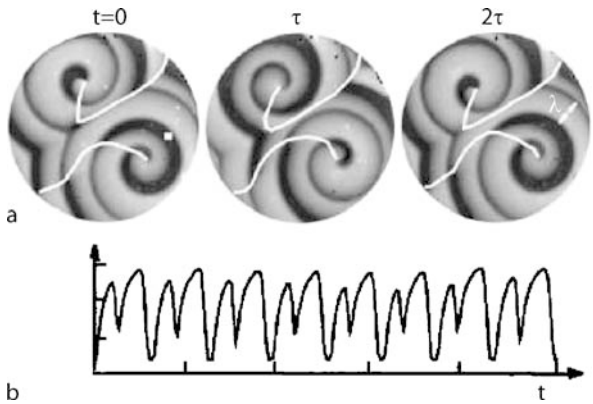

Patterns and Interfaces in Dissipative Dynamics, Figure 14
a A pair of period two spiral waves with the fundamental period τ and the average wavelength λ. The white solid lines are the synchronization defects. b A period two time series measured at the point marked by the *white filled square* ([63], reproduced with permission. Copyright by the American Physical Society)

where $k = \psi'(r)$ is the radial wavenumber. Stability analysis of plane waves applies also to far regions of spiral waves; one could expect therefore a transition to a turbulent state to occur under conditions when the selected asymptotic wavenumber k_∞ falls into the range where the corresponding plane wave solution of (23) is unstable. The respective stability limits in the parametric plane (η, ν) are presented in Fig. 13.

Another approach to constructing rotating spiral waves exploits kinematics of fronts of opposite polarity described by RDS (21), (22) [60]. The inhibitor diffusion can be neglected almost everywhere, except in the crucial tip region where the two fronts meet. Behavior of the spiral tip and its meandering instability has been elucidated analytically using a multiscale technique matching different approximations in overlapping regions [61]. Complex dynamics of a meandering tip, which exhibit quasiperiodic and chaotic motion in some parametric domains, can be well described with the help of a simpler phenomenological model [62]. A similar instability of spiral waves described by the CGL equation is the core acceleration instability [53], which may serve as a trigger of transition to spatio-temporal chaos alternative to instability of radiated waves.

A special kind of spiral wave patterns arises when the underlying dynamical system undergoes a period doubling transition. The period doubling causes the appearance of synchronization defect (SD) lines, which serve to reconcile the doubling of the oscillation period with the period of rotation of the spiral wave (see Fig. 14a). These lines are defined as the loci of those points in the medium where the two loops of the period two orbit exchange their positions in local phase space. The period two oscillations on the opposite sides of a SD are shifted relative to each other by 2π (i. e., a half of the full period), so that the dynamics projected on the rotation direction is effectively of period one, while it is of period two locally at any point in the medium (Fig. 14b).

A three-dimensional extension of a rotating spiral is a rotating scroll wave. The core filament of a scroll wave is a line vortex. A scroll wave with a straight-line core directed along the z-axis has identical spiral waves in each cross-section. Even then, the structure can be nontrivial if the spiral phases are given a phase twist, i. e. are shifted along the z-axis. A curved core filament may also close up into a ring or even form knots. A stable scroll structure evolves to decrease the filament curvature [64]. This kind of dynamics is similar to curvature-driven motion of interfaces, but may be reversed when the filament is unstable. The most dangerous perturbation modes are long-scale modes associated with meandering or translational core deformations [65]. Meandering instability usually saturates as a distorted scroll wave with a twisted rotating core (Fig. 15). Instability in the translation mode,

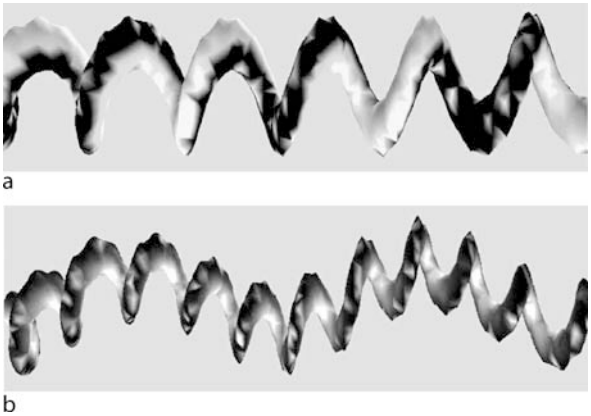

Patterns and Interfaces in Dissipative Dynamics, Figure 15
a A restabilized helical vortex; b A doubly periodic "superhelix" Isosurfaces of the modulus $\rho = 0.6$ shaded by phase field are shown. (CGL simulations [66], reproduced with permission. Copyright by the American Physical Society)

which causes spontaneous bending of the scroll axis, does not saturate, but gives rise to a scroll wave with a continuously extending core (Fig. 16a). This leads to a turbulent state visualized as a tangle of breaking wave fronts (Fig. 16b).

Spiral Patterns and Turbulence

Interaction of spiral waves is dominated by shocks – domain boundaries where waves emanating from different centers collide. The shocks effectively screen different spiral domains from radiation emitted by other spiral cores. A typical example of a spiral domain pattern in a stable parametric range obtained in a CGL simulation run starting from random initial conditions [59] is shown in Fig. 17. At the initial stage, the system tends to relax locally to the stable state with unity real amplitude, but, as the phases are random, the relaxation is frustrated, and a large number of defects – vortices of unit charge – are formed. At the following coarsening stage, oppositely charged vortices annihilate, so that the density of defects decreases. The coarsening process, however, stops halfway, leaving a certain number of single-charged spiral vortices with either sense of rotation. Vortices that failed to conquer a sufficiently large domain are reduced to "naked cores", left to satisfy the topological condition of conservation of circulation. The resulting stable spiral domain pattern is called vortex glass. The waves always propagate outwards from the vortex cores, so that the entire domain structure is generated when local order spreads out from centers to the periphery. Perturbations, also traveling outwards with the prevailing group velocity, are absorbed at shocks, and

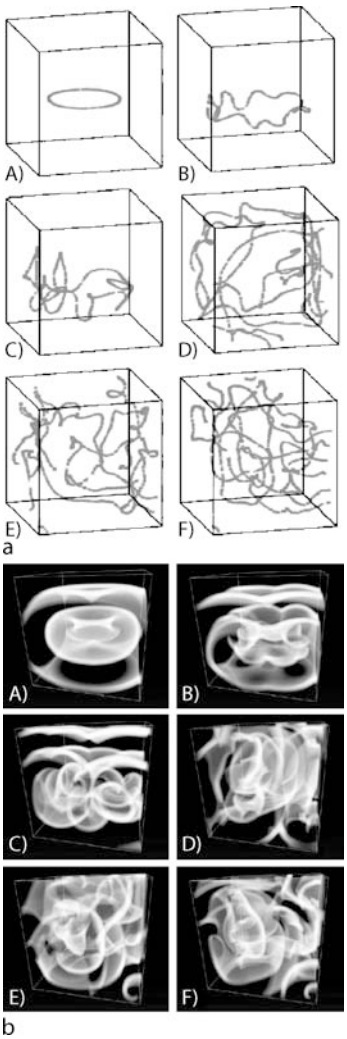

Patterns and Interfaces in Dissipative Dynamics, Figure 16
Transition to turbulence due to core filament extension and breakup of scroll waves. a Snapshots of the core filament, starting from a closed loop. b Respective snapshots of wave patterns showing semitransparent visualization of the activator fronts ([67], reproduced with permission. Copyright by the American Physical Society)

therefore the pattern may survive beyond the convective instability threshold. The turbulent state takes over only when the emanated waves become absolutely unstable, i. e., when some perturbations grow locally in the laboratory frame.

The overall structure of the pattern changes in the range of oscillatory spatial decay of waves emanated by the spiral cores (below the line OR in Fig. 13). Under these conditions, formation of stable bound spiral pairs becomes possible (see Fig. 18). Unlike the monotonic range, spiral domains may have in oscillatory range a wide size distri-

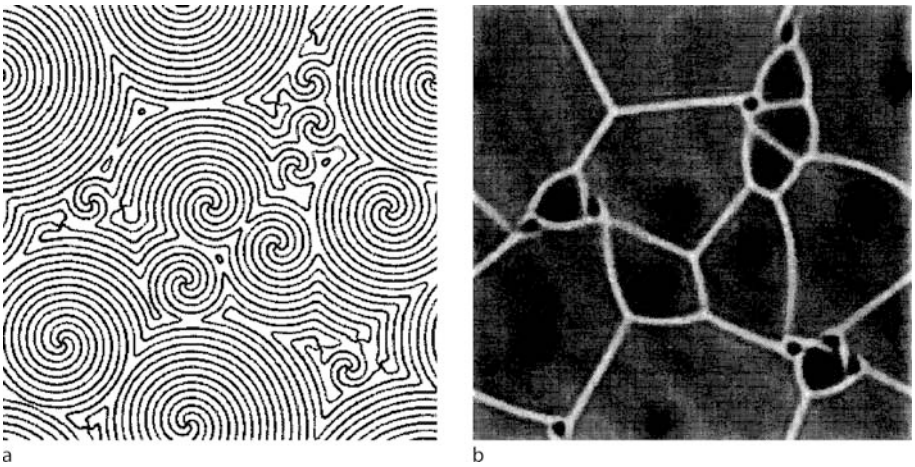

Patterns and Interfaces in Dissipative Dynamics, Figure 17
Spiral domains. *Left*: levels of constant phase. *Right*: grayscale amplitude map showing enhanced amplitudes at the shocks (CGL simulations [59], reproduced with permission from Elsevier Science)

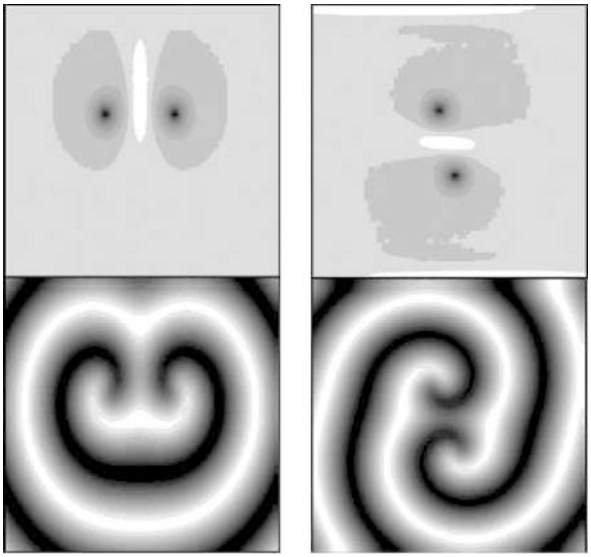

Patterns and Interfaces in Dissipative Dynamics, Figure 18
Bound states of oppositely (*left*) and likely (*right*) charged spirals (CGL simulations, $\eta = 0$, $\nu = 1.5$). The images show the modulus $\rho(x, y)$ (*top*) and Re(u) (*bottom*) ([53], reproduced with permission. Copyright by the American Physical Society)

bution, since shocks can be immobilized at different separations.

"Frozen" glassy patterns actually evolve on a very long time scale, as revealed in very long simulation runs [68]. In the monotonic range, spiral cores perform very slow diffusive motion; the apparent diffusivity increases with vortex density. In contrast, in the oscillatory range, spiral population spontaneously segregates after a very long transient into two distinct phases: large and almost immobile spirals and clusters of trapped small vortices. When the "liquid fraction" is small, the resulting pattern exhibits slow intermittent dynamics: bursts of activity separated by long quiescent intervals. The system keeps evolving on an extremely slow scale, which is consistent with exponentially weak repulsion between well separated spiral cores.

Another possibility, realized in a different parametric region, is a dynamic chaotic state that shows no persistent features. This state is attained under conditions when either spiral waves or vortex cores, or both, are unstable. One can distinguish between mild phase turbulence when no phase singularities occur, and defect chaos characterized by persistent creation and annihilation of vortex pairs. Phase turbulence may persist in the parametric region between the Benjamin–Feir line and the line L in Fig. 13 [59]. Beyond the line L, defects are created spontaneously, leading to defect chaos. Transition from vortex glass to defect turbulence in simulations starting from random initial conditions occurs at the numerically determined line T in Fig. 13 [59]. The transition occurs somewhat prior to the absolute instability limit determined by the linear stability analysis of plane waves emitted by spirals. This limit can be approached, however, by starting from carefully prepared initial conditions in the form of large spirals. Prior to the transition, one can observe transient defect turbulence which is unstable to spontaneous nucleation of spirals from the "turbulent sea", leading eventually to a vortex glass state.

Forced Systems

External forcing, including spatially as well as temporally variable inputs, can be used in a straightforward

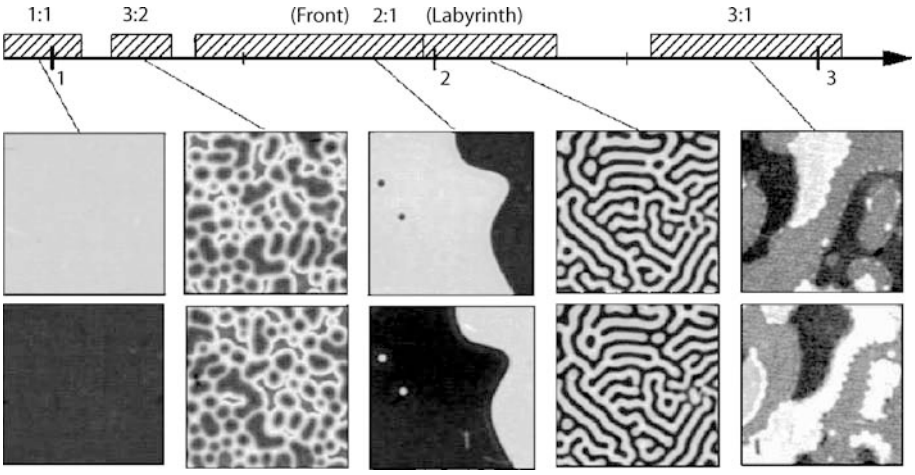

Patterns and Interfaces in Dissipative Dynamics, Figure 19
Frequency-locked regimes (experiment with light-sensitive reaction under periodic optical forcing [71], reproduced with permission). The *axis above* shows the ratio of the forcing to basic frequency. Patterns are shown in pairs, one above the other, at times separated by the forcing period $2\pi/\omega_c$, except for the 1 : 1 resonance where the interval is π/ω_c

way to enhance or suppress spontaneously emerging patterns [69]. Alternatively, it may enhance complexity by introducing additional spatial and temporal resonances, which may lead to formation of quasicrystalline structures [70]. Resonant forcing of oscillatory systems may drastically change the structure of wave patterns through phase locking. This happens when the CGL equation is forced on a frequency ω_c commensurate with the basic frequency ω_0 at the Hopf bifurcation. For an integer ratio $\omega_c/\omega_0 = n$, the amplitude equation amending (23) can be written by adding the forcing term possessing the required symmetry:

$$\partial_t u = (1+\mathrm{i}\eta)\nabla^2 u + (\mu+\mathrm{i}\omega)u - (1+\mathrm{i}\nu)|u|^2 u + \gamma \overline{u}^{n-1}\,, \tag{32}$$

where γ is the forcing amplitude and $\epsilon^2\omega$ is weak effective detuning, due to both parametric deviations from the Hopf bifurcation point and weak mismatch between ω_c/n and ω_0. The forcing term breaks the symmetry of the CGL equation to phase rotations, reducing it to discrete symmetry $u \to \mathrm{e}^{\mathrm{i}\pi m}u$, $m = 1, \ldots, n-1$. This changes the character of defects: instead of vortices, one can observed fronts separating alternative phase states.

Various patterns at different forcing frequencies, which can be modeled by (32), were observed both in experiments and simulations [71,72]. Some typical patterns are shown in Fig. 19. For the case of strong resonance ($n = 1$), this system provides a convenient tool for studying transitions between stationary and propagating fronts [73], labyrinthine patterns [74], and solitary structures [75]. These structures are not unlike those observed in the FitzHugh–Nagumo system, although they represent standing waves with the alternative phases interchanging within each domain. Higher resonances create still more complex dynamics involving interactions of different kinds of fronts [76].

Future Directions

The study of pattern formation is now a mature discipline based on well-established general theory and wealth of experimental evidence. The center of attention is turning to specific applications; among them, nonlinear optics and studies of granular media come to the forefront. Forcing and control of patterns, either enhancing or suppressing the complexity of behavior, are studied in detail. As a humble laptop turns into a supercomputer, more fascinating patterns, envy of abstract expressionists, are generated by model equations of increased complexity. Patterns showing dazzling mix of order and chaos are seen as well in various experimental setups. The ultimate aim of controlled creation of self-organized structures still remains elusive, and new ideas are awaited as the new century comes of age. The study of pattern formation, dealing with ubiquitous problems of order and chaos, is bound to find its way into basic curricula and wealth of practical applications.

Bibliography

Primary Literature

1. Faraday M (1831) Philos Trans R Soc Lond 121:299
2. Bénard H (1900) Ann Chim Phys 7 (Ser 23) 62

3. Turing AM (1952) Philos Trans R Soc Lond Ser B 237:37
4. Hamley IW (2003), Nanotechnology 14:R39
5. Burger M, Field RJ (1985) Oscillations and Traveling Waves in Chemical Systems. John Wiley, New York
6. Ouyang Q, Swinney HL (1991) Nature 352:610
7. Gierer A, Meinhard H (1972) Kybernetik 12:30
8. Murray JD (1981) J Theor Biol 88:161
9. Gilad E, Shachak M, Meron E (2007) Theor Popul Biol 72:214
10. Zeldovich YB (1985) The Mathematical Theory of Combustion and Explosions. Consultants Bureau, New York
11. Langer JS (1980) Rev Mod Phys 52:1
12. Arecchi FT (1999) Phys Rep 318:1
13. Josserand C, Pomeau Y, Rica S (2007) Eur Phys J Special Topics 146:47
14. Alexander S, McTague J (1978) Phys Rev Lett 41:702
15. Proctor MRE, Jones CA (1988) J Fluid Mech 188:301
16. Yang L, Dolnik M, Zhabotinsky AM, Epstein IR (2002) Phys Rev Lett 88:208303
17. Pampaloni E, Residori S, Soria S, Arecchi FT (1997) Phys Rev Lett 78:1042
18. Pismen LM, Rubinstein BY (1999) Chaos Soliton Fractal 10:761
19. Bowman C, Newell AC (1998) Rev Mod Phys 70:289
20. Newell AC, Whitehead JA (1969) J Fluid Mech 38:279
21. Segel LA (1969) J Fluid Mech 38:203
22. Pomeau Y, Manneville P (1979) J Phys Lett 40:L609
23. Cross MC, Newell AC (1984) Physica D 10:299
24. Mermin ND (1979) Rev Mod Phys 51:591
25. Bodenschatz E, Pesch W, Kramer L (1988) Physica D 32:135
26. Pismen LM, Rodriguez JD (1990) Phys Rev A 42:2471
27. Braun E, Steinberg V (1991) Europhys Lett 15:167
28. Nepomnyashchy AA, Pismen LM (1991) Phys Lett A 153:427
29. Newell AC, Passot T, Bowman C, Ercolani N, Indik R (1996) Physica D 97:185
30. Bodenschatz E, Pesch W, Ahlers G (2000) Annu Rev Fluid Mech 32:709
31. Rabinovich MI, Tsimring LS (1994) Phys Rev E 49:R35
32. Abou B, Wesfreid JE, Roux S (2000) J Fluid Mech 416:217
33. Assenheimer M, Steinberg V (1993) Phys Rev Lett 70:3888
34. van der Waals JD (1894) Z Phys Chem 13:657
35. Cahn JW, Hilliard JE (1958) J Chem Phys 28:258
36. Sagués F, Sancho JM, García-Ojalvo J (2007) Spatiotemporal order out of noise. Rev Mod Phys 79:829
37. Kolmogorov A, Petrovsky I, Piskunov N (1937) Bull Univ Moscow, Ser Int Sec A 1:1
38. van Saarloos W (2003) Phys Rep 386:29
39. Mullins WW, Sekerka RF (1963) J Appl Phys 34:323
40. Sivashinsky GI (1977) Acta Astronaut 4:1177
41. Kuramoto Y, Tsuzuki T (1976) Prog Theor Phys 55:356
42. Lifshitz IM, Slyozov VV (1958) Zh Eksp Teor Fiz 35:479
43. Ohta T, Mimura M, Kobayashi R (1989) Physica D 34:115
44. Coullet P, Riera C, Tresser C (2004) Chaos 14:193
45. Reynolds WN, Pearson JE, Ponce-Dawson S (1994) Phys Rev Lett 72:2797
46. Or-Guil M, Bode M, Schenk CP, Purwins H-G (1998) Phys Rev E 57:6432
47. Elphick C, Meron E, Spiegel EA (1988) Phys Rev Lett 61:496
48. Malomed BA, Nepomnyashchy AA, Tribelsky MI (1990) Phys Rev A 42:7244
49. Ben-Jacob E, Brand HR, Dee G, Kramer L, Langer JS (1985) Physica D 14:348
50. Pomeau Y (1986) Physica D 23:3
51. Aranson IS, Malomed BA, Pismen LM, Tsimring LS (2000) Phys Rev E 62:R5
52. Hagberg A, Yochelis A, Yizhak H, Elphick C, Pismen LM, Meron E (2006) Physica D 217:186
53. Aranson IS, Kramer L (2002) Rev Mod Phys 74:99
54. van Saarloos W, Hohenberg PC (1992) Physica D 56:303
55. Brusch L, Torcini A, van Hecke M, Zimmermann MG, Bär M (2001) Physica D 160:127
56. Knobloch E, de Luca J (1990) Nonlinearity 3:975
57. Pismen LM (1986) Dyn Stab Syst 1:97
58. Hagan PS (1982) SIAM J Appl Math 42:762
59. Chaté H, Manneville P (1996) Physica A 224:348
60. Tyson JJ, Keener JP (1988) Physica D 32:327
61. Hakim V, Karma A (1999) Phys Rev E 60:5073
62. Barkley D (1994) Phys Rev Lett 72:164
63. Park J-S, Lee KJ (2002) Phys Rev Lett 88:224501
64. Keener JP (1988) Physica D 31:269
65. Henry H, Hakim V (2002) Phys Rev E 65:046235
66. Rousseau G, Chaté H, Kapral R (1998) Phys Rev Lett 80:5671
67. Alonso S, Kähler R, Mikhailov AS, Sagués F (2004) Phys Rev E 70:056201
68. Brito C, Aranson IS, Chaté H (2003) Phys Rev Lett 90:068301
69. Nepomnyashchy AA, Golovin AA, Gubareva V, Panfilov V (2004) Physica D 199:61
70. Pismen LM (1987) Phys Rev Lett 59:2740
71. Petrov V, Ouyang Q, Swinney HL (1997) Nature 388:655
72. Lin AL, Hagberg A, Meron E, Swinney HL (2004) Phys Rev E 69:066217
73. Coullet P, Emilsson K (1992) Physica D 61:119
74. Yochelis A, Elphick C, Hagberg A, Meron E (2004) Physica D 199:201
75. Gomila D, Colet P, San Miguel M, Oppo G-L (2007) Eur Phys J Special Topics 146:71
76. Gallego R, Walgraef D, San Miguel M, Toral R (2001) Phys Rev E 64:056218

Books and Reviews

Cross MC, Hohenberg P (1993) Rev Mod Phys 65:851

Epstein IR, Pojman JA (1998) Introduction to Nonlinear Chemical Dynamics. Oxford University Press, New York

Fife PC (1979) Mathematical Aspects of Reacting and Diffusing Systems. Springer, Berlin

Haken H (2004) Synergetics: Introduction and Advanced Topics. Springer, Berlin

Hoyle R (2006) Pattern Formation. Cambridge UP, Cambridge

Kuramoto Y (1984) Chemical Oscillations, Waves and Turbulence. Springer, Berlin

Manneville P (1990) Dissipative Structures and Weak Turbulence. Academic Press, San Diego

Mikhailov AS (1991) Foundations of Synergetics I: Distributed Active Systems II (with AY Loskutov): Complex Patterns. Springer, Berlin

Murray JD (1989) Mathematical Biology. Springer, Berlin; 2nd edn 1993; 3rd edn 2002/2003

Pismen LM (1999) Vortices in Nonlinear Fields. Clarendon Press, Oxford

Pismen LM (2006) Patterns and Interfaces in Dissipative Dynamics. Springer, Berlin

Rabinovich MI, Ezersky AB, Weidman PD (2000) The Dynamics of Patterns. World Scientific, Singapore

Walgraef D (1997) Spatio-Temporal Pattern Formation. Springer, New York
Winfree AT (1987) When Time Breaks Down. Princeton University Press, Princeton

Pedestrian, Crowd and Evacuation Dynamics

Dirk Helbing[1,2], Anders Johansson[1]
[1] ETH Zurich, Zurich, Switzerland
[2] Institute for Advanced Study, Collegium Budapest, Budapest, Hungary

Article Outline

Glossary
Definition of the Subject
Introduction
Pedestrian Dynamics
Crowd Dynamics
Evacuation Dynamics
Future Directions
Acknowledgments
Bibliography

Glossary

Collective intelligence Emergent functional behavior of a large number of people that results from *interactions* of individuals rather than from individual reasoning or global optimization.

Crowd Agglomeration of many people in the same area at the same time. The density of the crowd is assumed to be high enough to cause continuous interactions with or reactions to other individuals.

Crowd turbulence Unanticipated and unintended irregular motion of individuals into different directions due to strong and rapidly changing forces in crowds of extreme density.

Emergence Spontaneous establishment of a qualitatively new behavior through non-linear interactions of many objects or subjects.

Evolutionary optimization Gradual optimization based on the effect of frequently repeated random mutations and selection processes based on some success function ("fitness").

Faster-is-slower effect This term reflects the observation that certain processes (in evacuation situations, production, traffic dynamics, or logistics) take more time if performed at high speed. In other words, waiting can often help to coordinate the activities of several competing units and to speed up the average progress.

Freezing-by-heating effect Noise-induced blockage effect caused by the breakdown of direction-segregated walking patterns (typically two or more "lanes" characterized by a uniform direction of motion). "Noise" means frequent variations of the walking direction due to nervousness or impatience in the crowd, e. g. also frequent overtaking maneuvers in dense, slowly moving crowds.

Panic Breakdown of ordered, cooperative behavior of individuals due to anxious reactions to a certain event. Often, panic is characterized by attempted escape of many individuals from a real or perceived threat in situations of a perceived struggle for survival, which may end up in trampling or crushing of people in a crowd.

Self-organization Spontaneous organization (i. e. formation of ordered patterns) not induced by initial or boundary conditions, by regulations or constraints. Self-organization is a result of non-linear interactions between many objects or subjects, and it often causes different kinds of spatio-temporal patterns of motion.

Social force Vector describing acceleration or deceleration effects that are caused by social interactions rather than by physical interactions or fields.

Definition of the Subject

The modeling of pedestrian motion is of great theoretical and practical interest. Recent experimental efforts have revealed quantitative details of pedestrian interactions, which have been successfully cast into mathematical equations. Furthermore, corresponding computer simulations of large numbers of pedestrians have been compared with the empirically observed dynamics of crowds. Such studies have led to a deeper understanding of how collective behavior on a macroscopic scale emerges from individual human interactions. Interestingly enough, the non-linear interactions of pedestrians lead to various complex, spatio-temporal pattern-formation phenomena. This includes the emergence of lanes of uniform walking direction, oscillations of the pedestrian flow at bottlenecks, and the formation of stripes in two intersecting flows. Such self-organized patterns of motion demonstrate that efficient, "intelligent" collective dynamics can be based on simple, local interactions. Under extreme conditions, however, coordination may break down, giving rise to critical crowd conditions. Examples are "freezing-by-heating" and "faster-is-slower" effects, but also the transition to "turbulent" crowd dynamics. These observations have im-

portant implications for the optimization of pedestrian facilities, in particular for evacuation situations.

Introduction

The emergence of new, functional or complex collective behaviors in social systems has fascinated many scientists. One of the primary questions in this field is how cooperation or coordination patterns originate based on elementary individual interactions. While one could think that these are a result of intelligent human actions, it turns out that much simpler models assuming automatic responses can reproduce the observations very well. This suggests that humans are using their intelligence primarily for more complicated tasks, but also that simple interactions can lead to intelligent patterns of motion. Of course, it is reasonable to assume that these interactions are the result of a previous learning process that has optimized the automatic response in terms of minimizing collisions and delays. This, however, seems to be sufficient to explain most observations.

In this contribution, we will start with a short history of pedestrian modeling and, then, introduce a simplified model of pedestrian interactions, the "social force model". Furthermore, we will discuss its calibration using video tracking data. Next, we will turn to the subject of crowd dynamics, as one typically finds the formation of large-scale spatio-temporal patterns of motion, when many pedestrians interact with each other. These patterns will be discussed in some detail before we will turn to evacuation situations and cases of extreme densities, where one can sometimes observe the breakdown of coordination. Finally, we will address possibilities to design improved pedestrian facilities, using special evolutionary algorithms.

Pedestrian Dynamics

Short History of Pedestrian Modeling

Pedestrians have been empirically studied for more than four decades [1,2,3]. The evaluation methods initially applied were based on direct observation, photographs, and time-lapse films. For a long time, the main goal of these studies was to develop a *level-of-service concept* [4], *design elements* of pedestrian facilities [5,6,7,8], or *planning guidelines* [9,10]. The latter have usually the form of *regression relations*, which are, however, not very well suited for the prediction of pedestrian flows in pedestrian zones and buildings with an exceptional architecture, or in challenging evacuation situations. Therefore, a number of simulation models have been proposed, e. g. *queuing models* [11], *transition matrix models* [12], and *stochastic models* [13], which are partly related to each other. In addition, there are models for the *route choice behavior* of pedestrians [14,15].

None of these concepts adequately takes into account the self-organization effects occurring in pedestrian crowds. These are the subject of recent experimental studies [8,16,17,18,19,20]. Most pedestrian models, however, were formulated before. A first modeling approach that appears to be suited to reproduce spatio-temporal patterns of motion was proposed by Henderson [21], who conjectured that pedestrian crowds behave similar to gases or fluids (see also [22]). This could be partially confirmed, but a realistic gas-kinetic or fluid-dynamic theory for pedestrians must contain corrections due to their particular interactions (i. e. avoidance and deceleration maneuvers) which, of course, do not obey momentum and energy conservation. Although such a theory can be actually formulated [23,24], for practical applications a direct simulation of *individual* pedestrian motion is favorable, since this is more flexible. As a consequence, pedestrian research mainly focuses on *agent-based models* of pedestrian crowds, which also allow one to consider local coordination problems. The "social force model" [25,26] is maybe the most well-known of these models, but we also like to mention *cellular automata* of pedestrian dynamics [27,28,29,30,31,32,33] and *AI-based models* [34,35].

The Social Force Concept

In the following, we shall shortly introduce the social force concept, which reproduces most empirical observations in a simple and natural way. Human behavior often seems to be "chaotic", irregular, and unpredictable. So, why and under what conditions can we model it by means of forces? First of all, we need to be confronted with a phenomenon of motion in some (quasi-)continuous space, which may be also an abstract behavioral space such as an opinion scale [36]. Moreover, it is favorable to have a system where the fluctuations due to unknown influences are not large compared to the systematic, deterministic part of motion. This is usually the case in pedestrian traffic, where people are confronted with standard situations and react "automatically" rather than taking complicated decisions, e. g. if they have to evade others.

This "automatic" behavior can be interpreted as the result of a *learning process* based on trial and error [37], which can be simulated with *evolutionary algorithms* [38]. For example, pedestrians have a preferred side of walking, since an asymmetrical avoidance behavior turns out to be profitable [25,37]. The related *formation of a behav-*

ioral convention can be described by means of *evolutionary game theory* [25,39].

Another requirement is the vectorial additivity of the separate force terms reflecting different environmental influences. This is probably an approximation, but there is some experimental evidence for it. Based on quantitative measurements for animals and test persons subject to separately or simultaneously applied stimuli of different nature and strength, one could show that the behavior in conflict situations can be described by a superposition of forces [40,41]. This fits well into a concept by Lewin [42], according to which behavioral changes are guided by so-called *social fields* or *social forces*, which has later on been put into mathematical terms [25,43]. In some cases, social forces, which determine the amount and direction of systematic behavioral changes, can be expressed as gradients of dynamically varying potentials, which reflect the social or behavioral fields resulting from the interactions of individuals. Such a social force concept was applied to opinion formation and migration [43], and it was particularly successful in the description of collective pedestrian behavior [8,25,26,37].

For reliable simulations of pedestrian crowds, we do not need to know whether a certain pedestrian, say, turns to the right at the next intersection. It is sufficient to have a good estimate what percentage of pedestrians turns to the right. This can be either empirically measured or estimated by means of route choice models [14]. In some sense, the uncertainty about the individual behaviors is averaged out at the macroscopic level of description. Nevertheless, we will use the more flexible microscopic simulation approach based on the social force concept. According to this, the temporal change of the location $\boldsymbol{r}_\alpha(t)$ of pedestrian α obeys the equation of motion

$$\frac{\mathrm{d}\boldsymbol{r}_\alpha(t)}{\mathrm{d}t} = \boldsymbol{v}_\alpha(t)\,. \tag{1}$$

Moreover, if $\boldsymbol{f}_\alpha(t)$ denotes the sum of social forces influencing pedestrian α and if $\boldsymbol{\xi}_\alpha(t)$ are individual fluctuations reflecting unsystematic behavioral variations, the velocity changes are given by the *acceleration equation*

$$\frac{\mathrm{d}\boldsymbol{v}_\alpha}{\mathrm{d}t} = \boldsymbol{f}_\alpha(t) + \boldsymbol{\xi}_\alpha(t)\,. \tag{2}$$

A particular advantage of this approach is that we can take into account the flexible usage of space by pedestrians, requiring a continuous treatment of motion. It turns out that this point is essential to reproduce the empirical observations in a natural and robust way, i. e. without having to adjust the model to each single situation and measurement site. Furthermore, it is interesting to note that, if the fluctuation term is neglected, the social force model can be interpreted as a particular *differential game*, i. e. its dynamics can be derived from the minimization of a special utility function [44].

Specification of the Social Force Model

The social force model for pedestrians assumes that each individual α is trying to move in a desired direction $\boldsymbol{e}^0_\alpha$ with a desired speed v^0_α, and that it adapts the actual velocity $\boldsymbol{v}_\alpha$ to the desired one, $\boldsymbol{v}^0_\alpha = v^0_\alpha \boldsymbol{e}^0_\alpha$, within a certain relaxation time τ_α. The systematic part $\boldsymbol{f}_\alpha(t)$ of the acceleration force of pedestrian α is then given by

$$\boldsymbol{f}_\alpha(t) = \frac{1}{\tau_\alpha}(v^0_\alpha \boldsymbol{e}^0_\alpha - \boldsymbol{v}_\alpha) + \sum_{\beta(\neq\alpha)} \boldsymbol{f}_{\alpha\beta}(t) + \sum_i \boldsymbol{f}_{\alpha i}(t)\,, \tag{3}$$

where the terms $\boldsymbol{f}_{\alpha\beta}(t)$ and $\boldsymbol{f}_{\alpha i}(t)$ denote the repulsive forces describing attempts to keep a certain safety distance to other pedestrians β and obstacles i. In very crowded situations, additional physical contact forces come into play (see Subsect.“ Force Model for Panicking Pedestrians”). Further forces may be added to reflect attraction effects between members of a group or other influences. For details see [37].

First, we will assume a simplified interaction force of the form

$$\boldsymbol{f}_{\alpha\beta}(t) = \boldsymbol{f}\big(\boldsymbol{d}_{\alpha\beta}(t)\big)\,, \tag{4}$$

where $\boldsymbol{d}_{\alpha\beta} = \boldsymbol{r}_\alpha - \boldsymbol{r}_\beta$ is the distance vector pointing from pedestrian β to α. Angular-dependent shielding effects may be furthermore taken into account by a prefactor describing the anisotropic reaction to situations in front of as compared to behind a pedestrian [26,45], see Subsect. “Angular Dependence”. However, we will start with a **circular specification** of the distance-dependent interaction force,

$$\boldsymbol{f}(\boldsymbol{d}_{\alpha\beta}) = A_\alpha \mathrm{e}^{-d_{\alpha\beta}/B_\alpha} \frac{\boldsymbol{d}_{\alpha\beta}}{\|\boldsymbol{d}_{\alpha\beta}\|}\,, \tag{5}$$

where $d_{\alpha\beta} = \|\boldsymbol{d}_{\alpha\beta}\|$ is the distance. The parameter A_α reflects the *interaction strength*, and B_α corresponds to the *interaction range*. While the dependence on α explicitly allows for a dependence of these parameters on the single individual, we will assume a homogeneous population, i. e. $A_\alpha = A$ and $B_\alpha = B$ in the following. Otherwise, it would be hard to collect enough data for parameter calibration.

Elliptical Specification Note that it is possible to express Eq. (5) as gradient of an exponentially decaying potential

$V_{\alpha\beta}$. This circumstance can be used to formulate a generalized, elliptical interaction force via the potential

$$V_{\alpha\beta}(b_{\alpha\beta}) = AB\mathrm{e}^{-b_{\alpha\beta}/B} , \tag{6}$$

where the variable $b_{\alpha\beta}$ denotes the semi-minor axis $b_{\alpha\beta}$ of the elliptical equipotential lines. This has been specified according to

$$2b_{\alpha\beta} = \sqrt{\begin{array}{l}(\|\boldsymbol{d}_{\alpha\beta}\| + \|\boldsymbol{d}_{\alpha\beta} - (\boldsymbol{v}_\beta - \boldsymbol{v}_\alpha)\Delta t\|)^2 \\ -\|(\boldsymbol{v}_\beta - \boldsymbol{v}_\alpha)\Delta t\|^2 ,\end{array}} \tag{7}$$

so that both pedestrians α and β are treated symmetrically. The repulsive force is related to the above potential via

$$\begin{aligned} \boldsymbol{f}_{\alpha\beta}(\boldsymbol{d}_{\alpha\beta}) &= -\nabla_{\boldsymbol{d}_{\alpha\beta}} V_{\alpha\beta}(b_{\alpha\beta}) \\ &= -\frac{\mathrm{d}V_{\alpha\beta}(b_{\alpha\beta})}{\mathrm{d}b_{\alpha\beta}} \nabla_{\boldsymbol{d}_{\alpha\beta}} b_{\alpha\beta}(\boldsymbol{d}_{\alpha\beta}) , \end{aligned} \tag{8}$$

where $\nabla_{\boldsymbol{d}_{\alpha\beta}}$ represents the gradient with respect to $\boldsymbol{d}_{\alpha\beta}$. Considering the chain rule, $\|\boldsymbol{z}\| = \sqrt{\boldsymbol{z}^2}$, and $\nabla_{\boldsymbol{z}}\|\boldsymbol{z}\| = \boldsymbol{z}/\sqrt{\boldsymbol{z}^2} = \boldsymbol{z}/\|\boldsymbol{z}\|$, this leads to the explicit formula

$$\begin{aligned} \boldsymbol{f}_{\alpha\beta}(\boldsymbol{d}_{\alpha\beta}) = A\mathrm{e}^{-b_{\alpha\beta}/B} \cdot \frac{\|\boldsymbol{d}_{\alpha\beta}\| + \|\boldsymbol{d}_{\alpha\beta} - \boldsymbol{y}_{\alpha\beta}\|}{2b_{\alpha\beta}} \\ \cdot \frac{1}{2}\left(\frac{\boldsymbol{d}_{\alpha\beta}}{\|\boldsymbol{d}_{\alpha\beta}\|} + \frac{\boldsymbol{d}_{\alpha\beta} - \boldsymbol{y}_{\alpha\beta}}{\|\boldsymbol{d}_{\alpha\beta} - \boldsymbol{y}_{\alpha\beta}\|}\right) \end{aligned} \tag{9}$$

with $\boldsymbol{y}_{\alpha\beta} = (\boldsymbol{v}_\beta - \boldsymbol{v}_\alpha)\Delta t$. We used $\Delta t = 0.5$ s. For $\Delta t = 0$, we regain the expression of Eq. (5).

The elliptical specification has two major advantages compared to the circular one: First, the interactions depend not only on the distance, but also on the relative velocity. Second, the repulsive force is not strictly directed from pedestrian β to pedestrian α, but has a lateral component. As a consequence, this leads to less confrontative, smoother ("sliding") evading maneuvers. Note that further velocity-dependent specifications of pedestrian interaction forces have been proposed [7,26], but we will restrict to the above specifications, as these are sufficient to demonstrate the method of evolutionary model calibration.

Evolutionary Calibration with Video Tracking Data

For parameter calibration, several video recordings of pedestrian crowds in different natural environments have been used. The dimensions of the recorded areas were known, and the floor tiling or environment provided something like a "coordinate system". The heads were automatically determined by searching for round moving structures, and the accuracy of tracking was improved by comparing actual with linearly extrapolated positions (so it would not happen so easily that the algorithm interchanged or "lost" close by pedestrians). The trajectories of the heads were then projected on two-dimensional space in a way correcting for distortion by the camera perspective. A representative plot of the resulting trajectories is shown in Fig. 1. Note that trajectory data have been obtained with infra-red sensors [47] or video cameras [48,49] for several years now, but algorithms that can simultaneously handle more than one thousand pedestrians have become available only recently [87].

For model calibration, it is recommended to use a hybrid method fusing empirical trajectory data and microscopic simulation data of pedestrian movement in space. In corresponding algorithms, a virtual pedestrian is assigned to each tracked pedestrian in the simulation domain. One then starts a simulation for a time period T (e. g. 1.5 s), in which one pedestrian α is moved according to a simulation of the social force model, while the others are moved exactly according to the trajectories ex-

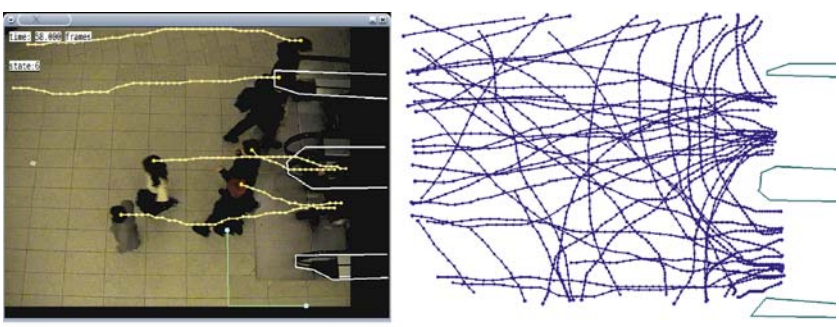

a b

Pedestrian, Crowd and Evacuation Dynamics, Figure 1
Video tracking used to extract the trajectories of pedestrians from video recordings close to two escalators (after [45]). a Illustration of the tracking of pedestrian heads. b Resulting trajectories after being transformed onto the two-dimensional plane

tracted from the videos. This procedure is performed for all pedestrians α and for several different starting times t, using a fixed parameter set for the social force model.

Each simulation run is performed according to the following scheme:

1. Define a starting point and calculate the state (position $\boldsymbol{r}_\alpha$, velocity $\boldsymbol{v}_\alpha$, and acceleration $\boldsymbol{a}_\alpha = \mathrm{d}\boldsymbol{v}_\alpha/\mathrm{d}t$) for each pedestrian α.
2. Assign a desired speed v_α^0 to each pedestrian, e. g. the maximum speed during the pedestrian tracking time. This is sufficiently accurate, if the overall pedestrian density is not too high and the desired speed is constant in time.
3. Assign a desired goal point for each pedestrian, e. g. the end point of the trajectory.
4. Given the tracked motion of the surrounding pedestrians β, simulate the trajectory of pedestrian α over a time period T based on the social force model, starting at the actual location $\boldsymbol{r}_\alpha(t)$.

After each simulation run, one determines the relative distance error

$$\frac{\|\boldsymbol{r}_\alpha^{\text{simulated}}(t+T) - \boldsymbol{r}_\alpha^{\text{tracked}}(t+T)\|}{\|\boldsymbol{r}_\alpha^{\text{tracked}}(t+T) - \boldsymbol{r}_\alpha^{\text{tracked}}(t)\|} \,. \tag{10}$$

After averaging the relative distance errors over the pedestrians α and starting times t, 1 minus the result can be taken as measure of the goodness of fit (the "fitness") of the parameter set used in the pedestrian simulation. Hence, the best possible value of the "fitness" is 1, but any deviation from the real pedestrian trajectories implies lower values.

One result of such a parameter optimization is that, for each video, there is a broad range of parameter combinations of A and B which perform almost equally well [45]. This allows one to apply additional goal functions in the parameter optimization, e. g. to determine among the best performing parameter values such parameter combinations, which perform well for *several* video recordings, using a fitness function which equally weights the fitness reached in each single video. This is how the parameter values listed in Table 1 were determined. It turns out that, in order to reach a good model performance, the pedestrian interaction force must be specified velocity dependent, as in the elliptical model.

Note that our evolutionary fitting method can be also used to determine interaction laws without prespecified interaction functions. For example, one can obtain the distance dependence of pedestrian interactions without a prespecified function. For this, one adjusts the values of the force at given distances $d_k = kd_1$ (with $k \in \{1, 2, 3, \ldots\}$) in an evolutionary way. To get some smoothness, linear interpolation is applied. The resulting fit curve is presented in Fig. 2 (left). It turns out that the empirical dependence of the force with distance can be well fitted by an exponential decay.

Pedestrian, Crowd and Evacuation Dynamics, Table 1
Interaction strength *A* and interaction range *B* resulting from our evolutionary parameter calibration for the circular and elliptical specification of the interaction forces between pedestrians (see main text). The calibration was based on three different video recordings, one for low crowd density, one for medium, and one for high density. The parameter values are specified as mean value ± standard deviation. The best fitness value obtained with the elliptical specification for the video with the lowest crowd density was as high as 0.9

Model	A	B	"Fitness"
Extrapolation	0	–	0.34
Circular	0.11 ± 0.06	0.84 ± 0.63	0.35
Elliptical	4.30 ± 3.91	1.07 ± 1.35	0.53

Angular Dependence

A closer study of pedestrian interactions reveals that these are not isotropic, but dependent on the angle $\varphi_{\alpha\beta}$ of the encounter, which is given by the formula

$$\cos(\varphi_{\alpha\beta}) = \frac{\boldsymbol{v}_\alpha}{\|\boldsymbol{v}_\alpha\|} \cdot \frac{-\boldsymbol{d}_{\alpha\beta}}{\|\boldsymbol{d}_{\alpha\beta}\|} \,. \tag{11}$$

Generally, pedestrians show little response to pedestrians behind them. This can be reflected by an angular-dependent prefactor $w(\varphi_{\alpha\beta})$ of the interaction force [45]. Empirical results are represented in Fig. 2 (right). Reasonable results are obtained for the following specification of the prefactor:

$$w(\varphi_{\alpha\beta}(t)) = \left(\lambda_\alpha + (1-\lambda_\alpha)\frac{1+\cos(\varphi_{\alpha\beta})}{2}\right) \,, \tag{12}$$

where λ_α with $0 \le \lambda_\alpha \le 1$ is a parameter which grows with the strength of interactions from behind. An evolutionary parameter optimization gives values $\lambda \approx 0.1$ [45], i. e. a strong anisotropy. With such an angle-dependent prefactor, the "fitness" of the elliptical force increases from 0.53 to 0.61, when calibrated to the same set of videos. Other angular-dependent specifications split up the interaction force between pedestrians into a component against the direction of motion and another one perpendicular to it. Such a description allows for even smoother avoidance maneuvers.

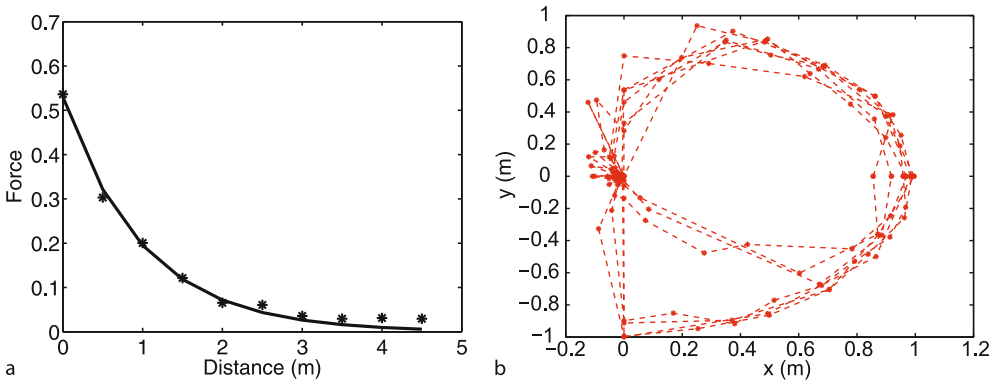

Pedestrian, Crowd and Evacuation Dynamics, Figure 2
Results of an evolutionary fitting of pedestrian interactions. a Empirically determined distance dependence of the interaction force between pedestrians (after [45]). An exponential decay fits the empirical data quite well. The dashed fit curve corresponds to Eq. (5) with the parameters $A = 0.53$ and $B = 1.0$. b Angular dependence of the influence of other pedestrians. The direction along the positive *x*-axis corresponds to the walking direction of pedestrians, *y* to the perpendicular direction

Crowd Dynamics

Analogies with Gases, Fluids, and Granular Media

When the density is low, pedestrians can move freely, and the observed crowd dynamics can be partially compared with the behavior of gases. At medium and high densities, however, the motion of pedestrian crowds shows some striking analogies with the motion of fluids:

1. Footprints of pedestrians in snow look similar to streamlines of fluids [15].
2. At borderlines between opposite directions of walking one can observe "viscous fingering" [50,51].
3. The emergence of pedestrian streams through standing crowds [7,37,52] appears analogous to the formation of river beds [53,54].

At high densities, however, the observations have rather analogies with driven granular flows. This will be elaborated in more detail in Sects. "Force Model for Panicking Pedestrians" and "Collective Phenomena in Panic Situations". In summary, one could say that fluid-dynamic analogies work reasonably well in normal situations, while granular aspects dominate at extreme densities. Nevertheless, the analogy is limited, since the self-driven motion and the violation of momentum conservation imply special properties of pedestrian flows. For example, one usually does not observe eddies, which typically occur in regular fluids at high enough Reynolds numbers.

Self-Organization of Pedestrian Crowds

Despite its simplifications, the social force model of pedestrian dynamics describes a lot of observed phenomena quite realistically. Especially, it allows one to explain various self-organized spatio-temporal patterns that are not externally planned, prescribed, or organized, e. g. by traffic signs, laws, or behavioral conventions. Instead, the spatio-temporal patterns discussed below emerge due to the non-linear interactions of pedestrians even without assuming strategical considerations, communication, or imitative behavior of pedestrians. Despite this, we may still interpret the forming cooperation patterns as phenomena that establish social order on short time scales. It is actually surprising that strangers coordinate with each other within seconds, if they have grown up in a similar environment. People from different countries, however, are sometimes irritated about local walking habits, which indicates that learning effects and cultural backgrounds still play a role in social interactions as simple as random pedestrian encounters. Rather than on particular features, however, in the following we will focus on the common, internationally reproducible observations.

Lane Formation In pedestrian flows one can often observe that oppositely moving pedestrians are forming lanes of uniform walking direction (see Fig. 3). This phenomenon even occurs when there is not a large distance to separate each other, e. g. on zebra crossings. However, the width of lanes increases (and their number decreases), if the interaction continues over longer distances (and if perturbations, e. g. by flows entering or leaving on the sides, are low; otherwise the phenomenon of lane formation may break down [55]).

Lane formation may be viewed as *segregation phenomenon* [56,57]. Although there is a weak preference for one side (with the corresponding behavioral convention

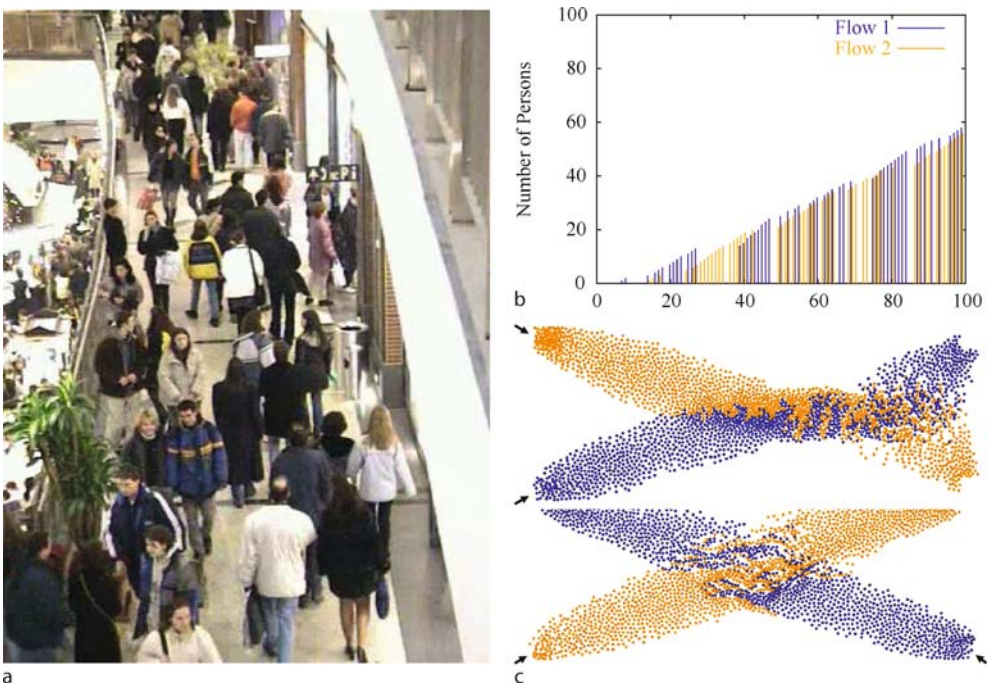

Pedestrian, Crowd and Evacuation Dynamics, Figure 3

Self-organization of pedestrian crowds. a Photograph of lanes formed in a shopping center. Computer simulations reproduce the self-organization of such lanes very well. b Evaluation of the cumulative number of pedestrians passing a bottleneck from different sides. One can clearly see that the narrowing is often passed by groups of people in an oscillatory way rather than one by one. c Multi-agent simulation of two crossing pedestrian streams, showing the phenomenon of stripe formation. This self-organized pattern allows pedestrians to pass the other stream without having to stop, namely by moving sidewards in a forwardly moving stripe. (After [8])

depending on the country), the observations can only be well reproduced when repulsive pedestrian interactions are taken into account. The most relevant factor for the lane formation phenomenon is the higher relative velocity of pedestrians walking in opposite directions. Compared to people following each other, oppositely moving pedestrians have more frequent interactions until they have segregated into separate lanes by stepping aside whenever another pedestrian is encountered. The most long-lived patterns of motion are the ones which change the least. It is obvious that such patterns correspond to lanes, as they minimize the frequency and strength of avoidance maneuvers. Interestingly enough, as computer simulations show, lane formation occurs also when there is no preference for any side.

Lanes minimize frictional effects, accelerations, energy consumption, and delays in oppositely moving crowds. Therefore, one could say that they are a pattern reflecting "collective intelligence". In fact, it is not possible for a single pedestrian to reach such a collective pattern of motion. Lane formation is a self-organized collaborative pattern of motion originating from simple pedestrian interactions. Particularly in cases of no side preference, the system behavior cannot be understood by adding up the behavior of the single individuals. This is a typical feature of complex, self-organizing systems and, in fact, a widespread characteristics of social systems. It is worth noting, however, that it does not require a conscious behavior to reach forms of social organization like the segregation of oppositely moving pedestrians into lanes. This organization occurs automatically, although most people are not even aware of the existence of this phenomenon.

Oscillatory Flows at Bottlenecks At bottlenecks, bidirectional flows of moderate density are often characterized by oscillatory changes in the flow direction (see Fig. 3). For example, one can sometimes observe this at entrances of museums during crowded art exhibitions or at entrances of staff canteens during lunch time. While these oscillatory flows may be interpreted as an effect of friendly behavior ("you go first, please"), computer simulations of the social force model indicate that the collective behavior may again

be understood by simple pedestrian interactions. That is, oscillatory flows occur even in the absence of communication. Therefore, they may be viewed as another self-organization phenomenon, which again reduces frictional effects and delays. That is, oscillatory flows have features of "collective intelligence".

While this may be interpreted as result of a learning effect in a large number of similar situations (a "repeated game"), our simulations suggest an even simpler, "many-particle" interpretation: Once a pedestrian is able to pass the narrowing, pedestrians with the same walking direction can easily follow. Hence, the number and "pressure" of waiting, "pushy" pedestrians on one side of the bottleneck becomes less than on the other side. This eventually increases their chance to occupy the passage. Finally, the "pressure difference" is large enough to stop the flow and turn the passing direction at the bottleneck. This reverses the situation, and eventually the flow direction changes again, giving rise to oscillatory flows.

Stripe Formation in Intersecting Flows In intersection areas, the flow of people often appears to be irregular or "chaotic". In fact, it can be shown that there are several possible collective patterns of motion, among them rotary and oscillating flows. However, these patterns continuously compete with each other, and a temporarily dominating pattern is destroyed by another one after a short time. Obviously, there has not evolved any social convention that would establish and stabilize an ordered and efficient flow at intersections.

Self-organized patterns of motion, however, are found in situations where pedestrian flows cross each other only in two directions. In such situations, the phenomenon of stripe formation is observed [58]. Stripe formation allows two flows to penetrate each other without requiring the pedestrians to stop. For an illustration see Fig. 3. Like lanes, stripes are a segregation phenomenon, but not a stationary one. Instead, the stripes are density waves moving into the direction of the sum of the directional vectors of both intersecting flows. Naturally, the stripes extend sidewards into the direction which is perpendicular to their direction of motion. Therefore, the pedestrians move forward *with* the stripes and sidewards *within* the stripes. Lane formation corresponds to the particular case of stripe formation where both directions are exactly opposite. In this case, no intersection takes place, and the stripes do not move systematically. As in lane formation, stripe formation allows to minimize obstructing interactions and to maximize the average pedestrian speeds, i. e. simple, repulsive pedestrian interactions again lead to an "intelligent" collective behavior.

Evacuation Dynamics

While the previous section has focused on the dynamics of pedestrian crowds in normal situations, we will now turn to the description of situations in which extreme crowd densities occur. Such situations may arise at mass events, particularly in cases of urgent egress. While most evacuations run relatively smoothly and orderly, the situation may also get out of control and end up in terrible crowd disasters (see Table 2). In such situations, one often speaks of "panic", although, from a scientific standpoint, the use of this term is rather controversial. Here, however, we will not be interested in the question whether "panic" actually occurs or not. We will rather focus on the issue of crowd dynamics at high densities and under psychological stress.

Evacuation and Panic Research

Computer models have been also developed for emergency and evacuation situations [32,60,61,62,63,64,65,66,67,68]. Most research into panic, however, has been of empirical nature (see, e. g. [69,70,71,72]), carried out by social psychologists and others.

With some exceptions, panic is observed in cases of scarce or dwindling resources [69,73], which are either required for survival or anxiously desired. They are usually distinguished into escape panic ("stampedes", bank or stock market panic) and acquisitive panic ("crazes", speculative manias) [74,75], but in some cases this classification is questionable [76].

It is often stated that panicking people are obsessed by short-term personal interests uncontrolled by social and cultural constraints [69,74]. This is possibly a result of the reduced attention in situations of fear [69], which also causes that options like side exits are mostly ignored [70]. It is, however, mostly attributed to social contagion [69,71,73,74,75,76,77,78,79,80,81], i. e., a transition from individual to mass psychology, in which individuals transfer control over their actions to others [75], leading to conformity [82]. This "herding behavior" is in some sense irrational, as it often leads to bad overall results like dangerous overcrowding and slower escape [70,75,76]. In this way, herding behavior can increase the fatalities or, more generally, the damage in the crisis faced.

The various socio-psychological theories for this contagion assume hypnotic effects, rapport, mutual excitation of a primordial instinct, circular reactions, social facilitation (see the summary by Brown [80]), or the emergence of normative support for selfish behavior [81]. Brown [80] and Coleman [75] add another explanation related to the prisoner's dilemma [83,84] or common goods dilemma [85], showing that it is reasonable to make one's

Pedestrian, Crowd and Evacuation Dynamics, Table 2
Incomplete list of major crowd disasters since 1970 after J. F. Dickie in [59], http://www.crowddynamics.com/Main/Crowddisasters.html, http://SportsIllustrated.CNN.com/soccer/world/news/2000/07/09/stadium_disasters_ap/, and other internet sources, excluding fires, bomb attacks, and train or plane accidents. The number of injured people was usually a multiple of the fatalities

Date	Place	Venue	Deaths	Reason
1971	Ibrox, UK	Stadium	66	Collapse of barriers
1974	Cairo, Egypt	Stadium	48	Crowds break barriers
1982	Moscow, USSR	Stadium	340	Re-entering fans after last minute goal
1988	Katmandu, Nepal	Stadium	93	Stampede due to hailstorm
1989	Hillsborough, Sheffield, UK	Stadium	96	Fans trying to force their way into the stadium
1990	New York City	Bronx	87	Illegal happy land social club
1990	Mena, Saudi Arabia	Pedestrian Tunnel	1426	Overcrowding
1994	Mena, Saudi Arabia	Jamarat Bridge	266	Overcrowding
1996	Guatemala City, Guatemala	Stadium	83	Fans trying to force their way into the stadium
1998	Mena, Saudi Arabia		118	Overcrowding
1999	Kerala, India	Hindu Shrine	51	Collapse of parts of the shrine
1999	Minsk, Belarus	Subway Station	53	Heavy rain at rock concert
2001	Ghana, West Africa	Stadium	> 100	Panic triggered by tear gas
2004	Mena, Saudi Arabia	Jamarat Bridge	251	Overcrowding
2005	Wai, India	Religious Procession	150	Overcrowding (and fire)
2005	Bagdad, Iraque	Religious Procession	> 640	Rumors regarding suicide bomber
2005	Chennai, India	Disaster Area	42	Rush for flood relief supplies
2006	Mena, Saudi Arabia	Jamarat Bridge	363	Overcrowding
2006	Pilippines	Stadium	79	Rush for game show tickets
2006	Ibb, Yemen	Stadium	51	Rally for Yemeni president

subsequent actions contingent upon those of others. However, the socially favorable behavior of walking orderly is unstable, which normally gives rise to rushing by everyone. These thoughtful considerations are well compatible with many aspects discussed above and with the classical experiments by Mintz [73], which showed that jamming in escape situations depends on the reward structure ("payoff matrix").

Nevertheless and despite of the frequent reports in the media and many published investigations of crowd disasters (see Table 2), a quantitative understanding of the observed phenomena in panic stampedes was lacking for a long time. In the following, we will close this gap.

Situations of "Panic"

Panic stampede is one of the most tragic collective behaviors [71,72,73,74,75,77,78,79,80,81], as it often leads to the death of people who are either crushed or trampled down by others. While this behavior may be comprehensible in life-threatening situations like fires in crowded buildings [69,70], it is hard to understand in cases of a rush for good seats at a pop concert [76] or without any obvious reasons. Unfortunately, the frequency of such disasters is increasing (see Table 2), as growing population densities combined with easier transportation lead to greater mass events like pop concerts, sport events, and demonstrations. Nevertheless, systematic empirical studies of panic [73,86] are rare [69,74,76], and there is a scarcity of quantitative theories capable of predicting crowd dynamics at extreme densities [32,60,61,64,65,68]. The following features appear to be typical [46,55]:

1. In situations of escape panic, individuals are getting nervous, i. e. they tend to develop blind actionism.
2. People try to move considerably faster than normal [9].
3. Individuals start pushing, and interactions among people become physical in nature.
4. Moving and, in particular, passing of a bottleneck frequently becomes incoordinated [73].
5. At exits, jams are building up [73]. Sometimes, intermittent flows or arching and clogging are observed [9], see Fig. 4.
6. The physical interactions in jammed crowds add up and can cause dangerous pressures up to 4,500 Newtons per meter [59,70], which can bend steel barriers or tear down brick walls.
7. The strength and direction of the forces acting in large crowds can suddenly change [87], pushing peo-

Pedestrian, Crowd and Evacuation Dynamics, Figure 4
Panicking football fans trying to escape the football stadium in Sheffield. Because of a clogging effect, it is difficult to pass the open door

ple around in an uncontrollable way. This may cause people to fall.

8. Escape is slowed down by fallen or injured people turning into "obstacles".
9. People tend to show herding behavior, i. e., to do what other people do [69,78].
10. Alternative exits are often overlooked or not efficiently used in escape situations [69,70].

The following quotations give a more personal impression of the conditions during crowd panic:

1. "They just kept pushin' forward and they would just walk right on top of you, just trample over ya like you were a piece of the ground." (After the panic at "The Who Concert Stampede" in Cincinnati.)
2. "People were climbin' over people ta get in … an' at one point I almost started hittin' 'em, because I could not believe the animal, animalistic ways of the people, you know, nobody cared." (After the panic at "The Who Concert Stampede".)
3. "Smaller people began passing out. I attempted to lift one girl up and above to be passed back … After several tries I was unsuccessful and near exhaustion." (After the panic at "The Who Concert Stampede".)
4. "I couldn't see the floor because of the thickness of the smoke." (After the "Hilton Hotel Fire" in Las Vegas.)
5. "The club had two exits, but the young people had access to only one", said Narend Singh, provincial minister for agriculture and environmental affairs. However, the club's owner, Rajan Naidoo, said the club had four exits, and that all were open. "I think the children panicked and headed for the main entrance where they initially came in,' he said." (After the "Durban Disco Stampede".)
6. "At occupancies of about 7 persons per square meter the crowd becomes almost a fluid mass. Shock waves can be propagated through the mass, sufficient to … propel them distances of 3 meters or more. … People may be literally lifted out of their shoes, and have clothing torn off. Intense crowd pressures, exacerbated by anxiety, make it difficult to breathe, which may finally cause compressive asphyxia. The heat and the thermal insulation of surrounding bodies cause some to be weakened and faint. Access to those who fall is impossible. Removal of those in distress can only be accomplished by lifting them up and passing them overhead to the exterior of the crowd." (J. Fruin in [88].)
7. "It was like a huge wave of sea gushing down on the pilgrims" (P. K. Abdul Ghafour, Arab News, after the sad crowd disaster in Mena on January 12, 2006).

Force Model for Panicking Pedestrians

Additional, physical interaction forces $\boldsymbol{f}^{\mathrm{ph}}_{\alpha\beta}$ come into play when pedestrians get so close to each other that they have physical contact (i. e. $d_{\alpha\beta} < r_{\alpha\beta} = r_\alpha + r_\beta$, where r_α means the "radius" of pedestrian α). In this case, which is mainly relevant to panic situations, we assume also a "*body force*" $k(r_{\alpha\beta} - d_{\alpha\beta})\boldsymbol{n}_{\alpha\beta}$ counteracting body compression and a "*sliding friction force*" $\kappa(r_{\alpha\beta} - d_{\alpha\beta})\Delta v^t_{\beta\alpha}\boldsymbol{t}_{\alpha\beta}$ impeding *relative* tangential motion. Inspired by the formulas for granular interactions [89,90], we assume

$$\boldsymbol{f}^{\mathrm{ph}}_{\alpha\beta}(t) = k\Theta(r_{\alpha\beta} - d_{\alpha\beta})\boldsymbol{n}_{\alpha\beta} + \kappa\Theta(r_{\alpha\beta} - d_{\alpha\beta})\Delta v^t_{\beta\alpha}\boldsymbol{t}_{\alpha\beta}\,, \tag{13}$$

where the function $\Theta(z)$ is equal to its argument z, if $z \geq 0$, otherwise 0. Moreover, $\boldsymbol{t}_{\alpha\beta} = (-n^2_{\alpha\beta}, n^1_{\alpha\beta})$ means the tangential direction and $\Delta v^t_{\beta\alpha} = (\boldsymbol{v}_\beta - \boldsymbol{v}_\alpha)\cdot\boldsymbol{t}_{\alpha\beta}$ the tangential velocity difference, while k and κ represent large constants. (Strictly speaking, friction effects already set in before pedestrians touch each other, because of the psychological tendency not to pass other individuals with a high relative velocity, when the distance is small.)

The interactions with the boundaries of walls and other obstacles are treated analogously to pedestrian interactions, i. e., if $d_{\alpha i}(t)$ means the distance to obstacle or boundary i, $\boldsymbol{n}_{\alpha i}(t)$ denotes the direction perpendicular to it, and $\boldsymbol{t}_{\alpha i}(t)$ the direction tangential to it, the corresponding interaction force with the boundary reads

$$\begin{aligned}\boldsymbol{f}_{\alpha i} = &\left\{A_\alpha \exp[(r_\alpha - d_{\alpha i})/B_\alpha] + k\Theta(r_\alpha - d_{\alpha i})\right\} \\ &\times \boldsymbol{n}_{\alpha i} - \kappa\Theta(r_\alpha - d_{\alpha i})(\boldsymbol{v}_\alpha\cdot\boldsymbol{t}_{\alpha i})\boldsymbol{t}_{\alpha i}\,.\end{aligned} \tag{14}$$

Finally, fire fronts are reflected by repulsive social forces similar those describing walls, but they are much stronger. The physical interactions, however, are qualitatively different, as people reached by the fire front become injured and immobile ($\boldsymbol{v}_\alpha = \mathbf{0}$).

Collective Phenomena in Panic Situations

In panic situations (e. g. in some cases of emergency evacuation) the following characteristic features of pedestrian behavior are often observed:

1. People are getting nervous, resulting in a higher level of fluctuations.
2. They are trying to escape from the source of panic, which can be reflected by a significantly higher desired velocity v_α^0.
3. Individuals in complex situations, who do not know what is the right thing to do, orient at the actions of their neighbors, i. e. they tend to do what other people do. We will describe this by an additional herding interaction.

We will now discuss the fundamental collective effects which fluctuations, increased desired velocities, and herding behavior can have. In contrast to other approaches, we do not assume or imply that individuals in panic or emergency situations would behave relentless and asocial, although they sometimes do.

Herding and Ignorance of Available Exits If people are not sure what is the best thing to do, there is a tendency to show a "herding behavior", i. e. to imitate the behavior of others. Fashion, hypes and trends are examples for this. The phenomenon is also known from stock markets, and particularly pronounced when people are anxious. Such a situation is, for example, given if people need to escape from a smoky room. There, the evacuation dynamics is very different from normal leaving (see Fig. 5).

Under normal visibility, everybody easily finds an exit and uses more or less the shortest path. However, when the exit cannot be seen, evacuation is much less efficient and may take a long time. Most people tend to walk relatively straight into the direction in which they suspect an exit, but in most cases, they end up at a wall. Then, they usually move along it in one of the two possible directions, until they finally find an exit [18]. If they encounter others, there is a tendency to take a decision for one direction

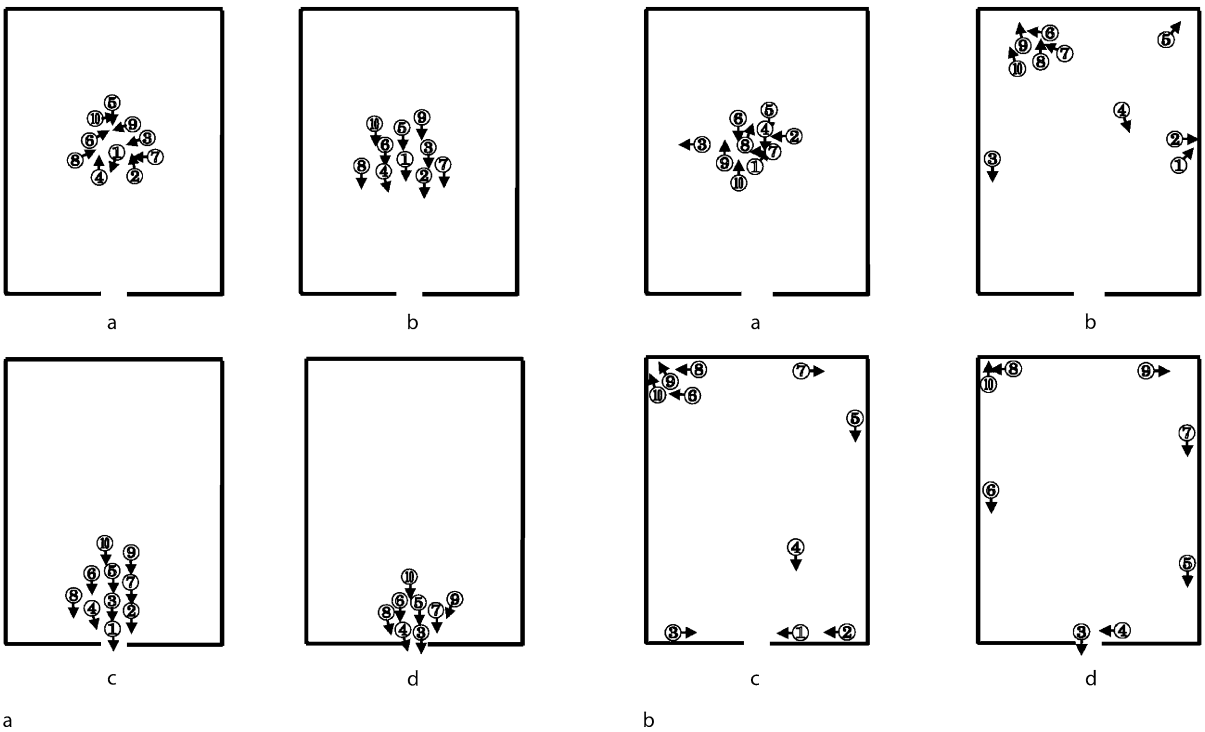

Pedestrian, Crowd and Evacuation Dynamics, Figure 5
a Normal leaving of a room, when the exit is well visible. b Escape from a room with no visibility, e. g. due to dense smoke or a power blackout. (After [18])

and move collectively. Also in case of acoustic signals, people may be attracted into the same direction. This can lead to over-crowded exits, while other exits are ignored. The same can happen even for normal visibility, when people are not well familiar with their environment and are not aware of the directions of the emergency exits.

Computer simulations suggest that neither individualistic nor herding behavior performs well [46]. Pure individualistic behavior means that each pedestrian finds an exit only accidentally, while pure herding behavior implies that the complete crowd is eventually moving into the same and probably congested direction, so that available emergency exits are not efficiently used. Optimal chances of survival are expected for a certain mixture of individualistic and herding behavior, where individualism allows *some* people to detect the exits and herding guarantees that successful solutions are imitated by small groups of others [46].

"Freezing by Heating" Another effect of getting nervous has been investigated in [55]. Let us assume the individual fluctuation strength is given by

$$\eta_\alpha = (1 - n_\alpha)\eta_0 + n_\alpha \eta_{\max} \,, \tag{15}$$

where n_α with $0 \le n_\alpha \le 1$ measures the nervousness of pedestrian α. The parameter η_0 means the normal and $\eta_{\max}$ the maximum fluctuation strength. It turns out that, at sufficiently high pedestrian densities, lanes are destroyed by increasing the fluctuation strength (which is analogous to the temperature). However, instead of the expected transition from the "fluid" lane state to a disordered, "gaseous" state, a solid state is formed. It is characterized by an at least temporarily blocked, "frozen" situation so that one calls this paradoxical transition "*freezing by heating*" (see Fig. 6). Notably enough, the blocked state has a *higher* degree of order, although the internal energy is *increased* [55].

The preconditions for this unusual freezing-by-heating transition are the driving term $v_\alpha^0 \boldsymbol{e}_\alpha^0/\tau_\alpha$ and the dissipative friction $-\boldsymbol{v}_\alpha/\tau_\alpha$, while the sliding friction force is not required. Inhomogeneities in the channel diameter

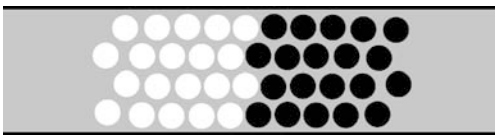

Pedestrian, Crowd and Evacuation Dynamics, Figure 6
Result of the noise-induced formation of a "frozen" state in a (periodic) corridor used by oppositely moving pedestrians (after [55])

or other impurities which temporarily slow down pedestrians can further this transition at the respective places. Finally note that a transition from fluid to blocked pedestrian counter flows is also observed, when a critical density is exceeded [31,55].

Intermittent Flows, Faster-Is-Slower Effect, and "Phantom Panic" If the overall flow towards a bottleneck is higher than the overall outflow from it, a pedestrian queue emerges [91]. In other words, a waiting crowd is formed upstream of the bottleneck. High densities can result, if people keep heading forward, as this eventually leads to higher and higher compressions. Particularly critical situations may occur if the arrival flow is much higher than the departure flow, especially if people are trying to get towards a strongly desired goal ("acquisitive panic") or away from a perceived source of danger ("escape panic") with an increased driving force $v_\alpha^0 \boldsymbol{e}_\alpha^0/\tau$. In such situations, the high density causes coordination problems, as several people compete for the same few gaps. This typically causes body interactions and frictional effects, which can slow down crowd motion or evacuation ("*faster is slower effect*").

A possible consequence of these coordination problems are intermittent flows. In such cases, the outflow from the bottleneck is not constant, but it is typically interrupted. While one possible origin of the intermittent flows are clogging and arching effects as known from granular flows through funnels or hoppers [89,90], stop-and-go waves have also been observed in more than 10 meter wide streets and in the 44 meters wide entrance area to the Jamarat Bridge during the pilgrimage in January 12, 2006 [87], see Fig. 7. Therefore, it seems to be important that people do not move continuously, but have minimum strides [25]. That is, once a person is stopped, he or she will not move until some space opens up in front. However, increasing impatience will eventually reduce the minimum stride, so that people eventually start moving again, even if the outflow through the bottleneck is stopped. This will lead to a further compression of the crowd.

In the worst case, such behavior can trigger a "*phantom panic*", i. e. a crowd disaster *without* any serious reasons (e. g., in Moscow, 1982). For example, due to the "faster-is-slower effect" panic can be triggered by small pedestrian counterflows [70], which cause delays to the crowd intending to leave. Consequently, stopped pedestrians in the back, who do not see the reason for the temporary slowdown, are getting impatient and pushy. In accordance with observations [7,25], one may model this by increasing the desired velocity, for example, by the formula

$$v_\alpha^0(t) = [1 - n_\alpha(t)]v_\alpha^0(0) + n_\alpha(t)v_\alpha^{\max} \,. \tag{16}$$

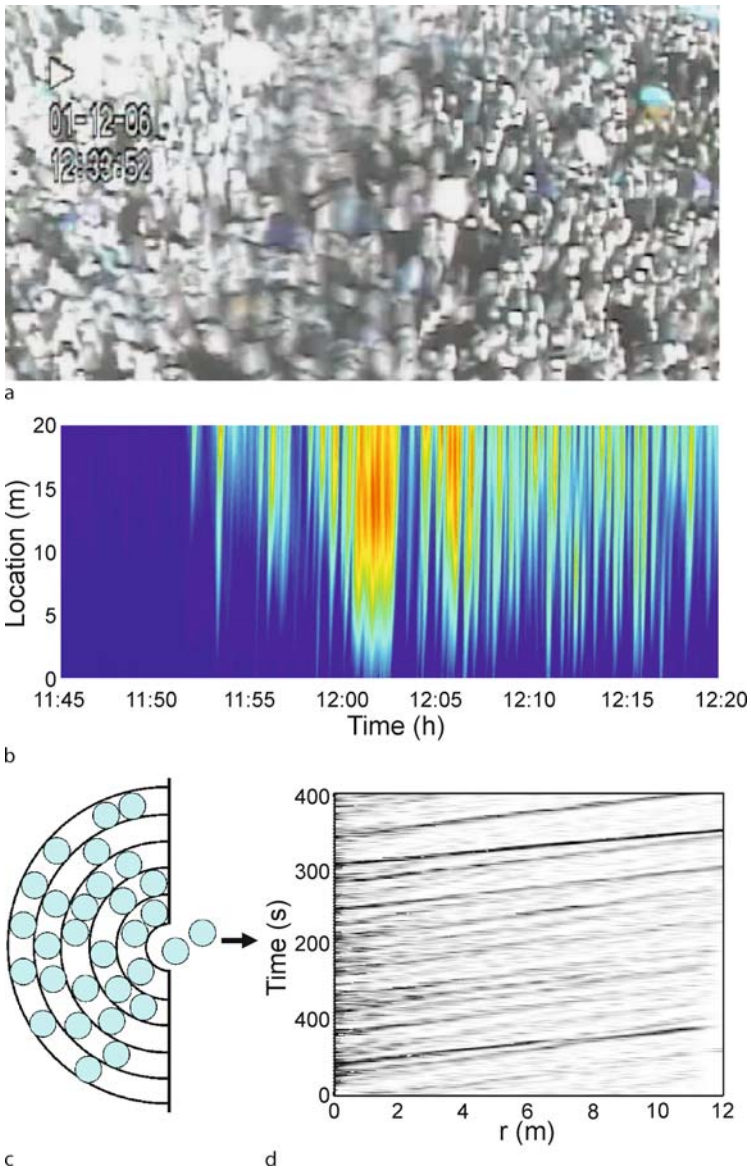

Pedestrian, Crowd and Evacuation Dynamics, Figure 7

a Long-term photograph showing stop-and-go waves in a densely packed street. While stopped people appear relatively sharp, people moving from right to left have a fuzzy appearance. Note that gaps propagate from *right* to *left*. b Empirically observed stop-and-go waves in front of the entrance to the Jamarat Bridge on January 12, 2006 (after [87]), where pilgrims moved from *left* to *right*. *Dark areas* correspond to phases of motion, *light colors* to stop phases. c Illustration of the "shell model", in particular of situations where several pedestrians compete for the same gap, which causes coordination problems. d Stop-and-go waves resulting from the alternation of forward pedestrian motion and backward gap propagation

Herein, $v_\alpha^{\max}$ is the maximum desired velocity and $v_\alpha^0(0)$ the initial one, corresponding to the expected velocity of leaving. The time-dependent parameter

$$n_\alpha(t) = 1 - \frac{\overline{v}_\alpha(t)}{v_\alpha^0(0)} \tag{17}$$

reflects the nervousness, where $\overline{v}_\alpha(t)$ denotes the average speed into the desired direction of motion. Altogether, long waiting times increase the desired speed v_α^0 or driving force $v_\alpha^0(t)\boldsymbol{e}_\alpha^0/\tau$, which can produce high densities and inefficient motion. This further increases the waiting times, and so on, so that this tragic feedback can eventually trig-

ger so high pressures that people are crushed or falling and trampled. It is, therefore, imperative, to have sufficiently wide exits and to prevent counterflows, when big crowds want to leave [46].

Transition to Stop-and-Go Waves Recent empirical studies of pilgrim flows in the area of Makkah, Saudi Arabia, have shown that intermittent flows occur not only when bottlenecks are obvious. On January 12, 2006, pronounced stop-and-go waves have been even observed upstream of the 44 m wide entrance to the Jamarat Bridge [87]. While the pilgrim flows were smooth and continuous ("laminar") over many hours, at 11:53 am stop-and-go waves suddenly appeared and propagated over distances of more than 30 m (see Fig. 7). The sudden transition was related to a significant drop of the flow, i. e. with the onset of congestion [87]. Once the stop-and-go waves set in, they persisted over more than 20 min.

This phenomenon can be reproduced by a recent model based on two continuity equations, one for forward pedestrian motion and another one for backward gap propagation [91]. The model was derived from a "shell model" (see Fig. 7) and describes very well the observed alternation between backward gap propagation and forward pedestrian motion.

Transition to "Crowd Turbulence" On the same day, around 12:19, the density reached even higher values and the video recordings showed a sudden transition from stop-and-go waves to *irregular* flows (see Fig. 8). These irregular flows were characterized by random, unintended displacements into all possible directions, which pushed people around. With a certain likelihood, this caused them to stumble. As the people behind were moved by the crowd as well and could not stop, fallen individuals were trampled, if they did not get back on their feet quickly enough. Tragically, the area of trampled people grew more and more in the course of time, as the fallen pilgrims became obstacles for others [87]. The result was one of the biggest crowd disasters in the history of pilgrimage.

How can we understand this transition to irregular crowd motion? A closer look at video recordings of the crowd reveals that, at this time, people were so densely packed that they were moved involuntarily by the crowd. This is reflected by random displacements into all possible directions. To distinguish these irregular flows from laminar and stop-and-go flows and due to their visual appearance, we will refer to them as "*crowd turbulence*".

As in certain kinds of fluid flows, "turbulence" in crowds results from a sequence of instabilities in the flow pattern. Additionally, one finds a sharply peaked probability density function of velocity increments

$$V_x^\tau = V_x(\boldsymbol{r}, t+\tau) - V_x(\boldsymbol{r}, t)\,, \tag{18}$$

which is typical for turbulence [92], if the time shift τ is small enough [87]. One also observes a power-law scaling of the displacements indicating self-similar behavior [87]. As large eddies are not detected, however, the similarity with *fluid* turbulence is limited, but there is still an analogy to turbulence at currency exchange markets [92]. Instead of vortex cascades like in turbulent fluids, one rather finds a hierarchical fragmentation dynamics: At extreme

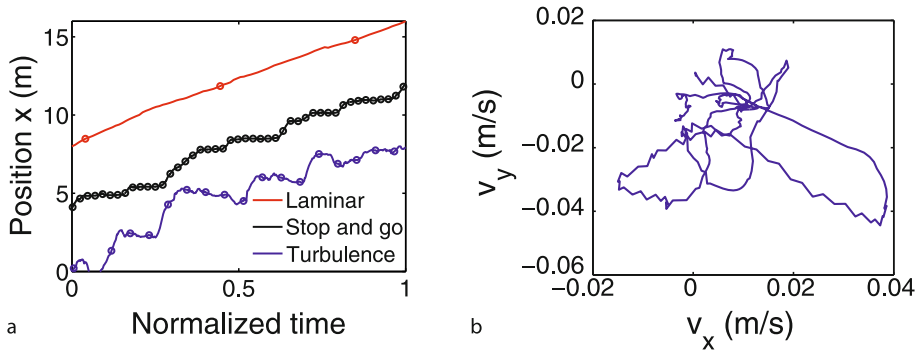

Pedestrian, Crowd and Evacuation Dynamics, Figure 8

Pedestrian dynamics at different densities. a Representative trajectories (space-time plots) of pedestrians during the laminar, stop-and-go, and turbulent flow regime. Each trajectory extends over a range of 8 meters, while the time required for this stretch is normalized to 1. To indicate the different speeds, symbols are included in the curves every 5 seconds. While the laminar flow (*top line*) is fast and smooth, motion is temporarily interrupted in stop-and-go flow (*medium line*), and backward motion can occur in "turbulent" flows (*bottom line*). b Example of the temporal evolution of the velocity components $v_x(t)$ into the average direction of motion and $v_y(t)$ perpendicular to it in "turbulent flow", which occurs when the crowd density is extreme. One can clearly see the irregular motion into all possible directions characterizing "crowd turbulence". For details see [87]

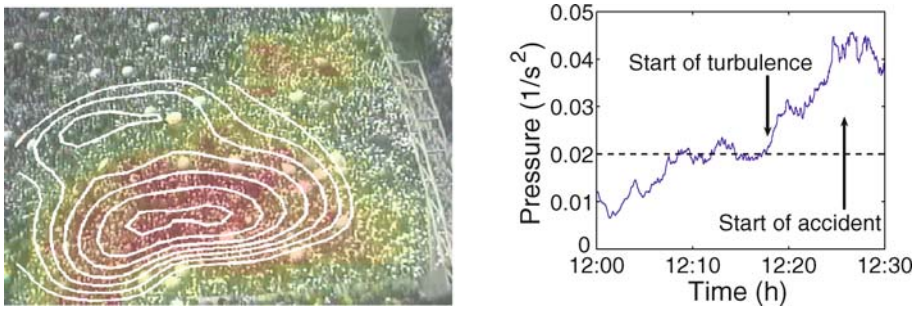

Pedestrian, Crowd and Evacuation Dynamics, Figure 9
***Left*: Snapshot of the on-line visualization of "crowd pressure". *Red colors* (see the lower ellipses) indicate areas of critical crowd conditions. In fact, the sad crowd disaster during the Muslim pilgrimage on January 12, 2006, started in this area. *Right*: The "crowd pressure" is a quantitative measure of the onset of "crowd turbulence". The crowd disaster started when the "crowd pressure" reached particularly high values**

densities, individual motion is replaced by mass motion, but there is a stick-slip instability which leads to "rupture" when the stress in the crowd becomes too large. That is, the mass splits up into clusters of different sizes with strong velocity correlations *inside* and distance-dependent correlations *between* the clusters.

"Crowd turbulence" has further specific features [87]. Due to the physical contacts among people in extremely dense crowds, we expect commonalities with granular media. In fact, dense driven granular media may form density waves, while moving forward [93], and can display turbulent-like states [94,95]. Moreover, under quasi-static conditions [94], force chains [96] are building up, causing strong variations in the strengths and directions of local forces. As in earthquakes [97,98] this can lead to events of sudden, uncontrollable stress release with power-law distributed displacements. Such a power-law has also been discovered by video-based crowd analysis [87].

Some Warning Signs of Critical Crowd Conditions

Turbulent waves are experienced in dozens of crowd-intensive events each year all over the world [88]. Therefore, it is necessary to understand why, where and when potentially critical situations occur. Viewing real-time video recordings is not very suited to identify critical crowd conditions: While the average density rarely exceeds values of 6 persons per square meter, the local densities can reach almost twice as large values [87]. It has been found, however, that even evaluating the local densities is not enough to identify the critical times and locations precisely, which also applies to an analysis of the velocity field [87]. The decisive quantity is rather the "crowd pressure", i. e. the density, multiplied with the variance of speeds. It allows one to identify critical locations and times (see Fig. 9).

There are even advance warning signs of critical crowd conditions: The crowd accident on January 12, 2006 started about 10 minutes after "turbulent" crowd motion set in, i. e. after the "pressure" exceeded a value of $0.02/s^2$ (see Fig. 9). Moreover, it occurred more than 30 min after stop-and-go waves set in, which can be easily detected in accelerated surveillance videos. Such advance warning signs of critical crowd conditions can be evaluated on-line by an automated video analysis system. In many cases, this can help one to gain time for corrective measures like flow control, pressure-relief strategies, or the separation of crowds into blocks to stop the propagation of shock-waves [87]. Such anticipative crowd control could increase the level of safety during future mass events.

Evolutionary Optimization of Pedestrian Facilities

Having understood some of the main factors causing crowd disasters, it is interesting to ask how pedestrian facilities can be designed in a way that maximizes the efficiency of pedestrian flows and the level of safety. One of the major goals during mass events must be to avoid extreme densities. These often result from the onset of congestion at bottlenecks, which is a consequence of the breakdown of free flow and causes an increasing degree of compression. When a certain critical density is increased (which depends on the size distribution of people), this potentially implies high pressures in the crowd, particularly if people are impatient due to long delays or panic.

The danger of an onset of congestion can be minimized by avoiding bottlenecks. Notice, however, that jamming can also occur at widenings of escape routes [46]. This surprising fact results from disturbances due to pedestrians, who try to overtake each other and expand in the wider area because of their repulsive interactions.

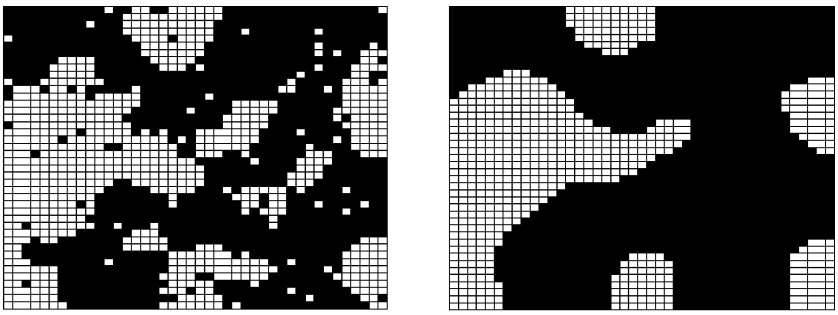

a b

Pedestrian, Crowd and Evacuation Dynamics, Figure 10
The evolutionary optimization based on Boolean grids [99] uses a two-stage algorithm. a In the randomization stage, obstacles are distributed over the grid with some randomness, thereby allowing for the generation of new topologies. b In the agglomeration stage, small nearby obstacles are clustered to form larger objects with smooth boundaries

These squeeze into the main stream again at the end of the widening, which acts like a bottleneck and leads to jamming. The corresponding drop of efficiency E is more pronounced,

1. if the corridor is narrow,
2. if the pedestrians have different or high desired velocities, and
3. if the pedestrian density in the corridor is high.

Obviously, the emerging pedestrian flows decisively depend on the geometry of the boundaries. They can be simulated on a computer already in the planning phase of pedestrian facilities. Their configuration and shape can be systematically varied, e. g. by means of evolutionary algorithms [28,100] and evaluated on the basis of particular mathematical performance measures [7]. Apart from the *efficiency*

$$E = \frac{1}{N} \sum_{\alpha} \frac{\boldsymbol{v}_\alpha \cdot \boldsymbol{e}_\alpha^0}{v_\alpha^0} \tag{19}$$

we can, for example, define the *measure of comfort* $C = (1 - D)$ via the discomfort

$$D = \frac{1}{N} \sum_{\alpha} \frac{\overline{(\boldsymbol{v}_\alpha - \overline{\boldsymbol{v}_\alpha})^2}}{\overline{(\boldsymbol{v}_\alpha)^2}} = \frac{1}{N} \sum_{\alpha} \left(1 - \frac{\overline{\boldsymbol{v}_\alpha}^2}{\overline{(\boldsymbol{v}_\alpha)^2}} \right) . \tag{20}$$

The latter is again between 0 and 1 and reflects the frequency and degree of sudden velocity changes, i. e. the level of discontinuity of walking due to necessary avoidance maneuvers. Hence, the optimal configuration regarding the pedestrian requirements is the one with the highest values of efficiency and comfort.

During the optimization procedure, some or all of the following can be varied:

1. the location and form of planned buildings,
2. the arrangement of walkways, entrances, exits, staircases, elevators, escalators, and corridors,
3. the shape of rooms, corridors, entrances, and exits,
4. the function and time schedule. (Recreation rooms or restaurants are often continuously frequented, rooms for conferences or special events are mainly visited and left at peak periods, exhibition rooms or rooms for festivities require additional space for people standing around, and some areas are claimed by queues or through traffic.)

In contrast to early evolutionary optimization methods, recent approaches allow to change not only the dimensions of the different elements of pedestrian facilities, but also to vary their topology. The procedure of such algorithms is illustrated in Fig. 10. Highly performing designs are illustrated in Fig. 11. It turns out that, for an emergency evacuation route, it is favorable if the crowd does not move completely straight towards a bottleneck. For example, a zigzag design of the evacuation route can reduce the pressure on the crowd upstream of a bottleneck (see Fig. 12). The proposed evolutionary optimization procedure can, of course, not only be applied to the design of new pedestrian facilities, but also to a reduction of existing bottlenecks, when suitable modifications are implemented.

Future Directions

In this contribution, we have presented a multi-agent approach to pedestrian and crowd dynamics. Despite the great effort required, pedestrian interactions can be well quantified by video tracking. Compared to other social interactions they turn out to be quite simple. Neverthe-

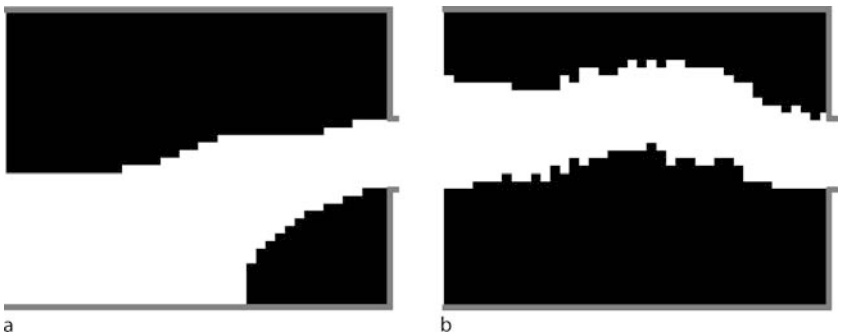

Pedestrian, Crowd and Evacuation Dynamics, Figure 11
Two examples of improved designs for cases with a bottleneck along the escape route of a large crowd, obtained with an evolutionary algorithm based on Boolean grids. People were assumed to move from left to right only. a Funnel-shaped escape route. b Zig-zag design

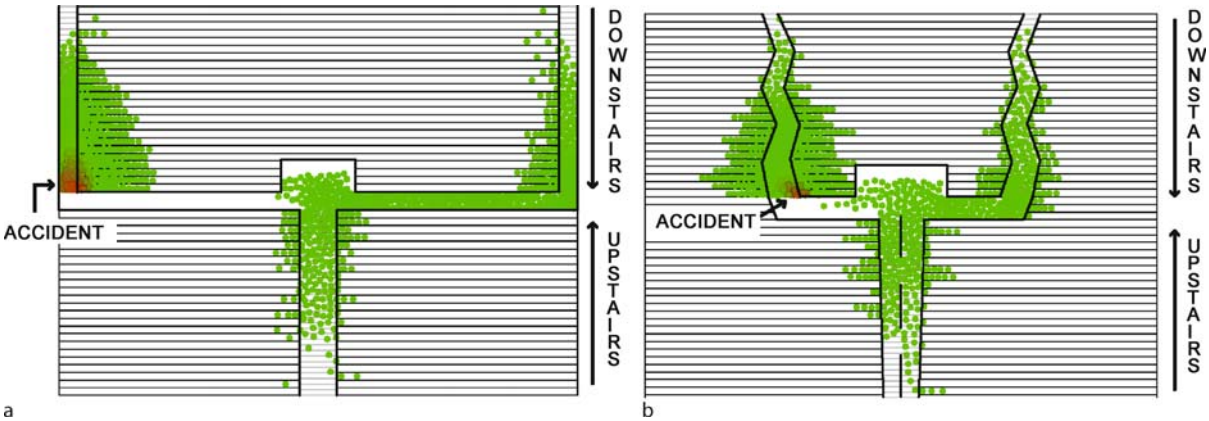

Pedestrian, Crowd and Evacuation Dynamics, Figure 12
a Conventional design of a stadium exit in an emergency scenario, where we assume that some pedestrians have fallen at the end of the downwards staircase to the left. The dark color indicates high pressures, since pedestrians are impatient and pushing from behind. b In the improved design, the increasing diameter of corridors can reduce waiting times and impatience (even with the same number of seats), thereby accelerating evacuation. Moreover, the zigzag design of the downwards staircases changes the pushing direction in the crowd. (After [8])

less, they cause a surprisingly large variety of self-organized patterns and short-lived social phenomena, where coordination or cooperation emerges spontaneously. For this reason, they are interesting to study, particularly as one can expect new insights into coordination mechanisms of social beings beyond the scope of classical game theory. Examples for observed self-organization phenomena in normal situations are lane formation, stripe formation, oscillations and intermittent clogging effects at bottlenecks, and the evolution of behavioral conventions (such as the preference of the right-hand side in continental Europe). Under extreme conditions (high densities or panic), however, coordination may break down, giving rise to "freezing-by-heating" or "faster-is-slower effects", stop-and-go waves or "crowd turbulence".

Similar observations as in pedestrian crowds are made in other social systems and settings. Therefore, we expect that realistic models of pedestrian dynamics will also promote the understanding of opinion formation and other kinds of collective behaviors. The hope is that, based on the discovered elementary mechanisms of emergence and self-organization, one can eventually also obtain a better understanding of the constituting principles of more complex social systems. At least the same underlying factors are found in many social systems: Non-linear interactions of individuals, time-dependence, heterogeneity, stochasticity, competition for scarce resources (here: Space and time), decision-making, and learning. Future work will certainly also address issues of perception, anticipation, and communication.

Acknowledgments

The authors are grateful for partial financial support by the German Research Foundation (research projects He 2789/7-1, 8-1) and by the "Cooperative Center for Communication Networks Data Analysis", a NAP project sponsored by the Hungarian National Office of Research and Technology under grant No. KCKHA005.

Bibliography

Primary Literature

1. Hankin BD, Wright RA (1958) Passenger flow in subways. Operat Res Q 9:81–88
2. Older SJ (1968) Movement of pedestrians on footways in shopping streets. Traffic Eng Control 10:160–163
3. Weidmann U (1993) Transporttechnik der Fußgänger. In: Schriftenreihe des Instituts für Verkehrsplanung, Transporttechnik, Straßen- und Eisenbahnbau. Institut für Verkehrsplanung, Transporttechnik, Straßen- und Eisenbahnbau, Zürich
4. Fruin JJ (1971) Designing for pedestrians: A level-of-service concept. In: Highway research record, Number 355: Pedestrians. Highway Research Board, Washington DC, pp 1–15
5. Pauls J (1984) The movement of people in buildings and design solutions for means of egress. Fire Technol 20:27–47
6. Whyte WH (1988) City. Rediscovering the center. Doubleday, New York
7. Helbing D (1997) Verkehrsdynamik. Springer, Berlin
8. Helbing D, Buzna L, Johansson A, Werner T (2005) Self-organized pedestrian crowd dynamics: Experiments, simulations, and design solutions. Transport Sci 39(1):1–24
9. Predtetschenski WM, Milinski AI (1971) Personenströme in Gebäuden – Berechnungsmethoden für die Projektierung. Müller, Köln-Braunsfeld
10. Transportation Research Board (1985) Highway Capacity Manual, Special Report 209. Transportation Research Board, Washington DC
11. Yuhaski SJ Jr, Macgregor Smith JM (1989) Modelling circulation systems in buildings using state dependent queueing models. Queueing Syst 4:319–338
12. Garbrecht D (1973) Describing pedestrian and car trips by transition matrices. Traffic Q 27:89–109
13. Ashford N, O'Leary M, McGinity PD (1976) Stochastic modelling of passenger and baggage flows through an airport terminal. Traffic Engin Control 17:207–210
14. Borgers A, Timmermans H (1986) City centre entry points, store location patterns and pedestrian route choice behaviour: A microlevel simulation model. Socio-Econ Plan Sci 20:25–31
15. Helbing D (1993) Stochastische Methoden, nichtlineare Dynamik und quantitative Modelle sozialer Prozesse. Ph.D. thesis University of Stuttgart, 1992 (published by Shaker, Aachen)
16. Helbing D, Isobe M, Nagatani T, Takimoto K (2003) Lattice gas simulation of experimentally studied evacuation dynamics. Phys Rev E 67:067101
17. Daamen W, Hoogendoorn SP (2003) Experimental research on pedestrian walking behavior (CDROM). In: Proceedings of the 82nd annual meeting at the transportation research board, Washington DC
18. Isobe M, Helbing D, Nagatani T (2004) Experiment, theory, and simulation of the evacuation of a room without visibility Phys Rev E 69:066132
19. Seyfried A, Steffen B, Klingsch W, Boltes M (2005) The fundamental diagram of pedestrian movement revisited. J Stat Mech P10002
20. Kretz T, Wölki M, Schreckenberg M (2006) Characterizing correlations of flow oscillations at bottlenecks. J Stat Mech P02005
21. Henderson LF (1974) On the fluid mechanics of human crowd motion. Transp Res 8:509–515
22. Hughes RL (2002) A continuum theory for the flow of pedestrians. Transp Res B 36:507–535
23. Helbing D (1992) A fluid-dynamic model for the movement of pedestrians. Complex Syst 6:391–415
24. Hoogendoorn SP, Bovy PHL (2000) Gas-kinetic modelling and simulation of pedestrian flows. Transp Res Rec 1710:28–36
25. Helbing D (1991) A mathematical model for the behavior of pedestrians. Behav Sci 36:298–310
26. Helbing D, Molnár P (1995) Social force model for pedestrian dynamics. Phys Rev E 51:4282–4286
27. Gipps PG, Marksjö B (1985) A micro-simulation model for pedestrian flows. Math Comp Simul 27:95–105
28. Bolay K (1998) Nichtlineare Phänomene in einem fluid-dynamischen Verkehrsmodell. Master's thesis, University of Stuttgart
29. Blue VJ, Adler JL (1998) Emergent fundamental pedestrian flows from cellular automata microsimulation. Transp Res Rec 1644:29–36
30. Fukui M, Ishibashi Y (1999) Self-organized phase transitions in cellular automaton models for pedestrians. J Phys Soc Japan 68:2861–2863
31. Muramatsu M, Irie T, Nagatani T (1999) Jamming transition in pedestrian counter flow. Physica A 267:487–498
32. Klüpfel H, Meyer-König M, Wahle J, Schreckenberg M (2000) Microscopic simulation of evacuation processes on passenger ships. In: Bandini S, Worsch T (eds) Theory and practical issues on cellular automata. Springer, London
33. Burstedde C, Klauck K, Schadschneider A, Zittartz J (2001) Simulation of pedestrian dynamics using a 2-dimensional cellular automaton. Physica A 295:507–525
34. Gopal S, Smith TR (1990) NAVIGATOR: An AI-based model of human way-finding in an urban environment. In: Fischer MM, Nijkamp P, Papageorgiou YY (eds) Spatial choices and processes. North-Holland, Amsterdam, pp 169–200
35. Reynolds CW (1994) Evolution of corridor following behavior in a noisy world. In: Cliff D, Husbands P, Meyer J-A, Wilson S (eds) From animals to animats 3: Proceedings of the third international conference on simulation of adaptive behavior. MIT Press, Cambridge, pp 402–410
36. Helbing D (1992) A mathematical model for attitude formation by pair interactions. Behav Sci 37:190–214
37. Helbing D, Molnár P, Farkas I, Bolay K (2001) Self-organizing pedestrian movement. Env Planning B 28:361–383
38. Klockgether J, Schwefel H-P (1970) Two-phase nozzle and hollow core jet experiments. In: Elliott DG (ed) Proceedings of the eleventh symposium on engineering aspects of magnetohydrodynamics. California Institute of Technology, Pasadena, pp 141–148

39. Helbing D (1992) A mathematical model for behavioral changes by pair interactions. In: Haag G, Mueller U, Troitzsch KG (eds) Economic evolution and demographic change. Formal models in social sciences. Springer, Berlin, pp 330–348
40. Miller NE (1944) Experimental studies of conflict. In: Mc Hunt VJ (ed) Personality and the behavior disorders, vol 1. Ronald, New York
41. Miller NE (1959) Liberalization of basic S-R-concepts: Extension to conflict behavior, motivation, and social learning. In: Koch S (ed) Psychology: A study of science, vol 2. McGraw Hill, New York
42. Lewin K (1951) Field theory in social science. Harper, New York
43. Helbing D (1994) A mathematical model for the behavior of individuals in a social field. J Math Sociol 19(3):189–219
44. Hoogendoorn S, Bovy PHL (2003) Simulation of pedestrian flows by optimal control and differential games. Optim Control Appl Meth 24(3):153–172
45. Johansson A, Helbing D, Shukla PK (2007) Specification of the social force pedestrian model by evolutionary adjustment to video tracking data. Adv Complex Syst 10:271–288
46. Helbing D, Farkas I, Vicsek T (2000) Simulating dynamical features of escape panic. Nature 407:487–490
47. Kerridge J, Chamberlain T (2005) Collecting pedestrian trajectory data in real-time. In: Waldau N, Gattermann P, Knoflacher H, Schreckenberg M (eds) Pedestrian and evacuation dynamics '05. Springer, Berlin
48. Hoogendoorn SP, Daamen W, Bovy PHL (2003) Extracting microscopic pedestrian characteristics from video data (CDROM). In: Proceedings of the 82nd annual meeting at the transportation research board. Mira Digital, Washington DC
49. Teknomo K (2002) Microscopic pedestrian flow characteristics: Development of an image processing data collection and simulation model. Ph D thesis, Tohoku University Japan
50. Kadanoff LP (1985) Simulating hydrodynamics: A pedestrian model. J Stat Phys 39:267–283
51. Stanley HE, Ostrowsky N (eds) (1986) On growth and form. Nijhoff, Boston
52. Arns T (1993) Video films of pedestrian crowds. Stuttgart
53. Stølum H-H (1996) River meandering as a self-organization process. Nature 271:1710–1713
54. Rodríguez-Iturbe I, Rinaldo A (1997) Fractal river basins: Chance and self-organization. Cambridge University, Cambridge
55. Helbing D, Farkas I, Vicsek T (2000) (2000) Freezing by heating in a driven mesoscopic system. Phys Rev Lett 84:1240–1243
56. Schelling T (1971) Dynamic models of segregation. J Math Sociol 1:143–186
57. Helbing D, Platkowski T (2000) Self-organization in space and induced by fluctuations. Int J Chaos Theory Appl 5(4):47–62
58. Ando K, Oto H, Aoki T (1988) Forecasting the flow of people. Railw Res Rev 45(8):8–13 (in Japanese)
59. Smith RA, Dickie JF (eds) (1993) Engineering for crowd safety. Elsevier, Amsterdam
60. Drager KH, Løvås G, Wiklund J, Soma H, Duong D, Violas A, Lanérés V (1992) EVACSIM – A comprehensive evacuation simulation tool. In: The proceedings of the 1992 Emergency Management and Engineering Conference. Society for Computer Simulation, Orlando, pp 101–108
61. Ebihara M, Ohtsuki A, Iwaki H (1992) A model for simulating human behavior during emergency evacuation based on classificatory reasoning and certainty value handling. Microcomput Civ Engin 7:63–71
62. Ketchell N, Cole S, Webber DM, Marriott CA, Stephens PJ, Brearley IR, Fraser J, Doheny J, Smart J (1993) The EGRESS code for human movement and behaviour in emergency evacuations. In: Smith RA, Dickie JF (eds) Engineering for crowd safety. Elsevier, Amsterdam, pp 361–370
63. Okazaki S, Matsushita S (1993) A study of simulation model for pedestrian movement with evacuation and queuing, In: Smith RA, Dickie JF (eds) Engineering for crowd safety. Elsevier, Amsterdam, pp 271–280
64. Still GK (1993) New computer system can predict human behaviour response to building fires. Fire 84:40–41
65. Still GK (2000) Crowd dynamics. Ph.D. thesis, University of Warwick
66. Thompson PA, Marchant EW (1993) Modelling techniques for evacuation. In: Smith RA, Dickie JF (eds) Engineering for crowd safety. Elsevier, Amsterdam, pp 259–269
67. Løvås GG (1998) On the importance of building evacuation system components. IEEE Trans Engin Manag 45:181–191
68. Hamacher HW, Tjandra SA (2001) Mathematical modelling of evacuation problems: A state of the art. In: Schreckenberg M, Sharma SD (eds) Pedestrian and evacuation dynamics. Springer, Berlin, pp 227–266
69. Keating JP (1982) The myth of panic. Fire J 57–61, 147
70. Elliott D, Smith D (1993) Football stadia disasters in the United Kingdom: Learning from tragedy? Ind Env Crisis Q 7(3):205–229
71. Jacobs BD, 't Hart P (1992) Disaster at Hillsborough Stadium: A comparative analysis. In: Parker DJ, Handmer JW (eds) Hazard management and emergency planning, Chapt 10. James and James Science, London
72. Canter D (ed) (1990) Fires and human behaviour. Fulton, London
73. Mintz A (1951) Non-adaptive group behavior. J Abnorm Norm Soc Psychol 46:150–159
74. Miller DL (1985) Introduction to collective behavior (Fig. 3.3 and Chap. 9). Wadsworth, Belmont
75. Coleman JS (1990) Foundations of social theory, Chaps. 9 and 33. Belkamp, Cambridge
76. Johnson NR (1987) Panic at "The Who Concert Stampede": An empirical assessment. Soc Probl 34(4):362–373
77. LeBon G (1960) The crowd. Viking, New York
78. Quarantelli E (1957) The behavior of panic participants Sociol Soc Res 41:187–194
79. Smelser NJ (1963) Theory of collective behavior. Free Press, New York
80. Brown R (1965) Social psychology. Free Press, New York
81. Turner RH, Killian LM (1987) Collective behavior, 3rd edn. Prentice Hall, Englewood Cliffs
82. Bryan JL (1985) Convergence clusters. Fire J 27–30, 86–90
83. Axelrod R, Hamilton WD (1981) The evolution of cooperation. Science 211:1390–1396
84. Axelrod R, Dion D (1988) The further evolution of cooperation. Science 242:1385–1390
85. Glance NS, Huberman BA (1994) The dynamics of social dilemmas. Scientific American 270:76–81
86. Kelley HH, Condry JC Jr, Dahlke AE, Hill AH (1965) Collective behavior in a simulated panic situation. J Exp Soc Psychol 1:20–54

87. Helbing D, Johansson A, Al-Abideen HZ (2007) The dynamics of crowd disasters: An empirical study. Phys Rev E 75:046109
88. Fruin JJ (1993) The causes and prevention of crowd disasters. In: Smith RA, Dickie JF (eds) Engineering for crowd safety. Elsevier, Amsterdam, pp 99–108
89. Ristow GH, Herrmann HJ (1994) Density patterns in two-dimensional hoppers. Phys Rev E 50:R5–R8
90. Wolf DE, Grassberger P (eds) (1997) Friction, arching, contact dynamics. World Scientific, Singapore
91. Helbing D, Johansson A, Mathiesen J, Jensen HM, Hansen A (2006) Analytical approach to continuous and intermittent bottleneck flows. Phys Rev Lett 97:168001
92. Ghashghaie S, Breymann W, Peinke J, Talkner P, Dodge Y (1996) Turbulent cascades in foreign exchange markets. Nature 381:767–770
93. Peng G, Herrmann HJ (1994) Density waves of granular flow in a pipe using lattice-gas automata. Phys Rev E 49:R1796–R1799
94. Radjai F, Roux S (2002) Turbulentlike fluctuations in quasistatic flow of granular media. Phys Rev Lett 89:064302
95. Sreenivasan KR (1990) Turbulence and the tube. Nature 344:192–193
96. Cates ME, Wittmer JP, Bouchaud J-P, Claudin P (1998) Jamming, force chains, and fragile matter. Phys Rev Lett 81:1841–1844
97. Bak P, Christensen K, Danon L, Scanlon T (2002) Unified scaling law for earthquakes. Phys Rev Lett 88:178501
98. Johnson PA, Jia X (2005) Nonlinear dynamics, granular media and dynamic earthquake triggering. Nature 437:871–874
99. Johansson A, Helbing D (2007) Pedestrian flow optimization with a genetic algorithm based on Boolean grids. In: Waldau N, Gattermann P, Knoflacher H, Schreckenberg M (eds) Pedestrian and evacuation dynamics 2005. Springer, Berlin, pp 267–272
100. Baeck T (1996) Evolutionary algorithms in theory and practice. Oxford University Press, New York

Books and Reviews

Decicco PR (ed) (2001) Evacuation from fires. Baywood, Amityville
Helbing D (2001) Traffic and related self-driven many-particle systems. Rev Mod Phys 73:1067–1141
Helbing D, Molnár P, Farkas I, Bolay K (2001) Self-organizing pedestrian movement. Environ Plan B 28:361–383
Helbing D, Buzna L, Johansson A, Werner T (2005) Self-organized pedestrian crowd dynamics: Experiments, simulations, and design solutions. Transp Sci 39(1):1–24
Le Bon G (2002) The Crowd. Dover, New York (1st edn: 1895)
Predtechenskii VM, Milinskii AI (1978) Planning for foot traffic flow in buildings. Amerind, New Delhi
Schreckenberg M, Sharma SD (eds) (2002) Pedestrian and evacuation dynamics. Springer, Berlin
Smith RA, Dickie JF (eds) (1993) Engineering for crowd safety. Elsevier, Amsterdam
Still GK (2000) Crowd Dynamics. Ph.D thesis, University of Warwick
Surowiecki J (2005) The Wisdom of Crowds. Anchor, New York
Tubbs J, Meacham B (2007) Egress design solutions: A guide to evacuation and crowd management planning. Wiley, New York
Waldau N, Gattermann P, Knoflacher H (eds) (2006) Pedestrian and evacuation dynamics 2005. Springer, Berlin
Weidmann U (1993) Transporttechnik der Fußgänger. In: Schriftenreihe des Institut für Verkehrsplanung, Transporttechnik, Straßen- und Eisenbahnbau 90. ETH Zürich

Percolation in Complex Networks

Reuven Cohen[1], Shlomo Havlin[2]
[1] Department of Mathematics, Bar-Ilan University, Ramat-Gan, Israel
[2] Department of Physics, Bar-Ilan University, Ramat-Gan, Israel

Article Outline

Glossary
Definition of the Subject
Introduction
Percolation Thresholds and Network Robustness
Epidemics and Immunization
The Generating Functions Method
Critical Exponents and Fractal Dimensions
Optimal Paths and Minimum Spanning Trees
Fragmentation of Social Networks
Future Directions
Bibliography

Glossary

Graph A set of nodes (sites) and edges (links or bonds) connecting them.

Weighted graph A graph where each edge is assigned a (usually non-negative) weight.

Random graph A graph selected from an ensemble (probability space) of graphs.

Degree Number of edges emanating from a node.

Scale free network A network whose nodes' degrees are distributed according to a power law.

Shortest path The path with minimum number of edges connecting two nodes.

Optimal path In a weighted graph – the path with minimum total weight connecting two nodes.

Loop A path that start and ends at the same node.

Tree A connected graph (a graph consisting of a single component) with no loops.

Minimum spanning tree In a weighted graph – the tree subgraph of the graph with the minimum total weight.

Component The set of nodes reachable from a given node. The nodes of a component are all reachable from each other.

Giant component The component of a graph with size (number of nodes) of order of the number of nodes in the graph.

Percolation theory The theory studying the connectivity behavior of networks when a fraction of the nodes or edges are removed. Site (or node) percolation involves occupying a fraction, p, of the nodes of the graph, or alternatively, removing a fraction $q = 1 - p$. In bond (or edge) percolation edges are occupied, or removed, with some probability. A combined site-bond percolation, where both processes occur simultaneously, is also considered.

Percolation threshold The fraction, p_c of occupied nodes or edges, under the graph is fragmented into small components, and above which a giant component emerges.

Definition of the Subject

In this chapter we survey the application of percolation theory to several random network classes, and in particular, to scale free networks. We show how ideas from percolation theory can be applied to the study of robustness and vulnerability of random networks. We show how percolation techniques can be applied also to understand phenomena such as immunization and epidemic spreading in populations and computer networks, minimum spanning trees and communication paths and fragmentation in social networks.

Introduction

In recent years considerable interest has been given to real world networks. The importance of technological networks such as the Internet and WWW, as well as the availability of large scale data sets on social, biological and technological networks made this subject approachable and popular.

The main model used in the study of complex networks was Erdős–Rényi (ER) random graphs [19,20,21] (also presented earlier by Rapaport [33]), which are graphs having N nodes (that is, sites or entities) and M edges (links, or connections between the nodes) distributed randomly between them, or alternatively, the almost identical model, having every pair of nodes connect with a constant probability. One of the key discoveries in recent years was that many real world networks, including the Internet, WWW and many biological and social networks, are not described well by the ER model [1,18,28,32]. One of the first deviations from the model to be noticed was the tendency of many real world networks to have high clustering [28], i. e., neighbors of the same node tend to connect between them (i. e., share an edge). This discovery has lead to the presentation of the small world model [37].

The small world model is based on some lattice, such as a one-dimensional ring or a higher dimensional grid, in which rewiring occurs. Some (usually small) fraction, φ of the links in the lattice are removed, and instead, new links are added randomly between the nodes. When φ is moderately small the generated graphs have the desired properties of high clustering, while the average distance between nodes is small (of the order $\log N$), as in random graphs (as opposed to $N^{1/d}$ as in a d-dimensional grid).

The second deviation to be noticed was the deviation of the degree sequence from the expectation of random graph theory. The degree of a node is the number of links (or edges) emanating from it, i. e., the number of neighbors it has in the graph. The number of nodes of degree k in a graph will be denoted $n(k)$, and the degree distribution, i. e., the probability that a random node has degree k, is $P(k) = n(k)/N$. In an ER random graph the expected degree sequence is a Poisson distribution [5], $P(k) = \mathrm{e}^{-C} C^k/k!$, where $P(k)$ is the probability of a node to have degree k, and c is the average degree ($C = 2M/N$). In many real world networks, including the Internet and WWW, it was observed that the degree distribution is actually a power law,

$$P(k) = ck^{-\gamma}\,, \quad m \le k \le U\,, \tag{1}$$

where c is a normalization factor, m and U are the minimum and maximum degrees, respectively, and γ is some exponent characterizing the distribution. In most networks studied γ has been found to lay in the range $2 < \gamma < 3$ [1,18,28,32]. These networks have become known as scale free networks, due to the lack of typical scale for the degree distribution. It should be noted that while m must be supplied externally for the distribution to be normalizable, U can be omitted, and will be determined naturally as the extreme value statistics of N variables, which gives $U \sim mN^{1/(\gamma-1)}$ in this case [13].

Several models have been developed for the understanding and study of scale free networks. The question of the reason for these networks' formation has been addressed by the Barabási–Albert model [3]. Many variants on the model have been studied since (see, e. g., [23]). Here we will focus on the configuration model, or the generalized random graph model [4], which is an equilibrium model for random graphs with a given degree distribution, producing all graphs having a given degree distribution with uniform probability. The model starts by having N distinct nodes and randomly selecting the degree of each of these nodes from the given degree distribution. Each node, i is then fitted with k_i "stubs", where k_i is its degree, drawn

from the distribution $P(k)$. After all nodes' degrees have been selected, a random matching of the stubs is selected, by choosing random pairs of stubs and pairing them (i. e., connecting the nodes by an edge and removing both stubs) until no stubs are left. In some cases a single stub is left, and also there may be edges connecting a node to itself or more than one edge connecting a given pair of nodes. These cases can be safely ignored as they only have a small effect on the final graphs obtained.

Percolation Thresholds and Network Robustness

One of the fundamental questions regarding a network's structure is its connectivity properties, i. e., what are the properties of the distinct components (or clusters) of the network. A component of a graph is the set of nodes reachable from a single node by following edges in the graph. Notice that in an (undirected) graph, the property of path connectedness is symmetric and transitive, i. e., if a node a is reachable from node b, then the inverse path also exists, and if c is reachable from a it is also reachable from b. Therefore, a component is uniquely defined by any node belonging to it. A network is said to be connected if all nodes in it belong to a single component, that is, each node is reachable from each other node.

Random graphs are locally tree-like, i. e., the number of closing a loop for a set of less than order N nodes is negligible. This also implies that below the percolation threshold, where all components (clusters) are small, almost all components are trees, i. e., posses of no loops.

To determine the percolation threshold it should be noticed that at the critical point every node reached by following a link from a previously visited node should have, on average, exactly one more link through which new nodes can be reached. If the average number of outgoing links is less than one, the uncovering process of the component will quickly decay, and only small components will be present. If the average number of outgoing links if larger than one, the size of the largest component will be proportional to that of the entire graph, i. e., a giant component, of size $O(N)$ will exist. This may lead to the conclusion that the average degree, $\langle k\rangle$, needs be two or more for a giant component to exist. However, the node reached by following a link is not chosen uniformly. The probability of reaching a node by following a link is proportional to its degree, $P_i = k_i/(N\langle k\rangle)$. The average outgoing degree of a node reached by following a link is therefore, $\sum(k-1)n(k)P_i = \langle k(k-1)\rangle/\langle k\rangle = \kappa - 1$, where $\kappa = \langle k^2\rangle/\langle k\rangle$ is the ratio of the first two moments of the degree distribution. It is this quantity that should be compared to one to determine whether a giant component exists [13,26]. Therefore, a giant component exists if and only if $\kappa > 2$.

In a percolation setting the nodes or edges are removed with probability q, or, alternatively, occupied with probability $p = 1 - q$. This model may represent random failures of nodes in the network, such as random failures of routers or links in the Internet [2]. The average number of outgoing links should be multiplied then by the probability, p_b, that the link is occupied, and by p_s, the probability that the node reached from the link is occupied, or not deleted. Therefore, the condition for the existence of a giant component becomes $p_b p_s(\kappa - 1) > 1$. Alternatively, p_c, the critical site or bond percolation threshold is given by [13]

$$p_c = \frac{1}{\kappa - 1} \,. \tag{2}$$

Notice that the critical threshold depends only on the first two moments of the degree distribution. Furthermore, in a scale free network with $\gamma \leq 3$ the second moment of the degree distribution diverges in the limit of infinite network size. Therefore, the critical threshold for this class of networks approaches 0, indicating that these networks are resilient to any finite fraction of random node failures [13]. Figure 1 presents the results of random node failure on the Internet as compared to an ER network.

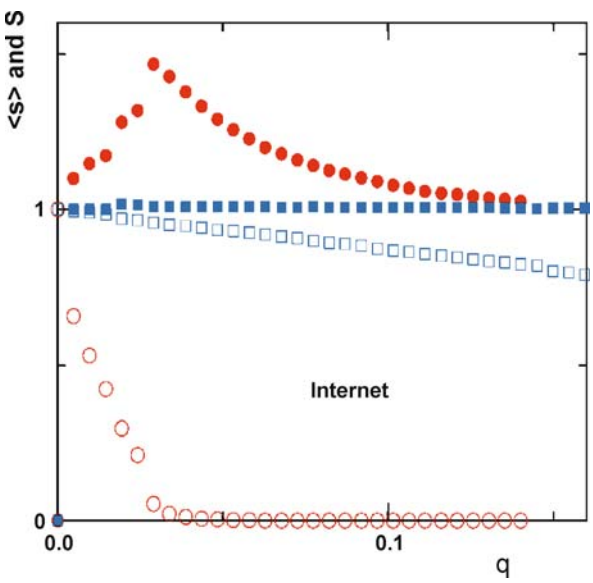

Percolation in Complex Networks, Figure 1
Results of targeted removal of a fraction q of the nodes from an ER graph (*circles*) and a partial Internet view (*squares*). *Full symbols* represent the size of the second largest component and *empty symbols* represent P_∞, the relative size of the largest component. After [2]

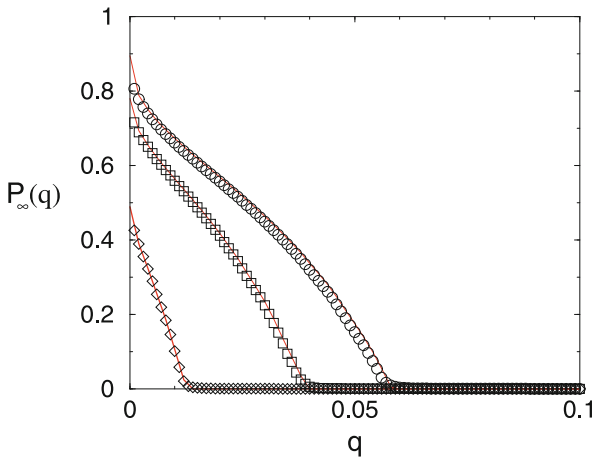

Percolation in Complex Networks, Figure 2
The relative size of the largest component, P_∞, as a function of the fraction of tergeted removed nodes, for scale free networks with $m = 1$ and $\gamma = 2.5$ (*circles*), $\gamma = 2.8$ (*squares*), and $\gamma = 3.3$ (*diamonds*). After [14]

In case the node removal is not random this situation may change drastically. The most well studied case is that of removal of the highest degree nodes, modeling an intentional, targeted attack on the most important nodes in the network. In this case calculations similar to the above lead to the conclusion that the percolation threshold is finite and small [2,9,14]. Figure 2 illustrates the results of targeted removal of a fraction q in scale free networks.

Epidemics and Immunization

The spread of epidemics in a population can be modeled as a dynamical process in a network. Each node represents an individual and the links represent interaction between individuals that allows transmission of the epidemic. Several models exist for epidemic transmission, depending on the type of epidemic. The most commonly used models are the Susceptible-Infected-Susceptible (SIS) model and the Susceptible-Infected-Removed (SIR) model. In both models it is assumed that nodes in the susceptible state are susceptible to the epidemic, i. e., may be infected when they come in contact with an infected individual. In the infected state the individual is infected by the epidemic and may infect other individuals, and in the removed (or recovered) state the individual is no more infected or infective and also is no longer susceptible to the disease. This state may occur due to recovery from the disease while the individual remains immunized against the disease or due to death of the infected individual. The SIS model assumes that recovered individuals are again susceptible and the SIR model assumes that each individual may only be infected once in a lifetime.

In the SIS model an epidemic can either quickly decay and vanish or prevail for a long period, during which a large (finite) fraction of the population is infected. In the SIS model it was shown that in scale free networks with $\gamma \leq 3$ the epidemic always prevails, regardless of the infection rate [30].

Consider epidemic spreading in the SIR model. Assume each infected node infects each of its neighbors with rate R, and has a constant infection time of T. For each neighbor, the probability of the neighbor being infected by an infected node is $p = 1 - e^{-RT}$. This can be viewed as the probability of the edge between the two nodes is occupied, i. e., can actually be used for transmission of the epidemic. Therefore, the SIR model can be mapped to a bond percolation model (Notice that in non homogeneous cases of node dependent rates a more complicated model is needed. See [27].)

The SIR model can therefore be solved by solving the bond percolation problem in the network [22,27,34]. Every edge is occupied with probability p, and the epidemic can reach an endemic state (i. e., infect $O(N)$ nodes with finite probability) if $p > p_c$, and will quickly decay, infecting only a negligible portion of the population if $p < p_c$. Furthermore, the distribution of the size of the epidemic outbreak is determined by the sizes of the graph components. If a giant component exists, the probability of a single infected individual to induce an endemic state of the population is P_∞, the size of the giant component, and the size of the outbreak is the size of the component to which this individual belongs.

To prevent epidemic outbreaks it is usually desirable to immunize the population and thus prevent the epidemic. In many cases it is difficult to immunize the entire population, and only a fraction is immunized. Each immunized individual is no longer susceptible to the disease, and can be viewed as removed from the network. The immunization process can therefore be viewed as a site percolation process, where each node is removed with probability q, or occupied with probability $p = 1 - q$. The epidemic progression can then be mapped into a site-bond percolation problem.

In order for the immunization to be highly efficient, it is desirable to surpass the percolation threshold in the immunization process, to ensure that the epidemic can not reach an endemic state. Since, as stated above, randomly immunizing a fraction of the population can be a highly inefficient process, requiring immunization of nearly 100% of the population, it has been suggested that a more efficient method for immunization is devised. The

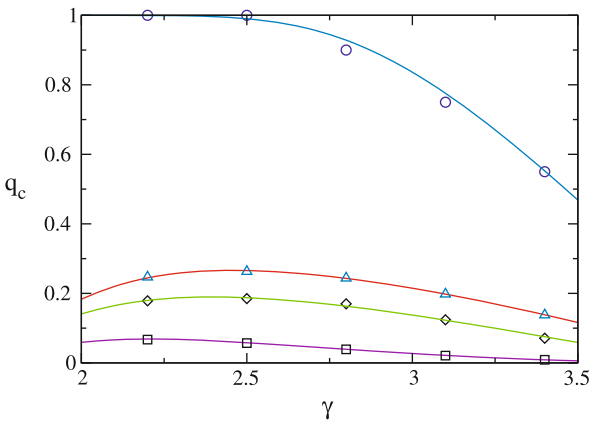

Percolation in Complex Networks, Figure 3
Critical immunization threshold, q_c, as a function of γ in scale-free networks (with $m = 1$), for the random immunization (∘), acquaintance immunization (△), double acquaintance immunization (⋄), and targeted immunization (□) strategies. Curves represent analytical results, while data points represent simulation data, for a population $N = 10^6$ (Due to the population's final size $q_c < 1$ for random immunization even when $\gamma < 3$). After [16]

simplest such method involves the immunization of the highest degree nodes in the population [31]. In this case immunizing a population will require vaccinating only a finite, and relatively small, fraction of the population.

In case only partial knowledge of the population exists it is sometimes also possible to immunize the population efficiently [17]. However, a different method, requiring no global knowledge exists. In this method, "acquaintance immunization", a fraction of randomly selected individuals are requested to point to one of their contacts, also randomly selected. The pointed contacts are then immunized. Although this is a seemingly random process a node having a high degree is immunized with very high probability, and the process behaves effectively as a targeted immunization of high degree nodes. See [16] for analytical treatment and Fig. 3 for illustration of the various immunization thresholds.

The Generating Functions Method

To allow the calculation of different properties of random networks it was proposed in [9,29] to use the generating function formalism (see, e. g., [38]). In this formalism a list of numbers A_i is treated as the coefficients of a formal power series $A(x) = \sum_i A_i x^i$. This treatment simplifies many equations regarding the variables A_i, and, in many cases, simplifies the calculation of the asymptotic behavior of the coefficients for large i.

In [9,29] a power series is built for the degree distribution,

$$G_0(x) = \sum_k P(k) x^k \,, \tag{3}$$

and for the distribution of outgoing links from a node reached by following a link,

$$G_1(x) = \sum_k \frac{kP(k)}{\langle k \rangle} x^{k-1} = \frac{G_0'(x)}{G_0'(1)} \,, \tag{4}$$

where $G_0' = \mathrm{d}G_0/\mathrm{d}x$.

A branch in the network is a link, traversed in one direction, and all the nodes reachable by following this link in this direction. This includes the node reached by following this link, and all the nodes reached by the branches emanating from the outgoing links of this node. An illustration of this recursive definition is in Fig. 4. A branch may be either finite or infinite, in which case, for a finite graph, it will reach $O(N)$ nodes and have many loops (in which case the branch description is no longer useful). A link and the node reached by following it are occupied with probability $p_b p_s$, and the generating function for the number of descendants of the reached node is G_1. Each of the descendants is a new branch. Therefore, the generating function for the size of a branch is given by

$$H_1(x) = 1 - p_s p_b + p_s p_b G_1(H_1(x)) \,. \tag{5}$$

A component in the graph is a node, and all the nodes reachable from it. Each of the links of a node leads to a branch. The generating function for the degree of a node is G_0, and the probability it is occupied is p_s. Therefore, the generating function for the component size distribution is,

$$H_0(x) = 1 - p_s + p_s G_0(H_1(x)) \,, \tag{6}$$

where H_1 is determined by Eq. (5). Again, as above, a component may be finite or infinite. Since H_0 is the generating function for the finite cluster distribution, its normalization $H_0(1)$ is the total probability that a node belongs to a finite component, and therefore its complement

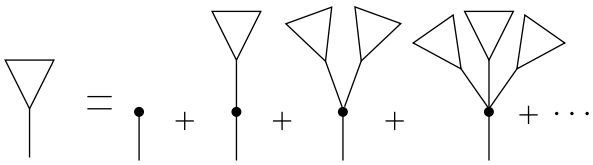

Percolation in Complex Networks, Figure 4
Illustration of the structure of a branch. A branch can contain either a link leading to a node with no outgoing links, or to a node having one or more outgoing links, each leading to another branch. After [29]

$1 - H_0(1)$ is the probability that a node belongs to the giant component. Thus, the size of the giant component (that is, the fraction of nodes belonging to the giant component) is given by

$$P_\infty = 1 - H_0(1) = 1 - p_s + p_s G_0(u) \,, \tag{7}$$

and $u = H_1(1)$ is given by the self consistent equation derived from Eq. (5)

$$u = 1 - p_s p_b + p_s p_b G_1(u) \,. \tag{8}$$

Critical Exponents and Fractal Dimensions

Percolation problems, as well as other critical phenomena, are known to present a universal behavior near and at the critical point. That means that many properties of the critical structure behave as power laws and that near the critical point many sizes behave as powers of $p - p_c$. The universality is pronounced by the fact that the exponents in the different power laws do not depend on the microscopic details of the problem, but only on the large scale details, in particular, the dimension of the space and the symmetries. In percolation, for example, it is known that slightly above the critical point the size of the giant component behaves as $P_\infty \sim (p - p_c)^\beta$ and that the number of components of size s at criticality decays as $n(s) \sim s^{-\tau}$. Both β and τ depend only on the dimensionality and not on the microscopic details (e. g., are the same for two dimensional square and triangular lattices, etc.).

Networks can be considered infinite dimensional objects. As mentioned above, the number of nodes at a distance (number of hops) at most ℓ from a node behaves as A^ℓ for some $A > 1$. For large values of ℓ this is larger than ℓ^d, obtained for any finite dimension d. This property also leads to the impossibility of embedding networks in any finite dimension.

For percolation theory it is known [8,35] that the upper critical dimension is 6, i. e., percolation on grids of any dimension larger than 6 behave similarly to percolation in infinite dimension (also known as mean field percolation, and usually studied using the Cayley tree model [8]). In mean field percolation it is know that $\beta = 1$ and $\tau = 2.5$. For dimensions less than 6 it is known that $\beta < 1$ and that $\tau < 2.5$.

To determine the critical exponents for random networks one can use the generating function formalism presented above. The size of the giant component is given by Eq. (7), where the value of u in this equation is the solution of Eq. (8). At and below the critical concentration, p_c, the size of the giant component is $P_\infty = 0$, since there are only finite components. This corresponds to a solution in which $u = 1$ and $H_0(1) = 1$. At $p = p_c + \delta$ it is expected that u is close to 1, i. e., $u = 1 - \epsilon$. Substituting this into Eq. (8) yields

$$1 - \epsilon = 1 - p_c - \delta + (p_c + \delta) G_1(1 - \epsilon) \,. \tag{9}$$

Expanding G_1 into a power series yields

$$\epsilon = (p_c + \delta)(1 - G_1(1) - G_1'(1)\epsilon - G_1''(1)\epsilon^2/2 - \cdots) \,. \tag{10}$$

This equation is self consistent and gives a non trivial solution only when $p_c = G_1'(1) = 1/(\kappa - 1)$ as obtained above. The solution obtained then is $\epsilon \propto \delta$. The solution of Eq. (7) then is $P_\infty \propto \delta = (p - p_c)$, similar to infinite dimensional percolation. Using similar expansions of Eqs. (7) and (8), now at the critical point, $p = p_c$ it can be shown that the probability of a node to belong to a component of size s is proportional to, $p_s \propto s^{-3/2}$, implying that the number of components of size s behaves as $n(s) \propto s^{-5/2}$. Both these exponents are the same as obtained for infinite dimension percolation.

The above treatment is correct, however, only assuming the sums in Eqs. (7) and (8) can be expanded in a power series. This is true only if the degree distribution, $P(k)$ decays quickly enough, say exponentially. In the case of a power law degree distribution, the series expansion is incorrect, and one needs to resort to other methods of obtaining the asymptotic behavior. The main mechanism for obtaining such asymptotics is using Abelian and Tauberian theorems. These theorems relate the decay of the coefficients of a power series and its behavior near a singular point of the function in the complex plane. Using these methods it can be shown [12] that the behavior of Eq. (8) near criticality becomes

$$\begin{aligned} 1 - \epsilon = 1 - p_c - \delta + (p_c + \delta)(G_1(1) + G_1'(1)\epsilon \\ + G_1''(1)\epsilon^2/2 + \cdots + C\epsilon^{\gamma-1} + \cdots \,. \end{aligned} \tag{11}$$

Thus, when $\gamma < 3$ the nonanalytic term $C\epsilon^{\gamma-1}$ dominates the linear term, and for $3 < \gamma < 4$ it dominates the quadratic term.

In the most interesting case of $3 < \gamma < 4$ the percolation threshold is finite, as seen from Eq. (2). However, using the expansion in Eq. (11) it can be seen that near the critical point $P_\infty \propto (p - p_c)^{1/(\gamma-3)}$, so $\beta = 1/(\gamma - 3)$. Similarly, it can be shown that $\tau = (2\gamma - 3)/(\gamma - 2)$ in this regime. Both exponents return to their mean field value for $\gamma > 4$.

Another common characteristic of critical phenomena is the fractal behavior at the critical point. For high dimensional percolation at the critical dimension $d_c = 6$ and above it is known that the fractal dimension of the

largest components is $d_f = 4$. This implies that the size of the largest component behaves as $S \sim L^4$, where L is the linear dimension of the grid. Since the total grid size is $N \sim L^6$ it follows that $S \sim N^{2/3}$. Also, it is known that each branch of the components at criticality behaves as an independent random walk. Since random walks have $\ell \sim L^2$, where ℓ is the number of random walk steps, representing the distance on the fractal itself, also known as the chemical distance. This implies that $S \sim \ell^2$. This dimension, $d_l = 2$ is known as the chemical dimension [23], and since in a network no embedding space is present, and therefore L is not well defined, the chemical dimension is the most appropriate measure to be used.

To deduce the critical dimensions of a percolating network one can observe the survivability of a branch in the network. Define

$$F(x) = 1 - p_c + p_c G_1(x) , \tag{12}$$

to be the degree distribution at criticality, i. e., the distribution of the number of occupied links leading to occupied nodes. Denote the number of nodes at a distance ℓ along a branch by N_ℓ. The distribution of such values can be fitted with a generating function $N_\ell(x)$. The generating functions for different layers satisfy

$$N_{\ell+1}(x) = F(N_\ell(x)) . \tag{13}$$

At criticality the average number of nodes on the ℓth layer along a branch is 1, since a lower branching factor will lead to fast extinction of all branches, and a higher branching factor will give an infinite branch with a high probability. Therefore, the number of nodes at a distance ℓ only along branches that survive at least ℓ layers equals the average number of nodes in the ℓth layer, which is 1, divided by the fraction of branches surviving at least ℓ layers. This number can be obtained by noticing that the probability of a branch to become extinct at the first ℓ layers is given by $N_\ell(0)$, the probability of having 0 nodes in the ℓth layer. The average number of nodes at the ℓth layer of surviving branches is therefore $m_\ell = 1/(1 - N_\ell(0))$, and the fractal dimension can be obtained from the total number of nodes up to the ℓth layer in surviving branches, $M_\ell = \sum_{i=1}^{\ell} m_\ell$. The asymptotic behavior of $N_\ell(0)$ can again be obtained and leads to $M_\ell \propto \ell^{(\gamma-2)/(\gamma-3)}$ for $3 < \gamma < 4$ and $M_\ell \propto \ell^2$ for $\gamma > 4$ and for ER networks. This implies

$$d_l = \begin{cases} \frac{\gamma-2}{\gamma-3} , & 3 < \gamma < 4 , \\ 2 , & \gamma > 4 . \end{cases} \tag{14}$$

The size of the largest component at criticality can be obtained using the other critical exponents. As presented above, the asymptotic behavior of the component size distribution is $P(s) = s^{-\tau+1}$, where $P(s)$ is the probability of a node to belong to a component of size s. There are N nodes in the graph, and therefore it is expected that the size of the largest component of a graph, S, will be such that $P(S) \sim N^{-1}$. This leads to $S \sim N^{1/(\tau-1)}$ and to

$$S \propto \begin{cases} N^{(\gamma-2)/(\gamma-1)} , & 3 < \gamma < 4 , \\ N^{2/3} , & \gamma > 4 . \end{cases} \tag{15}$$

Optimal Paths and Minimum Spanning Trees

Communication in a network usually follows the shortest path from source to destination. The network structure and function usually depends only weakly on the space the network exists in. Hence, the path length is usually defined by the intrinsic properties of the network. The simplest definition of the path length between two nodes, a and b, in a network is the hop distance, i. e., the minimum number of links that need to be traversed in order to arrive from a to b. The average path length in networks has been studied extensively, and is known to be logarithmic in the size of the network for ER [5] and scale free graphs with $\gamma > 3$ [29] and of order $\log\log N$ in scale free graphs with $\gamma < 3$ [15].

Another definition for the distance in a network can be given in case each link has some intrinsic length, or some intrinsic property (similar to energy in physics), termed "weight", measuring the cost of using it. When the link length distribution is narrow (the "weak disorder" limit) the behavior of the optimal (lowest cost) path is expected to be very similar to that of the shortest path. However, when the distribution of link costs becomes very wide (the "strong disorder" limit) the behavior of the optimal path becomes very different from that of the shortest path [12,38].

The limit of strong disorder is observed clearly when the weight distribution is so wide, that the weight of each link is expected to be at least twice as large as the next highest link weight. This implies also that the weight of each link is larger than that of *all* links with lower weight. In this case paths can be compared by sorting the link weights along each path, and then comparing the lists of link weights by lexicographic order. It should be noted that the graph of all shortest paths also called "optimal paths" is a tree, i. e., has no loops. It is similar to the minimum spanning tree (MST), which presents the same behavior regardless of the weight distribution. Another similar case is high bandwidth information transmission in communication networks, where it may be desirable to transmit

through the path having the highest *minimal* bandwidth, to avoid bottlenecks.

An alternative method for reaching the optimal path tree is the bombarding method. In this method the links in the network are removed one by one by order of decreasing weight. It is clear that removing a high weight link will not change any of the optimal paths, unless it makes the network disconnected, in which case it is not removed. Since the order of the link weights is random, so is the order of removal, and therefore this model is identical to random percolation, with the difference of refraining from removing links that make the network disconnected. Since at criticality percolation disintegrates the network into a collection of trees (or almost trees), the critical percolation component (cluster) is a subgraph of the optimal path tree.

This mapping of the optimal path and minimum spanning tree to a (restricted) percolation problem is very useful, since it allows the determination of the properties of these objects, based upon their similarity to the network at the critical point. From the results above in Sect. "Critical Exponents and Fractal Dimensions" the size of the largest component in ER networks is $S \propto N^{2/3}$. The chemical dimension is $d_l = 2$. Therefore, $S \propto \ell^2$, leading to the average hop distance between nodes on the critical components being $\ell \propto N^{1/3}$. Similarly for scale free networks with $3 < \gamma < 4$, $\ell \propto N^{(\gamma-3)/(\gamma-1)}$. This presents a lower bound on the length of the optimal path in ER and scale free networks [7]. In fact, it is observed that the critical components are connected in a compact way i. e., through a small number of components in the optimal path tree (or MST) and this lower bound is actually exact. See Fig. 5.

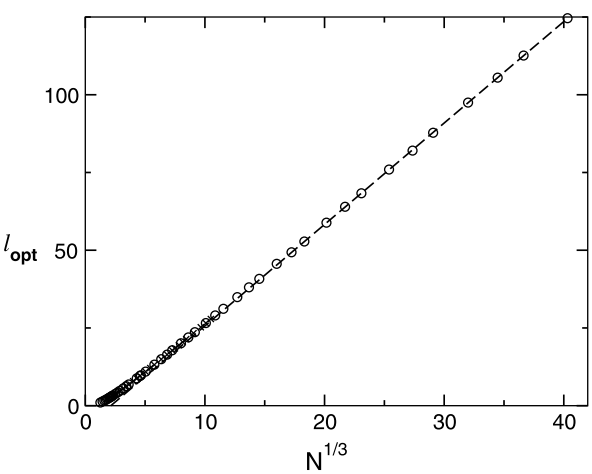

Percolation in Complex Networks, Figure 5
The optimal distance l_{opt} as a function of $N^{1/3}$ for ER graphs with strong disorder. After [7]

Fragmentation of Social Networks

One of the most interesting questions in sociological networks is quantifying the collapse process of a network. Under certain circumstances it may happen that a network of friendship or acquaintance is fragmented into several components. It is desirable in many cases to quantify the fragmentation using a measure that is sensitive to the different possible partitions into fragments. Such a measure, developed in [6] can distinguish between different partitions of a network, based on the sizes of all components. This is especially important in small and medium size networks, and in non random fragmentation processes, where the fragments may contain several similarly sized components.

Notice that the phenomenon of network fragmentation due to link removal (acquaintance separation) or node removal (individuals leaving the social network) is similar to bond and site percolation, respectively. One of the main shortcomings of the percolation description in this case is the focus of percolation theory on random processes and in the limit of large system size (the "thermodynamic limit"). Here we will discuss the proposed fragmentation measure and its relation to the standard percolation measures.

The suggested measure of fragmentation, F, is the number of pairs of nodes not reachable from each other, divided by the total number of pairs in the network [6]. Since nodes that are reachable from each other belong to the same component, this is equivalent to the following definition:

$$F = 1 - \frac{\sum_{i=1}^{n} s_i(s_i - 1)}{N(N-1)} \equiv 1 - C\,, \tag{16}$$

where s_i is the size of the ith component and n is the number of components in the graph.

In [10] the relation between percolation theory and this fragmentation measure was studied. As discussed above, in the limit of $N \to \infty$ above the percolation threshold there is a large gap between the size of the giant component, which is of order N and the size of the second largest component, which is usually of logarithmic size. At the threshold, the size of the largest component is of order $N^{2/3}$ for ER networks and some power of N for scale free networks, and the component size distribution is continuous. Below the percolation threshold the distribution is continuous again. However, the size of the largest component is even smaller (logarithmic in ER networks). Equation (16) therefore can be presented in the following equivalent form in the limit of large N

$$F \approx 1 - P_\infty^2 - \frac{\sum_s n(s)s(s-1)}{N^2} \approx 1 - P_\infty^2 - \frac{\langle s \rangle}{N}\,. \tag{17}$$

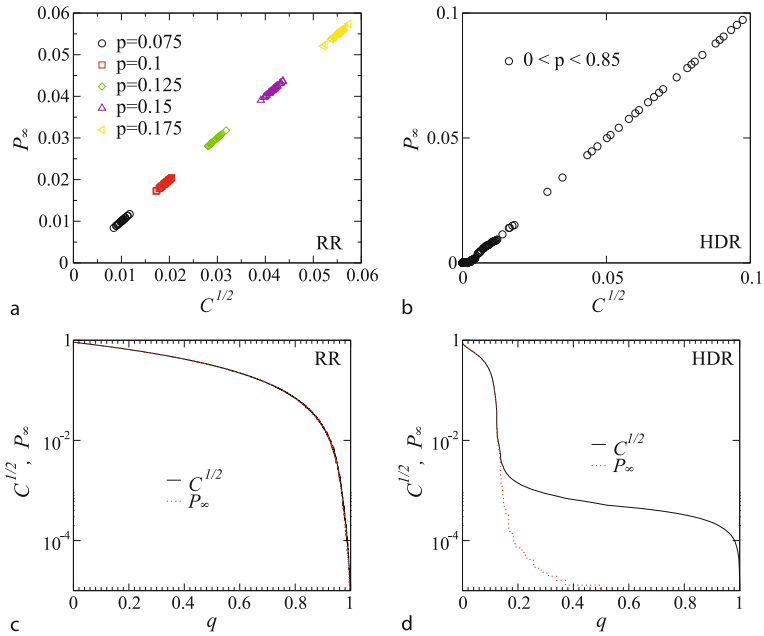

Percolation in Complex Networks, Figure 6

Comparison of P_∞ and $C \equiv 1 - F$ for random removal (a and c) and high degree removal (b and d) of links in a real social network of working relations in Sweden. After [10]

Notice that $\langle s \rangle$ represents the average component sampled over all nodes, rather then over all clusters. This gives a larger weight to larger components. In fact, since $\tau = 2.5$ in ER networks, and $\tau < 3$ for all $\gamma > 3$ and that $P(s)$, the probability of a node to belong to a component of size s scales as $P(s) \sim s^{-\tau+1}$, it follows that $\langle s \rangle$ diverges similarly to S at the threshold.

For finite networks in particular F has the advantage the it presents some measure of the fragmentation process both above and below the critical threshold. A comparison of P_∞ and F as measures of fragmentation can be seen in Fig. 6.

Future Directions

Different types of percolation can be defined and studied on random networks. Random percolation has special features in terms of the threshold and the critical exponents in scale free networks. The percolation theory of networks has many applications in epidemiology, network robustness, social networks analysis, and communication network efficiency.

Several open questions still remain regarding percolation theory in networks and its applications. For scale free networks with $2 \leq \gamma \leq 3$ it seems that a phase transition still exists with a threshold that approaches zero as a function of N. Although some progress has been made in this direction (see, e. g., [25]), the nature of this transition is not completely well understood and understanding it may enable the understanding of properties such as the optimal path behavior in such networks (which, from numerical results, seems to behave logarithmically).

Other important topics, which are not yet fully understood, is the question of optimal network design, and optimized attack strategies. In optimal network design, an attempt is made to find the parameters (such as degree distribution, correlations, or clustering) that produce a network class, with optimal percolation properties, including minimum random or targeted percolation thresholds, good near critical behavior, etc. Optimized attack strategies attempt to bring the network to the percolation threshold (or any other desired point) with the minimum number of removed nodes or links. Targeted attacks are usually very efficient in achieving percolation. However, more efficient methods can be conceived.

Bibliography

Primary Literature

1. Albert R, Barabási AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
2. Albert R, Jeong H, Barabási AL (2000) Error and attack tolerance of complex networks. Nature 406:378–382

3. Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512
4. Bollobás B (1980) A probabilistic proof of an asymptotic formula for the number of labeled regular graphs. Eur J Comb 1:311–316
5. Bollobás B (1985) Random graphs. Academic Press, London
6. Borgatti SP (2006) Identifying sets of key players in a network. Comput Math Organ Theor 12:21–34
7. Braunstein LA, Buldyrev SV, Cohen R, Havlin S, Stanley HE (2003) Optimal paths in disordered complex networks. Phys Rev Lett 91:168701
8. Bunde A, Havlin S (1996) Fractals and disordered system. Springer, New York
9. Callaway DS, Newman MEJ, Strogatz SH, Watts DJ (2000) Network robustness and fragility: percolation on random graphs. Phys Rev Lett 85:5468–5471
10. Chen Y, Paul G, Cohen R, Havlin S, Borgatti SP, Liljeros F, Stanley HE (2007) Percolation theory applied to measures of fragmentation in social networks. Phys Rev E 75:046107
11. Cieplak M, Maritan A, Banavar JR (1999) Optimal paths and growth process. Physica A 266:291–298
12. Cohen R, ben-Avraham D, Havlin S (2003) Percolation Critical Exponents in Scale Free Networks. Phys Rev E 66:036113
13. Cohen R, Erez K, ben-Avraham D, Havlin S (2000) Resilience of the Internet to Random Breakdown. Phys Rev Lett 85:4626–4628
14. Cohen R, Erez K, ben-Avraham D, Havlin S (2001) Breakdown of the Internet under Intentional Attack. Phys Rev Lett 86:3682–3685
15. Cohen R, Havlin S (2003) Scale free networks are ultrasmall. Phys Rev Lett 90:058701
16. Cohen R, Havlin S, ben-Avraham D (2003) Efficient Immunization Strategies for Computer Networks and Populations. Phys Rev Lett 91:247901
17. Dezső A, Barabási AL (2002) Halting viruses in scale free networks. Phys Rev E 65:055103
18. Dorogovtsev SN, Mendes JFF (2003) Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford University Press, Oxford
19. Erdős P, Rényi A (1959) On random graphs. Publ Math 6:290–297
20. Erdős P, Rényi A (1960) On the evolution of random graphs. Publ Math Inst Hung Acad Sci 5:17–61
21. Erdős P, Rényi A (1961) On the strength of connectedness of a random graph. Acta Math Sci Hung 12:261–267
22. Grassberger P (1983) On the critical behavior of the general epidemic process and dynamical percolation. Math Biosci 63:157–172
23. Krapivsky PL, Redner S (2002) A Statistical Physics Perspective on Web Growth. Comput Netw 39:261–276
24. Havlin S, Nossal R, (1984) Topological properties of percolation cluster. J Phys A 17:L427–L432
25. Lee DS, Goh KI, Kahng B, Kim D (2004) Evolution of scale-free random graphs: Potts model formulation. Nucl Phys B 696:351–380
26. Molloy M, Reed B (1995) A critical point for random graphs with a given degree sequence. Random Struct Algorithms 6:161–179
27. Newman MEJ (2002) The spread of epidemic disease on networks. Phys Rev E 66:016128
28. Newman MEJ (2003) Structure and function of complex networks. SIAM Rev 45:167–256
29. Newman MEJ, Strogatz SH, Watts DJ (2001) Random graphs with arbitrary degree distributions and their applications. Phys Rev E 64:026118
30. Pastor-Sattoras R, Vespignani A (2001) Epidemic spreading in scale free networks. Phys Rev Lett 86:3200–3203
31. Pastor-Sattoras R, Vespignani A (2001) Epidemic dynamics and endemic states in complex networks. Phys Rev E 63:066117
32. Pastor-Satorras R, Vespignani A (2003) Evolution and Structure of the Internet: A Statistical Physics Approach. Cambridge University Press, Cambridge
33. Rapoport A (1957) A contribution to the theory of random and biased nets. Bull Math Biophys 19:257–271
34. Sander LM, Warren CP, Sokolov IM, Simon C, Koopman J (2002) Percolation on heterogeneous networks as a model for epidemics. Math Biosci 180:293–305
35. Stauffer D, Aharony A (1991) Introduction to percolation theory. Taylor and Francis, London
36. Schwartz N, Porto M, Havlin S, Bunde A (1999) Optimal path in weak and strong disorder. Physica A 266:317–321
37. Watts DJ, Strogatz SH (1998) Collective dynamics of "small world" networks. Nature 393:440–442
38. Wilf HS (1994) Generatingfunctionology. Academic Press, San Diego

Books and Reviews

Albert R, Barabási AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97

Barabási AL (2002) Linked: The New Science of Networks. Perseus, Cambridge

Bornholdt S, Schuster HG (2002) Handbook of Graphs and Networks. Wiley-VCH, Berlin

Cohen R, Havlin S (2008) Complex Networks: Structure, Stability and Function. Cambridge University Press, Cambridge (in press)

Dorogovtsev SN, Mendes JFF (2002) Evolution of networks. Adv Phys 51:1079–1187

Dorogovtsev SN, Mendes JFF (2003) Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford University Press, Oxford

Newman MEJ (2003) Structure and function of complex networks. SIAM Rev 45:167–256

Newman MEJ, Barabási AL, Watts DJ (2006) The Structure and Dynamics of Networks. Princeton University Press, Princeton

Pastor-Satorras R, Vespignani A (2003) Evolution and Structure of the Internet: A Statistical Physics Approach. Cambridge University Press, Cambridge

Percolation, and Faults and Fractures in Rock

Pierre M. Adler[1], Jean-François Thovert[2], Valeri V. Mourzenko[2]

[1] UPMC-Sisyphe, Paris, France

[2] CNRS-LCD, Chasseneuil du Poitou, France

Article Outline

Glossary
Definition of the Subject
Introduction
Fracture Networks
Percolation of Fracture Networks
Determination of the Dimensionless Density from Experimental Data
Role of the Dimensionless Density in Other Geometrical Properties and Permeability
Future Directions
Bibliography

Glossary

Dimensionless density The dimensionless density is the number of objects per excluded volume.

Excluded volume The *excluded volume* V_{ex} of an object is defined as the volume surrounding it, in which the center of another object must be in order for them to intersect.

Fracture network A fracture network is generally defined as a set of individual fractures which may or may not intersect.

Percolation and percolation threshold Percolation is defined as the existence of a spanning connected cluster in the fracture network. Percolation occurs when the number of fractures per unit volume is equal or larger than a certain value called the percolation threshold.

Plane convex fractures A plane fracture is convex if for any points A and B which belong to the fracture, all the points of the segment AB belong to the fracture.

Definition of the Subject

The study of fractured porous media is of great practical and theoretical importance. It has been first generated by the fact that the presence of fractures can change completely the macroscopic properties of porous media which are present for instance in oil reservoirs, aquifers or waste repositories. The first contributions to this subject were made from very different standpoints by Barenblatt and coworkers [7], Conrad and Jacquin [13], and Witherspoon and coworkers (see for instance [22]).

In the eighties, these studies were renewed by concepts such as percolation and fractals. Fracture networks were first addressed in the framework of continuum percolation by [12] and [4].

Since the mid nineties, important progress have been made in this field thanks to systematic numerical experiments which can be rationalized by using the concept of excluded volume.

Introduction

Knowledge of geometrical properties of fracture networks is crucial to the understanding of flow and other transport processes in geological formations, both at small and large scales; introduction of fractures in a porous rock matrix seriously alters the macroscopic properties of the formation. Moreover, studies of fracture geometries during the last 30 years show that naturally occuring geological fractures exist on scales ranging from a few mm to several kilometers [33]. Therefore, fracture networks are likely to influence the transports on a large range of scales. Because of their importance, fracture networks are studied and applied in various areas such as oil and gas recovery, hydrology, nuclear waste storage and geothermal energy exploitation.

Geological fractures can be defined as discrete discontinuities within a rock mass; these breaks are characterized by the fact that their local aperture (defined as the local distance between the two surfaces which limit the fracture) is significantly smaller than their lateral extent; in other words, when they are viewed from far away, fractures can be assimilated to surfaces of discontinuity; in most cases, these surfaces are relatively plane. Fractures have varying degrees of aperture, and may in some cases be completely closed either because of deposition of material induced by fluid flow, or by displacements of the matrix.

An important property of these fracture sets or fracture network is their connectivity and their percolation properties. If a network percolates, fluid can circulate only through it and most likely much more rapidly than in the surrounding porous medium itself. Connectivity studies of fracture networks were initiated in 3D by Charlaix et al. [12] and Balberg [4].

The purpose of this paper is to provide a complete and updated view of the percolation properties of fracture networks in rocks. It is organized as follows. In Sect. "Fracture Networks", fractures are modeled as plane convex polygons which enables the introduction of the concept of excluded volume V_{ex}. This volume is a simple function of the surface and perimeter of the fractures, and it enables to introduce a dimensionless fracture density ρ' which is defined as the number of fractures per excluded volume. The tools necessary for the numerical study of the percolation thresholds are detailed and applied to mono- and poly- disperse fracture networks. It is shown that when expressed in terms of ρ', the percolation threshold does not

depend anymore on the fracture shapes. This crucial property is presented and discussed.

Section "Determination of the Dimensionless Density from Experimental Data" is devoted to the determination of the dimensionless density from experimental data. In most cases, these data are based on 1D and 2D measurements of fracture traces along boreholes or on exposed outcrops. These measurements necessitate extrapolation by stereological techniques to three dimensions. Significant progress can be made for plane convex fractures. Some recent applications of the methodology are given.

Finally, the independence of the dimensionless percolation threshold on the fracture shape can be extended to other properties such as other geometric properties and the macroscopic permeability of fractured rocks. These extensions are summarized in Sect. "Role of the Dimensionless Density in Other Geometrical Properties and Permeability".

Fracture Networks

A fracture network is generally defined as a set of individual fractures which may or may not intersect.

On a scale large with respect to the fracture aperture, fractures are usually modeled as convex, finite polygons possibly based on an embedding disk as shown in Fig. 1a. This is only a simplifying assumption which however provides a standard starting point for studying fracture networks. Convex polygons can be used to analyze shape and area dependencies of geometrical and topological features in the fracture systems in a systematic way.

The individual fractures are characterized by their orientation. This orientation is usually given by two unit vectors **n** and **m** (cf. Fig. 1). **n** is the normal to the fracture plane; **m** gives the orientation of the polygon in the fracture plane.

The simplest model consists of a network in which all fractures have the same shape and are inscribed in a circle with a given radius R. The normal vectors **n** are uniformly distributed on the unit sphere. The density ρ of this isotropic monodisperse network is defined as the number of fractures per unit volume. An illustration of such a fracture system is shown in Fig. 2a.

Next consider three-dimensional networks made up of polydisperse fractures with plane polygonal shapes. These polygons may be regular or not, but all their vertices are supposed to lie on their circumscribed circle, whose radius R provides a measure of their size. In agreement with many observations of fractured rocks [2], the statistical distribution of the fracture sizes is supposed to be given by a power law

$$n(R) = \alpha R^{-a} \tag{1}$$

where $n(R)\mathrm{d}R$ is the probability of fracture radii in the range $[R, R + \mathrm{d}R]$; α is a normalization coefficient, and the exponent a ranges between 1 and 5. In practice, R may vary over a large interval which can span five orders of magnitude, from the size R_m of the microcracks to the size R_M of the largest fractures in the system. The normalization condition implies that α verifies

$$\begin{aligned} \alpha &= \frac{a-1}{R_\mathrm{m}^{1-a} - R_\mathrm{M}^{1-a}} \quad (a \neq 1)\,; \\ \alpha &= \frac{1}{\ln R_\mathrm{M} - \ln R_\mathrm{m}} \quad (a = 1)\,. \end{aligned} \tag{2}$$

The definition of the network density ρ for polydisperse networks should be modified. To this end, we introduce the volumetric number density of fracture per frac-

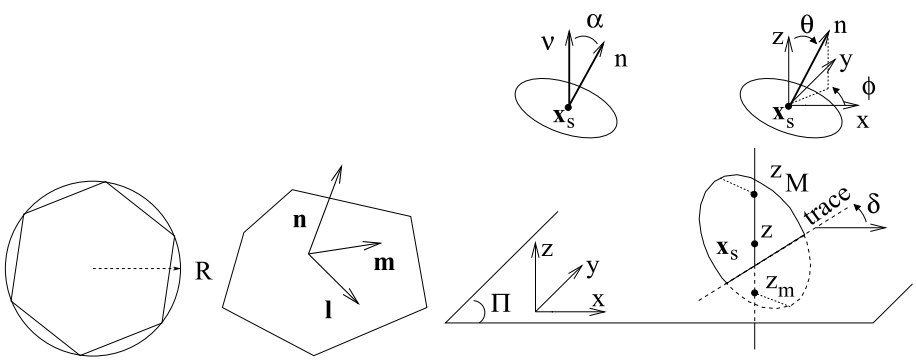

Percolation, and Faults and Fractures in Rock, Figure 1

Notations. Convex polygons such as hexagons are created within a circle of radius *R* (a). This polygon requires two unit vectors n and m to be oriented in space (b); l is a unit vector perpendicular to n and m. c illustrates the notations which are mostly used in Sect. "Determination of the Dimensionless Density from Experimental Data"

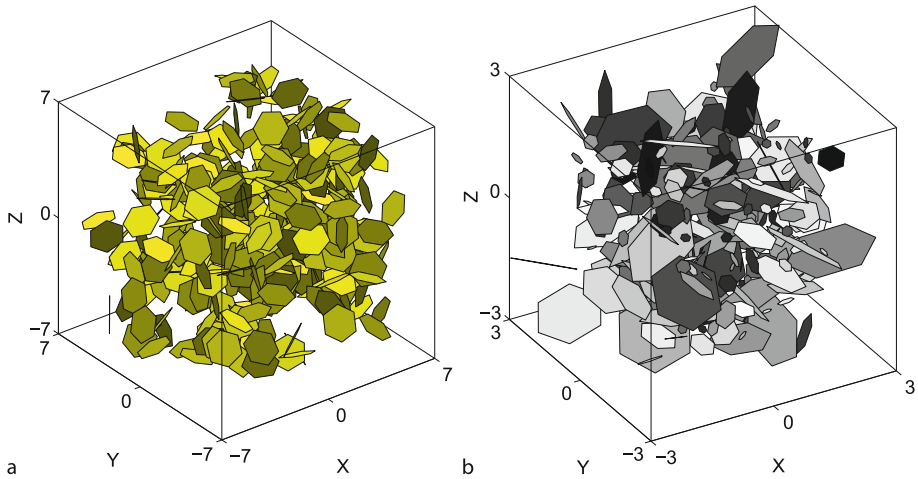

Percolation, and Faults and Fractures in Rock, Figure 2

Examples of three-dimensional fracture networks. a Monodisperse network made of identical polygons. The volume of size L^3 contains 495 hexagons; $L = 12R$ where R is the radius of the circle in which the hexagon is inscribed. b Polydisperse network of hexagonal fractures, with $L' = 4$, $a = 1.5$, $R'_m = 0.1$, which contains $N_{fr} = 300$ fractures ($\rho'_{21} = 1.25$, $\rho'_3 = 2.44$). The unit for the coordinates is R_M

ture size $F(R)$,

$$F(R) = \rho\, n(R) \tag{3}$$

where $F(R)\mathrm{d}R$ is the number of fractures with radius in the range $[R, R + \mathrm{d}R]$ per unit volume.

An example of such polydisperse networks is given in Fig. 2b.

Percolation of Fracture Networks

General Considerations on Continuum Percolation

Continuum Percolation Percolation, i. e., the existence of a spanning connnected cluster in the fracture network, is a crucial topological property which conditions many other geometrical or transport properties of the network.

Percolation of discrete sites or bonds lattices has been closely studied (see, e. g. [32,37],). In these lattices, the sites or bonds are occupied with a probability p, which can be interpreted as a concentration. In large systems, percolation occurs when p exceeds a critical value p_c, known as the percolation threshold, which depends on the underlying lattice structure. For p close to p_c, however, many geometrical or transport coefficients are known to scale as power laws of the difference $p - p_c$, according to the standard form

$$X \propto (p - p_c)^{\alpha} \ . \tag{4}$$

The quantity X may represent the correlation length, the fraction of sites connected to the infinite cluster, or the conductivity of the system. Different exponents are associated with the various quantities, but each is generally believed to be universal, i. e., insensitive to the details of the underlying lattice.

It is, of course, tempting to try to transpose this theoretical framework to the problem at hand. It is intuitively obvious that a fracture network will start percolating if some critical concentration is reached. The main difficulty is to define an equivalent of the probability p in discrete lattices. As shown below, this can be done by using the concept of excluded volume, introduced by [6] in the context of fracture networks.

Fracture networks belong to the general class of continuum percolation systems. Applications of continuum percolation concepts to geophysical problems have been reviewed by [9]. Continuum percolation differs from lattice percolation in several respects. First, the occupancy probability p in a discrete lattice ranges between 0 and 1, which means that there is a maximal concentration; the filling of the system can be defined relative to this upper bound. In continuum percolation, there is generally no such upper bound. For instance, there is ideally no upper limit to the degree of fracturation of a piece of rock. Consequently, the relative concentration p has to be replaced by a volumetric density. Second, any site or bond in lattice percolation cannot have more than a maximum number z of neighbors, called the lattice coordination number, whereas there is no limitation to the number of intersections for a fracture in a network. Other differences result from the variable lengths and orientations of the

bonds, in contrast with the discrete set of values imposed by a lattice, which may be significant for transport properties (see [5]).

Note that in this section we only consider "large" systems, i. e., the size of the objects in the percolation system may have a broad distribution, but the overall domain extension is supposed to widely exceed the size of the largest objects it contains. This condition may sometimes be difficult to fullfil; natural fracture networks often involve large-scale faults, which may in themselves ensure percolation if they cross the domain of investigation. [11] and later [24] considered such broad size-distributions where the probability of a spanning single fracture is non-zero.

In view of the previous considerations, two definitions of the system concentration appear possible. One is volumetric, quantified by the average number of objects in a reference volume; the other is topological, defined as the average number of connections with surrounding objects. These two definitions are nicely reconciled by the introduction of the concept of excluded volume.

The *excluded volume* V_{ex} of an object was defined by [6] as the volume surrounding it, in which the center of another object must be in order for them to intersect. We first discuss the simplest case of populations of identical objects, with volume V. For example, the excluded volumes of a sphere with volume V in 3D and of disks with area A in 2D are

$$\begin{aligned} V_{\mathrm{ex}} &= 8\,V \quad \text{for spheres}\,; \\ A_{\mathrm{ex}} &= 4\,A \quad \text{for disks in the plane}\,. \end{aligned} \tag{5}$$

These equations are also valid for any object with convex shape, if all the objects in the population have identical orientations.

If the objects are anisotropic and have distributed orientations, the excluded volume has to be averaged over all possible relative orientations of the intersecting objects.

Now suppose that the volumetric density of objects per unit volume is ρ. It is natural to use V_{ex} as a reference volume, and we may define the dimensionless density ρ' as the number of objects per volume V_{ex}

$$\rho' = \rho\, V_{\mathrm{ex}}\,. \tag{6}$$

On the other hand, the definition of V_{ex} implies that ρ' is also the average number of intersections per object, if they are randomly located according to a Poisson process. Therefore, given the shape of the object and its orientation distribution (and thus V_{ex}), the definition (6) incorporates both the volumetric and topologic aspects mentioned above.

It should be emphasized however, that the definition of the excluded volume is meaningful only if the object locations are uniformly distributed in space. If there are spatial correlations, they should be replaced by a spatial integral of the pair separation distribution function (see for instance [14] for applications to the physics of liquids).

Calculation of the Excluded Volume for Plane Convex Fractures A general expression for the excluded volume was established very early in the context of statistical mechanics by [19], for isotropically oriented objects. For two three-dimensional convex objects A and B with volumes V_A and V_B, areas A_A and A_B and surface averaged mean radius of curvature R_A and R_B, [19] obtained the mutual exclusion volume

$$V_{\mathrm{ex},AB} = V_A + V_B + (A_A R_B + A_B R_A)\,. \tag{7}$$

This expression can then be averaged over the distributions of object shapes and sizes. For equal spheres, Eq. (5) is obtained. For flat convex objects randomly oriented in space with perimeters P_A and P_B, it is reduced to [12]

$$V_{\mathrm{ex},AB} = \frac{1}{4}\,(A_A P_B + A_B P_A)\,. \tag{8}$$

On averaging (8) over the size distribution of objects with identical shapes, one obtains

$$V_{\mathrm{ex}} = \frac{1}{2}\,\langle A\rangle\,\langle P\rangle \tag{9}$$

where $\langle\cdot\rangle$ is the statistical average. If A and B are identical, (9) yields

$$V_{\mathrm{ex}}^{\mathrm{iso}} = \frac{1}{2}\,A\,P\,. \tag{10}$$

If the population of polygons is not isotropic and has a probability distribution $n(f)$, which may involve the shape or the size of the polygons, the average of (8) yields

$$\begin{aligned} V_{\mathrm{ex}}^{\mathrm{iso}} &= \frac{1}{4}\iint n(F_1)\,n(F_2)\,(A_1\,P_2 + A_2\,P_1)\,\mathrm{d}F_1\,\mathrm{d}F_2 \\ &= \frac{1}{2}\,\langle A\rangle\,\langle P\rangle \end{aligned} \tag{11}$$

where $\langle A\rangle$ and $\langle P\rangle$ are the average area and perimeter. Alternatively, the polygon orientation may be incorporated into $n(f)$ and a general expression of V_{ex} can be obtained.

Determinations of Continuum Percolation Thresholds The percolation thresholds of various simple continuous systems have been determined, since the pionneering papers of [35] and [27]. These early works were reviewed

Percolation, and Faults and Fractures in Rock, Table 1
Thresholds ρ'_c for various continuum percolation systems in d dimensions

d	Object type	ρ'_c	d	Object type	ρ'_c
2	Orthogonal sticks	3.2	3	Orthogonal elongated rods	0.7
2	Randomly oriented sticks	3.6	3	Randomly oriented elongated rods	1.4
2	Disks or parallel objects	4.5	3	Orthogonal squares	2.0
			3	Randomly oriented squares	2.46
			3	Spheres or parallel objects	2.80

by [6] and [5]. A few examples are given in Table 1, for monodisperse objects in a d-dimensional space. The critical concentration is described in terms of the average number ρ'_c of connections per object.

The influence of the orientation distribution was investigated by [29,30], and [6]. For sticks with constant length in the plane, [29] has shown that ρ'_c is identical for uniform orientation distributions in any angular sector and equal to the value 3.6 for an isotropic distribution. On the other hand, the value 3.2 for orthogonal sticks is also valid for any bimodal orientation distribution. By considering three-dimensional systems [6], also conclude that the total excluded volume at percolation is independent of the degree of anisotropy. [4] proposed a set of bounds which correspond to orthogonal and parallel object systems

$$3.2 \le \rho'_c \le 4.5 \quad d = 2; \quad 0.7 \le \rho'_c \le 2.8 \quad d = 3. \tag{12}$$

All these results were obtained by numerical simulations. One should also mention the heuristic criterion developed by [3]. They define the average "bonding distance" l as the mean distance between connected objects, which is essentially the gyration radius of the excluded volume

$$l^2 = \frac{1}{V_{\mathrm{ex}}} \int_{V_{\mathrm{ex}}} r^2 \, \mathrm{d}^3\mathbf{r} \,. \tag{13}$$

Note that l does not depend on the density of objects. They then postulate that percolation occurs when the average distance L_{d} between objects with at least two neighbors is smaller than or equal to $2l$. To evaluate L_{d}, they note that the number k of connections to a given object corresponds to a Poisson distribution

$$\Pr(k) = \frac{\rho'^k}{k!} \mathrm{e}^{-\rho'} \,. \tag{14}$$

Therefore, the density ρ_2 of objects with at least two neighbors is

$$\rho_2 = \rho \left[1 - \left(1 + \rho'\right) \mathrm{e}^{-\rho'} \right] . \tag{15}$$

Thus, an estimate of L_d follows from

$$\frac{4}{3} \pi \left(\frac{L_{\mathrm{d}}}{2} \right)^3 = \frac{1}{\rho_2} \,. \tag{16}$$

An equation for the critical concentration ρ'_c can be directly deduced from the statement that $L_{\mathrm{d}} = 2l$. Although the argument is not substantiated, it is quite successful. It yields directly $\rho'_c = 2.80$ for spheres. In two dimensions, L_{d} is replaced by the average distance between objects with at least 5 neighbors.

An interesting feature of this argument is that it can be easily generalized to account for spatial correlations. If $\rho g(r)$ denotes the probability density of finding an object center at a distance r from an object located at the origin, the bonding distance is defined by the weighted average

$$l^2 = \frac{\int_{V_{\mathrm{ex}}} r^2 \, g(r) \, \mathrm{d}^3\mathbf{r}}{\int_{V_{\mathrm{ex}}} g(r) \, \mathrm{d}^3\mathbf{r}} \,. \tag{17}$$

Similarly, the average number of bonds per object appears as

$$\rho' = \rho \int_{V_{\mathrm{ex}}} g(r) \mathrm{d}^3\mathbf{r} \,. \tag{18}$$

Using these two definitions, an equation for ρ'_c can be obtained. Its predictions were successfully compared by [3] to numerical simulations for systems of hard-core spheres with or without interaction potentials.

Only monodisperse objects have been addressed so far in this subsection. For polydisperse populations, there seems to be some confusion in the literature. For flat objects, the statistical derivation of the excluded volume in Sect. "Calculation of the Excluded Volume for Plane Convex Fractures" quite naturally yielded the averages (9) or (11), which account for the sizes of the two intersecting objects. For isotropic populations of segments with length l in the plane or disks with radius R in space, for instance, the averages can be expressed as

$$V_{\mathrm{ex}} = \frac{2}{\pi} \langle l \rangle^2 \ \text{segments}, \quad d = 2 \tag{19a}$$

$$V_{\mathrm{ex}} = \frac{\pi^2}{8} \langle R^2 \rangle \langle R \rangle \ \text{disks}, \quad d = 3 \,. \tag{19b}$$

However, another type of average has been proposed by [6], namely,

$$W_{\mathrm{ex}} = \frac{2}{\pi} \langle l^2 \rangle \ \text{segments}, \quad d = 2 \tag{20a}$$

$$W_{ex} = \frac{\pi^2}{8} \langle R^3 \rangle \quad \text{disks}, \quad d = 3 \,. \tag{20b}$$

On the basis of the numerical simulations of [29,30], [6] and others claim that the average bond number for polydisperse objects is not given by Eq. (6) but instead by

$$\rho'' = \rho \, W_{ex} \,. \tag{21}$$

However, a careful examination of [29]'s data shows that they correspond very accurately to Eq. (6) with (19a). Finally, the derivations of [8] are based on (19b) and yield consistent results.

Actually [29], and [30] showed that ρ'_c is not invariant for similar systems of segments in the plane with various degrees of polydispersity, while ρ''_c is. [11] also observed that ρ''_c is roughly constant for very broad power-law segment size distributions. The profound meaning of this observation is that continuum percolation is not determined only by the average coordination, when connections over various ranges may coexist. As suggested by [28], this is probably because contacts between objects too close to each other are redundant to percolation.

To summarize, the density ρ' based on V_{ex} resulting from the averages (9), (11), or (19a) is always equal to the mean number of intersections per object, but it cannot be used to relate the percolation thresholds of mono- and polydisperse systems. The alternative definition ρ'' in (21) is very successful in this respect and is going to be generalized as ρ'_3 in (29).

Methods

The three main tools necessary for the numerical study of the percolation properties of the fracture network models are summarized in this section.

First, the medium is assumed to be spatially periodic on a large scale. A detailed description of spatially periodic media is given by [1], and only the main characteristics of these models are briefly repeated here. The geometrical and physical properties of the system under investigation are invariant under the translations

$$\mathbf{R_i} = i_1 \mathbf{l}_1 + i_2 \mathbf{l}_2 + i_3 \mathbf{l}_3 \tag{22}$$

where $\mathbf{i} = (i_1, i_2, i_3) \in \mathcal{Z}^3$, and where $\mathbf{l}_1$, $\mathbf{l}_2$ and $\mathbf{l}_3$ define a unit cell where the system is studied. The entire space is tiled by replicas of this unit cell, translated by $\mathbf{R_i}$. All the studies presented in this chapter are performed in cubic unit cells where $|\mathbf{l}_1| = |\mathbf{l}_2| = |\mathbf{l}_3| = L$.

Spatial periodicity implies that fractures may cross the imaginary unit cell boundaries, and reach the neighboring cells of the periodic medium. Therefore, for polydisperse fractures, R_M should be at least smaller than $L/2$. Moreover, in order to represent a homogeneous medium by a periodic model, one has to set the unit cell size much larger than any finite characteristic length scale in the system. Practically speaking, because of the finite size effects which will be discussed in Sect. "Methods", R_M is at least smaller than $L/4$.

Second, the networks are characterized by a graph which provides all the necessary relations and information. This graph, denoted by Γ_1, consists of vertices which correspond to the polygons, and edges which correspond to the intersection between polygons. Γ_1 will be used to study network percolation as a function of fracture shape, distribution and density, as well as to characterize the topological features of the percolating components of the networks.

The information relative to the intersections is stored in the graph Γ_1. Since the networks considered here are spatially periodic, intersections of a polygon P_1 with the periodic replicas of a polygon P_2 in the 26 neighboring unit cells have to be checked as well. Once the intersections have been identified, the edges of Γ_1 are known, and Γ_1 can be set up.

Third, in order to estimate the percolation thresholds, the classical finite-size scaling method described in [37] is used. The percolating system is studied for various cell sizes L. For given values of L and ρ, the probability $\Pi_L(\rho)$ of having a percolating cluster is derived from numerous realizations of the system. Then, the numerical data are used to estimate ρ_{Lc} (the value for which $\Pi_L(\rho) = 1/2$) and an estimate of the width Δ_L of the transition region of $\Pi_L(\rho)$. $\Pi_L(\rho)$ was fitted with an error function of the form

$$\Pi_L(\rho) = \frac{1}{\sqrt{2\pi}\,\Delta_L} \int_{-\infty}^{\rho} \exp\left\{ -\frac{(\xi - \rho_{Lc})^2}{2(\Delta_L)^2} \right\} \mathrm{d}\xi \tag{23}$$

where ρ_{Lc} and Δ_L are fit parameters. Once they have been evaluated for several values of L, the asymptotic value ρ_{Lc} for infinite systems ρ_c can be derived from the two scaling relations

$$\rho_{Lc} - \rho_c \propto L^{-1/\nu} \qquad \Delta_L \propto L^{-1/\nu} \,. \tag{24}$$

Monodisperse Fractures

This case was addressed by [18]. Since the computer time increases proportionally to the square of the number N of objects, L (measured in units of the disk radius R) was kept below 16 in this early contribution. Despite the small cell sizes, the scaling laws (24) are well verified, which justifies

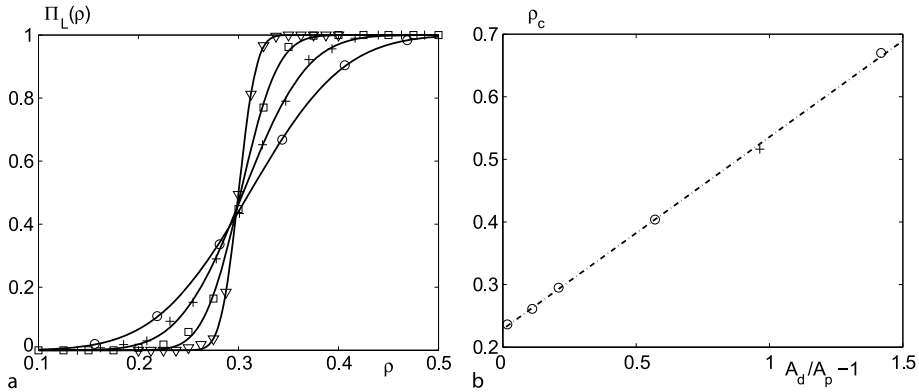

Percolation, and Faults and Fractures in Rock, Figure 3

a The probability of percolation $\Pi_L(\rho)$ *vs* the density ρ of fractures in fracture networks created by equal sized, regular hexagons. Data are for sample sizes $L/R = 4(\circ)$, 6 (+), 10 (□), 20 (▽). The solid lines are the fitted error functions. b The percolation thresholds ρ_c for regular polygons (∘) and rectangles with $a/b = 0.5(+)$ *vs* $(A_d/A_p - 1)$. The linear fit (— · — · —) yields $\rho_c = 0.231 \pm 0.002$ for disks

the extrapolations of ρ_c at $L \to \infty$. The polygons were created, and intersections identified. Percolation was checked in all possible directions x, y and z. Periodic boundary conditions were applied to the 3d graph during this search; this means that a cluster must touch two opposite faces of the unit cell, and in addition contain fractures intersecting one another across the faces.

An example of the plots of the estimated $\Pi_L(\rho)$ data points is given in Fig. 3a, together with the fitted error functions. Plots of $\ln(\Delta_L)$ *vs* $\ln(L/R)$ were used to obtain the critical exponent ν. The various polygons are expected to belong to the same universality class, and ν was expected to be the same in all cases. Values were in the range $\nu = 1.011 \pm 0.044$. The plots of ρ_{Lc} *vs* Δ_L were extrapolated for $\Delta_L \to 0$ to find ρ_c and these extrapolations are shown in Fig. 3b as functions of the shape factor $A_d/A_p - 1$ where A_d is the area πR^2 of the circumscribed disk.

These results can be analyzed in terms of the average number of intersections per fracture ρ'. (10) can be applied to networks made of identical polygons

$$\frac{V_{ex}}{R^3} = \pi^2 \left(\frac{N_\nu}{\pi}\right)^2 \cos\left(\frac{\pi}{N_\nu}\right) \sin^2\left(\frac{\pi}{N_\nu}\right), \quad \text{(regular } N_\nu\text{-polygons)} \tag{25a}$$

$$\frac{V_{ex}}{R^3} = \frac{8a(a+1)}{(a^2+1)^{3/2}}, \quad \text{(rectangles with aspect ratio } a) \tag{25b}$$

The resulting values of ρ'_c are remarkably constant (cf. [18]). For all the fracture networks, including the cases of anisotropic (rectangular) polygons, ρ'_c is within the range

$$\rho'_c = 2.26 \pm 0.04 . \tag{26}$$

Note that (26) concords with the limits (12) set up by [4] for 3d systems.

To summarize, this set of numerical results suggests that the percolation threshold of a network of identical Poissonian polygons has a universal value, expressed as Eq. (26).

Polydisperse Fractures

Since natural fracture networks are likely to have more complex size and shape distributions, the extension of (26) to these cases is of great interest. The key for this extension is the definition of a proper averaging procedure for the excluded volume.

The fracture size R is always supposed to follow the power law (1). Moreover, fractures of various shapes S are considered as well as mixtures of shapes. The three length scales R_m, R_M and L define two dimensionless ratios

$$R'_m = \frac{R_m}{R_M}, \quad L' = \frac{L}{R_M} . \tag{27}$$

Moreover, it will be shown below that global connectivity (percolation) is no longer controled solely by the local one (mean coordination), in the case of size polydispersity, and the definition of the percolation parameter has to be generalized. Since shape effects are well accounted for by $\langle V_{ex}\rangle$, it is useful to define the dimensionless shape factor $\langle v_{ex}\rangle$, for a set of fractures with identical shapes, but

possibly different sizes

$$\langle v_{\rm ex}\rangle = \frac{\langle V_{\rm ex}\rangle}{\langle R\rangle\langle R^2\rangle}\,. \tag{28}$$

It can then be used to define two dimensionless densities, with different weightings of the fracture sizes

$$\begin{aligned}\rho'_{21} &= \rho\,\langle v_{\rm ex}\rangle\,\langle R^2\rangle\langle R\rangle = \rho\langle V_{\rm ex}\rangle\,;\\ \rho'_{3} &= \rho\,\langle v_{\rm ex}\rangle\,\langle R^3\rangle\,.\end{aligned} \tag{29}$$

The subscripts are reminders of the statistical moments of R involved in each definition. ρ'_{21} is the generalization of ρ' for monodisperse networks, since it can be shown that it is still equal to the mean number of intersections per fracture [2]. Both ρ'_{21} and ρ'_3 reduce of course to ρ' in case of equal-sized fractures.

The main tools required to study the percolation of polydisperse networks model are similar to the ones described in Sect. "Methods".

For given values of the parameters, the probability Π of having a percolating cluster which spans the cell along the x-direction is derived from N_r random realizations of the system; then, the value ρ'_c for which $\Pi = 0.5$ is estimated. Π and ρ'_c depend on several parameters as summarized by the formulae

$$\Pi(R'_{\rm m}, L', a, S, \rho')\,, \quad \rho'_{\rm c}(R'_{\rm m}, L', a, S) \tag{30}$$

where ρ' denotes any one of the dimensionless densities defined in (29). For brevity, they will be often written as $\Pi(L', \rho')$ and $\rho'_{\rm c}(L')$.

In practice, $\Pi(L', \rho')$ was evaluated from sets of 500 realizations, for about 10 values of the network density, evenly distributed in a range where Π varies from 0.05 to 0.95. Since there is a correspondance between ρ'_{21} and ρ'_3, for given values of S, a and R_m, the same data sets can be used to determine $\rho'_{21c}(L)$ and $\rho'_{3c}(L)$. The 95% confidence interval is estimated to be about ± 0.04 in terms of $\rho'_{3c}(L)$.

The influence on ρ'_c of the four parameters in Eq. (30) was systematically studied in [24]. We only state here the main result, which is that in the range $1.5 \le a \le 4$, $R_m \ll L$ and for (almost) any fracture shape, ρ'_c depends only on the domain size, and that in the limit of infinite domains, a unique value of $\rho'_c(\infty)$ applies in all cases. The independence on the various parameters is illustrated in the following examples.

In the example of Fig. 4, L' and a are kept constant, but the range of size and the fracture shapes varied. The networks contain hexagons, squares or triangles, or mixtures of hexagons with triangles or rectangles with a four

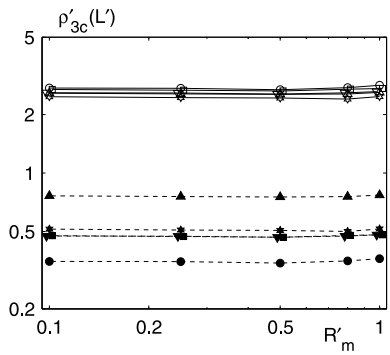

Percolation, and Faults and Fractures in Rock, Figure 4
The percolation thresholds ρ'_{3c} (*open symbols, solid lines*) and $\rho_c\langle R^3\rangle$ (*black symbols, broken lines*) for networks with $L' = 6$ and $a = 1.5$ for regular hexagons (○), squares (□), triangles (△), mixture of hexagons and triangles, 50%–50% (▽), and mixture of hexagons and rectangles with aspect ratio 4, 50%–50% (✡)

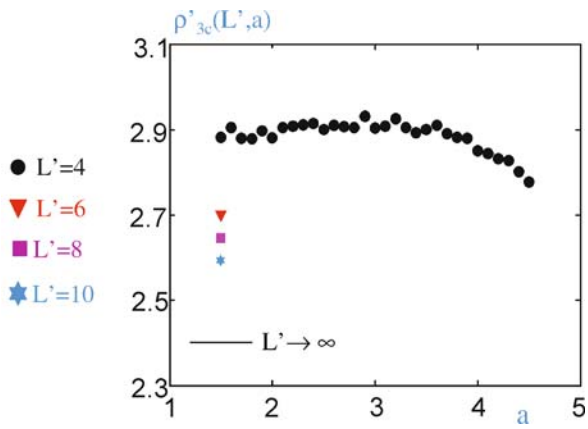

Percolation, and Faults and Fractures in Rock, Figure 5
The percolation threshold $\rho'_{3c}(L', a)$ for networks of hexagonal fractures with $R'_m = 0.1$, versus the exponent a, for various domain sizes L. The lower line is the extrapolation of the data for $a = 1.5$ when L' tends to infinity

to one aspect ratio. The upper set of curves shows that ρ'_c is indeed independent of R_m and S. Note that the rightmost points are actually monodisperse networks. For comparison, the thresholds $\rho_c\langle R^3\rangle$, which do not include the shape factor $\langle v_{\rm ex}\rangle$ (see Eq. 29), are also shown in the same figure and they are clearly much more scattered. It is the incorporation of $\langle v_{\rm ex}\rangle$ in the definition of ρ'_3 which unifies the results for the different shapes.

Conversely, the fracture shape (hexagonal) and the range of size ($R'_m = 0.1$) are kept constant in the example of Fig. 5, whereas the exponent a and the domain size L are varied. It is seen that ρ'_c does not vary when a ranges from 1.5 to 4. However, a definite dependence on the domain

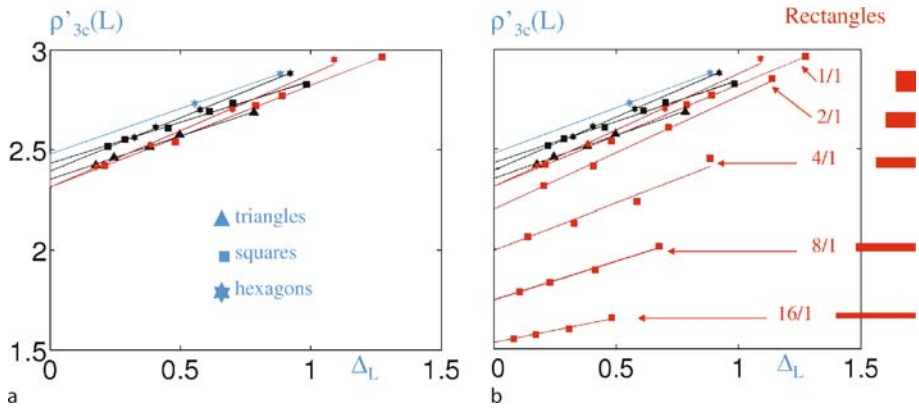

Percolation, and Faults and Fractures in Rock, Figure 6
The percolation threshold $\rho'_{3c}(L')$ for mono- or polydisperse networks of fractures with various shapes, versus the width Δ_L of the percolation transition. In a, the fractures are hexagons, squares or triangles. $\rho'_{3c}(\infty)$ is the extrapolation for $\Delta_L \to 0$, which falls in the range of Eq. 31. Data for monodisperse networks of rectangles with aspect ratios from 1 to 16 are added in b

size is observed, which corresponds to the well known finite size effects.

The data for increasing L' can be extrapolated for infinite systems by use of a classical technique. The combination of (23) and (24) shows that $\rho'_c(L) - \rho'_c(\infty)$ is proportional to the width Δ_L of the percolation transition zone. Hence, $\rho'_c(\infty)$ can be read on the vertical axis of the plot of $\rho'_c(L)$ versus Δ_L, which is shown in Fig. 6. The data for many cases, including various fracture shapes in monodisperse and polydisperse networks, are gathered in Fig. 6a. In all cases, the extrapolated values $\rho'_{3c}(\infty)$ fall in the narrow range

$$\rho'_{3c}(R'_m, a, S, L' \to \infty) = \rho'_{3c}(\infty) \approx 2.4 \pm 0.1 \,. \quad (31)$$

This applies for a variety of shapes, as well as for mixtures of fractures with different shapes (see Fig. 4).

However, when the polygons become elongated, $\rho'_{3c}(L')$ varies with the aspect ratio. Data for rectangles with aspect ratios A_r up to 16 are shown in Fig. 6b. It appears that $\rho'_{3c}(L')$ decreases significantly when A_r increases.

This can be taken into account by using the shape factor $\eta = 4R/P$ of the fractures. This ratio is minimum for disks, with $\eta = 2/\pi \approx 0.637$, and it increases up to one when the shape deviates from circularity. It turns out that a quadratic correction in terms of η is very successful for the representation of the data for very different and irregular fracture shapes.

All the thresholds obtained in cells with $L' = 6$ and mono- or polydisperse size distributions with $a = 1.5$ or 2 and $R_m = 0.1$ are plotted in Fig. 7 as functions of η. This includes networks of hexagons, squares, triangles, mixtures of hexagons with rectangles or triangles, and rectangles with h/w up to 16. The data are well fitted by the expression

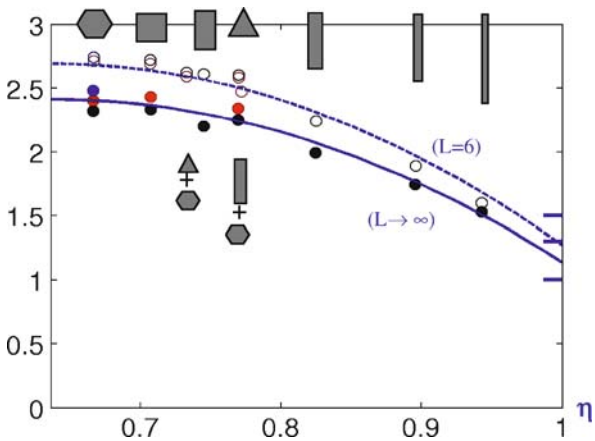

Percolation, and Faults and Fractures in Rock, Figure 7
The percolation thresholds $\rho'_{3c}(L' = 6)$ and $\rho'_{3c}(\infty)$ for a variety of fracture shapes and size distributions, in comparison with the expressions (32), (33). The marks on the right are the predictions of [15,16,31] for infinitely elongated objects. The fracture shapes are indicated by the icons above or below the data points

$$\rho'_{3c}(L') = 2.69\left[1 - 4\left(\eta - \frac{2}{\pi}\right)^2\right] \quad (L' = 6)\,. \quad (32)$$

The extrapolated data for infinite systems are also presented in Fig. 7, in comparison with the corrected version of Eq. (31),

$$\rho'_{3c}(\infty) = 2.41\left[1 - 4\left(\eta - \frac{2}{\pi}\right)^2\right]\,. \quad (33)$$

In both cases, the deviations never exceed ± 0.1. The corrective term becomes significant, i. e., larger than the error bar in (31), when $\eta > 3/4$, which corresponds for rectangles to aspect ratios larger than 2.

It can be noted that Eq. (33) predicts a threshold value 1.14 when h/w tends to infinity (i. e., when $\eta \to 1$), which is in the range of the predictions 1.5 for prolate ellipsoids [16], 1 for capped cylinders [15] and 1.3 for elongated prisms [31], in the limit of infinite slenderness.

Determination of the Dimensionless Density from Experimental Data

Since percolation properties are controlled by the dimensionless density ρ', it is theoretically and practically important to derive estimations of ρ' from field data. In most cases, these data are based on 1D and 2D measurements of fracture traces along boreholes or on exposed outcrops which necessitate extrapolation by stereological techniques to 3D. Such extrapolations have already been made for specific fracture shapes by Warburton [39,40], Piggott [26], Berkowitz and Adler [8] and Sisavath et al. [36] (see also the references therein).

Our general methodology which is detailed in [38] can be illustrated by the intersection of a family of convex fractures with a line of length L which is parallel to the unit vector $\mathbf{p}$. Consider a fracture of surface A, of normal $\mathbf{n}$ and of in-plane orientation ω; this object does not intersect the line when its center is located out of a surface of area A. Since this is valid for any in-plane orientation, the excluded volume of the line and of the surface is equal to $AL|\mathbf{p}.\mathbf{n}|$. Hence, the average number of intersections $\langle n_I \rangle$ per unit length between such a line and an isotropic network of a monodisperse family of fractures is

$$\langle n_I \rangle = \frac{1}{2} A\rho \,. \tag{34}$$

Of course, the major interests of this formula are that it does not depend on the precise shape S of the fractures and that ρ can be deduced from n_I and A. However, it depends in a crucial way on the convexity of the fractures.

Isotropic Networks

In order to derive the average number of intersections Σ_t of a family of convex fractures $\mathcal{F}(R)$ with a plane Π per unit area of the plane, define in Π a large convex region $\mathcal{R}$ of area $\mathcal{A}$ and perimeter $\mathcal{P}$. The excluded volume of $\mathcal{F}(R)$ and $\mathcal{R}$ is thus given by (8). The number of intersections $\mathrm{d}\Sigma_t$ of the fractures of size ranging from R to $R + \mathrm{d}R$ is proportional to the volumetric density of such fractures multiplied by the excluded volume of $\mathcal{F}(R)$ and $\mathcal{R}$ as expressed by (8); when $\mathcal{A} \to \infty$, $\mathcal{A} \gg \mathcal{P}$; therefore,

$$d\Sigma_t(R) \to \frac{1}{4}\rho P(R) n(R) \mathrm{d}R \quad \text{when} \quad \mathcal{A} \to \infty \,. \tag{35}$$

This relation can be averaged over the sizes R

$$\Sigma_t = \int \mathrm{d}\Sigma_t(R) = \frac{1}{4}\rho \langle P \rangle \,. \tag{36}$$

The intersections of the fractures with a plane are called *traces* or *chords*. Let c be the length of a trace as illustrated in Fig. 1c. Such an intersection of length $c(z, \mathbf{n}, \omega)$ exists if the vertical coordinate z of the center verifies

$$z_m(\mathbf{n}, \omega) \le z \le z_M(\mathbf{n}, \omega) \,. \tag{37}$$

For a given fracture of size R, the average trace length $\langle c \rangle_R$ when the intersection exists, can be expressed as

$$\langle c \rangle_R = \frac{\int \mathrm{d}\omega \int \mathrm{d}\mathbf{n} \int_{z_m}^{z_M} c \mathrm{d}z}{\int \mathrm{d}\omega \int \mathrm{d}\mathbf{n} \int_{z_m}^{z_M} \mathrm{d}z} \,. \tag{38}$$

Surprisingly, the numerator N_R of this fraction is easier to evaluate than its denominator D_R. The most internal integral $\int_{z_m}^{z_M} c\mathrm{d}z$ is equal to the area A of the fracture projected onto the plane perpendicular to Π which contains the trace, i. e., $A \sin\theta$. Therefore, N_R is equal to $\pi^3 A$. The derivation of D_R is slightly more involved Santalo [34]; it is proportional to the integral of the Feret (or caliper) diameter over ω. Frenet formulae are used to express this integral. Finally,

$$\langle c \rangle_R = \pi \frac{A(R)}{P(R)} \,. \tag{39}$$

For polydisperse fractures, the overall average $\langle c \rangle$ is given by

$$\langle c \rangle = \frac{\int \mathrm{d}R \Sigma_t(R) \langle c \rangle_R}{\int \mathrm{d}R \Sigma_t(R)} = \pi \frac{\langle A \rangle}{\langle P \rangle} \tag{40}$$

a formula which is again an obvious generalization of the disk formula (cf. (24a) of Berkowitz and Adler [8]).

The density of trace intersections Σ_p is defined as the number per unit surface in the observation plane of the points which are intersections of traces. Since the fractures are randomly oriented and distributed in space, the same properties are valid for the traces. Moreover, as a trivial extension of the concept of excluded volume, the excluded surface S_{ex} of two traces of random orientations and of lengths c_1 and c_2 is equal to (cf. [2])

$$S_{\mathrm{ex}} = \frac{2}{\pi} c_1 c_2 \,. \tag{41}$$

Let $\sigma_t(R, c)\mathrm{d}c\mathrm{d}R$ be the surface density of traces of length c ranging from c to $c + \mathrm{d}c$, for the fractures of size R ranging from R to $R + \mathrm{d}R$. Hence, the surface density of intersections of traces c_1 corresponding to fractures of size R_1 and of traces c_2 corresponding to fractures of size R_2 is

$$\sigma = \frac{1}{2}\sigma_t(R_1, c_1)\sigma_t(R_2, c_2)\frac{2}{\pi}c_1 c_2 \,. \tag{42}$$

As a direct consequence

$$\Sigma_p = \iiiint \sigma \,\mathrm{d}c_1\,\mathrm{d}R_1\,\mathrm{d}c_2\,\mathrm{d}R_2 \,. \tag{43}$$

This last expression can be split into a product of integrals since the populations 1 and 2 are independent. According to (35), (36), (43), Σ_p can be expressed as

$$\Sigma_p = \frac{1}{\pi}\frac{\pi^2}{16}\rho^2\langle A\rangle^2 = \frac{\pi}{16}\rho^2\langle A\rangle^2 \,. \tag{44}$$

Extensions

Let us now examine various possible extensions of the previous formulae.

The precise shape S of the fractures is never taken into account. Therefore, all the previous formulae are valid whatever the mixture of shapes S.

For anisotropic networks, the normal vector $\mathbf{n}$ is not uniformly distributed over the unit sphere. Let θ and φ be the two polar angles of $\mathbf{n}$ (cf. Fig. 1c); the probability that the end of $\mathbf{n}$ for fractures of sizes in the interval $[R, R + \mathrm{d}R]$ belongs to the interval $[\theta, \theta + \mathrm{d}\theta] \times [\phi, \phi + \mathrm{d}\phi]$ is given by $\rho n(R, \mathbf{n})\mathrm{d}\theta\mathrm{d}\phi\mathrm{d}R$. The statistical average $\langle\cdot\rangle$ can be calculated with this differential element.

The first quantity which can be easily generalized is $\langle n_I\rangle$ (cf. (34))

$$\langle n_I\rangle = \rho\iiint n(R, \mathbf{n})A\cos\theta\,\mathrm{d}\theta\,\mathrm{d}\phi\,\mathrm{d}R = \rho\langle A|\mathbf{p}\cdot\mathbf{n}|\rangle \,. \tag{45}$$

The other generalized formulae are summarized in Table 2. α is the angle between the normal $\boldsymbol{\nu}$ to the plane Π and $\mathbf{n}$; in most cases, by choosing the z-axis perpendicular to Π, α is equal to θ; β_{12} is the angle between the normals $\mathbf{n}_1$ and $\mathbf{n}_2$ to the two fractures 1 and 2.

These formulae can be specialized to networks of subvertical fractures with a horizontal observation plane Π. Then, α is equal to $\pi/2$ and β_{12} is equal to the angles between the two traces in Π. Such networks can be either isotropic (i. e., the directions of the traces in Π are isotropic), or anisotropic. The corresponding results are detailed in Table 2.

Percolation, and Faults and Fractures in Rock, Table 2
The major relations for the various kinds of networks. $\mathcal{B}_{12} = \langle A_1 A_2|\sin\beta_{12}|\rangle$

	Isotropic 3D	Anisotropic 3D	Subvertical isotropic	Subvertical anisotropic
$\langle n_I\rangle$	$\frac{1}{2}\rho\langle A\rangle$	$\rho\langle A\|\mathbf{p}\cdot\mathbf{n}\|\rangle$	$\frac{2}{\pi}\rho\langle A\rangle$	$\rho\langle A\|\mathbf{p}\cdot\mathbf{n}\|\rangle$
Σ_t	$\frac{1}{4}\rho\langle P\rangle$	$\frac{\rho}{\pi}\langle\|\sin\alpha\|P\rangle$	$\frac{\rho}{\pi}\langle P\rangle$	$\frac{\rho}{\pi}\langle P\rangle$
$\langle c\rangle$	$\pi\frac{\langle A\rangle}{\langle P\rangle}$	$\pi\frac{\langle A\|\sin\alpha\|\rangle}{\langle P\|\sin\alpha\|\rangle}$	$\pi\frac{\langle A\rangle}{\langle P\rangle}$	$\pi\frac{\langle A\rangle}{\langle P\rangle}$
Σ_p	$\frac{\pi}{16}\rho^2\langle A\rangle^2$	$\frac{1}{2}\rho^2\mathcal{A}_{12}$	$\frac{1}{\pi}\rho^2\langle A\rangle^2$	$\frac{\rho^2}{2}\mathcal{B}_{12}$

Discussion

Discrete Families of Fractures In many practical cases, the fractures are perpendicular to a finite set of normals $\{\mathbf{n}_i; i = 1, \ldots, m\}$ with probabilities $\{n(R, \mathbf{n}_i); i = 1, \ldots, m\}$. The integrals over $\mathrm{d}\theta\mathrm{d}\phi$ are thus replaced by the following summation for a function $f(R, \mathbf{n}_i)$

$$\rho\sum_{i=1}^{m} n(R, \mathbf{n}_i)f(R, \mathbf{n}_i) \,. \tag{46}$$

Practical Use of the Formulae The major interest of the formulae summarized in Table 2 is to try to use them to derive the macroscopic quantities ρ, $\langle A\rangle$ and $\langle P\rangle$. It is easy (and frustrating) to realize that only two of these quantities can be obtained. For instance, (40) implies that $\langle A\rangle = \pi^{-1}\langle P\rangle\langle c\rangle$; from (36), $\langle P\rangle = 4\rho^{-1}\Sigma_t$; therefore $\langle A\rangle = 4\pi^{-1}\rho^{-1}\Sigma_t\langle c\rangle$. When these expressions are introduced into (34) or (44), one obtains that the three following ratios should be equal to one

$$\kappa_1 = \frac{\pi}{2}\frac{\langle n_I\rangle}{\Sigma_t\langle c\rangle}\,, \quad \kappa_2 = \frac{\pi\Sigma_p}{\Sigma_t^2\langle c\rangle^2}\,, \quad \kappa_3 = \frac{\pi}{4}\frac{\langle n_I\rangle^2}{\Sigma_p}\,. \tag{47}$$

The third relation is derived by eliminating $\Sigma_t\langle c\rangle$ between κ_1 and κ_2. These relations provide consistency relations between the data, but not ρ.

In other words, only two of the three quantities ρ, $\langle A\rangle$ and $\langle P\rangle$ can be simultaneously derived from the average measured data. Note also that κ_1 is insensitive to the spatial organization, and that this is not true for κ_2 and κ_3 which depend on trace intersections.

One can go further if some geometrical information is available which could be $\langle V_{\mathrm{ex}}\rangle$. Here, we shall use a shape factor η which is defined as $\langle A\rangle\langle P\rangle^{-2}$. For 3D isotropic networks, this expression can be combined to (40) and to (36) to yield $\langle A\rangle$, $\langle P\rangle$ and ρ

$$\langle P\rangle = \frac{\langle c\rangle}{\pi\eta}\,, \quad \langle A\rangle = \frac{\langle c\rangle^2}{\pi^2\eta}\,, \quad \rho = 4\pi\eta\frac{\Sigma_t}{\langle c\rangle} \tag{48}$$

or for a set of fractures normal to $\mathbf{n}_i$

$$\rho_i = \frac{\pi^2 \eta_i}{|\sin\alpha_i|} \frac{\Sigma_{ti}}{\langle c \rangle_i} . \tag{49}$$

There are many equivalent ways to derive ρ. The choice of the adequate formula depends on the available data. Note that formulae which contains Σ_p cannot be applied to families of parallel fractures.

When ρ and therefore $\langle V_{\text{ex}} \rangle$ (cf. (8)) are known by one way or another, one can derive the dimensionless density $\rho' = \rho \langle V_{\text{ex}} \rangle$ for isotropic and anisotropic networks

$$\rho' = \frac{\rho}{\pi} \langle (A_1 P_2 + A_2 P_1) |\sin\beta_{12}| \rangle \tag{50}$$

$$\rho' = \frac{\rho}{2} \langle A \rangle \langle P \rangle \quad (3d) ; \quad \rho' = \frac{4\rho}{\pi^2} \langle A \rangle \langle P \rangle \quad (2d) . \tag{51}$$

Then, if the fracture network is not too polydisperse, one can use a classical mean field argument and approximate its properties by the properties of a monodisperse network of density ρ'.

Applications

Several applications have already been made of the previous methodology and they can be summarized as follows. [36] showed that when data relative to fractures are collected along a line (e. g. a road or a well), estimations can be given to the major geometrical properties of the corresponding fracture networks, such as the volumetric density of fractures and their percolation character. [38] used the two dimensional maps obtained by [25] for subvertical fractures. Among other results, some of the consistency relationships (47) are well verified by these data. As previously, ρ' is estimated.

Finally [17], reconstructed a three-dimensional fracture network in a granite block from a series of experimental serial sections provided by [21]. It was visualized and its most important geometrical characteristics were studied. Though the network mostly consists of two families of fractures, it is interesting to note that a simple model of randomly oriented, monodisperse hexagons often yields a good order of magnitude for the various geometrical properties, which have been measured on the real block.

Role of the Dimensionless Density in Other Geometrical Properties and Permeability

Though this article is focused on percolation properties, it is important to notice that the dimensionless densities which were introduced, play a crucial role in other properties as well. [18] studied two main other geometrical properties for monodisperse networks. Fracture networks partition the solid space into blocks; the block density is denoted by ρ_b. One can introduce the cyclomatic number of the graph Γ_1 which is the number of independent cycles of this graph, and more precisely the number of cycles $\overline{\beta}_1$ per unit volume. [18] showed that ρ_b and $\overline{\beta}_1$ when made dimensionless by the excluded volume are independent of the fracture shapes.

Similar properties are found for the macroscopic permeability of fracture networks [20] and of fractured porous media whether they are monodisperse [10] or polydisperse [23]. In a series of contributions, the corresponding dimensionless quantities were shown to depend only on the dimensionless density ρ'. This is illustrated in Fig. 8. The porous medium has a local permeability K_m and the monodisperse fractures a conductivity σ. The macroscopic permeability of this medium is denoted by K_{eff}. Dimensionless quantities denoted by primes can be defined as

$$\sigma = R\, K_m\, \sigma' , \quad K_{\text{eff}} = K_m\, K'_{\text{eff}} . \tag{52}$$

Figure 8 shows that the average macroscopic permeability $\langle K'_{\text{eff}} \rangle$ does not depend significantly on the fracture shape. The two major parameters are ρ' and σ'.

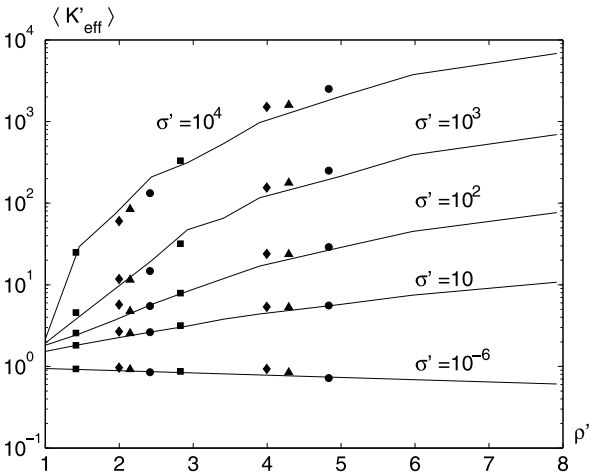

Percolation, and Faults and Fractures in Rock, Figure 8
Statistical averages of the permeability $\langle K'_{\text{eff}} \rangle$ for samples containing N_{fr}=16 or 32 fractures, with 4-, 6- or 20-gonal shapes, as functions of the network density ρ' and of the fracture conductivity σ'. The cell size is $L = 4R$. Data are for squares (□), rectangles with aspect ratios two to one (Δ) or four to one (◇), hexagons (lines) and icosagons (○)

Future Directions

The percolation properties of networks of random and convex plane fractures are successfully addressed by means of the excluded volume. Many important dimensionless properties of isotropic fracture networks only depend on the dimensionless density of fractures and not on the fracture shapes and sizes which represents a significant simplification.

These properties are not specific of fractures present in rocks and the same methodology can be applied for any other fracture system whatever the characteristic sizes and the nature of the material where it occurs.

These results should be generalized in several directions. In most cases, real fractures are not isotropically oriented and this feature should be incorporated in the next studies on this subject. The same is true for the homogeneous character of the network.

Bibliography

Primary Literature

1. Adler PM (1992) Porous Media: Geometry and Transports. Butterworth/Heinemann, Stoneham
2. Adler PM, Thovert J-F (1999) Fractures and fracture networks. Kluwer Academic Publishers, Dordrecht
3. Alon U, Balberg I, Drory A (1991) New, heuristic, percolation criterion for continuum systems. Phys Rev Lett 66:2879–2882
4. Balberg I (1985) Universal percolation threshold limits in the continuum. Phys Rev B31:4053–4055
5. Balberg I (1987) Recent developments in continuum percolation. Phil Mag B56:991–1003
6. Balberg I, Anderson CH, Alexander S, Wagner N (1984) Excluded volume and its relation to the onset of percolation. Phys Rev B30:3933–3943
7. Barenblatt GI, Zheltov IP, Kochina IN (1960) Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks. Soviet Appl Math Mech (PMM) 24:852–864
8. Berkowitz B, Adler PM (1998) Stereological analysis of fracture network structure in geological formations. J Geophys Res B103:15339–15360
9. Berkowitz B, Ewing RP (1998) Percolation theory and network modeling applications in soil physics. Survey Geophys 19:23–72
10. Bogdanov I, Mourzenko VV, Thovert J-F, Adler PM (2003) Effective permeability of fractured porous media in steady state flow. Water Resour Res 39. doi:10.1029/2001WR000756
11. Bour O, Davy P (1997) Connectivity of random fault networks following a power law fault length distribution. Water Resours Res 33:1567–1583
12. Charlaix E, Guyon E, Rivier N (1984) A criterion for percolation threshold in a random array of plates. Solid State Commun 50:999–1002
13. Conrad F, Jacquin C (1973) Représentation d'un réseau bidimensionnel de fractures par un modéle probabiliste. Application au calcul des grandeurs géométriques des blocs matriciels. Rev IFP 28:843–890
14. Drory A, Berkowitz B, Parisi G, Balberg I (1997) Theory of continuum percolation. III. Low-density expansion. Phys Rev E56:1379–1395
15. Florian R, Neda Z (2001) Improved percolation thresholds for rods in three-dimensional boxes. oai:arXiv.org:cond-mat/0110067
16. Garboczi EJ, Snyder KA, Douglas JF, Thorpe MF (1995) Geometrical percolation threshold of overlapping ellipsoids. Phys Rev E 52:819–828
17. Gonzalez Garcia R, Huseby O, Thovert J-F, Ledésert B, Adler PM (2000) Three-dimensional characterization of fractured granite and transport properties. J Geophys Res 105(B)21387–21401
18. Huseby O, Thovert J-F, Adler PM (1997) Geometry and topology of fracture systems. J Phys A30:1415–1444
19. Isihara A (1950) Determination of molecular shape by osmotic measurement. J Chem Phys 18:1446–1449
20. Koudina N, Gonzalez Garcia R, Thovert J-F, Adler PM (1998) Permeability of three-dimensional fracture networks. Phys Rev E 57:4466–4479
21. Ledésert B, Dubois J, Velde B, Meunier A, Genter A, Badri A (1993) Geometrical and fractal analysis of a three-dimensional hydrothermal vein network in a fractured granite. J Volcanol Geotherm Res 56:267–280
22. Long JCS, Remer JS, Wilson CR, Witherspoon PA (1982) Porous media equivalents for networks of discontinuous fractures. Water Resour Res 18:645–658
23. Mourzenko V, Thovert J-F, Adler PM (2004) Macroscopic permeability of three dimensional fracture networks with power law size distribution. Phys Rev E 69:066307
24. Mourzenko V, Thovert J-F, Adler PM (2004) Percolation of three-dimensional fracture networks with power-law size distribution. Phys Rev E 72:036103
25. Odling NE (1997) Scaling and connectivity of joint systems in sandstones from western Norway. J Struct Geol 19:1257–1271
26. Piggott AR (1997) Fractal relations for the diameter and trace length of disc-shaped fractures. J Geophys Res 102(B):18121–18125
27. Pike GE, Seager CH (1974) Percolation and conductivity: A computer study. I. Phys Rev B10:1421–1434
28. Rivier N, Guyon E, Charlaix E (1985) A geometrical approach to percolation through random fractured rocks. Geol Mag 122:157–162
29. Robinson PC (1983) Connectivity of fracture systems - A percolation theory approach. J Phys A16:605–614
30. Robinson PC (1984) Numerical calculations of critical densities for lines and planes. J Phys A17:2823–2830
31. Saar MO, Manga M (2002) Continuum percolation for randomly oriented soft-core prisms. Phys Rev E 65:056131
32. Sahimi M (1995) Flow and transport in porous media and fractured rocks. VCH, Weinheim
33. Sahimi M, Yortsos TL (1990) Applications of Fractal Geometry to Porous Media: A review. Society of Petroleum Engineers. Paper 20476
34. Santalo LA (1943) Sobre la distribution probable de corpusculos en un cuerpo, deducida de la distribution en sus secciones y problema analogos. Rev Unión Mat Argent 9:145–164
35. Sher H, Zallen R (1970) Critical density in percolation processes. J Chem Phys 53:3759–3761
36. Sisavath S, Mourzenko V, Genthon P, Thovert J-F, Adler PM (2004) Geometry, percolation and transport properties of frac-

ture networks derived from line data. Geophys J Int linebreak157:917–934
37. Stauffer D, Aharony A (1994) Introduction to Percolation Theory, 2nd edn. Taylor and Francis, Bristol
38. Thovert J-F, Adler PM (2005) Trace analysis for fracture networks of any convex shape. Geophys Res Lett 31:L22502
39. Warburton PM (1980a) A stereological interpretaion of joint trace data. Int J Rock Mech Min Sci Geomech Abstr 17:181–190
40. Warburton PM (1980b) Stereological interpretation of joint trace data: Influence of joint shape and implication for geological surveys. Int J Rock Mech Min Sci Geomech Abstr 17:305–316

Books and Reviews

Bear J, Tsang C-F, de Marsily G (1993) Flow and contaminant transport in fractured rock. Academic Press, San Diego
Myer LR, Tsang CF, Cook NGW, Goodman RE (1995) Fractured and jointed rock masses. Balkema, Rotterdam
van Golf-Racht TD (1982) Fundamentals of fractured reservoir engineering. Developments in Petroleum Science, vol 12. Elsevier, Amsterdam

Percolation, Introduction to

Muhammad Sahimi
Mork Family Department of Chemical Engineering and Materials Science, University of Southern California, Los Angeles, USA

Why is percolation theory relevant to the analysis of complex systems? The question can be answered only if we first define what we mean by a complex system. I spent a large amount of time in vain, searching for a "clean," generally-accepted definition of a complex system, until I finally realized that there are probably as many definitions as the number of scientists that deal with complex systems – there is not a clean *universal* definition of a complex system.

But, at the very least, we can agree that a complex system consists of a large number of interacting components, or parts. The interactions may be short- or long-ranged, and may or may not change with time. One type of such interactions is the *connectivity*, the way the components or parts of a complex system are connected with each other. Clearly, if the components or parts of a complex system are not connected, they do not interact with each other, at least not directly. Now, if the connectivity of the components or parts of a complex system plays an important role on the macroscopic, or effective, properties of the system, then, percolation theory plays a prominent – and in fact a decisive – role in quantifying the effect of the connectivity on the effective properties of that complex system, hence the inclusion of this section in the Encyclopedia.

Percolation processes are, in fact, the *opposite* of diffusion processes. In the latter case, the diffusant decides where to diffuse or move. The medium in which the diffusant is moving does not have any influence on the motion. This explains why many diffusion processes can be reduced to problems in essentially one-dimensional systems (all that matters is the distance r of the diffusant from the origin of its motion) which are, therefore, amenable to rigorous theoretical analysis and analytical solutions.

In a percolation process, on the other hand, it is the medium that decides where a species can go. Therefore, if the connectivity of the components or different parts of the medium – the complex system – is poor, the species cannot go far. If, on the other hand, the parts are well-connected, then the species is free to go almost anywhere. In political jargon, a complex system in which the connectivity of its different parts or components plays a decisive role in determining its effective properties – i. e., a percolation system – is like a corrupt society in which what matters is the connectivity to the powerful people! If one is well connected, one can advance rapidly; if not, there is little prospect for advancement. Since most media of interest are three- or at least two-dimensional systems, percolation problems are far more difficult to solve than the diffusion problems.

Why should the connectivity of a complex system be poor? Because natural complex systems are *disordered*. In fact, Nature, the most complex system that we know of, is disordered. Pure and geometrically perfect (periodic) systems are nowhere to be found, except perhaps in books and in our imaginations. One way that the disorder manifests itself is in the connectivity of the parts or components of a complex system. In some sectors of the system the parts are well connected, while in other sectors they are not. But, what matters most is the *overall* connectivity of the system, so that a given phenomenon can take place *across* the system.

An illuminating example is provided by porous materials. A piece of rock is a disordered porous medium: The shapes, sizes, and orientations of the pores are not identical, but vary greatly. If we attempt to characterize their statistical distributions, we find them to be broad. Thus, such a porous medium is a highly disordered and complex system. Clearly, the way the pores are connected plays a crucial role in flow of a fluid through the medium, which is why percolation is relevant to the description of fluid flow in disordered porous medium.

But, many of the man-made systems are also disordered, and the connectivity of their parts is a con-

trolling factor in determining their effective properties. For example, small molecules, or monomers, react, form chemical (covalent) bonds between themselves, and create a macromolecule. The structure of such a macromolecule is highly disordered. Clearly, the effective properties of the macromolecule is controlled by the connectivity of the monomers or the small molecules that formed it. If each monomer reacts with only two other monomers, its has a connectivity of two. The macromolecule has, therefore, no branches or loops. If, on the other hand, each monomer reacts with several others and has a higher connectivity, one obtains branched polymers and gels, the properties of which are completely different from those of linear polymers.

As another example, consider electrical conduction through a composite material which is a mixture of conducting and insulating phases. Assume that the two phases are randomly distributed in the composite. For simplicity, we model the composite material by a simple-cubic network in which each bond is either conducting with a finite conductivity, or insulating with zero conductivity. Suppose that we impose a voltage difference between two opposite faces of the network. The question then is: what fraction of the bonds must have a finite conductivity in order for the electrical current to flow through the material, so that it would have a nonzero macroscopic conductivity? This is clearly an important practical question, because its answer tells us, for example, what (volume) fraction of a composite material, such as carbon black composites that are used in many applications, must be conducting in order for the composite as a whole to be conducting.

If too many bonds are insulating, no macroscopic current will flow through the material, whereas for sufficiently large number of conducting bonds electrical current does flow in the material, so that its macroscopic effective conductivity is nonzero. If the fraction of the conducting bonds is p, then, there must be a minimum or critical value p_c of p, such that for $p \leq p_c$ no electrical current would flow through the material and, therefore, the material as a whole is insulating, whereas for $p > p_c$ the material becomes conducting.

The quantity p_c signals a phase transition: for $p \leq p_c$ there is no sample-spanning path of conducting bonds, so that the material is macroscopically insulating – the conducting bonds are not macroscopically connected. For $p > p_c$, on the other hand, the system becomes macroscopically conducting – the conducting bonds are macroscopically connected. Hence, p_c is the point at which a *geometrical* phase transition from a disconnected to a connected system takes place. Percolation theory, then, quantities the phase transition and its effect on the macroscopic properties of a complex system. p_c is called the *percolation threshold* of the medium.

Determination of the exact percolation thresholds of many 2D and all the 3D lattices remains an unsolved problem. John Wierman's article, ► Percolation Thresholds, Exact, describes and discusses the existing exact results for the percolation thresholds. Robert Ziff's article, ► Percolation Lattices, Efficient Simulation of Large, describes highly efficient numerical methods for estimating the percolation threshold and many other properties of percolation lattices, while Dietrich Stauffer's article, ► Scaling Properties, Fractals, and the Renormalization Group Approach to Percolation, describes the theoretical foundations for many important percolation properties near the percolation threshold.

The aforementioned articles describe percolation systems in which there is no correlations, and the complex system is completely random. However, disorder in many important heterogeneous materials is not completely random. There usually are correlations with an extent that may be finite but large. For example, in packing of solid particles, there are short-range correlations. Moreover, if the correlation function $C(r)$ decays as r^{-d} or faster, where d is the Euclidean dimensionality of the system, then many properties of the system are very similar with those of random percolation. This is not totally unexpected because even in random percolation, as p decreases toward p_c, correlations begin to build up and, therefore, the introduction of any type of correlation with a range shorter than the percolation correlation length cannot change the properties fundamentally. In many other cases, e. g., in some disordered elastic materials, there are very strong correlations. The article by Antonio Coniglio and Annalisa Fierro, ► Correlated Percolation, describes the major differences between percolation in random and correlated systems.

A particular type of percolation model with extended correlations is known as the *bootstrap percolation*. In this problem the sites of a lattice are initially randomly occupied. Then, those sites that do not have at least Z_c nearest-neighbor occupied sites are removed (note that $Z_c = 0$ is the usual random percolation). The interactions between the sites are short-ranged, but the correlations between them may build up as the distance between two occupied sites also increases. It now appears that bootstrap percolation possesses many unusual properties and, in fact, simulating and obtaining accurate estimates of this particular type of percolation systems have proven to be very difficult. The article by Paolo De Gregorio, Aonghus Lawlor, and Kenneth Dawson, ► Bootstrap Percolation, describes the progress in this important area of percolation prob-

lems, which appears to have applications to many important phenomena.

The example of conduction in composite materials that we described earlier also provided a hint on how complex percolation systems are modeled: They are represented by lattices or networks, such as the simple-cubic lattice. However, percolation in continua is also of great interest, since in most, if not all, of the practical applications one must deal with continuous systems. For example, continuum percolation is directly applicable to characterization and modeling of the morphology and effective transport properties of microemulsions, polymer blends, sintered materials, sol-gel transitions, and many other important practical problems. The article by Isaac Balberg, ► Continuum Percolation, describes the advances that have been made in understanding of the percolation effects in continuous systems.

Although percolation in lattice and continuous models has been studied extensively, in recent years another type of lattice model has attracted wide attention, as it appears to be applicable to a wide variety of phenomena, ranging from social dynamics to biological systems. These are lattices in which a small fraction of the sites – called the *hubs* – are highly connected, while a vast fraction of the remaining sites are connected only sparsely. Thus, the connectivity k of the nodes follows a statistical distribution. The connections are no longer necessarily between nearest-neighbor nodes. In fact, in such models the usual notion of nearest-neighbor nodes is not even useful or applicable. Some of such networks are *scale-free*. These are lattices in which the distribution of the nodal connectivities are of power-law type, $P(k) \sim k^{-\gamma}$, where γ can be quite large. Percolation in such scale-free networks has very unusual properties. The article by Reuven Cohen and Shlomo Havlin, ► Percolation in Complex Networks, provides a review of this rapidly developing field. They describe how the concepts of percolation can be used to study not only the robustness and vulnerability of random networks, but also such problems as immunization and epidemic spreading in populations and computer networks, communication paths, and fragmentation in social networks.

Other major applications of percolation theory include modeling of transport in disordered materials, and in particular composite solids, and porous media. The transport processes that have been studied include fluid flow, diffusion, conduction, and deformation (stress transport), including computation of the elastic moduli. One important application of the percolation concepts is to two-phase fluid flow in porous media. In such problems, one fluid is injected into a porous medium – that is, it invades the medium – in order to displace a second fluid already in the medium. A well-known example, practiced by the oil industry, is displacement of oil in an oil reservoir by water which is injected into the reservoir through some injection wells. Thus, this particular model is usually known as the invasion percolation. The article by Mark Knackstedt and Lincoln Paterson, ► Invasion Percolation, describes in detail how percolation is used to describe two-phase flow in porous media, although invasion percolation has proven to be relevant to many other phenomena. Other aspects of the application of percolation to problems that deal with fluid flow through porous media are described in the article by Peter King and Mohsen Masihi, ► Percolation in Porous Media.

The article by Barry Hughes, ► Conduction and Diffusion in Percolating Systems, provides a detailed description of diffusion and conduction in disordered and composite materials, including porous media, and presents a comprehensive account of the state-of-the-art of this important application of percolation to a problem of great practical importance.

Natural porous media are often fractured. For example, most oil reservoirs in the Middle East are fractured. The fractures often a large connected network, without which many oil reservoirs (such as those in Iran) could not produce any oil. Fractures also play an important role in fast transport and spreading of contaminants in groundwater aquifers. At the same time, natural porous media often contain large faults (one of the most famous of which, the San Andreas fault, is in California), which play the primary role in earthquakes. It now appears that percolation provides a powerful tool for modeling of the effect of the connectivity of fractures and faults on fluid flow and transport properties of rock, a highly complex set of phenomena. Fracture and faults also affect other important phenomena in rock, such as propagation of elastic and seismic waves. In their article, ► Percolation, and Faults and Fractures in Rock, Pierre Adler, Jean-Francois Thovert, and Valeri Mourzenko describe the recent progress in the application of percolation to this highly important set of problems.

Deformation and stress transport in disordered materials are important to a wide variety of phenomena in science and technology, ranging from elastic properties of solid materials, to viscoelastic properties of polymers and gels, and rigidity of biological materials (cells, proteins, bones, etc.). The application of percolation to modeling of such phenomena has proven to be highly successful and fruitful. The article by Phillip Duxbury, ► Elastic Percolation Networks, provides a comprehensive discussion of the subject, and describes the theoretical foundations and

computer simulation methods for stress transport in disordered percolation systems.

Two other articles expand on what Duxbury describes in his article. In his article, ► Networks, Flexibility and Mobility in, Michael Thorpe describes recent advances on generalization of the percolation model, and its application to modeling of proteins and other biological materials. The question of the rigidity of such materials is addressed. An important and well-established application of percolation is to modeling of the rheology of polymers and gels, particularly in the vicinity of the gelation point. Several variants of the percolation models have been developed in order to address this important problem. The article by Muhammad Sahimi, ► Percolation and Polymer Morphology and Rheology, describes the advances that have been made in this area.

Over the past three decades percolation theory has been applied to modeling of a wide variety of phenomena in disordered media and complex systems. It is impossible to describe and discuss all such applications. The applications that are described in this section of the Encyclopedia represent just the tip of the iceberg, but they do provide the reader with a good understanding of the power of percolation theory for describing a wide variety of important phenomena in complex systems. Have a good read!

Percolation Lattices, Efficient Simulation of Large

Robert M. Ziff
Department of Chemical Engineering,
University of Michigan, Ann Arbor, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Cluster Identification and Growth
Hull Walks and Hull-Generating Walks
Gradient Percolation
The Microcanonical-Canonical Method
Other Numerical Techniques
Conductivity and Backbones
Conclusions
Future Directions
Acknowledgments
Bibliography

Glossary

Hull The boundary of a percolation cluster, either internal or external.

Accessible hull The hull with pinched off "fjords" removed.

Hull-generating walk A way to generate the hull of a percolation cluster by a type of kinetic self-avoiding walk.

Queue A computer list construct in which the events are stored in such that the first in is the first out (also called "FIFO" or "breadth-first searching").

Stack A computer list construct in which the events are stored in such a way that the last in is the first out (also called "LIFO" or "depth-first searching").

Recursion A programming method in which a procedure calls itself, creating new local variables each time.

Tree A data structure in which points are connected in a tree-like structure with branches but no loops.

Stochastic Loewner evolution (SLE) A theoretical way to study conformally invariant random curves, including the hulls of percolation clusters, through a transformation of simple Brownian motion. Also called Schramm–Loewner Evolution.

Leath algorithm A technique where individual percolation clusters are "grown" from a seed by an epidemic type of process.

Hoshen–Kopelman algorithm A technique where a lattice (in 2d) is scanned one row at a time, and clusters are identified using information from the previous row only.

Newman–Ziff algorithm A way to efficiently generate microcanonical (fixed occupancy) states and from them to study all canonical (fixed p) states.

Definition of the Subject

Percolation is a simple model of the formation of long-range connectivity in random systems. While it can be solved exactly in a few cases of branched lattices, and while many results in two dimensions (2d) can be found exactly, most of the work in this field is intimately connected with computer simulation. Various algorithms have been developed over the years, and this article surveys some of them, especially related to cluster sizes and connectivity, and the hull.

Introduction

Percolation was introduced by Flory in 1941 [13] for branched networks (polymers) and Broadbent and Hammersley in 1957 for lattice networks, and its study by computer simulation began just a few years later [77]. The

overall development of the percolation field in the ensuing years has been intimately connected with advances in simulation and computer algorithms. Specific computer algorithms allow optimal simulation of different aspects of the percolating system, and the results of these simulations have provided information and ideas for theoretical developments and further understanding of this fundamental problem. Recent advances have allowed, for some examples, the determination of numerical thresholds to very high precision, the demonstration (later confirmed by theory) that universality applies to crossing and excess cluster properties, and the universality of scaling functions and their resulting amplitude ratios.

The basic percolation system is a lattice with sites (vertices) and/or bonds (edges) occupied with a given probability p, and the computational problem is to identify and characterize the clusters of adjacently occupied sites (site percolation) or of sites connected by the bonds (bond percolation), and determine properties such as the size distribution, conductivity, and crossing. In principle, these are generally rather straightforward problems, but in practice, the challenge is to do things efficiently so that large systems can be simulated over many times to get good numerical significance.

In this article, we will describe and summarize several algorithms that have been developed to simulate percolation. Explicit fragments of programming in C are given. The emphasis of the article is on the computational methods, and we use some recent examples of application to illustrate them. We do not address the large number of advances that have been made recently in the percolation field mainly by mathematicians, based upon Stochastic Loewner Evolution (SLE) [61] and related methods. For a recent review, see [24].

Another area of recent interest related to percolation is in the study of networks, including Erdős–Renyi random graphs [11], scale-free and small world networks. For these models, percolation corresponds to the formation of a "giant component". Because of the lack of loops over large distances, the percolation properties can generally be solved for analytically (i. e., [42,58]). Some of these results are closely related to percolation on the branch-free Bethe lattice, whose solution goes back to Flory and was analyzed in detail by Fisher and Essam [12]. However, in this review we will only consider regular lattices, and not the algorithms and results related to these systems.

Cluster Identification and Growth

Consider a lattice, say square for simplicity, and suppose that the sites have been "populated" by being made "occupied" (OCC) with probability p and empty or "vacant" (VAC) with probability $1-p$. (Here we are considering site percolation.) A basic problem is to identify the clusters, and to determine some property such as the size distribution n_s (the number of clusters of size s, divided by the total number of sites on the lattice) or whether crossing exists between two intervals on the boundary. Here we describe two neighbor-search methods to identify clusters: the depth-first or last-in, first-out (LIFO) method using recursion, and the breadth-first or first-in, first-out (FIFO) method using a queue.

In the following programs, we use a simple two-dimensional array `lat[x][y]` to represent a rectangular $W \times H$ system with a square lattice. For many problems it is more efficient to use a one-dimensional array and add $1, -1, W$, and $-W$ for the four directions (in 2d, for example), using wrap-around at the end to form "helical" boundary conditions, which for a large system is practically equivalent to a periodic one. (The helicity adds a "twist" to the boundaries, which can have an effect on some of the properties when it is large enough [93]). Later on, in Sect. "The Microcanonical-Canonical Method", we will give an example of a program which uses a one-dimensional array and also a neighbor array that can be used to program periodic boundary conditions precisely, with a bit more programming overhead however.

Recursive Search (LIFO)

In the recursive search method, the lattice (say of dimensions $W \times H$) is first scanned for new, unchecked sites:

```
for (xo = 0; xo < W; ++xo)
   for (yo = 0; yo < H; ++yo)
      if (lat[xo][yo] == OCC)
      { lat[xo][yo] = TAGGED;
        FindNeighbors(xo,yo); }
```
(P1)

where TAGGED means that the site has been checked and won't be checked again. Then the cluster belonging to that site is found using the `FindNeighbors` subroutine, which is given by

```
FindNeighbors(int x,y)
{ int dir;
  for (dir = 0; dir < 4; ++dir)
  { xp = x + dx[dir];
    yp = y + dy[dir];
    if (lat[xp][yp] == OCC)
    { lat[xp][yp] = TAGGED;
      FindNeighbors(xp,yp); } } }
```
(P2)

When an OCC neighbor is found for the first time, it is checked by calling the same routine over again.

Here we used the four nearest-neighbor direction vectors `dx[0]=1, dy[0]=0, dx[1]=0, dy[1]=1, dx[2]=-1, dy[2]=0, dx[3]=0, dy[3]=-1`. We have not dealt with the boundaries in this example. Open boundaries can be simulated by adding a perimeter of `VAC` sites; periodic boundary conditions can be simulated simply by writing `lat[xp&(W-1)][yp&(H-1)]` for `lat[xp][yp]`, where `&` is the bit-wise "and" operation, if `W` and `H` are exactly powers of two. As the clusters are identified, the number of occupied sites can be counted, moments determined, etc. Crossing can be determined if a single cluster touches two given boundaries. For periodic b.c., one typically considers not crossing but wraparound.

The above `FindNeighbors` program uses recursion and the compiler stores unchecked `xp,yp` and the local `dir` in the stack; recursion uses a "last in, first out" (LIFO) or "depth-first" [71] method that can cause problems (exhaust the available stack memory) for very large clusters.

Making a Queue (FIFO)

A more memory-efficient method to search for neighbors is to use a "queue" where one keeps a list of unchecked sites, and visits them in a first-in, first-out (FIFO) fashion. However, one must construct the list explicitly; the recursive method won't do it. We need functions to put and remove coordinates from the queue; here we use `%define` statement functions (which creates in-line text) for efficiency:

```
%define PutOnQueue(X,Y) \
{ xlist[putindex] = X; \
  ylist[putindex] = Y; \
  ++putindex; }
```
(P3)

and

```
%define GetFromQueue(X,Y) \
{ X = xlist[getindex]; \
  Y = ylist[getindex]; \
  ++getindex; }
```
(P4)

where we start the simulation with `getindex = putindex = 0`. The way the above lists are written, they must be dimensioned to be as large as the largest possible cluster; however, by making the list size `S` exactly a power of two, the list can be shortened and "recycled" simply by writing `xlist[putindex&(S-1)]` etc. The size `S` has only to be as large as the number of growth sites of a cluster, which grows as roughly the square root of the maximum size of the cluster. For example, for a lattice of size 1024×1024, it is more than sufficient to make $S = 4096$. To test for an error in the recycled queue, the line `if (putindex == getindex) ....` can be added after `++putindex;` in (P3).

To make use of the queue, (P1) is kept the same except that the last line is replaced by `PutOnQueue(xo,yo)`, followed by the loop

```
do
{ GetFromQueue(x,y)
  for (dir = 0; dir < 4; ++dir)
  { xp = x + dx[dir];
    yp = y + dy[dir];
    if ((lat[xp][yp]) == OCC)
    { lat[xp][yp] = TAGGED;
      PutOnQueue(xp,yp); } }
} while (getindex != putindex);
```
(P5)

When the two indices `putindex` and `getindex` are equal to each other, there are no more occupied sites to check and the search is complete.

Note that the lists `xlist` and `ylist` can also be used for the LIFO method by treating them as a stack rather than a queue – that is, by using only one index and to decrement the index when a set of coordinates is taken off the stack. This is a way to program the LIFO method without using the recursive feature of the C language.

Generating Occupied Sites or Bonds as you go – the Leath Method

In the above algorithms, the sites or bonds were made occupied or vacant ahead of time. An alternative scheme is to start with all sites in an `UNVISITED` state, and make them `OCC` or `VAC` when they are first encountered. For example, in (P5), starting from the sixth line, we would write instead

```
if ((lat[xp][yp]) == UNVISITED)
  if (random() < prob)
  { lat[xp][yp] = OCC;
    PutOnQueue(xp,yp); }
else lat[xp][yp] = VAC;
```
(P6)

where `random()` is the random number generator and `prob` is the probability. Here, the label `TAGGED` is not needed, because a site is added to the queue as soon as it is made occupied. Before putting the first site `xo,yo` on the queue, one also has to call the random number generator to determine whether that site is occupied.

When this program is applied to the growth of a single cluster (that is, starting with just one `xo` and `yo`), it is commonly called the Leath method, although it is not carried out in the same way as in [37], where larger diamond-shaped regions of the lattice around the cluster are

successively probed. In fact, the idea of looking at clusters wetting a single site was carried out in the earliest simulations of percolation [77].

The above growth scheme works particularly nicely in the case of bond percolation, especially when the clusters are characterized solely by the sites they connect, and not by the number or arrangement of the bonds that connect them. In this case, (P6) is kept exactly the same except the last line "`else lat[xp][yp] = VAC;`" is removed. This is because one can consider bond percolation as a spreading or epidemic process [18] from site to site along the occupied bonds; if a bond is not occupied, the fluid will not spread to the next site, but that does not preclude fluid from another cluster to visit that neighboring unvisited site. Sites with no bonds are considered to be clusters of size $s = 1$. Whether a bond is made occupied or not, that bond will never be considered again, so it state does not have to be remembered. Not all bonds of a cluster are generated in this method – for example, the fourth bond in a simple square arrangement will not be tested – but the four sites of the cluster will be sampled with the correct weight. In this process, one effectively generates a minimum spanning tree (with no loops) that connects every site on a cluster.

Note that this procedure checks (or grows) the cluster one growth shell at a time. Each growth shell is at a successive value of the minimum [50] or "chemical" [25] distance from the seed, so this method is useful for studying minimum distance problems including the fractal dimension $d_{\min}$. Like d_B, its value is not known exactly even in 2d and has to be determined by simulation. In 2d, its value is $d_{\min} = 1.1307(4)$ [20].

The Hoshen–Kopelman Algorithm

The Hoshen–Kopelman algorithm [29] has been a mainstay of work on percolation and is fairly well-known. It is described in some detail in the article of Stauffer in this work ► Scaling Properties, Fractals, and the Renormalization Group Approach to Percolation and so will just be described briefly here.

The main idea of this algorithm is that a percolating system (say in two dimensions) can be examined or created a row at a time, and the cluster statistics can be updated just from the knowledge of the connections in the previous row. In d-dimensions, one must remember the state of the previous surface in $(d-1)$-dimensions. The connections can be remembered in a look-up table, or a rooted-tree data structure can be used.

This algorithm can be used to analyze the clusters statistics of a given, fully populated system, or it can be used as a very memory efficient scheme (especially in 2d) to generate and analyze on the fly a large system, since only the previous row (in $2-d$) needs to be remembered. Using this scheme, Tiggemann has simulated a lattice of $4\,000\,000^2$ sites [73].

To actually identify all of the clusters on a given lattice, it is faster to sweep across the lattice using one of the neighbor-search algorithms above, because in the HK method the lattice would have to be swept a second time in order to get the most updated cluster labels.

Recently, Deng and Blöte [10] used a version of the Hoshen–Kopelman algorithm to determine $p_c = 0.5927465(4)$ for site percolation on the square lattice, and $p_c = 0.3116077(4)$ for site percolation on the simple cubic lattice, using a novel method of analysis based upon universal ratios of correlation functions and moment distributions. Tiggemann [73] has used a massively parallel version to study percolation on lattices in two, three, and four dimensions. A good discussion the the HK algorithm, its antecedents in the computer science field, and some of its modifications and extensions, is given by Martín–Herrero [41].

Example: Critical Density Plots

As an example of an application of the above simulation method, we show in Fig. 1 the average density of clusters anchored to a single point and simultaneously to two points at the boundary of a square system. Here the density is defined as the number of times a given site is connected to the corresponding boundary point, divided by the total number of realizations. We use bond percolation with the growth algorithm and the FIFO (queue) method, with `xlist` and `ylist` large enough to remember the largest possible cluster. This allows us to transfer the coordinates of the cluster to the appropriate array (touching one anchor, touching the other anchor, or touching both anchors) after the cluster is grown. In this simulation, we don't have to scan the entire lattice, but just use the two anchor points as seeds (`xo,yo`) for the cluster growth/indentification.

To analyze this problem theoretically, it is convenient to put the system on the complex plane, and consider a half-infinite system $y \geq 0$, with the boundary placed on the real axis, and anchors at x_a and x_b. The results of this calculation can be transformed to a square by a standard conformal map $w(z)$, using the transformation of the density $\rho(w) = (\mathrm{d}w/\mathrm{d}z)^{-h}(\mathrm{d}\overline{w}/\mathrm{d}\overline{z})^{-\overline{h}}\rho(z)$ where $h = \overline{h} = (2-D)/2 = 5/96$ for $2-d$ percolation, and D is the fractal dimension. Note that when the mesh size (lattice spacing) goes to zero with the boundaries fixed, the ac-

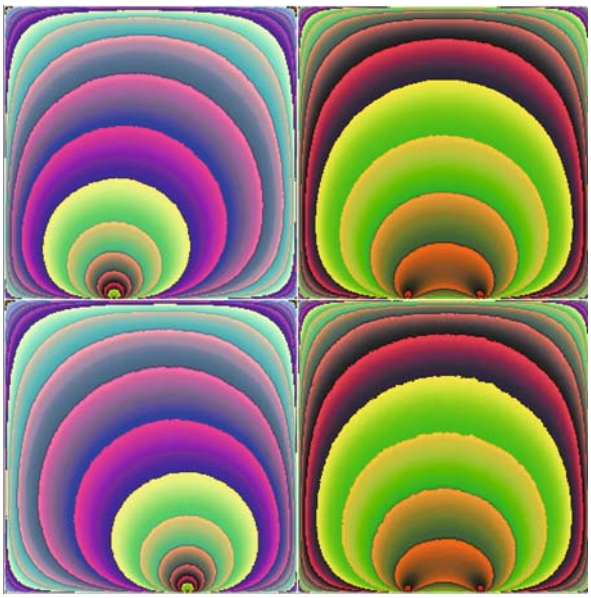

Percolation Lattices, Efficient Simulation of Large, Figure 1
Simulation results for the average density of clusters touching the left anchor (*top left*), right anchor (*bottom left*), both anchors simultaneously (*lower right*), and prediction from Eq. (2) (*upper right*). From [35]

tual density of the clusters, being fractal, goes to zero. The density we consider is effectively renormalized to remain finite and non-zero in that limit.

For a single anchor placed at x_a on the real axis, the density at a point $z = x + iy$ can be interpreted as the probability that the point z is connected to the point x_a – in other words, it is the two-point correlation function, $\mathcal{P}(z, x_a)$. This quantity is predicted to be given by [35]

$$\mathcal{P}(z, x_a) = \frac{c\, y^{11/48}}{|z - x_a|^{2/3}} \tag{1}$$

where $y = (z - \overline{z})/(2i)$ and c is a non-universal, lattice-dependent constant. It turns out that this density is closely related to that of a dipole in electrostatics – it is $y^{-5/48}$ multiplied by the potential of a dipole, raised to the 1/3 power.

Figure 1 shows the density contours in a square for single anchors $\mathcal{P}(z, x_a)$ and $\mathcal{P}(z, x_b)$, and and for two simultaneous anchors, which corresponds to the three-point correlation function $\mathcal{P}(z, x_a, x_b)$. Now, it was observed that the three-point function is proportional to the square root of the product of the two-point correlation functions and the probability $\mathcal{P}(x_a, x_b)$ that x_a and x_b are connected together [35]:

$$\mathcal{P}(z, x_a, x_b) = C\, \sqrt{\mathcal{P}(x_a, x_b)\mathcal{P}(z, x_a)\mathcal{P}(z, x_b)}\,, \tag{2}$$

where C is a constant, valid as long as z is at least several lattice spacing from the anchor points x_a and x_b. Near the anchor points for a finite mesh, C is not constant but is a function of x_a, x_b, and z. When z approaches an anchor point, say x_a, it follows (for bond percolation) that $\mathcal{P}(x_a, x_a, x_b) = \mathcal{P}(x_a, x_b)$, and $\mathcal{P}(x_a, x_a) = 1$, so at this point C is identically 1. However, when the mesh goes to zero, C is constant greater than 1 everywhere else.

Furthermore, rather surprisingly it was found that (away from the anchor points) C was the same for site and bond percolation, with the value $= 1.030 \pm 0.001$, and so appeared to be universal. After these numerical observations were made, it was shown theoretically that (2) indeed follows from conformal field theory using boundary operators [35], and the constant $C = C_{222}$ is universal and given explicitly by [65]

$$C_{222} = \sqrt{\frac{\Gamma(2/3)^3\Gamma(5/3)^2}{\Gamma(1/3)\Gamma(4/3)^3}} = \frac{2^{7/2}\pi^{5/2}}{3^{3/4}\Gamma(1/3)^{3/2}} = 1.02992679\ldots\,. \tag{3}$$

Excess Number of Clusters

Another example where precise simulations inspired a theoretical result is given by the problems of the *excess* number of clusters [92,93]. At the critical point, the number of clusters per lattice site n_c is a well-defined, finite, non-universal quantity. This quantity is known exactly for $2-d$ bond percolation on the triangular and square lattices; for the latter, Temperley and Lieb [72] found an integral expression for n_c which evaluates simply to $(3\sqrt{3} - 5)/2 = 0.098076211\ldots$ [92]. Counting the critical clusters N_c on finite lattices of size $W \times H$ with periodic boundary conditions, it was observed that

$$N_c = n_c HW + b(W/H) + \cdots \tag{4}$$

where $b(r)$ is a universal quantity that is a function of the aspect ratio $r = W/H$ but independent of the underlying percolation type (in contrast to n_c, which is not universal and varies from system to system). The universality of $b(r)$ was shown to follow as the singular part of the free energy [1], and $b(r)$ has been calculated exactly for some geometries. For example, in the limit that $r \gg 1$, that is, for a cylinder, $b(r)$ is proportional to r and is given simply by [34]

$$b(r) \sim \frac{5}{8\sqrt{3}}\, r\,. \tag{5}$$

This result is valid for all forms of critical $2-d$ percolation, is reminiscent of (but not identical to [32]) the num-

ber of wrapping critical clusters in a cylindrical system of length $r \gg 1$:

$$N_{\text{wrap}}(r) \sim \frac{1}{\sqrt{3}}\, r\,. \tag{6}$$

The universality of the excess cluster number also applies in higher dimensions, although no theoretical results are known there.

Finding p_c from the Leath Method

The Leath method can be used to generate single clusters in an empty lattice and find an unbiased measure of the size distribution. Say the procedure is started on a single seed. If the cluster grows to a size greater than or equal to some cutoff value $s_{\max}$, it is stopped. The cutoff is chosen sufficiently small so that the growth will always stop before the boundaries of the lattice are reached. In that case, the statistics of the upper cumulative size distribution are unbiased by any boundary effects, and the only finite-size effects are those imposed by the cutoff value.

We define as usual

$$n_s(p) = \text{the number of clusters containing } s \text{ sites, per lattice site} \tag{7}$$

then it follows that

$$P_s(p) = s n_s(p) = \text{the probability that a given site belongs to a cluster containing } s \text{ sites} \tag{8}$$

and

$$P_{\geq s}(p) = \sum_{s' \geq s} s' n_{s'}(p) = \text{the prob. that a given site belongs to a cluster containing } s \text{ or more sites}\,. \tag{9}$$

Here we are considering bond percolation; if it were site percolation, there would be an extra factor of p in P_s and $P_{\geq s}$ reflecting the probability that the selected site is an occupied one. The quantity $P_{\geq s}(p)$, unbiased for all $s < s_{\max}$, is determined directly by the growth of single clusters.

According to the usual scaling theory, in the scaling limit where $s \to \infty$ and $z = (p - p_c)s^\sigma$ is held constant,

$$n_s(p) \sim c_0 s^{-\tau} f(c_1(p - p_c)s^\sigma) \tag{10}$$

where c_0 and c_1 are the non-universal metric factors, while the exponents τ and σ, and the scaling function $f(z)$, are universal. (This scale variable z should not be confused with the complex coordinate z above.) To define c_0 and c_1 uniquely, one can assume for example $f(0) = 1$ and $\int_{-\infty}^{\infty} f(z)\mathrm{d}z = 1$. It follows from (10) that

$$P_{\geq s}(p) \sim \frac{c_0}{\tau - 2} s^{2-\tau} g(c_1(p - p_c)s^\sigma) \tag{11}$$

where $g(z) = [(\tau - 2)/\sigma] z^{(\tau-2)/\sigma} \int_z^\infty \zeta^{(2-\tau)/\sigma - 1} f(\zeta)\,\mathrm{d}\zeta$ for $p > p_c$, and similarly for $p < p_c$. Now it follows that if $f(z)$ is analytic near $z = 0$, then $g(z)$ is also, and we can carry out a Taylor-series expansion, yielding

$$P_{\geq s}(p) \sim s^{2-\tau}(A + B(p - p_c)s^\sigma + \cdots) \tag{12}$$

where A and B are constants. Thus, a plot of $s^{\tau-2}P_{\geq s}(p)$ vs. s^σ should give a straight line with a slope proportional to $p - p_c$.

p_c for the hcp and fcc Lattices

We illustrate this method for 3-d site percolation on the hcp and fcc lattices, from [40]. The lattice size was 2048^3, created by using a virtual-memory scheme [89] that only assigns physical memory to cubes of the space as the clusters grows into them, which works well for the growth method because only a small fraction of the lattice is accessed by an individual cluster. The cutoff was $s_{\max} = 2^{21} = 2\,048\,576$. The exponents τ and σ are known exactly in two dimensions but not three, and by fitting the data to (12), those exponents can be found by this method as well. In Fig. 2, we show a plot of $s^{\tau-2}P_{\geq s}(p)$ vs. s^σ for three close values of p, using $\tau = 2.189$ and $\sigma = 0.455$. These exponent values were consistent with this data and also that of other 3-d systems, and compare with the val-

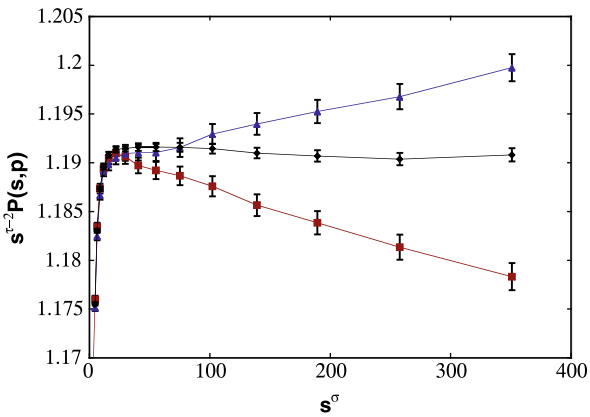

Percolation Lattices, Efficient Simulation of Large, Figure 2
Single-cluster growth statistics for site percolation on the hcp lattice, with $p = 0.1992600$, 0.1992555, and 0.1992500 (*top* to *bottom*). Here $P(s, p) \equiv P_{\geq s}(p)$. From [40]

ues 2.18906(8) and 0.4522(9) respectively found by Ballesteros et al. [3], and 2.18958(9) and 0.4535(2) by Deng and Blöte [10]. The plot slows clearly that for large s, the behavior predicted by (12) is well followed. For s less than about 1000, there are significant deviations due to the finite-size effects of the lattice discreteness. These finite-size effects at p_c can be fit asymptotically by an equation of the form $P_{\geq s}(p_c) \sim s^{\tau-2}(A + Cs^{-\Omega})$ with $\Omega \approx 0.64$ [39].

First the simulations were run at the two outside values of p, and then (12) was used to extrapolate $p_c = 0.1992555$, which was then verified by a third run at this value, which is seen to be horizontal (for large s) in the plot. Analyzing the errors yields $p_c = 0.1992555(10)$. A similar calculation for the closely related fcc lattice yields the slightly but statistically significant lower value $p_c = 0.1992365(10)$. For bond percolation, however, the thresholds for these two lattices at this level of precision are the same $p_c \approx 0.120164$, although they are likely to be different if measured to higher precision. Clearly, a very sensitive method, like the one presented here, is needed to distinguish such close thresholds.

Another application of single cluster growth has been to find amplitude ratios of the mean cluster size an equal distance below and above p_c, yielding a value $\Gamma^-/\Gamma^+ = 163 \pm 2$ in $2-d$ [33], much more precise than earlier determinations, and confirmed by high-order series expansions [33]. Finally, we note that a similar method of analysis has been applied to directed percolation in 2+1 dimensions [49] leading to the determination of the threshold and critical exponents to higher accuracy than previous works [22,74].

Hull Walks and Hull-Generating Walks

The hull in two-dimensional percolation is the boundary between occupied and vacant "perimeter" sites of a percolation cluster. A typical cluster has both external and internal hulls, and an infinite cluster at the critical point has an infinite number of hulls within hulls. The fractal dimension of critical hulls was first conjectured (based upon simulations) to be simply 7/4 [60], and then this conjecture was proven theoretically first from field theory [59] and more recently using Stochastic Loewner Evolution (SLE) [67].

The "accessible" or Grossman–Aharony hull is the hull (not necessarily external) in which closed-off inlets or "fjords" are bridged and the hull shortened [23]. It turns out that this hull is also a fundamental measure of percolation clusters and at the critical point has a fractal dimension of 4/3, identical to self-avoiding walks and the hulls of simple Brownian motion.

Hull-Walk Algorithm

Hulls can be identified on an existing percolation cluster by carrying out a walk that follows the edges of the cluster, as first studied by Voss [76]. On the other hand, just like in the cluster growth algorithm, one can start with an unvisited (undetermined) lattice and decide upon the occupancy of the sites or bonds as they are encountered, and thus generate the hull at the same time as it is being identified. This idea was proposed for site percolation on the square lattice in [89], site percolation on the triangular lattice in [78], and bond percolation on the square lattice in [19].

For bond percolation, the easiest formulation of the walk is to follow a path that jumps between the centers of the occupied and vacant bonds on the perimeter – or, equivalently, between the bonds on the lattice and the dual lattice. For bond percolation on a square lattice, the paths also follow a square lattice, rotated at 45° from the bond lattice. When an occupied bond is encountered, the walk turns clockwise, while when a vacant bond (or bond on the dual lattice) is encountered, it turns counter-clockwise.

In the hull-generating procedure, when an `UNVISITED` bond is encountered, that bond is made `OCC` with probability p and the walk turns clockwise, and made `VAC` with probability $1-p$, and the walk turns counter-clockwise. When an already visited site is encountered, the walk turns in such a way that the path always avoids itself. Following is a piece of a program that carries out these steps:

```
dir = 100; x = xo; y = yo;
do
{ x += dx[dir & 3];
  y += dy[dir & 3];
  switch (lat[x][y])
  { case UNVISITED:
      if (random() < prob)
      { lat[x][y] = OCC;
        ++nocc; --dir;
      } else
      { lat[x][y] = VAC;
        ++nvac; ++dir;
      } break;
      case VAC: ++dir; break;
      case OCC: --dir;
  }
} while ((x != xo) && (y != yo)
          && ((dir&3) != 0));
```
(P7)

Here we have rotated the original lattice by 45° so that the walk moves in horizontal and vertical directions and the bonds are effectively on the diagonals. The "`& 3`" construction is equivalent to "`% 4`" (modulo 4). The walk

ends when it returns to the starting point and it is going the same directions as it started out.

For lattices other than the square one, it is generally convenient to transform the lattice so that it fits on a square one. For example, a triangular lattice can be put on a square lattice with one diagonal bond put in. Then the hull walk has to be constructed between the centers of these bonds, with is rather intricate. An alternative approach is to remain on the simple square lattice with the walk moving in the four diagonal directions, but to make some bonds permanently occupied and/or vacant to simulate the particular lattices. For example, to create a triangular lattice, half of the horizontal bonds, alternating on each row, can be made permanently occupied. Making the same bonds permanently vacant gives the honeycomb lattice, etc.

For site percolation, one can use a variation of the above program in which the walk steps along occupied perimeter sites, always keeping the vacant sites on one side. Details are given in [89].

Finding p_c Directly from the Hull-Generating Walk

Starting from an `UNVISITED` lattice and a single seed, the hull-generating walk above will always close on itself, forming either external or internal hulls, depending upon the direction of closing. By making a simple hypothesis that at the critical point the internal and external hulls are equally likely (for large hulls), one can deduce an estimate for the critical point. For site percolation on a square lattice this method gave $p_c = 0.59275(3)$ in work from over 20 years ago [82]. This approach has not been pursued further, nor have questions of its finite-size effects been explored.

For p away from p_c, the statistics of the internal vs. external hulls will be much different. It was found that the average number of occupied sites in a hull $\langle s_H \rangle$, for site percolation on the square lattice, satisfies [89]

$$\begin{aligned} \langle s_H \rangle_{\text{ext}}^{-} &\sim \langle s_H \rangle_{\text{int}}^{+} \sim A|p - p_c|^{-2} \\ \langle s_H \rangle_{\text{int}}^{-} &\sim \langle s_H \rangle_{\text{ext}}^{+} \sim B|p - p_c|^{-2} \end{aligned} \tag{13}$$

with $A \approx 0.5$ and $B \approx 0.004$, where "int" and "ext" represent internal and external hulls, respectively, and + and − represent above and below p_c, respectively. (The simple exponent −2 above follows from scaling relations of the hull exponents and $D_H = 7/4$ [78,82].) These results imply that the average hull size shows an amplitude ratio of $A/B \approx 125$, reflecting the huge difference between the average size of these two kinds of hulls. While amplitude ratios play an important role in statistical mechanics [52], this ratio has not been studied further.

The Enclosed Area Distribution

A more recent application of the hull-generating method has been to find the enclosed area distribution $2 - d$. (Even though both the cluster and the hull are fractal, with dimension 91/48 and 7/4, respectively, the area enclosed by a hull is non-fractal and is proportional to the radius squared.) A simple argument from the scaling relation $n_s \sim c_0 s^{-\tau}$, the fractal relation $s \sim A^{D/d}$ and the hyperscaling relation $\tau - 1 = d/D$ implies that, in a critical system of total area A_{tot}, the number of clusters whose enclosed area is greater than A is given by [6]

$$N_{\geq A} \sim C A_{\text{tot}}/A \tag{14}$$

for A large compared to the mesh area A_0 and small compared to A_{tot}. (For $d > 2$, A represents the volume of an enclosing sphere or rectangular solid, and C is different.) The area distribution can be found from an ensemble of individual closed hulls (loops) generated by the hull-walk algorithm, and it is a simple addition to the program to calculate the enclosed area of the walk on the fly, while the walk is carried out. The coefficient C was found from simulations to be a universal constant 0.022976(5), and proven theoretically to be simply [6]

$$C = 1/(8\sqrt{3}\pi) = 0.022972037\ldots \tag{15}$$

via a conformal transformation to the problem of the number of clusters wrapping a cylinder given above in (6). Equation (14) represents a completely universal formulation of the size distribution at criticality (in contrast to $n_s \sim c_0 s^{\tau}$ which involves both the non-universal c_0, *and* the non-universal measure of the size, s). Another way to express (14) is in Zipf's-law form: if you rank-order all the hulls in the system by their enclosed area, then the area of the nth ranked hull is inversely proportional to n and is given by CA_{tot}/n for large A. Other forms of the universal size distribution are given in [93].

Applications of the Hull-Generating Walk to Crossing Problems

Hull generating walks can be used to efficiently test for crossing or spanning. For example, consider a rectangular system. A walk is started in the lower left-hand side, and represents the boundary between the occupied bonds above it and the vacant bonds below it. If the walk reaches the right boundary before reaching the top boundary, then there is horizontal crossing, while if it reaches the top before reaching the right-hand side, then there is no horizontal crossing. This process is more efficient than filling the entire lattice with clusters, because only the bonds along

the hull are simulated. It allowed a sensitive test [83,86] of the finite-size corrections to the crossing probability for a rectangle, which in an important development in percolation theory was found by Cardy to be given by [5]:

$$\Pi_h(r, p_c) = C_{112}^2\, \lambda^{1/3}\, {}_2F_1(1/3, 2/3; 4/3; \lambda)\ , \tag{16}$$

with $C_{112}^2 = 2\pi\sqrt{3}/\Gamma(\frac{1}{3})^3 = 0.56604668\ldots$, where ${}_2F_1$ is the hypergeometric function and the subscript h signifies the horizontal crossing probability. Here r is the aspect ratio of the rectangle, and is related to λ by $\lambda = ((1-k)/(1+k))^2$ and $r = 2K(k^2)/K(1-k^2)$, where $K(m)$ is the elliptic integral. This parametrization can be simplified to $r = K(1-\lambda)/K(\lambda)$ and inverted explicitly as [85] $\lambda = \vartheta_2^4(\exp(-\pi r))/\vartheta_3^4(\exp(-\pi r))$ where $\vartheta_n(q)$ are elliptic theta functions. Finally, the parameter λ, which represents the cross-ratio of the coordinates of the corners of the rectangle when mapped to the half plane, can be eliminated to yield a closed-form explicit expression for $\Pi_h(r, p_c)$, differentiated with respect to r [85]:

$$\begin{aligned}\frac{\partial \Pi_h(r, p_c)}{\partial r} &= -\frac{1}{3}\pi C_{112}^2 \vartheta_1'(e^{-\pi r})^{4/3} \\ &= -\frac{2^{4/3}}{3}\pi C_{112}^2 \eta(ir)^4\end{aligned} \tag{17}$$

where $\eta(\tau)$ is the Dedekind eta function. Equation (17) implies the series

$$\Pi_h(r, p_c) = \frac{2^{4/3}\pi C_{112}^2}{3}\left(e^{-\pi r/3} - \frac{4}{7}e^{-7\pi r/3}\ldots\right)\ . \tag{18}$$

The prediction for $\Pi_h(1, p_c) = 1/2$ the a system with a square boundary for site percolation on a square lattice of size $L \times L$ was verified in [83]. The hull-walk used in that work is illustrated in Fig. 3, here rotated so that crossing is considered in the vertical rather than horizontal direction. The behavior that crossing translates into the walk hitting one boundary before the other is exactly analogous to problems solved in SLE.

This work showed that the finite-size corrections to $\Pi_h(1, p_c) = 1/2$ are for large systems described by

$$\Pi_h(1, p_c) = 1/2 + 0.319/L + \cdots\ . \tag{19}$$

The numerical results were not consistent with a significant contribution from the "irrelevant" scaling variable $L^{-0.85}$ [84] and later it was shown that indeed because of the symmetry of the square system, the irrelevant term does not contribute here [30]. A consequence of (19) is that several estimates of p_c based upon measurements of $\Pi_h(1, p_c)$ do not converge with the usual scaling $\sim L^{1/\nu}$ but instead with the scaling $L^{-1-1/\nu}$ [83].

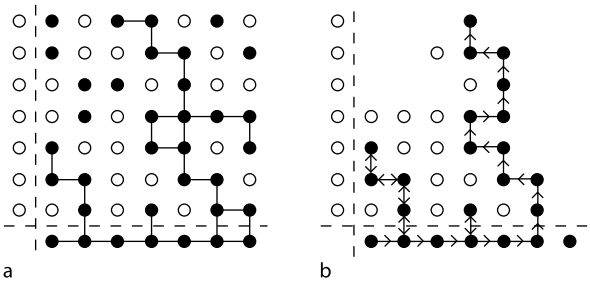

Percolation Lattices, Efficient Simulation of Large, Figure 3
Hull walk for a test for vertical crossing for site percolation on the square lattice. a all clusters connecting the bottom to the top. b the equivalent hull that would be generated from the site-percolation hull-generating walk with the same probability. *Solid circles*: occupied sites. *Open circles*: vacant sites. From [83]

The hull method also is very efficient for exact enumeration. Basically, by stepping through every possible hull, one can determine the polynomials for $\Pi_h(r, p)$ as a function of p, for a fixed r and system size. This way $\Pi_h(1, p)$ could be found for square system size $L \times L$ for L up to 7, [83,94] which would be very difficult to do with a complete exact enumeration, as this would require 2^{49} realizations of the lattice. For the first few values of L, $\Pi_h^{(L)}(1, p)$, written as a series in $p^i q^{L^2-i}$ where $q = 1-p$, is given by

$$\Pi_h^{(2)}(1, p) = 2p^2q^2 + 4p^3q + p^4 \tag{20}$$

$$\begin{aligned}\Pi_h^{(3)}(1, p) = {}& 3p^3q^6 + 22p^4q^5 + 59p^5q^4 \\ &+ 67p^6q^3 + 36p^7q^2 + 9p^8q + p^9\ .\end{aligned} \tag{21}$$

Polynomials for L up to 4 are given in [56] and those for L up to 7 can be found in [94]. These results can be used for a variety of studies; for example, looking at the kertosis of the distribution, it was shown [84] that the distribution of first crossings $(\partial/\partial p)\Pi_h(r, p)$ is not a gaussian curve as had been previously thought. This work was followed by a more general mathematical proof [4] that the tails behave like $\ln[(\partial/\partial p)\Pi_h(r, p)] \sim -L/\xi \sim -L|p - p_c|^{4/3}$ where $\xi \sim |p - p_c|^{-\nu}$ with $\nu = 4/3$ in 2d is the correlation length.

In recent work [66], the hull-generating method was used to test a generalization of Cardy's formula that describes the probability density that crossing clusters have lower edges at $y = a$ and $y = b$ on the left and right-hand boundaries, respectively, with various conditions on whether the cluster touches the bottom. Figure 4 shows how this problem is simulated by a hull walk, here for bond percolation with the bonds themselves horizontal

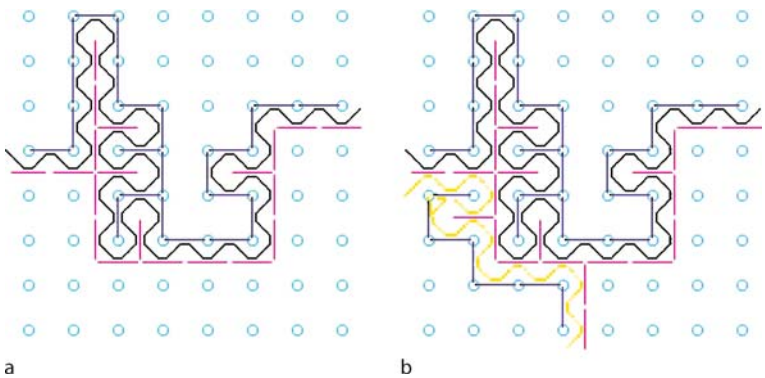

Percolation Lattices, Efficient Simulation of Large, Figure 4
a The hull-generating walk (in *black*) used to test for a cluster, whose lower edge is half-way up the left-hand side, crossing to the right-hand side, and to find the distribution of values *y* for where it hits on the right-hand side. b an additional walk (in *yellow*) to check that there are no other clusters crossing below the given walk. From [66]

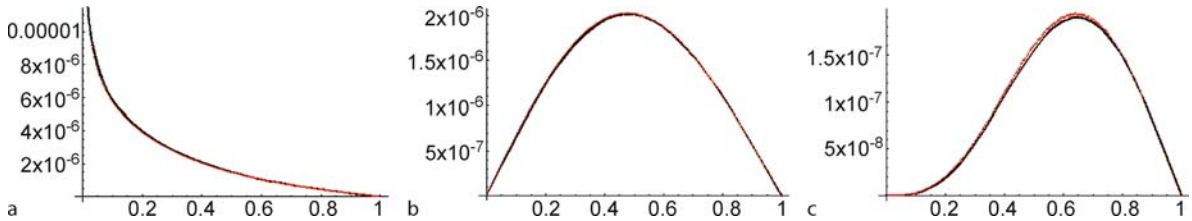

Percolation Lattices, Efficient Simulation of Large, Figure 5
Measurements (*points*) and theory (*lines*) for the distribution of the lower boundary on the right-hand side, of clusters whose lower boundary on the left-hand side is at $y = 1/2$. a Clusters that touch the bottom, b no restriction on the crossing of the clusters, and c clusters that cross from left to right but do not touch the bottom, and have no crossing clusters below them. From [66]

and vertical, and the steps along the diagonals. Figure 5 shows excellent agreement between the simulations and the theory. The hull-generating walk proved very efficient for this problem, since walks that hit one of the forbidden boundaries (used to enforce the crossing criteria) were stopped without generating the rest of the hull. Thus, in a few day's of computer work, it was possible to simulate $3.3 \cdot 10^{11}$ hulls on a lattice of 512×512 bonds, something that would be nearly impossible to carry out if all lattice bonds were considered in the simulation.

Gradient Percolation

Sapoval, Rosso and Gouyet [60] first considered percolation in a gradient, and showed that in 2d it is an efficient way to find that percolation threshold [57]. In this approach, a rectangular system is set up with a linear gradient in p, going from 0 to 1 as y goes from 0 to 1. There will be a percolating cluster connected to the upper boundary, and the hull of that cluster will sample values of p that are close to p_c. Sapoval et al. [60] showed that the hull will stay within a relatively small region $L^{4/7}$ of the lattice of inverse gradient $|\nabla p|^{-1} = L$. Therefore, as $L \to \infty$, the "frontier" will be localized about p_c.

For a finite L, two measures of $p_c(L)$ are (i), the average value of p of all the occupied+vacant sites (or bonds) of the hull, or, (ii) just the fraction of occupied to total bonds along the hull:

$$p_c^{\text{est}}(L) = \frac{n_{\text{occ}}}{n_{\text{occ}} + n_{\text{vac}}} \,. \tag{22}$$

For large systems (small gradients), these measures should be asymptotically equivalent. Simulating system of size up 1000×1000 for site percolation on the square lattice, Rosso et al. found that the estimates fell on a straight line when plotted as a function of the gradient $1/L$, with as extrapolated value of 0.592805(10), slightly higher than the values we have seen above.

By combining gradient percolation with the hull-generating walk, one can create a very efficient and simple method to determine percolation thresholds in two dimensions [90]. Basically, the program (P7) is used, with `prob` now a function of y, and periodic boundary conditions in the horizontal direction. The "front" (x coor-

dinate) of the walk is kept track of, and when it reaches a new value, the column (all values of y) with the value of x mapped on to the periodic lattice is cleared out and returned to the UNVISITED state. This is allowed because the walk snakes back a maximum distance of the order of its width in the y direction, so that sites or bonds behind that can be forgotten. Thus the simulation runs continuously, effectively simulating an infinitely wide system.

To start the walk, one has to make a vertical column on the left of all occupied sites or bonds above the starting point, and vacant ones below it. This will keep the walk from closing on itself at the beginning. Data from the early part of the simulation can be thrown away to eliminate any bias that it causes.

This method was used to find precise thresholds, some up to seven significant figures, for a variety of two-dimensional lattices, including the Archimedean lattices for site percolation [69], and the kagomé lattice for bond percolation [91]. These results are useful for understanding how thresholds depend upon the lattice structure, and to test conjectures for the values of the thresholds (see [62]).

The question of the convergence of the estimates is open: for many systems, the convergence behavior seems to change from $1/L$ for smaller L to a different behavior (or perhaps the same $1/L$ behavior but with a different coefficient) for larger L. In fact, for site percolation on the square lattice, it turns out that the linear behavior seen by Rosso et al. breaks down for L larger than about 1000, and the curve levels off, extrapolating to a value ≈ 0.5927465, close found by other methods. The understanding of the convergence of this method remains an open problem.

The errors can be determined easily by looking at batches of results, and are proportional to $(n_{\mathrm{occ}}+n_{\mathrm{vac}})^{-1/2}$. The proportionality constant is of order 1, indicating a very efficient method, and grows slowly with increasing L, implying that with increasing L, somewhat more work is needed to achieve the same level of precision.

Example: the Critical Surface for the Checkerboard Lattice

As an example of this method, we consider the checkerboard lattice, that is a square lattice with four different probabilities p_1, p_2, p_3, and p_4 around each colored square. According to a conjecture by Wu concerning the more general q-state Potts model, here specialized for $q = 1$, the critical surface satisfies the formula [80]

$$1 - (p_1p_2 + p_1p_3 + p_1p_4 + p_2p_3 + p_2p_4 + p_3p_4) + p_1p_2p_3 + p_1p_2p_4 + p_1p_3p_4 + p_2p_3p_4 = 0\,. \quad (23)$$

This result does not appear to follow directly from duality, in contrast to all other exact results known in percolation [70,79,95]. However, it reduces to the known exact results for the inhomogeneous honeycomb lattice (letting $p_4 = 0$), the inhomogeneous triangular lattice (letting $p_4 = 1$), and the dual checkerboard lattice ($p_2 = 1 - p_1$), and duality in the sense that $p_i \to 1 - p_i$ also satisfies this formula. It is the most general relation of this form, linear in all of the individual probabilites, that satisfies these requirements, but there is no obvious reason why it has to be linear in this way.

It seems that this result has not been tested numerically in the past. Here we investigate one case $p_1 = 73/90 = 0.811111\ldots$, and $p_2 = p_3 = p_4 = p$ using the hull-gradient method. According to the conjecture (23), p should equal 0.4 exactly.

In order to test this prediction, we fix the value of p_1 on every fourth bond, while for the rest of the bonds we allow p to follow the gradient, and use the n_{occ} and n_{vac} for these bonds to estimate the critical value of p by (22). Figure 6 shows the plot of $p_c^{\mathrm{est}}(L)$ that follows for different values of the inverse gradient L. The lattice was $16\,384 \times 16\,384$, but we were able to go to inverse gradients as large as $L = 524\,288$ without having $y_{\max} - y_{\min}$ or $x_{\mathrm{front}} - x$ exceed 16 384. We also used a periodic scheme in the vertical direction in order to have the walk automatically adjust to its own position. In this particular case, the linear dependence of p_c^{est} upon the gradient $1/L$ seems to hold, with an extrapolated value of 0.39999989(20). To achieve these very small error bars, a total of $n_{\mathrm{occ}} + n_{\mathrm{vac}} = 10^{14}$ random numbers were generated, one for each time the walk encountered an unvisited bond whose state was not previously determined. Thus, Wu's conjecture is numeri-

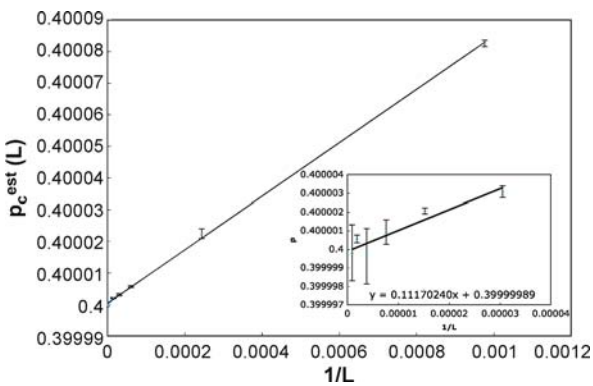

Percolation Lattices, Efficient Simulation of Large, Figure 6
Results of gradient percolation study for the checkerboard lattice with $p_1 = 73/90$, and $p_2 = p_3 = p_4$ predicted to be 0.4 by (23). From [63]

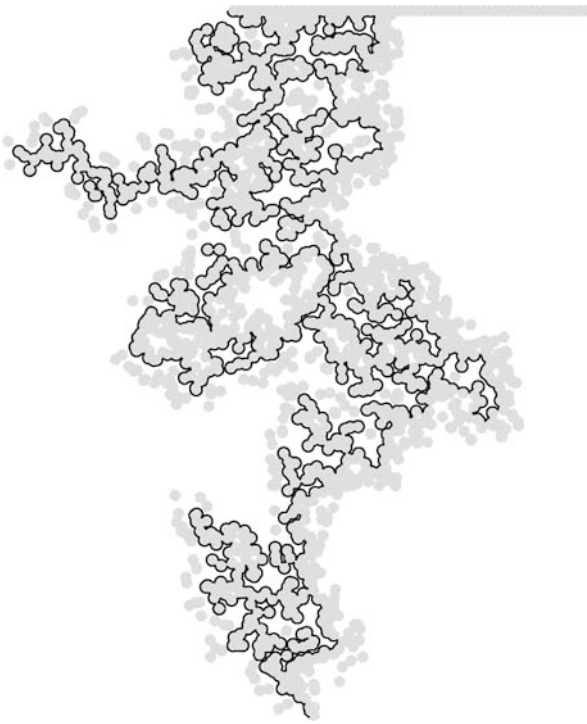

Percolation Lattices, Efficient Simulation of Large, Figure 7
Frontier (hull) of the percolating region for the continuum percolation of overlapping disks in a gradient in the horizontal direction. From [53]

cally confirmed to high accuracy for this point. Additional points are tested in [63].

The hull-gradient method has been generalized to continuum percolation, illustrated in Fig. 7, yielding the most precise known value for the fractional critical coverage $\phi_c = 0.6763475(6)$ [53,54].

Note, when doing work like this, it is imperative to use a high-quality random number generator, and not ones typically incorporated in computer languages or compilers. For much of our own work reviewed here, we used a four-tap shift-register sequence random number generator based upon the exclusive-or operation, with maximum lag of 9689 and cycle $2^{9689} - 1$ [87].

Simulating the Grossman–Aharony Accessible Hull

It does not seem possible to make a random walk process that generates the accessible hull of a percolation cluster (the hull in which all "fjords" are closed off) directly, because of the long range correlations. However, it is possible to generate this hull by carrying out two walks: the first to generate say the outside hull of a cluster, and then a second walk that encircles the first, and can jump across the fjords. In this way, samples of a walk that are equivalent to the two-dimensional self-avoiding walk can be generated easily.

In the same way, a second walk can be added to gradient percolation (delayed behind the first walk by the correlation length), and the second walk will trace out the accessible hull of the frontier in the gradient. In this, way, an infinitely long accessible hull can be made (essentially one dimensional, however, because of the effect of the gradient).

The Microcanonical-Canonical Method

Here we discuss the method of Newman and Ziff [44] to simulate percolation that, unlike other methods considered so far, allows one to simulate problems for all values of p through one simulation.

The idea is to start with an empty lattice, and add one site or bond at a time, and update the cluster connectivity on the fly, somewhat like the Hoshen–Kopelman method, but applied to clusters rather than rows. The quantity of interest is stored as a function of n, the number of added sites or bonds. Call this quantity Q_n, which represents the fixed-n, or "microcanonical" value of Q. Then, for a given probability, the "canonical" $Q(p)$ follows from convolving with the binomial distribution

$$Q(p) = \sum_{n=0}^{N} \binom{N}{n} p^n (1-p)^{N-n} Q_n \,, \tag{24}$$

where N is the total possible number of sites or bonds in the system. Note that these approaches have also been described as "canonical" and "grand-canonical", by considering s as representing the number of particles in the system [64]. Here, we are thinking in energetic terms, along the lines of the Potts model representation of percolation.

An example of this has already effectively been given above in the $p^n q^{N-n}$ series of $\Pi_h^{(L)}(1,p)$ (20) and (21), with $N = L^2$. Consider the case $L = 2$, and let $Q(p) = \Pi_h^{(2)}(1,p)$. The coefficients in (20) are precisely the number of ways of having horizontal crossing with n occupied sites and $N - n$ vacant sites. Because there are $\binom{N}{n}$ possibilities of having n sites in the system, it follows that the Q_n are just these coefficients divided by $\binom{N}{n}$. Therefore, for this system $Q_0 = Q_1 = 0$, $Q_2 = 1/3$, and $Q_3 = Q_4 = 1$ (the latter reflecting the fact that with three or four occupied sites, there will always be crossing). Then, the convolution in (24) is formally identical to $\Pi_h^{(2)}(1,p)$ given in (20).

The antecedents of this method in the literature are many. The idea of extrapolating results of simulations

to different values of p is reminiscent of the histogram method [31]. The representation of clusters as a tree structure and some of the update bookkeeping steps are reminiscent of the Hoshen–Kopelman method [29]. Similar tree structure representations of clusters have been used in kinetic gelation models [8]. Finally, the idea of adding one occupied site or bond at a time was suggested in a problem in [16]. But [44] seems to be the first place that all these ideas were put together along with the convolution (24) and used to find results for some quantity for all values of p. These ideas been incorporated in an extensive discussion of this method in are Gould, Tobochnik, and Christian [17].

For definiteness we consider bond percolation. Initially, no bonds exist, and all sites are clusters of size one and each has a different label. (Again, the size is the number of sites the cluster contains). When a new bond is added, it can either connect sites belonging to the same cluster, in which case nothing needs to be done, or it can connect sites from two different clusters. In the latter case, these two clusters are combined into one by a union operation. For efficiency, the smaller cluster is incorporated into the larger one.

A simple approach to carry out this operation is to label each site of the lattice by an index representing the cluster it belongs to, and having a look-up table that registers the number of sites in each cluster. When a new bond connects sites of two different indices, the look-up table tells which of the two is smaller, and then a neighbor search like (P2) or (P3) can be used to relabel all sites of the small

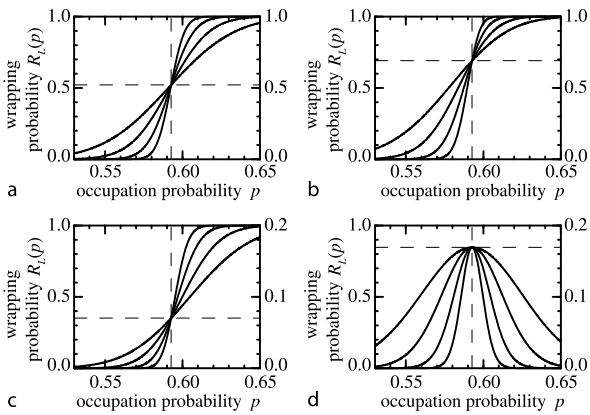

Percolation Lattices, Efficient Simulation of Large, Figure 8
Wrapping probabilities $R_L(p) \equiv \Pi_{\text{wrap}}(p_c)$ calculated using the Newman–Ziff algorithm, for $L \times L$ tori with $L = 32, 64, 128$, and 256, for wrapping (a) along a specified axis, (b) along either axis, (c) along both axes, and (d) along one axis but not the other. The dotted lines denote the expected values of p_c and $\Pi_{\text{wrap}}(p_c)$. The curves are sharper as L increases. From [44]

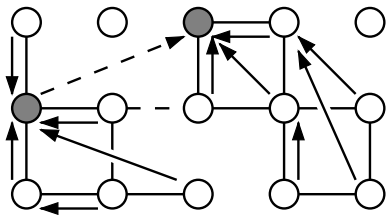

Percolation Lattices, Efficient Simulation of Large, Figure 9
Tree data structure, shown the merging off the cluster of six sites on the left with the cluster of seven sites on the right, due to the addition of the new bond (*dashed bond*). Arrows show the directions of the links. From [44]

cluster to the index of the larger cluster. The appropriate updates to the look-up table are then made.

However, this method is somewhat slow because a given site is relabeled several times, and it can be improved by having a cluster remembered as a tree structure and linking the root of the smaller cluster to that of the larger one, as shown in Fig. 9.

Here we describe a recursive program to carry out the "union-find" operation for bond percolation. More details are given in [45], which discusses the algorithm for site percolation.

First of all, in contrast to the programs given above, here we use a one-dimensional array `ptr[r]` to represent a system of any dimensions. An array `nn[i][dir]` is constructed ahead of time that tells the nearest-neighbors of every point `i` in the system, and thus can be set up for any boundary condition. This array is only used to decide which pair of sites a given bond connects. Sites are indexed with a single signed integer label for speed, taking values from 0 to $N - 1$.

The array `ptr[]` serves triple duty: for non-root occupied sites it contains the label for the site's parent in the tree (the "pointer"); root sites are recognized by a negative value of `ptr[]`, and that value is equal to minus the size of the cluster; for unoccupied sites `ptr[]` takes the value `EMPTY`, which is defined as some value such as $-N - 1$ that is never reached by any of the roots.

We define a function which performs the "find" operation, returning the label of the root site of a cluster, as well as accomplish path compression. The version we use is recursive:

```
int findroot(int i)
{
  if (ptr[i] < 0) return i;
  return ptr[i] = findroot(ptr[i]);
}
```

When the recursion is "unwound", all the sites of the links are relabeled to point to the new root. This seems to result in an optimal amount of relabeling to make this process run quickly.

The code to perform the actual algorithm is quite brief. Ahead of time, an ordered list of all the bonds is made. A given new bond connects the two neighboring sites `s1` and `s2`. The function `findroot()` is called to find the roots of each of the two sites. If amalgamation is needed, it is performed in a weighted fashion, smaller clusters being added to larger (bearing in mind that the value of `ptr[]` for the root nodes is *minus* the size of the corresponding cluster). Following is the main code to accomplish that:

```
r1 = findroot(s1);
r2 = findroot(s2);
if (r2!=r1)
  if (ptr[r1] > ptr[r2]) {
    ptr[r2] += ptr[r1];
    ptr[r1] = r2;
  } else {
    ptr[r1] += ptr[r2];
    ptr[r2] = r1;
  }
```

There are also easy techniques to check for crossing or wrapping during this process. One stores the number of such events as a function of n, the number of bonds put down, and then uses the convolution (24) to find the desired quantity as a function of p. Of course, the procedure must be repeated many times to get good statistics for all n. The result of a test of wrapping around a torus is given in Fig. 8.

The main point of this algorithm is that it can find the Q_n in a time that is very nearly linear in the number of lattice sites. Once the Q_n are found and stored in an appropriate array, they can be used to find various properties of the system for any p.

Other Numerical Techniques

In this section we briefly mention some other numerical techniques that have been applied to percolation.

The Binary Search Method

In this method, a random number p_i uniform in $(0, 1)$ is assigned to each site (say) of the lattice. One chooses a value of p such that sites with $p_i < p$ are assumed to be occupied, and checks for percolation. By making a binary search up and down in p (with the same p_i assigned to each site), in about 20 steps the probability where that particular sample first percolates can be found to six significant digits. Repeating this for many samples and averaging the results yields the *average* estimate for p_c.

If horizontal crossing is used for the criterion for percolation, then this average corresponds to $\int_0^\infty p(\partial/\partial p)\Pi_h(p, r)\mathrm{d}p$. When systems of different sizes are simulated, finite-size scaling of this quantity can be used to extrapolate the estimate to $L \to \infty$. In general, finite-size scaling implies that estimates converge to p_c as $L^{-1/\nu}$ [68], but for for certain symmetric systems, such as a square boundary for site percolation on a square lattice, the convergence goes as $L^{-1/\nu-1}$ [94] and even faster for wrapping around a periodic system [45].

This method is quite efficient in finding p_c since for a lattice of $N = L^2$ sites, the total amount of time to measure one sample grows only as $N \ln N$. It has been used in numerous studies of percolation in various dimensions [68].

Lattice-Less Methods

Vollmayr [75] introduced the idea that by using a kind of random number generator (effectively a very non-linear function) that takes as an input the coordinates of a site, and outputs a uniform random number that has no correlations with the random numbers that results from one of the neighbors, one does not have to remember the occupancy state of a given site. However, for most problems, one does have to remember whether that site has been visited (or checked as in the cluster algorithms), so techniques to remember the latter have to be used, some of which are discussed in [48]. There are several problems where one does not need to remember which sites have been visited, so for these problems this technique is particularly memory efficient. One example is the problem of finding the enclosed area of a hull-generating walk: here it is not necessary to remember if a site have been visited or not, and the method can be used to simulate any size system. Another example is just finding the end-to-end length of a very long walk, for a fractal measurement. One is only limited by the computational time available.

A drawback of this method is that it requires this special type of random number generator. Vollmayr uses a generator related to data encryption, which evidently has the necessary properties and is sufficiently uncorrelated and fast (though not as fast as typical random number generators used for Monte Carlo simulations). However, more work needs to be done to study this generator's quality for these types of problems.

Osterkamp, Stauffer and Aharony [46] introduced a related idea, using the feature of congruential random number generators that one can jump ahead or behind any

number of steps in the random number sequence by making an appropriate modification to the multiplier. Using this they were able to simulate diffusion on percolation clusters on high-dimensional (virtual) lattices as large as $420^7 \approx 2.3 \cdot 10^{18}$ sites. This is another problem for which there is no need to remember which site has been visited before, so no list is needed, and the program required very little memory.

Conductivity and Backbones

An important application of percolation is to flow and resistance problems. The conductivity of a percolating network (say with occupied bonds replace by identical resistors) goes to zero as the percolation threshold is approached from above, and much work has been done in studying that process. The simulations are based upon solving Kirchhoff's equations around each vertex of the lattice.

When considering conductivity of a percolation cluster, the role of different bonds becomes evident. One can define conductivity between two points far apart in a cluster, or alternatively between two opposite edges of a bounded system (the "bus-bar" problem). In either case there will be a backbone that carries the current, and dangling ends that are only singly connected to the backbone, which can be removed. Within the backbone, bonds can be classified into different categories, with the "red" bonds being ones the "hottest" in that cutting any one of them will break the flow of current [50].

To find the backbone, several methods have been used, including Tarjan's method [71], burning algorithms [27], and matching algorithms [43]. Grassberger has introduced an efficient hull-walk based algorithm which however works only in two dimensions [20]. Once the backbone is found, the conductivity can be estimated efficiently in 2d by the algorithm of Lobb and Frank [38], which reduces the lattice by successive use of a star-triangle transformation, or in general by finite-element methods to solve the Kirchhoff equations. Conductivity can also be studied by considering the properties of random walks on the percolation cluster [28].

In 2d at the critical point, the backbone has a fractal dimension $D_b = 1.6432(8)$ [21] or 1.6434(2) [9] such that in a system of length scale L, the number of backbone bonds grows as L^{D_b}. The red sites scale simply as the inverse of the correlation-length exponent, $D_{\mathrm{red}} = 1/\nu$ [7], which equals 3/4 in 2d. The conductivity at the critical point scales with the system size as L^t with $t/\nu = 0.9826(8)$ [21]. At one time there was a great deal of interest in studying this exponent because of the Alexander–Orbach conjecture [2] that t/ν was equal to 1, which however was proven (numerically) to be incorrect in five back-to-back papers in Physical Review B in 1984 [26,28,38,55,81].

Conclusions

This article describes a number of algorithms and programming techniques to study cluster statistics, crossing problems, area distributions, etc. of percolation. By no means did it cover all of them; having been a very active field of research for 50 years, there are many other methods and techniques that have been proposed and studied. Some applications of numerical techniques are also presented.

An example of the development of the substantial numerical work done in this field is provided by the determination of the threshold for site percolation on the square lattice, whose value has not been derived theoretically and must be found by simulation. After starting out at 0.581(15) in 1961 in the first Monte-Carlo determination [14], and after dozens of advances in the following three decades, by the early 1990s the six-digit value 0.592746 was achieved [83]. Yet, in the 16 years since then, while that value has been confirmed, the seventh digit still has not been agreed upon – the various determinations, many quoted above, fall in the range 0.5927460–0.5927466. Although one might think that with all the advances in computer power and algorithms that have occurred over these years, it would have been fairly easy to extend this result further, it turns out to be harder than might have been anticipated, because of uncertainties in the finite-size corrections and in the quality of random-number generators, which seem to be significant at this level of precision. Such high precision values are in fact necessary for precise studies of critical behavior, where simulations involving 10^{13}–10^{14} sites are not uncommon, and in 2d, site percolation on the simple square lattice remains one of the most popular models for various reasons, despite of the fact that exact thresholds are known for other $2 - d$ system.

Future Directions

Work in this field remains quite active and there are many interesting questions that are still unanswered. Convergence of many of the estimates, precise values of sub-leading (including irrelevant) exponents, more accurate calculation of the scaling functions and amplitude ratios are some such questions. For percolation thresholds, continued study of thresholds can perhaps lead to new exact results and in any case can help advance the understanding of why particular lattices have the thresholds that they

do. The combination of the efficient techniques that have been developed and improved upon over the years and the availability of powerful computers should allow many of these questions to be investigated fairly easily today. Here is a sampling of some specific questions for future study based upon the work discussed above:

- In the Newman–Ziff study of percolation around tori, it was found that the wrapping probabilities (for various situations, such as "either" of "one-way") approach their theoretical values [51] as L^{-2}, implying that the estimates for p_c converge to the actual value as $L^{-11/4}$. However, no theoretical justification was found for this result. Note that additional numerical work showing fair agreement with this scaling for a variety of latties was done recently by Parviainen [47].
- As mentioned above, the estimate of p_c for the hull-gradient method seems to converge as the reciprocal of the gradient for many systems (as in Fig. 6), while in others it changes its behavior and even is non-monotonic. Additional precise measurements, including tests with different random number generators, on a variety of systems can help elucidate this question. Another interesting question is the effect of the angle of the gradient with respect to the axis for various lattices. Some work along these lines for the kagomé lattice was done in [69].
- When written in terms of the enclosed area distribution, the size distribution follows the Zipf's-law form (15) above which is an entirely universal form with no metric factors. Only preliminary studies have been made on this quantity away from p_c, and a simple exponential scaling curve is possibly seen [88]. Clarification of this behavior (and in relation to the Kunz–Souillard [36] form of the percolation scaling function which predicts exponential behavior in the size, not the area) is needed. Note there are other measures of the area that can be used, such as that of the enclosing Grossman–Aharony hull, that would also be interesting to study.

Recent interest in percolation by mathematicians, following the developments in Stochastic Loewner Evolution, will undoubtedly lead to many more simulation studies and new algorithms in percolation, such as [15] which concerns geometry of clusters on the closely related Potts model. It is interesting to note that in the percolation case, SLE develops a theory for the continuum limit of precisely the hull walks which were first introduced as a computational technique in this field more than two decades ago [19,78,89].

Acknowledgments

This work was supported in part by the National Science Foundation under grant no. DMS-0553487. Comments by D. Stauffer and P. Kleban were highly appreciated.

Bibliography

Primary Literature

1. Aharony A Stauffer D (1997) Test of universal finite-size scaling in two-dimensional site percolation. J Phys A 30:L301–L306
2. Alexander S, Orbach R (1982) Density of states on fractals: fractons. J Phys Lett 43:L623
3. Ballesteros HG, Fernández LA, Martín-Mayor V, Muñoz Sudupe A, Parisi G, Ruiz-Lorenzo JJ (1999) Scaling corrections: site percolation and Ising model in three dimensions. J Phys A 32: 1–13
4. Berlyand L, Wehr J (1995) The probability distribution of the percolation threshold in a large system. J Phys A 28:7127–7133
5. Cardy JL (1992) Critical percolation in finite geometries. J Phys A 25:L201–206
6. Cardy J, Ziff RM (2003) Exact area distribution for critical percolation, Ising and Potts model clusters. J Stat Phys 110:1–33
7. Coniglio A (1982) Cluster structure near the percolation threshold. J Phys A 15:3829–3844
8. de Freitas JE, Lucena LS (2000) Equivalence between the FLR time-dependent percolation model and the Newman–Ziff algorithm. Int J Mod Phys C 8:1581–1584
9. Deng Y, Blöte HWJ, Nienhuis B (2004) Backbone exponents of the two-dimensional q-state Potts model: A Monte Carlo investigation. Phys Rev E 69:026114
10. Deng Y, Blöte HWJ (2005) Monte Carlo study of the site percolation model in two and three dimensions. Phys Rev E 72:016126
11. Erdős P, Rényi A (1959) On random graphs. Publ Math 6: 290–297
12. Fisher M, Essam JW (1961) Some cluster size and percolation problems. J Math Phys 2:609–619
13. Flory P (1941) Molecular Size Distribution in Three Dimensional Polymers I Gelation. J Am Chem Soc 63:3083–3090
14. Frisch HL, Sonnenblick E, Vyssotsky V, Hammersley JM (1961) Critical percolation probabilities (site problem) Phys Rev 124:1023–1022
15. Gamsa A, Cardy J (2007) SLE in the three-state Potts model – a numerical study. J Stat Mech P08020
16. Gould H, Tobochnik J (1996) An Introduction to Computer Simulation Methods, 2nd edn. Addison-Wesley, Reading, p 444
17. Gould H, Tobochnik J, Christian W (2006) An Introduction to Computer Simulation Methods, 3nd edn. Addison-Wesley, Reading
18. Grassberger P (1983) Critical behavior of the general epidemic process and dynamical percolation. Math Biosci 63:157–172
19. Grassberger P (1986) On the hull of two-dimensional percolation clusters. J Phys A 19:2675–2677
20. Grassberger P (1992) Spreading and backbone dimensions of 2D percolation. J Phys A 25:5475–5484
21. Grassberger P (1999) Conductivity exponent and backbone dimension in $2 - d$ percolation. Physica A 262:251–263
22. Grassberger P, Zhang YC (1996) Self-organized formulation of standard percolation phenomena. Physica A 224:169–179

23. Grossman T, Aharony A (1984) Structure and perimeters of percolation clusters. J Phys A 19:L745–L751
24. Gruzberg I (2006) Stochastic geometry of critical curves, Schramm-Loewner evolutions, and conformal field theory. J Phys A 39:12601–12656
25. Havlin S, Nossal R (1984) Topological properties of percolation clusters. J Phys A 17:L427–L432
26. Herrmann HJ, Derrida B, Vannimenus J (1984) Superconductivity exponents In: Herrmann HJ, Derrida B, Vannimenus J (eds) Two-and three-dimensional percolation. Phys Rev B 30: 4080–4082
27. Herrmann HJ, Hong DC, Stanley HE (1984) Backbone and elastic backbone of percolation clusters obtained by the new method of burning. J Phys A 17:L261–L266
28. Hong DC, Havlin S, Herrmann HJ, Stanley HE (1984) Breakdown of Alexander-Orbach conjecture for percolation: Exact enumeration of random walks on percolation backbones. Phys Rev B 30:4083–4086
29. Hoshen J Kopelman R (1976) Percolation and cluster distribution I Cluster multiple labeling technique and critical concentration algorithm. Phys Rev B 14:3438–3445
30. Hovi J-P, Aharony A (1996) Scaling and universality in the spanning probability for percolation. Phys Rev E 53:235–253
31. Hu CK (1992) Histogram Monte Carlo renormalization-group method for percolation problems. Phys Rev B 14:6592–6595
32. Hu CK, Chen J-A, Izmailian S Sh, Kleban P (2000) Recent developments in the Monte Carlo approach to percolation problems. Comp Phys Comm 126:77–81
33. Jensen I, Ziff RM (2006) Universal amplitude ratio Γ^-/Γ^+ for two-dimensional percolation. Phys Rev E 74:020101R
34. Kleban P, Ziff RM (1998) Exact results at the two-dimensional percolation point. Phys Rev B 57:R8075–R8078
35. Kleban P, Simmons JJH, Ziff RM (2006) Anchored critical percolation clusters and 2D electrostatics. Phys Rev Lett 97:115702
36. Kunz H, Souillard B (1978) Essential singularity in the percolation model. Phys Rev Lett 40:133–135
37. Leath P (1976) Cluster size and boundary distribution near percolation threshold. Phys Rev B 14:5046–5055
38. Lobb CJ, Frank DJ (1984) Percolation conduction and the Alexander-Orbach conjecture in two dimensions. Phys Rev B 30:4090–4092
39. Lorenz CD, Ziff RM (1998) Precise determination of the bond percolation thresholds and finite-size scaling corrections for the sc, fcc, and bcc lattices. Phys Rev E 57:230–236
40. Lorenz CD, May R, Ziff RM (2000) Similarity of percolation thresholds on the hcp and fcc lattices. J Stat Phys 98:961–970
41. Martín-Herrero J (2004) Hybrid cluster identification. J Phys A 37:9377–9386
42. Moore C, Newman MEJ (2000) Exact solution of site and bond percolation on small-world networks. Phys Rev E 62: 7059–7064
43. Moukarzel C (1998) A Fast Algorithm for Backbones. Int J Mod Phys C 9:887–895
44. Newman MEJ, Ziff RM (2000) Efficient Monte Carlo algorithm and high-precision results for percolation. Phys Rev Lett 85:4104–4107
45. Newman MEJ, Ziff RM (2001) A fast Monte Carlo algorithm for site or bond percolation. Phys Rev E 64:016706
46. Osterkamp D, Stauffer D, Aharony A (2003) Anomalous diffusion at percolation threshold in high dimensions of 10^{18} sites. Int J Mod Phys C 14:917–924
47. Parviainen R (2007) Estimates of the bond percolation thresholds on the Archimedean lattices. J Phys A 40:9253–9258
48. Paul G, Ziff RM, Stanley HE (2001) Percolation threshold, Fisher exponent, and shortest path exponent for four and five dimensions. Phys Rev E 64:026115
49. Perlsman E, Havlin S (2002) Method to estimate critical exponents using numerical studies. Europhys Lett 58:176–181
50. Pike R, Stanley HE (1981) Order propagation near the percolation threshold. J Phys A 14:L169
51. Pinson HT (1994) Critical percolation on the torus. J Stat Phys 75:1167–1177
52. Privman V, Hohenberg PC, Aharony A (1991) Universal Critical-Point Amplitude Relations. In: Domb C, Lebowitz JL (eds) Phase transition and critical phenomena, vol 14. Academic Press, New York
53. Quintanilla J, Torquato S, Ziff RM (2000) Efficient measurement of the percolation threshold for fully penetrable discs. J Phys A 33:L399–L407
54. Quintanilla J, Ziff RM (2007) Near symmetry of percolation thresholds of fully penetrable disks with two different radii. Phys Rev E 76:051115
55. Rammal R, Angles d'Auriac JC, Benoit A (1984) Universality of the spectral dimension of percolation clusters. Phys Rev B 30:4087–4089
56. Reynolds P, Stanley HE, Klein W (1980) Large-cell Monte Carlo renormalization group for percolation. Phys Rev B 21: 1223–1245
57. Rosso M, Gouyet JF, Sapoval B (1985) Determination of percolation probability from the use of a concentration gradient. Phys Rev B 32:6063–6054
58. Rozenfeld HD, ben-Avraham D (2007) Percolation in hierarchical scale-free nets. Phys Rev E 75:061102
59. Saleur H, Duplantier B (1987) Exact determination of the percolation hull exponent in two dimensions. Phys Rev Lett 58:2325–2328
60. Sapoval B, Rosso M, Gouyet J-F (1985) The fractal nature of a diffusion front and relation to percolation. J Phys Lett Paris 46:L149
61. Schramm O (1999) Scaling limits of loop-erased random walks and uniform spanning trees. Israel J Math 118:221–288
62. Scullard CR, Ziff RM (2006) Predictions of bond percolation thresholds for the kagomé and Archimedean (3, 12^2) lattices. Phys Rev E 73:045102(R)
63. Scullard CR, Ziff RM (2008) Critical surfaces of general bond percolation problems. Phys Rev Lett 100:185701
64. Shchur LN (2000) Incipient Spanning Clusters in Square and Cubic Percolation. In: Landau DP, Lewis SP, Schuettler HB (eds) Springer Proceedings in Physics, vol 85. Springer, Berlin
65. Simmons JJH, Kleban P, Ziff RM (2007) Exact factorization of correlation functions in 2-D critical percolation. 76:041106
66. Simmons JJH, Kleban P, Ziff RM (2007) Percolation crossing formulas and conformal field theory. J Phys A 40:F771–F784
67. Smirnov S, Werner W (2001) Critical exponents for two-dimensional percolation. Math Res Lett 8:729–744
68. Stauffer D, Aharony A (1994) An Introduction to Percolation Theory, revised 2nd edn. Taylor and Francis, London
69. Suding PN, Ziff RM (1999) Site percolation thresholds for Archimedean lattices. Phys Rev E 60:295–283
70. Sykes MF, Essam JW (1964) Exact critical percolation probabilities for site and bond problems in two dimensions. J Math Phys 5:1117–1127

71. Tarjan T (1972) Depth-first search and linear graph algorithms. SIAM J Comput 1:146–160
72. Temperley HNV, Lieb EH (1971) Relations between the 'percolation'and 'colouring' problem and other graph-theoretical problems associated with regular planar lattices: some exact results for the 'percolation' problem. Proc R Soc London A 322:251–280
73. Tiggemann D (2001) Simulation of percolation on massively-parallel computers. Int J Mod Phys C 12:871
74. Voigt CA, Ziff RM (1997) Epidemic analysis of the second-order transition in the Ziff-Gulari-Barshad surface-reaction model. Phys Rev E 56:R6241–R6244
75. Vollmayr H (1993) Cluster hull algorithms for large systems with small memory requirement. J Stat Phys 74:919–927
76. Voss RF (1984) The fractal dimension of percolation cluster hulls. J Phys A 17:L373–L377
77. Vyssotsky VA, Gordon SB, Frisch HL, Hammersley JM (1961) Critical Percolation Probabilities (Bond Problem). Phys Rev 123:1566–1567
78. Weinrib A, Trugman SA (1985) A new kinetic walk and percolation perimeters. Phys Rev B 31:2993–2997
79. Wierman JC(1984) A bond percolation critical probability determination based on the star-triangle transformation. J Phys A 17:1525–1530
80. Wu FY (1979) Critical point of planar Potts models. J Phys C 12:L645–L650
81. Zabolitzky JG (1984) Monte Carlo evidence against the Alexander-Orbach conjecture for percolation conductivity. Phys Rev B 30:4077–4079
82. Ziff RM (1986) Test of scaling exponents for percolation cluster perimeters. Phys Rev Lett 56:545–548
83. Ziff RM (1992) Spanning probability in 2D percolation. Phys Rev Lett 69:2670–2674
84. Ziff RM (1994) Reply to Comment on Spanning probability in 2D percolation. Phys Rev Lett 72:1942
85. Ziff RM (1995) Proof of crossing formula for 2D percolation. J Phys A 28:6479–6480
86. Ziff RM (1996) Effective boundary extrapolation length to account for finite-size effects in the percolation crossing function. Phys Rev E 54:2547–2554
87. Ziff RM (1998) Four-tap shift-register-sequence random-number generators. Comput Phys 12:385–392
88. Ziff RM (2004) Enclosed area distribution in percolation. Talk presented at StatPhys22. arXiv:cond-mat/0510633
89. Ziff RM, Cummings PT, Stell G (1984) Generation of percolation cluster perimeters by a random walk. J Phys A 17:3009–3017
90. Ziff RM, Sapoval B (1986) The efficient determination of the percolation threshold by a frontier-generating walk in a gradient. J Phys A 19:L1169–1172
91. Ziff RM, Suding PN (1997) Determination of the bond percolation threshold for the kagomé lattice. J Phys A 30:5351–5359
92. Ziff RM, Finch SR, Adamchik VS (1997) Universality of finite-size corrections to the number of critical percolation clusters. Phys Rev Lett 79:3447–3450
93. Ziff RM, Lorenz CD, Kleban P (1999) Shape-dependent universality in percolation. Physica A 266:17–26
94. Ziff RM, Newman MEJ (2002) Convergence of threshold estimates for two-dimensional percolation. Phys Rev E 66:016129
95. Ziff RM, Scullard CR (2006) Exact bond percolation thresholds in two dimensions. J Phys A 39:15083–15090

Books and Reviews

Bollobás B, Riordan O (2006) Percolation. Cambridge University Press, Cambridge

Percolation Phase Transition

Muhammad Sahimi
Mork Family Department of Chemical Engineering and Materials Science, University of Southern California, Los Angeles, USA

Article Outline

Glossary
Definition of the Subject
Introduction
What is Percolation?
The Percolation Phase Transition
Percolation Properties
Universal Scaling Properties of Percolation
Variants of Percolation Processes and Their Applications
Future Directions
Bibliography

Glossary

Accessible bonds or sites The retained bonds or sites in a percolation lattice that are connected to infinity via at least one path.

Backbone The set of (retained) bonds or sites in the sample-spanning percolation cluster that are connected to infinity by more than one independent path.

Bond percolation In a bond percolation model, a random lattice is formed from an infinite lattice by retaining each bond of the infinite lattice with probability p, and deleting the rest. When a bond is retained, so also are its two ends sites.

Cluster A connected set of bonds or sites that are retained in a percolation lattice with probability p.

Correlation length The length scale below which a disordered percolation system cannot be regarded as homogeneous.

Critical exponents At the percolation thresholds, many percolation properties follow universal power laws, and the exponents that characterize such power laws are called *critical exponents*.

Percolation transition The connectivity or geometrical transition between a system in which a sample-spanning cluster of retained sites (or bonds) exists, and one in which no such cluster exists.

Site percolation In a site percolation problem, a random lattice is formed from an infinite lattice by retaining each site of the infinite lattice with probability p and deleting the rest. A bond connecting two retained neighboring sites is also retained.

Definition of the Subject

This chapter describes the basic percolation problem. We begin by noting that most systems of scientific and practical applications are, at least at some scale, disordered. In such disordered media, then, the connectivity of the elementary or microscopic elements has a profound effect on the media's macroscopic properties. The percolation transition occurs at the percolation threshold, which is the point at which the microscopic elements become connected for the first time, and form a sample-spanning path across the system. Percolation theory aims to describe the effect of the connectivity of the microscopic elements on the effective macroscopic properties of disordered media, particularly in the vicinity of the percolation threshold.

Introduction

It is well known that *Nature is disordered*. Pure, perfectly characterized, and geometrically immaculate systems are nowhere to be found, except perhaps in books and papers on theoretical physics. Although the concept of an infinite, perfectly periodic crystal lattice is incredibly elegant, it is as remote from experimental reality as possible. Even the best experimentalist who focuses on the purest of substances, exemplified by carefully grown crystals, can hardly ever escape the effects of defects, trace impurities, and finite boundaries. Thus, we must come to terms with *disordered morphology*: variations in the shape and constitution that are often so ill-characterized that we must deem them to be random if we are to describe them, or have any hope of doing so. The morphology of a medium has two major aspects: the *topology*, the interconnectiveness of the system's individual microscopic elements, and the geometry – the shapes and sizes of the individual elements.

At the same time, we believe, at least above the quantum mechanical level, in the doctrine of determinism, yet important *continua* exist in which deterministic descriptions of many phenomena are beyond hope. A well-known example is diffusion where, at least over certain length scales, one observes an apparent random process – or *disordered dynamics*. The two types of disorder – morphological and dynamical – are often coupled and present simultaneously. An important example is fluid flow through a porous medium, where the interplay between the disordered morphology of the pore space and the dynamics of fluid motion gives rise to a rich variety of phenomena [6,8].

Research on understanding the macroscopic properties of materials did make remarkable progress by using statistical mechanics and taking advantage of periodic structures, or through the application of such equations as the Boltzmann's equation. However, due to the rather obvious randomness in Nature, and because, in the final analysis, one always must confront the real world which is disordered, it became apparent in the 1960s that a statistical physics of disordered media must be developed to provide methods for deriving macroscopic properties of such media from the laws governing the microscopic world or, alternatively, for deducing their microscopic properties from the macroscopic observations and experimental measurements. Such a statistical physics of disordered media must take into account the effect of *both* the topology and geometry of the media. Although the role of the geometry was appreciated as early as the beginning of the twentieth century, the effect of the topology was ignored for many decades, or was treated in an ad hoc manner, simply because it was thought to be too difficult to be taken into account rigorously.

As the history of science indicates, progress in any research field is not usually made with a constant rate, but rather in a sporadic manner. There are periods when a problem looks so difficult that we do not even know where or how to start attacking it, and periods when some seminal discoveries remove a great obstacle to progress and, thus, enable us to make a great leap forward. An excellent example is the discovery of a new class of superconducting materials by Bednorz and Müller [1], who showed that it is possible to have superconductivity in certain Cu alloys at temperatures $T > 30\,\mathrm{K}$. Since their discovery (which brought them the physics Nobel Prize in 1989), the field of superconductivity has advanced remarkably (after not making much progress for decades), so much so that we now have materials that can be superconducting at temperature well above 100 K.

Over the past three decades, the statistical physics of disordered media has been in a rapidly progressing phase, the reason for which is fourfold:

(i) Rigorous theoretical methods for calculating the macroscopic (average) properties of disordered media have been developed.
(ii) A large amount of accurate experimental data have been accumulated, thanks to many novel experimental techniques and instruments.
(iii) Advances in computer technology and computational strategies have enabled us to use precise nu-

merical simulations for obtaining accurate estimates of many properties of disordered materials.

(iv) The fifth, and perhaps the most important, reason for the rapid development of the statistical physics of disordered media is that, the effect of the interconnectivity of the microscopic elements of disordered media on their macroscopic properties has been understood and appreciated. This has become possible through the development and application of percolation theory, the subject of this section of the Encyclopedia.

What is Percolation?

Consider electrical (or thermal) conduction through a composite material which is a mixture of conducting and insulating constituents or phases. Assume also that the two phases are randomly distributed. As an idealization, we represent the composite material by a simple-cubic network in which each bond is either conducting with a finite conductivity, or insulating with zero conductivity. Suppose also that we impose a voltage difference between two opposite faces of the network. The question that we ask is: what fraction of the bonds must have a finite conductivity in order for the electrical current to flow through the material, so that it has a nonzero macroscopic conductivity? This is clearly an important question, because its answer tells us, for example, that what (volume) fraction of a composite material, such as carbon black composites that are routinely used in many applications, must be conducting in order for the composite as a whole to be conducting.

Consider a second example. Imagine that the bonds of a simple-cubic network represent the pore throats of a porous medium, e. g., an oil reservoir in porous rock. The pore throats are the narrow passages that connect the pore bodies. Most of the porosity of a porous medium (the void volume fraction in the porous medium) resides in the pore bodies. For brevity, we refer to the pore throats as pores. In reality, no porous medium looks as ordered as a simple-cubic network, but as an idealization the model is useful. Now, suppose that the pores (bonds) are filled with oil, and that there are two wells in the system, one at A on one face, and a second one at B on the opposite face of the network. We try to push the oil out of the network (porous medium) by injecting water into the system at A – the injection well – to produce oil at B – the production well. Oil and water do not mix with each other and, therefore, we assume that each pore is filled with either oil or water. We also assume that water wets the surface of the pores (the wetting fluid), whereas oil does not (the nonwetting fluid). In many oil reservoirs, such as carbonate oil reservoirs of the Middle East, the opposite is true, but this does not make any difference to our discussion.

When the water is injected and pushed into the reservoir, it tries, due to being the wetting fluid, to find the *smallest pores* that it can reach and expel the oil from it. In reality, the process is more complex than what we are describing, but we ignore all the complications. The displaced oil is produced at well B. The question that we ask is: what fraction of the pores are filled with water when it reaches the production well at B *for the first time* (this is called the *breakthrough point*)? In other words, we would like to know what fraction of the pores lose their oil and, thus, how much oil is produced at well B at the breakthrough point. This is clearly an important question, given that the price of oil is now around \$75/barrel.

The Percolation Phase Transition

In the example of composite materials described above, if too many bonds (or too much of the materials) are insulating, no macroscopic current will flow through the materials, whereas for sufficiently large number of conducting bonds electrical current does flow in the materials, so that their macroscopic effective conductivity is nonzero. Assume that the fraction of the conducting bonds is p. Therefore, there must be a minimum or critical value p_{cb} of p, such that for $p \leq p_{cb}$ no electrical current would flow through the material and, therefore, the materials as a whole are insulating, whereas for $p > p_{cb}$ the materials become conducting.

In the example of displacement of oil by water, p represents the fraction of pores from which oil has been expelled and replaced by water. Therefore, for $p \leq p_{cb}$, water flows only locally, and has not reached the production wall, whereas for $p > p_{cb}$ water flows between the injection and production wells. Thus, at any given time, p represents the fraction of the total oil in the reservoir that has been recovered, while p_{cb} represents its value at the breakthrough point.

Therefore, it should be clear that, in both examples, p_c signifies a phase transition: for $p \leq p_{cb}$ there is no sample-spanning path of conducting bonds, or pores filled by water, so that the system is macroscopically disconnected or *closed* (to electrical current or flow of water). But, for $p > p_{cb}$ the system becomes macroscopically connected. Hence, p_{cb} is the point at which a *geometrical* phase transition from a disconnected to a connected system takes place. Percolation theory, then, quantities the effect of the interconnectivity of the microscopic elements of a disordered medium (the conducting elements, or the pores

filled with water) on its macroscopic properties. p_{cb} is called the *bond percolation threshold* of the network.

We may also formulate the percolation problem in another way. Recall that most of the porosity of a porous medium resides in its pore bodies which connect the pore throats, as several pore throats meet at a pore body. Thus, in the pore network model of a porous medium described above, the nodes or sites of the network, which connect the pore throats or bonds, are the equivalent of the pore bodies. Then, in the example of the displacement of oil by water in a porous medium, the injected water pushes the oil from the sites (into the bonds) toward the production well. Since most of the porosity (the void space available for the fluids) resides in the pore bodies, to obtain a more accurate estimate of the volume of the oil recovered, we ask the question: At the breakthrough point, what fraction of the network's sites (pore bodies) are filled with water? Denoting this fraction by p_{cs}, it should be clear that it is the analogue of p_{cb}. p_{cs} is called the *site percolation threshold* of the network. For all two- and three-dimensional (3D) lattices, $p_{cs} > p_{cb}$.

Determination of exact percolation thresholds of many 2D and all the 3D lattices remains an unsolved problem. The article by Wierman describes and discusses the existing exact results for the percolation thresholds. Ziff's article describes highly efficient numerical methods for estimating the percolation threshold and many other properties of percolation lattices.

Percolation Properties

Some of the most important properties of percolation systems that describe their morphology are as follows. For simplicity, we use p_c to denote p_{cs} or p_{cb}.

1. The *percolation probability* $P_\infty(p)$ is the probability that, when the fraction of occupied bonds is p, a given site belongs to the sample-spanning (infinite) cluster of occupied bonds.
2. The *accessible fraction* $A(p)$ is that fraction of occupied bonds (or sites) that belong to the infinite cluster.
3. The *backbone fraction* $B(p)$ is the fraction of occupied bonds in the infinite cluster which actually participate in a transport process, such as conduction, since some of the bonds in the infinite cluster are dead-end and do not carry any current. Therefore, $A(p) \geq B(p)$.
4. The *correlation length* $\xi(p)$ is the typical radius of percolation clusters for $p < p_c$, and the typical radius of the "holes" above p_c that are generated by the vacant bonds or sites. For $p > p_c$, ξ is the length scale over which the system is macroscopically homogeneous.
5. The *average number of clusters of size s* (per lattice site) $n_s(p)$ is an important quantity in many of the problems of interest here because it corresponds to, for example, the number of conducting or insulating islands of a given size in a conductor-insulator composite solid.
6. The probability that two sites, one at the origin and another one at a distance $\mathbf{r}$, are both occupied and belong to the same cluster of occupied sites, is $p^2 P_2(\mathbf{r})$, where $P_2(\mathbf{r})$ is called the *pair-connectedness function*.
7. The *mean cluster size* S (also called the *site-averaged* cluster number) is the average number of sites in the cluster that contains a randomly-selected site, and is given by,

$$S = \frac{\sum_s s^2 n_s(p)}{\sum_s s n_s(p)} \,. \tag{1}$$

Essam [4] showed that S and the pair-connectedness function $P_2(\mathbf{r})$ are related through a simple relation:

$$S = 1 + p \sum_{\mathbf{r}} P_2(\mathbf{r}) \,. \tag{2}$$

8. Because a major application of percolation theory has been modeling of transport in disordered materials, and in particular composite solids, we must also consider the effective transport properties of percolation systems, namely, their conductivity, diffusivity, elastic moduli, and dielectric constant. We first consider the conductivity of a two-phase composite material modeled as a two-component network in which each (randomly-selected) bond has a conductance g_1 with probability p or g_2 with probability $q = 1 - p$. It is straightforward to show that the effective electrical (or thermal) conductivity σ_{eff} of the network is a *homogeneous function* and takes on the following form,

$$\sigma_{\text{eff}}(p, g_1, g_2) = g_1 F(p, h) \,, \tag{3}$$

where $h = g_2/g_1$. Due to the assumption of randomness of the material's morphology, σ_{eff} is invariant under the interchange of g_1 and g_2 (phase-inversion symmetry) and we must, therefore, have

$$\begin{aligned} \sigma_{\text{eff}}(p, g_1, g_2) &= \sigma_{\text{eff}}(q, g_2, g_1) \,, \quad \text{and} \\ F(p, h) &= h F(q, 1/h) \,. \end{aligned} \tag{4}$$

The limit in which $g_2 = 0$ and g_1 is finite corresponds to a conductor-insulator mixture, already described above. In this case, as $p \to p_c$, more and more bonds are insulating, the conduction paths become very tortuous and, therefore, σ_{eff} decreases; at p_c one

has $\sigma_{\text{eff}}(p_c) = 0$, since no sample-spanning conduction path exists any more. More generally, the conductance g_1 may follow a certain statistical distribution, which is in fact the case in most systems of practical importance, such as porous materials and composite solids.

The limit in which $g_1 = \infty$ and g_2 is finite a represents a *conductor-superconductor* mixture. All quantum-mechanical aspects of real superconductors are ignored in this definition, and we are concerned only with the effect of the local connectivity of the material on this conductivity. It is clear that the effective conductivity σ_{eff} of this system is dominated by the superconducting bonds. If $p < p_c$, then a sample-spanning cluster of the superconducting bonds does not exist, and σ_{eff} is finite. As $p \to p_c^-$, σ_{eff} increases until a sample-spanning cluster of the superconducting bonds is formed for the first time at $p = p_c$, where σ_{eff} *diverges*. Note that both limits (g_1 finite and $g_2 = 0$, and $g_1 = \infty$ and g_2 finite) correspond to $h = 0$. Therefore, the point $h = 0$ at $p = p_c$ is particularly important. The article by Hughes elaborates on these aspects, and provides a full account of the state-of-the-art of this problem.

9. In a similar manner, the elastic moduli of a two-phase composite solid, modeled by a percolation network, are defined. Consider a two-component network in which each bond is an elastic element (a spring or beam) which has an elastic constant e_1 with probability p or e_2 with probability $q = 1 - p$. The limit in which $e_2 = 0$ and e_1 is finite corresponds to composites of rigid materials and holes (for example, porous solids). In such networks, as $p \to p_c$, more bonds have no rigidity, the paths for transmission of stress or elastic forces become very tortuous and, therefore, the effective elastic moduli E (Young's, bulk, or shear moduli) decrease; at p_c one has $E(p_c) = 0$. In general, the elastic constant e_1 can be selected from a statistical distribution.

   The limit in which $e_1 = \infty$ and e_2 is finite represents mixtures of rigid-superrigid materials. In this case the effective elastic moduli E of the system are dominated by the superrigid bonds. If $p < p_c$, then a sample-spanning cluster of the superrigid bonds cannot form, and E is finite. As $p \to p_c^-$, the effective elastic moduli increase until the percolation threshold p_c of the rigid phase is reached at which a sample-spanning cluster of the superrigid bonds is formed for the first time, and the effective elastic moduli *diverge*. The article by Duxbury provides a comprehensive discussion of this subject.

10. The effective dielectric constant ε of a two-phase insulating composite material, modeled by a percolation network, may also be defined and, in fact, ε is closely related to the conductor-superconductor model described above (see, for example, [9]).

11. Finally, the effective diffusivity D of a porous material can also be defined in a similar manner; see the article ► Conduction and Diffusion in Percolating Systems by Hughes.

Universal Scaling Properties of Percolation

One of the most important characteristics of percolation systems is their *universal* properties. The behavior of many percolation quantities near p_c is insensitive to the microstructure (for example, the coordination number) of the network, and to whether the percolation process is a site or a bond problem. The quantitative statement of this universality is that many percolation properties follow power laws near p_c, and the *critical exponents* that characterize such power laws are universal and depend only on the Euclidean dimensionality d of the system. We first describe the universal properties of the quantities that characterize the morphology of percolation systems, and then present and discuss those of transport properties.

In general, the following power laws hold near p_c,

$$P_\infty(p) \sim (p - p_c)^\beta \,, \tag{5}$$

$$A(p) \sim (p - p_c)^\beta \,, \tag{6}$$

$$B(p) \sim (p - p_c)^{\beta_b} \,, \tag{7}$$

$$\xi(p) \sim |p - p_c|^{-\nu} \,, \tag{8}$$

$$S(p) \sim |p - p_c|^{-\gamma} \,, \tag{9}$$

$$P_2(\mathbf{r}) \sim \begin{cases} r^{2-d-\eta} \,, & p = p_c \,, \\ \exp(-r/\xi) \,, & \text{otherwise} \,, \end{cases} \tag{10}$$

where $r = |\mathbf{r}|$. For large clusters near p_c, the cluster size distribution $n_s(p)$ is described by the following scaling law,

$$n_s \sim s^{-\tau} f[(p - p_c)s^\sigma] \,, \tag{11}$$

where τ and σ are two more universal critical exponents, and $f(x)$ is a scaling function such that $f(0)$ is not singular. The article by Stauffer elaborates further on these.

Similar power laws are also followed by the transport properties of percolation composites. In particular,

$$\sigma_{\text{eff}}(p) \sim (p - p_c)^t \,, \quad \text{conductor-insulator composites} \tag{12}$$

$$\sigma_{\text{eff}}(p) \sim (p_c - p)^{-s} \,, \quad \text{conductor-superconductor composites} \tag{13}$$

$$E(p) \sim (p - p_c)^{T} \,, \quad \text{rigid-soft composites} \tag{14}$$

$$E(p) \sim (p_c - p)^{-S} \,, \quad \text{rigid-superrigid composites} \,. \tag{15}$$

For length scales $L < \xi$, the resistance R between two end points of a box of linear size L scales with L as $R \sim L^{\tilde{\zeta}}$. It is not difficult to show that

$$t = (d-2)\nu + \zeta \,, \tag{16}$$

where, $\zeta = \tilde{\zeta}\nu$. It has been shown [12] that in 2D, $t = s$.

The power law that characterizes the behavior of the effective diffusivity $D(p)$ near p_c is derived from that of $\sigma_{\text{eff}}(p)$, and is shown to be given by

$$D(p) \sim (p - p_c)^{t-\beta} \,. \tag{17}$$

The implied prefactors in all the above power laws depend on the type of lattice and are *not* universal.

Equations (12) and (13) can be unified by using the two-component resistor network described above. In the critical region, i. e., the region near p_c, where both $|p - p_c|$ and $h = g_2/g_1$ are small, the effective conductivity σ_{eff} follows the following scaling law [3,11]

$$\sigma_{\text{eff}} \sim g_1 |p - p_c|^t \Phi_{\pm}\left(h|p - p_c|^{-t-s}\right) \,. \tag{18}$$

where Φ_+ and Φ_- are two homogeneous functions corresponding, respectively, to the regions above and below p_c, and are, similar to t and s, universal. For any fixed and non-zero h, σ_{eff} has a smooth dependence on $p - p_c$. This becomes clearer if we rewrite Eq. (18) as

$$\sigma_{\text{eff}} \sim g_1 h^{t/(t+s)} \Psi\left[|p - p_c| h^{-1/(t+s)}\right] \,, \tag{19}$$

where $\Psi(x) = x^{\mu}\Phi_+(x^{-t-s}) = (-x)^t \Phi_-[(-x)^{-t-s}]$. Since the function $\Psi(x)$ is universal, the implication of Eq. (19) is that, if one plots $\sigma_{\text{eff}}/[g_1 h^{t/(t+s)}]$ versus $|p - p_c| h^{-1/(t+s)}$ for all networks (or randomly-disordered materials) that have the same Euclidean dimensionality, all the results (or measurements) should collapse onto a single universal curve. This provides a powerful tool for estimating the conductivity of a composite for any value of h, given the conductivities for two other values of h (by which the universal curve is constructed). Somewhat similar, but more complex, scaling equations can be developed for the elastic moduli, dielectric constant and other properties of percolation composites.

No exact relation is known between the transport and morphological exponents. This is, perhaps, because the transport exponents describe *dynamical* properties of disordered materials and media, whereas the morphological exponents characterize their *static* properties. In general, there is no reason to expect a direct relation between the two.

If two physical phenomena in heterogeneous media that contain percolation-type disorder are described by two different sets of critical exponents, then the physical laws governing the two phenomena must be fundamentally different. Thus, critical exponents help one to distinguish between different classes of problems and the physical laws that govern them. Moreover, since the numerical values of the percolation properties are not universal and vary from one system to another, but the scaling and power laws that they follow near p_c are universal and do not depend on the details of the system, estimates of the critical exponents for a certain phenomenon are used for establishing the relevance of a particular percolation model to that phenomenon in disordered materials.

Variants of Percolation Processes and Their Applications

In this section of the Encyclopedia, the theoretical aspects of percolation theory are described first and, then, some well-established applications are described.

It should be clear that a percolation network is created when sites or bonds are blocked or removed and, therefore, the macroscopic connectivity of the system is gradually lost. In the example of the composite materials, the bonds or sites are blocked to the conducting phase. In the example of displacement of oil by water in a porous medium, the bonds or sites are blocked to oil (since it is expelled from such sites or bonds). Therefore, percolation networks are also useful as simple models of any disordered medium in which the connectivity of the medium's microscopic elements influences its macroscopic properties. Moreover, as the articles by Stauffer and Ziff make clear, the main concepts of percolation theory are simple and, therefore, writing computer program for simulating a percolation process is conceptually straightforward and simple, if we do not wish to simulate very large networks. Thus, percolation networks may also serve as a simple tool for introducing students to computer simulation of disordered media. Stauffer and Aharony [10], Bunde and Havlin [2], and Hughes [5] emphasized the theoretical foundations of percolation theory, while Sahimi [7,8,9] described its important applications.

Although percolation in regular lattices – those in which the coordination number Z (the number of bonds connected to the same size) is the same everywhere – has been extensively invoked for studying the morphology and transport properties of many disordered materials, percolation in continua and in topologically-random networks – those in which the coordination number varies from site to site – are also of great interest, since in many practical situations one may have to deal with such irregular and continuous systems. For example, continuum percolation is directly applicable to characterization and modeling of the morphology and effective transport properties of microemulsions, polymer blends, sintered materials, sol-gel transitions, and many more. The article ► Continuum Percolation by Balberg describes the advances that have been made in understanding of the percolation effects in continuous systems, and in random networks.

In the percolation phenomena described so far, no correlations between various segments of the system (for example, bonds and/or sites, or their transport properties) were assumed. However, disorder in many important heterogeneous materials is not completely random. There usually are correlations of some extent that may be finite but large. For example, in packing of solid particles, there are short-range correlations. The universal scaling properties of percolation systems with finite-range correlations are the same as those of random percolation, if the length scale of interest is larger than the correlation length. Moreover, if the correlation function $C(r)$ decays as r^{-d} or faster, where d is the Euclidean dimensionality of the system, then the scaling properties of the system are identical with those of random percolation. This is not totally unexpected because even in random percolation, as p decreases toward p_c, correlations begin to build up and, therefore, the introduction of any type of correlations with a range shorter than the percolation correlation length ξ cannot change its scaling properties. In many other cases, e. g., in some disordered elastic materials, there are long-range correlations. The article by Coniglio describes the major differences between percolation in random and correlated systems.

A particular type of percolation model with extended correlations is known as the *bootstrap percolation*. In this problem sites of a lattice are initially randomly occupied. Then, those sites that do not have at least Z_c nearest-neighbor occupied sites are removed (note that $Z_c = 0$ is the usual random percolation). The interactions between the sites are short-ranged, but the correlations between them may build up as the distance between two occupied sites also increases. The original motivation for developing this model was to explain the behavior of some disordered materials in which magnetic impurities are randomly distributed in a host of non-magnetic metals. It is believed that in some of such materials an impurity atom cannot sustain a localized magnetic moment unless it is surrounded by a minimum number of magnetic neighbors. Bootstrap percolation has proven to be a complex problem with a rich variety of unusual properties that are a strong function of the parameter Z_c. For example, an important question is the nature of the percolation transition in this model. It now appears that for sufficiently high values of $Z_c \leq Z$ (where Z is the coordination number of the lattice), the percolation transition is *first-order*, i. e., discontinuous, whereas for low values of Z_c the transition is continuous and second-order. If the phase transition is first-order, then the percolation threshold of the system is, in fact, $p_c = 1$, the sample-spanning cluster is compact, and power laws (5)–(15) are no longer valid. The article ► Bootstrap Percolation by De Gregorio et al. describes this important area of percolation problems.

Over the past three decades percolation theory has been applied to modeling of a wide variety of phenomena in disordered media and systems. It is impossible to describe and discuss all such applications. In this section of the Encyclopedia, several well-established and well-understood applications are described and discussed.

The article ► Invasion Percolation by Knackstedt and Paterson describes in detail application of the percolation model to two-phase fluid flow in porous media, a simple example of which was already described above. Since, as described above, one fluid is injected into a porous medium – that is, it invades the medium – in order to displace the second fluid, this particular model is usually known as the *invasion percolation*. Other aspects of the application of percolation in problems on fluid flow through porous media are described in the article ► Percolation in Porous Media by King and Masihi.

It appears that percolation provides a powerful tool for modeling of the effect of the connectivity of fractures and faults on fluid flow and transport properties of rock, a highly complex set of phenomena. Thus, in their article ► Percolation, and Faults and Fractures in Rock, Adler et al. describe the recent application of percolation to this important problem.

The article ► Networks, Flexibility and Mobility in by Thorpe describes recent advances on generalization of the percolation model, and its application to modeling of network glasses and proteins. The question of the rigidity of such materials is addressed. Thrope's article is related to Duxbury's review ► Elastic Percolation Networks which describes elastic properties of percolation networks and their applications.

An important and well-established application of percolation is to modeling of the rheology of polymers and gels, particularly in the vicinity of the gelation point. Several variants of the percolation models have been developed in order to address this important problem. The article by Sahimi describes the advances that have been made in this area.

Finally, the article ► Percolation in Complex Networks by Cohen and Havlin describes application of percolation to problems in complex networks, particularly scale-free networks. They show how the concepts of percolation can be used to study not only the robustness and vulnerability of random networks, but also such problems as immunization and epidemic spreading in populations and computer networks, communication paths, and fragmentation in social networks.

Future Directions

Theoretically, most aspects of percolation are well-understood. Exact values of most of the critical exponents (in 2D), or their very accurate numerical estimates (in 3D), are known. Exact values of the bond and site percolation thresholds of several 2D lattices are also known, as are very accurate numerical estimates of the percolation thresholds of many 3D lattices, although we still do not know the exact value of, for example, the site percolation threshold of the square lattice. Thus, theoretically, the grand challenge is to develop general methods for obtaining the exact percolation thresholds of 3D lattices, and the exact values of the critical exponents for 3D systems, although the latter challenge may well be beyond reach.

Therefore, aside from the theoretical challenges described above, most of the future work on percolation will be concentrated on its applications to problems of practical importance, examples of which are described in this section of the Encyclopedia.

Bibliography

1. Bednorz JG, Müller KA (1986) Possible high T_c superconductivity in the Ba-La-Cu-O system. Z Phys B 64:189
2. Bunde A, Havlin S (eds) (1996) Fractals and Disordered Systems, 2nd edn. Springer, Berlin
3. Efros AL, Shklovskii BI (1976) Critical behavior of conductivity and dielectric constant near the metal-non-metal transition point. Phys Status Solidi B 46:475
4. Essam JW (1972) Percolation and cluster size. In: Domb C, Green MS (eds) Phase Transitions and Critical Phenomena, vol 2. Academic Press, London, p 197
5. Hughes BD (1996) Random Walks and Random Environments, vol 2. Oxford University Press, London
6. Sahimi M (1993) Flow phenomena in rocks: from continuum models to fractals, percolation, cellular automata, and simulated annealing. Rev Mod Phys 65:1393
7. Sahimi M (1994) Applications of Percolation Theory. Taylor and Francis, London
8. Sahimi M (1995) Flow and Transport in Porous Media and Fractured Rock. VCH, Weinheim
9. Sahimi M (2003) Heterogeneous Materials I & II. Springer, New York
10. Stauffer D, Aharony A (1994) Introduction to Percolation Theory, 2nd edn. Taylor and Francis, London
11. Straley JP (1976) Critical phenomena in resistor networks. J Phys C 9:783
12. Straley JP (1977) Critical exponents for the conductivity of random resistor lattices. Phys Rev B 15:5733

Percolation and Polymer Morphology and Rheology

Muhammad Sahimi
Mork Family Department of Chemical Engineering and Materials Science, University of Southern California, Los Angeles, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Percolation Models of Polymerization and Gelation
Branched Polymers
Rheology of Critical Gels
Resistor and Elastic Percolation Networks
Nearly-Critical Chemical Gels: Comparison of the Data with the Percolation Models
Physical Gels: Comparison of the Data with the Percolation Model
Future Directions
Bibliography

Glossary

Branched polymers Large polymers below the gel point with radii larger than the correlation length

Chemical gels Gel networks in which the monomers are co-valently bonded.

Critical gels The gel networks at the gel point

Elastic networks Networks in which each bond is an elastic element, such as a Hookean spring.

Gel point The point at which the critical gel network is formed for the first time.

Lattice animals Large percolation clusters below the percolation threshold with radii larger than the correlation length

Physical gels Gel networks in which the monomers or particles are connected through weak association.

Relaxation time spectrum The distribution of relaxation times that describes the linear viscoelastic behavior of liquids and solids.

Resistor networks Networks in which each bond is a resistor with a given conductivity.

Rigidity percolation Percolation networks in which each uncut bond is a Hookean spring, and there are no angle-changing forces.

Sol The solvent + finite polymer clusters below the gel point

Definition of the Subject

Elastic percolation networks described in the article by Duxbury (see also below), random resistor networks described in the articles by Hughes and Balberg, and a few continuum models [58] of heterogeneous materials provide a comprehensive understanding of their transport properties. The purpose of this chapter is to describe and discuss applications of such percolation models to predicting the structure and rheolgy of an important class of disordered materials, namely, polymers and gel networks, and test their validity by comparing their predictions with the relevant experimental data.

The formation of the polymeric materials that we consider in this chapter is characterized by the existence of a percolation-type transition point; see below. The rigidity and linear elastic properties of disordered materials, including polymers, that are far from their percolation threshold are well-described and predicted by mean-field theories, such as the effective-medium approximation [33,41]. However, the effective properties of polymeric materials that are near the percolation threshold deviate greatly from the predictions of mean-field theories and other approximations. It is the description of various properties of such polymeric materials that is best done by the percolation models, and is the focus of this chapter.

Introduction

Polymeric materials have wide applications in many branches of science and technology. In addition, they have many interesting, and in many cases unusual, properties that justify their study. One is that their relaxation modes are described by a wide spectrum, which provides clues to their structure (see below). Each mode is associated with a particular "event" or motion. In particular, motion of clusters of monomers or molecules is associated with the long modes, with the longest relaxation modes being due to the very large clusters. Such clusters are formed either by the formation of *chemical bonds* or *chemical crosslinking* between the monomers and also between monomers and small molecules or clusters which, when large enough, lead to *phase-separation*, or by *physical association* at the molecular or particulate level. It is the formation of such large clusters that is the root cause of an important class of phase transitions, namely, the *liquid–solid transition* (LST).

The significance of the LST cannot be over-emphasized, as it occurs in a wide variety of problems of practical importance. At the most basic level, it is important to be able to predict the point at which the LST occurs. The knowledge is sometimes necessary in order to design better polymer processing operations. An example is injection molding of semicrystalline polymers, the quality of the surface of which is a strong function of the location of the LST point. In other applications the knowledge of the LST point is necessary in order to *postpone* it or *avoid* it altogether. Since near the LST point the system is a mixture of a liquid and solid clusters, one may be able to design a wide variety of materials by changing the volume fraction of each phase and, therefore, study of the LST is important. At the same time, crosslinked polymers near the LST point are good adhesives, and have also been used as materials for membranes, absorbers, and many other applications.

In addition to chemists and chemical and polymer engineers who have traditionally studied polymers and their properties, a seminal work in the early 1970s attracted the attention of physicists, and opened the way to the application of modern methods of statistical mechanics to the study of polymers. De Gennes [71] demonstrated that there is a close connection between linear polymers – those in which the monomers have functionality or coordination number $Z = 2$ – and a statistical mechanical model, namely, the n-vector model. Clearly, no two monomers occupy the same point in space. If, in addition, there is no closed loop in the structure of the polymer, then, the result is a *linear polymer* which, as de Gennes showed, corresponds to the limit $n \to 0$ of n-vector model. The most suitable model for such polymers is the path of a self-avoiding walk in which a particle performs a random walk in s pace with the restriction that it never visits any point more than once (so that loop formation is avoided). De Gennes' discovery made it possible to apply modern methods of statistical mechanics, and in particular the renormalization group theory and the scaling concepts, to the study of linear polymers and, later, to branched polymers and gels.

The original work of de Gennes was restricted to linear polymers. However, if the monomers have functionalities

$Z > 2$, so that each of them may be connected with more than two neighboring monomers, then at least two other classes of polymers can be obtained:

(i) If the reaction time t is relatively short and below, but close to, a characteristic time t_g, then one obtains branched polymers in the solution, usually called *sol*, that form a viscous solution. It is called a sol because it is soluble in good solvents. Such branched polymers are large but finite clusters of monomers. t_g is called the *gelation time*.
(ii) If, however, the reaction time is larger than t_g, an infinitely large solid network of reacted monomers appears which is usually called a *chemical gel*, or simply a gel. There is clearly a LST as one passes from the sol to the gel phase, and the process, called *gelation*, has been described by the percolation models.

The gel can only swell, but not dissolve, in a solvent, even if finite clusters of reacted monomers still exist in the system. The point at which the gel network appears for the first time – which is, in fact, the point that signals the LST – is called the *gel point* (GP).

The gel network has interesting structural, mechanical, and rheological properties which are described in this chapter. Most of us are already familiar with such sol–gel transformations in our daily lives, since we all know about milk-to-cheese transition, pudding, gelatine, etc. However, materials that contain gels, or use their specific properties, are numerous. In addition to the examples mentioned above, another important example is the eye humor. Moreover, gels play an important role in laboratory technology (e. g., gel chromatography), in the fabrication of a wide variety of products, such as glues, cosmetics, contact lenses, etc., and in food technology. In addition, the sol–gel transition is a general phenomenon that has been utilized for producing a variety of ceramic materials [19].

Chemical reactions are responsible for the interconnectivity of the monomers in chemical gels. In general, there are three types of chemical gelation:

(i) *Polycondensation*, in which polymerization begins with either bifunctional units A-A or trifunctional ones B_3, or more generally Z-functional units B_Z. The A units are linked with the B units, with each elementary reaction being accompanied by the elimination of a molecule between units of A and B. Thus, a polymer network is formed in which the polymer chains are terminated by either A or B. No two units of the same class can participate in a reaction with each other and, therefore, there is always exactly *one* bifunctional unit between polyfunctional units in the polymer network.
(ii) *Vulcanization* begins with long linear polymer chains in a solution. The chains are then crosslinked by small units. An example is rubber, the elasticity of which is due to the introduction of S–S bonds between polyisoprene chains. Only a small number of bonds are needed to crosslink the chains and form an interconnected polymer network.
De Gennes [70] argued that for volcanizing polymers – those with high molecular weights – there exists only a very narrow region near the GP in which the percolation model may be applicable. In other words, such polymers exhibit the Flory–Stockmayer-type behavior [34,63] which is of the mean-field type. For this reason, we ignore volcanization in this chapter.
(iii) *Additive polymerization* is similar to polycondensation. The initial solution contains two types of units. The A=A units that are bifunctional when the double bond opens, and the B=D=B units that are quadrifunctional when the two double bonds open independently of each other. If the reaction polymerizes A=A units, one obtains A–A–A–A–$\cdots$ chains, whereas reaction between the A units and the B=D=B units reticulates the network. The length of the chains between two reticulation points is not fixed, but depends crucially on the initiation process and on the relative concentrations of the bi-and quadrifunctional units.

In addition to such chemical gels, one may also have *physical gels* in which the monomers or particles are attached to each other by relatively weak and *reversible* association, or by such physical processes as entanglement. A well-known example is silica aerogel. Another example is a solution of gelatin in water below a certain critical temperature where a coil-to-helix transition takes place, and bonds appear to form by winding of helixes of two adjacent chains. Such physical gels can be made and also destroyed by thermal treatment. Other important examples include liquid crystalline polymers at their nematic-to-smectic transition, suspensions and emulsions at the percolation threshold, partially crystalline polymers, and microphase-separating block copolymers.

In this chapter we describe modeling of structural, mechanical and viscoelastic properties of the sol and gel phases, especially near the GP. Modeling of gel network formation was pioneered by Flory [34] and Stockmayer [63], whose theory is essentially equivalent to percolation on the Bethe lattice, an endlessly branching net-

work without any closed loops. Stauffer [62] and de Gennes [72] emphasized the importance of the deviations from the Bethe lattice solution of Flory and Stockmayer, and proposed to replace it by percolation on three-dimensional (3D) lattices. This aspect of the problem, which can be described by a random percolation model or a variant of it, is now well-understood. De Gennes [72] also proposed that the elastic and viscoelastic properties of the gel and sol phases can be described by appropriate random resistor network models (see below). His suggestion was widely accepted for a long time, and was utilized for interpreting the experimental data. It was recognized in the 1980s that, while de Gennes' suggestion may be applicable to certain classes of polymeric materials, more general models are needed for several other important classes of such materials. This realization gave rise to the development of the elastic percolation models that are described in the article by Duxbury. We shall come back to this point shortly.

Percolation Models of Polymerization and Gelation

To see the connection between the sol–gel transition and the percolation model, consider a solution of molecules or monomers with functionality $Z \geq 3$. Suppose, for simplicity, that the monomers occupy the sites of a periodic lattice. With probability p, two nearest-neighbor monomers (sites) react and form a chemical bond between them. If p is small, only small polymers (clusters of reacted monomers) are formed. As p increases, increasingly larger polymers with a broad size distribution are formed. This mixture of clusters of reacted monomers and the isolated unreacted monomers represents the sol phase. For $p > p_c$, where p_c is a characteristic value that depends on the functionality Z (the number of nearest neighbors of a monomer in the lattice), an infinite cluster of reacted monomers is formed which represents the gel network described above. The gel network *at* the GP is usually called the *critical gel*. Near the GP the gel usually coexists with the sol such that the finite polymers are trapped in the interior of the gel. For $p \to 1$, almost all the monomers react, and the sol phase disappears completely. Thus, p_c signals a *connectivity* transition: For $p > p_c$, an infinite cluster (the get network), together (possibly) with a few finite-size clusters, exist and, thus, the system is mainly a rigid gel. The fraction of chemical bonds formed at the GP (which is related to the fraction of the reacted monomers at the GP) is obviously the analogue of the bond percolation threshold. Thus, it should be clear that the formation of branched polymers and gels is very similar to a percolation process.

In the earlier days of modeling the sol–gel transition, it was generally believed that the properties of the polymeric materials at the GP are independent of the structural details of the materials. But, despite decades of research, this is still an unproven hypothesis for the critical gel. In addition, the monomers do not react randomly. There are usually some correlations in the way the monomers react with one another.

Structural Properties of Branched Polymers and Gels

Studies of the sol–gel transition usually proceed by measuring the time evolution of the rheological (e. g., the viscosity) or mechanical properties (e. g., the elastic moduli) during the chemical reaction that leads to gelation, assuming that the experimental parameter – time or frequency – and the theoretical one – the number of crosslinks – are linearly related in the vicinity of the GP. If true, then, near the GP, $|p - p_c|$ is a measure of the distance from the GP. All the properties of nearly critical gels (i. e., those that are very near the GP) can be expanded in powers of $|p - p_c|$. This is completely similar to, for example, the vapor-liquid phase transition for which all the properties of the system near the critical temperature T_c can be expanded in powers of $|T - T_c|$. But, as the distance from the GP increases, the expansions breakdown.

Experimental rheological measurements are usually performed by using a cone and plate rheometer, or by the more accurate magnetic sphere rheometer. The ranges of shear rates, deformations, and times of measurements of such devices allow the determination of the steady-state zero-shear viscosity and the steady-state linear elastic moduli up to the vicinity of the LST at the GP, but it has proven to be almost impossible to do such measurements *at* the GP (see below).

The correlation or connectivity length ξ, which represents the typical size of the branched polymers below p_c, diverges as p_c is approached according to the power law

$$\xi \sim |p - p_c|^{-\nu} , \tag{1}$$

which is completely similar to ξ_p, the correlation length of percolation clusters. Below p_c, however, the polymers with radii much larger than ξ have completely different characteristics than those with a typical radius ξ. Therefore, we describe and discuss such polymers separately, and refer to them as the *branched polymers* to distinguish them from gel networks. Above the GP the correlation length of the polymers is taken as the mesh size of the gel network. For any length scale greater than ξ the gel network is homogeneous.

Scaling Properties of the Structure of Nearly Critical Gels

Several important structural properties of branched polymers and gel networks can be measured directly or indirectly. The *gel fraction* $f_g(p)$ is the fraction of the monomers that are in the gel network, and is measured by simply weighing the gel at different times during the polymerization process. Clearly, $f_g(p) > 0$ only if $p > p_c$. As far as the analogy with the percolation model is concerned, f_g is the analogue of percolation fraction or percolation probability $P(p)$ (see the article by Sahimi). Of particular interest to us is the behavior of $f_g(p)$ near p_c. In this region,

$$f_g(p) \propto (p - p_c)^{\beta} \ . \tag{2}$$

The number distribution of the polymers, i. e., the probability $Q(s, \epsilon)$ that a polymer in the sol phase contains s monomers at a distance $\epsilon = |p - p_c|$ from the GP, is clearly the analogue of the cluster size distribution n_s for percolation clusters (see the article by Sahimi). Thus, we may write

$$Q(s, \epsilon) \sim s^{-\tau} h_1(\epsilon s^{\sigma}) \ , \tag{3}$$

where h_1 is a universal scaling function. Instead of writing the distribution in terms of s, we may write down a power law for the cluster *mass distribution*, $N(M)$, in terms of the molecular weight M of the finite polymers. At the GP, one has,

$$N(M) \propto M^{-\tau} \ , \quad p = p_c \tag{4}$$

Then, near the GP, one has

$$N(M) \propto M^{-\tau} h_2(M/M_z) \ , \tag{5}$$

which is completely similar to Eq. (3), where h_2 is another scaling function, closely related to h_1.

Using $Q(s, \epsilon)$, we define two distinct mass averages. One, the *weight-average molecular weight*, is defined by

$$M_w = \frac{\int s^2 Q(s, \epsilon) \mathrm{d}s}{\int s Q(s, \epsilon) \mathrm{d}s} \sim |p - p_c|^{-\gamma} \ , \tag{6}$$

where $\gamma + 2\beta = \nu d$, with d being the dimensionality of the material. In the analogy with the percolation model, M_w is the analogue of the mean-cluster size (see the articles by Stauffer and Sahimi). In the polymer literature M_w is also called the *degree of polymerization*. The second mass average is defined by

$$M_z = \frac{\int s^3 Q(s, \epsilon) \mathrm{d}s}{\int s^2 Q(s, \epsilon) \mathrm{d}s} \sim \epsilon^{-1/\sigma} \sim |p - p_c|^{-1/\sigma} \ , \tag{7}$$

where M_z is the same quantity as in Eq. (5), and, $\sigma = (\tau - 2)/\beta$. Note, however, that the average

$$\langle M \rangle = \frac{\int s Q(s, \epsilon) ds}{\int Q(s, \epsilon) ds}$$

does *not* diverge at the GP. Note also that we may also express M_w in terms of ξ:

$$M_w \propto \xi^{\gamma/\nu} \ . \tag{8}$$

Recall (see the article by Stauffer) that percolation theory predicts that for 3D systems, $\nu \simeq 0.89$, $\beta \simeq 0.41$, $\tau \simeq 2.18$, $\sigma \simeq 0.46$, and $\gamma \simeq 1.82$. The mean-field theory of Flory and Stockmayer predicts the same type of power laws, but with, $\beta = \gamma = 1$, $\nu = \sigma = 1/2$, and $\tau = 5/2$.

At the GP the gel network is *not* homogeneous, but is a self-similar fractal object with a fractal dimension D_f that in d-dimensions is given by

$$D_f = d - \beta/\nu = d(\tau - 1)^{-1} \ , \tag{9}$$

Using numerical estimates of ν and β for the 3D percolation model, we obtain, $D_f \simeq 2.53$. On the other hand, the Flory–Stockmayer theory predicts that, $D_f = 4$, which is unphysical since D_f cannot be larger than 3.

Comparison of the Data with the Percolation Model

Since a main prediction of percolation is the existence of universal critical exponents and the fractal dimension D_f, and because the numerical value of any polymer property, such as it average molecular weight or the location of the GP, is not universal and depends on the structure of the polymer, we focus on a comparison between the measured exponents, such as β, γ and σ, and the predictions of the percolation model.

In their experiments with irradiated polystyrene solution in cyclopentane, Leibler and Schosseler [44] coupled gel permeation chromatography and light scattering to deduce the polymer size distribution which provides a direct means of measuring the exponent τ. Figure 1 presents their measurements from which one obtains, $\tau \simeq 2.3 \pm 0.1$, close to the percolation prediction of 2.18. Lapp et al. [43] further checked this result by carrying out similar experiments in a system made by chemical end-linking of polydimethylsiloxane, and Patton et al. [51] performed experiments in a system in which polyester was made by bulk condensation polymerization. The measurements of both groups were consistent with the value of τ predicted by the percolation model.

Adam et al. [3] tested the validity of the percolation model based on Eq. (6). They carried out static

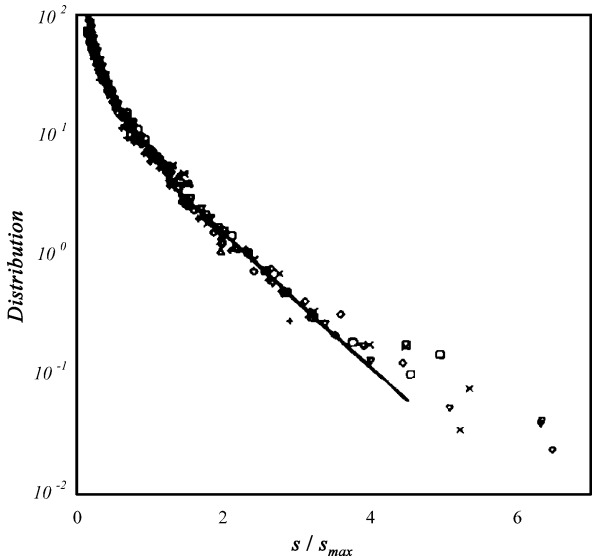

Percolation and Polymer Morphology and Rheology, Figure 1
Normalized polymer size distribution as a function of polymer size s. Percolation theory predicts the slope to be $1 - \tau \simeq -1.3 \pm 0.1$ (after Leibler and Schossler [44])

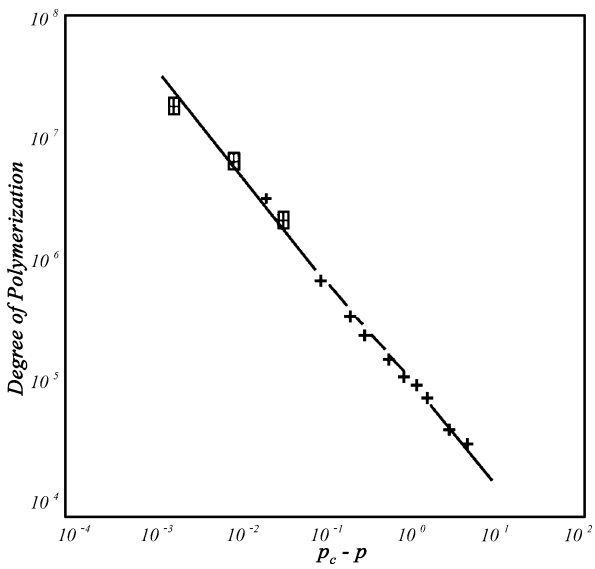

Percolation and Polymer Morphology and Rheology, Figure 2
Dependence of degree of polymerization M_w on $p_c - p$ for a polyurethane sol. The slope of the curve is $-\gamma$ (after Adam et al. [3])

light scattering measurements on a polyurethane sol. Candau et al. [20] performed their experiments on polystyrene systems crosslinked with divinylbenzene. Figure 2 displays the results of Adam et al. [3] from which one obtains, $\gamma \simeq 1.71 \pm 0.06$, only 5% smaller than the percolation prediction of 1.82. A similar estimate of γ was reported by Candau et al. [20]. On the other hand, one can also express the weight-average molecular weight in terms of the gel fraction near the GP, $M_w \sim f_g^{-\gamma/\beta}$ and, thus, a plot of $\log(M_w)$ versus $\log(f_g)$ yields an estimate of γ/β. Schmidt and Burchard [61] carried out anionic copolymerization of divinylbenzene with styrene and obtained both branched polymers (see below) and gels. Light scattering was used to measure the various properties of interest. When Schmidt and Burchard [61] plotted $\log(M_w)$ versus $\log(f_g)$, they obtained a straight line with the slope, $\gamma/\beta \simeq 4.5$, in good agreement with the percolation prediction, $\gamma/\beta \simeq 4.44$.

Branched Polymers

After a polymer is formed by crosslinking, the experimentalist usually analyzes its structure by diluting it in a good solvent. As mentioned above, branched polymers in a dilute solution of a good solvent may swell and have a radius *larger* than their extent at the end of the crosslinking. Thus, it is important to consider both typical polymers, which we already described above, and the swollen ones which we now consider.

Thus, consider a swollen branched polymer in a good solvent with a radius larger than the polymer correlation length ξ. The structural properties of such branched polymers are described by *lattice animals*, which are, in fact very large percolation clusters below the percolation threshold. Their radii are *larger* than the percolation correlation length ξ_p. But, the interesting and important point is that, although lattice animals are simply very large percolation clusters below p_c, their statistics are completely *different* from those of percolation clusters.

Statistics of Lattice Animals

To better understand the difference between percolation clusters and lattice animals, let us first define a few key statistics of lattice animals. Suppose that $A_s(p)$ is the average number (per lattice site) of the clusters, and a_{sm} the total number of geometrically different configurations for a cluster of s sites and perimeter m. Thus, $A_s(p) = \sum_m a_{sm} p^s (1-p)^m$. The asymptotic behavior of $A_s(p)$ for large values of s is described by the power law

$$A_s(p) \sim s^{-\theta} , \tag{10}$$

where θ is a universal exponent, independent of the coordination number of the lattice. Moreover, for large values of s a fractal dimension D_f, defined by

$$s \sim R^{D_f} , \tag{11}$$

describes the structure of the animals or the branched polymers, where R is the radius of the lattice animal. Note that the fractal dimension D_f is *distinct* from that of critical gels given by Eq. (9). Lubensky and Isaacson [46] and Family and Coniglio [31] showed that the exponents θ and D_f are not related to any of the percolation exponents defined above. Moreover, Parisi and Sourlas [50] showed that

$$\theta = \frac{d-2}{D_f} + 1 , \tag{12}$$

and that

$$D_f = 2 , \quad d = 3 . \tag{13}$$

One key difference between lattice animals and percolation clusters is that, the exponents θ and D_f are defined for *any* $p < p_c$ (recall that the percolation exponents are defined for $p \simeq p_c$), so long as R, the animals' radius, is much larger than the correlation length ξ. For this reason, they are called *non-critical* exponents.

We may also define a pair correlation function $C(r)$, i. e., the probability that two monomers (or sites)that are separated by a distance r belong to the same polymer (or cluster). For a d-dimensional branched polymer and large r, we expect the correlation function to decay as

$$C(r) \sim r^{D_f - d} . \tag{14}$$

The Fourier transform of $C(r)$ is proportional to the scattered intensity $I(q)$ in an X-ray or a neutron scattering experiment, where q is the magnitude of the scattering vector, given by, $q = (4\pi/\zeta)\sin(\theta/2)$, with θ being the wavelength of the radiation scattered by the material through an angle θ. Thus, by Fourier transforming Eq. (14), one obtains

$$I(q) \sim q^{-D_f} . \tag{15}$$

In practice, however, polymer solutions are almost always polydispersed and contain polymers of all sizes with radii that may be smaller or larger than the correlation length ξ. Thus, one must define *average* properties, where the averaging is taken over the polymer size distribution. An average polymer radius is defined by,

$$\langle R \rangle = \frac{\sum_s s^2 R(s) Q(s,\epsilon)}{\sum_s s^2 Q(s,\epsilon)} , \tag{16}$$

which, when combined with Eqs. (3) and (11), yields a relation between s and $\langle R \rangle$ [26]:

$$s \sim \langle R \rangle^{D_f(3-\tau)} , \tag{17}$$

so that, in analogy with Eq. (11), an *effective fractal dimension*, $D_f^e = D_f(3-\tau)$, may be defined. Note that, Eq. (17) mixes the branched polymers fractal dimension D_f with the gel exponent τ. If a percolation description of polymerization is correct (which the experiments described above confirmed it to be the case), we should have,

$$D_f^e \simeq 1.64 , \quad d = 3 , \tag{18}$$

indicating that the effective fractal dimension is smaller than that of a single branched polymer. Because an effective fractal dimension has been defined for a dilute polydisperse polymer solution, the scattering intensity for the same solution should also be modified to

$$I(q) \sim q^{-D_f(3-\tau)} . \tag{19}$$

Comparison of the Data with the Lattice Animal Model

Experimental evidence for Eq. (13) is actually provided through Eqs. (15) and (19). Bouchaud et al. [18] carried out small-angle neutron scattering experiments on a monodisperse polyurethane sample and measured the scattered intensity as a function of q. Figure 3 presents their data for the polymer from which one obtains, $D_f = 1.98 \pm 0.03$, in excellent agreement with Eq. (19). Bouchaud et al. [18] also synthesized a natural polydisperse polyurethane sample and carried out small-angle neutron scattering on a dilute solution of it. Their data

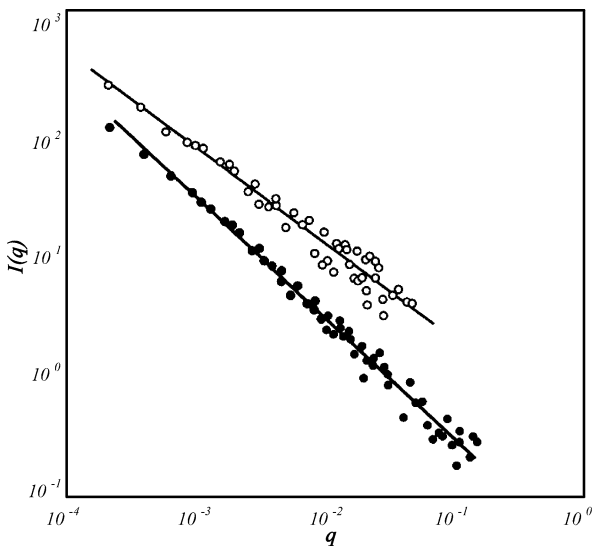

Percolation and Polymer Morphology and Rheology, Figure 3
Small-angle neutron scattering data for branched polymers. The *upper curve* is for a polydisperse polymer solution with a slope −1.6. The *lower curve* is for a single polymer in a good dilute solvent with a slope −1.98 (after Bouchaud et al. [18])

yielded the estimate, $D_f^e \simeq 1.6 \pm 0.05$, in good agreement with the theoretical prediction given by (18).

Adam et al. [3] carried out static light scattering experiments with dilute polydisperse polyurethane solutions and reported that, $D_f^e \simeq 1.62 \pm 0.08$, again in good agreement with (18). Leibler and Schosseler [44] measured the average radius of polystyrene, crosslinked by irradiation by elastic light scattering and found that, $D_f^e \simeq 1.72 \pm 0.09$, relatively close to the estimate obtained from Eq. (18). Patton et al. [51] performed both quasi-elastic and elastic light scattering experiments on branched polyesters and reported that, $D_f^e \simeq 1.52 \pm 0.1$, somewhat lower than the prediction (18), but still consistent with it.

Rheology of Critical Gels

One way of understanding what happens to the polymers as the GP is approached is through rheological experiments. Such experiments have been described in details by Winter and Mours [67] and, therefore, we provide only brief description of them and discuss their implications. In practice one imposes a small step shear strain ϵ_{zx} on a sample near the GP and measures the shear stress $\sigma_{zx}(t)$ as a function of time t. The key property is then the shear stress relaxation function, $G(t) = \sigma_{zx}/\epsilon_{zx}$, which is also referred to as the relaxation modulus. *Unlike* the elastic moduli, $G(t)$ can be measured for both liquids and solids and, therefore, it is a very useful property for studying the sol–gel transition and, more generally, any LST. Note that t, the time of crosslinking reaction, corresponds to the extent p of the reaction, the key property in percolation theory.

What happens to $G(t)$ as the GP is approached? The stress relaxes quickly at the early stages of crosslinking. As more chemical bonds are formed between the monomers, however, $G(t)$ stretches out further, since the relaxation process requires longer times. Exactly at the GP (and, more generally, at any LST point), the material is neither a liquid nor a solid yet (it has a tenuous fractal structure). The relaxation modulus follows a power law:

$$G(t) = G_0 t^{-n} , \quad t_0 < t < \infty \tag{20}$$

where G_0 is the gel stiffness. The parameters G_0, n, and t_0 all depend on the material structure at the GP. The exponent n is closely related to the exponents that characterize the power-law behavior of viscosity of the sol and the elastic moduli of the gel network near the GP (see below).

As the polymerization proceeds further and the GP p_c is passed, the material becomes a solid which has a finite relaxation modulus at long times, usually referred to as the equilibrium modulus G_e:

$$G_e = \lim_{t\to\infty} G(t) . \tag{21}$$

Under such conditions, the stresses can no longer relax completely.

The Relaxation Time Spectrum

Since the time dependence of a macroscopic relaxation process is always indicative of the underlying microscopic dynamics, one may look for kinetic equations that correctly describe the time-dependence of the observed responses of a material. In the simplest case, there is only a single characteristic time τ, the origin of which goes back to Debye who proposed it in his seminal work on the dielectric response of polar liquids. If we define a shear compliance (the inverse of shear modulus) $J(t)$ by

$$J(t) = \frac{\epsilon_{zx}}{\sigma_{zx}^0} ,$$

then, applying an oscillatory shear stress

$$\sigma_{zx}(t) = \sigma_{zx}^0 \exp(i\omega t) ,$$

on a polymer means imposing an oscillatory strain, $\epsilon_{zx}(t) = \Delta J \sigma_{zx}(t)$. Here, ΔJ is the *relaxation strength* of the material. The governing equation for $\epsilon_{zx}(t)$ is then given by

$$\frac{d\epsilon_{zx}(t)}{dt} = -\frac{1}{\tau}\left[\epsilon_{zx}(t) - \Delta J \sigma_{zx}^0 \exp(i\omega t)\right] . \tag{22}$$

Assuming a solution, $\epsilon_{zx}(t) = \sigma_{zx}^0 J^*(\omega)\exp(i\omega t)$, where $J^*(\omega)$ is the *complex shear compliance*, substituting it into Eq. (22) and solving it, yield

$$J^*(\omega) = \frac{\Delta J}{1 + i\omega\tau} , \tag{23}$$

which is usually referred to as a *Debye process*.

As discussed above, however, the dynamics of polymers and gels in the reaction bath cannot be described by a single relaxation time, rather by a statistical distribution of such characteristics times. For example, for the shear properties, we write the dynamic compliance $J^*(\omega)$ as a sum of the Debye processes with relaxation times τ_i and relaxation strengths ΔJ_i, so that

$$J^*(\omega) = J_u + \sum_i \frac{\Delta J_i}{1 + i\omega\tau_i} . \tag{24}$$

The sum is usually replaced by an integral, so that

$$J^*(\omega) = J_u + \int \frac{\mathcal{R}(\tau)}{1 + i\omega\tau} d\tau , \tag{25}$$

where $\mathcal{R}(\tau)$ is called the *retardation time spectrum* of the shear compliance J^*. One may also write these results in terms of the complex modulus $G^*_{\text{eff}}(\omega) = 1/J^*(\omega)$, which then yields

$$G^*(\omega) = G_u - \int \frac{H(\tau)}{1 + i\omega\tau} \, \mathrm{d}\tau \,, \tag{26}$$

where $H(t)$ is the *relaxation time spectrum* of the complex modulus $G^*(\omega)$ [or $G(t)$]. In practice, $H(t)$ cannot be directly measured, but is inferred only indirectly.

Dynamic Mechanical Experiments

The evolution of the molecular structure of a polymer during the gelation process has a profound effect on the molecular mobility, which can be monitored by probing the changes in the viscosity and elastic moduli. The initial ($p = 0$) liquid system has a steady shear viscosity η which increases with the extent of the reaction as the average molecular weight M_w increases. At the GP, the viscosity and the longest relaxation time τ_max diverge. Beyond p_c, the equilibrium elastic moduli increase until they attain their highest values, which is when the reaction is brought to completion, i. e., when $p \to 1$.

All the experimental data for the elastic moduli of the nearly-critical gel network indicate that the effective elastic moduli G_eff follow a power law:

$$G_\text{eff} \sim (p - p_\text{c})^z \,, \quad p > p_\text{c} \,. \tag{27}$$

On the other hand, near the GP, the viscosity of the sol phase also follows a power law, resulting in its divergence at the GP:

$$\eta \sim (p_\text{c} - p)^{-k} \,, \quad p < p_\text{c} \,, \tag{28}$$

while, as shown below, the longest relaxation time diverges as

$$\tau_\text{max} \sim |p - p_\text{c}|^{-k-z} \,, \quad |p - p_\text{c}| \ll 1 \,. \tag{29}$$

The divergence of the viscosity at the GP is precisely due to the divergence of the mean polymer (cluster) size at the GP which, near the GP, follows a power law similar to Eq. (6) for the weight-average molecular weight M_w with precisely the same exponent γ.

In practice, it is precisely the divergence of η that signals the formation of the critical gel network. Due to the divergence of τ_max measurements of the viscosity and elastic moduli fail at p_c, since steady-state conditions cannot be reached in a finite time. Another difficulty is that precise measurement of the GP is often difficult. Such difficulties are partially overcome by performing *dynamic mechanical experiments*. In such experiments the sample is exposed to a periodically varying stress field. For example, a tensile stress $\sigma_{zz}(t)$ is used

$$\sigma_{zz}(t) = \epsilon^0_{zz} \exp(i\omega t) \,, \tag{30}$$

which results in a time-dependent longitudinal strain $\epsilon_{zz}(t)$ that varies with the frequency of the stress, but shows, in general, a phase-lag φ, such that

$$\epsilon_{zz}(t) = \epsilon^0_{zz} \exp[-i(\omega t - \varphi)] \,. \tag{31}$$

We may then employ a *dynamic tensile modulus* $G^*(\omega)$, defined as

$$G^*(\omega) = \frac{\sigma_{zz}(t)}{\epsilon_{zz}(t)} = G'(\omega) + iG''(\omega) \,. \tag{32}$$

Analogous experiments can, of course, be carried out for other types of mechanical loading. Of particular interest are measurements under simple shear which determine the relation between the shear strain ϵ_{zx}, yielding the displacement along x per unit distance normal to the shear plane $z =$ constant, and the shear stress σ_{zx} that acts on the shear plane along x.

In any case, such dynamic mechanical experiments measure the small amplitude oscillatory shear behavior of evolving gels. Under this condition, the gel evolution is continuous (no singularity). But, even such experiments cannot entirely overcome the difficulties in the determination of the exponent k of the viscosity, because the measurements cannot be carried out *at* the GP and in the limit of *zero frequency*. To estimate k one usually measures the frequency-dependent complex modulus $G^*(\omega)$ at frequency ω. At the GP and for low frequencies, one has

$$G' \sim G'' \sim \omega^n \,, \quad p = p_\text{c} \,, \tag{33}$$

with

$$n = \frac{z}{z + k} \,, \tag{34}$$

where G' (the storage modulus) and G'' (the loss modulus) describe storage and dissipation in an oscillating strain field of constant amplitude. Note that the exponent n in Eq. (34) is the same as that in Eq. (20). Typical variations of G' and G'' with ω are shown in Fig. 4 for a polycondensed gel very close to the GP.

The complex modulus $G^*(\omega)$ is sometimes written as $G^* = G + i\omega\eta$, for which Durand et al. [29] proposed that

$$G^*(\omega, \epsilon) \sim \epsilon^z h_3\big(i\omega|p - p_\text{c}|^{z+k}\big) \,, \tag{35}$$

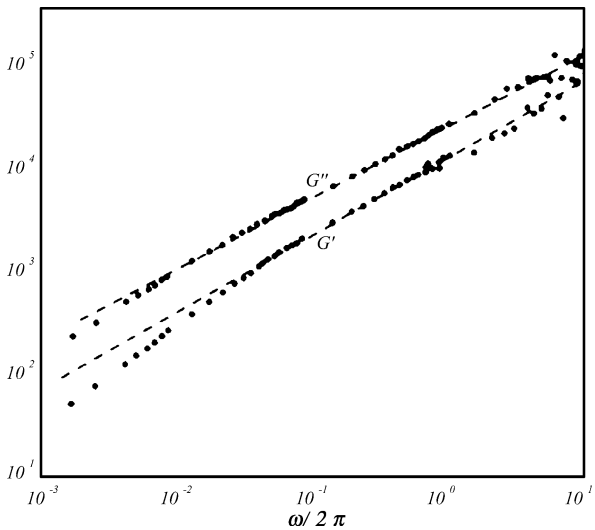

Percolation and Polymer Morphology and Rheology, Figure 4
Frequency-dependence of the storage modulus G' and loss modulus G'' for a polycondensed gel close to the gel point (after Durand et al. [29])

where $h_3(x)$ is a universal scaling function. The significance of scaling Eq. (35) is that it enables one to collapse the data for all values of $|p - p_c|$ and ω onto a single curve, usually called the *master curve* by polymer researchers. In the low-frequency regime, we do not expect $G^*(\omega)$ to depend on $|p - p_c|$, but only on ω, in which case one finds that, $G^* \sim (i\omega)^n$, which is equivalent to Eq. (33). Moreover, there is a loss angle δ defined by, $\tan\delta = G'/G''$. The remarkable property of δ is that at the GP it takes on a value δ_c given by

$$\delta_c = \frac{\pi}{2}(1-n) = \frac{\pi}{2}\frac{k}{z+k}\,, \tag{36}$$

so that, if the exponents z and k are universal, so will also be the loss angle δ_c.

Self-Similar Relaxation Time Spectrum

The observed power-law behavior of G' and G'', Eq. (33), implies a relaxation time spectrum which is self-similar (in time):

$$H(t) = \frac{G_0}{\Gamma(n)}\left(\frac{t}{\tau_0}\right)^{-n}\,, \tag{37}$$

where G_0 is the characteristic modulus, τ_0 is the characteristic *shortest relaxation time*, and Γ is the gamma function. The modulus of a fully crosslinked polymer network is typically 10^6–10^7 Pa, while the relaxation time of the network strand is about 10^{-7}–10^{-4} sec. The spectrum $H(t)$ extends from the shortest time, at which the strands are beginning to be probed, to the infinite relaxation time of the critical gel network. The parameters G_0 and τ_0 are material characteristics of the gel system. Most gel systems seem to possess the same value of n. However, there are also gels which exhibit no apparent universality in the value of n.

Viscosity and elastic moduli are rheological and mechanical properties of branched polymers and gels which characterize the dynamics of the polymerization, since we may measure indirectly the distribution of the relaxation times $H(t)$ in the reaction bath. The moments of $H(t)$ are directly related to the viscosity and the elastic moduli. Using Eqs. (35) and (37), we can back-calculate $H(t)$ [23]:

$$H(t) \sim t^{-n} h_4\left(t|p-p_c|^{z+k}\right)\,, \tag{38}$$

where h_4 is another universal scaling function. Equation (38), which indicates that in the scaling regime near the GP the relaxation time distribution is a slowly decaying power law, generalizes Eq. (37) to any value of p, the extent of the polymerization. Equations (37) and (38) indicate that *any* relaxation property in the intermediate time or frequency range is *not* exponential, but follows a *power law*. As pointed out by Daoud [23], two distinct averages or characteristic times may be defined. One is

$$\tau_1 = \frac{\int H(t)\mathrm{d}t}{\int [H(t)/t]\mathrm{d}t} \sim |p-p_c|^{-k} \propto \eta\,, \tag{39}$$

while the second one is given by

$$\tau_2 = \frac{\int tH(t)\mathrm{d}t}{\int H(t)\mathrm{d}t} \sim |p-p_c|^{-z-k}\,. \tag{40}$$

Note that τ_2 is in fact identical with $\tau_{\max}$, the longest relaxation time of the gel; see Eq. (29).

Determination of the Gel Point

As already mentioned above, an important problem in polymerization and gelation is accurate determination of the GP, either for avoiding it in order to prevent gelation (so that a branched polymer with specific properties can be prepared), or for making polymeric materials very close to the GP, as they have unusual properties in that region. The GP, which is the analogue of a percolation threshold, depends on the functionality Z of the polymer – the analogue of the coordination number in the lattice models – and decreases with increasing Z. Thus, polymers with crosslinks of high functionality gel very early. Holly et al. [39] proposed using the loss angle δ for locating the GP. They argued that, because as the GP is reached $\tan\delta$ becomes independent of the frequency [see Eq. (36)], then, if one plots

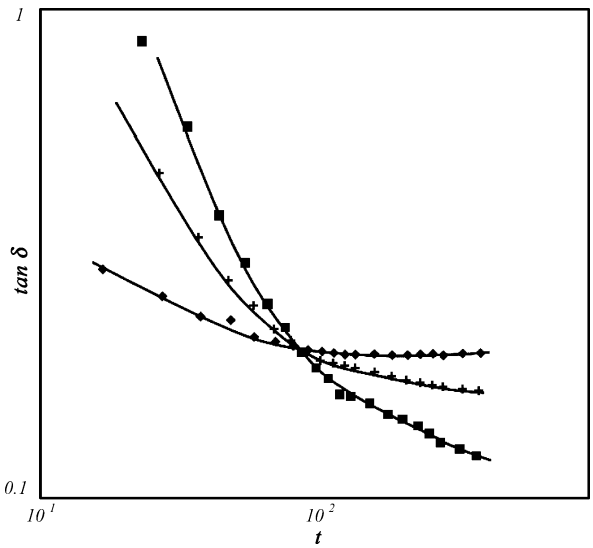

Percolation and Polymer Morphology and Rheology, Figure 5
Determination of gel point from data for loss angle δ. Time is in minutes. The data are for frequencies 31.6 rad s^{-1} (*diamonds*), 1.0 rad s^{-1} (+) and 0.0316 rad s^{-1} (*squares*) (after Lin et al. [45])

tan δ versus time at different frequencies, the intersection of all the curves should be at the GP. Figure 5 demonstrates how this method is used for locating the GP for a physical gel.

Resistor and Elastic Percolation Networks

Experimental data for the scaling properties of the elasticity and viscosity of gels are usually compared with those of the conductivity and elasticity of percolation networks. Thus, we first briefly summarize percolation models of the conductivity of a percolation network. Consider a two-component network in which each (randomly-selected) bond has a conductance g_1 with probability p or a conductance g_2 with probability $1-p$. The limit in which $g_2 = 0$ and g_1 is finite corresponds to a conductor-insulator mixture. As $p \to p_c$, more and more bonds are insulating, the conduction paths become very tortuous and, therefore, the effective conductivity g_{eff} of the network decreases. At p_c one has, $g_{\text{eff}}(p_c) = 0$, since no sample-spanning conduction path exists any more. In the critical region near p_c the effective conductivity follows a power laws:

$$g_{\text{eff}}(p) \sim (p-p_c)^t \quad \text{conductor-insulator networks.} \quad (41)$$

The limit in which $g_1 = \infty$ and g_2 is finite a represents a *conductor-superconductor* mixture. All quantum-mechanical aspects of real superconductors are ignored in this definition, as we are concerned only with the effect of the local connectivity of the material on g_{eff}. It should be clear that the effective conductivity g_{eff} of the network is dominated by the superconducting bonds. If $p < p_c$, then a sample-spanning cluster of the superconducting bonds does not exist, and σ_{eff} is finite. As $p \to p_c^-$, g_{eff} increases until a sample-spanning cluster of the superconducting bonds is formed for the first time at $p = p_c$, where g_{eff} *diverges*. As $p \to p_c^-$, the effective conductivity follows a power law in the critical region:

$$g_{\text{eff}}(p) \sim (p_c - p)^{-s} \quad \text{conductor-superconductor networks .} \quad (42)$$

The exponents t and s are mostly universal. The article by Balberg discusses the conditions under which they may be non-universal.

In a similar manner, the elastic moduli of a two-phase percolation network are defined. Consider a two-component network in which each bond is an elastic element (a spring or beam) with an elastic constant e_1 with probability p or e_2 with probability $1-p$. The limit in which $e_2 = 0$ and e_1 is finite corresponds to a mixture made of rigid materials and holes (for example, porous solids), or rigid and liquid materials. In such networks, as $p \to p_c$, an increasingly larger fraction of the bonds have no rigidity, the paths for transmission of stress or elastic forces become very tortuous and, therefore, the effective elastic moduli G_{eff} (Young's, bulk, or shear moduli) decrease. We refer to this model as the elastic percolation network. At p_c one has, $G_{\text{eff}}(p_c) = 0$, while near p_c in the critical region,

$$G_{\text{eff}}(p) \sim (p-p_c)^T \quad \text{rigid-soft two-phase networks.} \quad (43)$$

The limit in which $e_1 = \infty$ and e_2 is finite represents mixtures of rigid-superrigid materials. We refer to the model as the superelastic percolation network. In this case the effective elastic moduli G_{eff} of the network are dominated by the superrigid bonds. If $p < p_c$, then a sample-spanning cluster of the superrigid bonds cannot form, and G_{eff} is finite. As $p \to p_c^-$, the effective elastic moduli increase until the percolation threshold p_c of the rigid phase is reached, at which a sample-spanning cluster of the superrigid bonds is formed for the first time, and the effective elastic moduli *diverge*. In the critical region near p_c one has

$$G_{\text{eff}}(p) \sim (p_c - p)^{-S} \quad \text{rigid-superrigid networks.} \quad (44)$$

Unlike the exponents t and s for resistors networks, the exponents T and S may depend on the details and the type of the forces that are active in the networks. Thus, in what

follows, we first briefly describe various types of forces that have been included in such networks.

The Born Model

The Born model [17] is described by the following elastic Hamiltonian,

$$\mathcal{H} = \frac{1+\mathcal{P}}{1-\mathcal{P}} \sum_{ij} \mu[(\mathbf{u}_i - \mathbf{u}_j)\cdot \mathbf{R}_{ij}]^2 + \frac{1-3\mathcal{P}}{4(1-\mathcal{P})} \sum_{ij} (\mathbf{u}_i - \mathbf{u}_j)^2 , \quad (45)$$

where $\mathcal{P}$ is the Poisson's ratio, μ the shear modulus, $\mathbf{u}_i$ the displacement of site i, and $\mathbf{R}_{ij}$ the unit vector along the line (lattice bond) that connects sites i and j. The first term of Eq. (45) is the energy of a network of central-force (CF) springs, i. e., Hookean springs that transmit force only in the $\mathbf{R}ij$ direction, but do not transmit shear forces. The second term is a contribution analogous to, for example, the power dissipated in conduction, since $(\mathbf{u}_i - \mathbf{u}_j)^2$ represents the *magnitude* of the displacement difference $\mathbf{u}_i - \mathbf{u}_j$. The Born model may be considered as an analogue of a 3D solid in plane-stress with holes normal to the $x-y$ plane, or as a 2D solid with the Poisson's ratio defined as the negative of ratio of the strain in the $y-$direction to that in the $x-$direction, when a stress is applied in the $x-$direction but none is applied in the $y-$direction. Results for a 3D solid in plane-strain can be generated from those of this model using the transformation, $\mathcal{P}' = \mathcal{P}/(1+\mathcal{P})$, where $\mathcal{P}'$ is the Poisson's ratio for the plain strain.

The Born model does suffer from some pecularities. For example, it is not difficult to show (although it may not be obvious at first glance) that, except for $\mathcal{P} = 1/3$, the elastic energy $\mathcal{H}$ defined by Eq. (45) is *not* invariant with respect to arbitrary rigid body rotations, a fundamental requirement for any reasonable model of elastic properties of materials. In the limit $\mathcal{P} = 1/3$ the model reduces to a network of CF springs. When the elastic energy of a system is written in terms of an expansion in the displacement field $\mathbf{u}$, its rotational invariance is not easy to see. To demonstrate the lack of rotational invariance of the elastic energy, one substitutes an infinitesimal rotation $\boldsymbol{\omega} \times \mathbf{R}_i$ for the displacement vector $\mathbf{u}_i$, where $\mathbf{R}_i$ is the position vector of i. An arbitrary rotation of the solid should not contribute to its energy, but Eq. (45) indicates that, while the contribution of the CF part would indeed be zero, that of the scalar-like part would not be and, therefore, $\mathcal{H}$ is not rotationally invariant. Moreover, although materials do exist that have a Poisson's ratio as high as 1/2 ($\mathcal{P}$ can theoretically be as high as 1 in 2D materials [58]), the model fails to have a strictly positive energy for $\mathcal{P} > 1/3$ and, therefore, violates the thermodynamic requirement that the potential energy be a minimum at zero strain. Another example of displacements that contribute to the scalar-like portion of the energy $\mathcal{H}$ of the model, but not to the CF portion, arises when a significant fraction of the lattice's bonds is removed, i. e., a percolation network is generated. In such a lattice a site that is connected to only one bond can have an arbitrary displacement in the direction orthogonal to the direction of the bond without affecting the CF part of the elastic energy, as can a site which is connected to only a set of two collinear bonds.

In his original formulation of this model, Born [17] inserted the scalar-like part of the elastic energy (45) as a substitute for the many-body, angular and bending terms (see below) that normally arise in describing the elastic properties of materials, because the expansion of such scalar two-body terms is much simpler and more convenient than expanding the many-body terms that they replace. When viewed in this way, the coefficients of the model should be treated as fitting parameters. Hence, let us rewrite Eq. (45) as

$$\mathcal{H} = \tfrac{1}{2}\alpha_1 \sum_{ij} [(\mathbf{u}_i - \mathbf{u}_j)\cdot \mathbf{R}_{ij}]^2 + \tfrac{1}{2}\alpha_2 \sum_{ij} (\mathbf{u}_i - \mathbf{u}_j)^2 , \quad (46)$$

where α_1 and α_2 now represent two adjustable parameters. When introduced in this context, one may use the Born model for modeling and fitting elastic properties of certain materials.

Note that, so long as $\alpha_2 > 0$, the scalar-like term of Eq. (45) or (46) is the dominating contributing factor to the elastic energy $\mathcal{H}$. This implies immediately that, although the Born model is a vector model, the behavior of the elastic moduli in this model near the percolation threshold is effectively like that of a scalar (conductivity) model and, therefore,

(i) the percolation threshold of the Born model, at which the elastic moduli vanish or diverge, is the same as that of random percolation models, and

(ii) near the percolation threshold the elastic moduli of the Born model follow power law (43) or (44), but with $T = t$ and $S = s$. That is, the power-law behavior of the elastic moduli in the Born model is the same as that of the effective conductivity.

The Central-Force Model

Consider the limit $\mathcal{P} = 1/3$ of Eq. (45), i. e., a network of CF or Hookean springs. Since the elastic materials that we wish to consider are heterogeneous, the local shear modu-

lus μ varies spatially. Thus, writing $\mu = e_{ij}\alpha/4$ and taking $\mathcal{P} = 1/3$ reduce Eq. (45) to

$$\mathcal{H} = \tfrac{1}{2}\alpha \sum_{ij} [(\mathbf{u}_i - \mathbf{u}_j) \cdot \mathbf{R}_{ij}]^2 e_{ij} \tag{47}$$

where α is the CF constant.

We can easily compute the elastic moduli of the CF networks, if no percolation effect is present, i. e., if no bond is broken and all the e_{ij} are equal. Suppose that each spring has an unstretched length ℓ_0. Then, it is not difficult to show that the bulk modulus K of a triangular network is given by [58]

$$K = \tfrac{\sqrt{3}}{2}\alpha \quad \text{triangular network}\,, \tag{48}$$

whereas its shear modulus is given by [58]

$$\mu = \tfrac{\sqrt{3}}{4}\alpha \quad \text{triangular network,} \tag{49}$$

and, therefore, the Poisson's ratio of the network is, $\mathcal{P} = (K_e - \mu_e)/(K_e + \mu_e) = 1/3$. Similarly, we may show that [58],

$$K = \mu_\mathrm{p} = \tfrac{1}{2}\alpha \quad \text{square network,} \tag{50}$$

where μ_p is the shear modulus in *pure* shear (the network's shear modulus in *simple* shear is zero). As for the standard 3D cubic networks, one has [58]

$$K_e = \begin{cases} \frac{1}{3\ell_0}\alpha & \text{simple-cubic network}\,, \\ \frac{1}{\sqrt{3}\ell_0}\alpha & \text{BCC network}\,, \\ \frac{2\sqrt{2}}{\sqrt{3}\ell_0}\alpha & \text{FCC network.} \end{cases} \tag{51}$$

A simple-cubic network does not possess a shear modulus in simple shear. We emphasize that Eqs. (48)–(51) are valid at zero temperature and when the external stress is infinitesimally small.

In practice, however, all the experimental measurements are carried out at temperatures above $T = 0$ and, therefore, it is important to understand the temperature-dependence of the elastic moduli, at least in the context of the network models. In addition, in many practical situations, the material under study is exposed to a finite stress or tension (as opposed to an infinitesimal stress or tension considered above) and, thus, the role of such an external driving force in determining the elastic properties of materials must be understood. In principle, the role of the temperature can be understood by carrying out molecular dynamics simulations [54]. However, phenomenological calculations of the type described above can also be carried out for homogeneous networks at non-zero temperature. Suppose, then, that a 2D isotropic tension σ_i is imposed on a network at a non-zero temperature. One can show that for the triangular network [58]

$$K = \tfrac{1}{2}(\sqrt{3}\,\alpha - \sigma_i)\,, \tag{52}$$

$$\mu = \tfrac{\sqrt{3}}{4}(\alpha + \sqrt{3}\,\sigma_i)\,, \tag{53}$$

which reduce to Eqs. (48) and (49) in the limit $\sigma_i = 0$. Similar results are obtained for the square network [58]:

$$K = \tfrac{1}{2}(\alpha - \sigma_i)\,, \tag{54}$$

$$\mu_\mathrm{p} = \tfrac{1}{2}(\alpha + \sigma_i)\,, \tag{55}$$

$$\mu_\mathrm{s} = \sigma_i\,, \tag{56}$$

where μ_s is the shear modulus of the network in simple shear.

Rigidity Percolation

If the elastic constants e_{ij} of the bonds of a CF network take on either a finite value with probability p or vanish with probability $1 - p$, then one obtains an elastic percolation network. If e_{ij} is infinitely large with probability p or takes on a finite value with probability $1 - p$, then, one obtains a superelastic percolation network. Percolation on such networks of Hookean springs is called the *rigidity percolation*. Such networks are of both theoretical and practical interest. In addition to the polymeric materials described in this chapter, they are also useful models for describing the elastic properties of biological materials. Moreover, in many engineering problems, structures composed of bars or beams connected at nodes that are called trusses acquire their rigidity mainly from the tensile and compressive stiffness of the beams, and these are CF type of contributions. For example, in the absence of friction between the particles of a granular packing, which is a reasonable model of unconsolidated porous materials (such as powders), the mechanical behavior of the packing is similar to those of rigidity percolation. In contrast, those in which angular forces, e. g., covalent bonds at the molecular level, are the most important are usually referred to as *frames*. It is not difficult to see that rigid systems in which angular forces dominate their behavior support an applied stress, so long as they are simply connected. In contrast, the CF systems require higher degrees of connectivity. Therefore, the percolation thresholds of CF networks are much larger than those of random percolation networks; see the articles by Duxbury and Thorpe.

The Bond-Bending Model

Consider an elastic percolation network in which there are both central and bond-bending (angle-changing) forces, with the latter type representing the three-body interactions. One of the main advantages of such a model is that their percolation threshold can be the same as that of random percolation, if the many-body interactions are such that any deformation of the lattice is done at some costs to its elastic energy. In general, the elastic energy of such models is given by [40]

$$\mathcal{H} = \tfrac{1}{2}\alpha \sum_{\langle ij\rangle} [(\mathbf{u}_i - \mathbf{u}_j)\cdot \mathbf{R}_{ij}]^2 e_{ij} + \tfrac{1}{2}\chi \sum_{\langle jik\rangle} (\boldsymbol{\delta\theta}_{jik})^2 e_{ij} e_{ik} , \tag{57}$$

where α and χ are, respectively, the central and bond-bending (BB) force constants. Here, $\langle jik\rangle$ indicates that the sum is over all triplets in which the bonds j-i and i-k form an angle with its vertex at i. The first term on the right-hand side of Eq. (60) represents the usual CF contributions (see above), while the second term is due to the BB (angle-changing) forces. The precise form of $\boldsymbol{\delta\theta}_{jik}$ depends on the microscopic details of the model. In the version that is of interest to us, bending of the collinear bonds is allowed, in which case [13,66]

$$\boldsymbol{\delta\theta}_{jik} = \begin{cases} (\mathbf{u}_{ij} \times \mathbf{R}_{ij} - \mathbf{u}_{ik} \times \mathbf{R}_{ik}) \cdot \frac{(\mathbf{R}_{ij}\times\mathbf{R}_{ik})}{|\mathbf{R}_{ij}\times\mathbf{R}_{ik}|} , & \\ \qquad\qquad\qquad\qquad \mathbf{R}_{ij} \text{ not parallel to } \mathbf{R}_{ik} , \\ |(\mathbf{u}_{ij} + \mathbf{u}_{ik}) \times \mathbf{R}_{ij}| , & \mathbf{R}_{ij} \text{ parallel to } \mathbf{R}_{ik} , \end{cases} \tag{58}$$

where, $\mathbf{u}_{ij} = \mathbf{u}_i - \mathbf{u}_j$. For *all* 2D networks, Eq. (61) is simplified to

$$\boldsymbol{\delta\theta}_{jik} = (\mathbf{u}_i - \mathbf{u}_j) \times \mathbf{R}_{ij} - (\mathbf{u}_i - \mathbf{u}_k) \times \mathbf{R}_{ik} . \tag{59}$$

We refer to the model described by Eqs. (60)–(62) as the BB model. For most materials to which the BB model is applicable, one has, $\chi/\alpha \leq 0.3$ [49]. Sahimi [57] suggested that the critical exponent T of the elasticity in the BB model is related to t, the critical exponent of conductivity of percolation networks:

$$T = t + 2\nu , \tag{60}$$

where ν is the correlation length exponent of percolation. This relation is in excellent agreement with the available numerical estimates (see Table 1 below). The articles by Duxbury and Thorpe provide much more detailed discussions of the CF and BB models and, therefore, we do not discuss them any further.

Before embarking on a comparison between the experimental data for the rheological properties of gels near the GP, for convenience and as a basis for comparison with the experimental data, we summarize in Table 1 the current most accurate estimates of the various critical exponents for the conductivity and elasticity of percolation models near the percolation thresholds, including the CF and BB models.

We are now in a position to compare the predictions of the percolation models with the experimental data for the viscosity of the nearly critical sol, and the elastic moduli of the nearly critical gels.

Nearly-Critical Chemical Gels: Comparison of the Data with the Percolation Models

There are numerous experimental measurements of the elastic moduli of nearly-critical chemical gels and the associated exponent z. Examples include the measurements for hydrolyzed polyacrylamide [9,11], tetraethylorthosilicate reactions [38,64], gelatin solutions [28], polycondensation of polyoxypropylated trimethylolpropane with hexamethylenediisocyanate [29], and several other measurements [2,30,35,37,65]. These measurements yielded esti-

Percolation and Polymer Morphology and Rheology, Table 1

Estimates of the critical exponents of the conductivity and elastic moduli of percolation models in *d*-dimensions. Values of *T* and *S* for the CF model refer to bond percolation, whereas those of *t* and *s*, the conductivity exponents, are independent of the model. *ν* is the critical exponent of the percolation correlation length. Value of *ν* for the CF model is different from that of random percolation, whereas it is the same as that of random percolation for the BB model

d	t/ν	s/ν	T/ν	S/ν	Model
2	0.9745 ± 0.0015	0.9745 ± 0.0015	2.97 ± 0.03	0.92 ± 0.03	Bond bending
	—	—	2.95 ± 0.25	0.92 ± 0.02	Central force
3	2.27 ± 0.01	0.835 ± 0.005	4.3 ± 0.1	0.74 ± 0.04	Bond bending
			2.1 ± 0.1	0.80 ± 0.03	Central force

mates of z that are in the range 1.9–2.4, which do not agree with the value of the critical exponent T for the 3D BB model (see Table 1). In fact, if the size of a chemical gel network is large enough, the BB forces may not play any important role in determining the elastic properties of nearly critical gels and, therefore, the only important forces between the monomers are the central (stretching) forces. Therefore, these experimental data may be explained based on the elasticity exponent in the 3D CF percolation [14], $T \simeq 2.1 \pm 0.2$.

However, a value of z in the range 1.9–2.4 may also be interpreted in terms of two other models:

(i) As mentioned earlier, de Gennes [72] suggested that the scaling properties of the elastic moduli of nearly critical gels are in the universality class of the conductivity of percolation networks, implying that, $z = t \simeq 2.0$.

(ii) On the other hand, Alexander [7] argued that in some gels and rubbers that are under internal or external stresses, there are terms in their elastic energy that are similar to the Born model. As discussed above, the critical exponent of the elastic moduli in the Born model is equal to that of the conductivity, t, and in particular in 3D, $T = t \simeq 2.0$, because near the percolation threshold the contribution of the second term of the right side of Eq. (45) or (46), which is a purely scalar term, dominates that of the first term which is due to the CFs.

While the data mentioned above are more or less consistent with de Gennes' hypothesis, most of them are not precise enough to distinguish between $t \simeq 2.0$ for 3D percolation conductivity and $T \simeq 2.1$ for the CF percolation. However, there are also a few relatively precise sets of experimental data that seem to support de Gennes' conjecture. For example, Axelos and Kolb [15] measured the rheological properties of pectin biopolymers that consist of randomly connected $\alpha(1-4)$D-galacturonic acid units and their methyl esters. If the methyl ester content is low, pectin forms thermoreversible gels upon addition of cations, such as calcium. Axelos and Kolb [15] measured the frequency dependence of the storage modulus $G'(\omega)$ and loss modulus $G''(\omega)$ [see Eq. (33)] and reported that, $z \simeq 1.93$, $k \simeq 0.82$, and $n \simeq 0.71$. Their elasticity exponent is close to that of the conductivity, $t \simeq 2.0$, for 3D resistor networks. Less precise data, but still supportive of de Gennes' proposal, were reported by Adam et al. [5] for the complex modulus of end-linked poly(dimethylsiloxane) pregel polymer clusters, quenched at different distances from the gelation threshold. They reported that, $z \simeq 1.9 \pm 0.15$, consistent with the value of the conductivity exponent t for 3D percolation. However, the estimated error is large enough that one can easily interpret such a value of z in terms of the CF percolation model as well.

At first glance, de Gennes' proposal that the critical exponent of the effective moduli of gels, a vector transport property, should be equal to that of a scalar property, the effective electrical conductivity of a resistor network, may seem incorrect. However, to justify his proposal, de Gennes introduced the notion of an elastic chain between neighboring nodes or monomers that are the analogue of quasi-one-dimensional strands that percolation clusters possess near the percolation threshold (see the article by Coniglio). He then argued that if such chains are elongated, then their nodes carry an extra amount of energy. If we assume that the blobs – the multiply-connected parts – of the large cluster of monomers do not contribute significantly to the elastic moduli, then one must only consider the energetics of the links or the chains. If the extra energy of such chains is larger than $k_B T$ (k_B is the Boltzmann's constant), then, as Daoud [24] argued, one obtains de Gennes' proposal [72], $t = z$, although Daoud's analysis was a mean-field approximation, not a scaling one.

As for Alexander's proposal [7], rubbers and gels differ from the Born model in several important ways, such as the presence of non-linear terms in their elastic energy, and the possibility of negative as well as positive Born coefficients α_1 and α_2 in Eq. (46). Therefore, as discussed above, while one may use the Born model to *fit* the experimental data, it is not clear that, at a fundamental level, the Born model can actually describe the elastic properties of such gels, since its elastic energy is not rotationally invariant.

Enthalpic Versus Entropic Elasticity

There is yet another way of rationalizing the experimental data for the scaling properties of nearly critical gels. To describe it, we mention that several measurements of the elastic moduli of nearly critical gel and the associated exponent z deviate significantly from all the data described above. Examples include the measurements of Adam et al. [1] for polycondensation, $z \simeq 3.3 \pm 0.5$, those of Martin et al. [47] and Adolf et al. [6] for gels made from 89% (by weight) of the diglycidyl ether of bisphenol A cured with 11% (by weight) of diethanolamine which yielded, $z \simeq 3.3 \pm 0.3$, and the data reported by Colby et al. [22] for polyester gels, which have been argued to lie in the middle of the static crossover between the Flory–Stockmayer (Bethe lattice) model, which predicts that, $z = 3$, and the 3D percolation model. Colby et al. [22] reported that, $z \simeq 3.0 \pm 0.7$, which is inconsistent with both the CF percolation and the BB models, although one

might argue that their estimate is agreement with $z = 3$, the Flory–Stockmayer value of z. More recent measurements of the shear modulus of an end-linking polymer gel network by Takahashi et al. [64] yielded, $z \simeq 2.7$, which is again in the range of the above data.

One possible explanation for such data is that the elasticity of such gels is entropic rather than enthalpic. Plischke and Joós [54], Plischke et al. [55], Farago and Kantor [32], and Plischke [52,53] argued that the CF and the BB models, described above, are applicable to gels at temperature $T = 0$, and that for $T \neq 0$ there is an important contribution to the shear modulus that is entropic in nature. In analogy with the physics of rubber elasticity, Plischke et al. [55] argued that, near the percolation threshold or the GP, the polymer network consists essentially of long chains of singly-connected particles (monomers or sites), linked to each other at various junction points. Such chains are similar to the polymer chains that are crosslinked in rubber in order to produce a rigid amorphous material. Deformation of the sample changes the distance between the junctions points or crosslinks, as a result of which the entropy is generically decreased, resulting in an increase of free energy and, hence, a restoring force. There would be a net shear-restoring force, when the nearly critical gel is formed, implying that the connecting chain of particles acts as a stretched spring. Molecular dynamics simulations (for a review see, for example, [42]) of Plischke and co-workers, and computer simulations of Farago and Kantor [32], who used a model that consisted of hard spheres in which a fraction p of the neighbors are tethered by inextensible bonds, both yielded, $z \simeq t$.

On the other hand, del Gado et al. [74] proposed a different model, also purported to be appropriate for entropic gels, in which one begins with a random collection of monomers with concentration p. Each pair of the monomers are then linked with a probability p_b to form permanent bonds. Varying p_b produces a distribution of clusters of the linked monomers and, hence, gives rise to the possibility of forming a sample-spanning cluster. Monomers and the clusters then diffuse according to the bond fluctuation algorithm of Carmesin and Kremer [21]. In this algorithm, the monomers diffuse in the solution randomly but obeying the excluded-volume interaction (i. e., no two monomers can occupy the same point in space). Due to this random motion, the bonds may have to change their length in a set of allowed values and, thus, they may have to bend and take on many different values of the angles between the bonds, which then gives rise to a wide variety of polymer conformations. The mean-square displacement $\langle R^2(t)\rangle$ of the polymer's center of mass is then calculated which, due to the elastic potential that reduces the fluctuations proportionally to the effective elastic constant α, is given by

$$\langle R^2(t)\rangle \propto \alpha^{-1} , \tag{61}$$

and, therefore, the elastic constant and its power-law behavior near the percolation threshold p_c can be computed, from which the exponent z is extracted. Two-dimensional simulations of del Gado et al. [74] yielded the estimate, $z \simeq 2.7 \pm 0.1$, hence disagreeing with the results of Plischke and co-workers, and Farago and Kantor.

How can one interpret these results? If the elasticity of these gels is due to entropic effects, then, as Daoud and Coniglio [25] argued (see also Martin et al. [47]), the elastic free energy $\mathcal{H}$ per unit volume must be given by

$$\mathcal{H} \sim \xi_p^{-d} \alpha_\ell \xi_p^2 , \tag{62}$$

where ξ_p is the correlation length of percolation, and α_ℓ is the effective elastic constant of a long chain of length ξ_p connecting two nodes. Since $\alpha_\ell \sim \xi_p^{-2}$, we obtain

$$z = \nu d . \tag{63}$$

Since for 2D percolation, $\nu = 4/3$, Eq. (66) predicts that, $z = 8/3 \simeq 2.66$, quite different from the exponents T of the CF and BB models, and also the conductivity exponent, $t \simeq 1.3$ (see Table 1). However, it is in agreement with the numerical simulations of del Gado et al. [74]. Daoud [24] argued that Eq. (66) is valid when the energy of the chains is of the same order of magnitude as the thermal energy $k_B T$.

Since, $\nu \simeq 0.89$ for 3D percolation, Eq. (66) predicts that, $z \simeq 2.67$, consistent with the experimental data of Adam et al. [1], Martin et al. [47], Adolf et al. [6], Colby et al. [22] and Takahashi et al. [64] mentioned above, all of whom reported estimates of the elasticity exponent z in the range 2.7–3.3±0.5. Therefore, while these experimental data may be explained by Eq. (66), the numerical results of Plischke and co-workers [52,53,54,55], as well as those of Farago and Kantor [32], do not agree with the prediction of Eq. (66). Indeed, although the main argument of Plischke et al. [55] was that, the entropic effects are important at temperatures $T \neq 0$, where one should see a crossover to the value of the conductivity exponent, all of the above experiments were also carried out at finite temperatures, yet they did not indicate that, $z \simeq t$.

More recently, Xing et al. [69] studied the scaling of shear modulus near the gelation-vulcanization transition. They proposed that in a dense melt the sizes of the effective chains of the critical gel scale *sublinearly* with their counter length. The implication is that, the energy

that each chain contributes is of the order of $k_B T$, hence leading to Eq. (66). However, in *phantom* networks – those in which there is no repulsion between the particles (monomers) – the chains' sizes scale *linearly* with their contour length, which means that the elasticity exponent z crosses over to the conductivity exponent t. Thus, it may be that some of the conflicts between the various experimental estimates of z are due to the crossover effects, but the issue remains unsolved.

Viscosity of Nearly Critical Sol: Comparison of the Data with the Percolation Models

There is also a wealth of experimental data for the viscosity of the nearly critical sol solution and the associated critical exponent k, defined by Eq. (28). An important question that has been studied for over two decades is: how can one explain the extensive experimental data for the scaling behavior of the viscosity η of the sol phase near the GP? To begin with, it was proposed by Sahimi and Goddard [60] (see also [12,59]) that the power-law behavior of η near the GP is analogous to that of the shear modulus of a superelastic percolation network near p_c and, therefore, the same critical exponent characterizes both. To proceed further, we must first establish a rigorous relationship between the linear elasticity and the theory of viscous fluids, thus confirming the proposal of Sahimi and Goddard [60].

We consider a general time-dependent system and write down the equation of motion for a macroscopically-homogeneous material in terms of the displacement field $\mathbf{u}$:

$$(\lambda + \mu)\nabla(\nabla \cdot \mathbf{u}) + \mu\nabla^2\mathbf{u} + \mathbf{F} = \rho\frac{\partial^2\mathbf{u}}{\partial t^2} , \tag{64}$$

where ρ is the mass density, t the time, λ and μ the usual Lamé coefficients, and $\mathbf{F}$ an external force. For an *incompressible* material, i. e., one for which the bulk modulus K and the Lamé coefficient λ are both divergent, one has the solenoidal condition,

$$\nabla \cdot \mathbf{u} = 0 . \tag{65}$$

Due to the incompressibility condition, the first term of Eq. (67) is indeterminate. Equation (67) can be then written in terms of the reactive hydrostatic pressure P,

$$-\nabla P + \mu\nabla^2\mathbf{u} + \mathbf{F} = \rho\frac{\partial^2\mathbf{u}}{\partial t^2} . \tag{66}$$

On the other hand, let us write down the Navier–Stokes equations of motion for an incompressible and Newtonian viscous fluid,

$$-\nabla P + \eta\nabla^2\mathbf{u} + \mathbf{F} = \rho\left(\frac{\partial\mathbf{v}}{\partial t} + \mathbf{v}\cdot\nabla\mathbf{v}\right) , \tag{67}$$

where $\mathbf{v}$ is the fluid velocity field, and η the fluid's dynamic viscosity. For an incompressible fluid, the continuity equation is given by

$$\nabla \cdot \mathbf{v} = 0 , \tag{68}$$

which is similar to Eq. (68). For slow fluid flow, i. e., when the Reynolds number $\mathrm{Re} \ll 1$, the non-linear inertial term, $\mathbf{v}\cdot\nabla\mathbf{v}$, is very small and can be neglected, which means that the Navier–Stokes equations reduce to

$$-\nabla P + \eta\nabla^2\mathbf{v} + \mathbf{F} = \rho\frac{\partial\mathbf{v}}{\partial t} . \tag{69}$$

Thus, we see that, under steady-state condition, and when the flow of the fluid is slow, the governing equations for the displacement field $\mathbf{u}$ and the velocity field, $\mathbf{v} = \partial\mathbf{u}/\partial t$, are exactly identical, provided that there is a one-to-one correspondence between the shear modulus μ and the dynamic viscocity η.

In addition, under certain conditions, the effective viscosity η of a suspension of completely rigid spheres in creeping (very slow) flow of an incompressible fluid of viscosity η_1 is related to the *steady-state* effective shear modulus μ of a two-phase material composed of the same completely rigid spheres in an incompressible matrix with shear modulus μ_1. In this case, the working equation is given by

$$\frac{\eta_e}{\eta_1} = \frac{\mu_e}{\mu_1} . \tag{70}$$

Equation (70) is exact when, regardless of the configuration of the particles, hydrodynamic interactions between the particles can be neglected, which is the case when the system is infinitely dilute so that the volume fraction of the particles approaches zero. Even if the system is non-dilute, Eq. (70) would still be exact, provided that the configurations of the particles in the flow and the elasticity problems are identical.

Having established a theoretical connection between the viscosity η of a sol and the shear modulus of an appropriate two-phase material, the one-to-one correspondence between η and the shear modulus of a superelastic percolation network should be clear because,

(i) η and μ both diverge at p_c (the GP), and

(ii) the percolation models predict accurately the structure and elastic moduli of nearly critical gels.

On the other hand, de Gennes [73] (see also Allain et al. [8]) suggested an analogy between η and the effective conductivity of a conductor-superconductor percolation networks, the effective conductivity of which diverges at p_c according to power law (42), so that $k = s$.

However, even a one-to-one correspondence between η and the shear modulus of a superelastic percolation network is not nearly enough to explain the power law behavior of the viscosity of gelling solutions near the GP, because most of the experimental data indicate that the value of k is either in the range 0.6–0.9 (see, for example, [1,4,10,29]), or in the range 1.3–1.5 (see, for example, [28,47,48]), whereas the power-law behavior of the shear modulus of a 3D superelastic percolation network near p_c is characterized by a *unique* value of the exponent S defined by Eq. (44). The reason for having two distinct values of k is [12,59] that the dynamics of the sol solutions, that yield the two distinct values of k, may be completely different.

In one case, the solution may be close to the Zimm limit in which there is little or no polymer diffusion in the reaction bath, because there are strong hydrodynamic interactions between the monomers, and also between polymers of various sizes, that prevent diffusion in a nearly critical sol near the GP. Hence, a superelastic percolation network – a *static* system with fixed rigid clusters – may be suitable for simulating the Zimm regime. For this limit, Arbabi and Sahimi [12] suggested that,

$$S = \nu - \tfrac{1}{2}\beta \,, \tag{71}$$

where ν and β are the standard percolation exponents. Then, with $\nu \simeq 0.89$ and $\beta \simeq 0.41$ for 3D percolation, one obtains, $S \simeq 0.68$ for the divergence of the shear modulus of 3D superelastic percolation networks at the percolation threshold. Such a value of S seems to explain the estimates of k for those gelling solutions that have a viscosity exponent in the range 0.6–0.9, implying that the scaling behavior of their viscosity near the GP may be described by the shear modulus of a *static* superelastic percolation network.

On the other hand, the gelling solution may also be in, or near, the Rouse regime – one in which there are no hydrodynamic interactions between the polymers of various sizes – and, therefore, the finite polymers can diffuse essentially freely in the reaction bath. To simulate this regime the following model was proposed [12,59]. We consider a superelastic percolation network in which every cluster of the totally rigid bonds (the bonds in such clusters are totally rigid in order to distinguish them from those with a finite elastic constant) represents a finite polymer. Due to randomness of percolation networks, there is, of course, a wide distribution of such polymers or clusters in the network. The "soft" bonds – those with a finite elastic constant – represent the liquid solution in which the rigid clusters move randomly, with equal probability, in any direction of the network. The motion simulates diffusion of the finite polymers in the reaction bath. Two rigid clusters cannot overlap, but can temporarily join and form a larger cluster, which may also be broken up again at a later time. Thus, at every time step, a cluster and a direction for the motion are picked at random, and the cluster is moved by one step (one lattice bond) in that direction, taking into account the excluded-volume effect. This is then a *dynamic* superelastic percolation network, the shear modulus of which can be calculated at long times. Monte Carlo simulations [12,59] indicated that the shear modulus of such a dynamic superelastic percolation network diverges with an exponent S', which Arbabi and Sahimi [12] proposed it to be given by

$$S' = 2\nu - \beta \,. \tag{72}$$

Equation (72) predicts that in 3D, $S' \simeq 1.35$, which appears to explain the viscosity exponent for those gelling solutions that have an exponent k in the range 1.3–1.5, hence implying that the scaling behavior of their viscosity near the GP may be described by the shear modulus of a *dynamic* superelastic percolation network.

Daoud [24] also argued that, similar to the case of the elastic moduli of nearly critical gels described above, one must consider two distinct regimes for explaining the power-law behavior of the viscosity near the GP, except that his arguments were based on the energetics of the solution. According to him, if the elastic chains carry an energy which is of the same order of magnitude as the thermal energy $k_B T$, then the exponent S should be given by Eq. (71). On the other hand, if the elastic chains are stretched and have an extra energy larger than $k_B T$, then one should recover Eq. (72), which had also been conjectured by de Gennes [73], but based on the analogy between the viscosity and the effective conductivity of superconducting percolation networks described above.

We should point out that, as in the case of the elastic moduli of nearly critical gels discussed earlier, there are some experimental data that indicate some deviations of k from S or S'. However, as pointed out earlier, experimental determination of k (and the elasticity exponent z) involves (see, for example, [29]) measuring the complex modulus $G^*(\omega)$ for a series of frequencies ω. But, strictly speaking, the power laws for the elastic moduli of the elastic and superelastic percolation networks are valid only in the limit, $\omega \to 0$, whereas in practice it is essentially impossible to reach such a limit and, therefore, the measured values of k may exhibit some deviations from S or S'. Thus, such deviations are probably due to transient effects that should diminish as very low frequencies are accessed. Let us also mention that, Bergman [16] suggested that, $S = s$, but his suggestion is not currently supported by the estimates of the two exponents listed in Table 1.

Physical Gels: Comparison of the Data with the Percolation Model

In physical gels both inter- and intramolecular bondings are non-covalent. The presence of non-covalent bonding means that their numbers and positions fluctuate with time, as well as temperature, as such bonds are reversible. Moreover, the nature of the (physical) crosslinks is not completely understood. In many cases they involve hydrophobic, hydrogen bonding, and electrostatic interactions, the combination of which make gaining a better understanding of the properties of physical gels a very complex problem. This is particularly true for biopolymers.

Two well-known examples of physical gels are gelation of silica particles in NaCl solutions and in pure water [36], and silica aerogels [68]. As discussed above, the attachment of the particles in such gels is by relatively weak association. The BB forces are important in such gels since touching particles that form long chains, when deformed, roll on top of one another and this motion and the displacement of the centers of any 3 mutually-touching particles create forces that are equivalent to the BB forces.

Despite the many complications, percolation theory seems to be capable of providing rational explanations for the scaling behavior of at least some of the experimental data for the elastic moduli of physical gels near the GP. For example, measurements [36,68] of the elastic moduli of silica gels and aerogels yielded an estimate of the elasticity exponent, $z \simeq 3.8$, which is in excellent agreement with the critical exponent T for the 3D BB model (see Table 1), as well as with Eq. (63).

More recent measurements by Devreux et al. [27] indicate a crossover between the prediction of the BB model and another regime with a much smaller value of z. They measured the complex modulus G^* of silica gels formed by hydrolysis-condensation of a silicon alkoxide. For a restricted region near the GP they reported that, $z \simeq 2.0 \pm 0.1$, close to the exponent T of the CF percolation, whereas beyond this region they found $z \simeq 3.6 \pm 0.1$, which is close to the elasticity exponent for the 3D BB model. Devreux et al. interpreted their data for the region near the GP in terms of an analogy between elastic percolation networks and random resistor networks.

Future Directions

Percolation models explain qualitatively, and in many cases quantitatively, the structure, rheology, and mechanical behavior of branched polymers and gels. In particular, such models provide a rational explanation for the power-law behavior of many properties of such materials near the percolation threshold, which mean-field theories and effective-medium approximations fail to predict.

Despite their success, there are still many issues to be addressed. There is still doubt as to whether percolation models can explain the behavior of many types of nearly critical gels and branched polymers, particularly biopolymers (see, for example, [56]). Several sets of puzzling data on the elastic moduli of critical and nearly critical gels remain to be explained. If there is no unique universality class for the elastic moduli and the viscosity, but there is, instead, a few of them, the crossovers between the various universality classes remain to be clarified. Our understanding of various properties of physical gels is not as extensive as what we now know about chemical gels. These and other issues promise exciting research problems for the foreseeable future.

Bibliography

Primary Literature

1. Adam M, Delsanti M, Durand D (1985) Mechanical measurements in a reaction bath during the polycondensation reaction near the gelation threshold. Macromolecules 18:2285
2. Adam M, Delsanti M, Durand D, Hild G, Munch JP (1981) Mechanical properties near gelation threshold, comparison with classical and 3d percolation theories. Pure Appl Chem 53:1489
3. Adam M, Delsanti M, Munch JP, Durand D (1987) Size and mass determination of clusters obtained by polycondensation near the gelation threshold. J Phys 48:1809
4. Adam M, Delsanti M, Okasha R, Hild G (1979) Viscosity study in the reaction bath of the radical copolymerisation of styrene divinylbenzene. J Phys Lett 40:539
5. Adam M, Lairez D, Karpasas M, Gottlieb M (1997) Static and dynamic properties of cross-linked poly(dimethylsiloxane) pregel clusters. Macromolecules 30:5920
6. Adolf D, Martin JE, Wilcoxon JP (1990) Evolution of structure and viscoelasticity in an epoxy near the sol–gel transition. Macromolecules 23:527
7. Alexander S (1984) Is the elastic energy of amorphous materials rotationally invariant? J Phys 45:1939
8. Allain C, Limat L, Salomé L (1991) Description of the mechanical properties of gelling polymer solutions far from gelation threshold: generalized effective-medium calculations of the superconductor-conductor site percolation problem. Phys Rev A 43:5412
9. Allain C, Salomé L (1987) Hydrolysed polyacrylamide/Cr^3 gelation: Critical behaviour of the rheological properties at the sol–gel transition. Polym Commun 28:109
10. Allain C, Salomé L (1987) Sol-gel transition of hydrolyzed polyacrylamide + chromium III: Rheological behavior versus cross-link concentration. Macromolecules 20:2957
11. Allain C, Salomé L (1990) Gelation of semidilute polymer solutions by ion complexation: Critical behavior of the rheological properties versus cross-link concentration. Macromolecules 23:981
12. Arbabi S, Sahimi M (1990) Critical properties of viscoelasticity of gels and elastic percolation networks. Phys Rev Lett 65:725

13. Arbabi S, Sahimi M (1990) On three-dimensional elastic percolation networks with bond-bending forces. J Phys A 23:2211
14. Arbabi S, Sahimi M (1993) Mechanics of disordered solids. I Percolation on elastic networks with central forces. Phys Rev B 47:695
15. Axelos MAV, Kolb M (1990) Crosslinked biopolymers: Experimental evidence for scalar percolation theory. Phys Rev Lett 64:1457
16. Bergman DJ (2003) Exact relation between critical exponents for elastic stiffness and electrical conductivity of percolating networks. Physica B 338:240
17. Born M, Huang K (1954) Dynamical Theory of Crystal Lattices. Clarendom Press, Oxford
18. Bouchaud E, Delsanti M, Adam M, Daoud M, Durand D (1986) Gelation and percolation: swelling effect. J Phys Lett 47:539
19. Brinker CJ, Scherer GW (1990) The Physics and Chemistry of Sol-Gel Processing Academic, San Diego
20. Candau SJ, Ankrim M, Munch JP, Rempp P, Hild G, Osaka R (1985) Physical Optics of Dynamical Phenomena in Macromolecular Systems. Berlin, De Gruyter, p 145
21. Carmesin I, Kremer K (1988) The bond fluctuation method: A new effective algorithm for the dynamics of polymers in all spatial dimensions. Macromolecules 21:2819
22. Colby RH, Gilmor JR, Rubinstein M (1993) Dynamics of near-critical polymer gels. Phys Rev E 48:3712
23. Daoud M (1988) Distribution of relaxation times near the gelation threshold. J Phys A 21, L973
24. Daoud M (2000) Viscoelasticity near the sol–gel transition. Macromolecules 33:3019
25. Daoud M, Coniglio A (1981) Singular behaviour of the free energy in the sol–gel transition. J Phys A 14:301
26. Daoud M, Family F, Jannik G (1984) Dilution and polydispersity in branched polymers. J Phys Lett 45:199
70. de Gennes (1977) PG Critical behaviour for vulcanization processes. J Phys Lett 38:L355
71. de Gennes PG (1972) Exponents for excluded volume problem as derived by the Wilson method. Phys Lett A 38:339
72. de Gennes PG (1976) On a relation between percolation theory and the elasticity of gels. J Phys Lett 37:L1
73. de Gennes PG (1979) Incoherent scattering near a sol gel transition. J Phys Lett 40:197
74. del Gado E, de Arcangelis LE, Coniglio A (1999) Elastic properties at the sol–gel transition. Europhys Lett 46:288
27. Devreux F, Boilot JP, Chaput F, Malier L, Axelos MAV (1993) Crossover from scalar to vectorial percolation in silica gelation. Phys Rev E 47:2689
28. Djabourov M, Leblond J, Papon P (1988) Gelation of aqueous gelatin solutions. II Rheology of the sol–gel transition. J Phys Fr 49:333
29. Durand D, Delsanti M, Adam M, Luck JM (1987) Frequency dependence of viscoelastic properties of branched polymers near gelation threshold. Europhys Lett 3:297
30. Fadda GC, Lairez D, Pelta J (2001) Critical behavior of gelation probed by the dynamics of latex spheres. Phys Rev E 63:061405
31. Family F, Coniglio A (1980) Crossover from percolation to random animals and compact clusters. J Phys A 13:L403
32. Farago O, Kantor Y (2000) Entropic elasticity of two-dimensional self-avoiding percolation systems. Phys Rev Lett 85:2533
33. Feng S, Thorpe MF, Garboczi E (1985) Effective-medium theory of percolation on central-force elastic networks. Phys Rev B 31:276
34. Flory PJ (1941) Molecular size distribution in three dimensional polymers. Gelation I. J Am Chem Soc 63:3083
35. Gauthier MB, Guyon E (1980) Critical elasticity of polyacrylamide above its gel point. J Phys Lett 41:503
36. Gauthier MB, Guyon E, Roux S, Gits S, Lefaucheux F (1987) Critical viscoelastic study of the gelation of silica particles. J Phys 48:869
37. Gordon M, Torkington JA (1981) Scrutiny of the critical exponent paradigm, as exemplified by gelation. Pure Appl Chem 53:1461
38. Hodgson DF, Amis EJ (1990) Dynamic viscoelastic characterization of sol–gel reactions. Macromolecules 23:2512
39. Holly EE, Venkataraman SK, Chambon F, Winter HH (1988) Fourier transform mechanical spectroscopy of viscoelastic materials with transient structures. J Non-Newtonian Fluid Mech 27:17
40. Kantor Y, Webman (1984) I Elastic properties of random percolating systems. Phys Rev Lett 52:1891
41. Kirkpatrick S (1973) Percolation and conduction. Rev Mod Phys 45:574
42. Kremer K (1998) Numerical studies of polymer networks and gels. Philos Mag B 77:569
43. Lapp A, Leibler L, Schosseler F,Strazielle C (1989) Scaling behaviour of pregel sols obtained by end-linking of linear chains. Macromolecules 22:2871
44. Leibler L, Schosseler F (1985) Gelation of polymer solutions: an experimental verification of the scaling behavior of the size distribution function. Phys Rev Lett 55:1110
45. Lin YG, Mallin DT, Chien JCW, Winter HH (1991) Dynamical mechanical measurement of crystallization-induced gelation in thermoplastic elastomeric poly(propylene). Macromolecules 24:850
46. Lubensky TC, Isaacson J (1978) Field Theory for the statistics of branched polymers, gelation and vulcanization. Phys Rev Lett 41:829
47. Martin JE, Adolf D, Wilcoxon JP (1988) Viscoelasticity of near-critical gels. Phys Rev Lett 61:2620
48. Martin JE, Wilcoxon JP (1988) Critical dynamics of the sol–gel transition. Phys Rev Lett 61:373
49. Martins JL, Zunger A (1984) Bond lengths around isovalent impurities and in semiconductor solid solutions. Phys Rev B 30:6217
50. Parisi G, Sourlas N (1981) Critical behavior of branched polymers in the Lee-Yang edge singularity. Phys Rev Lett 46:891
51. Patton E, Wesson JA, Rubinstein M, Wilson JE, Oppenheimer LE (1989) Scaling properties of branched polyesters. Macromolecules 22:1946
52. Plischke M (2006) Critical behavior of entropic shear rigidity. Phys Rev. E 73:061406
53. Plischke M (2007) Rigidity of disordered networks with bond-bending forces. Phys Rev E 76:021401
54. Plischke M, Joós B (1998) Entropic elasticity of diluted central force networks. Phys Rev Lett 80:4907
55. Plischke M, Vernon DC, Joós B, Zhou Z (1999) Entropic rigidity of randomly diluted two- and three-dimensional networks. Phys Rev E 60:3129
56. Ross MSB (2007) Biopolymer gelation-exponents and critical exponents. Polym Bull 58:119
57. Sahimi M (1986) Relation between the critical exponent of elastic percolation networks and the dynamical and geometrical exponents. J Phys C 19:79

58. Sahimi M (2003) Heterogeneous Materials I. Springer, New York
59. Sahimi M, Arbabi S (1993) Mechanics of disordered solids. II Percolation on elastic networks with bond-bending forces. Phys Rev B 47:703
60. Sahimi M, Goddard JD (1985) Superelastic percolation networks and the viscosity of gels. Phys Rev B 32:1869
61. Schmidt M,Burchard W (1981) Critical exponents in polymers: A sol–gel study of anionically prepared styrene-divinylbenzene copolymers. Macromolecules 14:370
62. Stauffer D (1976) Gelation in concentrated critically branched polymer solutions. Percolation scaling theory of intramolecular bond cycles. J Chem Soc Faraday Trans II 72:1354
63. Stockmayer WH (1943) Theory of molecular size distribution and gel formation in branched-chain polymers. J Chem Phys 11:45
64. Takahashi M, Yokoyama K, Masuda T (1994) Dynamic viscoelasticity and critical exponents in sol–gel transition of an end-linking polymer. J Chem Phys 101:798
65. Tokita M, Niki R, Hikichi K (1984) Percolation theory and elastic modulus of gel. J Phys Soc Jpn 53:480
66. Wang J (1989) The bond-bending model in three dimensions. J Phys A 22:291
67. Winter HH, Mours M (1997) Rheology of polymers near liquid–solid transitions. Adv Polym Sci 134:165
68. Woignier T, Phalippou J, Sempere R, Pelons J (1988) Analysis of the elastic behaviour of silica aerogel taken as a percolating system. J Phys Fr 49:289
69. X Xing, Muthupadhyay S, Goldbart PM (2004) Scaling of entropic shear rigidity. Phys Rev Lett 93:225701

Books and Reviews

Brinker CJ, Scherer GW (1990) The Physics and Chemistry of Sol-Gel. Processing Academic, San Diego
de Gennes PG (1979) Scaling Concepts in Polymer Physics. Cornell University Press, Ithac
Rubinstein M, Colby RH (2003) Polymer Physics. Oxford University Press, New York
Stauffer D, Aharony A (1994) Introduction to Percolation Theory, 2nd revised ed. Taylor and Francis, London
Sahimi M (1994) Applications of Percolation Theory. Taylor and Francis, London
Sahimi M (2003) Heterogeneous Materials I & II. Springer, New York

Percolation in Porous Media

Peter King[1], Mohsen Masihi[2]
[1] Imperial College London, London, UK
[2] Sharif University of Technology, Tehran, Iran

Article Outline

Glossary
Definition of the Subject
Introduction
Porous Media
Application of Percolation to Pore Scale
Application of Percolation to the Field Scale
Future Directions
Bibliography

Glossary

Anisotropy There is anisotropy when the global physical property of the system is direction dependent.

Breakthrough time The time for convection of a single phase passive tracer between an injection well and a production well.

Connectivity The fraction of occupied sites belonging to the percolating clusters i. e. represents the strength of the percolating cluster.

Continuum percolation Percolation on continuum spaces with randomly distributed geometrical objects where there is no lattice at all.

Capillary dominated flow A flow regime in which the only dominant driving force is due to capillarity.

Fracture Any discontinuity within a rock mass which developed as a response to stress.

Field scale This represents large scale heterogeneities at reservoir level or the kilometer scale.

Finite size scaling A scaling law within percolation theory which deals with the effects of the finite boundaries.

Invasion percolation Another kind of percolation theory appropriate for describing the structure and amounts of two immiscible fluids at breakthrough.

Modeling Describing physical phenomena under nature's law in some mathematical relations, e. g. governing fluid flow, to better understand the system and to predict its behavior.

Porous media A medium consists of rock grains and disordered void spaces of approximately 10–100 μm across usually occupied by oil, water and gas in a typical hydrocarbon reservoir and characterized by porosity and permeability.

Pore scale This represents pore throat level or the micron scale.

Permeability The "conductance" of the rock to fluid flow determined from Darcy's Law that the flow rate is proportion to the applied pressure gradient and inversely proportional to the fluid viscosity, the constant of proportionality is the permeability.

Percolation threshold A particular value of occupancy probability at which one large cluster spans the whole region.

Simulation Numerical model for solution to the mathematical equations which be able to predict the physical behavior of the system.

Uncertainty The estimated amount or percentage by which an observed or calculated value may differ from the true value.

Definition of the Subject

Porous media are important in many areas including hydrocarbon reservoir engineering, hydrology and environmental engineering. They are also important in, for example, fuel cells, many industrial process and biological systems (lungs, bones, capillary networks and termite nests are all biological examples). Understanding the structure of porous media and the physics of fluid flow in porous media is of great interest. For example, choosing the efficient recovery techniques by reservoir engineers requires understanding of how different fluids and the porous media interact at different scales by simulating the fluid flow in the reservoir under variety of conditions. The exchange and transport of reagents in fuels cells governs their efficiency. Percolation theory which describes the connectivity of a system mathematically [22,100] has also many important applications from the spread of diseases and forest fires to the connectivity of geological entities (e. g., sandbodies or fractures) in porous media used for nuclear waste disposal or for hydrocarbon recovery [94]. Percolation concepts have been used to model fluid movement in porous media and fractured rocks at both pore scale (mm) and reservoir scale (km). At the pore scale a network of pore and throats is used to study the immiscible displacement and estimate the capillary pressure and relative permeabilities. At the field scale, the permeability map is split into either permeable (flow units e. g. good oil bearing sands) or impermeable (background e. g. shale) and assumes that the connectivity of flow units controls the flow movement. Then percolation theory is directly used to estimate static behavior (connectivity – i. e. connected fraction of good oil bearing sands) and dynamic behavior (i. e. effective permeability across the reservoir, breakthrough time between an injector and a producer or post breakthrough behavior). In particular, the percolation approach is able to estimate the uncertainty in the reservoir performance parameters which is not possible with a single detailed conventional simulation model.

Introduction

Percolation theory is a mathematical model of the connectivity and conductivity in geometrically complex systems [100] first developed by Broadbent and Hammersley in [22]. The full description of this theory and its applications to different disciplines can be found elsewhere [13, 94,100]. It links the global geometrical and physical properties of the system to the number density of geometrical objects (representing geological entities, e. g., fractures) placed randomly in space through algebraic universal laws. By universality we mean the large scale behavior of the system is independent of the small scale details of the system, i. e. local geometries.

On the simplest example is an infinite lattice of sites which are occupied with a probability p. Clusters are formed as the neighboring sites are occupied and they are identified by using standard algorithm [48]. The clusters grow in size as the occupancy probability increases. Then, at particular value of p, called the percolation threshold p_c^∞, one large cluster spans the whole region. There are also other small clusters which get absorbed as p further increases. For the infinite lattice there are some simple power law or scaling laws which describe the behavior of the system around the threshold p_c^∞ such as $P(p) \propto (p - p_c^\infty)^\beta$ and $\xi(p) \propto |p - p_c^\infty|^{-\nu}$ where $P(p)$ is the probability that an occupied site belongs to the spanning cluster (so called connected fraction or connectivity) and $\xi(p)$ is the correlation length (which is a measure of the "typical" size of the clusters, excluding the infinite cluster when above the threshold). Note that the correlation length ξ is related to the two point correlation function $g(r)$, which is the probability that two sites separated by a distance r are in the same cluster. The critical exponents β and ν are independent of the kind of the lattice or even if there is a lattice or not (continuum system) they only depends on the dimensionality of space (i. e. 2D or 3D). Values for $\beta = 5/36$ and 0.4 (in 2 and 3 dimensions respectively), and $\nu = 4/3$ and 0.88 [100]. This is known as *universality* and is an important concept in percolation theory which enables us to study and understand the behavior of a very wide range of systems without needing to worry too much about the small scale detail. However, the percolation threshold does depend on the detail of the lattice.

This is the basic framework of percolation theory. To be useful in the context of reservoir modeling we need to address some issues. Everything so far has been defined for an infinite lattice. What happens if there is no lattice at all (i. e. continuum systems with randomly distributed geometrical objects) and when the system is finite. There is also the issue of how to modify these percolation laws for anisotropic systems, variable size and orientation distribution of objects and spatial correlation between objects.

Porous Media

The flows of fluids in porous media are related to many important industrial and geological applications, such

as hydrocarbon recovery or ground water flow modeling [38,95]. The porous media are typically made from rock grains and disordered void spaces and are usually characterized by porosity ϕ (i. e. the storage capacity of a rock, in other words the volume fraction which is void or pore space) and permeability k (i. e. the "conductance" of the rock to fluid flow determined from Darcy's Law that the flow rate is proportion to the applied pressure gradient and inversely proportional to the fluid viscosity, the constant of proportionality is the permeability). The pore spaces are approximately 10–100 µm across and are usually occupied by oil, water and gas (in a typical hydrocarbon reservoir). However, the reservoir itself may be several kilometers in extent. Fluid displacement in porous media depends on the scale and can be controlled by several forces including capillary forces (mostly at the pore scale), viscous forces and gravity forces. Hence, the type of displacement observed depends on the capillary number, which is the ratio of the viscous pressure drop at the pore scale to the capillary pressure, and the Bond number, which is the ratio of the hydrostatic pressure drop over a pore to the capillary pressure. It should be noticed that the full description of displacement process in porous media is very difficult due to the variety of physical phenomena. For example, the flow of two immiscible fluids depends on the wetting properties of the two fluids, their viscosity ratio, their respective densities, and the displacement rate. Typical flow rates in reservoirs are of order of a few feet (10's of centimeters) a day. Hence, on the pore level the flow is controlled almost entirely by capillary forces between immiscible oil and water. However, over large distances viscous and buoyancy forces dominate [105]. To study the efficient recovery techniques (e. g. water flooding), it is necessary to simulate the fluid flow at the reservoir scale. Even with the today's modern computers, the simulation cannot be achieved at the pore scale (typically there would be of the order of 10^{21} pores in a reservoir). The conventional approach is to establish the simulation on a grid of roughly 100 m linear size which represents displacements occurring within millions of pores. The small scale physics is then represented by parameters in macroscopic partial differential equations to describe the transport of fluid in the field scale simulation. These parameters can be measured experimentally on representative core samples or they can be estimated from pore and throat network models (e. g. percolation based models).

The basic equation at the continuum scale is Darcy's law, which states that the flow rate is proportional to the applied pressure gradient and inversely proportional to the fluid viscosity

$$v = -\frac{K}{\eta}\nabla P \tag{1}$$

where K is the rock permeability. Some analysis shows that it has dimensions of area and for typical reservoir rocks it is of the order of $10^{-12}\,\mathrm{m}^2$, which is about the cross sectional area of a typical pore throat (although more detailed calculations are required to obtain better estimates than this). The permeability can vary by orders of magnitude over very short differences reflecting the heterogeneity in pore size arising from the complex processes of geological deposition. This law is used in conjunction with the assumed incompressibility of the fluid to give an equation for the fluid pressure.

$$\nabla \cdot v = \nabla \cdot K \nabla P = 0\,. \tag{2}$$

When there are two or more fluid phases present this basic law gets modified to,

$$v_i = -\frac{K k_i\,(S_i)}{\eta_i}\nabla P_i \tag{3}$$

where the k_i are the relative permeabilities of each phase which are usually assumed to be functions of the fluid saturations (S) only. Fluid saturation is the fraction of the volume of the pore space occupied by the phase. Along with the incompressibility condition for the total flux we also have a conservation law for each phase as,

$$\phi\frac{\partial S_i}{\partial t} + v_i \cdot \nabla S_i = 0 \tag{4}$$

where ϕ is the porosity. This set of equations is the most basic set of equations used to describe multi-phase flow in porous media. Typically the parameters (such as porosity, permeability and relative permeabilities) are determined empirically. There is really no microscopic averaging of the pore scale physics to determine these, although recent studies using percolation as one approach are an attempt to do this (as described in the next section). In reality more complex equations incorporating further physics (such as the phase behavior) are often included also.

The pore scale porous media can be simply modeled as a network of bonds i. e. pore throats and sites i. e. pore bodies to study the flow behavior. Then a series of displacement steps in each pore or throat are combined to simulate the flow movement. The idea was first used by Fatt [39] and since then, the capabilities of network and percolation based models have improved enormously [16,17,24,36,40,53,54,72,80,82,84,101,106]. Invasion percolation is a typical example for modeling capillary

dominated regime of flow in porous media [106] and will be more fully described in the next section. The models are then used to find capillary pressure and/or relative permeability [16,35,46,62,70,99]. For example, Soll & Celia [99] developed a pore scale capillary-dominated flow model by neglecting the viscous forces but considering the gravity effect (which modifies the local capillary pressures) to simulate capillary pressure-saturation relationships in a water-wet system. Moreover, Paterson et al. [81] used a percolation model with trapping to study the effects of spatial correlations in the pore size distributions on the relative permeabilities and residual saturations. They found lower residual saturations for correlated properties in comparison to uncorrelated ones.

At the field scale, heterogeneities which affect the flow behavior appear on all length scales from microns to tens of kilometers and have to be modeled to make reliable predictions of future reservoir performance. However, there exist very few direct measurements of the flow properties. Core plugs directly measure the permeability but they represent a volume of roughly 10^{-13} of a typical reservoir. Well logs and well tests measure large volumes (10^{-4} and 10^{-7} respectively) but the results have to be interpreted to infer flow properties. The flow itself takes place on the scale of the pores which are typically around 10^{-21} of the volume of the reservoir. The consequence is a great deal of uncertainty about the spatial distribution of the heterogeneities which influence the flow. The conventional approach to this is to build detailed reservoir models (note that the largest of these has around 10^7 grid blocks so they fall very short of the actual level heterogeneity that we know about), upscale or coarse grain them to around 10^4 or 10^5 grid blocks and then run flow simulations. These models need to be taken from a whole range of possible models with a suitable probability attached to each to determine the uncertainty in performance. The problem with this approach is that it is computationally very expensive. Therefore, there is a great incentive to produce much simpler models which can predict the uncertainty in performance much quicker. These models must be based on the dominant physics that control the displacement process. The percolation approach based on the connectivity of flow units is one very quick method to model flow and predicts uncertainty in the reservoir performance parameters. This will be further described in Sect. "Application of Percolation to the Field Scale".

Application of Percolation to Pore Scale

Let's start with a simple case of displacement of a fluid by second fluid in a two-phase system i. e. the problem of oil/water flooding. Fluid movement can be governed by viscous, gravity and capillary forces [8]. For systems without gravity we expect different flow regimes depending on the capillary number. Viscous forces in the two fluids may be different mainly because the viscosity of the fluids is different. The high viscosity of the displaced fluid can leads to a highly unstable displacement pattern with a rapid breakthrough of the non-wetting fluid into the wetting fluid called viscous fingering [26,47]. We neglect this by considering the situation that the displacing fluid has a higher viscosity or equal viscosity than the displaced fluid. Then, for slow displacements the invasion percolation can be used to describe the structure and amounts of fluids in a two-phase displacement at breakthrough when the invading fluid is completely nonwetting [24,106] [Lenormand and Bories 1980]. The displacement in the network (or model) is based on physical principles (i. e. the heterogeneity of the capillary pressures along the interface). Consider a lattice with sites and bonds representing pores and throats respectively (with the pores as spheres and the throats as cylinders in three dimensions). The throats can be classified into allowed (those can in principle be invaded by that phase ignoring the effect of surrounding bonds), occupied (those that are occupied by that phase) and accessible (those that are allowed by the phase but also the surrounding bonds will not inhibit the fluid to try). Two processes can be considered for an immiscible displacement. An event where a wetting phase (i. e. water) is displaced by a non-wetting phase with a positive capillary pressure is called drainage. The process where the wetting phase (i. e. water) enters the porous medium and displaces the non-wetting phase is called imbibition. In practice, the capillary pressure for imbibition is lower than that for drainage.

Here we will describe how invasion percolation works for drainage. This is the simplest situation and the extension to other process can be found in the literature. First consider a simple pore with a single interface between the fluids (Fig. 1).

In order to overcome the pressure caused by the interfacial pressure to drive the non-wetting phase (oil) into the wetting phase (water) occupied pore we need to apply a pressure of

$$P_c = \frac{\gamma \cos\theta}{r_{\text{throat}}} \tag{5}$$

where γ is the interfacial tension, θ is the contact angle and r_{throat} is the pore throat radius. We now imagine a network of pores linked by throats of varying radii. As we increase the pressure applied to the invading non-wetting phase it can be seen that the pores will fill from the largest first

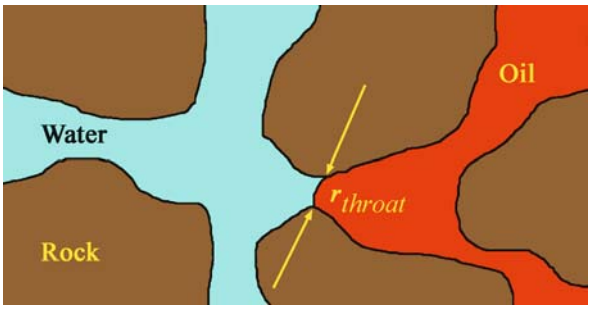

Percolation in Porous Media, Figure 1
A schematic representation of a simple pore with a single interface between the non-wetting phase (oil) and the wetting phase (water)

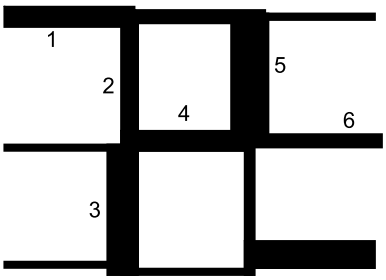

Percolation in Porous Media, Figure 2
A simple illustrative network of pores linked by throats of varying radii. The numbers represent the filling order of pores

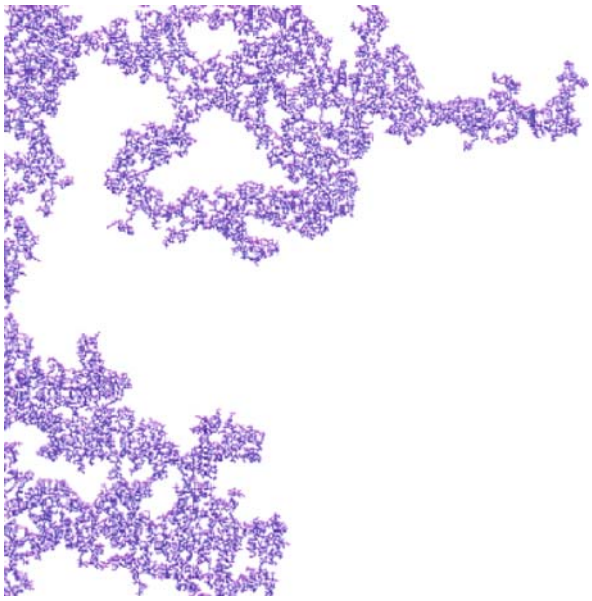

Percolation in Porous Media, Figure 3
A typical percolation structure obtained from invasion percolation simulations for drainage

(as there the entry pressure to fill is lowest). However, the pores can be filled only if they are connected to the inlet face of in contact with the non-wetting phase. Hence, in the simple network below (Fig. 2) the pores will fill in the order 1, 2, 3 and so on where the labels are in order of the throat radius (the invading fluid comes from the left).

If this process is continued we get a not fully occupied structure (Fig. 3).

It can be seen that the displacement process is not very efficient and there are large regions unswept. This process has picked out the percolation cluster of the network. If the bonds are ordered and a threshold applied such that (starting with largest first) a fraction of bonds equal to the percolation threshold are occupied a pattern such as this will be found, except that none of the clusters not connected to the infinite cluster are found. This represents the fact that the invading fluid cannot "jump" from a current occupied site to some other interior site. Flow can only take place through sites already connected to the inlet. This is the simple model of invasion percolation as introduced by Koplik and Wilkinson [54,106].

The main quantities of interest will be the fraction of sites which become occupied by the invader, and the distribution of random numbers of those sites. In invasion percolation with no trapping the clusters exhibit fractal character with fractal dimension (see [61] for fractals) $D = 1.89$ and 2.52 in 2D and 3D respectively which are similar to the results obtained from random percolation. The faction of volume occupied by the invader is proportional to the grid size to the power -0.11 and -0.48 for 2D and 3D cases respectively which are again consistent with the universal values derived from random percolation results. Hence, it has the same universality class as random percolation. Invasion percolation with trapping causes the phenomenon of residual oil. Fractal dimension of invader cluster is 1.82 in 2D which is less than 91/48 of random percolation with no significant difference in 3D. The faction of volume occupied by the invader is proportional to the grid size to power -0.18 in 2D. If we pursue with the invasion process beyond the point of percolation, a second percolation threshold is reached when the defender consists of isolated clusters only and the process stops. Then the system has reached saturation of residual oil. Properties of invasion percolation are believed to be consistent with that of random percolation. However, the spanning clusters are not precisely the same. Effects of gravity were characterized through dimensionless Bond number [15,71]. Percolation theory can then be used to improve our understanding of relative permeability and capillary pressure curves in porous media (e. g. [46]).

Whilst this process is not what happens in the typical oil recovery process it is a reasonable representation of the filling of an oil reservoir when the drops of oil formed in a source rock (typically some distance and much deeper than the final reservoir rock) moves under gravity and capillary forces to displace the water that originally in the reservoir rock.

Application of Percolation to the Field Scale

Here we describe the application of percolation theory first to low to intermediate net-to-gross conventional reservoirs and secondly to fractured reservoirs. Consider, for example, a meandering river which deposits sand over time as represented schematically in Fig. 4. The deposited sand creates a sandbody which covers the meander belt. Occasionally an event upstream, the river changes its path (called an evulsion by geologists) and deposits a new sandbody that may overlap a previous body. The process continues and forms a system of embedded sandbodies in an impermeable background. Although there may be other depositional and post-depositional events such as crevasse splays, mud drapes and shale layers which may alter this simple model, this simple model of overlapping sandbodies has been used [50,75,76] as the basic model for low to intermediate net-to-gross non fractured reservoirs.

Another example could be fracture networks. Field data obtained from large scale investigations show that the fracture network structure is close to the threshold and displays strong channeling patterns which can be explained by percolation theory [23]. It was found that in some natural fracture networks only a small percentage of fractures contributed to the permeability of the system ([95] and references therein). The large uncertainty associated with data and the lack of distinction between faults and joints makes it debatable whether or not natural fractures are well above the percolation threshold (highly interconnected) or near the threshold (poorly connected) ([14] and references therein). As pointed out by Berkowitz [12], in some circumstances fracture networks seems to be highly connected, whereas in many other cases, where the fractures are created as a results of stress, the network is poorly connected which indicates that it is near the percolation threshold. These studies indicate that the application of percolation theory in many fracture networks is reasonable. Simple percolation models would assume that the fractures are randomly oriented and independently located in space, however, in reality fractures show different orientation distribution [1,9,95,108] result from successive tectonic events and several length distributions such as power law [45,77,97,102], log normal (e. g. [77,93]) or exponential (e. g. [86,93]). There also exist spatial correlation between fractures ([18], and references therein) and cross correlation between fracture parameters such as correlation between aperture and the fracture size (e. g. [79,103]) or cross correlation between the position of a facture and its length (e. g. [21,29]). We shall first discuss the very simple model of fractures [4,5,11,19,20,25,65,89] made by constant length randomly oriented and/or orthogonal fractures (Fig. 5).

In these two examples, the objects (i. e. sandbodies and fractures) are not restricted to points on a fixed lattice

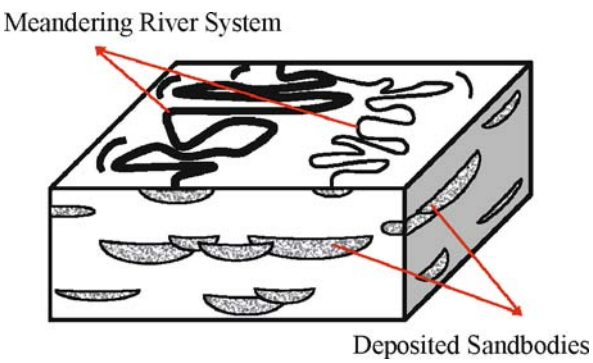

Percolation in Porous Media, Figure 4
A meandering river system deposits sandbodies over the ages [75]

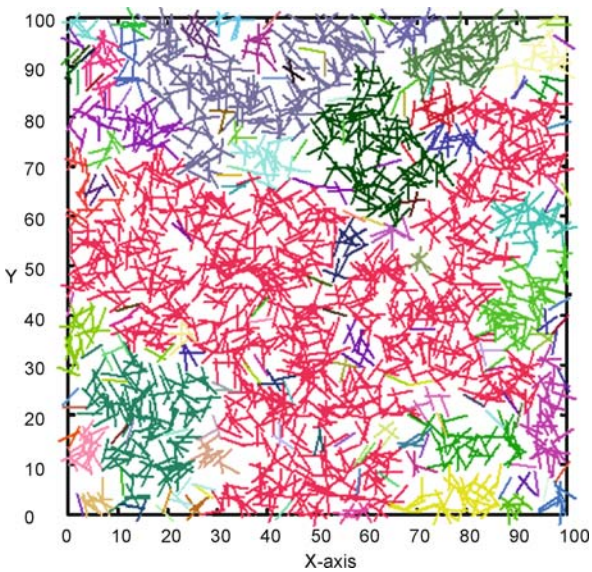

Percolation in Porous Media, Figure 5
A randomly oriented fracture network ($l = 5$) at the threshold with 2066 fractures where the spanning cluster shown in red consists of 897 fractures

so as there is no maximal concentration (notice that the upper bound of the occupancy probability in simple lattice percolation is one). In the case of fracture systems, for instance, there is theoretically no end to the degree of fracturing [95] and fractures can be of any shape with variable length, direction and number of interconnected bonds [11]. These lead us to use percolation on continuum spaces for these cases instead of using percolation on lattices.

Continuum Percolation

This is very straightforward because of the universality principle. We can place geometrical objects (e. g. rectangles representing sandbodies or line segments representing fractures in 2D space) randomly and independently (so called a *Poisson process*) in space. In place of the occupancy probability p we have the volume fraction of objects (or the probability that a point chosen at random lies within one of the objects) with the same notation, p. We get the same threshold phenomenon of a single cluster growing and dominating the system. The percolation threshold depends only on the shape of the objects, but for circles it is 0.678 ± 0.0024 and for squares it is 0.668 ± 0.0026 (similarly in 3D for spheres it is 0.288 ± 0.0016 and for cubes 0.276 ± 0.0013) so the difference is not very large and numerical experiments indicate that for reasonable convex (i. e. not very spiky) objects the threshold is around the same value. This is known as *continuum percolation*. Examples of applying percolation theory to uncorrelated (or even correlated) continuum systems that check the universality and determination of the percolation threshold of different models can be found elsewhere (e. g. [3,11,27,28,41,42,50,57,59,60,107]). Extensive studies have shown that fracture systems, for example, belong to the general class of continuum percolation systems (e. g. [1,5,11,55]). From the principle of universality, the critical exponents are then fixed but the percolation threshold depends on the network topology. Previous estimations of critical exponents were successfully close to those from lattice percolation (e. g. [1,5,19,20,25,90]). Hence, from the principle of universality, we can use the same scaling laws with the same numerical values of the critical exponents as in lattice percolation. This is a remarkable result that we can now use.

Finite Size Scaling

The problem of how to deal with finite size systems is known as *finite size scaling*. The simple scaling laws described in Sect. "Introduction to Percolation Theory" only apply to infinite-size systems. In a finite system because

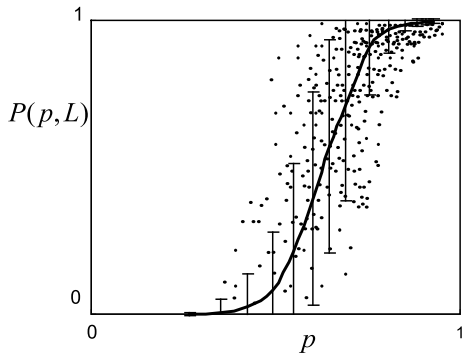

Percolation in Porous Media, Figure 6
A typical scatter with the curve and the lines represent respectively the mean $P(p, L)$ and the standard deviation $\Delta(p, L)$ of connected fraction determined over all realizations at the same occupancy probability of a finite size system

of a sample-size uncertainty there may be a connection at an occupancy very much less than the threshold value but still no connection at very high p values (greater than the threshold). This makes the definition of the percolation threshold of a finite system unclear, as a spanning cluster may appear in one realization and not in another at the same p. Therefore, for a finite system, an apparent threshold $\tilde{p}_c(L)$ which depends on the system size can be used. There have been several definitions of the apparent threshold in the literature [1,11,50,100] such as the occupancy probability at which half of the realizations percolate. The percolation probability (i. e. connectivity) $P(p, L)$ can be defined as the fraction of occupied sites belonging to the spanning clusters. This is the finite size analogue of $P(p)$. If we plot $P(p, L)$ as a function of p over a large number of realizations for a particular system size (Fig. 6), we get a scatter of points from which we can determine the mean connected fraction $P(p, L)$ (the same notation as before) and the standard deviation $\Delta(p, L)$ (the fluctuations about this mean value).

The effect of finite boundaries is to *smear out* the percolation transition (there is not a sharp transition in the connectivity as was exist in infinite systems). Plotting the mean connectivity $P(p, L)$ and the standard deviation of connectivity $\Delta(p, L)$ results obtained from different system sizes as a function of p gives different curves (Fig. 7a) which can be related to each other through the finite-size scaling law [100]:

$$P(p, L) = L^{-\beta/\nu} \mathfrak{I}\left[\left(p - p_c^{\infty}\right) L^{1/\nu}\right] \tag{6}$$

$$\Delta(p, L) = L^{-\beta/\nu} \mathfrak{R}\left[\left(p - p_c^{\infty}\right) L^{1/\nu}\right] \tag{7}$$

where $\mathfrak{I}$ and $\mathfrak{R}$ are two scaling functions for the mean and standard deviation of connectivity, respectively.

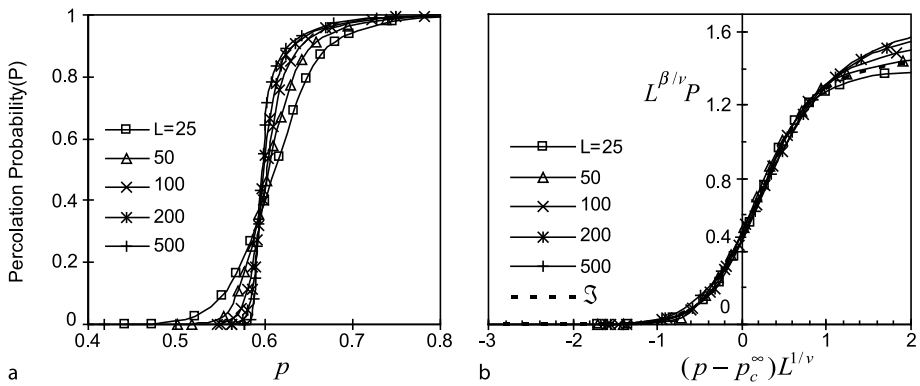

Percolation in Porous Media, Figure 7
Plot of the mean connected fraction $P(p, L)$ showing: a the effects of finite boundaries on the percolation transition and b the data collapse using finite size scaling transformations

This means that, for example, if we plot $L^{\beta/\nu}P$ against $(p - p_c^{\infty})L^{1/\nu}$ all the mean connectivity curves found previously should lie on top of each other to form a single universal curve $\Im$ (Fig. 7b).

It has been shown that all of the curves lie nearly on top of each other except as p approaches unity [65,75,76]. This corresponds to the region where simple scaling breaks down. This region may be small in some cases and could be treated by using effective medium theory [43,50,52,87,98]. It should be noticed that the scaling laws in Eqs. (6) and (7) are universal, but the scaling curves $\Im$ and $\Re$ depend on the model. These are very useful results, because once we get the two scaling curves from numerical simulations for a specific model, we can quickly predict the mean connectivity and its associated uncertainty for any other system sizes without performing any explicit realizations. Examples of the scaling master curves for the mean connectivity and the standard deviation of connectivity for fracture model and sandbodies model were given in Masihi et al. [65,66], and Nurafza et al. [75,76], respectively.

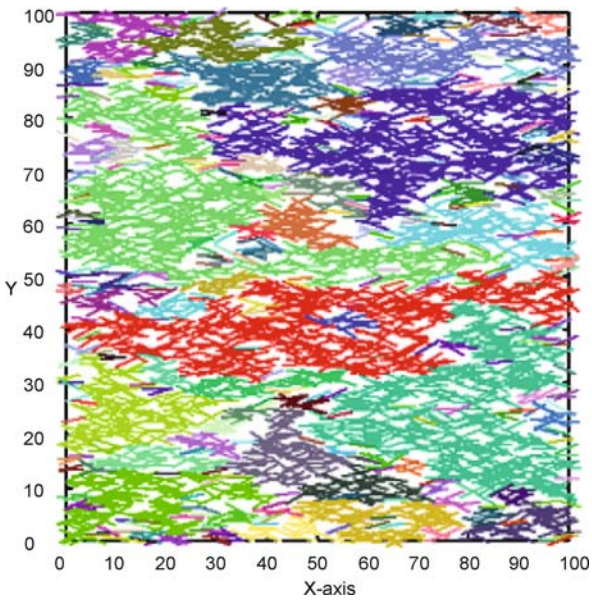

Percolation in Porous Media, Figure 8
An anisotropic fracture network ($l = 5$) at the threshold with random orientation in the range $\theta \in (-30^o, 30^o)$ around the horizontal showing the *easy* direction for connection along *x*-axis with 3802 fractures where the percolating cluster consists of 432 fractures. Connectivity along the *y*-axis, which is the *difficult* direction for connection, needs much more fractures to be placed in the regions

Anisotropy

The other problem that we have to deal with is due to anisotropy. By isotropy we mean that the horizontal connectivity is the same as the vertical connectivity on average if not for individual realizations. However, for many realistic systems, the objects or their orientation are rarely isotropic. For example in fractured rocks fracture sets with particular orientations are typically formed as a result of tectonic history [1,9,43,95,108]. This leads to the creation of an *easy* direction for connected paths which is in the short direction and a *difficult* direction which is along the long axis.

The question is how to apply finite size scaling to anisotropic systems. This requires understanding the anisotropic behavior in percolation. A survey of the literature shows few studies on the subject among which are Monetti and Albano [73] who performed Monte-Carlo simulations in an elongated geometry to obtain the dependency of the horizontal and vertical finite size perco-

lation threshold to the aspect ratio of the lattice; Marrink and Knackstedt [63] who assumed that an elongated lattice can be treated as a series of linked isotropic lattices; Hovi and Aharonyt [49] who used renormalization group theory and duality arguments; Langlands et al. [56] who found numerically the dependency of the crossing probability on the aspect ratio of rectangular systems.

Recently, Masihi et al. [68], have shown that it is possible to account for moderate anisotropy in finite size scaling within percolation by first using the apparent threshold $\tilde{p}_c$ in the principal coordinate directions of the anisotropy as the value of p when 50% of realizations connect in that direction instead of the infinite percolation threshold and then rescaling with the effective length L_x as,

$$P(p, L_x, \omega) = L_x^{-\beta/\nu} \Im\left[(p - \tilde{p}_c) L_x^{1/\nu}\right] \tag{8}$$

$$\Delta(p, L_x, \omega) = \omega^{1/2} L_x^{-\beta/\nu} \Re\left[(p - \tilde{p}_c) L_x^{1/\nu}\right] \tag{9}$$

where aspect ratio $\omega = L_x/L_y$ represents the anisotropy. The apparent threshold in each direction is given by,

$$\tilde{p}_c^i = p_c^\infty + \Lambda_i(\omega) L_x^{-1/\nu} \tag{10}$$

which has a symmetry property (see Fig. 9) where the constant of proportionality is $\Lambda(\omega) = c(\omega^{1/\nu} - 1)$ with $c \approx 0.92$, 0.58 and 0.41 for respectively elongated lattice [68], anisotropic fracture model [69] and anisotropic sandbody model [75,76]. This means that we can use the same isotropic universal curves ($\Im$ and $\Re$) for predicting connectivity of anisotropic cases.

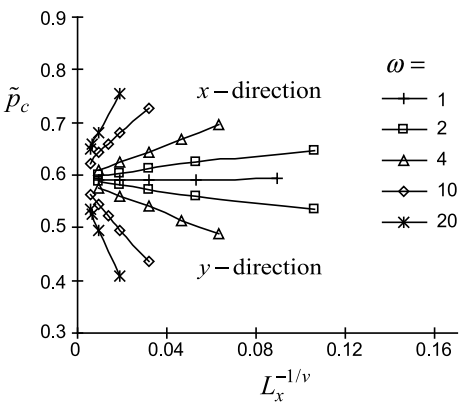

Percolation in Porous Media, Figure 9
Plot of apparent threshold in both the x and the y directions of anisotropic lattices as a function of $L_x^{-1/\nu}$, showing that the shift in the apparent thresholds is symmetrically placed about the isotropic case ($\omega = 1$)

Size Distribution

Another key parameter affecting the connectivity is the size distribution. In reality the sandbodies, for example, may have different sizes based on the sedimentological environment in which the sands were deposited. Also fractures usually have a length distribution depending on the degree of rock deformation (e. g. [88]) from negative-exponential and log-normal to power-law distributions (e. g. [18,45,78,97,102]).

The analysis of the connectivity based on finite size scaling that we have discussed so far assumes that objects (i. e. fractures or sandbodies) all have the same lengths. However the distribution of sizes introduces a new complication. The idea behind the finite size scaling is that the percolation behavior is controlled by two dimensionless lengths, the system size, L, and the correlation length, ξ (these are made dimensionless by scaling with the linear dimension of the geometrical object). If there is a distribution of lengths then this is changed. One might expect that the connectivity behavior of such a system with a distribution of lengths is identical to the connectivity behavior of a system with constant-length objects whose object length (called the effective length, l_{eff}) can be defined in an appropriate way. In the case of fracture model, for example, using the concept of effective length introduced by Robinson [89,90] and Balberg et al. [6], the two previously determined universal connectivity curves for the constant length fractures ($\Im$ and $\Re$) can be applied to fracture systems with a distribution of fracture lengths. This representative length can be based on either the first moment $\langle l \rangle$ or the second moment $\langle l^2 \rangle$ of the fracture length distribution [6,11,19,66,74,89,91]. From numerical studies Masihi et al. [66] found that the second moment gave a better fit to the data. In the case of very wide distribution of lengths (i. e. power law) the scaling exponents may be different from the standard values as there is a possibility for a single large fracture to connect both sides of the system and dominates the connectivity. In other words, in this case there is a break down in universality for the critical exponents. The dependency of the scaling exponents of percolation theory on the exponent of power law length distribution has been investigated numerically [19,66].

Orientation Distribution

In reality the sandbodies, for example, are not all aligned in one direction and are affected by the depositional process. Note that the orientational disorder of the bodies which is more relevant in three dimensions can greatly enhance the connectivity of the system, particularly for the systems with thin long bodies. In percolation terminology this is

a reduction in threshold which is not due to the finite size effects but a real shift in the infinite threshold. Very thin and long bodies, for example, if they are aligned they will not intersect so $p_c \approx 1$ otherwise if they have an angular dispersion they will intersect at any fractional concentration and so $p_c = 0$. There will also be another shift in apparent percolation due to finite size effects. Nurafza et al. [75,76] studied these effects numerically and showed that the only effect of the orientational disorder is to make the bodies appear a bit larger than they are and a bit less elongated. They defined a new aspect ratio and used the reduced percolation thresholds to account for the effects of orientation of sandbodies within the finite size scaling laws which makes the previously determined universal curves applicable for orientated sandbodies.

Spatial Correlation

The models discussed so far assumed the objects are distributed randomly and independently in the region (Poisson statistics). This is somehow in contradiction with the known existence of, for example, fracture sets or observed spatial correlation between natural fractures over scales (e. g. [18]). A survey of the literature shows that there has been little investigation of the percolation properties of systems with short- and long range correlations. Harter [44] has shown that the percolation threshold in Markov chain random field with short range correlation decreases as the correlation scale increases but the critical exponents are unaffected (they belong to the same universality class as uncorrelated percolation). For systems with long range correlations, on the other hand, it has been shown that the percolation behavior is drastically changed as even the critical exponents may become different [85,96].

In the study of fracture networks an investigation of the spatial correlation of fractures has concentrated on the long range fracture density correlations modeled by fractal geometry [14,31,33] [Watanabe and Takahashi, 1995] showed that the correlation pattern is likely to affect the connectivity behavior. When dealing with fracture correlation it is not straightforward to find the right percolation parameter p which is able to measure the connectivity of the fracture network. As pointed out by Darcel et al. [31] neither the mean fracture density $\rho = Nl^2/L^2$, proposed by Bour and Davy [19], nor the mean fractal fracture density $\rho_D = Nl^D/L^D$, suggested by Berkowitz et al. [14] are able to represent the connectivity state of constant-length fractal fracture networks. Darcel et al. [31] also emphasized that the transition width at the threshold for a large system with fractal correlation remains fixed (it does not vanish) which is in contrast to the second order phase transitions of percolation theory. Recently, Masihi and King [64] have presented a model of fractures which used a simulated annealing algorithm (with an objective function defined by the spatial correlation in the displacement of fractures) to generate realizations of correlated fracture networks and then used them in the percolation approach to investigate the effects of fracture spatial correlation. They have found that the scaling exponents of the connectivity are different from the conventional, uncorrelated values (see Masihi et al. [64,66] for more details).

With this background we now describe the percolation framework to model field scale reservoirs. We start with conventional reservoirs. Hydrocarbon reservoirs have a complicated geometry due to the complex sedimentary processes deposited them over the years. They consist of good sandstone (i. e. high permeability and porosity) containing oil within their pores, and poorer siltstones, mudstones and shales (i. e. low permeability). The main flow of oil during recovery is through these good sands and flow through poorer rock being too slow to be of economic consequence. Hence, connectivity of these sandbodies (also called flow units) across the reservoir or in between an injection and a production well is crucial. Note that the total sands gives the total oil in place and the fraction of connected sands between two wells shows the expected recoverable oil between the two wells. A part of the connected sand fraction is dead end and cannot contribute to the flow. The flowing part (backbone) of connected sands controls the flow through sands and so affects the effective permeability, sweep efficiency and breakthrough and post break through time behavior. Many important decisions for a given field can be made based on the knowledge of the connectivity and conductivity of these flow units. For example, at the exploration and appraisal stage decisions about initial well spacing and location can be made based on oil volumes connected by the wells and recovery factors; during plateau phase the decision about the end of plateau and the rate of increases of water or gas ratios depends on the knowledge of the geometry of the backbone or during the decline phase decisions about targeting infill wells to extend field life will be based on volumes of unswept oil or uncontacted oil.

The conventional approach to address these entails building detailed reservoir models which is very expensive in terms of human and CPU times. It has long been understood that flow in heterogeneous porous media is largely controlled by continuity of permeability contrasts, either flow barriers (e. g. shales or high permeability streaks) or faults. Although there are other influences these are the

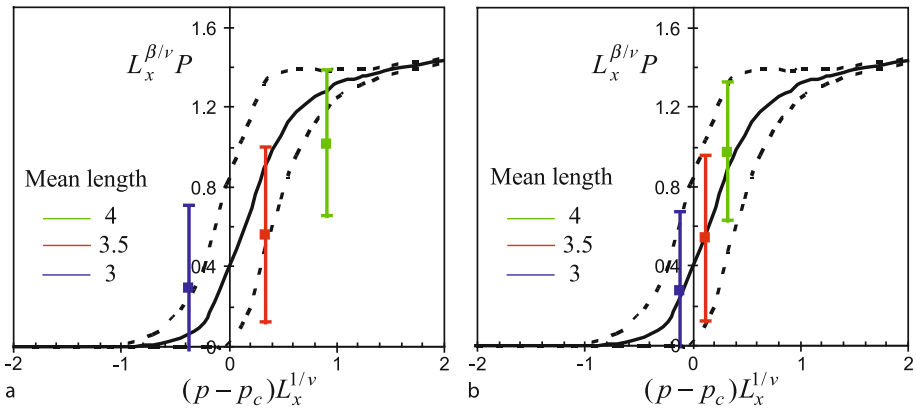

Percolation in Porous Media, Figure 10

A comparison of connectivity scaling of three correlated fracture networks with the average fracture length 4, 3.5 and 3 with the universal connectivity curves (solid and dotted curves are P and $P \pm \Delta$) using: a standard values for the exponent ($1/\nu = 0.75$ and $\beta/\nu = 0.105$) and b modified values $1/\nu = 0.4$ and $\beta/\nu = 0.09$

predominant features affecting flow. With this in mind we look to model reservoir flow which concentrates on the connectivity of permeability contrasts.

Imagine a typical reservoir model constructed with an object based technique [34]. That is geometrical objects (representing geological entities, e. g. shales, fractures, sand bodies etc.) are placed randomly in space. For example, the reservoir may have been deposited by meandering river belts in which case the good sand occurs as packages in a low permeability background. Then the connectivity and conductivity can be estimated directly by percolation theory. Note that the net to gross ratio is the volume fraction of the good sand and is, therefore, identical to the occupancy probability p. Suppose we have a reservoir of size L and a pair of wells separated by a Euclidean distance r. We can ask questions about the probability that the two wells are connected, or in percolation terminology, in the same cluster. This is just the two point correlation function defined previously. Suppose we want to know what fraction of the sand in contact with the wells is connected to both wells. This is just the connectivity function P defined earlier. We can use finite size scaling to estimate this fraction. Also we can use related scaling laws to estimate the uncertainty. Note that these are algebraic laws with no spare parameters. The percolation threshold is defined by the shape of the objects, but it is largely unimportant whether we model the sand units as rectangles or ellipsoids or other shapes (provided they are not too exotic). The scaling laws and exponents are determined from lattice models (and this has been done very extensively in the literature) and can be straightforwardly applied.

The percolation framework can be used to find the probability distribution for the breakthrough time for convection of a single phase passive tracer between an injector and a producer. Comparison of preliminary results [2,37,51,58] showed that the agreement between prediction from scaling law and the numerical simulation which is good enough for engineering purposes given the fact that the prediction from the scaling law took a fraction of a second of CPU times compared with the hours required for the conventional detailed simulation. Having estimated the breakthrough time, Roslien et al. [92] showed that log-log plot of the mean and variance in production results condition to any breakthrough time t_{br} against $(t - t_{\mathrm{br}})$ will lie on top of each other to produce two universal curves. Now with these master curves, one can make a rapid estimate of the mean and variance in future production (e. g. the time taken for production to fall by 50% or water cut to increase to 50%) good enough for engineering purposes.

In fractured reservoirs with low matrix permeability the flow behavior depends strongly on the spatial distribution of the fractures. In this case, the fractures are the flow units which need to be distributed randomly in the impermeable background (i. e. matrix). Then again the connectivity and conductivity can be estimated directly by percolation theory. The occupancy p can be interpreted as the density of fractures e. g. number density or the critical fracture length necessary to ensure percolation for a given number of fractures in a domain [83]. Equivalent terms to this can be volumetric base (i. e., average number of fractures in the region), topological base [1] (i. e., average number of connections with surrounding fractures) or a combination [67] based on average excluded area i. e., the area around a fracture in which the center of another fracture must lie in order for them to overlap over the distri-

bution of the fracture orientation and [6] length. Extensive studies have shown that constant-length fracture models belong to the general universality class of continuum percolation systems [5,11,19,20,25,90]. Recently, Belayneh et al. [10] have applied scaling laws from percolation theory to predict the connectivity of mineralized fractures with length distribution exposed on the southern margin of the Bristol Channel Basin.

Future Directions

There is obviously a need for further development to turn many concepts involved in this article into practical application. One area is to look for connectivity between two wells (represented by points or lines in 2D and 3D, respectively) similar to the previous works on the structure of the cluster connecting two given sites or lines of a 2D and 3D lattices [7,32]. Clearly the connectivity between two points will be lower than that between two sides. In determining the connectivity between two points the correlation function, $g(r)$ defined previously has a major role.

Percolation can do more than predict static connectivity. There are scaling laws for the effective permeability $K_{\text{eff}}(p - p_c^{\infty})^{\mu}$ for infinite size systems where the conductivity exponent μ is about 1.3 and 2 in two and three dimensions respectively [100]. It is clear that only a subset of the percolating cluster (called the backbone) can be swept whereas the dead ends are stagnant. It would be expected to see a similar finite size scaling and anisotropy effects to those for the connectivity. However, the detailed development for this is computationally very demanding as it requires solving the flow equations.

Bibliography

1. Adler PM, Thovert JF (1999) Fractures and fracture networks. Kluwer, London
2. Andrade JS, Buldyrev SV, Dokholyan NV, Havlin S, Lee Y, King PR, Paul G, Stanley HE (2000) Flow between two sites in percolation systems. Phys Rev E 62:8270–8281
3. Baker R, Paul G, Sreenivasan S, Stanley HE (2002) Continuum percolation threshold for interpenetrating squares and cubes. Phys Rev E 66:046136
4. Balberg I (1985) Universal percolation threshold limits in the continuum. Phys Rev B 31(6):4053–4055
5. Balberg I, Binenbaum N, Anderson CH (1983) Critical behaviour of the two dimensional sticks system. Phys Rev Lett B 51(18):1605–1608
6. Balberg I, Anderson CH, Alexander S, Wagner N (1984) Excluded volume and its relation to the onset of percolation. Phys Rev B 30(7):3933–3943
7. Barthelemy M, Buldyrev SV, Havlin S, Stanely HE (1999) Scaling for the critical percolation backbone. Phys Rev E 60(2):R1123–R1125
8. Bear J (1972) Dynamics of fluids in porous media. American Elsevier Publishing Company, New York
9. Bear J, Tsang CF, de Marsily G (1993) Flow and contaminant transport in fractured rock. Academic Press, San Diego, pp 169–231
10. Belayneh M, Masihi M, King PR, Matthäi SK (2006) Prediction of fracture uncertainty using percolation approach: model test with field data. J Geophys Eng 3:219–229
11. Berkowitz B (1995) Analysis of fracture network connectivity using percolation theory. Math Geol 27(4):467–483
12. Berkowitz B (2002) Characterizing flow and transport in fractured geological media: a review. Adv Water Resour 25:861–884
13. Berkowitz B, Balberg I (1993) Percolation theory and its application to ground water hydrology. Water Resour Res 29(4):775–794
14. Berkowitz B, Bour O, Davy P, Odling N (2000) Scaling of fracture connectivity in geological formations. Geophys Res Lett 27(14):2061–2064
15. Birovljev A, Furuberg L, Feder J, Jssang T, Mly KJ, Aharony A (1991) Gravity invasion percolation in two dimensions: Experiment and simulation. Phys Rev Lett 67:584–587
16. Blunt MJ, King PR (1990) Macroscopic parameters from simulations of pore scale flow. Phys Rev A 42(12):4780–4789
17. Blunt MJ, King M, Scher H (1992) Simulation and theory of two phase flow in porous media. Phys Rev A 46(12):7680–7699
18. Bonnet E, Bour O, Odling NE, Davy P, Main I, Cowie P, Berkowitz B (2001) Scaling of fracture systems in geological media. Rev Geophys 39(3):347–383
19. Bour O, Davy P (1997) Connectivity of random fault networks following a power law fault length distribution. Water Resour Res 33(7):1567–1583
20. Bour O, Davy P (1998) On the connectivity of three dimensional fault networks. Water Resour Res 34(10):2611–2622
21. Bour O, Davy P (1999) Clustering and size distributions of fault pattern: theory and measurements. Geophys Res Lett Press, 26:2001–2004
22. Broadbent I, Hammersley JM (1957) Percolation processes 1. Crystals and mazes. Proc Camb Philos Soc 53:629–641
23. Cacas MC, Ledoux E, de Marsily G, Tillie B, Barbreau A, Durand E, Feuga B, Peaudecerf P (1990) Modeling fracture flow with stochastic discrete fracture network: calibration and validation, 1. the flow model. Water Resour Res 26(3):479–489
24. Chandler R, Koplik J, Lerman K, Willemsen JF (1982) Capillary displacement and percolation in porous media. J Fluid Mech 119:249–267
25. Charlaix E (1986) Percolation threshold of random array of discs: a numerical simulation. J Phys A Math Gen 19(9):L533–L536
26. Chen JD, Wilkinson D (1985) Pore-scale viscous fingering in porous media. Phys Rev Lett 55:1892–1895
27. Choi HS, Talbot J, Tarjus G, Viot P (1995) Percolation and structural properties of particle deposits. Phys Rev E 51(2):1353–1363
28. Consiglio R, Zouain RNA, Baker DR, Paul G, Stanley HE (2004) Symmetry of the continuum percolation threshold in systems of two different size objects. Physica A 343:343–347
29. Darcel C, Bour O, Davy P (2003) Cross-correlation between length and position in real fracture networks. Geophys Res Lett 30(12):52.1–52.4

30. Darcel C, Bour O, Davy P (2003) Stereological analysis of fractal fracture networks. J Geophys Res 108(B9):ETG13.1–ETG13.14
31. Darcel C, Bour O, Davy P, De Dreuzy JR (2003) Connectivity properties of two dimensional fracture networks with stochastic fractal correlation. Water Resour Res 39(10):SBH1.1–SBH1.13
32. Da Silva LR, Paul G, Havlin S, Baker DR, Stanely HE (2003) Scaling of cluster mass between two lines in 3d percolation. Physica A 318:307–318
33. De Dreuzy JR, Darcel C, Davy P, Bour O (2004) Influence of spatial correlation of fracture centres on the permeability of two-dimensional fracture networks following a power law length distribution. Water Resour Res 40:W01502.1–W01502.11
34. Deutsch CV (2002) Geostatistical Reservoir Modeling. Oxford University Press, New York
35. Diaz CE, Chatzis I, Dullien FAL (1987) Simulation of capillary pressure curves using bond correlated site percolation on a simple cubic network. Transp Porous Media 2:215–240
36. Dijkstra T, Bartelds GA, Bruining J, Hassanizadeh M (1999) Dynamic Pore-Scale network for Two-Phase Flow. In: Proceedings on the international workshop on characterization and measurement of hydraulic properties of unsaturated porous media, Riverside, pp 63–71
37. Dokholyan NV, Lee Y, Buldyrew SV, Havlin S, King PR, Stanley HE (1998) Scaling of the distribution of shortest paths in percolation. J Stat Phys 93:603–613
38. Dullien FAL (1992) Porous media fluid transport and pore structure, 2nd edn. Academic Press, San Diego
39. Fatt (1956) The network model of porous media III. dynamic properties of networks with tube radius distribution. Trans Amer Inst Min, Metall, Petrol Eng 207:164–181
40. Fenwick DH, Blunt MJ (1998) Three-dimensional modeling of three phase imbibition and drainage. Adv Water Resour 21(2):121–143
41. Garboczi EJ, Snyder KA, Douglas JF, Thorpe MF (1995) Geometrical percolation threshold of overlapping ellipsoids. Phys Rev E 52(1):819–827
42. Gawlinski ET, Stanley HE (1981) Continuum percolation in two dimensions: Monte Carlo tests of scaling and universality for non-interacting discs. J Phys A Math Gen 14:L291–L299
43. Harris CK (1992) Effective medium treatment of flow through anisotropic fracture system-Improved permeability estimates using a new lattice mapping. Trans Porous Media 9:287–295
44. Harter T (2005) Finite size scaling analysis of percolation in three dimensional correlated binary Markov chain random fields. Phys Rev E 72:026120.1–26120.7
45. Heffer KJ, Bervan TG (1990) Scaling relationships in natural fractures-data, theory and applications. Paper SPE 20981 presented at Europec conference, The Hague, 20–24 Oct
46. Heiba AA, Sahimi M, Scriven LE, Davis HT (1992) Percolation theory of two-phase relative permeability. SPE Reserv Eng 7:123–132
47. Homsy GM (1987) Viscous fingering in porous media. Annu Rev Fluid Mech 19:271–311
48. Hoshen J, Kopelman R (1976) Percolation and cluster distribution, I, cluster multiple labelling technique and critical concentration algorithm. Phys Rev B 14(8):3438–3445
49. Hovi J-P, Aharony A (1996) Scaling and universality in the spanning probability for percolation. Phys Rev E 53:235
50. King PR (1990) The connectivity and conductivity of overlapping sandbodies. In: Buller AT (ed) North Sea Oil and Gas Reservoirs II. Graham and Trotman, London, pp 353–361
51. King PR, Buldyrev SV, Dokholyan NV, Havlin S, Lee Y, Paul G (2001) Predicting oil recovery using percolation theory. Petrol Geosci 7:S105–S107
52. Kirkpatrick S (1973) Percolation and conduction. Rev Mod Phys 45:574–588
53. Knackstedt MA, Sheppard AP, Sahimi M (2001) Pore network modelling of two-phase flow in porous rock: the effect of correlated heterogeneity. Adv Water Resour 24(3–4):257–277
54. Koplik J, Lasseter TJ (1985) Two-phase flow in random network models of porous media. SPE J February:89–100
55. Koudine N, Garcia RG, Thovert JF, Adler PM (1998) Permeability of three dimensional fracture networks. Phys Rev E 57(4):4466–4479
56. Langlands RP, Pichet C, Pouliot P, Saint Aubin Y (1992) On the universality of crossing probabilities in two-dimensional percolation. J Stat Phys 67:553
57. Lee SB, Torquato S (1990) Monte Carlo study of correlated continuum percolation: Universality and percolation thresholds. Phys Rev A 41(10):5338–5344
58. Lee Y, Andrade JS, Buldyrev SV, Dokholyan NV, Havlin S, King PR, Paul G, Stanley HE (1999) Traveling time and traveling length in critical percolation clusters. Phys Rev E 60:3425–3428
109. Lenormand R, Bories S (1980) CR Acad Sci, Paris B291:279
59. Lin C-Y, Hu C-K (1998) Universal finite size scaling functions for percolation on three dimensional lattices. Phys Rev E 58(2):1521–1527
60. Lorenz CD, Ziff RM (2001) Precise determination of the critical percolation threshold for the three dimensional Swiss cheese model using a growth algorithm. J Chem Phys 114(8):3659–3661
61. Mandelbrot BB (1982) The fractal geometry of nature. W.H. Freeman, New York, pp 468
62. Mani V, Mohanty KK (1998) Pore-level network modeling of three-phase capillary pressure and relative permeability curves. SPE J 3:238–248
63. Marrink SJ, Knackstedt MA (1999) Percolation thresholds on elongated lattices. J Phys A Math Gen 32:L461–L466
64. Masihi M, King PR (2007) A correlated fracture network: modeling and percolation properties. Water Resour Res J 43. doi: 10.1029/2006WR005331
65. Masihi M, King PR, Nurafza P (2005) Fast estimation of performance parameters in fractured reservoirs using percolation theory. Paper SPE 94186 presented at the 14th Biennial Conference, Madrid, 13–16 June
66. Masihi M, King PR, Nurafza P (2006) Connectivity prediction in fractured reservoirs with variable fracture size; analysis and validation. Paper SPE 100229 presented at the SPE Europec, Vienna, 12–15 June
67. Masihi M, King PR, Nurafza P (2006) Connectivity of fracture networks: the effects of anisotropy and spatial correlation. Paper CMWR XVI-99 presented at the Computational Methods in Water Resources conference, Copenhagen, 19–22 June
68. Masihi M, King PR, Nurafza P (2006) The Effect of Anisotropy on Finite Size Scaling in Percolation Theory. Phys Rev E 74:042102

69. Masihi M, King PR, Nurafza P (2007) Fast estimation of connectivity in fractured reservoirs using percolation theory. SPE J 12(2):167–178
70. Maximenko A, Kadet VV (2000) Determination of relative permeabilities using the network models of porous media. J Petrol Sci Eng 28(3):145–152
71. Méheust Y, Løvoll G, Måløy KJ, Schmittbuhl J (2002) Interface scaling in a 2d porous medium under combined viscous, gravity and capillary effects. Phys Rev E 66:51603–51615
72. Mohanty KK, Salter SJ (1982) Multiphase flow in porous media: II Pore-level modeling. Paper SPE 11018, Proceedings of the 57th SPE Annual Fall Technical Conference and Exhibition, New Orleans, 26–29 Sept
73. Monetti RA, Albano EV (1991) Critical behaviour of the site percolation model on the square lattice in a $L \times M$ geometry. Z Phys B Condensed Matter 82:129–134
74. Mourzenko VV, Thovert JF, Adler PM (2005) Percolation of three dimensional fracture networks with power law size distribution. Phys Rev E 72:036103.1–036103.14
75. Nurafza P, King PR, Masihi M (2006) Facies connectivity modelling; analysis and field study. Paper SPE 100333 presented at the SPE Europec, Vienna, 12–15 June
76. Nurafza P, Masihi M, King PR (2006) Connectivity modeling of heterogeneous systems: analysis and field study. Paper CMWR XVI-189 presented at the Computational Methods in Water Resources conference, Copenhagen, Denmark, 19–22 June
77. Odling NE (1997) Scaling and connectivity of joint systems in sandstones from western Norway. J Struct Geol 19(10):1257–1271
78. Odling NE, Gillespie P, Bourgine B, Castaing C, Chiles J-P, Christensen NP, Fillion E, Genter A, Olsen C, Thrane L, Trice R, Aarseth E, Walsh JJ, Watterson J (1999) Variations in fracture system geometry and their implications for fluid flow in fractured hydrocarbon reservoirs. Petrol Geosci 5:373–384
79. Olson JE (2003) Sublinear scaling of fracture aperture versus length: an exception or the rule? J Geophys Res 108(B9):ETG3.1–ETG3.11
80. Øren PE, Bakke S, Arntzen OJ (1998) Extending predictive capabilities to network models. SPE J 3(4):324–336
81. Paterson L, Lee JY, Pinczewski WV (1997) Three-phase relative permeability in heterogeneous formations. Paper SPE 38882, Proceedings of the SPE Annual Technical Conference and Exhibition, San Antonio, 5–8 Oct
82. Pereira GG, Pinczewski WV, Chan DYC, Paterson L, Øren PE (1996) Pore-scale network model for drainage-dominated three-phase flow in porous media. Trans Porous Media 24(2):167–201
83. Pike GE, Seager CH (1974) Percolation and conductivity: a computer study I. Phys Rev B 10(4):1421–1434
84. Piri M, Blunt MJ (2002) Pore-scale modeling of three-phase flow in mixed-wet systems. Paper SPE 77726, Proceedings of the SPE Annual Technical Conference and Exhibition, San Antonio, Texas, 29 September-2 October.
85. Prakash S, Havlin S, Schwartz M, Stanley HE (1992) Structural and long-range correlated percolation. Phys Rev A 46(4):R1724–1727
86. Priest SD, Hudson JA (1981) Estimation of discontinuity spacings and trace length using scan line surveys. Int J Rock Mech, Min Sci Geomech Abstr 18:185–197
87. Renshaw CE (1999) Connectivity of joint networks with power law length distributions. Water Resour Res 35(9):2661–2670
88. Rives T, Razack M, Pett JP, Rawnsley KD (1992) Joint spacing: analogue and numerical simulation. J Struct Geol 14(8/9):925–937
89. Robinson PC (1983) Connectivity of fracture systems-a percolation theory approach. J Phys A Math General 16:605–614
90. Robinson PC (1984) Numerical calculations of critical densities for lines and planes. J Phys A Math General 17(14):2823–2830
91. Rossen WR, Gu Y, Lake LW (2000) Connectivity and permeability in fracture networks obeying power law statistics. Paper SPE 59720, Proceedings of the SPE Permian Basin Oil and Gas Recovery Conference, Midland, Texas, 21–23 March
92. Roslien J, King PR, Buldyrev S, Lopez E, Stanley HE (2004) Prediction oil production conditioned on breakthrough time. Petrol Geosc (submitted)
93. Rouleau A, Gale JE (1985) Statistical characterization of the fracture system in the Stripa Granite, Sweden. International J Rock Mech, Min Sci Geomech Abstr 22(6):353–367
94. Sahimi M (1994) Applications of percolation theory. Taylor and Francis, London
95. Sahimi M (1995) Flow and transport in porous media and fractured Rock. VCH publication, London, pp 103–157
96. Schmittbuhl J, Vilotte JP, Roux S (1993) Percolation through self affine surfaces. J Phys A Math General 26:6115–6133
97. Segall P, Pollard DD (1983) Joint formation in granitic rock in the Sierra Nevada. Geol Soc Am Bull 94:563–575
98. Snow DT (1969) Anisotropic permeability of fractured media. Water Resour Res 5(6):1273–1289
99. Soll WE, Celia MA (1993) A modified percolation approach to simulating three-fluid capillary pressure-saturation relationships. Adv Water Resour 16:107–126
100. Stauffer D, Aharony A (1992) Introduction to percolation theory. Taylor and Francis, London
101. Valvatne PH, Blunt MJ (2003) Predictive pore-scale network modeling. Paper SPE 84550, Proceedings of the SPE Annual Technical Conference and Exhibition, Denver, Colorado, 5–8 Oct
102. Van Dijk JP, Bello M, Toscano C, Bersani A, Nardon S (2000) Tectonic model and three-dimensional fracture analysis of Monte Alpi-Southern Apennines. Tectonophysics 324:203–237
103. Vermilye JM, Scholz CH (1995) Relationship between vein length and aperture. J Struct Geol 17(3):423–434
110. Watanabe K, Takahashi H (1995) Fractal geometry characterization of geothermal reservoir fracture networks. J Geophys Res 100(B1):521–528
104. Watanabe H, Yukawa S, Ito N, Hu C-K (2004) Superscaling of percolation on rectangular domains. Phys Rev Lett 93(19):190601.1–190601.4
105. Wilkinson D (1984) Percolation model of immiscible displacement in the presence of buoyancy forces. Phys Rev A 34(1):520–531
106. Wilkinson D, Willemsen JF (1983) Invasion percolation: a new form of percolation theory. J Phys A Math General 16(16):3365–3376
107. Xia W, Thrope MF (1988) Percolation properties of random ellipses. Phys Rev A 38(5):2650–2656
108. Zhang X, Sanderson DJ (2002) Numerical modelling and analysis of fluid flow and deformation of fractured rock masses. Elsevier Science, Pergaman

Percolation Thresholds, Exact

John C. Wierman
Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Trees
Two-Dimensional Bond Models
Site Models in Two Dimensions
Random Voronoi Percolation
Multiparameter Critical Surfaces
Future Directions
Bibliography

Glossary

Archimedean lattices A *regular tiling* is a tiling of the plane which consists entirely of regular polygons. (A regular polygon is one in which all side lengths are equal and all interior angles are equal.) An *Archimedean lattice* is the graph of vertices and edges of a regular tiling which is vertex-transitive, i. e., for every pair of vertices, u and v, there is a graph isomorphism that maps u to v. There are exactly 11 Archimedean lattices. A notation for Archimedean lattices, which can also serve as a prescription for constructing them, is given in Grünbaum and Shephard [5]. Around any vertex (since all are equivalent, by vertex-transitivity), starting with the smallest polygon touching the vertex, list the number of edges of the successive polygons around the vertex. For convenience, an exponent is used to indicate that a number of successive polygons have the same size.

Bond percolation In a bond percolation model, a random subgraph is formed from an infinite graph G by retaining each edge of G with probability p, independently of all other edges.

Dual graph A graph is planar if it may be drawn in the plane with no edges intersecting except at their endpoints, thus dividing the plane into faces. Every planar graph G has a dual graph, denoted here by $D(G)$. $D(G)$ may be constructed by placing a vertex of $D(G)$ in each face of G and connecting two vertices of $D(G)$ by an edge if the corresponding faces in G share a common edge. Note that $D(D(G)) = G$.

Line graph The line graph, $L(G)$, of a graph G is constructed by placing a vertex of $L(G)$ on each edge of G and connecting two vertices of $L(G)$ if the corresponding edges of G share a common endpoint.

Matching graphs A pair of matching graphs may be constructed from an underlying planar graph. Select a set F of faces of the graph. Construct a graph G by adding an edge in each face of F between any pair of vertices that are not already connected by an edge. Construct the matching graph $M(G)$ of G by adding an edge between any pair of vertices in each face not in F that are not already connected by an edge. Note that $M(M(G)) = G$.

Percolation threshold In a percolation model with parameter p, there is a retention probability p_c, called the percolation threshold, above which the random subgraph contains an infinite connected component and below which all connected components are finite.

Periodic graph A periodic graph is an infinite graph that can be represented in d-dimensional space so that it is invariant under translations by all integer linear combinations of a fixed basis.

Site percolation In a site percolation model, a random graph is formed from an infinite graph G by retaining each vertex of G with probability p, independently of all other vertices. An edge of G is retained in the random graph if both its endpoint vertices are retained.

Definition of the Subject

Percolation models were introduced in the 1950s by Broadbent and Hammersley [1] to model the flow of fluid in a random medium. Since both terms, fluid and medium, may be broadly interpreted, percolation has a wide variety of applications, including thermal phase transitions, oil flow in sandstone, and the spread of epidemics. An important motivation for the development of percolation models was to provide an alternative to diffusion models, in which the randomness was associated with the fluid while the medium is relatively homogeneous. Since percolation models associate the randomness with the medium, it is possible for the fluid either to become trapped or to flow infinitely far. This presence of a phase transition is an important reason for the importance of percolation models. The percolation threshold is a critical probability in the percolation model which corresponds to the phase transition point. Since the emphasis in percolation theory is on the effect of the medium on the behavior of the model, it is important to understand how the percolation threshold depends on the characteristics of the medium.

The medium is often, but not always, modeled by a periodic graph, representing an atomic lattice structure. In the bond percolation model, each edge of the graph is retained with probability p, independently of all other

edges, to create a random subgraph. In the site percolation model, each vertex of the graph is retained with probability p, independently of all other vertices, and each edge is retained if and only if both its endpoints are retained. In both models, the focus is on the properties – in particular, the size – of the connected components, called clusters, of the random subgraph.

The most common definition is that the percolation threshold is a retention probability value p_c such that if $p > p_c$ there exists an infinite cluster in the graph and if $p < p_c$ there are only finite clusters in the graph. However, there exist other interpretations, which correspond to different definitions of the percolation threshold: (1) A percolation threshold p_H is the critical probability above which a specific vertex v is in an infinite cluster with positive probability. The percolation threshold p_H is independent of the specific vertex v if the graph is connected. (2) A percolation threshold p_T is defined as the critical probability above which the expected cluster size containing a specific vertex v is infinite, and is also independent of the choice of v if the graph is connected. (3) For periodic graphs, a percolation threshold p_S may be defined in terms of the limiting behavior of the probability that a cluster connects opposite sides of a rectangle in a sequence of similar rectangles whose areas are increasing to infinity. For the periodic graphs discussed in this article, these definitions provide equal values for the percolation threshold.

Due to the dependence of the percolation threshold on the features of the lattice, since the origins of percolation theory much research has been devoted to deriving exact values, computing simulation estimates, and proposing approximation formulas for the percolation threshold as a function of the lattice. This article focuses on the extent of knowledge of exact values of percolation thresholds for bond percolation and site percolation on various graphs.

Introduction

Although the percolation threshold problem is simply described and easily visualized, it has become recognized as extremely intractable, with the result that, after 50 years of research, exact percolation thresholds are known for few graphs. Besides the trivial one-dimensional case, and infinite regular trees, the only solutions are for two-dimensional graphs. There are no exact solutions for periodic graphs in three dimensions or higher.

We now provide a brief history of the development of exact percolation thresholds and the mathematical tools used.

The first major development was in 1960 by Harris [6], who proved a lower bound of 1/2 for the square lattice bond model threshold. The value was larger than simulation estimates at that time, and was believed to be sharp. Harris used the self-duality of the square lattice extensively in the proof, and established a lemma regarding covariant events which was later generalized by Fortuin, Kasteleyn, and Ginibre [4] and played a crucial role in later exact percolation threshold proofs.

Recognizing the importance of duality in the study of bond percolation, Sykes and Essam [15] developed a corresponding concept of matching graphs for site percolation models. Interpreting the percolation threshold as a singularity of a clusters-per-site function, they derived values of the bond percolation thresholds for the square, triangular, and hexagonal lattices. Their methods implied that site percolation thresholds of matching graphs sum to one, and bond percolation thresholds of dual graphs sum to one. A key transformation in the solutions for the triangular and hexagonal lattices was a star-triangle transformation relating the two graphs. However, their exact values would not be given mathematically rigorous proofs until much later.

Although they did not establish the exact percolation thresholds, Seymour and Welsh [14], in the context of the square lattice bond model, laid important groundwork for the solution that followed. They recognized and defined the p_H, p_T, and p_S interpretations of the percolation threshold, and proved relationships among them. Russo [11,12] independently established similar results for the square lattice site model.

In 1980, Kesten [7] rigorously established that the percolation threshold of the square lattice bond model is 1/2, using the self-duality of the square lattice, and proving that all versions of the percolation threshold are equal in this case.

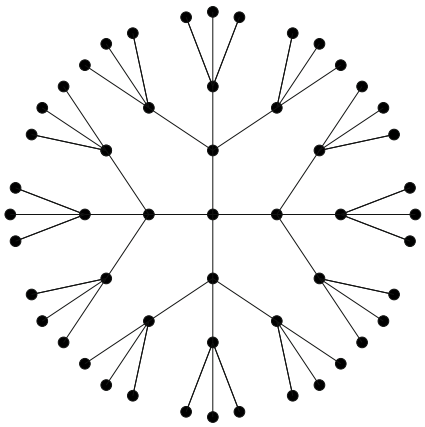

Percolation Thresholds, Exact, Figure 1
A portion of a Cayley tree with vertex degree four

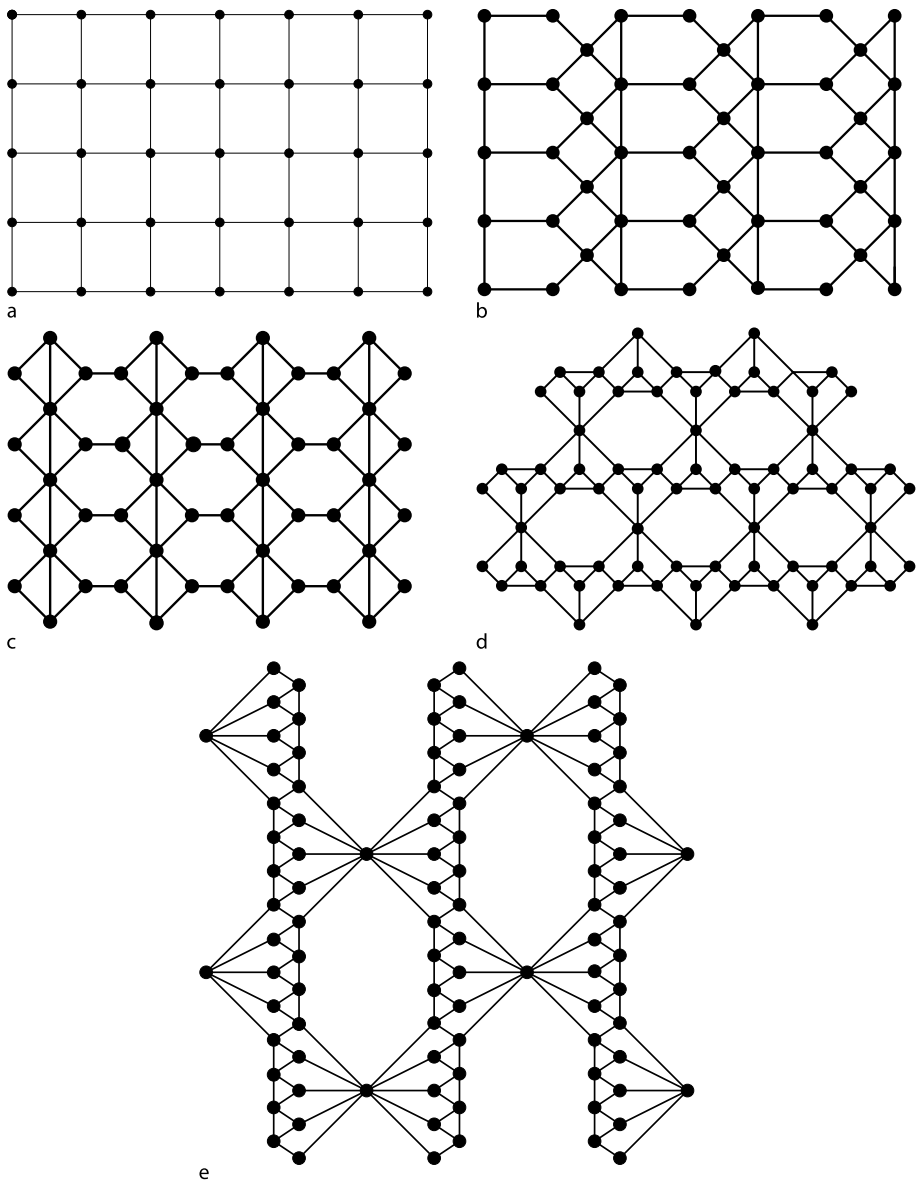

Percolation Thresholds, Exact, Figure 2
Portions of five self-dual periodic graphs. The square lattice is at the *upper left*. An example of the generalization to a family of self-dual graphs is at the *bottom*

Using Kesten's methods, the duality of the triangular and hexagonal lattices, and the star-triangle transformation between them, in 1981 Wierman [17] proved that the bond percolation threshold of the triangular lattice is the root of $1 - 3p + p^3$ in the interval $[0, 1]$, which is equal to $2\sin(\pi/18) \approx .347296$, and the bond percolation threshold of the hexagonal lattice is the complementary value, approximately .652704. In 1984, Wierman [19] discovered another pair of dual lattices for which the exact bond percolation threshold could be determined using a version of the star-triangle transformation. The bond threshold of the bowtie lattice is the root of the polynomial $1 - p - 6p^2 + 6p^3 - p^5$ in $[0, 1]$, which is approximately .404518, while the bond threshold of its dual graph is the complementary value, approximately .595482.

These results were generalized by Kesten [8] in his 1982 monograph, where he proved that the site percolation thresholds of a pair of periodic matching graphs

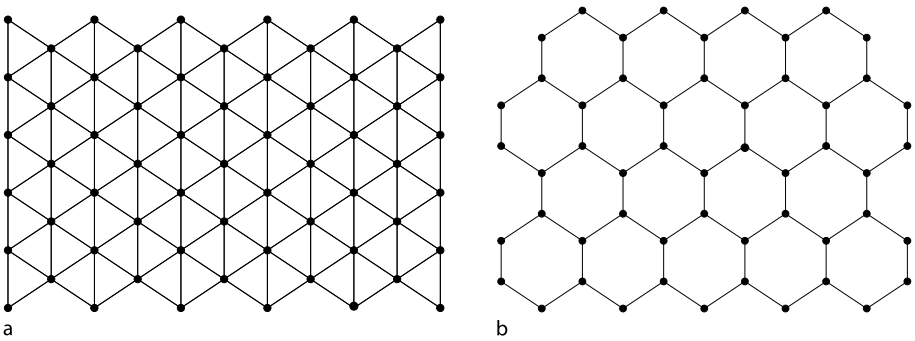

Percolation Thresholds, Exact, Figure 3
Portions of the triangular (a) and hexagonal or honeycomb (b) lattices

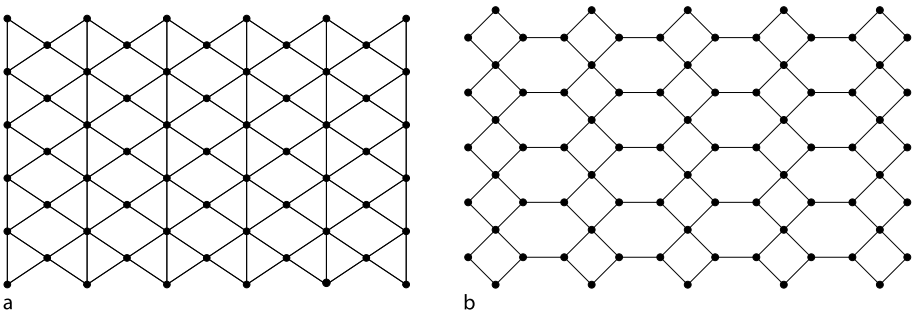

Percolation Thresholds, Exact, Figure 4
Portions of the bow-tie lattice (a) and its dual lattice (b)

sum to one. Since it is fully-triangulated and therefore self-matching, the triangular lattice has percolation threshold equal to one-half. The duality result for bond percolation thresholds is implied by this result via the bond-to-site transformation.

In 2006, Scullard and Ziff [13,21,22] derived exact bond percolation thresholds of additional periodic lattices based on the star-triangle transformation.

Exact percolation thresholds can be derived for additional graphs that are obtained by various transformations of graphs with exact solutions. Such solutions can be established via the bond-to-site transformation, subdivision of edges, and replacing edges with more complex decorations.

Trees

A tree is a graph which is connected and has no cycles, or equivalently, which has a unique path between every pair of vertices. An infinite tree in which every vertex has the same degree is called a Cayley tree or Bethe lattice.

The earliest non-trivial exact percolation threshold solutions were for Cayley trees. Let C_k denote the Cayley tree of degree k. Then, for all $k \geq 3$,

$$p_c(C_k) = \frac{1}{k-1} ,$$

for both bond percolation and site percolation.

It is easy to see that the bond and site percolation thresholds are equal. In a bond model on a Cayley tree, consider starting from a specific vertex, called the root, and moving outward. For each edge, consider the vertex at the end farthest from the root to be retained or not according to whether the edge is retained or not. This creates a site percolation model with the same parameter value as the original bond percolation model, in which there is an infinite cluster if and only if there is an infinite cluster in the bond model. Thus, the percolation thresholds of the two models are equal.

The values of the thresholds for Cayley trees can be determined either by calculation of the probability that a vertex is in an infinite component or by elementary theory of branching processes.

Lyons [9] has shown that, for rooted trees in general, an average number of branches per vertex, called the branching number, may be defined. The percolation

threshold of the tree is the reciprocal of the branching number. However, the definition of the branching number is rather intricate, so exact percolation thresholds are not easily computed.

Two-Dimensional Bond Models

Planarity plays an important role in establishing exact bond percolation thresholds in two-dimensional models. A graph is planar if it may be drawn in the plane with no edges intersecting except at their endpoints, thus dividing the plane into faces. Every planar graph G has a dual graph, denoted here by $D(G)$. $D(G)$ may be constructed by placing a vertex of $D(G)$ in each face of G and connecting two vertices of $D(G)$ by an edge if the corresponding faces in G share a common edge. Note that $D(D(G)) = G$.

The computational importance of dual graphs is that the bond percolation thresholds of a pair of periodic matching graphs sum to one, as implied via the bond-to-site transformation by a result Kesten in 1982.

Without an additional relationship between the pair of dual graphs, duality itself does not yield an exact percolation threshold solution. However, if the two graphs are isomorphic, the common graph is said to be self-dual, and its bond threshold must be one-half. Thus, the bond percolation threshold of the square lattice is exactly one-half.

Besides the square lattice, there are other periodic self-dual graphs, which, for the same reason, have a bond percolation threshold equal to one-half. Figure 2 shows additional self-dual graphs, and illustrates a construction of an infinite family of periodic self-dual graphs, given in [20].

Essentially all other exact bond threshold solutions are derived using a relationship called the star-triangle transformation, first used by Sykes and Essam. Notice that the set of edges of the triangular lattice may be decomposed into triangles that are similarly oriented. If each triangle is replaced by a three-pointed star with the points at the vertices of the triangle, the resulting graph is the hexagonal lattice, which is the dual graph of the triangular lattice. If retention probability parameters of the two lattices can be found so that the probabilities of all possible events involving connections of the three vertices on the boundary of the triangle are equal, then the exact percolation threshold can be determined. The solution is the root of a polynomial involving the retention probability of the triangular lattice: $1 - 3p + p^3$, giving the solution $2\sin(\pi/18) \approx .347296$. By duality, the bond threshold for the hexagonal lattice is $1 - 2\sin(\pi/18) \approx .652704$. While this solution was derived by Sykes and Essam in 1964, mathematical methods to rigorously prove it were not developed until later, with the result being proved by Wierman in 1981.

Modified versions of the star-triangle transformation were used to derive other exact bond percolation thresholds. Wierman discovered a pair of lattices, called the bow-tie lattice and its dual, which could be solved exactly. The bond threshold of the bow-tie lattice is the root of $1 - p - 6p^2 + 6p^3 - p^5$, which is approximately .404518, while the dual lattice has threshold approximately .595482.

Scullard and Ziff [13,21,22] used a modified star-triangle approach to find values for a lattice that they named the martini lattice, and applied the approach to other planar two-dimensional graphs. They use a triangle-triangle transformation in the derivation of their results. While their approach does produce correct exact percolation threshold results for some graphs, further study is needed to determine the complete range of validity of their method.

For any of the exact bond threshold solutions, additional exact thresholds may be determined for certain transformations of the graphs.

If each edge of a graph G is replaced by k edges in series, the resulting graph is called a k-subdivision of G. Since the k edges in series play the role of one edge of G, the bond percolation threshold of a k-subdivision of G is the kth root of the bond threshold of G.

More generally, instead of replacing each edge by a series of edges, one may replace it by some finite graph connecting only the two endpoints, which is called a decoration by Ord and Whittington [10]. By calculating the edge retention probability that makes the probability of connection through the decoration equal to the threshold of the original graph, the percolation threshold of the decorated graph may be exactly determined.

Site Models in Two Dimensions

A key concept for the understanding of exact percolation threshold solutions for two-dimensional site models is the idea of a matching pair of graphs, introduced by Sykes and Essam in 1964.

A pair of matching graphs is constructed from an underlying planar graph. Select a set F of faces of the graph. Construct a graph G by adding an edge between any pair of vertices in each face of F that are not already connected by an edge. Construct the matching graph $M(G)$ of G by adding an edge between any pair of vertices in each face not in F that are not already connected by an edge. Note that $M(M(G)) = G$. Note also that if the underlying planar graph has faces with more than three sides, then at least one of the graphs in the matching pair is nonplanar.

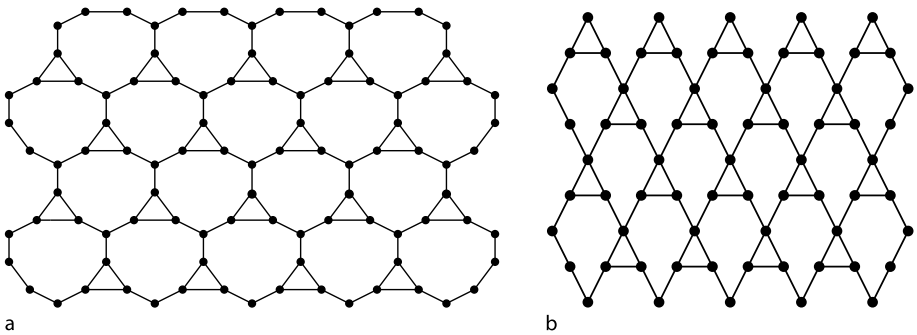

Percolation Thresholds, Exact, Figure 5
Portions of the martini lattice (a) and martini-A lattice (b)

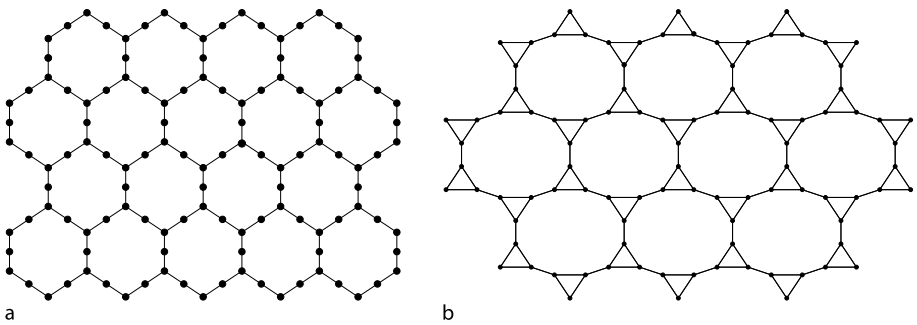

Percolation Thresholds, Exact, Figure 6
Portions of the 2-subdivision of the hexagonal lattice (a) and its line graph, the $(3, 12^2)$ lattice (b)

Percolation Thresholds, Exact, Table 1
Exact bond percolation thresholds of selected lattice graphs

Lattice	Bond Threshold	Equation
Triangular $= (3^6)$	$2\sin\pi/18 \approx .347296$	$p^3 - 3p + 1 = 0$
Bow-tie	.414518	$1 - p - 6p^2 + 6p^3 - p^5 = 0$
Square $= (4^4)$	.500000	$2p - 1 = 0$
Self-dual	.500000	$2p - 1 = 0$
D(Bow-tie)	.595482	$1 - p_c$(Bow-tie)
Martini-A	.625457	$p^5 - 4p^4 + 3p^3 + 2p^2 - 1 = 0$
Hexagonal $= (6^3)$	.652704	$1 - p_c(3^6)$
Martini	$1/\sqrt{2} \approx .707107$	$(2p^2 - 1)(p^4 - 3p^3 + 2p^2 + 1) = 0$

The importance of matching graphs is that the site percolation thresholds of a pair of periodic matching graphs sum to one, as proved by Kesten [8] in 1982. In fact, this result implied the result for bond percolation thresholds for dual graphs, via the bond-to-site transformation.

In general, additional information besides the matching property is needed to identify the exact percolation thresholds of the pair of graphs. However, if the matching graphs are identical, the graph is called self-matching, and the percolation threshold is necessarily one-half. Note that if a planar graph has all triangular faces, then it is self-matching. Therefore, the site percolation threshold is exactly one-half for the triangular lattice and the dual graphs of the $(4, 8^2)$, $(4, 6, 12)$, and $(3, 12^2)$ lattices. In addition, there are self-matching graphs that are not fully-triangulated, such as the line graph of the square lattice, shown in Fig. 7.

As a caution, however, note that a fully-triangulated graph may not have its site percolation threshold equal to one-half if it is not a periodic graph. An example of a fully-triangulated graph with site percolation threshold equal to one was given by van den Berg [16], and further discussion

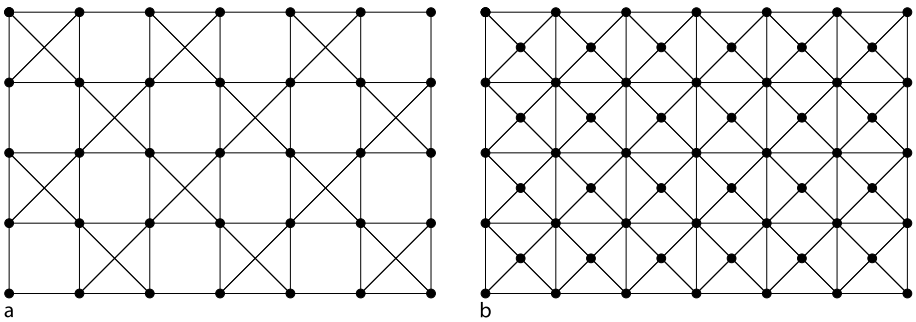

Percolation Thresholds, Exact, Figure 7
Two self-matching lattices: the line graph of the square lattice (a) and the dual of the $(4, 8^2)$ lattice (b)

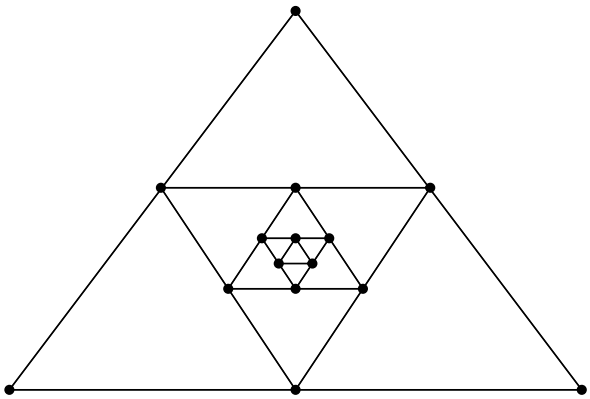

Percolation Thresholds, Exact, Figure 8
A portion of van den Berg's counterexample: The graph is fully-triangulated, and thus self-matching, but has percolation threshold equal to one – not one-half

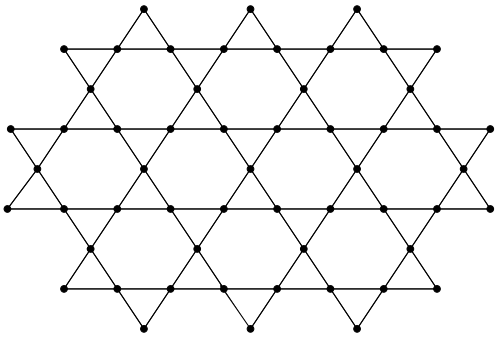

Percolation Thresholds, Exact, Figure 9
A portion of the Kagomé lattice, which is the line graph of the hexagonal lattice

Percolation Thresholds, Exact, Table 2
Exact bond percolation thresholds of selected lattice graphs

Lattice	Site Threshold	Equation
Triangular $= (3^6)$	.500000	$2p - 1 = 0$
Self -matching	.500000	$2p - 1 = 0$
Kagomé $= (3, 6, 3, 6)$	.652704	$p = p_c(6^3\text{bond})$
$(3, 12^2)$	.807901	$p = \sqrt{p_c(6^3\text{bond})}$

of similar counterexamples is provided in Wierman [18].

Scullard and Ziff [13,21,22] have also proposed exact site percolation thresholds for additional two-dimensional lattices.

Additional exact site percolation threshold solutions have been obtained by transformations of bond models with exact solutions. The line graph, $L(G)$, of a graph G is constructed by placing a vertex of $L(G)$ on each edge of G and connecting two vertices of $L(G)$ if the corresponding edges of G share a common endpoint. If there is a bond percolation model on G with each edge retained with probability p independently of the other edges, one may define a site percolation model on $L(G)$ in which each vertex is retained if and only if the corresponding edge of G is retained. Then, an infinite cluster in the bond model on G corresponds to an infinite cluster in the site model on $L(G)$, so the percolation thresholds of the two models are equal. This construction and relationship is called the bond-to-site transformation, and allows translation of all exact bond percolation threshold solutions into exact site percolation threshold solutions on line graphs. For example, the line graph of the hexagonal lattice, called the Kagomé lattice, has a site percolation threshold of exactly $1 - 2\sin(\pi/18) \approx .652704$, while its matching graph, the line graph of the triangular lattice, has exact site percolation threshold $2\sin(\pi/18) \approx .347296$. As another example, the site percolation threshold of the $(3, 12^2)$ lattice is exactly $\sqrt{1 - 2\sin(\pi/18)} \approx .807901$, since it is the line graph of the 2-subdivision of the hexagonal lattice.

Random Voronoi Percolation

In 2006, Bollobás and Riordan [2] provided the first exact percolation threshold solution for a continuum percolation model.

Consider the set of points P in a two-dimensional homogeneous Poisson point process. For a point $p \in P$, the Voronoi polygon of p is the set of all points in the plane that are closer to p than to any other point in P. Each Voronoi polygon is a convex polygon, and two Voronoi polygons either intersect in an edge or not at all. The edges of the collection of Voronoi polygons form an infinite planar graph called the Voronoi tessellation corresponding to P.

The dual of the Voronoi tessellation is called the Delaunay triangulation, since with probability one all faces are triangles. The remarkable result of Bollobás and Riordan [2] is that the site percolation threshold of the Delaunay triangulation is exactly one-half.

Multiparameter Critical Surfaces

For some applications, multi-parameter percolation models are considered. For example, in a bond percolation model, edges in different directions may have different retention probability parameters, giving a multi-dimensional parameter space. In such a parameter space, the role of the percolation threshold is played by the boundary between regions of the parameter space where infinite clusters occur and where all clusters are finite, called the critical surface.

Some multi-parameter bond percolation models are exactly solved. Two notable examples are: (1) In the square lattice with vertical edges retained with probability p and horizontal edges retained with probability q, the critical surface is the line segment $p + q = 1$. (2) For the triangular lattice with retention probabilities r, s, and t for bonds in the three different directions, the critical surface is the surface given by $1 - r - s - t + rst = 0$.

Future Directions

As seen above, the exact bond or site percolation threshold is known for relatively few lattices, with the exact solutions restricted to infinite trees and two-dimensional periodic graphs. Although they have been studied extensively, exact thresholds are not known for such common graphs as the square lattice site model, the hexagonal lattice site model, and the Kagomé lattice bond model. Although there are tools such as duality and matching, there is no general method available for providing exact threshold values. The grand challenge is to find the exact bond or site percolation threshold for a lattice in three dimensions or higher.

Bibliography

Primary Literature

1. Broadbent SR, Hammersley JM (1957) Percolation processes. I. Crystals and mazes. Proc Camb Philos Soc 53:629–641
2. Bollobás B, Riordan O (2006) The critical probability for random Voronoi percolation in the plane is $\frac{1}{2}$. Probab Theory Relat Fields 136:417–468
3. Bollobás B, Riordan O (2006) A short proof of the Harris-Kesten Theorem. Bull Lond Math Soc 38:470–484
4. Fortuin C, Kasteleyn PW, Ginibre J (1971) Correlation inequalities on some partially ordered sets. Commun Math Phys 22: 89–103
5. Grünbaum B, Shephard GC (1987) Tilings and Patterns. WH Freeman, New York
6. Harris TE (1960) A lower bound for the critical probability in a certain percolation process. Proc Camb Philos Soc 56:13–20
7. Kesten H (1980) The critical probability of bond percolation on the square lattice equals $\frac{1}{2}$. Commun Math Phys 74:41–59
8. Kesten H (1982) Percolation Theory for Mathematicians. Birkhäuser, Boston
9. Lyons R (1990) Random walks and percolation on trees. Ann Probab 18:931–958
10. Ord G, Whittington SG (1980) Lattice decorations and pseudo-continuum percolation. J Phys A: Math Gen 13:L307–L310
11. Russo L (1978) A note on percolation. Z Wahrscheinlichkeitstheorie Verwandte Geb 43:39–48
12. Russo L (1981) On the critical percolation probabilities. Z Wahrscheinlichkeitstheorie Verwandte Geb 56:229–237
13. Scullard CR (2006) Exact site percolation thresholds using a site-to-bond transformation and the star-triangle transformation. Phys Rev E 73:016107
14. Seymour PD, Welsh DJA (1978) Percolation probabilities on the square lattice. Ann Discret Math 3:227–245
15. Sykes MF, Essam JW (1964) Exact critical percolation probabilities for site and bond problems in two dimensions. J Math Phys 5:1117–1127
16. van den Berg J (1981) Percolation theory on pairs of matching lattices. J Math Phys 22:152–157
17. Wierman JC (1981) Bond percolation on the honeycomb and triangular lattices. Adv Appl Probab 13:298–313
18. Wierman JC (1984) Counterexamples in percolation: the site percolation critical probabilities p_H and p_T are unequal for a class of fully triangulated graphs. J Phys A: Math Gen 17: 637–646
19. Wierman JC (1984) A bond percolation critical probability determination based on the star-triangle transformation. J Phys A: Math Gen 17:1525–1530
20. Wierman JC (2006) Construction of infinite self-dual graphs. Proceedings of the 5th Hawaii International Conference on Statistics, Mathematics and Related Fields (CD-ROM). East West Council for Education, Honolulu
21. Ziff RM (2006) Generalized cell – dual-cell transformation and exact thresholds for percolation. Phys Rev E 73:016134
22. Ziff RM, Scullard CR (2006) Exact bond percolation thresholds in two dimensions. J Phys A: Math Gen 39:15083–15090

Books and Reviews

Bollobás B, Riordan O (2006) Percolation. Cambridge University Press, Cambridge

Grimmett G (1999) Percolation, 2nd edn. Springer, Berlin

Hughes BD (1996) Random Walks and Random Environments, vol 2: Random Environments. Oxford Science Publications, Oxford

Periodic Orbits of Hamiltonian Systems

Luca Sbano
Mathematics Institute, University of Warwick, Warwick, UK

Article Outline

Glossary
Definition
Introduction
Periodic Solutions
Poincaré Map and Floquet Operator
Hamiltonian Systems with Symmetries
The Variational Principles and Periodic Orbits
Further Directions
Acknowledgments
Bibliography

Glossary

Hamiltonian are called all those dynamical systems whose equations of motion form a vector field X_H defined on a symplectic manifold $(\mathcal{P}, \omega)$, and X_H is given by $i_{X_H}\omega = \mathrm{d}H$, where $H\colon \mathcal{P} \to \mathbb{R}$ is the Hamiltonian function.

Poisson systems These are dynamical systems whose vector field X_H can be described through a Poisson structure (Poisson brackets) defined on the ring of differentiable functions on a given manifold that is not necessarily symplectic (see "Hamiltonian Equations"). Note that on any symplectic manifold there is a natural Poisson structure such that any Hamiltonian system admits a Poisson formulation, but the contrary is false. The Poisson formulation of the dynamics is a generalization of the Hamiltonian one.

A periodic orbit $\phi(.)$ is a solution of the equations of motion that repeats itself after a certain time $T > 0$ called a *period*, that is, $\phi(t + T) = \phi(t)$ for every t.

Poincaré section/map Given a periodic orbit $\phi(.)$ a Poincaré section is a hyperplane S intersecting the curve $\{\phi(t)\colon t \in [0, T)\}$ transversely. The associated Poincaré map Π maps neighborhoods of S into itself by following the orbit $\phi(.)$ (see Definition 10).

A Hamiltonian system with symmetry is a Hamiltonian system in which there is a group G acting on $\mathcal{P}$, i. e., there is a map $\Phi\colon G \times \mathcal{P} \mapsto \mathcal{P}$, with Φ preserving the Hamiltonian and the symplectic form.

Relative periodic orbit Let G be a symmetry group for the dynamics. A path $\phi(.)$ is a relative periodic orbit if solves the equations of motion and repeats itself up to a group action after a certain time $T > 0$, that is, $\phi(t + T) = \Phi_g(\phi(t))$ for every t and for some $g \in G$.

Continuation Continuation is a procedure based on the implicit function theorem (IFT) that allows one to extend the solution of an equation for different values of the parameters. Let $f(x, \epsilon) = 0$ be an equation in $x \in \mathbb{R}^n$ where f is differentiable and $\epsilon \geq 0$ a parameter. Assume that $f(x_0, 0) = 0$; a curve $x(\epsilon)$ is called a *continued solution* if $x(0) = x_0$ and $f(x(\epsilon), \epsilon) = 0$ for some $\epsilon \geq 0$. In general $x(\epsilon)$ exists whenever the IFT can be applied, that is, if $D_x f(x, \epsilon)$ is invertible at $(x_0, 0)$.

Liapunov–Schmidt reduction Let f be a function on a Banach space. Liapunov–Schmidt reduction is a procedure that allows one to study $f(x, \epsilon) = 0$ under the condition that the kernel of Df is not empty but it is finite-dimensional.

Variational principles The principles which aim to translate the problem of solving the equations of motion of a dynamical system (e. g., Hamiltonian systems) into the problem of finding the critical points of certain functionals defined on spaces of all possible trajectories of the given system.

Definition

The study of periodic motions is very important in the investigations of natural phenomena. In particular the Hamiltonian formulation of the laws of motion has been able to formalize and solve many fundamental problems in mechanics and dynamical systems. This paper is focused on a selection of results in the study of periodic motions in Hamiltonian systems. We shall consider local problems (e. g., stability and continuation/bifurcation) and also the application of variational methods to study the existence in the large. Throughout the paper we give some details of the methods and proofs.

Introduction

Periodic motions and behaviors in Nature have always been of interest to mankind. All phenomena that have some cyclic nature have captured our attention because

they are a sign and a clue for regularity. Therefore, they are indications of the possibility of understanding the laws of Nature. Since the second century BC, Greek philosophers and astronomers have looked into the possibility of describing the motions of celestial bodies through the theory of epicycles, which are combinations of circular periodic motions. Notably from a modern point of view this theory can be interpreted as a clever geometrical application of Fourier expansions of the observed motions [23]. The development of mechanics, the discovery that the laws of Nature can be written in the language of calculus and that laws of motion can be described in terms of differential equations opened up the study of periodic solutions of equations of motion. In particular, since Newton and then Poincaré [46], the main interest has been the understanding of the planetary motions and the solution of the so-called *N-body problem*, the reader should refer to ► *n*-Body Problem and Choreographies in this encyclopedia. The interest in periodic motions has not been restricted to celestial mechanics but became a sort of paradigm in all areas where mechanics was successfully applied. In this article we shall illustrate some general aspects of the results regarding the theory of periodic orbits in Hamiltonian systems. Such systems are the modern formulation of those mechanical systems which are described by second-order differential equations and have an energy function. As an example the reader could think of Newton's equations for a point mass in potential field. The equations read

$$\begin{aligned} &m\,\ddot{x}(t) = -\nabla V(x(t))\,, \quad \text{with } x(t) \in \mathbb{R}^3 \\ &\text{for every } t\,, \quad m \text{ is the mass} \end{aligned} \tag{1}$$

$V(x)$ is the potential and $\nabla = (\partial/\partial x_1, \partial/\partial x_2, \partial/\partial x_3)$. The energy function

$$E = \frac{m}{2}\|\dot{x}(t)\|^2 + V(x(t))$$

is conserved along the trajectories solving (1). It is important to say that most of the systems of interest in physics can be naturally written in Hamiltonian form. The plan of this article is as follows. First we introduce the Hamiltonian formulations of the equations of motion for a classical mechanical system. It is well known that in many applications Hamiltonian systems derive from a Lagrangian formulation; therefore, this is also presented. Furthermore we introduce the Poisson formulation that is essentially the first generalization to the Hamiltonian point of view. Then we turn to the study of the properties of periodic solutions. In particular we focus on their "local properties", persistence and stability. In this analysis the main tool will be the implicit function theorem (IFT). In order to emphasize the utility of the IFT we present some of the proofs that contain typical calculations often scattered in the literature. Then we consider the problem of periodic orbit for Hamiltonian systems with symmetries, where we introduce the notion of relative equilibrium and relative periodic orbit. Inevitably we have also a short excursion about symmetry reduction, which is the natural theoretical setting to study systems with symmetries. The second part of the article is devoted to the exposition of the study of periodic orbits by variational methods. The discovery that the equations of motion of mechanical systems can be derived by a variational principle, the so-called *least action principle*, is usually attributed to Maupertuis (eighteenth century). According to this principle, the motions are critical points of a functional called the *action* defined in a suitable space of paths. Variational methods turned out to be one of the most effective methods to prove the existence of periodic orbits; notable is the case of the N-body problem (see [2] and ► *n*-Body Problem and Choreographies). The valuable feature of the variational methods is the possibility to study the existence problem by looking at the topology and geometry of the space of periodic paths without further restrictions. In the presentation of the results some elements of the proofs are illustrated in order to clarify the main ideas. In the final section there are some open problems and further directions of investigation, in particular a simple example of the so-called *multisymplectic* structures that has extended the possibility of applying the finite-dimensional Hamiltonian approach to multiperiodic problems for a large class of partial differential equations is presented. For centuries the study of periodic orbits has been one of the main centers of mathematical investigations and developments and still presents challenges and the capacity for producing new interesting mathematical ideas to understand the complexity of Nature.

Hamiltonian Equations

A Hamiltonian system is given by specifying a symplectic manifold $(\mathcal{P}, \omega)$, where $\mathcal{P}$ is a differentiable manifold of even dimension, ω is a closed differential two-form and a function $H\colon \mathcal{P} \to \mathbb{R}$ is called a Hamiltonian. In the language of the differential forms the Hamiltonian vector X_H field on $\mathcal{P}$ is written as

$$i_{X_H}\omega = \mathrm{d}H\,. \tag{2}$$

In the case $\mathcal{P} = \mathbb{R}^{2n}$, the symplectic structure is $\omega_0 = \sum_{i=1}^{n} \mathrm{d}x_i \wedge \mathrm{d}y_i$ and the Hamiltonian vector field is

$$X_H(z) \doteq J\,\nabla_z H(z)\,, \tag{3}$$

where $z = (x, y) \in \mathbb{R}^{2n}$ and J is the symplectic matrix

$$J = \begin{pmatrix} 0 & \mathbf{id}_n \\ -\mathbf{id}_n & 0 \end{pmatrix} . \tag{4}$$

The Hamiltonian equations are then

$$\frac{\mathrm{d}z(t)}{\mathrm{d}t} = X_H(z(t)) . \tag{5}$$

In the case $\mathcal{P} = \mathbb{R}^{2n}$ (5) reads

$$\dot{z}(t) = (\dot{x}(t), \dot{y}(t)) = \big(\nabla_y H(x(t), y(t)), -\nabla_x H(x(t), y(t))\big) .$$

For a mechanical systems described by (1) the Hamiltonian function is

$$H = \frac{\|y\|^2}{2m} + V(x) , \quad \text{where} \quad (x, y) \in \mathbb{R}^6 ,$$

and its Hamiltonian equations read

$$\begin{cases} \dot{x}(t) = \dfrac{y(t)}{m} \\ \dot{y}(t) = -\nabla V(x(t)) . \end{cases}$$

Note that the first equation corresponds to the classical definition of *momentum* in mechanics (here denoted with y) and the Hamiltonian function H coincides with the energy E. For more details the reader could consult [1,5] and also ▶ Dynamics of Hamiltonian Systems in this encyclopedia.

Lagrangian Formulation In many applications in physics and in particular in mechanics Hamiltonian systems arise from the Lagrangian description. In such a setting a mechanical system is described by prescribing a differentiable manifold $\mathcal{M}$ (the *configuration space*) and a Lagrangian function L defined on the tangent bundle $T\mathcal{M}$. Let $L\colon T\mathcal{M} \to \mathbb{R}$ be a Lagrangian on a manifold $\mathcal{M}$ of dimension n. If L is hyperregular (i. e., $\operatorname{rank}(D^2_{v_q} L(v_q, q)) = n$), then the Hamiltonian function is naturally constructed on the cotangent bundle $T^*\mathcal{M}$ by using the Legendre transform [1,5] as follows:

$$\begin{aligned} p_i &= \frac{\partial L}{\partial v_q^i} , \\ H(p, q) &= \sum_{i=1}^{n} v_q^i(p)\, q_i - L(v_q(p, q), q) . \end{aligned} \tag{6}$$

The Hamiltonian system is then defined on the cotangent bundle of $\mathcal{M}$ that is $\mathcal{P} = T^*\mathcal{M}$, which is endowed with the canonical symplectic form $\omega = \mathrm{d}\theta$, where $\theta = \sum_{i=1}^{n} p_i \mathrm{d}q_i$. In the Lagrangian description the equations of motion are

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t} \frac{\partial L(v_q(t), q(t))}{\partial v_{q_i}} - \frac{\partial L(v_q(t), q(t))}{\partial q_i} = 0 , \\ &i = 1, \ldots, n \text{ where } v_{q_i}(t) = \dot{q}_i(t) , \quad i = 1, \ldots, n . \end{aligned} \tag{7}$$

Note that (7) contains second order time-derivatives. For more details see [1,5].

Poisson Formulation Let $\mathcal{F}(\mathcal{P})$ be the space of differentiable functions on $(\mathcal{P}, \omega)$. On $\mathcal{F}(\mathcal{P})$ can be introduced a product $\{.,.\}$ [1,5,16]. The Poisson brackets

$$\omega(X_f, X_g) = \{f, g\} \quad \text{for } f, g \in \mathcal{F}(\mathcal{P}) . \tag{8}$$

In terms of the Poisson brackets the Hamiltonian equations can be written as a derivation acting on $\mathcal{F}(\mathcal{P})$:

$$X_H(f) = \{f, H\} \quad \text{for } f \in \mathcal{F}(\mathcal{P}) . \tag{9}$$

The Poisson brackets satisfy the following properties. For all $f, g \in \mathcal{F}(\mathcal{P})$

$$\begin{aligned} &\{f, g\} \text{ is bilinear with respect to } f \text{ and } g , \\ &\{f, g\} = -\{g, f\} \\ &\{fg, h\} = f\{g, h\} + g\{f, h\} \\ &\{\{f, g\}, h\} + \{\{h, f\}, g\} + \{\{g, h\}, f\} = 0 \text{ Jacobi identity} \end{aligned} \tag{10}$$

An easy consequence of (9) and (10) is

Proposition 1 *The Hamiltonian function H is a constant of motion.*

A manifold $\mathcal{P}$ endowed with the brackets $\{.,.\}$ is called a *Poisson manifold*. Any symplectic manifold is a Poisson manifold [1,27] but the contrary is false. In fact Poisson brackets on a symplectic manifold are always nondegenerate, namely, the condition $\{k, f\} = 0$ for all $f \in \mathcal{F}(\mathcal{P})$ implies that k is identically zero. In a general Poisson manifold there might exists nonvanishing k, which are then called *Casimir* functions. This is related to the fact that symplectic manifolds are always even-dimensional, whereas Poisson manifolds can be odd-dimensional. We can look at Poisson manifolds as a useful generalization of Hamiltonian systems; in fact in order to define Poisson brackets it is sufficient to have the ring of functions $\mathcal{F}(\mathcal{P})$. For a general and complete description of the Poisson structure the reader could see [1,5,16,27].

Periodic Solutions

Given a Hamiltonian vector field $X_H(z)$ on $(\mathcal{P}, \omega)$, one can consider the following Cauchy problem:

$$\begin{cases} \dfrac{\mathrm{d}z(t)}{\mathrm{d}t} = X(z(t)) \\ z(0) = z_0 \,. \end{cases} \tag{11}$$

Equation (11) is meant to be defined on a local chart in $\mathcal{P}$.

Definition 1 We call *flow* or *integral flow* the map $t \mapsto \phi(t, z_0)$ where $z(t) = \phi(t, z_0)$ solves problem (11).

Some simple consequences follow [36].

Remark 1 If X_H is complete, then the flow is defined for all $t \in \mathbb{R}$.

Remark 2 If the Hamiltonian vector field is autonomous, (that is, not explicitly dependent on time), then $\phi(., z_0)$ satisfies the composition property $\phi(t, \phi(s, z_0)) = \phi(t + s, z_0)$ and $\phi(., .)$ is called Hamiltonian flow.

Let us now introduce the main object of this exposition.

Definition 2 A flow $\phi : \mathbb{R} \to \mathcal{P}$, $z(t) = \phi(t, z_0)$ is said to be a T-periodic solution of (11) if there exists $T > 0$ such that $\phi(t + T, z_0) = \phi(t, z_0)$ for all t.

Remark 3 Note that if $\phi(., z_0)$ is a T-periodic solution, then $\phi(., z_0)$ is $n\,T$-periodic for any $n \in \mathbb{N}$. In fact from the definition $\phi(T, z_0) = z_0$ and using Remark 2, one can iterate

$$\begin{aligned} \phi(t + n\,T, z_0) &= \phi(t + (n-1)T, \phi(T, z_0)) \\ &= \phi(t + (n-1)T, z_0) \end{aligned}$$

and find $\phi(t + nT, z_0) = \phi(t, z_0)$.

Definition 3 T is called a minimal period of $\phi(t, z_0)$ if $T = \min_{\tau \in \mathbb{R}_+} \{\tau : \phi(t + \tau, z_0) = \phi(t, z_0) \text{ for all } t\}$.

Definition 4 A point $z^* \in \mathcal{P}$ such that $J\,\nabla_z H(z^*) = 0$ is called an equilibrium solution.

Obviously any equilibrium solution can be seen as a periodic solution with $T = 0$. One can easily show

Lemma 1 *$\phi(t, z_0)$ is periodic of period T if and only if $\phi(T, z_0) = z_0$.*

Stability of Periodic Orbits

Given a periodic orbit the first natural question is to study its stability. There are three possible stability criteria [1,36]:

Definition 5 (Liapunov-stable) A periodic orbit $\phi(., z_0)$ is Liapunov-stable if for all $\epsilon > 0$ there is $\delta(z_0, \epsilon)$ such that $\|z_0 - z'\| \le \delta(z_0, \epsilon)$ implies that $\|\phi(t, z_0) - \phi(t, z')\| \le \epsilon$ for all $t \ge 0$.

The previous definition is natural for a non-Hamiltonian system but in a Hamiltonian context it is very strong, since in the Hamiltonian systems periodic orbits are not isolated. It is useful though to compare Liapunov stability with the following weaker notions.

Definition 6 (Spectrally stable) A periodic orbit $\phi(t, z_0)$ is spectrally stable if the eigenvalues of $DX_H(z_0)$ lie all on the unit circle.

Definition 7 (Linearly stable) A periodic orbit $\phi(t, z_0)$ is linearly stable if it is spectrally stable and $DX_H(z_0)$ can be diagonalized.

One can show that spectral stability is implied by either linear stability or Liapunov stability. A natural notion of stability can be introduced by using the Poincaré map and will be presented in Sect. "Poincaré Map and Floquet Operator". The following results describe the structure of linear Hamiltonian systems, namely, systems whose Hamiltonian function is $H(z) = \frac{1}{2}\langle z, A\,z\rangle$ and the equations of motion read $\dot{z}(t) = J\,A\,z(t)$.

Theorem 1 ([36]) *Let A be time-independent. The characteristic polynomial of $J\,A$ is even and if λ is an eigenvalue then so are $-\lambda, \overline{\lambda}, -\overline{\lambda}$.*

A consequence of the previous result is that linear stability is equivalent to spectral stability for linear autonomous Hamiltonian systems.

Linear systems can depend on parameters; it is therefore interesting to have a notion of stability with respect to parametric changes.

Definition 8 A linear stable Hamiltonian system $H(z) = \frac{1}{2}\langle z, A\,z\rangle$ is said to be parametrically stable if for every symplectic matrix B such that $\|A - B\| < \epsilon$ the system $\dot{z} = J\,B\,z$ is linearly stable.

We finally give a characterization of linear Hamiltonian systems which are parametrically stable.

Theorem 2 ([36]) *If the Hamiltonian H is positive (or negative) definite or all the eigenvalues are simple, then A is parametrically stable.*

Continuation and Bifurcation of Equilibrium Solutions

Let $H(z, \epsilon)$ be a Hamiltonian function depending on a parameter $\epsilon \ge 0$. Let

$$Z_0 = \{z^* : J\nabla H(z^*, 0) = 0\}$$

be the set of equilibrium positions. An interesting problem is to study how Z_0 is modified when $\epsilon > 0$. The local properties of Z_0 depend on the spectrum of $J\,D^2H(z^*, 0)$. If $z^* \in Z_0$ is such that $JD^2H(z^*, 0)$ has nonzero eigenvalues, then by the IFT there exists a curve $z^*(\epsilon)$ of equilibrium points for ε positive and sufficiently small. If $J\,D^2H(z^*, 0)$ has at least one zero eigenvalue then bifurcation can occur, but the interesting point is that bifurcations are possible also when $J\,D^2H(z^*, 0)$ is not degenerate. In fact in Hamiltonian systems generically the spectrum of the linearization of the vector field contains couples of complex conjugated eigenvalues [5]. In such a case there is a theorem due to Liapunov which shows the existence of periodic orbits – so-called *nonlinear normal modes* in a neighborhood of z^*. Let us now present Liapunov's theorem. This result describes how a nondegenerate equilibrium can be continued into a periodic orbit. The types of orbits are called *nonlinear normal modes*.

Theorem 3 (Liapunov's center theorem [1,18]) *If the Hamiltonian system has a nondegenerate equilibrium at which the linearized vector field has eigenvalues* $\pm i\omega, \lambda_3, \ldots, \lambda_n$ *with* $\lambda_k/\omega \notin \mathbb{Z}$ *then there exists a one-parameter family of periodic orbits emanating from* z^*. *The period tends to* $2\pi/\omega$ *when the orbit radius tends to zero and the nontrivial multipliers tend to* $\exp(2\pi\lambda_k/\omega)$ *with* $k = 3 \ldots n$.

Proof Without loss of generality the nondegenerate equilibrium can be fixed at the origin. In a neighborhood of the origin the Hamiltonian vector field can be written as follows:

$$\dot{z}(t) = JA\,z(t) + r(z)\ ,$$

where $\|r(z)\| = o(\|z\|)$, that is, $\|r(z)\|$ is infinitesimal with respect to $\|z\|$.

The spectrum of JA is $\mathrm{Spec}(JA) = \{\pm i\omega, \lambda_3, \ldots, \lambda_n\}$. Let $y = \epsilon\, z$ with $\epsilon \in [0, 1]$, then

$$\dot{y}(t) = JA\,y(t) + r(y, \epsilon)\ ,$$

with $r(y, \epsilon)$ infinitesimal for $\epsilon \to 0$ uniformly for $\|y\|$ bounded. For $\epsilon = 0$ the system becomes

$$\dot{y}(t) = JA\,y(t) \tag{12}$$

and admits a periodic solution with period $T = 2\pi/\omega$:

$$y_0(t) = \exp(t\,JA)\,y_0,\ \ JA\,y_0 = \nu_0\,y_0\ .$$

Equation (12) is linear, and thereby coincides with its linearization. The Floquet multipliers are therefore

$$(1, 1, \exp(2\pi\,\lambda_3/\omega), \ldots, \exp(2\pi\,\lambda_n/\omega))$$

with $\exp(2\pi\,\lambda_k/\omega) \neq 1$ by hypothesis. Now we look for a solution of

$$\left(\frac{\mathrm{d}}{\mathrm{d}t} - JA\right) y(t) = r(y(t), \epsilon)\ . \tag{13}$$

On the space $C^1([0, T], \mathbb{R}^{2n})$ with periodic boundary condition the operator $\mathrm{d}/\mathrm{d}t - JA$ has a two-dimensional kernel. Therefore, one could solve (13) by looking for a solution in the form

$$y(t, \epsilon) = \exp(t\,JA)y_0 + u(t)\ ,$$

where $u(t) \in \mathrm{rank}(\mathrm{d}/\mathrm{d}t - JA)$. By the IFT one can show that $\|u(t)\| = o(1)$ and therefore the solution can be continued for small ϵ as

$$y(t, \epsilon) = \exp(t\,JA)(\mathbf{1} + o(1))\,y_0\ ,$$

that is,

$$z(t, \epsilon) = \epsilon\,\exp(t\,JA)(\mathbf{1}+o(1))\,y_0 \text{ with } \lim_{\epsilon\to 0} \|z(t, \epsilon)\| = 0.$$

□

Remark 4 The analysis of the operator $\mathrm{d}/\mathrm{d}t - JA$ used to prove the previous result is known as Liapunov–Schmidt reduction. In Sect. "Continuation of Periodic Orbits as Critical Points" we shall give an application of it in the study of critical points.

Normal Form Analysis Near Equilibrium Points

In [55] Liapunov's theorem was generalized to cases where the condition $\lambda/\omega \in \mathbb{Z}$ might hold. This corresponds to the so-called *resonance condition*.

Definition 9 (Resonance) The set of eigenvalues $\{\omega_l\}_{l=1}^k$ of the linearization $DX_H(z_0)$ are said to be resonant if $\mathbb{R}(\omega_i) = 0$ for all i and there exist $\{n_l\}_{l=1}^k \subset \mathbb{Z}$ such that

$$\sum_{l=1}^{k} \omega_l\,n_l = 0\ .$$

In [55] it is shown that around z_0 there are at least n periodic orbits. Because of the possible presence of resonance not all the orbits are a continuation of periodic orbits of the linearized vector field. In order to study this case one has to go beyond the linear approximation and analyze the Hamiltonian system in a neighborhood of the equilibrium taking into account the structure of terms of order higher than 2 in the canonical coordinates. To achieve this objective there is a general method, the *normal form* theory, to

expand the Hamiltonian function H using suitable coordinates in an ϵ-neighborhood of z_0. A given Hamiltonian H can be expanded as

$$H(z) = H_0(z) + \sum_{m=1} \epsilon^m H_m(z) \,, \tag{14}$$

where

$$\begin{aligned} H_0(z) &= \sum_{j=1} \frac{\omega_i}{2} \|z_j - z_{0,j}\|^2 \\ &= \sum_{j=1} \frac{\omega_i}{2} [(x_j - x_{0,j})^2 + (y_j - y_{0,j})^2] \end{aligned}$$

and $H_m(z)$ are polynomials of degree $m+2$ in the canonical coordinates. Normal form theory allows us to classify the possible form of expansions (14) according to the resonance condition. This classification is also very important in the study of small perturbations of integrable Hamiltonian systems (*KAM theory*) where it is still the resonance condition that causes the main difficulties [5]. For a detailed account of normal form theory and its applications the reader could refer to [49,54] and to ▶ Dynamics of Hamiltonian Systems for an introduction. Once the Hamiltonian is in normal form, one can study the bifurcation occurring when the resonance holds. Particularly interesting is to understand what conditions the frequencies have to fulfill in order for a system to have a number of periodic orbits exceeding the estimation given in [55]. There have been further generalizations to the results given in [55] and the reader could refer to [40,41].

Poincaré Map and Floquet Operator

To study a periodic orbit one could consider looking at it in a hyperplane which is transverse to its direction. This can be done locally and by constructing the *Poincaré map*.

Definition 10 Let $\phi(., z_0)$ be a T-periodic orbit. A Poincaré cross section of $\phi(., z_0)$ is

$$S_a = \{z \in \mathcal{P} \colon \langle a, (z - z_0)\rangle = 0, \text{ with } \langle a, J\nabla H(z_0)\rangle \neq 0\}.$$

Let $U \subset S_a$, a neighborhood of z_0 sufficiently small such that the first return time is

$$\mathcal{T} \colon U \to R_+, \; \phi(\mathcal{T}(z), z) \in S_a \text{ with } \mathcal{T}(z_0) = T \,.$$

The Poincaré map $\Pi \colon U \to S_a$ is defined as

$$U \ni z \to \Pi(z) \doteq \phi(\mathcal{T}(z), z) \in S_a \,.$$

Using the regularity properties of $\phi(t, z_0)$, one can show

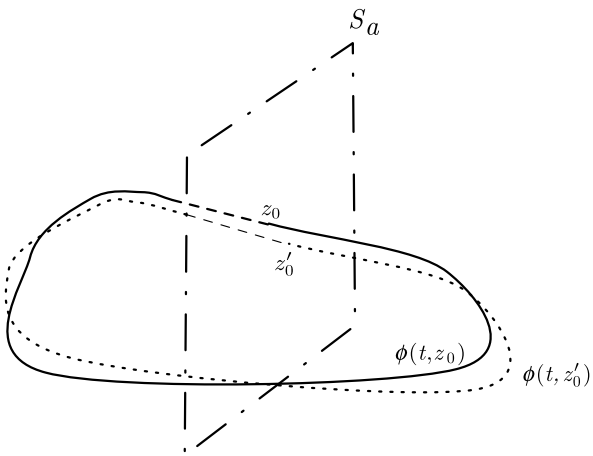

Periodic Orbits of Hamiltonian Systems, Figure 1
S_a is Poincaré section for $\phi(t, z_0)$, in this case $a = X_H(z_0)$. The continuation produces a new orbit with initial data z_0' close to z_0

Proposition 2 ([36]) *Let $\sigma(t, z) \doteq \langle a, \phi(t, z) - z\rangle$ be defined for $z \in S_a$. For $U \subset S_a$ sufficiently small the return times $\mathcal{T} \colon U \to R_+$ defined by the implicit equation $\sigma(\mathcal{T}(z), z) = 0$ and $\Pi \colon U \to S_a$ are smooth functions.*

Using this result, one can think of characterizing periodic orbits as fixed points of $\Pi(z) = \phi(\mathcal{T}(z), z)$. In fact if there exists $z^* \in S_a$ such that $\Pi(z^*) = z^*$, then the flow $\phi(t, z^*)$ is a periodic orbit with period $\mathcal{T}(z^*)$. An example of a Poincaré section is given in Fig. 1.

Definition 11 Let $\phi(t, z_0)$ be a periodic solution of (5). The Floquet operator

$$V(t) \doteq \frac{\partial \phi(t, z_0)}{\partial z_0}$$

solves the variational equation

$$\dot{V}(t) = D_z X_H(\phi(t, z_0))\, V(t) \tag{15}$$

with $V(0) = \mathbf{id}$. The matrix $V(T)$ is called a monodromy matrix and its eigenvalues are termed Floquet multipliers.

For a general discussion about Floquet theory one can refer to [36,49,59]. In general, the construction of the Poincaré map is not explicit. But knowledge of the monodromy matrix and Proposition 2 with the equation $\sigma(\mathcal{T}(z), z) = 0$ allow us to compute the Taylor expansion of $\Pi(z)$ in a neighborhood of z_0 in S_a. We shall examine this idea in the next section, where we also see the consequences of the fact that periodic orbits in Hamiltonian systems are not isolated. In fact the following result holds true:

Proposition 3 ([36]) *Periodic solutions are never isolated and $+1$ is always a multiplier whose eigenvector is $J\nabla H(z)$. If $F\colon \mathcal{P} \to \mathbb{R}$ is a first integral, then ∇F is a left eigenvector of the monodromy matrix with eigenvalue $+1$.*

Note that a consequence is that since in a Hamiltonian system the Hamiltonian H is always conserved, the multiplier $+1$ has algebraic multiplicity of at least 2.

In the presence of an integral of motion like the Hamiltonian, the Poincaré map can be restricted to the level set of H. The map Π turns out to be symplectic [5,36]. The stability of a periodic orbit can now be defined in terms of the stability of the fixed points of the Poincaré map. Let $\Pi^n(z)$ denote the nth iteration of the Poincaré map applied to $z \in S_a$. We can define

Definition 12 ([1,36,49]) A periodic orbit $\phi(., z_0)$ is stable for all $\epsilon > 0$ if there exists $\delta(\epsilon, z_0)$ such that

$$\|z - z_0\| < \delta \text{ implies } \|\Pi^n(z) - z_0\| < \epsilon \quad \text{for all} \quad n > 0 .$$

This stability criterion is difficult to verify. A weaker criterion is obtained by considering the linearization of $\Pi(z)$ at $z = z_0$:

Definition 13 ([1,36]) A periodic orbit $\phi(., z_0)$ is spectrally stable if the associated Poincaré map Π, which is restricted to the manifold defined by the integrals of motion, has a linearization $D_z\Pi(z_0)$ with a spectrum on the unit circle.

By a local change of coordinates one can show that the eigenvectors of $D_z\Pi(z_0)$ are equal to the eigenvectors of $V(T_0)$ different from $X_H(z_0) = J\nabla H(z_0)$ [36]. The Poincaré map removes the degeneracy of the monodromy matrix $V(T)$. To illustrate this point let us consider $x \in \mathbb{R}^n$ and an autonomous system

$$\dot{x}(t) = f(x(t)) \tag{16}$$

with $f\colon \mathbb{R}^n \mapsto \mathbb{R}^n$ differentiable. Let $\phi(t, x_0)$ be a T-periodic solution of (16) emanating from x_0. To study the trajectories near x_0 one can construct a local diffeomorphism $h\colon \mathbb{R}^n \mapsto \mathbb{R}^n$, $y = h(x)$ such that $y_0 = h(x_0)$ with (16) in the form

$$\begin{aligned} &\dot{y}(t) = Dh(h^{-1}(y(t)))\, f(h^{-1}(y(t))) = \tilde{f}(y(t)) \\ &\text{with} \quad \tilde{f}(y_0) = (1, 0, \ldots, 0) . \end{aligned} \tag{17}$$

In y coordinates the periodic orbit $\phi(t, x_0)$ reads $\psi(t, y_0) = h(\phi(t, h^{-1}(y_0)))$ and we can define a map σ as

$$\sigma(t, y) = \langle \tilde{f}(y_0), \psi(t, y_0) - y \rangle . \tag{18}$$

The form of $\tilde{f}(y_0)$ implies $\sigma(y, t) = \psi_1(t, y) - y_{0,1} = h_1(\phi(t, h^{-1}(y_0))) - y_{0,1}$. An easy calculation shows that

$$\left.\frac{\partial \sigma}{\partial t}\right|_{(0, y_0)} = \left.\sum_{l=1}^{n} \frac{\partial h_1}{\partial x_l} f_l(x)\right|_{(0, y_0)} = 1 ,$$

which implies the existence of a return time $\tau(y)$ satisfying $\sigma(\tau(y), y) = 0$ for y in a sufficiently small neighborhood of y_0. Now we define the map $\Pi(y) = \psi(\tau(y), y)$. By applying the IFT, we can compute

$$\frac{\partial \tau}{\partial y_j} = \delta_{1j} - \sum_{l,m} \frac{\partial h_1}{\partial x_l} \frac{\partial \phi_l}{\partial x_m} \frac{\partial x_m}{\partial y_j} \tag{19}$$

and also the Jacobian of Π:

$$\frac{\partial \Pi_i(y)}{\partial y_j} = \tilde{f}_i(y) \frac{\partial \tau}{\partial y_j} + \sum_{l,m} \frac{\partial y_i}{\partial x_l} \frac{\partial \phi_l}{\partial x_m} \frac{\partial x_m}{\partial y_j} . \tag{20}$$

Using (19), one can easily say that the matrix $\partial \Pi_i(y)/\partial y_j$ at $y = y_0$ has the first column equal to $\tilde{f}(y_0) = (1, 0, \ldots, 0)$. The diffeormophism h allowed us to "isolate" the direction of the vector field f at x_0. This implies that the map $\Pi_{S_{\tilde{f}(z_0)}}$, the restriction of Π on the Poincaré section $S_{\tilde{f}(z_0)} = \{y\colon \langle \tilde{f}(y_0), y - y_0 \rangle = 0\} \simeq \mathbb{R}^{n-1}$, maps $S_{\tilde{f}(z_0)}$ into $S_{\tilde{f}(z_0)}$ and its linearization has no eigenvectors in the direction of $\tilde{f}(y_0)$. The map $\Pi_{\tilde{f}(z_0)}$ describes the stability of the periodic orbit $\phi(t, x_0)$. A similar construction can be carried out when f is a Hamiltonian vector field. In that case it is necessary to take into account the existence of integrals of motion and construct the Poincaré map on the manifold where the integral of motions are fixed. A more complete study of the Poincaré map will be presented in the next section.

Let us now present some properties of the Floquet operator. Let us consider a Hamiltonian system with imaginary multipliers. In the linear time-independent case the monodromy matrix is given by

$$V(T_0) = \exp(T_0 JA) ,$$

where $T_0 = 2\pi/|\lambda|$, with λ an imaginary eigenvalue of JA. Now let $H = \frac{1}{2}\langle z, Az \rangle + h(z)$ be a Hamiltonian with $h(z) = o(\|z\|^2)$ near $z = 0$. The equation for the monodromy matrix is (15). Let $\phi_0(t)$ be another periodic orbit, then (15) can be written as follows:

$$\frac{\mathrm{d}V_\psi(s)}{\mathrm{d}s} = \frac{T_\phi}{T_0} DX_H(\psi_0(s))\, V_\psi(s) ,$$

where $t = T_0\, s/T_\phi$ and $\psi_0(s) = \phi_0(T_0\, s/T_\phi)$. Then one can show

Proposition 4 ([39]) *The Floquet operator $V_\psi(t)$ is C^∞ in ψ and T_ϕ.*

Corollary 1 ([39]) *If $\phi_0(t)$ is sufficiently close to $z = 0$, then $V_\phi(T_\phi)$ is arbitrarily close to* $\exp(T_0\, JA)$.

Corollary 1 is interesting because it allows us to use linearized dynamics. Now the analysis of the nonlinear stability can be carried out by using Krein's theory. For this typical references are [21,59]. Here we recall

Theorem 4 (Krein) *Let $V(T)$ be a spectrally stable monodromy matrix. Let Q_λ be a quadratic form $Q_\lambda(v) = \langle V(T)\, v, J\, v\rangle$ where v belongs to the λ eigenspace of $V(T)$. Then $V(T)$ is in an open set of spectrally stable matrices if and only if the quadratic Q_λ has definite sign.*

By combining Corollary 1 and Theorem 4, we could show that a solution $\phi_0(t)$ close enough to $z = 0$ is spectrally stable.

Continuation of Periodic Orbits in Hamiltonian Systems

In general it is difficult to prove the existence of and then construct periodic orbits. Sometimes, for certain specific values of the parameters characterizing the system it is possible to find particular solutions. Typically these are the equilibria and relative equilibria. In these cases the continuation method can be a useful approach. The idea is to look at how a given periodic orbit changes according to a small modification of the parameters. The method reduces the research of periodic orbits to the problem of finding fixed points of the Poincaré map that can be continued as a function of the parameters; see Fig. 1.

Definition 14 (Continuation of an orbit) Given a dynamical system and $\phi_0(t)$ one of its orbits, we say that $\phi_0(t)$ can be continued if there exists a family of orbits $\phi(t, \alpha)$ smoothly dependent on parameters α's and such that $\phi(t, 0) = \phi_0(t)$.

Let us consider a Hamiltonian vector field $X_H(z, \alpha)$ where $z \in \mathcal{P}$ and $\alpha \in \mathbb{R}^k$ are k parameters. We now write the equations of motion in a form where the period T appears explicitly. After $t \to t/T$ the equations read

$$\dot{z}(t) = T\, J\, \nabla H(z(t), \alpha) = T\, X_H(z(t), \alpha)\,, \tag{21}$$

with $t \in [0, 1]$.

Definition 15 Let $\phi(t, z, T, \alpha)$ be a solution of (21). The map

$$R(z, T, \alpha) \doteq \phi(1, z, T, \alpha) - z \tag{22}$$

is called a return map.

The orbit $\phi(t, z_0, \alpha_0)$ is T_0-periodic if $R(z_0, T_0, \alpha_0) = 0$. Now we are interested to see what is the fate of the orbit when z_0, T_0 and α_0 are varied; therefore, it is useful to determine the local behavior of the map R. This is collected in the following proposition.

Proposition 5 ([42]) *Let $\phi_0(t, z_0)$ be a periodic orbit with period T_0 and $\alpha = 0$, then the following relations hold*

- $D_z R(z_0, T_0, 0) = V(1) - \mathbf{id}$,
- $X_H(z_0, 0) \in \ker(V(1) - \mathbf{id})$,
- $D_T R(z_0, T_0, 0) = X_H(z_0, 0)$,
- $D_\alpha R(z_0, T_0, 0)$
  $= T_0\, V(T) \int_0^1 V^{-1}(s) D_\alpha X_H(\phi_0(s, z_0)) ds$,

together with the differential map

$$\begin{aligned} &DR(z_0, T_0, 0)(\xi, T, \alpha) \\ &= (V(1) - \mathbf{id}) \cdot \xi + T\, X_H(z_0, 0) + D_\alpha R(z_0, T_0, 0) \cdot \alpha\,, \end{aligned}$$

where $(\xi, T, \alpha) \in \mathbb{R}^{2n+k+1}$.

Initially let us consider a non-Hamiltonian dynamical system in $\mathbb{R}^n$:

$$\dot{x}(t) = f(x(t), \epsilon)\,, \quad \epsilon \in \mathbb{R}\,. \tag{23}$$

We now illustrate how the IFT is used to construct a new solution from a given one, i. e., by continuation. We present a detailed proof for the reader's convenience.

Proposition 6 ([3,42]) *Let $\Gamma_0 = \{\phi_{T_0}(t, x_0), t \geq 0\}$ be a periodic orbit of period 1 of $\dot{x}(t) = T_0\, f(x(t), \epsilon)$ for $\epsilon = 0$. If 0 is an eigenvalue of $D_x R(x_0, T_0, 0)$ with multiplicity 1, then orbit Γ_0 can be continued.*

Proof Let us consider the map $G\colon \mathbb{R}^n \times \mathbb{R}_+ \times [0, 1] \to \mathbb{R}^{n+1}$ defined by

$$G(x, T, \epsilon) = (R(x, T, \epsilon), \langle f(x_0, 0), x - x_0\rangle)\,. \tag{24}$$

Note that $\langle f(x_0), x - x_0\rangle = 0$ is the equation of the Poincaré section $S_{f(x_0)}$. Since Γ_0 is a periodic orbit with period T_0 emanating from x_0, then $G(x_0, T_0, 0) = 0$. As already stated the strategy is to employ the IFT to derive the continuation of Γ_0. Thus, it is necessary to compute $D_{x,T} G$ at $(x_0, T_0, 0)$:

$$D_{x,T} G(x_0, T_0, 0) = \begin{pmatrix} D_x R(x_0, T_0, 0) & f(x_0, 0) \\ f(x_0, 0) & 0 \end{pmatrix}\,.$$

In order to show that $D_{x,T}G(x_0, T_0, 0)$ is invertible one notes that the equation

$$D_{x,T}G(x_0, T_0, 0)(X, a) = (0, 0)$$

is equivalent to the system

$$\begin{cases} D_x R(x_0, T_0, 0)(X) + f(x_0, 0)\, a = 0 \\ \langle f(x_0, 0), X\rangle = 0\,. \end{cases} \tag{25}$$

Assume $X \neq 0$. The multiplicity of the 0 eigenvalue of $D_x R(x_0, T_0, 0)$ is 1 and

$$D_x R(x_0, T_0, 0)\, f(x_0, 0) = 0\,.$$

Now from (25) we derive $(D_x R(x_0, T_0, 0))^2\, X = 0$. Therefore, $D_x R(x_0, T_0, 0)(X) + f(x_0, 0)\, a = 0$ would be satisfied only for $a = 0$ and $X = c f(x_0, 0)$ with $c \in \mathbb{R}$. But this would imply $c\,\langle f(x_0, 0), f(x_0, 0)\rangle = 0$, which is possible only for $c = 0$. The kernel of $D_{x,T}G$ is empty at $(x_0, T_0, 0)$ and therefore the map is invertible and the application of the IFT provides the existence of $T(\epsilon)$ and $x(\epsilon)$ for ϵ in a neighborhood of zero such that $x(0) = x_0$, $T(0) = T_0$ and $G(x(\epsilon), T(\epsilon), \epsilon) = 0$. This corresponds to the existence of a new periodic orbit close to Γ_0. Upon the assumption of sufficient regularity for the vector field, the IFT provides also the possibility of approximating $x(\epsilon)$ and $T(\epsilon)$ by constructing a Taylor expansion in ϵ. Let $x(\epsilon) = x_0 + \xi(\epsilon)$ and $T(\epsilon) = T_0 + \tau(\epsilon)$, then (25) can be evaluated along the continuation curve defined by $(\xi(\epsilon), \tau(\epsilon), \epsilon)$:

$$\begin{cases} D_x R(x_0, T_0, 0)(\xi'(\epsilon)) + f(x_0, 0)\, \tau'(\epsilon) \\ \qquad + T_0 V(1) \displaystyle\int_0^1 V(s) \frac{\mathrm{d} f(\phi_{T_0}(s, x_0), 0)}{\mathrm{d}\epsilon} \mathrm{d}s = 0 \\ \langle f(x_0, 0), \xi'(\epsilon)\rangle = 0\,. \end{cases} \tag{26}$$

This set of equations allows us to compute an approximation for $\xi(\epsilon)$ and $\tau(\epsilon)$. □

For a general dynamical system in $\mathbb{R}^n$ (23) the possibility of constructing a Poincaré section is related to the notion of a nondegenerate periodic orbit.

Definition 16 A periodic orbit $\phi(t, z_0)$ is called nondegenerate if

$$\mathrm{rank}(V(1) - \mathbf{id}) \oplus \mathbb{R}\, f(z_0) = \mathbb{R}^n\,.$$

Now for Hamiltonian systems the time evolution is contained in the level set determined by the the Hamiltonian function (the energy) and all integrals of motion $\{F_i\}_{i=1}^k$

$$X_H(F_i) = \{H, F_i\} = 0\,, \quad i = 1, \ldots, k\,.$$

This requires a modification of the notion of nondegeneracy. In fact note that the set of integrals of motions $(F_1, \ldots, F_k)$ define a map $F\colon \mathcal{P} \mapsto \mathbb{R}^k$ and span $W = \{\nabla F_i, i = 1 \ldots k\}$. Now we have $W^\perp = \ker(D_z F(z))$. One can show $\mathrm{rank}(V(1) - \mathbf{id}) \subset \ker(D_z F(z))$ and $X_H(F) = 0$. The last condition is equivalent to $X_H(z) \in \ker(D_z F(z))$. Suppose that $\dim W^\perp = 2n - k$, if $\dim(\ker(V(1) - \mathbf{id})) = k$ then $\dim(\mathrm{rank}(V(1) - \mathbf{id}) = 2n - k$ and hence $\mathrm{rank}(V(1) - \mathbf{id}) = W^\perp$. Therefore, in the Hamiltonian case the natural notion of nondegeneracy has to involve the presence of integrals of motion. This is obtained by using the following definition.

Definition 17 ([42]) A periodic orbit $\phi(t, z_0)$ is called normal if

$$\mathrm{rank}(V(1) - \mathbf{id}) \oplus \mathbb{R}\, X_H(z_0) = W^\perp\,, \tag{27}$$

where all the gradients $\{\nabla H, \nabla F_1, \ldots, \nabla F_k\}$ are linearly independent.

Lemma 2 ([42]) *If the algebraic multiplicity of the zero eigenvalue of $V(1) - \mathbf{id}$ is $m_a = k + 1$, then condition (27) is satisfied.*

Finally we present how to construct the continuation of nontrivial periodic orbits in Hamiltonian systems.

Theorem 5 ([42]) *Let $\Gamma_0 = \{\phi_{T_0}(t, z_0)\colon t \in [0, 1)\}$ be an orbit of the Hamiltonian system H_0 on the energy level e_0. Let the algebraic multiplicity of eigenvalue 0 of $V(1) - \mathbf{id}$ be 2. Let $H_\epsilon = H_0 + \epsilon\, H_1$ be a smooth perturbation of H_0, then there exists a two-dimensional family of normal periodic orbits $\phi(t, z_0; \epsilon, e)$ where $\phi(t, z_0, e_0, 0) = \phi(t, z_0)$.*

Proof Let us consider the map $G\colon \mathbb{R}^{2n} \times \mathbb{R}_+ \times [0, 1] \to \mathbb{R}^{2n+1}$ defined by

$$G(x, T, e, \epsilon) = (R(z, T, \epsilon), \langle X_H(z_0, 0), z - z_0\rangle, H_\epsilon(z) - e). \tag{28}$$

Since Γ_0 is a periodic orbit with period T_0 emanating from z_0 then $G(z_0, T_0, e_0, 0) = 0$. The strategy is always to employ the IFT to derive the continuation of Γ_0. Thus, it is necessary to compute $D_{z,T}G$ at $(z_0, T_0, e_0, 0)$. A straightforward calculation gives

$$D_{z,T}G(z_0, T_0, 0) = \begin{pmatrix} D_z R(z_0, T_0, 0) & X_H(z_0, 0) \\ X_H(z_0, 0) & 0 \\ \nabla H_0(z_0) & 0 \end{pmatrix}.$$

In order to show that $D_{z,T}G(z_0, T_0, h_0, 0)$ is invertible one notes that the equation

$$D_{z,T}G(z_0, T_0, h_0, 0)(X, a, 0) = (0, 0, 0)$$

is equivalent to the system

$$\begin{cases} D_z R(z_0, T_0, e_0, 0)(X) + X_H(z_0, 0)\, a = 0 \\ \langle X_H(z_0, 0), X\rangle = 0 \\ \langle \nabla H_0(z_0), X\rangle = 0\,. \end{cases} \tag{29}$$

Since $X_H(z_0) \in \ker D_z R(z_0, T_0, e_0, 0)$ then from (29) we derive $(D_z R(z_0, T_0, e_0, 0))^2(X) = 0$ and therefore $a = 0$. Now Lemma 2 implies that $X \in W^{\perp}$; therefore, the third equation in (29) is satisfied and the first implies $X = b\, X_H(z_0, 0)$ for some $b \in \mathbb{R}$. But this would contradict $b\,\langle X_H(x_0, 0), X_H(x_0, 0)\rangle = 0$ unless $b = 0$. This implies that that kernel of $D_{x,T}G$ is empty at $(x_0, T_0, e_0, 0)$ and therefore the map G is invertible. The application of the IFT provides the existence of $T(e, \epsilon)$ and $z(e, \epsilon)$ for ϵ in a neighborhood of 0 and e in a neighborhood of e_0 such that $z(e_0, 0) = z_0$, $T(e_0, 0) = T_0$. The functions $T(e, \epsilon)$ and $z(e, \epsilon)$ satisfy $G(z(e, \epsilon), T(e, \epsilon), e, \epsilon) = 0$. This corresponds to the existence of a new periodic orbit close to Γ. Upon the assumption of sufficient regularity for the vector field, the IFT provides also the possibility of approximating $z(e, \epsilon)$ and $T(e, \epsilon)$ by following the same line of argument seen in Proposition 6. This concludes the Proof. □

Numerical Studies The analysis of periodic orbits is very important for its concrete applications, hence for the construction of numerical algorithms to construct periodic orbits and their continuation. Here we do not consider explicitly this problem but the reader is invited to consult, for example, [31,32]. Moreover there is some free and open-source software available, for example, AUTO (see http://indy.cs.concordia.ca/auto/ and [19,20]).

Example

On $\mathcal{P} = \mathbb{R}^4$ with coordinates $z = (p, q)$ and the standard symplectic form, we consider the Hamiltonian

$$H = \frac{1}{2}\|p\|^2 + V_0(q),\; V_0(q) = \frac{1}{2}\|p\|^2 - \frac{\lambda}{2}\|q\|^2 + \frac{1}{4}\|q\|^4 \tag{30}$$

with $\lambda > 0$. The reader could check that $V_0(q)$ has the shape of a "Mexican-hat." The Hamilton equationHamiltonians of motion read

$$\begin{cases} \dot{q} = p \\ \dot{p} = (\lambda - \|q\|^2)q\,. \end{cases} \tag{31}$$

Let $e(t)$ be a unit vector with $e(t + 2\pi/\nu_0) = e(t)$, then there is a periodic orbit of the form $q_0(t) = A_0\, e(t)$, $p_0(t) = A_0\, \dot{e}(t)$. This orbit is a relative equilibrium (see below for a formal definition). The initial conditions determine A_0, ν_0 and in particular the energy E, which in turn can be used to parameterize A_0, ν_0: $A_0^2 = \frac{2}{3}(\lambda + \sqrt{\lambda^2 + 3E})$, $\nu_0^2 = \frac{1}{3}(-\lambda + \sqrt{\lambda^2 + 3E})$; here $E \le 0$. Note that an easy calculation shows that the admissible values of the energy are $E \ge \min_q\{V_0(q)\} = -\lambda^2/4$. Let $\phi(t, z_0) = (p_0(t), q_0(t))$ be the periodic solution, then the Floquet operator is

$$V(t) = \frac{\partial \phi(t, z_0)}{\partial z_0}$$

and the linearized equations can be written as

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t} = M(t)V(t)\,, \text{ where } M(t) \text{ is a } 4 \times 4 \text{ matrix:}$$

$$M(t) = \begin{pmatrix} 0 & \mathbf{id} \\ -D^2 V_0(q_0(t)) & 0 \end{pmatrix}.$$

Now if we perform the transformation

$$W(t) = S(t)\, V(t) = \begin{pmatrix} R(t) & 0 \\ 0 & R(t) \end{pmatrix} X(t)$$

$$\text{where } R(t) = \begin{pmatrix} \cos \nu_0 t & -\sin \nu_0 t \\ \sin \nu_0 t & \cos \nu_0 t \end{pmatrix}$$

we obtain

$$\frac{\mathrm{d}W(t)}{\mathrm{d}t} = \hat{M} W(t)\,,$$

where

$$\hat{M} = \dot{S}(t)\, S^{-1}(t) + S(t)\, M(t)\, S^{-1}(t)\,.$$

$\hat{M}$ is not time-dependent. In fact $\hat{M}$ is the following matrix:

$$\hat{M} = \begin{pmatrix} \Omega & \mathbf{id} \\ -R^T(t) D^2 V_0(q_0(t) R(t) & \Omega \end{pmatrix},$$

where

$$\Omega = \begin{pmatrix} 0 & -\nu_0 \\ \nu_0 & 0 \end{pmatrix} \quad \text{and}$$

$$-R^T(t) D^2 V_0(q_0(t) R(t) = \begin{pmatrix} -3A_0^2 + \lambda & 0 \\ 0 & -A_0^2 + \lambda \end{pmatrix}.$$

Now at $t = T_0$

$$V(T_0) = S^{-1}(T_0)\, W(T_0) = \exp(\hat{M}\, T_0)\,.$$

The matrix $\hat{M}$ has spectrum

$$\mathrm{Spec}(\hat{M}) = \left\{0, 0, \pm i\sqrt{6A_0^2 - 4\lambda}\right\}$$

and the corresponding multipliers are:

$$\exp(\mathrm{Spec}(\hat{M})) = \left\{1, 1, \exp\left(\pm i\, T_0 \sqrt{6A_0^2 - 4\lambda}\right)\right\}$$
$$= \left\{1, 1, \exp\left(\pm i\, T_0 \sqrt[4]{\lambda^2 + 3E}\right)\right\} ,$$

where $T_0 = 2\pi/\nu_0$. The matrix $V(T_0) - \mathbf{id}$ has a zero eigenvalue with multiplicity 2, and the periodic orbit $\phi_0(t, z_0)$ can be continued by using Theorem 5.

Remark 5 Note that using the expression for ν_0, one can check that for $E = E_n$, where

$$E_n = \frac{\lambda^2}{3}\left[-1 + \frac{n^4}{(n^2 - 12)^2}\right] \text{ with } n \in \mathbb{Z} ,$$

the third and the fourth multipliers coalesce to 1 and therefore $q_0(.)$ loses its linear stability in the directions transverse to itself.

Hamiltonian Systems with Symmetries

In many applications there are Hamiltonian systems with symmetries, the best known example of which is the N-body problem (see [1,5] and ▶ *n*-Body Problem and Choreographies). A simpler example is (31), which is symmetric with respect the linear action of $SO(2)$. For such systems the Hamiltonian function is invariant under the action of a Lie group G. We shall see that symmetries can greatly simplify the study of the dynamics. A Hamiltonian system with symmetry is a quadruple $(\mathcal{P}, \omega, H, G)$ [1,5,48], where

- $(\mathcal{P}, \omega)$ is a symplectic manifold,
- $H: \mathcal{P} \to \mathbb{R}$ is a Hamiltonian function,
- G is a Lie group that acts smoothly on $\mathcal{P}$ according to $G \times \mathcal{P} \ni (g, z) \mapsto \Phi_g(z) \in \mathcal{P}$. The map Φ_g preserves the Hamiltonian (G-invariance) that is $H(\Phi_g(z)) = H(z)$.
- The action Φ is semisymplectic, namely, $\Phi_g^*\omega = \chi(g)\omega$, with $\chi(g) = \pm 1$. $\chi(.)$ is called temporal character.

In the sequel we consider symplectic actions whereby $\chi(g) = 1$ for all $g \in G$. One can show that for a system $(\mathcal{P}, \omega, H, G)$ the vector field $X_H(z)$ is *equivariant* [1], that is,

$$X_H(\Phi_g(z)) = D_z\Phi_g(z)\, X_H(z) . \tag{32}$$

With any element ξ in the Lie algebra $\mathfrak{g}$ of G we can associate a infinitesimal generator $\xi_{\mathcal{P}}(z)$ of the action defined by

$$\xi_{\mathcal{P}}(z) = \left.\frac{\mathrm{d}\Phi_{\exp(\xi\, t)}(z)}{\mathrm{d}t}\right|_{t=0} . \tag{33}$$

Remark 6 In many applications Hamiltonian systems are constructed from Lagrangian systems; therefore, a symmetry appears usually as a group action on the configuration space $\mathcal{M}$. The Hamiltonian symmetry is then the *lifted* action to the cotangent bundle $T^*\mathcal{M}$. For instance, if $SO(2)$ acts linearly on $\mathcal{M} = \mathbb{R}^2$ by $\Psi_R(q) = R\,q$ with $R \in SO(2)$, then its lifted action on $T^*\mathcal{M} \simeq \mathbb{R}^4$ is $\Phi_R(q, p) = (R\,q, R^T\,q)$, where R^T is the transpose of R. For more details see [1,36,48].

Symmetry and Reduction

Given a symplectic action of a group G there is a map $\mathbf{J}: \mathcal{P} \to \mathfrak{g}^*$ defined by

$$\begin{aligned} &\langle \mathrm{d}\mathbf{J}(v(z)), \xi\rangle = \omega(v(z), \xi_{\mathcal{P}}(z)) \\ &\text{for all} \quad z \in \mathcal{P},\ v \in T_z\mathcal{P} \quad \text{and} \quad \xi \in \mathfrak{g} . \end{aligned} \tag{34}$$

The map $\mathbf{J}$ is called a *momentum map*. This always exists locally, and its global existence requires conditions on G and the topology of $\mathcal{P}$ [27]. The crucial property of the momentum map is its encoding of the conserved quantities associated with the G action. This is the content of the famous Noether's theorem that reads in modern formulation as follows.

Theorem 6 (Noether [1,5]) *Let H be a G-invariant Hamiltonian on $\mathcal{P}$ with momentum map $\mathbf{J}$. Then $\mathbf{J}$ is conserved on the trajectories of the Hamiltonian vector field X_H.*

For instance, in a system like (30) with a $SO(2)$ symmetry action the momentum map is the classical angular momentum

$$\mathbf{J}(p, q) = p \wedge q .$$

The momentum map $\mathbf{J}$ has also the property of being equivariant with respect to the coadjoint action associated with G. In the case of G being semisimple or compact the result is as follows.

Theorem 7 (Souriau [27])

$$\mathbf{J}(\Phi_g(z)) = \mathrm{Ad}_g^*\, \mathbf{J}(z) .$$

The level sets of the momentum map are invariant with respect to the Hamiltonian flow; thus, it is natural to restrict the motion to $\mathbf{J}^{-1}(\mu)$. The construction of the dynamics reduced to the manifold defined by the conserved quantities is the origin of the theory of *symmetry reduction*. The first important result is the following.

Theorem 8 (Marsden–Weinstein) *Let $(\mathcal{P}, \omega, H, G)$ be a Hamiltonian system with a symplectic action of the Lie*

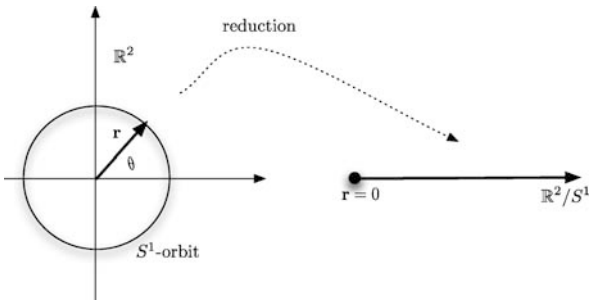

Periodic Orbits of Hamiltonian Systems, Figure 2
A simple example: the reduction of S^1 group action on the plane is singular. There are two strata $r = 0$ and $r \in (0, \infty)$

group G, then the triple $(\mathcal{P}_\mu, \omega_\mu, H_\mu)$ is called a reduced Hamiltonian system where

$$\pi : \mathcal{P} \to \mathcal{P}_\mu = \mathbf{J}^{-1}(\mu)/G_\mu \,, \;\; \pi^*\omega = \omega_\mu \,, \;\; H_\mu = H \circ \pi$$

and

$$G_\mu = \{g \in G \colon \; \mathrm{Ad}_g^*(\mu) = \mu\} \,.$$

The manifold $\mathcal{P}_\mu$ is symplectic with symplectic form ω_μ. The Hamiltonian flow on $\mathcal{P}_\mu$ is induced by the Hamiltonian vector field X_{H_μ} defined by

$$i_{X_{H_\mu}} \omega_\mu = dH_\mu \,.$$

Remark 7 Note that the reduced vector field X_{H_μ} is now defined on a manifold whose dimension is $\dim \mathcal{P}_\mu = \dim \mathcal{P} - \dim G - \dim G_\mu$.

For a detailed discussion of this theorem the reader should refer to [1,16,38]. In many cases it can be more useful to perform the reduction through a dual approach using the description of the dynamics in terms of functions on $\mathcal{P}$, namely, through the Poisson approach. This permits us to consider also cases where $\mathcal{P}_\mu$ is not a smooth manifold but rather a stratified space. A very simple example is shown in Fig. 2.

The theory of singular reduction can be found in [4]. A recent exposition and generalization is given in [45]. The reader could, in particular, consider the expositions given in [16,17], where it is shown through several examples that invariant theory and algebraic methods can be applied to describe reduced dynamics on spaces with singularities. The singular reduction can be summarized in the following result:

Theorem 9 (Singular reduction [45]) *Let $(\mathcal{P}, \{.,.\})$ be a Poisson manifold and let $\Phi : G \times \mathcal{P} \to \mathcal{P}$ be a smooth proper action preserving the Poisson bracket. Then the following holds:*

(i) The pair $(\mathcal{F}(\mathcal{P}/G), \{.,.\}_{\mathcal{P}/G})$ is a Poisson algebra, where the Poisson bracket $\{.,.\}_{\mathcal{P}/G}$ is characterized by $\{f, g\}_{\mathcal{P}/G} = \{f \circ \pi, g \circ \pi\}$; for any $f, g \in \mathcal{F}(\mathcal{P}/G)$; $\pi : \mathcal{P} \to \mathcal{P}/G$ denotes the canonical smooth projection.

(ii) Let h be a G-invariant function on M. The Hamiltonian flow $\phi(t, .)$ of X_h commutes with the G-action, so it induces a flow $\phi^{\mathcal{P}/G}(.,.)$ on $\mathcal{P}/G$ which is a Poisson flow and is characterized by $\phi(t, z) = \phi^{\mathcal{P}/G}(t, \pi(z))$.

(iii) The flow $\phi^{\mathcal{P}/G}(.,.)$ is the unique Hamiltonian flow defined by the function $[h] \in \mathcal{F}(\mathcal{P}/G)$ defined by $[h](\pi(z)) = h(z)$. We will call $H_{\mathcal{P}/G} = [h]$ the reduced Hamiltonian.

Relative Equilibria, Relative Periodic Orbits and their Continuation

In a Hamiltonian system with symmetry there are orbits that originated just from the symmetry invariance. These are the relative equilibria.

Definition 18 (Relative equilibria) A curve $z_e(t)$ in $\mathcal{P}$ is a relative equilibrium of $(\mathcal{P}, \omega, H, G)$ if $z_e(t) = \Phi_{g(t)}(w_e)$, where $g(t)$ is a curve in G and w_e is such that $X_H(z_e) = \dot{z}_e(t)$.

Note that if $g(t) = g(t + T)$ then $z_e(t)$ becomes a T-periodic orbit. For instance in the system (30) there is $z_e(t) = R(t)\, w_e = (A_0\, e(t), A_0\, \dot{e}(t))$ with $R \in SO(2)$. There are orbits that can be considered closed up to a G-action, i. e., they are closed on $\mathcal{P}/G_\mu$. These are the relative periodic orbits.

Definition 19 (Relative periodic orbit) A curve $z(t)$ in $\mathcal{P}$ is a relative periodic orbit of $(\mathcal{P}, \omega, H, G)$ with period T if $z(t + T) = \Phi_g(z(t))$, where $g \in G$ and $g \neq \mathbf{id}$.

An illustration is given in Fig. 3.

The reduction theory allows us to redefine the relative equilibrium. Let $\mu \in \mathfrak{g}$ and $z \in \mathcal{P}_\mu$ be an equilibrium for X_{H_μ}. Then z gives rise to a relative equilibrium in $\mathcal{P}$, for a closed curve $g(t)$ in G_μ and define $w(t) = \Phi_{g(t)}(w_e)$ with $w_e \in \pi_\mu^{-1}(z)$. If $z(t)$ is a relative equilibrium then $\pi(z(t))$ is an equilibrium of the reduced flow [1]. In general there are two possible ways to study the relative equilibrium. The first approach is to construct the reduction of the Hamiltonian system and then to study

$$X_{H_\mu}(z) = 0 \tag{35}$$

in Hamiltonian form and

$$\{H_G, f\}_{\mathcal{P}/G}(z) = 0 \quad \text{for all} \quad f \in \mathcal{F}(\mathcal{P}/G) \tag{36}$$

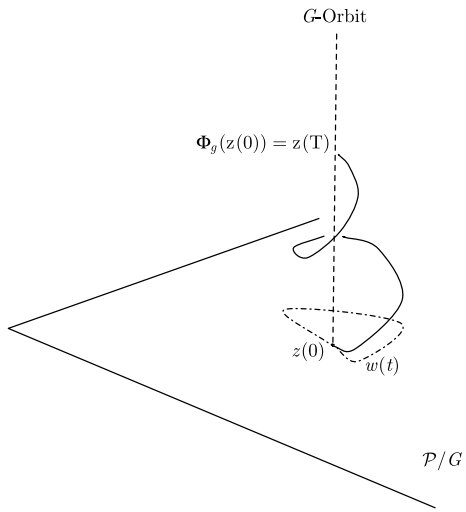

Periodic Orbits of Hamiltonian Systems, Figure 3
A relative periodic orbit. After a relative period *T* the orbit *z*(.) closes onto the group orbit. The projection of *z*(.) on $\mathcal{P}/G$ is closed

in the Poisson description. The second approach is to observe that if $z(t) = \Phi_{g(t)}(z^*)$ is a relative equilibrium, then the condition $\dot{z}(t) = X_H(z(t))$ implies

$$\mathrm{d}H(z^*) - \mathrm{d}\langle \mathbf{J}(z^*), \xi\rangle = 0\,, \quad \text{with; } \xi_{\mathcal{P}}(z^*) = \left.\frac{\mathrm{d}\Phi_{g(t)}(z^*)}{\mathrm{d}t}\right|_{t=0}. \tag{37}$$

On the three formulations, (35), (36) and (37), one can apply all the standard continuation techniques based on the IFT. The possible difficulty in the study of the continuation of relative equilibria and relative periodic orbits can be seen by looking at the linearization of (37). The symmetry contributes to the degeneracy of the linearized equations, in fact one can show

Proposition 7 ([37]) *Let $X(z)$ be an equivariant vector field with respect to a Lie group G. Then a relative equilibrium has multiplier $+1$ with multiplicity at least equal to* $\dim \mathfrak{g}$*; and a relative periodic orbit has multiplier $+1$ with multiplicity at least equal to* $\dim \mathfrak{g} + 1$.

In [39] relative periodic orbits are defined by looking at the fixed points of the action of $G \times S^1$ on the space of T-periodic paths. This reads

$$(g, \theta) \cdot z(t) = \Phi_g(z(t + \theta\, T))\,. \tag{38}$$

Remark 8 Note that (38) depends on T and it is easy to verify that T/k is a minimal period if the intersection of the isotropy subgroup of the action $G \times S^1$ with S^1 is equal to $\mathbb{Z}_k$.

Theorem 10 ([39]) *Let $(\mathcal{P}, \omega, H, G)$ be a symmetric Hamiltonian system. Let $z_e \in \mathcal{P}$ such that*

1. *$d^2H(z_e)$ is a nondegenerate quadratic form*
2. *$d^2H(z_e)$ is positive-definite if restricted to V_λ, which is the real part of the eigenspace associated with the eigenvalue $i\,\lambda$, $\lambda \in \mathbb{R}$.*

Then for every isotropy subgroup Γ of the $G \times S^1$-action on V_λ and ϵ sufficiently small there exist at least $\mathrm{Fix}(\Gamma, V_\lambda)$ *periodic trajectories with period near $2\pi/|\lambda|$, a symmetry group contained in Γ and $|H(z(t)) - H(z_e)| = \epsilon^2$.*

This result has been proved using a combination of a variational approach and the IFT, and will be considered in "Hamiltonian Viewpoint". We have seen that the Poincaré section is a useful construction to analyze the local structure of a periodic orbit. A very similar approach can be introduced for relative periodic orbits. A relative periodic orbit can be seen as a combination of motions in $\mathcal{P}$ and in the symmetry group G. It is therefore natural to look for a suitable decomposition of the motion. In [39] following decomposition was introduced:

$$T_z\mathcal{P} = W \oplus X \oplus Y \oplus Z\,, \tag{39}$$

where

$$\begin{aligned} W &= \ker D_z\mathbf{J}(z) \cap T_z(G\cdot z),\ \ G\cdot z = \{\Phi_g(z)\colon g \in G\}\,,\\ X &= T_z(G\cdot z)/W\,,\\ X &= \ker D_z\mathbf{J}(z)/W\,,\\ Z &= T_z\mathcal{P}/(\ker D_z\mathbf{J}(z) + T_z(G\cdot z))\,. \end{aligned}$$

Let $G_z = \{g \in\colon \Phi_g(z) = z\}$ and $G_\mu = \{g \in G\colon Ad^*_g(\mu) = \mu\}$ be the isotropy subgroups. It turns out that

- $\ker D_z\mathbf{J}(z)$ and $T_z(G\cdot z)$ are ω-orthogonal,
- ω restricted $\ker D_z\mathbf{J}(z) + T_z(G\cdot z)$ is singular and W is the kernel,
- ω induces a G_z-invariant symplectic form ω_X on X and ω_Y on Y,
- ω defines a G_z-invariant isomorphism between W and Z^* the dual of Z.

The splitting (39) allows us to decompose the vector field X_H and to analyze the motion along the group orbit and along the transverse directions. This is a tool used in [39] to study the Floquet operator in a neighborhood of the relative periodic orbit. For applications in the study of nonlinear normal modes and stability see [40,41]. In [58] it is shown that such a construction can be used to decompose the Poincaré section in a part which is tangent to the conserved momentum, another part which is tangent to

shape and a part parameterizing the momentum. This approach has also been applied to the study of the geometry of mechanical systems defined on the cotangent bundle of a differentiable manifold; for this see [50] and references therein.

The Variational Principles and Periodic Orbits

The idea behind variational principles is to transform a problem in differential equations into a question about critical points of a certain functional called *action*, whose domain is formed by trajectories of interest. Now in the study of T-periodic orbits we are interested in finding trajectories ϕ solving the equations of motion and satisfying $\phi(t) = \phi(t+T)$. This is a periodicity condition. The variational approach is particularly useful in the study of periodic problems because the periodicity condition is included in the definition of the space where the action is defined. Furthermore the method enables us to prove results not restricted to *small perturbations*. Let $\mathcal{A}\colon \Lambda \mapsto \mathbb{R}$ be a functional on a space of loops usually modeled on a Banach or Hilbert space. We shall see that the condition of vanishing of the first derivative of $\mathcal{A}$ is equivalent, in an appropriate sense, to solving the equations of motion. In order to describe the properties of $\mathcal{A}$ let us introduce

Definition 20 (Critical points) Let $\mathcal{A}[.]$ be a differentiable functional on Λ. A path $q(.)$ is a *critical point* of $\mathcal{A}[.]$ if

$$D\mathcal{A}[q](v) = 0 \quad \text{for all} \quad v(.) \in T\Lambda\,.$$

Definition 21 (Critical set) The set $K = \{q(.) \in \Lambda\colon D\mathcal{A}[q] = 0\}$ is the *critical set* of $\mathcal{A}[.]$.

Definition 22 (Critical value) A real number c is called a *critical value* if $K_c \doteq \mathcal{A}^{-1}[c] \cap K \neq \emptyset$.

Lagrangian Viewpoint

A typical way to write Newton's equations in a variational form is by using Lagrange's equations which are formulated through the *least action principle*. This can be achieved for all mechanical systems that have potential forces. Consider a mechanical system whose configuration space is a Riemannian manifold $\mathcal{M}$ of dimension n. We denote with (q, v_q) the local coordinates in $T\mathcal{M}$ and with $L\colon T\mathcal{M} \to \mathbb{R}$ the Lagrangian functionLagrangian. In particular in the case of the so-called *natural mechanical system* [1] the Lagrangian has the form

$$L(q, v_q) = \frac{1}{2}\langle v_q, v_q\rangle - V(q)\,, \tag{40}$$

where $\langle .,. \rangle$ is the metric on $T\mathcal{M}$ and $V\colon \mathcal{M} \to \mathbb{R}$ is the potential. Given a path $q\colon [0,T] \to \mathcal{M}$ with integrable time derivative one can define

Definition 23 (Action functional in Lagrangian form)

$$\mathcal{A}_L[q] = \int_0^T \mathrm{d}t\, L(q(t), \dot{q}(t))\,. \tag{41}$$

In what follows we will be mostly interested in closed paths. Let $C^2([0,T], \mathbb{R}^n)$ be the space of a closed path of period T with two continuous time derivatives. One can easily show that

Proposition 8 ([1,5]) *Let L be a smooth Lagrangian function on $\mathcal{M}$ of the form (40). Let $\mathcal{A}_L$ be defined over $\Lambda = C^2([0,T], \mathcal{M})$. Then*

$$D\mathcal{A}_L[q](v) = 0 \quad \textit{for every } v(.) \in T\Lambda \simeq C^2([0,T], T\mathcal{M})$$

is equivalent to Newton's equations with $q(0) = q(T)$. In particular the equations of motion in local coordinates are the Euler–Lagrange equations

$$\frac{d}{dt}\frac{\partial L(\dot{q}(t), q(t))}{\partial \dot{q}_i} - \frac{\partial L(\dot{q}(t), q(t))}{\partial q_i} = 0\,, \quad i = 1, \ldots, n\,.$$

Remark 9 For historical reasons the condition $D\mathcal{A}_L[q](v) = 0$ is called *least action* although it is only a condition for stationary points of $\mathcal{A}_L[.]$ in Λ.

In general the Lagrangian contains a quadratic form in v_q; in turn the action has a time integral of a quadratic form in $\dot{q}(t)$. This essentially shows that the natural domain for $\mathcal{A}_L[.]$ is the Sobolev space $H^1([0,T], \mathbb{R}^n)$. Actually the action is defined on $H^1([0,T], \mathcal{M})$, which is a Hilbert manifold [8,29]. This is locally described by absolutely continuous functions with the time derivative in the Lebesgue space L^2 [28]. The interest in variational methods is related to the possibility of using critical point theory to find critical points corresponding to certain type of trajectories and then to show that such trajectories are solutions of Newton's equations [2,21,33]. In particular this approach has been very successful in studying the problem of periodic orbits. Here is a general result in the Lagrangian setting:

Theorem 11 ([8]) *Let $\mathcal{M}$ be compact Riemannian manifold and let $L\colon T\mathcal{M} \to \mathbb{R}$ be a Lagrangian of the form (40) with $V \in C^1(\mathcal{M}, \mathbb{R})$. Then there exists a periodic orbit in any free homotopy class.*

We illustrate the ideas of the proof by considering the special case where $\mathcal{M}$ is the n-dimensional torus $\mathbb{T}^n \simeq \mathbb{R}^n/\mathbb{Z}^n$. The problem is now to show that the action $\mathcal{A}_L[.]$ attains a critical point, a minimum, in $\Lambda(\mathcal{M}) =$

$\{q(.) \in H^1([0,T],\mathcal{M})\colon q(.)$ is not null homotopic$\}$. Let us consider the sublevel of the action $A_k = \{q(.) \in \Lambda(\mathcal{M})\colon \mathcal{A}_L[q] \le k\}$. Now since V is smooth there exists $\min_{\mathcal{M}} V(q) = m > -\infty$ and therefore

$$\mathcal{A}_L[q] \ge \frac{1}{2}\int_0^T \mathrm{d}t\, \|\dot{q}(t)\|^2 - m\,.$$

Since $\mathcal{M}$ is compact, the norm of $q(.)$ in $\Lambda(\mathcal{M})$ is equivalent to $\|\dot{q}\|_2$. This allows us to show that the action $\mathcal{A}[.]$ is *coercive*, namely:

Definition 24 (Coercivity) $\mathcal{A}\colon \Lambda \to \mathbb{R}$ is coercive in Λ if for all $q_n(.)$ such that $\lim_{n\to\infty} \|q_n\|_\Lambda = \infty$. Then $\lim_{n\to\infty} \mathcal{A}[q_n] = 0$.

Moreover in A_k necessarily we have

$$\|\dot{q}\|_2^2 = \int_0^T \mathrm{d}t\, \|\dot{q}(t)\|^2 \le 2(k+m)\,.$$

This condition guarantees that A_k is weakly compact in the topology of Λ [2,33,53] and using the coercivity, one can obtain the existence of a minimizer $q_*(.)$. The minimizer $q_*(.)$ is attained in A_k and it is a *weak* solution of the equations of motion. Indeed $\mathcal{A}[q^*] = \min_\Lambda \mathcal{A}[q]$ and

$$D\mathcal{A}[q_*](v) = 0 \;\forall v(.) \in \Lambda(T\mathbb{R}^n)\,,$$

namely,

$$\int_0^T \mathrm{d}t \sum_{i=1}^n \left(\frac{\partial L(\dot{q}_*(t), q_*(t))}{\partial q_i}\, v_i(t) + \frac{\partial L(\dot{q}_*(t), q_*(t))}{\partial \dot{q}_i}\, \dot{v}_i(t) \right) = 0\,. \quad (42)$$

Using the fact that $q_*(.)$ is absolutely continuous and that L is regular, one can integrate by parts

$$\int_0^T \mathrm{d}t \sum_{i=1}^n \left(\frac{\partial L(\dot{q}_*(t), q_*(t))}{\partial \dot{q}_i} - \int^t \mathrm{d}s \frac{\partial L(\dot{q}_*(s), q_*(s))}{\partial q_i} \right) \dot{v}_i(t) = 0\,, \quad (43)$$

and obtain

$$\frac{\partial L(\dot{q}_*(t), q_*(t))}{\partial \dot{q}_i} - \int^t \mathrm{d}s \frac{\partial L(\dot{q}_*(s), q_*(s))}{\partial q_i} = c_i\,,$$

where c_i is constant in $L^2([0,T],\mathbb{R}^n)$.

By differentiating with respect to t, we obtain the Euler–Lagrange equations:

$$\frac{\mathrm{d}}{\mathrm{d}t} \frac{\partial L(\dot{q}_*(t), q_*(t))}{\partial \dot{q}_i} - \frac{\partial L(\dot{q}_*(t), q_*(t))}{\partial q_i} = 0 \text{ a.e. } \forall i. \quad (44)$$

It turns out that the equality in (44) holds for all times t because L is smooth. The reader should observe that the derivations of (42) and (44) are part of the proof of Proposition 8. It is necessary to note that constant paths are not in $\Lambda(\mathcal{M})$; indeed the number of minimizers is bounded from below by the Lusternik–Schnirelman category $\mathrm{Cat}(\mathcal{M}) = n+1$ [15,33]. Thus, we exclude possible minimizers which would be trivial periodic orbits. In this very simple example we can therefore appreciate the role of the definition of the space of paths $\Lambda(\mathcal{M})$ and its topology. More interesting cases can be found in the study of the N-body problem and in particular in the article ► n-Body Problem and Choreographies in this encyclopedia.

The result in [8] can be generalized to Lagrangian systems with symmetries. Assume there is a group action on the configuration space $G \times \mathcal{M} \mapsto \mathcal{M}$ – denoted by $(g,m) \mapsto g.m$ – which preserves the Lagrangian $L\colon T\mathcal{M} \to \mathbb{R}$ and consider the problem of finding relative periodic paths $\gamma(.)$ as critical points of the action $\mathcal{A}_L[.]$. We need to study the topology of

$$\Lambda^g(\mathcal{M}) = \{\gamma(.) \in H^1([0,T],\mathcal{M})\colon \gamma(t+T) = g.\gamma(t)\}\,. \quad (45)$$

In fact if the action $\mathcal{A}_L[.]$ is bounded from below on $\Lambda^g(\mathcal{M})$, then each connected component would contain at least a minimum that is a critical point and therefore a periodic orbit. The following analysis was presented in [35]. In what follows for notational simplicity we denote a path and its homotopy class by the same symbol and use $*$ to denote both concatenation of paths and the induced operations on homotopy classes. Assume that $\mathcal{M}$ is connected. Choose a base point $m \in \mathcal{M}$ and let

$$\Lambda_m^g(\mathcal{M}) \doteq \{\gamma \in \Lambda^g(\mathcal{M})\colon \gamma(0) = m\}\,,$$

the space of continuous paths from m to gm. Let $\Lambda_m(\mathcal{M}) \doteq \Lambda_m^{\mathrm{id}}(\mathcal{M})$ denote the space of continuous loops based at m. Note that the space of connected components of $\Lambda_m(\mathcal{M})$ is the fundamental group of $\mathcal{M}$: $\pi_0(\Lambda_m(\mathcal{M})) = \pi_1(\mathcal{M},m)$. Fix a particular path $\omega \in \Lambda_m^g(\mathcal{M})$. The map $\Phi_\omega(\gamma) = \omega^{-1} * \gamma$ is a bijection

$$\Phi_\omega\colon \pi_0(\Lambda_m^g(\mathcal{M})) \to \pi_0(\Lambda_m(\mathcal{M})) = \pi_1(\mathcal{M},m)\,,$$

where ω^{-1} is the path obtained by traversing ω "backwards." This bijection depends (only) on the homotopy class of ω in $\Lambda_m^g(\mathcal{M})$. For any $\alpha \in \Lambda_m(\mathcal{M})$ let $g.\alpha$ be the loop in $\Lambda_{gm}(\mathcal{M})$ obtained by applying the diffeomorphism g to α and define an automorphism of $\pi_1(\mathcal{M},m)$ by

$$\alpha \mapsto \alpha_g = \omega^{-1} * g.\alpha * \omega\,.$$

Again this depends (only) on the homotopy class of ω in $\Lambda^g_m(\mathcal{M})$. Now define the *g-twisted action* of $\pi_1(\mathcal{M}, m)$ on itself by

$$\alpha \cdot \beta \;=\; \alpha_g * \beta * \alpha^{-1} \qquad \alpha, \beta \in \pi_1(\mathcal{M}, m)\,. \quad (46)$$

The number of connected components of the relative loop space is given by the following result:

Theorem 12 *The map Φ_ω induces a bijection*

$$\pi_0(\Lambda^g(\mathcal{M})) \;\cong\; \overline{\pi_1(\mathcal{M}, m)}^g\,,$$

where $\overline{\pi_1(\mathcal{M}, m)}^g$ is the set of orbits of the g-twisted action of $\pi_1(\mathcal{M}, m)$ on itself.

Remark 10 If g is homotopic to the identity then $\Lambda^g(\mathcal{M})$ is homotopy-equivalent to the loop space $\Lambda(\mathcal{M}) \doteq \Lambda^{\mathrm{id}}(\mathcal{M})$ and the g-twisted action of $\pi_1(\mathcal{M}, m)$ on itself is just conjugation. This is therefore the well-known result that the connected components of the loop space map bijectively to the conjugacy classes of the fundamental group [29].

Remark 11 The g-twisted action of $\pi_1(\mathcal{M}, m)$ on itself induces an affine action of the first homology group $H_1(\mathcal{M})$ on itself:

$$\langle\alpha\rangle \cdot \langle\beta\rangle \;=\; g.\langle\alpha\rangle - \langle\alpha\rangle + \langle\beta\rangle\,,$$

where the brackets $\langle . \rangle$ denote the homology class and $g.\langle\alpha\rangle$ denotes the natural action of g on $H_1(\mathcal{M})$. When $\pi_1(\mathcal{M}, m)$ is Abelian this is the same as the action of $\pi_1(\mathcal{M}, m)$ on itself. More generally it is easier to calculate than the $\pi_1(\mathcal{M}, m)$ action and in typical applications may be used to describe relative periodic orbits in terms of winding numbers.

The analysis gives explicit results for systems whose configuration space has the property that all its homotopy groups except the fundamental group are trivial. In this case $\mathcal{M}$ is said to be $K(\pi, 1)$; for more details see [12,57]. Examples of $K(\pi, 1)$ spaces include tori, the plane $\mathbb{R}^2$ with N points removed, and the configuration spaces of planar N-body problems.

Theorem 13 *Assume $\mathcal{M}$ is a $K(\pi, 1)$. Then for any $\gamma \in \Lambda^g_m(\mathcal{M})$ the connected component of $\Lambda^g(\mathcal{M})$ containing γ, denoted $\Lambda^g_\gamma(\mathcal{M})$, is also a $K(\pi, 1)$ with*

$$\pi_1(\Lambda^g_\gamma(\mathcal{M})) \;\cong\; Z^g_{\pi_1(\mathcal{M})}(\Phi_\omega(\gamma)),$$

where

$$Z^g_{\pi_1(\mathcal{M})}(\Phi_\omega(\gamma)) \doteq \{\alpha \in \pi_1(\mathcal{M}):\; \alpha_g * \Phi_\omega(\gamma) * \alpha^{-1} = \Phi_\omega(\gamma)\}$$

i.e., the isotropy subgroup (or centralizer) at $\Phi_\omega(\gamma)$ of the g-twisted action of $\pi_1(\mathcal{M}, m)$ on itself.

We note that all $K(\pi, 1)$'s with isomorphic fundamental groups are homotopy-equivalent to each other [12,57], and so this result determines the homotopy types of connected components of relative loop spaces. The homology groups can be computed algebraically as the homology groups of the fundamental group [14].

A Simple Example Let $\mathcal{M} = T^1$, the circle, and consider first the loop space $\Lambda(T^1)$. The "g-twisted action" of $\pi_1(T^1)$ on itself is just conjugation, and since $\pi_1(T^1) \cong \mathbb{Z}$ is Abelian this is trivial. So $\pi_0(\Lambda(M)) \cong \mathbb{Z}$, the homotopy classes of loops being specified precisely by their winding numbers. Since T^1 is a $K(\pi, 1)$, Theorem 13 says that each component of relative loop space is also a $K(\pi, 1)$ with a fundamental group isomorphic to $\mathbb{Z}$, and therefore has the homotopy type of a circle.

Now consider $\Lambda^g(T^1)$ where $g\colon T^1 \to T^1$ is a reflection. Choose one of the two fixed points of the reflection to be the base point m. We may choose ω to be the trivial path from m to $g.m$. Then for each $\alpha \in \pi_1(T^1, m) \cong \mathbb{Z}$ we have $\alpha_g = -\alpha$ and so the "g-twisted action" (46) is the translation

$$\alpha.\beta \;=\; \beta - 2\alpha. \quad (47)$$

This has two orbits, $\overline{\pi_1(T^1)}^g \cong \mathbb{Z}_2$, and the isotropy subgroups are trivial. It follows from Theorems 12 and 13 that the space of relative loops $\Lambda^g(T^1)$ has two components, each of which is contractible.

Numerical Studies In many interesting problems, typically in celestial mechanics, the action functional is bounded from below and therefore the expected critical points are minimizers. In recent years, in connection with the discovery of the so-called *choreographic periodic orbits*, Simó [52] developed an algorithm to study how to perform a numerical minimization of the action in classes of loops with a definite symmetry type. The idea is based on the description of the orbit in terms of its Fourier coefficients and defining the action $\mathcal{A}_L[.]$ as a function on the Fourier space. The action $\mathcal{A}_L[.]$ is then minimized on the space of Fourier coefficients. The interested reader should consult ► *n*-Body Problem and Choreographies in this encyclopedia. It would be very interesting to generalize this method to different actions and to see whether one can impose constraints not only on the symmetry type but also on the topology of the space of loops.

Hamiltonian Viewpoint

The Hamilton principle gives a variational characterization to the Hamiltonian equation. For Hamiltonian systems in $\mathbb{R}^{2n}$ the formulation of the principle is very simple. Let $H\colon \mathbb{R}^{2n} \to \mathbb{R}$ be the Hamiltonian function, then the action functional is

Definition 25 (Action functional in Hamiltonian form)

$$\mathcal{A}_H[\gamma] \doteq \int_0^T \mathrm{d}t \sum_{i=1}^n p_i(t)\,\dot{q}_i(t) - \int_0^T \mathrm{d}t\, H(q(t), p(t))\,, \tag{48}$$

where $\gamma(t) = (q(t), p(t))$.

It is worth recalling that if the Lagrangian function is hyperregular, then the system can be transformed through the Legendre transform into a Hamiltonian system on the cotangent bundle of the configuration space. In any case, given $\mathcal{A}_H[.]$ one can show [1,5]

Proposition 9 *Let H a smooth Hamiltonian function on $\mathbb{R}^{2n}$. Let $\mathcal{A}_H$ be defined over $C^2([0,T],\mathbb{R}^{2n})$, then*

$$D\mathcal{A}_H[q,p](v) = 0 \quad \textit{for every } v(.) \in C^2([0,T],\mathbb{R}^{2n})$$

is equivalent to Hamiltonian equations with $q(0) = q(T)$, $p(0) = p(T)$.

Recall that in the Hamiltonian context p and q are independent variables and therefore to prove the preceding proposition it is necessary to compute the variations with respect to q's and p's independently. From the analytical point of view the functional $\mathcal{A}_H[.]$ is a difficult object to study since it is unbounded from below and above, namely, it is *indefinite* [33]. It is not difficult to give an example, consider $n = 1$ and a Hamiltonian like $H = p^2/2$. For indefinite functionals a whole theory has been developed to apply the *mountain pass* theorem [47]. There is also a generalization of the Legendre transform [21]. The new transform [21] can be applied directly to the action functional $\mathcal{A}_H[.]$ rather than to H. This allows us to work with a convex functional. We do not enter into the details but we want to recall one result which describes the conditions for the existence of at least n periodic orbits.

Theorem 14 ([21]) *Suppose that $H \in C^1(\mathbb{R}^{2n},\mathbb{R})$ and for some $\beta > 0$, assume that $C = \{z \in \mathbb{R}^{2n}\colon H(z) \le \beta\}$ is strictly convex, with boundary $\partial C = \{z \in \mathbb{R}^{2n}\colon H(z) = \beta\}$ satisfying $\langle z, \nabla H(z)\rangle > 0$ for all $z \in \partial C$. Suppose that for $r, R > 0$ with $r < R < \sqrt{2}\,r$ there are two balls $B_r(0), B_R(0)$ centered at the origin of $\mathbb{R}^{2n}$ such that $B_r(0) \subset allowbreakC \subset B_R(0)$, then there are at least n distinct periodic solutions on ∂C of the Hamilton system $\dot{z} = J\nabla H(z)$.*

The functional $\mathcal{A}_H[.]$ has been defined for Hamiltonian systems on $\mathbb{R}^{2n}$, but it admits a generalization to symplectic manifolds which are not tangent bundles:

Definition 26 Let $(\mathcal{P},\omega)$ be a symplectic manifold and $H\colon \mathcal{P} \to \mathbb{R}$ a smooth Hamiltonian function, let $\gamma\colon \mathcal{P} \to \mathbb{R}$ be a closed path which is the boundary of a two-dimensional connected region Σ, then

$$\mathcal{A}_H[\gamma] \doteq \int_\Sigma \omega - \int_0^T \mathrm{d}t\, H(z(t))\,. \tag{49}$$

The functional (49) is a very interesting object. It is naturally defined on closed paths which bound two-dimensional regions. The functional can be defined on "paths" but then it would become multivalued. In fact in $\mathcal{M}$ one has to introduce a 1-form α such that $\omega = \mathrm{d}\alpha$. The 1-form α is not unique and depends on the co-homology of $\mathcal{M}$. Although $\mathcal{A}_H$ is multivalued the differential $D\mathcal{A}_H$ is single-valued [34]. Let $\varphi(s)$ be a finite variation with $\frac{\mathrm{d}\varphi(s)}{\mathrm{d}s}|_{s=0} = \xi$, then

$$D\mathcal{A}_H[z](\xi) = \left.\frac{\mathrm{d}\mathcal{A}_H[\varphi(s)}{\mathrm{d}s}\right|_{s=0} = \int_\Sigma L_\xi(\omega) - \int_0^T \mathrm{d}t\langle \nabla H(z), \xi\rangle\,,$$

where L_ξ is the Lie derivative with respect to ξ. Now since $\mathrm{d}\omega = 0$ one can show that

$$D\mathcal{A}_H[z](\xi) = \int_0^T i_\xi\, i_{X_H}\omega - \int_0^T \mathrm{d}t\langle \nabla H(z), \xi\rangle\,,$$

which is single-valued. The theory for such multivalued functionals has been developed by many authors; here we would like to cite [22,43,56]. Note that functionals of the form (49) can result from *symmetry reduction*. In fact as shown in [34] a Lagrangian system with a non-Abelian symmetry G has a reduced dynamics determined by a variational principle of the form

$$\mathcal{A}_R[\gamma] \doteq \int_0^T \mathrm{d}t\, R(q(t),\dot{q}(t)) - \int_\Sigma \beta_\mu(q(t),\dot{q}(t))\,, \tag{50}$$

where R is the so-called *Routhian* and β_μ is a 2-form dependent on the conserved momentum μ. The construction of the Routhian and of the reduced action $\mathcal{A}_R[.]$ can be found in [34].

Fixed-Energy Problem, Hill's Region

Let us consider Hamiltonian systems $(\mathcal{P}, \omega, H)$ where the phase space a cotangent bundle $\mathcal{P} = T^*\mathcal{M} \simeq \mathbb{R}^{2n}$. The symplectic form is then given by $\omega = \mathrm{d}\theta$, where θ is the canonical form, and the Hamiltonian is

$$H(p, q) = \frac{1}{2}\langle p, p\rangle + V(q) \tag{51}$$

$\langle .,. \rangle$ is a Riemannian metric on $T_q^*\mathcal{M}$. The Hamiltonian flow preserves the Hamiltonian function; therefore, a natural problem is to restrict the dynamics to the manifold defined by a fixed constant value of the Hamiltonian that physically corresponds to fixing the energy. Letting E be the energy value, one can define this natural constraint by constructing the following submanifold of the phase space $\mathcal{P}$:

$$\Sigma_E = \{(p, q) \in \mathcal{P} \colon H(p, q) = E\}\,.$$

From Σ_E it is possible to construct a new manifold that corresponds to the image of the projection of Σ_E onto the configuration space $\mathcal{M}$. To obtain such a projection it is sufficient to look at (51) and observe that the norm of p on T_q^*M cannot be negative. From this, the new space, *Hill's region*, turns out to be defined as follows:

Definition 27 (Hill's region)

$$P_E \doteq \{q \in \mathcal{M} \colon E - V(q) \geq 0\}$$

Remark 12 The manifold P_E depends on the values of E and may have boundaries. For instance, consider the Hamiltonian

$$H = \frac{1}{2}\|p\|^2 - \frac{\lambda}{2}\|q\|^2 + \frac{1}{4}\|q\|^4 \text{ with } \lambda > 0\,,$$

which has a Hill region defined by

$$P_E \doteq \left\{q \in \mathcal{M} \colon E + \frac{\lambda}{2}\|q\|^2 - \frac{1}{4}\|q\|^4 \geq 0\right\}\,.$$

First note that P_E is a subset of $\mathbb{R}^2$ and that there are the following cases:

- $P_E = \emptyset$ for $E < -\lambda^2/4$,
- P_E is a circle for $E = -\lambda^2/4$,
- P_E is an annulus for $-\lambda^2/4 < E < 0$,
- P_E is a disk minus its center for $E = 0$,
- P_E is a disk for $E > 0$.

Note that if P_E is not empty, then the boundary of ∂P_E is given by

$$\partial P_E \doteq \left\{q \in \mathcal{M} \colon E + \frac{\lambda}{2}\|q\|^2 - \frac{1}{4}\|q\|^4 = 0\right\}$$

and is not empty. The boundary corresponds to the set of points where all momenta p vanish.

The Hill region has a topology which changes according to the values of the energy; hence, it is a natural problem to search for periodic orbits in it. What are the possible orbits in P_E? Assuming there is a generic Hill region with a nonempty boundary, there are two possible types of orbits:

(A) Orbits joining two points of the boundary, the *brake orbits*,
(B) Orbits not intersecting the boundary, the *internal periodic orbits*.

In general given a Hamiltonian system with a nonempty Hill region P_E we define

Definition 28 (Brake orbit) A solution $q(.)$ of the Hamilton equation is a *brake orbit* of period T if $q(t+T) = q(t)$, $q(t) \in P_E \backslash \partial P_E$ for all $t \in (0, T/2)$ and $q(0) \in \partial P_E$, $q(T/2) \in \partial P_E$.

Definition 29 (Internal periodic orbit) A solution $q(.)$ of the Hamilton equation is an *internal periodic orbit* of period T if $q(t + T) = q(t)$ and $q(t) \in P_E \backslash \partial P_E$ for all t.

Remark 13 In principle one could think that brake orbits could have more than two intersections with the boundary ∂P_E. This is not possible in systems with Hamiltonian (51) because the velocity on ∂P_E is zero and the Hamiltonian equations are "reversible", that is, if $q(t)$ is a solution then $q(-t)$ is a solution. These two properties imply that any solution $q(t)$ such that $q(t_i) \in \partial P_E$ follows the trajectory $q(-t)$ for $t > t_i$ with reversed velocity [24,30].

An example of Hill's region with brake orbits and internal periodic orbit is given in Fig. 4.

Jacobi Metric and Tonelli Functional The general approach to study the periodic orbits of type A or B is to use

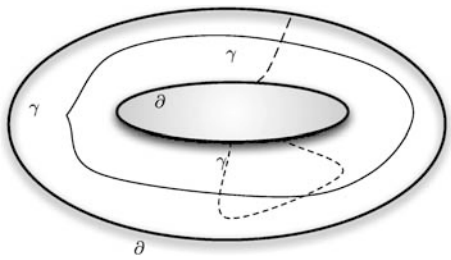

Periodic Orbits of Hamiltonian Systems, Figure 4
Example of a compact nonsimply connected Hill region with two brake orbits γ_{brake} and one internal orbit γ_{internal} that cannot be deformed into a point

a variational principle which is naturally defined on paths in P_E. Let us consider

Definition 30 (Jacobi metric)

$$\mathcal{J}_E[q] = \frac{1}{2}\int_0^1 \mathrm{d}s(E - V(q(s)) \left\| \frac{\mathrm{d}q(s)}{\mathrm{d}s} \right\|^2 . \tag{52}$$

The domain of $\mathcal{J}_E[.]$ is $\Lambda(P_E) \doteq \{q(.) \in H^1([0,1],\mathbb{R}^n)$: $q(s+1) = q(s)\ q(s) \in P_E\}$, and the following result holds:

Proposition 10 *Let $q(.)$ be a path in $\Lambda(P_E)$ where $\mathcal{J}_E[.]$ is differentiable and let $q(.) + \epsilon\, v(.) \in \Lambda(P_E \backslash \partial P_E)$ for all ϵ sufficiently small, then*

$$D\mathcal{J}_E[q](v) = 0 \quad \textit{for all} \quad v(.) \in T\Lambda(P_E)$$

is equivalent to

$$\frac{dq(t)}{dt} = -\nabla V(q(t))$$

after a time rescaling defined by

$$\frac{dt}{ds} = \left\| \frac{dq(s)}{ds} \right\| \frac{1}{\sqrt{E - V(q(s))}} .$$

The proof of this proposition is quite standard and can be found in [1,8,11,30,51]. Let us observe that in the regions where $E - V(q) > 0$ the functional $\mathcal{J}_E[.]$ can be derived from the Riemannian metric

$$\mathrm{d}^2\ell = (E - V(q)) \sum_{i,j} \delta_{ij} \mathrm{d}q_i \otimes \mathrm{d}q_j \tag{53}$$

and indeed from this its name originated. From the Jacobi metric a length can be defined

$$\ell[q] \doteq \int_0^1 \mathrm{d}s \sqrt{(E - V(q(s))\|\dot{q}(s)\|^2} .$$

The Jacobi metric (53) transformed the study of the classical Newton equations into the study of the geodesics on P_E. It is crucial to note that the Jacobi metric is vanishing on ∂P_E and therefore cannot be complete in P_E whenever the boundary is not empty. There is another functional that can be used to study periodic orbits in P_E. This is defined as follows:

Definition 31 (Tonelli functional)

$$\mathcal{T}_E[q] = \frac{1}{2}\int_0^1 \mathrm{d}s(E - V(q(s)) \int_0^1 \mathrm{d}s \left\| \frac{\mathrm{d}q(s)}{\mathrm{d}s} \right\|^2 .$$

Tonelli's functional is a product of two integrals and therefore does not have a natural geometrical interpretation like (53). One can observe that (by Schwartz inequality) the Tonelli functional is bounded from below by Jacobi length: $(\ell[q])^2 \leq 2\mathcal{T}_E[q]$. Note that also $\mathcal{T}_E[.]$ vanishes on ∂P_E. The functional $\mathcal{T}_E[.]$ is defined on $\Lambda(P_E)$ and provides another possible variational description of Newton's equations. Indeed one can easily show [2].

Proposition 11 *Let $q(.)$ be a path in $\Lambda(P_E)$ where $\mathcal{T}_E[.]$ is differentiable and let $q(.) + \epsilon\, v(.) \in \Lambda(P_E \backslash \partial P_E)$ for all ϵ sufficiently small, then*

$$D\mathcal{T}_E[q](v) = 0 \quad \textit{for all} \quad v(.) \in T\Lambda(P_E)$$

is equivalent to

$$\frac{dq(t)}{dt} = -\nabla V(q(t))$$

after time rescaling defined by

$$\left(\frac{dt}{ds}\right)^2 = \frac{\int_0^1 ds \left\| \frac{dq(s)}{ds} \right\|^2}{\int_0^1 ds(E - V(q(s))} .$$

Remark 14 Observe that in both the Jacobi and the Tonelli formulation there is a reparameterization of the time and therefore one can always restrict the consideration to trajectories with unit period.

In variational methods we aim to show that the critical set is not empty and, if possible, to estimate its cardinality. This is certainly affected by the topology of the space, where the paths are defined. Let us now consider the Jacobi metric $\mathcal{J}_E[.]$. This functional depends on the energy E, which affects the properties of P_E and $\Lambda(P_E)$. There are four possible cases:

- P_E is compact and $\partial P_E = \emptyset$,
- P_E is compact and $\partial P_E \neq \emptyset$,
- P_E is not compact and $\partial P_E = \emptyset$,
- P_E is not compact and $\partial P_E \neq \emptyset$.

In the next section we shall illustrate some results regarding the preceding four cases. We shall see that the main approaches have much in common with Riemannian geometry. For more details on variational methods the reader is encouraged to consult [2,15,28,33,53].

Orbits in Compact Hill's Regions Without a Boundary Let us consider a region P_E without a boundary for $E > \sup_{q\in\mathcal{M}} V(q)$. If $\mathcal{M}$ is compact, then the previous condition can be satisfied for some finite E. In this case

$P_E \simeq \mathcal{M}$ and the Jacobi metric $\mathcal{J}_E[.]$ becomes a Riemannian metric on $\mathcal{M}$; therefore, the problem of periodic orbits in $\mathcal{M}$ translates into the problem of closed geodesics in the Riemannian manifold $\mathcal{M}$. This is solved by

Theorem 15 (Lusternik and Fet) *Each compact Riemannian manifold contains a closed geodesic.*

For a proof see [28,29]. The main tool for proving the theorem is the so-called *curve shortening*. The manifold P_E is Riemannian with metric (53). A base for the topology is given by $B_r(q) = \{q' \in \mathcal{M}\colon \mathrm{d}(q, q') < r\}$, where

$$d(q, q') = \inf\{\ell[\gamma],\ \gamma(.) \text{ is a piecewise smooth curve such that } \gamma(0) = q \text{ and } \gamma(1) = q'\}\ .$$

Remark 15 The reader may observe that if P_E has a nonempty boundary then $d(.,.)$ turns into a pseudometric because $d(q, q') = 0$ for all $q, q' \in \partial P_E$. Certainly $d(.,.)$ is still a metric restricted to the open set $P_E \backslash \partial P_E$.

Let us now give a sketch of the curve-shortening method. A standard result in Riemannian geometry guarantees that given a point q_0 and a neighborhood $B_\delta(q_0)$, there exists δ such that any point in $\partial B_\delta(q_0)$ can be joined to q_0 by a unique geodesic [28]. Since $\mathcal{M}$ is compact one can use a finite family of such neighborhoods to join any two points $q_0, q_1 \in \mathcal{M}$ with a piecewise geodesic $\gamma(.)$. Now consider on $\mathcal{M}$ the space of closed paths of class $H^1([0,1], \mathcal{M})$ and consider a sequence $0 \le t_0^{(k)} < t_1^{(k)} < \cdots < t_{n-1}^{(k)} < t_n^{(k)} \le 1$ such that $t_{i+1} - t_i < \delta^2/(2c)$. Take a curve $\gamma(.) \in H^1([0,1], \mathcal{M})$ with $\mathcal{J}_E[\gamma] \le c$ then define

$$S^{(k)}(\gamma) = \begin{cases} \text{is a piecewise geodesic curve} \\ \quad \text{where } S^{(k)}(\gamma)(t) \text{ is restricted to } t_i^{(k)}, t_{i+1}^{(k)}, \\ \text{is a geodesic joining } \gamma t_i^{(k)} \text{ to } \gamma t_{i+1}^{(k)} \\ \quad \text{for } i = 1 \ldots n\ . \end{cases}$$

Now clearly $\mathcal{J}_E[S^{(k)}(\gamma)] \le \mathcal{J}_E[\gamma]$, $\ell[S^{(k)}(\gamma)] \le \ell[\gamma]$. The map $S^{(k+1)}(.)$ is constructed by taking a new partition of $[0,1]$ such that $t_i^{(k)} < t_i^{(k+1)} < t_{i+1}^{(k)}$. Since one can easily verify

$$\ell[S^{(k+1)}(\gamma)] \le \ell[S^{(k)}(\gamma)] \tag{54}$$

the iteration of $S^{(k)}(.)$ is called *curve shortening*. In [24,28], it is shown that $S^{(k)}(\gamma)$ converges uniformly to a geodesic. The limit could be just a point curve. To show that this cannot be the case one uses the fact that on a compact manifold of dimension n there is always a map $h\colon S^d \mapsto \mathcal{M}$ (with $1 \le d \le n$) which is not homotopic to a constant map.

There have been many generalizations of Lusternik–Fet theorem and the reader is invited to refer to [29].

Brake Orbits in Hill's Regions with a Boundary

Let P_E be a region with boundary for $\inf_{q\in\mathcal{M}} V(q) < E < \inf_{q\in\mathcal{M}} V(q)$. In this case there is the following result:

Theorem 16 ([11]) *Suppose that P_E is compact and there are no equilibrium positions in ∂P_E. Then the number of brake orbits in P_E is at least equal to the number of generators of $\pi_1(P_E \backslash \partial P_E)$.*

Since the Jacobi metric is singular on ∂P_E it is necessary to analyze the geodesic motion near the boundary. This type of analysis goes back to earlier works [9,10,51]. The main point is to construct a new region $P_{E-\epsilon} \subset P_E$ on which $\mathcal{J}_E[.]$ is positive-definite. The next step is to construct a geodesic joining two points on the new boundary $\partial P_{E-\epsilon}$. Finally the construction on $P_{E-\epsilon}$ is used to show that the geodesic $q^\epsilon(.)$ becomes a brake orbit in the limit $\epsilon \to 0$. In the case of noncompact P_E, the existence of a brake orbit was proved in [44]. This result is based on two assumptions:

(i) ∂P_E is not empty and it is formed by two connected components A_1 and A_2 such that if $x_n \in A_1$ and $y_n \in A_2$ with $\|x_n\| \to \infty$, $\|y_n\| \to \infty$ then $\|x_n - y_n\| \to \infty$.
(ii) Let $R_\delta = \{q \in \mathcal{M}\colon E - \delta < V(q) < E\}$ for $\delta > 0$. If either A_1 or A_2 is not compact, there is a number $r > 0$ with the following property: if $r_0 > r$ then there exist $r_1 > r_0$ and $\delta > \delta^* > 0$ such that for every $q \in R_\delta \cap \{q\colon \|q\| > r_1\}$ and $(q, p) \in H^{-1}(E)$ implies that the Hamiltonian flow $\phi(t, q, p)$ stays in $\{q\colon \|q\| > r_1\}$ for all $t \ge 0$ where defined.

Theorem 17 ([44]) *Suppose that P_E is connected, $\inf_{q\in\partial P_E} \|\nabla V(q)\| > 0$ and conditions i and ii hold true. Then there exists a periodic solution which is a brake orbit.*

In this result the lack of compactness of P_E does not allow us to use the curve shortening. For this reason in [44] the direct minimization was employed. First a new region $P_{E-\delta} \subset P_E$ is defined for $\delta > 0$. Such region now has a boundary with two different connected components, A_1^δ and A_2^δ, respectively. In [44] the following minimization problem was studied

$$\min \mathcal{J}_E[q] \text{ where } q(.) \in H^1$$
$$\text{and boundary conditions } q(0) \in A_1^\delta,\ q(0) \in A_1^\delta\ .$$

Condition i is able to control the lack of compactness of P_E and to show that the minimization problem admits a solution $q^\delta(.)$. Condition ii allows us to take the limit $\delta \to 0$ to obtain a brake orbit in P_E.

Orbits in Closed, Nonsimply Connected Hill's Regions with Boundaries Let us now consider a general Hill region which is not necessarily compact, with a boundary and with nontrivial homotopy group $\pi_1(P_E)$. The natural problem is to determine under which condition it is possible to prove the existence of a internal periodic orbit within a specific homotopy class of P_E. A partial answer is given by

Theorem 18 ([6,7]) *Suppose that P_E is closed and bounded and that $\nabla V(q) \neq 0$ for all $q \in \partial P_E$. Then there exists a periodic orbit $q(s) \in P_E$ for all s. The orbit $q(.)$ can be either a brake orbit or an internal periodic orbit.*

As we have already seen, the Jacobi metric (but also the Tonelli functional) is degenerate on the boundary ∂P_E, i. e., $\mathcal{J}_E[\gamma]$ is identically zero on every closed path $\gamma(.)$ such that $\gamma(s) \in \partial P_E$ for every $s \in [0,1]$. In order to avoid this problem, in [6,7], a modified functional was introduced:

$$\mathcal{J}^\epsilon[q] \doteq \mathcal{J}_E[q] + \int_0^1 \mathrm{d}s\, U^\epsilon[q(s)] \,.$$

The functional $\mathcal{J}^\epsilon[.]$ is called *penalized*. If $q_n(.)$ is a sequence of curves in $\overline{\Lambda(P_E)}$ (the closure of $\Lambda(P_E)$) weakly converging to a curve $q^*(.)$ intersecting the boundary, then the form of $U^\epsilon(.)$ implies that $\mathcal{J}^{(\epsilon)}[q_n] \to +\infty$ whenever $\epsilon > 0$. This type of penalization is similar to the so-called *strong force* potential used in the N-body problem (see [2,26] and also ► *n*-Body Problem and Choreographies). In this case there are two kinds of difficulties: P_E is no longer compact and the boundary ∂P_E is not empty. The strategy is as follows: prove that the critical level sets of $\mathcal{J}^\epsilon[.]$ are precompact. This is obtained by a generalization of the Palais–Smale condition. This prevents the sequences from converging on the boundary, and it is realized by taking a functional $\rho\colon C^1(\Lambda(P_E)) \to \mathbb{R}_+$ such that if $q_n(.)$ converges to a path intersecting the boundary, then $\rho(q_n) \to +\infty$. This allows us to introduce:

Definition 32 (Weighted Palais–Smale condition [6]) The action $\mathcal{J}^\epsilon[.]$ satisfies the weighted Palais–Smale condition if any sequence $q_n(.)$ fulfills one of the following alternatives:

1. $\rho(q_n)$ and $\mathcal{J}^\epsilon[q_n]$ are bounded and $D\mathcal{J}^\epsilon[q_n] \to 0$ and q_n has a convergent subsequence,
2. $\mathcal{J}^\epsilon[q_n]$ is convergent and $\rho(q_n) \to +\infty$ and there exists $\nu > 0$ such that $\|D\mathcal{J}^\epsilon[q_n]\| \geq \nu \|D\rho(q_n)\|$ for n sufficiently large.

The verification of the weighted Palais–Smale condition guarantees that the set $K_c \cap \{q\colon \rho(q) \leq M\}$ is compact for every $M > 0$. The next step in the proof is to construct a mini-max structure [47] that allows us to identify the critical values. This is achieved by constructing two manifolds Q and S such that:

(i) $S \cap \partial Q = \emptyset$,
(ii) If $u \in C^0(\overline{Q}, \Lambda(P_E))$ such that $u(q(.)) = q(.)$ for every $q(.) \in \partial Q$ then $h(\overline{Q}) \cap S \neq \emptyset$.

The manifolds Q and S are defined such that the critical level

$$c = \inf_{u \in U} \sup_{q \in Q} \mathcal{J}^\epsilon(h(q))$$

is finite. Here $U = \{u \in C^0(\overline{Q}, \overline{\Lambda(P_E)})\colon u(q(.)) = q(.) \text{ if } \mathcal{J}^\epsilon[q] \leq 0\}$. Then the weighted Palais–Smale condition allows us to construct a gradient flow and to prove the existence of a critical point $q^\epsilon(.)$. In [7] it is shown also that there exists $\alpha \leq \beta$ independent of ϵ such that

$$\alpha \leq \mathcal{J}^\epsilon[q^\epsilon] \leq \beta \,.$$

This uniform estimates make it possible to prove that for $\epsilon \to 0$ the path $q^\epsilon(.)$ converges uniformly to a closed path $q^0(.)$ such that the set $I \doteq \{t \in [0,1]\colon q^0(t) \in P_E \backslash \partial P_E\}$ is not empty. Since $q^0(.)$ is continuous then the interval I has connected components. Note that $q^0(.)$ has the same homotopy type as $q^\epsilon(.)$, but in general this is no longer true for the solution of the equations of motion. In fact $q^0(.)$ is a solution only on the connected component of I. Let Δ be one connected component of I. On Δ the path $q^0(.)$ solves the equations of motion. Therefore, the topological characterization of the periodic orbit depends on Δ. If $\Delta \subset [0,1]$ then $q^0(.)$ is a *brake orbit*, if $\Delta = [0,1]$ then $q^0(.)$ is an *internal periodic orbit*. The problem of finding internal periodic orbits in any prescribed homotopy class is still open.

Continuation of Periodic Orbits as Critical Points

Variational methods can also be very useful to study the problem of continuation of periodic orbits. Here we restrict ourselves to a few examples.

Lagrangian Variational Principle Let us consider a dynamical system in $\mathbb{R}^n$ that can be written in the Lagrangian form:

$$L(v_q, q) = \frac{1}{2}\|v_q\| - V(q)\,.$$

Let us assume

(i) There exists $a > 0$ such that $V(\lambda\, q) = \lambda^{-a}\, V(q)$ for any $\lambda > 0$,
(ii) There exists a 1-periodic orbit $q_1(.)$.

The orbit $q_1(.)$ is a critical point of $\mathcal{A}_L[q] = \int_0^1 \mathrm{d}s L(\dot{q}(s), q(s))$. Now let us consider another potential term $W(q)$ such that

$$W(\lambda\, q) = \lambda^{-b}\, W(q) \quad \text{with} \quad b > a\,.$$

Let us look for periodic orbits of period T for the system whose Lagrangian is

$$L(v_q, q) = \frac{1}{2}\|v_q\| - V(q) - W(q)\,.$$

This suggests searching for critical points $x(.)$ of the functional

$$\mathcal{A}_L^T[x] = \int_0^T \mathrm{d}t \left[\frac{\|\dot{x}(t)\|^2}{2} - V(x(t)) - W(x(t))\right]. \quad (55)$$

The scaling properties of V and W can be used to construct a perturbation argument. Define

$$x(t) = T^{-p}\, q(t/T)\,, \text{ where } q(s) \text{ is defined for } s \in [0,1]\,, \text{ and } p = 2/(a+2)\,. \quad (56)$$

The dynamical equations for $x(.)$ are equivalent to the Lagrangian equation for

$$\mathcal{A}_L[q;\epsilon] = \int_0^1 \mathrm{d}s \left[\frac{\|\dot{q}(s)\|^2}{2} - V(q(s)) - \epsilon\, W(q(s))\right], \quad (57)$$

where

$$\epsilon \doteq T^{\frac{-2(b-a)}{a+2}}\,.$$

The scaling properties of the potential allow us to transform the given problem into a perturbation problem. Note that there is the following correspondence: a small perturbation for the scaled action (57) corresponds to large periods for (55). Now $\mathcal{A}_L[.,0] = \mathcal{A}_L^1[.]$ and in general the critical point $q_1(.)$ is not isolated and therefore $D^2\mathcal{A}_L[q_1, 0]$ has some degeneracy. If the manifold of critical points is degenerate along normal directions (*normal degeneracy*), then it is possible to continue the critical point $q_1(.)$ into a T-periodic orbit by decomposing the continuation procedure. This approach is the so-called *Liapunov–Schmidt reduction*; it is used to generalize the IFT to situations where it cannot be applied directly. An example will be illustrated in the last part of this section in the study of Hamiltonian systems. For details about normal degeneracy the reader could consult [15].

Hamiltonian Variational Principle Hamiltonian equations can be thought of as a vector field on a space of loops. The geometrical construction was described in [56]. Let us consider the space $C^1([0,1], \mathbb{R}^n)$ of differentiable loops. We can define

$$\mathcal{X}(z) \doteq \nu \frac{\mathrm{d}z(s)}{\mathrm{d}s} - J\nabla H(z(s)),\ z(.) \in C^1([0,1], \mathbb{R}^{2n}). \quad (58)$$

The zero set of $\mathcal{X}(z)$ is formed by loops $z(.)$ such that

$$\nu \frac{\mathrm{d}z(s)}{\mathrm{d}s} = J\nabla H(z(s))\,.$$

This corresponds to a periodic orbit $z(.)$ with period $2\pi/\nu$. It turns out that $\mathcal{X}(.)$ is a Hamiltonian vector field whose Hamiltonian on $C^1([0,1], \mathbb{R}^{2n})$ [25,56] is given by

$$\mathcal{H}[z] = \frac{1}{2\pi}\int_0^{2\pi} \mathrm{d}s\, \{\langle \nu\, J\dot{z}(s), z(s)\rangle - H(z(s))\} \quad (59)$$

and with symplectic form

$$\Omega[z, w] = \frac{1}{2\pi}\int_0^{2\pi} \mathrm{d}s \langle Jz(s), w(s)\rangle\,. \quad (60)$$

Indeed a simple calculation shows that

$$\Omega(\mathcal{X}(z), w) = \mathrm{d}\mathcal{H}[z](w)\,.$$

Let us now suppose there is a periodic orbit $z_0(.)$. We want to illustrate how to study the continuation and bifurcation when the Hamiltonian is perturbed. In order to use the formulation in the loop space we can introduce the Liapunov–Schmidt reduction. In fact, in general, Hamiltonian vector fields at a periodic orbit have a linearization

with a nonempty kernel. The construction we now present is a very well known approach in the infinite-dimensional setting.

Liapunov–Schmidt Reduction for Hamiltonian Systems [3,25] Let $(\mathcal{H}, \mathcal{P}, \Omega)$ be a Hamiltonian system with $\mathcal{P} = C^1([0,1], \mathbb{R}^{2n})$. Let $\mathcal{X}: \mathcal{P} \to \mathcal{Y}$ be the Hamiltonian vector field such that $z_0(.)$ is a zero of $\mathcal{X}(.)$ and

(i) $L = D\mathcal{X}(z_0)$ is Fredholm,
(ii) $J(\ker(L)) = \ker(L)$.

Condition i implies that it is possible to make the following decomposition

1. $\mathcal{P} = \ker(L) \oplus W$, W closed subspace,
2. $\mathcal{Y} = \mathrm{rank}(L) \oplus N$, N closed subspace.

In fact any z can be written as $z = k + w$ with $k \in \ker(L)$ and $w \in W$. This allows us to reduce $\mathcal{X}(z) = 0$ to solving the following system of equations:

$$\begin{cases} \Pi\, \mathcal{X}(k+w) = 0 \in \mathrm{rank}(L) \\ (\mathbf{I} - \Pi)\, \mathcal{X}(k+w) = 0 \in N \end{cases} \tag{61}$$

where $\Pi: \mathcal{Y} \mapsto \mathrm{rank}(L)$. The first equation in (61) can be solved with respect to w by the IFT. The second equation becomes the so-called *bifurcation* equation:

$$X_g(k) = (\mathbf{I} - \Pi)\mathcal{X}(k + w(k)) = 0\,. \tag{62}$$

Note that since L is Fredholm $\dim\ker(L) < \infty$, (62) is a finite-dimensional problem. Now condition ii implies that X_g can be thought of as a map from $\ker(L)$ into itself. Furthermore this allows us to show that X_g is a Hamiltonian vector field with Hamiltonian

$$g(k) = \mathcal{H}(k + w(k))$$

and symplectic form $\Omega(.,.)$. If condition ii does not hold a further condition has to be included in order to guarantee that X_g is a Hamiltonian vector field. The bifurcation equation becomes therefore

$$\Omega(X_g(k), u) - \mathrm{d}\mathcal{H}(k + w(k))(u) = 0 \;\forall u\,. \tag{63}$$

In [56] the loop space approach and the reduction were used to show that if a Hamiltonian system has a manifold Σ of periodic orbits whose tangent space coincides with the kernel of $D^2\mathcal{A}_H$ (*nondegeneracy condition*) then small perturbations of the Hamiltonian cannot destroy all the periodic orbits:

Theorem 19 ([56]) *Let Σ be a compact, nondegenerate manifold of periodic orbits for a Hamiltonian system $(\mathcal{P}, \omega, H)$. Given any neighborhood U of Σ there exists $\epsilon_0 > 0$ such that for $|\epsilon| < \epsilon_0$, the number of periodic orbits in U for $(\mathcal{P}, \omega, H + \epsilon H_1)$ is no less than the Lusternik–Schnirelman category* $\mathrm{Cat}(\Sigma/S^1)$. *If Σ satisfies the additional condition that the algebraic multiplicity of 1 as an eigenvalue of the Poincaré map is uniformly equal to the* $\dim \Sigma$, *then* $\mathrm{Cat}(\Sigma/S^1) \geq (1 + \dim \Sigma)/2$.

In [25] Liapunov–Schmidt reduction was used to prove the Liapunov center theorem and also the Hamiltonian Hopf bifurcation.

Further Directions

This article has focused on periodic orbits in Hamiltonian systems. This problem can be seen as an organizing center in the history of the development of modern mathematics. In fact it is related to many theoretical aspects: bifurcation theory, symmetry reduction, variational methods and topology of closed curves on manifolds. Still, open problems remain, for example, one would like to prove the existence of internal periodic orbits in every homotopy class of a Hill region. Another interesting direction is the analytical study of the action functional that results from symmetry reduction. A further generalization of the Hamiltonian formalism is the study of the so-called multiperiodic patterns. Here is an example. Let ϕ, p be two functions defined on the two-dimensional torus $\mathbb{T}^2$ and valued in $\mathbb{R}$. The functions ϕ, p are 2-periodic because they satisfy $p(x+\tau_1, y+\tau_2) = p(x,y)$, $\quad \phi(x+\tau_1, y+\tau_2) = \phi(x)$, where $(x,y) + (\tau_1,\tau_2) \simeq (x,y)$ in $\mathbb{T}^2$. One can pose the problem to solve a couple of partial differential equations of the form

$$\begin{aligned} &\nabla^2\phi + \frac{\partial}{\partial_\phi} V(\phi, p, x, y) = 0 \quad \text{and} \\ &\nabla^2 p + \frac{\partial}{\partial_p} V(\phi, p, x, y) = 0\,. \end{aligned} \tag{64}$$

Here $V: \mathbb{R}^2 \mapsto \mathbb{R}$ is a smooth potential function. Note that the periodicity of ϕ and p is partial, that is why the term "multiperiodic patterns" is used. It has been shown that Eqs. (64) admits a description in terms of finite-dimensional multi-symplectic structure and the equations of motion can be cast into a general Hamiltonian variational principle. In fact one can define $Z = (\phi, u_1, u_2, p)$ defined as functions on $\mathbb{T}^2$ and the Hamiltonian

$$S(Z, x, y) = \frac{1}{2}(u_1^2 + u_2^2) + V(\phi, p, x, y)\,.$$

Then Eqs. (64) turn out to be equivalent to the variational equation

$$D\mathcal{L}(Z) = 0\,, \tag{65}$$

where $\mathcal{L}$ is the action functional

$$\mathcal{L}(Z) = \int_0^{\tau_1} \mathrm{d}x \int_0^{\tau_2} \mathrm{d}y \left(\frac{1}{2} \left\langle J_1 \frac{\partial Z}{\partial x} + J_2 \frac{\partial Z}{\partial y}, Z \right\rangle - S(Z, x, y) \right) \tag{66}$$

defined on 2-periodic maps $Z\colon \mathbb{T}^2 \mapsto \mathbb{R}^4$. The matrices J_1 and J_2 are two symplectic structures and form a multisymplectic structure. In this example J_1 and J_2 read

$$J_1 = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad \text{and} \quad J_2 = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}. \tag{67}$$

For Eqs. (64) one can pose the problem of finding solutions as "multiperiodic" critical points of the action $\mathcal{L}$, which is a generalization of the Hamiltonian action functional (48). An interesting reference for this problem is [13].

Acknowledgments

The author would like to thank Heinz Hanßmann for his critical and thorough reading of the manuscript and Ferdinand Verhulst for his useful suggestions.

Bibliography

1. Abraham R, Marsden J (1978) Foundations of Mechanics. Benjamin-Cummings, Reading
2. Ambrosetti A, Coti V (1994) Zelati Periodic Solutions of singular Lagrangian Systems. Birkhäuser, Boston
3. Ambrosetti A, Prodi G (1993) A primer on Nonlinear Analysis. CUP, Cambridge UK
4. Arms JM, Cushman R, Gotay MJ (1991) A universal reduction procedure for Hamiltonian group actions. In: Ratiu TS (ed) The Geometry of Hamiltonian Systems. Springer, New York, pp 33–51
5. Arnold VI (1989) Mathematical Methods of Classical Mechanics. Springer, New York
6. Benci V (1984) Normal modes of a Lagrangian system constrained in a potential well. Annales de l'IHP section C tome 5(1):379–400
7. Benci V (1984) Closed geodesics for the Jacobi metric and periodic solutions of prescribed energy of natural Hamiltonian systems. Annales de l'IHP section C tome 5(1):401–412
8. Benci V (1986) Periodic solutions for Lagrangian systems on a compact manifold. J Diff Eq 63:135–161
9. Birkhoff GD (1927) Dynamical Systems with two degrees of freedom. Trans Am Math Soc 18:199–300
10. Birkhoff GD (1927) Dynamical Systems. Colloq. Publ. 9. Amer. Math. Soc., Providence RI
11. Bolotin SV, Kozlov VV (1978) Libration in systems with many degrees of freedom. J Appl Math Mech 42:256–261
12. Bott R, Tu W (1995) Differential forms in algebraic topology. Springer, New York
13. Bridges T (2006) Canonical multi-symplectic structure on the total exterior algebra bundle. Proc R Soc A 462:1531–1551
14. Brown KS (1982) Cohomology of groups. Springer, New York
15. Chang K (1991) Infinite dimensional Morse theory and multiple solution problems. Birkhäuser, Boston
16. Cushman R, Bates L (1997) Global aspects of classical integrable systems. Birkhäuser, Boston
17. Cushmann R, Sadowski D, Efstathiou K (2005) No polar coordinates. In: Montaldi J, Ratiu T (eds) Geometric Mechanics and Symmetry: the Peyresq Lectures. Cambridge University Press, Cambridge UK
18. Dell Antonio G (1995) Agomenti scelti di meccanica SISSA/ISAS Trieste. ref. ILAS/FM-19/1995
19. Doedel EJ, Keller HB, Kernvez JP (1991) Numerical analysis and control of bifurcation problems: (I) Bifurcation in finite dimensions. Int J Bifurcat Chaos 3(1):493–520
20. Doedel EJ, Keller HB, Kernvez JP (1991) Numerical analysis and control of bifurcation problems: (II) Bifurcation in infinite dimensions. Int J Bifurcat Chaos 4(1):745–772
21. Ekeland I (1990) Convexity methods in Hamiltonian mechanics. Springer, Berlin
22. Farber M (2004) Topology of Closed One-Forms. AMS, Providence
23. Gallavotti G (2001) Quasi periodic motions from Hipparchus to Kolmogorov. Rend Mat Acc Lincei s. 9, 12:125–152
24. Gluck H, Ziller W (1983) Existence of periodic motions of conservative systems. In: Bombieri E (ed) Seminar on Minimal Submanifolds. Princeton University Press, Princeton NJ
25. Golubitsky M, Marsden JE, Stewart I, Dellnitz M (1995) The constrained Liapunov–Schmidt procedure and periodic orbits. Fields Institute Communications 4. Fields Institute, Toronto, pp 81–127
26. Gordon W (1975) Conservative dynamical systems involving strong force. Trans Am Math Soc 204:113–135
27. Guillemin V, Sternberg S (1984) Symplectic techniques in Physics. Cambridge University Press, Cambridge UK
28. Jost J (1995) Riemannian geometry and geometric analysis. Springer, Berlin
29. Klingenberg W (1978) Lectures on closed geodesics. Springer, Berlin
30. Kozlov VV (1985) Calculus of variations in the large and classical mechanics. Russ Math Surv 40:37–71
31. Kuznetsov YA (2004) Elements of Applied Bifurcation Theory. Springer, Berlin

32. Leimkuhler B, Reich S (2005) Simulating Hamiltonian Dynamics. CUP, Cambridge UK
33. Mawhin J, Willhelm M (1990) Critical point theory and Hamiltonian systems. Springer, Berlin
34. Marsden J, Ratiu TS, Scheurle J (2000) Reduction theory and the Lagrange–Routh equations. J Math Phys 41:3379–3429
35. McCord C, Montaldi J, Roberts M, Sbano L (2003) Relative Periodic Orbits of Symmetric Lagrangian Systems. Proc. Equadiff. World Scientific, Singapore
36. Meyer K, Hall GR (1991) Introduction to Hamiltonian Systems and the N-Body Problem. Springer, Berlin
37. Meyer K (1999) Periodic solutions of the *N*-body problem. LNM, vol 1719. Springer, Berlin
38. Montaldi J, Buono PL, Laurent F (2005) Poltz Symmetric Hamiltonian Bifurcations. In: Montaldi J, Ratiu T (eds) Geometric Mechanics and Symmetry: the Peyresq Lectures. Cambridge University Press, Cambridge UK
39. Montaldi J, Roberts M, Stewart I (1988) Periodic solutions near equilibria of symmetric Hamiltonian systems. Phil Trans R Soc Lon A 325:237–293
40. Montaldi J, Roberts M, Stewart I (1990) Existence of nonlinear normal modes of symmetric Hamiltonian systems. Nonlinearity 3:695–730
41. Montaldi J, Roberts M, Stewart I (1990) Stability of nonlinear normal modes of symmetric Hamiltonian systems. Nonlinearity 3:731–772
42. Munoz–Almaraz FJ, Freire E, Galan J, Doedel E, Vanderbauwhede (2003) A Continuation of periodic orbits in conservative and Hamiltonian systems. Physica D 181:1–38
43. Novikov SP (1982) The Hamiltonian formalism and a multivalued analogue of Morse theory. Russ Math Surv 37:1–56
44. Offin D (1987) A Class of Periodic Orbits in Classical Mechanics. J Diff Equ 66:9–117
45. Ortega J, Ratiu T (1998) Singular Reduction of Poisson Manifolds. Lett Math Phys 46:359–372
46. Poincaré H (1956) Le Méthodes Nouvelles de la Mécanique Céleste. Gauthiers-Villars, Paris
47. Rabinowitz PH (1986) Minimax Methods in Critical Point Theory with Applications to Differential Equations. In: CBMS Reg. Conf. Ser. No. 56, Amer. Math. Soc., Providence, RI
48. Ratiu T, Sousa Dias E, Sbano L, Terra G, Tudora R (2005) A crush course. In: Montaldi J, Ratiu T (eds) geometric mechanics in Geometric Mechanics and Symmetry: the Peyresq Lectures. Cambridge University Press, Cambridge UK
49. Sanders J, Verhulst F, Murdock J (2007) Averaging Methods in Nonlinear Dynamical Systems. Applied Mathematical Sciences, vol 59. Springer, Berlin
50. Schmah T (2007) A cotangent bundle slice theorem. Diff Geom Appl 25:101–124
51. Seifert H (1948) Periodische Bewegungen mechanischer Systeme. Math Z 51:197–216
52. Simó C (2001) New families of solutions in *N*-body problems. In: European Congress of Mathematics, vol I (Barcelona, 2000), pp 101–115, Progr. Math., 201. Birkhäuser, Basel
53. Struwe M (1990) Variational methods. Springer, Berlin
54. Verhulst F (1990) Nonlinear Dynamical Equations and Dynamical Systems. Springer, Berlin
55. Weinstein A (1973) Normal Modes for Nonlinear Hamiltonian Systems. Inventiones Math 20:47–57
56. Weinstein A (1978) Bifurcations and Hamilton Principle. Math Z 159:235–248
57. Whitehead GW (1995) Elements of homotopy theory, 3rd edn. Springer, Berlin
58. Wulf C, Roberts M (2002) Hamiltonian Systems Near Relative Periodic Orbits. Siam J Appl Dyn Syst 1(1):1–43
59. Yakubovich VA, Starzhinsky VM (1975) Linear Differential Equations with Periodic coefficients, vol 1:2. Wiley, UK

Periodic Solutions of Non-autonomous Ordinary Differential Equations

Jean Mawhin
Département de Mathématique, Université Catholique de Louvain, Maryland, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Poincaré Operator and Linear Systems
Boundedness and Periodicity
Fixed Point Approach: Perturbation Theory
Fixed Point Approach: Large Nonlinearities
Guiding Functions
Lower and Upper Solutions
Direct Method of the Calculus of Variations
Critical Point Theory
Future Directions
Bibliography

Glossary

Banach fixed point theorem If M is a complete metric space with distance d, and $f\colon M \to M$ is contractive, i. e. $d(f(u), f(v)) \leq \alpha d(u, v)$ for some $\alpha \in [0, 1)$ and all $u, v \in M$, then f has a unique fixed point u^* and $u^* = \lim_{k\to\infty} f^k(u_0)$ for any $u_0 \in M$.

Brouwer degree An integer $d_B[f, \Omega]$ which 'algebraically' counts the number of zeros of any continuous mapping $f\colon \overline{\Omega} \subset \mathbb{R}^n \to \mathbb{R}^n$ such that $0 \notin f(\partial\Omega)$, and is invariant for sufficiently small perturbations of f. If f is of class C^1 and its zeros are non degenerate, then $d_B[f, \Omega] = \sum_{x\in f^{-1}(0)} \operatorname{sign} \det f'(x)$.

Brouwer fixed point theorem Any continuous mapping $f\colon B \to B$, with B is homeomorphic to the closed unit ball in $\mathbb{R}^n$, has at least one fixed point.

Leray–Schauder degree The extension $d_{LS}[I - g, \Omega]$ of the Brouwer degree, where Ω is an open bounded subset of the Banach space X, and $g\colon \overline{\Omega} \to X$ is continuous, $g(\overline{\Omega})$ is relatively compact and $0 \notin (I - g)(\partial\Omega)$.

Leray–Schauder–Schaefer fixed point theorem If X is a Banach space, $g\colon X \to X$ is a continuous mapping taking bounded subsets into relatively compact ones, and if the set of possible fixed points of εg ($\varepsilon \in [0,1]$) is bounded independently of ε, then g has at least one fixed point.

Ljusternik–Schnirelmann category The Ljusternik–Schnirelmann category $cat\ (M)$ of a metric space M into itself is the smallest integer k such that M can be covered by k sets contractible in M.

Lower and upper solutions A lower (resp. upper) solution of the periodic problem $u'' = f(t,u)$, $u(0) = u(T)$, $u'(0) = u'(T)$ is a function α (resp. β) of class C^2 such that $\alpha''(t) \geq f(t,\alpha(t)), \alpha(0) = \alpha(T), \alpha'(0) \geq \alpha'(T)$ $\big(resp. \beta''(t) \leq f(t,\beta(t)), \beta(0) = \beta(T), \beta'(0) \leq \beta'(T)\big)$.

Palais–Smale condition for a C^1 function $\varphi\colon X \to \mathbb{R}$ Any sequence $(u_k)_{k\in\mathbb{N}}$ such that $(\varphi(u_k))_{k\in\mathbb{N}}$ is bounded and $\lim_{k\to\infty} \varphi'(u_k) = 0$ contains a convergent subsequence.

Poincaré operator The mapping defined in $\mathbb{R}^n$ by $P_T\colon y \mapsto p(T;y)$, where $p(t;y)$ is the unique solution of the Cauchy problem $x' = f(t,x)$, $x(0) = y$.

Schauder fixed point theorem If C is a closed bounded convex subset of a Banach spaxe X, any continuous mapping $g\colon C \to C$ such that $g(C)$ is relatively compact has at least one fixed point.

Sobolev inequality For any function $u \in L^2(0,T)$ such that $u' \in L^2(0,T)$ and $\int_0^T u(t)\mathrm{d}t = 0$, one has $\max_{t\in[0,T]} |u(t)| \leq (T^{1/2}/2\sqrt{3})[\int_0^T |u'(t)|^2\mathrm{d}t]^{1/2}$.

Wirtinger inequality For any function $u \in L^2(0,T)$ such that $u' \in L^2(0,T)$ and $\int_0^T u(t)\mathrm{d}t = 0$, one has $\int_0^T |u(t)|^2\mathrm{d}t \leq \left(T^2/4\pi^2\right)\int_0^T |u'(t)|^2\mathrm{d}t$.

Definition of the Subject

Many phenomena in nature can be modeled by systems of ordinary differential equations which depend periodically upon time. For example, a linear or nonlinear oscillator can be forced by a periodic external force, and an important question is to known if the oscillator can exhibit a periodic response under this forcing. This question originated from problems in classical and celestial mechanics, before receiving important applications in radioelectricity and electronics. Nowadays, it also plays a great role in mathematical biology and population dynamics, as well as in mathematical economics, where the considered systems are often subject to seasonal variations. The general theory originated with Henri Poincaré's work in celestial mechanics, at the end of the XIXth century, and has been constantly developed since.

Introduction

To motivate the problem and its difficulties, let us start with the simple linear oscillator with forcing (or input) $h \in L^2(0,2\pi)$

$$L_\lambda u := -u'' - \lambda u = h(t)\,, \tag{1}$$

where $\lambda \in \mathbb{R}$. We are interested in discussing the existence or non-existence, and the uniqueness or multiplicity of a 2π-periodic solution u of (1). Let

$$h(t) \sim c_0 + \sum_{k=1}^{\infty}[c_k \cos kt + d_k \sin kt]\,,$$

$$u(t) \sim a_0 + \sum_{k=1}^{\infty}[a_k \cos kt + b_k \sin kt]\,,$$

with

$$c_0 = \frac{1}{2\pi}\int_0^{2\pi} h(t)\mathrm{d}t\,,$$

$$\left\{\begin{array}{c} c_k \\ d_k \end{array}\right\} = \frac{1}{\pi}\int_0^{2\pi} h(t)\left\{\begin{array}{c} \cos kt \\ \sin kt \end{array}\right\}\mathrm{d}t \quad (k = 1,2,\ldots)$$

and similarly for u, be the Fourier series of h and u. If u is a possible 2π-periodic solution of (1), u is of class C^1 and $u'' \in L^2(0,2\pi)$, so that Fourier series of u and u' converge uniformly on $[0,2\pi]$ to u and u'. Furthermore, from Parseval equality

$$\|h\|_2^2 := \frac{1}{2\pi}\int_0^{2\pi} h^2(t)\mathrm{d}t = c_0^2 + \frac{1}{2}\sum_{k=1}^{\infty}[c_k^2 + d_k^2]\,,$$

and, as $u'' \in L^2(0,2\pi)$, $u''(t) \sim -\sum_{k=1}^{\infty} k^2[a_k \cos kt + b_k \sin kt]$. Therefore, finding the 2π-periodic solutions of (1) is equivalent to solving the infinite-dimensional linear system in l^2 with unknowns $(a_0, a_1, b_1, \ldots)$

$$-\lambda a_0 = c_0\,, \quad (k^2-\lambda)a_k = c_k, \quad (k^2-\lambda)b_k = d_k \quad (k = 1,2,\ldots)\,. \tag{2}$$

Letting

$$\Sigma := \{k^2\colon k = 0,1,2,\ldots\}\,,$$

we see that if $\lambda \notin \Sigma$ (non-resonance), system (2) has the unique solution

$$a_0 = -\frac{c_0}{\lambda}, \quad a_k = \frac{c_k}{k^2-\lambda}, \quad b_k = \frac{d_k}{k^2-\lambda} \quad (k = 1,2,\ldots),$$

which gives the unique 2π-periodic solution of (1)

$$u(t) = (L_\lambda^{-1}h)(t) = -\frac{c_0}{\lambda} + \sum_{k=1}^{\infty} \frac{1}{k^2-\lambda}[c_k \cos kt + d_k \sin kt].$$

Furthermore, using Parseval equality,

$$\|u\|_2^2 = \|L^{-1}h\|_2^2 = \frac{c_0^2}{\lambda^2} + \frac{1}{2}\sum_{k=1}^{\infty} \frac{1}{(k^2-\lambda)^2}[c_k^2 + d_k^2] \quad (3)$$

$$\leq \frac{1}{[\mathrm{dist}(\lambda, \Sigma)]^2}\left[c_0^2 + \frac{1}{2}\sum_{k=1}^{\infty}(c_k^2 + d_k^2)\right] = \frac{1}{[\mathrm{dist}(\lambda, \Sigma)]^2}\|h\|_2^2 . \quad (4)$$

Now, if $\lambda = j^2 \in \Sigma$ (resonance) and if $c_j \neq 0$ or $d_j \neq 0$, then (2) has no solution and (1) has no 2π-periodic solution. If $c_j = d_j = 0$, then (1) has the infinite family of 2π-periodic solutions

$$u(t) = \alpha + \sum_{k=1}^{\infty} \frac{1}{k^2}[c_k \cos kt + d_k \sin kt] \quad (\alpha \in \mathbb{R})$$

if $j = 0$ and

$$u(t) = \alpha \cos jt + \beta \sin jt - \frac{c_0}{j^2} + \sum_{\substack{k=1\\k\neq j}}^{\infty} \frac{1}{k^2-j^2}[c_k \cos kt + d_k \sin kt]$$

$(\alpha, \beta \in \mathbb{R})$ if $j \neq 0$.

For the nonlinear oscillator

$$-u'' = g(u) + h(t) \quad (5)$$

where $g\colon \mathbb{R} \to \mathbb{R}$ is continuous, a nonlinear version of the non-resonant linear situation can be obtained. Indeed, assume that there exist numbers $0 < \alpha \leq \beta$ such that, for all $u \neq v \in \mathbb{R}$ one has

$$\alpha \leq \frac{g(u) - g(v)}{u - v} \leq \beta , \quad [\alpha, \beta] \cap \Sigma = \emptyset \quad (6)$$

which means that there exists $j^2 \in \Sigma$ such that

$$j^2 < \alpha \leq \beta < (j+1)^2 . \quad (7)$$

If $\gamma := (\alpha + \beta)/2$, so that $\gamma \notin \Sigma$, we can write (5) in the equivalent form

$$-u'' - \gamma u = g(u) - \gamma u + h(t) , \quad (8)$$

and, if we define $F_\gamma\colon L^2(0, 2\pi) \to L^2(0, 2\pi)$ by $[F_\gamma(u)](t) = g(u) - \gamma u(t) + h(t)$, it follows from (6) that

$$\|F_\gamma(u) - F_\gamma(v)\|_2 \leq \delta\|u - v\|_2 , \quad (9)$$

with $\delta = (\beta - \alpha)/2$. Furthermore, finding the 2π-periodic solutions of (8) is equivalent to solving the equation in $L^2(0, 2\pi)$

$$u = L_\gamma^{-1}F_\gamma(u) . \quad (10)$$

Now, using estimates (3), (7) and (9), we get, for all $u, v \in L^2(0, 2\pi)$,

$$\|L_\gamma^{-1}[F_\gamma(u) - F_\gamma(v)]\|_2 \leq \frac{\delta}{\mathrm{dist}(\gamma, \Sigma)}\|u - v\|_2 ,$$

with $\delta/\mathrm{dist}(\gamma, \Sigma) < 1$. Banach fixed point theorem implies the existence of a unique fixed point u of $L_\gamma^{-1}F_\gamma$, and hence of a unique 2π-periodic solution of (5).

Such a result, proved in 1949 by Dolph [14] for Dirichlet boundary conditions and in 1976 in [31] for the periodic problem, has been extended in various directions. For example, it was proved in [8], using sophisticated topological methods and delicate a priori estimates, that if $G(u) = \int_0^u g(s)ds$, the forced nonlinear oscillator (5) has at least one 2π-periodic solution for each continuous $h\colon [0, 2\pi] \to \mathbb{R}$ if g is odd, $ug(u) \geq \theta G(u) > 0$ for some $\theta \geq 1$ and all large $|u|$, and if

$$\left[\liminf_{u\to+\infty} \frac{2G(u)}{u^2}, \limsup_{u\to+\infty} \frac{2G(u)}{u^2}\right] \neq \{k^2\}$$

for any positive integer k. However, it is still an open problem to know if the natural generalization of (6)

$$j^2 < \liminf_{|u|\to+\infty} \frac{2G(u)}{u^2} \leq \limsup_{|u|\to+\infty} \frac{2G(u)}{u^2} < (j+1)^2$$

is sufficient for the existence of a 2π-periodic solution to (5). The problem of obtaining nonlinear versions of the resonance situation is more difficult and also requires more sophisticated tools. It will be considered in Sects. "Lower and Upper Solutions" to "Critical Point Theory"

Poincaré Operator and Linear Systems

Let $T > 0$ be fixed, $f\colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$, $(t, x) \mapsto f(t, x)$ be locally Lipschitzian in x and continuous. For each $y \in \mathbb{R}^n$, there exists a unique solution $x(t) = p(t; y)$ of the Cauchy problem

$$x' = f(t, x) , \quad x(0) = y ,$$

defined and continuous on a maximal open set $G = \{(t, y) \in \mathbb{R} \times \mathbb{R}^n : \tau_-(y) < t < \tau_+(y)\}$, for some $-\infty \leq \tau_-(y) < 0 < \tau_+(y) \leq +\infty$.

A T-periodic solution of the differential system

$$x' = f(t, x) \tag{11}$$

is a solution of (11) defined at least over $[0, T]$ and such that $x(0) = x(T)$. If we assume in addition that $f(t, x) = f(t + T, x)$ for all $t \in \mathbb{R}$ and $x \in \mathbb{R}^n$, a T-periodic solution of (11) can be continued as a solution defined over $\mathbb{R}$ and such that $x(t) = x(t + T)$ for all $t \in \mathbb{R}$. Poincaré already observed at the end of the XIXth century [43] that $p(t; y)$ is a T-periodic solution of (11) if and only if $y \in \mathbb{R}^n$ is such that $\tau_+(y) > T$ and $y = p(T; y)$, i. e. y is a fixed point of Poincaré operator P_T defined by $P_T(y) = p(T; y)$ for $y \in \mathbb{R}^n$ such that $\tau_+(y) > T$.

A simple application is the case of a linear system

$$x' = A(t)x \tag{12}$$

where $A\colon [0, T] \to \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ is a continuous $(n \times n)$-matrix-valued function. Given $y \in \mathbb{R}^n$, the solution of (12) such that $x(0) = y$ can be written $x(t) = X(t)y$ for some $n \times n$ matrix such that $X(0) = I$, the fundamental matrix of (12). The corresponding Poincaré operator, given by $P_T(y) = X(T)y$, has non-trivial fixed points if and only if $I - X(T)$ is singular. If $h\colon [0, T] \to \mathbb{R}^n$ is continuous, the solution of the forced linear system

$$x' = A(t)x + h(t) \tag{13}$$

such that $x(0) = y$ being given by

$$p(t; y) = X(t)y + \int_0^t X(t)X^{-1}(s)h(s)\mathrm{d}s\,, \tag{14}$$

the corresponding Poincaré operator is

$$P_T(y) = X(T)y + \int_0^T X(T)X^{-1}(s)h(s)\mathrm{d}s\,. \tag{15}$$

Consequently, if $I - X(T)$ is non singular, i. e. if (12) only has the trivial T-periodic solution $x(t) \equiv 0$, (15) has the unique fixed point

$$y = [I - X(T)]^{-1} \int_0^T X(T)X^{-1}(s)h(s)\mathrm{d}s$$

and, by inserting this value of y in (14), (13) has the unique T-periodic solution

$$x(t) = \int_0^T G(t, s)h(s)\mathrm{d}s\,, \tag{16}$$

where $G(t, s)$ is the Green matrix explicitly given by

$$G(t, s) = \begin{cases} X(t)[I - X(T)]^{-1}X^{-1}(s) & \text{if } 0 \leq s \leq t \leq T \\ X(t)[I - X(T)]^{-1}X(T)X^{-1}(s) & \text{if } 0 \leq t < s \leq T \end{cases} \tag{17}$$

The situation is more complicated when $I - X(T)$ is singular, which always happens in the simple case where $A(t) \equiv 0$, to which we restrict ourself. In this case, $X(t) \equiv I$, and $P_T(y) = y + \int_0^T h(s)\mathrm{d}s$. It has fixed points if and only if h has mean value zero, namely

$$\overline{h} := \frac{1}{T}\int_0^T h(s)\mathrm{d}s = 0\,,$$

in which case (13) with $A(t) \equiv 0$ has the family of T-periodic solutions

$$x(t) = y + \int_0^t h(s)\mathrm{d}s \quad (y \in \mathbb{R}^n)\,.$$

Boundedness and Periodicity

When $n = 1$ and $f\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is T-periodic with respect to t, the local uniqueness assumption implies that two solutions $p(t; y)$ and $p(t; z)$ with $y < z$ are such that $p(t; y) < p(t; z)$ for all t were they are defined. Hence, if $p(t; y)$ is defined for all $t \geq 0$, $p(t + nT; y)$ is the solution of (11) equal to $y_n = p(nT; y)$ at $t = 0$ for any integer $n \geq 1$. If $y_1 = y$, $p(t; y)$ is T-periodic; if, say, $y_1 < y$, then $p(t + T; y) = p(t; y_1) < p(t; y)$ for all $t \geq 0$, and hence, $p(t + (n + 1)T; y) < p(t + nT; y)$ for all $t \geq 0$. Thus, for any $t \geq 0$, the sequence $(p(t + nT; y))_{n\in\mathbb{N}}$ is monotone. If $p(t; y)$ is bounded in the future, i. e. if there exists $M > 0$ such that $|p(t; y)| \leq M$ for all $t \geq 0$, the sequence above, monotone, bounded and equicontinuous (as $(p'(t + nT; y))_{n\in\mathbb{N}}$ is bounded), converges uniformly on each bounded interval to a continuous function $\xi(t)$. It follows from the identity

$$p(t + nT; y) = p(nT; y) + \int_0^t f(s, p(s + nT; y))\mathrm{d}s$$

that $\xi(t)$ is a solution of (11) defined for $t \geq 0$ and that

$$\xi(T) = \lim_{n\to\infty} p((n + 1)T; y) = \lim_{n\to\infty} p(nT; y) = \xi(0)\,.$$

This gives a result proved by Massera [29] in 1950: if $n = 1$ and f is T-periodic in t, locally Lipschitzian in x, and continuous, then (11) admits a T-periodic solution if and only if it admits a solution bounded in the future. Of

course, the same statement holds if we replace 'in the future' by 'in the past'. The result needs not to be true for $n \geq 2$, but, using delicate arguments of two-dimensional topology, Massera [29] has proved that if $n = 2$, f is T-periodic in t, locally Lipschitzian in x, continuous, if all solutions of (11) exist in the future and if one of them is bounded in the future, then (11) admits a T-periodic solution. Again, the same statement holds if we replace 'in the future' by 'in the past'. An important consequence of Massera's results is that the absence of T-periodic solutions in a one or two-dimensional T-periodic system implies the unboundedness of all its solutions in the past and in the future.

By reinforcing Massera's conditions, one can find criteria for the existence of T-periodic solutions valid for any n. A *set* G is a positively invariant set for (11) if, for each $y \in G$, $p(t; y) \in G$ for all $t \in [0, \tau_+(y)]$. In particular, if G is bounded, then $\tau_+(y) = +\infty$. Now, if G is invariant, and homeomorphic to the unit closed ball $B[0; 1]$ in $\mathbb{R}^n$, Poincaré operator maps continuously G into itself, and Brouwer fixed point theorem implies that (11) has at least one T-periodic solution with values in G.

Such a result, which can be traced to Lefschetz [24] and to Levinson [25] in the early 1940s has been widely used to study the periodic solutions of the periodically forced Liénard differential equation

$$y'' + h(y)y' + g(y) = e(t), \tag{18}$$

written as a two-dimensional first order system, under various conditions upon the friction coefficient $h(y)$, the restoring force $g(y)$ and the forcing term $e(t)$. Some of Levinson's results have been at the origin of the theory of chaos.

Fixed Point Approach: Perturbation Theory

Formula (16) suggests another approach for finding periodic solutions, which is independent of Cauchy's problem and requires less regularity. Consider the nonlinear differential system

$$x' = A(t)x + f(t, x) \tag{19}$$

where $A\colon [0, T] \to \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ and $f\colon [0, T] \times \mathbb{R}^n \to \mathbb{R}^n$ are continuous, and the corresponding linear system $x' = A(t)x$ only has the trivial T-periodic solution. Using formula (16), finding the T-periodic solutions of (19) is equivalent to solving the nonlinear integral equation

$$x(t) = \int_0^T G(t, s) f(s, x(s))\mathrm{d}s := [\mathcal{H}(x)](t) \qquad (t \in [0, T]), \tag{20}$$

with $G(t, s)$ defined in (17), in the Banach space $C_T^\#$ of continuous functions such that $x(0) = x(T)$, i. e. to finding the fixed points of the nonlinear operator $\mathcal{H}\colon C_T^\# \to C_T^\#$ defined in (20). Consider now the family of problems

$$x' = A(t)x + \varepsilon f(t, x), \quad x(0) = x(T) \quad (\varepsilon \in \mathbb{R}), \tag{21}$$

where $x' = A(t)x$ only has the trivial T-periodic solution. Solving (21) is equivalent to solving the equation in $C_T^\#$

$$\mathcal{K}(x, \varepsilon) := x - \varepsilon \mathcal{H}(x) = 0 .$$

Trivially, $\mathcal{K}(x, 0) = 0$ if and only if $x = 0$. If $f'_x(t, x)$ exists and is continuous, the Fréchet derivative $\mathcal{K}'_x(x, \varepsilon)$ at $x \in C_T^\#$ is given by $\mathcal{K}'_x(x, \varepsilon) = I - \varepsilon \mathcal{H}'(x)$, so that $\mathcal{K}'_x(0, 0) = I$ is invertible. It follows from the implicit function theorem in Banach spaces that for some $\varepsilon_0 > 0$ and each $|\varepsilon| \leq \varepsilon_0$, (21) has a unique solution x_ε such that $x_\varepsilon \to 0$ as $\varepsilon \to 0$. Such a result can be traced to Poincaré [43], who proved it using the associated Cauchy problem. From the proof of the implicit function theorem or from Banach fixed point theorem, it follows that x_ε can be obtained as the limit of the sequence of successive approximations given by

$$x_\varepsilon^0 = 0, \quad x_\varepsilon^{k+1}(t) = \varepsilon \int_0^T G(t, s) f(s, x_\varepsilon^k(s))\mathrm{d}s \qquad (k = 0, 1, 2, \ldots).$$

The situation is more complicated when $x' = A(t)x$ has nontrivial T-periodic solutions, and again we restrict ourself to the important special case where $A(t) \equiv 0$, i. e. to the family of problems

$$x' = \varepsilon f(t, x), \quad x(0) = x(T) \quad (\varepsilon \in \mathbb{R}), \tag{22}$$

which admits, for $\varepsilon = 0$, the n-parameter family of T-periodic solutions $x(t) = c$ ($c \in \mathbb{R}^n$). To reduce (22) to a fixed point problem, one can first notice that, as easily verified, the linear problem

$$Lx := x' + x(0) = h(t), \quad x(0) = x(T)$$

has, for each continuous $h\colon [0, T] \to \mathbb{R}^n$, the unique solution

$$[L^{-1}h](t) = \overline{h} + \int_0^t [h(s) - \overline{h}]\mathrm{d}s . \tag{23}$$

In particular, $L^{-1}\overline{h} = \overline{h}$. Now, for $\varepsilon \neq 0$, problem (22) is equivalent to

$$x' = (1 - \varepsilon)\overline{f(\cdot, x(\cdot))} + \varepsilon f(t, x), \quad x(0) = x(T), \tag{24}$$

which, using (23), is equivalent to the fixed point problem in $C_T^\#$

$$\begin{aligned} x(t) &= x(0) + \overline{f(\cdot, x(\cdot))} \\ &\quad + \varepsilon \int_0^t [f(s, x(s)) - \overline{f(\cdot, x(\cdot))}]\mathrm{d}s \\ &:= [\mathcal{M}(x, \varepsilon)](t) \end{aligned}$$

introduced by Mawhin in 1969 [30], i. e. to the equation in $C_T^\#$

$$\mathcal{N}(x, \varepsilon) := x - \mathcal{M}(x, \varepsilon) = 0 .$$

$\mathcal{N}(x, 0) = 0$ is equivalent to $x(t) = x(0) + \overline{f(\cdot, x(\cdot))}$, and hence to $x(t) \equiv c$, with $c \in \mathbb{R}^n$ such that

$$F(c) := \frac{1}{T} \int_0^T f(s, c)\mathrm{d}s = 0. \tag{25}$$

If c_0 is a solution of (25) satisfying condition

$$\operatorname{Jac} F(c_0) := \det \left[\frac{1}{T} \int_0^T f'_x(s, c_0)\mathrm{d}s \right] \neq 0, \tag{26}$$

the Fréchet derivative $\mathcal{N}'_x(c_0, 0)$ is invertible, and the implicit function theorem in Banach spaces implies the existence of $\varepsilon_0 > 0$ such that, for each $|\varepsilon| \leq \varepsilon_0$, equation $\mathcal{N}(x, \varepsilon) = 0$ has a unique solution x_ε such that $x_\varepsilon \to c_0$ as $\varepsilon \to 0$, and the same conclusion holds for (22).

This result is also a consequence of other perturbation methods for periodic solutions based upon Poincaré, averaging, Lyapunov–Schmidt or Cesari–Hale methods, and described in some of the books given in the Bibliography. The proof by iteration of the implicit function theorem implies that x_ε can be obtained as the limit of the sequence of successive approximations given by

$$\begin{aligned} x_\varepsilon^0 &= c_0 , \\ x_\varepsilon^{k+1} &= c + \varepsilon[\mathcal{I} - \mathcal{M}'_x(c_0, 0)]^{-1} \\ &\quad \cdot \Big[\mathcal{M}(x_\varepsilon^k, \varepsilon) - \mathcal{M}(c_0, 0) - \mathcal{M}'_x(c_0, 0)(x_\varepsilon^k - c)\Big] \\ &\qquad (k = 0, 1, 2, \ldots) . \end{aligned}$$

For example, the search of positive periodic solutions of Verhulst equation with seasonal variations in population dynamics

$$y' = \varepsilon[a(t)y - b(t)y^2] \quad (\varepsilon > 0) , \tag{27}$$

where $a, b\colon \mathbb{R} \to (0, +\infty)$ are continuous and have period T, is equivalent, through the transformation $y = \mathrm{e}^u$, to the problem

$$u' = \varepsilon[a(t) - b(t)\mathrm{e}^u] , \quad u(0) = u(T)$$

for which $F(c) = \overline{a} - \overline{b}\mathrm{e}^c$. This equation has the unique solution $c_0 = \log \overline{a}/\overline{b}$, and $F'(c_0) = -\overline{a} \neq 0$. Hence, for sufficiently small $\varepsilon > 0$, Verhulst equation with seasonal variations (27) has at least one positive T-periodic solution y_ε such that $y_\varepsilon \to \overline{a}/\overline{b}$ when $\varepsilon \to 0$.

Fixed Point Approach: Large Nonlinearities

The use of more sophisticated fixed point theorems provides existence results which are not of perturbation type. If f is continuous, $\mathcal{H}$ defined in (20) is continuous, and, using Arzelá-Ascoli theorem, $\mathcal{H}$ takes bounded sets into relatively compact sets, i. e. $\mathcal{H}$ is completely continuous on $C_T^\#$. Using Schauder fixed point theorem, $\mathcal{H}$ has at least one fixed point if it maps a closed ball of $C_T^\#$ into itself. It is the case in particular when f satisfies the growth condition

$$\|f(t, x)\| \leq \alpha\|x\| + \beta \quad ((t, x) \in [0, T] \times \mathbb{R}^n) ,$$

with $\beta \geq 0$ and $|\alpha| \int_0^T |G(t, s)|\mathrm{d}s < 1$ for all $t \in [0, T]$, and, in particular, when f is bounded on $[0, T] \times \mathbb{R}^n$.

More general existence conditions can be deduced from Leray–Schauder–Schaefer fixed point theorem, which, applied to $\mathcal{H}$, implies that (19) has at least one T-periodic solution if there exists $R > 0$ such that any possible solution x of the family of problems

$$x'(t) = A(t)x + \varepsilon f(t, x) , \quad x(0) = x(T) \quad (\varepsilon \in [0, 1])$$

is such that $\max_{t\in[0,T]} \|x(t)\| < R$. Special cases of this result, for particular equations, can be traced to Stoppelli [49], and the general form was first given by Reissig [46] and Villari [50].

For example, coming back to the equivalent form of the problem of positive periodic solutions of Verhulst equation with seasonal variations

$$u' = a(t) - b(t)\mathrm{e}^u , \quad u(0) = u(T) , \tag{28}$$

where $a, b\colon \mathbb{R} \to (0, +\infty)$ are continuous and have period T, we associate to (28) the family of problems

$$u' + u = \varepsilon[a(t) + b(t)\mathrm{e}^u + u], \quad u(0) = u(T), \quad (\varepsilon \in [0, 1]) , \tag{29}$$

which reduces to (28) for $\varepsilon = 1$, and only has the trivial solution for $\varepsilon = 0$. If, for some $\varepsilon \in [0, 1]$, a possible solution u of (29) reaches a positive maximum at τ, then

$$0 = u'(\tau) = \varepsilon[a(\tau) - b(\tau)\mathrm{e}^{u(\tau)}] - (1 - \varepsilon)u(\tau) ,$$

so that $\varepsilon[a(\tau) - b(\tau)\mathrm{e}^{u(\tau)}] = (1 - \varepsilon)u(\tau) \geq 0$ and

$$u(\tau) \leq \log \frac{a(\tau)}{b(\tau)} \leq \log \frac{\max_{[0,T]} a}{\min_{[0,T]} b} .$$

Similarly, if u reaches a negative minimum at τ', then

$$u(\tau') \geq \log \frac{a(\tau')}{b(\tau')} \geq \log \frac{\min_{[0,T]} a}{\max_{[0,T]} b} .$$

The existence of at least one T-periodic solution u for (27) follows from the previous theorem with

$$R > \max\left(\log \frac{\max_{[0,T]} a}{\min_{[0,T]} b}, -\log \frac{\min_{[0,T]} a}{\max_{[0,T]} b}\right) ,$$

and it gives the positive solution $y(t) = \mathrm{e}^{u(t)}$ for the original Verhulst equation with seasonal variations $y' = a(t)y - b(t)y^2$, which reduces to the positive equilibrium $y(t) \equiv \frac{a}{b}$ in the non-seasonal case where a and b are constant.

In the case of system (22), Schauder or Leray–Schauder–Schaefer fixed point theorems may be difficult to apply to $\mathcal{M}(x,\varepsilon)$ because of the presence of the $x(0)$ term. A generalization is required, based upon the concept of topological degree of some mappings $\mathcal{F}$ defined on the closure $\overline{\Omega}$ of a bounded open set Ω of a Banach space X, and such that $\mathcal{F}(x) \neq 0$ for $x \in \partial\Omega$. This topological degree, an integer counting 'algebraically' the number of zeros of $\mathcal{F}$ in Ω, is equal, for $\mathcal{F} = \mathcal{I}$ to 1 or 0 according to $0 \in \Omega$ or $0 \notin \overline{\Omega}$, implies the existence of a zero of $\mathcal{F}$ in Ω when it is nonzero, and the topological degree of $\mathcal{F}(\cdot,\varepsilon)$ is independent of ε when $\mathcal{F}(x,\varepsilon) \neq 0$ for all $(x,\varepsilon) \in \partial\Omega \times [0,1]$.

Applied to the family of mappings $\mathcal{I} - \mathcal{M}(x,\varepsilon)$ introduced in (22), Leray–Schauder and Brouwer degree theory implies the following continuation theorem, proved by Mawhin in 1969 [30] : if one can find an open bounded set $\Omega \subset C_T^{\#}$ such that

(i) for each $\varepsilon \in (0,1]$, problem (22) has no solution on $\partial\Omega$,
(ii) system $F(c) = 0$ defined in (25) has no solution on $\partial\Omega \cap \mathbb{R}^n$, with $\mathbb{R}^n$ identified with constant functions in $C_T^{\#}$,
(iii) the Brouwer degree $d_B[F, \Omega \cap \mathbb{R}^n]$ is different from zero,

then system (11) has at least one T-periodic solution in Ω. If condition (i) is dropped, the conclusion remains valid for (22) with ε sufficiently small. Notice that in the conditions of the perturbation result described above, condition (ii) holds for Ω a sufficient small open neighborhood of the zero c_0 of F, and condition (26) implies that $d_B[F, \Omega \cap \mathbb{R}^n] = \text{sign Jac } F(c_0) = \pm 1$.

For example, consider the complex-valued Riccati-type equation

$$z' = p(t) + q(t)z + r(t)\widehat{z} + \widehat{z}^2 \tag{30}$$

where $\widehat{z}$ denotes the complex conjugate of z, and $p, q, r\colon [0,T] \to \mathbb{C}$ are continuous. It is nothing but a concise writing for a system of two real differential equations. If $\varepsilon \in (0,1]$ and $z(t)$ is a possible T-periodic solution of

$$z' = \varepsilon[p(t) + q(t)z + r(t)\widehat{z} + \widehat{z}^2] , \tag{31}$$

then, multiplying each member of (31) by z^2 and integrating over $[0,T]$, we get

$$\begin{aligned} 0 &= \frac{1}{T}\int_0^T \left(\frac{z^3}{3}\right)' \mathrm{d}t = \frac{1}{T}\int_0^T z^2(t)z'(t)\mathrm{d}t \\ &= \varepsilon\left\{\frac{1}{T}\int_0^T \big[p(t)z^2(t) + q(t)z^3(t) \right. \\ &\qquad \left. + r(t)\widehat{z}(t)z^2(t) + |z(t)|^4\big]\,\mathrm{d}t\right\} . \end{aligned}$$

Hence letting, for $p \geq 1$,

$$\|u\|_p = \left[\frac{1}{T}\int_0^T |u(t)|^p \mathrm{d}t\right]^{\frac{1}{p}} ,$$
$$\|u\|_\infty = \max_{t\in[0,T]} |u(t)| ,$$

and using Hölder inequality, we obtain

$$\|z\|_4^4 \leq \|p\|_2\|z\|_4^2 + [\|q\|_4 + \|r\|_4]\|z\|_4^3$$

so that $\|z\|_4 \leq R_1$, where R_1 is the positive root of equation $r^2 = [\|q\|_4 + \|r\|_4]r + \|p\|_2$. From (31) follows then that

$$\|z'\|_1 \leq \|p\|_1 + [\|q\|_{4/3} + \|r\|_{4/3}]R_1 + R_1^2 = R_2 .$$

Inequalities upon $\|z\|_4$ and $\|z'\|_1$ imply that $\|z\|_\infty < R$ for some $R = R(R_1, R_2)$. Now, for $c \in \mathbb{C}$,

$$F(c) = \overline{p} + \overline{q}c + \overline{r}\widehat{c} + \widehat{c}^2$$

and hence, if $F(c) = 0$, we have

$$|c|^2 \leq (\|q\|_1 + \|r\|_1)|c| + \|p\|_1 \leq (\|q\|_4 + \|r\|_4)|c| + \|p\|_2$$

so that $|c| \leq R_1 < R$. Finally, $d_B[F, B(0,R)] = -2$. Hence all conditions of the second existence theorem hold for $\Omega = B(0,R)$, and (30) has at least one T-periodic solution, a result first proved, using a different approach, by Srzednicki [48] in 1994, the present proof being given in [35]. Notice that this result could not have been deduced from Leray–Schauder–Schaefer theorem, whose assumptions imply that the associated Leray–Schauder degree has absolue value one, although, for (30), its has absolue value two. The perturbed problem

$$z' = \varepsilon[p(t) + q(t)z + r(t)\overline{z} + z^2] , \quad z(0) = z(T)$$

satisfies conditions (ii) and (iii) of Mawhin's continuation theorem for $\Omega = B(0;R)$ with R sufficiently large, and has at least one T-periodic solution for $|\varepsilon|$ sufficiently small. But Lloyd [28] and Campos–Ortega [6] have shown by examples that the problem

$$z' = p(t) + q(t)z + r(t)\overline{z} + z^2 , \quad z(0) = z(T) ,$$

in contrast to (30), may have no T-periodic solution for some choice of p, q, r.

Finally, notice that the examples developed here illustrate two fundamental approaches to obtain a priori bounds for the possible T-periodic solutions : maximum principle-type argument and integral estimates.

Guiding Functions

A useful variant of Mawhin's continuation theorem given in Sect. "Fixed Point Approach: Large Nonlinearities" has been proved in 1992 by Capietto, Mawhin and Zanolin [6], with another proof in [5] based upon degree theory for S^1-equivariant mappings. It states, for $g\colon \mathbb{R}^n \to \mathbb{R}^n$ and $f\colon [0,T] \times \mathbb{R}^n \to \mathbb{R}^n$ continuous, that if there exists an open bounded set $\Omega \subset C_T^\#$ such that the family of problems

$$x' = (1-\varepsilon)g(x) + \varepsilon f(t,x), \quad x(0) = x(T) \quad (\varepsilon \in [0,1])$$

has no solution on $\partial\Omega$, and if $d_B[g, \Omega \cap \mathbb{R}^n] \neq 0$, then the problem

$$x' = f(t,x) , \quad x(0) = x(T) \tag{32}$$

has at least one solution in Ω. The proof is based upon the fact that $|d_{LS}[\mathcal{I} - \mathcal{M}, \Omega]| = |d_B[g, \Omega \cap \mathbb{R}^n]|$, with $\mathcal{M}$ the fixed point operator associated to the autonomous system $x' = g(x)$. Now a guiding function on $G \subset \mathbb{R}^n$ for (32) is a function $V\colon \mathbb{R}^n \to \mathbb{R}$ of class C^1 such that, with $(u|v)$ the inner product of u and v in $\mathbb{R}^n$,

$$\|V'(x)\| \neq 0 \quad \text{and} \quad (V'(x)|f(t,x)) \leq 0 \quad \text{when} \quad (t,x) \in [0,1] \times G .$$

This is an extension due to Mawhin–Ward [37] of a slight generalization given in [32] of a concept introduced in 1958 by Krasnosel'skii and Perov (see [22]) in the case where $G = \mathbb{R}^n \setminus B(0;\rho)$ for some $\rho > 0$. For example $V(x) = (1/2)\|x\|^2$ is a guiding function on $\partial B(0;r)$ if $(x|f(t,x)) \leq 0$ for $(t,x) \in [0,T] \times \partial B(0,r)$. We set $V^r = \{x \in \mathbb{R}^n\colon V(x) < r\}$. It is proved in [37] that if there exists a C^1 function $V\colon \mathbb{R}^n \to \mathbb{R}$ such that V^0 is non-empty and bounded, V is a guiding function on $V^{-1}(\{0\})$ for (32), and $d_B[V', V^0] \neq 0$, then problem (32) has at least one solution with values in $\overline{V_0}$. The proof can be based upon the continuation theorem mentioned above with $g(x) = -V'(x)$ and $\Omega = \{x \in C_T^\#\colon x(t) \in G \ (t \in [0,T])\}$, and its essential ingredient consists in showing that, for $\lambda \in [0,1)$, the family of problems

$$x' = -(1-\varepsilon)V'(x) + \varepsilon f(t,x) , \quad x(0) = x(T) ,$$

has no solution on $\partial\Omega$, which is done by studying the maximum of $V(x(t))$. In particular, if one takes $G = \mathbb{R}^n \setminus B(0;\rho)$ for some $\rho > 0$, and V coercive, i. e. $\lim_{\|x\|} V(x) = +\infty$, one can show that $d_B[V', V^0] = 1$, and the existence follows. This generalizes the result about positively invariant sets mentioned in Sect. "Boundedness and Periodicity".

The following generalization of the concept of guiding function is also given in [37]. An averaged guiding function on $G \subset \mathbb{R}^n$ for (32) is a function $V\colon G \to \mathbb{R}$ of class C^1 such that

$$\|V'(x)\| \neq 0 \quad \text{and} \quad \int_0^T (V'(x(t))|f(t,x(t))\mathrm{d}t) \leq 0$$

when $x \in C_T^\#$ and $x(t) \in G$ for all $t \in [0,T]$. It is proved in [37] that problem (32) has at least a solution if there exists a C^1 function $V\colon \mathbb{R}^n \to \mathbb{R}$ such that V^0 is non-empty, V^r is bounded for

$$r = \max\left\{ T \max_{u \in V^{-1}(\{0\})} \|V'(u)\|^2 , \max_{u \in V^{-1}(\{0\})} \int_0^T |(V'(u)|f(t,u))|\mathrm{d}t \right\} ,$$

V is an averaged guiding function on $\mathbb{R}^n \setminus V^0$ for (32) and $d_B[V', V^0] \neq 0$.

Of course, any guiding function is an averaged guiding function, but the converse is false. On the other hand, any V such that $(V'(x)|f(t,x)) \leq \alpha(t)$ for some nonnegative $\alpha \in L^1(0,T)$, all $(t,x) \in [0,T] \times \mathbb{R}^n$, and such that

$$\int_0^T \limsup_{\|x\| \to \infty} (V'(x)|f(t,x)))\mathrm{d}t < 0$$

is an averaged guiding function. In particular, taking $V(x) = \|x\| - \log(1 + \|x\|)$, so that $V'(x) = \frac{x}{1+\|x\|}$, it is easy to see that (32) has at least one solution if $\left(\frac{x}{\|x\|}|f(t,x)\right) \leq \alpha(t)$ and

$$\int_0^T \limsup_{\|x\| \to \infty} \left(\frac{x}{\|x\|}|f(t,x)\right) \mathrm{d}t < 0 .$$

Lower and Upper Solutions

A powerful way of proving the existence of T-periodic solutions for first or second order scalar differential equations is the *method of lower and upper solutions (or sub- and supersolutions)*, introduced by Scorza Dragoni [47] in 1931 for Dirichlet boundary conditions and by Knobloch [21] in 1967 for periodic solutions. We describe it here, following the approach initiated in [33], for problem

$$u'' = f(t,u)\,, \quad u(0) = u(T)\,, \quad u'(0) = u'(T)\,, \tag{33}$$

where $f\colon [0,T]\times\mathbb{R}\to\mathbb{R}$ is continuous. The function α (resp. β) $\in C^2([0,T])$ is a *lower* (resp. *upper*) solution for (33) if

$$\begin{aligned} &\alpha''(t) \geq f(t,\alpha(t)) \quad (t\in]0,T[)\,,\\ &\qquad \alpha(0)=\alpha(T)\,, \quad \alpha'(0)\geq\alpha'(T)\\ &(\text{resp. } \beta''(t) \leq f(t,\beta(x)) \quad (t\in]0,T[)\,,\\ &\qquad \beta(0)=\beta(T)\,, \quad \beta'(0)\leq\beta'(T))\,. \end{aligned}$$

The method of lower and upper solutions reduces the existence of a T-periodic solution to that of an ordered couple of lower and upper solutions: if there exists a lower solution α and an upper solution β for (33), such that $\alpha(t)\leq\beta(t)$ for all $t\in[0,T]$, then (33) has as least one solution u such that $\alpha(t)\leq u(t)\leq\beta(t)$ for all $t\in[0,T]$.

To prove such a result, one first introduce a modified problem whose solutions are solutions of (33), namely

$$\begin{aligned} u''-u = f(t,\gamma(t,u))-\gamma(t,u)\,,\\ u(0)=u(T)\,, \quad u'(0)=u'(T)\,, \end{aligned} \tag{34}$$

where $\gamma(t,u)=\alpha(t)$, u or $\beta(t)$ according to $u<\alpha(t)$, $\alpha(t)\leq u\leq\beta(t)$ or $u>\beta(t)$. Notice that (33) and (34) coincide when $\alpha(t)\leq u\leq\beta(t)$, that its linear part only has the trivial T-periodic solution, and that its right-hand member is continuous and bounded everywhere. A result in Sect. "Fixed Point Approach: Large Nonlinearities" implies that (34) has at least one solution, say $\widetilde{u}$. A contradiction argument, based upon simple characterizations of a maximum or a minimum, implies that $\alpha(t)\leq\widetilde{u}(t)\leq\beta(t)$ for all $t\in[0,T]$, so that $\widetilde{u}$ is a solution of (33).

A simple consequence is an elegant necessary and sufficient existence condition first proved by Kazdan and Warner [20] in 1975 in the Dirichlet case : if $f(t,c)$ is nondecreasing in c for each fixed t, problem (33) has at least one solution if and only if $\frac{1}{T}\int_0^T f(t,c)\mathrm{d}t=0$ for some $c\in\mathbb{R}$. Indeed, if (33) has a solution u, with $u_m:=\min_{[0,T]}u$, $u_M:=\max_{[0,T]}u$, integrating both members of (33) over $[0,T]$ and using the monotonicity of $f(t,\cdot)$ gives

$$\int_0^T f(t,u_m)\mathrm{d}t \leq 0 \leq \int_0^T f(t,u_M)\mathrm{d}t\,,$$

and the necessity follows. For the sufficiency, if $\int_0^T f(t,c)\mathrm{d}t=0$, and $v(t)$ is the unique solution of the linear problem

$$v''=f(t,c)\,, \quad v(0)=0=v(T)\,, \quad v'(0)=v'(T)\,,$$

then $\alpha(t):=c-\max_{[0,T]}v+v(t)\leq c\leq\beta(t):=c-\min_{[0,T]}v+v(t)$ are ordered lower and upper solutions for (33).

In particular, the problem

$$u''+g(u)=h(t)\,, \quad u(0)=u(T)\,, \quad u'(0)=u'(T)$$

with $g\colon\mathbb{R}\to\mathbb{R}$ continuous non increasing and $h\colon[0,T]\to\mathbb{R}$ continuous, has at least one solution if and only if $\overline{h}\in g(\mathbb{R})$. This is a first nonlinear generalization of resonance at the first eigenvalue zero. For example, the problem

$$u''-\arctan u=h(t)\,, \quad u(0)=u(T)\,, \quad u'(0)=u'(T)$$

has at least one solution if and only if $-\frac{\pi}{2}<\overline{h}<\frac{\pi}{2}$, and the problem

$$u''-u^+=h(t)\,, \quad u(0)=u(T)\,, \quad u'(0)=u'(T)$$

has at least one solution if and only if $\overline{h}\leq 0$. This last problem is an example of differential equation with asymmetric or jumping nonlinearities. Since the pioneering work of Fučik [18] in 1978, the study of problems of the type

$$\begin{aligned} u''+\alpha u^+-\beta u^-+g(u)=h(t)\,,\\ u(0)=u(T)\,, \quad u'(0)=u'(T) \end{aligned}$$

where the linear part in the nonlinear oscillator is replaced by a piecewise linear one, has been intensively studied. See for example [17] for references.

Combined with degree arguments, the method of lower and upper solutions also allows to obtain multiplicity results for T-periodic solutions. For example, Fabry, Mawhin and Nkashama [16] have proved in 1986 for the problem (with $c\in\mathbb{R}$, $s\in\mathbb{R}$)

$$\begin{aligned} u''+cu'+g(u)=h(t)+s\,,\\ u(0)=u(T)\,, \quad u'(0)=u'(T) \end{aligned} \tag{35}$$

with $\lim_{|u|\to\infty}g(u)=+\infty$, the existence of $s_0\in\mathbb{R}$ such that problem (35) has no solution for $s<s_0$, at least

one solution for $s = s_0$ and at least two solutions for $s > s_0$. Such a result is generally refered as an Ambrosetti–Prodi problem, after the pioneering work of Ambrosetti and Prodi [4], in 1969, for elliptic boundary value problems. Ortega has studied the stability of the T-periodic solutions when $c > 0$ (see [40]).

The example of

$$u'' + u = \sin t\,, \quad u(0) = u(2\pi)\,, \quad u'(0) = u'(2\pi)$$

which has no solution (resonance at the second eigenvalue 1) and has the (unordered) lower and upper solution $\alpha \equiv 1$ and $\beta \equiv -1$ shows that the method of lower and upper solutions fails in general in the case of unordered lower and upper solutions. However, Amann, Ambrosetti and Mancini [2] have proved in 1978, for some elliptic boundary value problems, that existence may still hold under further conditions, using the following approach. For each $\widetilde{h} \in \widetilde{C}_T^{\#} = \{h \in C_T^{\#} : \overline{h} = 0\}$, the linear problem

$$\widetilde{L}u := u'' = \widetilde{h}(t)\,, \quad u(0) = 0 = u(T)\,, \quad u'(0) = u'(T)$$

has the unique solution

$$u(t) = [\widetilde{L}^{-1}h](t) = \int_0^T K(t,s)\widetilde{h}(s)\mathrm{d}s\,,$$

where $K(t,s) = -s/T(T-t)$ if $0 \le s \le t \le T$ and $K(t,s) = -t/T(T-s)$ if $0 \le t < s \le T$. Hence u is a T-periodic solution of (33) if and only if $u \in C_T^{\#}$ is a solution of

$$u(t) - u(0) = \int_0^T K(t,s)[f(s,u(s)) - \overline{f(\cdot,u(\cdot))}]\mathrm{d}s \quad (36)$$

$$\overline{f(\cdot,u(\cdot))} = 0\,. \quad (37)$$

Letting $c = u(0)$, $y(t) = u(t) - u(0)$, so that $y \in \widehat{C}_T^{\#} = \{x \in C_T^{\#} : x(0) = 0\}$, the first equation in (36) becomes

$$\begin{aligned} y(t) &= \int_0^T K(t,s)[f(s,c+y(s)) - \overline{f(\cdot,c+y(\cdot))}]\mathrm{d}s \\ &:= [\mathcal{R}(y,c)](t)\,, \end{aligned} \quad (38)$$

with $\mathcal{R}$ is completely continuous in $\widehat{C}_T^{\#} \times \mathbb{R}$. Assuming now in addition that $|f|$ is bounded on $[0,T] \times \mathbb{R}$, say by M, Leray–Schauder degree theory applied to (38) as a fixed point problem in y with c as a parameter implies that the set $(y,c) \in \widehat{C}_T^{\#} \times \mathbb{R}$ satisfying (38) contains a continuum C whose projection on $\widehat{C}_T^{\#}$ is contained in a ball $B[0;R]$ with R depending only upon T and M, and whose projection on $\mathbb{R}$ is $\mathbb{R}$. On C, we have, differentiating (38),

$$\begin{aligned} y'' &= f(t,c+y) - \overline{f(\cdot,c+y(\cdot))}\,, \\ &\qquad y(0) = y(T)\,, \quad y'(0) = y'(T)\,. \end{aligned} \quad (39)$$

Hence, if there exists $(c,y) \in C$ such that $\overline{f(\cdot,c+y(\cdot))} = 0$, the second equation in (36) is also satisfied and $c + y$ is a solution of (33). If not, by connexity, either $\overline{f(\cdot,c+y(\cdot))} < 0$ or $\overline{f(\cdot,c+y(\cdot))} > 0$ for all $(c,y) \in C$. Assume now that (33) has a lower solution α and an upper solution β which are not ordered, and consider, say the case where $\overline{f(\cdot,c+y(\cdot))} < 0$ on C, the other one being similar. For each $c \in \mathbb{R}$, it follows from (39) that $c + y$ is a lower solution for (33) whenever $(c,y) \in C$. Taking c such that $c + y(t) \le \beta(t)$ for all $t \in [0,T]$ gives a couple of ordered lower and upper solution, and hence a T-periodic solution of (33). Proceeding like in the ordered case, we deduce from this result that if f is bounded and $f(t,\cdot)$ non-increasing for each fixed $t \in [0,T]$, then (33) has at least one solution if and only if $1/T\int_0^T f(t,c)\mathrm{d}t = 0$ for some $c \in \mathbb{R}$. In particular, the problem

$$u'' + g(u) = h(t)\,, \quad u(0) = u(T)\,, \quad u'(0) = u'(T)$$

with $g\colon \mathbb{R} \to \mathbb{R}$ continuous, bounded and non decreasing, and $h\colon [0,T] \to \mathbb{R}$ continuous, has at least one solution if and only if $\overline{h} \in g(\mathbb{R})$. This is another nonlinear generalization of resonance at the first eigenvalue 0. For example, the problem

$$u'' + \arctan u = h(t)\,, \quad u(0) = u(T)\,, \quad u'(0) = u'(T)$$

as at least one solution if and only if $-\frac{\pi}{2} < \overline{h} < \frac{\pi}{2}$, and the problem

$$u'' + \frac{u^+}{1+|u|} = h(t)\,, \quad u(0) = u(T)\,, \quad u'(0) = u'(T)$$

has at least one solution if and only if $0 \le \overline{h} < 1$. Notice that the boundedness condition upon f can be replaced by suitable linear growth conditions.

The case of constant lower and upper solutions gives a simple but useful existence condition: if $f(t,\alpha) \le 0 \le f(t,\beta)$ for some $\alpha \le \beta$ and all $t \in [0,T]$, problem (33) has at least one solution u with $\alpha \le u(t) \le \beta$ for all $t \in [0,T]$. The same conclusion holds for f bounded and $f(t,\beta) \le 0 \le f(t,\alpha)$ for all $t \in [0,T]$. For example, taking $\beta = R = -\alpha$ for sufficiently large $R > 0$, the problem

$$u'' = p(u) + h(t)\,, \quad u(0) = u(T)\,, \quad u'(0) = u'(T)$$

has at least one solution for each continuous h, when p is a real polynomial of odd order whose highest order term has a positive coefficient. This result can be traced to Lichtenstein [26], who proved it in 1915 using a variational method, and can be applied to the original Duffing's equation (with $a > 0$)

$$u'' + a\left(u - \frac{u^3}{6}\right) = h(t)\,, \qquad u(0) = u(T)\,, \quad u'(0) = u'(T)$$

introduced in 1918 by Duffing [15] as a nonlinear approximation to the forced pendulum equation.

However, for the exact pendulum equation

$$u'' + a\sin u = h(t)\,, \quad u(0) = u(T)\,, \quad u'(0) = u'(T) \tag{40}$$

a necessary condition for the existence of a solution is $-a \le \overline{h} \le a$, as follows from integrating both members over $[0, T]$. To obtain a necessary and sufficient existence condition for (40), Dancer [12] has used in 1982 an approach similar to the one described for the case of unordered lower and upper solutions. Writing $h = \overline{h} + \widetilde{h}$, he has proved, for each $\widetilde{h}$, the existence of a (possibly degenerate) closed interval $\mathcal{I}_{\widetilde{h}} \subset [-a, a]$ such that (40) has at least one solution if and only if $\overline{h} \in \mathcal{I}_{\widetilde{h}}$. In 1984, using degree theory, Mawhin and Willem [38] have proved the existence of at least two solutions for (40) when $\overline{h} \in \operatorname{int} \mathcal{I}$. The same authors have proved that the set of $\widetilde{h}$ for which $\mathcal{I}_{\widetilde{h}}$ has a non-empty interior is open and dense in the space of continuous functions with mean value zero, but it is still an open problem to know if there exists or not some $\widetilde{h}$ such that $\mathcal{I}_{\widetilde{h}}$ is a singleton. See [36] for a survey of the various results related to the forced pendulum equation.

Direct Method of the Calculus of Variations

When a differential equation or system can be written as the Euler–Lagrange equations of a problem of the calculus of variations, the direct method can be used to prove the existence of periodic solutions. Its fundamental result is that if X is a reflexive Banach space and if a weakly lower semi-continuous (w.l.s.c.) function $\varphi\colon X \to (-\infty, +\infty]$ has a bounded minimizing sequence, then it has a minimum on X. Recall that φ is weakly lower semi-continuous at $a \in X$ if $\liminf_{k\to\infty} \varphi(u_k) \ge \varphi(u)$ whenever $u_k \rightharpoonup u$, and that $(u_k)_{k\in\mathbb{N}}$ is a minimizing sequence for φ if $\varphi(u_k) \to \inf_X \varphi$. Another result, valid for an arbitrary Banach space X, is that a function $\varphi\colon X \to \mathbb{R}$ of class C^1, bounded from below and satisfying the Palais–Smale condition has a minimum on X. The Palais–Smale condition was introduced by Palais and Smale [42] in 1964.

For simplicity, only the case of Lagrangian systems of differential equations of the type

$$u'' = F'_u(t, u) + \widetilde{h}(t)\,, \quad u(0) = u(T)\,, \quad u'(0) = u'(T) \tag{41}$$

will be considered, where $F\colon [0, T] \times \mathbb{R}^n \to \mathbb{R}$ is continuous together with $F'_u\colon [0, T] \times \mathbb{R}^n \to \mathbb{R}^n$, $\widetilde{h}\colon [0, T] \to \mathbb{R}^n$ is continuous with $\int_0^T \widetilde{h}(t)\mathrm{d}t = 0$. We denote by H^1_T the Sobolev space

$$\{u \in L^2([0, T], \mathbb{R}^n)\colon u \text{ has a weak derivative } u' \in L^2([0, T], \mathbb{R}^n),\ u(0) = u(T)\},$$

with the inner product and norm

$$\langle u, v\rangle = \int_0^T [(u(t), v(t)) + (u'(t), v'(t))]\mathrm{d}t\,, \qquad \|u\|_{1,1} = \langle u, u\rangle^{\frac{1}{2}}\,.$$

It is easy to show that the action functional associated to (41) given by

$$\varphi(u) := \int_0^T \left[\frac{\|u'(t)\|^2}{2} + F(t, u(t)) + (\widetilde{h}(t), u(t))\right]\mathrm{d}t \tag{42}$$

$$= \int_0^T \left[\frac{\|u'(t)\|^2}{2} + F(t, u(t)) + (\widetilde{h}(t), \widetilde{u}(t))\right]\mathrm{d}t \tag{43}$$

is well defined, w.l.s.c. and of class C^1 on H^1_T, with

$$\varphi'(u)[v] = \int_0^T \left[(u'(t), v'(t)) + (F'_u(t, u(t)) + \widetilde{h}(t), v(t))\right]\mathrm{d}t$$

for all $u, v \in H^1_T$. Furthermore, if u is a critical point of φ, i. e. if $\varphi'(u) = 0$, then $u \in C^2([0, T], \mathbb{R}^n)$ and satisfies (41).

As a first application, following Mawhin–Willem [39], assume that $\|F'_u\|$ is bounded, say by M, on $[0, T] \times \mathbb{R}^n$ and that F satisfied the condition

$$\lim_{\|u\|\to\infty} \frac{1}{T}\int_0^T F(t, u)\mathrm{d}t = +\infty\,, \tag{44}$$

first introduced by Ahmad–Lazer–Paul [1] in 1976 for elliptic boundary value problems. Writing $u = \overline{u} + \widetilde{u}$, and

using Wirtinger's inequality, we obtain

$$\begin{aligned}\varphi(u) &= \int_0^T \frac{\|u'(t)\|^2}{2}\mathrm{d}t + \int_0^T F(t,\overline{u})\mathrm{d}t \\ &\quad + \int_0^T\int_0^1 (F'_u(t,\overline{u}+s\widetilde{u}(t)) + \widetilde{h}(t),\widetilde{u}(t))\mathrm{d}t \\ &\geq \int_0^T \frac{\|u'(t)\|^2}{2}\mathrm{d}t - \frac{M'T^{\frac{3}{2}}}{2\pi}\left[\int_0^T \frac{\|u'(t)\|^2}{2}\mathrm{d}t\right]^{\frac{1}{2}} \\ &\quad + \int_0^T F(t,\overline{u})\mathrm{d}t\end{aligned}$$

with $M' = M + \|\widetilde{h}\|_\infty$, which implies that $\varphi(u) \to +\infty$ as $\|u\|_{1,1} \to \infty$. Hence all minimizing sequences for φ are bounded and the minimum of φ is reached at some $u \in H^1_T$, and is a solution of (41).

For example, consider the problem

$$u'' + g(u) = h(t)\,, \quad u(0) = u(T)\,, \quad u'(0) = u'(T)\,,$$

where $h\colon [0,T] \to \mathbb{R}$ is continuous, $g\colon \mathbb{R} \to \mathbb{R}$ is continuous, bounded and, with $G(u) = \int_0^u g(s)\mathrm{d}s$, is such that

$$\widehat{g}(\pm\infty) := \lim_{u\to\pm\infty} \frac{G(u)}{u} \tag{45}$$

exist. For this problem

$$\lim_{|u|\to\infty} \int_0^T F(t,u)\mathrm{d}t = \lim_{|u|\to\infty} u\left[\overline{h} - \frac{G(u)}{u}\right] = +\infty$$

if $\widehat{g}(+\infty) < \overline{h} < \widehat{g}(-\infty)$. Such a condition, first introduced by Alonso and Ortega [2] in 1996, generalizes the one introduced by Lazer and Leach [23] in 1969 (with $g(u)$ instead of $G(u)/u$ in (45)), but usually refered as a Landesman–Lazer condition. For example the problem

$$u'' - \arctan u + a\sin u = h(t)\,,$$
$$u(0) = u(T)\,, \quad u'(0) = u'(T)$$

with $a \in \mathbb{R}$ has at least one solution if $-\frac{\pi}{2} < \overline{h} < \frac{\pi}{2}$.

As a second application, due to Willem [51], consider the case of a spatially periodic F, i. e. such that

$$F(t, u + T_i e_i) = F(t,u) \quad (1 \leq i \leq n) \tag{46}$$

for some $T_i > 0$ and all $(t,u) \in [0,T] \times \mathbb{R}^n$, where the e_i denote the unit vectors in $\mathbb{R}^n (1 \leq i \leq n)$. As F is bounded over $[0,T] \times \mathbb{R}^n$, it is easy to see, using Sobolev inequality, that

$$\begin{aligned}\varphi(u) &\geq \frac{1}{2}\int_0^T \|u'(t)\|^2\mathrm{d}t \\ &\quad - C_1\left[\int_0^T \|u'(t)\|^2\mathrm{d}t\right]^{\frac{1}{2}} - C_2\end{aligned}$$

for some $C_1, C_2 \geq 0$, which implies the existence of $C_3 > 0$ such that any minimizing sequence $(u_k)_{k\in\mathbb{N}}$ for φ satisfies $\int_0^T \|u_k(t)\|^2\mathrm{d}t \leq C_3$. On the other hand, it follows from (46) that φ is such that

$$\varphi(u + T_i e_i) = \varphi(u) \quad (u \in H^1_T,\ 1 \leq i \leq n) \tag{47}$$

and hence $(u_k + v_k)_{k\in\mathbb{N}}$ with $v_{k,i} = k_i T_i$, $(k_i \in \mathbb{Z}, 1 \leq i \leq n)$ is also a minimizing sequence. So there is always a minimizing sequence u_k such that $0 \leq \overline{u}_{k,i} \leq T_i (1 \leq i \leq n)$, and hence a bounded minimizing sequence in H^1_T, implying the existence of a T-periodic solution of (41) when (46) holds. In particular, this result implies that the periodic problem for the forced pendulum equation (40) has at least one solution for each $a \in \mathbb{R}$ and each h with $\overline{h} = 0$, so that $0 \in \mathcal{I}_{\widetilde{h}}$. Such a result, already proved by Hamel [19] in 1922 using calculus of variations, was independently rediscovered by Willem [51] in 1981 and Dancer [12] in 1982. Notice that, for $n = 1$, Dancer and Ortega [13] have proved in 2004 that if the minimum is isolated as a critical point of the action functional, then it is unstable in the sense of Lyapunov.

In this case of a periodic potential, multiplicity results can be proved using more sophisticated tools of the calculus of variations. Property (47) shows that one can consider naturally φ on the manifold $\mathbb{T}^n \times \widetilde{H}^1_T$, where $\mathbb{T}^n$ denotes the n-torus and $\widetilde{H}^1_T$ the subspace of H^1_T of functions with mean value zero. In such a case, a result of Palais [41] implies that if φ is bounded from below and satisfies the Palais–Smale condition, then φ has at least $\mathrm{cat}(\mathbb{T}^n \times \widetilde{H}^1_T)$ critical points. In this statement, $\mathrm{cat}(M)$ denotes the Lusternik–Schnirelmann category of the set M in itself, i. e. the least integer k such that M can be covered by k contractible subsets in M, a concept introduced by Ljusternik and Schnirelmann in 1934 [27]. Now it can be shown that

$$\mathrm{cat}(\mathbb{T}^n \times \widetilde{H}^1_T) = \mathrm{cat}(\mathbb{T}^n) = n + 1\,,$$

and hence, under condition (46), system (41) has at least $n + 1$ geometrically distinct T-periodic solutions. In particular, when $\overline{h} = 0$, the forced pendulum equation (40) has at least two T-periodic solutions, a result first obtained in 1984 by Mawhin and Willem [38], using another variational approach. For other results in the spirit of Lusternik–Schnirelmann category, see [9,34,45]. Using, instead of Ljusternik–Schnirelmann category, another variational technique, namely Morse theory (see [10]), one can prove the existence of at least 2^n geometrically distinct T-periodic solutions when they are non-degenerate.

In the case of a spatially periodic Hamiltonian system

$$Ju' = H'_u(t,u)\,, \quad u(0) = u(T)\,,$$

where $H\colon [0,T]\times\mathbb{R}^{2n}\to\mathbb{R}$, $H'_u\colon [0,T]\times\mathbb{R}^{2n}\to\mathbb{R}^{2n}$ are continuous, and $J=\begin{pmatrix}0 & -I\\ I & 0\end{pmatrix}$ is the symplectic matrix, a similar result, namely the existence of at least $2n+1$ T-periodic solutions, has been proved by Conley and Zehnder [11] in 1984 using some finite-dimensional reduction and Conley's index. It solves a conjecture of Arnold in symplectic geometry (see e. g. [10]).

Critical Point Theory

A function φ unbounded from below and from above needs not to have a minimum or a maximum, but may have a critical point of saddle point type. Many existence theorems for critical points of saddle point type of some functionals defined on a Banach space or a Banach manifold have been developed in the last thirty years, and have found many applications in the study of periodic solutions of differential systems having a variational structure, like Lagrangian or Hamiltonian systems. One such result, both simple and useful, is the saddle point theorem proved in 1978 by Rabinowitz [44]. Let $\varphi\colon X\to\mathbb{R}$ be of class C^1 on a Banach space X which can be splitted as a direct sum of closed subspaces $X^-\oplus X^+$, with X^+ finite-dimensional. Assume that, for some $R>0$, $\sup_{\partial B[0,R]\cap X^-}\varphi<\inf_{X^+}\varphi$, and let

$$M:=\{g\colon B[0,R]\cap X^-\to X^-,\ \text{continuous},\quad g(s)=s \text{ on } \partial B[0,R]\cap X^-\}\,.$$

If φ satisfies a Palais–Smale condition on X, φ has at least one critical point u such that $\varphi(u)=c=\inf_{g\in M}\sup_{s\in B[0,R]\cap X^-}\varphi(g(s))$.

As a simple application, we return to problem (41) with F'_u bounded over $[0,T]\times\mathbb{R}^n$, and we now assume instead of (44) that

$$\lim_{\|u\|\to\infty}\int_0^T F(t,u)\mathrm{d}t=-\infty. \tag{48}$$

It is easy to see that φ is unbounded from below on the subspace of constant functions and bounded from below but unbounded from above on the subspace of functions with mean value zero. Hence, in particular, $\inf_{\widetilde{H}^1_T}\varphi>-\infty$, and, for all sufficiently large $R>0$, $\sup_{\partial B[0,R]\cap\overline{H}^1_T}\varphi<\inf_{\widetilde{H}^1_T}\varphi$, where $\overline{H}^1_T$ is the subspace of constant functions, so that $H^1_T=\overline{H}^1_T\oplus\widetilde{H}^1_T$. The verification of the Palais–Smale condition follows from getting first a bound on $\widetilde{u}_k$ using the boundedness of F'_u and then a bound for $\overline{u}_k$ using condition (48). Hence (41) has at least one T-periodic solution. In particular, the problem

$$u''+g(u)=h(t)\,,\quad u(0)=u(T)\,,\quad u'(0)=u'(T)\,,$$

where $h\colon[0,T]\to\mathbb{R}$ is continuous and $g\colon\mathbb{R}\to\mathbb{R}$ is continuous, bounded and such that (with $\widehat{g}(\pm\infty)$ defined in (45)) $\widehat{g}(-\infty)<\overline{h}<\widehat{g}(+\infty)$, has at least one solution. For example, problem

$$u''+\arctan u+a\sin u=h(t)\,,\qquad u(0)=u(T)\,,\quad u'(0)=u'(T)\,,$$

has a solution if $a\in\mathbb{R}$ and $-\frac{\pi}{2}<\overline{h}<\frac{\pi}{2}$. This is another example of nonlinear generalization of the resonance condition at zero eigenvalue.

Using topological degree, variational methods, or symplectic techniques, existence results have also been obtained in the case of resonance at a nonzero eigenvalue. For example, the problem

$$u''+k^2u+g(u)=h(t)\,,\qquad u(0)=u(2\pi)\,,\quad u'(0)=u'(2\pi)$$

with $g\colon\mathbb{R}\to\mathbb{R}$ continuous and bounded and k a positive integer, has at least one solution for all continuous $h\colon[0,2\pi]\to\mathbb{R}$ such that the real function

$$\Phi(\theta):=2\left[\widehat{g}(+\infty)-\widehat{g}(-\infty)\right]-\int_0^{2\pi}h(t)\sin k(t+\theta)\mathrm{d}t$$

has no zero or more than two zeros, all simple, in $[0,2\pi/k[$. See [17] for the proof and references to earlier contributions of Lazer–Leach, Dancer, and Fabry–Fonda.

Future Directions

In the whole XXth century, the study of periodic solutions of non-autonomous ordinary differential equations has been highly influential in the creation and development of fundamental parts of present mathematics, like functional analysis (operator theory, iterative methods,...), algebraic topology (fixed point theorems, topological degree, Conley index,...), variational methods (dual action principle, minimax theorems, Morse theory,...), symplectic techniques (Poincaré-Birkhoff-type fixed point theorems,...). One can expect that further topological tools will be useful or developed in searching new existence and multiplicity theorems for periodic solutions.

The difficult problem of discussing the stability of those periodic solutions, still in infancy, is fundamental for the applications and must be developed. Most earlier studies of special differential equations have been devoted to models coming from mechanics and electronics, which are far from being completely understood, but the recent applications to biology, demography and economy will introduce new classes of differential equations and systems

with periodic time dependence, requiring the use of new analytical and topological tools. In this respect, the study of periodic solutions of nonlinear nonautonomous difference equations is of increasing importance.

Even if periodic solutions are not the easiest objects to be found numerically with a large degree of certitude, numerical methods and computers will play an increasing role in detecting the presence of periodic solutions. Some efficient softwares have already been developed in this respect.

An enormous gap still exists between the methods of approach and the results about what seems to be the natural generalization of periodic solutions, namely the almost periodic solutions. It is a paradoxal fact that some statements are true both for periodic solutions and for solutions bounded on the whole real line, but false for the intermediate case of almost periodic solutions!

Finally, like the equilibria in autonomous systems, the study of periodic solutions of non-autonomous equations will remain the unavoidable first step in trying to understand the complexity of the set of all solutions, and much remains to be done in this direction. Poincaré's famous sentence

'What renders these periodic solutions so precious is that they are, so to speak, the only breach through which we may try to penetrate a stronghold previously reputed to be impregnable'

keeps its full significance in the beginning of the XXIth century.

Bibliography

Primary Literature

1. Ahmad S, Lazer AC, Paul JL (1976), Elementary critical point theory and perturbations of elliptic boundary value problems. Indiana Univ Math J 25:933–944
2. Alonso JM, Ortega R (1996) Unbounded solutions of semilinear equations at resonance. Nonlinearity 9:1099–1111
3. Amann H, Ambrosetti A, Mancini G (1978) Elliptic equations with noninvertible Fredholm linear part and bounded nonlinearities. Math Z 158:179–194
4. Ambrosetti A, Prodi G (1972) On the inversion of some differentiable mappings with singularities between Banach spaces. Ann Mat Pura Appl 93(4):231–246
5. Bartsch T, Mawhin J (1991) The Leray–Schauder degree of S^1-equivariant operators associated to autonomous neutral equations in spaces of periodic functions. J Differ Equations 92:90–99
6. Campos J, Ortega R (1996) Nonexistence of periodic solutions of a complex Riccati equation. Differ Integral Equations 9:247–249
7. Capietto A, Mawhin J, Zanolin F (1992) Continuation theorems for periodic perturbations of autonomous systems. Trans Amer Math Soc 329:41–72
8. Capietto A, Mawhin J, Zanolin F (1995) A continuation theorem for periodic boundary value problems with oscillatory nonlinearities. Nonlinear Differ Equations Appl 2:133–163
9. Chang KC (1989) On the periodic nonlinearity and the multiplicity of solutions. Nonlinear Anal 13:527–537
10. Chang KC (1993) Infinite Dimensional Morse Theory and Multiple Solution Problems. Birkhäuser, Basel
11. Conley C, Zehnder E (1984) Morse type index for flows and periodic solutions for Hamiltonian operator. Comm Pure Appl Math 37:207–253
12. Dancer EN (1982) On the use of asymptotics in nonlinear boundary value problems. Ann Mat Pura Appl 131(4):67–187
13. Dancer EN, Ortega R (1994) The index of Lyapunov stable fixed points in two dimensions. J Dyn Differ Equations 6:631–637
14. Dolph CL (1949) Nonlinear integral equations of Hammerstein type. Trans Amer Math Soc 66:289–307
15. Duffing G (1918) Erzwungene Schwingungen bei veränderlicher Eigenfrequenz. Vieweg, Braunschweig
16. Fabry C, Mawhin J, Nkashama M (1986) A multiplicity result for periodic solutions of forced nonlinear second order differential equations. Bull London Math Soc 18:173–180
17. Fabry C, Mawhin J (2000) Oscillations of a forced asymmetric oscillator at resonance. Nonlinearity 13:493–505
18. Fučik S (1976) Boundary value problems with jumping nonlinearities. Časopis Pest Mat 101:69–87
19. Hamel G (1922) Ueber erzwungene Schwingungen bei endlichen Amplituden. Math Ann 86:1–13
20. Kazdan JL, Warner FW (1975) Remarks on some quasilinear elliptic equations. Comm Pure Appl Math 28:567–597
21. Knobloch HW (1963) Eine neue Methode zur Approximation periodischer Lösungen von Differentialgleichungen zweiter Ordnung. Math Z 82:177–197
22. Krasnosel'skii MA (1968) The operator of translation along the trajectories of differential equations. Amer Math Soc, Providence
23. Lazer AC, Leach DE (1969) Bounded perturbations for forced harmonic oscillators at resonance. Ann Mat Pura Appl 82(4):49–68
24. Lefschetz S (1943) Existence of periodic solutions of certain differential equations. Proc Nat Acad Sci USA 29:29–32
25. Levinson N (1943) On the existence of periodic solutions for second order differential equations with a forcing term. J Math Phys 22:41–48
26. Lichtenstein L (1915) Ueber einige Existenzprobleme der Variationsrechnung. J Reine Angew Math 145:24–85
27. Ljusternik L, Schnirelmann L (1934) Méthodes topologiques dans les problèmes variationnels. Hermann, Paris
28. Lloyd NG (1975) On a class of differential equations of Riccati type. J London Math Soc 10(2):1–10
29. Massera JL (1950) The existence of periodic solutions of systems of differential equations. Duke Math J 17:457–475
30. Mawhin J (1969) Équations intégrales et solutions périodiques de systèmes différentiels non linéaires. Bull Cl Sci Acad Roy Belgique 55(5):934–947
31. Mawhin J (1976) Contractive mappings and periodically perturbed conservative systems. Arch Math (Brno) 12:67–73
32. Mawhin J (1979) Topological Degree and Nonlinear Boundary Value Problems. CBMS Conf Math No 40. Amer Math Soc, Providence

33. Mawhin J (1983) Points fixes, points critiques et problèmes aux limites. Sémin Math Supérieures No 92, Presses Univ de Montréal, Montréal
34. Mawhin J (1989) Forced second order conservative systems with periodic nonlinearity. Ann Inst Henri Poincaré Anal non linéaire 6:415–434
35. Mawhin J (1994) Periodic solutions of some planar nonautonomous polynomial differential equations. Differ Integral Equations 7:1055–1061
36. Mawhin J (2004) Global results for the forced pendulum equation. In: Cañada A, Drabek P, Fonda A (eds) Handbook of Differential Equations. Ordinary Differential Equations, vol 1. Elsevier, Amsterdam, pp 533–590
37. Mawhin J, Ward JR (2002) Guiding-like functions for periodic or bounded solutions of ordinary differential equations. Discret Continuous Dyn Syst 8:39–54
38. Mawhin J, Willem M (1984) Multiple solutions of the periodic boundary value problem for some forced pendulum-type equations. J Differ Equations 52:264–287
39. Mawhin J, Willem M (1989) Critical point theory and Hamiltonian systems. Springer, New York
40. Ortega R (1995) Some applications of the topological degree to stability theory. In: Granas A, Frigon M (eds) Topological methods in differential equations and inclusions, NATO ASI C472. Kluwer, Amsterdam, pp 377–409
41. Palais R (1966) Ljusternik–Schnirelmann theory on Banach manifolds. Topology 5:115–132
42. Palais R, Smale S (1964) A generalized Morse theory. Bull Amer Math Soc 70:165–171
43. Poincaré H (1892) Les méthodes nouvelles de la mécanique céleste, vol 1. Gauthier–Villars, Paris
44. Rabinowitz P (1978) Some minimax theorems and applications to nonlinear partial differential equations. In: Cesari L, Kannan R, Weinberger H (eds) Nonlinear Analysis, A tribute to E. Rothe. Academic Press, New York
45. Rabinowitz P (1988) On a class of functionals invariant under a $\mathbb{Z}^n$-action. Trans Amer Math Soc 310:303–311
46. Reissig R (1964) Ein funktionenanalytischer Existenzbeweis für periodische Lösungen. Monatsber Deutsche Akad Wiss Berlin 6:407–413
47. Scorza Dragoni G (1931) Il problema dei valori ai limite studiate in grande per le equazione differenziale del secondo ordine. Math Ann 105:133–143
48. Srzednicki R (1994) On periodic planar solutions of planar polynomial differential equations with periodic coefficients. J Differ Equations 114:77–100
49. Stoppelli F (1952) Su un'equazione differenziale della meccanica dei fili. Rend Accad Sci Fis Mat Napoli 19(4):109–114
50. Villari G (1965) Contributi allo studio dell'esistenzadi soluzioni periodiche per i sistemi di equazioni differenziali ordinarie. Ann Mat Pura Appl 69(4):171–190
51. Willem M (1981) Oscillations forcées de systèmes hamiltoniens. Publications du Séminaire d'Analyse non linéaire, Univ Besançon

Books and Reviews

Bobylev NA, Burman YM, Korovin SK (1994) Approximation procedures in nonlinear oscillation theory. de Gruyter, Berlin
Bogoliubov NN, Mitropolsky YA (1961) Asymptotic methods in the theory of nonlinear oscillations. Gordon and Breach, New York
Cesari L (1971) Asymptotic behavior and stability problems for ordinary differential equations, 3rd edn. Springer, Berlin
Coddington EA, Levinson N (1955) Ordinary differential equations. McGraw-Hill, New York
Cronin J (1964) Fixed points and topological degree in nonlinear analysis. Amer Math Soc, Providence
Ekeland I (1990) Convexity methods in hamiltonian mechanics. Springer, Berlin
Farkas M (1994) Periodic motions. Springer, New York
Gaines RE, Mawhin J (1977) Coincidence degree and nonlinear differential equations. Lecture Notes in Mathematics, vol 568. Springer, Berlin
Hale JK (1969) Ordinary differential equations. Wiley Interscience, New York
Mawhin J (1993) Topological degree and boundary value problems for nonlinear differential equations. Lecture Notes in Mathematics, vol 1537. Springer, Berlin, pp 74–142
Mawhin J (1995) Continuation theorems and periodic solutions of ordinary differential equations. In: Granas A, Frigon M (eds) Topological methods in differential equations and inclusions, NATO ASI C472. Kluwer, Amsterdam, pp 291–375
Pliss VA (1966) Nonlocal problems of the theory of oscillations. Academic Press, New York
Rabinowitz P (1986) Minimax methods in critical point theory with applications to differential equations. CBMS Reg Conf Ser Math No 65. Amer Math Soc, Providence
Reissig R, Sansone G, Conti R (1963) Qualitative Theorie nichtlinearer Differentialgleichungen. Cremonese, Roma
Reissig R, Sansone G, Conti R (1974) Nonlinear differential equations of higher order. Noordhoff, Leiden
Rouche N, Mawhin J (1980) Ordinary differential equations. Stability and periodic solutions. Pitman, Boston
Sansone G, Conti R (1964) Non-linear differential equations. Pergamon, Oxford
Verhulst F (1996) Nonlinear differential equations and dynamical systems, 2nd edn. Springer, Berlin

Perturbation Analysis of Parametric Resonance

Ferdinand Verhulst
Mathematisch Instituut, University of Utrecht, Utrecht, The Netherlands

Article Outline

Glossary
Definition of the Subject
Introduction
Perturbation Techniques
Parametric Excitation of Linear Systems
Nonlinear Parametric Excitation
Applications
Future Directions
Acknowledgment
Bibliography

Glossary

Coexistence The special case when all the independent solutions of a linear, T-periodic ODE are T-periodic.

Hill's equation A second order ODE of the form $\ddot{x} + p(t)x = 0$, with $p(t)$ T-periodic.

Instability pockets Finite domains, usually intersections of instability tongues, where the trivial solution of linear, T-periodic ODEs is unstable.

Instability tongues Domains in parameter space where the trivial solution of linear, T-periodic ODEs is unstable.

Mathieu equation An ODE of the form $\ddot{x} + (a + b\cos(t))x = 0$.

Parametric resonance Resonance excitation arising for special values of coefficients, frequencies and other parameters in T-periodic ODEs.

Quasi-periodic A function of the form $\sum_{i=1}^{n} f_i(t)$ with $f_i(t)$ T_i-periodic, n finite, and the periods T_i independent over $\mathbb{R}$.

Sum resonance A parametric resonance arising in the case of at least three frequencies in a T-periodic ODE.

Definition of the Subject

Parametric resonance arises in mechanics in systems with external sources of energy and for certain parameter values. Typical examples are the pendulum with oscillating support and a more specific linearization of this pendulum, the Mathieu equation in the form

$$\ddot{x} + (a + b\cos(t))x = 0 \,.$$

The time-dependent term represents the excitation. Tradition has it that parametric resonance is usually not considered in the context of systems with *external* excitation of the form $\dot{x} = f(x) + \phi(t)$, but for systems where time-dependence arises in the coefficients of the equation. Mechanically this means usually periodically varying stiffness, mass or load, in fluid or plasma mechanics one can think of frequency modulation or density fluctuation, in mathematical biology of periodic environmental changes. The term 'parametric' refers to the dependence on parameters and certain resonances arising for special values of the parameters. In the case of the Mathieu equation, the parameters are the frequency ω ($a = \omega^2$) of the equation without time-dependence and the excitation amplitude b; see Sect. "Parametric Excitation of Linear Systems" for an explicit demonstration of resonance phenomena in this two parameters system.

Mathematically the subject is concerned with ODEs with periodic coefficients. The study of linear dynamics of this type gave rise to a large amount of literature in the first half of the 20th century and this highly technical, classical material is still accessible in textbooks. The standard equations are Hill's equation and the Mathieu equation (see Subsect. "Elementary Theory"). We will summarize a number of basic aspects. The reader is also referred to the article ► Dynamics of Parametric Excitation by Alan Champneys in this Encyclopedia.

Recently, the interest in nonlinear dynamics, new applications and the need to explore higher dimensional problems has revived the subject. Also structural stability and persistence problems have been investigated. Such problems arise as follows. Suppose that we have found a number of interesting phenomena for a certain equation and suppose we embed this equation in a family of equations by adding parameters or perturbations. Do the 'interesting phenomena' persist in the family of equations? If not, we will call the original equation structurally unstable. A simple example of structural instability is the harmonic equation which shows qualitative different behavior on adding damping. In general, Hamiltonian systems are structurally unstable in the wider context of dissipative dynamical systems.

Introduction

Parametric resonance produces interesting mathematical challenges and plays an important part in many applications. The linear dynamics is already nontrivial whereas the nonlinear dynamics of such systems is extremely rich and largely unexplored. The role of symmetries is essential, both in linear and in nonlinear analysis. A classical example of parametric excitation is the swinging pendulum with oscillating support. The equation of motion describing the model is

$$\ddot{x} + \left(\omega_0^2 + p(t)\right)\sin x = 0 \,, \tag{1}$$

where $p(t)$ is a periodic function. Upon linearization – replacing $\sin x$ by x – we obtain Hill's equation (Subsect. "Elementary Theory"):

$$\ddot{x} + \left(\omega_0^2 + p(t)\right)x = 0 \,.$$

This equation was formulated around 1900 in the perturbation theory of periodic solutions in celestial mechanics. If we choose $p(t) = \cos\omega t$, Hill's equation becomes the Mathieu equation. It is well-known that special tuning of the frequency ω_0 and the period of excitation (of $p(t)$) produces interesting instability phenomena (resonance). More generally we may study nonlinear parametric equations of the form

$$\ddot{x} + k\dot{x} + (\omega_0^2 + p(t))F(x) = 0 \,, \tag{2}$$

where $k > 0$ is the damping coefficient, $F(x) = x + bx^2 + cx^3 + \cdots$ and time is scaled so that $p(t)$ is a π-periodic function with zero average. We may also take for $p(t)$ a quasi-periodic or almost-periodic function. The books [48] cover most of the classical theory, but for a nice introduction see [38]. In [36], emphasis is placed on the part played by parameters, it contains a rich survey of bifurcations of eigenvalues and various applications. There are many open questions for Eqs. (1) and (2); we shall discuss aspects of the classical theory, recent theoretical results and a few applications.

As noted before, in parametric excitation we have an oscillator with an independent source of energy. In examples, the oscillator is often described by a one degree of freedom system but of course many more degrees of freedom may play a part; see for instance in Sect. "Applications" the case of coupled Mathieu-equations as studied in [33]. In what follows, ε will always be a small, positive parameter.

Perturbation Techniques

In this section we review the basic techniques to handle parametric perturbation problems. In the case of Poincaré–Lindstedt series which apply to periodic solutions, the expansions are in integer powers of ε. It should be noted that in general, other order functions of ε may play a part; see Subsect. "Elementary Theory" and [46].

Poincaré–Lindstedt Series

One of the oldest techniques is to approximate a periodic solution by the construction of a convergent series in terms of the small parameter ε. The method can be used for equations of the form

$$\dot{x} = f(t, x) + \varepsilon g(t, x) + \varepsilon^2 \cdots ,$$

with $x \in \mathbb{R}^n$ and (usually) assuming that the 'unperturbed' problem $\dot{y} = f(t, y)$ is understood and can be solved. Note that the method can also be applied to perturbed maps and difference equations. Suppose that the unperturbed problem contains a periodic solution, under what conditions can this solution be continued for $\varepsilon > 0$? The answer is given by the conditions set by the implicit function theorem, see for formulations and theorems [30] and [44]. Usually we can associate with our perturbation problem a parameter space and one of the questions is then to find the domains of stability and instability. The common boundary of these domains is often characterized by the existence of periodic solutions and this is where Poincaré-Lindstedt series are useful. We will demonstrate this in the next section.

Averaging

Averaging is a normalization method. In general, the term"normalization" is used whenever an expression or quantity is put in a simpler, standardized form. For instance, a $n \times n$-matrix with constant coefficients can be put in Jordan normal form by a suitable transformation. When the eigenvalues are distinct, this is a diagonal matrix.

Introductions to normalization can be found in [1, 15,19] and [13]. For the relation between averaging and normalization in general the reader is referred to [34] and [44]. For averaging in the so-called standard form it is assumed that we can put the perturbation problem in the form

$$\dot{x} = \varepsilon F(t, x) + \varepsilon^2 \cdots ,$$

and that we have the existence of the limit

$$\lim_{T \to \infty} \frac{1}{T} \int_0^T F(t, x) \mathrm{d}t = F^0(x) .$$

The analysis of the averaged equation $\dot{y} = F^0(y)$ produces asymptotic approximations of the solutions of the original equation on a long timescale; see [34]. Also, under certain conditions, critical points of the averaged equation correspond with periodic solutions in the original system. The choice to use Poincaré-Lindstedt series or the averaging method is determined by the amount of information one wishes to obtain. To find the location of stability and instability domains (the boundaries), Poincaré-Lindstedt series are very efficient. On the other hand, with somewhat more efforts, the averaging method will also supply this information with in addition the behavior of the solutions within the domains. For an illustration see Subsect. "Elementary Theory".

Resonance

Assume that $x = 0$ is a critical point of the differential equation and write the system as:

$$\dot{x} = Ax + f(t, x, \varepsilon) , \tag{3}$$

with $x \in \mathbb{R}^n$, A a constant $n \times n$-matrix; $f(t, x, \varepsilon)$ can be expanded in a Taylor series with respect to ε and in homogeneous vector polynomials in x starting with quadratic terms. Normalization of Eq. (3) means that by successive transformation we remove as many terms of Eq. (3) as possible. It would be ideal if we could remove all the nonlinear terms, i. e. linearize Eq. (3) by transformation. In general, however, some nonlinearities will be left and this is where

resonance comes in. The eigenvalues $\lambda_1, \ldots, \lambda_n$ of the matrix A are resonant if for some $i \in \{1, 2, \ldots, n\}$ one has:

$$\sum_{j=1}^{n} m_j \lambda_j = \lambda_i \,, \tag{4}$$

with $m_j \geq 0$ integers and $m_1 + m_2 + \cdots + m_n \geq 2$. If the eigenvalues of A are non-resonant, we can remove all the nonlinear terms and so linearize the system. However, this is less useful than it appears, as in general the sequence of successive transformations to carry out the normalization will be divergent. The usefulness of normalization lies in removing nonresonant terms to a certain degree to simplify the analysis.

Normalization of Time-Dependent Vectorfields

In problems involving parametric resonance, we have time-dependent systems such as equations perturbing the Mathieu equation. Details of proofs and methods to compute the normal form coefficients in such cases can be found in [1,18] and [34]. We summarize some aspects. Consider the following parameter and time dependent equation:

$$\dot{x} = F(x, \mu, t) \,, \tag{5}$$

with $x \in \mathbb{R}^m$ and the parameters $\mu \in \mathbb{R}^p$.

Here $F(x, \mu, t) : \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}^m$ is C^∞ in x and μ and T-periodic in the t-variable. We assume that $x = 0$ is a solution, so $F(0, \mu, t) = 0$ and, moreover assume that the linear part of the vectorfield $D_x F(0, 0, t)$ is time-independent for all $t \in \mathbb{R}$. We will write $L_0 = D_x F(0, 0, t)$. Expanding $F(x, \mu, t)$ in a Taylor series with respect to x and μ yields the equation:

$$\dot{x} = L_0 x + \sum_{n=2}^{k} F_n(x, \mu, t) + O(|(x, \mu)|^{k+1}) \,, \tag{6}$$

where the $F_n(x, \mu, t)$ are homogeneous polynomials in x and μ of degree n with T-periodic coefficients.

Theorem 1 *Let $k \in \mathbb{N}$. There exists a (parameter- and time-dependent) transformation*

$$x = \hat{x} + \sum_{n=2}^{k} P_n(\hat{x}, \mu, t) \,,$$

where $P_n(\hat{x}, \mu, t)$ are homogeneous polynomials in x and μ of degree n with T-periodic coefficients, such that Eq. (6) takes the form (dropping the hat):

$$\begin{aligned} \dot{x} &= L_0 x + \sum_{n=2}^{k} \tilde{F}_n(x, \mu, t) + O(|(x, \mu)|^{k+1}) \,, \\ \dot{\mu} &= 0 \,. \end{aligned} \tag{7}$$

The truncated vectorfield:

$$\dot{x} = L_0 x + \sum_{n=2}^{k} \tilde{F}_n(x, \mu, t) = \tilde{F}(x, \mu, t) \,, \tag{8}$$

which will be called the normal form of Eq. (5), has the following properties:

1. *$\frac{d}{dt} e^{L_0^* t} \tilde{F}(e^{-L_0^* t} x, \mu, t) = 0$, for all $(x, \mu) \in \boldsymbol{R}^{m+p}$, $t \in \mathbb{R}$.*
2. *If Eq. (5) is invariant under an involution (i. e. $SF(x, \mu, t) = F(Sx, \mu, t)$ with S an invertible linear operator such that $S^2 = I$), then the truncated normal form (8) is also invariant under S. Similarly, if Eq. (5) is reversible under an involution R (i. e. $RF(x, \mu, t) = -F(Rx, \mu, t)$), then the truncated normal form (8) is also reversible under R.*

For a proof, see [18].

The theorem will be applied to situations where L_0 is semi-simple and has only purely imaginary eigenvalues. We take $L_0 = diag\{i\lambda_1, \ldots, i\lambda_m)\}$. In our applications, $m = 2l$ is even and $\lambda_{l+j} = -\lambda_j$ for $j = 1, \ldots, l$. The variable x is then often written as $x = (z_1, \ldots, z_l, \bar{z}_1, \ldots, \bar{z}_l)$.

Assume $L_0 = diag\{i\lambda_1, \ldots, i\lambda_m)\}$ then:

- A term $x_1^{\gamma_1} \ldots x_m^{\gamma_n} e^{i \frac{2\pi}{T} k t}$ is in the jth component of the Taylor–Fourier series of $\tilde{F}(x, \mu, t)$ if:

$$-\lambda_j + \frac{2\pi}{T} k + \gamma_1 \lambda_1 + \cdots + \gamma_m \lambda_m = 0 \,. \tag{9}$$

  This is known as the resonance condition.
- Transforming the normal form through $x = e^{L_0 t} w$ leads to an autonomous equation for w:

$$\dot{w} = \sum_{n=2}^{k} \tilde{F}_n(w, \mu, 0) \,. \tag{10}$$

- An important result is this: If Eq. (5) is invariant (respectively reversible) under an involution S, then this also holds for Eq. (10).
- The autonomous normal form (10) is invariant under the action of the group $\mathcal{G} = \{g | gx = e^{jL_0 T} x, j \in \mathbb{Z}\}$, generated by $e^{L_0 T}$. Note that this group is discrete if the ratios of the λ_i are rational and continuous otherwise.

For a proof of the last two statements see [31].

By this procedure we can make the system autonomous. This is very effective as the autonomous normal form (10) can be used to prove the existence of periodic solutions and invariant tori of Eq. (5) near $x = 0$. We have:

Theorem 2 *Let $\varepsilon > 0$, sufficiently small, be given. Scale $w = \varepsilon \hat{w}$.*

1. *If $\hat{w}_0$ is a hyperbolic fixed point of the (scaled) Eq. (10), then Eq. (5) has a hyperbolic periodic solution $x(t) = \varepsilon \hat{w}_0 + O(\varepsilon^{k+1})$.*
2. *If the scaled Eq. (10) has a hyperbolic closed orbit, then Eq. (5) has a hyperbolic invariant torus.*

These results are related to earlier theorems in [3], see also the survey [45]. Later we shall discuss normalization in the context of the so-called sum-resonance.

Remarks on Limit Sets

In studying a dynamical system the behavior of the solutions is for a large part determined by the limit sets of the system. The classical limit sets are equilibria and periodic orbits. Even when restricting to autonomous equations of dimension three, we have no complete classification of possible limit sets and this makes the recognition and description of non-classical limit sets important. In parametrically excited systems, the following limit sets, apart from the classical ones, are of interest:

- Chaotic attractors. Various scenarios were found, see [31,32,42].
- Strange attractors without chaos, see [27]. The natural presence of various forcing periods in real-life models make their occurrence quite plausible.
- Attracting tori. These limit sets are not difficult to find; they arise for instance as a consequence of a Neimark–Sacker bifurcation of a periodic solution, see [19].
- Attracting heteroclinic cycles, see [20].

A large number of these phenomena can be studied both by numerics *and* by perturbation theory; using the methods simultaneously gives additional insight.

Parametric Excitation of Linear Systems

As we have seen in the introduction, parametric excitation leads to the study of second order equations with periodic coefficients. More in general such equations arise from linearization near T-periodic solutions of T-periodic equations of the form $\dot{y} = f(t, y)$. Suppose $y = \phi(t)$ is a T-periodic solution; putting $y = \phi(t) + x$ produces upon linearization the T-periodic equation

$$\dot{x} = f_x(t, \phi(t))x \,. \tag{11}$$

This equation often takes the form

$$\dot{x} = Ax + \varepsilon B(t)x \,, \tag{12}$$

in which $x \in \mathbb{R}^m$; A is a constant $m \times m$-matrix, $B(t)$ is a continuous, T-periodic $m \times m$-matrix, ε is a small parameter. For elementary studies of such an equation, the Poincaré-Lindstedt method or continuation method is quite efficient. The method applies to nonlinear equations of arbitrary dimension, but we shall demonstrate its use for equations of Mathieu type.

Elementary Theory

Floquet theory tells us that the solutions of Eq. (12) can be written as:

$$x(t) = \Phi(t, \varepsilon)\mathrm{e}^{C(\varepsilon)t} \,, \tag{13}$$

with $\Phi(t, \varepsilon)$ a T-periodic $m \times m$-matrix, $C(\varepsilon)$ a constant $m \times m$-matrix and both matrices having an expansion in order functions of ε. The determination of $C(\varepsilon)$ provides us with the stability behavior of the solutions. A particular case of Eq. (12) is Hill's equation:

$$\ddot{x} + b(t, \varepsilon)x = 0 \,, \tag{14}$$

which is of second order; $b(t, \varepsilon)$ is a scalar T-periodic function. A number of cases of Hill's equation are studied in [23]. A particular case of Eq. (14) which arises frequently in applications is the Mathieu equation:

$$\ddot{x} + (\omega^2 + \varepsilon \cos 2t)x = 0, \omega > 0 \,, \tag{15}$$

which is reversible. (In [23] one also finds Lamé's, Ince's, Hermite's, Whittaker–Hill and other Hill equations.) A typical question is: for which values of ω and ε in (ω^2, ε)-parameter space is the trivial solution $x = \dot{x} = 0$ stable?

Solutions of Eq. (15) can be written in the Floquet form (13), where in this case $\Phi(t, \varepsilon)$ will be π-periodic. The eigenvalues λ_1, λ_2 of C, which are called characteristic exponents and are ε-dependent, determine the stability of the trivial solution. For the characteristic exponents of Eq. (12) we have:

$$\sum_{i=1}^{n} \lambda_i = \frac{1}{T} \int_0^T Tr(A + \varepsilon B(t))\mathrm{d}t \,, \tag{16}$$

see Theorem 6.6 in [44]. So in the case of Eq. (15) we have:

$$\lambda_1 + \lambda_2 = 0 \,. \tag{17}$$

The exponents are functions of ε, $\lambda_1 = \lambda_1(\varepsilon)$, $\lambda_2 = \lambda_2(\varepsilon)$ and clearly $\lambda_1(0) = i\omega$, $\lambda_2(0) = -i\omega$. As $\lambda_1(\varepsilon) = -\lambda_2(\varepsilon)$, the characteristic exponents, which are complex conjugate, are purely imaginary or real. The implication is

that if $\omega^2 \neq n^2$, $n = 1, 2, \ldots$ the characteristic exponents are purely imaginary and $x = 0$ is stable near $\varepsilon = 0$. If $\omega^2 = n^2$ for some $n \in \mathbb{N}$, however, the imaginary part of $\exp(C(\varepsilon)t)$ can be absorbed into $\Phi(t, \varepsilon)$ and the characteristic exponents may be real. We assume now that $\omega^2 = n^2$ for some $n \in \mathbb{N}$, or near this value, and we shall look for periodic solutions of $x(t)$ of Eq. (15) as these solutions define the boundaries between stable and unstable solutions. We put:

$$\omega^2 = n^2 - \varepsilon\beta \,, \tag{18}$$

with β a constant, and we apply the Poincaré-Lindstedtj method to find the periodic solutions; see Appendix 2 in [44]. We find that periodic solutions exist for $n = 1$ if:

$$\omega^2 = 1 \pm \frac{1}{2}\varepsilon + \mathcal{O}(\varepsilon^2) \,.$$

In the case $n = 2$, periodic solutions exist if:

$$\begin{aligned} \omega^2 &= 4 - \frac{1}{48}\varepsilon^2 + \mathcal{O}(\varepsilon^4) \,, \\ \omega^2 &= 4 + \frac{5}{48}\varepsilon^2 + \mathcal{O}(\varepsilon^4) \,. \end{aligned} \tag{19}$$

The corresponding instability domains are called Floquet tongues, instability tongues or resonance tongues, see Fig. 1.

On considering higher values of n, we have to calculate to a higher order of ε. At $n = 1$ the boundary curves are intersecting at positive angles at $\varepsilon = 0$, at $n = 2$ ($\omega^2 = 4$)

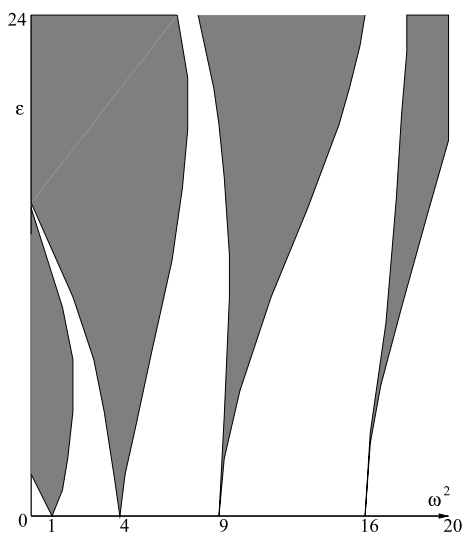

Perturbation Analysis of Parametric Resonance, Figure 1
Floquet tongues of the Mathieu Eq. (15); the instability domains are shaded

they are tangent; the order of tangency increases as $n - 1$ (contact of order n), making instability domains more and more narrow with increasing resonance number n.

Higher Order Approximation and an Unexpected Timescale The instability tongue of the Mathieu equation at $n = 1$ can be determined with more precision by Poincaré expansion. On using averaging, one also characterizes the flow outside the tongue boundary and this results in a surprise. Consider Eq. (15) in the form

$$\ddot{x} + (1 + \varepsilon a + \varepsilon^2 b + \varepsilon \cos 2t)x = 0,$$

where we can choose $a = \pm\frac{1}{2}$ to put the frequency with first order precision at the tongue boundary. The eigenvalues of the trivial solution are from first order averaging

$$\lambda_{1,2} = \pm\frac{1}{2}\sqrt{\frac{1}{4} - a^2},$$

which agrees with Poincaré expansion; $a^2 > \frac{1}{4}$ gives stability, the $<$ inequality instability. The transition value $a^2 = \frac{1}{4}$ gives the tongue location. Take for instance the $+$ sign. Second order averaging, see [46], produces for the eigenvalues of the trivial solution

$$\lambda_{1,2} = \pm\sqrt{\begin{array}{l} -\frac{1}{4}\left(\frac{1}{32} + b\right)\varepsilon^3 + \left(\frac{1}{64} + \frac{1}{2}b\right) \\ \cdot\left(\frac{7}{64} - \frac{1}{2}b\right)\varepsilon^4 \end{array}} \,.$$

So, if $\frac{1}{32} + b > 0$ we have stability, if $\frac{1}{32} + b < 0$ instability; at $b = -\frac{1}{32}$ we have the second order approximation of this tongue boundary. Note that near this boundary the solutions are characterized by eigenvalues of $O(\varepsilon^{\frac{3}{2}})$ and accordingly the time-dependence by timescale $\varepsilon^{\frac{3}{2}}t$.

The Mathieu Equation with Viscous Damping In real-life applications there is always the presence of damping. We shall consider the effect of its simplest form, small viscous damping. Eq. (15) is extended by adding a linear damping term:

$$\ddot{x} + \kappa\dot{x} + (\omega^2 + \varepsilon\cos 2t)x = 0 \,, \quad a, \kappa > 0 \,. \tag{20}$$

We assume that the damping coefficient is small, $\kappa = \varepsilon\kappa_0$, and again we put $\omega^2 = n^2 - \varepsilon\beta$ to apply the Poincaré-Lindstedt method.

We find periodic solutions in the case $n = 1$ if:

$$\omega^2 = 1 \pm \sqrt{\frac{1}{4}\varepsilon^2 - \kappa^2} \,. \tag{21}$$

Relation (21) corresponds with the curve of periodic solutions, which in (ω^2, ε)-parameter space separates stable and unstable solutions. We observe the following phenomena. If $0 < \kappa < \frac{1}{2}\varepsilon$, we have an instability domain

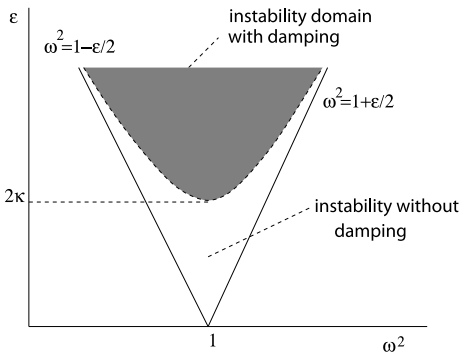

Perturbation Analysis of Parametric Resonance, Figure 2
First order approximation of instability domains without and with damping for Eq. (20) near $\omega^2 = 1$

which by damping has been lifted from the ω^2-axis; also *the width has shrunk*. If $\kappa > \frac{1}{2}\varepsilon$ the instability domain has vanished. For an illustration see Fig. 2.

Repeating the calculations for $n \geq 2$ we find no instability domains at all; damping of $O(\varepsilon)$ stabilizes the system for ε small. To find an instability domain we have to decrease the damping, for instance if $n = 2$ we have to take $\kappa = \varepsilon^2\kappa_0$.

Coexistence Linear periodic equations of the form (12) have m independent solutions and it is possible that all the independent solutions are periodic. This is called 'coexistence' and one of the consequences is that the instability tongues vanish. An example is Ince's equation:

$$(1 + a\cos t)\ddot{x} + \kappa \sin t\dot{x} + (\omega^2 + \varepsilon\cos t)x = 0\,,$$

see [23]. An interesting question is whether this phenomenon persists under nonlinear perturbations; we return to this question in Subsect. "Coexistence Under Nonlinear Perturbation".

More General Classical Results

The picture presented by the Mathieu equation resulting in resonance tongues in the ω, ε-parameter space, stability and instability intervals as parametrized by ω shown in Fig. 1, has been studied for more general types of Hill's equation. The older literature can be found in [39], see also [43].

Consider Hill's equation in the form

$$\ddot{x} + (\omega^2 + \varepsilon f(t))x = 0\,, \tag{22}$$

with $f(t)$ periodic and represented by a Fourier series. Along the ω^2-axis there exist instability intervals of size L_m, where m indicates the mth instability interval. In the case of the Mathieu equation, we have from [16]

$$L_m = O(\varepsilon^m)\,.$$

The resonance tongues become increasingly narrow.

For general periodic $f(t)$ we have weak estimates, like $L_m = O(\varepsilon)$, but if we assume that the Fourier series is finite, the estimates can be improved. Put

$$f(t) = \sum_{j=0}^{s} f_j \cos 2jt\,,$$

so $f(t)$ is even and π-periodic. From [22] we have the following estimates:

- If we can write $m = sp$ with $p \in \mathbb{N}$, we have

$$L_m = \frac{8s^2}{((p-1)!)^2}\left(\frac{|f_s\varepsilon|}{8s^2}\right)^p + O(\varepsilon^{p+1})\,.$$

- If we can not decompose m like this and $sp < m < s(p+1)$, we have

$$L_m = O(\varepsilon^{p+1})\,.$$

In the case of Eq. (22) we have no dissipation and then it can be useful to introduce canonical transformations and Poincaré maps. In this case, for example, put

$$\dot{x} = y,\ \ \dot{y} = -\frac{\partial H}{\partial x}\,,$$

with Hamiltonian function

$$H(x, y, t) = \frac{1}{2}y^2 + \frac{1}{2}(\omega^2 + \varepsilon f(t))x^2\,.$$

We can split $H = H_0 + \varepsilon H_1$ with $H_0 = \frac{1}{2}(y^2 + \omega^2 x^2)$ and apply canonical perturbation theory. Examples of this line of research can be found in [6] and [10]. Interesting conclusions can be drawn with respect to the geometry of the resonance tongues, crossings of tongues and as a possible consequence the presence of so-called instability pockets. In this context, the classical Mathieu equation turns out to be quite degenerate.

Hill's equation in the case of damping was considered in [35]; see also [36] where an arbitrary number of degrees of freedom is discussed.

Quasi-Periodic Excitation

Equations of the form

$$\ddot{x} + (\omega^2 + \varepsilon p(t))x = 0\,, \tag{23}$$

with parametric excitation $p(t)$ quasi-periodic or almost-periodic, arise often in applications. Floquet theory does not apply in this case but we can still use perturbation theory. A typical example would be two rationally independent frequencies:

$$p(t) = \cos t + \cos \gamma t \,,$$

with γ irrational. As an interesting example, in [7], $\gamma = \frac{1}{2}(1+\sqrt{5})$ was chosen, the golden number. It will be no surprise that many more complications arise for large values of ε, but for ε small (the assumption in this article), the analysis runs along similar lines producing resonance tongues, crossings of tongues and instability pockets. See also extensions in [11]. Detailed perturbation expansions are presented in [49] with a comparison of Poincaré expansion, the harmonic balance method and numerics; there is good agreement between the methods. Real-life models contain dissipation which inspired the authors of [49] to consider the equation

$$\ddot{x} + 2\mu\dot{x} + (\omega^2 + \varepsilon(\cos t + \cos \gamma t))x = 0, \quad \mu > 0 \,,$$

γ irrational. They conclude that

- The instability tongues become thinner and recede into the ω-axis as μ increases.
- High-order resonance tongues seem to be more affected by dissipation than low-order ones producing a dramatic loss of 'fine detail', even for small μ.
- The results of varying the parameter μ certainly needs more investigation.

Parametrically Forced Oscillators in Sum Resonance

In applications where more than one degree of freedom plays a part, many more resonances are possible. For a number of interesting cases and additional literature see [36]. An important case is the so-called sum resonance. In [17] a geometrical explanation is presented for the phenomena in this case using 'all' the parameters as unfolding parameters. It will turn out that four parameters are needed to give a complete description. Fortunately three suffice to visualize the situation. Consider the following type of differential equation with three frequencies

$$\dot{z} = Az + \varepsilon f(z, \omega_0 t; \lambda), \quad z \in \mathbb{R}^4, \quad \lambda \in \mathbb{R}^p \,, \tag{24}$$

which describes a system of two parametrically forced coupled oscillators. Here A is a 4×4 matrix, containing parameters, and with purely imaginary eigenvalues $\pm i\omega_1$ and $\pm i\omega_2$. The vector valued function f is 2π-periodic in $\omega_0 t$ and $f(0, \omega_0 t; \lambda) = 0$ for all t and λ. Equation (24) can be resonant in many different ways. We consider the sum resonance

$$\omega_1 + \omega_2 = \omega_0 \,,$$

where the system may exhibit instability. The parameter λ is used to control detuning $\delta = (\delta_1, \delta_2)$ of the frequencies (ω_1, ω_2) from resonance and damping $\mu = (\mu_1, \mu_2)$. We summarize the analysis from [17].

- The first step is to put Eq. (24) into normal form by normalization or averaging. In the normalized equation the time-dependence appears only in the higher order terms. But the autonomous part of this equation contains enough information to determine the stability regions of the origin. The linear part of the normal form is $\dot{z} = A(\delta, \mu)z$ with

$$A(\delta, \mu) = \begin{pmatrix} B(\delta, \mu) & 0 \\ 0 & \overline{B}(\delta, \mu) \end{pmatrix} , \tag{25}$$

and

$$B(\delta, \mu) = \begin{pmatrix} i\delta_1 - \mu_1 & \alpha_1 \\ \overline{\alpha}_2 & -i\delta_2 - \mu_2 \end{pmatrix} . \tag{26}$$

Since $A(\delta, \mu)$ is the complexification of a real matrix, it commutes with complex conjugation. Furthermore, according to the normal form Theorem 1 and if ω_1 and ω_2 are independent over the integers, the normal form of Eq. (24) has a continuous symmetry group.
- The second step is to test the linear part $A(\delta, \mu)$ of the normalized equation for structural stability i. e. to answer the question whether there exist open sets in parameter space where the dynamics is qualitatively the same. The family of matrices $A(\delta, \mu)$ is parametrized by the detuning δ and the damping μ. We first identify the most degenerate member N of this family and then show that $A(\delta, \mu)$ is its versal unfolding in the sense of [1]. The family $A(\delta, \mu)$ is equivalent to a versal unfolding $U(\lambda)$ of the degenerate member N.
- Put differently, the family $A(\delta, \mu)$ is structurally stable for $\delta, \mu > 0$, whereas $A(\delta, 0)$ is not. This has interesting consequences in applications as small damping and zero damping may exhibit very different behavior, see Sect. "Rotor Dynamics". In parameter space, the stability regions of the trivial solution are separated by a *critical surface* which is the hypersurface where $A(\delta, \mu)$ has at least one pair of purely imaginary complex conjugate eigenvalues. This critical surface is diffeomorphic to the *Whitney umbrella*, see Fig. 3 and for references [17]. It is the singularity of the Whitney umbrella that causes the discontinuous behavior of the stability diagram in

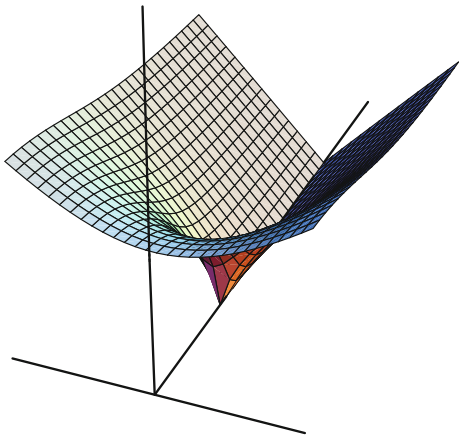

Perturbation Analysis of Parametric Resonance, Figure 3
The critical surface in (μ_+, μ_-, δ_+) space for Eq. (24). $\mu_+ = \mu_1 + \mu_2$, $\mu_- = \mu_1 - \mu_2$, $\delta_+ = \delta_1 + \delta_2$. Only the part $\mu_+ > 0$ and $\delta_+ > 0$ is shown. The parameters δ_1, δ_2 control the detuning of the frequencies, the parameters μ_1, μ_2 the damping of the oscillators (vertical direction). The base of the umbrella lies along the δ_+-axis

Sect. "Rotor Dynamics". The structural stability argument guarantees that the results are 'universally valid', i. e. they qualitatively hold for generic systems in sum resonance.

Nonlinear Parametric Excitation

Adding nonlinear effects to parametric excitation strongly complicates the dynamics. We start with adding nonlinear terms to the (generalized) Mathieu equation. Consider the following equation that includes dissipation:

$$\ddot{x} + \kappa\dot{x} + (\omega^2 + \varepsilon p(t))f(x) = 0\,, \tag{27}$$

where $\kappa > 0$ is the damping coefficient, $f(x) = x + bx^2 + cx^3 + \cdots$, and time is scaled so that:

$$p(t) = \sum_{l\in Z} a_{2l}\mathrm{e}^{2ilt}, \quad a_0 = 0, \quad a_{-2l} = \bar{a}_{2l}\,, \tag{28}$$

is an even π-periodic function with zero average. As we have seen in Sect. "Parametric Excitation of Linear Systems", the trivial solution $x = 0$ is unstable when $\kappa = 0$ and $\omega^2 = n^2$, for all $n \in \mathbb{N}$. Fix a specific $n \in \mathbb{N}$ and assume that ω^2 is close to n^2. We will study the bifurcations from the solution $x = 0$ in the case of primary resonance, which by definition occurs when the Fourier expansion of $p(t)$ contains nonzero terms $a_{2n}\mathrm{e}^{2int}$ and $a_{-2n}\mathrm{e}^{-2int}$. The bifurcation parameters in this problem are the detuning $\sigma = \omega^2 - n^2$, the damping coefficient κ and the Fourier coefficients of $p(t)$, in particular a_{2n}. The Fourier coefficients are assumed to be of equal order of magnitude.

The Conservative Case, $\kappa = 0$

An early paper is [21] in which Eq. (27) for $\kappa = 0$ is associated with the Hamiltonian

$$H(x, \dot{x}, t) = \frac{1}{2}\dot{x}^2 + \frac{\omega^2}{2}x^2 + p(t)\int_0^x f(s)\mathrm{d}s\,.$$

After transformation of the Hamiltonian, Lie transforms are implemented by MACSYMA to produce normal form approximations to $O(\varepsilon^2)$. A number of examples show interesting bifurcations.

A related approach can be found in [5]; as $p(t)$ is even, the equation is time-reversible. After construction of the Poincaré (timeperiodic) map, normal forms are obtained by equivariant transformations. This leads to a classification of integrable normal forms that are approximations of the family of Poincaré maps, a family as the map is parametrized by ω and the coefficients of $p(t)$.

Interestingly, the nonlinearity αx^3 is combined with the quasi-periodic Mathieu equation in [50] where global phenomena are described like resonance bands and chaos.

Adding Dissipation, $\kappa > 0$

Again time-periodic normal form calculations are used to approximate the dynamics; see [31], also [32] and the monograph [42]. The reflection symmetry in the normal form equations implies that all fixed points come in pairs, and that bifurcations of the origin will be symmetric (such as pitchfork bifurcations). We observe that the normal form equations show additional symmetries if either $f(x)$ in Eq. (27) is odd in x or if n is odd. The general normal form can be seen as a non-symmetric perturbation of the symmetric case. One finds pitchfork and saddle-node bifurcations, in fact all codimension one bifurcations; for details and pictures see Chap. 9 in [42].

Coexistence Under Nonlinear Perturbation

A model describing free vibrations of an elastica is described in [26]:

$$\left(1 - \frac{\varepsilon}{2}\cos 2t\right)\ddot{x} + \varepsilon\sin 2t\dot{x} + cx + \varepsilon\alpha x^2 = 0\,.$$

For $\alpha = 0$, the equation shows the phenomenon of coexistence. It is shown by second order averaging in [26] that for $\alpha \neq 0$ there exist open sets of parameter values for which the trivial solution is unstable.

An application to the stability problem of a family of periodic solutions in a Hamiltonian system is given in [29].

Other Nonlinearities

In applications various nonlinear terms play a part. In [25] one considers

$$\ddot{x} + (\omega^2 + \varepsilon\cos(t)) + \varepsilon(Ax^3 + Bx^2\dot{x} + Cx\dot{x}^2 + D\dot{x}^3) = 0\,,$$

where averaging is applied near the $2:1$-resonance. If $B, D < 0$ the corresponding terms can be interpreted as progressive damping. It turns out that for a correct description of the bifurcations second-order averaging is needed.

Nonlinear damping can be of practical interest. The equation

$$\ddot{x} + (\omega^2 + \varepsilon\cos(t)) + \mu|\dot{x}|\dot{x} = 0\,,$$

is studied with μ also a small parameter. A special feature is that an acceptable description of the phenomena can be obtained in a semi-analytical way by using Mathieu-functions as starting point. The analysis involves the use of Padé-approximants, see [28].

Applications

There are many applications of parametric resonance, in particular in engineering. In this section we consider a number of significant applications, but of course without any attempt at completeness. See also [36] and the references in the additional literature.

The Parametrically Excited Pendulum

Choosing the pendulum case $f(x) = \sin(x)$ in Eq. (27) we have

$$\ddot{x} + \kappa\dot{x} + (\omega^2 + \varepsilon p(t))\sin(x) = 0\,.$$

It is natural, because of the sin periodicity, to analyze the Poincaré map on the cylindrical section $t = 0 \bmod 2\pi\mathbb{Z}$. This map has both a spatial and a temporal symmetry. As we know from the preceding section, perturbation theory applied near the equilibria $x = 0, x = \pi$, produces integrable normal forms. For larger excitation (larger values of ε), the system exhibits the usual picture of Hamiltonian chaos; for details see [12,24].

The inverted case is intriguing. It is well-known that the upper equilibrium of an undamped pendulum can be stabilized by vertical oscillations of the suspension point with certain frequencies. See for references [8,9] and [37]. In [8] the genericity of the classical result is studied for (conservative) perturbations respecting the symmetries of the equation. In [9] genericity is studied for (conservative) perturbations where the spatial symmetry is broken, replacing $\sin x$ by more general 2π-periodic functions. Stabilization is still possible but the dynamics is more complicated.

Rotor Dynamics

When adding linear damping to a system there can be a striking discontinuity in the bifurcational behavior. Phenomena like this have already been observed and described in for instance [48] or [40]. The discontinuity is a fundamental structural instability in linear gyroscopic systems with at least two degrees of freedom and with linear damping. The following example is based on [41] and [33].

Consider a rigid rotor consisting of a heavy disk of mass M which is rotating with rotationspeed Ω around an axis. The axis of rotation is elastically mounted on a foundation; the connections which are holding the rotor in an upright position are also elastic. To describe the position of the rotor we have the axial displacement u in the vertical direction (positive upwards), the angle of the axis of rotation *with respect to* the z-axis and *around* the z-axis. Instead of these two angles we will use the projection of the center of gravity motion on the horizontal (x, y)-plane, see Fig. 4. Assuming small oscillations in the upright (u) posi-

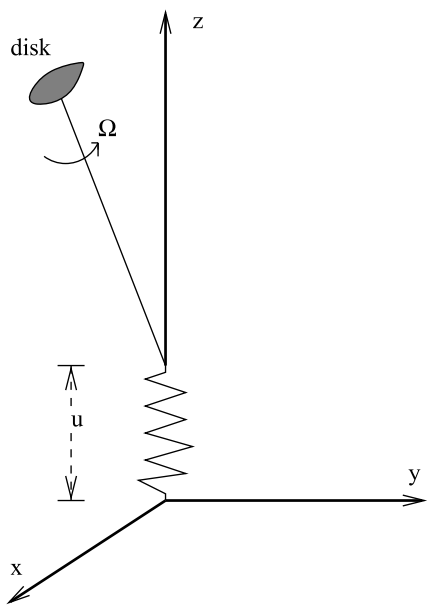

Perturbation Analysis of Parametric Resonance, Figure 4
Rotor with diskmass *M*, elastically mounted with axial (*u*) and lateral directions

tion, frequency 2η, the equations of motion become:

$$\begin{aligned} \ddot{x} + 2\alpha\dot{y} + (1 + 4\varepsilon\eta^2\cos 2\eta t)x = 0\,, \\ \ddot{y} - 2\alpha\dot{x} + (1 + 4\varepsilon\eta^2\cos 2\eta t)y = 0\,. \end{aligned} \tag{29}$$

System (29) constitutes a system of Mathieu-like equations, where we have neglected the effects of damping. Abbreviating $P(t) = 4\eta^2\cos 2\eta t$, the corresponding Hamiltonian is:

$$H = \frac{1}{2}(1+\alpha^2+\varepsilon P(t))x^2 + \frac{1}{2}p_x^2 + \frac{1}{2}(1+\alpha^2+\varepsilon P(t))y^2 + \frac{1}{2}p_y^2 + \alpha x p_y - \alpha y p_x\,,$$

where p_x, p_y are the momenta. The natural frequencies of the unperturbed system (29), $\varepsilon = 0$, are $\omega_1 = \sqrt{\alpha^2+1} + \alpha$ and $\omega_2 = \sqrt{\alpha^2+1} - \alpha$. By putting $z = x + iy$, system (29) can be written as:

$$\ddot{z} - 2\alpha i\dot{z} + (1 + 4\varepsilon\eta^2\cos 2\eta t)z = 0\,. \tag{30}$$

Introducing the new variable:

$$v = \mathrm{e}^{-i\alpha t}z\,, \tag{31}$$

and putting $\eta t = \tau$, we obtain:

$$v'' + \left(\frac{1+\alpha^2}{\eta^2} + 4\varepsilon\cos 2\tau\right)v = 0\,, \tag{32}$$

where the prime denotes differentiation with respect to τ. By writing down the real and imaginary parts of this equation, we get two identical Mathieu equations. We conclude that the trivial solution is stable for ε small enough, providing that $\sqrt{1+\alpha^2}$ is not close to $n\eta$, for some $n = 1, 2, 3, \ldots$. The first-order interval of instability, $n = 1$, arises if:

$$\sqrt{1+\alpha^2} \approx \eta\,. \tag{33}$$

If condition (33) is satisfied, the trivial solution of Eq. (32) is unstable. Therefore, the trivial solution of system (29) is also unstable. Note that this instability arises when:

$$\omega_1 + \omega_2 = 2\eta\,,$$

i. e. when the sum of the eigenfrequencies of the unperturbed system equals the excitation frequency 2η. This is known as a sum resonance of first order. The domain of instability can be calculated as in Subsect. "Elementary Theory"; we find for the boundaries:

$$\eta_b = \sqrt{1+\alpha^2}\,(1 \pm \varepsilon) + O(\varepsilon^2)\,. \tag{34}$$

The second order interval of instability of Eq. (32), $n = 2$, arises when:

$$\sqrt{1+\alpha^2} \approx 2\eta\,, \tag{35}$$

i. e. $\omega_1 + \omega_2 \approx \eta$. This is known as a sum resonance of second order. As above, we find the boundaries of the domains of instability:

$$\begin{aligned} 2\eta &= \sqrt{1+\alpha^2}\left(1 + \frac{1}{24}\varepsilon^2\right) + O(\varepsilon^4)\,, \\ 2\eta &= \sqrt{1+\alpha^2}\left(1 - \frac{5}{24}\varepsilon^2\right) + O(\varepsilon^4)\,. \end{aligned} \tag{36}$$

Higher order combination resonances can be studied in the same way; the domains of instability in parameter space continue to narrow as n increases. It should be noted that the parameter α is proportional to the rotation speed Ω of the disk and to the ratio of the moments of inertia.

Instability by Damping We add small linear damping to system (29), with positive damping parameter $\mu = 2\varepsilon\kappa$. This leads to the equation:

$$\ddot{z} - 2\alpha i\dot{z} + (1 + 4\varepsilon\eta^2\cos 2\eta t)z + 2\varepsilon\kappa\dot{z} = 0\,. \tag{37}$$

Because of the damping term, we can no longer reduce the complex Eq. (37) to two identical second order real equations, as we did in the previous section. In the sum resonance of the first order, we have $\omega_1 + \omega_2 \approx 2\eta$ and the solution of the unperturbed ($\varepsilon = 0$) equation can be written as:

$$z(t) = z_1\mathrm{e}^{i\omega_1 t} + z_2\mathrm{e}^{-i\omega_2 t}\,, \quad z_1, z_2 \in \mathbb{C}\,, \tag{38}$$

with $\omega_1 = \sqrt{\alpha^2+1} + \alpha$, $\omega_2 = \sqrt{\alpha^2+1} - \alpha$. Applying variation of constants leads to equations for z_1 and z_2:

$$\begin{aligned} \dot{z}_1 &= \frac{i\varepsilon}{\omega_1+\omega_2}\Big(2\kappa\big(i\omega_1 z_1 - i\omega_2 z_2 \mathrm{e}^{-i(\omega_1+\omega_2)t}\big) \\ &\quad + 4\eta^2\cos 2\eta t\big(z_1 + z_2\mathrm{e}^{-i(\omega_1+\omega_2)t}\big)\Big), \\ \dot{z}_2 &= \frac{-i\varepsilon}{\omega_1+\omega_2}\Big(2\kappa\big(i\omega_1 z_1 \mathrm{e}^{i(\omega_1+\omega_2)t} - i\omega_2 z_2\big) \\ &\quad + 4\eta^2\cos 2\eta t\big(z_1\mathrm{e}^{i(\omega_1+\omega_2)t} + z_2\big)\Big)\,. \end{aligned} \tag{39}$$

To calculate the instability interval around the value $\eta_0 = \frac{1}{2}(\omega_1 + \omega_2) = \sqrt{\alpha^2+1}$, we apply perturbation the-

ory to find for the stability boundary:

$$\eta_b = \sqrt{1+\alpha^2}\left(1 \pm \varepsilon\sqrt{1+\alpha^2-\frac{\kappa^2}{\eta_0^2}}+\cdots\right),$$
$$= \sqrt{1+\alpha^2}\left(1 \pm \sqrt{(1+\alpha^2)\varepsilon^2-\left(\frac{\mu}{\eta_0}\right)^2}+\cdots\right). \tag{40}$$

It follows that the domain of instability actually becomes *larger* when damping is introduced. The most unusual aspect of the above expression for the instability interval, however, is that there is a discontinuity at $\kappa = 0$. If $\kappa \to 0$, then the boundaries of the instability domain tend to the limits $\eta_b \to \sqrt{1+\alpha^2}(1 \pm \varepsilon\sqrt{1+\alpha^2})$ which differs from the result we found when $\kappa = 0:\ \eta_b = \sqrt{1+\alpha^2}(1 \pm \varepsilon)$.

In mechanical terms, the broadening of the instability-domain is caused by the coupling between the two degrees of freedom of the rotor in lateral directions which arises in the presence of damping. Such phenomena are typical for gyroscopic systems and have been noted earlier in the literature; see [2,4] and [36]. The explanation of the discontinuity and its genericity in [17], see Subsect. "Parametrically Forced Oscillators in Sum Resonance", is new. For hysteresis and phase-locking phenomena in this problem, the reader is referred to [33].

Autoparametric Excitation

In [42], autoparametric systems are characterized as vibrating systems which consist of at least two consisting subsystems that are coupled. One is a Primary System that can be in normal mode vibration. In the instability (parameter) intervals of the normal mode solution in the full, coupled system, we have *autoparametric resonance*. The vibrations of the Primary System act as parametric excitation of the Secondary System which will no longer remain at rest. An example is presented in Fig. 5.

In actual engineering problems, we wish sometimes to diminish the vibration amplitudes of the Primary System; sometimes this is called 'quenching of vibrations'. In other cases we have a coupled Secondary System which we would like to keep at rest. As an example we consider the following autoparametric system studied in [14]:

$$x'' + x + \varepsilon\left(k_1 x' + \sigma_1 x + a\cos 2\tau x + \frac{4}{3}x^3 + c_1 y^2 x\right) = 0$$
$$y'' + y + \varepsilon\left(k_2 y' + \sigma_2 y + c_2 x^2 y + \frac{4}{3}y^3\right) = 0 \tag{41}$$

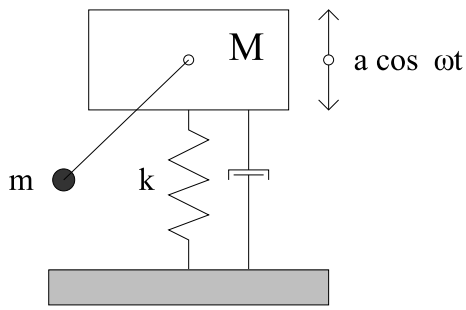

Perturbation Analysis of Parametric Resonance, Figure 5
Two coupled oscillators with vertical oscillations as Primary System and parametric excitation of the coupled pendulum (Secondary System)

where σ_1 and σ_2 are the detunings from the $1:1$-resonance of the oscillators. In this system, $y(t) = \dot{y}(t) = 0$ corresponds with a normal mode of the x-oscillator.

The system (41) is invariant under $(x, y) \to (x, -y)$, $(x, y) \to (-x, y)$, and $(x, y) \to (-x, -y)$. Using the method of averaging as a normalization procedure we investigate the stability of solutions of system (41). To give an explicit example we follow [14] in more detail. Introduce the usual variation of constants transformation:

$$x = u_1\cos\tau + v_1\sin\tau\ ; \quad x' = -u_1\sin\tau + v_1\cos\tau \tag{42}$$

$$y = u_2\cos\tau + v_2\sin\tau\ ; \quad y' = -u_2\sin\tau + v_2\cos\tau \tag{43}$$

After rescaling $\tau = \frac{\varepsilon}{2}\tilde{\tau}$ the averaged system of (41) becomes:

$$\begin{aligned}
u_1' &= -k_1 u_1 + \left(\sigma_1 - \frac{1}{2}a\right)v_1 + v_1\left(u_1^2 + v_1^2\right) \\
&\quad + \frac{1}{4}c_1 u_2^2 v_1 + \frac{3}{4}c_1 v_2^2 v_1 + \frac{1}{2}c_1 u_2 v_2 u_1 \\
v_1' &= -k_1 v_1 - \left(\sigma_1 + \frac{1}{2}a\right)u_1 - u_1\left(u_1^2 + v_1^2\right) \\
&\quad - \frac{3}{4}c_1 u_2^2 u_1 - \frac{1}{4}c_1 v_2^2 u_1 - \frac{1}{2}c_1 u_2 v_2 v_1 \\
u_2' &= -k_2 u_2 + \sigma_2 v_2 + v_2\left(u_2^2 + v_2^2\right) + \frac{1}{4}c_2 u_1^2 v_2 \\
&\quad + \frac{3}{4}c_2 v_1^2 v_2 + \frac{1}{2}c_2 u_1 v_1 u_2 \\
v_2' &= -k_2 v_2 - \sigma_2 u_2 - u_2\left(u_2^2 + v_2^2\right) - \frac{3}{4}c_2 u_1^2 u_2 \\
&\quad - \frac{1}{4}c_2 v_1^2 u_2 - \frac{1}{2}c_2 u_1 v_1 v_2\ .
\end{aligned} \tag{44}$$

This system is analyzed for critical points, periodic and quasi-periodic solutions, producing existence and stability diagrams in parameter space. The system also con-

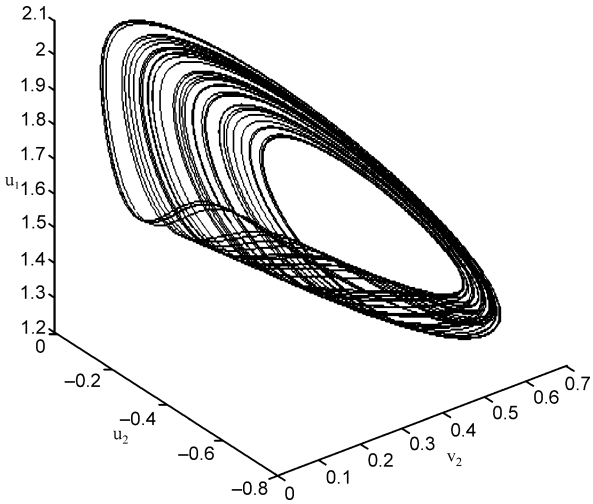

Perturbation Analysis of Parametric Resonance, Figure 6
The strange attractor of the averaged system (44). The phase-portraits in the (u_2, v_2, u_1)-space for $c_2 < 0$ at the value $\sigma_2 = 5.3$. The Kaplan–Yorke dimension for $\sigma_2 = 5.3$ is 2.29

tains a sequence of period-doubling bifurcations leading to chaotic solutions, see Fig. 6.

To prove the presence of chaos involves an application of higher dimensional Melnikov theory developed in [47]. A rather technical analysis in [14] shows the existence of a Šilnikov orbit in the averaged equation, which implies chaotic dynamics, also for the original system.

Future Directions

Ongoing research in dynamical systems includes nonlinear systems with parametric resonance, but there are a number of special features as these systems are non-autonomous. This complicates the dynamics from the outset. For instance a two degrees of freedom system with parametric resonance involves at least three frequencies, producing many possible resonances. The analysis of such higher dimensional systems with many more combination resonances, has begun recently, producing interesting limit sets and invariant manifolds. Also the analysis of PDEs with periodic coefficients will play a part in the near future. These lines of research are of great interest.

In the conservative case, the association with Hamiltonian systems, KAM theory etc. gives a natural approach. This has already produced important results. In real-life modeling, there will always be dissipation and it is important to include this effect. Preliminary results suggest that the impact of damping on for instance quasi-periodic systems, is quite dramatic. This certainly merits more research.

Finally, applications are needed to solve actual problems *and* to inspire new, theoretical research.

Acknowledgment

A number of improvements and clarifications were suggested by the editor, Giuseppe Gaeta. Additional references were obtained from Henk Broer and Fadi Dohnal.

Bibliography

Primary Literature

1. Arnold VI (1983) Geometrical methods in the theory of ordinary differential equations. Springer, New York
2. Banichuk NV, Bratus AS, Myshkis AD (1989) Stabilizing and destabilizing effects in nonconservative systems. PMM USSR 53(2):158–164
3. Bogoliubov NN, Mitropolskii Yu A (1961) Asymptotic methods in the theory of nonlinear oscillations. Gordon and Breach, New York
4. Bolotin VV (1963) Non-conservative problems of the theory of elastic stability. Pergamon Press, Oxford
5. Broer HW, Vegter G (1992) Bifurcational aspects of parametric resonance, vol 1. In: Jones CKRT, Kirchgraber U, Walther HO (eds) Expositions in dynamical systems. Springer, Berlin, pp 1–51
6. Broer HW, Levi M (1995) Geometrical aspects of stability theory for Hill's equation. Arch Rat Mech Anal 131:225–240
7. Broer HW, Simó C (1998) Hill's equation with quasi-periodic forcing: resonance tongues, instability pockets and global phenomena. Bol Soc Brasil Mat 29:253–293
8. Broer HW, Hoveijn I, Van Noort M (1998) A reversible bifurcation analysis of the inverted pendulum. Physica D 112:50–63
9. Broer HW, Hoveijn I, Van Noort M, Vegter G (1999) The inverted pendulum: a singularity theory approach. J Diff Eqs 157:120–149
10. Broer HW, Simó C (2000) Resonance tongues in Hill's equations: a geometric approach. J Differ Equ 166:290–327
11. Broer HW, Puig J, Simó C (2003) Resonance tongues and instability pockets in the quasi-periodic Hill-Schrödinger equation. Commun Math Phys 241:467–503
12. Broer HW, Hoveijn I, Van Noort M, Simó C, Vegter G (2005) The parametrically forced pendulum: a case study in $1\frac{1}{2}$ degree of freedom. J Dyn Diff Equ 16:897–947
13. Cicogna G, Gaeta G (1999) Symmetry and perturbation theory in nonlinear dynamics. Lecture Notes Physics, vol 57. Springer, Berlin
14. Fatimah S, Ruijgrok M (2002) Bifurcation in an autoparametric system in 1:1 internal resonance with parametric excitation. Int J Non-Linear Mech 37:297–308
15. Golubitsky M, Schaeffer D (1985) Singularities and groups in bifurcation theory. Springer, New York
16. Hale J (1963) Oscillation in nonlinear systems. McGraw-Hill, New York, 1963; Dover, New York, 1992
17. Hoveijn I, Ruijgrok M (1995) The stability of parametrically forced coupled oscillators in sum resonance. ZAMP 46:383–392
18. Iooss G, Adelmeyer M (1992) Topics in bifurcation theory. World Scientific, Singapore

19. Kuznetsov Yu A (2004) Elements of applied bifurcation theory, 3rd edn. Springer, New York
20. Krupa M (1997) Robust heteroclinic cycles. J Nonlinear Sci 7:129–176
21. Len JL, Rand RH (1988) Lie transforms applied to a non-linear parametric excitation problem. Int J Non-linear Mech 23:297–313
22. Levy DM, Keller JB (1963) Instability intervals of Hill's equation. Comm Pure Appl Math 16:469–476
23. Magnus W, Winkler S (1966) Hill's equation. Interscience-John Wiley, New York
24. McLaughlin JB (1981) Period-doubling bifurcations and chaotic motion for a parametrically forced pendulum. J Stat Phys 24:375–388
25. Ng L, Rand RH (2002) Bifurcations in a Mathieu equation with cubic nonlinearities. Chaos Solitons Fractals 14:173–181
26. Ng L, Rand RH (2003) Nonlinear effects on coexistence phenomenon in parametric excitation. Nonlinear Dyn 31:73–89
27. Pikovsky AS, Feudel U (1995) Characterizing strange nonchaotic attractors. Chaos 5:253–260
28. Ramani DV, Keith WL, Rand RH (2004) Perturbation solution for secondary bifurcation in the quadratically-damped Mathieu equation. Int J Non-linear Mech 39:491–502
29. Recktenwald G, Rand RH (2005) Coexistence phenomenon in autoparametric excitation of two degree of freedom systems. Int J Non-linear Mech 40:1160–1170
30. Roseau M (1966) Vibrations nonlinéaires et théorie de la stabilité. Springer, Berlin
31. Ruijgrok M (1995) Studies in parametric and autoparametric resonance. Thesis, Utrecht University, Utrecht
32. Ruijgrok M, Verhulst F (1996) Parametric and autoparametric resonance. Prog Nonlinear Differ Equ Their Appl 19:279–298
33. Ruijgrok M, Tondl A, Verhulst F (1993) Resonance in a rigid rotor with elastic support. ZAMM 73:255–263
34. Sanders JA, Verhulst F, Murdock J (2007) Averaging methods in nonlinear dynamical systems, rev edn. Appl Math Sci, vol 59. Springer, New York
35. Seyranian AP (2001) Resonance domains for the Hill equation with allowance for damping. Phys Dokl 46:41–44
36. Seyranian AP, Mailybaev AA (2003) Multiparameter stability theory with mechanical applications. Series A, vol 13. World Scientific, Singapore
37. Seyranian AA, Seyranian AP (2006) The stability of an inverted pendulum with a vibrating suspension point. J Appl Math Mech 70:754–761
38. Stoker JJ (1950) Nonlinear vibrations in mechanical and electrical systems. Interscience, New York, 1950; Wiley, New York, 1992
39. Strutt MJO (1932) Lamé-sche, Mathieu-sche und verwandte Funktionen. Springer, Berlin
40. Szemplinska-Stupnicka W (1990) The behaviour of nonlinear vibrating systems, vol 2. Kluwer, Dordrecht
41. Tondl A (1991) Quenching of self-excited vibrations. Elsevier, Amsterdam
42. Tondl A, Ruijgrok M, Verhulst F, Nabergoj R (2000) Autoparametric resonance in mechanical systems. Cambridge University Press, New York
43. Van der Pol B, Strutt MJO (1928) On the stability of the solutions of Mathieu's equation. Phil Mag Lond Edinb Dublin 7(5):18–38
44. Verhulst F (1996) Nonlinear differential equations and dynamical systems. Springer, New York
45. Verhulst F (2005) Invariant manifolds in dissipative dynamical systems. Acta Appl Math 87:229–244
46. Verhulst F (2005) Methods and applications of singular perturbations. Springer, New York
47. Wiggins S (1988) Global Bifurcation and Chaos. Appl Math Sci, vol 73. Springer, New York
48. Yakubovich VA, Starzhinskii VM (1975) Linear differential equations with periodic coefficients, vols 1 and 2. Wiley, New York
49. Zounes RS, Rand RH (1998) Transition curves for the quasi-periodic Mathieu equation. SIAM J Appl Math 58:1094–1115
50. Zounes RS, Rand RH (2002) Global behavior of a nonlinear quasi-periodic Mathieu equation. Nonlinear Dyn 27:87–105

Books and Reviews

Arnold VI (1977) Loss of stability of self-oscillation close to resonance and versal deformation of equivariant vector fields. Funct Anal Appl 11:85–92

Arscott FM (1964) Periodic differential equations. MacMillan, New York

Cartmell M (1990) Introduction to linear, parametric and nonlinear vibrations. Chapman and Hall, London

Dohnal F (2005) Damping of mechanical vibrations by parametric excitation. Ph D thesis, Vienna University of Technology

Dohnal F, Verhulst F (2008) Averaging in vibration suppression by parametric stiffness excitation. Nonlinear Dyn (accepted for publication)

Ecker H (2005) Suppression of self-excited vibrations in mechanical systems by parametric stiffness excitation. Fortschrittsberichte Simulation Bd 11. Argesim/Asim Verlag, Vienna

Fatimah S (2002) Bifurcations in dynamical systems with parametric excitation. Thesis, University of Utrecht

Hale J (1969) Ordinary differential equations. Wiley, New York

Kirillov ON (2007) Gyroscopic stabilization in the presence of nonconservative forces. Dokl Math 76:780–785; Orig Russian: (2007) Dokl Ak Nauk 416:451–456

Meixner J, Schäfke FW (1954) Mathieusche Funktionen und Sphäroidfunktionen. Springer, Berlin

Moon FC (1987) Chaotic vibrations: an introduction for applied scientists and engineers. Wiley, New York

Nayfeh AH, Mook DT (1979) Nonlinear Oscillations. Wiley Interscience, New York

Schmidt G (1975) Parametererregte Schwingungen. VEB Deutscher Verlag der Wissenschaften, Berlin

Schmidt G, Tondl A (1986) Non-linear vibrations. Akademie-Verlag, Berlin

Tondl A (1978) On the interaction between self-excited and parametric vibrations. In: Monographs and Memoranda, vol 25. National Res Inst Běchovice, Prague

Tondl A (1991) On the stability of a rotor system. Acta Technica CSAV 36:331–338

Tondl A (2003) Combination resonances and anti-resonances in systems parametrically excited by harmonic variation of linear damping coefficients. Acta Technica CSAV 48:239–248

Van der Burgh AHP, Hartono (2004) Rain-wind induced vibrations of a simple oscillator. Int J Non-Linear Mech 39:93–100

Van der Burgh AHP, Hartono, Abramian AK (2006) A new model for the study of rain-wind-induced vibrations of a simple oscillator. Int J Non-Linear Mech 41:345–358

Weinstein A, Keller JB (1985) Hill's equation with a large potential. SIAM J Appl Math 45:200–214
Weinstein A, Keller JB (1987) Asymptotic behaviour of stability regions for Hill's equation. SIAM J Appl Math 47:941–958
Wiggins S (1990) Introduction to applied nonlinear dynamical systems and chaos. Springer, New York

Perturbation of Equilibria in the Mathematical Theory of Evolution

Angel Sánchez[1,2]
[1] Grupo Interdisciplinar de Sistemas Complejos (GISC), Departamento de Matemáticas, Universidad Carlos III de Madrid, Madrid, Spain
[2] Instituto de Biocomputación y Física de Sistemas Complejos (BIFI), Universidad de Zaragoza, Zaragoza, Spain

Article Outline

Glossary
Definition of the Subject
Introduction
Evolution on a Fitness Landscape
Stability of Equilibria on a Fitness Landscape
Perturbation of Equilibria on a Fitness Landscape
Frequency Dependent Fitness: Game Theory
Equilibria in Evolutionary Game Theory
Perturbations of Equilibria in Evolutionary Game Theory
Future Directions
Bibliography

Glossary

Evolutionarily stable equilibria (ESS) An ESS is a set of frequencies of different types of individuals in a population that can not be invaded by the evolution of a single mutant. It is the evolutionary counterpart of a Nash equilibrium.

Fitness landscape A metaphorical description of fitness as a function of individual's genotypes or phenotypes in terms of a multivariable function that does not depend on any external influence.

Genetic locus The position of a gene on a chromosome. The different variants of the gene that can be found at the same locus are called alleles.

Nash equilibrium In classical game theory, a Nash equilibrium is a set of strategies, one for each player of the game, such that none of them can improve her benefits by unilateral changes of strategy.

Scale free network A graph or network such that the degrees of the nodes are taken from a power-law distribution. As a consequence, there is not a typical degree in the graph, i. e., there are no typical scales.

Small-world network A graph or network of N nodes such that the mean distance between nodes scales as $\log N$. It corresponds to the well-known "six degrees of separation" phenomenon.

Definition of the Subject

The importance of evolution can hardly be overstated. As the Jesuit priest Pierre Teilhard de Chardin put it,

> Evolution is a general postulate to which all theories, all hypotheses, all systems must hence forward bow and which they must satisfy in order to be thinkable and true. Evolution is a light which illuminates all facts, a trajectory which all lines of thought must follow – this is what evolution is.

Darwin's evolution theory is based on three fundamental principles: reproduction, mutation and selection, which describe how populations change over time and how new forms evolve out of old ones. Starting with W. F. R. Weldon, whom at the beginning of the 20th century realized that "the problem of animal evolution is essentially a statistical problem", and blooming in the 30's with Fisher, Haldane and Wright, numerous mathematical descriptions of the resulting evolutionary dynamics have been proposed, developed and studied. Deeply engraved in these frameworks are the mathematical concepts of equilibrium and stability, as descriptions of the observed population compositions and their lifetimes. Many results have been obtained regarding the stability of equilibria of evolutionary dynamics in idealized circumstances, such as infinite populations or global interactions. In the evolutionary context, stability is peculiar, in the sense that it is entangled with collective effects arising from the interaction of individuals. Therefore, perturbations of the idealized mathematical framework representing more realistic situations are of crucial importance to understand stability of equilibria.

Introduction

The idea of evolution is a simple one: Descent with modification acted upon by natural selection. Descent with modification means that we consider a population of replicators, entities capable of reproducing themselves, in which reproduction is not exact and allows for small differences between parents and offspring. Natural selection means

that different entities reproduce in different quantities because their abilities are also different: some are more resistant to external factors, some need less resources, some simply reproduce more... While in biology these replicators are, of course, living beings, the basic ingredients of evolution have by now trascended the realm of biology into the kingdom of objects such as computer codes (thus giving rise, e. g., to genetic algorithms).

The process of evolution is often described in short as "the survival of the fittest", meaning that those replicators that succeed and thrive are more fit than those which progressively disappear. This statement is not very appropriate, because evolution does not imply making organisms or entities more fit; it is simply a consequence of differential reproduction in the face of selection pressure. On the other hand, it leads to a tautology: The question of which are the fittest organisms is answered by saying that they are those that survive. It is then clear that a correct use of the concept of fitness is at the crux of any attempt to formalize mathematically evolution theory.

In this article we are going to discuss two manners to deal mathematically with the concept of fitness. The first and simplest one is to resort to a *fitness landscape*, whose basic feature is that the fitness of a given individual depends only on the individual's characteristics and not on external factors. We will present this approach in Sects. "Evolution on a Fitness Landscape" and "Stability of Equilibria on a Fitness Landscape" below, to subsequently discuss the effect of perturbations on the equilibria described by this picture in Sect. "Perturbation of Equilibria on a Fitness Landscape". To go beyond the fitness landscape picture one has to introduce frequency-dependent selection, i. e., to remove the independence of the fitness from external influences. In Sects. "Frequency Dependent Fitness: Game Theory" and "Equilibria in Evolutionary Game Theory" we consider the *evolutionary game theory* approach to this way to model evolution and, as before, analyze the perturbations of its equilibria in Sec "Perturbations of Equilibria in Evolutionary Game Theory". Sect. "Future Directions" summarizes the questions that remain open in this field.

Evolution on a Fitness Landscape

The metaphor of evolution on a "fitness landscape" reaches back at least to [26]: Drawing on the connection between fitness and adaptation, fitness is defined as the expected number of offspring of a given individual that reach adulthood, and thus represents a measure of its adaptation to the environment. In this context, fitness landscapes are used to visualize the relationship between genotypes

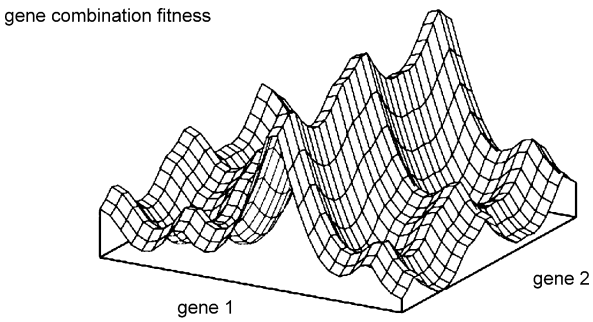

Perturbation of Equilibria in the Mathematical Theory of Evolution, Figure 1
Sketch of a fitness landscape

(or phenotypes) and reproductive success. It is assumed that every genotype has a well defined fitness, in the sense above, and that this fitness is the "height" of the landscape. Genotypes which are very similar are said to be "close" to each other, while those that are very different are "far" from each other. The two concepts of height and distance are sufficient to form the concept of a "landscape". The set of all possible genotypes, their degree of similarity, and their related fitness values is then called a fitness landscape.

Fitness landscapes are often conceived of as ranges of mountains. There exist local peaks (points from which all paths are downhill, i. e. to lower fitness) and valleys (regions from which most paths lead uphill). A fitness landscape with many local peaks surrounded by deep valleys is called rugged. If all genotypes have the same replication rate, on the other hand, a fitness landscape is said to be flat. A sketch of such a fitness landscape, showing the dependence of the fitness on two different "characteristics" or "genes", is shown in Fig. 1. Of course, the true fitness landscape would need a highly multidimensional space for its representation, as it would depend on all the characteristics of the organism, even those it still does not show. Therefore, the sketch is an extreme oversimplification, only to suggest the structure of a rugged fitness landscape.

Given the immense complexity of the genotype-fitness mapping, theoretical models have to make a variety of simplifying assumptions. Most models in biological literature focus on the effect of one or a few genetic loci on the fitness of individuals in a population, assuming that each of the considered loci can be occupied by a limited number of different alleles that have different effects on the fitness, and that the rest of the genome is part of the invariant environment. This approximation, the first attempt to obtain analytical results for changes in the gene pool of a population under the influence of inheritance,

selection and mutation is the pioneering work of Fisher, Haldane and Wright, who founded the field of population genetics. Their method of randomly drawing the genes of the daughter population from the pool of parent genes, with weights proportional to the fitness, proved to be very successful at calculating the evolution of allele frequencies from one generation to the next, or the chances of a new mutation to spread through a population, even taking into account various patterns of mating, dominance effects, nonlinear effects between different genes, etc. Population genetics has since then developed into a mature field with a sophisticated mathematical apparatus, and with wide-ranging applications.

Stability of Equilibria on a Fitness Landscape

The simplified pictures we have just described lead to a description in terms of dynamical systems, and therefore the stability of its equilibria can be studied by means of standard techniques. In principle, one can envisage the evolution of a population on a fitness landscape in the usual frame of particle dynamics on a potential. Every individual is a point in the space of phenotypes or genotypes, and evolves towards the maxima of the fitness; in the potential picture, the potential is given by minus the fitness. Furthermore, the dynamics is overdamped, i. e., there are no oscillations around the equilibria. Maxima of the fitness are therefore the equilibria of the evolutionary process. That this is so is a consequence of Fisher's theorem [3], whose original derivation is very general but quite complicated. Following [2] and [4], we prefer here to present two simpler situations: an asexual population, and a sexually reproducing population where the fitness is determined by a single gene with two alleles.

For an asexually reproducing population, the derivation of Fisher's theorem is straightforward: Let y_i be the number of genes i in the population, and y the total number of genes; then $p_i \equiv y_i/y$ is the frequency of genotype i in the population. If W_i is gene i's fitness, sticking to the interpretation of fitness in terms of offspring, the number of individuals carrying gene i in the next generation is $W_i y_i$ and, subsequently, the change in the frecuency p_i from one generation to the next is

$$\Delta p_i = p_i(W_i - \bar{W})/\bar{W} ,$$

leading to a change in mean fitness

$$\frac{\Delta \bar{W}}{\bar{W}} = \frac{\sum_i W_i \Delta p_i}{\bar{W}} = \frac{\sum_i p_i (W_i^2 - \bar{W}^2)}{\bar{W}^2} ,$$

which is proportional to the genetic variance in fitness. If the fitness changes from one generation to the next are small, this becomes an equation which states that the rate of change in fitness is identical to the genetic variance in fitness.

When reproduction is sexual, we note that p_i is the frequency of gene i. Assuming for simplicity that only two alleles are possible at the genetic locus of interest, the fitness of type 1 is $w_{11}p_1 + w_{12}p_2$, where w_{ij} is the fitness of an individual carrying alleles i and j, and hence the number of 1 genes will be $y_1(w_{11}p_1 + w_{12}p_2)$. We then have the differential equation

$$\dot{y}_1 = y_1(w_{11}p_1 + w_{12}p_2) . \tag{1}$$

It is enough then to differentiate the identity $\ln p_1 = \ln y_1 - \ln y$ and use a little algebra to show that y_1 obeys the *replicator equation*:

$$\dot{p}_1 = p_1(w_1 - \bar{w}) \tag{2}$$

Fisher's theorem then states that fitness increases along trajectories of this equation: Indeed, by noting that the average fitness is

$$\bar{w} = \sum_{i,j=1,2} w_{ij} p_i p_j , \tag{3}$$

differentiating and using the replicator equation we arrive at the final result

$$\dot{\bar{w}} = 2 \sum_i p_i (w_i - \bar{w})^2 . \tag{4}$$

Fisher's theorem thus means that an evolving population will typically climb uphill in the fitness landscape, by a series of small genetic changes, until a local optimum is reached. This is due to the fact that the average fitness of the population always increases, as we have just shown; hence the analogy with overdamped dynamics on a (inverted) potential function. Furthermore, because of this result, the population remains there, at the equilibrium point, because it cannot reduce its fitness. We then realize that all equilibria in a fitness landscape within the interpretation of fitness as the reproductive success are stable.

Perturbation of Equilibria on a Fitness Landscape

In order to understand the possible breakdowns of stability in the fitness landscape picture, one has to look carefully at the hypothesis of Fisher's theorem. We have not stated it in a formal manner, hence it is important to summarize here the main ones:

- Population is infinite.
- There are no mutations (i. e., the only source of every gene or species is reproduction).
- There is only one population (i. e., there are no population fluxes or migrations between separate groups).
- Fitness depends only on the individual's genotype and not on the other individuals.

It is then clear that, even if it is mathematically true, the applicability of Fisher's theorem is a completely different story, and as a consequence, the conclusion that maxima of the fitness landscape are stable may be wrong when discussing real systems. A detailed discussion of all these issues can be found in [2], and we refer the reader to her paper for a thorough discussion of all these factors. For our present purposes, namely to show that these perturbations can change the stability of the equilibria, it will suffice to present a few ideas about the case of finite population size. Afterwards, the rest of the paper will proceed along the idea of fitness depending on other individuals, giving up the paradigm of a fixed fitness landscape.

The subject of finite size populations is the subject of fluctuations and its main consequences, genetic drift and stochastic escape. Regarding the first concept, as compared to natural selection, i. e., to the tendency of beneficial alleles to become more common over time (and detrimental ones less common), genetic drift is the fundamental tendency of any allele to vary randomly in frequency over time due to statistical variation alone, so long as it does not comprise all or none of the distribution. In other words, even when individuals face the same odds, they will differ in their success. A rare succession of chance events can thus bring a trait to predominance, causing a population or species to evolve (in fact, this idea is at the core of the neutral theory of evolution, first proposed by [10]). On the other hand, stochastic escape refers to the situation in which a population of individuals placed at a maximum of the fitness landscape may leave this maximum due to fluctuations. Obviously, both genetic drift and stochastic escape affect the stability of the maxima as predicted by Fisher's theorem.

One consequence of finite population sizes and fluctuations in the composition of a population is that genes get lost from the gene pool. If there is no new genetic input through mutation or migration, the genetic variability within a population decreases with time. After sufficiently many generations, all individuals will carry the same allele of a given gene. This allele is said to have become fixed. In the absence of selection, the probability that a given allele will become fixed is proportional to the number of copies in the initial population. Thus, if a new mutant arises that has no selective advantage or disadvantage, this mutant will spread through the entire population with a probability $1/M$, M being the population size. If the individuals of the population are diploid, each carries two sets of genes, and M must be taken as the number of sets of genes, i. e., as twice the population size. On the other hand, it can be shown [2] that the probability that a mutant that conveys a small fitness increase by a factor $1 + s$ has as probability of the order s to spread through a population. In populations of sizes much smaller than $1/s$, this selective advantage is not felt, because mutations that carry no advantage become fixed at a similar rate. In the same manner, a mutation that decreases the fitness of its carrier by a factor $1 - s$, is not felt in a population much smaller than $1/s$. An interesting consequence of these results is that the rate of neutral (or effectively neutral) substitutions is independent of the population size. The reason is that the probability that a new mutant is generated in the population is proportional to M, while its probability of becoming fixed is $1/M$.

Frequency Dependent Fitness: Game Theory

In the preceding sections we have considered the case when the fitness depends on the genotype, but is independent of the composition of the population, i. e., the presence of inviduals of the same genotype or of other genotypes does not change the fitness of the focal one. This assumption, that allows for an intuitive picture in terms of a fitness landscape, is clearly an over-simplification, as was already mentioned above. For instance, consider an homogeneous population in a closed environment. The population will grow at a pace given by the fitness of its individuals until it eventually exhausts the available resources or even physically fills the environment. Therefore, even if the individuals are all equal, their fitness will not be the same if there are only a few of them or if there are very many. Another trivial example is the effect on the fitness of the presence or absence of predators of the species of interest; clearly, predators will reduce the fitness (understood as above in a reproductive sense) of their prey.

Therefore, individuals will evolve subject not only to external influences but also to their mutual competition, both intra-specific and inter-specific. This leads us to consider frequency-dependent selection, which can be described by very many, different theoretical approaches. These include game theory as well as discrete and continuous genetic models, and the concepts of kin selection, group selection, and sexual selection. Among the possible dynamical patterns arising, there are single fixed points, lines of fixed points, runaway, limit cycles, and chaos. A re-

view of all these descriptions, whose use to model evolution depends on the specific issues one is interested in, is clearly far beyond the scope of this article and, hence, we have focused on evolutionary game theory as a particularly suited case study to show the effects of perturbations on equilibria.

Before going into the study of evolutionary game theory, we need to summarize briefly a few key concepts about its originating theory, namely (classical) game theory. Pioneered in the early XIX century by the economist Cournot, game theory was introduced by the brilliant, multi-faceted mathematician John von Neumann in 1928, and it was first presented as a specific subject in von Neumann's book (with Oskar Morgenstern) *Theory of Games and Economic Behavior* in 1944. Since then, game theory has been used to model strategic situations, i. e., situations in which actors or agents follow different strategies (meaning that they choose among different possible actions or behaviors) to maximize their benefit, usually referred to as *payoff*. These arise in very many different contexts, from biology and psychology to philosophy through politics, economics or computer science.

The central concept of game theory is the *Nash equilibrium* [14], introduced by the mathematician John Nash in 1955, awarded with a Nobel Prize in Economics almost forty years later for this work. A set of strategies, one for each participant in the game, is a Nash equilibrium if every strategy is the best response (in terms of maximizing the player's payoff) to the subset of the strategies of the rest of the players. In this case, if all players use strategies belonging to a Nash equilibrium, none of them will have any incentive to change her behavior. In this situation we indeed have the equivalent of the traditional concept of equilibrium in dynamical systems: players keep playing the same strategy as, given the behavior of the others, they follow the optimal strategy (note that this does not mean the strategy is optimal in absolute terms: it is only optimal in view of the actions of the rest).

Equilibria in Evolutionary Game Theory

In the seventies, game theory, which as proposed by von Neumann and Nash was to be used to understand economic behavior, entered the realm of biology through the pioneering work of John Maynard-Smith and George Price [12], who introduced the evolutionary version of the theory. The key contribution of their work was a new interpretation of the general framework of game theory in terms of populations instead of individual players. While traditional game theoretical players behaved following some strategy and could change it to improve their performance, in the picture of Maynard-Smith and Price individuals had a fixed strategy, determined by their genotype, and different strategies were represented by subpopulations of individuals. In this representation, changes of strategies correspond to the replacement of the individuals by their offspring, possibly with mutations. Payoffs obtained by individuals in the game are accordingly understood as fitness, the reproductive rate that governs how the replacement occurs.

There is a large degree of arbitrariness as to the evolutionary dynamics of the populations. All we have said so far is that fitness, obtained through the game, determines the composition of the population at the next time step (or instant, if we think of continuous time). Probably the most popular choice (but by no means the only one, see [18] for different evolutionary proposals and their relationships, see also [9] for other dynamics) is to use the replicator equation we have previously found to describe the evolution of the frequency y_i of strategists of type i:

$$\dot{y}_i = y_i(w_i - \bar{w}) \,. \tag{5}$$

It is important to note that the steps leading to the derivation of this equation are the same as above, and therefore for it to be applicable in principle one must keep in mind the same hypotheses. The difference is that now the fitness is not a constant but rather it is determined by a game, which enters the equation in the following way.

Let us call $\mathbf{A}$ the payoff matrix of the game (for simplicity, we will consider only symmetric games), whose entries a_{ij} are the payoffs to an individual using strategy i facing another using strategy j. Assuming the frequencies $y_i(t)$ are differentiable functions, if individuals meet randomly and then engage in the game, and this takes place very many (infinite) times, then $(\mathbf{Ay})_i$ is the expected payoff for type i individuals in a population described by the vector $\mathbf{y}$, whose components are the frequencies of each type. By the same token, the average payoff in the population is $\bar{w} = \mathbf{y}^{\mathrm{T}}\mathbf{Ay}$, so substituting in (5) we are left with

$$\dot{y}_i = y_i\big[(\mathbf{Ay})_i - \mathbf{y}^{\mathrm{T}}\mathbf{Ay}\big] \,, \tag{6}$$

where we now see explicitly how the game affects the evolution. Nevertheless, it is also clear that this rule is arbitrary, and there are many other options one can use to postulate how the population evolves. We will come back to this issue when considering perturbations of the equilibria.

If Nash equilibrium is the key concept in game theory, evolutionarily stable strategy is the relevant one in its evolutionary counterpart. [11] defined evolutionarily stable strategy (ESS) as a strategy such that, if every individual

in the population uses it, no other (mutant) strategy could invade the population by natural selection. It is trivial to show that, in terms of the payoffs of the game, for strategy i to be an ESS, one of the following two conditions must hold:

$$a_{ii} > a_{ij} \;\; \forall j \neq i \,, \quad \text{or} \tag{7}$$

$$a_{ii} = a_{ij} \;\; \text{for some } j \quad \text{and} \quad a_{ij} > a_{jj} \,. \tag{8}$$

If the first condition is fulfilled, we speak of *strict* ESS. It is important to realize that this concept is absolutely general and, in particular, it does not depend on the evolutionary dynamics of choice (in so far as it favors the strategies that receive the best payoffs).

Of course, the two concepts, Nash equilibrium and ESS, are related. This is in fact one of the reasons why evolutionary game theory ended up appealing to the economists, who faced the question as to how individuals ever get to play the Nash equilibrium strategies: They now had a dynamical way that might precisely describe that process and, furthermore, to decide which Nash equilibrium was selected if there were more than one. To show the connection, one must decide on a dynamical rule, for which we will stay within the framework of the replicator dynamics. For this specific evolutionary dynamics, it can be rigorously shown that (see, e. g., [7,9])

1. if y_0 is a Nash equilibrium, it is a rest point (a zero of the rhs of (5);
2. if y_0 is a strict Nash equilibrium, it is asymptotically stable;
3. if y_0 is a rest point and is the limit of an interior orbit for $t \to \infty$, then it is a Nash equilibrium; and
4. if y_0 is a stable rest point, it is a Nash equilibrium.

This means that there indeed is a relationship between Nash equilibria and ESS, but more subtle that could appear at first. Probably, the most important non trivial aspect of this result is that not all ESS are Nash equilibria, as stability is required in addition.

Perturbations of Equilibria in Evolutionary Game Theory

The evolutionary viewpoint on game theory allows to study Nash equilibria/ESS within the standard framework of dynamical systems theory, by using the concepts of stability, asymptotical stability, global stability and related notions. In fact, one can do more than that: the problem of invasion by a mutant, the biological basis of the ESS concept of Maynard-Smith, can always be formulated in terms of a dynamical coupling of the mutant and the incumbent species and hence studied in terms of the stability of a rest point of a dynamical system. In principle, the same idea can be generalized to simultaneous invasion by more than one mutant and, although the problem may be technically much more difficult, the basic procedure remains the same.

As we did with the fitness landscape concept, when considering perturbations of equilibria, our interest goes beyond this traditional stability ideas, and once again, we need to focus on the deviations from the framework that allows to derive the replicator equation. There are a number of such deviations. The simplest ones are the inclusion of mutations or migrations, leading to the so-called replicator-mutator equation [18], that can be subsequently studied as a dynamical system. Other deviations affect much more, and in a way more difficult to aprehend, to the evolutionary dynamics and its equilibria, such as considering finite size populations, alternative learning/reproduction dynamics, or the non-universality of interactions among individuals. In this section we will choose this last point as our specific example, and analyze the consequences of relaxing the hypothesis that every player plays every other one. This hypothesis is needed to substitute the payoff earned by a player by what she would have obtained facing the average player of the population (an approach that has been traditionally used in physics under the name of mean-field approximation). However, interactions may not be universal after all, either because of spatial or temporal limitations. We will address both in what follows. The reader is referred to [15] for discussions of other perturbations.

Spatial Perturbations

One of the reasons why maybe not all individuals interact with all others is that they could not possibly meet. In biological terms this may occur because the population is very sparsely distributed and every individual meets only a few others within its living range, or else in a very numerous population where it is impossible in practice to meet all individuals. In social terms, an alternative view is the existence of a social network or network or contacts that prescribes who interacts with whom.

This idea was first introduced in a famous paper by [16] on the evolutionary dynamics of the Prisoner's Dilemma on a square lattice. In the Prisoner's Dilemma two players simultaneously decide cooperate or to defect. Cooperation results in a benefit b to the recipient but incurs costs c to the donor ($b > c > 0$). Thus, mutual cooperation pays a net benefit of $R = b - c$ whereas mutual

defection results in $P = 0$. However, unilateral defection yields the highest payoff $T = b$ and the cooperator has to bear the costs $S = -c$. It immediately follows that it is best to defect regardless of the opponents decision. For this reason defection is the evolutionarily stable strategy even though all individuals would be better of if all would cooperate (mutual cooperation is better than mutual defection because $R > P$). Mutual defection is also the only Nash equilibrium of the game. All this translates into the following payoff matrix:

$$\begin{array}{cc} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \begin{pmatrix} R & S \\ T & P \end{pmatrix} \end{array} . \tag{9}$$

For this matrix to correspond to a Prisoner's Dilemma game, the ordering of payoffs must be $T > R > P > S$. As we will see below, other orderings define different games.

What Nowak and May did was to set the individuals on the nodes of a square lattice, where they played the game only with their nearest and next-nearest neighbors (Moore neighborhood). They ran simulations with the following dynamics: every individual played the game with her neighbors and collected the corresponding payoff, and afterwards she updated her strategy by imitating that of her most successful (in terms of payoff) neighbor. In their simulations, Nowak and May found that if they started with a population with a majority of cooperators, a large fraction of them remained cooperators instead of changing their behavior towards the ESS, namely, defection. The reason is that the structured interaction allowed cooperators to do well and avoid exploitation by defectors by grouping into clusters, inside which they interacted mostly with other cooperators, whereas defectors at the boundaries of those clusters, interacting mostly with other defectors, did not fare as well and therefore did not induce cooperators to defect.

This result is partly due to the imitation dynamics, which, if postulated to rule the evolution of a population of individuals that interact with all others, does not lead to the replicator equation. As we mentioned in the preceding section, the update rule for the strategies is arbitrary and can be chosen at will (preferrably with some specific modelling in my mind). To reproduce the behavior of the replicator equation, a probabilistic rule has to be used [4], and with this rule the equilibrium is not changed and the population evolves to full defection. However, [16] opened the door to a number of more detailed studies that considered also different dynamical rules including the one corresponding to the replicator dynamics in which it was shown that the structure of a population definitely had a strong influence on the game equilibria.

One such study, perhaps the most systematic to date, was carried out by [5], who compared the equilibrium frequencies of cooperators and defectors in populations with and without spatial structuring (square lattices), finding two important results: First, including spatial extension has indeed significant effects on the equilibrium frequencies of cooperators and defectors. In some parameter regions spatial extension promotes cooperative behavior while inhibiting it in others; and, second, differences in the initial frequencies of cooperators are readily leveled out and hardly affect the equilibrium frequencies except for $T < 1$, $S < 0$. This choice is not the Prisoner's Dilemma anymore, it corresponds to the so-called Stag Hunt game [22], and in the replicator dynamics is a bistable system where the initial frequencies determine the long term behavior, a feature that is generally preserved for the spatial setting. Of course, [5] also observed that the size of the neighborhood obviously affects the spreading speed of successful strategies. Interestingly, although the message seems to be that the strategy of cooperation is favored over the strategy of defection by the presence of a spatial structure, this is not always the case, and in games where the equilibrium population has a certain percentage of both types of strategists, the network of interactions makes the frequency of cooperators decrease [6]. Therefore, the effect of this perturbation is not all trivial and needs careful consideration.

All the results discussed so far correspond to a square lattice as substrate to define the interaction pattern, but this is certainly a highly idealized setup that can hardly correspond to any real, natural system. Recent studies have shown that the results also depend on the type of graph or network used. Thus, [21] have shown that in more realistic, heterogeneous populations, modeled by random graphs of different types, the sustainability of cooperation (implying the departure of the equilibrium predicted by the replicator equation) is simpler to achieve than in homogeneous populations, a result which is valid irrespective of the dilemma or game adopted as a metaphor of cooperation. Therefore, heterogeneity constitutes a powerful mechanism for the emergence of cooperation (and consequently an important perturbation of the dynamics), since even for mildly heterogeneous populations it leads to sizeable effects in the evolution of cooperation. The overall enhancement of cooperation obtained on single-scale and scale-free graphs [1] may be understood as resulting from the interplay of two mechanisms: The existence of many long-range connections in random and small-world networks [25], which precludes the formation of compact clusters of cooperators, and the heterogeneity exhibited by these networks, which opens a new

route for cooperation to emerge and contributes to enhance cooperation (which increases with heterogeneity), counteracting the previous effect. This result depends also on the intricate ties between individuals, even for the same class of graphs, features absent in the replicator dynamics.

We have thus seen that removing the hypothesis of universal interaction is a strong perturbation to equilibrium as understood from the replicator dynamics. There have been many other studies following those we can possibly review here; a recent, very comprehensive summary can be found in [23]. It must be realized that, intermixed with the effect of the network of interactions, the different dynamical rules one can think of have also a relevant influence on the equilibria. While we have not considered them as perturbations of the replicator dynamics here, because they do not behave as described by that equation even in the presence of universal interactions, it must be kept in mind that they do affect the equilibria and therefore they must be properly specified in any serious study of evolutionary game theory.

Time Scales

Let us now come back to the situation in which there is no spatial structure, and every agent can in principle play the game against every other one. Afterwards, reproduction proceeds according to the payoff earned during the game stage. As we have already said, for large populations, this amounts to saying that every player gains the payoff of the game averaged in the current distribution of strategies. In terms of time scales, such an evolution corresponds to a regime in which reproduction-selection events take place at a much slower rate than the interaction between agents. However, these two time scales need not be different in general and, in fact, for many specific applications they can arguably be of the same order [7].

To study different rates of selection we can consider the following new dynamics [19,20,24]. There is a population with N players. A pair of individuals is randomly selected for playing, earning each one an amount of fitness according to the rules of the game. This game act is repeated s times, choosing a new random pair of players in each occasion.

Afterwards, selection takes place. Following [17], we have chosen Moran dynamics [13] as the most suitable to model selection in a finite population. This is necessary because the replicator equation is posed for continuous values of the populations and here we need to consider discrete values, i. e., individual by individual, in order to pinpoint the existing time scales. However, it can be shown [19] that the equilibria of Moran dynamics are the same as those of the replicator equation, and in fact, the whole evolution is the same except for a rescaling of time. Moran dynamics is defined as follows: One individual among the population of N players is chosen for reproduction proportionally to its fitness, and its offspring replaces a randomly chosen individual. As the fitness of all players is set to zero before the following round of s games, the overall result is that all players have been replaced by one descendant, but the player selected for reproduction has had a reproductive advantage of doubling its offspring a the expense of the randomly selected player. It is worth noting that the population size N is therefore constant along the evolution.

The parameter s controls the time scales of the model, i. e. reflects the relation between the rate of selection and the rate of interaction. For $s \ll N$ selection is very fast and very few individuals interact between reproduction events. Higher values of s represent proportionally slower rates of selection. Thus, when $s \gg N$ selection is very slow and population is effectively well-mixed and we recover the behavior predicted by the replicator equation.

The most striking example of the influence of the selection is the so-called Harmony game, a trivial one that has henceforth never been studied, and that is determined by $R > T > S > P$. The *only* Nash equilibrium or ESS of this game is mutual cooperation, as it is obvious from the payoffs: The best option for both players is to cooperate, which yields the maximum payoff for each one. Let us denote by $0 \leq n \leq N$ the number of cooperators present in the population, and look at the probability x_n of ending up in state $n = N$ (i. e., all players cooperate) when starting in state $n < N$. For $s = 1$ and $s \to \infty$, an exact, analytical expression for x_n can be obtained [19]. For arbitrary values of s, such a closed form cannot be found; however, it is possible to carry out a combinatorial analysis of the possible combinations of rounds and evaluate, numerically but exactly, x_n.

In Fig. 2a, we show that the rationally expected outcome of a population consisting entirely of cooperators is not achieved for small and moderate values of s, our selection rate parameter. For the smallest values, only when starting from a population largely formed by cooperators there is some chance to reach full cooperation; most of the times, defectors will eventually prevail and invade the whole population. This counterintuitive result may arise even for values of s comparable to the population size, by choosing suitable payoffs. Interestingly, the main result that defection is selected for small values of s does not depend on the population size N; only details such as the shape of the curves (cf. Fig. 2b) are modified by N.

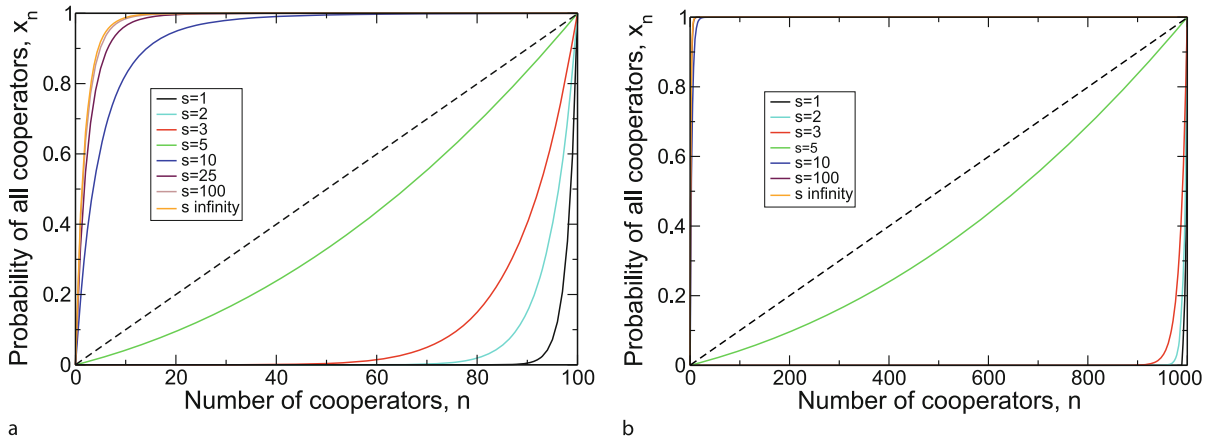

Perturbation of Equilibria in the Mathematical Theory of Evolution, Figure 2

Probability of ending up with all cooperators starting from n cooperators, x_n for different values of s. a For the smallest values of s, full cooperation is only reached if almost all agents are initially cooperators. Values of s of the order of 10 show a behavior much more favorable to cooperators. In this plot, the population size is $N = 100$. b Taking a population of $N = 1000$, we observe that the range of values of s for which defectors are selected does not depend on the population size, only the shape of the curves does. Parameter choices are: Number of games between reproduction events, s, as indicated in the plots; payoffs for the Harmony game, $R = 11$, $S = 2$, $T = 10$, $P = 1$. The dashed line corresponds to a probability to reach full cooperation equal to the initial fraction of cooperators and is shown for reference

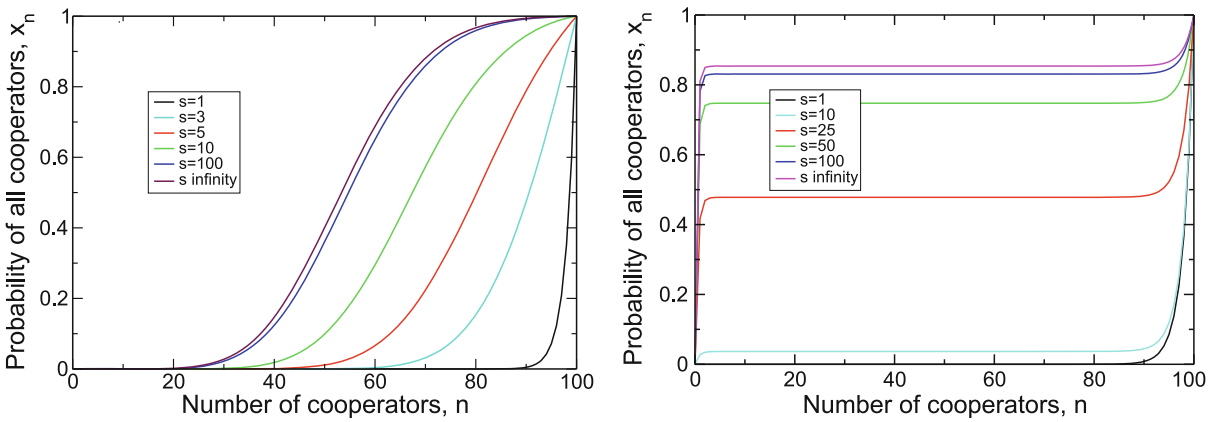

Perturbation of Equilibria in the Mathematical Theory of Evolution, Figure 3

***Left*: Same as Fig. 2 for the Stag-Hunt game. The probability of ending up with all cooperators starting from n cooperators, x_n, is very low when s is small, and as s increases it tends to a quasi-symmetric distribution around 1/2. Payoffs for the Stag-Hunt game, $R = 6$, $S = 1$, $T = 5$, $P = 2$. *Right*: Same as Fig. 2 for the Snowdrift game. The probability of ending up with all cooperators starting from n cooperators is almost independent of n except for very small or very large values. Small s values lead once again to selection of defectors, whereas cooperators prevail more often as s increases. Payoffs for the Snowdrift game, $R = 1$, $S = 0.35$, $T = 1.65$, $P = 0$. Other parameter choices are: Population, $N = 100$; number of games between reproduction events, s, as indicated in the plot**

In the preceding paragraph we have chosen the Harmony game to discuss the effect of the rate of selection, but this effect is very general and appears in many other games. Consider the example of the already mentioned Stag-Hunt game [22], with payoffs $R > T > P > S$. This is the paradigmatic situation of game with two Nash equilibria in pure strategies, mutual cooperation and mutual defection, each one with its own basis of attraction in the replicator equation framework (in general, which of these equilibria is selected has been the subject of a long argument in the past, and rationales for both of them can be provided [22]). As Fig. 3 (left) shows, simulation results for finite s are largely different from the curve obtained for $s \to \infty$: Indeed, we see that for $s = 1$, all agents become defectors except for initial densities close to 1. Even for values of s as large as N evolution will more

likely lead to a population entirely consisting of defectors.

Yet another example of the importance of the selection rate is provided by the Snowdrift game, with payoffs $T > R > S > P$. Figure. 3 (right) shows that for small values of s defectors are selected for almost any initial fraction of cooperators. When s increases, we observe an intermediate regime where both full cooperation and full defection have nonzero probability, which, interestingly, is almost independent of the initial population. And, for large enough s, full cooperation is almost always achieved.

With these examples, it is clear that considering independent interaction and selection time scales may lead to highly non-trivial, counter-intuitive results. [19] showed that out of the 12 possible different games with two players, six were severely changed by the introduction of the game time scale, whereas the other six remained with the same equilibrium structure. Of course, the extent of the modifications of the replicator dynamics picture depends on the structure of the unperturbed phase space. Thus, rapid selection perturbations show up in changes of the asymptotically selected equilibria, i. e., of the asymptotically stable one, to changes of the basins of attraction of equilibria, or to suppression of long-lived metastable equilibria. As in the case of spatial perturbations, we are thus faced with a most relevant influence on the equilibria of the evolutionary game.

Future Directions

As we have seen in this necessarily short excursion, the simplest mathematical models of evolution allow for a detailed, analytical study of their equilibria (which are supposed to represent stable states of populations) but, when leaving aside some of the hypothesis involved in the derivation of those simple models, the structure of equilibria may be seriously modified and highly counter-intuitive results may arise. We have not attempted to cover all possible perturbations but we believe we have provided evidence enough that their effect is certainly very relevant. When trying to bridge the gap between simple models and reality, other hypotheses will be broken, maybe more than one simultaneously, and subsequently the equilibria will be severely affected.

In the future, we believe that this line of research will undergo very interesting developments, particularly in the framework of evolutionary game theory, as the fitness landscape picture seems to be rather well understood and, on the other hand, is felt to be a much too simple model. In the case of evolutionary games, while some of the results we have collected here are analytical, there are many others which are only numerical, in particular when perturbations depending on space (evolutionary game theory on graphs) are considered. A lot of research needs to be devoted in the next few years to understand this problem analytically, more so because there is now a "zoo" of results that are even hard to classify or interpret within a common basis. This will probably require a combined effort from different mathematical disciplines, ranging from discrete mathematics to dynamical systems through, of course, graph theory.

On the other hand, our examples have consisted of two-player, two-strategy, symmetric games, i. e., the simplest possible scenario. There are practically no results about games with more than two strategies or, even worse, with more than two players. In fact, even the classification of the phase portraits within the replicator equations for those situations is far from understood, the more so the higher the dimensionality of the problem. Much remains to be done in this direction. Asymmetrical games are a different story; for those, the replicator equation is not (6) anymore but rather one has to take into account that payoffs when playing as player 1 or player 2 are not the same, and the corresponding equation is more complicated. Again, this line of research is still in its infancy and awaiting for dedicated work.

Finally, a very interesting direction is the application of the results to problems in social or biological contexts. The evolutionary game theory community has been relying strongly on the predictions from the replicator equation which we now see may not agree with reality or at least with what occurs when some of its hypotheses are not fulfilled. This has led to a number of conundrums, particularly prominent among those being the problem of the emergence of cooperation. A recent study [24] have shown that, in a scenario described by the so-called Ultimatum game, taking into consideration the possible separation of time scales leads to results compatible with the experimental observations on human subjects, observations that the replicator equation is not able to reproduce. We envisage that similar results will be ubiquitous when trying to match the predictions of the replicator equation with actual systems or problems. Understading the effect of perturbations in a comprehensive manner will then be the key to the fruitful development of the theory as a "natural" or "physical" one.

Bibliography

Primary Literature

1. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512

2. Drossel B (2001) Biological evolution and statistical physics. Adv Phys 50:209–295
3. Fisher RA (1958) The Genetical Theory of Natural Selection, 2nd edn. Dover, New York
4. Gintis H (2000) Game Theory Evolving. Princeton University, Princeton
5. Hauert C (2002) Effects of space in 2 × 2 games. Int J Bifurc Chaos 12:1531–1548
6. Hauert C, Doebeli M (2004) Spatial structure often inhibits the evolution of cooperation in the snowdrift game. Nature 428:643–646
7. Hendry AP, Kinnison MT (1999) The pace of modern life: Measuring rates of contemporary microevolution. Evolution 53:1637–1653
8. Hofbauer J, Sigmund K (1998) Evolutionary Games and Population Dynamics. Cambridge University, Cambridge
9. Hofbauer J, Sigmund K (2003) Evolutionary game dynamics. Bull Am Math Soc 40:479–519
10. Kimura M (1983) The Neutral Theory of Molecular Evolution. Cambridge University, Cambridge
11. Maynard-Smith J (1982) Evolution and the Theory of Games. Cambridge University, Cambridge
12. Maynard-Smith J, Price GR (1973) The logic of animal conflict. Nature 246:15–18
13. Moran PAP (1962) The Statistical Processes of Evolutionary Theory. Clarendon, Oxford
14. Nash JF (1950) Equilibrium points in n-person games. Proc Natl Acad Sci USA 36:48–49
15. Nowak MA (2006) Evolutionary Dynamics: Exploring the Equations of Life. Harvard Univ Press, Harvard
16. Nowak MA, May R (1992) Evolutionary games and spatial chaos. Nature 359:826–829
17. Nowak MA, Sasaki A, Taylor C, Fudenberg D (2004) Emergence of cooperation and evolutionary stability in finite populations. Nature 428:646–650
18. Page K, Nowak MA (2002) A unified evolutionary dynamics. J Theor Biol 219:93–98
19. Roca CP, Cuesta JA, Sánchez A (2006) Time scales in evolutionary dynamics. Phys Rev Lett 97: art. no. 158701
20. Roca CP, Cuesta JA, Sánchez A (2007) The importance of selection rate in the evolution of cooperation. Eur Phys J Special Topics 143:51–58
21. Santos FC, Pacheco JM, Lenaerts T (2006) Evolutionary dynamics of social dilemmas in structured heterogeneous populations. Proc Natl Acad Sci USA 103:3490–3494
22. Skyrms B (2003) The Stag Hunt and the Evolution of Social Structure. Cambridge University, Cambridge
23. Szabó G, Fáth G (2007) Evolutionary games on graphs. Phys Rep 446:97–216
24. Sánchez A, Cuesta JA (2005) Altruism may arise by individual selection. J Theor Biol 235:233–240
25. Watts DJ, Strogatz SH (1998) Collective dynamics of "Small World" Networks. Nature 393:440–442
26. Wright S (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. Proc 6th Int Cong Genet 1:356–366

Books and Reviews

Nowak M, Sigmund K (2004) Evolutionary dynamics of biological games. Science 303:793–799

Taylor PD, Jonker L (1978) Evolutionarily stable strategies and game dynamics. J Math Biosci 40:145–156

von Neumann J, Morgenstern O (1944) Theory of Games and Economic Behavior. Princeton University Press, Princeton

Perturbation of Systems with Nilpotent Real Part

Todor Gramchev
Dipartimento di Matematica e Informatica, Università di Cagliari, Cagliari, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Complex and Real Jordan Canonical Forms
Nilpotent Perturbation and Formal Normal Forms of Vector Fields and Maps Near a Fixed Point
Loss of Gevrey Regularity in Siegel Domains in the Presence of Jordan Blocks
First-Order Singular Partial Differential Equations
Normal Forms for Real Commuting Vector Fields with Linear Parts Admitting Nontrivial Jordan Blocks
Analytic Maps near a Fixed Point in the Presence of Jordan Blocks
Weakly Hyperbolic Systems and Nilpotent Perturbations
Bibliography

Glossary

Perturbation Typically, one starts with an "initial" system S_0, which is usually simple and/or well understood. We perturb the system by adding a (small) perturbation R so that the new object becomes $S_0 + R$. In our context the typical examples for S_0 will be systems of linear ordinary differential equations with constant coefficients in $\mathbb{R}^n$ or the associated linear vector fields.

Nilpotent linear transformation Let $A\colon \mathbb{K}^n \mapsto \mathbb{K}^n$ be a linear map, where $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$. We call A nilpotent if there exists a positive integer r such that the rth iteration A^r become the zero map, in short $A^r = 0$.

Gevrey spaces Let Ω be an open domain in $\mathbb{R}^n$ and let $\sigma \geq 1$. The Gevrey space $G^\sigma(\Omega)$ stands for the set of all functions $f \in C^\infty(\Omega)$ such that for every compact subset $K \subset\subset \Omega$ one can find $C = C_{K,f} > 0$ such that

$$\sup_{x\in K} |\partial_x^\alpha f(x)| \leq C^{|\alpha|+1}\alpha!^\sigma \tag{1}$$

for all $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{Z}_+^n$, $\alpha! = \alpha_1! \ldots \alpha_n!$, $|\alpha| := \alpha_1 + \ldots + \alpha_n$. If $\sigma = 1$ we recapture the space of real analytic functions in Ω while the scale $G^\sigma(\Omega)$, $\sigma > 1$, serves as an intermediate space between the real analytic functions and the set of all C^∞ functions in Ω. By the Stirling formula one may replace $\alpha!^\sigma$ by $|\alpha|!^\sigma$, $|\alpha|^{\sigma|\alpha|}$ or $\Gamma(\sigma|\alpha|)$, where $\Gamma(z)$ stands for the Euler Gamma function cf. the book of Rodino [41] for more details on the Gevrey spaces.

One associates also Gevrey index to formal power series, namely, given a (formal) power series

$$f(x) = \sum_\alpha f_\alpha x^\alpha$$

this is in the formal Gevrey space $G_f^\sigma(\mathbb{K}^n)$ if there exist $C > 0$ and $R > 0$ such that

$$|f_\alpha| \leq C^{|\alpha|+1} |\alpha|!^{\sigma-1} \tag{2}$$

for all $\alpha \in \mathbb{Z}_+^n$.

In fact, one can find in the literature another definition of the formal Gevrey spaces G_f^τ of index τ, namely replacing $\sigma - 1$ by τ (see e. g. Ramis [40]).

Definition of the Subject

The main goal of this article is to dwell upon the influence of the presence (explicit and/or hidden) of nontrivial real nilpotent perturbations appearing in problems in Dynamical Systems, Partial Differential Equations and Mathematical Physics. Under the term nilpotent perturbation we will mean, broadly speaking, a classical linear algebra type setting: we start with an object (vector field or map near a fixed point, first-order singular partial differential equations, system of evolution partial differential equations) whose "linear part" A is semisimple (diagonalizable) and we add a (small) nilpotent part N. The problems of interest might be summarized as follows: are the "relevant properties" (in suitable functional framework) of the initial "object" stable under the perturbation N. If not, to classify, if possible, the novel features of the perturbed systems.

Broadly speaking, the cases when the instabilities occur are rare, they form some kind of exceptional sets. However, they appear in important problems (both in mathematical and physical contests) when degeneracies (bifurcations) occur.

Introduction

We will focus our attention on topics where the presence of nontrivial Jordan blocks in the linear parts changes the properties of the original systems (i. e., instabilities occurs unless additional restrictions are imposed):

(i) Convergence/divergence issues for the normal form theory of vector fields and maps near a singular (fixed) point in the framework of spaces of analytic functions and Gevrey classes.
(ii) (Non)solvability for singular partial differential equations near a singular point.
(iii) Cauchy problems for hyperbolic systems of partial differential equations with multiple characteristics.

Some basic features of the normal form theory for vector fields near a point will be recalled with an emphasis on the difficulties appearing in the presence of nontrivial Jordan blocks in the classification and computational aspects of the normal forms. For more details and various aspects of perturbation theory in Dynamics we refer to other articles in the Perturbation Theory Section of this Encyclopedia: cf. Bambusi ▶ Perturbation Theory for PDEs, Gaeta ▶ Non-linear Dynamics, Symmetry and Perturbation Theory in, Gallavotti ▶ Perturbation Theory, Broer ▶ Normal Forms in Perturbation Theory, Broer and Hanssmann ▶ Hamiltonian Perturbation Theory (and Transition to Chaos), Teixeira ▶ Perturbation Theory for Non-smooth Systems, Verhulst ▶ Perturbation Analysis of Parametric Resonance, Walcher ▶ Perturbative Expansions, Convergence of.

We start by outlining some motivating examples.

Consider the nilpotent planar linear system of ordinary differential equations

$$\dot{x} = \begin{pmatrix} 0 & \varepsilon \\ 0 & 0 \end{pmatrix} x\,, \quad x(0) = x^0 \in \mathbb{R}^2\,,$$

where $\varepsilon \in \mathbb{R}$. The explicit solution is given by $x_1(t) = x_1^0 + x_2^0 \varepsilon t$, $x_2(t) = x_2^0$ and clearly the equilibrium $(0, 0)$ is not stable if $\varepsilon \neq 0$. On the other hand, if $U(x_1)$ is a smooth real valued analytic function, satisfying $U(0) = 0$, $U(x_1) > 0$ for $x_1 \neq 0$, then it is well known that for $\varepsilon > 0$ the Newton equation

$$\dot{x} = \begin{pmatrix} 0 & \varepsilon \\ 0 & 0 \end{pmatrix} x - \begin{pmatrix} 0 \\ U'(x_1) \end{pmatrix}$$

is stable at $(0, 0)$.

Another example, which enters in the framework considered here, is given by a conservative (i. e. Hamiltonian) dynamical system perturbed by a friction term.

Next, we illustrate the influence of the real nilpotent perturbations in the realm of the normal form theory and in the general theory of singular partial differential equations. Consider the linear PDE

$$(x_1 + \varepsilon x_2)\partial_{x_1} u + x_2 \partial_{x_2} u - \sqrt{2} x_3 \partial_{x_3} u - u = f(x)\,,$$

where $\varepsilon \in \mathbb{R}$, and f stands for a convergent power series in a neighborhood of the origin of $\mathbb{R}^3$, at least quadratic at $x = 0$. Such equations appear in the so-called homological equations for the reduction to Poincaré–Dulac linear normal form of systems of analytic ODEs having an equilibrium at the origin. It turns out that in the semisimple case $\varepsilon = 0$ we can solve the equation in the space of convergent power series while for $\varepsilon \neq 0$ (i. e., when a nontrivial Jordan block appears) the equation is solvable only formally, namely, divergent solutions appear. One is led in a natural way to study the Gevrey index of divergent solutions.

Complex and Real Jordan Canonical Forms

We start by revisiting the notion of complex and real canonical Jordan forms.

Recall that each linear map is decomposed uniquely into the sum of a semisimple map and a nilpotent one. We state a classical result in linear algebra

Lemma 1 *Let A be an $n \times n$ matrix with real or complex entries. Then it is uniquely decomposed as*

$$A = A_s + A_{\text{nil}}\,, \tag{3}$$

where A_s is semisimple (i. e., diagonalizable over $\mathbb{C}$) and A_{nil} is nilpotent, i. e., $A_{\text{nil}}^r = 0_{n\times n}$ for some positive integer r. Here $0_{n\times n}$ stands for the zero $n \times n$ matrix.

Denote by $\operatorname{spec}(A) = \{\lambda_1, \ldots, \lambda_n\} \subset \mathbb{C}$ the set of all eigenvalues of A counted with their multiplicity.

The complex Jordan canonical form (JCF) is defined by the following assertion cf. Gantmacher [21]:

Theorem 2 *Let A be an $n \times n$ matrix over $\mathbb{K}$, $\mathbb{K} = \mathbb{C}$ or $\mathbb{K} = \mathbb{R}$. Then there exist positive integers $m, k_1, \ldots, k_m$, $m \geq n$, $k_1 + \ldots + k_m = n$ and a matrix $S \in GL(n;\mathbb{C})$ such that*

$$S^{-1}AS = J_A := \begin{pmatrix} \lambda_1^A I_{k_1} + N_{k_1} & 0_{k_1\times k_2} & \ldots & 0_{k_1\times k_p} \\ 0_{k_2\times k_1} & \lambda_2^A I_{k_2} + N_{k_2} & \ldots & 0_{k_2\times k_3} \\ \vdots & \vdots & \vdots & \vdots \\ 0_{k_m\times k_1} & 0_{k_m\times k_2} & \ldots & \lambda_m^A I_{k_m} + N_{k_m} \end{pmatrix} \tag{4}$$

where $\lambda_1^A, \ldots, \lambda_m^A$ are the eigenvalues of A, which need not all be distinct, and N_r, when $r \geq 2$, stands for the square $r \times r$ matrix

$$\begin{pmatrix} 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & 0 & \ldots & 0 & 0 \end{pmatrix} \tag{5}$$

with the convention $N_1 = 0$. Moreover, if $A_{\text{nil}} \neq 0$, i. e., $k_j \geq 2$ for at least one $j \in \{1, \ldots, m\}$, then for every $\varepsilon \in \mathbb{C} \setminus 0$ the matrix $S(\varepsilon) \in GL(n;\mathbb{C}) = \operatorname{diag}\{S_1(\varepsilon), \ldots, S_m(\varepsilon)\}$, $S_j(\varepsilon) = 1$ if $k_j = 1$, $S_j(\varepsilon) = \operatorname{diag}\{1, \ldots, \varepsilon^{k_j-1}\}$, provided $k_j \geq 2$, $j = 1, \ldots, m$, satisfies the identity

$$S^{-1}(\varepsilon)AS(\varepsilon) = J_A(\varepsilon) = \begin{pmatrix} \lambda_1^A I_{k_1} + \varepsilon N_{k_1} & 0_{k_1\times k_2} & \ldots & 0_{k_1\times k_p} \\ 0_{k_2\times k_1} & \lambda_2^A I_{k_2} + \varepsilon N_{k_2} & \ldots & 0_{k_2\times k_3} \\ \vdots & \vdots & \vdots & \vdots \\ 0_{k_m\times k_1} & 0_{k_m\times k_2} & \ldots & \lambda_m^A I_{k_m} + \varepsilon N_{k_m} \end{pmatrix}. \tag{6}$$

One may define in an obvious way another JCF (lower triangular) replacing J by its transposed J^{T}.

Note that λ_k^A, $k = 1, \ldots, m$, are not necessarily distinct.

If μ is an eigenvalue of A with algebraic multiplicity d, i. e., it is a zero of multiplicity d of the characteristic polynomial

$$P_A(\lambda) = |A - \lambda I|\,,$$

one can have different Jordan block structures. For example, a 3×3 matrix with a triple eigenvalue μ can be reduced to one of the three JCF:

- the matrix μI_3, i. e., μ does not admit nontrivial Jordan blocks;
- $\begin{pmatrix} \mu & 1 & 0 \\ 0 & \mu & 1 \\ 0 & 0 & \mu \end{pmatrix}$;
- $\begin{pmatrix} \mu & 0 & 0 \\ 0 & \mu & 1 \\ 0 & 0 & \mu \end{pmatrix}$

For higher dimensions the description of all JCF becomes more involved, cf. [3,21].

Remark 3 We can choose $|\varepsilon|$ arbitrarily small but never zero if the nilpotent part is nonzero. The columns of the conjugating matrix S are formed by eigenvectors and generalized eigenvectors. The smallness of $|\varepsilon|$ leads in a natural way to view the presence of nontrivial nilpotent parts as a perturbation.

Next, if A is a real matrix, using the real and the imaginary parts of the eigenvectors for the complex eigenvalues, one introduces the real JCF.

Theorem 4 *Let $A \in M_{n\times n}(\mathbb{R})$. Then there exist nonnegative integers p, q, $1 \leq p + 2q \leq n$, p (if $p \geq 1$) positive integers $k_1, \ldots, k_p$, q (if $q \geq 1$) positive integers $\ell_1, \ldots, \ell_q$*

satisfying $k_1 + \cdots + k_p + 2(\ell_1 + \cdots + \ell_q) = n$, *and* $S \in SL(n;\mathbb{R})$, *such that*

$$S^{-1}AS = \begin{pmatrix} J_A^{\mathbf{R}} & 0_{k\times 2\ell} \\ 0_{2\ell\times k} & J_A^{\mathbf{C}} \end{pmatrix} \tag{7}$$

with

$$J_A^{\mathbf{R}} = \begin{pmatrix} \lambda_1^A I_{k_1} + \varepsilon_1^A N_{k_1} & 0_{k_1\times k_2} & \dots & 0_{k_1\times k_p} \\ 0_{k_2\times k_1} & \lambda_2^A I_{k_2} + \varepsilon_2^A N_{k_2} & \dots & 0_{k_2\times k_3} \\ \vdots & \vdots & \vdots & \vdots \\ 0_{k_p\times k_1} & 0_{k_p\times k_2} & \dots & \lambda_p^A I_{k_p} + \varepsilon_p^A N_{k_p} \end{pmatrix}, \tag{8}$$

for some $\lambda_j \in \mathbb{R}$, $j = 1, \dots, p$, *and*

$$J_A^{\mathbf{C}} = \begin{pmatrix} D_1^A & 0_{2\ell_1\times 2\ell_2} & \dots & 0_{2\ell_1\times 2\ell_q} \\ 0_{2\ell_2\times 2\ell_1} & D_2^A & \dots & 0_{2\ell_2\times 2\ell_q} \\ \vdots & \vdots & \vdots & \vdots \\ 0_{2\ell_q\times 2\ell_1} & 0_{2\ell_q\times 2\ell_2} & \dots & D_q^A \end{pmatrix}, \tag{9}$$

with D_μ^A *being* $2\ell_\mu \times 2\ell_\mu$ *matrices, written as* $\ell_\mu \times \ell_\mu$ *block matrices of* 2×2 *matrices of the following* 2×2 *block matrix form:*

$$D_\mu^A = \begin{pmatrix} \alpha_\mu & -\beta_\mu \\ \beta_\mu & \alpha_\mu \end{pmatrix} I_{\ell_\mu}(2) + \begin{pmatrix} \gamma_\mu^A & -\delta_\mu^A \\ \delta_\mu^A & \gamma_\mu^A \end{pmatrix} N_\mu(2), \tag{10}$$

for some $\alpha_\mu, \beta_\mu \in \mathbb{R}$, $\beta_\mu \neq 0$, *with* $\alpha_k \pm \beta_k i \in \operatorname{spec}(A)$. *Here,* $I_k(2) = diag\{I_2, \dots, I_2\}$ *denotes the* $2k \times 2k$ *matrix written as a* $k \times k$ *matrix with* 2×2 *block matrices while* $N_k(2)$ *stands for the following* $2k\times 2k$ *nilpotent matrix written as* $k\times k$ *matrix with* 2×2 *block matrices as entries:*

$$N_k(2) = \begin{pmatrix} 0_{2\times 2} & I_2 & 0_{2\times 2} & \dots & 0_{2\times 2} \\ 0_{2\times 2} & 0_{2\times 2} & I_2 & \dots & 0_{2\times 2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0_{2\times 2} & 0_{2\times 2} & 0_{2\times 2} & \dots & I_2 \\ 0_{2\times 2} & 0_{2\times 2} & 0_{2\times 2} & \dots & 0_{2\times 2} \end{pmatrix} \tag{11}$$

and $0_{r\times s}$ *stands for the zero* $r\times s$ *matrix.*

The smallness of the parameter ε and the explicit form of the conjugating matrices $S(\varepsilon)$ are instrumental in showing some useful estimates for the study of the dynamics of the linear maps A which are not semisimple. Let $r(A) := \max\{|\lambda| : \lambda \in \operatorname{spec}(A)\}$ (the spectral radius of A). Then the following estimate, useful in different branches of Dynamical Systems, holds (cf. [27])

Lemma 5 *For every* $\eta > 0$ *there exists a norm in* $\mathbb{R}^n$ *such that* $\|A\| \le r(A) + \eta$. *Moreover, if*

$$\{\lambda \in \operatorname{spec}(A) : |\lambda| = r(A) \quad \textit{do not admit Jordan blocks}\}$$

one has $\|A\| = r(A)$ *for some norm. In particular, if* A *is semisimple, the last conclusion holds.*

Next, consider the linear autonomous systems of ordinary differential equations

$$\dot{x} = Ax\,. \tag{12}$$

where $A \in M_n(\mathbb{R})$. We recall that if $x(0) = \xi \in \mathbb{R}^n$, then the unique solution is defined by

$$x(t) = \exp(tA)\xi := \sum_{k=0}^{\infty} \frac{t^k A^k}{k!}\xi\,, \tag{13}$$

e. g., cf. Arnold [3], Coddington and Levinson [14].

We exhibit an assertion where the structure of the real nilpotent perturbation A_{nil} in the linear part plays a crucial role for the stability for $t \ge 0$ of the solutions of the linear system (12). We recall that the origin is stable if for every $\varepsilon > 0$ one can find $\delta > 0$ such that $\|\xi\| < \delta$ implies $\|\exp(tA)\xi\| < \varepsilon$ for $t \ge 0$.

Proposition 6 *The zero solution of (12) is stable for* $t \ge 0$ *if and only if the following two conditions hold:*

i) $\operatorname{spec}(A) \subset \{\lambda \in \mathbb{C}; \operatorname{Re}\lambda \le 0\}$;
ii) if $\lambda \in \operatorname{spec}(A)$ *and* $\operatorname{Re}\lambda = 0$ *then* λ *does not admit a nontrivial Jordan block.*

On the other hand, the origin is asymptotically stable for $t \to +\infty$ *if and only if*

$$\operatorname{spec}(A) \subset \{\lambda \in \mathbb{C}; \operatorname{Re}\lambda < 0\}\,.$$

The proof is straightforward in view of the explicit formula for the exponent of the matrix $\exp(tA)$ by means of the Jordan canonical form.

In particular, we get

Corollary 7 *Let* A *be a real matrix such that all eigenvalues lie on the pure imaginary axis. Then the zero solution of (12) is stable if and only if* A *is semisimple (i. e.,* $A_{\text{nil}} = 0$*).*

Nilpotent Perturbation and Formal Normal Forms of Vector Fields and Maps Near a Fixed Point

Normal form theory (originating back to Poincaré's thesis) has proven to be one of the most useful tools for the local

analysis of dynamical systems near an equilibrium (singular) point x^0 for autonomous systems of ODE

$$\dot{x} = X(x) \tag{14}$$

or the associated vector field

$$\widetilde{X}(x) = \langle X(x), \partial_x \rangle = \sum_{j=1}^{n} X_j(x)\partial_{x_j} \tag{15}$$

(see [3,7,9,12,19,20,43], and the references therein). Without loss of generality (after a translation) one may assume that x^0 coincides with the origin and write

$$X(x) = Ax + R(x)\,, \quad A = \nabla X(0)\,, \qquad R(x) = O\left(|x|^2\right)\,, \quad |x| \to 0\,. \tag{16}$$

Denote by $\mathrm{spec}(A) = \{\lambda_1, \ldots, \lambda_n\}$ the spectrum of A. The basic idea, going back to Poincaré, is to find a (formal) change of the coordinates defined as a (at least) quadratic perturbation of the identity

$$x = u(y) = y + v(y)\,, \tag{17}$$

which transforms $\widetilde{X}$ into a new vector field $\widetilde{Y}(y)$ which has a "simpler" form (Poincaré–Dulac normal form).

The original idea of Poincaré regards the possibility to linearize $\widetilde{X}$, i. e., $\widetilde{Y}(y) = Ay$. Straightforward calculations show that the linearization of $\widetilde{X}$ means that $v(y)$ satisfies (at least formally) a system of first order semilinear partial differential equations, called the system of the homological (difference) equations

$$L_A v(y) = R(y + v(y)) \tag{18}$$

In fact, it is the system above where the first substantial technical difficulty appears if the nilpotent part A_{nil} is not zero, i. e., the matrix A is not diagonalizable. Indeed, if A is diagonalizable and we choose (after a linear change of the variables in $\mathbb{C}^n$) $A = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$, then the system (18) is written as

$$\sum_{k=1}^{n} \lambda_k \partial_{y_k} v_j(y) - \lambda_j v_j = R_j(y + v(y))\,, \quad j = 1, \ldots, n \tag{19}$$

We recall that $\widetilde{X}$ (or spec$A$) is said to be in the Poincaré domain (respectively, Siegel domain) if the convex hull of $\{\lambda_1, \ldots, \lambda_n\}$ in the complex plane does not contain (respectively, contains) 0. Further, spec(A) is called nonresonant iff

$$\langle \lambda, \alpha \rangle - \lambda_j \neq 0\,, \quad j = 1, \ldots, n,\ \alpha \in \mathbb{Z}_+^n(2)\,, \tag{20}$$

where $\langle \lambda, \alpha \rangle = \sum_{j=1}^{n} \lambda_j \alpha_j$, $\mathbb{Z}_+^n(2) := \{\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{Z}_+^n \colon |\alpha| := \alpha_1 + \cdots + \alpha_n \geq 2\}$.

By the Poincaré–Dulac theorem, if spec(A) is in the Poincaré domain, then there are at most finitely many resonances and there exists a convergent transformation (in some neighborhood of the singular point) which reduces $\widetilde{X}$ to a (finitely resonant) normal form.

Theorem 8 *Let the linear part A of the complex (respectively, real) field above be nonresonant. Then the vector field is formally linearizable by a complex (respectively, real) transformation.*

Denote by

$$\mathrm{Res}_j^A = \{\alpha \in \mathbb{Z}_+^n(2) \colon \lambda_j = \langle \lambda, \alpha \rangle\}\,, \quad j = 1, \ldots, n\,.$$

Clearly the nonresonance hypothesis is equivalent to $\mathrm{Res}_j^A = \emptyset$ for $j = 1, \ldots, n$.

Additional technical complications appear if we consider real vector fields.

Theorem 9 *Every formal vector field with a singular point at the origin is transformed by a formal complex change of variables to a field of the form*

$$\langle J_A z, \partial_z \rangle + \sum_{j=1}^{n} \sum_{\alpha \in \mathrm{Res}_j^A} q_{j;\alpha} z^\alpha \partial_{z_j} \tag{21}$$

The coefficients $q_{j;\alpha}$ may be complex even though the original vector field is real.

We note that if the linear part is nilpotent, i. e., spec$(A) = \{0\}$, then the theorem above gives no simplification. In that case Belitskii [7] has classified completely the formal normal forms.

Let A be a nilpotent matrix (i. e. the semisimple part A_s is the zero matrix). The Poincaré–Dulac NF does not provide any information. The following theorem is due to Belitskii [7] (see also Arnold and Ilyashenko [4]).

Theorem 10 *Let A be a nilpotent matrix and let $X(x)$ be a formal vector field with a linear part given by Ax. Then X is transformed by a formal complex change of variables $x \mapsto z$ to a field of the form*

$$\langle J_A z, \partial_z \rangle + \langle B(z), \partial_z \rangle \tag{22}$$

with $B(z)$ being at least quadratic near the origin, where the nonlinear vector field $\langle B(z), \partial_z \rangle$ commutes with $\langle J_A^* z, \partial_z \rangle$. (Here * stands for the Hermitian conjugation).

We point out that another important problem is the computation of the normal form. Here the presence of the Jordan blocks leads to substantial difficulties.

The description and the computation of nilpotent normal forms, based on an algebraic approach and the systematic use of the theory of invariance, with particular emphasis on the equivalence between different normal forms, has been developed in a body of papers (see [16,36,37], and the references therein).

Finally, we mention that nilpotent perturbations appear in the classification and Casimir invariants of Lie–Poisson brackets that are formed by Lie algebra extensions for physical systems admitting Hamiltonian structure to such brackets (e. g. cf. Thiffeault and Morison [49]).

Loss of Gevrey Regularity in Siegel Domains in the Presence of Jordan Blocks

The convergence question in the Siegel domain is more difficult since small divisors appear. In a fundamental paper Bruno [9] succeeded in proving a deep result of the following type: a formal normal form is convergent under an (optimal) arithmetic condition on the small divisors $|\langle\lambda,\alpha\rangle-\lambda_j|^{-1}$ and a condition on the formal normal form, called the A condition. It should be pointed out that while in the original paper [9] the condition A allows in some cases nontrivial Jordan blocks of the linear part, in the subsequent works the linear part A is required to be semisimple (diagonalizable).

We recall that for the finitely smooth and C^∞ local normal forms the presence of the real nilpotent part does not influence the convergence: e. g., in the Sternberg theorem (cf. [45]) the small divisors and the presence of Jordan blocks play no role (cf. [7], where many other references can be found).

Little is known about convergence–divergence problems in the analytic category if spec(A) is in the Siegel domain and A is *not* semisimple.

Even in the cases where the presence of the nilpotent part does not influence the assertions, the proofs become more involved. Here we outline various aspects of normal forms when nilpotent perturbations are present. We stress especially the real case.

The combined influence of the Jordan blocks and the small divisors on the convergence of the formal linearizing transformation for analytic vector fields in $\mathbb{R}^3$ has been studied in Gramchev [23] (see also [53] for some examples). Let $n=3$ and consider a nonsemisimple linear part $A\in GL(3;\mathbb{K})$ satisfying spec$(A)=\{\lambda,\mu,\mu\}$, $\lambda\neq\mu$. This means that we can reduce A to the ε Jordan normal form

$$A_\varepsilon=\begin{pmatrix}\lambda&0&0\\0&\mu&\varepsilon\\0&0&\mu\end{pmatrix},\quad \varepsilon\neq 0\,. \tag{23}$$

We recall that one can make $|\varepsilon|$ arbitrarily small by linear change of the variables (but never 0). Then spec(A) is nonresonant and in the Siegel domain iff

$$\rho:=\frac{\lambda}{\mu}<0\,,\quad \rho\notin\mathbb{Q}\,. \tag{24}$$

The typical example is a real vector field with a linear part given by

$$A^0_\varepsilon=\begin{pmatrix}\rho&0&0\\0&1&\varepsilon\\0&0&1\end{pmatrix},\quad \varepsilon\neq 0\,. \tag{25}$$

One observes that such real vector fields are nonresonant and hyperbolic and therefore, by the Chen theorem, linearizable by smooth transformations.

We recall that an irrational number ρ is said to be diophantine of order $\tau>0$, and write $\rho\in D(\tau)$, if there exist $C>0$ such that

$$\min_{p\in\mathbb{Z}}|q\rho+p|\geq\frac{C}{q^\tau}\,,\quad q\in\mathbb{N}\,. \tag{26}$$

By a classical result in number theory $D(\tau)\neq\emptyset$ iff $\tau\geq 1$. An irrational number ρ is called Liouville iff it is not diophantine.

Given an irrational number ρ we set

$$\tau_0=\tau_0(\rho)=\inf\{\tau>0:\ \text{such that}\ \rho\in D(\tau)\}\,,\ \text{before} \tag{27}$$

with the convention $\tau_0=+\infty$ if ρ is a Liouville number.

Theorem 11 *Let* spec(A) *be in the Siegel domain. Then L_A is not solvable in the space of convergent power series, namely, we can find RHS f which is analytic but the unique formal power series solution is divergent. Moreover, we can always find a convergent RHS f such that the unique formal solution u satisfies*

$$u\notin\bigcup_{1\leq\sigma<2+\tau_0}G^\sigma_f(\mathbb{K}^3) \tag{28}$$

In particular, if $\tau_0=+\infty$, then

$$u\notin G^\sigma\,,\quad \sigma\geq 1\,. \tag{29}$$

In case a diophantine condition is satisfied, estimates on the Gevrey character of the divergent power series are derived.

The nonsolvability of the LHE in the spaces of the convergent power series does note exclude a priori that the vector field is linearizable by analytic transformations.

However, using a fundamental result of Pérez Marco [39] one can state the following assertion for hyperbolic vector fields which are not in the Poincaré domain.

Theorem 12 *Let A be a real $n \times n$ matrix which is hyperbolic, not semisimple and is not in the Poincaré domain. Let $R(x) = (R_1(x), \ldots, R(x))$ be a real valued analytic function near the origin, at least quadratic near $x = 0$, i. e. $R(0) = 0$, $\nabla R(0) = 0$. Then there exists $\eta_0 > 0$ such that for almost all $\eta \in]0, \eta_0]$ in the sense of the Lebesgue measure, the vector field defined by*

$$X_\varepsilon(x) = Ax + \eta R(x) \tag{30}$$

is not linearizable by convergent transformations.

Remark 13 In fact, the result can be made more precise, using the notion of capacity instead of the Lebesgue measure and allowing polynomial dependence on η (cf. [39]).

Another direction where the presence of real nilpotent perturbations in the linear parts presents challenging obstacles is the study of the dynamics in a neighborhood of a fixed point carried out via a normalization up to finite order and the issue of optimal truncation

$$\sum_{|\alpha| \le N_{\text{opt}}} u_\alpha x^\alpha$$

of normal form transformations for analytic vector fields. In an impressive work Iooss and Lombardi [33] demonstrate in particular that for large classes of real analytic vector fields with semisimple linear parts one can truncate the formal normal form transformation for $|x| \le \delta$, $\delta_0 > 0$ arbitrarily small, in such a way that the reminder in the normal form $R(x)$ satisfies the following estimates

$$\sup_{|x| \le \delta} |R_{N_{\text{opt}}}(x)| \le M\delta^2 \exp\left(-\frac{w}{\delta^b}\right) \tag{31}$$

where $b = 1 + \tau$. Here either $\tau > n - 1$, in which case $\tau \ne 0$ is the diophantine index of the small divisor type estimates modulo the resonance set, namely for some $\gamma > 0$ the following estimates hold for the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A

$$|\langle \lambda, \alpha \rangle - \lambda_j| \ge \frac{\gamma}{|\alpha|^\tau}, \quad \text{if} \quad \langle \lambda, \alpha \rangle - \lambda_j \ne 0 \tag{32}$$

for $j = 1, \ldots, n$, $\alpha \in \mathbb{Z}_+^n(2)$, or $\tau = 0$ and λ satisfies the nonresonant type estimates

$$|\langle \lambda, \alpha \rangle - \lambda_j| \ge \gamma, \quad \text{if} \quad \langle \lambda, \alpha \rangle - \lambda_j \ne 0 \tag{33}$$

for $j = 1, \ldots, n, \alpha \in \mathbb{Z}_+^n(2)$.

The question of the validity of such results for analytic vector fields with nonsemisimple linearization is far more intricate. Iooss and Lombardi [33] give two examples of non-semisimple linearizations (nilpotent perturbations of size 2 and 3) for which the result is still true. The question remains totally open for other non-semisimple linearizations.

First-Order Singular Partial Differential Equations

This section deals with the study of formal power series solutions to singular linear first-order partial differential equations with analytic coefficients of the form

$$\sum_{j=1}^{d} a_j(x)\partial_{x_j} u(x) + b(x)u(x) = f(x), \tag{34}$$

where $a_j(x)$ (with $j = 1, \ldots, d$), $b(x)$ and the right-hand side $f(x)$ are analytic in a neighborhood of the origin of $\mathbf{C}^d$, and $a_j(0) = 0$ for $j = 1, \ldots, d$.

In an interesting paper Hibino [29] allows a Jacobian matrix $\nabla a(0)$ without a Poincaré condition. More precisely, the Jacobi matrix at the origin can be reduced via a conjugation with a nonsingular matrix S to the following form: for some nonnegative integers, m, d, p, and d positive integers $k_j \ge 2$ (with $j = 1, \ldots, d$), if $d \ge 1$, such that $m + r_1 + \ldots r_d + p = n$, one can write as follows:

$$S^{-1}\nabla a(0)S = \begin{pmatrix} A & & & & \\ & N_{r_1} & & & \\ & & \vdots & & \\ & & & N_{r_d} & \\ & & & & 0_{p\times p} \end{pmatrix} \tag{35}$$

where N_r, $r \ge 2$, stands for the $r \times r$ nilpotent Jordan block (5) and A is an $m \times m$ satisfying the Poincaré condition (the convex hull in $\mathbb{C}$ of the eigenvalues $\lambda_1, \ldots, \lambda_m$ of A does not contain the origin, provided $m \ge 1$).

One observes that the hypothesis $d \ge 1$ implies that $\nabla a(0)$ does not satisfy the Poincaré condition.

The fundamental hypothesis on the zero order term $b(x)$ reads as follows

$$\left|\sum_{r=1}^{m} \lambda\alpha_r + b(0)\right| \ne 0, \quad \alpha \in \mathbb{Z}_+^m \quad \text{if} \quad m \ge 1 \tag{36}$$

$$|b(0)| \ne 0 \quad \text{if} \quad m = 0 \tag{37}$$

In particular, by (37) one gets that necessarily $b(0) \ne 0$.

It should be stressed that the classes of singular partial differential equations above do not capture the systems of homological equations when small divisors occur, but

they outline some interesting features in the presence of nontrivial Jordan blocks even in the lack of small divisors phenomena.

Set $\tau_0 = \max_{\ell=1,\dots,d} r_\ell$ if $d \geq 1$. Then the main result in [29] reads as follows

Theorem 14 *Under the conditions (35)–(37) for every RHS*

$$f(x) = \sum_{\alpha \in \mathbb{Z}_+^n} f_\alpha x^\alpha$$

which converges in a neighborhood of the origin the equation (34) has a unique formal solution which belongs to the formal Gevrey space $G_f^\sigma(\mathbb{K}^n)$ with

$$\sigma = \begin{cases} 2\tau_0 & \text{if} \quad d \geq 1 \\ 2 & \text{if} \quad d = 0, p \geq 1 \\ 1 & \text{if} \quad d = p = 0 \end{cases} . \tag{38}$$

The Gevrey index is determined by a Newton polyhedron, a generalization of the notion of the Newton polygon for singular ordinary differential equations. The arguments of the proof rely on subtle Gevrey combinatorial estimates.

Extensions for first-order quasilinear singular partial differential equations are done by Hibino [31].

It should be pointed out that the loss of Gevrey regularity comes not only from the nilpotent Jordan blocks, but from the nonlinear (at least quadratic) terms in $a(x)$ as well. Indeed, for the equation

$$(1 - x_1^2 \partial_{x_1} - x_2^2 \partial_{x_2}) u(x) = x_1 + x_2$$

the unique formal solution is defined as follows

$$u(x) = \sum_{j=0}^{\infty} (j-1)!(x_1^j + x_2^j)$$

cf. [29]. For further investigations on the loss of Gevrey regularity for solutions of singular ordinary differential equations of irregular type see Gramchev and Yoshino [26]. For characterizations of the Borel summability of a divergent formal power series solution of classes of first-order linear singular partial differential equation of nilpotent type see [30] and the references therein.

Solvability in classical Sobolev spaces and Gevrey spaces for linear systems of singular partial differential equations with real coefficients in $\mathbb{R}^n$ with nontrivial real Jordan blocks are derived by Gramchev and Tolis [24].

Normal Forms for Real Commuting Vector Fields with Linear Parts Admitting Nontrivial Jordan Blocks

The main goal of this section is to exhibit the influence of nontrivial linear nilpotent parts for the simultaneous reduction to convergent normal forms of commuting vector fields with a common fixed point.

Consider a family of commuting $n \times n$ matrices $A_1, \dots, A_d$. If all matrices are semisimple, then they can be simultaneously diagonalized over $\mathbb{C}$ or put into a block-diagonal form over $\mathbb{R}$, if $A_1, \dots, A_d$ are real, by a linear transformation S (e. g., [12,21]). This property plays a crucial role in the study of the normal forms of commuting vector fields with semisimple linear parts (cf. [46,47]).

However, if the matrices have nontrivial nilpotent parts, then it is not possible to transform simultaneously $A_1, \dots, A_d$ in Jordan canonical forms. This is a consequence of the characterization of the centralizer of a matrix in a JCF (see [3,21]).

Apparently the first examples of simultaneous reduction to (formal) normal forms of commuting vector fields with nonsemisimple linear parts are due to Cicogna and Gaeta [12]) for two commuting vector fields using the set up of the symmetries. More precisely, first, the definition of Semisimple Joint Normal Form (SJNF) is introduced: let

$$X = \langle f(x), \partial_x \rangle , \quad Y = \langle g(x), \partial_x \rangle ,$$

$A = \nabla f(0)$, $B = \nabla g(0)$ and $f(x) = Ax + F(x)$, $g(x) = Bx + G(x)$. Then X and Y are said to be in SJNF if both F and G belong to $\mathrm{Ker}(\mathcal{A}_s) \bigcap \mathrm{Ker}(\mathcal{B}_s)$. Here $\mathcal{A}_s$ stands for the homological operator associated to the semisimple part A_s of A.

Next, X, Y are in X-Joint NF iff

$$F \in \mathrm{Ker}(\mathcal{A}^+) \bigcap \mathrm{Ker}(\mathcal{B}_s) , \quad \text{and} \quad g \in \mathrm{Ker}(\mathcal{A}_s) \bigcap \mathrm{Ker}(\mathcal{B}_s) .$$

Clearly in Y-JNF the role of f and g is reversed.

The main assertion is:

Theorem 15 *Let $[X, Y] = 0$. Then X and Y can be reduced to a SJNF by means of a formal change of the variables. They can also be reduced (formally) to X-Joint or Y-Joint NF.*

Example 16

$$A = \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & \varepsilon \\ 0 & 0 & \lambda \end{pmatrix} \quad B = \begin{pmatrix} p & 0 & q \\ r & s & t \\ 0 & 0 & s \end{pmatrix} , \tag{39}$$

where $\lambda, \varepsilon, p, q, r, s, t \in \mathbb{C} \setminus 0$. Then A and B commute but it is impossible, in general, to reduce B to the Jordan block structure of A, preserving that of A.

We point out that the problem of finding conditions guaranteeing simultaneous reduction of commuting matrices to Jordan normal forms is related to questions in the Lie group theory (e. g., see Chap. IV of [12] and the references therein).

Next, we outline recent results on simultaneous reductions to normal forms of commuting analytic vector fields admitting nontrivial Jordan blocks in their linear parts following Yoshino and Gramchev [54].

Let $\mathbb{K}$ be $\mathbb{K} = \mathbb{C}$ or $\mathbb{K} = \mathbb{R}$, and $B = \infty$, $B = \omega$ or $B = k$ for some $k > 0$. Let $\mathcal{G}_B^n$ denote a d-dimensional Lie algebra of germs at $0 \in \mathbb{K}^n$ of C^B vector fields vanishing at 0. Let ρ be a germ of singular infinitesimal $\mathbb{K}^d$-actions of class C^B $(d \geq 2)$

$$\rho\colon \mathbb{K}^d \longrightarrow \mathcal{G}_B^n \,. \tag{40}$$

Denote by $\mathrm{Act}^B(\mathbb{K}^d : \mathbb{K}^n)$ the set of germs of singular infinitesimal $\mathbb{K}^d$-actions of class C^B at $0 \in \mathbb{K}^n$. By choosing a basis $e_1, \dots, e_d \in \mathbb{K}^n$, the infinitesimal action can be identified with a d-tuple of germs at 0 of commuting vector fields $X^j = \rho(e_j)$, $j = 1, \dots, d$ (cf. [18,46,47,55]). We can define, in view of the commutativity relation, the action

$$\begin{aligned} &\widetilde{\rho}\colon \mathbb{K}^d \times \mathbb{K}^n \longrightarrow \mathbb{K}^n \,, \\ &\widetilde{\rho}(s; z) = X^1_{s_1} \circ \cdots \circ X^d_{s_d}(z) = X^{\sigma_1}_{s_{\sigma_1}} \circ \cdots X^{\sigma_d}_{s_{\sigma_d}}(z) \,, \\ &s = (s_1, \dots, s_d) \,, \end{aligned} \tag{41}$$

for all permutations $\sigma = (\sigma_1, \dots, \sigma_d)$ of $\{1, \dots, d\}$, where X^j_t denotes the flow of X^j. We denote by ρ_{lin} the linear action formed by the linear parts of the vector fields defining ρ.

A natural question is to investigate necessary and sufficient conditions for the linearization of ρ (allowing nilpotent perturbations in the linear parts) namely, whether there exists a C^B diffeomorphism g preserving 0 such that g conjugates $\widetilde{\rho}$ and $\widetilde{\rho_{\mathrm{lin}}}$

$$\widetilde{\rho}(s; g(z)) = g(\widetilde{\rho_{\mathrm{lin}}}(s, z)) \quad (s, z) \in \mathbb{K}^d \times \mathbb{K}^n \,. \tag{42}$$

It is well known (e. g., cf. [35]) that there exists a positive integer $m \leq n$ such that $\mathbb{K}^n$ is decomposed into a direct sum of m linear subspaces invariant under all $A^\ell = \nabla X_\ell(0)$ $(\ell = 1, \dots, d)$:

$$\begin{aligned} \mathbb{K}^n = \mathbb{I}^{s_1} + \cdots + \mathbb{I}^{s_m} \,, \quad \dim \mathbb{I}^{s_j} = s_j \,, \quad j = 1, \dots, m, \\ s_1 + \cdots + s_m = n. \end{aligned} \tag{43}$$

The matrices $A^1, \dots, A^d$ can be simultaneously brought into upper triangular form, and we write again A^ℓ for the matrices

$$A^\ell = \begin{pmatrix} A^\ell_1 & 0_{s_1 \times s_2} & \dots & 0_{s_1 \times s_m} \\ 0_{s_2 \times s_1} & A^\ell_2 & \dots & 0_{s_2 \times s_m} \\ \vdots & \vdots & \vdots & \vdots \\ 0_{s_m \times s_1} & 0_{s_m \times s_2} & \dots & A^\ell_m \end{pmatrix}, \qquad \ell = 1, \dots, d \,. \tag{44}$$

If $\mathbb{K} = \mathbb{C}$, the matrix A^ℓ_j is given by

$$A^\ell_j = \begin{pmatrix} \lambda^\ell_j & A^\ell_{j,12} & \dots & A^\ell_{j,1s_j} \\ 0 & \lambda^\ell_j & \dots & A^\ell_{j,2s_j} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \lambda^\ell_j \end{pmatrix}, \tag{45}$$

with $\lambda^\ell_j, A^\ell_{j,\nu\mu} \in \mathbb{C}$, $\ell = 1, \dots, d$, $j = 1, \dots, m$. On the other hand, if $\mathbb{K} = \mathbb{R}$, then we have, for every $1 \leq j \leq m$ two possibilities: firstly, all A^ℓ_j $(\ell = 1, \dots, d)$ are given by (45) with $\lambda^\ell_j \in \mathbb{R}$. Secondly, $s_j = 2\widetilde{s_j}$ is even and A^ℓ_j is a $\widetilde{s_j} \times \widetilde{s_j}$ square block matrix given by

$$A^\ell_j = \begin{pmatrix} R_2(\lambda^\ell_j, \mu^\ell_j) & A^{12}_{\ell,j} & \dots & A^{1\widetilde{s_j}}_{\ell j} \\ 0 & R_2(\lambda^\ell_j, \mu^\ell_j) & \dots & A^{2\widetilde{s_j}}_{\ell j} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & R_2(\lambda^\ell_j, \mu^\ell_j) \end{pmatrix}, \qquad \ell = 1, \dots, d \,, \tag{46}$$

where

$$R_2(\lambda, \mu) := \begin{pmatrix} \lambda & \mu \\ -\mu & \lambda \end{pmatrix}, \quad \lambda, \mu \in \mathbb{R} \,, \tag{47}$$

and $A^{rs}_{\ell j}$ are appropriate real matrices.

Following the decomposition (45) (respectively, (46)) we define $\widetilde{\lambda^j}$ by

$$\widetilde{\lambda^k} = {}^t(\lambda^k_1, \dots, \lambda^k_m) \in \mathbb{K}^m \,, \quad k = 1, \dots, d \,. \tag{48}$$

Then we assume

$$\widetilde{\lambda^1}, \cdots, \widetilde{\lambda^d} \quad \text{are linearly independent in} \quad \mathbb{K}^m \,. \tag{49}$$

One can easily see that (49) is invariantly defined.

By (44) we define

$$\vec{\lambda}_j = {}^t(\lambda^1_j, \cdots, \lambda^d_j) \in \mathbb{K}^d \,, \quad j = 1, \dots, m \,, \tag{50}$$

and

$$\Lambda_m := \{\vec{\lambda}_1, \dots, \vec{\lambda}_m\} . \tag{51}$$

We define the cone $\Gamma[\Lambda_m]$ by

$$\Gamma[\Lambda_m] = \left\{ \sum_{j=1}^{m} t_j \vec{\lambda}_j \in \mathbb{K}^d;\ t_j \geq 0, j = 1, \dots, m, \sum_{j=1}^{m} t_j \neq 0 \right\} . \tag{52}$$

Definition 17 A $\mathbb{K}^d$-action ρ is called a Poincaré morphism if there exists a basis $\Lambda_m \subset \mathbb{K}^m$ such that $\Gamma[\Lambda_m]$ is a proper cone in $\mathbb{K}^m$, namely it does not contain a straight real line. If the condition is not satisfied, then, we say that the $\mathbb{K}^d$-action is in a Siegel domain.

Note that the definition is invariant under the choice of the basis Λ_m.

Remark 18 The geometric definition above is equivalent to the notion of Poincaré morphism given by Stolovitch (Definition 6.2.1 in [46]).

Next, we need to introduce the notion of simultaneous resonance. For $\alpha = (\alpha_1, \dots, \alpha_m) \in \mathbb{K}^m$, $\beta = (\beta_1, \dots, \beta_m) \in \mathbb{K}^m$, we set $\langle \alpha, \beta \rangle = \sum_{\nu=1}^{m} \alpha_\nu \beta_\nu$. For a positive integer k we define $\mathbb{Z}_+^m(k) = \{\alpha \in \mathbb{Z}_+^m;\ |\alpha| \geq k\}$. Put

$$\omega_j(\alpha) = \sum_{\nu=1}^{d} |\langle \widetilde{\lambda}^\nu, \alpha \rangle - \lambda_j^\nu| , \quad j = 1, \dots, m , \tag{53}$$

$$\omega(\alpha) = \min\{\omega_1(\alpha), \dots, \omega_m(\alpha)\} . \tag{54}$$

Definition 19 The cone Λ_m is called simultaneously nonresonant (or, in short ρ is simultaneously nonresonant), if

$$\omega(\alpha) \neq 0 , \quad \forall \alpha \in \mathbb{Z}_+^m(2) . \tag{55}$$

If (55) does not hold, then Λ_m is said to be simultaneously resonant.

Clearly, the simultaneously nonresonant condition (55) is invariant under a change of the basis Λ_m.

The next assertion provides a geometrically invariant condition guaranteeing that the simultaneous reduction to normal form does not depend on (small) nilpotent perturbation of the linear part.

Theorem 20 *Let ρ be a Poincaré morphism. Then ρ is conjugated to a polynomial action by a convergent change of variables.*

Remark 21 As a corollary of Theorem 20 for vector fields having linear parts with nontrivial Jordan blocks one obtains generalizations of results for the existence of convergent normal forms for analytic vector fields admitting symmetries cf. [5,12,13].

Example 22 Let ρ be a $\mathbb{R}^2$-action in $\mathbb{R}^n$, $n \geq 4$ with $m = 3$. Choose a basis Λ_2 of $\mathbb{R}^3$ such that

$$\Lambda_2 = \left\{ {}^t(1, 1, \nu), {}^t(0, 1, \mu) \right\} , \quad \nu, \mu \in \mathbb{R} . \tag{56}$$

By (52), $\Gamma[\Lambda_2]$ is generated by the set of vectors $\{(1, 0), (1, 1), (\nu, \mu)\}$. Hence the action is a Poincaré morphism if and only if these vectors generate a proper cone, namely (ν, μ) is not in the set $\{(\nu, \mu) \in \mathbb{R}^2; \nu \leq \mu \leq 0\}$. We note that the interesting case is $\mu < \nu \leq 0$, where every generator in (56) is in a Siegel domain.

Next, given a two-dimensional Lie algebra, choose a basis X_1, X_2 with linear parts $A_j \in GL(4; \mathbb{C})$ satisfying $\mathrm{spec}(A_1) = \{1, 1, \nu, \nu\}$ and $\mathrm{spec}(A_2) = \{0, 1, \mu, \mu\}$, respectively, where $\nu < \mu < 0$, $(\nu, \mu) \notin \mathbb{Q}^2$, and

$$A_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \nu & 0 \\ 0 & 0 & 0 & \nu \end{pmatrix} , \quad A_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \mu & \varepsilon \\ 0 & 0 & 0 & \mu \end{pmatrix} , \tag{57}$$

where $\varepsilon \neq 0$.

We show a refinement of the divergence result in Gevrey classes in [54] for the solution v of the overdetermined systems of linear homological equations $L_j v := \nabla v(x) A_j x - A_j v = f^j$ $(j = 1, 2)$, with the compatibility conditions for the RHS (see [54] for more details).

Theorem 23 *Let $1/2 \leq \tau_0 < \infty$. Then there exists*

$$E_0 \subset \{(\nu, \mu) \in (\mathbb{R} \setminus \mathbb{Q})^2; \nu < \mu < 0 ,\ \nu \text{ does not satisfy the Bruno condition}\}$$

with the density of continuum such that for every $(\nu, \mu) \in E_0$, there exists analytic $f = {}^t(f^1, f^2) \in (\mathbb{C}_2^4\{x\})^2$, satisfying the compatibility condition for the overdetermined system and such that the unique formal solution $v(x)$ is not in $\bigcup_{1 \leq \sigma < 2+\tau} G^\sigma(\mathbb{C}^4)$. Moreover, for every analytic f we can find $C > 0$ such that the unique formal solution satisfies the anisotropic Gevrey estimates

$$|v_\alpha| \leq C^{|\alpha|+1} (\alpha_3 + \alpha_4)^{(1+\tau)\alpha_4} , \quad \alpha \in \mathbb{Z}_+^4(2) , \tag{58}$$

Analytic Maps near a Fixed Point in the Presence of Jordan Blocks

As for the real analytic local diffeomorphisms preserving the origin in $\mathbb{R}^n$, one has in fact to deal necessarily with hyperbolic maps (cf. [4]).

Recall first the complex analytic case cf. [3,4]. Let $\Phi(x)$ be a biholomorphic map of $\mathbb{C}^n$ preserving the origin and $\Phi'(0)$ the Jacobian matrix at the origin. Denote by $\mathrm{spec}(\Phi'(0)) = \{\lambda_1, \ldots, \lambda_n\}$ the spectrum of $\Phi'(0)$. Clearly $\lambda_j \neq 0$ for all $j = 1, \ldots, n$. We define the set of all resonance multiindices of Φ (actually it depends only on $\mathrm{spec}(\Phi'(0))$) as follows:

$$\begin{aligned} \mathrm{Res}[\lambda_1, \ldots, \lambda_n] &= \bigcup_{j=1}^{n} \mathrm{Res}^j[\lambda_1, \ldots, \lambda_n] \\ \mathrm{Res}^j[\lambda_1, \ldots, \lambda_n] &= \{\alpha \in \mathbb{Z}_+^n(2) \colon \lambda^\alpha - \lambda_j = 0\}\,; \end{aligned} \tag{59}$$

where $\mathbb{Z}_+^n(k) := \{\alpha \in \mathbb{Z}_+^n \colon |\alpha| \geq k\}$, $\lambda = (\lambda_1, \ldots, \lambda_n)$ and $\lambda^\alpha = \lambda_1^{\alpha_1} \ldots \lambda_n^{\alpha_n}$ for $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{Z}_+^n$.

Given $\Lambda \in GL(n, \mathbb{C})$ define $\mathcal{O}[\Lambda]$ as the germ of all local complex analytic diffeomorphisms (biholomorphic maps) $\Phi(x)$ of $\mathbb{C}^n$ with 0 as a fixed point and such that $\Phi'(0)$ is in the $GL(n, \mathbb{C})$-conjugacy class of Λ. We stress that if Λ is diagonalizable (semisimple) then the germ is determined by the spectrum of Λ, i. e., by n nonzero complex numbers $\lambda_1, \ldots, \lambda_n$.

We will say that $\Phi(x)$ is formally linearizable if there exists a formal series $u(x) = x + \sum_{\alpha \in \mathbb{Z}_+^n(2)} u_\alpha x^\alpha$ such that

$$u^{-1} \circ \Phi \circ u = \Phi'(0) \quad \text{formally in} \quad (\mathbb{C}[x])^n\,, \tag{60}$$

where $\mathbb{C}[x]$ stands for the set of all formal power series with complex coefficients. It is well known (the Poincaré–Dulac theorem, cf. [3]) that under the nonresonance hypothesis $\mathrm{Res}[\lambda_1, \ldots, \lambda_n] = \emptyset$, i. e.,

$$\lambda^\alpha - \lambda_j \neq 0\,, \quad \alpha \in \mathbb{Z}_+^n(2)\,, \quad j = 1, \ldots, n\,, \tag{61}$$

Φ is formally linearizable. In fact, (61) is a necessary and sufficient condition in order that every holomorphic $\Phi(x)$ with $\mathrm{spec}\Phi'(0) = \{\lambda_1, \ldots, \lambda_n\}$ is formally linearizable. We refer to Gramchev and Walcher [25] for formal and algebraic aspects of normal forms of maps.

The formal solution of (60) involves expressions of the form $(\lambda^\alpha - \lambda_j)^{-1}$; so when $\inf_\alpha |\lambda^\alpha - \lambda_j| = 0$ for some $j \in \{1, \ldots, n\}$, the convergence of u becomes a subtle question.

One of the main problems, starting from the pioneering work of Siegel [44], has been (and still is) to find general conditions which guarantee that $u(x)$ converges, i. e., that $\Phi(x)$ is linearizable. We recall the state of the art of this subject. If the linear part $\Phi'(0)$ is semisimple (i. e., $\Phi'(0)$ has no nontrivial Jordan blocks), then it is well known that for convergence we need arithmetic (Diophantine) conditions on

$$\omega(m) := \min_{2 \leq |\alpha| \leq m, 1 \leq j \leq n} |\lambda^\alpha - \lambda_j|\,, \quad m \in \mathbb{Z}_+(2)\,. \tag{62}$$

We refer for the history and references to the survey paper by Herman [28]. The best condition that implies linearizability for all maps with a given semisimple linear part is due to Bruno [9], and can be expressed as (following Herman, p. 143 in [28])

$$\sum_{k=1}^{\infty} 2^{-k} \ln(\omega^{-1}(2^{k+1})) < \infty\,. \tag{63}$$

If one assumes the Poincaré condition

$$\max_{1 \leq j \leq n} |\lambda_j| < 1 \quad \text{or} \quad \min_{1 \leq j \leq n} |\lambda_j| > 1\,, \tag{64}$$

then Φ is always analytically equivalent to its linear part $\Phi'(0)$, provided the nonresonance hypothesis holds. More generally, (64) implies that there are finitely many resonances and, according to the Poincaré–Dulac theorem, we can find a local biholomorphic change of the variables u bringing Φ to normal form

$$(u^{-1} \circ \Phi \circ u)(x) = \Phi'(0)x + P_{\mathrm{res}}(x)\,, \tag{65}$$

where the remainder $P_{\mathrm{res}}(x)$ is a polynomial map containing only resonant terms.

Little is known about the (non)linearizability of Φ in the analytic category if the Poincaré condition doesn't hold and the matrix $\Phi'(0)$ has at least one nontrivial Jordan block. When $n = 2$ and A has a double eigenvalue $\lambda_1 = \lambda_2$, $|\lambda_1| = 1$ and a nontrivial Jordan block, then in general Φ is not linearizable. This result is contained in Proposition 3, p. 143 in [28], which is a consequence of results of Ilyashenko [32] and Yoccoz (the latter proved in 1978; for published proofs we refer to Appendix, pp. 86–87 in [52]). It is not difficult to extend this negative result to $\mathbb{C}^n$, $n \geq 3$.

In a recent paper of DeLatte and Gramchev [17] biholomorphic maps in $\mathbb{C}^n$, $n = 3$ and $n = 4$ having a single nontrivial Jordan block of the linear part have been studied. We mention also the paper of Abate [1], where nondiagonalizable discrete holomorphic dynamical systems have been investigated using geometrical tools.

One observes that in the real case, as the matrix $\Phi'(0)$ is real, the nonresonance condition excludes eigenvalues on the unit circle, i. e., the map is hyperbolic.

As a corollary from results of Delatte and Gramchev [17] on nonsolvability of the LHE in the presence of Jordan blocks in the linear parts of biholomorphic maps preserving the origin of $\mathbb{C}^3$ and the fundamental results of Pérez Marco [39] one gets readily the following assertion for hyperbolic analytic maps preserving the origin in $\mathbb{R}^n$ and having nondiagonalizable linear parts at the origin.

Theorem 24 *Let A be a real $n \times n$ matrix such that its eigenvalues are nonresonant and lie outside the unit circle in $\mathbb{C}$, and A is neither expansive nor contractive (i. e., the Poincaré condition (64) is not satisfied). Let $R(x) = (R_1(x), \ldots, R(x))$ be a real valued analytic function near the origin, at least quadratic near $x = 0$, i. e. $R(0) = 0$, $\nabla R(0) = 0$. Then there exists $\eta_0 > 0$ such that for almost all $\eta \in]0, \eta_0]$ in the sense of the Lebesgue measure, the local analytic diffeomorphism defined by*

$$\Phi(x) = Ax + \eta R(x) \tag{66}$$

is not linearizable by convergent transformations.

Weakly Hyperbolic Systems and Nilpotent Perturbations

The presence of nondiagonalizable matrices appear as a challenging problem in the framework of the well-posedness of the Cauchy problem for evolution partial differential equations. In order to illustrate the main features we focus on linear hyperbolic systems with constant coefficients in space dimension one

$$\partial_t u + A\partial_x u + Bu = 0\,, \quad t > 0, \quad x \in \mathbb{R} \tag{67}$$

$$u(0, x) = u^0(x) \tag{68}$$

where A and B are real $m \times m$ matrices, $u = (u_1, \ldots, u_m)$ stands for a vector-valued smooth function. One is interested in the C^∞ well-posedness of the Cauchy problem (67), (68), namely, for every initial data $u^0 \in (C^\infty(\mathbb{R}))^m$ there exists a unique solution $u \in C^\infty(\mathbb{R}^2)$ of (67), (68).

We recall that the system is hyperbolic if the characteristic equation

$$|A - \lambda I| = 0 \tag{69}$$

has m real solutions $\lambda_1, \ldots, \lambda_m$. If the roots are distinct, the system is called strictly hyperbolic. The strict hyperbolicity implies that A is diagonalizable, which in turn implies that, after a linear change of variables, one can assume A to be diagonal and reduce essentially the problem to m first-order scalar equations. In that case the Cauchy problem is well-posed in the classical sense, namely for every smooth initial data u^0 (it is enough $u^0 \in C^1(\mathbb{R})$) there exists a unique smooth solution $u(t, x)$ to the Cauchy problem. In fact, it is enough to require that A is semisimple (i. e., allowing multiple eigenvalues but excluding nilpotent parts) in order to have well-posedness (e. g., cf. the book of Taylor [48] and the references therein for more details on strictly hyperbolic systems with variable coefficients, and more general set-up in the framework of pseudodifferential operators).

If the system is hyperbolic but not strictly hyperbolic (it is called also weakly hyperbolic), it means that the matrix A has multiple eigenvalues. Here the influence of the Jordan block structure is decisive and one has non existence theorems in the C^∞ category unless one imposes additional restrictions on the lower-order term B. In many applications one encounters weakly hyperbolic systems. One example is in the so-called water waves problem, concerning the motion of the free surface of a body of an incompressible irrotational fluid under the influence of gravity (see [15] and the references therein), where for the linearized system one encounters a 2×2 matrix of the type

$$A = \begin{pmatrix} c & \varkappa \\ 0 & c \end{pmatrix}\,,$$

where c stands for the velocity, and $\varkappa$ is a nonzero real number.

The assertions, even in the seemingly simple model cases, are not easy to state in simple terms. The first results on such systems are due to Kajitani [34]. The classification problem was completely settled for the so-called hyperbolic systems of constant multiplicities by Vaillant [50] using subtle linear algebra arguments with multiparameter dependence if the space dimension is greater then 1.

We illustrate the assertions for $m = 2$ and $m = 3$, where the influence of the nilpotent part is somewhat easier to describe in details. In what follows we rewrite the results in [34,50] by means of the Jordan block structures.

The case $m = 2$ is easy.

Proposition 25 *Let*

$$A = \begin{pmatrix} \lambda_0 & 1 \\ 0 & \lambda_0 \end{pmatrix}\,, \tag{70}$$

for some $\lambda_0 \in \mathbb{R}$. Then the Cauchy problem is C^∞ well posed iff $b_{21} = 0$, with

$$B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}\,.$$

Let the 3×3 real matrix A have a triple eigenvalue λ_0 and be not semisimple. Then we are reduced (modulo conjugation with an invertible matrix) to two possibilities: either

the JCF of A is a maximal Jordan block

$$\begin{pmatrix} \lambda_0 & 1 & 0 \\ 0 & \lambda_0 & 1 \\ 0 & 0 & \lambda_0 \end{pmatrix}, \tag{71}$$

or the degenerate case of one 2×2 and one 1×1 elementary Jordan blocks

$$A = \begin{pmatrix} \lambda_0 & 1 & 0 \\ 0 & \lambda_0 & 0 \\ 0 & 0 & \lambda_0 \end{pmatrix}, \tag{72}$$

Proposition 26 *Let $n = 3$. Then the following assertions hold:*

(i) let A be defined by (71). Then the Cauchy problem is well-posed in C^∞ iff the entries of the matrix

$$B = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix}$$

satisfy the identities

$$b_{31} = b_{21} + b_{32} = b_{11} - b_{13} = 0\,; \tag{73}$$

(ii) suppose that A is given by (72). The well-posedness in C^∞ holds iff

$$b_{21} = b_{21} b_{32} = 0\,. \tag{74}$$

In an interesting work, Petkov [38], using real Jordan block structures depending on parameters and reduction to normal forms of matrices depending on parameters (cf. [2]), derived canonical microlocal forms for the full symbol of a pseudodifferential system with real characteristics of constant multiplicity and applies them to study the propagation of singularities of solutions of certain systems.

More generally, the Cauchy problem for hyperbolic systems with multiple characteristics have been studied by various authors where (implicitly) conditions on the nilpotent perturbations and the lower order term are imposed (e. g., cf. [8,34,51], and the references therein).

It would be interesting to write down such conditions in terms of conditions on the nilpotent perturbations on the principal part and the lower-order terms.

Finally, we mention also the work of Ghedamsi, Gourdin, Mechab, and Takeuchi [22], concerned with the Cauchy problem for Schrödinger-type systems with characteristic roots of multiplicity two admitting nontrivial Jordan blocks.

Bibliography

1. Abate M (2000) Diagonalization of nondiagonalizable discrete holomorphic dynamical systems. Amer J Math 122:757–781
2. Arnold VI (1971) Matrices depending on parameters. Uspekhi Mat Nauk 26:101–114 (in Russian); Russ Math Surv 26:29–43
3. Arnold VI (1983) Geometrical methods in the theory of ordinary differential equations. Springer, New York
4. Arnold VI, Ilyashenko YU (1988) In: Anosov DV, Arnold VI (eds) Encyclopedia of Math Sci, vol 1. Dynamical Systems I. Springer, New York, pp 1–155
5. Bambusi D, Cicogna G, Gaeta G, Marmo G (1998) Normal forms, symmetry and linearization of dynamical systems. J Phys A 31:5065–5082
6. Belitskii GR (1978) Equivalence and normal forms of germs of smooth mappings. Uspekhi Mat Nauk 33:95–155, 263 (in Russian); Russ Math Surv 33:107–177
7. Belitskii GR (1979) Normal forms, invariants, and local mappings. Naukova Dumka, Kiev (in Russian)
8. Bove A, Nishitani T (2003) Necessary conditions for hyperbolic systems. II. Jpn J Math (NS) 29:357–388
9. Bruno AD (1971) The analytic form of differential equations. Tr Mosk Mat O-va 25:119–262; (1972) 26:199–239 (in Russian); See also (1971) Trans Mosc Math Soc 25:131–288; (1972) 26:199–239
10. Bruno AD, Walcher S (1994) Symmetries and convergence of normalizing transformations. J Math Anal Appl 183:571–576
11. Chen KT (1965) Diffeomorphisms: C^∞-realizations of formal properties. Amer J Math 87:140–157
12. Cicogna G, Gaeta G (1999) Symmetry and perturbation theory in nonlinear dynamics. Lecture Notes in Physics. New Series m: Monographs, vol 57. Springer, Berlin
13. Cicogna G, Walcher S (2002) Convergence of normal form transformations: the role of symmetries. (English summary) Symmetry and perturbation theory. Acta Appl Math 70:95–111
14. Coddington EA, Levinson N (1955) Theory of ordinary differential equations. McGraw-Hill, New York
15. Craig W (1987) Nonstrictly hyperbolic nonlinear systems. Math Ann 277:213–232
16. Cushman R, Sanders JA (1990) A survey of invariant theory applied to normal forms of vector fields with nilpotent linear part. In: Stanton D (ed) Invariant theory and tableaux. IMA Vol Math Appl, vol 19. Springer, New York, pp 82–106
17. DeLatte D, Gramchev T (2002) Biholomorphic maps with linear parts having Jordan blocks: Linearization and resonance type phenomena. Math Phys Electron J 8(2):1–27
18. Dumortier F, Roussarie R (1980) Smooth linearization of germs of R^2-actions and holomorphic vector fields. Ann Inst Fourier Grenoble 30:31–64
19. Gaeta G, Walcher S (2005) Dimension increase and splitting for Poincaré-Dulac normal forms. J Nonlinear Math Phys 12(1):327–342
20. Gaeta G, Walcher S (2006) Embedding and splitting ordinary differential equations in normal form. J Differ Equ 224:98–119
21. Gantmacher FR (1959) The theory of matrices, vols 1, 2. Chelsea, New York
22. Ghedamsi M, Gourdin D, Mechab M, Takeuchi J (2002) Équations et systèmes du type de Schrödinger à racines caractéristiques de multiplicité deux. Bull Soc Roy Sci Liège 71:169–187

23. Gramchev T (2002) On the linearization of holomorphic vector fields in the Siegel Domain with linear parts having nontrivial Jordan blocks. In: Abenda S, Gaeta G, Walcher S (eds) Symmetry and perturbation theory, Cala Gonone, 16–22 May 2002. World Sci. Publ., River Edge, pp 106–115
24. Gramchev T, Tolis E (2006) Solvability of systems of singular partial differential equations in function spaces. Integral Transform Spec Funct 17:231–237
25. Gramchev T, Walcher S (2005) Normal forms of maps: formal and algebraic aspects. Acta Appl Math 85:123–146
26. Gramchev T, Yoshino M (2007) Normal forms for commuting vector fields near a common fixed point. In: Gaeta G, Vitolo R, Walcher S (eds) Symmetry and Perturbation theory, Oltranto, 2–9 June 2007. World Sci Publ, River Edge, pp 203–217
27. Hasselblatt B, Katok A (2003) A first course in dynamics: with a panorama of recent developments. Cambridge University Press, Cambridge
28. Herman M (1987) Recent results and some open questions on Siegel's linearization theorem of germs of complex analytic diffeomorphisms of C^n near a fixed point. VIIIth international congress on mathematical physics, Marseille 1986. World Sci Publ, Singapore, pp 138–184
29. Hibino M (1999) Divergence property of formal solutions for first order linear partial differential equations. Publ Res Inst Math Sci 35:893–919
30. Hibino M (2003) Borel summability of divergent solutions for singular first order linear partial differential equations with polynomial coefficients. J Math Sci Univ Tokyo 10:279–309
31. Hibino M (2006) Formal Gevrey theory for singular first order quasi–linear partial differential equations. Publ Res Inst Math Sci 42:933–985
32. Il'yashenko Y (1979) Divergence of series reducing an analytic differential equation to linear form at a singular point. Funct Anal Appl 13:227–229
33. Iooss G, Lombardi E (2005) Polynomial normal forms with exponentially small remainder for vector fields. J Differ Equ 212:1–61
34. Kajitani K (1979) Cauchy problem for non–strictly hyperbolic systems. Publ Res Inst Math 15:519–550
35. Katok A, Katok S (1995) Higher cohomology for Abelian groups of toral automorphisms. Ergod Theory Dyn Syst 15:569–592
36. Murdock J (2002) On the structure of nilpotent normal form modules. J Differ Equ 180:198–237
37. Murdock J, Sanders JA (2007) A new transvectant algorithm for nilpotent normal forms. J Differ Equ 238:234–256
38. Petkov VM (1979) Microlocal forms for hyperbolic systems. Math Nachr 93:117–131
39. Pérez Marco R (2001) Total convergence or small divergence in small divisors. Commun Math Phys 223:451–464
40. Ramis J-P (1984) Théorèmes d'indices Gevrey pour les équations différentielles ordinaires. Mem Amer Math Soc 48:296
41. Rodino L (1993) Linear partial differential operators in Gevrey spaces. World Science, Singapore
42. Sanders JA (2005) Normal form in filtered Lie algebra representations. Acta Appl Math 87:165–189
43. Sanders JA, Verhulst F, Murdock J (2007) Averaging methods in nonlinear dynamical systems, 2nd edn. Applied Mathematical Sciences, vol 59. Springer, New York
44. Siegel CL (1942) Iteration of analytic functions. Ann Math 43:607–614
45. Sternberg S (1958) The structure of local homeomorphisms. II, III. Amer J Math 80:623–632, 81:578–604
46. Stolovitch L (2000) Singular complete integrability. Publ Math IHES 91:134–210
47. Stolovitch L (2005) Normalisation holomorphe d'algèbres de type Cartan de champs de vecteurs holomorphes singuliers. Ann Math 161:589–612
48. Taylor M (1981) Pseudodifferential operators. Princeton Mathematical Series, vol 34. Princeton University Press, Princeton
49. Thiffeault J-L, Morison PJ (2000) Classification and Casimir invariants of Lie–Poisson brackets. Physica D 136:205–244
50. Vaillant J (1999) Invariants des systèmes d'opérateurs différentiels et sommes formelles asymptotiques. Jpn J Math (NS) 25:1–153
51. Yamahara H (2000) Cauchy problem for hyperbolic systems in Gevrey class. A note on Gevrey indices. Ann Fac Sci Toulouse Math 19:147–160
52. Yoccoz J-C (1995) A remark on Siegel's theorem for nondiagonalizable linear part. Manuscript, 1978; See also Théorème de Siegel, nombres de Bruno e polynômes quadratic. Astérisque 231:3–88
53. Yoshino M (1999) Simultaneous normal forms of commuting maps and vector fields. In: Degasperis A, Gaeta G (eds) Symmetry and perturbation theory SPT 98, Rome, 16–22 December 1998. World Scientific, Singapore, pp 287–294
54. Yoshino M, Gramchev T (2008) Simultaneous reduction to normal forms of commuting singular vector fields with linear parts having Jordan blocks. Ann Inst Fourier (Grenoble) 58:263–297
55. Zung NT (2002) Convergence versus integrability in Poincaré-Dulac normal form. Math Res Lett 9:217–228

Perturbation Theory

Giovanni Gallavotti
Dipartimento di Fisica and I.N.F.N., sezione Roma-I, Università di Roma I "La Sapienza", Roma, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Poincaré's Theorem and Quanta
Mathematics and Physics. Renormalization
Need of Convergence Proofs
Multiscale Analysis
A Paradigmatic Example of PT Problem
Lindstedt Series
Convergence. Scales. Multiscale Analysis
Non Convergent Cases
Conclusion and Outlook
Future Directions
Bibliography

Glossary

Formal power series a power series, giving the value of a function $f(\varepsilon)$ of a parameter ε, that is derived assuming that f is analytic in ε.

Renormalization group method for multiscale analysis and resummation of formal power series. Usually applied to define a systematic collection of terms to organize a formal power series into a convergent one.

Lindstedt series an algorithm to develop formal power series for computing the parametric equations of invariant tori in systems close to integrable.

Multiscale problem any problem in which an infinite number of scales play a role.

Definition of the Subject

Perturbation Theory: Computation of a quantity depending on a parameter ε starting from the knowledge of its value for $\varepsilon = 0$ by deriving a power series expansion in ε, under the assumption of its existence, and if possible discussing the interpretation of the series. Perturbation theory is very often the only way to get a glimpse of the properties of systems whose equations cannot be "explicitly solved" in computable form.

The importance of Perturbation Theory is witnessed by its applications in Astronomy, where it led not only to the discovery of new planets (Neptune) but also to the discovery of Chaotic motions, with the completion of the Copernican revolution and the full understanding of the role of Aristotelian Physics formalized into uniform rotations of deferents and epicycles (today Fourier representation of quasi periodic motions). It also played an essential role in the development of Quantum Mechanics and the understanding of the periodic table. The successes of Quantum Field Theory in Electrodynamics first, then in Strong interactions and finally in the unification of the elementary forces (strong, electromagnetic, and weak) are also due to perturbation theory, which has also been essential in the theoretical understanding of the critical point universality. The latter two themes concern the new methods that have been developed in the last fifty years, marking a kind of new era for perturbation theory; namely dealing with singular problems, via the techniques called, in Physics, "Renormalization Group" and, in Mathematics, "Multiscale Analysis".

Introduction

Perturbation theory, henceforth PT, arises when the value of a function of interest is associated with a problem depending on a parameter, here called ε. The value has to be a simple, or at least explicit and rigorous, computation for $\varepsilon = 0$ while its computation for $\varepsilon \neq 0$, small, is attempted by expressing it as the sum of a power series in ε which will be called here the "solution".

It is important to say since the beginning that a real PT solution of a problem involves two distinct steps: the first is to show that assuming that there is a convergent power series solving the problem then the *coefficients* of the nth power of ε exist and can be computed via finite computation. The resulting series will be called *formal solution* or *formal series* for the problem. The second step, that will be called *convergence theory*, is to prove that the formal series converges for ε small enough, or at least find a "summation rule" that gives a meaning to the formal series thus providing a real solution to the problem. None of the two problems is trivial, in the interesting cases, although the second is certainly the key and a difficult one.

Once Newton's law of universal gravitation was established it became necessary to develop methods to find its implications. Laplace's "Mécanique Céleste" [19], provided a detailed and meticulous exposition of a general method that has become a classic, if not the first, example of perturbation theory, quite different from the parallel analysis of Gauss which can be more appropriately considered a "non perturbative" development.

Since Laplace one can say that many applications along his lines followed. In the XIX century wide attention was dedicated to extend Laplace's work to cover various astronomical problems: tables of the coefficients were dressed and published, and algorithms for their construction were devised, and planets were discovered (Neptune, 1846). Well known is the "Lindstedt algorithm" for the computation of the nth order coefficients of the PT series for the non resonant quasi periodic motions. The algorithm provides a power series representation for the quasi periodic motions with non resonant frequencies which is extremely simple: however it represents the nth coefficient as a sum of many terms, some of size of the order of a power on $n!$. Which of course is a serious problem for the convergence.

It became a central issue, known as the "small denominators problem" after Poincaré's deep critique of the PT method, generated by his analysis of the three-body problem. It led to his "non-integrability theorem" about the generic nonexistence of convergent power series in the perturbation parameter ε whose sum would be a constant of motion for a Hamiltonian H_ε, member of a family of Hamiltonians parametrized by ε and reducing to an integrable system for $\varepsilon = 0$. The theorem suggested (to some) that even the PT series of Lindstedt (to which Poincaré's theorem does not apply) could be meaningless even though formally well defined [23].

A posteriori, it should be recognized that PT was involved also in the early developments of Statistical Mechanics in the XIX century: the virial theorem application to obtain the Van der Waals equation of state can be considered a first-order calculation in PT (although this became clear only a century later with the identification of ε as the inverse of the space dimension).

Poincaré's Theorem and Quanta

With Poincaré begins a new phase: the question of convergence of series in ε becomes a central one in the Mathematics literature. Much less, however, in the Physics literature where the new discoveries in the atomic phenomena attracted the attention. It seems that in the Physics research it was taken for granted that convergence was not an issue: atomic spectra were studied via PT and early authoritative warnings were simply disregarded (for instance, explicit by Einstein, in [1], and clear, in [3], but "timid" being too far against the mainstream, for his young age). In this way quantum theory could grow from the original formulations of Bohr, Sommerfeld, Eherenfest relying on PT to the final formulations of Heisenberg and Schrödinger quite far from it. Nevertheless, the triumph of quantum theory was quite substantially based on the technical development and refinement of the methods of formal PT: the calculation of the Compton scattering, the Lamb shift, Fermi's weak interactions model and other spectacular successes came in spite of the parallel recognition that some of the series that were being laboriously computed not only could not possibly be convergent but their very existence, to all orders n, was in doubt.

The later Feynman graphs representation of PT was a great new tool which superseded and improved earlier graphical representations of the calculations. Its simplicity allowed a careful analysis and understanding of cases in which even *formal* PT seemed puzzlingly failing.

Renormalization theory was developed to show that the convergence problems that seemed to plague even the computation of the individual coefficients of the series, hence the formal PT series at fixed order, were, in reality, often absent, in great generality, as suspected by the earlier treatments of special (important) cases, like the higher-order evaluations of the Compton scattering, and other quantum electrodynamics cross sections or anomalous characteristic constants (e. g. the magnetic moment of the muon).

Mathematics and Physics. Renormalization

In 1943 the first important result on the convergence of the series of the Lindstedt kind was obtained by Siegel [25]: a formal PT series, of interest in the theory of complex maps, was shown to be convergent. Siegel's work was certainly a stimulus for the later work of Kolmogorov who solved [18], a problem that had been considered not soluble by many: to find the convergence conditions and the convergence proof of the Lindstedt series for the quasi periodic motions of a generic analytic Hamiltonian system, in spite of Poincaré's theorem and actually avoiding contradiction with it. Thus, showing the soundness of the comments about the unsatisfactory aspects of Poincaré's analysis that had been raised almost immediately by Weierstrass, Hadamard and others.

In 1956 not only Kolmogorov theorem appeared but also convergence of another well known and widely used formal series, the virial series, was achieved in an unnoticed work by Morrey [21], and independently rediscovered in the early 1960's.

At this time it seems that all series with well-defined terms were thought to be either convergent or at least asymptotic: for most Physicists convergence or asymptoticity were considered of little interest and matters to be left to Mathematicians.

However, with the understanding of the formal aspects of renormalization theory the interest in the convergence properties of the formal PT series once again became the center of attention.

On the one hand mathematical proofs of the existence of the PT series, for interesting quantum fields models, to all orders were investigated settling the question once and for all (Hepp's theorem [16]); on the other hand it was obvious that even if convergent (like in the virial or Meyer expansions, or in the Kolmogorov theory) it was well understood that the radius of convergence would not be large enough to cover all the physically interesting cases. The sum of the series would in general become singular in the sense of analytic functions and, even if admitting analytic continuation beyond the radius of convergence, a singularity in ε would be eventually hit. The singularity was supposed to correspond to very important phenomena, like the critical point in statistical mechanics or the onset of chaotic motions (already foreseen by Poincaré in connection with his non convergence theorem). Thus, research developed in two direction.

The first aimed at understanding the nature of the singularities from the formal series coefficients: in the 1960s many works achieved the understanding of the scaling laws (i. e. some properties of the divergences appearing at the singularities of the PT series or of its analytic continuation, for instance in the work of M. Fisher, Kadanoff, Widom and may others). This led to trying to find *resummations*, i. e. to collect terms of the formal series to trans-

form them into *convergent series* in terms of *new parameters*, the *running couplings*.

The latter would be singular functions of the original ε thus possibly reducing the study of the singularity to the singularities of the running couplings. The latter could be studied by independent methods, typically by studying the iterations of an auxiliary dynamical system (called the *beta function flow*). This was the *approach* or *renormalization group method* of Wilson [10,29].

The second direction was dedicated to finding out the real meaning of the PT series in the cases in which convergence was doubtful or a priori excluded: in fact already Landau had advanced the idea that the series could be just illusions in important problems like the paradigmatic quantum field theory of a scalar field or the fundamental quantum electrodynamics [4,15].

In a rigorous treatment the function that the series were supposed to represent would be in fact a trivial function with a dependence on ε unrelated to the coefficients of the well defined and non trivial but formal series. It was therefore important to show that there were at least cases in which the perturbation series of a nontrivial problem had a meaning determined by its coefficients. This was studied in the scalar model of quantum field theory and a proof of "non triviality" was achieved after the ground-breaking work of Nelson on two-dimensional models [22,26]: soon followed by similar results in two dimensions and the difficult extension to three-dimensional models by Glimm and Jaffe [14], and generating many works and results on the subject which took the name of "constructive field theory" [6].

But Landau's *triviality conjecture* was actually dealing with the "real problem", i. e. the 4-dimensional quantum fields. The conjecture remains such at the moment, in spite of very intensive work and attempts at its proof. The problem had relevance because it could have meant that not only the simple scalar models of constructive field theory were trivial but also the QED series which had received strong experimental support with the correct prediction of fine structure phenomena could be illusions, in spite of their well-defined PT series: which would remain as mirages of a non existing reality.

The work of Wilson made clear that the "triviality conjecture" of Landau could be applied only to theories which, after the mentioned resummations, would be controlled by a beta function flow that could not be studied perturbatively, and introduced the new notion of *asymptotic freedom*. This is a property of the beta function flow, implying that the running couplings are bounded and small so that the resummed series are more likely to have a meaning [29].

This work revived the interest in PT for quantum fields with attention devoted to new models that had been believed to be non renormalizable. Once more the apparently preliminary problem of developing a formal PT series played a key role: and it was discovered that many Yang–Mills quantum field theories were in fact renormalizable in the ultraviolet region [27,28], and an exciting period followed with attempts at using Wilson's methods to give a meaning to the Yang–Mills theory with the hope of building a theory of the strong interactions. Thus, it was discovered that several Yang–Mills theories were asymptotically free as a consequence of the high symmetry of the model, proving that what seemed to be strong evidence that no renormalizable model would have asymptotic freedom was an ill-founded belief (that in a sense slowed down the process of understanding, and not only of the strong interactions).

Suddenly understanding the strong interactions, until then considered an impossible problem became possible [15], as solutions could be written and *effectively computed* in terms of PT which, although not proved to be convergent or asymptotic (still an open problem in dimension $d = 4$) were immune to the argument of Landau. The impact of the new developments led a little later to the unification of all interactions into the *standard model* for the theory of elementary particles (including the electromagnetic and weak interactions). The standard model was shown to be asymptotically free *even in the presence of symmetry breaking*, at least if a few other interactions in the model (for instance the Higgs particle self interaction) were treated heuristically while waiting for the discovery of the "Higgs particle" and for a better understanding of the structure of the elementary particles at length scales intermediate between the Fermi scale ($\sim 10^{-15}$ cm (the weak interactions scale)) and the Planck scale (the gravitational interaction scale, 15 orders of magnitude below).

Given that the very discovery of renormalizability of Yang–Mills fields and the birth of a strong interactions theory had been firmly grounded on experimental results [15], the latter "missing step" was, and still is, considered an acceptable gap.

Need of Convergence Proofs

The story of the standard model is paradigmatic of the power of PT: it should convince anyone that the analysis of formal series, including their representation by diagrams, which plays an essential part, is to be taken seriously. PT is certainly responsible for the revival and solution of problems considered by many as hopeless.

In a sense PT in the elementary particles domain can only, so far, partially be considered a success. Different is the situation in the developments that followed the works of Siegel and Kolmogorov. Their relevance for Celestial Mechanics and for several problems in applied physics (particle accelerator design, nuclear fusion machines for instance) and for statistical mechanics made them too the object of a large amount of research work.

The problems are simpler to formulate and often very well posed but the possibility of existence of chaotic motions, always looming, made it imperative not to be content with heuristic analysis and imposed the quest of mathematically complete studies. The lead were the works of Siegel and Kolmogorov. They had established convergence of certain PT series, but there were other series which would certainly be not convergent even though formally well defined and the question was, therefore, which would be their meaning.

More precisely it was clear that the series could be used to find approximate solutions to the equations, representing the motion for very long times under the assumption of "small enough" ε. But this could hardly be considered an understanding of the PT series in Mechanics: the estimated values of ε would have to be too small to be of interest, with the exception of a few special cases. The real question was what could be done to give the PT series the status of exact solution.

As we shall see the problem is deeply connected with the above-mentioned asymptotic freedom: this is perhaps not surprising because the link between the two is to be found in the "multiscale analysis" problems, which in the last half century have been the core of the studies in many areas of Analysis and in Physics, when theoretical developments and experimental techniques became finer and able to explore nature at smaller and smaller scales.

Multiscale Analysis

To illustrate the multiscale analysis in PT it is convenient to present it in the context of Hamiltonian mechanics, because in this field it provides us with nontrivial cases of almost complete success.

We begin by contrasting the work of Siegel and that of Kolmogorov: which are based on radically different methods. The first being much closer in spirit to the developments of renormalization theory and to the Feynman graphs.

Most interesting formal PT series have a common feature: namely their nth order coefficients are constructed as sums of many "terms" and the first attempt to a complete analysis is to recognize that their sum, which gives the uniquely defined nth coefficient is much smaller than the sum of the absolute values of the constituent terms. This is a property usually referred to as a "cancellation" and, as a rule, it reflects some symmetry property of the problem: hence one possible approach is to look for expressions of the coefficients and for cancellations which would reduce the estimate of the nth order coefficients, very often of the order of a power of $n!$, to an exponential estimate $O(\varrho^{-n})$ for some $\varrho > 0$ yielding convergence (parenthetically in the mentioned case of Yang–Mills theories the reduction is even more dramatic as it leads from divergent expressions to finite ones, yet of order $n!$).

The multiscale aspect becomes clear also in Kolmogorov's method because the implicit functions theorem has to be applied over and over again and deals with functions implicitly defined on smaller and smaller domains [5,7]. But the method purposedly avoids facing the combinatorial aspects behind the cancellations so much followed, and cherished [28], in the Physics works.

Siegel's method was developed to study a problem in which no grouping of terms was eventually needed, even though this was by no means clear a priori [24]; and to realize that no cancellations were needed forced one to consider the problem as a multiscale one because the absence of rapid growth of the nth order coefficients became manifest after a suitable "hierarchical ordering" of the terms generating the coefficients. The approach establishes a strong connection with the Physics literature because the technique to study such cases was independently developed in quantum field theory with renormalization, as shown by Hepp in [16], relying strongly on it. This is very natural and, in case of failure, it can be improved by looking for "resummations" turning the power series into a convergent series in terms of functions of ε which are singular but controllably so. For details see below and [8].

What is "natural", however, is a very personal notion and it is not surprising that what some consider natural is considered unnatural or clumsy or difficult (or the three qualifications together) by others.

Conflict arises when the same problem can be solved by two different "natural" methods and in the case of PT for Hamiltonian systems close to integrable ones (closeness depending on the size of a parameter ε), the so-called "small denominators" problem, the methods of Siegel and Kolmogorov are antithetic and an example of the just mentioned dualism.

The first method, that will be called here "Siegel's method" (see below for details), is based on a careful analysis of the structure of the various terms that occur at a given PT order achieving a proof that the nth order co-

efficient which is represented as the sum of many terms some of which *might* have size of order of a power of n! has in fact a size of $O(\varrho^{-n})$ so that the PT series is convergent for $|\varepsilon| < \varrho$. Although strictly speaking the original work of Siegel does not immediately apply to the Hamiltonian Mechanics problems (see below), it can nevertheless be adapted and yields a solution, as made manifest much later in [2,8,24].

The second method, called here "Kolmogorov's method", instead does not consider the individual coefficients of the various orders but just regards the sum of the series as a solution of an implicit function equation (a "Hamilton–Jacobi" equation) and devises a recursive algorithm approximating the unknown sum of the PT series by functions analytic in a disk of fixed radius ϱ in the complex ε-plane [5,7].

Of course the latter approach implies that no matter how we achieve the construction of the nth order PT series coefficient there will have to be enough cancellations, if at all needed, so that it turns out bounded by $O(\varrho^{-n})$. And in the problem studied by Kolmogorov cancellations would be necessarily present if the nth order coefficient was represented by the sum of the terms in the Lindstedt series.

That this is not obvious is supported by the fact that it was considered an open problem, for about thirty years, to find a way to exhibit explicitly the cancellation mechanism in the Lindstedt series implied by Kolmogorov's work. This was done by Eliasson [2], who proved that the coefficients of the PT of a given order n as expressed by the construction known as the "Lindstedt algorithm" yielded coefficients of size of $O(\varrho^{-n})$: his argument, however, did not identify in general which term of the Lindstedt sum for the nth order coefficient was compensated by which other term or terms. It proved that the sum had to satisfy suitable relations, which in turn implied a total size of $O(\varrho^{-n})$. And it took a few more years for the complete identification [8], of the rules to follow in collecting the terms of the Lindstedt series which would imply the needed cancellations.

It is interesting to remark that, aside from the example of Hamiltonian PT, multiscale problems have dominated the development of analysis and Physics in recent time: for instance they appear in harmonic analysis (Carleson, Fefferman), in PDE's (DeGiorgi, Moser, Caffarelli–Kohn–Ninberg), in relativistic quantum mechanics (Glimm, Jaffe, Wilson), in Hamiltonian Mechanics (Siegel, Kolmogorov, Arnold, Moser), in statistical mechanics and condensed matter (Fisher, Wilson, Widom) … Sometimes, although not always, studied by PT techniques [10].

A Paradigmatic Example of PT Problem

It is useful to keep in mind an example illustrating technically what it means to perform a multiscale analysis in PT. And the case of quasi periodic motions in Hamiltonian mechanics will be selected here, being perhaps the simplest.

Consider the motion of ℓ unit masses on a unit circle and let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_\ell)$ be their positions on the circle, i. e. $\boldsymbol{\alpha}$ is a point on the torus $\mathcal{T}^\ell = [0, 2\pi]^\ell$. The points interact with a potential energy $\varepsilon f(\boldsymbol{\alpha})$ where ε is a strength parameter and f is a trigonometric even polynomial, of degree N: $f(\boldsymbol{\alpha}) = \sum_{\boldsymbol{\nu}\in\mathcal{Z}^\ell, |\boldsymbol{\nu}|\le N} f_{\boldsymbol{\nu}} e^{i\boldsymbol{\nu}\cdot\boldsymbol{\alpha}}$, $f_{\boldsymbol{\nu}} = f_{-\boldsymbol{\nu}} \in \mathcal{R}$, where $\mathcal{Z}^\ell$ denotes the lattice of the points with integer components in $\mathcal{R}^\ell$ and $|\boldsymbol{\nu}| = \sum_j |\nu_j|$.

Let $t \to \boldsymbol{\alpha}_0 + \boldsymbol{\omega}_0 t$ be the motion with initial data, at time $t = 0$, $\boldsymbol{\alpha}(0) = \boldsymbol{\alpha}_0, \dot{\boldsymbol{\alpha}}(0) = \boldsymbol{\omega}_0$, in which all particles rotate at constant speed with rotation velocity $\boldsymbol{\omega}_0 = (\omega_{01}, \ldots, \omega_{0\ell}) \in \mathcal{R}^\ell$. This is a solution for the equations of motion for $\varepsilon = 0$ and it is a quasi periodic solution, i. e. each of the angles α_j rotates periodically at constant speed ω_{0j}, $j = 1, \ldots, \ell$.

The motion will be called *non resonant* if the components of the rotation speed $\boldsymbol{\omega}_0$ are rationally independent: this means that $\boldsymbol{\omega}_0 \cdot \boldsymbol{\nu} = 0$ with $\boldsymbol{\nu} \in \mathcal{Z}^\ell$ is possible only if $\boldsymbol{\nu} = \mathbf{0}$. In this case the motion $t \to \boldsymbol{\alpha}_0 + \boldsymbol{\omega}_0 t$ covers, $\forall \boldsymbol{\alpha}_0$, densely the torus $\mathcal{T}^\ell$ as t varies. The PT problem that we consider is to find whether there is a family of motions "of the same kind" for each ε, small enough, solving the equations of motion; more precisely whether there exists a function $\mathbf{a}_\varepsilon(\boldsymbol{\varphi})$, $\boldsymbol{\varphi} \in \mathcal{T}^\ell$, such that setting

$$\boldsymbol{\alpha}(t) = \boldsymbol{\varphi} + \boldsymbol{\omega}_0 t + \mathbf{a}_\varepsilon(\boldsymbol{\varphi} + \boldsymbol{\omega}_0 t)\,, \quad \text{for } \boldsymbol{\varphi} \in \mathcal{T}^\ell \tag{1}$$

one obtains, $\forall \boldsymbol{\varphi} \in \mathcal{T}^\ell$ and for ε small enough, a solution of the equations of motion for a force $-\varepsilon \partial_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha})$: i. e.

$$\ddot{\boldsymbol{\alpha}}(t) = -\varepsilon \partial_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}(t))\,. \tag{2}$$

By substitution of Eq. (1) in Eq. (2), the condition becomes $(\boldsymbol{\omega}_0 \cdot \partial_{\boldsymbol{\varphi}})^2 \mathbf{a}(\boldsymbol{\varphi} + \boldsymbol{\omega}_0 t) = -\partial_{\boldsymbol{\alpha}} f(\boldsymbol{\varphi} + \boldsymbol{\omega}_0 t)$. Since $\boldsymbol{\omega}_0$ is assumed rationally independent $\boldsymbol{\varphi} + \boldsymbol{\omega}_0 t$ covers densely the torus $\mathcal{T}^\ell$ as t varies: hence the equation for $\mathbf{a}_\varepsilon$ is

$$(\boldsymbol{\omega}_0 \cdot \partial_{\boldsymbol{\varphi}})^2 \mathbf{a}_\varepsilon(\boldsymbol{\varphi}) = -\varepsilon\, \partial_{\boldsymbol{\alpha}} f(\boldsymbol{\varphi} + \mathbf{a}_\varepsilon(\boldsymbol{\varphi}))\,. \tag{3}$$

Applying PT to this equation means to look for a solution $\mathbf{a}_\varepsilon$ which is analytic in ε small enough and in $\boldsymbol{\varphi} \in \mathcal{T}^\ell$. In colorful language one says that the perturbation effect is slightly deforming a nonresonant torus with given frequency spectrum (i. e. given $\boldsymbol{\omega}_0$) on which the motion develops, without destroying it and keeping the quasi periodic motion on it with the same frequency spectrum.

Lindstedt Series

As it follows from a very simple special case of Poincaré's work, Eq. (3) cannot be solved if also $\boldsymbol{\omega}_0$ is considered variable and the dependence on $\varepsilon, \boldsymbol{\omega}_0$ analytic. Nevertheless if $\boldsymbol{\omega}_0$ is *fixed and non resonant* and if $\mathbf{a}_\varepsilon$ is supposed analytic in ε small enough and in $\boldsymbol{\varphi} \in \mathcal{T}^\ell$, then there can be at most one solution to the Eq. (3) with $\mathbf{a}_\varepsilon(\mathbf{0}) = \mathbf{0}$ (which is not a real restriction because if $\mathbf{a}_\varepsilon(\boldsymbol{\varphi})$ is a solution also $\mathbf{a}_\varepsilon(\boldsymbol{\varphi} + \mathbf{c}_\varepsilon) + \mathbf{c}_\varepsilon$ is a solution for any constant $\mathbf{c}_\varepsilon$). This so because the coefficients of the power series in ε, $\sum_{n=1}^\infty \varepsilon^n \mathbf{a}_n(\boldsymbol{\varphi})$, are uniquely determined if the series is convergent. In fact they are trigonometric polynomials of order $\le nN$ which will be written as

$$\boldsymbol{\alpha}_n(\boldsymbol{\varphi}) = \sum_{0<|\boldsymbol{\nu}|\le Nn} \boldsymbol{\alpha}_{n,\boldsymbol{\nu}} e^{i\boldsymbol{\nu}\cdot\boldsymbol{\varphi}} . \tag{4}$$

It is convenient to express them in terms of graphs. The graphs to use to express the value $\mathbf{a}_{n,\boldsymbol{\nu}}$ are

(i) trees with n *nodes* $v_1, \ldots, v_n$,
(ii) one root r,
(iii) n lines joining pairs (v_i', v_i) of nodes or the root and one node, *always one and not more*, (r, v_i);
(iv) the lines will be different from each other and distinguished by a mark label, $1, \ldots, n$ attached to them. The connections between the nodes that the lines generate have to be *loopless*, i. e. the graph formed by the lines must be a tree.
(v) The tree lines will be imagined oriented towards the root: hence a partial order is generated on the tree and the line joining v to v' will be denoted $\lambda_{v'v}$ and v' will be the node closer to the root immediately following v, hence such that $v' > v$ in the partial order of tree.

The number of such trees is large and exactly equal to n^{n-1}, as an application of Cayley's formula implies: their collection will be denoted T_n^0.

To compute $\mathbf{a}_{n,\boldsymbol{\nu}}$ consider all trees in T_n^0 and attach to each node v a vector $\boldsymbol{\nu}_v \in \mathcal{Z}^\ell$, called "mode label", such that $f_{\boldsymbol{\nu}_v} \ne 0$, hence $|\boldsymbol{\nu}_v| \le N$. To the root we associate one of the coordinate unit vectors $\boldsymbol{\nu}_r \equiv \mathbf{e}_r$. We obtain a set T_n of *decorated trees* (with $\le (2N+1)^{\ell n} n^{n-1}$ elements, by the above counting analysis).

Given $\theta \in T_n$ and $\lambda = \lambda(v', v) \in \theta$ we define the *current* on the line λ to be the vector $\boldsymbol{\nu}(\lambda) \equiv \boldsymbol{\nu}(v', v) \stackrel{\text{def}}{=} \sum_{w \le v} \boldsymbol{\nu}_w$: i. e. we imagine that the node vectors $\boldsymbol{\nu}_{v_i}$ represent currents entering the node v_i and flowing towards the root. Then $\boldsymbol{\nu}(\lambda)$ is, for each λ, the sum of the currents which entered all the nodes not following v, i. e. current accumulated after passing the node v.

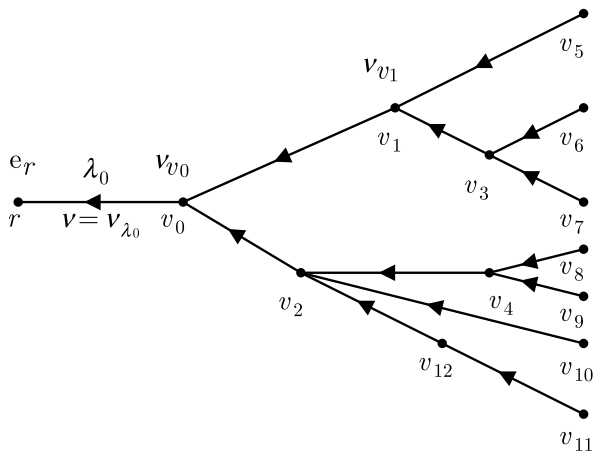

Perturbation Theory, Figure 1
A tree $\boldsymbol{\theta}$ with $m_{v_0} = 2, m_{v_1} = 2, m_{v_2} = 3, m_{v_3} = 2, m_{v_4} = 2, m_{v_{12}} = 1$ lines entering the nodes v_i, $k = 13$. Some labels or decorations explicitly marked (on the lines $\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1$ and on the nodes v_1, v_2); the number labels, distinguishing the branches, are not shown. The *arrows* represent the partial ordering on the tree

The current flowing in the root line $\boldsymbol{\nu} = \sum_v \boldsymbol{\nu}_v$ will be denoted $\boldsymbol{\nu}(\theta)$.

Let T_n^* be the set trees in T_n in which *all lines* carry a *non zero* current $\boldsymbol{\nu}(\lambda) \ne \mathbf{0}$. A *value* $\mathrm{Val}(\theta)$ will be defined, for $\theta \in T_n^*$, by a product of node factors and of line factors over all nodes and

$$\mathrm{Val}(\theta) = \frac{i(-1)^n}{n!} \prod_{v\in\theta} f_{\boldsymbol{\nu}_v} \prod_{\lambda=(v',v)} \frac{\boldsymbol{\nu}_{v'} \cdot \boldsymbol{\nu}_v}{(\boldsymbol{\omega}_0 \cdot \boldsymbol{\nu}(v', v))^2} . \tag{5}$$

The coefficient $\mathbf{a}_{n,\boldsymbol{\nu}}$ will then be

$$\mathbf{a}_{n,\boldsymbol{\nu}} = \sum_{\substack{\theta\in T_n^* \\ \boldsymbol{\nu}(\theta)=\boldsymbol{\nu}}} \mathrm{Val}(\theta) \tag{6}$$

and, when the coefficients are imagined to be constructed in this way, the formal power series $\sum_{n=1}^\infty \varepsilon^n \sum_{|\boldsymbol{\nu}|\le Nn} \mathbf{a}_{n,\boldsymbol{\nu}}$ is called the "Lindstedt series". Eq. (5) and its graphical interpretation in Fig. 1 should be considered the "Feynman rules" and the "Feynman diagrams" of the PT for Eq. (3) [9,10].

Convergence. Scales. Multiscale Analysis

The Lindstedt series is well defined because of the non resonance condition and the nth term is not even a sum of too many terms: if $F \stackrel{\text{def}}{=} \max_{\boldsymbol{\nu}} |f_{\boldsymbol{\nu}}|$, each of them can be bounded by $F^n/n! \prod_{\lambda\in\theta} N^2/(\boldsymbol{\omega}_0 \cdot \boldsymbol{\nu}(\lambda)^2)$; hence their sum can be bounded, if G is such that $((2N+1)^{\ell n} n^{n-1} F^n)/n! \le G^n$, by $G^n \prod_{\lambda\in\theta} N^2/(\boldsymbol{\omega}_0 \cdot \boldsymbol{\nu}(\lambda)^2)$.

Thus, all $\mathbf{a}_n$ are well defined and finite *but* the problem is that $|\boldsymbol{\nu}(\lambda)|$ can be large (up to Nn at given order n) and therefore $\boldsymbol{\omega}_0 \cdot \boldsymbol{\nu}(\lambda)$ although never zero can become very small as n grows. For this reason the problem of convergence of the series is an example of what is called a *small denominators problem*. And it is necessary to assume more than just non resonance of $\boldsymbol{\omega}_0$ in order to solve it in the present case: a simple condition is the *Diophantine* condition, namely the existence of $C, \tau > 0$ such that

$$|\boldsymbol{\omega}_0 \cdot \boldsymbol{\nu}| \geq \frac{1}{C\,|\boldsymbol{\nu}|^{\tau}}\,, \quad \forall \mathbf{0} \neq \boldsymbol{\nu} \in \mathcal{Z}^{\ell}\,. \tag{7}$$

But this condition is not sufficient in an obvious way: because it only allows us to bound individual tree-values by $n!^a$ for some $a > 0$ related to τ; furthermore it is not difficult to check that there are single graphs whose value is actually of "factorial" size in n. Although non trivial to see (as mentioned above) this was only apparently so in the earlier case of Siegel's problem but it is the new essential feature of the terms generating the nth order coefficient in Eq. (6).

A resummation is necessary to show that the tree-values can be grouped so that the sum of the values of each group can be bounded by ϱ^{-n} for some $\varrho > 0$ and $\forall n$, although the group may contain (several) terms of factorial size. The terms to be grouped have to be ordered hierarchically according to the sizes of the line factors $1/(\boldsymbol{\omega}_0 \cdot \boldsymbol{\nu}(\lambda))^2$, which are called *propagators* in [8,12].

A similar problem is met in quantum field theory where the graphs are the *Feynman graphs*: such graphs can only have a small number of lines that converge into a node but they can have loops, and to show that the perturbation series is well defined to all orders it is also necessary to collect terms hierarchically according to the propagators sizes. The systematic way was developed by Hepp [16,17], for the PT expansion of the Schwinger functions in quantum field theory of scalar fields [6]. It has been used on many occasions later and it plays a key role in the renormalization group methods in Statistical Mechanics (for instance in theory of the ground state of Fermi systems) [10,11].

However, it is in the Lindstedt series that the method is perhaps best illustrated. Essentially because it ends up in a convergence proof, while often in the field theory or statistical mechanics problems the PT series can be only proved to be well defined to all orders, but they are seldom, if ever, convergent so that one has to have recourse to other supplementary analytic means to show that the PT series are asymptotic (in the cases in which they are such).

The path of the proof is the following.

(1) Consider only trees in which no two lines λ_+ and λ_-, with λ_+ following λ_- in the partial order of the tree, have the *same* current $\boldsymbol{\nu}_0$. In this case the maximum of the $\prod_\lambda 1/(\boldsymbol{\omega}_0 \cdot \boldsymbol{\nu}(\lambda))^2$ over all tress $\theta \in T_n^*$ can be bounded by ${G_1}^n$ for some G_1.
This is an immediate consequence and the main result in Siegel's original work [25], which dealt with a different problem with small denominators in its formal PT solution: the coefficients of the series could also be represented by tree graphs, very similar too the ones above: but the only allowed $\boldsymbol{\nu} \in \mathcal{Z}^{\ell}$ were the non zero vectors with all components ≥ 0.
The latter property automatically guarantees that the graphs contain no pair of lines λ_+, λ_- following each other as above in the tree partial order and having the same current. Siegel's proof also implies a multiscale analysis [24]: but it requires no grouping of the terms unlike the analogue Lindstedt series, Eq. (6).

(2) Trees which contain lines λ_+ and λ_-, with λ_+ following λ_-, in the partial order of the tree, and having the *same* current $\boldsymbol{\nu}_0$ can have values which have size of order $O(n!^a)$ with some $a > 0$. Collecting terms is therefore essential.
A line λ of a tree is said to have scale k if $2^{-k-1} \leq 1/C|\boldsymbol{\omega}_0 \cdot \boldsymbol{\nu}| < 2^{-k}$. The lines of a tree $\theta \in T_n^*$ can then be collected in *clusters*.[1]
A cluster of scale p is a maximal connected set of lines of scale $k \geq p$ with at least one line of scale p. Clusters are connected to the rest of the tree by lines of lower scale which can be *incoming* or *outgoing* with respect to the partial ordering. Clusters also contain nodes: a node is in a cluster if it is an extreme of a line contained in a cluster; such nodes are said to be *internal* to the cluster.
Of particular interest are the *self energy* clusters. These are clusters with only one incoming line and only one outgoing line which *furthermore* have the same current $\boldsymbol{\nu}_0$.
To simplify the analysis the Diophantine condition can be strengthened to insure that if in a tree graph the line incoming into a self energy cluster and ending in an internal node v is detached from the node v and reattached to another node internal to the same cluster which is not in a self-energy subcluster (if any) then the new tree nodes are still enclosed in the same clusters. Alternatively the definition of scale of a line can be modified slightly to achieve the same goal.

[1] The scaling factor 2 is arbitrary: any scale factor > 1 could be used.

(3) Then it makes sense to sum together all the values of the trees whose nodes are collected into the same families of clusters and differ only because the lines entering the self energy clusters are attached to a different node internal to the cluster, but external to the inner self energy subclusters (if any). Furthermore, the value of the trees obtained by changing simultaneously sign to the $\boldsymbol{\nu}_v$ of the nodes inside the self energy clusters have also to be added together.
After collecting the terms in the described way it is possible to check that each sum of terms so collected is bounded by ϱ_0^{-n} for some ϱ_0 (which can also be estimated explicitly). Since the number of addends left is not larger than the original one the bound on $\sum_{\boldsymbol{\nu}} |\mathbf{a}_{n,\boldsymbol{\nu}}|$ becomes $\le (F^n(2N+1)^\ell n^n N^{2n})/n!\,\varrho_0^{-n} \le \varrho^{-n}$, for suitable ϱ_0, ϱ, so that convergence of the formal series for $\mathbf{a}_\varepsilon(\boldsymbol{\varphi})$ is achieved for $|\varepsilon| < \varrho$, see [8].

Non Convergent Cases

Convergence is *not* the rule: very interesting problems arise in which the PT series is, or is believed to be, only asymptotic. For instance in quantum field theory the PT series are well defined but they are not convergent: they can be proved, in the scalar φ^4 theories in dimension 2 and 3 to be asymptotic series for a function of ε which is *Borel summable*: this means in particular that the solution can be in principle recovered, for $\varepsilon > 0$ and small, just from the coefficients of its formal expansion.

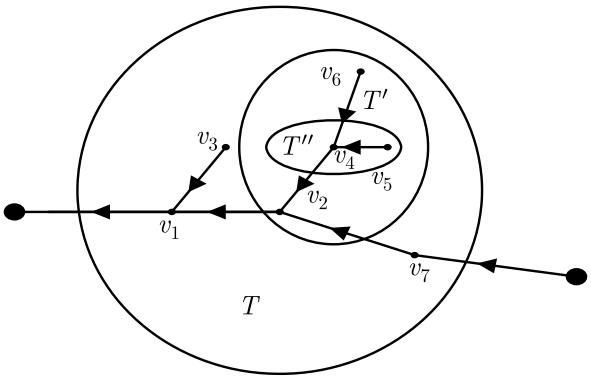

Perturbation Theory, Figure 2
An example of three clusters symbolically delimited by circles, as visual aids, inside a tree (whose remaining branches and clusters are not drawn and are indicated by the bullets); not all labels are explicitly shown. The scales (not marked) of the branches increase as one crosses inward the circles boundaries: recall, however, that the scale labels are integers ≤ 1 (hence typically ≤ 0). The $\boldsymbol{\nu}$ labels are not drawn (but must be imagined). If the $\boldsymbol{\nu}$ labels of (v_4, v_5) add up to 0 the cluster T'' is a self-energy graph. If the $\boldsymbol{\nu}$ labels of (v_2, v_4, v_5, v_6) add up to 0 the cluster T' is a self-energy graph and such is T if the $\boldsymbol{\nu}$ labels of $(v_1, v_2, v_3, v_4, v_5, v_6, v_7)$ add up to 0. The cluster T' is maximal in T

Other non convergent expansions occur in statistical mechanics, for example in the theory of the ground state of a Fermi gas of particles on a lattice of obstacles. This is still an open problem, and a rather important one. Or occur in quantum field theory where sometimes they can be proved to be Borel summable.

The simplest instances again arise in Mechanics in studying *resonant quasi periodic motions*. A paradigmatic case is provided by Eqs. (1),(2) when $\boldsymbol{\omega}_0$ has some vanishing components: $\boldsymbol{\omega}_0 = (\omega_1, \ldots, \omega_r, 0, \ldots, 0) = (\widetilde{\boldsymbol{\omega}}_0, \mathbf{0})$ with $1 < r < \ell$. If one writes $\boldsymbol{\alpha} = (\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}}) \in \mathcal{T}^r \times \mathcal{T}^{\ell-r}$ and looks motions like Eq. (1) of the form

$$\begin{aligned} \widetilde{\boldsymbol{\alpha}}(t) &= \widetilde{\boldsymbol{\varphi}} + \widetilde{\boldsymbol{\omega}}_0 t + \widetilde{\mathbf{a}}_\varepsilon(\widetilde{\boldsymbol{\varphi}} + \widetilde{\boldsymbol{\omega}}_0 t) \\ \widetilde{\boldsymbol{\beta}}(t) &= \boldsymbol{\beta}_0 + \widetilde{\mathbf{b}}_\varepsilon(\widetilde{\boldsymbol{\varphi}} + \widetilde{\boldsymbol{\omega}}_0 t) \end{aligned} \tag{8}$$

where $\widetilde{\mathbf{a}}_\varepsilon(\widetilde{\boldsymbol{\varphi}}), \widetilde{\mathbf{b}}_\varepsilon(\widetilde{\boldsymbol{\varphi}})$ are functions of $\widetilde{\boldsymbol{\varphi}} \in \mathcal{T}^r$, analytic in ε and $\widetilde{\boldsymbol{\varphi}}$.

In this case the analogue of the Lindstedt series can be devised provided $\boldsymbol{\beta}_0$ is chosen to be a stationary point for the function $\widetilde{f}(\widetilde{\boldsymbol{\beta}}) = \int f(\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}}) \frac{d\widetilde{\boldsymbol{\alpha}}}{(2\pi)^r}$, and provided $\widetilde{\boldsymbol{\omega}}_0$ satisfies a Diophantine property $|\widetilde{\boldsymbol{\omega}}_0 \cdot \widetilde{\boldsymbol{\nu}}| > 1/(C|\widetilde{\boldsymbol{\nu}}|^\tau)$ for all $\mathbf{0} \ne \widetilde{\boldsymbol{\nu}} \in \mathcal{Z}^r$ and for τ, C suitably chosen.

This time the series is likely to be, in general, non convergent (although there is not a proof yet). And the terms of the Lindstedt series can be suitably collected to improve the estimates. Nevertheless, the estimates cannot be improved enough to obtain convergence. Deeper resummations are needed to show that in some cases the terms of the series can be collected and rearranged into a convergent series.

The resummation is deeper in the sense that it is not enough to collect terms contributing to a given order in ε but it is necessary to collect and sum terms of different order according to the following scheme.

(1) The terms of the Lindstedt series are first "regularized" so that the new series is manifestly analytic in ε with, however, a radius of convergence depending on the regularization. For instance one can consider only terms with lines of scale $\le M$.
(2) Terms of different orders in ε are then summed together and the series becomes a series in powers of functions $\lambda_j(\varepsilon; M)$ of ε with very small radius of convergence in ε, *but* with an M-independent radius of convergence ϱ in the $\lambda_j(\varepsilon, M)$. The labels $j = 0, 1, \ldots, M$ are scale labels whose value is determined by the order in which they are generated in

the hierarchical organization of the collection of the graphs according to their scales.

(3) One shows that the functions $\lambda_j(\varepsilon; M)$ ("running couplings") can be analytically continued in ε to an *M-independent domain* $\mathcal{D}$ containing the origin in its closure and where they remain smaller than ϱ for all M. Furthermore, $\lambda_j(\varepsilon; M) \xrightarrow[M\to\infty]{} \lambda_j(\varepsilon)$, for $\varepsilon \in \mathcal{D}$.

(4) The convergent power series in the running couplings admits an asymptotic series in ε at the origin which coincides with the formal Lindstedt series. Hence in the domain $\mathcal{D}$ a meaning is attributed to the sum of Lindstedt series.

(5) One checks that the functions $\widetilde{\mathbf{a}}_\varepsilon, \widetilde{\mathbf{b}}_\varepsilon$ thus defined are such that Eq. (8) satisfies the equations of motion Eq. (2).

The proof can be completed if the domain $\mathcal{D}$ contains real points ε.

If $\widetilde{\boldsymbol{\beta}}_0$ is a maximum point the domain $\mathcal{D}$ contains a circle tangent to the origin and centered on the positive real axis. So in this case the $\widetilde{\mathbf{a}}_\varepsilon, \widetilde{\mathbf{b}}_\varepsilon$ are constructed in $\mathcal{D} \cap \mathcal{R}_+$, $\mathcal{R}_+ \stackrel{\text{def}}{=} (0, +\infty)$.

If instead $\widetilde{\boldsymbol{\beta}}_0$ is a minimum point the domain $\mathcal{D}$ exists but $\mathcal{D} \cap \mathcal{R}_+$ touches the positive real axis on a set of points with positive measure and density 1 at the origin. So $\widetilde{\mathbf{a}}_\varepsilon, \widetilde{\mathbf{b}}_\varepsilon$ are constructed only for ε in this set which is a kind of "Cantor set"[13].

Again the multiscale analysis is necessary to identify the tree values which have to be collected to define $\lambda_j(\varepsilon; M)$. In this case it is an analysis which is much closer to the similar analysis that is encountered in quantum field theory in the "self energy resummations", which involve collecting and summing graph values of graphs contributing to different orders of perturbation.

The above scheme can also be applied when $r = \ell$, i. e. in the case of the classical Lindstedt series when it is actually convergent: this leads to an alternative proof of the Kolmogorov theorem which is interesting as it is even closer to the renormalization group methods because it expresses the solution in terms of a power series in running couplings [Chaps. 8, 9 in 12].

Conclusion and Outlook

Perturbation theory provides a general approach to the solution of problems "close" to well understood ones, "closeness" being measured by the size of a parameter ε. It naturally consists of two steps: the first is to find a formal solution, under the assumption that the quantities of interest are analytic in ε at $\varepsilon = 0$. If this results in a power series with well-defined coefficients then it becomes necessary to find whether the series thus constructed, called a *formal series*, converges.

In general the proof that the formal series exists (when it really does) is nontrivial: typically in quantum mechanics problems (quantum fields or statistical mechanics) this is an interesting and deep problem giving rise to renormalization theory. Even in classical mechanics PT of integrable systems it has been, historically, a problem to obtain (in wide generality) the Lindstedt series (of which a simple example is discussed above).

Once existence of a PT series is established, very often the series is not convergent and at best is an asymptotic series. It becomes challenging to find its meaning (if any, as there are cases, even interesting ones, on which conjectures exist claiming that the series have no meaning, like the quantum scalar field in dimension 4 with "φ^4-interaction" or quantum electrodynamics).

Convergence proofs, in most interesting cases, require a multiscale analysis: because the difficulty arises as a consequence of the behavior of singularities at infinitely many scales, as in the case of the Lindstedt series above exemplified.

When convergence is not possible to prove, the multiscale analysis often suggest "resummations", collecting the various terms whose sums yields the formal PT series (usually the algorithms generating the PT series give its terms at given order as sums of simple but many quantities, as in the discussed case of the Lindstedt series). The collection involves adding together terms of different order in ε and results in a new power series, the *resummed series*, in a family of parameters $\lambda_j(\varepsilon)$ which are functions of ε, called the "running couplings", depending on a "scale index" $j = 0, 1, \ldots$.

The running couplings are (in general) singular at $\varepsilon = 0$ as functions of ε but C^∞ there, and obey equations that allow one to study and define them independently of a convergence proof. If the running couplings can be shown to be so small, as ε varies in a suitable domain $\mathcal{D}$ near 0, to guarantee convergence of the resummed series and therefore to give a meaning to the PT for $\varepsilon \in \mathcal{D}$ then the PT program can be completed.

The singularities in ε at $\varepsilon = 0$ are therefore all contained in the running couplings, usually very few and the same for various formal series of interest in a given problem.

The idea of expressing the sum of formal series as sum of convergent series in new parameters, the running couplings, determined by other means (a recursion relation denominated the beta function flow) is the key idea of the renormalization group methods: PT in mechanics is a typical and simple example.

On purpose attention has been devoted to PT in the analytic class: but it is possible to use PT techniques in problems in which the functions whose value is studied are not analytic; the techniques are somewhat different and new ideas are needed which would lead quite far away from the natural PT framework which is within the analytic class.

Of course there are many problems of PT in which the formal series are simply convergent and the proof does not require any multiscale analysis. Greater attention was devoted to the novel aspect of PT that emerged in Physics and Mathematics in the last half century and therefore problems not requiring multiscale analysis were not considered. It is worth mentioning, however, that even in simple convergent PT cases it might be convenient to perform resummations. An example is Kepler's equation

$$\ell = \xi - \varepsilon \sin \xi \,, \quad \xi, \ell \in T^1 = [0, 2\pi] \tag{9}$$

which can be (easily) solved by PT. The resulting series has a radius of convergence in ε rather small (Laplace's limit): however if a resummation of the series is performed transforming it into a power series in a "running coupling" $\lambda_0(\varepsilon)$ (only 1, because no multiscale analysis is needed, the PT series being convergent) given by [Vol. 2, p. 321 in 20]

$$\lambda_0 \stackrel{\text{def}}{=} \frac{\varepsilon\, e^{\sqrt{1-\varepsilon^2}}}{1 + \sqrt{1-\varepsilon^2}} \,, \tag{10}$$

then the resummed series is a power series in λ_0 with radius of convergence 1 and when ε varies between 0 and 1 the parameter λ_0 corresponding to it goes from 0 to 1. Hence in terms of λ_0 it is possible to invert *by power series* the Kepler equation for all $\varepsilon \in [0, 1)$, i. e. in the entire interval of physical interest (recall that ε has the interpretation of eccentricity of an elliptic orbit in the 2-body problem). Resummations can improve convergence properties.

Future Directions

It is always hard to indicate future directions, which usually turn to different paths. Perturbation theory is an ever evolving subject: it is a continuous source of problems and its applications generate new ones. Examples of outstanding problems are understanding the triviality conjectures of models like quantum φ^4 field theory in dimension 4 [6]; or a development of the theory of the ground states of Fermionic systems in dimensions 2 and 3 [11]; a theory of weakly coupled Anosov flows to obtain information of the kind that it is possible to obtain for weakly coupled Anosov maps [12]; uniqueness issues in cases in which PT series can be given a meaning, but in a priori non unique way like the resonant quasi periodic motions in nearly integrable Hamiltonian systems [12].

Bibliography

1. Einstein A (1917) Zum Quantensatz von Sommerfeld und Epstein. Verh Dtsch Phys Ges 19:82–102
2. Eliasson LH (1996) Absolutely convergent series expansions for quasi-periodic motions. Math Phys Electron J (MPEJ) 2:33
3. Fermi E (1923) Il principio delle adiabatiche e i sistemi che non ammettono coordinate angolari. Nuovo Cimento 25:171–175. Reprinted in Collected papers, vol I, pp 88–91
4. Frölich J (1982) On the triviality of $\lambda\varphi_d^4$ theories and the approach to the critical point in $d\,(\geq 4)$ dimensions. Nucl Phys B 200:281–296
5. Gallavotti G (1985) Perturbation Theory for Classical Hamiltonian Systems. In: Fröhlich J (ed) Scaling and self similarity in Physics. Birkhauser, Boston
6. Gallavotti G (1985) Renormalization theory and ultraviolet stability for scalar fields via renormalization group methods. Rev Mod Phys 57:471–562
7. Gallavotti G (1986) Quasi integrable mechanical systems. In: Phenomènes Critiques, Systèmes aleatories, Théories de jauge (Proceedings, Les Houches, XLIII (1984) vol II, pp 539–624) North Holland, Amsterdam
8. Gallavotti G (1994) Twistless KAM tori. Commun Math Phys 164:145–156
9. Gallavotti G (1995) Invariant tori: a field theoretic point of view on Eliasson's work. In: Figari R (ed) Advances in Dynamical Systems and Quantum Physics. World Scientific, pp 117–132
10. Gallavotti G (2001) Renormalization group in Statistical Mechanics and Mechanics: gauge symmetries and vanishing beta functions. Phys Rep 352:251–272
11. Gallavotti G, Benfatto G (1995) Renormalization group. Princeton University Press, Princeton
12. Gallavotti G, Bonetto F, Gentile G (2004) Aspects of the ergodic, qualitative and statistical theory of motion. Springer, Berlin
13. Gallavotti G, Gentile G (2005) Degenerate elliptic resonances. Commun Math Phys 257:319–362. doi:10.1007/s00220-005-1325-6
14. Glimm J, Jaffe A (1981) Quantum Physics. A functional integral point of view. Springer, New York
15. Gross DJ (1999) Twenty Five Years of Asymptotic Freedom. Nucl Phys B (Proceedings Supplements) 74:426–446. doi:10.1016/S0920-5632(99)00208-X
16. Hepp K (1966) Proof of the Bogoliubov–Parasiuk theorem on renormalization. Commun Math Phys 2:301–326
17. Hepp K (1969) Théorie de la rénormalization. Lecture notes in Physics, vol 2. Springer, Berlin
18. Kolmogorov AN (1954) On the preservation of conditionally periodic motions. Dokl Akad Nauk SSSR 96:527–530 and in: Casati G, Ford J (eds) (1979) Stochastic behavior in classical and quantum Hamiltonians. Lecture Notes in Physics vol 93. Springer, Berlin
19. Laplace PS (1799) Mécanique Céleste. Paris. Reprinted by Chelsea, New York
20. Levi-Civita T (1956) Opere Matematiche, vol 2. Accademia Nazionale dei Lincei and Zanichelli, Bologna
21. Morrey CB (1955) On the derivation of the equations of hydrodynamics from Statistical Mechanics. Commun Pure Appl Math 8:279–326

22. Nelson E (1966) A quartic interaction in two dimensions. In: Goodman R, Segal I (eds) Mathematical Theory of elementary particles. MIT, Cambridge, pp 69–73
23. Poincaré H (1892) Les Méthodes nouvelles de la Mécanique céleste. Paris. Reprinted by Blanchard, Paris, 1987
24. Pöschel J (1986) Invariant manifolds of complex analytic mappings. In: Osterwalder K, Stora R (eds) Phenomènes Critiques, Systèmes aleatories, Théories de jauge (Proceedings, Les Houches, XLIII (1984); vol II, pp 949–964) North Holland, Amsterdam
25. Siegel K (1943) Iterations of analytic functions. Ann Math 43:607–612
26. Simon B (1974) The $P(\varphi)_2$ Euclidean (quantum) field theory. Princeton University Press, Princeton
27. 't Hooft G (1999) When was asymptotic freedom discovered? or The rehabilitation of quantum field theory. Nucl Phys B (Proceedings Supplements) 74:413–425. doi:10.1016/S0920-5632(99)00207-8
28. 't Hooft G, Veltman MJG (1972) Regularization and renormalization of gauge fields, Nucl Phys B 44:189–213
29. Wilson K, Kogut J (1973) The renormalization group and the ε-expansion. Phys Rep 12:75–199

Perturbation Theory in Celestial Mechanics

Alessandra Celletti
Dipartimento di Matematica, Università di Roma Tor Vergata, Roma, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Classical Perturbation Theory
Resonant Perturbation Theory
Invariant Tori
Periodic Orbits
Future Directions
Bibliography

Glossary

KAM theory Provides the persistence of quasi-periodic motions under a small perturbation of an integrable system. KAM theory can be applied under quite general assumptions, i. e. a non-degeneracy of the integrable system and a diophantine condition of the frequency of motion. It yields a constructive algorithm to evaluate the strength of the perturbation ensuring the existence of invariant tori.

Perturbation theory Provides an approximate solution of the equations of motion of a nearly-integrable system.

Spin-orbit problem A model composed of a rigid satellite rotating about an internal axis and orbiting around a central point-mass planet; a spin-orbit resonance means that the ratio between the revolutional and rotational periods is rational.

Three-body problem A system composed by three celestial bodies (e. g. Sun-planet-satellite) assumed to be point-masses subject to the mutual gravitational attraction. The restricted three-body problem assumes that the mass of one of the bodies is so small that it can be neglected.

Definition of the Subject

Perturbation theory aims to find an approximate solution of nearly-integrable systems, namely systems which are composed by an integrable part and by a small perturbation. The key point of perturbation theory is the construction of a suitable canonical transformation which removes the perturbation to higher orders. A typical example of a nearly-integrable system is provided by a two-body model perturbed by the gravitational influence of a third body whose mass is much smaller than the mass of the central body. Indeed, the solution of the three-body problem greatly stimulated the development of perturbation theories. The solar system dynamics has always been a testing ground for such theories, whose applications range from the computation of the ephemerides of natural bodies to the development of the trajectories of artificial satellites.

Introduction

The two-body problem can be solved by means of Kepler's laws, according to which for negative energies the point-mass planets move on ellipses with the Sun located in one of the two foci. The dynamics becomes extremely complicated when adding the gravitational influence of another body. Indeed Poincaré showed [34] that the three-body problem does not admit a sufficient number of prime integrals which allow to integrate the problem. Nevertheless, the so-called *restricted* three-body problem deserves special attention, namely when the mass of one of the three bodies is so small that its influence on the others can be neglected. In this case one can assume that the primaries move on Keplerian ellipses around their common barycenter; if the mass of one of the primaries is much larger than the other (as it is the case in any Sun-planet sample), the motion of the minor body is governed by nearly-integrable equations, where the integrable part

represents the interaction with the major body, while the perturbation is due to the influence of the other primary. A typical example is provided by the motion of an asteroid under the gravitational attraction of the Sun and Jupiter. The small body may be taken not to influence the motion of the primaries, which are assumed to move on elliptic trajectories. The dynamics of the asteroid is essentially driven by the Sun and perturbed by Jupiter, since the Jupiter-Sun mass-ratio amounts to about 10^{-3}. The solution of this kind of problem stimulated the work of many scientists, especially in the XVIII and XIX centuries. Indeed, Lagrange, Laplace, Leverrier, Delaunay, Tisserand and Poincaré developed perturbation theories which are the basis of the studies of the dynamics of celestial bodies, from the computation of the ephemerides to the recent advances in flight dynamics. For example, on the basis of perturbation theory Delaunay [16] developed a theory of the Moon, providing very refined ephemerides. Celestial Mechanics greatly motivated the advances of perturbation theories as witnessed by the discovery of Neptune: its position was theoretically predicted by John Adams and by Jean Urbain Leverrier on the basis of perturbative computations; following the suggestion provided by the theoretical investigations, Neptune was finally discovered on 23 September 1846 by the astronomer Johann Gottfried Galle.

The aim of perturbation theory is to implement a canonical transformation which allows one to find the solution of a nearly-integrable system within a better degree of approximation (see Sect. "Classical Perturbation Theory" and references [3,6,20,24,32,37,38]. Let us denote the frequency vector of the system by $\underline{\omega}$ (see "Normal Forms in Perturbation Theory", "Kolmogorov–Arnol'd"Moser (KAM Theory)"), which we assume to belong to $\mathbf{R}^n$, where n is the number of degrees of freedom of the system. Classical perturbation theory can be implemented provided that the frequency vector satisfies a non-resonant relation, which means that there does not exist a vector $\underline{m} \in \mathbf{Z}^n$ such that $\underline{\omega} \cdot \underline{m} \equiv \sum_{j=1}^{n} \omega_j m_j = 0$. In case there exists such commensurability condition, a resonant perturbation theory can be developed as outlined in Sect. "Resonant Perturbation Theory". In general, the three-body problem (and, more extensively, the N-body problem) is described by a *degenerate* Hamiltonian system, which means that the integrable part (i. e., the Keplerian approximation) depends on a subset of the action variables. In this case a degenerate perturbation theory must be implemented as explained in Subsect. "Degenerate Perturbation Theory". For all the above perturbation theories (classical, resonant and degenerate) an application to Celestial Mechanics is given: the precession of the perihelion of Mercury, orbital resonances within a three-body framework, the precession of the equinoxes.

Even if the non-resonance condition is satisfied, the quantity $\underline{\omega} \cdot \underline{m}$ can become arbitrarily small, giving rise to the so-called *small divisor* problem; indeed, these terms appear in the denominator of the series defining the canonical transformations necessary to implement perturbation theory and therefore they might prevent the convergence of the series. In order to overcome the small divisor problem, a breakthrough came with the work of Kolmogorov [26], and was later extended to different mathematical settings by Arnold [2] and Moser [33]. The overall theory is known as the acronym KAM theory. As far as concrete estimates on the allowed size of the perturbation are concerned, the original versions of the theory gave discouraging results, which were extremely far from the physical measurements of the parameters involved in the proof. Nevertheless the implementation of computer-assisted KAM proofs allowed one to obtain results which are in good agreement with reality. Concrete estimates with applications to Celestial Mechanics are reported in Sect. "Invariant Tori".

In the framework of nearly-integrable systems, a very important role is provided by periodic orbits, which might be used to approximate the dynamics of quasi-periodic trajectories; for example, a truncation of the continued fraction expansion of an irrational frequency provides a sequence of rational numbers, which are associated to periodic orbits eventually approximating a quasi-periodic torus. A classical computation of periodic orbits using a perturbative approach is provided in Sect. "Periodic Orbits", where an application to the determination of the libration in longitude of the Moon is reported.

Classical Perturbation Theory

The Classical Theory

Consider a nearly-integrable Hamiltonian function of the form

$$H(\underline{I}, \underline{\varphi}) = h(\underline{I}) + \varepsilon f(\underline{I}, \underline{\varphi}) \,, \tag{1}$$

where h and f are analytic functions of $\underline{I} \in V$ (V is an open set of $\mathbf{R}^n$) and $\underline{\varphi} \in \mathbf{T}^n$ ($\mathbf{T}^n$ is the standard n-dimensional torus), while $\varepsilon > 0$ is a small parameter which measures the strength of the perturbation. The aim of perturbation theory is to construct a canonical transformation, which allows to remove the perturbation to higher orders in the perturbing parameter. To this end, let us look for a canonical change of variables (i. e., with symplectic Jacobian matrix) $C\colon (\underline{I}, \underline{\varphi}) \to (\underline{I}', \underline{\varphi}')$, such that the Hamiltonian (1)

takes the form

$$H'(\underline{I}', \underline{\varphi}') = H \circ C(\underline{I}, \underline{\varphi}) \equiv h'(\underline{I}') + \varepsilon^2 f'(\underline{I}', \underline{\varphi}') , \quad (2)$$

where h' and f' denote the new unperturbed Hamiltonian and the new perturbing function, respectively. To achieve such a result we need to proceed along the following steps: build a suitable canonical transformation close to the identity, perform a Taylor series expansion in the perturbing parameter, require that the unknown transformation removes the dependence on the angle variables up to second order terms, and expand in a Fourier series in order to get an explicit form of the canonical transformation.

The change of variables is defined by the equations

$$\begin{aligned} \underline{I} &= \underline{I}' + \varepsilon \frac{\partial \Phi(\underline{I}', \underline{\varphi})}{\partial \underline{\varphi}} \\ \underline{\varphi}' &= \underline{\varphi} + \varepsilon \frac{\partial \Phi(\underline{I}', \underline{\varphi})}{\partial \underline{I}'} , \end{aligned} \quad (3)$$

where $\Phi(\underline{I}', \underline{\varphi})$ is an unknown generating function, which is determined so that (1) takes the form (2). Decompose the perturbing function as

$$f(\underline{I}, \underline{\varphi}) = f_0(\underline{I}) + \tilde{f}(\underline{I}, \underline{\varphi}) ,$$

where f_0 is the average over the angle variables and $\tilde{f}$ is the remainder function defined through $\tilde{f}(\underline{I}, \underline{\varphi}) \equiv f(\underline{I}, \underline{\varphi}) - f_0(\underline{I})$. Define the *frequency vector* $\underline{\omega} = \underline{\omega}(\underline{I})$ as

$$\underline{\omega}(\underline{I}) \equiv \frac{\partial h(\underline{I})}{\partial \underline{I}} .$$

Inserting the transformation (3) in (1) and expanding in a Taylor series around $\varepsilon = 0$ up to the second order, one gets

$$\begin{aligned} & h\left(\underline{I}' + \varepsilon \frac{\partial \Phi(\underline{I}', \underline{\varphi})}{\partial \underline{\varphi}}\right) + \varepsilon f\left(\underline{I}' + \varepsilon \frac{\partial \Phi(\underline{I}', \underline{\varphi})}{\partial \underline{\varphi}}, \varphi\right) \\ &= h(\underline{I}') + \underline{\omega}(\underline{I}') \cdot \varepsilon \frac{\partial \Phi(\underline{I}', \underline{\varphi})}{\partial \underline{\varphi}} + \varepsilon f_0(\underline{I}') \\ &\qquad + \varepsilon \tilde{f}(\underline{I}', \underline{\varphi}) + O(\varepsilon^2) . \end{aligned}$$

The new Hamiltonian is integrable up to $O(\varepsilon^2)$ provided that the function Φ satisfies:

$$\underline{\omega}(\underline{I}') \cdot \frac{\partial \Phi(\underline{I}', \underline{\varphi})}{\partial \underline{\varphi}} + \tilde{f}(\underline{I}', \underline{\varphi}) = \underline{0} . \quad (4)$$

In such case the new integrable part becomes

$$h'(\underline{I}') = h(\underline{I}') + \varepsilon f_0(\underline{I}') ,$$

which provides a better integrable approximation with respect to (1). The solution of (4) yields the explicit expression of the generating function. In fact, let us expand Φ and $\tilde{f}$ in Fourier series as

$$\begin{aligned} \Phi(\underline{I}', \underline{\varphi}) &= \sum_{\underline{m} \in \mathbf{Z}^n \setminus \{\underline{0}\}} \hat{\Phi}_{\underline{m}}(\underline{I}') e^{i \underline{m} \cdot \underline{\varphi}}, \\ \tilde{f}(\underline{I}', \underline{\varphi}) &= \sum_{\underline{m} \in \mathcal{I}} \hat{f}_{\underline{m}}(\underline{I}') e^{i \underline{m} \cdot \underline{\varphi}} , \end{aligned} \quad (5)$$

where $\mathcal{I}$ denotes the set of integer vectors corresponding to the non-vanishing Fourier coefficients of $\tilde{f}$. Inserting the above expansions in (4) one obtains

$$i \sum_{\underline{m} \in \mathbf{Z}^n \setminus \{\underline{0}\}} \underline{\omega}(\underline{I}') \cdot \underline{m} \hat{\Phi}_{\underline{m}}(\underline{I}') e^{i \underline{m} \cdot \underline{\varphi}} = - \sum_{\underline{m} \in \mathcal{I}} \hat{f}_{\underline{m}}(\underline{I}') e^{i \underline{m} \cdot \underline{\varphi}} ,$$

which provides

$$\hat{\Phi}_{\underline{m}}(\underline{I}') = - \frac{\hat{f}_{\underline{m}}(\underline{I}')}{i \underline{\omega}(\underline{I}') \cdot \underline{m}} .$$

Casting together the above formule, the generating function is given by

$$\Phi(\underline{I}', \underline{\varphi}) = i \sum_{\underline{m} \in \mathcal{I}} \frac{\hat{f}_{\underline{m}}(\underline{I}')}{\underline{\omega}(\underline{I}') \cdot \underline{m}} e^{i \underline{m} \cdot \underline{\varphi}} . \quad (6)$$

We stress that this algorithm is constructive in the sense that it provides an explicit expression for the generating function and for the transformed Hamiltonian. We remark that (6) is well defined unless there exists an integer vector $\underline{m} \in \mathcal{I}$ such that

$$\underline{\omega}(\underline{I}') \cdot \underline{m} = 0 .$$

On the contrary, if $\underline{\omega}$ is rationally independent, there are no zero divisors in (6), though these terms can become arbitrarily small with a proper choice of the vector $\underline{m}$. This problem is known as the *small divisor problem*, which can prevent the implementation of perturbation theory (see "Normal Forms in Perturbation Theory", "Kolmogorov–Arnol'd–Moser (KAM Theory)", "Perturbation Theory").

The Precession of the Perihelion of Mercury

As an example of the implementation of classical perturbation theory we consider the computation of the precession of the perihelion in a (restricted, planar, circular) three-body model, taking as a sample the planet Mercury. The computation requires the introduction of Delaunay action-angle variables, the definition of the three-body Hamiltonian, the expansion of the perturbing function and the implementation of classical perturbation theory (see [7,39]).

Delaunay Action-Angle Variables We consider two bodies, say P_0 and P_1 with masses, respectively, m_0, m_1; let $M \equiv m_0 + m_1$ and let $\mu > 0$ be a positive parameter. Let r be the orbital radius and φ be the longitude of P_1 with respect to P_0; let (I_r, I_φ) be the momenta conjugated to (r, φ). In these coordinates the two-body problem Hamiltonian takes the form

$$H_{2b}(I_r, I_\varphi, r, \varphi) = \frac{1}{2\mu}\left(I_r^2 + \frac{I_\varphi^2}{r^2}\right) - \frac{\mu M}{r} . \tag{7}$$

On the orbital plane we introduce the planar Delaunay action-angle variables $(\Lambda, \Gamma, \lambda, \gamma)$ as follows [12]. Let E denote the total mechanical energy; then:

$$I_r = \sqrt{2\mu E + \frac{2\mu^2 M}{r} - \frac{I_\varphi^2}{r^2}} .$$

Since (7) does not depend on φ, setting $\Gamma = I_\varphi$ and $\Lambda = \sqrt{-(\mu^3 M^2)/(2E)}$, we introduce a generating function of the form

$$F(\Lambda, \Gamma, r, \varphi) = \int \sqrt{-\frac{\mu^4 M^2}{\Lambda^2} + \frac{2\mu^2 M}{r} - \frac{\Gamma^2}{r^2}}\, \mathrm{d}r + \Gamma\varphi .$$

From the definition of Λ the new Hamiltonian H_{2D} becomes

$$H_{2D}(\Lambda, \Gamma, \lambda, \gamma) = -\frac{\mu^3 M^2}{2\Lambda^2} ,$$

where (Λ, Γ) are the Delaunay action variables; by Kepler's laws one finds that (Λ, Γ) are related to the semi-major axis a and to the eccentricity e of the Keplerian orbit of P_1 around P_0 by the formula:

$$\Lambda = \mu\sqrt{Ma}, \qquad \Gamma = \Lambda\sqrt{1 - e^2} .$$

Concerning the conjugated angle variables, we start by introducing the eccentric anomaly u as follows: build the auxiliary circle of the ellipse, draw the line through P_1 perpendicular to the semi-major axis whose intersection with the auxiliary circle forms at the origin an angle u with the semi-major axis. By the definition of the generating function, one finds

$$\begin{aligned}\lambda = \frac{\partial F}{\partial \Lambda} &= \int \frac{\mu^4 M^2}{\Lambda^3 \sqrt{-\frac{\mu^4 M^2}{\Lambda^2} + \frac{2\mu^2 M}{r} - \frac{\Gamma^2}{r^2}}}\, \mathrm{d}r \\ &= u - e \sin u ,\end{aligned}$$

which defines the mean anomaly λ in terms of the eccentric anomaly u.

In a similar way, if f denotes the true anomaly related to the eccentric anomaly by $\tan f/2 = \sqrt{(1+e)/(1-e)}\, \tan u/2$, then one has:

$$\begin{aligned}\gamma = \frac{\partial F}{\partial \Gamma} &= \varphi - \int \frac{\Gamma}{r^2 \sqrt{-\frac{\mu^4 M^2}{\Lambda^2} + \frac{2\mu^2 M}{r} - \frac{\Gamma^2}{r^2}}}\, \mathrm{d}r \\ &= \varphi - f ,\end{aligned}$$

which represents the argument of the perihelion of P_1, i. e. the angle between the perihelion line and a fixed reference line.

The Restricted, Planar, Circular, Three-Body Problem Let P_0, P_1, P_2 be three bodies with masses m_0, m_1, m_2, respectively. We assume that m_1 is much smaller than m_0 and m_2 (restricted problem) and that the motion of P_2 around P_0 is circular. We also assume that the three bodies always move on the same plane. We choose the free parameter μ as $\mu \equiv 1/m_0^{2/3}$, so that the two-body Hamiltonian becomes $H_{2D} = -1/(2\Lambda^2)$, while we introduce the perturbing parameter as $\varepsilon \equiv m_2/m_0^{2/3}$ [12]. Set the units of measure so that the distance between P_0 and P_2 is one and so that $m_0 + m_2 = 1$. Taking into account the interaction of P_2 on P_1, the Hamiltonian function governing the three-body problem becomes

$$\begin{aligned}H_{3b}(\Lambda, \Gamma, \lambda, \gamma, t) &= -\frac{1}{2\Lambda^2} \\ &+ \varepsilon\left(r_1 \cos(\varphi - t) - \frac{1}{\sqrt{1 + r_1^2 - 2r_1 \cos(\varphi - t)}}\right) ,\end{aligned}$$

where r_1 is the distance between P_0 and P_1. The first term of the perturbation comes out from the choice of the reference frame, while the second term is due to the interaction with the external body. Since $\varphi - t = f + \gamma - t$, we perform the canonical change of variables

$$\begin{aligned} L &= \Lambda & \ell &= \lambda \\ G &= \Gamma & g &= \gamma - t ,\end{aligned}$$

which provides the following two degrees-of-freedom Hamiltonian

$$H_{3D}(L, G, \ell, g) = -\frac{1}{2L^2} - G + \varepsilon R(L, G, \ell, g) , \tag{8}$$

where

$$\begin{aligned}&R(L, G, \ell, g) \\ &\quad \equiv r_1 \cos(\varphi - t) - \frac{1}{\sqrt{1 + r_1^2 - 2r_1 \cos(\varphi - t)}}\end{aligned} \tag{9}$$

where r_1 and $\varphi - t$ must be expressed in terms of the Delaunay variables (L, G, ℓ, g). Notice that when $\varepsilon = 0$ one obtains the integrable Hamiltonian function $h(L, G) \equiv -1/(2L^2) - G$ with associated frequency vector $\underline{\omega} = (\partial h/\partial L, \partial h/\partial G) = (1/L^3, -1)$.

Expansion of the Perturbing Function We expand the perturbing function (9) in terms of the Legendre polynomials P_j obtaining

$$R = -\frac{1}{r_1}\sum_{j=2}^{\infty} P_j(\cos(\varphi - t))\frac{1}{r_1^j} .$$

The explicit expressions of the first few Legendre polynomials are:

$$\begin{aligned}
P_2(\cos(\varphi - t)) &= \frac{1}{4} + \frac{3}{4}\cos 2(\varphi - t)\\
P_3(\cos(\varphi - t)) &= \frac{3}{8}\cos(\varphi - t) + \frac{5}{8}\cos 3(\varphi - t)\\
P_4(\cos(\varphi - t)) &= \frac{9}{64} + \frac{5}{16}\cos 2(\varphi - t)\\
&\quad + \frac{35}{64}\cos 4(\varphi - t)\\
P_5(\cos(\varphi - t)) &= \frac{15}{64}\cos(\varphi - t) + \frac{35}{128}\cos 3(\varphi - t)\\
&\quad + \frac{63}{128}\cos 5(\varphi - t) .
\end{aligned}$$

We invert Kepler's equation $\ell = u - e\sin u$ to the second order in the eccentricity as

$$u = \ell + e\sin\ell + \frac{e^2}{2}\sin(2\ell) + O(e^3) ,$$

from which one gets

$$\begin{aligned}
\varphi - t &= g + \ell + 2e\sin\ell + \frac{5}{4}e^2\sin 2\ell + O(e^3)\\
r_1 &= a\left(1 + \frac{1}{2}e^2 - e\cos\ell - \frac{1}{2}e^2\cos 2\ell\right) + O(e^3) .
\end{aligned}$$

Then, up to inessential constants the perturbing function can be expanded as

$$\begin{aligned}
R = {} & R_{00}(L, G) + R_{10}(L, G)\cos\ell + R_{11}(L, G)\cos(\ell + g)\\
& + R_{12}(L, G)\cos(\ell + 2g) + R_{22}(L, G)\cos(2\ell + 2g)\\
& + R_{32}(L, G)\cos(3\ell + 2g) + R_{33}(L, G)\cos(3\ell + 3g)\\
& + R_{44}(L, G)\cos(4\ell + 4g) + R_{55}(L, G)\cos(5\ell + 5g)\\
& + \dots ,
\end{aligned} \tag{10}$$

where the coefficients R_{ij} are given by the following expressions (recall that $e = \sqrt{1 - G^2/L^2}$):

$$\begin{aligned}
R_{00} &= -\frac{L^4}{4}\left(1 + \frac{9}{16}L^4 + \frac{3}{2}e^2\right) + \dots ,\\
R_{10} &= \frac{L^4 e}{2}\left(1 + \frac{9}{8}L^4\right) + \dots\\
R_{11} &= -\frac{3}{8}L^6\left(1 + \frac{5}{8}L^4\right) + \dots ,\\
R_{12} &= \frac{L^4 e}{4}\left(9 + 5L^4\right) + \dots\\
R_{22} &= -\frac{L^4}{4}\left(3 + \frac{5}{4}L^4\right) + \dots ,\\
R_{32} &= -\frac{3}{4}L^4 e + \dots\\
R_{33} &= -\frac{5}{8}L^6\left(1 + \frac{7}{16}L^4\right) + \dots ,\\
R_{44} &= -\frac{35}{64}L^8 + \dots\\
R_{55} &= -\frac{63}{128}L^{10} + \dots .
\end{aligned} \tag{11}$$

Computation of the Precession of the Perihelion We identify the three bodies P_0, P_1, P_2 with the Sun, Mercury, and Jupiter, respectively. Taking ε as perturbing parameter, we implement a first order perturbation theory, which provides a new integrable Hamiltonian function of the form

$$h'(L', G') = -\frac{1}{2L'^2} - G' + \varepsilon R_{00}(L', G') .$$

From Hamilton's equations one obtains

$$\dot{g} = \frac{\partial h'(L', G')}{\partial G'} = -1 + \varepsilon\frac{\partial R_{00}(L', G')}{\partial G'} ;$$

neglecting $O(e^3)$ in R_{00} and recalling that $g = \gamma - t$, one has

$$\dot{\gamma} = \varepsilon\frac{\partial R_{00}(L', G')}{\partial G'} = \frac{3}{4}\varepsilon L'^2 G' .$$

Notice that to the first order in ε one has $L' = L$, $G' = G$. The astronomical data are $m_0 = 2 \cdot 10^{30}$ kg, $m_2 = 1.9 \cdot 10^{27}$ kg, which give $\varepsilon = 9.49 \cdot 10^{-4}$; setting to one the Jupiter–Sun distance one has $a = 0.0744$, while $e = 0.2056$. Taking into account that the orbital period of Jupiter amounts to about 11.86 years, one obtains

$$\dot{\gamma} = 154.65\frac{\text{arcsecond}}{\text{century}} ,$$

which represents the contribution due to Jupiter to the precession of perihelion of Mercury. The value found by

Leverrier on the basis of the data available in the year 1856 was of 152.59 arcsecond/century [14].

Resonant Perturbation Theory

The Resonant Theory

Let us consider an Hamiltonian system with n degrees of freedom of the form

$$H(\underline{I}, \underline{\varphi}) = h(\underline{I}) + \varepsilon f(\underline{I}, \underline{\varphi})$$

and let $\omega_j(\underline{I}) = (\partial h(\underline{I}))/(\partial I_j)$ $(j = 1, \ldots, n)$ be the frequencies of the motion, which we assume to satisfy ℓ, $\ell < n$, resonance relations of the form

$$\underline{\omega} \cdot \underline{m}_k = 0 \qquad \text{for } k = 1, \ldots, \ell \,,$$

for suitable rational independent integer vectors $\underline{m}_1$, …, $\underline{m}_\ell$. A resonant perturbation theory can be implemented to eliminate the non-resonant terms. More precisely, the aim is to construct a canonical transformation $C\colon (\underline{I}, \underline{\varphi}) \to (\underline{J}', \underline{\vartheta}')$ such that the transformed Hamiltonian takes the form

$$H'(\underline{J}', \underline{\vartheta}') = h'(\underline{J}', \vartheta'_1, \ldots, \vartheta'_\ell) + \varepsilon^2 f'(\underline{J}', \underline{\vartheta}') \,, \tag{12}$$

where h' depends only on the resonant angles $\vartheta'_1, \ldots, \vartheta'_\ell$. To this end, let us first introduce the angles $\underline{\vartheta} \in \mathbf{T}^n$ as

$$\begin{aligned} \vartheta_j &= \underline{m}_j \cdot \underline{\varphi} \qquad & j &= 1, \ldots, \ell \\ \vartheta_k &= \underline{m}_k \cdot \underline{\varphi} \qquad & k &= \ell + 1, \ldots, n \,, \end{aligned}$$

where the first ℓ angle variables are the resonant angles, while the latter $n - \ell$ angle variables are defined as suitable linear combinations so to make the transformation canonical together with the following change of coordinates on the actions $\underline{J} \in \mathbf{R}^n$:

$$\begin{aligned} I_j &= \underline{m}_j \cdot \underline{J} \qquad & j &= 1, \ldots, \ell \\ I_k &= \underline{m}_k \cdot \underline{J} \qquad & k &= \ell + 1, \ldots, n \,. \end{aligned}$$

The aim is to construct a canonical transformation which removes (to higher order) the dependence on the short-period angles $(\vartheta_{\ell+1}, \ldots, \vartheta_n)$, while the lowest order Hamiltonian will necessarily depend upon the resonant angles. Let us decompose the perturbation as

$$f(\underline{J}, \underline{\vartheta}) = \overline{f}(\underline{J}) + f_r(\underline{J}, \vartheta_1, \ldots, \vartheta_\ell) + f_n(\underline{J}, \underline{\vartheta}) \,, \tag{13}$$

where $\overline{f}$ is the average of the perturbation over the angles, f_r is the part depending on the resonant angles and f_n is the non-resonant part. In analogy to the classical perturbation theory, we implement a canonical transformation of the form

$$\begin{aligned} \underline{J} &= \underline{J}' + \varepsilon \frac{\partial \Phi}{\partial \underline{\vartheta}}(\underline{J}', \underline{\vartheta}) \\ \underline{\vartheta}' &= \underline{\vartheta} + \varepsilon \frac{\partial \Phi}{\partial \underline{J}'}(\underline{J}', \underline{\vartheta}) \,, \end{aligned}$$

such that the new Hamiltonian takes the form (12). Taking into account (13) and developing up to the second order in the perturbing parameter, one obtains:

$$\begin{aligned} & h\left(\underline{J}' + \varepsilon \frac{\partial \Phi}{\partial \underline{\vartheta}}\right) + \varepsilon f(\underline{J}', \underline{\vartheta}) + O(\varepsilon^2) \\ &= h(\underline{J}') + \varepsilon \sum_{k=1}^{n} \frac{\partial h}{\partial J_k} \frac{\partial \Phi}{\partial \vartheta_k} + \varepsilon \overline{f}(\underline{J}') + \varepsilon f_r(\underline{J}', \vartheta_1, \ldots, \vartheta_\ell) \\ &\quad + \varepsilon f_n(\underline{J}', \underline{\vartheta}) + O(\varepsilon^2) \,. \end{aligned}$$

Equating same orders of ε one gets that

$$h'(\underline{J}', \vartheta'_1, \ldots, \vartheta'_\ell) = h(\underline{J}') + \varepsilon \overline{f}(\underline{J}') + \varepsilon f_r(\underline{J}', \vartheta'_1, \ldots, \vartheta'_\ell) \,, \tag{14}$$

provided that

$$\sum_{k=1}^{n} \omega'_k \frac{\partial \Phi}{\partial \vartheta_k} = -f_n(\underline{J}', \underline{\vartheta}) \,, \tag{15}$$

where $\omega'_k = \omega'_k(\underline{J}') \equiv (\partial h(\underline{J}'))/(\partial J'_k)$. The solution of (15) gives the generating function, which allows one to reduce the Hamiltonian to the required form (12); as a consequence, the conjugated action variables, say $J_{\ell+1'}, \ldots, J'_n$, are constants of the motion up to the second order in ε. We conclude by mentioning that using the new frequencies ω'_k, the resonant relations take the form $\omega'_k = 0$ for $k = 1, \ldots, \ell$.

Three-Body Resonance

We consider the three-body Hamiltonian (8) with perturbing function (10)–(11) and let $\underline{\omega} \equiv (\omega_\ell, \omega_g)$ be the frequency of motion. We assume that the frequency vector satisfies the resonance relation

$$\omega_\ell + 2\omega_g = 0 \,.$$

According to the theory described in the previous section we perform the canonical change of variables

$$\begin{aligned} \vartheta_1 &= \ell + 2g \qquad & J_1 &= \frac{1}{2} G \\ \vartheta_2 &= 2\ell \qquad & J_2 &= \frac{1}{2} L - \frac{1}{4} G \,. \end{aligned}$$

In the new coordinates the unperturbed Hamiltonian becomes

$$h'(\underline{J}) \equiv -\frac{1}{2(J_1+2J_2)^2} - 2J_1 ,$$

with frequency vector $\underline{\omega}' \equiv \frac{\partial h'(\underline{J})}{\partial \underline{J}}$, while the perturbation takes the form

$$\begin{aligned} R(J_1, J_2, \vartheta_1, \vartheta_2) &\equiv R_{00}(\underline{J}) + R_{10}(\underline{J}) \cos\left(\frac{1}{2}\vartheta_2\right) \\ &+ R_{11}(\underline{J}) \cos\left(\frac{1}{2}\vartheta_1 + \frac{1}{4}\vartheta_2\right) + R_{12}(\underline{J}) \cos(\vartheta_1) \\ &+ R_{22}(\underline{J}) \cos\left(\vartheta_1 + \frac{1}{2}\vartheta_2\right) + R_{32}(\underline{J}) \cos(\vartheta_1 + \vartheta_2) \\ &+ R_{33}(\underline{J}) \cos\left(\frac{3}{2}\vartheta_1 + \frac{3}{4}\vartheta_2\right) + R_{44}(\underline{J}) \cos(2\vartheta_1 + \vartheta_2) \\ &+ R_{55}(\underline{J}) \cos\left(\frac{5}{2}\vartheta_1 + \frac{5}{4}\vartheta_2\right) + \dots \end{aligned}$$

with the coefficients R_{ij} as in (11). Let us decompose the perturbation as $R = \overline{R}(\underline{J}) + R_r(\underline{J}, \vartheta_1) + R_n(\underline{J}, \underline{\vartheta})$, where $\overline{R}(\underline{J})$ is the average over the angles, $R_r(\underline{J}, \vartheta_1) = R_{12}(\underline{J}) \cos(\vartheta_1)$ is the resonant part, while R_n contains all the remaining non-resonant terms. We look for a canonical transformation close to the identity with generating function $\Phi = \Phi(\underline{J}', \underline{\vartheta})$ such that

$$\underline{\omega}'(\underline{J}') \cdot \frac{\partial \Phi(\underline{J}', \underline{\vartheta})}{\partial \underline{\vartheta}} = -R_n(\underline{J}', \underline{\vartheta}) ,$$

which is well defined since $\underline{\omega}'$ is non-resonant for the Fourier components appearing in R_n. Finally, according to (14) the new unperturbed Hamiltonian is given by

$$h'(\underline{J}', \vartheta_1') \equiv h(\underline{J}') + \varepsilon R_{00}(\underline{J}') + \varepsilon R_{12}(\underline{J}') \cos \vartheta_1' .$$

Degenerate Perturbation Theory

A special case of resonant perturbation theory is obtained when considering a degenerate Hamiltonian function with n degrees of freedom of the form

$$H(\underline{I}, \underline{\varphi}) = h(I_1, \dots, I_d) + \varepsilon f(\underline{I}, \underline{\varphi}), \qquad d < n ; \quad (16)$$

notice that the integrable part depends on a subset of the action variables, being degenerate in $I_{d+1}, \dots, I_n$. In this case we look for a canonical transformation $C\colon (\underline{I}, \underline{\varphi}) \to (\underline{I}', \underline{\varphi}')$ such that the transformed Hamiltonian becomes

$$H'(\underline{I}', \underline{\varphi}') = h'(\underline{I}') + \varepsilon h_1'(\underline{I}, \varphi_{d+1}', \dots, \varphi_n') + \varepsilon^2 f'(\underline{I}', \underline{\varphi}'), \quad (17)$$

where the part $h' + \varepsilon h_1'$ admits d integrals of motion. Let us decompose the perturbing function in (16) as

$$f(\underline{I}, \underline{\varphi}) = \overline{f}(\underline{I}) + f_d(\underline{I}, \varphi_{d+1}, \dots, \varphi_n) + \tilde{f}(\underline{I}, \underline{\varphi}) , \quad (18)$$

where $\overline{f}$ is the average over the angle variables, f_d is independent on $\varphi_1, \dots, \varphi_d$ and $\tilde{f}$ is the remainder. As in the previous sections we want to determine a near-to-identity canonical transformation $\Phi = \Phi(\underline{I}', \underline{\varphi})$ of the form (3), such that in view of (18) the Hamiltonian (16) takes the form (17). One obtains

$$\begin{aligned} &h(I_1', \dots, I_d') + \varepsilon \sum_{k=1}^{d} \frac{\partial h}{\partial I_k} \frac{\partial \Phi}{\partial \varphi_k} + \varepsilon \overline{f}(I') \\ &\quad + \varepsilon f_d(\underline{I}', \varphi_{d+1}, \dots, \varphi_n) + \varepsilon \tilde{f}(\underline{I}', \underline{\varphi}) + O(\varepsilon^2) \\ &= h'(\underline{I}') + \varepsilon h_1'(\underline{I}', \varphi_{d+1}, \dots, \varphi_n) + O(\varepsilon^2) , \end{aligned}$$

where

$$\begin{aligned} h'(\underline{I}') &= h(I_1', \dots, I_d') + \varepsilon \overline{f}(I') \\ h_1'(\underline{I}', \varphi_{d+1}, \dots, \varphi_n) &= f_d(\underline{I}', \varphi_{d+1}, \dots, \varphi_n) , \end{aligned}$$

while Φ is determined solving the equation

$$\sum_{k=1}^{d} \frac{\partial h}{\partial I_k} \frac{\partial \Phi}{\partial \varphi_k} + \tilde{f}(\underline{I}', \underline{\varphi}) = 0 .$$

Expanding Φ and $\tilde{f}$ in Fourier series as in (5) one obtains that Φ is given by (6) where $\underline{\omega} \cdot \underline{m} = \sum_{k=1}^{d} m_k \omega_k$, being $\omega_k = 0$ for $k = d+1, \dots, n$. The generating function is well defined provided that $\underline{\omega} \cdot \underline{m} \neq 0$ for any $\underline{m} \in \mathcal{I}$, which is equivalent to requiring that

$$\sum_{k=1}^{d} m_k \omega_k \neq 0 \qquad \text{for } \underline{m} \in \mathcal{I} .$$

The Precession of the Equinoxes

An example of the application of the degenerate perturbation theory in Celestial Mechanics is provided by the computation of the precession of the equinoxes.

We consider a triaxial rigid body moving in the gravitational field of a primary body. We introduce the following reference frames with a common origin in the barycenter of the rigid body: $(O, \underline{i}_1^{(i)}, \underline{i}_2^{(i)}, \underline{i}_3^{(i)})$ is an inertial reference frame, $(O, \underline{i}_1^{(b)}, \underline{i}_2^{(b)}, \underline{i}_3^{(b)})$ is a body frame oriented along the direction of the principal axes of the ellipsoid, $(O, \underline{i}_1^{(s)}, \underline{i}_2^{(s)}, \underline{i}_3^{(s)})$ is the spin reference frame with the vertical axis along the direction of the angular momentum. Let (J, g, ℓ) be the Euler angles formed by the body and spin

frames, and let $(K, h, 0)$ be the Euler angles formed by the spin and inertial frames. The angle K is the obliquity (representing the angle between the spin and inertial vertical axes), while J is the non-principal rotation angle (representing the angle between the spin and body vertical axes).

This problem is conveniently described in terms of the following set of action-angle variables introduced by Andoyer in [1] (see also [17]). Let $\underline{M}_0$ be the angular momentum and let $M_0 \equiv |\underline{M}_0|$; the action variables are defined as

$$\begin{aligned} G &\equiv \underline{M}_0 \cdot \underline{i}_3^{(s)} = M_0 \\ L &\equiv \underline{M}_0 \cdot \underline{i}_3^{(b)} = G \cos J \\ H &\equiv \underline{M}_0 \cdot \underline{i}_3^{(i)} = G \cos K \,, \end{aligned}$$

while the corresponding angle variables are the quantities (g, ℓ, h) introduced before.

We limit ourselves to consider the gyroscopic case in which $I_1 = I_2 < I_3$ are the principal moments of inertia of the rigid body E around the primary S; let m_E and m_S be their masses and let $|E|$ be the volume of E. We assume that E orbits on a Keplerian ellipse around S with semi-major axis a and eccentricity e, while λ_E and $\underline{r}_E$ denote the longitude and instantaneous orbital radius (due to the assumption of Keplerian motion λ_E and $\underline{r}_E$ are known functions of the time). The Hamiltonian describing the motion of E around S is given by [15]

$$\begin{aligned} &\mathcal{H}(L, G, H, \ell, g, h, t) \\ &\quad = \frac{G^2}{2I_1} + \frac{I_1 - I_3}{2I_1 I_3} L^2 + V(L, G, H, \ell, g, h, t) \,, \end{aligned}$$

where the perturbation is implicitly defined by

$$V \equiv -\int_E \frac{\tilde{G} m_S m_E}{|\underline{r}_E + \underline{x}|} \frac{\mathrm{d}\underline{x}}{|E|} \,,$$

$\tilde{G}$ being the gravitational constant. Setting $r_E = |\underline{r}_E|$ and $x = |\underline{x}|$, we can expand V using the Legendre polynomials as

$$\begin{aligned} V = -\frac{\tilde{G} m_S m_E}{r_E} \int_E \frac{\mathrm{d}\underline{x}}{|E|} \Bigg[1 - \frac{\underline{x} \cdot \underline{r}_E}{r_E^2} + \frac{1}{2 r_E^2} \\ \cdot \left(3 \frac{(\underline{x} \cdot \underline{r}_E)^2}{r_E^2} - x^2 \right) \Bigg] + O\left(\left(\frac{\underline{x}}{\underline{r}_E} \right)^3 \right) . \end{aligned}$$

We further assume that $J = 0$ (i. e. $G = L$) so that E rotates around a principal axis. Let G_0 and H_0 be the initial values of G and H; if α denotes the angle between $\underline{r}_E$ and $\underline{i}_3^{(b)}$, the perturbing function can be written as

$$V = \frac{3}{2} \eta \omega \frac{G_0^2}{H_0} \frac{(1 - \mathrm{e} \cos \lambda_E)^3}{(1 - \mathrm{e}^2)^3} \cos^2 \alpha$$

with $\eta = (I_3 - I_1)/I_3$ and $\omega = (\tilde{G} m_S)/a^3 I_3 H_0 / G_0^2$. Elementary computations show that

$$\cos \alpha = \sin(\lambda_E - h) \sqrt{1 - \frac{H^2}{G^2}} \,.$$

Neglecting first order terms in the eccentricity, we approximate $(1 - e \cos \lambda_E)^3/(1 - e^2)^3$ with one. A first order degenerate perturbation theory provides that the new unperturbed Hamiltonian is given by

$$\mathcal{K}(G, H) = \frac{G^2}{2I_3} + \frac{3}{2} \eta \omega \frac{G_0^2}{H_0} \frac{G^2 - H^2}{2G^2} \,.$$

Therefore the average angular velocity of precession is given by

$$\dot{h} = \frac{\partial \mathcal{K}(G, H)}{\partial H} = -\frac{3}{2} \eta \omega \frac{G_0^2}{H_0} \frac{H}{G^2} \,.$$

At $t = 0$ it is

$$\dot{h} = -\frac{3}{2} \eta \omega = -\frac{3}{2} \eta \omega_y^2 \omega_d^{-1} \cos K \,, \tag{19}$$

where we used $\omega = \omega_y^2 \omega_d^{-1} \cos K$ with ω_y being the frequency of revolution and ω_d the frequency of rotation.

In the case of the Earth, the astronomical measurements show that $\eta = 1/298.25$, $K = 23.45°$. The contribution due to the Sun is thus obtained by inserting $\omega_y = 1$ year, $\omega_d = 1$ day in (19), which yields $\dot{h}^{(S)} = -2.51857 \cdot 10^{-12}$ rad/sec, corresponding to a retrograde precessional period of 79 107.9 years. A similar computation shows that the contribution of the Moon amounts to $\dot{h}^{(L)} = -5.49028 \cdot 10^{-12}$ rad/sec, corresponding to a precessional period of 36 289.3 years. The total amount is obtained as the sum of $\dot{h}^{(S)}$ and $\dot{h}^{(L)}$, providing an overall retrograde precessional period of 24 877.3 years.

Invariant Tori

Invariant KAM Surfaces

We consider an n-dimensional nearly-integrable Hamiltonian function

$$H(\underline{I}, \underline{\varphi}) = h(\underline{I}) + \varepsilon f(\underline{I}, \underline{\varphi}) \,,$$

defined in a $2n$-dimensional phase space $\mathcal{M} \equiv V \times \mathbf{T}^n$, where V is an open bounded region of $\mathbf{R}^n$. A *KAM torus* associated to H is an n-dimensional invariant surface on which the flow is described parametrically by a coordinate $\underline{\theta} \in \mathbf{T}^n$ such that the *conjugated flow* is linear, namely

$\underline{\theta} \in \mathbf{T}^n \to \underline{\theta} + \underline{\omega} t$ where $\underline{\omega} \in \mathbf{R}^n$ is a Diophantine vector, i. e. there exist $\gamma > 0$ and $\tau > 0$ such that

$$|\underline{\omega} \cdot \underline{m}| \geq \frac{\gamma}{|\underline{m}|^\tau}\,, \qquad \forall \underline{m} \in \mathbf{Z}^n \backslash \{0\}\,.$$

Kolmogorov's theorem [26] (see also "Kolmogorov–Arnol'd–Moser (KAM Theory)") ensures the persistence of invariant tori with diophantine frequency, provided ε is sufficiently small and provided the unperturbed Hamiltonian is non-degenerate, i. e. for a given torus $\{\underline{I}_0\} \times \mathbf{T}^n \subset \mathcal{M}$

$$\det h''(\underline{I}_0) \equiv \det\left(\frac{\partial^2 h}{\partial I_i \partial I_j}(\underline{I}_0)\right)_{i,j=1,\ldots,n} \neq 0\,. \tag{20}$$

The condition (20) can be replaced by the isoenergetic non-degeneracy condition introduced by Arnold [2]

$$\det\begin{pmatrix} h''(\underline{I}_0) & h'(\underline{I}_0) \\ h'(\underline{I}_0) & \underline{0} \end{pmatrix} \neq 0\,, \tag{21}$$

which ensures the existence of KAM tori on the energy level corresponding to the unperturbed energy $h(\underline{I}_0)$, say $\mathcal{M}_0 \equiv \{(\underline{I}, \underline{\varphi}) \in \mathcal{M} : H(\underline{I}, \underline{\varphi}) = h(\underline{I}_0)\}$. In the context of the n-body problem Arnold [2] addressed the question of the existence of a set of initial conditions with positive measure such that, if the initial position and velocities of the bodies belong to this set, then the mutual distances remain perpetually bounded. A positive answer is provided by Kolmogorov's theorem in the framework of the planar, circular, restricted three-body problem, since the integrable part of the Hamiltonian (8) satisfies the isoenergetic non-degeneracy condition (21); denoting the initial values of the Delaunay's action variables by (L_0, G_0), if ε is sufficiently small, there exist KAM tori for (8) on the energy level $\mathcal{M}_0 \equiv \{H_{3D} = -1/(2L_0^2) - G_0\}$. In particular, the motion of the perturbed body remains forever bounded from the orbits of the primaries. Indeed, a stronger statement is also valid: due to the fact that the two-dimensional KAM surfaces separate the three dimensional energy levels, any trajectory starting between two KAM tori remains forever trapped in the region between such tori.

In the framework of the three-body problem, Arnold [2] stated the following result: "*If the masses, eccentricities and inclinations of the planets are sufficiently small, then for the majority of initial conditions the true motion is conditionally periodic and differs little from Lagrangian motion with suitable initial conditions throughout an infinite interval time* $-\infty < t < \infty$". Arnold provided a complete proof for the case of three coplanar bodies, while the spatial three-body problem was investigated by Laskar and Robutel in [27,35] using Poincaré variables, the Jacobi's "reduction of the nodes" (see, e. g., [11]) and Birkhoff's normal form [3,4,38]. The full proof of Arnold's theorem was provided in [19], based on Herman's results on the planetary problem; it makes use of Poincaré variables restricted to the symplectic manifold of vertical total angular momentum.

Explicit estimates on the perturbing parameter ensuring the existence of KAM tori were given by M. Hénon [25]; he showed that direct applications of KAM theory to the three-body problem lead to analytical results which are much smaller than the astronomical observations. For example, the application of Arnold's theorem to the restricted three-body problem is valid provided the mass-ratio of the primaries is less than 10^{-333}. This result can be improved up to 10^{-48} by applying Moser's theorem, but it is still very far from the actual Jupiter-Sun mass-ratio which amounts to about 10^{-3}. In the context of concrete estimates, a big improvement comes from the synergy between KAM theory and computer-assisted proofs, based on the application of *interval arithmetic* which allows to keep rigorously track of the rounding-off and propagation errors introduced by the machine. Computer-assisted KAM estimates were implemented in a number of cases in Celestial Mechanics, like the three-body problem and the spin-orbit model as briefly recalled in the following subsections.

Another interesting example of the interaction between the analytical theory and the computer implementation is provided by the analysis of the stability of the triangular Lagrangian points; in particular, the stability for exponentially long times is obtained using Nekhoroshev theory combined with computer-assisted implementations of Birkhoff normal form (see, e. g., [5,13,18,21,22,23,28,36]).

Rotational Tori for the Spin-Orbit Problem

We study the motion of a rigid triaxial satellite around a central planet under the following assumptions [8]:

i) The orbit of the satellite is Keplerian,
ii) The spin-axis is perpendicular to the orbital plane,
iii) The spin-axis coincides with the smallest physical axis,
iv) External perturbations as well as dissipative forces are neglected.

Let $I_1 < I_2 < I_3$ be the principal moments of inertia; let a, e be the semi-major axis and eccentricity of the Keplerian ellipse; let r and f be the instantaneous orbital radius and the true anomaly of the satellite; let x be the angle between the longest axis of the triaxial satellite and the periapsis line. The equation of motion governing the spin-

orbit model is given by:

$$\ddot{x} + \frac{3}{2}\frac{I_2 - I_1}{I_3}\left(\frac{a}{r}\right)^3 \sin(2x - 2f) = 0\,. \tag{22}$$

Due to *assumption* i, the quantities r and f are known functions of the time. Expanding the second term of (22) in Fourier–Taylor series and neglecting terms of order 6 in the eccentricity, setting $y \equiv \dot{x}$ one obtains that the equation of motion corresponds to Hamilton's equations associated to the Hamiltonian

$$\begin{aligned} H(y,x,t) \equiv \frac{y^2}{2} - \varepsilon\Bigg[& \left(-\frac{e}{4} + \frac{e^3}{32} - \frac{5}{768}e^5\right)\cos(2x - t) \\ & + \left(\frac{1}{2} - \frac{5}{4}e^2 + \frac{13}{32}e^4\right)\cos(2x - 2t) \\ & + \left(\frac{7}{4}e - \frac{123}{32}e^3 + \frac{489}{256}e^5\right)\cos(2x - 3t) \\ & + \left(\frac{17}{4}e^2 - \frac{115}{12}e^4\right)\cos(2x - 4t) \\ & + \left(\frac{845}{96}e^3 - \frac{32525}{1536}e^5\right)\cos(2x - 5t) \\ & + \frac{533}{32}e^4\cos(2x - 6t) \\ & + \frac{228347}{7680}e^5\cos(2x - 7t)\Bigg]\,, \end{aligned} \tag{23}$$

where $\varepsilon \equiv 3/2(I_2 - I_1)/I_3$ and we have chosen the units so that $a = 1$, $2\pi/T_{\text{rev}} = 1$, where T_{rev} is the period of revolution. Let p, q be integers with $q \neq 0$; a $p:q$ resonance occurs whenever $\langle \dot{x} \rangle = \frac{p}{q}$, meaning that during q orbital revolutions, the satellite makes on average p rotations. Since the phase-space is three-dimensional, the two-dimensional KAM tori separates the phase-space into invariant regions, thus providing the stability of the trapped orbits. In particular, let $\mathcal{P}(\frac{p}{q})$ be the periodic orbit associated to the $p:q$ resonance; its stability is guaranteed by the existence of trapping rotational tori with frequencies $\mathcal{T}(\omega_1)$ and $\mathcal{T}(\omega_2)$ with $\omega_1 < p/q < \omega_2$. For example, one can consider the sequences of irrational rotation numbers

$$\Gamma_k^{(p/q)} \equiv \frac{p}{q} - \frac{1}{k+\alpha}\,, \quad \Delta_k^{(p/q)} \equiv \frac{p}{q} + \frac{1}{k+\alpha}\,, \qquad k \in \mathbf{Z}, k \geq 2$$

with $\alpha \equiv (\sqrt{5} - 1)/2$. In fact, the continued fraction expansion of $1/(k+\alpha)$ is given by $1/(k+\alpha) = [0, k, 1^\infty]$. Therefore, both $\Gamma_k^{(p/q)}$ and $\Delta_k^{(p/q)}$ are *noble* numbers (i. e. with continued fraction expansion definitely equal to one); by number theory they satisfy the diophantine condition and bound $\frac{p}{q}$ from below and above.

As a concrete sample we consider the synchronous spin-orbit resonance ($p = q = 1$) of the Moon, whose physical values of the parameters are $\varepsilon \equiv 3.45 \cdot 10^{-4}$ and $e = 0.0549$. The stability of the motion is guaranteed by the existence of the surfaces $\mathcal{T}(\Gamma_{40}^{(1)})$ and $\mathcal{T}(\Delta_{40}^{(1)})$, which is obtained by implementing a computer-assisted KAM theory for the realistic values of the parameters. The result provides the confinement of the synchronous periodic orbit in a limited region of the phase space [8].

Librational Tori for the Spin-Orbit Problem

The existence of invariant librational tori around a spin-orbit resonance can be obtained as follows [9]. Let us consider the 1:1 resonance corresponding to Hamilton's equations associated to (23). First one implements a canonical transformation to center around the synchronous periodic orbit; after expanding in Taylor series, one diagonalizes the quadratic terms, thus obtaining a harmonic oscillator plus higher degree (time-dependent) terms. Finally, it is convenient to transform the Hamiltonian using the action-angle variables (I, φ) of the harmonic oscillator. After these symplectic changes of variables one is led to a Hamiltonian of the form

$$H(I,\varphi,t) \equiv \omega I + \varepsilon\overline{h}(I) + \varepsilon R(I,\varphi,t), \qquad I \in \mathbf{R}, (\varphi,t) \in T^2\,,$$

where $\omega \equiv \omega(\varepsilon)$ is the frequency of the harmonic oscillator, while $\overline{h}(I)$ and $R(I,\varphi,t)$ are suitable functions, precisely polynomials in the action (of order of the square of the action). Then apply a Birkhoff normal form (see –Normal Forms in Pertubation Theory–) up to the order k ($k = 5$ in [9]) to obtain the following Hamiltonian:

$$H_k(I',\varphi',t) = \omega I' + \varepsilon h_k(I';\varepsilon) + \varepsilon^{k+1}R_k(I',\varphi',t)\,.$$

Finally, implementing a computer-assisted KAM theorem one gets the following result: consider the Moon–Earth case with $\varepsilon_{\text{obs}} = 3.45 \cdot 10^{-4}$ and $e = 0.0549$; there exists an invariant torus around the synchronous resonance corresponding to a libration of 8.79° for any $\varepsilon \leq \varepsilon_{\text{obs}}/5.26$. The same strategy applied to different samples, e. g. the Rhea–Saturn pair, allows one to prove the existence of librational invariant tori around the synchronous resonance for values of the parameters in full agreement with the observational measurements [9].

Rotational Tori for the Restricted Three-Body Problem

The planar, circular, restricted three-body problem has been considered in [12], where the stability of the asteroid 12 Victoria has been investigated under the gravitational influence of the Sun and Jupiter. On a fixed energy level invariant KAM tori trapping the motion of Victoria have been established for the astronomical value of the Jupiter–Sun mass-ratio (about 10^{-3}). After an expansion of the perturbing function and a truncation to a suitable order (see [12]), the Hamiltonian function describing the motion of the asteroid is given in Delaunay's variables by

$$H(L,G,\ell,g) \equiv -\frac{1}{2L^2} - G - \varepsilon f(L,G,\ell,g)\,,$$

where setting $a \equiv L^2$, $e = \sqrt{1-G^2/L^2}$, the perturbation is given by

$$\begin{aligned} f(L,G,\ell,g) = 1 &+ \frac{a^2}{4} + \frac{9}{64}a^4 + \frac{3}{8}a^2e^2 \\ &- \left(\frac{1}{2} + \frac{9}{16}a^2\right)a^2e\cos\ell \\ &+ \left(\frac{3}{8}a^3 + \frac{15}{64}a^5\right)\cos(\ell+g) \\ &- \left(\frac{9}{4} + \frac{5}{4}a^2\right)a^2e\cos(\ell+2g) \\ &+ \left(\frac{3}{4}a^2 + \frac{5}{16}a^4\right)\cos(2\ell+2g) \\ &+ \frac{3}{4}a^2e\cos(3\ell+2g) \\ &+ \left(\frac{5}{8}a^3 + \frac{35}{128}a^5\right)\cos(3\ell+3g) \\ &+ \frac{35}{64}a^4\cos(4\ell+4g) \\ &+ \frac{63}{128}a^5\cos(5\ell+5g)\,. \end{aligned}$$

For the asteroid Victoria the orbital elements are $a_V \simeq 2.334$ AU, $e_V \simeq 0.220$, which give the observed values of the Delaunay's action variables as $L_V = 0.670$, $G_V = 0.654$. The energy level is taken as

$$\begin{aligned} E_V^{(0)} &\equiv -\frac{1}{2L_{V^2}} - G_V \simeq -1.768\,, \\ E_V^{(1)} &\equiv -\langle f(L_V,G_V,\ell,g)\rangle \simeq -1.060\,, \\ E_V(\varepsilon) &\equiv E_V^{(0)} + \varepsilon E_V^{(1)}\,. \end{aligned}$$

The osculating energy level of the Sun–Jupiter–Victoria model is defined as

$$E_V^* \equiv E_V(\varepsilon_J) = E_V^{(0)} + \varepsilon_J E_V^{(1)} \simeq -1.769\,.$$

We now look for two invariant tori bounding the observed values of L_V and G_V. To this end, let $\tilde{L}_\pm = L_V \pm 0.001$ and let

$$\underline{\tilde{\omega}}_\pm = \left(\frac{1}{\tilde{L}_\pm^3}, -1\right) \equiv (\tilde{\alpha}_\pm, -1)\,.$$

To obtain diophantine frequencies, the continued fraction expansion of $\tilde{\alpha}_\pm$ is modified adding a tail of ones after the order 5; this procedure gives the diophantine numbers $\alpha_\pm$ which define the bounding frequencies as $\underline{\omega}_\pm = (\alpha_\pm, -1)$. By a computer-assisted KAM theorem, the stability of the asteroid Victoria is a consequence of the following result [12]: for $|\varepsilon| \le 10^{-3}$ the unperturbed tori can be analytically continued into invariant KAM tori for the perturbed system on the energy level $H^{-1}(E_V(\varepsilon))$, keeping fixed the ratio of the frequencies. Therefore the orbital elements corresponding to the semi-major axis and to the eccentricity of the asteroid Victoria stay forever ε-close to their unperturbed values.

Planetary Problem

The dynamics of the planetary problem composed by the Sun, Jupiter and Saturn is investigated in [29,30,31]. In [29] the secular dynamics of the following model is studied: after the Jacobi's reduction of the nodes, the 4-dimensional Hamiltonian is averaged over the fast angles and its series expansion is considered up to the second order in the masses. This procedure provides a Hamiltonian function with two degrees of freedom, describing the slow motion of the parameters characterizing the Keplerian approximation (i. e., the eccentricities and the arguments of perihelion). Afterwards, action-angle coordinates are introduced and a partial Birkhoff normalization is performed. Finally, a computer-assisted implementation of a KAM theorem yields the existence of two invariant tori bounding the secular motions of Jupiter and Saturn for the observed values of the parameters.

The approach sketched above is extended in [31] so to include the description of the fast variables, like the semi-major axes and the mean longitudes of the planets. Indeed, the preliminary average on the fast angles is now performed without eliminating the terms with degree greater or equal than two with respect to the fast actions. The canonical transformations involving the secular coordinates can be adapted to produce a good initial approximation of an invariant torus for the reduced Hamiltonian of the three-body planetary problem. This is the starting point of the procedure for constructing the Kolmogorov's normal form which is numerically shown to be convergent. In [30] the same result of [31] has been obtained for

a fictitious planetary solar system composed by two planets with masses equal to 1/10 of those of Jupiter and Saturn.

Periodic Orbits

Construction of Periodic Orbits

One of the most intriguing conjectures of Poincaré concerns the pivotal role of the periodic orbits in the study of the dynamics; more precisely, he states that given a particular solution of Hamilton's equations one can always find a periodic solution (possibly with very long period) such that the difference between the two solutions is small for an arbitrary long time. The literature on periodic orbits is extremely wide (see, e. g., [4,7,24,38,39] and references therein); here we present the construction of periodic orbits implementing a perturbative approach (see [10]) as shown by Poincaré in [34]. We describe such a method taking as an example the spin-orbit Hamiltonian (23) that we write in a compact form as $H(y, x, t) \equiv y^2/2 - \varepsilon f(x, t)$ for a suitable function $f = f(x, t)$; the corresponding Hamilton's equations are

$$\begin{aligned} \dot{x} &= y \\ \dot{y} &= \varepsilon f_x(x, t) \,. \end{aligned} \tag{24}$$

A spin-orbit resonance of order $p : q$ is a periodic solution of period $T = 2\pi q$ ($q \in \mathbf{Z}\backslash\{0\}$), such that

$$\begin{aligned} x(t + 2\pi q) &= x(t) + 2\pi p \\ y(t + 2\pi q) &= y(t) \,. \end{aligned} \tag{25}$$

From (24) the solution can be written in integral form as

$$\begin{aligned} y(t) &= y(0) + \varepsilon \int_0^t f_x(x(s), s)\mathrm{d}s \\ x(t) &= x(0) + y(0)t + \varepsilon \int_0^t \int_0^\tau f_x(x(s), s)\mathrm{d}s \, \mathrm{d}\tau \\ &= x(0) + \int_0^t y(s)\mathrm{d}s \,; \end{aligned}$$

combining the above equations with (25) one obtains

$$\begin{aligned} \int_0^{2\pi q} f_x(x(s), s)\mathrm{d}s &= 0 \\ \int_0^{2\pi q} y(s)\mathrm{d}s - 2\pi p &= 0 \,. \end{aligned} \tag{26}$$

Let us write the solution as the series

$$\begin{aligned} x(t) &\equiv \overline{x} + \overline{y}t + \varepsilon x_1(t) + \ldots \\ y(t) &\equiv \overline{y} + \varepsilon y_1(t) + \ldots \,, \end{aligned} \tag{27}$$

where $x(0) = \overline{x}$ and $y(0) = \overline{y}$ are the initial conditions, while $x_1(t)$, $y_1(t)$ are the first order terms in ε. Expanding the initial conditions in power series of ε, one gets:

$$\begin{aligned} \overline{x} &= \overline{x}_0 + \varepsilon\overline{x}_1 + \varepsilon^2\overline{x}_2 + \ldots \\ \overline{y} &= \overline{y}_0 + \varepsilon\overline{y}_1 + \varepsilon^2\overline{y}_2 + \ldots \end{aligned} \tag{28}$$

Inserting (27) and (28) in (24), equating same orders in ε and taking into account the periodicity condition (26), one can find the following explicit expressions for $x_1(t)$, $y_1(t)$, $\overline{y}_0$, $\overline{y}_1$:

$$\begin{aligned} y_1(t) &= y_1(t; \overline{y}, \overline{x}) = \int_0^t f_x(\overline{x}_0 + \overline{y}_0 s, s)\mathrm{d}s \\ x_1(t) &= x_1(t; \overline{y}, \overline{x}) = \int_0^t y_1(s)\mathrm{d}s \\ \overline{y}_0 &= \frac{p}{q} \\ \overline{y}_1 &= -\frac{1}{2\pi q} \int_0^{2\pi q} \int_0^t f_x(\overline{x}_0 + \overline{y}_0 s, s)\mathrm{d}s\mathrm{d}t \,. \end{aligned}$$

Furthermore, $\overline{x}_0$ is determined as a solution of

$$\int_0^{2\pi q} f_x(\overline{x}_0 + \overline{y}_0 s, s)\mathrm{d}s = 0 \,,$$

while $\overline{x}_1$ is given by

$$\begin{aligned} \overline{x}_1 = &-\frac{1}{\int_0^{2\pi q} f_{xx}^0 \mathrm{d}t} \\ &\cdot \left[\overline{y}_1 \int_0^{2\pi q} t f_{xx}^0 \mathrm{d}t + \int_0^{2\pi q} f_{xx}^0 x_1(t)\mathrm{d}t \right] , \end{aligned}$$

where, for shortness, we have written $f_{xx}^0 = f_{xx}(\overline{x}_0 + \overline{y}_0 t, t)$.

The Libration in Longitude of the Moon

The previous computation of the $p : q$ periodic solution can be used to evaluate the libration in longitude of the Moon. More precisely, setting $p = q = 1$ one obtains

$$\begin{aligned} \overline{x}_0 &= 0 \\ \overline{y}_0 &= 1 \\ x_1(t) &= 0.232086t - 0.218318\sin(t) \\ &\quad - 6.36124 \cdot 10^{-3}\sin(2t) - 3.21314 \cdot 10^{-4}\sin(3t) \\ &\quad - 1.89137 \cdot 10^{-5}\sin(4t) - 1.18628 \cdot 10^{-6}\sin(5t) \\ y_1(t) &= 0.232086 - 0.218318\cos(t) - 0.0127225\cos(2t) \\ &\quad - 9.63942 \cdot 10^{-4}\cos(3t) - 7.56548 \cdot 10^{-5}\cos(4t) \\ &\quad - 5.93138 \cdot 10^{-6}\cos(5t) \\ \overline{x}_1 &= 0 \\ \overline{y}_1 &= -0.232086 \,, \end{aligned}$$

where we used $e = 0.05494$, $\varepsilon = 3.45 \cdot 10^{-4}$. Therefore the synchronous periodic solution, computed up to the first order in ε, is given by

$$\begin{aligned} x(t) &= \overline{x}_0 + \overline{y}_0 t + \varepsilon x_1(t) = t - 7.53196 \cdot 10^{-5} \sin(t) \\ &\quad - 2.19463 \cdot 10^{-6} \sin(2t) - 1.10853 \cdot 10^{-7} \sin(3t) \\ &\quad - 6.52523 \cdot 10^{-9} \sin(4t) - 4.09265 \cdot 10^{-10} \sin(5t) \\ y(t) &= \overline{y}_0 t + \varepsilon y_1(t) = 1 - 7.53196 \cdot 10^{-5} \cos(t) \\ &\quad - 4.38926 \cdot 10^{-6} \cos(2t) - 3.3256 \cdot 10^{-7} \cos(3t) \\ &\quad - 2.61009 \cdot 10^{-8} \cos(4t) - 2.04633 \cdot 10^{-9} \cos(5t) \,. \end{aligned}$$

It turns out that the libration in longitude of the Moon, provided by the quantity $x(t) - t$, is of the order of $7 \cdot 10^{-5}$ in agreement with the observational data.

Future Directions

The last decade of the XX century has been greatly marked by astronomical discoveries, which changed the shape of the solar system as well as of the entourage of other stars. In particular, the detection of many small bodies beyond the orbit of Neptune has moved the edge of the solar system forward and it has increased the number of its population. Hundreds objects have been observed to move in a ring beyond Neptune, thus forming the so-called Kuiper's belt. Its components show a great variety of behaviors, like resonance clusterings, regular orbits, scattered trajectories. Furthermore, far outside the solar system, the astronomical observations of extrasolar planetary systems have opened new scenarios with a great variety of dynamical behaviors. In these contexts classical and resonant perturbation theories will deeply contribute to provide a fundamental insight of the dynamics and will play a prominent role in explaining the different configurations observed within the Kuiper's belt as well as within extrasolar planetary systems.

Bibliography

1. Andoyer H (1926) Mécanique Céleste. Gauthier-Villars, Paris
2. Arnold VI (1963) Small denominators and problems of stability of motion in classical and celestial mechanics. Uspehi Mat Nauk 6 18(114):91–192
3. Arnold VI (1978) Mathematical methods of classical mechanics. Springer, Berlin
4. Arnold VI (ed) (1988) Encyclopedia of Mathematical Sciences. Dynamical Systems III. Springer, Berlin
5. Benettin G, Fasso F, Guzzo M (1998) Nekhoroshev-stability of L_4 and L_5 in the spatial restricted three-body problem. Regul Chaotic Dyn 3(3):56–71
6. Boccaletti D, Pucacco G (2001) Theory of orbits. Springer, Berlin
7. Brouwer D, Clemence G (1961) Methods of Celestial Mechanics. Academic Press, New York
8. Celletti A (1990) Analysis of resonances in the spin-orbit problem. In: Celestial Mechanics: The synchronous resonance (Part I). J Appl Math Phys (ZAMP) 41:174–204
9. Celletti A (1993) Construction of librational invariant tori in the spin-orbit problem. J Appl Math Phys (ZAMP) 45:61–80
10. Celletti A, Chierchia L (1998) Construction of stable periodic orbits for the spin-orbit problem of Celestial Mechanics. Regul Chaotic Dyn (Editorial URSS) 3:107–121
11. Celletti A, Chierchia L (2006) KAM tori for N-body problems: a brief history. Celest Mech Dyn Astron 95 1:117–139
12. Celletti A, Chierchia L (2007) KAM Stability and Celestial Mechanics. Mem Am Math Soc 187:878
13. Celletti A, Giorgilli A (1991) On the stability of the Lagrangian points in the spatial restricted problem of three bodies. Celest Mech Dyn Astron 50:31–58
14. Chebotarev AG (1967) Analytical and Numerical Methods of Celestial Mechanics. Elsevier, New York
15. Chierchia L, Gallavotti G (1994) Drift and diffusion in phase space. Ann l'Inst H Poincaré 60:1–144
16. Delaunay C (1867) Mémoire sur la théorie de la Lune. Mém l'Acad Sci 28:29 (1860)
17. Deprit A (1967) Free rotation of a rigid body studied in the phase space. Am J Phys 35:424–428
18. Efthymiopoulos C, Sandor Z (2005) Optimized Nekhoroshev stability estimates for the Trojan asteroids with a symplectic mapping model of co-orbital motion. MNRAS 364(6):253–271
19. Féjoz J (2004) Démonstration du "théorème d'Arnold" sur la stabilité du système planétaire (d'après Michael Herman). Ergod Th Dynam Syst 24:1–62
20. Ferraz-Mello S (2007) Canonical Perturbation Theories. Springer, Berlin
21. Gabern F, Jorba A, Locatelli U (2005) On the construction of the Kolmogorov normal form for the Trojan asteroids. Nonlinearity 18:1705–1734
22. Giorgilli A, Skokos C (1997) On the stability of the trojan asteroids. Astron Astroph 317:254–261
23. Giorgilli A, Delshams A, Fontich E, Galgani L, Simó C (1989) Effective stability for a Hamiltonian system near an elliptic equilibrium point, with an application to the restricted three-body problem. J Diff Eq 77:167–198
24. Hagihara Y (1970) Celestial Mechanics. MIT Press, Cambridge
25. Hénon M (1966) Explorationes numérique du problème restreint IV: Masses egales, orbites non periodique. Bull Astron 3(1, fasc 2):49–66
26. Kolmogorov AN (1954) On the conservation of conditionally periodic motions under small perturbation of the Hamiltonian. Dokl Akad Nauk SSR 98:527–530
27. Laskar J, Robutel P (1995) Stability of the planetary three-body problem I Expansion of the planetary Hamiltonian. Celest Mech and Dyn Astron 62(3):193–217
28. Lhotka Ch, Efthymiopoulos C, Dvorak R (2008) Nekhoroshev stability at L_4 or L_5 in the elliptic restricted three body problem-application to Trojan asteroids. MNRAS 384:1165–1177
29. Locatelli U, Giorgilli A (2000) Invariant tori in the secular motions of the three-body planetary systems. Celest Mech and Dyn Astron 78:47–74
30. Locatelli U, Giorgilli A (2005) Construction of the Kolmogorov's normal form for a planetary system. Regul Chaotic Dyn 10:153–171
31. Locatelli U, Giorgilli A (2007) Invariant tori in the Sun–Jupiter–Saturn system. Discret Contin Dyn Syst-Ser B 7:377–398

32. Meyer KR, Hall GR (1991) Introduction to Hamiltonian dynamical systems and the *N*-body problem. Springer, Berlin
33. Moser J (1962) On invariant curves of area-preserving mappings of an annulus. Nach Akad Wiss Göttingen. Math Phys Kl II 1:1
34. Poincarè H (1892) Les Méthodes Nouvelles de la Mécanique Céleste. Gauthier-Villars, Paris
35. Robutel P (1995) Stability of the planetary three-body problem II KAM theory and existence of quasi-periodic motions. Celest Mech Dyn Astron 62(3):219–261
36. Robutel P, Gabern F (2006) The resonant structure of Jupiter's Trojan asteroids – I Long-term stability and diffusion. MNRAS 372(4):1463–1482
37. Sanders JA, Verhulst F (1985) Averaging methods in nonlinear dynamical systems. Springer, Berlin
38. Siegel CL, Moser JK (1971) Lectures on Celestial Mechanics. Springer, Heidelberg
39. Szebehely V (1967) Theory of orbits. Academic Press, New York

Perturbation Theory, Introduction to

Giuseppe Gaeta
Dipartimento di Matematica,
Università di Milano, Milan, Italy

The idea behind Perturbation Theory is that when we are not able to determine exact solutions to a given problem, we might be able to determine approximate solutions to our problem starting from solutions to an approximate version of the problem, amenable to exact treatment. Thus, in a way, we use *exact solutions to an approximate problem* to get *approximate solutions to an exact problem*.

It goes without saying that many mathematical problems met in realistic situations, in particular as soon as we leave the linear framework, are not exactly solvable–either for an inherent impossibility or for our insufficient skills. Thus, Perturbation Theory is often the only way to approach realistic nonlinear systems.

It is implicit in the very nature of perturbation theory that it can only worke once a problem which is both solvable–one also says "integrable"–and in some sense "*near*" to the original problem can be identified (it should be mentioned in this respect that the issue of "how near is near enough" is a delicate one).

Quite often, the integrable problem to be used as a starting point is a linear one–maybe obtained as the first-order expansion around a trivial or however known solution–and nonlinear corrections can be computed term by term via a recursive procedure based on expansion in a small parameter (usually denoted as ε by tradition); the point is that at each stage of this procedure one should only solve *linear* equations, so that the procedure can–at least in principle–be carried over up to any desired order. In practice, one is limited by time, computational power, and the increasing dimension of the linear systems to be solved.

But limitations are not only due to the limits of the humans–or the computers–performing the actual computations: in fact, some delicate points arise when one considers the convergence of the ε series involved in the computations and in the expression of the solutions obtained by Perturbation Theory.

These points–i. e. the power of Perturbation Theory, its basic features and tools, and its limitations in particular with regard to convergence issues–are discussed in the article ► Perturbation Theory by Gallavotti. This article also stresses the role which problems originating in Physics had in the development of Perturbation Theory; and this not only in historical terms (the computation of planetary orbits), but also in more recent times through the work of Poincaré first and then via Quantum Theory.

The modern setting of Perturbation Theory was laid down by Poincaré, and goes through the use of what is today known as *Poincaré normal forms*; these are a cornerstone of the whole theory and hence, implicitly or explicitly, of all the articles presented in this section of the Encyclopedia. But, they are also discussed in detail, together with their application, in the article ► Normal Forms in Perturbation Theory by Broer.

The latter deals with the general problem, i. e. with evolution differential equations (Dynamical Systems) with no special structure; or, in applications originating from Physics or Engineering one is often dealing with systems that (within a certain approximation) preserve Energy and can be written in *Hamiltonian* form. In this case, as emphasized by Birkhoff, one can more efficiently consider perturbations of the Hamiltonian rather than of the equations of motion (the advantage originating in the fact that the Hamiltonian is a scalar function, while the equations of motion are a system of $2n$ equations in $2n$ dimensions). The normal form approach for Hamiltonian systems, and more generally Hamiltonian perturbation theory, is discussed in the article ► Hamiltonian Perturbation Theory (and Transition to Chaos) by Broer and Hanßmann. This also discusses the problem of transition–as some control parameter, often the Energy, is varied–from the regular behavior of the unperturbed system to the chaotic ("turbulent" if we deal with fluid motion) behavior displayed by many relevant Hamiltonian as well as non-Hamiltonian systems.

As mentioned above, in all the matters connected with Perturbation Theory and its applications, convergence issues play an extremely important role. They are discussed in the article ► Perturbative Expansions, Convergence of

by Walcher, both in the general case and for Hamiltonian systems.

The interplay between perturbations–and more generally changes in some relevant parameter characterizing the system within a more general family of system–and qualitative (not only quantitative) changes in its behavior is of course of general interest not only in the "extreme" case of transition from integrable to chaotic behavior, but also when the qualitative change in the behavior of the system is somehow more moderate. Such a change is also known as a *bifurcation*. Albeit there is no article specifically devoted to these, the reader will note the concept of bifurcation appears in many, if not most, of the articles.

The behavior of a "generically perturbed" system depends on what is meant by "generically". In particular if we deal with an unperturbed system which has some degree of symmetry, this may be an "accidental" feature–maybe due to the specially simple nature of integrable systems such as the one chosen as an unperturbed one–but might also correspond to a requirement by the very problem we are modeling; this is often the case when we deal with problems of physical or engineering origin, just because the fundamental equations of Physics have some degree of symmetry. The presence of symmetry can be quite helpful–e. g. reducing the effective degrees of freedom of a given problem–and should be taken into account in the perturbative expansion. Moreover, the perturbative expansions can be made to have some degree of symmetry which can be used in the solution of the resulting equations. These matters are discussed at length in the article ► Non-linear Dynamics, Symmetry and Perturbation Theory in by Gaeta.

A special–but widely applicable and very interesting–framework for the occurrence of bifurcation is provided by systems exhibiting *parametric resonance*. This is, for example, the case for an ample class of coupled oscillator systems, which would per se suffice to guarantee the phenomenon is of special interest in applications, beside its theoretical appeal. The analysis of parametric resonance from the point of view of Perturbation Theory is discussed in the article ► Perturbation Analysis of Parametric Resonance by Verhulst.

As mentioned above, the transition from fully regular (integrable) to chaotic behavior is discussed in general terms in the article ► Hamiltonian Perturbation Theory (and Transition to Chaos) by H. Broer and H. Hanßmann. However, quite remarkably, in some cases a perturbation will only moderately destroy the integrable behavior. This should be meant in the following sense: integrable behavior is characterized by the fact that whatever the initial conditions of the system, we are able to predict its behavior after an arbitrary long time. It may happen that albeit this is not true, we are still able to predict either (a) the arbitrarily long time behavior for a dense subset of all the possible initial conditions; or (b) the exponentially long time behavior for a subset of full measure of possible initial conditions (usually, those "sufficiently near" to the exactly integrable case).

In the first case, the meaning of the statement is that any possible initial condition is "near" to an initial condition leading to an integrable-type behavior over all times (which does not imply its behavior will be near to integrable over arbitrary times, but only for sufficiently small–albeit this "small" could be extremely long on human scale–times). This kind of situation is investigated by the *KAM theory* (named after the initials of Kolmogorov, Arnold, and Moser), discussed in the article ► Kolmogorov–Arnold–Moser (KAM) Theory by Chierchia.

In the second case, the statements about stable behavior are valid only for a finite (albeit exponentially long, hence again often extremely long on human scale) time, but apply to an open set of initial conditions. This approach was taken by Nekhoroshev, and is presently known–together with the results obtained in this direction–as *Nekhoroshev theory*; this is the subject of the article ► Nekhoroshev Theory by Niederman.

As mentioned above, the problem of planetary motions was historically at the origin of Perturbation Theory, since Ancient Greece; actually more recent results, including those due to the work of Poincaré–and those embodied in KAM and Nekhoroshev theories–also have roots in Celestial Mechanics (albeit then being used in completely different fields, e. g. in the study of electron motion in a crystal). The application of Perturbation Theory in Celestial Mechanics is a very active field of research, and the subject of the article ► Perturbation Theory in Celestial Mechanics by Celletti.

In this context, one often considers reduced problems where not all the planets are taken into account; this is the origin of the "three-body problem," the three bodies being, for example the Sun, Jupiter, and the Earth; or the Earth, the Moon, and an artificial satellite. Much effort has been recently devoted to the study of special solutions for the N-body problem (that is, N bodies mutually attracting via potential forces) after the discovery of remarkable special solutions–termed "choreographies"–in which the bodies move along one or few common trajectories. This theory has not yet found applications in concrete physical systems or technology, but on the one hand these special solutions provide an organizing center for general nonlinear dynamics, and on the other the applicative potential of such collective motions (say in micro-devices) is rather

obvious. The N-body problem and these special solutions are discussed in the article ► *n*-Body Problem and Choreographies by Terracini.

The reader has probably noted that all these problems correspond to smooth *conservative* systems, or at least to perturbation of such systems. When this is relaxed–which is not appropriate in studying the motion of planetary objects, but may be appropriate in a number of contexts–the situation is both different and less well understood. The article ► Perturbation of Systems with Nilpotent Real Part by Gramchev studies perturbation of linear systems with a nilpotent linear part, which is the case for dissipative unperturbed systems. The article ► Perturbation Theory for Non-smooth Systems by Teixeira discusses the case of nonsmooth systems, which is the case–quite relevant in real-world engineering applications–of systems with impacts.

It was also mentioned that Quantum Theory was another major source of motivation for the development of Perturbation Theory, both historically and still in recent times. As for the latter, one should remark how a standard tool in quantum perturbation theory, i. e. the technique of *Feynman diagrams*–or more generally, diagrammatic expansions–was incorporated into classical perturbation theory only in relatively recent times. The use of diagrammatic expansions in classical perturbation theory is discussed in the article ► Diagrammatic Methods in Classical Perturbation Theory by Gentile.

Apart from this, knowledge of the perturbation-theoretic techniques developed in the framework of quantum theory–both in general and for the study of atoms and molecules–is of general interest, both as a source of inspiration for tackling problems in different contexts and for the intrinsic interest of microscopic systems; while well known to physicists, this theory is maybe less known to mathematicians and engineers. The articles provided in this section of the Encyclopedia can be an excellent entry point for those not familiar with this theory.

In the article ► Perturbation Theory in Quantum Mechanics by Picasso, Bracci, and D'Emilio, the general setting and results are described, together with some selected special topics; the role of symmetries–and hence degeneracies–within quantum perturbation theory is paramount and also discussed here.

The article ► Perturbation Theory and Molecular Dynamics by Panati focuses instead on the specific aspects of the perturbative approach to the quantum dynamics of molecules; this is a remarkable example of how taking into account the separation between slow and fast degrees of freedom allows one to deal with seemingly intractable problems.

A bridge between quantum and classical Perturbation Theory is provided by the *semiclassical* case, corresponding to taking into account the smallness of the energy scale set by Planck's constant h with respect to the energy scale involved in many (macro- or meso-scopic) problems. This is discussed in the article ► Perturbation Theory, Semiclassical by Sacchetti.

The quantum framework is also very interesting in connection with Bifurcation Theory; in this framework the "qualitative changes in the dynamics" which characterize bifurcations corresponds to qualitative changes in the spectrum. This in turn is related to *monodromy* on the mathematical side; and to the problem of an atom in crossed magnetic and electric fields on the physical side. These matters, strongly related to several of those mentioned above, are discussed in the article ► Quantum Bifurcations by Zhilinskii.

From the mathematical point of view, in the quantum case one deals with a Partial Differential Equation–the Schrödinger equation–rather than with a system of Ordinary Differential Equations (it should be noted that when dealing with the spectrum only, one is actually not requiring to study the full set of solutions to the concerned PDE).

Needless to say, this is not the only case where one has to deal with PDEs in the applications, continuum mechanics providing a classical framework where one is obliged to deal with PDEs. Rigorous results in Perturbation Theory for PDEs are not at the same level as for ODEs (and the insight provided by the quantum case is henceforth specially valuable); research in this direction is very active, and faces rather difficult problems despite the progresses obtained in recent years.

Some of these results, together with an overview of the field, are described in the article ► Perturbation Theory for PDEs by Bambusi. Quite appropriately, this article ends up stating that in several applied fields a sound understanding of PDEs rigorous Perturbation Theory would be relevant for applications, mentioning in particular the water wave problem, quantum mechanics, electromagnetic theory and magnetohydrodynamics, and elastodynamics.

This could also be a convenient way to conclude this Introduction, but in this case the reader would unavoidably remain with the impression that Perturbation Theory is mainly dealing with nonlinear problems originating in Physics or Engineering. While this is historically the origin of the most striking developments of the theory–and the realm in which it proved most successful–such characterization is by no means a built-in restriction.

Perturbation Theory can also deal effectively with problems originating in different fields and having a rather

different mathematical formulation, such as those arising in certain fields of Biology (beside fields where the mathematical formulation is anyway in terms of Dynamical Systems). This is shown concretely in the article ► Perturbation of Equilibria in the Mathematical Theory of Evolution by Sanchez; in this case the problem is formulated in terms of *Evolutionary Game Theory*. It should be noted that this is interesting not only for the intrinsic interest of the Darwin theory of Evolution, but also because Game Theory is increasingly used in rather diverse contexts.

Finally, I would like to warmly thank all the Authors for providing the remarkable articles making up this section of the *Encyclopedia*, as well as the Referees who checked them anonymously and in several cases gave suggestions leading to improvements.

Perturbation Theory and Molecular Dynamics

Gianluca Panati
Dipartimento di Matematica, Università di Roma "La Sapienza", Roma, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
The Framework
The Leading Order Born–Oppenheimer Approximation
Beyond the Leading Order
Future Directions
Bibliography

Glossary

Adiabatic decoupling In a complex system (either classical or quantum), the dynamical decoupling between the slow and the fast degrees of freedom.

Adiabatic perturbation theory A mathematical algorithm which exploits the adiabatic decoupling of degrees of freedom in order to provide an approximated (but yet accurate) description of the slow part of the dynamics. In the framework of QMD, it is used to approximately describe the dynamics of nuclei, the perturbative parameter ε being related to the small electron/nucleus mass ratio.

Electronic structure problem The problem consisting in computing, at fixed positions of thenuclei, the energies (eigenvalues) and eigenstates corresponding to the electrons. An approximate solution is usually obtained numerically.

Molecular dynamics The dynamics of the nuclei in a molecule. While a first insight in the problem can be obtained by using classical mechanics (Classical Molecular Dynamics), a complete picture requires quantum mechanics (Quantum Molecular Dynamics) ► Perturbation Theory in Quantum Mechanics. This contribution focuses on the latter viewpoint.

Definition of the Subject

In the framework of Quantum Mechanics the dynamics of a molecule is governed by the (time-dependent) Schrödinger equation, involving nuclei and electrons coupled through electromagnetic interactions. While the equation is mathematically well-posed, yielding the existence of a unique solution, the complexity of the problem makes the exact solution unattainable. Even for small molecules, the large number of degrees of freedom prevents from *direct* numerical simulation, making an approximation scheme necessary.

Indeed, one may exploit the smallness of the electron/nucleus mass ratio to introduce a convenient computational scheme leading to approximate solutions of the original time-dependent problem. In this article we review the standard approximation scheme (**dynamical Born–Oppenheimer approximation**) together with its ramifications and some recent generalizations, focusing on mathematically rigorous results.

The success of this approximation scheme is rooted in a clear separation of time-scales between the motion of electrons and nuclei. Such separation provides the prototypical example of **adiabatic decoupling** between the fast and the slow part of a quantum dynamics. More generally, adiabatic separation of time-scales plays a fundamental role in the understanding of complex system, with applications to a wide range of physical problems.

Introduction

Through the discovery of the Schrödinger equation the theoretical physics and chemistry community attained a powerful tool for computing atomic spectra, either exactly or in perturbation expansion. Born and Oppenheimer [2] immediately strived for a more ambitious goal, namely to understand the excitation spectrum of molecules on the basis of the new wave mechanics. They accomplished to exploit the small electron/nucleus mass ratio as an expansion parameter, which then leads to

the ***stationary*** Born–Oppenheimer approximation. Since then it has become a standard and widely used tool in quantum chemistry, now supported by rigorous mathematical results [4,5,18,19,27].

Beyond molecular structure and excitation spectra, dynamical processes have gained in interest. Examples are the scattering of molecules, chemical reactions, or the decay of an excited state of a molecule through tunneling processes. Such problems require a ***dynamical*** version of the Born–Oppenheimer approximation, which is the topic of this article. At the leading order, the electronic energy at fixed positions of the nuclei serves as an effective potential between the nuclei. We call this the zeroth order Born–Oppenheimer approximation. The resulting effective Schrödinger equation can be used for both statical and dynamical purposes. The input is an electronic structure calculation, which for the purpose of our article we regard as given by other means.

While there are many physical and chemical properties of molecules explained by the zeroth order Born–Oppenheimer approximation, there are cases where higher order corrections are required. Famous examples are the dynamical Jahn–Teller effect and the dynamics of singled out nuclear degrees of freedom near the conical intersection of two energy surfaces. The first order Born–Oppenheimer approximation involves geometric phases, which are of great interest also in other domains of Quantum Mechanics ([41], ► Quantum Bifurcations).

Finally, we mention that some dynamical processes can be modeled as scattering problems. In such cases it is convenient to combine the Born–Oppenheimer scheme together with scattering theory, a topic which goes beyond the purpose of this contribution (see [28,29] and references therein).

A complete overview of the vast literature on the subject of the dynamical Born–Oppenheimer approximation is provided in [24].

The Framework

We consider a molecule consisting of K nuclei, whose positions are denoted as $x = (x_1, \dots, x_K) \in \mathbb{R}^{3K} =: X$, and N electrons, with positions $y = (y_1, \dots, y_N) \in \mathbb{R}^{3N} =: Y$. The wavefunction of the molecule is therefore a square-integrable function Ψ depending on all these coordinates.

Molecular dynamics is described through the Schrödinger equation

$$\imath\hbar \frac{\mathrm{d}}{\mathrm{d}s}\Psi_s = H_{\mathrm{mol}}\Psi_s \,, \tag{1}$$

where s denotes time measured in microscopic units and the Hamiltonian operator is given by

$$H_{\mathrm{mol}} = -\sum_{k=1}^{K} \frac{\hbar^2}{2M_k}\Delta_{x_k} - \sum_{i=1}^{N} \frac{\hbar^2}{2m_{\mathrm{e}}}\Delta_{y_i} + V_{\mathrm{e}}(y) + V_n(x) + V_{\mathrm{en}}(x, y) \,. \tag{2}$$

Here $\hbar$ is the Planck constant, m_{e} is the mass of the electron and M_k the mass of the kth nucleus, and the interaction terms are explicitly given by

$$V_n(x) = \sum_{k=1}^{K}\sum_{l\neq k}^{K} \frac{e^2 Z_k Z_l}{|x_k - x_l|}, \quad V_{\mathrm{e}}(y) = \sum_{i=1}^{N}\sum_{j\neq i}^{N} \frac{e^2}{|y_i - y_j|},$$

and

$$V_{\mathrm{en}}(x, y) = \sum_{k=1}^{K}\sum_{i=1}^{N} -\frac{e^2 Z_k}{|x_k - y_i|} \,,$$

where eZ_k, for $Z_k \in \mathbb{Z}$, is the electric charge of the kth nucleus. In some cases, to obtain rigorous mathematical results one needs to slightly smear out the charge distribution of the nuclei. This is in agreement with the physical picture that nuclei are not point like but extended objects. Hereafter we will assume, for sake of a simpler notation, that all the nuclei have the same mass M. The subsequent discussion is still valid in the general case, with the appropriate choice of the adiabatic parameter indicated below.

As mentioned above, the large number of degrees of freedom makes convenient to elaborate an approximation scheme, exploiting the smallness of the parameter

$$\varepsilon := \sqrt{\frac{m_{\mathrm{e}}}{M}} = 10^{-2} \cdots 10^{-3} \,. \tag{3}$$

In the general case, one has to choose $\varepsilon = \max\{\sqrt{m_{\mathrm{e}}/M_k} : 1 \le k \le K\}$.

By introducing atomic units ($\hbar = 1, m_{\mathrm{e}} = 1$) and making explicit the role of the **adiabatic parameter** ε, the Hamiltonian H_{mol} reads (up to a change of energy scale)

$$H_\varepsilon = -\sum_{k=1}^{K} \frac{\varepsilon^2}{2}\Delta_{x_k} + V_n(x) + \underbrace{\sum_{i=1}^{N} -\frac{1}{2}\Delta_{y_i} + V_{\mathrm{e}}(y) + V_{\mathrm{en}}(x, y)}_{H_{\mathrm{el}}(x)} \,. \tag{4}$$

Notice that, for each fixed nuclei configuration $x = (x_1, \dots, x_K) \in X$, the operator $H_{\mathrm{el}}(x)$ is an operator acting

on the Hilbert space $\mathcal{H}_{\rm el}$ corresponding to the electrons alone.

If the kinetic energies of the nuclei and the electrons are comparable, as it happens in the vast majority of physical situations in view of energy equipartition, then the velocities scale as

$$|v_n| \approx \sqrt{\frac{m_{\rm e}}{M}}|v_{\rm e}| = \varepsilon|v_{\rm e}| \,, \tag{5}$$

where v_n and $v_{\rm e}$ denotes respectively the typical velocity of nuclei and electrons. Therefore, in order to observe a nontrivial dynamics for the nuclei, one has to wait a microscopically long time, namely a time of order $\mathcal{O}(\varepsilon^{-1})$. This scaling fixes the **macroscopic time scale** t, together with the relation $t = \varepsilon s$, where s is the microscopic time appearing in Eq. (1). We are therefore interested in the behavior of the solutions of the equation

$$\imath\varepsilon\frac{\mathrm{d}}{\mathrm{d}t}\Psi(t) = H_\varepsilon\Psi(t) \tag{6}$$

in the limit $\varepsilon \to 0$.

We assume as an input a solution of the **electronic structure problem**, i. e. that for every fixed configuration of the nuclei $x = (x_1, \ldots, x_K)$ one knows the solution of the eigenvalue problem

$$H_{\rm e}(x)\chi_j(x) = E_j(x)\chi_j(x) \,, \tag{7}$$

with $E_j(x) \in \mathbb{R}$ and $\chi_j(x) \in \mathcal{H}_{\rm el}$. Since electrons are fermions, one has $\mathcal{H}_{\rm el} = S_{\rm a}L^2(\mathbb{R}^{3N})$ with $S_{\rm a}$ projecting onto the antisymmetric wave functions. The eigenvectors in Eq. (7) are normalized as

$$\langle\chi_j(x)\,,\,\chi_\ell(x)\rangle_{\mathcal{H}_{\rm el}} \equiv \int_Y \chi_j^*(x,y)\,\chi_\ell(x,y)\mathrm{d}y = \delta_{j\ell}$$

with respect to the scalar product in $\mathcal{H}_{\rm el}$. Note that the eigenvectors are determined only up to a phase $\vartheta_j(x)$. Generically, in addition to the bound states, $H_{\rm e}(x)$ has continuous spectrum. We label the eigenvalues in Eq. (7) as

$$E_1(x) \le E_2(x) \le \cdots \,, \tag{8}$$

including multiplicity. The graph of E_j is called the **jth energy surface** or **energy band**, see Fig. 1.

Generically, in realistic examples such energy bands cross each other, and the possible structures of band crossing have been classified [22,35]. Figure 2 illustrates a realistic example of energy bands, showing in particular the typical conical intersection of two energy surfaces.

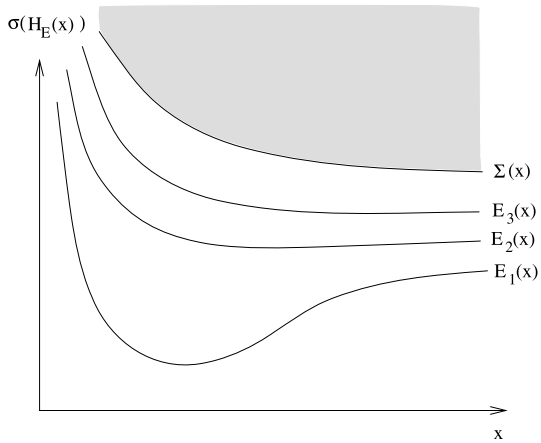

Perturbation Theory and Molecular Dynamics, Figure 1
A schematic representation of energy bands. At each fixed nuclei configuration $x = (x_1, \ldots, x_K)$ the electronic Hamiltonian $H_{\rm el}(x)$ exhibits point spectrum (corresponding to states with all the electrons bound to the nuclei) and continuous spectrum (corresponding to states in which one or more electrons are quasi-free, i. e. the molecule is ionized)

Let $\psi(x)$ be a nucleonic wave function, $\psi \in \mathcal{H}_n$, the Hilbert space corresponding to the nuclei. For simplicity we take $\mathcal{H}_n = L^2(X)$, remembering that to impose the physically correct statistics for the nuclei requires extra considerations [33]. States Ψ of the molecules with the property that the electrons are precisely in the jth eigenstate are then of the form

$$\Psi(x,y) = \psi(x)\,\chi_j(x,y)\,. \tag{9}$$

We can think of Ψ either as a wave function in the total Hilbert space $\mathcal{H} = \mathcal{H}_n \otimes \mathcal{H}_{\rm el}$, or as a wave function for the electrons (i. e. an element of $\mathcal{H}_{\rm el}$) depending parametrically on x. In the common jargon, a state in the form Eq. (9) is said a state *concentrated on the jth band*. We denote as P_j the projector on the subspace consisting of states of the form Eq. (9); since the $\{\chi_j(x)\}_{j\in\mathbb{N}}$ are orthonormal, P_j is indeed an orthogonal projection in $\mathcal{H}_n \otimes \mathcal{H}_{\rm el}$.

The Leading Order Born–Oppenheimer Approximation

We focus now on an a specific energy band, say E_n, assuming that it is globally isolated from the rest of the spectrum (the behavior of the wavefunction at the crossing points will be addressed later).

Under such assumption, a state Ψ_0 which is initially concentrated on the nth band will stay localized in the same band up to errors of order $\mathcal{O}(\varepsilon)$: more specifically,

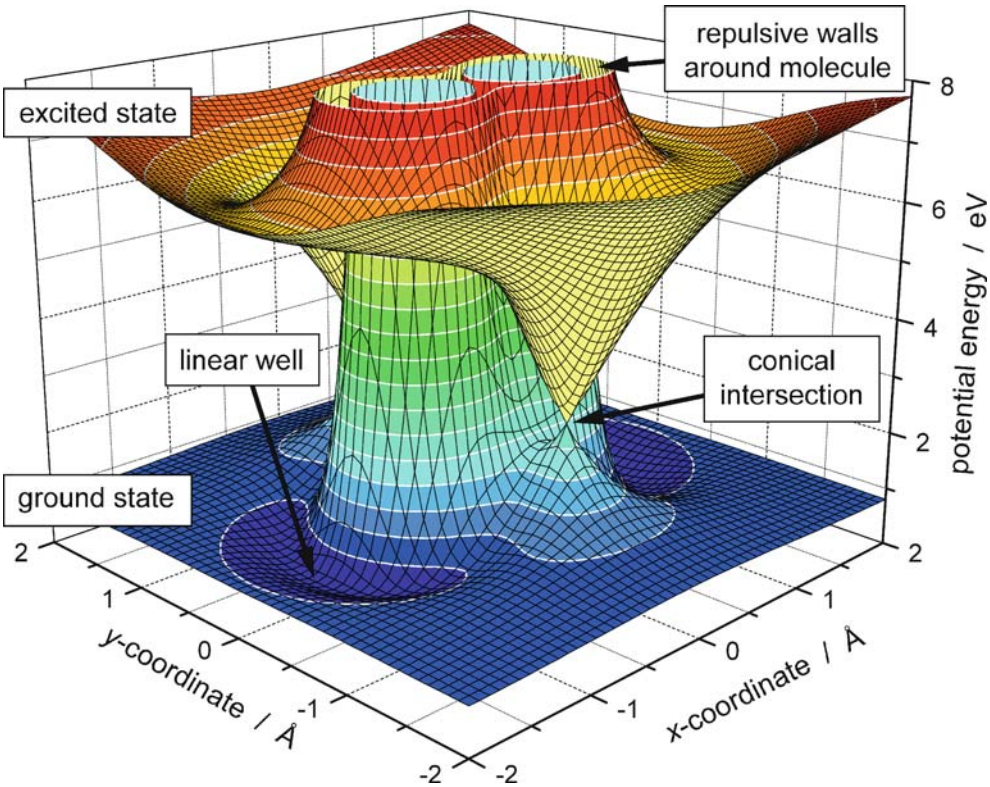

Perturbation Theory and Molecular Dynamics, Figure 2

The first two energy bands for the hydrogen quasi-molecule H_3, i. e. the system consisting of three protons and three electrons. The picture shows the restriction of such bands over a 2-dimensional subspace of the configuration space. Two hydrogen nuclei are located on the *x*-axis with a fixed separation of 1.044 Angstrom Å, and the energy bands are plotted as a function of the relative position (*x*, *y*) of the third nucleus. Notice the conical intersection between the ground and the first excited state, which appears at equilateral triangular geometries. (© Courtesy of Eckart Wrede, Durham University. The plot is generated using the analytic representation of the H_3 energy bands obtained in [40])

one shows that

$$\left\|(1-P_n)\,\mathrm{e}^{-\imath H_\varepsilon t/\varepsilon}\,P_n\Psi_0\right\|_{\mathcal{H}} = \mathcal{O}(\varepsilon)\,. \tag{10}$$

The space Ran P_n, consisting of wavefunctions in the form Eq. (9), is usually called the **adiabatic subspace** corresponding to the nth band.

Since Ran P_n is approximately invariant under the dynamics, one may wonder whether there is a simple and convenient way to approximately describe the dynamics inside such subspace. Indeed, one may argue that for an initial state in Ran P_n the dynamics of the nuclei is governed by the reduced Hamiltonian

$$P_n H_\varepsilon P_n = -\frac{\varepsilon^2}{2}\sum_{k=1}^{K}\Delta_{x_k} + V_n(x) + E_n(x) + \mathcal{O}(\varepsilon) \tag{11}$$

acting in Ran $P_n \cong \mathcal{H}_n = L^2(X)$. The **dynamical Born–Oppenheimer approximation** consists, at the leading order, in replacing the original Hamiltonian Eq. (4) by the Hamiltonian

$$H_{\mathrm{BO}} = -\frac{\varepsilon^2}{2}\sum_{k=1}^{K}\Delta_{x_k} + V_n(x) + E_n(x) \tag{12}$$

acting in $L^2(X)$, getting thus an impressive dimensional reduction. In other words: let $\Psi(t,x,y)$ be the solution of Eq. (6) with initial datum $\Psi_0(x,y) = \varphi_0(x)\chi_n(x,y)$; then $\Psi(t,x,y) = \varphi(t,x)\chi_n(x,y) + \mathcal{O}(\varepsilon)$ where $\varphi(t,x)$ is the solution of the effective equation

$$\imath\frac{\mathrm{d}}{\mathrm{d}t}\varphi(t) = H_{\mathrm{BO}}\varphi(t) \tag{13}$$

with initial datum φ_0.

To prove mathematically the previous claim, one has to bound the difference

$$\left(\mathrm{e}^{-\mathrm{i}H_\varepsilon t/\varepsilon} - \mathrm{e}^{-\mathrm{i}P_n H_\varepsilon P_n t/\varepsilon}\right) P_n\,.$$

A proof of this fact is not immediate as one might expect. Indeed, the Duhamel method (consisting essentially in rewriting a function as the integral of its derivative) yields

$$\begin{aligned}
&\left(\mathrm{e}^{-\mathrm{i}H_\varepsilon\, t/\varepsilon} - \mathrm{e}^{-\mathrm{i}P_n H_\varepsilon P_n\, t/\varepsilon}\right) P_n \\
&= \mathrm{i}\mathrm{e}^{-\mathrm{i}H_\varepsilon t/\varepsilon} \int_0^{t/\varepsilon} \mathrm{d}s\, \mathrm{e}^{\mathrm{i}H_\varepsilon s} \left(P_n H_\varepsilon P_n - H_\varepsilon\right) \mathrm{e}^{-\mathrm{i}P_n H_\varepsilon P_n s}\, P_n \\
&= \mathrm{i}\mathrm{e}^{-\mathrm{i}H_\varepsilon t/\varepsilon} \int_0^{t/\varepsilon} \mathrm{d}s\, \mathrm{e}^{\mathrm{i}H_\varepsilon s} \left(P_n H_\varepsilon P_n - H_\varepsilon\right) P_n\, \mathrm{e}^{-\mathrm{i}P_n H_\varepsilon P_n s} \\
&= \mathrm{i}\mathrm{e}^{-\mathrm{i}H_\varepsilon t/\varepsilon} \int_0^{t/\varepsilon} \mathrm{d}s\, \mathrm{e}^{\mathrm{i}H_\varepsilon s} \underbrace{[P_n, H_\varepsilon]\, P_n}_{\mathcal{O}(\varepsilon)}\, \mathrm{e}^{-\mathrm{i}P_n H_\varepsilon P_n s} \ .
\end{aligned}$$

The commutator appearing in the last line is estimated as

$$[P_n, H_\varepsilon]P_n = \left[|\chi_n(x)\rangle\langle\chi_n(x)|, -\frac{\varepsilon^2}{2}\Delta_x\right] P_n = \mathcal{O}(\varepsilon)\ , \tag{14}$$

but the integration interval diverges as $\mathcal{O}(\varepsilon^{-1})$. Thus the *naïve approach* fails. A rigorous proof has been provided in [39], elaborating on [26], exploiting the fact that the integral in Eq. (14) is, roughly speaking, an oscillatory operator integral. A more direct approach, based on the evolution of generalized Gaussian wavepackets, has been followed in the pioneering papers by Hagedorn [17,20].

Beyond the Leading Order

The dynamics of a state initially concentrated on an isolated energy band is described, up to errors of order $\mathcal{O}(\varepsilon)$, by the Born–Oppenheimer dynamics Eq. (13). It is physically interesting to find an effective dynamics which approximates the original dynamics with an higher degree of accurancy. At first sight one might think that this goal can be simply reached by expanding the operator $P_n\, H_\varepsilon\, P_n$ to the next order in ε. (Notice that the first term appearing in Eq. (11) does contribute as a term of $\mathcal{O}(1)$, since we are considering states such that the kinetic energy of the nuclei is not vanishing, i. e. $\| - \varepsilon^2\Delta_x\Psi\| = \mathcal{O}(1)$, in agreement with the mentioned energy equipartition). However such *naïf* expansion has no physical meaning since it makes no sense to compute the operator appearing in Eq. (11) with greater accuracy if the space Ran P_n itself is invariant only up to terms of order $\mathcal{O}(\varepsilon)$.

To get a deeper insight in the problem, one has to investigate the origin of the $\mathcal{O}(\varepsilon)$ term appearing in the Eq. (10): either (a) there is a part of the wavefunction of order $\mathcal{O}(\varepsilon)$ which is scattered in all the directions in the Hilbert space, or (b) still there is a subspace invariant up to smaller errors, which is however tilted with respect to Ran P_n by a term of order $\mathcal{O}(\varepsilon)$.

Therefore, two natural questions arise:

(i) **Almost-invariant subspace:** is there a subspace of $\mathcal{H} = \mathcal{H}_n \otimes \mathcal{H}_{\mathrm{el}}$ which is invariant under the dynamics up to errors ε^N, for $N > 1$?
(ii) **Intra-band dynamics:** in the affirmative case, is there any simple and convenient way to accurately describe the dynamics inside this subspace?

As for the first question, one may show that to any globally **isolated** energy band E_n corresponds a subspace of the Hilbert space which is almost-invariant under the dynamics. More precisely, one constructs an orthogonal projector $\Pi_{n,\varepsilon} \in \mathcal{B}(\mathcal{H})$, with $\Pi_{n,\varepsilon} = P_n + \mathcal{O}(\varepsilon)$, such that for any $N \in \mathbb{N}$ there exists C_N such that

$$\begin{aligned}
&\left\|(1 - \Pi_{n,\varepsilon})\, \mathrm{e}^{-\mathrm{i}H_\varepsilon t/\varepsilon}\, \Pi_{n,\varepsilon}\Psi_0\right\| \\
&\qquad \le C_N \varepsilon^N (1 + |t|)(1 + \mathcal{E})\|\Psi_0\| \ .
\end{aligned} \tag{15}$$

Here $\mathcal{E}$ denotes a cut-off on the kinetic energy of the nuclei, which corresponds to the physical assumption that the kinetic energies of nuclei and electrons are comparable. Equation (15) shows that if the molecule is initially in a state $\Psi_0 \in$ Ran $\Pi_{n,\varepsilon}$, then after a macroscopic time t the molecule is in a state which is still in Ran $\Pi_{n,\varepsilon}$ up to an error smaller than any power of ε, with the error scaling linearly with respect to time and to the kinetic energy cut-off. For this reason the space Ran $\Pi_{n,\varepsilon}$ is called **superadiabatic subspace** or **almost-invariant subspace**.

We emphasize that the adiabatic decoupling, as formulated in Eq. (15), holds on a long time-scale, as opposed to the semiclassically limit which is known to hold on a time scale of order $\mathcal{O}(\ln\varepsilon)$. Indeed the adiabatic decoupling is a pure quantum phenomenon, conceptually and mathematically independent from the semiclassical limit.

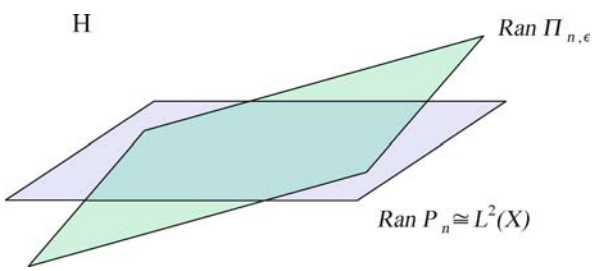

Perturbation Theory and Molecular Dynamics, Figure 3
A schematic illustration of the superadiabatic subspace Ran $\Pi_{n,\varepsilon}$, tilted by a correction of order ε with respect to the usual adiabatic subspace Ran P_n

The previous result is based on a long history of mathematical research, starting with pioneering ideas of Sjöstrand [3,9,31,32,34,37,38]. It has been formulated in the form above in [36].

As for question (ii), one has to face the problem that there is no natural identification between the super-adiabatic subspace Ran $\Pi_{n,\varepsilon}$ and $\mathcal{H}_n \cong L^2(X)$, and therefore no evident reduction in the number of degrees of freedom. This difficulty can be circumvented by constructing a unitary operator which intertwines the previous two spaces, namely

$$U_{n,\varepsilon}\colon \operatorname{Ran}\Pi_{n,\varepsilon} \to \mathcal{H}_n \cong L^2(X)\,. \tag{16}$$

Notice that such a unitary operator is not unique. With the help of $U_{n,\varepsilon}$ we can map the intraband dynamics to the nuclei Hilbert space, obtaining an effective Hamiltonian $\hat{H}_{\mathrm{eff},\varepsilon} := U_{n,\varepsilon}\,\Pi_{n,\varepsilon}\,H_\varepsilon\,\Pi_{n,\varepsilon}\,U_{n,\varepsilon}^{-1}$ acting in $L^2(X)$. It follows from Eq. (15) that for every $N \in \mathbb{N}$ there exist C_N such that

$$\left\| \left(\mathrm{e}^{-\mathrm{i}H_\varepsilon t/\varepsilon} - U_{n,\varepsilon}^{-1}\, \mathrm{e}^{-\mathrm{i}\hat{H}_{\mathrm{eff},\varepsilon}\, t/\varepsilon}\, U_{n,\varepsilon} \right) \Pi_{n,\varepsilon}\Psi_0 \right\|_{\mathcal{H}} \le C_N\, \varepsilon^N\, (1+|t|)(1+\mathcal{E})\|\Psi_0\|\,.$$

However, with an arbitrary choice of $U_{n,\varepsilon}$ the effective Hamiltonian $\hat{H}_{\mathrm{eff},\varepsilon}$ does not appear simpler than $\Pi_{n,\varepsilon}\,H_\varepsilon\,\Pi_{n,\varepsilon}$. On the other side, the non-uniqueness of $U_{n,\varepsilon}$ can be conveniently exploited to simplify the structure of the effective Hamiltonian. It has been proved in [36] that the unitary operator $U_{n,\varepsilon}$ can be explicitly constructed in such a way that $\hat{H}_{\mathrm{eff},\varepsilon}$ has a simple structure, namely it is (close to) the ε-Weyl quantization of a function

$$H_{\mathrm{eff},\varepsilon}\colon X \times \mathbb{R}^{3K} \to \mathbb{R}\,, \quad (q,p) \mapsto H_{\mathrm{eff},\varepsilon}(q,p)\,,$$

defined over the classical phase space. We recall that the ε-Weyl quantization maps a (smooth) function over $X \times \mathbb{R}^{3K}$ into a (possibly unbounded) operator acting in $L^2(X)$. The correspondence is such that any function $f(q)$ is mapped into the multiplication operator times $f(x)$, and any function $g(p)$ is mapped into $g(\imath\varepsilon\nabla_x)$; for a generic function $f(q,p)$ the ordering ambiguity is fixed by choosing

$$\mathrm{e}^{\imath\alpha\cdot q}\mathrm{e}^{\imath\beta\cdot p} \mapsto \mathrm{e}^{\imath(\alpha\cdot x+\beta\cdot(\imath\varepsilon\nabla_x))}\,.$$

For readers interested in the mathematical structure of Weyl quantization, we recommend [15].

Equipped with this terminology, we come back to the effective Hamiltonian. It turns out that, with the appropriate choice of $U_{n,\varepsilon}$, $\hat{H}_{\mathrm{eff},\varepsilon}$ is the ε-Weyl quantization of the function

$$H_{\mathrm{eff},\varepsilon}(q,p) = h_0(q,p) + \varepsilon h_1(q,p) + \varepsilon^2 h_2(q,p) + \mathcal{O}(\varepsilon^3) \tag{17}$$

where

$$\begin{aligned} h_0(q,p) &= \tfrac{1}{2}p^2 + E_n(q) + V_n(q) \\ h_1(q,p) &= -\imath p\cdot\langle \chi_n(q), \nabla_q\chi_n(q)\rangle =: -p\cdot\mathcal{A}_n(q) \end{aligned} \tag{18}$$

and

$$\begin{aligned} h_2(q,p) = {} & \tfrac{1}{2}\mathcal{A}_n^2(q) + \tfrac{1}{2}\langle \nabla_q\chi_n(q), (1-P_n(q)) \\ & \cdot \nabla_q\chi_n(q)\rangle_{\mathcal{H}_{\mathrm{el}}} \\ & - \langle p\cdot\nabla_q\chi_n(q)\,; \\ & \big(H_{\mathrm{el}}(q) - E_n(q)\big)^{-1}\big(1-P_n(q)\big) \\ & \times p\cdot\nabla_q\chi_n(q)\rangle_{\mathcal{H}_{\mathrm{el}}}\,. \end{aligned}$$

The Weyl quantization of h_0 provides the leading order Born–Oppenheimer Hamiltonian Eq. (13). The term h_1 has a geometric origin, involving the Berry connection $\mathcal{A}_n(x)$, a quantity appearing in a variety of adiabatic problems [41]; this term is responsible for the screening of magnetic fields in atoms [42]. Geometric effects in molecular systems (and more generally in adiabatic systems) are an active field of research [10,11,12], see ► Quantum Bifurcations and references therein. As for the second order correction h_2, the first term completes the square $(p-\mathcal{A}_n(x))^2$ showing that the dynamics involves a covariant derivative; the second term is known as the Born–Huang term; the last term contains the reduced resolvent (i. e. the resolvent in the orthogonal complement of Ran P_n) and is due to the fact that the superadiabatic subspace Ran $\Pi_{n,\varepsilon}$ is tilted with respect to Ran P_n.

The third term in h_2, namely

$$\begin{aligned} \mathcal{M}(q,p) = \big\langle p\cdot\nabla\chi_n(q), & (H_{\mathrm{el}}(q) - E_n(q))^{-1} \\ & \times (1-P_n(q))\, p\cdot\nabla\chi_n(q)\big\rangle_{\mathcal{H}_{\mathrm{el}}}\,, \end{aligned} \tag{19}$$

appeared firstly in [36], as a consequence of the rigorous adiabatic perturbation theory developed there. This term is responsible for an $\mathcal{O}(\varepsilon^2)$-correction to the effective mass of the nuclei. Indeed, since different quantization schemes differ by a term of order $\mathcal{O}(\varepsilon)$, we may replace the Weyl quantization with the simpler symmetric quantization, namely we consider

$$\begin{aligned} (\widehat{\mathcal{M}}\psi)(x) = \sum_{\ell,k=1}^{3K} \frac{1}{2}\Big(& \mathfrak{m}_{\ell k}(x)(-\mathrm{i}\varepsilon\partial_{x_\ell})(-\mathrm{i}\varepsilon\partial_{x_k}) \\ & + (-\mathrm{i}\varepsilon\partial_{x_\ell})(-\mathrm{i}\varepsilon\partial_{x_k})\mathfrak{m}_{\ell k}(x)\Big)\psi(x)\,, \end{aligned} \tag{20}$$

where $\mathfrak{m}$ is the x-dependent matrix

$$\mathfrak{m}_{\ell k}(x) = \big\langle \partial_\ell \chi_n(x), (H_e(x) - E_n(x))^{-1} \times (1 - P_n(x))\, \partial_k \chi_n(x) \big\rangle_{\mathcal{H}_{el}} .$$

It is clear from Eq. (20) that this term induces a correction of order $\mathcal{O}(\varepsilon^2)$ to the Laplacean, i. e. to the inertia of the nuclei.

Finally, we point out that the effective Hamiltonian $\hat{H}_{\mathrm{eff},\varepsilon}$ can be conveniently truncated at any order in ε, getting corresponding errors in the effective quantum dynamics: if we pose

$$\hat{H}_{\mathrm{eff},\varepsilon} = \sum_{j=0}^{N} \varepsilon^j \hat{h}_j + \mathcal{O}(\varepsilon^{N+1}) =: \hat{h}_{(N),\varepsilon} + \mathcal{O}(\varepsilon^{N+1}) , \quad (21)$$

then there exist a constant $\tilde{C}_N$ such that

$$\left\| \left(\mathrm{e}^{-\mathrm{i}H_\varepsilon t/\varepsilon} - U_{n,\varepsilon}^{-1}\, \mathrm{e}^{-\mathrm{i}\hat{h}_{(N),\varepsilon} t/\varepsilon}\, U_{n,\varepsilon} \right) \Pi_{n,\varepsilon} \Psi_0 \right\|_{\mathcal{H}} \le \tilde{C}_N\, \varepsilon^{N+1} (1 + |t|)(1 + \mathcal{E}) \|\Psi_0\| .$$

The determination of the effective Hamiltonian, here described following [36,40], has been investigated earlier in [31,41] with different but related techniques. The result in [36] is based on an iterative algorithm inspired by classical perturbation theory (see ▶ Kolmogorov–Arnold–Moser (KAM) Theory and ▶ Normal Forms in Perturbation Theory).

Future Directions

Generically energy surfaces cross each other (see Fig. 2), and a globally isolated energy band is just a mathematical idealization. On the other side, if the initial datum $\Psi_0(x, y) = \varphi_0(x)\chi_n(x, y)$ contains a nucleonic wavefunction φ_0 localized far away from the crossing points, the adiabatic approximation is still valid, up to the time when the wavefunction becomes relevant in a neighborhood of radius $\sqrt{\varepsilon}$ of the crossing points. This "hitting-time" can be estimated semiclassically, as done for example in [39].

When the wavefunction reaches the region around the crossing point a relevant part of it might undergo a transition to the other crossing band. (The simultaneous crossing of more than two bands is not generic, see [35], so we focus on the crossing of two bands). The understanding of the dynamics near a conical crossing is a very active field of research.

The first step is a convenient classification of the possible structures of band crossings. Since the early days of Quantum Mechanics [35], it has been realized that eigenvalue crossings occurs on submanifolds of various codimension, according to the symmetry of the problem. In the case of a molecular Hamiltonian in the form Eq. (2), generic crossings of bands with the minimal multiplicity allowed by the symmetry group have been classified in [22].

The second step consist in an analysis of the propagation of the wavefunction near the conical crossing, assuming that the initial state is concentrated on a single band, say the nth band. A pioneering work [23] shows that the qualitative picture is the following: for crossings of codimension 1 the wavefunction follows, at the leading order, the analytic continuation of the nth band, as if there was no crossing. In the higher codimension case, a part of the wavefunction of order $\mathcal{O}(1)$ undergoes a transition to the other band. More recently, propagation through conical crossings has been investigated with new techniques [6,7,8,13,14] opening the way to future research.

Alternatively, one may consider a family of energy bands which cross each other, but which are separated by an energy gap from the rest of the spectrum. Indeed, in a molecular collision or in excitations through a laser pulse only a few energy surfaces take part in the subsequent dynamics. Thus we take a set I of adjacent energy surfaces and call

$$P_{\mathrm{I}} = \sum_{j \in I} P_j \quad (22)$$

the projection onto the relevant subspace (or subspace of physical interest). To ensure that other bands are not involved, we assume them to have a spectral gap of size $a_{\mathrm{gap}} > 0$ away from the energy surfaces in I, i. e.

$$\sup_{x \in X} |E_i(x) - E_j(x)| \ge a_{\mathrm{gap}} \quad \text{for all } j \in I ,\ i \in I^c . \quad (23)$$

Also the continuous spectrum is assumed to be at least a_{gap} away from the relevant energy surfaces. Under such assumption, the **multiband adiabatic theory** assures that the subspace Ran P_I is adiabatically protected against transitions, i. e.

$$\|(1 - P_{\mathrm{I}})\, \mathrm{e}^{-iH_\varepsilon t/\varepsilon}\, P_{\mathrm{I}} \Psi_0\|_{\mathcal{H}} = \mathcal{O}(\varepsilon) .$$

Analogously to the case of a single band, one may construct the corresponding superadiabatic projector. The effective Hamiltonian $\widehat{H}_{\mathrm{eff},\varepsilon}$ corresponding to a family of m bands becomes, in this context, the ε-Weyl quantization of *matrix-valued* function over the classical phase space [36,40].

A deeper understanding of nuclear dynamics near conical crossings and a further developments of multiband adiabatic perturbation theory are, in the opinion of the author, two of the main directions for future research.

Bibliography

Primary Literature

1. Berry MV, Lim R (1990) The Born–Oppenheimer electric gauge force is repulsive near degeneracies. J Phys A 23:L655–L657
2. Born M, Oppenheimer R (1927) Zur Quantentheorie der Molekeln. Ann Phys (Leipzig) 84:457–484
3. Brummelhuis R, Nourrigat J (1999) Scattering amplitude for Dirac operators. Comm Partial Differ Equ 24(1–2):377–394
4. Combes JM (1977) The Born–Oppenheimer approximation. Acta Phys Austriaca 17:139–159
5. Combes JM, Duclos P, Seiler R (1981) The Born–Oppenheimer approximation. In: Velo G, Wightman A (eds) Rigorous Atomic and Molecular Physics. Plenum, New York, pp 185–212
6. de Verdière YC (2004) The level crossing problem in semi-classical analysis. II. The Hermitian case. Ann Inst Fourier (Grenoble) 54(5):1423–1441
7. de Verdière YC, Lombardi M, Pollet C (1999) The microlocal Landau–Zener formula. Ann Inst H Poincaré Phys Theor 71:95–127
8. de Verdière YC (2003) The level crossing problem in semi-classical analysis. I. The symmetric case. Proceedings of the international conference in honor of Frédéric Pham (Nice, 2002). Ann Inst Fourier (Grenoble) 53(4):1023–1054
9. Emmrich C, Weinstein A (1996) Geometry of the transport equation in multicomponent WKB approximations. Commun Math Phys 176:701–711
10. Faure F, Zhilinskii BI (2000) Topological Chern indices in molecular spectra. Phys Rev Lett 85:960–963
11. Faure F, Zhilinskii BI (2001) Topological properties of the Born–Oppenheimer approximation and implications for the exact spectrum. Lett Math Phys 55:219–238
12. Faure F, Zhilinskii BI (2002) Topologically coupled energy bands in molecules. Phys Lett 302:242–252
13. Fermanian–Kammerer C, Gérard P (2002) Mesures semi-classiques et croisement de modes. Bull Soc Math France 130:123–168
14. Fermanian–Kammerer C, Lasser C (2003) Wigner measures and codimension 2 crossings. J Math Phys 44:507–527
15. Folland GB (1989) Harmonic analysis in phase space. Princeton University Press, Princeton
16. Hagedorn GA (1980) A time dependent Born–Oppenheimer approximation. Commun Math Phys 77:1–19
17. Hagedorn GA (1986) High order corrections to the time-dependent Born–Oppenheimer approximation. I. Smooth potentials. Ann Math (2) 124(3):571–590
18. Hagedorn GA (1987) High order corrections to the time-independent Born–Oppenheimer approximation I: Smooth potentials. Ann Inst H Poincaré Sect. A 47:1–16
19. Hagedorn GA (1988) High order corrections to the time-independent Born–Oppenheimer approximation II: Diatomic Coulomb systems. Comm Math Phys 116:23–44
20. Hagedorn GA (1988) High order corrections to the time-dependent Born–Oppenheimer approximation. II. Coulomb systems. Comm Math Phys 117(3):387–403
21. Hagedorn GA (1989) Adiabatic expansions near eigenvalue crossings. Ann Phys 196:278–295
22. Hagedorn GA (1992) Classification and normal forms for quantum mechanical eigenvalue crossings. Méthodes semi-classiques, vol 2 (Nantes, 1991). Astérisque 210(7):115–134
23. Hagedorn GA (1994) Molecular propagation through electron energy level crossings. Mem Amer Math Soc 111:1–130
24. Hagedorn GA, Joye A (2007) Mathematical analysis of Born–Oppenheimer approximations. Spectral theory and mathematical physics: a Festschrift in honor of Barry Simon's 60th birthday, In: Proc Sympos Pure Math 76, Part 1, Amer Math Soc, Providence, RI, pp 203–226
25. Herrin J, Howland JS (1997) The Born–Oppenheimer approximation: straight-up and with a twist. Rev Math Phys 9:467–488
26. Kato T (1950) On the adiabatic theorem of quantum mechanics. Phys Soc Jap 5:435–439
27. Klein M, Martinez A, Seiler R, Wang XP (1992) On the Born–Oppenheimer expansion for polyatomic molecules. Commun Math Phys 143:607–639
28. Klein M, Martinez A, Wang XP (1993) On the Born–Oppenheimer approximation of wave operators in molecular scattering theory. Comm Math Phys 152:73–95
29. Klein M, Martinez A, Wang XP (1997) On the Born–Oppenheimer approximation of diatomic wave operators II. Sigular potentials. J Math Phys 38:1373–1396
30. Lasser C, Teufel S (2005) Propagation through conical crossings: an asymptotic transport equation and numerical experiments. Commun Pure Appl Math 58:1188–1230
31. Littlejohn RG, Flynn WG (1991) Geometric phases in the asymptotic theory of coupled wave equations. Phys Rev 44:5239–5255
32. Martinez A, Sordoni V (2002) A general reduction scheme for the time-dependent Born–Oppenheimer approximation. Comptes Rendus Acad Sci Paris 334:185–188
33. Mead CA, Truhlar DG (1979) On the determination of Born–Oppenheimer nuclear motion wave functions including complications due to conical intersections and identical nuclei. J Chem Phys 70:2284–2296
34. Nenciu G, Sordoni V (2004) Semiclassical limit for multistate Klein–Gordon systems: almost invariant subspaces and scattering theory. J Math Phys 45:3676–3696
35. von Neumann J, Wigner EP (1929) Über das Verhalten von Eigenwerten bei adiabatischen Prozessen. Phys Z 30:467–470
36. Panati G, Spohn H, Teufel S (2003) Space-adiabatic perturbation theory. Adv Theor Math Phys 7:145–204
37. Sjöstrand J (1993) Projecteurs adiabatiques du point de vue pseudodifferéntiel. Comptes Rendus Acad Sci Paris, Série I 317:217–220
38. Sordoni V (2003) Reduction scheme for semiclassical operator-valued Schrödinger type equation and application to scattering. Comm Partial Differ Equ 28(7–8):1221–1236
39. Spohn H, Teufel S (2001) Adiabatic decoupling and time-dependent Born–Oppenheimer theory. Commun Math Phys 224:113–132
40. Varandas AJC, Brown FB, Mead CA, Truhlar DG, Blais NC (1987) A double many-body expansion of the two lowest-energy potential surfaces and nonadiabatic coupling for H_3. J Chem Phys 86:6258–6269
41. Weigert S, Littlejohn RG (1993) Diagonalization of multicomponent wave equations with a Born–Oppenheimer example. Phys Rev A 47:3506–3512
42. Yin L, Mead CA (1994) Magnetic screening of nuclei by electrons as an effect of geometric vector potential. J Chem Phys 100:8125–8131

Books and Reviews

Bohm A, Mostafazadeh A, Koizumi A, Niu Q, Zwanziger J (2003) The geometric phase in quantum systems. Texts and monographs in physics. Springer, Heidelberg

Teufel S (2003) Adiabatic perturbation theory in quantum dynamics. Lecture notes in mathematics, vol 1821. Springer, Berlin

Perturbation Theory for Non-smooth Systems

Marco Antônio Teixeira
Department of Mathematics, Universidade Estadual de Campinas, Campinas, Brazil

Article Outline

Glossary
Definition of the Subject
Introduction
Preliminaries
Vector Fields near the Boundary
Generic Bifurcation
Singular Perturbation Problem in 2D
Future Directions
Bibliography

Glossary

Non-smooth dynamical system Systems derived from ordinary differential equations when the non-uniqueness of solutions is allowed. In this article we deal with discontinuous vector fields in R^n where the discontinuities are concentrated in a codimension-one surface.

Bifurcation In a k-parameter family of systems, a bifurcation is a parameter value at which the phase portrait is not structurally stable.

Typical singularity Are points on the discontinuity set where the orbits of the system through them must be distinguished.

Definition of the Subject

In this article we survey some qualitative and geometric aspects of non-smooth dynamical systems theory. Our goal is to provide an overview of the state of the art on the theory of contact between a vector field and a manifold, and on discontinuous vector fields and their perturbations. We also establish a bridge between two-dimensional non-smooth systems and the geometric singular perturbation theory. Non-smooth dynamical systems is a subject that has been developing at a very fast pace in recent years due to various factors: its mathematical beauty, its strong relationship with other branches of science and the challenge in establishing reasonable and consistent definitions and conventions. It has become certainly one of the common frontiers between mathematics and physics/engineering. We mention that certain phenomena in control systems, impact in mechanical systems and nonlinear oscillations are the main sources of motivation for our study concerning the dynamics of those systems that emerge from differential equations with discontinuous right-hand sides. We understand that non-smooth systems are driven by applications and they play an intrinsic role in a wide range of technological areas.

Introduction

The purpose of this article is to present some aspects of the geometric theory of a class of non-smooth systems. Our main concern is to bring the theory into the domain of geometry and topology in a comprehensive mathematical manner.

Since this is an impossible task, we do not attempt to touch upon all sides of this subject in one article. We focus on exploring the local behavior of systems around typical singularities. The first task is to describe a generic persistence of a local theory (structural stability and bifurcation) for discontinuous systems mainly in the two- and three-dimensional cases. Afterwards we present some striking features and results of the regularization process of two-dimensional discontinuous systems in the framework developed by Sotomayor and Teixeira in [44] and establish a bridge between those systems and the fundamental role played by the Geometric Singular Perturbation Theory (GSPT). This transition was introduced in [10] and we reproduce here its main features in the two-dimensional case. For an introductory reading on the methods of geometric singular perturbation theory we refer to [16,18,30]. In Sect. "Definition of the Subject" we introduce the setting of this article. In Sect. "Introduction" we survey the state of the art of the contact between a vector field and a manifold. The results contained in this section are crucial for the development of our approach. In Sect. "Preliminaries" we discuss the classification of typical singularities of non-smooth vector fields. The study of non-smooth systems, via GSPT, is presented in Sect. "Vector Fields near the Boundary". In Sect. "Generic Bifurcation" some theoretical open problems are presented.

One aspect of the qualitative point of view is the problem of structural stability, the most comprehensive of many different notions of stability. This theme was studied in 1937 by Andronov–Pontryagin (see [3]). This prob-

lem is of obvious importance, since in practice one obtains a lot of qualitative information not only on a fixed system but also on its nearby systems.

We deal with non-smooth vector fields in R^{n+1} having a codimension-one submanifold M as its discontinuity set. The scheme in this work toward a systematic classification of typical singularities of non-smooth systems follows the ideas developed by Sotomayor–Teixeira in [43] where the problem of contact between a vector field and the boundary of a manifold was discussed. Our approach intends to be self-contained and is accompanied by an extensive bibliography. We will try to focus here on areas that are complimentary to some recent reviews made elsewhere.

The concept of structural stability in the space of non-smooth vector fields is based on the following definition:

Definition 1 Two vector fields Z and $\tilde{Z}$ are C^0 equivalent if there is an M-invariant homeomorphism $h\colon R^{n+1} \to R^{n+1}$ that sends orbits of Z to orbits of $\tilde{Z}$.

A general discussion is presented to study certain unstable non-smooth vector fields within a generic context. The framework in which we shall pursue these unstable systems is sometimes called generic bifurcation theory. In [3] the concept of kth-order structural stability is also presented; in a local approach such setting gives rise to the notion of a codimension-k singularity. In studies of classical dynamical systems, normal form theory has been well accepted as a powerful tool in studying the local theory (see [6]). Observe that, so far, bifurcation and normal form theories for non-smooth vector fields have not been extensively studied in a systematic way.

Control Theory is a natural source of mathematical models of these systems (see, for instance, [4,8,20,41,45]). Interesting problems concerning discontinuous systems can be formulated in systems with hysteresis ([41]), economics ([23,25]) and biology ([7]). It is worth mentioning that in [5] a class of relay systems in R^n is discussed. They have the form:

$$X = Ax + \operatorname{sgn}(x_1)k$$

where $x = (x_1, x_2, \ldots, x_n)$, $A \in M_R(n, n)$ and $k = (k_1, k_2, \ldots, k_n)$ is a constant vector in R^n. In [28,29] the generic singularities of reversible relay systems in $4D$ were classified. In [54] some properties of non-smooth dynamics are discussed in order to understand some phenomena that arise in chattering control. We mention the presence of chaotic behavior in some non-smooth systems (see for example [12]). It is worthwhile to cite [17], where the main problem in the classical calculus of variations was carried out to study discontinuous Hamiltonian vector fields. We refer to [14] for a comprehensive text involving non-smooth systems which includes many models and applications. In particular motivating models of several non-smooth dynamical systems arising in the occurrence of impacting motion in mechanical systems, switchings in electronic systems and hybrid dynamics in control systems are presented together with an extensive literature on impact oscillators which we do not attempt to survey here. For further reading on some mathematical aspects of this subject we recommend [11] and references therein. A setting of general aspects of non-smooth systems can be found also in [35] and references therein. Our discussion does not focus on continuous but rather on non-smooth dynamical systems and we are aware that the interest in this subject goes beyond the approach adopted here.

The author wishes to thank R. Garcia, T.M. Seara and J. Sotomayor for many helpful conversations.

Preliminaries

Now we introduce some of the terminology, basic concepts and some results that will be used in the sequel.

Definition 2 Two vector fields Z and $\tilde{Z}$ on R^n with $Z(0) = \tilde{Z}(0)$ are *germ-equivalent* if they coincide on some neighborhood V of 0.

The equivalent classes for this equivalence are called germs of vector fields. In the same way as defined above, we may define germs of functions. For simplicity we are considering the germ notation and we will not distinguish a germ of a function and any one of its representatives. So, for example, the notation $h\colon R^n, 0 \to R$ means that the h is a germ of a function defined in a neighborhood of 0 in R^n. Refer to [15] for a brief and nice introduction of the concepts of *germ* and *k-jet* of functions.

Discontinuous Systems

Let $M = h^{-1}(0)$, where h is (a germ of) a smooth function $h\colon R^{n+1}, 0 \longrightarrow R$ having $0 \in R$ as its regular value. We assume that $0 \in M$.

Designate by $\chi(n+1)$ the space of all germs of C^r vector fields on R^{n+1} at 0 endowed with the C^r-topology with $r > 1$ and large enough for our purposes. Call $\Omega(n+1)$ the space of all germs of vector fields Z in $R^{n+1}, 0$ such that

$$Z(q) = \begin{cases} X(q)\,, & \text{for } h(q) > 0\,, \\ Y(q)\,, & \text{for } h(q) < 0\,, \end{cases} \tag{1}$$

The above field is denoted by $Z = (X, Y)$. So we are considering $\Omega(n+1) = \chi(n+1) \times \chi(n+1)$ endowed with the product topology.

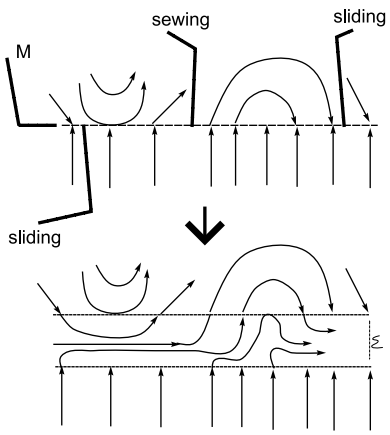

Perturbation Theory for Non-smooth Systems, Figure 1
A discontinuous system and its regularization

Definition 3 We say that $Z \in \Omega(n+1)$ is structurally stable if there exists a neighborhood U of Z in $\Omega(n+1)$ such that every $\tilde{Z} \in U$ is C^0-equivalent with Z.

To define the orbit solutions of Z on the switching surface M we take a pragmatic approach. In a well characterized open set O of M (described below) the solution of Z through a point $p \in O$ obeys the Filippov rules and on $M - O$ we accept it to be multivalued. Roughly speaking, as we are interested in studying the structural stability in $\Omega(n+1)$ it is convenient to take into account all the leaves of the foliation in R^{n+1} generated by the orbits of Z (and also the orbits of X and Y) passing through $p \in M$. (see Fig. 1)

The trajectories of Z are the solutions of the autonomous differential system $\dot{q} = Z(q)$.

In what follows we illustrate our terminology by presenting a simplified model that is found in the classical electromagnetism theory (see for instance [26]):

$$\ddot{x} - \dddot{x} + \alpha \operatorname{sign} x = 0 \,.$$

with $\alpha > 0$.

So this system can be expressed by the following objects: $h(x, y, z) = x$ and $Z = (X, Y)$ with $X(x, y, z) = (y, z, z + \alpha)$ and $Y(x, y, z) = (y, z, z - \alpha)$.

For each $X \in \chi(n+1)$ we define the smooth function $Xh\colon R^{n+1} \to R$ given by $Xh = X \cdot \nabla h$ where $\cdot$ is the canonical scalar product in R^{n+1}.

We distinguish the following regions on the discontinuity set M:

(i) M_1 is the *sewing region* that is represented by $h = 0$ and $(Xh)(Yh) > 0$;
(ii) M_2 is the *escaping region* that is represented by $h = 0$, $(Xh) > 0$ and $(Yh) < 0$;
(iii) M_3 is the *sliding region* that is represented by $h = 0$, $(Xh) < 0$ and $(Yh) > 0$.

We set $\mathcal{O} = \bigcup_{i=1,2,3} M_i$.

Consider $Z = (X, Y) \in \Omega(n+1)$ and $p \in M_3$. In this case, following Filippov's convention, the solution $\gamma(t)$ of Z through p follows, for $t \geq 0$, the orbit of a vector field tangent to M. Such system is called *sliding vector field* associated with Z and it will be defined below.

Definition 4 The sliding vector field associated to $Z = (X, Y)$ is the smooth vector field Z^s tangent to M and defined at $q \in M_3$ by $Z^s(q) = m - q$ with m being the point where the segment joining $q + X(q)$ and $q + Y(q)$ is tangent to M.

It is clear that if $q \in M_3$ then $q \in M_2$ for $-Z$ and then we define the *escaping vector field* on M_2 associated with Z by $Z^e = -(-Z)^s$. In what follows we use the notation Z^M for both cases.

We recall that sometimes Z^M is defined in an open region U with boundary. In this case it can be C^r extended to a full neighborhood of $p \in \partial U$ in M.

When the vectors $X(p)$ and $Y(p)$, with $p \in M_2 \bigcup M_3$ are linearly dependent then $Z^M(p) = 0$. In this case we say that p is a simple singularity of Z. The other singularities of Z are concentrated outside the set O.

We finish this subsection with a three-*dimensional* example:

Let $Z = (X, Y) \in \Omega(3)$ with $h(x, y, z) = z$, $X = (1, 0, x)$ and $Y = (0, 1, y)$. The system determines four quadrants around 0, bounded by $\tau_X = \{x = 0\}$ and $\tau_Y = \{y = 0\}$. They are: $Q_1^+ = \{x > 0, y > 0\}$, $Q_1^- = \{x < 0, y < 0\}$, $Q_2 = \{x < 0, y > 0\}$ (sliding region) and $Q_3 = \{x > 0, y < 0\}$ (escaping region). Observe that $M_1 = Q_1^+ \bigcup Q_1^-$.

The sliding vector field defined in Q_2 is expressed by:

$$Z^s(x, y, z) = (y - x)^{-1} \left(x + y, \frac{y + x}{8}, 0 \right) .$$

Such a system is (in Q_2) equivalent to $G(x, y, z) = (x + y, \frac{y+x}{8}, 0))$. In our terminology we consider G a smooth extension of Z^s, that is defined in a whole neighborhood of 0. It is worthwhile to say that G is in fact a system which is equivalent to the original system in Q_2.

In [50] a generic classification of one-parameter families of sliding vector fields is presented.

Singular Perturbation Problem

A singular perturbation problem is expressed by a differential equation $z' = \alpha(z, \varepsilon)$ (refer to [16,18,30]) where $z \in R^{n+m}$, ε is a small non-negative real number and α is a C^∞ mapping.

Let $z = (x, y) \in R^{n+m}$ and $f\colon R^{m+n} \to R^m$, $g\colon R^{m+n} \to R^n$ be smooth mappings. We deal with equations that may be written in the form

$$\begin{cases} x' = f(x, y, \varepsilon) \\ y' = \varepsilon g(x, y, \varepsilon) \end{cases} \quad x = x(\tau), y = y(\tau)\,. \tag{2}$$

An interesting model of such systems can be obtained from the singular van der Pol's equation

$$\varepsilon x'' + (x^2 + x)x' + x - a = 0\,. \tag{3}$$

The main trick in the geometric singular perturbation (GSP) is to consider the family (2) in addition to the family

$$\begin{cases} \varepsilon \dot{x} = f(x, y, \varepsilon) \\ \dot{y} = g(x, y, \varepsilon) \end{cases} \quad x = x(t), y = y(t) \tag{4}$$

obtained after the time rescaling $t = \varepsilon\tau$.

Equation (2) is called the *fast system* and (4) the *slow system*. Observe that for $\varepsilon > 0$ the phase portrait of fast and slow systems coincide.

For $\varepsilon = 0$, let S be the set of all singular points of (2). We call S the slow manifold of the singular perturbation problem and it is important to notice that Eq. (4) defines a dynamical system on S called the *reduced problem*.

Combining results on the dynamics of these two limiting problems (2) and (4), with $\varepsilon = 0$, one obtains information on the dynamics for small values of ε. In fact, such techniques can be exploited to formally construct approximated solutions on pieces of curves that satisfy some limiting version of the original equation as ε goes to zero.

Definition 5 Let $A, B \subset R^{n+m}$ be compact sets. The Hausdorff distance between A and B is $D(A, B) = \max_{z_1 \in A, z_2 \in B}\{d(z_1, B), d(z_2, A)\}$.

The main question in GSP-theory is to exhibit conditions under which a singular orbit can be approximated by regular orbits for $\varepsilon \downarrow 0$, with respect to the Hausdorff distance.

Regularization Process

An approximation of the discontinuous vector field $Z = (X, Y)$ by a one-parameter family of continuous vector fields will be called a regularization of Z. In [44], Sotomayor and Teixeira introduced the regularization procedure of a discontinuous vector field. A transition function is used to average X and Y in order to get a family of continuous vector fields that approximates the discontinuous one. Figure 1 gives a clear illustration of the regularization process.

Let $Z = (X, Y) \in \Omega(n+1)$.

Definition 6 A C^∞ function $\varphi\colon R \longrightarrow R$ is a transition function if $\varphi(x) = -1$ for $x \le -1$, $\varphi(x) = 1$ for $x \ge 1$ and $\varphi'(x) > 0$ if $x \in (-1, 1)$. The ϕ-regularization of $Z = (X, Y)$ is the one-parameter family $X_\varepsilon \in C^r$ given by

$$Z_\varepsilon(q) = \left(\frac{1}{2} + \frac{\varphi_\varepsilon(h(q))}{2}\right) X(q) + \left(\frac{1}{2} - \frac{\varphi_\varepsilon(h(q))}{2}\right) Y(q). \tag{5}$$

with h given in the above Subsect. "Discontinuous Systems" and $\varphi_\varepsilon(x) = \varphi(x/\varepsilon)$, for $\varepsilon > 0$.

As already said before, a point in the phase space which moves on an orbit of Z crosses M when it reaches the region M_1. Solutions of Z through points of M_3, will remain in M in forward time. Analogously, solutions of Z through points of M_2 will remain in M in backward time. In [34,44] such conventions are justified by the regularization method in dimensions two and three respectively.

Vector Fields near the Boundary

In this section we discuss the behavior of smooth vector fields in R^{n+1} relative to a codimension-one submanifold (say, the above defined M). We base our approach on the concepts and results contained in [43,53]. The principal advantage of this setting is that the generic contact between a smooth vector field and M can often be easily recognized. As an application the typical singularities of a discontinuous system can be further classified in a straightforward way.

We say that $X, Y \in \chi(n+1)$ are M-equivalent if there exists an M-preserving homeomorphism $h\colon R^{n+1}, 0 \longrightarrow R^{n+1}, 0$ that sends orbits of X into orbits of Y. In this way we get the concept of M-structural stability in $\chi(n+1)$.

We call $\Gamma_0(n+1)$ the set of elements X in $\chi(n+1)$ satisfying one of the following conditions:

0) $Xh(0) \neq 0$ (0 is a regular point of X). In this case X is transversal to M at 0.
1) $Xh(0) = 0$ and $X^2h(0) \neq 0$ (0 is a 2-*fold* point of X;)
2) $Xh(0) = X^2h(0) = 0$, $X^3h(0) \neq 0$ and the set $\{Dh(0), DXh(0), DX^2h(0)\}$ is linearly independent (0 is a *cusp* point of X;)

   ...

n) $Xh(0) = X^2h(0) = \cdots = X^nh(0) = 0$ and $X^{n+1}h(0) \neq 0$. Moreover the set $\{Dh(0), DXh(0),$

$DX^2h(0), \ldots, DX^nh(0)\}$ is linearly independent, and 0 is a regular point of the mapping $Xh_{|M}$.

We say that 0 is an M-singularity of X if $h(0) = Xh(0) = 0$. It is a *codimension-zero* M-singularity provided that $X \in \Gamma_0(n+1)$.

We know that $\Gamma_0(n+1)$ is an open and dense set in $\chi(n+1)$ and it coincides with the M-structurally stable vector fields in $\chi(n+1)$ (see [53]).

Denote by $\tau_X \subset M$ the M-singular set of $X \in \chi(n+1)$; this set is represented by the equations $h = Xh = 0$. It is worthwhile to point out that, generically, all two-folds constitute an open and dense subset of τ_X. Observe that if $X(0) = 0$ then $X \notin \Gamma_0(n+1)$.

The M-bifurcation set is represented by $\chi_1(n+1) = \chi(n+1) - \Gamma_0(n+1)$

Vishik in [53] exhibited the normal forms of a *codimension-zero* M-singularity. They are:

I) Straightened vector field

$$X = (1, 0, \ldots, 0)$$

and

$$h(x) = x_1^{k+1} + x_2x_1^{k-1} + x_3x_1^{k-2} + \cdots + x_{k+1}, \quad k = 0, 1, \ldots, n$$

or

II) Straightened boundary

$$h(x) = x_1$$

and

$$X(x) = (x_2, x_3, \ldots, x_k, 1, 0, 0, \ldots, 0)$$

We now discuss an important interaction between vector fields near M and singularities of mapping theory. We discuss how singularity-theoretic techniques help the understanding of the dynamics of our systems.

We outline this setting, which will be very useful in the sequel. The starting point is the following construction.

A Construction

Let $X \in \chi(n+1)$. Consider a coordinate system $x = (x_1, x_2, \ldots, x_{n+1})$ in $R^{n+1}, 0$ such that

$$M = \{x_1 = 0\}$$

and

$$X = (X^1, X^2, \ldots, X^{n+1})$$

Assume that $X(0) \neq 0$ and $X^1(0) = 0$. Let N_0 be any transversal section to X at 0.

By the implicit function theorem, we derive that:

for each $p \in M, 0$ there exists a unique $t = t(p)$ in $R, 0$ such that the orbit-solution $t \mapsto \gamma(p, t)$ of X through p meets N_0 at a point $\tilde{p} = \gamma(p, t(p))$.

We define the smooth mapping $\rho_X \colon R^n, 0 \longrightarrow R^n, 0$ by $\rho_X(p) = \tilde{p}$. This mapping is a powerful tool in the study of vector fields around the boundary of a manifold (refer to [21,42,43,46,53]). We observe that τ_X coincides with the singular set of ρ_X.

The late construction implements the following method. If we are interested in finding an equivalence between two vector fields which preserve M, then the problem can be sometimes reduced to finding an equivalence between ρ_X and ρ_Y in the sense of singularities of mappings.

We recall that when 0 is a fold M-singularity of X then associated to the fold mapping ρ_X there is the symmetric diffeomorphism β_X that satisfies $\rho_X \circ \beta_X = \rho_X$.

Given $Z = (X, Y) \in \Omega(n+1)$ such that ρ_X and ρ_Y are fold mappings with $X^2h(0) < 0$ and $Y^2h(0) > 0$ then the composition of the associated symmetric mappings β_X and β_Y provides a first return mapping β_Z associated to Z and M. This situation is usually called a *distinguished fold-fold* singularity, and the mapping β_Z plays a fundamental role in the study of the dynamics of Z.

Codimension-one M-Singularity in Dimensions Two and Three

Case $n = 1$ In this case the unique codimension-zero M-singularity is a fold point in $R^2, 0$. The codimension-one M-singularities are represented by the subset $\Gamma_1(2)$ of $\chi_1(2)$ and it is defined as follows.

Definition 7 A codimension-one M-singularity of $X \in \Gamma_1(2)$ is either a *cusp* singularity or an M-hyperbolic critical point p in M of the vector field X. A cusp singularity (illustrated in Fig. 2) is characterized by $Xh(p) = X^2h(p) = 0$, $X^3h(p) \neq 0$. In the second case this means that p is a hyperbolic critical point (illustrated in Fig. 3) of X with distinct eigenvalues and with invariant manifolds (stable, unstable and strong stable and strong unstable) transversal to M.

In this subsubsection we consider a coordinate system in $R^2, 0$ such that $h(x, y) = y$.

The next result was proved in [46]. It presents the normal forms of the codimension-one singularities defined above.

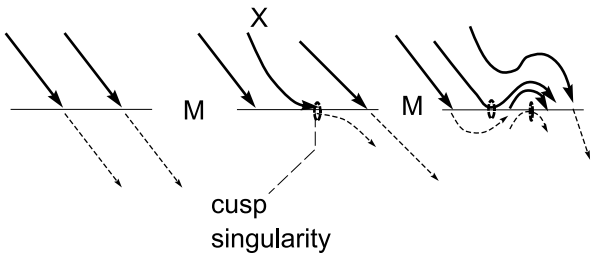

Perturbation Theory for Non-smooth Systems, Figure 2
The cusp singularity and its unfolding

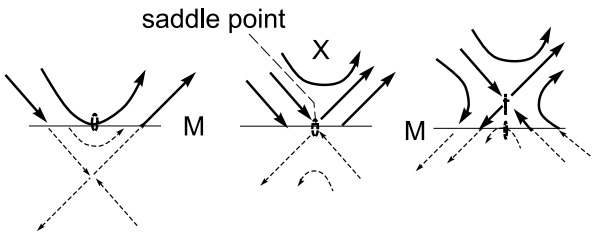

Perturbation Theory for Non-smooth Systems, Figure 3
The saddle point in the boundary and its unfolding

Theorem 8 *Let $X \in \chi_1(2)$. The vector field X is M-structurally stable relative to $\chi_1(2)$ if and only if $X \in \Gamma_1(2)$. Moreover, $\Gamma_1(2)$ is an embedded codimension-one sub manifold and dense in $\chi_1(2)$. We still require that any one-parameter family X_λ, $(\lambda \in (-\varepsilon, \varepsilon))$ in $\chi(1)$ transverse to $\Gamma_1(2)$ at X_0, has one of the following normal forms:*

0.1: $X_\lambda(x, y) = (1, 0)$ *(regular point);*
0.2: $X_\lambda(x, y) = (1, x)$ *(fold singularity);*
1.1: $X_\lambda(x, y) = (1, \lambda + x^2)$ *(cusp singularity);*
1.2: $X_\lambda(x, y) = (ax, x + by + \lambda)$, $a = \pm 1, b = \pm 2$;
1.3: $X_\lambda(x, y) = (x, x - y + \lambda)$;
1.4: $X_\lambda(x, y) = (x + y, -x + y + \lambda)$.

Case $n = 2$

Definition 9 A vector field $X \in \chi(3)$ belongs to the set $\Gamma_1(a)$ if the following conditions hold:

(i) $X(0) = 0$ and 0 is a hyperbolic critical point of X;
(ii) the eigenvalues of $DX(0)$ are pairwise distinct and the corresponding eigenspaces are transversal to M at 0;
(iii) each pair of non complex conjugate eigenvalues of $DX(0)$ has distinct real parts.

Definition 10 A vector field $X \in \chi(3)$ belongs to the set $\Gamma_1(b)$ if $X(0) \neq 0, Xh(0) = 0, X^2h(0) = 0$ and one of the following conditions hold:

(1) $X^3h(0) \neq 0, rank\{Dh(0), DXh(0), DX^2h(0)\} = 2$ and 0 is a non-degenerate critical point of $Xh_{|M}$.
(2) $X^3h(0) = 0, X^4h(0) \neq 0$ and 0 is a regular point of $Xh_{|M}$.

The next results can be found in [43].

Theorem 11 *The following statements hold:*

(i) $\Gamma_1(3) = \Gamma_1(a) \cup \Gamma_2(b)$ *is a codimension-one submanifold of* $\chi(3)$.
(ii) $\Gamma_1(3)$ *is open and dense set in* $\chi_1(3)$ *in the topology induced from* $\chi_1(3)$.
(iv) For a residual set of smooth curves $\gamma : R, 0 \to \chi(3), \gamma$ *meets* $\Gamma_1(3)$ *transversally.*

Throughout this subsubsection we fix the function $h(x, y, z) = z$.

Lemma 12 (Classification Lemma) *The elements of $\Gamma_1(3)$ are classified as follows:*

(a_{11}) *Nodal M-Singularity:* $X(0) = 0$, *the eigenvalues of* $DX(0), \lambda_1, \lambda_2, and \lambda_3$, *are real, distinct,* $\lambda_1\lambda_j > 0$, $j = 2, 3$ *and the eigenspaces are transverse to M at 0;*

(a_{12}) *Saddle M-Singularity:* $X(0) = 0$, *the eigenvalues of* $DX(0), \lambda_1, \lambda_2$ *and* λ_3, *are real, distinct,* $\lambda_1\lambda_j < 0$, $j = 2$ *or* 3 *and the eigenspaces are transverse to M at 0;*

(a_{13}) *Focal M-Singularity:* *0 is a hyperbolic critical point of X, the eigenvalues of* $DX(0)$ *are* $\lambda_{12} = a \pm ib, \lambda_3 = c$, *with* a, b, c *distinct from zero and* $c \neq a$, *and the eigenspaces are transverse to M at 0.*

(b_{11}) *Lips M-Singularity:* *presented in* ***Definition 8, item 1****, when* $Hess(Fh_{/S}(0)) > 0$:

(b_{12}) *Bec to Bec M-Singularity:* *presented in* ***Definition 8, item 1****, when* $Hess(Fh_{/S}(0)) < 0$;

(b_{13}) *Dove's Tail M-Singularity:* *presented in* ***Definition 8, item 2****.*

The next result is proved in [38]. It deals with the normal forms of a codimension-one singularity.

Theorem 13 *i) (Generic Bifurcation and normal forms) Let $X \in \chi(3)$. The vector field X is M-structurally stable relative to $\chi_1(3)$ if and only if $X \in \Gamma_1(3)$. ii) (Versal unfolding) In the space of one-parameter families of vector fields X_α in $\chi(3), \alpha \in (-\varepsilon, \varepsilon)$ an everywhere dense set is formed by generic families such that their normal forms are:*

- $X_\alpha \in \Gamma_0(3)$
  0.1: $X_\alpha(x, y, z) = (0, 0, 1)$
  0.2: $X_\alpha(x, y, z) = (z, 0, \pm x)$
  0.3: $X_\alpha(x, y, z) = (z, 0, x^2 + y)$

- $X_0 \in \Gamma_1(3)$
  1.1: $X_\alpha(x,y,z) = (z, 0, \frac{-3x^2+y^2+\alpha}{2})$
  1.2: $X_\alpha(x,y,z) = (z, 0, \frac{-3x^2-y^2+\alpha}{2})$
  1.3: $X_\alpha(x,y,z) = (z, 0, \frac{4\delta x^3+y+\alpha x}{2})$, *with* $\delta = \pm 1$
  1.4: $X_\alpha(x,y,z) = (axz, byz, \frac{ax+by+cz^2+\alpha}{2})$, *with* $(a,b,c) = \delta(3,2,1)$, $\delta = \pm 1$
  1.5: $X_\alpha(x,y,z) = (axz, byz, \frac{ax+by+cz^2+\alpha}{2})$, *with* $(a,b,c) = \delta(1,3,2)$, $\delta = \pm 1$
  1.6: $X_\alpha(x,y,z) = (axz, byz, \frac{ax+by+cz^2+\alpha}{2})$, *with* $(a,b,c) = \delta(1,2,3)$, $\delta = \pm 1$
  1.7: $X_\alpha(x,y,z) = (xz, 2yz, \frac{x+2y-cz^2+\alpha}{2})$
  1.8: $X_\alpha(x,y,z) = ((-x+y)z, (-x-y)z, \frac{-3x-y+z^2+\alpha}{2})$

Generic Bifurcation

Let $Z = (X,Y) \in \Omega^r(n+1)$. Call by $\Sigma_0(n+1)$ (resp. $\Sigma_1(n+1)$) the set of all elements that are structurally stable in $\Omega^r(n+1)$ (resp. $\Omega^r_1(n+1) = \Omega^r(n+1)_{\setminus \Sigma_0(n+1)}$) in $\Omega^r(n+1)$. It is clear that a pre-classification of the generic singularities is immediately reached by:

If $Z = (X,Y) \in \Sigma_0(n+1)$ (resp. $Z = (X,Y) \in \Sigma_1(n+1)$) then X and Y are in $\Gamma_0(n+1)$ (resp. $X \in \Gamma_0(n+1)$ and $Y \in \Gamma_1(n+1)$ or vice versa). Of course, the case when both X and Y are in $\Gamma_1(n+1)$ is a-codimension-two phenomenon.

Two-Dimensional Case

The following result characterizes the structural stability in $\Omega^r(2)$.

Theorem A *(see [31,44]): $\Sigma_0(2)$ is an open and dense set of $\Omega^r(2)$. The vector field $Z = (X,Y)$ is in $\Sigma_0(2)$ if and only one of the following conditions is satisfied:*

i) Both elements X and Y are regular. When $0 \in M$ is a simple singularity of Z then we assume that it is a hyperbolic critical point of Z^M.
ii) X is a fold singularity and Y is regular (and vice-versa).

The following result still deserves a systematic proof. Following the same strategy stipulated in the generic classification of an M-singularity, Theorem 11 could be very useful. It is worthwhile to mention [33] where the problem of generic bifurcation in $2D$ was also addressed.

Theorem B (Generic Bifurcation) *(see [36,43]) $\Sigma_1(2)$ is an open and dense set of $\Omega^r_1(2)$. The vector field $Z = (X,Y)$ is in $\Sigma_1(2)$ provided that one of the following conditions is satisfied:*

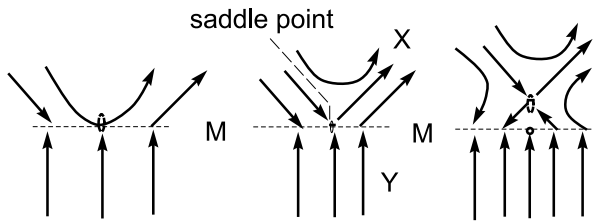

Perturbation Theory for Non-smooth Systems, Figure 4
***M*-critical point for *X*, *M*-regular for *Y* and its unfolding**

i) Both elements X and Y are M-regular. When $0 \in M$ is a simple singularity of Z then we assume that it is a codimension-one critical point (saddle-node or a Bogdanov–Takens singularity) of Z^M.
ii) 0 is a codimension-one M-singularity of X and Y is M-regular. This case includes when 0 is either a cusp M-singularity or a critical point. Figure 4 illustrates the case when 0 is a saddle critical point in the boundary.
iii) Both X and Y are fold M-singularities at 0. In this case we have to impose that 0 is a hyperbolic critical point of the C^r-extension of Z^M provided that it is in the boundary of $M_2 \cup M_3$ (see example below). Moreover when 0 is a distinguished fold-fold singularity of Z then 0 is a hyperbolic fixed point of the first return mapping β_Z.

Consider in a small neighborhood of 0 in R^2, the system $Z = (X,Y)$ with $X(x,y) = (1 - x^3 + y^2, x)$, $Y(x,y) = (1+x+y, -x+x^2)$ and $h(x,y) = y$. The point 0 is a *fold-fold*-singularity of Z with $M_2 = \{x < 0\}$ and $Z^s(x,0) = (2x - x^2)^{-1}(2x - x^4 + x^5)$. Observe that 0 is a hyperbolic critical point of the extended system $G(x,y) = 2x - x^4 + x^5$.

The classification of the codimension-two singularities in $\Omega^r(2)$ is still an open problem. In this direction [51] contains information about the classification of codimension-two M-singularities.

Three-Dimensional Case

Let $Z = (X,Y) \in \Omega^r(3)$.

The most interesting case to be analyzed is when both vector fields, X and Y are fold singularities at 0 and the tangency sets τ_X and τ_Y in M are in general position at 0. In fact they determine (in M) four quadrants, two of them are M_1-regions, one is an M_3-region and the other is an M_2-region (see Fig. 5). We emphasize that the sliding vector field Z^M can be C^r-extended to a full neighborhood of 0 in M. Moreover, $Z^M(0) = 0$. Inside this class the *distinguished fold-fold* singularity (as defined in Sub-

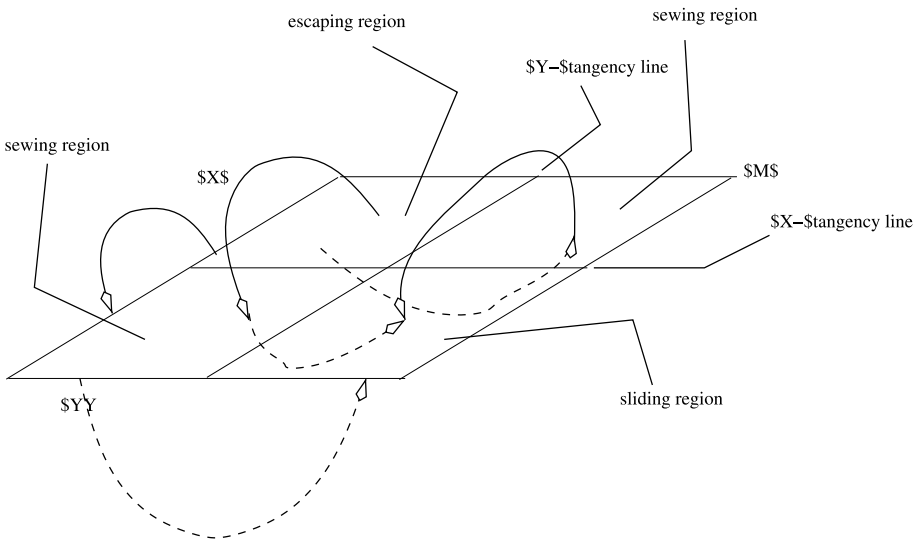

Perturbation Theory for Non-smooth Systems, Figure 5
The distinguished fold-fold singularity

sect. "A Construction") must be taken into account. Denote by A the set of all *distinguished fold-fold singularities* $Z \in \Omega^r(3)$. Moreover, the eigenvalues of $D\beta_Z(0)$ are $\lambda = a \pm \sqrt{(a^2-1)}$. If $\lambda \in R$ we say that Z belongs to A_s. Otherwise Z is in A_e. Recall that β_Z is the first return mapping associated to Z and M at 0 as defined in Subsect. "A Construction".

It is evident that the elements in the open set A_e are structurally unstable in $\Omega^r(3)$. It is worthwhile to mention that in A_e we detect elements which are asymptotically stable at the origin [48]. Concerning A_s few things are known.

We have the following result:

Theorem C *The vector field $Z = (X, Y)$ belongs to $\Sigma_0(3)$ provided that one of the following conditions occurs:*

i) Both elements X and Y are regular. When $0 \in M$ is a simple singularity of Z then we assume that it is a hyperbolic critical point of Z^M.
ii) X is a fold singularity at 0 and Y is regular.
iii) X is a cusp singularity at 0 and Y is regular.
*iv) Both systems X and Y are of fold type at 0. Moreover: a) the tangency sets τ_X and τ_Y are in general position at 0 in M; b) The eigenspaces associated with Z^M are transverse to τ_X and τ_Y at $0 \in M$ and c) Z **is not in** A. Moreover the real parts of non conjugate eigenvalues are distinct.*

We recall that bifurcation diagrams of sliding vector fields are presented in [50,52].

Singular Perturbation Problem in 2D

Geometric singular perturbation theory is an important tool in the field of continuous dynamical systems. Needless to say that in this area very good surveys are available (refer to [16,18,30]). Here we highlight some results (see [10]) that bridge the space between discontinuous systems in $\Omega^r(2)$ and singularly perturbed smooth systems.

Definition 14 Let $U \subset R^2$ be an open subset and $\varepsilon \geq 0$. A singular perturbation problem in U (SP-Problem) is a differential system which can be written as

$$x' = \frac{\mathrm{d}x}{\mathrm{d}}\tau = f(x, y, \varepsilon), \quad y' = \frac{\mathrm{d}y}{\mathrm{d}}\tau = \varepsilon g(x, y, \varepsilon) \quad (6)$$

or equivalently, after the time re-scaling $t = \varepsilon\tau$

$$\varepsilon\dot{x} = \varepsilon\frac{\mathrm{d}x}{\mathrm{d}}t = f(x, y, \varepsilon), \quad \dot{y} = \frac{\mathrm{d}y}{\mathrm{d}}t = g(x, y, \varepsilon), \quad (7)$$

with $(x, y) \in U$ and f, g smooth in all variables.

Our first result is concerned with the transition between non-smooth systems and GSPT.

Theorem D *Consider $Z \in \Omega^r(2)$, Z_ε its φ-regularization, and $p \in M$. Suppose that φ is a polynomial of degree k*

in a small interval $I \subseteq (-1,1)$ with $0 \in I$. Then the trajectories of Z_ε in $V_\varepsilon = \{q \in R^2, 0\colon h(q)/\varepsilon \in I\}$ are in correspondence with the solutions of an ordinary differential equation $z' = \alpha(z,\varepsilon)$, satisfying that α is smooth in both variables and $\alpha(z,0) = 0$ for any $z \in M$. Moreover, if $((X-Y)h^k)(p) \neq 0$ then we can take a C^{r-1}-local coordinate system $\{(\partial/\partial x)(p), (\partial/\partial y)(p)\}$ such that this smooth ordinary differential equation is a SP-problem.

The understanding of the phase portrait of the vector field associated to a SP-problem is the main goal of the *geometric singular perturbation-theory* (GSP-theory). The techniques of GSP-theory can be used to obtain information on the dynamics of (6) for small values of $\varepsilon > 0$, mainly in searching minimal sets.

System (6) is called the *fast system*, and (7) the *slow system* of the SP-problem. Observe that for $\varepsilon > 0$ the phase portraits of the fast and the slow systems coincide.

Theorem D says that we can transform a discontinuous vector field in a SP-problem. In general this transition cannot be done explicitly. Theorem E provides an explicit formula of the SP-problem for a suitable class of vector fields. Before the statement of such a result we need to present some preliminaries.

Consider $C = \{\xi\colon R^2, 0 \to R\}$ with $\xi \in C^r$ and $L(\xi) = 0$ where $L(\xi)$ denotes the linear part of ξ at $(0,0)$.

Let $\Omega_d \subset \Omega^r(2)$ be the set of vector fields $Z = (X,Y)$ in $\Omega^r(2)$ such that there exists $\xi \in C$ that is a solution of

$$\nabla\xi(X-Y) = \Pi_i(X-Y)\,, \tag{8}$$

where $\nabla\xi$ is the gradient of the function and Π_i denote the canonical projections, for $i = 1$ or $i = 2$.

Theorem E *Consider $Z \in \Omega_d$ and Z_ε its φ-regularization. Suppose that φ is a polynomial of degree k in a small interval $I \subset R$ with $0 \in I$. Then the trajectories of Z_ε on $V_\varepsilon = \{q \in R^2, 0\colon h(q)/\varepsilon \in I\}$ are solutions of a SP-problem.*

We remark that the singular problems discussed in the previous theorems, when $\varepsilon \searrow 0$, defines a dynamical system on the discontinuous set of the original problem. This fact can be very useful for problems in Control Theory.

Our third theorem says how the fast and the slow systems approximate the discontinuous vector field. Moreover, we can deduce from the proof that whereas the fast system approximates the discontinuous vector field, the slow system approaches the corresponding sliding vector field.

Consider $Z \in \Omega^r(2)$ and $\rho\colon R^2, 0 \longrightarrow R$ with $\rho(x,y)$ being the distance between (x,y) and M. We denote by $\widehat{Z}$ the vector field given by $\widehat{Z}(x,y) = \rho(x,y)Z(x,y)$.

In what follows we identify $\widehat{Z}_\varepsilon$ and the vector field on $\{\{R^2,0\} \setminus M \times R\} \subset R^3$ given by $\widehat{Z}(x,y,\varepsilon) = (\widehat{Z}_\varepsilon(x,y), 0)$.

Theorem F *Consider $p = 0 \in M$. Then there exists an open set $U \subset R^2$, $p \in U$, a three-dimensional manifold M, a smooth function $\Phi\colon M \longrightarrow R^3$ and a SP-problem W on M such that Φ sends orbits of $W|_{\Phi^{-1}(U\times(0,+\infty))}$ in orbits of $\widehat{Z}|_{(U\times(0,+\infty))}$.*

Examples

1. Take $X(x,y) = (1,x)$, $Y(x,y) = (-1,-3x)$, and $h(x,y) = y$. The discontinuity set is $\{(x,0) \mid x \in R\}$. We have $Xh = x$, $Yh = -3x$, and then the unique non-regular point is $(0,0)$. In this case we may apply Theorem E.
2. Let $Z_\varepsilon(x,y) = \big(y/\varepsilon, 2xy/\varepsilon - x\big)$. The associated partial differential equation (refer to Theorem E) with $i = 2$ given above becomes $2(\partial\xi/\partial x) + 4x(\partial\xi/\partial y) = 4x$. We take the coordinate change $\overline{x} = x, \overline{y} = y - x^2$. The trajectories of X_ε in these coordinates are the solutions of the singular system
$$\varepsilon\dot{\overline{x}} = \overline{y} + \overline{x}^2\,, \quad \dot{\overline{y}} = -\overline{x}\,.$$
3. In what follows we try, by means of an example, to present a rough idea on the transition from non-smooth systems to GSPT. Consider $X(x,y) = (3y^2 - y - 2, 1)$, $Y(x,y) = (-3y^2 - y + 2, -1)$ and $h(x,y) = x$. The regularized vector field is
$$Z_\varepsilon(x,y) = \left(\frac{1}{2} + \frac{1}{2}\varphi\left(\frac{x}{\varepsilon}\right)\right)(3y^2 - y - 2, 1) + \left(\frac{1}{2} - \frac{1}{2}\varphi\left(\frac{x}{\varepsilon}\right)\right)(-3y^2 - y + 2, -1)\,.$$

After performing the polar blow up coordinates $\alpha\colon [0,+\infty) \times [0,\pi] \times R \to R^3$ given by $x = r\cos\theta$ and $\varepsilon = r\sin\theta$ the last system is expressed by:

$$r\dot{\theta} = -\sin\theta(-y + \varphi(\cot\theta)(3y^2 - 2))\,, \quad \dot{y} = \varphi(\cot\theta)\,.$$

So the slow manifold is given implicitly by $\varphi(\cot\theta) = y/(3y^2 - 2)$ which defines two functions $y_1(\theta) = (1 + \sqrt{1 + 24\varphi^2(\cot\theta)})/(6\varphi(\cot\theta))$ and $y_2(\theta) = (1 - \sqrt{1 + 24\varphi^2(\cot\theta)})/(6\varphi(\cot\theta))$. The function $y_1(\theta)$ is increasing, $y_1(0) = 1, \lim_{\theta\to\pi/2-} y_1(\theta) = +\infty$, $\lim_{\theta\to\pi/2+} y_1(\theta) = -\infty$ and $y_1(\pi) = -1$. The function

$y_2(\theta)$ is increasing, $y_2(0) = -2/3, \lim_{\theta\to\pi/2} y_2(\theta) = 0$ and $y_2(\pi) = 2/3$. We can extend y_2 to $(0, \pi)$ as a differential function with $y_2(\pi/2) = 0$.

The fast vector field is $(\theta', 0)$ with $\theta' > 0$ if (θ, y) belongs to

$$\left[\left(0, \frac{\pi}{2}\right) \times (y_2(\theta), y_1(\theta)) \bigcup \left(\frac{\pi}{2}, \pi\right) \times (y_2(\theta), +\infty) \bigcup \left(\frac{\pi}{2}, \pi\right) \times (-\infty, y_1(\theta))\right]$$

and with $\theta' < 0$ if (θ, y) belongs to

$$\left[\left(0, \frac{\pi}{2}\right) \times (y_1(\theta), +\infty) \bigcup \left(0, \frac{\pi}{2}\right) \times (-\infty, y_2(\theta)) \bigcup \left(\frac{\pi}{2}, \pi\right) \times (y_1(\theta), y_2(\theta))\right].$$

The reduced flow has one singular point at $(0, 0)$ and it takes the positive direction of the y-axis if $y \in (-\frac{2}{3}, 0) \cup (1, \infty)$ and the negative direction of the y-axis if $y \in (-\infty, -1) \cup (0, \frac{2}{3})$.

One can see that the singularities $(\theta, y, r) = (0, 1, 0)$ and $(\theta, y, r) = (0, -1, 0)$ are not normally hyperbolic points. In this way, as usual, we perform additional blow ups. In Fig. 6 we illustrate the fast and the slow dynamics of the SP-problem. We present a phase portrait on the blowing up locus where a double arrow over a trajectory means that the trajectory belongs to the fast dynamical system, and a simple arrow means that the trajectory belongs to the slow dynamical system.

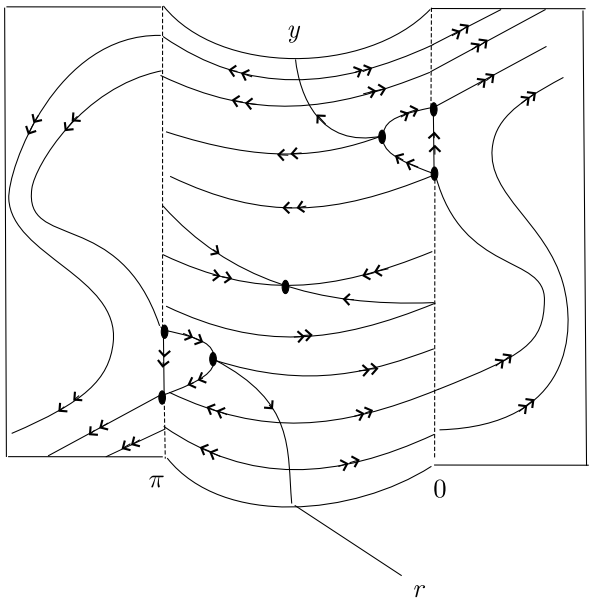

Perturbation Theory for Non-smooth Systems, Figure 6
Example of fast and slow dynamics of the SP-Problem

Future Directions

Our concluding section is devoted to an outlook. Firstly we present some open problems linked with the setting that point out future directions of research. The main task for the future seems to bring the theory of non-smooth dynamical systems to a similar maturity as that of smooth systems. Finally we briefly discuss the main results in this text.

Some Problems

In connection to this present work, some theoretical problems remain open:

1. The description of the bifurcation diagram of the codimension-two singularities in $\Omega(2)$. In this last class we find some models (see [37]) where the following questions can also be addressed. a) When is a typical singularity topologically equivalent to a regular center? b) How about the isochronicity of such a center? c) When does a polynomial perturbation of such a system in $\Omega(2)$ produce limit cycles? The articles [9,13,21,22,47] can be useful auxiliary references.
2. Let $\Omega(N)$ be the set of all non-smooth vector fields on a two-dimensional compact manifold N having a codimension-one compact submanifold M as its discontinuity set. The problem is to study the global generic bifurcation in $\Omega(N)$. The articles [31,33,40,46] can be useful auxiliary references.
3. Study of the bifurcation set in $\Omega^r(3)$. The articles [38,40,43,50] can be useful auxiliary references.
4. Study of the dynamics of the distinguished *fold-fold* singularity in $\Omega^r(n+1)$. The article [48] can be a useful auxiliary reference.
5. In many applications examples of non-smooth systems where the discontinuities are located on algebraic varieties are available. For instance, consider the system $\ddot{x} + x sign(x) + sign(\dot{x}) = 0$. Motivated by such models we present the following problem. Let 0 be a non-degenerate critical point of a smooth mapping $h\colon R^{n+1}, 0 \to R, 0$. Let $\Phi(n+1)$ be the space of all vector fields Z on $R^{n+1}, 0$ defined in the same way as $\Omega(n+1)$. We propose the following. i) Classify the typical singularities in that space. ii) Analyze the elements of $\Phi(2)$ by means of "regularization processes" and the methods of GSPT, similarly to Sect. "Vector Fields near

the Boundary". The articles [1,2] can be very useful auxiliary references.

6. In [27,29] classes of $4D$-relay systems are considered. Conditions for the existence of one-parameter families of periodic orbits terminating at typical singularities are provided. We propose to find conditions for the existence of such families for *n-dimensional* relay systems.

Conclusion

In this paper we have presented a compact survey of the geometric/qualitative theoretical features of non-smooth dynamical systems. We feel that our survey illustrates that this field is still in its early stages but enjoying growing interest. Given the importance and the relevance of such a theme, we have pointed above some open questions and we remark that there is still a wide range of bifurcation problems to be tackled. A brief summary of the main results in the text is given below.

1. We firstly deal with two-*dimensional* non-smooth vector fields $Z = (X, Y)$ defined around the origin in R^2, where the discontinuity set is concentrated on the line $\{y = 0\}$. The first task is to characterize those systems which are structurally stable. This characterization is a starting point with which to establish a bifurcation theory as indicated by the Thom–Smale program.
2. In higher dimension the problem becomes much more complicated. We have presented here sufficient conditions for the three-*dimensional* local structural stability. Any further investigation on bifurcation in this context must pass through a deep analysis of the so called *fold-fold* singularity.
3. We have established a bridge between discontinuous and singularly perturbed smooth systems. Many similarities between such systems were observed and a comparative study of the two categories is called for.

Bibliography

Primary Literature

1. Alexander JC, Seidman TI (1998) Sliding modes in intersecting switching surfaces I: Blending. Houston J Math 24(3):545–569
2. Alexander JC, Seidman TI (1999) Sliding modes in intersecting switching surfaces II: Hysteresis. Houston J Math 25:185–211
3. Andronov A, Pontryagin S (1937) Structurally stable systems. Dokl Akad Nauk SSSR 14:247–250
4. Andronov AA, Vitt AA, Khaikin SE (1966) Theory of ocillators, Dover, New York
5. Anosov DV (1959) Stability of the equilibrium positions in relay systems. Autom Remote Control XX(2):135–149
6. Arnold VI (1983) Methods in the theory of ordinary differential equations. Springer, New York
7. Bazykin AD (1998) Nonlinear dynamics of interacting populations. World Sc. Publ. Co. Inc., River–Edge, NJ
8. Bonnard B, Chyba M (2000) Singular trajectories and their role in control theory. Mathématiques and Applications, vol 40. Springer, Berlin
9. Broucke ME, Pugh CC, Simić SN (2001) Structural stability of piecewise smooth systems. Comput Appl Math 20(1–2):51–89
10. Buzzi C, Silva PR, Teixeira MA (2006) A singular approach to discontinuous vector fields on the plane. J Differ Equ 23:633–655
11. Chillingworth DR (2002) J.(4-SHMP) Discontinuity geometry for an impact oscillator. Dyn Syst 17(4):389–420
12. Chua LO (1992) The genesis of Chua's circuit. AEU 46:250
13. Coll B, Gasull A, Prohens R (2001) Degenerate Hopf bifurcations in discontinuous planar systems. J Math Anal Appl 253(2):671–690
14. di Bernardo M, Budd C, Champneys AR, Kowalczyk P, Nordmark AB, Olivar G, Piiroinen PT (2005) Bifurcations in non-smooth dynamical systems. Bristol Centre for Applied Nonlinear Mathematics, N. 2005-4, Bristol
15. Dumortier F (1977) Singularities of vector fields. IMPA, Rio de Janeiro
16. Dumortier F, Roussarie R (1996) Canard cycles and center manifolds. Mem Amer Mat Soc 121:x+100
17. Ekeland I (1977) Discontinuités de champs hamiltoniens et existence de solutions optimales en calcul des variations. Inst Hautes Études Sci Publ Math 47:5–32
18. Fenichel N (1979) Geometric singular perturbation theory for ordinary differential equations. J Differ Equ 31:53–98
19. Filippov AF (1988) Differential equations with discontinuous righthand sides. Kluwer Academic, Dordrecht
20. Flúgge-Lotz I (1953) Discontinuous automatic control. Princeton University, Princeton, pp vii+168
21. Garcia R, Teixeira MA (2004) Vector fields in manifolds with boundary and reversibility-an expository account. Qual Theory Dyn Syst 4:311–327
22. Gasull A, Torregrosa J (2003) Center-focus problem for discontinuous planar differential equations. Int J Bifurc Chaos Appl Sci Eng 13(7):1755–1766
23. Henry P (1972) Diff. equations with discontinuous right-hand side for planning procedures. J Econ Theory 4:545–551
24. Hogan S (1989) On the dynamics of rigid-block motion under harmonic forcing. Proc Roy Soc London A 425:441–476
25. Ito T (1979) A Filippov solution of a system of diff. eq. with discontinuous right-hand sides. Econ Lett 4:349–354
26. Jackson JD (1999) Classical electrodynamics, 3rd edn. Wiley, New York, pp xxii+808
27. Jacquemard A, Teixeira MA (2003) Computer analysis of periodic orbits of discontinuous vector fields. J Symbol Comput 35:617–636
28. Jacquemard A, Teixeira MA (2003) On singularities of discontinuous vector fields. Bull Sci Math 127:611–633
29. Jacquemard A, Teixeira MA (2005) Invariant varieties of discontinuous vector fields. Nonlinearity 18:21–43
30. Jones C (1995) Geometric singular perturbation theory. C.I.M.E. Lectures, Montecatini Terme, June 1994, Lecture Notes in Mathematics 1609. Springer, Heidelberg

31. Kozlova VS (1984) Roughness of a discontinuous system. Vestinik Moskovskogo Universiteta, Matematika 5:16–20
32. Kunze M, Kupper T (1997) Qualitative bifurcation analysis of a non-smooth friction oscillator model. Z Angew Math Phys 48:87–101
33. Kuznetsov YA et al (2003) One-parameter bifurcations in planar Filippov systems. Int J Bifurac Chaos 13:2157–2188
34. Llibre J, Teixeira MA (1997) Regularization of discontinuous vector fields in dimension 3. Discret Contin Dyn Syst 3(2):235–241
35. Luo CJ (2006) Singularity and dynamics on discontinuous vector fields. Monograph Series on Nonlinear Science and Complexity. Elsevier, New York, pp i+310
36. Machado AL, Sotomayor J (2002) Structurally stable discontinuous vector fields in the plane. Qual Theory Dyn Syst 3:227–250
37. Manosas F, Torres PJ (2005) Isochronicity of a class of piecewise continuous oscillators. Proc of AMS 133(10):3027–3035
38. Medrado J, Teixeira MA (1998) Symmetric singularities of reversible vector fields in dimension three. Physica D 112:122–131
39. Medrado J, Teixeira MA (2001) Codimension-two singularities of reversible vector fields in 3D. Qualit Theory Dyn Syst J 2(2):399–428
40. Percell PB (1973) Structural stability on manifolds with boundary. Topology 12:123–144
41. Seidman T (2006) Aspects of modeling with discontinuities. In: N'Guerekata G (ed) Advances in Applied and Computational Mathematics Proc Dover Conf, Cambridge. http://www.umbc.edu/~seideman
42. Sotomayor J (1974) Structural stability in manifolds with boundary. In: Global analysis and its applications, vol III. IEAA, Vienna, pp 167–176
43. Sotomayor J, Teixeira MA (1988) Vector fields near the boundary of a 3-manifold. Lect Notes in Math, vol 331. Springer, Berlin-Heidelberg, pp 169–195
44. Sotomayor J, Teixeira MA (1996) Regularization of discontinuous vector fields. International Conference on Differential Equations, Lisboa. World Sc, Singapore, pp 207–223
45. Sussmann H (1979) Subanalytic sets and feedback control. J Differ Equ 31:31–52
46. Teixeira MA (1977) Generic bifurcations in manifolds with boundary. J Differ Equ 25:65–89
47. Teixeira MA (1979) Generic bifurcation of certain singularities. Boll Unione Mat Ital 16-B:238–254
48. Teixeira MA (1990) Stability conditions for discontinuous vector fields. J Differ Equ 88:15–24
49. Teixeira MA (1991) Generic Singularities of Discontinuous Vector Fields. An Ac Bras Cienc 53(2):257–260
50. Teixeira MA (1993) Generic bifurcation of sliding vector fields. J Math Anal Appl 176:436–457
51. Teixeira MA (1997) Singularities of reversible vector fields. Physica D 100:101–118
52. Teixeira MA (1999) Codimension-two singularities of sliding vector fields. Bull Belg Math Soc 6(3):369–381
53. Vishik SM (1972) Vector fields near the boundary of a manifold. Vestnik Moskovskogo Universiteta, Matematika 27(1):13–19
54. Zelikin MI, Borisov VF (1994) Theory of chattering control with applications to astronautics, robotics, economics, and engineering. Birkhäuser, Boston

Books and Reviews

Agrachev AA, Sachkov YL (2004) Control theory from the geometric viewpoint. Encyclopaedia of Mathematical Sciences, 87. Control theory and optimization, vol II. Springer, Berlin, pp xiv+412

Barbashin EA (1970) Introduction to the theory of stability. Wolters–Noordhoff Publishing, Groningen. Translated from the Russian by Transcripta Service, London. Edited by T. Lukes, pp 223

Brogliato B (ed) (2000) Impacts in mechanical systems. Lect Notes in Phys, vol 551. Springer, Berlin, pp 160

Carmona V, Freire E, Ponce E, Torres F (2005) Bifurcation of invariant cones in piecewise linear homogeneous systems. Int J Bifurc Chaos 15(8):2469–2484

Davydov AA (1994) Qualitative theory of control systems. American Mathematical Society, Providence, RI. (English summary) Translated from the Russian manuscript by V. M. Volosov. Translations of Mathematical Monographs, 141, pp viii+147

Dercole F, Gragnani F, Kuznetsov YA, Rinaldi S (2003) Numerical sliding bifurcation analysis: an application to a relay control system. IEEE Trans Circuit Systems – I: Fund Theory Appl 50:1058–1063

Glocker C (2001) Set-valued force laws: dynamics of non-smooth systems. Lecture Notes in Applied Mechanics, vol 1. Springer, Berlin

Hogan J, Champneys A, Krauskopf B, di Bernardo M, Wilson E, Osinga H, Homer M (eds) Nonlinear dynamics and chaos: Where do we go from here? Institute of Physics Publishing (IOP), Bristol, pp xi+358, 2003

Huertas JL, Chen WK, Madan RN (eds) (1997) Visions of nonlinear science in the 21st century, Part I. Festschrift dedicated to Leon O. Chua on the occasion of his 60th birthday. Papers from the workshop held in Sevilla, June 26, 1996. Int J Bifurc Chaos Appl Sci Eng 7 (1997), no. 9. World Scientific Publishing Co. Pte. Ltd., Singapore, pp i–iv, pp 1907–2173

Komuro M (1990) Periodic bifurcation of continuous piece-wise vector fields. In: Shiraiwa K (ed) Advanced series in dynamical systems, vol 9. World Sc, Singapore

Kunze M (2000) Non-smooth dynamical systems. Lecture Notes in Mathematics, vol 1744. Springer, Berlin, pp x+228

Kunze M, Küpper T (2001) Non-smooth dynamical systems: an overview. In: Fiedler B (ed) Ergodic theory, analysis, and efficient simulation of dynamical systems. Springer, Berlin, pp 431–452

Llibre J, Silva PR, Teixeira MA (2007) Regularization of discontinuous vector fields on R^3 via singular perturbation. J Dyn Differ Equ 19(2):309–331

Martinet J (1974) Singularités des fonctions et applications différentiables. Pontifícia Universidade Catolica do Rio de Janeiro, Rio de Janeiro, pp xiii+181 (French)

Minorsky N (1962) Nonlinear oscillations. D Van Nostrand Co. Inc., Princeton, N.J., Toronto, London, New York, pp xviii+714

Minorsky N (1967) Théorie des oscillations. Mémorial des Sciences Mathématiques Fasc, vol 163. Gauthier-Villars, Paris, pp i+114 (French)

Minorsky N (1969) Theory of nonlinear control systems. McGraw-Hill, New York, London, Sydney, pp xx+331

Ostrowski JP, Burdick JW (1997) Controllability for mechanical systems with symmetries and constraints. Recent developments in robotics. Appl Math Comput Sci 7(2):305–331

Peixoto MC, Peixoto MM (1959) Structural Stability in the plane with enlarged conditions. Anais da Acad Bras Ciências 31:135–160
Rega G, Lenci S (2003) Nonsmooth dynamics, bifurcation and control in an impact system. Syst Anal Model Simul Arch 43(3), Gordon and Breach Science Publishers, Inc. Newark, pp 343–360
Seidman T (1995) Some limit problems for relays. In: Lakshmikantham V (ed) Proc First World Congress of Nonlinear Analysis, vol I. Walter de Gruyter, Berlin, pp 787–796
Sotomayor J (1974) Generic one-parameter families of vector fields on 2-manifolds. Publ Math IHES 43:5–46
Szmolyan P (1991) Transversal heteroclinic and homoclinic orbits in singular perturbation problems. J Differ Equ 92:252–281
Teixeira MA (1985) Topological invariant for discontinuous vector fields. Nonlinear Anal TMA 9(10)1073–1080
Utkin V (1978) Sliding modes and their application in variable structure systems. Mir, Moscow
Utkin V (1992) Sliding Modes in Control and Optimization. Springer, Berlin

Perturbation Theory for PDEs

Dario Bambusi
Dipartimento di Matematica, Università degli Studi di Milano, Milano, Italia

Article Outline

Glossary
Definition of the Subject
Introduction
The Hamiltonian Formalism for PDEs
Normal Form for Finite Dimensional Hamiltonian Systems
Normal Form for Hamiltonian PDEs: General Comments
Normal Form for Resonant Hamiltonian PDEs and its Consequences
Normal Form for Nonresonant Hamiltonian PDEs
Non Hamiltonian PDEs
Extensions and Related Results
Future Directions
Bibliography

Glossary

Perturbation theory The study of a dynamical systems which is a perturbation of a system whose dynamics is known. Typically the unperturbed system is linear or integrable.

Normal form The normal form method consists of constructing a coordinate transformation which changes the equations of a dynamical system into new equations which are as simple as possible. In Hamiltonian systems the theory is particularly effective and typically leads to a very precise description of the dynamics.

Hamiltonian PDE A Hamiltonian PDE is a partial differential equation (abbreviated PDE) which is equivalent to the Hamilton equation of a suitable Hamiltonian function. Classical examples are the nonlinear wave equation, the Nonlinear Schrödinger equation, and the Kortweg–de Vries equation.

Resonance vs. Non-Resonance A frequency vector $\{\omega_k\}_{k=1}^n$ is said to be non-resonant if its components are independent over the relative integers. On the contrary, if there exists a non-vanishing $K \in \mathbb{Z}^n$ such that $\omega \cdot K = 0$ the frequency vector is said to be resonant. Such a property plays a fundamental role in normal form theory. Non-resonance typically implies stability.

Actions The action of a harmonic oscillator is its energy divided by its frequency. It is usually denoted by I. The typical issue of normal form theory is that in nonresonant systems the actions remain approximatively unchanged for very long times. In resonant systems there are linear combinations of the actions with such properties.

Sobolev space Space of functions which have weak derivatives enjoying suitable integrability properties. Here we will use the spaces H^s, $s \in \mathbb{N}$ of the functions which are square integrable together with their first s weak derivatives.

Definition of the Subject

Perturbation theory for PDEs is a part of the qualitative theory of differential equations. One of the most effective methods of perturbation theory is the normal form theory which consists of using coordinate transformations in order to describe the qualitative features of a given or generic equation. Classical normal form theory for ordinary differential equations has been used all along the last century in many different domains, leading to important results in pure mathematics, celestial mechanics, plasma physics, biology, solid state physics, chemistry and many other fields.

The development of effective methods to understand the dynamics of partial differential equations is relevant in pure mathematics as well as in all the fields in which partial differential equations play an important role. Fluidodynamics, oceanography, meteorology, quantum mechanics, and electromagnetic theory are just a few examples of potential applications. More precisely, the normal form theory allows one to understand whether a small nonlinearity can change the dynamics of a linear PDE or not.

Moreover, it allows one to understand how the changes can be avoided or forced. Finally, when the changes are possible it allows to predict the behavior of the perturbed system.

Introduction

The normal form method was developed by Poincaré and Birkhoff between the end of the 19th century and the beginning of the 20th century. During the last 20 years the method has been successfully generalized to a suitable class of partial differential equations (PDEs) in finite volume (in the case of infinite volume dispersive effects appear and the theory is very different. See e. g. [58]). In this article we will give an introduction to this recent field. We will almost only deal with Hamiltonian PDEs, since on the one hand the theory for non Hamiltonian systems is a small variant of the one we will present here, and on the other hand most models are Hamiltonian.

We will start by a generalization of the Hamiltonian formalism to PDEs, followed by a review of the classical theory and by the actual generalization of normal form theory to PDEs.

In the next section we give a generalization of the Hamiltonian formalism to PDEs. The main new fact is that in PDEs the Hamiltonian is usually a smooth function, but the corresponding vector field is nonsmooth (it is an operator extracting derivatives). So the standard formalism has to be slightly modified [10,29,45,49,52,61]. Here we will present a version of the Hamiltonian formalism which is enough to cover the models of interest for local perturbation theory. To clearly illustrate the situation we will start the article with an introduction to the Lagrangian and Hamiltonian formalism for the wave equation. This will lead to the introduction of the paradigm Hamiltonian which is usually studied in this context. This will be followed by a few results on the Hamiltonian formalism that are needed for perturbation theory.

Subsequently, we shortly present the standard Birkhoff normal form theory for finite dimensional systems. This is useful since all the formal aspects are equal in the classical case and the case of PDEs.

Then we come to the generalization of normal form theory to PDEs. In the present paper we will concentrate almost only on the case of 1-dimensional semilinear equations. This is due to the fact that the theory of higher dimensional and quasilinear equations is still quite unsatisfactory.

In PDEs one essentially meets two kinds of difficulties. The first one is related to the existence of non smooth vector fields. The second difficulty is due to the fact that in the infinite dimensional case there are small denominators which are much worse than in the finite dimensional one.

We first present the theory for completely resonant systems [10,14] in which the difficulties related to small denominators do not appear. It turns out that it is quite easy to obtain a normal form theorem for resonant PDEs, but the kind of normal form one gets is usually quite poor. In order to extract dynamical informations from the normal form one can only compute and study it explicitly. Usually this is very difficult. Nevertheless in some cases it is possible and leads to quite strong results. We will illustrate such a situation by studying a nonlinear Schrödinger equation [2,25].

For the general case there is a theorem ensuring that a generic system admits at least one family of "periodic like trajectories" which are stable over exponentially long times [15]. We will give its statement and an application to the nonlinear wave equation

$$u_{tt} - u_{xx} + \mu^2 u + f(u) = 0\,, \tag{1}$$

with $\mu = 0$ and the Dirichlet boundary conditions on a segment [55].

Then we turn to the case of nonresonant PDEs. The main difficulty is that small denominators accumulate to zero already at order 3. Such a problem has been overcome in [4,6,9,12,43] by taking advantage of the fact that the nonlinearities appearing in PDEs typically have a special form. In this case one can deduce a very precise description of the dynamics and also some interesting results of the kind of almost global existence of smooth solutions [46]. To illustrate the theory we will make reference to the nonlinear wave Eq. (1) with almost any μ, and to the nonlinear Schrödinger equation.

Another aspect of the theory of close to integrable Hamiltonian PDEs concerns the extension of KAM theory to PDEs. We will not present it here. We just recall the most celebrated results which are those due to Kuksin [48], Wayne [60], Craig–Wayne [34], Bourgain [24,26], Kuksin–Pöschel [50], Eliasson–Kuksin [39]. All these results ensure the existence of families of quasiperiodic solutions, i. e. solutions lying on finite dimensional manifolds. We also mention the papers [21,57] where some Cantor families of full dimension tori are constructed. We point out that in the dynamics on such ∞-dimensional tori the amplitude of oscillation of the linear modes decreases super exponentially with their index. A remarkable exception is provided by the paper [27] where the tori constructed are more "thick" (even if of course they lie on Cantor families).

On the contrary, the results of normal form theory describe solutions starting on opens subsets of the phase space, and do not have particularly strong localizations properties with respect to the index. The price one has to pay is that the description one gets turns out to be valid only over long but finite times.

Finally we point out a related research stream that has been carried on by Bourgain [21,22,23,25] who studied intensively the behavior of high Sobolev norms in close to integrable Hamiltonian PDEs (see also [28]).

The Hamiltonian Formalism for PDEs

The Gradient of a Functional

Definition 1 Consider a function $f \in C^\infty(\mathcal{U}_s, \mathbb{R})$, $\mathcal{U}_s \subset H^s(\mathbb{T})$ open, $s \geq 0$ a fixed parameter and $\mathbb{T} := \mathbb{R}/2\pi\mathbb{Z}$ is the 1-dimensional torus. We will denote by $\nabla f(u)$ the gradient of f with respect to the L^2 metric, namely the unique function such that

$$\langle \nabla f(u), h\rangle_{L^2} = \mathrm{d}f(u)h\,, \quad \forall h \in H^s \tag{2}$$

where

$$\langle u, v\rangle_{L^2} := \int_{-\pi}^{\pi} u(x)v(x)\mathrm{d}x \tag{3}$$

is the L^2 scalar product and $\mathrm{d}f(u)$ is the differential of f at u. The gradient is a smooth map from H^s to H^{-s} (see e. g. [3]).

Example 2 Consider the function

$$f(u) := \int_{-\pi}^{\pi} \frac{u_x^2}{2}\mathrm{d}x\,, \tag{4}$$

which is differentiable as a function from $H^s \to \mathbb{R}$ for any $s \geq 1$. One has

$$\mathrm{d}f(u)h = \int_{-\pi}^{\pi} u_x h_x \mathrm{d}x = \int_{-\pi}^{\pi} -u_{xx} h\mathrm{d}x = \langle -u_{xx}, h\rangle_{L^2} \tag{5}$$

and therefore in this case one has $\nabla f(u) = -u_{xx}$.

Example 3 Let $\mathcal{F}: \mathbb{R}^2 \to \mathbb{R}$ be a smooth function and define

$$f(u) = \int_{-\pi}^{\pi} \mathcal{F}(u, u_x)\mathrm{d}x \tag{6}$$

then the gradient of f coincides with the so called functional derivative of $\mathcal{F}$:

$$\nabla f \equiv \frac{\delta \mathcal{F}}{\delta u} := \frac{\partial \mathcal{F}}{\partial u} - \frac{\partial}{\partial x}\frac{\partial \mathcal{F}}{\partial u_x}\,. \tag{7}$$

Lagrangian and Hamiltonian Formalism for the Wave Equation

Until Subsect. "Basic Elements of Hamiltonian Formalism for PDEs" we will work at a formal level, without specifying the function spaces and the domains.

Definition 4 Let $L(u, \dot{u})$ be a Lagrangian function, then the corresponding Lagrange equations are given

$$\nabla_u L - \frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\dot{u}} L = 0 \tag{8}$$

where $\nabla_u L$ is the gradient with respect to u only, and similarly $\nabla_{\dot{u}}$ is the gradient with respect to $\dot{u}$.

Example 5 Consider the Lagrangian

$$L(u, \dot{u}) := \int_{-\pi}^{\pi} \left(\frac{\dot{u}^2}{2} - \frac{u_x^2}{2} - \mu^2\frac{u^2}{2} - F(u)\right)\mathrm{d}x\,. \tag{9}$$

then the corresponding Lagrange equations are given by (1) with $f = -F'$.

Given a Lagrangian system with a Lagrangian function L, one defines the corresponding Hamiltonian system as follows.

Definition 6 Consider the momentum $v := \nabla_{\dot{u}} L$ conjugated to u; assume that L is convex with respect to $\dot{u}$, then the Hamiltonian function associated to L is defined by

$$H(v, u) := \left[\langle v; \dot{u}\rangle_{L^2} - L(u, \dot{u})\right]_{\dot{u}=\dot{u}(u,v)}\,. \tag{10}$$

Definition 7 Let $H(v, u)$ be a Hamiltonian function, then the corresponding Hamilton equations are given by

$$\dot{v} = -\nabla_u H\,, \quad \dot{u} = \nabla_v H\,. \tag{11}$$

As in the finite dimensional case, one has that the Lagrange equations are equivalent to the Hamilton equation of H.

An elementary computation shows that for the wave equation one has $v = \dot{u}$ and

$$H(v, u) = \int_{-\pi}^{\pi} \left(\frac{v^2 + u_x^2 + \mu^2 u^2}{2} + F(u)\right)\mathrm{d}x \tag{12}$$

Canonical Coordinates

Consider a Lagrangian system and let $\mathbf{e}_k$ be an orthonormal basis of L^2, write $u = \sum_k q_k \mathbf{e}_k$ and $\dot{u} = \sum_k \dot{q}_k \mathbf{e}_k$, then one has the following proposition.

Proposition 8 *The Lagrange Eqs. (8) are equivalent to*

$$\frac{\partial L}{\partial q_k} - \frac{d}{dt}\frac{\partial L}{\partial \dot{q}_k} = 0 \tag{13}$$

Proof Taking the scalar product of (8) with $\mathbf{e}_k$ one gets

$$\langle \mathbf{e}_k; \nabla_u L\rangle_{L^2} - \frac{\mathrm{d}}{\mathrm{d}t}\langle \mathbf{e}_k; \nabla_{\dot{u}} L\rangle_{L^2} = 0$$

but one has $\langle \mathbf{e}_k; \nabla_u L\rangle = \frac{\partial L}{\partial q_k}$ and similarly for the other term. Thus the thesis follows. □

This proposition shows that, once a basis has been introduced, the Lagrange equations have the same form as in the finite dimensional case.

In the Hamiltonian case exactly the same result holds. Precisely, denoting $v := \sum_k p_k \mathbf{e}_k$ one has the following proposition.

Proposition 9 *The Hamilton equations of a Hamiltonian function H are equivalent to*

$$\dot{p}_k = -\frac{\partial H}{\partial q_k}\,, \quad \dot{q}_k = \frac{\partial H}{\partial p_k}\,. \tag{14}$$

In the case of the nonlinear wave equation, in order to get a convenient form of the equations, one can choose the Fourier basis. Such a basis is defined by

$$\hat{e}_k := \begin{cases} \frac{1}{\sqrt{\pi}}\cos kx & k > 0 \\ \frac{1}{\sqrt{2\pi}} & k = 0 \\ \frac{1}{\sqrt{\pi}}\sin -kx & k < 0 \end{cases} \tag{15}$$

Thus the Hamiltonian (12) takes the form

$$H(p,q) = \sum_{k\in\mathbb{Z}} \frac{p_k^2 + \omega_k^2 q_k^2}{2} + \int_{-\pi}^{\pi} F\left(\sum_k q_k \hat{e}_k(x)\right)\mathrm{d}x\,, \tag{16}$$

where $\omega_k^2 := k^2 + \mu^2$. For the forthcoming developments it is worth to rescale the variables by defining

$$p'_k := \frac{p_k}{\sqrt{\omega_k}}\,, \quad q'_k := \sqrt{\omega_k}\, q_k\,, \tag{17}$$

so that, omitting primes, the Hamiltonian takes the form

$$H(p,q) = \sum_k \omega_k \frac{p_k^2 + q_k^2}{2} + H_P(p,q) \tag{18}$$

where H_P has a zero of order higher than 2. *In the following we will always study systems of the form* (18). Moreover, by relabeling the variables and the frequencies it is possible to reduce the problem to the case where k varies in $\mathbb{N} \equiv \{1, 2, 3 \ldots\}$. This is what we will assume in developing the abstract theory.

Example 10 An example of a different nature in which the Hamiltonian takes the form (18) is the nonlinear Schrödinger equation

$$-\mathrm{i}\dot{\psi} = \psi_{xx} + f(|\psi|^2)\psi\,, \tag{19}$$

where f is a smooth function. Eq. (19) has the conserved energy functional

$$H(\psi, \bar{\psi}) := \int_{-\pi}^{\pi} \left(|\psi|^2 + F(|\psi|^2)\right)\mathrm{d}x\,, \tag{20}$$

where F is such that $F' = f$. Introduce canonical coordinates (p_k, q_k) by

$$\psi = \sum_{k\in\mathbb{Z}} \frac{p_k + \mathrm{i}q_k}{\sqrt{2}}\hat{e}_k\,, \tag{21}$$

then the energy takes the form (18) with $\omega_k = k^2$ and the NLS is equivalent to the corresponding Hamilton equations.

Example 11 Consider the Kortweg–de Vries equation

$$u_t + u_{xxx} + uu_x = 0\,, \tag{22}$$

in the space of functions with zero mean value. The conserved energy is given by

$$H(u) = \int_{-\pi}^{\pi}\left(\frac{u_x^2}{2} + \frac{u^3}{6}\right)\mathrm{d}x\,, \tag{23}$$

which again is also the Hamiltonian of the system. Canonical coordinates are here introduced by

$$u(x) = \sum_{k>0} \sqrt{k}(p_k \hat{e}_k + q_k \hat{e}_{-k})\,, \tag{24}$$

in which the Hamiltonian takes the form (18) with $\omega_k = k^3$.

Remark 12 It is also interesting to study some of these equations with Dirichlet boundary conditions (DBC), typically on $[0, \pi]$. This will always be done by identifying the space of the functions fulfilling DBC with the space of the function fulfilling periodic boundary conditions on $[-\pi, \pi]$ which are skew symmetric. Similarly, Neumann boundary conditions will be treated by identifying the corresponding functions with periodic even functions. In some cases (e. g. in Eq. (1) with DBC and an f which does not have particular symmetries) the equations do not extend naturally to the space of skew symmetric and this has some interesting consequences (see [7,13]).

Basic Elements of Hamiltonian Formalism for PDEs

A suitable topology in the phase space is given by a Sobolev like topology.

For any $s \in \mathbb{R}$, define the Hilbert space ℓ^2_s of the sequences $x \equiv \{x_k\}_{k\geq 1}$ with $x_k \in \mathbb{R}$ such that

$$\|x\|_s^2 := \sum_k |k|^{2s}|x_k|^2 < \infty \tag{25}$$

and the phase spaces $\mathcal{P}_s := \ell^2_s \oplus \ell^2_s \equiv z \ni (p, q) \equiv (\{p_k\}, \{q_k\})$. In $\mathcal{P}_s$ we will sometimes use the scalar product

$$\langle (p, q), (p^1, q^1)\rangle_s := \langle p, p^1\rangle_{\ell^2_s} + \langle q, q^1\rangle_{\ell^2_s} . \tag{26}$$

In the following we will always assume that

$$|\omega_k| \leq C|k|^d \tag{27}$$

for some d.

Remark 13 Defining the operator $A_0 \colon D(A_0) \to \mathcal{P}_s$ by $A_0(p, q) = (\omega_k p_k, \omega_k q_k)$ one can write $H_0 = \frac{1}{2}\langle A_0 z; z\rangle_0$, $D(A_0) \supset \mathcal{P}_{s+d}$.

Given a smooth Hamiltonian function $\chi \colon \mathcal{P}_s \supset \mathcal{U}_s \to \mathbb{R}$, $\mathcal{U}_s$ being an open neighborhood of the origin, we define the corresponding Hamiltonian vector field $X_\chi \colon \mathcal{U}_s \mapsto \mathcal{P}_{-s}$ by

$$X_\chi \equiv \left(-\frac{\partial \chi}{\partial q_k}, \frac{\partial \chi}{\partial p_k}\right) . \tag{28}$$

Remark 14 Corresponding to a function χ as above we will denote by $\nabla\chi$ its gradient with respect to the $\ell^2 \equiv \ell^2_0$ metric. Defining the operator J by $J(p, q) := (-q, p)$ one has $X_\chi = J\nabla\chi$.

Definition 15 The Poisson Bracket of two smooth functions χ_1, χ_2 is formally defined by

$$\{\chi_1; \chi_2\} := d\chi_1 X_{\chi_2} \equiv \langle \nabla\chi_1; J\nabla\chi_2\rangle_0 . \tag{29}$$

Remark 16 As the example $\chi_1 = \sum_k kq_k$, $\chi_2 := \sum_k kp_k$ shows, there are cases where the Poisson Bracket of two functions is not defined.

For this reason a crucial role is played by the functions whose vector field is smooth.

Definition 17 A function $\chi \in C^\infty(\mathcal{U}_s, \mathcal{P}_s)$, $\mathcal{U}_s \subset \mathcal{P}_s$ open, is said to be of class Gen_s, if the corresponding Hamiltonian vector field X_χ is a smooth map from $\mathcal{U}_s \to \mathcal{P}_s$. In this case we will write $\chi \in \mathrm{Gen}_s$

Proposition 18 *Let $\chi_1 \in Gen_s$. If $\chi_2 \in C^\infty(\mathcal{U}_s, \mathbb{R})$ then $\{\chi_1, \chi_2\} \in C^\infty(\mathcal{U}_s, \mathbb{R})$. If $\chi_2 \in Gen_s$ then $\{\chi_1, \chi_2\} \in Gen_s$.*

Definition 19 A smooth coordinate transformation $\mathcal{T} \colon \mathcal{P}_s \supset \mathcal{U}_s \to \mathcal{P}_s$ is said to be canonical if for any Hamiltonian function H one has $X_{H\circ\mathcal{T}} = \mathcal{T}^* X_H \equiv \mathrm{d}\mathcal{T}^{-1} X_H \circ \mathcal{T}$, i. e. it transforms the Hamilton equations of H into the Hamilton equations of $H \circ \mathcal{T}$.

Proposition 20 *Let $\chi_1 \in Gen_s$, and let $\Phi^t_{\chi_1}$ be the corresponding time t flow (which exists by standard theory). Then $\Phi^t_{\chi_1}$ is a canonical transformation.*

Normal Form for Finite Dimensional Hamiltonian Systems

Consider a system of the form (18), but with finitely many degrees of freedom, namely a system with a Hamiltonian of the form (18) with

$$H_0(p, q) = \sum_{k=1}^n \omega_k \frac{p_k^2 + q_k^2}{2} , \quad \omega_k \in \mathbb{R} \tag{30}$$

and H_P which is a smooth function having a zero of order at least 3 at the origin.

Definition 21 A polynomial Z will be said to be in normal form if $\{H_0; Z\} \equiv 0$.

Theorem 22 (Birkhoff) *For any positive integer $r \geq 0$, there exist a neighborhood $\mathcal{U}^{(r)}$ of the origin and a canonical transformation $\mathcal{T}_r \colon \mathbb{R}^{2n} \supset \mathcal{U}^{(r)} \to \mathbb{R}^{2n}$ which puts the*

system (18) in Birkhoff Normal Form up to order r, namely s.t.

$$H^{(r)} := H \circ \mathcal{T}_r = H_0 + Z^{(r)} + \mathcal{R}^{(r)} \tag{31}$$

where $Z^{(r)}$ is a polynomial of degree $r+2$ which is in normal form, $\mathcal{R}^{(r)}$ is small, i. e.

$$|\mathcal{R}^{(r)}(z)| \le C_r \|z\|^{r+3}\,, \quad \forall z \in \mathcal{U}^{(r)}\,; \tag{32}$$

moreover, one has

$$\|z - \mathcal{T}_r(z)\| \le C'_r \|z\|^2\,, \quad \forall z \in \mathcal{U}^{(r)}\,. \tag{33}$$

An inequality identical to (33) is fulfilled by the inverse transformation $\mathcal{T}_r^{-1}$.

If the frequencies are nonresonant at order $r+2$, namely if

$$\omega \cdot K \neq 0\,, \quad \forall K \in \mathbb{Z}^n\,, \quad 0 < |K| \le r+2 \tag{34}$$

the function $Z^{(r)}$ depends on the actions

$$I_j := \frac{p_j^2 + q_j^2}{2}$$

only.

Remark 23 If the nonlinearity is analytic and the frequencies are Diophantine, i. e. there exist $\gamma > 0$ and τ such that

$$|\omega \cdot K| \ge \frac{\gamma}{|K|^\tau}\,, \quad \forall K \in \mathbb{Z}^n - \{0\}\,, \tag{35}$$

then one can compute the dependence of the constant C_r (cf. Eq. (32)) on r and optimize the value of r as a function of $\|z\|$. This allows one to improve (32) and to show that there exists and r_{opt} such that (see e. g. [40])

$$|\mathcal{R}^{(r_{\text{opt}})}(z)| \le C \exp\left(-\frac{c}{\|z\|^{1/(\tau+1)}}\right)\,. \tag{36}$$

In turn, such an estimate is the starting point for the proof of the celebrated Nekhoroshev's theorem [53].

The idea of the proof is to construct a canonical transformation putting the system in a form which is as simple as possible, namely the normal form. More precisely one constructs a canonical transformation $\mathcal{T}^{(1)}$ pushing the non normalized part of the Hamiltonian to order four followed by a transformation $\mathcal{T}^{(2)}$ pushing it to order five and so on, thus getting $\mathcal{T}_r = \mathcal{T}^{(1)} \circ \mathcal{T}^{(2)} \circ \cdots \circ \mathcal{T}^{(r)}$. Each of the transformations $\mathcal{T}^{(j)}$ is constructed as the time one flow of a suitable auxiliary Hamiltonian function say χ_j (Lie transform method). It turns out that χ_j is determined as the solution of the Homological equation

$$Z_j := \{\chi_j, H_0\} + H^{(j)} \tag{37}$$

where $H^{(j)}$ is constructed recursively and Z_j has to be determined together with χ_j in such a way that $\{Z_j; H_0\} = 0$ and (37) holds. In particular $H^{(1)}$ coincides with the first non vanishing term in the Taylor expansion of H_P.

The algorithm of solution of the Homological Eq. (37) involves division by the so called small denominators $i\omega \cdot K$, where $K \in \mathbb{Z}^n - \{0\}$, fulfills $|K| \le j+2$ and $\omega \cdot K \neq 0$.

The above construction is more or less explicit: provided one has at disposal enough time, he can explicitly compute $Z^{(r)}$ up to any given order. In the case of nonresonant frequencies this is not needed if one wants to understand the dynamics over long times. Indeed its features are an easy consequence of the fact that $Z^{(r)}$ depends on the actions only. A precise statement will be given in the case of PDEs. It has to be noticed that the normal form can be used also as a starting point for the construction of invariant tori through KAM theory. To this end however one has to verify a nondegeneracy condition and this requires the explicit computation of the normal form.

In the resonant case the situation is more complicated, however, it is often enough to compute the first non vanishing term of $Z^{(r)}$ in order to get relevant information on the dynamics. This usually requires only the ability to compute the function Z_1, defined by (37) with $H^{(1)}$ coinciding with the first non vanishing term of the Taylor expansion of H_P. For a detailed analysis we refer to other sections of the Encyclopedia.

A particular case where one can use a coordinate independent formula for the computation of Z_j and χ_j is the one in which the frequencies are completely resonant.

Assume that there exists $\nu > 0$ and integer numbers $\ell_1, .., \ell_n$ such that

$$\omega_k = \nu \ell_k \quad \forall k = 1, \ldots, n\,. \tag{38}$$

Remark 24 Denote by Ψ^t the flow of the linear system with Hamiltonian H_0, then one has

$$\Psi^{t+T} = \Psi^t\,, \quad T := \frac{2\pi}{\nu}\,, \quad t \in \mathbb{R}\,. \tag{39}$$

Moreover in this case one has $\omega \cdot K \neq 0 \Longrightarrow |\omega \cdot K| \ge \nu > 0$, so there are no small denominators.

In this case one has an interesting coordinate independent formula for the solution of the homological Eq. (37).

Lemma 25 *Let f be smooth function, defined in neighborhood of the origin. Define*

$$Z(z) \equiv \langle f \rangle (z) := \frac{1}{T} \int_0^T f(\Psi^t(z)) dt \,,$$
$$\chi(z) := \frac{1}{T} \int_0^T t \left[f(\Psi^t(z)) - Z(\Psi^t(z)) \right] dt \,, \tag{40}$$

then such quantities fulfill the equation $\{H_0, \chi\} + f = Z$.

Normal Form for Hamiltonian PDEs: General Comments

As anticipated in the introduction there are two problems in order to generalize Birkhoff's theorem to PDEs: (1) the existence of nonsmooth vector fields and (2) the appearance of small denominators accumulating at zero already at order 3.

There are two reasons why (1) is a problem. The first one is that if the vector field of χ_1 were not smooth then it would be nontrivial to ensure that it generates a flow, and thus that the normalizing transformation exists. The second related problem is that, if a transformation could be generated, then the Taylor expansion of the transformed Hamiltonian would contain a term of the form $\{H_1; \chi_1\} = \mathrm{d}H_1 X_{\chi_1}$, which is typically not smooth if X_χ is not smooth. Thus one has to show that the construction involves only functions which are of class Gen_s for some s (see Definition 17).

The difficulty related to small denominators is the following: In the finite dimensional case, $\{\omega \cdot K \neq 0, |K| \leq r + 2\}$ implies $|\omega \cdot K| \geq \gamma > 0$. Thus division by $\omega \cdot K$ is a harmless operation in the finite dimensional case. In the infinite dimensional case this is no longer true. For example, when $\omega_k = \sqrt{k^2 + \mu^2}$ one already has

$$\inf_{0 \neq |K| \leq 3} |\omega \cdot K| = 0 \,.$$

In order to solve such a problem one has to take advantage of a property of the nonlinearity which typically holds in PDEs and is called having *localized coefficients*. By also assuming a suitable nonresonance property for the frequency vector, one can deduce a normal form theorem identical to Theorem 22. The main difficulty consists in verifying the assumptions of the theorem. We will show how to verify such assumptions by applying this method to some typical examples.

Normal Form for Resonant Hamiltonian PDEs and its Consequences

In the case of resonant frequencies and smooth vector field it is possible to obtain a normal form up to an exponentially small remainder.

Consider the system (18) in the phase space $\mathcal{P}_s$ with some fixed s. Assume that the frequencies are completely resonant, namely that (38) holds (with $k \in \mathbb{N}$); assume that $H_P \in \mathrm{Gen}_s$ and that its vector field extends to a complex analytic function in a neighborhood of the origin. Finally assume that H_P has a zero of order $n \geq 3$ at the origin. Then we have the following theorem.

Theorem 26 ([10,14]) *There exists a neighborhood of the origin $\mathcal{U}_s \subset \mathcal{P}_s$ and an analytic canonical transformation $\mathcal{T} : \mathcal{U}_s \to \mathcal{P}_s$ with the following properties: $\mathcal{T}$ is close to identity. Namely, it satisfies*

$$\|z - \mathcal{T}(z)\|_s \leq C \|z\|_s^{n-1} \,. \tag{41}$$

$\mathcal{T}$ puts the Hamiltonian in resonant normal form up to an exponentially small remainder, namely one has

$$H \circ \mathcal{T} = H_0 + \langle H_P \rangle + Z_2 + \mathcal{R} \tag{42}$$

where $\langle H_P \rangle$ is the average (defined by (40)) of H_P with respect to the unperturbed flow; Z_2 is in normal form, namely $\{Z_2; H_0\} \equiv 0$, and has a zero of order $2n - 2$ at the origin; $\mathcal{R}$ is an exponentially small remainder whose vector field is estimated by

$$\|X_{\mathcal{R}}(z)\|_s \leq C \|z\|_s^{n-1} \exp\left(-\frac{C}{\|z\|_s^{n-2}} \right) \,.$$

Example 27 The nonlinear Schrödinger equation (19). Here one has $\ell_k = k^2$ and $\nu = 1$. The Sobolev embedding theorems ensure that the vector field of the nonlinearity is analytic if f is analytic in a neighborhood of the origin. Thus Theorem 26 applies to the NLS. To deduce dynamical consequences it is convenient to explicitly compute $\langle H_P \rangle$. Assuming $f(0) = 0$ and $f'(0) = 1$ this was done in [2] using formula (40) which gives

$$\langle H_P \rangle (z) = \frac{1}{2} \left(\sum_k I_k \right)^2 - \frac{1}{8} \sum_k |I_k|^2 \tag{43}$$

where $I_k = (p_k^2 + q_k^2)/2$ are the linear actions. Thus one has that $H_0 + \langle H_P \rangle$ is a function of the actions only, and thus it is an integrable system. It is thus natural to study the system (42) as a perturbation of such an integrable system. This was done in [2] and [25] obtaining the results we are going to state. For simplicity we will concentrate here on

the case of Dirichlet boundary conditions, thus the function ψ will always be assumed to be skew symmetric with respect to the origin. Define

$$\epsilon_s := \left(\frac{1}{2}\int_{-\pi}^{\pi} |\partial_x^s \psi^{(0)}(x)|^2\right)^{1/2}, \tag{44}$$

i. e., the H^s norm of the initial datum $\psi^{(0)}$, $s \geq 0$, and denote by $I_k(0)$ the initial value of the linear actions.

Theorem 28 ([2]) *Fix $N \geq 1$, then there exists a constant ϵ_*, with the property that, if the initial datum $\psi^{(0)}$ is such that*

$$\epsilon_1 < \epsilon_* , \qquad \sum_{k \geq N+1} I_k(0)^2 \leq C\epsilon_1^4 \epsilon_1^{2-1/N} , \tag{45}$$

then along the corresponding solution of (19) one has

$$\sum_{k \geq 1} |I_k(t) - I_k(0)|^2 \leq C' \epsilon_1^{4+1/N} \tag{46}$$

for the times t fulfilling

$$|t| \leq C'' \exp\left(\frac{\epsilon_*}{\epsilon_1}\right)^{1/N} .$$

This result in particular allows one to control the distance in energy norm of the solution from the torus given by the intersection of the level surfaces of the actions taken at the initial time.

Theorem 29 ([25]) *Fix an arbitrarily large r, then there exists s_r such that for any $s \geq s_r$ there exists ϵ_{*s}, such that the following holds true: most of the initial data with $\epsilon_s < \epsilon_{*s}$ give rise to solutions with*

$$\|\psi(t)\|_s \leq C\epsilon_s , \quad \forall |t| \leq \frac{C}{\epsilon_s^r} . \tag{47}$$

For the precise meaning of "most of the initial data," we refer to the original paper. The result is based on the proof that the considered solutions remain close in the H^s topology to an infinite dimensional torus. In particular the uniform estimate of the Sobolev norm is relevant for applications to numerical analysis [30,44].

Example 30 Consider the nonlinear wave Eq. (1) with $\mu = 0$. Here the frequencies are given by $|k|$ and thus they are completely resonant. Again the smoothness of the nonlinearity is ensured by Sobolev embedding theorem. In the case of DBC in order to ensure smoothness one has also to assume that the nonlinearity is odd, namely that $f(-u) = -f(u)$, then Theorem 26 applies. However in this case the computation of $\langle H_P \rangle$ is nontrivial. It has been done (see [55]) in the case of $f(u) = \pm u^3 + O(u^4)$ and Dirichlet boundary conditions. The result however is that the function $\langle H_P \rangle$ does not have a particularly simple structure, and thus it is not easy to extract informations on the dynamics.

In order to extract information on the dynamics, consider the simplified system in which the remainder is neglected, namely the system with Hamiltonian

$$H_S := H_0 + \langle H_P \rangle + Z_2 . \tag{48}$$

Such a system has two integrals of motion, namely H_0 and $\langle H_P \rangle + Z_2$. Let γ_ϵ be the set of the z's at which $\langle H_P \rangle + Z_2$ is restricted to the surface $S_\epsilon := \{z\colon H_0(z) = \epsilon^2\}$ has an extremum, say a maximum. Then γ_ϵ is an invariant set for the dynamics of H_S. By the invariance under the flow of H_0, one has that γ_ϵ is the union of one dimensional closed curves, but generically it is just a single closed curve. In such a situation it is also a stable periodic orbit of (48) (see [35]). Actually it is very difficult to compute $(\langle H_P \rangle + Z_2)|_{S_\epsilon}$, but a maximum of such a function can be easily constructed by applying the implicit function theorem to a non degenerate maximum of $\langle H_P \rangle|_{S_\epsilon}$. The addition of the remainder then modifies the dynamics only after an exponentially long time. We are now going to state the corresponding theorem.

First remark that a critical point of $\langle H_P \rangle|_{S_1}$ is a solution z_a of the system

$$\lambda_a A_0 z_a + \nabla \langle H_P \rangle(z_a) = 0 , \quad H_0(z_a) = 1 \tag{49}$$

where we used the notations of Remarks 13 and 14. Here λ_a is clearly the Lagrange multiplier. The closed curve $\gamma_a := \bigcup_t \Psi^t(z_a)$ is (the trajectory of) a periodic solution of $H_0 + \langle H_P \rangle$. Consider now the linear operator $B_a := \mathrm{d}(\nabla \langle H_P \rangle)(z_a)$.

Definition 31 The critical point z_a is said to be non degenerate if the system

$$\lambda_a A_0 h + B_a h = 0 , \quad \langle A_0 z_a ; h \rangle_0 = 0 \tag{50}$$

has at most one solution.

Under the assumptions of Theorem 32 below it is easy to prove that γ_a is a smooth curve and that its tangent vector $h_a := \frac{\mathrm{d}}{\mathrm{d}t} \Psi^t(z_a)\big|_{t=0}$ is a solution of (50).

Theorem 32 ([14,15]) *Assume that $H_P \in Gen_s$ for any s large enough, assume also that there exists a non degenerate maximum z_a of $\langle H_P \rangle|_{S_1}$, then there exists a constant ϵ_*, such that the following holds true: consider a solution z(t)*

of the Hamilton equation of (18) with initial datum z_0; if there exists $\epsilon < \epsilon_*$, such that

$$d_E(\epsilon\gamma_a, z_0) \leq C\epsilon^n \,, \tag{51}$$

then one has

$$d_E(\epsilon\gamma_a, z(t)) \leq C'\epsilon^n \,, \tag{52}$$

for all times t with $|t| \leq \frac{C}{\epsilon^{r-1}} \exp\left(\frac{\epsilon_*}{\epsilon}\right)^{n-1}$. Here d_E is the distance in the energy norm.

Such a theorem does not ensure that there exist periodic orbits of the complete system, but just a family of closed curves with the property that starting close to it one remains close to it for exponentially long times. Some results concerning the existence of true periodic orbits close to such periodic like trajectories can also be proved (see e. g. [16,19,20,42,51]).

Example 33 In the paper [55] it has been proved that the non degeneracy assumption (50) of Theorem 32 holds for the Eq. (1) with $f(u) = \pm u^3 +$higher order terms and Dirichlet boundary conditions. In the case of such an equation an extremum of $\langle H_P \rangle|_{S_1}$ is given by

$$u(x) = V_m \mathrm{sn}(\mathrm{w}x, \mathrm{m}) \,, \quad v(x) \equiv 0 \,,$$

with V_m, w and m suitable constants, and sn the Jacobi elliptic sine. Therefore the curve γ_a is the phase space trajectory of the solution of the linear wave equation with such an initial datum. There are no other extrema of $\langle H_P \rangle|_{S_1}$. Thus the Theorem 32 ensures that solutions starting close to a rescaling of such a curve remain close to it for very long times.

Normal Form for Nonresonant Hamiltonian PDEs

A Statement

We turn now to the nonresonant case. The theory we will present has been developed in [6,9,43], and is closely related to the one of [4,11,37]. First we introduce the class of equations to which the theory applies. To this end it is useful to treat the p's and the q's exactly on an equal footing so we will denote by $z \equiv (z_k)_{k\in\tilde{\mathbb{Z}}}$, $\tilde{\mathbb{Z}} := \mathbb{Z} - \{0\}$ the set of all the variables, where

$$z_{-k} := p_k \,, \quad z_k := q_k \quad k \geq 1 \,.$$

Given a polynomial function $f\colon \mathcal{P}_\infty \to \mathbb{R}$ of degree r one can decompose it as follows

$$f(z) = \sum_{k_1,\ldots,k_r} f_{k_1,\ldots,k_r} z_{k_1} \ldots z_{k_r} \,. \tag{53}$$

We will assume suitable localization properties for the coefficients $f_{k_1,\ldots,k_r}$ as functions of the indexes $k_1, \ldots, k_r$.

Definition 34 Given a multi-index $k \equiv (k_1, \ldots, k_r)$, let $(k_{i_1}, k_{i_2}, k_{i_3} \ldots, k_{i_r})$ be a reordering of k such that

$$|k_{i_1}| \geq |k_{i_2}| \geq |k_{i_3}| \geq \cdots \geq |k_{i_r}| \,.$$

We define $\mu(k) := |k_{i_3}|$ and

$$S(k) := \mu(k) + ||k_{i_1}| - |k_{i_2}|| \,. \tag{54}$$

Definition 35 Let $f\colon \mathcal{P}_\infty \to \mathbb{R}$ be a polynomial of degree r. We say that f has *localized coefficients* if there exists $\nu \in [0, +\infty)$ such that $\forall N \geq 1$ there exists C_N such that for any choice of the indexes $k_1, \ldots, k_r$, the following inequality holds:

$$\left|f_{k_1,\ldots,k_r}\right| \leq C_N \frac{\mu(k)^{\nu+N}}{S(k)^N} \,. \tag{55}$$

Definition 36 A function $f \in \mathrm{Gen}_s$ for any s large enough, is said to have localized coefficients if all the terms of its Taylor expansion have localized coefficients.

Some important properties of functions with localized coefficients are given by Theorem 37.

Theorem 37 *Let $f\colon \mathcal{P}_\infty \to \mathbb{R}$ be a polynomial of degree r with localized coefficients, then there exists s_0 such that for any $s \geq s_0$ the vector field X_f extends to a smooth map from $\mathcal{P}_s$ to itself; moreover the following estimate holds:*

$$\|X_f(z)\| \leq C\|z\|_s \|z\|_{s_0}^{r-2} \,. \tag{56}$$

In particular it follows that a function with localized coefficients is of class Gen_s for any $s \geq s_0$.

Theorem 38 *The Poisson Bracket of two functions with localized coefficients has localized coefficients.*

In order to develop perturbation theory we also need a quantitative nonresonance condition.

Definition 39 Fix a positive integer r. The frequency vector ω is said to fulfill the *property (r-NR) if there exist* $\gamma > 0$, and $\alpha \in \mathbb{R}$ such that for any N large enough one has

$$\left|\sum_{k\geq 1} \omega_k K_k\right| \geq \frac{\gamma}{N^\alpha} \,, \tag{57}$$

for any $K \in \mathbb{Z}^\infty$, fulfilling $0 \neq |K| := \sum_k |K_k| \leq r+2$, $\sum_{k>N} |K_k| \leq 2$.

It is easy to see that under this condition one can solve the homological equation and that, if the known term of the equation has localized coefficients, then the solution also has localized coefficients.

Theorem 40 ([9,12]) *Fix $r \geq 1$, assume that the frequencies fulfill the nonresonance condition (r-NR); assume that H_P has localized coefficients. Then there exists a finite s_r, a neighborhood $\mathcal{U}_{s_r}^{(r)}$ of the origin in $\mathcal{P}_{s_r}$, and a canonical transformation $\mathcal{T} : \mathcal{U}_{s_r}^{(r)} \to \mathcal{P}_{s_r}$ which puts the system in normal form up to order $r+3$, namely*

$$H^{(r)} := H \circ \mathcal{T} = H_0 + Z^{(r)} + \mathcal{R}^{(r)} \tag{58}$$

where $Z^{(r)}$ has localized coefficients and is a function of the actions I_k only; $\mathcal{R}^{(r)}$ has a small vector field, i. e.

$$\|X_{\mathcal{R}^{(r)}}(z)\|_{s_r} \leq C\|z\|_{s_r}^{r+2}\,, \quad \forall z \in \mathcal{U}_{s_r}^{(r)}\,; \tag{59}$$

thus, one has

$$\|z - \mathcal{T}_r(z)\|_{s_r} \leq C\|z\|_{s_r}^2\,, \quad \forall z \in \mathcal{U}_{s_r}^{(r)}\,. \tag{60}$$

An inequality identical to (60) is fulfilled by the inverse transformation $\mathcal{T}_r^{-1}$. Finally for any $s \geq s_r$ there exists a subset $\mathcal{U}_s^{(r)} \subset \mathcal{U}_{s_r}^{(r)}$ open in $\mathcal{P}_s$ such that the restriction of the canonical transformation to $\mathcal{U}_s^{(r)}$ is analytic also as a map from $\mathcal{P}_s \to \mathcal{P}_s$ and the inequalities (59) and (60) hold with s in place of s_r.

This theorem allows one to give a very precise description of the dynamics.

Proposition 41 *Under the same assumptions of Theorem 37, $\forall s \geq s_r$ there exists ϵ_{*s} such that, if the initial datum fulfills $\epsilon := \|z_0\|_s < \epsilon_{*s}$, then one has*

$$\|z(t)\|_s \leq 4\epsilon\,, \quad \sum_k k^{2s}\,|I_k(t) - I_k(0)| \leq C\epsilon^3 \tag{61}$$

for all the times t fulfilling $|t| \leq \epsilon^{-r}$. Moreover there exists a smooth torus $\mathbb{T}_0$ such that, $\forall M \leq r$

$$d_s(z(t), \mathbb{T}_0) \leq C\epsilon^{(M+3)/2}\,, \quad \textit{for } |t| \leq \frac{1}{\epsilon^{r-M}} \tag{62}$$

where $d_s(.,.)$ is the distance in $\mathcal{P}_s$.

A generalization to the resonant or partially resonant case is easily obtained and can be found in [11].

Verification of the Property of Localization of Coefficients

The property of localization of coefficients is quite abstract. We illustrate via a few examples some ways to verify it.

Example 42 Consider the nonlinear wave Eq. (1) with Neumann boundary conditions on $[0, \pi]$. We recall that the corresponding space of functions will be considered as a subset of the space of periodic functions.

Consider the Taylor expansion of the nonlinearity, i. e. write $F(u) = \sum_{r\geq 3} c_r \int_{-\pi}^{\pi} u^r$. Then one has to prove that the functions $f_r(u) \equiv \int_{-\pi}^{\pi} u^r$ have localized coefficients. The coefficients $f_{k_1,\ldots,k_r}$ are given by

$$f_{k_1,\ldots,k_r} = \int_{-\pi}^{\pi} \cos(k_1 x)\cos(k_2 x)\ldots\cos(k_r x)\mathrm{d}x\,. \tag{63}$$

One could compute and estimate such a quantity directly, but it is easier to proceed in a different way: to show that f_3 has localized coefficients and then to use Theorem 38 to show that each f_r has localized coefficients for any r. This is the path we will follow. Consider

$$f_{k_1,k_2,k_3} = \int_{-\pi}^{\pi} \cos(k_1 x)\cos(k_2 x)\cos(k_3 x)\mathrm{d}x\,. \tag{64}$$

Since the estimate (55) is symmetric with respect to the indexes, we can assume that they are ordered, $k_1 \geq k_2 \geq k_3$, so that (64)$= \pi\delta_{k_1}^{k_2+k_3}/2$, $\mu(k) = k_3$, $S(k) = k_3 + k_1 - k_2$ from which one immediately sees that (55) holds with $\nu = 0$. As a consequence one also gets that the function $g_3 := \int_{-\pi}^{\pi} \nu u^2$ has localized coefficients. Since $\{g_3; f_r\} = r f_{r+1}$, by induction Theorem 38 ensures that f_r has localized coefficients for any r.

Often it is impossible to explicitly compute the coefficients $f_{k_1,\ldots,k_r}$, so one needs a different way to verify the property.

Example 43 Consider the nonlinear wave equation

$$u_{tt} - u_{xx} + Vu = f(u) \tag{65}$$

with Neumann boundary conditions. Here V is a smooth, even, periodic potential. The Hamiltonian reduces to the form (18) by introducing the variables q_k by $u(x) = \sum_k q_k \varphi_k(x)$ where $\varphi_k(x)$ are the eigenfunctions of the Sturm Liouville operator $-\partial_{xx} + V$, and similarly for v. In such a case one has

$$f_{k_1,k_2,k_3} = \int_{-\pi}^{\pi} \varphi_{k_1}(x)\varphi_{k_2}(x)\varphi_{k_3}(x)\mathrm{d}x\,. \tag{66}$$

Here the idea is to consider (66) as the matrix element L_{k_1,k_2} of the operator L of multiplication by $\varphi_{k_3}(x)$ on the basis of the eigenfunctions of the operator $S := -\partial_{xx} + V$. The key idea is to proceed as follows.

Let L be a linear operator which maps $D(S^r)$ into itself for all $r \geq 0$, and define the sequence of operators

$$L_N := [S, L_{N-1}]\,, \quad L_0 := L\,. \tag{67}$$

Lemma 44 *Let S be as above, then, for any $N \geq 0$ one has*

$$\left|L_{k_1,k_2}\right| = \left|\langle L\varphi_{k_1};\varphi_{k_2}\rangle\right| \leq \frac{1}{|\lambda_{k_1}-\lambda_{k_2}|^N}\left|\langle L_N\varphi_{k_1};\varphi_{k_2}\rangle\right| \tag{68}$$

where λ_{k_j} is the eigenvalue of S corresponding to φ_{k_j}.

Then, in order to conclude the verification of the property of localization of the coefficients, one has just to compute L_N and to estimate the scalar product product of the r.h.s. All the computations can be found in [6].

Example 45 A third example where the verification of the property of localization of coefficients goes almost in the same way is the nonlinear Schrödinger equation

$$\mathrm{i}\dot{\psi} = -\psi_{xx} + V\psi + \frac{\partial F(\psi,\bar{\psi})}{\partial\bar{\psi}} \tag{69}$$

with Dirichlet Boundary conditions on $[0,\pi]$. Here one has to assume that V is a smooth even potential and that F is smooth and fulfills $F(-\psi,-\bar{\psi}) = F(\psi,\bar{\psi})$ (this is required in order to leave the space of skew symmetric functions invariant, see Remark 12). Here the variables (p,q) are introduced by

$$\psi = \sum_{k\in\mathbb{Z}} \frac{p_k + \mathrm{i}q_k}{\sqrt{2}}\varphi_k\,, \tag{70}$$

where ψ_k are the eigenfunctions of S with Dirichlet boundary conditions. Here the Taylor expansion of the nonlinearity has only even terms. Thus the building block for the proof of the property of localization of coefficients is the operator L of multiplication by $\varphi_{k_3}\varphi_{k_4}$. Then, mutatis mutandis the proof goes as in the previous case.

Verification of the Nonresonance Property

Finally in order to apply Theorem 40 one has to verify the nonresonance property (r-NR). As usual in dynamical systems, this is done by tuning the frequencies using parameters. In the case of the nonlinear wave Eq. (1) one can use the mass μ.

Theorem 46 ([4,12,36]) *There exists a zero measure set $S \subset \mathbb{R}$ such that, if $\mu \in \mathbb{R} - S$, then the frequencies $\omega_k = \sqrt{k^2+\mu^2}$, $k \geq 1$ fulfill the condition (r-NR) for any r.*

Thus the Theorem 40 applies to the Eq. (1) with almost any mass.

A similar result holds for the Eq. (65), where the role of the mass is played by the average of the potential.

The situation of the nonlinear Schrödinger is more difficult. Here one can use the Fourier coefficients of the potential as parameters.

Fix $\sigma > 0$ and, for any positive R define the space of the potentials, by

$$\mathcal{V}_R := \left\{ V(x) = \sum_{k\geq 1} v_k \cos kx \mid v_k' := v_k R^{-1}\mathrm{e}^{\sigma k} \in \left[-\frac{1}{2},\frac{1}{2}\right] \text{ for } k \geq 1 \right\} \tag{71}$$

Endow such a space with the product normalized probability measure.

Theorem 47 ([12], **see also** [21]) *For any r there exists a positive R and a set $S \subset \mathcal{V}_R$ such that property (r-NR) holds for any potential $V \in S$ and $|\mathcal{V}_R - S| = 0$.*

So, provided the potential is chosen in the considered set, Theorem 40 also applies to the Eq. (69).

We point out that the proof of Theorem 46 and of Theorem 47 essentially consists of two steps. First one proves that for most values of the parameters one has

$$\left|\sum_{k\geq 1}\omega_k K_k\right| \geq \frac{\gamma}{N^\alpha}\,, \qquad \forall K \in \mathbb{Z}^\infty\,, \quad \text{with } 0 \neq |K| := \sum_k |K_k| \leq r+2\,. \tag{72}$$

Then one uses the asymptotic of the frequencies, namely $\omega_k \sim ak^d$ with $d \geq 1$, in order to get the result.

Non Hamiltonian PDEs

In this section we will present some results for the non Hamiltonian case.

It is useful to complexify the phase space. Thus, in this section we will always denote the space of the complex sequences $\{z_k\}$ whose norm (defined by (25)) is finite by $\mathcal{P}_s$.

In the space $\mathcal{P}_s$ consider a system of differential equations of the form

$$\dot{z}_k = \lambda_k z_k + P_k(z)\,, \quad k \in \mathbb{Z} - \{0\} \tag{73}$$

where λ_k are complex numbers and $P(z) \equiv \{P_k(z)\}$ has a zero of order at least 2 at the origin. Moreover we will assume P to be a complex analytic map from a neighborhood of the origin of $\mathcal{P}_s$ to $\mathcal{P}_s$. The quantities λ_k are clearly the eigenvalues of the linear operator describing the linear part of the system, and for this reason they will be called "the eigenvalues".

Example 48 A system of the form (18) with H_P having a vector field which is analytic. The corresponding Hamilton equations have the form (73) with $\lambda_k = -\lambda_{-k} = \mathrm{i}\omega_k$, $k \geq 1$.

Example 49 Consider the following nonlinear heat equation with periodic boundary conditions on $[-\pi, \pi]$:

$$u_t = u_{xx} - V(x)u + f(u) \,. \tag{74}$$

If f is analytic then it can be given the form (73) by introducing the basis of the eigenfunctions ψ_k of the Sturm Liouville operator $S := -\partial_{xx} + V$, i. e. denoting $u = \sum_k z_k \varphi_k$. In this case the the eigenvalues λ_k are the opposite of the periodic eigenvalues of S. Thus in particular one has $\lambda_k \in \mathbb{R}$ and $\lambda_k \sim -k^2$.

In this context one has to introduce a suitable concept of nonresonance:

Definition 50 A sequence of eigenvalues is said to be resonant if there exists a sequence of integer numbers $K_k \geq 0$ and an index i such that

$$\sum_k \lambda_k K_k - \lambda_i = 0 \,. \tag{75}$$

In the finite dimensional case the most celebrated results concerning systems of the form (73) are the Poincaré theorem, the Poincaré–Dulac theorem and the Siegel theorem [1]. The Poincaré theorem is of the form of Birkhoff's Theorem 22, while the Poincaré–Dulac and Siegel theorems guarantee (under suitable assumptions) that there exists an analytic coordinate transformation reducing the system to its normal form or linear part (no remainder!).

At present there is not a satisfactory extension of the Poincaré–Dulac theorem to PDEs (some partial results have been given in [41]). We now state a known extension of the Siegel theorem to PDEs.

Theorem 51 ([54,63]) *Assume that the eigenvalues fulfill the Diophantine type condition*

$$\left| \sum_k \lambda_k K_k - \lambda_i \right| \geq \frac{\gamma}{|K|^\tau} \,, \quad \forall i, K \quad \text{with} \, 2 \leq |K| \,, \tag{76}$$

where $\gamma > 0$ and $\tau \in \mathbb{R}$ are suitable parameters; then there exists an analytic coordinate transformation defined in a neighborhood of the origin, such that the system (73) is transformed into its linear part

$$\dot{z}_k = \lambda_k z_k \,. \tag{77}$$

The main remark concerning this theorem is that the condition (76) is only exceptionally satisfied. If $\lambda \in \mathbb{C}^n$ then condition (76) is generically satisfied only if $\tau > (n-2)/2$. Nevertheless some examples where such an equation is satisfied are known [54].

The formalism of Sect. "Normal Form for Hamiltonian Nonresonant PDEs" can be easily generalized to the non Hamiltonian case giving rise to a generalization of Poincaré's theorem, which we are going to state.

Given a polynomial map $P : \mathcal{P}_\infty \to \mathcal{P}_{-\infty}$ one can expand it on the canonical basis $\mathbf{e}_k$ of $\mathcal{P}_0$ as follows:

$$P(z) = \sum_{k_1,..,k_r,i} P^i_{k_1,..,k_r} z_{k_1} \ldots z_{k_r} \mathbf{e}_i \,, \quad P^i_{k_1,..,k_r} \in \mathbb{C} \,. \tag{78}$$

Definition 52 A polynomial map P is said to have localized coefficients if there exists $\nu \in [0, +\infty)$ such that $\forall N \geq 1$ there exists C_N such that for any choice of the indexes $k_1, \ldots, k_r, i$ following the inequality holds:

$$\left| P^i_{k_1,\ldots,k^i_r} \right| \leq C_N \frac{\mu(k,i)^{\nu+N}}{S(k,i)^N} \,, \tag{79}$$

where $(k, i) = (k_1, .., k_r, i)$. A map is said to have localized coefficients if for any s large enough, it is a smooth map from $\mathcal{P}_s$ to itself and each term of its Taylor expansion has localized coefficients.

Definition 53 The eigenvalues are said to be strongly nonresonant at order r if for any N large enough, any $K = (K_{k_1}, \ldots, K_{k_r})$ and any index i such that $|K| \leq r$ and there are at most two of the indexes $k_1, .., k_r, i$ larger than N the following inequality holds:

$$\left| \sum_k \lambda_k K_k - \lambda_i \right| \geq \frac{\gamma}{N^\alpha} \,. \tag{80}$$

Theorem 54 *Assume that the nonlinearity has localized coefficients and that the eigenvalues are strongly nonresonant at order r, then there exists an s_r and an analytic coordinate transformation $\mathcal{T}_r : \mathcal{U}_{s_r} \to \mathcal{P}_{s_r}$ which transforms the system (73) to the form*

$$\dot{z}_k = \lambda_k z_k + \mathcal{R}_k(z) \,, \tag{81}$$

where the following inequality holds

$$\|\mathcal{R}(z)\|_{s_r} \leq C \|z\|^r_{s_r} \,. \tag{82}$$

Extensions and Related Results

The theory presented here applies to quite general semilinear equations in one space dimension. At present a satisfactory theory applying to quasilinear equations and/or equations in more than one space dimensions is not available. The main difficulty for the extension of the theory to

semilinear equations in higher space dimension is related to the nonresonance condition. The general theory can be easily extended to the case where the differences $\omega_k - \omega_l$ between frequencies accumulate only at a set constituted by isolated points.

Example 55 Consider the nonlinear wave equation on the d-dimensional sphere

$$u_{tt} - \Delta_g u + \mu^2 u = f(x, u)\,, \quad x \in S^d \tag{83}$$

with Δ_g the Laplace Beltrami operator; the frequencies are given by $\omega_k = \sqrt{k(k+d-1)+\mu^2}$ and their differences accumulate only at integers. A version of Theorem 40 applicable to (83) was proved in [9]. As a consequence in particular one has a lower bound on the existence time t of the small amplitude solutions of the form $|t| \geq \epsilon^{-r}$, where ε is proportional to a high Sobolev norm of the initial datum. An extension to $u_{tt} - \Delta_g u + Vu = f(x, u)$ is also known.

Example 56 A similar result was proved in [11] for the nonlinear Schrödinger equation

$$-\mathrm{i}\dot{\psi} = -\Delta\psi + V * \psi + f(|\psi|^2)\psi\,, \quad x \in \mathbb{T}^d\,, \tag{84}$$

where the star denotes convolution.

The only general result available at present for quasilinear system is the following theorem.

Theorem 57 ([5]) *Fix $r \geq 1$, assume that the frequency vector fulfills condition (72) and that there exists d_1 such that the vector field of H_P is a smooth map from $\mathcal{P}_{s+d_1}$ to $\mathcal{P}_s$ for any s large enough. Then the same result of Theorem 40 holds, but the functions do not necessarily have localized coefficients and, for any s large enough* ***the remainder is estimated by***

$$\|X_{\mathcal{R}^{(r)}}(z)\|_s \leq C\|z\|_{s+d_r}^{r+2}\,, \quad \forall z \in \mathcal{U}_s^{(r)}\,; \tag{85}$$

where d_r is a large positive number.

The estimate (85) shows that the remainder is small only when considered as an operator extracting a lot of derivatives. In particular it is non trivial to use such a theorem in order to extract information on the dynamics. Following the approach of [8] and [5] this can be done using the normal form to construct approximate solutions and suitable versions of the Gronwall lemma to compare it with solutions of the true system. This however allows to control the dynamics only over times of the order of ϵ^{-1}, ε being again a measure of the size of the initial datum. Such a theory has been applied to quasilinear wave equations in [5] and to the Fermi Pasta Ulam problem in [17]. Among the large number of papers containing related results we recall [41,47,56,59]. A stronger result for the nonlinear wave equation valid over times of order ϵ^{-2} can be found in [36].

Future Directions

Future directions of research include both purely theoretical aspects and applications to other sciences.

From a purely theoretical point of view, the most important open problems pertain to the validity of normal form theory for equations in which the nonlinearity involves derivatives, and for general equations in more than one space dimension.

These results would be particularly important since the kind of equations appearing in most domains of physics are quasilinear and higher dimensional.

Concerning applications, we would like to describe a few of them which would be of interest.

- Water wave problem. The problem of description of the free surface of a fluid has been shown to fit in the scheme of Hamiltonian dynamics [62]. Normal form theory could allow one to extract the relevant informations on the dynamics in different situations [31,38], ranging from the theory of Tsunamis [32] to the theory of fluid interface, which is relevant e. g. to the construction of oil platforms [33].
- Quantum mechanics. A Bose condensate is known to be well described by the Gross Pitaevskii equation. When the potential is confining, such an equation is of the form (18). Normal form theory has already been used for some preliminary results [18], but a systematic investigation could lead to interesting new results.
- Electromagnetic theory and magnetohydrodynamics. The equations have a Hamiltonian form; normal form theory could help to describe some instability arising in plasmas.
- Elastodynamics. Here one of the main theoretical open problems is the stability of equilibria which are a minimum of the energy. The problem is that in higher dimensions the conservation of energy does not ensure enough smoothness of the solution to ensure stability. Such a problem is of the same kind as the one solved in [9] when dealing with the existence times of the nonlinear wave equation.

Bibliography

1. Arnold V (1984) Chapitres supplémentaires de la théorie des équations différentielles ordinaires. Mir, Moscow
2. Bambusi D (1999) Nekhoroshev theorem for small amplitude solutions in nonlinear Schrödinger equation. Math Z 130:345–387

3. Bambusi D (1999) On the Darboux theorem for weak symplectic manifolds. Proc Amer Math Soc 127(11):3383–3391
4. Bambusi D (2003) Birkhoff normal form for some nonlinear PDEs. Comm Math Phys 234:253–283
5. Bambusi D (2005) Galerkin averaging method and Poincaré normal form for some quasilinear PDEs. Ann Sc Norm Super Pisa Cl Sci 4(5):669–702
6. Bambusi D (2008) A Birkhoff normal form theorem for some semilinear PDEs. In: Craig W (ed) Hamiltonian dynamical systems and applications. Springer
7. Bambusi D, Carati A, Penati T (2007) On the relevance of boundary conditions for the FPU paradox. Preprint Instit Lombardo Accad Sci Lett Rend A (to appear)
8. Bambusi D, Carati A, Ponno A (2002) The nonlinear Schrödinger equation as a resonant normal form. DCDS-B 2:109–128
9. Bambusi D, Delort JM, Grébert B, Szeftel J (2007) Almost global existence for Hamiltonian semi-linear Klein–Gordon equations with small Cauchy data on Zoll manifolds. Comm Pure Appl Math 60(11):1665–1690
10. Bambusi D, Giorgilli A (1993) Exponential stability of states close to resonance in infinite-dimensional Hamiltonian systems. J Statist Phys 71(3–4):569–606
11. Bambusi D, Grebert B (2003) Forme normale pour NLS en dimension quelconque. Compt Rendu Acad Sci Paris 337:409–414
12. Bambusi D, Grébert B (2006) Birkhoff normal form for partial differential equations with tame modulus. Duke Math J 135(3):507–567
13. Bambusi D, Muraro D, Penati T (2008) Numerical studies on boundary effects on the FPU paradox. Phys Lett A 372(12):2039–2042
14. Bambusi D, Nekhoroshev NN (1998) A property of exponential stability in the nonlinear wave equation close to main linear mode. Phys D 122:73–104
15. Bambusi D, Nekhoroshev NN (2002) Long time stability in perturbations of completely resonant PDE's, Symmetry and perturbation theory. Acta Appl Math 70(1–3):1–22
16. Bambusi D, Paleari S (2001) Families of periodic orbits for resonant PDE's. J Nonlinear Sci 11:69–87
17. Bambusi D, Ponno A (2006) On metastability in FPU. Comm Math Phys 264(2):539–561
18. Bambusi D, Sacchetti A (2007) Exponential times in the one-dimensional Gross–Petaevskii equation with multiple well potential. Commun Math Phys 234(2):136
19. Berti M, Bolle P (2003) Periodic solutions of nonlinear wave equations with general nonlinearities. Commun Math Phys 243:315–328
20. Berti M, Bolle P (2006) Cantor families of periodic solutions for completely resonant nonlinear wave equations. Duke Math J 134(2):359–419
21. Bourgain J (1996) Construction of approximative and almost-periodic solutions of perturbed linear Schrödinger and wave equations. Geom Funct Anal 6:201–230
22. Bourgain J (1996) On the growth in time of higher Sobolev norms of smooth solutions of Hamiltonian PDE. Int Math Res Not 6:277–304
23. Bourgain J (1997) On growth in time of Sobolev norms of smooth solutions of nonlinear Schrödinger equations in R^D. J Anal Math 72:299–310
24. Bourgain J (1998) Quasi-periodic solutions of Hamiltonian perturbations of 2D linear Schrödinger equation. Ann Math 148:363–439
25. Bourgain J (2000) On diffusion in high-dimensional Hamiltonian systems and PDE. J Anal Math 80:1–35
26. Bourgain J (2005) Green's function estimates for lattice Schrödinger operators and applications. In: Annals of Mathematics Studies, vol 158. Princeton University Press, Princeton
27. Bourgain J (2005) On invariant tori of full dimension for 1D periodic NLS. J Funct Anal 229(1):62–94
28. Bourgain J, Kaloshin V (2005) On diffusion in high-dimensional Hamiltonian systems. J Funct Anal 229(1):1–61
29. Chernoff PR, Marsden JE (1974) Properties of infinite dimensional Hamiltonian systems In: Lecture Notes in Mathematics, vol 425. Springer, Berlin
30. Cohen D, Hairer E, Lubich C (2008) Long-time analysis of nonlinearly perturbed wave equations via modulated Fourier expansions. Arch Ration Mech Anal 187(2)341–368
31. Craig W (1996) Birkhoff normal forms for water waves In: Mathematical problems in the theory of water waves (Luminy, 1995), Contemp Math, vol 200. Amer Math Soc, Providence, RI, pp 57–74
32. Craig W (2006) Surface water waves and tsunamis. J Dyn Differ Equ 18(3):525–549
33. Craig W, Guyenne P, Kalisch H (2005) Hamiltonian long-wave expansions for free surfaces and interfaces. Comm Pure Appl Math 58(12):1587–1641
34. Craig W, Wayne CE (1993) Newton's method and periodic solutions of nonlinear wave equations. Comm Pure Appl Math 46:1409–1498
35. Dell'Antonio GF (1989) Fine tuning of resonances and periodic solutions of Hamiltonian systems near equilibrium. Comm Math Phys 120(4):529–546
36. Delort JM, Szeftel J (2004) Long-time existence for small data nonlinear Klein–Gordon equations on tori and spheres. Int Math Res Not 37:1897–1966
37. Delort J-M, Szeftel J (2006) Long-time existence for semi-linear Klein–Gordon equations with small Cauchy data on Zoll manifolds. Amer J Math 128(5):1187–1218
38. Dyachenko AI, Zakharov VE (1994) Is free-surface hydrodynamics an integrable system? Phys Lett A 190:144–148
39. Eliasson HL, Kuksin SB (2006) KAM for non-linear Schroedinger equation. Ann of Math. Preprint (to appear)
40. Fassò F (1990) Lie series method for vector fields and Hamiltonian perturbation theory. Z Angew Math Phys 41(6):843–864
41. Foias C, Saut JC (1987) Linearization and normal form of the Navier–Stokes equations with potential forces. Ann Inst H Poincaré Anal Non Linéaire 4:1–47
42. Gentile G, Mastropietro V, Procesi M (2005) Periodic solutions for completely resonant nonlinear wave equations with Dirichlet boundary conditions. Comm Math Phys 256(2):437–490
43. Grébert B (2007) Birkhoff normal form and Hamiltonian PDES. Partial differential equations and applications, 1–46 Sémin Congr, 15 Soc Math France, Paris
44. Hairer E, Lubich C (2006) Conservation of energy, momentum and actions in numerical discretizations of nonlinear wave equations
45. Kappeler T, Pöschel J (2003) KdV & KAM. Springer, Berlin
46. Klainerman S (1983) On almost global solutions to quasilinear

wave equations in three space dimensions. Comm Pure Appl Math 36:325–344
47. Krol MS (1989) On Galerkin–averaging method for weakly nonlinear wave equations. Math Meth Appl Sci 11:649–664
48. Kuksin SB (1987) Hamiltonian perturbations of infinite-dimensional linear systems with an imaginary spectrum. Funct Anal Appl 21:192–205
49. Kuksin SB (1993) Nearly integrable infinite-dimensional Hamiltonian systems. Springer, Berlin
50. Kuksin SB, Pöschel J (1996) Invariant Cantor manifolds of quasi-periodic oscillations for a nonlinear Schrödinger equation. Ann Math 143:149–179
51. Lidskij BV, Shulman EI (1988) Periodic solutions of the equation $u_{tt} - u_{xx} + u^3 = 0$. Funct Anal Appl 22:332–333
52. Marsden J (1972) Darboux's theorem fails for weak symplectic forms. Proc Amer Math Soc 32:590–592
53. Nekhoroshev NN (1977) Exponential estimate of the stability of near integrable Hamiltonian systems. Russ Math Surv 32(6): 1–65
54. Nikolenko NV (1986) The method of Poincaré normal form in problems of integrability of equations of evolution type. Russ Math Surv 41:63–114
55. Paleari S, Bambusi D, Cacciatori S (2001) Normal form and exponential stability for some nonlinear string equations. ZAMP 52:1033–1052
56. Pals H (1996) The Galerkin–averaging method for the Klein–Gordon equation in two space dimensions. Nonlinear Anal TMA 27:841–856
57. Pöschel J (2002) On the construction of almost-periodic solutions for a nonlinear Schrödinger equation. Ergod Th Dyn Syst 22:1–22
58. Soffer A, Weinstein MI (1999) Resonances, radiation damping and instability in Hamiltonian nonlinear wave equations. Invent Math 136(1):9–74
59. Stroucken ACJ, Verhulst F (1987) The Galerkin–averaging method for nonlinear, undamped continuous systems. Math Meth Appl Sci 335:520–549
60. Wayne CE (1990) Periodic and quasi-periodic solutions of nonlinear wave equations via KAM theory. Comm Math Phys 127:479–528
61. Weinstein A (1969) Symplectic structures on Banach manifolds. Bull Amer Math Soc 75:1040–1041
62. Zakharov VE (1968) Stability of periodic waves of finite amplitude on the surface of a deep fluid. Appl Mech Tech Phys 2:190–194
63. Zehnder E (1978) C L Siegel's linearization theorem in infinite dimensions. Manuscripta Math 23:363–371

Perturbation Theory in Quantum Mechanics

Luigi E. Picasso[1,2], Luciano Bracci[1,2], Emilio d'Emilio[1,2]
[1] Dipartimento di Fisica, Università di Pisa, Pisa, Italy
[2] Istituto Nazionale di Fisica Nucleare, Sezione di Pisa, Pisa, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Presentation of the Problem and an Example
Perturbation of Point Spectra: Nondegenerate Case
Perturbation of Point Spectra: Degenerate Case
The Brillouin–Wigner Method
Symmetry and Degeneracy
Problems with the Perturbation Series
Perturbation of the Continuous Spectrum
Time Dependent Perturbations
Future Directions
Bibliography

Glossary

Hilbert space A Hilbert space $\mathcal{H}$ is a normed complex vector space with a Hermitian scalar product. If $\varphi, \psi \in \mathcal{H}$ the scalar product between φ and ψ is written as $(\varphi, \psi) \equiv (\psi, \varphi)^*$ and is taken to be linear in ψ and antilinear in φ: if $a, b \in \mathbb{C}$, the scalar product between $a\varphi$ and $b\psi$ is $a^* b(\varphi, \psi)$. The norm of ψ is defined as $\|\psi\| \equiv \sqrt{(\psi, \psi)}$. With respect to the norm $\|\cdot\|$, $\mathcal{H}$ is a complete metric space. In the following $\mathcal{H}$ will be assumed to be separable, that is any complete orthonormal set of vectors is countable.

States and observables In quantum mechanics the states of a system are represented as vectors in a Hilbert space $\mathcal{H}$, with the convention that proportional vectors represent the same state. Physicists mostly use Dirac's notation: the elements of $\mathcal{H}$ are represented by $|\cdot\rangle$ ("ket") and the scalar product between $|\varphi\rangle$ and $|\psi\rangle$ is written as $\langle\varphi \mid \psi\rangle$ ("braket"). The observables, i. e. the physical quantities that can be measured, are represented by linear Hermitian (more precisely: self-adjoint) operators on $\mathcal{H}$. The eigenvalues of an observable are the only possible results of the measurement of the observable. The observables of a system are generally the same of the corresponding classical system: energy, angular momentum, etc., i. e. they are of the form $f(q, p)$, with $q \equiv (q_1, \ldots, q_n)$, $p \equiv (p_1, \ldots, p_n)$ the position and momentum canonical variables of the system: q_i and p_i are observables, i. e. operators, which satisfy the commutation relations $[q_i, q_j] \equiv q_i q_j - q_j q_i = 0$, $[p_i, p_j] = 0$, $[q_i, p_j] = i\hbar\,\delta_{ij}$, with $\hbar$ the Planck's constant h divided by 2π.

Representations Since separable Hilbert spaces are isomorphic, it is always possible to represent the elements of $\mathcal{H}$ as elements of l_2, the space of the sequences $\{u_i\}$, $u_i \in \mathbb{C}$, with the scalar product

$(v, u) \equiv \sum_i v_i^* u_i$. This can be done by choosing an orthonormal basis of vectors e_i in $\mathcal{H} : (e_i, e_j) = \delta_{ij}$ and defining $u_i = (e_i, u)$; with Dirac's notations $|A\rangle \to \{a_i\}$, $a_i = \langle e_i \mid A\rangle$. Linear operators ξ are then represented by $\{\xi_{ij}\}, \xi_{ij} = (e_i, \xi\, e_j) \equiv \langle e_i \mid \xi \mid e_j\rangle$. The ξ_{ij} are called "matrix elements" of ξ in the representation e_i. If $\xi^\dagger$ is the Hermitian-conjugate of ξ, then $(\xi^\dagger)_{ij} = \xi_{ji}^*$. If the e_i are eigenvectors of ξ then the (infinite) matrix ξ_{ij} is diagonal, the diagonal elements being the eigenvalues of ξ.

Schrödinger representation A different possibility is to represent the elements of $\mathcal{H}$ as elements of $L^2[\mathbb{R}^n]$, the space of the square-integrable functions on $\mathbb{R}^n$, where n is the number of degrees of freedom of the system. This can be done by assigning how the operators q_i and p_i act on the functions of $L^2[\mathbb{R}^n]$: in the Schrödinger representation the q_i are taken to act as multiplication by x_i and the p_i as $-i\hbar\partial/\partial x_i$: if $|A\rangle \to \psi_A(x_1, \ldots, x_n)$, then

$$q_i|A\rangle \to x_i \psi_A(x_1, \ldots, x_n)\,,$$
$$p_i|A\rangle \to -i\hbar\partial\psi_A(x_1, \ldots, x_n)/\partial x_i\,.$$

Schrödinger equation Among the observables, the Hamiltonian H plays a special role. It determines the time evolution of the system through the time dependent Schrödinger equation

$$i\hbar\frac{\partial\psi}{\partial t} = H\psi\,,$$

and its eigenvalues are the energy levels of the system. The eigenvalue equation $H\psi = E\psi$ is called the Schrödinger equation.

Definition of the Subject

In the investigation of natural phenomena a crucial role is played by the comparison between theoretical predictions and experimental data. Those practicing the two arts of the trade continuously put challenges to one another either presenting data which ask for an explanation or proposing new experimental verifications of a theory. Celestial mechanics offers the first historical instance of this interplay: the elliptical planetary orbits discovered by Kepler were explained by Newton; when discrepancies from the elliptical paths definitely emerged it was necessary to add the effects of the heavier planets to the dominant role of the sun, until persistent discrepancies between theory and experiment asked for the drastic revision of the theory of gravitation put forth by Einstein, a revision which in turn offered a lot of new effects to observe, some of which have been verified only recently.

In this dialectic interaction between theory and experiment only the simplest problem, that of a planet moving in the field of the sun within Newton's theory, can be solved exactly. All the rest was calculated by means of perturbation theory. Generally speaking, perturbation theory is the technique of finding an approximate solution to a problem where to a dominant factor, which allows for an exact solution (zeroth order solution), other "perturbing" factors are added which are outweighed by the dominant factor and are expected to bring small corrections to the zeroth order solution.

Perturbation theory is ever-pervasive in physics, but an area where it plays a major role is quantum mechanics. In the early days of this discipline, the interpretation of atomic spectra was made possible only by a heavy use of perturbation theory, since the only exactly soluble problem was that of the hydrogen atom without external fields. The explanation of the Stark spectra (hydrogen in a constant electric field) and of the Zeeman spectra (atom in a magnetic field) was only possible when a perturbation theory tailored to the Schrödinger equation, which rules the atomic world, was devised. As for heavier atoms, in no case an exact solution for the Schrödinger equation is available: they could only be treated as a perturbation of simpler "hydrogenoid" atoms. Most of the essential aspects of atomic and molecular physics could be explained quantitatively in a few years by recourse to suitable forms of perturbation theory. Not only did it explain the position of the spectral lines, but also their relative intensities, and the absence of some lines which showed the impossibility of the corresponding transitions (selection rules) found a convincing explanation when symmetry considerations were introduced. When later more accurate measurements revealed details in the hydrogen spectrum (the Lamb shift) that only quantum field theory was able to explain, perturbation theory gained a new impetus which sometimes resulted in the anticipation of theory (quantum electrodynamics) over experiment as to the accuracy of the effect to be measured.

An attempt to describe all the forms that perturbation theory assumes in the various fields of physics would be vain. We will limit to illustrate its role and its methods in quantum mechanics, which is perhaps the field where it has reached its most mature development and finds its widest applications.

Introduction

An early example of the use of perturbation theory which clearly illustrates its main ideas is offered by the study of

the free fall of a body [29]. The equation of motion is

$$\dot{\vec{v}} = \vec{g} + 2\vec{v} \times \vec{\Omega} + \vec{\Omega} \times (\vec{r} \times \vec{\Omega}) \tag{1}$$

where $\vec{g}$ is the constant gravity acceleration and $\vec{\Omega}$ the angular velocity of the rotation of the earth about its axis. Ω is the parameter characterizing the perturbation. If we wish to find the eastward deviation of the trajectory to first order in Ω we can neglect the third term in the RHS of Eq. (1), whose main effect is to cause a southward deviation (in the northern hemisphere). The ratio of the second term to the first one in the RHS of Eq. (1) (the effective perturbation parameter) is $\Omega\sqrt{h/g} \simeq 10^{-4}$ for the fall from a height $h \sim 100$ m, so we can find the effect of Ω by writing $\vec{v} = \vec{v}_0 + \vec{v}_1$ in Eq. (1), where $\vec{v}_0$ is the zeroth order solution ($\vec{v}_0 = \vec{g}t$ if $\vec{v}_0(0) = 0$) and $\vec{v}_1$ obeys

$$\dot{\vec{v}}_1 = 2\vec{v}_0 \times \vec{\Omega} = 2t\vec{g} \times \vec{\Omega} \,. \tag{2}$$

The solution is $\vec{r} = \vec{h} + \frac{1}{2}\vec{g}t^2 + \frac{1}{3}t^3\vec{g} \times \vec{\Omega}$. The eastward deviation is the deviation in the direction of $\vec{g} \times \vec{\Omega}$ and its value is $\delta = \frac{1}{3}\bar{t}^{\,3} g\Omega \cos\theta$, where θ is the latitude and $\bar{t}$ the zeroth order time of fall, $\bar{t} = \sqrt{2gh}$.

While the above example is a nice illustration of the main features of perturbation theory (identification of a perturbation parameter whose powers classify the contributions to the solution, existence of a zeroth order exact solution) the beginning of modern perturbation theory can be traced back to the work of Rayleigh on the theory of sound [33]. In essence, he wondered how the normal modes of a vibrating string

$$\rho(x)\frac{\partial^2 v}{\partial t^2} = \frac{\partial^2 v}{\partial x^2} \,, \quad v(0,t) = v(\pi,t) = 0 \tag{3}$$

are modified when passing from a constant density $\rho = 1$ to a perturbed density $\rho + \epsilon\sigma(x)$. To solve this problem he wrote down most of the formulae [10,33] which are still in use to calculate the first order correction to non-degenerate energy levels in quantum mechanics.

The equation for the normal modes is

$$u''(x) + \lambda\rho(x)u(x) = 0 \,, \quad u(0) = u(\pi) = 0 \,. \tag{4}$$

Let $u_n^{(0)} \equiv \sqrt{2/\pi}\sin nx$ be the unperturbed solution for the nth mode, $\lambda_n = n^2$, and $u_n^{(0)} + \epsilon u_n^{(1)}$ the perturbed solution through first order, corresponding to a frequency $\lambda_n + \epsilon\mu_n$. By writing the equation for $u_n^{(1)}$

$$\frac{\mathrm{d}^2 u_n^{(1)}}{\mathrm{d}x^2} + \lambda_n u_n^{(1)} + \mu_n u_n^{(0)} + \lambda_n \sigma u_n^{(0)} = 0 \,, \qquad u_n^{(1)}(0) = u_n^{(1)}(\pi) = 0 \tag{5}$$

after multiplying by $u_r^{(0)}$ and using Green's theorem he found

$$\mu_n = -\lambda_n \int_0^\pi \sigma(x)\left(u_n^{(0)}\right)^2 \mathrm{d}x \,, \tag{6}$$

$$a_{rn} \equiv \int_0^\pi u_r^{(0)} u_n^{(1)} \mathrm{d}x = \frac{\lambda_n}{\lambda_r - \lambda_n}\int_0^\pi \sigma u_r^{(0)} u_n^{(0)} \mathrm{d}x \qquad (r \neq n) \tag{7}$$

$$\int_0^\pi u_n^{(0)} u_n^{(1)} \mathrm{d}x = 0 \,. \tag{8}$$

As an application Rayleigh found the position $\pi/2 + \delta x \equiv \pi/2 + \epsilon\tau$ of the nodal point of the perturbed mode $n = 2$ when the perturbation to the density is $\sigma = \kappa\delta(x - \pi/4)$. The vanishing of $u_2^{(0)} + \epsilon u_2^{(1)}$ determines $2\sqrt{2/\pi}\,\tau = u_2^{(1)}(\pi/2)$. By Eq. (7) the function $u_2^{(1)}$ has an expansion $u_2^{(1)} = \sum_{n\neq 2} a_{n2} u_n^{(0)}$, $a_{n2} = \frac{4\kappa}{n^2-4}\sin n\pi/4$. The result for τ is

$$\tau = -\frac{2\kappa}{\pi\sqrt{2}}\left(1 + \frac{1}{3} - \frac{1}{5} - \frac{1}{7} + \frac{1}{9} + \frac{1}{11} - \cdots\right) = -\frac{\kappa}{2} \,.$$

(The series in brackets is equal to $\int_0^1 (1+x^2)/(1+x^4)\mathrm{d}x = 1/2\int_0^\infty (1+x^2)/(1+x^4)\mathrm{d}x$, which can be calculated by contour integration.)

Perturbation theory was revived by Schrödinger, who introduced it into quantum mechanics in a pioneering work of 1926 [44]. There, he applied the concepts and methods which Rayleigh had put forth to the case where the zeroth order problem was a partial differential equation with non-constant coefficients, and he wrote down, in the language of wave mechanics, all the relevant formulae which yield the correction to the energy levels and to the wave functions for the case of both non-degenerate and degenerate energy levels. As an application he calculated the shift of the energy levels of the hydrogen atom in a constant electric field by two different methods. First he observed that in parabolic coordinates the wave equation is separable also with a constant electric field, which implies that in the subspace of the states with equal zeroth order energy the perturbation is diagonal in the basis of the parabolic eigenfunctions, thus circumventing the intricacies of the degenerate case. Later, he used the spherical coordinates, which entails a non diagonal perturbation matrix and calls for the full machinery of the perturbation theory for degenerate eigenvalues.

It is of no use to repeat here Schrödinger's calculations, since the methods which they use are at the core of modern perturbation theory, which is referred to as the

Rayleigh–Schrödinger (RS) perturbation theory. It rapidly superseded other approaches (as that by Born, Heisenberg and Jordan [7], who worked in the framework of the matrix quantum mechanics), and will be presented in the following sections.

Presentation of the Problem and an Example

The most frequent application of perturbation theory in quantum mechanics is the approximate calculation of point spectra. The Hamiltonian H is split into an exactly solvable part H_0 (the unperturbed Hamiltonian) plus a term V (the perturbation) which, in a sense to be specified later, is small with respect to H_0: $H = H_0 + V$. In many cases the perturbation contains an adjustable parameter which depends on the actual physical setting. For example, for a system in an external field this parameter is the field strength. For weak fields one expects the spectrum of H to differ only slightly from the spectrum of H_0. In these cases it is convenient to single out the dependence on a parameter by setting

$$H(\lambda) \equiv H_0 + \lambda V \,. \tag{9}$$

Accordingly we will write the Schrödinger equation as

$$H(\lambda)\psi(\lambda) = E(\lambda)\psi(\lambda) \,. \tag{10}$$

We will retain the form Eq. (9) of the Hamiltonian even when H does not contain a variable parameter, thereby understanding that the actual eigenvalues and eigenvectors are the values at $\lambda = 1$.

The basic idea of the RS perturbation theory is that the eigenvalues and eigenvectors of H can be represented as power series

$$\psi(\lambda) = \sum_0^\infty \lambda^n \psi^{(n)} \,, \quad E(\lambda) = \sum_0^\infty \lambda^n \epsilon_n \,, \tag{11}$$

whose coefficients are determined by substituting expansions Eq. (11) into Eq. (10) and equating terms of equal order in λ. Generally, only the first few terms of the series can be explicitly computed, and the primary task of the RS perturbation theory is their calculation. The practicing scientist who uses perturbation theory never has to tackle the mathematical problem of the convergence of the series. This problem, however, or more generally the connection between the truncated perturbation sums and the actual values of the energy and the wave function, is fundamental for the consistency of perturbation theory and will be touched upon in a later section.

Before expounding the technique of the RS perturbation theory we will consider a simple (two-dimensional) problem which can be solved exactly, since in its discussion several features of perturbation theory will emerge clearly, concerning both the behavior of the energy $E(\lambda)$ and the behavior of the Taylor expansion of this function. From the physical point of view a system with two-dimensional Hilbert space $\mathbb{C}^2$ can be thought of as a particle with spin 1/2 when the translational degrees of freedom are ignored.

Let us write the Hamiltonian $H = H_0 + \lambda V$ in a representation where H_0 is diagonal:

$$\begin{aligned} H &= \begin{pmatrix} E_1^0 & 0 \\ 0 & E_2^0 \end{pmatrix} + \lambda \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \\ &= \begin{pmatrix} E_1^0 + \lambda V_{11} & \lambda V_{12} \\ \lambda V_{12}^* & E_2^0 + \lambda V_{22} \end{pmatrix} . \end{aligned} \tag{12}$$

We consider first the case $E_1^0 \neq E_2^0$, $V_{12} \neq 0$. The exact eigenvalues $E_{1,2}(\lambda)$ of H are found by solving the secular equation:

$$\begin{aligned} &E_{1,2}(\lambda) \\ &\quad = \frac{1}{2}\left[\left(E_1^0 + \lambda V_{11}\right) + \left(E_2^0 + \lambda V_{22}\right) \pm \sqrt{\Delta(\lambda)} \right] , \end{aligned} \tag{13}$$

$$\Delta(\lambda) \equiv \left(\left(E_1^0 + \lambda V_{11}\right) - \left(E_2^0 + \lambda V_{22}\right)\right)^2 + 4\lambda^2 |V_{12}|^2 . \tag{14}$$

The corresponding eigenvectors, in the so called intermediate normalization defined by $(\psi(0), \psi(\lambda)) = 1$, are

$$\psi_1(\lambda) = \left(1, \frac{\sqrt{\Delta(\lambda)} - \left(E_1^0 - E_2^0\right) - \lambda\left(V_{11} - V_{22}\right)}{2\lambda V_{12}}\right) \tag{15}$$

$$\psi_2(\lambda) = \left(-\frac{\sqrt{\Delta(\lambda)} - \left(E_1^0 - E_2^0\right) - \lambda\left(V_{11} - V_{22}\right)}{2\lambda V_{21}}, 1\right) . \tag{16}$$

Expanding $E_{1,2}(\lambda)$ through order λ^3 we get:

$$\begin{aligned} E_1(\lambda) = E_1^0 + \lambda V_{11} + \lambda^2 \frac{|V_{12}|^2}{E_1^0 - E_2^0} \\ - \lambda^3 \frac{|V_{12}|^2 (V_{11} - V_{22})}{\left(E_1^0 - E_2^0\right)^2} + O(\lambda^4) \end{aligned} \tag{17}$$

$$\begin{aligned} E_2(\lambda) = E_2^0 + \lambda V_{22} - \lambda^2 \frac{|V_{12}|^2}{E_1^0 - E_2^0} \\ + \lambda^3 \frac{|V_{12}|^2 (V_{11} - V_{22})}{\left(E_1^0 - E_2^0\right)^2} + O(\lambda^4) \,. \end{aligned} \tag{18}$$

At order 1 only the diagonal matrix elements of V contribute to $E_{1,2}$. The validity of the approximation requires $\lambda|V_{12}| \ll |E_1^0 - E_2^0|$. If this condition is not satisfied, that is if the eigenvalues E_1^0, E_2^0 are "quasi-degenerate", all terms of the expansion can be numerically of the

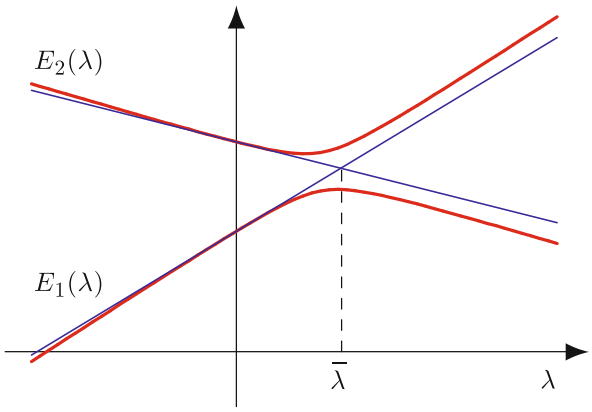

Perturbation Theory in Quantum Mechanics, Figure 1
The behavior of the exact eigenvalues $E_{1,2}(\lambda)$ when $V_{12} = 0$ (*blue lines*) and when $V_{12} \neq 0$ (*red lines*)

same order of magnitude and no approximation of finite order makes sense.

Note that, within the first order approximation, "level crossing" ($E_1(\lambda) = E_2(\lambda)$) occurs at

$$\bar{\lambda} = -\left(E_1^0 - E_2^0\right) / \left(V_{11} - V_{22}\right) . \tag{19}$$

On the other hand Eq. (13) shows that level-crossing is impossible, unless $V_{12} = 0$, in which case the first order approximation yields the exact result. If $V_{12} \neq 0$ the behavior of the levels $E_1(\lambda)$ and $E_2(\lambda)$ near $\bar{\lambda}$ is shown in Fig. 1: the two levels "repel" each other [49].

At first order the eigenvectors $\psi_{1,2}(\lambda)$ are

$$\psi_1^{[1]} = \left(1,\ -\lambda V_{21} / \left(E_2^0 - E_1^0\right)\right) \tag{20}$$

$$\psi_2^{[1]} = \left(-\lambda V_{12} / \left(E_1^0 - E_2^0\right), 1\right) . \tag{21}$$

The expectation value $(\psi_1^{[1]}, H\psi_1^{[1]})/(\psi_1^{[1]}, \psi_1^{[1]})$ of the Hamiltonian over $\psi_1^{[1]}$, for example, is

$$E_1^0 + \lambda V_{11} + \lambda^2 \frac{|V_{12}|^2}{E_1^0 - E_2^0} - \lambda^3 \frac{|V_{12}|^2 (V_{11} - V_{22})}{\left(E_1^0 - E_2^0\right)^2} - \lambda^4 \frac{|V_{12}|^4}{\left(E_1^0 - E_2^0\right)^3} + O(\lambda^5) \tag{22}$$

which agrees with $E_1(\lambda)$ up to the λ^3 terms (the correct fourth order term contains also $|V_{12}|^2 (V_{11} - V_{22})^2 / (E_1^0 - E_2^0)^3$). This is an example of Wigner's $(2n + 1)$-theorem [52], see Subsect. "Wigner's Theorem".

The power expansions of $E_{1,2}(\lambda)$ and $\psi_{1,2}(\lambda)$ converge in the disk $|\lambda| < |E_1^0 - E_2^0| / \sqrt{(V_{11} - V_{22})^2 + 4|V_{12}|^2}$. The denominator is just twice the infimum over a of the operator norm of $V - aI$. Since adding to V a multiple of the identity does not affect the convergence properties of the Taylor's series of $E(\lambda)$, we see that the convergence domain always contains the disk $|\lambda| < |E_1^0 - E_2^0| / 2\|V\|$, a property which holds true for any bounded perturbation in Hilbert space (see Sect. "Problems with the Perturbation Series").

If H_0 is degenerate, that is $E_1^0 = E_2^0 \equiv E^0$, then the eigenvalues are obtained by diagonalizing V. The degeneracy is removed and the corrections to the eigenvalues are of first order in λ:

$$E_{1,2}(\lambda) = E^0 + \frac{1}{2}\lambda\Big(V_{11} + V_{22} \pm \sqrt{(V_{11} - V_{22})^2 + 4|V_{12}|^2}\Big) \tag{23}$$

while the eigenvectors are λ independent.

The infinite dimensional case is much more involved. In particular, in most cases the perturbation series does not converge at all, that is its radius of convergence vanishes. However, we shall meet again the three situations discussed above: the case of non-degenerate eigenvalues E_n^0 such that $|E_n^0 - E_m^0| \gg |\lambda V_{nm}|$, the case of degenerate eigenvalues and finally the case of "quasi-degenerate" eigenvalues, i. e. groups of eigenvalues $E_{n_i}^0$ such that $|E_{n_i}^0 - E_{n_j}^0| \lesssim |\lambda V_{n_i n_j}|$. As discussed above, in this last case $H_{n_i n_j}$ must be diagonalized exactly prior to applying perturbation theory.

Perturbation of Point Spectra: Nondegenerate Case

In this section we consider an eigenvector ψ_0 of H_0 belonging to a non-degenerate eigenvalue E_0 and apply the RS theory to determine the power expansions Eq. (11) such that Eq. (10) is satisfied, the Hamiltonian $H(\lambda)$ being given by Eq. (9). The case of a degenerate eigenvalue will be considered in Sect. "Perturbation of Point Spectra: Degenerate Case". For both cases the starting point is the substitution of the expansions Eq. (11) into Eq. (10), which, upon equating terms with equal powers, yields the following system of equations

$$(H_0 - E_0)\psi^{(n)} + V\psi^{(n-1)} = \sum_{k=0}^{n-1} \psi^{(k)} \epsilon_{n-k} , \qquad n = 1, 2, \ldots \tag{24}$$

A perturbative calculation of the energy and the wave function through order h amounts to calculating ϵ_n and $\psi^{(n)}$ up to $n = h$ and truncating the series in Eq. (11) at $n = h$.

Corrections to the Energy and the Eigenvectors

In the following let ψ_k, E_k be the normalized eigenvectors and the eigenvalues of H_0, and let $\Delta E_{k0} \equiv E_k - E_0$, $V_{hk} \equiv (\psi_h, V\psi_k)$. The correction ϵ_n is recursively defined in terms of the lower order corrections to the energy and the wave function: by left multiplying Eq. (24) by ψ_0 we find

$$\epsilon_n = \left(\psi_0, V\psi^{(n-1)}\right) - \sum_{h=1}^{n-1} \left(\psi_0, \psi^{(h)}\right) \epsilon_{n-h} \,. \tag{25}$$

Similarly, the components $(\psi_k, \psi^{(n)})$, $k \neq 0$, are found by left-multiplying by ψ_k, $k \neq 0$:

$$\left(\psi_k, \psi^{(n)}\right) = -\left(\psi_k, V\psi^{(n-1)}\right) \Delta E_{k0}{}^{-1} + \sum_{h=1}^{n-1} \left(\psi_k, \psi^{(h)}\right) \epsilon_{n-h} \Delta E_{k0}{}^{-1} \,. \tag{26}$$

Note that, even if the functions $\psi^{(k)}$'s for $k < n$ were known, still $(\psi_0, \psi^{(n)})$ is intrinsically undefined, since to any solution of Eq. (24) we are allowed to add any multiple of ψ_0. The reason of this indeterminacy is that Eq. (10) defines $\psi(\lambda)$ only up to a multiplicative factor $\alpha(\lambda)$. Even the normalization condition $(\psi(\lambda), \psi(\lambda)) = 1$ leaves $\psi(\lambda)$ undetermined by a phase factor $\exp(i\varphi(\lambda))$, $\varphi(\lambda) \in \mathbb{R}$. On the contrary, the corrections ϵ_n as well as all the expectation values (up to order n) are unaffected by these modifications of the wave function $\psi(\lambda)$ [16].

We can turn to our advantage the indeterminacy of $(\psi_0, \psi^{(n)})$ by requiring that in the expression of ϵ_n, Eq. (25), the dependence on the values of $(\psi_0, \psi^{(k)})$, $k \leq n-1$, disappears. For example, after writing

$$\left(\psi_0, V\psi^{(n-1)}\right) = V_{00} \left(\psi_0, \psi^{(n-1)}\right) + \sum_{h\neq 0} V_{0h} \left(\psi_h, \psi^{(n-1)}\right)$$

the independence of $(\psi_0, \psi^{(n-1)})$ implies $\epsilon_1 = V_{00}$. Next, requiring ϵ_n to be independent of $(\psi_0, \psi^{(n-2)})$ determines ϵ_2 and so on, until finally Eq. (25) gives ϵ_n. As an example we carry through this procedure for $n = 3$. Starting from

$$\epsilon_3 = V_{00} \left(\psi_0, \psi^{(2)}\right) + \sum_{k\neq 0} V_{0k} \left(\psi_k, \psi^{(2)}\right) - \epsilon_1 \left(\psi_0, \psi^{(2)}\right) - \epsilon_2 \left(\psi_0, \psi^{(1)}\right)$$

we first find

$$\epsilon_1 = V_{00} \,. \tag{27}$$

Next, from Eq. (26) for $n = 2$, we get

$$\epsilon_3 = -\sum_{k\neq 0} \frac{|V_{0k}|^2}{\Delta E_{k0}} \left(\psi_0, \psi^{(1)}\right) - \sum_{h,k\neq 0} \frac{V_{0k}}{\Delta E_{k0}} V_{kh} \left(\psi_h, \psi^{(1)}\right) + \sum_{k\neq 0} \frac{V_{0k}}{\Delta E_{k0}} \left(\psi_k, \psi^{(1)}\right) \epsilon_1 - \epsilon_2 \left(\psi_0, \psi^{(1)}\right) \,.$$

The independence of $(\psi_0, \psi^{(1)})$ implies

$$\epsilon_2 = -\sum_{k\neq 0} \frac{|V_{0k}|^2}{\Delta E_{k0}} \,. \tag{28}$$

Finally, by using Eq. (26) for $n = 1$ we find

$$\epsilon_3 = \sum_{h,k\neq 0} \frac{V_{0k}}{\Delta E_{k0}} \frac{V_{kh}}{\Delta E_{h0}} V_{h0} - \epsilon_1 \sum_{k\neq 0} \frac{|V_{0k}|^2}{\Delta E_{k0}{}^2} \,. \tag{29}$$

Note that if ϵ_n is required, the lower order corrections being known, one can use a simplified version of Eqs. (25) and (26) where the terms $(\psi_0, \psi^{(k)})$ are omitted since the beginning. Once the values of ϵ_k, $k \leq n$, have been determined, Eq. (26) yields $(\psi_k, \psi^{(n)})$. For example, for the first order correction to the wave function we have

$$\left(\psi_k, \psi^{(1)}\right) = -\frac{V_{k0}}{\Delta E_{k0}} \,. \tag{30}$$

By suitably choosing the arbitrary factor $\alpha(\lambda)$ we referred to after Eq. (26) we can impose $(\psi_0, \psi(\lambda)) = 1$. With this choice (known as the "intermediate normalization", since $\psi(\lambda)$ is not normalized) we have $(\psi_0, \psi^{(k)}) = 0$ for any $k > 0$. As a result, for the wave function through order n we find

$$\psi^{[n]} \equiv \psi_0 + \sum_{k=1}^{n} \lambda^k \psi^{(k)} \equiv \psi_0 + \delta_n \psi \tag{31}$$

with

$$\left(\psi_0, \psi^{[n]}\right) = 1 \,. \tag{32}$$

Using the intermediate normalization the expression of ϵ_n is

$$\epsilon_n = \left(\psi_0, V\psi^{(n-1)}\right) \,, \tag{33}$$

while the value of $(\psi_k, \psi^{(n-1)})$ can be read immediately in the expression of ϵ_n: $(\psi_k, \psi^{(n-1)})$ is obtained from ϵ_n by omitting in each term the factor V_{0k} and the sum over k.

For example the wave function $\psi^{[2]} \equiv \psi_0 + \lambda\psi^{(1)} + \lambda^2\psi^{(2)}$ in the intermediate normalization by Eqs. (28) and (29) is

$$\psi^{[2]} = \psi_0 - \lambda \sum_{k=1} \psi_k \frac{V_{k0}}{\Delta E_{k0}} + \lambda^2 \sum_{h,k=1} \psi_k \frac{V_{kh}}{\Delta E_{k0}} \frac{V_{h0}}{\Delta E_{h0}} - \lambda^2 \epsilon_1 \sum_{k=1} \psi_k \frac{V_{k0}}{\Delta E_{k0}{}^2} . \tag{34}$$

In order to calculate expectation values, transition probabilities and so on one needs the normalized wave function

$$\psi_N^{[n]} = N^{1/2} \left(\psi_0 + \delta_n \psi\right) \tag{35}$$

with $N^{-1} = 1 + (\delta_n\psi, \delta_n\psi)$. N can be chosen real. Note that the wave function $\psi^{[1]}$ is correctly normalized up to first order.

From the above equations one sees in which sense the perturbation V must be small with respect to the unperturbed Hamiltonian H_0: the separation between the unperturbed energy levels must be large with respect to the matrix elements of the perturbation between those levels and the total correction δE to E_0 should be small with respect to $|E_i - E_0|$, E_i standing for any other level of the spectrum of H_0.

Wigner's Theorem

From Eq. (24) it follows that

$$H\psi^{[n]} = E^{[n]}\psi^{[n]} + O\left(\lambda^{n+1}\right) ,$$

whence one should infer that, if E is the exact energy, $E - (\psi_N^{[n]}, H\psi_N^{[n]}) = O(\lambda^{n+1})$. It is therefore remarkable Wigner's result that the knowledge of $\psi^{[n]}$ allows the calculation of the energy up to order $2n+1$ (Wigner's $2n+1$ theorem) [52]. Indeed, he proved that, if E is the exact energy,

$$E - \frac{\left(\psi^{[n]}, H\psi^{[n]}\right)}{\left(\psi^{[n]}, \psi^{[n]}\right)} = O\left(\lambda^{2n+2}\right) .$$

To this purpose, let

$$\chi^{(n+1)} = \psi - \frac{\psi^{[n]}}{\sqrt{\left(\psi^{[n]}, \psi^{[n]}\right)}}$$

where ψ is the normalized exact wave function, $H\psi = E\psi$. Then

$$\chi^{(n+1)} = O\left(\lambda^{n+1}\right) ,$$
$$\left(\psi, \chi^{(n+1)}\right) + \left(\chi^{(n+1)}, \psi\right) = -\left(\chi^{(n+1)}, \chi^{(n+1)}\right) = O\left(\lambda^{2n+2}\right) .$$

As a consequence

$$\frac{\left(\psi^{[n]}, H\psi^{[n]}\right)}{\left(\psi^{[n]}, \psi^{[n]}\right)} - (\psi, H\psi) = O\left(\lambda^{2n+2}\right) .$$

We make explicit this point with an example. Since

$$\psi_N^{[1]} = \frac{\psi_0 + \lambda\psi^{(1)}}{\sqrt{1 + \lambda^2\left(\psi^{(1)}, \psi^{(1)}\right)}} ,$$

by using Eq. (30) and recalling Eq. (28) and Eq. (29) we have

$$\left(\psi_N^{[1]}, H\psi_N^{[1]}\right) = E_0 + \lambda\epsilon_1 + \frac{\lambda^2\epsilon_2 + \lambda^3\epsilon_3}{1 + \lambda^2\left(\psi^{(1)}, \psi^{(1)}\right)} = E_0 + \lambda\epsilon_1 + \lambda^2\epsilon_2 + \lambda^3\epsilon_3 + O\left(\lambda^4\right) .$$

The Feynman–Hellmann Theorem

The RS perturbative expansion rests on the hypothesis that both the eigenvalues $E(\lambda)$ and the corresponding eigenvectors $\psi(\lambda)$ admit a power series expansion, in short, that they are analytic functions of λ in a neighborhood of the origin. As we shall see in Sect. "Problems with the Perturbation Series", as a rule it is not so and the perturbative expansion gives rise only to a formal series. For this reason it is advisable to derive the various terms of the perturbation expansion without assuming analyticity. If we need $E(\lambda)$ and $\psi(\lambda)$ through order n it is sufficient to assume that, as functions of λ, they are C^{n+1}, that is continuously differentiable $(n+1)$ times. The procedure consists in taking the derivatives of Eq. (10) [15,28]: at the first step we get

$$H\psi'(\lambda) + V\psi(\lambda) = E'(\lambda)\psi(\lambda) + E(\lambda)\psi'(\lambda) \tag{36}$$

and by left multiplication by $\psi(\lambda)$, with $(\psi(\lambda), \psi(\lambda)) = 1$, we get

$$E'(\lambda) = \left(\psi(\lambda), V\psi(\lambda)\right) , \tag{37}$$

which is a special case of the Feynman–Hellmann theorem [17,23]:

$$\frac{\partial E}{\partial \lambda} = \left(\psi(\lambda), \frac{\partial H}{\partial \lambda}\psi(\lambda)\right) . \tag{38}$$

For $\lambda = 0$ we find

$$E'(0) = V_{00} , \tag{39}$$

whence $\epsilon_1 = V_{00}$, in agreement with Eq. (27). Next, after left multiplying Eq. (36) by ψ_k and taking $\lambda = 0$ we get

$$\left(\psi_k, \psi'\right) = -\frac{V_{k0}}{\Delta E_{k0}} \tag{40}$$

which, again, agrees with Eq. (30). Taking now the derivative of (37) at $\lambda = 0$ and using Eq. (40) we obtain

$$E''(0) = 2\sum_{k=1} V_{0k}\left(\psi_k, \psi'\right) = -2\sum_{k=1} \frac{|V_{0k}|^2}{\Delta E_{k0}} \tag{41}$$

whence $\epsilon_2 = \frac{1}{2}E''(0)$, in agreement with Eq. (28).

It is clear that the procedure can be pursued to any allowed order, and that the results for the energy corrections, as well as for the wave functions, are the same we obtained earlier by the RS technique. However, the conceptual difference, that no analyticity hypothesis is required, is important since in many cases this hypothesis is not satisfied.

As to the relation of $E^{[n]} \equiv E_0 + \lambda\epsilon_1 + \cdots + \lambda^n\epsilon_n$ with $E(\lambda)$ we recall that, since by assumption $E(\lambda)$ is C^{n+1}, we can write Taylor's formula with a remainder:

$$E(\lambda) = \sum_0^n \frac{E^{(p)}}{p!}\lambda^p + \frac{E^{(n+1)}(\theta\lambda)}{(n+1)!}\lambda^{n+1}, \quad 0 < \theta < 1 . \tag{42}$$

As observed in [28], since for small λ the sign of the remainder is the sign of $E^{(n+1)}(0)\lambda^{n+1}$, Eq. (42) allows to establish whether the sum in Eq. (42) underestimates or overestimates $E(\lambda)$. Moreover, if two consecutive terms, say q and $q+1$, have opposite sign, then (for sufficiently small λ) $E(\lambda)$ is bracketed between the partial sums including and excluding the qth term. It is a pity that no one can anticipate how small such a λ should be. (Of course these remarks apply to the RS truncated series as well.)

Perturbation of Point Spectra: Degenerate Case

The case when the unperturbed energy E_0 is a degenerate eigenvalue of H_0, i. e. in the Hilbert space there exists a subspace W_0 generated by a set $\{\psi_0^{(i)}\}, 1 \le i \le n_0$, of orthogonal normalized states, such that each ψ_0 in W_0 obeys $(H_0 - E_0)\psi_0 = 0$, deserves a separate treatment. The main problem is that, if $\psi(\lambda)$ is an eigenstate of the exact Hamiltonian $H = H_0 + \lambda V$, we do not know beforehand which state of W_0 $\psi(0)$ is.

In order to use a more compact notation it is convenient to introduce the projection P_0 onto the subspace W_0 and its complement Q_0

$$P_0\psi = \sum_{i=0}^{n_0} \psi_0^{(i)}\left(\psi_0^{(i)}, \psi\right), \quad Q_0 \equiv I - P_0 , \tag{43}$$

where $\psi_0^{(i)}$, $1 \le i \le n_0$, is any orthonormal basis of W_0. The Hamiltonian $H = H_0 + \lambda V$ can be written as

$$\begin{aligned} H &= (P_0 + Q_0)(H_0 + \lambda V)(P_0 + Q_0) \\ &= E_0P_0 + \lambda V_{PP} + \lambda V_{PQ} + \lambda V_{QP} + \lambda V_{QQ} \\ &\quad + Q_0H_0Q_0 , \end{aligned} \tag{44}$$

where

$$\begin{aligned} V_{PP} &= P_0VP_0 , \quad V_{PQ} = P_0VQ_0 , \\ V_{QP} &= Q_0VP_0 , \quad V_{QQ} = Q_0VQ_0 . \end{aligned} \tag{45}$$

After projecting the Schrödinger equation onto W_0 and its orthogonal complement $W_0^\perp$, we find

$$(E_0 + \lambda V_{PP})P_0\psi + \lambda V_{PQ}Q_0\psi = EP_0\psi \tag{46}$$

$$\lambda V_{QP}P_0\psi + Q_0H_0Q_0\psi + \lambda V_{QQ}Q_0\psi = EQ_0\psi . \tag{47}$$

Letting

$$H_{QQ} = Q_0HQ_0 = Q_0H_0Q_0 + \lambda V_{QQ} \tag{48}$$

$Q_0\psi$ can be extracted from Eq. (47):

$$Q_0\psi = \lambda(E - H_{QQ})^{-1}V_{QP}P_0\psi . \tag{49}$$

Note that in Eq. (47) the operator H_{QQ} acts on vectors of $W_0^\perp$ and that $E - H_{QQ}$ does possess an inverse in $W_0^\perp$. Indeed, the existence of a vector ζ in $W_0^\perp$ such that

$$(H_{QQ} - E)\zeta = 0 \tag{50}$$

contradicts the assumptions which perturbation theory is grounded in: the separation between $E(\lambda)$ and $E(0)$ should be negligible with respect to the separation between different eigenvalues of H_0. Actually, if ψ_k is such that $H_0\psi_k = E_k\psi_k$, $E_k \ne E_0$, by left multiplying Eq. (50) by ψ_k we would find

$$(E - E_k)(\psi_k, \zeta) = \lambda(\psi_k, V\zeta)$$

where the LHS is of order 0 in λ, whereas the RHS of order 1.

By substituting Eq. (49) into Eq. (46) we have

$$\begin{aligned} (E_0 + \lambda V_{PP})P_0\psi + \lambda^2 V_{PQ}(E - H_{QQ})^{-1}V_{QP}P_0\psi \\ = EP_0\psi . \end{aligned} \tag{51}$$

The energy shifts $\Delta E \equiv E - E_0$ appear as eigenvalues of an operator $A(E)$ acting in W_0

$$A(E) \equiv \lambda V_{PP} + \lambda^2 V_{PQ}(E - H_{QQ})^{-1}V_{QP} \tag{52}$$

which however still depends on the unknown exact energy E. A calculation of the energy corrections up to a given order is possible, starting from Eq. (52), provided we expand the term $(E - H_{QQ})^{-1}$ as far as is necessary to include all terms of the requested order.

Corrections to the Energy and the Eigenvectors

The contributions ϵ_i are extracted from Eq. (51) by expanding

$$E = E_0 + \lambda\epsilon_1 + \lambda^2\epsilon_2 + \cdots$$
$$P_0\psi = \varphi_0 + \lambda\varphi_1 + \lambda^2\varphi_2 + \cdots$$

and equating terms of equal order. At the first order, since the second term in the LHS of Eq. (51) is of order 2 or larger, we have

$$V_{PP}\varphi_0 = \epsilon_1\varphi_0 \,. \tag{53}$$

The first order corrections to the energy are the eigenvalues of the matrix V_{PP} and the corresponding zeroth order wave function is the corresponding eigenvector.

In the most favorable case the eigenvalues of V_{PP} are simple, and the degeneracy is completely removed since the first order of perturbation theory. In this case, in order to get the higher order corrections, we can avail ourselves of the arbitrariness in the way of splitting the exact Hamiltonian into a solvable unperturbed Hamiltonian plus a perturbation by putting

$$H = (H_0 + \lambda V_{PP}) + \lambda(V - V_{PP}) \equiv H_0' + \lambda V' \,. \tag{54}$$

The eigenvectors of H_0' are the solutions of Eq. (53), with eigenvalues $E_0 + \lambda\epsilon_1^{(i)}, 1 \le i \le n_0$, plus the eigenvectors ψ_j of H_0 with eigenvalues $E_j \neq E_0$. Since the eigenvalues $E_0 + \lambda\epsilon_1^{(i)}$ are no longer degenerate, the formalism of non-degenerate perturbation theory can be applied, but a warning is in order. When in higher perturbation orders a denominator ΔE_{k0} occurs with the index k referring to another vector of the basis of W_0, this denominator is of order λ and consequently the order of the term containing this denominator is lower than the naive V-counting would imply. In each such term, the effective order is the V-counting order minus the number of these denominators. As shown below, this situation occurs starting from terms of order 4 in the perturbation V. Note that, also in the case of non-complete removal of the degeneracy, the procedure outlined above, with obvious modifications, can be applied to search the higher order corrections to those eigenvalues which at first order turn out to be non-degenerate.

If a residual degeneracy still exists, i. e. an eigenvalue ϵ_1 of Eq. (53) is not simple, we must explore the higher order corrections until the degeneracy, if possible, is removed. First of all we must disentangle the contributions of different order in λ from $(E - H_{QQ})^{-1}$. Since

$$E - H_{QQ} = (E - Q_0H_0Q_0) \cdot \left[1 - \lambda(E - Q_0H_0Q_0)^{-1}V_{QQ}\right] \,,$$

we have

$$(E - H_{QQ})^{-1} = \sum_0^\infty \lambda^n \left[(E - Q_0H_0Q_0)^{-1}V_{QQ}\right]^n \cdot (E - Q_0H_0Q_0)^{-1} \,. \tag{55}$$

As the energy E still contains contributions of any order, the operator $(E - Q_0H_0Q_0)^{-1}$ must in turn be expanded into a series in λ. To make notations more readable, we define

$$\frac{Q_0}{a^n} \equiv (E_0 - Q_0H_0Q_0)^{-n} \,. \tag{56}$$

The second order terms from Eqs. (51) and (55) give

$$V_{PP}\varphi_1 + V_{PQ}\frac{Q_0}{a}V_{QP}\varphi_0 = \epsilon_2\varphi_0 + \epsilon_1\varphi_1 \,. \tag{57}$$

Let $P_0^{(i)}$ be the projections onto the subspaces $W_0^{(i)}$ of W_0 corresponding to the eigenvalues $\epsilon_1^{(i)}$:

$$\begin{aligned} &P_0 = \sum_i P_0^{(i)} \,, \quad V_{PP} = \lambda\sum \epsilon_1^{(i)}P_0^{(i)} \,, \\ &P_1 \equiv P_0^{(1)} \,, \quad \epsilon_1 \equiv \epsilon_1^{(1)} \,. \end{aligned} \tag{58}$$

By projecting onto $W_1 \equiv W_0^{(1)}$ and recalling that φ_0 is in W_1 we get

$$P_1V_{PQ}\frac{Q_0}{a}V_{QP}\varphi_0 = \epsilon_2\varphi_0 \,, \tag{59}$$

whence ϵ_2 is an eigenvalue of the operator

$$V_1 \equiv P_1V_{PQ}\frac{Q_0}{a}V_{QP}P_1 = P_1V\frac{Q_0}{a}VP_1 \,. \tag{60}$$

Again, if the eigenvalue ϵ_2 is non-degenerate, we can use the previous theory by splitting the Hamiltonian as

$$\begin{aligned} H &= (H_0 + \lambda V_{PP} + \lambda V_1) + \lambda(V - V_{PP} - V_1) \\ &\equiv H_0'' + \lambda V'' \,. \end{aligned} \tag{61}$$

The vectors which make V_1 diagonal belong to non-degenerate eigenvalues of H_0'', hence the non-degenerate theory can be applied. If, on the contrary, the eigenvalue ϵ_2 of V_1 is still degenerate, the above procedure can be carried out one step further, with the aim of removing the residual degeneracy. We work out the calculation for ϵ_3, since a new aspect of degenerate perturbation theory emerges: a truly third order term which is the ratio of a term of order 4 in the potential and a term of first order (see Eq. (67) below).

From Eq. (51) and (55) we extract the contribution of order 3:

$$V_{PP}\varphi_2 + V_{PQ}\frac{Q_0}{a}V_{QP}\varphi_1 - \epsilon_1 V_{PQ}\frac{Q_0}{a^2}V_{QP}\varphi_0 + V_{PQ}\frac{Q_0}{a}V_{QQ}\frac{Q_0}{a}V_{QP}\varphi_0 = \epsilon_1\varphi_2 + \epsilon_2\varphi_1 + \epsilon_3\varphi_0 \,. \tag{62}$$

We want to convert this equation into an eigenvalue problem for ϵ_3. In analogy with Eq. (58) we have

$$P_1 = \sum_i P_1^{(i)} \,, \quad V_1 = \sum \epsilon_2^{(i)} P_1^{(i)} \,, \qquad P_2 \equiv P_1^{(1)} \,, \quad \epsilon_2 \equiv \epsilon_2^{(1)} \,. \tag{63}$$

Since $P_2 V_{PP} = \epsilon_1 P_2$, first we eliminate φ_2 by applying P_2 to Eq. (62):

$$P_2 V_{PQ}\frac{Q_0}{a}V_{QP}\varphi_1 - \epsilon_1 P_2 V_{PQ}\frac{Q_0}{a^2}V_{QP}\varphi_0 + P_2 V_{PQ}\frac{Q_0}{a}V_{QQ}\frac{Q_0}{a}V_{QP}\varphi_0 = \epsilon_3 P_2\varphi_0 + \epsilon_2 P_2\varphi_1 \,. \tag{64}$$

Writing $\varphi_1 = \sum_i P_0^{(i)}\varphi_1$, since $P_2 V_1 = \epsilon_2 P_2$ the contribution with $i = 1$ of the first term in the LHS of Eq. (64) is $P_2 V_1 \varphi_1 = \epsilon_2 P_2 \varphi_1$. Hence, Eq. (64) reads

$$\sum_{i\neq 1} P_2 V_{PQ}\frac{Q_0}{a}V_{QP}P_0^{(i)}\varphi_1 - \epsilon_1 P_2 V_{PQ}\frac{Q_0}{a^2}V_{QP}\varphi_0 + P_2 V_{PQ}\frac{Q_0}{a}V_{QQ}\frac{Q_0}{a}V_{QP}\varphi_0 = \epsilon_3 P_2\varphi_0 + \epsilon_2 \sum_{i\neq 1} P_0^{(i)}\varphi_1 \,. \tag{65}$$

Finally, $P_0^{(i)}\varphi_1$, $i \neq 1$, is extracted from Eq. (57) by projecting with $P_0^{(i)}$, $i \neq 1$, and recalling that $P_0^{(i)}\varphi_0 = 0$ if $i \neq 1$:

$$P_0^{(i)}\varphi_1 = P_0^{(i)} V_{PQ}\frac{Q_0}{a}V_{QP}\varphi_0 / \left(\epsilon_1 - \epsilon_1^{(i)}\right) \,, \quad i \neq 1 \,. \tag{66}$$

Substituting into Eq. (65) we see that ϵ_3 is defined by the eigenvalue equation for the operator

$$V_2 \equiv P_2 V \frac{Q_0}{a} V \frac{Q_0}{a} V P_2 - \epsilon_1 P_2 V \frac{Q_0}{a^2} V P_2 + \sum_{i\neq 1} P_2 V \frac{Q_0}{a} V \frac{P_0^{(i)}}{\epsilon_1 - \epsilon_1^{(i)}} V \frac{Q_0}{a} V P_2 \,. \tag{67}$$

Despite the presence of four factors in the potential, the last term is actually a third order term due to the denominators $\epsilon_1 - \epsilon_1^{(i)}$.

The procedure outlined above, which essentially embodies the Rayleigh–Schrödinger approach, can be pursued until the degeneracy is (if possible, see below Sect. "Symmetry and Degeneracy") completely removed, after which the theory for the non-degenerate case can be used. Rather than detailing the calculations, we present an alternative iterative procedure due to Bloch [4] which allows a more systematic calculation of the corrections to the energy and the wave function.

Bloch's Method

In equations Eq. (51) and (52) we have seen that the energy corrections ΔE and the projections onto W_0 of the vectors $\psi_k(\lambda)$ are eigenvalues and eigenvectors of an operator acting in W_0. This observation is not immediately useful since the operator depends on the unknown exact energy $E(\lambda)$. However, it is possible to produce an operator $B(\lambda)$, which can be calculated in terms of known quantities and has the property that, if $E_k(\lambda)$, $\psi_k(\lambda)$ are eigenvalues and eigenvectors of Eq. (10) such that $E_k(0) = E_0$, then

$$B(\lambda)P_0\psi_k(\lambda) = \Delta E_k P_0 \psi_k(\lambda) \,. \tag{68}$$

First of all, note that the vectors $P_0\psi_k(\lambda)$ are a basis for the subspace W_0. Indeed, it is implicit in the assumption that perturbation theory does work that the perturbing potential should produce only slight modifications of the unperturbed eigenvectors of the Hamiltonian, so that the vectors $P_0\psi_k(\lambda)$ are linearly independent (although not orthogonal). Since their number equals the dimension of W_0, they are a basis for this subspace.

Following [4], we define a λ dependent operator U in this way:

$$UP_0\psi_k(\lambda) = \psi_k(\lambda) \,; \qquad UQ_0 = 0 \,. \tag{69}$$

As a consequence we have

$$U = UP_0 \,, \qquad P_0 U = P_0 \,, \tag{70}$$

$$U\psi_k(\lambda) = \psi_k(\lambda) \,. \tag{71}$$

The former of Eq. (70) follows immediately from the definition of U. Hence $P_0 U = P_0 U P_0$, which implies the latter of Eq. (70). Equation (71) is verified by applying the former of Eq. (70) to $\psi_k(\lambda)$.

Let

$$B(\lambda) \equiv \lambda P_0 V U \,. \tag{72}$$

We verify that, if $\Delta E_k \equiv E_k - E_0$, then

$$B(\lambda)P_0\psi_k(\lambda) = \Delta E_k P_0 \psi_k(\lambda) \,. \tag{73}$$

Indeed, by Eq. (69) we have $P_0 V U P_0 \psi_k(\lambda) = P_0 V \psi_k(\lambda)$. Writing Eq. (10) as

$$(H_0 - E_0 + \lambda V)\psi_k(\lambda) = \Delta E_k \psi_k(\lambda)$$

and multiplying by P_0 we find

$$\lambda P_0 V \psi_k(\lambda) = \Delta E_k P_0 \psi_k(\lambda) , \tag{74}$$

hence Eq. (73) is satisfied.

A practical use of Eq. (73) requires an iterative definition of U in terms of known quantities. From Eqs. (69) and (70) we have

$$U = P_0 U + Q_0 U = P_0 + Q_0 U P_0 . \tag{75}$$

We calculate the latter term of Eq. (75) on the vectors $P_0\psi_k(\lambda)$. Since

$$(\lambda V - \Delta E_k)\psi_k = (E_0 - H_0)\psi_k ,$$

recalling Eq. (69) we have

$$\begin{aligned} Q_0 U P_0 \psi_k(\lambda) &= Q_0 \psi_k(\lambda) \\ &= \frac{Q_0}{a}(\lambda V - \Delta E_k)\psi_k(\lambda) \\ &= \lambda \frac{Q_0}{a} V U \psi_k(\lambda) - \Delta E_k \frac{Q_0}{a} U \psi_k(\lambda) \\ &= \lambda \frac{Q_0}{a} V U \psi_k(\lambda) - \Delta E_k \frac{Q_0}{a} U P_0 \psi_k(\lambda) . \end{aligned}$$

By Eq. (74)

$$\begin{aligned} Q_0 U P_0 \psi_k(\lambda) &= \lambda \frac{Q_0}{a} V U \psi_k(\lambda) - \lambda \frac{Q_0}{a} U P_0 V \psi_k(\lambda) \\ &= \lambda \frac{Q_0}{a} V U \psi_k(\lambda) - \lambda \frac{Q_0}{a} U P_0 V U \psi_k(\lambda) \\ &= \lambda \frac{Q_0}{a} \big(V U - U V U\big) P_0 \psi_k(\lambda) . \end{aligned}$$

As a consequence the desired iterative equation for U is

$$U = P_0 + \lambda \frac{Q_0}{a}(V U - U V U) . \tag{76}$$

Equation (76) in turn allows an iterative definition of the operator $B(\lambda)$ of Eq. (72) depending only on quantities which can be computed in terms of the known spectral representation of H_0. Knowing U through order $n-1$ gives $B^{[n]}(\lambda) \equiv \sum_{i=1}^{n} B^{(i)}(\lambda)$, whose eigenvalues are the energy corrections through order n. In fact, if $B = \sum_{i=1}^{\infty} B^{(i)}$ and $P_0\psi_k = \sum_{s=0}^{\infty} \lambda^s \varphi_s$, the order r contribution to Eq. (73) is

$$\sum_{i=1}^{r} B^{(i)} \varphi_{r-i} = \sum_{i=1}^{r} \epsilon_i \varphi_{r-i} . \tag{77}$$

Defining $P_0 \psi_k^{[n]} \equiv \sum_0^n \lambda^r \varphi_r \equiv \varphi^{[n]}$, $\Delta E^{[n]} \equiv \sum_1^n \lambda^r \epsilon_r$, we see that the sum of Eq. (77) for values of r through n gives

$$B^{[n]} \varphi^{[n]} = \Delta E^{[n]} \varphi^{[n]} + O(\lambda^{n+1}) . \tag{78}$$

Once $P_0 \psi_k^{[n]}(\lambda)$ has been found, Eq. (69) gives the component of $\psi_k(\lambda)$ in $W_0^{\perp}$ through order $n+1$. As an example, for $n = 3$ we have

$$\begin{aligned} U^{[2]} = P_0 + \lambda \frac{Q_0}{a} V P_0 + \lambda^2 \frac{Q_0}{a} V \frac{Q_0}{a} V P_0 \\ - \lambda^2 \frac{Q_0}{a^2} V P_0 V P_0 , \end{aligned} \tag{79}$$

$$\begin{aligned} B^{[3]} = \lambda P_0 V P_0 + \lambda^2 P_0 V \frac{Q_0}{a} V P_0 + \lambda^3 P_0 V \frac{Q_0}{a} V \frac{Q_0}{a} V P_0 \\ - \lambda^3 P_0 V \frac{Q_0}{a^2} V P_0 V P_0 . \end{aligned} \tag{80}$$

If W_0 is one dimensional, Eq. (80) gives for $\lambda\epsilon_1 + \lambda^2\epsilon_2 + \lambda^3\epsilon_3$ the same result as Eqs. (27), (28) and (29).

The main difference between the RS perturbation theory and Bloch's method is that within the former the energy corrections through order n are calculated by means of a sequential computation starting from ϵ_1, with the consequence that at each step the dimension of the matrix to be diagonalized is smaller. Conversely, within Bloch's method one has to diagonalize the matrix $B^{[n]}(\lambda)$, which has the dimension of W_0. However, as noted above, for $n > 1$ the eigenvalues of $B^{[n]}(\lambda)$ are different from $\lambda\epsilon_1 + \lambda^2\epsilon_2 + \cdots + \lambda^n\epsilon_n$ by terms of order at least $n+1$. Similarly, the eigenvectors of Eq. (78) differ from the component in W_0 of $\psi^{[n]} = \psi_0 + \lambda\psi^{(1)} + \cdots + \lambda^n \psi^{(n)}$ by terms of order larger than n.

It is instructive to reconsider the calculation of ϵ_2 and ϵ_3 in the light of Bloch's method. If $P_0 = P_1 + P_1'$, then $P_0 V P_0 = \epsilon_1 P_1 + P_1' V P_1'$ and

$$\begin{aligned} B^{[2]}(\lambda) = \lambda\epsilon_1 P_1 + \lambda P_1' V P_1' + \lambda^2 P_1 V \frac{Q_0}{a} V P_1 \\ + \lambda^2 P_1' V \frac{Q_0}{a} V P_1' + \lambda^2 P_1 V \frac{Q_0}{a} V P_1' + \lambda^2 P_1' V \frac{Q_0}{a} V P_1 . \end{aligned}$$

The last two terms represent off-diagonal blocks which can be omitted for the calculation of $\lambda\epsilon_1 + \lambda^2\epsilon_2$, since the lowest order contribution to the eigenvalues of a matrix X from the off-diagonal terms X_{ij} is $|X_{ij}|^2/(X_{ii} - X_{jj})$. For a second order expansion as $B^{[2]}$ this yields third order contributions of the type

$$\lambda^3 P_1 V \frac{Q_0}{a} V \frac{P_1'}{\epsilon_1 - \epsilon_1'} V \frac{Q_0}{a} V P_1 .$$

These are just the contributions to ϵ_3 which we met in the RS approach: the expression of V_2 given in Eq. (67) combines the block-diagonal term of order 3 with the off-diagonal terms of order 2 giving a third order contribution.

The Quasi–Degenerate Case

There are cases, in both atomic and molecular physics, where the energy levels of H_0 present a multiplet structure: the energy levels are grouped into "multiplets" whose separation ΔE is large compared to the energy separation δE between the levels belonging to the same multiplet. For instance, in atomic physics this is the case of the fine structure (due to the the so called spin–orbit interaction) or of the hyperfine structure (due to the interaction of the nuclear magnetic moment with the electrons); in molecular physics typically this is the case of the rotational levels associated with the different and widely separated vibrational levels.

If a perturbation V is such that its matrix elements between levels of the same multiplet are comparable to δE, while being small with respect to ΔE, then naive perturbation theory fails because of the small energy denominators pertaining to levels belonging to the same multiplet. To solve this problem, named the problem of quasi-degenerate levels, once again we can exploit the arbitrariness in the way of splitting the Hamiltonian H into an unperturbed Hamiltonian and a perturbation. Let

$$E_0^{(1)} \equiv E_0 + \delta E^{(1)} , \quad E_0^{(2)} \equiv E_0 + \delta E^{(2)} , \quad \ldots , \quad E_0^{(n)} \equiv E_0 + \delta E^{(n)} ,$$

be the unperturbed energies within a multiplet, with E_0 any value close to the $E_0^{(i)}$'s (for instance their mean value), and $P_0^{(i)}$ the projections onto the corresponding eigenspaces. Let

$$H_0^0 \equiv H_0 - \sum_i \delta E^{(i)} P_0^{(i)} , \quad \widetilde{V} \equiv \lambda V + \sum_i \delta E^{(i)} P_0^{(i)} ,$$

so that

$$H = H_0^0 + \widetilde{V} . \tag{81}$$

We consider H_0^0 as the unperturbed Hamiltonian and $\widetilde{V}$ as the perturbation. From the physical point of view this procedure, if applied to all multiplets, is just the inclusion into the perturbation of those terms of H_0 that are responsible for the multiplet structure. With the splitting of the Hamiltonian as in Eq. (81) we can apply the methods of degenerate perturbation theory. The most efficient of these techniques is Bloch's method, which yields a simple prescription for the calculation of the corrections of any order. If for example we are content with the lowest order, we must diagonalize the matrix $P_0 \widetilde{V} P_0$, or equivalently $P_0 H P_0$, that is the energies through first order are the eigenvalues of the equation

$$P_0 H P_0 \psi = E P_0 \psi , \tag{82}$$

where $P_0 = \sum_i P_0^{(i)}$ is the projection onto W_0, the eigenspace of H_0^0 corresponding to the eigenvalue E_0. These eigenvalues are algebraic functions of λ, and no finite order approximation is meaningful, since all terms can be numerically of the same order, due to the occurrence of small denominators $(\delta E^{(i)} - \delta E^{(j)})^n$.

The Brillouin–Wigner Method

Equations (52) and (55) yield an alternative approach to the calculation of the energy shift ΔE due to a perturbation to a non-degenerate energy level E_0, the so called Brillouin–Wigner method [9,22,52]. In this case W_0, the space spanned by the unperturbed eigenvector ψ_0, is one-dimensional. The correction ΔE obeys the equation

$$\Delta E = (\psi_0, A(E)\psi_0) , \tag{83}$$

where the operator $A(E)$ is defined in Eq. (52).

Substituting into the expression of A the expansion Eq. (55) for $(E - H_{QQ})^{-1}$ and noting that, if $\{E_k\}$ is the spectrum of H_0,

$$\left(\psi_0, V_{PQ}(E - Q_0 H_0 Q_0)^{-1} V_{QP} \psi_0\right) = \sum_{k \neq 0} \frac{|V_{0k}|^2}{E - E_k} ,$$

$$\left(\psi_0, V_{PQ}(E - Q_0 H_0 Q_0)^{-1} V_{QQ}(E - Q_0 H_0 Q_0)^{-1} V_{QP} \psi_0\right) = \sum_{k,h \neq 0} V_{0k}(E - E_k)^{-1} V_{kh}(E - E_h)^{-1} V_{h0}$$

and so on, we find the following implicit expression for the exact energy E:

$$E = E_0 + \lambda(\psi_0, V\psi_0) + \lambda^2 \sum_{k \neq 0} \frac{|V_{0k}|^2}{E - E_k} + \lambda^3 \sum_{k,h \neq 0} \frac{V_{0k}}{E - E_k} \frac{V_{kh}}{E - E_h} V_{h0} + \cdots . \tag{84}$$

Consistently with the assumption that perturbation theory does work, the denominators in Eq. (84) are non-vanishing. The equation can be solved by arresting the expansion to a given power n in the potential and searching

a solution iteratively starting with $E = E_0$. However, the result differs from the energy $E^{[n]} = E_0 + \lambda\epsilon_1 + \lambda^2\epsilon_2 + \cdots + \lambda^n\epsilon_n$, calculated by means of the RS perturbation theory, by terms of order $n+1$ in the potential. The result of the RS perturbation theory can be recovered from the Brillouin–Wigner approach by substituting in the denominators $E = E_0 + \lambda\epsilon_1 + \lambda^2\epsilon_2 + \cdots + \lambda^n\epsilon_n$, expanding the denominators in powers of $\lambda^k\epsilon_k/E_0$ and equating terms of equal orders in both sides of Eq. (84).

As for the perturbed wave function, if the intermediate normalization is used, by Eq. (49) we have:

$$\psi = \psi_0 + Q_0\psi = \psi_0 + \lambda(E - H_{QQ})^{-1} V_{QP}\psi_0 \,. \quad (85)$$

Again, using the expansion Eq. (55) we find

$$\psi = \psi_0 + \lambda \sum_{k\neq 0} \psi_k \frac{V_{k0}}{E - E_k} + \lambda^2 \sum_{k,h\neq 0} \psi_k \frac{V_{kh}}{E - E_k} \frac{V_{h0}}{E - E_h} + \cdots \,. \quad (86)$$

As for the energy, if we arrest this expression to order n and substitute for E the value calculated by using Eq. (84), the result will differ from the one of Rayleigh–Schrödinger perturbation theory by terms of order $n+1$.

A major drawback of the Brillouin–Wigner method is its lack of size-consistency: for a system consisting of non-interacting subsystems, the perturbative correction to the energy of the total system is not the sum of the perturbative corrections to the energies of the separate subsystems through any finite order. This is best illustrated by the simple case of two systems a, b with unperturbed eigenvectors, energies and interactions ψ_0^a, E_0^a, λV^a and ψ_0^b, E_0^b, λV^b respectively. If for example the expansion Eq. (84) is arrested at order 2, by noting that the matrix elements $V_{0,ij}$ between the unperturbed state and the states $\psi_i^a\psi_j^b$ are

$$V_{0,ij} \equiv \left(\psi_0^a\psi_0^b, \left(V^a + V^b\right)\psi_i^a\psi_j^b\right) = \left(\psi_0^a, V^a\psi_i^a\right)\delta_{0j} + \left(\psi_0^b, V^b\psi_j^b\right)\delta_{0i} \,,$$

for the second order equation defining E we find

$$E = E_0^a + E_0^b + \lambda\epsilon_1^a + \lambda\epsilon_1^b + \lambda^2 \sum_j \frac{\left|V_{0j}^b\right|^2}{E - E_0^a - E_j^b} + \lambda^2 \sum_i \frac{\left|V_{0i}^a\right|^2}{E - E_0^b - E_i^a} \,. \quad (87)$$

On the other hand, for the energy of each system at second order we find

$$E^a = E_0^a + \lambda\epsilon_1^a + \lambda^2 \sum_i \frac{\left|V_{0i}^a\right|^2}{E_a - E_i^a} \,; \qquad E^b = E_0^b + \lambda\epsilon_1^b + \lambda^2 \sum_j \frac{\left|V_{0j}^b\right|^2}{E_b - E_j^b} \,. \quad (88)$$

It is apparent that the sum of the expression reported in Eq. (88) does not equal the expression of the energy reported in Eq. (87). This pathology is absent in the RS perturbation theory, where for non-interacting systems $E(\lambda) = E^a(\lambda) + E^b(\lambda)$, hence, for any j, $\epsilon_j = (1/j!)D^jE(\lambda)|_{\lambda=0} = \epsilon_j^a + \epsilon_j^b$.

Symmetry and Degeneracy

In Sect. "Perturbation of Point Spectra: Degenerate Case" we applied perturbation theory to the case of degenerate eigenvalues with special emphasis on the problem of the removal of the degeneracy at a suitable order of perturbation theory. The main problem is to know in advance whether the degeneracy can be removed completely, or a residual degeneracy is to be expected. The answer is given by group theory [21,51,53].

The very existence of degenerate eigenvalues of a Hamiltonian H is intimately connected with the symmetry properties of this operator. Generally speaking, a group G is a symmetry group for a physical system if there exists an associated set $\{T(g)\}$ of transformations in the Hilbert space of the system such that $|(T(g)\varphi, T(g)\psi)|^2 = |(\varphi, \psi)|^2$, $g \in G$ [53]. It is proven that the operators $T(g)$ must be either unitary or antiunitary [2,53]. We will consider the most common case that they are unitary and can be chosen in such a way that

$$T(g_1)T(g_2) = T(g_1g_2) \,, \quad g_1, g_2 \in G \,, \quad (89)$$

so that the operators $\{T(g)\}$ are a representation of G.

A system described by a Hamiltonian H is said to be invariant under the group G if the time evolution operator commutes with $T(g)$. Under fairly wide hypotheses this implies

$$\left[H, T(g)\right] = 0 \,, \quad g \in G \,. \quad (90)$$

A consequence is that, for any $g \in G$,

$$H\psi = E\psi \;\Rightarrow\; HT(g)\psi = ET(g)\psi \,, \quad (91)$$

that is the restrictions $T(g)|_W$ of the operators $T(g)$ to the space W corresponding to a given energy E are a representation of G. Given an orthonormal basis $\{\psi_i\}$ in W, we

have

$$T(g)\psi_i = \sum_j t_{ji}(g)\psi_j \tag{92}$$

and the vectors ψ_i are said to transform according to the representation of G described by the matrices t_{ji}.

This representation, apart from the occurrence of the so called accidental degeneracy (which in most cases actually is a consequence of the invariance of the Hamiltonian under additional transformations) is irreducible: no subspace of W is invariant under all the transformations of the group. As a consequence, knowing the dimensions d_j of the irreducible representations of G allows to predict the possible degree of degeneracy of a given energy level, since the dimension of W must be equal to one of the numbers d_j. If the group of invariance is Abelian, all the irreducible representations are one dimensional, and degeneracy can only be accidental.

Two irreducible representations are equivalent if there are bases which transform with the same matrix $t_{ji}(g)$. Otherwise they are inequivalent. The following orthogonality theorems hold. If a and b are inequivalent representations and $\psi_i^{(a)}, \varphi_j^{(b)}$ transform according these representations, then

$$\left(\psi_i^{(a)}, \varphi_j^{(b)}\right) = 0 \tag{93}$$

while, if a_r and a_s are equivalent, for the basis vectors $\psi_i^{(a_r)}$, $\varphi_j^{(a_s)}$ we have

$$\left(\psi_i^{(a_r)}, \varphi_j^{(a_s)}\right) = K_{rs}^{(a)}\delta_{ij}\,. \tag{94}$$

Moreover, if A_{rs} is a matrix which commutes with all the matrices $t_{ij}^{(a)}$ of an irreducible representation b, then $A_{rs} = a\delta_{rs}$ (Schur's lemma).

Symmetry and Perturbation Theory

If $H = H_0 + \lambda V$, let G_0 be the group under which H_0 is invariant. Although it is not the commonest case, we start with assuming that also the perturbation V commutes with $T(g)$ for any element g of G_0. As a rule W_0, the space of eigenvectors of H_0 with energy E_0, hosts an irreducible representation $T(g)$ of G_0. In this case the degeneracy cannot be removed at any order of perturbation theory. While this follows from general principles (for any value of λ, $\psi(\lambda)$ and $T(g)\psi(\lambda)$ are eigenvectors of $H(\lambda)$, and by continuity the eigenspace W_λ will have the same dimension as W_0), it is interesting to understand how the symmetry properties affect the mechanism of perturbation theory.

If $\{\psi_i^0\}$ is a basis of W_0 transforming according to an irreducible representation a of G_0, then the matrix $V_{ij} = (\psi_i^0, V\psi_j^0)$ commutes with all the matrices $t_{ji}^{(a)}(g)$ and, according to Schur's lemma, $(\psi_i^0, V\psi_j^0) = \nu\delta_{ij} = \epsilon_1\delta_{ij}$. No splitting occurs at the level of first order perturbation theory, neither can it occur at any higher order. Indeed, when V commutes with the operators $T(g)$, then Bloch's operator U, and consequently the operator $B(\lambda)$ of Eq. (72), both commute with the $T(g)$'s too. Again by Schur's lemma, the operator $B(\lambda)$ is a multiple of the identity. At any order of perturbation the degeneracy of the level is not removed.

In most of the cases, however, the perturbation V does not commute with all the operators $T(g)$. The set

$$G = \left\{g : g \in G_0, \left[T(g), V\right] = 0\right\}$$

is a subgroup G of G_0 and the group of invariance for the Hamiltonian H is reduced to G. W_0 generally contains G-irreducible subspaces W_i, $1 \le i \le n$: the operators $T(g)|_{W_0}$ are a reducible representation of G. The decomposition into irreducible representations of G is unique up to equivalence.

The crucial information we gain from group theory is the following: the number of energy levels which the energy E_0 is split into is the number of irreducible representations of G which the representation of G_0 in W_0 is split into. The degrees of degeneracy are the dimensions of these representations. What is relevant is that we only need to study the eigenspace W_0 of H_0, which is known by hypothesis.

In fact, let $W(\lambda)$ be the space spanned by the eigenvectors $\psi_k(\lambda)$ of $H(\lambda)$ such that $\psi_k(0) \in W_0$. $W(\lambda)$ is invariant under the operators $T(g)$, $g \in G$, since Bloch's operator U commutes with the operators $T(g)$, $g \in G$. $W(\lambda)$ can be decomposed into G-irreducible subspaces $W_k(\lambda)$, and in each of them by Schur's lemma the Hamiltonian $H(\lambda)$ is represented by a matrix $E_k(\lambda)I_{W_k(\lambda)}$. The projections $P_0W_k(\lambda)$ span the space W_0 and transform with the same representation of G as $W_k(\lambda)$, since P_0 commutes with $T(g)$ for any g in G_0, hence for any g in G. Thus, the space W_0 hosts as many irreducible representations of G as $W(\lambda)$. Assuming that the eigenvalues $E_k(\lambda)$ are different from one another, we see that the decomposition of the representation of G_0 in W_0 into irreducible representations of G determines the number and the degeneracy of the eigenvalues of $H(\lambda)$ such that the corresponding eigenvectors $\psi(\lambda)$ are in W_0 for $\lambda = 0$. The possibility that some of the $E_k(\lambda)$ are equal will be touched upon in the next subsection.

Examples where the above mechanism is at work are common in atomic physics. When an atom, whose unper-

turbed Hamiltonian H_0 is invariant under $O(3)$, is subjected to a constant electric field $\vec{E} = E\hat{z}$ (Stark effect), the invariance group G of its Hamiltonian is reduced to the rotations about the z axis ($SO(2)$) and the reflections with respect to planes containing the z axis. The irreducible representations of this group have dimension at most 2, and the G-irreducible subspaces of W_0 (the space generated by the eigenvectors $\psi_{E_0 lm}$ of H_0 corresponding to the energy E_0) are generated by $\psi_{E_0 l0}$ (one dimensional representation) and $\psi_{E_0 l\pm m}$ (two dimensional representations). Hence, the level E_0 is split into $l+1$ levels, the states with m and $-m$ remaining degenerate since reflections transform a vector with a given m into the vector with opposite m. Instead, if the atom is subjected to a constant magnetic field $\vec{B} = B\hat{z}$ (Zeeman effect), the surviving invariance group G consists of $SO(2)$ plus the reflections with respect to planes $z = z_0$. G being Abelian, the degeneracy is completely removed and this occurs at the first order of perturbation theory.

In the rather special case that W_0 contains subspaces transforming according to inequivalent representations of G_0, also a G_0-invariant perturbation V can separate in energy the states belonging to inequivalent representations. For example, the spectrum of alkali atoms can be calculated by considering in a first approximation an electron in the field of the unit charged atomic rest, which is treated as pointlike. In this problem the obvious invariance group of the Hamiltonian of the optical electron is $O(3)$, the group of rotations and reflections, and the space W_0 corresponding to the principal quantum number $n > 1$ contains n inequivalent irreducible representations which are labeled by the angular momentum $l \leq n-1$. When the finite dimension of the atomic rest is taken into account as a perturbation, its invariance under $O(3)$ splits the levels with given n and different l into n sublevels. A more careful consideration, however, shows that also the Lenz vector commutes with the unperturbed Hamiltonian [5], and that the space W_0 is irreducible under a larger group, the group $SO(4)$ [18], which is generated by the angular momentum and the Lenz vector. As a consequence the l degeneracy is by no means accidental: a space irreducible under a given group can turn out to be reducible with respect to one of its subgroups.

Group theory is a valuable tool in degenerate perturbation theory to search the correct vectors $\psi_k(0)$ which make the operator $P_0 V P_0$ diagonal. In fact, let $\psi_i^{(a)}$ be vectors which reduce the representation T of G in W_0 into its irreducible components $T^{(a)}$. The vectors $\psi_i^{(a)}$ and $V\psi_i^{(a)}$ transform according to the same irreducible representation $T^{(a)}$. Hence, by Eqs. (93) and (94) we find that the $P_0 V P_0$ is a diagonal block matrix with respect to inequivalent representations:

$$\left(\psi_i^{(a_r)}, V\psi_j^{(b_s)}\right) = K_{rs}^{(a)} \delta_{ij} \delta_{ab} \,, \tag{95}$$

with $\delta_{ab} = 1$ if representations a and b are equivalent, $\delta_{ab} = 0$ otherwise. The matrices $K_{rs}^{(a)}$ are generally much smaller than the full matrix of the potential. Thus, the operation of diagonalizing V is made easier by finding the G-irreducible subspaces W_a. Conversely, the reduction of an irreducible representation of a group G_0 in a space W_0 into irreducible representations of a subgroup G can be achieved by the following trick: find an operator V whose symmetry group is just G and interpret W_0 as the degeneracy eigenspace of a Hamiltonian H_0. The G-irreducible subspaces of W_0 are the eigenspaces of $P_0 V P_0$.

Level Crossing

As shown in the foregoing section, the existence of a non-Abelian group of symmetry for the Hamiltonian entails the existence of degenerate eigenvalues. The problem naturally arises as to whether there are cases when, on the contrary, the degeneracy is truly "accidental", that is it cannot be traced back to symmetry properties. The problem was discussed by J. Von Neumann and E.P. Wigner [49], who showed that for a generic $n \times n$ Hermitian matrix depending on real parameters $\lambda_1, \lambda_2, \ldots$, three real values of the parameters have to be adjusted in order to have the collapse of two eigenvalues (level crossing).

When passing to infinite dimension, arguments valid for finite dimensional matrices might fail. Moreover, often the Hamiltonian is not sufficiently "generic" so that level crossing may occur. As a consequence, we look for necessary conditions in order that, given the Hamiltonian $H(\lambda) = H_0 + \lambda V$, two eigenvalues collapse for some (real) value $\bar{\lambda}$ of the parameter λ: $E_1(\bar{\lambda}) = E_2(\bar{\lambda}) \equiv \bar{E}$. In this case, if $\psi_1(\bar{\lambda})$ and $\psi_2(\bar{\lambda})$ are any two orthonormal eigenvectors of $H(\bar{\lambda}) = H_0 + \bar{\lambda} V$ belonging to the eigenvalue $\bar{E}$, the matrix

$$H_{ij}\left(\bar{\lambda}\right) \equiv \left(\psi_i\left(\bar{\lambda}\right), \left(H_0 + \bar{\lambda} V\right) \psi_j\left(\bar{\lambda}\right)\right) \,, \quad i, j = 1, 2$$

must be a multiple of the identity:

$$H_{11}\left(\bar{\lambda}\right) = H_{22}\left(\bar{\lambda}\right) \,, \tag{96}$$

$$H_{12}\left(\bar{\lambda}\right) = 0 \,. \tag{97}$$

Equations (96) and (97) are three real equations for the unknown $\bar{\lambda}$; hence, except for special cases, level crossing cannot occur.

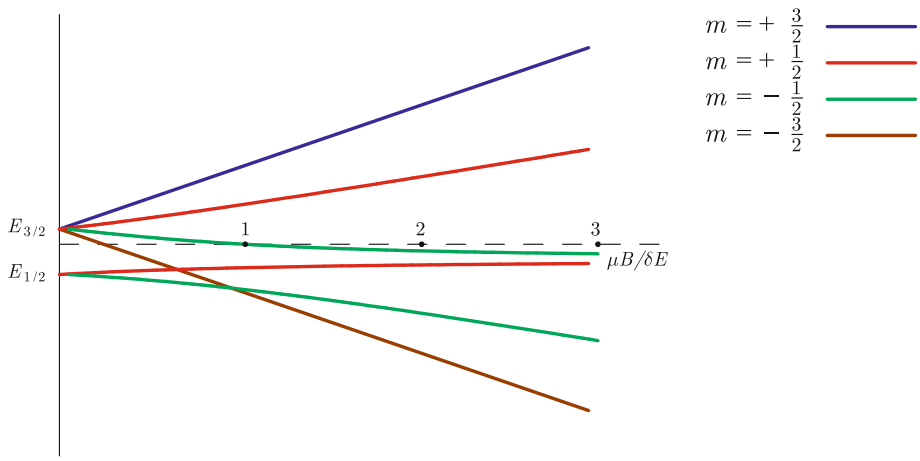

Perturbation Theory in Quantum Mechanics, Figure 2
The effect of a magnetic field on the doublet $2p_{1/2}, 2p_{3/2}$ of the lithium whose degeneracies, in the absence of the magnetic field, are respectively 2 and 4. μ is the magnetic moment of the electron and $\mu B/\delta E = 1$ for $B \approx 1.4\,\mathrm{T}$

The condition expressed by Eq. (97) is satisfied if the states corresponding to the eigenvalues $E_1(\lambda)$ and $E_2(\lambda)$ possess different symmetry properties, that is if they belong to inequivalent representations of the invariance group of the Hamiltonian or, equivalently, if they are eigenvectors with different eigenvalues of an operator which for any λ commutes with the Hamiltonian $H(\lambda)$ (hence it commutes with both H_0 and V). In this case $H_{12} = 0$ and the occurrence of level crossing depends on whether Eq. (96) has a real solution. This explains the statement that level crossing can occur only for states with different symmetry, while states of equal symmetry repel each other. Indeed, if Eq. (97) is not satisfied, the behavior of two close eigenvalues as functions of λ is illustrated in Fig. 1 (Sect. "Presentation of the Problem and an Example").

Figure 2 illustrates the behavior of the quasi-degenerate energy levels $2p_{1/2}, 2p_{3/2}$ of the lithium atom in the presence of an external magnetic field $\vec{B}$. In the absence of the magnetic field they are split by the spin-orbit interaction, with a separation $\delta E \equiv E_{3/2} - E_{1/2} = 0.4 \times 10^{-4}$ eV, to be compared with the separation in excess of 1 eV from the adjacent $2s$ and $3s$ levels. This justifies treating the effect of the magnetic field by means of the first order perturbation theory for quasi-degenerate levels.

When the magnetic field is present, the residual symmetry is the (Abelian) group of rotations about the direction of $\vec{B}$. Hence, the Hamiltonian commutes with the component of the angular momentum along the direction of $\vec{B}$, whose eigenvalues are denoted with m. In Fig. 2 the energies of states with equal symmetry, that is with the same value of m, are depicted with the same color. No crossing occurs between states with equal m, while the level with $m = -3/2$ does cross both the levels with $m = 1/2$ and with $m = -1/2$ which the $2p_{1/2}$ level is split into.

Problems with the Perturbation Series

So far we have assumed that all the power expansions appearing in the calculations were converging for $|\lambda| \leq 1$, that is we assumed analyticity in λ of $E(\lambda)$. Actually, it is only for rather special cases that analyticity can be proved. For most of the cases of physical interest, even if the terms of the perturbation series can be shown to exist, the series does not converge, or, when it converges, the limit is not $E(\lambda)$. In spite of this, special techniques have been devised to extract a good approximation to $E(\lambda)$ from the (generally few) terms of the perturbation series which can be computed. We will outline the main results existing in the field, without delving into mathematical details, for which we refer the reader to the books of Kato [25] and Reed–Simon [34] and the references therein.

The most favorable case is that of the so called regular perturbations [36,37,38,39], where the perturbation series does converge to $E(\lambda)$. More precisely, if E_0 is a nondegenerate eigenvalue of H_0, for λ in a suitable neighborhood of $\lambda = 0$ the Hamiltonian $H = H_0 + \lambda V$ has a nondegenerate eigenvalue $E(\lambda)$ which is analytic in λ and equals E_0 for $\lambda = 0$. The same property holds for the eigenvector $\psi(\lambda)$. A sufficient condition for this property to hold is expressed by the Kato–Rellich theorem [26,36,37,38,39], which essentially states that if the perturbation V is H_0-bounded, in the sense that constants a, b exist such that

$$\|V\psi\| \leq a\|H_0\psi\| + b\|\psi\| \tag{98}$$

for any ψ in the domain of V (which must include the domain of H_0) then the perturbation is regular. A lower bound to the radius r such that the perturbation series converges to the eigenvalue $E(\lambda)$ for $|\lambda| < r$ can be given in terms of the parameters a, b appearing in Eq. (98) and the distance δ of the eigenvalue E_0 from the rest of the spectrum of H_0. We have

$$r = \left[a + \frac{2}{\delta}\left[b + a\left(|E_0| + \frac{\delta}{2}\right)\right]\right]^{-1} . \tag{99}$$

It must be stressed, however, that the constants a and b are not uniquely determined by V and H_0. If the perturbation V is bounded ($a = 0$, $b = \|V\|$) condition Eq. (99) reads $r = \delta/(2\|V\|)$, which implies that the perturbation series for $H = H_0 + V$ with V bounded converges if $\|V\| < \delta/2$ (Kato bound [26]). The analysis of the two-level system (Sect. "Presentation of the Problem and an Example") shows that the figure 1/2 cannot be improved. Still, Kato bound is only a lower bound to r.

A similar statement holds for degenerate eigenvalues [34]: if E_0 has multiplicity m there are m single valued analytic functions $E_k(\lambda)$, $k = 1, \ldots, m$ such that $E_k(0) = E_0$ and, for λ in a neighborhood of 0, $E_k(\lambda)$ are eigenvalues of $H(\lambda) = H_0 + \lambda V$. Some of the functions $E_k(\lambda)$ may be coincident, and in a neighborhood of E_0 there are no other eigenvalues of $H(\lambda)$.

Regular perturbations are in fact exceedingly rare, a notable case being that of helium-like atoms [47]. Actually, there are cases where, although on physical grounds $H_0 + \lambda V$ does possess bound states, the relationship between $E(\lambda)$ and the RS expansion is far more complicated than for regular perturbations. As pointed out by Kramers [27], with an argument similar to an observation by Dyson [12] for quantum electrodynamics, the quartic anharmonic oscillator with Hamiltonian

$$H = H_0 + \lambda V \equiv \frac{p^2}{2m} + \frac{m\,\omega^2 x^2}{2} + \lambda \frac{m^2\omega^3}{\hbar} x^4 \tag{100}$$

is such an example. In fact, on the one hand bound states exist only for $\lambda \geq 0$; on the other hand, if a power series converges for $\lambda > 0$, then the series should converge also for negative values of λ. But for $\lambda < 0$ no bound state exists. Still worse, by estimating the coefficients of the RS expansion it has been proved that the series has vanishing radius of convergence [3].

In spite of this negative result, in this case it has been proved [46] that the perturbation series is an asymptotic series. This means that, for each n, if $\sum_0^n \epsilon_k \lambda^k$ is the sum through order n of the perturbation series, then

$$\lim_{\lambda \to 0} \frac{\sum_0^n \epsilon_k \lambda^k - E(\lambda)}{\lambda^n} = 0 . \tag{101}$$

We recall the difference between an asymptotic and an absolutely converging series, such as occurs with regular perturbations. For the latter one, given any λ in the convergence range of the series, the distance $|\sum_0^n \epsilon_k \lambda^k - E(\lambda)|$ can be made arbitrarily small provided n is sufficiently large (so that a converging series is also an asymptotic series). On the contrary, for an asymptotic series $|\sum_0^n \epsilon_k \lambda^k - E(\lambda)|$ is arbitrarily small only if λ is sufficiently near 0, but for a definite value of λ the quantity $|\sum_0^n \epsilon_k \lambda^k - E(\lambda)|$ might decrease to a minimum, attained for some value N, and then it could start to oscillate for $n > N$ (this is indeed the case for the anharmonic oscillator). As a consequence, for asymptotic series it is not expedient to push the calculation of the terms of the series beyond the limit where wild oscillations set in.

Any C^∞ function has an asymptotic series, as can be seen by inspection of the Taylor's formula with a remainder (see Eq. (42)). By this means Krieger [28] argued that, if ϵ_k (or equivalently the kth derivative of $E(\lambda)$) exists for any k, the RS series is asymptotic. However, generally there is not a range where the series converges to $E(\lambda)$, that is $E(\lambda)$ is not analytic. An asymptotic series may fail to converge at all for $\lambda \neq 0$, as noted for the anharmonic oscillator. The asymptotic series of a function, if it exists, is unique, but the converse is not true. For example, for the C^∞ function defined for real x as $f(x) = \exp(-1/x^2)$ if $x \neq 0$, $f(0) = 0$, the asymptotic series vanishes. There are also cases when the perturbation series is asymptotic for $\arg\lambda$ lying in a range $[\alpha, \beta]$. This occurs for example for the generalized anharmonic oscillator with perturbation $V \propto \lambda x^{2n}$. It has been proved that its perturbation series is asymptotic for $|\arg\lambda| \leq \theta < \pi$ [46] (note that the domain does not include negative values of λ). The result was later extended to multidimensional anharmonic oscillators [19]. General theorems stating sufficient hypotheses for the perturbation series to be asymptotic can be found in the literature. As a rule, however, they do not cover most of the cases of physical interest.

Even in the felicitous case when the perturbation series is asymptotic, it is only known that a partial sum approaches $E(\lambda)$ as much as desired provided λ is sufficiently small. This is not of much help to the practicing scientist, who generally is confronted with a definite value of the parameter λ, which can always be considered $\lambda = 1$ by an appropriate rescaling of the potential V. Recalling that different functions can have the same asymptotic series, it seems hopeless to try to recover the function $E(\lambda)$ from its asymptotic series, but this is possible for the so called strong asymptotic series. A function $E(\lambda)$ analytic in a sectorial region ($0 < |\lambda| < B$, $|\arg\lambda| < \pi/2 + \delta$) is said to have strong asymptotic series $\sum_0^\infty a_k \lambda^k$ if for all λ

in the sector

$$\left| E(\lambda) - \sum_0^n a_k \lambda^k \right| \leq C\sigma^{n+1} |\lambda|^{n+1} (n+1)! \tag{102}$$

for some constants C, σ. For strong asymptotic series it is proved that the function $E(\lambda)$ is uniquely determined by the series. Conditions that ensure that the RS series is a strong asymptotic series have been given [34].

The problem of actually recovering the function $E(\lambda)$ from its asymptotic series can be tackled by several methods. The most widely used procedure is the Borel summation method [6], which amounts to what follows. Given the strongly asymptotic series $\sum_0^\infty a_k \lambda^k$, one considers the series $F(\lambda) \equiv \sum_0^\infty (a_k/k!)\lambda^k$. This is known as the Borel transform of the initial series, which, by the hypothesis of strong asymptotic convergence, can be proved to have a non-vanishing radius of convergence and to possess an analytic continuation to the positive real axis. Then the function $E(\lambda)$ is given by

$$E(\lambda) = \int_0^\infty F(\lambda x) \exp(-x)\, \mathrm{d}x \,. \tag{103}$$

The above statement is Watson's theorem [50]. Roughly speaking, it yields the function $E(\lambda)$ as if the following exchange of the series with the integral were allowed:

$$\begin{aligned} E(\lambda) \sim \sum_0^\infty a_k \lambda^k &= \sum_0^\infty \frac{a_k}{k!} \int_0^\infty \exp(-x) x^k \,\mathrm{d}x \lambda^k \\ &= \int_0^\infty \exp(-x) \sum_0^\infty \frac{a_k}{k!} (x\lambda)^k \mathrm{d}x \\ &= \int_0^\infty F(\lambda x) \exp(-x) \mathrm{d}x \,. \end{aligned}$$

A practical problem with perturbation theory is that, apart from a few classroom examples, one is able to calculate only the lower order terms of the perturbation series. Although in principle it is impossible to divine the rest of a series by knowing its terms through a given order, a technique which in some cases turned out to work is that of Padé approximants [1,32]. A Padé $[M, N]$ approximant to a series is a rational function

$$R_{MN}(z) = \frac{P_M(z)}{Q_N(z)} \tag{104}$$

whose power expansion near $z = 0$ is equal to the first $M + N$ terms of the series. It has been proved [31] that the Padé $[N, N]$ approximants converge to the true eigenvalue of the anharmonic oscillator with x^4 or x^6 perturbation. The Padé $[M, N]$ approximant to a function $f(z)$ is unique, but its domain of analyticity is generally larger. Even for asymptotic series whose first terms are known one can write the Padé approximants. One can either use directly the Padé approximant as the value of $E(\lambda)$ for the desired value of λ, or can insert it into the Borel summation method. For the case of the quartic anharmonic oscillator (Eq. (100)) both methods have been proved to work (at the cost of calculating some tens of terms of the series).

Another approach to the problem is the method of self-similar approximants [55], whereby approximants to the function $E(\lambda)$ for which the asymptotic series is known are sought by means of products

$$f_{2p}(\lambda) = \prod_{i=1}^{p} (1 + A_i \lambda)^{n_i} \,. \tag{105}$$

The $2p$ parameters A_i, n_i, $1 \leq i \leq p$, are determined by equating the Taylor expansion of $f_{2p}(\lambda)$ with the asymptotic series through order $2p$ ($a_0 = 1$ can be assumed, with no loss of generality, see [55]). Also, odd order approximants f_{2p+1} are possible. For the anharmonic oscillator (Eq. (100)) the calculations exhibit a steady convergence to the correct value of the energy of both the even order and the odd order approximants also for $\lambda = 200$.

The problem with the above approaches is that their efficiency seems limited to toy models as the anharmonic oscillator. For realistic problems it is difficult to establish in advance that the method converges to the correct answer.

Perturbation of the Continuous Spectrum

In this section we consider the effect of a perturbing potential V on states belonging to the continuous spectrum. Since the problem is interesting mainly for the theory of scattering, we will assume that the unperturbed Hamiltonian H_0 is the free Hamiltonian of a particle of mass m. Also, assuming that the potential $V(\vec{r})$ vanishes at infinity, the spectrum of the free Hamiltonian H_0 and the continuous spectrum of the exact Hamiltonian $H = H_0 + \lambda V$ are equal and consist of the positive real semi-axis. Given an energy $E = \hbar^2 k^2/2m$, the problem is how the potential V affects that particular eigenfunction ψ_0 of H_0 which would represent the state of the system if the interaction potential were absent.

Letting $\psi = \psi_0 + \delta\psi$, the Schrödinger equation reads

$$(E - H_0)\delta\psi = \lambda V(\psi_0 + \delta\psi) \,. \tag{106}$$

In the spirit of the perturbation approach, $\delta\psi$ can be calculated by an iterative process provided we are able to find the solution of the inhomogeneous equation

$$(E - H_0)\delta\psi = \zeta \tag{107}$$

in the form

$$\delta\psi = \widetilde{G}_0\zeta\,, \tag{108}$$

$\widetilde{G}_0$ being the Green's function of Eq. (107). Assuming that $\widetilde{G}_0$ is known, we find

$$\begin{aligned}\delta\psi &= \lambda\widetilde{G}_0 V\psi\\ &= \lambda\widetilde{G}_0 V\psi_0 + \lambda\widetilde{G}_0 V\delta\psi\\ &= \lambda\widetilde{G}_0 V\psi_0 + \lambda\widetilde{G}_0 V(\lambda\widetilde{G}_0 V\psi_0 + \lambda\widetilde{G}_0 V\delta\psi)\\ &= \lambda\widetilde{G}_0 V\psi_0 + \lambda^2\widetilde{G}_0 V\widetilde{G}_0 V\psi_0 + \lambda^2\widetilde{G}_0 V\widetilde{G}_0 V\delta\psi\\ &= \cdots\,,\end{aligned} \tag{109}$$

that is $\delta\psi$ is written as a power expansion in λ in terms of the free wave function ψ_0.

Scattering Solutions and Scattering Amplitude

One has to decide which eigenfunction of H_0 must be inserted into the above expression, and which Green function $\widetilde{G}_0$ must be used, since, of course, the solution of Eq. (107) is not unique. The questions are strongly interrelated, and the answers depend on which solution of the exact Schrödinger equation one wishes to find. Since the study of the perturbation of the continuous spectrum is relevant mainly for the theory of potential scattering, we will focus on this aspect. In the theory of scattering it is shown [24] that, for a potential $V(\vec{r})$ vanishing faster than $1/r$ for $r\to\infty$, a wave function ψ which in the asymptotic region is an eigenfunction of the momentum operator plus an outgoing wave

$$\psi \overset{r\to\infty}{\longrightarrow} \exp\left(i\vec{k}\cdot\vec{r}\right) + f_{\vec{k}}(\theta,\varphi)\,\frac{\exp(ikr)}{r} \tag{110}$$

(θ, φ being the polar angles with respect to the $\vec{k}$ axis) is suitable for describing the process of diffusion of a beam of free particles with momentum $\vec{k}$ which impinge onto the interaction region and are scattered according the amplitude $f_{\vec{k}}(\theta,\varphi)$. (The character of outgoing wave of the second term in Eq. (110) is apparent when the time factor $\exp(-iEt)$ is taken into account.) The differential cross section $\mathrm{d}\sigma/\mathrm{d}\Omega$ is the ratio of the flux of the probability current density due to the outgoing wave to the flux due to the impinging plane wave. One finds

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \left|f_{\vec{k}}(\theta,\varphi)\right|^2\,. \tag{111}$$

In conclusion, we require that ψ_0 is a plane wave, and that the Green function $\widetilde{G}_0$ has to be chosen in such a way as to yield an outgoing wave for large r.

Thus we need to solve the equation

$$\begin{aligned}\left(\vec{k}^2+\Delta\right)\delta\psi &= \lambda\frac{2m}{\hbar^2}V\left(\exp\left(i\vec{k}\cdot\vec{r}\right)+\delta\psi\right)\\ &\equiv U\left(\vec{r}\right)\left(\exp\left(i\vec{k}\cdot\vec{r}\right)+\delta\psi\right)\end{aligned} \tag{112}$$

with the asymptotic condition $\delta\psi\to f_{\vec{k}}\exp(ikr)/r$ for $r\to\infty$. In terms of the Green's function $G_0(\vec{r},\vec{r}\,')$, which satisfies the equation

$$\left(\Delta+\vec{k}^2\right)G_0\left(\vec{r},\vec{r}\,'\right) = \delta\left(\vec{r}-\vec{r}\,'\right)\,, \tag{113}$$

the solution of Eq. (112) can be written as

$$\begin{aligned}\delta\psi\left(\vec{r}\right) = \int G_0\left(\vec{r},\vec{r}\,'\right)U\left(\vec{r}\,'\right)\\ \left[\exp\left(i\vec{k}\cdot\vec{r}\,'\right)+\delta\psi\left(\vec{r}\,'\right)\right]\mathrm{d}\vec{r}\,'\,,\end{aligned} \tag{114}$$

which is a form of the Lippmann–Schwinger equation [30]. The integral operator G_0 with kernel $G_0(\vec{r},\vec{r}\,')$ is connected to the operator $\widetilde{G}_0$ of Eq. (108) by the equation $\widetilde{G}_0 = 2mG_0/\hbar^2$.

The leading term of $\delta\psi$ for $r\to\infty$ is determined by the leading term of $G_0(\vec{r},\vec{r}\,')$, so we look for a solution of Eq. (113) with the behavior of outgoing wave for $r\to\infty$. Due to translation and rotation invariance (if both the incoming beam and the scattering potential are translated or rotated by the same amount, the scattering amplitude $f_{\vec{k}}(\theta,\varphi)$ is unchanged), we require for the solution a dependence only on $|\vec{r}-\vec{r}\,'|$.

Recalling that $\Delta 1/r = -4\pi\delta(\vec{r})$, we look for a solution of Eq. (113) with $\vec{r}\,'=0$ of the form $-F(r)/(4\pi r)$, with $F(0)=1$. The function $G_0(\vec{r},\vec{r}\,')$ then will be

$$G_0\left(\vec{r},\vec{r}\,'\right) = \frac{F\left(\left|\vec{r}-\vec{r}\,'\right|\right)}{\left|\vec{r}-\vec{r}\,'\right|}\,. \tag{115}$$

The equation for $F(r)$ is

$$F'' + k^2F = 0\,, \tag{116}$$

whose solutions are $\exp(\pm ikr)$ (outgoing and incoming wave respectively). In conclusion for G_0 we find

$$G_0(\vec{r},\vec{r}\,') = -\frac{1}{4\pi}\frac{\exp\left(ik\left|\vec{r}-\vec{r}\,'\right|\right)}{\left|\vec{r}-\vec{r}\,'\right|}\,. \tag{117}$$

The solution of the Schrödinger equation with the Green function given in Eq. (117) is denoted as $\psi^+_{\vec{k}}$ and obeys the integral equation known as the Lippmann–Schwinger

equation [30]:

$$\psi_{\vec{k}}^{+}(\vec{r}) = \exp\left(i\vec{k}\cdot\vec{r}\right) - \frac{1}{4\pi}\int \frac{\exp\left(ik\left|\vec{r}-\vec{r}\,'\right|\right)}{\left|\vec{r}-\vec{r}\,'\right|} U\left(\vec{r}\,'\right)\psi_{\vec{k}}^{+}\left(\vec{r}\,'\right)\,\mathrm{d}\vec{r}\,' . \quad (118)$$

The behavior for $r \to \infty$ can be easily checked to be as in Eq. (110) by inserting the expansion

$$\left|\vec{r}-\vec{r}\,'\right| = r - \frac{\vec{r}\cdot\vec{r}\,'}{r} + O(1/r) \quad (119)$$

into the Green function G_0. We find ($\hat{r} \equiv \vec{r}/r$)

$$-\frac{1}{4\pi}\frac{\exp\left(ik\left|\vec{r}-\vec{r}\,'\right|\right)}{\left|\vec{r}-\vec{r}\,'\right|} \overset{r\to\infty}{\longrightarrow} -\frac{1}{4\pi}\frac{\exp\left[ik\left(r-\hat{r}\cdot\vec{r}\,'\right)\right]}{r}\left[1+\frac{\vec{r}\cdot\vec{r}\,'}{r^2}\right] , \quad (120)$$

which yields for $\psi_{\vec{k}}^{+}(\vec{r})$ ($\vec{k}_f \equiv k\hat{r}$)

$$\psi_{\vec{k}}^{+}(\vec{r}) \overset{r\to\infty}{\longrightarrow} \exp\left(i\vec{k}\cdot\vec{r}\right) - \frac{1}{4\pi}\frac{\exp(ikr)}{r} \int \exp\left(-i\vec{k}_f\cdot\vec{r}\,'\right) U\left(\vec{r}\,'\right)\psi_{\vec{k}}^{+}\left(\vec{r}\,'\right)\,\mathrm{d}\vec{r}\,' . \quad (121)$$

The solutions $\psi_{\vec{k}}^{+}(\vec{r})$ are normalized as the plane waves $\exp(i\vec{k}\cdot\vec{r})$:

$$\left(\psi_{\vec{k}}^{+}, \psi_{\vec{k}\,'}^{+}\right) = (2\pi)^3\delta\left(\vec{k}-\vec{k}\,'\right) . \quad (122)$$

In addition, they are orthogonal to any possible bound state solution of the Schrödinger equation with the Hamiltonian $H = H_0 + \lambda V$. Together with the bound state solutions they constitute a complete set. On a par with the solutions $\psi_{\vec{k}}^{+}(\vec{r})$ one can also envisage solutions $\psi_{\vec{k}}^{-}(\vec{r})$ with asymptotic behavior of incoming wave. They are obtained using for H (see Eq. (116)) the solution $\exp(-ikr)$. The normalization and orthogonality properties of the functions $\psi_{\vec{k}}^{-}(\vec{r})$ are the same as for the $\psi_{\vec{k}}^{+}(\vec{r})$ functions.

From Eq. (121) we derive an implicit expression for the scattering amplitude $f_{\vec{k}}(\theta,\varphi)$:

$$f_{\vec{k}}(\theta,\varphi) = -\frac{1}{4\pi}\int \exp\left(-i\vec{k}_f\cdot\vec{r}\,'\right) U\left(\vec{r}\,'\right)\psi_{\vec{k}}^{+}\left(\vec{r}\,'\right)\,\mathrm{d}\vec{r}\,' \quad (123)$$

where the unknown function $\psi_{\vec{k}}^{+}(\vec{r})$ still appears. Letting $\varphi_f \equiv \exp\left(i\vec{k}_f\cdot\vec{r}\right)$, Eq. (123) can also be written as

$$f_{\vec{k}}(\theta,\varphi) = -\frac{1}{4\pi}\left(\varphi_f, U\psi_{\vec{k}}^{+}\right) . \quad (124)$$

The Born Series and its Convergence

Equations (118) and (124) are the starting point to obtain the expression of the exact wave function $\psi_{\vec{k}}^{+}(\vec{r})$ and the scattering amplitude $f_{\vec{k}}(\theta,\varphi)$ as a power series in λ, in the spirit of the perturbation approach. Recalling that $U = (2m/\hbar^2)V$ (see Eq. (112)), if $\varphi_{\vec{k}} \equiv \exp(i\vec{k}\cdot\vec{r})$ for $\psi_{\vec{k}}^{+}(\vec{r})$ we find

$$\begin{aligned}\psi_{\vec{k}}^{+} &= \varphi_{\vec{k}} + \lambda G_0\frac{2m}{\hbar^2}V\varphi_{\vec{k}} + \lambda^2 G_0\frac{2m}{\hbar^2}VG_0\frac{2m}{\hbar^2}V\varphi_{\vec{k}} + \cdots \\ &= \exp\left(i\vec{k}\cdot\vec{r}\right) + \lambda\int G_0\left(\vec{r},\vec{r}\,'\right)\frac{2m}{\hbar^2}V\left(\vec{r}\,'\right) \\ &\exp\left(i\vec{k}\cdot\vec{r}\,'\right)\mathrm{d}\vec{r}\,' + \lambda^2\int \mathrm{d}\vec{r}\,'\int \mathrm{d}\vec{r}\,'' G_0\left(\vec{r},\vec{r}\,'\right)\frac{2m}{\hbar^2} \\ &V\left(\vec{r}\,'\right)G_0\left(\vec{r}\,',\vec{r}\,''\right)\frac{2m}{\hbar^2}V\left(\vec{r}\,''\right)\exp\left(i\vec{k}\cdot\vec{r}\,''\right) + \cdots .\end{aligned} \quad (125)$$

Inserting the above expansion into Eq. (124), for the scattering amplitude $f_{\vec{k}}(\theta,\varphi)$ we find

$$f_{\vec{k}}(\theta,\varphi) = \sum_{n=1}^{\infty} f_{\vec{k}}^{(n)}(\theta,\varphi) \quad (126)$$

where $f_{\vec{k}}^{(n)}$, the contribution of order n in λ to $f_{\vec{k}}(\theta,\varphi)$, is obtained by substituting $\psi_{\vec{k}}^{+}$ in Eq. (124) with the contribution of order $n-1$ of the expansion Eq. (125). The term of order 1 is called the Born approximation [8], and is given by

$$f_{\vec{k}}^{B}(\theta,\varphi) \equiv f_{\vec{k}}^{(1)}(\theta,\varphi) = -\frac{1}{4\pi}\left(\frac{2m\lambda}{\hbar^2}\right) \int \exp\left[i\left(\vec{k}-\vec{k}_f\right)\cdot\vec{r}\,'\right] V\left(\vec{r}\,'\right)\,\mathrm{d}\vec{r}\,' . \quad (127)$$

The term of order 2 is

$$f_{\vec{k}}^{(2)}(\theta,\varphi) = -\frac{1}{4\pi}\left(\frac{2m\lambda}{\hbar^2}\right)^2\left(\varphi_f, VG_0V\varphi_{\vec{k}}\right) \quad (128)$$

and the general term of order n is

$$f_{\vec{k}}^{(n)}(\theta,\varphi) = -\frac{1}{4\pi}\left(\frac{2m\lambda}{\hbar^2}\right)^n\left(\varphi_f, VG_0V\cdots G_0V\varphi_{\vec{k}}\right) \quad (n \text{ times } V) . \quad (129)$$

The scattering amplitude through order n is

$$f_{\vec{k}}^{[n]}(\theta,\varphi) = \sum_{i=1}^{n} f_{\vec{k}}^{(i)}(\theta,\varphi) \quad (130)$$

with $f_{\vec{k}}^{(1)}(\theta,\varphi) \equiv f_{\vec{k}}^{B}(\theta,\varphi)$, and the series $\sum_1^\infty f_{\vec{k}}^{(i)}(\theta,\varphi)$ is known as the Born series [8]. Of course, when using Eq. (130) for calculating the differential cross section $d\sigma/d\Omega$ only terms of order not exceeding n should be consistently retained.

For a discussion of the range of validity and the convergence of the expansions Eqs. (125) and (126) it is convenient to pose the problem in the framework of integral equations in the Hilbert space L^2 [40,54], which provides a natural notion of convergence. To this purpose, since $\psi_{\vec{k}}^{+}(\vec{r})$ is not square integrable, we start assuming that the potential $V(\vec{r})$ is summable

$$\int \left|V\left(\vec{r}\right)\right| d\vec{r} < \infty \tag{131}$$

and multiply Eq. (118) by $|V(\vec{r})|^{\frac{1}{2}}$ [20,41,45]. Letting $\epsilon_V(\vec{r}) \equiv V(\vec{r})/|V(\vec{r})|$ ($\epsilon_V(\vec{r}) \equiv 0$ if $V(\vec{r}) = 0$) and defining

$$\zeta_{\vec{k}}^{+}\left(\vec{r}\right) \equiv \left|V\left(\vec{r}\right)\right|^{\frac{1}{2}} \psi_{\vec{k}}^{+}\left(\vec{r}\right) \tag{132}$$

$$K'\left(\vec{r},\vec{r}\,'\right) \equiv -\frac{2m}{\hbar^2} G_0\left(\vec{r},\vec{r}\,'\right) \left|V(\vec{r})\right|^{\frac{1}{2}} \left|V(\vec{r}\,')\right|^{\frac{1}{2}} \epsilon_v\left(\vec{r}\,'\right) , \tag{133}$$

Eq. (118) reads:

$$\zeta_{\vec{k}}^{+}\left(\vec{r}\right) = \left|V(\vec{r})\right|^{\frac{1}{2}} \exp\left(i\vec{k}\cdot\vec{r}\right) + \lambda \int K'\left(\vec{r},\vec{r}\,'\right) \zeta_{\vec{k}}^{+}\left(\vec{r}\,'\right) d\vec{r}\,' . \tag{134}$$

Now the function in front of the integral is square integrable and the kernel $K'(\vec{r},\vec{r}\,')$ is square integrable too

$$\int d\vec{r} \int d\vec{r}\,' \left|K'(\vec{r},\vec{r}\,')\right|^2 = \int d\vec{r} \int d\vec{r}\,' \frac{\left|V\left(\vec{r}\right)\right| \left|V\left(\vec{r}\,'\right)\right|}{\left|\vec{r}-\vec{r}\,'\right|^2} < \infty \tag{135}$$

provided the potential $V(\vec{r})$ obeys the additional condition

$$\int d\vec{r}\,' \frac{\left|V\left(\vec{r}\,'\right)\right|}{\left|\vec{r}-\vec{r}\,'\right|^2} < \infty . \tag{136}$$

Equation (134) can be formally written as

$$\zeta_{\vec{k}}^{+} = \zeta_0 + \lambda \hat{K}' \zeta_{\vec{k}}^{+} , \quad \zeta_0 \equiv \left|V(\vec{r})\right|^{\frac{1}{2}} \exp\left(i\vec{k}\cdot\vec{r}\right) , \tag{137}$$

$\hat{K}'$ being the integral operator with kernel K' given in Eq. (133). The function $\zeta_{\vec{k}}^{+}$ is formally given as

$$\zeta_{\vec{k}}^{+} = \left(I - \lambda \hat{K}'\right)^{-1} \zeta_0 \tag{138}$$

where the inverse operator $(I - \lambda \hat{K}')^{-1}$ exists except for those values of λ (singular values) for which $I - \lambda \hat{K}'$ has the eigenvalue 0.

Since by Eq. (135) $\hat{K}'$ is a compact operator, the singular values are isolated points which obey the inequality $|\lambda| \geq \|\hat{K}'\|^{-1}$, since the spectrum of an operator is contained in the closed disc of radius equal to the norm of the operator. Thus, when $\|\lambda \hat{K}'\| < 1$ the inverse operator $(I - \lambda \hat{K}')^{-1}$ exists and is given by the Neumann series

$$\left(I - \lambda \hat{K}'\right)^{-1} = I + \lambda \hat{K}' + \lambda^2 \hat{K}'^2 + \cdots \equiv I + R_\lambda^{K'} , \tag{139}$$

which is clearly norm convergent. By the inequality

$$\left\|\lambda \hat{K}'\right\|^2 \leq \lambda^2 \operatorname{Tr}\left(\hat{K}'^{\dagger} \hat{K}'\right) = \lambda^2 \int d\vec{r} \int d\vec{r}\,' \left|K'\left(\vec{r},\vec{r}\,'\right)\right|^2 \tag{140}$$

we see that, if

$$\lambda^2 \frac{m^2}{4\pi^2\hbar^4} \int d\vec{r} \int d\vec{r}\,' \frac{\left|V(\vec{r})\right| \left|V\left(\vec{r}\,'\right)\right|}{\left|\vec{r}-\vec{r}\,'\right|^2} < 1 , \tag{141}$$

the condition $\|\lambda \hat{K}'\| < 1$ is satisfied and consequently the inverse operator $(I - \lambda \hat{K}')^{-1}$ exists. In conclusion, a sufficient condition for the convergence of the expansion

$$\zeta_{\vec{k}}^{+} = \zeta_0 + \lambda \hat{K}' \zeta_0 + \lambda^2 K'^2 \zeta_0 + \cdots \tag{142}$$

in the L^2 norm is that Eq. (141) holds [43]. Since $\|\hat{K}'\|^4 \leq \operatorname{Tr}(\hat{K}'^{\dagger} \hat{K}' \hat{K}'^{\dagger} \hat{K}')$, by the Riemann–Lebesgue lemma it is possible to prove [56] that for any given λ the condition $\|\lambda \hat{K}'\| < 1$ is satisfied provided the energy $\hbar^2 k^2/2m$ is sufficiently large.

The implications for the convergence of the expansion Eq. (126) of the scattering amplitude are immediate, once Eq. (124) is written in the form

$$f_{\vec{k}}(\theta,\varphi) = -\lambda \frac{m}{2\pi\hbar^2} \left(\left|V\left(\vec{r}\right)\right|^{\frac{1}{2}} \varphi_f\left(\vec{r}\right) \epsilon_V\left(\vec{r}\right) , \left|V\left(\vec{r}\right)\right|^{\frac{1}{2}} \psi_{\vec{k}}^{+}\left(\vec{r}\right)\right) . \tag{143}$$

The Born series converges whenever the iterative solution of Eq. (137) converges, that is if Eq. (135) is satisfied. As noted above, for any given λ this happens for sufficiently large energy. An additional useful result is that the Born approximation $f_{\vec{k}}^{B}(\theta,\varphi)$ (or the expansion $f_{\vec{k}}^{[n]}(\theta,\varphi)$ through any n) converges to the exact scattering amplitude $f_{\vec{k}}(\theta,\varphi)$ when the energy grows to infinity. More precisely [48], if Eq. (131) holds then

$$\left|f_{\vec{k}}(\theta,\varphi) - f_{\vec{k}}^{[n]}(\theta,\varphi)\right| \overset{k\to\infty}{\longrightarrow} 0 . \tag{144}$$

If $\|\lambda\hat{K}'\| \geq 1$ the Neumann series Eq. (139) does not converge and the perturbation approach is no longer viable. However, if λ is not a singular value Eq. (137) can be solved by reducing it to an integral equation with a kernel D of norm less than 1 plus a problem of linear algebra [54]. In fact, for any positive value L the operator $\hat{K}'$ can be approximated by a finite rank operator F

$$F\zeta = \sum_{i=1}^{n} \alpha_i\left(\vec{r}\right)\left(\beta_i\left(\vec{r}\,'\right), \zeta\left(\vec{r}\,'\right)\right) \tag{145}$$

such that, if $D \equiv \hat{K}' - F$, $\|D\| < 1/L$. Equation (137) then reads

$$(I - \lambda D)\zeta_{\vec{k}}^{+} = \zeta_0 + \lambda F\zeta_{\vec{k}}^{+} \,. \tag{146}$$

Since for $|\lambda| < L$ we have $\|\lambda D\| < 1$, $\zeta_{\vec{k}}^{+}$ can be written in terms of the appropriate Neumann series $I + R_\lambda^D$ (see Eq. (139)):

$$\zeta_{\vec{k}}^{+} = \zeta_0 + R_\lambda^D\zeta_0 + F\zeta_{\vec{k}}^{+} + R_\lambda^D F\zeta_{\vec{k}}^{+} \,. \tag{147}$$

The unknown quantities $(\beta_i, \zeta_{\vec{k}}^{+})$ which appear in the RHS of Eq. (147) are determined by solving the linear-algebraic problem obtained by left-multiplying both sides of the equation by β_r, $1 \leq r \leq n$. The values of λ in the range $|\lambda| < L$ for which the algebraic problem is not soluble are the singular values of Eq. (137) in that range. Thus, Eq. (137) can be solved for any non-singular value.

Time Dependent Perturbations

A rather different problem is presented by the case that a time independent Hamiltonian H_0, for which the spectrum and the eigenfunctions are known, is perturbed by a time dependent potential $V(t)$. This occurs, for example, when an atom or a molecule interacts with an external electromagnetic field. For the total Hamiltonian $H = H_0 + \lambda V(t)$ stationary states no longer exist, and the relevant question is how the perturbation affects the time evolution of the system. We assume that the state ψ is known at a given time, which can be chosen as $t = 0$, and search for $\psi(t)$. Obviously, any time t_0 prior to the setting on of the perturbation $\lambda V(t)$ could be chosen instead of $t = 0$.

The time dependent Schrödinger equation reads

$$i\hbar\frac{\partial\psi}{\partial t} = H\psi(t) = H_0\psi(t) + \lambda V(t)\psi(t) \,. \tag{148}$$

At any t, $\psi(t)$ can be expanded in the basis of the eigenfunctions $\varphi_n(t)$ of H_0:

$$H_0\varphi_n(t) = E_n\varphi_n(t) \tag{149}$$

$$\varphi_n(t) = \varphi_n(0)\exp(-iE_n t/\hbar) \equiv \zeta_n \exp(-iE_n t/\hbar) \,. \tag{150}$$

For the sake of simplicity we treat H_0 as if it only had discrete spectrum, but the presence of a continuous component of the spectrum does not create any problem.

We can write [11,42]

$$\psi(t) = \sum_n a_n(t)\varphi_n(t) \,. \tag{151}$$

Note that the basis vectors $\varphi_n(t)$ are themselves time dependent (by the phase factor given in Eq. (150)), whereas the vectors ζ_n are time independent. The isolation of the contribution of H_0 to the time evolution as a time dependent factor allows a simpler system of equations for the unknown coefficients $a_n(t)$.

Substituting expansion Eq. (151) into Eq. (148) we find

$$\begin{aligned} i\hbar\sum_n \dot{a}_n(t)\varphi_n(t) + \sum_n a_n(t)E_n\varphi_n(t) \\ = \sum_n a_n(t)E_n\varphi_n(t) + \lambda\sum_n a_n(t)V(t)\varphi_n(t) \,. \end{aligned}$$

By left multiplying by $\varphi_k(t)$, for the coefficients $a_k(t)$ we find the system of equations

$$\begin{aligned} i\hbar\dot{a}_k(t) &= \lambda\sum_n a_n(t)(\varphi_k(t), V(t)\varphi_n(t)) \\ &\equiv \lambda\sum_n V_{kn}^I(t)a_n(t) \,, \end{aligned} \tag{152}$$

where we have defined

$$V_{kn}^I(t) = (\varphi_k(t), V(t)\varphi_n(t)) \,. \tag{153}$$

The matrix elements $V_{kn}^I(t)$ are the matrix elements of an operator $V^I(t)$ between the time independent vectors ζ_k, ζ_n. Indeed, since

$$\varphi_n(t) = \zeta_n \exp(-iE_n t/\hbar) = \exp(-iH_0 t/\hbar)\zeta_n \,, \tag{154}$$

Eq. (153) can be written as

$$\begin{aligned} V_{kn}^I(t) &= \left(\exp\left(-iH_0 t/\hbar\right)\zeta_k \,, V(t)\exp\left(-iH_0 t/\hbar\right)\zeta_n\right) \\ &\equiv \left(\zeta_k, V^I(t)\zeta_n\right) \,, \end{aligned} \tag{155}$$

where the operator $V^I(t)$ is defined as follows:

$$V^I(t) \equiv \exp(iH_0 t/\hbar)V(t)\exp(-iH_0 t/\hbar) \,. \tag{156}$$

System Eq. (152) must be supplemented with the appropriate initial conditions, which depend on the particular problem. The commonest application of time dependent perturbation theory is the calculation of transition

probabilities between eigenstates of H_0. Thus, we assume that at $t = 0$ the system is in an eigenstate of the unperturbed Hamiltonian H_0, say the state φ_1. In this case $a_1(0) = 1$, $a_n(0) = 0$ if $n \neq 1$.

In the spirit of perturbation theory, each a_n is expanded into powers of λ

$$a_1(t) = 1 + \sum_{r=1} \lambda^r a_1^{(r)}(t) \,, \qquad a_n(t) = \sum_{r=1} \lambda^r a_n^{(r)}(t) \quad n \neq 1 \,, \tag{157}$$

and terms of equal order are equated. For $a_k^{(r)}$ we find

$$i\hbar \dot{a}_k^{(r)} = \sum_n V_{kn}^I a_n^{(r-1)} \,, \quad r > 0 \,. \tag{158}$$

By Eq. (157), for any $r > 0$, $a_n^{(r)}(0) = 0$. As a consequence, for $r = 1$ we have

$$\begin{aligned} a_k^{(1)} &= \frac{-i}{\hbar} \int_0^t V_{k1}^I(t_1) \mathrm{d}t_1 \\ &= \frac{-i}{\hbar} \int_0^t (\zeta_k, V(t_1)\zeta_1) \exp\left(i\Delta E_{k1} t_1/\hbar\right) \mathrm{d}t_1 \,, \end{aligned} \tag{159}$$

where $\Delta E_{k1} \equiv E_k - E_1$. For $r = 2$ we find

$$\begin{aligned} i\hbar \dot{a}_k^{(2)} &= \sum_n V_{kn}^I(t) a_n^{(1)}(t) \\ &= \frac{-i}{\hbar} \sum_n V_{kn}^I(t) \int_0^t V_{n1}^I(t_1) \mathrm{d}t_1 \end{aligned}$$

whose solution is

$$\begin{aligned} a_k^{(2)}(t) &= \left(\frac{-i}{\hbar}\right)^2 \int_0^t \mathrm{d}t_2 \int_0^{t_2} \mathrm{d}t_1 \sum_n V_{kn}^I(t_2) V_{n1}^I(t_1) \\ &= \left(\frac{-i}{\hbar}\right)^2 \int_0^t \mathrm{d}t_2 \int_0^{t_2} \mathrm{d}t_1 \sum_n (\zeta_k, V(t_2)\zeta_n) \\ &\quad \exp(i\Delta E_{kn} t_2/\hbar)(\zeta_n, V(t_1)\zeta_1) \exp(i\Delta E_{n1} t_1/\hbar) \,. \end{aligned} \tag{160}$$

It is clear how the calculation proceeds for higher values of r. The general expression is

$$\begin{aligned} a_k^{(r)}(t) &= \left(\frac{-i}{\hbar}\right)^r \int_0^t \mathrm{d}t_r \int_0^{t_r} \mathrm{d}t_{r-1} \cdots \int_0^{t_3} \mathrm{d}t_2 \\ &\quad \int_0^{t_2} \mathrm{d}t_1 \sum V_{kn_r}^I(t_r) V_{n_r n_{r-1}}^I(t_{r-1}) \cdots \\ &\quad V_{n_3 n_2}^I(t_2) V_{n_2 1}^I(t_1) \,. \end{aligned} \tag{161}$$

By the completeness of the vectors ζ_n, the sums over the intermediate states can be substituted by the identity and the expression of $a_k^{(r)}$ is simplified into

$$\begin{aligned} a_k^{(r)}(t) &= \left(\frac{-i}{\hbar}\right)^r \int_0^t \mathrm{d}t_r \int_0^{t_r} \mathrm{d}t_{r-1} \cdots \int_0^{t_3} \mathrm{d}t_2 \\ &\quad \int_0^{t_2} \mathrm{d}t_1 \left(\zeta_k, V^I(t_r) V^I(t_{r-1}) \cdots V^I(t_2) V^I(t_1) \zeta_1\right) \,. \end{aligned} \tag{162}$$

It is customary to write Eq. (162) in a different way. The r-dimensional cube $0 \leq t_i \leq t$, $1 \leq i \leq r$, can be split into $r!$ subdomains

$$0 \leq t_{p_1} \leq t_{p_2} \leq \cdots \leq t_{p_{r-1}} \leq t_{p_r} \leq t \,, \tag{163}$$

with $\{p_1, p_2, \ldots, p_{r-1}, p_r\}$ a permutation of $\{1, 2, \ldots, r-1, r\}$. The time ordered product of r (non-commuting) operators $V^I(t_{p_1}), V^I(t_{p_2}), \ldots, V^I(t_{p_r})$ is introduced according to the definition

$$T\left[V^I(t_{p_1}) \cdots V^I(t_{p_r})\right] \equiv V^I(t_r) \cdots V^I(t_1) \,, \qquad t_1 \leq t_2 \leq \cdots \leq t_r \,. \tag{164}$$

If $(-i\lambda/\hbar)^r (\zeta_k, T\left[V^I(t_{p_1}) \cdots V^I(t_{p_r})\right] \zeta_1)$ is integrated over the r-cube, then each of the $r!$ subdomains defined by Eq. (163) yields the same contribution. As a consequence Eq. (161) can be written as

$$\begin{aligned} a_k^{(r)}(t) &= \frac{1}{r!} \left(\frac{-i}{\hbar}\right)^r \int_0^t \mathrm{d}t_r \int_0^t \mathrm{d}t_{r-1} \cdots \int_0^t \mathrm{d}t_2 \\ &\quad \int_0^t \mathrm{d}t_1 \left(\zeta_k, T\left[V^I(t_r) V^I(t_{r-1}) \cdots V^I(t_2) V^I(t_1)\right] \zeta_1\right) \,. \end{aligned} \tag{165}$$

The amplitudes $a_k(t)$ can then be written as

$$a_k(t) = \left(\zeta_k, T\left[\exp\left(\frac{-i\lambda}{\hbar}\right) \int_0^t V^I(t') \mathrm{d}t'\right] \zeta_1\right) \tag{166}$$

with obvious significance of the T-exponential: each monomial in the V^I operators which appear in the expansion of the exponential is to be time ordered according to the T-prescription. If the initial state is given at time t_0 the integral appearing in the T-exponential should start at t_0. We define

$$U^I(t, t_0) \equiv T\left[\exp\left(\frac{-i\lambda}{\hbar}\right) \int_{t_0}^t V^I(t') \mathrm{d}t'\right] \,. \tag{167}$$

The expansion of the T-exponential into monomials in the V^I operators is called the Dyson series [13,14]. It is extensively employed in perturbative quantum field theory.

From Eqs. (166) and (167) it is easy to derive an expression for the time evolution operator $U(t, t_0)$ such that

$$U(t, t_0)\psi(t_0) = \psi(t) \,. \tag{168}$$

Indeed, $\psi(t)$ and $\psi(t_0)$ can be expanded in the basis of the vectors $\varphi_n(t)$ and $\varphi_n(t_0)$ respectively as in Eq. (151). By the linearity of the Schrödinger equation it suffices to determine $(\varphi_k(t), U(t, t_0)\varphi_n(t_0))$, which we already know to be $(\zeta_k, U^I(t, t_0)\zeta_n)$. We have

$$\begin{aligned}
&(\varphi_k(t), U(t, t_0)\varphi_n(t_0)) \\
&\quad = (\exp(-iH_0 t/\hbar)\zeta_k, U(t, t_0)\exp(-iH_0 t_0/\hbar)\zeta_n) \\
&\quad = (\zeta_k, \exp(iH_0 t/\hbar)U(t, t_0)\exp(-iH_0 t_0/\hbar)\zeta_n) \,.
\end{aligned}$$

As a consequence we find the equation

$$\begin{aligned}
&U(t, t_0) \\
&= \exp(-iH_0 t/\hbar)\, T\left[\exp\left(\frac{-i\lambda}{\hbar}\right)\int_{t_0}^{t} V^I(t')\,\mathrm{d}t'\right] \\
&\qquad \exp(iH_0 t_0/\hbar) \,,
\end{aligned} \tag{169}$$

which provides the perturbation expansion of the evolution operator $U(t, t_0)$ in powers of λ.

It can be proved that if the operator function $V(t)$ is strongly continuous and the operators $V(t)$ are bounded, then the expansion which defines the T-exponential is norm convergent to a unitary operator, as expected [35]. The restriction to bounded operators $V(t)$ does not detract from the range of applications of Eqs. (166) and (169), since time dependent perturbation theory is almost exclusively used for treating interactions of a system with external fields, which generate bounded interactions.

Future Directions

The long and honorable service of perturbation theory in every sector of quantum mechanics must be properly acknowledged. Its future is perhaps already in our past: the main achievement is its application to quantum field theory where, just to quote an example, the agreement between the measured value and the theoretical prediction of the electron magnetic moment anomaly to ten significant digits has no rivals.

Despite its successes, still perturbation theory is confronted with fundamental questions. In most of realistic problems it is unknown whether the perturbation series is convergent or at least asymptotic. In non-relativistic quantum mechanics this does not represent a practical problem since only a limited number of terms can be calculated, but in quantum field theory, where higher order terms are in principle calculable, this calls for dedicate investigations. There, in particular, conditions for recovering the exact amplitudes from the first terms of the series by such techniques as the Padé approximants or the self similar approximants, and the estimate on the bound of the error, deserve further investigation.

Somewhat paradoxically, it can be said that the future of perturbation theory is in the non-perturbative results (analyticity domains, large coupling constant behavior, tunneling effect ...) – an issue where much work has already been done – since they have proved to be complementary to the use of perturbation theory.

Bibliography

Primary Literature

1. Baker GA, Graves-Morris P (1996) Padé approximants. Cambridge Univ. Press, Cambridge
2. Bargmann V (1964) Note on Wigner's theorem on symmetry operations. J Math Phys 5:862–868
3. Bender CM, Wu TT (1969) Anharmonic oscillator. Phys Rev 184:1231–1260
4. Bloch C (1958) Sur la théorie des perturbations des états liées. Nucl Phys 6:329–347
5. Böhm A (1993) Quantum mechanics, foundations and applications. Springer, New York, pp 208–215
6. Borel E (1899) Mémoires sur le séries divergentes. Ann Sci École Norm Sup 16:9–136
7. Born M, Heisenberg W, Jordan P (1926) Zur Quantenmechanik, II. Z Phys 35:557–615
8. Born M (1926) Quantenmechanik der Stossvorgänge. Z Phys 38:803–827
9. Brillouin L (1932) Perturbation problem and self consistent field. J Phys Radium 3:373–389
10. Courant R, Hilbert D (1989) Methods of mathematical physics, vol I. Wiley, New York, pp 343–350
11. Dirac PAM (1926) On the theory of quantum mechanics. Proc Roy Soc A112:661–677
12. Dyson FJ (1952) Divergence of perturbation theory in quantum electrodynamics. Phys Rev 85:631–632
13. Dyson FJ (1949) The radiation theories of Tomonaga, Schwinger and Feynman. Phys Rev 75:486–502
14. Dyson FJ (1949) The S-matrix in quantum electrodynamics. Phys Rev 75:1736–1755
15. Epstein ST (1954) Note on perturbation theory. Amer J Phys 22:613–614
16. Epstein ST (1968) Uniqueness of the energy in perturbation theory. Amer J Phys 36:165–166
17. Feynman RP (1939) Forces in molecules. Phys Rev 56:340–343
18. Fock VA (1935) Zur Theorie des Wasserstoffatoms. Z Phys 98:145–154
19. Graffi S, Grecchi V, Simon B (1970) Borel summability: Application to the anharmonic oscillator. Phys Lett B32:631–634
20. Grossman A (1961) Schrödinger scattering amplitude I. J Math Phys 3:710–713
21. Hamermesh M (1989) Group theory and its application to physical problems. Dover, New York, pp 32–114

22. Hannabuss K (1997) Introduction to quantum theory. Clarendon, Oxford, pp 131–136
23. Hellmann H (1937) Einführung in die Quantenchemie. Deuticke, Leipzig
24. Joachain J (1983) Quantum collision theory. North–Holland, Amsterdam
25. Kato T (1966) Perturbation theory for linear operators. Springer, New York
26. Kato T (1949) On the convergence of the perturbation method I. Progr Theor Phys 4:514–523
27. Kramers HA (1957) Quantum mechanics. North-Holland, Amsterdam, pp 198–202
28. Krieger JB (1968) Asymptotic properties of perturbation theory. J Math Phys 9:432–435
29. Landau LD, Lifshitz EM (1960) Mechanics. Pergamon Press, Oxford, p 129
30. Lippmann BA, Schwinger J (1950) Variational principles for scattering processes I. Phys Rev 79:469–480
31. Loeffel J, Martin A, Wightman A, Simon B (1969) Padé approximants and the anharmonic oscillator. Phys Lett B 30:656–658
32. Padé H (1899) Sur la représentation approchée d'une fonction pour des fonctions rationelles. Ann Sci Éco Norm Sup Suppl 9(3):1–93
33. Rayleigh JW (1894–1896) The theory of sound, vol I. Macmillan, London, pp 115–118
35. Reed M, Simon B (1975) Methods of modern mathematical physics, vol II. Academic Press, New York, pp 282–283
34. Reed M, Simon B (1978) Methods of modern mathematical physics, vol IV. Academic Press, New York, pp 10–44
36. Rellich F (1937) Störungstheorie der Spektralzerlegung I. Math Ann 113:600–619
37. Rellich F (1937) Störungstheorie der Spektralzerlegung II. Math Ann 113:677–685
38. Rellich F (1939) Störungstheorie der Spektralzerlegung III. Math Ann 116:555–570
39. Rellich F (1940) Störungstheorie der Spektralzerlegung IV. Math Ann 117:356–382
40. Riesz F, Sz.-Nagy B (1968) Leçons d'analyse fonctionelle. Gauthier-Villars, Paris, pp 143–188
41. Rollnik H (1956) Streumaxima und gebundene Zustände. Z Phys 145:639–653
42. Sakurai JJ (1967) Advanced quantum mechanics. Addison-Wesley, Reading, pp 39–40
43. Scadron M, Weinberg S, Wright J (1964) Functional analysis and scattering theory. Phys Rev 135:B202-B207
44. Schrödinger E (1926) Quantisierung als Eigenwertproblem. Ann Phys 80:437–490
45. Schwartz J (1960) Some non-self-adjoint operators. Comm Pure Appl Math 13:609–639
46. Simon B (1970) Coupling constant analyticity for the anharmonic oscillator. Ann Phys 58:76–136
47. Simon B (1991) Fifty years of eigenvalue perturbation theory. Bull Am Math Soc 24:303–319
48. Thirring W (2002) Quantum mathematical physics. Springer, Berlin, p 177
49. von Neumann J, Wigner E (1929) Über das Verhalten von Eigenwerten bei adiabatischen Prozessen. Phys Z 30:467–470
50. Watson G (1912) A theory of asymptotic series. Philos Trans Roy Soc Lon Ser A 211:279–313
51. Weyl H (1931) The theory of groups and quantum mechanics. Dover, New York
52. Wigner EP (1935) On a modification of the Rayleigh–Schrödinger perturbation theory. Math Natur Anz (Budapest) 53:477–482
53. Wigner EP (1959) Group theory and its application to the quantum mechanics of atomic spectra. Academic Press, New York
54. Yosida K (1991) Lectures on differential and integral equations. Dover, New York, pp 115–131
55. Yukalov VI, Yukalova EP (2007) Methods of self similar factor approximants. Phys Lett A 368:341–347
56. Zemach C, Klein A (1958) The Born expansion in non-relativistic quantum theory I. Nuovo Cimento 10:1078–1087

Books and Reviews

Hirschfelder JO, Byers Brown W, Epstein ST (1964) Recent developments in perturbation theory. In: Advances in quantum chemistry, vol 1. Academic Press, New York, pp 255–374

Killingbeck J (1977) Quantum-mechanical perturbation theory. Rep Progr Phys 40:963–1031

Mayer I (2003) Simple theorems, proofs and derivations in quantum chemistry. Kluwer Academic/Plenum Publishers, New York, pp 69–120

Morse PM, Feshbach H (1953) Methods of theoretical physics, part 2. McGraw-Hill, New York, pp 1001–1106

Wilcox CH (1966) Perturbation theory and its applications in quantum mechanics. Wiley, New York

Perturbation Theory, Semiclassical

Andrea Sacchetti
Dipartimento di Matematica Pura ed Applicata, Universitá di Modena e Reggio Emilia, Modena, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
The WKB Approximation
Semiclassical Approximation in Any Dimension
Propagation of Quantum Observables
Future Directions
Bibliography

Glossary

Agmon metric In the classically forbidden region where the potential energy $V(x)$ is larger than the total energy E, i. e. $V(x) > E$, we introduce a notion of distance based on the *Agmon metric* defined as $[V(x) - E]\mathrm{d}x^2$, where $\mathrm{d}x^2$ is the usual Riemann metric. We emphasize that such an Agmon metric is the "semiclassical" equivalent of the "classical" Jacobi met-

ric $[E - V(x)]\mathrm{d}x^2$ introduced in classical mechanics in the classically permitted region where $V(x) < E$.

Asymptotic series The notion of *asymptotic series* goes back to Poincaré. We say that a formal power series $\sum_r a_r z^{-r}$ is taken to be the *asymptotic power series* for a function $f(z)$, as $|z| \to \infty$ in a given infinite sector $S = \{z \in \mathbb{C} : \alpha \leq \arg z \leq \beta\}$, if for any fixed N the remainder term

$$R_N(z) = f(z) - \sum_{r=0}^{N-1} a_r z^{-r}$$

is such that $R_N(z) = \mathcal{O}(z^{-N})$, that is

$$|R_N(z)| \leq C_N \left|z^{-N}\right| \ , \quad \forall z \in S \ ,$$

for some positive constant C_N depending on N.

Classically allowed and forbidden regions – turning points We distinguish two different regions: the region $V(x) < E$ where the *classical motion is allowed* and the region $V(x) > E$ where the *classical motion is forbidden*. The points that separate these two regions, that is such that $V(x) = E$, will play a special role and they are named *turning points*.

Semiclassical limit Cornerstone of Quantum Mechanics is the time-dependent Schrödinger equation

$$i\hbar \frac{\partial \psi}{\partial t} = -\frac{\hbar^2}{2m} \Delta \psi + V(x)\psi \ , \tag{1}$$

where $V(x)$ is assumed to be a real-valued function, usually it represents the potential energy, and m is the mass of the particle. The solution of this equation defines the density of probability $|\psi(x,t)|^2$ to find the particle in some region of space. The parameter $\hbar$ in Eq. (1) is related to Planck's constant h

$$\hbar = \frac{h}{2\pi} = 1.0545 \times 10^{-27} \text{erg sec}$$
$$= 6.5819 \times 10^{-22} \ \text{MeV sec} \ .$$

According to the correspondence principle, when Planck's constant can be considered small with respect to the other parameters, such as masses and distances, then quantum theory approaches classical Newton theory. Thus, roughly speaking, we expect that classical mechanics is contained in quantum mechanics as a limiting form (i. e. $\hbar \to 0$). The limit of small $\hbar$, when compared with the other parameters, is the so-called *semiclassical limit*. We should emphasize that making Planck's constant small in Eq. (1) is a rather singular limit and many difficult mathematical problems occur.

Stokes lines There is some misunderstanding in the literature concerning the name of the curves $\Im\left[\xi(z)\right] = 0$, where $\Im(\xi)$ denotes the imaginary part of

$$\xi(z) = \frac{i}{\hbar} \int_{x_E}^{z} \sqrt{E - V(q)} \mathrm{d}q \ , \quad z \in \mathbb{C} \ ,$$

and $x_E \in \mathbb{R}$ is a turning point (i. e. $V(x_E) = E$), the potential V is assumed to be an analytic function in the complex plane. As usually physicists do, and in agreement with Stokes' original treatment, we adopt here the convention to name as *Stokes line* any path coming from the turning point x_E such that $\Im[\xi(z)] = 0$; reserving the name of *anti-Stokes lines*, or *regular lines*, for the curves such that $\Re[\xi(z)] = 0$, where $\Re(\xi)$ denotes the real part of ξ (we should emphasize that most mathematicians adopt the opposite rule, that is they call Stokes lines the paths such that $\Re[\xi(z)] = 0$).

Tunnel effect At a real (simple) turning point x_E where $E = V(x)$ (and $V'(x) \neq 0$), classical particles incident from an accessible region [where $E > V(x)$] reverse their velocity and return back; the adjacent region [where $E < V(x)$] is forbidden to these particles according to classical mechanics. In fact, quantum mechanically exponentially growing or damped waves can exist in forbidden regions and a quantum particle can pass through a potential barrier. This is the so-called *tunnel effect*.

WKB The WKB method, named after the contributions independently given by Wentzel, Kramers and Brillouin, consists of connecting the approximate solutions of the time independent Schrödinger equation across turning points.

Definition of the Subject

Several kinds of perturbation methods are commonly used in quantum mechanics ► Perturbation Theory and Molecular Dynamics, ► Perturbation Theory in Quantum Mechanics, ► Perturbation Theory. This chapter deals with *semiclassical approximations* where expressions for energy levels and wave functions are obtained in the limiting case of small values of Planck's constant.

We emphasize that wave functions are highly singular as the parameter $\hbar$ goes to zero. In fact the semiclassical limit is a singular perturbation problem; namely, Eq. (1) suffers a reduction of order setting $\hbar$ equal to zero and the resulting equation is not a differential equation and does not give the classical limit correctly. Therefore, ordinary perturbation methods, which usually give energy lev-

els and wave-functions as convergent power series of the small parameter, cannot be applied.

The main goal of *semiclassical methods* consists of obtaining *asymptotic series*, in the limit of small $\hbar$, for the quantum-mechanical quantities. For instance, semiclassical methods permit us to solve the eigenvalues problem

$$-\frac{\hbar^2}{2m}\Delta\psi + V(x)\psi = E\psi\,, \quad x \in \mathbb{R}^n\,, \quad n \geq 1\,, \tag{2}$$

where the energy levels E and the wave-functions $\psi(x)$ are given by means of asymptotic series in the limit of vanishing $\hbar$, or to obtain a direct link between the time evolution of a quantum observable and the Hamiltonian flux of the associated classical quantity.

Introduction

The first contributions to this theory, in the framework of Quantum Mechanics, go back to Wentzel (1926), Kramers (1926) and Brillouin (1926). In their papers they independently developed the determination of connection formulas linking exponential and oscillatory approximations of the solution of Eq. (2) in dimension $n = 1$ across a turning point. This method, explained in Sect. The WKB Approximation, is usually called WKB approximation, or also JWKB method to acknowledge that the approximate connection formula was previously discovered by Jeffrey. Actually, oscillatory and exponential approximation formulas, obtained far from turning points, were independently used by Green (1837) and Liouville (1837) and they can be already found in an investigation on the motion of a planet in an unperturbed elliptic orbit by the Italian astronomer Carlini (1817). For historical notes on the WKB-method we refer to [7,11,23].

The WKB method can be also applied to three-dimensional problems only under some particular circumstances; for instance when the potential is spherically symmetric and the radial differential equation can be separated.

In general WKB approximation is not suitable for problems in dimension n higher than 1. Actually, semiclassical methods in higher dimension require new sophisticated tools such as the *Agmon metric*, *microlocal calculus* and *$\hbar$-pseudodifferential operators*. These tools have been developed by Agmon, Hörmander and Maslov in the 1970s and since then a large number of mathematicians have contributed to this subject. In Sects. "Semiclassical Approximation in Any Dimension" and "Propagation of Quantum Observables" we briefly introduce the reader to the basic concepts of these theories and we resume the most important results such as the exponential decay of wave-functions and the Egorov Theorem, which asymptotically describes the quantum evolution of an observable by means of the classical evolution of its classical counterpart.

Actually, these methods may be applied to many other fields, where $\hbar$ is a small quantity not related to the Planck constant but it may represent a different small physical quantity such as the adiabatic parameter in adiabatic problems, the (square root of the) inverse of the heavy mass in the Born–Oppenheimer approximation, etc.

Notation

Hereafter, for the sake of definiteness, let us fix $2m = 1$.

The WKB Approximation

If the potential $V(x)$ does not have a very simple form then the solution of the time-independent Schrödinger equation even in one dimension

$$-\hbar^2\psi'' + V(x)\psi = E\psi\,, \quad x \in (x_1, x_2)\,, \tag{3}$$

is a quite complicated problem which requires the use of a sort of approximation method (hereafter $' = \mathrm{d}/\mathrm{d}x$ denotes the derivative with respect x); here (x_1, x_2) is a given finite or infinite one-dimensional interval.

We distinguish different regions: the classically allowed regions where $V(x) < E$ and the classically forbidden regions where $V(x) > E$. Approximation formulas for the solution of Eq. (3) in these separate regions have been studied since Carlini, Green, Liouville and Jacobi. The problem to connect these approximated solutions across the turning points was raised in the framework of Quantum Mechanics and it was independently solved by Jensen, Wentzel, Kramers and Brillouin. For the general treatment of semiclassical approximation in dimension one and for physical applications we refer to [2,3,10,21,24,25,33].

Semiclassical Solutions in the Classically Allowed Region

The basic idea is quite simple: if $V = \text{const.}$ is a constant smaller than E then Eq. (3) has solutions $\mathrm{e}^{\pm ikx/\hbar}$ for a suitable real-valued constant k. This fact suggests to us that if the potential, while no longer constant, varies only slowly with x and it is such that $V(x) < E$ for any $x \in (x_1, x_2)$, we might try a solution of the form

$$\psi(x) = a(x)\mathrm{e}^{iS(x)/\hbar}\,, \tag{4}$$

except that the *amplitude* $a(x)$ is not constant and the *phase* $S(x)$ is not simply proportional to x. Substituting (4)

into (3) and separating the real and imaginary parts of the coefficients of $e^{iS(x)/\hbar}$, then we get a system of equations for the x-dependent real-valued phase $S(x)$ and amplitude $a(x)$:

$$\begin{cases} u^2 - p^2(x) = \hbar^2 a''/a \\ (a^2 u)' = 0 \end{cases}$$

where we set

$$u(x) = S'(x) \quad \text{and} \quad p(x) = \sqrt{E - V(x)}\,.$$

Hence, $a(x) = C[u(x)]^{-1/2}$ for some constant C.

Since the complex conjugation $a(x)e^{-iS(x)/\hbar}$ of (4) is a solution of the same equation and the Wronskian between these two solutions is not zero then the general solution of Eq. (3) can be written as

$$\begin{aligned} \psi(x) &= b_+\psi_+(x) + b_-\psi_-(x)\,, \\ \psi_\pm(x) &= \frac{1}{\sqrt{u(x)}} e^{\pm i/\hbar \int u(x)\mathrm{d}x}\,. \end{aligned} \tag{5}$$

Here, $b_\pm$ are arbitrary constants and $u = u(x;\hbar)$ is a real-valued solution, depending on the variable x and on the semiclassical parameter $\hbar$, of the nonlinear ordinary differential equation

$$\mathcal{F}(u) = u^2 - p^2 - \hbar^2 u f(u) = 0 \tag{6}$$

where

$$f(u) = u^{-1/2}\left(u^{-1/2}\right)''\,.$$

The fact that (6) is a nonlinear equation, whereas the Schrödinger equation (3) is linear, would be usually regarded as a drawback, but we shall take advantage of the nonlinearity to develop a simple approximation method for solving (6).

Taking the limit of $\hbar$ small in (6) then it turns out that the leading order of u is simply given by p. Actually, it is possible to prove that (3) has twice continuously differentiable solutions $\psi_\pm$ satisfying the following asymptotic behavior

$$\psi_\pm(x) = \frac{1}{\sqrt{p(x)}} e^{\pm i/\hbar \int_a^x p(q)\mathrm{d}q}\left[1 + \delta_\pm(x)\right]\,, \qquad x \in (x_1, x_2)\,, \tag{7}$$

where a is an arbitrary point in (x_1, x_2) and where

$$|\delta_\pm(x)| \le e^{\frac{1}{2}\hbar\left|\int_a^x |f[p(q)]|\mathrm{d}q\right|} - 1 = \mathcal{O}(\hbar)\,.$$

Henceforth, the dominant term of the solution (5) has an oscillating behavior of the form

$$\psi(x) = b_+ \frac{e^{i/\hbar \int_a^x \sqrt{E-V(q)}\mathrm{d}q}}{\sqrt[4]{E - V(x)}} + b_- \frac{e^{-i/\hbar \int_a^x \sqrt{E-V(q)}\mathrm{d}q}}{\sqrt[4]{E - V(x)}} + \mathcal{O}(\hbar)\,.$$

In order to compute also the other terms of the asymptotic expansion of the solutions $\psi_\pm(x)$ we formally solve the nonlinear Eq. (6) by means of the formal power series in $\hbar^2$:

$$u = u(x,\hbar) = \sum_{n=0}^{\infty} \hbar^{2n} u_{2n}(x)\,. \tag{8}$$

The functions $u_{2n}(x)$ are determined explicitly, order by order, by formally substituting (8) into (6) and requiring that the coefficients of the same terms $\hbar^{2n}$ are zero for any $n \ge 0$. In such a way and assuming that the potential $V(x)$ admits derivatives at any order then we obtain that

$$\begin{aligned} u_0(x) &= p(x) \\ u_2(x) &= -\frac{1}{4}\frac{p''(x)}{p^2(x)} + \frac{3}{8}\frac{[p'(x)]^2}{p^3(x)} \\ u_4(x) &= \frac{1}{16}\frac{p''''(x)}{p^4(x)} - \frac{5}{8}\frac{p'''(x)p'(x)}{p^5(x)} - \frac{13}{32}\frac{\left[p''(x)\right]^2}{p^5(x)} \\ &\quad + \frac{99}{32}\frac{p''(x)\left[p'(x)\right]^2}{p^6(x)} - \frac{297}{128}\frac{\left[p'(x)\right]^4}{p^7(x)}\,, \end{aligned}$$

and so on. Thus, the asymptotic behavior (7) can be improved up to any order. In particular, let

$$\psi_{N,\pm}(x) = \frac{1}{\sqrt{p_N(x)}} e^{\pm i/\hbar \int_a^x p_N(q)\,\mathrm{d}q}\,, \qquad x \in (x_1, x_2)\,,$$

where $a \in (x_1, x_2)$ is arbitrary and fixed and $p_N(x) = \sum_{n=0}^{N} u_{2n}(x)\hbar^{2n}$, in particular $p_0(x) = p(x)$, let us introduce the *error-control function*

$$\epsilon_N(x) = \frac{1}{\hbar p_N(x)}\mathcal{F}\left[p_N(x)\right] = \mathcal{O}(\hbar^{2N+1})$$

where $\epsilon_0(x) = \hbar f[p(x)]$; then the two functions $\psi_{N,\pm}(x)$ approximate the solutions $\psi_\pm(x)$ of Eq. (3) up to the order $2N+1$, that is

$$\begin{aligned} |\psi_\pm(x) - \psi_{N,\pm}(x)| &\le \left[\exp\left(\frac{1}{2}\left|\int_a^x |\epsilon_N(q)|\,\mathrm{d}q\right|\right) - 1\right] \\ &= \mathcal{O}(\hbar^{2N+1})\,, \quad \forall x \in (x_1, x_2)\,. \end{aligned}$$

We underline that this result is valid whether or not x_1 and x_2 are finite, also whether or not $V(x)$ is bounded: it suffices that the *error-control function* is absolutely integrable: $\epsilon_N \in L^1(x_1, x_2)$.

Semiclassical Solutions in the Classically Forbidden Region

The previous asymptotic computation of the solutions $\psi_\pm$ applies also in the classically forbidden region where $V(x) > E$, provided that $p(x)$ is one of the two purely imaginary determination of $\sqrt{E - V(x)}$. By assuming that $\Im p(x) < 0$ then Eq. (3) has twice continuously differentiable solutions of the form

$$\psi_\pm(x) = \frac{1}{\sqrt{|p(x)|}} \mathrm{e}^{\pm 1/\hbar \int_a^x |p(q)| \mathrm{d}q} \left[1 + \delta_\pm(x)\right] , \qquad x \in (x_1, x_2) , \quad (9)$$

where a is an arbitrary point in (x_1, x_2) and where the remainder terms $\delta_\pm$ are bounded as follows

$$|\delta_\pm(x)| \le \mathrm{e}^{\frac{1}{2}\hbar \left| \int_{a_\pm}^x |f[p(q)]| \mathrm{d}q \right|} - 1 = \mathcal{O}(\hbar)$$

where $a_+ = x_1$ and $a_- = x_2$. In particular, if $V(x) > E$ for any $x \in (-\infty, x_2)$, that is $x_1 = -\infty$ (and similarly we can consider the case where $x_2 = +\infty$), and the error-control function $f[p(x)] \in L^1(-\infty, x_2)$ then the above asymptotic estimate holds for any $x \in (-\infty, x_2)$. We also emphasize that the asymptotic behavior (9) could be also improved up to any order, as done for the approximated solutions in the classically allowed regions.

We underline that the two solutions $\psi_\pm$ play here a different role. Indeed, solution $\psi_-(x)$ is an exponentially decreasing function as $x > a$ grows, while $\psi_+(x)$ is exponentially increasing. The first solution is usually called *recessive* solution and it is *uniquely determined* by the asymptotic power expansion as $\hbar \to 0$. The other solution is called *dominant* solution and it *is not uniquely determined* by the asymptotic expansion; in fact, $\psi_+(x) + c\psi_-(x)$, $x > a$, is still a solution of Eq. (3) which has the same asymptotic behavior of $\psi_+(x)$ for any $c \in \mathbb{C}$.

Connection Formula

It is not possible to use the semiclassical solutions (7) and (9) at turning points x_E since the term $1/p$ diverges and there is no guarantee that the same combination of simple semiclassical solutions will fit the same particular solution on both sides of the turning point. This is the so-called *connection problem*: if the values $b_\pm$ of the semiclassical solution are known in a given region then the problem consists of computing the values of $b_\pm$ in a different region separated from the first region by one (or more) turning points.

The main two methods used to treat this problem are the following ones:

- The *complex method*, where the two regions surrounding the turning point are joined by a path in the complex plane which is sufficiently far from the turning points. In such a case the semiclassical solutions in different regions are connected by means of holomorphic extension arguments. Thus, in order to apply this method we have to assume that the potential $V(x)$ can be holomorphically extended to the complex plane.
- The second method employs the technique of the *uniform approximation*, where the required solution is mapped on the solution of a simpler and suitable equation which has the same disposition of turning points as the original one. As well as enabling the connection problem to be solved, this method also provides the wave function in the neighborhood of the turning points, which was bypassed in the complex method.

We are going now to explain these methods in detail.

The Complex Method By assuming that $V(x)$ is the restriction on the real axis of a holomorphic function $V(z)$, $z \in \mathbb{C}$, we turn now to the approximate solution of the differential equation

$$-\hbar^2 \frac{\mathrm{d}^2 \psi(z)}{\mathrm{d}z^2} = \left[E - V(z)\right] \psi(z)$$

in a complex domain $\mathcal{D}$ in which $V(z)$ is holomorphic and $p^2(z) = E - V(z)$ does not vanish. Then, this equation has two holomorphic solutions

$$\psi_\pm(z) = \frac{1}{\sqrt{p(z)}} \mathrm{e}^{\pm i/\hbar \int_a^z p(q) \mathrm{d}q} \left[1 + \mathcal{O}(\hbar)\right] , \quad z \in \mathcal{D} , \quad (10)$$

where a is an arbitrary point; in particular, for our purposes it is convenient to choose it coinciding with the turning point, i. e.: $a = x_E$. Actually, this asymptotic behavior holds under a technical assumption: given a reference point a_+ (respectively a_-) then (10) is true for $\psi_+(z)$ (resp. $\psi_-(z)$) for any z such that there exists a *progressive* (resp. *regressive*) path Γ_+ (resp. Γ_-) contained in $\mathcal{D}$ and connecting z with a_+ (resp. a_-); that is as q passes along Γ_+ (resp. Γ_-) from a_+ (resp. a_-) to z then $\Re[\xi(z)]$ is non-

decreasing (resp. nonincreasing) where

$$\xi(z) = \frac{i}{\hbar}\int_{x_E}^{z} p(q)\mathrm{d}q\,.$$

Now, we denote as a *Stokes line* any path coming from the turning point x_E such that $\Im[\xi(z)] = 0$; we denote as an *anti-Stokes line*, or *principal line*, any path coming from the turning point x_E such that $\Re[\xi(z)] = 0$. Classically forbidden (resp. allowed) regions are particular cases of Stokes (resp. anti-Stokes) lines, which lie on the real axis.

Let us consider now the case, as in Fig. 1, where the turning point x_E is simple (that is $V'(x_E) \neq 0$), $p^2(x) < 0$ for $x < x_E$ and $p^2(x) > 0$ for $x > x_E$. Since the turning point x_E is simple then in a neighborhood of the turning point we have that $p(z) \approx c\,[z - x_E]^{\frac{1}{2}}$, for some $c \in \mathbb{R}^+$, and $\xi(z) \approx \frac{2ic}{3\hbar}[z - x_E]^{\frac{3}{2}}$. Thus, Stokes lines are three different paths coming from the turning point and with asymptotic directions $\arg(z - x_E) = \frac{1}{3}\pi, \pi, \frac{5}{3}\pi$; while the asymptotic directions of the anti-Stokes lines are given by $\arg(z - x_E) = 0, \frac{2}{3}\pi, \frac{4}{3}\pi$.

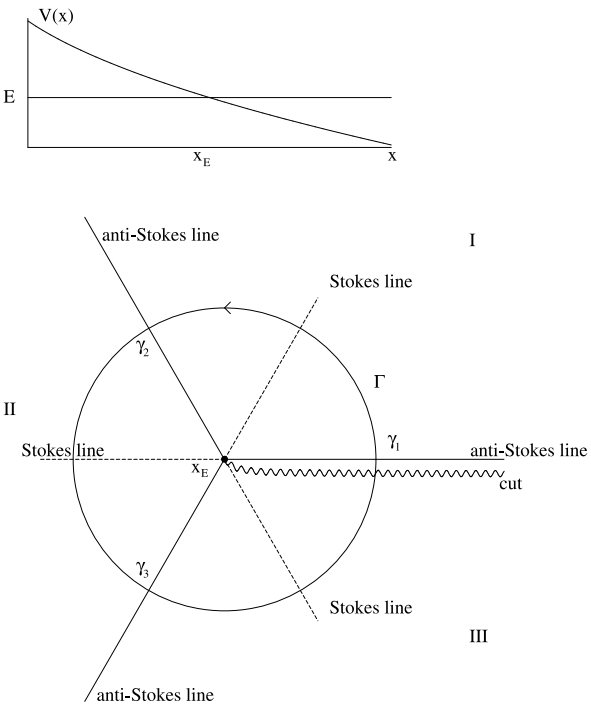

Perturbation Theory, Semiclassical, Figure 1
Stokes and anti-Stokes lines in a neighborhood of a simple turning point x_E where the left-hand side of the turning point is classically forbidden and the right-hand side is classically allowed. The wavy line denotes the *cut* of the multi-valued function $p(z)$

Along the three anti-Stokes lines γ_1, γ_2 and γ_3 the solution ψ can be written as

$$\psi(z) = b_{+,j}\psi_{+,j}(z) + b_{-,j}\psi_{-,j}(z), \quad z \in \gamma_j, \quad j = 1, 2, 3,$$

where $\psi_{\pm,j}$ have the asymptotic behavior given by (10) and where the coefficients $b_{\pm,j}$ are asymptotically fixed along the anti-Stokes lines. Let F_j be the 2×2 matrix which connects the coefficients:

$$\begin{pmatrix} b_{+,j+1} \\ b_{-,j+1} \end{pmatrix} = F_j \begin{pmatrix} b_{+,j} \\ b_{-,j} \end{pmatrix}.$$

In the region I, with boundaries given by the anti-Stokes lines γ_1 and γ_2, the solution ψ_+ is *dominant* and thus the coefficient b_+ cannot change in this region, similarly in the region III with boundaries γ_3 and the cut. In contrast, in the region II, with boundaries given by the anti-Stokes lines γ_2 and γ_3, the solution ψ_- is *dominant* and thus the coefficient b_- cannot change in this region. Then the matrices F_j can be written as

$$F_1 = \begin{pmatrix} 1 & 0 \\ r & 1 \end{pmatrix}, \quad F_2 = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}$$
$$\text{and} \quad F_3 = \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix}$$

where r, s and t have to be determined. To this end let us consider the closed anti-clockwise path Γ surrounding the simple turning point x_E; along this path the exponential term gains a phase $\nu = \frac{1}{\hbar}\oint_\Gamma p(z)\mathrm{d}z$ and the argument of $p(z)$, which appears in (10), decreases of π. Hence, the holomorphic extension of $\psi_\pm$, around this closed path, are thus proportional to the solution $\psi_\mp$:

$$\psi_+ \to i\mathrm{e}^{i\nu}\psi_- \quad \text{and} \quad \psi_- \to i\mathrm{e}^{-i\nu}\psi_+\,.$$

Therefore, the numbers r, s and t are such that

$$F_1\,F_2\,F_3 = \begin{pmatrix} 1 + st & s \\ r + (rs+1)t & rs + 1 \end{pmatrix} \sim \begin{pmatrix} 0 & i\mathrm{e}^{-i\nu} \\ i\mathrm{e}^{i\nu} & 0 \end{pmatrix}$$

from which it follows that

$$s = i\mathrm{e}^{-i\nu}\,, \quad r = t = i\mathrm{e}^{i\nu}\,.$$

In particular, since $\int p(z)\mathrm{d}z$ does not diverge at x_E then we can take Γ in a small neighborhood of x_E and $\nu \sim 0$. Thus, the *Stokes' rule* follows: *the connection between the two coefficients from one Stokes line to the (anti-clockwise) adjacent one follows the following rule*

$$b_{\text{dominant}} \to b_{\text{dominant}} \quad \text{and}$$
$$b_{\text{recessive}} \to b_{\text{recessive}} + ib_{\text{dominant}}\,.$$

The Method of Comparison Equations With this method we obtain a local approximate solution, in a neighborhood of the turning points, in terms of the known solutions of an equivalent equation

$$\frac{\mathrm{d}^2\phi}{\mathrm{d}y^2} + q(y)\phi(y) = 0\,, \tag{11}$$

where $q(y)$ is chosen to be similar in some way to $p^2(x)$, but simpler in order to have an explicit solution. For the sake of definiteness, we restrict here our analysis to the case of a simple turning point x_E.

In order to get an approximate solution to Eq. (3) in a small neighborhood of a simple zero x_E of $p^2(x)$ we introduce the following changes of variable

$$x \to y(x) = \left[\frac{3}{2}\xi(x)\right]^{\frac{2}{3}}\,, \quad \xi(x) = \frac{i}{\hbar}\int_{x_E}^{x} p(q)\mathrm{d}q\,.$$

Equation (3) takes the form of the approximate Eq. (11) where $q(y) \approx -y$ in a neighborhood of the turning point, which admits solutions given by the Airy functions $Ai(y)$ and $Bi(y)$. Then, the approximate solution of the Schrödinger equation (3) in a neighborhood of the turning point has the form

$$\psi(x) \sim \alpha\left[\frac{\pi^2 y(x)}{p^2(x)}\right]^{\frac{1}{4}}\{\cos(\mu)Ai[y(x)] + \sin(\mu)Bi[y(x)]\}$$

where α and β are constants. The Airy functions are well understood and exact connection formulas can be established obtaining that

$$\psi(x) \to \frac{\alpha}{2\sqrt{|p(x)|}}\left[\cos(\mu)\mathrm{e}^{-|\xi|} + 2\sin(\mu)\mathrm{e}^{|\xi|}\right]$$

from the turning point to the classically forbidden region, and

$$\psi(x) \to \frac{\alpha}{\sqrt{|p(x)|}}\cos\left(|\xi| + \mu - \frac{1}{4}\pi\right)$$

from the turning points to the classically allowed region. Thus, we have established a connection between oscillatory and exponentially increasing and decreasing solutions. Clearly, this entire approach breaks down if the energy is too close to a value corresponding to a stationary point of the potential. In this case a different approximation must be implemented. For instance, in the case of a turning point with multiplicity 2 then the approximate solution of the Schrödinger equation (3) is given by means of Weber parabolic cylinder functions.

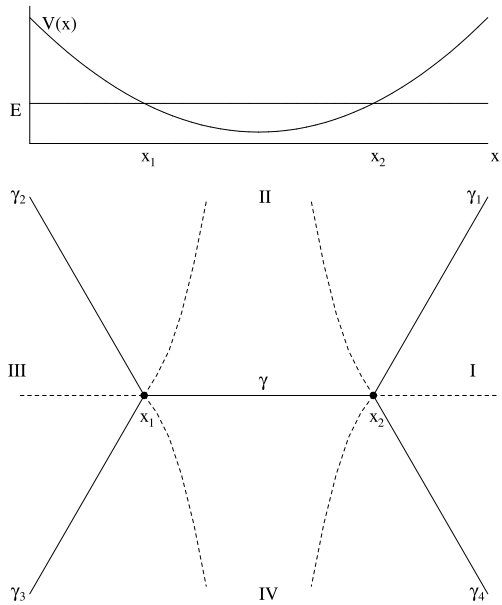

Perturbation Theory, Semiclassical, Figure 2
Stokes and anti-Stokes lines in a neighborhood of the bottom of a single well

Bound States for a Single Well Potential

We apply now the previous techniques in order to compute the stable states, that is normalized solutions of the Eq. (2), when the potential $V(x)$ has a single well shape; that is it has a simple minimum point $x_{\min}$, with minimum value $V_{\min}$ such that $V(x) > V_{\min}$ for any $x \neq x_{\min}$ and $\liminf_{|x|\to+\infty} V(x) > V_{\min}$. Then, for $E > V_{\min}$ close enough to the bottom of the well, equation $p(x) = 0$ has only two simple solutions $x_1 < x_{\min} < x_2$ and the typical picture of the Stokes lines appears as in Fig. 2. Here, $\psi_{\pm,x_1}$ (resp. $\psi_{\pm,x_2}$) denotes the fundamental solutions with asymptotic behaviors (7) and (9) in a neighborhood of x_1 (resp. x_2). Since we are looking for normalized solutions then in the classically forbidden region $x > x_2$, contained in the region I, the *dominant* term ψ_{+,x_2} of the solution ψ should have coefficient b^{I}_{+,x_2} exactly zero, that is

$$\psi(z) = \psi_{-,x_2}(z)\,, \quad z \in \mathrm{I}\,,$$

where the coefficients b_{-,x_2} of the *recessive* solution ψ_{-,x_2} is chosen equal to 1. Turning around the turning point x_2 and crossing the anti-Stokes line it follows that the coefficients of the solution $\psi = b^{\mathrm{II}}_{+,x_2}\psi_{+,x_2}(x) + b^{\mathrm{II}}_{-,x_2}\psi_{-,x_2}(x)$ in the region II take the form

$$\begin{pmatrix} b^{\mathrm{II}}_{+,x_2} \\ b^{\mathrm{II}}_{-,x_2} \end{pmatrix} = F_2\begin{pmatrix} 0 \\ 1 \end{pmatrix}\,,$$

where

$$F_2 \sim \begin{pmatrix} 1 & i \\ 0 & 1 \end{pmatrix}$$

according to the Stokes' rule.

In order to study the matching condition around the other turning point x_1 we have to transport the origin of integration at x_1 obtaining

$$\psi(z) = b^{\mathrm{II}}_{+,x_1}\psi_{+,x_1}(z) + b^{\mathrm{II}}_{-,x_1}\psi_{-,x_1}(z)$$

where

$$b^{\mathrm{II}}_{+,x_1} = \psi_{+,x_2}(x_1)b^{\mathrm{II}}_{+,x_2} \quad \text{and} \quad b^{\mathrm{II}}_{-,x_1} = \psi_{-,x_2}(x_1)b^{\mathrm{II}}_{-,x_2}.$$

On the other hand, also in region III the *dominant* term ψ_{+,x_1} of the solution ψ should have coefficient b^{III}_{+,x_1} exactly zero; from this fact and from the Stokes' rule applied to the turning point x_1 it follows that the equation for the dominant term of the bound states is

$$\psi_{+,x_2}(x_1) + \psi_{-,x_2}(x_1) = 0\,,$$

which implies

$$\cos\left[\frac{1}{\hbar}\left(\int_{x_1}^{x_2} p(x)\mathrm{d}x + \mathcal{O}(\hbar)\right)\right] = 0\,.$$

Hence, we obtain the well known Bohr–Sommerfeld rule

$$\int_{x_1}^{x_2} p(x)\mathrm{d}x = \hbar\left[\frac{1}{2}\pi + n\pi\right] + \mathcal{O}(\hbar^2)\,, \quad n \in \mathbb{N}\,.$$

Double Well Model: Estimate of the Splitting and the "Flea of the Elephant"

We consider now the case of a symmetric double well potential; that is $V(-x) = V(x)$ and it has two simple minima at $x_+ > 0$ and $x_- = -x_+$ separated by a barrier, we assume also that the minimum value $V_{\min} < \liminf_{|x|\to\infty} V(x)$. For instance, $V(x) = x^4 - 2\alpha x^2$ for some $\alpha > 0$; in this case the shape of the potential has two symmetric wells where $x_\pm = \pm\sqrt{\alpha}$ and $V_{\min} = -\alpha^2 < 0$, the two wells are separated by an energy barrier with top $V_{\max} = 0$ at $x = 0$. The semiclassical double well model is not only a very enlightening pedagogical problem, but it is also the basic argument explaining many relevant physical questions, see, e. g., [5,12,16,32,34].

In the interval $(V_{\min}, V_M)$, where

$$V_M = \min\left[V_{\max}, \liminf_{|x|\to\infty} V(x)\right]\,,$$

the stable states appear as *doublets* $E_\pm$ whose distance ω, named *splitting*, is quite small. The associated eigenvectors are even and odd (real-valued) wave-functions, that is

$$\psi_\pm(-x) = \pm\psi_\pm(x)\,. \tag{12}$$

The splitting $\omega = E_- - E_+$ can be computed as

$$\omega = \hbar^2 \frac{\psi_+(0)\psi'_-(0)}{\int_0^\infty \psi_-(x)\psi_+(x)\mathrm{d}x}$$

and it turns out to be exponentially small as $\hbar \to 0$. More precisely, if E_+ is the ground state then $\omega = \mathcal{O}(\mathrm{e}^{-S_0/\hbar})$ where

$$S_0 = \int_{x_-}^{x_+} \sqrt{V(x) - V_{\min}}\mathrm{d}x$$

is the Agmon distance between the two wells.

We would emphasize that the eigenvectors $\psi_\pm$ are asymptotically localized on both wells because of the symmetry property (12). The effect on the ground state of a small perturbation $W(x)$ that breaks the symmetry is worth mentioning. In such a case the property (12) does not work and, even if the small perturbation $W(x)$ is supported only on one side and far from the bottom of the well, the ground state, instead of being asymptotically supported on both wells, may be localized on just one well. According to Barry Simon we may state that *the perturbation $W(x)$ is a small flea on the elephant $V(x)$. The flea does not change the shape of the elephant* – in the sense that the splitting is still exponentially small – *but it can irritate the elephant enough so that it shifts its weight* – in the sense that the ground state is localized on just one well.

Semiclassical Approximation in Any Dimension

Here we consider the eigenvalue problem (2) in any dimension $n \geq 1$ where the potential V is assumed to be a *multi-well potential* [6,17,18,31]. More precisely, we assume that

$$V_{\min} := \inf V(x) < V_\infty = \liminf_{|x|\to+\infty} V(x)$$

and for any $E \in (V_{\min}, V_\infty)$ we can write the decomposition of

$$V^{-1}\left((-\infty, E]\right) = \cup_{j=1}^N U_j$$

as the union of N disjoint, compact and connected sets U_j. Inside these sets, named *wells*, the classical motion is allowed; outside the classical motion is forbidden.

From well-known results we have that for energies in the interval $(V_{\min}, V_\infty)$ the eigenvalue problem (2) admits only discrete spectrum, that is we have only isolated eigenvalues with finite multiplicity. In order to compute these eigenvalues we'll consider, at first, the solutions of Eq. (2) inside any single well and then we'll take into account the tunneling effect among the adjacent wells as a perturbation.

Actually, it is also possible to consider the case where $E > V_\infty$. In such a case we don't expect to have isolated eigenvalues but, under some circumstances, resonant states [13,19,20]. However, we don't fix our attention here on this problem.

Semiclassical Eigenvalues at the Bottom of a Well

Let x_0 be a local nondegenerate minimum for the potential $V(x)$. For the sake of definiteness we assume that $x_0 = 0$ and the local minimum is such that $V(0) = \nabla V(0) = 0$ and the Hessian matrix $\mathrm{Hess} = \left(\partial^2 V(0)/\partial x_i \partial x_j\right)_{i,j=1\ldots n}$ is positively defined with eigenvalues $2\lambda_j > 0$, $j = 1, \ldots, n$; furthermore, we choose the system of coordinates such that the Hessian matrix has diagonal form, i. e. $\mathrm{Hess} = \mathrm{diag}(2\lambda_j)$.

In order to compute the eigenvalues at the bottom of the well containing the minimum x_0 we approximate the potential by means of the n-dimensional harmonic oscillator. Thus, Eq. (2) takes the form

$$\sum_{j=1}^{n}\left[-\hbar^2\frac{\partial^2\psi}{\partial x_j^2} + \lambda_j x_j^2 - E\right]\psi = 0$$

which has exact eigenvalues

$$\hbar\left[\sum_{j=1}^{n}\sqrt{\lambda_j}(2\alpha_j+1)\right], \quad \alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n ,$$

with normalized eigenvectors

$$(\hbar\pi)^{-n/4}\prod_{j=1}^{n}\left(2^{\alpha_j}\alpha_j!\right)^{-1/2}\lambda_j^{1/8} \cdot \mathrm{e}^{-\sqrt{\lambda_j}x_j^2/2\hbar}H_{\alpha_j}\left(\lambda_j^{1/4}\hbar^{-1/2}x\right) ,$$

where H_m is the Hermite polynomial of degree m.

It is clear that when $\hbar$ is small enough then the first energy level of the harmonic oscillator is very close to the bottom of the potential and thus we expect that such an approximation gives the leading term of the first energy level and wave-function of (2). In fact, there exist two formal power series

$$E(\hbar) \sim \sum_{j=0}^{\infty}\hbar^j e_j \quad \text{and} \quad a(x;\hbar) \sim \sum_{r=0}^{\infty}\hbar^r a_r(x) , \tag{13}$$

where $e_0 = V_{\min} = 0$, $\hbar e_1 = \hbar\sum_{j=1}^{n}\sqrt{\lambda_j}$ is the first eigenvalue of the associated harmonic oscillator and $a_0 = \left[\prod_{j=1}^{n}\lambda_j/\pi^2\right]^{1/8}$, such that the function

$$\psi(x;\hbar) = \hbar^{-n/4}a(x;\hbar)\mathrm{e}^{-\varphi(x)/\hbar} \tag{14}$$

is such that

$$-\hbar^2\Delta\psi + \left[V(x) - E(\hbar)\right]\psi = \mathcal{O}(\hbar^\infty)\mathrm{e}^{-\varphi(x)/\hbar}$$

in a neighborhood of $x_0 = 0$, where $\varphi(x) = \frac{1}{2}\sum_{j=1}^{n}\sqrt{\lambda_j}\, x_j^2 + \mathcal{O}(|x|^2)$ is the positive solution of the Eikonal equation

$$|\nabla\varphi|^2 = V - e_0$$

in a neighborhood of $x_0 = 0$. The way to obtain this result essentially consists of inserting the formal power series (14) in Eq. (2) and expanding in powers of $\hbar$ the coefficients of $\mathrm{e}^{-\varphi(x)/\hbar}$, then of requiring the cancelation of the coefficients of $\hbar^j$ for any $j \in \mathbb{N}$.

These approximate solutions are valid in a neighborhood of the nondegenerate minimum and they are the natural candidates to become the true eigenvalue and eigenfunction of the problem (2). In fact, we'll extend this single well approximate solution (14) to a larger domain and then we'll solve the *connection* problem.

To this end we fix an open, sufficiently small, regular bounded set Ω containing the minimum point x_0; then Eq. (2) with Dirichlet boundary condition on Ω, that is

$$\psi_\Omega|_{\partial\Omega} = 0 , \quad \|\psi_\Omega\|_{L^2(\Omega)} = 1 , \tag{15}$$

admits one simple eigenvalue $E_\Omega(\hbar)$ at the bottom of the well, close to the first eigenvalue of the harmonic oscillator $e_0 + \hbar e_1$, which admits the asymptotic expansion (13).

Actually, such an eigenvalue will depend on the choice of the domain Ω. More precisely, if $E_{\Omega'}(\hbar)$ denotes the first eigenvalue of Eq. (2) with Dirichlet boundary condition on Ω' for a different domain Ω' containing x_0, then $E_{\Omega'}(\hbar)$ differs from $E_\Omega(\hbar)$ for an exponentially small term. Furthermore, it is also possible to prove that, modulo an exponentially small error, the spectrum of the Schrödinger equation (2) in some interval depending on $\hbar$ is the same as the spectrum of the direct sum of the Dirichlet single well problems in the same interval.

Agmon Metric

In order to study the exponential behavior of the solution (14) when x is far from the minimum x_0 we introduce now the Agmon distance between two points $x_1, x_2 \in \mathbb{R}^n$ defined as

$$\mathrm{d}_E(x_1, x_2) = \inf_{\gamma} \int_{\gamma} \sqrt{[V(x) - E]_+} \mathrm{d}x \,,$$

where

$$[V(x) - E]_+ = \max\{V(x) - E, 0\} \,,$$

γ is any regular path connecting the two points x_1 and x_2 and $E \geq V_{\min}$ is fixed. We underline that the Agmon distance defines a (pseudo)-metric on the Euclidean space $\mathbb{R}^n$. Indeed, it satisfies the triangle inequality

$$\mathrm{d}_E(x_1, x_2) \leq \mathrm{d}_E(x_1, x_3) + \mathrm{d}_E(x_3, x_2), \quad \forall x_1, x_2, x_3 \in \mathbb{R}^n \,,$$

but it is degenerate on each well U where $V < E$, in particular $\mathrm{d}_E(x_1, x_2) = 0$ for any couple of points x_1 and x_2 belonging to the same well and the *diameter* of each well U is zero with respect to this metric.

This kind of (pseudo-)metric has been used to obtain precise exponential decay of wave-functions of Schrödinger operators. Indeed, the exponential term $\varphi(x)$ in (14) is simply given by

$$\varphi(x) = \mathrm{d}_{e_0}(x_0, x)$$

for any x in a neighborhood of x_0.

Furthermore, by means of the Agmon metric we are also able to give an estimate of the exponential decay of wavefunctions not only for x in a neighborhood of the minimum x_0, but also for x far from this point. In particular, let $V_{\min} < E < V_\infty$ be fixed and let U be one of the wells, we consider now the eigenvalue problem for the Schrödinger equation with Dirichlet boundary conditions on a regular open set Ω containing the well U. Let $E_\Omega < E$ be an eigenvalue of Eq. (2) with Dirichlet boundary condition (15) with associated eigenfunction ψ_Ω. Then, the Agmon theorem enables us to obtain the following decay estimate for the eigenfunction ψ_Ω: for any fixed and positive ϵ then

$$\left\| \nabla \left[\mathrm{e}^{\mathrm{d}_E(x,U)/\hbar} \psi_\Omega \right] \right\|_{L^2(\Omega)} + \left\| \mathrm{e}^{\mathrm{d}_E(x,U)/\hbar} \psi_\Omega \right\|_{L^2(\Omega)} \leq C_\epsilon \mathrm{e}^{\epsilon/\hbar}$$

for some positive constant C_ϵ depending on ϵ (of course we expect that C_ϵ will grow as ϵ goes to zero). That is we have a good a priori estimate of the wavefunction ψ_Ω in the weighted $H^1(\Omega)$ space with weight $\mathrm{e}^{\mathrm{d}_E(x,U)/\hbar}$. Since $d(x, U) = 0$ for any x belonging to the well U then it follows that the solution ψ_Ω is exponentially decreasing outside the classical allowed region, as already seen in the one-dimensional problems.

Tunneling Between Wells

In order to consider the tunneling effect among the wells $\{U_j\}_{j=1}^N$ for any fixed $E \in (V_{\min}, V_\infty)$ we define

$$\mathrm{d}_E(U_i, U_j) = \inf_{x \in U_i, y \in U_j} \mathrm{d}_E(x, y)$$

as the Agmon distance between the two wells U_i and U_j, by construction it turns out that this distance is strictly positive for any $i \neq j$. A special role will be played by the minimal distance among these wells

$$S_0 = \min_{i \neq j} \mathrm{d}_E(U_i, U_j) \,.$$

In fact, the discrete spectrum of the Dirichlet realization of the Schrödinger equation on the boundary of the wells give, up to error of the order $\mathrm{e}^{-S_0/\hbar}$, the discrete spectrum of the original Schrödinger equation (2). In order to be more definite, let $M_j^{S,\eta}$ be the open set containing the well U_j defined as

$$M_j^{S,\eta} = \{x \in \mathbb{R}^n : \mathrm{d}_E(x, U_j) < S \text{ and } \mathrm{d}_E(x, U_k) > \eta, k \neq j\}$$

that is $M_j^{S,\eta}$ is, essentially, a ball (with respect to the Agmon pseudo-metric) large enough centered in the well U_j where we have eliminated the points from the other wells. If necessary we can regularize the boundary of $M_j^{S,\eta}$. In the following we take $\eta > 0$ small enough and S large enough, in particular we require that $S > 2S_0$. Let σ_j be the discrete spectrum of the Schrödinger equation (2) with Dirichlet condition on $M_j^{S,\eta}$ and let σ be the discrete spectrum of the Schrödinger equation (2). Then, for any interval $I(\hbar) = [\alpha(\hbar), \beta(\hbar)]$, where $\alpha(\hbar), \beta(\hbar) \to e_0$ as $\hbar \to 0$, there exists a bijection

$$b : \sigma \cap I(\hbar) \to \left[\cup_{j=1}^N \sigma_j \right] \cap I(\hbar)$$

such that for any $\rho < S_0 - 2\eta$

$$|b(\lambda) - \lambda| \leq C_\rho \mathrm{e}^{-\rho/\hbar} \tag{16}$$

for some $C_\rho > 0$.

Actually, this result could be improved by assuming some regularity properties on the potential V and estimate (16) can be replaced by the precise asymptotic behavior.

If we consider now the *symmetric double well* problem where $N = 2$ and with symmetric potential, e. g. $V(x_1, x_2, \ldots, x_n) = V(-x_1, x_2, \ldots, x_n)$, then, by symmetry, we have that the two Dirichlet realizations coincide, up to the inversion $x_1 \to -x_1$. Then $\sigma_1 = \sigma_2$. If we denote by E_1 and E_2 the first two eigenvalues of σ then $E_1 < E_2$ (in fact, the first eigenvalue is always nondegenerate) and the splitting between them can be estimated as

$$|E_2 - E_1| \le |E_2 - b(E_2)| + |b(E_2) - b(E_1)| + |b(E_1) - E_1| \le C_\rho \mathrm{e}^{-\rho/\hbar} ,$$

since $b(E_1) = b(E_2)$, for any $\rho < S_0$ since $\eta > 0$ is arbitrary.

Propagation of Quantum Observables

So far we have restricted our analysis to the semiclassical computation of the stable states of the time-independent Schrödinger equation (2). However, there exist some relevant results for what concerns the time evolution operator $\mathrm{e}^{-itH/\hbar}$, which is the formal solution of the time-dependent Schrödinger equation (1) with Hamiltonian $H = -\hbar^2\Delta + V$. In other words, this result is connected to the time-evolution of quantum observables [8,29]. The main tool is the $\hbar$-pseudodifferential calculus we briefly review below.

Brief Review of $\hbar$-Pseudodifferential Calculus

Here, we briefly review some basic results of the semiclassical pseudodifferential (also called $\hbar$-pseudodifferential) calculus. We refer to the books by Folland [9], Grigis and Sjöstrand [15], Martinez [22] and Robert [28] for a detailed treatment.

In this brief review, in order to avoid some technicalities, we restrict ourselves to the simpler case of bounded potentials; however, some of the following results hold in the general case of unbounded potentials too (see, e. g. [28]). To this end we consider *symbols* $a(x, y, p) \in S_{3n}(\langle p\rangle^m)$ for some positive m, that is the function a defined on $\mathbb{R}^{3n}$ depends smoothly on p and for any $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$ one has

$$\left|\nabla_p^\alpha a(x, y, p)\right| = \mathcal{O}\left(\langle p\rangle^m\right) \tag{17}$$

uniformly. That is

$$\left|\nabla_p^\alpha a(x, y, p)\right| \le C\langle p\rangle^m = C\left[1 + |p|\right]^{m/2}$$

for some positive constant C independent of x, y and p, where $\nabla_p^\alpha = (\partial_{p_1}^{\alpha_1}, \ldots, \partial_{p_n}^{\alpha_n})$ and $|p| = |p_1| + \ldots + |p_n|$.

We associate to a symbol $a \in S_{3n}(\langle p\rangle^m)$ the *semiclassical pseudodifferential operator* of degree m defined for $u \in C_0^\infty(\mathbb{R}^n)$ as the Fourier integral operator

$$\left[\mathrm{Op}_\hbar(a)u\right](x) = \frac{1}{[2\pi\hbar]^n} \cdot \int_{\mathbb{R}^n\times\mathbb{R}^n} \mathrm{e}^{i(x-y)\cdot p/\hbar} a(x, y, p)u(y)\mathrm{d}y\mathrm{d}p .$$

We notice that it is formally self-adjoint on the Hilbert space $L^2(\mathbb{R}^n)$ when the symbol a is such that $a(x, y, p) = \overline{a(y, x, p)}$.

The above pseudodifferential operator can be extended in a unique way to a linear continuous operator on the space of smooth rapidly decreasing functions $S(\mathbb{R}^n)$ and its dual space of tempered distributions $S'(\mathbb{R}^n)$. Furthermore, when the symbol $a \in S_{3n}(1)$ then the associate pseudodifferential operator is continuous on $L^2(\mathbb{R}^n)$ (this result is the so-called Calderón–Villancourt theorem).

It is easy to see that this class of pseudodifferential operators contains the usual differential operators. For instance, in the particular case where the symbol a has the form

$$a(x, y, p) = \sum_{|\alpha|\le m} b_\alpha(x)p^\alpha$$

then its associated operator is the differential operator given by

$$\left[\mathrm{Op}_\hbar(a)u\right](x) = \left[\mathrm{Op}_\hbar\left(\sum_{|\alpha|\le m} b_\alpha(x)p^\alpha\right)u\right](x) = \sum_{|\alpha|\le m} b_\alpha(x)\left(i\hbar\nabla_x\right)^\alpha .$$

The class of pseudodifferential operator is closed with respect to the composition. That is: given two symbols $a \in S_{3n}(\langle p\rangle^m)$ and $b \in S_{3n}(\langle p\rangle^{m'})$ then the composition of the associated pseudodifferential operators is still a pseudodifferential operator:

$$\mathrm{Op}_\hbar(a) \circ \mathrm{Op}_\hbar(b) = \mathrm{Op}_\hbar(c)$$

where the symbol $c \in S_{3n}\left(\langle p\rangle^{m+m'}\right)$ depends on $\hbar$ and it is given by the *Moyal product*

$$c(x,y,p) = (a\sharp b)(x,y,p) = \frac{1}{[2\pi\hbar]^n} \cdot \int_{\mathbb{R}^n\times\mathbb{R}^n} e^{i(x-z)\cdot(\eta-p)/\hbar} a(x,z,\eta)b(z,y,p)\mathrm{d}z\,\mathrm{d}\eta \sim \sum_{|\alpha|\geq 0} \frac{\hbar^{|\alpha|}}{i^{|\alpha|}\alpha!}\nabla_z^\alpha\,\nabla_\eta^\alpha\left[a(x,z,\eta)b(z,y,p)\right]\Big|_{z=x,\eta=p}$$

as $\hbar$ goes to zero; we should underline that the above asymptotic formula becomes an exact formula when the symbol a is polynomial with respect to p since the sum becomes finite. As a result it follows that for any *elliptic symbol* $a \in S_{3n}\left(\langle p\rangle^m\right)$, that is such that $|a(x,y,p)| \geq C\langle p\rangle^m$ for some positive constant C, then its associate pseudodifferential operator is invertible in the sense that there exists a symbol $b \in S_{3n}\left(\langle p\rangle^{-m}\right)$, depending on $\hbar$, where

$$\mathrm{Op}_\hbar(a)\circ\mathrm{Op}_\hbar(b) = 1 + \mathrm{Op}_\hbar(r)$$

and

$$\mathrm{Op}_\hbar(b)\circ\mathrm{Op}_\hbar(a) = 1 + \mathrm{Op}_\hbar(s)\,,$$

with $r, s = \mathcal{O}(\hbar^\infty)$ in S_{3n} (1). The symbol $b \sim \sum_j \hbar^j b_j$ is iteratively defined where $b_0 = \frac{1}{a}$.

Now, we are ready to define the *quantization of a classical observable*. Classical observables are functions $a(x,p)$ of the position $x \in \mathbb{R}^n$ and of the momenta $p \in \mathbb{R}^n$, if we assume that a is a smooth function such that

$$\left|\nabla_p^\alpha a(x,p)\right| \leq C\langle p\rangle^m$$

then, for any $\rho \in [0,1]$, it follows that

$$a^\rho(x,y,z) = a\left[(1-\rho)x + \rho y, p\right] \in S_{3n}\left(\langle p\rangle^m\right)\,.$$

We define the *quantization of the observable a* as:

$$\mathrm{Op}_\hbar^\rho(a) = \mathrm{Op}_\hbar(a^\rho)$$

where the values $\rho = 0, \frac{1}{2}, 1$ will play a special role; in particular, for $\rho = 0$ we have the *standard* (also called *left*) quantization, for $\rho = 1$ we have the *right* quantization and for $\rho = \frac{1}{2}$ we have the *Weyl* quantization

$$\mathrm{Op}_\hbar^W(a) := \mathrm{Op}_\hbar^{1/2}(a) = \mathrm{Op}_\hbar(a^{1/2})\,.$$

We emphasize that the Weyl quantization is particularly important in quantum mechanics because when the classical observable a is a real valued function then the associate pseudodifferential operator $\mathrm{Op}_\hbar^W(a)$ is formally self-adjoint on $L^2(\mathbb{R}^n)$.

We close this review by recalling the following important result which connects the commutator $[A,B] = AB - BA$ of the pseudodifferential operators $A = \mathrm{Op}_\hbar^\rho(a)$ and $B = \mathrm{Op}_\hbar^\rho(b)$, where a, b are two classical observables, and the *Poisson bracket* of a and b: let c be the unique symbol such that $\mathrm{Op}_\hbar^\rho(c) = [A,B]$, then

$$c = -i\hbar\,\{a,b\} + \mathcal{O}(\hbar^2)\,.$$

Egorov Theorem

Now, we are ready to compare the classical evolution of a classical observable with the quantum evolution of its quantum counterpart. With more details let $h(x,p)$ be a classical Hamiltonian where $x \in \mathbb{R}^n$ denotes the spatial variable and $p \in \mathbb{R}^n$ denotes the momentum. Let

$$\phi^t\colon \mathbb{R}^{2n} \to \mathbb{R}^{2n}$$

be the (classical) Hamiltonian flux. Thus, for any classical observable function $b = b(x,p) \in C^\infty(\mathbb{R}^{2n};\mathbb{R})$ let

$$b^t(x,p) = \left(b\circ\phi^t\right)(x,p) = b[\phi^t(x,p)]$$

be the classical evolution of this observable.

It is well known that we can associate to any real-valued classical observable $b(x,p)$ a symmetric *$\hbar$-pseudodifferential linear operator* denoted by $\mathrm{Op}_\hbar^W(b)$ by means of the semiclassical (Weyl) quantization rule formally defined as

$$\left[\mathrm{Op}_\hbar^W(b)u\right](x) := \frac{1}{[2\pi\hbar]^n} \cdot \int_{\mathbb{R}^n\times\mathbb{R}^n} e^{i(x-y)\cdot p/\hbar} b\left(\frac{x+y}{2}, p\right) u(y)\mathrm{d}y\mathrm{d}p$$

where $(x-y)\cdot p = \sum_{i=1}^n (x_i - y_i)p_i$. In order to properly define this integral operator on the Hilbert space $L^2(\mathbb{R}^n)$ we require that estimate (17) holds; actually, to the present purposes it is sufficient to assume the following weaker condition on the observable b:

$$\left|\nabla_{x,p}^\alpha b(x,p)\right| \leq C\left[1 + |x|^2 + |p|^2\right]^{m/2} \tag{18}$$

for some $m \geq 0$ and any $\alpha \in \mathbb{N}^{2n}$.

For instance, if h_0 is the Hamiltonian associated to a harmonic oscillator, then it is a quadratic function with respect to both position and momentum variables

$$h_0(x,p) = \sum_{j=1}^n \left[p_j^2 + \omega_j^2 x_j^2\right] \tag{19}$$

and the associated Hamiltonian operator (let $n = 1$ for the sake of simplicity) takes the usual form

$$\begin{aligned}[H_0 u](x) &= \left[\mathrm{Op}_\hbar^W(h_0)u\right](x) \\ &= \frac{1}{2\pi\hbar}\int_{\mathbb{R}\times\mathbb{R}} \mathrm{e}^{i(x-y)p/\hbar}\left[p^2+\omega^2\left(\frac{x+y}{2}\right)^2\right] \\ &\quad \cdot u(y)\mathrm{d}y\mathrm{d}p \\ &= -\hbar^2\frac{\mathrm{d}^2}{\mathrm{d}x^2}+\omega^2x^2\end{aligned}$$

by integrating by parts twice and since $1/(2\pi\hbar)\int_{\mathbb{R}} \mathrm{e}^{i(x-y)p/\hbar}\mathrm{d}p = \delta(x-y)$.

In such a way it is possible to associate a classical observable b to a quantum operator B. If we denote by $B = \mathrm{Op}_\hbar^W(b)$ and $H = \mathrm{Op}_\hbar^W(h)$ the operators associated to the classical observable b and to the Hamiltonian h then the time quantum evolution $B^t = \mathrm{e}^{itH/\hbar}B\mathrm{e}^{-itH/\hbar}$ of the observable B solves the Heisenberg equation

$$\frac{\mathrm{d}B^t}{\mathrm{d}t} = \frac{i}{\hbar}\left[H, B^t\right]$$

and it is, in some sense, related with the classical evolution b^t as we are going to explain.

In the very particular case where the Hamiltonian h_0 is given by the harmonic oscillator (19) then we have that the quantum evolution B^t of the observable B is a $\hbar$-pseudodifferential operator with symbol b^t given by means of the classical flux Hamiltonian; that is

$$\mathrm{e}^{itH_0/\hbar}\mathrm{Op}_\hbar^W(b)\,\mathrm{e}^{-itH_0/\hbar} = \mathrm{Op}_\hbar^W\left(b\circ\phi^t\right)$$

where ϕ^t is the Hamiltonian flux generated by the Hamiltonian (19).

In other words, *for the harmonic oscillator Hamiltonian then quantum evolution and Weyl quantization commute*. This property is specific for this problem and it comes from the fact that the flux Hamiltonian is a linear function with respect to (x, p). This fact is not true in a general case. However, it is possible to see that a generalization of such a result holds when H_0 is replaced by any symmetric $\hbar$-pseudodifferential operator H; this result is the so-called *Egorov theorem* (see [8] for the original statement, see also [29] for a detailed review). In particular, under some assumptions, it is possible to prove that the zeroth order remainder term

$$R(t) := \mathrm{e}^{itH/\hbar}\mathrm{Op}_\hbar^W(b)\,\mathrm{e}^{-itH/\hbar} - \mathrm{Op}_\hbar^W\left(b\circ\phi^t\right)$$

is a bounded operator such that $\|R(t)\| = \mathcal{O}(\hbar)$ in the sense that the norm of the operator $R(t)$ is bounded

$$\|R(t)\| \le C\hbar$$

for some $C > 0$ and for any $t \ge 0$.

Furthermore, it is possible to extend this asymptotic result to any order $N \ge 1$ for a suitable choice of $p_{n,t}$, that is the classical and quantum evolution coincides up to a term of order $\hbar^{N+1}$ in the semiclassical limit:

$$\begin{aligned}\mathrm{e}^{itH/\hbar}\mathrm{Op}_\hbar^W(b)\,\mathrm{e}^{-itH/\hbar} - \mathrm{Op}_\hbar^W\left(b\circ\phi^t\right) \\ - \sum_{n=1}^{N}\hbar^n\mathrm{Op}_\hbar^W\left(p_{n,t}\right) = \mathcal{O}(\hbar^{N+1})\end{aligned}$$

in the norm sense.

Future Directions

Semiclassical methods are a field of research where theoretical results are rapidly evolving. Just to name some active research topics: $\hbar$-pseudodifferential operators, Weyl functional calculus, frequency sets, semiclassical localization of eigenfunctions, semiclassical resonant states, Born–Oppenheimer approximation, stability of matter and Scott conjecture, semiclassical Lieb–Thirring inequality, Peierls substitution rule, etc. Furthermore, semiclassical methods have been also successfully applied in different contests such as superfluidity and statistical mechanics.

Looking forward, we see new emerging research fields in the area of semiclassical methods: *numerical WKB interpolation techniques* and *semiclassical nonlinear Schrödinger equations*.

Indeed, the recent researches in the area of nanosciences and nanotechnologies have opened up new fields where models for semiconductor devices of increasingly small size and electric charge transport along nanotubes cannot be fully understood without considering their quantum nature. Since the oscillating behavior of the solutions of the Schrödinger equation induces serious difficulties for standard numerical simulations, then new numerical approaches based on WKB interpolation are required [1,4,26].

Although the nonlinear Schrödinger equation has been an argument of theoretical research since the 1970s, only in the last few years, with the successful experiments on Bose–Einstein condensate states, has an increasing interest been shown. When we add a nonlinear term to the time-dependent Schrödinger equation (1) then the dynamics of the model drastically changes and new peculiar features, such as the *blow-up effect* and *stability of*

stationary states, appear. Semiclassical arguments applied to nonlinear Schrödinger equations justify their reduction to finite dimensional dynamical systems and thus it is possible to obtain an approximate solution, at least for nonlinear time-dependent Schrödinger equations in small dimensional spaces, typically for $n = 1$ and $n = 2$. The extension of this technique to the case $n > 2$ and the validity of such an approximation for large times is still an open problem [14,27,30].

Bibliography

1. Ben Abdallah N, Pinaud O (2006) Multiscale simulation of transport in an open quantum system: Resonances ans WKB interpolation. J Comp Phys 213:288–310
2. Berezin FA, Shubin MA (1991) The Schrödinger equation. Kluwer, Dordrecht
3. Berry MV, Mount KE (1972) Semiclassical approximation in wave mechanics. Rep Prog Phys 35:315–397
4. Bonnaillie-Noël V, Nier F, Patel Y (2006) Computing the steady states for an asymptotic model of quantum transport in resonant heterostructures. J Comp Phys 219:644–670
5. Claviere P, Jona Lasinio G (1986) Instability of tunneling and the concept of molecular structure in quantum mechanics: The case of pyramidal molecules and the enantiomer problem. Phys Rev A 33:2245–2253
6. Dimassi M, Sjöstrand J (1999) Spectral asymptotics in the semiclassical limit. In: London Math Soc Lecture Note Series 268. Cambridge University Press, Cambridge
7. Dingle RB (1973) Asymptotic expansion: Their derivation and interpretation. Academic, London
8. Egorov YV (1971) Canonical transformation of pseudodifferential operators. Trans Moscow Math Soc 24:1–24
9. Folland G (1988) Harmonic analysis in phase space. Princeton University Press, Princeton
10. Fröman N, Fröman PO (1965) JWKB approximation. North Holland, Amsterdam
11. Fröman N, Fröman PO (2002) Physical problems solved by the phase integral methods. Cambridge University Press, Cambridge
12. Graffi S, Grecchi V, Jona-Lasinio G (1984) Tunneling instability via perturbation theory. J Phys A: Math Gen 17:2935–2944
13. Grecchi V, Martinez A, Sacchetti A (1996) Splitting instability: The unstable double wells. J Phys A: Math Gen 29:4561–4587
14. Grecchi V, Martinez A, Sacchetti A (2002) Destruction of the beating effect for a non-linear Schrödinger equation. Comm Math Phys 227:191–209
15. Grigis B, Sjöstrand J (1994) Microlocal analysis for differential operators. An introduction. In: London Math. Soc. Lecture Note Series 196. Cambridge University Press, Cambridge
16. Harrell EM (1980) Double wells. Commun Math Phys 75:239–261
17. Helffer B (1988) Semi-classical analysis for the Schrödinger operator and applications. Lecture Notes in Mathematics 1336. Springer, Berlin
18. Helffer B, Sjöstrand J (1984) Multiple wells in the semiclassical limit I. Comm Part Diff Eq 9:337–408
19. Helffer B, Sjöstrand J (1986) Resonances en limite semi-classique. Mém Soc Math France (N.S.) 24–25
20. Hislop P, Sigal IM (1996) Introduction to spectral theory. In: Appl Math Sci, vol. 113. Springer, New York
21. Landau LD, Lifshitz EM (1959) Quantum mechanics. Course of theoretical physics. Pergamon, Oxford
22. Martinez A (2002) An introduction to semiclassical and microlocal analysis. Springer, New York
23. McHugh JAM (1971) An historical survey of ordinary linear differential equations with a large parameter and turning points. Arch Hist Exact Sci 7:277–324
24. Merzbacher E (1970) Quantum mechanics, 2nd edn. Wiley, New York
25. Olver FWJ (1974) Asymptotics and Special Functions. Academic, New York
26. Presilla C, Sjöstrand J (1996) Transport properties in resonant tunneling heterostructures. J Math Phys 37:4816–4844
27. Raghavan S, Smerzi A, Fantoni S, Shenoy SR (1999) Coherent oscillations betwene two weakly coupled Bose–Einstein condensates: Josephson effects, π oscillations, and macroscopic quantum self-trapping. Phys Rev A 59:620–633
28. Robert D (1987) Autour de l'Approximation Semiclassique. Birkhäuser, Basel
29. Robert D (1988) Semi-classical approximation in quantum mechanics. A survey of old and recent mathematical results. Helv Phys Acta 71:44–116
30. Sacchetti A (2005) Nonlinear double well Schrödinger equations in the semiclassical limit. J Stat Phys 119:1347–1382
31. Simon B (1983) Semiclassical limit of low lying Eigenvalues I: Non degenerate minima. Ann H Poincaré 38:295–307
32. Simon B (1985) Semiclassical limit of low lying Eigenvalues IV: The flea of the elephant. J Funct Anal 63:123–136
33. Voros A (1982) Spectre de l'Équation de Schrödinger et Méthode BKW. Publications Mathmatiques d'Orsay 81.09
34. Wilkinson M, Hannay JH (1987) Multidimensional tunneling between excited states. Phys D: Nonlin Phenom 27:201–212

Perturbative Expansions, Convergence of

Sebastian Walcher
Lehrstuhl A für Mathematik, RWTH Aachen, Aachen, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Poincaré–Dulac Normal Forms
Convergence and Convergence Problems
Lie Algebra Arguments
NFIM and Sets of Analyticity
Hamiltonian Systems
Future Directions
Bibliography

Glossary

Resonant eigenvalues Let B be a linear endomorphism of $\mathbb{C}^n$, with eigenvalues $\lambda_1, \ldots, \lambda_n$ (counted according to multiplicity). One calls these eigenvalues *resonant* if there are integers $d_j \geq 0$, $\sum d_j \geq 2$, and some $k \in \{1, \ldots, n\}$ such that

$$d_1\lambda_1 + \cdots + d_n\lambda_n - \lambda_k = 0\,.$$

If B is represented by a matrix in Jordan canonical then the associated vector monomial $x_1^{d_1} \ldots x_n^{d_n} e_k$ will be called a *resonant monomial*. (Here $e_1, \ldots, e_n$ denote the standard basis, and the x_i are the corresponding coordinates.)

Poincaré–Dulac normal form Let $f = B + \ldots$ be a formal or analytic vector field about 0, and let B_s be the semisimple part of B. Then one says that f is in Poincaré–Dulac normal form (PDNF) if $[B_s, f] = 0$. An equivalent characterization, if B is in Jordan form, is to say that only resonant monomials occur in the series expansion.

Normalizing transformations and convergence A relatively straightforward argument shows that any formal vector field $f = B + \ldots$ can be transformed to a formal vector field in PDNF via a formal power series transformation. But for analytic vector fields, the existence of a convergent transformation is not assured. There are two obstacles to convergence: First, the possible existence of small denominators (roughly, this means that the eigenvalues satisfy "near-resonance conditions"); and second, "algebraic" obstructions due to the particular form of the normalized vector field.

Lie algebras of vector fields The vector space of analytic vector fields on an open subset U of $\mathbb{C}^n$, with the bracket $[p, q]$ defined by

$$[p, q](x) := Dq(x)\,p(x) - Dp(x)\,q(x)$$

becomes a Lie algebra, as is well known. Mutatis mutandis, this also holds for local analytic and for formal vector fields. As noted previously, PDNF is most naturally defined via this Lie bracket. Moreover, the "algebraic" obstructions to convergence are most appropriately discussed within the Lie algebra framework.

Normal form on invariant manifolds While there may not exist a convergent transformation to PDNF for a given vector field f, one may have a convergent transformation to a "partially normalized" vector field, which admits a certain invariant manifold and is in PDNF when restricted to this manifold. This observation is of some practical importance.

Definition of the Subject

It seems appropriate to first clarify what types of perturbative expansions are to be considered here. There exist various types of such expansions in various settings (a very readable introduction is given in Verhulst's monograph [41]), but for many of these settings the question of convergence is not appropriate or irrelevant. Therefore we restrict attention to the scenario outlined, for instance, in the introductory chapter of the monograph [13] by Cicogna and Gaeta, which means consideration of normal forms and normalizing transformation for local analytic vector fields.

Normal forms are among the most important tools for the local analysis and classification of vector fields and maps near a stationary point. (See ► Normal Forms in Perturbation Theory.) Convergence problems arise here, and they turn out to be surprisingly complex. While the first contributions date back more than a century, some very strong and very deep results are just a few years old, and this remains an active area of research. Clearly convergence questions are relevant for the analytic classification of local vector fields, but they are also of practical relevance in applications, e. g. for stability questions and for the existence of particular types of solutions.

Introduction

The theory of normal forms was initiated by Poincaré [35], and later extended by Dulac [17], and by Birkhoff [5] to Hamiltonian vector fields. There exist various types of normal forms, depending on the specific problem one wants to address. Bruno (see [6,7]) in the 1960s and 1970s performed a comprehensive and deep investigation of Poincaré–Dulac normal forms, which are defined with respect to the semisimple part of the linearization. Such normal forms are very important in applications, and moreover they have certain built-in symmetries, which allows a well-defined reduction procedure.

We will first give a quick review of normalization procedures and normal forms, and then discuss convergence problems (which mostly refers to convergence or divergence of normalizing transformations). We will present fundamental convergence and divergence results due to Poincaré, Siegel, Bruno and others. We then proceed to discuss the relevance of certain Lie algebras of analytic vector fields for these matters, including results by Cicogna and the author of this article on the influence of symmetries, and the far-reaching generalization of Bruno's theorems (among others) due to Stolovitch. Variants of normal forms which guarantee convergence on certain subsets (due to Bibikov and Bruno) are then discussed, and ap-

plications are mentioned. Finally, the Hamiltonian setting, which deserves a discussion in its own right, is presented, starting with results of Ito and recently culminating in Zung's convergence theorem. For Hamiltonian systems there is also work due to Perez–Marco on divergence of normal forms.

Within the space limitations of this contribution, and in view of some very intricate and space-consuming technical questions and conditions, the author tried to find an approach that, for some problems, should provide some insight into a result, the arguments in its proof or its relevance, without exhibiting all the technicalities, or without giving the most general statement. The author hopes to have been somewhat successful in this, and apologizes to the creators of the original theorems for presenting just "light" versions.

Poincaré–Dulac Normal Forms

We will start with a coordinate-free approach to normal form theory. Our objects are local ordinary differential equations (over $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$)

$$\dot{x} = F(x)\,, \qquad F(0) = 0$$

with F analytic, thus we have a convergent series expansion

$$F(x) = Bx + \sum_{j\geq 2} f_j(x) = Bx + f_2(x) + f_3(x) + \cdots \qquad \text{near} \quad 0 \in \mathbb{K}^n\,.$$

Here $B = DF(0)$ is linear, and each f_j is homogeneous of degree j. Our objective is to simplify the Taylor expansion of F. For this purpose, take an analytic "near-identity" map

$$H(x) = x + h_2(x) + \cdots$$

Since H is locally invertible, there is a unique

$$F^*(x) = Bx + \sum_{j\geq 2} f_j^*(x)$$

such that the identity

$$DH(x)F^*(x) = F(H(x)) \qquad \text{(R)}$$

holds. H "preserves solutions" in the sense that parametrized solutions of $\dot{x} = F^*(x)$ are mapped to parametrized solutions of $\dot{x} = F(x)$ by H. It is convenient to introduce the following abbreviation:

$$F^* \xrightarrow{H} F \quad \text{if (R) holds}\,.$$

Given the expansion

$$F(x) = Bx + f_2(x) + \cdots + f_{r-1}(x) + f_r(x) + \cdots\,,$$

assume that $f_2, \ldots, f_{r-1}$ are already deemed "satisfactory" (according to some specified criterion). Then the ansatz $H(x) = x + h_r(x) + \ldots$ yields

$$F^*(x) = Bx + f_2(x) + \cdots + f_{r-1}(x) + f_r^*(x) + \cdots$$

(with terms of degree $< r$ unchanged), and at degree r one obtains the so-called *homological equation*:

$$[B, h_r] = f_r - f_r^*$$

(Here $[p, q](x) = Dq(x)p(x) - Dp(x)q(x)$ denotes the usual Lie bracket of vector fields).

The space $\mathcal{P}_r$ of homogeneous vector polynomials of degree r is finite dimensional, and $\mathrm{ad}B = [B, \cdot]$ sends $\mathcal{P}_r$ to $\mathcal{P}_r$. Thus the homological equation poses a linear algebra problem on a finite dimensional vector space: Given B and f_r, determine f_r^* so that the equation can be solved and let h_r be a solution. How can f_r^* be chosen? Let $\mathcal{W}$ be any subspace of $\mathcal{P}_r$ such that

$$\text{image}\,(\mathrm{ad}B) + \mathcal{W} = \mathcal{P}_r\,.$$

Then one may choose $f_r^* \in \mathcal{W}$. If the sum is direct then $f_r^* \in \mathcal{W}$ is uniquely determined by f_r.

Generally, the type of normal form is specified by the choice of a subspace $\mathcal{W}_r$ for each degree r such that $\text{image}\,(\mathrm{ad}B) + \mathcal{W}_r = \mathcal{P}_r$. The Poincaré–Dulac choice is as follows: Given the decomposition

$$B = B_\mathrm{s} + B_\mathrm{n}$$

into semisimple and nilpotent part, then $\mathrm{ad}\,B = \mathrm{ad}B_\mathrm{s} + \mathrm{ad}B_\mathrm{n}$ is known to be the corresponding decomposition on $\mathcal{P}_r$. Choose $\mathcal{W}_r = \mathrm{Ker}(\mathrm{ad}\,B_\mathrm{s})$. By linear algebra

$$\mathcal{W}_r + \text{image}(\mathrm{ad}\,B) = \mathcal{P}_r\,;$$

and the sum is direct if B is semisimple. In any case we have $[B_\mathrm{s}, f_r^*] = 0$.

Definition 1 The vector field F^* is in *Poincaré–Dulac normal form* if $[B_\mathrm{s}, f_j^*] = 0$ for all j; equivalently if $[B_\mathrm{s}, F^*] = 0$.

If one sets aside convergence questions for the moment, an immediate consequence of the considerations above is:

Proposition 1 *For any $F = B + \cdots$ there are formal power series $H(x) = x + \cdots$, $F^*(x) = Bx + \ldots$ such that $F^* \xrightarrow{H} F$ and F^* is in Poincaré–Dulac normal form.*

One should note that the homological equation is of relevance beyond the setting of Poincaré–Dulac, and forms the foundation for various other (or more refined) types of normal form. See also the entry in this section of the Encyclopedia by T. Gramchev on normal forms with respect to a nilpotent linear part.

Now let us turn to the standard approach via suitable coordinates. Given F be as above, let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of B. Complexify, if necessary, and assume that B is in Jordan canonical form with respect to the given coordinates $x_1, \ldots, x_n$ (with corresponding basis $e_1, \ldots, e_n$ of $\mathbb{K}^n$). In particular,

$$B_s = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) .$$

In the following we will use x_i and e_i to denote eigencoordinates, resp. eigenbasis elements, and reserve $\lambda_1, \ldots, \lambda_n$ for the eigenvalues of B, without explicitly saying so in every instance.

Lemma 1 *The "vector monomial" $p(x) = x_1^{m_1} \ldots x_n^{m_n} e_j$ satisfies*

$$[B_s, p] = (m_1\lambda_1 + \cdots + m_n\lambda_n - \lambda_j) \cdot p$$

Thus, the vector monomials form an eigenbasis of $\operatorname{ad} B_s$ on the space $\mathcal{P}_r$, with eigenvalues $m_1\lambda_1 + \cdots + m_n\lambda_n - \lambda_j$ (m_j nonnegative integers, $\sum m_j = r$).

The eigenvalues play a crucial role both for the classification of formal normal forms and for convergence issues. The following distinction is pertinent here:

Definition 2 One calls $(\lambda_1, \ldots, \lambda_n)$ *resonant* if there are integers $d_j \geq 0$, $\sum d_j \geq 2$, and some $k \in \{1, \ldots, n\}$ such that

$$d_1\lambda_1 + \cdots + d_n\lambda_n - \lambda_k = 0 .$$

In this case, the vector monomial $x_1^{d_1} \ldots x_n^{d_n} e_k$ is also called a *resonant monomial*. One calls $(\lambda_1, \ldots, \lambda_n)$ *nonresonant* otherwise.

Given eigencoordinates, one can characterize a Poincaré–Dulac normal form by the property that only resonant monomials occur in the series expansion. Moreover, evaluation of the homological equation shows that the nonzero eigenvalues of $\operatorname{ad} B$ will occur as denominators in its solution. To iterate the normalization, one may proceed degree by degree with a series of transformations of the form $\exp(h_r)$, using the solution of the homological equation; see [44]. One may also choose a different iterative approach to compute a normalizing transformation, such as the important"distinguished transformation" of Bruno [6]. In any case this will yield coefficients whose denominators are products of nonzero terms $m_1\lambda_1 + \cdots + m_n\lambda_n - \lambda_j$. This is the source of convergence problems caused by *small denominators*:

There are formal series H and F^* such that $F^* \xrightarrow{H} F$ and F^* is in Poincaré–Dulac normal form, but does there exist a convergent H? (Here "convergent" means: convergent in some neighborhood of 0.)

To answer this question, it is necessary to specify a particular type of transformation: Transformations to Poincaré–Dulac normal form are not necessarily unique, even if one stipulates a near-identity transformation $H(x) = x + \cdots$. Non-uniqueness occurs whenever the eigenvalues of B are resonant, because then some homological equation will not have a unique solution: In eigencoordinates of B_s, the series of the transformation is fixed only up to resonant monomial terms. In particular, if F itself is in normal form then any formal power series $H(x) = x + \cdots$ that contains only resonant monomials will provide a normal form F^*, and a suitable nontrivial choice of H will even force $F^* = F$. Thus there may be divergent transformations sending an equation in normal form to itself.

Convergence and Convergence Problems

Let us note at the start that for given analytic F there may not exist any convergent transformation to normal form; thus the convergence question has no simple answer.

An early positive convergence result is due to Poincaré [35], with a later improvement due to Dulac [17]:

Theorem 1 (Poincaré–Dulac) *If $\lambda_1, \ldots, \lambda_n$ lie in an open half-plane in $\mathbb{C}$ which does not contain 0 then there exists a convergent transformation.*

For example, one may think of the open left half-plane. The proof is relatively straightforward, employing natural majorants. (Due to the hypothesis, the $|\sum_{i=1}^n m_i\lambda_i - \lambda_j|$ are unbounded for $\sum m_i \to \infty$.) Poincaré's condition does not preclude the existence of resonant monomials but it ensures that there are at most finitely many of these.

One main technical difficulty in proving convergence was to replace the direct majorant arguments by more efficient, and more sophisticated, tools, so that small denominator problems could be tackled. The following result, due to C.L. Siegel [38], may be seen as the start of the "modern phase" for convergence results. Characteristically, this result goes back to a mathematician who also worked in analytic number theory. Siegel assumed that the eigenvalues satisfy a certain arithmetic condition.

Condition S: The eigenvalues are pairwise different and there are constants $C > 0, \nu > 0$ such that for all nonnegative integer tuples (m_i), $\sum m_i > 1$, the following inequality holds:

$$\left|\sum_{i=1}^{n} m_i\lambda_i - \lambda_j\right| \geq C \cdot (m_1 + \cdots + m_n)^{-\nu}$$

Theorem 2 (Siegel) *If Condition S holds then there is a convergent transformation to normal form.*

Proof A very rough sketch of the proof is as follows. One works in eigencoordinates. By scaling one may assume that all coefficients in the expansion of F are absolutely bounded by some constant $M \geq 1$. For the transformation one writes

$$H(x) = x + \sum_{r\geq 2}\left(\sum \alpha_{m_1,\ldots,m_n,k} \cdot x_1^{m_1} \cdots x_n^{m_n} e_k\right)$$

where the sum inside the bracket extends over all nonnegative integer tuples with $\sum m_i = r$ and $1 \leq k \leq n$. (Since Condition S precludes resonances, the coefficients of the series are uniquely determined.) Now set

$$A_{m_1,\ldots,m_n} := \sum_k |\alpha_{m_1,\ldots,m_n,k}| \,.$$

From the homological equations one finds by recursion

$$\left|\sum_i m_i\lambda_i - \lambda_k\right| \cdot |\alpha_{m_1,\ldots,m_n,k}| \leq M \cdot \sum A_{d_{1,1},\ldots,d_{1,n}} \cdots A_{d_{s,1},\ldots,d_{s,n}}$$

where the summation on the right hand side extends over all tuples $(d_{i,1},\ldots,d_{i,n})$ that add up to $(m_1,\ldots,m_n)$. From this one may obtain an estimate for $A_{m_1,\ldots,m_n}$, and invoking Condition S one eventually arrives at the conclusion that $\sum A_{m_1,\ldots,m_n} x_1^{m_1} \ldots x_n^{m_n}$ is majorized by the series of

$$\frac{x_1 + \cdots + x_n}{1 - K \cdot (x_1 + \cdots + x_n)}, \quad \text{some } K > 0 \,.$$

□

Example Let $\dot{x} = Bx + \cdots$ be given in dimension two, and assume that the eigenvalues λ_1, λ_2 of B are nonresonant, and are algebraic irrational numbers. (This is the case when the entries of B are rational but the characteristic polynomial is irreducible over the rationals.) Then λ_2/λ_1 is algebraic but not rational, and (λ_1, λ_2) satisfies Condition S, due to a celebrated number-theoretic result of Thue, Siegel and Roth. Thus there exists a convergent transformation to normal form.

While Siegel's convergence proof uses majorizing series, the approach is not as straightforward as in the Poincaré setting. Siegel's result is strong in the sense that Condition S is satisfied by Lebesgue – almost all tuples $(\lambda_1,\ldots,\lambda_n) \in \mathbb{C}^n$. But the condition forces the normal form to be uninteresting: One necessarily has $F^* = B = B_s$. In the same paper [38], Siegel also notes that divergence is possible, even for a set of "eigenvalue vectors" that is everywhere dense in n-space.

In the resonant case there is a second source of obstacles to convergence. An early example for this is due to Horn (about 1890); see [6]:

Example The system

$$\dot{x}_1 = x_1^2$$
$$\dot{x}_2 = x_2 - x_1$$

(with eigenvalues (0, 1) for the linear part) admits no convergent transformation to normal form. A detailed proof for this can be found in [14]. The underlying reason is that the ansatz for a transformation – unavoidably – leads to the differential equation

$$x^2 \cdot y' = y - x \,,$$

(which goes back to Euler) with divergent solution $\sum_{k\geq 1}(k-1)!\, x^k$. There are no small denominators here. The problem actually lies within the normal form, which can be computed as

$$\dot{y}_1 = y_1^2$$
$$\dot{y}_2 = y_2 \,.$$

The single nonlinear term is sufficient to obstruct convergence.

Pliss [34] showed that Siegel's theorem still holds if there are no such nonlinear obstructions in the normal form:

Theorem 3 (Pliss) *Assume that:*

(i) *The nonzero elements among the $\sum_{i=1}^{n} m_i\lambda_i - \lambda_j$ satisfy Condition S.*

(ii) *Some formal normal form of F is equal to $B = B_s$.*

Then there exists a convergent transformation to normal form.

While it seemingly extends Siegel's result only to a rather narrow special setting, Pliss' theorem proved to be quite important for future developments. Pliss uses a different approach to proving convergence, via a generalized Newton method.

Fundamental insights into normal forms, and in particular into convergence and divergence problems were achieved by Bruno, starting in the 1960s; see [6,7]. His results included or surpassed much of the earlier work. Let us take a closer look at Bruno's conditions. As above, let B be in Jordan form, with eigenvalues $\lambda_1, \ldots, \lambda_n$. For $k \geq 1$ set

$$\omega_k := \min \left\{ \left| \sum m_i \lambda_i - \lambda_j \right| \neq 0 : \right.$$
$$\left. 1 \leq j \leq n,\ m_i \in \mathbb{Z}_+,\ \sum m_i < 2^k \right\}$$

Bruno introduced two arithmetic conditions:

Condition ω:

$$-\sum_{k=1}^{\infty} \frac{\ln \omega_k}{2^k} < \infty$$

Condition $\overline{\omega}$:

$$\limsup_k -\frac{\ln \omega_k}{2^k} < \infty$$

Condition ω can be shown to imply Condition S. Clearly Condition ω implies Condition $\overline{\omega}$.

Possibly in view of Pliss' theorem, Bruno introduced the following algebraic condition on the normal form:

Condition A: Some formal normal form is of the type

$$F^* = \sigma \cdot B_s$$

with σ a scalar formal power series.

Pliss' condition corresponds to Condition A with $\sigma = 1$. Moreover, one can show that Condition A (as well as Pliss' condition) is satisfied by every (formal) normal form if it is satisfied by one.

Now let us turn to Bruno's main theorems. The convergence theorem here may be expected in view of the previous results, but the divergence theorem – as well as its proof – conquers new ground. (We do not state the most general version of the divergence theorem.)

Theorem 4 (Bruno's convergence theorem) *If Condition ω and Condition A are satisfied then a convergent normal form transformation exists.*

Theorem 5 (Bruno's divergence theorem) *Assume that $\lambda_1, \ldots, \lambda_n$ do not lie in a complex half-plane with 0 in its boundary (in particular they do not satisfy the Poincaré condition). Moreover assume that an analytic vector field F^* in normal form does not satisfy a weaker version of Condition A, or that Condition $\overline{\omega}$ is not satisfied. Then there exists an analytic F with normal form F^* such that no transformation of F to (any) normal form converges.*

As for the "weaker version of Condition A", see Remark (c) below. Bruno also discusses the scenario when all eigenvalues are contained in some complex half-plane with 0 in its boundary.

Example There exists an analytic vector field

$$F(x) = \begin{pmatrix} \sqrt{2} & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{pmatrix} + \cdots$$

which admits no convergent transformation to normal form (which is just the linear part). The arithmetic conditions on the eigenvalues are satisfied, but the nontrivial Jordan block violates Condition A. For the related problem of normalizing local analytic diffeomorphisms (rather than vector fields), see e. g. DeLatte and Gramchev [16], and Gramchev [21], for similar divergence results.

Remark (a) Bruno's divergence theorem cannot be applied directly to prove divergence of all normalizing transformations for a specific given vector field. (Therefore any particular example still has to be worked "by hand".) Rather, the divergence theorem provides a generic result. Since the arithmetic conditions (both ω and $\overline{\omega}$) are very weak, one sees that in absence of the Poincaré condition, the algebraic obstructions from the formal normal form are mostly responsible for divergence.

Remark (b) Bruno's divergence theorem starts from the assumption that a convergent normal form exists for a given vector field; thus he deals with convergence or divergence of normalizing transformations. Little seems to be known about analytic vector fields that admit only divergent normal forms.

Remark (c) The version of Condition A stated here is taken from Bruno's monograph (see Chapter III, § in [7]). The original paper (p. 140 ff. in [6]). contains somewhat different versions. In the interesting case when the hypothesis of Theorem 5 holds, Bruno's original requirement on the normal form in eigencoordinates is as follows:

$$\dot{x}_j = x_j \left(\lambda_j \sigma + \overline{\lambda}_j \tau \right), \quad 1 \leq j \leq n,$$

with scalar series σ and τ. This condition seems to have been stated with special regard to real systems, because otherwise complex conjugation plays no distinguished role. To illustrate this point, consider the complexification of a real system, and multiply this through by some scalar $\exp(\mathrm{i}\theta)$ which is neither real nor purely imaginary. Convergence or divergence of normalizing transformations is not affected by this, but the shape of the condition

above changes considerably. (The appropriate setting for this seems to be the one developed by Stolovitch [39]; see below.) In later publications, notably in [7], Bruno himself mostly used Condition A in the simple form we stated above.

Lie Algebra Arguments

Bruno's theorems set the standard against which later results are measured. However, they do not directly address this basic question: Given a particular local analytic vector field F, characterize properties that are necessary (and, ideally, also sufficient) for the existence of a convergent normal form transformation. (Obviously, Condition A is not generally necessary for the existence of a convergent normalizing transformation.) To answer this basic question and related problems, it turns out helpful to consider Lie algebras of analytic vector fields.

A possible approach is based on the earlier observation that normal forms admit symmetries. (See also the entry on Symmetry and Perturbation Theory in this section.) The coordinate-free approach proves to be quite suitable here. First, let us formalize the observation:

Lemma 2 *If F admits a convergent normalizing transformation then there exists a nontrivial G (i. e., $G \notin \mathbb{K} \cdot F$) such that $[G, F] = 0$.*

Proof There exist a convergent Ψ and a convergent F^* in normal form such that

$$F^* \xrightarrow{\Psi} F$$

Now Ψ sends B_s to some analytic $G = B_s + \ldots$, and $[B_s, F^*] = 0$ implies $[G, F] = 0$. If $F^* \neq B_s$, we are done. If $F^* = B_s$ then take some linear map $C \notin \mathbb{K} \cdot B_s$ that commutes with B_s, and define G by $C \xrightarrow{\Psi} G$. □

In other words, if there is a convergent normalizing transformation then there exists a nontrivial infinitesimal symmetry at the stationary point. One can try to turn this necessary condition around and thus obtain sufficient convergence criteria. This has been done, with some success, since the early 1990s. Among the relevant contributions are those by Markhashov [29], Bruno and Walcher [9], Cicogna [11,12], Bambusi et al. [3], and, from a different starting point, Stolovitch [39] (see below and see also Gramchev and Yoshino [24] for maps). The survey paper [14] by Cicogna and Walcher collects the development up to 2001.

Here we will present a few results to give the reader an impression of the arguments employed.

The objects to deal with are $C_{\text{for}}(F)$ and $C_{\text{an}}(F)$, i. e., the formal, respectively analytic, centralizer of F, which by definition consist of all formal, respectively analytic, vector fields H such that $[H, F] = 0$. The first result is due to Cicogna [11,12] and Walcher [45]. Note that Pliss' theorem (and thus Condition A) plays a crucial role in the proof.

Theorem 6 *Given the analytic vector field F with formal normal form $\widehat{F}$, assume that*

$$\dim C_{for}(\widehat{F}) = k < \infty \,.$$

If the eigenvalues of B satisfy Condition ω and $\dim C_{an}(F) = k$ then there exists a convergent transformation to normal form.

Proof We have $\dim C_{\text{for}}(F) = \dim C_{\text{for}}(\widehat{F})$, so $\dim C_{\text{an}}(F) = k$ implies $C_{\text{an}}(F) = C_{\text{for}}(F)$. Given a formal transformation Ψ with $\widehat{F} \xrightarrow{\Psi} F$, there exists an analytic H such that $B \xrightarrow{\Psi} H$ and $[H, F] = 0$, since $B \in C_{\text{for}}(\widehat{F})$. Note that $H = B + \cdots$, so B is a normal form of H. Due to Pliss' theorem, there is a convergent Φ with $B \xrightarrow{\Phi} H$. Now $\widetilde{F} \xrightarrow{\Phi} F$ for some $\widetilde{F}$, and $[B, \widetilde{F}] = 0$, so $\widetilde{F}$ is in normal form. □

The dimension of the formal centralizer is computable in many cases. The requirement on Condition ω can be relaxed; see [11,45] and [14]. The next result is due to Markhashov [29] for the non-resonant case, and to Bruno and Walcher [9] in the resonant case. One may base a proof on the observation that $\dim C_{\text{for}}(F)$ is infinite only if Condition A holds, and use Theorem 6 otherwise.

Theorem 7 *In dimension $n = 2$, there is a convergent transformation of F to normal form if and only if F admits a nontrivial commuting vector field in 0.*

In dimension two, there are other, very precise, characterizations of resonant vector fields (for λ_2/λ_1 a negative rational number) admitting a convergent normalization, which can be drawn from the work of Martinet and Ramis [30]. Beyond this, building on work of Ecalle [18,19] and Voronin [42], Martinet and Ramis succeeded in giving an analytical classification of germs of such vector fields, and those admitting a convergent normalizing transformation can be characterized by the vanishing of infinitely many analytical invariants. In this sense, the convergence problem was settled earlier, at least for the interesting cases. But Theorem 7 approaches the question from a different perspective, gives an algebraic characterization and provides structural insight that is not directly available from [30].

Lie algebra arguments also play a fundamental role, from a different perspective, in the work of Stolovitch [39].

We will here give a simplified and incomplete account of his important results; a full presentation would take up much more space. To motivate Stolovitch's approach, note that in many cases there are natural decompositions $B = B_1 + \cdots + B_\ell$, with all $[B_i, B_j] = 0$ which come from eigenvalues splitting up into groups of pairwise commensurable ones, with pairwise incommensurability between the groups. The resonance conditions $\sum m_j \lambda_j - \lambda_k = 0$ then split up into conditions involving only the groups. We illustrate this with a simple example for such a decomposition:

$$\operatorname{diag}(1, -1, \sqrt{2}, -\sqrt{2}) = \operatorname{diag}(1, -1, 0, 0) + \operatorname{diag}(0, 0, \sqrt{2}, -\sqrt{2}) .$$

The setting considered by Stolovitch is as follows: Given (complex) analytic vector fields

$$F_1 = B_1 + \cdots , \quad \ldots , \quad F_\ell = B_\ell + \cdots$$

that commute pairwise, thus all $[F_i, F_j] = 0$, ask about simultaneous analytic normalization of these vector fields. There are sensible notions of normal form of a vector field with respect to a linear Lie algebra, and in particular there is a natural extension of Poincaré–Dulac for abelian Lie algebras of diagonal matrices; see [39], Sect. "Convergence and Convergence Problems": Each semisimple linear part $B_{i,s}$ commutes with each F_j in such a normal form. To avoid trivial scenarios, Stolovitch requires the semisimple parts $B_{i,s}$ to be linearly independent. To formulate diophantine conditions extending Bruno's Condition ω, one may proceed as follows: Let $\lambda_{i,1}, \ldots, \lambda_{i,n}$ denote the eigenvalues of B_i. For nonnegative integers $m_1, \ldots, m_n$, and for $1 \le d \le n$ set

$$\gamma_{m_1,\ldots,m_n,d} := \sum_i \left| \sum_j m_j \lambda_{i,j} - \lambda_d \right|$$

and (for instance; Stolovitch gives a more general formulation)

$$\omega_k(B_1, \ldots, B_\ell) := \inf \left\{ \gamma_{m_1,\ldots,m_n,d} \neq 0 : 1 \le d \le n,\ 2 \le \sum m_j \le 2^k \right\} .$$

This leads to an appropriate diophantine condition which extends Bruno's Condition ω.

Condition $\omega^\#$:

$$-\sum_{k=1}^{\infty} \frac{\omega_k(B_1, \ldots, B_\ell)}{2^k} < \infty$$

We remark briefly that there appear to be some issues of well-definedness here. For instance, the choice of the F_i, and hence of the B_i, is not unique, and one has to verify that the important notions, like Condition $\omega^\#$, do not depend on these choices. Stolovitch [39] works in an invariant setting from the start; as a consequence, no such questions arise.

Finally, Stolovitch introduces the notion of *formal complete integrability*. Disregarding technical subtleties (even though these are relevant and of interest), one may informally characterize this property as follows: The system $F_1, \ldots, F_\ell$ is formally completely integrable if it has as many formal integrals as admissible by the semisimple linear parts $B_{1,s}, \ldots, B_{\ell,s}$. To cast at least some light on this, we note that every formal integral of the system in normal form is also a simultaneous first integral of the $B_{i,s}$. See Walcher [43] for the case $\ell = 1$ and Stolovitch [39], Sect. "Lie Algebra Arguments" for the general case. This notion of complete integrability is the appropriate extension of Bruno's Condition A.

The following is a simplified representative of the results in [39]. The system $B_1, \ldots, B_\ell$ is said to have *small divisors* if 0 is a limit point of the $\gamma_{m_1,\ldots,m_n,d} \neq 0$.

Theorem 8 (Stolovitch) *In the presence of small divisors, if Condition $\omega^\#$ holds then every formally completely integrable system $F_1, \ldots, F_\ell$ admits a convergent transformation to normal form.*

The principal value of Stolovitch's results is that they open up a unified approach to a number of applications. One almost immediate application is a recovery of Bruno's convergence theorem. (See also Remark (c) following Theorem 5.) We give two more applications. For the first one, compare also Bambusi et al. [3].

Corollary 1 (Linearization) *In the presence of small divisors, assuming that Condition $\omega^\#$ holds, every formally linearizable system $F_1, \ldots, F_\ell$ admits a convergent linearization.*

The second application is a short and clear proof of a theorem due to Vey [40]:

Corollary 2 (Volume-preserving vector fields) *Assume that $\ell = n - 1$ and that $F_1, \ldots, F_{n-1}$ are commuting volume-preserving vector fields, with diagonal and linearly independent linear parts $B_i = B_{i,s}$. Then the F_i are simultaneously analytically normalizable.*

Remark (a) For similar results see Zung [46]. Zung uses a different approach based on a convergence result (a precursor of [47]) for Hamiltonian vector fields; see also the section on Hamiltonian vector fields below. His argument seems somewhat sketchy.

Remark (b) There is a nontrivial overlap of Stolovitch's results with the approach to convergence via symmetries outlined above: Some results can be proven by either method, as the references indicate.

We finish this section with a third aspect of Lie algebra arguments in convergence proofs. The following is taken from Walcher [44]:

Theorem 9 *Let $\mathcal{L}$ be a finite dimensional Lie algebra of polynomial vector fields which is graded in the sense that it contains all homogeneous parts of each of its elements, and let $F \in \mathcal{L}$ and $F(0) = 0$. Then F admits a convergent transformation to normal form F^*, and one may take $F^* \in \mathcal{L}$.*

The basic idea for the proof is to take suitable solutions h_r of the homological equations, which can be chosen in $\mathcal{L}$, and transformations $\exp(h_r)$; see [44], Prop. 2.7. Due to finite dimension of $\mathcal{L}$, finitely many such transformations suffice. There are several interesting Lie algebras among the finite dimensional graded ones:

Example (a) Every projective vector field, as well as every conformal vector field in dimension ≥ 3, admits a convergent transformation to Poincaré–Dulac normal form.

Example (b) Matrix Riccati equations: Every matrix differential equation of the form

$$\dot{x} = xax + bx + xc\,,$$

with x, a, b, c matrices of appropriate sizes, admits a convergent transformation to normal form.

NFIM and Sets of Analyticity

As we have seen, convergence problems for normalizing transformations are unsurmountable in many instances. But there are sensible strategies to achieve convergence by relaxing the requirements on the normalized vector field. Moreover, such strategies frequently provide interesting information, e. g. on stability or on the existence of periodic solutions. There are two related, but not equivalent, approaches to be discussed here: Bibikov's *normal form on an invariant manifold (NFIM)*, see [4]; and Bruno's *sets of analyticity*, see [7].

We will first discuss Bibikov's work, including a coordinate-free approach proposed in [44].

Definition 3 Let C be a semisimple linear map. A vector subspace of $\mathbb{K}^n$ is called *strongly C-stable* if it is invariant for every vector field $F = B + \cdots$, with $B_s = C$, in normal form.

Strongly C-stable spaces are in particular C-stable. A coordinate-dependent characterization is as follows.

Lemma 3 *Assume that C is in diagonal form, and let $1 < r < n$. Then the subspace $U := \mathbb{K}e_1 + \cdots + \mathbb{K}e_r$ is strongly C-stable if and only if*

$$m_1\lambda_1 + \cdots + m_r\lambda_r - \lambda_j \neq 0$$

for all nonnegative integers m_i with $\sum m_i > 0$, and all $j \in \{r+1, \ldots, n\}$.

Here, the choice of indices $1, \ldots, r$ is just for the sake of convenience. The proof is simple: The condition ensures that no monomial

$$x_1^{m_1} \ldots x_r^{m_r} e_j, \quad j > r$$

will occur in the normal form, and therefore the subspace U, characterized by $x_{r+1} = \cdots = x_n = 0$, is invariant for any normal form $F = C + B_n + \cdots$. Examples include the stable, unstable and center subspaces of C.

Now one can introduce the notion of NFIM:

Definition 4 Assume that $U = \mathbb{K}e_1 + \cdots + \mathbb{K}e_r$ is strongly C-stable. A vector field $F = B + \cdots$ with $B_s = C$ is said to be in *normal form on the invariant manifold U (NFIM on U)* if U is invariant for F and furthermore

$$[B_s|_U, F|_U] = 0\,.$$

Example The two-dimensional vector field

$$F = \begin{pmatrix} x_1^2 - 2x_1x_2 + 3x_1^3 \\ x_2(1 + x_1 + x_2^2) \end{pmatrix}, \text{ with } Bx = \begin{pmatrix} 0 \\ x_2 \end{pmatrix},$$

is in NFIM on $U = \mathbb{K}e_1$.

Bibikov also introduced a refined version of NFIM, which he calls quasi-normal form (QNF). The above example is not in quasi-normal form; in a QNF the first entry would contain only functions of x_1.

Because normal forms are, in particular, in NFIM on every strongly B_s-stable subspace, it is obvious that formal transformations to NFIM exist. But there is more freedom to construct such transformations, which may be utilized to force convergence. To give a flavor of Bibikov's results, we present a weakened version of one of his theorems.

Theorem 10 (Bibikov) *Given a vector field $F = B + \cdots$, assume that the subspace $U := \mathbb{K}e_1 + \cdots + \mathbb{K}e_r$ is strongly B_s-stable, and moreover that:*

(i) *There is an $\epsilon > 0$ such that*

$$|m_1\lambda_1 + \cdots + m_r\lambda_r - \lambda_j| \geq \epsilon$$

for all nonnegative integers m_i, $\sum m_i > 0$, and for all $j > r$.

(ii) *Some formal normal form $\widehat{F}|_U$ satisfies the Pliss condition on U.*
Then there exists a convergent transformation to NFIM on U.

Remark In Bibikov's original theorems (see [4], Theorem 3.2 and Theorem 10.2), one finds a more general (quite technical) condition instead of (ii). Thus the range of Bibikov's theorems is wider than our statement indicates. Condition (i), or some related condition, cannot be discarded completely: There are examples of strongly C-stable subspaces which do not correspond to analytic invariant manifolds for certain vector fields. (This is another incarnation of the small denominator problem.)

Applications of Bibikov's theorems include the existence of analytic stable and unstable manifolds; stability of the stationary point in case $\lambda_1 = 0$ when all other λ_i have negative real parts, and $\widehat{F}|_U = 0$ on $U = \mathbb{K}e_1$ ("transcendental case"); and the existence of certain periodic solutions.

Let us now turn to Bruno's method of analytic invariant sets; see Bruno Part I, Ch. III, and Part II in [7]. To motivate the approach, one may use the following observation: For an analytic vector field $\widehat{F} = B + \ldots$ in normal form the set

$$\mathcal{A} = \left\{z : \widehat{F}(z) \text{ and } B_s z \text{ linearly dependent in } \mathbb{K}^n\right\}$$

is invariant for $\widehat{F}$. (This is a consequence of $[B_s, \widehat{F}] = 0$: The set of points for which a vector field and an infinitesimal symmetry "point in the same direction" is invariant.) It is clearly possible to write down analytic functions such that $\mathcal{A}$ is their common zero set: For instance, take suitable 2×2-determinants. One may now introduce the notion that a vector field $\tilde{F}$ is *normalized on* $\mathcal{A}$: In the coordinate version this means that for each entry of the right-hand side the sum of all nonresonant parts lies in the defining ideal of $\mathcal{A}$.

If F is not in normal form but some formal normal form $\widehat{F}$ satisfies Condition A then – assuming some mild diophantine conditions – $\mathcal{A}$ is a whole neighborhood of the stationary point, according to Bruno's convergence theorem. Bruno now refines this by investigating whether the generally "formal" set $\mathcal{A}$ is analytic (or at least certain subsets are), and what can be said about the solutions on such subsets. To be more precise: Given a not necessarily convergent normal form $\widehat{F}$ of F, one can still write down formal power series whose "common zero set" defines $\mathcal{A}$, and it makes sense to ask what of this can be salvaged for analyticity. The following is a sample of Bruno's results (see Part I, Ch. III, Theorem 2 and Theorem 4 in [7]). As above, we do not write down the technical conditions completely.

Theorem 11 (Bruno) *Let the analytic vector field $F = B + \cdots$ be given.*

(a) *If the eigenvalues $\lambda_1, \ldots, \lambda_n$ of B are commensurable (i. e., pairwise linearly dependent over the rationals) then the set $\mathcal{A}$ is analytic and there is a convergent transformation to a vector field $\tilde{F}$ that is normalized on $\mathcal{A}$.*
(b) *Generally, there exists a (formal) subset $\mathcal{B}$ of $\mathcal{A}$ which is analytic, and for $\mathcal{B}$ the same conclusion as above holds.*

For applications see Bruno [7], Part II, and the recent papers by Edneral [20], and Bruno and Edneral [8], on existence of periodic solutions for certain equations.

Hamiltonian Systems

Hamiltonian systems have a special position among differential equations (see e. g. [2] for an overview, and ► Hamiltonian Perturbation Theory (and Transition to Chaos)); in view of their importance it is appropriate to give them particular attention. When discussing normal forms of Hamiltonian systems (where Poincaré–Dulac becomes Birkhoff; see [5]) it is natural to consider canonical transformations only. In view of the correspondence between integrals of F and Hamiltonian vector fields commuting with F, it is furthermore natural to consider integrals in the Hamiltonian setting. As for convergence results, we first state two theorems by H. Ito [26,27], from around 1990:

Theorem 12 (Ito; non-resonant case) *Let $F = B + \cdots$ be Hamiltonian and let $(\omega_1, -\omega_1, \ldots, \omega_r, -\omega_r)$ be the eigenvalues of B. Moreover assume that $\omega_1, \ldots, \omega_r$ are non-resonant, thus $\sum m_j\omega_j = 0$ for integers $m_1, \ldots, m_r$ implies $m_1 = \cdots = m_r = 0$. If F possesses r independent integrals in involution (i. e. with vanishing Poisson brackets) then there exists a convergent canonical transformation of f to Birkhoff normal form.*

This condition is also necessary in the non-resonant case: If there is a convergent transformation to analytic normal form $\widehat{F}$ then there are r linearly independent linear Hamiltonian vector fields that commute with $\widehat{F}$, and these, in turn, correspond to r independent quadratic integrals of $\widehat{F}$. For the "single resonance" case one has:

Theorem 13 (Ito; a simple-resonance case) *Let $F = B + \cdots$ be Hamiltonian and let $(\omega_1, -\omega_1, \ldots, \omega_r, -\omega_r)$ be the eigenvalues of B. Moreover assume that there are*

nonzero integers n_1, n_2 such that $n_1\omega_1 + n_2\omega_2 = 0$, but there are no further resonances. If F possesses r independent integrals in involution then there exists a convergent canonical transformation to Birkhoff normal form.

Again, the condition is also necessary. Ito's proofs are quite long and intricate. Kappeler, Kodama and Nemethi [28] proved a generalization of Theorem 13 for more general single-resonance cases. Moreover, they showed that there is a natural obstacle to further generalizations, since there exist non-integrable (polynomial) Hamiltonian systems in normal form. Thus, a complete integrability condition is not generally necessary for convergence.

Nguyen Tien Zung [47] recently succeeded in a far-reaching generalization of Ito's theorems. Considering the existence of non-integrable normal forms, this seems the best possible result on integrability and convergence.

Theorem 14 (Zung) *Any analytically integrable Hamiltonian system near a stationary point admits a convergent transformation to Birkhoff normal form.*

One remarkable feature of Zung's proof is its relative shortness, compared with the proofs by Ito.

Finally, turning to divergent normal forms (rather than normalizing transformations) of Hamiltonian systems, Perez–Marco [33] recently established a theorem about convergence or generic divergence of the normal form in the non-resonant scenario. Although numerical computations indicate the existence of analytic Hamiltonian vector fields which admit only divergent normal forms, there still seems to be no example known. Perez–Marco showed that if there is one example then divergence is generic.

Future Directions

There are various ways to extend the approaches and results presented above, and clearly there are unresolved questions. In the following, some of these will be listed, respectively, recalled.

The convergence problem for normal forms and normalizing transformations is part of the much bigger problem of analytic classification of germs of vector fields. Except for dimension two (see the references [18,19,30] mentioned above) little seems to be known in the case of nontrivial formal normal forms. Going beyond analyticity, an interesting extension would be towards Gevrey spaces. Some work on this topic exists already; see e. g. [22].

Passing from vector fields to maps, matters turn out to be much more complicated in the case of nontrivial formal normal forms, and even one-dimensional maps show very rich behavior; see [31,32]. A brief introduction is given in the survey [23] mentioned above.

A complete ("algebraic") understanding of Bruno's Condition A would be desirable; this could also provide an approach to a non-Hamiltonian version of Zung's Theorem 14. It seems well possible that all the necessary ingredients for this endeavor are contained in Stolovitch's work [39].

An extension or refinement of Bibikov's and Bruno's results on the existence of certain invariant sets (in the case of non-convergence) would obviously be interesting. There seems to be a natural guideline here: Check what invariant sets are forced onto a system in PDNF and see which ones can be salvaged. (The existence of a commuting vector field, for instance, has more consequences than those exploited by Bruno in the arguments leading up to Theorem 11.)

Finally, one could turn to more refined versions of normal forms, such as normal forms with respect to a nilpotent linear part (see [15]), and quite general constructions such as presented by Sanders [36,37]. It seems that little attention has been paid to convergence questions for such types of normal forms. For normal forms with respect to a nilpotent linear part, there are obviously no small denominator problems, but algebraic obstructions abound. There exists a precise algebraic characterization for such normal forms, involving the representation theory of $sl(2)$ (see [15]). Thus there may be some hope for an algebraic characterization of convergently normalizable vector fields.

Bibliography

1. Arnold VI (1982) Geometrical methods in the theory of ordinary differential equations. Springer, Berlin
2. Arnold VI, Kozlov VV, Neishtadt AI (1993) Mathematical aspects of classical and celestial mechanics. In: Arnold VI (ed) Dynamical Systems III, Encyclop Math Sci, vol 3, 2nd edn. Springer, New York
3. Bambusi D, Cicogna G, Gaeta G, Marmo G (1998) Normal forms, symmetry and linearization of dynamical systems. J Phys A 31:5065–5082
4. Bibikov YN (1979) Local theory of nonlinear analytic ordinary differential equations. Lecture Notes in Math 702. Springer, New York
5. Birkhoff GD (1927) Dynamical systems, vol IX. American Mathematical Society, Colloquium Publications, Providence, RI
6. Bruno AD (1971) Analytical form of differential equations. Trans Mosc Math Soc 25:131–288
7. Bruno AD (1989) Local methods in nonlinear differential equations. Springer, New York
8. Bruno AD, Edneral VF (2006) The normal form and the integrability of systems of ordinary differential equations. (Rus-

sian) Programmirovanie 3:22–29; translation in Program Comput Software 32(3):139–144
9. Bruno AD, Walcher S (1994) Symmetries and convergence of normalizing transformations. J Math Anal Appl 183:571–576
10. Chow S-N, Li C, Wang D (1994) Normal forms and bifurcations of planar vector fields. Cambridge Univ Press, Cambridge
11. Cicogna G (1996) On the convergence of normalizing transformations in the presence of symmetries. J Math Anal Appl 199:243–255
12. Cicogna G (1997) Convergent normal forms of symmetric dynamical systems. J Phys A 30:6021–6028
13. Cicogna G, Gaeta G (1999) Symmetry and perturbation theory. In: Nonlinear Dynamics, Lecture Notes in Phys 57. Springer, New York
14. Cicogna G, Walcher S (2002) Convergence of normal form transformations: The role of symmetries. Acta Appl Math 70:95–111
15. Cushman R, Sanders JA (1990) A survey of invariant theory applied to normal forms of vectorfields with nilpotent linear part, IMA vol Math Appl 19. Springer, New York, pp 82–106
16. DeLatte D, Gramchev T (2002) Biholomorphic maps with linear parts having Jordan blocks: linearization and resonance type phenomena. Math Phys Electron J 8(2):27 (electronic)
17. Dulac H (1912) Solutions d'un systéme d'équations différentielles dans le voisinage de valeurs singuliéres. Bull Soc Math Fr 40:324–383
18. Ecalle J (1981) Sur les fonctions résurgentes I. Publ Math d'Orsay 81(5):1–247
19. Ecalle J (1981) Sur les fonctions résurgentes II. Publ Math d'Orsay 81(6):248–531
20. Edneral VF (2005) Looking for periodic solutions of ODE systems by the normal form method. In: Differential equations with symbolic computation. Trends Math, Birkhäuser, Basel, pp 173–200
21. Gramchev T (2002) On the linearization of holomorphic vector fields in the Siegel domain with linear parts having nontrivial Jordan blocks SPT. In: Symmetry and perturbation theory (Cala Gonone). World Sci Publ, River Edge, NJ, pp 106–115
22. Gramchev T, Tolis E (2006) Solvability of systems of singular partial differential equations in function spaces. Integral Transforms Spec Funct 17:231–237
23. Gramchev T, Walcher S (2005) Normal forms of maps: Formal and algebraic aspects. Acta Appl Math 87(1–3):123–146
24. Gramchev T, Yoshino M (1999) Rapidly convergent iteration method for simultaneous normal forms of commuting maps. Math Z 231:745–770
25. Iooss G, Adelmeyer M (1992) Topics in Bifurcation Theory and Applications. World Scientific, Singapore
26. Ito H (1989) Convergence of Birkhoff normal forms for integrable systems. Comment Math Helv 64:412–461
27. Ito H (1992) Integrability of Hamiltonian systems and Birkhoff normal forms in the simple resonance case. Math Ann 292:411–444
28. Kappeler T, Kodama Y, Nemethi A (1998) On the Birkhoff normal form of a completely integrable system near a fixed point in resonance. Ann Scuola Norm Sup Pisa Cl Sci 26:623–661
29. Markhashov LM (1974) On the reduction of an analytic system of differential equations to the normal form by an analytic transformation. J Appl Math Mech 38:788–790
30. Martinet J, Ramis J-P (1983) Classification analytique des équations différentielles non linéaires résonnantes du premier ordre. Ann Sci Ecole Norm Sup 16:571–621
31. Perez Marco R (1995) Nonlinearizable holomorphic dynamics having an uncountable number of symmetries. Invent Math 119:67–127
32. Perez–Marco R (1997) Fixed points and circle maps. Acta Math 179:243–294
33. Perez–Marco R (2003) Convergence and generic divergence of the Birkhoff normal form. Ann Math 157:557–574
34. Pliss VA (1965) On the reduction of an analytic system of differential equations to linear form. Differ Equ 1:153–161
35. Poincaré H (1879) Sur les propriétés des fonctions deéfinies par les équations aux differences partielles. These, Paris
36. Sanders JA (2003) Normal form theory and spectral sequences. J Differential Equations 192:536–552
37. Sanders JA (2005) Normal form in filtered Lie algebra representations. Acta Appl Math 87:165–189
38. Siegel CL (1952) Über die Normalform analytischer Differentialgleichungen in der Nähe einer Gleichgewichtslösung. Nachr Akad Wiss Göttingen, Math-Phys Kl, pp 21–30
39. Stolovitch L (2000) Singular complete integrability. IHES Publ Math 91:133–210
40. Vey J (1979) Algébres commutatives de champs de vecteurs isochores. Bull Soc Math France 107(4):423–432
41. Verhulst F (2005) Methods and applications of singular perturbations Boundary layers and multiple timescale dynamics. In: Texts in Applied Mathematics, vol 50. Springer, New York
42. Voronin SM (1981) Analytic classification of germs of conformal mappings $(\mathbb{C}, 0) \to (\mathbb{C}, 0)$. Funct Anal Appl 15:1–13
43. Walcher S (1991) On differential equations in normal form. Math Ann 291:293–314
44. Walcher S (1993) On transformations into normal form. J Math Anal Appl 180(2):617–632
45. Walcher S (2000) On convergent normal form transformations in presence of symmetries. J Math Anal Appl 244:17–26
46. Zung NT (2002) Convergence versus integrability in Poincaré–Dulac normal form. Math Res Lett 9(2–3):217–228
47. Zung NT (2005) Convergence versus integrability in Birkhoff normal form. Ann Math (2) 161(1):141–156

Phase Transitions in Cellular Automata

Nino Boccara[1,2]
[1] Department of Physics, University of Illinois, Chicago, USA
[2] CE Saclay, Gif-sur-Yvette, France

Article Outline

Glossary
Definition of the Subject
Introduction
The Domany–Kinzel Cellular Automaton
Car Traffic Models

Epidemic Models
Future Directions
Bibliography

Glossary

Phase transition In statistical physics, a phase transition is the transformation of a macroscopic system from one phase to another. Phase transitions are divided into two types. First-order phase transitions are characterized by discontinuities of first-order derivatives of the Gibbs free energy, such as the entropy or the volume whereas second-order phase transitions are characterized by power-law behaviors of second-order derivatives of the Gibbs free energy such as the specific heat or the magnetic susceptibility. These singular behaviors occur for specific values of intensive variables such as the temperature or the magnetic field. Similar behaviors have also been discovered in many-component systems.

Critical behavior Critical behavior manifests itself in many-component systems and is characteristic of a cooperative behavior of the various components. This notion has been introduced in equilibrium statistical physics for many-body systems that exhibit second-order phase transitions. In the vicinity of the transition temperature T_c, a singular physical quantity Q has a power-law behavior, that is, it behaves as $|T - T_c|^{\varepsilon}$, where ε is the critical exponent characterizing the critical behavior of Q.

Universality Despite the great variety of physical systems that exhibit equilibrium second-order phase transitions, their critical behaviors, characterized by a set of critical exponents, fall into a small number of universality classes that only depend on the symmetry of the order parameter and the space dimension. Critical behavior is universal in the sense that it does not depend upon details whose characteristic size is much less than the correlation length, such as lattice structure, range of interactions (as long as this range is finite), spin length, and so on. Since time is involved in nonequilibrium critical phenomena, universality classes are expected to offer a richer variety.

Cellular automaton A cellular automaton is a fully discrete dynamical system. It consists of a regular finite-dimensional lattice of cells, each one in a state belonging to a finite set of states. The state of each cell evolves in discrete time steps. At time t the state of a given cell is a function of the states of a finite number of neighboring cells at time $t - 1$. The neighborhood of a given cell may or may not include the cell itself. All cells evolve according to the same evolution rule which may be either deterministic or probabilistic.

Model A model is a simplified mathematical representation of a system. In the actual system, although many features are likely to be important, not all of them, however, should be included in the model. Only a few relevant features which are thought to play an essential role in the interpretation of the observed phenomena should be retained. If it captures the key elements of a complex system, a simple model, may elicit highly relevant questions.

Mean-field approximation In a many-agent system the mean-field approximation is a first attempt to understand the behavior of the system. Although rather crude, it is often useful. Ignoring space correlations between the agents and replacing local interactions by uniform long-range ones, the mean-field approximation only deals with average quantities. Historically, under the name "molecular field theory", it was first used by Pierre Weiss (1869–1940) in 1907 to build up the first simple theory of the para-ferromagnetic second-order phase transition.

Definition of the Subject

Phase transitions in cellular automata are non-equilibrium phase transitions observed in probabilistic cellular automata with absorbing states [1,2], that is, states that can be reached by the dynamics but cannot be left. Directed percolation is one of the most studied example of a non-equilibrium phase transition but many other examples such as epidemic or traffic models have attracted, since the 1990s, a considerable degree of attention.

Introduction

In this article, the notion of critical behavior will often play an important role. Critical behavior manifests itself in many-component systems and is characteristic of a cooperative behavior of the various components. This notion has been introduced in statistical physics for many-body systems such as ferromagnetic materials, alloys, or liquid helium, that exhibit second-order phase transitions; that is, a phase change as a function of a tuning parameter, such as the temperature. A ferromagnetic material (i. e., a system having a spontaneous nonzero magnetization), becomes, as its temperature is (in general) increased, paramagnetic (i. e., its magnetization in the absence of an external magnetic field is equal to zero). In an ordered alloy such as β-brass – a 50 % copper and 50 % zinc alloy – the atoms of copper and zinc are located on two identical sublattices, one sublattice containing more copper and the other more zinc. As the temperature is increased, the alloy

becomes disordered (i. e., both sublattices contain equal fractions of copper and zinc). Liquid helium, which behaves as an ordinary liquid at temperatures above 2.19 K, becomes superfluid at lower temperatures. The temperature at which these phase transitions occur is called the critical temperature, and the system at the critical temperature – more generally at the critical point – is said to be in a critical state.

At the critical point, physical quantities, such as entropy, volume, or magnetization, that are first derivatives of the Gibbs free energy are continuous, in contrast with second-order derivatives, such as the specific heat or the magnetic susceptibility, which are singular. These singular behaviors reflect the long-range nature of the correlations in the vicinity of the critical point.

Close to a second-order phase transition, correlation functions of fluctuating quantities (such as spins in the case of a para-ferromagnetic phase transition) at two different points decrease exponentially with a characteristic correlation length ξ. As the temperature T approaches the critical temperature T_c, ξ diverges as $(T - T_c)^{-\nu}$ if $T > T_c$ and $(T - T_c)^{-\nu'}$ if $T < T_c$.

This cooperative effect is characteristic of criticality. It implies that certain physical quantities either vanish or diverge as powers of $|T - T_c|$ as T approaches T_c. Today, when a many-agent system displays a power-law behavior for some observable, most researchers agree that this is a sign of some cooperative effect and a manifestation of the system complexity [3].

Despite the great variety of physical systems that exhibit second-order phase transitions, their critical behaviors, characterized by a set of critical exponents, fall into a small number of universality classes that only depend on the symmetry of the order parameter (such as the magnetization for a ferromagnet) and space dimension. Critical behavior is universal in the sense that it does not depend upon details whose characteristic size is much less than the correlation length, such as lattice structure, range of interactions (as long as this range is finite), spin length, and so on. Moreover, for a given second-order phase transition, one needs to know only a rather small number of critical exponents to determine all other exponents. For instance, in the case of a para-ferromagnetic second-order phase transition, the specific heat at constant magnetic field C_B diverges as $(T - T_c)^{-\alpha}$ if $T > T_c$ and $(T - T_c)^{-\alpha'}$ if $T < T_c$, the magnetization M, which is identically equal to zero for $T > T_c$, goes to zero as $(T_c - T)^{\beta}$ if $T < T_c$, and the isothermal susceptibility χ_T diverges as $(T - T_c)^{-\gamma}$ if $T > T_c$ and $(T - T_c)^{-\gamma'}$ if $T < T_c$. If we assume that the free energy F, close to the critical point, is a generalized homogeneous function of $T - T_c$ and M, that is, a function satisfying, for all values of λ, the relation

$$F(\lambda(T - T_c), \lambda^{\beta} M) \equiv \lambda^{2-\alpha} F(T - T_c, M) .$$

Choosing $\lambda = 1/|T - T_c|$, we can write $F(T - T_c, M)$ under the form

$$F(T - T_c, M) = |T - T_c|^{2-\alpha} f\left(\frac{M}{|T - T_c|^{\beta}}\right) ,$$

where f is a function of only one variable. From this expression it can be shown (see [4]) that the critical exponents satisfy the following so-called scaling relations

$$\alpha = \alpha', \quad \gamma = \gamma', \quad \text{and} \quad \alpha + 2\beta + \gamma = 2 .$$

An important distinguishing feature of the power-law behavior of physical quantities in the neighborhood of a critical point is that these quantities have no intrinsic scale. The function $x \mapsto \exp(-x/\xi)$ has an intrinsic scale ξ, whereas the function $x \mapsto x^a$ has no intrinsic scale: power laws are self-similar.

Quantities exhibiting a power-law behavior have been observed in a variety of disciplines ranging from linguistics and geography to medicine and economics. As mentioned above, the emergence of such a behavior is regarded as the signature of a collective mechanism.

Second-order phase transitions are always associated with a broken symmetry. That is, the symmetry group of the ordered phase (the phase characterized by a nonzero value of the order parameter) is a subgroup of the disordered phase (the phase characterized by an order parameter identically equal to zero). To an order parameter, we can always associate a symmetry-breaking field. In the presence of such a field, the order parameter has a nonzero value, and, in this case, the system cannot exhibit a second-order phase transition. In the case of an ideally simple para-ferromagnetic phase transition, the paramagnetic phase is invariant under the tridimensional rotation group whereas the ferromagnetic phase is no more invariant under that group. The corresponding broken symmetry is characterized by the nonzero value of the vector magnetization $\mathbf{M}$ which plays the role of the order parameter, and the symmetry-breaking field is the magnetic field $\mathbf{B}$ intensive conjugate parameter of $\mathbf{M}$.

The nature of the broken symmetry is not always obvious as, for instance, in the case of the normal-superfluid or normal-superconductor phase transitions, but it does exist (see [5]).

Depending upon the nature of the order parameter, a second-order phase transition exists only above a critical space dimensionality, called the lower critical space

dimension. Critical exponents, which depend upon space dimensionality, take their mean-field values (i. e., when space dependence and correlations are ignored) above another critical space dimensionality, called the upper critical space dimension. In the case of the Ising model, the lower and upper critical space dimensions are, respectively, equal to 1 and 4.

In what follows we focus on phase transitions in cellular automata. Deterministic one-dimensional cellular automaton rules are defined as follows. Let Q denote the finite set of integers $\{0, 1, \ldots, m\}$ and $s(i, t) \in Q$ represent the state at site $i \in \mathbb{Z}$ and time $t \in \mathbb{N}$; a local evolution rule is a map $f\colon Q^{r_\ell + r_r + 1} \to Q$ such that

$$s(i, t+1) = f\big(s(i-r_\ell, t), s(i-\ell+1, t), \ldots, s(i+r_r, t)\big), \tag{1}$$

where the integers r_ℓ and r_r are, respectively, the left radius and right radius of the rule f; if $r_\ell = r_r = r$, r is called the radius of the rule. The local rule f, which is a function of $n = r_\ell + r_1 + 1$ arguments, is often said to be an n-input rule. The function $S_t\colon i \mapsto s(i, t)$ is the state of the cellular automaton at time t; S_t belongs to the set $Q^{\mathbb{Z}}$ of all configurations. Since the state S_{t+1} at $t+1$ is entirely determined by the state S_t at time t and the local rule f, there exists a unique mapping $F_f\colon S \to S$, such that

$$S_{t+1} = F_f(S_t)\,, \tag{2}$$

called the cellular automaton global rule or the cellular automaton evolution operator induced by the local rule f.

Most models presented in this article will be formulated in terms of probabilistic cellular automata. In this case, the image by the evolution rule of any $(r_\ell + r_r + 1)$-block, is a discrete random variable with values in Q.

The simplest cellular automata are the so-called elementary cellular automata in which the finite set of states is $Q = \{0, 1\}$ and the rule's radii are $r_\ell = r_r = 1$. Sites in a nonzero state are sometimes said to be active. It is easy to verify that there exist $2^{2^3} = 256$ different elementary cellular automaton local rules $f\colon \{0, 1\}^3 \to \{0, 1\}$. The local rule of an elementary cellular automaton can be specified by its look-up table, giving the image of each of the eight three-site neighborhoods. That is, any sequence of eight binary digits specifies an elementary cellular automaton rule. Here is an example:

111	110	101	100	011	010	001	000
1	0	1	1	1	0	0	0

Following Wolfram [6], a code number may be associated with each cellular automaton rule. If $Q = \{0, 1\}$, this code number is the decimal value of the binary sequence of images. For instance, the code number of the rule above is 184 since

$$10111000_2 = 2^7 + 2^5 + 2^4 + 2^3 = 184_{10}\,.$$

More generally, the code number $N(f)$ of a one-dimensional $|Q|$-state n-input cellular automaton rule f is defined by

$$N(f) = \sum_{(x_1, x_2, \ldots, x_n) \in Q^n} f(x_1, x_2, \ldots, x_n) \cdot |Q|^{|Q|^{n-1} x_1 + |Q|^{n-2} x_2 + \cdots + |Q|^0 x_n}\,.$$

The Domany–Kinzel Cellular Automaton

Directed percolation refers to lattice models that mimic coffee brewing, that is, causing water to pass through a bed of ground coffee. Consider the square lattice represented in Fig. 1 in which open bonds are randomly distributed with a probability p. In contrast with the usual bond percolation problem, here bonds are directed downwards, as indicated by the arrows.

If we imagine a fluid flowing downwards from wet sites in the first row, one problem is to find the probability $P(p)$ that, following directed open bonds, the fluid will reach sites on an infinitely distant last row. There clearly exists a threshold value p_c above which $P(p)$ is nonzero (see [7]). If the downwards direction is considered to be the time direction, the directed bond percolation process may be viewed as the evolution of a two-input one-dimensional cellular automaton rule f such that

$$s(i, t+1) = f\big(s(i, t), s(i+1, t)\big) = \begin{cases} 0, & \text{if } s(i,t) + s(i+1,t) = 0\,, \\ 1, & \text{with probability } p, \text{ if } s(i,t) + s(i+1,t) = 1\,, \\ 1, & \text{with probability } 1 - (1-p)^2, \\ & \quad \text{if } s(i,t) + s(i+1,t) = 2\,. \end{cases}$$

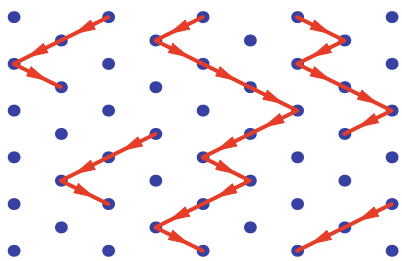

Phase Transitions in Cellular Automata, Figure 1
A configuration of directed bond percolation on a square lattice

The density of active (wet) sites ρ, which is equal to $P(p)$, plays the role the order parameter of the second-order phase transition. Using the image of the flowing fluid, it can be shown that there exists a directed bond percolation threshold (or a directed bond percolation probability) p_c^{DBP} above which the fluid has a nonzero probability of reaching an infinitely distant last row. In the vicinity of the critical probability p_c^{DBP},f if $\xi_\parallel$ and $\xi_\perp$ denote, respectively, the correlation length in the flow direction and perpendicular to it, we have

$$\xi_\parallel \sim \xi_\perp^\theta \sim \left(p - p_c^{DBP}\right)^{-\nu_\parallel} \sim \left(p - p_c^{DBP}\right)^{-\nu_\perp \theta} ,$$

where θ is the anisotropy exponent and $\nu_\parallel$ and $\nu_\perp$ the correlation length exponents in the longitudinal and transverse directions respectively. Using the finite-size renormalization group technique Kinzel and Yeomans, assuming free boundary conditions, found [8]

$$p_c^{DBP} = 0.644 \pm 0.001 \quad \theta = 1.582 \pm 0.001$$
$$\nu_\parallel = 1.739 \pm 0.002 \quad \nu_\perp 1.099 \pm 0.001 .$$

A cellular automaton n-input rule is said to be totalistic if it only depends upon the sum of the n inputs. The most general two-input one-dimensional totalistic probabilistic cellular automaton rule, called the Domany–Kinzel cellular automaton rule [9], may be written

$$s(i,t+1) = f\big(s(i,t), s(i+1,t)\big)$$
$$= \begin{cases} 0, & \text{if } s(i,t) + s(i+1,t) = 0 , \\ 1, & \text{with probability } p_1, \text{ if } s(i,t) + s(i+1,t) = 1 , \\ 1, & \text{with probability } p_2, \text{ if } s(i,t) + s(i+1,t) = 2 . \end{cases}$$

Directed bond percolation corresponds to $p_1 = p$ and $p_2 = 2p - p^2$. The case $p_1 = p_2 = p$ is also interesting; it describes the *directed site percolation* process. In this case, numerical simulations show that the directed site percolation probability $p_c^{DSP} = 0.7058 \pm 0.0005$ (values of the critical exponents characterizing the singular behavior close to the critical probability p_c^{SDP} can be found in [8]).

Domany and Kinzel showed that, in the infinite time limit, there exist two phases, an active phase in which a macroscopic fraction of all sites are occupied (state value equal to 1) and a phase in which all sites become empty (state value equal to 0). The domains of existence of these two phases in the (p_1, p_2)-plane are separated by a second-order phase transition line (see [9]). Along this line, all phase transitions belong to the same universality class characterized by a critical exponent $\beta = 0.273 \pm 0.002$, characterizing the singular behavior of the order parameter in the vicinity of the critical temperature.

A few years later, Martins, Verona de Resende, Tsallis, and de Magalhães [10] considered a generalized version of the Domany–Kinzel cellular automaton whose evolution rule is given by

$$s(i,t+1) = f\big(s(i,t), s(i+1,t)\big)$$
$$= \begin{cases} 0, & \text{if } s(i,t) = 0 \text{ and } s(i+1,t) = 0 , \\ 1, & \text{with probability } p_1, \\ & \quad \text{if } s(i,t) = 0 \text{ and } s(i+1,t) = 1 , \\ 1, & \text{with probability } p_3, \\ & \quad \text{if } s(i,t) = 1 \text{ and } s(i+1,t) = 0 , \\ 1, & \text{with probability } p_2, \\ & \quad \text{if } s(i,t) = 1 \text{ and } s(i+1,t) = 1 , \end{cases}$$

that is, a two-input one-dimensional probabilistic cellular automaton rule which is no more totalistic. In the three-dimensional (p_1, p_2, p_3)-phase space they found, in the infinite time limit, a new chaotic phase. The corresponding phase transition does not belong to the same universality class as the Domany–Kinzel phase transition. In particular, the critical exponent $\beta = 0.5 \pm 0.02$. The boundaries between the three phases of the Domany–Kinzel probabilistic cellular automaton have been determined with high accuracy in [11]. For a renormalization group approach of the Domany–Kinzel cellular automaton refer to [12].

A richer phase diagram of a simple three-input one-dimensional totalistic cellular automaton has been studied by Bagnoli, Boccara, and Rechtman [13]. These authors studied the phase diagram and the critical behavior of the one-dimensional radius-1 totalistic probabilistic cellular automaton whose evolution rule is defined as follows. If $s(i,t)$ denotes the state of the ith cell at time t, then

$$s(i,t+1)$$
$$= \begin{cases} 0, & \text{if } s(i-1,t) + s(i,t) + s(i+1,t) = 0 , \\ X_1, & \text{if } s(i-1,t) + s(i,t) + s(i+1,t) = 1 , \\ X_2, & \text{if } s(i-1,t) + s(i,t) + s(i+1,t) = 2 , \\ 1, & \text{if } s(i-1,t) + s(i,t) + s(i+1,t) = 3 , \end{cases}$$

where X_j $(j = 1, 2)$ is a Bernoulli random variable equal to 1 with probability p_j, and to 0 with probability $1 - p_j$. In the (p_1, p_2)-plane, the line $p_1 + p_2 = 1$ is a symmetry axis of the phase diagram. The evolution rule implies that states in which the cells are either all empty or all occupied

are absorbing. There exists a first-order phase transition between these two phases along the line $p_1 + p_2 = 1$ ending at a multicritical point where two second-order phase transition lines meet. These second-order transition lines separate the absorbing states mentioned above from a stable phase having a density ρ of occupied sites such that $0 < \rho < 1$, (i. e., $\rho \neq 0$ and $\rho \neq 1$). The two second-order phase transitions belong to the universality class of the directed percolation phase transition. Finally there exists a chaotic phase, located in the neighborhood of the point $(p_1, p_2) = (1, 0)$ in the (p_1, p_2)-phase plane of the type discovered by Martins, Verona de Resende, Tsallis, and de Magalhães.

Car Traffic Models

Vehicular traffic can be treated as a system of interacting particles driven far from equilibrium. The particle-hopping model describes car traffic in terms of probabilistic cellular automata. An interesting model of this type was proposed by Nagel and Schreckenberg [14]. These authors consider a finite lattice of length L with periodic boundary conditions. A finite one-dimensional cellular automaton of length L is said to satisfy periodic boundary conditions if the set of vertices is $\mathbb{Z}_L$, that is, the set of integers modulo L. In the case of a one-lane traffic model, these conditions are equivalent to assuming that cars are moving on a circular one-lane highway with neither entries nor exits. Each cell is either empty (i. e., in state e) or occupied by a car (i. e., in state v), where $v = 0, 1, \ldots, v_{\max}$ denotes the car velocity (cars are moving to the right). If d_i is the distance between cars i and $i+1$, car velocities are updated in parallel according to the following subrules.

$$v_i\left(t + \tfrac{1}{2}\right) = \min(v_i(t) + 1, d_i(t) - 1, v_{\max})$$
$$v_i(t+1) = \begin{cases} \max\left(v_i\left(t + \tfrac{1}{2}\right) - 1, 0\right), & \\ & \text{with probability } p, \\ v_i\left(t + \tfrac{1}{2}\right), & \text{with probability } 1 - p, \end{cases} \tag{3}$$

where $v_i(t)$ is the velocity of car i at time t. Then, if $x_i(t)$ is the position of car i at time t, cars are moving according to the rule

$$x_i(t+1) = x_i(t) + v_i(t+1) .$$

That is, at each time step, each car increases its speed by one unit (acceleration $a = 1$), respecting the safety distance and the speed limit. The model also includes noise: with a probability p, each car decreases its speed by one unit. Although rather simple, the model exhibits features observed in real highway traffic, that is, with increasing vehicle density, it shows a phase transition from laminar traffic flow to start-stop waves as illustrated in Fig. 2.

In order to understand, in the case of a second-order phase transition in a highway car traffic cellular automaton model, the nature of the order parameter, show how it is related to symmetry-breaking, determine the symmetry-breaking field conjugate to the order parameter, define the

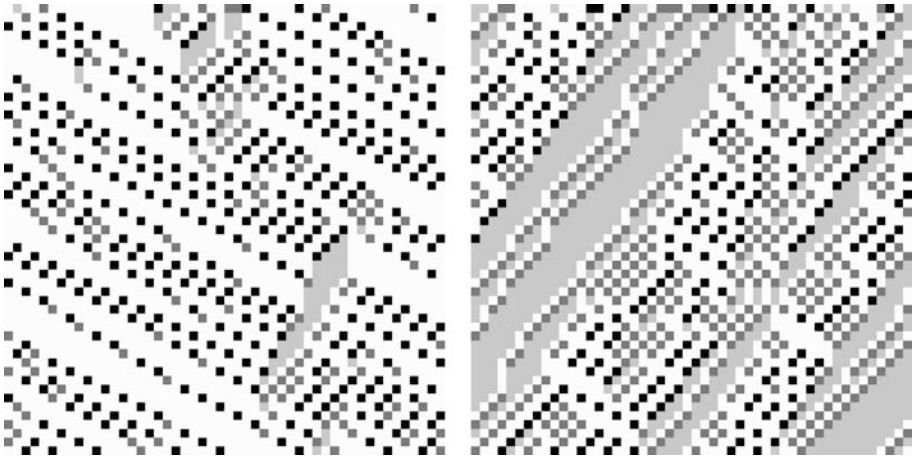

Phase Transitions in Cellular Automata, Figure 2
First 50 iterations of the Nagel–Schreckenberg probabilistic cellular automaton traffic flow model. The initial configuration is random with a density equal to 0.2 in the *left* figure and 0.5 in the *right* one. In both cases $v_{\max} = 2$ and $p = 0.2$. The number of lattice sites is equal to 50. Empty cells are white while cells occupied by a car with velocity v equal to 0, 1, and 2 have *darker shades of gray*. Time increases downwards

analogue of the susceptibility, study the critical behavior, and find scaling laws, Boccara and Fukś studied in details a deterministic version of the Nagel–Schreckenberg highway traffic model [15].

If $p = 0$, the Nagel–Schreckenberg model is deterministic and the average velocity over the whole lattice is exactly given by

$$\langle v \rangle = \min\left(v_{\max}, \frac{1}{\rho} - 1\right) . \tag{4}$$

This expression shows that, below a critical car density

$$\rho_c = 1/(v_{\max} + 1) ,$$

all cars move with a velocity equal to $v_{\max}$, while above ρ_c, the average velocity is less than $v_{\max}$.

To further simplify the Nagel–Schreckenberg model, we assume that the acceleration, which is equal to 1 in the Nagel–Schreckenberg model, has the largest possible value (less or equal to $v_{\max}$) as in the Fukui–Ishibashi model [16]. That is, in our model, we just replace (3) by

$$v_i(t + 1/2) = \min(d_i(t) - 1, v_{\max}) \tag{5}$$

Deterministic cellular automaton rules modeling traffic flow on one-lane highways are number-conserving (i. e., $\rho =$ constant). Limit sets of number-conserving cellular automata have, in most cases, a very simple structure and, these limit sets are reached after a number of time steps proportional to the lattice size [17,18,19] as illustrated in Fig. 3.

If $\rho \le \rho_c$, any configuration in the limit set consists of "perfect tiles" of $v_{\max} + 1$ cells as shown below

$v_{\max}$	e	...	e	e

in a sea of cells in state e.

If $\rho > \rho_c$, a configuration belonging to the limit set only consists of a mixture of tiles containing $v + 1$ cells of the type

v	e	...	e	e

where $v = 0, 1, \ldots, v_{\max}$. For $v < v_{\max}$, all these are said to be "defective." If $\{\rho_v \mid v = 0, 1, 2, \ldots, v_{\max}\}$ is the velocities distribution, we have

$$\rho = \sum_{v=0}^{v_{\max}} \rho_v$$

$$1 = \sum_{v=0}^{v_{\max}} (v + 1)\rho_v$$

$$\langle v \rangle = \frac{1}{\rho}\left(\sum_{v=0}^{v_{\max}} v\rho_v\right) .$$

Note that Relation (4) is a simple consequence of these relations.

If we introduce random braking, then, even at low density, some tiles become defective, which causes the average velocity to be less than $v_{\max}$. The random-braking parameter p, which is an essential ingredient of all cellular automaton traffic flow model, can, therefore, be viewed as a symmetry-breaking field, and the order parameter, conjugate to that field is

$$m = v_{\max} - \langle v \rangle \tag{6}$$

This point of view implies that the phase transition characterized by m will be smeared out in the presence of random braking as a para-ferromagnetic phase transition in the presence of a magnetic field.

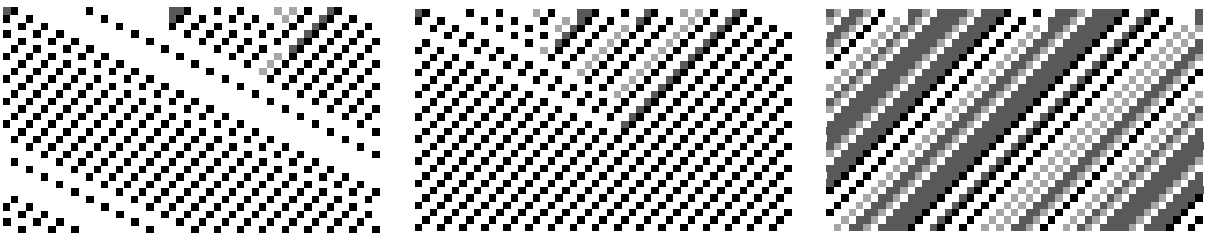

Phase Transitions in Cellular Automata, Figure 3

First 30 iterations of the deterministic cellular automaton traffic model for $v_{\max} = 2$. The critical density ρ_c is equal to 1/3. Initial configurations are random with a density exactly equal to 0.26 in the *left* figure, 1/3 in the *central* one, and 0.6 in the *right* one. The number of lattice sites is equal to 50 in the *left* and *right* figures and to 51 in the *central* figure. Empty cells are *white* whereas cells occupied by a car with velocity v equal to either 0, 1, or 2 have *darker shades of gray*. Time increases downwards. Note that for $\rho \le 1/3$, all cars move at the speed limit $v_{\max} = 2$

From (6) and (4), it follows that, for $p = 0$,

$$m = \begin{cases} 0 & \text{if } \rho \le \rho_c \,, \\ \frac{\rho - \rho_c}{\rho \rho_c} & \text{otherwise} \,. \end{cases} \tag{7}$$

The critical exponent β is, therefore, equal to 1.

The other critical exponents cannot be found exactly but can be determined using numerical simulations. In our simulations, we took $v_{\max} = 2$, used a lattice size equal to 1000 and we averaged our results over 1000 runs of 1000 iterations. For $p = 0.0005$ we took a lattice of 10 000 sites and averaged 500 runs of 10 000 iterations [15].

The susceptibility χ_ρ at constant ρ, defined by

$$\chi_\rho = \lim_{p \to 0} \frac{\partial m}{\partial p} \,. \tag{8}$$

In the limit $p \to 0$, the susceptibility diverges as $(\rho_c - \rho)^{-\gamma}$ for $\rho < \rho_c$, and as $(\rho - \rho_c)^{-\gamma'}$ for $\rho > \rho_c$. Our simulations yield

$$\gamma = 0.86 \pm 0.05 \quad \text{and} \quad \gamma' = 0.94 \pm 0.05 \,.$$

Another exponent of interest is δ. It characterizes the behavior of m as a power of p for $\rho = \rho_c$. Here again we have determined the value of

$$\lim_{p \to 0} \frac{m(\rho_c, 0) - m(\rho_c, p)}{p}$$

using numerical simulations. Our result is

$$1/\delta = 0.53 \pm 0.02$$

In is interesting to note that the values

$$\beta = 1, \quad \gamma \approx 1, \quad \delta \approx 2$$

obtained for the critical exponents are found in equilibrium statistical physics in the case of second-order phase transitions characterized by nonnegative order parameters above the upper critical dimensionality.

Close to the phase transition point, critical exponents obey scaling relations. If we assume that, in the vicinity of the critical point ($\rho = \rho_c$, $p = 0$), the order parameter m is a generalized homogeneous function of $\rho - \rho_c$ and p of the form

$$m = |\rho - \rho_c|^\beta f\left(\frac{p}{|\rho - \rho_c|^{\beta\delta}}\right) \,, \tag{9}$$

where the function f is such that $f(0) \neq 0$, then, differentiating f with respect to p and taking the limit $p \to 0$, we readily obtain

$$\gamma = \gamma' = (\delta - 1)\beta \,, \tag{10}$$

in agreement with our numerical simulations.

Boccara [20] has shown that this highway traffic flow model satisfies, with other deterministic traffic flow models, a variational principle.

Epidemic Models

The general epidemic process (see Fig. 4) describes, according to Grassberger [21,22], who studied its critical properties, "the essential features of a vast number of population growth phenomena." In its simplest version the growth process can be described as follows. Initially the cluster consists of the seed site located at the origin. At the next time step, a nearest-neighboring site is randomly chosen. This site is either added to the cluster with a probability p or rejected with a probability $1 - p$. At all subsequent time steps, the same process is repeated: a nearest-neighboring site of any site belonging to the cluster is selected at random, and it is either added to the cluster with a probability p or rejected with a probability $1 - p$. It is clear that there exists a critical probability p_c such that for $p > p_c$, the seed site has a nonzero probability of belonging to an infinite cluster. In order to determine the critical behavior of the general epidemic model, Grassberger performed extensive numerical simulations on a slightly different model that

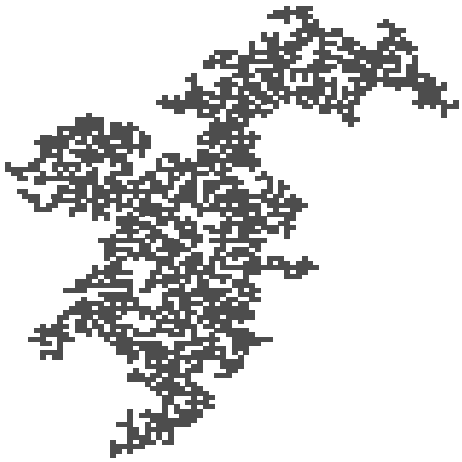

Phase Transitions in Cellular Automata, Figure 4
Grassberger's general epidemic process. The cluster represents the spread of the epidemic to 3000 sites for $p = 0.6$

belongs to the same universality class and whose static properties are identical to a bond percolation model. In this model, every lattice site of a two-dimensional square lattice is occupied by only one individual, who cannot move away from it. The individuals are either susceptible, infected, or immune. At each time step, every infected individual infects each nearest-neighboring susceptible site with a probability p and becomes immune with probability 1. For this model, the critical probability p_c is exactly equal to 1/2. If, at time $t = 0$, all the sites of one edge of the lattice are infected, among other results, Grassberger found that, at p_c, the average number of immune sites per row parallel to the initial infected row increases as a function of time as t^x, where the critical exponent x is equal to 0.807 ± 0.01 (for more detailed numerical results, refer to Grassberger [21]).

Epidemic models have a long history that started in 1927 with the publication by Kermack and McKendrick of their celebrated "threshold theorem" stating that the spread of a disease occurs only if the density of individuals susceptible to catch the disease is greater than a threshold value [23]. Here we shall only describe a model formulated in terms of cellular automata due to Boccara and Cheong [24,25] that exhibits a phase transition. In this so-called SIS epidemic model, individuals are divided into two disjoint groups:

1. *susceptible* individuals capable of contracting the disease and becoming infective, and
2. *infective* individuals capable of transmitting the disease to susceptibles.

If p_i denotes the probability for a susceptible to be infected and p_r the probability for an infective to recover and return to the susceptible group, the possible evolution of an individual may be represented by the following transfer diagram:

$$S \xrightarrow{p_i} I \xrightarrow{p_r} S\,.$$

In a two-dimensional cellular automaton model, with periodic boundary conditions, in which the sites are elements of the space $\mathbb{Z}_L \times \mathbb{Z}_L$, each site is either empty or occupied by a susceptible or an infective.

The spread of the disease is governed by the following rules.

1. Susceptible individuals become infected by contact (i. e., a susceptible may become infective with a probability p_i if, and only if, it is in the neighborhood of an infective). This hypothesis neglects incubation and latent periods: an infected susceptible becomes immediately infective.
2. Infective individuals recover and become susceptible again with a probability p_r. This assumption states that recovery is equally likely among infective individuals but does not take into account the length of time the individual has been infective.
3. The time unit is the time step. During one time step, the two preceding rules are applied synchronously, and the individuals move on the lattice according to a specific rule.
4. An individual selected at random performs a move. That is, a site occupied by an individual is selected at random and swapped with another site (either empty or occupied) also selected at random. If the second site is a nearest neighbor of the first site, the resulting move of the individual is said to be short-range whereas it is said to be long-range if it is any site of the lattice. This operation is repeated $\lfloor m \times \rho \times L^2 \rfloor$ times, where m is a positive real number called the degree of mixing and ρ the density of occupied sites at time t. When two occupied sites care swapped the move is not effective; m is therefore the average number of tentative moves. The notation $\lfloor x \rfloor$ represents the largest integer less than or equal to x.

The model assumes that the population is closed; it therefore ignores births, deaths by other causes, immigrations, or emigrations.

The critical behavior of this model has been studied by means of numerical simulations [24]. The total density of individuals was equal to 0.6, slightly above the site percolation threshold in two dimensions for the square lattice in order to be able to observe cooperative effects. Most simulations were performed on a 100×100 lattice and some on a 200×200 lattice to check possible size effects.

In the case of short-range moves, for given values of p_r and m, there exists a critical value p_i^c of the probability for a susceptible to become infected. At this transition point, the stationary density of infective individuals $I_\infty(m)$ behaves as $(p_i - p_i^c)^\beta$. When $m = 0$ (i. e., if individuals do not move), the exponent β is close to 0.6, which is the value for the two-dimensional directed percolation.

For a given value of p_r, the variations of β and p_i^c as functions of m are found to exhibit two regimes reminiscent of crossover phenomena. In the small m regime (i. e., for $m \lessapprox 10$), p_i^c and particularly β have their $m = 0$ values. In the large m regime (i. e., for $m \gtrapprox 300$), p_i^c and β have their mean-field values. In agreement with what is known in phase transition theory, the exponent β does not

seem to depend upon p_r; i. e., its value does not change along the second-order transition line.

For given values of p_i and p_r, the asymptotic behaviors of the stationary density of infective individuals $I_\infty(m)$ for small and large values of m may be characterized by the exponents

$$\alpha_0 = \lim_{m\to 0} \frac{\log\left(I_\infty(m) - I_\infty(0)\right)}{\log m} ,$$
$$\alpha_\infty = \lim_{m\to \infty} \frac{\log\left(I_\infty(\infty) - I_\infty(m)\right)}{\log m} .$$

It is found that $\alpha_0 = 0.177 \pm 0.15$ and $\alpha_\infty = -0.945 \pm 0.065$.

The fact that α_0 is rather small shows the importance of motion in the spread of a disease. The stationary number of infective individuals increases dramatically when the individuals start to move. In other words, the response $\partial I_\infty(m)/\partial m$ of the stationary density $I_\infty(m)$ to the degree of mixing m tends to ∞ when m tends to 0.

In the case of long-range moves, for a fixed value of p_r, the variations of p_i^c and β are very different from those for short-range moves. Whereas for short-range moves β and p_i^c do not vary in the small-m regime, for long-range moves, on the contrary, the derivatives of β and p_i^c with respect to m tend to ∞ as m tends to 0. For small m, the asymptotic behaviors of β and p_i^c may therefore be characterized by the exponents

$$\alpha_\beta = \lim_{m\to \infty} \frac{\log\left(\beta(m) - \beta(0)\right)}{\log m} ,$$
$$\alpha_{p_i^c} = \lim_{m\to 0} \frac{\log\left(p_i^c(m) - p_i^c(0)\right)}{\log m} .$$

Both exponents are found to be close to 0.5.

Future Directions

In the article we tried to focus on the essential characteristics of phase transitions in cellular automata illustrating our discussion with representative examples. In this section we briefly present other examples and list some articles that go deeper into the examples we chose to discuss.

One of the earliest example of a phase transition in cellular automata has been studied in 1984 by Grassberger et al. [26] who showed that the spatial patterns of two probabilistic cellular automata exhibit a transition from stability to instability of kinks between ordered states. As a function of the probability p the cellular automaton rules 94 and 50 for $p = 0$ are continuously modified to become, for $p = 1$, rules 22 and 122 respectively. In both cases the authors determined the critical probabilities and a few critical exponents.

In 1985 Kinzel [27] investigated phase transitions of probabilistic two-state three-input one-dimensional cellular automata with absorbing states. Using a transfer matrix technique, he determined phase diagrams and critical exponents. He also studied a special three-state probabilistic cellular automaton that could be mapped onto a two-state cellular automaton modeling the spread of an epidemic taking into account immunization. For a field theoretic treatment of this epidemic model, see Cardy [28].

A particular class of probabilistic two-dimensional two-state cellular automata defined on $\mathbb{Z}_L \times \mathbb{Z}_L$ (i. e., on a square lattice of size L with periodic boundary conditions) with nearest-neighbor interactions were investigated by Kaneko and Akutsu [29] who considered the evolution rule

$$s(t+1, i, j) = \begin{cases} f\left(\sigma(t, i, j), s(t, i, j)\right) , \\ \qquad\qquad \text{with probability } 1-p , \\ 1 - f\left(\sigma(t, i, j), s(t, i, j)\right) , \\ \qquad\qquad \text{with probability } p , \end{cases}$$

where $s(t, i, j)$ denotes the state at time t of the cell at site (i, j), $\sigma(t, i, j) = s(t, i, j-1) + s(t, i, j+1) + s(t, i-1, j) + s(t, i+1, j)$, and f is a function of two variables which takes the value 0 or 1. For small values of the probability p, they found a rich variety of phases.

In epidemic models we stressed the importance of individuals' motion. In our cellular automaton models, motion was modeled by a site-exchange process. That is, we considered cellular automata whose evolution rule consists of two subrules; the first one, applied synchronously, is a usual cellular automaton rule, whereas the second, applied sequentially, is a local or nonlocal exchange of two site values [30,31]. The evolution of a probabilistic site-exchange cellular automaton depends, therefore, upon two parameters, the probability p characterizing the probabilistic cellular automaton rule, and the degree of mixing m resulting from the exchange process (for the precise definition of m, refer above). Depending upon the values of these two parameters, the system exhibits a second order phase transition characterized by a nonnegative order parameter, whose role is played by the stationary density of occupied sites. When m is very large, the correlations created by the application of the probabilistic cellular automaton rule are destroyed and, as expected, the behavior of the system is then correctly described by

a mean-field-type approximation. According to whether the exchange of site values is local or nonlocal, the critical behavior is qualitatively different as m varies [32]. In [33], the authors found in the (p, m)-plane that, along the transition line, increasing m, the order of the phase transition changes from second- to first-order at a tricritical point characterized by a different critical behavior.

The Domany–Kinzel cellular automaton is a popular toy model that has attracted a lot of attention. A significant number of papers devoted to its study have been published. It has been shown that the qualitative features of the phase diagram, including the new phase found by Martins, Verona de Resende, Tsallis, and de Magalhães can be predicted analytically by going one-step beyond the mean-field approximation [34]. Although there is a clear numerical evidence that the critical behavior along the critical line found by Domany and Kinzel is that of directed percolation [1,35], this is not the case at terminal point (1/2, 1) [36,37]. The Domany–Kinzel model has also been used to illustrate the breakdown of universality in transitions to spatiotemporal chaos [38] and, recently, a few limit theorems have been rigorously established [39].

Since the early 1990s, traffic problems have drawn considerable attention and a great number of cellular automaton models of traffic flow have been proposed to deal with many diverse situations such as the existence of a jamming transition in two dimensions [40], the influence of two-level crossings on traffic jams [41], the effect of traffic accidents on the jamming transition [42], the crossing of two roads [43], the existence of a roadblock and the resulting number of stopped cars [44,45], and building up models of city traffic [46]. In the case of deterministic cellular automaton traffic flow models generalizing the cellular automaton rule 184 [47], that is, models in which the maximum speed is greater than 1, Fukś [48] has been able to derive exactly the flow diagram, that is, the graph of the car flow as a function of the car density. There exist also quite a few cellular automaton models of pedestrian traffic that exhibit phase transitions similar to those found in car traffic [49,50]. A very simple cellular automaton pedestrian model exhibiting self-organized motion in a multilane passageway with pedestrians moving in opposite directions is described in [3] page 204. For a detailed recent review on the application of statistical physics to traffic see [51]. Concerning "realistic" traffic, there exists an agent-based simulation project at Los Alamos National Laboratory called TRANSIMS (TRansportation ANalysis and SIMulation System) [52] "capable of simulating the second-by-second movements of every person and every vehicle through the transportation network of a large metropolitan area" [53].

Bibliography

Primary Literature

1. Ódor G (2004) Universality classes in nonequilibrium lattice systems. Rev Mod Phys 76:663–724
2. Hinrichsen H (2006) Non-equilibrium phase transitions. Physica A 369:1–28
3. Boccara N (2004) Modeling Complex Systems. Springer, New York
4. Boccara N (1976) Symétries Brisées. Hermann, Paris
5. Boccara N (1972) On the microscopic formulation of Landau theory of phase transition. Solid State Commun 11:131–141
6. Wolfram S (1983) Statistical physics of cellular automata. Rev Mod Phys 55:601–644
7. Kinzel W (1983) Directed percolation in Percolation, Structures and Processes. Ann Isr Phys Soc 5:425–445
8. Kinzel W, Yeomans J (1981) Directed percolation: a finite-size scaling renormalization group approach. J Phys A 14: L163–L168
9. Domany E, Kinzel W (1984) Equivalence of cellular automata to Ising models and directed percolation. Phys Rev Lett 53: 311–314
10. Martins ML, Verona de Resende HF, Tsallis C, Magalhães ACN (1991) Evidence of a new phase in the Domany–Kinzel cellular automaton. Phys Rev Lett 66:2045–2047
11. Zebende GF, Penna TJP (1994) The Domany–Kinzel cellular automaton phase diagram. J Stat Phys 74:1274–1279
12. Tomé T, de Oliveira J (1997) Renormalization group of the Domany–Kinzel cellular automaton. Phys Rev E 55:4000–4004
13. Bagnoli F, Boccara N, Rechtman R (2001) Nature of phase transitions in a probabilistic cellular automaton with two absorbing states. Phys Rev E 63:046 116
14. Nagel K, Schreckenberg M (1992) A cellular automaton model for freeway traffic. J Phys I 2:2221–2229
15. Boccara N, Fukś H (2000) Critical behavior of a cellular automaton highway traffic model. J Phys A: Math Gen 33:3407–3415
16. Fukui M, Ishibashi Y (1996) Traffic flow in a 1D cellular automaton model including cars moving with high speed. J Phys Soc Jpn 65:1868–1870
17. Boccara N, Nasser J, Roger M (1991) Particlelike structures and their interactions in spatiotemporal patterns generated by one-dimensional deterministic cellular-automaton rules. Phys Rev A 44:866–875
18. Boccara N, Fukś H (1998) Cellular automaton rules conserving the number of active sites. J Phys A: Math Gen 31:6007–6018
19. Boccara N, Fukś H (2002) Number-conserving cellular automaton rules. Fundamenta Informaticae 52:1–13
20. Boccara N (2001) On the existence of a variational principle for deterministic cellular automaton odels of highway traffic flow. Int J Mod Phys C 12:1–16
21. Grassberger P (1983) On the critical behavior of the general epidemic process and dynamical percolation. Math Biosci 63:157–172

22. Cardy JL, Grassberger P (1985) Epidemic models and percolation. J Phys A: Math Gen 18:L267–L271
23. Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. Proc Royal Soc A 115: 700–721
24. Boccara N, Cheong K (1993) Critical behaviour of a probabilistic automata network SIS model for the spread of an infectious disease in a population of moving individuals. J Phys A: Math Gen 26:3707–3717
25. Boccara N, Cheong K (1993) Automata network epidemic models. In: Boccara N, Goles E, Martínez S, Picco P (eds) Cellular Automata and Cooperative Systems. Kluwer, Dordrecht, pp 29–44
26. Grassberger P, Krause F, Von der Twer T (1984) A new type of kinetic critical phenomenon. J Phys A: Math Gen 17:L105–L109
27. Kinzel W (1985) Phase transitions in cellular automata. Z Phys B 58:229–244
28. Cardy J (1983) Field theoretic treatment of an epidemic process with immunisation. J Phys A 16:L709–L712
29. Kaneko K, Akutsu Y (1986) Phase transitions in two-dimensional stochastic cellular automata. J Phys A: Math Gen 19: L69–L75
30. Boccara N, Roger M (1993) Site-exchange cellular automata. In: Tirapegui E, Zeller W (eds) Instabilities and Nonequilibrium Structures, IV. Kluwer Dordrecht, pp 109–118
31. Boccara N, Roger M (1994) Some properties of local and nonlocal site-exchange deterministic cellular automata. Int J Mod Phys C 5:581–588
32. Boccara Nasser J, Roger M (1994) Critical behavior of a probabilistic local and nonlocal site-exchange cellular automaton. Int J Mod Phys C 5:537–545
33. Ódor G, Boccara N, Szabo G (1993) Phase-transition study of a one-dimensional probabilistic site-exchange cellular automaton. Phys Rev E 48:3168–3171
34. Kohring GA, Schreckenberg M (1992) The Domany–Kinzel cellular automaton revisited. J Phys I (France) 2:2033–2037
35. Lübeck S (2004) Universal scaling of non-equilibrium phase transitions. Int J Mod Phys B 18:3977–4118
36. Essam J (1989) Directed compact percolation: cluster size and hyperscaling. J Phys A: Math Gen 22:4927–4937
37. Dickman R, Tretyakov A (1995) Hyperscaling in the Domany–Kinzel cellular automaton. Phys Rev E 52:3218–3220
38. Bohr T, van Hecke M, Mikkelsen R, Ipsen M (2001) Breakdown of universality in transitions to spatiotemporal chaos. Phys Rev Lett 24:5482–5485
39. Katori M, Konno N, Tanemura H (2002) Limit theorems for the nonattractive Domany–Kinzel model. Ann Probab 30:933–947
40. Biham O, Middleton A, Levine D (1992) Self-organization and a dynamical transition in traffic flow models. Phys Rev A 46:R6124–R6127
41. Nagatani T (1993) Jamming transition in the traffic-flow model with two-level crossings. Phys Rev E 48:3290–3294
42. Nagatani T (1993) Effect of traffic accident on jamming transition in traffic-flow model. J Phys A: Math Gen 26:L1015–L1020
43. Ishibashi Y, Fukui M (1996) Phase diagram for the traffic model of two one-dimensional roads with a crossing. J Phys Soc Jpn 65:2793–2795
44. Boccara N, Fukś H, Zeng Q (1997) Car accidents and number of stopped cars due to road blockade on a one-lane highway. J Phys A: Math Gen 30:3329–3332
45. Sakakibara T, Honda Y, Horiguchi T (2000) Effect of obstacles on formation of traffic jam in a two-dimensional traffic network. Phys A: Stat Mech Appl 276:316–337
46. Schadschneider A, Chowdhury D, Brockfeld E, Klauck K, Santen L, Zittartz J (2000) A new cellular automata model for city traffic. In: Helbing D, Herrmann H, Schreckenberg M, Wolf DE (eds) Traffic and Granular Flow 1999: Social, Traffic, and Granular Dynamics. Springer, Berlin
47. Fukś H, Boccara N (1998) Generalized deterministic traffic rules. Int J Mod Phys C 9:1–12
48. Fukś H (1999) Exact results for deterministic cellular automata traffic models. Phys Rev E 60:197–202
49. Fukui M, Ishibashi Y (1999) Self-orgsanized phase transitions in cellular automaton models for pedestrians. J Phys Soc Jpn 68:2861–2863
50. Fukui M, Ishibashi Y (1999) Jamming transition in cellular automaton models for pedestrians on passageway. J Phys Soc Jpn 68:3738–3739
51. Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. Phys Rep 329:199–329
52. Rickert M, Nagel K (2001) Traffic simulation: Dynamic traffic assignment on parallel computers in TRANSIMS. Future Gener Comput Syst 17:637–648
53. Refer to the TRANSIMS Web site: http://transims.tsasa.lanl.gov/

Books and Reviews

Here are a few articles on different problems of phase transitions in cellular automata that might be of interest to readers wishing to go in more details.

Bagnoli F, Franci F, Rechtman R (2002) Opinion formation and phase transitions in a probabilistic cellular automaton with two absorbing states. Lecture Notes in Computer Science, vol 2493. Springer, pp 249–258

Behera L, Schweitzer F (2003) On spatial consensus formation: Is the Sznajd model different from a voter model? Int J Mod Phys C 14:1331–1354

Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. Phys Rep 329:199–329

Hołyst JA, Kacperski K, Schweitzer F (2000) Phase transitions in social impact models of opinion formation. Physica A 285: 199–210

Kerner BS, Klenov SL, Wolf DE (2002) Cellular automata approach to three-phase traffic theory. J Phys A 35:9971–10013

Maerivoet S, De Moor B (2007) Non-concave fundamental diagrams and phase transitions in a stochastic traffic cellular automaton. Eur Phys J 42:131–140

Takeuchi K (2006) Can the Ising critical behavior survive in non-equilibrium synchronous cellular automata? Physica D 223:146–150

van Wijland F (2002) Universality class of nonequilibrium phase transition with infinitely many absorbing states. Phys Rev Lett 89:190 602

Phase Transitions on Fractals and Networks

DIETRICH STAUFFER
Institute for Theoretical Physics, Cologne University, Köln, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Ising Model
Fractals
Diffusion on Fractals
Ising Model on Fractals
Networks
Future Directions
Bibliography

Glossary

Cluster Clusters are sets of occupied neighboring sites.

Critical exponent At a critical point or second-order phase transition, many quantities diverge or vanish with a power law of the distance from this critical point; the critical exponent is the exponent for this power law.

Diffusion A random walker decides at each time step randomly in which direction to proceed. The resulting mean square distance normally is linear in time.

Fractals Fractals have a mass varying with some power of their linear dimension. The exponent of this power law is called the fractal dimension and is smaller than the dimension of the space.

Ising model Each site carries a magnetic dipole which points up or down; neighboring dipoles "want" to be parallel.

Percolation Each site of a large lattice is randomly occupied or empty.

Definition of the Subject

At a phase transition, as a function of a continuously varying parameter (like the temperature), a sharp singularity happens in infinitely large systems, where quantities (e. g. the density) jump, vanish, or diverge.

Introduction

Some phase transitions, like the ferromagnetic Curie point where the spontaneous magnetization vanishes, happen in solids, and experiments often try to grow crystals very carefully such that the solid in which the transition will be observed is periodic with very few lattice faults. Other phase transitions like the boiling of water, or the liquid-vapor critical point where the density difference between a liquid and its vapor vanishes, happen in a continuum without any underlying lattice structure. Nevertheless, the critical exponents of the Ising model on a simple-cubic lattice agree well with those of liquid-vapor experiments. Impurities, which are either fixed ("quenched dilution") or mobile ("annealed dilution"), are known to change these exponents slightly, e. g. by a factor $1-\alpha$, if the specific heat diverges in the undiluted case at the critical point, i. e. if the specific heat exponent α is positive. In this review we deal neither with regular lattices nor with continuous geometry, but with phase transitions on fractal and other networks. We will compare these results with the corresponding phase transitions on infinite periodic lattices like the Ising model.

Ising Model

Ernst Ising in 1925 (then pronounced EEsing, not EYEsing) published a model which is, besides percolation, one of the simplest models for phase transitions. Each site i is occupied by a variable $S_i = \pm 1$ which physicists often call a spin but which may also be interpreted as a trading activity [10] on stock markets, as a "race" or other ethnic group in the formation of city ghettos [26], as the type of molecule in binary fluid mixtures like isobutyric acid and water, as occupied or empty in a lattice-gas model of liquid-vapor critical points, as an opinion for or against the government [29], ► Opinion Dynamics and Sociophysics, or whatever binary choice you have in mind. Also models with more than two choices, like $S_i = -1, 0$ and 1 have been investigated both for atomic spins as well as for races, opinions, ... Two spins i and k interact with each other by an energy $-JS_iS_k$ which is $-J$ if both spins are the same and $+J$ if they are the opposite of each other. Thus $2J$ is the energy to break one bond, i. e. to transform a pair of equal spins to a pair of opposite spins. The total interaction energy is thus

$$E = -J \sum_{<i,k>} S_i S_k \ , \tag{1a}$$

with a sum over all neighbor pairs. If you want to impress your audience, you call this energy a Hamiltonian or Hamilton operator, even though most Ising model publications ignore the difficulties of quantum mechanics ex-

cept for assuming the discrete nature of the S_i. (If instead of these discrete one-dimensional values you want to look at vectors rotating in two- or three-dimensional space, you should investigate the XY or Heisenberg models instead of the Ising model.)

Different configurations in thermal equilibrium at absolute temperature T appear with a Boltzmann probability proportional to $\exp(-E/k_B T)$, and the Metropolis algorithm of 1953 for Monte Carlo computer simulations flips a spin with probability $\exp(-\Delta E/k_B T)$, where k_B is Boltzmann's constant and $\Delta E = E_{\text{after}} - E_{\text{before}}$ the energy difference caused by this flip. If one starts with a random distribution of half the spins up and half down, using this algorithm at positive but low temperatures, one sees growing domains [28]. Within each domain, most of the spins are parallel, and thus a computer printout shows large black domains coexisting with large white domains. Finally, one domain covers the whole lattice, and the other spin orientation is restricted to small clusters or isolated single spins within that domain. This self-organization (biologist may call it "emergence") of domains and of phase separation appears only for positive temperatures below the critical temperature T_c and only in more than one dimension. For $T > T_c$ (or at all positive temperatures in one dimension) we see only finite domains which no longer engulf the whole lattice. This phase transition between long-range order below and short-range order above T_c is called the Curie or critical point; we have $J/k_B T_c = \frac{1}{2}\ln(1+\sqrt{2})$ on the square lattice and 0.221655 on the simple cubic lattice with interactions to the z nearest lattice neighbors; $z = 4$ and 6, respectively. The mean field approximation becomes valid for large z and gives $J/k_B T_c = 1/z$. Near $T = T_c$ the difference between the number of up and down spins vanishes as $(T_c - T)^\beta$ with $\beta = 1/8$ in two, $\simeq 0.32$ in three, and 1/2 in six and more dimensions and in mean field approximation.

We may also influence the Ising spins though an external field h by adding

$$-h\sum_i S_i \tag{1b}$$

to the energy of Eq. (1a). This external field then pushes the spins to become parallel to h. Thus we no longer have emergence of order from the interactions between the spins, but imposition of order by the external field. In this simple version of the Ising model there is no sharp phase transition in the presence of this field; instead the spontaneous magnetization (fraction of up spins minus fraction of down spins) smoothly sinks from one to zero if the temperature rises from zero to infinity.

Fractals

Fractals obey a power law relating their mass M to their radius R:

$$M \propto R^D \tag{2}$$

where D is the fractal dimension. An exactly solved example are random walks (= polymer chains without interaction) where $D = 2$ if the length of the walk is identified with the mass M. For self-avoiding walks (= polymer chains with excluded volume interaction), the Flory approximation gives $D = (d+2)/3$ in $d \le 4$ dimensions ($D(d \ge 4) = 2$ as for random walks), which is exact in one, two and four dimensions, and too small by only about two percent in three dimensions.

We now discuss the fractal dimension of the Ising model. In an infinite system at temperatures T close to T_c, the difference M between the number of up and down spins varies as $(T_c - T)^\beta$ while the correlation length ξ varies as $|T - T_c|^{-\nu}$. Thus, $M \propto \xi^{-\beta/\nu}$. The proportionality factor varies as the system size L^d in d dimensions since all spins are equivalent. In a finite system right at the critical temperature T_c we replace ξ by L and thus have $M \propto L^{d-\beta/\nu} = L^D$ with the fractal dimension

$$D = d - \beta/\nu \quad (d \le 4)\,. \tag{3a}$$

Warning: one should not apply these concepts to spin clusters if clusters are simply defined as sets of neighboring parallel spins; to be fractals at $T = T_c$ the clusters have to be sets of neighboring parallel spins connected by active bonds, where bonds are active with probability $1 - \exp(-2J/k_B T)$. Then the largest cluster at $T = T_c$ is a fractal with this above fractal dimension.

This warning is superfluous for percolation theory (see separate reviews in this encyclopedia) where each lattice site is occupied randomly with probability p and clusters are defined as sets of neighboring occupied sites. For $p > p_c$ one has an infinite cluster spanning from one side of the sample to the other; for $p < p_c$ one has no such spanning cluster; for $p = p_c$ one has sometimes such spanning clusters, and then the number of occupied sites in the largest or spanning cluster is

$$M \propto L^D; \quad D = d - \beta/\nu \quad (d \le 6) \tag{3b}$$

with the critical exponents β, ν of percolation instead of Ising models.

These were probabilistic fractal examples, as opposed to deterministic ones like the Sierpinski carpets and gaskets, which approximate in their fractal dimensions the percolation problem. We will return to them in the Sect. "Ising Model on Fractals".

Now, instead of asking how phase transitions produce fractals we ask what phase transitions can be observed on these fractals.

Diffusion on Fractals

Unbiased Diffusion

The most thoroughly investigated phase transitions on fractals are presumably random walkers on percolation clusters [17], particularly at $p = p_c$. This research was started by Brandt [7] but it was the later Nobel laureate de Gennes [16] who gave it the catchy name "ant in the labyrinth". The anomalous diffusion [5,14,22] then made it famous a few years later and may have also biological applications [13].

We put an ant onto a randomly selected occupied site in the middle of a large lattice, where each site is permanently occupied (randomly with probability p) or empty $(1 - p)$. At each time step, the ant selects randomly a neighbor direction and moves one lattice unit in this direction if and only if that neighbor site is occupied. We measure the mean distance

$$R(t) = \left\langle r(t)^2 \right\rangle^{1/2} \quad \text{or} \quad = \langle r(t) \rangle \,, \tag{4}$$

where $\mathbf{r}$ is the vector from the starting point of the walk and the present position, and $r = |\mathbf{r}|$ its length. The average $\langle \ldots \rangle$ goes over many such walking ants and disordered lattices. These ants are blind, that means they do not see from their old place whether or not the selected neighbor site is accessible (occupied) or prohibited (empty). (Also myopic ants were grown which select randomly always an occupied neighbor since they can see over a distance of one lattice unit.) The squared distance r^2 is measured by counting how often the ant moved to the left, to the right, to the top, to the bottom, to the front, or to the back on a simple cubic lattice.

The problem is simple enough to be given to students as a programming project. They then should find out by their simulations that for $p < p_c$ the above R remains finite while for $p > p_c$ is goes to infinity as $\sqrt{t}$, for sufficiently long times t. But even for $p > p_c$ it may happen that for a single ant the distance remains finite: If the starting point happened to fall on a finite cluster, then $R(t \to \infty)$ measures the radius of that cluster. Let μ be the

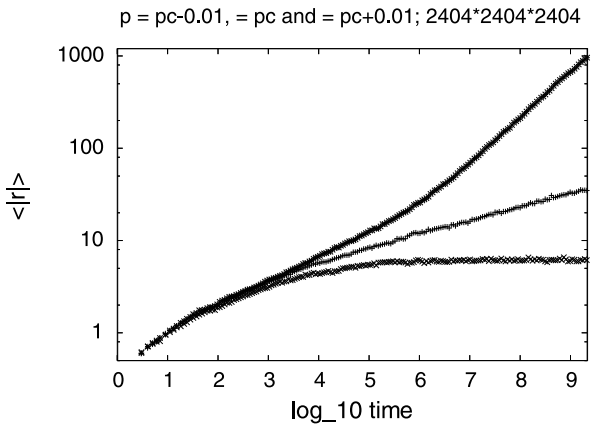

Phase Transitions on Fractals and Networks, Figure 1
Log-log plot for unbiased diffusion at (middle curve), above (upper data) and below (lower data) the percolation threshold p_c. We see the phase transition from limited growth at $p_c - 0.01$ to diffusion at $p_c + 0.01$, separated by anomalous diffusion at p_c. Average over 80 lattices with 10 walks each

exponent for the conductivity if percolation is interpreted as a mixture of electrically conducting and insulating sites. Then right at $p = p_c$, instead of a constant or a square-root law, we have anomalous diffusion:

$$R \propto t^k, \quad k = (\nu - \beta/2)/(2\nu + \mu - \beta) \,, \tag{5}$$

for sufficiently long times. This exponent k is close to but not exactly 1/3 in two and 1/5 in three dimensions. β and ν are the already mentioned percolation exponents. If we always start the ant walk on the largest cluster at $p = p_c$ instead of on any cluster, the formula for the exponent k simplifies to $\nu/(2\nu + \mu - \beta)$. The theory is explained in detail in the standard books and reviews [8,17]. We see here how the percolative phase transition influences the random walk and introduces there a transition between diffusion for $p > p_c$ and finite motion for $p < p_c$, with the intermediate "anomalous" diffusion (exponent below 1/2) at $p = p_c$. Figure 1 shows this transition on a large cubic lattice.

Biased Diffusion

Another type of transition is seen in biased diffusion, also for $p > p_c$. Instead of selecting all neighbors randomly, we do that only with probability $1 - B$, while with probability B the ant tries to move in the positive x-direction. One may think of an electron moving through a disordered lattice in an external electric field. For a long time experts discussed whether for $p > p_c$ one has a drift behavior (dis-

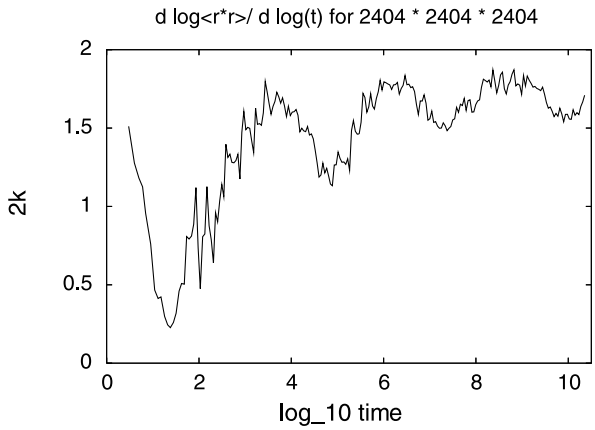

Phase Transitions on Fractals and Networks, Figure 2
Log-periodic oscillation in the effective exponent k for biased diffusion; $p = 0.725$, $B = 0.98$. The limit $k = 1$ corresponds to drift. 80 lattices with 10 walks each

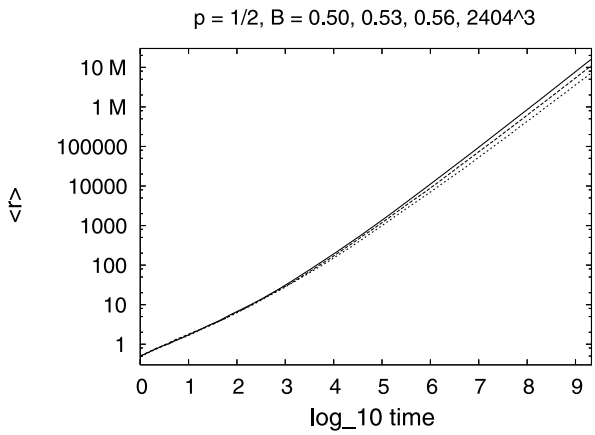

Phase Transitions on Fractals and Networks, Figure 3
Difficulties at transition from drift (small bias, upper data) to slower motion (large bias, lower data); 80 lattices with 10 walks each

tance proportional to time) for small B, and a slower motion for larger B, with a sharp transition at some p-dependent B_c. In the drift regime one may see log-periodic oscillations $\propto \sin(\text{const} \log t)$ in the approach towards the long-time limit, Fig. 2. Such oscillations have been predicted for stock markets [19], where they could have made us rich, but for diffusion they hamper the analysis. They come presumably from sections of occupied sites which allow motion in the biased direction and then end in prohibited sites [20].

Even in a region without such oscillations, Fig. 3 shows no clear transition from drift to no drift; that transition could only be seen by a more sophisticated analysis which

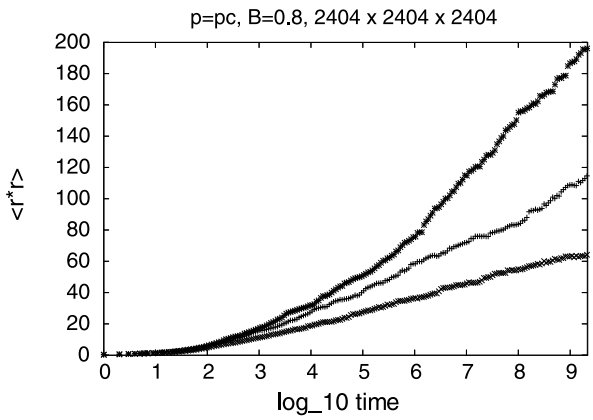

Phase Transitions on Fractals and Networks, Figure 4
Biased diffusion at $p = pc$ (middle curve) and $p = p_c \pm 0.01$ (upper and lower data) for bias $B = 0.8$; 80 lattices with 10 walks each

showed for the p of Fig. 3 that the reciprocal velocity, plotted vs. log(time), switches from concave to convex shape at $B_c \simeq 0.53$. Fortunately, only a few years after these simulations [11] the transition was shown to exist mathematically [6].

These simulations were made for $p > p_c$; at $p = p_c$ with a fractal largest cluster, drift seems impossible, and for a fixed B the distance varies logarithmically, with a stronger increase slightly above p_c and a limited distance slightly below p_c, Fig. 4.

Ising Model on Fractals

What happens if we set Ising spins onto the sites of a fractal? In particular, but also more generally, what happens to Ising spins on the occupied site of a percolation lattice, when each site is randomly occupied with probability p? In this case one expects three sets of critical exponents describing how the various quantities diverge or vanish at the Curie temperature $T_c(p)$. For $p = 1$ one has the standard Ising model with the standard exponents. If p_c is the percolation threshold where an infinite cluster of occupied sites starts to exist, then one has a second set of exponents for $p_c < p < 1$, where $0 < T_c(p) < T_c(p = 1)$. Finally, for zero temperature as a function of $p - p_c$ one has the percolation exponents as a third set of critical exponents. (If $p = p_c$ and the temperature approaches $T_c(p_c) = 0$ from above, then instead of powers of $T - T_c$ exponential behavior is expected.) In computer simulations, the second set of critical exponents is difficult to observe; due to limited accuracy the effective exponents have a tendency to vary continuously with p.

The behavior at zero temperature is in principle trivial: each cluster of occupied neighbors has parallel spins, the spontaneous magnetization is given by the largest cluster while the many finite clusters cancels each other in their magnetization. However, the existence of several infinite clusters at $p = p_c$ disturbs this argument there; presumably the total magnetization (i. e. not normalized at magnetization per spin) is a fractal with the fractal dimension of percolation theory.

Deterministic fractals, instead of the random "incipient infinite cluster" at the percolation threshold, may have a positive T_c and then allow a more usual study of critical exponents at that phase transition. Koch curves and Sierpinski structures have been intensely studied in that aspect since decades. To build a Sierpinski carpet we take a square, divide each side into three thirds such that the whole square is divided into nine smaller squares, and then we take away the central square. On each of the remaining eight smaller squares this procedure is repeated, diving each into nine squarelets and omitting the central squarelet. This procedure is repeated again and again, mathematically ad infinitum. Physicists like more to think in atoms of a fixed distance and would rather imagine each square to be enlarged in each direction by a factor three with the central square omitted; and then again and again this enlargement is repeated. In this way we grow a large structure built by squares of unit area.

Unfortunately, the phase transitions on these fractals depend on details and are not already fixed if the fractal dimension is fixed. Also other properties of the fractals like their "ramification" are important [15]; see [3] for recent work. This is highly regrettable since modern statistical physics is not restricted to three dimensions. Models were studied also in seven and in minus two dimensions, in the limits of dimensionality going to infinity or to zero, or for non-integral dimensionality. (Similarly, numbers were generalized from positive counts to negative integers, to rational and irrational numbers, and finally to imaginary/complex numbers.) It would have been nice if these fractals would have been models for these non-integral dimensions, giving one and the same set of critical exponents once their fractal dimension is known. Regrettably, we had to give up that hope.

Many other phase transitions, like those of Potts or voter models, were studied on such deterministic fractals, but are not reviewed here. We mention that also percolation transitions exist on Sierpinski structures [24]. Also, various hierarchical lattices different from the above fractals show phase transitions, if Ising spins are put on them; the reader is referred to [18,25] for more literature. As the to our knowledge most recent example we mention [21] that Ising spins were also thrown into the sandpiles of Per Bak, which show self-organized criticality.

Networks

Definitions

While fractals were a big physics fashion in the 1980s, networks are now a major physics research field. Solid state physics requires nice single crystals where all atoms sit on a periodic lattice. In fluids they are ordered only over shorter distances but still their forces are restricted to their neighbors. Human beings, on the other hand, form a regular lattice only rarely, e. g. in a fully occupied lecture hall. In a large crowd they behave more like a fluid. But normally each human being may have contacts with the people in neighboring residences, with other neighbors at the work place, but also via phone or internet with people outside the range of the human voice. Thus social interactions between people should not be restricted to lattices, but should allow for more complex networks of connections.

One may call Flory's percolation theory of 1941 a network, and the later random graphs of Erdös and Rényi (where everybody is connected with everybody, albeit with a low probability) belong to the same "universality class" (same critical exponents) as Flory's percolation. In Kauffman's random Boolean network of 1969, everybody has K neighbors selected randomly from the N participants. Here we concentrate on two more recent network types, the small-world [31] and the scale-free [1] networks of 1998 and 1999 respectively (with a precursor paper of economics Nobel laureate Simon [27] from 1955).

The small-world or Watts–Strogatz networks start from a regular lattice, often only a one-dimensional chain. Then each connection of one lattice site to one of its nearest neighbors is replaced randomly, with probability p, by a connection to a randomly selected other site anywhere in the lattice. Thus the limits $p = 0$ and 1 correspond to regular lattices and roughly random graphs, respectively.

In this way everybody may have exactly two types of connections, to nearest neighbors and to arbitrarily far away people. This unrealistic feature of small-world networks is avoided by the scale-free networks of Barabási and Albert [1], defined only through topology with (normally) no geometry involved:

We start with a small set of fully connected people. Then more people join the network, one after the other. Each new member selects connections to exactly m already

existing members of the network. These connections are not random but follow preferential attachment: The more people have selected a person to be connected with in the past, the higher is the probability that this person is selected by the newcomer: the rich get richer, famous people attract more followers than normal people. In the standard Barabási–Albert network, this probability is proportional to the number of people who have selected that person. In this case, the average number of people who have been selected by k later added members varies as $1/k^3$. A computer program is given e. g. in [28].

These networks can be undirected (the more widespread version) or directed (used less often.) For the undirected or symmetric case, the connections are like friendships: If A selects B as a friend, then B also has A as a friend. For directed networks, on the other hand, if A has selected B as a boss, then B does not have A as a boss, and the connection is like a one-way street. Up to 10^8 nodes were simulated in scale-free networks. We will now check for phase transitions on both directed and undirected networks.

Phase Transitions

The Ising model in one dimension does not have a phase transition at some positive critical temperature T_c. However, its small-world generalization, i. e. the replacement of small fraction of neighbor bonds by long-range bonds, produces a positive T_c with a spontaneous magnetization proportional to $(T_c - T)^\beta$, and a β smaller than the 1/8 of two dimensional lattices [4].

The Solomon network is a variant of the small-world network: Each person has one neighborhood corresponding to the workplace and another neighborhood corresponding to the home [23]. It was suggested and simulated by physicists Solomon and Malarz, respectively, before Edmonds [12] criticized physicists for not having enough "models which explicitly include actions and effects within a physical space as well as communication and action within a social space". Even in one dimension a spontaneous magnetization was found.

On Barabási–Albert (scale-free) networks, Ising models were found [2] for small m and millions of spins to have a spontaneous magnetization for temperatures below some critical temperature T_c which increases logarithmically with the number N of spins: $k_B T_c/J \simeq 2.6 \ln(N)$ for $m = 5$.

Here we had undirected networks with symmetric couplings between spins: actio = –reactio, as required by Newton. Ising spins on directed networks, on the other hand, have no well-defined total energy, though each single spin may be influenced as usual by its m neighbor spins. If in an isolated pair of spins i and k we have a directed interaction in the sense that spin k tries to align spin i into the direction of spin k, while i has no influence on k, then we have a perpetuum mobile: Starting with the two spins antiparallel, we first flip i into the direction of k, which gives us an energy $2J$. Then we flip spin k which does not change the energy. Then we repeat again and again these two spin flips, and gain an energy $2J$ for each pair of flips: too nice to be true. The violations of Newton's symmetry requirement makes this directed network applicable to social interactions between humans, but not to forces between particles in physics.

On this directed Barabási–Albert network, the ferromagnetic Ising spins gave no spontaneous magnetization, but the time after which the magnetization becomes zero (starting from unity) becomes very long at low temperatures, following an Arrhenius law [30]: time proportional to $\exp(\text{const}/T)$. Also on a directed lattice such Arrhenius behavior was seen while for directed random graphs and for directed small-world lattices a spontaneous magnetization was found [30]. A theoretical understanding for these directed cases is largely lacking.

Better understood is the percolative phase transition on scale-free networks (see end of Sect. "Introduction" for definition of percolation). If a fraction $1 - p$ of the connections in an undirected Barabási–Albert network is cut randomly, does the remaining fraction p keep most of the network together? It does, for large enough networks [9], since the percolation threshold p_c below which no large connected cluster survives, goes to zero as $1/\log(N)$ where N counts the number of nodes in the network. This explains why inspite of the unreliability of computer connections, the internet still allows most computers to reach most other computers in the word: If one link is broken, some other link may help even though it may be slower [1]. (For intentional [9] cuts in hierarchical networks one may have a finite percolation threshold [32].)

Future Directions

We reviewed here a few phase transitions, and ignored many others. At present most interesting for future research seem to be the directed networks, since they have been investigated with methods from computational physics even though they are not part of usual physics, not having a global energy. A theoretical (i. e. not numerical) understanding would help.

We thank K. Kulakowski for comments.

Bibliography

1. Albert R, Barabási AL (2002) Rev Mod Phys 74:47; Boccalettia S, Latora V, Moreno Y, Chavez M, Hwang D-U (2006) Phys Repts 424:175
2. Aleksiejuk A, Hołyst JA, Stauffer D (2006) Physica A 310:260; Dorogovtsev SN, Goltsev AV, Mendes JFF (2002) Phys Rev E 66:016104. arXiv:0705.0010 [cond-mat.stat-mech]; Bianconi G (2002) Phys Lett A 303:166
3. Bab MA, Fabricius G, Albano EV (2005) Phys Rev E 71:036139
4. Barrat A, Weigt M (2000) Eur Phys J B 13:547; Gitterman M (2000) J Phys A 33:8373; Pękalski A (2001) Phys Rev E 64: 057104
5. Ben-Avraham D, Havlin S (1982) J Phys A 15:L691
6. Berger N, Ganten N, Peres Y (2003) Probab Theory Relat Fields 126:221
7. Brandt WW (1975) J Chem Phys 63:5162
8. Bunde A, Havlin S (1996) Fractals and Disordered Systems. Springer, Berlin
9. Cohen R, Erez K, ben-Avraham D, Havlin S (2000) Phys Rev Lett 85:4626; Cohen R, Erez K, ben-Avraham D, Havlin S (2001) Phys Rev Lett 86:3682; Callaway DS, Newman MEJ, Strogatz SH, Watts DJ (2000) Phys Rev Lett 85:5468
10. Cont R, Bouchaud J-P (2000) Macroecon Dyn 4:170
11. Dhar D, Stauffer D (1998) Int J Mod Phys C 9:349
12. Edmonds B (2006) In: Billari FC, Fent T, Prsakwetz A, Scheffran J (eds) Agent-based computational modelling. Physica-Verlag, Heidelberg, p 195
13. Frey E, Kroy K (2005) Ann Physik 14:20
14. Gefen Y, Aharony A, Alexander S (1983) Phys Rev Lett 50:77
15. Gefen Y, Mandelbrot BB, Aharony A (1980) Phys Rev Lett 45:855
16. de Gennes PG (1976) Rech 7:916
17. Havlin S, Ben Avraham D (1987) Adv Phys 36:395; Havlin S, Ben Avraham D (2002) Adv Phys 51:187
18. Hinczewski M, Berker AN (2006) Phys Rev E 73:066126
19. Johansen A, Sornette D (1999) Int J Mod Phys C 10:563
20. Kirsch A (1999) Int J Mod Phys C 10:753
21. Koza Z, Ausloos M (2007) Physica A 375:199
22. Kutner R, Kehr K (1983) Phil Mag A 48:199
23. Malarz K (2003) Int J Mod Phys C 14:561
24. Monceau P, Hsiao PY (2004) Phys Lett A 332:310
25. Rozenfeld HD, Ben-Abraham D (2007) Phys Rev E 75:061102
26. Schelling TC (1971) J Math Sociol 1:143
27. Simon HA (1955) Biometrika 42:425
28. Stauffer D, Moss de Oliveira S, de Oliveira PMC, Sá Martins JS (2006) Biology, Sociology, Geology by Computational Physicists. Elsevier, Amsterdam
29. Sznajd-Weron K, Sznajd J (2000) Int J Mod Phys C 11:1157
30. Sánchez AD, López JM, Rodríguez MA (2002) Phys Rev Lett 88:048701; Sumour MA, Shabat MM (2005) Int J Mod Phys C 16:585; Sumour MA, Shabat MM, Stauffer D (2006) Islamic Univ J (Gaza) 14:209; Lima FWS, Stauffer D (2006) Physica A 359:423; Sumour MA, El-Astal AH, Lima FWS, Shabat MM, Khalil HM (2007) Int J Mod Phys C 18:53; Lima FWS (2007) Comm Comput Phys 2:522 and (2008) Physica A 387:1545; 3503
31. Watts DJ, Strogatz SH (1998) Nature 393:440
32. Zhang Z-Z, Zhou S-G, Zou T (2007) Eur Phys J B 56:259

Philosophy of Science, Mathematical Models in

Zoltan Domotor
University of Pennsylvania, Philadelphia, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Mathematical Models: What Are They?
Philosophical and Mathematical Structuralism
Three Approaches to Applying Mathematical Models
Validating Mathematical Models
Future Directions
Bibliography

Glossary

Philosophy of science Broadly understood, *philosophy of science* is a branch of philosophy that studies and reflects on the presuppositions, concepts, theories, arguments, methods and aims of science. Philosophers of science are concerned with general questions which include the following: What is a scientific theory and when can it be said to be confirmed by its predictions? What are mathematical models and how are they validated? In virtue of what are mathematical models representations of the structure and behavior of their target systems? In sum, a major task of philosophy of science is to analyze and make explicit common patterns that are implicit in scientific practice.

Mathematical model Stated loosely, *models* are simplified, idealized and approximate representations of the structure, mechanism and behavior of real-world systems. From the standpoint of set-theoretic model theory, a *mathematical model* of a target system is specified by a nonempty set – called the model's *domain*, endowed with some operations and relations, delineated by suitable *axioms* and intended empirical *interpretation*. No doubt, this is the simplest definition of a model that, unfortunately, plays a limited role in scientific applications of mathematics. Because applications exhibit a need for a large variety of vastly different mathematical structures – some topological or smooth, some algebraic, order-theoretic or combinatorial, some measure-theoretic or analytic, and so forth, no useful overarching definition of a mathematical model is known even in the edifice of modern category theory. It is difficult to come up with a workable

concept of a mathematical model that is adequate in most fields of applied mathematics and anticipates future extensions.

Target system There are many definitions of the concept of 'system'. Here by a *target system* we mean an effectively isolated (physical, biological, or other empirical) part of the universe – made to function or run by some internal or external causes, whose interactions with the universe are strictly delineated by a fixed (input and output) interface, and whose structure, mechanism, or behavior are the objects of mathematical modeling. Changes produced in the target system are presumed to be externally detectable via measurements of the system's characterizing quantitative properties.

Definition of the Subject

Models are indispensable scientific tools in generating information about the world. Concretely, scientists rely on models for explanation and prediction. Recent years have seen intensive attempts to extend the art of modeling and simulation to a vast range of application areas. Fostered by inexpensive computers and software, and the accelerating growth of mathematical knowledge, in contemporary applied research reference to models is far more frequent than reference to theories or principles. In applications, the term 'model' is used in a wide variety of senses, including mathematical, physical, mental, and computational models. Our focus here is on mathematical models, their structure, representational role, and validation.

Currently, many philosophers of science are interested in two major aspects of mathematical models: their *nature* and representational *role*. The question "What are mathematical models?" is answered in two fundamentally different but closely related ways:

1. In a seemingly natural way, by viewing mathematical models as families of *equations* of some kind, accompanied with certain empirical interpretations that link them to their target systems. It turns out that this simple conception, called the *received view* or the *syntactic* approach, presents several troubling representational and interpretational problems.
2. The so-called *structuralist* or *semantic* answer takes mathematical models of target systems to be suitable set-theoretic structures or generally objects in a specific dynamical (or other) category. In order to pursue this approach, model builders and users need to be able to understand how complex notions, relevant to modeling, are defined in terms of the model's structure.

One of the most useful general results on the nature of mathematical models is the following. If a given semantic formulation of a model involves also a specification of the *solution space* of some equations, then in this case the equational and structural conceptions of models are formally equivalent. Indeed, equations uniquely characterize their solution spaces and these in turn determine the associated semantic model. Conversely, if the semantic model specifies an abstract solution space, then the latter delineates a system of characterizing equations. This type of *Galois correspondence* between equations and their solutions has been established in many linear and even in some nonlinear settings.

There is a major counter to this syntactic-semantic dilemma. Structuralists are quick to point out that the mathematical models of exchange economy, *n*-person games, probability and decision, and so forth, are *autonomous* set-theoretic structures that are not associated with any system of equations. For example, recall that a classical probability model of a statistical experiment is defined by a triple, consisting of a set together with a designated field of its subsets – forming an underlying measurable space, and a probability measure thereon. Remarkably, even though in this case there are no equations to consider, thanks to the powerful *Stone–Gelfand duality* result, every probability model has an associated algebraic counterpart model, given by a linear space (thought of as the space of bounded random variables on the model's underlying measurable space) and a positive linear functional on it (viewed as the expectation functional induced by the measure). Statisticians in particular (e. g., [21]) prefer to build their models of experiments in a 'dual', computationally stronger, algebraic setting. It happens quite often that autonomous mathematical models arise in 'adjoint' or paired geometric/algebraic formulations. Characteristically, models of this nature tend to form impressively versatile 'mathematical universes' for all of the mathematics that model users may need in a given area of application.

The question "What is the role of mathematical models in scientific practice?" is answered by describing how models are conceived, constructed, and used in various applications. Since models are mathematical constructs, in addition to being objects of a formal inquiry, they are also involved in epistemic relations, expressing intended uses.

Models do not and need not match reality in all of its aspects and details to be adequate. A mathematical

model is usually developed for a specific class of target systems, and its validity is determined relative to its intended applications. A model is considered *valid* within its intended domain of applicability provided that its predictions in that domain fall within an acceptable range of error, specified prior to the model's development or identification.

To construct a mathematical model of a target system, it is typically necessary to formalize the system's decisive causal mechanisms and behavior with special regards to the model's *structural stability* and *mutability* into a larger network of models that includes not only additional cause-effect relations but also a broad range of deterministic and stochastic perturbations. In the next section we consider a simple class of dynamical models in physics that is also relevant to modeling in other disciplines.

Introduction

The subject of mature mathematical models in the form of *equations* has its roots in post-Newtonian developments of classical mechanics, hydrodynamics, electromagnetism, and kinetic theory of gases. It came on the scene of applied mathematics gradually, during the *analytic period* before 1880, thanks to the innovative efforts of great scientists, including, among many others, the Swiss mathematician Leonhard Euler (1707–1783), Italian–French mathematician Louis Joseph Lagrange (1736–1813), French astronomer-physicist Pierre Simon de Laplace (1749–1827), Scottish physicist James Clerk Maxwell (1831–1879), English physicist Lord John William Strutt Rayleigh (1842–1919), and the Austrian physicist Ludwig Boltzmann (1844–1906). It was the genius of the French mathematician Henri Poincaré (1854–1912) that generated many of our current topological and differential methods of mathematical modeling in the world of dynamical systems. Over the past 100 years or so, mathematical models have evolved to become the basic tools in a wide variety of disciplines, including not only most of physical sciences, but also chemical kinetics, population dynamics, economics, sociology, and psychology.

The many different theoretical areas of natural and social sciences have led to the development of a large assortment of mathematical models, including but not limited to descriptive vs. normative, static vs. dynamic, phenomenological vs. process-based, discrete vs. continuous, deterministic vs. stochastic, linear vs. nonlinear, finite- vs. infinite-dimensional, difference vs. differential, and topological vs. measure-theoretic models. Using category theory, efforts have been made to construct general theories of mathematical models of which models of logical systems and dynamical systems are special cases.

There has been a renewed interest also among philosophers of science in the problems of structure and function of abstract models. (See, for example, [3,5,10,13,25], and [33].) Prime questions about abstract models good many philosophers ask and attempt to answer include the following:

(i) *Ontology of models:* What, precisely, are mathematical models and how are they used in science? Are they structures in the sense of classical set theory, modern category theory, or something else, belonging, e. g., to the nebulous world of fictional entities or human constructs?

(ii) *Semantics of models:* In virtue of what are mathematical models representations of the structure, causal mechanisms and behavioral regimes of their target systems? Is it in virtue of some (possibly partial) 'isomorphisms' holding between mathematical and physical domains or by reason of certain designated 'similarities', analogies, or resemblance relations between models and aspects of the world, or because of 'empirical interpretations' of (parts of) the representing model's mathematical vocabulary – implying quantitative claims about the world that can be corroborated by empirical data, or in virtue of yet something else, such as homology or physical instantiation? Since models appear to be the main vehicles in the pursuit of scientific knowledge, philosophers are also interested in analyzing the truth conditions of semantic relations of a more general nature, such as "Researcher $\mathcal{R}$ uses model $\mathfrak{M}$ to represent target system **S** for purpose Π".

(iii) *Validation of models:* How are mathematical models validated? Is validation just a procedure in which the model's predictions are simply compared with a set of observations within the model's domain of applicability, or is it a comprehensive, all-out testing to determine the degree of agreement between the model and its target system in terms of internal structure, cause-effect relationships, and predictions?

Needless to add, these and many other questions to be examined below do not belong to science per se; they are *about* science. In this sense, philosophy of science is a second-order discipline, addressing the practices, methods and aims of the various sciences. However, it is clear from [24] that second-order questions about science are investigated also by scientists themselves.

Before considering the details of *models* of mathematical models, we begin by characterizing a particular use of the term *mathematical model* which, although not representative of the more careful formulations of the majority of philosophers of science, is nevertheless the most common practice encountered in physics, engineering and economics. Most physicists, engineers and mathematically oriented social scientists understand mathematical models to be systems of (algebraic, difference, differential, integral or other) *equations* (linking time- and spacetime-dependent quantities, and parameters), derived from first principles under various idealizing and simplifying scenarios of target systems or induced by available observational data and 'empirical laws'.

As an example illustrating the equational conception of mathematical models, we shall consider briefly the most familiar simple classical planar *pendulum model*, describing an undamped pendulum's dynamical behavior. (Additional examples can be found in [12].) It is specified by the autonomous, deterministic second-order nonlinear differential equation

$$\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} + \frac{g}{\ell}\sin\theta = 0\,,$$

in which the time-dependent *indeterminate* θ represents the target pendulum's variable angle from its downward vertical to the pivoting rod, to which a bob of mass m is attached at its swinging end. Coefficient g denotes the homogeneous gravitational acceleration acting on the pendulum's bob downward, and coefficient ℓ captures the length of the pendulum's idealized 'massless' and perfectly stiff rod. These constant coefficients are needed for individuating the target pendulum in its *classical* gravitational ambience. Under the accompanying physical interpretation, fixed by the pendulum's *idealizing scenario* (involving significant idealizations of friction, torque, resistance, and elasticity) and first-principle framework, the all-important observable quantity is the pendulum's total energy (Hamiltonian)

$$H\left(\theta, \frac{\mathrm{d}\theta}{\mathrm{d}t}\right) =_{\mathrm{df}} m\ell^2\left(\frac{1}{2}\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2 - \frac{g}{\ell}\cos\theta\right)\,.$$

The first term in the differential equation above encodes inertia and the second term stands for gravity. Recall that the pendulum's rod is suspended from a pivot point, around which it oscillates or rotates in a vertical plane without surface resistance and forcing, so that there are no extra *additive* terms in the equation for the effects of friction and torque. Coarsely speaking, in general a model is a simplified representation of a real-world system of interest for designated scientific purposes. Although the target system has many important features, not all of them can and should be included in the model for reasons of tractability and limited epistemic import. And those that are included, often involve drastic idealizations, known to be empirically false.

It has long been known that in general the foregoing differential equation does not have a closed-form solution in terms of traditional elementary functions. For general analytic solutions, Jacobi's periodic elliptic functions are needed, with values knowable only with specified degrees of accuracy from mathematical tables or computations performed by special computer programs. The gap between theoretically granted solutions and their approximate variants has led Harald Atmanspacher and Hans Primas [2] to advocate a dichotomy between *states of reality* and *states of knowledge*. In a nutshell, derivation of highly theoretical equations of motion (providing maximal information) offers a so-called *ontic* (endophysical) view of modes of being of target systems, whereas (statistical) approximation and measurement procedures give an *epistemic* (exophysical) perspective on real-world systems, involving errors and updating. It is well known that the nonlinear equation above has several geometric types of solution that capture all sorts of swinging and rotating motions, and states of rest – for the most part knowable only approximately.

Moving beyond the important ontic vs. epistemic dichotomy in modeling, note that because the foregoing pendulum equation's nonlinear component is representable by the infinite series $\sin\theta = \theta - \theta^3/(3!) + \theta^5/(5!) - \cdots$ and since for *small* deflections (less than 5°) of the pendulum's rod from the vertical the values $\sin\theta(t)$ of the angle quantity are very nearly equal to $\theta(t)$, we can substitute θ for $\sin\theta$ in the equation and forget about the higher-order polynomial terms in the infinite series. So, at the cost of obvious approximation errors, we obtain the *basic linear pendulum* differential equation $\mathrm{d}^2\theta/\mathrm{d}t^2 + g/\ell\theta = 0$ that is easy to solve. It is an elementary exercise to show that its smooth closed-form solutions are given by parametrized trigonometric position functions of the form

$$\theta(t) = \theta(0)\cos(\omega t) + \frac{\dot\theta(0)}{\omega}\sin(\omega t)$$

for all time instants t, with initial conditions $\theta(0)$ and $\dot\theta(0) =_{\mathrm{df}} \frac{\mathrm{d}\theta}{\mathrm{d}t}(t)\big|_{t=0}$, and parameter $\omega =_{\mathrm{df}} \sqrt{\ell : g}$, characterizing the pendulum's frequency of oscillation. In the study of differential equations and their solutions it is stan-

dard to adopt the so-called *flow* point of view. Simply, instead of studying a particular solution for a given initial value, the entire *solution space*

$$\left\{\theta \in \mathbf{C}^{\infty}(\mathbb{T}) \;\middle|\; \frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} + \frac{g}{\ell}\theta = 0\right\}$$

is studied, preferably in a parametrized form. Granted that the model is 'correct', this subspace of the space $\mathbf{C}^{\infty}(\mathbb{T})$ of smooth real-valued functions on a designated continuum-time domain $\mathbb{T}$ provides the investigator with complete information about the target pendulum's possible angular positions and dynamical behavior. Clearly, the foregoing solution space includes many relativistically forbidden solutions that encode physically impossible behaviors – individuated by the pendulum's superluminal velocities. The key idea here is that in the absence of complete knowledge of the model's empirical *domain of applicability*, a researcher may be in an epistemic danger of trying to physically interpret (in terms of behavior) some of the solutions – countenanced by the model, that are actually meaningless in the physical world.

To remedy the situation, the foregoing classical simple pendulum model must be extended to the *relativistic* case (studied, e. g., in [11]), having the form

$$\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} + \left(1 - \left(\frac{\ell}{\mathsf{c}}\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2\right)\frac{g}{\ell}\sin\theta = 0\,,$$

where parameter c denotes the speed of light. Because formal models usually possess limited empirical domains of applicability, in comprehensive treatments of target system behavior several different, closely related models may become necessary.

Real-world systems tend to have many representing models and these models often possess a surplus content, which supports their mutations and extensions in unexpected ways. For example, in the presence of randomness or 'noise', a classical stochastic pendulum model provides the most appropriate representation of randomly perturbed motion. A typical model of the behavior of simple pendulums affected by 'noise' is presented by the stochastic differential equation

$$\frac{\mathrm{d}^2\mathsf{X}}{\mathrm{d}t^2} + (1 + \varepsilon\mathsf{W})\omega^2\sin\mathsf{X} = 0\,,$$

where X denotes the stochastic position indeterminate, $\varepsilon > 0$ is a parameter with small values, and W is a (e. g., Gaussian) stochastic process, capturing the pendulum's random perturbation. Naturally, to extract information from a stochastic pendulum model, the investigator has to calculate the moments of target pendulum's positions or consider the transition probability density that allows to calculate the conditional probability of a future position of the pendulum's bob, given its position at a designated starting time.

In many applications, the same basic pendulum differential equation drops out of a wholly different idealizing scenario of, say, a coupled mass-spring system, consisting of a (point) mass attached to a spring at one end, where the other end of the spring is tied to a fixed frame. Other prominent examples of systems, whose dynamical behavior is also modeled by 'pendulum equations', include electric circuits, chemical reactions and interacting biological populations with oscillatory behaviors. Of course, in each application, the indeterminate and individuating parameters are interpreted differently.

These examples illustrate clearly that mathematical models are characteristically generic or *protean* with respect to real-world systems, meaning that often the same mathematical model applies to different empirical situations, admits several different empirical interpretations, and is functioning both as an investigative instrument and as an object of mathematical inquiry. Although, broadly speaking, models can be representations of a particular *token* target system (canonical examples are cosmological models) or of a *stereotype* class of systems (e. g., pendulums), they are always representations of some particular structure of a phenomenon, mechanism or behavior. Needless to add, empirically meaningful deductions from the representing equations are guided not just by the free-standing equational structure of pure mathematics but also by the accompanying empirical interpretation, encapsulated in part by the target system's idealizing scenario. Along these lines, in [23] Saunders Mac Lane asserts that "mathematics is protean science; its subject matter consists of those structures which appear (unchanged) in different scientific contexts".

On this picture, scientists construct formal models in the form of (differential, partial differential, stochastic differential, etc.) equations, proceeding in a simple *top-down* analytic fashion – using first principles (e. g., Newton's, Kirchhoff's and other laws) and idealizing scenarios or 'empirical rules' that are not part of any ambient theory, or in a more involved *bottom-up, data-driven* synthetic manner, starting from experimental (e. g., time series) data, the extant stage of knowledge, analogies with other systems, and background assumptions. In a total absence of any data or prior knowledge pertaining to the target system, it is inappropriate to consider mathematical models at all. This raises several additional foundational questions:

(i) What is at stake, if anything, in conceiving mathematical models as empirically interpreted equations of some sort?
(ii) How is it that the causal mechanisms and behaviors of real-world systems can be so effectively studied in terms of solutions of various equations that employ highly idealized assumptions, known to be false?
(iii) Since most (nonlinear) equations arising in science are solvable only approximately and in a discrete range of values, what is their epistemic import? Recall that predictions from these equations may have to be obtained by brute-force numerical methods that require more computational power than the scientists are likely to have. Thus, only a Laplacean superscientist with unlimited cognitive capacities and computational resources could make full use of such equations. In contrast, although a real-world scientist does start with such "true-in-heaven" equations but then he or she quickly modifies them to make them computationally easier to extract predictions from them.

We are now ready to start discussing some of the answers to these and other previously listed philosophical questions about mathematical models.

Mathematical Models: What Are They?

We begin, as a way of entering the subject of mathematical models by clarifying the dichotomy between *equational* (syntactic) and *structural* (semantic) conceptions of models. The standard view among most theoretical physicists, engineers and economists is that mathematical models are *syntactic* (linguistic) items, identified with particular systems of equations or relational statements. From this perspective, the process of solving a designated system of (algebraic, difference, differential, stochastic, etc.) equations of the target system, and interpreting the particular solutions directly in the context of predictions and explanations are primary, while the mathematical structures of associated state and orbit spaces, and quantity algebras – although conceptually important, are secondary.

This is a good place to recall that, contrary to the above, philosophical structuralists (e. g., Landry [18]) defend the claim that mathematical models are *structures* and category theory is the correct framework for their study. Structuralists have two major arguments against the equational (syntactic) view of mathematical models.

The first reason why the popular syntactic conception of models is troubling to philosophers of science is the following. Even though a stipulated system of equations of a target system admits infinitely many alternative formulations – linked by scale, coordinate, change-of-variable and other transformations, researchers do not presume of having different models presented by these inessentially different formulations. It is important to bear in mind that the structure of the associated *solution space*, serving as a storehouse for all model-based information about the target system's causal mechanisms and behavior, remains basically the same. Each type of *solution* of a given system of differential (difference) equations is not defined by the system of equations *per se*, but generally by a much larger (prime) *differential* (difference) *ideal* of equations, generated by it. Since solutions and their mathematical semantics are more important than change-of-variable transformations, systems of equations are related to each other in a considerably deeper way by solution-preserving transformations than by, say, scale transformations. A case in point is the familiar physical equivalence of Hamiltonian and Lagrangean formulations of equations of motion in mechanics. Though the underlying language in each case is different, nevertheless the modeling results are the same. In brief, models in the sense suggested by the practice of science possess properties equations do not seem to have.

It is well known that the basic linear second-order pendulum differential equation $\mathrm{d}^2\theta/\mathrm{d}t^2 + g/\ell\theta = 0$ can be put in the form of two linear *first-order* differential equations

$$\upsilon = \frac{\mathrm{d}\theta}{\mathrm{d}t} \quad \text{and} \quad \frac{\mathrm{d}\upsilon}{\mathrm{d}t} + \frac{g}{\ell}\theta = 0\,,$$

illustrating the notion of an equation-based *state space model*. As an aside, we note that this type of transformation is general in that any system of higher-order differential equations can be rewritten as a first-order system of higher dimensionality.

In this system of equations, the position-velocity pairs $\langle\theta(t), \upsilon(t)\rangle$ are thought to encode the target pendulum's *dynamical state* (or physical mode of being) of interest at time t. More importantly, the equations' smooth, two-dimensional parametric solutions are now given by the position-velocity function pair

$$\langle\theta_{a,b}(t), \upsilon_{a,b}(t)\rangle = \left\langle a\cos(\omega t) + \frac{b}{\omega}\sin(\omega t), b\cos(\omega t) - a\omega\sin(\omega t)\right\rangle,$$

with parameters $a = \theta(0)$ and $b = \upsilon(0)$. Thus, to specify a particular (position-velocity) state-space trajectory for the pendulum model, all we need is the *initial state* $\langle a, b\rangle$, arrived at via first and second integration of the original equation. We mention in passing that if the differential equation characterizing a target system were of order n,

then generally we would need n independent parameter values for a unique individuation of any of its solutions in the form of specific time-functions.

The point of this simple exercise in pendulum modeling is to make a headway in the direction of a powerful alternate *structuralist* formulation of the pendulum model that has the following *smooth group action* (smooth flow or smooth group representation) form

$$\mathbb{T} \times (\mathbb{R}_{/2\pi\mathbb{Z}} \times \mathbb{R}) \xrightarrow{\langle\theta,\upsilon\rangle} \mathbb{R}_{/2\pi\mathbb{Z}} \times \mathbb{R} \ ,$$

to be defined next. Here and below, $\mathbb{T}$ denotes a continuum-time domain, isomorphic to the group $\langle\mathbb{R}, 0, +\rangle$ of real numbers, and $\mathbb{R}_{/2\pi\mathbb{Z}} \times \mathbb{R}$ denotes the two-dimensional cylinder-shaped state space, generated by the unit circle and intended to encode all physically relevant states in terms of values of the parametrized position-velocity function. The customary geometric interpretation of state-parametrized solutions consists of smooth curves, covering the cylinder state space and thereby forming the target system's *phase portrait*.

By a *continuous group action* of a topological group $\langle\mathbb{T}, 0, +\rangle$ (interpreted as a continuum-time domain or time-group) on a topological space X (thought of as a *state space*) we mean a jointly continuous map of the form $\delta\colon \mathbb{T} \times X \longrightarrow X$ (also known as a dynamical *transition map*) such that the following axioms of *group action* hold for all time instants $t, t' \in \mathbb{T}$ and for all states x in X:

(i) *Identity property:* $\delta(0, \mathsf{x}) = \mathsf{x}$, and
(ii) *Group property:* $\delta(t, \delta(t', \mathsf{x})) = \delta(t + t', \mathsf{x})$.

Because in what follows, group actions will be used mainly to represent the temporal evolution of target systems' states, the above-introduced group action $\delta\colon \mathbb{T} \times X \longrightarrow X$ is alternatively called a (deterministic) *topological dynamical model* and is symbolized more succinctly by the curved group-action arrow $\mathbb{T} \underset{\delta}{\curvearrowright} X$. An impressively large class of deterministic dynamical models arises from autonomous systems of ordinary first-order differential equations by simple state-parametrizations of their solution spaces. If the time domain $\mathbb{T}$ and the state space X are both smooth manifolds, and if the transition map δ is also smooth (i. e., infinitely differentiable), then $\mathbb{T} \underset{\delta}{\curvearrowright} X$ is called a (deterministic) *smooth dynamical model*. In particular, the structural variant of the earlier discussed smooth pendulum dynamical model has the form $\mathbb{T} \underset{\delta,\upsilon}{\curvearrowright} (\mathbb{R}_{/2\pi\mathbb{Z}} \times \mathbb{R})$, obtained directly from the solutions of a pair of first-order pendulum equations. Because the time group's operation is an action on its own underlying space, we automatically obtain the 'clock' dynamical model $\mathbb{T} \underset{+}{\curvearrowright} |\mathbb{T}|$.

Passage from equations to group action formulations of their solutions represents a significant change of viewpoint. From this new *structuralist* perspective, in formulating a state-based mathematical model or simply a *dynamical model* of a target system's behavior, the modeler needs to specify the following two conceptual ingredients: a state space X (endowed with additional smooth, topological, or measurable structure, assumed to be respected by all maps on it) and a time-group action $\delta\colon \mathbb{T} \times X \longrightarrow X$, satisfying the identity and group properties. A defining property of a state space is its being the domain of real-valued, observable quantities. A prime example of an observable quantity in the pendulum model is the Hamiltonian, representing the pendulum's energy levels.

Structuralists argue that the particular system of equations (if any) that generates the group action, is of secondary concern. In a dynamical model, comprised of a state space and a time-group action on it, each state-space point is an abstract, information-bearing encoding of the target system's physical mode of being at a given time. Assuming that the system under consideration has a physical state structure, its adequate state space model comes with a state space comprised of points that are presumed to contain *complete information* about the past history of the system, relevant to its future behavior. Thus, if the target system's instantaneous physical state – encoded by a point in its representing state model, is known, then the system's subsequent temporal evolution can be predicted with the help of the model's time-group action, without any additional knowledge of what has happened to the system. The group action need not be continuous or smooth. It can also be measurable, computable, discrete, and local (i. e., only partially defined).

The structuralist point of view of mathematical models also includes the so-called *behavioral* approach, proposed by Jan Willems [35]. It turns out to be just as effective as the group action approach. Recall from the basic pendulum model example above that the group action pair $\langle\theta, \upsilon\rangle$ has a function space transpose

$$\mathbb{R}_{/2\pi\mathbb{Z}} \times \mathbb{R} \xrightarrow{\widehat{\langle\theta,\upsilon\rangle}} (\mathbb{R}_{/2\pi\mathbb{Z}} \times \mathbb{R})^{\mathbb{T}} \ ,$$

defined by the time function $\left[\widehat{\langle\theta, \upsilon\rangle}(\langle a, b\rangle)\right](t) \ =_{\mathrm{df}} \left\langle\theta_{a,b}(t), \upsilon_{a,b}(t)\right\rangle$ for all time instants t and initial states $\langle a, b\rangle$ in $\mathbb{R}_{/2\pi\mathbb{Z}} \times \mathbb{R}$. Note that here the image function space $\widehat{\langle\theta, \upsilon\rangle}(\mathbb{R}_{/2\pi\mathbb{Z}} \times \mathbb{R})$ is comprised of those time functions (pictured as state trajectories) in $(\mathbb{R}_{/2\pi\mathbb{Z}} \times \mathbb{R})^{\mathbb{T}}$ that satisfy the initially given pair of first-order pendulum equations.

In engineering applications, it is often conceptually advantageous to work directly with the solution space

$$\left\{\theta \;\middle|\; \frac{\mathrm{d}^2\theta}{\mathrm{d}t^2} + \frac{g}{\ell}\theta = 0\right\} \subseteq \mathbf{C}^\infty(\mathbb{T})$$

of higher-order (linear) equations – interpreted as the subspace of *behaviors*, without any passage to a parametrizing state space. In this behavioral setting, a *behavioral topological dynamical model* is a triple $\langle \mathbb{T}, X, \mathcal{B}\rangle$, consisting of a continuum-time domain $\mathbb{T}$, a topological space X (interpreted as a *signal* space), and a function subspace $\mathcal{B} \subseteq X^{\mathbb{T}}$ of maps (encoding behaviors), satisfying the system's dynamical equations. (Here and below, $X^{\mathbb{T}}$ denotes the space of all continuous time-functions, i. e., continuous maps from $\mathbb{T}$ to X, endowed with the product topology.) This notion of a model is an outgrowth of the traditional input-output method, popular in engineering. The proponents of the behavioral approach in algebraic systems theory emphasize that in general the space of time-functions $X^{\mathbb{T}}$ need not be just the space $\mathbf{C}^\infty(\mathbb{T})$ of all smooth functions. Instead, it can also be the space of compactly supported smooth functions, locally integrable functions, distributions, and so forth. Polderman and Willems [27] provide a very readable introduction to the behavioral conception of mathematical models in systems theory. The foregoing lengthy digression into structuralist conceptions of models completes the first reason why the syntactic definition of mathematical models is unsatisfactory. We are now ready for the second reason.

The second reason why the equational conception in science is inadequate is because equations are unable to deal fully and directly with intended empirical interpretations, representational power, denotation, model networks, and many other semantic and representational functions of models. To place the equations on a mathematically rigorous ground, the model builder must choose a *host* ring or field for the possible values of indeterminates and observable quantities. Beyond that, several other choices have to be made, including the 'correct' choice of a function space for indeterminates and quantities, and that of parameter spaces for constants. Depending on the interpretation, solutions can be everywhere defined and smooth, or they can be generalized functions (distributions), and so forth. There is also the problem of parameter identification and the question of how various quantities are to be measured. These functions of models are typically determined by the intended interpretation of solution spaces and phase portraits. Last but not least, parent models come with various structural enrichments and impoverishments – forming a model network. But of course in many applications, mathematical models involve both linguistic and structural (non-linguistic) elements. In representing natural systems, usually equations provide an all-important starting point. However, in view of various inevitable abstractions and idealizations, the corresponding group-action and behavioral models – having "more meat", serve uniquely well as equation-world *mediators* or *proxies*, and *stand-ins* for the real-world target systems. Indeed, statements derived from well-confirmed equations are strictly true in the associated group-action and behavioral models, but they are only indirectly and approximately true of the actual system's behavior. The applied mathematics literature usually leaves implicit many of the details needed for understanding the equations.

So what is the merit of philosophical objections to the equational treatment of mathematical models? In their seminal work, Oberst [26], Röhrl [28], and Walcher [34] have established a natural *Galois correspondence* between the two (syntactic and structural) approaches, leading to a deeper understanding of the relations between the properties of (differential and difference) equations and the properties of their solutions, representing behavior. This fundamental relationship between the world of equations and that of solution spaces (or input-output behaviors) has been established for a large class of linear (and also for some nonlinear) cases, in the form of category-theoretic duality. In parallel with the above, classical models of statistical experiments, decision theory, game theory, and so forth, built over underlying smooth, topological, metric and measurable spaces, possess computationally convenient algebraic counterparts, granted by Stone–Gelfand duality results, discussed, e. g., in [17].

Having studied the ways in which equations lead to group-actions on state spaces and function spaces of trajectories representing behaviors, we now discuss how these and other types of models can be put together from more basic mathematical structures, and how certain universal constructions support the construction of complex models from simple ones.

Philosophical and Mathematical Structuralism

Since much of what is taken as distinctive of mathematical models and their connections to target systems is tied to mathematical structuralism, a brief description of the concept of structure is in order. We begin by reviewing the principal theses of *philosophical structuralism*. Within the current philosophical literature (e. g., [8]) on the status of scientific realism, two major positions can be discerned. The first, resurrected by Worrall [36], is *epistemic structuralism* that places a special restriction on scientific

knowledge. Its central thesis is that we can have knowledge of structures without knowledge of natures. Here structures are identified with relations – broadly understood, and natures are the intrinsic modes of being of objects that are related. This view is motivated by the need to handle the ontological discontinuity across theory-change in terms of a well-grounded knowledge of structures. For example, although conceptions of the nature of light have changed drastically through history (from a particle ontology to a wave ontology, and then again to a wave-particle duality), nevertheless, the old equation-based models describing the propagation of light have not been abandoned. Rather, at each stage of knowledge, they have been incorporated into more refined successor models. The second position, called *ontic structuralism*, initiated by Ladyman [16], holds that since structure is all there is to reality, what we can have at best is knowledge of the structural aspects of real-world systems. Inspired by the peculiar nature of quantum theory, ontic structuralism argues for the thesis that there are no objects. This object-free ontology leaves its proponents with the burden of showing how physico-chemical entities (e. g., superconductors and radioactive chemicals), endowed with astonishing causal powers, can be dissolved into structures, without anything left over.

In the field of mathematical ontology, structuralism is regarded as one of the most successful approaches to the philosophy of mathematics. For example, Shapiro [31] states that mathematics is the study of independently existing structures, and not of collections of mathematical objects per se. This picture may seem to be a bit distorted to some, since not all of mathematics is concerned with structures. For example, the distribution of prime numbers in the set $\mathbb{N}$ of natural numbers and arguments as to why π must be a transcendental real are hardly structural matters.

Be that as it may, from the standpoint of classical model theory, a mathematical structure is simply a list of operations and relations on a set, together with their required properties, commonly stated in terms of equational axioms. The most familiar examples of such structures are Bourbaki's *mother structures* on sets: *algebraic* (e. g., time groups and quantity rings), *topological* (e. g., metric, measurable and topological state spaces), and *order-theoretic* (e. g., partially ordered sets of observables and event algebras). Of course, there are many vastly more elaborate *composite* types of set-theoretic structures, including ordered groups, topological groups, ordered topological groups, group actions, linear spaces, differentiable manifolds, bundles, sheaves, stacks, schemes, and so forth. The world of sets is granted by a fixed maximal (possibly Platonist) mathematical ontology out of which all structures of mainstream mathematics are presumed to be built.

The foregoing idea of mathematical structure is due mainly to the influence of Nicolas Bourbaki [7], who (together with other mathematicians around him) noted that mathematical structures of practical interest can be generated by three universal operations on sets: The (Cartesian) *product* $X \times Y$ of two sets X and Y, the *power set* $\mathfrak{P}(X)$ of X, and the exponentiation-type *function set* Y^X (consisting of all functions that map X into Y). For example, a group structure on a set X can be viewed as a designated element of the function set $X^{(X\times X)}$ with the usual equational properties of the group operation, together with a suitable element of X^X (for the inverse operation) and a designated element $0 \in X$. Similarly, a topological structure on a set X is given by a designated element of the iterated power set construct $\mathfrak{P}(\mathfrak{P}(X))$, satisfying the axioms of topology. Thus, Bourbaki's concept of a mathematical structure is given by an underlying set together with some higher-order set-theoretic data, forming a string of designated elements of suitable product, power or function set constructions on it, and satisfying certain axioms. Unfortunately, Bourbaki's definition of the concept of mathematical structure includes many pathological examples of no known utility. Furthermore, in general, Bourbaki does not consider the notion of structure-preserving mappings between mathematical structures, except isomorphisms that are always suitable functions, treated as subsets of the Cartesian product of their domains and codomains.

The plurality of set theories with divergent properties in particular and that of universes of mathematical objects in general with wildly different and mutually inconsistent formulations have led to new ideas in philosophical structuralism. Among other things, these ideas suggest that in the presence of an abstract product $\times$, exponentiation $(\cdot)^{(\cdot)}$ and other universal operations, any entity whatever can serve as a proxy for the underlying set-theoretic domain or carrier of a structure. This undermines the Platonist commitment to sets, believed to underly mathematical reality. The plurality of set theories has been replaced by a plurality of *toposes* – a special class of categories with products, function space constructions and subobject classification, in which most set-like mathematical constructions can be carried out. Category theory is seen by many as providing a structuralist framework for mathematics and hence also for mathematical models per se. Simply, structures are determined up to isomorphism, making the particular nature, individuality or constitution of their underlying domains irrelevant. Domains are only 'positions' in structured systems. For example, real numbers are fully specified by the structure of a complete Archimedean ordered

field that has many realizations. Mathematics is not about objects, either empirical or mathematical. It is about the axiomatic presentation of the structure of such objects in general, and of no objects in particular. If mathematics discusses objects, it does so only by construing them as 'positions' in structured systems, devoid of any special ontology. In sum, mathematical objects need not be sets and mappings need not be functions. However, mathematical structuralism does not always translate into physical structuralism. For example, the quantum structure of particles and the physical structure of liquids is empirically meaningless when viewed in isolation from their concrete physical carriers.

Present-day mathematics uses structure-preserving mappings to specify different kinds of sets-with-structure. For example, instead of describing groups traditionally in terms of elements and higher-order data (i. e., operations satisfying the group axioms) on a set, mathematical practice demonstrates that it is far more effective to treat groups as abstract black-box type objects together with distinguished maps, emulating group homomorphisms between them. Upon engaging product and other operations in the construction of these abstractly given objects and maps, algebraists quickly arrive at a crucial universe of discouse, namely the category **Grp** of groups and group homomorphisms. For a thorough discussion of categories, see [22].

In [30] Dana Schlomiuk specifies the classical notion of a topological structure strictly in terms of abstract spaces and maps of a category **Top** – called the category of topological spaces, independently of the usual set-based means of specification, employing elements and closure operators, or algebras of open or closed subsets. Even far more sophisticated structures can be treated in this fundamentally structuralist manner. For example, a smooth structure of a differentiable manifold can be given category-theoretically by specifying exactly which continuous maps must be smooth in the category of abstractly conceived smooth manifolds. Here the utility of category theory is in providing a uniform treatment of the concept of structure, solely in terms of holistically given 'structure-preserving' maps that serve remarkably well also the needs of empirical applications.

Having now indicated the ways category theory specifies mathematical structures, we may profitably revisit the earlier discussed syntactic *equational* and *structuralist* approaches to mathematical models from the point of view we have just developed. Here we only give one illustration of this approach out of many possibilities, namely the category $\mathbf{Top}^{\mathbb{T}}$ of topological dynamical models in the form of a topological group-action on state spaces by a designated time group $\langle \mathbb{T}, 0, + \rangle$, discussed in the previous Section. This category consists of topological dynamical models of the form $\mathbb{T} \underset{\delta}{\curvearrowright} X$, serving as its objects, where the map $\delta\colon \mathbb{T} \times X \longrightarrow X$ is a continuous action of time-group $\mathbb{T}$ on its codomain topological state space X. Next, maps between dynamical models of the form $\varphi\colon (\mathbb{T} \underset{\delta}{\curvearrowright} X) \longrightarrow (\mathbb{T} \underset{\delta'}{\curvearrowright} X')$, called *dynamorphisms*, are given by continuous mappings $\varphi\colon X \longrightarrow X'$ between the underlying state spaces such that the diagram

$$\begin{array}{ccc} \mathbb{T} \times X & \xrightarrow{\delta} & X \\ {\scriptstyle 1_{\mathbb{T}} \times \varphi}\downarrow & & \downarrow{\scriptstyle \varphi} \\ \mathbb{T} \times X' & \xrightarrow[\delta']{} & X' \end{array}$$

commutes, i. e., the compositions of constituent maps along both paths in the diagram give the same map. In other words, we have the equality $\varphi \circ \delta = \delta' \circ (1_{\mathbb{T}} \times \varphi)$. Because the identity maps of state spaces are trivially dynamorphisms and since the composition of two dynamorphisms is again a dynamorphism, by definition, $\mathbf{Top}^{\mathbb{T}}$ is indeed a category.

A dynamorphism φ from $\mathbb{T} \underset{\delta}{\curvearrowright} X$ to $\mathbb{T} \underset{\delta'}{\curvearrowright} X'$ is said to be an *isomorphic dynamorphism* or simply a *conjugacy* provided that $\varphi\colon X \longrightarrow X'$ is a homeomorphism. From the standpoint of mathematical systems theory, two conjugate dynamical models are structurally identical. Any dynamical property (i. e., a property defined in terms of a group action on its underlying state space) possessed by one is also possessed by the other. Concretely, a dynamical model $\mathbb{T} \underset{\delta}{\curvearrowright} X$ is said to be *structurally stable* provided that any dynamical model $\mathbb{T} \underset{\delta'}{\curvearrowright} X$ with transition map δ' sufficiently close to δ in a suitable topological sense is conjugated to $\mathbb{T} \underset{\delta}{\curvearrowright} X$.

Surjective (onto) dynamorphisms are called *factor* dynamorphisms. Each factor dynamorphism $\varphi\colon X \longrightarrow X'$ comes with its natural equivalence relation, given by $x_1 \equiv x_2$ iff $\varphi(x_1) = \varphi(x_2)$, and induced *quotient* dynamical model $\mathbb{T} \underset{\delta_\varphi}{\curvearrowright} X_{/\equiv}$. Injective (one-to-one) dynamorphims specify dynamical *submodels*. For example, the trajectory passing through a state x at time zero, defined by $\mathbb{T}(x) = \{\delta(t, x) \mid t \in T\}$, can be viewed as a member of the family of the smallest nontrivial submodels of $\mathbb{T} \underset{\delta}{\curvearrowright} X$. Many other notions and constructions available in the category **Top** of topological spaces automatically transfer to the category $\mathbf{Top}^{\mathbb{T}}$ of topological dynamical models and dynamorphisms. For example, the *product* $(\mathbb{T} \underset{\delta}{\curvearrowright} X) \times (\mathbb{T} \underset{\delta'}{\curvearrowright} X')$ of two dynamical models $\mathbb{T} \underset{\delta}{\curvearrowright} X$ and $\mathbb{T} \underset{\delta'}{\curvearrowright} X'$ is defined in the same way as in topology, namely by

$$(\mathbb{T} \underset{\delta}{\curvearrowright} X) \times (\mathbb{T} \underset{\delta'}{\curvearrowright} X') =_{\mathrm{df}} \mathbb{T} \underset{\delta;\delta'}{\curvearrowright} (X \times X') ,$$

where the diagonal action is given by $[\delta;\delta'](t,(x,x')) = \langle\delta(t,x),\delta'(t,x')\rangle$. Disjoint sum (coproduct) of two dynamical models is defined similarly. For more details regarding category-theoretic constructions, see [20] that includes an elaborate category-theoretic treatment of dynamical models. Similar constructions work with varying degrees of success in many other host categories. For example, if **Top** is replaced by the category **Mes** of measurable spaces and measurable maps between them, we obtain the much studied category $\mathbf{Mes}^{\mathbb{T}}$ of measurable dynamical models. Another significant mutation of dynamical models arises, when the continuum-time domain is replaced by a discrete-time domain, isomorphic to the group $\langle\mathbb{Z},0,+\rangle$ of integers or the monoid $\langle\mathbb{N},0,+\rangle$ of natural numbers.

Topological (smooth or measurable) dynamical models provide complete support also for the representation of *perturbed, controlled, nonautonomous, and random* dynamical systems, in terms of various skew-product constructions. In more detail, let $\mathbb{T}\underset{\delta}{\curvearrowright}X$ be a (topological, smooth or measurable) *base* dynamical model that represents a *driving* system (i. e., perturbation, control or environmental noise process), affecting the target system's dynamical behavior. The model of this behavior, subject to the influence of added perturbation, control, and so on, is based on a so-called (continuous, differentiable or measurable) *cocycle* map $\alpha\colon \mathbb{T}\times X\times Y \longrightarrow Y$, acting on the *driven* target system's principal state space Y and satisfying the following *skew-product action* (or *cocycle*) axioms for all t,t' in $\mathbb{T}$, $x\in X$, and $y\in Y$:

(i) *Identity property:* $\alpha(0,x,y)=y$, and
(ii) *Cocycle property:* $\alpha(t+t',x,y) = \alpha\big(t,\delta(t',x),\alpha(t',x,y)\big)$.

The space $X\times Y$ should be thought of, informally, as a *trivial bundle*, made up of fibers $\{x\}\times Y$, indexed by points $x\in X$ and "glued together" by the topology. In the same way, for time instants t,t', the cocycle map should be viewed as an indexed family

$$\{x\}\times Y \xrightarrow{\alpha(t,\bullet,\cdot)} \{\delta(t,x)\}\times Y \xrightarrow{\alpha(t',\delta(t,\bullet),\cdot)} \{\delta(t'+t,x,)\}\times Y$$

of maps between fiber state spaces. As the extant base state x at time t is shifted by the base model's dynamics to $\delta(t,x)$, the restricted cocycle map $\alpha(t,x,\cdot)$ moves each system state y in the fiber $\{x\}\times Y$ over x to the state $\alpha(t,x,y)$ belonging to the fiber $\{\delta(t,x)\}\times Y$ over $\delta(t,x)$.

The *skew-product* $(\mathbb{T}\underset{\delta}{\curvearrowright}X)\underset{\alpha}{\ltimes}Y$ of a base dynamical model $\mathbb{T}\underset{\delta}{\curvearrowright}X$ and a principal state space Y under cocycle α acting on Y is defined by the dynamical model

$$(\mathbb{T}\underset{\delta}{\curvearrowright}X)\underset{\alpha}{\ltimes}Y =_{\mathrm{df}} \mathbb{T}\underset{\phi}{\curvearrowright}(X\times Y)$$

on the trivial bundle $X\times Y$, where $\phi\colon \mathbb{T}\times(X\times Y)\longrightarrow X\times Y$ is specified by $\phi(t,(x,y)) =_{\mathrm{df}} \langle\delta(t,x),\alpha(t,x,y)\rangle$ for all time instants t and states x,y. It is easy to check that the transition map ϕ is fiber-preseving and hence a bundle morphism. Since each group-action induces a skew-product action, product dynamical models are special cases of skew-product dynamical models. More importantly, because the second projection map $\pi_Y\colon (\mathbb{T}\underset{\delta}{\curvearrowright}X)\underset{\alpha}{\ltimes}Y \longrightarrow (\mathbb{T}\underset{\delta}{\curvearrowright}X)$ is a factor dynamorphism, skew-products are best viewed as objects of the so-called slice category $\mathbf{Top}^{\mathbb{T}}_{/\mathbb{T}\underset{\delta}{\curvearrowright}X}$ that fuses topological (smooth, measurable, etc.) bundle theory with the theory of dynamical systems. *Rohlin's representation theorem* states that the domains of factor morphisms in the category of probability spaces and maps are essentially skew-products. Thus, all stochastic representations of perturbed systems are tractable uniformly by skew-product constructions.

Nonautonomous differential equations (describing nonautonomous systems) of the form $\mathrm{d}y/\mathrm{d}t = f(t,y)$ with $t\in\mathbb{T}$ and $y\in Y$ (involving an explicit time dependence), are known to have solutions that induce skew-product dynamical models of the form $(\mathbb{T}\underset{+}{\curvearrowright}|\mathbb{T}|)\underset{\alpha}{\ltimes}Y$ over a suitable cocycle α. Quite simply, the model enlarges the principal state space Y by the space $|\mathbb{T}|$ of time coordinates. Because solutions of stochastic differential equations have the form of cocycles over base measurable dynamical models, once again, smooth skew-products, called *random dynamical models*, provide the correct structuralist framework for them. For a thorough treatment of some of these topics, see [1] and [9].

Three Approaches to Applying Mathematical Models

Since one of the most perplexing philosophical questions is how models relate to their target systems, in this Section we review three major approaches to the problem of model-world relationship.

Internal Approach

Set-theoretic models are applicable to real-world systems simply because physical objects can literally be members of legitimate sets and there are special functions between sets comprised of physical objects (e. g., particles and bodies) and pure mathematical objects (e. g., reals). As a classical example of a physical application of this nature, recall the application of Newton's laws to the solar system. The proponents of the internal view (including, e. g., Steiner [32]) argue that planets together with the sun form a perfectly meaningful finite set on which real-valued mass and position functions can be defined in the usual way. In this man-

ner, set-theoretic models are applied to physics (and other disciplines) using only *internal* (i. e., internal to set theory) relations between, say, bodies and numbers. The defenders of the internal view treat the names of (say) planets as *rigid designators*, always picking the same object in every possible situation in which they exist at all. It is now easy to see how the intended empirical interpretation of mathematical models leads directly to claims about the physical world.

It must be stressed, however, that there is a good basis for questioning this fairly popular approach. One obvious concern is this: For particles, bodies and other physical objects to form classical sets, it must be assumed that they can never be annihilated, split or changed in some other ways, leading to a loss of their identity. Under these types of implicit comprehensive idealizations, it is better to think of such 'physical' or 'empirical sets' as ordinary sets involving mathematical encodings. Simply, the actual physical objects under study are encodable in terms of suitable abstract elements (e. g., numbers) of a set, so that the informal semantic assumptions of the sort "$\mathcal{B}$ is a set of physical bodies", "X is a set of physical states", "$\mathbb{T} \underset{\delta}{\curvearrowright} X$ is a dynamical system", and so forth, are treated as succinct abbreviations of the formal details of fallible mathematical encodings and their empirical interpretation. A more serious philosophical concern pertains to the nature of (e. g., differentiable) maps on various 'empirical sets', consisting of fluids and other objects of continuum mechanics.

External Approach

Some philosophers suggest to draw a thick conceptual line between timeless mathematical models and actual physical systems. But then, for a model to be a model of something else, a relationship between a model and what it models is required. Specifically, applications of mathematical models involve certain external (preferably intensional) relations between models and real-world systems. In more detail, philosophical structuralists argue that the model-world relation is best described by structure-preserving maps (usually isomorphisms or embeddings) between representing mathematical models and their intended physical (or generally empirical) situations. As a concrete illustration of this viewpoint, consider the representational model of weight measurement, studied in great detail by Kranz et al. in [15]. The empirical domain $\mathcal{D}$ of the 'physical structure' of weight measurement is a set of material objects. This domain is equipped with a binary 'heavier-than' weak-order relation $\succ$, where the empirical statement $d \succ d'$ – stating that object d is heavier than object d', is operationally established by placing d and d' on the two pans of an equal-arm balance scale and then observing which pan drops. In addition, there is a binary weakly commutative and weakly associative composition operation $\oplus$ on $\mathcal{D}$, where for objects d and d' the value $d \oplus d'$ denotes the composite object obtained by placing both objects in the same pan with d beneath d'. Now, here the world-model relation is given by an algebraic homomorphism from the so-called physical *extensive measurement structure* $\langle \mathcal{D}, \succ, \oplus \rangle$ of weight measurement to the ordered additive structure $\langle \mathbb{R}, >, + \rangle$ of reals. Such homomorphisms are guaranteed to exist provided that the extensive measurement structure satisfies certain relatively simple axioms, similar to those used in the definitions of Archimedean-ordered semigroups.

We see that for a structure-preserving world-to-model map to exist, numerous idealizing restrictions must be placed on the target physical domain's structure. For example, in the case of weight measurement, the Archimedean property forces the empirical domain $\mathcal{D}$ to be infinite, even though in applications experimenters always work with finitely many objects. Because idealizations typically involve known-to-be-false descriptions, ignorance of causes, and deliberate stipulations of objects that do not exist in the actual physical situation, a structure-preserving map can be guaranteed to exist only between the idealization itself and a representing mathematical model, but not necessarily between the actual real-world situation (enjoying unlimited degrees of freedom) and the model.

Finally, one must question the ontological status of the structure-preserving maps (isomorphisms). If these maps go from a mathematical domain to a real-world physical domain or conversely, then their graphs, being subsets of a 'hybrid set' of physical-mathematical object pairs, take us back straight to the internal viewpoint, questioned earlier.

A rival and equally popular approach to this general topic is to replace the isomorphism (or homomorphism) relation between models and the world by some other, technically less demanding devices. For example, Giere [13] proposes to solve the problem of structure-preserving maps by passing to considerably simpler graded context-dependent *similarity* relations between models and aspects of the real world. Thus, if the representing model is sufficiently similar to its target system, then the analysis of the model is also (indirectly) an analysis of the system. This solution comes at quite a price, since it is hard to see what exactly, if anything, is similar between the familiar Lotka–Volterra differential equations and the predator-prey two-species system of finitely many sharks and fishes in the Adriatic Sea. Another major concern is that under similarity anything can be a model of anything

else, since any two things can always be claimed to be similar in some respect and to some degree or other. In general, model builders cannot reduce the problem of application of models to similarity relations, because models usually come with falsifying idealizations that simply violate any similarity relation that may perhaps exist between intractable 'perfect' models and their systems. Last but not least, mathematical models tend to possess a 'surplus structure' and features that do not correspond to anything in particular, similar or not, in the actual system (e. g., recall the applications of complex analysis to the behavior of electric circuits or the role of undetectable measure-zero attractors in a dynamical model).

Model Network Approach

The relation between a mathematical model and its target system is not as straightforward as it is often thought to be. Models interact with other models, both globally and locally in complex ways, and their relationship to the world is not a static affair. If connections between models and their target systems were tractable by rigid relations, then a large bulk of scientific research would become redundant. However, as mathematical models are structurally mutated, their representational capacities often undergo unexpected refinements and changes. As a result, in subsequent applications new epistemic links between models and systems emerge, and accrue along the way. Thus, from the standpoint of scientific practice, an essential aspect of mathematical modeling is extendability, enrichment and open-endedness of model constructions. Model builders cannot realistically constrain models to be just single frozen structures. Instead, applied mathematical models are best viewed as complex *networks* of structures with distributed empirical interpretations that guide model users in the network-world interface to gather more informative data and make better predictions. (For related views but motivated by different considerations, see [4] and [5].)

Applications of category theory to mathematical models has opened up the possibility of building new models from old ones via functorial constructions. In Sect. "Philosophical and Mathematical Structuralism", we have given a brief introduction to some concepts, intended for the class of topological dynamical models. But there is more. For example, the study of deterministic dynamical systems operating in a chaotic regime – hard to understand in deterministic terms, calls for a passage to *probabilistic* dynamical models. These are obtained by introducing *probability monads* – special functors discussed by Giry [14], on suitable categories of topological and measurable spaces. In brief, the parent deterministic dynamical model $\mathbb{T} \underset{\delta}{\curvearrowright} X$ is used in conjunction with the associated logically higher-level probabilistic dynamical model $\mathbb{T} \underset{\delta^*}{\curvearrowright} \boldsymbol{\mathcal{D}}(X)$ with time-group action on the convex space $\boldsymbol{\mathcal{D}}(X)$ of probability density functions, representing the target system's stochastic states. Here the transition map δ^* is defined by the so-called *Perron–Frobenius integral equation*, studied in great detail by Lasota and Mackey [19]. Mutation of $\mathbb{T} \underset{\delta}{\curvearrowright} X$ into nondeterministic, fuzzy and other dynamical models is also obtained by the monad approach. Since the time evolution of states of a target system is observed only indirectly via measurable real-valued quantities, defined on the representing model's state space X, the *parent* model $\mathbb{T} \underset{\delta}{\curvearrowright} X$ is used in parallel with another associated higher-level dynamical model $\mathbb{T} \underset{\delta^*}{\curvearrowright} \mathbf{C}(X)$ on the Banach algebra $\mathbf{C}(X)$ of (bounded) real-valued continuous functions on X. The induced time-group action δ^*, defined by the fundamental *Gelfand duality* result, characterizes the time evolution of observable quantities, crucial in time-series analysis.

Because in applications the evaluation of the analytic parent model's transition map tends to involve roundoff and other truncation operations that may have long-term pathological effects, and since generally the model's time and space points cannot be identified with arbitrarily fine degrees of accuracy, several forms of (temporal, spatial and parametric) *discretizations* become necessary. *Time discretizations* of a continuum-time model $\mathbb{T} \underset{\delta}{\curvearrowright} X$ is solved by passing to discrete-time dynamical models $\tau\mathbb{Z} \underset{\delta_\tau}{\curvearrowright} X$, where $\tau\mathbb{Z} = \{\cdots, -2\tau, -\tau, 0, \tau, 2\tau, \cdots\}$ is the group of discrete timesteps with sampling period $\tau > 0$. *Spatial discretization* of $\mathbb{T} \underset{\delta}{\curvearrowright} X$ is captured by a nested family of coarse-grained dynamical models of the form $\mathbb{T} \underset{\delta_\epsilon}{\curvearrowright} X_\epsilon$, where the underlying state space X_ϵ is comprised of a uniform grid (lattice) of finitely many points in X with mesh size $\epsilon = \frac{1}{n}$, tractable with finitary resources. Spatially discretized dynamical models provide the formal meeting ground for measurement results and the parent model's predictions. To understand the gap between *ontologically* and *epistemologically* motivated dynamical models, it is important to know how well a discretized space mimics the properties of the parent model's state space. Remarkably, Hausdorff topological state spaces are known to be approximable by inverse limits of nested sequences of T_0 finite topological spaces. For more details on this subject, see [6].

Validating Mathematical Models

Although many models are built for scientific research purposes, applications demand that models be validated iteratively by comparisons of their predictions with mea-

surement data. Validation is intended to demonstrate that the representing model generates predictions and claims in its domain of applicability that are consistent with its intended application – within an acceptable range of accuracy. Because it is not true that good predictions can only be obtained from a model that is causally sound and since data may not portray the target system accurately, full *verification* of a model requires various testing procedures for agreement with cause-effect relationships (operating in the target system) and analysis of inter-model relations in a model network to which it belongs.

Validation tends to be limited by the available domain of measurement data – to be used in model-world comparisons, and generally is not sufficient to demonstrate that the model is adequate in its entire state space, modulo locally granted margins of error. As a matter of fact, mathematical models are seldom valid in all regions of their phase portraits. As we have seen, the most popular example is the Newtonian dynamical model of a simple pendulum that gradually fails as its angular velocity approaches the speed of light. In scientific practice, a dynamical model is typically valid only in limited regions of its underlying state space that are of particular interest. Problems arise when the model fails to be valid in most parts of the state space regions under consideration. Deficiencies in the model may be traced to a wrong choice of differential equations (that ignores higher degrees of nonlinearity), a crude choice of time and/or space discretization parameters, and reliance on data samples contaminated with gross errors.

In many applications, it is far from sufficient to know that the model is *statistically* valid. The reason is that a statistically valid theoretical model may turn out to be *dynamically* (qualitatively) invalid, meaning that it may fail to represent correctly the dynamical invariants of the target system (including equilibrium states, periodic, aperiodic, chaotic or strange attractors, and their basins) under various choices of parameter values. Note that even if a dynamical model provides highly accurate predictions in its tested state space regions, it does not necessarily follow that the model is dynamically valid, since it may include spurious dynamics in untested state space regions that are also of interest. The spectrum of actual behaviors of the target system remains largely unknown from the perspective of its representing model. Further observation and theoretical research may be needed to obtain a better knowledge about the dynamical invariants and characteristics of the system.

These considerations lead us to reason about the validation of dynamical models geometrically, in terms of suitable conditional geometric measures of adequacy.

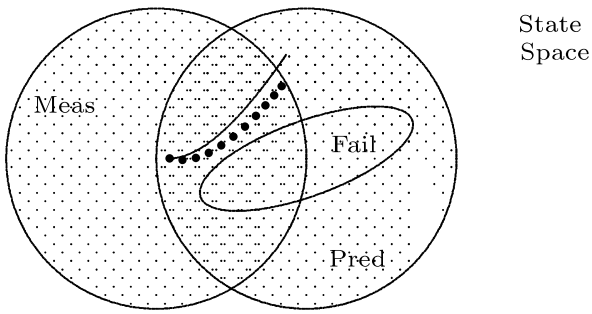

Philosophy of Science, Mathematical Models in, Figure 1
Geometry of model validation

Given a parent dynamical model of a target system, suppose the associated body of actual observation and *measurement* data, available at a given stage of knowledge and collected independently of the model under consideration, forms a subset Meas of a finite grid of the parent model's underlying state space, as illustrated by a shaded circle in Fig. 1 below. The size of the set Meas is constrained by various resources, available measuring instruments, their accuracy, and research interests of the validating scientist. Of course, Meas varies with time and it can form a multiply connected or scattered subset of the state space. For simplicity, we assume that the discretization embodied in the grid captures the uniform admissible margin of error.

Along related lines, let Pred be a subset of a finite grid of the parent model's state space, given by all quantitative *predictions* calculated from the representing model at a given stage of research and depicted by the second shaded circle in Fig. 1 that overlaps with Meas. Naturally, the set Pred grows in time as brand-new predictions are generated by the model. It should be obvious that the intersection Meas $\cap$ Pred represents those predictions that have been checked by observation or measurement.

Since the model of interest need not be perfect or fully reliable, it is likely to generate some predictions that falsify the model. In Fig. 1, the set of *falsifiers* is indicated by the ellipse Fail, forming a subset of Pred. Note that at the current stage of research, only a proper subset of incorrect predictions in Fail is comparable with measurement results. The other potentially falsifying predictions in Fail have not yet been verified. Finally, note also that subsequent discrete measurements of a state trajectory (indicated in Fig. 1 by a dotted curve segment) over a longer period of time may gradually diverge from the corresponding parent model-given continuous state trajectory, both generated by the same presumed initial state.

We are now ready to use the geometric scheme developed above to *measure the adequacy of dynamical models*.

Even though measurement results and predictions form finite sets, for the sake of simplicity we choose to measure the *adequacy* of the given dynamical model by the ratio

$$\frac{\Lambda(\mathtt{Meas} \cap (\mathtt{Pred} - \mathtt{Fail}))}{\Lambda(\mathtt{Meas})},$$

where Λ denotes the Lebesgue measure defined on the model's underlying state space. The idea is that the ratio of the volume of fully verified *correct* predictions and the volume of all available measurement results is a good indicator of the model's adequacy at a given stage of knowledge. In any case, it should be clear that a *semantic* approach to confirmation theory must directly engage the structure of dynamical models under consideration. Because adequacy and reliability should be judged according to the volume of phase portrait regions on which measurements and predictions agree, modulo admissible errors, measures of the sort displayed above are in the ballpark of measuring adequacy. Note, however, that just like many other measures in dynamical systems theory (including topological and Kolmogorov–Sinai entropy functions), the foregoing measure of model adequacy is largely conceptual and not readily implementable in all instances of real-life models.

Testing models for adequacy is conceptually quite similar to testing statistical hypotheses. Recall that a given representative sample of data validates a statistical hypothesis about the target population only with a certain degree of confidence; the larger the sample, the higher the degree of confidence in the correctness of the hypothesis. Likewise, representative measurement data validate the dynamical model under consideration via its predictions pertaining to the target system's addressed behavior only locally, specified by a region in its phase portrait. The larger the body of validating data in its phase portrait, the higher the degree of confidence in the model's global adequacy.

Future Directions

We do not give a detailed list but briefly mention two major directions of current research.

Galois Correspondence Between Equational and Structuralist Approaches

There are several algebraic methods in the realm of linear differential equations that demonstrate a *Galois correspondence* between differential or difference ideals of equations (modules of differential operators) and group-action (behavioral) models. There are reasons to conjecture that these results remain valid also for a large class of *nonlinear* differential or difference equations. Although some examples of this are studied by Walcher [34], a more complete understanding of the equation-solution relationship is needed. A similar Galois connection must be addressed also in the context of stochastic differential equations and random dynamical models or stochastic flows.

Validation and Verification of Mathematical Models

It is a problem of considerable interest to formalize the processes of model validation and testing. In recent years efforts have been made to come to grips with this problem in the field of validation of computer simulation models. (See, for example, the influential work of Sargent [29].) However, a substative theory of dynamical model validation that circumvents the inadequacies of classical confirmation theory is not yet available.

Bibliography

Primary Literature

1. Arnold L (1998) Random dynamical systems. Springer, New York
2. Atmanspacher H, Primas H (2003) Epistemic and ontic realities. In: Castell L and Ischebeck O (eds) Time, Quantum and Information. Springer, New York, pp 301–321
3. Bailer-Jones DM (2002) Scientists' thoughts on scientific models. Perspectives Sci 10:275–301
4. Balzer W, Moulines CU, Sneed J (1987) An Architectonic for science: The structuralist program. Reidel, Dordrecht
5. Balzer W, Moulines CU (1996) Structuralist theory of science. Focal issues, new results. de Gruyter, New York
6. Batitsky V, Domotor Z (2007) When good theories make bad predictions. Synthese 157:79–103
7. Bourbaki N (1950) The architecture of mathematics. Am Math Mon 57:221–232
8. Chakravartty A (2004) Structuralism as a form of scientific realism. Int Stud Philos Sci 18:151–171
9. Colonius F, Kliemann W (1999) The dynamics of control. Birkhäuser, Boston
10. Da Costa NCA, French S (2003) Science and partial truth. A unitary approach to models and scientific reasoning. Oxford University Press, New York
11. Erkal C (2000) The simple pendulum: a relativistic revisit. Eur J Phys 21:377–384
12. Fulford G, Forrester P and Jones A (1997) Modelling with differential and difference equations. Cambridge University Press, New York
13. Giere R (2004) How models are used to represent reality. Philos Sci (Proc) 71:742–752
14. Giry M (1981) A categorical approach to probability theory. In: Banaschewski B (ed) Categorical Aspects of Topology and Analysis, Lecture Notes in Mathematics, vol 915. Springer-Verlag, New York, pp 68–85
15. Krantz DH, Luce RD, Suppes P, Tversky A (1971) Foundations of measurement. Academic Press, New York
16. Ladyman J (1998) What is structural realism? Stud Hist Philos Sci 29:409–424

17. Lambek J, Rattray BA (1979) A general Stone-Gelfand duality. Trans Am Math Soc 248:1–35
18. Landry E (1999) Category theory: The language of mathematics. In: Howard DA (ed) Proceeedings of the 1998 Biennial Meeting of the Philosophy of Science Association. The University of Chicago Press, Chicago, pp S14–S26
19. Lasota A, Mackey MC (1994) Chaos, Fractals and Noise. Stochastic Aspects of Dynamics, 2nd edn. Springer-Verlag, New York
20. Lawvere FW (2005) Taking categories seriously. Theor Appl Categ 8:1–24
21. Le Cam L (1986) Asymptotic Methods in Statistical Decision Theory. Springer-Verlag, New York
22. Mac Lane S (1971) Categories for the working mathematician. Springer-Verlag, New York
23. Mac Lane S (1996) Structure in Mathematics. Philosophia Mathematica 20:1–175
24. Moore C (1991) Generalized shifts: unpredictability and undecidability in dynamical systems. Nonlinearity 4:199–230
25. Morgan M (2005) Experiments versus models: New phenomena, inference and surprise. J Econ Methodol 12:317–329
26. Oberst U (1990) Multidimensional constant linear systems. Acta Appl Math 68:59–122
27. Polderman JW, Willems JC (1998) Introduction to mathematical systems theory: A behavioral approach. Texts in Applied Mathematics 26. Springer-Verlag, New York
28. Röhrl H (1977) Algebras and differential equations. Nagoya Math J 68:59–122
29. Sargent RG (2003) Verification and validation of simulation models. In: Chick S, Sanchez PJ, Ferrin E, and Morrice DJ (eds) Proceedings of the 2003 Winter Simulation Conference. IEEE, Piscataway, pp 37–48
30. Schlomiuk DI (1970) An elementary theory of the category of topological spaces. Trans Am Math Soc 149:259–278
31. Shapiro S (1997) Philosophy of Mathematics. Structure and Ontology. Oxford University Press, Oxford
32. Steiner M (1998) The applicability of mathematics as a philosophical problem. Harvard University Press, New York
33. Suarez M (2003) Scientific representation: Against similarity and isomorphism. Int Stud Philos Sci 17:225–244
34. Walcher S (1991) Algebras and differential equations. Hadronic Press, Palm Harbor
35. Willems JC (1991) Paradigms and puzzles in the theory of dynamical systems. IEEE Trans Automatic Control 36:259–294
36. Worrall J (1996) Structural realism: the best of both worlds? In: Papineau D (ed) The Philosophy of Science, Oxford University Press, New York, pp 139–165

Books and Reviews

Brin M, Stuck G (2002) Introduction to Dynamical Systems. Cambridge University Press, New York
Hellman G (2001) Three varieties of mathematical structuralism. Philosophia Mathematica 9:184–211
Margani L, Nersessian N, Thagard P (eds) (1999) Model-based Reasoning in Scientific Discovery. Kluwer, Dordrecht
Margani L, Nersessian N (eds) (2002) Model-based Reasoning: Science, Technology, Values. Kluwer, Dordrecht
Morgan M, Morrison M (1999) Models as Mediators. Perspectives on Natural and Social Science. Cambridge University Press, Cambridge
Psillos S (1999) Scientific Realism: How Science Tracks Truth. Routledge, London
Suppes P (2002) Representation and Invariance of Scientific Structures. CSLI Publications, Stanford University, Stanford

Physics and Mathematics Applications in Social Science

Dietrich Stauffer[1], Sorin Solomon[2]
[1] Institute for Theoretical Physics, Cologne University, Köln, Germany
[2] Racah Institute of Physics, Hebrew University, Jerusalem, Israel

Article Outline

Glossary
Definition of the Subject
Introduction
Some Models and Concepts
Applications
Future Directions
Bibliography

Glossary

Cellular automata Discrete variables on a discrete lattice change in discrete time steps.

Ising model Neighboring variables prefer to be the same but exceptions are possible. The probability for such exceptions is an exponential function of "temperature".

Percolation Each site is randomly either occupied or empty, leading to random clusters. At the percolation threshold for the first time an infinite cluster is formed.

Universality Certain properties are the same for a whole set of models or of real objects.

Definition of the Subject

Herein we introduce the section of this Encyclopedia devoted to Social Sciences, edited by A. Nowak, which concentrates on the application of mathematics and physics to this field. Under "mathematics" we include also all computer simulations if they are not taken from physics; while physics applications include model simulations derived from physics that were applied to social simulations. Thus, obviously there is no sharp border between applications from physics and from mathematics in the sense of our definition. Also social science is not defined precisely. Included are economics and linguistics, but not social insects

or fish swarms, nor human epidemics or demography. It should further be noted that the section of the Encyclopedia on agent-based modeling edited by F. Castiglione also contains articles of social interest.

Introduction

If mathematical/physical methods are applied to social sciences, a major problem is the mutual lack of literature knowledge. Take for example the Schelling model of racial segregation in cities [1]. Sociologists don't cite the better and simpler Ising model, physicists ignored the Schelling model for decades, and sociologists also ignored better sociology work [2]. For simulations of financial markets, many econophysicists thought that they had introduced Monte Carlo and agent-based simulations to finance, not knowing of earlier work from some forward-looking Nobel laureates in economics [3,4]. For inter-community relations, already 25 centuries ago analogies with liquids were pointed out by Empedokles in Sicily ► Opinion Dynamics and Sociophysics or [5]. More recently, Ettore Majorana [6] around 1940 suggested application of quantum-mechanical uncertainty to socio-economic questions. With emphasis shifted to statistical physics, sociophysics and econophysics became fashionable around the change of the millennium, but continuous lines of research by some physicists had already begun in 1971 [7]. In the same year the Journal of Mathematical Sociology started and published Schelling's model of urban segregation [1], which is a modification of the Ising magnet at zero temperature. The year of 1982 saw the start of two other lines of research by physicists on socio-economic questions [8,9].

Languages have been simulated on computers for decades, while the interest of physicists in this area is more recent [10,11], triggered mostly by a model of language competition [12].

We do not mention chemists since at present they play no major role in this field. However, the 1921 chemistry Nobel laureate F. Soddy [13], to whom we owe the "isotope" concept, had already worked on economic, social and political theories, and his finance work of the 1930s was still being cited in 2007. The present authors are trying this the other way round: First apply physics to social sciences, and then get the Nobel prize (for literature: science fiction).

Some Models and Concepts

Physicist Albert Einstein said that models should be as simple as possible, but not simpler. In this spirit we now introduce some basic physics models and concepts for readers from social sciences. All models are complex in the sense that the behavior of large systems cannot be predicted from the properties of the single element.

Cellular Automata

Mathematicians denote cellular automata often as "interacting particle systems", but since many other models or methods in physics use interacting particles, we do not use this term here. A large d-dimensional lattice of L^d sites carries variables S_i $(i = 1, 2, \ldots, L^d)$ which can be either zero or one; more generally, they are small integers between 1 and $Q > 2$. The lattice may be square (four nearest neighbors), triangular (six nearest neighbors), or simple cubic (also six nearest neighbors, but in $d = 3$ dimensions); many other choices are also possible. Time $t = 1, 2, \ldots$ increases in steps. At each time step, each $S_i(t + 1)$ is calculated anew, one i after the other, from a deterministic or probabilistic rule involving the neighboring $S_k(t)$ of the previous time step. This way of updating is called "simultaneous" or "parallel"; one may also use sequential updating where S_i depends on the current values of the neighbors S_k; then the order of updating is important: random sequential, or regular like a typewriter.

An example is a biological infection process: Each site i becomes permanently infected, $S_i = 1$, if at least one of its nearest neighbors is already infected. (Computers handle that efficiently if each computer word of, say, 32 bits stores 32 sites, and if then 32 possible infections are treated at once by bit-by-bit logical-OR operations [14].)

Temperature

We know temperature T from the weather reports, but in physics it enters according to Boltzmann into the probability

$$p \propto \exp(-E/k_B T) \tag{1}$$

to observe some configuration with an energy E. Here T is the temperature measured in Kelvin (about 273 + the Celsius or centigrade temperature), and k_B the Boltzmann constant relating the scales of energy and temperature. For simplicity we now set $k_B = 1$, i. e., we measure temperature and energy in the same units. If g different configurations have the same energy, then $S = \ln(g)$ is called the entropy, and the probability to observe this energy is $\propto g \exp(-E/T) = \exp(-F/T)$ with the "free energy" $F = E - TS$.

In a social application we may think of peer pressure or herding: If your neighbors drink Pepsi Cola, they influence you to also drink Pepsi, even though at present you drink Coca Cola. Thus, let E be the number of nearest neighbors drinking Pepsi Cola, minus the number of

Coke-drinking nearest neighbors. The probability for you to switch then is given by the energy difference and equal to $\exp(-2E/T)$ (or 1 if $E < 0$) in the Metropolis algorithm, or $1/(1 + \exp(2E/T)$ in the Glauber or Heat Bath algorithm. In both cases there is a tendency to decrease E. In the limit $T = 0$ one never makes a change which increases E, while for small positive T one increases E with a low but finite probability. In the opposite limit of infinite temperature, the energy becomes unimportant and all possible configurations become equally probable. Neither zero nor infinite temperature are usually realistic.

In this sense, decreasing the energy E is the most simple or most plausible choice, and the temperature measures the willingness or ability to deviate from this simplest option, e. g. to withstand peer pressure. But temperature also incorporates all those random accidents of life that influence us but are not part of the social model. For example, it may happen that there is no Pepsi Cola available even though all your neighbors drink Pepsi and you want to follow them. Investors have to make their financial choices under the influence of their clients, whose life is shaped by births, marriages, deaths, or other personal events which are not included explicitly into a financial market model. These accidents are then simulated by a finite temperature, entering the probability that one does not follow the usual rule.

The ability to withstand peer pressure and the randomness of personal lives are in principle two different things, and if one wants to include them both one needs two different temperatures T_1 and T_2 [15], which do not exist in traditional physics.

Ising Model

In the model published by Ernst Ising in 1925, the variables S_i are not 0 or 1, but ± 1:

$$E = -\sum_{i,k} S_i S_k - B \sum_i S_i \tag{2}$$

and for $B = 0$ this corresponds to the above Coke versus Pepsi example. The first summation runs over all neighbor pairs, the second over all sites. Thus, if site i considers changing its variable, the energy change is $\pm\Delta E = 2(\sum_k S_k - B)$ and enters through $\exp(-\Delta E/T)$ into the probabilities to flip S_i; now k runs over the neighbors of i only. (If instead of flipping one S_i one wants to exchange two different variables S_i and S_j, moving S_i into site j and S_j into site i, then one has to calculate the energy changes for both sites i and j in this "Kawasaki" kinetics.) A computer program and pictures from its application are given elsewhere in this Encyclopedia ► Opinion Dynamics and Sociophysics or in [5].

In physics, the S_i are magnetic dipole moments of the atoms, often called spins, and B is proportional to the magnetic field. Usually, physicists write an exchange constant J before the first sum, but we set $J = 1$ for simplicity here. The model was invented to describe ferromagnetism, like in the elements iron, cobalt or nickel. Later it was found to describe liquid–vapour equilibria and other phase transitions. We know that iron at room temperature is magnetic, and this corresponds to the fact that for $0 < T < T_c$ and under zero field B, the majority of its spins point in one direction (either mostly +1 or mostly −1), while for $T > T_c$ half of the spins point in one and the other half in the opposite direction. The magnetization $M = \sum_i S_i$, often normalized by the number L^d of spins, is therefore an order parameter. The critical temperature T_c is often named after Pierre Curie.

In one dimension, we have $T_c = 0$; in the square lattice in two dimensions we know $T_c = 2/\ln(1 + \sqrt{2})$ exactly, while on the simple cubic lattice $T_c \simeq 4.5115$ is estimated only numerically. Of course, one has generalized the model to more than nearest neighbors, to more than two states ± 1 for each spin, and to disordered lattices and networks.

Percolation

Simpler than the Ising model but less useful is percolation theory, reviewed more thoroughly in this Encyclopedia in the section edited by M. Sahimi. Each site of a large lattice is randomly occupied with probability p, empty with probability $1 - p$, and clusters are sets of occupied neighboring sites. There is one percolation threshold p_c such that for $p < p_c$ only finite clusters exist, for $p > p_c$ also one infinite cluster, and at $p = p_c$ several infinite clusters may co-exist, which are fractal: The number of occupied sites belonging to the infinite clusters varies at p_c as L^D where D is the fractal dimension. Here "infinite" means: spanning from one end of the sample of L^d sites to the opposite end, or: increasing in average number of sites with a positive power of L. In one dimension, again one has no phase transition ($p_c = 1$), on the square lattice $p_c \simeq 0.5927462$ and on the simple cubic lattice $p_c \simeq 0.311608$ are known only numerically, with a fractal dimension of 1, 91/48 and $\simeq 2.5$ in one to three dimensions.

In the resulting disordered lattices, each site has from 0 to z neighbors, where z is the number of neighbors in the ordered lattice $p = 1$. If one neglects the possibility of cyclic links one finds $p_c = 1/(z - 1)$ in this Bethe lattice or Cayley tree. Near this percolation threshold the critical

exponents with which several quantities diverge or vanish are the same as in the random graphs of Erdös and Rényi. But this percolation theory was published nearly two decades earlier, in 1941 by P. Flory, later to become a chemistry Nobel laureate.

Mean Field Approximations

What is called "mean field" is similar to the "representative agent" theory in economics, and is widespread in chemistry where the changes in the concentrations of various reacting compounds are approximated as functions of these time-dependent concentrations. A particularly simple example is Verhulst's logistic equation $\mathrm{d}x/\mathrm{d}t = ax(1-x)$, known as Bass diffusion in economics. We now explain why this approximation is unreliable.

Let us return to the above-mentioned Ising model of Eq. (2) and replace the S_k in that equation by it's average $\langle S_k \rangle = m = M/L^d$ which is a real number between -1 and $+1$ instead of being just -1 or $+1$; m is the normalized magnetization. Then the total energy E is approximated as the sum over single energies E_i:

$$E = \sum_i E_i \ , \quad E_i = \left(-\sum_k \langle S_k \rangle - B \right) S_i = -B' S_i$$

with a mean magnetic field $B' = B + \sum_k \langle S_k \rangle = B + mz$ where z again is the number of lattice neighbors. The system now behaves as if each spin S_i is in an effective field B' influenced only by the average magnetization m and no longer directly by its neighbors S_k. The two possible orientations of S_i have the energies $\pm B'$, giving an average

$$m = \langle S_i \rangle = \tanh(B'/T) = \tanh[(B + zm)/T] \qquad (3)$$

and thus a self-consistency equation for m. Expanding the hyperbolic tangent into a Taylor series for small m and B we get

$$B = (1 - z/T)m + m^3/3 + \ldots \qquad (4)$$

which gives a Curie temperature $T_c = z$, since for $T < T_c$ the magnetization is $m = \pm[3(z/T - 1)]^{1/2} \propto (T_c - T)^{1/2}$. Similar approximations for liquid–vapour equilibria lead to the Van der Waals equation of 1872, which may be regarded as the first quantitative theory of a complex phenomenon. (m in that case is the difference between the liquid and the vapour density.) Nowhere in Eqs. (2) and (3) have we put in that there is a phase transition to ferromagnetism; it just arises from the very simple interaction energy $S_i S_k$ between neighboring spins, and similarly the formation of raindrops emerges from the interaction between the molecules of water vapour. The water molecule is the same H_2O in the vapour, the liquid or the ice phase.

But this nice approximation contradicts the results mentioned above. For the chain, square and simple cubic lattice it predicts $T_c = z = 2, 4$ and 6 while the correct values are 0, 2.2 and 4.5. Particularly in one dimension it predicts a phase transition at a positive T_c while no such transition is possible: $T_c = 0$. This was the main result of Ernst Ising's thesis in 1925. And even in three dimensions, where the difference in T_c between 4.5 and 6 is less drastic, the above square-root law for m is wrong, since m varies for T slightly below T_c roughly as $(T_c - T)^{0.32}$. Thus, mean field theory, Van der Waals equation, and similar approximations averaging over many particles are at best qualitatively correct. They become exact when each particle interacts equally with all other particles.

Analogously for percolation, Flory's approximation of neglecting cyclic links and the Erdös-Rényi random graphs lead to results corresponding to mean field approximations and should not be relied upon in one, two or three dimensions with links between nearest neighbors only.

For cellular automata a particularly drastic failure of analogous mean field approximations (differential equations) was given by Shnerb et al. [16] for a biological problem. Even simpler, many cellular automata on the square lattice lead to blinking pairs of next-nearest neighbors: at even times one site of the pair is 1 and the other is 0, while at odd times the first is 0 and the second is 1. Averaging over many sites destroys these local correlations which keep the blinking pair alive.

Applications

A thorough review of "sociophysics" was given recently by Castellano et al. [17], and a long list of references by Carbone et al. [18]. In this Encyclopedia some work by social scientists is reviewed by Davidsson and Verhagen in the section on agent-based simulations in sociology, while Troitzsch in this section reviews both social scientists and physicists. His book with Gilbert [29] is, of course, more complete. Thus, we merely sketch here some of the areas covered in greater detail in the other articles or in the cited literature.

Elections

A social scientist may be interested in predicting the fate of one particular party or candidate in one particular election, or to explain it after this election. A physicist, accustomed to electrons, hydrogen atoms and water molecules

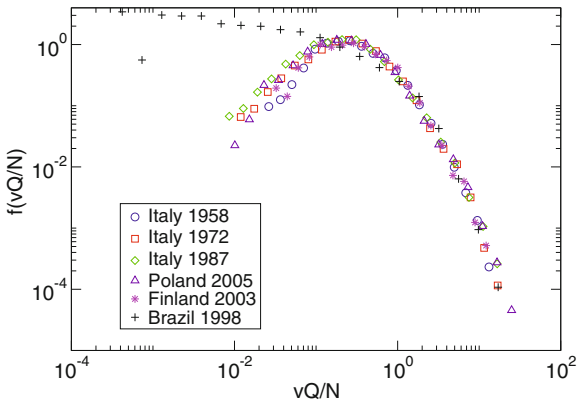

Physics and Mathematics Applications in Social Science, Figure 1
The vote distribution in several countries and elections is a function only of the scaled variable *vQ/N*. From [19]

being the same all over the world may be more interested to find which universal properties all elections have in common. Figure 1, kindly provided by Santo Fortunato, is an example. Let v be the number of votes that a candidate obtained, Q the number of candidates in that election, and N the total number of votes cast. Then the probability distribution $P(v, Q, N)$ for the number of votes is actually a function $f(vQ/N)$ of only one scaled variable, and that variable vQ/N is the ratio of the actual number v of votes per candidate to the average number N/Q of votes per candidate. Various countries and various elections, all using a proportional election system, gave the same curve $f(vQ/N)$ which is a parabola on this double-logarithmic plot and thus corresponds to a log-normal distribution. In Brazil, however, where the personality of a candidate plays a major role, with the party affiliation of the candidate playing a lesser role, the results were different. These authors also present a model to explain the log-normal distribution [19].

Other models for opinion dynamics are reviewed elsewhere in this Encyclopedia ► Opinion Dynamics and Sociophysics or in [5]. A more cognitive approach for interacting agents is their realization by neural network models [20,21,22], ► Social Cognitive Complexity.

Financial Markets

Agent-based simulation of stock markets [23] are a typical example of complex systems applications: In these models not the single agent but their (unconscious) cooperation produces the ups and downs on the stock market, the bubbles and the crashes. These models deal with the more or less random fluctuations, not with well founded market changes due to new inventions or major natural catastrophes.

Real markets give at each time interval a return r which is the relative change of the price. Typically, an index of the whole market like Dow Jones changes each trading day by about one percent. Much larger fluctuations are rarer, and the probability to have a change larger than r decays for large r as $1/r^3$: Fat tails, compared with normal Gaussian distributions. The sign of the change is barely predictable, but its absolute value is: Volatility clustering. Thus, in calm times when $|r|$ was small, tomorrow's $|r|$ probably is also small, whereas for turbulent times with high $|r|$ in the past one should also expect a large $|r|$ tomorrow. The daily weather behaves similarly: presumably tomorrow will be like today.

A simple model, going back to Bachelier more than a century ago, would throw a coin to determine whether the market tomorrow will go up or down. This simple random-walk or diffusion model was shown by Mandelbrot in the 1960s not to describe a real market; it lacks fat tails and volatility clustering but may be good for monthly changes. Many better agent-based models have been invented during the last decade and reproduce these real properties; the Cont–Bouchaud model is based on the above percolation theory (see Fig. 2) [24], while the Minority Game is based on the hypothesis that it is better not to be with the big crowd [25].

Languages

The versatility of human languages distinguishes us from the simpler communication systems of other living beings. With computers or mathematically exact solutions [26] models have been studied for the learning of a language by children or for the evolution of human languages out of simpler forms.

Closer to simulations in biology with the Darwinian selection of the fittest are the models of competition between various languages of adult humans: Will the Welsh language survive against English in Great Britain? Similar to Lotka–Volterra equations for prey and predator in biology, some nonlinear differential equations [12] seem to describe the extinction of the weaker language. Statistics better than in [12] are available for the size distribution of languages, where "size" is the number of people speaking this language, and for the number of languages in one language family. Here one model from de Oliveira et al. [28] found good agreement with reality, see Fig. 3; other models [27] were less successful, in spite of many simulations from physicists.

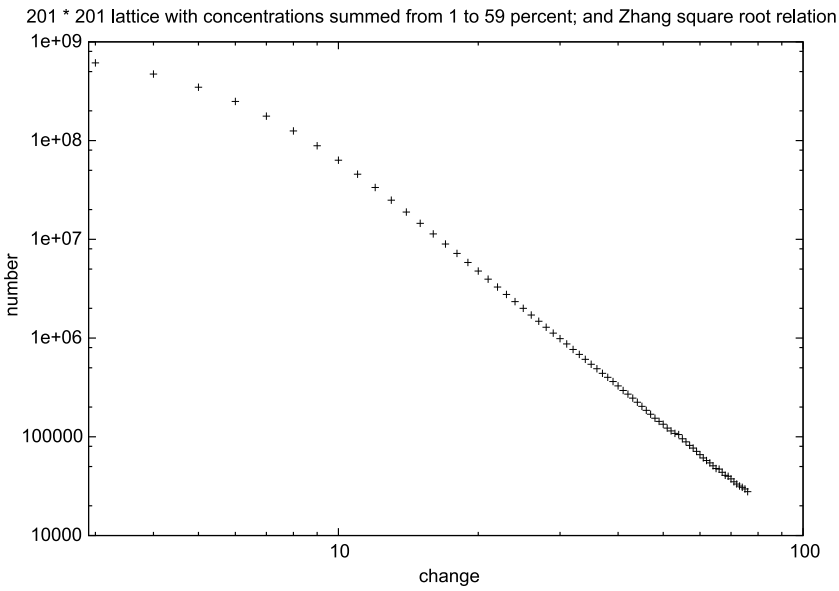

Physics and Mathematics Applications in Social Science, Figure 2
Simulated return distribution in the Cont–Bouchaud percolation model of stock markets [24]. The asymptotic slope to the right is about — 2.9. We plot the number of simulated events for a given absolute value of the market change (in arbitrary units) versus that value

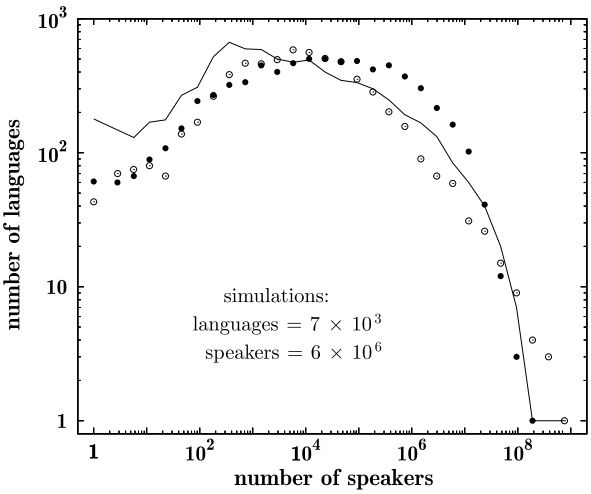

Physics and Mathematics Applications in Social Science, Figure 3
Simulated size distributions for human languages (*full circles*, and *line*), compared with reality (*open circles*). From [28]

Future Directions

The future should see more work on what we have shown here through our three figures: Searching for universal properties, or the lack of them, in the multitudes of models and in reality. Biology became a real science when the various living beings were classified into horses, mammals, vertebrates etc. Within each such taxonomic set all animals have certain things in common, which animals in other taxonomic sets do not share. This check for universality is different from improving our ability to ride horses. Thus, making money on the stock market, or explaining the crash of 1987, is nice, but investigating the exponents of the fat tails, Fig. 2, of all markets may give us more insight into what drives a market and what differences exist between different markets. Winning one particular election and predicting the winner is important, but universal scaling properties as in Fig. 1 may help us to understand democracy better. Preventing the extinction of the French language in Canada is important for the people there, but explaining the overall statistics of languages in Fig. 3 is relevant globally.

It is in these general aspects where the methods of mathematics and physics seem to be most fruitful. One specific problem is better solved by the local people who know that problem best, not by general simplified models.

We thank G. Weisbuch for comments on this manuscript.

Bibliography

Primary literature

1. Schelling TC (1971) J Math Sociol 1:143
2. Dethlefsen E, Moody C (1982) Byte 7:178; Jones FL (1985) Aust N Z J Sociol 21:431

3. Stigler GJ (1964) J Bus 37:117
4. Kim GW, Markowitz HM (1989) J Portf Manag 16:45
5. Stauffer D (2008) Opinion Dynamics and Sociophysics. arXiv:0705.0891
6. Mantegna RN (2005) Presentation of the English translation of Ettore Majorana's paper: The value of statistical laws in physics and social sciences. Quant Finance 5:133
7. Weidlich W (2000) Sociodynamics; A Systematic Approach to Mathematical Modelling in the Social Sciences. Harwood Academic Publishers, Amsterdam; reprint: Dover, New York
8. Galam S (2008) Int J Mod Phys C 19:409; Galam S, Gefen Y, Shapir Y (1982) J Math Sociol 9:1
9. Roehner B, Wiese KE (1982) Environ Plan A14:1449
10. Gomes MAF, Vasconcelos GL, Tang IJ, Tang IR (1999) Physica A 271:489
11. Zanette D (2001) Adv Complex Syst 4:281
12. Abrams MH, Strogatz SH (2003) Nature 424:900
13. http://nobelprize.org/nobel_prizes/chemistry/laureates/1921/soddy-bio.html
14. Stauffer D (1991) J Phys A 24:909
15. Ódor G (2008) Int J Mod Phys C 19:393
16. Shnerb NM, Louzoun Y, Bettelheim E, Solomon S (2000) Proc Natl Acad Sci USA 97:10322
17. Castellano C, Fortunato S, Loreto V (2009) Statistical physics of social dynamics. Rev Mod Phys (to be published). arXiv:0710.3256
18. Carbone A, Kaniadakis G, Scarfone AM (2007) Eur Phys J B 57:121
19. Fortunato S, Castellano C (2007) Phys Rev Lett 99:138701
20. Weisbuch G, Duchateau-Nguyen G (1998) J Artif Soc Soc Simul 1(2), paper 2 (electronic only: jasss.soc.surrey.ac.uk)
21. Wischmann S, Hulse M, Knabe JF, Pasemann F (2006) Adapt Behav 14:117
22. Stauffer D (2007) Opinion Dynamics with Hopfield Neural Networks. arXiv:0712.4364
23. Levy M, Levy H, Solomon S (2000) Microscopic Simulation of Financial Markets. Academic Press, New York; Samanidou E, Zschichang R, Stauffer D, Lux T (2007) Rep Progr Phys 70:404; Lux T (2007) Stochastic behavioral asset pricing models and the stylized facts. In: Hens T, Schenk-Hoppé K (eds) Handbook on Financial Economics. (draft)
24. Stauffer D (2001) Adv Complex Syst 4:19
25. Challet D, Marsili M, Zhang YC (2004) Minority Games. Oxford University Press, Oxford
26. Komarova NL (2004) J Theor Biol 230:227
27. Schulze S, Stauffer D, Wichmann S (2008) Commun Comput Phys 3:271
28. de Oliveira PMC, Stauffer D, Wichmann S, de Oliveira MS (2008) A computer simulation of language families. J Linguist 44:659; arXiv:0709.0868
29. Gilbert N, Troitzsch KG (2005) Simulation for the Social Scientist, 2nd edn. Open University Press, Maidenhead

Books and Reviews

Stauffer D, de Oliveira SM, de Oliveira PMC, Sá Martins JS (2006) Biology, Sociology, Geology by Computational Physicists. Elsevier, Amsterdam

Billari FC, Fent T, Prskawetz A, Scheffran J (2006) Agent-based computational modelling. Physica, Heidelberg

Polymer Physics

T. C. B. McLeish

UK Polymer IRC, Department of Physics and Astronomy, University of Leeds, Leeds, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Single Polymer Chain Physics
Equilibrium Properties of Many-Chain Fluids
Dynamics of Polymeric Fluids
Multi-Phase Polymeric Fluids
Future Directions
Bibliography

Glossary

$\boldsymbol{R}$ Radius of gyration of a polymer chain

$\boldsymbol{N}, \boldsymbol{M}$ Degree of polymerization or number of monomers in a polymer chain, molecular weight.

$\boldsymbol{\xi}$ The screening length or correlation length in semi-dilute polymer solutions; the length over which local density is dominated by a single chain.

$\boldsymbol{N_e}, \boldsymbol{M_e}$ Entanglement degree of polymerization and molecular weight.

$\boldsymbol{l_p}$ Persistence length of polymer chain.

$\boldsymbol{\nu}$ "Flory" exponent of a polymer chain relating R and N so that $R \sim N^{\nu}$.

$\boldsymbol{S(k)}$ Scattering structure factor from a polymeric fluid as function of scattering vector $k = 4\pi \sin\theta/\lambda$ where λ is the wavelength and θ the scattering angle of the experiment.

$\boldsymbol{\Pi(c)}$ Osmotic pressure of a solution as a function of concentration c.

$\boldsymbol{\sigma_{ij}}$ Components of the stress tensor.

$\boldsymbol{d}, \boldsymbol{D}$ Dimensions of a macromolecular object and its embedding space.

$\mathbf{R}(\boldsymbol{n}, \boldsymbol{t})$ Functional description of a macromolecular contour as functions of monomer number n and time t.

$\boldsymbol{G(t)}$ Time dependent relaxation modulus.

$\boldsymbol{\eta}$ Viscosity.

$\boldsymbol{k_B}$ Boltzmann's constant.

$\boldsymbol{\chi}$ Flory interaction parameter between monomers of different chemistry.

Definition of the Subject

Physics is uniquely endowed among the sciences with complete freedom from restriction to any particular do-

main of the physical world. It is able to turn its particular outlook on the scientific program and its special set of experimental and theoretical tools to most material phenomena. In particular it is not limited to any particular length scale, but is sensitive to the emergence of new structures and processes of any size. So macromolecular science, born of the chemistry of the early 20th century, soon gave rise to a branch of physics that seeks to understand the special phenomena of polymer molecules and polymeric matter. Polymers are giant, usually linear molecules constructed as covalently-bonded chains of identical units, or monomers. Individual polymer molecules may contain hundreds or even millions of monomers. Initially disfavored by an organic chemistry community that prized exactitude, polymers were largely ignored since their molecular weight was inexact within a single sample. However, biology knew long before we did that the properties of polymers met the essential requirements of living organisms. Polymers constitute nature's scaffolds from the macroscopic (bone, collagen) to the microscopic (filamentous actin, polymerized tubulin), her force-generators (myosin, dynein and kinesin protein polymers) her information-processing networks (peptide-binding proteins, polysaccharides) and her instruction sets (the nucleic acid family of polymers including DNA). Much of the reason for this lies in the unique set of physical properties of polymeric matter. Some of these have now familiar application in plastic materials from packaging to high-performance fiber and even addressable polymeric electronics. Yet polymer physics is much more than the application of statistical mechanics and spectroscopy to a class of molecular matter; it has taught our discipline about some of its own deep structural connections. Path-integral techniques at the heart of theoretical polymer physics were adopted from readily developed tools in quantum field theory and magnetism. Flexible linear-like objects occur in many other avenues of the subject, superconductivity and plasma turbulence to name two. Above all polymer physics has provided us with a classic example of emergent simplicity from bewildering complexity, so beloved of our subject. It shows every indication of providing in future an essential Ariadne's thread to guide us in our exploration of complexity itself.

Introduction

The fascinating physics of flexible polymers flows from both necessity and beauty. Born of the early investigations into the phenomenon of the elasticity of natural rubber [1,2], then out of the rapid growth in synthetic polymer materials in the post-war years, the need to understand and control the processing of such highly viscoelastic liquids as polymer melts, and to understand the properties of the resulting materials led rapidly to fundamental investigations led by physicists who saw an opportunity to explore the fundamental structures of a new class of matter. Flory [3], Zimm and Stockmayer [4] and Edwards [5] asked how large would macromolecules, linear or branched, be, while Zimm [6] and Rouse [7] asked how such giant molecules would move. Paralleling developments in solid-state many-body physics, the focus of investigations moved from single-chain to many-chain systems, which we review below in Sects. "Single Polymer Chain Physics" and "Equilibrium Properties of Many-Chain Fluids". These pioneers were already using a beautiful notion that was to take hold of condensed-matter physics in the mid 20th century – that of *universality*, or the independence of physical phenomena from local, small-scale details. The emergence of universal properties is usually associated with "critical phenomena" [8], since near phase transitions, the spatial scale of correlated fluctuations may hugely exceed molecular dimensions. Any properties that depend on these fluctuations (an example would be compressibility of a fluid near its critical point, and especially the *exponent* with which it vanishes as the temperature tends to its critical value) will then be insensitive to molecular detail. In field theories of both condensed and high-energy matter, the field-fluctuations "renormalize" microscopic constants into new emergent numbers on which the physics at coarser length scales (or lower energies) may be built (a famous example is the charge of the electron). Although there is at first glance no apparent neighboring critical point in the case of polymeric fluids, both universality in exponents and renormalized quantities appear in abundance. Moreover, there is a natural large number associated with mesoscopic, rather than microscopic lengthscales. The defining feature of a polymer is, after all, its large "degree of polymerization", N, the number of monomers linked together covalently to form the polymer chain. (The literature discusses interchangeably N and the *molecular weight* M of the chains, given in terms of the monomer molecular weight m_0 by $M = Nm_0$). At the most basic level of inquiry into polymer structure, experiments and simulations asking how the average end-to-end distance R of a polymer molecule in solution depends on its degree of polymerization N, began to suggest a universal scaling behavior

$$R \sim N^{\nu} \tag{1}$$

with an exponent ν, dependent only on the embedding dimension d and first calculated by Edwards to be $\nu = \frac{d}{d+2}$. It assumes a rather larger value in solution ($\simeq 0.59$) than

the simple random walk value of 0.5 [9], due to the self-exclusion of the monomers. More phenomena reminiscent of other areas of condensed matter appeared at the level of many-body effects. In the dense limit of polymer melts and concentrated solutions, where chains are highly overlapped (and in embedding dimensions greater than 4, the "upper critical dimension" of the self-exclusion problem), the exponent ν reassumes the value of 1/2 of the ideal Gaussian random walk ("Gaussian" because the ensemble of spatial end-to end vectors of the polymer chains is normally distributed). Closer inspection revealed this to be true above a "screening length", introduced into polymer physics by Edwards [10]. The screening length ξ itself may be directly measured by neutron scattering, and depends on concentration via another universal scaling exponent, related to ν [11]

$$\xi \sim c^{\frac{-\nu}{3\nu-1}} . \tag{2}$$

The picture we have build up of a many-chain polymer solution so far is summarized in Fig. 1, where atomic detail at the monomer level is far below the resolution of the diagram. As the polymer concentration increases, so the screening length or "mesh size" decreases. A typical strand of chain, whose end-to-end distance is ξ, dominates the monomer concentration within the volume it spans.

Both experimental and theoretical evidence of universality continued to build up. Even in the case of dynamics, the many-chain system of a polymer melt followed the ideal, local-dissipation theory of Rouse [7] for sufficiently low molecular weight chains (see below Sect. "Dynamics of Polymeric Fluids") that assumed ideal Gaussian chains. It became clear that Rouse's result can be seen as a fixed "point" of all theories of polymer dynamics in which linear connected objects are ideal and are subject to local dissipation (in dilute solution, far-field hydrodynamics destroys this locality [6]). For example, lattice models of polymer dynamics with local update rules renormalize to the continuum Rouse theory at large enough length scales [12].

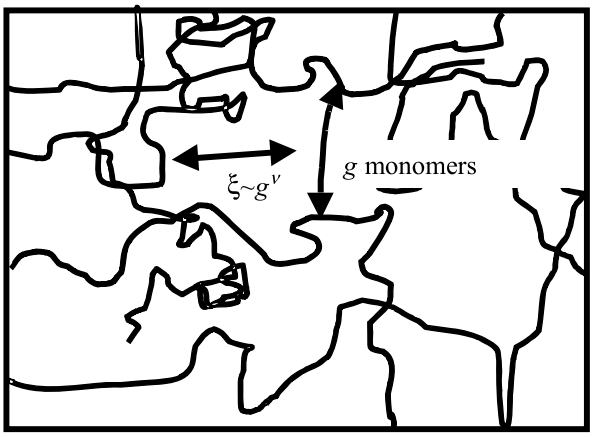

Polymer Physics, Figure 1
Schematic picture of universal structures of screening (overlap) length ξ and the number of monomers g that just spans ξ

It seemed as though the huge connectivity of macromolecules acts to freeze-in long-range order, even though there is no true thermodynamic transition nearby. Such suspicions were confirmed by the demonstration of direct isomorphisms of the calculation of statistical mechanical partition functions of polymers, both dilute and concentrated, onto idealized spin-lattice models of magnetism [9]. It is indeed the high molecular connectivity, as the inverse of the degree of polymerization, N^{-1}, that plays the part of proximity to the distance from a critical point in the spin model

$$N^{-1} \sim \varepsilon \equiv \frac{T - T_c}{T_c} . \tag{3}$$

So by exhibiting physics in which an ensemble of macromolecules of polystyrene (PS) exhibits the same emergent behavior as polyisoprene (PI) or polybutadiene (PB), following scaling laws, and tractable by application of statistical mechanical field theories [8], polymer physics drew together many of the strongest conceptual strands of the century.

More, however, has proved to be true in the realm of *topological* effects. The polymer melts of industrial polymer processing are very highly overlapped on the molecular level, where it becomes immediately apparent that molecular relaxation processes controlling elastic stress are prolonged to very long times indeed. All the important phenomenology is covered in Ferry's seminal survey of polymer viscoelasticity [13]. Mechanical experiments restricted to a range of intermediate timescales of the plateau are hardly able to distinguish between the polymer melt and a rubber, in which the chains are permanently cross-linked to each other at very rare points, sufficiently for each chain to be permanently immobilized from large-scale diffusion. Conceptually, the absent "cross-links" were replaced in the minds of engineers and physicists alike by "entanglements" [13]. These loosely-defined objects were assumed to represent the topological constraint that covalently-bonded molecular chains may not pass through each other. The effective distance between these objects could be calculated, employing rubber elasticity theory (see below), to deduce the degree of polymerization between entanglements N_e, or the equivalent "entanglement molecular weight", M_e. The number

N_e consistently turned out to be of order 10^2, indicating a length-scale for an "entanglement spacing" of 50–100 Å, depending on the particular chemistry. This is highly significant for us, because it shows that small chains on the threshold of feeling topological interactions are real polymers, already long enough to show to a good approximation all the universal properties of statistical connected chains. It also suggests that the role of topology in highly-entangled ($N \gg N_e$) polymer fluids has the potential to be treated universally. Further evidence of universality in entanglements came from experiments in which the polymers were diluted to a volume fraction ϕ_p by a compatible solvent, indicating that $M_e \sim \phi_p^{-\alpha}$ where the scaling exponent $\alpha \simeq 1$ [11]. The dependence of melt viscosity η (at fixed temperature) on molecular weight also exhibits remarkable universality over very many different polymer chemistries [13] closely matched by $\eta \sim M^{3.4}$, providing that the molecular weight lay well above the entanglement threshold M_e suggested by their high-frequency elastic modulus. We review this rapidly-progressing area in Sect. "Dynamics of Polymeric Fluids".

New phenomena arise when different chemistries of monomer are introduced into the same chain. Effective repulsive interactions between heterogeneous monomers create a tendency for strong spatial correlations of chemical type. In blends of more than one type of homogeneous polymer, the result is often a demixing transition with near-universal structure and dynamics [17]. When the chemical species are combined into the same chain in regular "blocks" of controlled molecular weight, demixing occurs on the scale of the chains themselves, giving an extremely rich variety of spatially-periodic nanoscopic structures, self-assembled micellar structures, and controlled collapsed forms. Both chain composition and temperature act as control parameters of a structural space that becomes increasingly biomimetic as the information content of the macromolecular sequence increases. Experiment and theory are reviewed in Sect. "Multi-Phase Polymeric Fluids".

Single Polymer Chain Physics

It should not be surprising that the notion of complexity should arise even in the context of the "single particle" domain of polymer physics: that of a single molecule. For already a macromolecule contains many degrees of freedom that are coupled in non-linear ways. Not only this, but also at the single-chain level emergent co-operative properties arise. We briefly survey three important cases of single-chain physics: the non-interacting chain, the effect of excluded volume and the role of charge.

Ideal Non-interacting Chains

The first and most fundamental of these underlies the physics of rubber elasticity: it is the emergence of an entropic Hookean spring for macrostates of a single polymer chain in thermal equilibrium defined in terms of its end-to-end vector $\mathbf{R}$. We recap briefly here the statistical physics of a polymer chain, modeled as a *random walk* in space and subject to some local rule for spatial links. An example is the freely jointed chain, in which the orientations of a set of linked rods are uncorrelated. The step length of the chain corresponds to the Kuhn length, b, of the polymer (the shortest independently oriented segment length). It is not as small as a monomer length, but usually 4 or 5 monomers long.

For polymer statistics suppose a whole walk has N links. Let the end to end displacement of an individual chain be $\mathbf{R}(N)$. From the theory of random walks: $\langle R^2(N)\rangle = Nb^2$ and the probability density for the end to end vector $G(\mathbf{R})$ must have Gaussian form (from the law of large numbers, since each vector step is an independent random variable whose sum is $\mathbf{R}(N)$). So

$$G(\mathbf{R}, N) = \left(\frac{3}{2\pi N b^2}\right)^{3/2} \mathrm{e}^{-3R^2/2Nb^2} . \tag{4}$$

The macrostate of an ensemble of such chains is defined by the chain end-to-end vector $\mathbf{R}$. The microstates are the different specific paths through space that have $\mathbf{R}$ as their end-to-end displacement. Each individual path, or microstate of the chain, will be specified if the spatial position of each link is known. We will use the notation $\mathbf{R}(n)$ for the position of the nth link. The full time-dependence of the chain would then be described by the function $\mathbf{R}(n, t)$, extended to the two dependent variables of contour position n and time t. The role of Brownian motion can be cast in the form of Langevin equations for $\mathbf{R}(n, t)$ (see Rouse model, Sect. "Dynamics of Polymeric Fluids"), but here we exploit it as a generator of ergodic exploration of all chain configurations in the ensemble. The number of configurations with fixed end to end vector $\mathbf{R}$ is just the corresponding fraction of total microstates $\Omega(\mathbf{R}) = \Omega_{\text{TOT}} P(\mathbf{R})$. Since the entropy of the walk $S = k_B \ln \Omega(\mathbf{R})$ we have $S(\mathbf{R}) = \text{const.} - \frac{3k_B R^2}{2Nb^2}$. The conformational free energy of the chain $F(\mathbf{R}) = U - TS$ has $U = 0$ since there are no sources of internal energy. This yields for the free energy of a chain of fixed end-to end vector $F(\mathbf{R}) = \frac{3k_B T R^2}{2Nb^2}$. Finally we may derive the thermodynamic force (or "Brownian tension") on the chain end-to-end vector as

$$f = -\nabla F(\mathbf{R}) = -\frac{3k_B T}{Nb^2}\mathbf{R} \tag{5}$$

and recognize a linear elastic spring law. I. e., a random walking polymer at finite T is a Hookean spring with spring constant $\propto T/N$.

The analogies between the statistical mechanics of random walks and the quantum mechanics of spinless particles also emerges from this analysis. Applying Eq. (4) to small subchains of path length ΔN allows the (careful) taking of a limit so that

$$G(\mathbf{R}, \Delta N) = \left(\frac{3}{2\pi \Delta N b^2}\right)^{3/2} \mathrm{e}^{-\frac{3}{2b^2}\left(\frac{\partial \mathbf{R}}{\partial N}\right)^2 \Delta N} .$$

The notation G for the probability distribution is suggestive: this descriptor of the chain has the structure of a propagator. The complete end-to-end propagation of (4) can be written as the sum over all possible intermediate positions of the meeting points of all smaller subchains, which in turn is just an example of a Feynman path integral over all possible paths of the polymer contour $\mathbf{R}(n)$

$$G(\mathbf{R}, N) = \int_{\mathbf{R}(0)=\mathbf{0}}^{\mathbf{R}(N)=\mathbf{R}} \mathrm{e}^{-\frac{3}{2b^2}\int_0^N \left(\frac{\partial \mathbf{R}}{\partial N}\right)^2 \mathrm{d}N} \text{Đ}\left[\mathbf{R}(n)\right] . \quad (6)$$

The propagator structure arises because the properties of the chain at equilibrium is governed by its partition function, which in turn is a sum over microstates that are in this case just the possible paths of the chain. An analogous integral arises in Feynman's form of quantum mechanics because the sum over all trajectories that gives the (complex) amplitude for particle propagation is just the same geometrical set of paths. The difference is that in polymer statistical mechanics the phase angle attributed to the path is imaginary, giving a real argument of the exponential in the path integral. A rich set of techniques flow naturally from this analogy: since the propagator also obeys Schrödinger's equation, the equilibrium configuration of ideal chains in confined geometries and external potentials can be solved by any technique developed for the quantum mechanical case [9].

Experiments on single chains have until recently been indirect ensemble measurements. However, neutron-scattering can give averaged single chain properties because of the very different scattering lengths of hydrogen and deuterium nuclei. It is relatively straightforward to replace chemically some or all of the hydrogen atoms in a fraction of the chains in a polymeric fluid. When the chains themselves are monodisperse, the small-angle scattering pattern is identical to that of the population of labeled chains. Figure 2 shows an example of scattering from a 7% labeled fraction of polystyrene chains with a narrow distribution of molecular weight. Early experiments of this type showed the remarkable result illustrated here that the

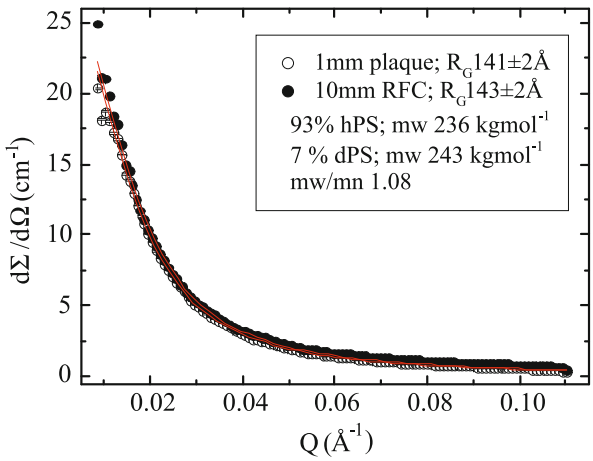

Polymer Physics, Figure 2
Small angle neutron scattering (SANS) data averaged over angles for a monodisperse polystyrene melt. The theoretical curve is that calculated from a Gaussian propagator

single chains in a densely packed melt actually assume an ideal (effectively non-self-interacting) set of configurations described by the Gaussian propagator of Eq. (4) (see Sect. "Equilibrium Properties of Many-Chain Fluids" below). Much more recently the advent of recombinant DNA, fluorescent labeling and video confocal microscopy has begun to make direct inspection of individual polymer molecules possible in restricted circumstances. Even very high molecular weight DNA has a random-walk molecular dimension ($bN^{1/2}$) below the resolution limit of optics, but if the chain is stretched out much larger dimensions are accessible approaching bN, so that some of the predictions of single-chain elasticity can be explored.

Figure 3 shows one famous example that actually illustrates chain response as the maximum elongation is approached. In this limit the Gaussian approximation breaks down badly and a divergence of force with extension is measured. The form of the divergence is, unlike the linear entropy-dominated range of elasticity, not universal among local polymer chemistries, but depends on the form of the local structure and its response to tension. Another analogy with high-energy physics arises here: *strong* applied forces measure structure at *smaller* length scales. This is powerfully illustrated by the high-force asymptote of two models of flexible polymers, the Freely Jointed Chain (FJC) and the Worm–Like Chain (WLC). The first models the local structure of a chain as$N$ freely-hinged but infinitely stiff rods each of length b. Such a chain has a maximum extension of its own contour length $L_0 = Nb$, but at low forces responds with the linear behavior of the Gaussian chain so that the force f with extension L follows

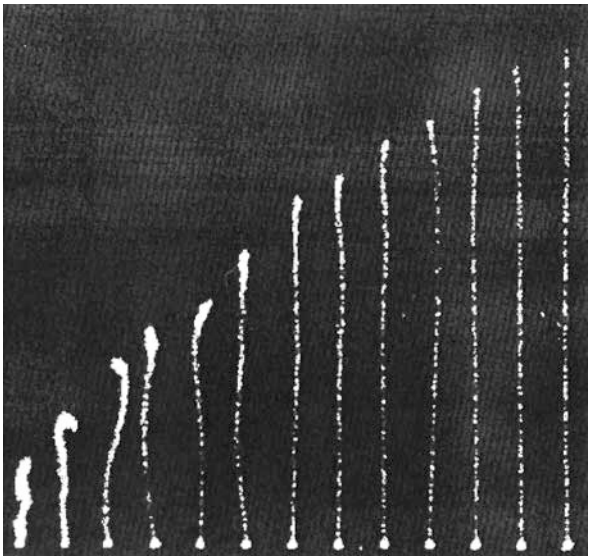

Polymer Physics, Figure 3
Images of a 64μ long DNA molecule held in place at one end by optical tweezers and stretched out by hydrodynamic flow of increasing velocity from *left* to *right*. Reprinted with permission from [19]

$f \sim k_B T(L/Nb^2)$. The force diverges as $L \to L_0$ with the asymptotic form $f(L) \sim (1 - L/L_0)^{-1}$. The WLC introduces a finite bending rigidity everywhere along the chain so that the internal energy of a configuration $\mathbf{R}(n)$ can be written

$$E = k_B T l_p \int_{n=0}^{N} \left(\frac{\partial^2 \mathbf{R}(n)}{\partial n^2} \right)^2 \mathrm{d}n \tag{7}$$

with the constraint that $\left| \frac{\partial \mathbf{R}(n)}{\partial n} \right| = 1$. The stiffness is written as $k_B T l_p$ in terms of the "persistence length" l_p because at equilibrium the chain is locally stiff at smaller lengthscales and flexible over longer lengths. This statement can be made exact by considering the correlation function of the orientation of the chain:

$$\left\langle \frac{\partial \mathbf{R}(n_1)}{\partial n} \frac{\partial \mathbf{R}(n_2)}{\partial n} \right\rangle = \mathrm{e}^{-|n_1 - n_2|/l_p} .$$

Although this model also shares the Gaussian linear response of the FJC, it possesses quite different asymptotics at high force [20], following $f(L) \sim (1 - L/L_0)^{-1/2}$. The calculation takes the response under the force of all harmonic normal modes of the Hamiltonian (7), the more gentle divergence arising from the successive suppression of contortions of the chain at smaller and smaller wavelength as the force increases.

Self-Interacting Chains: Excluded Volume

As pointed out in the introduction, real polymer chains in solution are not typically Gaussian. The reason is that the path integral of (6) overcounts the allowable configurations of the chains, including those that cross themselves. The modification to the Hamiltonian that achieves the monomeric self-exclusion with maximum simplicity and generality is the "Edwards Hamiltonian" [21]

$$G(\mathbf{R}, N) = \int_{\mathbf{R}(0)=\mathbf{0}}^{\mathbf{R}(N)=\mathbf{R}} \mathrm{e}^{-\frac{3}{2b^2}\int_0^N \left(\frac{\partial \mathbf{R}}{\partial N}\right)^2 \mathrm{d}N - w \int_0^N \int_0^N \delta[\mathbf{R}(n) - \mathbf{R}(n')] \mathrm{d}n \mathrm{d}n'} \text{Đ}\left[\mathbf{R}(n)\right] .$$

Although formally the delta-function potential removes only a volume of phase space of zero measure from the path integral, its anticipated use within field-theoretic tools for the solution of the model mean that it represents a renormalized local repulsion between monomers of the chain at the level of this universal coarse-grained theory. From this starting point one can proceed by several methods: [21] self-consistent field treatment of the excluded volume term, mapping onto problems in critical phenomena [23], direct renormalization-group calculation [24] and Monte Carlo numerical enumeration [22]. For a comprehensive and technical review see [25]. The essential structure within the self-consistent field methods is the same as that of an early calculation by Flory [3] who balanced the scaling forms of the free-energy contributions from chain elasticity (R^2/N) and excluded volume (N^2/R^d) in d-dimensional space and minimized to give the dependence of the scaling exponent for chain size ν as

$$\nu = \frac{d}{d+2} . \tag{8}$$

This is fortuitously accurate in all dimensions from $d = 1$ (where it is exact) to $d = 4$, which it correctly identifies as the upper critical dimension of the problem since the Gaussian exponent of $\nu = 1/2$ is recovered here. However, the method is unreliable for the calculation of other quantities because in dimensions less than four the excluded volume term is not perturbative for any strength of the parameter w in the large-N limit. This really forces a renormalization approach to the problem, already achieved in the case of spin interactions in magnets at the ferromagnetic critical point. A formal and beautiful exact mapping first pointed out by de Gennes exits between the excluded volume chain and an analytic continuation of the Heisenberg model in which the number of spin components goes to zero [23]. The analogy arises because the calculation of

the correlation function between the magnetic spins on any two sites of the lattice in the magnet model can be written as a weighted sum over all non-intersecting paths that connect the two sites within the lattice. So formally

$$\langle S_i S_j \rangle = \sum_N \Omega(N) \mathrm{e}^{-\epsilon N} , \tag{9}$$

where $\Omega(N)$ counts the number of self-excluding walks of length N connecting the sites and ε measures the dimensionless temperature difference from the ferromagnetic critical point as in Eq. (3). The analogy led directly to the first calculated values for the Flory Exponent ν using diagrammatic expansion methods developed originally from Feynman's perturbation tools for quantum electrodynamics. In three dimensions the problem is non-perturbative – the exponent departs from the mean-field value for all values of the excluded volume parameter w no matter how small, providing that the chains are long enough. The calculation of ν requires first taking the problem in 4 dimensions, where it becomes perturbative, then taking a double expansion in w and in the analytic continuation of dimension below 4. For short enough segments on the other hand, the chains do not depart strongly from ideal behavior. There is a characteristic length scale at which the energy of self-exclusion equals the thermal energy kT below which statistics are near-ideal and beyond which the chains are swollen. Subchains of this intermediate lengthscale are known as "thermal blobs". In $d = 3$ the value of $\nu \simeq 0.588$, in close agreement with the value predicted (fortuitously) by (8).

The behavior of chains under an attractive 2-body interaction constitutes an emergent phenomenon that mirrors that under repulsion discussed above. The experimental case is that of a "poor solvent" where monomers of the polymer now enjoy a favorable interaction. Now the chains are densely packed for high enough molecular weights so that $\nu = 1/d$ above the thermal blob size. In the limit of infinite molecular weight the transition from a swollen to collapsed chain becomes thermodynamic. This is possible to realize experimentally in some cases by simply controlling the solvent quality though control of temperature. There exists in these cases a critical point known as the "theta temperature" at which the effective 2-body monomer-monomer interactions from excluded volume and solvent-induced attraction exactly cancel, leaving only 3-body and higher terms. The "coil-collapse" transition is not sharp (only becoming so in the thermodynamic limit of infinitely long chains).

The collapsed and swollen chain configurations of real chains leave their traces on the emergent elastic properties of single chains that we examined above in the ideal case. For example, now the effective Hookean spring force-distance relation $f(R)$ is modified to $f \sim R^{\frac{\nu}{1-\nu}}$ in general.

The Role of Electrostatic Charge

A recently very active area in polymer physics that has produced a number of surprises is the form of emergent behavior arising from the combination of electrostatic charge, polymeric connectivity and counter-charges in solution. This is the case of "polyelectrolytes" – polymers containing charged monomers. Complex emergent behavior is perhaps unsurprising since all three have the propensity to generate long-ranged interactions that are candidates for generators of qualitatively different physics from the local interactions examined in the last section.

The static configurations of a polyelectrolyte are radically different from those of a neutral flexible chain in solution. Constructing a mean-field "Flory" type theory for a chain containing a fraction f of monomers carrying charge e that, balancing electrostatic energy with configurational entropy yields the result [26]

$$R \simeq N b f^{2/3} \left(\frac{l_\mathrm{B}}{b} \right)^{1/2} \tag{10}$$

showing that at this level the chains will be completely stretched at the scaling level. This result contains one of the several new lengthscales that appear in complex fluid electrostatics, the Bjerrum length

$$l_\mathrm{B} \equiv e^2 / (\varepsilon k T) \ .$$

This is the distance between charges at which the electrostatic and thermal energies are comparable.

Charged configurations of this extreme kind are not realized globally, since counter-ions are universally present to ensure overall neutrality. Their presence in solution screens the "bare" long range electrostatic repulsion beyond the "Debye screening length"

$$l_\mathrm{D} \equiv \left(\frac{\varepsilon k T}{n_0 e^2} \right)^{1/2} \ .$$

This is not the length-scale on which the charged chains adopt random-walk configurations, however, since there is typically a strong repulsion (many kT) between chain segments adjacent by l_D, providing the charge fraction is great enough. The emergent persistence length is another new scale

$$l_\mathrm{P} = l_\mathrm{D} \left(\frac{l_\mathrm{D} l_\mathrm{B}}{b^2} \right) f^2 \ .$$

A different set of structures arises if the energy competition is between electrostatic repulsion and chain-collapse in a poor solvent. Without electrostatics the chain will collapse into a compact globular form with $\nu = 1/d$ to minimize the total contact between monomers and solvent, restricting it to the surface of the globule. The surface energy rises as γR^2, but is eventually overcome by the electrostatic self-repulsion from the globule's charge, which rises (in three dimensions) as $N^2 f^2 e^2/R \sim R^5$. This instability was identified as a cause of droplet breakup in simple fluids by Lord Rayleigh. The free-energy minimum is achieved in the polymeric case by a configuration resembling a string of pearls, in which the polyelectolyte breaks up into small globules that closely balance the surface and electrostatic energies, connected by stretched strands of chain. It is a rare example of heterogeneous configurations minimizing the free energy at the level of a single chain [27].

The combination of persistent, or rod-like configurations and counter-charges leads to other remarkable properties of charged polymers. The entropy of confinement of counter ions within a cylindrical region around an extended polymer of radius R decreases as $kT \log R$ per ion. In this geometry this has the same functional form as the electrostatic attraction (providing that R is less than the screening length), which depends on the charge per unit length of the polymer ρ as $(\rho e/\varepsilon) \log R$. Providing $(\rho e/\varepsilon) > kT$, most of the counter-ions in solution balancing the charge on the polymer will be closely-bound to it, the phenomenon of "Manning condensation" [28].

The charge clouds condensed in the vicinity of neighboring polymers sustain thermal fluctuations that correlate via the long-range electrostatic interaction in an analogous way to the electronic origin of the Van der Waals interaction. Providing the fluctuations are large enough (these are enhanced by ions of high valency), then the fluctuation-induced attraction between the counter-ion clouds may actually overcome the electrostatic repulsion of the bare polyelectrolyte charge. For polymers in solution like charges may indeed attract! This effect is responsible for the attraction and bundling of polymers of DNA, which carries a strong negative charge [29].

Equilibrium Properties of Many-Chain Fluids

We have seen that the thermodynamic properties of single polymer chains exhibit both richness and a high degree of universality. Both aspects of the emergent properties of polymer physics extend to many-chain systems in which the molecular chains become strongly overlapped. When this happens the quality of the solvent (controlled with temperature) that for single chains gave rise to the coil-collapse transition, swollen statistics and the pseudo-ideal theta temperature, now control co-operative phenomena such as the emergence of an osmotic pressure, and phase separation.

Semi-dilute Solutions

The existence of the coil-size itself $R \sim N^\nu$ sets a new concentration regime, termed "semi-dilute", that only exists in the case of polymeric fluids. We have already encountered it, and its key attributes, in the introductory discussion of Fig. 1. The semi-dilute concentration range covers those cases where the chains themselves are strongly overlapped but where the volume fraction taken up by monomer rather than solvent molecules is still small. The structure of the solution when the polymer-solvent interaction is favorable is entirely dominated in this regime by the new lengthscale of the screening length ξ. The physics of this lengthscale is readily detected in two ways: by direct structural probes such as neutron scattering, and by the solution property of osmotic pressure. If scattering contrast exists between monomer and solvent then the scattering intensity and scattering wavevector k measures the fluctuations in monomer concentration within volumes of fluid of size k^{-1}. For large lengthscales $k \ll \xi^{-1}$, there are no correlations between one element bounded by a screening length and its neighbors: the identity of a single chain is lost at these scales where it is highly overlapped with others. At the other limit, all scattering for $k \gg \xi^{-1}$ is governed by correlations along single chain segments within such correlation volumes, and the scattering is identical to that from a single chain. Since chains have a spatial scaling structure whether in theta or good solvents, power-law behavior is induced in the scattering as well. The physics is captured by a generalization of the Ornstein–Zernicke scattering function:

$$S(k) \simeq \frac{S(0)}{1 + (k\xi)^{1/\nu}} \tag{11}$$

illustrated in the figure [30]. The exponent ν is the same as that determining coil size for single chains discussed in the previous section.

The osmotic pressure as a function of monomer concentration $\Pi(c)$ of a semi-dilute polymer solution is also connected with the structure of closely-packed correlation volumes that emerges from the screening-length picture. This is because Π measures the change of free energy with concentration, itself dominated by the balance of chain entropy and chain contact-energy. Since all sub-chain con-

figurations within each correlation volume ξ^3 are sampled ergodically, but all correlations at greater lengthscales lost, the consequence is that the free-energy density carries the structure of kT/ξ^3, or one thermal degree of freedom per correlation volume. Since the concentration of the correlation length is calculable in terms, once more, of the Flory exponent ν, from Eq. (2), the concentration dependence of the osmotic pressure follows $\Pi(c) \sim c^{3\nu/(3\nu-1)}$ in the semi-dilute regime. The same result can be arrived at by anticipating a scaling structure that crosses-over to the correct ideal-gas form under truly dilute conditions, writing

$$\Pi(c) = \frac{kT}{b^3}\frac{c}{N} f\left(\frac{c}{c^*}\right) . \tag{12}$$

Here $f(x)$ is a scaling function that tends to unity for small argument, and to a power law $f(x) \sim x^z$ for largest. The concentration c^* is the overlap threshold separating dilute and semidilute regimes, where individual coils just begin to overlap. Insisting that for $c \gg c^*$ the osmotic pressure should not depend on chain length N (this is equivalent to the loss of correlations beyond ξ) fixes the power $z = \frac{1}{3\nu-1}$ and the result $\Pi(c) \sim c^{3\nu/(3\nu-1)}$ is recovered. Experiments at different molecular weights and chemistries in good solvents collapse well onto the scaling form of (12) [31]. A formal route to the calculation of the solution properties of correlation, free energy and osmotic pressure was discovered by Des Cloiseaux as a generalization of the magnetic spin-analogy of de Gennes. A diagrammatic expansion of n-body chain interactions is performed for the zero-spin limit of the model, but this time in the presence of an external field. This acts as a fugacity for chains, and creates a system where renormalization-group methods may be used, again in expansion around 4 spatial dimensions, to calculate exponents such as z [25].

Complex Topology Polymers and Gelation The emergent properties of polymers in solution, it should be clear by now, arise from the connectivity of chains, either on its own, or in the presence of other physical interactions such as excluded volume or electrostatics. It is also the essentially topological properties of connectivity that endow polymer physics with its universality. It is therefore of interest to explore the consequences of altering the chain topology in more complex ways. One very practical way of doing this is to introduce cross-links chemically into a polymer solution (industrially this is the route to preparation of rubbers). In the limit of high cross-link density the result is a system of chains that have a significant number of their degrees of freedom "quenched" (chemically connected monomers on different chains are perpetually forced into proximity). That the resulting physical object is a solid when the original system was fluid is by no means an obvious result, and emerges only delicately from the treatment of the statistical mechanics of quenched disorder. Historically, methods originally developed to treat the model systems of "spin-glasses" were the first to treat the cross-linked polymeric fluid from this point of view, among them the "replica method" of Edwards [33]. Here the formal free energy of the rubber as an average over the logarithms of quenched partition functions with different configurations of cross-links $\{N_x\}$:

$$F(N, N_x, \Lambda) = kT \left\langle \ln Z\left(N, \{N_x\}, \Lambda\right)\right\rangle_{\{N_x\}}$$

is treated using the formal limit $\log Z = \lim_{n\to 0}\left(\frac{Z^n-1}{n}\right)$. This amounts to the statistical mechanics of an unquenched ensemble of replicas of the original network, but in the limit (analytically-continued) of the number of replicas tending to zero. The liquid-solid transition emerges as the physical consequence of a mathematical symmetry-breaking between the replicas in the evaluation of the free energy, and a shear-modulus grows as the third power of the difference between the cross-link density and a critical value for the onset of the solid [34].

Although the formal treatment of rubber elasticity is very subtle, there are good approximations that have generated a sequence of semi-empirical models capable of capturing not only stress generation in cross-linked rubbers, but also the temporary stress generated in entangled polymeric fluids (see below). These begin with identifying a length-scale at which the deformation Λ is imposed on chain segments affinely, and below which the chains are assumed to be able to explore all microstates. Applying the result for the effective elasticity of the chain segments (5) above yields an expression for the bulk stress tensor in terms of the average configuration of the subchains:

$$\sigma_{ij} = \frac{3k_{\mathrm{B}}T}{b^2}\mathbb{C}\left\langle \frac{\partial R_i}{\partial n}\frac{\partial R_j}{\partial n}\right\rangle . \tag{13}$$

These results obtain for the case in which the cross-links are suddenly imposed upon an equilibrium semi-dilute or concentrated polymer solution. If they are introduced gradually a richer structure arises in which chains of increasingly branched nature are created well before the critical liquid-solid transition. If this process is imposed mathematically on an ensemble of ideal chains, an ensemble of very compact branched polymers results with a Flory exponent of $\nu = 1/4$. From the definition of ν, (1), we see that its inverse $1/\nu \equiv D$ plays the role of a dimension. Ideal linear chains are in this sense two-dimensional ob-

jects, but ideal branched polymers are four-dimensional! This does not present a necessary difficulty at fixed spatial scales, since the individual molecules of finite molecular weight are very sparse, but at high enough molecular weight there will not be sufficient space in three embedding dimensions to contain such objects without overlap. This must occur in successive cross-linking, because calculations of the distribution of molecular weights that result from random insertion of cross-links (this may be done elegantly by the use of generating functions [9]) gives a result of the form

$$P(M) = M^{-\tau} f(M/M_x)\,, \tag{14}$$

where τ is a "Fisher exponent" and $f(x)$ a cut-off function that limits the distribution function to a highest molecular weight of M_x. This upper molecular weight itself diverges as another scaling function of the difference in cross-link density from the critical value $M_x \sim |p - p_c|^{-1/\sigma}$. This class of critical phenomena generated by topological connectivity is called "percolation" and generates beautiful physical effects in polymers. Like other critical phenomena associated with thermodynamic phase transitions it possesses universality classes in which the values of the exponents are independent of the geometry of local interactions. We can see that the mean-field (ideal) value of $D = 4$ (and $\tau = 5/2$) is not sustainable in 3-dimensional space: the branched polymers must swell by excluded volume so that they do not overlap strongly with themselves (or with subclusters of larger molecules) [35]. Applying this requirement on the non-overlap of clusters of all molecular weights generates a relationship between the exponents τ, D and the embedding dimension d called the hyperscaling relation:

$$\frac{d}{D} = \tau - 1\,. \tag{15}$$

This holds for a wide range of "bare" fractal dimensions D since the excluded volume drives systems to marginal overlap at all length scales stably: if the overlap reduces for larger chains then they suffer less excluded volume and swell less, consequently increasing overlap once more. The values for percolation in 3-dimensional space are $D \cong 2.53$ and $\tau \cong 2.18$. Should the overlap increase then the consequent increased self-repulsion increases swelling with the opposite effect. The structure implied by (15) is self-similar on a grand scale: not only are the individual molecules self-similar, but the ensemble also satisfies self-similarity at all length-scales in the marginal overlap of clusters of all sizes. There is no correlation length and the scattering function is a power law.

Dynamics of Polymeric Fluids

Complex, emergent phenomena extend from static structure of polymeric fluids into their rich dynamic properties. This is especially true of the semi-dilute and concentrated cases when interactions between distinct chains are very strong and the conformational dynamics highly coupled. But even in the case of dilute or effectively uncoupled dynamics the scaling structures we saw in equilibrium pattern the behavior out of equilibrium. The experimental tools deployed also mirror those used for statics: direct structural probes of light or neutron scattering possess dynamic counterparts in photon correlation spectroscopy [36] and neutron spin-echo [37]. Specific averages over bond correlation dynamics are available from dielectric spectroscopy [39] or NMR relaxation [38]. Emergent effects at the level of the fluid are most striking in their rheology [40]. At the level of linear response, the rubber-elastic stress generated by small deformations decays with a function $G(t)$ characteristic of the molecular structures present: the phenomenon of viscoelasticity. In strongly non-linear flows other effects appear, such as the effective decrease in viscosity with shear rate, complex transient behavior in stress response and the generation of rich stress-fields and non-inertial elastic instabilities in non-trivial flows. We first review briefly the dynamic properties of single polymer chains before treating more extensively the complex dynamics of entangled systems.

Single Chain Dynamics

In the simplest fundamental model of polymer dynamics, due to Rouse, we make three key simplifying assumptions [7]. Their physical validity depends on the effectiveness of screening [10] of both static and hydrodynamic quantities in a melt. In the case of concentrated solutions, screening will not be operative at lengthscales below the mesh size ξ, but will hold at larger scales. But these contain the lengthscales of entanglement effects, so in the solution case also, coarse-grained local dynamics are expected to follow the Rouse model locally. Even in concentrated polymeric fluids the dynamics of chains (and subchains) below a characteristic molecular weight are not strongly coupled to the presence of other chains. The central assumptions are:

- Gaussian Chains: in which the force on a subchain segment n is the net entropic force from its neighbors. In the continuum representation we have adopted, this is equivalent to a thermodynamic force at each point of the chain of $-\frac{\partial}{\partial n}\left(\kappa \frac{\partial \mathbf{R}}{\partial n}\right) = -\kappa \frac{\partial^2 \mathbf{R}}{\partial n^2}$ with $\kappa = \frac{3k_B T}{b^2}$.

- Local drag: the drag force on a subchain segment comes from frictional drag against background without long-range hydrodynamic effects of backflow (this works in melts where all long-range mediated backflows are screened). This force is $\zeta \frac{\partial \mathbf{R}}{\partial t}$ with ζ a drag coefficient per segment.
- Brownian motion: a random force $\boldsymbol{f}(n, t)$ acts on each subchain with correlation times much faster than any polymer dynamics to be modeled by the theory.

The monomeric drag constant ζ_0 will parametrize all our theoretical models, setting the timescale for both Rouse and, subsequently, entangled, motion. The balance of entropic, drag and random forces on the chain of N subchains is the Rouse equation:

$$\zeta_0 \frac{\partial \mathbf{R}}{\partial t} = \frac{3k_\mathrm{B}T}{b^2} \frac{\partial^2 \mathbf{R}}{\partial n^2} + \boldsymbol{f}(n, t) \,. \tag{16}$$

The noise on each subchain is related to its frictional drag by the generalized Einstein relation as above:

$$\left\langle \boldsymbol{f}(n, t) \boldsymbol{f}(m, t') \right\rangle = 2\zeta_0 k_\mathrm{B} T \, \mathbf{I} \, \delta(n - m) \delta(t - t') \,. \tag{17}$$

The Rouse dynamical Eq. (16) is diagonalized by the transformation:

$$\mathbf{R}(n, t) = \boldsymbol{X}_0(t) + 2 \sum_{p=1}^{\infty} \boldsymbol{X}_p(t) \cos\left(\frac{p\pi n}{N}\right) \,. \tag{18}$$

The $\boldsymbol{X}_p(t)$ are the time-dependent amplitudes of the "Rouse modes" of the polymer chain. These are just the (vector amplitude) Fourier modes of the chain path $\mathbf{R}(n, t)$ with respect to the arclength coordinate n. The key result for us is the time correlation function of the mode amplitudes, which is:

$$\left\langle \boldsymbol{X}_p(t) \boldsymbol{X}_q(t') \right\rangle = \mathbf{I} \, \frac{k_\mathrm{B}T}{k_p} \delta_{pq} \mathrm{e}^{-|t-t'|/\tau_p} \quad \text{with } k_p = \frac{6k_\mathrm{B}Tp^2\pi^2}{Nb^2} \,. \tag{19}$$

Each mode has its own relaxation time $\tau_p = \zeta/k_p$ that decrease rapidly (as $1/p^2$) with mode index p. The longest of these relaxation times $\tau_1 = \frac{\zeta N^2 b^2}{3\pi^2 k_\mathrm{B} T}$ has special significance. It is known as the *Rouse Time*, and often given the notation τ_R. It is the time for relaxation of the overall shape of the molecule (it is the relaxation time of the amplitude of the normal mode with fewest nodal points, $\cos \frac{\pi n}{N}$), and is also the time for a Gaussian Rouse chain to diffuse its own radius of gyration.

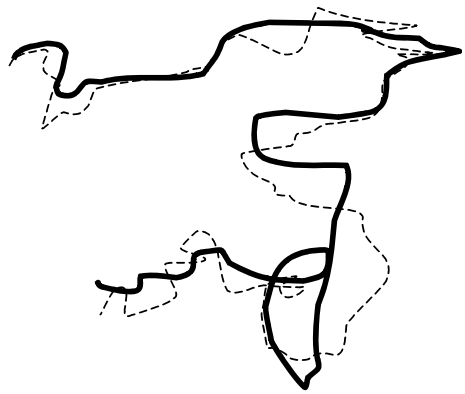

Polymer Physics, Figure 4
A Rouse chain changes its configuration (from solid to dashed curve) locally but not globally in times shorter than τ_R

What does the local motion of this model chain look like? We expect for short intervals that the chain contour may have adjusted locally, but retain a very similar global configuration (see Fig. 4). In order to answer this question, and to compare with local diffusion probes of NSE and NMR on unentangled dynamics, we need to calculate the correlation function $\phi_n(t) \equiv \left\langle (\mathbf{R}(n, t) - \mathbf{R}(n, 0))^2 \right\rangle$ that describes the mean displacement over time of monomers on the chain. Summing over the independent normal modes gives

$$\phi_n(t) = 6D_{CM}t + \frac{Nb^2}{3\pi^2} \left(\frac{t}{\tau_\mathrm{R}}\right)^{1/2} \alpha \,.$$

Here $\alpha = \frac{1}{2} \int_0^\infty z^{-3/2} (1 - \mathrm{e}^{-z}) \, \mathrm{d}z \simeq 1.77$ is a purely numerical constant. The result is remarkable: each monomer executes an "anomalous" or sub-Fickian diffusion, such that its mean square displacement goes as $t^{1/2}$ rather than t (as for ordinary diffusion). This behavior persists until times longer than the Rouse time, after which each monomer is carried by the (faster) center of mass motion of the whole molecule. This anomalous diffusion is simply a consequence of chain connectivity: the further a monomer travels under Brownian motion, the greater is the length of chain that must be correlated with it and the greater the effective drag over that lengthscale.

The (deviatoric) stress formula

$$\sigma_{ij} = \frac{3k_\mathrm{B}T}{b^2} \mathbb{C} \left\langle \frac{\partial R_i}{\partial n} \frac{\partial R_j}{\partial n} \right\rangle$$

we derived above Eq. (13) leads, via representation in terms of the Rouse modes [97], to an expression for the time-dependent modulus function following a step strain $G(t) = \sigma_{xy}(t)/\gamma$. Each mode decays back to equilibrium

anisotropy with its own characteristic time. The power-law distribution of modes then gives:

$$G(t) = \frac{\mathbb{C} k_B T}{N} \sum_p e^{-2p^2 t/\tau_R}$$
$$\approx \mathbb{C} k_B T \left(\frac{t}{\tau_1}\right)^{-1/2} e^{-t/\tau_R} . \quad (20)$$

So we find that, until a final crossover to an exponential decay beyond the Rouse time, the Rouse model has a relaxation modulus which is a power-law of $G(t) \sim t^{-1/2}$. Note that the longest relaxation time scales with molecular weight as N^2, but the viscosity scales as $\mathbb{C} k_B T N$. This is because at the longest relaxation time that sets the value of the viscosity, the stress is carried only by the lowest Rouse mode; the density of these modes is just one per chain, or $\frac{\mathbb{C}}{N}$.

Beautiful generalizations of such dynamical scaling behavior arise in the case of marginally overlapped fractal clusters, such as those arising from gelation transitions. Long-range hydrodynamic interactions are effectively screened, but entanglement effects are negligible. So now the Rouse Eq. (16) generalizes to solving the eigenvalues of the Laplacian (right hand side of (16)) on a fractal cluster with high degree of branching. Locally (for high frequency modes) this is indistinguishable from the case of linear chains, but at longer wavelengths than the chain length between branch points, the effective dimensionality of the cluster rises. Just as in many cases it is possible to represent the mass distribution with an effective dimensional exponent $D = 1/\nu$, so the Laplacian eigenfunctions of a self-similar branched object possess a "spectral dimension" d_s and the density of states in k-space can be written

$$g(k)\mathrm{d}k \cong k^{d_s - 1}\mathrm{d}k . \quad (21)$$

The combination of the scaling dimension of individual clusters, and the self-similarity of the molecular weight distribution via the Fisher exponent yields a generalized power-law for stress relaxation [35]. For Rouse dynamics (local friction) and hyperscaling the result depends only on the fractal dimension of the clusters

$$G(t) \simeq \frac{kT}{b^3} \left(\frac{t}{\tau_0}\right)^{-\frac{3}{D+2}} . \quad (22)$$

The value of the dynamic exponent z in $G(t) \sim t^{-z}$ for three-dimensional percolation is therefore predicted to be $z \cong 0.66$. Experiments on critical gels of unentangled chains confirm this [41].

Entangled Polymer Dynamics

Richer behavior still emerges in the realm of inter-chain *topological* effects that dominate in fluids where chain overlap is very strong. The polymer melts of industrial polymer processing are very highly overlapped on the molecular level, where it becomes immediately apparent that molecular relaxation processes controlling elastic stress are prolonged to very long times indeed. The classic "relaxation modulus" $G(t)$ measuring stress linear-response to a step strain records a "plateau" value before a terminal relaxation time that increases rapidly with molecular weight, in strong contrast to the power-law decays discussed above. Experiments restricted to the timescales of the plateau are hardly able to distinguish between the polymer melt and a rubber, in which the chains are permanently cross-linked to each other at very rare points, sufficiently for each chain to be permanently immobilized from large-scale diffusion. Conceptually, the absent "cross-links" were replaced in the minds of engineers and physicists alike by "entanglements" [13]. These loosely-defined objects were assumed to represent the topological constraint that covalently-bonded molecular chains may not pass through each other. The effective distance between these objects could be calculated, employing rubber elasticity theory,

$$G_N^{(0)} = k_G \frac{RT\rho}{M_e}$$

(with the constant $k_G = 1$ for "affine" and 1/2 for "junction fluctuation" models of elasticity) to deduce the degree of polymerization between entanglements N_e, or the equivalent "entanglement molecular weight", M_e. The number N_e consistently turned out to be of order 10^2, indicating a length-scale for an "entanglement spacing" of 50–100 Å, depending on the particular chemistry. This is highly significant for us, because it shows that small chains on the threshold of feeling topological interactions are real polymers, already long enough to show to a good approximation all the universal properties of statistical connected chains. It also suggests that the role of topology in highly-entangled ($N \gg N_e$) polymer fluids has the potential to be treated universally. Further evidence of universality in entanglements came from experiments in which the polymers were diluted to a volume fraction ϕ_p by a compatible solvent. The apparent entanglement molecular weight $M_e \sim \phi_p^{-\alpha}$ where the scaling exponent $\alpha \simeq 1$ [11].

Other experiments had pointed to the existence of a topological feature at this coarse-grained scale of structure. Careful measurements on rubbers of controlled synthesis had shown that the shear modulus was higher for

a network of long chains than a model incorporating cross-links alone would predict [14]. Other "trapped entanglements" on the same scale as the melt value of N_e seemed to contribute to the elasticity. Advanced theories of rubber elasticity have been able to treat rubber networks in terms of the two distinct constraints of physical cross-links and trapped entanglements [16,42]. A remarkable universality also emerged in measurements of the scaling of melt viscosity η on the molecular weight M of very many different polymer chemistries [13]:

$$\begin{aligned} \eta &\sim M^1 \qquad M < M_c \\ \eta &\sim M^{3.4} \qquad M > M_c \,. \end{aligned} \tag{23}$$

For each material, a critical molecular weight, M_c emerged, above which the viscosity rises very steeply with molecular weight. Furthermore, within experimental error, this explicitly dynamical observation was linked phenomenologically to the essentially static measurements of the plateau modulus by the correlation

$$M_c \simeq 2M_e \,. \tag{24}$$

This connection between essentially dynamic (M_c) and static (M_e) experiments, observed over a wide range of chemistries, is strong evidence that topological interactions dominate both the molecular dynamics and the viscoelasticity at the 10 nm scale in polymer melts (and at correspondingly larger scales for concentrated solutions).

Without going beyond rheological measurements on bulk samples, there has long been other very strong evidence that molecular topology is the dominant physics in melt dynamics. This emerges from the phenomenology of "long chain branched" (LCB) melts. These materials, commonly used in industry, possess identical molecular structure to their linear cousins on the local scale, but contain rare molecular branches (they differ from the "polymeric fractals" discussed above in that their linear sections are long enough to be entangled). The density of branching varies from one branched carbon in every 10 000 to 1 in 1000. This level is chemically all but undetectable, yet the melt rheology is changed out of all recognition if the molecular weight is high enough [43]. Providing that $M \gg M_e$, the limiting low-shear viscosity may be much higher for the same molecular weight. Moreover in strong extensional flows the melt responds with a much higher apparent viscosity than in linear response. This phenomenon, vital for the stable processing properties of branched melts, is called "extension hardening". The effect is all the more remarkable because in shear flows, branched, as well as linear, melts exhibit a lower stress than would be predicted by a continuation of their linear response [44] (they are "shear-thinning"). A fascinating example of the difference between linear and branched entangled melts is well-known from flow-visualization experiments. The velocity field in a strong "contraction flow" of a linear polymer melt resembles that of a Newtonian fluid, while that of a branched polymer sets up large vortices situated in the corners of the flow field. Slight changes to the topology of the molecules themselves give rise to qualitatively different features in the macroscopic fluid response.

The most successful accounts of these phenomena have been given by the *tube model*. The idea is to deploy the theoretical physicists favorite strategy of replacing a difficult many-body problem with a tractable single-body problem in an effective field. In this case the "single body" is the single polymer chain, and the effective field becomes a tubelike region of constraint along the contour of the chain. The tube is invoked to represent the sum of all topological non-crossing constraints active with neighboring chains, and the tube radius, a, is of the order of the end-to-end length of a chain of molecular weight M_e. In this way, only chains of higher molecular weight than M_e are strongly affected by the topological constraints (see Fig. 5).

The tube was first invoked by Edwards [45] in an early model for the trapped entanglements in a rubber network. The consequences of the idea for dynamics were first explored by de Gennes [46], again in the context of networks. A free chain in a network would be trapped by neighbor-

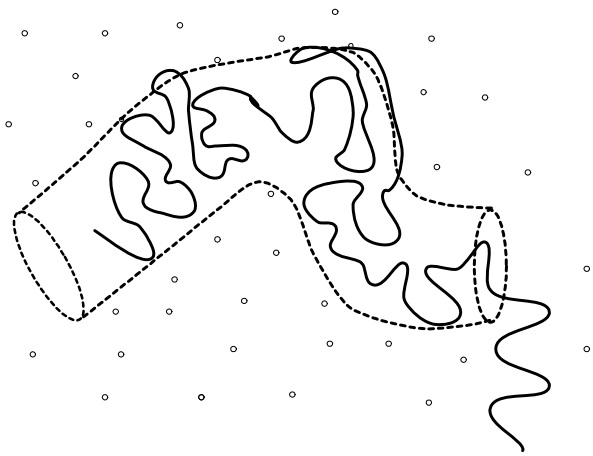

Polymer Physics, Figure 5
A tubelike region of constraint arises around any selected polymer chain in a melt due to the topological constraints of other chains (*small circles*) in its neighborhood (diagram courtesy of R. Blackwell)

ing chains into tube of radius a defined by its own contour, suppressing motion perpendicular to the tube's local axis beyond a distance of a, but permitting both local curvilinear chain motions and center-of-mass diffusion along the tube. de Gennes coined the term "reptation" for this snake-like wriggling of the chain under Brownian motion. The theory gives immediately a characteristic timescale for disengagement from the tube by curvilinear center-of-mass diffusion. This disengagement time τ_d is naturally proportional to the cube of the molecular weight of the trapped chain (this arises from combining the Fickian law of diffusive displacement of length L with time τ, $\tau \sim L^2$, recognizing that path length $L \sim M$, with one extra power arising from the proportionality of the total drag on molecular weight). Very significantly, de Gennes also realized that a tubelike confining field would endow a dangling arm, fixed to the network at one end, or belonging to a star-shaped polymer in a network, with exponentially slow relaxations. In this topology reptation would be suppressed by the immobile branch point [47], and only exponentially-rare retractions of the dangling arm would disengage it from its original tube (see Fig. 6 below). In the late 1970s, S.F. Edwards and M. Doi developed the tube concept into a theory of entangled melt dynamics and rheology for monodisperse, linear chains [97], finding extensions to flow instabilities [48], blends [49] and polymers of controlled architecture that go beyond the star topology [50] to H-polymers [51] and combs [52].

Fully atomistic simulations of polymer melts are now able to examine entanglement effects. It is now possible to conduct molecular dynamics simulations of, for example, elastically-connected Lennard–Jones polymers that contain 50 chains each of 10 000 monomers well into the regime in which entanglements dominate the dynamics [53]. This technology is now at the point at which direct comparisons to experimental results such as NSE is now possible. The other advantage of large simulations is that they may mimic the "ideal" experiment in which everything may in principle be measured. This has been exploited in tests of fundamental theories of entanglements (see below) [54].

The growing quantity of data on branched molecules of controlled molecular weight and topology has provided severe tests of the tube concept at a level beyond that probed by linear chains [40]. The hierarchical nature of configurational relaxation at the molecular level in particular has been turned from speculation into orthodoxy. In the simplest case of entangled star polymers, the theory suggests that chains escape from their confining tubes not by reptation, which is suppressed by virtue of the immobile branch point, but by a process of *arm retraction*, present but largely eclipsed in the case of linear polymers (see Fig. 6).

The effect on the viscosity of replacing linear molecules with those of identical molecular weight, but of star topology, is striking: now

$$\eta \sim e^{\alpha (M_a/M_e)} \quad M_a > M_c \tag{25}$$

is the dominant form of the molecular weight dependence (where M_a is the molecular weight of the dangling *arm*), rather than $\eta \sim M^{3.4}$ (in the case of linear polymers the entire effect of these fluctuations is to change the apparent exponent of this relation from 3 to 3.4 up to a high molecular weight of order 10^3 entanglements, where it sat-

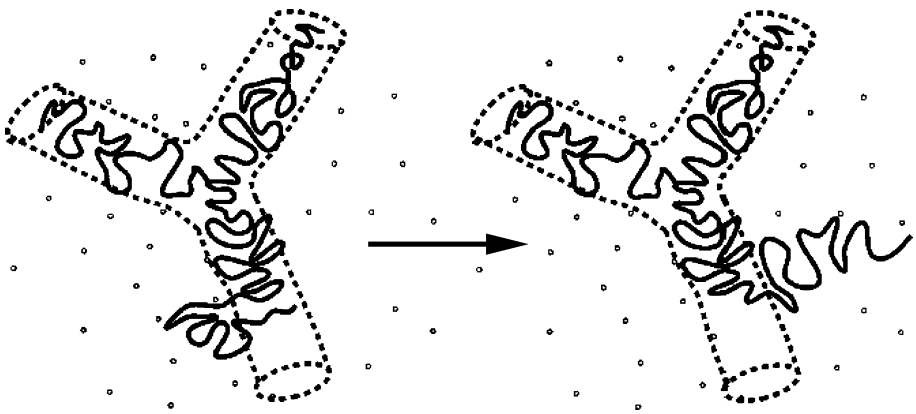

Polymer Physics, Figure 6
The process of arm retraction predicted by the tube model for the case of dangling entangled arms, as from the branch point of a star polymer. Unlike in reptation, reconfiguration of the outer parts of the arm occurs many times for one relaxation of deeper segments

urates at the "bare reptation" value of 3). In entangled melts with repeated levels of branching, the retraction dynamics of outer levels generate the slower retraction of inner branches in a hierarchical way. The most advanced application of this process has been on materials that combine the complexity of the gelation-point critical ensemble (see Subsect."Single Chain Dynamics" above) with the physics of entanglement. The broad rheological spectrum of branching chemistry that builds molecules by grafting back repeatedly previously synthesized molecules [55] is captured by these calculations. When the percolation transition is approached by cross-linking entangled polymers, there exists a measurable region in which mean-field statistics ($D = 4$) actually hold. Stress relaxation from such an ensemble is logarithmic in this regime, but is closely modeled by a dynamical scaling form in which the dynamic exponent z is small (being inversely proportional to the number of entanglements between branch points) [56].

The wealth of experimental and simulation data has sharpened the theoretical picture. Without exploding with new parameters, it has been possible to capture, in a single model, modes of entangled motion beyond pure reptation. In linear response *contour length fluctuation* (CLF), the Brownian fluctuation of the length of the entanglement path through the melt, modifies early-time relaxation. Similarly the process of *constraint release* (CR), by which the reptation of surrounding chains endows the tube constraints on a probe chain with finite lifetimes, contributes to the conformational relaxation of chains at longer times. Both the processes of CLF and CR contribute to the quantitative understanding of linear rheology, such that the $\eta \sim M^{3.4}$ law is no longer a mystery [57,58], but much of the newer data still need to be examined quantitatively as sensitive tests of the detailed physics, and many puzzles remain. These two additional processes are visualized in Fig. 7.

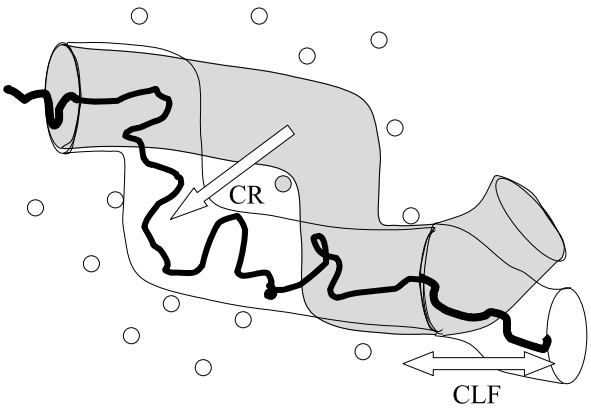

Polymer Physics, Figure 7
A cartoon of the processes of contour length fluctuation (CLF) and constraint release (CR) on a linear polymer in a constraining tube. In CLF the chain end retracts via longitudinal fluctuations of the entangled chain, but without requiring center-of-mass (reptation) motion. Re-extension of the chain end may explore new topological constraints, reconfiguring the tube. In CR, an entanglement with a neighboring chains (shown hatched) may disappear, allowing effective conformational relaxation of that part of the tube, again without reptation of the test chain itself. In both cases the former tube configuration is shown *dark*, the new, *light*

Non-linear Flow of Polymeric Fluids

In strong deformations the additional processes of *chain stretch, chain retraction*, and *branch-point withdrawal* emerge on the level of single chains (the latter exclusively in the branched case), and *convective constraint-release* (CCR) at the level of co-operative motion [59,60,61]. The most advanced formulation of the tube model for linear polymers keeps the coarse grained coordinates of the chain, and allowing CCR events to generate local Rouse jumps of the tube [61]. The idea is to retain full information about average chain trajectories instead of working indirectly with dynamic equations for the stress and orientation tensor. This approach also allows quantitative predictions about the single chain scattering function $S(\boldsymbol{q})$, and to develop a *local* description of CCR events. The main assumptions of the first version of the theory (valid when there is no chain stretch) are: (i) that CCR operates locally in reorienting chain segments both into and away from the flow direction, and (ii) that neither the number of entanglements per chain $Z = M/M_e$ nor the tube diameter a changes. The first assumption endows the tube itself with a Rouse-like motion in which the local hopping rate is coupled to the global deformation rate via a single new parameter. The second (constant length) assumption introduces a difference from ordinary Rouse-chain motion, and limits the range of validity at first (but see below) to $0 < \dot{\gamma} < 1/(\tau_e Z^2)$.

No single set of variables will be able to diagonalize the essential entangled modes of motion, namely (i) chain reptation, (ii) chain retraction, (iii) tube-length fluctuation and the new mode (iv) Rouse-tube motion. However, the theory is conventionally cast in a real-space notation for the tube trajectory $\mathbf{R}(s, t)$ and its tangent curve $\mathbf{R}' \equiv \frac{\partial \mathbf{R}}{\partial s}$, functions of curvilinear distance s from along the tube and time t. Our chains are monodisperse containing Z entanglements of tube diameter a. The (stochastic) equation of

motion becomes [62]

$$\mathbf{R}(s, t+\Delta t) = \kappa\, \mathbf{R}\Delta t + \mathbf{R}(s+\Delta\xi(t), t) + \Delta t\left(\frac{3\nu}{2}\frac{\partial^2\mathbf{R}}{\partial s^2} + \mathbf{g}(s,t)\right) + \Delta t\lambda\left(\frac{Z}{2} - s\right)\frac{\partial\mathbf{R}}{\partial s} + 3D_c Z\frac{(\mathbf{R}''\,\mathbf{R}')}{|\mathbf{R}'|^2}\mathbf{R}' . \tag{26}$$

The terms of this formulation describe, in order, affine convection of the tube, reptation, CR Rouse motion of the tube, retraction of the chain within the tube and chain stretch. This model predicts in shear flow a near-plateau of $\sigma_{xy}(\dot\gamma)$ between $\dot\gamma\tau_d$ and $\dot\gamma\tau_R$, and an increase proportional to $\dot\gamma^{1/2}$ beyond that (so that the "shear-dependent viscosity" $\eta(\dot\gamma) \sim \dot\gamma^{-1/2}$ (see Fig. 8). Both this feature, and a prediction of strong extension hardening at rates faster than the inverse chain stretch time, have been quantitatively matched to experiment.

The strongest tests of the predictions for chain conformations have been performed in "neutron flow mapping" experiments [64,65]. Here a partially-deuterated (for neutron scattering contrast) polymer melt of monodisperse or controlled architecture, is passed continuously through a complex flow field such as a contraction, bounded by windows transparent to both laser illumination and thermal neutrons. The stress field is measured in optical birefringence, while single chain conformations are reported by the neutron small angle scattering. Scanning the apparatus across the neutron beam allows the experiment to probe regions of the flow with varying strain histories. The experiments are compared to calculations in which the model of (26) is used both to calculated the flow field itself, then the scattering function of chains subjected to the stream lines that flow through the neutron beams.

Figure 9 shows the results of one such experiment. The key result is that relaxation to equilibrium structure takes place in the flow at timescales that depend on the chain lengthscale examined. It also shows that no model containing just one viscoelastic relaxation time is even qualitatively able to account for the non-linear physics of entangled melts. The simplest must possess at least two relaxation times, corresponding to chain stretch (fast – by Rouse motion) and chain orientation (slow – by reptation). At the purely phenomenological level of the stress tensor the emergent property is a rapidly-relaxing trace, and a slower traceless part to the stress.

At a more approximate level, it is possible to combine the complexities of non-linear response and complex branched topologies in entangled melts. All the linear and non-linear molecular processes of chains in tube-like entanglement fields are present, together with one additional process, that of "branch point withdrawal". When a strand

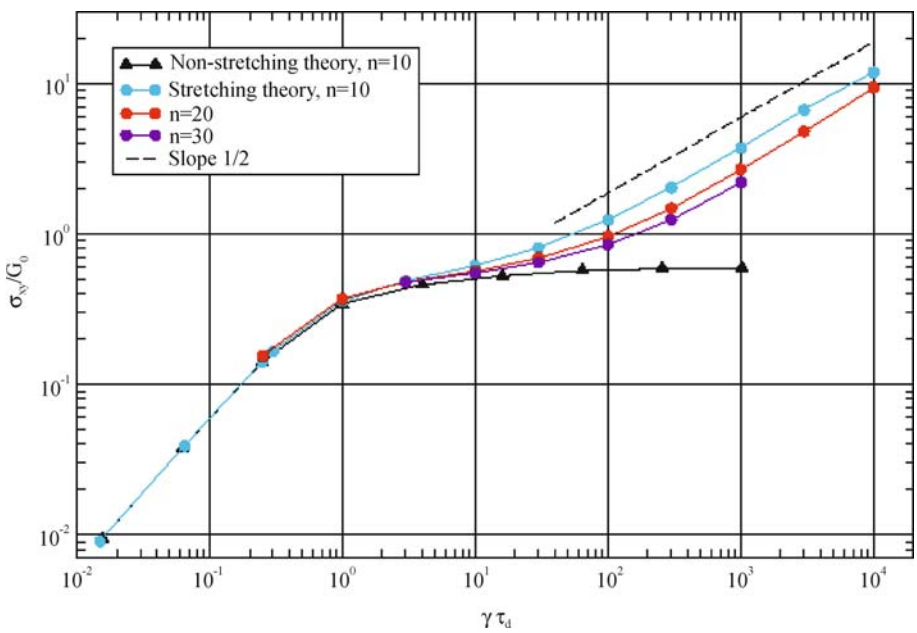

Polymer Physics, Figure 8
Predictions of the local CCR model with chain stretch using $c_v = 0.1$ for values of $Z = 10, 20, 30$. A comparison to the non-stretching version is given

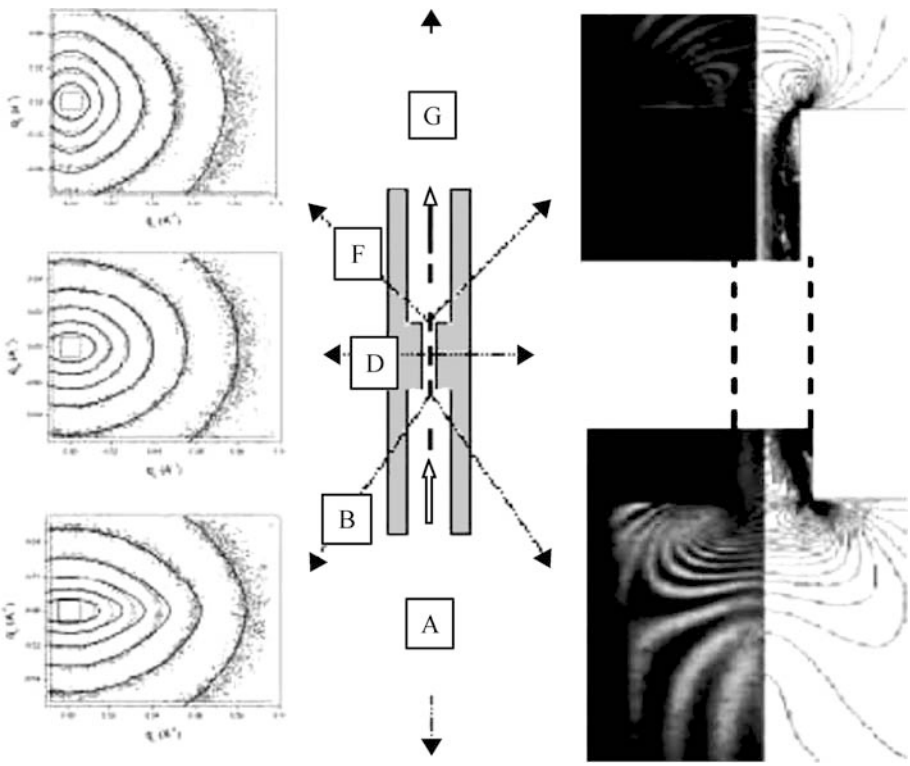

Polymer Physics, Figure 9
A monodisperse linear entangled melt in a neutron flow mapping experiment. Flow is from bottom to top. Single chain scattering functions from windows B, D and F are compared with contour predictions on the *left*. Chain configurations relax during passage through the channel from an initial strain. On the *right*, contours of principal stress (*left*) are compared with calculations (*right*)

connecting two branch points is very highly stretched, the finite size of the clusters it connects (in contrast to the network case) becomes apparent, and one of the clusters may topologically collapse into the deforming tube of the central strand. In the simplest case of the H-polymer the two outer arms are drawn into the tube of the "cross-bar" segment (see Fig. 10).

There are consequences of the process for both chain conformation (scattering) and stress response (rheology). If the retracting arms are labeled in a scattering experiment, very strong signal enhancement is seen in the direction of the deformation [51,66]. At the same time, the growing extensional stress, until that point resembling that of a cross-linked network, reaches a near-plateau. The effect of this at the level of the process engineering is to endow strongly branched melts with both extension hardening (leading to flow stabilization) and good processability. A useful modeling tool for branched melts has been derived from the ideal "pom-pom" architecture [67]. This generalization of the H-polymer allows the number of arms attached at either end of the cross-bar, q, to vary. So q becomes a molecular parameter controlling

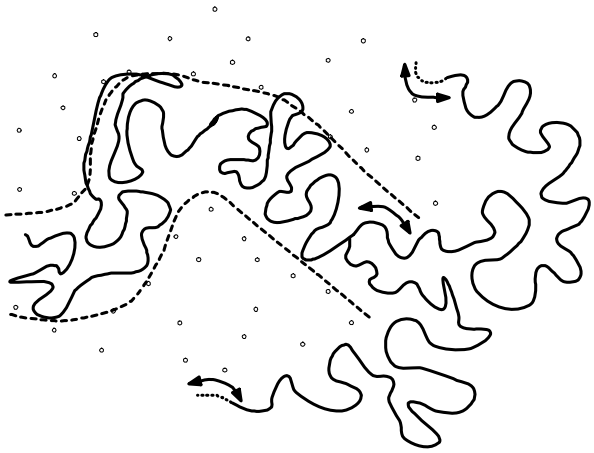

Polymer Physics, Figure 10
The process of branch point withdrawal: a segment with greater than equilibrium tension pulls attached dangling arms some distance into its own tube, thus shortening their effective entangled path length

the degree of strain-hardening. Creating multiple modes from this model allows accurate models to be built for commercial, highly-branched melts [68,69,70]. Computing in complex geometries with these models has successfully predicted features of the flows particular to branched melts, such as recirculating vortices upstream of a contraction and very persistent sheets of high birefringence downstream from re-entrant corners, sometimes known as "stress-fangs" [71].

Multi-Phase Polymeric Fluids

In the foregoing, the spatial structure of polymeric fluids has remained essentially homogeneous, all heterogeneities remaining at the level of entanglements, or correlation volumes, in solution. The hidden reason for this is that we have considered systems containing just one chemistry of monomer. Yet one of the great attractions of polymers is their tendency to develop complex spatial structures when different chemistries are combined, either in the case of entire chains, when we create *polymer blends*, or within single chains, referred to as *co-polymers*. In the latter case, especially interesting structures arise when the distinct monomers are correlated in sequence along the chain, forming *block co-polymers*. The simplest example would be a string of n monomers of chemistry A followed by m of chemistry B. This is an A–B diblock co-polymer. The essential reason for the sensitivity to local chemistry is that the entropy of spatial translation per monomer in a polymeric fluid is extremely low. The usual Van't-Hoff term $S_{\text{trans}} = -kT \log \phi$ is divided by the degree of polymerization of the chain so that it effectively competes with the mild Van der Waals dominated enthalpic interaction between monomer units, that tend to favor proximity of identical chemical units: polymers tend therefore towards demixing. In blends the demixing competes with the slow, viscoelastic dynamics we discussed in the last section, creating spatially complex morphologies that are determined kinetically. In the case of block co-polymers, the demixing occurs locally, confined to spatial scales of the block subchain radii. The demixing now competes with the chain elasticity we discussed in Sect. "Single Polymer Chain Physics". The connection of polymer sequence with emergent structure illustrates the high potential information content of a macromolecule. It is an example of the genotype-phenotype pattern developed to a much higher degree in the case of the DNA-embedded genetic code. This field, like that of controlled architecture dynamics, is an area of polymer physics where the complex emergent phenomena have required a parallel implementation of careful synthetic chemistry [72], experimental physics [73] and advanced theoretical techniques [75] to explore, and is growing extremely rapidly.

Polymer Blends

The starting point for a conceptual understanding of the physics of polymer blends is a mean-field model for the free energy of mixing of two species such that one occupies a total volume fraction ϕ. Known as the Flory–Huggins free-energy [17], it balances the entropy of mixing of the two components against the energy difference of mixed and demixed states:

$$\Delta F_{\text{mix}} = k_{\text{B}} T \left[\frac{\phi}{N_{\text{A}}} \ln \phi + \frac{1-\phi}{N_{\text{B}}} \ln (1-\phi) + \chi \phi (1-\phi) \right]. \tag{27}$$

The control parameters for this theory are the two molecular weights and the (temperature dependent) interaction ("Flory") parameter χ. The whole system is therefore three-dimensional: the monomer interaction is best normalized with the mean molecular weight so that $\chi \vec{N}$ is the essential parameter that controls interaction strength. Then $N_{\text{A}}/N_{\text{B}}$ becomes the asymmetry parameter while ϕ controls the composition. The possibility of phase separation means that ΔF_{mix} can be minimized by forming regions with different values of ϕ if the curvature $\partial^2 \Delta F_{\text{mix}}/\partial \phi^2$ is anywhere less than zero. This in turn occurs for all $\chi > \chi_{\text{c}}$. Phase separation then occurs for all systems whose mean composition falls between the two minima of $\Delta F_{\text{mix}}(\phi)$, a region that broadens as χ moves further (as temperature is changed) from χ_{c}.

This behavior is mapped in Fig. 11. The region between the two curves (binodal and spinodal) of the plot corresponds to compositions and temperatures where the curvature of $\Delta F_{\text{mix}}(\phi)$ is positive, producing a fluid that is *locally* stable to composition fluctuations though globally unstable to phase separation. In this case droplets of the separated phase have to nucleate, giving an initially disconnected morphology. Within the spinodal curve on the other hand, the fluid is unstable to local and infinitesimal changes in composition, so that natural thermal fluctuations of composition are amplified. The presence of a fastest-growing wavelength leads to a connected (or "spinodal") morphology for most of this region. Since the final minimum in free-energy is total phase separation, whatever intermediate structure evolves eventually coarsens with time with well-known growth laws depending on whether hydrodynamics or diffusion is dominating [76]. This beguilingly simple map hides a great deal of latent complexity, however. More recently, the recognition that the spinodal process leads to a morphology dependent

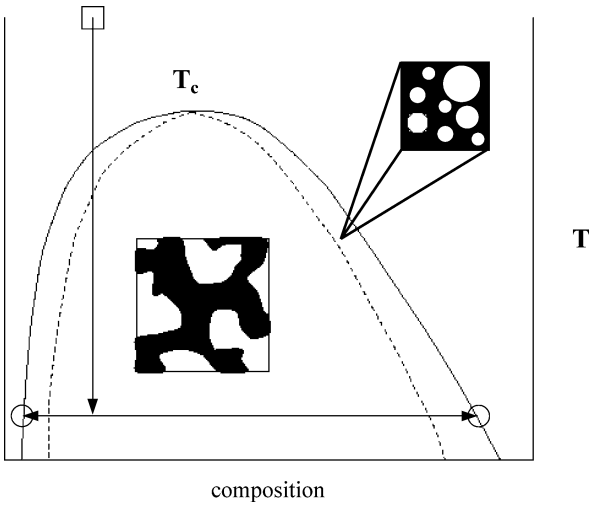

Polymer Physics, Figure 11
Regions of phase-separation in polymer blends with the Flory–Huggins model. Phase separation occurs within the binodal curve (*solid curve*), but is unstable locally only within the spinodal (*dashed curve*)

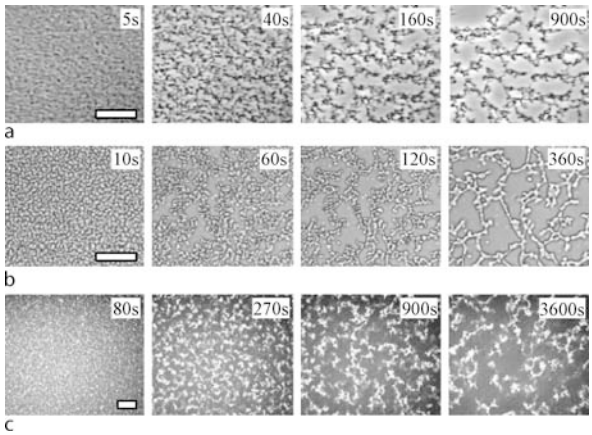

Polymer Physics, Figure 12
Confocal imaging of viscoelastic phase separation from [78] showing the time development of structure from a dilute polymer solution, b protein solution, c concentrated polymer solution

on both the local free energy and the distribution of initial fluctuations has generated experimental and theoretical explorations of the wider control space that includes trajectories across the composition diagram [77], variations on boundary conditions and viscoelastic effects [78]. A striking example is given by the "target" structures that result from kinetic trajectories that spend some time in the nucleation region of the composition space before (cooling) into the spinodal region. During the first period, small nuclei are allowed to form but sustain limited growth. Instead, the growth occurs in the locally unstable region using the nuclei as sources of unstable fluctuations. Because the Fourier wavelet decomposition of a spherical nucleus with a sharp boundary consists of spherical waves centered on the nucleus, it is these that are amplified by the growth process, or reverse-diffusion, and structures of nested spheres of varying composition appear during it. Since phase separating polymers typically also have a glass transition temperature, it is possible to quench the composition structure at any point in the phase separation.

Viscoelasticity may play a modifying role at both early and late stages of the phase separation. At early times, any dominant viscoelastic mode, such as reptation, can undergo mode-mixing with the dominant phase separation time to produce to two-timescale process of separation [79]. This may result in non-monotonic concentration growth with time, since of the two resulting modes one typically decays while the other grows. At later stages, a strong viscoelasticity in at least one phase causes the heterogeneous fluid to retain the mutual connectivity of both phases for much longer than in Newtonian fluids [78].

Figure 12 illustrates the high degree of universality of this effect, contrasting dilute and concentrated polymer solutions with a protein solution in which the protein molecules behave more as colloidal particles than polymer chains. Although the structures coarsen, they retain connectivity of the minority phase rather than suffer it breaking into droplets. A similar effect is generated by the natural composition-dependence of mobility. It is unlikely that the mobilities of the two demixing species will remain independent of the local composition, since the natural drag constants within the two final demixed phases will typically differ. The phase of lower mobility then becomes kinetically quenched earlier than that of higher mobility. This nonlinearity affects the connectivity of the morphology at long times [80].

Block Co-Polymers

The demixing tendencies of polymers of different chemistries becomes very rich when it competes with the feature at the heart of polymer physics itself: that of connectivity. By connecting a chain of A-monomers to one of B-monomers they are forced by the spatial correlations that chain connectivity induces into proximity. Within a mean field picture, the repulsive interaction in the free energy density $k_B T \chi \phi_A \phi_B$ (from the last term in (27)) competes with the entropic chain elasticity $F_{elas} = k_B T \left(R_A^2/N_A b^2 + R_B^2/N_B b^2\right)$. In contrast to previous sections, we consider first the case of dense chains

(melts), then solutions and finally the effects within single chains of such controlled chemical architecture.

Consider first the case of the *diblock* co-polymer just outlined. In a melt, when the interaction parameter χ is large enough, the chains must locally segregate into regions rich in one or other of the polymers. The molecules themselves clearly minimize their contribution by situating their midpoints where the two chemistries join at the interface between the two regions. Topology dictates that at most two A-chains may span an A-rich region, so consequently the length scale of the morphology must be of the same order as the radius of gyration of the chains. In the symmetric case the system spontaneously forms *lamellae* of alternating composition whose width increases as the interaction parameter does though with a weak power law of $\chi^{1/6}$. The width of the interface region in which both A and B monomers are present decreases with χ. In a fully quantitative "strong segregation theory" of this physics [81] the lamellum width L and interface width w obey

$$L = 2\left(\frac{8}{3\pi^4}\right)^{1/6} N^{2/3} b\chi^{1/6}\,; \quad w = \frac{2b}{(6\chi)^{1/2}}\,. \tag{28}$$

But additional physics comes into play as soon as the two blocks are of different length, for now the entropy of confinement of the chains in the plane of the interface does not balance (we did not consider that in the above), with the result that the interface tends to curve away from the domains containing the longer chains. Other more complex morphologies become candidates for the minimum in free energy: periodic cylinders (C) and spheres (S) as well as the more exotic forms of the gyroid (G), perforated lamellae (PL) and double-diamond lattices (D) of Fig. 13.

Experiments on carefully synthesized block co-polymers, especially the model polystyrene-polyisoprene system, have mapped out the morphology diagram in terms of interaction (via temperature, as in blends) and composition of the diblock. Calculations were first performed near the critical point, where segregation of the two species is only weak, and the free energy can be expanded in powers of the difference of the local mean concentrations $\phi = (\phi_A - \phi_B)$ [83], but can be extended into strong segregation by a fully continuous self-consistent mean field theory [82] that ignores only the fluctuations in composition. Results and comparison with experiment are shown in Fig. 14.

Clearly the calculations are qualitatively, and in some aspects quantitatively in agreement, especially far from the disordered (fully mixed) state. However, fluctuations clearly have an important effect near the boundary. The

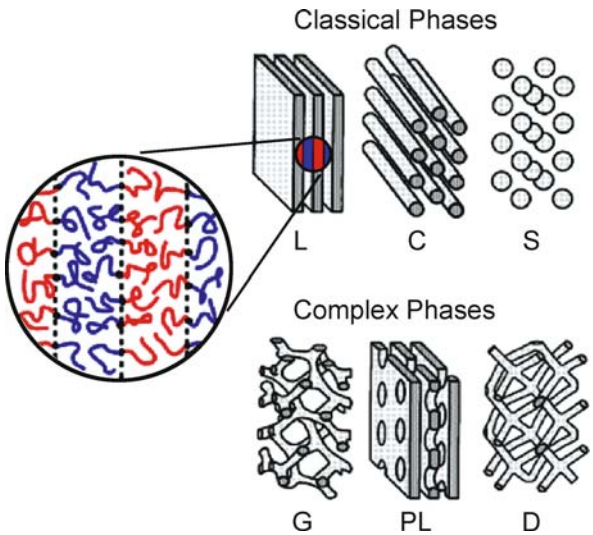

Polymer Physics, Figure 13
Potential candidate morphologies for block co-polymer melts, from [82]

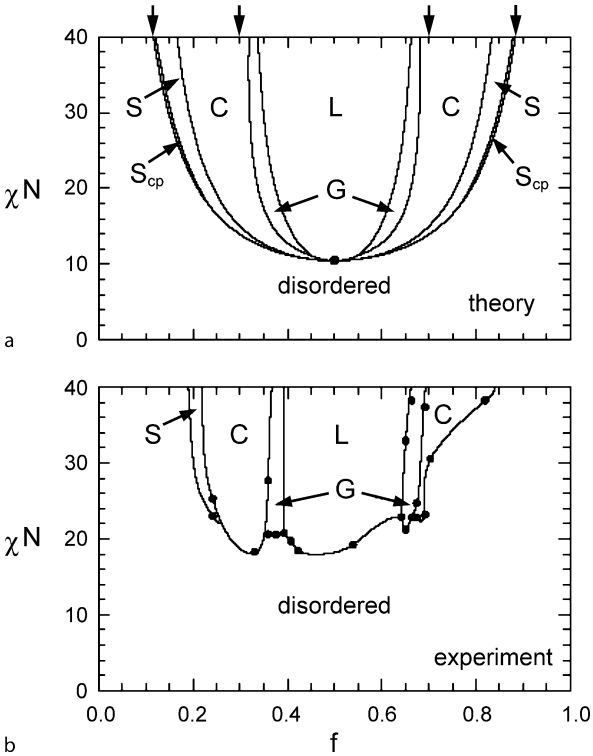

Polymer Physics, Figure 14
The morphology map for diblocks as a function of interaction parameter and composition by a self-consistent field calculation and b experiment on PS-PI diblocks. From [82]

topology of the diagram at the critical point is actually changed by them, stabilizing a finite region of the lamellar phase. Quite recently, field-theoretical methods for computing the full effect of fluctuations in the partition function of block co-polymers have been introduced [84]. Other natural extensions include increasing the complexity of the chemical structure of the block co-polymer. Taking only one step to the tri-block permits a very wide class of morphologies that mix those of the diblock, so that for example spheres of *A* may decorate cylinders of *B* within a matrix of *C*. Such materials are by no means only of academic interest, as they constitute the fundamental technology behind toughened plastics, in which the minority, rubbery phase prevents crack propagation within a majority glassy phase [85]. We can begin to see the sequence of the chain corresponding to an information content ("genotype") that implicitly codes for the morphology ("phenotype") of the emergent morphology.

Block copolymers in solution, either of a low molecular weight species or of a simple homopolymer, may act as polymeric versions of the familiar surfactants. Their self-assembly picks out the same structures as we have already seen in the context of bulk microphase separation (lamellae, cylinders and spheres) with now the difference that these self-assembled structures exist in isolation rather than in a regular array. The corresponding structures are vesicles, wormlike micelles and spherical micelles [86].

Examples are given in Fig. 15. These phases have a visibly complex form, since they are largely determined kinetically: because of the long timescales for diffusion and collision of the individual micelles, and of the high activation energy for mutual rearrangement, merger and breaking, true equilibrium is very hard to reach in these systems. So vesicles of widely different radius may coexist, together with nested structures and complexes. The transitions between linear and spherical structures, which may be driven by changes in temperature or pH depending on the chemistry of the polymers, may exhibit a rich kinetics in which cylindrical structures emerge from spherical and vice versa.

We finally consider the generalization of the coil-collapse transition we saw in the case of dilute chains in poor solvents in Sect. "Single Polymer Chain Physics". When all monomers are identical in their interaction with the solvent the form of the resultant globule will be on average spherical, together with natural thermal fluctuations of the interface. When the polymer contains blocks of heterogeneous interaction with solvent, as well as self-interaction, the globule may become a much more complex object. Experiments on diblocks indicate that a range of non-spherical geometries may be generated by controlling the architecture of the primary chain [87]. Calculations balancing chain stretch, mutual interaction and surface tension against the solvent [88] indicate that it is theoretically possible, using only two monomers, to code for transitions between near-spherical to prolate and finger-like forms of the collapsed globule. More complex features such as budded and pearled structures also emerge from the same mean-field level of theory that successfully treats block co-polymer melts. This single-molecule form of the coding of emergent phenotype from the information coded as a polymer sequence is very suggestive of nature's own method of constructing the functional single molecules of enzymes, motors and cellular structures. For proteins are "just" co-polymers (of a possible 20 amino acid monomer set) that code in their sequence for absolutely specific forms of their collapsed state. The protein-folding problem has generated a huge literature [89], although rather little of it actually exploits polymer physics, with its natural high order of dimensionality and degrees of freedom [90] to understand this ultimate refinement of the art of coded morphology.

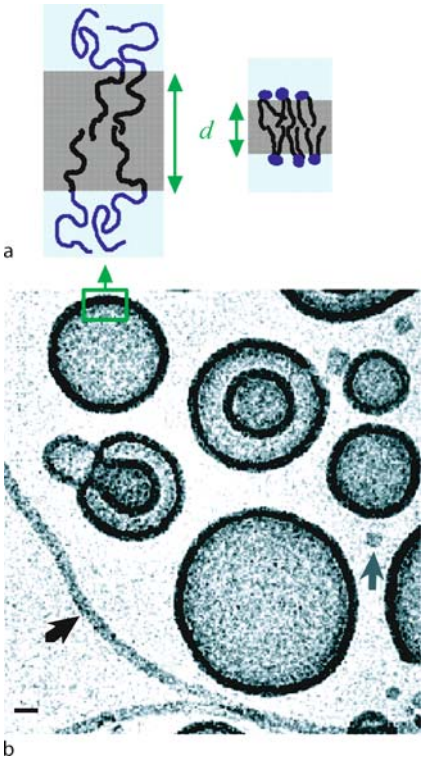

Polymer Physics, Figure 15
Schematic of chain configuration in the wall of a block co-polymer vesicle and its architectural control (a) and transmission electron micrograph of polymeric vesicles (b), from [86]

Future Directions

Several current themes suggest a rich future for polymer physics, notwithstanding the experience of history that the richest veins of research to come will be those currently unforeseen. The general area of biomimetic, or perhaps better, bio-suggestive polymer physics is bound to be an area of growth. Biology has already mastered the art of synthesizing single polymers that act as self-assembling machines, chemical reactors and separation systems. Artificial polymers that act in these ways are certainly possible, and may not need to be as exactly structured as proteins are in order to deliver function. We have already seen that theoretical schemes exist for designing block co-polymers that will collapse into pre-determined shapes and forms [88]. Future designs of active versions of protein-like polymers may also go beyond the chemical "fuel" of ATP dephosphorylation, perhaps using light as both an energy and signaling source. Optically-activated mechanical transitions in polymers have already been demonstrated [91]. Effectively equally fast response can be elicited from pressure changes. In the long term one might hope for advanced therapeutics from this route, especially in combination with polymer-based encapsulation systems such as triggered micelles of block co-polymers.

Structural materials properties also have many things to learn from evolution. As it becomes possible to moderate and control microstructure, both in terms of crystallinity and microseparation, so polymeric nano-composites will be able to realize combinations of strength and stiffness currently existing as ideals [74].

A related direction brings polymer physics into biological research directly. Dynamic neutron scattering by neutron spin echo [92], as well as advanced NMR relaxation techniques will assist the current move towards exploring the role of dynamics in molecular biological function on a similar footing to the achievements in the area of structure. Even thermal dynamics is beginning to be recognized as a generator of functions such as signaling, in an analogous way to the emergence of rubber elasticity from the same source [93]. Technologically, the use of artificial polymers to create scaffolds for tissue engineering will require a balance of local biochemical interaction and global mechanical and topological structure formation.

The key underpinning science of biocompatible and biomimetic polymers is the control of self-assembly. Already there are a number of "suprachemical" options of non-covalent, reversible polymerization [94]. Playing with nature's alphabet of peptide-forming amino acids, that naturally self-assemble via main-chain hydrogen bonds, gives a rich system in which chirality becomes a new control parameter for the equilibrium structure [95]. Combining main-chain self-assembly with side-chain functionality may prove to open up new classes of functional polymeric materials.

The growth of conducting and semi-conducting polymeric materials within a burgeoning new sector of the electronics industry is already well under way. But this area has been driven largely be technology, and much of fundamental science remains to be understood, especially at the level of many chain, materials physics [96]. The changing demands of the world's energy economy are bound to put pressure on developments of organic, polymer-based photovoltaic materials, as well as lightweight polymer gel energy storage. The combination of information-processing and advanced structural properties within new polymeric materials is a tempting prospect.

Bibliography

Primary Literature

1. Kuhn W (1936) Koll Z 76:258
2. Guth E, Mark H (1934) Monatshefte Chemie 65:93
3. Flory PJ (1953) Principles of Polymer Chemistry. Cornell University Press, Ithaca
4. Zimm BH, Stockmayer WH (1949) J Chem Phys 17:1301
5. Edwards SF (1976) The configuration and dynamics of polymer chains. In: Balian R, Weill G (eds) Molecular Fluids. Gordon and Breach, London, pp 151–208
6. Zimm BH (1956) J Chem Phys 24:269
7. Rouse PE (1953) J Chem Phys 21:1272
8. Zinn-Justin J (1993) Field Theory and Critical Phenomena. Clarendon Press, Oxford
9. De Gennes PG (1986) Scaling Concepts in Polymer Physics. Cornell University Press, Ithaca
10. Edwards SF (1966) Proc Phys Soc Lond 88:265
11. Adam M, Delsanti M (1984) J Phys France 45:1513
12. Verdier PH, Stockmayer WH (1962) J Chem Phys 36:227
13. Ferry JD (1986) Viscoelastic Properties of Polymers. Wiley, New York
14. Treloar LRG (1975) The Physics of Rubber Elasticity. Clarendon, Oxford
15. James H, Guth E (1947) J Chem Phys 15:669
16. Ball RC, Doi M, Edwards SF, Warner M (1981) Polymer 1010:22
17. Koningsvelt R, Stockmayer WH, Nies E (2001) Polymer Phase Diagrams. Oxford University Press, Oxford
18. Feynman R (1965) Hibbs Quantum Mechanics and Path Integrals. McGraw Hill, Kogakusha
19. Perkins TT, Smith DE, Larson RG, Chu S (1995) Stretching of a single tethered polymer in a uniform flow. Science 268:83
20. Marko JF, Siggia ED (1995) Stretching DNA. Macromolecules 28:8759
21. Edwards SF (1965) Proc Phys Soc 85:613
22. Mazur J, McCrakin FL (1968) J Chem Phys 49:648
23. de Gennes PG (1972) Phys Lett 38A:339–341
24. des Cloiseaux J (1981) J Phys 42:635
25. des Cloiseaux J, Jannink G (1990) Polymers in Solution. Oxford University Press, Oxford

26. Rubinstein M, Colby RH (2003) Polymer Physics. Oxford University Press, Oxford
27. Dobrynin AV, Rubinstein M (2005) Theory of polyelectrolytes in solutions and at surfaces. Prog Polym Sci 30(11):1049–1118
28. Manning GS (1969) Limiting Laws and Counterion Condensation in Polyelectrolyte Solutions I. Colligative Properties. J Chem Phys 51(3):924–933
29. Ha BY, Liu Andrea J (1998) Effect of Non-Pairwise-Additive Interactions on Bundles of Rodlike Polyelectrolytes. Phys Rev Lett 81:1011
30. Daoud M et al (1975) Macromolecules 8:804
31. Noda I et al (1981) Macromolecules 14:668
32. des Cloiseaux J (1975) J Phys 36:281
33. Deam RT, Edwards SF (1976) Phil Trans Roy Soc A 280:317–353
34. Castillo HE, Goldbart P (2000) Phys Rev E 62:8159
35. Cates ME (1985) Excluded volume and hyperscaling in polymeric systems. J Phys Lett 46:L837–L843
36. Jian T, Vlassopoulos D, Fytas G, Pakula T, Brown W (1996) Colloid Polym Sci 274:1033
37. Higgins JS, Benoit H (1994) Polymers and Neutron Scattering. Clarendon Press, Oxford
38. Klein PG, Adams CH, Brereton MG, Ries ME, Nicholson TM, Hutchings LR, Richards RW (1998) Macromolecules 31:8871
39. Watanabe H (1999) Prog Polym Sci 24:1253
40. McLeish TCB (2002) Tube Theory of Entangled Polymer Dynamics Adv Phys 51:1379–1527
41. Lusignan CP et al (1995) Phys Rev E 52:6271
42. Rubinstein M, Panyukov S (2002) Macromoelcules 35:6670
43. Small PA (1975) Adv Polym Sci 18:1
44. Meissner J (1975) Pure Appl Chem 42:551
45. Edwards SF (1967) Proc Roy Soc London 92:9
46. de Gennes PG (1971) J Chem Phys 55:572
47. de Gennes PG (1975) J Phys (Paris) 36:1199
48. McLeish TCB, Ball RC (1986) J Polym Sci Polym Phys Edn 24:1755; McLeish TCB (1987) J Polym Sci Polym Edn Phys 25:2253
49. Watanabe H, Kotaka T (1984) Macromolecules 17:2316
50. Pearson DS, Halfand E (1984) Macromolecules 17:888
51. McLeish TCB, Allgaier J, Bick DK, Bishko G, Biswas P, Blackwell R, Blottière B, Clarke N, Gibbs B, Groves DJ, Hakiki A, Heenan R, Johnson JM, Kant R, Read DJ, Young RN (1999) Macromolecules 32:6734–6758
52. Inkson NJ, Graham RS, McLeish TCB, Groves DJ, Fernyhough CM (2006) Macromolecules 39:4217–4227
53. Pütz M, Kremer K, Grest GS (2000) Europhys Lett 49:735
54. Everaers R (1998) Eur Phys J B 4:341
55. Das C, Inkson NJ, Read DJ, Kelmanson Mark A, McLeish TCB (2006) J Rheol 50:207
56. Das C, Read DJ, Kelmanson MA, McLeish TCB (2006) Phys Rev E 74:011404
57. Milner ST, McLeish TCB (1998) Phys Rev Lett 81:725
58. Likhtman AE, McLeish TCB (2002) Macromolecules 35:6332–6343
59. Marrucci G, Non-Newt J (1996) Fluid Mech 62:279
60. Mead DW, Doi M, Larson RG (1998) Macromolecules 31:7895
61. Likhtman AE, McLeish TCB, Milner ST (2000) Phys Rev Lett 85:4550
62. Graham RS, Likhtman AE, Milner ST, McLeish TCB (2003) J Rheol 47:1171
63. Fetters LJ, Lohse DJ, Graessley WW (1999) J Polym Sci Polym Phys Edn 37:1023
64. Bent J et al (2003) Science 301:1691–1695
65. Graham RS (2006) Macromolecules 39:2700–2709
66. Heinrich M, Pyckhout-Hintzen W, Allgaier J, Richter D, Straube E, McLeish TCB, Wiedenmann A, Blackwell RJ, Read DJ (2004) Small-Angle Neutron Scattering Study of the Relaxation of a Melt of Polybutadiene H-Polymers Following a Large Step Strain. Macromolecules 37:5054–5064
67. McLeish TCB, Larson RG (1998) J Rheol 42:81
68. Inkson NJ, McLeish TCB, Groves DJ, Harlen OG (1999) J Rheol 43:873
69. Blackwell RJ, McLeish TCB, Harlen OG (2000) J Rheol 44:121
70. Graham RS, McLeish TCB, Harlen OG (2001) J Rheol 45:275
71. Lee K, Mackley MR, McLeish TCB, Nicholson TM, Harlen OG (2001) Experimental observation and numerical simulation of transient stress fangs within flowing molten polyethyelene. J Rheol 45:1261–1277
72. Ruokolainen J, Mezzenga R, Fredrickson GH, Kramer EJ, Hustad PD, Coates GW (2005) Morphology and thermodynamic behaviour of syndiotactic polypropylene-poly(ethylene-co-propylene) block polymers prepared by living olefin polymerization. Macromolecules 38(3):851–60
73. Adhikari R, Michler GH (2004) Influence of molecular architecture on Morphology and Micromechanical behaviour of styrene/butadiene block copolymer systems. Prog Polym Sci 29:949–86
74. Leibler L, Ajdari A, Mourran A, Coulon G, Chatenay D (1994) Ordering in macromolecular systems. Springer, Berlin
75. Matsen MW (2005) In: Gompper G, Schick M (eds) Soft Condensed Matter. Wiley-VCH, Berlin
76. Onuki A (2002) Phase Transition Dynamics. Cambridge University Press, Cambridge
77. Henderson IC, Clarke N (2004) Macromolecules 37:1952–1959
78. Tanaka H, Araki T, Koyama T, Nishikawa Y (2005) Universality of viscoelastic phase separation in soft matter. J Phys Cond Mat 17:S3195–S3204
79. Clarke N, McLeish TCB, Pavawongsak S, Higgins JS (1997) Viscoelastic Effects on the Phase-Separation of Polymer Blends. Macromolecules 30(15):4459–4463
80. Mao Y, McLeish TCB, Teixeira PIC, Read DJ (2001) Asymmetric landscapes of early spinodal decomposition. Eur Phys J E 6:69–77
81. Semenov AN (1985) Sov Phys–JETP 61:733
82. Matsen MW (2002) The standard Gaussian model for block copolymer melts. J Phys Cond Mat 14:R21–R47
83. Leibler L (1980) Macromolecules 13:1602
84. Fredrickson GH (2007) Computational field theory of polymers: opportunities and challenges. Soft Matter 3:1329–1334
85. Leibler L (2005) Nanostructured plastics: Joys of self-assembling. Prog Polym Sci 30:898–914
86. Discher DE, Eisenberg A (2002) Science 297:967–973
87. Govorun EN, Ivanov VA, Khokhlov AR, Khalatur PG, Borovinsky AL, Grosberg AYU (2001) Phys Rev E 64:R40903
88. Khokhlov AR, Semenov AN, Subbotin AV (2005) Shape transformations of protein-like copolymer globules. Eur Phys J E 17:283–306
89. Dinner AR, Šali A, Smith LJ, Dobson CM, Karplus M (2000) Understanding protein folding via free-energy surfaces from theory and experiment. TIBS 25:331–339
90. McLeish TCB (2005) Protein Folding in High-Dimensional Spaces: Hypergutters and the Role of Nonnative Interactions. Biophys J 88:172–183

91. Hugel T, Holland NB, Cattani A, Moroder L, Seitz M, Gaub HE (2002) Single-Molecule Optomechanical Cycle. Science 296:1103–1106
92. Bu Z, Biehl R, Monkenbusch M, Richter D, Callaway DJE (2005) Coupled protein domain motion in Taq polymerase revealed byneutron spin-echo spectroscopy. PNAS 102:17646–17651
93. Hawkins RJ, McLeish TCB (2004) Coarse-Grained Model of Entropic Allostery. Phys Rev Lett 93:098104
94. Lehn JM, Mascal M, DeCian A, Fischer J (1990) J Chem Soc Chem Commun pp 479
95. Davies RPW, Aggeli A, Beevers AJ, Boden N, Carrick LM, Fishwick CWG, McLeish TCB, Nyrkova I, Semonov AN (2006) Self-assembling β-Sheet Tape Forming Peptides. Supramol Chem 18:435–443
96. Barford W (2005) Electronic and Optical Properties of Conjugated Polymers. Oxford University Press, Oxford
97. Doi M, Edwards SF (1986) The Theory of Polymer Dynamics. Oxford University Press, Oxford

Books and Reviews

Larson RG (1999) The Structure and Dynamics of Complex Fluids. Clarendon Press, Oxford
Fredrickson GH (2006) The Equilibrium Theory of Inhomogeneous Polymers. Oxford University Press, Oxford

Polymers, Non-linearity in

Koh-hei Nitta
Division of Material Sciences, Graduate School of Natural Science and Technology, Kanazawa University, Kanazawa, Japan

Article Outline

Glossary
Definition of the Subject
Introduction
Topological Analysis of Branched Molecules
Ideal Chain Models
Chain Statistics of Nonlinear Polymers
Chain Dynamics of Nonlinear Polymers
Future Directions
Bibliography

Glossary

Connectivity Connectivity extends the concept of adjacency and is essentially a form (and measure) of concatenated adjacency. If it is possible to establish a path from any vertex to any other vertex of a graph, the graph is said to be connected; otherwise, the graph is disconnected. A graph is totally disconnected if there is no path connecting any pair of vertices. This is just another name to describe an empty graph or independent set.

Distance The distance d_{uv} between two vertices u and v in a graph is the length of a shortest path between them. When u and v are identical, their distance is 0. When u and v are unreachable from each other, their distance is defined to be infinity ∞.

Ideal chain The interactions between adjacent structural units along the chain are called short-range interaction, whereas the interactions between units which are far removed from each other along the chain are called long-range interaction. Note that the long-range interactions are typically short-range in space. An ideal chain is the unperturbed chain without the intramolecular long-range interaction between structural units where only short rage interaction is considered. A basic theoretical model for ideal flexible chains is the Gaussian chain, which assumes a number of ideal beads with intramolecular distance between them following a Gaussian distribution.

Long chain branching The clearest definition of this is that to be long a branch needs to have a molecular weight at least greater than the entanglement molecular weight which is defined in terms of the plateau modulus, which can be directly measured from the linear viscoelasticity of the polymer.

Nonlinear polymer These are branched and cross-linked polymers which contains some polyfunctional units. This term is reserved for functionalities exceeding two and the branches formed from the polyfunctional units sufficiently are long or large.

Polymer A chain like molecule make up of repetition of a particular atomic group joined together by covalent bonds. The chain-like molecule is usually called the polymer if the entanglement and intertwining interaction occurs in the melt state and in the concentrated solution. The basic unit of this sequence is called the 'structural unit', and the number of units in the sequence is the degree of polymerization.

Tree A tree is a connected acyclic simple graph. A vertex of degree 1 is called a leaf, or pendant vertex. An edge incident to a leaf is a leaf edge, or pendant edge. (Some people define a leaf edge as a leaf and then define a leaf vertex on top of it. These two sets of definitions are often used interchangeably.) A non-leaf vertex is an internal vertex. Sometimes, one vertex of the tree is distinguished, and called the root. A rooted tree is a tree with a root. Rooted trees are often treated as directed acyclic graphs with the edges pointing away from the root.

Definition of the Subject

Branching formation is known from the beginning of polymer chemistry and modern synthesis methods make it possible to prepare a great variety of nonlinear polymers with specific branching structure. Branched polymers are nowadays becoming more and more important so that analytical techniques to reveal the role of branching in macroscopic properties of polymers are desired. This article demonstrates that the graph-theoretical approaches are most effective when providing topological and physical insights into the nonlinearity in polymer architecture. Thus, the problems of the dynamics and statistics of any branched molecule are shown to be completely reduced to the problem of the eigen-polynominal of graph, resulting in that various ideas and concepts, thus, obtained from the graph theory can be applied directly to the topological analysis for architecture in nonlinear polymers. This will leads to a new kind of paradigm in polymer science and engineering.

Introduction

A polymer is a chain-like molecule that comprises huge number of repeating structural units or atoms connected by chemical bonds. Modern polymerization techniques for the preparation of chain-like molecules can produce chain branching and these branched polymers possess practically important properties and specific phenomena involving processability and rheological properties [19,34,50,55,87,109] which cannot be reached for linear polymers.

Branching formation was suspected almost from the beginning of the study of polymer chemistry in the 1930s. Flory [41] first pointed out the possibility of the occurrence of branching in the free-radical polymerization of diene monomers. The occurrence of branching had been well recognized by the early 1940s and since then, many theoretical and experimental studies related to the effects of branching on the properties of polymers have been carried out. In 1953, several important papers [2,7,108,115] showed that polyethylene materials produced by the free-radical process have not only a significant number of short branches but also long-chain branches.

In the 1960s, several research groups found that the "living polymer" procedure of anionic polymerization makes it possible to synthesize definitely known branched polymers [88]. During the 1970s and 1980s in addition to random branching that occurs in industrial polymerization, more rational synthetic methods based on living anionic polymers have facilitated the preparation of a great variety of polymers with specific branching structures such as star-shaped polymers having several branches attached to a single polyfunctional branching point [58] and comb-shaped polymers having a given number of branching points randomly distributed along a backbone [93,124]. Comb-shaped polymers with branching points of functionality greater than three are also sometimes called polymeric brushes [80]. During the period from the end of the 1980s to the 1990s, it has been possible to build structures possessing regular "hyper-branched" polymers [74] or starburst polymers with radial symmetry, which is generally called a "dendrimer" [127]. The physical and chemical properties of such dendritic polymers, which essentially differ from those of not only linear polymers but also comb- and star-shaped polymers, enables unpredictable wide applicability [49]. These specific branched polymers are illustrated in Fig. 1.

In order to clarify the effects of long chain branching on polymer properties, we need some quantitative measures that reflect molecular branching: one of the quantitative factors is the average size of isolated single polymers characterized through the measurement of dilute solution properties. The smaller expansion in the space of isolated branched molecules as compared with isolated linear molecules of the same molecular weight is the basis for the most fundamental methods for estimating branching as demonstrated by Zimm–Stockmayer [139].

For the purpose of estimating branching, therefore, considerable effort has been extended to obtain the chain dimensions such as the mean square radius of gyration $\langle s^2 \rangle$ estimated by performing the light scattering (LS) measurements and/or the mean Stokes radius r_H, estimated by using the intrinsic viscosity $[\eta]$ data. In addition, a gel permeation chromatograph (GPC) [20,26,60] is one of the most popular devices used for the fractionation of a polymer according to the volume dimension $[\eta]M$ where M is the molecular weight. Therefore, the effort to relate $[\eta]$ to $\langle s^2 \rangle$ or r_H, has been continuing for a long time by a combination of GPC and LS or $[\eta]$ techniques, although they cannot satisfactorily describe the structural details of branched polymers such as the number of branches, the branch length, and their position along the backbone.

Under ordinary conditions, in a dilute solution, there are significant long-range intramolecular correlations that "perturb" the conformation of the chains. However, under special conditions in the Θ state [84], the effects of the excluded volume of a structural unit vanish and the dimensions of the macromolecule adopt their unperturbed values which are appropriate for evaluating the branching degree based on conformational statistics.

Consequently, extensive research has been carried out in calculating the various conformational properties corresponding to well-defined branching architectures. How-

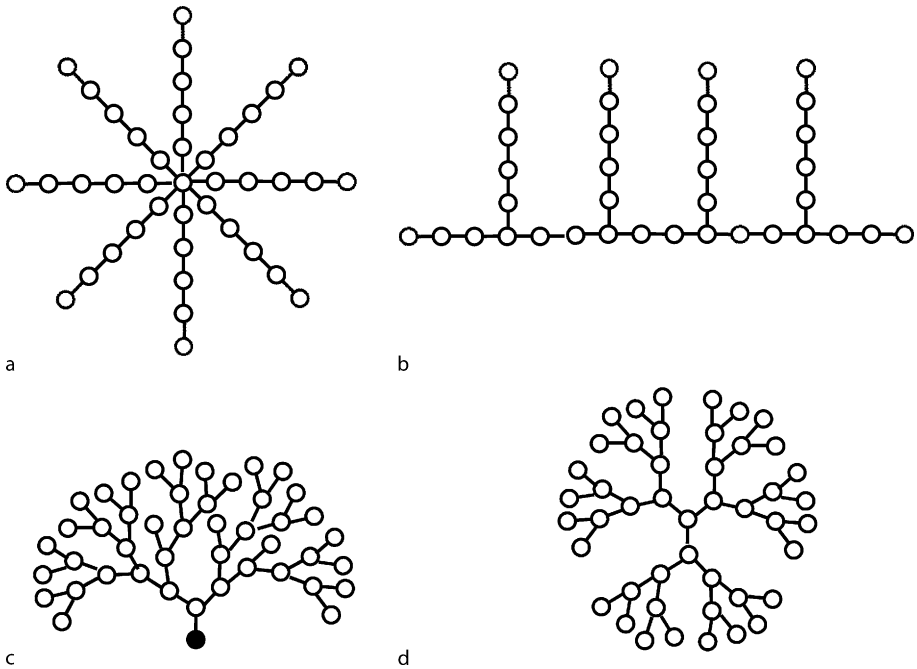

Polymers, Non-linearity in, Figure 1
Structural images of a regular star; b regular comb; c hyperbranch; d dendrimer

ever some mathematical difficulties often arise in applying statistical treatments to unperturbed branched polymers. The Monte Carlo simulation methods [6,48] have been powerful computational tools in the treatment of branched chains more complicated than those generally treated by statistical methods. The simulations suggest that intrinsic properties of polymers are influenced not only by the number of branches but also by the branch length and their position along the backbone. Quantitative data concerning the local dynamics and local structure can be obtained from the molecular dynamics simulation studies. However, it is hard to provide the topological or physical insights into these computer simulation results.

The specification of a linear polymer requires only one parameter; i. e. the degree of polymerization or molecular weight; however, additional key parameters are necessary for the specification of any branched polymer. For understanding molecular branching from a general point of view, systematic studies on the effect of branching on the physical properties in the frame of a homologous series of polymeric materials are desirable. For this purpose, it is also necessary to develop the topological or graph-theoretical methods for analyzing the branching and/or skeletal nature of polymers such as the branch length, the number of branches, and their positions.

The final goal of our work is to set up a rigorous molecular theory for linear as well as nonlinear flexible polymers and to give a description and a reliable prediction of the materials properties in the melt and solid states. This article demonstrates that the graph-theoretical approaches are most effective when establishing such a universal framework for polymers with any structural architecture.

This article deals with the topological nature of tree-like chain molecules as typical nonlinear polymers. This article is organized as follows: The graph-theoretical approach for characterizing branched molecules is described in brief in Sect. "Topological Analysis of Branched Molecules". In Sect. "Ideal Chain Models", the chain conformational statistics in the unperturbed state is discussed in the framework of ideal chain models based on the spring-beads and rod-beads models. The graph-theoretical approach to chain statistics is presented in Sect. "Chain Statistics of Nonlinear Polymers". The relations between the conformational statistics of ideal chains and the graphic representation are also presented. In Sect. "Chain Dynamics of Nonlinear Polymers", it is shown that the dynamics of various types of branched chains are obtained by solving the polynomial equation derived from the graph-theoretical representations. The final section provides a brief summary of this article and future problems.

Topological Analysis of Branched Molecules

An alkane molecule is one of the most basic organic compounds, and it is a set of carbon and hydrogen atoms that are connected to one another by covalent bonds. The topological analysis of a branched alkane begins with a drawing where the atoms are depicted as points and the bonds linking them are depicted as straight lines.

If only the carbon atoms of alkanes are depicted as points and their hydrogen atoms are omitted, there is a one-to-one correspondence between these isomers and the drawings whose points have at most four neighbors. This type of drawings is called the chemical graph [122]. In the graphs, the hydrogen atoms do not normally play a major role in discriminating the structure of a molecule and in enumerating the isomers of alkanes. It was major problem for chemists to predict the number of isomers of any alkane on the basis of simple graphical constructions. The enumeration of the chemical isomers, in particular the constructional isomers of alkanes, has been treated as the subject matter of a mathematical discipline known as graph theory since the pioneering work by Cayley [21] and Sylvester [123].

In graph theory, points (atoms) are generally referred to as *vertices* and lines (bonds) are referred to as *edges*, and the functionality of the atoms is called the "*vertex degree*". We consider only the connected graphs in which every vertex has more than one neighbor, and no loops or multiple edges are involved. A hydrocarbon molecule can be represented as a tree graph G, in which a carbon atom (or vertex) and a bond (or edge) are arbitrarily numbered, and a digraph D, in which each edge of the graph G is arbitrarily directed, as exemplified in Figs. 2a and 2b. If the chemical graph is allowed to have a branch point with more than four degrees, the extended chemical graph, which may be called the molecular graph, could be used to describe the structural diversity in branched polymers by offering quantitative descriptors, called topological indices.

In graph theory, the algebraic expressions using several matrices reflecting the connectivity in graph G [69] are important devices for determining the topological feature of graphs, and the algebraic properties of the characteristic polynomials have been extensively examined. For graph G, the adjacency matrix $\mathbf{A}$ is the most fundamental matrix for the representation of graphs [69] and it is defined in graph theory as a square matrix with the following elements:

$$a_{ij} = \begin{cases} 1, & \text{if the vertexes } i \text{ and } j \text{ are adjacent,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The matrix $\mathbf{A}$ is often called a topological matrix [68]. The order of $\mathbf{A}$ is identical with the total number of vertices in G. The adjacency matrix $\mathbf{A}$ of G has been useful for characterizing and encoding the skeletal structure of the corresponding molecules.

The characteristic polynomial of a molecular graph represents an important, even if not unique, molecular invariant. It is defined as

$$\Phi(\mathbf{A};\lambda) = \mathrm{Det}\,|\mathbf{A} - \lambda\mathbf{E}|\,, \tag{2}$$

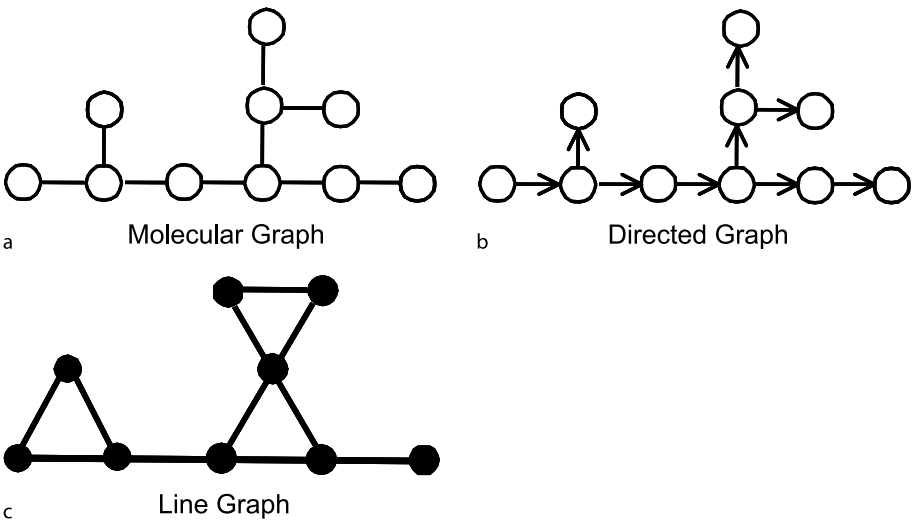

Polymers, Non-linearity in, Figure 2
Representation of a nonlinear molecule a the ordinary graph; b the digraph; c the line graph. The line graph (c) is transformed from the graph (a)

where $\mathbf{E}$ is a unit matrix of the same order as $\mathbf{A}$. The list of eigenvalues of a matrix calculated from the characteristic polynomial is called the spectrum of the matrix which includes much quantitative information on topological nature of the molecules.

Another matrix describing a graph is the incidence matrix [70], which represents a linear mapping and determines the homology of the graph. The incidence matrix $\mathbf{B} = (b_{ij})$ of a digraph D is defined by the following:

$$b_{ij} = \begin{cases} +1, & \text{if edge } j \text{ starts from vertex } i\,, \\ -1, & \text{if edge } j \text{ terminates in vertex } i\,, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The incidence matrix may be constructed for any graph having n vertices and m edges by setting up an $n \times m$ matrix array; the rows and columns of the matrix then correspond to the vertices and the edges of the digraph, respectively. Also, the signless incidence matrix $\mathbf{C} = (c_{ij})$ of an ordinary graph G can be defined and its elements are the absolute values of those of $\mathbf{B}$. Thus, the number 1 is inserted in the (i, j)th position in the array, if the jth edge is coincident with the ith vertex and all the other entries in the array are zeros.

Kirchhoff [75] discovered graphs while solving problems involving the calculation of currents in electrical networks where a connectivity matrix $\mathbf{K} = \mathbf{C}^\top\mathbf{C}$ was considered for admittance conductivity. The superscript "T" indicates the transpose of a matrix and a vector. He found the following formula relating the vertex adjacency matrix to the incidence matrix for tree graphs:

$$\mathbf{A}_\mathrm{L} = \mathbf{K} - 2\mathbf{E}\,, \tag{4}$$

where $\mathbf{A}_\mathrm{L}$ represents the adjacency matrix for the line graph of G. The line graph $L(G)$ of G is formed by replacing the edges of G by vertices in a manner such that the vertices in $L(G)$ are connected whenever the corresponding edges in G are adjacent. The example of a line graph is shown in Fig. 2c. Furthermore, a combination of (4) and (2) gives

$$\Phi(\mathbf{K};\lambda) = \Phi(\mathbf{A}_\mathrm{L};\lambda - 2)\,. \tag{5}$$

Consequently, the eigenvalues λ_i of $\mathbf{K}$ of a graph can be calculated from the eigenvalues μ_i of $\mathbf{A}_\mathrm{L}$ of its line graph; i. e. $\lambda_i = \mu_i + 2$.

The topological analysis of graphs has widely been performed by the Laplacian matrices being related to both the adjacency and incidence matrices and they are defined by

$$\mathbf{L} = \mathbf{V} - \mathbf{A}\,, \tag{6}$$

$$\mathbf{L}^+ = \mathbf{V} + \mathbf{A}\,, \tag{7}$$

where $\mathbf{V}$ is a diagonal matrix whose entries are the vertex degrees, $\mathbf{L}^+$ is the signless Laplacian matrix [116,131], the entries of which is the absolute values of the entries of $\mathbf{L}$. The Laplacian matrices can be represented using the incidence matrices [24,130]

$$\mathbf{L} = \mathbf{B}\mathbf{B}^\top\,, \tag{8}$$

$$\mathbf{L}^+ = \mathbf{C}\mathbf{C}^\top\,. \tag{9}$$

The Laplacian matrix $\mathbf{L}$ is in agreement with the Zimm matrix $\mathbf{Z}$ which has been used for molecular dynamics of linear as well as branched polymers [137,138]. The details of the molecular dynamics are described in Sect. "Chain Dynamics of Nonlinear Polymers". In addition, one can find the relation

$$\Phi(\mathbf{L};\lambda) = \Phi(\mathbf{L}^+;\lambda)\,, \tag{10}$$

indicating that the eigenvalues of the signless matrix $\mathbf{L}^+$ are identical with those of the Laplacian matrix $\mathbf{L}$. Furthermore, comparing the characteristic polynomials of $\mathbf{L}^+ = \mathbf{C}\mathbf{C}^\top$ and $\mathbf{K} = \mathbf{C}^\top\mathbf{C}$, we can find the following relation:

$$\Phi(\mathbf{L}^+;\lambda) = \lambda\,\Phi(\mathbf{K};\lambda)\,. \tag{11}$$

The eigenvalues of $\mathbf{L}^+$ contain one zero eigenvalue and the non-zero eigenvalues of $\mathbf{L}^+$ are identical with those of $\mathbf{K}$.

Another important invariant of a molecular graph is its distance matrix $\mathbf{D}$ [70] which contains the shortest paths (distances d_{ij}) between every pair of connected vertices. The entries in the distance matrix are related to the entries in powers of the adjacency matrix $\mathbf{A}$.

A secular determinant giving the Hückel molecular orbitals for the π electrons of an unsaturated hydrocarbon is reduced to the same form as the determinant of $\mathbf{A}$ [3,56,57]. In other words, the problem of the Hückel orbital energies can be completely reduced to the eigenvalue problem of $\mathbf{A}$. Thus, the spectrum of the matrix $\mathbf{A}$ yields the energy levels of molecular orbitals.

One of the earliest graph invariants or topological indices is known to mathematicians as the vertex number and to chemists as the carbon number. The carbon number is an appropriate index only for linear chain molecules; however, it is not well suited to branched molecules, which may have considerably different skeletal structures as compared to one another even if they have the same number of carbon atoms. Because there exists many different molecules with the same carbon number, the carbon number seems to be an index with low discriminating power.

Apparently, it is necessary to develop other indices that can effectively distinguish various types of isomers having different branching structures.

The first topological index capable of characterizing the branchedness of molecules was proposed by Wiener [133,134] to predict the boiling points of isomeric alkanes. The Wiener index W, named by Platt [102,103], is defined as the sum of the distances between any two carbon atoms in a hydrocarbon molecule:

$$W = \sum_{i=1}^{N-1} i\, k_i \,, \tag{12}$$

where k_i stands for the total number of pairs of atoms whose separation is i. Platt attempted to interpret this index that $W^{1/3}$ is a sort of the mean molecular diameter [103].

The Wiener index of a hydrocarbon molecule is like the carbon number because it becomes generally larger for molecules with higher molecular weights; however, W also provides some measure of a molecule's branching structure. Since Wiener devised it, several other researchers found that the Wiener index correlated surprisingly well with properties for certain types of hydrocarbon molecules, as well as conjugated polymers [11,12,13,14,89,90] such as heat capacity, viscosity, surface tension, refractive index and electron energy.

Hosoya [71] found that W can also be obtained by the half sum of the off-diagonal elements of a distance matrix whose element d_{ij} is the number of edges for the shortest path between the ith and jth vertices: this not only offers an alternative method to determine W but also allows a particular extension of W to cyclic structures. The W is given by

$$W = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij} \,, \tag{13}$$

where N is the total number of vertices. Rouvray examined the sum of the elements of the distance matrix, independently and considered it as a topological index [111]. The fact that Rouvray's index is equal to $2W$ was soon recognized. Bonchev et al. [15] studied the Wiener index of any branched graph and succeeded in formulating structural features such as the branching point and the branch length.

The Laplacian matrix is a real symmetric matrix. The diagonalization of the Laplacian matrix for a graph G with N vertices produces N real eigenvalues, $\lambda_1 > \lambda_2 > \ldots > \lambda_N = 0$, where the smallest eigenvalue is always zero. Let a graph G be a tree; then, the Wiener index of the tree can be obtained in terms of its Laplacian eigenvalues as follows [91,95]:

$$W = N \sum_{i=1}^{N-1} \frac{1}{\lambda_i} \,. \tag{14}$$

In chemical graph theory, the distance matrix $\mathbf{D}$ accounts for the bond interactions of atom in molecules. However, these interactions decrease as the distance between atoms increases so that the research groups of Balaban [73] and Trinajistić [104] respectively, proposed the reciprocal distance matrix whose entries are given by d_{ij}^{-1}. The reciprocal distance matrix enables the calculation of a Wiener index analog, as the half sum of its entries:

$$H = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^{-1} \,. \tag{15}$$

This index H is called the "Harary index". Here we generalize this index for its applicability to polymers as follows:

$$H_\varepsilon \equiv \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^{-\varepsilon} \,. \tag{16}$$

Clearly, one can find $W = H_{-1}$ and $H = H_1$.

Hosoya et al. [72] have introduced a topological index Z; in the case of acyclic molecules, Z is equal to the sum of the absolute values of the polynomial coefficients of the adjacency matrix and it was used to describe the boiling points of various molecules. Recently, in a number of papers, the problem of molecular branching was related to the properties of the characteristic polynomial. Lovász and Pelikán [85] have found that the maximal eigenvalue of a tree graph is a fairly reliable measure of branching.

The Wiener index of a molecule is essentially based on the topological concept of *distance* so that it can be recognized as a measure of the molecular size rather than molecular shape and connectivity. Whereas Randić proposed a different type of index which is based on the topological concept of *degree*. The Randić index [107], or the connectivity index χ, can be defined in terms of the atomic contributions and relates also to the elements of the normalized Laplace matrix. The vertex-connectivity index is defined as

$$\chi = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} (d(v_i)\, d(v_j))^{-0.5} \,, \tag{17}$$

where $d(v_i)$ is the degree of vertex i. Estrada [37] introduced the edge-connectivity index on the basis of the Randić index analog. These connectivity indices were found to be surprisingly correlated with the density, solu-

bility in water, molar volume, molar refractivity, and various types of biological responses.

In this section we introduced typical topological indices which have a possibility to be suited to branched polymers. At present, the mathematical characterization of molecules has led to a large number of molecular descriptor, and their number continues to grow. These descriptors, often referred to as topological indices, will play an important role in structure-property and structure-activity studies. However, the important task is to find the physical meanings of these topological descriptors.

Ideal Chain Models

In this section, we deal with only the statistical properties of ideal polymer chains in the unperturbed state and only short-rage interactions are considered. Flory proposed that the polymer molecules in the molten state are unperturbed as they are in a Θ solvent [42]. Small angle neutron scattering data [39,64] support the Flory theorem. An elegant explanation for the Flory theorem was given by de Gennes [29]. Therefore, the concept of ideal chain model plays a central role in analyzing the fundamental topological feature of any branched polymer.

Let us consider the simple homologous series of normal alkane hydrocarbons designated as linear polyethylenes. These polymers have a general structure where the number of $-CH_2$-units is connected by chemical single bonds. Although chemical bonds are fairly rigid with respect to stretching and to bending of the valence angles between adjacent bonds, a single-polymer molecule has many internal degrees of rotational freedom about each C-C bond in the polyethylene molecule, resulting in that it can adopt many different configurations, thereby necessitating the use of statistical mechanics.

Here we consider three models of varying levels of complexity and reality for flexible chains. The simplest model is called the "Kramers chain" [81] by considering the polymer to be composed of points or units that are freely joined by bonds of a fixed length. For the random flight chain the bond angles are fixed; however, there is free rotation about the bonds. This is called the "Kirkwook–Riseman chain" [76] (see Fig. 3a). The bond angle θ of typical vinyl polymers is approximately 111.5° and the bond length is approximately 0.154 nm. More realistic models include the hindered rotations about the bonds that form the chain backbone, and they also possibly include the steric hindrances which result in the interdependence of the internal rotational degrees of freedom. Three types of micro-conformations, i. e. one trans and two gauche states, occur every C-C bond.

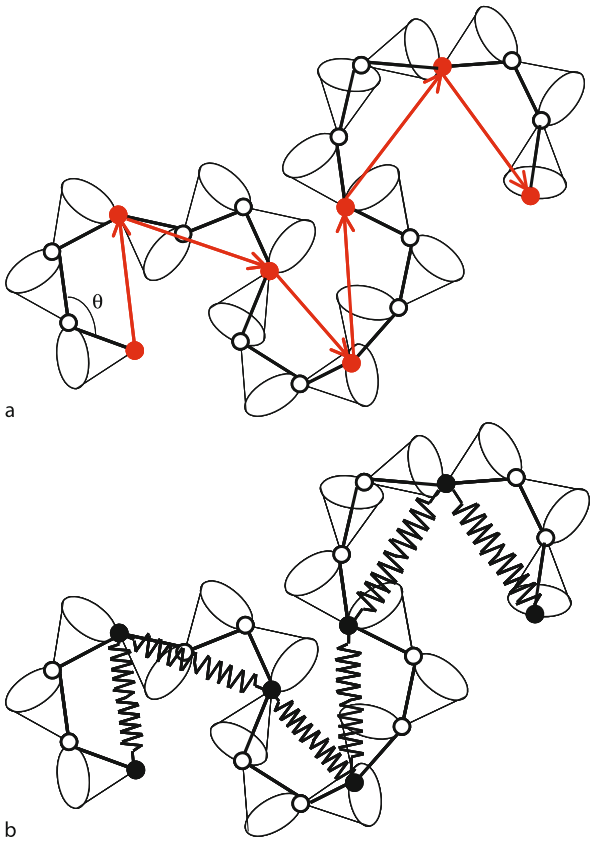

Polymers, Non-linearity in, Figure 3
Kirkwood–Riseman chain and a its equivalent chain model; b its Gaussian model

Each conformation exists for only a very short time: the observed proportions of the microconformations, and thus the resulting macroconformations, are temporal averages over all molecules, resulting in that the polymer chain forms a random coil. Consider the end-to-end vector $\mathbf{r}$ joining one end of the linear alkane chain to the other: the average length of the alkane chain can be considered as an indicator of the spreading out or the size of the polymer. If the chain is made up of bonds labeled from numeral 1 through ν, atoms labeled from 1 through N, and if $\boldsymbol{\alpha}_i$ is the bond vector of the ith bond ($i = 1, 2, 3, \ldots, \nu$), we have

$$\mathbf{r} = \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 + \ldots + \boldsymbol{\alpha}_\nu \,. \tag{18}$$

It is apparent that we obtain the relation $\nu = N - 1$ for any chain possessing no loops or rings. Since the average value of $\mathbf{r}$ is zero, its average length can be obtained by taking the square root of $r^2 = \mathbf{r} \cdot \mathbf{r}$:

$$\langle r^2 \rangle_0 = \nu\alpha^2 + \left\langle 2 \sum_{i>j}\sum \mathbf{r}_i \cdot \mathbf{r}_j \right\rangle , \tag{19}$$

where the angular bracket indicates the average of the spatial quantities, the subscript 0 signifies the Θ conditions, and the double sum is over all pairs $i < j$. The end-to-end distance is the spatial size of a linear chain, but it has no significance for a branched chain. The square radius of gyration $\langle s^2 \rangle$ and hydrodynamic Stokes radius r_H are an appropriate measure of the spatial size of linear and branched coils or particles [119,120].

The square radius of gyration is the means of all position of atoms from the center of mass vector. The mean square radius of gyration for unperturbed chains can be derived from the Lagrange theorem [30,43,59]:

$$\langle s^2 \rangle_0 = \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \langle \mathbf{r}_{ij}^2 \rangle \tag{20}$$

and the hydrodynamic radius or the Stokes radius of the coil is obtained from the mean reciprocal distance:

$$\langle r_H^{-1} \rangle_0 = \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \langle |\mathbf{r}_{ij}|^{-1} \rangle , \tag{21}$$

where $\mathbf{r}_{ij}$ is the vector from atoms i to j.

In the random-flight Kramers model, the second term on the right-hand side of (19) disappears. Therefore we have

$$\langle r^2 \rangle_0 = \nu \alpha^2 . \tag{22}$$

It follows that the size of the polymer is proportional to the square root of the bond number. The chain behaves as a ghost or phantom chain that can intersect or cross itself freely and is often called the "Markov chain". Note that in the model α is not always 0.154 nm, which is the actual length of the C-C bond. Mathematically, the possible maximum length of the freely joint chain is undoubtedly given by $\nu\alpha$ which corresponds to the contour length r_{cont} because it paces off the contour of the chain. The contour length of a chain is given by geometry in an all-trans conformation with a fixed bond angle θ as $r_{cont} = n_C a \sin(\theta/2)$ where a is 0.154 nm and n_C is the number of C-C bonds. Thus the end-to-end distance of a fully extended real chain should be the maximum length, $\nu\alpha$ of the straightly aligned Kramers chain. The value of α may be the length of the virtual bond which is a projection of the real bonds onto the end-to-end vector. In this case, the virtual bond length is given by $\alpha = 0.254$ nm and $\nu = n_C/2$ because $a = 0.154$ nm and $\theta = 111.5°$.

The freely rotating chain is identical to the freely joined chain in every respect, except that the angle between two adjoining bonds is held fixed at a predetermined value. In this Karkwood–Risemann chain (see Fig. 3a), (19) becomes

$$\langle r^2 \rangle_0 = \left(\frac{1 - \cos\theta}{1 + \cos\theta} + \frac{2\cos\theta\,[1 - (-\cos\theta)^\nu]}{\nu\,(1 + \cos\theta)^2} \right) \nu\alpha^2 . \tag{23}$$

In a more realistic model, the symmetric hindered rotation model, the conformations of the bonds do not occur with equal probability; however, they are determined by a torsion angle-dependent potential function $u(\phi)$, where $\phi = 0°$ for the tarns state and $\phi = \pm 120°$ for the gauche state. At equilibrium, the bond conformations are distributed according to the Boltzmann distribution, and we obtain

$$\langle r^2 \rangle_0 = \left(\frac{1 - \cos\theta}{1 + \cos\theta} + \frac{2\cos\theta\,[1 - (-\cos\theta)^\nu]}{\nu(1 + \cos\theta)^2} \right) \times \left(\frac{1 - \langle\cos\phi\rangle}{1 + \langle\cos\phi\rangle} \right) \nu\alpha^2 , \tag{24}$$

where the average value of $\langle\cos\phi\rangle$ can be readily obtained from the assumed potential energy function.

Consequently both these realistic models in the unperturbed state adapt to the same type of equation as

$$\langle r^2 \rangle_0 = C\nu\alpha^2 , \tag{25}$$

where C is called the characteristic ratio and depends on the nature of the polymer. As is evident from (23) and (24), the C is slightly dependent on the value ν but C may be considered to be constant for flexible polymers or large ν. The chain composed of ν bonds of length α can be regarded as a freely joined chain consisting of n "effective bonds" of length b by taking $b = C\alpha$ and $n = \nu/C$. The two parameters n and b are determined under the restriction that the contour length must be the same as that of the real chain, i. e., $\nu\alpha = nb$. Then, the end-to-end distance of any unperturbed chain molecule can be renormalized as

$$\langle r^2 \rangle_0 = nb^2 . \tag{26}$$

Such a coarse-graining polymer chain is often called the *equivalent chain*. The effects of the chemical bulkiness of units, the inclusions of double bonds and short chain branches can be attributed to the parameter C. Thus, a real flexible polymer may be effectively treated as a freely jointed chain; i. e. Kramers chain, of n bonds with a length of b, and the details of the chemical structure of real chains can be smeared out (see Fig. 3a).

The equilibrium distribution function in freely joined chains is assumed to be identical to the random walk. It should be noted here that the random walk distribution

is inconsistent with the equilibrium statistical mechanics [52] but the differences are probably inconsequential for large N. As described previously, a polymer chain can take many different macroconformations above a molecular glass temperature. This corresponds to the micro-Brownian motion and the chains are rapidly converted into other macroconformations. The instantaneous shape, which is obtained by time-averaging over many conformations, can be described by a Gaussian distribution for large N. The distribution function $W(\mathbf{b}_i)$ of the vector $\mathbf{b}_i$ between adjacent units i and $i+1$ can be defined either as the time-averaged incidence of $\mathbf{b}_i$ within the specified range for a given molecule or as the average incidence for an ensemble of many identical units subject to identical conditions:

$$W(\mathbf{b}_i) = \left(\frac{3}{2\pi\langle\mathbf{b}^2\rangle}\right)^{3/2} \exp\left(-\frac{3\mathbf{b}_i^2}{2\langle\mathbf{b}^2\rangle}\right) , \tag{27}$$

where $\langle\mathbf{b}^2\rangle$ is the time-averaged square bond length. Let the position vector of the ith effective unit be $\mathbf{r}_i$. Then, the distribution of the effective bond vector $\mathbf{b}_i = \mathbf{r}_{i+1} - \mathbf{r}_i$ is given by a Gaussian distribution function (27); hence the probability distribution of the set of position vectors $\{\mathbf{r}_i\} = \{\mathbf{r}_1, \ldots, \mathbf{r}_N\}$ is proportional to

$$P(\{\mathbf{r}_i\}) = \left(\frac{3}{2\pi\langle\mathbf{b}^2\rangle}\right)^{3/2} \exp\left(-\frac{3}{2\langle\mathbf{b}^2\rangle}\sum_{i=1}^{N-1}(\mathbf{r}_{i+1}-\mathbf{r}_i)^2\right) . \tag{28}$$

The equilibrium state of this chain is described by a distribution function proportional to $\exp(-V/k_\mathrm{B}T)$, where V is the potential energy, k_B is the Boltzmann constant and T is the absolute temperature; therefore, if we choose

$$V = \frac{1}{2}k\sum_{i=1}^{N-1}(\mathbf{r}_{i+1}-\mathbf{r}_i)^2 , \tag{29}$$

where $k = 3k_\mathrm{B}T/\langle\mathbf{b}^2\rangle$, then the chain's equilibrium distribution is given by (28). Consequently, the Gaussian chain can be modeled as a chain of beads connected by a Hookean spring with a spring constant k of (29) as shown in Fig. 3b. In the model, if the square root of $\langle\mathbf{b}^2\rangle$ is identical to the effective bond length b, $\langle\mathbf{b}^2\rangle^{1/2} = b$, then the end-to-end distance of Gaussian chain may be identical to the value given by (26). In this case, the spring of the Gaussian chain is called the "segment" (see Fig. 3b). The Gaussian chain plays a central role in the study of chain dynamics [17,110,137]. The applicability of the Gaussian chain model to any type of branched polymers was soon attempted [22,23,32,67,79,138]. The length of any bond and the angles between the bonds are insignificant because a connection between segments, rather than the precise nature of the connection of atoms, is important in topological analysis for flexible polymers.

Chain Statistics of Nonlinear Polymers

One of the quantitative measures that reflect molecular branching is the average size of isolated single branched polymers. The expansion in the space of isolated branched polymers is smaller than that of linear chains with the same molecular weight. As shown in the previous section, we can treat any flexible polymer as a random-flight chain model which can be regarded as a molecular graph composed of $N-1$ statistical bonds (edges) of a length b joining N points (vertices) of unit mass. The quality b is the segment length and it depends on the conformational characteristics of the polymer species.

When the branched chains obey random-flight statistics, the average of the scalar product extending over all sets of bond vectors $\mathbf{b}_i$ is given by $\langle\mathbf{b}_i\cdot\mathbf{b}_j\rangle = b^2\delta_{ij}$ where δ_{ij} is the Kronecker delta function. Thus, the mean square of $\mathbf{r}_{ij}$ becomes

$$\langle\mathbf{r}_{ij}^2\rangle = \sum_{i,j} b^2 = d_{ij}\, b^2 , \tag{30}$$

because the number of bonds between the ith and jth bead is identical to the graph-theoretical distance d_{ij}. Then, combination of (13) and (20) with (30) yields

$$\langle s^2\rangle_0 = \left(\frac{b}{N}\right)^2 W . \tag{31}$$

This is a graph-theoretical expression for the mean square radius of gyration for any unperturbed random-flight chain [96,132]. The radius of branched chains becomes proportional to $W^{1/2}N^{-1}$.

As a measure of branching, Zimm–Stockmayer proposed a parameter g [139] which is defined by the ratio of $\langle s^2\rangle$ for a given branched chain to that for the linear chain at the same number of statistical units. Consequently, the g-factor is equal to the Wiener index normalized by the Wiener index of linear chain proposed by Bonchev et al. [9,10]:

$$g = \frac{\langle s^2\rangle_0}{\langle s^2\rangle_0^\mathrm{L}} = \frac{W}{W_\mathrm{L}} , \tag{32}$$

where the superscript L indicates the linear chain. The Wiener index of linear chains W_L can be easily evaluated

by substituting the relation $k_i = N - i$ into (12):

$$W_{\mathrm{L}} = \sum_{i=1}^{N-1} i\,(N-i) = \frac{1}{6}\,N\big(N^2-1\big)\,. \tag{33}$$

Substituting (33) into Eq. (31) yields the equation of $\langle s^2 \rangle$ of linear chains, which was first derived by Kramers [31,81,112]. Substitution of (33) into (32) results in that the g-factor of any branched polymer can be represented by the Wiener index W and the number of beads N:

$$g = 6\,\frac{W}{N(N^2-1)} \cong 6N^{-3}W\,. \tag{34}$$

Here we present the graph theoretical expressions of typical types of branched random-flight chains to evaluate their mean square radius of gyration according to the mathematical procedure by Bonchev–Trinajstić [15]. The branched chains are here classified into two types: one is a type of a non-linear chain with a fixed type of branching: referred to as a "specifically branched chain", and the other is an unfixed type of branching: referred to as a "randomly branched chain".

Specifically Branched Chains

Let us consider a chain with any specified branching geometry in which various linear side chains are arbitrarily connected with a linear main chain. Bonchev–Trinajstić [15] showed that the Wiener index is given by the sum of the total path number between any two beads in the main chain, W_{ch}, that between any two beads in a common side chain, W_{br}, that between a bead in the main chain and a bead in a side chain, $W_{\mathrm{ch-br}}$ and that between a bead in a side chain and a bead in another side chain, $W_{\mathrm{br-br}}$.

The molecular graph of such a specifically branched chain is illustrated in Fig. 4, in which N_0 is the total number of beads in the main chain and the beads are labeled with the numerals 1 through N_0; p is the total number of branched beads; m_k is the label numeral of the kth branched bead, where k enumerates the branched beads ($1 \le k \le p$); n_k is the total number of branches connected with the labeled bead m_k; and $N_{k,j}$ is the number of beads in the jth branch, where j enumerates the branches attached to the bead m_k. Then, the Wiener index of graphs with linear branches is given by

$$W = W_{\mathrm{ch}} + W_{\mathrm{br}} + W_{\mathrm{ch-br}} + W_{\mathrm{br-br}}\,, \tag{35}$$

where

$$W_{\mathrm{ch}} = \frac{1}{6}\,N_0\,\big(N_0^2-1\big)\,, \tag{36}$$

$$W_{\mathrm{br}} = \sum_{k=1}^{p}\left[\sum_{j=1}^{n_k}\frac{1}{6}\,N_{k,j}\,\big(N_{k,j}^2-1\big)\right]\,, \tag{37}$$

$$W_{\mathrm{ch-br}} = \sum_{k=1}^{p}\sum_{j=1}^{n_k}\frac{1}{2}\,N_{k,j}\Big[\big(N_0-m_k+1\big)^2 + \big(m_k^2-1\big) + N_0\,N_{k,i}\Big]\,, \tag{38}$$

$$W_{\mathrm{br-br}} = \sum_{k=1}^{p}\left[\sum_{i=1}^{n_k-1}\sum_{j=1+i}^{n_k} N_{k,i}\,N_{k,j} \times \left(\frac{N_{k,i}+N_{k,j}}{2}+1\right)\right]$$

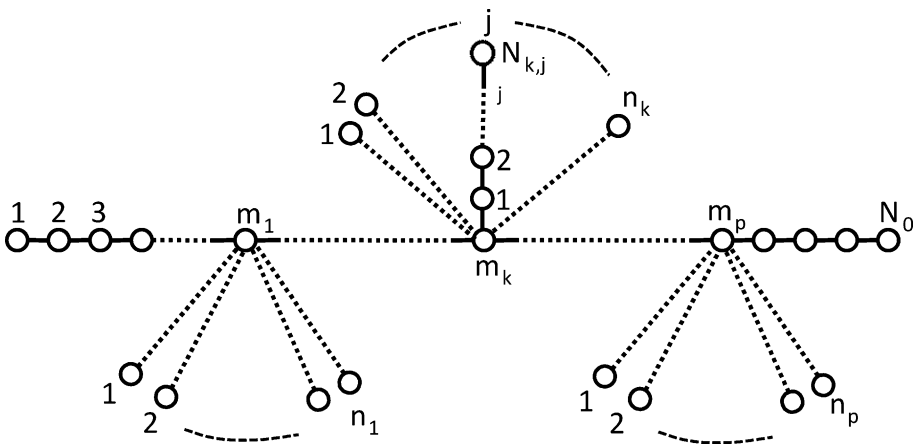

Polymers, Non-linearity in, Figure 4
The molecular graph of a specifically branched chain with linear branches

$$+\sum_{k=1}^{p-1}\sum_{l=1+k}^{p}\left[\sum_{i=1}^{n_l}\sum_{j=1}^{n_k}N_{l,i}\,N_{k,j}\times\left(\frac{N_{l,i}+N_{k,j}}{2}+1+m_l-m_k\right)\right]. \tag{39}$$

Here the total number of beads N becomes

$$N=N_0+\sum_{k=1}^{p}\sum_{i=1}^{n_k}N_{k,i}\,. \tag{40}$$

The first term on the right-hand side of (39) is the sum of the path number extending over all pairs of branches connected with a common branched bead and the second term is the sum extending over all pairs of branches connected with different branched beads. Thus, the first term vanishes for tri-functional comb-shaped molecules and the second term vanishes for star-shaped molecules.

Explicit expressions for the mean square radius of gyration $\langle s^2\rangle_0$ or the g-factor can be obtained by the substitution of (35)–(39) into (31) or (34). Furthermore, the present formula from (36) to (39) for specifically branched chains with linear branches can be recursively extended to include more complicated types of specifically branched chains having any branched branches as well as reduced to simplified types of specifically branched chains. The changes in the Wiener index due to the linkage of two graphs are formulated in general by Polansky et al. [105,106].

Here we exemplify the expressions for the Wiener indices of two specifically branched chains that are of great interest in practical use, that is, star-shaped and tri-functional comb-shaped chains.

Setting $p=1$ in from (36) to (39), we obtain the Wiener index of stars:

$$\begin{aligned}W_{\text{Star}}&=\frac{1}{6}N_0\left(N_0^2-1\right)+\sum_{j=1}^{n_1}\frac{1}{6}N_{1,j}\left(N_{1,j}^2-1\right)\\&+\frac{1}{2}\sum_{j=1}^{n_1}N_{1,j}\Big[(N_0+2-m_1)(N_0+1-m_1)\\&+(m_1-1)(m_1+2)+N_0\left(N_{1,j}-1\right)\Big]\\&+\sum_{i=1}^{n_1-1}\sum_{j=1+i}^{n_1}N_{1,i}\,N_{1,j}\left(\frac{N_{1,i}+N_{1,j}}{2}+1\right).\end{aligned} \tag{41}$$

Likewise, setting $n_k=1$ in (36)–(39), we obtain that of combs:

$$\begin{aligned}W_{\text{Comb}}&=\frac{1}{6}N_0\left(N_0^2-1\right)+\sum_{k=1}^{p}\frac{1}{6}N_{k,1}\left(N_{k,1}^2-1\right)\\&+\frac{1}{2}\sum_{k=1}^{p}N_{k,1}\Big[(N_0+2-m_k)(N_0+1-m_k)\\&+(m_k-1)(m_k+2)+N_0\left(N_{k,1}-1\right)\Big]\\&+\sum_{k=1}^{p-1}\sum_{j=1+k}^{p}N_{l,1}\,N_{k,1}\left(\frac{N_{l,1}+N_{k,1}}{2}\right.\\&\left.+1+m_l-m_k\right).\end{aligned} \tag{42}$$

The topological method gives the results numerically identical with that calculated by usual statistical methods [5,99]. Advantages in the formulation based on the Wiener index are that the mean square radius of gyration or the g-factor not only can be evaluated in the homologous process but also can be expressed as a function of topological parameters such as the position of branches and the length of main or side chains. For example, let us consider a simple specifically branched chain with one linear side chain, the so-called Y-shaped chain. The Wiener index of the Y-shaped chain W_{Y} can be readily evaluated from (41) or (42). The g-factor of Y-shaped chain g_{Y} can be expressed as a function of the position of branch point or the length of branches.

$$g_{\text{Y}}=\frac{6N_{1,1}}{N\left(N^2-1\right)}\Big[(m_1-1)^2-(N_0-1)(m_1-1)\Big]+1, \tag{43}$$

where $N_{1,1}$ is the number of beads in the branch, N_0 is the number of beads in the main chain, $N=N_{1,1}+N_0$, and m_1 is the position of branched bead on the main chain. When $N_{1,1}$ and N_0 are fixed and discrete, m_1 is transformed into continuous m_1 because of the large N_0 and g becomes a quadratic function of the branch position m_1 in which g has the minimum at $m_1=N_0/2$. Likewise, when N and m_1 are fixed, g becomes another quadratic function of N_0 or $N_{1,1}$ and has a minimum at $N_0=(N+m_1)/2$ or $N_{1,1}=(N-m_1)/2$.

Block copolymers [126] have also been given much attention to offer the potential for obtaining materials that incorporate the properties of different homopolymers and for improving the compatibility of the immiscible polymer blends and the interfacial adhesion in polymeric composites. Recent techniques of living anionic copolymer-

ization [66] make it possible to produce various complicated nonlinear copolymers such as graft copolymers with the different chemical nature of the backbone [117,118] and the attached branches and the miktoarm (μ-star), or heteroarm star polymers consisting of chemically different arms [65]. The graph-theoretical expressions for linear block copolymers has been studied by Yang [136].

The nonlinear copolymer thus considered is a chain consisting of branch units with edge bond length b attached to the backbone chain with edge length a. The length of edge of backbone and branches is different so that the path number can be determined on the basis of the concept of Altenburg polynomial [1]. The path number of the nonlinear graph can be obtained

$$\Omega_{\mathrm{ch}} = \frac{1}{6} N_0 \left(N_0^2 - 1\right) a^2 , \tag{44}$$

$$\Omega_{\mathrm{br}} = \sum_{k=1}^{p} \left[\sum_{j=1}^{n_k} \frac{1}{6} N_{k,j} \left(N_{k,j}^2 - 1\right) b^2 \right] , \tag{45}$$

$$\Omega_{\mathrm{ch\text{-}br}} = \sum_{k=1}^{p} \sum_{j=1}^{n_k} \frac{1}{2} N_{k,j} \left[\left\{ \left(N_0 - m_k + 1\right)^2 + \left(m_k^2 - 1\right) \right\} a^2 + N_0 N_{k,i} b^2 \right] , \tag{46}$$

$$\Omega_{\mathrm{br\text{-}br}} = \sum_{k=1}^{p} \left[\sum_{i=1}^{n_k - 1} \sum_{j=1+i}^{n_k} N_{k,i} N_{k,j} \left(\frac{N_{k,i} + N_{k,j}}{2} + 1 \right) b^2 + a^2 \right] + \sum_{k=1}^{p-1} \sum_{l=1+k}^{p} \left[\sum_{i=1}^{n_l} \sum_{j=1}^{n_k} N_{l,i} N_{k,j} \times \left\{ \left(\frac{N_{l,i} + N_{k,j}}{2} \right) b^2 + \left(1 + m_l - m_k\right) a^2 \right\} \right] . \tag{47}$$

The total path number statistically weighted by a and b becomes $\Omega = \Omega_{\mathrm{ch}} + \Omega_{\mathrm{br}} + \Omega_{\mathrm{ch\text{-}br}} + \Omega_{\mathrm{br\text{-}br}}$ and the mean square radius of gyration for any nonlinear copolymers is given by $\langle s^2 \rangle = \Omega / N^2$. Of course (44)–(47) of Ω are reduced to (36)–(39) of the Wiener index W for branched homopolymers when $a = b = 1$.

Randomly Branched Chains

Any real polymeric material will have a distribution in chain lengths, number and position of branches, and so on, depending on polymerization and/or fractionation; the actual polymers may be a mixture of such isomers. For the proceeding to such actual polymers, Zimm–Stockmayer [139] investigated the ensemble of random mixture of structural isomers, in which it was assumed that the total number of beads (or molecular weight) and the number of subchains per chain are fixed but all possible arrangements of subchains of various lengths occur with equal frequency. Here the term "subchain" refers to a portion of the molecule between two adjacent branched beads or between adjacents pair of an end and a branched bead.

An example of a randomly branched chain is shown in Fig. 5a. Let n_k be the number of the beads within the kth subchain, then the total number of beads is given by $N = n_1 + n_2 + \ldots + n_p$ where p is the total number of subchains per molecule. Let us consider the g-factor of randomly branched chains with a random distribution of n_k which is defined such that all possible sets of the numbers, $(n_1, n_2, \ldots, n_p)$, occur with equal frequency under the restriction that N is fixed. Then the g-factor of the randomly branched chains can be expressed as a function of a set of $(n_1, n_2, \ldots, n_p)$. Consequently the g-factor of any branched chain can be determined by averaging the g value over all arrangements of n_k under the fixed N and p as follows

$$\bar{g} = \mathrm{Av}[g] = \frac{\int_0^N \mathrm{d}n_{p-1} \int_0^{N-n_{p-1}} \mathrm{d}n_{p-2} \ldots \int_0^{N-\sum_{k=1}^{p-1} n_k} \mathrm{d}n_1 \, g(n_1, n_2, \ldots, n_p)}{\int_0^N \mathrm{d}n_{p-1} \int_0^{N-n_{p-1}} \mathrm{d}n_{p-2} \ldots \int_0^{N-\sum_{k=1}^{p-1} n_k} \mathrm{d}n_1} . \tag{48}$$

According to Kataoka and Saito [83,135], the g-factor can be expressed using the total number of subchains p as

$$\bar{g} = \frac{6}{p\,(p+1)\,(p+2)} \left(p^2 + \sum_{(\alpha,\beta)} \nu_{\alpha\beta} \right) , \tag{49}$$

where $\nu_{\alpha\beta}$ represents the number of subchains lying between the specified subchains α and β, and $\sum_{(\alpha,\beta)}$ is the sum extending over all distinguishable pairs of subchains.

According to Nitta [96], the g-factor of any randomly branched chain can be estimated from the Wiener index ω of its reduced graph in which subchains are transformed into single edges of unit length and both junctions and ends into vertices of unit mass, as shown in Fig. 5b. Thus,

$$\bar{g} = \mathrm{Av}[g] = \mathrm{Av}\left[\frac{W}{W_{\mathrm{L}}} \right] = \frac{\omega}{\omega_{\mathrm{L}}} , \tag{50}$$

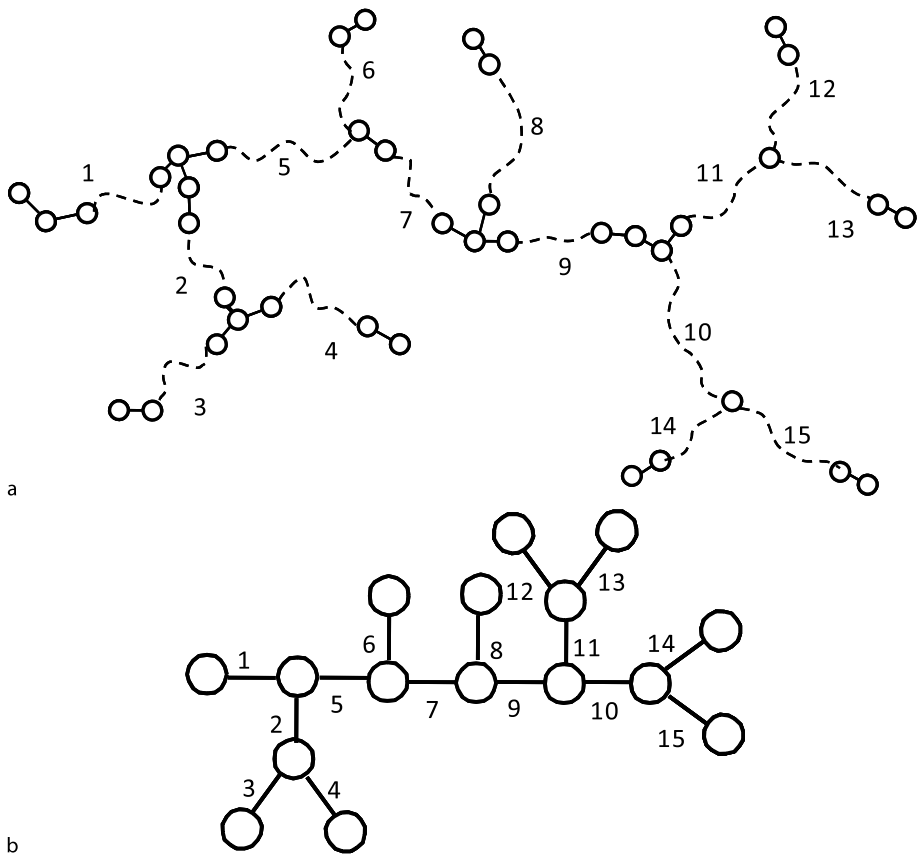

Polymers, Non-linearity in, Figure 5
A randomly branched molecule and its reduced molecular graph in which branched and end beads are transformed to uniform vertices and subchains are transformed to uniform edges

where ω_L is the Wiener index of a linear graph with the same number of edges which can be readily evaluated from (33) by setting $N = p + 1$. Thus,

$$\omega_L = \frac{p\,(p+1)\,(p+2)}{6}\,. \tag{51}$$

Moreover, since W_L may be considered to be constant for the averaging process in (48), the average Wiener index can be readily obtained as follows:

$$\bar{W} = \mathrm{Av}[W] = \mathrm{Av}[g]\,W_L = \frac{N\,(N+1)\,(N+2)}{p\,(p+1)\,(p+2)}\,\omega. \tag{52}$$

Consequently, the mean Wiener index of any randomly branched chain can be calculated by the total number of subchains p, the total number of beads N, and the Wiener index ω of the corresponding reduced graph where all subchains are transformed into single edges (see Fig. 5b). This indicates that the mathematical difficulty involved with the treatment of highly branched molecules is largely reduced by the use of the Wiener index. The reduced graph corresponds to the *proper graph*, which contains only two types of vertices-terminal and branched ones, as introduced by Bonchev et al. [10]. Specific cases of randomly branched chains are exemplified below.

In the case of "star" chains, the Wiener index can be evaluated from (41) by putting $N_{1,i} = 1$, $N_0 = 3$, $m_1 = 2$ and $n_1 = q - 2$:

$$\omega_S = q^2\,. \tag{53}$$

In the case of the "comb" type chains, we obtain

$$\omega_C = \frac{1}{12}\left(q^3 + 9q^2 - q + 3\right) \tag{54}$$

by substituting the relations that $N_{k,1} = 1$, $m_k = k + 1$, $p = (q-1)/2$, and $N_0 = (q+3)/2$ into (42). Of course, these equations are identical with the results derived from Kurata and Fukatsu's statistical treatments [83]. Gutman et al. [62,63] presented the formula for the Wiener index of

regular dendrimer graphs of three and four degrees. The more generalized Wiener index for a regular dendrimer graph of degree f is given by

$$\omega_{\mathrm{D}}(n,f) = \frac{f}{(f-2)^3}\left[\left\{nf^2 - 2(1+n)f + 1\right\}(f-1)^{2n} + 2f(f-1)^n - 1\right], \quad (55)$$

where n is the generation number.

Intrinsic Viscosity and Hydrodynamic Radius

The most precise and direct experimental method for characterizing chain architecture is through the measurement of dilute solution properties. The most useful measure of branching is the Zimm–Stockmayer ratio g of the radius of gyration of the branched polymer to that of the linear polymer at the same molecular weight as shown in (32). Other commonly employed branching parameters are the ratio of the intrinsic viscosity g' [138] and the ratio of hydrodynamic radius h [120] for branched polymers to the linear ones, respectively. The non-free draining intrinsic viscosity of suspended spheres with hydrodynamic radius was given by Einstein as

$$[\eta]_{\mathrm{ND}} = \frac{5}{2}\left(\frac{N_{\mathrm{A}}}{M}\right)\frac{4}{3}\pi\, r_{\mathrm{H}}^3\,, \quad (56)$$

where M is the molecular weight and N_{A} is Avogadro's number.

According to (21), the mean reciprocal hydrodynamic radius r_{H}^{-1} can be calculated from the mean reciprocal of $\mathbf{r}_{ij}$ given by [59,120]

$$\left\langle |\mathbf{r}_{ij}|^{-1}\right\rangle = \left(\frac{6}{\pi}\right)^{1/2} d_{ij}^{-1/2}\, b^{-1}\,. \quad (57)$$

Using the generalized Harary index (16) with $\varepsilon = 1/2$, one obtains the reciprocal hydrodynamic radius given by

$$\left\langle r_{\mathrm{H}}^{-1}\right\rangle_0 = \frac{1}{bN^2}\left(\frac{6}{\pi}\right)^{1/2} H_{1/2} \quad (58)$$

and its branching factor can then be written as

$$h = \frac{\langle r_{\mathrm{H}}\rangle_0^{-1}}{\langle r_{\mathrm{H}}\rangle_0^{\mathrm{L}-1}} = \frac{H_{1/2}^{\mathrm{L}}}{H_{1/2}}\,, \quad (59)$$

where the superscript L represents the linear chain. In the non-draining condition, we have $g' = h^3$. Consequently the hydrodynamic radius and the branching parameter h or g' can be evaluated from the distance matrix.

Assuming that the Stokes radius r_{H} is proportional to the square root of the mean-square radius of gyration with proportional constant χ

$$r_{\mathrm{H}} = \chi\left(6\left\langle s^2\right\rangle\right)^{1/2} = \chi\left(\frac{b}{N}\right)(6W)^{1/2}\,. \quad (60)$$

Consequently the intrinsic viscosity in the non-draining condition is also related to the Wiener index. Consequently we have [44]

$$[\eta]_{\mathrm{ND}} = \frac{5}{2}\,\Phi\,\frac{\left(6\left\langle s^2\right\rangle\right)^{3/2}}{N} = \frac{5}{2}\,\Phi\,\frac{b^3}{N^4}\,(6W)^{3/2}\,, \quad (61)$$

where $\Phi = (N_{\mathrm{A}}/M_0)\,4/3\,\pi\chi^3$ and M_0 is the molar mass of the unit, i. e. $M_0 = M/N$.

Chain Dynamics of Nonlinear Polymers

As mentioned before, Gaussian chains have played a central role on the statistics and dynamics of flexible chain molecules with and without branching [8,53,78,100]. According to the central limit theorem in statistical physics [59], the random-flight statistics of a flexible polymer can be described by a Gaussian chain which is mathematically simpler to handle [47]. Rouse [110] and Bueche [17] demonstrated that the dynamics of dilute solutions of linear polymers can be characterized by considering a Gaussian chain suspended in a flowing viscous liquid. Subsequently, the extension of this model to any branched molecule was made by Ham [67]. In fact, the Rouse model provides a good description for semi-dilute solutions and for melts below the entanglement limit rather than dilute solutions [38]. The mathematical representation of graph theory is helpful to generalize the statistics and dynamics of Gaussian chains to include any type of branching [36,45,46,136]. In this section, we present a graph-theoretical method for calculating the relaxation spectra of flexible chain molecules with any type of branching. The significance of this approach lies in providing the relationship between the statistics and the dynamics of chain molecules being reformulated into a more convenient algebraic form.

Rouse–Ham Theory

Branched Gaussian chains containing no loops or circles contains N beads and $N-1$ segments acting as Hookean springs with spring constant of $3k_BT/b^2$ as shown in (29). The laws, which govern the behavior of linear flexible chain, are assumed to hold for nonlinear flexible polymers. Considering a Gaussian chain suspended in a flow liquid,

one finds the motion equation given by [101]

$$-\zeta_0 \dot{\mathbf{r}} = \frac{3k_B T}{b^2} \mathbf{Z} \cdot \mathbf{r} , \tag{62}$$

where $\mathbf{r}$ is a $3 \times N$ matrix whose rows contain the dimensional component of the N position vectors of beads, $\dot{\mathbf{r}}$ is the time derivative of $\mathbf{r}$, ζ_0 is the friction constant of the beads, and $\mathbf{Z}$ is the $N \times N$ connectivity matrix which is called the Zimm matrix in polymer physics [137] and is identical to the Laplacian matrix $\mathbf{L} (= \mathbf{V} - \mathbf{A})$ in graph theory. According to Rouse theory [110], one can rewrite (62) in terms of bond vectors $\mathbf{b}$ as

$$-\zeta_0 \dot{\mathbf{b}} = \frac{3k_B T}{b^2} \mathbf{R} \cdot \mathbf{b} , \tag{63}$$

where $\mathbf{b}$ is a $3 \times (N-1)$ matrix whose rows contain the dimensional component of the $N-1$ bond vectors. The matrix $\mathbf{R}$ is the $(N-1) \times (N-1)$ connectivity matrix which is called the Rouse matrix [110]. Comparing the characteristic polynomials of $\mathbf{Z}$ and $\mathbf{R}$, we can find the following relation:

$$\Phi(\mathbf{Z}; \lambda) = \lambda \, \Phi(\mathbf{R}; \lambda) . \tag{64}$$

The eigenvalues of $\mathbf{Z}$ contain one zero eigenvalue and, hence, $\mathbf{Z}$ do not possess an ordinary inverse. (64) shows that the non-zero eigenvalues of $\mathbf{Z}$ are identical with those of $\mathbf{R}$. The springs and beads are assigned in any arbitrary fashion. The eigenvalues of $\mathbf{Z}$ and $\mathbf{R}$ are, however, independent of how their elements are numbered [46,51,94]. Also, this relation (64) corresponds to the relation (11).

The zero eigenvalue of $\mathbf{Z}$ represents the mode of chain translation [61]. Each eigenvalue λ_i $(i = 1, 2, \ldots, N-1)$ of $\mathbf{R}$ or non-zero eigenvalue of $\mathbf{Z}$ is associated with the relaxation times τ_i of the ith mode for dynamic molecular motions [46,51]:

$$2\tau_i = \frac{\zeta_0 \, b^2}{3k_B T} \lambda_i^{-1} . \tag{65}$$

According to the theory of linear viscoelasticity, the relaxation time spectrum, $H(\tau)$, is given by

$$H(\tau) = \frac{c k_B T}{N} \sum_{i=1}^{N-1} \delta(\ln \tau - \ln \tau_i) , \tag{66}$$

where c is the concentration of beads per unit volume and $\delta(x)$ is a Dirac delta function. Various rheological functions describing the bulk properties of polymers can be determined using the relaxation spectrum $H(\tau)$.

The Zimm matrix $\mathbf{Z}$ or the Rouse matrix $\mathbf{R}$ can also be constructed in a different manner by making use of the incident matrix of a digraph D. Forsman [46] showed that $\mathbf{Z}$ and $\mathbf{R}$ are given by the incidence matrix $\mathbf{B}$ of D:

$$\mathbf{Z} = \mathbf{B}\mathbf{B}^\top \quad \text{and} \quad \mathbf{R} = \mathbf{B}^\top \mathbf{B} . \tag{67}$$

It should be here noted that all elements of a connectivity matrix $\mathbf{K}$ formed by $\mathbf{C}^\top \mathbf{C}$ are the absolute values of those of $\mathbf{R}$. Combination of the relations (10), (11), (64), $\mathbf{Z} = \mathbf{L}$ gives

$$\Phi(\mathbf{R}; \lambda) = \Phi(\mathbf{K}; \lambda) . \tag{68}$$

Furthermore from the Kirchhoff's relation (4), we have

$$\Phi(\mathbf{R}; \lambda) = \Phi(\mathbf{A}_L; \lambda - 2) . \tag{69}$$

Consequently, the eigenvalues λ_i of $\mathbf{R}$ can be calculated from the eigenvalues μ_i of $\mathbf{A}_L$ as follows:

$$\lambda_i = \mu_i + 2 . \tag{70}$$

Remembering (65), we have

$$\tau_i = \frac{\zeta_0 b^2}{6k_B T} (\mu_i + 2)^{-1} . \tag{71}$$

It is interesting to note that the relaxation spectrum of a chain molecule can be determined entirely by the set of eigenvalues of the adjacency matrix $\mathbf{A}_L$ of its line graph.

According to Forsman [46], the square radius of gyration is given by $s^2 = N^{-1} \mathbf{b}^\top \cdot \mathbf{F} \cdot \mathbf{b}$ where $\mathbf{F}$ is called the Kramers matrix and it is equal to the inverse of $\mathbf{R}$. When the chains obey random-flight statistics, the mean square of radius of gyration can be related to the reciprocal of eigenvalues of $\mathbf{R}$ and we have

$$\langle s^2 \rangle_0 = \frac{b^2}{N} \operatorname{Tr}\left(\mathbf{R}^{-1}\right) = \frac{b^2}{N} \sum_{i=1}^{N-1} \frac{1}{\lambda_i} = \frac{b^2}{N} \sum_{i=1}^{N-1} \frac{1}{\mu_i + 2} , \tag{72}$$

where Tr denotes the trace of a matrix. The mean radius of gyration is related to the sum of the reciprocal eigenvalues of Rouse matrix so that it largely depends on the minimum eigenvalue of $\mathbf{R}$ or the second smallest one of $\mathbf{L}$. This corresponds to Mohar's suggestion that the second smallest one of $\mathbf{L}$ is related to the diameter and mean distance in a graph [94].

Comparing (72) and (31) gives

$$W = N \operatorname{Tr}\left(\mathbf{R}^{-1}\right) = N \sum_{i=1}^{N-1} \frac{1}{\lambda_i} = N \sum_{i=1}^{N-1} \frac{1}{\mu_i + 2} . \tag{73}$$

Moreover, a combination of (65) and (73) gives a relation between the Wiener index and relaxation times:

$$W = \frac{6Nk_{\mathrm{B}}T}{\zeta b^2} \sum_{i=1}^{N-1} \tau_i \,. \tag{74}$$

This may be a physical meaning of the Wiener index. The steady-state viscosity in a dilute solution can be obtained from the relaxation time spectrum (66):

$$\eta = \eta_{\mathrm{s}} + c\,\frac{N_{\mathrm{A}}}{M}\,k_{\mathrm{B}}T \sum_{i=1}^{N-1} \tau_i \,. \tag{75}$$

Consequently, the intrinsic viscosity in the freely draining condition can be written as

$$[\eta]_{\mathrm{FD}} = \frac{N_{\mathrm{A}}}{M\eta_{\mathrm{s}}}\,k_{\mathrm{B}}T \sum_{i=1}^{N-1} \tau_i \,, \tag{76}$$

which turns out to be proportional to the mean square of the radius of gyration:

$$[\eta]_{\mathrm{FD}} = \frac{N_{\mathrm{A}}\zeta}{6M_0\eta_{\mathrm{s}}}\,\langle s^2\rangle = \frac{N_{\mathrm{A}}\zeta b^2}{6M_0\eta_{\mathrm{s}}N^2}\,W \,, \tag{77}$$

where η_s is the viscosity of pure solvent. It can be shown that the intrinsic viscosity in freely draining conditions is proportional to W/N^2 for branched chains and to the first power of N for linear chains. (77) may be referred to as "Staudinger's relation". This suggests that the Wiener index of any polymer molecule can be evaluated from the steady-flow viscosity of dilute solution in the freely draining condition.

The non free-draining intrinsic viscosity η_{ND} as shown in (61) is proportional to $W^{3/2}/N^4$. Zimm and Kilb [138] introduced a branching parameter g' defined as the ratio of $[\eta]$ of a branched polymer to that of a linear polymer in the Θ condition. Therefore, we have $g' = g^{1.0}$ for the freely draining condition and $g' = g^{1.5}$ for the non-draining condition. For various branched polymers, however, measured exponents fall between 0.5 and 2.0, and they are strongly dependent on solvents and molecular weight [82,125].

Sheridan et al. [114] examined the relation between the intrinsic viscosity and the Wiener index for hyperbranched polymers. According to their computer simulations, they found the relation $[\eta] \sim W^a N^b$ where $a = 1.0$ and $b = -2.2$ which is similar to the result of (77) in the freely draining condition.

High Moments of Relaxation Time and Radius of Gyration

The potential energy of the chain molecule given by (29) can be rewritten as

$$V = \frac{1}{2}\,k\,\mathrm{Tr}\left[\mathbf{r}\cdot\mathbf{r}^{\top}\right]. \tag{78}$$

Fixman [40] derived that the distribution function of the square radius of gyration s^2 as follows:

$$P(s^2) \propto \int \delta\left(s^2 - N^{-1}\,\mathrm{Tr}\left[\mathbf{r}\cdot\mathbf{r}^{\top}\right]\right) \mathrm{e}^{-V/k_{\mathrm{B}}T} \,. \tag{79}$$

In order to obtain the higher moments of radius of gyration $\langle s^{2n}\rangle$, we introduces a Laplace transform of $P(s^2)$ with respect to s^2 according to Eichinger's treatment [35]. Then we obtain

$$\mathcal{L}\left[P(s^2)\right] = \int_0^{\infty} \mathrm{e}^{-zs^2}\,P(s^2)\,\mathrm{d}s^2 = \varphi(z)^{-3/2} \tag{80}$$

and

$$\varphi(z) = \mathrm{Det}\,|\mathbf{E} + \gamma z\,\mathbf{R}^{-1}| \tag{81}$$

where $\gamma = 2b^2/(3N)$. The Laplace transform of $P(s^2)$, i. e., the generating function of $P(s^2)$, provides the average $\langle \mathrm{e}^{-zs^2}\rangle$ and, therefore, the average of the powers of s^2 can be computed by making use of the expansion in the form of a power series in z. Thus we have

$$\langle s^{2n}\rangle_0 = (-1)^n\,\frac{\partial^n}{\partial z^n}\,\varphi(z)^{-3/2}\bigg|_{z\to 0} \,. \tag{82}$$

Since we have $\mathrm{Det}\,|\mathbf{R}| = N$ for any tree graph, (81) can be rewritten as

$$\varphi(z) = \frac{1}{N}\,\mathrm{Det}\,|\mathbf{R} + \gamma z\,\mathbf{E}| = \frac{1}{N}\,\Phi(\mathbf{R};-\gamma z) \,. \tag{83}$$

Using (4) and (68), we have

$$\varphi(z) = \frac{1}{N}\,\mathrm{Det}\,|\mathbf{A}_{\mathrm{L}} + (\gamma z + 2)\,\mathbf{E}| = \frac{1}{N}\,\Phi(\mathbf{A}_{\mathrm{L}};-\gamma z - 2) \,. \tag{84}$$

It was shown that the characteristic polynomial of the line graph gives the general equation for calculating the radius of gyration of a Gaussian chain with any type of branching [97].

According to Eichinger's mathematical treatments [35], we showed the relation between eigenvalues of the matrix $\mathbf{R}$ and the moments of the radius of gyration: ex-

pansion of (83) and comparing the coefficients of z give

$$\langle s^4\rangle_0 - \langle s^2\rangle_0^2 = \frac{2b^4}{3N^2}\,\mathrm{Tr}\left(\mathbf{R}^{-1}\right)^2 = \frac{2b^4}{3N^2}\sum_{i=1}^{N-1}\frac{1}{\lambda_i^2}$$
$$= \frac{2b^4}{3N^2}\sum_{i=1}^{N-1}\frac{1}{(\mu_i+2)^2}\,, \tag{85}$$
$$\langle s^6\rangle_0 - 3\langle s^4\rangle_0\langle s^2\rangle_0 + 2\langle s^2\rangle_0^3 = \frac{8b^6}{9N^3}\,\mathrm{Tr}\left(\mathbf{R}^{-1}\right)^3$$
$$= \frac{8b^6}{9N^3}\sum_{i=1}^{N-1}\frac{1}{\lambda_i^3}$$
$$= \frac{8b^6}{9N^3}\sum_{i=1}^{N-1}\frac{1}{(\mu_i+2)^3}\,. \tag{86}$$

The above formulas also yield the relation between viscoelasticity and chain dimensions through the Rouse matrix.

The significance of the present graph-theoretical approach is to provide the general equations for the relaxation spectrum and the radius of gyration of any tree-like chain. In particular, it is noteworthy that the mathematical method has the potential to provide an algorithmic method of calculating high-order moments of the radius of gyration and the relaxation time for any tree-like chain. These values can hardly be calculated from the usual statistical methods because of a great difficulty in enumerating the distribution function.

The characteristic function (81) can be rewritten in terms of the $N-1$ eigenvalues of the Rouse matrix as [35]

$$\varphi(z) = \prod_{i=1}^{N-1}\left(1+\frac{\gamma}{\lambda_i}z\right)\,. \tag{87}$$

Making use of the relation

$$\ln\varphi(z)$$
$$= \sum_{i=1}^{N-1}\ln\left(1+\frac{\gamma}{\lambda_i}z\right)$$
$$= \sum_{i=1}^{N-1}\left[\frac{\gamma}{\lambda_i}z - \frac{1}{2}\left(\frac{\gamma}{\lambda_i}\right)^2 z^2 + \frac{1}{3}\left(\frac{\gamma}{\lambda_i}\right)^3 z^3 - \ldots\right]\,, \tag{88}$$

the sums of the reciprocal powers of the eigenvalues are easily determined from the following equation:

$$(n-1)!\sum_{i=1}^{N-1}\left(\frac{\gamma}{\lambda_i}\right)^n = (-1)^{n-1}\frac{\partial^n}{\partial z^n}\ln\varphi(z)\bigg|_{z\to 0}\,. \tag{89}$$

Using (65), we have

$$\sum_{i=1}^{N-1}\tau_i^n = \frac{(-1)^{n-1}}{(n-1)!}\left(\frac{N\zeta_0}{4k_\mathrm{B}T}\right)^n\frac{\partial^n}{\partial z^n}\ln\varphi(z)\bigg|_{z\to 0}\,. \tag{90}$$

The coefficient of z in the characteristic function (81) can be related to a topological index such as the Wiener index of the molecular graph G, i. e., a total sum of the elements of its distance matrix, which is potentially useful in the correlation of molecular topology to thermodynamic properties for alkanes. Therefore, any high-order coefficient in (81) for a tree-like graph, or any coefficient of the characteristic polynomial for its line graph, has the potential to be a new topological index. From the coefficients, we can define the high-ordered Wiener indices as

$$W_k \equiv (k-1)!\left(\frac{2}{3}\right)^{k-1}N^k\sum_{i=1}^{N-1}\left(\frac{1}{\lambda_i}\right)^k\,. \tag{91}$$

One can easily see that the first index W_1 becomes the original W. For example, the second and third Wiener indices, W_2 and W_3, can be used for calculating the higher moments of radius of gyration and mechanical relaxation spectrum of branched chains. The second and third Wiener indices provide the fourth and sixth moments of radius of gyration.

$$\langle s^4\rangle_0 = \left(\frac{b}{N}\right)^4\left(W_2 + W_1^2\right)\,, \tag{92}$$

$$\langle s^6\rangle_0 = \left(\frac{b}{N}\right)^6\left(W_3 + 3W_2W_1 + W_1^3\right)\,. \tag{93}$$

Applications

In this section, the relationship between the characteristic polynomial and the mechanical relaxation spectra is examined. Then the zero-shear-rate viscosity η_0 and the steady-state compliance J_e^0 [38], a measure of the elastic energy stored under steady flow, are

$$\eta_0 = \frac{ck_\mathrm{B}T}{N}\sum_{i=1}^{N-1}\tau_i\,,$$
$$J_\mathrm{e}^0 = \frac{N}{ck_\mathrm{B}T}\left(\frac{\sum_{i=1}^{N-1}\tau_i^2}{\left(\sum_{i=1}^{N-1}\tau_i\right)^2}\right) \tag{94}$$

so that these basic rheological functions can be rewritten using the Wiener indices as

$$\eta_0 = \frac{c\zeta_0}{6}\left(\frac{b}{N}\right)^2 W_1\,,$$
$$J_\mathrm{e}^0 = \frac{3N}{2ck_\mathrm{B}T}\left(\frac{W_2}{W_1^2}\right)\,. \tag{95}$$

Thus, the rheological properties can be calculated from the distance matrix of the corresponding tree-graph. This is a very important result because we can realize the effects of the branching feature such as the branch length, branch number, and branch position on these rheological behaviors. In addition, the intrinsic dynamic moduli of storage $[G']$ and loss $[G'']$ can be obtained by using the following relations [129]:

$$[G'] = \sum_{i=1}^{N-1} \frac{\omega^2 \tau_i^2}{1 + \omega^2 \tau_i^2}, \quad [G''] = \sum_{i=1}^{N-1} \frac{\omega \tau_i}{1 + \omega^2 \tau_i^2}, \quad (96)$$

where the ω is the frequency of strain oscillation.

It is possible to obtain the relaxation spectra through (71) from the eigenvalues of the adjacency matrix of its line graph. Calculations of relaxation spectra were exemplified for a linear, Y-shape star, comb, and dendrimer with a fixed vertex number of $N = 94$ in Fig. 6. The corresponding frequency dependences of $[G']$ and $[G'']$ are plotted in Fig. 7 with the reduced variables. As seen in these figures, the low frequency slopes of $[G']$ and $[G'']$ are 2 and 1 on the logarithmic plot as expected for any linear polymer liquid but the higher frequency slope of $[G']$ for the dendrimer was found to be larger than the 1/2 of the linear chain. It is evident that the comb and Y-star are almost intermediate between the linear and the dendrimer. In the flow region, $\omega \to 0$, $[G']$ and $[G'']$ is going to $\omega^2 \tau_R$ and $\omega \tau_R$ as is obvious from (96) where τ_R is the highest relaxation time or Rouse time. The Rouse time τ_R can be estimated from the smallest eigenvalue of $\mathbf{K}$ or $\mathbf{R}$ or the second smallest one of $\mathbf{L}^+$ or $\mathbf{Z}$. The order of τ_R was in linear > comb > Y-star > dendrimer.

Future Directions

The formulation of conformational statistics and dynamics due to the graph-theoretical expressions is based on the fact that the random-flight model and/or Gaussian chain model may be characterized by the so-called Markov nature such that the mean-square radius is proportional to the number of bonds in the chain in the unperturbed state without excluded volume. However, the Markov na-

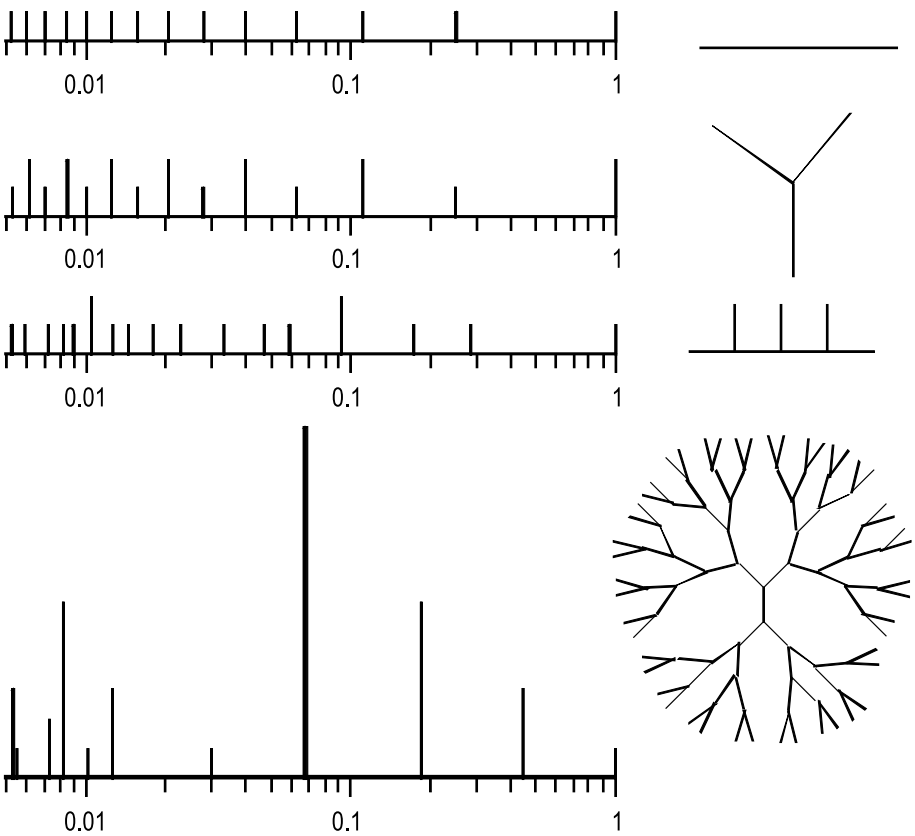

Polymers, Non-linearity in, Figure 6
Relaxation spectra of linear, Y-star, comb, and dendrimer chains with a fixed vertex number of 94

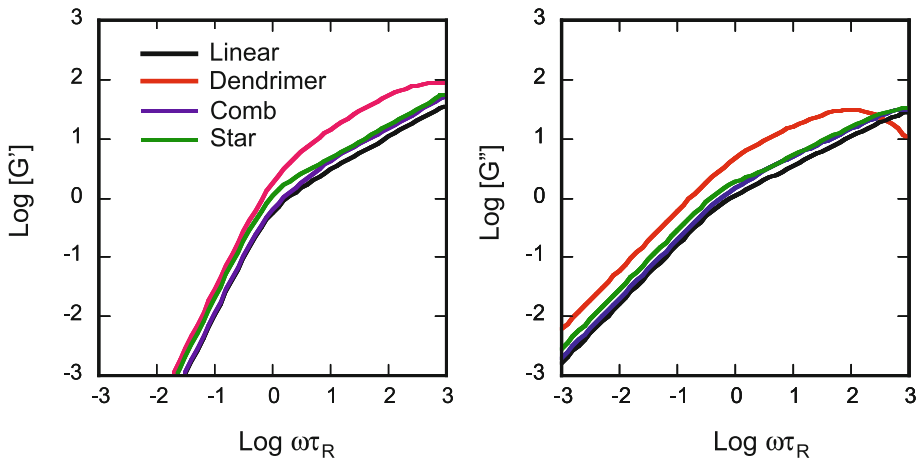

Polymers, Non-linearity in, Figure 7

Logarithmic plot of reduced intrinsic moduli [G'] and [G''] against $\omega\tau_R$ for linear, comb, star, and dendrimer chains with $N = 94$, where τ_R is the maximum relaxation time in Rouse dynamics

ture breaks down for real short chains and semiflexible or stiff chains [135]. This non-Markov nature arises from a sort of stiffness or static rigidity, as introduced by the constraints on the internal degrees of freedom such as fixed bond lengths, fixed bond angles, and hindered internal rotations. In other words, the Wiener index reflecting the radius of gyration based on the Markov nature can be considered to be an appropriate measure of molecular size and viscosity of flexible polymers rather than small molecules. For the study of such problems, the graph-theoretical studies for nonlinear non-Markov chains will play a central role in presenting the more useful devices and descriptors not only for semi-flexible chains and liquid crystalline polymers [98,135] but also for small molecules. The modification on the original bead-spring (Gaussian) chain or Markov chain model for nonlinear chains is to include the 'excluded volume effect' [129] because a substantial increase in the density of structural units close to branching point causes a stronger excluded-volume effects in comparison with other regions of the chain [25]. Regarding such theoretical difficulties inherent to branched structures, simulation work as well as the statistic-based methods such as renormalizing group theory [113] and self-avoiding random work (SAW) procedure [77] will be a very powerful tool in the study of this type of branched polymers [48]. Furthermore, the combination with the graph-theoretical formulation will make it possible to find key parameters expressing the nature and degree of branching, which inevitably makes the mathematics more complicated.

The chain entanglements play an important role in the melt viscosity and rheological properties [54]. The extend of the theoretical framework for entanglements presented by de Gennes [27,28] and Doi–Edwards [33] to branched polymers could be important for understanding the rheological properties of various branched polymers. Bonchev et al. [16] propose the topological descriptors of coarse-graining polymer graphs obtained by regarding the single edge as entanglement length. In particular the discrete mathematics such as graph theory and knot theory [121] seems to be important in the deeply understanding of the interplay between the entanglement and branching effects.

In addition, the effects of molecular weight distribution will be also the inevitable problems in industrial polymer products [4,92]. In random branching and hyperbranching processes the polydispersity index increases with the molecular weight [18,119]. The shape and width of the molar mass distribution curve remain extremely important also for branched polymers and markedly affect the melt properties of the polymers [86,87]. The concept of "*forest*" in graph theory will be appropriate to this problem.

Throughout this article, we have demonstrated that the graph-theoretical approach provides a topological insight into the nonlinearity in polymer architecture. This article deals with only the flexible polymers containing no loops nor rings are treated because of the lack of sufficient data on the statistics and dynamics of nonlinear chains with loops and rings. Thus, the problems of the dynamics and statistics of a tree-shaped molecule, which are very important for practical applications, were found to be completely reduced to the problem of the eigen-polynominal of the chain graph. This suggests that various ideas and concepts thus obtained from graph theory can be applied directly to the topological analysis for architecture in nonlinear polymers. For example, advantages in the formulation due

to the topological invariants are that the conformational statistics and dynamics of any branched polymer not only can be simply and universally evaluated, but also can be expressed as a function of topological parameters such as the position of branches and the length of main or side chains.

In conclusion, polymer chemistry would need a sense of the *mathematical chemistry* defined by Trinajstić–Gutman [128]: "*Mathematical chemistry is part of theoretical chemistry which is concerned with applications of mathematical methods to the chemical problems.*". The topological descriptor widely used in polymer chemistry is the molecular weight or the degree of polymerization which seems to be an index with low discriminating power and appropriate only to linear chain molecules. The author believes that the topological sense resulting from discrete mathematics such as graph theory provides powerful devises for leading to the quantitative structure-property and structure-activity relationships in the respectable form.

Bibliography

Primary Literature

1. Altenburg VK (1960) Zur Berechnung des Radius verzweigter Moleküle. Kolloid-Z 178:112–119
2. Baesley JK (1953) The molecular structure of polyethylene. IV. Kinetic calculations of the effect of branching on molecular weight distribution. J Am Chem Soc 75:6123–6127
3. Balakrishnan R (2004) The energy of a graph. Linear Algebr Appl 387:287–295
4. Berger L, Meissner J (1992) Linear viscoelasticity, simple and planar melt extension of linear polybutadienes with bimodal molar mass distributions. Rheol Acta 31:63–74
5. Berry GC, Orofino TA (1964) Branched polymers. III. Dimensions of chains with small excluded volume. J Chem Phys 40:1614–1621
6. Bicerano J (1989) Molecular level calculations of the structures and properties of non-crystalline polymers. Dekker, New York
7. Billmeyer FW (1953) The molecular structure of polyethylene. III. Determination of long chain branching. J Am Chem Soc 75:6118–6122
8. Biswas P, Kant R, Blumen A (2000) Polymer dynamics and topology: Extension of stars and dendrimers in external fields. Macromol Theor Simul 9:56–67
9. Bonchev D, Markel E, Dekmezian AH (2001) Topological analysis of long-chain branching patterns in polyolefins reciprocal distance matrix. J Chem Inf Comput Sci 41:1274–1285
10. Bonchev D, Markel EJ, Dekmezian AH (2002) Long chain branch polymer chain dimensions: Application of topology to the Zimm–Stockmayer model. Polymer 43:203–222
11. Bonchev D, Mekenyan O (1980) Topological approach to the calculation of the π-electron energy and energy gap of infinite conjugated polymers. Z Naturforsch 35a:739–747
12. Bonchev D, Mekenyan O, Polansky OE (1981) Topological approach to the predicting of the electron energy characteristics of conjugated infinite polymers. II. PPP-calculations. Z Naturforsch 36a:643–646
13. Bonchev D, Mekenyan O, Polansky OE (1981) Topological approach to the predicting of the electron energy characteristics of conjugated infinite polymers. III. The influence of some structural modifications of polymers. Z Naturforsch 36a: 647–650
14. Bonchev D, Mekenyan O, Protić G, Tranajstić N (1979) Application of topological indices to gas chromatographic data: Calculation of the retention indices of isomeric alkylbenzenes. J Chromatogr 176:149–156
15. Bonchev D, Trinajstic N (1977) Information theory, distance matrix, and molecular branching. J Chem Phys 67:4517–4533
16. Bonvhev D, Dekmezian AH, Markel E, Faldi A (2003) Topology-rheology regression models for monodisperse linear and branched polyethylenes. J Appl Polym Sci 90:2648–265
17. Bueche F (1954) The viscoelastic properties of plastics. J Chem Phys 22:603–609
18. Burchard W (1972) Angular distribution of Rayleigh scattering from branched polycondensates. Amylopectin and Glycogen types. Macromol 5:604–610
19. Burchard W (1999) Solution properties of branched macromolecules. Adv Polym Sci 143:113–194
20. Casassa EF, Tagami Y (1969) An equilibrium theory for exclusion chromatography of branched and linear polymer chains. Macromol 2:14–26
21. Cayley A (1874) On the mathematical theory of isomers. Phylos Mag 67:444–446
22. Chompff AJ (1970) Normal modes of branched polymers. I. Simple ring and star-shaped molecules. J Chem Phys 53:1566–1576
23. Chompff AJ (1970) Normal modes of branched polymers. II. Complex branched molecules and ring systems. J Chem Phys 53:1577–1584
24. Cvetković D (2005) Signless Laplacian and line graph. Bull Acad Serbe Sci Arts Cl Sci Math Natur Sci Math 31:85–92
25. Daout M, Cotton JP (1982) Star shaped polymers: A model for the conformation and its concentration dependence. J Phys 43:531–538
26. Dawkins JV, Maddock JW, Coupe D (1970) Gel-permeation chromatography: Examination of universalcalibration procedures for Polydimethylsiloxane in a poor solvent. J Polym Sci A-2 8:1803–1821
27. De Gennes PG (1971) Reptation of a polymer chain in the presence of fixed obstacles. J Chem Phys 55:572–579
28. De Gennes PG (1975) Reptation of stars. J Phys 36:1199–1203
29. De Gennes PG (1979) Scaling concepts in polymer physics. Cornell Univ, New York
30. Debye P (1946) The intrinsic viscosity of polymer solutions. J Chem Phys 14:636–639
31. Dobson GR, Gordon M (1964) Configurational statistics of highly branched polymer systems. J Chem Phys 41: 2389–2398
32. Doi M (1974) Relaxation spectra of nonlinear polymers. Polym J 6:108–120
33. Doi M, Edwards SF (1978) Dynamics of concentrated polymer systems. J Chem Soc Faraday Trans 2:1789–1832
34. Ebrahimi KG, Takahashi M, Arai O, Masuda T (1995) Effects of molecular weight distribution on dynamic viscoelasticity and biaxial extensinal flow behavior of polystyrene melts. J Rheol 39:1385–1397

35. Eichinger BE (1980) Configuration statistics of Gaussian molecules. Macromol 13:1–11
36. Eichinger BE (1976) Molecules as graphs. J Polym Sci Symp 54:127–134
37. Estrada E (1995) Edge adjacency relationships and a novel topological index related to molecular volume. J Chem Inf Comput Sci 35:31–33
38. Ferry JD (1980) Viscoelastic properties of polymers, 3rd edn. Wiley, New York
39. Fischer EW, Hahn K, Kugler J, Struth U, Born R (1984) An estimation of the number of tie molecules in semicrystalline polymers by means of neutron scattering. J Polym Sci 22:1491–1513
40. Fixman M (1962) Radius of gyration of polymer chains. J Chem Phys 36:306–318
41. Flory PJ (1937) Mechanism of vinyl polymerizations. J Am Chem Soc 59:236–241
42. Flory PJ (1953) Principles of polymer chemistry. Cornell, New York
43. Flory PJ (1969) Statistical mechanics of chain molecules. Hanser, New York
44. Flory PJ, Fox TG (1954) Treatment of intrinsic viscosities. J Am Chem Soc 73:1904–1908
45. Forsman WC (1968) Matrix methods for determining the dimensions of branched random-flight chains. Macromol 1:343–347
46. Forsman WC (1976) Graph theory and the statistics and dynamics of polymer chains. J Chem Phys 65:4111–4115
47. Freed KF (1972) Functional integrals and polymer statistics. Adv Chem Phys 22:1–128
48. Freire JJ (1999) Conformational properties of branched polymers: Theory and simulations. Adv Polym Sci 143:35–112
49. Fréchet JMJ (1994) Functional polymers and dendrimers: Reactivity, molecular architecture, and interfacial energy. Science 263:1710–1715
50. Fujimoto T, Narukawa H, Nagasawa M (1970) Viscoelastic properties of comb-shaped polystyrenes. Macromol 3:57–65
51. Gordon M (1979) From Riemann's metric to the graph metric, or applying Occam's razor to entanglements. Polymer 20:1349–1356
52. Gottlieb M, Bird BR (1976) A molecular dynamics calculation to confirm the incorrectness of the random-walk distribution for describing the Kramers freely jointed bead-rod chain. J Chem Phys 65:2467–2468
53. Graessley WW (1975) Statistical mechanics of random coil networks. Macromol 8:186–190
54. Graessley WW (1976) The entanglement concept in polymer rheology. Adv Polym Sci 16:1–179
55. Graessley WW, Masuda T, Roovers JEL, Hadjichristidis N (1976) Rheological properties of linear and branched polyisoprene. Macromol 9:127–141
56. Graovac A, Gutman I, John PE, Vidović, Vlah I (2001) On statistics of graph energy. Z Naturforsch 56a:307–311
57. Graovac A, Gutman I, Trinajistić N, Bonchev D (1972) Graph theory and molecular orbitals. Application of Sachs theorem. Theoret Chim Acta 26:67–78
58. Grest GS, Fetters LJ, Juang JS, Richter D (1996) Star polymers: Experiment, theory, and simulation. Adv Chem Phys 94: 67–163
59. Grosberg AY, Khokhlov AR (1994) Statistical physics of macromolecules. AIP, New York
60. Grubisic Z, Rempp P, Benoit H (1967) A universal calibration for gel permeation chromatography. J Polym Sci B 5:753–759
61. Guenza M, Perico A (1992) A reduced description of the local dynamics of star polymers. Macromol 25:5942–5949
62. Gutman I, Yeh Y-N, Lee S-L, Chen J-C (1994) Wiener numbers of dendrimers. Commun Math Chem (MATCH) 30:103–115
63. Gutman I, Yeh Y-N, Lee S-L, Luo Y-L (1993) Some recent results in the theory of the Wiener number. Ind J Chem 32A:651–661
64. Guttman CM, Hoffman JD, DiMarzio EA (1979) Monte Carlo calculation of SANS for various models of semicrystalline polyethylene. Faraday Discuss Chem Soc 68:297–309
65. Hadjichristidis N (1999) Synthesis of miktoarm star polymers. J Polym Sci Polym Chem 37:857–871
66. Hadjichristidis N, Pitsikalis M, Pispas S, Iatrou H (2001) Polymers with complex architecture by living anionic polymerization. Chem Rev 101:3747–3792
67. Ham JD (1957) Viscosity theory of branched and cross-linked polymers. J Chem Phys 26:625–633
68. Ham NS, Ruedenberg K (1958) Energy levels, atom populations, bond populations in the LCAO MO model and in the FE MO model. A quantitative analysis. J Chem Phys 29: 1199–1214
69. Harary F (1962) The determinant of the adjacency matrix of a graph. SIAM Rev 4:202–210
70. Harary F (1969) Graph theory. Addison-Wesley, Reading
71. Hosoya H (1971) Topological index, A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. Bull Chem Soc Jpn 44: 2332–2339
72. Hosoya H, Kawasaki K, Mizutani K (1972) Topological index and thermodynamic properties K Empirical rules on the boiling point of saturated hydrocarbons. Bull Chem Soc Jpn 45:3415–3421
73. Ivanciuc O, Balaban T, Balaban AT (1993) Reciprocal distance matrix, related local vertex incariants and topological indices. J Math Chem 12:309–318
74. Kim YH, Webster OW (1990) Water-soluble hyperbranched polyphenylene: A unimolecular micelle? J Am Chem Soc 112:4592–4593
75. Kirchhoff G (1847) Über die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Verteilung galvanischer Ströme geführt wird. Ann Phys Chem 72: 497–508
76. Kirkwood JG, Riseman J (1948) The intrinsic viscosities and diffusion constants of flexible macromolecules in solution. J Chem Phys 16:565–573
77. Klein J (1980) Self-avoiding walks constrained to strips, cylinders, and tubes. J Stat Phys 23:561–586
78. Klein J, Fletcher D, Fetters L (1983) Dynamics of entangled star-branched polymers. Faraday Symp Chem Soc 18: 159–171
79. Kloczknowski A (2002) Application of statistical mechanics to the analysis of various physical properties of elastomeric networks. Polymer 43:1503–1525
80. Kosmas MK, Gaunt DS, Whittington SG (1989) Dimensions of the branched of a uniform brush polymer. J Phys A Math Gen 22:5109–5116
81. Kramers HA (1946) The behavior of macromolecules in inhomogeneous flow. J Chem Phys 14:415–425

82. Kurata M, Abe M, Iwama M, Matsushima M (1972) Randomly branched polymers. I. Hydrodynamic properties. Polym J 3:729–738
83. Kurata M, Fukatsu M (1964) Unperturbed dimension and translational friction constant of branched polymers. J Chem Phys 41:2934–2944
84. Kurata M, Stockmayer WH (1963) Intrinsic viscosity and unperturbed dimensions of long chain molecules. Adv Polym Sci 3:196–312
85. Lovász L, Pelikán J (1973) On the eigenvalues of trees. Period Math Hung 3:175–182
86. Masuda T, Kitagawa K, Inoue T, Onogi S (1970) Rheological properties of anionic polystyrenes. II. Dynamic viscoelasticity of blends of narrow-distribution polystyrene. Macromol 3:116–125
87. Masuda T, Ohta Y, Onogi S (1981) Rheological properties of anionic polystyrenes. III. Characterization and rheological properties of four-branch polystyrenes. Macromol 4:763–768
88. McGraph JE (ed) (1981) Anionic polymerization: Kinetics, mechanism and synthesis, vol 166. ACS, Washington
89. Mekenyan O, Bopnchev D, Trinajistić N (1980) Chemical graph theory: Modeling the thermodynamic properties of molecules. Int J Quantum Chem 18:369–380
90. Mekenyan O, Dimitrov S, Bonchev D (1963) Graph-theoretical approach to the calculation of physicohemical properties of polymers. Eur Polym J 19:1185–1193
91. Merris R (1990) The distance spectrum of a tree. J Graph Theory 14:365–369
92. Mills NJ (1969) The rheological properties and molecular weight distribution of polydimethylsiloxane. Eur Polym J 5:675–695
93. Milner ST (1991) Polymer brushes. Science 251:905–914
94. Mohar B (1991) Eigenvalues, diameter, and mean distance in graphs. Graph Comb 7:53–64
95. Mohar B (1993) A novel definition of the Wiener index for trees. J Chem Inf Comput Sci 33:153–154
96. Nitta K (1994) A topological approach to statistics and dynamics of chain molecules. J Chem Phys 101:4222–4228
97. Nitta K (1999) A graph-theoretical approach to statistics and dynamics of tree-like molecules. J Math Chem 25:133–143
98. Noda I, Hearst JE (1971) Polymer dynamics. V. The shear dependent properties of linear polymers including intrinsic viscosity, flow dichroism and birefringence, relaxation, and normal stresses. J Chem Phys 54:2342–2354
99. Orofino TA (1961) Branched polymers. II. Dimensions in noninteracting media. Polymer 2:305–314
100. Pearson DS, Rju VR (1982) Configurational and viscoelastic properties of branched polymers. Macromol 15:294–298
101. Peticolas WL (1963) Introduction to the molecular viscoelastic theory of polymers and its applications. Rubber Chem Technol 36:1422–1458
102. Platt JR (1947) Influence of neighbor bonds on additive bond properties in paraffins. J Chem Phys 15:419–420
103. Platt JR (1952) Prediction of isomeric differences in paraffin properties. J Phys Chem 56:328–336
104. Plavšić D, Nikolić S, Trinajstić N, Mihalić Z (1993) On the Harary index for the characterization of chemical graphs. J Math Chem 12:235–250
105. Polansky OE, Bonchev D (1987) The Wiener number of graphs. I. General theory and changes due to graph operations. Commun Math Chem (MATCH) 21:133–186
106. Polansky OE, Bonchev D (1990) Theory of the Wiener number of graphs. II. Transfer graphs and some of their metric properties. Commun Math Chem (MATCH) 25:3–40
107. Randić M (1975) On characterization of molecular branching. J Am Chem Soc 97:6609–6615
108. Roedal M (1953) The molecular structure of polyethylene. I. Chain branching in polyethylene during polymerization. J Am Chem Soc 75:6110–6112
109. Roovers J, Graessley WW (1981) Melt rheology of some model comb polystyrenes. Macromol 14:766–733
110. Rouse PE (1953) A theory of the linear viscoelastic properties of dilute solutions of coiling polymers. J Chem Phys 21: 1272–1280
111. Rouvray DH (1975) The value of topological indices in chemistry. Commun Math Chem (MATCH) 1:125–134
112. Sack RA (1953) Mean square radius of randomly coiled molecular chain. Nature 171:310
113. Schäfer L (1999) Excluded volume effects in polymer solutions. Springer, Berlin
114. Sheridan PF, Adolf DB, Lyulin AV, Neelov I, Davies GR (2002) Computer simulations of hyperbranched polymers: The influence of the Wiener index on the intrinsic viscosity and radius of gyration. J Chem Phys 117:7802–7812
115. Sperati CA, Franta WA, Strarkweather HW (1953) The molecular structure of polyethylene. V. The effect of chain branching and molecular weight on physical properties. J Am Chem Soc 75:6127–6133
116. Spialter L (1968) The atom connectivity matrix and its characteristic polynomial: A new computer-oriented chemical nomenclature. J Am Chem Soc 85:2012–2013
117. Stejskal J, Horská J, Kratochvil P (1984) Graft copolymer statistics. Macromol 17:920–926
118. Stejskal J, Kratochvil P, Jenkins AD (1987) Graft copolymer statistics. 2. Application to graft copolymers prepared from macromonomers. Macromol 20:181–185
119. Stockmayer WH (1943) Theory of molecular size distribution and gel formation in branched chain polymers. J Chem Phys 11:45–55
120. Stockmayer WH, Fixman M (1953) Dilute solutions of branched polymers. Ann NY Acad Sci 57:334–352
121. Summers DW (1987) Konts, macromolecules and chemical dynamics. In: King RB, Rouvray DH (eds) Graph theory and topology in chemistry. Elsevier, Amsterdam, pp 3–22
122. Sylvester JJ (1878) Chemistry and algebra. Nature 17:284
123. Sylvester JJ (1878) On an application of the new atomic theory to graphical representation of the invariants and covariants of binary quantics. Amer J Math 1:64–125
124. Szleifer I, Carignano MA (1996) Tethered polymer layers. Adv Chem Phys 94:165–260
125. Tackx P, Tacx JCJF (1998) Chain architecture of LDPE as a function of molar mass using size exclusion chromatography and multi-angle laser light scattering (SEC-MALLS). Polymer 39:3109–3113
126. Tang H (1996) Rouse dynamics of block copolymers. Macromol 29:2633–2640
127. Tomalia DA, Baker H, Dewald J, Hall M, Kallos G, Martin S (1985) A new class of polymers: Starburst-dendric macromolecules. Polym J 17:117–132

128. Trinajistić N, Gutman I (2002) Mathematical chemistry. Croatica Chem Act 75:329–356
129. Tschoegl NW (1964) Influence of hydrodynamic interaction on the viscoelastic behavior of dilute polymer solutions in good solvents. J Chem Phys 40:473–479
130. Ugi T, Marquarding D, Klusacek H, Gokel G, Gillespie P (1970) Chemie und logische Strukturen. Angew Chem 82:741–771
131. Van Dam ER, Haemers WH (2003) Which graphs are determined by their spectrum? Linear Algebr Appl 373:241–272
132. Widmann AH, Davies GR (1998) Simulation of the intrinsic visocisty of hydrobranched polymers with varing topology. 1. Dendric polymers built by sequential addition. Comput Theor Polym Sci 8:191–199
133. Wiener H (1947) Correlation of heats of isomerization and differences in heats of vaporization of isomers among the paraffin hydrocarbons. J Am Chem Soc 69:2636–2638
134. Wiener H (1947) Structural determination of parafin boiling points. J Am Chem Soc 69:17–20
135. Yamakawa H (1971) Modeln theory of polymer solutions. Harper and Row, New York
136. Yang Y (1998) Graph theory of viscoelastic and configurational properties of Gaussian chains. Macromol Theory Simul 7:521–549
137. Zimm BH (1953) Dynamics of polymer molecules in dilute solution: Viscoelasticity, flow birefringence and dielectric loss. J Chem Phys 24:269–278
138. Zimm BH, Kilb RW (1959) Dynamics of branched polymer molecules in dilute solution. J Polym Sci 37:19–42
139. Zimm BH, Stockmayer WH (1949) The dimensions and chain molecules containing branches and rings. J Chem Phys 17:1301–1314

Books and Reviews

Balaban AT (1976) Chemical applications of graph theory. Academic Press, London
Beineka LW, Wilson RJ (2004) Topics in algebraic graph theory. Cambridge Univ, Cambridge
Bonchev D, Rouvray DH (1991) Chemical graph theory: Introduction and fundamentals. Abasuc Press, Gordon and Breach Sci Publ, New York
Doi M, Edwards SF (1986) The theory of polymer dynamics. Clarendon, Oxford
Flory PJ (1969) Statistical mechanics of chain molecules. Wiley, New York
King RB (1983) Chemical applications of topology and graph theory. Elsevier, Amsterdam
Small PA (1975) Long-chain branching in polymers. Adv Polym Sci 18:1–64
Trinajstić N (1992) Chemical graph theory, 2nd edn. CRC Press, Florida

Popular Wavelet Families and Filters and Their Use

Ming-Jun Lai
Department of Mathematics, The University of Georgia, Athens, USA

Article Outline

Glossary
Introduction
Definition of Wavelets
Definition of Filters
Multi-Resolution Analysis
Wavelet Decomposition and Reconstruction
Refinable Functions
Compactly Supported Orthonormal Wavelets
Parameterization of Orthonormal Wavelets
Biorthogonal Wavelets
Prewavelets
Tight Wavelet Frames
Tight Wavelet Frames over Bounded Domain
q-Dilated Orthonormal Wavelets
Multiwavelets and Balanced Multiwavelets
Multivariate Orthonormal Wavelets
Biorthogonal Box Spline Wavelets
Multivariate Prewavelets
Multivariate Tight Wavelet Frames and Bi-Frames
Spherical Tight Wavelet Frames
Wavelets for Image Processing
Future Directions
Bibliography

Glossary

B-splines N_d is the uniform B-spline of order d based on integer knot sequence. It is a function of piecewise polynomial of degree $d-1$ and smoothness $d-2$. Trigonometric B-splines T_d will also be used.

Box splines $B_{\ell,m,n}$ is a box spline of degree $\ell+m+n-2$ on three direction mesh. $B_{k,\ell,m,n}$ is a box spline of degree $k+\ell+m+n-2$ on four direction mesh. They are bivariate piecewise polynomial functions of certain smoothness dependent on integers k, ℓ, m, n.

Filter A filter is a sequence of real numbers. For example, a FIR filter is a finite sequence of real numbers. An IIR filter is a sequence of real numbers whose discrete Fourier transform is a rational function in $z = \mathrm{e}^{i\omega}$.

Filter process A filter process is to convolute a digital signal with a filter, converting an input digital signal to an output digital signal. A subband coding scheme is a synthetic filter process which convert an input signal to several output signals.

Image compression A procedure to use less bytes of information to represent the same image (within tolerance). That is, for an image of size 512×512 and standard integer gray level $[0, 255]$, the image needs a file of $512 \times 512 \times 8$ bytes to store in a computer or to be sent over the internet. If one can use a file of some bytes less

than $512 \times 512 \times 8$ to represent this image (storage or transmission), then the file is a compressed image.

Image denoise A procedure to remove noises from a noised image to make the image sharper and clearer.

Image edge detection A procedure to find features, skeleton or segmentation of images.

L_2 spaces A space of all square integrable functions.

Mask A mask is a finite sequence of real numbers. A mask polynomial is the discrete Fourier transform of a mask. Sometimes, a mask polynomial is also called the symbol of a mask.

MRA MRA stands for multi-resolution approximation (of a L_2 space).

Wavelet A function or a group of functions which can generate a basis for Hilbert space by its translations and dilations is called wavelet. Many generalized versions of wavelets will be discussed including orthonormal wavelets, biorthogonal wavelets, prewavelets, wavelet frames, multi-wavelets, q-dilated wavelets, multivariate nonseparable wavelets.

Introduction

Wavelets are one or a few functions whose integer translations and dilations can generate a basis for a Hilbert space. The concept of wavelets was introduced in the 1980's and has since been generalized and extended in many directions. The theory and applications have been continuously developed. One of its significant features is that it provides a systematical approach for designing various filters and filter banks for signal and image processing. Another feature is that wavelets leads to the theory of multi-resolution approximation (MRA). Wavelets and MRA have found many applications in most areas of science and technology, e. g., astronomy, electric engineering, fuzzy logic, geoscience, medical imaging, physics, and statistics. Wavelets have become an important subject in applied mathematics, approximation theory, numerical analysis and harmonic analysis.

In this article we present only the discrete wavelet transform, omitting the discussion of continuous wavelet transform. We shall describe various kinds of wavelets, outline their construction, and present some examples. To simplify the study of the regularity properties of wavelets, we restrict our discussion to primarily those wavelets whose construction is based on the well-known refinable functions, B-splines and box splines whose regularity is fully understood. Additionally, some wavelets without smoothness will be given as examples. Also, we shall describe the concepts of filters, filter banks, filtering processes, and their connection to wavelets. For the purpose of applications, only finite impulse response filters and realizable infinite impulse response filters (some rational filters) can be implemented. Thus, we shall restrict our attention to compactly supported wavelet functions with the exception of those associated to stable rational filters. All globally supported wavelet functions such as Meyer's wavelet, Battle–Lemarie's wavelets, Shannon's wavelets, and Freeden's spherical wavelets will not be mentioned further in this article. Also we only discuss real-valued wavelets although there are complex-valued wavelets available in the literature.

For applications of wavelets, we will consider three applications in image processing. Other applications in signal processing, in solution of integral equations and partial differential equations, in subdivision algorithms for curves and surfaces, and in statistics, have to be omitted due to the space and time limitation.

Definition of Wavelets

Let us start with a well-known Hilbert space, $L_2(\mathbb{R})$, the space of all square integrable functions on $\mathbb{R}$. That is,

$$L_2(\mathbb{R}) = \left\{ f: \quad \int_{-\infty}^{\infty} |f(x)|^2 \mathrm{d}x < +\infty \right\} .$$

A function $\psi \in L_2(\mathbb{R})$ is a *wavelet* if the integer translations and dilations

$$\psi_{jk}(x) := 2^{j/2}\psi(2^j x - k) , \quad j, k \in \mathbb{Z} \tag{1}$$

of ψ form a basis for $L_2(\mathbb{R})$. That is, these $\psi_{j,k}$ are linearly independent, and any function $f \in L_2(\mathbb{R})$ can be represented as a linear combination of ψ_{jk}.

It is clear that such a function is useful for computation since one only needs to store one function ψ in computer to represent any function in $L_2(\mathbb{R})$. In particular, a compactly supported wavelet ψ is more appropriate for the evaluation of any f represented in terms of these ψ_{jk}.

There are many kinds of wavelet functions. When these ψ_{jk} are orthonormal, i. e.,

$$\int_{-\infty}^{\infty} \psi_{j,k}(x)\psi_{m,n}(x)\mathrm{d}x = \begin{cases} 1 , & \text{if } j = m , \quad k = n \\ 0 , & \text{otherwise} \end{cases} \tag{2}$$

ψ is called *orthonormal wavelet*.

Let $W_j = \mathrm{span}\{\psi_{jk}, k \in \mathbb{Z}\}$ for each j. When W_j is orthogonal to W_m for $m \neq j$ while the integer translates $\psi_{jk}, k \in \mathbb{Z}$ may not be orthonormal to each other, ψ is called *prewavelet* or *semi-orthonormal wavelet*.

Since ψ_{jk}, $j, k \in \mathbb{Z}$ form a basis, for each $f \in L_2(\mathbb{R})$

$$f = \sum_{j,k\in\mathbb{Z}} c_{jk}\psi_{jk}$$

for some coefficients c_{jk}. Suppose that there are two positive constants A and B such that

$$A \sum_{j,k\in\mathbb{Z}} |c_{jk}|^2 \le \left\| \sum_{j,k\in\mathbb{Z}} c_{jk}\psi_{jk} \right\|_2^2 \le B \sum_{j,k\in\mathbb{Z}} |c_{jk}|^2$$

for all coefficients c_{jk}. Then ψ is called *Riesz wavelet*.

Suppose that there exists another function $\widetilde{\psi} \in L_2(\mathbb{R})$ associated with ψ such that

$$\int_{-\infty}^{\infty} \psi_{j,k}(x)\widetilde{\psi}_{m,n}(x)\mathrm{d}x = \begin{cases} 1\,, & \text{if } j = m \quad \text{and} \quad k = n \\ 0\,, & \text{otherwise} \end{cases}$$

for $j, k, m, n \in \mathbb{Z}$, where $\widetilde{\psi}_{jk}(x) = 2^{j/2}\widetilde{\psi}(2^j x - k)$. Suppose that both ψ and $\widetilde{\psi}$ are Riesz wavelets. Then ψ is called *biorthogonal wavelet* and $\widetilde{\psi}$ is a dual of ψ. In this case,

$$f = \sum_{j,k\in\mathbb{Z}} \langle f, \widetilde{\psi}_{jk}\rangle \psi_{jk} = \sum_{j,k\in\mathbb{Z}} \langle f, \psi_{jk}\rangle \widetilde{\psi}_{jk}$$

for all $f \in L_2(\mathbb{R})$, where

$$\langle f, g\rangle = \int_{-\infty}^{\infty} f(x)g(x)\mathrm{d}x$$

is the standard inner product in $L_2(\mathbb{R})$. Clearly, when $\widetilde{\psi} = \psi$, ψ is an orthonormal wavelet.

Furthermore, let us now assume that ψ_{jk}, $j, k \in \mathbb{Z}$, form a redundant basis for $L_2(\mathbb{R})$ in the sense that any function $f \in L_2(\mathbb{R})$ can be expressed by these ψ_{jk} in the following sense

$$f = \sum_{j,k\in\mathbb{Z}} c_{jk}\psi_{jk}\,,$$

where these ψ_{jk} may not be linearly independent and c_{jk} not unique. We say that these ψ_{jk} form a *tight wavelet frame* if

$$\sum_{j,k\in\mathbb{Z}} |\langle f, \psi_{jk}\rangle|^2 = \|f\|^2$$

for all $f \in L_2(\mathbb{R})$, where $\|f\|^2 := \langle f, f\rangle$.

Next, let us consider a Sobolev space $H^k(\mathbb{R})$ consisting of all functions f whose derivatives up to $k > 0$ are in $L_2(\mathbb{R})$. That is,

$$H^k(\mathbb{R}) = \{f\colon\ f^{(r)} \in L_2(\mathbb{R}), r = 0, \ldots, k\}\,,$$

where $f^{(r)}$ denotes the rth derivative of f. Certainly, if we replace $L_2(\mathbb{R})$ by a Sobolev space $H^k(\mathbb{R})$ above, we will get corresponding Sobolev wavelets. In particular, if we let ψ be a function in $H^k(\mathbb{R})$ and ψ_{jk} defined as in (1), and if

$$\langle \psi_{jk}, \psi_{mn}\rangle_k = \begin{cases} 1, & \text{if } j = m \quad \text{and} \quad k = n \\ 0, & \text{otherwise}\,, \end{cases}$$

then ψ is called *orthonormal Sobolev wavelet*. Here $\langle f, g\rangle_k$ stands for the inner product in $H^k(\mathbb{R})$ defined by

$$\langle f, g\rangle_k = \sum_{r=0}^{k} \langle f^{(r)}, g^{(r)}\rangle$$

for all $f, g \in H^k(\mathbb{R})$. Similarly we can define biorthogonal wavelets in Sobolev spaces (cf. [3,4,41]).

Furthermore, we can generalize the concepts of wavelets in several directions. If we consider the d-dimensional Euclidean space $\mathbb{R}^d$ with $d > 1$, then all of the above concepts for various wavelets can be defined in $L_2(\mathbb{R}^d)$ or in $H^k(\mathbb{R}^d)$ accordingly. These concepts will be explained in detail in later sections.

Another direction of generalization is to consider a group of wavelet functions. Let Ψ be a function vector of finite length $r > 1$. That is,

$$\Psi = [\psi_1, \ldots, \psi_r]^{\mathrm{T}}\,.$$

For example, Ψ contains a symmetric function ψ_1, antisymmetric function ψ_2, a sufficiently smooth function ψ_3 and a discontinuous function ψ_4. In fact, a wavelet vector Ψ consisting of functions of various shapes and properties is very useful to represent any function $f \in L_2(\mathbb{R})$.

A new direction of generalization is to consider a dilation factor $s > 2$. That is, we define

$$\psi_{s,j,k}(x) = s^{j/2}\psi(s^j x - k), \quad j, k \in \mathbb{Z}\,.$$

In this case, one wavelet function is usually not enough. For example, when $s = 3$ we will need 2 wavelet functions.

The above two generalizations lead to the following concept of multi-wavelets. Let $\Psi_{s,j,k}$ be the kth translation and jth dilation of Ψ with integer dilation factor $s \ge 2$. That is,

$$\Psi_{s,j,k} = [s^{j/2}\psi_1(s^j x - k), \ldots, s^{j/2}\psi_r(2^j x - k)]^{\mathrm{T}}\,.$$

Suppose that any function $f \in L_2(\mathbb{R})$ can be expressed by using the entries of $\Psi_{s,j,k}$, i. e., there exist coefficient vectors $c_{j,k}$ of length r such that

$$f = \sum_{j,k\in\mathbb{Z}} c_{j,k}^{\mathrm{T}}\Psi_{s,j,k}\,.$$

Suppose that we have

$$\int_{-\infty}^{\infty} \Psi_{s,j,k}(x)\Psi_{s,m,n}(x)^{\mathrm{T}}\mathrm{d}x = I_t\delta_{jm}\delta_{kn}$$

with I_r being the identity matrix of size $r \times r$ and δ_{jm} being the Kronecker delta, i. e., $\delta_{jm} = 1$ if $j = m$ and $\delta_{jm} = 0$ otherwise. Then Ψ is called an *orthonormal multi-wavelet*. Other concepts of wavelets can be generalized in this multi-wavelet setting as well. We leave the discussion to later sections.

Yet, another direction of generalization is to consider wavelet functions over a bounded domain. Let $\Omega \subset \mathbb{R}^d$ be a bounded domain, e. g., $\Omega = [0, 1]$ when $d = 1$. One would like to have locally supported wavelet functions over Ω.

Definition of Filters

We begin with the following basic concepts in signal processing. A continuous-time signal, or sometimes called an analog signal, $u(t)$, is a piecewise continuous function of the time variable t, where t ranges from $-\infty$ to ∞. It is called a band-limited signal if

$$u(t) = \int_{-\omega_0}^{\omega_0} \sigma(\omega)\mathrm{e}^{i\omega t}\,\mathrm{d}\omega \tag{3}$$

for a function $\sigma(\omega)$ in $L^1(-\omega_0, \omega_0)$, where ω_0 is a positive number. Certainly we shall consider only signals $u(t)$ of finite energy, i. e.,

$$\int_{-\infty}^{\infty} |u(t)|^2 < \infty\,.$$

In mathematical language, we shall restrict ourselves to functions in $L_2(\mathbb{R})$. If a band-limited analog signal $u(t)$ happens to be in $L_1(\mathbb{R})$, the space of all absolutely integrable functions over $\mathbb{R}$, then the Fourier transform of $u(t)$, defined by

$$\widehat{u}(\omega) = \int_{-\infty}^{\infty} u(t)\mathrm{e}^{-i\omega t}\mathrm{d}t$$

is given by

$$\widehat{u}(\omega) = \begin{cases} 2\pi\sigma(\omega), & \text{for} \quad -\omega_0 \le \omega \le \omega_0 \\ 0, & \text{otherwise} \end{cases}.$$

The Fourier transform takes $u(t)$ defined on the *time domain* to $\widehat{u}(\omega)$ defined on the *frequency domain*. Here, ω will be reserved for the frequency variable and i for the pure imaginary number $\sqrt{-1}$. The length of the smallest subinterval of $(-\omega_0, \omega_0)$ outside which $\sigma(\omega)$ vanishes identically is called the *bandwidth* of $u(t)$. Hence, the bandwidth of $u(t)$ as defined in (3) does not exceed $2\omega_0$. That is, the signal $u(t)$ contains no frequency higher than $\frac{\omega_0}{\pi}$ cycles per second.

Any analog signal $u(t)$ can be converted to a discrete-time signal u_n $(n = \ldots, -1, 0, 1, \ldots)$ by first sampling $u(t)$ periodically with sampling time $t_0 > 0$ and then quantizing by rounding off the values of $u(nt_0)$.

It is important to note, however, that if the sampling time t_0 is not chosen small enough, then the analog signal would not be well represented. For a band-limited analog signal, the following famous result gives us a guide-line for choosing t_0.

Theorem 1 (Shannon's Sampling Theorem) *Let $u(t)$ be a band-limited analog signal in $L_2(-\infty, \infty)$ with bandwidth $2\omega_0$, and let*

$$0 < t_0 \le \frac{\pi}{\omega_0}.$$

Then $u(t)$, $-\infty < t < \infty$, can be recovered from its values $u(nt_0)$, $n = \ldots, -1, 0, 1, \ldots$, by using the formula

$$u(t) = \sum_{n=-\infty}^{\infty} u(nt_0)\frac{\sin \pi(t/t_0 - n)}{\pi(t/t_0 - n)}\,,$$

where the convergence is uniform in $t \in (-\infty, \infty)$.

Of course, most analog signals are not band-limited. However, we have the following result:

Theorem 2 (Plancherel's Theorem) *There is a linear isometry Ψ of $L_2(\mathbb{R})$ onto $L_2(\mathbb{R})$ which is uniquely determined by the requirement that*

$$\Psi f = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt$$

for every function $f \in L_2(\mathbb{R}) \cap L_1(\mathbb{R})$ such that for every $f, g \in L_2(\mathbb{R})$,

$$\int_{-\infty}^{\infty} f(t)g(t)dt = \int_{-\infty}^{\infty} \Psi f(\omega)\Psi g(\omega)\,\mathrm{d}\omega\,.$$

By the above theorem, if $u(t)$ is in $L_2(\mathbb{R})$ and has a Fourier transform (e. g., its frequency can be seen or measured), then its Fourier transform $\hat{u}(\omega)$ is in $L_2(\mathbb{R})$. If $u(t) \in L_2(\mathbb{R}) \cap L^1(\mathbb{R})$, then $\widehat{u}(\omega)$ is a continuous function

on $(-\infty, \infty)$. Thus, $\hat{u}(\omega)$ decays to 0 as ω tends to $\pm\infty$, and $\widehat{u}(\omega)$ can be approximated by the truncated functions

$$\sigma_b(\omega) = \begin{cases} \frac{1}{2\pi}\hat{u}(\omega) & \text{for} \quad -b < \omega < b \\ 0, & \text{otherwise}\,, \end{cases}$$

and this in turn implies that $u(t)$ can be approximated by the band-limited signals

$$u_b(t) = \int_{-b}^{b} \sigma_b(\omega) e^{j\omega t}\, d\omega\,,$$

where $b \in (0, \infty)$ is usually determined by some practical criteria.

Next we introduce the z-transform of a digital signal $\{u_n : n \in \mathbb{Z}\}$ by

$$Z(\{u_n\}) = \sum_{n\in\mathbb{Z}} u_n z^{-n}$$

for $z = re^{i\omega}$ for some $0 < r < \infty$, where $u_n = u(nt_0)$ for a band-limited signal with bandwidth $2\pi/t_0$. Sometimes, it is called the discrete Fourier transform when $r = 1$.

The connection between an analog signal $u(t)$ in the frequency domain and its digital signal $\{u(nt_0)\}$ in the frequency domain is the following

Theorem 3 *Suppose that $u(t) \in L^1(-\infty, \infty)$ and its Fourier transform $\hat{u}$ of u is in $L^1(-\infty, \infty)$. Then*

$$U^*(\omega) = \sum_n u(nt_0) e^{j\omega n t_0} = \frac{1}{t_0}\sum_n \widehat{u}\left(\frac{2n\pi}{t_0} - \omega\right)$$

where $t_0 > 0$ is a sampling time.

This can be verified by using the well-known Poisson's summation formula (cf. [8]).

With the above knowledge of analog and digital signals, let us explain filters and filtering process. Although a digital filter is a (linear or nonlinear) transformation that takes any digital signal $\{u_n\}$, called an input signal, to a digital signal $\{v_n\}$, called the corresponding output signal, we are only interest in those filters which are stable, time-invariant, and linear filters. Such a filter is uniquely determined by a sequence of complex numbers

$$\ldots, h_{-2}, h_{-1}, h_0, h_1, h_2, \ldots$$

with $\sum_{n\in\mathbb{Z}} |h_n| < \infty$ and by the *filtering process* defined by convolution:

$$\{v_n\} = \{h_n\} * \{u_n\} \text{ with } \quad v_n = \sum_{j=-\infty}^{\infty} h_j u_{n-j}\,. \qquad (4)$$

By using the z-transform to both sides of the equation above, a digital filter is associated with a transfer function $H(z)$ such that

$$Z(\{v_n\}) = H(z) Z(\{u_n\})$$

with $H(z) = \sum_{n\in\mathbb{Z}} h_n z^{-n}$.

For example, a filter such that for any input signal $\{u_n\}$, an output signal $\{v_n\}$ is obtained as defined by

$$v_n = \frac{1}{k}\sum_{i=1}^{k} u_{n-i}$$

is an example of moving average filter. As another example, a filter which takes input signal $\{u_n\}$ and output a digital signal $\{v_n\}$ with $v_n = u_{-n}$ for all n is called a mirror filter. And as another example, a transfer function $H(z)$ satisfying $|H(z)|^2 + |H(-z)|^2 = 1$ is called *conjugate filter*.

If $\{h_n\}$ is a finite sequence (i. e. only finitely many $h_n \neq 0$), it is called a *finite impulse response* (FIR) digital filter. If infinitely many h_n are nonzero, the filter is called an *infinite impulse response* (IIR) digital filter.

An FIR digital filter is easy to implement. For example, if h_j, $j = 0, \ldots, N$ are only nonzero, then

$$v_n = \sum_{j=0}^{N} h_j u_{n-j} = h_0 u_n + \ldots + h_N u_{n-N}\,, \quad \forall n \in \mathbb{Z}\,.$$

Of course the operations described above can be considered as a weighted average with weights $h_0, \ldots, h_N$, an FIR filter is also called a *moving average* (MA) digital filter.

On the other hand, an IIR digital filter cannot be implemented in the same manner, simply because it is not possible to implement infinitely many scalar multiplications except for those filters whose z-transform is a rational function in z. That is, we will see that a rational function $H(z)$ indeed provides a realizable IIR digital filter. Suppose that

$$H(z) = \frac{a_0 + a_1 z^{-1} + \ldots + a_M z^{-M}}{1 - b_1 z^{-1} - \ldots - b_N z^{-N}}\,,$$

where $a_0, \ldots, a_M$, $b_1, \ldots, b_N$ are complex numbers and M, N are nonnegative integers. In terms of z-transform, we may write

$$\left(1 - \sum_{n=1}^{N} b_n z^{-n}\right)\left(\sum_{n=0}^{\infty} v_n z^{-n}\right) = \left(\sum_{n=0}^{M} a_n z^{-n}\right)\left(\sum_{n=0}^{\infty} u_n z^{-n}\right)\,.$$

Comparing with the coefficients from both sides of the above equation, we obtain

$$v_n - \sum_{j=1}^{N} b_j v_{n-j} = \sum_{j=0}^{M} a_j u_{n-j}$$

or

$$v_n = \sum_{j=0}^{M} a_j u_{n-j} + \sum_{j=1}^{N} b_j v_{n-j} \, .$$

Observe that the filtered outputs v_{n-j}, $j = 1, \ldots, N$, are used again to give the output v_n. This shows that a rational transfer function provides a realizable digital filter.

However, an important question is that this filter must be stable, i. e., it transforms any bounded input digital signal to a bounded output digital signal. The following result provides a criterion (cf. [8]).

Theorem 4 (Stability Criterion for IIR Causal Digital Filters) *An IIR digital filter with transfer function $H(z)$ is stable if and only if all the poles of the rational function $H(z)$ lie in the open unit disk $|z| < 1$.*

Since any FIR filter is stable, more FIR filters than realizable IIR filters have been designed in practice.

The most important role that the transfer function $H(z)$ of a digital filter plays is to decide what frequencies to pass and what frequencies to stop. In practice, a whole range of frequencies must be filtered. For instance, if we wish to stop all frequencies ω in the range $a < \omega < b$ and pass those frequencies ω in the range $c < \omega < d$, (where the intervals (a, b) and (c, d) lie in $(0, \pi)$ and do not overlap), then we require

$$H^*(\omega) = \begin{cases} 0, & \text{for} \quad a < \omega < b \\ 1, & \text{for} \quad c < \omega < d \, . \end{cases}$$

The intervals (a, b) and (c, d) are called the *stopband* and *passband* of the digital filter, respectively. See Figs. 1 and 2

Unfortunately, any ideal amplitude filter characteristic $|H_I^*(\omega)|$ which has a stopband (consisting of at least one interval) can not have a causal representation nor be an FIR filter. Thus we have to use FIR or realizable IIR filters to approximate the ideal filters $H_I^*(\omega)$. There are many ways to do such approximation. Wavelets are one of the modern approaches.

Subband coding is a good example of digital filtering application in signal processing. In Fig. 3, a subband coding scheme is shown, where $\boxed{H}$ denotes the filter H which is convoluted with input signal $\{s_n\}$. $\boxed{G}$, $\boxed{H^\circ}$, and $\boxed{G^\circ}$ denote filters similarly. $\boxed{2\downarrow}$ denotes a downsampling by 2 of an input signal. That is, retain every even indexed sample values and delete every odd indexed sample values of an input digital signal. And $\boxed{2\uparrow}$ denotes a upsampling by 2 of an input signal. That is, insert zero between every sample values of input signal. $\oplus$ stands for a simple addition of two input digital signals.

A typical subband coding involves filtering a broadband signal into two frequency bands (high-pass and lower-pass bands) so that the resulting subbands can be independently encoded for transmission. The encoders can be optimized to the statistics of the input signal, and made to give preferential coding weight to signal components in the parts of the spectrum that are perceptually most signif-

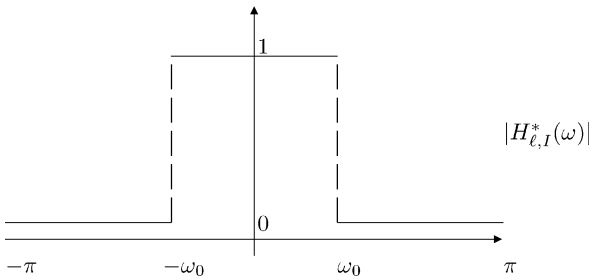

Popular Wavelet Families and Filters and Their Use, Figure 1
Ideal low–pass filter

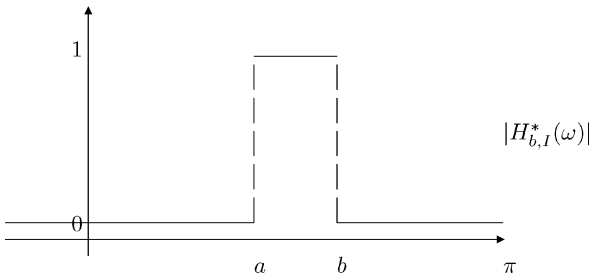

Popular Wavelet Families and Filters and Their Use, Figure 2
Ideal band–pass filter

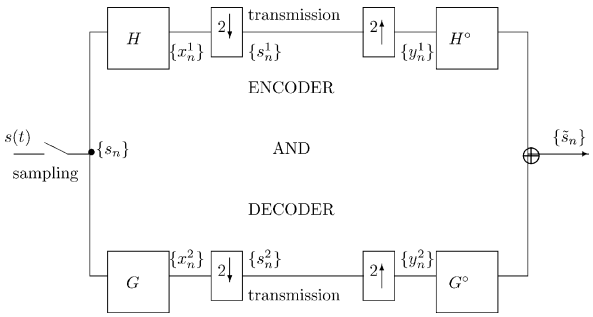

Popular Wavelet Families and Filters and Their Use, Figure 3
A subband coding schematic

icant. The received signals are decoded and the subband components recombined to give an output signal which, for a given bit rate, is subjectively better than an equivalent broad-band waveform encoded signal. Idealy, $\tilde{s}_n = s_{n+k}$ for all n by using four filters H, G, H°, and G°, where k is a certain time delay. This is called a perfect reconstruction. For application purposes, we say those four filters realize a perfect reconstruction if $S(z) = \tilde{S}(z)$, where $S(z)$ and $\tilde{S}(z)$ denote the z-transform of $\{s_n\}$ and $\{\tilde{s}_n\}$.

Let $X^1(z), X^2(z), Y^1(z), Y^2(z), S^1(z)$, and $S^2(z)$ denote the z-transform of digital signals $\{x_n^1\}$, $\{x_n^2\}$, $\{y_n^1\}$, $\{y_n^2\}$, $\{s_n^1\}$, and $\{s_n^2\}$, respectively. Write

$$H(z) = \sum_k h_k z^k \quad \text{and} \quad H^\circ(z) = \sum_k h_k^\circ z^k$$

and

$$G(z) = \sum_k g_k z^k \quad \text{and} \quad G^\circ(z) = \sum_k g_k^\circ z^k \,.$$

Then $X^1(z) = H(z)S(z)$ and $X^2(z) = G(z)S(z)$. After down-sampling by 2,

$$s_n^1 = \sum_k h_{k-2n} s_k \quad \text{and} \quad s_n^2 = \sum_k g_{k-2n} s_k \,.$$

In terms of z-transform,

$$S^1(z^2) = \frac{1}{2}(X^1(z) + X^1(-z)) \quad \text{and}$$
$$S^2(z^2) = \frac{1}{2}(X^2(z) + X^2(-z)) \,.$$

The received signal, after decoding, is up-sampled by 2 by inserting a zero valued sample between each received sample. That is,

$$y_n^1 = \begin{cases} s_{n/2}^1, & \text{if } n \text{ is even} \\ 0, & \text{if } n \text{ is odd} \end{cases}$$

and

$$y_n^2 = \begin{cases} s_{n/2}^2, & \text{if } n \text{ is even} \\ 0, & \text{if } n \text{ is odd} \,. \end{cases}$$

Or, $Y^1(z) = S^1(z^2)$ and $Y^2(z) = S^2(z^2)$. Then the two subbands are filtered by H° and G°, respectively and are added into an output signal $\{\tilde{s}_n\}$. That is,

$$\tilde{S}(z) = H^\circ(z)Y^1(z) + G^\circ(z)Y^2(z).$$

Equivalently,

$$\tilde{s}_n = \sum_k h_{2k-n}^\circ s_k^1 + g_{2k-n}^\circ s_k^2, \forall n \,.$$

It can be easily seen that

$$\tilde{S}(z) = \frac{1}{2}[H^\circ(z)H(z) + G^\circ(z)G(z)]S(z) + \frac{1}{2}[H^\circ(z)H(-z) + G^\circ(z)G(-z)]S(-z) \,.$$

Thus, in order to achieve the perfect reconstruction, we need to choose H, G, H°, and G° such that $\tilde{S}(z) = S(z)$. In terms of those four filters, we need to have

$$\begin{cases} H^\circ(z)H(z) + G^\circ(z)G(z) & = 2 \\ H^\circ(z)H(-z) + G^\circ(z)G(-z) & = 0 \,. \end{cases} \tag{5}$$

Hence, the above set of equations in (5) is called the perfect reconstruction condition.

Example 1 If $G(z) = H(-z)$, $H^\circ(z) = H(z)$, and $G^\circ(z) = -H(-z)$, then the second equation of the above system can be easily seen to satisfy. The first equation becomes

$$(H(z))^2 + (H(-z))^2 = 2 \,.$$

Such a group of four filters is called a *Quadrature Mirror Filter Bank*. It is not easy to obtain such a bank of four filters. In practice, one designs filters so that

$$\left|(H(z))^2 + (H(-z))^2 - 2\right| \approx 0 \,.$$

Example 2 Consider $G(z) = z\overline{H(-z)}$. Choosing $H^\circ(z) = \overline{H(z)}$ and $G^\circ(z) = \overline{G(z)}$ with $z = \mathrm{e}^{-i\omega}$, we have those four filters satisfied the second equation of the system (5) above. The first equation becomes

$$|H(z)|^2 + |H(-z)|^2 = 2 \,.$$

Such a group of four filters is called a *Conjugate Quadrature Filter Bank*. In later sections we shall use wavelets to give many examples of such a bank of filters.

In the above, we have discussed a simple subband splitting. Certainly, we can further split each of bandpass signals before the encoder and transmission in the above diagram. If we split only low-pass signal in the above diagram and keep the high-pass signal unsplitted, we get a nonuniform subband split. In this way, we shall obtain a wavelet decomposition of a signal if we use the filters associated with a wavelet. If we split both low- and high-pass signal further, we obtain a wavelet packet decomposition (cf. [91]).

It is easy to see that the concepts of filters, filtering process, subband coding scheme, and perfect reconstruction

have counterparts for multi-dimensional signals, e. g., 2D images (cf., e. g., [90]).

Multi-Resolution Analysis

A simple example of orthonormal wavelet is *Haar wavelet* ψ defined by

$$\psi(x) = \begin{cases} 1, & \text{if} \quad x \in [0, 1/2) \\ -1, & \text{if} \quad x \in [1/2, 1) \\ 0, & \text{otherwise} . \end{cases}$$

It is easy to check that $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$ are orthonormal and are dense in $L_2(\mathbb{R})$.

To construct a wavelet with more smoothness than that of the Haar wavelet we need a concept called multiresolution approximation of $L_2(\mathbb{R})$. It is natural to let $W_j = \text{span}\{\psi_{j,k}, k \in \mathbb{Z}\}$ and $V_j = \bigcup_{-\infty<i\le j} W_j$ for all $j \in \mathbb{Z}$. We can easily see that $V_j \subset V_{j+1}$. In fact, we have $V_{j+1} = \{f(2x-k), \forall f \in V_j, k \in \mathbb{Z}\}$. Suppose that V_0 is spanned by translates of one function ϕ, i. e.,

$$V_0 = \text{span}\left\{\sum_{k\in\mathbb{Z}} c_k \phi(x-k), \quad c_k \in \mathbb{R}\right\} .$$

Then since $V_0 \subset V_1$, we have

$$\phi(x) = \sum_{k\in\mathbb{Z}} p_k \phi(2x-k) \tag{6}$$

which is called *dilation equation* and ϕ is called *refinable function* or *scaling function*. In particular, if ϕ is also orthonormal in the sense that

$$\int_{-\infty}^{\infty} \phi(x)\phi(x-k)\mathrm{d}x = \delta_{0k}, \quad \forall k \in \mathbb{Z} , \tag{7}$$

then ϕ is called *father wavelet*. In this case, a wavelet function ψ, sometimes called *mother wavelet* can be found directly as follows. Since $V_1 = V_0 \oplus W_0$, let $W_0 = \text{span}\{\sum_{k\in\mathbb{Z}} d_k \psi(x-k), d_k \in \mathbb{R}\}$. If we can find a function $\psi \in V_1$ which is orthonormal to V_0 and integer translates of ψ are orthonormal, then ψ can be shown to be an orthonormal wavelet. Indeed, we can easily see that $\psi_{jk} \in V_{j+1}$. To see ψ_{jk} are orthonormal, let us consider the inner product of ψ_{jk} and ψ_{mn}. Without loss of generality, let us assume that $j < m$. Then ψ_{jk} is in $V_{j+1} \subset V_m$ and $\psi_{mn} \in V_m$ which is orthonormal to V_{j+1} and hence, they are orthogonal to each other. Once $V_j, j \in \mathbb{Z}$, are dense in $L_2(\mathbb{R})$, it immediately follows that ψ is an orthonormal wavelet.

In order to find such a function $\psi \in V_1$, we take Fourier transform of Eq. (6) to get

$$\widehat{\phi}(\omega) = P(\omega/2)\widehat{\phi}(\omega/2) , \tag{8}$$

where $P(\omega) = \frac{1}{2}\sum_{k\in\mathbb{Z}} p_k \mathrm{e}^{-ik\omega}$. Using Parseval's equality to Eq. (7), we have

$$\begin{aligned} \delta_{0k} &= \frac{1}{2\pi}\int_{-\infty}^{\infty} |\widehat{\phi}(\omega)|^2 \mathrm{e}^{-ik\omega}\,\mathrm{d}\omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \sum_{n\in\mathbb{Z}} |\widehat{\phi}(\omega+2n\pi)|^2 \mathrm{e}^{-ik\omega}\,\mathrm{d}\omega . \end{aligned}$$

It follows that

Theorem 5 *ϕ is orthonormal if and only if*

$$\sum_{n\in\mathbb{Z}} |\widehat{\phi}(\omega+2n\pi)|^2 = 1 .$$

In terms of (8), we have

$$|P(\omega)|^2 + |P(\omega+\pi)|^2 = 1 . \tag{9}$$

Since $\psi \in V_1$, let us write $\psi(x) = \sum_{k\in\mathbb{Z}} q_k \phi(2x-k)$. In terms of Fourier transform, we have $\widehat{\psi}(\omega) = Q(\omega/2)\widehat{\phi}(\omega/2)$. The analysis similar to the above shows that if $\psi(x-k), k \in \mathbb{Z}$ are orthonormal to each other, then

$$|Q(\omega)|^2 + |Q(\omega+\pi)|^2 = 1 .$$

Also, a similar analysis of $\int_{-\infty}^{\infty} \psi(x)\phi(x-k)\mathrm{d}x = 0$ for all $k \in \mathbb{Z}$ implies that

$$P(\omega)\overline{Q(\omega)} + P(\omega+\pi)\overline{Q(\omega+\pi)} = 0 .$$

That is, the matrix

$$\begin{bmatrix} P(\omega) & P(\omega+\pi) \\ Q(\omega) & Q(\omega+\pi) \end{bmatrix}$$

is a unitary matrix which is a necessary condition for ψ to be orthonormal. The solution of Q is trivial and it is $Q(\omega) = \mathrm{e}^{i\omega}\overline{P(\omega+\pi)}$. We remark here that after multiplying them by $\sqrt{2}$, the four Laurent polynomials $P(\omega)$, $P(\omega+\pi)$, $Q(\omega)$, and $Q(\omega+\pi)$ form a conjugate quadrature filter bank (see Example 2 in the previous section).

The above discussion explains an excellent approach to construct wavelet functions. We can summarize as follows.

Theorem 6 (Multiresolution Approximation) *Let V_j, $j \in \mathbb{Z}$ be a nested sequence of subspaces of $L_2(\mathbb{R})$. That is, $V_j \subset V_{j+1}, \forall j \in \mathbb{Z}$. Suppose that $\bigcup_{j\in\mathbb{Z}} V_j$ is dense in $L_2(\mathbb{R})$ and $\bigcap_{j\in\mathbb{Z}} = \{0\}$. If there exists a scaling or refinable function $\phi \in V_0$ whose integer translates span V_0 such that there exist positive constants A and B*

$$A \le \sum_{k\in\mathbb{Z}} |\widehat{\phi}(\omega + 2k\pi)|^2 \le B\,,$$

then there exists an orthonormal wavelet function $\psi \in V_1$ such that $\{\psi_{jk}, j, k \in \mathbb{Z}\}$ is an orthonormal basis for $L_2(\mathbb{R})$.

A proof of this result is provided in [70]. We now explain the usefulness when a nested sequence $\{V_j, j \in \mathbb{Z}\}$ forms a multiresolution approximation of $L_2(\mathbb{R})$. To approximate a function $f \in L_2(\mathbb{R})$, we can use functions in V_j for some integer j. If functions in V_j can not approximate f very well, we look for functions in the next level V_{j+1} to approximate f by adding functions in W_j. We can continue to add functions in $W_{j+1}, W_{j+2}, \ldots$ to get better and better approximation of f. Thus we can build several levels of approximations of f. In other word, f can be approximated in different resolutions.

When $A \neq B$ in Theorem 6, ψ is not in general compactly supported. To obtain a compactly supported orthonormal wavelet, we have to study the properties of ϕ to satisfy the necessary and sufficient condition in Theorem 5 There are several equivalent conditions which can be found in [24]. We shall discuss the construction of compactly supported orthonormal wavelets as well as other wavelet functions in later sections.

Wavelet Decomposition and Reconstruction

Suppose that we have a compactly supported refinable function $\phi \in L_2(\mathbb{R})$ which generates an MRA of $L_2(\mathbb{R})$. Let $V_j = \text{span}\{\phi_{j,k}, k \in \mathbb{Z}\}$ for all $j \in \mathbb{Z}$ and W_j be the the orthogonal complement of V_j in V_{j+1}, where $\phi_{jk} = 2^{j/2}\phi(2^j x - k)$ as usual. Also let $\psi \in W_0$ be a wavelet function which spans W_0 in the sense that $W_0 = \text{span}\{\psi(x-k), \forall k \in \mathbb{Z}\}$ and

$$A \le \sum_{k\in\mathbb{Z}} |\widehat{\psi}(\omega + 2k\pi)|^2 \le B$$

for two positive constants A and B. Because $V_0 \subset V_1$, $W_0 \subset V_1$ and $V_1 = V_0 + W_0$, we have the following decomposition and reconstruction sequences:

$$\begin{aligned} \phi(x) &= \sum_{k\in\mathbb{Z}} p_k 2^{1/2}\phi(2x-k) \\ \psi(x) &= \sum_{k\in\mathbb{Z}} q_k 2^{1/2}\phi(2x-k) \end{aligned} \tag{10}$$

and

$$2^{1/2}\phi(2x-k) = \sum_{j\in\mathbb{Z}} (a_{2j-k}\phi(x-j) + b_{2j-k}\psi(x-j)), \tag{11}$$

where only finitely many p_k, q_k, a_j, b_j are nonzero.

Let $f \in L_2(\mathbb{R})$ be a band-limited signal as explained in Sect. "Definition of Filters". We first approximate f by the MRA generated by ϕ at a fine level. That is,

$$A_j f(x) = \sum_{k\in\mathbb{Z}} s_{j,k}\phi_{jk}(x)\,,$$

for some coefficients $s_{j,k}$. For example, when ϕ is orthonormal, we may use orthogonal projection,

$$s_{j,k} := \langle f, \phi_{jk}\rangle\,.$$

As another example, we may choose $s_{j,k} = f(k/2^j)$ to be the digital samples of f if j is sufficiently large such that $1/2^j < \pi/\omega_0$, where ω_0 is the the half of the bandwidth of f. All we need is to make sure that $A_j f$ converges to f in $L_2(\mathbb{R})$ as $j \longrightarrow \infty$. Now we can decompose $A_j f$ into

$$A_{j-1} f = \sum_{k\in\mathbb{Z}} s_{j-1,k}\phi_{j-1,k} \in V_{j-1}$$

and

$$D_{j-1} f = \sum_{k\in\mathbb{Z}} \mathrm{d}_{j-1,k}\psi_{j-1,k} \in W_{j-1}$$

by using (11) with

$$\begin{aligned} s_{j-1,k} &= \sum_{m\in\mathbb{Z}} s_{j,m} a_{2k-m} \\ d_{j-1,k} &= \sum_{m\in\mathbb{Z}} s_{j,m} b_{2k-m}\,. \end{aligned}$$

Note that the digital signal $\{s_{j-1,k}\}$ is computed from $\{s_{j,k}\}$ by first convolution with a digital filter $\{a_k\}$ and then downsampling by 2. Similar for $\{d_{j-1,k}\}$. We can certainly continue such a decomposition step. That is, we get $A_j f, A_{j-1} f, \ldots, A_{j-\ell} f$ and $D_{j-1} f, D_{j-2} f, \ldots, D_{j-\ell} f$ for some $\ell \ge 1$.

On the other hand, we can reconstruct $A_j f$ back from these $A_{j-\ell} f, D_{j-\ell}, D_{j-\ell+1}, \ldots, D_{j-1} f$. Indeed, let us say $\ell = 1$. Then

$$\begin{aligned} A_{j-1} f &= \sum_{k\in\mathbb{Z}} s_{j-1,k}\phi_{j-1,k} = \sum_{k\in\mathbb{Z}} s_{j-1,k} \sum_{m\in\mathbb{Z}} p_{m-2k}\phi_{j,m} \\ &= \sum_{m\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} s_{j-1,k} p_{m-2k}\phi_{j,m}\,. \end{aligned}$$

Similar for $D_{j-1}f$. Thus, $A_j f = A_{j-1}f + D_{j-1}f$ implies that

$$s_{j,m} = \sum_{k\in\mathbb{Z}} s_{j-1,k}p_{m-2k} + \sum_{k\in\mathbb{Z}} \mathrm{d}_{j-1,k}q_{m-2k}\,.$$

In terms of signal processing, the digital signal $\{s_{j,k}\}$ is obtained from $\{s_{j-1,k}\}$ and $\{d_{j-1,k}\}$ by first upsampling $\{s_{j-1,k}\}$ by 2 then convoluting with digital filter $\{p_k\}$, secondly upsampling $\{d_{j-1,k}\}$ by 2 and convoluting with $\{q_k\}$, and finally adding the resulting signals together.

The above decomposition/reconstruction procedures are exactly the same as the well-known subband coding of signals described in Sect. "Definition of Filters". It is clear that the decomposition/reconstruction can be generalized to deal with 2D signals. In other words, 2D images can be decomposed and reconstructed in the same fashion. This decomposition and reconstruction process is very popular in applications such as image/signal denoising, image/signal feature extraction, and image/signal compression. A less popular, but useful idea to use wavelets is to detect singularities from signal by examining the wavelet coefficients $D_{j-1}f, \ldots, D_{j-\ell}f$. Another less popular, but useful, method is to use wavelets to build more accurate solutions. That is, if we have $A_j f$ and if we compute $D_j f$ from f directly, then we can build up $A_{j+1}f$ by letting $A_{j+1}f = A_j f + D_j f$. In particular, when $D_j f$ may be computed using the same expense as $A_j f$, $A_{j+1}f$ can be done in parallel to achieve the efficiency.

Refinable Functions

As we saw in the previous section, we need a refinable function ϕ to build a multiresolution approximation of $L_2(\mathbb{R})$. That is, we want to have

$$\widehat{\phi}(\omega) = P(\omega/2)\widehat{\phi}(\omega/2)$$

for some Laurant polynomial function $P(\omega)$. Fortunately, there are many known functions satisfying the above dilation relation. Fix any integer $n \geq 1$ and let

$$\widehat{N_n}(\omega) = \left(\frac{1-\mathrm{e}^{i\omega}}{i\omega}\right)^n. \tag{12}$$

This is the well-known *uniform B-spline* function of order n over integer knot sequence $\{0, 1, \ldots, n\}$. Since $1-\mathrm{e}^{i\omega} = (1+\mathrm{e}^{i\omega/2})(1-\mathrm{e}^{i\omega/2})$, this N_n is refinable with $P(\omega) = (1+\mathrm{e}^{i\omega})^n/2^n$. See [6] and [84] for the properties of B-splines.

To construct other refinable functions, we have to study what kinds of sequence $\{p_k, k \in \mathbb{Z}\}$ such that $P(\omega) = \sum_{k\in\mathbb{Z}} p_k \mathrm{e}^{ik\omega}$ can generate a refinable function ϕ in the following sense:

$$\widehat{\phi}(\omega) = P(\omega/2)\widehat{\phi}(\omega/2) = \ldots = \prod_{j=1}^{\infty} P(\omega/2^j)\widehat{\phi}(0)\,.$$

It follows that $P(0)$ must be equal to 1 in order that the above infinite product converges. When $P(0) = 1$ and $P(\omega)$ is a Laurent polynomial in the sense that only finitely many p_k are nonzero, $\widehat{\phi}(\omega)$ is a continuous function. Furthermore, the distribution ϕ is compactly supported by using Paley–Wiener's Theorem. It turns out that the condition (9) ensures that $\widehat{\phi}$ belongs to $L_2(\mathbb{R})$ (cf. [70]). Hence, when $P(0) = 1$, $P(\omega)$ is a Laurent polynomial, and $P(\omega)$ satisfies (9), ϕ is a compactly supported function in $L_2(\mathbb{R})$. We still need to check if ϕ satisfies the necessary and sufficient condition for orthonormality in Theorem 5. There is an example that P satisfies the above mentioned three properties, but P is not orthonormal (cf. [24]).

If ϕ is orthonormal, we can use the constructive procedure in Sect. "Definition of Filters", that is, Theorem 6 to find the associated wavelet function. This constitutes a powerful method of construction of univariate orthonormal wavelets. All the construction of various wavelets are based this idea and started with the construction of refinable functions.

Once we have wavelets, we would like to know the smoothness property of the wavelets. Clearly, the smoothness of ϕ determines the smoothness of the associated wavelet function. In terms of Fourier transform, $\phi \in H^k(\mathbb{R})$ if and only if

$$\int_{-\infty}^{\infty} (1+|\omega|^2)^k |\widehat{\phi}(\omega)|^2\,\mathrm{d}\omega < +\infty$$

for some $k > 0$. Thus, $\widehat{\phi}$ needs a decay factor $O(|\omega|^{-(k+1)})$. One way to get it is to use the Fourier transform of B-spline of degree k (or order $k+1$). Using this factor,

$$\widehat{\phi}(\omega) = \left(\frac{1-\mathrm{e}^{i\omega}}{i\omega}\right)^{k+1} \widetilde{\phi}(\omega)$$

for some function $\widetilde{\phi} \in L_2(\mathbb{R})$. In this case, $\widehat{\phi}(2\omega) = P(\omega)\widehat{\phi}(\omega)$ with mask $P(\omega) = \left(\frac{1+\mathrm{e}^{i\omega}}{2}\right)^{k+1} \tilde{P}(\omega)$ for some Laurent polynomial $\hat{P}(\omega)$. It turns out that the factor is related to the polynomial reproductivity by the linear combination of integer translates of ϕ. By Strang–Fix conditions ([89] or [60]), a polynomial q of degree $\leq k$ can be

expressed by

$$q(x) = \sum_{j\in\mathbb{Z}} c_j(q)\phi(x-j)$$

for some coefficients $c_j(q)$ if and only if $\widehat{\phi}$ contains this factor.

Compactly Supported Orthonormal Wavelets

As we discussed in previous sections, the necessary conditions for ϕ to be an orthonormal scaling function are placed on the mask polynomial $P(\omega)$, where we recall that $\widehat{\phi}(2\omega) = P(\omega)\widehat{\phi}(\omega)$. They are

- 1) $P(0) = 1$;
- 2) $|P(\omega)|^2 + |P(\omega+\pi)|^2 = 1$;
- 3) $P(\omega)$ contains a factor $\left(\frac{1+e^{i\omega}}{2}\right)^k$ for some $k > 0$.

Let $P(\omega) = \left(\frac{1+e^{i\omega}}{2}\right)^k \tilde{p}(\omega)$ and $|P(\omega)|^2 = |\cos(\omega/2)|^{2k}$ $|\tilde{p}(\omega)|^2$. For simplicity we write $x = \cos^2(\omega/2)$ and $|\tilde{p}(\omega)|^2 = p(1-x)$. Then the requirement 2) is equal to

$$x^k p(1-x) + (1-x)^k p(x) = 1\,. \tag{13}$$

It is easy to see that

$$\begin{aligned} 1 = (1-x+x)^{2k+1} &= \sum_{j=0}^{k}\binom{2k+1}{j}(1-x)^j x^{2k-j} \\ &+ \sum_{j=0}^{k}\binom{2k+1}{2k+1-j}(1-x)^{2k-j}x^j \\ &= x^k\sum_{j=0}^{k}\binom{2k+1}{j}(1-x)^j x^{k-j} \\ &+ (1-x)^k\sum_{j=0}^{k}\binom{2k+1}{j}(1-x)^{k-j}x^j\,. \end{aligned} \tag{14}$$

Thus, we may choose $p(x) = \sum_{j=0}^{k}\binom{2k+1}{j}(1-x)^{k-j}x^j$. Note that for $x = \cos^2(\omega/2) \in [0,1]$, $p(x) = p(\cos^2(\omega/2)) = p(1/2 + (e^{i\omega}+e^{-i\omega})/4)$ is a Laurent polynomial. Since $p(x) \geq 0$, there exists $\tilde{p_k}(e^{i\omega})$ such that

$$p(x) = |\tilde{p_k}(e^{i\omega})|^2$$

by Fejér–Riesz's Lemma. It is easy to see that $p(1-x) = |\tilde{p_k}(-e^{i\omega})|^2$. Thus, letting

$$P_k(\omega) = \left(\frac{1+e^{i\omega}}{2}\right)^k \tilde{p_k}(e^{i\omega})\,,$$

we can see that P_k satisfies 2). Note that P_k satisfies 1). as well. We define a compactly supported function $\phi \in L_2(\mathbb{R})$, in terms of Fourier transform, by

$$\widehat{\phi_k}(\omega) = \prod_{j=1}^{\infty} P_k(\omega/2j)\,.$$

One can prove (see [24]) that this function ϕ_k is an orthonormal scaling function and generates a multiresolution approximation of $L_2(\mathbb{R})$ when $k \geq 1$. Then we define the associated wavelet ψ_k by its Fourier transform:

$$\widehat{\psi_k}(\omega) = Q_k(\omega/2)\widehat{\phi_k}(\omega/2)\,,$$

where $Q_k(\omega) = e^{i\omega}\overline{P_k(\omega+\pi)}$. As explained in the previous sections, ψ_k is an orthonormal wavelet. These wavelets are called Daubechies wavelets which were invented in 1988 (cf. [23]).

Example 3 Let $k = 1$. Then $P_1(\omega) = \frac{1+e^{\omega}}{2}$ and hence, ϕ is the uniform B-spline of order 1 which is the characteristic function

$$\phi(x) = \begin{cases} 1, & x \in [0,1) \\ 0, & \text{otherwise}\,. \end{cases}$$

The associated wavelet is the well-known Haar wavelet:

$$\psi(x) = \begin{cases} 1, & x \in [0,1/2) \\ -1, & x \in [1/2,1) \\ 0, & \text{otherwise}\,. \end{cases}$$

When $2 \leq k \leq 10$, the coefficients of $P(z)$ associated with Daubechies wavelets are given in [24] which can be immediately used for wavelet decomposition and reconstruction.

Next we look at the smoothness of these scaling functions. Unfortunately, ϕ_k does not belong to $H^k(\mathbb{R})$ since the $\prod_{j=1}^{\infty}\widetilde{p_k}(e^{i\omega/2^j})$ does not belong to $L_2(\mathbb{R})$. One has to use a part of the decay factor $\left(\frac{1+e^{i\omega}}{2}\right)^k$ to ensure $\prod_{j=1}^{\infty}\left(\frac{1+e^{\omega/2^j}}{2}\right)^{k-k'}\tilde{p_k}(e^{i\omega/2^j})$ to be in $L_2(\mathbb{R})$ and thus, $\phi_k \in H^{k'}(\mathbb{R})$ for some $k' < k$ (see [24] for k').

Parameterization of Orthonormal Wavelets

In the previous section we obtained special solutions to the requirements 1), 2), and 3). We now look for general solutions. For convenience, let $P_n(\omega) = c_0 + c_1 z + \ldots + c_{n-1}z^{n-1}$ with $z^{i\omega}$. We look for $P_n(\omega)$ satisfying 1), 2), and 3) with $k = 1$ discussed in the previous section. That is, $P_n(0) = 1$, $P(\pi) = 0$, and

$$|P_n(\omega)|^2 + |P_n(\omega+\pi)|^2 = 1, \quad \forall\omega \in \mathbb{R}\,.$$

In this section, we shall show the solutions $P_n(\omega)$ for $n = 4$ and $n = 6$. For general $n \geq 8$ see [59]. As in the previous section, once we have P_n, we can define a refinable function $\phi_n \in L_2(\mathbb{R})$ and its associated wavelet function ψ_n in terms of Fourier transform, by

$$\widehat{\phi_n}(\omega) = \prod_{j=1}^{\infty} P_n\left(\omega/2^j\right)$$

and with $Q_n(\omega) = z\overline{P_n(\omega+\pi)}$ and $z = \mathrm{e}^{i\omega}$,

$$\widehat{\psi_n}(\omega) = Q_n(\omega/2)\widehat{\phi_n}(\omega/2)\,.$$

We have to point out that not all the refinable functions ϕ_n are orthonormal. However, the associated wavelet functions generate tight wavelet frames to be defined in a later section.

We first consider $n = 4$. Write $P_4(\omega) = a_0 + b_0 z + a_1 z^2 + b_1 z^3$.

Lemma 1 *$P_4(\omega)$ satisfies $P_4(0) = 1$ and*

$$|P_4(\omega)|^2 + |P_4(\omega+\pi)|^2 = 1, \quad \forall \omega \in \mathbb{R}$$

if and only if

$$a_0 = \frac{1}{4} + \frac{1}{2\sqrt{2}}\cos\alpha, \quad b_0 = \frac{1}{4} + \frac{1}{2\sqrt{2}}\sin\alpha,$$
$$a_1 = \frac{1}{4} - \frac{1}{2\sqrt{2}}\cos\alpha, \quad b_1 = \frac{1}{4} - \frac{1}{2\sqrt{2}}\sin\alpha\,,$$

for any $\alpha \in \mathbb{R}$.

Example 4 When $\alpha = \frac{\pi}{4}$, we get $P_4(\omega) = \frac{1+z}{2}$, which is associated with the Haar wavelet.

Example 5 When $\alpha = \frac{5\pi}{12}$, we get $a_0 = \frac{1+\sqrt{3}}{8}$ and $a_1 = \frac{3-\sqrt{3}}{8}$ as well as $b_0 = \frac{3+\sqrt{3}}{8}$, $b_1 = \frac{1-\sqrt{3}}{8}$. Then $P_4(\omega)$ is associated with the Daubechies D4 wavelet.

Next, we look for choices of α where $P_4(\omega)$ has a second-order vanishing moment, that is, $P_4(\omega) = \left(\frac{1+z}{2}\right)^2 \tilde{p}(z)$ where $\tilde{p}(z)$ is some trigonometric polynomial.

In this case $\frac{\mathrm{d}}{\mathrm{d}z}P_4(\omega)|_{z=-1} = 0$ which is equivalent to $\alpha = \frac{5\pi}{12}$ or $\alpha = \frac{13\pi}{12}$. Both are associated with Daubechies' D4 wavelet. Thus, D4 is the only member of this family with two vanishing moments.

Next we consider $P_6(\omega) = a_0 + b_0 z + a_1 z^2 + b_1 z^3 + a_2 z^4 + b_2 z^5$.

Lemma 2 *$P_6(\omega)$ satisfies $P_6(0) = 1$ and*

$$|P_6(\omega)|^2 + |P_6(\omega+\pi)|^2 = 1, \forall \omega \in \mathbb{R}$$

if and only if

$$a_0 = \frac{1}{8} + \frac{1}{4\sqrt{2}}\cos\alpha + \frac{p}{2}\cos\beta$$
$$a_1 = \frac{1}{4} - \frac{1}{2\sqrt{2}}\cos\alpha$$
$$a_2 = \frac{1}{8} + \frac{1}{4\sqrt{2}}\cos\alpha - \frac{p}{2}\cos\beta$$
$$b_0 = \frac{1}{8} + \frac{1}{4\sqrt{2}}\sin\alpha + \frac{p}{2}\sin\beta$$
$$b_1 = \frac{1}{4} - \frac{1}{2\sqrt{2}}\sin\alpha$$
$$b_2 = \frac{1}{8} + \frac{1}{4\sqrt{2}}\sin\alpha - \frac{p}{2}\sin\beta\,,$$

where

$$p = \frac{1}{2}\sqrt{1 + \sin\left(\alpha + \frac{\pi}{4}\right)}$$

for any $\alpha, \beta \in \mathbb{R}$.

Example 6 If $\alpha = \frac{\pi}{4}$ and $\beta = \frac{\pi}{4}$, then $P_6(\omega)$ is associated with the Haar wavelet.

Example 7 If $\alpha = \frac{5\pi}{12}$ and $\beta = \frac{\pi}{3}$, then $P_6(\omega)$ is associated with the Daubechies D4 wavelet.

Example 8 If

$$\cos\alpha = -\frac{1}{4}\sqrt{8 + \sqrt{15 + 12\sqrt{10}}},$$
$$\sin\alpha = \frac{1}{4}\sqrt{8 - \sqrt{15 + 12\sqrt{10}}},$$
$$\cos\beta = \frac{1}{4}\sqrt{8 - 8\sqrt{-25 + 8\sqrt{10}}},$$
$$\sin\beta = \frac{1}{4}\sqrt{8 + 8\sqrt{-25 + 8\sqrt{10}}}\,,$$

then $P_6(\omega)$ is the filter associated with Daubechies D6 wavelet.

Example 9 To see when $P_6(\omega)$ has two vanishing moments, we require $P'(\pi) = 0$ which implies

$$\sin(\alpha+\pi) = \frac{1}{8\sin^2(\beta - \pi/4)} - 1\,.$$

When the ordered pair (α, β) satisfy the above equation, $P_6(\omega)$ has two vanishing moments.

Example 10 With $\alpha = 1.4288992721907328$, $\beta = 1.1071487177940904$, $P_6(\omega)$ is associated with the most smooth length-four filter as given in [24]. When $\alpha = 1.9886461158096038$ and $\beta = 1.0934936891036087$, $P_6(\omega)$ is associated with the most smooth length-six filter as given in [24].

Biorthogonal Wavelets

In this section we explain biorthogonal wavelets. Suppose that ϕ and $\widetilde{\phi}$ are two refinable functions in $L_2(\mathbb{R})$ which are dual to each other in the following sense:

$$\int_{-\infty}^{\infty} \phi(x-j)\widetilde{\phi}(x-k)\mathrm{d}x = \begin{cases} 1\,, & \text{if } j=k, \\ 0\,, & \text{otherwise} \end{cases} \tag{15}$$

for all $j, k \in \mathbb{Z}$. As usual, define $\phi_{jk}(x) = 2^{j/2}\phi(2^j x - k)$ and $\widetilde{\phi}_{jk}(x) = 2^{j/2}\widetilde{\phi}(2^j x - k)$ for all $j, k \in \mathbb{Z}$. Also we let $V_j = \mathrm{span}\{\phi_{j,k}, k \in \mathbb{Z}\}$ and $\widetilde{V}_j = \mathrm{span}\{\widetilde{\phi}_{j,k}, k \in \mathbb{Z}\}$. We look for $\psi \in V_1$ and $\widetilde{\psi} \in \widetilde{V}_1$ such that

$$\langle \phi(\cdot), \widetilde{\psi}(\cdot - k)\rangle = 0,\ \langle \psi(\cdot), \widetilde{\phi}(\cdot - k)\rangle = 0,$$
$$\langle \psi(\cdot), \widetilde{\psi}(\cdot - k)\rangle = \delta_k, \tag{16}$$

for all $k \in \mathbb{Z}$, where $\delta_k = 1$ when $k = 0$ and $= 0$ for other k. Then ψ and $\widetilde{\psi}$ will be a pair of biorthogonal wavelets under the assumptions that both families $\{\phi_{jk}, j, k \in \mathbb{Z}\}$ and $\{\widetilde{\phi}_{jk}, j, k \in \mathbb{Z}\}$ are multiresolution approximations of $L_2(\mathbb{R})$.

Indeed, let $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$ and $\widetilde{\psi}_{jk}(x) = 2^{j/2}\widetilde{\psi}(2^j x - k)$ for all $j, k \in \mathbb{Z}$. For any j, k and m, n we claim that

$$\int_{-\infty}^{\infty} \psi_{j,k}(x)\widetilde{\psi}_{m,n}(x)\mathrm{d}x = \delta_{j,m}\delta_{k,n}\,.$$

For $j < m$, we know from (16) that $\widetilde{\psi}_{m,n}$ is orthogonal to $\phi_{m,k}$ for all k and hence is orthogonal to V_m which contains ψ_{j+1} since $\psi_{j,\ell} \in V_{j+1} \subset \ldots \subset V_m$. Similar for $j > m$. When $j = m$, we use (16) again. We now need the following concept:

Definition 1 A family $\{\phi_{j,k}, j, k \in ZZ\}$ is a Riesz basis for $L_2(\mathbb{R})$ if

1. $\phi_{j,k}, j, k \in \mathbb{Z}$ are linearly independent, and
2. there exist two strictly positive constants A and B such that for any $f \in L_2(\mathbb{R})$,

$$A\|f\|_2^2 \le \sum_{j,k} \left| \int_{-\infty}^{\infty} f(x)\phi_{j,k}(x)\mathrm{d}t \right|^2 \le B\|f\|_2^2.$$

Let P and $\widetilde{P}$ be the mask polynomials associated with ϕ and $\widetilde{\phi}$. That is,

$$\widehat{\phi}(\omega) = P(\omega/2)\widehat{\phi}(\omega/2) \quad \text{and} \quad \widehat{\widetilde{\phi}}(\omega) = \widetilde{P}(\omega/2)\widehat{\widetilde{\phi}}(\omega/2)\,.$$

Then the condition (15) implies

$$P(\omega)\overline{\widetilde{P}(\omega)} + P(\omega+\pi)\overline{\widetilde{P}(\omega+\pi)} = 1\,. \tag{17}$$

Let us define the associated wavelet functions ψ and $\widetilde{\psi}$ in terms of Fourier transform by

$$\widehat{\psi}(\omega) = Q(\omega/2)\widehat{\phi}(\omega/2) \quad \text{and} \quad \widehat{\widetilde{\psi}}(\omega) = \widetilde{Q}(\omega/2)\widehat{\widetilde{\phi}}(\omega/2),$$

where $Q(\omega) = \mathrm{e}^{i\omega}\overline{\widetilde{P}(\omega+\pi)}$ and $\widetilde{Q}(\omega) = \mathrm{e}^{i\omega}\overline{P(\omega+\pi)}$. The conditions in (16) give

$$\begin{bmatrix} P(\omega) & P(\omega+\pi) \\ Q(\omega) & Q(\omega+\pi) \end{bmatrix} \begin{bmatrix} \widetilde{P}(\omega) & \widetilde{P}(\omega+\pi) \\ \widetilde{Q}(\omega) & \widetilde{Q}(\omega+\pi) \end{bmatrix}^* = I_2$$

where I_2 denotes the identity matrix of size 2×2.

In fact, the Q and $\widetilde{Q}$ so defined as above, one only needs to solve (17). If P and $\widetilde{P}$ satisfy (17) and if the ϕ and $\widetilde{\phi}$ generate two Riesz bases for $L_2(\mathbb{R})$, then one can show that ψ and $\widetilde{\psi}$ are biorthogonal wavelet functions (cf. [22])

In this setting, for any $f \in L_2(\mathbb{R})$, we have

$$f(x) = \sum_{j,k\in\mathbb{Z}} \langle f, \widetilde{\psi}_{j,k}\rangle \psi_{j,k}(x) = \sum_{j,k\in\mathbb{Z}} \langle f, \psi_{jk}\rangle \widetilde{\psi}_{jk}(x)\,.$$

To construct some examples, we begin with the uniform B-spline ϕ_k whose Fourier transform is $\left(\frac{1-\mathrm{e}^{i\omega}}{i\omega}\right)^k$. It is well-known that ϕ_k is a refinable and generate a multiresolution approximation of $L_2(\mathbb{R})$. Let $P_k(\omega) = \left(\frac{1+\mathrm{e}^{i\omega}}{2}\right)^k$ be its mask polynomial. Next we construct a dual $\widetilde{\phi^{k,n}}$ whose mask polynomial

$$\widetilde{P_{k,n}}(\omega) = \left(\frac{1+\mathrm{e}^{-i\omega}}{2}\right)^{2n-k} \mathrm{e}^{in\omega}\overline{p_n(\sin^2(\omega/2))}$$

where $p_n(x)$ is a polynomial in x defined in (14). So that

$$\begin{aligned} & P(\omega)\overline{\widetilde{P_{k,n}}(\omega)} + P(\omega+\pi)\overline{\widetilde{P_{k,n}}(\omega+\pi)} \\ &= \left(\frac{1+\mathrm{e}^{i\omega}}{2}\right)^{2n} \mathrm{e}^{-in\omega} p_n(\sin^2(\omega/2)) + \\ &\quad \left(\frac{1-\mathrm{e}^{i\omega}}{2}\right)^{2n} \mathrm{e}^{-in\omega}(-1)^n p_n(\cos^2(\omega/2)) \\ &= (\cos^2(\omega/2))^n p_n(\sin^2(\omega/2)) \\ &\quad + (\sin^2(\omega/2))^n p_n(\cos^2(\omega/2)) = 1 \end{aligned}$$

by using (14) in a previous section. If we choose n large enough we can ensure that $\widetilde{\phi^{k,n}} \in L_2(\mathbb{R})$ and satisfies the duality relation (15) with ϕ_k by using the equation above.

There are many other choices of biorthogonal wavelets. In the following we explain a well-known biorthogonal wavelet called CDF 9/7 wavelet (cf. [50]). It was famous for finger print compression employed by FBI in the beginning of 90's. Now it is exclusively used by the state-of-art JPEG2000 for image compression.

Example 11 Letting $z = e^{i\omega}$, we choose

$$P(\omega) = \left(\frac{1+z}{2}\right)^4 (c_0 + c_1 z + (1 - 2(c_0 + c_1))z^2 + c_1 z^3 + c_0 z^4)$$

$$\widetilde{P}(\omega) = \left(\frac{1+z}{2}\right)^4 z(\tilde{c}_0 + \tilde{c}_1 z + \tilde{c}_0 z^2)\,.$$

The reason to choose this pattern of coefficients is to make ϕ and $\widetilde{\phi}$ symmetric.

The necessary biorthogonal condition (17) implies that the coefficients c_0, c_1 and $\tilde{c}_0, \tilde{c}_1$

$$c_0 = \frac{5}{24} + \frac{\sqrt{15}-5}{48}\alpha + \frac{2\sqrt{15}-5}{336}\alpha^2,$$
$$c_1 = -\frac{3}{2} + \frac{6-\sqrt{15}}{12}\alpha - \frac{\sqrt{15}-3}{24}\alpha^2,$$
$$\tilde{c}_0 = -\frac{1}{3} - \frac{1}{12}\alpha + \frac{3\sqrt{15}-11}{168}\alpha^2,$$
$$\tilde{c}_1 = \frac{5}{3} + \frac{\alpha}{6} - \frac{3\sqrt{15}-11}{84}\alpha^2\,,$$

where $\alpha = (154 + 42\sqrt{15})^{1/3}$. This leads to the well-known CDF biorthogonal 9/7 wavelets. Writing $P(z) = \sum_{i=0}^{8} p_i z^i$, $Q(z) = \sum_{i=0}^{6} q_i z^i$, $\widetilde{P}(z) = \sum_{i=1}^{7} \tilde{p}_i z^i$ and $\widetilde{Q}(z) = \sum_{i=-1}^{7} \widetilde{q}_i z^i$ to be the mask polynomials associated with the CDF 9/7 wavelets, their coefficients are listed in Table 1 which can be directly used for computation.

Popular Wavelet Families and Filters and Their Use, Table 1
Numerical values of $\{p_k\}_{k=0}^{8}$, $\{q_k\}_{k=0}^{6}$, $\{\widetilde{p}_k\}_{k=1}^{7}$, and $\{\widetilde{q}_k\}_{k=-1}^{7}$

p_0	.0534975148216202	q_0	.0912717631143501
p_1	−.0337282368857499	q_1	−.0575435262285002
p_2	−.1564465330579805	q_2	−.5912717631142501
p_3	.5337282368857499	q_3	1.1150870524570004
p_4	1.2058980364727207	q_4	−.5912717631142501
p_5	.5337282368857499	q_5	−.0575435262285002
p_6	−.1564465330579805	q_6	.0912717631142501
p_7	−.0337282368857499		
p_8	.0534975148216202		
		$\widetilde{q}_{-1}$	.0534975148216202
$\widetilde{p}_0$	.0	$\widetilde{q}_0$	.0337282368857499
$\widetilde{p}_1$	−.0912717631142501	$\widetilde{q}_1$	−.1564465330579805
$\widetilde{p}_2$	−.0575435262285002	$\widetilde{q}_2$	−.5337282368857499
$\widetilde{p}_3$	.5912717631142501	$\widetilde{q}_3$	1.2058980364727207
$\widetilde{p}_4$	1.1150870524570004	$\widetilde{q}_4$	−.5337282368857499
$\widetilde{p}_5$	.5912717631142501	$\widetilde{q}_5$	−.1564465330579805
$\widetilde{p}_6$	−.0575435262285002	$\widetilde{q}_6$	.0337282368857499

Prewavelets

In this section, we first construct a compactly supported pre-wavelet ψ_n associated with B-spline $\phi_n = N_n$ of order n. After that we describe a general approach to construct compactly supported pre-wavelets from any refinable function ϕ who generates a multiresolution approximation of $L_2(\mathbb{R})$.

It is known that the B-spline N_n of order $n \geq 1$ generates a multiresolution approximation of $L_2(\mathbb{R})$ (cf. [12]). Let

$$V_0 = \mathrm{span}_{L_2(\mathbb{R})}\{N_n(x-k)\,, \quad k \in \mathbb{Z}\}$$

and $V_j = \{f(2^j\cdot), \forall f \in V_0\}$. Since N_n is a refinable function, $V_j \subset V_{j+1}$. Let W_j be the orthogonal complement of V_j in V_{j+1}. If a function $\psi_n \in W_j$ such that $W_j = \mathrm{span}_{L_2(\mathbb{R})}\{\psi_n(x-m), m \in \mathbb{Z}\}$, then ψ_n is called a prewavelet associated with N_n.

According to [18], let

$$\psi_n(x) = \frac{1}{2^{n-1}} \sum_{i=0}^{2n-2} (-1)^i N_{2n}(i+1) N_{2n}^{(n)}(2x-j)\,, \quad (18)$$

where N_{2n} is the uniform B-spline of order $2n$. Here $N_{2n}^{(n)}$ denotes the nth derivative of $N_{2n}(x)$ which is a linear combination of B-splines $N_n(x-k)$ for finitely many k. In fact,

$$N_{2n}^{(n)}(x) = \sum_{i=0}^{n} (-1)^i \binom{n}{i} N_n(x-j)\,.$$

It follows that

$$\psi_n(x) = \sum_{i=0}^{3n-2} \frac{(-1)^i}{2^{n-1}} \cdot \sum_{k=0}^{n} \binom{n}{k} N_{2n}(i-k+1) N_n(2x-i) \in V_1\,. \quad (19)$$

We can verify that ψ_n is a pre-wavelet associated with N_n (cf. [18]).

Theorem 7 *The spline function $\psi_n(x)$ defined in (18) with support $[0, 2n-1]$ is a pre-wavelet associated with the B-spline $N_n(t)$ of order n.*

Let us determine two sequences $\{g_n\}$ and $\{h_n\}$ such that

$$N_n(2x-k) = \sum_{m\in\mathbb{Z}} g_{k-2m} N_n(x-m) + \sum_{m\in\mathbb{Z}} h_{k-2m} \psi_n(x-m) \quad (20)$$

for all $k \in \mathbb{Z}$.

Let G and H be the z-transform of the sequences $\{g_i\}_{i\in\mathbb{Z}}$ and $\{h_i\}_{i\in\mathbb{Z}}$. That is,

$$G(z) = \sum_{m\in\mathbb{Z}} g_m z^{-m} \quad \text{and} \quad H(z) = \sum_{m\in\mathbb{Z}} h_m z^{-m} .$$

As we know from a previous section,

$$\widehat{N_n}(\omega) = \frac{1}{2} P(\omega)\widehat{N_n}(\omega/2) .$$

We find

$$\begin{aligned}
\widehat{\psi_n}(\omega) &= \int_{-\infty}^{\infty} \psi_n(x) \mathrm{e}^{-ix\omega} \mathrm{d}x \\
&= 2^{-n+1} \sum_{i=0}^{2n-2} (-1)^i N_{2n}(i+1) \\
&\quad \int_{-\infty}^{\infty} N_{2n}^{(n)}(2x-i) \mathrm{e}^{-ix\omega} \mathrm{d}x \\
&= 2^{-n} \sum_{j=0}^{2n-2} (-1)^j N_{2n}(j+1) \mathrm{e}^{-ij\omega/2} \\
&\quad \int_{-\infty}^{\infty} N_{2n}^{(n)}(x) \mathrm{e}^{-ix\omega/2} \mathrm{d}x \\
&= 2^{-n} \sum_{j=0}^{2n-2} (-1)^j N_{2n}(j+1) \mathrm{e}^{-ij\omega/2} (i\omega/2)^n \\
&\quad \int_{-\infty}^{\infty} N_{2n}(x) \mathrm{e}^{-ix\omega/2} \mathrm{d}x \\
&= 2^{-n} \sum_{j=0}^{2n-2} (-1)^j N_{2n}(j+1) \mathrm{e}^{-ij\omega/2} \\
&\quad (1-\mathrm{e}^{-i\omega/2})^n \hat{N}_n(\omega/2) \\
&= 2^{-n} \sum_{j=0}^{2n-2} (-1)^j N_{2n}(j+1) z^j (1-z)^n \hat{N}_n(\omega/2) \\
&= \frac{1}{2} Q(z) \widehat{N_n}(\omega/2)
\end{aligned}$$

where $z = \mathrm{e}^{-i\omega/2}$ and

$$Q(z) = 2^{-n+1}(1-z)^n \sum_{j=0}^{2n-2} (-1)^j N_{2n}(j+1) z^j .$$

Note that the Fourier transform of (20) and the orthogonality between W_0 and V_0 can be recast in the following matrix format:

$$\begin{cases} P(z)G(z) + Q(z)H(z) & = 2 \\ P(-z)G(z) + Q(-z)H(z) & = 0 . \end{cases}$$

The solution are

$$\begin{aligned}
G(z) &= \frac{2Q(-z)}{P(z)Q(-z) - P(-z)Q(z)} ; \\
H(z) &= \frac{-2P(-z)}{P(z)Q(-z) - P(-z)Q(z)} .
\end{aligned}$$

We have to verify that $P(z)Q(-z) - P(-z)Q(z) \neq 0$ for all z. This follows easily from the following (cf. [18])

Lemma 3

$$P(z)Q(-z) - P(-z)Q(z) = \frac{2^{2n} z \tau_{2n-1}(z^2)}{2^{2n-2}(2n-1)!} .$$

where τ_{2n-1} is the Euler–Frobenius polynomial which is defined by

$$\tau_n(t) := n! \sum_{i=0}^{n-1} N_{n+1}(i+1) t^i , \quad \forall n \geq 0 ,$$

which is never zero for z (cf. [83]).

Therefore, ϕ is a pre-wavelet associated with $N_n(x)$. It follows from its definition (19) that ψ_n is of compact support.

Example 12 Consider $m = 1$. Then

$$N_1(x) = \begin{cases} 1 & x \in [0,1) \\ 0 & \text{otherwise} \end{cases}$$

and

$$\psi_1(x) = N_2(1) N_2'(2x) = N_2'(2x) = \begin{cases} 1 & x \in [0, \frac{1}{2}) \\ -1 & x \in [\frac{1}{2}, 1] \end{cases}$$

which is the Haar wavelet.

Example 13 Consider $m = 2$.

$$N_2(x) = \begin{cases} x, & x \in [0,1] \\ 2-x, & x \in [1,2] \\ 0, & \text{otherwise} . \end{cases}$$

Then

$$\begin{aligned}
\psi_2(x) &= \frac{1}{2} \sum_{j=0}^{2} N_4(j+1) N_4^{(2)}(2x-j) \\
&= \frac{1}{12} (N_4^{(2)}(2x) - 4N_4^{(2)}(2x-1) + N_4^{(2)}(2x-2)) \\
&= \frac{1}{12} (N_2(2x) - 6N_2(2x-1) + 10N_2(2x-2) \\
&\quad - 6N_2(2x-3) + N_2(2x-4)) .
\end{aligned}$$

We remark that both $G(z)$ and $H(z)$ are rational filters. As pointed in Sect. "Definition of Filters", these are realizable (recursive) filters which can be used for digital signal reconstruction procedure. On the other hand, the zeros $\lambda_1, \ldots, \lambda_{n-1}$ of $\tau_n(z)$ are located in $(-\infty, 0)$ and satisfy $\lambda_{n-1}\lambda_1 = \lambda_{n-2}\lambda_2 = \ldots = 1$ (cf. [83]). Thus, G and H are not stable IIR filters.

We now consider a new constructive method for finding compactly supported prewavelet function for any given refinable function ϕ which generates a multiresolution approximation of $L_2(\mathbb{R})$. Let $V_j = \mathrm{span}_{L_2(\mathbb{R})}\{\phi(2^j x - k), k \in \mathbb{Z}\}$ for all $j \in \mathbb{Z}$. We are looking for compactly supported functions ψ in V_1 such that

$$V_1 = V_0 \bigoplus W_0 ,$$

where W_0 is the closure of the linear span of integer translates of $\psi(2^j x - m), m \in \mathbb{Z}$ and $\psi(\cdot - m), m \in \mathbb{Z}$ form a stable basis for W_0.

To do so, we first introduce a function

$$\Phi(z) := \sum_{m\in\mathbb{Z}} \langle \phi(x), \phi(x-m)\rangle z^m ,$$

where $\langle f, g\rangle$ stands for the standard inner product for $L_2(\mathbb{R})$. This function Φ may be called the generalized Euler–Frobenius polynomial.

Next we need a necessary and sufficient condition for the orthogonality. Writing

$$g(x) = \sum_{m\in\mathbb{Z}} c_m 2^{1/2}\phi(2x-m) \in V_1 ,$$

to be a general function in V_1 and

$$G(z) = \frac{1}{\sqrt{2}} \sum_{m\in\mathbb{Z}} c_m z^m ,$$

i. e., the Fourier transform $\widehat{g}(\omega) = G(\mathrm{e}^{i\omega/2})\widehat{\phi}(\omega/2)$, we let

$$\mathcal{G} = \mathrm{closure}_{L_2(\mathbb{R})}\{g(x-m), \quad m \in \mathbb{Z}\}$$

be the closure of the linear span of integer translates of g. Then we have the following (see, e. g., [49] for a proof)

Theorem 8 *Let $P(z)$ be the mask polynomial of ϕ. That is, $\widehat{\phi}(\omega) = P(e^{i\omega/2})\widehat{\phi}(\omega/2)$. Then $\mathcal{G}$ is orthogonal to V_0 if and only if*

$$G(z)\overline{P(z)}\Phi(z) + G(-z)\overline{P(-z)}\Phi(-z) = 0 .$$

Our first step is to construct compactly supported $g_k \in V_1, k = 1, 2$ such that the closure $\mathcal{G}_k$ of the linear span of integer translates g_k is orthogonal to V_0 for $k = 1$ and $k = 2$ and $V_1 = V_0 \oplus (\mathcal{G}_1 + \mathcal{G}_2)$. The second step is to find a sufficient condition that one of them can be written in terms of integer translates of the other. For example, $\mathcal{G}_2$ is contained in $\mathcal{G}_1$. In this case we will have $V_1 = V_j \oplus \mathcal{G}_1$. To be more precise, we suppose that $g_k \in V_1$ satisfy

$$g_k(x-m) \perp V_0, \quad m \in \mathbb{Z}$$

for $k = 1, 2$ and

$$2^{1/2}\phi(2x) = \sum_{m\in\mathbb{Z}} \big(a_{1,m}\phi(x-m) + b_{1,m}g_1(x-m)\big)$$
$$2^{1/2}\phi(2x-1) = \sum_{m\in\mathbb{Z}} \big(a_{2,m}\phi_j(x-m) + b_{2,m}g_2(x-m)\big) .$$

In terms of Fourier transform, the above equations can be rewritten as

$$\begin{aligned}\frac{1}{2^{1/2}}\widehat{\phi}(\frac{\omega}{2}) &= A_1(\omega)\widehat{\phi}(\omega) + B_1(\omega)\widehat{g_1}(\omega)\\ &= A_1(\omega)P\left(\frac{\omega}{2}\right)\widehat{\phi}\left(\frac{\omega}{2}\right)\\ &\quad + B_1(\omega)G_1\left(\frac{\omega}{2}\right)\widehat{\phi}\left(\frac{\omega}{2}\right)\end{aligned}$$

and

$$\begin{aligned}\frac{1}{2^{1/2}}\mathrm{e}^{i\omega/2}\widehat{\phi}\left(\frac{\omega}{2}\right) &= A_2(\omega)\widehat{\phi}(\omega) + B_2(\omega)\widehat{g_1}(\omega)\\ &= A_2(\omega)P\left(\frac{\omega}{2}\right)\widehat{\phi}\left(\frac{\omega}{2}\right)\\ &\quad + B_2(\omega)G_2\left(\frac{\omega}{2}\right)\widehat{\phi}\left(\frac{\omega}{2}\right) ,\end{aligned}$$

where $A_k(\omega) = \sum_{m\in\mathbb{Z}} a_{k,m}\mathrm{e}^{im\omega}$ and $B_k(\omega) = \sum_{m\in\mathbb{Z}} b_{k,m}\mathrm{e}^{im\omega}$. Here, we have abused the notation of G_k, that is, we use $G_k(\omega)$ instead of $G_k(z)$ with $z = \mathrm{e}^{i\omega}$ just for convenience.

It follows that

$$A_1(2\omega)P(\omega) + B_1(2\omega)G_1(\omega) = 1,$$
$$A_2(2\omega)P(\omega) + B_2(2\omega)G_2(\omega) = \mathrm{e}^{i\omega} .$$

Using Theorem 8, the solution of A_k, B_k and G_k can be easily found. Indeed, let E be the operator which maps a Laurent polynomial f into a Laurent polynomial $E(f)$ which contains all the even index terms of f. That is, $E(f) = (f(z) + f(-z))/2$. One simple property of E is $E(f(z)) = E(f(-z))$.

Theorem 9 *Suppose that $E(P(z)\overline{P(z)}(z)) \neq 0$ for all z with $|z| = 1$. Let*

$$A_1(2\omega) := \frac{E(\overline{P(z)}\Phi(z))}{E(P(z)\overline{P(z)}\Phi(z))},$$
$$B_1(2\omega) := \frac{1}{E(P(\omega)\overline{P(\omega)}\Phi(z))},$$
$$G_1(\omega) := E(P(\omega)\overline{P(\omega)}\Phi(z)) - E(\overline{P(\omega)}\Phi(z))P(\omega)$$

and

$$A_2(2\omega) := \frac{E(e^{i\omega}\overline{P(\omega)}\Phi(z))}{E(P(\omega)\overline{P(\omega)}\Phi(z))},$$
$$B_2(2\omega) := \frac{1}{E(P(\omega)\overline{P(\omega)}\Phi(z))},$$
$$G_2(\omega) := E(P(\omega)\overline{P(\omega)}\Phi(z))e^{i\omega} - E(e^{i\omega}\overline{P(\omega)}\Phi(z))P(\omega)\,.$$

Then $\mathcal{G}_k$ is orthogonal to V_0 for all $k = 1, 2$ and

$$V_1 = V_0 \bigoplus (\mathcal{G}_1 + \mathcal{G}_2)\,.$$

Proof Using the assumption of Theorem 9, we know that A_k, B_k are well-definited. It is clear that V_1 is the direct sum of V_0 and $\mathcal{G}_k$, $k = 1, 2$. To see $\mathcal{G}_k$ is orthogonal to V_0, we use Theorem 8 to see $E(G_k(z)\overline{P(z)}\Phi(z)) = 0$. Since $B_k(2\omega) \neq 0$ and

$$E(B_k(2\omega)G_k(z)\overline{P(z)}\Phi(z)) = B_k(2\omega)E(G_k(z)\overline{P(z)}\Phi(z))\,,$$

we may consider

$$\begin{aligned} &E(B_1(2\omega)G_1(z)\overline{P(z)}\Phi(z)) \\ &\quad = E((1 - A_1(2\omega)P(z))\overline{P(z)}\Phi(z)) \\ &\quad = E(\overline{P(z)}\Phi(z)) - A_1(2\omega)E(P(z)\overline{P(z)}\Phi(z)) = 0 \end{aligned}$$

by the construction of A_1. Similar for the second equation. This completes the proof. □

Let us make a remark on $E(P(\omega)\overline{P(\omega)}\Phi(z))$. The following result is known (cf. [12]).

Lemma 4

$$E(P(\omega)\overline{P(\omega)}\Phi(z)) = \frac{1}{2}\Phi(z^2)\,.$$

Next we show that g_k, $k = 1, 2$ are linearly dependent if P satisfies another condition. Let us write P in its polyphase form, i. e.,

$$P(z) = P_0(z^2) + zP_1(z^2)\,.$$

Theorem 10 *Suppose that $P_0(z)$ is not zero for any z with $|z| = 1$. Then there exist non-zero coefficients f_m's such that*

$$g_1(x) = \sum_{m\in\mathbb{Z}} f_m g_2(x - m)\,.$$

Furthermore, the integer translates of g_2 form a Riesz basis for $W_0 := V_1 \ominus V_0$ which is the closure of the linear span of $g_2(2^j \cdot -m)$, $m \in \mathbb{Z}$. If $P_1(z) \neq 0$ for all z, then

$$g_2(x) = \sum_{m\in\mathbb{Z}} f'_{j,m} g_1(x - m)$$

for some coefficients f'_m and the integer translates of g_1 form a Riesze basis for W_0.

Without loss of generality, we may assume that P_1 is not zero. The letting $\psi = g_1$ and $W_0 := \mathcal{G}_1$, ψ is a prewavelet associated with ϕ. It is easy to see that ψ is compactly support if ϕ is since $\psi = g_1$ is a finitely linear combination of integer translates of ϕ because

$$\widehat{g_1}(\omega) = G_1(\omega)\widehat{\phi}(\omega)$$

with Laurent polynomial $G_1(\omega)$. Let $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$. We can verify that all these ψ_{jk} form a Riesz basis for $L_2(\mathbb{R})$ (cf. [49]). We have

Theorem 11 *Suppose that ϕ is a compactly supported refinable function generating an MRA for $L_2(\mathbb{R})$. Denote $\widehat{\phi}(\omega) = P(\omega/2)\widehat{\phi}(\omega/2)$. Let Φ be the generalized Euler–Frobenius polynomial associated with ϕ. Suppose that*

$$E(P(\omega)\overline{P(\omega)}\Phi(\omega)) \neq 0, \quad \forall e^{i\omega}$$

and suppose that at least one of the two polyphases of P is not zero, i. e., $P_0(\omega) \neq 0$ or $P_1(\omega) \neq 0$ for all $\omega \in [0, 2\pi]$. Then there exist a compactly supported function ψ such that the closure W_0 of the linear span of integer translates $\psi(x - m)$, $m \in \mathbb{Z}$ is orthogonal to V_0, $V_1 = V_0 \bigoplus W_0$ and the integer translates of ψ form a Riesz basis for W_0. All of them forms a Riesz basis for $L_2(\mathbb{R})$.

Let us use B-splines to give some examples of prewavelets for $L_2(\mathbb{R})$. Recall B-splines from (12). Fix an integer $n > 0$. Let $\phi = N_n$ the n-th order B-spline. It is easy to see

$$\widehat{N_n}(\omega) = P^{(n)}(\omega/2)\widehat{N_n}(\omega/2)$$

with $P^{(n)}(\omega) = \left(\frac{1+e^{-i\omega}}{2}\right)^n$.

It is clear that

$$P^{(n)}(\omega) - P^{(n)}(\omega + \pi) = \left(\frac{1 + e^{-i\omega}}{2}\right)^n - \left(\frac{1 - e^{-i\omega}}{2}\right)^n \neq 0$$

for any $\omega \in [0, 2\pi]$. It follows that the polyphase $P_1^{(n)}$ associated with ϕ is never zero.

Next by Lemma 4, we know

$$E(P^{(n)}(\omega)\overline{P^{(n)}(\omega)}\Phi(\omega)) = \frac{1}{2}\Phi(2\omega)\,.$$

By Poisson summation formula,

$$\Phi(\omega) = \sum_{m\in\mathbb{Z}} \langle N_n(\cdot), N_n(\cdot - m)\rangle e^{-im\omega} = \sum_{m\in\mathbb{Z}} |\widehat{N_n}(\omega + 2m\pi)|^2$$

which is never zero (cf. [83]). That is, the conditions in Theorem 11 are satisfied.

The above discussions verify that all B-spline functions can be used to construct prewavelets for $L_2(\mathbb{R})$. Note that our B-spline prewavelets have a larger support than those constructed in the first half of this section. The purpose of the examples is to show the detail of the constructive procedure. The advantages of the second construction are (1). it enables us to construct prewavelets from any refinable functions and (2). the construction can be easily generalized to the multivariate setting to be discussed later. Note that the first construction has no multivariate generalization.

Example 14 Consider linear B-spline N_2 with $P^{(2)}(z) = (1+z)^2/4$. It is easy to see that

$$\Phi(z) = \frac{1}{6}z^{-1} + \frac{4}{6} + \frac{1}{6}z\,.$$

Indeed, by using the symmetric property of B-splines, i. e., $N_2(x) = N_2(2-x)$. It is easy to see

$$\Phi(z) = \sum_{m\in\mathbb{Z}} N_4(2+m)z^m = \frac{1}{24z^2}(10z^2 + z^4 + 1)\,.$$

Thus, we know $E(P^{(2)}(z)P^{(2)}(1/z)\Phi(z)) = \frac{1}{2}\Phi(z^2) \neq 0$. Using the formulas in Theorem 9, we have

$$G(z) = \frac{1}{96z^2}(z^6 - 6z^5 + 11z^4 - 12z^3 + 11z^2 - 6z + 1)\,.$$

That is, the prewavelet associated with linear B-spline $\phi := N_2$ is

$$\psi(x) = \frac{1}{96}(\phi(2x+2) - 6\phi(2x+1) + 11\phi(2x) - 12\phi(2x-1) + 11\phi(2x-2) - 6\phi(2x-3) + \phi(2x-4))\,.$$

Example 15 Consider cubic B-spline N_4 with $P^{(4)}(z) = (1+z)^4/16$. We have

$$\Phi(z) = \frac{1}{5040z^3} + \frac{1}{42z^2} + \frac{397}{1680z} + \frac{151}{315} + \frac{397}{1680}z + \frac{1}{42}z^2 + \frac{1}{5040}z^3.$$

Thus, $E(P^{(2)}(z)P^{(2)}(1/z)\Phi(z)) = \frac{1}{2}\Phi(z^2) \neq 0$.

$$E(zP^{(4)}(1/z)\Phi(z)) = \frac{1}{80640z^6}(18482z^4 + z^{10} + 18482z^6 + 1677z^2 + 1 + 1677z^8).$$

Thus, by using the formula in Theorem 9, $G(z) = \sum_{k=-6}^{8} g_k z^k$ with coefficients g_k as follows:

$$g_{-6} = \frac{-1}{1290240}, \quad g_{-5} = \frac{31}{322560}, \quad g_{-4} = \frac{-187}{143360}$$
$$g_{-3} = \frac{1081}{161280}, \quad g_{-2} = \frac{-1903}{86016}, \quad g_{-1} = \frac{17953}{322560},$$
$$g_0 = \frac{-131051}{1290240}, \quad g_1 = \frac{1441}{11520}, \quad g_2 = \frac{-131051}{1290240}$$
$$g_3 = \frac{17953}{322560}, \quad g_4 = \frac{-1903}{86016}, \quad g_5 = \frac{1081}{161280},$$
$$g_6 = \frac{-187}{143360}, \quad g_7 = \frac{3}{322560}, \quad g_8 = \frac{-1}{1290240}\,.$$

That is, the prewavelet associated with cubic B-spline $\phi := N_4$ is

$$\begin{aligned}\psi(x) = \frac{1}{1290240}(&\phi(2x+6) - 124\phi(2x+5) \\ &+ 1683\phi(2x+4) - 8648\phi(2x+3) \\ &+ 28545\phi(2x+2) - 71812\phi(2x+1) \\ &+ 131051\phi(2x)) - 161392\phi(2x-1) \\ &+ 131051\phi(2x-2) - 71812\phi(2x-3) \\ &+ 28545\phi(2x-4) - 8648\phi(2x-5) \\ &+ 1683\phi(2x-6) - 124\phi(2x-7) + \phi(2x-8))\,.\end{aligned}$$

We can easily verify that ψ is orthogonal to the integer translates of ϕ using the computer program MAPLE.

Tight Wavelet Frames

In this section we consider the tight wavelet frames. First of all, let us introduce some notation and explain their usefulness for representation of functions in $L_2(\mathbb{R})$. We will show the stricking feature of tight wavelet frames: their representation of any function $f \in L_2(\mathbb{R})$ is just like that by using an orthonormal wavelet basis although they can contain redundancy. Also, the representation is most economic in the sense that the ℓ^2 norm of the coefficients in the representation is the same as the L_2 norm of the function.

Definition 2 A family of functions $\{\phi_j, j \in J\}$ in $L_2(\mathbb{R})$ is called a *frame*, if there exist constants $A > 0, B > 0$ such that

$$A\|f\|^2 \le \sum_{j\in J} |\langle f, \phi_j\rangle|^2 \le B\|f\|^2, \quad \forall f \in L_2(\mathbb{R})\,,$$

where $\|f\|$ denotes the norm of f in $L_2(\mathbb{R})$. If $A = B$, $\{\phi_j\}_{j\in J}$ is called *tight wavelet frame*.

Thus for a tight wavelet frame we may normalize so that $A = B = 1$. In this case, we will have

$$\|f\|^2 = \sum_{j\in J} |\langle f, \phi_j\rangle|^2, \quad \forall f \in L_2(\mathbb{R}) . \tag{21}$$

Thus for any f and g in $L_2(\mathbb{R})$, we have

$$\|f+g\|^2 = \sum_{j\in J} |\langle f+g, \phi_j\rangle|^2,$$
$$\|f-g\|^2 = \sum_{j\in J} |\langle f-g, \phi_j\rangle|^2 .$$

Differencing the two equations, we have

$$4\langle f, g\rangle = 4\sum_{j\in J} \langle f, \phi_j\rangle\langle \phi_j, g\rangle \tag{22}$$

for all $g \in L_2(\mathbb{R})$. That is,

$$f = \sum_{j\in J} \langle f, \phi_j\rangle \phi_j, \quad \text{weakly} \quad \forall f \in L_2(\mathbb{R}) . \tag{23}$$

Thus, a tight wavelet frame can represent any $f \in L_2(\mathbb{R})$ just like an orthonormal basis. Note that the representation (23) is simply derived from (21) and the technique is called polarization.

Compared to the orthonormal wavelet basis representation of functions in $L_2(\mathbb{R})$, a tight wavelet frame needs no orthonormality nor linear independence among the functions ϕ_j, $j \in J$. In fact, it allows redundancy in ϕ_j, $j \in J$. Thus, we have more degrees of freedom to construct tight wavelet frames. In the following we shall present three methods.

In general we shall construct a finitely many compactly supported functions $\psi_\ell \in L_2(\mathbb{R})$ such that

$$\Lambda(\psi_\ell, \ell = 1, \ldots, r) = \{\psi_{\ell:j,k}(x) := 2^{j/2}\psi_\ell(2x-k) , \quad j, k \in \mathbb{Z} , \quad \ell = 1, 2, \ldots, r\}$$

is a tight wavelet frame.

We start with a refinable function $\phi \in L_2(\mathbb{R})$. Let $P(\omega)$ be the mask polynomial of ϕ. That is,

$$\widehat{\phi}(\omega) = P(\omega/2)\widehat{\phi}(\omega/2) .$$

Suppose that ϕ generates a multiresolution approximation of $L_2(\mathbb{R})$. For Laurent polynomials $Q_\ell(\omega)$, $\ell = 1, \ldots, r$, let ψ_ℓ be a function in $L_2(\mathbb{R})$ defined in terms of Fourier transform by

$$\widehat{\psi_\ell}(\omega) = Q_\ell(\omega/2)\widehat{\phi}(\omega/2), \quad \ell = 1, \ldots, r .$$

The following condition is called Unitary Extension Principle (UEP) developed in [76].

$$P(\omega)\overline{P(\omega+\nu)} + \sum_{\ell=1}^{r} Q^\ell(\omega)\overline{Q^\ell(\omega+\nu)} = \begin{cases} 1 & \text{if} \nu = 0 , \\ 0 & \text{if} \nu = \pi . \end{cases} \tag{24}$$

Theorem 12 *Suppose that $\phi \in L_2(\mathbb{R})$ is a compactly supported continuous function with Hölder continuity $\alpha > 0$ and suppose that ϕ generates a multiresolution approximation of $L_2(\mathbb{R})$. Assume that there exist Laurent polynomials Q_ℓ, $\ell = 1, \ldots, r$ such that P and Q_ℓ satisfy the UEP condition. Then defining ψ_ℓ as above, $\Lambda(\psi_\ell, \ell = 1, \ldots, r)$ is a tight wavelet frame for $L_2(\mathbb{R})$.*

The proof of the above theorem can be derived based on the following three lemmas.

Lemma 5 *Let $\phi \in L_2(\mathbb{R})$. Suppose that $\widehat{\phi}(0) = 1$ and that for some constant $B > 0$,*

$$\sum_{m\in 2\pi\mathbb{Z}} |\widehat{\phi}(\omega+m)|^2 \le B < +\infty, \quad a.\ e., \quad \omega \in \mathbb{R} .$$

Define

$$\beta_j(f, \omega) = 2^{j/2} \sum_{m\in 2\pi\mathbb{Z}} \widehat{f}(2^j(\omega+m))\overline{\widehat{\phi}(\omega+m)}$$

for any fixed $f \in L_2(\mathbb{R})$ and $j \in \mathbb{Z}$. Then

$$\lim_{j\to+\infty} \int_{[0,2\pi]} |\beta_j(f,\omega)|^2 \, d\omega = \|f\|^2 \quad \textit{and}$$
$$\lim_{j\to-\infty} \int_{[0,2\pi]} |\beta_j(f,\omega)|^2 \, d\omega = 0 .$$

Lemma 6 *For the refinable function ϕ satisfying the same condition in Lemma 5 and ψ_ℓ defined above, we have the following equations:*

$$\sum_{k\in\mathbb{Z}} |\langle f, \phi_{j,k}\rangle|^2 = \int_{[0,2\pi]} |\beta_j(f,\omega)|^2 \, d\omega$$

$$\sum_{k\in\mathbb{Z}} |\langle f, \phi_{j-1,k}\rangle|^2 = \frac{1}{2}\int_{[0,2\pi]} \left| \sum_{\nu\in\{0,1\}\pi} \overline{P\left(\frac{\omega}{2}+\nu\right)} \beta_j\left(f, \frac{\omega}{2}+\nu\right) \right|^2 d\omega$$

$$\sum_{k\in\mathbb{Z}} |\langle f, \psi^\ell_{j,k}\rangle|^2 = \frac{1}{2}\int_{[0,2\pi]} \left| \sum_{\nu\in\{0,1\}\pi} \overline{Q_\ell\left(\frac{\omega}{2}+\nu\right)} \beta_j\left(f, \frac{\omega}{2}+\nu\right) \right|^2 d\omega$$

where $\phi_{j,k}(\cdot) = 2^{j/2}\phi(2^j \cdot -k)$, $\psi_{\ell:j,k}(\cdot) = 2^{j/2}\psi_\ell(2^j \cdot -k)$.

Lemma 7 *Suppose that we can find* Q_ℓ, $\ell = 1, \ldots, r$ *satisfying (24). Let* ψ_ℓ *be the functions defined by its Fourier transform using* $\widehat{\psi_\ell}(\omega) = Q_\ell(\omega/2)\widehat{\phi}(\omega/2)$. *Then the collection of functions* $\{\psi_{\ell:j,k} : \ell = 1, \ldots, r, j, k \in \mathbb{Z}\}$ *is a tight wavelet frame.*

We now rewrite the UEP in a matrix format.

Lemma 8 *Let* $\mathcal{P} = [P(\omega), P(\omega+\pi)]$ *be a vector and* $\mathcal{Q} = [Q_\ell(\omega), Q_\ell(\omega+\pi)]_{\ell=1,\ldots,r}$ *be a matrix of size* $r \times 2$. *Then the UEP condition in (24) is equivalent to*

$$\mathcal{Q}^*\mathcal{Q} = I_2 - \mathcal{P}\mathcal{P}^* . \tag{25}$$

Here I_2 *is the identity matrix of size* 2×2.

Multiplying $\mathcal{P}^*$ from the left of (25) and $\mathcal{P}$ from the right, we immediately get

$$\mathcal{P}^*\mathcal{P} - (\mathcal{P}^*\mathcal{P})^2 \geq 0 .$$

It follows that

$$\mathcal{P}^*\mathcal{P} \leq 1 \quad \text{or} \quad |P(\omega)|^2 + |P(\omega+\pi)|^2 \leq 1 . \tag{26}$$

That is, P is necessary to satisfy the above (26) in order to have the associated tight wavelet frame. On the other hand, if a mask polynomial P satisfies (26), we now show that there is a set of frame generators $\psi_\ell, \ell = 1, 2, 3$ such that their dilations and integer translates form a tight wavelet frame for $L_2(\mathbb{R})$. That is, the condition (26) is necessary and sufficient to have a tight wavelet frame associated with the refinable function ϕ (cf. [14] and [72]).

Indeed, let $P(\omega) = P_1(2\omega) + zP_2(2\omega)$ be the polyphase format of P, where $z = \mathrm{e}^{i\omega}$. Since

$$|P(\omega)|^2 + |P(\omega+\pi)|^2 = 2|P_1(2\omega)|^2 + 2|P_2(2\omega)|^2 \leq 1, \tag{27}$$

we let $P_3(\omega)$ be a polynomial in z such that

$$2|P_1(2\omega)|^2 + 2|P_2(2\omega)|^2 + |P_3(2\omega)|^2 = 1 . \tag{28}$$

This can be done using Riesz–Fejér factorization. We now define a matrix of 3×3 by

$$\widetilde{Q}(\omega) = I_3 - \widetilde{P}(\omega)\widetilde{P}(\omega)^*$$

with $\widetilde{P}(\omega) = [\sqrt{2}P_1(\omega), \sqrt{2}P_2(\omega), P_3(\omega)]^{\mathrm{T}}$ being a vector of 3×1. It is easy to check that

$$\widetilde{Q}(\omega)^*\widetilde{Q}(\omega) = I_3 - \widetilde{P}(\omega)\widetilde{P}(\omega)^*$$

by using (28). If we choose the top 2×2 principal block from the right-hand side of the above equation, we will have, letting $\widetilde{Q} = [\tilde{Q}_{jk}]_{1\leq j,k\leq 3}$,

$$\left[\tilde{Q}_{jk}\right]^*_{\substack{1\leq j\leq 3\\1\leq k\leq 2}} \left[\tilde{Q}_{jk}\right]_{\substack{1\leq j\leq 3\\1\leq k\leq 2}} = I_2 - \begin{bmatrix}\sqrt{2}P_1(\omega)\\ \sqrt{2}P_2(\omega)\end{bmatrix} \begin{bmatrix}\sqrt{2}P_1(\omega)^* & \sqrt{2}P_2(\omega)^*\end{bmatrix} .$$

Replacing ω by 2ω in the above equation, multiplying the following unitary matrix from the left of the equation

$$U := \frac{1}{\sqrt{2}}\begin{bmatrix}1 & z\\ 1 & -z\end{bmatrix}$$

and multiplying U^* from the right of the equation, we obtain

$$U\left[\tilde{Q}_{jk}\right]^*_{\substack{1\leq j\leq 3\\1\leq k\leq 2}} \left[\tilde{Q}_{jk}\right]_{\substack{1\leq j\leq 3\\1\leq k\leq 2}} U^* = I_2 - \begin{bmatrix}P(\omega)\\ P(\omega+\pi)\end{bmatrix} \begin{bmatrix}P(\omega)^* & P(\omega+\pi)^*\end{bmatrix} .$$

Then $Q_\ell(\omega) = \sqrt{2}(\tilde{Q}_{\ell,1}(2\omega) + z\tilde{Q}_{\ell,2}(2\omega))/2$ for $\ell = 1, 2, 3$ together with $P(\omega)$ satisfy the matrix format UEP (cf. Lemma 8), and thus ψ_ℓ defined by using these Q_ℓ are tight wavelet frame generators.

Another method to construct tight wavelet frames is to do matrix extension. That is, let $\widetilde{P}$ be the vector of size 3×1 defined above. Note that $\widetilde{P}^*\widetilde{P} = 1$. By a method in [62], one can find $\widetilde{Q}_1 = [\tilde{Q}_{11}, \tilde{Q}_{12}, \tilde{Q}_{13}]^{\mathrm{T}}$ and $\widetilde{Q}_2 = [\tilde{Q}_{21}, \tilde{Q}_{22}, \tilde{Q}_{23}]^{\mathrm{T}}$, two vectors of size 3×1 such that the matrix $\left[\widetilde{P}\ \widetilde{Q}_1\ \widetilde{Q}_2\right]$ is unitary. See also [14] for detail construction. Once we have such unitary extension, we know that the first two rows are unitary. Rewriting the row vectors in column format, we have

$$\begin{bmatrix}\sqrt{2}P_1(\omega) & \sqrt{2}P_2(\omega)\\ \tilde{Q}_{11}(\omega) & \tilde{Q}_{12}(\omega)\\ \tilde{Q}_{21}(\omega) & \tilde{Q}_{22}(\omega)\end{bmatrix}^* \begin{bmatrix}\sqrt{2}P_1(\omega) & \sqrt{2}P_2(\omega)\\ \tilde{Q}_{11}(\omega) & \tilde{Q}_{12}(\omega)\\ \tilde{Q}_{21}(\omega) & \tilde{Q}_{22}(\omega)\end{bmatrix} = I_2.$$

Substituting ω by 2ω in the above equation and multiplying $\overline{U}$ from the right and $\overline{U}^*$ from the left of the above equation both sides, we obtain

$$\begin{bmatrix}P(\omega) & P(\omega+\pi)\\ Q_1(\omega) & Q_1(\omega+\pi)\\ Q_2(\omega) & Q_2(\omega+\pi)\end{bmatrix}^* \begin{bmatrix}P(\omega) & P(\omega+\pi)\\ Q_1(\omega) & Q_1(\omega+\pi)\\ Q_2(\omega) & Q_2(\omega+\pi)\end{bmatrix} = I_2$$

which is just an another form of the UEP (see (24)). Here $Q_\ell(\omega) = \tilde{Q}_{\ell,1}(2\omega) + z\tilde{Q}_{\ell,2}(2\omega)$ for $\ell = 1$ and $\ell = 2$. Hence, we can define two tight wavelet frame generators ψ_ℓ using the above Q_ℓ for $\ell = 1, 2$.

We now discuss a new method how to construct tight wavelet frames. Let

$$\widehat{Q}(\omega) = I_2 - \widehat{P}(\omega)\widehat{P}(\omega)^*$$

where $\widehat{P}(\omega) = [\sqrt{2}P_1(\omega), \sqrt{2}P_2(\omega)]^{\mathrm{T}}$. We claim that $\widehat{Q}$ is nonnegative definite under the assumption (26). Clearly, $\widehat{Q}(\omega)$ is symmetric. 1 is an eigenvalue of $\widehat{Q}$ since any vector $v(\omega)$ orthogonal to $\widehat{P}(\Omega)$ is an eigenvector of $\widehat{Q}$. $1 - \widehat{P}(\omega)^*\widehat{P}(\omega) \geq 0$ is another eigenvalue of $\widehat{Q}$. Thus, by matrix-valued Riesz–Fejér factorization (cf. [34] and [30]) we have

$$\widehat{Q}(\omega) = \widetilde{Q}(\omega)\widetilde{Q}(\omega)^*$$

for a Laurent polynomial matrix $\widetilde{Q}(\omega)$ of size 2×2. That is,

$$\widehat{P}(\omega)\widehat{P}(\omega)^* + \widetilde{Q}\widetilde{Q}^* = I_2 \,.$$

Replacing ω by 2ω in the above equation, multiplying U from the left and U^* from the right of the equation we have

$$\begin{aligned}
&\begin{bmatrix} P(\omega) \\ P(\omega+\pi) \end{bmatrix} \left[P(\omega)^* \; P(\omega+\pi)^*\right] \\
&\quad + \begin{bmatrix} Q_1(\omega) & Q_1(\omega+\pi) \\ Q_2(\omega) & Q_2(\omega+\pi) \end{bmatrix} \\
&\quad \times \begin{bmatrix} Q_1(\omega)^* & Q_2(\omega) \\ Q_1(\omega+\pi) & Q_2(\omega+\pi) \end{bmatrix} = I_2 \,.
\end{aligned}$$

We now use B-splines to give some examples of tight wavelet frames (cf. [14]).

Example 16 Consider linear B-splines. The mask polynomial $P(\omega) = (1+z)^2/4$ with $z = \mathrm{e}^{i\omega}$. It is easy to find $Q_1(z) = -1/4 + z/2 - z^2/4$ and $Q_2(z) = \sqrt{2}(1-z^2)/2$. One of the tight wavelet frame generators defined by using Q_1 and Q_2 is symmetric and the other is anti-symmetric.

Example 17 Consider quadratic B-splines. The mask polynomial $P(\omega) = (1+z)^3/8$. One can find $Q_1(z) = \sqrt{3}(1-z)/4$ and $Q_2(z) = (1+3z-3z^2-z^3)/8$. The tight wavelet generators associated with these Q_1 and Q_2 are all anti-symmetric.

Example 18 Consider cubic B-splines. The mask polynomial $P(\omega) = (1+z)^4/16$. Denote

$$a = \frac{\sqrt{8-2\sqrt{14}}}{8}, b = \frac{\sqrt{8+2\sqrt{14}}}{8}, r = \frac{\sqrt{16+2\sqrt{14}}}{8}.$$

Let

$$\begin{aligned}
Q_1(z) &= 4ar^2 + \left(r - \frac{1}{16r}\right)\frac{z}{\sqrt{2}} + \frac{1-2r^2}{\sqrt{2}r}z^2 \\
&\quad - \frac{b}{\sqrt{2}}z^3 - \frac{b}{4\sqrt{2}}z^4 \\
Q_2(z) &= \frac{r}{4\sqrt{2}} + (a + \frac{b}{\sqrt{2}})z \\
&\quad - \frac{b}{2\sqrt{2}}z^2 - \frac{b}{\sqrt{2}r}z^3 - \frac{b^2}{\sqrt{2}r}z^4 \,.
\end{aligned}$$

One can define tight wavelet frame generators ψ_ℓ using these Q_ℓ.

It is also possible to find symmetric and anti-symmetric tight wavelet frame generators using B-splines. We refer to [14,72,73] for the detail.

Tight Wavelet Frames over Bounded Domain

We discuss the construction of tight wavelet frames over bounded domain. The following discussion works for a bounded domain over any Euclidean space. For convenience, we restrict our attention to the univariate setting.

Let $\Omega \subset \mathbb{R}$ be a bounded domain, e. g., an interval. A tight wavelet frame for $L_2(\Omega)$ is based on a half infinite sequence of nested subspaces over Ω. Suppose that we have a sequence of nested subspaces $\{V_k\}_{k\in\mathbb{Z}_+} \subset L_2(\Omega)$ satisfying

$$\begin{aligned}
&V_1 \subset V_2 \subset \ldots \subset V_k \subset \ldots \to L_2(\Omega) \quad \text{and} \\
&\bigcup_{k=1}^{\infty} V_k \quad \text{is dense in} \quad L_2(\Omega) \,.
\end{aligned}$$

Let $\Phi_k := (\phi_{k,1}, \ldots, \phi_{k,m_k})^{\mathrm{T}}$ be a column vector of locally supported functions in V_k which generate V_k, i. e., $V_k = \mathrm{span}\{\phi_{k,1}, \ldots, \phi_{k,m_k}\}$.

Because V_k is a subspace of V_{k+1}, the vector Φ_k in V_k can be generated by the column vector Φ_{k+1} which spans V_{k+1}. Thus there exists a matrix P_k of size $m_k \times m_{k+1}$ with $m_k \leq m_{k+1}$ such that

$$\Phi_k = P_k \Phi_{k+1} \,. \tag{29}$$

The matrix P_k is often called a refinement matrix. Let Q_k be another matrix of size $n_k \times m_{k+1}$. Define

$$\Psi_k := Q_k \Phi_{k+1} \,. \tag{30}$$

Definition 3 The family of vectors $\{\Psi_k\}_{k\in\mathbb{Z}_+}$ defined in (30) is a *tight wavelet frame* associated with $\{\Phi_k\}_{k\in\mathbb{Z}_+}$ in

$L_2(\Omega)$ if

$$\|f\|^2 = \sum_{j=1}^{m_1} |\langle f, \phi_{1,j}\rangle|^2 + \sum_{k=1}^{\infty}\sum_{j=1}^{n_k} |\langle f, \psi_{k,j}\rangle|^2, \qquad \forall f \in L_2(\Omega)\,,$$

where $\langle f, g\rangle = \int_\Omega f(x)g(x)\mathrm{d}x$ be the standard inner product on $L_2(\Omega)$.

If $\phi_{1,j}, j = 1, \ldots, m_1$ and $\psi_{k,j}, j = 1, \ldots, n_k, k = 1, \ldots$ generate a tight wavelet frame, then one can prove that

$$f = \sum_{j=1}^{m_1} \langle f, \phi_{1,j}\rangle \phi_{1,j} + \sum_{k=1}^{\infty}\sum_{j=1}^{n_k} \langle f, \psi_{k,j}\rangle \psi_{k,j}$$

for any $f \in L_2(\Omega)$ by using the polarization technique mentioned before.

We show that a matrix Q_k satisfying

$$I_{m_{k+1}} = P_k^{\mathrm{T}} P_k + Q_k^{\mathrm{T}} Q_k\,, \tag{31}$$

for a given refinement matrix P_k in (29) is a key step for constructing a tight wavelet frame. Here $I_{m_{k+1}}$ is the standard identity matrix of size m_{k+1}.

Clearly, each component in the function vector Ψ_k is in V_{k+1}. We want to have

$$\langle f, \Phi_{k+1}\rangle^{\mathrm{T}} \Phi_{k+1} = \langle f, \Phi_k\rangle^{\mathrm{T}} \Phi_k + \langle f, \Psi_k\rangle^{\mathrm{T}} \Psi_k, \qquad \forall f \in L_2(\Omega)\,. \tag{32}$$

Let $c_{k,i} := \langle f, \phi_{k,i}\rangle$ for all $i = 1, \ldots, m_k$ and $C_k := (c_{k,1}, \ldots, c_{k,m_k})^{\mathrm{T}}$ be a column vector of size $m_k \times 1$ for any $k \in \mathbb{Z}_+$. In the same way, let $d_{k,j} := \langle f, \psi_{k,j}\rangle$ for all $j = 1, \ldots, n_k$ and $D_k := (d_{k,1}, \ldots, d_{k,n_k})^{\mathrm{T}}$. Then we know

$$C_k = \langle f, \Phi_k\rangle = \langle f, P_k \Phi_{k+1}\rangle = P_k C_{k+1}\,, \tag{33}$$

$$D_k = \langle f, \Psi_k\rangle = \langle f, Q_k \Phi_{k+1}\rangle = Q_k C_{k+1}\,. \tag{34}$$

Thus condition in (32) can be expressed in the following form

$$C_{k+1}^{\mathrm{T}} \Phi_{k+1} = C_k^{\mathrm{T}} P_k \Phi_{k+1} + D_k^{\mathrm{T}} Q_k \Phi_{k+1}\,.$$

That is, $C_{k+1}^{\mathrm{T}} = C_k^{\mathrm{T}} P_k + D_k^{\mathrm{T}} Q_k$. By using (34), we get

$$C_{k+1}^{\mathrm{T}} C_{k+1} = C_{k+1}^{\mathrm{T}} \left(P_k^{\mathrm{T}} P_k + Q_k^{\mathrm{T}} Q_k\right) C_{k+1}\,.$$

This implies that Q_k must satisfy (31) for all $k \geq 1$. On the other hand, if we find Q_k satisfying (31) for all $k \geq 1$, then we have the above equation and hence, by using (34),

$$C_{k+1}^{\mathrm{T}} C_{k+1} = C_k^{\mathrm{T}} C_k + D_k^{\mathrm{T}} D_k\,.$$

It follows for any $\ell \in \mathbb{Z}_+$ with $\ell < k$,

$$C_{k+1}^{\mathrm{T}} C_{k+1} = C_\ell^{\mathrm{T}} C_\ell + \sum_{j=\ell}^{k} D_j^{\mathrm{T}} D_j\,. \tag{35}$$

The condition (31) implies $C_{k+1}^{\mathrm{T}} = C_{k+1}^{\mathrm{T}}(P_k^{\mathrm{T}} P_k + Q_k^{\mathrm{T}} Q_k) = C_k^{\mathrm{T}} P_k + D_k^{\mathrm{T}} Q_k$ and hence,

$$C_{k+1}^{\mathrm{T}} \Phi_{k+1} = C_k^{\mathrm{T}} \Phi_k + Q_k^{\mathrm{T}} \Psi_k = \ldots = C_\ell^{\mathrm{T}} \Phi_\ell + \sum_{j=\ell}^{k} D_j^{\mathrm{T}} \Psi_j.$$

If $C_{k+1}^{\mathrm{T}} \Phi_{k+1}$ converges to f in $L_2(\Omega)$, for any $\ell \in \mathbb{Z}_+$, we have

$$\begin{aligned}
\|f\|^2 &= \left\langle f, \lim_{k\to+\infty} C_{k+1}^{\mathrm{T}} \Phi_{k+1}\right\rangle \\
&= \lim_{k\to+\infty} \left\langle f, C_\ell^{\mathrm{T}} \Phi_\ell + \sum_{j=\ell}^{k} D_j^{\mathrm{T}} \Psi_j\right\rangle \\
&= C_\ell^{\mathrm{T}} C_\ell + \sum_{j=\ell}^{\infty} D_j^{\mathrm{T}} D_j \\
&= \sum_{j=1}^{m_\ell} |\langle f, \phi_{\ell,j}\rangle|^2 + \sum_{k=\ell}^{\infty}\sum_{j=1}^{n_k} |\langle f, \psi_{k,j}\rangle|^2\,.
\end{aligned}$$

Theorem 13 *Suppose that Φ_k is a given refinable vector which spans V_k for all $k \geq 1$ with refinable matrix P_k, i. e., $\Phi_k = P_k \Phi_{k+1}$. Suppose that the projections f to V_k converge to f, i. e., $C_k^{\mathrm{T}} \Phi_k \longrightarrow f$ for any $f \in L_2(\Omega)$. Suppose Q_k satisfies (31). Let $\Psi_k = Q_k \Phi_k$. Then $\Psi_k, k \in \mathbb{Z}_+$ form a tight wavelet frame. Hence, any $f \in L_2(\Omega)$ can be generated by using Φ_ℓ and Ψ_k with $k \geq \ell$ for any $\ell \geq 1$ using the formula above.*

We now explain how to compute Q_k satisfying (31).

Theorem 14 *Let $\{V_k\}_{k=1}^{\infty}$ be a nested sequence and V_k be generated by a family of functions Φ_k. Denote by P_k the refinable matrix, i. e., $\Phi_k = P_k \Phi_{k+1}$. If $I_{m_k} - P_k P_k^{\mathrm{T}}$ is positive semi-definite, then there exists a matrix Q_k satisfying (31). Moreover, if each component function $\phi_{k,j}$ of Φ_k is locally supported then each component function $\psi_{k,j}$ of the vector Ψ_k is locally supported.*

We note that the condition $I_{m_k} - P_k P_k^{\mathrm{T}} \geq 0$ is different from (31) which may be rewritten in $I_{m_{k+1}} - P_k^{\mathrm{T}} P_k = Q_k^{\mathrm{T}} Q_k \geq 0$ since the the size of the matrix, e. g., I_{m_k} is smaller than that of $I_{m_{k+1}}$. The proof of Theorem 14 can be found in [54].

Next we use B-splines to construct tight wavelet frames over a bounded interval $[0, b]$ for an integer $b > 0$. Fix

$m \geq 1$ and consider B-spline N_m which satisfies the following refinement equation:

$$N_m(x) = \sum_{j\in\mathbb{Z}} c_j^m N_m(2x - j)\,,$$

where

$$c_j^m = \begin{cases} 2^{-m+1}\binom{m}{j}, & \text{for} \quad 0 \leq j \leq m \\ 0 & \text{otherwise}\,. \end{cases} \tag{36}$$

We now define scaling functions $\phi_{j,k}(x) = 2^{j/2}N_m(2^j x - k)|_{[0,b]}$ which are nonzero when $k = 1-m, \ldots, 2^j b - m + 1$. Thus, let

$$V_j := \{\phi_{j,k} : -1 \leq k \leq 2^j b - m + 1\}$$

for all j. It is easy to see that $V_j \subset V_{j+1}$ by using (36). Also, we can easily verify that $\bigcup_{j=1}^{\infty} V_j$ is dense in $L_2([0, b])$. Let $m_j = 2^j b - m + 2$ and $\Phi_j = [\phi_{j,-1}, \ldots, \phi_{j,m_j-1}]^{\mathrm{T}}$ be a vector of basis functions for V_j. We can find a refinement matrix P_j^m of size $m_j \times m_{j+1}$ satisfying $\Phi_j^m = P_j^m \Phi_{j+1}^m$ for each $j \in \mathbb{Z}_+$ using the coefficients in (36). Based on the discussion above we can check the positive semi-definite property of the matrix

$$I_{m_j} - P_j^m \cdot P_j^{m\,\mathrm{T}}$$

for the identity matrix I_{m_j}.

Lemma 9 *The above matrix of size $m_j \times m_j$ associated with B-spline N_m of order m is positive semi-definite for each $j \in \mathbb{Z}$ and $m \geq 1$.*

Proof Let us denote $(p_{i,k}^{m,j}) := P_j^m$. Then for each $i = 1, \ldots, m_j$

$$0 \leq \sum_{k=1}^{m_{j+1}} p_{i,k}^{m,j} \leq \frac{1}{2}\sum_{k=0}^{m} c_k^m = 1\,, \tag{37}$$

where $c_k{}^m$ is in (36). Let us express $[g_{i,k}^{m,j}]_{1\leq i,k\leq m_j} = P_j^m \cdot P_j^{m\,\mathrm{T}}$. We claim that the matrix $I_{m_k} - P_j^m \cdot P_j^{m\,\mathrm{T}}$ is diagonally dominant. Indeed, it is sufficient to check $|1 - g_{i,i}^{m,j}| \geq \sum_{k\neq i} |\, g_{i,k}^{m,j}\,|$ for $i \leq m_k$. Notice that

$$g_{i,k}^{m,j} = \sum_{\ell=1}^{m_{j+1}} p_{i,\ell}^{m,j} p_{\ell,k}^{m,j}\,.$$

Then for each $k \in \mathbb{Z}_+$,

$$\begin{aligned} 1 - |g_{i,i}^{m,j}| - \sum_{k\neq i} |\, g_{i,k}^{m,j}\,| &= 1 - \sum_{k=1}^{m_{j+1}}\sum_{\ell=1}^{m_{j+1}} p_{i,\ell}^{m,k} p_{k,\ell}^{m,j} \\ &= 1 - \left(\sum_{\ell=1}^{m_{j+1}} p_{i,\ell}^{m,j}\right)^2\,. \end{aligned}$$

Since (37), $1 - |g_{i,i}^{m,j}| \geq \sum_{k\neq i} |\, g_{i,k}^{m,j}\,|$ for all $i = 1, \ldots, m_k$. Therefore the symmetry matrix $I_{m_k} - P_k^m \cdot P_k^{m\,\mathrm{T}}$ is positive semi-definite. This completes the proof. □

That is, we can use B-splines of any order to construct tight wavelet frames over any intervals.

q-Dilated Orthonormal Wavelets

In the previous sections, we consider refinable functions ϕ which are dilated by 2. That is,

$$\phi(x) = \sum_n h_n \sqrt{q}\phi(qx - n), \quad x \in \mathbb{R} \tag{38}$$

for $q = 2$. In general, we can consider refinable function ϕ which can be dilated by $q > 2$. All the previous theories of various wavelets have a generalization in this setting of dilation factor $q > 2$. One important feature of wavelets in the new setting is that it requires more than one wavelet function. Let us give a brief explanation of orthonormal wavelets based on dilation factor $q = 3$.

Suppose we have a multiresolution approximation of $L_2(\mathbb{R})$ based on dilation factor q. That is, we have

Definition 4 A sequence of closed subspaces $\ldots V_{-2}, V_{-1}, V_0, V_1, V_2, \ldots$ of $L_2(\mathbb{R})$ is a multiresolution approximation of $L_2(\mathbb{R})$ if they satisfy the following

1. $V_j \subset V_{j+1}, \forall j \in \mathbb{Z}$;
2. $\overline{\bigcup_{j\in\mathbb{Z}} V_j} = L_2(\mathbb{R})$ and $\bigcap_{j\in\mathbb{Z}} V_j = \{0\}$;
3. $f(\cdot) \in V_0 \iff f(q^j \cdot -k) \in V_j$ for all $j, k \in \mathbb{Z}$;
4. There is a function $\phi \in L_2(\mathbb{R})$ called scaling function such that $\{\phi(x - k),\ k \in \mathbb{Z}\}$ forms an orthonormal basis of V_0.

By taking the Fourier transform both sides of (38), we get that

$$\widehat{\phi}(\xi) = m_0(\omega/q)\widehat{\phi}(\omega/q), \quad \xi \in \mathbb{R}\,, \tag{39}$$

where

$$m_0(\omega) = \frac{1}{\sqrt{q}}\sum_n h_n \mathrm{e}^{-in\omega}\,. \tag{40}$$

m_0 is a Laurent polynomial in $z = \mathrm{e}^{i\omega}$ and is called the mask of ϕ.

Since ϕ is orthonormal, we know

$$\begin{aligned} \delta_{n0} &= \int_{\mathbb{R}} \phi(x)\overline{\phi(x-n)}\mathrm{d}x = \frac{1}{2\pi}\int_{\mathbb{R}} |\hat{\phi}(\xi)|^2 \mathrm{e}^{in\xi}\,\mathrm{d}\xi \\ &= \frac{1}{2\pi}\int_0^{2\pi}\left(\sum_{k\in\mathbb{Z}} |\hat{\phi}(\xi + 2k\pi)|^2\right)\mathrm{e}^{in\xi}\,\mathrm{d}\xi\,, \end{aligned}$$

and hence,

$$\sum_{k\in\mathbb{Z}} |\widehat{\phi}(\xi + 2k\pi)|^2 = 1 \,. \tag{41}$$

Since m_0 is 2π-periodic, by (39) and (41) we find that

$$\begin{aligned} 1 &= \sum_{k\in\mathbb{Z}} |\widehat{\phi}(q\xi + 2k\pi)|^2 \\ &= \sum_{k\in\mathbb{Z}} |m_0(\xi + 2k\pi/q)|^2 |\widehat{\phi}(\xi + 2k\pi/q)|^2 \\ &= |m_0(\xi)|^2 \sum_{k\in\mathbb{Z}} |\widehat{\phi}(\xi + 2k\pi)|^2 + |m_0(\xi + 2\pi/q)|^2 \\ &\quad \sum_{k\in\mathbb{Z}} |\widehat{\phi}(\xi + 2\pi/q + 2k\pi)|^2 \\ &\quad + \ldots + |m_0(\xi + 2(q-1)\pi/q)|^2 \\ &\quad \sum_{k\in\mathbb{Z}} |\widehat{\phi}(\xi + 2(q-1)\pi/q + 2k\pi)|^2 \,. \end{aligned}$$

Hence

$$|m_0(\xi)|^2 + |m_0(\xi + 2\pi/q)|^2 + \ldots + |m_0(\xi + 2(q-1)\pi/q)|^2 = 1 \,. \tag{42}$$

Consider $j = 0$ and let W_0 be the orthogonal complement of V_0 in V_1. Since V_0 is generated by integer translates of $\phi(x)$ and V_1 is generated by integer translates of q functions, $\phi(qx)$, $\phi(qx-1)$, …, $\phi(qx-q+1)$, we need $q-1$ functions whose integer translates to generate W_0. That is, we look for $\psi_1, \psi_2, \ldots, \psi_{q-1}$ such that

$$\{\psi_i(x-k) : i = 1, 2, \ldots, q-1, k \in \mathbb{Z}\} \tag{43}$$

forms an orthonormal basis for W_0. For convenience, let

$$W_0^i := span\{\psi_i(x-k), k \in Z\} \quad \text{for} \quad i = 1, 2, \ldots, q-1. \tag{44}$$

Then we have $W_0 = W_0^1 \bigoplus W_0^2 \bigoplus \ldots \bigoplus W_0^{q-1}$. In general, we set

$$W_j^i = \text{span}\{\psi_i(q^j x-k), k \in Z\} \quad \text{for} \quad i = 1, 2, \ldots, q-1.$$

Then we have

$$V_{j+1} = V_j \oplus W_j^1 \oplus W_j^2 \oplus \ldots \oplus W_j^{q-1} \tag{45}$$

and

$$W_j^i \perp W_j^{i'} \quad \text{if} \quad i \neq i' \,. \tag{46}$$

It implies

$$L_2(\mathbb{R}) = \bigoplus_{j\in\mathbb{Z}} \bigoplus_{i=1}^{q-1} W_j^i \,. \tag{47}$$

That is, ψ_i, $i = 1, 2, \ldots, q-1$ form a set of orthonormal q-wavelet functions.

Since $\psi_j \in V_1$ and $\psi_j \perp V_0$, we have

$$\psi_j(x) = \sqrt{q} \sum_n g_n^j \phi(qx - n) \,, \tag{48}$$

with $g_n^j = \langle \psi_j, \phi_{1,n} \rangle$ and $\{g_n^j\} \in \ell^2(\mathbb{Z})$. This implies

$$\hat{\psi}_j(\xi) = m_j(\xi/q)\hat{\phi}(\xi/q) \,, \tag{49}$$

where

$$m_j(\xi) = \frac{1}{\sqrt{q}} \sum_n g_n^j e^{-in\xi} \,. \tag{50}$$

We can prove the following

Lemma 10 *Let ϕ be an orthonormal refinable function associated with dilation factor $q \geq 2$ and m_0 be the corresponding mask. Let $\psi_1, \psi_2, \ldots, \psi_{q-1}$ be functions defined by in terms of Fourier transform using take $m_1, m_2, \ldots, m_{q-1}$ as above. Then the family $\{\psi_j(x-n), j = 1, 2, \ldots, q-1, n \in \mathbb{Z}\}$ is an orthonormal basis for the orthogonal complement W_0 of V_0 in V_1 if and only if the matrix*

$$\begin{bmatrix} m_0(\omega) & m_0(\omega + 2\pi/q) & \ldots & m_0(\omega + 2(q-1)\pi/q) \\ m_1(\omega) & m_1(\omega + 2\pi/q) & \ldots & m_1(\omega + 2(q-1)\pi/q) \\ \vdots & \vdots & \ddots & \vdots \\ m_{q-1}(\omega) & m_{q-1}(\omega + 2\pi/q) & \ldots & m_{q-1}(\omega + 2(q-1)\pi/q) \end{bmatrix} \tag{51}$$

is unitary.

As usual, the first problem is to find a Laurent polynomial m_0 such that $\sum_{i=0}^{q-1} |m_0(\omega + 2i\pi/q)|^2 = 1$. Once such m_0 is found, we can find the other m_j, $j = 1, \ldots, q-1$ by using unitary matrix extension technique in [62].

To give some examples, we now restrict ourselves to $q = 3$. Consider a mask m_0 which must satisfy

$$m_0(0) = 1$$

in order to make $\prod_{j=1}^{\infty} m_0(\omega/3^j)$ converges so that it defines a distribution ϕ. m_0 must satisfy

$$|m_0(\omega)|^2 + |m_0(\omega + 2\pi/3)|^2 + |m_0(\omega + 4\pi/3)|^2 = 1$$

so that ϕ is in $L_2(\mathbb{R})$. Note that the above two conditions implies that $m_0(\omega) = \left(\frac{1+e^{i\omega}+e^{2i\omega}}{3}\right)\mathcal{L}(\omega)$ for some Laurent polynomial in $e^{i\omega}$. The following can be found in [92].

Lemma 11 *A Laurent polynomial* $m_0(\omega) = \sum_{k=0}^{5} p_k e^{ik\omega}$ *satisfies above two conditions if and only if*

$$p_0 = \frac{1}{6} + \frac{\sqrt{3}}{6}\cos\theta, \qquad p_1 = \frac{1}{6} + \frac{\sqrt{3}}{6}\sin\theta\cos\alpha$$
$$p_2 = \frac{1}{6} + \frac{\sqrt{3}}{6}\sin\theta\sin\alpha, \quad p_3 = \frac{1}{6} - \frac{\sqrt{3}}{6}\cos\theta$$
$$p_4 = \frac{1}{6} - \frac{\sqrt{3}}{6}\sin\theta\cos\alpha, \quad p_5 = \frac{1}{6} - \frac{\sqrt{3}}{6}\sin\theta\sin\alpha$$

for some θ, α *in* $[0, 2\pi]$.

Example 19 A scaling function associated with m_0 in Lemma 11 is symmetric if and only if the coefficients given above with $\tan\theta = -1$, $\sin\alpha = 1$ or $\tan\theta = 1$, $\sin\alpha = -1$.

Example 20 m_0 in Lemma 11 has two order of vanishing moments, i. e., m_0 contains a factor $(1 + e^{i\omega} + e^{i2\omega})^2$ if $\alpha = \arccos\left(\sqrt{\frac{19}{50+4\sqrt{57}}}\right)$ and $\theta = \arcsin\left(\frac{\sqrt{50+4\sqrt{57}}}{9}\right)$.

Next we consider $m_0(z) = p_0 + p_1 z + p_2 z^2 + p_3 z^3 + p_4 z^4 + p_5 z^5 + p_6 z^6 + p_7 z^7 + p_8 z^8$ be the symbol of a scaling function supported on $[0, 4]$, where $z = e^{i\omega}$.

Lemma 12 $m_0(z)$ *satisfies* $m_0(0) = 1$ *and*

$$|m_0(z)|^2 + |m_0(z\tau)|^2 + \left|m_0(z\tau^2)\right|^2 = 1\,, \qquad \forall z = e^{i\omega}, \quad \omega \in \mathbb{R}, \tag{52}$$

with $\tau = e^{i\frac{2\pi}{3}}$ *if and only if*

$$p_1 = \frac{1}{12} + \frac{\sqrt{3}}{12}\sin\alpha + \frac{1}{6}r\sin\theta$$
$$p_2 = \frac{1}{12} + \frac{\sqrt{3}}{12}\cos\alpha\sin\beta + \frac{1}{6}r\cos\theta\sin\gamma$$
$$p_3 = \frac{1}{6}(1 - \sqrt{3}\cos\alpha\cos\beta)$$
$$p_4 = \frac{1}{6}(1 - \sqrt{3}\sin\alpha)$$
$$p_5 = \frac{1}{6}(1 - \sqrt{3}\cos\alpha\sin\beta)$$
$$p_6 = \frac{1}{12} + \frac{\sqrt{3}}{12}\cos\alpha\cos\beta - \frac{1}{6}r\cos\theta\cos\gamma$$
$$p_7 = \frac{1}{12} + \frac{\sqrt{3}}{12}\sin\alpha - \frac{1}{6}r\sin\theta$$
$$p_8 = \frac{1}{12} + \frac{\sqrt{3}}{12}\cos\alpha\sin\beta - \frac{1}{6}r\cos\theta\sin\gamma,$$

where

$$r = \sqrt{\frac{1}{2} + \frac{\sqrt{3}}{6}(\cos\alpha\cos\beta + \cos\alpha\sin\beta + \sin\alpha)} \tag{53}$$

for some α, β, γ *in* $[0, 2\pi]$.

Example 21 A scaling function associated with m_0 in Lemma 12 is symmetric if and only if the coefficients given above with $\sin\theta = 0$, $\cos\alpha = 0$, $\tan\beta = 1$, and $\tan\gamma = -1$.

Example 22 m_0 in Lemma 12 has three order of vanishing moments, i. e., m_0 contains a factor $(1 + e^{i\omega} + e^{i2\omega})^3$ if $\alpha = \gamma = 0$, $\beta = \arcsin(\frac{7\sqrt{6}}{36}) - \frac{\pi}{4}$, and $\theta = \pi + \arcsin(\frac{3}{5})$.

Multiwavelets and Balanced Multiwavelets

We shall explain a construction of multiwavelets in this section. Typical multiwavelets are DGHM multiwavelets (cf. [28]), Chui-Lian wavelets (cf. [16]) and multiwavelets based on B-splines (cf. [29]). In the following we present a newer construction. An advantage is that the number of wavelets is always 3 no matter how smooth the wavelets are.

Fix integer $r > 1$. Let $\Phi = [\phi_1, \ldots, \phi_r]^{\mathrm{T}}$ be a vector of compactly supported real-valued functions in $\mathbb{R}$. We suppose that Φ is refinable. That is, there exist matrices A_k's of size $r \times r$ such that

$$\Phi(x) = \sum_{k\in\mathbb{Z}^d} A_k \Phi(2x - k), \quad x \in \mathbb{R}\,.$$

Also, we say Φ is orthonormal if

$$\int_{\mathbb{R}} \phi_i(x)\phi_j(x-k)\mathrm{d}x = \begin{cases} 1\,, & \text{if} \quad i = j \quad \text{and} \quad k = 0\,, \\ 0\,, & \text{otherwise} \end{cases}$$

for all $i, j = 1, \ldots, r$. Φ generates a space S if S consists of all finitely linear combination of integer translates of entries of Φ.

Next we define a Grammian matrix $G = (G_{ij})_{i,j=1,\ldots,r}$ of size $r \times r$ associated with Φ by

$$G_{ij}(z) = \sum_{k\in\mathbb{Z}} z^k \int_{\mathbb{R}} \phi_i(x)\phi_j(x-k)\mathrm{d}x$$

for all $i, j = 1, \ldots, r$ with $z \in \mathbf{C} \setminus \{0\}$. We note that Φ is orthonormal if and only if its Grammian matrix G is the identity of size $r \times r$.

We suppose that Φ generates a space S. Then for any compactly supported functions $\psi_1, \ldots, \psi_s$ in S, there exists a finitely many nonzero matrices C_k of size $s \times r$ such

that

$$\Psi(x) = [\psi_1(x), \ldots, \psi_s(x)]^{\mathrm{T}} = \sum_{k\in\mathbb{Z}} C_k \Phi(x-k) .$$

In terms of Fourier transform, we have

$$\widehat{\Psi}(\omega) = C(z)\widehat{\Phi}(\omega)$$

where $C(z)$ denotes the $s \times r$ matrix of Laurent polynomials, i. e.,

$$C(z) := \sum_{k\in\mathbb{Z}} C_k z^k .$$

A square matrix $C(z)$ is said to be invertible if $\det(C(z))$ is a monomial of z, e. g., αz^m for a scalar $\alpha \neq 0$ and an integer $m \in \mathbb{Z}$. It is clear that if $C(z)$ is invertible, Ψ generates the same S. A proof of the following result can be found in literature (cf. e. g., [31]).

Lemma 13 *Suppose that Ψ is a vector of compactly supported functions and generates a space S. Let $\mathcal{G}(z) = (G_{ij}(z))_{i,j=1,\ldots,r}$ of size $r \times r$ by*

$$G_{ij}(z) = \sum_{k\in\mathbb{Z}} z^k \int_{\mathbb{R}} \psi_i(x)\psi_j(x-k)dx$$

for all $i, j = 1, \ldots, r$ be the Grammian matrix associated with Ψ. If the determinant of the Grammian matrix $\mathcal{G}(z)$ is a nonzero constant, then there exists a Φ which is orthonormal and generates S. The converse is also true.

The above lemma reveals a key for constructing orthonormal vector of scaling functions: find $\psi_1, \ldots, \psi_r$ which generate the same space S such that its Grammian matrix has a constant determinant.

We now explain how to use B-splines for constructing an orthonormal vector of scaling functions with $r = 3$. Let N_m be the uniform B-spline of order m, in terms of Fourier transform,

$$\widehat{N}_m(\omega) = \left(\frac{1-\mathrm{e}^{-i\omega}}{i\omega}\right)^m .$$

Let $V_0 = \mathrm{span}\{N_m(x-k), k \in \mathbb{Z}\}$ be the spline space. Since N_m is a refinable function, for V_1 being spanned by the integer translates of $N_m(2x-k), k \in \mathbb{Z}$, we have $V_0 \subset V_1$. Thus, letting $\psi_1(x) = N_m(2x)$ and $\psi_2(x) = N_m(2x-1)$, ψ_1 and ψ_2 generate V_1. On the other hand, by the dilation equation, there exist two finite sequences a_{2k} and a_{2k+1} such that

$$N_m(x) = \sum_{k\in\mathbb{Z}} a_{2k}\psi_1(x-k) + \sum_{k\in\mathbb{Z}} a_{2k+1}\psi_2(x-k) .$$

Note that the Fourier transform of the above equation is

$$\widehat{N}_m(2\omega) = \frac{1}{2}A(z)\widehat{N}_m(\omega)$$

and

$$\begin{aligned}\widehat{N}_m(\omega) &= A_0(z)\widehat{\psi}_1(\omega) + A_1(z)\widehat{\psi}_2(\omega)\\ &= A_0(z)\frac{1}{2}\widehat{N}_m\left(\frac{\omega}{2}\right) + A_1(z)\frac{1}{2}z^{\frac{1}{2}}\widehat{N}_m\left(\frac{\omega}{2}\right)\end{aligned}$$

where

$$A_0(z) = \sum_{k\in\mathbb{Z}} a_{2k}z^k \quad \text{and} \quad A_1(z) = \sum_{k\in\mathbb{Z}} a_{2k+1}z^k .$$

It follows that

$$A(z) = A_0(z^2) + zA_1(z^2) .$$

It is known that $A(z) = 2\left(\frac{1+z}{2}\right)^m$. It is easy to see that there exist two Laurent polynomials $B_0(z)$ and $B_1(z)$ of degree $\leq m$ such that

$$A_0(z)B_0(z) + A_1(z)B_1(z) = 1 .$$

We now define a new spline function in terms of Fourier transform by

$$\widehat{M}_m(\omega) = -B_1(z)\widehat{\psi}_1(\omega) + B_0(z)\widehat{\psi}_2(\omega) .$$

Recall that

$$\widehat{N}_m(\omega) = A_0(z)\widehat{\psi}_1(\omega) + A_1(z)\widehat{\psi}_2(\omega) .$$

It follows that N_m and M_m generate V_1 since the determinant of the foll owing matrix

$$\begin{bmatrix}\widehat{N}_m(\omega)\\ \widehat{M}_m(\omega)\end{bmatrix} = \begin{bmatrix}A_0(z) & A_1(z)\\ -B_1(z) & B_0(z)\end{bmatrix}\begin{bmatrix}\widehat{\psi}_1(\omega)\\ \widehat{\psi}_2(\omega)\end{bmatrix}$$

is constant 1. Furthermore, $N_m(2x), N_m(2x-1)$, $M_m(2x), M_m(2x-1)$ generate V_2.

Define $\psi_3 = \sum_{k\in\mathbb{Z}} \alpha_k M_m(2x-k)$ for some finitely many nonzero coefficients α_k. We will show how to find such α_k that the Grammian matrix associated with $\{\psi_1, \psi_2, \psi_3\}$,

$$\mathcal{G}(z) = \left(\sum_{k\in\mathbb{Z}} z^k \int_{\mathbb{R}} \psi_i(x)\psi_j(x-k)\mathrm{d}x\right)_{i,j=1,2,3}$$

has a constant determinant. Let

$$r(z) = \sum_{k\in\mathbb{Z}} \alpha_k z^k .$$

The computation in [31] shows

$$4\det \mathcal{G}(z^2) = D(z)r(z)r(1/z) + D(-z)r(-z)r(-1/z) \quad (54)$$

where

$$D(z) = \frac{1}{2}(a(z^2) - zb(z^2))\,(a(z)^2 - zb(z)^2)$$

with

$$a(z) = \sum_{k\in\mathbb{Z}} z^k \int N_m(2x)N_m(2x-k)\mathrm{d}x,$$
$$b(z) = \sum_{k\in\mathbb{Z}} z^k \int N_m(2x)N_m(2x-2k-1)\mathrm{d}x\,.$$

Now we claim that there exists a polynomial $p(z) \geq 0$ such that

$$D(z)p(z) + D(-z)p(-z) = 1\,.$$

Once we have such a $p(z)$, it follows from the Riesz–Féjer lemma that there exists a polynomial $r(z)$ such that $r(z)r(1/z) = p(z)$. This $r(z)$ is the polynomial we look for such that the determinant (54) of Grammian matrix $\mathcal{G}(z)$ is a nonzero constant.

The existence of $p(z)$ satisfying the above properties is guaranteed by the following lemma (see a constructive proof from [32]).

Lemma 14 *Let q be a polynomial of degree n with all its zeros in $[1,\infty)$ having a positive leading coefficient. Then there exists a unique polynomial p with real coefficients of degree $n-1$ such that*

$$p(x)q(x) + p(1-x)q(1-x) = 1$$

for $x \in [0,1]$. Moreover, $(-1)^n p(x) > 0$ for $x \in (0,1)$.

By letting $x = 1-(z+1/z)/2$ we can convert $D(z)$ into a polynomial in x. By studying the zeros of $D(z)$ we can see that $q(x) = D(z)$ satisfies the conditions in the above lemma (for details, see [31] and [9]). This shows that the Grammian matrix $\mathcal{G}(z)$ associated with $\Psi = (\psi_1, \psi_2, \psi_3)^{\mathrm{T}}$ is nonzero monomial. Hence, it can be factored into $\mathcal{G}(z) = B(z)^* B(z)$ with invertible polynomial matrix $B(z)$ by matrix-valued Fejér–Riesz lemma (cf. [34]), where $B(z)^*$ stands for the transpose and conjugate of $B(z)$. A straightforward computational method to do such factorizations can be found in [9] based on the theory developed in [30]. Letting

$$\widehat{\Phi}(z) = B(z)^{-1}\widehat{\Psi}(z)\,,$$

we know that the Grammian matrix of Φ is $B(z)^{-1}\mathcal{G}(z)B^{-1}(z)^*$ which is the identity matrix and hence $\Phi = [\phi_1, \phi_2, \phi_3]^{\mathrm{T}}$ is an orthonormal refinable function vector. Next we explain how to compute the associated wavelets. We begin with

Lemma 15 *Φ is refinable. That is, letting*

$$\widetilde{\Phi}(x) = \sqrt{2}(\phi_1(2x), \phi_2(2x), \phi_3(2x), \phi_1(2x-1), \phi_2(2x-1), \phi_3(2x-1))^{\mathrm{T}}\,,$$

there exists matrix coefficients p_i of size 3×6 such that

$$\Phi(x) = \sum_{k\in\mathbb{Z}} p_k \widetilde{\Phi}(x-k) \quad \text{or} \quad \widehat{\Phi}(\omega) = P(z)\widehat{\widetilde{\Phi}}(\omega), \quad (55)$$

where $P(z)$ is a matrix mask of size 3×6.

Since Φ is of compact compact, we may assume that only $m+1$ terms $p_0, p_1, \ldots, p_m$ are nonzero matrix coefficients. Then the orthogonal condition implies

$$\begin{aligned} 0 &= \int_{\mathbb{R}} \Phi(x)\Phi(x-k)^{\mathrm{T}}\mathrm{d}x \\ &= \sum_{i,j=1}^{m} p_i \int_{\mathbb{R}} \widetilde{\Phi}(x-i)\widetilde{\Phi}^{\mathrm{T}}(x-j-k)\mathrm{d}x\, p_j^{\mathrm{T}} \\ &= \sum_{i,j=1}^{m} p_i \delta_{i,j+k} I_{6\times 6} p_j^{\mathrm{T}} = \sum_{i=k}^{m} p_i p_{i-k}^{\mathrm{T}}\,, \end{aligned}$$

for $k = 1, \ldots, m$. In particular, we have

$$p_m p_0^{\mathrm{T}} = 0\,. \quad (56)$$

We now use induction on m to explain how to construct three compactly supported orthonormal wavelets $h_1, h_2, h_3 \in S_1$ such that

$$\mathcal{W} := \mathrm{span}\{h_1(\cdot - i), h_2(\cdot - j), h_3(\cdot - k), i, j, k \in \mathbb{Z}\}$$

is the orthogonal completement of S in S_1. It is trivial when $m = 0$. Indeed, in this case, $P(z) = p_0$ is a scalar matrix. We simply choose $Q(z)$ to be a scalar matrix which is an orthonormal extension of p_0. Assume that for $m \geq 1$, when $P_m(z) = \sum_{k=0}^{m} p_k z^k$ is an orthonormal matrix of 3×6, we can find $Q_m(z)$ such that

$$\begin{bmatrix} P_m(z) \\ Q_m(z) \end{bmatrix}$$

is unitary. We now consider the case of $m+1$: $P_{m+1}(z) = \sum_{k=0}^{m+1} p_k z^k$ satisfying orthonormal properties. In particular, (56) implies that there exists a unitary matrix U_0 of size 6×6 such that $p_0 U_0 = [0_{3\times 3}\ \tilde{p}_0^b]$ and $p_{m+1}U_0 = [\tilde{p}_{m+1}^a\ 0_{3\times 3}]$, where $\tilde{p}_0^b$ is of size 3×3 and the

same for $\tilde{p}^a_{m+1}$. Writing $p_k U_0 = [\tilde{p}^a_k, \tilde{p}^b_k]$ with $\tilde{p}^a_k$ and $\tilde{p}^b_k$ being of size 3×3. Then

$$P_{m+1}(z)U_0 = \left[\sum_{k=1}^{m+1} \tilde{p}^a_k z^k, \sum_{k=0}^{m} \tilde{p}^b_k z^k\right].$$

Let

$$U_1 := \begin{bmatrix} \frac{1}{z} I_{3\times 3} & 0_{3\times 3} \\ 0_{3\times 3} & I_{3\times 3} \end{bmatrix}.$$

Then it follows that

$$\begin{aligned} P_{m+1}(z)U_0U_1 &= \left[\sum_{k=0}^{m} \tilde{p}^a_{k+1} z^k, \sum_{k=0}^{m} \tilde{p}^b_k z^k\right] \\ &= \sum_{k=0}^{m} \left[\tilde{p}^a_{k+1}, \tilde{p}^b_k\right] z^k . \end{aligned}$$

That is, $\tilde{P}_m(z) := P_{m+1}(z)U_0U_1$ has only $m+1$ terms and is unitary. By induction, we can find an unitary extension $\tilde{Q}_m(z)$ such that

$$\begin{bmatrix} P_{m+1}(z)U_0U_1 \\ \tilde{Q}_m(z) \end{bmatrix}$$

is unitary. Clearly,

$$\begin{bmatrix} P_{m+1}(z)U_0U_1 \\ \tilde{Q}_m(z) \end{bmatrix} U_1^* U_0^* = \begin{bmatrix} P_{m+1}(z) \\ \tilde{Q}_m(z)U_1^* U_0^* \end{bmatrix}$$

is also unitary. It follows that $Q_{m+1}(z) := \tilde{Q}_m(z)U_1^* U_0^*$ is an unitary extension of $P_{m+1}(z)$. This completes the induction procedure. Several examples of multiwavelets and filters associated with these multiwavelets are given in [9].

Before we apply these multiwavelets, we need to balance them. The concept of balance was initially proposed in [63]. Mainly we need to make sure that a constant signal is reproduced by using balanced multiwavelets with constant coefficient vectors. The discussion and computation of balancing the above multiwavelets are presented in [38].

Multivariate Orthonormal Wavelets

Univariate wavelets have found successful applications in signal processing. To apply wavelet methods to digital image processing, we have to construct bivariate wavelets. The most commonly used method is the tensor product of univariate wavelets. This construction leads to a separable wavelet which has a disadvantage of giving a particular importance to the horizontal and vertical directions. Much effort has been spent on constructing non-separable bivariate wavelets in the last ten years. In this and the following three sections, we survey some methods for constructing bivariate and multivariate non-separable compactly supported orthonormal, biorthogonal, pre-wavelets as well as tight wavelet frames. All the discussion is based on the commonly used uniform dilation matrix $2I_d$ d-dimensional Euclidean space $\mathbb{R}^d$, where I_d is the identity of $d \times d$ matrix. Due to the space limitation, we are not able to include all the constructions. We refer the reader to the following literature [21,33,45,46,87,88], for other methods of constructing nonseparable compactly supported orthonormal wavelets (cf. [48]).

All the construction of compactly supported orthonormal wavelets is based multiresolution analysis (MRA). For convenience, we restrict our attention to $\mathbb{R}^2$. To construct bivariate wavelets, we need to solve the following two mathematical problems.

(1) Find a Laurent polynomial

$$m_0(x, y) = \sum_{\substack{-M \le j \le M \\ -N \le k \le N}} c_{jk} x^j y^k$$

so normalized that $m_0(1, 1) = 1$ satisfying

$$|m_0(x, y)|^2 + |m_0(-x, y)|^2 + |m_0(x, -y)|^2 + |m_0(-x, -y)|^2 = 1 \quad (57)$$

for $x = e^{i\xi}$ and $y = e^{i\eta}$ with $\xi, \eta \in \mathbb{R}$.

(2) Find another three Laurent polynomials $m_j(x, y)$, $j = 1, 2, 3$ such that the following matrix $\mathcal{M}$ is unitary, i. e.,

$$\mathcal{M} = \begin{bmatrix} m_0(x, y) & m_0(-x, y) & m_0(x, -y) & m_0(-x, -y) \\ m_1(x, y) & m_1(-x, y) & m_1(x, -y) & m_1(-x, -y) \\ m_2(x, y) & m_2(-x, y) & m_2(x, -y) & m_2(-x, -y) \\ m_3(x, y) & m_3(-x, y) & m_3(x, -y) & m_3(-x, -y) \end{bmatrix} \quad (58)$$

is unitary.

The condition (58) is called the perfect reconstruction condition as in Sect. "Definition of Filters".

Since $m_0(1, 1) = 1$, let

$$\widehat{\phi}(\xi, \eta) = \prod_{k=1}^{\infty} m_0(e^{i\xi/2^k}, e^{i\eta/2^k}) \quad (59)$$

be the refinable function associated with m_0. Then the condition (57) implies that $\phi \in L_2(\mathbb{R}^2)$.

In order to see that ϕ is orthonormal, we may apply the multidimensional generalization of Cohen's condition or Lawton's condition (cf. [24]). For simplicity, let us assume that ϕ is orthonormal in the sense that

$$\int_{\mathbb{R}^2} \phi(x,y)\phi(x-k,y-j)\mathrm{d}x\mathrm{d}y = \begin{cases} 1, & \text{if} \quad j=k=0 \\ 0, & \text{otherwise} \end{cases}$$

for all $j,k \in \mathbb{Z}$ and ϕ generate a multiresolution approximation of $L_2(\mathbb{R}^2)$. That is, let

$$V_j := \operatorname{span}\{2^j\phi(2^jx-m, 2^jy-n), m,n \in Z\}$$

for $j \in Z$. $\bigcup_{j\in\mathbb{Z}} V_j$ is dense in $\mathbb{R}^2$ and $\bigcap_{j\in\mathbb{Z}} V_j = \{0\}$.

Let $\widehat{\psi}_k(\xi,\eta) = m_k(\mathrm{e}^{i\xi/2}, \mathrm{e}^{i\eta/2})\widehat{\phi}(\xi/2,\eta/2)$ and

$$W_{0,k} := \operatorname{span}\{2^j\psi(2^jx-m, 2^jy-n), m,n \in Z\}$$

for $k = 1,2,3$. In order for ψ_k to be orthonormal wavelets, we need to have

1) ψ_k is orthonormal for $k = 1,2,3$ and
2)

$$V_1 = V_0 \oplus W_{0,1} \oplus W_{0,2} \oplus W_{0,3}\,.$$

These conditions 1) and 2) are equivalent to (58) (cf. [40]). Therefore, we mainly review recent methods for solving (57) and (58).

Method of Tensor Product

Starting with univariate wavelets, we can use the method of tensor product to construct separable wavelets. Let us recall the method below.

Let $P_a(x)$ and $P_b(x)$ be two univariate Laurent polynomials satisfying $P_a(1) = 1$, $P_b(1) = 1$,

$$|P_a(x)|^2+|P_a(-x)|^2 = 1 \quad \text{and} \quad |P_b(x)|^2+|P_b(-x)|^2 = 1.$$

Letting $\widetilde{Q_a} = xP_a(-1/x)$ and $\widetilde{Q_b}(x) = xP_b(-1/x)$ be the Laurent polynomials associated with P_a and P_b, respectively, we know

$$P_a(x)\overline{\widetilde{Q_a}(x)} + P_a(-x)\overline{\widetilde{Q_a}(-x)} = 0$$
$$P_b(y)\overline{\widetilde{Q_b}(y)} + P_b(-y)\overline{\widetilde{Q_b}(-y)} = 0\,.$$

With these relations, we define

$$\begin{aligned} m_0(x,y) &= P_a(x)P_b(y), \\ m_1(x,y) &= \widetilde{Q_a}(x)P_b(y), \\ m_2(x,y) &= P_a(x)\widetilde{Q_b}(y), \\ m_3(x,y) &= \widetilde{Q_a}(x)\widetilde{Q_b}(y)\,. \end{aligned}$$

Then, m_0 satisfies (57) and m_j, $j = 0,1,2,3$ satisfy (58). This is an easy method. However, the method emphasizes horizontal and vertical directions which may not be desirable for applications. Therefore, these prompt for construction of nonseparable wavelets.

The Ayache Method

In [1], Ayache proposed two methods for constructing bivariate nonseparable compactly supported orthonormal wavelets. The wavelets constructed using one of the methods is called semi-separable wavelets.

Starting with a separable filter $m(x,y) = P_a(x)P_b(y)$, we write

$$m(x,y) = P_a(x)\frac{P_b(y)+P_b(-y)}{2} + P_a(x)\frac{P_b(y)-P_b(-y)}{2}\,.$$

Let

$$\begin{aligned} (P_b)_e(y) &= \frac{1}{2}(P_b(y)+P_b(-y)), \\ (P_b)_o(y) &= \frac{1}{2}(P_b(y)-P_b(-y))\,. \end{aligned}$$

be the even and odd part of $P_b(x)$. That is, $m(x,y) = P_a(x)(P_b)_e(y) + P_a(x)(P_b)_o(y)$. Let $P_c(x)$ be another univariate CQF which is different from $P_a(x)$. We define

$$\begin{aligned} m_0(x,y) &= P_a(x)(P_b)_e(y) + P_c(x)(P_b)_o(y) \\ m_1(x,y) &= \widetilde{P_a}(x)(P_b)_e(y) + \widetilde{P_c}(x)(P_b)_o(y) \\ m_2(x,y) &= P_a(x)(\widetilde{P_b})_e(y) + P_c(x)(\widetilde{P_b})_o(y) \\ m_3(x,y) &= \widetilde{P_a}(x)(\widetilde{P_b})_e(y) + \widetilde{P_c}(x)(\widetilde{P_b})_o(y)\,. \end{aligned}$$

Then we have

Lemma 16 *Let $m_0(x,y)$ be a Laurent polynomial defined above. Then $m_0(x,y)$ satisfies (57) and m_j, $j = 0,1,2,3$ satisfy (58).*

Let us give some simple examples.

Example 23 Let $P_a(x) = P_b(x) = \frac{1+x}{2}$ and $P_c(x) = \frac{1+x^3}{2}$. Then

$$m(x,y) = \frac{1+x}{4} + \frac{1+x^3}{4}y$$

is a simple bivariate nonseparable filter satisfying (57). It can be easily verified that $m(x,y)$ satisfies the bivariate Cohen condition for the orthonormality. Thus, $m(x,y)$ generates a bona fide scaling function ϕ.

Example 24 Let $P_a(x) = P_b(x) = \frac{1+x}{2}$ and

$$P_c(x) = \frac{1+\sqrt{3}}{8} + \frac{3+\sqrt{3}}{8}x + \frac{3-\sqrt{3}}{8}x^2 + \frac{1-\sqrt{3}}{8}x^3 .$$

Then $m(x, y) = \frac{1+x}{4} + \frac{y}{2}P_c(x)$ is another simple bivariate nonseparable filter satisfying (57).

When P_a and P_c are masks associated with scaling functions ϕ_a and ϕ_c, if P_b is sufficiently close to P_a, then $m_0(x, y)$ defined above will generate an orthonormal refinable function ϕ. Indeed, m_0 will satisfy the bivariate generalization of Cohen's orthonormal condition when P_b is sufficiently close to P_a (cf. [1] and [2]).

The Maass–Ayache Method

This is the second method that Ayache proposed in [1]. Let $\lambda(x, y)$ and $\mu(x, y)$ be two even Laurent polynomials satisfying

$$|\lambda(x, y)|^2 + |\mu(x, y)|^2 = 1 ,$$

where $\lambda(x, y)$ also satisfies $\lambda(1, 1) = 1$. We define

$$\begin{aligned}
m_0(x, y) &= \lambda(x, y)P_a(x)P_b(x) + \mu(x, y)\widetilde{P_a}(x)P_b(y)\\
m_1(x, y) &= \overline{\mu(x, y)}P_a(x)P_b(y) - \overline{\lambda(x, y)}\widetilde{P_a}(x)P_b(y)\\
m_2(x, y) &= \lambda(x, y)P_a(x)\widetilde{P_b}(y) + \mu(x, y)\widetilde{P_a}(x)\widetilde{P_b}(y)\\
m_3(x, y) &= \overline{\mu(x, y)}P_a(x)\widetilde{P_b}(y) - \overline{\lambda(x, y)}\widetilde{P_a}(x)\widetilde{P_b}(y) .
\end{aligned}$$

Lemma 17 *$m(x, y)$, $m_1(x, y)$, $m_2(x, y)$, $m_3(x, y)$ so defined above satisfy the perfect reconstruction condition (58).*

Next let $q(x)$ be a nonzero Laurent polynomial with $0 \le q(x) \le 1$ and $q(1) = 0$. We define

$$\lambda(x) = 1 - \frac{1}{4}q(x)$$

and $\nu(x)$ such that

$$|\lambda(x)|^2 + |\nu(x)|^2 = 1 . \tag{60}$$

We now define $m_{N,0}(x, y)$ by

$$m_{N,0}(x, y) = D_N(y)(\lambda(x^2)D_N(x) + \nu(x^2)\tilde{D}_N(x)) , \tag{61}$$

where D_N is the filter associated with Daubechies' wavelet and $\widetilde{D}_N(x) = x\overline{D_N(-x)}$ is a conjugate filter of $D_N(x)$.

We refer [1] for the discussion of the orthonormality and regularity of the refinable function ϕ_N generated by $m_{N,0}$.

Also, in [69], Maass proposed a similar method for $m_0(x, y)$ such that the refinable function ϕ associated with m_0 satisfies the orthonormal condition under the dilation matrix $\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$: i. e.,

$$|m_0(x, y)|^2 + |m_0(-x, -y)|^2 = 1 . \tag{62}$$

His method can be given as follows: Let λ and ν be two even Laurent polynomials satisfying (60) with $\lambda(1, 1) = 1$. Then

$$m_0(x, y) = \lambda(x, y)D_N(x) + \nu(x, y)\widetilde{D}_N(x)$$

satisfies (62). Comparing with (61), we can see that the Maass method and the method Ayache used here are very similar. Thus, we call the construction the Maass–Ayache method.

The Belogay and Wang Method

In [5], Belogay and Wang constructed nonseparable compactly supported orthonormal wavelets using dilation matrix $\begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix}$. However, their method can be modified to construct nonseparable wavelets using the dilation matrix $\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$. We begin with

Lemma 18 *Let $P_a(x)$ and $P_b(x)$ be two Laurent polynomials of $x = e^{i\omega}$. Suppose that P_a and P_b satisfy that $P_a(1) = 1$,*

$$|P_a(x)|^2 + |P_a(-x)|^2 + |P_b(x)|^2 + |P_b(-x)|^2 = 1 \tag{63}$$

and

$$m_a(x)m_b(1/x) + m_a(-x)m_b(-1/x) = 0 . \tag{64}$$

Then, letting

$$m_0(x, y) = (P_a(x) + yP_b(x))(P_a(y) + x^2P_b(y)) ,$$

$m_0(1, 1) = 1$ and $m_0(x, y)$ satisfies (57).

Next we define

$$\begin{aligned}
m_1(x, y) &= y(P_a(x) + yP_b(x))(P_a(-1/y)\\
&\quad + x^2P_b(-1/y)),\\
m_2(x, y) &= x(P_a(-1/x) + 1/yP_b(-1/x))\\
&\quad (P_a(y) + x^2P_b(y)),\\
m_3(x, y) &= xy(P_a(-1/x) + 1/yP_b(-1/x))\\
&\quad (P_a(-1/y) + x^2P_b(-1/y)) .
\end{aligned}$$

Then it is easy to verify the following

Lemma 19 *Let $m(x, y)$ and $m_i(x, y)$, $i = 1, 2, 3$ be Laurent polynomials defined above. They satisfy the perfect reconstruction condition (58).*

Next we define these P_a and P_b which satisfy (63) and (64). Let $T(x) = 2^{N-1}$ and $S(x)$ such that

$$|S(x)|^2 = \sum_{k=0}^{N-1} \binom{N}{k} \left(\sin^2 \frac{\omega}{2}\right)^k \left(\cos^2 \frac{\omega}{2}\right)^{N-k} + \frac{3}{4}\left(\sin^2 \frac{\omega}{2}\right)^N$$

with $x = e^{i\omega}$. Then S and T satisfy

$$|S(x)|^2 + |T(1/x)|^2 \left|\frac{1-x}{4}\right|^{2N} = 1 .$$

Lemma 20 *Let*

$$P_a(x) = \left(\frac{1+x}{2}\right)^N \mathcal{L}_N(x) S(x^2) x^\nu$$

$$P_b(x) = \left(\frac{1+x}{2}\right)^N \left(\frac{1-x}{2}\right)^{2N} \mathcal{L}_N(-1/x) T(1/x^2) ,$$

where $|\mathcal{L}_N(x)|^2 = \sum_{k=0}^{N-1} \binom{n-1+k}{k} \left(\sin^2 \frac{\omega}{2}\right)^k$, ν is an integer such that $\nu + N$ is odd. Then $P_a(x)$ and $P_b(x)$ satisfy (63) and (64).

This is the main step for constructing the Belogay–Wang nonseparable wavelets associated with dilation matrix $2I_2$. We omit the discussion on the orthonormality and regularity here. See their paper for details.

The Karoui Method

In [43], Karoui proposed the following method to design nonseparable wavelets. Starting with a Laurant polynomial $m(x, y)$ which satisfies

$$|m(x, y)|^2 + |m(-x, -y)|^2 = 1 , \tag{65}$$

we define

$$m_0(x, y) = m(x, y) m(x/y, xy) .$$

Then we have the following

Lemma 21 *Let m_0 be defined above. Then $m_0(x, y)$ satisfies (57) for $|x| = 1$ and $|y| = 1$.*

Next we define

$$m_1(x, y) = x/y\, m(x, y) \overline{m(-x/y, -xy)}$$
$$m_2(x, y) = x \overline{m(-x, -y)} m(x/y, xy),$$
$$m_3(x, y) = 1/y \overline{m(-x, -y) m(-x/y, -xy)} .$$

Then we have

Lemma 22 *The four filters m_j, $j = 0, 1, 2, 3$ satisfy the perfect reconstruction condition (58).*

Note that (65) is the same as (62) before. By the Maass method, we can construct m satisfy (65).

The He and Lai Method

In [35], He and Lai construct many examples of nonseparable orthonormal wavelets. The constructive method starts with Laurent polynomial

$$m_0(x, y) = \sum_{0 \le j \le 3, 0 \le k \le 3} c_{j,k} x^j y^k$$

with $x = e^{i\xi}$ and $y = e^{i\eta}$. First we write $m(x, y)$ in its polyphase form:

$$m(x, y) = f_0(x^2, y^2) + x f_1(x^2, y^2) + y f_2(x^2, y^2) + xy f_3(x^2, y^2) ,$$

where

$$f_\nu(x, y) = a_\nu + b_\nu x + c_\nu y + d_\nu xy , \quad \nu = 0, 1, 2, 3 .$$

Lemma 23 *$m(x, y)$ satisfies (57) if and only if the following 5 nonlinear equations hold*

$$\sum_{\nu=0}^{3} (a_\nu b_\nu + c_\nu d_\nu) = 0 , \quad \sum_{\nu=0}^{3} (a_\nu c_\nu + b_\nu d_\nu) = 0 ,$$

$$\sum_{\nu=0}^{3} a_\nu d_\nu = 0 , \quad \sum_{\nu=0}^{3} b_\nu c_\nu = 0 ,$$

and

$$\sum_{\nu=0}^{3} \left(a_\nu^2 + b_\nu^2 + c_\nu^2 + d_\nu^2\right) = \frac{1}{4} .$$

The requirements $m(1, 1) = 1$ and $m(-1, y) = 0 = m(x, -1)$ imply the following 5 linear equations

$$\sum_{\nu=0}^{3} a_\nu + b_\nu + c_\nu + d_\nu = 1 ,$$

$$a_\nu + b_\nu + c_\nu + d_\nu = \frac{1}{4} , \quad \nu = 0, 1, 2, 3 .$$

We were able to find a complete solution for $c_{j,k}, 0 \le j, k \le 3$.

Theorem 15 *Let*

$$m(x,y) = \sum_{0\le j\le 3, 0\le k\le 3} c_{j,k} x^j y^k = \frac{(1+x)(1+y)}{16} \times \left(a_{00} + a_{10}x + a_{01}y + a_{11}xy + a_{20}x^2 + a_{21}x^2y + a_{12}xy^2 + a_{22}x^2y^2 + a_{02}y^2\right) \quad (66)$$

with

$$\begin{cases} a_{00} &= 1 + \sqrt{2}(\cos\alpha + \cos\beta) + 2\cos\theta\cos\xi \\ a_{10} &= \sqrt{2}(\sin\alpha - \cos\alpha) - 2\cos\theta\cos\xi \\ & \quad +2\cos\theta\sin\xi \\ a_{01} &= \sqrt{2}(\sin\beta - \cos\beta) - 2\cos\theta\cos\xi \\ & \quad +2\sin\theta\cos\eta \\ a_{11} &= 2(\cos\theta\cos\xi + \sin\theta\sin\eta - \cos\theta\sin\xi \\ & \quad -\sin\theta\cos\eta) \\ a_{20} &= 1 + \sqrt{2}(\cos\beta - \sin\alpha) - 2\cos\theta\sin\xi \\ a_{02} &= 1 + \sqrt{2}(\cos\alpha - \sin\beta) - 2\sin\theta\cos\eta \\ a_{21} &= \sqrt{2}(\sin\beta - \cos\beta) - 2\sin\theta\sin\eta \\ & \quad +2\cos\theta\sin\xi \\ a_{12} &= \sqrt{2}(\sin\alpha - \cos\alpha) - 2\sin\theta\sin\eta \\ & \quad +2\sin\theta\cos\eta \\ a_{22} &= 1 - \sqrt{2}(\sin\alpha + \sin\beta) + 2\sin\theta\sin\eta\,. \end{cases}$$

Suppose that $\alpha, \beta, \theta, \xi, \eta$ satisfy the following

$$\cos\theta\cos\xi + \cos\theta\sin\xi + \sin\theta\cos\eta + \sin\theta\sin\eta = 2\sin\left(\alpha + \frac{\pi}{4}\right)\sin\left(\beta + \frac{\pi}{4}\right)\,. \quad (67)$$

Then $m_0(x,y)$ satisfies (57). On the other hand, it $m_0(x,y)$ satisfies (57), then the coefficients c_{ij} can be given in the above form and satisfy (67).

Theorem 16 *For any given $m(x,y)$ in (66) satisfying (57), one can construct Laurant polynomials m_1, m_2, and m_3 such that m_j, $j = 0,1,2,3$ satisfy the perfect reconstruction condition (58).*

Their proof is constructive and has been implemented in MATLAB. See [35] for details. Let us give two examples of $m(x,y)$.

Example 25 The following are two examples which has rational coefficients. One can construct m_j, $j = 1,2,3$ associated with any of the following such that they satisfy the perfect reconstruction condition (58).

$$m(x,y) = \frac{(1+x)(1+y)}{100}(11 + 6x - 2x^2 + 6y + 13xy - 4x^2y - 2y^2 - 4xy^2 + x^2y^2)$$

$$m(x,y) = \frac{(1+x)(1+y)}{3468}(544 + 120x - 52x^2 + 120y + 416xy - 128x^2y - 52y^2 - 128xy^2 + 27x^2y^2)\,.$$

There are many other examples including those which are continuous. We refer to [35] for detail.

Method of Symmetry

Another easy method to construct bivariate compactly supported wavelets is to use the method of symmetry. Suppose that a Laurent polynomial $m_0(x,y)$ satisfies a symmetry property: $m_0(1/x, 1/y) = x^{-M}y^{-N}m_0(x,y)$. Suppose that $m_0(x,y)$ satisfies (57). Then the other three filters m_j, $j = 1,2,3$ can be easily obtained by

$$m_1(x,y) = m_0(-x,y)$$
$$m_2(x,y) = x\cdot m_0(x,-y)$$
$$m_3(x,y) = x\cdot m_0(-x,-y)\,.$$

In [57], we found the complete solutions for

$$m_0(x,y) = \sum_{0\le j,k\le 5} c_{j,k}x^jy^k$$

which satisfies the orthonormal condition (57), symmetry condition

$$m_0(1/x, 1/y) = x^{-5}y^{-5}m_0(x,y)$$

and low-pass feature $m_0(-1,y) = m_0(x,-1) = 0$. We leave the detail to [57].

The Multiwavelet Method

Assume that ϕ is a given compactly supported scaling function which generates an MRA. We now discuss how to construct compactly supported orthonormal wavelets ψ_j, $j = 1,\ldots,n$ with $n \ge 3$ associated with ϕ. For simplicity, let

$$m(\xi,\eta) = \sum_{0\le j\le 5, 0\le k\le 5} c_{j,k}e^{i(j\xi + k\eta)}\,.$$

Consider a function vector

$$\Phi(x,y) = \begin{bmatrix} 2\phi(2x,2y) \\ 2\phi(2x-1,2y) \\ 2\phi(2x,2y-1) \\ 2\phi(2x-1,2y-1)\,. \end{bmatrix}$$

Sine ϕ is orthonormal, so is $\Phi(x,y)$, i. e.,

$$\int_{\mathbb{R}^2} \Phi(x-\ell, y-k)\Phi(x,y)^{\mathrm{T}}\mathrm{d}x\mathrm{d}y = I_{4\times 4}\delta_{0,\ell}\delta_{0,k}\,.$$

Writing $\Phi = (\phi_1,\phi_2,\phi_3,\phi_4)^{\mathrm{T}}$, we let

$$\widetilde{V}_k = \mathrm{span}_{L_2}\{\phi_j(2^kx - \ell, 2^ky - m)\,, \quad \ell, m \in Z\,, \quad j = 1,2,3,4\}\,.$$

Note that is $\tilde{V}_0 = V_1$. It follows that $\{\tilde{V}_k, k \in Z\}$ forms an MRA and hence Φ generates an MRA. It is clear that $\Phi(x, y)$ is a refinable vector. That is,

$$\Phi(x, y) = \sum_{\ell,m\in Z} M_{\ell,m}\Phi(2x - \ell, 2y - m) .$$

In terms of Fourier transform, we have

$$\hat{\Phi}(\xi, \eta) = M(\xi/2, \eta/2)\hat{\Phi}(\xi/2, \eta/2) ,$$

where $M(\xi, \eta) = \frac{1}{4}\sum_{\ell,m} M_{\ell,m}e^{i(\ell\xi+m\eta)}$.

It follows that

$$\sum_{\ell,m\in Z} \hat{\Phi}(\xi+2\ell\pi, \eta+2m\pi)\hat{\Phi}(\xi+2\ell\pi, \eta+2m\pi)^* = I_{4\times 4}.$$

Thus, we have

$$\begin{aligned} &M(\xi, \eta)M(\xi, \eta)^* + M(\xi + \pi, \eta)M(\xi + \pi, \eta)^* \\ &+ M(\xi, \eta + \pi)M(\xi, \eta + \pi)^* + M(\xi + \pi, \eta + \pi) \\ &\qquad M(\xi + \pi, \eta + \pi)^* = I_{4\times 4} . \end{aligned}$$

Let $\tilde{W}_k$ be the orthogonal complement of $\tilde{V}_k$ in $\tilde{V}_{k+1}$. We will construct three compactly supported orthonormal multi-wavelet vectors $\Psi_1, \Psi_2, \Psi_3 \in \tilde{W}_0$ such that

$$\tilde{V}_1 = \tilde{V}_0 \oplus \tilde{W}_0 ,$$

and for $j, k = 1, 2, 3$,

$$\int_{\mathbb{R}^2} \Psi_j(x - \ell, y - m)\Psi_k(x, y)^{\mathrm{T}}\mathrm{d}x\mathrm{d}y = 0 , \quad j \neq k ,$$

where

$$\begin{aligned} \tilde{W}_0 = \mathrm{span}_{L_2}\{&\Psi_{j,k}(x - \ell, y - m) , \\ &\ell, m \in Z, j = 1, 2, 3 , \quad k = 1, \ldots, 4\} \end{aligned}$$

and $\Psi_j = (\Psi_{j1}, \Psi_{j2}, \Psi_{j3}, \Psi_{j4})^{\mathrm{T}}$, $j = 1, 2, 3$. Writing

$$\hat{\Psi}_j(\xi, \eta) = M_j(\xi/2, \eta/2)\hat{\Phi}(\xi/2, \eta/2) , \quad j = 1, 2, 3 ,$$

we need to find matrices $M_j(\xi, \eta)$ with polynomial entries in $(\mathrm{e}^{i\xi}, \mathrm{e}^{i\eta})$ such that

$$\begin{bmatrix} M(\xi, \eta) & M(\xi + \pi, \eta) & M(\xi, \eta + \pi) & M(\xi + \pi, \eta + \pi) \\ M_1(\xi, \eta) & M_1(\xi + \pi, \eta) & M_1(\xi, \eta + \pi) & M_1(\xi + \pi, \eta + \pi) \\ M_2(\xi, \eta) & M_2(\xi + \pi, \eta) & M_2(\xi, \eta + \pi) & M_2(\xi + \pi, \eta + \pi) \\ M_3(\xi, \eta) & M_3(\xi + \pi, \eta) & M_3(\xi, \eta + \pi) & M_3(\xi + \pi, \eta + \pi) \end{bmatrix} \tag{68}$$

is a unitary matrix.

By the above properties and the fact that Φ generates an MRA, we know that $\{\Psi_{jk}, j = 1, 2, 3, k = 1, 2, 3, 4\}$ are compactly supported orthonormal wavelet functions for $L_2(\mathbb{R}^2)$ if we have (68).

To solve this matrix extension problem, i. e., finding M_j, $j = 1, 2, 3$ such that (68) holds, we need the following lemma whose proof is based on the ideas in [35].

Theorem 17 *Let $M(\xi, \eta)$ be a matrix of size 4×4 with polynomial entries in $e^{i\xi}$ and $e^{i\eta}$ with coordinate degrees $\leq (3, 3)$. Suppose that*

$$\begin{aligned} I_{4\times 4} = M(\xi, \eta)M(\xi, \eta)^* &+ M(\xi + \pi, \eta)M(\xi + \pi, \eta)^* \\ &+ M(\xi, \eta + \pi)M(\xi, \eta + \pi)^* \\ &+ M(\xi + \pi, \eta + \pi)M(\xi + \pi, \eta + \pi)^* . \end{aligned}$$

Then there exist polynomial matrix M_j, $j = 1, 2, 3$ such that the matrix in (68) is unitary.

Therefore we can have

Theorem 18 *Suppose that $\phi(x, y) \in L_2(\mathbb{R}^2)$ is a scaling function associated with dilation matrix $2I_2$. Let*

$$m(\xi, \eta) = \frac{1}{4}\sum_{\substack{0\leq j\leq 5\\ 0\leq k\leq 5}} c_{jk}e^{i(j\xi+k\eta)}$$

be the mask associated with ϕ. Then there exist 12 compactly supported orthonormal wavelets ψ_{jk}, $j = 1, 2, 3, k = 1, 2, 3, 4$ such that translates dilation of these $\psi_{j,k}$'s form an orthonormal basis for $L_2(\mathbb{R}^2)$.

The detail of the proof is contained in [49]. What happens when the support size of ϕ is bigger. Suppose that

$$m(\xi, \eta) = \sum_{\substack{0\leq j\leq 9\\ 0\leq j\leq 9}} c_{jk}\mathrm{e}^{i(j\xi+k\eta)}$$

is a trigonometric polynomial associated with a scaling function ϕ. We will let

$$\Phi(x,y) = \begin{bmatrix} 4\phi(4x,4y) \\ 4\phi(4x+1,4y) \\ 4\phi(4x,4y+1) \\ 4\phi(4x+1,4y+1) \\ 4\phi(4x+2,4y) \\ 4\phi(4x+2,4y+1) \\ 4\phi(4x+2,4y+2) \\ 4\phi(4x+1,4y+2) \\ 4\phi(4x,4y+2) \\ 4\phi(4x+3,4y) \\ 4\phi(4x+3,4y+1) \\ 4\phi(4x+3,4y+2) \\ 4\phi(4x+3,4y+3) \\ 4\phi(4x+2,4y+3) \\ 4\phi(4x+1,4y+3) \\ 4\phi(4x,4y+3) \end{bmatrix}_{16\times 1}.$$

Then $\Phi = (\phi_1, \ldots, \phi_{16})^{\mathrm{T}}$ is an orthonormal scaling vector. Let

$$\widehat{V}_0 = \operatorname{span}_{L_2}\{\phi_j(x-\ell, y-m), \ell, m \in \mathbb{Z}, j = 1, \ldots, 16\}.$$

It is easy to see that $\widehat{V}_0 = V_2$. Thus, Φ generates a bona fide MRA. The above construction procedure can be simply extended for this Φ. We can construct three multiwavelet vectors of size 16×1. The details can be found in [49]. Therefore, we conclude

Theorem 19 *Suppose that $\phi(x,y) \in L_2(\mathbb{R}^2)$ is a scaling function associated with dilation matrix $2I_2$. Then there exist compactly supported orthonormal wavelets ψ_{jk}, $j = 1,2,3$, $k = 1, \ldots, n$ with appropriate n dependent on the size of the support of ϕ such that translates dilation of these $\psi_{j,k}$'s form an orthonormal basis for $L_2(\mathbb{R}^2)$.*

Biorthogonal Box Spline Wavelets

In a previous section, we have used B-spline function N_n to construct biorthogonal dual function $\tilde{B}_n$ and the associated compactly supported biorthogonal wavelets. Since bivariate box splines are a natural generalization of B-spline functions, we shall present a constructive method to find biorthogonal wavelets associated with box spline functions.

Let $B_{l,m,n}$ be the bivariate box spline function whose Fourier transform is

$$\widehat{B}_{l,m,n}(\omega_1,\omega_2) = \left(\frac{1-\mathrm{e}^{i\omega_1}}{i\omega_1}\right)^l \left(\frac{1-\mathrm{e}^{i\omega_2}}{i\omega_2}\right)^m \left(\frac{1-\mathrm{e}^{i(\omega_1+\omega_2)}}{i(\omega_1+\omega_2)}\right)^n.$$

(For properties of box spline functions, see [7,11,60] For computation of these bivariate box spline functions, see [47]). It is known that $B_{l,m,n}$ generates a multi-resolution approximation of $L_2(\mathbb{R}^2)$ (cf. [74]). We are interested in constructing a compactly supported function $\tilde{B}_{l,m,n}$ generating a multi-resolution approximation of $L_2(\mathbb{R}^2)$ which is a biorthogonal dual to $B_{l,m,n}$ in the following sense:

$$\iint_{\mathbb{R}^2} B_{l,m,n}(x-j, y-k)\tilde{B}_{l,m,n}(x-j', y-k')\mathrm{d}x\mathrm{d}y = \delta_{j,j'}\delta_{k,k'} \quad (69)$$

for all integers $j, k \in \mathbb{Z}$, where $\delta_{j,k}$ is the standard Kronecker notation defined by $\delta_{j,k} = 0$ if $j \neq k$ and $\delta_{j,k} = 1$ if $j = k$ and $\mathbb{Z}$ is the collection of all integers.

We are furthermore interested in constructing compactly supported biorthogonal wavelets ψ_j, $j = 1,2,3$ and $\tilde{\psi}_j$, $j = 1,2,3$ and two families of FIR filters $\{M_j, j = 0,1,2,3\}$ and $\{J_j, j = 0,1,2,3\}$ with

$$\widehat{\psi}_j(\omega_1,\omega_2) = M_j(\mathrm{e}^{i\frac{\omega_1}{2}}, \mathrm{e}^{i\frac{\omega_2}{2}})\widehat{B}_{l,m,n}\left(\frac{\omega_1}{2}, \frac{\omega_2}{2}\right), \quad j = 1,2,3, \quad (70)$$

and

$$\widehat{\tilde{\psi}}_j(\omega_1,\omega_2) = J_j\left(\mathrm{e}^{i\frac{\omega_1}{2}}, \mathrm{e}^{i\frac{\omega_2}{2}}\right)\widehat{\tilde{B}}_{l,m,n}\left(\frac{\omega_1}{2}, \frac{\omega_2}{2}\right), \quad j = 1,2,3, \quad (71)$$

such that the dilations and translates of the ψ_j's and $\tilde{\psi}_j$'s form two dual Riesz bases for $L_2(\mathbb{R}^2)$ and the two families form an exact reconstruction of synthesis/analysis filter bank for image/data processing.

In the following we mainly follow the construction of biorthogonal box spline wavelets in [36]. Denote $z_1 = \mathrm{e}^{i\omega_1}$ and $z_2 = \mathrm{e}^{i\omega_2}$. Let

$$M_0(z_1,z_2) = \left(\frac{1+z_1}{2}\right)^l \left(\frac{1+z_2}{2}\right)^m \left(\frac{1+z_1z_2}{2}\right)^n$$

be a mask associated with the box spline function $B_{l,m,n}$. We look for a mask $J_0(z_1,z_2)$ in the form

$$\overline{J_0(z_1,z_2)} = \left(\frac{1+z_1}{2}\right)^{\tilde{n}-l} \left(\frac{1+z_2}{2}\right)^{\tilde{n}-m} \times \left(\frac{1+z_1z_2}{2}\right)^{\tilde{m}-n} H(z_1,z_2)D(z_1z_2) \quad (72)$$

with $\tilde{n} > l$, $\tilde{n} > m$ and odd integer $\tilde{m} > n$ such that

$$M_0(z_1,z_2)\overline{J_0(z_1,z_2)} + M_0(-z_1,z_2)\overline{J_0(-z_1,z_2)} + M_0(z_1,-z_2)\overline{J_0(z_1,-z_2)} + M_0(-z_1,-z_2)\overline{J_0(-z_1,-z_2)} = 1\,. \quad (73)$$

Recall from (14) that there exists a polynomial $P_N(y)$ of degree $< N$ such that

$$(1-y)^N P_N(y) + y^N P_N(1-y) = 1\,, \quad (74)$$

In fact, we have

$$P_N(y) = \sum_{k=0}^{N-1} \binom{2N-1}{k} (1-y)^{N-1-k} y^k\,.$$

Theorem 20 *Let $\tilde{n} > n$ and $\tilde{m} = 2\widehat{m}+1$. Let $J_0(z_1,z_2)$ be defined in (72) with H and D defined by*

$$H(z_1,z_2) = \sum_{k=0}^{\tilde{n}-1} \binom{2\tilde{n}-1}{k} \left(\frac{1+z_1}{2}\frac{1+z_2}{2}\right)^{\tilde{n}-1-k} \left(\frac{1-z_1}{2}\frac{1-z_2}{2}\right)^{k}\,, \quad (75)$$

and

$$D(e^{i(\omega_1+\omega_2)}) = e^{-i(\omega_1+\omega_2)N} P_{\tilde{n}+\tilde{m}}\left(\sin^2\left(\frac{\omega_1+\omega_2}{2}\right)\right)\,. \quad (76)$$

Then J_0 is a dual of M_0 satisfying (73).

Proof We first note that

$$\frac{1+z_1z_2}{2} = \frac{1+z_1}{2}\frac{1+z_2}{2} + \frac{1-z_1}{2}\frac{1-z_2}{2}\,.$$

By letting $H(z_1,z_2)$ be defined in (75), we have, similar to the derivation of $P_N(y)$ in (14)

$$\begin{aligned}
&\left(\frac{1+z_1z_2}{2}\right)^{2\tilde{n}-1} \\
&= \left(\frac{1+z_1}{2}\frac{1+z_2}{2} + \frac{1-z_1}{2}\frac{1-z_2}{2}\right)^{2\tilde{n}-1} \\
&= \sum_{k=0}^{\tilde{n}-1} \binom{2\tilde{n}-1}{k} \cdot \left(\frac{1+z_1}{2}\frac{1+z_2}{2}\right)^{2\tilde{n}-1-k} \left(\frac{1-z_1}{2}\frac{1-z_2}{2}\right)^{k} \\
&\quad + \sum_{\ell=0}^{\tilde{n}-1} \binom{2\tilde{n}-1}{2\tilde{n}-1-\ell} \left(\frac{1+z_1}{2}\frac{1+z_2}{2}\right)^{\ell} \cdot \left(\frac{1-z_1}{2}\frac{1-z_2}{2}\right)^{2\tilde{n}-1-\ell} = \left(\frac{1+z_1}{2}\frac{1+z_2}{2}\right)^{\tilde{n}} \\
&\qquad \cdot H(z_1,z_2) + \left(\frac{1-z_1}{2}\frac{1-z_2}{2}\right)^{\tilde{n}} H(-z_1,-z_2)
\end{aligned}$$

and similarly,

$$\begin{aligned}
&\left(\frac{1-z_1z_2}{2}\right)^{2\tilde{n}-1} = \left(\frac{1-z_1}{2}\frac{1+z_2}{2} + \frac{1+z_1}{2}\frac{1-z_2}{2}\right)^{2\tilde{n}-1} = \left(\frac{1-z_1}{2}\frac{1+z_2}{2}\right)^{\tilde{n}} \\
&\qquad \cdot H(-z_1,z_2) + \left(\frac{1+z_1}{2}\frac{1-z_2}{2}\right)^{\tilde{n}} H(z_1,-z_2)\,.
\end{aligned}$$

With the definition of J_0 in (73), (73) may be simplified as follows:

$$\begin{aligned}
&M_0(z_1,z_2)\overline{J_0(z_1,z_2)} + M_0(-z_1,-z_2)\overline{J_0(-z_1,-z_2)} \\
&+ M_0(-z_1,z_2)\overline{J_0(-z_1,z_2)} + M_0(z_1,-z_2)\overline{J_0(z_1,-z_2)} \\
&= \left[\left(\frac{1+z_1}{2}\frac{1+z_2}{2}\right)^{\tilde{n}} H(z_1,z_2) + \left(\frac{1-z_1}{2}\frac{1-z_2}{2}\right)^{\tilde{n}} H(-z_1,-z_2)\right] \times \left(\frac{1+z_1z_2}{2}\right)^{\tilde{m}} D(z_1z_2) \\
&+ \left[\left(\frac{1-z_1}{2}\frac{1+z_2}{2}\right)^{\tilde{n}} H(-z_1,z_2) + \left(\frac{1+z_1}{2}\frac{1-z_2}{2}\right)^{\tilde{n}} H(z_1,-z_2)\right] \times \left(\frac{1-z_1z_2}{2}\right)^{\tilde{m}} D(-z_1z_2) \\
&= \left(\frac{1+z_1z_2}{2}\right)^{2\tilde{n}+\tilde{m}-1} D(z_1z_2) + \left(\frac{1-z_1z_2}{2}\right)^{2\tilde{n}+\tilde{m}-1} D(-z_1z_2)\,.
\end{aligned}$$

Let $\tilde{m} = 2\widehat{m}+1$ and $N = \tilde{n}+\widehat{m}$. Recall $z_1 = \mathrm{e}^{i\omega_1}$ and $z_2 = \mathrm{e}^{i\omega_2}$. Then the last equation may be simplified further:

$$\begin{aligned}
&\left(\cos^2\frac{\omega_1+\omega_2}{2}\right)^N \mathrm{e}^{i(\omega_1+\omega_2)N} D(\mathrm{e}^{i(\omega_1+\omega_2)}) \\
&+ \left(\sin^2\frac{\omega_1+\omega_2}{2}\right)^N (-1)^N \mathrm{e}^{i(\omega_1+\omega_2)N} D(-\mathrm{e}^{i(\omega_1+\omega_2)}).
\end{aligned}$$

Let $y = \sin^2\left(\frac{\omega_1+\omega_2}{2}\right)$ and recognize that $\mathrm{e}^{i(\omega_1+\omega_2)N} D(\mathrm{e}^{i(\omega_1+\omega_2)}) = P_N(y)$. We can see that the above equa-

tion is just the left-hand side of (74). Therefore, we have established the results of Theorem 20. □

We remark here that the filter $J_0(z_1, z_2)$ is a linear phase filter. It is known that the Fourier transform of box spline $B_{l,m,n}$ is

$$\widehat{B}_{l,m,n}(\omega_1, \omega_2) = \prod_{k=1}^{\infty} M_0\left(e^{i\frac{\omega_1}{2^k}}, e^{i\frac{\omega_2}{2^k}}\right) \in L_2(\mathbb{R}^2) .$$

We now construct the dual functions $\tilde{B}_{l,m,n}$ in terms of its Fourier transform by

$$\widehat{\tilde{B}}_{l,m,n}(\omega_1, \omega_2) = \prod_{k=1}^{\infty} J_0\left(e^{i\frac{\omega_1}{2^k}}, e^{i\frac{\omega_2}{2^k}}\right) .$$

To see $\tilde{B}_{l,m,n}$ is in $L_2(\mathbb{R}^2)$ and generates a MRA in the bivariate setting, we have the following (cf. [36])

Theorem 21 *Let $\tilde{n}$ and $\tilde{m}$ be large enough. Then $\tilde{B}_{l,m,n}$ is a well-defined compactly supported L_2 function. Furthermore, for any $\alpha > 0$, $\tilde{B}_{l,m,n} \in C^{\alpha}(\mathbb{R}^2)$ if $\tilde{n}$ and $\tilde{m}$ sufficiently large, e. g.,*

$$\tilde{n} > \frac{2(\max(l, m) + 2 + \alpha)}{2 - \log(3)/\log(2)},$$
$$\tilde{m} > 2\frac{\tilde{n}\log(3)/\log(2) + n - 1}{2 - \log(3)/\log(2)} + 1 .$$

We next show that $\tilde{B}_{l,m,n}$ defined above is a biorthogonal dual to $B_{l,m,n}$ in the sense of (69). We first see that $\widehat{\tilde{B}}_{l,m,n}$ is continuous and $\widehat{\tilde{B}}_{l,m,n}(0,0)\widehat{B}_{l,m,n}(0,0) = 1$. It is straightforward to prove

Theorem 22 *For any sufficiently large integers $\tilde{n}$ and $\tilde{m}$,*

$$\sum_{\ell \in z^2} \left|\widehat{B}_{l,m,n}((\omega_1, \omega_2) + 2\pi\ell)\widehat{\tilde{B}}_{l,m,n}((\omega_1, \omega_2) + 2\pi\ell)\right|^2 \geq C_2 > 0 .$$

The same arguments in the proof of Theorem 22 can also show that

$$\sum_{\ell \in \mathbb{Z}^2} \left|\widehat{\tilde{B}}_{l,m,n}((\omega_1, \omega_2) + 2\pi\ell)\right|^2 \geq C_1 . \tag{77}$$

It follows from Theorem 21 that

$$\sum_{\ell \in \mathbb{Z}^2} \left|\widehat{\tilde{B}}_{l,m,n}((\omega_1, \omega_2) + 2\pi\ell)\right|^2 \leq C_2 . \tag{78}$$

Thus, letting

$$V_0 = \operatorname{span}\{\tilde{B}_{l,m,n}(x - j, y - k), (j, k) \in \mathbb{Z}^2\} ,$$

the inequalities (77) and (78) imply that $\{\tilde{B}_{l,m,n}(x - j, y - k), (j, k) \in \mathbb{Z}^2\}$ is a Riesz basis for V_0. Letting $V_k := \{f(x/2^k, y/2^k) : \forall f(x, y) \in V_0\}$ for $k \in \mathbb{Z}$, we can show that $\bigcup_k V_k$ is dense in $L_2(\mathbb{R}^2)$ and $\bigcap_k V_k = \{0\}$. We leave the detail to the interested reader. Thus, we conclude that $\tilde{B}_{l,m,n}$ generates a multi-resolution approximation of $L_2(\mathbb{R}^2)$. These complete the proof of the following

Theorem 23 *Let $\tilde{n}$ and $\tilde{m}$ be sufficiently large. Then $\tilde{B}_{l,m,n}$ generates a multi-resolution approximation of $L_2(\mathbb{R}^2)$. Also, $\tilde{B}_{l,m,n}$ is a biorthogonal dual of $B_{l,m,n}$.*

Next we work on constructing biorthogonal wavelets associated with $B_{l,m,n}$ and $\tilde{B}_{l,m,n}$. As before we start with the construction of two families of Laurent polynomials satisfying the following

$$\begin{bmatrix} M_0(z_1, z_2) & M_1(z_1, z_2) & M_2(z_1, z_2) & M_3(z_1, z_2) \\ M_0(-z_1, z_2) & M_1(-z_1, z_2) & M_2(-z_1, z_2) & M_3(-z_1, z_2) \\ M_0(z_1, -z_2) & M_1(z_1, -z_2) & M_2(z_1, -z_2) & M_3(z_1, -z_2) \\ M_0(-z_1, -z_2) & M_1(-z_1, -z_2) & M_2(-z_1, -z_2) & M_3(-z_1, -z_2) \end{bmatrix} \times \begin{bmatrix} J_0(z_1, z_2) \\ J_1(z_1, z_2) \\ J_2(z_1, z_2) \\ J_3(z_1, z_2) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{79}$$

For convenience, let us denote by $A(M_0, M_1, M_2, M_3)$ the coefficient matrix in (79). In order to have compactly supported wavelets we need $M_j, J_j, j = 1, 2, 3$ to be Laurent polynomials in (z_1, z_2) and the matrix $A(M_0, M_1, M_2, M_3)$ must have a nonzero monomial determinant, i. e., $Cz_1^j z_2^k$.

To this end, we rewrite $M_j, j = 0, 1, 2, 3$ in its polyphase form

$$M_j(z_1, z_2) = f_{j0}\left(z_1^2, z_2^2\right) + z_1 f_{j1}\left(z_1^2, z_2^2\right) + z_2 f_{j2}\left(z_1^2, z_2^2\right) + z_1 z_2 f_{j3}\left(z_1^2, z_2^2\right) .$$

Similarly we have

$$\begin{aligned} M_j(-z_1, z_2) &= f_{j0}\left(z_1^2, z_2^2\right) - z_1 f_{j1}\left(z_1^2, z_2^2\right) \\ &\quad + z_2 f_{j2}\left(z_1^2, z_2^2\right) - z_1 z_2 f_{j3}\left(z_1^2, z_2^2\right) \\ M_j(z_1, -z_2) &= f_{j0}\left(z_1^2, z_2^2\right) + z_1 f_{j1}\left(z_1^2, z_2^2\right) \\ &\quad - z_2 f_{j2}\left(z_1^2, z_2^2\right) - z_1 z_2 f_{j3}\left(z_1^2, z_2^2\right) \\ M_j(-z_1, -z_2) &= f_{j0}\left(z_1^2, z_2^2\right) - z_1 f_{j1}\left(z_1^2, z_2^2\right) \\ &\quad - z_2 f_{j2}\left(z_1^2, z_2^2\right) + z_1 z_2 f_{j3}\left(z_1^2, z_2^2\right) . \end{aligned}$$

We can easily check

$$A(M_0, M_1, M_2, M_3) = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & z_1 & 0 & 0 \\ 0 & 0 & z_2 & 0 \\ 0 & 0 & 0 & z_1 z_2 \end{bmatrix} \times \begin{bmatrix} f_{00} & f_{10} & f_{20} & f_{30} \\ f_{01} & f_{11} & f_{21} & f_{31} \\ f_{02} & f_{12} & f_{22} & f_{32} \\ f_{03} & f_{13} & f_{23} & f_{33} \end{bmatrix}, \tag{80}$$

where $f_{jk} := f_{jk}(z_1^2, z_2^2)$'s. It is easy to see that for a given M_0, the existence of the matrix $A(M_0, M_1, M_2, M_3)$ such that its determinant is a monomial $Cz_1^{2\mu} z_2^{2\nu}$ is equivalent to the existence of $[f_{jk}]_{0\le j,k\le 3}$ whose determinant is a monomial.

It is clear from the expression of $M_0(z_1, z_2)$ associated with box spline $B_{l,m,n}$ that $M_0(z_1, z_2)$, $M_0(-z_1, z_2)$, $M_0(z_1, -z_2)$, $M_0(-z_1, -z_2)$ have no common zeros in $\mathbf{C}^2$, where $\mathbf{C}$ denotes the usual complex space. It follows that $f_{00}, f_{01}, f_{02}, f_{03}$ have no common zeros.

We further show that the first three polyphase terms $f_{00}(z_1^2, z_2^2)$, $f_{01}(z_1^2, z_2^2)$, $f_{02}(z_1^2, z_2^2)$ have no common zero in $(\mathbf{C})^2$ (cf. [36]).

Lemma 24 *Suppose that f_{0j}, $j = 0, \ldots, 3$ are polynomials in (z_1, z_2). Suppose that f_{00}, f_{01}, f_{02} have no common zeros in $(\mathbf{C})^2$. Then there exist $f_{k,j}$, $j = 0, 1, 2, 3$ and $k = 1, 2, 3$ such that the matrix $[f_{k,j}]_{0\le k,j\le 3}$ is of determinant ± 1.*

Proof By the well-known Hilbert Nullstellensatz there exist polynomials p_0, p_1, p_2 such that $p_0 f_{00} + p_1 f_{01} + p_2 f_{02} = 1$. Then it is easy to check that

$$\begin{bmatrix} f_{00} & 1 & 0 & 0 \\ f_{01} & 0 & 1 & 0 \\ f_{02} & 0 & 0 & 1 \\ f_{03} & -p_0(1-f_{03}) & -p_1(1-f_{03}) & -p_2(1-f_{03}) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -p_0(1-f_{03}) & -p_1(1-f_{03}) & -p_2(1-f_{03}) & 1 \end{bmatrix} \times \begin{bmatrix} f_{00} & 1 & 0 & 0 \\ f_{01} & 0 & 1 & 0 \\ f_{02} & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

which is obviously of determinant -1. We choose this matrix for $[f_{kj}]_{0\le k,j\le 3}$. □

By Lemma above, for M_0, we can find $\tilde{M}_1, \tilde{M}_2, \tilde{M}_3$ such that $A(M_0, \tilde{M}_1, \tilde{M}_2, \tilde{M}_3)$ has a determinant which is a nonzero monomial $Cz_1^j z_2^k$. A computation of the determinant of matrix $A(M_0, \tilde{M}_1, \tilde{M}_2, \tilde{M}_3)$ from the right-hand side of (80) yields the determinant is of even power, i. e., $j = 2j'$ and $k = 2k'$. Without loss of generality, we may simply assume

$$\det(A(M_0, \tilde{M}_1, \tilde{M}_2, \tilde{M}_3)) = 1$$

by absorbing $Cz_1^{2j'} z_2^{2k'}$ into $\tilde{M}_1$. Let us invert the matrix $(A(M_0, \tilde{M}_1, \tilde{M}_2, \tilde{M}_3))^{\mathrm{T}}$. From the definition of the inverse matrix, we know there exist polynomial entries $\tilde{J}_0, \bar{J}_1, \bar{J}_2$, and $\bar{J}_3$ such that

$$\left(A\left(M_0, \tilde{M}_1, \tilde{M}_2, \tilde{M}_3\right)^{\mathrm{T}}\right)^{-1} = A\left(\tilde{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3\right)$$

or equivalently,

$$\left(A\left(\tilde{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3\right)\right)^{-1} = \left(A\left(M_0, \tilde{M}_1, \tilde{M}_2, \tilde{M}_3\right)\right)^{\mathrm{T}}.$$

Since the determinant is 1, we know that, by Cramer's rule, M_0 is equal to the cofactor of $\tilde{J}_0$ in matrix $A(\tilde{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3)$. In particular, we have

$$M_0(z_1, z_2) = \det \begin{bmatrix} \bar{J}_1(-z_1, z_2) & \bar{J}_2(-z_1, z_2) & \bar{J}_3(-z_1, z_2) \\ \bar{J}_1(z_1, -z_2) & \bar{J}_2(z_1, -z_2) & \bar{J}_3(z_1, -z_2) \\ \bar{J}_1(-z_1, -z_2) & \bar{J}_2(-z_1, -z_2) & \bar{J}_3(-z_1, -z_2) \end{bmatrix}. \tag{81}$$

Note that expanding according to the first column of $A(\tilde{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3)$ and by using the definition of the inverse matrix, we have

$$\begin{aligned} 1 &= \det(A(\tilde{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3) \\ &= \tilde{J}_0(z_1, z_2) M_0(z_1, z_2) + \tilde{J}_0(-z_1, z_2) M_0(-z_1, z_2) \\ &\quad + \tilde{J}_0(z_1, -z_2) M_0(z_1, -z_2) + \tilde{J}_0(-z_1, -z_2) \\ &\quad M_0(-z_1, -z_2). \end{aligned}$$

Replacing the first column of matrix $A(\tilde{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3)$ by a column $[\overline{J_0(z_1, z_2)}, \overline{J_0(-z_1, z_2)}, \overline{J_0(z_1, -z_2)}, \overline{J_0(-z_1, -z_2)}]^{\mathrm{T}}$ with J_0 being defined in (72), we get a new matrix $A(\bar{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3)$ whose determinant is

$$\begin{aligned} \det(A(\bar{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3) &= \bar{J}_0(z_1, z_2) M_0(z_1, z_2) \\ &+ \bar{J}_0(-z_1, z_2) M_0(-z_1, z_2) + \bar{J}_0(z_1, -z_2) M_0(z_1, -z_2) \\ &+ \bar{J}_0(-z_1, -z_2) M_0(-z_1, -z_2) = 1 \end{aligned}$$

by (73). We compute the inverse of $A(\bar{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3)$ and write

$$A(\bar{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3)^{-1} = A(q_0, M_1, M_2, M_3)^{\mathrm{T}}.$$

By the definition of the inverse matrices, it is now easy to recognize that $q_0 = M_0$ since (81). That is, we have

$$A(\bar{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3)A(M_0, M_1, M_2, M_3)^{\mathrm{T}} = I_4$$

where I_4 stands for the identity matrix of 4×4. Hence,

$$A(M_0, M_1, M_2, M_3)A(\bar{J}_0, \bar{J}_1, \bar{J}_2, \bar{J}_3)^{\mathrm{T}} = I$$

which implies (79). Therefore, we have obtained the following

Theorem 24 *Let $M_0(z_1, z_2) = \left(\frac{1+z_1}{2}\right)^l \left(\frac{1+z_2}{2}\right)^m \left(\frac{1+z_1 z_2}{2}\right)^n$ and J_0 defined in (72). Then there exist M_1, M_2, M_3 and J_1, J_2, J_3 such that the exact reconstruction condition (79) holds.*

We are finally ready to define biorthogonal wavelet functions ψ_j and $\widetilde{\psi}_j$ in terms of Fourier transform by

$$\widehat{\psi}_j(\omega_1, \omega_2) = J_j\left(\mathrm{e}^{i\frac{\omega_1}{2}}, \mathrm{e}^{i\frac{\omega_2}{2}}\right)\widehat{B}_{l,m,n}\left(\frac{\omega_1}{2}, \frac{\omega_2}{2}\right),$$

and

$$\widehat{\widetilde{\psi}}_j(\omega_1, \omega_2) = M_j\left(\mathrm{e}^{i\frac{\omega_1}{2}}, \mathrm{e}^{i\frac{\omega_2}{2}}\right)\widehat{\widetilde{B}}_{l,m,n}\left(\frac{\omega_1}{2}, \frac{\omega_2}{2}\right),$$

for $j = 1, 2, 3$. By using a generalization of the proof in [22] we can prove

Theorem 25 *Let ψ_j, $j = 1, 2, 3$ and $\widetilde{\psi}_j$, $j = 1, 2, 3$ be defined above. Let*

$$\begin{aligned}\psi_{j,k,(\ell_1,\ell_2)}(x, y) &= 2^{-k}\psi_j(2^{-k}x - \ell_1, 2^{-k}y - \ell_2)\\ \widetilde{\psi}_{j,k,(\ell_1,\ell_2)}(x, y) &= 2^{-k}\widetilde{\psi}_j(2^{-k}x - \ell_1, 2^{-k}y - \ell_2)\end{aligned}$$

for $(\ell_1, \ell_2) \in \mathbb{Z}^2$, $k \in \mathbb{Z}$, and $j = 1, 2, 3$. Then the $\psi_{j,k,(\ell_1,\ell_2)}$'s and $\tilde{\psi}_{j,k,(\ell_1,\ell_2)}$'s constitute two dual Riesz bases of $L_2(\mathbb{R}^2)$. They satisfy biorthogonal conditions.

$$\int_{\mathbb{R}^2} \psi_{j,k,\ell_1,\ell_2}(x, y)\widetilde{\psi}_{i,m,n_1,n_2}(x, y)dxdy = \delta_{i,j}\delta_{k,m}\delta_{\ell_1,n_1}\delta_{\ell_2,n_2},$$

for all $i, j = 1, 2, 3$, $k, m \in \mathbb{Z}$, and $\ell_1, \ell_2, n_1, n_2 \in \mathbb{Z}$. That is, ψ_j and $\widetilde{\psi}_j$ are biorthogonal wavelets.

Although the above theory is beautiful, examples are difficult to give and hence, are omitted here. Trivariate biorthogonal box spline wavelets were discussed in [37]. The interested reader may refer to [37] for details.

Multivariate Prewavelets

The general construction of prewavelets in the univariate setting (see Sect. "Prewavelets") can easily be generalized to the multivariate setting. We shall outline such a generalization here.

We start the definition of multi-resolution approximation of $L_2(\mathbb{R}^d)$.

Definition 5 A multi-resolution approximation (MRA) of $L_2(\mathbb{R}^d)$ is a sequence of subspaces V_j, $j \in \mathbb{Z}$ of $L_2(\mathbb{R}^d)$ such that

(i) $V_j \subset V_{j+1}$;
(ii) $\bigcup_{j=-\infty}^{\infty} V_j$ is dense in $H^s(\mathbb{R}^d)$;
(iii) $\bigcap_{j=-\infty}^{\infty} V_j = \{0\}$;
(iv) there is a function $\phi \in V_0$ such that the integer translates, $\phi(x - m)$, $m \in \mathbb{Z}^d$ form a Riesz basis for V_0, i. e., there exist two positive numbers A and B such that

$$A\left\|\{c_m, m \in \mathbb{Z}^d\}\right\|_2^2 \le \left\|\sum_{m\in\mathbb{Z}^d} c_m\phi_j(x-m)\right\|_2^2 \le B\left\|\{c_m, m \in \mathbb{Z}^d\}\right\|_2^2,$$

for all square summable sequence $\{c_m, m \in \mathbb{Z}^d\}$.

We shall say ϕ generates a multiresolution approximation of $L_2(\mathbb{R}^d)$ if letting $V_j = \mathrm{span}\{\phi(2^j x - k), k \in \mathbb{Z}^d\}$, V_j is a nested subspace satisfying the above conditions (i)–(iv). Next we need

Definition 6 A collection $\psi_{j,k}$, $k = 2, \ldots, 2^d$ of functions in $L_2(\mathbb{R}^d)$ satisfying the following five properties are called prewavelets:

1. the closure W_k of the linear span of integer translates of ψ_k is orthogonal to the closure V_0 of the linear span of integer translates of ϕ;
2. W_k is orthogonal each other among $k = 2, \ldots, 2^d$,
3. V_1 is the direct sum V_0 and W_k, $k = 2, \ldots, 2^d$;
4. the integer translates of ψ_k form a Riesz basis for W_k; That is, there exist two positive constants A and B such that

$$A\sum_{k=2}^{2^d}\sum_{m\in\mathbb{Z}^d}|c_{k,m}|^2 \le \left\|\sum_{k=2}^{2^d}\sum_{m\in\mathbb{Z}^d} c_{k,m}2^{jd/2}\psi_k(2^j\cdot - m)\right\|_2^2 \le B\sum_{k=2}^{2^d}\sum_{m\in\mathbb{Z}^d}|c_{k,m}|^2$$

for all square summable sequence $\{c_{k,m}, k = 2, \ldots, 2^d, m \in \mathbb{Z}^d\}$.

In this section, we assume that there exists a compactly supported function ϕ which generates an MRA of $L_2(\mathbb{R}^d)$. Let P be the mask polynomial defined by

$$\widehat{\phi}(2\omega) = P(z)\widehat{\phi}(\omega) ,$$

where $\widehat{\phi}$ denotes the Fourier transform of ϕ and $z = \exp(i\omega)$. We are looking for compactly supported functions $\psi_k, k = 2, \ldots, 2^d$ in V_1 such that

$$V_1 = V_0 \bigoplus \bigoplus_{k=2}^{2^d} W_k ,$$

and $\phi(\cdot - m), \psi_k(\cdot - m), m \in \mathbb{Z}^d, k = 2, \ldots, 2^d$ form a stable basis for V_1, where W_k is the closure of the linear span of integer translates of $\psi_k(x - m), m \in \mathbb{Z}^d$.

To do so, we first introduce a Laurent polynomial

$$\Phi(z) := \sum_{m\in\mathbb{Z}^d} \left\langle 2^{d/2}\phi(2x), 2^{d/2}\phi_j(2x - m)\right\rangle z^m .$$

This function Φ is called the generalized Euler–Frobenius polynomial.

Next we need a necessary and sufficient condition for the orthogonality. Writing

$$g_k(2x) = \sum_{m\in\mathbb{Z}^d} c_{k,m} 2^{d/2}\phi(2x - m) \in V_1 ,$$

and

$$G_k(z) = \frac{1}{2^{d/2}} \sum_{m\in\mathbb{Z}^d} c_{k,m} z^m ,$$

the Fourier transforms of g_k and ϕ are related by

$$\widehat{g_k}(2\omega) = G_k(z)\widehat{\phi}(\omega) .$$

Let

$$\mathcal{G}_k = \text{closure}_{L_2(\mathbb{R}^d)}\{g_k(x - m), \quad m \in \mathbb{Z}^d\}$$

be the closure of the linear span of integer translates of g_k.

Recall a special operator E which maps any Laurent polynomial f into such a Laurent polynomial $E(f)$ which contains all the even index terms of f. For example, when $d = 2$ and $z = (z_1, z_2)$,

$$E(f(z)) = \frac{1}{4}(f(z_1, z_2) + f(-z_1, z_2) + f(z_1, -z_2) + f(-z_1, -z_2)) .$$

One important property is

$$E(P(z)\overline{P(z)}\Phi(z)) = 2^{-d}\Phi(z^2)$$

(see a proof in [19].)

We have the following generalization of Theorem 8.

Theorem 26 *$\mathcal{G}_k$ is orthogonal to $\mathcal{G}_{k'}$ for $k' \neq k$ if and only if*

$$E(G_k(z)\overline{G_{k'}(z)}\Phi(z)) = 0 . \tag{82}$$

We divide the construction of compactly supported prewavelets into two steps. The first step is to construct compactly supported $g_k \in V_1, k = 1, \ldots, 2^d$ such that the closure $\mathcal{G}_k$ of the linear span of integer translates g_k is orthogonal to V_0 for each k. The second step is to use a technique like Gram–Schmidt orthonormal procedure to orthogonalize these g_k's for different k.

To be more precise, we let $\{n_1, \ldots, n_{2^d}\} = \{0, 1\}^d$ with $n_k \in \mathbb{Z}^{2^d}$ and $g_k \in V_1$ satisfy

$$g_k(x - m) \perp V_0, \quad m \in \mathbb{Z}^d$$

and

$$\begin{aligned} &2^{d/2}\phi(2x - n_k) \\ &= \sum_{m\in\mathbb{Z}^d} \left(a_{k,m} 2^{d/2}\phi(x - m) + b_{k,m} 2^{d/2} g_k(2^j x - m)\right) \end{aligned}$$

for each $k \in \{1, \ldots, 2^d\}$. That is, we want to have $\mathcal{G}_k$ is orthogonal to V_0 and $V_1 = V_0 \bigoplus (\mathcal{G}_1 + \ldots + \mathcal{G}_{2^d})$. In terms of Fourier transform, the above equations can be rewritten as

$$\begin{aligned} \frac{1}{2^{d/2}} e^{in_k\omega/2}\widehat{\phi}\left(\frac{\omega}{2}\right) &= A_k(\omega)\widehat{\phi}(\omega) + B_k(\omega)\widehat{g_k}(\omega) \\ &= A_k(\omega)P(\omega/2)\widehat{\phi}(\omega/2) + B_k(\omega)G_k(\omega/2)\widehat{\phi}(\omega/2) , \end{aligned}$$

where $A_k(\omega) = \sum_{m\in\mathbb{Z}^d} 2^{d/2} a_{k,m} e^{im\omega}$ and $B_k(\omega) = \sum_{m\in\mathbb{Z}^d} 2^{d/2} b_{k,m} e^{im\omega}$. Here, we have abused the notation of $P(z)$ $G_k(z)$, that is, we use $P(\omega)$ instead of $P(z)$ and $G_k(\omega)$ instead of $G_k(z)$ with $z = e^{i\omega}$ just for convenience. It follows that

$$A_k(2\omega)P(z) + B_k(2\omega)G_k(z) = \frac{e^{in_k\omega}}{2^{d/2}} , \tag{83}$$

for $k = 1, \ldots, 2^d$. Using Theorem 26, the solution of A_k, B_k and G_k can be easily found as shown in the following.

Lemma 25 *Suppose that $\Phi(z^2) = 2^d E(P(z)\overline{P(z)}\Phi(z)) \neq 0$ for all z in torus T^d. Let*

$$\begin{aligned} A_k(2\omega) :=& \frac{E(e^{in_k\omega}\overline{P(z)}\Phi(z))}{E(P(z)\overline{P(z)}\Phi(z))}, \\ B_k(2\omega) :=& \frac{1}{E(P(\omega)\overline{P(z)}\Phi(z))}, \\ G_k(z) :=& \frac{1}{2^{d/2}} E(P(z)\overline{P(z)}\Phi(z))e^{in_k\omega} \\ & - \frac{1}{2^{d/2}} E(e^{in_k\omega}\overline{P(z)}\Phi(z))P(z) . \end{aligned}$$

Then $\mathcal{G}_k$ is orthogonal to V_0 for all $k = 1, \ldots, 2^d$ and

$$V_1 = V_0 \bigoplus (\mathcal{G}_1 + \ldots + \mathcal{G}_{2^d}) .$$

Next we can show that integer translates of $g_k, k = 1, \ldots, 2^d$ are linearly dependent. However, under a certain condition on P, integer translates of $g_k, k = 2, \ldots, 2^d$ are linearly independent. Let us write P in its polyphase form, i. e.,

$$P(z) = \sum_{k=1}^{2^d} \mathrm{e}^{in_k\omega} P_k(z^2) , \tag{84}$$

where $n_k, k = 1, \ldots, 2^d$ are the multi-integers in the collection $\{0, 1\}^d$ as defined above. We refer to [49] for a proof of the following theorem.

Theorem 27 *Suppose that $\Phi(z^2) \neq 0$. Suppose that there exists an integer $k \in \{0, 1\}^d$ such that $P_k(z^2) \neq 0$ for all z on the torus T^d. For simplicity, let us assume that $P_1(z^2) \neq 0$ for all z on T^d. Then the integer translates of $g_k, k = 2, \ldots, 2^d$ form a Riesz basis for $V_1 \ominus V_0$.*

The second step is to use a technique like the well-known Gram–Schmidt orthonormalization to construct $\psi_{j,k}$ from $g_{j,k}$ such that $\psi_{j,k}$ are orthogonal among each other. It is a standard technique (cf. e. g., [39]).

We first choose $\psi_2 = g_2$. Let

$$\psi_3(x) = \sum_{m\in\mathbb{Z}^d} (c_{1,m}\psi_2(x-m) + c_{2,m}g_3(x-m))$$

for some coefficients $c_{1,m}$ and $c_{2,m}$. To compute these coefficients, we write them in terms of Fourier transform

$$\begin{aligned} \widehat{\psi_3}(2\omega) =& C_1(z^2)\widehat{\psi_2}(2\omega) + C_2(z^2)\widehat{g_3}(2\omega) \\ =& (C_1(z^2)G_2(z) + C_2(z^2)G_3(z))\widehat{\phi}(\omega/2) , \end{aligned}$$

where C_1 and C_2 are discrete Fourier transform of sequences $c_{1,m}$'s and $c_{2,m}$'s. For convenience, we let $Q_2(z) = G_2(z)$ and

$$Q_3(z) = C_1(z^2)G_2(z) + C_2(z^2)G_3(z) .$$

In order to have $W_3 \perp W_2$, the orthogonal condition in Theorem 26 implies that

$$\begin{aligned} C_1(z^2)E(G_2(z)\overline{G_2(z)}\Phi(z) \\ + C_2(z^2)E(G_3(z)\overline{G_2(z)}\Phi(z)) = 0 . \end{aligned} \tag{85}$$

By choosing

$$\begin{aligned} C_1(z^2) =& E(G_3(z)\overline{G_2(z)}\Phi(z)), \\ C_2(z^2) =& - E(G_2(z)\overline{G_2(z)}\Phi(z)) , \end{aligned}$$

we know that the the Eq. (85) holds and W_3 is perpendicular to W_2. We continue this procedure above. To be more precise, let us show how to construct ψ_4. That is, let

$$\begin{aligned} \psi_4(x) = \sum_{m\in\mathbb{Z}^d} & (d_{1,m}\psi_2(x-m) + d_{2,m}\psi_3(x-m)) \\ & + d_{3,m}g_4(x-m)) . \end{aligned}$$

In terms of Fourier transform, we have

$$\begin{aligned} \widehat{\psi_4}(2\omega) =& D_1(z^2)\widehat{\psi_2}(\omega) + D_2(z^2)\widehat{\psi_3}(\omega) + D_3(z^2)\widehat{g_4}(\omega) \\ =& (D_1(z^2)Q_2(z) + D_2(z^2)Q_3(z) \\ & + D_3(z^2)G_4(z)))\widehat{\phi}(\omega/2) . \end{aligned}$$

In order to have $W_4 \perp W_2$ and $W_4 \perp W_3$, we have the following two equations with three unknowns:

$$\begin{aligned} & D_1(z^2)E(Q_2(z)\overline{Q_2(z)}\Phi(z)) \\ & \quad + D_3(z^2)E(G_4(z)\overline{Q_2(z)}\Phi(z)) = 0 \\ & D_2(z^2)E(Q_3(z)\overline{Q_3(z)}\Phi(z)) \\ & \quad + D_3(z^2)E(G_4(z)\overline{Q_3(z)}\Phi(z)) = 0 \end{aligned} \tag{86}$$

which is an upper triangular homogeneous linear system. It can be solved easily. A solution may be given below. Let

$$\begin{aligned} D_1(z^2) =& E(Q_3(z)\overline{Q_3(z)}\Phi(z))E(G_4(z)\overline{Q_2(z)}\Phi(z)) \\ D_2(z^2) =& E(Q_2(z)\overline{Q_2(z)}\Phi(z))E(G_4(z)\overline{Q_3(z)}\Phi(z)) \\ D_3(z^2) =& - E(Q_2(z)\overline{Q_2(z)}\Phi(z))E(Q_3(z)\overline{Q_3(z)}\Phi(z)) . \end{aligned}$$

With these Laurent polynomials D_1, D_2, D_3, the two equations in (86) are satisfied simultaneously. Thus we obtain the desired function ψ_4. Repeating the above constructive steps when $d > 2$, we find $\psi_k, k = 2, \ldots, 2^d$. It is easy to see that ψ_k's are compactly support when ϕ are compactly supported. The above construction shows that the integer translates of ψ_k form a Riesz basis for W_k for $k = 2, \ldots, 2^d$ and W_k's are mutually orthogonal. We have thus obtained the following

Theorem 28 *If a refinable function ϕ generates an MRA for $L_2(\mathbb{R}^d)$. If $\Phi(z) \neq 0$ for all $z = e^{i\omega}$ and one of the polyphases of the mask $P(\omega)$ of ϕ is not zero for all $e^{i\omega}$. Then the functions $\psi_{j,k}$ constructed above are prewavelets for $L_2(\mathbb{R}^d)$ satisfying the conditions 1.–4. in Definition 6.*

Next we show how to use box splines to construct prewavelets in $L_2(\mathbb{R}^d)$ since multivariate box splines are a very important class of refinable functions. Let us recall the definition of box splines. Let D be a set of nonzero vectors in $\mathbb{R}^d$ (counting multiple of a same vector) which span $\mathbb{R}^d$. The box spline ϕ_D associates with the direction set D is the function whose Fourier transform is defined by

$$\hat{\phi}_D(\omega) = \prod_{y \in D} \frac{1 - \mathrm{e}^{-i\xi\cdot\omega}}{iy \cdot \omega} .$$

It is well-known that box spline ϕ_D is a piecewise polynomial function of degree $\leq \#D - d$, where $\#D$ denotes the cardinality of D. For more properties of box splines, see [7,11,60]. In particular, for $d = 2$, $e_1 = (1,0)^{\mathrm{T}}$, $e_2 = (0,1)^{\mathrm{T}}$, and

$$D = \{\underbrace{e_1, \ldots, e_1}_{\ell}, \underbrace{e_2, \ldots, e_2}_{m}, \underbrace{e_1 + e_2, \ldots, e_1 + e_2}_{n}\} ,$$

the box spline $\phi_{\ell mn}$ based on such direction set D is called 3-direction box spline whose Fourier transform is

$$\hat{\phi}_{\ell mn}(\omega_1, \omega_2) = \left(\frac{1 - \mathrm{e}^{-i\omega_1}}{i\omega_1}\right)^{\ell} \left(\frac{1 - \mathrm{e}^{-i\omega_2}}{i\omega_2}\right)^{m} \left(\frac{1 - \mathrm{e}^{-i(\omega_1+\omega_2)}}{i(\omega_1 + \omega_2)}\right)^{n} .$$

It is well-known that box spline ϕ_D generates a bona fide MRA of $L_2(\mathbb{R}^d)$ (cf. [74]) when the direction set D is unimodular, i. e., the determinant of any d directions which span $\mathbb{R}^d$ is 1 or -1 (cf. [7]). The unimodularity also implies $\Phi(\omega) > 0$. Let Φ_D be the Euler–Frobenius polynomial associated with ϕ_D and P_D be the mask associated with ϕ_D, i. e., $\widehat{\phi_D}(2\omega) = P_D(z)\widehat{\phi_D}(\omega)$.

Theorem 29 *Consider the linear box spline in $\mathbb{R}^d$. That is, let*

$$D = \{e_1, \ldots, e_d, -(e_1 + \ldots + e_d)\} ,$$

where e_i denotes the standard unit vector in $\mathbb{R}^d$ which is 1 in the ith component while zero in the rest of the components for $i = 1, \ldots, d$. Then $\Phi_D(z) \neq 0$ for all z with $|z| = 1$ and $E(P_D(z)) = \frac{1}{2^d}$.

Proof Since the D is unimodular, we have $\Phi_D(z) \neq 0$ for all z with $|z| = 1$. Next it is easy to see that $P_D(z) = \prod_{i=1}^{d} \left(\frac{1+z_i}{2}\right)\left(\frac{1+1/(z_1 \ldots z_d}{2}\right)$. Then we can see that the even index term $E(P_D(z))$ is only the constant term which is $2/2^{d+1} = 1/2^d$. This completes the proof. □

Example 26 Consider $\phi_{2,2,1}$. Since $P_{2,2,1}(z) = (1 + z_1)^2(1 + z_2)^2(1 + z_1 z_2)/32$, it is easy to check that

$$E(P_{2,2,1}(z)) = \frac{1}{32}\left(5z_1^2 z_2^2 + z_1^2 + z_2^2 + 1\right) .$$

Since

$$\begin{aligned} 32|E(P_{2,2,1}(z)| &= |5 + (z_1 z_2)^{-2} + (z_1)^{-2} + (z_2)^{-2}| \\ &> 5 - |(z_1 z_2)^{-2}| - |(z_1)^{-2}| - |(z_2)^{-2}| = 2 , \end{aligned}$$

we know that $E(P_{2,2,1}(z) \neq 0$ for $z = (z_1, z_2)$ with $|z_1| = |z_2| = 1$.

Example 27 Consider $\phi_{2,2,2}$. Similar to the examples above, we have

$$\begin{aligned} E(P_{2,2,2}(z)) = &\frac{1}{64x^2y^2}(10x^2y^2 + x^2 + y^2 + 1 + y^4x^2 \\ &+ x^4y^4 + x^4y^2) . \end{aligned}$$

We can easily see that $E(P_{2,2,2}(z)) \neq 0$. □

Example 28 We use box spline $\tilde{B}_{1,1,1} = \phi_D$ based on $D = \{\mathrm{e}^1, \mathrm{e}^2, -(\mathrm{e}^1 + \mathrm{e}^2)\}$ to construct compactly supported pre-wavelets in $L_2(\mathbb{R}^2)$. Note that our prewavelets have a larger support than those constructed in [44] and [39]. The purpose of this example is to show the detail of our constructive procedure. Clearly,

$$P(z) = \frac{1 + z_1}{2}\frac{1 + z_2}{2}\frac{1 + 1/(z_1 z_2)}{2}$$

and $\Phi_D(z) = \frac{1}{2} + \frac{1}{12}(z_1 + z_2 + 1/z_1 + 1/z_2 + z_1 z_2 + 1/(z_1 z_2)) = \frac{1}{2} + \frac{1}{6}(\cos(\omega_1) + \cos(\omega_2) + \cos(\omega_1 + \omega_2)) \neq 0$ for any ω_1 and ω_2. Using a computer algebra program Maple, we obtain the Laurent polynomials for $G_1, \ldots, G_4$ and Q_2, Q_3, Q_4. They are as follows:

$$768G_1(z_1, z_2) = \begin{bmatrix} z_1^{-3} \\ z_1^{-2} \\ z_1^{-1} \\ 1 \\ z_1 \\ z_1^2 \\ z_1^3 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} -1 & -1 & -1 & -1 & 0 & 0 & 0 \\ -1 & 14 & -2 & 14 & -1 & 0 & 0 \\ -1 & -2 & -19 & -19 & -2 & -1 & 0 \\ -1 & 14 & -19 & 60 & -19 & 14 & -1 \\ 0 & -1 & -2 & -19 & -19 & -2 & -1 \\ 0 & 0 & -1 & 14 & -2 & 14 & -1 \\ 0 & 0 & 0 & -1 & -1 & -1 & -1 \end{bmatrix} \begin{bmatrix} z_2^{-3} \\ z_2^{-2} \\ z_2^{-1} \\ 1 \\ z_2 \\ z_2^2 \\ z_2^3 \end{bmatrix} ,$$

$$768G_2(z_1,z_2)=\begin{bmatrix} z_1^{-1}\\ 1\\ z_1\\ z_1^2\\ z_1^3\end{bmatrix}^{\mathrm{T}}\begin{bmatrix} -2 & 14 & -10 & 6 & 0 & 0 & 0\\ -2 & -4 & -12 & -20 & -10 & 0 & 0\\ 0 & 14 & -12 & 76 & -12 & 14 & 0\\ 0 & 0 & -10 & -20 & -12 & -4 & -2\\ 0 & 0 & 0 & 6 & -10 & 14 & -2\end{bmatrix}\begin{bmatrix} z_2^{-3}\\ z_2^{-2}\\ z_2^{-1}\\ 1\\ z_2\\ z_2^2\\ z_2^3\end{bmatrix},$$

$$768G_3(z_1,z_2)=\begin{bmatrix} z_1^{-3}\\ z_1^{-2}\\ z_1^{-1}\\ 1\\ z_1\\ z_1^2\\ z_1^3\end{bmatrix}^{\mathrm{T}}\begin{bmatrix} -2 & -2 & 0 & 0 & 0\\ 14 & -4 & 14 & 0 & 0\\ -10 & -12 & -12 & -10 & 0\\ 6 & -20 & 76 & -20 & 6\\ 0 & -10 & -12 & -12 & -10\\ 0 & 0 & 14 & -4 & 14\\ 0 & 0 & 0 & -2 & -2\end{bmatrix}\begin{bmatrix} z_2^{-1}\\ 1\\ z_2\\ z_2^2\\ z_2^3\end{bmatrix},$$

and

$$768G_4(z_1,z_2)=\begin{bmatrix} z_1^{-1}\\ 1\\ z_1\\ z_1^2\\ z_1^3\end{bmatrix}^{\mathrm{T}}\begin{bmatrix} 6 & -10 & 14 & -2 & 0\\ -10 & -20 & -12 & -4 & -2\\ 14 & -12 & 76 & -12 & 14\\ -2 & -4 & -12 & -20 & -10\\ 0 & -2 & 14 & -10 & 6\end{bmatrix}\begin{bmatrix} z_2^{-1}\\ 1\\ z_2\\ z_2^2\\ z_2^3\end{bmatrix}.$$

Since $Q_2 = G_2$, we now give Q_3 as follows. Let

$$10616832Q_3 = \left[z_1^{-7},\ldots,z_1^{-1},1,z_1,\ldots,z_1^{-7}\right]\mathcal{Q}\left[z_2^{-7},\ldots,z_2^{-1},1,z_2,\ldots,z_2^{-9}\right]^{\mathrm{T}}$$

with matrix $\mathcal{Q}$ being a of size 15×17 defined by $\mathcal{Q} = [\mathcal{Q}_1\mathcal{Q}_2]$ and $\mathcal{Q}_1$ of size 15×9 and $\mathcal{Q}_2$ of size 15×8, where

$$\mathcal{Q}_1=\begin{bmatrix} -1 & 5 & -1 & 9 & 1 & 3 & 1 & -1 & 0\\ 1 & -2 & -26 & -2 & -52 & 2 & -22 & 2 & 3\\ -2 & 7 & 11 & 73 & 71 & 63 & 85 & -17 & 27\\ 2 & -4 & -46 & 24 & -516 & 144 & -628 & 168 & -158\\ -1 & -1 & 33 & 37 & 369 & 271 & 773 & 15 & 480\\ 1 & -2 & -22 & 70 & -672 & 714 & -3110 & 1402 & -2285\\ 0 & -3 & 21 & -45 & 385 & -63 & 1713 & 291 & 2057\\ 0 & 0 & -2 & 44 & -228 & 700 & -2462 & 2712 & -6520\\ 0 & 0 & 0 & -18 & 70 & -174 & 746 & -260 & 2057\\ 0 & 0 & 0 & 0 & -4 & 96 & -340 & 792 & -2285\\ 0 & 0 & 0 & 0 & 0 & -28 & 74 & -150 & 480\\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 52 & -158\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -14 & 27\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\end{bmatrix},$$

and

$$\mathcal{Q}_2=\begin{bmatrix} 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ -14 & 0 & 0 & 0\\ 52 & 2 & 0 & 0\\ -150 & 74 & -28 & 0\\ 792 & -340 & 96 & -4\\ -260 & 746 & -174 & 70\\ 2712 & -2462 & 700 & -228\\ 291 & 1713 & -63 & 385\\ 1402 & -3110 & 714 & -672\\ 15 & 773 & 271 & 369\\ 168 & -628 & 144 & -516\\ 2 & -22 & 2 & -52\\ -1 & 1 & 3 & 1\end{bmatrix}$$

$$\begin{matrix} 0 & 0 & 0 & \\ 0 & 0 & 0 & \\ 0 & 0 & 0 & \\ 0 & 0 & 0 & \\ 0 & 0 & 0 & \\ 0 & 0 & 0 & \\ -18 & 0 & 0 & \\ 44 & -2 & 0 & 0 \\ -45 & 21 & -3 & 0 \\ 70 & -22 & -2 & 1 \\ 37 & 33 & -1 & -1 \\ 24 & -46 & -4 & 2 \\ -2 & -26 & -2 & 1 \\ 9 & -1 & 5 & -1 \end{matrix}\Bigg] .$$

The expression for Q_4 involves a matrix of size about 51×51. Due to the space limit, we omit the details for Q_4. □

Multivariate Tight Wavelet Frames and Bi-Frames

In this section, we generalize one of three constructive methods in Sect. "Tight Wavelet Frames" to give a construction of tight wavelet frames in the multivariate setting. We first introduce some notation.

Given a function $\psi \in L_2(\mathbb{R}^d)$, we let

$$\psi_{j,k}(y) = 2^{jd/2}\psi(2^j y - k), \forall j \in \mathbb{Z}, k \in \mathbb{Z}^d .$$

Let Ψ be a finite subset of $L_2(\mathbb{R}^d)$ and

$$\Lambda(\Psi) := \{\psi_{j,k}; \psi \in \Psi, j \in \mathbb{Z}, k \in \mathbb{Z}^d\}$$

where $\mathbb{Z}$ is the set of all integers.

Definition 7 We say $\Lambda(\Psi)$ is a frame if there exist two positive numbers A and B such that

$$A\|f\|^2_{L_2(\mathbb{R}^d)} \le \sum_{g\in\Lambda(\Psi)} |\langle f, g\rangle|^2 \le B\|f\|^2_{L_2(\mathbb{R}^d)}$$

for all $f \in L_2(\mathbb{R}^d)$. $\Lambda(\Psi)$ is a tight frame if it is a frame with $A = B$. In this case, after a renormalization of the g's in Ψ, we have

$$\sum_{g\in\Lambda(\Psi)} |\langle f, g\rangle|^2 = \|f\|^2_{L_2(\mathbb{R}^d)} \tag{87}$$

for all $f \in L_2(\mathbb{R}^d)$.

By polarization technique (as in Sect. "Tight Wavelet Frames"), the Eq. (87) implies that when $\Lambda(\Psi)$ is a tight frame, any $f \in L_2(\mathbb{R}^d)$ can be represented as

$$f = \sum_{g\in\Lambda(\Psi)} \langle f, g\rangle g .$$

We start with a compactly supported refinable function $\phi \in L_2(\mathbb{R}^d)$ which generates a MRA of $L_2(\mathbb{R}^d)$ under a standard dilation matrix $2I_d$, where I_d is the identity matrix in $\mathbb{R}^d$. Since $\phi \in L_2(\mathbb{R}^d)$ is compactly supported and refinable,

$$\hat{\phi}(\omega) = P(\omega/2)\hat{\phi}(\omega/2)$$

where mask $P(\omega)$ is a trigonometric polynomial. Suppose P satisfies

$$\sum_{j\in\{0,1\}^d\pi} |P(\omega + j)|^2 \le 1 . \tag{88}$$

Note that this condition is necessary, but may not be sufficient in the multivariate setting. We have to assume that there exist Laurent polynomials $\widetilde{P}_\ell$ $\ell = 1, \ldots, N$ such that

$$1 - \sum_{j\in\{0,1\}^d\pi} |P(\omega + j)|^2 = \sum_{k=1}^{N} |\widetilde{P}_k(2\omega)|^2 . \tag{89}$$

Here N is a nonnegative integer which is dependent on P. We shall show that for all multivariate box spline functions the condition (89) will be satisfied.

To construct tight wavelet frames we use the unitary extension principle (UEP) (cf. [76,77]). That is, we look for Q_i (trigonometric polynomial) such that

$$\begin{aligned} P(\omega)\overline{P(\omega+\ell)} + \sum_{i=0}^{r} Q_i(\omega)\overline{Q_i(\omega+\ell)} \\ = \begin{cases} 1 & \text{if} \quad \ell = 0, \\ 0, & \ell \in \{0,1\}^d\pi\backslash\{0\} . \end{cases} \end{aligned} \tag{90}$$

With these Q_i's we can define wavelet frame generators $\psi^{(i)}$, in terms of their Fourier transforms, by

$$\hat{\psi}^{(i)}(\omega) = Q_i(\omega/2)\hat{\phi}(\omega/2), \quad i = 1, \ldots, r . \tag{91}$$

Then, if ϕ is continuous and Lip α, with $\alpha > 0$, and the UEP is satisfied, the family $\Psi = \{\psi^{(i)}, i = 1, \ldots, r\}$ generates a tight frame, i. e., $\Lambda(\Psi)$ is a tight wavelet frame. This result can be proved by using the same proof of Theorem 12 in Sect. "Tight Wavelet Frames". (See [20] and [25] for different proofs.)

For convenience, we rewrite (90) in an equivalent matrix form as follows:

Lemma 26 *Let $\mathcal{P} = (P(\omega + \ell); \ell \in \{0,1\}^d\pi)^{\mathrm{T}}$ be a vector of size $2^d \times 1$ and $\mathcal{Q} = (Q_i(\omega + \ell); \ell \in \{0,1\}^d\pi, i = 1, \ldots, r,)$ be a matrix of size $2^d \times r$. Then (90) is equivalent to*

$$\mathcal{Q}\mathcal{Q}^* = I_{2^d} - \mathcal{P}\mathcal{P}^* , \tag{92}$$

where $\mathcal{P}^*$ denotes the complex conjugate transpose of the column vector $\mathcal{P}$.

Proof This can be verified directly. □

For example, when $d = 2$, $r = 4$ and $\omega = (\xi, \eta)$, we have

$$\mathcal{Q} = \begin{bmatrix} Q_1(\xi,\eta) & Q_1(\xi+\pi,\eta) & Q_1(\xi,\eta+\pi) & Q_1(\xi+\pi,\eta+\pi) \\ Q_2(\xi,\eta) & Q_2(\xi+\pi,\eta) & Q_2(\xi,\eta+\pi) & Q_2(\xi+\pi,\eta+\pi) \\ Q_3(\xi,\eta) & Q_3(\xi+\pi,\eta) & Q_3(\xi,\eta+\pi) & Q_3(\xi+\pi,\eta+\pi) \\ Q_4(\xi,\eta) & Q_4(\xi+\pi,\eta) & Q_4(\xi,\eta+\pi) & Q_4(\xi+\pi,\eta+\pi) \end{bmatrix}^{\mathrm{T}},$$

$\mathcal{P} = (P(\xi,\eta), P(\xi+\pi,\eta), P(\xi,\eta+\pi), P(\xi+\pi,\eta+\pi))^{\mathrm{T}}$, and

$$\mathcal{Q}\mathcal{Q}^* = I_{2^d} - \mathcal{P}\mathcal{P}^* . \tag{93}$$

Our construction of tight wavelet frames is mainly based on the matrix form (92).

In general, for P satisfying (88) we write P in its polyphase form (cf. (84) with a normalized constant) and let $\widehat{\mathcal{P}} = (P_m(2\omega); m \in \{0,1\}^d)^T = \mathcal{M}^*\mathcal{P}$, where $\mathcal{M}$ is the polyphase matrix

$$\mathcal{M} = 2^{-d/2}(\mathrm{e}^{im\cdot(\omega+\ell)})_{\substack{\ell\in\{0,1\}^d\pi \\ m\in\{0,1\}^d}} \tag{94}$$

which is unitary and $\mathcal{P} = (P(\omega+\ell); \ell \in \{0,1\}^d\pi)^{\mathrm{T}}$. Here we have abused notation by writing $P(\omega)$ in terms of $P(z)$ with $z = \mathrm{e}^{i\omega}$. Then (88) is equivalent to

$$\widehat{\mathcal{P}}^*\widehat{\mathcal{P}} = \sum_{m\in\{0,1\}^d} |\widehat{P}_m(2\omega)|^2 \le 1 . \tag{95}$$

Theorem 30 *Suppose that P satisfies the condition (88). Suppose that there exist Laurent polynomials $\widetilde{P}_1, \ldots, \widetilde{P}_N$ such that*

$$\sum_{m\in\{0,1\}^d} |P_m(\omega)|^2 + \sum_{i=1}^{N} |\widetilde{P}_i(\omega)|^2 = 1 . \tag{96}$$

Then there exist $2^d + N$ compactly supported tight frame generators with wavelet masks Q_m, $m = 1, \ldots, 2^d + N$, such that P, Q_m, $m = 1, \ldots, 2^d + N$, satisfy (90).

Proof We define the combined column vector $\widetilde{\mathcal{P}} = (P_m(2\omega);\ m \in \{0,1\}^d, \widetilde{P}_i(2\omega);\ 1 \le i \le N)^{\mathrm{T}}$ of size $(2^d + N)$ and the matrix

$$\widetilde{\mathcal{Q}} := I_{(2^d+N)} - \widetilde{\mathcal{P}}\widetilde{\mathcal{P}}^* .$$

Note that all entries of $\widetilde{\mathcal{P}}$ and $\widetilde{\mathcal{Q}}$ are π-periodic. Identity (96) implies that $\widetilde{\mathcal{Q}}\widetilde{\mathcal{Q}}^* = \widetilde{\mathcal{Q}}$, and this gives

$$\widetilde{\mathcal{P}}\widetilde{\mathcal{P}}^* + \widetilde{\mathcal{Q}}\widetilde{\mathcal{Q}}^* = I_{(2^d+N)} .$$

Restricting to the first principle $2^d \times 2^d$ blocks in the above matrices, we have

$$\widehat{\mathcal{P}}\widehat{\mathcal{P}}^* + \widehat{\mathcal{Q}}\widehat{\mathcal{Q}}^* = I_{2^d} , \tag{97}$$

where $\widehat{\mathcal{P}} = \mathcal{M}^*\mathcal{P}$ was already defined above and $\widehat{\mathcal{Q}}$ denotes the first $2^d \times (2^d + N)$ block matrix of $\widetilde{\mathcal{Q}}$. By (94), we have $\mathcal{P} = \mathcal{M}\widehat{\mathcal{P}}$, and (97) yields

$$\mathcal{P}\mathcal{P}^* + \mathcal{M}\widehat{\mathcal{Q}}(\mathcal{M}\widehat{\mathcal{Q}})^* = I_{2^d} ,$$

which is (93). Thus we let

$$\mathcal{Q} = \mathcal{M}\widehat{\mathcal{Q}} .$$

Then the first row $[Q_1, \ldots, Q_{2^d+N}]$ of $\mathcal{Q}$ gives the desired trigonometric functions for compactly supported tight wavelet frame generators. The form $\mathcal{Q} = [Q_i(\omega+\ell)]$ is inherited from $\mathcal{M}$, since the entries of $\widehat{\mathcal{Q}}$ are π-periodic. This completes the proof. □

In general, we do not know if (96) holds for any given mask P. The problem is related to Hilbert's 17th problem. For any multivariate nonnegative polynomial, it is not know that if it can be written as a sum of square of finitely many polynomials (cf. [30,80]). However, we have the following

Theorem 31 *Let $P(\omega) = \sum_{k\in\mathbb{Z}^d} c_k z^k$ be a Laurant polynomial, with N nonzero coefficients c_k. Suppose that all c_k are nonnegative. Furthermore, writing*

$$P(\omega) = \sum_{k=1}^{2^d} e^{i\pi n_k} P_k(2\omega)$$

in its polyphase form with $\{n_k, k = 1, \ldots, 2^d\} = \{0,1\}^d$, suppose that $P_k(0) = 1/2^d$, $k = 1, \ldots, 2^d$. Then there exist at most $r = N^2$ polynomials Q_j such that (90) holds.

A proof of this result can be found in [10]. In particular, for a box spline function ϕ, its mask polynomial satisfies the condition in Theorem 31. Thus we can always use box spline function to construct tight wavelet frames.

We shall use the constructive scheme above to find compactly supported tight wavelet frames based on multivariate box splines, in particular, bivariate box splines on three and four directional meshes.

Let us recall the definition of box spline ϕ_D from Sect. "Multivariate Prewavelets", where D is a set of non zero vectors in $\mathbb{R}^d$ (allowing multiples of the same vector)

which span $\mathbb{R}^d$. It is well-known that ϕ_D is refinable and its Fourier transform satisfies $\widehat{\phi}_D(\omega) = P_D(\frac{\omega}{2})\widehat{\phi}_D(\frac{\omega}{2})$ that

$$P_D(\omega) = \prod_{\xi\in D} \frac{1+\mathrm{e}^{-i\xi\cdot\omega}}{2}\,.$$

Thus P_D is a Laurent polynomial and $|P_D(\omega)|^2 = \prod_{\xi\in D}(\cos\frac{\xi\cdot\omega}{2})^2$. Then we have the following result.

Lemma 27 *Suppose that a given direction set $D \subset \mathbb{Z}^d$ contains all of the standard unit vectors e_i of $\mathbb{R}^d$, $i = 1, \ldots, d$. Then P_D satisfies (88).*

Proof Since $|P_D(\omega)|^2 \le \prod_{i=1}^d \cos^2\frac{\omega_i}{2}$, with $\omega = (\omega_1, \ldots, \omega_s)^{\mathrm{T}} \in \mathbb{R}^d$, we have

$$\sum_{\ell\in\{0,1\}^d\pi} |P_D(\omega+\ell)|^2 \le \prod_{i=1}^d \left(\cos^2\frac{\omega_i}{2} + \sin^2\frac{\omega_i}{2}\right) = 1\,.$$

This completes the proof. □

We now give some examples that the mask polynomial P_D associated with bivariate box splines on three and four direction meshes satisfy (89).

Example 29 Consider a three directional box spline $\phi_{1,1,1}$. It is easy to see that

$$\begin{aligned}1 &- \sum_{\ell\in\{0,1\}^2\pi} |P_{1,1,1}(\omega+\ell)|^2\\ &= \frac{3}{8} - \frac{1}{8}\cos(2\omega_1) - \frac{1}{8}\cos(2\omega_2) - \frac{1}{8}\cos(2\omega_1+2\omega_2).\end{aligned}$$

Thus, we let

$$\begin{aligned}\widetilde{P}_1(\omega) &= \frac{\sqrt{6}}{8}(1-\mathrm{e}^{i\omega_1}), \quad \text{and}\\ \widetilde{P}_2(\omega) &= \frac{\sqrt{2}}{8}(2-\mathrm{e}^{i\omega_2}-\mathrm{e}^{i(\omega_1+\omega_2)}).\end{aligned}$$

Clearly, we have

$$\sum_{\ell\in\{0,1\}^2\pi} |P_{1,1,1}(\omega+\ell)|^2 + \sum_{i=1}^2 |\widetilde{P}_i(2\omega)|^2 = 1\,.$$

Thus, one can apply the constructive steps in the proof of Theorem 30 to get 6 tight frame masks Q_i, $i = 1, \ldots, 6$. □

Example 30 Consider box spline $\phi_{2,2,1}$. We find that

$$\begin{aligned}1 &- \sum_{\ell\in\{0,1\}^2\pi} |P_{2,2,1}(\omega+\ell)|^2\\ &= \frac{19}{32} - \frac{7}{32}\cos(2\omega_1) - \frac{7}{32}\cos(2\omega_2)\\ &\quad - \frac{1}{64}\cos(2\omega_1-2\omega_2) - \frac{9}{64}\cos(2\omega_1+2\omega_2)\,.\end{aligned}$$

Let

$$\begin{aligned}\widetilde{P}_1(\omega) &= \frac{\sqrt{21}}{12} - \frac{\sqrt{102}+2\sqrt{21}}{48}\mathrm{e}^{i\omega_1}\\ &\quad + \frac{\sqrt{102}-2\sqrt{21}}{48}\mathrm{e}^{i\omega_2}\\ \widetilde{P}_2(\omega) &= -\frac{\sqrt{42}+2\sqrt{51}}{48} + \frac{\sqrt{42}}{24}\mathrm{e}^{i\omega_2}\\ &\quad - \frac{\sqrt{42}-2\sqrt{51}}{48}\mathrm{e}^{i(\omega_1+\omega_2)}\,.\end{aligned}$$

It is easy to check that

$$\sum_{\ell\in\{0,1\}^2\pi} |P_{2,2,1}(\omega+\ell)|^2 + \sum_{i=1}^2 |\widetilde{P}_i(2\omega)|^2 = 1\,.$$

Hence, the constructive steps in the proof of Theorem 30 yield 6 tight frame masks and thus, 6 tight frame generators. □

Example 31 For box spline $\phi_{1,1,1,1}$, we have

$$\begin{aligned}1 &- \sum_{\ell\in\{0,1\}^2\pi} |P_{1,1,1,1}(\omega+\ell)|^2\\ &= \frac{5}{8} - \frac{1}{8}\left(\mathrm{e}^{i2\omega_1}+\mathrm{e}^{-i2\omega_1}\right) - \frac{1}{8}\left(\mathrm{e}^{2i\omega_2}+\mathrm{e}^{-2i\omega_2}\right)\\ &\quad - \frac{1}{32}\left(\mathrm{e}^{2i(\omega_1+\omega_2)}+\mathrm{e}^{-2i(\omega_1+\omega_2)}\right)\\ &\quad - \frac{1}{32}\left(\mathrm{e}^{2i(\omega_1-\omega_2)}+\mathrm{e}^{-2i(\omega_1-\omega_2)}\right) = \sum_{i=1}^2 |\widetilde{P}_i(2\omega)|^2\,,\end{aligned}$$

where $\widetilde{P}_1(\omega) = \frac{\sqrt{6}}{8}(1-\mathrm{e}^{i(\omega_1-\omega_2)}$, and

$$\widetilde{P}_2(\omega) = -\frac{1}{4} + \frac{\sqrt{6}}{8} + \frac{1}{4}(\mathrm{e}^{i\omega_1}+e^{i\omega_2}) - \frac{2+\sqrt{6}}{8}\mathrm{e}^{i(\omega_1+\omega_2)}.$$

Hence, the constructive steps in the proof of Theorem 30 yield 6 tight frame masks and hence, 6 tight frame generators. □

In the rest of this section, we construct compactly supported bi-frames. For refinable functions ϕ and ϕ^{dual}, let P and P^{dual} denote the symbols of the respective refinement masks. We use the superscript *dual* in order to point to duality of the respective frames. Moreover, let ψ_j and ψ_j^{dual} be functions associated with ϕ and ϕ^{dual} defined by

$$\begin{aligned}\hat{\psi}_j(\omega) &= Q_j(\omega/2)\hat{\phi}(\omega/2) \quad \text{and}\\ \widehat{\psi_j^{\mathrm{dual}}}(\omega) &= Q_j^{\mathrm{dual}}(\omega/2)\widehat{\phi^{\mathrm{dual}}}(\omega/2)\,,\end{aligned} \tag{98}$$

where Q_j, Q_j^{dual}, $j = 1, \ldots, r$, are two families of masks. Let $\Psi = \{\psi_j, j = 1, \ldots, r$ and $\Psi^{\mathrm{dual}} = \{\psi_j^{\mathrm{dual}}, j = 1,$

$\ldots, r\}$ be the corresponding families of framelets and their shifts and dilates are collected into the following two sets:

$$\begin{aligned}\Lambda(\Psi) &:= \{2^{jd/2}\psi_i(2^j x - k)\,;\ \ j \in \mathbb{Z}\,, k \in \mathbb{Z}^d\,,\\ &\qquad i = 1,\ \ldots, r\},\\ \Lambda(\Psi^{\text{dual}}) &:= \{2^{jd/2}\psi_i^{\text{dual}}(2^j x - k)\,;\ \ j \in \mathbb{Z},\ k \in \mathbb{Z}^d\,,\\ &\qquad i = 1,\ \ldots, r\}\,.\end{aligned} \tag{99}$$

Recall that a family of functions $\{\phi_j, j \in J\}$ is a Bessel family if

$$\left\| \sum_{j\in J} c_j \phi_j \right\|_2^2 \le B \sum_{j\in J} |c_j|^2$$

for any coefficients c_j (see [17]).

Definition 8 The two families $\Lambda(\Psi)$ and $\Lambda(\Psi^{\text{dual}})$ are called bi-frames if they are Bessel families, and the duality relation holds for all $f, g \in L_2(\mathbb{R}^d)$

$$\langle f, g\rangle = \sum_{i=1}^{n} \sum_{\substack{j\in\mathbb{Z}\\ k\in\mathbb{Z}^d}} \langle f, \psi_{i;j,k}\rangle \left\langle \psi_{i;j,k}^{\text{dual}}, g\right\rangle\,. \tag{100}$$

The functions ψ_i and ψ_i^{dual} are called bi-framelets or bi-frame generators.

It follows that

$$f = \sum_{i=1}^{n} \sum_{\substack{j\in\mathbb{Z}\\ k\in\mathbb{Z}^d}} \langle f, \psi_{i;j,k}\rangle \psi_{i;j,k}^{\text{dual}} \tag{101}$$

and

$$f = \sum_{i=1}^{n} \sum_{\substack{j\in\mathbb{Z}\\ k\in\mathbb{Z}^d}} \left\langle \psi_{i;j,k}^{\text{dual}}, f\right\rangle \psi_{i;j,k} \tag{102}$$

weakly for all $f \in L_2(\mathbb{R}^2)$.

To construct the bi-frames, we need the following theorem (cf. Proposition 5.2 in [25]).

Theorem 32 *Suppose that ϕ and ϕ^{dual} are compactly supported refinable functions. Suppose that there are Q_i, Q_i^{dual}, $i = 1, \ldots, r$, satisfying*

$$P(\omega)\overline{P^{\text{dual}}(\omega+\ell)} + \sum_{i=0}^{r} Q_i(\omega)\overline{Q_i^{\text{dual}}(\omega+\ell)} = \delta_\ell \tag{103}$$

for $\ell \in \{0,1\}^d\pi$. Suppose that $Q_i(\omega)$ and $Q_i^{\text{dual}}(\omega)$ have a zero at $\omega = 0$. Let ψ_i and ψ_i^{dual} be the functions defined by their Fourier transform in (98). Then the two families $\Lambda(\Phi)$ and $\Lambda(\Phi^{\text{dual}})$ are bi-frames.

It is easy to see that the relations in (103) can be recast as

Lemma 28 *Let $\mathcal{P} = (P(\omega+\ell); \ell \in \{0,1\}^d\pi)$ be a vector of size $2^d \times 1$, $\mathcal{Q} = (Q_i(\omega+\ell); \ell \in \{0,1\}^d\pi, i = 1, \ldots, r)$ be a matrix of size $2^d \times r$, and $\mathcal{P}^{\text{dual}}$, $\mathcal{Q}^{\text{dual}}$ be given analogously. Then (103) is equivalent to*

$$\mathcal{Q}\left(\mathcal{Q}^{\text{dual}}\right)^* = I_r - \mathcal{P}\left(\mathcal{P}^{\text{dual}}\right)^*\,. \tag{104}$$

Proof This can be verified directly. □

We will construct compactly supported bi-frames for those masks P and P^{dual} which satisfy

$$\sum_{\ell\in\{0,1\}^d\pi} P(\omega+\ell)\overline{P^{\text{dual}}(\omega+\ell)} = 1 \tag{105}$$

and $P(0) = 1 = P^{\text{dual}}(0)$. Let $\mathcal{P}$ and $\mathcal{P}^{\text{dual}}$ be given as in Lemma 28. Recall the unitary matrix $\mathcal{M}$ defined as before. Then we have

Theorem 33 *Define*

$$\mathcal{Q} := (Q_i(\omega+\ell))_{\substack{\ell\in\{0,1\}^d\pi\\ i=1,\ldots,2^d}} = (I_{2^d\times 2^d} - \mathcal{P}(\mathcal{P}^{\text{dual}})^*)\mathcal{M}$$

and

$$\mathcal{Q}^{\text{dual}} := (Q_i^{\text{dual}}(\omega+\ell))_{\substack{\ell\in\{0,1\}^d\pi\\ i=1,\ldots,2^d}} = (I_{2^d\times 2^d} - \mathcal{P}^{\text{dual}}\mathcal{P}^*)\mathcal{M}.$$

Then $\mathcal{P}, \mathcal{P}^{\text{dual}}, \mathcal{Q}$, and $\mathcal{Q}^{\text{dual}}$ satisfy (104), Let ψ_i and ψ_i^{dual} be defined by (98), with these Q_i's and Q_i^{dual}'s. Then $\{\psi_i, i = 1, \ldots, 2^d\}$ and $\{\psi_i^{\text{dual}}, i = 1, \ldots, 2^d\}$ are bi-framelets.

Proof It is trivial to verify that

$$\mathcal{Q}(\mathcal{Q}^{\text{dual}})^* = I_{2^d\times 2^d} - \mathcal{P}(\mathcal{P}^{\text{dual}})^*$$

which is (104) with $S(\omega) = 1$. Since both $\mathcal{M}$ and $\mathcal{P}$ have the desired form, and since $(\mathcal{P}^{\text{dual}})^*\mathcal{M}$ is a row vector whose entries are π periodic, the matrix $\mathcal{Q}$ has the desired form as well. Analogous statements hold for $\mathcal{Q}^{\text{dual}}$.

Next we need to verify the vanishing moment conditions for Q_i and Q_i^{dual}. Let $(\widehat{P}_m(2\omega); m \in \{0,1\}^d) = \mathcal{M}^*\mathcal{P}$ be the polyphase components of P. Then

$$Q_m^{\text{dual}}(\omega) = 2^{-d/2}e^{im\cdot\omega} - P^{\text{dual}}(\omega)\overline{\widehat{P}_m(2\omega)}$$

and $\widehat{P}_m(0) = 2^{-d/2}$. Note that $P^{\text{dual}}(0) = 1$. Therefore, $Q_m^{\text{dual}}(0) = 0$ for $m \in \{0,1\}^d$. Analogous statements show

that $Q_m(0) = 0$ for $m \in \{0,1\}^d$. Using Theorem 32, we conclude that ψ_i and ψ_i^{dual} defined above, using these Q_m's and Q_m^{dual}'s, are bi-framelets. This completes the proof. □

We have the following example of bi-frame generators based on bivariate box splines.

Example 32 For the mask $P_{\ell,m,n}$, associated with the bivariate box spline $B_{\ell,m,n}$ on the three direction mesh, many dual masks $P_{\ell,m,n}^{\text{dual}}$ were given in Sect. "Biorthogonal Box Spline Wavelets" satisfying (105) with $d = 2$. Then the formulae for $\mathcal{Q}$ and $\mathcal{Q}^{\text{dual}}$ given in Theorem 33 provide an explicit representation of bi-framelets or bi-frame generators. □

Next we consider a general refinable function ϕ. Let P be the mask associated with ϕ. Note that $P(0) = 1$. Assume that $P(\ell) = 0$ for $\ell \in \{0,1\}\pi^d \backslash \{0\}$. To ensure (105) for any given mask P, we may use the celebrated Hilbert Nullstellensatz. Indeed, we let $P_m(2\omega)$ be the polyphase components of P, i. e.,

$$(P_m(2\omega); m \in \{0,1\}^d) = \mathcal{M}^*(P(\omega+\ell); \ell \in \{0,1\}^d) .$$

Similarly, for the dual mask P_{dual}, let $P_m^{\text{dual}}(2\omega)$ be the polyphase components of P_{dual}. Then (105) is equivalent to

$$\sum_{m\in\{0,1\}^d} P_m(\omega)\overline{P_m^{\text{dual}}}(\omega) = 1 .$$

By the Hilbert Nullstellensatz, we have

Lemma 29 *Let P be the mask of a refinable function ϕ. Write $\hat{P}_m(z) := P_m(\omega)$ in terms of $z = e^{i\omega} := (e^{i\omega_1}, \ldots, e^{i\omega_d}) \in \mathbb{C}^d$ for $m \in \{0,1\}^d$. If the Laurent polynomials $\hat{P}_m$ have no common zero in $(\mathbb{C}\backslash\{0\})^d$, then there exist Laurent polynomials $\hat{Q}_m(z)$ such that*

$$\sum_{m\in\{0,1\}^d} \hat{P}_m(z)\hat{Q}_m(z) = 1 . \tag{106}$$

Thus, we let $P^{\text{dual}}(\omega) = 2^{-d/2}\sum_{m\in\{0,1\}^d} \mathrm{e}^{im\cdot\omega}\hat{Q}_m(\mathrm{e}^{i2\omega})$. Then P and P^{dual} satisfy (105). In order to apply our Theorem 33, we only need to make sure that $P^{\text{dual}}(0) = 1$. Using the fact $P(0) = 1$ and the assumption $P(\ell) = 0$ for $\ell \in \{0,1\}^d\pi\backslash\{0\}$, we conclude from (105) that $P^{\text{dual}}(0) = 1$. Hence, we obtain the following

Theorem 34 *Given a mask P, suppose that $P(0) = 1$ and $P(\ell) = 0$ for $\ell \in \{0,1\}^d\pi\backslash\{0\}$. Let $P_m(\omega)$, $m \in \{0,1\}^d$, be the polyphase components of P. Writing $\hat{P}_m(z) := P_m(\omega)$ in terms of $z = e^{i\omega}$, suppose that the Laurent polynomials $\hat{P}_m$ have no common zero in $z \in \boldsymbol{C}^d\backslash\{0\}$. Then there exists a pair of bi-frames $\{\psi_i; i = 1,\ldots,2^d\}$ and $\{\psi_i^{\text{dual}}; i = 1,\ldots,2^d\}$ associated with P.*

The tight wavelet frames and bi-frames constructed above have one order of vanishing moment. In order to increase the order of vanishing moment, we have to introduce vanishing moment recovery function S or use *Oblique Extension Principle*. We refer the reader to [20] and [25] for the details.

Spherical Tight Wavelet Frames

Let $S \in \mathbb{R}^3$ be the unit spherical surface and $L_2(S)$ be the space of all square integrable functions over S. For any function $F(x,y,z) \in L_2(S)$ with $|x|^2 + |y|^2 + |z|^2 = 1$, we can use a standard transform $x = \cos(\theta)\cos(\phi)$, $y = \cos(\theta)\sin(\phi)$, and $z = \sin(\theta)$ to convert F int o a function f over $[-\pi/2, \pi/2] \times [0, 2\pi]$ by

$$f(\theta,\phi) = F(x,y,z) .$$

Note that $f(\theta,\phi)$ is not an ordinary function over rectangular domain $[-\pi/2, \pi/2] \times [0, 2\pi]$. When F is continuous at the north and south poles, we have

$$f(\pm\pi/2, \phi) = F(0,0,\pm 1), \quad \forall\phi \in [0, 2\pi] . \tag{107}$$

To have C^1 continuity or continuous tangent plane at the both poles, we have

$$\frac{\partial f}{\partial\theta}(\pm\pi/2,\phi) = -F_1(0,0,\pm 1)\cos(\phi) - F_2(0,0,\pm 1)\sin(\phi), \ \forall\phi \in [0,2\pi] . \tag{108}$$

The conditions (107) and (108) can be found in [26,66,85].

Furthermore, using the standard transform, we have

$$\int_S |F(x,y,z)|^2 \mathrm{d}S = \int_{-\pi/2}^{\pi/2}\int_0^{2\pi} |f(\theta,\phi)|^2\cos(\theta)\,\mathrm{d}\phi\,\mathrm{d}\theta . \tag{109}$$

Thus, we consider a special L_2 space over rectangular domain $[-\pi/2, \pi/2] \times [0, 2\pi]$. Let $L_2^*([-\pi/2, \pi/2]\times[0, 2\pi])$ be the space of all bivariate functions f on $[-\pi/2, \pi/2] \times [0, 2\pi]$ such that the right-hand side of the Eq. (109) is finite.

To build a multiresolution approximation of L_2^* $([-\pi/2, \pi/2] \times [0, 2\pi])$, we may use tensor product of two sets of univariate refinable functions. Let $V_0 = \text{span}\{B_1(\theta), \ldots, B_m(\theta)\}$ and $\widehat{V_0} = \text{span}\{T_1(\phi), \ldots,$

$T_n(\phi)\}$ be two basic spaces. Consider functions in the following form:

$$f(\theta,\phi)=\sum_{i=1}^{m}\sum_{j=1}^{n}c_{ij}B_i(\theta)T_j(\phi)$$

for coefficients c_{ij}. Assume that the basis functions B_i and T_j are locally supported, that is, B_i does not contain a neighborhood $[-\pi/2,-\pi 2/+\delta)$ except for $i=1$ and B_i does not contain $(\pi/2-\delta,\pi/2]$ except for $i=m$, where $\delta>0$ is sufficiently small. In this case, we have

$$f(-\pi/2,\phi)=\sum_{j=1}^{n}c_{1j}B_1(-\pi/2)T_j(\phi)$$

$$f(\pi/2,\phi)=\sum_{j=1}^{n}c_{mj}B_m(\pi/2)T_j(\phi)\,.$$

By (107), we have

$$F(0,0,-1)=\sum_{j=1}^{n}c_{1j}B_1(-\pi/2)T_j(\phi)$$

$$F(0,0,1)=\sum_{j=1}^{n}c_{mj}B_m(\pi/2)T_j(\phi)$$

Similarly, by using (108), we have

$$\begin{aligned}-F_1(0,0,-1)\cos(\phi)-F_2(0,0,-1)\sin(\phi)&\\ =\sum_{j=1}^{n}c_{1j}B_1'(-\pi/2)T_j(\phi)&\\ -F_1(0,0,1)\cos(\phi)-F_2(0,0,1)\sin(\phi)&\\ =\sum_{j=1}^{n}c_{mj}B_m'(\pi/2)T_j(\phi)\,.&\end{aligned}$$

Therefore, the space span$\{T_j, j=1,\ldots,n\}$ has to be able to reproduce 1 and both $\cos(\phi)$ and $\sin(\phi)$. Fortunately, trigonometric splines of even degrees have such a property (based on trigonometric Marsden's identity).

This is why in [66], a tensor of C^1 quadratic B-splines and C^1 trigonometric splines are used to build a multiresolution approximation of $L_2(S)$ and to construct prewavelets. To have a higher smoothness at the both north and south poles, one has to use trigonometric splines of higher order (cf. [53]).

Trigonometric splines were first studied in [82] and they are much like ordinary B-splines. More literature on trigonometric splines can be found in [67,68,84]. Let us give a brief explanation of trigonometric B-splines. Let

$$s(x)=\sin(x),\quad c(x)=\cos(x)\,.$$

Let $\mathcal{T}_d$ be the space of all trigonometric polynominals of degree $\le d$. That is, $\mathcal{T}_d$ is a collection of functions of form

$$f(x)=\begin{cases}\frac{a_0}{2}+\sum_{j=1}^{m}(a_{2j}c(2jx)+b_{2j}s(2jx)),\\ \text{if } d=2m\\ \sum_{j=1}^{m}(a_{2j-1}c((2j-1)x)+b_{2j-1}s((2j-1)x)),\\ \text{if } d=2m-1\,.\end{cases}$$

It is easy to see that $\mathcal{T}_d$ is a linear vector space of dimension $d+1$. It is easy to verify that functions $c(2jx)$, $s(2jx)$, $j=0,\ldots,m$ are orthogonal with respect to L_2 inner product on interval $[0,2\pi]$ when $d=2m$ is even. Similar for the case when $d=2m-1$ is odd. Any function $f\in\mathcal{T}_d$ satisfies $f(x+2\pi)=(-1)^d f(x)$ for all x and d. That is, f is periodic on $[0,2\pi]$ if d is even.

Trigonometric splines of degree d are piecewise trigonometric functions, which each piece belongs to the space $\mathcal{T}_d$ of trigonometric polynomials of degree $\le d$. For simplicity, we consider $[0,2\pi]$. Suppose that $\mathbf{t}=(t_j, j=0,\ldots,n+1)$ is a knot sequence with length at least $n\ge d+2$ such that $0<t_{j+1+d}-t_j<\pi$ for all possible j. Starting with

$$T_{0,j}(x)=\begin{cases}1, & \text{if}\quad t_j\le x<t_{j+1}\\ 0, & \text{otherwise}\,,\end{cases}\tag{110}$$

for $j=0,\ldots,n$, and for $k=1,2,\ldots,d$, we recursively define

$$\begin{aligned}T_{k,j}(x)=&\frac{s(x-t_j)}{s(t_{j+k}-t_j)}T_{k-1,j}(x)\\ &+\frac{s(t_{j+1+k}-x)}{s(t_{j+1+k}-t_{j+1})}T_{k-1,j+1}(x)\end{aligned}\tag{111}$$

for $j=0,\ldots,n-k$, where terms with zero denominators are defined to be zero. $T_{k,j}$ is called the jth trigonometric B-spline of degree k. Repeating (111) we have

Lemma 30 *A trigonometric B-spline $T_{k,j}$ is a piecewise trigonometric polynomial. In fact,*

$$T_{k,j}(x)=\sum_{i=j}^{j+k}P_{i,j,k}(x)T_{0,i}(x)\,,$$

where $P_{i,j,k}(x)$ are trigonometric polynomials $\in\mathcal{T}_k$.

More properties of trigonometric B-splines can be found in the references mentioned above. Especially, we can show that $T_{k,j}$ is refinable. That is, after uniformly refining the knot sequence $\mathbf{t}$, we can define trigonometric B-spline $T_{k,j}^{h/2}$ as above and show that $T_{k,j}$ is a finitely linear combination of $T_{k,j'}^{h/2}$.

We are now ready to present a method to find a tight wavelet frame for $L_2(S)$. We begin with

Definition 9 The family $\{\Psi_k\}_{k\in\mathbb{Z}_+}$ of function in $L_2(S)$ is *a (MRA) tight wavelet frame* for $L^2(S)$ if

$$\|F\|^2 = \sum_{j\in\mathbb{Z}_+} |\langle F, \Psi_j\rangle|^2, \quad \forall f \in L^2(S)\,.$$

Using the polarization technique as in Sect. "Tight Wavelet Frames" we have

$$F(x,y,z) = \sum_{k\in\mathbb{Z}_+} \langle f, \Psi_j\rangle \Psi_j(x,y,z)$$

for all $F \in L_2(S)$.

By (109) it is equivalent to build a tight wavelet frame for $L_2^*([-\pi/2,\pi/2]\times[0,2\pi])$. That is, we need to find a family $\{\widetilde{\Psi}_k\}_{k\in\mathbb{Z}_+}$ of function in $L_2^*([-\pi/2,\pi/2]\times[0,2\pi])$ such that

$$\|f\|_*^2 = \sum_{j\in\mathbb{Z}_+} |\langle F, \widetilde{\Psi}_j\rangle_*|^2\,,$$

where $\|f\|_*$ and $\langle f,g\rangle_*$ are the norm and the inner product associated with the weighted L_2 space $L_2([-\pi/2,\pi/2]\times[0,2\pi])$. The following result tells us how to do.

Theorem 35 *Suppose that $\{\omega_j, j\in J\}$ is a tight frame for $L_2(0,2\pi)$ and $\{\psi_k, k\in K\}$ is a tight frame for $L_2(-\pi/2,\pi/2)$ with respect to a nonnegative weight function $\cos(x)$. Then $\{\omega_j(\theta)\psi_k(\phi), j\in J, k\in K\}$ is a tight frame for $L_2^*([-\pi/2,\pi/2]\times[0,2\pi])$.*

Proof Since $f \in L_2^*([-\pi/2,\pi/2]\times[0,2\pi])$

$$\int_{-\pi/2}^{\pi/2} |f(\theta,\phi)|^2 \cos(\theta)\,\mathrm{d}\theta$$

is essentially bounded by Fubini's theorem. For such a $\phi\in[0,2\pi]$ that

$$\int_{-\pi/2}^{\pi/2} |f(\theta,\phi)|^2 \cos(\theta)\,\mathrm{d}\theta < +\infty\,,$$

we have

$$\int_{-\pi/2}^{\pi/2} |f(\theta,\phi)|^2 \cos(\theta)\,\mathrm{d}\theta = \sum_{k\in K}\left|\int_{-\pi/2}^{\pi/2} f(\theta,\phi)\psi_k(\theta)\cos(\theta)\,\mathrm{d}\theta\right|^2 \tag{112}$$

because that $\{\psi_k, k\in K\}$ is a tight frame. Since

$$\left|\int_{-\pi/2}^{\pi/2} f(\theta,\phi)\psi_k(\theta)\cos(\theta)\,\mathrm{d}\theta\right|^2 \le \int_{-\pi/2}^{\pi/2} |f(\theta,\phi)|^2\cos(\theta)\,\mathrm{d}\theta\,,$$

it follows that $\int_0^{2\pi}\left|\int_{-\pi/2}^{\pi/2} f(\theta,\phi)\psi_k(\theta)\cos(\theta)\mathrm{d}\theta\right|^2 \mathrm{d}y < \infty$. For each $k\in K$, the function

$$\int_0^{2\pi} f(\theta,\phi)\psi_k(\theta)d\theta \in L_2(0,2\pi)\,.$$

The tight frame of $\{\omega_j, j\in J\}$ implies that

$$\int_0^{2\pi}\left|\int_{-\pi/2}^{\pi/2} f(\theta,\phi)\psi_k(\theta)\cos(\theta)\,\mathrm{d}\theta\right|^2 \mathrm{d}\phi = \sum_{j\in J}\left|\int_0^{2\pi}\int_{-\pi/2}^{\pi/2} f(\theta,\phi)\psi_k(\theta)\cos(\theta)\mathrm{d}\theta\omega_j(\phi)\,\mathrm{d}\phi\right|^2.$$

Therefore, integrating (112) from 0 to 2π and using Lebegues dominant convergence theorem, we have

$$\begin{aligned}
&\int_0^{2\pi}\int_{-\pi/2}^{\pi/2} |f(\theta,\phi)|^2\cos(\theta)\,\mathrm{d}\theta\,\mathrm{d}\phi \\
&= \sum_{k\in K}\int_0^{2\pi}\left|\int_0^{\pi} f(\theta,\phi)\psi_k(\theta)\cos(\theta)\,\mathrm{d}\theta\right|^2 \mathrm{d}\phi \\
&= \sum_{k\in K}\sum_{j\in J}\left|\int_0^{2\pi}\int_{-\pi/2}^{\pi/2} f(\theta,\phi)\psi_k(\theta)\cos(\theta)\,\mathrm{d}\theta\omega_j(\phi)\,\mathrm{d}\phi\right|^2
\end{aligned}$$

by (113). Therefore, $\{\omega_j(\theta)\psi_k(\phi), j\in J, k\in K\}$ is a tight frame in $L_2^*([-\pi/2,\pi/2]\times[0,2\pi])$. □

Let $-\pi/2 = t_{1,0} < t_{1,1} < t_{1,2} < \ldots < t_{1,m} = \pi/2$ be a knot sequence with $t_{1,i} = -\pi/2 + ih$ with $h = \pi/m$ and $B_{1,i}$ be B-spline of order $d > 1$ based on knots $t_{1,i},\ldots,t_{1,i+d}$, where $t_{1,i+d} = t_{1,i+d-m}$ when $i+d > m$. Here, due to the equally-spaced knots $t_{1,i}$, $B_i(x)$ is just a shift of scaled version of uniform B-spline N_d. We uniformly refine the knots $t_{1,i}$, $i = 0,\ldots,m$ to have $t_{2,i} = -\pi/2 + i\pi/(2m)$ and $B_{2,i}$ be the B-spline of order d based

on knots $t_{2,i}, \ldots, t_{2,i+d}$. Let $V_1 = \text{span}\{B_{1,0}, \ldots, B_{1,m-1}\}$ and $V_2 = \text{span}\{B_{2,0}, \ldots, B_{2,2m-1}\}$. It is easy to see that V_1 is refinable in the sense that $V_1 \subset V_2$. We repeat adding new knots and use the uniform B-splines on the new knots to form finer and finer spline spaces. Thus, we will have $V_1 \subset V_2 \subset V_3 \subset \ldots \subset L_2([-\pi/2, \pi/2])$. It is known that $\bigcup_{j=1}^{\infty} V_j$ is dense in $L_2([-\pi/2, \pi/2])$ with respect to nonnegative weight function $\cos(\theta)$. Similar to the construction in Sect. "Tight Wavelet Frames over Bounded Domain, we can find tight wavelet frames associated with these B-spline subspaces with respect to the weighted L_2 inner product. For simplicity, let $\{\psi_j, j \in J\}$ stands for the tight wavelet frame.

Similarly, we can find tight wavelet frame $\{\omega_k, k \in K\}$ associated with trigonometric B-splines of even degree $d > 0$. See [53] for detail. Then $\{\psi_j \omega_k, j \in J, k \in K\}$ form a tight wavelet frame by Theorem 35.

Wavelets for Image Processing

In this section we explain how to use wavelets to find the edges of images, remove noises from images, and compress images. Mainly we use the wavelet decomposition and reconstruction as explained in Sect. "Wavelet Decomposition and Reconstruction". That is, we first use a wavelet or a wavelet frame to decompose an image into several subimages. Then we treat these subimages by some methods. Finally we reconstruct the image back using treated subimages based on the same wavelet or wavelet frame. Certainly, different wavelets give us different reconstructed images. One of the purposes of the study is to find the best one from all wavelets constructed so far.

A Wavelet/Wavelet Frame Method for Edge Detection

We shall use tensor product of some orthonormal wavelets and the tight wavelet frame based on box spline B_{2211} to find the edges of images.

The wavelet/wavelet frame method for edge detection can be described as follows. We first use a wavelet or wavelet-frame to decompose an image into many levels of subimages which consist of a low-pass part and several high-pass parts of the image. Then we set the low-pass part to be zero and reconstruct the image back using zero low-pass part and the original high-pass parts. Such reconstructed image contains all the edges of the image. (See also [13].) For box spline wavelet frame, we only do one level of decomposition. For other standard wavelets (Haar, D4, D6, biorthogonal 9/7 wavelets), we do 1, 2, 3 levels of decomposition dependent on the images. For some images, e. g. the finger print image, we must do 3 levels of decomposition while for many other images, one or two levels of decomposition are enough. We choose the best edge representation by visual inspection among three levels of decompositions to present here.

To present the edges clearly, we normalize the reconstructed image into the standard grey level between 0 to 255 and use a threshold to divide the pixel values into two major groups. That is, if a pixel value is bigger than the threshold, it is set to be 1. Otherwise, it is set to be zero.

In the following we present several sets of images for comparison. The top two figures are the original image and the edges based on the box spline B_{2211}. The two figures in the middle row are the edges computed by using the Haar and Daubechies D4 wavelets. The two figures in the last row are based on Daubechies D6 and wavelet 9/7 wavelet. From these figures, it is clear that box spline B_{2211} on four direction mesh does an excellent job to reveal the edges of images. (See more examples in [71].)

A Wavelet/Wavelet Frame Method for Image Denoising

Here we shall use the tight wavelet frames based on box splines $\phi_{111}, \phi_{221}, \phi_{222}, \phi_{1111}, \phi_{2211}$ to remove the noises from images. To compare the effectiveness of image denoising, we also use tensor products of the Haar wavelet, Daubechies D4 and D6 wavelets, and biorthogonal 9/7 wavelet to do the denoising. The main idea to use a wavelet or wavelet frame to denoise is we first decompose an image into one level of subimages which consist of a low-pass part and several high-pass parts of the image, then use the soft-thresholding [27,42] to treat each high-pass subimage by shrinking wavelet coefficients, and finally reconstruct the image by using the original low-pass part and shranked high-pass parts. The reconstructed image is an denoised image.

The soft-thresholding method is to set each pixel value z of an image to be nz according to the following

$$nz = \begin{cases} 0, & \text{if} \quad \text{abs}(z) \le \epsilon \\ \text{sign}(z)(\text{abs}(z) - \epsilon), & \text{if} \quad \text{abs}(z) > \epsilon \end{cases}$$

where ϵ is a thresholding value. An heuristic reason for this wavelet denoising method is that the noise of the image will be in contents of the high-pass subimages after the wavelet decomposition. To remove the noise, we reduce the high frequency contents by ϵ.

To measure the quality of denoised images, we use the peak signal to noise ratio (PSNR) which can be explained as follows. Let $x_{ij}, 1 \le i, j \le 512$ be the pixel values of the original image with pixel values in the range $[0, 255]$. Let y_{ij} be the pixel value of the denoised image. Then we mea-

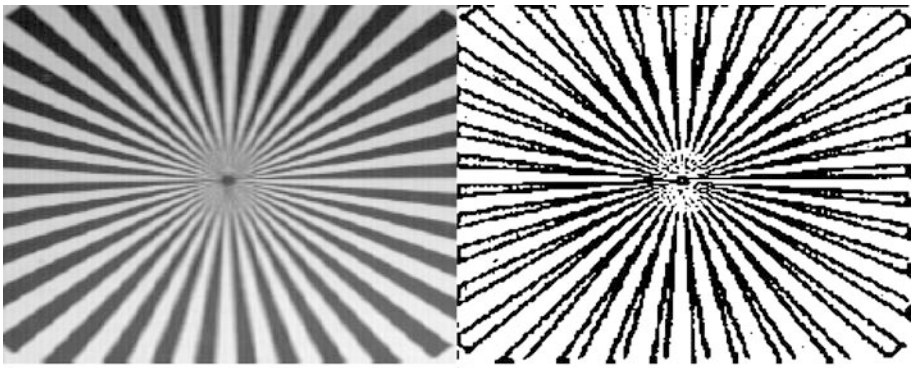

a b

Popular Wavelet Families and Filters and Their Use, Figure 4
a Original image. b Edges by box spline B_{2211}

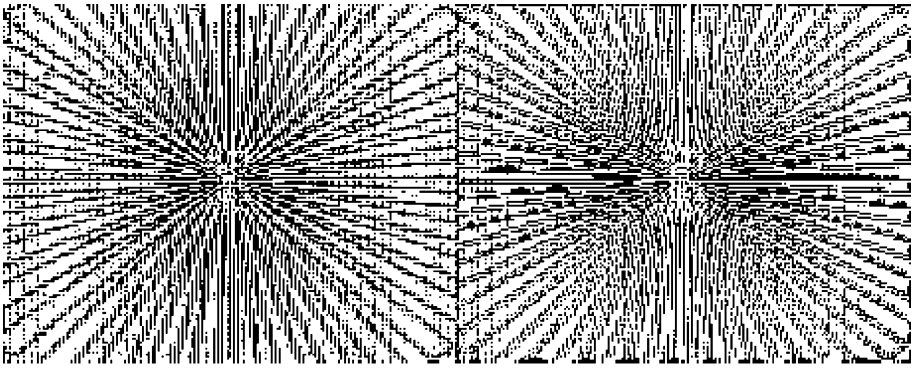

a b

Popular Wavelet Families and Filters and Their Use, Figure 5
a Edges by Haar wavelet. b Edges by Daubechies D4 wavelet

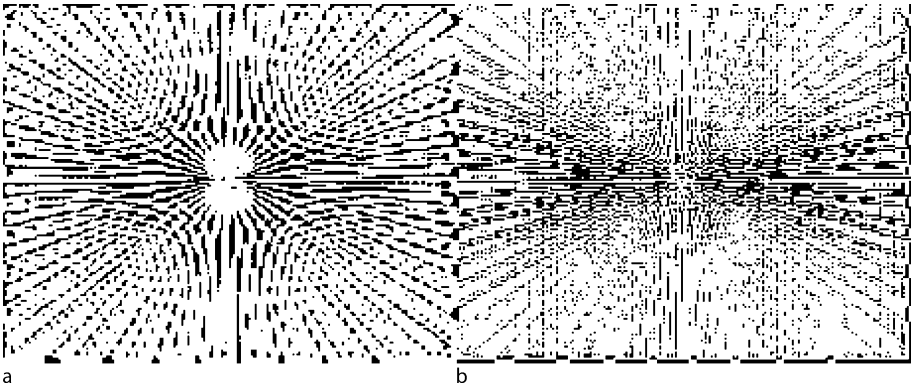

a b

Popular Wavelet Families and Filters and Their Use, Figure 6
a Edges by D6 wavelet. b Edges by CDF 9/7 wavelet

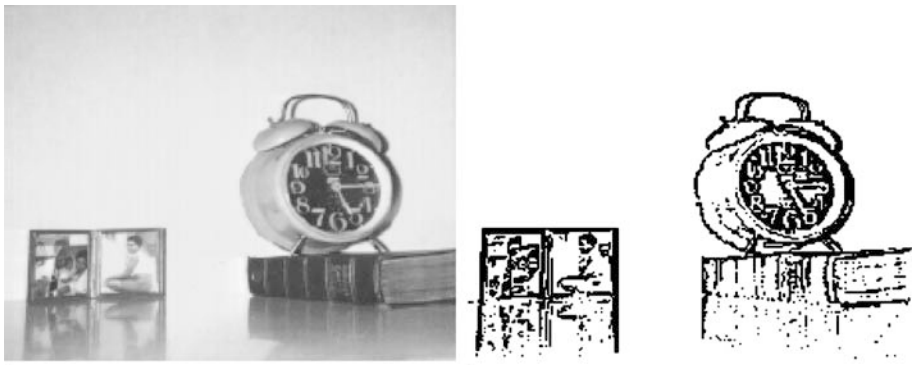

a b

Popular Wavelet Families and Filters and Their Use, Figure 7
a Original image. b Edges by box spline B_{2211}

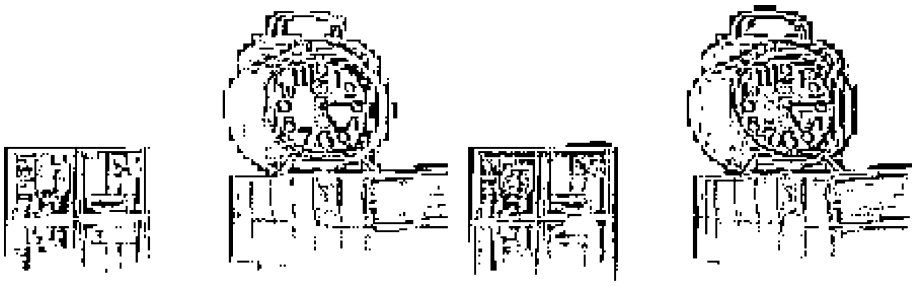

a b

Popular Wavelet Families and Filters and Their Use, Figure 8
a Edges by Haar wavelet. b Edges by D4 wavelet

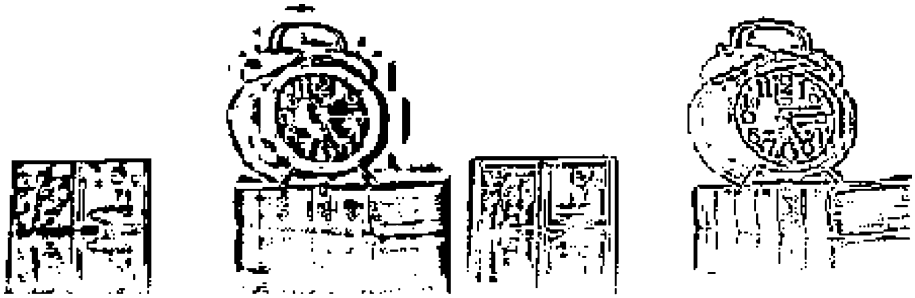

a b

Popular Wavelet Families and Filters and Their Use, Figure 9
a Edges by D6 wavelet. b Edges by CDF 9/7 wavelet

sure the PSNR of the denoised image by

$$\mathrm{PSNR} = -10\log_{10}\frac{\sum_{i,j=1}^{512}(y_{ij}-x_{ij})^2}{255^2\times 512^2}.$$

Certainly, the thresholding value ϵ is dependent on images and noises. We should look for an optimal ϵ such that the PSNR is the best. It is known that the thresholding values are different for high-pass subimages at different levels of decomposition. It is also known that the most of the noise will be in the high frequency contents at the first level of decomposition. one level of the wavelet decomposition.

To test the effectiveness of the wavelet/wavelet frame method, we start with an original image with pixel values x_{ij} and then add a Gaussian noise with zero mean and various variances σ to obtain a noisy image. That is, the pixel values of the noisy image is

$$z_{ij} = x_{ij} + \sigma\delta_{ij}$$

Popular Wavelet Families and Filters and Their Use, Table 2
The PSNR comparison for standard image Bank

Images	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$	$\sigma = 25$
Noised	28.28	24.81	22.36	20.50
Haar wavelet	29.94	26.68	24.38	22.59
D4 wavelet	29.92	26.69	24.39	22.61
D6 wavelet	29.92	26.70	24.40	22.62
Biorth. 9/7	29.93	26.70	24.40	22.62
BSTF111	**31.59**	**28.99**	27.32	26.05
BSTF221	31.52	28.99	**27.40**	**26.24**
BSTF222	31.14	28.55	26.91	25.73
BSTF1111	31.34	28.78	27.20	26.04

with Gaussian noise $\delta_{ij} \sim N(0, 1)$, where σ will be chosen at levels 10, 15, 20 and 25. We choose a popular image Lena to test all the standard wavelets and box spline tight frames. In Table 2, we first list the PSNR numbers for noised image of Lena with different variances $\sigma = 10, 15, 20, 25$ of noises. $BSTF111$ stands for box spline tight frame using box spline ϕ_{111}. Similar for other box spline tight frames. Then we present the noisy images and denoised images to show how well the box spline tight frames can perform.

Wavelets for Image Compression

The scheme to use wavelets for image compression can be described as follows.

(1) We use a wavelet to decomposing the gray-scale values of an images to a maximum number of levels.
(2) Encoding the decomposed image using an embedded zero-tree encoder (cf. [81]) to a specified file size. For images of size 512 × 512, the file size of 262,159 bytes which is approximately one byte per pixel. The actual compressed file sizes are 32,793 bytes (8:1), 16,409 bytes (16:1), 8,217 bytes (32:1), and 4,121 bytes (64:1).
(3) Decoding the compressed file.
(4) Reconstructing the image using the wavelet transform and rounding the values to the nearest integer.
(5) Calculating the peak signal to noise ratio (PSNR) as in the previous section.

The PSNR larger the better. That is, the compressed image can be recovered better. In general, a PSNR ≥ 32 is considered to be visually indistincted from the original and reconstructed images by normal people's eyes at a comfortable viewing distance away from the images. In the following we list a table of PSNR by using tensor product of Haar, Daubechies D4, D6, and a Lai–Roach wavelet (cf. [59]). Test images are standard and given in Fig. 14.

Popular Wavelet Families and Filters and Their Use, Table 3
PSNR of five images by various wavelets

Image: Lena 512 × 512				
Wavelet	8:1	16:1	32:1	64:1
Haar	36.2258	32.6462	29.5685	27.5420
D4	38.4440	34.9209	31.6733	28.8185
D6	38.7819	35.3234	32.0479	29.0727
9/7	39.6450	36.3379	32.9495	29.9049
LR6	38.8167	35.4208	32.1585	29.1588
Image: Barbara 512 × 512				
Wavelet	8:1	16:1	32:1	64:1
Haar	30.4954	26.8119	24.6329	22.7409
D4	32.8675	28.6364	25.6853	23.3821
D6	33.4311	29.0735	25.9431	23.4715
9/7	34.7288	30.0708	26.4626	24.1701
LR6	33.6441	29.2804	26.1074	23.5789
Image: Boat 512 × 512				
Wavelet	8:1	16:1	32:1	64:1
Haar	34.7720	30.7103	27.5762	25.4130
D4	35.6517	31.5910	28.5165	26.0746
D6	35.8593	31.8088	28.6402	26.1880
9/7	37.7118	33.1534	29.7557	27.1292
LR6	35.9301	31.9080	28.7187	26.2733
Image: Finger-print 512 × 512				
Wavelet	8:1	16:1	32:1	64:1
Haar	32.4415	29.9961	28.5828	27.3396
D4	33.3224	30.8632	29.4448	27.9762
D6	33.8077	31.2097	29.8049	28.2562
9/7	34.7043	32.3327	30.3877	28.8153
LR6	33.9236	31.3387	29.9687	28.3928
Image: Crowd 512 × 512				
Wavelet	8:1	16:1	32:1	64:1
Haar	33.3109	29.2032	26.1346	23.7222
D4	34.7214	30.6307	27.4101	24.8170
D6	35.1740	31.0444	27.7601	25.0348
9/7	37.0871	32.4109	28.8652	25.9603

In the above tables, D4 and D6 denote Daubechies wavelets of order 4 and 6, respectively. 9/7 stands for CDF 9/7 biorthogonal wavelet as in Sect. "Biorthogonal Wavelets". As we can see that CDF 9/7 does much better for image compression for all levels. This is why that CDF 9/7 wavelets were implemented in JPEG 2000 standard.

Future Directions

We have discussed many constructive methods of various wavelets in the univariate and multivariate settings. There

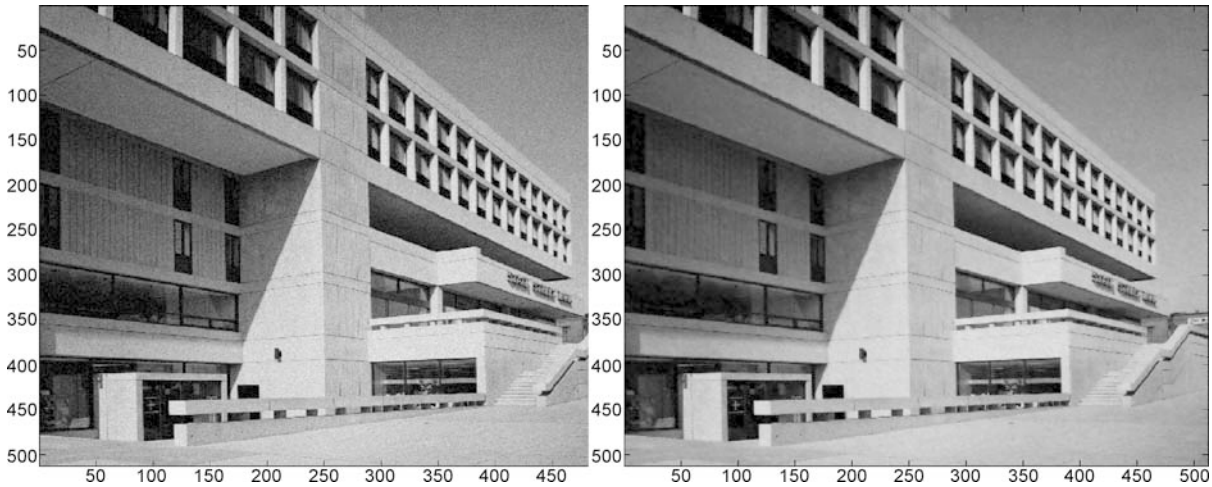

Popular Wavelet Families and Filters and Their Use, Figure 10
The noisy image when $\sigma = 10$ and a denoised image by using BSTF1111

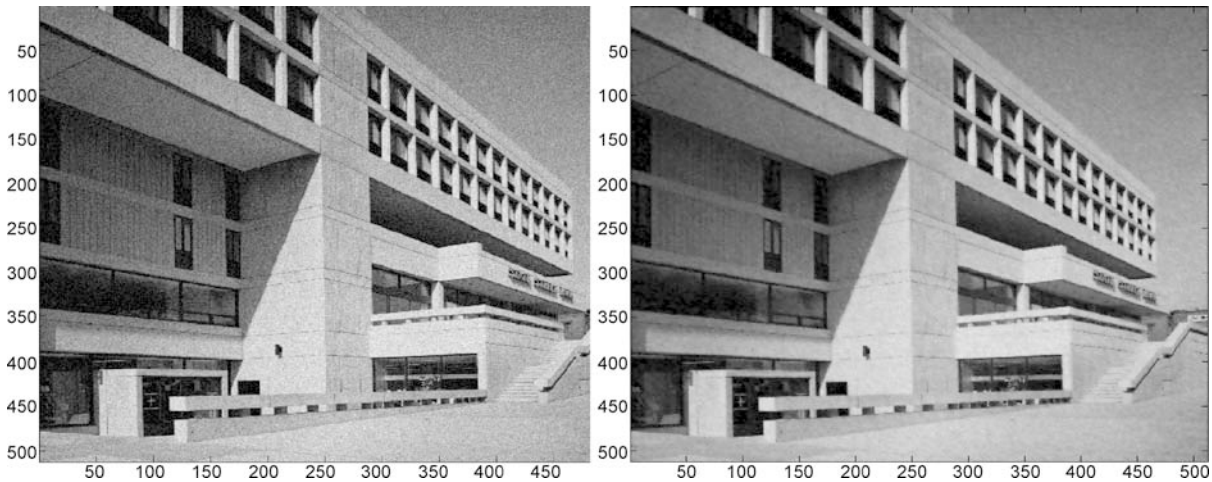

Popular Wavelet Families and Filters and Their Use, Figure 11
The noisy image when $\sigma = 15$ and a denoised image by using BSTF1111

are still many research problems remaining open after about twenty years of development of wavelets since the successful construction of compactly supported orthonormal wavelets with arbitrary regularity in 1988 [23]. We list some of important open problems below.

1. Construction of multivariate nonseparable compactly supported wavelets with high order smoothness. In the multivariate setting, how to find a mask polynomial P such that

$$\sum_{k\in\{0,1\}^d} |P(\omega + k\pi)|^2 = 1$$

and ϕ defined by its Fourier transform

$$\widehat{\phi}(\omega) = \prod_{j=1}^{\infty} P(\omega/2^j) \tag{113}$$

is compactly supported and has any desired smoothness. Belogay–Wang's method in Sect. "Multivariate Orthonormal Wavelets" provides a way to do it in the bivariate setting. How can we do in the multivariate setting? Are there any other methods available?
2. In Sect. "Multiwavelets and Balanced Multiwavelets", we describe the Goodman method for compactly supported orthonormal multi-wavelet using B-splines. One open problem is to generalize the Goodman

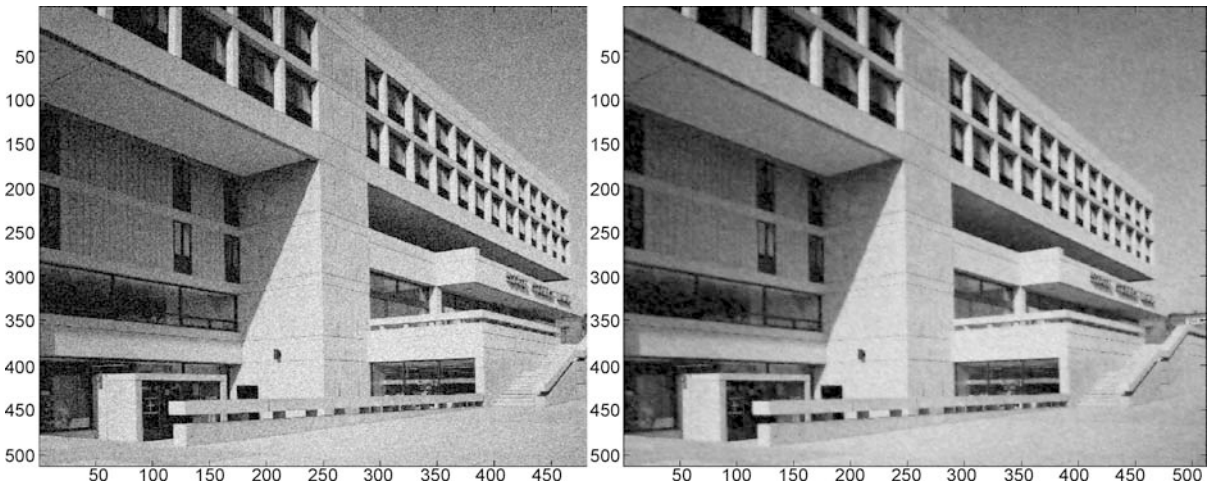

Popular Wavelet Families and Filters and Their Use, Figure 12
The noisy image when $\sigma = 20$ and a denoised image by using BSTF2211

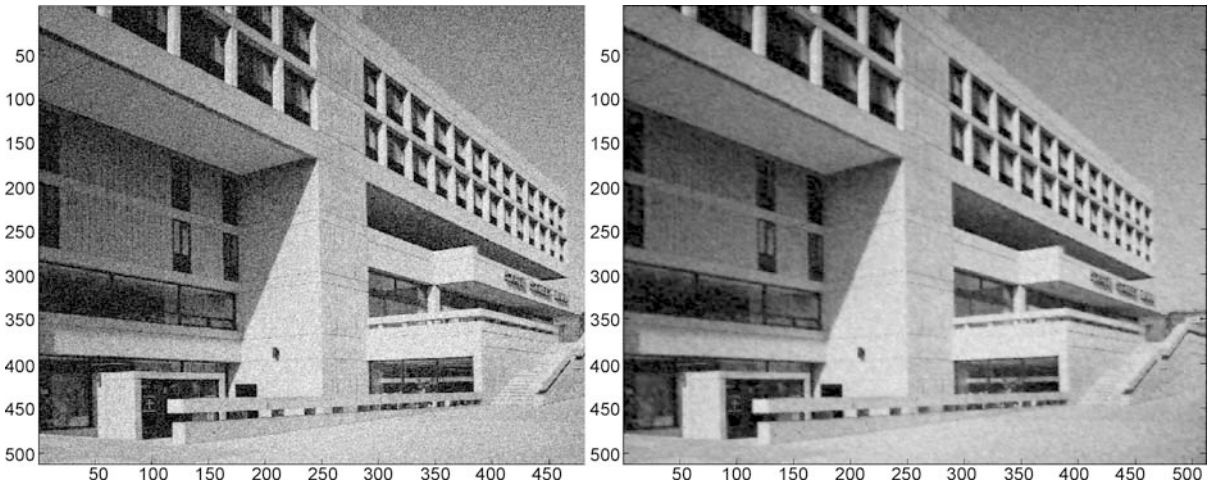

Popular Wavelet Families and Filters and Their Use, Figure 13
The noisy image when $\sigma = 25$ and a denoised image by using BSTF2211

method to the multivariate setting. That is, can we use box splines to construct compactly supported orthonormal multi-wavelets?

3. Even in the univariate setting, when using B-splines to construct compactly supported orthonormal wavelets based on Goodman's method, it is interesting to know if we can use B-splines to construct symmetric compactly supported orthonormal multiwavelets.
4. In the univariate setting, consider using dilation factor $q > 2$ to construct orthonormal wavelets. Suppose we have an orthonormal scaling function. That is, we have the mask polynomial P. It is known that we can use Lawton–Lee–Shen's method [62] to construct the associated wavelets. That is, we can find $Q_k, k = 1, \ldots, q-1$. The open question is to find a formula for Q_k's like the case when $q = 2$.
5. When constructing biorthogonal wavelets using box splines, the dual wavelets by the He–Lai method needs box splines of very high degree. Thus, the support of dual wavelets is very large. An open problem is to find multivariate biorthogonal wavelets of arbitrary smoothness with relatively small support. For example, for lower smoothness, say, C^r for $r = 0, 1, 2$ how can we construct biorthogonal compactly supported wavelets of C^r with the support which is much smaller than the one by using the He–Lai method?
6. Suppose that there is a mask polynomial P satisfying (113). An open problem is if and how we can find

Popular Wavelet Families and Filters and Their Use, Figure 14
Images: a Barbara. b Finger-print. c Boat and d Crowd

$Q_k, k = 1, \ldots, 2^d - 1$ such that

$$\begin{bmatrix} P(\omega + n_1\pi) & P(\omega + n_2\pi) & \ldots & P(\omega + n_{2^d}\pi) \\ Q_1(\omega + n_1\pi) & Q_1(\omega + n_2\pi) & \ldots & P(\omega + n_{2^d}\pi) \\ \vdots & \ddots & \ddots & \vdots \\ Q_{2^d-1}(\omega + n_1\pi) & Q_{2^d-1}(\omega + n_2\pi) & \ldots & Q_{2^d-1}(\omega + n_{2^d}\pi) \end{bmatrix}$$

is unitary, where $\{n_k, k = 1, \ldots, 2^n\} = \{0, 1\}^d \subset \mathbb{Z}_+^d$.

7. Construction of multivariate compactly supported orthonormal wavelets, prewavelets, and tight wavelet frames in Sobolev spaces is a challenge for two decades. So far we still do not have a standard wavelet tool for numerical solution of partial differential equations. One initial step toward to this problem is [52]. Compactly supported prewavelets under the norm in $H_0^1(\Omega)$ with Ω being a rectangular domain or triangular domain were constructed and tested to solve Poisson equation. More work is needed for a wavelet method for biharmonic equations and nonlinear equations.
8. As we have seen from the previous sections, tight wavelet frames are much easier to construct than orthonormal wavelets and they perform much better than orthonormal wavelets for image edge detection and denoising. However, one can not straightforwardly apply the wavelet compression schemes to do image compression based on tight wavelet frames. Mainly the number of wavelet frame generators is more than the number of wavelet functions and hence one needs more bit budget allocated to code the coefficients of wavelet frame representation of an image than to code the coefficients of orthonormal wavelet representation. This is one of major problems that wavelet researchers face today.

Regarding tight wavelet frames, one of problems is to increase the approximation power and order of vanishing moments of wavelet frames. This difficulty may be overcome by using the method of virture components (cf. [55,56]). Construction of tight wavelet frames based on refinable function vectors will be reported in [51].

The wavelet research has already stimulated the interest from both pure and applied mathematicians. The problems proposed above require a deep knowledge of algebra and algebraic geometry (e. g., Problem 6) and an extensive knowledge of applied mathematics (e. g., Problem 8). It is not know if Problem 6 has a solution in the multivariate setting. The Problem 8 may needs some knowledge of optimization and nonlinear sparse approximation. Thus, they are extremely difficult to solve. It may take time to further develop various linear and nonlinear approaches before we can see some hope to answer these questions.

Bibliography

Primary Literature

1. Ayache A (1999) Construction of non-separable dyadic compactly supported orthonormal wavelet bases for $L^2(R^2)$ of arbitrarily high regularity. Rev Mat Iberoamericana 15:37–58
2. Ayache A (2001) Some methods for constructing nonseparable, orthonormal, compactly supported wavelet bases. Appl Comput Harmon Anal 10(1):99–111
3. Bastin F, Boigelot C (1998) Biorthogonal wavelets in $H^m(\mathbb{R})$. J Fourier Anal Appl 4:749–768
4. Bastin F, Laubin P (1997) Regular compactly supported wavelets in Sobolev spaces. Duke Math J 87:481–508
5. Belogay E, Wang Y (1999) Arbitrarily smooth orthogonal nonseparable wavelets in R^2. SIAM J Math Anal 30:678–697
6. de Boor C (1978) A practical guide to splines. Springer, New York
7. de Boor C, Hölig K, Riemenschneider S (1993) Box splines. Springer, New York
8. Chen G, Chui CK (1992) Signal processing and system theory. Springer, New York
9. Cho O, Lai MJ (2006) A class of compactly supported orthonormal B-Spline wavelets. In: Chen G, Lai MJ (eds) Wavelets and Splines. Nashboro Press, Brentwood, pp 123–151
10. Chrina M, Stöckler J (2007) Tight wavelet frames for irregular multiresolution analysis. Appl Comput Harmonic Anal (accepted for publication)
11. Chui CK (1988) Multivariate splines. SIAM Publications, Philadelphia
12. Chui CK (1992) An introduction to wavelets. Academic Press, San Diego
13. Chui CK (1997) Wavelets: A Mathematical Tool for Signal Analysis. SIAM Publication, Philedalphia
14. Chui CK, He W (2000) Compactly supported tight frames associated with refinable functions. Appl Comp Harmonic Anal 8:293–319
15. Chui CK, He W (2001) Construction of multivariate tight frames via Kronecker products. Appl Comp Harmonic Anal 11:305–312
16. Chui CK, Lian JA (1996) A study of orthonormal multi-wavelets. Appl Numer Math 20:273–298
17. Chui CK, Shi XL (1993) Bessel sequences and affine frames. Appl Comp Harmonic Anal 1:29–49
18. Chui CK, Wang JZ (1992) On compactly supported spline-wavelets and a duality principle. Trans Amer Math Soc 330:903–915
19. Chui CK, Stöckler J, Ward JD (1992) On compactly supported box-spline wavelets. Approx Theory Appl 8:77–100
20. Chui CK, He W, Stöckler J (2002) Compactly supported tight and sibling frames with maximum vanishing moments. Appl Comp Harmonic Anal 13:224–262
21. Cohen A, Daubechies I (1993) Nonseparable dimensional wavelet bases. Revista Math Ibereamericana 9:51–137
22. Cohen A, Daubechies I, Feauveau J-C (1992) Biorthogonal bases of compactly supported wavelets. Commun Pure Appl Math XLV:485–560
23. Daubechies I (1988) Orthonormal bases of compactly supported wavelets. Comm Pure Appl Math 41:909–996
24. Daubechies I (1992) Ten lectures on wavelets. SIAM Publications, Philadelphia
25. Daubechies I, Han B, Ron A, Shen ZW (2003) Framelets: MRA-based constructions of wavelet frames. Appl Comp Harmonic Anal 14:1–46
26. Dierckx P (1986) An algorithm for fitting data over a circle using tensor product splines. J Comp Appl Math 15:161–173
27. Donoho D (1995) De-noising by soft-thresholding. IEEE Trans Inform Theory 41:613–627
28. Donovan GC, Geronimo JS, Hardin DP, Massopust PR (1996) Construction of orthogonal wavelets using fractal interpolation functions. SIAM J Math Anal 27:1158–1192
29. Donovan GC, Geronimo JS, Hardin DP (1996) Interwining multiresolution analyses and the construction of piecewise polynomial wavelets. SIAM J Math Anal 27:1791–1815
30. Geronimo J, Lai MJ (2006) Factorization of multivariate positive Laurent polynomials. J Approx Theory 139:327–345
31. Goodman T (2003) A class of orthonormal refinable functions and wavelets. Constr Approx 19:525–540
32. Goodman T, Micchelli CA (1994) Orthonormal cardinal functions. In: Chui CK et al (eds) Wavelets: Theory, algorithms and applications. Academic, San Diego, pp 53–88
33. Han B (2007) Construction of wavelets and framelets by the projection method. Int J Math Sci, to appear
34. Hardin D, Hogan A, Sun Q (2004) The matrix-valued Riesz lemma and local orthonormal bases in shift-invariant spaces. Adv Comput Math 20:367–384
35. He W, Lai MJ (1997) Examples of bivariate non-separable continuous compactly supported orthonormal wavelets. Proc SPIE 3169:303–314. (It also appears in IEEE Trans Image Proces 9:949–953 (2000)
36. He W, Lai MJ (1999) Construction of bivariate compactly supported biorthogonal box spline wavelets with arbitrarily high regularities. J Appl Comput Harmonic Anal 6:53–74
37. He W, Lai MJ (2003) Construction of trivariate compactly supported biorthogonal box wavelets. J Approx Theor 120:1–19
38. Hong J, Lai MJ (2007) New constructions of orthonormal multiwavelets. Manuscript
39. Jia RQ, Micchelli CA (1992) Using the refinement equation for the construction of pre-wavelets, V. Extensibility of trigonometric polynomials. Comput 48:61–72

40. Jia RQ, Shen ZW (1994) Multiresolution and wavelets. Proc Edinburgh Math Soc 37:271–300
41. Jia R-Q, Wang J, Zhou D-X (2003) Compactly supported wavelet bases for Sobolev spaces. Appl Comput Harmon Anal 15(3):224–241
42. Johnstone I, Donoho D (1994) Ideal spatial adaptation via wavelet shrinkage. Biometrika 81:425–455
43. Karoui A (2003) A note on the construction of nonseparable wavelet bases and multiwavelet matrix filters of $L_2(\mathbb{R}^n)$, where $n \geq 2$. Electron Res Announc Amer Math Soc 9:32–39
44. Kotyczka U, Oswald P (1995) Piecewise linear prewavelets of small support. In: Chui CK, Schumaker LL (eds) Approximation Theory VIII, vol 2. World Scientific, Singapore, pp 235–242
45. Kovačević J, Vetterli M (1992) Nonseparable multidimensional perfect reconstruction filter banks and wavelet bases for $(R)^n$. IEEE Trans Info Theory 38:533–555
46. Kovačević J, Vetterli M (1995) Nonseparable two- and three-dimensional wavelets. IEEE Trans Signal Proc 43:1260–1273
47. Lai M-J (1992) Fortran subroutines for B-nets of box splines on three and four directional meshes. Numerical Algo 2:33–38
48. Lai M-J (2002) Methods for constructing nonseparable compactly supported orthonormal wavelets. In: Zhou DX (ed) Wavelet analysis: Twenty year's development. World Scientific, Singapore, pp 231–251
49. Lai M-J (2006) Construction of multivariate compactly supported orthonormal wavelets. Adv Comput Math 25:41–56
50. Lai M-J, Lian J-A (2007) A private communication on CDF9/7, May 2007
51. Lai M-J, Lian J-A (2007) Construction of tight multi-wavelet frames (under preparation)
52. Lai M-J, Liu HP (2007) Prewavelet solution to Poisson equation. Manuscript
53. Lai M-J, Lyche T (2007) Tight wavelets frames using trigonometric B-splines. Manuscript
54. Lai M-J, Nam K (2006) Tight wavelet frames over bounded domains. In: Chen G, Lai M-J (eds) Wavelets and splines. Nashboro Press, Athens, pp 313–326
55. Lai M-J, Petukhov A (2007) Method of virtual components for constructing redundant filter banks and wavelet frames. Appl Comput Harmonic Anal 22:304–318
56. Lai M-J, Petukhov A (2007) Method of virtual components in the multivariate setting. (submitted)
57. Lai M-J, Roach DW (1999) Nonseparable symmetric wavelets with short support. In: Proceedings of SPIE Conference on Wavelet Applications in Signal and Image Processing VII, vol 3813, pp 132–146
58. Lai M-J, Roach DW (2001) Construction of bivariate symmetric orthonormal wavelets with short support. In: Kopotun K, Lyche T, Neamtu M (eds) Trends in approximation theory. Vanderbilt University Press, Nashville, pp 213–223
59. Lai M-J, Roach D (2002) Parameterizations of univariate orthogonal wavelets with short support. In: Chui CK, Schumaker LL, Stoeckler J (eds) Approximation theory X: Wavelets, splines, and applications. Vanderbilt University Press, Nashville, pp 369–384
60. Lai M-J, Schumaker LL (2007) Spline functions over triangulations. Cambridge University Press, Cambridge
61. Lai M-J, Stoeckler J (2006) Construction of multivariate compactly supported tight wavelet frames. Appl Comput Harmonic Anal 21:324–348
62. Lawton W, Lee SL, Shen ZW (1996) An algorithm for matrix extension and wavelet construction. Math Comp 65:723–737
63. Lebrun J, Vetterli M (1998) Balanced multiwavelets: Theory and design. IEEE Trans Signal Process 46:1119–1125
64. Lyche T (1999) Trigonometric splines: A survey with new results. In: Peńa J (ed) Shape preserving representations in computer-aided geometric design. Noa Science Publishers, New York
65. Lyche T, Schumaker LL (1994) L-spline wavelets. In: Chui C, Montefusco L, Puccio L (eds) Wavelets: Theory, algorithms, and applications. Academic Press, New York, pp 197–212
66. Lyche T, Schumaker LL (2000) A multiresolution tensor spline method for fitting functions on the sphere. SIAM J Sci Comput 22:724–746
67. Lyche T, Winther R (1979) A stable recurrence relation for trigonometric B-splines. J Approx Theory 3:266–279
68. Lyche T, Schumaker LL, Stanley S (1998) Quasi-interpolants based on trigonometric splines. J Approx Theory 95:280–309
69. Maass P (1997) Families of orthogonal two-dimensional wavelets. SIAM J Math Anal 27:1454–1481
70. Mallat S (1989) Multi-resolution approximations and wavelet orthonormal bases of $L_2(\mathbb{R})$. Trans Amer Math Soc 315:69–87
71. Nam K (2005) Box spline tight frames and their applications for image processing, Ph.D. Dissertation. University of Georgia, Athens
72. Petukhov A (2001) Explicit construction of framelets. Appl Comp Harmonic Anal 11:313–327
73. Petukhov A (2003) Symmetric framelets. Constr Approx 19:309–328
74. Riemenschneider S, Shen Z (1991) Box splines, cardinal series, and wavelets. In: Chui CK (ed) Approximation theory and functional analysis. Academic, Boston, pp 133–149
75. Riemenschneider S, Shen Z (1992) Wavelets and pre-wavelets in low dimensions. J Approx Theory 71:18–38
76. Ron A, Shen ZW (1997) Affine systems in $L_2(\mathbb{R}^d)$: The analysis of the analysis operator. J Func Anal 148:408–447
77. Ron A, Shen ZW (1997) Affine system in $L_2(\mathbb{R}^d)$, II. Dual systems. J Fourier Anal Appl 3:617–637
78. Ron A, Shen ZW (1998) Compactly supported tight affine spline frames in $L_2(R^d)$. Math Comp 67:191–207
79. Ron A, Shen ZW (1998) Construction of compactly supported affine frames in $L_2(R^d)$. In: Lau KS (ed) Advances in Wavelets. Springer, New York, pp 27–49
80. Rudin W (1963) The existence problem for positive definite functions. Illinois J Math 7:532–539
81. Said A, Pearlman WA (1996) A new fast and efficient image codec based on set partitioning in hierarchical trees. IEEE Trans Circ Syst Video Technol 6:243–250
82. Schoenberg IJ (1964) On trigonometric spline interpolation. J Math Mech 13:795–825
83. Schoenberg IJ (1973) Cardinal spline interpolation. SIAM Publication, Philadelphia
84. Schumaker LL (2007) Spline functions, basic theory. Cambridge University Press, Cambridge
85. Schumaker LL, Traas C (1991) Fitting scattered data on sphere like surface using tensor products of trigonometric and polynomial splines. Num Math 60:133–144
86. Simoncelli EP, Adelson EH (1990) Non-separable extensions of quadrature mirror filters to multiple dimensions. Proc IEEE 78:652–664

87. Stanhill D, Zeevi YY (1996) Two dimensional orthogonal wavelets with vanishing moments. IEEE Trans Signal Proces 46:2579–2590
88. Stanhill D, Zeevi YY (1998) Two dimensional orthogonal filter banks and wavelets with linear phase. IEEE Trans Signal Proces 46:183–190
89. Strang G, Fix G (1973) A Fourier analysis of the finite element variational method. In: Geymonat G (ed) Constructive aspects of functional analysis, C.I.M.E. II Ciclo 1971; 793–840
90. Vetterli M (1984) Multidimensional subband coding: some theory and algorithms. Signal Proces 6:97–112
91. Wickerhauser MV (1994) Adapted wavelet analysis from theory to software. AK Peters Ltd, Wellesley
92. Zhou J (2006) Construction of orthonormal wavelets of dilation factor 3 with application in image compression. Ph.D. dissertation, University of Georgia, Athens

Books and Reviews

Antoine JP, Murenzi R, Vandergheynst P, Ali ST (2004) Two-dimensional wavelets and their relatives. Cambridge University Press, Cambridge
Chen G, Lai M-J (2006) Wavelets and Splines: Athens, 2005. Nashboro Press, Brentwood
Cohen A (2003) Numerical analysis of wavelet methods. North-Holland Publishing Co, Amsterdam
Hernández E, Weiss G (1996) A first course on wavelets, with a foreword by Yves Meyer. CRC Press, Boca Raton
Mallat S (1998) Wavelet tour of signal processing. Academic Press, San Diego
MathWorks, Inc (2007) MATLAB wavelet toolbox
Meyer Y, Coifman R (1997) Wavelets, Calderon-Zygmund and multilinear operators. Hermann, Paris, 1990 and Cambridge University University Press, Cambridge
Strang G, Nguyen T (1996) Wavelets and filter banks. Wellesley-Cambridge Press, Wellesley
Vetterli M and Kovačević J (1995) Wavelets and subband coding. Prentice Hall, New Jersey
Vidakovic B (1999) Statistical modeling by wavelets. A Wiley-Interscience Publication. Wiley, New York

Positional Analysis and Blockmodeling

Patrick Doreian
Department of Sociology, University of Pittsburgh, Pittsburgh, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Conventional Blockmodeling
Generalized Blockmodeling
Examples of Generalized Blockmodeling
Future Directions
Bibliography

Glossary

Social network A simple social network is a set of social actors with one or more social relations defined over them. A network for one relation can be represented by a graph which is defined as an ordered triple $\boldsymbol{G} = (V, E, A)$. In this representation, V denotes the set of vertices (or nodes) that represent the social actors, E the set of edges (undirected lines) representing reciprocated or symmetric ties and A the set of arcs (directed lines) representing unreciprocated or directed ties. The cardinality of V, usually denoted by n, gives the size of the network. The sets E and A are pairwise disjoint and their union is the set of all the ties for the social relation, denoted by R. The graph can also be denoted by $\boldsymbol{G} = (V, R)$ given that $R = E \cup A$. Such network data are also called one-mode data (because there is only one type of social unit over which social relations are defined). If there are r relations, these are denoted by $R_1, R_2, \ldots, R_r$ and the social network is denoted by $\boldsymbol{G} = (V, R_1, R_2, \ldots, R_r)$. For ease of exposition, a network with one relation is used below. The ties, where the arcs and edges are viewed as lines in the pictorial representation of a graph, can be binary or valued.

Location in a network The location of an actor (v_i) is the set of ties that v_i has with all of the other actors in the network including the absence of ties.

Partitioning networks A partition of a social network is a simultaneous partitioning of the actors into *positions* and the social ties into *blocks*. For blockmodeling, this simultaneous partitioning is done solely in terms of the relational ties in R.

Position When the set of vertices (V) is partitioned into a set of clusters $\boldsymbol{C} = (C_1, C_2, \ldots, C_k)$ where $C_i \cap C_j = \varphi$, the empty set, for distinct i and j and $\cup_i C_i = V$, then each $C_i \in \boldsymbol{C}$ (and the vertices it contains) forms a *position*. (See Fig. 2, below). The number of positions is denoted by k. This partition determines an equivalence relation and the units within each cluster are said to be equivalent to each other (and not equivalent to any the remaining vertices).

Positional analysis A positional analysis of a social network is based on the locations of all vertices (representing actors) and positions (occupied by vertices).

Block $\boldsymbol{C}$ also partitions the relation, R, into k^2 blocks: $R(C_i, C_j) = R \cap (C_i \times C_j)$ where $\times$ denotes the Cartesian product. Each *block*, $R(C_i, C_j)$, consists of all of the arcs from vertices in C_i to vertices in C_j. The set of these arcs and edges is used to 'summarize' the relation between C_i and C_j by an arithmetic

calculation using the values of the arcs and edges in the block. A block is a relation between two clusters of vertices. (This is illustrated below in Sects. "Introduction", "Conventional Blockmodeling" and "Generalized Blockmodeling" and diagrammed in Fig. 2). When $i = j$ the blocks are called diagonal blocks and there are k of them, one for each position. The remaining $k(k-1)$ blocks are called nondiagonal blocks.

Blockmodel A *blockmodel* consists of structures obtained by: (i) identifying all vertices within a cluster (in the clustering $\boldsymbol{C}$) and representing each cluster as a vertex to construct k vertices for another graph and (ii) combining all of the ties in a block into a single tie between positions and constructing one tie for each block. If there are no ties in a block there is no tie between the two positions defining the block.

Blockmodel image The blockmodel is another network (graph) with many fewer vertices than $\boldsymbol{G}(k \ll n)$ and many fewer ties. When this blockmodel is represented as a simpler (and smaller) matrix or graph the result is labeled the *blockmodel image*. Even if there are no loops (ties from vertices to themselves) in $\boldsymbol{G}$ there can be loops in the blockmodel image depending on the ties present in diagonal blocks.

Blockmodeling Blockmodeling refers to the set of techniques used to discern network structure as blockmodels and representing them.

Two-mode network data A two-mode network $\boldsymbol{N} = (V_1, V_2, R, w)$ has one set of social units $V_1 = (v_{11}, v_{12}, \ldots, v_{1p})$ and a second set of units $V_2 = (v_{21}, v_{22}, \ldots, v_{2q})$ where $V_1 \cap V_2$ is empty. The social relation R is a subset of $V_1 \times V_2$ and is a relation *between* the two sets of vertices. The item, w, is a set of weights (viewed as a mapping of the relational ties of R to the real numbers).

Definition of the Subject

There are many types of social actors ranging from individuals to groups to organizations to institutions to societies. Many different social relations can be defined for all of these types of social actors. Depending on the substantive concerns of a researcher, many different and diverse social networks can be studied. All these networks can be represented by graphs as described above. It commonplace to talk of the structure of these networks: communication networks, transportation networks, organizational networks, trading networks all have structures. Two ideas are central to social network analysis. The first is that the structure *of* a social network, as a whole, is important for collective outcomes at the level of the network. The second is that the location occupied *in* a network is important for outcomes at the actor level – regardless of the network that is studied. In order to study these collective and individual outcomes it is *essential* to know how a network is structured. Blockmodeling is an approach that is highly effective in identifying the overall structure of a network and describing it. Large and/or complex networks are mapped into simpler structures – called blockmodel images – that are viewed as structural summaries of the large and/or complex social networks. In one type of imagery, the structure depicted in a blockmodel image is the fundamental structure of the network and the observed network is an instantiation of the fundamental structure. Knowing the fundamental structure of a network is the primary goal. In turn, this permits the understanding of many network phenomena at both the collective and local levels.

Introduction

The foundational paper for blockmodeling [19] introduced the concept of structural equivalence. Two actors (vertices) are structurally equivalent if they are connected to exactly the same other actors in the network. Their locations are identical. In this sense they are structurally identical and occupy exactly the same location in the network. Grouping together all of the vertices having the same network location defines a position occupied jointly by these (structurally equivalent) vertices. If all vertices can be grouped into sets of structurally equivalent clusters, the way is open to reduce a network to a simpler structure defined over these positions. The first widely accepted generalization of structural equivalence equivalence was regular equivalence as proposed in [21] where two vertices are regularly equivalent if they are equivalently connected to equivalent others. The difference between the two conceptions of equivalence is illustrated in Fig. 1 which shows a simple hierarchy twice. On the left, an exact structural equivalence partition is shown where the vertices have been color coded so that vertices in the same equivalence class have the same color. Note that in the lower two levels of the hierarchy not all of the vertices at a given level have the same color, a representation that appears to ignore a basic feature of this network. On the right, a regular equivalence partition is shown with three positions, one for each level. (As an illustration of the notion of regular equivalence, vertices j, k, and l are connected to d in the same way as h and i are connected to c while c and d are connected in the same way to their subordinates.) There is an additional subtlety here because every network has a lattice of regular equivalence partitions [4]. Both of the

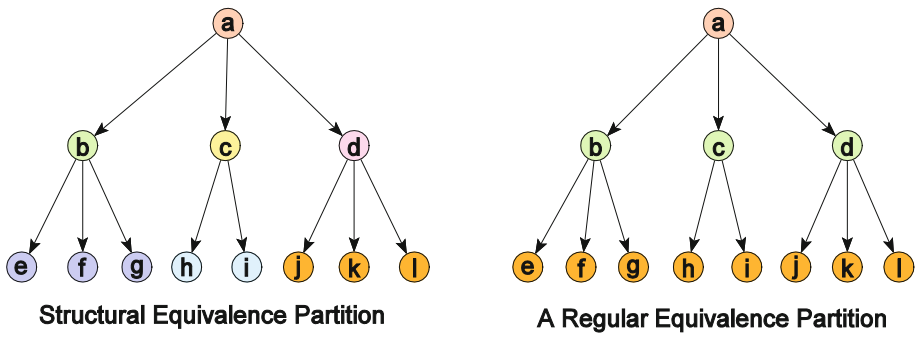

Positional Analysis and Blockmodeling, Figure 1
Two partitions of a hierarchy

partitions shown in Fig. 1 are located in the lattice of regular partitions of the network shown in the figure.

Of course, most empirical networks cannot be partitioned *exactly* in terms of structural or regular equivalence in a useful fashion. (Extremely fine grained partitions are possible for structural equivalence where k approaches n with the result that the blockmodel image is almost as large as the empirical network, an outcome that is not very useful. And, in terms of regular equivalence, there are the trivial partitions with 1 or n clusters for a network without isolates). So, in practice, partitions are sought that come as 'close as is possible' to the underlying conception of equivalence used to partition the network. In short, the notion of an exact equivalence partition is approximated when partitioning empirical networks to get a blockmodel. Conventional blockmodeling and generalized blockmodeling provide two approaches to establishing blockmodels empirically where the concept of an exact equivalence partition is approximated.

Conventional Blockmodeling

Structural equivalence is defined as follows: x and y are structurally equivalent iff (i) xRy iff yRx; (ii) xRx iff yRy; (iii) For all $z \in V \backslash \{x, y\}$, xRz iff yRz and (iv) For all $z \in V \backslash \{x, y\}$, zRx iff zRy. In order to have a practical empirical procedure, conventional blockmodeling operationalizes an approximation to exact structural equivalence in terms of a metric. For example, if two vertices are structurally equivalent, the correlation of the vectors giving their locations is 1 because the vectors are identical. Similarly, the Euclidean distance between the two such vectors is 0. The correlation between "almost equivalent" locations will be close to 1 and the Euclidean distance between them will be close to 0. So "exactly equivalent" becomes "almost exactly equivalent" and clustering procedures can be applied to matrices containing (dis)similarities that have been constructed from the relational data. This strategy has been called this 'the indirect approach' because the structural network data have been replaced by the (dis)similarity measures used for the clustering procedures [14]. One widely used algorithm, CONCOR [7], uses iterated correlations while another widely used algorithm, STRUCTURE [8], uses (corrected) Euclidean distances. Variants of both are implemented in UCINET [6].

An equivalence $\approx$ on V is a regular equivalence on $\boldsymbol{G} = (V, R)$ iff for all $x, y, z \in V$, $x \approx y$ implies both (i) xRz implies there exists $w \in V$ such that (yRw and $w \approx z$) and (ii) zRx implies there exists $w \in V$ such that (wRy and $w \approx z$). As for structural equivalence, regular equivalence does not apply exactly for most empirical networks. REGGE [20] has been used to relax exact regular equivalence in the form of an iterative algorithm to locate empirical partitions based on approximate regular equivalence. Thus far, no satisfactory metric for approximating exact regular equivalence has been established and, as a result, indirect methods for establishing partitions based on regular equivalence have been limited in their applicability. For a more general discussion of positions and equivalence, including automorphic equivalence, see [5].

Generalized Blockmodeling

Generalized blockmodeling has been proposed as an alternative approach to conventional blockmodeling for establishing blockmodels empirically [14]. (See also [1]). While it shares the goal of discerning the structure of a network via homomorphisms of the network to the blockmodel image, it has many distinctive features: (i) it is a direct approach that works with the relational data; (ii) it generalizes the notion of equivalence in a way that permits a series of indefinite extensions; (iii) it uses an optimizational approach with an explicit criterion function; (iv) this cri-

terion function is used as measure of fit for established blockmodels; and (v) generalized blockmodeling can be used both inductively and deductively.

Direct Approach

The direct approach is facilitated by translating the notion of equivalence into the set of block types that are consistent with the equivalence. Such block types are called *permitted* or *ideal* block types for the equivalence. Thus, structural equivalence is turned into the four block types that it implies: null block where every element is 0, complete blocks where every element is 1, diagonal blocks that have 0 only on the diagonal and 1 elsewhere, and diagonal blocks with only 1 on the diagonal and 0 elsewhere. (Of these block types, the fourth is empirically rare and is seldom used.) Under exact structural equivalence no other block types are possible. Regular equivalence is treated in a similar fashion: The only block types that regular equivalence permits are null blocks and 1-covered blocks [3]. A 1-covered block has at least one 1 in every row and every column. (Note that a complete block is a special case of a 1-covered block, consistent with structural equivalence being a special case of regular equivalence.) The implication of redefining equivalence in terms of a set of permitted block types means that an empirical procedure can be constructed by focusing on the extent to which the permitted block types are present in a network instead of transforming the structural data into (dis)similarities.

Generalizing Equivalence Ideas

By focusing instead on permitted block types the way is cleared for defining new block types and new types of blockmodels. One way of doing this is to modify the definition of extant block types. The idea of 1-covered blocks can be relaxed from regular to row-regular blocks and column-regular blocks. The former have at least one 1 in every row while the latter has at least one 1 in each column. It follows that if a block is both row-regular and column-regular then it is regular. A row-dominant block is one where there is at least one row consisting entirely of 1s. If the clusters of the blockmodel partition are C_i and C_j, then at least one element in C_i has a tie directed all vertices in C_j. Column-dominant blocks can be defined in a similar fashion. (See pp. 211–215 in [14] for further examples of new block types).

New types of block types can be defined from scratch. An example of doing this stems from structural balance theory [17]. One generalization of Heider's theory [9] showed that if a signed network is structurally balanced according to Heider, there will be a clear partition structure: the vertices will be partitioned into two clusters so that all of the positive ties will be within the clusters and all of the negative ties will be between the clusters. By modifying the definition of balance, this can be generalized so that there would be two or more clusters with the same partition structure [10]. This suggests a natural pair of block types and a new type of blockmodel [11]. The new block types are positive block (where the only elements are positive or null ties) and negative block (where the only possible ties are negative and null ties). According to the theory, the positive blocks will be on the diagonal and the negative blocks will be nondiagonal blocks. An algorithm for detecting partitions that are as close as possible to exact structural balance is suggested in [11]. Another example is the ranked clusters blockmodel where the diagonal blocks have vertices that are linked only by edges (or by null ties) and all of the arcs either go up the ranking or down the ranking [12]. Each of these blockmodels is an example of defining new block types and, with them, new types of blockmodels that can be fitted to empirical data.

The general scheme for generalized blockmodeling is shown in the diagram shown in Fig. 2. There, vertices in the cluster C_i are shown, generically, as x and the vertices of C_j are shown as y. The set of arcs from vertices in C_i to vertices in C_j is shown as a_{xy} in the figure. The cluster C_i of the original network is mapped under a mapping μ to the position i (as a vertex in the blockmodel) and the cluster C_j is mapped to j (another position as

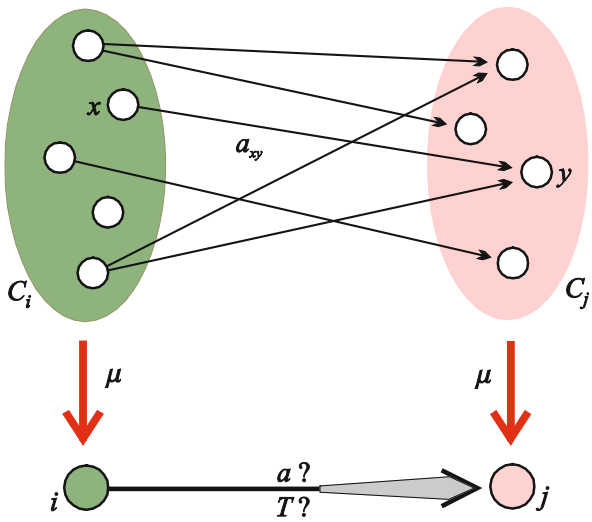

Positional Analysis and Blockmodeling, Figure 2
A diagram of the generalized blockmodeling scheme

a vertex in the blockmodel) also under μ. The set a_{xy} is summarized, via arithmetic rules, as a in the blockmodel and T is a set of predicates that are used to describe the block types of the blockmodel. Thus, for structural equivalence, $T = \{\text{null, complete}\}$ and for regular equivalence $T = \{\text{null, 1-covered}\}$. For structural balance, $T = \{\text{positive, negative}\}$.

Criterion Functions

When permitted block types are defined, it is straightforward to define departures from the permitted (ideal) block types. For structural equivalence, every 1 in what is thought to be a null block is inconsistent with structural equivalence. Similarly, every 0 in what is thought to be a complete block is inconsistent with structural equivalence. The more inconsistencies there are in a block the less it conforms to the corresponding ideal block type. And the more inconsistencies there are, when the block inconsistencies are combined for the whole network, the worse the fitted empirical block model. Similarly, for regular equivalence, every row or column in a block lacking a 1 is not consistent with the block being 1-covered and every 1 in an otherwise null block is an inconsistency with regular equivalence. The more inconsistencies there are, when they are summed across the blocks, the worse the fitted blockmodel when regular equivalence has been used. Note that, in general, the count of inconsistencies for structural equivalence will differ from the count of inconsistencies for regular equivalence because the block types differ. Counting inconsistencies in blocks and combining them to give the total number of inconsistencies for a blockmodel is the core of the idea of a criterion function that can be used to establish blockmodels.

Let $\boldsymbol{G} = (V, R)$ be a network and let Θ denote the set of all equivalence relations of a given type (for example structural equivalence, or regular equivalence, or balance theoretic, or ranked clusters). Every equivalence, say $\sim$, on V determines a partition $\boldsymbol{C}$ of V and vice versa. Let Φ denote the set of all partitions corresponding to the relations from Θ. In general, we need a criterion function, denoted $P(\boldsymbol{C})$, to satisfy two criteria: (i) $P(\boldsymbol{C}) \geqslant 0$ and (ii) $P(\boldsymbol{C}) = 0$ iff $\sim \in \theta$. With such a $P(\boldsymbol{C})$ we can set up a clustering problem to minimize $P(\boldsymbol{C})$ (given the equivalence specified for a blockmodel). If there are exact equivalences of a given type then $P(\boldsymbol{C})$ is 0. However, if there are no such partitions (i. e. Θ is empty) then an optimization approach gives the solution(s) that differ(s) the least from an ideal case – provided that the criterion function is compatible with and sensitive to the type of equivalence that is used.

The intuition behind using these criterion functions is that of a *pair of blockmodels* that are compared in terms of the criterion function. One is an ideal blockmodel with only the permitted block types for a given equivalence and the other is a corresponding empirical blockmodel (i. e. with the same partition of the vertices) for a clustering $\boldsymbol{C}$. Let $B(C_i, C_j)$ denote the set of all of the ideal blocks (in the blockmodel) and let $p(C_i, C_j)$ denote the inconsistency between the empirical block defined by C_i and C_j and the corresponding nearest ideal block. The criterion function for the blockmodel is expressed as $P(\boldsymbol{C}) = \Sigma\, p(C_i, C_j)$ where the summation is over all k^2 blocks. At the block level, the block inconsistency is $p(C_i, C_j) = \min \delta(R(C_i, C_j), B)$ where the minimization is over all possible blocks for $B(C_i, C_j)$ and $\delta(R(C_i, C_j), B)$ measures the deviation (number of inconsistencies) between $R(C_i, C_j)$ and the nearest ideal block B. For structural equivalence, the simplest inconsistency measure for a block whose nearest ideal block is null is the number of 1s in it. Put differently, $\delta(R(C_i C_j), B)$ is the number of 1s that are present where they ought not be. Similarly, the simplest inconsistency ($\delta(R(C_i, C_j), B)$) for a block whose nearest ideal block is complete is the number of 0s in it. The total number of inconsistencies is obtained by summing the inconsistencies over all of the blocks in the blockmodel. This gives us $P(\boldsymbol{C})$. This can be generalized by using differential weights for the types of inconsistencies between the ideal and empirical blockmodels. For example, for structural equivalence partitions, it is possible to view 1s in what is specified as a null block as more serious than 0s in what is thought to be a complete block. If so, not all inconsistencies are viewed as being equally important and penalties can be imposed on some of them. Of course, this modifies the definition of, and values returned for, the criterion function, $P(\boldsymbol{C})$. The 'best' partitions $\boldsymbol{C}$ are not known ahead of time and have to be identified empirically. A simple relocation algorithm can be used determine one (or more) partitions that minimize the criterion function.

Optimization and Measures of Fit

The relocation algorithm is defined and mobilized as follows. To locate a partition (as consistent as possible with the set of permitted block types defining the equivalence used) into k clusters, the network is partitioned randomly into k clusters. The value of $P(\boldsymbol{C})$ is then computed. Of course, given that the starting partition is random, this value will be very large. The starting (and any) partition can be changed in two ways: (i) interchanging a pair of vertices between two different clusters or (ii) moving

a vertex from one cluster to another. These two types of change (transformations) determine the neighborhood of any clustering, $\boldsymbol{C}$. The algorithm is expressed as follows:

Repeat:
  if in the neighborhood of the current clustering, $\boldsymbol{C}$,
    there exists a clustering, $\boldsymbol{C}'$, such that $P(\boldsymbol{C}') < P(\boldsymbol{C})$
  then move to the clustering $P(\boldsymbol{C}')$.

This continues until a smaller value of $P(\boldsymbol{C})$ cannot be found. The relocation algorithm has to be repeated many times (hundreds or thousands of times) for different random starting partitions for a given value of k in order to obtain solutions with the minimal values of $P(\boldsymbol{C})$.

A criterion function is defined *explicitly for a particular type of blockmodel*. As noted above, for regular equivalence there are two ideal block types. The value of $\delta(R(C_i, C_j), B)$ is defined differently for each of these types. If B is a null block, then one specification of $\delta(R(C_i, C_j), B)$ is the number of 1-covered rows and if B is a regular block, $\delta(R(C_i, C_j), B)$ is the number of rows or columns that have only 0s. (Other specifications are possible.) For a structural balance theoretic partition, there are two block types – positive and negative blocks. For positive blocks (on the diagonal) every -1 (for binary signed networks) contributes to $\delta(R(C_i, C_j), B)$ and for negative blocks (off-diagonal) every $+1$ (for binary signed networks) contributes to $\delta(R(C_i, C_j), B)$ for that block. If P is the total number of 'positive inconsistencies' and N is the total number of 'negative inconsistencies' then one specification for the criterion function is $P(\boldsymbol{C}) = P + N$. Alternatively, these two inconsistency types can be weighted differently as in $P(\boldsymbol{C}) = \alpha P + (1 - \alpha)N$ with $0 \leqslant \alpha \leqslant 1$. This formulation extends naturally to valued signed networks. A more extended discussion is contained in [13] and generalized blockmodeling is implemented in [3].

Deductive and Inductive Uses of Blockmodeling

Within conventional blockmodeling, using the indirect approach, blockmodeling is essentially inductive. A specific type of equivalence is specified, some (dis)similarity measure is specified (often implicitly) and a clustering algorithm is used to identify a clustering. For example, in using structural equivalence as the selected equivalence, corrected Euclidean distances can be computed and used in conjunction with the Ward criterion in a hierarchical clustering procedure. This usage results in a dendrogram or a cluster diagram and it is then necessary for a researcher to identify a clustering given the dendrogram. Neither the number of clusters is specified in advance nor is the location of the permitted types in a blockmodel specified. This is purely inductive. In another variant of using structural equivalence, CONCOR can be used to give successive splits of previously established clusters (starting with the whole network as a cluster) where the number of splits is specified in advance. This, too, is inductive and the researcher accepts (or not) the clusters that happen to be returned by the application of the algorithm. In contrast, generalized blockmodeling can be used in a pre-specified (and hence deductive use of blockmodeling) fashion. The different options are given in Table 1.

A simple pre-specified model is one where the block types and their location in the blockmodel are specified in advance but the actual clustering, C, is not known and is to be "discovered". Two examples of pre-specified blockmodels for $k = 3$ are given in Table 2 where both the types

Positional Analysis and Blockmodeling, Table 1
Inductive and deductive uses of generalized blockmodeling

		Clustering	Blockmodel
Inductive		Unknown (in advance)	Unknown (in advance)
Deductive	Pre-specification	Unknown	Given
	Constraints	Given	Unknown
	Constrained Pre-specification	Given	Given

Positional Analysis and Blockmodeling, Table 2
Two examples of pre-specified blockmodels

Model 1			Model 2		
Complete	Complete	Null	Row Regular	Row Dominant	Complete
Complete	Regular	Null	Null	Complete	Null
Regular	Null	Null	Null	Regular	Symmetric

of blocks and their place in the blockmodel are specified. These specifications state that the type of blockmodel is known or conjectured but not its exact realization.

It is permissible to formulate alternative blockmodels for the same network as alternative hypothetical models. However, the two criterion functions cannot be compared directly (in magnitude) because different block types define different inconsistencies. It is possible – and interesting – for different blockmodels to fit the same network.

In general, the criterion function for structural equivalence is the most stringent with every 0 and 1 in the 'wrong' places (blocks of the 'opposite' type) contributing while the number of inconsistencies under regular equivalence is likely to be much smaller. So comparing the criterion functions for different blockmodels based on structural and regular equivalence for the same data is inadmissible. In short, the metrics differ and the size of the criterion function matters only within a *given* blockmodel type for a specific network data set – where the aim is to find the closest empirical blockmodel(s) consistent with the selected type of blockmodel for the network studied. What matters is the appropriate type of blockmodel given the problem studied. The use of criterion functions is to identify the best fitting blockmodels of a given type and they should never be compared across different types of blockmodels.

An alternative option is to specify the clustering and leave open the exact nature of the blockmodel. The constraint takes the form of specifying which vertices go in which cluster and, given that, seeing the blockmodel type that results (in terms of the permitted block types). In the first example used below the presence of two types of organizations in the network suggested a partition of the organizations into the two types. Under this strategy, it is possible also to identify only partially the membership of the clusters by specifying that some vertices must be clustered together and other pairs of vertices are never clustered together. The third approach listed in Table 1 is the constrained pre-specification where both the blockmodel and the clustering are specified. (In the organizational example just mentioned, the partition into two types of organizations could be coupled to a specification of one type of a core-periphery blockmodel.)

Engaging in pre-specifying blockmodels requires social knowledge about the specific network or the type of network that it represents. And it may be that a lot of such knowledge is required to make blockmodel specifications with confidence. However, often a network analyst knows more about a network than what is implied by a completely inductive approach. As such, a pre-specified blockmodel is one that can be *tested* rather than one that is discovered. All of the inductive, partially deductive and fully deductive uses of blockmodeling are useful and the decision over which approach to take depends on the knowledge that is available about a network. For example, a type of blockmodel that is discovered for a network of a given type (e. g. corporate networks in the United States) becomes a candidate for a blockmodel to be tested in another network of the same broad type in another country.

Examples of Generalized Blockmodeling

Example 1. An Interorganizational Network

The first example features a small network with 15 organizations and is taken from [15]. The studied organizations are involved in a specific collaborative enterprise and the relation shown in Fig. 3 is the presence (or not) of reciprocated ties prior to the start of the collaborative enterprise. Even though this is a small network, the structure is not immediately apparent from the diagram. Prior to the analysis, the basic intuition was that these organizations form some kind of a core-periphery structure. Of course, there are many types of core-periphery structures so something more has to be specified. Those organizations thought to be in the core were expected to be tied to each other and to most of the other organizations in the network. These organizations occupy the core as Position 1. Another set of organizations was each thought to be tied to most of the core organizations but not to each other. These organiza-

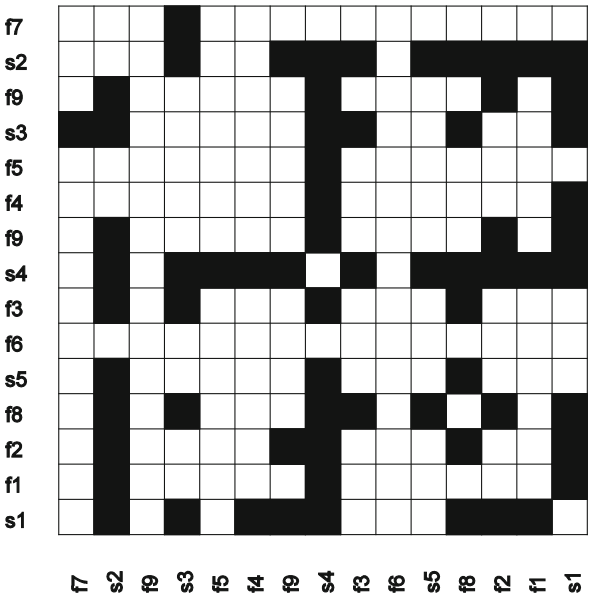

Positional Analysis and Blockmodeling, Figure 3
An interorganizational network matrix in an arbitrary order

Positional Analysis and Blockmodeling, Table 3
A pre-specified blockmodel

	Position 1	Position 2	Position 3
Position 1	Complete	Regular	Null
Position 2	Regular	Null	Null
Position 3	Null	Null	Null

tions occupy Position 2 and can be called the semi-periphery. Finally, there was the suspicion that one organization largely unknown prior to the collaborative effort. This organization (or possibly more than one organization) occupies Position 3. Reflecting these arguments, the specification used in [15] is shown in Table 3.

This pre-specified model was fitted with 3 clusters and the rows of the matrix were permuted so that members of the same position (cluster) are grouped together. The matrix array in Fig. 4 resulted. The blue lines (that extend beyond the matrix on the left and at the bottom) separate the three positions. While the model fits the data well, there are some inconsistencies: among the core organizations in Position 1, s6 and s1 do not have a reciprocated tie (and contribute 2 inconsistencies to the criterion function as 0s in a specified complete block) and, among the organizations in the semi-periphery (Position 2), f9 and f2 do share a tie when the expectation expressed in the model was there would be no ties there. This brings the number of inconsistencies to 4. There is a lone organization in Position 3 that has no reciprocated ties with the other

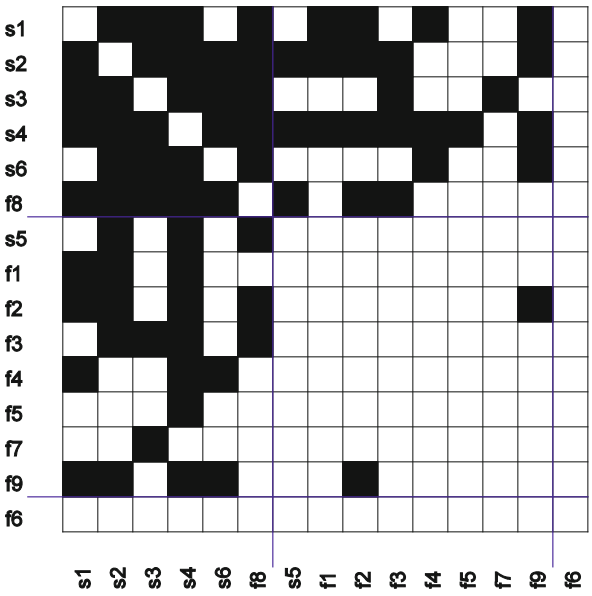

Positional Analysis and Blockmodeling, Figure 4
Blockmodel of the interorganizational network with 3 positions

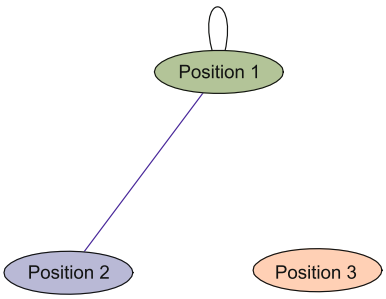

Positional Analysis and Blockmodeling, Figure 5
Blockmodeling image for Fig. 4

organizations: the blocks defined in terms of this position contribute no inconsistencies. So the optimized value of the criterion function is 4 and this partition is unique. Figure 5 shows the blockmodel image for this network gotten by using the pre-specified model shown in Table 3. The loop for Position 1 represents a complete block and the edge between Position 1 and Position 2 represents a regular link.

This example raises another interesting issue. As noted above in Sect. "Conventional Blockmodeling", the organizations could be grouped into two clusters based on what they do. One was made up of 'practicing' organizations (with regard to health service delivery) and the other was made up of organizations 'supporting' the collaborative effort. The former organizations are labeled f1 through f9 and the second set s1 though s6. An alternative blockmodel was an obvious candidate and is consistent with the third deductive use of blockmodeling listed in Table 1. It has the same specification of block types with three positions: Position 1 made up of the support organizations; Position 2 made up of all of the practicing organizations. (Position 3 could be specified for f6 as an isolate.) Much as this blockmodel made sense a priori, it has a much worse fit. Looking at Fig. 4, it is clear that if s5 and f6 are interchanged between Positions 1 and 2, the count of inconsistencies would jump by 12 to 16. The structure of the second specification is right but the composition of the positions was inferior – which shows, for this network, that the constrained pre-specification did not fit as well as a simpler pre-specification with only the blockmodel stated ahead of the model fitting. It suggests also that *testing* specifications is important and that the assumed social knowledge for specified blockmodel need not be correct.

Example 2. Structural Balance and Signed Networks

The data for this example are taken from [18] for a group of 21 women living in a college residence. Measurement for signed social relations was done by using four rela-

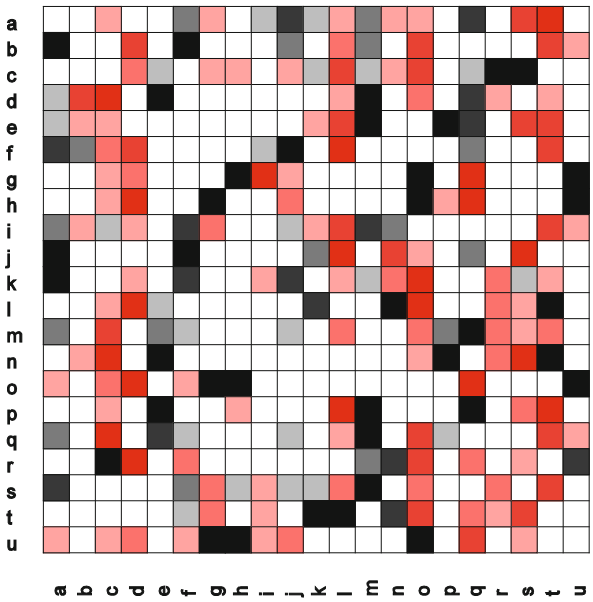

Positional Analysis and Blockmodeling, Figure 6
Original signed and valued matrix

tions that tapped an underlying dimension of affect. Each woman was asked to provide signed data about the other women in the residence in the form of who they would like to do activities with and those with whom they did not want to do these things. The items used were: having them as a room mate; going on a double date with them; taking them home to visit the family for a weekend and being a friend after college. These relations have been combined into counts of activities to be shared and counts of activities not to be shared are shown in Fig. 6.

Black squares indicate the presence of positive ties for all four relations with red squares showing where all four ties were negative. Shades of gray show lesser counts of positive ties (+3, +2 and +1) and shades of pink show lesser counts of negative ties (−3, −2 and −1). White squares indicate no preference either way. (There were no cases where, for a pair or women, there was a positive tie on one relation and a negative tie on another relation). If structural balance holds exactly then a signed relation will partition the actors into clusters (also called plus-sets) where all of the ties within a cluster are positive and all ties between the clusters are negative. In practice, structural balance does not hold exactly and there are some negative ties within plus-sets and some positive ties between them. These are all inconsistencies with structural balance and the total count of them gives the value of a criterion function. Partitions were sought that minimize this criterion function, $P(\boldsymbol{C})$, and this depends on the number of clusters (k) used. Table 4 summarizes the outcomes leading to the choice of the partition shown in Fig. 6. (The criterion function, $P(\boldsymbol{C})$, is larger for higher values of k).

Positional Analysis and Blockmodeling, Table 4
Values of $P(C)$ and the number of solutions for values of k

k	2	3	4	5	6
$P(\boldsymbol{C})$	48.5	37	32	32.5	33
Solutions	1	1	1	2	1

The best fitting (and unique) partition is with 4 clusters. The pre-specified blockmodel for this case is:

Positive (+, 0)	Negative (−, 0)	Negative (−, 0)	Negative (−, 0)
Negative (−, 0)	Positive (+, 0)	Negative (−, 0)	Negative (−, 0)
Negative (−, 0)	Negative (−, 0)	Positive (+, 0)	Negative (−, 0)
Negative (−, 0)	Negative (−, 0)	Negative (−, 0)	Positive (+, 0)

To locate the best fitting partition(s), the relocation algorithm described above was used with $P(\boldsymbol{C}) = \alpha P + (1-\alpha)N$ with $\alpha = 0.5$. The unique best fitting partition is shown in Fig. 7 where the blue lines mark the boundaries between the four plus-sets. Any red or pink square in the diagonal blocks shows the presence of negative ties, in what is specified as positive blocks, as inconsistencies contributing to $P(\boldsymbol{C})$. Similarly, black and gray squares in the off-diagonal blocks show positive ties that are present in what are specified as negative blocks and they also con-

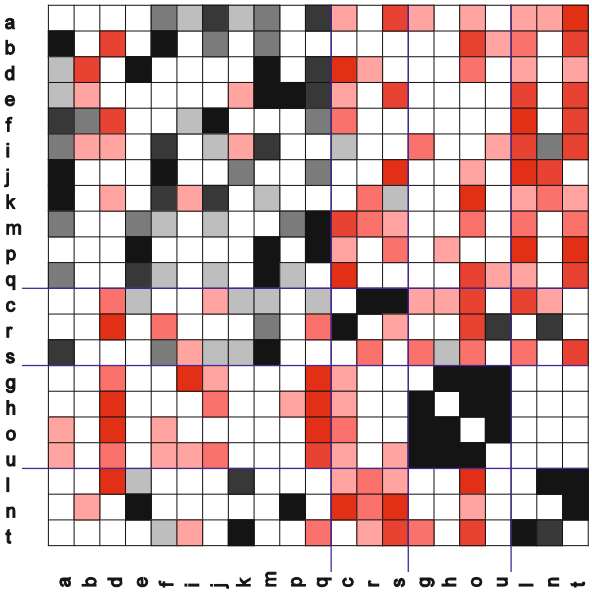

Positional Analysis and Blockmodeling, Figure 7
Permuted signed and valued matrix with 4 plus-sets

tribute to $P(\boldsymbol{C})$. The plus-set {g, h, o, u} is particularly cohesive with all their positive ties being distributed among themselves and all their negative ties being sent to women in other plus-sets. Furthermore, the women in this plus-set receive no positive ties from women in other plus-sets. The plus-set {l, n, t} has only positive ties within their plus-set but two positive ties are sent to women in other plus-sets. All of the negative ties are sent outside the plus-set. The other three person plus-set is less consistent with the ideal structural balance partition having two negative ties within it and positive ties to women in other plus-sets. The blocks associated with the large plus-set have both kinds of inconsistencies. In summary, the generalized blockmodeling discerned a structural pattern as a partition of a valued signed relation that is readily interpretable.

Example 3. The Supreme Court as Two-Mode Data

Two-mode data involves two sets of social objects. Examples of such data include companies employing people, people attending events, organizations joining alliances and Supreme Court Justices voting on cases they hear. While the row objects and column objects belong to different sets, it is still possible to apply blockmodeling tools to two-mode data. Of course, the rows and columns will be partitioned differently. For the example of the voting of Supreme Court Justices there are two possible weights: 1 for voting in favor of a decision and 0 for not so voting. (An alternative weighting scheme has three values where 1 is voting for a decision, 0 is not voting at all (for justices recusing themselves from a case) and -1 for voting in dissent). The simpler weight scheme is used in this example.

Instead of seeking a clustering $\boldsymbol{C}$ as used for one-mode data, a two-clustering is sought where $\boldsymbol{C} = (\boldsymbol{C}_1, \boldsymbol{C}_2)$ with $\boldsymbol{C}_1$ a partition of V_1 and $\boldsymbol{C}_2$ a partition of V_2. As for one-mode data, block types, together with their associated inconsistency counts, can be defined and an optimization problem formulated as (Φ, $P(\boldsymbol{C})$, minimize). The solution set is made up of two-clusterings $\boldsymbol{C}^* = (\boldsymbol{C}_1^*, \boldsymbol{C}_2^*)$ for which $P(\boldsymbol{C}^*)$ is minimized over all $\boldsymbol{C} \in \Phi$, the set of feasible two-clusterings. The value of the criterion function, $P(\boldsymbol{C}^*) = P(\boldsymbol{C}_1, \boldsymbol{C}_2)$ is obtained by summing all of the inconsistencies for each block of the two-mode blockmodel. Structural equivalence blocks (null and complete) was used in [14] to examine the voting of the Supreme Court for the 2000–1 term for the 26 most important decisions identified in [16](a). The justices were partitioned into 4 clusters and the cases into 7 clusters with 13 inconsistencies. Their optimal solution partition is shown in Fig. 8 where the four clusters of justices are in the center

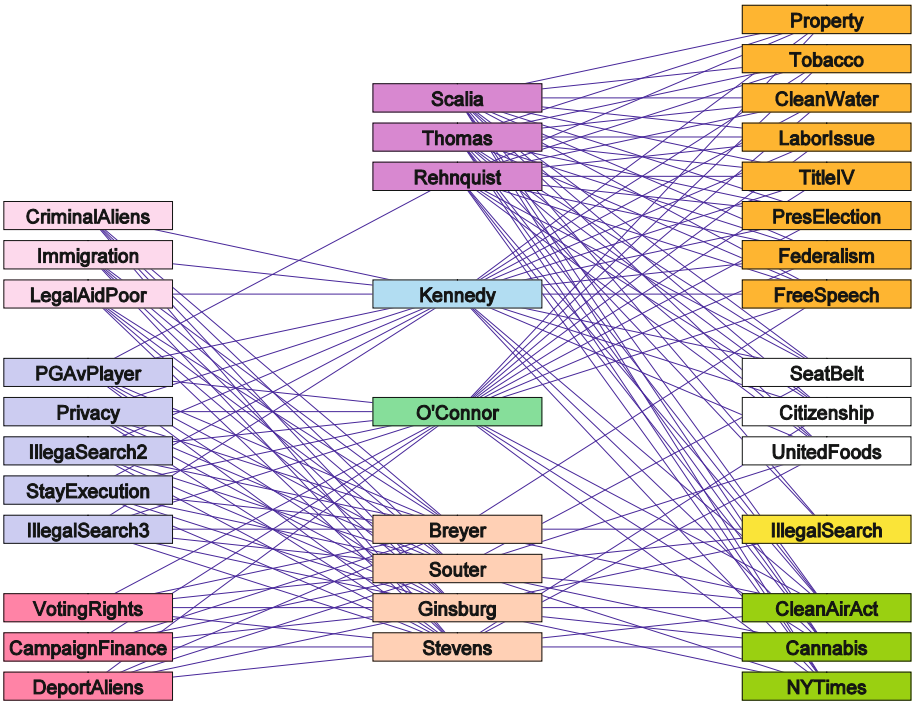

Positional Analysis and Blockmodeling, Figure 8
A partition of the Supreme Court voting for 2000–2001

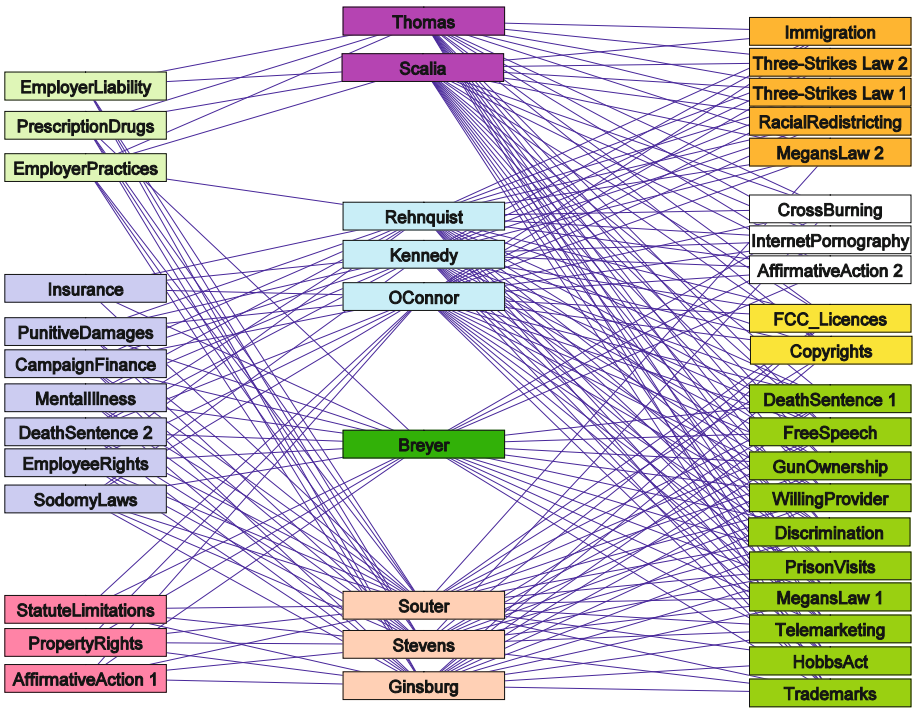

Positional Analysis and Blockmodeling, Figure 9
A partition of the Supreme Court voting for 2002–2003

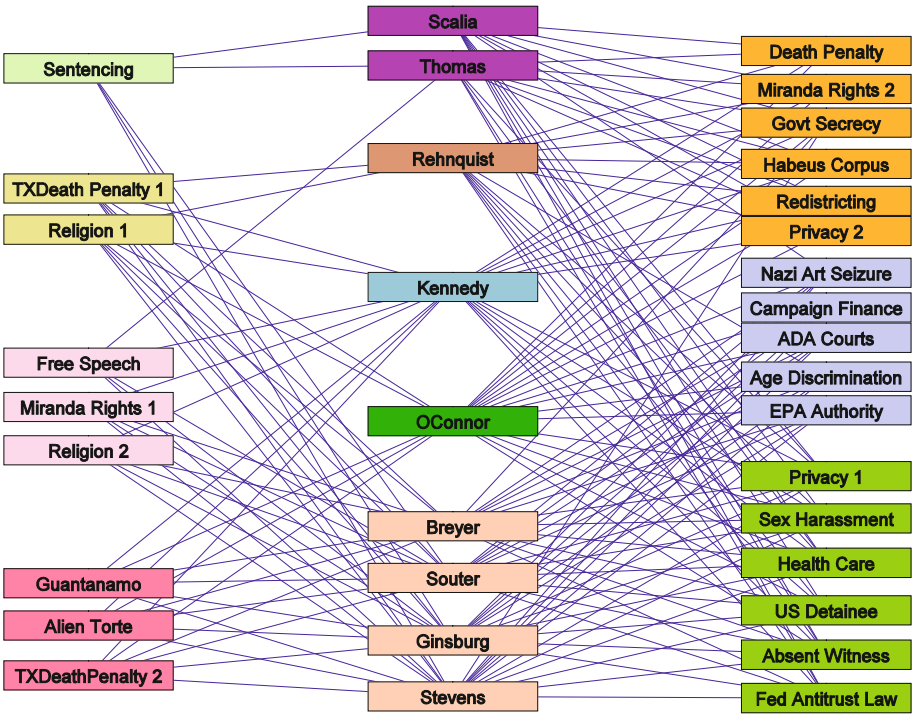

Positional Analysis and Blockmodeling, Figure 10
A partition of the Supreme Court voting for 2003–2004

column of the figure and the clusters of cases are shown on the two sides. Both sets of clusters are color coded according to cluster membership. The blue lines show the votes of the justices for their decisions. The partition of the justices shows the familiar ideological cleavage of the Supreme Court with the wing of 'conservative' justices at the top and the 'liberal' wing at the bottom.

Figures 9 and 10 show the corresponding results for the next two years of the Supreme Court's decisions based on the important cases identified in [16](b, c). There are subtle differences between the three two-mode blockmodel partitions within the same broad description of the voting patterns of the Supreme Court across the years. The methodological point here is that generalized blockmodeling can be applied to two-mode data and that the resulting partitions are both clear and interpretable.

The benefits from using a generalized blockmodeling approach include the use of precise criterion functions and an optimization approach for identifying blockmodels empirically, having a well defined measure of fit for a blockmodel, being able to define new types of blocks and new blockmodel types, being able to use blockmodeling in a deductive fashion (when our knowledge merits doing so), and being able to differentially weight the different types of inconsistencies in blocks and doing so within a very flexible, coherent and broad framework for blockmodeling. Further, in cases where the direct and indirect approaches have been compared, for both structural and regular equivalence, the generalized approach has never been outperformed by the indirect approach and most often it returns blockmodels that fit better for the same equivalence type. However, even though these benefits are clear, it does not follow that the generalized approach to blockmodeling totally dominates the conventional approach nor that the major problems have been solved.

Future Directions

One drawback to generalized blockmodeling is that the combinatorial computational burden is considerable even when a local optimization method is used. This constrains the size of networks that can be modeled, a major disadvantage. Much larger networks can be analyzed using the indirect approach and, even though attention is confined to structural and regular equivalence with it, the network size restriction for generalized blockmodeling is problematic. This issue becomes even more acute when three-mode networks are considered. For these problems, some combination of indirect methods, direct methods and graph theoretical methods will be needed. So, one open problem set stems the need to create more efficient algorithms together with the formulation of better heuristics for partitioning networks in general.

Positional analysis gives priority to network locations and network positions. Indeed, this is its hallmark. Because the whole network matters, both approaches to blockmodeling assume that the boundary of the network has been located correctly. In many cases, the assumption is appropriate but, as the ambition of network analysts expands to consider networks of (much) larger size, this assumption becomes questionable. At the moment, we simply do not know the vulnerability (or instability) of blockmodeling solutions when the boundary has been draw inaccurately. (Leaving out crucial network vertices is far more problematic than including vertices of little structural relevance). The need to learn the vulnerabilities of blockmodeling to this type of boundary problem is acute and creates a second open problem set. One data analytic response to this problem is to represent the wider network environment of the network being studied in some fashion so as to reduce the vulnerability that comes from simply ignoring the wider network environment.

Equally acute is the problem of missing network data when the boundary has been correctly identified. That is, the data may contain all of the actors but the information from and about them is incomplete, inaccurate or missing. This sets up the third set of open problems: dealing with inadequate data. The missing data part of this problem is easy to sweep under the rug by assuming that the missing relational data is the same as null relational data. In reality, the 0 in a network data set can be a null relation or that no data for the tie is available and it is folly to treat them as the same. There are three broad approaches to this problem that show promise. One is to experiment with complete extant data by systematically deleting or otherwise corrupting the data and reanalyzing the corrupted data with the same blockmodeling tools to learn more about which types of corrupted data matter the most and how much missing or corrupted data will distort the results of blockmodeling the network. A second approach is to use simulation methods to generate data where the sources of missing or incomplete data is built in the generating process and can be studied. The disadvantage to both of these approaches is that the derived knowledge may be restricted to the situations featured in the experiments and simulations. The third approach deals with the analysis of data directly by marking the missing data as such and incorporating the missing data by defining new block types defined by the missing data in them. All three approaches are suggestions for tackling the third open problem set arising from the need to consider 'bad' data seriously – far

more seriously than the current cavalier practice of assuming network data are all 'good' when positional approaches are used.

Most of the examples used in [14] feature binary networks and there is a clear need to extend generalized blockmodeling to consider valued networks, the fourth set of open problems. The balance theoretic partitioning can deal with valued signed network – as the example discussed above demonstrates. In one respect, conventional blockmodeling does this because many of the (dis)similarity measures that are used are applicable for valued networks. But for generalized blockmodeling this is a real problem. A crude way of dealing with valued networks is to impose some threshold to convert a valued network into one (or more) binary networks. The work in [22] suggests that this strategy is fraught with hazard and he has introduced a potentially fruitful approach for a generalized blockmodeling approach to valued network data.

A fifth open problem set is ushered in by the criterion function used to fit generalized blockmodels. For a given set of block types and a specific blockmodel type, we can say that the optimization approach provides empirical blockmodels that fit the best – subject to the caveat the relocation method is a local optimization procedure. If the criterion function is well defined (i. e. it is compatible with and sensitive to the equivalence used), then identifying partitions minimizing it all but guarantees the claim. And if the values of the criterion function are small (as in the examples used above) this claim is strengthened. However, this does not address the issue of how large a criterion function has to be to say that the specified blockmodel does not fit the data well even though the partition returns the lowest possible value of the criterion function *for the data at hand*. The 'best' may not be good enough. Some statistical theory is needed for fully testing the fit of blockmodels to data in a principled fashion. Assembling this constitutes a solution set for the fifth set of open problems.

Dealing with multiple relations constitutes the sixth open problem set discussed here. This is acute for both the conventional and generalized approaches to the blockmodeling problem even though, at face value, it seems more acute for generalized blockmodeling. Conventional blockmodeling handles this issue by stacking the relations and computing the (dis)similarity measure across the full set of relations and mobilizing a clustering algorithm. However, this strategy implies that every relationship has the same set of block types for each relation. The example shown in Fig. 11 demonstrates the serious drawback with this assumption.

Two relations are shown. On the left is a hierarchical relation for a hypothetical bureaucratic structure, similar to the one in Fig. 1. The relation shown is 'has authority over' and a regular equivalence is shown with the levels that are color coded according to the partition. On the right is a hypothetical social relationship of advice seeking. Suppose, further, in this hypothetical organization that there are definite norms about advice seeking. People at the same bureaucratic level are free to seek advice from each other and it is acceptable for lower ranked subordinates to see advice from their bosses. In addition, suppose that the informal norms are such that bosses seeking advice from their subordinates is unacceptable and they

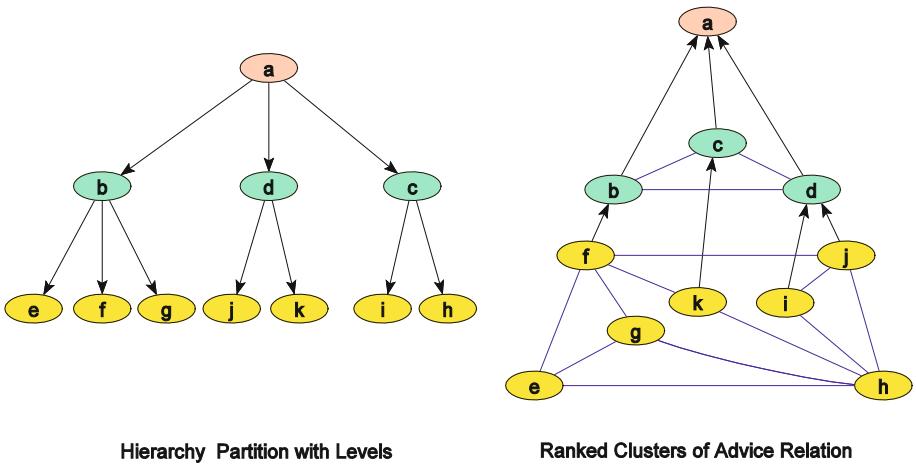

Positional Analysis and Blockmodeling, Figure 11
Two relations with the same partition but different blockmodels

never do so (or admit to doing so). The data on the right hand network are fully consistent with a ranked clusters model [12]. Within levels of the hierarchy, advice seeking, when it occurs, is fully symmetrical and the only directed ties are from subordinates at the lowest level to some of their bosses at the middle level and from the middle level to the top company official. This ranked clusters partition is exact and the clusters are identical for each relation in Fig. 11. The problem with the 'stacking' of relations approach is clear for this example. The blockmodels for each relationship are totally different even though the vertex partition for each relation is the same. Assuming, for example, that a single equivalence holds for both relations and computing some (dis)similarity will simply confound the two types of blockmodel with the result that the actual partitions will not be identified and, even if they were, they would be misrepresented in terms of the predicates of the blockmodel.

Analyzing multiple relations where the different relations can have quite different blockmodels is a deep problem, one that cannot be solved by the use of (dis)similarity measures as if they apply to each relation in the same fashion.

The seventh, and final, open problem set considered here has to do with the notion of evolving social networks. One of the basic assumptions of the positional approach, with blockmodeling as the primary strategy, is that there is a large and/or complex social network that we observe and the primary task is understand the structure of the network. This is done by delineating and describing the simpler blockmodel image. Another part of the basic assumption is that the underlying blockmodel image is the fundamental structure and the observed (surface) structure is an indicator only of the fundamental structure. There is a hint in the Supreme Court example that the voting structure did change over time. Of course, this could be due to the nature of the cases that happened to be considered in each of the terms studied. But even if there is some systematic evolution of the structure, the images depicted in Figs. 8 through 10 are no more than a description of change and does not deal with evolution as a structural process. The (deep) problem is to formulate models of network evolution for the fundamental structure(s) represented by the blockmodel while all of the data depict the observed surface model.

Blockmodeling methods, whether in the conventional or generalized mode, form a powerful set of tools for modeling, representing and understanding the structure of social networks. The many benefits stemming from the use of these tools, especially in the generalized blockmodeling variant, inspire confidence in this approach and suggest an enormous potential for using these tools. However, there are many serious open problem sets that need to be solved before this potential is fully realized.

Bibliography

1. Batagelj V (1997) Notes on blockmodeling. Soc Netw 19: 143–53
2. Batagelj V, Mrvar A (1998) Pajek – A Program for large network analysis. Connections 21:47–57
3. Batagelj V, Doreian P, Ferligoj A (1992) An optimizational approach to regular equivalence. Soc Netw 14:121–35
4. Borgatti SP, Everett MG (1989) The class of all regular equivalences: Algebraic structure and computation. Soc Netw 11:159–72
5. Borgatti SP, Everett MG (2002) Notions of position in social network analysis. In: Marsden PV (ed) Sociological Methodology. Am Soc Assoc, Washington DC, pp 1–36
6. Borgatti SP, Everett MG, Freeman LC (2002) Ucinet for Windows: Software for Social Network Analysis, Analytic Technologies. Harvard, Massachussets
7. Breiger RL, Boorman SA, Arabie P (1975) An algorithm for clustering relational data with applications to social network analysis and comparison to multidimensional scaling. J Math Psychol 12:328–83
8. Burt RS (1976) Positions in networks. Soc Forces 55:93–122
9. Cartwright D, Harary F (1956) Structural balance: A generalization of Heider's theory. Psychol Rev 63:277–292
10. Davis JA (1967) Clustering and structural balance in graphs. Hum Relat 20:181–7
11. Doreian P, Mrvar A (1996) A partitioning approach to structural balance. Soc Netw 18:149–68
12. Doreian P, Batagelj V, Ferligoj A (2000) Symmetric-acyclic decomposition of networks. J Classif 17:3–28
13. Doreian P, Batagelj V, Ferligoj A (2005) Positional analysis of sociometric data. In: Carrington PJ, Scott J, Wasserman S (eds) Models and Methods in Social Network Analysis. Cambridge University Press, New York
14. Doreian P, Batagelj V, Ferligoj A (2005) Generalized Blockmodeling. Cambridge University Press, New York
15. Gold M, Doreian P, Taylor EF (2008) Understanding a collaborative effort to reduce racial and ethnic disparities in health care: Contributions from social network analysis. Soc Sci Med 67:1018–1027
16. Greenhouse L (2001, 2003, 2004) (a) In year of Florida vote, Supreme Court did much other work. New York Times, July 2; (b) In momentous term, justices remake the law and the court. New York Times, July 1; (c) The year Rehnquist may have lost his court. New York Times, July 4
17. Heider F (1946) Attitudes and cognitive organization. J Psychol 21:107–12
18. Lemann TB, Solomon RL (1952) Group Characteristics as Revealed in Sociometric Patterns and Personality Ratings. Sociometry Monographs, vol 27. Beacon House, Boston
19. Lorrain FP, White H (1971) Structural equivalence of individuals in networks. J Math Soc 1:49–80
20. White DR (1984) REGGE: A regular graph equivalence algorithm for computing role distances prior to blockmodeling. University of California, Irvine (Unpublished manuscript)

21. White DR, Reitz KP (1983) Graph and semigroup homomorphisms on networks and relations. Soc Netw 5:193–235
22. Žiberna A (2005) Generalized blockmodeling of valued networks. Soc Netw 29:105–26

Possibility Theory

Didier Dubois, Henri Prade
IRIT-CNRS, Universite Paul Sabatier,
Toulouse Cedex, France

Article Outline

Glossary
Definition of the Subject
Introduction
Historical Background
Basic Notions of Possibility Theory
Qualitative Possibility Theory
Quantitative Possibility Theory
Probability-Possibility Transformations
Applications and Future Directions
Bibliography

Glossary

Possibility distribution A possibility distribution restricts a set of possible values for a variable of interest in an elastic way. It is represented by a mapping from a universe gathering the potential values of the variable to a scale such as the unit interval of the real line, or a finite linearly ordered set, expressing to what extent each value is possible for the variable. Thus, a possibility distribution restricts a set of more or less possible values belonging to a universe that may be also ordered such as a subpart of real line for a numerical variable, or not ordered if for instance the variable takes its value in the set of interpretations of a logical language. This may be used for representing uncertainty if the restriction pertains to possible values for an ill-known state of the world, or for representing preferences if the restriction encodes a set of values that are considered as more or less satisfactory for some purpose.

Possibility measure A possibility measure is a set function (increasing in the wide sense) that returns the maximum of a possibility distribution over a subset representing an event.

Necessity measure A necessity measure is a set function, associated by duality to a possibility measure through a relation expressing that an event is all the more necessarily true (all the more certain) as the opposite event is less possible. A necessity measure estimates to what extent the information represented by the underlying possibility distribution entails the occurrence of the event.

Guaranteed possibility A guaranteed possibility measure is a set function (decreasing in the wide sense) that returns the minimum of a possibility distribution over a subset representing an event. While possibility measures evaluate the consistency of the information between an event and the available information represented by the underlying possibility distribution, guaranteed possibility measures capture another view of the idea of possibility related to the idea of (guaranteed) feasibility, or sufficiency condition.

Possibilistic logic Standard possibilistic logic is a weighted logic where formulas are pairs made of a classical logical formula and a weight that acts as a lower bound of the necessity of the logical formula. Extended possibilistic logics may include formulas weighted in terms of lower bounds of possibility or guaranteed possibility measures.

Definition of the Subject

Possibility theory is the simplest uncertainty theory devoted to the modeling of incomplete information. It is characterized by the use of two basic dual set functions that respectively grade the possibility and the necessity of events. Possibility theory lies at the crossroads between fuzzy sets, probability and non-monotonic reasoning. Possibility theory is closely related to fuzzy sets if one considers that a possibility distribution is a particular fuzzy set (of mutually exclusive) possible values. However fuzzy sets and fuzzy logic are primarily motivated by the representation of gradual properties while possibility theory handles the uncertainty of classical (or fuzzy) propositions. Possibility theory can be cast either in an ordinal or in a numerical setting. Qualitative possibility theory is closely related to belief revision theory, and common-sense reasoning with exception-tainted knowledge in Artificial Intelligence. It has been axiomatically justified in a decision-theoretic framework in the style of Savage, thus providing a foundation for qualitative decision theory. Quantitative possibility theory is the simplest framework for statistical reasoning with imprecise probabilities. As such it has close connections with random set theory and confidence intervals, and can provide a tool for uncertainty propagation with limited statistical or subjective information.

Introduction

Possibility theory is an uncertainty theory devoted to the handling of incomplete information. To a large extent, it

is similar to probability theory because it is based on set-functions. It differs from the latter by the use of a pair of dual set functions (possibility and necessity measures) instead of only one. Besides, it is not additive and makes sense on ordinal structures. The name "Theory of Possibility" was coined by Zadeh [1], who was inspired by a paper by Gaines and Kohout [2]. In Zadeh's view, possibility distributions were meant to provide a graded semantics to natural language statements. However, possibility and necessity measures can also be the basis of a full-fledged representation of partial belief that parallels probability. It can be seen either as a coarse, non-numerical version of probability theory, or a framework for reasoning with extreme probabilities, or yet a simple approach to reasoning with imprecise probabilities [3].

After reviewing pioneering contributions to possibility theory, we recall its basic concepts and present the two main directions along which it has developed: the qualitative and quantitative settings. Both approaches share the same basic "maxitivity" axiom. They differ when it comes to conditioning, and to independence notions.

Historical Background

Zadeh was not the first scientist to speak about formalizing notions of possibility. The modalities *possible* and *necessary* have been used in philosophy at least since the Middle-Ages in Europe, based on Aristotle's works. More recently they became the building blocks of Modal Logics that emerged at the end of the first decade of the XXth century from the works of C.I. Lewis (see Hughes and Cresswell [5]). In this approach, possibility and necessity are all-or-nothing notions, and handled at the syntactic level. More recently, and independently from Zadeh's view, the notion of possibility, as opposed to probability, was central in the works of one economist, Shackle, and in those of two philosophers, D. Lewis and L.J. Cohen.

G.L.S. Shackle

A graded notion of possibility was introduced as a full-fledged approach to uncertainty and decision in the 1940–1970's by the English economist G.L.S. Shackle [6], who called *degree of potential surprise* of an event its degree of impossibility, that is, the degree of necessity of the opposite event. Shackle's notion of possibility is basically epistemic, it is a "character of the chooser's particular state of knowledge in his present". Impossibility is understood as disbelief. Potential surprise is valued on a disbelief scale, namely a positive interval of the form $[0, y^*]$, where y^* denotes the absolute rejection of the event to which it is assigned. In case everything is possible, all mutually exclusive hypotheses have zero surprise. At least one elementary hypothesis must carry zero potential surprise. The degree of surprise of an event, a set of elementary hypotheses, is the degree of surprise of its least surprising realization. The disbelief notion introduced later by Spohn [7] employs the same type of convention as potential surprise, but using the set of natural integers as a disbelief scale. Shackle also introduces a notion of conditional possibility, whereby the degree of surprise of a conjunction of two events A and B is equal to the maximum of the degree of surprise of A, and of the degree of surprise of B, should A prove true.

D. Lewis

In his 1973 book [8] the philosopher David Lewis considers a graded notion of possibility in the form of a relation between possible worlds he calls *comparative possibility*. He equates this concept of possibility to a notion of similarity between possible worlds. This non-symmetric notion of similarity is also comparative, and is meant to express statements of the form: *a world j is at least as similar to world i as world k is*. Comparative similarity of j and k with respect to i is interpreted as the comparative possibility of j with respect to k viewed from world i. Such relations are assumed to be complete pre-orderings and are instrumental in defining the truth conditions of counterfactual statements. Comparative possibility relations $\geq_\Pi$ obey the key axiom: for all events A, B, C,

$$A \geq_\Pi B \text{ implies } C \cup A \geq_\Pi C \cup B\,.$$

This axiom was later independently proposed by the first author [9] in an attempt to derive a possibilistic counterpart to comparative probabilities. Interestingly, the connection between numerical possibility and similarity is currently investigated by Sudkamp [10].

L.J. Cohen

A framework very similar to the one of Shackle was proposed by the philosopher L.J. Cohen [11] who considered the problem of legal reasoning. He introduced so-called *Baconian probabilities* understood as degrees of provability. The idea is that it is hard to prove someone guilty at the court of law by means of pure statistical arguments. The basic feature of degrees of provability is that a hypothesis and its negation cannot both be provable together to any extent (the contrary being a case for inconsistency). Such degrees of provability coincide with necessity measures.

L.A. Zadeh

In his seminal paper [1] Zadeh proposed an interpretation of membership functions of fuzzy sets as possibility distributions encoding flexible constraints induced by natural language statements. Zadeh articulated the relationship between possibility and probability, noticing that what is probable must preliminarily be possible. However, the view of possibility degrees developed in his paper refers to the idea of graded feasibility (degrees of ease, as in the example of "how many eggs can Hans eat for his breakfast") rather than to the epistemic notion of plausibility laid bare by Shackle. Nevertheless, the key axiom of "maxitivity" for possibility measures is highlighted. In two subsequent articles [12,13], Zadeh acknowledged the connection between possibility theory, belief functions and upper/lower probabilities, and proposed their extensions to fuzzy events and fuzzy information granules.

Basic Notions of Possibility Theory

The basic building blocks of possibility theory were first extensively described in the authors' books [14,15] (see also [16]). Let S be a set of states of affairs (or descriptions thereof), or states for short. A possibility distribution is a mapping π from S to a totally ordered scale L, with top 1 and bottom 0, such as the unit interval. The function π represents the state of knowledge of an agent (about the actual state of affairs) distinguishing what is plausible from what is less plausible, what is the normal course of things from what is not, what is surprising from what is expected. It represents a flexible restriction on what is the actual state with the following conventions (similar to probability, but opposite to Shackle's potential surprise scale):

- $\pi(s) = 0$ means that state s is rejected as impossible;
- $\pi(s) = 1$ means that state s is totally possible (= plausible).

If S is exhaustive, at least one of the elements of S should be the actual world, so that $\exists s, \pi(s) = 1$ (normalization). Distinct values may simultaneously have a degree of possibility equal to 1.

Possibility theory is driven by the principle of minimal specificity. It states that any hypothesis not known to be impossible cannot be ruled out. A possibility distribution π is said to be at least as specific as another π' if and only if for each state of affairs s: $\pi(s) \leq \pi'(s)$ (Yager [17]). Then, π is at least as restrictive and informative as π'.

In the possibilistic framework, extreme forms of partial knowledge can be captured, namely:

- Complete knowledge: for some $s_0, \pi(s_0) = 1$ and $\pi(s) = 0, \forall s \neq s_0$ (only s_0 is possible)
- Complete ignorance: $\pi(s) = 1, \forall s \in S$, (all states are possible).

Given a simple query of the form "does event A occur?" where A is a subset of states, the response to the query can be obtained by computing degrees of possibility and necessity, respectively (if the possibility scale $L = [0, 1]$):

$$\Pi(A) = \sup_{s \in A} \pi(s); N(A) = \inf_{s \notin A} 1 - \pi(s) .$$

$\Pi(A)$ evaluates to what extent A is consistent with π, while $N(A)$ evaluates to what extent A is certainly implied by π. The possibility-necessity duality is expressed by $N(A) = 1 - \Pi(A^c)$, where A^c is the complement of A. Generally, $\Pi(S) = N(S) = 1$ and $\Pi(\emptyset) = N(\emptyset) = 0$. Possibility measures satisfy the basic "maxitivity" property $\Pi(A \cup B) = \max(\Pi(A), \Pi(B))$. Necessity measures satisfy an axiom dual to that of possibility measures, namely $N(A \cap B) = \min(N(A), N(B))$. On infinite spaces, these axioms must hold for infinite families of sets.

Human knowledge is often expressed in a declarative way using statements to which belief degrees are attached. It corresponds to expressing constraints the world is supposed to comply with. Certainty-qualified pieces of uncertain information of the form "A is certain to degree α" can then be modelled by the constraint $N(A) \geq \alpha$. The least specific possibility distribution reflecting this information is [15]:

$$\pi_{(A,\alpha)}(s) = \begin{cases} 1 , & \text{if } s \in A \\ 1 - \alpha , & \text{otherwise} . \end{cases} \quad (1)$$

Acquiring further pieces of knowledge leads to updating $\pi_{(A,\alpha)}$ into some $\pi < \pi_{(A,\alpha)}$.

Apart from Π and N, a measure of *guaranteed possibility* or *sufficiency* can be defined [18]: $\Delta(A) = \inf_{s \in A} \pi(s)$. In contrast, Π appears to be a measure of *potential* possibility. It estimates to what extent *all* states in A are actually possible according to evidence. $\Delta(A)$ can be used as a degree of evidential support for A. Uncertain statements of the form "A is possible to degree β" often mean that all realizations of A are possible to degree β. They can then be modelled by the constraint $\Delta(A) \geq \beta$. It corresponds to the idea of observed evidence. This type of information is better exploited by assuming an informational principle opposite to the one of minimal specificity, namely, any situation not yet observed is tentatively considered as impossible. This is similar to closed-world assumption. The most specific distribution $\delta_{(A,\beta)}$ in agreement with $\Delta(A) \geq \beta$ is:

$$\delta_{(A,\beta)}(s) = \begin{cases} \beta , & \text{if } s \in A \\ 0 , & \text{otherwise} . \end{cases}$$

Acquiring further pieces of evidence leads to updating $\delta_{(A,\beta)}$ into some wider distribution $\delta > \delta_{(A,\beta)}$. Such evidential support functions do not behave with the same conventions as possibility distributions: $\delta(s) = 1$ means that S is guaranteed to be possible, because of a high evidential support, while $\delta(s) = 0$ only means that S has not been observed yet (hence is of unknown possibility). Distributions δ are generally not normalized to 1, and serve as lower bounds to possibility distributions π (because what is observed must be possible). Such a bipolar representation of information using pairs (δ, π) may provide a natural interpretation of interval-valued fuzzy sets. Note that possibility distributions induced from certainty-qualified pieces of knowledge combine conjunctively, by discarding possible states, while evidential support distributions induced by possibility-qualified pieces of evidence combine disjunctively, by accumulating possible states.

Notions of conditioning and independence were studied for possibility measures. Conditional possibility is defined similarly to probability theory using a Bayesian like equation of the form [15]

$$\Pi(B \cap A) = \Pi(B|A) \star \Pi(A) .$$

However, in the ordinal setting the operation $\star$ cannot be a product and is changed into the minimum. In the numerical setting, there are several ways to define conditioning, not all of which have this form [19]. There are several variants of possibilistic independence [20,21,22]. Generally, independence in possibility theory is neither symmetric, nor insensitive to negation. For non Boolean variables, independence between events is not equivalent to independence between variables. Joint possibility distributions on Cartesian products of domains can be represented by means of graphical structures similar to Bayesian networks for joint probabilities (see [23,24]. Such graphical structures can be taken advantage of for evidence propagation [25] or learning [26].

Qualitative Possibility Theory

This section is restricted to the case of a finite state space S, supposed to be the set of interpretations of a formal propositional language. In other words, S is the universe induced by Boolean attributes. A plausibility ordering is a complete pre-order of states denoted by $\geq_\pi$, which induces a well-ordered partition $\{E_1, \ldots, E_n\}$ of S. It is the comparative counterpart of a possibility distribution π, i. e., $s \geq_\pi s'$ if and only if $\pi(s) \geq \pi(s')$. Indeed it is more natural to expect that an agent will supply ordinal rather than numerical information about his beliefs. By convention E_1 contains the most normal states of fact, E_n the least plausible, or most surprising ones. Denoting argmax(A) any most plausible state $s_0 \in A$, ordinal counterparts of possibility and necessity measures [9] are then defined as follows: $\{s\} \geq_\Pi \emptyset$ for all $s \in S$ and

$$A \geq_\Pi B \text{ if and only if } \max(A) \geq_\pi \max(B)$$
$$A \geq_N B \text{ if and only if } \max(B^c) \geq_\pi \max(A^c) .$$

Possibility relations $\geq_\Pi$ are those of Lewis and satisfy the characteristic property

$$A \geq_\Pi B \text{ implies } C \cup A \geq_\Pi C \cup B$$

while necessity relations can also be defined as $A \geq_N B$ if and only if $B^c \geq_\Pi A^c$, and satisfy a similar axiom:

$$A \geq_N B \text{ implies } C \cap A \geq_N C \cap B .$$

The latter coincide with epistemic entrenchment relations in the sense of belief revision theory [27,28]. Conditioning a possibility relation $\geq_\Pi$ by an non-impossible event $C >_\Pi \emptyset$ means deriving a relation $\geq_\Pi^C$ such that

$$A \geq_\Pi^C B \text{ if and only if } A \cap C \geq_\Pi B \cap C .$$

The notion of independence for comparative possibility theory was studied in Dubois et al. [22], for independence between events, and Ben Amor et al. [29] between variables.

Non-monotonic Inference

Suppose S is equipped with a plausibility ordering. The main idea behind qualitative possibility theory is that the state of the world is always believed to be as normal as possible, neglecting less normal states. $A \geq_\Pi B$ really means that there is a normal state where A holds that is at least as normal as any normal state where B holds. The dual case $A \geq_N B$ is intuitively understood as "A is at least as certain as B", in the sense that there are states where B fails to hold that are at least as normal as the most normal state where A does not hold. In particular, the events accepted as true are those which are true in all the most plausible states, namely the ones such that $A >_N \emptyset$. These assumptions lead us to interpret the plausible inference $A| \approx B$ of a proposition B from another A, under a state of knowledge $\geq_\Pi$ as follows: *B should be true in all the most normal states were A is true*, which means $B >_\Pi^A B^c$ in terms of ordinal conditioning, that is, $A \cap B$ is more plausible than $A \cap B^c$. $A| \approx B$ also means that the agent considers B as an accepted belief in the context A.

This kind of inference is non-monotonic in the sense that $A| \approx B$ does not always imply $A \cap C| \approx B$ for any

additional information C. This is similar to the fact that a conditional probability $P(B|A \cap C)$ may be low even if $P(B|A)$ is high. The properties of the consequence relation $|\approx$ are now well-understood, and are precisely the ones laid bare by Lehmann and Magidor [30] for their so-called "rational inference". Monotonicity is only partially restored: $A| \approx B$ implies $A \cap C| \approx B$ holds provided that $A| \approx C^c$ does not hold (i. e. that states were A is true do not typically violate C). This property is called *rational monotony*, and, along with some more standard ones (like closure under conjunction), characterizes default possibilistic inference $|\approx$. In fact, the set $\{B, A| \approx B\}$ of accepted beliefs in the context A is deductively closed, which corresponds to the idea that the agent reasons with accepted beliefs in each context as if they were true, until some event occurs that modifies this context. This closure property is enough to justify a possibilistic approach [31] and adding the rational monotonicity property ensures the existence of a single possibility relation generating the consequence relation $|\approx$ [32].

Rather than being constructed from scratch, plausibility orderings can be generated by a set of if-then rules tainted with unspecified exceptions. This set forms a knowledge base supplied by an agent. Each rule "if A then B" is understood as a constraint of the form $A \cap B >_{\Pi} A \cap B^c$ on possibility relations. There exists a single minimally specific element in the set of possibility relations satisfying all constraints induced by rules (unless the latter are inconsistent). It corresponds to the most compact plausibility ranking of states induced by the rules [32]. This ranking can be computed by an algorithm originally proposed by Pearl [33].

Possibilistic Logic

Qualitative possibility relations can be represented by (and only by) possibility measures ranging on any totally ordered set L (especially a finite one) [9]. This absolute representation on an ordinal scale is slightly more expressive than the purely relational one. When the finite set S is large and generated by a propositional language, qualitative possibility distributions can be efficiently encoded in possibilistic logic [34]. A possibilistic logic base K is a set of pairs (ϕ, α), where ϕ is a Boolean expression and α is an element of L. This pair encodes the constraint $N(\phi) \geq \alpha$ where $N(\phi)$ is the degree of necessity of the set of models of ϕ. Each prioritized formula (ϕ, α) has a fuzzy set of models (described in Sect. "Basic Notions of Possibility Theory") and the fuzzy intersection of the fuzzy sets of models of all prioritized formulas in K yields the associated plausibility ordering on S.

Syntactic deduction from a set of prioritized clauses is achieved by refutation using an extension of the standard resolution rule, whereby $(\phi \vee \psi, \min(\alpha, \beta))$ can be derived from $(\phi \vee \xi, \alpha)$ and $(\psi \vee \neg\xi, \beta)$. This rule, which evaluates the validity of an inferred proposition by the validity of the weakest premiss, goes back to Theophrastus, a disciple of Aristotle. Possibilistic logic is an inconsistency-tolerant extension of propositional logic that provides a natural semantic setting for mechanizing nonmonotonic reasoning [35], with a computational complexity close to that of propositional logic. See [36] for a detailed introduction to possibilistic logic, its syntactic and semantic aspects, different extensions that involve time, for instance, or that encode constraints on lower bounds of possibility or guaranteed possibility measures, or that perform multiple source information fusion. See [37] for a sketch of further extensions for handling groups of agents' beliefs and mutual beliefs.

Another compact representation of qualitative possibility distributions is the possibilistic directed graph, which uses the same conventions as Bayesian nets, but relies on an ordinal notion of conditional possibility [15]

$$\Pi(B|A) = \begin{cases} 1\,, & \text{if } \Pi(B \cap A) = \Pi(A) \\ \Pi(B \cap A)\,, & \text{otherwise}\,. \end{cases}$$

Joint possibility distributions can be decomposed into a conjunction of conditional possibility distributions (using minimum) in a way similar to Bayes nets [38]. It is based on a symmetric notion of qualitative independence $\Pi(B \cap A) = \min(\Pi(A), \Pi(B))$ that is weaker than the causal-like condition $\Pi(B|A) = \Pi(B)$ [22]. Ben Amor and Benferhat [39] investigate the properties of qualitative independence that enable local inferences to be performed in possibilistic nets.

Decision-Theoretic Foundations

Zadeh [1] hinted that "since our intuition concerning the behavior of possibilities is not very reliable", our understanding of them "would be enhanced by the development of an axiomatic approach to the definition of subjective possibilities in the spirit of axiomatic approaches to the definition of subjective probabilities". Decision-theoretic justifications of qualitative possibility were recently devised, in the style of Savage [40]. On top of the set of states, assume there is a set X of consequences of decisions. A decision, or act, is modelled as a mapping f from S to X assigning to each state S its consequence $f(s)$. The axiomatic approach consists in proposing properties of a preference relation $\succeq$ between acts so that a representation of this

relation by means of a preference functional $W(f)$ is ensured, that is, act f is as good as act g (denoted $f \succeq g$) if and only if $W(f) \geq W(g)$. $W(f)$ depends on the agent's knowledge about the state of affairs, here supposed to be a possibility distribution π on S, and the agent's goal, modelled by a utility function u on X. Both the utility function and the possibility distribution map to the same finite chain L. A pessimistic criterion $W_\pi^-(f)$ is of the form:

$$W_\pi^-(f) = \min_{s \in S} \max(n(\pi(s)), u(f(s)))$$

where n is the order-reversing map of L. $n(\pi(s))$ is the degree of certainty that the state is not s (hence the degree of surprise of observing s), $u(f(s))$ the utility of choosing act f in state s. $W_\pi^-(f)$ is all the higher as all states are either very surprising or have high utility. This criterion is actually a prioritized extension of the Wald maximin criterion. The latter is recovered if $\pi(s) = 1$ (top of L) $\forall s \in S$. According to the pessimistic criterion, acts are chosen according to their worst consequences, restricted to the most plausible states $S^* = \{s, \pi(s) \geq n(W_\pi^-(f))\}$. The optimistic counterpart of this criterion is:

$$W_\pi^+(f) = \max_{s \in S} \min(\pi(s)), u(f(s)))\ .$$

$W_\pi^+(f)$ is all the higher as there is a very plausible state with high utility. The optimistic criterion was first proposed by Yager [41] and the pessimistic criterion by Whalen [42]. These optimistic and pessimistic possibilistic criteria are particular cases of a more general criterion based on the Sugeno integral [43] specialized to possibility and necessity of fuzzy events [1,14]:

$$S_{\gamma,u}(f) = \max_{\lambda \in L} \min(\lambda, \gamma(F_\lambda))$$

where $F_\lambda = \{s \in S, u(f(s)) \geq \lambda\}$, γ is a monotonic set function that reflects the decision-maker attitude in front of uncertainty: $\gamma(A)$ is the degree of confidence in event A. If $\gamma = \Pi$, then $S_{\Pi,u}(f) = W_\pi^+(f)$. Similarly, if $\gamma = N$, then $S_{N,u}(f) = W_\pi^-(f)$.

For any acts f, g, and any event A, let fAg denote an act consisting of choosing f if A occurs and g if its complement occurs. Let $f \wedge g$ (resp. $f \vee g$) be the act whose results yield the worst (resp. best) consequence of the two acts in each state. Constant acts are those whose consequence is fixed regardless of the state. A result in [44,45] provides an act-driven axiomatization of these criteria, and enforces possibility theory as a "rational" representation of uncertainty for a finite state space S:

Theorem 1 *Suppose the preference relation $\succeq$ on acts obeys the following properties:*

1. *$(X^S, \succeq)$ is a complete preorder.*
2. *There are two acts such that $f \succ g$.*
3. *$\forall A, \forall g$ and h constant, $\forall f, g \succeq h$ implies $gAf \succeq hAf$.*
4. *If f is constant, $f \succ h$ and $g \succ h$ imply $f \wedge g \succ h$.*
5. *If f is constant, $h \succ f$ and $h \succ g$ imply $h \succ f \vee g$.*

Then there exists a finite chain L, an L-valued monotonic set-function γ on S and an L-valued utility function u, such that $\succeq$ is representable by a Sugeno integral of $u(f)$ with respect to γ. Moreover γ is a necessity (resp. possibility) measure as soon as property (iv) (resp. (v)) holds for all acts. The preference functional is then $W_\pi^-(f)$ (resp. $W_\pi^+(f)$).

Axioms (4–5) contradict expected utility theory. They become reasonable if the value scale is finite, decisions are one-shot (no compensation) and provided that there is a big step between any level in the qualitative value scale and the adjacent ones. In other words the preference pattern $f \succ h$ always means that f is significantly preferred to h, to the point of considering the value of h negligible in front of the value of f. The above result provides decision-theoretic foundations of possibility theory, whose axioms can thus be tested from observing the choice behavior of agents. See [46] for another approach to comparative possibility relations, more closely relying on Savage axioms but giving up any comparability between utility and plausibility levels. The drawback of these and other qualitative decision criteria is their lack of discrimination power [47]. To overcome it, refinements of possibilistic criteria were recently proposed, based on lexicographic schemes [48]. These new criteria turn out to be representable by a classical (but big-stepped) expected utility criterion.

Quantitative Possibility Theory

The phrase "quantitative possibility" refers to the case when possibility degrees range in the unit interval. In that case, a precise articulation between possibility and probability theories is useful to provide an interpretation to possibility and necessity degrees. Several such interpretations can be consistently devised. See [49] for a detailed survey. A degree of possibility can be viewed as an upper probability bound [50], and a possibility distribution can be viewed as a likelihood function [51]. A possibility measure is also a special case of a Shafer plausibility function [52]. Following a very different approach, possibility theory can account for probability distributions with extreme values, infinitesimal [7] or having big steps [53]. There are finally close connections between possibility theory and idempotent analysis [54,55]. The theory of large deviations in probability theory [56] also handles set-functions that look

like possibility measures [57]. Here we focus on the role of possibility theory in the theory of imprecise probability.

Possibility as Upper Probability

Let π be a possibility distribution where $\pi(s) \in [0,1]$. Let $\mathbf{P}(\pi)$ be the set of probability measures P such that $P \leq \Pi$, i.e. $\forall A \subseteq S, P(A) \leq \Pi(A)$. Then the possibility measure Π coincides with the upper probability function P^* such that $P^*(A) = \sup\{P(A), P \in \mathbf{P}(\pi)\}$ while the necessity measure N is the lower probability function P_* such that $P_*(A) = \inf\{P(A), P \in \mathbf{P}(\pi)\}$; see [50,58] for details. P and π are said to be consistent if $P \in \mathbf{P}(\pi)$. The connection between possibility measures and imprecise probabilistic reasoning is especially promising for the efficient representation of non-parametric families of probability functions, and it makes sense even in the scope of modelling linguistic information [59].

A possibility measure can be computed from a set of nested confidence subsets $\{A_1, A_2, \ldots, A_m\}$ where $A_i \subset A_{i+1}, i = 1 \ldots m-1$. Each confidence subset A_i is attached a positive confidence level λ_i interpreted as a lower bound of $P(A_i)$, hence a necessity degree. It is viewed as a certainty-qualified statement that generates a possibility distribution π_i according to Sect. "Basic Notions of Possibility Theory". The corresponding possibility distribution is

$$\pi(s) = \min_{i=1,\ldots,m} \pi_i(s) = \begin{cases} 1\,, & \text{if } u \in A_1 \\ 1-\lambda_{j-1}\,, & \text{if } j = \max\{i : s \notin A_i\} > 1\,. \end{cases}$$

The information modelled by π can also be viewed as a nested random set $\{(A_i, \nu_i), i = 1, \ldots, m\}$, where $\nu_i = \lambda_i - \lambda_{i-1}$. This framework allows for imprecision (reflected by the size of the A_i's) and uncertainty (the ν_i's). And ν_i is the probability that the agent only knows that A_i contains the actual state (it is not $P(A_i)$). The random set view of possibility theory is well adapted to the idea of imprecise statistical data, as developed in [60,61]. Namely, given a bunch of imprecise (not necessarily nested) observations (called focal sets), π supplies an approximate representation of the data, as $\pi(s) = \sum_{i:\, s \in A_i} \nu_i$.

The set $\mathbf{P}(\pi)$ contains many probability distributions, arguably too many. Neumaier [62] has recently proposed a related framework, in a different terminology, for representing smaller subsets of probability measures using two possibility distributions instead of one. He basically uses a pair of distributions (δ, π) (in the sense of Sect. "Basic Notions of Possibility Theory") of distributions, he calls "cloud", where δ is a guaranteed possibility distribution (in our terminology) such that $\pi \geq \delta$. A cloud models the (generally non-empty) set $\mathbf{P}(\pi) \cap \mathbf{P}(1-\delta)$, viewing $1-\delta$ as a standard possibility distribution.

Conditioning

There are two kinds of conditioning that can be envisaged upon the arrival of new information E. The first method presupposes that the new information alters the possibility distribution π by declaring all states outside E impossible. The conditional measure $\pi(.|E)$ is such that $\Pi(B|E) \cdot \Pi(E) = \Pi(B \cap E)$. This is formally Dempster rule of conditioning of belief functions, specialized to possibility measures. The conditional possibility distribution representing the weighted set of confidence intervals is,

$$\pi(s|E) = \begin{cases} \frac{\pi(s)}{\Pi(E)}\,, & \text{if } s \in E \\ 0\,, & \text{otherwise}\,. \end{cases}$$

De Baets et al. [63] provide a mathematical justification of this notion in an infinite setting, as opposed to the min-based conditioning of qualitative possibility theory. Indeed, the maxitivity axiom extended to the infinite setting is not preserved by the min-based conditioning. The product-based conditioning leads to a notion of independence of the form $\Pi(B \cap E) = \Pi(B) \cdot \Pi(E)$ whose properties are very similar to the ones of probabilistic independence [21].

Another form of conditioning [64,65], more in line with the Bayesian tradition, considers that the possibility distribution π encodes imprecise statistical information, and event E only reflects a feature of the current situation, not of the state in general. Then the value $\Pi(B\|E) = \sup\{P(B|E), P(E) > 0, P \leq \Pi\}$ is the result of performing a sensitivity analysis of the usual conditional probability over $\mathbf{P}(\pi)$ (Walley [66]). Interestingly, the resulting set-function is again a possibility measure, with distribution

$$\pi(s\|E) = \begin{cases} \max\left(\pi(s), \frac{\pi(s)}{\pi(s)+N(E)}\right)\,, & \text{if } s \in E \\ 0\,, & \text{otherwise}\,. \end{cases}$$

It is generally less specific than π on E, as clear from the above expression, and becomes non-informative when $N(E) = 0$ (i.e. if there is no information about E). This is because $\pi(\cdot\|E)$ is obtained from the focusing of the generic information π over the reference class E. On the contrary, $\pi(\cdot|E)$ operates a revision process on π due to additional knowledge asserting that states outside E are impossible. See De Cooman [65] for a detailed study of this form of conditioning.

Probability-Possibility Transformations

The problem of transforming a possibility distribution into a probability distribution and conversely is meaningful in the scope of uncertainty combination with heterogeneous sources (some supplying statistical data, other linguistic data, for instance). It is useful to cast all pieces of information in the same framework. The basic requirement is to respect the consistency principle $\Pi \geq P$. The problem is then either to pick a probability measure in $\mathbf{P}(\pi)$, or to construct a possibility measure dominating P.

There are two basic approaches to possibility/probability transformations, which both respect a form of probability-possibility consistency. One, due to Klir [67,68] is based on a principle of information invariance, the other [69] is based on optimizing information content. Klir assumes that possibilistic and probabilistic information measures are commensurate. Namely, the choice between possibility and probability is then a mere matter of translation between languages "neither of which is weaker or stronger than the other" (quoting Klir and Parviz [70]). It suggests that entropy and imprecision capture the same facet of uncertainty, albeit in different guises. The other approach, recalled here, considers that going from possibility to probability leads to increase the precision of the considered representation (as we go from a family of nested sets to a random element), while going the other way around means a loss of specificity.

From Possibility to Probability

The most basic example of transformation from possibility to probability is the Laplace principle of insufficient reason claiming that what is equally possible should be considered as equally probable. A generalized Laplacean indifference principle is then adopted in the general case of a possibility distribution π: the weights ν_i bearing the sets A_i from the nested family of levels cuts of π are uniformly distributed on the elements of these cuts A_i. Let P_i be the uniform probability measure on A_i. The resulting probability measure is $P = \sum_{i=1,\ldots,m} \nu_i \cdot P_i$. This transformation, already proposed in 1982 [71] comes down to selecting the center of gravity of the set $\mathbf{P}(\pi)$ of probability distributions dominated by π. This transformation also coincides with Smets' pignistic transformation [72] and with the Shapley value of the "unamimity game" (another name of the necessity measure) in game theory. The rationale behind this transformation is to minimize arbitrariness by preserving the symmetry properties of the representation. This transformation from possibility to probability is one-to-one. Note that the definition of this transformation does not use the nestedness property of cuts of the possibility distribution. It applies all the same to non-nested random sets (or belief functions) defined by pairs $\{(A_i, \nu_i), i = 1, \ldots, m\}$, where ν_i are non-negative reals such that $\sum_{i=1,\ldots,m} \nu_i = 1$.

From Objective Probability to Possibility

From probability to possibility, the rationale of the transformation is not the same according to whether the probability distribution we start with is subjective or objective [73]. In the case of a statistically induced probability distribution, the rationale is to preserve as much information as possible. This is in line with the handling of Δ-qualified pieces of information representing observed evidence, considered in Sect. "Basic Notions of Possibility Theory"; hence we select as the result of the transformation of a probability measure P, the most specific possibility measure in the set of those dominating P [69]. This most specific element is generally unique if P induces a linear ordering on S. Suppose S is a finite set. The idea is to let $\Pi(A) = P(A)$, for these sets A having minimal probability among other sets having the same cardinality as A. If $p_1 > p_2 > \ldots > p_n$, then $\Pi(A) = P(A)$ for sets A of the form $\{s_i, \ldots, s_n\}$, and the possibility distribution is defined as $\pi_P(s_i) = \sum_{j=i,\ldots,m} p_j$. Note that π_P is a kind of cumulative distribution of P. If there are equiprobable elements, the unicity of the transformation is preserved if equipossibility of the corresponding elements is enforced. In this case it is a bijective transformation as well. Recently, this transformation was used to prove a rather surprising agreement between probabilistic indeterminateness as measured by Shannon entropy, and possibilistic non-specificity. Namely it is possible to compare probability measures on finite sets in terms of their relative *peakedness* (a concept adapted from Birnbaum [74]) by comparing the relative specificity of their possibilistic transforms. Namely let P and Q be two probability measures on S and π_P, π_Q the possibility distributions induced by our transformation. It can be proved that if $\pi_P \geq \pi_Q$ (i. e. P is less peaked than Q) then the Shannon entropy of P is higher than the one of Q [75]. This result give some grounds to the intuitions developed by Klir [67], without assuming any commensurability between entropy and specificity indices.

Possibility Distributions Induced by Prediction Intervals

In the continuous case, moving from objective probability to possibility means adopting a representation of uncertainty in terms of prediction intervals around the mode viewed as the "most frequent value". Extracting a predic-

tion interval from a probability distribution or devising a probabilistic inequality can be viewed as moving from a probabilistic to a possibilistic representation. Namely suppose a non-atomic probability measure P on the real line, with unimodal density p, and suppose one wishes to represent it by an interval I with a prescribed level of confidence $P(I) = \gamma$ of hitting it. The most natural choice is the most precise interval ensuring this level of confidence. It can be proved that this interval is of the form of a cut of the density, i. e. $I_\gamma = \{s, p(s) \geq \theta\}$ for some threshold θ. Moving the degree of confidence from 0 to 1 yields a nested family of prediction intervals that form a possibility distribution π consistent with P, the most specific one actually, having the same support and the same mode as P and defined by ([69]):

$$\pi(\inf I_\gamma) = \pi(\sup I_\gamma) = 1 - \gamma = 1 - P(I_\gamma) \,.$$

This kind of transformation again yields a kind of cumulative distribution according to the ordering induced by the density p. Similar constructs can be found in the statistical literature (Birnbaum [74]). More recently Mauris et al. [76] noticed that starting from any family of nested sets around some characteristic point (the mean, the median,...), the above equation yields a possibility measure dominating P. Well-known inequalities of probability theory, such as those of Chebyshev and Camp–Meidel, can also be viewed as possibilistic approximations of probability functions. It turns out that for symmetric uni-modal densities, each side of the optimal possibilistic transform is a convex function. Given such a probability density on a bounded interval $[a, b]$, the triangular fuzzy number whose core is the mode of p and the support is $[a, b]$ is thus a possibility distribution dominating P regardless of its shape (and the tightest such distribution). These results justify the use of symmetric triangular fuzzy numbers as fuzzy counterparts to uniform probability distributions. They provide much tighter probability bounds than Chebyshev and Camp–Meidel inequalities for symmetric densities with bounded support. This setting is adapted to the modelling of sensor measurements [77]. These results are extended to more general distributions by Baudrit et al., [78], and provide a tool for representing poor probabilistic information.

Subjective Possibility Distributions

The case of a subjective probability distribution is different. Indeed, the probability function is then supplied by an agent who is in some sense forced to express beliefs in this form due to rationality constraints, and the setting of exchangeable bets. However his actual knowledge may be far from justifying the use of a single well-defined probability distribution. For instance in case of total ignorance about some value, apart from its belonging to an interval, the framework of exchangeable bets enforces a uniform probability distribution, on behalf of the principle of insufficient reason. Based on the setting of exchangeable bets, it is possible to define a subjectivist view of numerical possibility theory, that differs from the proposal of Walley [66]. The approach developed by Dubois Prade and Smets [79] relies on the assumption that when an agent constructs a probability measure by assigning prices to lotteries, this probability measure is actually induced by a belief function representing the agents actual state of knowledge. We assume that going from an underlying belief function to an elicited probability measure is achieved by means of the above mentioned pignistic transformation, changing focal sets into uniform probability distributions. The task is to reconstruct this underlying belief function under a minimal commitment assumption. In the paper [79], we pose and solve the problem of finding the least informative belief function having a given pignistic probability. We prove that it is unique and consonant, thus induced by a possibility distribution. This result exploits a simple partial ordering between belief functions comparing their information content, in agreement with the expected cardinality of random sets. The obtained possibility distribution can be defined as the converse of the pignistic transformation (which is one-to-one for possibility distributions). It is subjective in the same sense as in the subjectivist school in probability theory. However, it is the least biased representation of the agents state of knowledge compatible with the observed betting behavior. In particular it is less specific than the one constructed from the prediction intervals of an objective probability. This transformation was first proposed in [80] for objective probability, interpreting the empirical necessity of an event as summing the excess of probabilities of realizations of this event with respect to the probability of the most likely realization of the opposite event.

Possibility Theory and Defuzzification

Possibilistic mean values can be defined using Choquet integrals with respect to possibility and necessity measures [65,81], and come close to defuzzification methods [82]. A fuzzy interval is a fuzzy set of reals whose membership function is unimodal and upper-semi continuous. Its α-cuts are closed intervals. Interpreting a fuzzy interval M, associated to a possibility distribution μ_M, as a family of probabilities, upper and lower mean values $E^*(M)$

and $E_*(M)$, can be defined as [83]:

$$E_*(M) = \int_0^1 \inf M_\alpha \, d\alpha; \quad E^*(M) = \int_0^1 \sup M_\alpha \, d\alpha$$

where M_α is the α-cut of M.

Then the mean interval $E(M) = [E_*(M), E^*(M)]$ of M is the interval containing the mean values of all random variables consistent with M, that is $E(M) = \{E(P) | P \in \mathbf{P}(\mu_M)\}$, where $E(P)$ represents the expected value associated to the probability measure P. That the "mean value" of a fuzzy interval is an interval seems to be intuitively satisfactory. Particularly the mean interval of a (regular) interval $[a, b]$ is this interval itself. The upper and lower mean values are linear with respect to the addition of fuzzy numbers. Define the addition $M + N$ as the fuzzy interval whose cuts are $M_\alpha + N_\alpha = \{s + t, s \in M_\alpha, t \in N_\alpha\}$ defined according to the rules of interval analysis. Then $E(M + N) = E(M) + E(N)$, and similarly for the scalar multiplication $E(aM) = aE(M)$, where aM has membership grades of the form $\mu_M(s/a)$ for $a \neq 0$. In view of this property, it seems that the most natural defuzzication method is the middle point $\hat{E}(M)$ of the mean interval (originally proposed by Yager [84]). Other defuzzification techniques do not generally possess this kind of linearity property. $\hat{E}(M)$ has a natural interpretation in terms of simulation of a fuzzy variable [85], and is the mean value of the pignistic transformation of M. Indeed it is the mean value of the empirical probability distribution obtained by the random process defined by picking an element α in the unit interval at random, and then an element s in the cut M_α at random.

Applications and Future Directions

Possibility theory has not been the main framework for engineering applications of fuzzy sets in the past. However, on the basis of its connections to symbolic artificial intelligence, to decision theory and to imprecise statistics, we consider that it has significant potential for further applied developments in a number of areas, including some where fuzzy sets are not yet al.ways accepted. Only some directions are pointed out here.

1. Rules with exceptions can be modelled by means of conditional possibility [32], based on its capability to account for non-monotonic inference, as shown in Sect. "Non-monotonic Inference". Possibility theory has also enabled a typology of fuzzy rules to be laid bare, distinguishing rules whose purpose is to propagate uncertainty through reasoning steps, from rules whose main purpose is similarity-based interpolation [86], depending on the choice of a many-valued implication connective that models a rule. The bipolar view of information based on (δ, π) pairs sheds new light on the debate between conjunctive and implicative representation of rules [87]. Representing a rule as a material implication focuses on counterexamples to rules, while using a conjunction between antecedent and consequent points out examples of the rule and highlights its positive content. Traditionally in fuzzy control and modelling, the latter representation is adopted, while the former is the logical tradition. Introducing fuzzy implicative rules in modelling accounts for constraints or landmark points the model should comply with (as opposed to observed data) [88]. The bipolar view of rules in terms of examples and counterexamples may turn out to be very useful when extracting fuzzy rules from data [89].
2. Possibility theory also offers a framework for preference modeling in constraint-directed reasoning. Both prioritized and soft constraints can be captured by possibility distributions expressing degrees of feasibility rather than plausibility [90]. Possibility offers a natural setting for fuzzy optimization whose aim is to balance the levels of satisfaction of multiple fuzzy constraints (instead of minimizing an overall cost) [91]. Qualitative decision criteria are particularly adapted to the handling of uncertainty in this setting. Applications of possibility theory-based decision-making can be found in scheduling [92,93,94,95].
3. Quantitative possibility theory is the natural setting for a reconciliation between probability and fuzzy sets. An important research direction is the comparison between fuzzy interval analysis [96] and random variable calculations with a view to unifying them [97]. Indeed, a current major concern, in for instance risk analysis studies, is to perform uncertainty propagation under poor data and without independence assumptions (see the papers in the special issue [98]). Finding the potential of possibilistic representations in computing conservative bounds for such probabilistic calculations is certainly a major challenge [99] The active area of fuzzy random variables is also connected to this question [100].
4. One might also mention the well-known possibilistic clustering technique [101,102]. However, it is only loosely related to possibility theory. This name is due to the use of fuzzy clusters with (almost) unrestricted membership functions, that no longer form a usual fuzzy partition. But one might use it to generate genuine possibility distributions, where possibility derives from similarity, a point of view already mentioned before.

Other applications of possibility theory can be found in fields such as data analysis [103,104,105], database querying [106], diagnosis [107,108], belief revision [109], argumentation [110] case-based reasoning [111,112]. Lastly, possibility theory is also being studied from the point of view of its relevance in cognitive psychology. Experimental results [113] suggest that there are situations where people reason about uncertainty using the rules or possibility theory, rather than with those of probability theory.

Bibliography

1. Zadeh LA (1978) Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets Syst 1:3–28
2. Gaines BR, Kohout L (1975) Possible automata. In: Proc Int Symp Multiple-Valued Logic, Bloomington May 13–16. IEEE Press, pp 183–196
3. Dubois D, Prade H (1998) Possibility theory: Qualitative and quantitative aspects. In: Gabbay DM, Smets PP (eds) Handbook of Defeasible Reasoning and Uncertainty Management Systems, vol 1. Kluwer, Dordrecht, pp 169–226
4. Dubois D, Nguyen HT, Prade H (2000) Fuzzy sets and probability: Misunderstandings, bridges and gaps. In: Dubois D, Prade H (eds) Fundamentals of Fuzzy Sets. Kluwer, Boston, pp 343–438
5. Hughes GE, Cresswell MJ (1968) An introduction to modal logic. Methuen, London
6. Shackle GLS (1961) Decision, order and time in human affairs, 2nd edn. Cambridge University Press, Cambridge
7. Spohn W (1990) A general, nonprobabilistic theory of inductive reasoning. In: Shachter RD et al (eds) Uncertainty in Artificial Intelligence, vol 4. North Holland, Amsterdam, pp 149–158
8. Lewis DL (1973) Counterfactuals. Basil Blackwell, Oxford
9. Dubois D (1986) Belief structures, possibility theory and decomposable measures on finite sets. Comput AI 5:403–416
10. Sudkamp T (2002) Similarity and the measurement of possibility. Actes Rencontres Francophones sur la Logique Floue et ses Applications (Montpellier, France). Cepadues Editions, Toulouse, pp 13–26
11. Cohen LJ (1977) The probable and the provable. Clarendon, Oxford
12. Zadeh LA (1979) Fuzzy sets and information granularity. In: Gupta MM, Ragade R, Yager RR (eds) Advances in fuzzy set theory and applications. North-Holland, Amsterdam, pp 3–18
13. Zadeh LA (1982) Possibility theory and soft data analysis. In: Cobb L, Thrall R (eds) Mathematical frontiers of social and policy sciences. Westview Press, Boulder, pp 69–129
14. Dubois D, Prade H (1980) Fuzzy sets and systems: Theory and applications. Academic Press, New York
15. Dubois D, Prade H (1988) Possibility theory. Plenum, New York
16. Klir GJ, Folger T (1988) Fuzzy sets, uncertainty and information. Prentice Hall, Englewood Cliffs
17. Yager RR (1983) An introduction to applications of possibility theory. Hum Syst Manag 3:246–269
18. Dubois D, Hajek P, Prade H (2000) Knowledge-driven versus data-driven logics. J Log Lang Inf 9:65–89
19. Walley P (1996) Measures of uncertainty in expert systems. Artif Intell 83:1–58
20. De Cooman G (1997) Possibility theory. Part I: Measure- and integral-theoretic groundwork; Part II: Conditional possibility; Part III: Possibilistic independence. Int J Gen Syst 25:291–371
21. De Campos LM, Huete JF (1999) Independence concepts in possibility theory. Fuzzy Sets Syst 103:127–152 & 487–506
22. Dubois DD, Farinas del Cerro L, Herzig A, Prade H (1997) Qualitative relevance and independence: A roadmap. In: Proc of the 15h Inter Joint Conf on Artif Intell, Nagoya, 23–29 August, 1997. Morgan Kaufmann, San Mateo, pp 62–67
23. Borgelt C, Gebhardt J, Kruse R (2000) Possibilistic graphical models. In: Della Riccia G et al (eds) Computational intelligence in data mining. Springer, Wien, pp 51–68
24. Benferhat S, Dubois D, Garcia L, Prade H (2002) On the transformation between possibilistic logic bases and possibilistic causal networks. Int J Approx Reason 29(2):135–173
25. Ben Amor N, Benferhat S, Mellouli K (2003) Anytime propagation algorithm for min-based possibilistic graphs. Soft Comput 8(2):150–161
26. Borgelt C, Kruse R (2003) Operations and evaluation measures for learning possibilistic graphical models. Artif Intell 148(1–2):385–418
27. Gärdenfors P (1988) Knowledge in flux. MIT Press, Cambridge
28. Dubois D, Prade H (1991) Epistemic entrenchment and possibilistic logic. Artif Intell 50:223–239
29. Ben Amor N et al (2002) A theoretical framework for possibilistic independence in a weakly ordered setting. Int J Uncert Fuzz & Knowl-B Syst 10:117–155
30. Lehmann D, Magidor M (1992) What does a conditional knowledge base entail? Artif Intell 55:1–60
31. Dubois D, Fargier H, Prade H (2004) Ordinal and probabilistic representations of acceptance. J Artif Intell Res 22:23–56
32. Benferhat S, Dubois D, Prade H (1997) Nonmonotonic reasoning, conditional objects and possibility theory Artif Intell 92:259–276
33. Pearl J (1990) System Z: A natural ordering of defaults with tractable applications to default reasoning. In: Proc 3rd Conf Theoretical Aspects of Reasoning About Knowledge. Morgan Kaufmann, San Francisco, pp 121–135
34. Dubois D, Lang J, Prade H (1994) Possibilistic logic. In: Gabbay DM et al (eds) Handbook of logic in AI and logic programming, vol 3. Oxford University Press, Oxford, pp 439–513
35. Benferhat S, Dubois D, Prade H (1998) Practical handling of exception-tainted rules and independence information in possibilistic logic. Appl Intell 9:101–127
36. Dubois D, Prade H (2004)Possibilistic logic: A retrospective and prospective view. Fuzzy Sets Syst 144:3–23
37. Dubois D, Prade H (2007) Toward multiple-agent extensions of possibilistic logic. In: Proc IEEE Inter Conf on Fuzzy Systems (FUZZ-IEEE (2007)), London 23–26 July, 2007. pp 187–192
38. Benferhat S, Dubois D, Garcia L, Prade H (2002) On the transformation between possibilistic logic bases and possibilistic causal networks. Int J Approx Reason 29:135–173
39. Ben Amor N, Benferhat S (2005) Graphoid properties of qualitative possibilistic independence relations. Int J Uncert Fuzz & Knowl-B Syst 13:59–97
40. Savage LJ (1972) The foundations of statistics. Dover, New York
41. Yager RR (1979) Possibilistic decision making. IEEE Trans Syst Man Cybern 9:388–392

42. Whalen T (1984) Decision making under uncertainty with various assumptions about available information. IEEE Trans Syst Man Cybern 14:888–900
43. Grabisch M, Murofushi T, Sugeno M (eds) (2000) Fuzzy measures and integrals theory and applications. Physica, Heidelberg
44. Dubois D, Prade H, Sabbadin R (2000) Qualitative decision theory with Sugeno integrals. In: Grabisch M, Murofushi T, Sugeno M (eds) Fuzzy measures and integrals theory and applications. Physica, Heidelberg, pp 314–322
45. Dubois D, Prade H, Sabbadin R (2001) Decision-theoretic foundations of possibility theory. Eur J Oper Res 128:459–478
46. Dubois D, Fargier H, Perny P, Prade H (2003) Qualitative decision theory with preference relations and comparative uncertainty: An axiomatic approach. Artif Intell 148:219–260
47. Dubois D, Fargier H (2003) Qualitative decision rules under uncertainty. In: Della Riccia G et al (eds) Planning based on decision theory. CISM courses and Lectures, vol 472. Springer, Wien, pp 3–26
48. Fargier H, Sabbadin R (2005) Qualitative decision under uncertainty: Back to expected utility. Artif Intell 164:245–280
49. Dubois D (2006) Possibility theory and statistical reasoning. Comput Stat Data Anal 51(1):47–69
50. Dubois D, Prade H (1992) When upper probabilities are possibility measures. Fuzzy Sets Syst 49:s 65–74
51. Dubois D, Moral S, Prade H (1997) A semantics for possibility theory based on likelihoods. J Math Anal Appl 205:359–380
52. Shafer G (1987) Belief functions and possibility measures. In: Bezdek JC (ed) Analysis of fuzzy information, vol I: Mathematics and Logic. CRC Press, Boca Raton, pp 51–84
53. Benferhat S, Dubois D, Prade H (1999) Possibilistic and standard probabilistic semantics of conditional knowledge bases. J Log Comput 9:873–895
54. Maslov V (1987) Méthodes Opératorielles. Mir Publications, Moscow
55. Kolokoltsov VN, Maslov VP (1997) Idempotent analysis and applications. Kluwer, Dordrecht
56. Puhalskii A (2001) Large deviations and idempotent probability. Chapman and Hall, Boca Raton
57. Nguyen HT, Bouchon-Meunier B (2003) Random sets and large deviations principle as a foundation for possibility measures. Soft Comput 8:61–70
58. De Cooman G, Aeyels D (1999) Supremum-preserving upper probabilities. Inf Sci 118:173–212
59. Walley P, De Cooman G (1999) A behavioural model for linguistic uncertainty. Inf Sci 134:1–37
60. Gebhardt J, Kruse R (1993) The context model. I Int J Approx Reason 9:283–314
61. Joslyn C (1997) Measurement of possibilistic histograms from interval data. Int J Gen Syst 26:9–33
62. Neumaier A (2004) Clouds, fuzzy sets and probability intervals. Reliab Comput 10:249–272
63. De Baets B, Tsiporkova E, Mesiar R (1999) Conditioning in possibility with strict order norms. Fuzzy Sets Syst 106:221–229
64. Dubois D, Prade H (1997) Bayesian conditioning in possibility theory. Fuzzy Sets Syst 92:223–240
65. De Cooman G (2001) Integration and conditioning in numerical possibility theory. Ann Math AI 32:87–123
66. Walley P (1991) Statistical reasoning with imprecise probabilities. Chapman and Hall, Boca Raton
67. Klir GJ (1990) A principle of uncertainty and information invariance. Int J Gen Syst 17:249–275
68. Geer JF, Klir GJ (1992) A mathematical analysis of information-preserving transformations between probabilistic and possibilistic formulations of uncertainty. Int J Gen Syst 20:143–176
69. Dubois D, Prade H, Sandri S (1993) On possibility/probability transformations. In: Lowen R, Roubens M (eds) Fuzzy logic: State of the art. Kluwer, Dordrecht, pp 103–112
70. Klir GJ, Parviz B (1992) Probability B-possibility transformations: A comparison. Int J Gen Syst 21:291–310
71. Dubois D, Prade H (1982) On several representations of an uncertain body of evidence. In: Gupta M, Sanchez E (eds) Fuzzy information and decision processes. North-Holland, Amsterdam, pp 167–181
72. Smets P (1990) Constructing the pignistic probability function in a context of uncertainty. In: Henrion M et al (eds) Uncertainty in artificial intelligence, vol 5. North-Holland, Amsterdam, pp 29–39
73. Dubois D, Prade H (2001) Smets new semantics for quantitative possibility theory. In: Proc ESQARU (2001), Toulouse, LNAI 2143. Springer, pp 410–421
74. Birnbaum ZW (1948) On random variables with comparable peakedness. Ann Math Stat 19:76–81
75. Dubois D, Huellermeier E (2005) A Notion of comparative probabilistic entropy based on the possibilistic specificity ordering. In: Godo L (ed) Symbolic and Quantitative Approaches to Reasoning with Uncertainty. Proc of 8th European Conference, ECSQARU 2005, Barcelona, 6–8. Lecture Notes in Computer Science, vol 3571. Springer, Berlin
76. Dubois D, Foulloy L, Mauris G, Prade H (2004) Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. Reliab Comput 10:273–297
77. Mauris G, Lasserre V, Foulloy L (2000) Fuzzy modeling of measurement data acquired from physical sensors. IEEE Trans Meas Instrum 49:1201–1205
78. Baudrit C, Dubois D, Fargier H (2004) Practical representation of incomplete probabilistic information. In: Lopz-Diaz M et al (eds) Soft methods in probability and statistics. Proc 2nd Int Conf. Springer, Oviedo, pp 149–156
79. Dubois D, Prade H, Smets P (2003) A definition of subjective possibility. Badania Operacyjne i Decyzije (Wroclaw) 4:7–22
80. Dubois D, Prade H (1983) Unfair coins and necessity measures: A possibilistic interpretation of histograms. Fuzzy Sets Syst 10(1):15–20
81. Dubois D, Prade H (1985) Evidence measures based on fuzzy information. Automatica 21:547–562
82. Van Leekwijck W, Kerre EE (2001) Defuzzification: Criteria and classification. Fuzzy Sets Syst 108:303–314
83. Dubois D, Prade H (1987) The mean value of a fuzzy number. Fuzzy Sets Syst 24:279–300
84. Yager RR (1981) A procedure for ordering fuzzy subsets of the unit interval. Inf Sci 24:143–161
85. Chanas S, Nowakowski M (1988) Single value simulation of fuzzy variable. Fuzzy Sets Syst 25:43–57
86. Dubois D, Prade H (1996) What are fuzzy rules and how to use them. Fuzzy Sets Syst 84:169–185
87. Dubois D, Prade H, Ughetto L (2003) A new perspective on reasoning with fuzzy rules. Int J Intell Syst 18:541–567
88. Galichet S, Dubois D, Prade H (2004) Imprecise specification of ill-known functions using gradual rules. Int J Approx Reason 35:205–222

89. Dubois D, Huellermeier E, Prade H (2003) A note on quality measures for fuzzy association rules. In: De Baets B, Bilgic T (eds) Fuzzy sets and systems. Proc of the 10th Int Fuzzy Systems Assoc World Congress IFSA, Istanbul, 2003, LNAI 2715. Springer, pp 346–353
90. Dubois D, Fargier H, Prade H (1996) Possibility theory in constraint satisfaction problems: Handling priority, preference and uncertainty. Appl Intell 6:287–309
91. Dubois D, Fortemps P (1999) Computing improved optimal solutions to max-min flexible constraint satisfaction problems. Eur J Oper Res 118:95–126
92. Dubois D, Fargier H, Prade H (1995) Fuzzy constraints in job-shop scheduling. J Intell Manuf 6:215–234
93. Slowinski R, Hapke M (eds) (2000) Scheduling under fuzziness. Physica, Heidelberg
94. Chanas S, Zielinski P (2001) Critical path analysis in the network with fuzzy activity times. Fuzzy Sets Syst 122:195–204
95. Chanas S, Dubois D, Zielinski P (2002) Necessary criticality in the network with imprecise activity times. IEEE Trans Man Mach Cybern 32:393–407
96. Dubois D, Kerre E, Mesiar R, Prade H (2000) Fuzzy interval analysis. In: Dubois D, Prade H (eds) Fundamentals of fuzzy sets. Kluwer, Boston, pp 483–581
97. Dubois D, Prade H (1991) Random sets and fuzzy interval analysis. Fuzzy Sets Syst 42:87–101
98. Helton JC, Oberkampf WL (eds) (2004) Alternative Representations of Epistemic Uncertainty. Reliability Engineering and Systems Safety, vol 85. Elsevier, Amsterdam, p 369
99. Guyonnet D et al (2003) Hybrid approach for addressing uncertainty in risk assessments. J Env Eng 129:68–78
100. Gil M (ed) (2001) Fuzzy random variables. Inf Sci 133
101. Krishnapuram R, Keller J (1993) A possibilistic approach to clustering. IEEE Trans Fuzzy Syst 1:98–110
102. Bezdek J, Keller J, Krishnapuram R, Pal N (1999) Fuzzy models and algorithms for pattern recognition and image processing. In: The Handbooks of Fuzzy Sets Series. Kluwer, Boston
103. Wolkenhauer O (1998) Possibility theory with applications to data analysis. Research Studies Press, Chichester
104. Tanaka H, Guo PJ (1999) Possibilistic data analysis for operations research. Physica, Heidelberg
105. Borgelt C, Gebhardt J, Kruse R (2000) Possibilistic graphical models. In: Della Riccia G et al (eds) Computational intelligence in data mining. CISM Series, vol N408. Springer, Berlin
106. Bosc P, Prade H (1997) An introduction to the fuzzy set and possibility theory-based treatment of soft queries and uncertain of imprecise databases. In: Smets P, Motro A (eds) Uncertainty management in information systems. Kluwer, Dordrecht, pp 285–324
107. Cayrac D, Dubois D, Prade H (1996) Handling uncertainty with possibility theory and fuzzy sets in a satellite fault diagnosis application. IEEE Trans Fuzzy Syst 4:251–269
108. Boverie S et al (2002) Online diagnosis of engine dyno test benches: A possibilistic approach. Proc 15th Eur Conf on Artificial Intelligence, Lyon. IOS Press, Amsterdam, p 658–662
109. Benferhat S, Dubois D, Prade H, Williams M-A (2002) A practical approach to revising prioritized knowledge bases. Stud Log 70:105–130
110. Amgoud L, Prade H (2004) Reaching agreement through argumentation: A possibilistic approach. In: Proc of the 9th Int Conf on Principles of Knowledge Representation and Reasoning (KR'04), Whistler. AAAI Press, Cambridge, pp 175–182
111. Dubois D, Huellermeier E, Prade H (2002) Fuzzy set-based methods in instance-based reasoning. IEEE Trans Fuzzy Syst 10:322–332
112. Huellermeier E, Dubois D, Prade H (2002) Model adaptation in possibilistic instance-based reasoning. IEEE Trans Fuzzy Syst 10:333–339
113. Raufaste E, Da Silva R, Neves C (2003) Mariné testing the descriptive validity of possibility theory in human judgements of uncertainty. Artif Intell 148:197–218

Pressure and Equilibrium States in Ergodic Theory

Jean-René Chazottes[1], Gerhard Keller[2]

[1] Centre de Physique Théorique, CNRS/École Polytechnique, Palaiseau, France

[2] Department Mathematik, Universität Erlangen-Nürnberg, Erlangen, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Warming Up: Thermodynamic Formalism for Finite Systems
Shift Spaces, Invariant Measures and Entropy
The Variational Principle: A Global Characterization of Equilibrium
The Gibbs Property: A Local Characterization of Equilibrium
Examples on Shift Spaces
Examples from Differentiable Dynamics
Nonequilibrium Steady States and Entropy Production
Some Ongoing Developments and Future Directions
Bibliography

Glossary

Dynamical system In this article: a continuous transformation T of a compact metric space X. For each $x \in X$, the transformation T generates a trajectory $(x, Tx, T^2x, \dots)$.

Invariant measure In this article: a probability measure μ on X which is invariant under the transformation T, i. e., for which $\langle f \circ T, \mu\rangle = \langle f, \mu\rangle$ for each continuous $f\colon X \to \mathbb{R}$. Here $\langle f, \mu\rangle$ is a short-hand notation for $\int_X f \, d\mu$. The triple (X, T, μ) is called a measure-preserving dynamical system.

Ergodic theory Ergodic theory is the mathematical theory of measure-preserving dynamical systems.

Entropy In this article: the maximal rate of information gain per time that can be achieved by coarse-grained observations on a measure-preserving dynamical system. This quantity is often denoted $h(\mu)$.

Equilibrium state In general, a given dynamical system $T: X \to X$ admits a huge number of invariant measures. Given some continuous $\phi: X \to \mathbb{R}$ ("potential"), those invariant measures which maximize a functional of the form $F(\mu) = h(\mu) + \langle \phi, \mu \rangle$ are called "equilibrium states" for ϕ.

Pressure The maximum of the functional $F(\mu)$ is denoted by $P(\phi)$ and called the "topological pressure" of ϕ, or simply the "pressure" of ϕ.

Gibbs state In many cases, equilibrium states have a local structure that is determined by the local properties of the potential ϕ. They are called "Gibbs states".

Sinai–Ruelle–Bowen measure Special equilibrium or Gibbs states that describe the statistics of the attractor of certain smooth dynamical systems.

Definition of the Subject

Gibbs and equilibrium states of one-dimensional lattice models in statistical physics play a prominent role in the statistical theory of chaotic dynamics. They first appear in the ergodic theory of certain differentiable dynamical systems, called "uniformly hyperbolic systems", mainly Anosov and Axiom A diffeomorphisms (and flows). The central idea is to "code" the orbits of these systems into (infinite) symbolic sequences of symbols by following their history on a finite partition of their phase space. This defines a nice shift dynamical system called a subshift of finite type or a topological Markov chain. Then the construction of their "natural" invariant measures and the study of their properties are carried out at the symbolic level by constructing certain equilibrium states in the sense of statistical mechanics which turn out to be also Gibbs states. The study of uniformly hyperbolic systems brought out several ideas and techniques which turned out to be extremely fruitful for the study of more general systems. Let us mention the concept of Markov partition and its avatars, the very important notion of SRB measure (after Sinai, Ruelle, and Bowen) and transfer operators. Recently, there was a revival of interest in Axiom A systems as models to understand nonequilibrium statistical mechanics.

Introduction

Our goal is to present the basic results on one-dimensional Gibbs and equilibrium states viewed as special invariant measures on symbolic dynamical systems, and then to describe without technicalities a sample of results they allow to obtain for certain differentiable dynamical systems. We hope that this contribution will illustrate the symbiotic relationship between ergodic theory and statistical mechanics, and also information theory.

We start by putting Gibbs and equilibrium states in a general perspective. The theory of Gibbs states and equilibrium states, or Thermodynamic Formalism, is a branch of rigorous Statistical Physics. The notion of a Gibbs state dates back to R.L. Dobrushin (1968–1969) [17,18,19,20] and O.E. Lanford and D. Ruelle (1969) [41] who proposed it as a mathematical idealization of an equilibrium state of a physical system which consists of a very large number of interacting components. For a finite number of components, the foundations of statistical mechanics were already laid in the nineteenth century. There was the well-known Maxwell–Boltzmann–Gibbs formula for the equilibrium distribution of a physical system with given energy function. From the mathematical point of view, the intrinsic properties of very large objects can be made manifest by performing suitable limiting procedures. Indeed, the crucial step made in the 1960s was to define the notion of a Gibbs measure or Gibbs state for a system with an infinite number of interacting components. This was done by the familiar probabilistic idea of specifying the interdependence structure by means of a suitable class of conditional probabilities built up according to the Maxwell–Boltzmann–Gibbs formula [29]. Notice that Gibbs states are often called "DLR states" in honor of Dobrushin, Lanford, and Ruelle. The remarkable aspect of this construction is the fact that a Gibbs state for a given type of interaction may fail to be unique. In physical terms, this means that a system with this interaction can take several distinct equilibria. The phenomenon of nonuniqueness of a Gibbs measure can thus be interpreted as a phase transition. Therefore, the conditions under which an interaction leads to a unique or to several Gibbs measures turns out to be of central importance. While Gibbs states are defined locally by specifying certain conditional probabilities, equilibrium states are defined globally by a *variational principle*: they maximize the entropy of the system under the (linear) constraint that the mean energy is fixed. Gibbs states are always equilibrium states, but the two notions do not coincide in general. However, for a class of sufficiently regular interactions, equilibrium states are also Gibbs states.

In the effort of trying to understand phase transitions, simplified mathematical models were proposed, the most famous one being undoubtedly the Ising model. This is an example of a lattice model. The set of configurations of a lattice model is $X := A^{\mathbb{Z}^d}$, where A is a finite set, which is invariant by "spatial" translations. For the physical in-

terpretation, X can be thought, for instance, as the set of infinite configurations of a system of spins on a crystal lattice $\mathbb{Z}^d$ and one may take $A = \{-1, +1\}$, i. e., spins can take two orientations, "up" and "down". The Ising model is defined by specifying an interaction (or potential) between spins and studying the corresponding (translation-invariant) Gibbs states. The striking phenomenon is that for $d = 1$ there is a unique Gibbs state (in fact a Markov measure) whereas if $d \geq 2$, there may be several Gibbs states although the interaction is very simple [29].

Equilibrium states and Gibbs states of one-dimensional lattice models ($d = 1$) played a prominent role in understanding the ergodic properties of certain types of differentiable dynamical systems, namely uniformly hyperbolic systems, Axiom A diffeomorphisms in particular. The link between one-dimensional lattice systems and dynamical systems is made by *symbolic dynamics*. Informally, symbolic dynamics consists of replacing the orbits of the original system by its history on a finite partition of its phase space labeled by the elements of the "alphabet" A. Therefore, each orbit of the original system is replaced by an infinite sequence of symbols, i. e., by an element of the set $A^{\mathbb{Z}}$ or $A^{\mathbb{N}}$, depending on whether the map describing the dynamics is invertible or not. The action of the map on an initial condition is then easily seen to correspond to the translation (or shift) of its associated symbolic sequence. In general there is no reason to get all sequences of $A^{\mathbb{Z}}$ or $A^{\mathbb{N}}$. Instead one gets a closed invariant subset X (a subshift) which can be very complicated. For a certain class of dynamical systems the partition can be successfully chosen so as to form a *Markov partition*. In this case, the dynamical system under consideration can be coded by a *subshift of finite type* (also called a *topological Markov chain*) which is a very nice symbolic dynamical system. Then one can play the game of statistical physics: for a given continuous, real-valued function (a "potential") on X, construct the corresponding Gibbs states and equilibrium states. If the potential is regular enough, one expects uniqueness of the Gibbs state and that it is also the unique equilibrium state for this potential. This circle of ideas – ranging from Gibbs states on finite systems over invariant measures on symbolic systems and their (Shannon-)entropy with a digression to Kolmogorov–Chaitin complexity to equilibrium states and Gibbs states on subshifts of finite type – is presented in the next four sections.

At this point it should be remembered that the objects which can actually be observed are not equilibrium states (they are measures on X) but individual symbol sequences in X, which reflect more or less the statistical properties of an equilibrium state. Indeed, most sequences reflect these properties very well, but there are also rare sequences that look quite different. Their properties are described by *large deviations principles* which are not discussed in the present article. We shall indicate some references along the way.

In Sects. "Examples on Shift Spaces" and "Examples from Differentiable Dynamics" we present a selection of important examples: measure of maximal entropy, Markov measures and Hofbauer's example of nonuniqueness of equilibrium state; uniformly expanding Markov maps of the interval, interval maps with an indifferent fixed point, Anosov diffeomorphisms and Axiom A attractors with Sinai–Ruelle–Bowen measures, and Bowen's formula for the Hausdorff dimension of conformal repellers. As we shall see, Sinai–Ruelle–Bowen measures are the only physically observable measures and they appear naturally in the context of nonuniformly hyperbolic diffeomorphisms [71].

A revival of the interest to Anosov and Axiom A systems occurred in statistical mechanics in the 1990s. Several physical phenomena of nonequilibrium origin, like entropy production and chaotic scattering, were modeled with the help of those systems (by G. Gallavotti, P. Gaspard, D. Ruelle, and others). This new interest led to new results about old Anosov and Axiom A systems, see, e. g., [15] for a survey and references. In Sect. "Nonequilibrium Steady States and Entropy Production", we give a very brief account of *entropy production* in the context of Anosov systems which highlights the role of *relative entropy*.

This article is a little introduction to a vast subject in which we have tried to put forward some aspects not previously described in other expository texts. For readers willing to deepen their understanding of equilibrium and Gibbs states, there are the classic monographs by Bowen [6] and by Ruelle [58], the monograph by one of us [38], and the survey article by Chernov [15] (where Anosov and Axiom A flows are reviewed). Those texts are really complementary.

Warming Up: Thermodynamic Formalism for Finite Systems

We introduce the thermodynamic formalism in an elementary context, following Jaynes [34]. In this view, entropy, in the sense of information theory, is the central concept.

Incomplete knowledge about a system is conveniently described in terms of probability distributions on the set of its possible states. This is particularly simple if the set of states, call it X, is finite. Then the equidistribution on X describes complete lack of knowledge, whereas a probability vector that assigns probability 1 to one single state and

probability 0 to all others represents maximal information about the system. A well-established measure of the amount of uncertainty represented by a probability distribution $\nu = (\nu(x))_{x\in X}$ is its *entropy*

$$H(\nu) := -\sum_{x\in X} \nu(x) \log \nu(x) ,$$

which is zero if the probability is concentrated in one state and which attains its maximum value $\log|X|$ if ν is the equidistribution on X, i. e., if $\nu(x) = |X|^{-1}$ for all $x \in X$. In this completely elementary context we will explore two concepts whose generalizations are central to the theory of equilibrium states in ergodic theory:

- Equilibrium distributions – defined in terms of a variational problem.
- The Gibbs property of equilibrium distributions.

The only mathematical prerequisite for this section are calculus and some elements from probability theory.

Equilibrium Distributions and the Gibbs Property

Suppose that a finite system can be observed through a function $U\colon X \to \mathbb{R}$ (an "observable"), and that we are looking for a probability distribution μ which maximizes entropy among all distributions ν with a prescribed expected value $\langle U, \nu\rangle := \sum_{x\in X} \nu(x)U(x)$ for the observable U. This means we have to solve a variational problem under constraints:

$$H(\mu) = \max\{H(\nu) : \langle U, \nu\rangle = E\} . \tag{1}$$

As the function $\nu \mapsto H(\nu)$ is strictly concave, there is a unique maximizing probability distribution μ provided the value E can be attained at all by some $\langle U, \nu\rangle$. In order to derive an explicit formula for this μ we introduce a Lagrange multiplier $\beta \in \mathbb{R}$ and study, for each β, the unconstrained problem

$$H(\mu_\beta) + \langle \beta U, \mu_\beta\rangle = p(\beta U) := \max_\nu (H(\nu) + \langle \beta U, \nu\rangle) . \tag{2}$$

In analogy to the convention in ergodic theory we call $p(\beta U)$ the *pressure* of βU and the maximizer μ_β the corresponding *equilibrium distribution* (synonymously *equilibrium state*).

The equilibrium distribution μ_β satisfies

$$\mu_\beta(x) = \exp(-p(\beta U) + \beta U(x)) \quad \text{for all } x \in X \tag{3}$$

as an elementary calculation using Jensen's inequality for the strictly convex function $t \mapsto -t \log t$ shows:

$$\begin{aligned} H(\nu) + \langle \beta U, \nu\rangle &= \sum_{x\in X} \nu(x) \log \frac{e^{\beta U(x)}}{\nu(x)} \\ &\le \log \sum_{x\in X} \nu(x) \frac{e^{\beta U(x)}}{\nu(x)} \\ &= \log \sum_{x\in X} e^{\beta U(x)} , \end{aligned}$$

with equality if and only if $e^{\beta U}$ is a constant multiple of ν. The observation that $\nu = \mu_\beta$ is a maximizer proves at the same time that $p(\beta U) = \log \sum_{x\in X} e^{\beta U(x)}$.

The equality expressed in (3) is called the *Gibbs property* of μ_β, and we say that μ_β is a Gibbs distribution if we want to stress this property.

In order to solve the constrained problem (1) it remains to show that there is a unique multiplier $\beta = \beta(E)$ such that $\langle U, \mu_\beta\rangle = E$. This follows from the fact that the map $\beta \mapsto \langle U, \mu_\beta\rangle$ maps the real line monotonically onto the interval $(\min U, \max U)$ which, in turn, is a direct consequence of the formulas for the first and second derivative of $p(\beta U)$ w.r.t β:

$$\frac{dp}{d\beta} = \langle U, \mu_\beta\rangle , \quad \frac{d^2p}{d\beta^2} = \langle U^2, \mu_\beta\rangle - \langle U, \mu_\beta\rangle^2 . \tag{4}$$

As the second derivative is nothing but the variance of U under μ_β, it is strictly positive (except when U is a constant function), so that $\beta \mapsto \langle U, \mu_\beta\rangle$ is indeed strictly increasing. Observe also that $dp/d\beta$ is indeed the directional derivative of $p\colon \mathbb{R}^{|A|} \to \mathbb{R}$ in direction U. Hence the first identity in (4) can be rephrased as: μ_β is the gradient at βU of the function p.

A similar analysis can be performed for an $\mathbb{R}^d$-valued observable U. In that case a vector $\beta \in \mathbb{R}^d$ of Lagrange multipliers is needed to satisfy the d linear constraints.

Systems on a Finite Lattice

We now assume that the system has a lattice structure, modeling its extension in space, for instance. The system can be in different states at different positions. More specifically, let $\mathbb{L}_n = \{0, 1, \ldots, n-1\}$ be a set of n positions in space, let A be a finite set of states that can be attained by the system at each of its sites, and denote by $X := A^{\mathbb{L}_n}$ the set of all configurations of states from A at positions of $\mathbb{L}_n$. It is helpful to think of X as the set of all words of length n over the alphabet A. We focus on observables U_n which are sums of many local contributions in the sense that $U_n(a_0 \ldots a_{n-1}) = \sum_{i=0}^{n-1} \phi(a_i \ldots a_{i+r-1})$ for

some "local observable" $\phi\colon A^r \to \mathbb{R}$. (The index $i+r-1$ has to be taken modulo n.) In terms of ϕ the maximizing measure can be written as

$$\mu_\beta(a_0 \dots a_{n-1}) = \exp\left(-nP(\beta\phi) + \beta \sum_{i=0}^{n-1} \phi(a_i \dots a_{i+r-1})\right) , \quad (5)$$

where $P(\beta\phi) := n^{-1}p(\beta U_n)$. A first immediate consequence of (5) is the invariance of μ_β under a cyclic shift of its argument, namely $\mu_\beta(a_1 \dots a_{n-1}a_0) = \mu_\beta(a_0 \dots a_{n-1})$. Therefore, we can restrict the maximizations in (1) and (2) to probability distributions ν which are invariant under cyclic translations which yields

$$\begin{aligned} P(\beta\phi) &= \max_\nu (n^{-1}H(\nu) + \langle \beta\phi, \nu\rangle) \\ &= n^{-1}H(\mu_\beta) + \langle \beta\phi, \mu_\beta\rangle . \end{aligned} \quad (6)$$

If the local observable ϕ depends only on one coordinate, μ_β turns out to be a product measure:

$$\mu_\beta(a_0 \dots a_{n-1}) = \prod_{i=0}^{n-1} \exp\left(-P(\beta\phi) + \beta\phi(a_i)\right) .$$

Indeed, comparison with (3) shows that μ_β is the n-fold product of the probability distribution μ_β^{loc} on A that maximizes $H(\nu) + \beta\nu(\phi)$ among all distributions ν on A. It follows that $n^{-1}H(\mu_\beta) = H(\mu_\beta^{\mathrm{loc}})$ so that (6) implies $P(\beta\phi) = p(\beta\phi)$ for observables ϕ that depend only on one coordinate.

Shift Spaces, Invariant Measures and Entropy

We now turn to *shift dynamical systems* over a finite alphabet A.

Symbolic Dynamics

We start by fixing some notation. Let $\mathbb{N}$ denote the set $\{0, 1, 2, \dots\}$. In the sequel we need

- a finite set A (the "alphabet"),
- the set $A^{\mathbb{N}}$ of all infinite sequences over A, i. e., the set of all $\underline{x} = x_0x_1 \dots$ with $x_n \in A$ for all $n \in \mathbb{N}$,
- the translation (or shift) $\sigma\colon A^{\mathbb{N}} \to A^{\mathbb{N}}$, $(\sigma\underline{x})_n = x_{n+1}$, for all $n \in \mathbb{N}$,
- a shift invariant subset $X = \sigma(X)$ of $A^{\mathbb{N}}$. With a slight abuse of notation we denote the restriction of σ to X by σ again.

We mention two interpretations of the dynamics of σ: it can describe the evolution of a system with state space X in discrete time steps (this is the prevalent interpretation if $\sigma\colon X \to X$ is obtained as a symbolic representation of another dynamical system), or it can be the spatial translation of the configuration of a system on an infinite lattice (generalizing the point of view from Subsect. "Systems on a Finite Lattice" above). In the latter case one usually looks at the shift on the two-sided shift space $A^{\mathbb{Z}}$, for which the theory is nearly identical.

On $A^{\mathbb{N}}$ one can define a metric d by

$$d(\underline{x}, \underline{y}) := 2^{-N(\underline{x},\underline{y})} \quad \text{where } N(\underline{x}, \underline{y}) := \min\{k \in \mathbb{N} : x_k \neq y_k\} . \quad (7)$$

Hence $d(\underline{x}, \underline{y}) = 1$ if and only if $x_0 \neq y_0$, and $d(\underline{x}, \underline{x}) = 0$ upon agreeing that $N(\underline{x}, \underline{x}) = \infty$ and $2^{-\infty} = 0$. Equipped with this metric, $A^{\mathbb{N}}$ becomes a compact metric space and σ is easily seen to be a continuous surjection of $A^{\mathbb{N}}$. Finally, if X is a closed subset of $A^{\mathbb{N}}$, we call the restriction $\sigma\colon X \to X$, which is again a continuous surjection, a shift dynamical system. We remark that d generates on $A^{\mathbb{N}}$ the product topology of the discrete topology on A, just as many variants of d do. For more details ▶ Symbolic Dynamics. As usual, $C(X)$ denotes the space of real-valued continuous functions on X equipped with the supremum norm $\|\cdot\|_\infty$.

Invariant Measures

A probability distribution ν (or simply distribution) on X is a Borel probability measure on X. It is unambiguously specified by its values $\nu[a_0 \dots a_{n-1}]$ ($n \in \mathbb{N}$, $a_i \in A$) on *cylinder sets*

$$[a_0 \dots a_{n-1}] := \{\underline{x} \in X : x_i = a_i \text{ for all } i = 0, \dots, n-1\} .$$

Any bounded and measurable $f\colon X \to \mathbb{R}$ (in particular any $f \in C(X)$) can be integrated by any distribution ν. To stress the linearity of the integral in both, the integrand and the integrator, we use the notation

$$\langle f, \nu\rangle := \int_X f \mathrm{d}\nu .$$

In probabilistic terms, $\langle f, \nu\rangle$ is the expectation of the observable f under ν. The set $\mathcal{M}(X)$ of all probability distributions is compact in the weak topology, the coarsest topology on $\mathcal{M}(X)$ for which $\nu \mapsto \langle f, \nu\rangle$ is continuous for all $f \in C(X)$, ▶ Measure Preserving Systems, Subsect. "Existence of Invariant Measures". (Note that in functional analysis this is called the weak-* topology.) Henceforth we will use both terms, "measure" and "distribution", if we talk about probability distributions.

A measure ν on X is *invariant* if expectations of observables are unchanged under the shift, i. e., if

$$\langle f \circ \sigma, \nu \rangle = \langle f, \nu \rangle \quad \text{for all bounded measurable } f\colon X \to \mathbb{R}\,.$$

The set of all invariant measures is denoted by $\mathcal{M}_\sigma(X)$. As a closed subset of $\mathcal{M}(X)$ it is compact in the weak topology. Of special importance among all invariant measures ν are the ergodic ones which can be characterized by the property that, for all bounded measurable $f\colon X \to \mathbb{R}$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f(\sigma^k \underline{x}) = \langle f, \nu \rangle \quad \text{for } \nu\text{-a. e. (almost every) } \underline{x}\,, \tag{8}$$

i. e., for a set of $\underline{x}$ of ν-measure one. They are the indecomposable "building blocks" of all other measures in $\mathcal{M}_\sigma(X)$, ► Measure Preserving Systems or ► Ergodic Theorems. The almost everywhere convergence in (8) is Birkhoff's ergodic Theorem ► Ergodic Theorems, the constant limit characterizes the ergodicity of ν.

Entropy of Invariant Measures

We give a brief account of the definition and basic properties of the entropy of an invariant measure ν. For details and the generalization of this concept to general dynamical systems we refer to ► Entropy in Ergodic Theory or [37], and to [36] for an historical account.

Let $\nu \in \mathcal{M}_\sigma(X)$. For each $n > 0$ the cylinder probabilities $\nu[a_0 \dots a_{n-1}]$ give rise to a probability distribution on the finite set $A^{\mathbb{L}_n}$, see Sect. "Warming Up: Thermodynamic Formalism for Finite Systems", so

$$H_n(\nu) := -\sum_{a_0,\dots,a_{n-1}\in A} \nu[a_0 \dots a_{n-1}] \log \nu[a_0 \dots a_{n-1}]$$

is well defined. Invariance of ν guarantees that the sequence $(H_n(\nu))_{n>0}$ is *subadditive*, i. e., $H_{k+n}(\nu) \le H_k(\nu) + H_n(\nu)$, and an elementary argument shows that the limit

$$h(\nu) := \lim_{n\to\infty} \frac{1}{n} H_n(\nu) \in [0, \log|A|] \tag{9}$$

exists and equals the infimum of the sequence. We simply call it the *entropy* of ν. (Note that for general subshifts X many of the cylinder sets $[a_0 \dots a_{n-1}] \subseteq X$ are empty. But, because of the continuity of the function $t \mapsto t \log t$ at $t = 0$, we may set $0 \log 0 = 0$, and, hence, this does not affect the definition of $H_n(\nu)$.)

The entropy $h(\nu)$ of an ergodic measure ν can be observed along a "typical" trajectory. That is the content of the following theorem, sometimes called the "ergodic theorem of information theory" ► Entropy in Ergodic Theory.

Theorem (Shannon–McMillan–Breiman Theorem)

$$\lim_{n\to\infty} \frac{1}{n} \log \nu[x_0 \dots x_{n-1}] = -h(\nu) \quad \textit{for } \nu\textit{-a. e. } \underline{x}. \tag{10}$$

Observe that (9) is just the integrated version of this statement. A slightly weaker reformulation of this theorem (again for ergodic ν) is known as the "asymptotic equipartition property".

Asymptotic Equipartition Property

Given (arbitrarily small) $\epsilon > 0$ and $\alpha > 0$, one can, for each sufficiently large n, partition the set A^n into a set $\mathcal{T}_n$ of typical words and a set $\mathcal{E}_n$ of exceptional words such that each $a_0 \dots a_{n-1} \in \mathcal{T}_n$ satisfies

$$\mathrm{e}^{-n(h(\nu)+\alpha)} \le \nu[a_0 \dots a_{n-1}] \le \mathrm{e}^{-n(h(\nu)-\alpha)} \tag{11}$$

and the total probability $\sum_{a_0\dots a_{n-1}\in\mathcal{E}_n} \nu[a_0 \dots a_{n-1}]$ of the exceptional words is at most ε.

A Short Digression on Complexity

Kolmogorov [40] and Chaitin [14] introduced the concept of complexity of an infinite sequence of symbols. Very roughly it is defined as follows: First, the complexity $K(x_0 \dots x_{n-1})$ of a finite word in A^n is defined as the bit length of the shortest program that causes a suitable general purpose computer (say a PC or, for the mathematically minded reader, a Turing machine) to print out this word. Then the complexity of an infinite sequence is defined as $K(\underline{x}) := \limsup_{n\to\infty} \frac{1}{n} K(x_0 \dots x_{n-1})$. Of course, the definition of $K(x_0 \dots x_{n-1})$ depends on the particular computer, but as any two general purpose computers can be programmed to simulate each other (by some finite piece of software), the limit $K(\underline{x})$ is machine independent. It is the optimal compression factor for long initial pieces of a sequence $\underline{x}$ that still allows complete reconstruction of $\underline{x}$ by an algorithm. Brudno [8] showed:

If $X \subseteq A^{\mathbb{N}}$ and $\nu \in \mathcal{M}_\sigma(X)$ is ergodic, then $K(\underline{x}) = \frac{1}{\log 2} h(\nu)$ for ν-a. e. $\underline{x} \in X$.

Entropy as a Function of the Measure

An important technical remark for the further development of the theory is that the entropy function

$h\colon \mathcal{M}_\sigma(X) \to [0,\infty)$ is *upper semicontinuous*. This means that all sets $\{\nu : h(\nu) \geq t\}$ with $t \in \mathbb{R}$ are closed and hence compact. In particular, upper semicontinuous functions attain their supremum. Indeed, suppose a sequence $\nu_k \in \mathcal{M}_\sigma(X)$ converges weakly to some $\nu \in \mathcal{M}_\sigma(X)$ and $h(\nu_k) \geq t$ for all k so that also $\frac{1}{n}H_n(\nu_k) \geq t$ for all n and k. As $H_n(\nu)$ is an expression that depends continuously on the probabilities of the finitely many cylinders $[a_0 \dots a_{n-1}]$ and as the indicator functions of these sets are continuous, $\frac{1}{n}H_n(\nu) = \lim_{k\to\infty} \frac{1}{n}H_n(\nu_k) \geq t$, hence $h(\nu) \geq t$ in the limit $n \to \infty$.

A word of caution seems in order: the entropy function is rarely continuous. For example, on the full shift $X = A^{\mathbb{N}}$ each invariant measure, whatever its entropy is, can be approximated in the weak topology by equidistributions on periodic orbits. But all these equidistributions have entropy zero.

The Variational Principle: A Global Characterization of Equilibrium

Usually, a dynamical systems model of a "physical" system consists of a state space and a map (or a differential equation) describing the dynamics. An invariant measure for the system is rarely given a priori. Indeed, many (if not most) dynamical systems arising in this way have uncountably many ergodic invariant measures. This limits considerably the "practical value" of Birkhoff's ergodic theorem (8) or the Shannon–McMillan–Breiman theorem (10): not only do the limits in these theorems depend on the invariant measure ν, but also the sets of points for which the theorems guarantee almost everywhere convergence are practically disjoint for different ν and ν' in $\mathcal{M}_\sigma(X)$. Therefore, a choice of ν has to be made which reflects the original modeling intentions. We will argue in this and the next sections that a variational principle with a judiciously chosen "observable" may be a useful guideline – generalizing the observations for finite systems collected in the corresponding section above. As announced earlier we restrict again to shift dynamical systems, because they are rather universal models for many other systems.

Equilibrium States

We define the *pressure* of an observable $\phi \in C(X)$ as

$$P(\phi) := \sup\{h(\nu) + \langle \phi, \nu\rangle : \nu \in \mathcal{M}_\sigma(X)\}\,. \tag{12}$$

Since $\mathcal{M}_\sigma(X)$ is compact and the functional $\nu \mapsto h(\nu) + \langle \phi, \nu\rangle$ is upper semicontinuous, the supremum is attained – not necessarily at a unique measure as we will see (which is remarkably different from what happens in finite systems). Each measure ν for which the supremum is attained is called an *equilibrium state* for ϕ. Here the word "state" is used synonymously with "distribution" or "measure" – a reflection of the fact that in "well-behaved cases", as we will see in the next section, this measure is uniquely determined by the constraint(s) under which it maximizes entropy, and that means by the *macroscopic state* of the system. (In contrast, the word "state" was used in the above section on finite systems to designate microscopic states.)

As, for each $\nu \in \mathcal{M}_\sigma(X)$, the functional $\phi \mapsto h(\nu) + \langle \phi, \nu\rangle$ is affine on $C(X)$, the pressure functional $P\colon C(X) \to \mathbb{R}$, which, by definition, is the pointwise supremum of these functionals, is convex. It is therefore instructive to fit equilibrium states into the abstract framework of convex analysis [32,38,45,68]. To this end recall the identities in (4) that identify, for finite systems, equilibrium states as gradients of the pressure function $p\colon \mathbb{R}^{|A|} \to \mathbb{R}$ and guarantee that p is twice differentiable and strictly convex. In the present setting where P is defined on the Banach space $C(X)$, differentiability and strict convexity are no more guaranteed, but one can show:

Equilibrium states as (sub)-gradients

$\phi \in \mathcal{M}_\sigma(X)$ is an equilibrium state for ϕ if and only if μ is a subgradient (or tangent functional) for P at ϕ, i. e., if $P(\phi+\psi)-P(\phi) \geq \langle \psi, \mu\rangle$ for all $\psi \in C(X)$. In particular, ϕ has a unique equilibrium state μ if (13) and only if P is differentiable at ϕ with gradient μ, i. e., if $\lim_{t\to 0} \frac{1}{t}\left(P(\phi + t\psi) - P(\phi)\right) = \langle \psi, \mu\rangle$ for all $\psi \in C(X)$.

Let us see how equilibrium states on $X = A^{\mathbb{N}}$ can directly be obtained from the corresponding equilibrium distributions on finite sets A^n introduced in Subsect. "Systems on a Finite Lattice". Define $\phi^{(n)}\colon A^n \to \mathbb{R}$ by $\phi^{(n)}(a_0 \dots a_{n-1}) := \phi(a_0 \dots a_{n-1} a_0 \dots a_{n-1} \dots)$, denote by U_n the corresponding global observable on A^n, and let μ_n be the equilibrium distribution on A^n that maximizes $H(\mu) + \langle U_n, \mu\rangle$. Then all weak limit points of the "approximative equilibrium distributions" μ_n on A^n are equilibrium states on $A^{\mathbb{N}}$.

This can be seen as follows: Let the measure μ on $A^{\mathbb{N}}$ be any weak limit point of the μ_n. Then, given $\epsilon > 0$ there exists $k \in \mathbb{N}$ such that

$$\begin{aligned} h(\mu) + \langle \phi, \mu\rangle &\geq \frac{1}{k}H_k(\mu) + \langle \phi, \mu\rangle - \epsilon \\ &\geq \frac{1}{k}H_k(\mu_n) + \langle \phi^{(n)}, \mu_n\rangle - 2\epsilon \end{aligned}$$

for arbitrarily large n, because $\|\phi - \phi^{(n)}\|_\infty \to 0$ as $n \to \infty$ by construction of the $\phi^{(n)}$. As the μ_n are invariant under cyclic coordinate shifts (see Subsect. "Systems on a Finite Lattice"), it follows from the subadditivity of the entropy that

$$h(\mu) + \langle \phi, \mu \rangle \geq \frac{1}{n}(H_n(\mu_n) + \langle U_n, \mu_n \rangle) - 2\epsilon - \frac{k}{n} \log |A| \,.$$

Hence, for each $\nu \in \mathcal{M}_\sigma(X)$,

$$h(\mu) + \langle \phi, \mu \rangle \geq \frac{1}{n}(H_n(\nu) + \langle U_n, \nu \rangle) - 2\epsilon - \frac{k}{n} \log |A| \to h(\nu) + \langle \phi, \nu \rangle - 2\epsilon$$

as $n \to \infty$, and we see that μ is indeed an equilibrium state on $A^{\mathbb{N}}$.

The Variational Principle

In Subsect. "Equilibrium Distributions and the Gibbs Property", the pressure of a finite system was defined as a certain supremum and then identified as the logarithm of the normalizing constant for the Gibbsian representation of the corresponding equilibrium distribution. We are now going to approximate equilibrium states by suitable Gibbs distributions on finite subsets of X. As a by-product the pressure $P(\phi)$ is characterized in terms of the logarithms of the normalizing constants of these approximating distributions. Let $S_n\phi(\underline{x}) := \phi(\underline{x}) + \phi(\sigma\underline{x}) + \cdots + \phi(\sigma^{n-1}\underline{x})$. From each cylinder set $[a_0 \ldots a_{n-1}]$ we can pick a point $\underline{z}$ such that $S_n\phi(\underline{z})$ is the maximal value of $S_n\phi$ on this set. We denote the collection of the $|A|^n$ points we obtain in this way by E_n. Observe that E_n is not unambiguously defined, but any choice we make will do.

Theorem (Variational Principle for the Pressure)

$$P(\phi) = \limsup_{n\to\infty} \frac{1}{n} P_n(\phi) \quad \text{where} \quad P_n(\phi) := \log \sum_{\underline{z}\in E_n} \mathrm{e}^{S_n\phi(\underline{z})} \,. \tag{14}$$

To prove the "$\leq$" direction of this identity we just have to show that $\frac{1}{n}H_n(\nu) + \langle \phi, \nu \rangle \leq \frac{1}{n}P_n(\phi)$ for each $\nu \in \mathcal{M}_\sigma(X)$ or, after multiplying by n, $H_n(\nu) + \langle S_n\phi, \nu \rangle \leq P_n(\phi)$. But Jensen's inequality implies:

$$\begin{aligned} H_n(\nu) &+ \langle S_n\phi, \nu \rangle \\ &\leq \sum_{a_0,\ldots,a_{n-1}\in A} \nu[a_0 \ldots a_{n-1}] \\ &\quad \log\left(\frac{\sup\left\{\mathrm{e}^{S_n\phi(\underline{x})} : \underline{x} \in [a_0 \ldots a_{n-1}]\right\}}{\nu[a_0 \ldots a_{n-1}]} \right) \\ &\leq \log \sum_{a_0,\ldots,a_{n-1}\in A} \sup\left\{\mathrm{e}^{S_n\phi(\underline{x})} : \underline{x} \in [a_0 \ldots a_{n-1}]\right\} \\ &= \log \sum_{\underline{z}\in E_n} \mathrm{e}^{S_n\phi(\underline{z})} = P_n(\phi) \,. \end{aligned}$$

For the reverse inequality consider the discrete Gibbs distributions

$$\pi_n := \sum_{\underline{z}\in E_n} \delta_{\underline{z}} \exp(-P_n(\phi) + S_n\phi(\underline{z}))$$

on the finite sets E_n, where $\delta_{\underline{z}}$ denotes the unit point mass in $\underline{z}$. One might be tempted to think that all weak limit points of the measures π_n are already equilibrium states. But this need not be the case because there is no good reason that these limits are shift invariant. Therefore, one forces invariance of the limits by passing to measures μ_n defined by $\langle f, \mu_n \rangle := \langle \frac{1}{n}\sum_{k=0}^{n-1} f \circ \sigma^k, \pi_n \rangle$. Weak limits of these measures are obviously shift invariant, and a more involved estimate we do not present here shows that each such weak limit μ satisfies $h(\mu) + \langle \phi, \mu \rangle \geq P(\phi)$.

We note that the same arguments work for any other sequence of sets E_n which contain exactly one point from each cylinder. So there are many ways to approximate equilibrium states, and if there are more than one equilibrium state, there is generally no guarantee that the limit is always the same.

Nonuniqueness of Equilibrium States: An Example

Before we turn to sufficient conditions for the uniqueness of equilibrium states in the next section, we present one of the simplest nontrivial examples for nonuniqueness of equilibrium states. Motivated by the so-called Fisher–Felderhof droplet model of condensation in statistical mechanics [23,25], Hofbauer [31] studies an observable ϕ on $X = \{0, 1\}^{\mathbb{N}}$ defined as follows: Let (a_k) be a sequence of negative real numbers with $\lim_{k\to\infty} a_k = 0$. Set $s_k := a_0 + \cdots + a_k$. For $k \geq 1$ denote $M_k := \{\underline{x} \in X : x_0 = \cdots = x_{k-1} = 1, x_k = 0\}$ and $M_0 := \{\underline{x} \in X : x_0 = 0\}$, and define

$$\phi(\underline{x}) := a_k \quad \text{for } \underline{x} \in M_k \text{ and } \phi(11\ldots) = 0 \,.$$

Then $\phi\colon X \to \mathbb{R}$ is continuous, so that there exists at least one equilibrium state for ϕ. Hofbauer proves that there is more than one equilibrium state if and only if $\sum_{k=0}^{\infty} e^{s_k} = 1$ and $\sum_{k=0}^{\infty}(k+1)e^{s_k} < \infty$. In that case $P(\phi) = 0$, so one of these equilibrium states is the unit mass $\delta_{11\ldots}$, and we denote the other equilibrium state by μ_1, so $h(\mu_1) + \langle \phi, \mu_1 \rangle = 0$. In view of (13) the pressure function is not differentiable at ϕ.

What does the pressure function $\beta \mapsto P(\beta\phi)$ look like? As $h(\delta_{11\ldots}) + \langle \beta\phi, \delta_{11\ldots} \rangle = 0$ for all β, $P(\beta\phi) \geq 0$ for all β. Observe now that $\phi(\underline{x}) \leq 0$ with equality only for $\underline{x} = 11\ldots$ This implies that $\langle \phi, \mu \rangle < 0$ for all $\mu \in \mathcal{M}_\sigma(X)$ different from $\delta_{11\ldots}$. From this we can conclude:

- $P(\beta\phi) \leq P(\phi) = 0$ for $\beta > 1$, so $P(\beta\phi) = 0$ for $\beta \geq 1$.
- $P(\beta\phi) \geq h(\mu_1) + \langle \beta\phi, \mu_1 \rangle = h(\mu_1) + \langle \phi, \mu_1 \rangle - (1-\beta)\langle \phi, \mu_1 \rangle = -(1-\beta)\langle \phi, \mu_1 \rangle$.

It follows that, at $\beta = 1$, the derivative from the right of $P(\beta\phi)$ is zero, whereas the derivative from the left is at most $-\langle \phi, \mu_1 \rangle < 0$.

More on Equilibrium States

In more general dynamical systems the entropy function is not necessarily upper semicontinuous and hence equilibrium states need not exist, i. e., the supremum in (12) need not be attained by any invariant measure. A well-known sufficient property that guarantees the upper semicontinuity of the entropy function is the *expansiveness* of the system, see, e. g., [53]: a continuous transformation T of a compact metric space is *positively expansive*, if there is a constant $\gamma > 0$ such that for any two points x and y from the space there is some $n \in \mathbb{N}$ such that $T^n x$ and $T^n y$ are at least a distance γ apart. If T is a homeomorphism one says it is *expansive*, if the same holds for some $n \in \mathbb{Z}$. The previous results carry over without changes (although at the expense of more complicated proofs) to general expansive systems. The variational principle (14) holds in the very general context where T is a continuous action of $\mathbb{Z}_+^d$ on a compact Hausdorff space X. This was proved in [44] in a simple and elegant way. In the monograph [45] it is extended to amenable group actions.

The Gibbs Property: A Local Characterization of Equilibrium

In this section we are going to see that, for a sufficiently regular potential ϕ on a topologically mixing subshift of finite type, one has a unique equilibrium state which has the "Gibbs property". This property generalizes formula (5) that we derived for finite lattices. Subshifts of finite type are the symbolic models for Axiom A diffeomorphisms, as we shall see later on.

Subshifts of Finite Type

We start by recalling what is a subshift of finite type and refer the reader to ▶ Symbolic Dynamics or [43] for more details. Given a "transition matrix" $M = (M_{ab})_{a,b\in A}$ whose entries are 0's or 1's, one can define a subshift X_M as the set of all sequences $\underline{x} \in A^{\mathbb{N}}$ such that $M_{x_i x_{i+1}} = 1$ for all $i \in \mathbb{N}$. This is called a subshift of finite type or a topological Markov chain. We assume that there exists some integer p_0 such that M^p has strictly positive entries for all $p \geq p_0$. This means that M is irreducible and aperiodic. This property is equivalent to the property that the subshift of finite type is topologically mixing. A general subshift of finite type admits a decomposition into a finite union of transitive sets, each of which being a union of cyclically permuted sets on which the appropriate iterate is topologically mixing. In other words, topologically mixing subshifts of finite type are the building blocks of subshifts of finite type.

The Gibbs Property for a Class of Regular Potentials

The class of regular potentials we consider is that of "summable variations". We denote by $\mathrm{var}_k(\phi)$ the modulus of continuity of ϕ on cylinders of length $k \geq 1$, that is,

$$\mathrm{var}_k(\phi) := \sup\{|\phi(\underline{x}) - \phi(\underline{y})| : \underline{x} \in [y_0 \ldots y_{k-1}]\} .$$

If $\mathrm{var}_k(\phi) \to 0$ as $k \to \infty$, this means that ϕ is (uniformly) continuous with respect to the distance (7). We impose the stronger condition

$$\sum_{k=1}^{\infty} \mathrm{var}_k(\phi) < \infty . \tag{15}$$

We can now state the main result of this section.

The Gibbs state of a summable potential Let X_M be a topologically mixing subshift of finite type. Given a potential $\phi\colon X_M \to \mathbb{R}$ satisfying the summability condition (15), there is a (probability) measure μ_ϕ supported on X_M, that we call a Gibbs state. It is the unique σ-invariant measure which satisfies the following property:

There exists a constant $C > 0$ such that, for all $\underline{x} \in X_M$ and for all $n \geq 1$,

$$C^{-1} \leq \frac{\mu_\phi[x_0 \ldots x_{n-1}]}{\exp(S_n\phi(\underline{x}) - nP(\phi))} \leq C. \quad \text{(“Gibbs property”)} \tag{16}$$

Moreover, the Gibbs state μ_ϕ is ergodic and is also the unique equilibrium state of ϕ, i. e., the unique invariant measure for which the supremum in (12) is attained.

We now make several comments on this theorem.

- The Gibbs property (16) gives a *uniform control* of the measure of *all cylinders* in terms of their “energy”. This strengthens considerably the asymptotic equipartition property (11) that we recover if we restrict (16) to the set of μ_ϕ measure 1 where Birkhoff's ergodic Theorem (8) applies, and use the identity $\langle \phi, \mu_\phi \rangle - P(\phi) = -h(\mu_\phi)$.
- Gibbs measures on topologically mixing subshifts of finite type are ergodic (and actually mixing in a strong sense) as can be inferred from Ruelle's Perron–Frobenius Theorem see the next subsection.
- Suppose that there is another invariant measure μ' satisfying (16), possibly with a constant C' different from C. It is easy to verify that $\mu' = f\mu$ for some μ-integrable function f by using (16) and the Radon–Nikodym Theorem. Shift invariance imposes that, μ-a. e., $f = f \circ \sigma$. Then the ergodicity of μ implies that f is a constant μ-a. e., thus $\mu' = \mu$; see [6].
- One could define a Gibbs state by saying that it is an invariant measure μ satisfying (16) for a given continuous potential ϕ. If one does so, it is simple to verify that such a μ must also be an equilibrium state. Indeed, using (16), one can deduce that $\langle \phi, \mu \rangle + h(\mu) \geq P(\phi)$. The converse need not be true in general, see Subsect. “More on Hofbauer's Example” below. But the summability condition (15) is indeed sufficient for the coincidence of Gibbs and equilibrium states. A proof of this fact can be found in [58] or [38].

Ruelle's Perron–Frobenius Theorem

The powerful tool behind the theorem in the previous subsection is a far-reaching generalization of the classical Perron–Frobenius theorem for irreducible matrices. Instead of a matrix, one introduces the so-called transfer operator, also called the “Perron–Frobenius operator” or “Ruelle's operator”, which acts on a suitable Banach space of observables. It is D. Ruelle [52] who first introduced this operator in the context of one-dimensional lattice gases with exponentially decaying interactions. In our context, this corresponds to Hölder continuous potentials: these are potentials satisfying $\mathrm{var}_k(\phi) \leq c\theta^k$ for some $c > 0$ and $\theta \in (0, 1)$. A proof of “Ruelle's Perron–Frobenius Theorem” can be found in [4,6]. It was then extended to potentials with summable variations in [67]. We refer to the book of V. Baladi [1] for a comprehensive account on transfer operators in dynamical systems.

We content ourselves to define the transfer operator and state Ruelle's Perron–Frobenius Theorem. Let $\mathcal{L} : C(X_M) \to C(X_M)$ be defined by

$$\begin{aligned}(\mathcal{L}f)(\underline{x}) &:= \sum_{\underline{y} \in \sigma^{-1}\underline{x}} \mathrm{e}^{\phi(\underline{y})} f(\underline{y}) \\ &= \sum_{a \in A : M(a, x_0) = 1} \mathrm{e}^{\phi(a\underline{x})} f(a\underline{x}) \,.\end{aligned}$$

(Obviously, $a\underline{x} := ax_0x_1 \ldots$)

Theorem (Ruelle's Perron–Frobenius Theorem) *Let X_M be a topologically mixing subshift of finite type. Let ϕ satisfy condition (15). There exist a number $\lambda > 0$, $h \in C(X_M)$, and $\nu \in \mathcal{M}(X)$ such that $h > 0$, $\langle h, \nu \rangle = 1$, $\mathcal{L}h = \lambda h$, $\mathcal{L}^*\nu = \lambda\nu$, where $\mathcal{L}^*$ is the dual of $\mathcal{L}$. Moreover, for all $f \in C(X_M)$,*

$$\|\lambda^{-n}\mathcal{L}^n f - \langle f, \nu \rangle \cdot h\|_\infty \to 0, \quad \textit{as } n \to \infty \,.$$

By using this theorem, one can show that $\mu_\phi := h\nu$ satisfies (16) and $\lambda = \mathrm{e}^{P(\phi)}$.

Let us remark that for potentials which are such that $\phi(\underline{x}) = \phi(x_0, x_1)$ (i. e., potentials constant on cylinders of length 2), $\mathcal{L}$ can be identified with a $|A| \times |A|$ matrix and the previous theorem boils down to the classical Perron–Frobenius theorem for irreducible aperiodic matrices [63]. The corresponding Gibbs states are nothing but Markov chains with state space A (Chapter 3 in [29]). We shall take another point of view below (Subsect. “Markov Chains over Finite Alphabets”).

Relative Entropy

We now define the relative entropy of an invariant measure $\nu \in \mathcal{M}_\sigma(X_M)$ given a Gibbs state μ_ϕ as follows. We first define

$$\begin{aligned}&H_n(\nu|\mu_\phi) \\ &\quad := \sum_{a_0, \ldots, a_{n-1} \in A} \nu[a_0 \ldots a_{n-1}] \log \frac{\nu[a_0 \ldots a_{n-1}]}{\mu_\phi[a_0 \ldots a_{n-1}]}\end{aligned} \tag{17}$$

with the convention $0\log(0/0) = 0$. Now the relative entropy of ν given μ_ϕ is defined as

$$h(\nu|\mu_\phi) := \limsup_{n\to\infty} \frac{1}{n} H_n(\nu|\mu_\phi) .$$

(By applying Jensen's inequality, one verifies that $h(\nu|\mu_\phi) \geq 0$.) In fact the limit exists and can be computed quite easily using (16):

$$h(\nu|\mu_\phi) = P(\phi) - \langle\phi, \nu\rangle - h(\nu) . \tag{18}$$

To prove this formula, we first make the following observation. It can be easily verified that the inequalities in (16) remain the same when $S_n\phi$ is replaced by the "locally averaged" energy $\tilde{\phi}_n := (\nu[x_0 \ldots x_{n-1}])^{-1} \int_{[x_0\ldots x_{n-1}]} S_n\phi(\underline{y})\mathrm{d}\nu(\underline{y})$ for any cylinder with $\nu[x_0 \ldots x_{n-1}] > 0$. Cylinders with ν measure zero does not contribute to the sum in (17).

We can now write that

$$\begin{aligned} -\frac{1}{n}\log C \\ &\leq -\frac{1}{n} H_n(\nu|\mu_\phi) + \left(P(\phi) - \frac{1}{n}\langle S_n\phi, \nu\rangle - \frac{1}{n} H_n(\nu)\right) \\ &\leq \frac{1}{n}\log C . \end{aligned}$$

To finish we use that $\langle S_n\phi, \nu\rangle = n\langle\phi, \nu\rangle$ (by the invariance of ν) and we apply (9) to obtain

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} H_n(\nu|\mu_\phi) &= P(\phi) - \langle\phi, \nu\rangle - \lim_{n\to\infty} \frac{1}{n} H_n(\nu) \\ &= P(\phi) - \langle\phi, \nu\rangle - h(\nu) \end{aligned}$$

which proves (18).

The variational principle revisited We can reformulate the variational principle in the case of a potential satisfying the summability condition (15):

$$h(\nu|\mu_\phi) = 0 \quad \text{if and only if } \nu = \mu_\phi , \tag{19}$$

i. e., given μ_ϕ, the relative entropy $h(\cdot|\mu_\phi)$, as a function on $\mathcal{M}_\sigma(X_M)$, attains its minimum only at μ_ϕ.

Indeed, by (18) we have $h(\nu|\mu_\phi) = P(\phi) - \langle\phi, \nu\rangle - h(\nu)$. We now use (12) and the fact that μ_ϕ is the unique equilibrium state of ϕ to conclude.

More Properties of Gibbs States

Gibbs states enjoy very good statistical properties. Let us mention only a few. They satisfy the "Bernoulli property", a very strong qualitative mixing condition [4,6,67]. The sequence of random variables $(f \circ \sigma^n)_n$ satisfies the central limit theorem [15,16,49] and a large deviation principle if f is Hölder continuous [21,38,39,70]. Let us emphasize the central role played by relative entropy in large deviations. (The deep link between thermodynamics and large deviations is described in [42] in a much more general context.) Finally, the so-called "multifractal analysis" can be performed for Gibbs states, see, e. g., [48].

Examples on Shift Spaces

Measure of Maximal Entropy and Periodic Points

If the observable ϕ is constant zero, an equilibrium state simply maximizes the entropy. It is called *measure of maximal entropy*. The quantity $P(0) = \sup\{h(\nu) : \nu \in \mathcal{M}_\sigma(X)\}$ is called the *topological entropy* of the subshift $\sigma : X \to X$. When X is a subshift of finite type X_M with irreducible and aperiodic transition matrix M, there is a unique measure of maximal entropy, see, e. g., [43]. As a Gibbs state it satisfies (16). By summing over all cylinders $[x_0 \ldots x_{n-1}]$ allowed by M, it is easy to see that the topological entropy $P(0)$ is the asymptotic exponential growth rate of the number of sequences of length n that can occur as initial segments of points in X_M. This is obviously identical to the logarithm of the largest eigenvalue of the transition matrix M.

It is not difficult to verify that the total number of periodic sequences of period n equals the trace of the matrix M^n, i. e., we have the formula

$$\mathrm{Card}\left\{\underline{x} \in X_M : \sigma^n\underline{x} = \underline{x}\right\} = \mathrm{tr}(M^n) = \sum_{i=1}^{m} \lambda_i^n ,$$

where $\lambda_1, \ldots, \lambda_m$ are all the eigenvalues of M. Asymptotically, of course, $\mathrm{Card}\{\underline{x} \in X_M : \sigma^n\underline{x} = \underline{x}\} = \mathrm{e}^{nP(0)} + O(|\lambda'|^n)$, where λ' is the second largest (in absolute value) eigenvalue of M.

The measure of maximal entropy, call it μ_0, describes the distribution of periodic points in X_M: one can prove [3,37] that for any cylinder $B \subset X_M$

$$\lim_{n\to\infty} \frac{\mathrm{Card}\left\{\underline{x} \in B : \sigma^n\underline{x} = \underline{x}\right\}}{\mathrm{Card}\left\{\underline{x} \in X_M : \sigma^n\underline{x} = \underline{x}\right\}} = \mu_0(B) .$$

In other words, the finite atomic measures that assign equal weights $1/\mathrm{Card}\{\underline{x} \in X_M : \sigma^n\underline{x} = \underline{x}\}$ to each periodic point in X_M with period n weakly converges to μ_0, as $n \to \infty$. Each such measure has zero entropy while $h(\mu_0) = P(0) > 0$, so the entropy is not continuous on the space of invariant measures. It is, however, upper-semicontinuous (see Subsect. "Entropy as a Function of the Measure").

In fact, it is possible to approximate any Gibbs state μ_ϕ on X_M in a similar way by finite atomic measures on periodic orbits, if one assigns weights properly (see, e. g., Theorem 20.3.7 in [37]).

Markov Chains over Finite Alphabets

Let $Q = (q_{a,b})_{a,b\in A}$ be an irreducible stochastic matrix over the finite alphabet A. It is well known (see, e. g., [63]) that there exists a unique probability vector π on A that defines a stationary Markov measure ν_Q on $X = A^{\mathbb{N}}$ by $\nu_Q[a_0 \dots a_{n-1}] = \pi_{a_0} q_{a_0 a_1} \dots q_{a_{n-2} a_{n-1}}$. We are going to identify ν_Q as the *unique Gibbs distribution* $\mu \in \mathcal{M}_\sigma(X)$ *that maximizes entropy under the constraints* $\mu[ab] = \mu[a]q_{ab}$, *i. e.,* $\langle \phi^{ab}, \mu\rangle = 0$ $(a, b \in A)$, where $\phi^{ab} := \mathbb{1}_{[ab]} - q_{ab}\mathbb{1}_{[a]}$. Indeed, as μ is a Gibbs measure, there are $\beta_{ab} \in \mathbb{R}$ $(a, b \in A)$ and constants $P \in \mathbb{R}$, $C > 0$ such that

$$C^{-1} \le \frac{\mu[x_0 \dots x_{n-1}]}{\exp\left(\sum_{a,b\in A} \beta_{ab}\phi_n^{ab}(\underline{x}) - nP\right)} \le C \tag{20}$$

for all $\underline{x} \in A^{\mathbb{N}}$ and all $n \in \mathbb{N}$. Let $r_{ab} := \exp(\beta_{ab} - \sum_{b'\in A} \beta_{ab'} q_{ab'} - P)$. Then the denominator in (20) equals $r_{x_0 x_1} \dots r_{x_{n-2} x_{n-1}}$, and it follows that μ is equivalent to the stationary Markov measure defined by the (nonstochastic) matrix $(r_{ab})_{a,b\in A}$. As μ is ergodic, μ is this Markov measure, and as μ satisfies the linear constraints $\mu[ab] = \mu[a]q_{ab}$, we conclude that $\mu = \nu_Q$.

The Ising Chain

Here the task is to characterize all "spin chains" in $\underline{x} \in \{-1, +1\}^{\mathbb{N}}$ (or, more commonly, $\{-1, +1\}^{\mathbb{Z}}$) which are as random as possible with the constraint that two adjacent spins have a prescribed probability $p \ne \frac{1}{2}$ to be identical. With $\phi(\underline{x}) := x_0 x_1$ this is equivalent to requiring that $\underline{x}$ is typical for a Gibbs distribution $\mu_{\beta\phi}$ where $\beta = \beta(p)$ is such that $\langle \phi, \mu_{\beta\phi}\rangle = 2p - 1$. It follows that there is a constant $C > 0$ such that for each $n \in \mathbb{N}$ and any two "spin patterns" $\underline{a} = a_0 \dots a_{n-1}$ and $\underline{b} = b_0 \dots b_{n-1}$

$$\left|\log \frac{\mu_{\beta\phi}[a_0 \dots a_{n-1}]}{\mu_{\beta\phi}[b_0 \dots b_{n-1}]} - \beta(N_{\underline{a}} - N_{\underline{b}})\right| \le C\,,$$

where $N_{\underline{a}}$ and $N_{\underline{b}}$ are the numbers of identical adjacent spins in $\underline{a}$ and $\underline{b}$, respectively.

More on Hofbauer's Example

We come back to the example described in Subsect. "Nonuniqueness of Equilibrium States: An Example". It is easy to verify that in that example $\mathrm{var}_{k+1}(\phi) = |a_k|$. For instance, if $a_k = -1/(k+1)^2$ there is a unique Gibbs/equilibrium state. If $a_k = -3\log((k+1)/k)$ for $k \ge 1$ and $a_0 = -\log\sum_{j=1}^{\infty} j^{-3}$, then from [31] we know that ϕ admits more than one equilibrium state, one of them being $\delta_{11\dots}$, which cannot be a Gibbs state for any continuous ϕ.

Examples from Differentiable Dynamics

In this section we present a number of examples to which the general theory developed above does not apply directly but only after a transfer of the theory from a symbolic space to a manifold. We restrict to examples where the *results* can be transferred because those aspects of the smooth dynamics we focus on can be studied as well on a shift dynamical system that is obtained from the original one via symbolic coding. (We do not discuss the coding process itself which is sometimes far from trivial, but we focus on the application of the Gibbs and equilibrium theory.) There are alternative approaches where instead of the results the *concepts and (partly) the strategies of proofs* are transferred to the smooth dynamical systems. This has led both to an extension of the range of possible applications of the theory and to a number of refined results (because some special features of smooth systems necessarily get lost by transferring the analysis to a completely disconnected metric space).

In the following examples, T denotes a (possibly piecewise) differentiable map of a compact smooth manifold M. Points on the manifold are denoted by u and v. In all examples there is a Hölder continuous coding map $\pi : X \to M$ from a subshift of finite type X onto the manifold which respects the dynamics, i. e., $T \circ \pi = \pi \circ \sigma$. This *factor map* π is "nearly" invertible in the sense that the set of points in M with more than one preimage under π has measure zero for all T-invariant measures we are interested in. Hence such measures $\tilde{\mu}$ on M correspond unambiguously to shift invariant measures $\mu = \tilde{\mu} \circ \pi^{-1}$. Similarly observables $\tilde{\phi}$ on M and $\phi = \tilde{\phi} \circ \pi$ on X are related.

Uniformly Expanding Markov Maps of the Interval

A transformation T on $M := [0, 1]$ is called a *Markov map*, if there are $0 = u_0 < u_1 < \dots < u_N = 1$ such that each restriction $T|_{(u_{i-1},u_i)}$ is strictly monotone, C^{1+r} for some $r > 0$, and maps (u_{i-1}, u_i) onto a union of some of these N monotonicity intervals. It is called *uniformly expanding* if there is some $k \in \mathbb{N}$ such that $\lambda := \inf_x |(T^k)'(x)| > 1$. It is not difficult to verify that the symbolic coding of such a system leads to a topological Markov chain over the alphabet $A = \{1, \dots, N\}$. To simplify the discussion we assume that the transition ma-

trix M of this topological Markov chain is irreducible and aperiodic.

Our goal is to find a T-invariant measure $\tilde{\mu}$ represented by $\mu \in \mathcal{M}_\sigma(X_M)$ which minimizes the relative entropy to Lebesgue measure on $[0,1]$

$$h(\tilde{\mu}|m) := \lim_{n\to\infty} \frac{1}{n} \sum_{a_0,\ldots,a_{n-1}\in\{1,\ldots,N\}} \mu[a_0\ldots a_{n-1}] \log \frac{\mu[a_0\ldots a_{n-1}]}{\nu_n[a_0\ldots a_{n-1}]} ,$$

where $\nu_n[a_0\ldots a_{n-1}] := |I_{a_0\ldots a_{n-1}}|$. (Recall that, without insisting on invariance, this would just be the Lebesgue measure itself.) The existence of the limit will be justified below – observe that m is not a Gibbs state as ν is in Eq. (17). The argument rests on the simple observation (implied by the uniform expansion and the piecewise Hölder-continuity of T') that T has *bounded distortion*, i. e., that there is a constant $C > 0$ such that for all $n \in \mathbb{N}$, $a_0\ldots a_{n-1} \in \{1,\ldots,N\}^n$ and $u \in I_{a_0\ldots a_{n-1}}$ holds

$$\begin{aligned} &C^{-1} \le |I_{a_0\ldots a_{n-1}}| \cdot |(T^n)'(u)| \le C\,, \quad \text{or, equivalently,} \\ &C^{-1} \le \frac{|I_{a_0\ldots a_{n-1}}|}{\exp\left(S_n\tilde{\phi}(u)\right)} \le C\,, \end{aligned} \tag{21}$$

where $\tilde{\phi}(u) := -\log|T'(u)|$. (Observe the similarity between this property and the Gibbs property (16).) Assuming bounded distortion we have at once

$$\begin{aligned} h(\tilde{\mu}|m) &= \lim_{n\to\infty} \frac{1}{n}\left(-H_n(\mu) - \sum_{k=0}^{n-1}\langle\phi\circ\sigma^k,\mu\rangle\right) \\ &= -h(\mu) - \langle\phi,\mu\rangle\,, \end{aligned}$$

and minimizing this relative entropy just amounts to maximizing $h(\mu) + \langle\phi,\mu\rangle$ for $\phi = -\log|T'|\circ\pi$. As the results on Gibbs distributions from Sect. "The Gibbs Property: A Local Characterization of Equilibrium" apply, we conclude that

$$C^{-1} \le \frac{\mu[a_0\ldots a_{n-1}]}{|I_{a_0}\ldots a_{n-1}|} \le C$$

for some $C > 0$. So the unique T-invariant measure $\tilde{\mu}$ that minimizes the relative entropy $h(\tilde{\mu}|m)$ is equivalent to Lebesgue measure m. (The existence of an invariant probability measure equivalent to m is well known, also without invoking entropy theory. It is guaranteed by a "Folklore Theorem" [33].)

Interval Maps with an Indifferent Fixed Point

The presence of just one point $x \in [0,1]$ such that $T'(x) = 1$ dramatically changes the properties of the system. A canonical example is the map $T_\alpha : x \mapsto x(1+2^\alpha x^\alpha)$ if $x \in [0,1/2[$ and $x \mapsto 2x-1$ if $x \in [1/2,1]$. We have $T'(0) = 1$, i. e., 0 is an indifferent fixed point. For $\alpha \in [0,1[$ this map admits an absolutely continuous invariant probability measure $d\mu(x) = h(x)dx$, where $h(x) \sim x^{-\alpha}$ when $x \to 0$ [66]. In the physics literature, this type of map is known as the "Manneville–Pomeau" map. It was introduced as a model of transition from laminar to intermittent behavior [50]. In [28] the authors construct a piecewise affine version of this map to study the complexity of trajectories (in the sense of Subsect. "A Short Digression on Complexity"). This gives rise to a countable state Markov chain. In [69] the close connection to the Fisher–Felderhof model and Hofbauer's example (see Subsect. "Nonuniqueness of Equilibrium States: An Example") was realized. We refer to [61] for recent developments and a list of references.

Axiom A Diffeomorphisms, Anosov Diffeomorphisms, Sinai–Ruelle–Bowen Measures

The first spectacular application of the theory of Gibbs measures to differentiable dynamical systems was Sinai's approach to Anosov diffeomorphisms via Markov partitions [64] that allowed one to code the dynamics of these maps into a subshift of finite type and to study their invariant measures by methods from equilibrium statistical mechanics [65] that had been developed previously by Dobrushin, Lanford, and Ruelle [17,18,19,20,41]. Not much later this approach was extended by Bowen [2] to Smale's Axiom A diffeomorphisms (and to Axiom A flows by Bowen and Ruelle [7]); see also [54]. The interested reader can consult, e. g., [71] for a survey, and either [6] or [15] for details.

Both types of diffeomorphisms act on a smooth compact Riemannian manifold M and are characterized by the existence of a compact T-invariant *hyperbolic set* $\Lambda \subseteq M$. Their basic properties are described in detail in the contribution ► Ergodic Theory: Basic Examples and Constructions. Very briefly, the tangent bundle over Λ splits into two invariant subbundles – a stable one and an unstable one. Correspondingly, through each point of Λ there passes a local stable and a local unstable manifold which are both tangent to the respective subspaces of the local tangent space. The unstable derivative of T, i. e., the derivative DT restricted to the unstable subbundle, is uniformly expanding. Its Jacobian determinant, denoted by $J^{(u)}$, is Hölder continuous as a function on Λ. Hence the

observable $\phi^{(u)} := -\log|J^{(u)}| \circ \pi$ is Hölder continuous, and the Gibbs and equilibrium theory apply (via the symbolic coding) to the diffeomorphism T (modulo possibly a decomposition of the hyperbolic set into irreducible and aperiodic components, called basic sets, that can be modeled by topologically mixing subshifts of finite type). The main results are:

Characterization of attractors The following assertions are equivalent for a basic set $\Omega \subseteq \Lambda$:

(i) Ω is an attractor, i. e., there are arbitrarily small neighborhoods $U \subseteq M$ of Ω such that $TU \subset U$.
(ii) The union of all stable manifolds through points of Ω is a subset of M with positive volume.
(iii) The pressure $P_{T|_\Omega}(\phi^{(u)}) = 0$.

In this case the unique equilibrium and Gibbs state μ^+ of $T|_\Omega$ is called the *Sinai–Ruelle–Bowen (SRB) measure* of $T|_\Omega$. It is uniquely characterized by the identity $h_{T|_\Omega}(\mu^+) = -\langle\phi^{(u)}, \mu^+\rangle$. (For all other T-invariant measures on Ω one has "<" instead of "=".)

Further properties of SRB measures Suppose $P_{T|_\Omega}(\phi^{(u)}) = 0$ and let μ^+ be the SRB measure.

(a) For a set of points $u \in M$ of positive volume we have:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k u) = \langle f, \mu^+\rangle \,.$$

(Indeed, because of (ii) of the above characterization, this holds for almost all points of the union of the stable manifolds through points of Ω.)
(b) Conditioned on unstable manifolds, μ^+ is absolutely continuous to the volume measure on unstable manifolds.

In the special case of transitive Anosov diffeomorphisms, the whole manifold is a hyperbolic set and $\Omega = M$. Because of transitivity, property (ii) from the characterization of attractors is trivially satisfied, so there is always a unique SRB measure μ^+. As T^{-1} is an Anosov diffeomorphism as well – only the roles of stable and unstable manifolds are interchanged – T^{-1} has a unique SRB measure μ^- which is the unique equilibrium state of T^{-1} (and hence also of T) for $\phi^{(s)} := \log|J^{(s)}|$. One can show:

SRB measures for Anosov diffeomorphisms The following assertions are equivalent:

(i) $\mu^+ = \mu^-$.
(ii) μ^+ or μ^- is absolutely continuous w.r.t the volume measure on M.
(iii) For each periodic point $u = T^n u \in M$, $|J(u)| = 1$, where J denotes the determinant of DT.

We remark that, similarly to the case of Markov interval maps, the unstable Jacobian of T^n at u is asymptotically equivalent to the volume of the "n-cylinder" of the Markov partition around u. So the maximization of $h(\mu) + \langle\phi^{(u)}, \mu\rangle$ by the SRB measure μ^+ can again be interpreted as the minimization of the relative entropy of invariant measures with respect to the normalized volume, and the fact that $P(\phi^{(u)}) = 0$ in the Anosov (or more generally attractor) case means that μ^+ is as close to being absolutely continuous as it is possible for a singular measure. This is reflected by the above properties (a) and (b).

We emphasize the meaning of property (a) above: it tells us that the SRB measure μ^+ is the only *physically observable* measure. Indeed, in numerical experiments with physical models, one picks an initial point $u \in M$ "at random" (i. e., with respect to the volume or Lebesgue measure) and follows its orbit $T^k u$, $k \geq 0$.

Bowen's Formula for the Hausdorff Dimension of Conformal Repellers

Just as nearby orbits converge towards an attractor, they diverge away from a repeller. Conformal repellers form a nice class of systems which can be coded by a subshift of finite type. The construction of their Markov partitions is much simpler than that of Anosov diffeomorphisms, see, e. g., [72].

Let us recall the definition of a conformal repeller before giving a fundamental example. Given a holomorphic map $T\colon V \to \mathbb{C}$ where $V \subset \mathbb{C}$ is open and J a compact subset of $\mathbb{C}$, one says that (J, V, T) is a conformal repeller if

(i) there exist $C > 0$, $\alpha > 1$ such that $|(T^n)'(z)| \geq C\alpha^n$ for all $z \in J$, $n \geq 1$;
(ii) $J = \bigcap_{n\geq 1} T^{-n}(V)$, and
(iii) for any open set U such that $U \cap J \neq \emptyset$, there exists n such that $T^n(U \cap J) \supset J$.

From the definition it follows that $T(J) = J$ and $T^{-1}(J) = J$.

A fundamental example is the map $T\colon z \to z^2 + c$, $c \in \mathbb{C}$ being a parameter. It can be shown that for $|c| < \frac{1}{4}$ there exists a compact set J, called a (hyperbolic) *Julia set*, such that $(J, \mathbb{C}, T)$ is a conformal repeller.

Conformal repellers J are in general fractal sets and one can measure their "degree of fractality" by means of their Hausdorff dimension, $\dim_{\mathrm{H}}(J)$. Roughly speaking, one computes this dimension by covering the set J by balls

with radius less than or equal to δ. If $N_\delta(J)$ denotes the cardinality of the smallest such covering, then we expect that

$$N_\delta(J) \sim \delta^{-\dim_{\mathrm{H}}(J)}, \quad \text{as } \delta \to 0 .$$

We refer the reader to ► Ergodic Theory: Fractal Geometry or [22,46] for a rigorous definition (based on Carathéodory's construction) and for more information on fractal geometry.

Bowen's formula relates $\dim_{\mathrm{H}}(J)$ to the unique zero of the pressure function $\beta \mapsto P(\beta\tilde{\phi})$ where $\tilde{\phi} := -(\log|T'|)|_J$. It is not difficult to see that indeed this map has a unique zero for some positive β.

By property (i), $S_n\tilde{\phi} \le \text{const} - n\log\alpha$, which implies (by (13)) that $\frac{\mathrm{d}}{\mathrm{d}\beta}P(\beta\tilde{\phi}) = \langle\tilde{\phi}, \mu_\beta\rangle \le -\log\alpha < 0$. As $P(0)$ equals the topological entropy of J, i. e., the logarithm of the largest eigenvalue of the matrix M associated to the Markov partition, $P(0)$ is strictly positive. Therefore, (recall that the pressure function is continuous) there exists a unique number $\beta_0 > 0$ such that $P(\beta_0\tilde{\phi}) = 0$.

It turns out that this unique zero is precisely $\dim_{\mathrm{H}}(J)$:

Bowen's formula The Hausdorff dimension of J is the unique solution of the equation $P(\beta\tilde{\phi}) = 0, \beta \in \mathbb{R}$; in particular

$$P(\dim_{\mathrm{H}}(J)\tilde{\phi}) = 0 .$$

This formula was proven in [55] for a general class of conformal repellers after the seminal paper [5]. The main tool is a distortion estimate very similar to (21). A simple exposition can be found in [72].

Nonequilibrium Steady States and Entropy Production

SRB measures for Anosov diffeomorphisms and Axiom A attractors have been accepted recently as conceptual models for *nonequilibrium steady states* in nonequilibrium statistical mechanics. Let us point out that the word "equilibrium" is used in physics in a much more restricted sense than in ergodic theory. Only diffeomorphisms preserving the natural volume of the manifold (or a measure equivalent to the volume) would be considered as appropriate toy models of physical equilibrium situations. In the case of Anosov diffeomorphisms this is precisely the case if the "forward" and "backward" SRB measures μ^+ and μ^- coincide. Otherwise, the diffeomorphism models a situation out of equilibrium, and the difference between μ^+ and μ^- can be related to entropy production and irreversibility.

Gallavotti and Cohen [26,27] introduced SRB measures as idealized models of nonequilibrium steady states around 1995. In order to have as firm a mathematical basis as possible they made the "*chaotic hypothesis*" that the systems they studied behave like transitive Anosov systems. Ruelle [56] extended their approach to more general (even nonuniformly) hyperbolic dynamics; see also his reviews [57,59] for more recent accounts discussing also a number of related problems; see by [51], too. The importance of the Gibbs property of SRB measures for the discussion of entropy production was also highlighted in [35], where it is shown that for transitive Anosov diffeomorphisms the relative entropy $h(\mu^+|\mu^-)$ equals the average entropy production rate $\langle\log|J|, \mu^+\rangle$ of μ^+ where J denotes again the Jacobian determinant of the diffeomorphism. In particular, the entropy production rate is zero if, and only if, $h(\mu^+|\mu^-) = 0$, i. e., using coding and (19), if, and only if, $\mu^+ = \mu^-$. According to Subsect. "Axiom A Diffeomorphisms, Anosov Diffeomorphisms, Sinai–Ruelle–Bowen Measures", this is also equivalent to μ^+ or μ^- being absolutely continuous with respect to the volume measure.

Some Ongoing Developments and Future Directions

As we saw, many dynamical systems with uniform hyperbolic structure (e. g., Anosov maps, axiom A diffeomorphisms) can be modeled by subshifts of finite type over a finite alphabet. We already mentioned in Subsect. "Interval Maps with an Indifferent Fixed Point" the typical example of a map of the interval with an indifferent fixed point, whose symbolic model is still a subshift of finite type, but with a countable alphabet. The thermodynamic formalism for such systems is by now well developed [24,30,60,61,62] and used, e. g., for multidimensional piecewise expanding maps [13]. An active line of research is related to systems admitting representations by symbolic models called "towers" constructed by using "inducing schemes". The fundamental example is the class of one-dimensional unimodal maps satisfying the "Collet–Eckmann condition". A first attempt to develop thermodynamic formalism for such systems was made in [10] where existence and uniqueness of equilibrium measures for the potential function $\tilde{\phi}_\beta(u) = -\beta\log|T'(u)|$ with β close to 1 was established. Very recently, new developments in this direction appeared, see, e. g., [11,12,47].

A largely open field of research concerns a new branch of nonequilibrium statistical mechanics, the so-called "chaotic scattering theory", namely the analysis of chaotic systems with various openings or holes in phase space, and the corresponding repellers on which interesting invariant measures exist. We refer the reader to [15] for a brief account and references to the physics litera-

ture. The existence of (generalized) steady states on repellers and the so-called "escape rate formula" have been observed numerically in a number of models. So far, little has been proven mathematically, except for Anosov diffeomorphisms with special holes [15] and for certain nonuniformly hyperbolic systems [9].

Bibliography

1. Baladi V (2000) Positive transfer operators and decay of correlations. Advanced Series in Nonlinear Dynamics, vol 16. World Scientific, Singapore
2. Bowen R (1970) Markov partitions for Axiom A diffeomorphisms. Amer J Math 92:725–747
3. Bowen R (1974/1975) Some systems with unique equilibrium states. Math Syst Theory 8:193–202
4. Bowen R (1974/75) Bernoulli equilibrium states for Axiom A diffeomorphisms. Math Syst Theory 8:289–294
5. Bowen R (1979) Hausdorff dimension of quasicircles. Inst Hautes Études Sci Publ Math 50:11–25
6. Bowen R (2008) Equilibrium states and the ergodic theory of Anosov diffeomorphisms. Lecture Notes in Mathematics, vol 470, 2nd edn (1st edn 1975). Springer, Berlin
7. Bowen R, Ruelle D (1975) The ergodic theory of Axiom A flows. Invent Math 29:181–202
8. Brudno AA (1983) Entropy and the complexity of the trajectories of a dynamical system. Trans Mosc Math Soc 2:127–151
9. Bruin H, Demers M, Melbourne I (2007) Existence and convergence properties of physical measures for certain dynamical systems with holes. Preprint
10. Bruin H, Keller G (1998) Equilibrium states for S-unimodal maps. Ergodic Theory Dynam Syst 18(4):765–789
11. Bruin H, Todd M (2007) Equilibrium states for potentials with $\sup\varphi - \inf\varphi < h_{top}(f)$. Commun Math Phys doi:10.1007/s00220-0-008-0596-0
12. Bruin H, Todd M (2007) Equilibrium states for interval maps: the potential $-t\log|Df|$. Preprint
13. Buzzi J, Sarig O (2003) Uniqueness of equilibrium measures for countable Markov shifts and multidimensional piecewise expanding maps. Ergod Theory Dynam Syst 23(5):1383–1400
14. Chaitin GJ (1987) Information, randomness & incompleteness. Papers on algorithmic information theory. World Scientific Series in Computer Science, vol 8. World Scientific, Singapore
15. Chernov N (2002) Invariant measures for hyperbolic dynamical systems. In: Handbook of Dynamical Systems, vol 1A. North-Holland, pp 321–407
16. Coelho Z, Parry W (1990) Central limit asymptotics for shifts of finite type. Isr J Math 69(2):235–249
17. Dobrushin RL (1968) The description of a random field by means of conditional probabilities and conditions of its regularity. Theory Probab Appl 13:197–224
18. Dobrushin RL (1968) Gibbsian random fields for lattice systems with pairwise interactions. Funct Anal Appl 2:292–301
19. Dobrushin RL (1968) The problem of uniqueness of a Gibbsian random field and the problem of phase transitions. Funct Anal Appl 2:302–312
20. Dobrushin RL (1969) Gibbsian random fields. The general case. Funct Anal Appl 3:22–28
21. Eizenberg A, Kifer Y, Weiss B (1994) Large deviations for $\mathbb{Z}^d$-actions. Comm Math Phys 164(3):433–454
22. Falconer K (2003) Fractal geometry. Mathematical foundations and applications, 2nd edn. Wiley, San Francisco
23. Fisher ME, Felderhof BU (1970) Phase transition in one-dimensional clusterinteraction fluids: IA. Thermodynamics, IB. Critical behavior. II. Simple logarithmic model. Ann Phys 58:177–280
24. Fiebig D, Fiebig U-R, Yuri M (2002) Pressure and equilibrium states for countable state Markov shifts. Isr J Math 131:221–257
25. Fisher ME (1967) The theory of condensation and the critical point. Physics 3:255–283
26. Gallavotti G (1996) Chaotic hypothesis: Onsager reciprocity and fluctuation-dissipation theorem. J Stat Phys 84:899–925
27. Gallavotti G, Cohen EGD (1995) Dynamical ensembles in stationary states. J Stat Phys 80:931–970
28. Gaspard P, Wang X-J (1988) Sporadicity: Between periodic and chaotic dynamical behaviors. In: Proceedings of the National Academy of Sciences USA, vol 85, pp 4591–4595
29. Georgii H-O (1988) Gibbs measures and phase transitions. In: de Gruyter Studies in Mathematics, 9. de Gruyter, Berlin
30. Gurevich BM, Savchenko SV (1998) Thermodynamic formalism for symbolic Markov chains with a countable number of states. Russ Math Surv 53(2):245–344
31. Hofbauer F (1977) Examples for the nonuniqueness of the equilibrium state. Trans Amer Math Soc 228:223–241
32. Israel R (1979) Convexity in the theory of lattice gases. Princeton Series in Physics. Princeton University Press
33. Jakobson M, Świątek (2002) One-dimensional maps. In: Handbook of Dynamical Systems, vol 1A. North-Holland, Amsterdam, pp 321–407
34. Jaynes ET (1989) Papers on probability, statistics and statistical physics. Kluwer, Dordrecht
35. Jiang D, Qian M, Qian M-P (2000) Entropy production and information gain in Axiom A systems. Commun Math Phys 214:389–409
36. Katok A (2007) Fifty years of entropy in dynamics: 1958–2007. J Mod Dyn 1(4):545–596
37. Katok A, Hasselblatt B (1995) Introduction to the modern theory of dynamical systems. Encyclopaedia of Mathematics and its Applications, vol 54. Cambridge University Press, Cambridge
38. Keller G (1998) Equilibrium states in ergodic theory. In: London Mathematical Society Student Texts, vol 42. Cambridge University Press, Cambridge
39. Kifer Y (1990) Large deviations in dynamical systems and stochastic processes. Trans Amer Math Soc 321:505–524
40. Kolmogorov AN (1983) Combinatorial foundations of information theory and the calculus of probabilities. Uspekhi Mat Nauk 38:27–36
41. Lanford OE, Ruelle D (1969) Observables at infinity and states with short range correlations in statistical mechanics. Commun Math Phys 13:194–215
42. Lewis JT, Pfister C-E (1995) Thermodynamic probability theory: some aspects of large deviations. Russ Math Surv 50:279–317
43. Lind D, Marcus B (1995) An introduction to symbolic dynamics and coding. Cambridge University Press, Cambridge
44. Misiurewicz M (1976) A short proof of the variational principle for a $\mathbb{Z}_+^N$ action on a compact space. In: International Conference on Dynamical Systems in Mathematical Physics (Rennes, 1975), Astérisque, No. 40, Soc Math France, pp 147–157
45. Moulin-Ollagnier J (1985) Ergodic Theory and Statistical Mechanics. In: Lecture Notes in Mathematics, vol 1115. Springer, Berlin

46. Pesin Y (1997) Dimension theory in dynamical systems. Contemporary views and applications. University of Chicago Press, Chicago
47. Pesin Y, Senti S (2008) Equilibrium Measures for Maps with Inducing Schemes. J Mod Dyn 2(3):1–31
48. Pesin Y, Weiss (1997) The multifractal analysis of Gibbs measures: motivation, mathematical foundation, and examples. Chaos 7(1):89–106
49. Pollicott M (2000) Rates of mixing for potentials of summable variation. Trans Amer Math Soc 352(2):843–853
50. Pomeau Y, Manneville P (1980) Intermittent transition to turbulence in dissipative dynamical systems. Comm Math Phys 74(2):189–197
51. Rondoni L, Mejía-Monasterio C (2007) Fluctuations in nonequilibrium statistical mechanics: models, mathematical theory, physical mechanisms. Nonlinearity 20(10):R1–R37
52. Ruelle D (1968) Statistical mechanics of a one-dimensional lattice gas. Commun Math Phys 9:267–278
53. Ruelle D (1973) Statistical mechanics on a compact set with $\mathbb{Z}^{\nu}$ action satisfying expansiveness and specification. Trans Amer Math Soc 185:237–251
54. Ruelle D (1976) A measure associated with Axiom A attractors. Amer J Math 98:619–654
55. Ruelle D (1982) Repellers for real analytic maps. Ergod Theory Dyn Syst 2(1):99–107
56. Ruelle D (1996) Positivity of entropy production in nonequilibrium statistical mechanics. J Stat Phys 85:1–23
57. Ruelle D (1998) Smooth dynamics and new theoretical ideas in nonequilibrium statistical mechanics. J Stat Phys 95:393–468
58. Ruelle D (2004) Thermodynamic formalism: The mathematical structures of equilibrium statistical mechanics. Second edition. Cambridge Mathematical Library. Cambridge University Press, Cambridge
59. Ruelle D (2003) Extending the definition of entropy to nonequilibrium steady states. Proc Nat Acad Sc 100(6):3054–3058
60. Sarig O (1999) Thermodynamic formalism for countable Markov shifts. Ergod Theory Dyn Syst 19(6):1565–1593
61. Sarig O (2001) Phase transitions for countable Markov shifts. Comm Math Phys 217(3):555–577
62. Sarig O (2003) Existence of Gibbs measures for countable Markov shifts. Proc Amer Math Soc 131(6):1751–1758
63. Seneta E (2006) Non-negative matrices and Markov chains. In: Springer Series in Statistics. Springer
64. Sinai Ja G (1968) Markov partitions and C-diffeomorphisms. Funct Anal Appl 2:61–82
65. Sinai Ja G (1972) Gibbs measures in ergodic theory. Russ Math Surv 27(4):21–69
66. Thaler M (1980) Estimates of the invariant densities of endomorphisms with indifferent fixed points. Isr J Math 37(4):303–314
67. Walters P (1975) Ruelle's operator theorem and *g*-measures. Trans Amer Math Soc 214:375–387
68. Walters P (1992) Differentiability properties of the pressure of a continuous transformation on a compact metric space. J Lond Math Soc (2) 46(3):471–481
69. Wang X-J (1989) Statistical physics of temporal intermittency. Phys Rev A 40(11):6647–6661
70. Young L-S (1990) Large deviations in dynamical systems. Trans Amer Math Soc 318:525–543
71. Young L-S (2002) What are SRB measures, and which dynamical systems have them? Dedicated to David Ruelle and Yasha Sinai on the occasion of their 65th birthdays. J Statist Phys 108(5–6):733–754
72. Zinsmeister M (2000) Thermodynamic formalism and holomorphic dynamical systems. SMF/AMS Texts and Monographs, 2. American Mathematical Society

Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation

Mie Ichihara[1], Takeshi Nishimura[2]
[1] Earthquake Research Institute, University of Tokyo, Tokyo, Japan
[2] Tohoku University, Sendai, Japan

Article Outline

Glossary
Definition of the Subject
Introduction
Elementary Processes in a Single-Bubble Dynamics
Bubbly Magma in an Elastic Rock as a Pressure Source
Acoustic Bubbles in Hydrothermal Systems
Future Directions
Acknowledgments
Bibliography

Glossary

Impulse The word *impulse* is used in many areas in different ways. In classical mechanics, the impulse is the integral of force with respect to time. It is also used to refer to a fast-acting force, which is often idealized by a step function or a delta function. In this text, it is used to represent any functional form of pressure increase, either static or transient, which can generate observable signals.

Magma, melt, liquid Magma is a general name for molten rock. It is fluid but contains solid and gas inclusions in liquid matrix. The matrix in magma is silicate melt (which is often called just melt), and that in a hydrothermal system is water.

Volatile Volatile is compound in silicate melt. The major component is H_2O, of which concentration is 1–5 wt% depending mainly on pressure and composition of the melt. It exsolves from melt and forms gas bubbles at relatively low pressure (ca. 100 MPa corresponding to the litho-static pressure around several kilo-meters). The second major component is CO_2. Although its

concentration is usually several ppm, some kinds of melts may dissolve 3–30 wt% of CO_2 at several GPa.

Long period seismic events Long-period (or very long-period) seismic events are dominant in the period from about 1 s to more than a few tens of seconds. These signals at volcanoes are considered to be generated by interaction or resonance between volcanic fluid and the surrounding medium.

Ground deformation Ground deformation is often observed at volcanoes when magma chambers inflate or deflate. Such ground deformation is detected by geodetic measurements such as GPS, tilt or strain meters, and the deformations often continue for a few tens of minutes to days or even months.

Magma chamber A magma chamber is a storage system of molten magma. It is generally hard to detect, but is probably located at from a shallow depth (ca. 1 km) to a few tens of km beneath the volcanoes. The shape and size have not been confirmed yet, but it is usually assumed to be rather round and hundreds to thousands of meters in scale. A magma storage system which has a horizontal extent is called a sill, and one which has a vertical extent is called a dike.

Rectified diffusion and rectified heat transfer Rectified diffusion is a mechanism which can push dissolved volatiles into bubbles in a sound field. Bubbles take in more volatiles during expansion than they discharge during contraction, mainly because of the following two non-linear effects. Firstly, during expansion the bubble radius becomes larger so that the bubble surface is also larger than the surface during contraction. Secondly, radial bubble expansion tangentially stretches the diffusion layer and sharpens the radial gradient of the volatile concentration in the diffusion layer, so that the volatile flux into the bubble. The mechanism also works to push heat into bubbles and enhances evaporation in a liquid-vapor system. The rectified diffusion and heat transfer have been known and studied in mechanical and chemical engineering.

Bubble collapse When the bubble is compressed, oscillates, or loses its mass by diffusion or phase change, it contracts to a very small size and sometimes disappear. Bubble collapse indicates the contraction of a bubble and does not necessarily indicate its disappearance.

Definition of the Subject

A volcano consists of solids, liquids, gases, and intermediate materials of any two of these phases. Mechanical and thermo-dynamical interactions of these phases are essential in generating a variety of volcanic activities. In particular, the gas phase is mechanically distinct from the other phases and plays important roles in the dynamic phenomena of volcanoes. When we work on volcanic activities, we are almost certainly confronted with physics problems associated with bubbles.

The roles of bubbles in volcanic activities have been investigated mainly in three aspects. Firstly, the nucleation, growth, and expansion of bubbles is considered to be the main force that brings the magma to the surface [62,88]. Secondly, a single bubble, if it is sufficiently large, may generate seismic waves when it rapidly expands or accelerates in the volcanic conduit [11,35,81,95], and may generate acoustic waves in the air when it oscillates or bursts at the magma surface [36,82,95,96]. Thirdly, the existence of bubbles can significantly reduce the sound velocity [16,

Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Table 1
List of important variables and constants

Notation	Unit	Definition
C_{eq}	–	Equilibrium volatile concentration (weight fraction) in the liquid
c_l	$m\,s^{-1}$	Sound speed in the liquid
c_{pg}	$J\,kg^{-1}K^{-1}$	Heat capacity of the gas at constant pressure
c_{pl}	$J\,kg^{-1}K^{-1}$	Heat capacity of the liquid at constant pressure
K_g	Pa	Effective bulk modulus of the bubble
K_l	Pa	Bulk modulus of the liquid
L	$J\,kg^{-1}$	Latent heat
p_g	Pa	Pressure in the bubble
p_g	Pa	Pressure in the liquid far from the bubble
R	m	Bubble radius
S	m	Outer radius of the cellular bubble
T	K	Temperature
U	$m\,s^{-1}$	Translational velocity of the bubble
γ	–	Specific heat ratio
η_l	Pa s	Liquid viscosity
κ_{gl}	$m^2\,s^{-1}$	Diffusivity of the volatile in the liquid
κ_T	$m^2\,s^{-1}$	Thermal diffusivity in the bubble
κ_{Tl}	$m^2\,s^{-1}$	Thermal diffusivity in the liquid
$\bar{\mu}$	Pa	Effective stiffness of the magma chamber
μ_l	Pa	Shear elasticity of the liquid
ρ_l	$kg\,m^{-3}$	Liquid density
Σ	$m\,s^{-1}$	Surface tension
σ_∞	Pa	Ambient stress change given to the magma chamber
τ	s	Maxwell relaxation time
ω	$rad\,s^{-1}$	Angular frequency

42] and increase the attenuation and dispersion of the waves [15,30,44]. This effect is considered to be relevant to many spectral features of seismic waves and air-waves associated with volcanic activities [3,10,22,45].

Studies on bubble dynamics relevant to the volcanology are spread over many research fields and cannot be covered by a single paper. Good review papers and textbooks have already been published on bubble phenomena in sound fields [61,70,73,74] and on the nucleation and growth of bubbles in magma [62]. In this paper, we discuss several bubble dynamics phenomena selected from a particular point of view that the bubbly fluid works as an impulse generator. Here the term impulse means a pressure increase, either static or transient, which can generate any observable signal (e. g. earthquakes, ground deformations, airwaves, and an eruption itself). Especially, we focus on the processes that the impulse is excited by nonlinear coupling between the internal processes of a bubbly fluid and an external perturbation. The importance of these processes have recently become noticed as a possible triggering mechanism of eruptions, earthquakes, and inflation of a volcano [57,64]. Although it is generally considered that stress perturbation caused by preceding events is important, exact mechanisms to generate a pressure increase, which is required to trigger the subsequent events, are yet under discussion. In the first place, factors controlling single bubble dynamics are summarized as the elementary processes in the bubbly fluid. Then two distinct liquid-bubble systems are considered, both of which are included in a volcano. The one is a body of bubbly magma confined in an elastic chamber, where elasticity of the chamber, melt viscosity, and gas diffusion are important. The other is a hydrothermal system, where bubble oscillation, evaporation, and heat transfer are important.

Introduction

Elementary Processes in Single-Bubble Dynamics

Radial motion of a single bubble interacting with ambient pressure perturbation is the elementary process controlling behaviors of the liquid-bubble mixtures. Although it appears quite simple, it contains various mechanisms in plenty. The great variety of behaviors of a single bubble has attracted many scientists, among whom is Leonardo da Vinci [76]. Nowadays, knowledge of the single bubble dynamics is used and studies are continued in many academic and industrial areas such as mechanical engineering, chemical engineering, medical science, and earth science.

Factors which may control the radial motion of a bubble are the pressure difference inside and outside the bubble, inertia and stress associated with the deformation of the surrounding liquid, propagation of pressure waves, heat and mass transport, phase transition at the bubble wall, chemical reactions, relative translational motion between the bubble and the liquid, and so on. Because including all these mechanisms at the same time in order to calculate the behavior of a bubble is unrealistic, we need to make adequate simplification and assumptions. Each mechanism has its own characteristic time scale in which the effect is dominant (Table 2), and its own effect on the bubble dynamics. Knowing the individual time scales and features is important when we want to understand and simulate a certain phenomenon correctly and efficiently. A brief review of some representative mechanisms with linearized analyses are presented in Sect. "Elementary Processes in a Single–Bubble Dynamics" for this purpose. Based on the results, geophysical phenomena and proposed models are discussed in the latter sections.

Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Table 2
List of characteristic times

Notation	Equation	Definition
τ_c	(2)	Inertia-controlled bubble collapse
τ_g	(25)	Mass diffusion in the liquid around the bubble
τ_T	(24)	Thermal diffusion in the bubble
$\tau_v <$	(11)	Viscosity-controlled bubble expansion
ω_o	(9)	Natural frequency of a bubble

Bubbly Magma in an Elastic Rock as a Pressure Source

We consider a body of bubbly magma confined in an elastic rock. Pressure perturbations to the system are caused by a change of tectonic stress due to local earthquakes, surface unloading by dome collapse, passing seismic waves from a near or distant source, or depressurization of the chamber by degassing or magma leakage. Dynamic response of the system may be relevant to subsequent activities of the volcano as follows.

Nishimura [64] investigated pressure re-equilibration between the bubbles, the melt, and the surrounding elastic medium. It is assumed that the pressure of the system is suddenly decreased. After re-equilibration, the original magma pressure is partially or completely recovered or even exceeded, depending on the size of the bubbles, stiffness of the elastic container, and the confining pressure. His model is used to explain rapid pressurization of a magma chamber triggered by the lava-dome collapse at Soufriere Hills Volcano [97], and pressure recovery in

magma filling the chamber after explosive degassing to continue activities at Popocatepetl Volcano [12]. Shimomura et al. [87] extended the formulation of [64] to calculate the time profile of the pressure recovery after sudden decompression. They showed that the time scale of the pressure recovery is strongly controlled by the system parameters, which include stiffness of the elastic container, bubble number density, diffusivity of the volatile in the melt, ambient pressure, and properties of the melt. Chouet et al. [13] also calculated the time profile assuming the system parameters for Popocatepetl Volcano, and compared the results with a particular source time function of a very-long-period seismic signal. Furthermore, in the same year, Lensky et al. [53] independently developed a mathematically equivalent model considering magma with CO_2 bubbles in mantle rock. They interpreted the results as a possible pressurization mechanism to initiate dikes in mantle which allow the fast transport of magma. There are in fact quite a few documented cases in which eruptions were triggered by local tectonic earthquakes (e. g. [47,68]) and wave propagation from a distant earthquake (e. g. [6,54]). Recently, Manga and Brodsky [57] have given a comprehensive review on the phenomena and possible mechanisms. Brodsky et al. [6] investigated the possibility that a strain wave from a distant earthquake can increase the pressure in bubbly magma by rectified diffusion, which is the mechanism by which volatiles are pumped into a bubble by cyclic expansion and contraction. However, it has turned out that the mechanism by itself can cause a negligibly small pressure increase [6,29]. Several other mechanisms for long-range triggering have been proposed, which include pressure increase from rising bubbles [55], sub-critical crack growth [7], and fracture unclogging [8].

Acoustic Bubbles in Hydrothermal Systems

A hydrothermal system is another major source of pressure increase, long-period volcano seismic events [46], and triggered seismicity [57,89]. Behaviors of a single bubble and liquid-bubble mixtures in a hydrothermal system are quite different from those in a magmatic system, mainly because of the water viscosity which is less than that of magma by several orders of magnitude. We introduce several phenomena which are particular to the hydrothermal systems in Sect. "Acoustic Bubbles in Hydrothermal Systems".

Geyser activity is well known for intermittent activity of hot-water effusion. The effusion process looks quite similar to volcanic eruptions, and some geysers are characterized by regular intervals of time and duration, which are also recognized in particular types of eruptions and seismic activities. Consequently, the geysers have been widely studied using seismological and geophysical techniques, as well as field observations, not only for clarifying the mechanism of the geysers but also for understanding the volcanic activities (e. g., [39,40,43,65]). Kedar et al. [39,40] conducted a unique experiment at Old Faithful Geyser, Yellowstone. They measured pressure within the geyser's water column simultaneously with seismic measurements on the surface. The data demonstrated that the tremor observed at Old Faithful results from impulsive events in the geyser. The impulsive events were modeled by a collapse of a spherical bubble by cooling that occurred when the water column reached a critical temperature. Their data are reviewed in Sect. "Acoustic Bubbles in Hydrothermal Systems" in relation to other studies on the dynamics of gas and vapor bubbles.

Elementary Processes in a Single-Bubble Dynamics

Equation of Motion for the Bubble Radius

Motion of a bubble is in fact a fluid dynamical problem for the liquid surrounding the bubble. The simplest model describing the behavior of a bubble is based on three assumptions:

(1) The bubble is spherical,
(2) The liquid is incompressible, and
(3) The motion is radial.

Using the basic equations of fluid mechanics, which are the continuity equation and the momentum equation, and the force balance at the bubble surface, the first equation of motion for the bubble radius was obtained by Rayleigh [80]:

$$\rho_l \left(R\ddot{R} + \frac{3}{2}\dot{R}^2 \right) = p(R) - p_l \,, \tag{1}$$

where R is the bubble radius, ρ_l is the liquid density, $p(R)$ is the pressure in the liquid at the bubble surface, and p_l is the pressure in the liquid at a large distance from the bubble. Using Eq. (1), Rayleigh [80] solved the problem of the collapse of an empty cavity in a large body of liquid at a constant p_l and showed the characteristic collapse time is

$$\tau_c = R_o \sqrt{\rho_l / p_l} \,. \tag{2}$$

The time τ_c is called the Rayleigh collapse time and is one of the most important time scales in the bubble dynamics.

Plesset [69] extended Eq. (1) including the effects of surface tension and time-dependent pressure field, and

Proitsky [71] included the effect of viscosity. The generalized Rayleigh equation for bubble dynamics is known as the Rayleigh–Plesset equation. liquid viscosity and surface tension. The generalized Rayleigh equation for bubble dynamics is called the Rayleigh–Plesset equation [70]:

$$\rho_l\left(R\ddot{R}+\frac{3}{2}\dot{R}^2\right)=p_g-p_l-4\eta_l\frac{\dot{R}}{R}-\frac{2\Sigma}{R}\,, \tag{3}$$

where η_l is the liquid viscosity, and Σ is the surface tension. Equation (3) is valid for a Newtonian liquid under conditions of negligible mass exchange at the bubble surface. A further generalized equation to which these two restrictions do not apply is [72]:

$$\rho_l\left(R\dot{u}_l+\frac{3}{2}u_l^2\right)-J\left[2u_l+J\left(\frac{1}{\rho_g}-\frac{1}{\rho_l}\right)\right]=p_g-p_l+\int_R^\infty\frac{3\tau_{rr}}{r}\mathrm{d}r-\frac{2\Sigma}{R}\,, \tag{4}$$

where u_l is the radial liquid velocity at the bubble surface, $J=\rho_l(u_i-\dot{R})$ is the outgoing mass flux through the bubble wall, ρ_g is the density of the gas in the bubble, and τ_{rr} is the normal radial stress. When the interfacial mass flux J vanishes, $u_l=\dot{R}$ as in the left-hand side of the original Eq. (3).

While the above equations consider a single bubble in an infinite melt, magmatic systems often contain bubbles with some finite spacing. Cellular models of packing which include a finite volume of melt in interaction with each bubble have been employed for closely spaced bubbles [79]. When the elementary cell is represented by a sphere with an outer radius of S, the equation corresponding to (3) is [76]:

$$\rho_l\left[R\ddot{R}\left(1-\frac{R}{S}\right)+\frac{3}{2}\dot{R}^2\left(1-\frac{4}{3}\frac{R}{S}+\frac{1}{3}\frac{R^4}{S^4}\right)\right]=p_g-p_l-4\eta_l\frac{\dot{R}}{R}\left(1-\frac{R^3}{S^3}\right)-\frac{2\Sigma}{R}\,. \tag{5}$$

Equation (5) agrees with Eq. (3) for $S\to\infty$.

When there is no transport of heat or mass between the liquid and the bubble, the pressure in the bubble is determined by the instantaneous bubble radius alone. Using the ideal gas approximation, we have

$$p_gR^{3\gamma}=p_{go}R_o^{3\gamma}\,, \tag{6}$$

where γ is the specific heat ratio, and the subscript o indicates the equilibrium value of the variable. Substituting Eq. (6) into Eq. (3) for p_g and linearizing the equation, we obtain a damped oscillator equation:

$$\ddot{X}+2b_\nu\dot{X}+\omega_o^2X=-\frac{p_l'}{\rho_lR_o^2}\,, \tag{7}$$

$$b_\nu=\frac{2\eta_l}{\rho_lR_o^2}\,, \tag{8}$$

$$\omega_o=\frac{1}{R_o}\sqrt{\frac{3\gamma p_{go}-2\Sigma/R_o}{\rho_l}}\,, \tag{9}$$

where X and p_l' are defined by $R=R_o(1+X)$ and $p_l=p_{go}-2\Sigma/R_o+p_l'$, respectively. Equation (7) is useful to see the characteristic behaviors of a bubble and their time scales. The resonant frequency of the bubble is ω_o (rad s^{-1}). When the second term in the left-hand side of Eq. (7) dominates the first one in the time scale of the resonant oscillation, namely when $\omega_o<b_\nu$, the resonant oscillation is damped. In the case of a gas bubble with a radius of 10^{-3} m in magma ($\rho_l=2500\,\mathrm{kg\,m^{-3}}$) at 10 MPa ($p_{go}-2\Sigma/R_o=10^7$ Pa), the frequency ($\omega_o/(2\pi)$) is about 20 kHz. The oscillation is damped when the viscosity is larger than 160 Pa s. This viscosity is relatively small for magma. According to these estimations, we see that the free oscillation of a bubble in magma is possible in the limited cases that the viscosity is small and the bubble is large. We also see that the bubble oscillation is easily excited in water which has a viscosity about 10^{-3} Pa s.

Liquid Rheology

The shear rheology of the liquid surrounding the bubble is one of the controlling factors for the bubble dynamics. According to experimental results, magma has viscoelastic nature, which is the most simply represented by a linear Maxwell model [99]. Then the normal radial stress τ_{rr} in Eq. (4) is related to the corresponding strain rate $\dot{e}_{rr}$ by

$$\tau_{rr}=\mu_l\int_0^t\exp\left(-\frac{t-t'}{\tau}\right)\dot{e}_{rr}\mathrm{d}t'\,, \tag{10}$$

where μ_l is the shear elasticity and τ is the relaxation time. In the limit of $t\ll\tau$, the Maxwell relation (10) is reduced to a linear elastic stress-strain relation as $\tau_{rr}=\mu_le_{rr}$. While in the limit of $t\gg\tau$, it is reduced to a Newtonian viscous relation, that is a linear stress-strain rate relation as $\tau_{rr}=\mu_l\tau\dot{e}_{rr}$, where $\mu_l\tau$ corresponds to the Newtonian viscosity η_l.

Fogler and Goddard [21] first used the viscoelastic relation (10) in the generalized Rayleigh–Plesset Eq. (4) without mass flux, and demonstrated that the influence of the viscoelastic effects on the radial motion of a bubble is characterized by a dimensionless parameter called the Deborah number $De=\tau/\tau_c$, which compares the relaxation time τ and Rayleigh collapse time τ_c defined in Eq. (2): the influence is large when $\tau\gg\tau_c$. Extending the formulation by Fogler and Goddard [21] to a cellular bubble, Ichihara [28] investigated its characteristic behaviors

in magmatic conditions. It is shown that the elastic oscillation of a bubble, which occurs in the case of $\tau \gg \tau_c$, is in a frequency of order of MHz and with very small displacement of the bubble wall because of the large shear modulus of the magma.

Change of the bubble radius in magma is mainly controlled by the viscosity except in magma with very small viscosity in which the bubble oscillation is possible. Therefore, in most of the studies for bubble growth in magma, effects of liquid inertia and viscoelasticity are not considered, and Eq. (3) or (5) for a Newtonian fluid is used, neglecting the left-hand side terms representing the inertia [2,62,79,88]. Barclay et al. [2] analytically solved the problem of the viscosity-controlled bubble expansion for instantaneous decompression, and showed the characteristic expansion time is

$$\tau_v = \frac{4\eta_l}{3p_l}, \tag{11}$$

where p_l is the pressure in the liquid. The time τ_v is one of the most important time scales of the bubble dynamics in magma [2,30], while the Rayleigh collapse time τ_c, which is controlled by the inertia, is important in low-viscosity fluids including hydrothermal systems.

Definition of the viscous expansion time corresponding to Eq. (11) is different depending on which problems and literature are being referenced. The time scale of the entire expansion of a bubble for instantaneous decompression is represented by Eq. (11) using the reduced pressure for p_l [2]. Volumetric oscillation of a bubble in an acoustic field is prevented by the viscous resistance if the period is shorter than τ_v, in this case with the initial static pressure for p_l [30]. When the bubble expansion is driven by a constant gas pressure, which occurs at the initial stage of diffusion-drive gas expansion when the gas is efficiently supplied from the liquid, the bubble grows approximately as $R \sim R_o \exp[t\Delta p/(4\eta_l)]$ [62,93], where Δp is the pressure difference. In this case, $\tau_v = 4\eta_l/\Delta p$. The last case is discussed again later in the section of mass transport.

Liquid Compressibility

The effect of liquid compressibility on radial motion of a bubble was first considered in connection with underwater explosions [14,41]. Liquid compressibility allows energy transport as a pressure wave so that it causes radiation damping. Noting that the effect is considerable in the case of a violent oscillation or collapse of a bubble, several mathematical approaches were proposed to include the effect in the equation of bubble radius. According to mathematical and numerical studies by Prosperetti and Lezzi [78], which compared the proposed equations, the following Keller's equation [41] is widely accepted as the most adequate form.

$$\rho_l\left[\left(1-\frac{\dot{R}}{c_l}\right)R\ddot{R}+\frac{3}{2}\left(1-\frac{\dot{R}}{3c_l}\right)\dot{R}^2\right] = \left(1+\frac{\dot{R}}{c_l}+\frac{R}{c_l}\frac{\mathrm{d}}{\mathrm{d}t}\right)\left(p_g-p_l-4\eta_l\frac{\dot{R}}{R}-\frac{2\Sigma}{R}\right), \tag{12}$$

where c_l is the sound speed in the liquid. Although Prosperetti and Lezzi [78] further proposed to use the liquid enthalpy at the bubble wall instead of the pressure for the best accuracy, Eq. (12) is generally used in the literature.

Comparing Eqs. (12) and (3), we can see that the correction terms due to the liquid compressibility have the order of $\dot{R}/c_l$. It means that the correction is considerable only when the bubble wall velocity becomes as large as the sound speed of the liquid. By applying $[1+(\dot{R}/c_l)+(R/c_l)d/\mathrm{d}t]^{-1}$ to both sides of Eq. (12) and linearizing the equation, Prosperetti [75] derived the acoustic damping coefficient, which corresponds to b_v in Eq. (8) for the viscous damping, as b_{ac}:

$$b_{ac} = \frac{\omega^2 R_o}{2c_l}, \tag{13}$$

where ω is the angular frequency of the oscillation. The acoustic damping is more significant when the bubble is larger and the oscillation frequency is higher.

Effects of liquid compressibility on bubble dynamics in a highly viscous liquid are not understood comprehensively. Derivation of Eq. (12) and related studies were performed thinking of liquids with ordinary viscosities like water. Therefore, the Reynolds number $Re = \rho_l R_o c_l/\eta_l$ was presupposed to be large. On the other hand, the viscosity of magma can be large enough to make Re very small. In this case, the same mathematical approximation is not necessarily applicable. Yamada et al. [100] pointed out this problem and solved the equations including the viscous force associated with the volumetric strain rate. Although the equation of motion for the bubble radius appears not to be affected by the compressibility when the system is initially hydrostatic, the velocity field around the bubble is different from the incompressible solution, even if the wall velocity is much smaller than the acoustic velocity.

There is argument whether the equation of radial motion of a bubble surrounded by a finite volume of liquid needs correction terms for the compressibility. However, it seems to be negligible in magma, which is evaluated as follows [28]. In the case that the bubble is surrounded by an

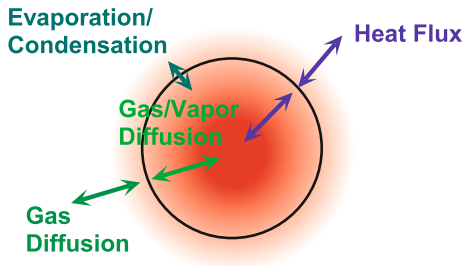

Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 1

The internal heat and mass transport processes which control pressure change and energy loss associated with the bubble oscillation

elastic shell, contribution of the compressibility to the bubble expansion is $\delta_c R = R_o(p_g - p_l)(1 - R_o^3/S_o^3)^{-1}(3K_l)^{-1}$, where K_l is the bulk modulus of the liquid, which is the reciprocal of the compressibility [48]. We can evaluate $\delta_c R/R_o < 10^{-3}$, because $K_l \sim 10^{10} - 10^{11}$ Pa for magma [99], the realistic pressure difference, $p_g - p_l$ is not much larger than 10^7 Pa, and the volume fraction of the bubbles, R_o^3/S_o^3, is reasonably assumed to be smaller than the close-packing limit (~ 0.74). The change of p_g due to $\delta_c R$ is $\delta_c p_g/p_{go} \sim -3\delta_c R/R_o$, which is also in the same order. When the shell deforms viscously, displacement due to non-volumetric deformation grows, while that from the volumetric deformation remains in the same order.

Heat and Mass Transport

Each equation of motion for the bubble radius includes p_g, which is the pressure of the gas in the bubble, as we see in Eqs. (3), (5), and (12). Equation (6) is available to calculate p_g only for an adiabatic process. In fact, when a bubble expands, the pressure and temperature in the bubble decreases. Then heat and volatile components flow into the bubble from the surrounding liquid. The opposite occurs when a bubble shrinks. The internal processes which control p_g are schematically shown in Fig. 1. These transport effects are essential in most of actual systems including magmatic and hydrothermal systems.

Growth of a bubble by the mass diffusion in an oversaturated liquid is one of the fundamental problems. Based on purely dimensional considerations, an approximate growth law is given by

$$R\dot{R} = \frac{\kappa_{gl}\rho_l(C_o - C_{eq})}{\rho_g}\,, \tag{14}$$

where κ_{gl} is the diffusivity of the volatile in the liquid, C_o is the dissolved volatile concentration at a large distance from the bubble, and C_{eq} is the equilibrium concentration at the given pressure [62,70]. From Eq. (14) we find that, asymptotically, $R \sim \sqrt{2\kappa_{gl}\rho_l(C_o - C_{eq})t/\rho_g}$. This expression is not valid at $t \to 0$ making $\dot{R} \to \infty$. In the initial stage, the diffusion is very efficient and the bubble growth is controlled by viscous resistance [62,93]. In this stage, $R \sim R_o \exp[t\Delta p/(4\eta_l)]$ as discussed in the section of liquid rheology. The approximate time of the transition from the viscosity-controlled exponential solution to the diffusion-controlled square-root solution is found by Navon and Lyakhovsky [62] to be

$$\tau_{vd} \sim [-15 - 10\log(Pe)]\eta_l/\Delta p\,, \tag{15}$$

where $Pe = \Delta p R_o^2 \eta_l^{-1}\kappa_{gl}^{-1}$ is the Pecret number that compares the time scales of viscous expansion and diffusion. It is noted that Eq. (15) is validated for $Pe < 10^{-2}$, that is for relatively large viscosity and small bubble radius [62]. If we consider $\Delta p \sim 10^6$ Pa and $\kappa_{gl} \sim 10^{-11}\,\mathrm{m^2\,s^{-1}}$, this condition is satisfied when $R_o^2\eta_l^{-1} < 10^{-19}$, that is $\eta_l > 10^7$ Pa s when $R_o \sim 10^{-6}$ m, and $\eta_l > 10^9$ Pa s when $R_o \sim 10^{-5}$ m. Then the corresponding times are $\tau_{vd} > 50$ s and $\tau_{vd} > 5000$ s, respectively. Lensky et al. [51] have suggested that the change of the characteristic behavior of the bubble expansion over the time scale τ_{vd} generates a non-linear response of the liquid-bubble mixture to the pressure perturbation, which may cause amplification of a pressure wave. Coupling of effects of viscosity and diffusion on the bubble expansion also occurs through the material properties, because magma viscosity and water diffusivity are strongly influenced by the amount of dissolved water, which is the major volatile component in magma [4,50].

Matsumoto and Takemura [59] numerically solved a complete set of equations for the radial dynamics of a bubble including the conservation equation for mass, momentum, and energy in the bubble, heat and mass diffusion in the liquid, and heat and mass exchange between the gas and the liquid by diffusion and evaporation/condensation. Except in extremely rapid phenomena as the cases they treated, approximation of a spatially uniform pressure in the bubble is adequate [75]. With this approximation, the computational load is considerably reduced [9,38,63].

It is necessary to consider non-uniform temperature distribution and compositions, in order to quantify the amounts of energy exchange between the bubble and liquid and energy loss associated with the non-equilibrium process. Time scales required to recover uniform temperature and composition in the bubble are controlled by diffusion processes and are much longer than that to attain uniform pressure, which is controlled by the pressure

wave propagation in the bubble. Assuming representative values of the thermal diffusivity, $\kappa_T \sim 10^{-5}$ ($m^2\,s^{-1}$), and the inter-diffusivity of the components in the gas phase, $\kappa_{gi} \sim 10^{-7}$ ($m^2\,s^{-1}$) [38], development of thermal and material diffusion layers all over the bubble with radius of 10^{-3} m takes ~ 0.1 s and ~ 10 s, respectively. The time range of 0.1–10 s is exactly what studies on seismoacoustic phenomena in volcanology are mainly concerned with. It takes an even longer amount of time to recover uniform concentration of volatile components in the liquid around the bubble. Therefore approximation of uniform temperature and compositions are not always adequate. Again we introduce results from the linearized theory for a periodic acoustic field. The bulk modulus of a bubble, K_g, is defined as:

$$K_g = -\frac{R}{3}\frac{\partial p_g}{\partial R}\,, \tag{16}$$

which is generally a complex number. The elasticity and the energy loss associated with volumetric change of a bubble are represented by the real and imaginary parts of K_g, respectively. Then the damping factor and the resonant frequency for the bubble oscillation, which correspond to Eqs. (8) and (9), respectively, are [75]:

$$b_t = \frac{3\mathrm{Im}(K_g)}{2\rho_l \omega R_o^2} \tag{17}$$

$$\omega_o = \frac{1}{R_o}\sqrt{\frac{3\mathrm{Re}(K_g) - 2\Sigma/R_o}{\rho_l}}\,. \tag{18}$$

In the case of an adiabatic process for an ideal gas, where Eq. (6) holds, $K_g = \gamma p_{go}$ and Eq. (18) agree with Eq. (9). While in the case of an isothermal process, $K_g = p_{go}$. In these two extreme conditions, $\mathrm{Im}(K_g) = 0$ and there is no thermal damping.

In order to include the effect of non-uniform temperature distribution in the bubble, we have to solve the energy equation with the continuity of temperature at the bubble surface. Assuming that the pressure in the bubble is uniform, the temperature at the bubble wall is constant, which is supported by the large heat capacity of the liquid compared with that of the gas, and the pressure perturbation is periodic ($\propto e^{i\omega t}$), the effective bulk modulus, K_g, is represented as [75]:

$$\frac{p_{go}}{K_g} = \frac{1}{\gamma} - \frac{3(\gamma-1)}{\gamma} i\chi\left[\sqrt{\frac{i}{\chi}}\coth\left(\sqrt{\frac{i}{\chi}}\right) - 1\right], \tag{19}$$

$$\chi = \frac{\kappa_T}{\omega R^2}\,. \tag{20}$$

When the mass transfer is controlled by the diffusion of the volatile component in the liquid phase, the diffusion equation in the liquid and the equilibrium condition at the bubble surface are added. Then the effective bulk modulus which includes both the heat and mass transport is

$$\frac{p_{go}}{K_g} = \frac{1}{\gamma} - \frac{3(\gamma-1)}{\gamma} i\chi\left[\sqrt{\frac{i}{\chi}}\coth\left(\sqrt{\frac{i}{\chi}}\right) - 1\right] - 3A_g\sqrt{\alpha_g}\, i\chi\left(\sqrt{\frac{i}{\chi}} + \sqrt{\alpha_g}\right), \tag{21}$$

$$A_g = \frac{\rho_l p_{go}}{\rho_{go}}\frac{\partial C_{eq}}{\partial p}\,, \tag{22}$$

$$\alpha_g = \frac{\kappa_{gl}}{\kappa_T}\,, \tag{23}$$

where C_{eq} is the saturation concentration at p_{go} [30]. Equation (21) has the last term in addition to Eq. (19), which represents the effect of the mass transfer. The dimensionless parameter, A_g, represents the ratio of the volatile mass going into the gas phase from a unit volume of the liquid phase by decompression to the mass change in a unit volume of the gas phase due to expansion.

Figure 2 shows the relevant range of the dimensionless parameters, A_g and α_g, for an H_2O bubble in magma [30]. As temperature decreases or pressure increases, A_g decreases (Fig. 2) because of the following two reasons. With decreasing temperature, ρ_{go}^{-1} decreases. In the case of magma, $C_{eq}(p)$ is approximately proportional to $\sqrt{p}$ [26] so that $\partial C_{eq}/\partial p$ decreases with increasing pressure.

The effective bulk modulus of a bubble for some selected values of the parameters in the range is presented in Fig. 3 [30]. The thick broken lines in the figure are obtained by Eq. (19), which includes only the heat transport. In this case, the real part approaches the isothermal bulk modulus and the adiabatic one in the low and high frequencies, respectively. The mass transport makes the bubble stiffness ($\mathrm{Re}(K_g)$) smaller, which is the more significant in the lower frequency regime. It is because the bubble has more time to take in and out the volatile from the liquid in a cycle of the pressure perturbation. The imaginary part for each parameter set has a local peak around

$$\omega \sim \tau_T^{-1} = 15\kappa_T R_o^{-2}\,, \tag{24}$$

which is the characteristic frequency of the energy loss due to heat transfer. In the case of $R_o = 10^{-3}$ (m) and $\kappa_T = 4\times10^{-6}$ ($m^2\,s^{-1}$), which is the value for H_2O at 10 MPa and 1273 K [5], the corresponding frequency ($\omega/(2\pi)$) is 9.5 Hz. The imaginary part of K_g including the diffusion effect has another peak at the characteristic frequency of the energy loss due to the mass transport. The frequency is approximately represented by

$$\omega \sim \tau_g^{-1} = 9\alpha_g\kappa_T A_g^2 R_o^{-2}\,, \tag{25}$$

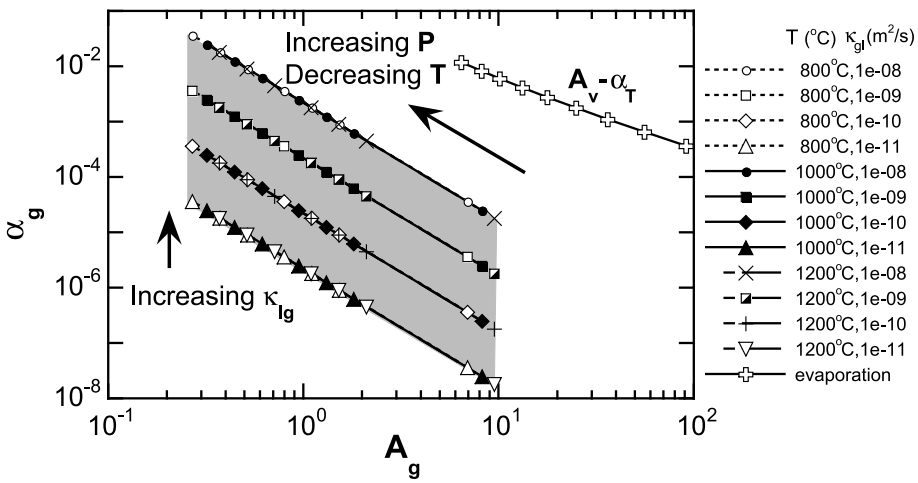

Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 2

The relevant range of the dimensionless parameters representing the effect of the volatile transfer in the magmatic system is shown as the *gray area*. The parameter A_g and α_g are defined in Eqs. (22) and (23), respectively. The temperature (T) and the volatile diffusivity (κ_{gl}) are assumed as shown in the legend, and the pressure is varied from 0.1 to 100 MPa. The *open crosses* are the corresponding parameters for a vapor-bubble system, A_v and α_T given in Eqs. (27) and (28), respectively. The temperature is varied from 380 K to 500 K and the pressure is the saturation pressure at each temperature. (Modified from Fig. 1 in [30])

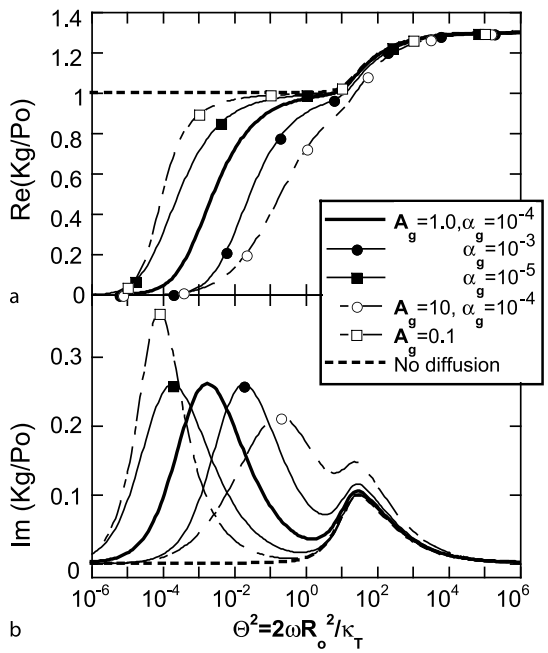

Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 3

The effective bulk modulus of a gas bubble with heat and mass diffusion (calculated by Eq. (21)) as a function of the dimensionless frequency. The real and the imaginary parts are presented in a and b, respectively. The *thick broken lines* indicate no diffusion and include only thermal effects (which is calculated by Eq. (19)). (Fig. 2 in [30])

which is usually smaller than τ_T^{-1} [30]. It is noted that the above model assumes a single bubble in an infinite liquid. When the oscillation period is very long, interaction of the diffusion layers of the adjacent bubbles has to be considered [15].

When the mechanism of the mass exchange between the liquid and the bubble is the evaporation/condensation at the bubble wall, the latent heat plays an important role. Then the thermal diffusion equation in the liquid and the balance between the heat flux through the bubble surface and generation of the latent heat should be added [19,24]. The corresponding bulk modulus of the bubble is represented as [24]:

$$\frac{p_{go}}{K_g} = \frac{1}{\gamma} - \frac{3(\gamma-1)}{\gamma}\left(1-\frac{c_{pg}T_o}{L}\right)^2 i\chi\left[\sqrt{\frac{i}{\chi}}\coth\left(\sqrt{\frac{i}{\chi}}\right)-1\right] - 3A_v\sqrt{\alpha_T}\,i\chi\left(\sqrt{\frac{i}{\chi}}+\sqrt{\alpha_T}\right), \quad (26)$$

$$A_v = \frac{\rho_l c_{pl} T_o p_{go}}{(\rho_{go} L)^2}, \quad (27)$$

$$\alpha_T = \frac{\kappa_{Tl}}{\kappa_T}, \quad (28)$$

where c_{pg} and c_{pl} are the heat capacity at constant pressure in the gas and the liquid phases, respectively, L is the latent heat, and κ_{Tl} is the thermal diffusivity in the liquid. In obtaining Eq. (26), the Clausius-Clapeyron re-

lation: $(\mathrm{d}p/\mathrm{d}T)_{\mathrm{sat}} = L\rho_{\mathrm{g}}/T$, and thermodynamic relations for an ideal gas are used.

Equation (26) has the same form as Eq. (21) except $c_{\mathrm{pg}}T_o/L$. This difference is due to the temperature change at the bubble wall. The dimensionless parameter, A_v, corresponds to A_{g} and has a similar physical meaning, which represents the ratio of the mass going through a phase change in a unit volume of the liquid phase to the mass change in a unit volume of the gas phase due to expansion. Although the equation is similar, the possible range of the parameter is different (Fig. 2). As a result, the frequency dependence of K_{g} is also different as is shown in Fig. 4. Comparing the figure with Fig. 3, we can see that the energy loss of the vapor bubble due to the phase change is significant in a frequency range higher than that due to diffusion. The frequency range is comparable to that of the heat transfer, but the amount of energy loss is much larger. The vapor bubble loses its elasticity, which is represented by $\mathrm{Re}(K_{\mathrm{g}})$, in the lower frequency as well.

Under the action of the sound field, there is a net transport of heat into the bubble by a non-linear process called the rectified heat transfer [98]. In the evaporation/condensation system, the order of the non-linear effect is so large that it affects the amplitude and damping of the oscillation in the linear regime [24]. Equation (26) does not include the effect. Some works investigating the effect of rectified diffusion process in triggering an eruption or an earthquake are introduced in the later sections.

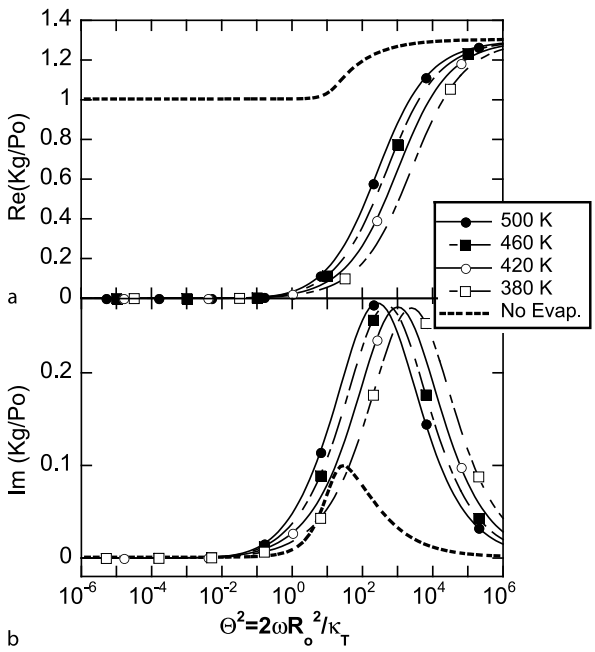

Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 4

The effective bulk modulus of a vapor bubble with thermal and evaporation effects (calculated by Eq. (26)) as a function of the dimensionless frequency. The real and the imaginary parts are presented in a and b, respectively. The *thick broken lines* indicate no evaporation and include only thermal effects (which is calculated by Eq. (19))

Translational Motion

So far we have neglected the translational motion of a bubble relative to the liquid. The translational motion is considered to be negligible when the translational displacement is smaller than the diffusion layer in the liquid surrounding the bubble. In an acoustic field with frequency ω, this condition is represented by $U/\omega < \sqrt{\kappa/\omega}$, where U is the translational speed and κ is the relevant diffusivity. The translational velocity of a spherical bubble driven by buoyancy is estimated by $U = k\rho_{\mathrm{l}}R_o^2 g\eta_{\mathrm{l}}^{-1}$, where g is the gravitational acceleration. Although $k = 1/3$ for a pure liquid, $k = 2/9$ is used for most of actual liquid, which is not perfectly pure, because the pro-surface components concentrate on the bubble surface to make the surface less mobile [49]. These approximations hold for relatively slow velocity, which satisfies $Re_{\mathrm{t}} = 2\rho_{\mathrm{l}}R_o U\eta_{\mathrm{l}}^{-1} \leq 1$. Then the condition in which the translational motion has a minor effect on the heat and mass transfer is

$$\omega > \frac{R^4}{\kappa}\left(\frac{k\rho_{\mathrm{l}}g}{\eta_{\mathrm{l}}}\right)^2 . \tag{29}$$

Assuming $k = 2/9$, $R = 10^{-3}$ (m), $\kappa = \kappa_{\mathrm{gl}} = 10^{-11}$ ($\mathrm{m^2\,s^{-1}}$), $\rho_{\mathrm{l}} = 2500(\mathrm{kg\,m^{-3}})$, $\eta_{\mathrm{l}} = 10^5$ (Pa s) for a magma-H_2O system, $\omega > 3\times10^{-4}$ ($\mathrm{rad\,s^{-1}}$) and $U = 6\times10^{-6}(\mathrm{m\,s^{-1}})$. For a water-vapor system, on the other hand, we assume $k = 1/3$, $\kappa = \kappa_{\mathrm{Tl}} = 10^{-7}$, $\rho_{\mathrm{l}} = 1000$, $\eta_{\mathrm{l}} = 10^{-3}$. Then, if $R = 10^{-5}$, $\omega > 1$ and $U = 3\times10^{-4}$, and if $R = 10^{-4}$, $\omega > 10^4$ and $U = 3\times10^{-2}$. According to these estimations, we can see that the translational motion is negligible for most cases with magma except basalt, which has relatively small viscosity ($\eta_{\mathrm{l}} < 10^2$) and large diffusivity ($\kappa_{\mathrm{gl}} \sim 10^{-9}$), while it is considerable in hydrothermal systems, except for very small bubbles and the time scale is very short. As an example, the effect on the thermal collapse of a vapor bubble is introduced later.

Another effect of the translational motion is its mechanical coupling with the radial motion. Because a bubble has to move the surrounding liquid in order to make itself move, it is subject to the inertial force of the liquid, which depends on its volume [49]. Therefore when the bubble volume changes, the force also changes. By this consideration, an equation of the translational motion of the bubble

is approximately represented as [101]:

$$\dot{U} = -\frac{3}{R}\dot{R}U + 2g - \frac{3}{4}\frac{C_D}{R}|U|U\,, \tag{30}$$

where C_D is the non-dimensional drag coefficient, which is given as a function of Re_t. From the first term in the right-hand side, we see that the translational motion is decelerated by the bubble expansion. Although it had been theoretically recognized for long time, it was quantitatively justified by experiments recently [66]. On the other hand, the effect of the translational motion on the radial motion is represented by the term, $\rho U^2/2$. This term can usually be neglected [37,66] except in cases with very strong oscillation [17].

Bubbly Magma in an Elastic Rock as a Pressure Source

Model Overview

Here we consider a magma chamber filled with compressible viscous melt and numerous tiny H_2O gas bubbles (Fig. 5) [79,87]. The magma chamber is confined in an elastic rock. When perturbation is given to the system, pressure may increase by interaction of the elastic deformation of the chamber, expansion of the bubbles, and gas diffusion from the melt to the bubble. Recently, the process has been discussed in the literature in relation to the observed volcanic phenomena [6,13,29,53,64,87,97], which are presented in the introduction.

The melt and bubbles are expressed by the cell model [79], in which multiple spherical bubbles of a constant radius are uniformly packed. Each bubble is surrounded by a finite volume of the melt, represented by an elementary cell. The elementary cell is spherical, in which a single gas bubble is located at the center. It is assumed there is no interaction between neighboring elementary cells such that all gas bubbles grow in the same manner. This simplification enables us to examine bubble growth processes in the entire chamber by studying the growth of just a single bubble, which is represented by Eq. (5).

The main mechanism for increasing the pressure is diffusion of the volatile. It is the slowest process of the bubble dynamics as is described in the previous section. It is certainly longer than the period of resonant oscillation of the individual bubbles so that the inertia terms in Eq. (5) are neglected. It is also longer than the time scale of the heat transport within the bubble as is shown in Fig. 3 so that we may assume uniform and constant temperature within the bubble. Then the mathematical model for the elementary cell consists of three equations, which represent the radial motion of the bubble, volatile diffusion in the melt, and ideal gas approximation, respectively, and three boundary conditions, which are phase equilibrium and mass flux at the bubble surface and no mass flux at the external boundary of the cell element.

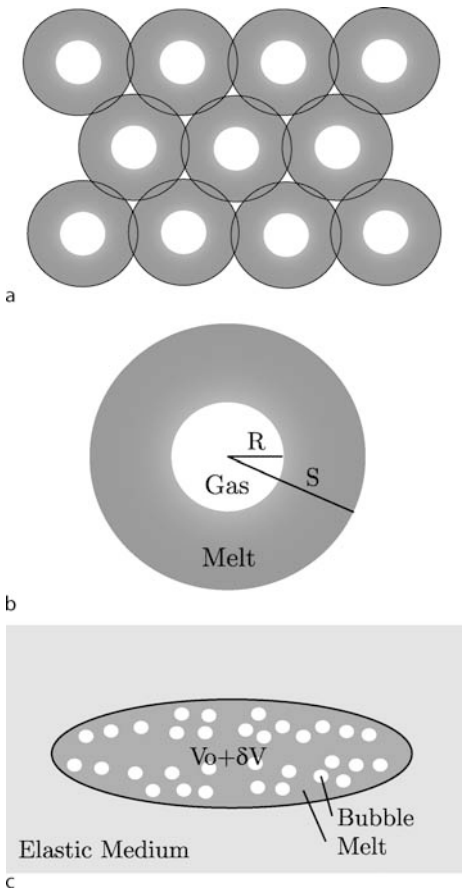

Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 5
Schematic illustrations of a cell model [79], b an elementary cell, and c a chamber surrounded by an elastic medium. The chamber is filled with compressible viscous melt and numerous tiny spherical gas bubbles. Magma is represented by a combination of many elementary cells. *R* and *S* is the radius of the elementary cell and gas bubble, respectively, and $V_o + \delta V$ is the volume of the chamber. (Modified from Fig. 1 in [87])

Interaction Between Melt and Elastic Medium

The volumetric change of the bubbles and the melt due to the pressure change is compensated by the elastic deformation of the chamber. Here we consider the initial pressure and stress conditions in relation to the physical process which brings about the condition, since the relations have not always been mentioned clearly in previous litera-

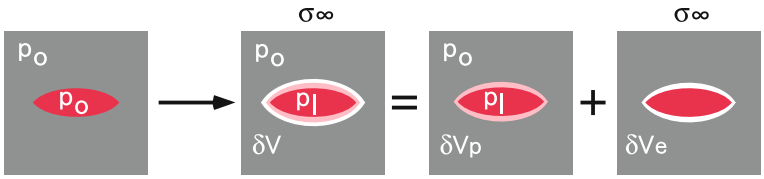

Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 6
Mathematical representation for the interaction of melt pressure p_l, stress in the ambient rock σ_∞, and volume change of the chamber, δV. Volume change due to internal overpressure $p_l - p_o$ and that due to external stress is considered separately

ture. We assume quasi-static deformation of the chamber, where the pressure of the melt is balanced by the elastic stress applied by the wall of the chamber. The volumetric change can be caused by (a) pressure change within the chamber and (b) stress change in the surrounding rock (Fig. 6). Each process is individually represented by

$$p_l - p_o = \bar{\mu}_p \delta V_p / V_o \,, \tag{31}$$

$$-\sigma_\infty = \bar{\mu}_e \delta V_e / V_o \,, \tag{32}$$

where V_o is the initial equilibrium volume of the chamber, σ_∞ is the ambient stress change, δV_p and δV_e are the volumetric change due to (a) and (b), respectively, and $\bar{\mu}_p$ and $\bar{\mu}_e$ are the corresponding effective stiffness of the wall. Each effective stiffness depends on the elasticity of the rock, the geometry of the chamber, and the applied stress field. For the simplicity, we approximate $\bar{\mu}_p = \bar{\mu}_e = \bar{\mu}$. Then the total volumetric change, δV, is given by

$$p_l - p_o - \sigma_\infty = \bar{\mu} \delta V / V_o \,. \tag{33}$$

The two perturbations which cause the volumetric change have not been clearly distinguished in previous literature. The mathematical treatment by [87] assumed that $p_l - p_o = -\Delta p$ is given at $t = 0$. They considered that this pressure drop is caused by a decrease of the ambient stress field by a certain amount, say $\sigma_\infty = -\Delta\sigma$. On the other hand, the assumption of [13] is that the pressures in all of the bubbles, the melt, and the rock are lower by Δp than the saturation pressure for the dissolved volatile concentration at $t = 0$. Strictly speaking, the consequent processes are different depending on what causes the pressure perturbation. If the pressure drop of Δp occurred first within the chamber, that is $p_l - p_o = -\Delta p$ and $\sigma_\infty = 0$, the chamber would initially shrink according to Eq. (33). If it is caused by an ambient stress drop first, that is $p_l - p_o = 0$ and $\sigma_\infty < 0$. Then the chamber would initially expand. In either case, the initial response is almost instantaneous, which is controlled by elasticity of the rock and compressibility of the melt. The major deformation occurs later and is controlled by volumetric change of the bubbles.

Response to Sudden Decompression and Characteristic Time for Pressure Recovery

There are three important parameters to characterize the response of the system to the pressure drop which are useful for comparing the model and the field observations. The first one is the re-equilibrated pressure p_f. The second is the final bubble radius, R_f. The third is the characteristic time of the recovery process, T_{growth}.

The first and the second are calculated by consideration of the equilibrium condition alone and can be calculated semi-analytically [64]. On the other hand, T_{growth} is determined by numerical calculation of the set of equations described above. Shimomura et al. [87] investigated the recovery processes and presented that T_{growth} depends on the stiffness of the chamber ($\bar{\mu}$), initial bubble radius (R_o), number density of the bubbles (N), volatile diffusivity in the melt (κ_{gl}), initial pressure (p_o), the pressure drop (Δp), and the melt properties in a complicated manner.

The corresponding study for the bubble growth in an open space, where the pressure is constant regardless of the bubble expansion, was presented by Prousevitch et al. [79]. They assumed an initially supersaturated condition, in which both p_{lo} and p_{go} are lower than the saturation pressure of the volatile dissolved in the melt. They investigated the final bubble radius and the time to reach it, which correspond to R_f and T_{growth}, respectively. They also presented effects of initial bubble radius (R_o), number density of the bubbles (N), volatile diffusivity in the melt (κ_{gl}), initial pressure (p_o), and initial super-saturation.

A simple theory to estimate the time scale of re-equilibration is useful to compare the model with observations, but has not been determined yet. Here we test two hypotheses.

1. The recovery time is comparable with the time scale in which the diffusion layer develops over the entire shell, that is $\tau_1 = (S_f - R_f)^2/\kappa_{gl}$.

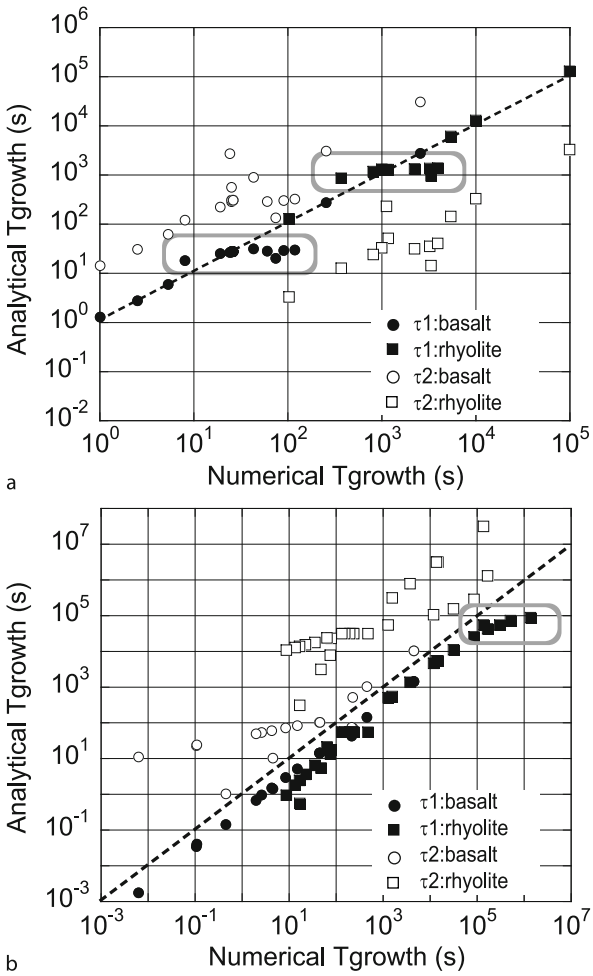

Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 7

Numerical results of the pressure recovery time for magma in an elastic rock a and the bubble growth time in an open system b for various system parameters are compared with analytical approximations: τ_1 is the time scale of mass diffusion across the final shell thickness, τ_2 is approximation by [53]. Agreement between the numerical results and τ_1 is better, but some systematic discrepancy remains, as indicated by *gray frames*. The numerical results for a are from [87], and those for b are from [79]

2. Based on a dimensional analysis of the simplified diffusion Eq. (14), Lensky et al. [53] proposed $\tau_2 = (R_f^2/\kappa_{gl})(\rho_{gf}/\rho_l)(C_o - C_f)^{-1}$, where ρ_{gf} and C_f are the final gas density in the bubble and the volatile concentration remained in the melt, both of which are functions of p_f. They obtained this equation based on the approximation that the quasi-static mass flux through the interface is $(C_o - C_{eq}(p_g))/R$.

The re-equilibration times, T_{growth}, obtained by Shimomura et al. [87] and Prousevitch et al. [79] are compared with the above models in Fig. 7. Comparing T_{growth}-τ_1 (black symbols) and T_{browth}-τ_2 (white symbols), we see that the general trend of T_{growth} is better estimated by τ_1 than by τ_2 in both confined and open systems. However, it should also be noted that τ_1 still has systematic errors which are indicated by gray frames. The errors are more dominant in the confined system (Fig. 7a). It is indicated that the simple estimation does not include all the factors relevant to the re-equilibration time and it is not necessarily applicable to the wider range of the parameters.

Re-equilibration Processes

Here we discuss different re-equilibration processes depending on the cause of the pressure drop and the relevant initial conditions. Three representative solutions are presented in Figs. 8–10. They are obtained for the standard basaltic system [87], but only viscosity is varied from 50 Pa s for (a) to 10^6 Pa s for (b).

Figure 8 is the case in which the stress drop occurs in the ambient rock first. It is generated by, for example, surface unloading by dome collapse [97] and stress change after a local earthquake. The condition is represented by $\sigma_\infty = -\Delta\sigma$ at $t \geq 0$ while $p_l = p_g - 2\Sigma/R_o = p_o$ at $t = 0$. According to Eq. (33), the chamber expands instantaneously, and p_l drops by Δp. Then the initial condition assumed by [64,87] is attained. Response of p_g is not instantaneous [62]. Due to the difference between p_g and p_l, the bubble expands according to Eq. (5) to decrease p_g. Then the difference between p_g and the equilibrium pressure for the volatile concentration in the melt occurs to make the volatile flow into the bubble to re-increase p_g. As the bubbles grow, the entire volume of the magma (δV) increases to enlarge the chamber elastically. Then the elastic stress $\bar{\mu}\delta V/V_o$ increases p_l according to Eq. (33). The re-equilibration proceeds in this way [87].

Figure 9 is the case in which the pressure in the bubble as well as those in the melt and the ambient rock is lower than the saturation pressure for the initial volatile concentration in the melt at $t = 0$. This condition occurs if bubbles are mixed with the supersaturated melt instantaneously, or if the bubbles are kept in the supersaturated mixture without interaction and suddenly allows the diffusion. Mathematically, the initial condition is equivalent to those assumed by [13] and [79]. The condition is represented by $p_l - p_o = p_g - 2\Sigma/R_o - p_o = \sigma_\infty = -\Delta p$ at $t \geq 0$. Diffusion of the volatile into the bubble starts, which increases p_g first. Then $p_g - p_l$ expands the bubble and the chamber to increase p_l in the same way as the previous case. Practically, the difference between Fig. 8 and Fig. 9 occurs only during a very short period in the beginning,

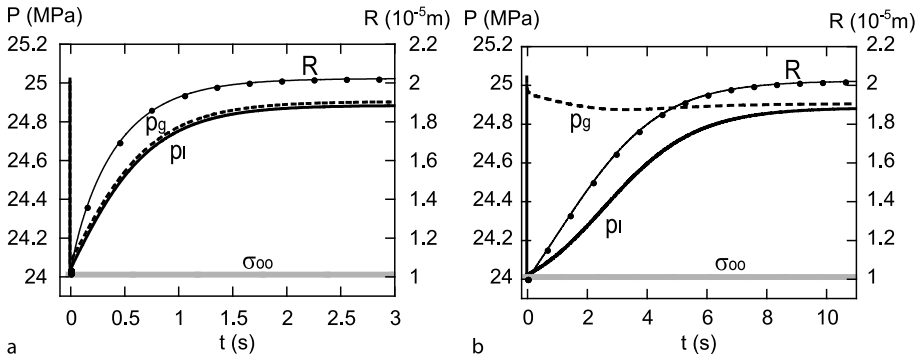

Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 8

Pressure recovery in a bubbly magma in an elastic chamber after sudden unloading $\sigma_\infty = -1$ MPa. The initial condition is $p_l = p_g - 2\Sigma/R = 25$ MPa and $R = 10^{-5}$m. The bubble radius on the right axis is plotted with a *line* and *points*. The stress and pressures on the left axis are plotted with a *solid line* for p_l, a *dotted line* for p_g, and *gray line* for σ_∞. The system parameters are $\kappa_{gl} = 10^{-8}$ m^2 s^{-1}, the bubble number density is 10^{11} m^{-3}, and $\eta_l = 50$ Pa s for a and 10^6 Pa s for b. The others are the same as those for the basaltic system by [87]

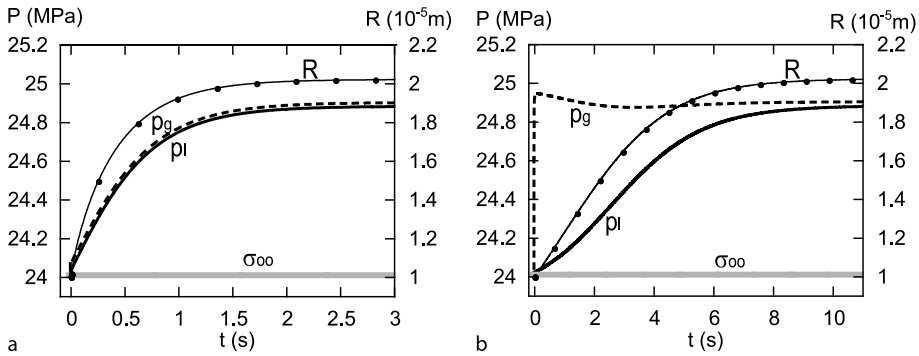

Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 9

Similar to Fig. 8, but the initial condition is $\sigma_\infty = p_l - p_o = p_g - 2\Sigma/R - p_o = -1$ MPa, with $p_o = 25$ MPa

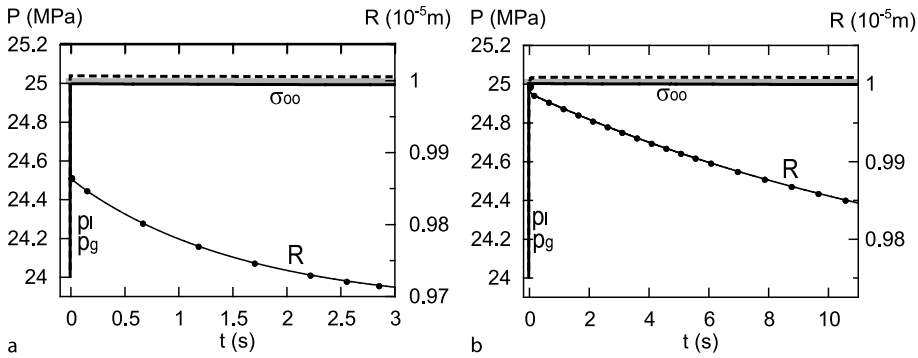

Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 10

Similar to Fig. 8, but the initial condition is $\sigma_\infty = 0$ and $p_l - p_o = p_g - 2\Sigma/R - p_o = -1$ MPa, with $p_o = 25$ MPa

and the subsequent increase of the pressure and volume of the chamber may look the same from outside.

Figure 10 is the case in which the pressure drop occurs in the melt and in the bubble, while the stress in the ambient rock is unchanged. The condition is represented by $p_l - p_o = p_g - 2\Sigma/R_o - p_o = -\Delta p$ and $\sigma_\infty = 0$. Although the assumed initial condition is rather imaginary, this case is presented in order to demonstrate how the response can be different depending on the way the system is decompressed. In fact, it is more realistic that p_l drops

first, while p_g remains at the initial value. This situation may occur by small leakage of the melt from the system. In this case, the melt pressure just recovers almost instantaneously, because the container compresses the melt according to Eq. (33) and bubbles also compress the melt. No other significant change is expected. On the other hand, if both p_l and p_g drop, as in Fig. 10, the pressure still recovers rapidly, but bubbles are compressed. Because the mechanical balance of the bubble and the melt is attained with $p_g - 2\Sigma/R = p_l$ according to Eq. (5), p_g has to become larger when R decreases. Then p_g exceeds the equilibrium pressure for the volatile in the melt to make the volatile dissolve into the melt.

Rectified Diffusion

So far, we discussed responses of the system to a stepwise pressure drop. When the perturbation is caused by a seismic wave from an external source, the system is subject to a cyclic disturbance. Rectified diffusion is a mechanism which can push dissolved volatiles into bubbles in a sound field. Bubbles take in more volatiles during expansion than they discharge during contraction, mainly because of the following two non-linear effects [18,27]. Firstly, the interface is larger during expansion than during contraction. Secondly, radial bubble expansion tangentially stretches the diffusion layer and sharpens the radial gradient of the volatile concentration in the diffusion layer, so that the volatile flux into the bubble is enhanced.

Brodsky et al. [6] discussed the possible pressure increase of a bubbly magma confined in an elastic rock by this mechanism. Using the solution by Hsieh and Plesset [27] for a periodic system, they considered that, even though the net pressure changes are determined by the pre-existing oversaturation, the rectified diffusion accelerates the pressure increase and may break the balance which had been stabilized the system prior to the oscillation. Ichihara and Brodsky [29] improved the solution by including resorption of gas as the pressure increase and development of the diffusion layer around the bubble in a self-consistent way. It is then shown that rectified diffusion is not faster than the ordinary diffusion and its contribution to the net pressure change is at the most 2×10^{-9} of the initial pressure regardless of the pre-existing oversaturation.

Acoustic Bubbles in Hydrothermal Systems

Pressure Impulses Generated in a Geyser

Here we consider a mixture of water and vapor bubbles, in which we expect effects of bubble oscillations and evaporation.

Kedar et al. [39,40] conducted field experiments at Old Faithful Geyser, Yellowstone. They measured pressure within the geyser's water column simultaneously with seismic measurements on the surface. The data show a distinct cause-and-effect relationship between the impulsive pressure source and the impulse response of the rock surrounding the water column. In addition, the pressure pulse, which is strongest at the top transducer, strongly attenuates downward. Considering that the pulse is generated by the oscillation of a single bubble, they compared one selected signal with a solution of the equation of motion for the bubble radius (Eq. (3)). In order to fit the measured oscillation with a reasonable bubble radius, they had to assume a very small ambient pressure to lower the frequency, and a very large viscosity to increase the damping. For $R_o = 0.055$ m, for example, they used $p_{go} = 0.02$ MPa, with which Eq. (9) gives the resonant frequency close to the observation: ~ 20 Hz. The viscosity was assumed as $\eta_l = 40$ Pa s, which is larger than the actual value by more than four orders. They compared the damping coefficient with those from radiation and heat transfer, though these effects were not included in the calculation of Eq. (3), and concluded that mechanisms other than acoustic, thermal, or viscous damping are required to explain the strong damping observed.

We have already introduced the damping coefficient b_t with the evaporation effect in Eq. (17) with Eq. (26). Then, assuming the similar bubble radius and frequency ranges as [40], let us see the damping coefficient by evaporation in comparison with the other coefficients, which are for viscous, acoustic and thermal damping, represented by Eqs. (8), (13), and (17) with Eq. (19), respectively. Their values are compared in Fig. 12b, assuming $p_{go} = 0.13$ MPa (the saturation pressure at 380 K). We can see that the evaporation effect significantly increases the damping and dominates the other damping mechanisms in the frequency range of the geyser oscillation. The evaporation effect also decreases $\mathrm{Re}(K_g)$ (Fig. 12a) in the range. It is thus suggested that the evaporation effect is significant for the bubble dynamics in the hydrothermal system.

Inertial and Thermal Collapse of a Bubble

When we heat water in a kettle, we hear strong intermittent pulses before boiling starts. The phenomenon is explained in terms of the bubble dynamics as follows [1]. In the first regime (which initiates above approximately 40 °C), small bubbles form slowly out of dissolved air in the liquid, rising silently to the surface as they break off the side of the vessel. At higher temperature (~ 70 °C), vapor

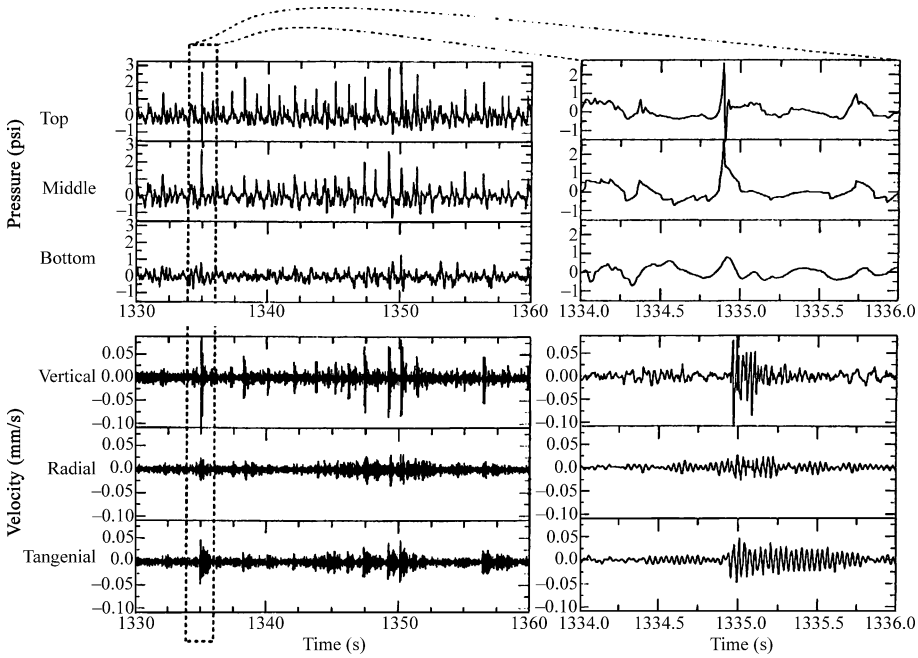

Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 11

Simultaneous pressure records (through a high-pass filter at 1 Hz) and seismic traces at Old Faithful geyser, Yellowstone. The geyser's eruptions are 2–5 min long with the interval between them ranging from 30–100 min. The figure shows a 30 s data about 27 min after the previous eruption and about 52 min before the next eruption. The conduit of the geyser is 22 m deep, where the bottom sensor was located. The bottom, middle and top sensors were connected 3 m apart. The seismic station was located at ∼ 25 m from the geyser. The data show a direct correspondence between the pressure pulses and the seismic signals that follow them. (Fig. 3 in [39])

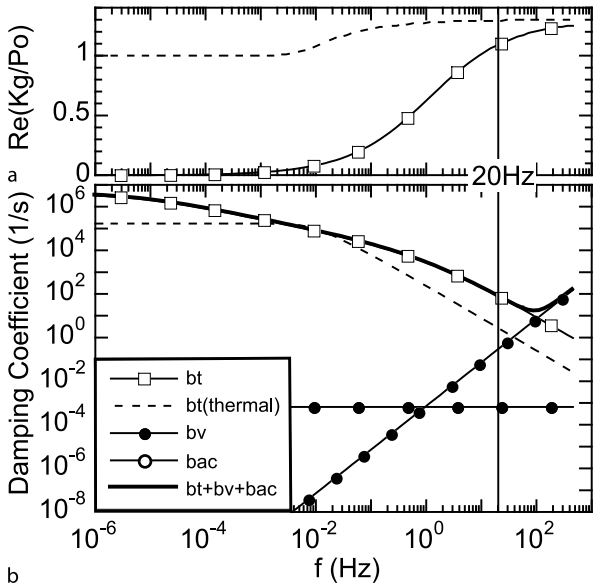

Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 12

The bubble elasticity (a) and damping factors (b) for a single vapor bubble with radius 0.055 m at 380 K, 0.13 MPa (saturation pressure)

bubbles start to nucleate at various sites at the heated bottom surface of the container. Vapor bubbles are different from the air bubbles in that their formation and collapse (at the bottom of the vessel) occurs explosively, producing pressure impulses that traverse the liquid and cause much of the sound we hear. In the third stage (between 90 °C and 100 °C), vapor bubbles grow, coalesce, and survive their ascent through the liquid. Bursting of vapor bubbles at the top surface is considered to be the sound source in this regime. Finally, the transition to full boil is characterized by large bubble formation throughout the bulk of the liquid.

We consider whether and how the impulse generation by the bubble collapse occurs in a geyser, where water that is already boiling is injected and cooled from the surface [43]. The collapse (or growth) of a bubble is classified into two modes: the inertia mode, which is controlled by the liquid inertia and driven by the pressure difference between the liquid and the bubble, and the thermal mode, which is controlled by the heat transfer and driven by the temperature difference [20,103]. The former is more violent than the latter and is responsible for the impulse generation.

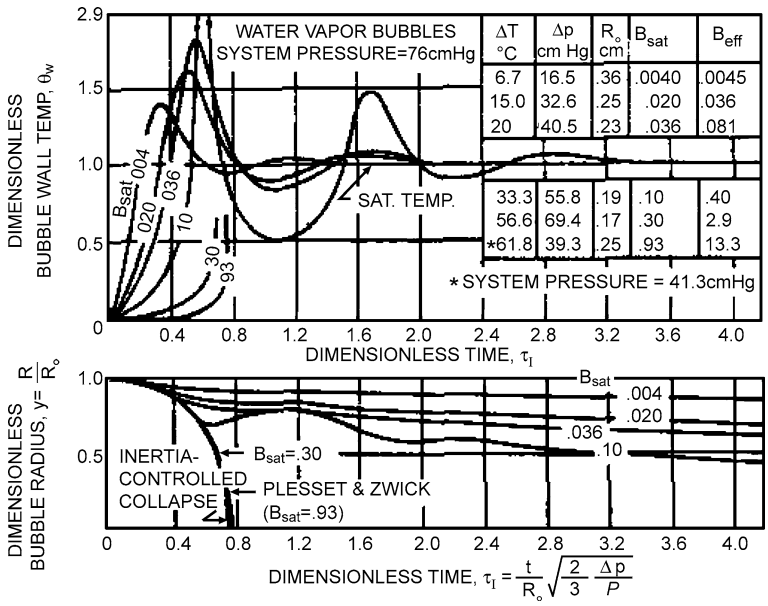

Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 13

Variation of wall temperature and radius during collapse of water vapor bubbles. B_{sat} is the dimensionless parameter given by Eq. (35), which determines relative importance of the liquid inertia and the heat transfer (Fig. 3 in [20])

Based on theoretical and experimental studies, Florschuetz and Chao [20] proposed that the relative importance of the inertia and the heat transfer is evaluated by a dimensionless parameter, B, defined by

$$Ja = \frac{\rho_l c_{pl}(T_{sat}(p_{lo}) - T_o)}{\rho_{go} L}, \tag{34}$$

$$B = Ja^2 \frac{\kappa_{Tl}}{R_o} \sqrt{\frac{\rho_l}{p_{lo} - p_{sat}(T_o)}}, \tag{35}$$

where $T_{sat}(p_{lo})$ is the saturation temperature for the ambient pressure (p_{lo}) and $p_{sat}(T_o)$ is the saturation pressure for the system temperature (T_o). The dimensionless parameter Ja is called the Jacob number, which represents the degree of subcooling. Figure 13 displays their calculation results, which clearly shows that for $B \geq 0.3$, the collapse rate is dominated by liquid inertia effect, while for $B \leq 0.03$ it is much slower and is recognized as the thermal mode. For an intermediate value of B, oscillation is observed.

In these works [20,77], it is often assumed that the pressure in the bubble is initially equal to the saturation pressure of the subcooled liquid, that is $p_{go} = p_{sat}(T_o)$, which is less than the ambient pressure (p_{lo}) [20,77]. Experimentally, it is achieved by preparing for a thermally equilibrium water-vapor system at a low pressure and suddenly increasing the system pressure to p_{lo} [20]. Then the initial collapse is relatively violent and continues by inertia until the vapor heating at the bubble wall increases the vapor pressure above the ambient pressure to such an extent that the liquid is momentarily brought to rest and its motion actually reverses before the vapor pressure again drops below the system pressure [20,77]. Although the oscillation is difficult to see on the radius change curves in Fig. 13 for small B, the beginning inertia controlled stage is evidenced by that all the curves start along the inertia curve.

On the other hand, in case that a bubble suddenly enters cold water, $p_{go} = p_{lo} > p_{sat}(T_o)$ while temperature in the bubble T_{go} is larger than T_o and is close to $T_{sat}(p_{go})$. Then the collapse begins in the gentle mode controlled by the heat transfer. It can turn into the inertia mode only if the rate of heat transport and condensation to decrease the vapor pressure is so large that inward motion of the surrounding liquid cannot follow. Prosperetti and Hao [77] presented that relative translational motion between the bubble and the liquid significantly increases the rate of heat transport and accelerates the bubble collapse. Furthermore, they pointed out the coupling effect between the translational and the radial motions. As Eq. (30) suggests, the decreasing bubble radius ($\dot{R} < 0$) works to accelerate the translational motion. Although the drag force ($\propto C_D R^{-1}|U|U$) increases as R decreases and U increases, there are cases in which the contribution of the first term is so large as to make $\dot{U} > 0$. Then the collapse and the translational motion of the bubble accelerate each other [77].

Rectified Heat Transfer

In the same way as the rectified diffusion discussed in the previous section, rectified heat transfer works for a vapor bubble in an acoustic field [24,98]. When the bubble is compressed, some vapor condenses, the surface temperature rises, and heat is conducted away into the adjacent liquid. When the bubble expands during the following half cycle, evaporation causes a temperature drop of the bubble surface, with a consequent heat flux from the liquid. The imbalance of the heat flux and the interface area between the compression phase and the expansion phase causes the net energy flux into the bubble.

Sturtevant et al. [89] investigated the effect of rectified diffusion on pressure increase in a hydrothermal system, as a possible mechanism for triggered seismicity by a distant earthquake. They modeled the system as a two-component H_2O-CO_2 system, and considered rectified mass diffusion. As is mentioned in the previous section, the net pressure change due to rectified mass diffusion is very small, if it is evaluated in a self-consistent way [29].

Although rectified heat transfer has a similar mechanism as the rectified mass transfer, it is much more intense since the thermal diffusivity of liquids typically exceeds the mass diffusivity by two orders of magnitude [77]. It can grow a vapor bubble quickly within several cycles of oscillation at the beginning (Fig. 14), and thus can be effective even with low-frequency pressure waves. Moreover the effect is reinforced by the coupling effect of the bubble growth and translational motion [25], and coupling of evaporation and diffusion of another gas component [38]. The net energy flux into the bubble increases the temperature in the bubble, which may change the liquid static pressure [24]. It is noted that numerical results in the literature cannot be directly used to estimate the pressure increase, because they were obtained for an open system in which the bubble continues to grow instead of increasing system pressure. However, it might be worth while to re-evaluate effects of rectified processes in a hydrothermal system taking account of these recent results.

Non-linear Oscillation of a Spherical Cloud of Bubbles

Oscillation of a group of bubbles generates particular signals as well as oscillation of a single bubble. The presence of bubbles can lower the acoustic speed of the fluid by an order of magnitude [16,42], if the liquid viscosity is small enough [30,31]. Therefore, there is a sharp impedance contrast between a region of bubbly liquid and a region of pure liquid. The boundary of a bubble cloud acts like an elastic boundary that traps acoustic energy in the bubbly region so that the bubble cloud has characteristic frequency of resonance [56,67,102]. Chouet [10] considered that it is the mechanism for the harmonic oscillations observed at volcanoes having relatively low-viscosity magma. He showed the typical values for the oscillation in the few Hz range may be generated by a columnar bubble cloud

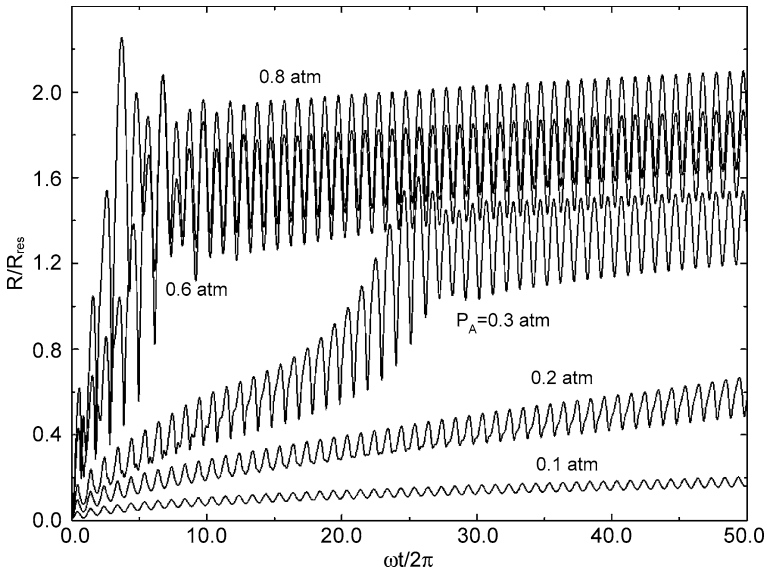

Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 14
Growth of a vapor bubble by rectified heat transfer. Bubble radius normalized by the linear resonant radius R_r = 2.71 mm versus time for saturate water at 1 atm. The sound frequency is 1 kHz and the amplitude is attached to each profile (Fig. 4 in [24])

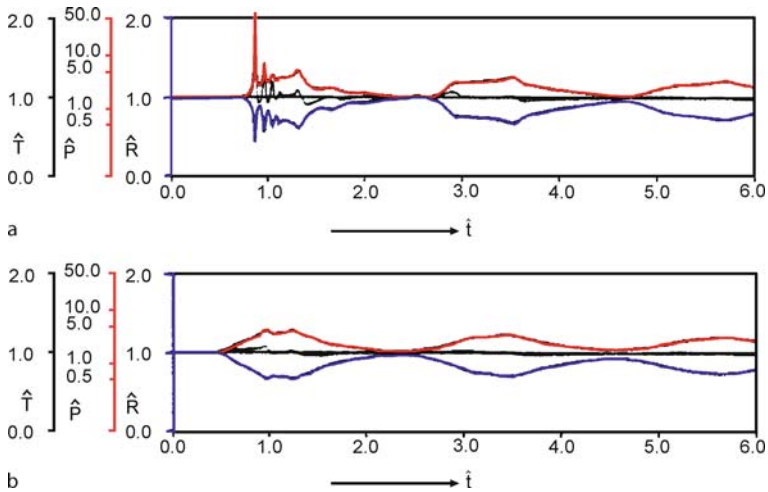

Pressure Impulses Generated by Bubbles Interacting with Ambient Perturbation, Figure 15
Change of bubble radius *(blue lines)* and pressure *(red lines)* in a spherical cloud of bubbles after a sudden pressure rise. The initial pressure and temperature are 3×10^4 Pa, 293 K, radii of the bubble and the cloud are 2.5×10^{-4} m and 3×10^{-2}, the void fraction is 3%, and the pressure rise is 2.4×10^4 Pa. Values at the center (a) and at a half the radius from the center (b) are presented in dimensionless form (a unit of the dimensionless time corresponds to ∼ ms). (Modified from Fig. 9 in [67])

with void fraction of 1%, ∼ 100 m in length, and ∼ 1 m in radius in a bubble-free magma with ∼ 5 m in radius. The radius of each bubble is assumed as 10^{-3} m, which has its own resonance frequency at 8.9 kHz [10].

Omta [67] conducted numerical calculation for oscillation of a spherical cloud of bubbles with relatively large amplitude. Figure 15 shows one of his results, in which the cloud oscillation was excited by an external pressure increase. Three cycles of the oscillation are presented. We can see high-frequency strong pulsation at the center of the bubble cloud (Fig. 15a). It is explained as follows [67]. The pressure perturbation is amplified and sharpened toward the center of the cloud because of the spherical geometry. Then the bubbles are excited at their resonance frequency.

The strong pulsation is observed only near the center in a spherical bubble cloud (Fig. 15). If a bubble cloud is hemispherical with its cross section attached on a solid wall, the pulsation generated at the center of the hemisphere may strongly hit the wall [86]. Generation of strong high-frequency pulses by a bubble cloud interacting with a pressure perturbation with lower frequencies is actually observed in experiments and now the phenomenon is going to be applied to medical treatment under controlled condition [33,60]. Similar mechanisms may work in a hydrothermal system, in which a low-frequency perturbation generates strong pressure impulses hitting the walls to be observed as seismic waves. In fact pressure oscillations in the bubble cloud (Fig. 15) and in the geyser (Fig. 11) have quite similar features, though their time scales are different by three orders of magnitude.

Future Directions

We have summarized theoretical bases of the bubble dynamics, mainly based on radial motion equations of a single bubble. These theories have been established and verified by experiments for simple systems. For volcanic systems, these theories have mainly been applied to the nucleation and growth of bubbles in magma. This subject takes an important role of volcanology, though it is not included in this review paper. By comparing the theory with observation of bubbles left in natural volcanic rock [58, 94] and re-producing the process by laboratory experiments [23,52,90], researchers have determined physical parameters of the volcanic processes, which are the temperature, pressure, volatile saturation, ascent rate, and so on. The bubble dynamics theories have also been applied to explain geophysical observations, as we have reviewed several possible mechanisms of bubbles that generate pressure impulses. However, determining the effects or even existence of bubbles is more difficult in these phenomena than in the bubble growth problems, because direct observation of bubbles and re-production of the process in a laboratory are more difficult. Here we discuss how we can go forward to confirm the models and apply them to determine useful physical parameters.

It would be effective to focus on relatively simple volcanic phenomena in which the bubble dynamics theory appears to work. Especially, for some volcanoes which erupt frequently, data taken by modern geophysical methods are being accumulated and the phenomenological cause-and-result relations between an eruption and pressure impulses before and during the eruption are well documented. For example, at Stromboli Volcano all the sequences of the repetitive small eruptions and a few proximal explosions have been taken by multi-parameter monitoring systems [84,85]. Geophysical data taken close to active craters at Sakurajima, Suwanosejima, and Semeru volcanoes have revealed common features in the pressure change before and during an explosion [32]. At Sakurajima Volcano, behaviors of a shallow gas pocket are discussed in the sequences of seismic and explosion events based on analyses of seismic data [91,92]. It might be possible and useful to have a common backbone model for these eruptions, based on which we can explain the particular detail of each case as a result of different parameters of the system.

The bubble dynamics theory which is used in the models needs to be updated, too. Responses of a bubbly fluid are sensitive to the system parameters, which determine possibilities, features, and time scales of the individual mechanisms, as are shown in the text. Although many models assume uniform system for simplicity, the natural system is considered to be non-uniform. In other words, regions having different physical parameters may coexist. Interaction of these subsystems may enhance the characteristic response, but may diminish with one another, or generate completely different effects. The inhomogeneities and their interactions may occur in various scales and manners. For example, in Sect. "Bubbly Magma in an Elastic Rock as a Pressure Source", behaviors of a single uniform magma body in an elastic rock have been discussed. In the real system, the bubble size and the chemical composition are likely to be non-uniform within a small region and/or over the entire magma body. If the system is large, the hydrostatic pressure gradient is significant. Moreover, if there are multiple magma containers which have different physical parameters, each magma body will respond to the pressure perturbation differently, and pressure gradient may be generated between the two adjacent containers. Developing a model including these subscale interactions may be a subject of modern multi-scale multi-physics studies.

Laboratory experiments using analogous materials are useful in verifying and improving the models. In the procedure to construct a model system, we frequently find factors which are important in the real system but have been neglected in the idealized mathematical model. Although there may be some processes which can be realized in the nature more easily, and there always be a scaling problem, experiments will give us more concrete idea about the mechanism of the models. Although most of the previous analogue experiments are designed to be compared with geological and petrological observations, there are some which investigate generation of pressure impulses by bubbles and are intended to explain seismic and/or acoustic observations [34,35,83]. According to the results and implications obtained by these preceding works, laboratory studies in this direction are promising.

It is also important to connect geophysical observation and geological data to understand the bubble dynamics phenomena in volcanology. Compared with other geophysical processes which occur in the earth, there is larger possibility that the source of activity appears to the surface after relatively short time. The geological samples (e. g., pylocrasts during eruptions and volcanic gasses) can inform us physical and chemical properties of the materials generating pressure impulses, and would make useful constraints on the model. On the other hand, these constraints can be tested and verified by geophysical signals of seismic, geodetic and acoustic measurements when the models are established.

By combining these theoretical, experimental, geophysical, and geological approaches, we will get better understanding on the processes which generate volcanic activities.

Acknowledgments

The authors are grateful to Dr. B. Chouet, and Dr. H. Kawashima for useful information and advise. We also thank Dr. M. Kameda and two anonymous reviewers for their help to improve the manuscript.

Bibliography

1. Aljishi S, Tatarkievicz J (1991) Why does heating water in a kettle produce sound? Am J Phys 59:628–632
2. Barclay J, Riley DS, Sparks RSJ (1995) Analytical models for bubble growth during decompression of high viscosity magmas. Bull Volcanol 57:422–431
3. Benoit JP, McNutt SR (1997) New constraints on source processes of volcanic tremor at Arenal volcano, Costa Rica, using broadband seismic data. Geophys Res Lett 24:449–452
4. Blower JD, Mader HM, Wilson SDR (2001) Coupling of viscous and diffusive controls on bubble growth during explosive volcanic eruptions. Earth Planet Sci Lett 193:47–56
5. Bowers TS (1995) Pressure-volume-temperature properties of H2O CO2 fluids. In: Ahrens TJ (ed) A Handbook of Physical Constants: Rock Physics and Phase Relations. AGU Reference Shelf Series 3, AGU, pp 45–72

6. Brodsky EE, Sturtevant B, Kanamori H (1998) Earthquakes, volcanoes, and rectified diffusion. J Geophys Res 103:23827–23838
7. Brodsky EE, Karakostas V, Kanamori H (2000) A new observation of dynamically triggered regional seismicity: Earthquakes in Greece following the august, 1999 Izmit, Turkey earthquake. Geophys Res Lett 27:2741–2744
8. Brodsky EE, Roeloffs E, Woodcock D, Gall I, Manga M (2003) A mechanism for sustained groundwater pressure changes induced by distant earthquakes. J Geophys Res 108(B8):2390. doi: 10.1029/2002JB002321
9. Campos FB, Lage PLC (2000) Heat and mass transfer modeling during the formation and ascension of superheated bubbles. Int J Heat Mass Transf 43:2883–2894
10. Chouet BA (1996) New methods and future trends in seismological volcano monitoring. In: Scarpa R, Tilling R (eds) Monitoring and Mitigation of Volcano Hazards. Springer, Berlin, pp 23–97
11. Chouet B, Dawson P, Ohminato T, Martini M (2003) Source mechanisms of explosions at Stromboli volcano, Italy, determined from moment-tensor inversions of very-long-period data. J Geophys Res 108
12. Chouet B, Dawson P, Arciniega-Ceballos A (2005) Source mechanism of vulcanian degassing at Popocatepetl volcano, Mexico, determined from waveform inversions of very long period signals. J Geophys Res 110:B07301
13. Chouet B, Dawson P, Nakano M (2006) Dynamics of diffusive bubble growth and pressure recovery in a bubbly rhyolitic melt embedded in an elastic solid. J Geophys Res 111: B07310
14. Cole RH (1948) Underwater Explosions. Dover, New York
15. Collier L, Neuberg JW, Lensky N, Lyakhovsky V, Navon O (2006) Attenuation in gas-charged magma. J Volcanol Geotherm Res 153:21–36
16. Commander KW, Prosperetti A (1989) Linear pressure waves in bubbly liquid – comparison between theory and experiments. J Acoust Soc Am 85:732–746
17. Doinikov AA (2005) Equations of coupled radial and translational motions of a bubble in a weakly compressible liquid. Phys Fluids 17:128101
18. Eller A, Flynn HG (1965) Rectified diffusion during nonlinear pulsations of cavitation bubbles. J Acoust Soc Am 37:493–503
19. Finch RD, Neppiras EA (1973) Vapor bubble dynamics. J Acoust Soc Am 53:1402–1410
20. Florschuetz LW, Chao BT (1965) On the mechanics of vapor bubble collapse. Trans ASME, J Heat Transf 87:209–220
21. Fogler HS, Goddard JD (1970) Collapse of spherical cavities in viscoelastic fluids. Phys Fluids 13:1135–1141
22. Garces MA, McNutt SR (1997) Theory of the airborne sound field generated in a resonant magma conduit. J Volcanol Geotherm Res 78:155–178
23. Gardner JE, Hilton M, Carroll MR (1999) Experimental constraints on degassing of magma: isothermal bubble growth during continuous decompression from high pressure. Earth Planet Sci Lett 168:201–218
24. Hao Y, Prosperetti A (1999) The dynamics of vapor bubbles in acoustic pressure fields. Phys Fluids 11:2008–2019
25. Hao Y, Prosperetti A (2002) Rectified heat transfer into translating and pulsating vapor bubbles. J Acoust Soc Am 112:1787–1796
26. Holloway JR, Blank JG (1994) Application of experimental results to C-O-H species in natural melts. In: Carroll MR, Holloway JR (eds) Volatiles in Magma. Rev Miner, vol 30. Mineral Soc Am, Washington, pp 187–230
27. Hsieh DY, Plesset MS (1961) Theory of rectified diffusion of mass into gas bubbles. J Acoust Soc Am 33:206–215
28. Ichihara M (2007) Dynamics of a spherical viscoelastic shell: Implications to a criterion for fragmentation/expansion of bubbly magma. Earth Planet Sci Lett 265:18–32
29. Ichihara M, Brodsky EE (2006) A limit on the effect of rectified diffusion in volcanic systems. Geophys Res Lett 33:L02316
30. Ichihara M, Kameda M (2004) Propagation of acoustic waves in a visco-elastic two-phase system: influences of the liquid viscosity and the internal diffusion. J Volcanol Geotherm Res 137:73–91
31. Ichihara M, Ohkunitani H, Ida Y, Kameda M (2004) Dynamics of bubbly oscillation and wave propagation in viscoelastic liquids. J Volcanol Geotherm Res 129:37–60
32. Iguchi M, Tameguri T, Yakiwara H (2006) Source mechanisms of volcanic explosion revealed by geophysical observations at Sakurajima, Suwanosejima and Semeru volcanoes. Eos Trans AGU 87 (Fall Meet Suppl); Abstract V31G–03
33. Ikeda T, Yoshizawa S, Tosaki M, Allen JS, Takagi S, Ohta N, Kitamura T, Matsumoto Y (2006) Cloud cavitation control for lithotripsy using high intensity focused ultrasound. Ultrasound Med Biol 32:1383–1397
34. James MR, Lane SJ, Chouet B, Gilbert JS (2004) Pressure changes associated with the ascent and bursting of gas slugs in liquid-filled vertical and inclined conduits. J Volcanol Geotherm Res 129:61–82
35. James MR, Lane SJ, Chouet BA (2006) Gas slug ascent through changes in conduit diameter: Laboratory insights into a volcano-seismic source process in low-viscosity magmas. J Geophys Res 111:B05201
36. Johnson JB, Aster RC, Kyle PR (2004) Triggering of volcanic eruptions. Geophys Res Lett 31:L14604. doi:10.1029/2004GL020020
37. Kameda M, Matsumoto Y (1996) Shock waves in a liquid containing small gas bubbles. Phys Fluids 8:322–335
38. Kawashima H, Ichihara M, Kameda M (2001) Oscillation of a vapor/gas bubble with heat and mass transport. Trans JSME, Ser B 67:2234–2242
39. Kedar S, Sturtevant B, Kanamori H (1996) The origin of harmonic tremor at old faithful geyser. Nature 379:708–711
40. Kedar S, Kanamori H, Sturtevant B (1998) Bubble collapse as the source of tremor at old faithful geyser. J Geophys Res 103:24283–24299
41. Keller JB, Kolodner II (1956) Damping of underwater explosion bubble oscillations. J Appl Phys 27:1152–1161
42. Kieffer SW (1977) Sound speed in liquid-gas mixtures: water-air and water-steam. J Geophys Res 82:2895–2904
43. Kieffer SW (1989) Geologic nozzles. Rev Geophysics 27:3–38
44. Kumagai H, Chouet BA (2000) Acoustic properties of a crack containing magmatic or hydrothermal fluids. J Geophys Res 105:25493–25512
45. Kumagai H, Chouet BA (2001) The dependence of acoustic properties of a crack on the resonance mode and geometry. Geophys Res Lett 28:3325
46. Kumagai H, Chouet BA, Nakano M (2002) Temporal evolution of a hydrothermal system in Kusatsu–Shirane volcano, Japan,

inferred from the complex frequencies of long-period events. J Geophys Res 107:2236
47. La Femina PC, Connor CB, Hill BE, Strauch W, Saballos JA (2004) Magma-tectonic interactions in Nicaragua: the 1999 seismic swarm and eruption of Cerro Negro volcano. J Volcanol Geotherm Res 137:187–199
48. Landau LD, Lifshitz EM (1986) Theory of Elasticity, 3rd edn. Butterworth, Oxford
49. Landau LD, Lifshitz EM (1987) Fluid Mechanics, 2nd edn. Pergamon Press, Oxford
50. Lensky NG, Lyakhovsky V, Navon O (2001) Radial variations of melt viscosity around growing bubbles and gas overpressure in vesiculating magmas. Earth Planet Sci Lett 186:1–6
51. Lensky NG, Lyakhovsky V, Navon O (2002) Expansion dynamics of volatile-supersaturated liquids and bulk viscosity of bubbly magmas. J Fluid Mech 460:39–56
52. Lensky NG, Navon O, Lyakhovsky V (2004) Bubble growth during decompression of magma: experimental and theoretical investigation. J Volcano Geotherm Res 129:7–22
53. Lensky NG, Niebo RW, Holloway JR, Lyakhovsky V, Navon O (2006) Bubble nucleation as a trigger for xenolith entrapment in mantle melts. Earth Planet Sci Lett 245:278–288
54. Linde AT, Sacks I (1998) Triggering of volcanic eruptions. Nature 395:888–890
55. Linde AT, Sacks I, Johnston MJS, Hill DP, Bilham RG (1994) Increased pressure from rising bubbles as a mechanism for remotely triggered seismicity. Nature 371:408–410
56. Lu NQ, Prosperetti A, Yoon SW (1990) Underwater noise emissions from bubble clouds. IEEE J Ocean Eng 15:275–281
57. Manga M, Brodsky E (2006) Seismic triggering of eruptions in the far field: Volcanoes and geysers. Ann Rev Earth Planet Sci 34:263–291
58. Mangan MT, Cashman KV (1996) The structure of basaltic scoria and reticulite and inferences for vesiculation, foam formation, and fragmentation in lava fountains. J Volcanol Geotherm Res 73:1–18
59. Matsumoto Y, Takemura F (1994) Influence of internal phenomena on gas bubble motion (effects of thermal diffusion, phase change on the gas–liquid interface and mass diffusion between vapor and noncondensable gas in the collapsing phase). JSME Int J, Ser B 37:288–296
60. Matsumoto Y, Allen JS, Yoshizawa S, Ikeda T, Kaneko Y (2005) Medical ultrasound with microbubbles. Exp Therm Fluid Sci 29:255–265
61. Nakoryakov VE, Pokusaev BG, Shreiber IR (1993) Wave Propagation in Gas-Liquid Media, 2nd edn. CRC Press, Boca Raton
62. Navon O, Lyakhovsky V (1998) Vesiculation processes in silicic magmas. In: Gilbert JS, Sparks RSJ (eds) The Physics of Explosive Volcanic Eruption. Geol Soc, Special Publications, 145, London, pp 27–50
63. Nigmatulin RI, Khabeev NS, Nagiev FB (1981) Dynamics, heat and mass-transfer of vapor-gas bubbles in a liquid. Int J Heat Mass Trans 24:1033–1044
64. Nishimura T (2004) Pressure recovery in magma due to bubble growth. Geophys Res Lett 31:L12613
65. Nishimura T, Ichihara M, Ueki S (2006) Investigation of the Onikobe geyser, NE Japan, by observing the ground tilt and flow parameters. Earth Planets Space 58:e21–e24
66. Ohl CD, Tijink A, Prosperetti A (2003) The added mass of an expanding bubble. J Fluid Mech 482:271–290
67. Omta R (1987) Oscillations of a cloud of bubbles of small and not so small amplitude. J Acoust Soc Am 82:1018–1033
68. Oura A, Yoshida S, Kudo K (1992) Rupture process of the Ito-Oki Japan earthquake of 1989 July 9 and interpretation as a trigger of volcanic-eruption. Geophys J Int 109:241–248
69. Plesset MS (1949) The dynamics of cavitation bubbles. J Appl Mech 16:277–282
70. Plesset MS, Prosperetti A (1977) Bubble dynamics and cavitation. Ann Rev Fluid Mech 9:145–185
71. Poritsky H (1952) The collapse or growth of a spherical bubble or cavity in a viscous fluid. Proc First Nat Cong Appl Mech 813–821
72. Prosperetti A (1982) A generalization of the rayleigh-plesset equation of bubble dynamics. Phys Fluids 25:409–410
73. Prosperetti A (1984) Acoustic cavitation series: part three, bubble phenomena in sound fields: part two. Ultrasonics 222:115–124
74. Prosperetti A (1984) Acoustic cavitation series: part two, bubble phenomena in sound fields: part one. Ultrasonics 222:69–78
75. Prosperetti A (1991) The thermal behavior of oscillating gas bubbles. J Fluid Mech 222:587–616
76. Prosperetti A (2004) Bubbles. Phys Fluids 16:1852–1865
77. Prosperetti A, Hao Y (2002) Vapor bubbles in flow and acoustic fields. Ann NY Acad Sci 974:328–347
78. Prosperetti A, Lezzi A (1986) Bubble dynamics in a compressible liquid, 1. 1st-order theory. J Fluid Mech 168:457–478
79. Prousevitch AA, Sahagian DL, Anderson AT (1993) Dynamics of diffusive bubble growth in magmas: Isothermal case. J Geophys Res 98:22283–22307
80. Rayleigh L (1917) On the pressure developed in a liquid during the collapse of a spherical cavity. Philos Mag 34:94–98
81. Ripepe M, Gordeev E (1999) Gas bubble dynamics model for shallow volcanic tremor at Stromboli. J Geophys Res 104:10639–10654
82. Ripepe M, Poggi P, Braun T, Gordeev E (1996) Infrasonic waves and volcanic tremor at Stromboli. Geophys Res Lett 23:181–184
83. Ripepe M, Ciliberto S, Della Schiava M (2001) Time constraints for modeling source dynamics of volcanic explosions at Stromboli. J Geophys Res 106:8713–8727
84. Ripepe M, Marchetti E, Poggi P, Harris A, Fiaschi AJL, Ulivieri G (2004) Seismic, acoustic and thermal network monitors the 2003 eruption of Stromboli volcano. EOS, Trans AGU 85:329
85. Ripepe M, Marchetti E, Ulivieri G, DelleDonne D, Genco R, Lacanna G (2007) Monitoring the 2007 Stromboli effusive eruption by an integrated geophysical network. The 21st Century COE Earth Sci Int Symp Abstr Z 327
86. Shimada M, Matsumoto Y, Kobayashi T (2006) Influence of the nuclei size distribution on the collapsing behavior of the cloud cavitation. JSME Int J Ser B 155:307–322
87. Shimomura Y, Nishimura T, Sato H (2006) Bubble growth processes in magma surrounded by an elastic medium. J Volcanol Geotherm Res 155:307–322
88. Sparks RSJ (1978) Dynamics of bubble formation and growth in magmas – review and analysis. J Volcanol Geotherm Res 3:1–37
89. Sturtevant B, Kanamori H, Brodsky EE (1996) Seismic triggering by rectified diffusion in geothermal systems. J Geophys Res 101:25269–25282

90. Suzuki Y, Gardner JE, Larsen JF (2007) Experimental constraints on syneruptive magma ascent related to the phreatomagmatic phase of the 2000AD eruption of Usu volcano, Japan. Bull Volcanol 69:423–444
91. Tameguri T, Iguchi M, Ishihara K (2002) Mechanism of explosive eruptions from moment tensor analyses of explosion earthquakes at Sakurajima volcano. Bull Volcanol Soc Japan 47:197–215
92. Tameguri T, Maryanto S, Iguchi M (2007) Source mechanisms of harmonic tremors at Sakurajima volcano. Bull Volcanol Soc Japan 52:273–279
93. Toramaru A (1995) Numerical study of nucleation and growth of bubbles in viscous magmas. J Geophys Res 100:1913–1931
94. Toramaru A (2006) BND (bubble number density) decompression rate meter for explosive volcanic eruptions. J Volcanol Geotherm Res 154:303–316
95. Vergniolle S, Brandeis G (1994) Origin of the sound generated by Strombolian explosions. Geophys Res Lett 21:1959–1962
96. Vergniolle S, Brandeis G (1996) Strombolian explosions. 1. A large bubble breaking at the surface of a lava column as a source of sound. J Geophys Res 101:20433–20447
97. Voight B, Linde AT, Sacks IS, Mattioli GS, Sparks RSJ, Elsworth D, Hidayat D, Malin PE, Shalev E, Widiwijayanti C, Young SR, Bass V, Clarke A, Dunkley P, Johnston W, McWhorter N, Neuberg J, Williams P (2006) Unprecedented pressure increase in deep magma reservoir triggered by lava-dome collapse. Geophys Res Lett 33:L03,312
98. Wang T (1974) Rectified heat transfer. J Acoust Soc Am 56:1131–1143
99. Webb SL (1997) Silicate melts: Relaxation, rheology, and the glass transition. Rev Geophys 35:191–218
100. Yamada K, Emori H, Nakasawa N (2006) Bubble expansion rates in viscous compressible liquid. Earth Planets Space 58:865–872
101. Yang B, Prosperetti A, Takagi S (2003) The transient rise of a bubble subject to shape or volume changes. Phys Fluids 15:2640–2648
102. Yoon SW, Crum LA, Prosperetti A, Lu NQ (1991) An investigation of the collective oscillations of a bubble cloud. J Acoust Soc Am 89:700–706
103. Zuber N (1961) The dynamics of vapor bubbles in nonuniform temperature fields. Int J Heat Mass Transf 2:83–98

Principal-Agent Models

Inés Macho-Stadler, David Pérez-Castrillo
Universitat Autònoma de Barcelona, Barcelona, Spain

Article Outline

Glossary
Definition of the Subject
Introduction
The Base Game
Moral Hazard
Adverse Selection
Future Directions
Bibliography

Glossary

Information economics Information economics studies how information and its distributions among the players affect economic decisions.

Asymmetric information In a relationship or a transaction, there is asymmetric information when one party has more or better information than other party concerning relevant characteristics of the relationship or the transaction. There are two types of asymmetric information problems: Moral Hazard and Adverse Selection.

Principal–agent The principal-agent model identifies the difficulties that arise in situations where there is asymmetric information between two parties and finds the best contract in such environments. The "principal" is the name used for the contractor while the "agent" corresponds to the contractee. Both principal and agent could be individuals, institutions, organizations, or decision centers. The optimal solutions propose mechanisms that try to align the interests of the agent with those of the principal, such as piece rates or profit sharing; or that induce the agent to reveal the information, such as self-reporting contracts.

Moral hazard (hidden action) The term moral hazard initially referred to the possibility that the redistribution of risk (such as insurance which transfers risk from the insured to the insurer) changes people's behavior. This term, which has been used in the insurance industry for many years, was studied first by Kenneth Arrow.

In principal-agent models, the term moral hazard is used to refer to all environments where the ignorant party lacks information about the behavior of the other party *once* the agreement has been signed, in such a way that the asymmetry arises *after* the contract is settled.

Adverse selection (hidden information) The term adverse selection was originally used in insurance. It describes a situation where, as a result of private information, the insured are more likely to suffer a loss than the uninsured (such as offering a life insurance contract at a given premium may imply that only the people with a risk of dying over the average take it).

In principal-agent models, we say that there is an adverse selection problem when the ignorant party lacks information while negotiating a contract, in such a way that the asymmetry is *previous* to the relationship.

Definition of the Subject

Principal-Agent models provide the theory of *contracts under asymmetric information*. Such a theory analyzes the characteristics of optimal contracts and the variables that influence these characteristics, according to the behavior and information of the parties to the contract. This approach has a close relation to *game theory* and *mechanism design*: it analyzes the strategic behavior by agents who hold private information and proposes mechanisms that minimize the inefficiencies due to such strategic behavior. The costs incurred by the principal (the contractor) to ensure that the agents (the contractees) will act in her interest are some type of *transaction cost*. These costs include the tasks of investigating and selecting appropriate agents, gaining information to set performance standards, monitoring agents, bonding payments by the agents, and residual losses.

Principal-agent theory (and *Information Economics* in general) is possibly the area of economics that has evolved the most over the past twenty-five years. It was initially developed in parallel with the new economics of *Industrial Organization* although its applications include now almost all areas in economics, from finance and political economy to growth theory.

Some early papers centered on incomplete information in insurance contracts, and more particularly on moral hazard problems, are Spence and Zeckhauser [88] and Ross [81]. The theory soon generalized to dilemmas associated with contracts in other contexts [38,46]. It was further developed in the mid-seventies by authors such as Pauly [72,73], Mirrlees [66], Harris and Raviv [39], and Holmström [40]. Arrow [7] worked on the analysis of the optimal incentive contract when the agent's effort is not verifiable.

A particular case of adverse selection is the one where the type of the agent relates to his valuation of a good. Asymmetric information about buyers' valuation of the objects sold is the fundamental reason behind the use of auctions. Vickrey [92] provides the first formal analysis of the first and second-prize auctions. Akerlof [3] highlighted the issue of adverse selection in his analysis of the market for second-hand goods. Further analyzes include the early work of Mirrlees [65], Spence [87], Rothschild and Stiglitz [83], Mussa and Rosen [68], as well as Baron and Myerson [10], and Guesnerie and Laffont [36].

The importance of the topic has also been recognized by the Nobel Foundation. James A. Mirrlees and William Vickrey were awarded with the Nobel Prize in Economics in 1996 "for their fundamental contributions to the economic theory of incentives under asymmetric information". Five years later, in 2001, George A. Akerlof, A. Michael Spence and Joseph E. Stiglitz also obtained the Nobel Prize in Economics "for their analyzes of markets with asymmetric information".

Introduction

The objective of the Principal-Agent literature is to analyze situations in which a contract is signed under asymmetric information, that is, when one party knows certain relevant things of which the other party is ignorant. The simplest situation concerns a bilateral relationship: the contract between one principal and one agent. The objective of the contract is for the agent to carry out actions on behalf of the principal; and to specify the payments that the principal will pass on to the agent for such actions.

In the literature, it is always assumed that the principal is in charge of designing the contract. The agent receives an offer and decides whether or not to sign the contract. He will accept it whenever the utility obtained from it is greater than the utility that the agent would get from not signing. This utility level that represents the agent's outside opportunities is his *reservation utility*. In order to simplify the analysis, it is assumed that the agent cannot make a counter offer to the principal. This way of modeling implicitly assumes that the principal has all the bargaining power, except for the fact that the reservation utility can be high in those cases where the agent has excellent outside opportunities.

If the agent decides not to sign the contract, the relationship does not take place. If he does accept the offer, then the contract is implemented. It is crucial to notice that the contract is a reliable promise by both parties, stating the principal and agent's obligations for all (contractual) contingencies. It can only be based on *verifiable variables*, that is, those for which it is possible for a third party (a court) to verify whether the contract has been fulfilled. When some players know more than others about relevant variables, we have a situation with asymmetric information. In this case, *incentives* play an important role.

Given the description of the game played between principal and agent, we can summarize its timing in the following steps:

(i) The principal designs the contract (or set of contracts) that she will offer to the agent, the terms of which are not subject to bargaining.
(ii) The alternatives opened to the agent are to accept or to reject the contract. The agent accepts it if he desires so, that is, if the contract guarantees him greater expected utility than any other (outside) opportunities available to him.

(iii) The agent carries out an action or effort on behalf of the principal.
(iv) The outcome is observed and the payments are done.

From these elements, it can be seen that *the agent's objectives may be in conflict with those of the principal*. When the information is asymmetric, the informed party tries to take advantage, while the uninformed player tries to control this behavior via the contract. Since a Principal-Agent problem is a sequential game, the solution concept to use is *Subgame (Bayesian) Perfect Equilibrium*.

The set-up gives rise to three possible scenarios:

1. The *Symmetric Information* case, where the two players share the same information, even if they both may ignore some important elements (some elements may be uncertain).
2. The *Moral Hazard* case, where the asymmetry of information arises once the contract has been signed: the decision or the effort of agent is not verifiable and hence it cannot be included in the contract.
3. The *Adverse Selection* case, where the asymmetry of information is previous to the signature of the contract: a relevant characteristic of the agent is not verifiable and hence the principal cannot include it in the contract.

To see an example of moral hazard, consider a laboratory or research center (the principal) that contracts a researcher (the agent) to work on a certain project. It is difficult for the principal to distinguish between a researcher who is thinking about how to push the project through, and a researcher who is thinking about how to organize his evening. It is precisely this difficulty in controlling effort inputs, together with the inherent uncertainty in any research project, what generates a moral hazard problem, which is a non-standard labor market problem.

For an example of adverse selection, consider a regulator who wants to set the price of the service provided by a public monopoly equal to the average costs in the firm (to avoid subsidies). This policy (as many others) is subject to important informational requirements. It is not enough that the regulator asks the firm to reveal the required information in order to set the adequate price, since the firm would attempt to take advantage of the information. Therefore, the regulator should take this problem into account.

The Base Game

Consider a contractual relationship between a principal and an agent, who is contracted to carry out a task. The relationship allows a certain *result* to be obtained, whose monetary value will be referred to as x. For the sake of exposition, the set of possible results X is assumed to be finite, $X = \{x_1, \ldots, x_n\}$. The final result depends on the *effort* that the agent devotes to the task, which will be denoted by e, and the value of a random variable for which both participants have the same prior distribution. The probability of result x_i conditional on effort e can be written as:

$$\text{Prob}[x = x_i | e] = p_i(e) \,, \quad \text{for} \quad i \in \{1, 2, \ldots, n\} \,,$$

with $\sum_{i=1}^{n} p_i(e) = 1$. Let us assume that $p_i(e) > 0$ for all e, i, which implies that no result can be ruled out for any given effort level.

The Base Game is the reference situation, where principal and agent have the same information (even the one concerning the random component that affects the result). Since uncertainty exists, participants react to risk. Risk preferences are expressed by the shape of their *utility functions* (of the von Neumann–Morgenstern type). The principal, who owns the result and must pay the agent, has preferences represented by the utility function

$$B(x - w) \,,$$

where w represents the payoff made to the agent. $B(.)$ is assumed to be increasing and concave: $B' > 0$, $B'' \leq 0$ (where the primes represent, respectively, the first and second derivatives). The concavity of the function $B(.)$ indicates that the principal is either risk-neutral or risk averse.

The agent receives a monetary pay-off for his participation in the relationship, and he supplies an effort which implies some cost to him. For the sake of simplicity we represent his utility function as:

$$U(w, e) = u(w) - v(e) \,,$$

additively separable in the components w and e. This assumption implies that the agent's risk-aversion does not vary with the effort he supplies (many results can be generalized for more general utility functions). The utility derived from the wage, $u(w)$, is increasing and concave:

$$u'(w) > 0 \,, \quad u''(w) \leq 0 \,,$$

thus the agent may be either risk-neutral, $u''(w) = 0$, or risk-averse $u''(w) < 0$. In addition, greater effort means greater disutility. We also assume that the marginal disutility of effort is not decreasing: $v'(e) > 0$, $v''(e) \geq 0$.

A *contract* can only be based in verifiable information. In the base game, it can depend on the characteristics of the principal and the agent and includes both the effort e

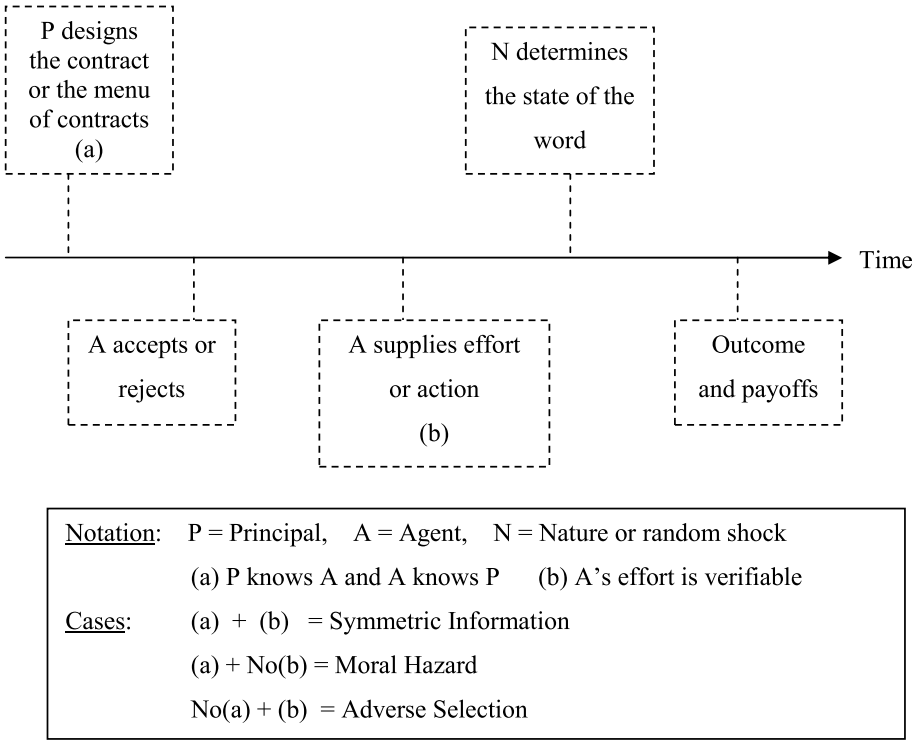

Principal-Agent Models, Figure 1
The figure summarizes the timing of the relationship and the three cases as a function of the information available to the participants

that the principal demands from the agent, and the wages $\{w(x_i)\}_{i=1,\ldots,n}$.

If the agent rejects the contract, he will have to fall back on the outside opportunities that the market offers him. These other opportunities, that by comparison determine the limit for participation in the contract, are summarized in the agent's *reservation utility*, denoted by $\underline{U}$. So the agent will accept the contract as long as he earns an expected utility equal or higher than his reservation utility. Since, the principals problem is to design a contract that the agent will accept, (by backward induction) the optimal contract must satisfy the *participation constraint* and it is the solution to the following maximization problem:

$$\max_{[e,\{w(x_i)\}_{i=1,\ldots,n}]} \sum_{i=1}^{n} p_i(e)B(x_i - w(x_i))$$
$$\text{s.t.} \sum_{i=1}^{n} p_i(e)u(w(x_i)) - v(e) \geq \underline{U}\,.$$

The above problem corresponds to a *Pareto Optimum* in the usual sense of the term. The solution to this problem is conditional on the value of the parameter $\underline{U}$, so that even those cases where the agent can keep a large share of the surplus are taken into account.

The principal's program is well behaved with respect to payoffs given the assumptions on $u(w)$. Hence the Kuhn–Tucker conditions will be both necessary and sufficient for the global solution of the problem. However, we cannot ascertain the concavity (or quasi-concavity) of the functions with respect to effort given the assumptions on $v(e)$, because these functions also depend on all the $p_i(e)$. Hence it is more difficult to obtain global conclusions with respect to this variable.

Let us denote by e° the efficient effort level. From the first-order Kuhn–Tucker conditions with respect to the wages in the different contingencies, we can analyze the associated pay-offs $\{w^\circ(x_i)_{i=1,\ldots,n}\}$. We obtain the following condition:

$$\lambda^\circ = \frac{B'(x_i - w^\circ(x_i))}{u'(w^\circ(x_i))}\,, \quad \text{for all} \quad i \in \{1, 2, \ldots, n\}\,,$$

where λ° is the multiplier associated with the participation constraint. When the agent's utility is additively separable, the participation constraint binds (λ° is positive). The previous condition equates marginal rates of substitution and indicates that the optimal distribution of risk requires that the ratio of marginal utilities of the principal and the agent to be constant irrespective of the final result.

If the principal is risk-neutral ($B''(.) = 0$), then the optimal contract has to be such that $u'(w^\circ(x_i)) =$ constant for all i. In addition, if the agent is risk-averse ($u''(.) < 0$), he receives the same wage, say w°, in all contingencies. This wage only depends on the effort demanded and is determined by the participation constraint. If the agent is risk-neutral ($u''(.) = 0$) and the principal is risk-averse ($B''(.) < 0$), then we are in the opposite situation. In this scenario, the optimal contract requires the principal's profit to be independent of the result. Consequently, the agent bears all the risk, insuring the principal against variations in the result. When both the principal and the agent are risk-averse, each of them needs to accept a part of the variability of the result. The precise amount of risk that each of them supports depends on their degrees of risk-aversion. Using the Arrow–Pratt measure of absolute risk-aversion $r_\mathrm{p} = -B''/B'$ and $r_\mathrm{a} = -u''/u'$ for the principal and the agent respectively, we can show that:

$$\frac{\mathrm{d}w^\circ}{\mathrm{d}x_i} = \frac{r_\mathrm{p}}{r_\mathrm{p} + r_\mathrm{a}},$$

which indicates how the agent's wage changes given an increase in the result x_i. Since $r_\mathrm{p}/(r_\mathrm{p} + r_\mathrm{a}) \in (0, 1)$, when both participants are risk-averse, the agent only receives a part of the increased result via a wage increase. The more risk-averse is the agent, that is to say, the greater is r_a, the less the result influences his wage. On the other hand, as the risk-aversion of the principal increases, greater r_p, changes in the result correspond to more important changes in the wage.

Moral Hazard

Basic Moral Hazard Model

Here we concentrate on *moral hazard*, which is the case in which the informational asymmetry relates to the agent's behavior during the relationship. We analyze the optimal contract when the agent's effort is not verifiable. This implies that effort cannot be contracted upon, because in case of the breach of contract, no court of law could know if the contract had really been breached or not. There are many examples of this type of situation. A traditional example is accident insurance, where it is very difficult for the insurance company to observe how careful a client has been to avoid accidents.

The principal will state a contract based on any signals that reveal information on the agent's effort. We will assume that only the result of the effort is verifiable at the end of the period and, consequently, it will be included in the contract. However, if possible, the contract should be contingent on many other things. Any information related to the state of nature is useful, since it allows better estimations of the agent's effort thus reducing the risk inherent in the relationship. This is known as the *sufficient statistic result*, and it is perhaps the most important conclusion in the moral hazard literature [40]. The empirical content of the sufficient statistic argument is that a contract should exploit all available information in order to filter out risk optimally.

The timing of a moral hazard game is the following. In the first place, the principal decides what contract to offer the agent. Then the agent decides whether or not to accept the relationship, according to the terms of the contract. Finally, if the contract has been accepted, the agent chooses the effort level that he most desires, given the contract that he has signed. This is a free decision by the agent since effort is not a contracted variable. Hence, the principal must bear this in mind when she designs the contract that defines the relationship.

To better understand the nature of the problem faced by the principal, consider the case of a risk-neutral principal and a risk-averse agent, which implies that, under the symmetric information, the optimal contract is to completely insure the agent. However, if the principal proposes this contract when the agent's effort is not a contracted variable, once he has signed the contract the agent will exert the effort level that is most beneficial for him. Since the agent's wage does not depend on his effort, he will use the lowest possible effort.

The idea underlying an incentive contract is that the principal can make the agent interested in the consequences of his behavior by making his pay-off dependent on the result obtained. Note that this has to be done at the cost of distorting the optimal risk sharing among both participants. The trade-off between *efficiency*, in the sense of the optimal distribution of risk, and *incentives* determines the optimal contract.

Formally, since the game has to be solved by backwards induction, the *optimal contract* under moral hazard is the solution to the maximization problem:

$$\max_{[e,\{w(x_i)\}_{i=1,\ldots,n}]} \sum_{i=1}^{n} p_i(e)B(x_i - w(x_i))$$

$$\text{s.t.} \sum_{i=1}^{n} p_i(e)u(w(x_i)) - v(e) \geq \underline{U}$$

$$e \in \operatorname{Arg\,Max}_{\hat{e}} \left\{ \sum_{i=1}^{n} p_i(\hat{e})u(w(x_i)) - v(\hat{e}) \right\}$$

The second restriction is the *incentive compatibility constraint* and the first restriction is the participation constraint. The incentive compatibility constraint, and not the principal as under symmetric information, determines the effort of the agent.

The first difficulty in solving this program is related to the fact that the incentive compatibility constraint is a maximization problem. The second difficulty is that the expected utility may fail to be concave in effort. Hence, to use the first order condition of the incentive compatibility constraint may be incorrect. In spite of this, there are several ways to proceed when facing to this problem. (a) Grossman and Hart [33] propose to solve it in steps, identifying first the optimal payment mechanism for any effort and then, if possible, the optimal effort. This can be done since the problem is concave in payoffs. (b) The other possibility is to consider situations where the agent's maximization problem is well defined. One possible scenario is when the set of possible efforts is finite, in which case the incentive compatibility constraint takes the form of a finite set of inequalities. Another scenario is to write the incentive compatibility as the first-order condition of the maximization problem, and introduce assumptions that allow doing it. The last solution is known as the *first-order approach*.

Let us assume that the first-order approach is adequate, and substitute the incentive compatibility constraint in the previous program by

$$\sum_{i=1}^{n} p_i'(\hat{e})u(w(x_i)) - v'(\hat{e}) = 0 \,.$$

Solving the principals program with respect to the payoff scheme, and denoting by λ (resp., μ) the Lagrangean multiplier of the participation constraint (resp., the incentive compatibility constraint), we obtain that for all i:

$$\frac{1}{u'(w(x_i))} = \lambda + \mu \frac{p_i'(e)}{p_i(e)} \,.$$

This condition shows that the wage should not depend at all on the value that the principal places on the result. It depends on the results as a measure of how informative they are as to effort, in order to serve as an incentive for the agent. The wage will be increasing in the result as long as the result is increasing in effort. Hence, it is optimal that the wage will be increasing in the result only in particular cases. The necessary condition for a wage to be increasing with results is $p_i'(e)/p_i(e)$ to be decreasing in i. In statistics, this is called the *monotonous likelihood quotient property*. It is a strong condition; for example, first-order stochastic dominance does not guarantee the monotonous likelihood property.

Extensions of Moral Hazard Models

The basic moral hazard setup, with a principal hiring and an agent performing effort, has been extended in several directions to take into account more complex relationships.

Repeated Moral Hazard Certain relationships in which a moral hazard problem occurs do not take place only once, but they are repeated over time (for example, work relationships, insurance, etc.). The duration aspect (the repetition) of the relationship gives rise to new elements that are absent in static models.

Radner [76] and Rubinstein and Yaari [84] consider infinitely repeated relationships and show that frequent repetition of the relationship allows us to converge towards the efficient solution. Incentives are not determined by the payoff scheme contingent on the result of each period, but rather on average effort, and the information available is very precise when the number of periods is large. A sufficiently threatening punishment, applied when the principal believes that the agent on average does not fulfill his task, may be sufficient to dissuade him from shirking.

When the relationship is repeated a finite number of times, the analysis of the optimal contract concentrates on different issues relating *long-term* agreements and *short-term* contracts. Note that in a repeated set up, the agent's wage and the agent's consumption in a period need not be equal. Lambert [54], Rogerson [80], and Chiappori and Macho-Stadler [15] show that long-term contracts have memory (i. e., the pay-offs in any single period will depend on the results of all previous periods) since they internalize agent's consumption over time, which depends on the sequence of payments received (as a function of the past contingencies). Malcomson and Spinnewyn [58], Fudenberg, Holmström, and Milgrom [30], and Rey and Salanié [77] study when the optimal long-term contract can be implemented through the sequence of optimal short-term contracts. Chiappori, Macho-Stadler, Rey, and Salanié [16] show that, in order for the sequence of optimal short-term contracts to admit the same solution as the long-term contract, two conditions must be met. First, the optimal sequence of single-period contracts should have memory. That is why, when the reservation utility is invariant (is not history dependent), the optimal sequence of short-term contracts will not replicate the long-term optimum unless there exist means of smoothing consumption, that is, the

agent has access to credit markets. Second, the long-term contract must be renegotiation-proof. A contract is said to be renegotiation-proof if at the beginning of any intermediate period, no new contract or renegotiation that would be preferred by all participants is possible. When the long term contract is not renegotiation-proof (i. e., if it is not possible for participants to change the clauses of the contract at a certain moment of time even if they agree), it cannot coincide with the sequence of short term contracts.

One Principal and Several Agents When a principal contracts with more than on agent, the stage where agents exert their effort, which is translated into the incentive compatibility, depends on the *game among the agents*. If the agents behave as a coordinated and cooperating group, then the problem is similar to the previous one where the principal hires a team. A more interesting case appears when agents play a non-cooperative game and their strategies form a Nash equilibrium.

Holmström [40] and Mookherjee [67], in models where there is *personalized information* about the output of each agent, show that the principal is interested in paying each agent according to his own production and that of the other agents if these other results can inform on the actions of the agent at hand. Only if the results of the other agents do not add information or, in other words, if an agent's result is a *sufficient statistic* for his effort, then he will be paid according to his own result.

When the only verifiable outcome is the final result of teamwork (*joint production models*), the optimal contract can only depend on this information and the conclusions are similar to those obtained in models with only one agent. Alchian and Demsetz [5] and Holmström [41] show that joint production cannot lead to efficiency when all the income is distributed amongst the agents, i. e., if the budget constraint always binds. Another player should be contracted to control the productive agents and act as the residual claimant of the relationship.

Tirole [90] and Laffont [49] have studied the effect of coalitions among the agents in an organization on their payment scheme. If collusion is bad for the organization, it adds another dimension of moral hazard (the colluding behavior). The principal may be obliged to apply rules that are collusion-proof, which implies more constraints and simpler contracts (more bureaucratic). When coordination can improve the input of a group of agents, the optimal contract has to find payment methods that strengthen group work (see [44,56]).

Another principal's decision when she hires several agents is the organization with which she will relate. This includes such fundamental decisions as how many agents to contract and how should they be structured. These issues have been studied by Demski and Sappington [23], Melumad and Reichelstein [63], and Macho-Stadler and Pérez-Castrillo [57].

Holmström and Milgrom [42] analyze a situation in which the agent carries out *several tasks*, each one of which gives rise to a different result. They study the optimal contract when tasks are complementary (in the sense that exerting effort in one reduces the costs of the other) or substitutes. Their model allows to build a theory of job design and to explain the relationship among responsibility and authority.

Several Principals and One Agent When one agent works for (or signs his contracts with) several principals simultaneously (*common agency* situation), in general, the principals are better off if they cooperate. When the principals are not able to achieve the coordination and commitment necessary to act as a single individual and they do not value the results in the same way, they each demand different efforts or actions from the agent. Bernheim and Whinston [12] show that the effort that principals obtain when they do not cooperate is less than the effort that would maximize their collective profits. However, the final contract that is offered to the agent minimizes the cost of getting the agent to choose the contractual effort.

Adverse Selection

Basic Adverse Selection Model

Adverse selection is the term used to refer to problems of asymmetric information that appear before the contract is signed. The classic example of Akerlof [3] illustrates very well the issue: the buyer of a used car has much less information about the state of the vehicle than the seller. Similarly, the buyer of a product knows how much he appreciates the quality, while the seller only has statistical information about a typical buyer's taste [68]; or the regulated firm has more accurate information about the marginal cost of production than the regulator.

A convenient way to model adverse selection problems is to consider that the agent can be of different *types*, and that the agent knows his type before any contract is signed while the principal does not know it. In the previous examples, the agent's type is the particular quality of the used car, the level of appreciation of quality, or the firm's marginal cost. How can the principal deal with this informational problem? Instead of offering just one contract for every (or several) types of agents, she can pro-

pose several contracts so that each type of agent chooses the one that is best for him. A useful result in this literature is the *revelation principle* [31,32,69] that states that any mechanism that the principal can design is equivalent to a *direct revelation mechanism* by which the agent is asked to reveal his type and a contract is offered according to his declaration. That is, a direct revelation mechanism offers a *menu of contracts* to the agent (one contract for each possible type), and the agent can choose any of the proposed contracts. Clearly, the mechanism must give the agent the right incentives to choose the appropriate contract, that is, it must be a *self-selection mechanism*. Menus of contracts are not unusual. For instance, insurance companies offer several possible insurance contracts between which clients may freely choose their most preferred. For example, car insurance contracts can be with or without deductible clauses. The second goes to more risk averse or more frequent drivers while deductibles attract less risk averse or less frequent drivers.

Therefore, the timing of an adverse selection game is the following. In the first place, the agent's characteristics (his "type") are realized, and only the agent learns them. Then, the principal decides the menu of contracts to offer to the agent. Having received the proposal, the agent decides which one of the contracts (if any) to accept. Finally, if one contract has been accepted, the agent chooses the predetermined effort and receives the corresponding payment.

A simple model of adverse selection is the following. A risk-neutral principal contracts an agent (who could be risk-neutral or risk-averse) to carry out some verifiable effort on her behalf. Effort e provides an expected payment to the principal of $\Pi(e)$, with $\Pi'(e) > 0$ and $\Pi''(e) < 0$. The agent could be either of *two types* that differ with respect to the disutility of effort, which is $v(e)$ for type G(good), and $kv(e)$, with $k > 1$ for type B(bad). Hence, the agent's utility function is either $U^G(w, e) = u(w) - v(e)$ or $U^B(w, e) = u(w) - kv(e)$. The principal considers that the probability for an agent to be type-G is q, where $0 < q < 1$.

The principal designs a menu of contracts $\{(e^G, w^G), (e^B, w^B)\}$, where (e^G, w^G) is directed towards the most efficient type of agent, while (e^B, w^B) is intended for the least efficient type. For the menu of contracts to be a sensible proposal, the agent must be better off by truthfully revealing his type than by deceiving the principal. The principals problem, is therefore to maximize her expected profits subject to the restrictions that (a) after considering the contracts offered, the agent decides to sign with the principal (*participation constraints*), and (b) each agent chooses the contract designed for his particular type (*incentive compatibility constraints*):

$$\begin{aligned}
\underset{[(e^G,w^G),(e^B,w^B)]}{\text{Max}} \quad & q[\Pi(e^G) - w^G] + (1-q)[\Pi(e^B) - w^B] \\
\text{s.t.} \quad & u(w^G) - v(e^G) \geq \underline{U} \\
& u(w^B) - kv(e^B) \geq \underline{U} \\
& u(w^G) - v(e^G) \geq u(w^B) - v(e^B) \\
& u(w^B) - kv(e^B) \geq u(w^G) - kv(e^G)
\end{aligned}$$

The main characteristics of the optimal contract menu $\{(e^G, w^G), (e^B, w^B)\}$ are the following:

(i) The contract offered to the good agent (e^G, w^G) is efficient ('*non distortion at the top*'). The optimal salary w^G however is higher than under symmetric information: this type of agent receives an *informational rent*. That is, the most efficient agent profits from his private information and in order to reveal this information he has to receive a utility greater than his reservation level.

(ii) The participation condition binds for the agent when he has the highest costs (he just receives his reservation utility). Moreover, a distortion is introduced into the efficiency condition for this type of agent. By distorting, the principal loses efficiency with respect to type-B agents, but she pays less informational rent to the G-types.

Principals Competing for Agents in Adverse Selection Frameworks

Starting with the pioneer work by Rothschild and Stiglitz [83] on insurance markets, there have been many studies on markets with adverse selection problems where there is competition among principals to attract agents. We move from a model where one principal maximizes her profits subject to the above constraints, to a game theory environment where each principal has to take into account the actions by others when deciding which contract to offer. In this case, the adverse selection problem may be so severe that we may find ourselves in situations in which no equilibrium exists.

To highlight the main results in this type of models, consider a simple case in which there are two possible risk-averse agent types: good (G) and bad (B) with G being more productive than B. In particular, we assume that G is more careful than B, in the sense that he commits fewer errors. When the agent exerts effort, the result could be either a success (S) or a failure (F). The probability that it is successful is p^{G} when the agent is type-G and p^{B} when he

is type B, where $p^G > p^B$. The principal values a successful result more than a failure. The result is observable, so that the principal can pay the agent according to the result, if she so desires.

There are several risk-neutral principals. Therefore, we look for the set of equilibrium contracts in the game played by principals competing to attract agents. Equilibrium contracts must satisfy that there does not exist a principal who can offer a different contract that would be preferred by all or some of the agents and that gives that principal greater expected profits. This is why, if information was symmetric, the equilibrium contracts would be characterized by the following properties: (i) principals' expected profits are zero; and (ii) each contract must be efficient. Hence the agent receives a fixed contract insuring him against random events. In particular, the equilibrium salary that the agent receives under symmetric information is higher when he is of type G than when he is of type B.

When the principals cannot observe the type of the agent, the previous contracts cannot be longer an equilibrium: all the agents would claim to be a good type. An equilibrium contract pair $\{C^G, C^B\}$ must satisfy the condition that no principal can add a new contract that would give positive expected profits to the agents that prefer this new contract to C^G and C^B. If the equilibrium contracts for the two agent types turn out to be the same, that is, there is only one contract that is accepted by both agent types, then the equilibrium is said to be *pooling*. On the other hand, when there is a different equilibrium contract for each agent type, then we have a *separating* equilibrium. In fact, pooling equilibria never exist, since pooling contracts always give room for a principal to propose a profitable contract that would only be accepted by the *G*-types (the best agents). If an equilibrium does exist, it must be such that each type of agent is offered a different contract.

If the probability that the agent is "good" is large enough, then a separating equilibrium does not exist either. That is, an adverse selection problem in a market may provoke the absence of any equilibrium in that market. When, a separating equilibria does exist. In, the results are similar to the ones under moral hazard in spite of the differences in the type of asymmetric information and in the method of solving. That is, contingent pay-offs are offered to the best agent to allow the principal to separate them from the less efficient ones. In this equilibrium, the least efficient agents obtain the same expected utility (and even sign the same contract) as under symmetric information, while the best agents lose expected utility due to the asymmetric information.

Extensions of Adverse Selection Models

Repeated Adverse Selection In this extension, we consider whether the repetition of the relationship during several periods helps the principal and how it influences the form of the optimal contract. Note first that if the agent's private information is different in each period and the information is not correlated among periods, then any current information revealed does not affect the future and hence the repeated problem is equivalent to simple repetition of the initial relationship. The optimal intertemporal contract will be the sequence of optimal single-period contracts.

Consider the opposite situation where the agent's type is constant over time. If the agent decides to reveal his type truthfully in the first period, then the principal is able to design efficient contracts that extract all surpluses from the agent. Hence, the agent will have very strong incentives to misrepresent his information in the early periods of the relationships. In fact, Baron and Besanko [8] show that if the principal can commit herself with a contract that covers all the periods, then the optimal contract is the repetition of the optimal static contract. This implies that the contract is not *sequentially rational*, and it is also non *robust to* renegotiation: once the first period is finished, the principal "knows" the agent's type and a better contract for both parties is possible.

It is often the case that the principal cannot commit not to renegotiate a long-term contract. Laffont and Tirole [52] show that, in this case, it may be impossible to propose perfect revelation (separating) contracts in the first periods. This is known as the *ratchet effect*. Also, Freixas, Guesnerie, and Tirole [29], and Laffont and Tirole [51] have proven that, even when separating contracts exist, they may be so costly that they are often non optimal and we should expect that information be revealed progressively over time. Baron and Besanko [9] and Laffont and Tirole [53] also introduce frameworks in which it is possible to propose perfect revelation contracts but they are not optimal.

Relationships with Several Agents: Auctions One particularly interesting case of relationship among one principal and several agents is that of a seller who intends to sell one or several items to several interested buyers, where buyers have private information about their valuation for the item(s). A very popular selling mechanism in such a case is an *auction*. As Klemperer [48] would put it, auction theory is one of economics' success stories in both practical and theoretical terms. Art galleries generally use *English auctions:* the agents bid "upwards";

while fish markets are generally examples of *Dutch auctions*: the seller reduces the price of the good until someone stops the auction by buying. Public contracts are generally awarded through (*first price* or *second price*) *sealed-bid auctions* where buyers introduce their bid in a closed envelope, the good is sold to the highest bidder and the prize is either the winner's own bid or the second highest bid.

Vickrey [92,93] was the first to establish the key result in auction theory, the *Revenue Equivalence Theorem* which, subject to some reasonable conditions, says that the seller can expect equal profits on average from all the above (and many other) types of auctions, and that buyers are also indifferent among them all. Auctions are efficient, since the buyer who ends up with the object is the one with the highest valuation. Hence, the existence of private information does not generate any distortions with respect to who ends up getting the good, but the revenue of the seller is less than under symmetric information.

Myerson [70] solves the general mechanism design problem of a seller who wants to maximize her expected revenue, when the bidders have independent types and all agents are risk-neutral. In general, the optimal auction is more complex than the traditional English (second price) or Dutch (first price) auction. His work has been extended by many other authors. When the buyers' types are affiliated (i. e., they are not negatively correlated in any subset of their domain), Milgrom and Weber [64] show that the revenue equivalence theorem breaks down. In fact, in this situation, McAfee, McMillan, and Reny [61] show that the seller may extract the entire surplus from the bidders as if there was no asymmetric information. Starting with Maskin and Riley [60], several authors have also analyzed auctions of multiple units.

Finally, Clarke [19] and Groves [34] initiated another group of models in which the principal contracts with several agents simultaneously, but does not attempt to maximize her own profits. This is the case of the provision of a public good through a mechanism provided by a benevolent regulator.

Relationships with Several Agents: Other Models and Organizational Design Adverse selection models have attempted to analyze the optimal task assignment, the advantages of delegation, or the optimal structure of contractual relationships, when the principal contracts with several agents. Riordan and Sappington [79] analyze a situation where two tasks have to be fulfilled and show that if the person in charge of each task has private information about the costs associated with the task, then the assignment of tasks within the organization is an important decision. For example, when the costs are positively correlated, then the principal will prefer to take charge of one of the phases herself while she will prefer to delegate the task when the costs are negatively correlated.

In a very general framework, Myerson [71] shows a powerful result: in adverse selection situations, centralization cannot be worse than decentralization, since it is always possible to replicate a decentralized contract with a centralized one. This result is really a generalization of the revelation principle. Baron and Besanko [11] and Melumad, Mookherjee, and Reichelstein [62] show that, if the principal can offer complex contracts in a decentralized organization, then a decentralized structure can replicate a centralized organization. When there are problems of communication between principal and agents, the equivalence result does not hold: Melumad and Reichelstein [63] show that *delegation* of authority can be preferable if communication between the principal and the agents is difficult. Still concerning the optimal design of the organization, Dana [22] analyzes the optimal hierarchical structure in industries with several productive phases, when firms have private information related to their costs. They show that structures that concentrate all tasks to a single agent are superior since, the incentives to dishonestly reveal the costs of each of the phases are weaker. Da-Rocha-Alvarez and De-Frutos [20] argue that the absolute advantage of the centralized hierarchy is not maintained if the differences in costs between the different phases are sufficiently important.

Several Principals Stole [89] and Martimort [59] point out the difficulty of extending the revelation principle to situations where an agent with private information is contracted by several principals who act separately. Given that not only one contract (or menu of contracts) is offered to the agent, but several contracts coming from different principals, it is not longer necessarily true that the best a principal can do is to offer a "truth-telling mechanism".

Consider a situation with two principals that are hiring a single agent. If we accept that agent's messages are restricted to the set of possible types that the agent may have, we can obtain some conclusions. If the activities or efforts that the agent carries out for the two principals are substitutes (for instance, a firm producing for two different customers), then the usual result on the distortion of the decision holds: the most efficient type of agent supplies the efficient level of effort while the effort demanded from the least efficient type is distorted. However, due to the lack of cooperation between principals, the distortion induced in the effort demanded from the less efficient type of agent is lower than the one maximizing the principals'

aggregate profits. On the other hand, if the activities that the agent carries out for the principals are complementary (for example, the firm produces a final good that requires two complementary intermediate goods in the production process), then the comparison of the results under cooperation and under no cooperation between the principals reveals that: if a principal reduces the effort demanded from the agent, in the second case, this would imply that it is also profitable for the other principal to do the same. Therefore, the distortion in decisions is greater to that produced in the case in which principals cooperate.

Models of Moral Hazard and Adverse Selection The analysis of principal-agent models where there are simultaneously elements of moral hazard and adverse selection is a complex extension of classic agency theory. Conclusions can be obtained only in particular scenarios. One common class of models considers situations where the principal cannot distinguish the part corresponding to effort from the part corresponding to the agent's efficiency characteristic because both variables determine the production level. Picard [74] and Guesnerie, Picard, and Rey [37] propose a model with risk-neutral participants and show that, if the effort demanded from the different agents is not decreasing in the characteristic (if a higher value of this parameter implies greater efficiency), then the optimal contract is a menu of distortionary deductibles designed to separate the agents. The menu of contracts includes one where the principal sells the firm to the agent (aiming at the most efficient type), and another contract where she sells only a part of the production at a lower prize (aiming at the least efficient type). However, there are also cases where fines are needed to induce the agents to honestly reveal their characteristic.

In fact, the main message of the previous literature is that the optimal solution for problems that mix adverse selection and moral hazard does not imply efficiency losses with respect to the pure adverse selection solution when the agent's effort is observable. However, in other frameworks (see [50]), a true problem of asymmetric information appears only when both problems are mixed when, and efficiency losses are evident. Therefore, the same solution as when only the agent's characteristic is private information cannot be achieved.

Future Directions

Empirical Studies of Principal-Agent Models

The growing interest on empirical issues related to asymmetric information started in the mid nineties (see the survey by Chiappori and Salanie [18]). A very large part of the literature is devoted to test the predictions of the canonical models of moral hazard and adverse selection, where there is only one dimension in which information is asymmetric. A great deal of effort is devoted to try to ascertain whether it is moral hazard, or adverse selection, or both prevalent in the market. This is a difficult task because both adverse selection and moral hazard generate the same predictions in a cross section. For instance, a positive correlation between insurance coverage and probability of accident can be due to either the intrinsically riskier drivers selecting into contracts with better coverage (as the [83] model of adverse selection will predict) or to drivers with better coverage exerting less effort to drive carefully (as the canonical moral hazard model will predict). Chiappori, Jullien, Salanié and Salanié [17] have shown that the positive correlation between coverage and risk holds more generally that in the canonical models as long as the competitive assumption is maintained.

Future empirical approaches are likely to incorporate market power (as in [17]), multiple dimensions of asymmetric information (as in [28]), as well as different measures of asymmetric information (as in [91]). These advances will be partly possible thanks to richer surveys which collect subjective information regarding agents' attributes usually unobserved by principals or agent's subjective probability distributions. The wider availability of panel data will mean that it will become easier to disentangle moral hazard from adverse selection (as in [1]). Much is to be learnt by using field experiments that allow randomly varying contract characteristics offered to individuals and hence disentangling moral hazard from adverse selection (as in [47]).

Contracts and Social Preferences

Although principal-agent theory has proved fundamental in expanding our understanding of contract situations, real-life contracts frequently do not exactly match its predictions. Many contracts are linear and simpler, incentives are often stronger and wage gaps more compressed than expected. One possible explanation is that the theory has mainly focused on economic agents exclusively motivated by their own monetary incentives. However, this assumption leaves aside issues such as social ties, team spirit or work morale, which the human resources literature highlights. A recent strand of economic literature, known as "behavioral contract theory", has tried to incorporate social aspects into the economic modeling of contracts.

Such theory has been motivated by two types of empirical support. On the one hand, extensive interview studies with firm managers and employees [13] has shown not

only that agents care about social comparisons such as internal pay equity or effort diversity, but that their incentives to work hard are affected by them and that principals are aware of it and design their contracts accordingly. On the other hand, one of the most influential contributions of the experimental literature has been to show that, assuming that economic agents are not completely selfish (but exhibit some form of social preferences), helps organizing many laboratory data. Experiments replicating labor markets (starting with Fehr's [26]) confirm Akerlof's [4] insight that contracts may be understood as a form of gift exchange in which principals may offer a "generous" wage and agents may respond with more than the minimum effort required.

Incorporating social and psychological aspects in a systematic manner into agents' motivations has given rise to several forms of utility functions reflecting inequality aversion [14,24,27], fairness [75] and reciprocity [25]. More recently, such utility functions have been included into standard contract theory models and have helped in shortening the gap between theory predictions and real-life contracts. In particular, issues such as employees' feelings of envy or guilt towards their bosses [45], utility comparisons among employees [35,78] or peer-pressure motivating effort decisions [43,55] have proved important in widening the scope of issues principal-agent theory can help to understand.

Principal-Agent Markets

The literature has been treating each principal-agent relation as an isolated entity. Thus, it normally takes a given relationship between a principal and an agent (or among several principals and/or several agents), and analyzes the optimal contract. In particular, the principal assumes all the bargaining power as she has the right to offer the contract she likes the most, and agent's payoff is determined by his exogenously given reservation utility. However, in markets there is typically not a single partnership but there are several. It is then interesting to consider the simultaneous determination of both the identity of the pairs that meet (i. e., the *matching* between principals and agents) and the contracts these partnerships sign. The payoffs to each principal and agent will then depend on the other principal-agent relationships being formed in the market. This analysis requires a general equilibrium-like model.

Game theory provides a very useful tool to deal with the study of markets where heterogeneous players from one side can form partnerships with heterogeneous players from the other side: the two-sided matching models. Examples of classic situations studied in two-sided matching models (see [82,86]) are the marriage market, the college admissions model, or the assignment market (where buyers and sellers transact). Several papers extend this game theory models to situations where each partnership involves contracts and show that the simultaneous consideration of matching and contracts has important implications. Dam and Pérez-Castrillo [21] show that, in an economy where landowners contract with tenants, a government willing to improve the situation of the tenants can be interested in creating wealth asymmetries among them. Otherwise, the landowners would appropriate all the incremental money that the government is willing to provide to the agents. Serfes [85] shows that higher-risk projects do not necessarily lead to lower incentives, which is the prediction in the standard principal-agent theory, and Alonso-Paulí and Pérez-Castrillo [6] apply the theory to markets where contracts (between shareholders and managers) can include Codes of Best Practice. On the empirical side, Ackerberg and Botticini [2] find strong evidence for endogenous matching between landlords and tenants and that risk sharing is an important determinant of contract choice.

Future research will extend the general equilibrium analysis of principal-agent contracts to other markets. In addition, the literature has only studied one-to-one matching models. This should be extended to situations where each principal can hire several agents, or where each agent deals with several principals. The interplay between (external) market competition and (internal) collaboration between agents or principals can provide useful insights about the characteristics of optimal contracts in complex environments.

Bibliography

Primary Literature

1. Abbring J, Chiappori PA, Heckman JJ, Pinquet J (2003) Adverse Selection and Moral Hazard in Insurance: Can Dynamic Data Help to Distinguish? J European Econ Ass 1:512–521
2. Ackerberg DA, Botticini M (2002) Endogenous Matching and the Empirical Determinants of Contract Form. J Political Econ 110:564–592
3. Akerlof G (1970) The Market for 'Lemons': Qualitative Uncertainty and the Market Mechanism. Quarterly J Econ 89:488–500
4. Akerlof G (1982) Labor Contracts as a Partial Gift exchange. Quarterly J Econ 97:543–569
5. Alchian A, Demsetz H (1972) Production, Information Costs, and Economic Organization. Am Econ Rev 62:777–795
6. Alonso-Paulí E, Pérez-Castrillo D (2007) Codes of Best Practice in Competitive Markets. Mimeo
7. Arrow K (1985) The Economics of Agency. In: J Pratt, R Zeckhauser (eds) Principals and agents: The Structure of business. Harvard University Press, Boston

8. Baron D, Besanko D (1984) Regulation and Information in a Continuing Relationship. Information Econ Policy 1:267–302
9. Baron D, Besanko D (1987) Commitment and Fairness in a Dynamic Regulatory Relationship. Rev Econ Studies 54:413–436
10. Baron D, Myerson R (1982) Regulating a Monopoly with Unknown Costs. Econometrica 50:911–930
11. Baron D, Besanko D (1992) Information, Control and Organizational Structure. J Econ Management Strategy 1(2):237–275
12. Bernheim BD, Whinston MD (1986) Common Agency. Econometrica 54:923–942
13. Bewley T (1999) Why Rewards Don't Fall During a Recession. Harvard University Press, Cambridge
14. Bolton G, Ockenfels A (2000) ERC: A Theory of Equity, Reciprocity and Competition. Am Econ Rev 90:166–193
15. Chiappori PA, Macho-Stadler I (1990) Contrats de Travail Répétés: Le Rôle de la Mémoire. Annales Econ Stat 17:4770
16. Chiappori PA, Macho-Stadler I, Rey P, Salanié B (1994) Repeated Moral Hazard: The Role of Memory, Commitment and Access to Credit Markets. Eur Econ Rev 38:1527–1553
17. Chiappori PA, Jullien B, Salanié B, Salanié F (2006) Asymmetric Information In Insurance: General Testable Implications. Rand J Econ 37:783–798
18. Chiappori PA, Salanié B (2003) Testing Contract Theory: a Survey of Some Recent Work. In: Dewatripont, Hansen, Turnovsky (eds) Advances in Economics and Econometrics, vol. 1. Cambridge University Press, Cambridge, pp 115–149
19. Clarke E (1971) Multipart Pricing of Public Goods. Public Choice 11:17–33
20. Da-Rocha-Alvarez JM, De-Frutos MA (1999) A Note on the Optimal Structure of Production. J Econ Theory 89:234–246
21. Dam K, Pérez-Castrillo D (2006) The Principal-Agent Matching Market. Frontiers Econ Theory. Berkeley Electro 2(1):1–34
22. Dana JD (1993) The Organization and Scope of Agents: Regulating Multiproduct Industries. J Econ Theory 59:288–310
23. Demski JS, Sappington D (1986) Line-item Reporting, Factor Acquisition and Subcontracting. J Accounting Research 24:250–269
24. Desiraju R, Sappington D (2007) Equity and Adverse Selection. J Econ Management Strategy 16:285–318
25. Dufwenberg M, Kirchsteiger G (2004) A Theory of Sequential Reciprocity. Games Econ Behavior 47:268–298
26. Fehr E, Kirchsteiger G, Riedl A (1993) Does Fairness Prevent Market Clearing? Quart J Econ 108:437–460
27. Fehr E, Schmidt K (1999) A Theory of Fairness, Competition and Cooperation. Quart J Econ 114:817–868
28. Finkelstein A, McGarry K (2006) Multiple Dimensions of Private Information: Evidence from the Long-term Care Insurance Market. Am Econ Rev 96:938–958
29. Freixas X, Guesnerie R, Tirole J (1985) Planning under Information and the Ratchet Effect. Rev Econ Studies 52:173–192
30. Fudenberg D, Holmström B, Milgrom B (1990) Short-term Contracts and Long-term Agency Relationships. J Econ Theory 51:1–31
31. Gibbard A (1973) Manipulation for Voting Schemes. Econometrica 41:587–601
32. Green JR, Laffont JJ (1977) Characterization of Satisfactory Mechanisms for the Revelation of Preferences for Public Goods. Econometrica 45:427–438
33. Grossman SJ, Hart OD (1983) An Analysis of the Principal-Agent Problem. Econometrica 51:7–45
34. Groves T (1973) Incentives in Teams. Econometrica 41:617–631
35. Grund C, Sliwka D (2005) Envy and Compassion in Tournaments. J Econ Manag Strateg 14:187–207
36. Guesnerie R, Laffont JJ (1984) A Complete Solution to a Class of Principal-Agent Problems with Application to the Control of a Self-Managed Firm. J Public Econ 25:329–369
37. Guesnerie R, Picard P, Rey P (1989) Adverse Selection and Moral Hazard with Risk Neutral Agents. Eur Econ Rev 33:807–823
38. Harris M, Raviv A (1978) Some Results on Incentive Contracts with Applications to Education and Employment, Health Insurance and Law Enforcement. Am Econ Rev 68:20–30
39. Harris M, Raviv A (1979) Optimal Incentive Contracts with Imperfect Information. J Econ Theory 2:231–259
40. Holmström B (1979) Moral Hazard and Observability. Bell J Econ 10:74–91
41. Holmström B (1982) Moral Hazard in Teams. Bell J Econ 13:324–340
42. Holmström B, Milgrom P (1991) Multitask Principal-Agent Analysis: Incentive Contracts, Assets Ownership, and Job Design. J Law Econ Organization 7(Suppl):24–52
43. Huck S, Rey-Biel P (2006) Endogenous Leadership in Teams. J Inst Theoretical Econ 162:1–9
44. Itoh H (1990) Incentives to Help in Multi-Agent Situations. Econometrica 59:611–636
45. Itoh H (2004) Moral Hazard and Other-Regarding Preferences. Jpn Econ Rev 55:18–45
46. Jensen M, Meckling W (1976) The Theory of the Firm, Managerial Behavior, Agency Costs and Ownership Structure. J Finan Econ 3:305–360
47. Karlan D, Zinman J (2006) Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment. Mimeo
48. Klemperer P (2004) Auctions: Theory and Practice. Princeton University Press, Princeton
49. Laffont JJ (1990) Analysis of Hidden Gaming in a Three Levels Hierarchy. J Law Econ Organ 6:301–324
50. Laffont JJ, Tirole J (1986) Using Cost Observation to Regulate Firms. J Polit Econ 94:614–641
51. Laffont JJ, Tirole J (1987) Comparative Statics of the Optimal Dynamic Incentive Contract. Eur Econ Rev 31:901–926
52. Laffont JJ, Tirole J (1988) The Dynamics of Incentive Contracts. Econometrica 56:1153–1176
53. Laffont JJ, Tirole J (1990) Adverse Selection and Renegotiation in Procurement. Rev Econ Studies 57:597–626
54. Lambert R (1983) Long Term Contrats and Moral Hazard. Bell J Econ 14:441–452
55. Lazear E (1995) Personnel Economics. MIT Press, Cambridge
56. Macho-Stadler I, Pérez-Castrillo D (1993) Moral Hazard with Several Agents: The Gains from Cooperation. Int J Ind Organ 11:73–100
57. Macho-Stadler I, Pérez-Castrillo D (1998) Centralized and Decentralized Contracts in a Moral Hazard Environment. J Ind Econ 46:489–510
58. Malcomson JM, Spinnewyn F (1988) The Multiperiod Principal-Agent Problem. Rev Econ Stud 55:391–408
59. Martimort D (1996) Exclusive Dealing, Common Agency and Multiprincipals Incentive Theory. Rand J Econ 27:1–31
60. Maskin E, Riley J (1989) Optimal Multi-Unit Auctions. In: Hahn F (ed) The Economics of Missing Markets, Information, and Games. Oxford University Press, Oxford, pp 312–335

61. McAfee P, McMillan J, Reny P (1989) Extracting the Surplus in the Common Value Auction. Econometrica 5:1451–1459
62. Melumad N, Mookherjee D, Reichestein S (1995) Hierarchical Decentralization of Incentive Contracts. Rand J Econ 26:654–672
63. Melumad N, Reichelstein S (1987) Centralization vs Delegation and the Value of Communication. J Account Res 25:1–18
64. Milgrom P, Weber RJ (1982) A Theory of Auctions and Competitive Bidding. Econometrica 50:1089–1122
65. Mirrlees J (1971) An Exploration in the Theory of Optimum Income Taxation. Rev Econ Studies 38:175–208
66. Mirrlees J (1975) The Theory of Moral Hazard and Unobservable Behavior, Part I. WP Nuffield College, Oxford
67. Mookherjee D (1984) Optimal Incentive Schemes with Many Agents. Rev Econ Studies 51:433–446
68. Mussa M, Rosen S (1978) Monopoly and Product Quality. J Econ Theory 18:301–317
69. Myerson R (1979) Incentive Compatibility and the Bargaining Problem. Econometrica 47:61–73
70. Myerson R (1981) Optimal Auction Design. Math Op Res 6:58–73
71. Myerson R (1982) Optimal Coordination Mechanisms in Generalized Principal-Agent Models. J Math Econ 10:67–81
72. Pauly MV (1968) The Economics of Moral Hazard. Am Econ Rev 58:531–537
73. Pauly MV (1974) Overinsurance and Public Provision of Insurance: The Roles of Moral Hazard and Adverse Selection. Quart J Econ 88:44–62
74. Picard P (1987) On the Design of Incentive Schemes under Moral Hazard and Adverse Selection. J Public Econ 33:305–331
75. Rabin M (1993) Incorporating Fairness into Game Theory and Economics. Am Econ Rev 83:1281–1302
76. Radner R (1981) Monitoring Cooperative Agreements in a Repeated Principal-Agent Relationship. Econometrica 49:1127–1148
77. Rey P, Salanié B (1990) Long term, Short term and Renegotiation. Econometrica 58:597–619
78. Rey-Biel P (2008) Inequity Aversion and Team Incentives. ELSE WP, Scandinavian J Econ 110:297–320
79. Riordan MH, Sappington DE (1987) Information, Incentives, and the Organizational Mode. Quart J Econ 102:243–263
80. Rogerson W (1985) Repeated Moral Hazard. Econometrica 53:69–76
81. Ross SA (1973) The Economic Theory of Agency: The Principal's Problem. Am Econ Rev 63:134–139
82. Roth AE, Sotomayor M (1990) Two-sided matching: A study in game-theoretic modeling and analysis. Cambridge University Press, New York
83. Rothschild M, Stiglitz J (1976) Equilibrium in Competitive Insurance Markets: An Essay in the Economics of Imperfect Information. Quart J Econ 90:629–650
84. Rubinstein A, Yaari ME (1983) Repeated Insurance Contrats and Moral Hazard. J Econ Theory 30:74–97
85. Serfes K (2008) Endogenous Matching in a Market with Heterogeneous Principals and Agents. Int J Game Theory 36:587–619
86. Shapley LS, Shubik M (1972) The Assignment Game I: The Core. Int J Game Theory 1:111–130
87. Spence M (1974) Market Signaling. Harvard University Press, Cambridge
88. Spence M, Zeckhauser R (1971) Insurance Information, and Individual Action. Am Econ Rev 61:380–387
89. Stole L (1991) Mechanism Design under Common Agency. WP MIT, Cambridge
90. Tirole J (1986) Hierarchies and Bureaucracies: On the Role of Collusion in Organizations. J Law Econ Organ 2:181–214
91. Vera-Hernandez M (2003) Structural Estimation of a Principal-Agent Model: Moral Hazard in Medical Insurance. Rand J Econ 34:670–693
92. Vickrey W (1961) Counterspeculation, Auctions and Competitive Sealed Tenders. J Finance 16:8–37
93. Vickrey W (1962) Auction and Bidding Games. In: Recent Advances in Game Theory. The Princeton University Conference, Princeton. pp 15–27

Books and Reviews

Hart O, Holmström B (1987) The Theory of Contracts. In: Bewley T (ed) Advances in Economic Theory, Fifth World Congress. Cambridge University Press, Cambridge

Hirshleifer J, Riley JG (1992) The Analytics of Uncertainty and Information. Cambridge University Press, Cambridge

Laffont JJ, Martimort D (2002) The Theory of Incentives: The Principal-Agent Model. Princeton University Press, Princeton

Laffont JJ, Tirole J (1993) A Theory of Incentives in Procurement and Regulation. MIT Press, Cambridge

Macho-Stadler I, Pérez-Castrillo D (1997) An Introduction to the Economics of Information: Incentives and Contracts. Oxford University Press, Oxford

Milgrom P, Roberts J (1992) Economics, Organization and Management. Prentice-Hall, Englewood Cliffs

Probability Densities in Complex Systems, Measuring

Gunnar Pruessner
Department of Mathematics, Imperial College London, London, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Simulation Techniques
Scaling
Histogram Data Representation
Future Directions
Bibliography

Glossary

Apparent exponent The apparent exponent is the power-law exponent of the probability density function $P(s; s_c)$ in the scaling region. The scaling region is the

range of event sizes s which is closest to a straight line in a double-logarithmic plot of $P(s; s_c)$, the intermediate range $s_c \gg s \gg s_0$.

Consistent estimator An estimator is consistent if it converges to the quantity estimated (the expectation value of the population) as the sample size is increased. For example, the lack of independence of consecutive measurements generated in a numerical simulation can render an estimator inconsistent.

Corrections to scaling In general, pure power-law behavior is found in an observable only to leading order, for example $\langle s^2 \rangle = aL^\alpha + bL^\beta + \ldots$ with $\alpha > \beta$. While the sub-leading terms can have great physical relevance more emphasis is normally given to the leading order. The quality of the data analysis when determining the leading order can improve significantly by allowing for correction terms. The exponents found in these correction terms are usually expected to be universal as well.

Correlation time Fitting the autocorrelation function of an observable to an exponential $\exp(-t/\tau)$ produces the correlation time τ. Although correlations are in general more complicated than the single exponential suggests, the standard deviation of the estimator of the nth moment from N measurements is often estimated to be $\sqrt{(2\tau+1)/N}$ times the estimated standard deviation of the nth moment,

$$\overline{\sigma}^2\left(\overline{s^n}\right) = \frac{2\tau+1}{N}\overline{\sigma}^2\left(s^n\right) , \tag{1}$$

as if the number of independent measurements was only $N/(2\tau+1)$.

Estimator A numerical estimator is any function that provides an estimate from the sample, that is the set of all measurements taken. A good estimator is unbiased, consistent and efficient. Very often, such an estimator coincides with the definition of the observable as taken from the exact distribution, for example $\overline{s^2} = \frac{1}{N}\sum_i^N s_i^2$ for estimating the second moment from an uncorrelated sample $s_1, s_2, \ldots, s_N$, with exact value $\langle s^2 \rangle = \int \mathrm{d}s s^2 P(s; s_c)$. However, generally, a function of observables to be estimated, is not well estimated by taking the function of the estimates. For example, the square of the first moment $\langle s \rangle^2$ is not well estimated by the numerical estimate $\overline{s}^2 = (\sum s_i/N)^2$, as this estimator would be biased.

Finite size scaling Observables in complex systems that display a power-law dependence on a parameter, often diverge in the thermodynamic limit. In finite systems they remain finite and their value is expected to diverge as a power-law of the system size. The relation between the value of the observable and the system size is known as "finite size scaling" (abbreviated FSS).

Gap exponent Given that a system displays scaling, the exponents γ'_n characterizing the dependence of the nth moment on a parameter, such as the system size in case of finite size scaling, often are linear in n. The gap exponent is the gap between consecutive moments, $\gamma'_{n+1} - \gamma'_n$.

Importance sampling Importance sampling is a numerical technique to bias the frequency which with configurations are generated, so that states of greater importance, e. g. large observables, are generated more often than others, less important ones. Using a Markov chain to generate the states of an Ising model as opposed to generating them at random and applying a Boltzmann–Gibbs weight can be regarded as a form of importance sampling.

Lower cutoff The lower cutoff in a probability density function of event sizes in complex systems is a value of the event size above which the distribution displays universal behavior. Below this value the system is governed by microscopic details. For many systems, the probability density functions for different system sizes coincide for event sizes below the lower cutoff.

Markov chain Monte Carlo A Monte Carlo technique whereby configurations are generated by transforming the state of the system according to a transition probability. The stationary probability distribution of the different states corresponds to the target probability distribution, i. e. the distribution to be modeled. In complex systems, Markov Chain Monte Carlo (abbreviated MCMC) is the natural method to study a model: Configurations of the system are generated with the frequency they occur in the exact distribution.

Moment analysis In general it is a difficult to identify and quantify scaling behavior in probability density functions. The most general analysis is a data collapse, the quality of which is not easily determined. The most widely used method to determine scaling exponents and moment ratios of the universal scaling function therefore is a moment analysis. Based on the scaling assumption of the probability density function, moments scale as a power of the upper cutoff with amplitudes certain ratios of which are universal.

Monte Carlo Technique to calculate numerical estimates for expectation values in stochastic models by generating configurations at random. More generally, Monte Carlo (abbreviated MC) is a stochastic technique to numerically integrate a high-dimensional in-

tegral, here corresponding to the calculation of an expectation value by integrating over all degrees of freedom, constituting the phase space of the system.

Parameter space and phase space The parameter space of a complex system is the space spanned by all parameters of the model, such as system size and couplings of the interacting agents. A numerical study usually aims to probe the model throughout a large part of the parameter space. The number of parameters therefore needs to be as small as possible. Often only a single parameter exists. Leaving the parameters fixed, an individual numerical simulation samples the phase space available to the system. The phase space is the set of all possible configurations or states of the model. This space is very high dimensional and it is virtually impossible to sample this space homogeneously. A numerical simulation relies on the assumption that the sample taken nevertheless is sufficiently representative to allow for reliable estimates.

Stationary distribution and transient Most complex systems studied possess a limiting distribution, i. e. the probability density distribution in phase space converges. This is the stationary distribution. A free random walker, for example, does not possess a stationary distribution, while a random walker in a harmonic potential does. Due to correlations, the probability distribution of states after any finite time generally depends on the initial condition, which is therefore often chosen to be random. The measurements discarded due to these correlation are called the transient.

Unbiased estimator An estimator is unbiased if the population average of the estimator is independent of the sample size. For example, $\bar{s} = \sum_i^N s_i/N$ from an uncorrelated sample $s_1, \ldots, s_N$ is an unbiased estimator of the expected first moment. Estimating the variance of s as $\overline{\sigma}^2(s) = \overline{s^2} - \bar{s}^2$ however is biased, because the population mean of $\overline{s^2} - \bar{s}^2$ is $((N-1)/N)(\langle s^2\rangle - \langle s\rangle^2)$, which depends on the sample size N.

Upper cutoff The upper cutoff is the characteristic scale of the universal part of the event size distribution. It is a measure of the event size at which the scaling function of the event size distribution breaks up. Moments $\langle s^n\rangle$ with sufficiently large n are to leading order a power of the upper cutoff. The upper cutoff itself is expected to be a power law of the system parameter, i. e. the system size in case of finite size scaling. The exponent controlling the relation between upper cutoff and system size is the gap exponent.

Definition of the Subject

Observables in complex systems usually obtain a broad range of values. They are random variables of a stochastic process and the system explores a wide phase space. The observables are therefore characterized by a probability density function (PDF), representing the probability (density) to find the complex system in a state with a particular value of the observable. As in other areas of statistical mechanics, complex systems are often studied in computer simulations, most prominently **Monte Carlo** [1,2,3] and the probability density function is recorded in form of a histogram, which frequently has power-law asymptotes. Historically, the probability density function itself plays a dominant rôle in the characterization of complex system, while more recently derived quantities, in particular moments, are used more frequently.

Introduction

There are three main stages in the analysis of the PDF of a complex system: The first step is to generate in a computer simulation estimates of the PDF itself or derived quantities, such as moments. This is a general, technical problem considered in computational physics. In a second step, the expected behavior of the observables is to be derived, a process that often feeds to the stage of data generation, as it might suggest new observables to be estimated. Thirdly, the estimates are to be analyzed and compared to the expected behavior.

The following material highlights main aspects of these steps. In practice, the first and the second step go hand in hand and might even change places, but as the simulation techniques are so clearly distinct from the theoretical expectations and the data analysis, which, in turn, are closely related, simulation techniques are presented first.

The following subsection describes a key-example in complexity which has been analyzed in great detail using the techniques described in this article. The subsequent sections present some principal ideas and techniques from the three areas introduced above.

Bak–Tang–Wiesenfeld Model

The Bak–Tang–Wiesenfeld (BTW) Model [4,5] is the first model of Self Organized Criticality studied because of the peculiar features of the probability density function of its key observable. The BTW model is a cellular automaton, sometimes also called a "stochastic cellular automaton" when it is updated in a random fashion. The BTW model evolves as follows: In two dimension, height variables z_i, which can take integral values set to 0 initially, are assigned

to every site i of a square lattice of size $L \times L$. Whenever a height variable exceeds a certain critical value, its value is decreased by 4 and the variables of the surrounding four nearest neighboring sites are incremented by 1. These rules are modified for sites on the open boundary, where particles are lost. To drive the model, the height variable at either a single, specific site or at a randomly chosen site is increased by 1 and the relaxation rules described above are applied until none of the sites exceeds the critical value, which gives rise to an "avalanche" of size s, measuring the number of times the updating rule has been applied.

The histogram $\overline{P(s)}$, calculated by repeating the above procedure millions of times, measures the frequency with which a certain avalanche size has been observed. It represents a numerical estimate of the exact, discrete probability function $\langle P(s) \rangle$. In the following, an over-line as in $\overline{P(s)}$ indicates a numerical estimate, while brackets, such as in $\langle P(s) \rangle$ indicate the exact population average.

Strictly, the distribution $\langle P(s) \rangle$ is usually inaccessible, because even a finite complex system has so many different states and therefore so many different outcomes, some with very small frequencies, that it is virtually impossible to calculate the distribution exactly or sample all outcomes with the exact weight.

As shown in Fig. 1, the PDF of the avalanche size has a very long, power law tail. This is typical for complex systems and distinguishes it from what is often found in simpler problems. Most importantly, the PDF deviates from a simple Gaussian which is, by the central limit theorem [6], the expected PDF, if the observable was a sum or an average of many independent random variables. The lack of convergence to a Gaussian therefore indicates the presence of **correlations**, the hallmark of a complex system. *The aim of a numerical estimate and analysis of the PDF of a complex system is to characterize it quantitatively in the form of exponents, moment ratios and other features, in order to compare it to other complex systems*. To improve the numerical results, various simulation and analysis techniques are employed, many of which are aiming to minimize the amount of CPU-time spent in the simulation.

In the following, some standard simulation techniques are described. These methods have been developed since the early 1950s [2] and represent themselves a branch of

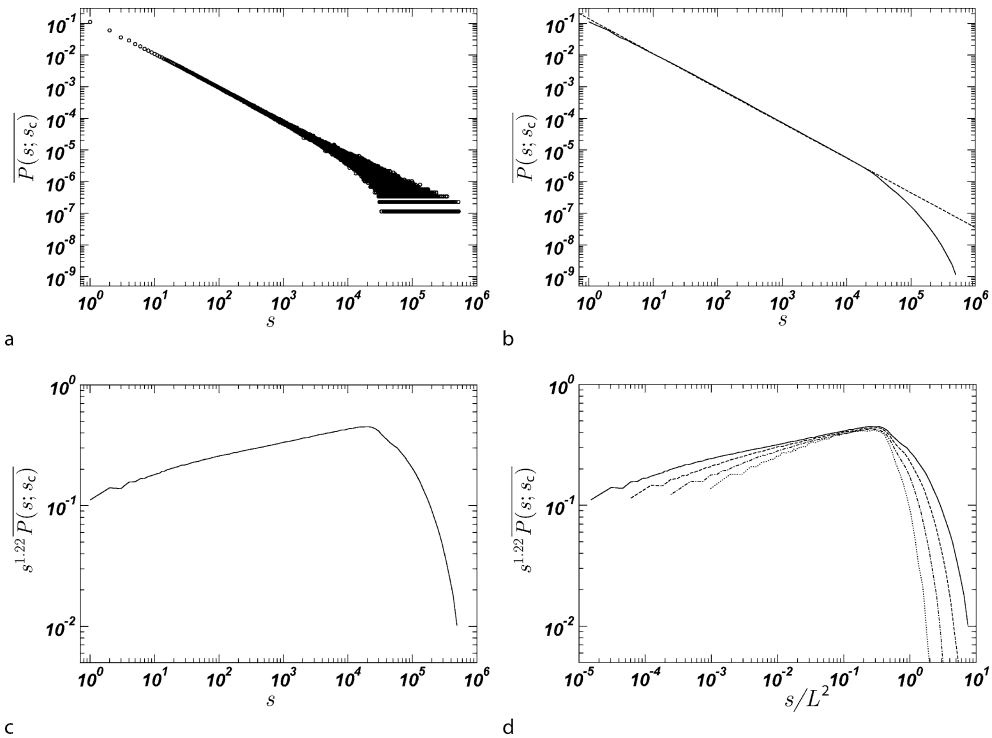

Probability Densities in Complex Systems, Measuring, Figure 1

BTW simulation of a system of size 256 × 256, with $1 \cdot 10^7$ iterations used in the transient and $2 \cdot 10^7$ iterations used for statistics: a Raw histogram, b binned histogram, (*dashed line* shows exponent − 1.1) c rescaled histogram, using $\tau = 1.22$ and d (attempt) of a data collapse including very small system sizes, $L = 32, 64, 128, 256$. The different ways to generate, process, present and analyze the raw data are discussed in this article

numerical analysis. The data analysis is based on the **sample**, i. e. the set of **measurements** produced in numerical simulations, subject of the following section. The analysis can either focus on the PDF itself or on its moments, revealing different features of the system.

Simulation Techniques

In general, the complex system to be studied numerically models a more complicated phenomenon observed in nature, in economy, society etc. Typically, such a model is composed of a large number of interacting elements, which might be divisible in different classes. These elements might be agents, particles, organisms, locations etc. and the classes might represent different rôles or species. Depending on the class, the type of interaction might be very different.

Only in very rare cases the model strictly reflects the original natural, economical or social observation it has been derived from. A complete representation of the original problem, even if possible, might be numerically intractable, because of too large a number of parameters that characterize the model. These parameters generally are **couplings**, i. e. they parametrize the interaction between the individual elements, and generally stay fixed during the simulation. The immediate aim of the simulation is to determine observables as functions of the parameters. Only if the number of parameters is small enough, the **parameter space** can be covered sufficiently densely to allow for a reliable statement of their various rôles and effects.

The parameter space is to be distinguished from the **phase space**, which is the space or set of possible configurations of the system. Here, a **configuration** or **state** is the (smallest) set of values describing the system in such a way that they suffice at any point to prescribe the further evolution of the system. Each of these values is associated with a **degree of freedom** in phase space.

Reducing the degrees of freedom generally has little impact on the quality of the numerical estimates. Even the simplest computer models usually have so many states that it becomes virtually impossible to calculate any property by realizing all such states. For example, a two dimensional square lattice of size 10×10 sites, which each can be in one out of two states, has 2^{100} possible states and it would take thousands of billions of years to visit each state, even if a billion states could be realized every second. Such a comprehensive study of this rather small system therefore is out of scope. Moreover, the original process to be modeled, realized in nature or society, cannot be thought of being appropriately represented in this form.

Monte Carlo is by far the most widely used technique for measuring probability densities in a complex system. Other approaches, in particular deterministic **molecular dynamics**, are far less popular. This section focuses on the practice of Monte Carlo and various standard techniques used in conjunction with it.

Monte Carlo Methods

In its most general form, a Monte Carlo algorithm estimates an expectation value of an observable by averaging over a random sample. In the following, this is illustrated for a problem with discrete configurations σ, each of which describing the state of the entire system. In case of interacting particles at sites, this would be a set of numbers representing the state of every individual site. Assuming the exact, but usually unknown probability of state σ to be $\mathcal{P}(\sigma)$, which in systems with continuous degrees of freedom is replaced by a PDF, the expectation value for an observable $A(\sigma)$ is

$$\langle A \rangle = \sum_{\{\sigma\}} \mathcal{P}(\sigma) A(\sigma) \tag{2}$$

where the sum runs over the set of all states $\{\sigma\}$. The exact expectation value can be approximated by the mean $\overline{A}$ over a (small) **random sample** of N configurations $\sigma_1, \sigma_2, \ldots, \sigma_N$,

$$\overline{A} = \frac{1}{N} \sum_{i=1}^{N} A(\sigma_i) \tag{3}$$

if the configurations σ_i occur with $\mathcal{P}(\sigma)$. By the (weak) law of large numbers, the estimate $\overline{A}$ converges to the exact value $\langle A \rangle$ in the limit of a large **sample size** N.

In most classical models of statistical mechanics, the probability $\mathcal{P}(\sigma)$ is known exactly and often corresponds to the Boltzmann–Gibbs weight. In some rare cases, such as percolation [7], configurations of the system are all equally important and are therefore generated easily at random and independently. For a much greater class of problems, for example in case of the Ising model [8], the configurations differ greatly in statistical weight and it is a hard computational problem to generate them in a sensible way. In these cases, Monte–Carlo sampling is mainly concerned with finding *important* configurations, which is often achieved by constructing a pseudo-dynamical process that generates the configurations with the correct probabilities. In complex systems, on the other hand, usually the dynamics is given while the probability is unknown, so that the first concern is to produce measurements with the correct probability (density). However,

both issues lead to the same solution, namely **Markov Chain Monte Carlo**, which can be regarded as a form of **importance sampling**.

Importance sampling is a method for producing measurements with a frequency optimized for reducing the number of measurements required to estimate an expectation value within a prescribed error. The sampling method therefore is, ideally, adapted to the observable. In the vast majority of numerical simulations of complex systems, however, the sampling scheme used corresponds to the simplest form of a Markov chain, where the configuration of the system evolves from one stage to another in discrete steps given by the dynamics, so that the frequency with which a configuration occurs during the simulation converges to the exact probability of that configuration [9]. The Markov condition means that the probability to observe a certain configuration of the system depends *only on the directly preceding configuration*. By including a sufficient amount of information in the configuration, in principle every process can be rendered Markovian. For example, while the sequence of particle coordinates in a classical gas is not Markovian, the same sequence together with the particles' moments is so.

More sophisticated approaches bias the sampling so that regions in phase space which have a greater contribution $A(\sigma)\mathcal{P}(\sigma)$ to the expectation value than other regions are visited more frequently, which is discussed further in Subsect. "Rare Events".

Monte Carlo Applied to Complex Systems

In the following, it will be assumed that the complex system has its dynamics given in the form of a set of rules specifying how the model evolves from one configuration to another. Furthermore, it will be assumed that these rules involve some randomness, so that a given starting configuration does not necessarily lead to one particular configuration in the next step. The evolution of the system is subject to certain **transition probabilities** encoded in the transition matrix $W(\sigma, \sigma')$, which gives the probability to go from configuration σ' to configuration σ. This leads to the notion of the system's **trajectory in phase space**, which is a random variable, traced out by application of the transition matrix in every **Monte Carlo time step** or **update**. **Markov Chain Monte Carlo** uses the phase space trajectory to sample the phase space in a fair fashion.

By construction, consecutive configurations are generally correlated, i. e. the probability to find the system in a certain configuration depends on the preceding configuration. Given the Markovian nature of the process, the probability depends only on the previous configuration, nevertheless correlations can be very long-lived, because the previous configuration's probability depends in turn on its preceding configuration and so on. The **correlation time** is a measure for the number of updates over which these correlations decay, the numerical consequences of which are discussed in Subsect. "Correlation Time".

At the beginning of the simulation the system is initialized in a particular form which in practice is often a rather artificial state. Due to correlations the influence of the initial condition can last for very many updates. The Markov-chain approach relies on the insight that the sampling frequency after a sufficiently long **transient** converges to the **stationary distribution** $\mathcal{P}(\sigma)$, which solves for all states σ [6]

$$0 = \sum_{\{\sigma'\}} \mathcal{P}(\sigma')W(\sigma, \sigma') - \sum_{\{\sigma'\}} \mathcal{P}(\sigma)W(\sigma', \sigma) \tag{4}$$

where $W(\sigma, \sigma')$ is the transition probability from σ' to σ introduced above. Unless relaxation processes are to be studied, the stationary regime is sampled after a sufficiently transient, which is ignored, i. e. configurations generated during the transient do not enter the numerical estimates. It is general practice to save the configuration of a system regularly at later stages, which might subsequently be used as a starting point for other trajectories in phase space. This method also allows a systematic study of the influence of the transient. Eq. (4) is to be distinguished from **detailed balance**,

$$0 = \mathcal{P}(\sigma')W(\sigma, \sigma') - \mathcal{P}(\sigma)W(\sigma', \sigma) \tag{5}$$

which solves (4) term by term. Detailed balance is the hallmark of equilibrium thermodynamics and has far reaching consequences for the topology of the network spanned by configurations and the possible transitions between them [10].

Poisson Processes Many complex systems consist of a set of concurrent **Poisson processes**, for example particles at sites that decay with one particular rate and interact with another rate. Maintaining a list of all possible processes and treating the fastest as deterministic is a widely adopted approach. If n processes with rate $r_1, \ldots, r_n$ are possible with the fastest having the largest rate e, one process, say i, out of n is picked at random with uniform probability, performed with probability r_i/e [11] and the continuous time incremented by $1/(en)$. Within time T on average enT picks are made, each process is selected eT times and performed $r_i T$ times, which is precisely the average number of times the Poissonian event should occur. This implementation therefore correctly reproduces the averages. On the other hand, an event controlled by a Poisson

process can take place an arbitrary number of times within any finite time-span, which this implementation is incapable of. If correct time-keeping is expected to play an important rôle, as for example in the case of temporal correlations, the time increment itself is to be drawn randomly from a waiting time distribution [6,12].

Rare Events In the presence of rare events, [13] reliably estimating expectation values might become impossible. The problem affect bounded and unbounded observables, although an unbounded observable is potentially more prone to this phenomenon.

The problem is caused by events that occur only very rarely, where, however, the observable is particularly large. If, for example, an event occurs with frequency $f = 10^{-6}$ where the observable A has a value of $A^* = 10^4 a$, with a being the average of A over all other realizations, then including the event will change the estimate of the average of A only by $fA^*/a = 10^{-2}$. Assuming that the second moment over all other events is a^2, including the rare event will shift the estimate by $fA^{*2}/a^2 = 10^2$.

An even more dramatic effect is expected for higher moments, the details of which are discussed in Sect. "Scaling". Anticipating Subsect. "Moment Analysis" somewhat, the difficulty of estimating higher moments reliably lies in their higher demand of independent measurements (also discussed in Subsect. "Estimators"). To prevent systematic underestimation, at least a certain fraction of measurements must be taken from above $\langle s^n \rangle^{1/n}$. The number of measurements required increases significantly with n and is further increased in cases of long time correlation and pathological distributions. In the presence of rare events more sophisticated importance sampling techniques have to be used [14,15] to calculate reliable estimates.

In financial applications, rare events are often discussed in the context of so-called "fat tails" [16], which refer to PDFs that are close to a normal distribution but still have significant support at about five to ten standard deviations away from the mean. Fat tails are sometimes characterized by power laws, which are often the focus of the analysis of complex systems, as discussed in this article.

Rare events are sometimes associated [17] with multiscaling, which is a form of power law scaling to be discussed in Subsect. "Gap Scaling Versus Multiscaling".

Random Number Generators The simulation of a complex system that is essentially discrete and can be represented by integers usually spends a significant fraction of time producing (pseudo) random numbers. If, on the other hand, the model is essentially continuous, i. e. requires floating point numbers and involves functions from the mathematical library, those are likely to consume most of the CPU-time.

This observation suggests two paths of optimization: Firstly, floating point operations should be avoided as much as possible. For example, as most random number generators produce random integers, a frequent comparison of the form `rand()/RAND_MAX < q` where `q` is a constant floating point number, should be replaced by `rand() < i`, where `i` is `q` rescaled by a factor `RAND_MAX`. Similarly, Boolean random variables, true with probability 1/2, should be replaced by random bits, obtained using a bit-mask on an integer random variable, which is shifted on many platforms most efficiently by adding it to itself.

The second obvious path of optimization addresses the (pseudo) random number generator. All modern random number generators, such as `rand2` and others from [18], various generators discussed in [19] or the Mersenne twister [20], are of comparable quality and usually have passed all standard tests [21,22]. Their periods, that is the number of pseudo random numbers generated until they repeat, are usually long enough for modern requirements. Different generators mainly differ in CPU-time consumptions and general acceptance. Linear congruential random number generators and some of those found in the Standard C Library, on the other hand, generally lack the quality required in modern computer simulations. They often fail standard tests, yet some of them are very fast.

Different numerical simulations are independent only if the random number sequences used are independent. Only very few random number generators, mostly those designed for the use in parallelized simulations, guarantee the independence of sequences for different seeds, which in some cases have to be generated separately. It is good practice to seed the random number generator in a controlled way that ensures that none of the seeds is used more than once.

Estimators

Similar to experiments, numerical simulations suffer from different sources of error, many of which can be reliably estimated or even cured. In a Monte-Carlo simulation, expectation values are estimated using an **estimator**, for example the estimator for the nth moment

$$\langle s^n \rangle = \int \mathrm{d}s\, s^n \mathcal{P}(s) \tag{6}$$

from a simulation producing a sequence $s_1, s_2, \ldots, s_N$ is

$$\overline{s^n} = N^{-1} \sum_{i=1}^{N} s^n \,. \tag{7}$$

For the sake of the argument, in the following it will be assumed that all moments exist.

An estimator is called **unbiased** if the expectation value of the estimator coincides with the expectation value of the object to be estimated, independent of the sample size N. Provided the sequence $s_1, \ldots, s_N$, introduced above, was taken from the stationary regime, the above example, Eq. (7), estimating $\overline{s^n}$, is unbiased, because $\langle \overline{s^n} \rangle = \langle s^n \rangle$ since $\langle s_i^n \rangle = \langle s^n \rangle$ for every element in the sequence. On the other hand, $\overline{s^n}^2$ is not an unbiased estimator of $\langle s^n \rangle^2$, even if the elements in the sequence are independent. In this case,

$$\langle \overline{s^n}^2 \rangle = \langle s^n \rangle^2 + N^{-1} \left(\langle s^{2n} \rangle - \langle s^n \rangle^2 \right) \tag{8}$$

which has an explicit dependence on N unless the variance of s^n vanishes. Consequently, the unbiased estimator of the variance of s^n, denoted as $\overline{\sigma}^2(s^2)$ is

$$\overline{\sigma}^2(s^n) = \frac{N}{N-1} \left(\overline{s^{2n}} - \overline{s^n}^2 \right) \tag{9}$$

assuming mutual independence of the measurements.

An estimator is called **consistent** if it converges to the quantity to be estimated as the sample size is increased. The law of large numbers ensures this property for simple means, however, more complicated objects to be estimated might not posses an estimator that is obviously consistent.

Numerical Error The central aim of a numerical simulation is to produce, efficiently, large sample sizes, with small or even vanishing correlation time, estimating observables using unbiased and consistent estimators. The quality of the numerical estimate, i. e. its error is usually measured by means of the standard deviation of the estimator, which, in turn, is to be estimated as well. The standard deviation measures the width of the Gaussian describing the expected normal distribution of the estimator.

The standard deviation, $\sigma(\cdot)$, is defined as the square root of variance $\sigma^2(\cdot)$. The variance of the estimator of $\langle s^n \rangle$, introduced in Eq. (7) is

$$\sigma^2 \left(\overline{s^n} \right) = \left\langle \overline{s^n}^2 \right\rangle - \left\langle \overline{s^n} \right\rangle^2 = N^{-1} \left(\langle s^{2n} \rangle - \langle s^n \rangle^2 \right) = N^{-1} \sigma^2(s^n) \tag{10}$$

using Eq. (8), assuming that the sample is uncorrelated. A naïve estimator of this variance is $N^{-1}(\overline{s^{2n}} - \overline{s^n}^2)$, which is, however, a biased estimator, because of Eq. (8):

$$\left\langle N^{-1} \left(\overline{s^{2n}} - \overline{s^n}^2 \right) \right\rangle = N^{-1} \sigma^2(s^n) \frac{N-1}{N} \tag{11}$$

so that the unbiased estimator for the variance of the estimator of $\langle s^n \rangle$ is in fact

$$\overline{\sigma}^2 \left(\overline{s^n} \right) = (N-1)^{-1} \left(\overline{s^{2n}} - \overline{s^n}^2 \right) . \tag{12}$$

When estimating functions of means, such as the variance

$$f \left(\langle s \rangle, \langle s^2 \rangle \right) = \langle s^2 \rangle - \langle s \rangle^2 \tag{13}$$

finding an unbiased estimator for it might already be a difficult task. The most natural choice is of course $f(\overline{s}, \overline{s^2})$ but as pointed out above, Eq. (9), this choice is biased.

The task of finding an estimator for the variance of the estimator might be even more problematic. The most naïve approach is to assume independence and approximate the variance by error propagation

$$\sigma^2(f) \approx \sigma^2(\overline{s}) \left(\left. \frac{\partial f}{\partial \overline{s}} \right|_{\langle \overline{s} \rangle, \langle \overline{s^2} \rangle} \right)^2 + \sigma^2 \left(\overline{s^2} \right) \left(\left. \frac{\partial f}{\partial \overline{s^2}} \right|_{\langle \overline{s} \rangle, \langle \overline{s^2} \rangle} \right)^2 . \tag{14}$$

A first attempt to include correlations is

$$\sigma^2(f) \approx \sigma^2(\overline{s}) \left(\frac{\partial f}{\partial \overline{s}} \right)^2 + \sigma^2 \left(\overline{s^2} \right) \left(\frac{\partial f}{\partial \overline{s^2}} \right)^2 + 2 \operatorname{covar} \left(\overline{s}, \overline{s^2} \right) \frac{\partial f}{\partial \overline{s}} \frac{\partial f}{\partial \overline{s^2}} \tag{15}$$

where $\operatorname{covar} \left(\overline{s}, \overline{s^2} \right)$ denotes the covariance

$$\operatorname{covar} \left(\overline{s}, \overline{s^2} \right) = \left\langle \overline{s} \overline{s^2} \right\rangle - \left\langle \overline{s} \right\rangle \left\langle \overline{s^2} \right\rangle . \tag{16}$$

Similar schemes are often used to estimate variances of very complicated functions of the observables, sometimes assuming that the expectation value of the function coincides with the function of the expectation values of the observables, which clearly is only the case for linear functions.

One widely used class of methods available to reduce bias in estimators and the estimators of their variances are the Bootstrap and the Jackknife [23,24]. Both methods are so-called "resampling plans", which prescribe a method to construct estimators and estimate their variances from the distribution of estimates based on sub-samples. In case of the Jackknife, a sample containing N individual measurements produces N sub-samples by taking all measurements apart from the first, second, third and so on. The estimator of the desired quantity is applied to each of the N sub-samples, each of which containing $N-1$ measurements. More elaborate schemes are available particularly suited to small sample sizes.

A very similar approach is to measure the desired quantity in, say M, independent measurements, such as different Monte-Carlo simulations with different random number sequences, each consisting of N individual measurements. If N is large enough, the central limit theorem renders the distribution of the M estimates indistinguishable from a normal distribution. The error of the mean is then easily derived. This scheme lends itself to a simulation consisting of many independent runs, if they are necessary anyway to determine a quantity reliably. It can be applied in a weighted fashion if individual runs consists of different sample sizes.

Correlation Time If the sequence $s_1, s_2, \ldots, s_N$ used in Eq. (7) is observed in a Markov process, the individual measurements will not be mutually independent: If $P(s_2|s_1)$ denotes the probability to observe s_2 given the directly preceding sample was s_1, then $P(s_2|s_1)$ is, in general, not independent of s_1. The same is true for $n > 1$ in $P(s_n|s_1)$, which is generally not independent of s_1, and even $P(s_n|s_{n-1}, s_{n-2}, \ldots, s_1)$ or any other probability of s_n given a previous (sub-)sequence. Only if the probability of s_n depends only on the directly preceding sample s_{n-1} but not on any earlier one, then the sequence $s_1, \ldots, s_n$ *itself* is a Markov process and fully specified by the transition probabilities between successive s_i.

The **independence** of two measurements s_i and s_j means that their joint probability factorizes. It implies that they are **uncorrelated**, which means that the expectation value of their product factorizes, $\langle s_i s_j \rangle = \langle s_i \rangle \langle s_j \rangle$. The converse is generally not true, but frequently assumed.

The correlation time in a sample is derived from the **autocorrelation function** in the stationary state,

$$C(j) = \frac{\langle s_i s_{j+i} \rangle - \langle s \rangle^2}{\sigma^2(s)} \tag{17}$$

which is independent of i because of stationarity. The autocorrelation function is normalized with the variance of the observable s,

$$\sigma^2(s) = \langle s^2 \rangle - \langle s \rangle^2 \tag{18}$$

so that $C(0) = 1$. Motivated by the study of Markovian processes [6], the **correlation time τ** of a sequence is estimated by fitting $C(j)$ to an exponential $\exp(-j/\tau)$. One can show that the mean calculated from a finite sequence s_i, $i = 1, 2, \ldots, N$, with correlation time τ has a variance equal to the variance of the mean derived from a set of $N/(2\tau + 1)$ uncorrelated measurements.

Obviously, in numerical simulations a sample with large correlation time is inferior to one with a smaller correlation time. The estimated mean from an uncorrelated sample comprising of N measurements has variance $\sigma^2(\bar{s}) = \sigma^2(s)/N$, so that along the simple arguments presented above, the variance of the mean increases (almost) linearly in the correlation time and decreases inversely in the sample size. As discussed further below, the square root of the variance of the estimator is usually used as a measure of the statistical error. When comparing different numerical techniques in terms of their computational costs, i. e. their CPU-time, one therefore has to compare the product of the square of the error and CPU-time.

Reducing the Correlation Time A sample of **sample size** N and correlation time τ contains τ correlated subsamples which have only correlation time 1, $s_1, s_{1+\tau}, s_{1+2\tau}, \ldots$. By pruning the original sample or, equivalently, reducing the sampling rate in a numerical simulation, one can reduce the correlation time to arbitrarily small values. This, however, is achieved only with the additional computational cost for the intermediate states of the system which are produced but not sampled. Such a technique pays off only if the cost of the sampling is high compared to the production of a new state of the system, as for example if the (expensive) Fourier transform is to be taken in a system the configuration of which evolves at virtually no costs. Often, it is more efficient to include all correlated sub-samples.

Scaling

Scaling is the hallmark of universality [25,26]. One of the aims of the analysis of a complex system is to determine whether it displays scaling and if so, whether the characteristics of the scaling behavior suggest that it belongs to a certain universality class. The **universality hypothesis** makes it possible to extend the predictive power of a simple model to more realistic situations and entire classes of stochastic processes.

Finite Size Scaling Hypothesis

The PDF $\langle P(s; s_c) \rangle$ of an observable expected to display scaling is tested against the scaling hypothesis also known as **simple scaling**

$$\langle P(s; s_c) \rangle = \begin{cases} a s^{-\tau} \mathcal{G}(s/s_c) & \text{for} \quad s > s_0 \quad (19a) \\ f(s; s_c) & \text{otherwise} \quad (19b) \end{cases}$$

where a is a **metric factor** (constant and in particular independent of s_c), τ denotes a **(critical) scaling exponent** (not to be confused with the correlation time) and $\mathcal{G}(x)$ is the **scaling function** usually assumed to be universal up to

a pre-factor. The **upper cutoff** s_c is the typical scale of the observable for which the density function rapidly drops to 0. It is also called "characteristic event size" and coincides, up to a pre-factor, with the first moment $\langle s \rangle$ in the case $\tau = 1$.

The **lower cutoff** s_0 separates a range of values of the observable s where the histogram is governed predominantly by the non-universal PDF $f(s; s_c)$. Equation (19) applies only approximately in discrete systems and in general only asymptotically, i. e. for sufficiently large s, s_c. Taking this limit appropriately, quickly becomes very technical and for that reason will not be discussed in detail in the following.

The upper cutoff s_c is often assumed to diverge with the system size L as a power-law, $s_c = bL^D$ with D being another (critical) exponent, which can be regarded as the spatial dimension of s_c and therefore of the observable s. The pre-factor b is a second metric factor. Together with a, these two factors are the only **non-universal** parameters entering the universal part of the PDF, Eq. (19). In order to test for universal behavior, one has to impose that $\mathcal{G}(s/s_c)$ is a dimensionless function of a dimensionless argument. For dimensional consistency, a and b both are dimensionful, unless $\tau = 1$ in which case a is a pure number which can in principle be absorbed into $\mathcal{G}$, or if $D = 1$, in which case b is a number which can be absorbed into $\mathcal{G}$ as well.

If s_c depends only on L, (19) describes **finite size scaling**. In standard critical phenomena [25] this is observed only exactly at the critical point. Away from it, the upper cutoff depends on an external tuning parameter, such as the reduced temperature t, which vanishes at the critical point of a ferromagnetic phase transition. The PDF then is still expected to display scaling as in (19), to be investigated in the thermodynamic limit, i. e. $L \to \infty$. In this case one expects $s_c = b'\xi^D$ with b' being another metric factor, ξ the correlation length and D the same spatial dimensionality as above. Similarly τ is expected to be the same, while the scaling function differs in finite size scaling and critical (point) scaling. Most importantly, the finite size scaling function depends on more details of the system than the critical scaling function, such as the shape, topology and aspect ratio of the system as well boundary conditions [26].

The difference between finite size scaling function and critical scaling function is generally explained to be caused by the difference in correlation length: The statistics of a finite system at the critical point displaying finite size scaling depends on the geometric properties of the boundaries of the system as the correlation length reaches them. In case of critical scaling the system can often be treated as being composed of many independent sub-systems, not probing the boundaries. Therefore, the distribution of some observables, such as the order parameter, is normal if the thermodynamic limit is taken when the system is not at the critical point. Other PDFs, such as the cluster size distribution in percolation, remain non-trivial but different from finite L, as they are not subject to the central limit theorem [6,7].

Complex systems and most notably self-organized criticality very often display finite size scaling and lack an explicit tuning parameter. Therefore, the following focuses on finite size scaling.

Data Collapse By plotting the numerical estimate of $\langle P(s; s_c)\rangle(s)$ in a double logarithmic plot versus s, the slope of the resulting graph in the scaling region, i. e. the region dominated by the power-law behavior, $s_c \gg s \gg s_0$, gives the **apparent exponent** $\tilde{\tau}$. This is illustrated in Fig. 1b. According to Eq. (19a) the slope is $-\tau$ plus any contribution due to the scaling function. Only if the scaling function converges to a constant for small arguments, $\lim_{x\to 0} \mathcal{G}(x) > 0$, will the apparent exponent $\tilde{\tau}$ coincide with the (actual) scaling exponent τ.

From the slope of the PDF in a double-logarithmic plot a first estimate for τ can be derived and can subsequently be used in a **data collapse**, which consists in plotting the estimates of the PDF in the rescaled form $s^\tau \overline{P(s; s_c)}$ versus s/s_c. For each individual measurement s_c is to be estimated, but the same τ is assumed to apply to all of them. If the result follows simple scaling, then there is a value of s_c for each measurement so that all data $s^\tau \overline{P(s; s_c)}$ for $s > s_0$ collapse onto the same curve, representing the scaling function in the form $a\mathcal{G}(s/s_c)$, as exemplified by the so-called Oslo model [27] in Fig. 2. A failed data collapse is shown for small system sizes of the BTW model in Fig. 1d. The construction of a data collapse is discussed further in Subsect. "Binning During Data Analysis".

The different measurements with s_c varying between them are usually taken from different system sizes. Plotting s_c versus the system size in a double-logarithmic plot reveals the spatial dimension or gap exponent (see Subsect. "Gap Scaling Versus Multiscaling") D, since $s_c = bL^D$, possibly with corrections to scaling (see Subsect. "Corrections to Scaling"). The same strategy is applied if s_c is expected to depend on a different parameter, such as a temperature.

There is no established method for quantifying the quality of a data collapse and it therefore remains somewhat subjective. There is, however, a simple method to rule out simple scaling, for example if the tail of the scaling function, i. e. the region of large arguments, changes shape for different values of s_c, as in the BTW model at least in the case of small system sizes, Fig. 1d, or if the lower cut-

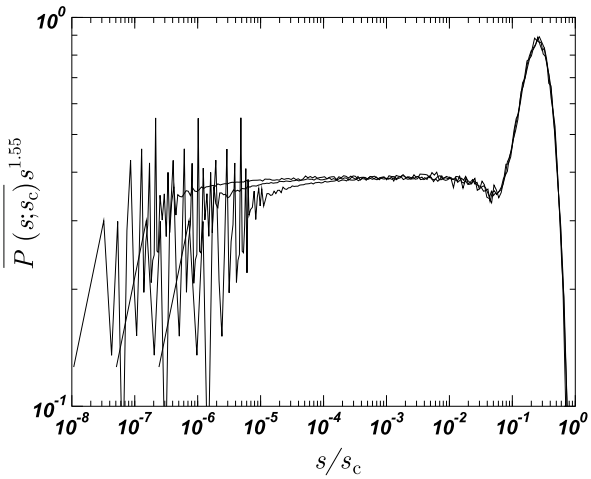

Probability Densities in Complex Systems, Measuring, Figure 2
Example of a data collapse, here for the avalanche size distribution of the one-dimensional Oslo model, [27], driven at a single site, with system sizes $L = 640, 1280, 2560$. The upper cutoff is expected to scale like $s_c \propto L^{2.25}$. The data collapses for large values of s. For small s, the behavior is non-universal

off changes significantly with increasing s_c as in the Forest Fire Model [28]. It is particularly important to probe for this phenomenon, because the quantitative analysis based on moments of the PDF is generally unable to detect it, see Subsect. "Variance and Numerical Error of Higher Moments".

The Lower Cutoff The lower cutoff s_0 appears in Eq. (19) to distinguish the universal part of the PDF shaped by $\mathcal{G}(s/s_c)$ from the non-universal part given by $f(s;s_c)$. Below the lower cutoff complex and critical systems are expected to be governed by microscopic physics, i. e. the specific details of the process. In this region, the PDF might depend on additional parameters. More often however, the processes on a very small scale are expected to be similar, so that often $f(s;s_c) \approx f(s;s_c')$ even for s_c very different from s_c'. This is found, for example, in critical percolation, where s_c is solely determined by the system size and s is chosen to be so small that the finiteness of the system is irrelevant. This feature is also visible in the patterns at low s in Fig. 2. Different definitions of the observable usually have an impact in this region, but not for the asymptotic, large s behavior.

Some complex systems do not possess a lower cutoff, so that the entire region of accessible values of the observable s is governed by the universal behavior. Although discrete systems have a natural lower cutoff given by the smallest possible event, this might be hidden in the definition range of the observable, as for example in the case of the avalanche size distribution of the one-dimensional BTW model, which follows $\langle P(s)\rangle = s^{-1}(s/s_c)\theta(1-s/s_c)$ with $s \in \{1,2,\ldots,s_c\}$, $s_c = L$ and θ denoting the Heaviside step function.

A continuous system without a lower cutoff is physically equivalent to one without an upper cutoff, because by suitable rescaling the range of observables s tested can be made arbitrarily large.

Corrections to Scaling Equation (19a) describes only the leading order behavior of the PDF, which is only asymptotically valid as $s_c \gg s_0$ and $s \gg s_0$. The former condition, $s_c \gg s_0$, suppresses contributions from the non-universal part and can in case of finite size scaling be achieved by increasing the system size. The latter condition, $s \gg s_0$, is reflected in the condition $s > s_0$, however the larger s_0 is chosen, the smaller the error of Eq. (19a).

The corrections accounting for the failure of the exact scaling behavior are known as **corrections to scaling** [29], originally introduced in the context of ferromagnetic phase transitions. In principle, the scaling form Eq. (19a) can be generalized to

$$as^{-\tau}\mathcal{G}(s/s_c) + a_1 s^{-(\tau+\omega_1)}\mathcal{G}_1(s/s_c) + \ldots \tag{20}$$

with $\omega_1 > 0$. It is much more common to account for corrections to scaling on the level of individual observables, such as moments or s_c, instead of the entire distribution. For example, the standard finite size scaling ansatz including corrections to scaling is

$$s_c = bL^D(1 + b_1 L^{-\omega_1'} + b_2 L^{-\omega_2'} + \ldots) \tag{21}$$

with the exponents ω_i and ω_i' expected to be universal.

Moments

The data collapse described in the previous section suffers from the problem that its quality is not easily quantifiable. Only the obvious failure gives a strong indication of the absence of finite size scaling in the form Eq. (19).

A more quantitative tool is the **moment analysis**. In principle, moments can be derived from the histogram after the simulation, but the error introduced by binning methods, which are almost always a necessity, or rounding errors in case of continuous event sizes can be very big. They are easily avoided by calculating the moments during the simulation.

Numerics of Moments Moments are usually calculated during the simulation and provide access to the scaling be-

havior of the system without large memory or computational requirements. However, as individual contributions to moments can vary greatly in value, the quality of the estimate relies crucially on the precision of the numerical calculation.

The standard method to calculate the nth moment of $\overline{P(s;s_c)}$ is to sum all nth powers of individual event sizes s and finally divide by the number of contributions. If the process takes place in continuous time, each event size might carry an additional weight with it, and the normalization required is the sum over all weights. This weight usually corresponds to the amount of time the observable had a particular value now entering the estimator.

Depending on the type of process, measuring the moments can contribute significantly to the overall computing time. One method to minimize it, is to reduce the number of additions and multiplications. An implementation could read

```
// Normalization
power=weight;
moment[0]+=power;
// First moment
power*=event_size;
moment[1]+=power;
// Second moment, etc.
power*=event_size;
moment[2]+=power;
power*=event_size;
moment[3]+=power;
...
```

which can be further simplified in the form of a loop. Moments of very high order can be generated in the form

```
// Normalization
moment[0]+=weight;
// First moment
power=event_size;
moment[1]+=power*weight;
// Second moment
power*=power;
moment[2]+=power*weight;
// Fourth moment
power*=power;
moment[3]+=power*weight;
...
```

Precision is crucial in particular when it is a priori unknown whether the main contribution to a moment is due to many small events or a few very big ones. Integer variables are an option only if the weight is integer-valued or can be rendered so. They are ideal, because they do not suffer from rounding errors. To avoid overruns, many platforms offer 64 bit integers. Integers are often computationally, i. e. in terms of CPU-time, advantageous as well.

Where floating point numbers are a necessity, they have to be of appropriate size. A central criterion is whether small events can still enter with sufficient accuracy at the end of the simulation, when the various variables holding *sums* of different powers of the observables contain very large numbers. Adding a sufficiently small number to a large floating point number might not actually change it, depending on the size of the mantissa. The IEEE 754 standard [30] describes floating point numbers with 24, 53 and 64 bit mantissae.

It is computationally very inefficient to study moments $\langle s^n \rangle$ for $n \notin \mathbb{N}$, because it requires a floating point operation, such as `sqrt` or `pow`. Where such moments are unavoidable, these operations can often be "recycled", by, say, calculating `pow(event_size, 1./3.)`, taking its square and constructing from these two values plus `event_size`, all powers 1/3, 2/3, 1, 4/3, 5/3, ... using a minimal number of multiplications.

Moment Analysis Simple finite size scaling, Eq. (19), implies that the moments $\langle s^n \rangle$ of the PDF scale like a power of the upper cutoff s_c,

$$\lim_{s_c\to\infty} \frac{\langle s^n \rangle}{s_c^{n+1-\tau}} = a \lim_{y\to 0} g_n(y) \quad \text{for} \quad n > \tau - 1 \tag{22}$$

where the **amplitude** $g_n(0)$ is a moment of the universal scaling function

$$g_n(y) = \int_y^\infty \mathrm{d}x\, x^{n-\tau} \mathcal{G}(x)\,, \tag{23}$$

which is universal up to a pre-factor, i. e. any ratio of these amplitudes is universal. Assuming that the scaling function $\mathcal{G}(x)$ is continuous and has no singularity at finite argument, one can prove that the limit $\lim_{y\to 0} g_n(y)$ exists. Moreover, as $g_0(0) > 0$, the exponent τ cannot be less than unity.

Simplifying Eq. (22)

$$\langle s^n \rangle \propto s_c^{\gamma_n} \quad \text{with} \quad \gamma_n = 1+n-\tau \quad \text{for} \quad n > \tau - 1 \tag{24}$$

the scaling exponents γ_n are usually determined for n ranging from 1 up to typically 8 as the slope in a double-logarithmic plot of the moment versus the system parameter, which in case of finite size scaling is the linear size of the system. If no such parameter is quantifiable, moments can be measured with respect to each other, noting that

$$\langle s^n \rangle \propto \langle s^m \rangle^{\frac{\gamma_n}{\gamma_m}} \quad \text{for} \quad n > \tau - 1\,. \tag{25}$$

A priori the lower cutoff and the corrections to scaling are unknown and in order to determine the (asymptotic) finite size scaling behavior, a numerical study has to probe system sizes as large as possible. The true asymptote will, by definition, strictly remain inaccessible.

Gap Scaling Versus Multiscaling Moments following the scaling behavior prescribed by (22) are said to display **gap scaling** with D being the **gap exponent** [31]. This term refers to the constant gap between the scaling exponents for consecutive moments as a function not of s_c itself but of the tuning parameter, usually the system size. Assuming Eq. (21) and using the simplified notation Eq. (24) finite size scaling implies

$$\langle s^n \rangle \propto L^{\gamma'_n} \quad \text{with} \quad \gamma'_n = D(1 + n - \tau) \tag{26}$$

so that two consecutive moments have an exponent differing by $\gamma'_{n+1} - \gamma'_n = D$, visible as a constant slope when plotting the scaling exponent as a function of n, see Fig. 3. Measuring this slope is the standard method for calculating the exponent D.

The intersection of the linear continuation of the γ_n with the abscissa gives $\tau - 1$, which is a standard method of estimating τ. An exponent $\gamma_n > 0$ indicates that the corresponding moment is diverging as s_c diverges. For $\tau > 1$ there are some moments for non-negative n which do not diverge. If $\tau = 1$ the only non-negative moment that does not diverge is the normalization $n = 0$.

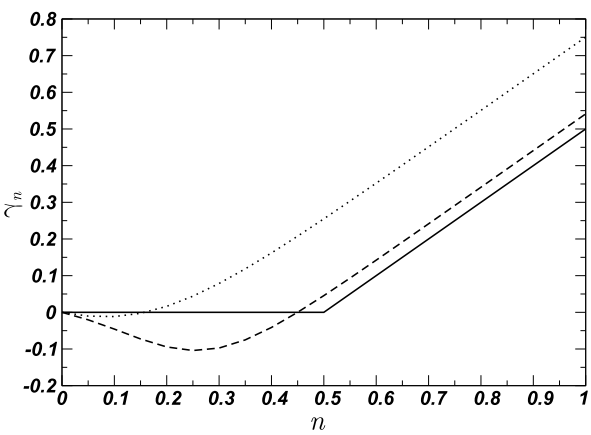

Probability Densities in Complex Systems, Measuring, Figure 3
The exponents γ_n estimated from the moments calculated from $\langle P(s; s_c)\rangle$ as introduced in Fig. 4. The *dotted line* shows the exponents estimated from s_c ranging from 25 to $1 \cdot 10^4$, the *dashed line* shows exponents based on s_c ranging from $1 \cdot 10^4$ to $6.4 \cdot 10^5$. The *full line* shows the exact values the numerical data converges to. The rounding in the numerical estimates can be mistaken as multiscaling

If the exponents γ_n do not increase linearly in n yet the moments still display scaling behavior, the system is said to exhibit multiscaling. In this case, the scaling form Eq. (22) is often extended to

$$\lim_{s_c \to \infty} \frac{\langle s^n \rangle}{\ln(s_c)^{\lambda_n} s_c^{\gamma_n}} \tag{27}$$

to allow for logarithmic contributions to the scaling behavior. The presence of these logarithmic "corrections" are often interpreted as a sign of multiscaling, as is the presence of **rare events**, see Subsect. "Rare Events".

Multiscaling often means that the γ_n converge for large n and only the λ_n continue to increase. Multiscaling is sometimes confused with the absence of scaling altogether, or with the rounding effect observed close to $n = \tau - 1$, as shown in Fig. 3.

The standard method of estimating τ and the gap exponent D is to fit the moments $\langle s^n \rangle$ to a power-law, see (24) and Eq. (26), possibly including corrections to scaling (Subsect. "Corrections to Scaling") and usually by plotting not versus s_c but versus some system parameter such as the size L. The resulting γ_n or γ'_n are then fitted against $1 + n - \tau$ or $D(1 + n - \tau)$ respectively to determine τ and possibly D.

Variance and Numerical Error of Higher Moments The variance of the nth moment is $\langle s^{2n} \rangle - \langle s^n \rangle^2 = \langle (s^n - \langle s^n \rangle)^2 \rangle$. This difference is always non-negative and comparing the scaling exponents for both contributions from Eq. (24) confirms $\gamma_{2n} \geq 2\gamma_n$, where the equality holds only if $\tau = 1$. Thus, unless $\tau = 1$, the standard deviation of the estimator of the nth moment is expected to scale with exponent $(1/2)\gamma_{2n} = nD + (1/2)(1 - \tau)$. The relative variance, also known as the relative fluctuations, therefore is expected to scale as

$$\frac{\langle s^{2n} \rangle}{\langle s^n \rangle^2} \propto s_c^{\gamma_{2n} - 2\gamma_n} = s_c^{\tau - 1} \tag{28}$$

and therefore diverges asymptotically for $\tau > 1$ as the upper cutoff increases.

For $\tau = 1$ the relative variance might not change with s_c, although the ratio $\sqrt{\langle s^{2n} \rangle / \langle s^n \rangle^2}$ might have a strong dependence on n. Because (24) determines only the leading order scaling of the moments but makes no statement about the respective amplitude, the difference $\langle s^{2n} \rangle - \langle s^n \rangle^2$ might have a leading order that scales slower than γ_{2n} if $\tau = 1$. In this case the relative variance might asymptotically vanish, which is known as **self-averaging** [32].

Self-averaging is more commonplace in the context of scaling with system size away from a critical point. In this

case, spatial densities usually follow a Gaussian distribution and the system can be decomposed into finite patches, the number of which grows linearly in the volume. Therefore, the relative variance of a density that converges in the thermodynamic limit to a finite value, decreases like L^{-d}. This effect is known as **strong self-averaging** [33]. Where the variance decays with increasing L, but not as fast as L^{-d}, the effect is known as **weak self-averaging**. The variance can decrease faster than L^{-d} only in the (rare) presence of anti-correlations.

A Monte–Carlo simulation can be regarded as a means to integrate a function, so that estimating $\langle s^n \rangle$ means in fact to perform the integral

$$\overline{s^n} = \int \mathrm{d}s\, \overline{P(s; s_c)} s^n \,. \tag{29}$$

The most efficient method to estimate $\langle s^n \rangle$ samples exactly with a density $\propto \langle P(s; s_c) \rangle s^n$, which in principle can be achieved using (ideal) importance sampling. In practice, this is rarely possible and only very crude approximations to $\langle P(s; s_c) \rangle s^n$ are available as sampling frequencies. In the context of complex systems, $\langle P(s; s_c) \rangle$ itself is the sampling rate, in the vast majority of problems.

The comparison to ideal importance sampling suggests an alternative, much stronger criterion to assess the quality of a numerical estimate. The product $\langle P(s; s_c) \rangle s^n$ usually has its maximum at some $s_{\max}$ close to s_c and one might therefore compare the density of measurements around $s_{\max}$ to the weight this range of s enters the integral (29) [34].

Correspondingly, one can derive the number of measurements needed to perform the integral (29) reliably by imposing a lower bound on the density of measurements around s as a fraction of $\langle P(s; s_c) \rangle s^n$. Demanding that the density of measurements around s is nowhere less than $\rho^* \overline{P(s; s_c)} s^n$, then gives rise to a scale s^*, so that at least one sample is to be taken above the "characteristic largest event" s^*,

$$\int_{s^*}^{\infty} \mathrm{d}s\, \rho^* \overline{P(s; s_c)} s^n = 1 \tag{30}$$

which depends on n and is to be compared to the total weight in the sample actually produced in the simulation $\tilde{P}(s; s_c)$,

$$\tilde{N}(s^*) = N \int_{s^*}^{\infty} \mathrm{d}s \tilde{P}(s; s_c) \tag{31}$$

where N is the total number of measurements taken and $\tilde{P}(s; s_c) = \overline{P(s; s_c)}$ is the sampling frequency unless importance sampling is used. The event size s^* usually is of the order of the upper cutoff s_c and the number of measurements needed typically scales itself like s_c^n.

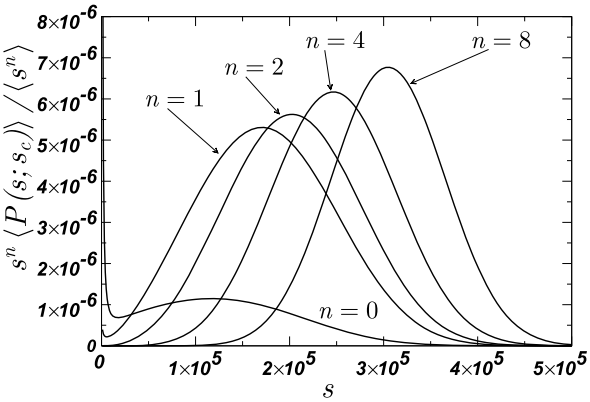

Probability Densities in Complex Systems, Measuring, Figure 4
Moments for larger n draw more weight from larger s. A moment analysis investigating large moments might possibly miss features of the distribution at small arguments. The distribution used here is $\langle P(s; s_c) \rangle = a s^{-3/2} \mathcal{G}(s/s_c)$ for $s \in [1, \infty$ with $\mathcal{G}(x) = (1 + x^2) \exp(-x^3/s_c)$ and $s_c = 10\,000$

Thus, this much stronger criterion, constraining the sampling density even in the extreme region of very large s produces a much stronger constraint on the sample size than suggested by the relative variance, which increases with s_c only as $s_c^{\tau-1}$, see Eq. (28).

Figure 4 shows the product $s^n \langle P(s; s_c) \rangle / \langle s^n \rangle$ for a range of n as a function of s. As indicated above, the larger n, the more weight enters from the tail of the distribution, which implies that small s features are more and more ignored with increasing n. While increasing n are needed in order to derive the **gap exponent** (see Subsect. "Gap Scaling Versus Multiscaling") reliably, an analysis that relies solely on moments can miss a lack of finite size scaling for small s, such as divergent lower cutoff.

This problem can be avoided either by careful inspection of the data collapse, or by accurate determination of the scaling of moments for very small $n > \tau - 1$, which might necessitate the investigation of fractional moments, see Fig. 3. The closer the moment is to $\tau - 1$, the slower the convergence to the asymptotic behavior. This is illustrated in Fig. 5a which shows the scaling of the first moment $\langle s \rangle$ in a system with $\tau = 1.5$.

Whether displaying and analyzing the histogram or moments, it is generally significantly more informative to plot the data with the leading order divided out, for example $\bar{s} s_c^{-1/2}$ if $\bar{s} \propto s_c^{1/2}$ is expected as shown in Fig. 5b. If the numerical data converges to a constant, this confirms the scaling, while the shape of $\bar{s} s^{-1/2}$ indicates the rôle of corrections to scaling. Moreover, the spread of the numerical

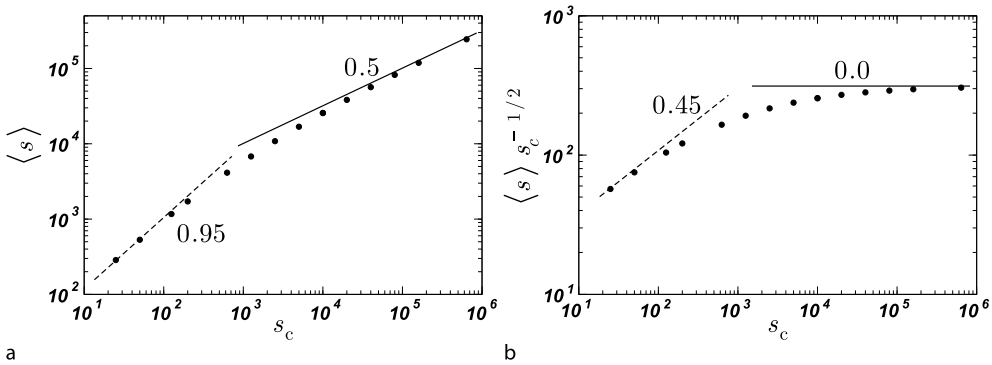

Probability Densities in Complex Systems, Measuring, Figure 5

Scaling of the first moment $\langle s \rangle$ of $\langle P(s; s_c) \rangle$ (see Fig. 4). a In a double-logarithmic plot, the slope ranges from 0.95 (*dashed line*) to about 0.5 (asymptotically exact; *full line*) as indicated. Correction terms (corrections to scaling) or very large system sizes are necessary for the estimated exponent to be sufficiently close to the asymptote, which is $\gamma_1 = 0.5$ known from construction. This becomes more apparent when plotted in the form shown in b: The same data but plotted as $\langle s \rangle s_c^{-1/2}$. In this form it is easier to see that the asymptotic behavior is only reached by probing system sizes at least as big as the largest one shown

data across the ordinate is reduced, allowing for a more detailed inspection of the details. This is discussed further in Subsect. "Binning During Data Analysis" for the scaling of histograms.

Critical Slowing Down Large upper cutoffs are needed to isolate the asymptotic behavior on the one hand, but they are numerically very demanding on the other hand. Another effect which reduces the effectiveness of Monte–Carlo simulations, known from ferromagnetic phase transitions, is **critical slowing down**.

The (auto-)correlation time discussed in Subsect. "Correlation Time" effectively reduces the number of independent measurements produced in the simulation. Critical slowing down describes the behavior of the correlation time at the critical point or in finite size scaling: The correlation time τ, see Subsect. "Correlation Time", diverges like a power-law of the tuning parameter, which in finite size scaling means that $\tau \propto L^z$, where z is known as the dynamical exponent.

Taking all different effects into account, the increasing relative variance for $\tau > 1$ (exponent τ, see Eq. (19)), the minimal number of measurements required for a certain minimal coverage and finally critical slowing down, large system sizes always represent a challenge to the quality of the estimate of moments, in particular for large n. These are, however, necessary for a reliable estimate of the asymptotic behavior of the PDF.

Moment Ratios A priori the amplitude a on the RHS of (22) is unknown; definition Eq. (19a) fixes only τ and otherwise states only that there exist quantities a and $\mathcal{G}(x)$, fixed only up to a pre-factor, so that the universal part of the PDF obeys Eq. (19a).

By taking ratios of quantities containing non-universal pre-factors these are removed. For example,

$$\frac{g_n(0)\, g_1^{n-2}(0)}{g_2^{n-1}(0)} = \frac{\langle s^n \rangle \langle s \rangle^{n-2}}{\langle s^2 \rangle^{n-1}} \tag{32}$$

contains only moments of the scaling function with any non-universal pre-factor removed. Moreover, one can show that (32) converges to a non-zero value if $\langle s^n \rangle$ follows Eq. (24).

For $\tau = 1$ the metric-factor a becomes dimensionless and can therefore be fully absorbed into the scaling function. In ferromagnetic critical phenomena and related phase transitions, such as percolation, where $\tau = 1$, ratios of the form $\langle s^{n\,m} \rangle / \langle s^m \rangle^n$ are universal already.

Histogram Data Representation

Naïvely, collecting the histogram in a simulation of a complex system with integer valued event sizes amounts only to incrementing a variable `histogram [event_size]++`. While this procedure is very efficient, provided that the number of processor cycles needed to access `histogram[event_size]` is minimal, the memory requirements quickly exceed that of standard computers as s_c increases. Moreover, histogram entries for small event sizes might overflow, necessitating very large data types to accommodate the largest entry in the histogram. At the same time, entries for very large event sizes are very sparse. Using this technique for continuous event sizes means that they have to be rounded.

Various methods of **binning** are established to handle these problems. Binning is usually also used at the stage of data analysis, i. e. in the preparation of data after the actual simulation has taken place. The key is to map the complete range of event sizes to the range available in the representation of the histogram. In the context of computing, the map is effectively a hash-function, in the following $h(s)$. Assuming that the event size lies within a certain, finite range and is non-negative, $s \in [0, s_{\max}] = \Omega$, not necessarily discrete, sets $\mathbb{S}_l \subset \Omega$ are defined so that all $s \in \mathbb{S}_l$ are equivalent under the hash algorithm and distinct otherwise, i. e. $s, s' \in \mathbb{S}_l$ implies $h(s) = h(s') = l$ and vice versa. The sets are usually continuous, so that the hash algorithm $h(s)$ can be represented by a set of ranges, for example $h(s) = l$ for $s_l \leq s < s_{l+1}$. The hash function is then used in the form `histogram[h(event_size)]++`.

After collecting the data into bins, they are normalized by the bin size and often shown with respect to the geometric mean of the bin ranges $\sqrt{s_l s_{l+1}}$. There is obviously some freedom of choice, but if the choice makes a qualitative or quantitative difference, this implies that the binning is too coarse.

Binning Schemes Within the Simulation

Binning can take place either during the simulation or in the data analysis. In general, it is advisable to keep the simulation as simple as possible: If the binning scheme within the simulation fails, all data might be lost or rendered useless. If it fails at the stage of data analysis, it can easily be fixed and repeated. However, due to memory constraints, it is very often necessary to use binning within the simulation. In these cases, it is vital to choose the most efficient binning scheme to avoid a negative impact on the CPU time.

Power Law Binning The aim of binning generally is to reduce the histogram's memory requirements, while avoiding potential overflows. The first aim can be achieved by choosing the size of the ranges $s_{l+1} - s_l$ according to the (expected) PDF, which a priori is, however, unknown. Assuming a pure power law, $\langle P(s; s_c)\rangle = s^{-\tau}$, the ranges would need to be $s_l = (c(M - l))^{1/(1-\tau)}$ for $\tau > 1$, where c determines the size of the range and M is the maximum l. In case of discrete event sizes, obviously the ranges need to be rounded.

Even distribution of the events across the histogram has the additional advantage of equal error for each bin, provided that the error is only a function of the number of events. This might not be the case for very small events, where correlations might be significantly larger than for larger event sizes.

Although the binning ranges described above would produce ideal bins, it is generally not advisable to use an expected result, such as the exponent τ, in the process of collecting the raw data to measure this quantity.

Exponential Binning The scheme most suitable for presenting the data in a double-logarithmic plot is exponential binning. In this case the range limits are powers of some base s_b, so that $s_l = r s_b^l$ with some pre-factor r. As a result, the data points are evenly spread with a spacing of $\ln(s_b)$ after taking the logarithm of the abscissa. The error, on the other hand, increases towards larger s because of the smaller number of events collected in bins for larger event sizes. Approximating the PDF by a pure power law, the number of events in the bin with hash l is, up to the normalization

$$\int_{s_l}^{s_{l+1}} \mathrm{d}s\, s^{-\tau} = \frac{(r s_b^l)^{1-\tau}}{\tau - 1}(1 - s_b^{1-\tau}) \tag{33}$$

for $\tau > 1$ which decreases with increasing l. The width of each bin is $s_{l+1} - s_l = s_l(s_b - 1)$.

Exponential binning is frequently used and the one most suited for a direct calculation of the hash value during the simulation. Two simple methods can be distinguished for determining the hash value: Firstly, a function $\tilde{h}(s)$ can be devised so that, for example, its integer part gives the hash value. In case of exponential binning, this function is $\tilde{h}(s) = \ln(s/r)/\ln(s_b)$. Using functions from the mathematical library, such as `log` is, however, computationally very expensive.

Secondly, one can compare s to the various s_l until l is found so that $s_l \leq s < s_{l+1}$. Since small events are the most frequent ones, this might be implemented by simply linearly increasing l. A crude estimate of the expected number of comparisons shows that this number generally quickly converges with increasing s_c, as $\langle \ln(s)\rangle$ is finite.

The constant number of expected comparisons it to be compared to the obvious "divide and conquer" or tree-search approach: Given M bins, s is compared to $s_{M/2}$, in the next step, depending on the outcome of the previous comparison, to $s_{M/4}$ or $s_{3M/4}$ etc. This method can be improved further by choosing the values to compare to so that the probability for s to be above or below are roughly equal, again assuming a certain, simplified form of $\langle P(s; s_c)\rangle$. The expected number of comparisons in this scheme is of the order of $\ln(M)$. Usually, the (logarithmic) bin size s_b is fixed, so that M increases with s_c like $\ln(s_c)$, i. e. the expected number of comparisons is proportional to the double logarithm of s_c. Even though asymptotically

tree-search has a greater number of comparisons than the linear search discussed above, in practice, the double logarithms means that tree-search will almost always by far outperform a linear search. Similarly, even though direct calculation of the bin l using a function $\tilde{h}(s)$ has constant computational costs, they will typically be greater than for the tree search, in particular if the observable is discrete and provisions must be made in the region of its small values.

Coarse Graining The simplest binning scheme consists of dividing s by a constant, $h(s) = s/q$ so that the resulting range of hash values is small enough to fit in the memory available. In case of integer valued event sizes, the simplest and potentially the fasted way of determining the hash value is a bit shift, i. e. a division by a power of 2.

The key problem is the same as the initial motivation for binning, namely that entries for small s might overflow, while entries for large s might be very sparse. The problem is more serious in case of coarse graining if the factor q is so large that the majority of events arrives in the first bin. Moreover, not fully resolving the distribution for small s means that most of the non-universal part of the distribution is lost, potentially hiding problems of the behavior of the lower cutoff.

A very powerful alternative is a combination of coarse graining and exponential binning. A small number of thresholds is introduced and within each pair, a different

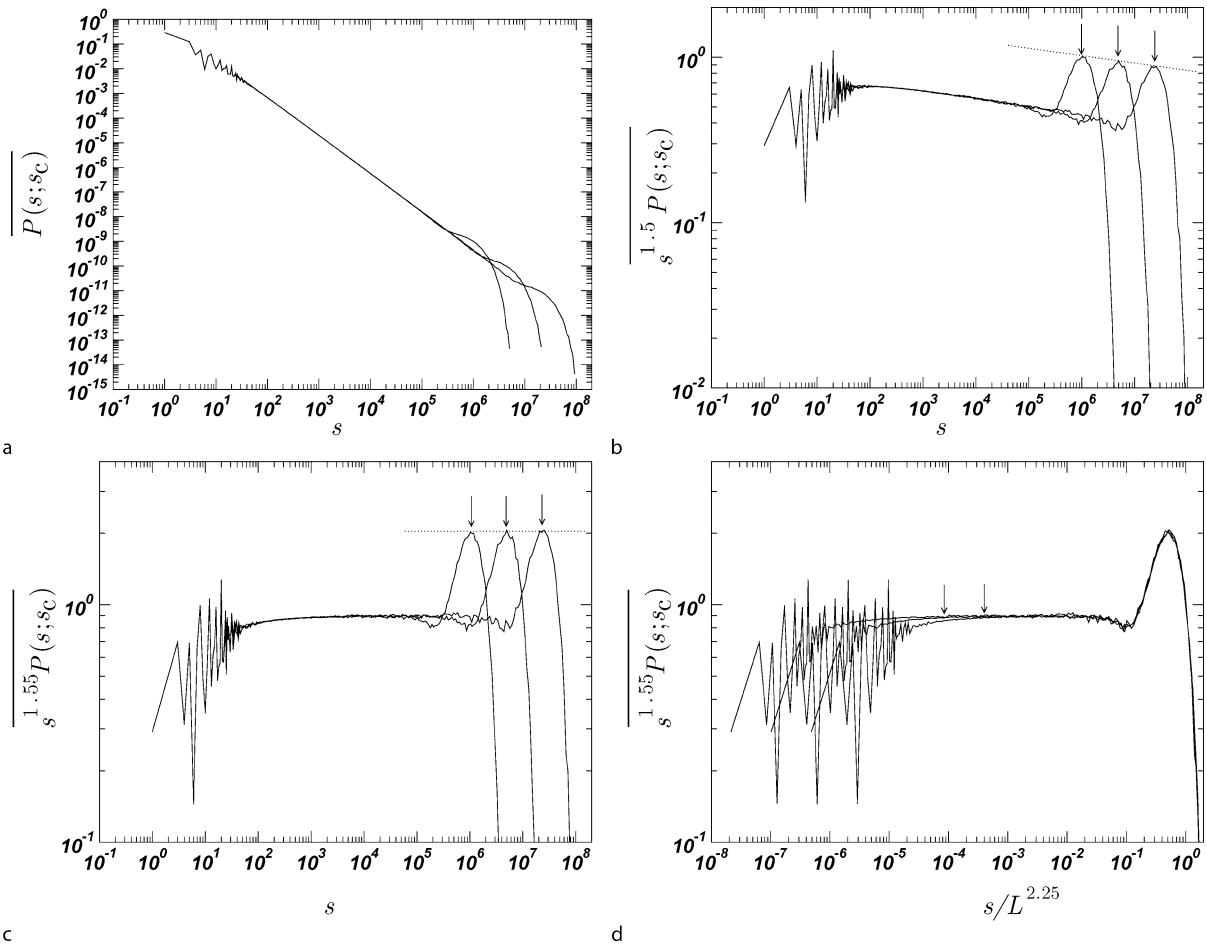

Probability Densities in Complex Systems, Measuring, Figure 6

A data collapse in detail. Using the data shown in Fig. 2, a shows the binned data. In b a preliminary estimate has been made for the scaling exponent τ and the data is plotted in the form $s^{1.5}\overline{P(s;s_C)}$. A "landmark" has been chosen (the maxima, indicated by *arrows*), which connected by a line (*dotted*) indicate that the exponent τ is to be chosen larger than the preliminary choice. In c $\tau = 1.55$ is chosen to make all maxima line up (*dotted line*). The position of the maxima on the abscissa (*arrows*) indicate the value of s_c. d shows the final collapse, after shifting the data horizontally by dividing s by the respective s_c. The approximate value for s_0, estimated for the two smaller system sizes as the point where the data approximately coincide is indicated by an *arrow*

rescaling q factor is used to implement different degrees of coarse graining. This method also allows the use of different variables types in different regions of the histogram, i. e. large types for small s and smaller less memory consuming ones for smaller values.

Binning During Data Analysis

To simplify visual inspection, in practice the raw histogram produced in the simulation is always binned afterwards. The most widely used method is exponential binning, as it provides equally spaced data points in a double-logarithmic plot.

To perform a data collapse, the data is plotted in the form $s^\tau \overline{P(s;s_c)}$ versus s/s_c. If Eq. (19) applies, the data for different s_c collapses on the same line, approximating $\mathcal{G}(s/s_c)$ up to an error due to **finite size corrections** and the presence of the non-universal part. An example is shown in Fig. 2. If Eq. (19) applies, the data collapses for any function $f(x)$ when plotted in the form $s^\tau f(s/s_c)\overline{P(s;s_c)}$ versus s/s_c and shows $\mathcal{G}(s/s_c)f(s/s_c)$. A particularly important case is $f(x) = x^{-\tau}$, which leads to $s_c^\tau \overline{P(s;s_c)}$ versus s/s_c. To expose as many details as possible of the assumed scaling function, it is generally advisable to choose $f(x)$ so that the ordinate of the resulting data spreads as little as possible (see also Fig. 5). If $\lim_{x\to 0} \mathcal{G}(x) \neq 0$, this is usually the case for $f(x) = 1$. If $\mathcal{G}(x)$ is expected to have a power-law dependence on the argument, $\mathcal{G}(x) = x^\alpha \tilde{\mathcal{G}}(x)$ with $\lim_{x\to 0} \tilde{\mathcal{G}}(x) \neq 0$, then $f(x) = x^{-\alpha}$ is the appropriate choice.

If the exponent τ and the values s_c are unknown, the first step to perform a data collapse is to collapse the binned data (Fig. 6a) approximately with some preliminary choices for τ and (Fig. 6b) possibly also for the different values of s_c and determine an outstanding feature in the resulting data, such as the maximum (arrows in Fig. 6b). A different choice for τ will change the relative vertical position of the "landmark", which is to be chosen so that all landmarks line up at the same value on the ordinate, Fig. 6c. Next, the s_c for the different simulations are chosen to collapse the data horizontally, usually by estimating s_c as the position of the landmark with respect to the abscissa, Fig. 6d. In this last figure the lower cutoff s_0 can be estimated as well, as the value of s from where on the data roughly coincide. A double-logarithmic plot of both cutoffs indicates roughly constant s_0 and a gap-exponent $D = 2.25$ for s_c. A data collapse often reveals that much greater system sizes are needed, as scaling applies to the *asymptotic* behavior only. The attempt of a data collapse for the BTW model, Fig. 1d, shows an example for a failed collapse in the case of small system sizes.

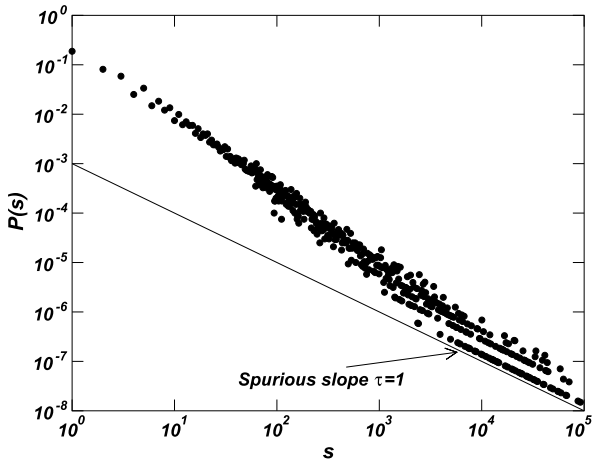

Probability Densities in Complex Systems, Measuring, Figure 7
If the sample is too sparse in the tail of the distribution, exponential binning produces a spurious slope $\tau = 1$

While binning greatly improves the visual quality of the data, it assumes that $\langle P(s;s_c)\rangle$ does not change too suddenly on the scale of the size of the bin, which therefore needs to be chosen small enough. However, choosing s_b too small results in too many bins and therefore too much statistical noise. For large s most of those will be empty, some will contain a single entry, some two etc. Because the width of the bin is $(s_b - 1)s_l$ the density within bins containing a single entry scales like s_l^{-1}, leading in a double-logarithmic plot to a sequence of points with slope -1. Parallel to this line, lies a set of points corresponding to bins with two entries etc. The resulting plot displays a spurious scaling exponent $\tau = 1$, see Fig. 7.

Future Directions

The numerical methods discussed above are commonly used in other areas of statistical physics and have been mainly developed there. That applies to the Monte Carlo methods as well as to the data analysis. Within complexity, the development of methods hardly forms a research branch in its own right. Most methods are developed by practitioners for the specific problem at hand.

As complexity is a very diverse field, only very few models have received so much attention that specialized algorithms have been developed and discussed broadly in the literature, for example [11,35]. Most models are defined by a dynamic process which normally makes a simple implementation readily available. Algorithmic improvements mainly focus on efficiency and resources, as both CPU time and memory requirements are the two key factors limiting the quality of results in terms of sample size

and system size respectively. In other areas, as for example in the context of networks, new observables often require new, efficient algorithms.

Even though almost all models are constrained by the computer hardware available and therefore benefit significantly from improved algorithms, optimization often clashes with the universality of the algorithm and, on a more technical level, with the elegance and readability of the code.

With respect to data analysis, some literature is concerned with new approaches to established observables, such as the moment analysis discussed in Subsect. "Moment Analysis" to derive the exponents of the PDF.

While algorithms and methods have advanced significantly, there seems to be a maximum amount of information about a complex system accessible through a given amount of CPU time spent in a computer simulation. The most significant general advancement in computational physics therefore comes from the continuous improvement of computer hardware available, producing an ever increasing sample size within a certain amount of CPU time. While this relieves the research from many constraints, it does not imply that sophisticated algorithms and well thought-out methods become less relevant. It is the combined effect of methods, software and hardware that moves the subject forward continuously. A good algorithm to cope with today's technical constraints will do an even better job in the years to come.

Bibliography

Primary Literature

1. Metropolis N, Ulam S (1949) The Monte Carlo Method. J Am Stat Ass 44:335–341
2. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of State Calculations by Fast Computing Machines. J Chem Phys 21:1087–1092
3. Landau DP, Binder K (2000) A Guide to Monte Carlo Simulations in Statistical Physics. Cambridge University Press, Cambridge
4. Bak P, Tang C, Wiesenfeld K (1987) Self-Organized Criticality: An Explanation of 1/f Noise. Phys Rev Lett 59:381–384
5. Bak P, Tang C, Wiesenfeld K (1988) Self-Organized Criticality. Phys Rev A 38:364–374
6. van Kampen NG (1992) Stochastic Processes in Physics and Chemistry, 3rd impression 2001, enlarged and revised edn. Elsevier, Amsterdam
7. Stauffer D, Aharony A (1994) Introduction to Percolation Theory. Taylor, London
8. Binney JJ, Dowrick NJ, Fisher AJ, Newman MEJ (1998) The Theory of Critical Phenomena. Clarendon Press, Oxford
9. Binder K, Heermann DW (1997) Monte Carlo Simulation in Statistical Physics. Springer, Berlin
10. Zia RKP, Schmittmann B (2007) Probability currents as principal characteristics in the statistical mechanics of non-equilibrium steady states. J Stat Mech P07012
11. de Oliveira MM, Dickman R (2005) How to simulate the quasistationary state. Phys Rev E 71:016129
12. Spitzer F (1970) Interaction of Markov processes. Adv Math 5:256–290
13. Bouchaud JP, Comtet A, Georges A, Le Doussal P (1990) Classical Diffusion of a Particle in a One-Dimensional Random Force Field. Ann Phys 201:285–341
14. Grassberger P (2002) Go with the winners: a general Monte Carlo strategy. Comp Phys Comm 147:64–70
15. Pradhan P, Dhar D (2006) Sampling rare fluctuations of height in the oslo ricepile model. J Phys A 40:2639–2650
16. Bouchaud JP, Potters M (2003) Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management. Cambridge University Press, Cambridge
17. De Menech M, Stella AL, Tebaldi C (1998) Rare events and breakdown of simple scaling in the Abelian sandpile model. Phys Rev E 58:R2677–R2680
18. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical Recipes in C, 2nd edn. Cambridge University Press, New York
19. Knuth DE (1997) The Art of Computer Programming, vol 3. Seminumerical Algorithms, 2nd edn. Addison Wesley, Massachusetts
20. Matsumoto M, Nishimura T (1998) Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudorandom Number Generator. ACM Trans Model Comp Sim 8:3–30
21. Gentle JE (1998) Random Number Generation and Monte Carlo Methods. Springer, Berlin
22. Ferrenberg AM, Landau DP (1992) Monte Carlo Simulations: Hidden Errors from "Good" Random Number Generators. Phys Rev Lett 69:3382–3384
23. Efron B (1982) The Jackknife, the Bootstrap and Other Resampling Plans. SIAM, Philadelphia
24. Berg BA (1992) Double Jackknife bias-corrected estimators. Comp Phys Com 69:7–14
25. Fisher ME (1967) The Theory of equilibrium critical phenomena. Rep Prog Phys 30:615–730
26. Privman V, Hohenberg PC, Aharony A (1991) Universal Critical-Point Amplitude Relations. In: Domb C, Lebowitz JL (eds) Phase Transitions and Critical Phenomena, vol 14. Academic Press, New York, pp 1–134
27. Christensen K, Corral A, Frette V, Feder J, Jøssang T (1996) Tracer Dispersion in a Self-Organized Critical System. Phys Rev Lett 77:107–110
28. Pruessner G, Jensen HJ (2002) Broken scaling in the forest-fire model. Phys Rev E 65:056707–1–8. Preprint cond-mat/0201306
29. Wegner FJ (1972) Corrections to Scaling Laws. Phys Rev B 5:4529–4536
30. Dowd K, Severance C (1998) High Performance Computing, 2nd edn. O'Reilly, Sebastopol
31. Pfeuty P, Toulouse G (1977) Introduction to the Renormalization Group and to Critical Phenomena. Wiley, Chichester
32. Ferrenberg AM, Landau DP, Binder K (1991) Statistical and Systematic Errors in Monte Carlo Sampling. J Stat Phys 63:867–882
33. Milchev A, Binder K, Heermann DW (1986) Fluctuations and Lack of Self-Averaging in the Kinetics of Domain Growth. Z Phys B 63:521–535

34. Christensen K, Moloney NR (2005) Complexity and Criticality. Imperial College Press, London
35. Pruessner G, Jensen HJ (2004) Efficient algorithm for the forest fire model. Phys Rev E 70:066707–1–25. Preprint cond-mat/0309173

Books and Reviews

Anderson TW (1964) The Statistical Analysis of Time Series. Wiley, London
Barenblatt GI (1996) Scaling, self-similarity, and intermediate asymptotics. Cambridge University Press, Cambridge
Berg BA (2004) Markov Chain Monte Carlo Simulations and Their Statistical Analysis. World Scientific, Singapore
Brandt S (1998) Data Analysis. Springer, Berlin
Jensen HJ (1998) Self-Organized Criticality. Cambridge University Press, New York
Liggett TM (2005) Stochastic Interacting Systems: Contact, Voter and Exclusion Processes. Springer, Berlin
Marro J, Dickman R (1999) Nonequilibrium Phase Transitions in Lattice Models. Cambridge University Press, New York
Newman MEJ, Barkema GT (1999) Monte Carlo Methods in Statistical Physics. Oxford University Press, New York
Stanley HE (1971) Introduction to Phase Transitions and Critical Phenomena. Oxford University Press, New York

Probability Distributions in Complex Systems

Didier Sornette
Department of Management, Technology and Economics, ETH Zurich, Zurich, Switzerland

Article Outline

Glossary
Definition of the Subject
Introduction
The Fascination with Power Laws
Testing for Power Law Distributions in your Data
Beyond Power Laws: "Kings"
Future Directions
Bibliography

Glossary

Complex system A system with a large number of mutually interacting parts, often open to its environment, which self-organizes its internal structure and its dynamics with novel and sometimes surprising macroscopic "emergent" properties.

Criticality (in physics) A state in which spontaneous fluctuations of the order parameter occur at all scales, leading to diverging correlation length and susceptibility of the system to external influences.

Power law distribution A specific family of statistical distribution appearing as a straight line in a log-log plot; exhibits the property of scale invariance and therefore does not possess characteristic scales.

Self-organized criticality Occurs when the system dynamics are attracted spontaneously, without any obvious need for parameter tuning to a critical state with infinite correlation length and power law statistics.

Stretched-exponential distribution A specific family of sub-exponential distribution interpolating smoothly between the exponential distribution and the power law family.

Definition of the Subject

This core article for the Encyclopedia of Complexity and System Science (Springer Science) reviews briefly the concepts underlying complex systems and probability distributions. The latter are often taken as the first quantitative characteristics of complex systems, allowing one to detect the possible occurrence of regularities providing a step toward defining a classification of the different levels of organization (the "universality classes"). A rapid survey covers the Gaussian law, the power law and the stretched exponential distributions. The fascination for power laws is then explained, starting from the statistical physics approach to critical phenomena, out-of-equilibrium phase transitions, self-organized criticality, and ending with a large, but not exhaustive, list of mechanisms leading to power law distributions. A checklist for testing and qualifying a power law distribution from data is described in seven steps. This essay enlarges the description of distributions by proposing that "kings", i. e., events even beyond the extrapolation of the power law tail, may reveal information which is complementary and perhaps sometimes even more important than the power law distribution. We conclude with a list of future directions.

Introduction

Complex Systems

The study of out-of-equilibrium dynamics (e. g. dynamical phase transitions) and of heterogeneous systems (e. g., spin-glasses) has progressively made popular in physics the concept of complex systems and the importance of systemic approaches: systems with a large number of mutually interacting parts, often open to their environment, self-organize their internal structure and their dynamics with novel and sometimes surprising macroscopic ("emergent") properties. The complex system approach,

which involves "seeing" inter-connections and relationships, i. e., the whole picture as well as the component parts, has become pervasive in modern control of engineering devices and business management. It also plays an increasing role in most of the scientific disciplines, including biology (biological networks, ecology, evolution, origin of life, immunology, neurobiology, molecular biology, etc.), geology (plate-tectonics, earthquakes and volcanoes, erosion and landscapes, climate and weather, environment, etc.), economics and social sciences (cognition, distributed learning, interacting agents, etc.). There is a growing recognition that progress in most of these disciplines, in many of the pressing issues for our future welfare as well as for the management of our everyday life, will need such a systemic complex system and multidisciplinary approach.

A central property of a complex system is the possible occurrence of coherent large-scale collective behaviors with very rich structure, resulting from the repeated nonlinear interactions among its constituents: the whole turns out to be much more than the sum of its parts. Most complex systems around us exhibit rare and sudden transitions that occur over time intervals which are short compared to the characteristic time scales of their prior evolution. Such extreme events express more than anything else the underlying "forces" usually hidden by almost perfect balance and thus provide the potential for a better scientific understanding of complex systems. These crises have fundamental societal impacts and range from large natural catastrophes such as earthquakes, volcanic eruptions, hurricanes and tornadoes, landslides, avalanches, lightning strikes, and catastrophic events of environmental degradation, to the failure of engineering structures, crashes in the stock market, social unrest leading to large-scale strikes and upheaval, economic drawdowns on national and global scales, regional power blackouts, traffic gridlock, diseases and epidemics, etc.

Given the complex dynamics of these systems, a first standard attempt to quantify and classify the characteristics and the possible different regimes consists of

1. identifying discrete events,
2. measuring their sizes,
3. constructing their probability distribution.

The interest in probability distributions in complex systems has the following roots.

- They offer a natural metric of the relative rate of occurrence of small versus large events, and thus of the associated risks.
- As such, they constitute essential components of risk assessment and prerequisites of risk management.
- Their mathematical form can provide constraints and guidelines to identify the underlying mechanisms at their origin and thus at the origin of the behavior of the complex system under study.
- This improved understanding may lead to better forecasting skills, and even to the option (or illusion (?)) of (a certain degree of) control [1,2].

Probability Distributions

Let us first establish some notation and vocabulary. Consider a process X whose outcome is a real number. The probability density function $P(x)$ of X (also called probability distribution or pdf) is such that the probability that X is found in a small interval Δx around x is $P(x)\Delta x$. The probability that X is between a and b is therefore given by the integral of $P(x)$ between a and b:

$$\mathcal{P}(a < X < b) = \int_a^b P(x)\mathrm{d}x \,. \tag{1}$$

The pdf $P(x)$ depends on the units used to quantify the variable x and has the dimension of the inverse of x, such that $P(x)\Delta x$, being a probability, i. e., a number between 0 and 1, is dimensionless. In a change of variable, say $x \to y = f(x)$, the probability is invariant. Thus, the invariant quantity is the probability $P(x)\Delta x$ and not the pdf $P(x)$. We thus have

$$P(x)\Delta x = P(y)\Delta y \,, \tag{2}$$

leading to $P(y) = P(x)|\mathrm{d}f/\mathrm{d}x|^{-1}$, taking the limit of infinitesimal intervals. By definition, $P(x) \geq 0$. It is normalized, $\int_{x_{\min}}^{x_{\max}} P(x)\mathrm{d}x = 1$, where $x_{\min}$ and $x_{\max}$ (often $\pm\infty$) are the smallest and largest possible values for x, respectively.

The empirical estimation of the pdf $P(x)$ is usually plotted with the horizontal axis scaled as a graded series for the measure under consideration (the magnitude of the earthquakes, etc.) and the vertical axis scaled for the number of outcomes or measures in each interval of horizontal value (the number of earthquakes of magnitude between 1 and 2, between 2 and 3, etc.). This implies a "binning" into small intervals. If the data is sparse, the number of events in each bin becomes small and can fluctuate, leading to a poor representation of the data. In this case, it is useful to construct the cumulative distribution $P_{\leq}(x)$ defined by

$$\mathcal{P}_{\leq}(x) = \mathcal{P}(X \leq x) = \int_{-\infty}^{x} P(y)\mathrm{d}y \,, \tag{3}$$

which is much less sensitive to fluctuations. $\mathcal{P}_{\leq}(x)$ gives the fraction of events with values less than or equal to x. $\mathcal{P}_{\leq}(x)$ increases monotonically with x from 0 to 1. Similarly, we can define the so-called complementary cumulative (or survivor) distribution $\mathcal{P}_{>}(x) = 1 - \mathcal{P}_{\leq}(x)$.

For random variables which take only discrete values $x_1, x_2, \ldots, x_n$, the pdf is made of a discrete sum of Dirac functions $(1/n)[\delta(x - x_1) + \delta(x - x_2) + \cdots + \delta(x - x_n)]$. The corresponding cumulative distribution function (cdf) $\mathcal{P}_{\leq}(x)$ is a staircase. There are also more complex distributions made of a continuous cdf but which are singular with respect to the Lebesgue measure dx. An example is the Cantor distribution constructed from the Cantor set (see for instance Chap. 5 in [3]). Such a singular cdf is continuous but has its derivative which is zero almost everywhere: the pdf does not exist (see e. g. [4]).

Brief Survey of Probability Distributions

Statistical physics is rich with probability distributions. The most famous is the Boltzmann distribution, which describes the probability that the configuration of a system in thermal equilibrium has a given energy. Its extension to out-of-equilibrium systems is the subject of intense scrutiny [5]; see also Chap. 7 in [3] and references therein. Special cases include the Maxwell–Boltzmann distribution, the Bose–Einstein distribution and the Fermi–Dirac distribution.

In the quest to characterize complex systems, two distributions have played a leading role: the normal (or Gaussian) distribution and the power law distribution. The Gaussian distribution is the paradigm of the "mild" family of distributions. In contrast, the power law distribution is the representative of the "wild" family. The contrast between "mild" and "wild" is illustrated by the following questions:

- What is the probability that someone has twice your height? Essentially zero! The height, weight and many other variables are distributed with "mild" pdfs with a well-defined typical value and relatively small variations around it. The Gaussian law is the archetype of "mild" distributions.
- What is the probability that someone has twice your wealth? The answer of course depends somewhat on your wealth but in general, there is a non-vanishing fraction of the population twice, ten times or even one hundred times as wealthy as you are. This was noticed at the end of the 19th century by Pareto, after whom the Pareto law has been named, which describes the power law distribution of wealth [6,7], a typical example of "wild" distributions.

The Normal (or Gaussian) Distribution The expression of the Gaussian probability density function of a random variable x with mean x_0 and standard deviation σ reads

$$P_G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - x_0)^2}{2\sigma^2}\right) \quad \text{defined for } -\infty < x < +\infty\,. \tag{4}$$

The importance of the normal distribution as a model of quantitative phenomena in the natural and social sciences can be in large part attributed to the central limit theorem. Many measurements of physical as well as social phenomena can be well approximated by the normal distribution. While the mechanisms underlying these phenomena are often unknown, the use of the normal model can be theoretically justified by assuming that many small, independent effects additively contribute to each observation. The Gaussian distribution is also justified as the most parsimonious choice in absence of information other than just the mean and the variance: it maximizes the information entropy among all distributions with known mean and variance. As a result of the central limit theorem, the normal distribution is the most widely used family of distributions in statistics and many statistical tests are based on the assumption of asymptotic normality of the data. In probability theory, the standard Gaussian distribution arises as the limiting distribution of a large class of distributions of random variables (with suitable centering and normalization) characterized by a finite variance, which is nothing but the statement of the central limit theorem (see Chapter 2 in [3]).

At the beginning of the twenty-first century, when power laws are often taken as the hallmark of complexity, it is interesting to reflect on the fact that the previous giants of science in the eighteenth and nineteenth centuries (Halley, Laplace, Quetelet, Maxwell and so on) considered that the Gaussian distribution expressed a kind of universal law of nature and of society. In particular, the Belgian astronomer Adolphe Quetelet was instrumental in popularizing the statistical regularities discovered by Laplace in the frame of the Gaussian distribution, which influenced the likes of John Herschel and John Stuart Mill and led Comte to define the concept of "social physics."

The Power Law Distribution A probability distribution function $P(x)$ exhibiting a power law tail is such that

$$P(x) \propto \frac{C_\mu}{x^{1+\mu}}, \quad \text{for } x \text{ large}\,, \tag{5}$$

possibly up to some large limiting cut-off. The exponent μ (also referred to as the "index") characterizes the nature of

the tail: for $\mu < 2$, one speaks of a "heavy tail" for which the variance is theoretically not defined. For power laws, the scale factor C_μ plays a role analogous to the role of variance in Gaussian distributions (see Chapter 4 in [3]). In particular, it enjoys the additivity property: the scale factor of the distribution of the sum of several independent random variables, each with a distribution exhibiting a power law tail with the same exponent μ, is equal to the sum of the scale factors characterizing each distribution of each random variable in the sum.

A more general form is

$$P(x) \propto \frac{L(x)}{x^{1+\mu}}, \quad \text{for } x \text{ large}, \tag{6}$$

where $L(x)$ is a slowly varying function defined by $\lim_{x\to\infty} L(tx)/L(x) = 1$ for any finite t (typically, $L(x)$ is a logarithm $\ln(x)$ or power of a logarithm such as $(\ln(x))^n$ with n finite). In mathematical language, a function such as (6) is said to be "regularly varying". This more general form means that the power law regime is an asymptotic statement holding only as a better and better approximation as one considers larger and larger x values.

Power laws obey the symmetry of scale invariance, that is, they verify the following defining property that, for an arbitrary real number λ, there exists a real number γ such that

$$P(x) = \gamma P(\lambda x), \quad \forall x. \tag{7}$$

Obviously, $\gamma = \lambda^{1+\mu}$. The relation (7) means that the ratio of the probabilities of occurrence of two sizes x_1 and x_2 depends only on their ratio x_1/x_2 and not on their absolute values. For instance, according to the Zipf law ($\mu = 1$) for the distribution of city sizes, the ratio of the number of cities with more that 1 million inhabitants to those with more than 100,000 persons is the same as the ratio of the number of cities with more than 100,000 inhabitants to those with more than 10,000 persons, both ratios being equal to 1/10. The symmetry of scale invariance (7) extends to the space of functions the concept of scale invariance which characterizes fractal geometric objects.

It should be stressed that, when they exhibit a power-law-like shape, most empirical distributions do so only over a finite range of event sizes, either bounded between a lower and an upper cut-off [8,9,10,11], or above a lower threshold, i. e., only in the tail of the observed distribution [12,13,14,15]. Power law distributions and, more generally, regularly varying distributions remain robust functional forms under a large number of operations, such as linear combinations, products, minima, maxima, order statistics, and powers, which may also explain their ubiquity and attractiveness. Jessen and Mikosch [16] give the conditions under which transformations of power law distributions are also regularly varying, possibly with a different exponent (see also Sect. 4.4 in [3] for an heuristic presentation of similar results).

The Stretched Exponential Distribution The so-called stretched exponential (SE) distribution has been found to be a versatile intermediate distribution interpolating between "thin tail" (Gaussian, exponential, ...) and very "fat tail" distributions. In particular, Laherrère and Sornette [17] have found that several examples of fat-tailed distributions in the natural and social sciences, often considered to be good examples of power laws, could sometimes be represented as well as or even better by an SE distribution. Malevergne et al. [18] present systematic statistical tests comparing the SE family with the power law distribution in the context of financial return distributions. The SE family is defined by the following expression for the survival distribution (also called the complementary cumulative distribution function):

$$\mathcal{P}_{\geq u}(x) = 1 - \exp\left[-\left(\frac{x}{d}\right)^c + \left(\frac{u}{d}\right)^c\right], \quad \text{for } x \geq u. \tag{8}$$

The constant u is a lower threshold that can be changed to increasingly emphasize the tail of the distribution as u is increased. The structural exponent c controls the "thin" versus "heavy" nature of the tail.

1. For $c = 2$, the SE distribution (8) has the same asymptotic tail as the Gaussian distribution.
2. For $c = 1$, expression (8) recovers the pure exponential distribution.
3. For $c < 1$, the tail of $\mathcal{P}_u(x)$ is fatter than an exponential, and corresponds to the regime of sub-exponentials (see Chapter 6 in [3]).
4. For $c \to 0$ with
$$c \cdot \left(\frac{u}{d}\right)^c \to \mu, \tag{9}$$
the SE distribution converges to the Pareto distribution with tail exponent μ.

Indeed, we can write

$$\begin{aligned}
&\frac{c}{d^c} \cdot x^{c-1} \cdot \exp\left(-\frac{x^c - u^c}{d^c}\right) \\
&= c\left(\frac{u}{d}\right)^c \cdot \frac{x^{c-1}}{u^c} \exp\left[-\left(\frac{u}{d}\right)^c \cdot \left(\left(\frac{x}{u}\right)^c - 1\right)\right], \\
&\simeq \mu \cdot x^{-1} \exp\left[-c\left(\frac{u}{d}\right)^c \cdot \ln\frac{x}{u}\right], \quad \text{as } c \to 0 \\
&\simeq \mu \cdot x^{-1} \exp\left[-\mu \cdot \ln\frac{x}{u}\right], \\
&\simeq \mu \frac{u^\mu}{x^{\mu+1}},
\end{aligned} \tag{10}$$

which is the pdf of the Pareto power law model with tail index μ. This implies that, as $c \to 0$, the characteristic scale d of the SE model must also go to zero with c to ensure its convergence towards the Pareto distribution.

This shows that the Pareto model can be approximated with any desired accuracy on an arbitrary interval $(u > 0, U)$ by the (SE) model with parameters (c, d) satisfying Eq. (9) where the arrow is replaced by an equality. The limit $c \to 0$ provides any desired approximation to the Pareto distribution uniformly on any finite interval (u, U). This deep relationship between the SE and power law models allows us to understand why it can be very difficult to decide, on a statistical basis, which of these models best fits the data [17,18]. This insight can be made rigorous to develop a formal statistical test of the (SE) hypothesis versus the Pareto hypothesis [18,19].

From a theoretical viewpoint, this class of distributions (8) is motivated in part by the fact that large deviations of multiplicative processes are generically distributed with stretched exponential distributions [20]. Stretched exponential distributions are also parsimonious examples of the important subset of sub-exponentials, that is, of the general class of distributions decaying slower than an exponential [21]. This class of sub-exponentials share several important properties of heavy-tailed distributions [22] not shared by exponentials or distributions decreasing faster than exponentials: for instance, they have "fat tails" in the sense of the asymptotic probability weight of the maximum compared with the sum of large samples [4] (see also Chaps 1 and 6 in [3]).

Notwithstanding their fat-tailness, stretched exponential distributions have only finite moments, in contrast with regularly varying distributions for which moments of order equal to or larger than the tail index μ are not defined. However, they do not admit an exponential moment, which leads to problems in the reconstruction of the distribution from the knowledge of their moments [23]. In addition, the existence of all moments is an important property allowing for an efficient estimation of any high-order moment, since it ensures that the estimators are asymptotically Gaussian. In particular, for stretched-exponentially distributed random variables, the variance, skewness and kurtosis can be accurately estimated, contrarily to random variables with regularly varying distribution with tail index smaller than about 5.

The Fascination with Power Laws

Probability distribution functions with a power law dependence in terms of event or object sizes seem to be ubiquitous statistical features of natural and social systems. It has repeatedly been argued that such an observation relies on an underlying self-organizing mechanism, and therefore power laws should be considered as the statistical imprints of complex systems. It is often claimed that the observation of a power law relation in data often points to specific kinds of mechanisms at its origin, that can often suggest a deep connection with other, seemingly unrelated systems. In complex systems, the appearance of power law distributions is often thought to be the signature of hierarchy and robustness. In the last two decades, such claims have been made, for instance, for earthquakes, weather and climate changes, solar flares, the fossil record, and many other systems, to promote the relevance of self-organized criticality as an underlying mechanism for the organization of complex systems [24]. This claim is often unwarranted as there are many non-self-organizing mechanisms that produce power law distributions [3,25,26,27].

Research on the origins of power law relations and efforts to observe and validate them in the real world are extremely active in many fields of modern science, including physics, geophysics, biology, medical sciences, computer science, linguistics, sociology and economics. This section briefly summarizes the present understanding.

Statistical Physics in General and the Theory of Critical Phenomena

The study of critical phenomena in statistical physics suggests that power laws emerge close to special critical or bifurcation points separating two different phases or regimes of the system. In systems at thermodynamic equilibrium modeled by general spin models, renormalization group theory [28] has demonstrated the existence of universality, so that diverse systems exhibit the same critical exponents and identical scaling behavior as they approach criticality, i. e., they share the same fundamental macroscopic properties. For instance, the behavior of water and CO_2 at their boiling points at a certain critical pressure and that of a magnet at its Curie point fall in the same universality class because they can be characterized by order parameter s with the same symmetries and dimensions in the same space dimension. In fact, almost all material phase transitions are described by a small set of universality classes.

From this perspective, the fascination with power laws reflects the fact that they characterize the many coexisting and delicately interacting scales at a critical point. The existence of many scales leading to complex geometrical properties is often associated with fractals [12]. While it is true that critical points and fractals share power law relations, power law relations and power law distributions

are not the same. The latter, which is the subject of this essay, describes the probability density function or frequency of occurrence of objects or events, such as the frequency of earthquakes of a given magnitude range. In contrast, power law relations between two variables (such as the magnetization and temperature in the case of the Curie point of a magnet) describe a functional abstraction belonging to or characteristic of these two variables. Both power law relations and power law distributions can result from the existence of a critical point. A simple example in percolation is (i) the power law dependence of the size of the larger cluster as a function of the distance from the percolation threshold and (ii) the power law distribution of cluster sizes at the percolation threshold [29].

Out-of-Equilibrium Phase Transition and Self-Organized Critical Systems (SOC)

In the broadest sense, self-organized criticality (SOC) refers to the spontaneous organization of a system driven from the outside into a globally stationary state, which is characterized by self-similar distributions of event sizes and fractal geometrical properties. This stationary state is dynamical in nature and is characterized by statistical fluctuations, which are generically referred to as "avalanches."

The term "self-organized criticality" contains two parts. The word "criticality" refers to the state of a system at a critical point at which the correlation length and the susceptibility become infinite in the infinite size limit as in the preceding section. The label "self-organized" is often applied indiscriminately to pattern formation among many interacting elements. The concept is that the structuration, the patterns and large scale organization appear spontaneously. The notion of self-organization refers to the absence of control parameters.

In this class of mechanisms, where the critical point is the attractor, the situation becomes more complicated as the number of universality classes proliferates. In particular, it is not generally well-understood why sometimes local details of the dynamics may change the macroscopic properties completely while in other cases, the universality class is robust. In fact, the more we learn about complex out-of-equilibrium systems, the more we realize that the concept of universality developed for critical phenomena at equilibrium has to be enlarged to embody a more qualitative meaning: the critical exponents defining the universality classes are often very sensitive to many (but not all) details of the models [30].

Of course, one of the hailed hallmarks of SOC is the existence of power law distributions of "avalanches" and of other quantities [3,24,31].

Non-exhaustive List of Mechanisms Leading to Power Law Distributions

There are many physical and mathematical mechanisms that generate power law distributions and self-similar behavior. Understanding how a mechanism is affected by microscopic laws constitutes an active field of research. We can propose the following non-exhaustive list of mechanisms that have been found to operate in different complex systems, and which can lead to power law distribution of avalanches or cluster sizes. For most of these mechanisms, we refer the reader to Chaps. 14 and 15 in [3] and to [27,32] for detailed explanations and the relevant bibliography. However, some of the mechanisms mentioned here have not been reviewed in these three references and are thus new to the list developed in particular in [3]. We should also stress that some of the mechanisms in this list are actually different incarnations of the same underlying idea (for instance preferential attachment which is a re-discovery of the Yule process, see [33] for an informative historical account).

1. percolation, fragmentation and other related processes,
2. directed percolation and its universality class of so-called "contact processes,"
3. cracking noise and avalanches resulting from the competition between frozen disorder and local interactions, as exemplified in the random field Ising model, where avalanches result from hysteretic loops [34],
4. random walks and their properties associated with their first passage statistics [35] in homogeneous as well as in random landscapes,
5. flashing annihilation in Verhulst kinetics [36],
6. sweeping of a control parameter towards an instability [25,37],
7. proportional growth by multiplicative noise with constraints (the Kesten process [38] and its generalization, for instance in terms of generalized Lotka–Volterra processes [39]), whose ancestry can be traced to Simon and Yule,
8. competition between multiplicative noise and birth-death processes [40],
9. growth by preferential attachment [32],
10. exponential deterministic growth with random times of observations (which gives the Zipf law) [41],
11. constrained optimization with power law constraints (HOT for "highly optimized tolerant"),
12. control algorithms, which employ optimal parameter estimation based on past observations, shown to generate broad power law distributions of fluctuations

and of their corresponding corrections in the control process [42,43],

13. on-off intermittency as a mechanism for power law pdf of laminar phases [44,45],
14. self-organized criticality which comes in many flavors as explained in Chapter 15 of [3]:
    - cellular automata sandpiles with and without conservation laws,
    - systems made of coupled elements with threshold dynamics,
    - critical de-synchronization of coupled oscillators of relaxation,
    - nonlinear feedback of the order parameter onto the control parameter
    - generic scale invariance,
    - mapping onto a critical point,
    - extremal dynamics.

If there is one lesson to extract from this impressive list it is that, when observing an approximate linear trend in the log-log plot of some data distribution, one should refrain from jumping to hasty conclusions on the implications of this approximate power law behavior. Another lesson is that power laws appear to be so ubiquitous perhaps because many roads lead to them!

Testing for Power Law Distributions in your Data

Although power law distributions are attractive for their simplicity (they are straight lines on log-log plots) and may be justified from theoretical reasons as discussed above, demonstrating that data do indeed follow a power law distribution requires more than simple fitting. Indeed, several alternative functional forms can appear to follow a power law form over some extent, such as stretched exponentials and log-normal distributions. Thus, validating that a given distribution is a power law is not easy and there is no silver bullet.

Clauset et al. [46] have recently summarized some statistical techniques for making accurate parameter estimates for power-law distributions based on maximum likelihood methods and the Kolmogorov–Smirnov statistic. They illustrate these statistical methods on 24 real-world data sets from a range of different disciplines. In some cases, they find that power laws are consistent with the data while in others the power law is ruled out. The log-likelihood ratio that they propose is however not warranted for non-nested models [47]

Here, we offer some advice for the characterization of a power law distribution as a possible adequate representation of a given data set. We emphasize good sense and practical aspects.

1. **Survivor distribution** First, the survival distribution should be constructed using raw data by ranking the values in increasing order. Then, rank versus values gives immediately a non-normalized survival distribution. The advantage of this construction is that it does not require binning or kernel estimation, which is a delicate art, as we have alluded to.
2. **Probability density function** The previous construction of the complementary cumulative (or survivor) distribution function should be complemented with that of the density function. Indeed, it is well-known that the cumulative distribution, being a "cumulative" integral of the density function as its name indicates, may be contaminated by disturbances at one end of the density function, leading to rather long cross-overs that may hide or perturb the power law. For instance, if the generating density distribution is a power law truncated by an exponential, as found for critical systems not exactly at their critical point or in the presence of finite-size effects [48], the power law part of the cumulative distribution will be strongly distorted leading to a spurious estimation of the exponent μ. This problem can be in large part alleviated by constructing the pdf using binning or, even better, kernel methods (see the very readable article [49] and references therein). By testing and comparing the survival and the probability density distributions, one obtains either a confirmation of power law scaling or an understanding of the origin(s) of the deviations from the power law.
3. **Structural analysis by visual inspection** Given that these first two steps have been performed, we recommend a preliminary visual exploration by plotting the survival and density distributions in (i) linear-linear coordinates, (ii) log-linear coordinates (linear abscissa and logarithmic ordinate) and (iii) log-log coordinates (logarithmic abscissa and logarithmic ordinate). The visual comparison between these three plots provides a fast and intuitive view of the nature of the data.
    - A power law distribution will appear as a convex curve in the linear-linear and log-linear plots and as a straight line in the log-log plot.
    - A Gaussian distribution will appear as a bell-shaped curve in the linear-linear plot, as an inverted parabola in the log-linear plot and as a strongly concave sharply falling curve in the log-log plot.
    - An exponential distribution will appear as a convex curve in the linear-linear plot, as a straight line in the log-linear plot and as a concave curve in the log-log plot.

    Having in mind the shape of these three reference distributions in these three representations provides fast

and useful reference points to classify the unknown distribution under study. For instance, if the log-linear plot shows a convex shape (upward curvature), we can conclude that the distribution has a tail fatter than an exponential. Then, the log-log plot will confirm if a power law is a reasonable description. If the log-log plot shows a downward curvature (concave shape), together with the information that the log-linear plot shows a convex shape, we can conclude that the distribution has a tail fatter than an exponential but thinner than a power law. For example, it could be a gamma distribution ($\sim x^n \exp[-x/x_0]$ with $n > 0$) or a stretched distribution (expression (8) with $c < 1$). Only more detailed quantitative analysis will allow one to refine the diagnostic, often with less-than-definite conclusions (see as an illustration the detailed statistical analysis comparing the power law to the stretched exponential distributions to describe the distribution of financial returns [18]).

The deviations from linearity in the log-log plot suggest the boundaries within which the power law regime holds. We say "suggest," as a visual inspection is only a first step which can be actually misleading. While we recommend a first visual inspection, it is only a first indication, not a proof. It is a necessary step to convince oneself (and the reviewers and journal editors) but certainly not a sufficient condition. It is a standard rule of thumb that power law scaling is thought to be meaningful if it holds over at least two to three decades on both axes and is bracketed by deviations on both sides whose origins can be understood (for instance, due to insufficient sampling and/or finite-size effects).

As an illustration of the potential errors stemming from visual inspection, we refer to the discussion of Sornette et al. [50], on the claim of Pacheco et al. [51] of the existence of a break in the Gutenberg–Richter distribution of earthquake magnitudes at $m = 6.4$ for California. This break was claimed to reveal the finiteness of the crust thickness according to Pacheco et al. [51]. This claim has subsequently been shown to be unsubstantiated, as the Gutenberg–Richter law (which is a power law when expressed in earthquake energies or seismic moments) seems to remain valid up to magnitudes of 7.5 in California and up to magnitude about 8–8.5 worldwide. This visual break at $m = 6.4$ turned out to be just a statistical deviation, completely expected from the nature of power law fluctuations [15,52].

4. **OLS fitting** The next step is often to perform an ordinary least-square (OLS) regression of the data (survival distribution or kernel-reconstructed density) on the logarithm of the variables, in order to estimate the parameters of the power law. These parameters are the exponent μ, the scale factor C_μ and possibly an upper threshold or other parameters controlling the crossover to other behaviors outside the scaling regime. Using logarithms ensures that all the terms in the sum of squares over the different data points contribute approximately similarly in the OLS. Otherwise, without logarithms, given the large range of values spanned by a typical power law distribution, a relative error of say 1% around a value of the order of 10^4 would have a weight in the sum ten thousand times larger than the weight due to the same relative error of 1% around a value of the order of 10^2, biasing the estimation of the parameters towards fitting preferentially the large values. In addition, in logarithm units, the estimation of the exponent μ of a power law constitutes a linear problem which is solved analytically. When performing the OLS estimation on the survival distribution, it is optimal to shift the ranks by 1/2 [53]. With this improvement, the OLS method is typically more robust to deviations from a pure power law form than the Hill estimator discussed below.
5. **Maximum likelihood estimation** Using an OLS method to estimate the parameters of a power law assumes implicitly that the distribution of the deviations from the power law (actually the difference between the logarithm of the data and the logarithm of the power law distribution) are normally distributed. This may not be a suitable approximation. An estimation which removes this assumption consists in using the likelihood method, in which the parameters of the power law are chosen so as to maximize the likelihood function. When the data points are independent, the likelihood function is nothing but the product $\prod_{i=1}^{N} P(x_i)$ over the N data points $x_1, x_2, \ldots, x_N$ of the power law distribution $P(x)$. In this case, the exponent μ which maximizes this likelihood (or equivalently and more conveniently its logarithm called the log-likelihood) is called the Hill estimator [54]. It reads

$$\frac{1}{\mu} = \frac{1}{n} \sum \ln \left[\frac{x_j}{x_{\min}} \right] , \tag{11}$$

where $x_{\min}$ is the smallest value among the n values used in the data set for the estimation of μ. Since power laws are often asymptotic characteristics of the tail, it is appropriate not to use the full data set but only the upper tail with data values above a lower threshold. Then, plotting $1/\mu$ or μ as a function of the lower threshold usually provides a good sense of the existence of a power law regime: one should expect an approximate stability of $1/\mu$ over some scaling regime. Note

that the Hill estimator provides an unbiased estimate of $1/\mu$ while μ obtained by inverting $1/\mu$ is slightly biased (see e. g., Chapter 6 in [3]). We refer to [55,56] for improved versions and procedures of the Hill estimator which deal with finite ranges and dependence.

6. **Non-parametric methods** Methods testing for a power law behavior in a given empirical distribution which are not parametric and not sensitive to the value of the exponent provide useful complements of the above fitting and parametric estimation approaches. Pisarenko et al. [57] and Pisarenko and Sornette [58] have developed new statistics such that a power law behavior is associated with a zero value of the statistics *independently of the numerical value of the exponent* μ and with a non-zero value otherwise. Plotting these statistics as a function of the lower threshold of the data sample allows one to detect subtle deviations from a pure power law. Lasocki [59] and Lasocki and Papadimitriou [60] have developed another non-parametric approach to detect deviations from a power law, the smoothed bootstrap test for multimodality, which makes it possible to test the complexity of the distribution without specifying any particular probabilistic model. The method relies on testing the hypotheses that the number of modes or the number of bumps exhibited by the distribution function equal 1. Rejection of one of these hypotheses indicate that the distribution has more complexity than described by a simple power law.

Once the evidence for a power law distribution has been reasonably demonstrated, the most difficult task remains: finding a mechanism and model which can explain the data. Note that the term "explain" refers to different meanings depending on the expert you are speaking to. For a statistician, having been unable to reject the power law function (5) given the data amounts to saying that the power law model "explains" the data. The emphasis of the statistician will be on refining parametric and non-parametric procedures to test the way the power law "fits" or deviates from the empirical data. In contrast, a physicist or a natural scientist sees this as only a first step, and attributes the word "explain" to the stage where a mechanism described in terms of a more fundamental process or first principles can derive the power law. But even among natural scientists, there is no consensus on what is a suitable "explanation." The reason stems from the different cultures and levels of study in different fields, well addressed in the famous paper "More is different" by Anderson [61]: a suitable explanation for a physicist will frustrate a chemist whose explanation, in turn, will not satisfy a biologist. Each scientific discipline's concepts are anchored in its characteristic fundamental scientific level, which provides the underpinning for the next scientific level of description (think for instance of the hierarchy: physics → chemistry → molecular biology → cell biology → animal biology → ethology → sociology → economics → ...).

Once a model at a given scientific description level has been proposed, the action of the model on inputs gives outputs which are compared with the data. Verifying that the model, inspired by the preliminary power law evidence, adequately fits this power law is a first step. Unfortunately, much too often, scientists stop there and are happy to report that they have a model that fits their empirical power law data. This is not good science. Keeping in mind the many possible mechanisms at the origin of power law distributions reviewed above, a correct procedure is to run the candidate model to get other predictions that can themselves be put to the test. This validation is essential to determine the degree to which the model is an accurate representation of the real world from the perspective of its intended uses. Reviewing a large body of literature devoted to the problem of validation, Sornette et al. [62] have proposed a synthesis in which the validation of a given model is formulated as an iterative construction process that mimics the often implicit process occurring in the minds of scientists. Validation is nothing but the progressive build-up of trust in the model, based on testing the model against non-redundant novel experiments or data, that allows one to make a decision and act on that basis. The applications of the validation program to a cellular automaton model for earthquakes, to a multifractal random walk model for financial time series, to an anomalous diffusion model for solar radiation transport in the cloudy atmosphere, and to a computational fluid dynamics code for the Richtmyer–Meshkov instability, exemplify the importance of going beyond the simple qualification of a power law.

Beyond Power Laws: "Kings"

The Standard View

Power law distributions embody the notion that extreme events are not exceptional 9-sigma events (to refer to the terminology using the Gaussian bell curve and its standard deviation σ as the metric to quantify deviations from the mean). Instead, extreme events should be considered as rather frequent and part of the same organization as other events. In this view, a great earthquake is just an earthquake that started small... and did not stop; it is inherently unpredictable due to its sharing of all the properties and characteristics of smaller events (except for its size),

so that no genuinely informative precursor can be identified [63]. This is the view expounded by Bak and coworkers in their formulation of the concept of self-organized criticality [24,64]. In the following, we outline several promising directions of research that expand on these ideas.

Self-Organized Criticality Versus Criticality

However, there are many suggestions that inherent unpredictability does not need to be the case. One argument is that criticality and self-organized criticality (SOC) can actually co-exist. The hallmark of criticality is the existence of specific precursory patterns (increasing susceptibility and correlation length) in space and time. Continuing with the example of earthquakes, the idea that a great earthquake could result from a critical phenomenon has been put forward by different groups, starting almost three decades ago [65,66,67]. Attempts to link earthquakes and critical phenomena find support in the evidence that rupture in heterogeneous media is similar to a critical phenomenon (see Chapter 13 of [3] and references therein). Also indicative is the often-reported observation of increased intermediate magnitude seismicity before large events [68,69]. An illustration of the coexistence of criticality and of SOC is found in a simple sandpile model of earthquakes on a hierarchical fault structure [70]. Here, the important ingredient is to take into account both nonlinear dynamics and complex geometry. From the point of view of self-organized criticality, this is surprising news: large earthquakes do not lose their identity. In the model of Huang et al. [70], a large earthquake is different from a small one, a very different story than the one told by common SOC wisdom in which any precursory state of a large event is essentially identical to a precursory state of a small event and an earthquake does not "know" how large it will become. The difference comes from the absence of geometry in standard SOC models. Reintroducing geometry is essential. In models with hierarchical fault structures, one finds a degree of predictability of large events.

Beyond Power Laws: Five Examples of "Kings"

Are power laws the whole story? The following examples suggest that some extreme events are even "wilder" than predicted by the extrapolation of power law distributions. They can be termed "outliers" or even better "kings" [17]. According to the definition of the Engineering Statistical Handbook [71], "An outlier is an observation that lies an abnormal distance from other values in a random sample from a population." Here, we follow Laherrère and Sornette [17] and use the term "king" to refer to events which are even beyond the extrapolation of the fat tail distribution of the rest of the population.

- **Material failure and rupture processes** There is now ample evidence that the distribution of damage events, for instance quantified by the acoustic emission radiated by micro-cracking in heterogeneous systems, is well-described by a Gutenberg–Richter like power law [72,73,74,75]. But consider now the energy released in the final global event rupturing the system in pieces! This release of energy is many, many times larger than the largest ever recorded event in the power law distribution. Material rupture exemplifies the co-existence of a power law distribution and a catastrophic event lying beyond the power law.
- **Gutenberg–Richter law and characteristic earthquakes** In seismo-tectonics, the situation is muddy because of the difficulties with unambiguously defining the spatial domain of influence of a given fault. Researchers who have delineated a spatial domain surrounding a clearly mapped large fault claim to find a Gutenberg–Richter distribution up to a large magnitude region characterized by a bump or anomalous rate of large earthquakes. These large earthquakes have rupture lengths comparable with the fault length [76,77]. If proven valid, this concept of a characteristic earthquake provides another example in which a "king" coexists with a power law distribution of smaller events. Others have countered that this bump disappears when removing the somewhat artificial partition of the data [78,79], so that the characteristic earthquake concept may be a statistical artifact. In this view, a particular fault may appear to have characteristic earthquakes, but the stress-shedding region, as a whole, behaves according to a pure scale-free power law distribution.
  Several theoretical models have been offered to support the idea that, in some seismic regimes, there is a coexistence between a power law and a large size regime (the "king" effect). Gil and Sornette [80] reported that this occurs when the characteristic rate for local stress relaxations is fast compared with the diffusion of stress within the system. The interplay between dynamical effects and heterogeneity has also been shown to change the Gutenberg–Richter behavior to a distribution of small events combined with characteristic system size events [81,82,83,84]. On the empirical side, progress should be made in testing the characteristic earthquake hypothesis by using the prediction of the models to identify independently of seismicity those seismic regions in which the king effect is expected. This remains to be done [85].

- **Extreme king events in the pdf of turbulent velocity fluctuations** The evidence for kings does not require, and is not even synonymous in general with, the existence of a break or a bump in the distribution of event sizes. This point is well-illustrated in shell models of turbulence which are believed to capture the essential ingredient of these flows, while being amenable to analysis. Such "shell" models replace the three-dimensional spatial domain by a series of uniform onion-like spherical layers with radii increasing as a geometrical series $1, 2, 4, 8, \ldots, 2^n$ and communicating mostly with nearest neighbors. The quantity of interest is the distribution of velocity variations between two instants at the same position or between two points simultaneously. L'vov et al. [86] have shown that they could collapse the pdf's of velocity fluctuations for different scales only for small velocity fluctuations, while no scaling held for large velocity fluctuations. The conclusion is that the distributions of velocity increments seems to be composed of two regions, a region of so-called "normal scaling" and a domain of extreme events. They could also show that these extreme fluctuations of the fluid velocity correspond to intensive peaks propagating coherently (like solitons) over several shell layers with a characteristic bell-like shape, approximately independent of their amplitude and duration (up to a rescaling of their size and duration). One could summarize these findings by saying that "characteristic" velocity pulses decorate an otherwise scaling probability distribution function.
- **Outliers and kings in the distribution of financial drawdowns** In a series of papers, Johansen and Sornette [87,88,89] have shown that the distribution of drawdowns in financial markets exhibits the coexistence of a fat tail with a characteristic regime with "kings" (called "outliers" in the papers). The analysis encompasses exchange markets (US dollar against the Deutsch Mark and against the Yen), the major world stock markets, the U.S. and Japanese bond markets and commodity markets. Here, drawdowns are defined as a continuous decrease in the value of the price at the close of each successive trading day. The results are found robust with using "coarse-grained drawdowns," which allows for a certain degree of fuzziness in the definition of cumulative losses. Interestingly, the pdf of returns at a fixed time scale, usually the daily returns, does not exhibit any anomalous king behavior in the tail: the pdf of financial returns at fixed time scales seems to be adequately described by power law tails [90]. The interpretation proposed by Johansen and Sornette is that these drawdown kings are associated with crashes, which occur due to a global instability of the market which amplifies the normal behavior via strong positive feedback mechanisms [91].
- **Paris as the king in the Zipf distribution of French city sizes** Since Zipf [92], it is well-documented that the distribution of city sizes (measured by the number of inhabitants) is, in many countries, a power law with an exponent μ close to 1. France is not an exception as it exhibits a nice power law distribution of city sizes… except for Paris which is completely out of range, a genuine king with a size several times larger than expected from the distribution of the rest of the populations of cities [17]. This king effect reveals a particular historical organization of France, whose roots are difficult to unravel. Nevertheless, we think that this king effect embodied by Paris is a significant signal to explain in order to understand the competition between cities in Europe.

Kings and Crises in Complex Systems

We propose that these kings may reveal information which is complementary and perhaps sometimes even more important than the power law pdf.

Indeed, it is essential to realize that the long-term behavior of complex systems is often controlled in large part by rare catastrophic events: the universe was probably born during an extreme explosion (the "big-bang"); the nucleosynthesis of all important atomic elements constituting our matter results from the colossal explosion of supernovae; the largest earthquake in California repeating about once every two centuries accounts for a significant fraction of the state's total tectonic deformation; landscapes are shaped more by the "millennium" flood that moves large boulders than by the action of all other eroding agents; the largest volcanic eruptions lead to major topographic changes as well as severe climatic disruptions; evolution is characterized by phases of quasi-statis interrupted by episodic bursts of activity and destruction; financial crashes can destroy trillions of dollars in an instant; political crises and revolutions shape the long-term geopolitical landscape; even our personal life is shaped on the long run by a few key "decisions/happenstances."

The outstanding scientific question is thus how such large-scale patterns of catastrophic nature might evolve from a series of interactions from the smallest to increasingly larger scales. In complex systems, it has been found that the organization of spatial and temporal correlations does not stem, in general, from a nucleation phase diffusing across the system. It results, rather, from a progressive and more global cooperative process occurring over

the whole system by repetitive interactions. An instance would be the many occurrences of simultaneous scientific and technical discoveries signaling the global nature of the maturing process.

Standard models and simulations of scenarios of extreme events are subject to numerous sources of error, each of which can have a negative impact on the validity of the predictions [93]. Some of the uncertainties are under control in the modeling process; they usually involve trade-offs between faithful descriptions and manageable calculations. Other sources of errors are beyond control as they are inherent in the modeling methodology of the specific disciplines. The two known strategies for modeling are both limited in this respect: analytical theoretical predictions are out of reach for most complex problems, while brute force numerical resolution of the equations (when they are known) or of scenarios is reliable only in the "center of the distribution", i. e., in the regime far from the extremes where good statistics can be accumulated. Crises are extreme events that occur rarely, albeit with extraordinary impact, and are thus completely under-sampled and poorly constrained. Even the introduction of teraflop (or even petaflops in the near future) supercomputers does not qualitatively change this fundamental limitation.

Recent developments suggest that non-traditional approaches, based on the concepts and methods of statistical and nonlinear physics could provide a middle way to direct the numerical resolution of more realistic models and the identification of relevant signatures of impending catastrophes. Enriching the concept of self-organizing criticality, the predictability of crises would then rely on the fact that they are fundamentally outliers, e. g., large earthquakes are not scaled-up versions of small earthquakes but the result of specific collective amplifying mechanisms. To address this challenge, the available theoretical tools comprise in particular bifurcation and catastrophe theories, dynamical critical phenomena and the renormalization group, nonlinear dynamical systems, and the theory of partially (spontaneously or not) broken symmetries. Some encouraging results have been gathered on concrete problems, such as the prediction of the failure of complex engineering structures, the detection of precursors of stock market crashes and of human parturition, with exciting potential for earthquakes. At the beginning of the third millennium, it is tempting to extrapolate and forecast that a larger multidisciplinary integration of the physical sciences together with artificial intelligence and soft-computational techniques, fed by analogies and fertilization across the natural sciences, will provide a better understanding of the limits of predictability of catastrophes and adequate measures of risks for a more harmonious and sustainable future for our complex world.

Future Directions

Our exposition has focused mainly on the concept of distributions of event sizes as a first approach to characterizing the organization of complex systems. But probability distribution functions are just one-point statistics and thus provide only an incomplete picture of the organization of complex systems. This opens the road to several better measures of the organization of complex systems.

- Statistical estimations of probability distribution functions is a delicate art. An active research field in mathematical statistics which is insufficiently used by practitioners of other sciences is the domain of "robust estimation." Robust estimation techniques are methods which are insensitive to small departures from the idealized assumptions which have been used to optimize the algorithm. Such techniques include M-estimates (which follow from maximum likelihood considerations), L-estimates (which are linear combinations of order statistics), and R-estimates (based on statistical rank tests) [94,95,96].
- Ideally, one would like to measure the full multivariate distribution of events, which can be in full generality decomposed into the set of marginal distributions discussed above and of the copula of the system. A copula embodies completely the entire dependence structure of the system [97,98]. Copulas have recently become fashionable in financial mathematics and in financial engineering [19]. Their use in other fields in the natural sciences is embryonic but can be expected to blossom.
- When analyzing a complex system, a common trap is to assume without critical thinking and testing that the statistics are stationary, implying that monovariate (marginal) and multivariate distribution functions are sufficient to fully characterize the system. It is indeed a common experience that the dependencies estimated and predicted by standard models change dramatically at certain times. In other words, statistical properties are conditional on specific regimes. The existence of regime-dependent statistical properties has been discussed in particular in climate science, in medical sciences and in financial economics. In the latter, a quite common observation is that investment strategies, which have some moderate beta (coefficient of regression to the market) for normal times, can see their beta jump to a much larger value (close to 1 or larger depending on the leverage of the investment) at certain times when the market collectively dives. Said dif-

ferently, investments which are thought to be hedged against negative global market trends may actually lose as much or more than the global market at certain times when a large majority of stocks plunge simultaneously. In other words, the dependency structure and the resulting distributions at different time scales may change in certain regimes.

The general problem of the application of mathematical statistics to non-stationary data (including non-stationary time series) is very important, but alas, not much can be done. There are only a few approaches which may be used and only in specific conditions, which we briefly mention.

1. Use of algorithms and methods which are robust with respect to possible non-stationarity in data, such as normalization procedures or the use of quantile samples instead of initial samples.
2. Modeling non-stationarity by some low-frequency random processes, such as a narrow-band random process $X(t) = A(t)\cos(\omega t + \phi(t))$ where $\omega \ll 1$ and $A(t)$ and $\phi(t)$ are slowly varying amplitude and phase. In this case, the Hilbert transform can be very useful to characterize $\phi(t)$ non-parametrically.
3. The estimation of the parameters of a low-frequency process based on a "short" realization is often hopeless. In this case, the only quantity which can be evaluated is the uncertainty (or scatter) of the results due to non-stationarity.
4. Regime Switching popularized by Hamilton [99] for autoregressive time series models is a special case of non-stationarity, which can be handled with specific methods.

- We already discussed the problem of "kings." One key issue that needs more scrutiny is that these outliers are often identified only with metrics adapted to take into account transient increases of the time dependence in the time series, as for instance in the case of returns of individual financial assets [88] (see also Chap. 3 of [91]). These outliers seem to belong to a statistical population which is different from the bulk of the distribution and require some additional amplification mechanisms active only at special times. The presence of such outliers both in marginal distributions and in concomitant events, together with the strong impact of crises and of crashes in complex systems, suggests the need for novel measures of dependence, different definitions of events and other time-varying metrics across different variables. This program is part of the more general need for a joint multi-time-scale and multi-variate approach to the statistics of complex systems.
- The presence of outliers poses the problem of exogeneity versus endogeneity. An event identified as anomalous could perhaps be cataloged as resulting from exogenous influences. The concept of exogeneity is fundamental in statistical estimation [100,101]. Here, we refer to the question of exogeneity versus endogeneity in the broader context of self-organized criticality, inspired in particular by the physical and natural sciences. As we already discussed, according to self-organized criticality, extreme events are seen to be endogenous, in contrast with previous prevailing views (see for instance the discussion in [64,102]). But, how can one assert with 100% confidence that a given extreme event is really due to an endogenous self-organization of the system, rather than a response to an external shock? Most natural and social systems are indeed continuously subjected to external stimulations, noises, shocks, solicitations, and forcing, which can vary widely in amplitude. It is thus not clear a priori if a given large event is due to a strong exogenous shock, to the internal dynamics of the system, or to a combination of both. Addressing this question is fundamental for understanding the relative importance of self-organization versus external forcing in complex systems and underpins much of the problem of dependence between variables. The concepts of endogeneity and exogeneity have many applications in the natural and social sciences (see [103] for a review) and we expect this viewpoint to develop into a general strategy of investigation.

Bibliography

1. Satinover JB, Sornette D (2007) "Illusion of Control" in Minority and Parrondo Games. Eur Phys J B 60:369-384
2. Satinover JB, Sornette D (2007) Illusion of Control in a Brownian Game. Physica A 386:339-344
3. Sornette D (2006) Critical Phenomena in Natural Sciences. Chaos, Fractals, Self-organization and Disorder: Concepts and Tools, 2nd edn. Springer Series in Synergetics. Springer, Heidelberg
4. Feller W (1971) An Introduction to Probability Theory and its Applications, vol II. John Wiley, New York
5. Ruelle D (2004) Conversations on Nonequilibrium Physics With an Extraterrestrial. Phys Today 57(5):48–53
6. Zajdenweber D (1976) Hasard et Prévision. Economica, Paris
7. Zajdenweber D (1997) Scale invariance in Economics and Finance. In: Dubrulle B, Graner F, Sornette D (eds) Scale Invariance and Beyond. EDP Sciences and Springer, Berlin, pp 185–194
8. Malcai O, Lidar DA, Biham O, Avnir D (1997) Scaling range and cutoffs in empirical fractals. Phys Rev E 56:2817–2828
9. Biham O, Malcai O, Lidar DA, Avnir D (1998) Is nature fractal? Response Sci 279:785–786
10. Biham O, Malcai O, Lidar DA, Avnir D (1998) Fractality in nature. Response Sci 279:1615–1616

11. Mandelbrot BB (1998) Is nature fractal? Sci 279:783–784
12. Mandelbrot BB (1982) The fractal Geometry of Nature. Freeman WH, San Francisco
13. Aharony A, Feder J (eds) (1989) Fractals in Physics. Phys D 38(1–3). North Holland, Amsterdam
14. Riste T, Sherrington D (eds) (1991) Spontaneous Formation of Space-Time Structures and Criticality. Proc NATO ASI, Geilo, Norway. Kluwer, Dordrecht
15. Pisarenko VF, Sornette D (2003) Characterization of the frequency of extreme events by the Generalized Pareto Distribution. Pure Appl Geophys 160(12):2343–2364
16. Jessen AH, Mikosch T (2006) Regularly varying functions. Publ l'Inst Math, Nouvelle serie 79(93):1–23. Preprint http://www.math.ku.dk/~mikosch/Preprint/Anders/jessen_mikosch.pdf
17. Laherrère J, Sornette D (1999) Stretched exponential distributions in nature and economy: Fat tails with characteristic scales. Eur Phys J B 2:525–539
18. Malevergne Y, Pisarenko VF, Sornette D (2005) Empirical Distributions of Log-Returns: between the Stretched Exponential and the Power Law? Quant Fin 5(4):379–401
19. Malevergne Y, Sornette D (2006) Extreme Financial Risks (From Dependence to Risk Management). Springer, Heidelberg
20. Frisch U, Sornette D (1997) Extreme deviations and applications. J Phys I, France 7:1155–1171
21. Willekens E (1988) The structure of the class of subexponential distributions. Probab Theory Relat Fields 77:567–581
22. Embrechts P, Klüppelberg CP, Mikosh T (1997) modeling Extremal Events. Springer, Berlin
23. Stuart A, Ord K (1994) Kendall's advances theory of statistics. John Wiley, New York
24. Bak P (1996) How Nature Works: the Science of Self-organized Criticality. Copernicus, New York
25. Sornette D (1994) Sweeping of an instability: an alternative to self-organized criticality to get power laws without parameter tuning. J Phys I Fr 4:209–221
26. Sornette D (2002) Mechanism for Power laws without Self-Organization. Int J Mod Phys C 13(2):133–136
27. Newman MEJ (2005) Power laws, Pareto distributions and Zipf's law. Contemp Phys 46:323–351
28. Wilson KG (1979) Problems in physics with many scales of length. Sci Am 241:158–179
29. Stauffer D, Aharony A (1994) Introduction to Percolation Theory, 2nd edn. Taylor & Francis, London, Bristol, PA
30. Gabrielov A, Newman WI, Knopoff L (1994) Lattice models of Fracture: Sensitivity to the Local Dynamics. Phys Rev E 50:188–197
31. Jensen HJ (2000) Self-Organized Criticality: Emergent Complex Behavior. In: Physical and Biological Systems. Cambridge Lecture Notes in Physics, Cambridge University Press
32. Mitzenmacher M (2004) A brief history of generative models for power law and lognormal distributions. Internet Math 1:226–251
33. Simkin MV, Roychowdhury VP (2006) Re-inventing Willis. Preprint http://arXiv.org/abs/physics/0601192
34. Sethna JP (2006) Crackling Noise and Avalanches: Scaling, Critical Phenomena, and the Renormalization Group. Lecture notes for Les Houches summer school on Complex Systems, summer 2006
35. Redner S (2001) A Guide to First-Passage Processes. Cambridge University Press, New York
36. Zygadło R (2006) Flashing annihilation term of a logistic kinetic as a mechanism leading to Pareto distributions. Phys Rev E 77, 021130
37. Stauffer D, Sornette D (1999) Self-Organized Percolation Model for Stock Market Fluctuations. Phys A 271(3–4):496–506
38. Kesten H (1973) Random difference equations and renewal theory for products of random matrices. Acta Math 131:207–248
39. Solomon S, Richmond P (2002) Stable power laws in variable economies. Lotka-Volterra implies Pareto-Zipf. Eur Phys J B 27:257–261
40. Saichev A, Malevergne A, Sornette D (2007) Zipf law from proportional growth with birth-death processes (working paper)
41. Reed WJ, Hughes BD (2002) From Gene Families and Genera to Incomes and Internet File Sizes: Why Power Laws are so Common in Nature. Phys Rev E 66:067103
42. Cabrera JL, Milton JG (2004) Human stick balancing: Tuning Lévy flights to improve balance control. Chaos 14(3):691–698
43. Eurich CW, Pawelzik K (2005) Optimal Control Yields Power Law Behavior. Int Conf Artif Neural Netw 2:365–370
44. Platt N, Spiegel EA, Tresser C (1993) On-off intermittency: A mechanism for bursting. Phys Rev Lett 70:279–282
45. Heagy JF, Platt N, Hammel SM (1994) Characterization of on-off intermittency. Phys Rev E 49:1140–1150
46. Clauset A, Shalizi CR, Newman MEJ (2007) Power-law distributions in empirical data. Preprint http://arxiv.org/abs/0706.1062
47. Gourieroux C, Monfort A (1994) Testing non-nested hypotheses. In: Engle RF, McFadden DL (eds) Handbook of Econometrics, Volume IV. Elsevier Science, pp 2583-2637
48. Cardy JL (1988) Finite-Size Scaling. North Holland, Amsterdam
49. Cranmer K (2001) Kernel estimation in high-energy physics. Comput Phys Commun 136(3):198–207
50. Sornette D, Knopoff L, Kagan YY, Vanneste C (1996) Rank-ordering statistics of extreme events: application to the distribution of large earthquakes. J Geophys Res 101:13883–13893
51. Pacheco JF, Scholz C, Sykes L (1992) Changes in frequency-size relationship from small to large earthquakes. Nature 355:71–73
52. Main I (2000) Apparent Breaks in Scaling in the Earthquake Cumulative Frequency-magnitude Distribution: Fact or Artifact? Bull Seismol Soc Am 90:86–97
53. Gabaix X, Ibragimov R (2008) Rank-1/2: A simple way to improve the OLS estimation on tail exponents. Work Paper NBER
54. Hill BM (1975) A simple general approach to inference about the tail of a distribution. Ann Stat 3:1163–1174
55. Drees H, de Haan L, Resnick SI (2000) How to Make a Hill Plot. Ann Stat 28(1):254–274
56. Resnik SI (1997) Discussion of the Danish Data on Large Fire Insurance Losses. Astin Bull 27(1):139–152
57. Pisarenko VF, Sornette D, Rodkin M (2004) A new approach to characterize deviations in the seismic energy distribution from the Gutenberg–Richter law. Comput Seism 35:138–159
58. Pisarenko VF, Sornette D (2006) New statistic for financial return distributions: power law or exponential? Phys A 366:387–400
59. Lasocki S (2001) Quantitative evidences of complexity of magnitude distribution in mining-induced seismicity: Implications for hazard evaluation. 5th International Symposium

on Rockbursts and Seismicity in Mines. In: van Aswegen G, Durrheim RJ, Ortlepp WD (eds) Dynamic Rock Mass Response to Mining, Symp Ser, vol S27, S Afr Inst Min Metall, Johannesburg, pp 543–550
60. Lasocki S, Papadimitriou EE (2006) Magnitude distribution complexity revealed in seismicity from Greece. J Geophys Res B11309(111). doi:10.1029/2005JB003794
61. Anderson PW (1972) More is different (Broken symmetry and the nature of the hierarchical structure of science). Science 177:393–396
62. Sornette D, Davis AB, Ide K, Vixie KR, Pisarenko VF, Kamm JR (2007) Algorithm for Model Validation: Theory and Applications. Proc Nat Acad Sci USA 104(16):6562–6567
63. Geller RG, Jackson DD, Kagan YY, Mulargia F (1997) Earthquakes cannot be predicted. Science 275(5306):1616–1617
64. Bak P, Paczuski M (1995) Complexity, contingency and criticality. Proc Nat Acad Sci USA 92:6689–6696
65. Allègre CJ, Le Mouel JL, Provost A (1982) Scaling rules in rock fracture and possible implications for earthquake predictions. Nature 297:47–49
66. Keilis-Borok V (1990) The lithosphere of the Earth as a large nonlinear system. Geophys Monogr Ser 60:81–84
67. Sornette A, Sornette D (1990) Earthquake rupture as a critical point: Consequences for telluric precursors. Tectonophysics 179:327–334
68. Bowman DD, Ouillon G, Sammis CG, Sornette A, Sornette D (1996) An observational test of the critical earthquake concept. J Geophys Res 103:24359–24372
69. Sammis SG, Sornette D (2002) Positive Feedback, Memory and the Predictability of Earthquakes. Proc Nat Acad Sci USA (SUPP1) 99:2501–2508
70. Huang Y, Saleur H, Sammis CG, Sornette D (1998) Precursors, aftershocks, criticality and self-organized criticality. Europhys Lett 41:43–48
71. National Institute of Standards and Technology (2007) Engineering Statistical Handbook, National Institute of Standards and Technology. Preprint http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm
72. Pollock AA (1989) Acoustic Emission Inspection. In: Metal Handbook, 9th edn, vol 17, Nondestructive Evaluation and Quality Control. ASM International, pp 278–294
73. Omeltchenko A, Yu J, Kalia RK, Vashishta P (1997) Crack Front Propagation and Fracture in a Graphite Sheet: a Molecular-dynamics Study on Parallel Computers. Phys Rev Lett 78:2148–2151
74. Fineberg J, Marder M (1999) Instability in Dynamic Fracture. Phys Rep 313:2–108
75. Lei X, Kusunose K, Rao MVMS, Nishizawa O, Sato T (2000) Quasi-static Fault Growth and Cracking in Homogeneous Brittle Rock Under Triaxial Compression Using Acoustic Emission Monitoring. J Geophys Res 105:6127–6139
76. Wesnousky SG (1994) The Gutenberg-Richter or characteristic earthquake distribution, which is it? Bull Seismol Soci Am 84(6):1940–1959
77. Wesnousky SG (1996) Reply to Yan Kagan's comment On: The Gutenberg-Richter or characteristic earthquake distribution, which is it? Bull Seismol Soc Am 86(1A):286–291
78. Kagan YY (1993) Statistics of characteristic earthquakes. Bull Seismol Soc Am 83(1):7–24
79. Kagan YY (1996) Comment On: The Gutenberg–Richter or characteristic earthquake distribution, which is it? by Wesnousky SG. Bull Seismol Soc Am 86:274–285
80. Gil G, Sornette D (1996) Landau–Ginzburg theory of self-organized criticality. Phys Rev Lett 76:3991–3994
81. Fisher DS, Dahmen K, Ramanathan S, Ben-Zion Y (1997) Statistics of Earthquakes in Simple Models of Heterogeneous Faults. Phys Rev Lett 78:4885–4888
82. Ben-Zion Y, Eneva M, Liu Y (2003) Large Earthquake Cycles And Intermittent Criticality On Heterogeneous Faults Due To Evolving Stress And Seismicity. J Geophys Res B6(108):2307. doi:10.1029/2002JB002121
83. Hillers G, Mai PM, Ben-Zion Y, Ampuero J-P (2007) Statistical Properties of Seismicity Along Fault Zones at Different Evolutionary Stages. Geophys J Int 169:515–533
84. Zöller G, Ben-Zion Y, Holschneider M (2007) Estimating recurrence times and seismic hazard of large earthquakes on an individual fault. Geophys J Int 170:1300–1310
85. Ben-Zion (2007) private communication
86. L'vov VS, Pomyalov A, Procaccia I (2001) Outliers, Extreme Events and Multiscaling. Phys Rev E 6305(5):6118, U158-U166
87. Johansen A, Sornette D (1998) Stock market crashes are outliers. Eur Phys J B 1:141–143
88. Johansen A, Sornette D (2001) Large Stock Market Price Drawdowns Are Outliers. J Risk 4(2):69–110. http://arXiv.org/abs/cond-mat/0010050
89. Johansen A, Sornette D (2007) Shocks, Crash and Bubbles in Financial Markets. In press, In: Brussels Economic Review on Non-linear Financial Analysis 149–2/Summer 2007. Preprint http://arXiv.org/abs/cond-mat/0210509
90. Gopikrishnan P, Meyer M, Amaral LAN, Stanley HE (1998) Inverse cubic law for the distribution of stock price variations. Eur Phys J B 3:139–140
91. Sornette D (2003) Why Stock Markets Crash, Critical Events in Complex Financial Systems. Princeton University Press, Princeton, NJ
92. Zipf GK (1949) Human behavior and the principle of least-effort. Addison-Wesley, Cambridge, MA
93. Karplus WJ (1992) The Heavens are Falling: The Scientific Prediction of Catastrophes in Our Time. Plenum, New York
94. Hubert PJ (2003) Robust Statistics. Wiley-Interscience, New York
95. van der Vaart AW, Gill R, Ripley BD, Ross S, Silverman B, Stein M (2000) Asymptotic Statistics. Cambridge University Press, Cambridge
96. Wilcox RR (2004) Introduction to Robust Estimation and Hypothesis Testing, 2nd edn. Academic Press, Boston
97. Joe H (1997) Multivariate models and dependence concepts. Chapman & Hall, London
98. Nelsen RB (1998) An Introduction to Copulas, Lectures Notes in statistic 139. Springer, New York
99. Hamilton JD (1989) A New Approach to the Economic Analysis of Non-stationary Time Series and the Business Cycle. Econometrica 57:357–384
100. Engle RF, Hendry DF, J Richard F (1983) Exogeneity. Econometrica 51:277–304
101. Ericsson N, Irons JS (1994) Testing exogeneity, Advanced Texts in Econometrics. Oxford University Press, Oxford
102. Sornette D (2002) Predictability of catastrophic events: material rupture, earthquakes, turbulence, financial crashes and human birth. Proc Nat Acad Sci USA 99(SUPP1):2522–2529

103. Sornette D (2005) Endogenous versus exogenous origins of crises, in the monograph entitled: Extreme Events in Nature and Society. In: Albeverio S, Jentsch V, Kantz H (eds) Series: The Frontiers Collection. Springer, Heidelberg (e-print at http://arxiv.org/abs/physics/0412026)

Probability and Statistics in Complex Systems, Introduction to

Henrik Jeldtoft Jensen[1,2]
[1] Institute for Mathematical Sciences, London, UK
[2] Department of Mathematics, Imperial College London, London, UK

Is the nature of complex systems such that they are in need of a special treatment in terms of statistics and probability. Yes, they are in the sense that a particular focus suggests itself. This becomes clear if we try to specify what we mean by complex systems. Although no consensus exists for the definition of what constitutes a complex system, the following summary will probably be accepted by most people.

- **Complex Systems** consist of a large number of interacting components. The interactions give rise to emergent hierarchical structures. The components of the system and properties at systems level typically change with time. A complex system is inherently open and its boundaries often a matter of convention.

Large numbers of components and interaction between these are central and this has immediate consequences for the nature of the relevant statistics. The interactions between the many components will typically be so strong that correlations cannot be neglected (► Correlations in Complex Systems) and hence, when we sum up contributions from the individual parts, we may not have the central limit theorem to ensure that the macroscopic quantities are Gaussian distributed. Instead of peaked distributions with well-defined average and higher moments, one typically encounters very broad heavy tailed distributions, frequently the tail reaches so far out that the average doesn't even exist. When this is the case, a description in terms of the typical—on the average scenario—is not possible (► Probability Distributions in Complex Systems). This makes not only fluctuations significant, it makes a description and understanding of the fluctuations absolutely essential (► Fluctuations, Importance of: Complexity in the View of Stochastic Processes). To illustrate this point think of, say, flood protection. It is no good to protect oneself against some imaginary average event, say a "typical" flooding, if effectively any size of flooding may occur with a non negligible probability. This makes it important to be able to deal with broad distributions and distributions of extreme events (► Extreme Value Statistics).

The hierarchical nature of complex systems (► Hierarchical Dynamics) leads to situations where the existence of many time scales cannot be ignored and as a result it may be inappropriate to assume the statistics to be stationary. The time dependence of the statistics may come about for a number of reasons. One is that the emergent hierarchical components change their internal state with time. An alternative reason may be that the strong interaction between components, themselves with no internal time evolution, prevents the collective set of components from reaching an equilibrium or asymptotic stationary state. In such cases one needs to be able to understand how to describe and predict the behavior of a system that is in a transient for all the relevant time scales, such as, say, the age of life on earth, if one is dealing with the history of the biosphere.

It seems natural to divide the discussion of probability and statistics for complex systems into at least three aspects:

1. Analyzing data from complex systems,
2. The phenomenology and
3. Modeling.

When analyzing data it is important to be able to handle the effect of long memory, correlations and exceptionally strong fluctuations. These effects make it necessary to exert special care when trying to identify probability distributions extracted from observational or experimental data or from data generated by computer simulations (► Probability Densities in Complex Systems, Measuring). The often very large amounts of data, and the lack of fundamental theory derived from first principle, have made it important to develop methods to identify structures in data sets, in this respect Bayesian statistics is often very relevant to complex systems (► Bayesian Statistics).

From experiments and observations we know that one signature of complexity may be power law like probability distributions. To determine the exponent can be complicated as the exponent one reads off from the slope in a log-log plot of the distribution may only be an "apparent" exponent. Not that this makes the exponent less important, but it does make the exponent more specific. An apparent exponent might very well depend on systems size and on the amount of collected data. In contrast we know from the lesson of statistical mechanics of critical phenomena that the asymptotic behavior of an infinite system may be the same even when microscopic details differ. This encourages the study of simplistic models in the hope that, though simplistic, the models might nevertheless capture

the essential mechanisms relevant for the phenomena under consideration. With this in mind, particular emphasis is placed on how apparent exponents converge towards their "universal" values in the limit of long distance and long time scales in the limit of a system composed of infinitely many components. Of course we don't know how applicable this form of universality is in the case of non-equilibrium complex systems. It might very well be that the many power laws observed in experiments are in fact only approximate and only persist for a limited range of support for the stochastic variable being studied. To settle this, careful experiments and theoretical studies are needed and will make use of some of the methods relevant to power laws discussed in the present section of the encyclopedia.

The focus on the functional form of the probability distributions becomes complicated by the fact that many complex systems never enter a stationary state during time scales accessible to observations or simulations. This doesn't exclude that the system relaxes towards a stationary sate in the mathematically asymptotic limit of infinite time, but this limit might be of little physical relevance. Then it becomes essential to study the very nature of how systems for all observational times is relaxing towards the stationary state. Inspired by the slow dynamics in glassy systems, and the very widespread intermittent dynamics encountered when many components interact, it has been suggested that the statistics of records may be of relevance to the intermittent relaxation of complex systems. The analysis of the statistics of the time instances marked by the occurrence of abrupt activity can be interesting. It might allow insight into the question concerning, how relevant record dynamics is for the complex dynamics and can in this way help to provide understanding of the collective dynamics of the components (► Record Statistics and Dynamics).

To construct and analyze theoretical models of complex systems a number of methods have been developed. At the intuitive level we have attempts to develop phenomenological theories by generalizing the well-studied branching process originating in Galton's and Watson's sociological studies (► Branching Processes). Among attempts to formulate theory of more basic foundation we encounter methods developed in physics to deal with phase transitions in materials where many interacting particles enter into a critical state. Here a critical state is taken to mean that no particular length or time scale can be identified, one talks about scale invariance, to indicate that all length and time scales are involved in the phenomena. To analyze such systems field theoretic methods are particularly useful as a tool to study the asymptotic behavior, i. e., long length and time scales (► Field Theoretic Methods). Stochastic analysis is particularly relevant since we are dealing with large numbers of components and, moreover, a stochastic element is often present even at the basic level, e. g. through thermal fluctuations in the case of physical systems (► Stochastic Processes).

In traditional statistical mechanics the concept of entropy has played a very important role. It is therefore only natural that work is being done to try to develop the definition of entropy to make it applicable beyond the traditional areas. In particular strong interaction makes the entropy non-extensive and appropriate generalizations are needed (► Entropy in Ergodic Theory).

The ordinary random walker is a bit of a workhorse in statistical mechanics. When random walks are used to model aspects of complex systems, the walker needs often to be dressed up with a broad distribution of step sizes (► Levy Statistics and Anomalous Transport: Levy Flights and Subdiffusion) or to walk in a background that makes the step size distribution location dependent (► Random Walks in Random Environment). This more sophisticated walker is not any longer necessarily controlled by the central limit theorem and new mathematics are needed.

The theory of random matrices is another example of methods developed to understand statistical aspects of physical systems that now have become of much broader importance. This is a field developed in response to the need of physicists when they, a while ago, started to try to understand the energy spectrum of heavy nuclei. The formalism has since then been used in a range of very different fields including the analysis of the stability of ecosystems. It is very likely that random matrices also will become a standard tool in complexity (► Random Matrix Theory). One reason for this is that complex systems often can be represented in terms of networks and that random matrix theory naturally relates to network analysis.

Finally, I feel it is in place to explain why this section contains an article about the stochastic Löwner equation (► Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems). The topic is included, as a spectacular example of how successful fairly abstract mathematics, from the realms of the pure end of the mathematical spectrum, can sometimes be in providing a detailed quantitative understanding of the intricacies of systems that are complex in the sense that they are far from equilibrium and in a non-stationary state. One might hope that this degree of detail in the mathematical understanding of the statistics of complex systems may become the norm in the future, as the research in to the statistical and probabilistic analysis of complex systems develop.

Protein Mechanics at the Single-Molecule Level

Mariano Carrión-Vázquez[1], Marek Cieplak[2], Andrés F. Oberhauser[3]

[1] Cajal Institute, CSIC & CIBERNED (Network on Degenerative Deseases), Madrid, Spain

[2] Institute of Physics, Polish Academy of Sciences, Warsaw, Poland

[3] Department of Neuroscience and Cell Biology, Department of Biochemistry and Molecular Biology and Sealy Center for Structural Biology and Molecular Biophysics, University of Texas Medical Branch, Galveston, USA

Article Outline

Glossary
Definition of the Subject
Background
Methodological Bases
Main Findings
Future Directions
Acknowledgments
Bibliography

Glossary

Atomic force microscopy (AFM) A near-field type of microscopy that uses a mechanical force sensor (*cantilever*) and a nanopositioner to either scan a surface, obtaining a *topography* of the sample ("imaging" configuration) and/or to stretch the sample to obtain a *force spectrum* ("force spectroscopy" configuration). This acronym is also used throughout the text for the instrument itself, the atomic force microscope.

Bionanomachines Nanoscopic structures, usually constituted by protein complexes, which perform most cell functions with a few exceptions like genetic information storage.

Enthalpic elasticity Elasticity derived from the breaking of non-covalent bonds after stretching.

Entropic elasticity In polymer science, this type of elasticity results from the entropic behavior of monomers in a polymer after its stretching. This recoiling is driven by thermal energy (second principle of thermodynamics) and results in the so-called *restoring force*.

Functionalization Process by which specific chemical groups are added typically to a surface and/or the sample to gain certain control of sample immobilization (attachment, orientation, coverage, etc.)

Hookean spring A spring that shows a linear force-extension relationship as is the case of an AFM cantilever.

Mechanical stability It can be operationally defined as the amplitude of the highest force peak (F_{max}) observed during the stretching of a single protein molecule using the length-clamp mode of SMFS, averaged over many unraveling events. Most proteins show just one force peak.

Molecular chaperons Protein complexes that help protein folding and complex assembly. Chaperonins are a well characterized subclass of chaperons that assist in the folding of newly-made proteins in all cells. These bionanomachines use chemical energy, in the form of adenosine triphosphate (ATP).

Molecular dynamics A specialized discipline of *molecular modeling* and *computer simulation* based on statistical mechanics. These techniques are used to simulate the behavior of molecules from the physicochemical principles.

Polyprotein In protein engineering, artificial polymeric protein formed by repeats (oriented or not) of a protein or a protein domain linked by covalent (*peptide*, *isopeptide* or *disulfide*) bonds. Polymerization can be achived at the DNA (by genetic engineering techniques: in vivo) or protein (by using biochemical techniques: in vitro) level.

Protein Natural biopolymer composed of up to 20 different monomers, *amino acids*, linked by so-called *peptide bonds* (a planar covalent bond), which typically acquires a unique 3D (*fold*) structure. The sequence of amino acids of a polypeptide is its *primary structure*. Proteins with *quaternary* structure are formed by several polypeptides, which are linked by non-covalent bonds, other covalent bonds, or both.

Protein engineering Discipline at the crossroads of molecular biology (includes genetic engineering, a technology to manipulate DNA), biochemistry, structural biology, and bioinformatics that aims to either the modification of proteins to improve or study its properties (redesign), or to design new proteins de novo. Usually it involves making changes in the sequence of a gene coding for a protein (usually by *polymerase chain reaction* and *directed mutagenesis*) in order to bring about desirable changes in its structure and/or function.

Protein folding Process by which a polypeptide acquires its native 3D structure (*fold*).

Protein fold Unique 3D structure of proteins (also called *superior structure* or *protein conformation*), achieved by either self-assembly alone or with the help or specific proteins (*molecular chaperons*), which is neces-

sary for its biological function. There are several structural levels of protein conformation: *secondary* (the main models are α-helix and β-sheet; the latter formed by β-strands), *tertiary* (final fold of a polypeptide; e. g. β-sandwich β-barrel), *quaternary* (resulting from the association of several polypeptides usually by non-covalent bonds). Protein structures (folds) resolved at atomic resolution are unique having an ascribed file specifying their atomic coordinates (i. e., Protein Data Bank, PDB file) and are classified somewhat artificially into discrete classes (e. g., immunoglobulin fold).

Protein nanomechanics New discipline in charge of measuring forces, distances, motions, energies, and deformations involved in the manipulation of individual proteins or protein complexes, which are typically in the sub-micrometer and sub-nanonewton ranges.

Protein unfolding Process by which a native protein loses its native folding becoming "*denatured*".

Proteasomes Large bionanomachines that degrade damaged or unneeded proteins by *proteolysis* (a chemical reaction that breaks peptide bonds). They are present in all kinds of cells, belonging to the class of enzymes called *proteases*, and they use ATP as a source of chemical energy.

Single-molecule force spectroscopy (SMFS) Technique carried out by several instruments (*AFM*, *optical tweezers* or *biomembrane force probe*) consisting in stretching single molecules to measure the resistance forces (*length-clamp* mode) or distances traveled between resistance barriers (*force-clamp* mode).

Unfolding (folding) pathway Energetic representation of the pathway of an unfolding (folding) reaction.

Definition of the Subject

Proteins can be considered as machine-like devices that function through complex structural changes in their intra- or intermolecular bonding. Understanding the dynamics of the inner workings of proteins is still one of the major challenges in biology.

Many proteins are nanomachines that use mechanical forces to fulfill a variety of cellular functions from replication to cell adhesion to cell crawling. The nanomachinery involved in these processes (i. e. the internal parts of these bionanomachines) is still poorly understood. Protein mechanics has emerged as a new multidisciplinary field to directly apply and measure mechanical forces through an array of recently developed dynamic techniques for manipulating single molecules, both in real time and under physiological conditions. After a decade, this field is still maturing fast and exciting developments await just around the corner.

AFM (atomic force microscopy) single-molecule force spectroscopy (SMFS) is one of the main technical pillars of this new discipline and it is particularly suited to directly quantify the forces involved in both intra- and intermolecular protein interactions. In combination with protein engineering and computer simulations, this technique has been used to characterize the unfolding and refolding reactions in a variety of protein structures, both with and without "mechanical functions".

Protein engineering allows the unequivocal identification of single molecules, through polyprotein analysis, and a careful experimental dissection of the experimental variables involved, through mutational analysis. Computer simulations based on molecular dynamics allow the process of the mechanical unfolding/folding of a protein to be modeled at the atomic level in order to obtain detailed structural information on its dynamics. These descriptions have been particularly useful in predicting and understanding the complexity behind the experimental results obtained by SMFS.

This review summarizes the concepts underpinning this field and some of the main findings to date.

Background

Mechanical Force: A New Biochemical Parameter

In contemporary biology we tend to regard the cell as a factory-like system crowded with a variety of specialized molecular "machines" of nanometer size (i. e., nanomachines), mainly proteins that exist either as single polypeptides or as complex assemblies of protein "parts" [5]. Understanding the inner workings of these biological machines remains one of the more active frontiers in biology.

The function of an individual protein or protein complex often involves the conversion of chemical energy (stored or supplied) into mechanical work through conformational changes. These "bionanomachines" that make use of mechanical forces are located throughout the cell (from the cell nucleus to the extracellular matrix) and they are involved in processes as diverse as replication, transcription, translation, protein folding, protein and nucleic acid unfolding, protein degradation, nucleic acid and protein translocation, organelle transport, muscle elasticity, cell adhesion, membrane fusion, or cell crawling [19,53] (Fig. 1).

Mechanical forces are also crucial to regulate the structure and function of cells and tissues. Thus, the shape of eukaryotic cells, and by extension that of the multicellular organisms they form, is the result of mechanosen-

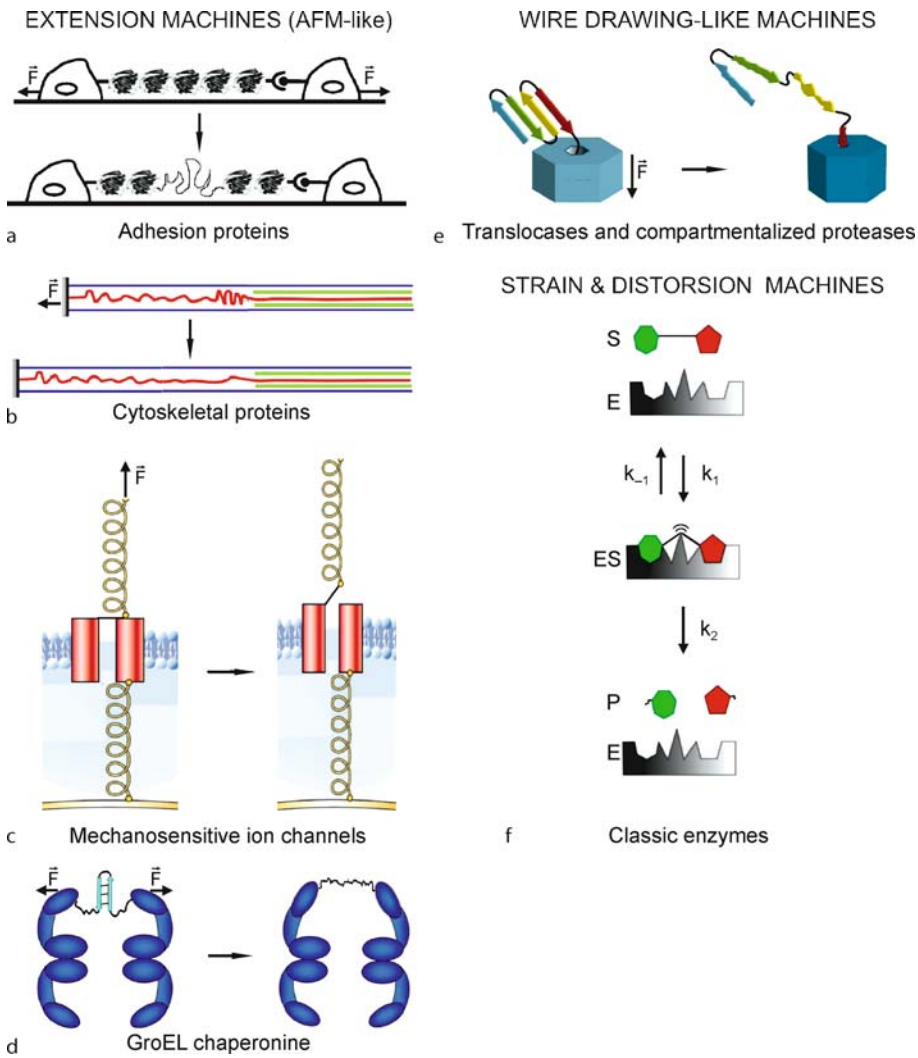

Protein Mechanics at the Single-Molecule Level, Figure 1

Mechanical biomachines. a Modular cell adhesion proteins may function as shock absorbers by increasing the range and lifetime of adhesion bonds. b Cytoskeletal proteins such as titin (in *red*, myosin filament in *green*, actin filaments in *blue*) may function as adjustable elastic springs. c Mechanosensitive ion channels are present in many biological systems some of which, like the auditory system have not yet been identified. The gating spring is a critical proteinaceous component of this machinery. d Chaperonines such as GroEL may induce conformational changes to mechanically unfold the substrate protein before refolding. e Compartmental proteases: the AAA+ (hexameric ring) ATPase from the proteasome and other related proteases unfold proteins, presumable by force, in an ATP-dependent manner, prior to their translocation to the catalytic chamber for degradation. f The strain and distortion hypothesis by Haldane and Pauling postulates that enzyme catalysis may work by inducing mechanical tension in the enzyme-substrate (ES) complex. (Modified from [27])

sory, mechanotransduction and mechanoresponse cycles. Responses to mechanical forces also underlie many biological processes from normal morphogenesis to carcinogenesis, cardiac hypertrophy, wound healing, and bone homeostasis. Indeed, recent studies show that several signaling pathways are involved in force transduction, including MAP kinases, small GTPases, and tyrosine kinases/phosphatases [6,48,49,121,122].

The molecular mechanisms by which mechanical forces influence these processes have been elusive due to the lack of appropriate tools. However, with the recent advent of single-molecule manipulation techniques we can now investigate these new biochemical pathways by directly probing bond dynamics in real time and under physiological conditions. These new techniques allow us to use mechanical force as an additional parameter in a biochem-

ical reaction, which can dramatically affect its rates in both directions.

Single Molecule Biology: A New Scientific Revolution

Single-molecule biology (alternatively referred to as single-molecule, biophysics or biochemistry) is an entirely new field of science, at the crossroads of several disciplines (namely biology, physics, chemistry, material science, and computer science), which overcomes the restrictions of the traditional bulk biochemical studies by focusing not on a population of molecules but on the molecule itself. Single-molecule methods may be considered a "paradigm shift" as they allow the behavior of individual molecules to be analyzed directly (under thermodynamic equilibrium or nonequilibrium conditions) at their real "nanoscale" and in their own "nanoworld" where thermal motion is a dominant force (see below).

Single-molecule techniques allow us to test the so-called ergodic principle (see ► Ergodic Theorems) of molecular populations: the measurement of a property in an ensemble at a given time should be equivalent to the average of the property measured on a single molecule over long periods of time. Moreover, these methods are revealing important information regarding:

a) Intermolecular variations in the experimental parameter of interest, which can arise from chemical (e. g. extent of glycosylation) or non-chemical (often called "static disorder") differences,
b) Rare events hidden in the ensemble,
c) The distribution underlying the ensemble average,
d) Fluctuations over time (the so-called "dynamic disorder"), and
e) Molecule kinetics without requirement for synchronization [61,96,134].

An additional technical advantage of single-molecule techniques is that the data can be directly compared to that obtained *in silico* by molecular dynamics simulations, as these methods also deal with single molecules.

There are two main types of single-molecule methods: 1) those that do not use an external force such as single-molecule fluorescence microscopy [61,96,134]; and 2) those imposing an external force to the system through an electric field (e. g. patch-clamp [83]) or a mechanical manipulation (through tension or torsion). The latter subtype, the so-called single-molecule manipulation techniques, offers a unique opportunity to study the behavior of molecules under an external mechanical force, applied either directly using flexible beams (AFM, microneedles, optical fibers) and vesicle membranes (biomembrane force probe), or through external-field manipulators (optical tweezers, magnetic tweezers, flow-field apparatus) [20,84]. These techniques have been used to examine the nanomechanics of the main biopolymers (DNA, RNA, polysaccharides, and proteins) [61,96,134]. Proteins are in charge of virtually every process that occurs in modern cells and are the subject of this review.

Protein Nanomechanics

Nanomanipulation techniques can be used to study all kinds of proteins, although pioneering experiments concentrated for obvious reasons (namely function, modularity and size) on the so-called "mechanical proteins". These are proteins with a mechanical function that can generate, transmit or use mechanical forces during their normal activity in the organism. They fall into two main subclasses: proteins that generate mechanical forces (biomolecular motors, which and have been mainly probed by optical tweezers) and proteins that are subjected to the mechanical forces generated by biomolecular motors or from the environment (mainly probed by AFM).

In order to identify the molecular mechanisms involved in the activity of both types of proteins it is important to analyze their mechanics at the single-molecule level. "Protein nanomechanics" achieves this by studying the forces, distances, motions, energies, and deformations involved in individual proteins or protein complexes, typically in the sub-micrometer and sub-nanonewton ranges.

Mechanical forces have also been used to probe the mechanical strength of "non-mechanical proteins" (i. e., those with no known mechanical function) to better understand the thermodynamics and kinetics of protein unfolding/refolding. Moreover, intermolecular interactions in protein complexes, some of which are subjected to mechanical force in vivo (e. g. adhesion proteins), have also been studied by mechanical stretching.

We shall focus on the response of individual proteins to the mechanical forces applied mainly by AFM and will provide an overview highlighting the principles, ideas, achievements, and perspectives of this fast-growing multidisciplinary field. For further reading and details, the reader is recommended to consult other reviews on protein nanomechanics from the general ones, which include those focusing on molecular biomotors [19,34,53,79], to others restricted to specific types of interactions: inter- and intramolecular [27,80,133], intermolecular [51,66,125], intramolecular [44,86,98,132], and molecular interactions of native membrane proteins [60,82].

Protein Mechanics at the Single-Molecule Level, Table 1

Relevant forces in protein nanomechanics. Ubiquitous thermal forces help to overcome the energy activation barriers of biochemical reactions and are the basis for the entropic elasticity displayed by many proteins in solution, which is on the range of a few pN. Breaking the non-covalent bonds (van der Waals, hydrogen, electrostatic) that maintain protein folds and protein interactions needs higher forces, typically below 300 pN, while the rupture of covalent bonds requires stronger forces, above 1000 pN. Forces accessible to the AFM techniques are highlighted in italics

Force type	Force range (pN)	Origin	Biological role	Protein examples
Langevin (thermal agitation)	~0.001	thermal energy	activation of energy barrier of reactions	typical enzymatic reactions
entropic	~0–10	thermal energy (on a polymer)	entropic elasticity (recoil) of biopolymers	titin passive elasticity (physiological range)
enthalpic (*non covalent*)	~10–300	non -covalent bonds	folding/interactions in biopolymers	unfolding/unbinding in proteins
enthalpic (*covalent*)	~1000–3000	covalent bonds	biopolymer synthesis	Proteolysis (enzymatic hydrolysis of proteins)

Methodological Bases

Range of Relevant Mechanical Forces in Biology

What are the magnitudes of the biologically relevant forces that affect protein structure? Proteins are subject to thermal forces, which are random in nature. These forces are in the femtonewton ($fN = 10^{-15}$ N) range and when they act on small objects like bionanomachines in solution, they result in what is called Brownian motion (Table 1). It is through thermal energy that proteins reach the high-energy transition states that are essential in biochemical reactions. To understand life at a fundamental level. it is important to know how these protein machines move their parts and change shape in response to the thermal and mechanical forces present in their nanoenvironment. The energies involved in protein conformational changes (the "signals" of our experiments) are just above thermal energy levels ("noise"), typically ranging from 1 $k_B T$ (thermal energy; $k_B T = 4.1$ pNnm $= 0.6$ kcal/mol, at room temperature; where k_B is the Boltzmann constant and T is the absolute temperature) to 25 $k_B T$ (the energy released by ATP hydrolysis) such that the structures are stable enough to prevail at physiological temperatures. Given that changes in protein conformation are measured in the Ångstrom-nanometer range (Å-nm, 1Å $= 10^{-10}$ m, 1 nm $= 10^{-9}$ m) the relevant biological forces are expected to be in the piconewton range (1 pN $= 10^{-12}$ N).

Because proteins are subject to thermal forces, the number of possible conformations (entropy) reaches a maximum when a protein forms a random coil or is denatured. Conformational entropy becomes progressively reduced with the formation of secondary and tertiary structures. Stretching random-coiled proteins in the low force regime to overcome "entropic forces" requires the application of forces in the order of a few pN, which has been achieved experimentally using single-molecule manipulation techniques. Several molecular motors such as myosin, kinesin, and RNA or DNA polymerases also generate forces in this range.

The next group are the "enthalpic forces", which includes the forces needed to unfold the folded domains of proteins (i. e. intramolecular interactions) as well as those required to overcome specific intermolecular interactions such as ligand/receptor or antigen/antibody. These forces are typically in the 50–300 pN range, when measured at a high loading rate. It must be noted that protein mechanical unfolding is typically a non-equilibrium dynamic process and therefore, these forces depend on the loading rate (see definition below). The typical loading rates in vivo may in some cases be much lower and accordingly the corresponding forces may also be lower.

Finally, the forces needed to break covalent bonds apart are almost two orders of magnitude larger, in the range of a few nanonewtons (1 nN $= 10^{-9}$ N) [134].

Single-Molecule Force Spectroscopy of Proteins: The Underlying Principle and Modes

AFM is the main technique used to characterize the mechanical resistance of both individual polypeptides (intramolecular interactions) and protein-biomolecule bonds (intermolecular interactions) [20]. Thus most of studies on single protein stretching have used this technique [27] with the exception of three studies performed with optical tweezers [8,28,62]. The mechanical resistance of intermolecular interactions in protein pairs have also been studied using the biomembrane force probe [41].

The AFM was originally developed as a high-resolution imaging tool [15] before it began to be used to probe and manipulate atoms and molecules. The so-called

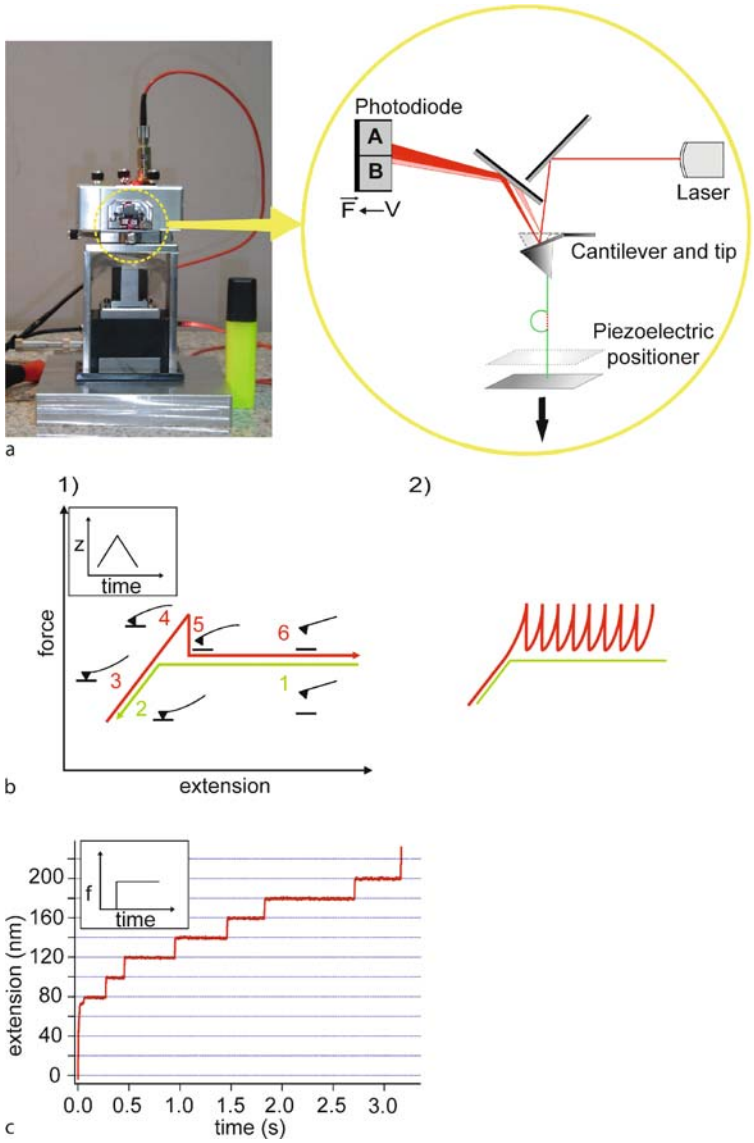

Protein Mechanics at the Single-Molecule Level, Figure 2

Single-molecule force spectroscopy (SMFS) by AFM: physical principle and modes of operation. a Schematic diagram of the typical AFM used in pulling experiments. The *line* depicts the laser light path before (*dashed line; pale red*) and after (*solid lines; red*) pulling on the system (i. e., protein, *green*). The protein connects mechanically the tip and the substrate, which is in turn bound to a piezoelectric positioner in this specific setup. The movement of the positioner along the *z*-axis results in bending of the cantilever along the same axis. This bending is tracked by changes in the reflected angle of a laser beam bounced off the cantilever, which in turn is detected by a split photodiode as a voltage difference between the two channels and is converted into force using Hooke's law. b 1 Typical force curve diagram in length-clamp SMFS mode showing different snapshots of the movement of the cantilever and tip as the positioner completes an approach-retraction cycle: it starts with the substrate not in contact with the tip (1), then it contacts with it, which bends the cantilever (2) increasingly (3); afterwards it is withdrawn from the tip, which bends the cantilever the other way (4) as it adheres to the tip, originating a force peak (4) on "jumping off contact" from the tip (5) which ends with the substrate again not in contact with the tip (6). 2 Schematic force-extension diagram showing a recording of a typical sawtooth pattern obtained by stretching of a multidomain protein molecule. c Force-clamp mode of SMFS showing the typical staircase extension-time recording. This particular example shows a polyubiquitin protein (N-C linked) being stretched at a constant force (110 pN). (Modified from [27], © Springer-Verlag 2006, with permission from Springer Science and Business Media)

"force spectroscopy" or "force-measuring" configuration of the AFM (Fig. 2) was designed to record force-extension curves obtained by pulling in a single direction (*z*-axis). Single molecules can be readily analyzed in this way by "single-molecule force spectroscopy" (SMFS).

SMFS is a very sensitive technique that can measure forces of tens of piconewtons and changes in length at nanometer resolution. A common problem is that force peaks can originate from a variety of sources other than the interaction of interest (detachment of other molecules from any of the two anchoring points, protein-protein interactions, disentanglement of molecules, etc.), or from multiple molecules in parallel. This serious drawback was overcome in pioneering studies of protein nanomechanics by using modular proteins [87,94] or homomeric recombinant polyproteins [25], in which their pseudo-periodicity or pure periodicity, respectively, was used to infer single molecules unequivocally (i. e., a single-molecule reporter). These protein molecules are first immobilized between the substrate (a glass coverslip) and the force sensor so that a "mechanical circuit" is established that connects these two points (Fig. 1). Typically, in these experiments pro-

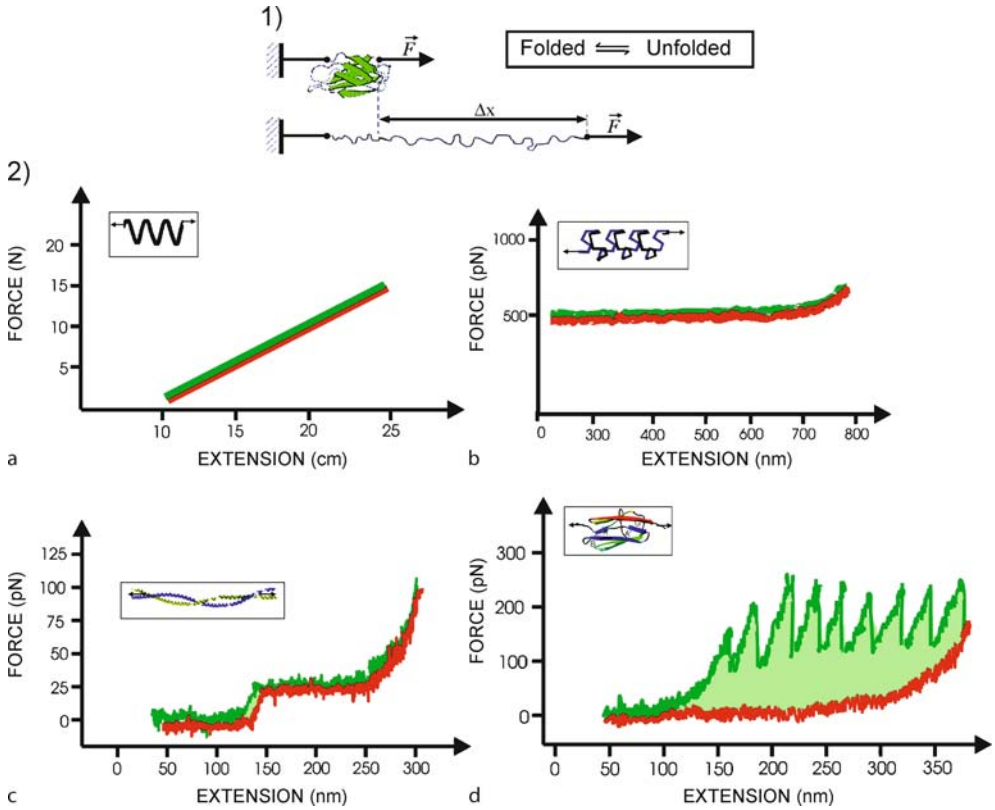

Protein Mechanics at the Single-Molecule Level, Figure 3

Elastic behavior of proteins under SMFS (length clamp mode). 1 Cartoon representation of the process of mechanical unfolding for a protein module using an applied force (*F*) that results in an extension of the protein (Δx). 2 Equilibrium and nonequilibrium in protein unfolding-refolding (*green* unfolding process, *red* retraction, *pale green* hysteresis; insets show the corresponding structures). a Force-extension curve of a "Hookean" spring such as the AFM cantilever (a macroscopic system). b Force-extension curve of an elastomeric protein: elastin. This protein behaves as an entropic spring in equilibrium, showing no hysteresis. c Force-extension curve of an elastic protein (*inset*): the tail of myosin II. This structure behaves close to a truly elastic protein showing little hysteresis when relaxed, which reflects the small amount of energy dissipated between extension and retraction. d Force-extension curve of a modular protein: a region of 8 Ig modules from the elastic region of titin. This structure shows a characteristic entropic-enthalpic force spectrum. Stretching the ends of a multimodular protein sequentially unfolds the domains, generating a typical saw-tooth pattern. The forced unfolding of this structure is a nonequilibrium process (note the marked hysteresis), which makes this protein a perfect shock absorber: it dissipates as heat part of the energy put into the system by the external mechanical work. *Inserts:* structure of a typical macroscopic spring (a), putative structure of elastin (b), structure of myosin II tail (c), and structure of the titin I27 Ig module (d). (Modified from [27])

teins become attached by physisorption (i. e., nonspecific adsorption via "physical forces") to the two elements, although sometimes specific functionalization methods are also used (e. g., terminal cysteine residues in the protein to covalently link it to the gold coated surface of the substrate or/and the cantilever tip). Often, proteins in solution are attached nonspecifically to the substrate forming a dense layer of molecules from which the cantilever tip can pick out single molecules at random. Protein molecules are then stretched as in a lilliputian medieval rack, by moving apart the AFM positioner (bound to the substrate), which applies a mechanical force that unfolds the individual domains. This imposes a specific reaction coordinate (i. e. the end-to-end distance) on the unfolding process. By retracting the AFM positioner, the protein can also be refolded in the presence or absence of mechanical force. These experiments are typically performed under nonequilibrium conditions (Fig. 3). There are two basic modes currently being used depending on the variable being controlled: the most common, length-clamp, yielding a "force-extension" curve; and force-clamp, which yields an "extension-time" curve (Fig. 2).

What Kind of Information Can Be Extracted from Stretching Proteins?

When the length-clamp mode is used on a modular protein or a polyprotein, the first source of resistance to extension typically comes from entropic forces, as these are "polymeric" proteins formed by several "pseudo-repeats" or perfect repeats, respectively. Entropic elasticity is a general property of polymers that results from the tendency of a chain to form a coil in order to maximize the conformational freedom (entropy) of its constituent monomers under the drive of thermal fluctuations. Stretching this polymeric molecule reduces its entropy producing a restoring force that bends the cantilever. In contrast to the linear relationship that governs cantilever bending (common mechanical springs follow Hooke's law, Fig. 3), the entropic elasticity of a protein follows a non-linear relationship that can be formally described by the so-called "worm-like chain" model of polymer elasticity (WLC, see Sect. "Mechanical Dissection of Titin I27 Module: A Model System").

Further extension of the protein may unravel it in a typically all-or-none fashion, which can be described by a two-state model (Fig. 4, see below). This exposes previously "force hidden" amino acid residues to force, resulting in an increase in the end-to-end length of the protein trapped between the tip and the positioner (ΔL_c), while relaxing the force acting on the cantilever to near

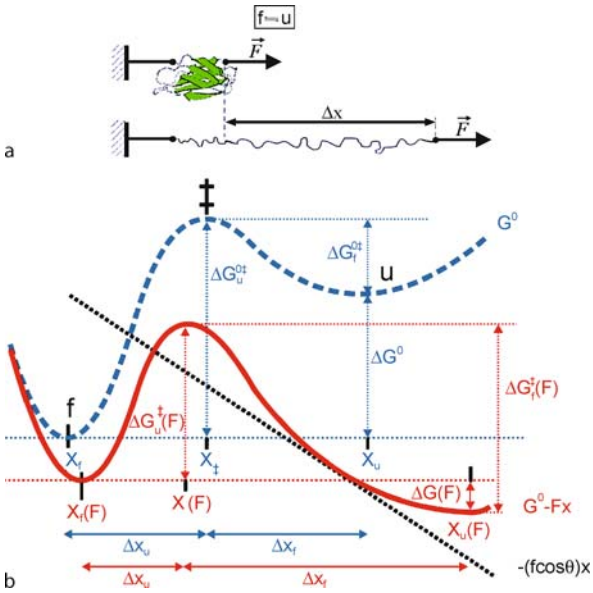

Protein Mechanics at the Single-Molecule Level, Figure 4
The effect of a mechanical force on the energy diagram of a protein. a Cartoon representation of the process of mechanical unfolding as in Fig. 3a. b The effect of a mechanical force on the free energy diagram of a protein that unfolds following a two-state model (f folded, u unfolded). The *dashed blue curve* represents the process in the absence of an applied force. An applied force (*black dashed line*) tilts the energy diagram of the process, decreasing the barrier to the transition state, ‡ [$\Delta G^{\ddagger}(F) < \Delta G0^{\ddagger}$] (*red line*). The application of force also lowers the energy of the unfolded state relative to that of the folded state [$\Delta G(F) < 0$]. The mechanical reaction coordinate is *x*. With an applied force, the positions of the free-energy minima (x_f and x_u) and the maximum ($x^{\ddagger}$) shift such that Δx_u becomes shorter (Δx_u, *red* < Δx_u, *blue*) and Δx_f longer (Δx_f, *red* > Δx_f, *blue*). Local curvature of the free-energy surface dictates these relative shifts in position. (Modified from [19] © Annual Reviews 2004, with permission)

zero (see Sect. "Mechanical Dissection of Titin I27 Module: A Model System"). Thus, unfolding events are detected as sudden changes in the end-to-end distance of the molecule. For a modular protein or a polyprotein, further extension repeats this cycle for each of the remaining folded modules, giving rise to a characteristic sawtooth pattern in the force-extension curve (Fig. 2b). When the force-clamp mode is used, the extension-time curve adopts a staircase pattern (Fig. 2c).

SMFS permits the direct measurement of both the mechanical stability of the barriers that a protein offers to its stretching (i. e., the mean force, as the process of unfolding is thermally driven and is stochastic) and the location of these barriers with single amino acid resolution [24]. As we shall see, the mechanical stability of a protein is a property that cannot be inferred from thermodynamic stability

measurements. Additionally, and based on conventional transition state theory, we can infer the kinetic parameters of the forced unfolding/refolding reaction [9,40]. An applied mechanical force tilts the energy diagram of the process, decreasing the barrier to the transition state and increasing the rate of the forward reaction (Fig. 4).

It is possible to calculate the probability density for unfolding, which predicts the most likely force of unfolding in terms of the spontaneous unfolding rate constant as follows:

$$F_u = (k_B T/\Delta x_u) \ln(r \Delta x_u / k_u^0 k_B T) \,.$$

By using this analytical solution we can calculate the kinetic parameters for the unfolding reaction: k_u^0 (spontaneous rate of unfolding) and Δx_u (width of the activation energy barrier, i. e. the distance on the reaction coordinate over which the force must be applied to reach the transition state). This equation predicts that the mechanical stability of a protein (F) depends on the unfolding distance, Δx_u, the height of the barrier ($\Delta G_u^{\ddagger}$, which depends on k_u^o), thermal energy, and the loading rate used during the extension of the protein ($r = \mathrm{d}F/\mathrm{d}t = k.v$; where v is the pulling speed).

Also, by relaxing the tethered protein before it breaks and waiting for appropriate periods of time, we can perform refolding experiments (in the presence of force or in its virtual absence) from which we can extract the equivalent parameters of the folding process.

Two recent SMFS studies show that it is now possible to gather both detailed information on protein structure, through a method called "mechanical triangulation" [38], and to probe dynamic rearrangements within the active site of an enzyme with unprecedented resolution [129].

Computer Simulations of the Mechanical Unfolding of Proteins

SMFS experiments can measure changes in length at single amino acid resolution [24]. However, the typical methods used cannot provide a detailed structural resolution nor a "microscopic" (in the physicist sense, "nanoscopic" for biologists) interpretation of these processes. To this end computer simulations by molecular dynamics have proven to be very important for the atomic analysis of this process. They allow us to follow how mechanical forces change the structure of proteins under stress at the atomic level. These simulations are relatively simple to implement since the denaturing agent (force) and the reaction coordinate (N-C distance, if the protein is pulled by its termini) can be readily simulated. As these techniques also deal with single molecules, they are especially suitable for direct comparison with SMFS results, which constitutes one of the main appeals of SMFS for theoreticians. The methods that have been used are all-atom (with and without explicit water solvent) and coarse-grained (e. g., on- and off-lattice simulations) models. The latter approaches give less detail but allow massive studies to be carried out at near experimental pulling speeds.

Classical all-atom simulations have been used for decades as an approach to interpret the behavior of proteins near their native states over time scales up to the order of 100 ns, with and without an explicit water solvent [46,50,93,109]. The usual approach is to resolve Newton's equations (for instance, by using the leap-frog or prediction-correction algorithms) for all of the atoms in the system, monitor trajectories, and calculate time averages. However, interatomic forces have many sources. The forces are typically described by about 1000 parameters, which are collectively known as the 'force field' and they are determined through studies of model peptides. The primary contributions are Coulombic interactions between partial charges placed on the individual atoms. The all-atom models are often considered as being realistic representations of proteins, however, they incorporate several important approximations. Indeed, they neglect the dynamics of electrons as described by Schrödinger's equation and what is often more important, the force field is 'trained' to be valid near expected native states of simple proteins and not in their non-native conformations.

By applying all-atom simulations to protein stretching we can obtain a wealth of information about the atomic details of conformation, which often enables the nature of mechanical stability to be elucidated (i. e., to explain which modules give rise to resistance to pulling). However, in order to fit the process to the computationally available time scales, very fast pulling speeds are used. These simulations are generally performed at pulling speeds many orders of magnitude faster than those of the experiment (eight or more, typically 1–10 mm/ms vs. 0.1–1 nm/ms). Because of this reduction in time, additional irreversible work is done to stretch the molecule, which typically results in force peaks one order of magnitude higher than those observed in AFM experiments. Hence, as simulations not fully reproduce the experimental conditions, it is not clear whether they precisely model the real process in these short periods of time in which new physical phenomena may arise. It should be noted that the peak heights at the computational speeds are usually substantially larger not only than the experimental unfolding force peaks but also than their extrapolation, assuming a logarithmic dependence on the speed that is often found.

These simulation techniques have a much higher time resolution than experimental single-molecule techniques (which can only capture slow conformational motions). Computational times are also shorter, in the range of nanoseconds or tenths of nanoseconds, which are achieved by increasing the pulling speed in the case of constant-velocity simulations (or by increasing the force in the case of constant-force simulations). However, a direct comparison is not possible because these techniques simulate the unfolding process over a very short time period (picosecond to nanosecond range), whereas experimental SMFS data are obtained over much longer times (millisecond-second range).

In spite of these difficulties, the synergy between all-atom simulations and experiments is very powerful. It is expected that such simulations will be important for the future development of the field in order to obtain the necessary high resolution picture of the mechanical unfolding of proteins. The first protein for which stretching was simulated was the titin I27 domain, which has become a model system in the field (see Sect. "Mechanical Dissection of Titin I27 Module: A Model System"). So far, about 30 protein structures have been studied using all-atom simulations, representing about half of the proteins folds studied by SMFS. Both numbers are only a tiny fraction of proteins stored in the PDB or existing in any organism. In this proteomic era there is a need for approaches to investigate the elastic landscape of large numbers of proteins and coarse-graining has recently offered such a tool.

Coarse-grained molecular dynamics models provide access to the experimental timescales of protein processes, in an approximate way, not only by reducing the number of degrees of freedom (e. g., by dealing only with the C^{α} atoms) but also by introducing simple effective interactions that pertain to the larger scale level of description. The coarse-grained models are generally well suited for studies of large conformational changes, such as those occurring during folding, stretching and thermal melting. They offer further advantages when considering biomolecular complexes, such as the multiple linkages of proteins and ribosomes [116]. Their high-throughput character allows the comparison of the properties of a large number of proteins. A further simplification of the model may be implemented by making the conformational space discrete through constraining the locations of the C^{α} atoms to certain sites of a lattice. This step also requires replacing Newton-equations based dynamics by Monte Carlo sampling [74].

Coarse-grained models of proteins have gained popularity in the last decade, although biochemists tend to dismiss them because of their reduced 'realism'. Among the coarse-grained models of proteins, the so-called Go-like systems are currently the most widely used [17,32,35,52,59,63,120,126], particularly since they are easy to implement and yet specific to a protein. The general idea is to formulate a model with effective couplings that is consistent with the experimentally established structure of the native state [1,114]. Clearly, there is no unique prescription for how to implement it. However, there is a host of simple choices that produce fairly good correlations with the experimental SMFS data. Go-like models have questionable features to study folding, but they are expected to be more reliable for studies of stretching where much of the relevant dynamics of the process takes place near the native state.

The construction of a C^{α}-based Go-like model starts by representing each amino acid by a bead located at the C^{α} site. Consecutive beads are tethered by harmonic potentials with minima at 3.8 Å. The next step is to introduce bead-bead interactions that would minimize the potential energy of the system in the experimentally established native structure. This is done by first determining which interactions should be operational in the native state and which should not. This results in the generation of the so-called contact map, which lists the native contacts (i. e. interactions). A sensible way to construct the contact map is to read in the all-atom native structure and represent each amino acid by a cluster of 'grapes'. Each grape has a radius equal to the van der Waals radius of the corresponding atom multiplied by a factor to account for attraction. If such clusters overlap in the native state a native contact arises, which represents specific interactions such as hydrogen bonds, ionic bonds, disulfide bonds, and so on. Otherwise the contact is non-native and the corresponding interaction is made repulsive to prevent entanglement.

Once the contact map has been determined, one needs to decide on how to pick an effective potential that represents a contact whatever its chemical nature. Among the simple choices, the Lennard–Jones potential, $V_{\mathrm{LJ}} = 4\varepsilon[(\sigma/r)^{12} - (\sigma/r)^{6}]$, appears to work particularly well. Here, r is the distance between interacting C^{α} atoms, σ is the length parameter, which is determined contact-by-contact so that the minimum in the potential coincides with that in the experimental structure. ε is the energy scale. In its simplest version, the energy scale is uniform and equal to 1.3–1.6 kcal/mol (so that the unit of force, ε/Å, is 67–110 pN). Disulfide bonds are covalent in nature so their effective coefficient ε is an order of magnitude larger. The model can be improved by terms that represent local backbone stiffness. They favor the native values of the bond and dihedral angles. In order to mimic the thermostating effects of the solvent, one also introduces random

forces whose amplitude is proportional to temperature, T, (Langevin noise) and a damping force that produces overdamping. One can take $k_B T/\varepsilon \sim 0.3$ to represent room temperature situations. The characteristic time scale (τ), in the resulting molecular dynamics simulations is of the order of ns instead of ps since it relates to the diffusion time across a distance of a typical σ, instead of an oscillatory time.

It should be noted that temperature exerts a profound effect on the force-displacement patterns since thermal fluctuations contribute to unraveling. The higher the T, the smaller the force peaks. At sufficiently high temperatures ($k_B T/\varepsilon \sim 0.8$), in the so-called entropic limit, the force peaks due to the contact potentials disappear altogether. It is in this limit that the physics of an entropic chain applies.

The stretching simulations in all-atom models are usually implemented by making the molecular dynamics "steered", which means that pulling is done directly through the attachment to a selected amino acid, usually a terminus, and this amino acid is made to move in a specific way. In coarse-grained models, stretching is usually accomplished by attaching two amino acids to elastic springs that imitate the elasticity of the cantilever on one end and that of the attachment to a substrate on another end. These simulations can be done at the experimental pulling speeds although, in order to survey the PDB at reasonable times, they are typically done two orders of magnitude faster.

Protein Engineering is a Fine Tool for Protein Nanomechanics

Protein engineering has been critical for the progress of this field not only because it allows the construction of polyproteins (single-molecule markers for SMFS) but also for permitting the introduction of mutations.

Mutational analysis is a high-resolution tool that permits minimal perturbations of the system to be introduced in order to analyze the underlying molecular details. It has offered a unique way to study the molecular basis of how proteins respond to mechanical force. Several types of mutant proteins of the titin I27 domain (See Sect. "Mechanical Dissection of Titin I27 Module: A Model System") have been used in SMFS studies. Loop insertion using glycine residues was used to show the existence of a "mechanical clamp" in I27 formed by backbone hydrogen bonds, demonstrating also the amino acid resolution of the technique [24]. Proline mutagenesis was used to show the existence of a mechanical folding intermediate [75] and to alter the mechanical stability of the domain [69]. This approach takes advantage of the fact that proline (a known breaker of secondary structure) is an imino acid and as such, cannot form backbone hydrogen bonds. Conservative substitution and deletion mutagenesis have been used to demonstrate the existence of the I27 folding intermediate [45]. Furthermore, analysis of the so-called mechanical Φ-value holds promise for a more detailed examination of the structure of the transition state in a forced unfolding reaction [11]. This is the mechanical equivalent to the well established Φ-value analysis [43], which is used to examine the conformational effect of a mutation based on the relative changes in free energy of the native, transition and denatured states ($\Delta\Delta G_{N-\ddagger}$, $\Delta\Delta G_{U-\ddagger}$). The Φ-value is obtained from the changes in thermodynamic stability measured from the shifts in the equilibrium denaturation curves. The unfolding Φ-value, defined as $\Phi_u = \Delta\Delta G_{N-\ddagger}/\Delta\Delta G_{N-D}$, is determined by comparing the effect of the mutation on the transition state and it measures the amount of native structure that is present around the mutated residue in the transition state. The $\Delta\Delta G_{N-D}$ is determined using equilibrium denaturation experiments. This analysis can only be applied when Δx_u remains the same for wild type and mutant proteins. Using these methods it has been shown that mechanical unfolding of titin I27 is a 3-state process in which the first transition state (between the native and the intermediate states) is very similar to the native state (Fig. 5).

Main Findings

Which Proteins Have Been Studied?

Over the last decade both mechanical and non-mechanical proteins have been studied by mechanical unfolding and folding using SMFS [27,86] (Table 2). The protein structures analyzed so far include the following:

a) the "all-β" protein E2Lip3 [17];
b) several "all-β" structures of the β-sandwich type including "Ig-like" domains (from titin, fibronectin, tenascin, filamin, polycystin-1, Sls-kettin and projectin) and a "C2 domain-like" module (from synaptotagmin I) [18,26,70,87,89,92,105];
c) a protein of the β-barrel type (GFP: green flourescent protein) [37,90];
d) several "$\alpha + \beta$" proteins: ankyrin B, T4 lysozyme, barnase, DHFR (dihydrofolate reductase), β-grasp proteins (ubiquitin, protein L, GB1), RNase H, maltose binding protein, and Top7 [3,8,10,21,23,28,67,103,106,131];

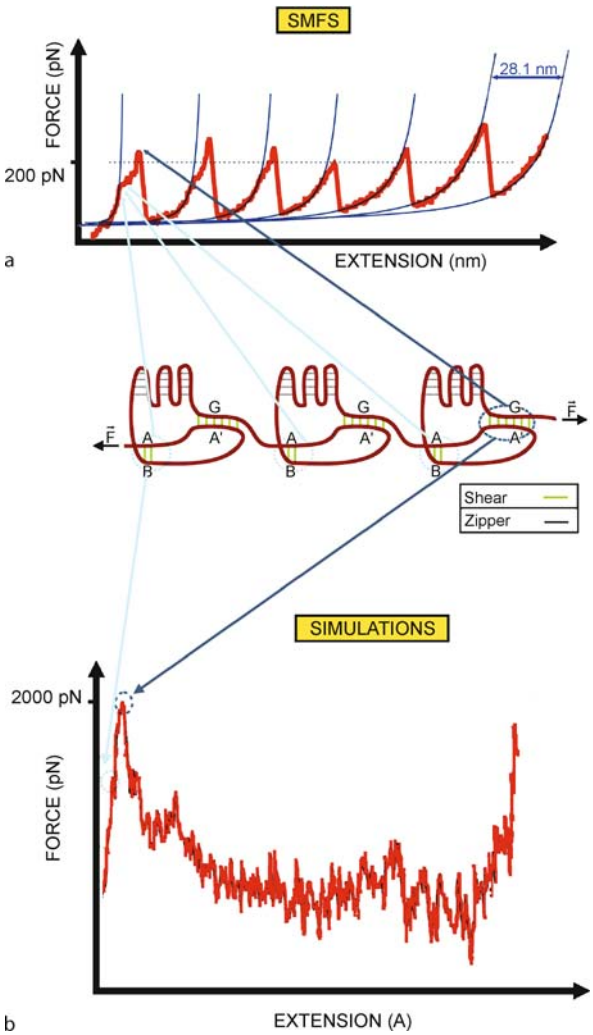

◀ **Protein Mechanics at the Single-Molecule Level, Figure 5**
A model system in protein nanomechanics: the I27 module. Schematic representation of the mechanical architecture of the I27 module from titin. A three repeats polyprotein (center panel), showing patches of backbone hydrogen bonds in both "zipper" (*dashed grey lines)* and "shear" (AB, A′G; *solid green lines)* mechanical topologies. a A force extension-curve (force spectrum) obtained after stretching an I27 polyprotein by SMFS-AFM. b Force spectrum predicted by steered molecular dynamics (*SMD*) simulations, by an all-atom method. The main stability determinants of this module are a minor mechanical barrier (A-B patch of backbone hydrogen bonds in the polypeptide backbone) and a major one (A′G patch of backbone hydrogen bonds). The major resistance barrier corresponds to the force peaks in the force–extension spectrum (a). The minor barrier is detected in SMFS as a deviation of the WLC ("hump"), which is more apparent in the first peak of the spectrum since it collects the contribution of all the events from the stretched modules (*pink arrows*). This intermediate refolds after each unfolding event, as the force relaxes. These results are correlated with SMD simulations, where a low force peak and a high one can be observed when stretching a single domain (b). Although the extension data are similar (28.1 nm in AFM vs. 25–30 nm in SMD), the force data are different. Thus, the AFM mean peak value is about 200 pN, whereas the main force peak in the SMD simulation is about 2,000 pN. This discrepancy can be attributed to the different timescale of the experiments; simulations are much shorter processes (1 ns to tenths of 1 ns) than experiments (milliseconds to seconds) and thus they cannot fully reproduce the experimental conditions. (Modified from [27], © Springer-Verlag 2006, with permission from Springer Science and Business Media)

e) several "all α" structures (calmodulin, spectrin, and myosin II tail) [7,26,64,95,104]; and
f) several unstructured proteins (elastin, titin PEVK, and titin N2B) [65,70,71,100,118].

Most proteins analyzed to date (i. e., modular proteins and recombinant polyproteins) have been pulled in the N-C direction, which is the natural direction provided by the peptide bond as synthesized at the ribosome. Exceptions are naturally occurring proteasomal polyubiquitins (K48-C linked) [23], E2Lip3 domain (K41-C) [17], lyzozyme polyproteins cysteine-linked in solid state (C21-C124) [131] and GFP polyproteins cysteine-linked in solution (C3-C132, C3-C212, and C132-C212) [38]. These alternative linkages offer unique opportunities to apply force to different points of the protein. As we will see, these alternative geometries of pulling greatly alter the mechanical stability of proteins.

In addition to single polypeptides, the mechanical properties of some protein complexes as a whole have been also analyzed by SMFS (e. g., myosin II tails [104] and *E. coli* adhesive pili [81]).

Molecular Determinants of the Mechanical Stability of Proteins

The number of proteins analyzed so far by SMFS is still very small (about 23 different proteins, roughly 55 PDB structures) and they have been studied with differing detail (systematization of the conditions in which these studies are done would be highly desirable, as already proposed [97]). Although this prevents us from being able to extract general principles, regarding mechanical stability some tendencies have been observed that can be summarized as follows [27,86]:

Proteins have widely different mechanical stabilities, measured by the most probable unfolding force (F_u) when pulled in the N-C direction: they range from less than the limit of resolution of AFM (typically ~20 pN; e. g. calmodulin) to 330 pN (e. g., titin Ig domains). Interestingly, mechanical proteins intended to resist force tend to be more mechanically stable than non-mechanical ones or the so-

Protein Mechanics at the Single-Molecule Level, Table 2

Mechanical stability of representative proteins. Mechanical stability (measured in pN) is defined as the average force of unfolding. The proteins listed here were all pulled apart from their N- and C-termini at comparable loading rates

Protein fold		Examples	Mechanical stability, F (pN)
α-helix		calmodulin	< 25
β-spiral		α -elastin, PEVK (titin)	< 25
α-helical structures			
bundles		spectrin, dystrophin, myosin tail, ankyrin B repeats	25–50
solenoids		ankyrin B	360 ?
$\alpha + \beta$ structures	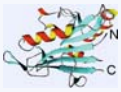	dihydrofolate reductase, barnase	27–70
β-grasp	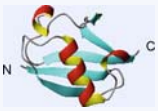	ubiquitin, GB-1, protein L	136–200
β-barrel	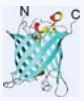	GFP	100
β-sandwich			
"zipper"	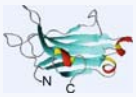	C2A domain	60
"shear" (PKD)		polycystin-1	50–250
"shear" (fn)		titin A band, fibronectin, tenascin, projectin (fn), myomesin (fn)	80–200
"shear" (Ig)		titin I band, sls -kettin, projectin (Ig), myomesin (Ig), filamin A	150–330

called elastomeric proteins, although there are some exceptions.

Unstructured and β-spiral proteins (e. g. elastomeric proteins like elastin and the PEVK and N2B regions from human titin) are among the less mechanically stable proteins. Such proteins are entropic springs and behave reversibly. α-helical proteins (e. g., calmodulin, T4 lysozyme) also have relatively low mechanical stability although α-helical bundles (e. g., spectrin, myosin II tail) and solenoids (e. g., ankyrin B) are more stable. β-stranded proteins tend to unfold at higher unfolding forces than α-helical ones.

The mechanical stability of most mechanical proteins (e. g., Ig-like β-sandwich and the ubiquitin-like β-grasp folds) tends to be determined by highly-localized mechanical clamps at the breakpoint formed by shear hydrogen bonds between the backbones of the β-strands. However, in some protein folds (e. g., fibronectin type III, fnIII, from human tenascin- TNfnIII3) the hydrophobic core also contributes to mechanical resistance [85]. Exceptions to the shear mechanical clamp topology are GFP, ankyrin B, and the de novo designed Top7 fold. In addition to secondary-structure based elasticity, there are two more types of structure elasticity: tertiary (e. g., the solenoid of ankyrin

B [67]) and quaternary (e. g. the helical rod of *E. coli* adhesive pili [81] and the myosin II tail [104]).

Since force is a vector quantity, the mechanical stability and the mechanical unfolding pathway depend on the pulling geometry, which is affected by both the topology at the breakpoint and the point of application of the force. Hence, β-stranded proteins with a shear mechanical topology at the breakpoint (i. e. the force vector is orthogonal to the hydrogen bonds) are more mechanically stable than zipper β-stranded proteins (where the force vector is parallel to the hydrogen bonds). The points of application of the force to a protein are also relevant as they can substantially alter its mechanical stability [17,23], implying that proteins have "Achilles' heels" in their structure. For instance, GFP pulled from geometries other than N-C shows mechanical stabilities of up to ~550 pN [39].

Mechanical stability is a kinetic property which in general is not correlated with thermodynamic stability (ΔG) or with melting temperature ($Tm = \Delta G/\Delta S$) [26]. Mechanical stability seems to be roughly predicted by the unloaded unfolding rate constant [27].

Chemical and mechanical unfolding have been shown to follow different pathways [10,45] and have different unfolding barriers [10,16].

Moreover, it has been demonstrated that mechanical stability can be modulated by ligand binding [3,58] and by disulfide bond formation [4,14,22,128,129].

Taking together, these findings show that proteins display a broad range of responses to mechanical stress, which cannot easily be rationalized in terms of predictors of mechanical resistance. Thus, although the molecular basis underlying the mechanical resistance of proteins is still unclear, several determinants have been identified through these studies: amino acid sequence, mechanical topology, unloaded unfolding rate constant and pulling geometry.

The molecular structure of a protein, poses constraints on the location of the transition state in mechanical unfolding. It has been suggested that tertiary interactions have shorter distances to their transition states than secondary structures, and they tend to be more brittle (i. e. they break at high forces and after small deformations) than secondary interactions, which are more compliant (breaking at low forces and after large deformations). Furthermore, tertiary interactions may require more time to equilibrate than secondary ones and therefore, they often present hysteresis in the pulling-relaxation cycle [19]. Most proteins show a high degree of connectivity and as a result, their unfolding seems to be highly cooperative with the stability of secondary structures depending on their tertiary context and often presenting no intermediates. Still, due to the local action of the applied force, their mechanical stability tends to be related to highly localized molecular structures near the mechanical "breakpoint" rather than to the global structure [19,73]. A massive survey has been carried out recently to identify these mechanical clamps in all protein modules (up to 150 amino acids long) for which there are atomic structures available [110,111].

Mechanical Dissection of Titin I27 Module: A Model System

The model system most commonly used to study mechanical unfolding/refolding in proteins is the I27 module from titin (Fig. 5), a gigantic multimodular protein responsible for the so-called passive elasticity of muscle (see Sect. "The Elasticity of Muscle Explained at the Single-Molecule Level) [117]. This module has an 89 amino acids long Ig-like β-sandwich fold (Fig. 3d and Table 2). All-atom molecular dynamics simulations of its stretching identified two patches of backbone hydrogen bonds as true "structural" barriers with different mechanical resistances: a low force barrier involving 2 hydrogen bonds between β-strands A and B (AB patch), and a high force barrier between β-strands A′ and G involving 6 hydrogen bonds (A′G patch) [73]. The hydrogen bonds in both patches are perpendicular to the direction of the force vector (a "shear" mechanical topology: bonds are arranged "in parallel" in the mechanical circuit), whereas the remaining hydrogen bonds in the structure are parallel to the force vector (a "zipper" mechanical topology where the bonds are arranged in series) and like the hydrophobic core, seem to offer little resistance to extension (Fig. 5). Interestingly, coarse-grained models yield similar results [32,126].

These predictions were remarkably consistent with the experimental data obtained from SMFS using polyproteins: an intermediate found at low force (~100 pN, at 0.3–05 nm/ms) was associated to the rupture of the AB patch [75], whereas a high force peak (~200 pN, at 0.6 nm/ms) was found to depend exclusively on the A′G patch [25] and the associated side-chain packing (i. e. hydrophobic) interactions between the A′ and G strands. The hydrophobic core of this structure plays no role in resisting force [13]. The hypothetical role of the A′G patch as a mechanical clamp was tested by loop insertion [24] and proline mutagenesis [69], these mutants providing the first mechanical phenotypes.

Unveiling Intermediates and Rare Misfolding Events

As mentioned in the last section, experimental stretching of I27 also confirmed the existence of a weaker mechanical barrier in the AB patch, which appeared as a small

deviation in the force-extension recordings (at ∼100 pN and at 0.3–0.5 nm/ms) from the pure entropic behavior described by the WLC model (a "hump" of decreasing intensity on each of the saw teeth; Fig. 5). This deviation was interpreted as evidence for the existence of an unfolding intermediate involving the rupture of the AB patch. The existence of this unfolding intermediate in the wild type I27 module was later confirmed by using mutant proteins [45,75].

During refolding experiments using polyproteins or modular proteins, "skip" events are occasionally observed, whose contour lengths suggest they are due to the formation of a "superfold" which includes two consecutive domains plus the linker region between them. This misfolded "superdomain" unfolds at similar forces to those for a single domain and refolds back to the dimensions and mechanical stability of two normal domains. SMFS of the titin poly-I27, the fnIII domains of tenascin, and the R1617 tandem repeat of spectrin have demonstrated that these single-molecule studies can detect rare events in as little as 2, 4 and 3% of the population, respectively. Hence, the efficiency (i. e. fidelity) of refolding after mechanical unfolding can be estimated [21,88], which has been found to be exceptionally high for the non-mechanical protein GB1, a structural homologue of protein L (above 99.8%) [21]. Such rare unfolding events could not be detected using ensemble techniques and they open the door to investigate the means of reversing the undesirable misfolding that occurs in a number of pathological conditions.

The Elasticity of Muscle Explained at the Single-Molecule Level

The basic contractile unit of muscle is the sarcomere. This relatively simple contraction/extension machine (i. e. it works in a single dimension) is highly elastic [55]. The pioneering work by Wang et al. [124] and Maruyama et al. [76] demonstrated that the so-called "passive elasticity" of muscle (i. e. generation of restoring forces that resist stretch independently of ATP) is mainly mediated by titin, a giant protein (> 3MDa, the longest polypeptide known to date) that spans half a sarcomere (∼ 1 µm; from the *Z* disk to the *M* line) and that acts as a molecular spring (Fig. 6). Passive elasticity plays an important role in muscle function since, typically, a muscle actively contracts against the elastic strain of a passively elongating muscle. This property ensures that the sarcomere recovers its initial dimensions on muscle relaxation [117].

A remarkable feat achieved using SMFS was the reconstruction of the passive elasticity of intact myofibrils by simply scaling up from the mechanical properties of single titin molecules [70] (Fig. 6). These were reconstructed from the mechanical properties of representative elements of its elastic region (i. e. N2B, PEVK, and the proximal and distal Ig regions relative to the N-terminus of the protein). These results showed that titin behaves very differently from a Hookean spring. Instead, in response to axial tension, titin behaves as a multistage non-linear spring that adjusts both its length and apparent stiffness by virtue of its particular modular design. At low forces the entropic springs dominate (i. e. N2B, PEVK, and Ig straightening) while at high force (i. e. in non-physiological conditions) the enthalpic springs (partial and total unfolding of a few Ig domains) act as "shock absorbers" to prevent damage of the sarcomere. Thus, titin Ig shock absorbers are essentially a "safety mechanism" for the sarcomere.

Through this reductionist approach, it has been possible to explain muscle passive elasticity (a macroscopic biological property) from the additive mechanical properties of a single sarcomeric protein at the single-molecule level. The example of titin elasticity shows how single-molecule experiments can be used to elucidate, at a more fundamental level, the physiological function of a protein through a pure biophysical dissection of a complex hierarchical system.

Molecular Shock Absorbers Galore

Mechanical proteins tend to be modular, and are often composed of assortments of modules of the same type (e. g. immunoglobulin, fibronectin), which frequently display distinct mechanical stabilities. The shock absorber effect of the Ig domains from titin has also been found in cell adhesion proteins like tenascin (in its fnIII modules), where it was proposed to extend the range and lifetime of cell-cell interactions [87]. In the case of fibronectin, another extracellular matrix protein, most of its (Ig-like) fnIII domains remain folded during the stretching of its matrix [2], similar to titin.

Protein domains from muscle (titin, myomesin, projectin and Sls-kettin), cell adhesion (tenascin, fibronectin, polycystin-1), cytoskeletal (filamin), and surface receptor proteins tend to belong to the Ig-like β-sandwich family of proteins and probably have somewhat similar mechanical topologies at the breakpoint. This superfamily of Ig folds, which includes Ig (titin, projectin, Sls-kettin), fnIII (tenascin, fibronectin, titin, projectin), E-set (filamin), and PKD (polycystin-1) types, consist of 7 stranded β-sandwich structures in which the N- and C-terminal strands are parallel to each other, pointing in completely different directions (i. e., 180°, Table 2, Fig. 3). These modules seem to have evolved to withstand forces, when con-

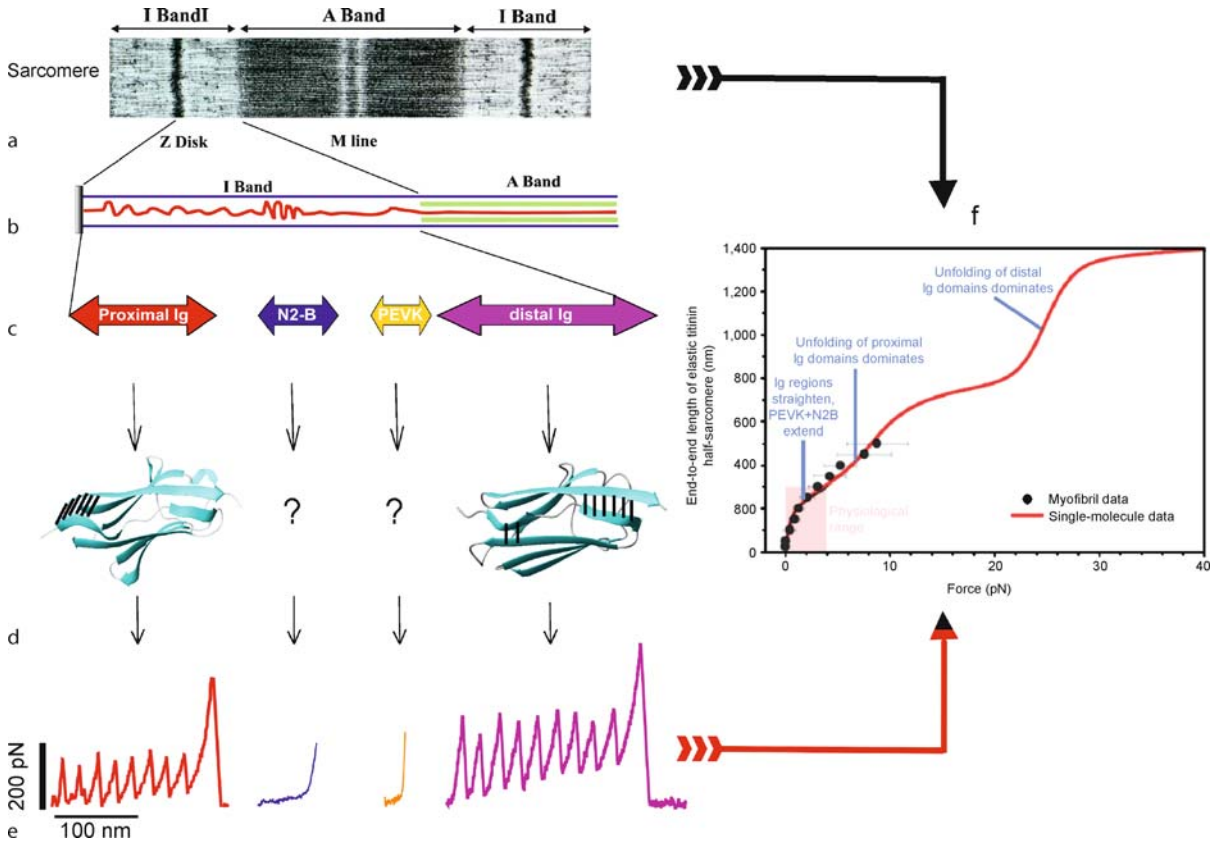

Protein Mechanics at the Single-Molecule Level, Figure 6

Reverse engineering of titin. a–f Reconstruction of muscle passive elasticity from the mechanical properties of single titin proteins. The mechanical properties of representative elements of its elastic region, I band (b). c These elements are the proximal Ig domains (the crystal structure of I1 is shown in d, *left panel*), the N2B and PEVK (hypothetical random coils) and the distal Ig domains (the NMR structure of I27 is shown in d, *right panel*). e Representative force-extension recordings obtained after stretching I4-I11 (in *red*), an N2B (in *blue*), a PEVK (*yellow*) and $I27_8$ proteins (*violet*). f According to this model, within the physiological range of sarcomere extension (i. e. forces below 4 pN) unfolding would rarely happen and most of the elasticity of titin would result from the entropic elasticity of straightening the Ig domains in the I band, and of extending the random coils (PEVK and N2B). (Modified from [70,86])

nected in series, thanks to a shear mechanical topology of the hydrogen bonds at the breakpoint. This arrangement may provide these domains with considerable resistance to mechanical stretching. Thus, the Ig-like β-sandwich fold seems to be a platform that can tolerate higher mechanical stress than other folds. Within this class, the Ig (forces ranging from 150–330 pN, at 0.6 nm/ms) and the PKD (~200 pN; at 1 nm/ms) domains appear to be more mechanically stable than the fibronectin type III (fnIII; 75–220 pN, at 0.6 nm/ms) or the E-set (to which filamin belongs; 50–220 pN, at 0.37 nm/ms) domains [27].

As discussed, SMFS is often used for the mechanical analysis of multi-modular proteins and polyproteins. In these proteins, stretching results in the unwinding of many segments. Often the patterns are interpreted as a serial process. This seems justified in many cases (especially for polyI27 or polyubiquitin), but there are examples when this assumption fails (e. g., polycalmodulin). Temperature also seems to play an important role in these experiments. At high temperatures, thermal fluctuations dominate and all segments unwind in parallel (i. e. simultaneously, because such fluctuations are independent of the segment number although they are sensitive to the sequential distance between amino acids that are in contact so that the short range contacts unravel last). Thus, in these conditions the weakest spots will unravel first, irrespective of whether they are hidden, exposed, or near the termini. On the other hand, at very low temperatures strong spots resist simultaneous unwinding, resulting in a serial process except at the very initial stages when all segments yield somewhat. At room temperature, there is a competition between serial and parallel unwinding pathways and either

one may win, depending on the nature of the modules. Thus, in general, one should always expect some degree of mutual influence between modules [29,31,56].

On the Biological Meaning of Protein "Robustness": Selected Trait or Epiphenomenon?

How well do SMFS experiments mimic protein mechanics in vivo? In the case of extension machines like the sarcomere, SMFS seems to adequately mimic the natural linear pulling geometry, since proteins are pulled apart from both ends of the polypeptide chain. This may also be the case for other cytoskeletal machines (the case of spectrin may not be as clear as it forms a mesh-like structure), the adhesion machinery, and some mechanosensitive ion channels. Furthermore, it has been suggested that some chaperonins may pull their protein substrates apart in a similar way prior to their refolding [107].

However, the case of the unfoldases is not so clear-cut (Fig. 1). The accepted model for protein translocases (from the mitochondrion, chloroplast and endoplasmic reticulum) and compartmental proteases (such as the proteasome), is also a mechanical one. Nevertheless, rather than a linear pulling geometry with two attachment points, the evidence here favors a different geometry that involves a single attachment point from which the pulling would be done by threading the protein towards the entrance of a narrow channel present in these nanomachines (i. e., in similar way to wire drawing in metallurgy). This model is mainly based on the fact that the susceptibility of substrate proteins to be unfolded by these nanomachines in vivo correlates more closely with the mechanical stability obtained by mechanical unfolding (using SMFS) than with thermodynamic or kinetic stability (measured in vitro by bulk chemical or heat denaturation). In the case of compartmental proteases, the AAA+ ATPase motor involved in the pulling process seems to unfold the structure adjacent to the degradation tag by trapping local unfolding fluctuations. Global unfolding then occurs immediately, driven by the cooperativity of the protein unfolding process [77,91,102]. As seen from SMFS findings, in this "local stability" model the structure and pulling geometry at the attachment point (i. e. local stability) are more important than their global counterparts. Similarly, protein import by the mitochondrial translocase depends on the N-terminal targeting sequence and the local structure of the adjacent protein, more akin to the vectorial nature of AFM pulling experiments than to solution or heat denaturation experiments [130]. Indeed, mechanical hypomorphic mutations can also accelerate mitochondrial import under specific conditions [101]. The vectorial nature of protein translocation highlights the importance of the existence of Achilles' heels in proteins. Accordingly, in order to unfold their protein substrates more economically, mechanical "unfoldases" might have evolved specific pulling mechanisms to take advantage of the presence of weak spots in the structure of the latter.

If the mechanical model is confirmed for "unfoldases" (molecular chaperones, compartmental proteases such as proteasomes, and the protein translocases from the protein import machinery of the mitochondria, chloroplasts, and endoplasmic reticulum), then most proteins in the cell (i. e., "mechanical substrates" of unfoldases) would be mechanically unfolded at some point or another during their lifespan, which would make their mechanical properties physiologically relevant.

Mechanochemical enzymes may not be alone in their capacity able to generate mechanical forces. Thus, according to the hypothesis of the tension-induced catalysis proposed by Haldane and Pauling, mechanical forces may also underlie the activity of other enzymes [19]. This hypothesis postulates that enzyme catalysis may work by inducing mechanical tension in the enzyme-substrate complex.

The correlation between mechanical stability and mechanical function holds well at high forces but it is less true below ~100 pN [27]. This may reflect the possibility that for some proteins with no known mechanical function (e. g. protein L, GB1, GFP) their relatively high mechanical stability (in some cases derived from a shear mechanical topology at the breakpoint) may just be an epiphenomenon, unrelated to the biological function of the protein. Indeed, a number of the proteins studied may never experience mechanical forces in the cell and if they do, the pulling geometry in situ might be different to that of SFMS experiments. Alternatively, mechanical stability may simply be a neutral trait for certain proteins, not directly selected by evolution (i. e. a remnant by-product of their evolutionary history). For instance, the mechanical stability of fnIII domains from the titin A-band (non-elastic region bound to the myosin thick filament of the sarcomere and therefore probably not subject to axial stress) is only slightly lower than those of the I-band (180 vs. 200 pN, at similar pulling speed). Hence, if our current model of how titin works is correct, this trait does not appear to have been evolutionarily selected in the case of the A-band domains (non-elastic) but rather, it may be an epiphenomenon originated through constraints imposed during the evolution of this elastic protein. Alternatively, it is also possible that the anchoring geometry of these domains to myosin may require such a high mechanical stability.

In conclusion, by using SMFS only a certain type of mechanical stability in proteins can be accessed at present.

This mechanical stability (defined as F_{max} when they are pulled from the N- to C-termini, using the length-clamp mode) would be expected to be an evolutionarily selected trait for components and substrates of mechanical biomachines that work like extension machines (i. e., similar to a medieval rack with two attachment points, Fig. 1a,b,c,d). However for other mechanical biomachines (e. g., wire drawing-like machines. Figure 1e) and for "non-mechanical" proteins, this parameter may well be a mere epiphenomenon. Thus, in each case careful comparative and functional analysis should be done to distinguish between these two possibilities.

Predicting the Robustness of Proteins from Their Structure

As we have already pointed out, the nanomechanics of only a few proteins and protein modules has been studied to date. Considering how vast proteomes are, there are therefore plenty of mechanical proteins and unfoldase substrates waiting to be mechanically characterized. In the mean time, simulations provide a nice short cut in order to help us to predict and understand protein robustness, and to access pulling geometries that are not yet experimentally accessible (e. g., "wire drawing" pulling, [54,115,127]).

As mentioned above, using a simple variant of a Go-model a massive survey [110,111] has been recently conducted to identify mechanical clamps in 7510 protein modules, of up to 150 amino acids long, for which uncorrupted atomic structures were available at the time when the study was undertaken. Additionally, 239 substantially longer proteins were also included in the survey. This study has identified new mechanical clamps and has enabled the available protein structures to be ordered on the basis of their predicted mechanical stability. This model correlates fairly well (correlation coefficient of 89%) with the available experimental values of the mechanical stability, F_{max}. The theoretical and experimental data points are listed in Table 3 and displayed in Fig. 7. The comparison is restricted to 28 experimentally stretched proteins for which the PDB structure is available.

This survey used a simple Go-like model and stretching was implemented at a constant speed and at a temperature meant to correspond to room temperature. The values of resistance to pulling, F_{max}, together with the force-displacement patterns are available at info.ifpan.edu.pl/BSDB (Bio-molecule Stretching DataBase). The survey is restricted to stretching at the termini (i. e., N-t and C-t amino acids), although it should be kept in mind that there

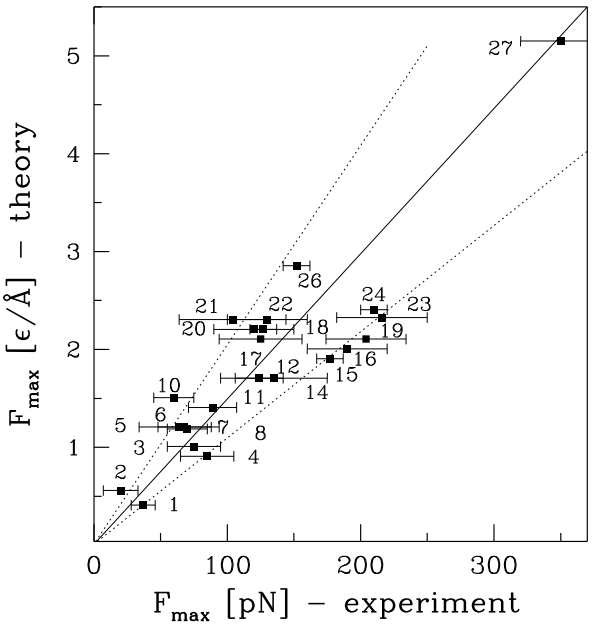

Protein Mechanics at the Single-Molecule Level, Figure 7
Modeling mechanical stability with coarse-grained molecular dynamics. Correlation between the experimental and theoretical values of F_{max} for a simple Go-like model with uniform Lennard–Jones potentials in the native contacts. The contact map is obtained by studying heavy atom overlaps and by eliminating the $i, i+2$ contacts which usually are weak. The chain of amino acids is also endowed with local terms that mimic the backbone stiffness. The numbers in the top left panel indicate particular proteins. These are: 1(1n11), 2(1cfc), 3(1hci), 4(^{10}FNfnIII), 5(1u4q), 6(1aj3), 7(B), 8(1ubq(48-N)), 9(1b6i), 10(1rsy), 11(^{13}FNfnIII), 12(^{12}FNfnIII), 14(^{3}TNfnIII), 15(1qjo(N-41)), 16(G), 17(^{1}FNfnIII), 18(I1), 19(I27), 20(1emb), 21(1emb(132-212)), 22(1emb(3-212)), 23(1ubq), 24(1nct), 25(1g1c), 26(L), 27(1emb(3-132)), 28(1vsc). B, L, and G denote barnase, protein L, and protein G respectively. The *solid line* corresponds to ε/Å = 67 pN and the lower *dotted line* to 111 pN. This figure also illustrates the best pick for ε (1.3 kcal/mol, the unit of force 67 pN) and this translation of the theoretical results into values in pN is shown in Table 3. Adapted from [113]

is a significant linkage dependence, i. e. F_{max} depends on the particular pulling geometry.

The first observation of this study is that the distribution of F_{max} across the PDB is non-Gaussian and has a pronounced tail at the high end (the end corresponding to "strong" proteins). The I27 domain of titin is a fairly strong mechanical protein, with its F_{max} being about twice as high as the average. However, there are a substantial number of proteins that are predicted to have twice the strength of this module. In 80% of cases, the maximum force is found at the early stages of stretching. The force was found not to depend on the number of amino acids (N) as one can find weak and strong proteins for any value

Protein Mechanics at the Single-Molecule Level, Table 3

Predicted and measured mechanical stability in proteins. Comparison between experimentally measured values F_{max} with theoretical predictions in the Lennard–Jones Go-like model. The theoretical results are averaged over 10 trajectories to account for several pathways if any. Proteins were pulled by their termini except for the ones in which the amino acids being pulled are indicated in brackets. The first column shows the PDB code and the second one the number of amino acids. Adapted from [113]

PDB	N	F_{max}[pN] -experiment	F_{max}[pN] -theory	
1tit	89	204 ± 30	144	I27*8
1nct	98	210 ± 10	161 ± 13	I54-I59
1g1c	97	127 ± 10	154 ± 13	I1 titin
1b6i	164	64 ± 30	80	T4 lysozyme(21-141)
1aj3	106	68 ± 20	82	spectrin R16
1qjo	80	15 ± 10	80	eE2lip3(N-C)
1qjo	40	177 ± 10	134	E2lip3(N-41)
1dqv	127	60 ± 15	100	calcium binding C2A
1rsy	127	60 ± 15	100 ± 13	calcium binding C2A
1byn	127	60 ± 15	94	calcium binding C2A
1cfc	148	< 20 pN	55 ± 20	calmodulin
1n11	33	37 ± 9	27	ankyrin*1
1bni	108	70 ± 15	94, 114	barnase/I27
1bnr	108	70 ± 15	70	barnase/I27
1bny	108	70 ± 15	74, 87	barnase/I27
1hz6	67	152 ± 10	235	protein L
1hz5	67	152 ± 10	188	protein L
2ptl	67	152 ± 10	147 ± 13	protein L
1ksr	100	45 ± 20	134 ± 20	DdFLN -4
2rn2	155	19 ± 10	121 ± 13	ribonuclease H
1ubq	76	230 ± 34	155	ubiquitin
1ubq	76	203 ± 35	155	ubiqutin(N-C)*9
1ubq	28	85 ± 20	60	ubiquitin(K48-C)*(2-7)
1emb	129	350 ± 30	345 ± 25	GFP(3-132)
1emb	219	130 ± 30	154, 288	GFP(3-212)
1emb	80	120 ± 30	147 ± 13	GFP(132-212)
1emb	235	104 ± 40	154 ± 13	GFP(N-C)
1fnf	94	75 ± 20	107, 121	^{10}FNfnIII
1ttf	94	75 ± 20	47, 80	^{10}FNfnIII
1ttg	94	75 ± 20	47, 67	^{10}FNfnIII
1fnh	92	124 ± 18	121	^{12}FNfnIII
1fnh	89	89 ± 18	94, 114	^{13}FNfnIII
1oww	93	125 ± 31	141 ± 13	^{1}FNfnIII
1ten	90	135 ± 40	114	^{3}TNfnIII
1pga	56	190 ± 20	161 ± 13	protein G
1gb1	56	190 ± 20	111 ± 13	protein G

of N. However, the larger the N the higher the probability that force was higher. The α-class of proteins usually showed weak forces. There are other specific correlations between types of structure, as described by the CATH classification (Class, Architecture, Topology, Homology), so that the strongest proteins correspond to two types of specific architectures: 30% α, β-rolls (ubiquitin-like) and 60% β-sandwiches (titin-like).

The unraveling process can be represented by "scenario diagrams" that show for how long a given native contact holds. Accumulation of such events results in the appearance of a force peak. Thus, it is possible to determine which contacts give rise to a force peak and then asses their dynamic impact by removing them in small sets and determining what effect this has on F_{max}. The contacts whose removal reduces F_{max} substantially correspond to

a mechanical clamp (i. e., a resistance point). In the top strongest proteins, 95% have mechanical clamps corresponding to long stretches of parallel β-strands that shear on pulling (like titin modules). Particularly large forces arise when such parallel strands are further stabilized by neighboring parts of the structure. However, there is also another type of mechanical clamps, those formed by antiparallel β-strands that are unstructured and even delocalized (where the clamping action sits in disjoint places).

Finally, there is a small but interesting group of proteins (about 300 examples so far) which contain simple knots in their native structures, usually trefoil knots. Knots are attractive topological features of proteins biologically relevant (e. g., toxicity), which are predicted to have important mechanical properties. These proteins could be pulled by a non-terminal amino acid to unravel the knot whereas pulling at the termini would tighten the knot. The latter case has been studied theoretically using the Go-like model for 20 proteins [112]. The simulations suggest that the knot tightening in a stretched protein proceeds through jumps, i. e. sudden displacements of the ends of the knot along the sequence. (The ends of a knot can be identified by removing the C^{α} atoms as long as the backbone does not intersect a triangle set by the atom under consideration and its two immediate sequential neighbors.) These jumps have definite lengths and together with the final location of a tightened knot they are specified by the local geometry of a protein chain (sharp turns are favored). The larger the size of a knot the larger the number of jumps observed before its final tightening. However, such jumps are not observed in the dynamics of knot motion on stretched polymers. In this case, the motion has a diffusive character and usually results in sliding of the knot off the chain. Another possible way to manipulate such a protein is to pull it to a certain extension and then release it abruptly. If the stretching stage lasts sufficiently long (so that several force peaks are observed and the knot gets tightened substantially) then the protein misfolds upon release and the knot ends up residing at metastable locations (this is predicted to happen for 2etl ($N = 223$), 1vho ($N = 157$), and 1v2x ($N = 191$) and in 80% of trajectories for 1o6d ($N = 147$). However, the knot in protein 1j85 ($N = 156$) was usually found (for most trajectories) to return reversibly to its native location.

Go-models can also be used to simulate the pulling of membrane proteins out of membranes (an important topic that is out of the scope of this review as it involves not only unfolding but also unbinding from the membrane components), resulting in a multipeak force pattern for bacteriorhodopsin that was remarkably similar to the experimental one [33]. All these results eagerly encourage the use of Go-like models for comparative studies in a variety of proteins.

Insights into Protein Folding from Forced Unfolding/Folding

Protein nanomechanics is multifaceted. One attractive element of our new capability to unfold/refold proteins by force is that it may provide new insights into the protein folding problem for any protein (i. e., mechanical proteins, "mechanical protein substrates", and non-mechanical proteins).

Ten years ago the only way to measure the stability of a protein was to change its physical (temperature or pressure) or chemical environment (using guanidinium chloride or urea, or varying the pH), and to monitor the loss of protein conformation by spectroscopic techniques in order to determine the change in the Gibbs free energy (ΔG). Most of these folding studies are typically done using chemical denaturants acting on untethered proteins (i. e. in solution). However, a considerable number of cytoskeletal and extracellular matrix proteins, as well as unfoldase substrates from chaperones, translocases and protein degradation machines, are likely to be subjected to mechanical forces and are tethered (Fig. 1). Thus, for those proteins at least, SMFS experiments (providing that the pulling geometry is the right one) may more closely mimic the physiological conditions in which they function in the cell.

Furthermore, the reaction coordinate in SMFS experiments has a well-defined physical meaning (i. e., protein length) and it is a "natural" one for some mechanical proteins. This is in clear contrast with the less well physically defined kinetic "m-values" used in chemical folding experiments. The m-value is defined as the derivative of the natural logarithm of the folding, or unfolding, rate constant with respect to the denaturant concentration and it measures the sensitivity of the rate of the process to denaturant concentration, being generally interpreted as a measure of the change in solvent exposure of the lateral chains of amino acid residues.

Thermodynamic comparisons of the ΔG of the process between both methods should in principle give identical results, if the entropic contribution of tethering the ends of the molecule is properly corrected for. At equilibrium, from a single molecule, it is possible to determine ΔG (which as a state function, it depends only on the initial and final stages of the process and can be obtained by integrating a reversible force-extension curve), the equilibrium constant, the reaction kinetics and their dependence on force [19]. Most single-molecule unfolding

experiments in proteins are performed under nonequilibrium conditions, where the molecule is pulled or relaxed at a faster rate than its spontaneous rate of equilibration (i. e. its molecular relaxation time). Consequently, a marked hysteresis is apparent (i. e. the extension and relaxation curves do not overlap) and at the same time the unfolding forces become speed dependent (Fig. 3). This indicates that not all the mechanical work carried out during the unfolding reaction is converted into a change in the free energy of the molecule (i. e., the efficiency of the process is less than 100%). It has recently been demonstrated that it is possible to recover the free energy of unfolding RNA molecules not only from near-equilibrium conditions [72] but also in far-from-equilibrium conditions [36]. The latter method is a promising tool to extract the equilibrium free energy of protein unfolding, since most protein unfolding reactions occur far from equilibrium.

In contrast, the kinetics of the reaction depends on the pathway and SMFS unfolding experiments impose a reaction coordinate to the molecule distinct to that of the bulk (chemical) experiments. Force acts along a single dimension in specific regions of the protein, typically the N- and C-termini, while conventional denaturants have a more global effect. As a result, the kinetic parameters of unfolding obtained by both methods may differ. In fact, a poor correlation was found between mechanical and chemical unfolding kinetic rates (a measure of kinetic stability) in a recent survey of 19 proteins [27]. In contrast, there was a remarkable agreement between these unfolding rate constants for modules I27 and I28 from the distal I-band region of titin [25,68]. This is noteworthy and raises interesting questions about the mechanical design of these particular domains. It has been shown that the A′G region is the only region in the I27 fold that is critical in both pathways and it is responsible for kinetic stability in both cases [44,69]. This may explain why the two unfolding rates are so close in this module.

While unfolding rates along the force-unfolding pathway can be easily estimated, refolding rates along the force-folding pathway cannot usually be measured because the high forces applied often prevent the folding of proteins. Using the regular length-clamp mode it has been possible to measure refolding rates during relaxation for several proteins, such as I27, projectin, ankyrin B, myosinII and filamin [18,25,67,104,105]. For example, ankyrin B repeats were shown to refold against force (20–25 pN) with complete reversibility, which makes them true elastic enthalpic springs [67]. The myosin II tail (a dimer that forms a coiled coil supramolecular structure) is also an elastic protein structure able to refold against forces of up to 30 pN, although the transition presented a little hysteresis, indicating that the process is not fully reversible [104].

Finally, force-clamp SMFS has been used to directly examine the mechanical folding pathways of I27, ubiquitin and projectin molecules [18,42,47,123] (Fig. 2c). In these experiments, the protein is first unfolded and extended at a high force and then relaxed at lower forces so that refolding can be monitored by measuring changes in the end-to-end length of the protein. Under these conditions, folding is marked by large fluctuations in the end-to-end length of the protein, which have been interpreted as folding of the chain through many continuous steps. By controlling the end-to-end length of a single protein with subnanometer resolution these studies provide us with a new perspective on how to analyze protein folding trajectories (Fig. 8). It should be noted that refolding using the force-clamp mode may proceed along different pathways than those in the absence of such restraints, as shown for ubiquitin using a Go-like model [30]. Finally, optical tweezers (that use lower spring constants and hence, correspondingly lower loading rates than AFM) have been used to study the mechanical unfolding/folding of RNase H and maltose binding protein [8,28]. These studies are providing new molecular insights into folding intermediates, the energy landscape of the process, and how chaperone interactions affect protein folding pathways at the single molecule level.

Future Directions

It has been just a decade since the first individual protein molecule was pulled by SMFS. This impressive feat has since grown into a new discipline providing a wealth of information on the molecular elasticity of proteins, a fundamental property for many biological processes. This new methodology is unveiling the mechanical properties of more and more proteins, and it is providing new insights into the problem of protein folding.

However, only a few protein folds have been analyzed to date and many improvements are still required, such as more specific functionalization methods for immobilization of soluble proteins in order to improve the efficiency and sample control of these experiments. These immobilization methods should ideally be compatible with a *quasi*-simultaneous imaging of proteins in the same sample [119]. We also need single-molecule reporters to conduct more reliable studies on intermolecular interactions between proteins.

It would be also very interesting to use optical tweezers and the biomembrane force probe instrument (which allow even lower loading rates to be applied) on some of the proteins already studied by AFM and to compare the

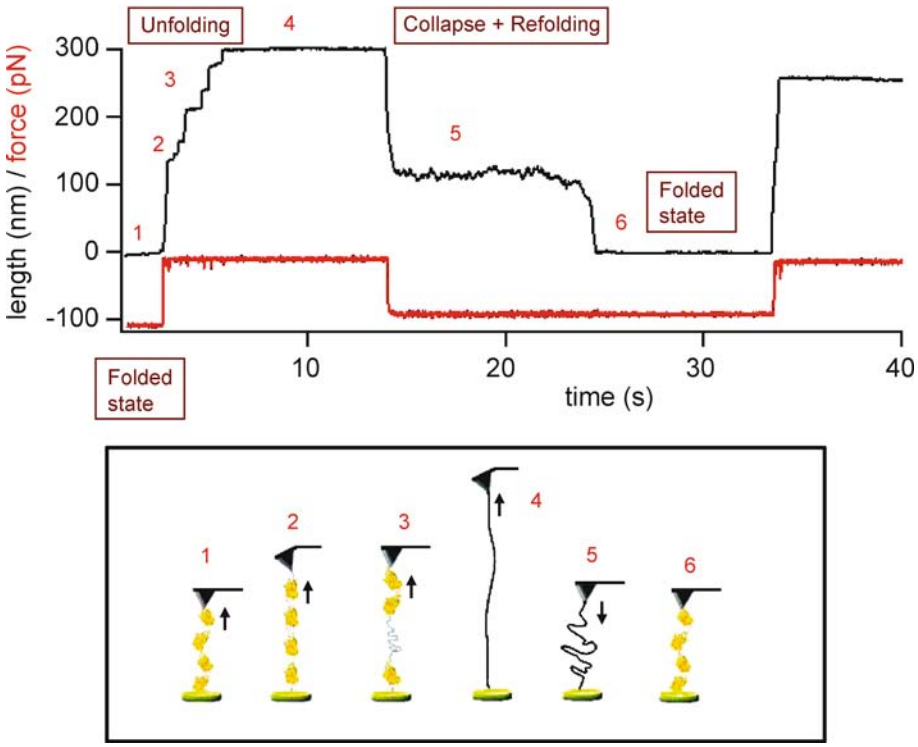

Protein Mechanics at the Single-Molecule Level, Figure 8

Collapse of unfolded titin-like domains under force (mechanical refolding). A titin like protein (projectin) molecule is first unfolded and extended at a high force (97 pN) using force-clamp SMFS. We observed 10 unfolding events. There was an initial large step elongation of ∼100 nm upon application of force and this initial phase most likely corresponds to the length of the folded polypeptide chain plus a few already unfolded domains. Then after ∼12 s the protein was relaxed to a force of 15 pN and before the protein reached its fully collapsed state there was a dramatic increase in the noise level with length fluctuations of up to 10 nm peak-to-peak. The source of this noise is not clear, but the phenomenon may reflect the transient formation of secondary structures or intermediate folded conformations. There are three main phases: i) a fast phase (< 200 ms) corresponding to the elastic recoil of the unfolded polypeptide chain and accounting for ∼60% of the unfolded length of the protein; ii) a slow phase (∼1–8 nm/s) characterized by large fluctuations in end-to-end length (up to 10 nm in this example); and iii) again a fast phase (> 100 nm/s) that corresponds to the final collapse of the polypeptide chain to its folded length. In order to test whether the domains are folded, the protein was unfolded by applying a second stretching pulse to 97 pN after 30 s. This experiment demonstrates that titin domains can refold under force; this suggests that titin-like proteins could function according to a folding-based-spring mechanism. (After [18])

results obtained at these low loading rates. For the sake of comparison researchers should also follow a standard set of experimental conditions for SMFS experiments [12,97], as recently proposed in the field of chemical protein folding [78]. Moreover, there is a need for a sensor that could report forces inside the living cell and important advances along these lines were recently reported [57,99,108].

Single-molecule mechanical techniques are still in their infancy, but they are maturing fast. These techniques are providing us more and more fundamental information on the structure and function of proteins. Accordingly, they are becoming an indispensable tool to understand how proteins fold and work in real time. With the newfound capacity to manipulate and look at the "secret life" of single molecules, we should be prepared for many surprises from the mechanochemistry of proteins. Through the information unveiled by these techniques we are entering a new and exciting time in biology which, in combination with the knowledge generated in this proteomic era, is likely to move us closer to understanding the logic behind protein design.

Acknowledgments

This work was funded by grants from the Spanish Ministry of Science and Education (BIO2007-67116), the Consejería de Educación of the Madrid Community (S-0505/MAT/0283), and the Spanish Research Council (200620F00) to M.C.-V., from the NIH (R01DK073394), the John Sealy Memorial Endowment Fund for Biomedical

Research, and the Polycystic Kidney Foundation (116a2r) to A.F.O., and from the Ministry of Science and Higher Education (N N202 0852 33) to M.C. We apologize to all researchers whose pioneering work was not cited due to limitations of space.

Bibliography

Primary Literature

1. Abe H, Go N (1981) Noninteracting local-structure model of folding and unfolding transition in globular proteins, II, Application to two-dimensional lattice proteins. Biopolymers 20:1013–1031
2. Abu-Lail NI, Ohashi T, Clark RL, Erickson HP, Zauscher S (2005) Understanding the elasticity of fibronectin fibrils: unfolding strengths of FN-III and GFP domains measured by single-molecule force spectroscopy. Matrix Biol 25:175–184
3. Ainavarapu SR, Li L, Badilla CL, Fernandez JM (2005) Ligand binding modulates the mechanical stability of dihydrofolate reductase. Biophys J 89:3337–3344
4. Ainavarapu SR, Brujic J, Huang HH, Wiita AP, Lu H, Li L, Walther KA, Carrion-Vazquez M, Li H, Fernandez JM (2007) Contour length and refolding rate of a small protein controlled by engineered disulfide bonds. Biophys J 92:225–233
5. Alberts B (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell 92:291–294
6. Bao G, Suresh S (2003) Cell and molecular mechanics of biological materials. Nat Mater 2:715–725
7. Batey S, Randles LG, Steward A, Clarke J (2005) Cooperative folding in a multi-domain protein. J Mol Biol 349:1045–1059
8. Bechtluft P, van Leeuwen RG, Tyreman M, Tomkiewicz D, Nouwen N, Tepper HL, Driessen AJ, Tans SJ (2007) Direct observation of chaperone-induced changes in a protein folding pathway. Science 318:1458–1461
9. Bell GI (1978) Models for the specific adhesion of cells to cells. Science 200:618–627
10. Best RB, Li B, Steward A, Daggett V, Clarke J (2001) Can non-mechanical proteins withstand force? Stretching barnase by atomic force microscopy and molecular dynamics simulation. Biophys J 81:2344–2356
11. Best RB, Fowler SB, Toca-Herrera JL, Clarke J (2002) A simple method for proving the mechanical unfolding pathway of proteins in detail. Proc Natl Acad Sci USA 99:12143–12148
12. Best RB, Brockwell DJ, Toca-Herrera JL, Blake AW, Smith DA, Radford SE, Clarke J (2003) Force mode atomic force microscopy as a tool for protein folding studies. Anal Chim Acta 479:87–105
13. Best RB, Fowler SB, Toca-Herrera JL, Steward A, Paci E, Clarke J (2003) Mechanical unfolding of a titin Ig domain: Structure of transition state revealed by combining atomic force microscopy, protein engineering and molecular dynamics simulations. J Mol Biol 330:867–877
14. Bhasin N, Carl P, Harper S, Feng G, Lu H, Speicher DW, Discher DE (2004) Chemistry on a single protein, vascular cell adhesion molecule-1, during forced unfolding. J Biol Chem 279:45865–45874
15. Binnig G, Quate CF, Gerber C (1986) Atomic force microscope. Phys Rev Lett 56:930–933
16. Brockwell DJ, Beddard GS, Clarkson J, Zinober RC, Blake AW, Trinick J, Olmsted PD, Smith DA, Radford SE (2002) The effect of core destabilization on the mechanical resistance of I27. Biophys J 83:458–472
17. Brockwell DJ, Paci E, Zinober RC, Beddard GS, Olmsted PD, Smith DA, Perham RN, Radford SE (2003) Pulling geometry defines the mechanical resistance of a beta-sheet protein. Nat Struct Biol 10:731–737
18. Bullard B, Garcia T, Benes V, Leake MC, Linke WA, Oberhauser AF (2006) The molecular elasticity of the insect flight muscle proteins projectin and kettin. Proc Natl Acad Sci USA 103:4451–4456
19. Bustamante C, Chemla YR, Forde NR, Izhaky D (2004) Mechanical processes in biochemistry. Annu Rev Biochem 73:705–748
20. Bustamante C, Macosko JC, Wuite GJ (2000) Grabbing the cat by the tail: Manipulating molecules one by one. Nat Rev Mol Cell Biol 1:130–136
21. Cao Y, Li H (2007) Polyprotein of GB1 is an ideal artificial elastomeric protein. Nat Mater 6:109–114
22. Carl P, Kwok CH, Manderson G, Speicher DW, Discher DE (2001) Forced unfolding modulated by disulfide bonds in the Ig domains of a cell adhesion molecule. Proc Natl Acad Sci USA 98:1565–1570
23. Carrion-Vazquez M, Li H, Lu H, Marszalek PE, Oberhauser AF, Fernandez JM (2003) The mechanical stability of ubiquitin is linkage dependent. Nat Struct Biol 10:738–743
24. Carrion-Vazquez M, Marszalek PE, Oberhauser AF, Fernandez JM (1999) Atomic force microscopy captures length phenotypes in single proteins. Proc Nat Acad Sci USA 96:11288–11292
25. Carrion-Vazquez M, Oberhauser AF, Fowler SB, Marszalek PE, Broedel SE, Clarke J, Fernandez JM (1999) Mechanical and chemical unfolding of a single protein: A comparison. Proc Nat Acad Sci USA 96:3694–3699
26. Carrion-Vazquez M, Oberhauser AF, Fisher TE, Marszalek PE, Li H, Fernandez JM (2000) Mechanical design of proteins studied by single-molecule force spectroscopy and protein engineering. Prog Biophys Mol Biol 74:63–91
27. Carrión-Vázquez M, Oberhauser AF, Díez H, Hervás R, Oroz J, Fernández J, Martínez-Martín D (2006) Protein nanomechanics, as studied by AFM single-molecule force spectroscopy. In: Arrondo JLR, Alonso A (eds) Advanced Techniques in Biophysics. Springer, Berlin, pp 163–245
28. Cecconi C, Shank EA, Bustamante C, Marqusee S (2005) Direct observation of the three-state folding of a single protein molecule. Science 309:2057–2060
29. Cieplak M (2005) Mechanical stretching of proteins: calmodulin and titin. Physica A 352:28–42
30. Cieplak M, Szymczak P (2006) Protein folding in a force clamp. J Chem Phys 124:194901
31. Cieplak M, Hoang TX, Robbins MO (2002) Folding and stretching in a Go-like model of titin. Proteins 49:114–124
32. Cieplak M, Hoang TX, Robbins MO (2004) Thermal effects in stretching of Go-like models of titin and secondary structures. Proteins: Struct Funct Bio 56:285–297
33. Cieplak M, Filipek S, Janovjak H, Krzysko KA (2006) Pulling single bacteriorhodopsin out of a membrane: Comparison of simulation and experiment. Biochim Biophys Acta 1758:537–544

34. Clausen-Schaumann H, Seitz M, Krautbauer R, Gaub HE (2000) Force spectroscopy with single bio-molecules. Curr Opin Chem Biol 4:524–530
35. Clementi C, Nymeyer H, Onuchic JN (2000) Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. J Mol Biol 298:937–953
36. Collin D, Ritort F, Jarzynski C, Smith SB, Tinoco I Jr, Bustamante C (2005) Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies. Nature 437:231–234
37. Dietz H, Rief M (2004) Exploring the energy landscape of GFP by single-molecule mechanical experiments. Proc Natl Acad Sci USA 101:16192–16197
38. Dietz H, Rief M (2006) Protein structure by mechanical triangulation. Proc Natl Acad Sci USA 103:1244–1247
39. Dietz H, Berkemeier F, Bertz M, Rief M (2006) Anisotropic deformation response of single protein molecules. Proc Natl Acad Sci USA 103:12724–12728
40. Evans E, Ritchie K (1997) Dynamic strength of molecular adhesion bonds. Biophys J 72:1541–1555
41. Evans E, Ritchie K, Merkel R (1995) Sensitive force technique to probe molecular adhesion and structural linkages at biological interfaces. Biophys J 68:2580–2587
42. Fernandez JM, Li H (2004) Force-clamp spectroscopy monitors the folding trajectory of a single protein. Science 303:1674–1678
43. Fersht AR, Matouschek A, Serrano L (1992) The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. J Mol Biol 224:771–782
44. Forman JR, Clarke J (2007) Mechanical unfolding of proteins: insights into biology, structure and folding. Curr Opin Struct Biol 17:58–66
45. Fowler SB, Best RB, Toca-Herrera JL, Rutherford TJ, Steward A, Paci E, Karplus M, Clarke J (2002) Mechanical unfolding of a titin Ig domain: Structure of unfolding intermediate revealed by combining AFM, molecular dynamics simulations, NMR and protein engineering. J Mol Biol 322:841–849
46. Gao M, Sotomayor M, Villa E, Lee EH, Schulten K (2006) Molecular mechanisms of cellular mechanics. Phys Chem Chem Phys 8:3692–3706
47. Garcia-Manyes S, Brujic J, Badilla CL, Fernandez JM (2007) Force-clamp spectroscopy of single-protein monomers reveals the individual unfolding and folding pathways of I27 and ubiquitin. Biophys J 93:2436–2446
48. Geiger B, Bershadsky A (2002) Exploring the neighborhood: adhesion-coupled cell mechanosensors. Cell 110:139–142
49. Giannone G, Sheetz MP (2006) Substrate rigidity and force define form through tyrosine phosphatase and kinase pathways. Trends Cell Biol 16:213–223
50. Grubmuller H, Heymann B, Tavan P (1996) Ligand binding: Molecular mechanics calculation of the streptavidin biotin rupture force. Science 271:997–999
51. Hinterdorfer P (2002) Molecular recognition studies using the atomic force microscope. Methods Cell Biol 68:115–139
52. Hoang TX, Cieplak M (2000) Molecular dynamics of folding of secondary. structures in Go-like models of proteins. J Chem Phys 112:6851–6862
53. Howard J (2001) Mechanics of Motor Proteins and the Cytoskeleton. Sinauer Associates, Sunderland
54. Huang L, Kirmizialtin S, Makarov DE (2005) Computer simulations of the translocation and unfolding of a protein pulled mechanically through a pore. J Chem Phys 123:124903
55. Huxley H, Hanson J (1954) Changes in the cross-striations of muscle during contraction and stretch and their structural interpretation. Nature 173:973–976
56. Hyeon CB, Thirumalai D (2003) Can energy landscape roughness of proteins and RNA be measured by using mechanical unfolding experiments? Proc Natl Acad Sci USA 100:10249–10253
57. Johnson CP, Tang HY, Carag C, Speicher DW, Discher DE (2007) Forced unfolding of proteins within cells. Science 317:663–666
58. Junker JP, Hell K, Schlierf M, Neupert W, Rief M (2005) Influence of substrate binding on the mechanical stability of mouse dihydrofolate reductase. Biophys J 89:L46–48
59. Karanicolas J, Brooks CL III (2002) The origins of asymmetry in the folding transition states of protein L and protein G. Prot Sci 11:2351–2361
60. Kedrov A, Janovjak H, Sapra KT, Müller DJ (2007) Deciphering molecular interactions of native membrane proteins by single-molecule force spectroscopy. Annu Rev Biophys Biomol Struct 36:233–260
61. Kellermayer MS (2005) Visualizing and manipulating individual protein molecules. Physiol Meas 26:R119–R153
62. Kellermayer MS, Smith SB, Granzier HL, Bustamante C (1997) Folding-unfolding transitions in single titin molecules characterized with laser tweezers. Science 276:1112–1116
63. Klimov DK, Thirumalai D (2000) Native topology determines force-induced unfolding pathways in globular proteins. Proc Natl Acad Sci USA 97:7254–7259
64. Law R, Carl P, Harper S, Dalhaimer P, Speicher DW, Discher DE (2003) Cooperativity in forced unfolding of tandem spectrin repeats. Biophys J 84:533–544
65. Leake MC, Grutzner A, Kruger M, Linke WA (2006) Mechanical properties of cardiac titin's N2B-region by single-molecule atomic force spectroscopy. J Struct Biol 155:263–272
66. Lee C-K, Wang Y-M, Huang L-S, Lin S (2007) Atomic force microscopy: Determination of unbinding force, off rate and energy barrier for protein-ligand interaction. Micron 38:446–461
67. Lee G, Abdi K, Jiang Y, Michaely P, Bennett V, Marszalek PE (2006) Nanospring behaviour of ankyrin repeats. Nature 440:246–249
68. Li H, Oberhauser AF, Fowler SB, Clarke J, Fernandez JM (2000) Atomic force microscopy reveals the mechanical design of a modular protein. Proc Natl Acad Sci USA 97:6527–6531
69. Li H, Carrion-Vazquez M, Oberhauser AF, Marszalek PE, Fernandez JM (2000) Point mutations alter the mechanical stability of immunoglobulin modules. Nat Struct Biol 7:1117–1120
70. Li H, Linke WA, Oberhauser AF, Carrion-Vazquez M, Kerkvliet JG, Lu H, Marszalek PE, Fernandez JM (2002) Reverse engineering of the giant muscle protein titin. Nature 418:998–1002
71. Li H, Oberhauser AF, Redick SD, Carrion-Vazquez M, Erickson HP, Fernandez JM (2001) Multiple conformations of PEVK proteins detected by single-molecule techniques. Proc Natl Acad Sci USA 98:10682–10686
72. Liphardt J, Dumont S, Smith SB, Tinoco I Jr, Bustamante C (2002) Equilibrium information from nonequilibrium mea-

surements in an experimental test of Jarzynski's equality. Science 296:1832–1853
73. Lu H, Isralewitz B, Krammer A,Vogel V, Schulten K (1998) Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation. Biophys J 75:662–671
74. Makarov DE, Hansma PK, and Metiu H (2001) Kinetic Monte Carlo simulation of titin unfolding. J Chem Phys 114:9663–9673
75. Marszalek PE, Lu H, Li H, Carrion-Vazquez M, Oberhauser AF, Schulten K, Fernandez JM (1999) Mechanical unfolding intermediates in titin modules. Nature 402:100–103
76. Maruyama K, Kimura S, Ohashi K, Kuwano Y (1981) Connectin, an elastic protein of muscle. Identification of "titin" with connectin. J Biochem (Tokyo) 89:701–709
77. Matouschek A (2003) Protein unfolding-an important process *in vivo*? Curr Opin Struct Biol 13:98–109
78. Maxwell KL, Wildes D, Zarrine-Afsar A, de los Rios MA, Brown AG, et al (2005) Protein folding: Defining a "standard" set of experimental conditions and a preliminary kinetic data set of two-state proteins. Protein Sci 14:602–616
79. Mehta AD, Rief M, Spudich JA (1999) Biomechanics, one molecule at a time. J Biol Chem 274:14517–14520
80. Merkel R (2001) Force spectroscopy on single passive biomolecules and single biomolecular bonds. Phys Rep 346:344–385
81. Miller E, Garcia T, Hultgren S, Oberhauser AF (2006) The mechanical properties of *E. coli* type 1 pili measured by atomic force microscopy techniques. Biophys J 91:3848–3856
82. Müller DJ, Engel A (2007) Atomic force microscopy and spectroscopy of native membrane proteins. Nat Protoc 2:2191–2197
83. Neher E, Sakmann B (1976) Single-channel currents recorded from membrane of denervated frog muscle fibres. Nature 260:799–802
84. Neuman KC, Lionnet T, Allemand J-F (2007) Single-molecule micromanipulation techniques. Annu Rev Mater Res 37:33–67
85. Ng SP, Rounsevell RWS, Steward A, Geierhaas CD, Williams PM, Paci E, Clarke J (2005) Mechanical unfolding of TNfn3: The unfolding pathway of a fnIII domain probed by protein engineering, AFM and MD simulation. J Mol Biol 350:776–789
86. Oberhauser AF, Carrión-Vázquez M (2008) Mechanical biochemistry of proteins one molecule at a time. J Biol Chem 283:6617–6621
87. Oberhauser AF, Marszalek PE, Erickson HP, Fernandez JM (1998) The molecular elasticity of tenascin, an extracellular matrix protein. Nature 393:181–185
88. Oberhauser AF, Marszalek PE, Carrion-Vazquez M, Fernandez JM (1999) Single protein misfolding events captured by atomic force microscopy. Nat Struct Biol 6:1025–1028
89. Oberhauser AF, Badilla-Fernandez C, Carrion-Vazquez M, Fernandez JM (2002) The mechanical hierarchies of fibronectin observed with single-molecule AFM. J Mol Biol 319:433–447
90. Perez-Jimenez R, Garcia-Manyes S, Ainavarapu SR, Fernandez JM (2006) Mechanical unfolding pathways of the enhanced yellow fluorescent protein revealed by single molecule force spectroscopy. J Biol Chem 281:40010–40014
91. Prakash S, Matouschek A (2004) Protein unfolding in the cell. Trends Biochem Sci 29:593–600
92. Qian F, Wei W, Germino G, Oberhauser A (2005) The nanomechanics of polycystin-1 extracellular region. J Biol Chem 280:40723–40730
93. Rief M, Grubmuller H (2002) Force spectroscopy of single biomolecules. Chemphyschem 3:255–261
94. Rief M, Gautel M, Oesterhelt F, Fernandez JM, Gaub HE (1997) Reversible unfolding of individual titin immunoglobulin domains by AFM. Science 276:1109–1112
95. Rief M, Pascual J, Saraste M, Gaub HE (1999) Single molecule force spectroscopy of spectrin repeats: low unfolding forces in helix bundles. J Mol Biol 286:553–561
96. Ritort F (2006) Single-molecule experiments in biological physics: Methods and applications. J Phys: Condens Matter 18:R531–R583
97. Rounsevell R, Forman JR, Clarke J (2004) Atomic force microscopy: mechanical unfolding of proteins. Methods 34:100–111
98. Samorì B, Zuccheri G, Baschieri R (2005) Protein unfolding and refolding under force: Methodologies for nanomechanics. Chemphyschem 6:29–34
99. Sarkar A, Robertson RB, Fernandez J M (2004) Simultaneous atomic force microscope and fluorescence measurements of protein unfolding using a calibrated evanescent wave. Proc Natl Acad Sci USA 101:12882–12886
100. Sarkar A, Caamano S, Fernandez JM (2005) The elasticity of individual titin PEVK exons measured by single molecule atomic force microscopy. J Biol Chem 280:6261–6264
101. Sato T, Esaki M, Fernandez JM, Endo T (2005) Comparison of the protein-unfolding pathways between mitochondrial protein import and atomic-force microscopy measurements. Proc Natl Acad Sci USA 102:17999–18004
102. Sauer RT, Bolon DN, Burton BM, Burton RE, Flynn JM, Grant RA, Hersch GL, Joshi SA, Kenniston JA, Levchenko I, Neher SB, Oakes E, Siddiqui SM, Wah DA, Baker TA (2004) Sculpting the proteome with AAA(+) proteases and disassembly machines. Cell 119:9–18
103. Schlierf M, Li H, Fernandez JM (2004) The unfolding kinetics of ubiquitin captured with single-molecule force-clamp techniques. Proc Natl Acad Sci USA 101:7299–7304
104. Schwaiger I, Sattler C, Hostetter DR, Rief M (2002) The myosin coiled-coil is a truly elastic protein structure. Nat Mater 1:232–235
105. Schwaiger I, Schleicher M, Noegel AA, Rief M (2005) The folding pathway of a fast-folding immunoglobulin domain revealed by single-molecule mechanical experiments. EMBO Rep 6:1–6
106. Sharma D, Perisic O, Peng Q, Cao Y, Lam C, Lu H, Li H (2007) Single-molecule force spectroscopy reveals a mechanically stable protein fold and the rational tuning of its mechanical stability. Proc Natl Acad Sci USA104:9278–9283
107. Shtilerman M, Lorimer GH, Englander SW (1999) Chaperonin function: folding by forced unfolding. Science 284:822–825
108. Smith ML, Gourdon D, Little WC, Kubow KE, Eguiluz RA, Luna-Morris S, Vogel V (2007) Force-induced unfolding of fibronectin in the extracellular matrix of living cells. PLoS Biol 5:e268
109. Sotomayor M, Schulten K (2007) Single-molecule experiments *in vitro* and *in silico*. Science 316:1144–1148
110. Sulkowska JI, Cieplak M (2007) Mechanical stretching of proteins- A theoretical survey of the Protein Data Bank – a topical review. J Phys Cond Mat 19:283201

111. Sulkowska JI, Cieplak M (2008) Stretching to understand proteins-A survey of the Protein Data Bank. Biophys J 94:6–13
112. Sulkowska JI, Sulkowski P, Szymczak P, Cieplak M (2008) Tightening of knots in proteins. Phys Rev Lett 100:058106
113. Sulkowska JI, Cieplak M (2008) Selection of optimal variants of Go-like models of protein through studies of stretching. Biophys J 95:3174–3191
114. Takada S (1999) Go-ing for the prediction of protein folding mechanisms. Proc Natl Acad Sci USA 96:11609–11700
115. Tian P, Andricioaei I (2005) Repetitive pulling catalyzes co-translocational unfolding of barnase during import through a mitochondrial pore. J Mol Biol 350:1017–1034
116. Trylska J, Tozzini V, McCammon JA (2005) Exploring global motions and correlations in the ribosome. Biophys J 89:1455–1463
117. Tskhovrebova L, Trinick J (2003) Titin: properties and family relationships. Nat Rev Mol Cell Biol 4:679–789
118. Urry DW, Hugel T, Seitz M, Gaub HE, Sheiba L, Dea J, Xu J, Parker T (2002) Elastin: a representative ideal protein elastomer. Philos Trans R Soc Lond B Biol Sci 357:169–184
119. Valbuena A, Oroz J, Vera AM, Gimeno A, Gómez-Herrero J, Carrión-Vázquez M (2007) *Quasi*-simultaneous imaging/pulling analysis of polyprotein molecules by AFM. Rev Sci Instrum 78:113707
120. Veitshan T, Klimov D, Thirumalai D (1997) Protein folding kinetics: Timescales, pathways and energy landscapes in terms of sequence-dependent properties. Folding Des 2:1–22
121. Vogel V (2006) Mechanotransduction involving multimodular proteins: converting force into biochemical signals. Annu Rev Biophys Biomol Struct 35:459–488
122. Vogel V, Sheetz M (2006) Local force and geometry sensing regulate cell functions. Nat Rev Mol Cell Biol 7:265–275
123. Walther KA, Grater F, Dougan L, Badilla CL, Berne BJ, Fernandez JM (2007) Signatures of hydrophobic collapse in extended proteins captured with force spectroscopy. Proc Natl Acad Sci USA 104:7916–7921
124. Wang K, McClure J, Tu A (1979) Titin: major myofibrilar components of striated muscle. Proc Natl Acad Sci USA 76:3698–3702
125. Weisel JW, Shuman H, Litvinov RI (2003) Protein-protein unbinding induced by force: Single-molecule studies. Curr Opin Struct Biol 13:227–235
126. West DK, Brockwell DJ, Olmsted PD, Radford SE, Paci E (2006) Mechanical resistance of proteins explained using simple molecular models. Biophys J 90:287–297
127. West DK, Brockwell DJ, Paci E (2006) Prediction of the translocation kinetics of a protein from its mechanical properties. Biophys J 91:L51–53
128. Wiita AP, Ainavarapu SR, Huang HH, Fernandez JM (2006) Force-dependent chemical kinetics of disulfide bond reduction observed with single-molecule techniques. Proc Natl Acad Sci USA 103:7222–7227
129. Wiita AP, Perez-Jimenez R, Walther KA, Grater F, Berne BJ, Holmgren A, Sanchez-Ruiz JM, Fernandez JM (2007) Probing the chemistry of thioredoxin catalysis with force. Nature 450:124–127
130. Wilcox AJ, Choy J, Bustamante C, Matouschek A (2005) Effect of protein structure on mitochondrial import. Proc Natl Acad Sci USA 102:15435–15440
131. Yang G, Cecconi C, Baase WA, Vetter IR, Breyer WA, Haack JA, Matthews BW, Dahlquist FW, Bustamante C (2000) Solid-state synthesis and mechanical unfolding of polymers of T4 lysozyme. Proc Natl Acad Sci USA 97:139–144
132. Zhuang X, Rief M (2003) Single-molecule folding. Curr Opin Struct Biol 13:88–97
133. Zlatanova J, Lindsay SM, Leuba SH (2000) Single molecule force spectroscopy in biology using the atomic force microscope. Prog Biophys Mol Biol 74:37–61
134. Zlatanova J, van Holde K (2006) Single-molecule biology: What is it and how does it work? Mol Cell 24:317–329

Books and Reviews

Boal D (2002) Mechanics of the Cell. Cambridge University Press, Cambridge
Brande C, Tooze J (1999) Introduction to Protein Structure, 2nd edn. Garland Publishing, New York
Goodsell DS (2004) Bionanotechnology: Lessons from Nature. Wiley, Hoboken
Grosberg AY, Khokhlov AR (1997) Giant Molecules: Here, There and Everywere. Academic Press, San Diego
Grubmüller H, Schulten K (eds) (2007) Advances in molecular dynamics simulations. J Struct Biol 157(special issue):443–615
Leuba SH, Zlatanova J (2000) Biology at the Single-Molecule Level. Pergamon Press, Oxford
Wainwright SA, Biggs WD, Currey JD, Gosline JM (1976) Mechanical Design in Organisms. Princeton University Press, Princeton

Public Policy, System Dynamics Applications to

David F. Andersen[1], Eliot Rich[2], Roderick MacDonald[3]
[1] Rockefeller College of Public Affairs and Policy, University at Albany, Albany, USA
[2] School of Business, University at Albany, Albany, USA
[3] Initiative for System Dynamics in the Public Sector, University at Albany, Albany, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Medical Malpractice: A System Dynamics and Public Policy Vignette
What Is System Dynamics Modeling?
How Is System Dynamics Used to Support Public Policy and Management?
System Dynamics and Models: A Range of Analytic Scope and Products
What Are the Arenas in Which System Dynamics Models Are Used?

What Are some of the Substantive Areas Where System Dynamics Has Been Applied?
Evaluating the Effectiveness of System Dynamics Models in Supporting the Public Policy Process
Summary: System Dynamics – A Powerful Tool to Support Public Policy
Future Directions
Bibliography

Glossary

Causal loop diagram A diagrammatic artifact that captures the causal model and feedback structure underlying a problem situation. Commonly used as a first-cut tool to identify major stakeholder concerns and interactions. These diagrams are often precursors to formal models.

Dynamic modeling Formal examination of the behavior of a system over time. Contrast with point-estimation, which attempts to predict an average outcome.

Feedback A relationship where two or more variables are linked over time so that the influence of one variable on a second will later affect the state of the first. If the influence is such as to increase the state of the first over time, the feedback is termed reinforcing. If the influence is such as to decrease the state of the first, it is termed balancing.

Formal model The representation of a system structure in mathematical form. Contrast with causal model, which represents structure without the underlying mathematics.

Mental model The representation of a problem's structure as possessed by an expert in a particular domain. Mental models are often intangible until explicated by the expert.

Public policy Any and all actions or non-actions, decisions or non-decisions taken by government, at all levels, to address problems. These actions, non-actions, decisions or non-decisions are implemented through laws, regulations and the allocation of resources.

Group model building (GMB) An approach to problem definition that asks multiple experts and major stakeholders to provide collective insights into the structure and behavior of a system through facilitated exercises and artifacts. GMB is often used to explicate the contrasting mental models of stakeholders.

Stakeholder An individual or group that has significant interest or influence over a policy problem.

System dynamics An analytic approach to problem definition and solution that focuses on endogenous variables linked through feedback, information and material delays, and non-linear relationships. The structure of these linkages determines the behavior of the modeled system.

Definition of the Subject

System dynamics is an approach to problem understanding and solution. It captures the complexity of real-world problems through the explication of feedback among endogenous variables. This feedback, and the delays that accompany it, often drive public sector programs towards unanticipated or unsatisfactory results. Through formal and informal modeling, System Dynamics-based analysis explicates and opens these feedback structures to discussion, debate and consensus building necessary for successful public sector policymaking.

Introduction

In the 50 years since its founding, System Dynamics has contributed to public policy thought in a number of areas. Major works, such as *Urban Dynamics* [35] and *Limits to Growth* [61] have sparked controversy and debate. Other works in the domains of military policy, illegal drugs, welfare reform, health care, international development, and education have provided deep insight into complex social problems. The perspective of System Dynamics, with its emphasis on feedback, changes over time, and the role of information delays, helps inform policy makers about the intended and unintended consequences of their choices. The System Dynamics method includes a problem-oriented focus and the accommodation of multiple stakeholders, both crucial to the development of sound policy. Through the use of formal simulation, decision makers may also use System Dynamics models to consider the effects of their choices on short- and long-term outcomes. We illustrate this process with real life examples, followed by a review of the features of System Dynamics as they relate to public policy issues. We then describe the conjunction of System Dynamics and Group Model Building as a mechanism for policy ideation and review. We identify some of the historical and current uses of System Dynamics in the public sector, and discuss techniques for evaluating its effects on policy and organizations.

Medical Malpractice: A System Dynamics and Public Policy Vignette

The year was 1987 and New York's medical malpractice insurance system was in a state of crisis. Fueled by unprecedented levels of litigation, total settlements were soaring as were the malpractice insurance rates charged to hospitals

and physicians. Obstetricians stopped taking on new patients. Doctors threatened to or actually did leave the state. Commercial insurance carriers had stopped underwriting malpractice insurance policies, leaving state-sponsored risk pools as the only option. The Governor and the Legislature were under pressure to find a solution and to find it soon. At the center of this quandary was the state's Insurance Department, the agency responsible for regulating and setting rates for the state's insurance pools. The agency's head found himself in just the kind of media hot seat one seeks to avoid in the public service.

An in-house SWAT team of actuaries, lawyers, and analysts had been working to present a fiscally sound and politically viable set of options for the Agency to consider and recommend to the legislature. They had been working with a team of System Dynamics modelers to gain better understanding of the root causes of the crisis. Working as a group, they had laid out a whole-system view of the key forces driving malpractice premiums in New York State. Their simulation model, forged in the crucible of group consensus, portrayed the various options on a "level playing field," each option being analyzed using a consistent set of operating assumptions. One option stood out for its ability to offer immediate malpractice insurance premium relief, virtually insuring a rapid resolution to the current crisis. An actuarial restructuring of future liabilities arising from future possible lawsuits relieved immediate pressure on available reserve funds. Upward pressure on premiums would vanish; a showdown in the legislature would be averted. Obviously, the Commissioner was interested in this option–who would not be?

"But what happens in the later years, after our crisis is solved?" he asked. As the team pored over the simulation model, they found that today's solution sowed the seeds for tomorrow's problems. Ten, fifteen, or maybe more years into the future, the deferred liabilities piled up in the system creating a secondary crisis, quite literally a second crisis caused by the resolution of the first crisis.

"Take that option off the table – it creates an unacceptable future," was the Commissioner's snap judgment. At that moment a politically appointed official had summarily dismissed a viable and politically astute "silver bullet" cure to a current quandary because he was thinking dynamically, considering both short-term and long-term effects of policy.

The fascinating point of the medical malpractice vignette is that the option taken off the table was indeed, in the short run, a "silver bullet" to the immediate crisis. The System Dynamics model projected that the solution's unraveling would occur long after the present Commissioner's career was over, as well as after the elected life span of the Governor who had appointed him and the legislators whose votes would be needed to implement the solution. His decision did not define the current problem solely in terms of the current constellation of stakeholders at the negotiations, each with their particular interests and points of view. His dynamic thinking posed the current problem as the result of a system of forces that had accumulated in the past. Symmetrically, his dynamic thinking looked ahead in an attempt to forecast what would be the future dynamic consequences of each option. Might today's solution become tomorrow's problem?

This way of thinking supported by System Dynamics modeling invites speculation about long-run versus short-run effects. It sensitizes policy makers to the pressure of future possible stakeholders, especially future generations who may come to bear the burden of our current decisions. It draws attention into the past seeking causes that may be buried at far spatial and temporal distances from current symptoms within the system. It seeks to understand the natural reaction time of the system, the period during which problems emerge and hence over which they need to be solved. System Dynamics-based analysis in the public sector draws analytic attention away from the riveting logic of the annual or biannual budget cycle, often focusing on options that will play themselves out years after current elected officials have left office. Such work is hard to do, but critical if one wants to think in systems terms.

What Is System Dynamics Modeling?

While other papers in this series may provide a more expanded answer to this basic question, it may be useful to begin this discussion of System Dynamics and public policy with a brief description of what System Dynamics is.

System Dynamics is an approach to policy analysis and design that applies to problems arising in complex social, managerial, economic, or ecological systems [31,33,74,95]. System Dynamics models are built around a particular problem. The problem defines the relevant factors and key variables to be included in the analysis. This represents the model's boundary, which may cross departmental or organizational boundaries. One of the unique advantages of using System Dynamics models to study public policy problems is that assumptions from a variety of stakeholders are explicitly stated, can be tested through simulation, and can be examined in context.

System Dynamics models rely on three sources of information: numerical data, the written database (reports, operations manuals, published works, etc.), and the expert knowledge of key participants in the system [36]. The numerical database of most organizations is very small,

the written database is larger, and the expert knowledge of key participants is vast. System Dynamicists rely on all three sources, with particular attention paid to the expert knowledge of key participants because it is only through such expert knowledge that we have any knowledge of the structure of the system. The explicit capturing of accumulated experience from multiple stakeholders in the model is one of the major differences between System Dynamics models and other simulation paradigms. An understanding of the long term effects of increased vigilance on the crime rate in a community needs to account for the reaction of courts, prisons, and rehabilitation agencies pressed to manage a larger population. This knowledge is spread across experts in several fields, and is not likely to be found in any single computer database. Rather, insight requires a process that makes these factors visible and explicit. For public sector problems, in particular, this approach helps move conflict out of the realm of inter-organizational conflict and towards a problem-solving focus.

Through the use of available data and by using the verbal descriptions of experts to develop mathematical relationships between variables, we expose new concepts and/or previously unknown but significant variables. System Dynamics models are appropriate to problems that arise in closed-loop systems, in which conditions are converted into information that is observed and acted upon, changing conditions that influence future decisions [69].

This idea of a "closed loop" or "endogenous" point of view on a system is really important to all good System Dynamics models. A simple example drawn from everyday life may help better to understand what an endogenous (versus exogenous) point of view means. If a father believes that his teenage daughter is always doing things to annoy him and put him in a bad mood, then he has an exogenous or "open loop" view of his own mood because he is seeing his mood as being controlled by forces outside of or exogenous to his own actions. However, if the father sees that his daughter and her moods are reacting to his own actions and moods while in turn his daughter's actions shape and define his moods, then this father has an endogenous point of view on his own mood. He understands how his mood is linked in a closed loop with another member of his family. Of course, the father with an endogenous view will be in a better position to more fully understand family dynamics and take actions that can prevent bad moods from spreading within the family.

Using an SD Model to Develop a Theory

A System Dynamics model represents a theory about a particular problem. Since models in the social sciences

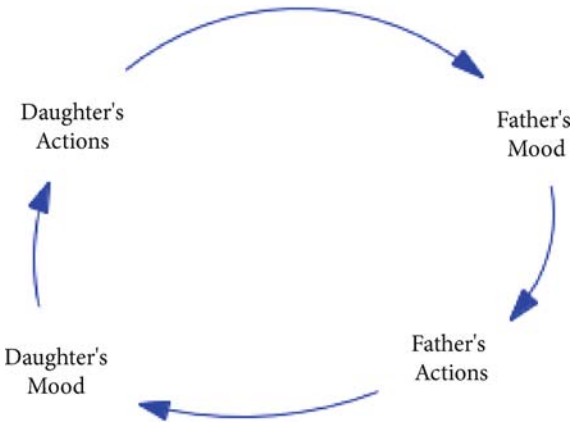

Public Policy, System Dynamics Applications to, Figure 1
Closed loop diagram of fathers and daughters

represent a theory, the most we can hope for from all these models, mental or formal, is that they be useful [94]. System Dynamics models are useful because the mathematical underpinning needed for computer simulation requires that the theory be precise. The process of combining numerical data, written data, and the knowledge of experts in mathematical form can identify inconsistencies about how we think the system is structured and how it behaves over time [38].

In policymaking it is often easy and convenient to blame other stakeholders for the problem state. Often, though, the structure of the system creates the problem by, for example, shifting resources to the wrong recipient or by inclusion of policies that intervene in politically visible but ineffective ways. The use of inclusive SD models educates us by identifying these inconsistencies through an iterative process involving hypotheses about system structure and tests of system behavior. Simulation allows us to see how the complex interactions we have identified work when they are all active at the same time. Furthermore, we can test a variety of policies quickly to see how they play out in the long run. The final result is a model that represents our most insightful and tested theory about the endogenous sources of problem behavior.

Behavior over Time Versus Forecasts

People who take a systems view of policy problems know that behavior generated by complex organizations cannot be well understood by examining the parts. By taking this holistic view, System Dynamicists capture time delays, amplification, and information distortion as they exist in organizations. By developing computer simulation models that incorporate information feedback, systems modelers seek to understand the internal policies and decisions, and

the external dynamic phenomena that combine to generate the problems observed. They seek to predict dynamic implications of policy, not forecast the values of quantities at a given time in the future.

System Dynamics models are tools that examine the behavior of key variables over time. Historical data and performance goals provide baselines for determining whether a particular policy generates behavior of key variables that is better or worse, when compared to the baseline or other policies. Furthermore, models incorporating rich feedback structure often highlight circumstances where the forces governing a system may change in a radical fashion. For example, in early phases of its growth a town in an arid region may be driven by a need to attract new jobs to support its population. At some future point in time, the very fact of successful growth may lead to a water shortage. Now the search for more water, not more jobs, may be what controls growth in the system. Richardson [69] has identified such phenomena as shifts in loop dominance that provide endogenous explanations for specific outcomes. Simulation allows us to compress time [95] so that many different policies can be tested, the outcomes explained, and the causes that generate a specific outcome can be examined by knowledgeable people working in the system, before policies are actually implemented.

Excellent short descriptions of System Dynamics methodology are found in Richardson [69,70] and Barlas [9]. Furthermore, Forrester's [33] detailed explanation of the field in *Industrial Dynamics* is still relevant, and Richardson and Pugh [74], Roberts et al. [78], Coyle [22], Ford [31], Maani and Cavana [53], Morecroft [65] and Sterman [95] are books that describe the field and provide tools, techniques and modeling examples suitable for the novice as well as for experienced System Dynamics modelers.

An Application of System Dynamics – The Governor's Office of Regulatory Assistance (GORA) Example

When applied to public policy problems, the "nuts and bolts" of this System Dynamics process consist of identifying the problem, examining the behavior of key variables over time, creating a visualization of the feedback structure of the causes of the problem, and developing a formal simulation model. A second case illustration may assist in understanding the process. The New York State Governor's Office of Regulatory Assistance (GORA) is a governmental agency whose mission it is to provide information about government rules and regulations to entrepreneurs who seek to start up new businesses in the state. The case was described by Andersen et al. [7] and is often used as a teaching case introducing System Dynamics to public managers.

Figure 2 below illustrates three key feedback loops that contribute both to the growth and eventual collapse of citizen service requests at GORA. The reinforcing feedback loop labeled "R1" illustrates how successful completion of citizen orders creates new contacts from word-of-mouth by satisfied citizens which in turn leads to more requests for service coming into the agency. If only this loop were working, a self-reinforcing process would lead to continuing expansion of citizen requests for services at GORA. The balancing loop labeled "B2" provides a balancing effect. As workers within the agency get more and more work to complete, the workload within the agency goes up with one effect being a possible drop in the quality in the work completed. Over time, loop B2 tells a story of how an increased workload can lead to a lower quality of work, with the effect of that lower quality being fewer incoming requests in the future. So over time, too many incoming requests set off a process that limits future requests by driving down quality. Many public managers who have worked with the GORA model find these two simple feedback loops to be realistic and powerful explanations of many of the problems that their agencies face on a day-to-day basis. The full GORA model has many other feedback loops and active variables not shown in the aggregated Fig. 2.

Once all of the variables have been represented by mathematical equations, a computer simulation is able to recreate an over time trajectory possible future values for all of the variables in the model. Figure 3 shows a graph over time of simulated data for key indicators in the GORA case study. The simulation begins when GORA comes into existence to provide services to the public and runs for 48 months. Initially, there is adequate staff and the amount of work to do is low, so the Workload Ratio, shown as part of loops B1 and B2 in the previous figure, is very low. With a low Workload Ratio GORA employees are able to devote additional time to each task they perform and the Quality of Work[1] is thus relatively high. The Backlog of Requests and the Average Completions Per Year begin at 0 and then increase and level off over time to approximately 4,500 and 41,000 respectively. The Fraction Experienced Staff mea-

[1] The Workload Ratio and Quality of Work are normalized variables. This means that they are measured against some predetermined standard. Therefore, when these two variables are equal to 1 they are operating in the desired state. Depending on the definition of the variable, values below or above 1 indicate when they are operating in a desired or undesired state. For example, Quality of Work above 1 indicates that quality is high, relative to the predetermined normal. However, Quality of Work below 1 indicates an undesirable state.

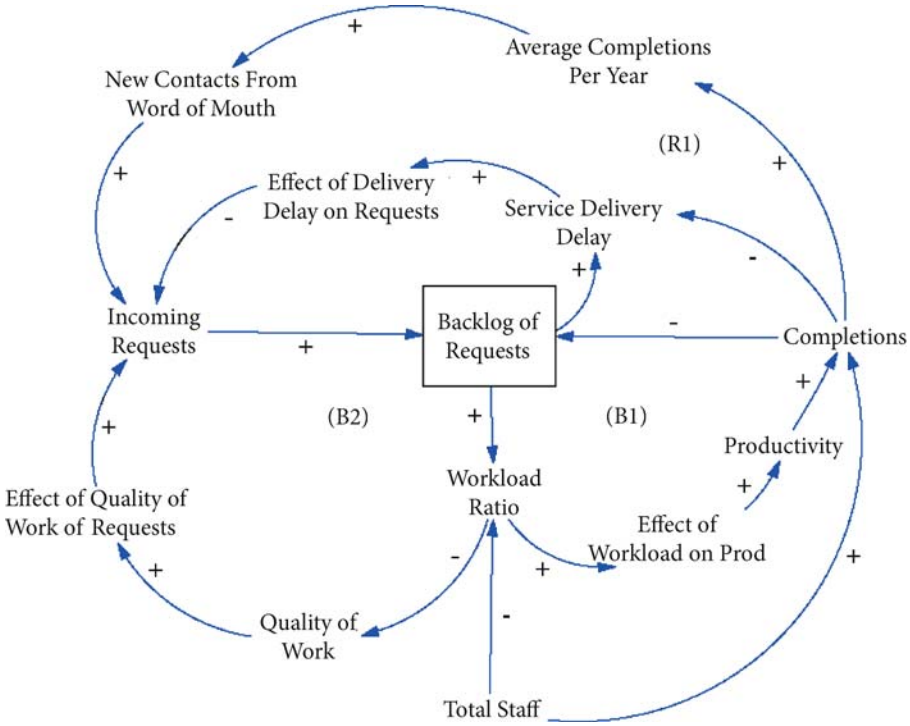

Public Policy, System Dynamics Applications to, Figure 2
Key feedback loops in a simulation of workflow in the Governor's Office of Regulatory Administration (GORA)

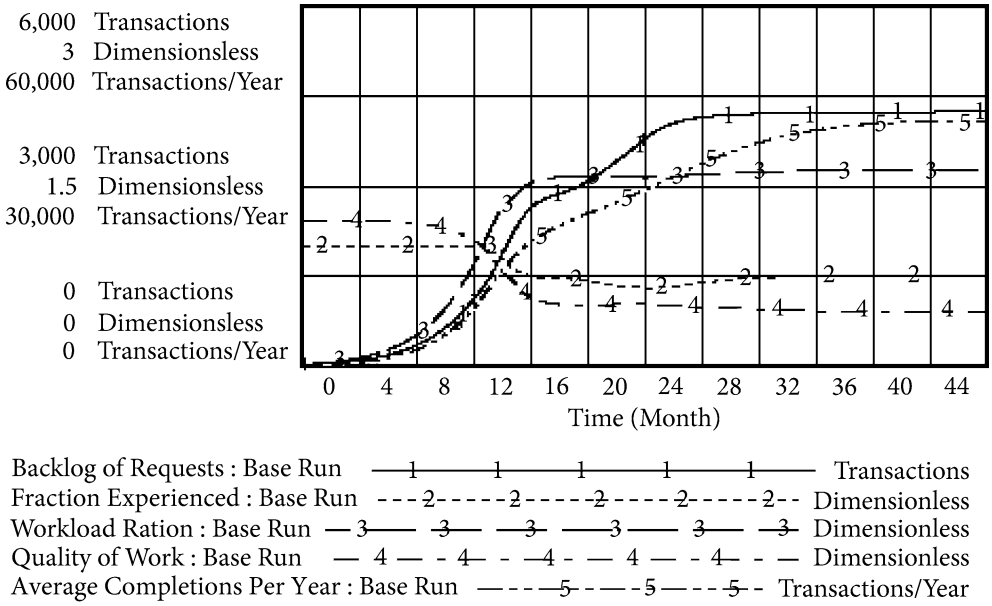

Public Policy, System Dynamics Applications to, Figure 3
Simulated performance of key variables in the GORA case study

sures what portion of the overall workers are experienced and hence more efficient at doing their jobs. As shown in Fig. 3, the Fraction Experienced begins at 1 and then falls and increases slightly to .75 indicating that GORA is having a harder time retaining experienced staff and is experiencing higher employee turnover. (The full GORA model has a theory of employee burnout and turnover not shown in Fig. 2.)

The combination of the visualization in Fig. 2 with a formal model capable of generating the dynamic output shown in Fig. 3 illustrates the power of System Dynamics modeling for public policy issues. Linking behavior and structure helps stakeholders understand why the behavior of key variables unfolds over time as it does. In the GORA case, the program is initially successful as staff are experienced, are not overworked, and the quality of the services they provide is high. As clients receive services the R1 feedback loop is dominant and this attracts new clients to GORA. However, at the end of the first year the number of clients requesting services begins to exceed the ability of GORA staff to provide the requested services in a timely manner. The Workload Ratio increases, employees are very busy, the Quality of Work falls, and the B2 feedback loop works to limit the number of people seeking services. Furthermore, people are waiting longer to receive services and some are discouraged from seeking services due to the delay. The initial success of the program cannot be sustained and the program settles down into an unsatisfactory situation where the Workload Ratio is high, Quality of Work is low, clients are waiting longer for services and staff turnover is high as indicated by the Fraction Experienced.

The model tells a story of high performance expectations, initial success and later reversal, all explained endogenously. Creating and examining the simulation helps managers consider possible problems before they occur – before staff are overtaxed, before turnover climbs, and before the agency has fallen behind. Having a model to consider compresses time and provides the opportunity for a priori analysis. Finally, having a good model can provide managers with a test bed for asking "what if" questions, allowing public managers to spend simulated dollars and make simulated errors all the while learning how to design better public policies at relatively low cost and without real (only simulated) risk.

How Is System Dynamics Used to Support Public Policy and Management?

The Medical Malpractice vignette that opened this chapter involving the New York State Commissioner of Insurance is more fully documented by Reagan-Cirincione et al. [68] and is one of the first published examples of the results of a team of government executives working in a face-to-face group model building session to create a System Dynamics model to support critical policy decisions facing the group. The combined group modeling and simulation approach had a number of positive effects on the policy process. Those positive effects are:

Make Mental Models of Key Players Explicit

When the Commissioner drew together his team, the members of this group held different pieces of information and expertise. Much of the most important information was held in the minds, in the mental models, of the Commissioner's staff, and not in data tabulations. The System Dynamics modeling process made it possible for managers to explicitly represent and manipulate their shared mental models in the form of a System Dynamics simulation model. This process of sharing and aligning mental models, as done during a System Dynamics modeling intervention, is an important aspect of a "learning organization" as emphasized by Senge [87].

Create a Formal and Explicit Theory of the Public Policy Situation Under Discussion

The formal model of malpractice insurance contained an explicit and unambiguous theory of how the medical malpractice system in New York State functions. The shared mental models of the client team implied such a formal and model-based theory, but the requirements of creating a running simulation forced the group to be much more explicit and clear about their joint thinking. As the modeling team worked with the group, a shared consensus about how the whole medical malpractice system worked was cast, first into a causal-loop diagram, and later into the equations of the formal simulation model [74,95].

Document all Key Parameters and Numbers Supporting the Policy Debate

In addition to creating a formal and explicit theory, the System Dynamics model was able to integrate explicit data and professional experience available to the Department of Insurance. Recording the assumptions of the model in a clear and concise way makes possible review and examination by those not part of the model's development. Capturing these insights and their derivation provides face validity to the model's constructs.

Building confidence in the utility of a System Dynamics model for use in solving a public policy problem in-

volves a series of rigorous tests that probe how the model behaves over time as well as how available data, both numerical and tacit structural knowledge, have been integrated and used in the model. Forrester and Senge [39] detail 17 tests for building confidence in a System Dynamics model. Sterman [95] identifies 12 model tests, the purpose or goal of each test, and the steps that modelers should follow in undertaking those tests. Furthermore, Sterman [95] also lists questions that model consumers should ask in order to generate confidence in a model. This is particularly important for public policy issues where the ultimate goal or outcome for different stakeholders may be shared, but underlying assumptions of the stakeholders may be different.

Create a Formal Model that Stimulates and Answers Key "what if" Questions

Once the formal model was constructed, the Commissioner and his policy team were able to explore "what if" scenarios in a cost-free and risk-free manner. Significant cost overruns in a simulated environment do not drive up real tax rates, nor do they lead to an elected official being voted out of office, nor to an appointed official losing her job. Quite the contrary, a simulated cost overrun or a simulated failed program provides an opportunity to learn how better to implement or manage the program or policy (or to avoid trying to implement the policy). Public managers get to experiment quickly with new policies or programs in a risk-free simulated environment until they "get it right" in the simulated world. Only then should they take the risk of implementation in a high stakes policy environment.

Bringing a complex model to large groups sometimes requires the development of a more elaborate simulation, so that those who were not part of the initial analysis can also derive insight from its results. Iterative development and discussion provides an additional validation of the constructs and conclusions of the model, Zagonel et al. [108] have described a case where local managers responsible for implementing the 1996 federal welfare reform legislation used a simulation model to explore such "what if" futures before taking risks of actual implementation.

Public policy problems are complex, cross organizational boundaries, involve stakeholders with widely different perspectives, and evolve over time. Changes in police procedures and/or resources may have an effect on prison and parole populations many years into the future. Health care policies will determine how resources are allocated at local hospitals and the types of treatments that can be obtained. Immigration policies in one country may influence the incomes and jobs of people in a second country. Miyakawa [64] has pointed out that public policies are systemically interdependent. Solutions to one problem often create other problems. Increased enforcement of immigration along the U.S. borders has increased the workload of courts [26]. Besides being complex these examples also contain stakeholders with different sets of goals. In solving public policy problems, how diverse stakeholders work out their differences is a key component of successful policy solutions. System Dynamics modeling interventions, and in particular the techniques of group model building [2,6,72,98], provide a unique combination of tools and methods to promote shared understanding by key stakeholders within the system.

System Dynamics and Models: A Range of Analytic Scope and Products

In the malpractice insurance example, the Commissioner called his advisors into a room to explicitly engage in a group model building session. These formal group model-building sessions involve a specialized blend of projected computer support plus professional facilitation in a face-to-face meeting of public managers and policy analysts. Figure 4 is an illustration of a team of public managers working together in a group model building project. In this photograph, a facilitator is working on a hand drawn view of a simulation model's structure while projected views of computer output can be used to look at first cut simulation runs or refined images of the model being built by the group. Of course, the key feature of this whole process is facilitated face-to-face conversations between the key stakeholders responsible for the policy decisions being made.

Richardson and Andersen [72], Andersen and Richardson [6], Vennix [98], and Luna-Reyes et al. [52] have provided detailed descriptions of how this kind of group model building process actually takes place. In addition to these group model building approaches, the System Dynamics literature describes five other ways that teams of modelers work with client groups. They are (1) the Reference Group approach [91], (2) the Strategic Forum [75], (3) The stepwise approach [104], (4) strategy dynamics [100,101,102], and (5) the "standard method" of Hines [67].

Some System Dynamics-oriented analyzes of public policies completed by groups of public managers and policy analysts stop short of building a formal simulation model. The models produced by Wolstenholme and Coyle [107], Cavana, Boyd and Taylor [14] and the system

Public Policy, System Dynamics Applications to, Figure 4
A team of public managers working together to build a System Dynamics model of welfare reform policies

archetypes promoted by Senge [87] have described how these qualitative system mapping exercises, absent a formal running simulation model, can add significant value to a client group struggling with an important public policy problem. The absence of a formal simulation limits the results to a conceptual model, rather than a tool for systematic experimentation.

Finally, a number of public agencies and Non Governmental Organizations are joining their counterparts in the private sector by providing broad-based systems thinking training to their top leaders and administrative staff. A number of simulation-based management exercises such as the production-distribution game (also known as the "beer game") [93] and the People's Express Flight Simulator [92] have been developed and refined over time to support such training and professional development efforts. In addition, Cavana and Clifford [11] have used GMB to develop a formal model and flight simulator to examine the policy implications of an excise tax policy on tobacco smoking.

What Are the Arenas in Which System Dynamics Models Are Used?

The malpractice insurance vignette and the GORA example represented cases where a model was developed for a single problem within one agency. Naill [66] provides an example of how a sustained modeling capability can be installed within an agency to support a range of ongoing policy decisions (in this case the model was looking at transitional energy policies at the federal level). Barney [10] developed a class of System Dynamics simulation models to support economic development and planning in developing nations. Wolstenholme [105] reported on efforts to support health planning within the British Health Service.

Addressing a tactical problem within a single public sector agency, while quite common, is only one of the many types of decision arenas in which System Dynamics models can be and are used to support public policy. Indeed, how a model is used in a public policy debate is largely determined by the unique characteristics of the specific decision-making arena in which the model is to be used. Some of the more common examples follow.

Models Used to Support Inter-Agency and Inter-Governmental Collaborative Efforts

A quite different arena for the application of System Dynamics models to support the policy process occurs when an interagency or inter-governmental network of program managers must cooperate to meet a common mission. For example, Rohrbaugh [79] and Zagonel et al. [108] report a case where state and local officials from social services, labor, and health agencies combined their efforts with private and non-profit managers of day care services, health care services, and worker training and education services to plan for comprehensive reform of welfare policies in the late 1990s. These teams were seeking strategies to blend financial and program resources across a myriad of stovepipe regulations and reimbursement schemes to provide a seamless system of service to clients at the local level. To complete this task, they created a simulation model containing a wide range of system-level interactions

and tested policies in that model to find out what blend of policies might work. Policy implementation followed this model-based and simulation-supported policy design.

Models Used to Support Expert Testimony in Courtroom Litigation

Cooper [20] presented one of the first published accounts of a System Dynamics model being used as a sort of expert witness in courtroom litigation. In the case he reported, Litton Industries was involved in a protracted lawsuit with the U. S. Navy concerning cost and time overruns in the construction of several naval warships. In a nutshell, the Navy contended that the cost overruns were due to actions taken (or not taken) by Litton Industries as primary contractor on the project and as such the Navy should not be responsible for covering cost overruns. Litton maintained that a significant number of change orders made by the Navy were the primary drivers of cost overruns and time delays. A simulation model was constructed of the shipbuilding process and the simulation model then built two simulated ships without any change orders. A second set of "what if" runs subsequently built the same ships except that the change orders from the Navy were included in the construction process. By running and re-running the model, the analysts were able to tease out what fractions of the cost overrun could reasonably be attributed to Litton and what fraction should be attributed to naval change orders. Managers at Litton Industries attribute their receipt of hundreds of millions dollars of court-sanctioned payments to the analysis supported by this System Dynamics simulation model. Ackermann, Eden and Williams [1] have used a similar approach involving soft systems approaches combined with a System Dynamics model in litigation over cost overruns in the channel tunnel project.

Models Used as Part of the Legislative Process

While System Dynamics models have been actively used to support agency-level decision making, inter-agency and inter-governmental task forces and planning, and even courtroom litigation, their use in direct support of legislative processes has a more uneven track record. For example, Ford [30] reports successes in using System Dynamics modeling to support regulatory rule making in the electric power industry, and Richardson and Lamitie [73] report on how System Dynamics modeling helped redefine a legislative agenda relating to the school aid formula in the U.S. state of Connecticut. However, Andersen [4] remains more pessimistic about the ability of System Dynamics models to directly support legislative decision making, especially when the decisions involve zero-sum tradeoffs in the allocation of resources (such as formula-driven aid involving local municipal or education formulas). This class of decisions appears to be dominated by short-term special interests. A longer-term dynamic view of such immediate resource allocation problems is less welcome. The pathway to affecting legislative decision making appears to be by working through and with public agencies, networks of providers, the courts, or even in some opinions, by directly influencing public opinion.

Models Used to Inform the Public and Support Public Debate

In addition to using System Dynamics modeling to support decision making in the executive, judicial, and legislative branches of government (often involving Non-Governmental Organizations and private sector support), a number of System Dynamics studies appeal directly to the public. These studies intend to affect public policy by shaping public opinion in the popular press and the policy debate. In the 1960s, Jay Forrester's *Urban Dynamics* [34] presented a System Dynamics model that looked at many of the problems facing urban America in the latter half of the 20th century. Several years later in response to an invitation from the Club of Rome, Forrester put together a study that led to the publication of *World Dynamics* [35], a highly aggregate System Dynamics model that laid out a feedback-oriented view of a hypothesized set of relationships between human activity on the planet, industrialization, and environmental degradation. Meadows et al. [61] followed on this study with a widely hailed (and critiqued) System Dynamics simulation study embodied in the best-selling book, *Limits to Growth*. Translated into over 26 languages, this volume coalesced a wide range of public opinion leading to a number of pieces of environmental reform in the decade of the 1970s. The debate engendered by that volume continues even 30 years later [63]. Donella Meadows continued in this tradition of appealing directly to public opinion through her syndicated column, The Global Citizen, which was nominated for the Pulitzer Prize in 1991. The column presented a System Dynamics-based view of environment matters for many years (http://www.pcdf.org/meadows/).

What Are some of the Substantive Areas Where System Dynamics Has Been Applied?

The International System Dynamics Society (http://www.systemdynamics.org) maintains a comprehensive bibliography of over 8,000 scholarly books and articles documenting a wide variety of applications of System Dynamics modeling to applied problems in all sectors. MacDon-

ald et al. [54] have created a bibliography extracted from this larger database that summarizes some of the major areas where System Dynamics modeling has been applied to public policy. Below, we summarize some of the substantive areas where System Dynamics has been applied, giving one or two sample illustrations for each area.

Health Care

System Dynamicists have been applying their tools to analyze health care issues at both the academic and practitioner level for many years. *The System Dynamics Review*, the official journal of the System Dynamics Society, devoted a special issue to health care in 1999 due to the importance of health care as a critical public policy issue high on the political agenda of many countries and as an area where much System Dynamics work has been performed. The extensive System Dynamics work performed in the health care area fell into three general categories: patient flow management, general health policy, and specific health problems.

The patient flow management category is exemplified by the work of Wolstenholme [106], Lane and Rosenhead [48], and Van Ackere and Smith [97]. The articles written by these authors focused on issues and policies relating to patient flows in countries where health care service is universal.

The general health policy category is rather broad in that these articles covered policy and decision making from the micro level [96] to the macro level [88]. There were also many articles that showed how the process of modeling resulted in better understanding of the problem and issues facing health care providers and policy makers [12].

The last category dealt with specific health-related problems such as the spread of AIDS [43,76], smoking [42], and malaria control [29], as well as many other health-related conditions.

Education

The education articles touched on various topics relating to education ranging from using System Dynamics in the classroom as a student-centered teaching method to models that dealt with resource allocations in higher education. Nevertheless, many of the articles fell into five categories that could be labeled management case studies or flight simulators, teaching technology, research, teaching, and education policy.

The management case study and flight simulator articles are best exemplified by Sterman's [93] article describing the Beer Game and Graham, Morecroft et al. [41] article on "Model Supported Case Studies for Management Education." The emphasis of these works is on the use of case studies in higher education, with the addition of games or computer simulations. This is related to the teaching technology category in that both emphasize using System Dynamics models/tools to promote learning. However, the teaching technology category of articles stresses the introduction of computer technology, specifically System Dynamics computer technology, into the classroom. Steed [90] has written an article that discusses the cognitive processes involved while using Stella to build models, while Waggoner [99] examined new technologies versus traditional teaching approaches.

In addition to teaching technology are articles that focus on teaching. The teaching category is very broad in that it encompasses teaching System Dynamics in K-12 and higher education as subject matter [37,77] as well as ways to integrate research into the higher-education classroom [71]. System Dynamics models are also used to introduce advanced mathematical concepts through simulation and visualization, rather than through equations [27,28]. In addition, lesson plans for the classroom are also part of this thread [44]. The Creative Learning Exchange (http://www.clexchange.com) provides a central repository of lessons and models useful for pre-college study of System Dynamics, including a selfstudy roadmap to System Dynamics principles [23].

There are also a number of articles that pertain to resource allocation [15] at the state level for K-12 schools along with articles that deal with resource-allocation decisions in higher education [32,40]. Saeed [83] and Mashayekhi [56] cover issues relating to higher education policy in developing countries.

The last education category involved research issues around education. These articles examined whether the System Dynamics methodology and simulation-based education approaches improved learning [24,47,55].

Defense

System Dynamics modeling work around the military has focused on manpower issues, resource allocation decisions, decision making and conflict. Coyle [21] developed a System Dynamics model to examine policies and scenarios involved in sending aircraft carriers against land-based targets. Wils, Kamiya et al. [103] have modeled internal conflicts as a result of outbreaks of conflict over allocation and competition of scarce resources. The manpower articles focused on recruitment and retention policies in the armed forces and are represented in articles by Lopez and Watson [51], Andersen and Emmerichs [5],

Clark [18], Clark, McCullough et al. [19] and Cavana et al. [14]. The resource allocation category deals with issues of money and materials, as opposed to manpower, and is represented by Clark [16,17]. Decision making in military affairs from a System Dynamics perspective is represented in the article by Bakken and Gilljam [8].

Environment

The System Dynamics applications dealing with environmental resource issues can be traced back to when the techniques developed in *Industrial Dynamics* were beginning to be applied to other fields. The publication of Forrester's *World Dynamics* in 1971 and the follow-up study *Limits to Growth* [61,62,63] used System Dynamics methodology to address the problem of continued population increases on industrial capital, food production, resource consumption and pollution. Furthermore, specific studies dealing with DDT, mercury and eutrophication of lakes were part of the Meadows et al. [59] project and appeared as stand-alone journal articles prior to being published as a collection in Meadows and Meadows [60].

The environmental applications of System Dynamics have moved on since that time. Recent work has combined environmental and climate issues with economic concerns thorough simulation experiments [25] as well as stakeholder participation in environmental issues [89]. In 2004, the *System Dynamics Review* ran a special issue dedicated to environmental issues. Cavana and Ford [13] were the editors and did a review of the System Dynamics bibliography in 2004, identifying 635 citations with the key words "environmental" or "resource." Cavana and Ford broke the 635 citations into 11 categories they identified as resources, energy, environmental, population, water, sustainable, natural resources, forest, ecology, agriculture, pollution, fish, waste, earth, climate and wildlife.

General Public Policy

The System Dynamics field first addressed the issue of public policy with Forrester's *Urban Dynamics* [34] and the follow-up work contained in *Readings in Urban Dynamics* [58] and Alfeld and Graham's *Introduction to Urban Dynamics* [3]. The field then branched out into the previously mentioned *World Dynamics* and the follow-up studies related to that work. Moreover, the application of System Dynamics to general public policy issues began to spread into areas as diverse as drug policy [50], and the causes of patient dropout from mental health programs [49], to ongoing work by Saeed [82,84,85] on development issues in emerging economies. More recently, Saysel et al. [86] have examined water scarcity issues in agricultural areas, Mashayekhi [57] reports on the impact on public finance of oil exports in countries that export oil and Jones et al. [46] cover the issues of sustainability of forests when no single entity has direct control.

This brief review of the literature where System Dynamics modeling has been used to address public policy issues indicates that the field is making inroads at the micro level (within government agencies) and at the macro level (between government agencies). Furthermore, work has been performed at the international level and at what could truly be termed the global level with models addressing public policy issues aimed at climate change.

Evaluating the Effectiveness of System Dynamics Models in Supporting the Public Policy Process

System Dynamics modeling is a promising technology for policy development. But does it really work? Over the past several decades, a minor cottage industry has emerged that purports to document the successes (and a few failures) of System Dynamics models by reporting on case studies. These case studies report on successful applications and sometimes analyze weaknesses, making suggestions for improvement in future practice. Rouwette et al. [81] have compiled a meta-analysis of 107 such case-based stories.

However, as compelling as such case stories may be, case studies are a famously biased and unsystematic way to evaluate effectiveness. Presumably, failed cases will not be commonly reported in the literature. In addition, such a research approach illustrates in almost textbook fashion the full litany of both internal and external threats to validity, making such cases an interesting but unscientific compilation of war stories. Attempts to study live management teams in naturally occurring decision situations can have high external validity but almost always lack internal controls necessary to create scientifically sound insights.

Huz et al. [45] created an experimental design to test for the effectiveness of a controlled series of group-based System Dynamics cases in the public sector. They used a wide battery of pre- and post survey, interview, archival, administrative data, and qualitative observation techniques to evaluate eight carefully matched interventions. All eight interventions dealt with the integration of mental health and vocational rehabilitation services at the county level. Four of the eight interventions contained System Dynamics modeling sessions and four did not. These controlled interventions were designed to get at the impact of System Dynamics modeling on the public policy process.

Overall, Huz et al. [45] envisioned that change could take place in nine domains measured across three separate levels of analysis as illustrated in Table 1 below.

Using the battery of pre- and post test instruments, Huz found important and statistically significant results in eight of the nine domains measured. The exceptions were in domain 9 where they did not measure client outcomes, in domain 5 where "participants were not significantly more aligned in their perceptions on strategies for changes" (but were more aligned in goals), and in domain 7 where "no significant change was found with respect to structural conditions within the network" (but two other dimensions of organizational relationships did change).

In their meta-analysis of 107 case studies of System Dynamics applications, Rouwette et al. [81] coded case studies with respect to eleven classes of outcomes, sorted into individual level, group level, and organizational level. The 107 cases were dominated by for-profit examples with 65 such cases appearing in the literature followed by 21 cases in the non-profit sector, 18 cases in governmental settings, and three cases in mixed settings. While recognizing possible high levels of bias in reported cases as well as difficulties in coding across cases and a high number of missing categories, they found high percentages of positive outcomes along all 11 dimensions of analysis. For each separate dimension, they analyzed between 13 and 101 cases with the fraction of positive outcomes for each dimension ranging from a low of 83% to several dimensions where 100% of the cases reporting on a dimension found positive results. At the individual level, they coded for overall positive reactions to the work, insight gained from the work, and some level of individual commitment to the results emerging from the study. At the group level, they coded for increased levels of communication, the emergence of shared language, and increases in consensus or mental model alignment. Organizational level outcomes included implementation of system level change. With respect to this important overall indicator they "found 84 projects focused on implementation, which suggests that in half (42) of the relevant cases changes are implemented. More than half (24) of these changes led to positive results"(see p. 20 in [81]).

Rouwette [80] followed this meta-analysis with a detailed statistical analysis of a series of System Dynamics-based interventions held mostly in governmental settings in the Netherlands. He was able to estimate a statistical model that demonstrated how System Dynamics group model building sessions moved both individuals and groups from beliefs to intentions to act, and ultimately on to behavioral change.

In sum, attempts to evaluate System Dynamics interventions in live settings continue to be plagued by methodological problems that researchers have struggled to overcome with a number of innovative designs. What is emerging from this body of study is a mixed, "good news and bad news" picture. All studies that take into account a reasonable sample of field studies show some successes and some failures. About one-quarter to one-half of the System Dynamics studies investigated showed low impact on decision making. On the other hand, roughly half of the studies have led to system-level implemented change with approximately half of the implemented studies being associated with positive measures of success.

Public Policy, System Dynamics Applications to, Table 1
Domains of measurement and evaluation used to assess impact of systems-dynamics interventions (see p. 151 in [45])

Level I	**Reflections of the modeling team**
Domain 1	Modeling team's assessment of the intervention
Level II	**Participant self-reports of the intervention**
Domain 2	Participants' perceptions of the intervention
Domain 3	Shifts in participants' goal structures
Domain 4	Shifts in participants' change strategies
Domain 5	Alignment of participant mental models
Domain 6	Shifts in understanding how the system functions
Level III	**Measurable system change and "bottom line" results**
Domain 7	Shifts in network of agencies that support services integration
Domain 8	Changes in system-wide policies and procedures
Domain 9	Changes in outcomes for clients

Summary: System Dynamics – A Powerful Tool to Support Public Policy

While recognizing and respecting the difficulties of scientific evaluation of System Dynamics studies in the public sector, we remain relentlessly optimistic about the method's utility as a policy design and problem-solving tool. Our glass is half (or even three-quarters) full. A method that can deliver high decision impact up to three-quarters of the time and implement results in up to half of the cases examined (and in a compressed time frame) is a dramatic improvement over alternative approaches that can struggle for months or even years without coming to closure on important policy directions.

System Dynamics-based modeling efforts are effective because they join the minds of public managers and policy makers in an emergent dialog that relies on formal modeling to integrate data, other empirical insights, and

mental models into the policy process. Policy making begins with the pre-existing mental models and policy stories that managers bring with them into the room. Policy consensus and direction emerge from a process that combines social facilitation with technical modeling and analysis. The method blends dialog with data. It begins with an emergent discussion and ends with an analytic framework that moves from "what is" baseline knowledge to informed "what if" insights about future policy directions.

In sum, we believe that a number of the process features related to building System Dynamics models to solve public policy problems contribute to their appeal for frontline managers:

- **Engagement** Key managers can be in the room as the model is evolving, and their own expertise and insights drive all aspect of the analysis.
- **Mental models** The model-building process uses the language and concepts that managers bring to the room with them, making explicit the assumptions and causal mental models managers use to make their decisions.
- **Complexity** The resulting nonlinear simulation models lead to insights about how system structure influences system behavior, revealing understandable but initially counterintuitive tendencies like policy resistance or "worse before better" behavior.
- **Alignment** The modeling process benefits from diverse, sometimes competing points of view as stakeholders can have a chance to wrestle with causal assumptions in a group context. Often these discussions realign thinking and are among the most valuable portions of the overall modeling effort.
- **Refutability** The resulting formal model yields testable propositions, enabling managers to see how well their implicit theories match available data about overall system performance.
- **Empowerment** Using the model managers can see how actions under their control can change the future of the system.

System Dynamics modeling projects merge managers' causal and structural thinking with the available data, drawing upon expert judgment to fill in the gaps concerning possible futures. The resulting simulation models provide powerful tools to develop a shared understanding and to ground what-if thinking.

Future Directions

While the field of System Dynamics has reached its half-centenary in 2007, its influence on public policy continues to grow. Many of the problems defined by the earliest writers in the field continue to challenge us today. The growing literature base of environmental, social, and education policy is evidence of continued interest in the systems perspective. In addition, System Dynamics modeling is growing in popularity for defense analysis, computer security and infrastructure planning, and emergency management. These areas have the characteristic problems of complexity and uncertainty that require the integration of multiple perspectives and tacit knowledge that this method supports. Researchers and practitioners will continue to be attracted to the open nature of System Dynamics models as a vehicle for consensus and experimentation.

We anticipate that the tool base for developing and distributing System Dynamics models and insights will also grow. Graphical and multimedia-based simulations are growing in popularity, making it possible to build clearer models and disseminate insights easily. In addition, the development of materials for school-age learners to consider a systems perspective to social problems gives us optimism for the future of the field, as well as for future policy.

Bibliography

Primary Literature

1. Ackermann F, Eden C, Williams T (1997) Modeling for Litigation: Mixing Qualitative and Quantitative Approaches. Interfaces 27(2):48–65
2. Akkermans H, Vennix J (1997) Clients' Opinions on Group Model-Building: An Exploratory Study. Syst Dyn Rev 13(1):3–31
3. Alfeld L, Graham A (1976) Introduction to Urban Dynamics. Wright-Allen Press, Cambridge
4. Andersen D (1990) Analyzing Who Gains and Who Loses: The Case of School Finance Reform in New York State. Syst Dyn Rev 6(1):21–43
5. Andersen D, Emmerichs R (1982) Analyzing US Military Retirement Policies. Simulation 39(5):151–158
6. Andersen D, Richardson GP (1997) Scripts For Group Model Building. Syst Dyn Rev 13(2):107–129
7. Andersen D, Bryson J, Richardson GP, Ackermann F, Eden C, Finn C (2006) Integrating Modes of Systems Thinking into Strategic Planning Education and Practice: The Thinking Persons' Institute Approach. J Public Aff Educ 12(3):265–293
8. Bakken B, Gilljam M (2003) Dynamic Intuition in Military Command and Control: Why it is Important, and How It Should be Developed. Cogn Technol Work (5):197–205
9. Barlas Y (2002) System Dynamics: Systemic Feedback Modeling for Policy Analysis. In: Knowledge for Sustainable Development, an Insight into the Encyclopedia of Life Support Systems, vol 1. UNESCO-EOLSS, Oxford, pp 1131–1175
10. Barney G (1982) The Global 2000 Report to the President: Entering the Twenty-First Century. Penguin, New York
11. Cavana RY, Clifford L (2006) Demonstrating the utility for system dynamics for public policy analysis in New Zealand:

the case for excise tax policy on tobacco. Syst Dyn Rev 22(4):321–348
12. Cavana RY, Davies P et al (1999) Drivers of Quality in Health Services: Different Worldviews of Clinicians and Policy Managers Revealed. Syst Dyn Rev 15(3):331–340
13. Cavana RY, Ford A (2004) Environmental and Resource Systems: Editor's Introduction. Syst Dyn Rev 20(2):89–98
14. Cavana RY, Boyd D, Taylor R (2007) A Systems Thinking Study of Retnetion and Recruitment Issues for the New Zealand Army Electronic Technician Trade Group. Syst Res Behav Sci 24(2):201–216
15. Chen F, Andersen D et al (1981) A Preliminary System Dynamics Model of the Allocation of State Aid to Education. Dynamica 7(1):2–13
16. Clark R (1981) Readiness as a Residual of Resource Allocation Decisions. Def Manag J 1:20–24
17. Clark R (1987) Defense Budget Instability and Weapon System Acquisition. Public Budg Financ 7(2):24–36
18. Clark R (1993) The Dynamics of US Force Reduction and Reconstitution. Def Anal 9(1):51–68
19. Clark T, McCullough B et al (1980) A Conceptual Model of the Effects of Department of Defense Realignments. Behav Sci 25(2):149–160
20. Cooper K (1980) Naval Ship Production: A Claim Settled and Framework Built. Interfaces 10(6):20
21. Coyle RG (1992) A System Dynamics Model of Aircraft Carrier Survivability. Syst Dyn Rev 8(3):193–213
22. Coyle RG (1996) System Dynamics Modelling: A Practical Approach. Chapman and Hall, London
23. Creative Learning Exchange (2000) Road Maps: A Guide to Learning System Dynamics. Available at http://sysdyn.clexchange.org/road-maps/home.html
24. Davidsen P (1996) Educational Features of the System Dynamics Approach to Modelling and Learning. J Struct Learn 12(4):269–290
25. Fiddaman TS (2002) Exploring Policy Options with a Behavioral Climate-Economy Model. Syst Dyn Rev 18(2):243–267
26. Finely B (2006) Migrant Cases Burden System: Rise in Deportations Floods Detention Centers, Courts. Denver Post, 10/2/06: http://www.denverpost.com/immigration/ci_4428563
27. Fisher D (2001) Lessons in Mathematics: A Dynamic Approach. iSee Systems, Lebanon
28. Fisher D (2004) Modeling Dynamic Systems: Lessons for a First Course. iSee Systems, Lebanon
29. Flessa S (1999) Decision Support for Malaria-Control Programmes – a System Dynamics Model. Health Care Manag Sci 2(3):181–91
30. Ford A (1997) System Dynamics and the Electric Power Industry. Syst Dyn Rev 13(1):57–85
31. Ford A (1999) Modeling the Environment: An Introduction to System Dynamics Modeling of Environmental Systems. Island Press, Washington, DC
32. Forsyth B, Hirsch G et al (1976) Projecting a Teaching Hospital's Future Utilization: A Dynamic Simulation Approach. J Med Educ 51(11):937–9
33. Forrester J (1961) Industrial Dynamics. Pegasus Communications, Cambridge
34. Forrester J (1969) Urban Dynamics. Pegasus Communications, Waltham
35. Forrester J (1971) World Dynamics. Pegasus Communications, Waltham
36. Forrester J (1980) Information Sources for Modeling the National Economy. J Am Stat Assoc 75(371):555–566
37. Forrester J (1993) System Dynamics as an Organizing Framework for Pre-College Education. Syst Dyn Rev 9(2):183–194
38. Forrester J (1994) Policies, Decisions, and Information Sources for Modeling. Modeling for Learning Organizations. In: Morecroft J, Sterman J (eds) Productivity Press. Portland, OR, pp 51–84
39. Forrester J, Senge P (1980) Tests for Building Confidence in System Dynamics Models. In: Legasto Jr. AA et al (eds) System Dynamics. North-Holland, New York, 14, pp 209–228
40. Galbraith P (1989) Mathematics Education and the Future: A Long Wave View of Change. Learn Math 8(3):27–33
41. Graham A, Morecroft J et al (1992) Model Supported Case Studies for Management Education. Eur J Oper Res 59(1):151–166
42. Homer J, Roberts E et al (1982) A Systems View of the Smoking Problem. Int J Biomed Comput 13 69–86
43. Homer J, St Clair C (1991) A Model of HIV Transmission Through Needle Sharing. A Model Useful in Analyzing Public Policies, Such as a Needle Cleaning Campaign. Interfaces 21(3):26–29
44. Hopkins P (1992) Simulating Hamlet in the Classroom. Syst Dyn Rev 8(1):91–100
45. Huz S, Andersen D, Richardson GP, Boothroyd R (1997) A Framework for Evaluating Systems Thinking Interventions: An Experimental Approach to Mental Health System Change. Syst Dyn Rev 13(2)149–169
46. Jones A, Seville D et al (2002) Resource Sustainability in Commodity Systems: The Sawmill Industry in the Northern Forest. Syst Dyn Rev 18(2):171–204
47. Keys B, Wolfe J (1996) The Role of Management Games and Simulations in Education Research. J Manag 16(2):307–336
48. Lane DC, Monefeldt C, Rosenhead JV (1998) Emergency – But No Accident – A System Dynamics Study of an accident and emergency department. OR Insight 11(4):2–10
49. Levin G, Roberts E (eds) (1976) The Dynamics of Human Service Delivery. Ballinger, Cambridge
50. Levin G, Hirsch G, Roberts E (1975) The Persistent Poppy: A Computer Aided Search for Heroin Policy. Ballinger, Cambridge
51. Lopez T, Watson J Jr (1979) A System Dynamics Simulation Model of the U. S. Marine Corps Manpower System. Dynamica 5(2):57–78
52. Luna-Reyes L, Martinez-Moyano I, Pardo T, Creswell A, Richardson GP, Andersen D (2007) Anatomy of a Group Model Building Intervention: Building Dynamic Theory from Case Study Research. Syst Dyn Rev 22(4):291–320
53. Maani KE, Cavana RY (2007) Systems Thinking, System Dynamics: Managing Change and Complexity. Pearson Education (NZ) Ltd, Auckland
54. MacDonald R et al (2007) System Dynamics Public Policy Literature. Syst Dyn Soc, http://www.systemdynamics.org/short_bibliography.htm
55. Mandinach E, Cline H (1993) Systems, Science, and Schools. Syst Dyn Rev 9(2):195–206
56. Mashayekhi A (1977) Economic Planning and Growth of Education in Developing Countries. Simulation 29(6):189–197
57. Mashayekhi A (1998) Public Finance, Oil Revenue Expenditure and Economic Performance: A Comparative Study of Four Countries. Syst Dyn Rev 14(2–3):189–219

58. Mass N (ed) (1974) Readings in Urban Dynamics. Wright-Allen Press, Cambridge
59. Meadows D, Beherens W III, Meadows D, Nail R, Randers J, Zahn E (ed) (1974) Dynamics of Growth in a Finite World. Pegasus Communications, Waltham
60. Meadows D, Meadows D (1977) Towards Global Equilibrium: Collected Papers. MIT Press, Cambridge
61. Meadows D, Meadows D, Randers J (1972) The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind. Universe Books, New York
62. Meadows D, Meadows D, Randers J (1992) Beyond the Limits. Chelsea Green Publishing Company, Post Mills
63. Meadows D, Randers J, Meadows D (2004) Limits to Growth: The 30-Year Update. Chelsea Green Publishing Company, White River Junction
64. Miyakawa T (ed) (1999) The Science of Public Policy: Essential Readings in Policy Sciences 1. Routledge, London
65. Morecroft J (2007) Strategic Modelling and Business Dynamics: A Feedback Systems Approach. Wiley, West Sussex
66. Naill R (1977) Managing the Energy Transition. Ballinger Publishing Co, Cambridge
67. Otto P, Struben J (2004) Gloucester Fishery: Insights from a Group Modeling Intervention. Syst Dyn Rev 20(4):287–312
68. Reagan-Cirincione P, Shuman S, Richardson GP, Dorf S (1991) Decision modeling: Tools for Strategic Thinking. Interfaces 21(6):52–65
69. Richardson GP (1991) System Dynamics: Simulation for Policy Analysis from a Feedback Perspective. In: Fishwick P, Luker P (eds) Qualitative Simulation Modeling and Analysis. Springer, New York
70. Richardson GP (1996) System Dynamics. In: Gass S, Harris C (eds) Encyclopedia of Operations Research and Management Science. Kluwer Academic Publishers, Norwell
71. Richardson G, Andersen D (1979) Teaching for Research in System Dynamics. Dynamica 5(3)
72. Richardson G, Andersen D (1995) Teamwork in Group Model Building. Syst Dyn Rev 11(2):113–137
73. Richardson G, Lamitie R (1989) Improving Connecticut School Aid: A Case Study with Model-Based Policy Analysis. J Educ Financ 15(2):169–188
74. Richardson G, Pugh J (1981) Introduction to System Dynamics Modeling. Pegasus Communications, Waltham
75. Richmond B (1997) The Strategic Forum: Aligning Objectives Strategy and Process. Syst Dyn Rev 13(2):131–148
76. Roberts C, Dangerfield B (1990) Modelling the Epidemiological Consequences of HIV Infection and AIDS: a Contribution from Operational Research. J Oper Res Soc 41(4):273–289
77. Roberts N (1983) An Introductory Curriculum in System Dynamics. Dynamica 9(1):40–42
78. Roberts N, Andersen DF, Deal RM, Grant MS, Schaffer WA (1983) Introduction to Computer Simulation: a System Dynamics Modeling Approach. Addison Wesley, Reading
79. Rohrbaugh J (2000) The Use of System Dynamics in Decision Conferencing: Implementing Welfare Reform in New York State. In: Garson G (ed) Handbook of Public Information Systems. Marcel Dekker, New York, pp 521–533
80. Rouwette E (2003) Group Model Building as Mutual Persuasion. Wolf Legal Publishers, Nijmegen
81. Rouwette E, Vennix J, Van Mullekom T (2002) Group Model Building Effectiveness: a Review of Assessment Studies. Syst Dyn Rev 18(1):5–45
82. Saeed K (1994) Development Planning and Policy Design: A System Dynamics Approach. Ashgate, Aldershot
83. Saeed K (1996) The Dynamics of Collegial Systems in the Developing Countries. High Educ Policy 9(1):75–86
84. Saeed K (1998) Towards Sustainable Development, 2nd Edition: Essays on System Analysis of National Policy. Ashgate, Aldershot
85. Saeed K (2003) Articulating Developmental Problems for Policy Intervention: A System Dynamics Modeling Approach. Simul Gaming 34(3):409–436
86. Saysel A, Barlas Y et al (2002) Environmental Sustainability in an Agricultural Development Project: A System Dynamics Approach. J Environ Manag 64(3):247–260
87. Senge P (1990) The Fifth Discipline: the Art and Practice of the Learning Organization. Doubleday/Currency, New York
88. Senge P, Asay D (1988) Rethinking the Healthcare System. Healthc Reform J 31(3):32–34, 44–45 5
89. Stave K (2002) Using SD to Improve Public Participation in Environment Decisions. Syst Dyn Rev 18(2):139–167
90. Steed M (1992) Stella, a Simulation Construction Kit: Cognitive Process and Educational Implications. J Comput Math Sci Teach 11(1):39–52
91. Stenberg L (1980) A Modeling Procedure for the Public Policy Scene. In: Randers J (ed) Elements of the System Dynamics Method. Pegasus Communications, Waltham, pp 257–288
92. Sterman J (1988) People Express Management Flight Simulator: Simulation Game, Briefing Book, and Simulator Guide. http://web.mit.edu/jsterman/www/SDG/MFS/PE.html
93. Sterman J (1992) Teaching Takes Off: Flight Simulators for Management Education. OR/MS Today (October):40–44
94. Sterman J (1996) A Skeptic's Guide to Computer Models. In: Richardson GP (ed) Modelling for Management. Dartmouth Publishing Company, Aldershot
95. Sterman J (2000) Business Dynamics: Systems Thinking and Modeling for a Complex World. Irwin/McGraw-Hill, Boston
96. Taylor K, Lane D (1998) Simulation Applied to Health Services: Opportunities for Applying the System Dynamics Approach. J Health Serv Res Policy 3(4):226–232
97. van Ackere A, Smith P (1999) Towards a Macro Model of National Health Service Waiting Lists. Syst Dyn Rev 15(3):225–253
98. Vennix J (1996) Group model building: Facilitating team learning using system dynamics. Wiley, Chichester
99. Waggoner M (1984) The New Technologies versus the Lecture Tradition in Higher Education: Is Change Possible? Educ Technol 24(3):7–13
100. Warren K (1999) Dynamics of Strategy. Bus Strateg Rev 10(3):1–16
101. Warren K (2002) Competitive Strategy Dynamics. Wiley, Chichester
102. Warren K (2005) Improving Strategic Management with the Fundamental Principles of System Dynamics. Syst Dyn Rev 21(4):329–350
103. Wils A, Kamiya M et al (1998) Threats to Sustainability: Simulating Conflict Within and Between Nations. Syst Dyn Rev 14(2–3):129–162
104. Wolstenholme E (1992) The Definition and Application of a Stepwise Approach to Model Conceptualization and Analysis. Eur J Oper Res 59(1):123–136
105. Wolstenholme E (1993) A Case Study in Community Care Using Systems Thinking. J Oper Res Soc 44(9):925–934

106. Wolstenholme E (1999) A Patient Flow Perspective of UK Health Services: Exploring the Case for New Intermediate Care Initiatives. Syst Dyn Rev 15(3):253–273
107. Wolstenholme E, Coyle RG (1983) The Development of System Dynamics as a Methodology for System Description and Qualitative Analysis. J Oper Res Soc 34(7):569–581
108. Zagonel A, Andersen D, Richardson GP, Rohrbaugh J (2004) Using Simulation Models to Address "What If" Questions about Welfare Reform. J Policy Anal Manag 23(4):890–901

Books and Reviews

Coyle RG (1998) The Practice of System Dynamics: Milestones, Lessons and Ideas from 30 Years of Experience. Syst Dyn Rev 14(4):343–365

Forrester J (1961) Industrial Dynamics. Pegasus Communications, Waltham

Maani KE, Cavana RY (2007) Systems Thinking, System Dynamics: Managing Change and Complexity. Pearson Education (NZ) Ltd, Auckland

MacDonald R (1998) Reducing Traffic Safety Deaths: A System Dynamics Perspective. In: 16th International Conference of the System Dynamics Society, Quebec '98, System Dynamics Society Quebec City

Morecroft JDW, Sterman JD (eds) (1994) Modeling for Learning Organizations, System Dynamics Series. Productivity Press, Portland

Wolstenholme EF (1990) System Enquiry: A System Dynamics Approach. Wiley, Chichester

Q

QSAR Modeling and QSAR Based Virtual Screening, Complexity and Challenges of Modern

Alexander Tropsha
School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, USA

Article Outline

Glossary
Definition of the Subject
Introduction
The Complexity of Modern Datasets and QSAR Methods
Recent Progress in Chemical Descriptor Research: 2D Chirality and Fragment Descriptors
Critical Importance of Model Validation
Applicability Domains of QSAR Models
Combinatorial QSAR and Model Acceptability Criteria
Predictive QSAR Modeling Workflow and its Application to Virtual Screening
Computational Chemical Toxicology
Conclusions. Rapid Growth of Publicly Available Databases and Emerging QSAR Research Strategies
Acknowledgments
Bibliography

Glossary

QSAR – quantitative structure activity relationships a method to predict biological activity from chemical structure

Combi-QSAR – combinatorial QSAR implies concurrent generation of QSAR models using all possible binary combinations of different descriptor types and model optimization techniques

QSAR modeling workflow a hierarchy of QSAR model development and validation protocols that should be followed to establish validated and externally predictive model

kNN – k nearest neighbors a pattern recognition approach used in deriving non-linear QSAR models

Model validation a set of computational routines used to establish internal and external predictive power of QSAR models

Applicability domain restriction on the chemistry space occupied by compounds for which the prediction of their activity from training set QSAR model is considered reliable.

Definition of the Subject

In the early days of Quantitative Structure Activity Relationship (QSAR) modeling the experimental datasets were relatively small and chemically congeneric and the techniques employed were relatively unsophisticated. Since then, the size and complexity of experimental datasets has increased dramatically, and so had the complexity and challenges of data analytical approaches. This chapter examines the strategy and the output of the modern QSAR modeling approaches especially as applied to complex biomolecular datasets. We discuss a data-analytical modeling workflow developed in our laboratory that incorporates modules for combinatorial QSAR model development (i. e., using all possible binary combinations of available descriptor sets and statistical data modeling techniques), rigorous model validation, and virtual screening of available chemical databases to identify novel biologically active compounds. Our approach places particular emphasis on model validation as well as on the need to define model applicability domains in the chemistry space. We present examples of studies where the application of rigorously validated QSAR models for virtual screening identified computational hits that were confirmed by subsequent experimental investigations. The emerging focus of QSAR modeling on target property forecasting brings it forward as predictive, as opposed to evaluative, modeling approach.

Introduction

QSAR methodology was introduced by Hansch et al. in early 1960th [1]. The original approach stemmed from linear free-energy relationships and is based upon the assumption that the difference in structural properties accounts for the difference in biological activities of compounds. According to this approach, the structural changes that affect the biological activities of a set of congeners are of three major types: electronic, steric, and hydrophobic [2]. These structural properties are often described by Hammett electronic constants [2], Verloop STERIMOL parameters [3], hydrophobic constants [2], etc. The relationship between a biological activity (or chemical property) and the structural parameters is obtained using linear or multiple linear regression (MLR) analysis. The fundamentals and applications of this method in chemistry and biology have been summarized by Hansch and Leo [2].

The original QSAR method was relatively straightforward; the datasets were small, and so was the number of relatively simple physical chemical descriptors used in modeling building. Today, even a person familiar with the basic principles of QSAR could be easily confused by the diversity of methodologies and naming conventions used in modern QSAR studies. Two-dimensional (2D) or three-dimensional (3D) QSAR, variable selection or Artificial Neural Network methods, Comparative Molecular Field Analysis (CoMFA) or binary QSAR present examples of various terms that may appear to describe totally independent approaches, which can not be generalized or even easily compared to each other. Thus, modern QSAR modeling is a very complex and complicated field requiring deep understanding and thorough practicing to develop robust models. Nevertheless, an attempt can be made to provide some unifying concepts that underlie practically any QSAR methodology.

Indeed, any QSAR method can be generally defined as an application of mathematical and statistical methods to the problem of finding empirical relationships (QSAR models) of the form $P_i = \hat{k}(D_1, D_2, \ldots, D_n)$, where P_i are biological activities (or other properties of interest) of molecules, $D_1, D_2, \ldots, D_n$ are calculated (or, sometimes, experimentally measured) structural properties (molecular descriptors) of compounds, and $\hat{k}$ is some empirically established mathematical transformation that should be applied to descriptors to calculate the property values for all molecules. The relationship between values of descriptors D and target properties P can be linear or non-linear. The example of the former relationship is given by multiple linear regression (MLR) common to the Hansch QSAR approach [1], where target property can be predicted directly from the descriptor values. On the contrary, nearest neighbor QSAR methods serve as examples of non-linear techniques where descriptor values are used in characterizing chemical similarities between molecules, which are then used to infer compound activity. The goal of QSAR modeling is to establish a trend in the descriptor values, which parallels the trend in biological activity. In essence, all QSAR approaches imply, directly or indirectly, a simple similarity principle, which for a long time has provided a foundation for the experimental medicinal chemistry: compounds with similar structures are expected to have similar biological activities.

The subsequent sections of this chapter present a brief overview of the modern QSAR modeling field without going into specific details of any particular technique; introduce the predictive QSAR modeling workflow developed in our group; present examples of successful applications of the workflow to several datasets resulting in experimentally confirmed computational predictions of biologically active compounds by the means of virtual screening; address the issue of fruitful collaborations between QSAR modelers in developing and supporting "best practices" in QSAR modeling; and summarize most important challenges that the field of QSAR modeling is facing today.

The Complexity of Modern Datasets and QSAR Methods

Traditionally, QSAR approaches have been applied to modeling datasets tested against a single target, e. g., in specific enzymatic or receptor-binding assays. Recent experimental advances in high-throughput screening and multi-target testing of compound libraries have led to the establishment of datasets of biologically active compounds (often publicly available) that we shall define as complex. A complex dataset could include a library of compounds tested against multiple targets, or have the target property measured in the form of gene or protein expression profiles across many genes (chemical genomics), or could be formed by diverse compounds tested against a complex assay where multiple mechanisms leading to the measured response could be involved (e. g., carcinogenicity or mutagenicity). The examples of complex datasets include Pubchem [4], GPCR ligands [5], NCI [6], US FDA [7], NIEHS [8], and EPA DSS-Tox [9] (see more examples in a recent review [10]). Naturally, the complex datasets call for the development of more sophisticated computational tools and corresponding models.

Modern QSAR approaches are characterized by the use of multiple descriptors of chemical structure com-

bined with the application of both linear and non-linear optimization approaches, and a strong emphasis on rigorous model validation to afford robust and predictive models. The most important recent developments in the field concur with a substantial increase in the size of experimental datasets available for the analysis and an increased application of QSAR models as virtual screening tools to discover biologically active molecules in chemical databases and/or virtual chemical libraries [11]. The latter focus differs substantially from the traditional emphasis on developing so called explanatory QSAR models characterized by high statistical significance but only as applied to training sets of molecules with known chemical structure and biological activity.

The differences in various QSAR methodologies can be understood in terms of the types of *target property* values, *descriptors*, and *optimization* algorithms used to relate descriptors to the target properties and generate statistically significant models. *Target* properties (regarded as dependent variables in statistical data modeling sense) can be generally of three types: *continuous* (i. e., real values covering certain range, e. g., IC_{50} values, or binding constants); *categorical related, or rank-based* (e. g., classes of rank-ordered target properties covering certain range of values, e. g., classes of metabolic stability such as unstable, moderately stable, stable); and *categorical unrelated* (i. e., classes of target properties that do not relate to each other in any continuum, e. g., compounds that belong to different pharmacological classes). As simple as it appears, understanding this classification is actually very important since the choice of descriptor types and modeling techniques as well as model accuracy metrics is often dictated by the type of the target properties. Thus, in general the latter two types require *classification* modeling approaches whereas the former type of the target properties allows the use of (multi)linear regression type modeling. The corresponding methods of data analysis are referred to as either classification or continuous property QSAR.

Many QSAR approaches have been developed during the past few decades (e. g., see recent reviews [12,13]). The major differences between various approaches are due to structural parameters (descriptors) used to characterize molecules and the mathematical approaches used to establish a correlation between descriptor values and biological activity. Most of the modeling techniques assume a linear relationship between molecular descriptors and a target property, which may be an adequate methodology for many datasets. However, the advances in combinatorial chemistry and high throughput screening technologies have resulted in the explosive growth of the amount of structural and biological data making the problem of developing robust QSAR models more challenging. This progress has provided an impetus for the development of fast, nonlinear QSAR methods that can capture structure-activity relationships for large and complex data. New nonlinear methods of multivariate analysis such as different types of Artificial Neural Networks [14,15,16,17], Generalized Linear Models [15,18,19,20], Classification and Regression Trees [18,21,22,23,24], Random Forests [25,26,27], MARS (Multivariate Adaptive Regression Splines) [27,28], Support Vector Machines [29,30,31,32], and some other methods have become routine tools in QSAR studies. Interesting examples of applications have been reported for all types of the above methods. In some cases the comparisons between different techniques as applied to the same dataset have been made but in general there appears to be no universal QSAR approach that produces the best QSAR models for any datasets.

For instance, several types of ANNs have been used in QSAR studies, including feed forward back propagation neural networks (FFBPNN), counter propagation neural networks (CPNN), radial basis function neural networks (RBFNN), Bayesian regularized neural networks (BRNN), etc. In [33], analysis of FFBPNN was carried out based on artificial and real data. It has been shown that for neural nets capable of generating linear models, increasing the number of nodes in the hidden layer can improve statistics for the training sets. However, leave-one-out and leave-some-out cross-validation statistics as well as prediction power for the test sets start decreasing when the number of hidden layer nodes passes some threshold that depends on the dataset. On the whole, according to [33], linear FFBPNNs compare unfavorably with more simple MLR methodology. No conclusions have been made for neural nets including quadratic and indicator variables as well as ANNs with other architectures.

In [34], probabilistic and generalized regression neural networks (PNN and GRNN, respectively), which are variants of RBFNN, have been used for classification and continuous QSAR modeling for a data set of soluble epoxide hydrolase inhibitors and prediction of aqueous solubility of some classes of small organic molecules. Final models appeared to have comparable predictive power to kNN classification, MLR and feed-forward neural networks, but they included significantly lower number of descriptors.

Overall, the advantage of ANNs is that they do not explicitly define the relationship between descriptors and target property making these techniques applicable to datasets of complex chemical nature and/or complex response variable where the relationship between descriptor values and the response variable are likely to be non-linear. On the other hand, for the same reason the interpretation

of ANN QSAR models is very difficult, if not impossible. ANNs can be used as part of the combi-QSAR strategy discussed below and may prove to be particularly useful (as with any other QSAR approach) for specific datasets.

Our laboratory among others has concentrated on the development of automated QSAR approaches based on variable selection and stochastic optimization. The examples of methods developed or implemented in our laboratory include k-Nearest Neighbors (kNN) [31,35,36], Simulated Annealing-Partial Least Squares (SA-PLS) [37], Support Vector Machines (SVM) [38,39,40], and the Automated Lazy Learning QSAR (ALL-QSAR) [41].

Recent Progress in Chemical Descriptor Research: 2D Chirality and Fragment Descriptors

In most cases the accurate formal description of molecules is the key to developing successful QSAR models. Many approaches to generating descriptors have been developed during the long history of QSAR modeling; programs such as Dragon [42] or MolConnZ [43] could calculate 100s or even 1000s of molecular descriptors. The molecular descriptors that are commonly used in QSAR studies can be divided into several groups. Physicochemical descriptors reflect electronegativity, partial charges, hydrogen bond acceptor and donor ability, molecular weight, logP, surface area, etc. Another class of descriptors are three-dimensional (3D) descriptors which are derived from spatial structures of molecules (e. g. CoMFA [44], CoMSIA [45], QSiAR [46] etc.). Most of QSAR studies based on 3D descriptors require exhaustive conformational analysis and spatial alignment of molecules, and thus large computational and human resources (e. g., [47]). Approaches such as GRIND [48] and VolSurf [49] generate descriptors based on 3D interaction maps and do not require alignment, but still require conformational search. On the other hand, descriptors based on molecular graphs (e. g. molecular connectivity indices [50], molecular shape indices [51], E-state indices [52], etc.) are naturally insensitive to problems related to conformation or alignment. Since chemical graphs are planar, these descriptors are often referred to as two-dimensional (2D) descriptors. An important note is that one should not be confused by the perceived low dimensionality of 2D descriptors: a molecule is actually described in multidimensional 2D descriptor space, with the number of dimensions (i. e., descriptors) frequently higher than that for 3D descriptors. Thus any modern "2D" or "3D" QSAR model actually operates on a complex high-dimensional space of multiple chemical descriptors. In this regard, the tasks and challenges of modern QSAR are much more similar to those faced by, e. g., bioinformaticians trying to analyse multidimensional data on protein or gene expression profiles of say patients vs. healthy individuals. 2D descriptors have certain advantages over 3D descriptors: they can be easily calculated even for very large datasets and in most case they seem to capture sufficient information about chemical structures to enable important modeling tasks such as similarity searching or QSAR modeling. For instance, it was shown [53] that 2D descriptors outperformed 3D descriptors in chemical similarity and diversity analyses. Our studies have shown that the predictive power of QSAR models based on 2D descriptors is generally comparable (and sometimes superior) to that of the models based on 3D descriptors [35,54]. Finally, recent comparison of protein structure based vs. ligand based virtual screening results [55] demonstrated that the latter approach helped recover a higher number of confirmed hits than the former one.

Overall, the field of chemical descriptor development has probably reached its saturation (as far as active development of new generalized descriptors is concerned) although there are continuing reports on novel 3D descriptors such as for instance inductive descriptors [56]. Still, there remain challenges in developing specialized descriptors such as topological chirality descriptors and fragment descriptors that have been of special interest to several groups including ours. We briefly review recent advances in such specialized descriptor development research.

Chirality Descriptors

Many biologically-active compounds are in fact enantiospecific, and their chiralities are believed to directly influence their bioactivities because these compounds are recognized differently by their corresponding receptors. In 1999, the worldwide annual sales of chiral drugs exceeded $100 billion that constituted almost one-third of all drug sales, and in 2000, these numbers were $133 billion and 40%, respectively. It was projected that the sales for chiral drugs in 2008 could reach $200 billion [57].

In QSAR studies, taking chirality into account has become possible only after the development of the 3D QSAR methods such as the comparative molecular field analysis (CoMFA) [44]. Subsequently, many CoMFA-like methods have appeared (e. g., [45,46,58]). Evidently, the chirality in such methods is taken into account by default, since 3D molecular field values of chiral isomers are different. At the same time as mentioned above 2D QSAR methods offer a clear advantage over 3D methods since they require no conformational analysis, no alignment, and no 3D pharmacophore hypothesis and, as a result, can be easily fully automated. However, one of the main drawbacks

of QSAR approaches utilizing 2D molecular descriptors such as molecular connectivity indices (calculated with the MolconnZ program [43]) or atom pairs [59] has been their inability to take into account the chirality of atoms since the latter is a true 3D property. Naturally, this deficiency has severely limited the range of applications of 2D QSAR methods, especially as compared to 3D QSAR approaches. We shall discuss two approaches that circumvent this limitation of 2D descriptors.

Topological Chirality Indices

The idea of modifying conventional 2D descriptors to make them sensitive to chirality was put forward as early as in 1995 [60]. The authors introduced a so-called chiral factor equal to + 1 or − 1 for an atom in R- or S-configuration, respectively. This factor was used to derive several chirality descriptors which appeared to have different values for linear hexose isomers. The first attempt to consider chirality descriptors in QSAR studies was reported by Ortiz et al. in 1998 [61] when several chirality-sensitive molecular and charge topological indices were introduced. Both chirality descriptors and conventional MolconnZ descriptors have been implemented in QSAR studies of the series of D_2 dopamine and σ receptor ligands. It was shown that the resulting QSAR models had better statistics and predictive power for IC_{50} values of the ligands than those obtained with conventional descriptors only. In addition, using these chirality descriptors, a set of chiral barbiturates were correctly classified as sedatives or stimulants [61].

Several series of chirality molecular topological descriptors have been developed in our laboratory earlier [62]. These descriptors have been developed on the basis of conventional 2D topological descriptors of molecular graphs. They include modified molecular connectivity indices, Zagreb group indices, extended connectivity, overall connectivity, and topological charge indices. These modified descriptors make use of a term called chirality correction, which is added to the vertex degrees of asymmetric atoms in a molecular graph. They have been used to build predictive QSAR models for several datasets [63]. Similar approach was used to develop cis-trans (ZE) isomerism descriptors as well [64].

Chirality-Sensitive Atom Pair Descriptors

Atom pair descriptors were introduced by Carhart et al. [59] and implemented in the GenAP program in our laboratory. The conventional atom pair descriptors can not be applied for a data set of chiral compounds since their numerical values are identical for enantiomeric molecules. Thus, we have recently introduced chiral atom types and thereby augmented the conventional AP descriptors with the novel chiral atom pair (cAP) descriptors [65].

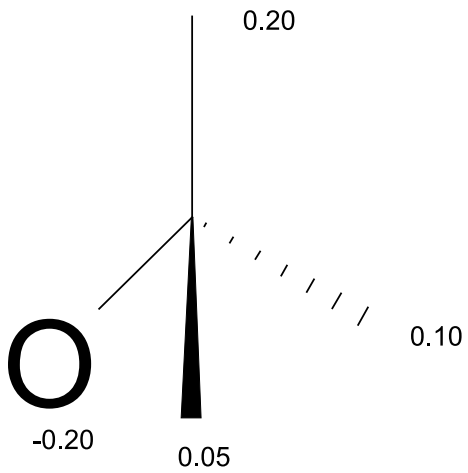

QSAR Modeling and QSAR Based Virtual Screening, Complexity and Challenges of Modern, Figure 1
Example of the chiral atom type definition based on partial charges of the substituents (here, three carbon atoms and one oxygen atom) at the chiral center

A cAP descriptor is defined as a pair of atoms separated by certain chemical graph distance (for 2D representation of a molecule) or a real physical distance (in 3D) where at least one atom of the pair is chiral. Our definition of a chiral atom is similar to the conventional IUPAC nomenclature but different in how we define the seniority of the substituents at the chiral atom. The difference is that the relative priorities of the substituents are not determined by their atomic numbers, but values of a certain property, such as atomic partial charge, van der Waals volume, etc. In the previous studies [65], Gasteiger–Huckel charges as implemented in the SYBYL software were used. In addition, our current implementation does not take into consideration properties of any atoms besides the substituents. Figure 1 shows an example of a chiral atom where all four substituents at a carbon atom have different partial charges.

Since the oxygen atom has the most negative partial charge, we assign the lowest seniority to it, and since the charges on the substituent increase in a counterclockwise order, the central atom is in R_q configuration (index q means that the configuration was defined using partial charges). (Note that the same carbon atom would not be considered chiral based on the standard nomenclature!) Using different atomic properties to assign chirality, we can introduce many different definitions of R and S con-

figurations of chiral atoms; this work is in progress in our laboratory.

Fragment Descriptors

2D descriptors such as molecular connectivity indices have one significant drawback: in most cases models based on 2D descriptors are difficult (if not impossible) to interpret, because most of 2D descriptors have no clear physicochemical meaning. To overcome this drawback, molecular structural keys, molecular fingerprints and molecular holograms are used. Unlike molecular descriptors based on physicochemical properties, 2D and 3D descriptors, structural keys, molecular fingerprints and molecular holograms can provide a mechanistic explanation to the target property of active molecules.

Structural keys and molecular fingerprints were initially introduced by chemical information system companies for querying chemical databases. A structural key of a molecule is defined as a bit string; each bit in state "on" (one) or "off" (zero) represents the presence or absence of a certain atom or group (fragment) in a molecule [66,67]. Alternatively, it is a string of numbers; each number represents how many times an atom or a fragment is represented in a molecule. A predefined library of fragments is used. Usually bit strings are very long and for small molecules almost all bits are zeros.

To reduce the memory necessary to store structural keys, hashing [68] is used to map structural keys on a shorter string of a predefined length. For example, typical Daylight fingerprints have 512 or 1024 bits, but any power of two can be generated. The MACCS public fingerprints include 166 or 320 keys which encapsulate 966 original (private) bits [69]. Barnard Chemical Information Systems (BCI) fingerprints are generated combining both the Daylight and MDL approaches, and molecular fingerprint bit lengths are about 5,000.

Molecular holograms used in HQSAR are based on the structural keys containing the number of times each bit was set. This information is stored in the hologram string which is a string of integers rather than a bit string. As fingerprints, molecular holograms are based on a predefined library of fragments (chemical groups). The development of frequent common subgraph based molecular descriptors represent an important and interesting area of current descriptor research. For instance, recurring substructures in a group of chemicals with similar activity can be identified by finding frequent subgraphs in their related graphical representations. The recurring substructures can implicate chemical features responsible for compounds' biological activities [70].

Critical Importance of Model Validation

It should sound almost axiomatic that validation should be natural part of any model development process. Indeed, what is the (ultimate) purpose of any modeling approach such as QSAR, if not developing models with a significant external predictive power? Unfortunately, as we and others have indicated in many publications (e. g., [46,71,72]), the entire field of QSAR modeling counting nearly 50 years of a rich history has been plagued with insufficient attention paid to the subject of external validation. Indeed, most practitioners have merely presumed that internally cross-validated models built from available training set data should be externally predictive. However, the overwhelming prevalence of QSAR publications exploring small to medium size datasets to produce models with little statistical significance led to the recent editorial published by the *J. Chem. Info. Model.* two years ago [73] that explicitly discouraged researchers from submitting the "introspective" QSAR/QSPR publications and requested that "evidence that any reported QSAR/QSPR model has been properly validated using data not in the training set must be provided". We and others have demonstrated (as we detail below) that the training set statistics using most common internal validation techniques such as leave-one-out or even leave-many-out cross-validation approaches is insufficient and the statistical figures of merit of such models serve as misleading indicators of the external predictive power of QSAR models [72]. We shall refer to studies limited to exploring training sets (and associated model statistical parameters) as *"narcissistic modeling"*.

In our highly cited publication "Beware of q^2!" [71], we have demonstrated the insufficiency of the training set statistics for developing externally predictive QSAR models and formulated the main principles of model validation. At the time of that publication in 2002, the majority of papers on QSAR analysis ignored any model validation except for the cross-validation, performed during model development. Despite earlier observations of several authors [74,75,76] warning that high cross-validated correlation coefficient $R^2(q^2)$ is the necessary, but not the sufficient condition for the model to have high predictive power, many authors continued to consider q^2 as the only parameter characterizing the predictive power of QSAR models. In [71] we have shown that the predictive power of QSAR models can be claimed only if the model was successfully applied for prediction of the external test set compounds, which were not used in the model development. We have demonstrated that the majority of the models with high q^2 values have poor predictive power when applied for prediction of compounds in

the external test set. We believe that paying attention only to the training set statistics equates to "*narcissistic*" modeling in a sense that such models appear "beautiful" only in the eyes of their developers but provide little if any utility to potential users ("viewers") of these models. In another publication [72] the importance of rigorous validation was again emphasized as a crucial, integral component of model development. Several examples of published QSPR models with high fitted accuracy for the training sets, which failed rigorous validation tests, have been considered. We presented a set of simple guidelines for developing validated and predictive QSPR models and discussed several validation strategies such as the randomization of the response variable (Y-randomization) external validation using rational division of a dataset into training and test sets. We highlighted the need to establish the domain of model applicability in the chemical space to flag molecules for which predictions may be unreliable, and discussed some algorithms that can be used for this purpose. We advocated the broad use of these guidelines in the development of predictive QSPR models [72,77,78].

Nowadays, most of the QSAR modeling studies include validation of QSAR models. However, some authors still publish QSAR models which lack proper validation. For example, [79] developed a model to predict gastrointestinal absorption of drugs, but did not validate it using a test set. Verma et al. [80] developed QSAR models for predicting cytotoxicity of a group of compounds with antiovarian cancer activity which were validated only using cross-validation and Y-randomization.

At the 37th Joint Meeting of Chemicals Committee and Working Party on Chemicals, Pesticides & Biotechnology, held in Paris on 17–19 November 2004, the OECD (Organization for Economic Co-operation and Development) member countries adopted the following five principles that valid (Q)SAR models should follow to allow their use in regulatory assessment of chemical safety. (i) a defined endpoint; (ii) an unambiguous algorithm; (iii) a defined domain of applicability; (iv) appropriate measures of goodness-of-fit, robustness and predictivity; (v) a mechanistic interpretation, if possible. Since then, most of the European authors publishing in QSAR area include a statement that their models fully comply with OECD principles (e. g., see [81,82,83,84]). For instance, two aspects of QSAR modeling outlined in the OECD principles are considered by Estrada and Patlewicz [85]. The first aspect concerns the theoretical approaches used in chemistry in general, and in QSAR in particular, i. e., which method should be selected for theoretical studies: more sophisticated and complex or more simple. The authors criticize the common belief that applying more sophisticated methods should always lead to significantly better results. They considered an example of polycyclic aromatic hydrocarbons (PAHs) the toxicity of which is believed to depend on the energy gap between HOMO and LUMO values. The authors show that a simple Hückel Molecular Orbital theory gives practically the same values of HOMO and LUMO as the sophisticated ab initio methods yet the calculations are 10^{-4} to 10^{-7} times faster. They reach the conclusion that if a more simple method is capable of giving results better or similar to those of more sophisticated method, one should naturally use a more simple method!

The second aspect relates to so called "mechanistic" QSAR. Some authors (e. g., [86]) prefer descriptors which are mechanistically interpretable. On the other hand, Estrada and Patlewicz [85] argue that in many cases a biological response is a result of a multitude of different processes, some of which can be even not known, and its a posteriori mechanistic interpretation is difficult if not impossible. The authors suggest an alternative approach where a biological system is considered as a black box, when considering several possible mechanisms would be more productive. At the same time, some variables included in the model can describe several different mechanisms simultaneously, e. g. $\log P$, so in many cases it makes no sense to suggest that the use of this descriptor in QSAR models affords any mechanistic interpretation (see also [87]). We would add that descriptors which give better models in terms of their predictive power are actually preferable. We consider building predictive models as the main goal of QSAR analysis. Of course, interpretation of the model is also important, and if it is possible, it should be done. However, in many cases it is impossible, even when models with high predictive power have been obtained (e. g., best models were found to be those built using the molecular connectivity indices but these models were disregarded by the authors for the lack of mechanistic interpretability [86]). We believe that mechanistic interpretation of the *externally validated* QSAR model is an important a posteriori exercise that should be done after the model has been internally and externally validated, and descriptors that afford models with the highest predictive power should be always used preferentially.

Validation of QSAR models remains one of the most critical problems of QSAR. Recently, we have extended our requirements for the validation of multiple QSAR models selected by acceptable statistics criteria of prediction of the test set [88]. Additional studies in this critical component of QSAR modeling should establish reliable and commonly accepted "good practices" for model development.

Applicability Domains of QSAR Models

One of the most important problems in QSAR analysis is establishing the models' domain of applicability in the chemistry space. In the absence of the applicability domain, each model can formally predict the activity of any compound, even with a completely different structure from those included in the training set. Thus, the absence of the model applicability domain as a mandatory component of any QSAR model would lead to the unjustified extrapolation of the model in the chemistry space and, as a result, a high likelihood of inaccurate predictions. In our research we have always paid particular attention to this issue [36,41,62,72,89,90,91,92]. The need for establishing the applicability domain for every model adds another critical degree of complexity to the model building process.

The applicability domain problem has been addressed by many researchers. Mandel [93] introduced the so called Effective Prediction Domain which was based on the ranges of descriptors included in the regression equation. Afantitis et al. [94] built a multiple linear regression model for a dataset of apoptotic agents. They defined the applicability domain for each compound as a leverage defined as a corresponding diagonal element of the hat matrix. In fact, it is a method for detecting possible leverage outliers. If for some compound leverage is higher than $3K/N$, where K is the number of descriptors and N is the number of compounds, the compound is an outlier. To use this approach, for each external compound it would be necessary to recalculate the leverage. Netzeva et al. [95] and Saliner et al. [83] defined the applicability domain by ranges of descriptors, i. e., in fact, as a subspace occupied by representative points in the descriptor space. This definition of the applicability domain has a significant drawback, because the representative points could be found only in a small part of the hyper-parallelepiped corresponding to descriptor ranges rather than distributed uniformly. A similar definition of the applicability domain was proposed by Tong et al. [96]. The authors built QSAR models for two datasets of estrogen receptor ligands using the Decision Forest method and studied the dependence of the model predictive power vs. the applicability domain threshold. The prediction accuracy within the domain is defined as a ratio of the number of correct predictions to the total number of compounds in the domain. The accuracy was changing from about 90% for the initial applicability domain to about 50% when the applicability domain increased by 30%. Interestingly, for one of the datasets the prediction accuracy was increasing until the domain was extended by about 20%. Another important aspect of this study was that the authors defined the confidence level of prediction. The probability that a compound belongs to a certain class was defined as the percentage of active compounds in the leaf node that the compound belongs to. The authors found (as expected), that the confidence level correlated with the prediction accuracy.

In [97] a lazy learning kNN-like method was applied for the prediction of rodent carcinogenicity and Salmonella mutagenicity. The applicability domain was defined by a so-called confidence index. A compound was assigned to one of the two classes by a weighted majority vote of its nearest neighbors. The confidence index is the weighted majority quote divided by the number of nearest neighbors. If the absolute value of the confidence index is low (< 0.05) a compound is said to be out of the applicability domain. This definition of the applicability domain captures the areas in the descriptor space where compounds of both classes are close to each other, and possibly mixed. In this area the precise and accurate prediction of a compound's class is impossible. A Tanimoto-like coefficient is used as a similarity measure. Nearest neighbors are defined by the value of this coefficient higher than 0.3, which limits the possibility of over-extrapolation.

In most of our QSAR studies we have defined the applicability domain as the distance cutoff value $D_{\text{cutoff}} = \langle D \rangle + zs$, where Z is a similarity threshold parameter defined by a user, and $\langle D \rangle$ and s are the average and standard deviation of all Euclidian distances in the multidimensional descriptor space between each compound and its nearest neighbors for all compounds in the training set (e. g., see [78]). This definition of the applicability domain has several major drawbacks which we continue to address in our ongoing studies: (i) Currently, applicability domain is direction-independent in the descriptor space. We shall consider the directions in the descriptor space in which the distribution of representative points has smaller spread as less important than those that have higher spread. Thus, the applicability domain will be represented as a multidimensional ellipsoid in the principal component space. (ii) Too strict definition of the applicability domain: if a compound is outside of the model applicability domain, we currently do not predict its activity. Naturally, we shall establish the lower and upper bounds for the applicability domain. (iii) Finally, it seems reasonable to introduce a confidence level of prediction, which will depend on the distance of the compound under prediction from its nearest neighbor of the training set. These considerations provide just a few examples that illustrate the importance of ongoing research in this area of QSAR modeling. Not surprisingly, the model applicability domain was the subject of a special symposium organized at the most recent 235th

meeting of the American Chemical Society in New Orleans, LO.

Combinatorial QSAR and Model Acceptability Criteria

The chief hypothesis of the combi-QSAR approach that we introduced in recent publications [39,98] is that if an implicit structure-activity relationship exists for a given data set, it can be formally manifested via a variety of QSAR models obtained with different descriptors and optimization protocols. Our experience indicates that there is no universal QSAR method that is guaranteed to give the best results for any dataset. Thus we believe that multiple alternative QSAR models should be developed (as opposed to a single model using some favorite QSAR method) for each dataset to identify the most successful technique in the context of the given dataset. Since QSAR modeling is relatively fast, these alternative models could be explored simultaneously when making predictions for external data sets. The consensus predictions of biological activity for novel test set compounds on the basis of several QSAR models, especially when they converge, are more reliable and provide better justification for the experimental exploration of hits.

Our current approach to combi-QSAR modeling is summarized on the workflow diagram (Fig. 2). Our experience suggests that QSAR is a highly experimental area of statistical data modeling where it is impossible to decide a priori as to which particular QSAR modeling method will prove most successful. To achieve QSAR models of the highest internal, and most importantly, *external* accuracy, the combi-QSAR approach explores all possible binary combinations of various descriptor types and optimization methods along with external model validation. Each combination of descriptor sets and optimization techniques is likely to capture certain unique aspects of the structure-activity relationship. Since our ultimate goal is to use the resulting models as reliable activity (property) predictors, application of different combinations of modeling techniques and descriptor sets will increase our chances for success. All types of descriptors and modeling techniques are available within our laboratory and are described in detail in our recent publications on the implementation of the combi-QSAR strategy [39,98].

In our critical publications [71,72] we have recommended a set of statistical criteria which must be satisfied by a predictive model. For continuous QSAR, criteria that we will follow in developing activity/property predictors are as follows: (i) correlation coefficient R between the predicted and observed activities; (ii) coefficients of determination [99] (predicted versus observed activities R_0^2, and observed versus predicted activities $R_0'^2$ for regressions through the origin); (iii) slopes k and k' of regression lines through the origin. We consider a QSAR model *predictive*, if the following conditions are satisfied (i) $q^2 > 0.5$; (ii) $R^2 > 0.6$; (iii) $(R^2 - R_0^2)/R^2 < 0.1$ and $0.85 \le k \le 1.15$ or $(R^2 - R_0'^2)/R^2 < 0.1$ and $0.85 \le k' \le 1.15$; (iv) $\left|R_0^2 - R_0'^2\right| < 0.3$ where q^2 is the cross-validated correlation coefficient calculated for the training set, but all other criteria are calculated for the test set.

Predictive QSAR Modeling Workflow and its Application to Virtual Screening

Our experience in QSAR model development and validation has led us to establishing a complex strategy that is

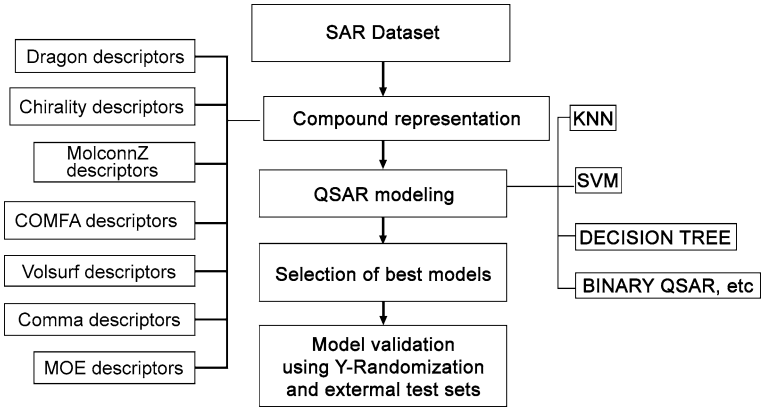

QSAR Modeling and QSAR Based Virtual Screening, Complexity and Challenges of Modern, Figure 2
Flow-chart of the combinatorial QSAR methodology. All descriptor sets and methods currently implemented in our laboratory are listed

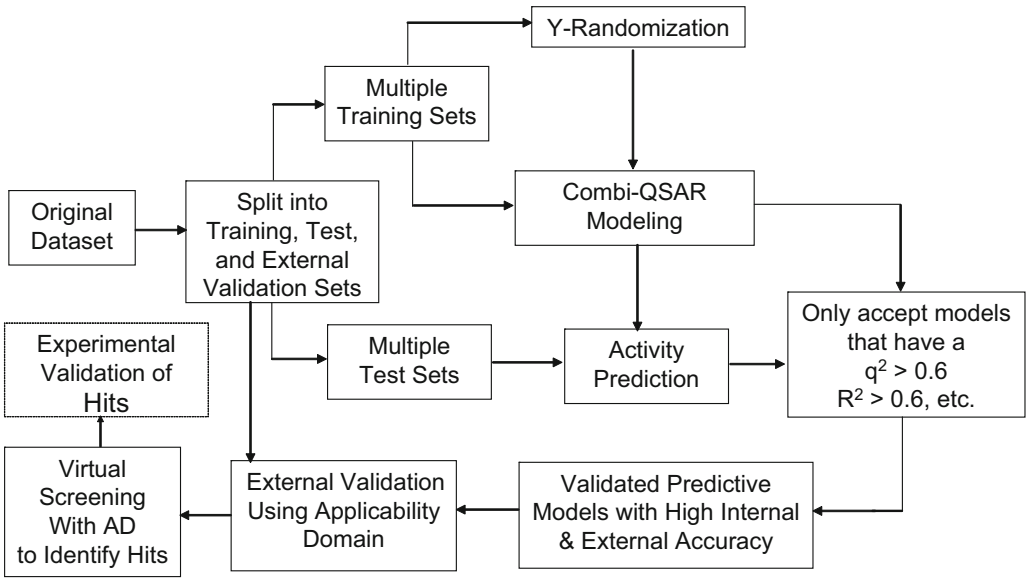

QSAR Modeling and QSAR Based Virtual Screening, Complexity and Challenges of Modern, Figure 3
Flowchart of predictive QSAR modeling framework based on the validated combi-QSAR models

summarized in Fig. 3. It describes the predictive QSAR modeling workflow focused on delivering validated models and ultimately, computational hits confirmed for the experimental validation. We start by randomly selecting a fraction of compounds (typically, 10–15%) as an external validation set. The remaining compounds are then divided rationally (using the Sphere Exclusion protocol implemented in our laboratory [78]) into multiple training and test sets that are used for model development and validation, respectively using criteria discussed in more detail below. We employ multiple QSAR techniques based on the combinatorial exploration of all possible pairs of descriptor sets coupled with various statistical data mining techniques (combi-QSAR) and select models characterized by high accuracy in predicting both training and test sets data. Validated models are finally tested using the evaluation set. The critical step of the external validation is the use of applicability domains. If external validation demonstrates the significant predictive power of the models we use all such models for virtual screening of available chemical databases (e. g., ZINC [100]) to identify putative active compounds and work with collaborators who could validate such hits experimentally. The entire approach is described in detail in several recent papers and reviews (e. g., [11,13,72]).

In our recent studies we were fortunate to recruit experimental collaborators who have validated computational hits identified through our modeling of anticonvulsants [36], HIV-1 reverse transcriptase inhibitors [101], D1 antagonists [31], antitumor compounds [102], and beta-lactamase inhibitors [103]. Thus, models resulting from this workflow could be used to prioritize the selection of chemicals for the experimental validation. However, since we can not generally guarantee that every prediction resulting from our modeling effort will be validated experimentally we can not include the experimental validation step as a mandatory part of the workflow on Fig. 3, which is why we used the dotted line for this component. We note that our approach shifts the emphasis on ensuring good (best) statistics for the model that fits known experimental data towards generating testable hypothesis about purported bioactive compounds. Thus, the output of the modeling has exactly same format as the input, i. e., chemical structures and (predicted) activities making model interpretation and utilization completely seamless for medicinal chemists.

The development of truly validated and predictive QSAR models affords their growing application in chemical data mining and combinatorial library design [37,104]. For example, three-dimensional (3D) stereoelectronic pharmacophore based on QSAR modeling was used recently to search the National Cancer Institute Repository of Small Molecules [6] to find new leads for inhibiting HIV type 1 reverse transcriptase at the non-nucleoside binding site [105].

Our studies have shown that QSAR models could be used successfully as virtual screening tools to discover compounds with the desired biological activity in chemical

databases or virtual libraries [11,31,36,102,103,106]. The discovery of novel bioactive chemical entities is the primary goal of computational drug discovery, and the development of validated and predictive QSAR models is critical to achieve this goal.

Computational Chemical Toxicology

Chemical toxicity can be associated with many hazardous biological effects such as gene damage, carcinogenicity, or induction of lethal rodent or human diseases. Toxicity presents an example of complex biological property (cf. the discussion of complex datasets above) where the underlying mechanisms are most frequently unknown and in fact multiple mechanisms are likely to be involved in mediating the end point response. For this reason, toxicity modeling is a very challenging QSAR problem yet alternative approaches such as protein structure based modeling can not be explored in most cases for the same reason of complex or unknown molecular mechanisms.

Although the experimental protocols for toxicity testing have been developed for many years and the cost of compound testing has reduced significantly, computational chemical toxicology continues to be a viable approach to reduce both the amount of efforts and the cost of experimental toxicity assessment [107]. Significant savings could be achieved if accurate predictions of potential toxicity could be used to prioritize compound selection for experimental testing. Many Quantitative Structure Activity Relationship (QSAR) studies have been conducted for different toxicity endpoints to address this challenge, [108,109,110,111]. However, the most critical limitation of many traditional QSAR studies has been their low *external* predictive power, i. e., their ability to predict accurately the underlying end point toxicity for compounds that were not used for model development. We discuss below the results of a recent important study of aquatic toxicity [112] since in our opinion this particular study may serve as a useful example to illustrate the complexity and power of modern QSAR modeling approaches.

The combinational QSAR modeling approach has been applied to a diverse series of organic compounds tested for aquatic toxicity in *Tetrahymena pyriformis* in the same laboratory over nearly a decade [113,114,115,116,117,118,119]. The unique aspect of this research was that it was conducted in collaboration between six academic groups specializing in cheminformatics and computational toxicology. The common goals for our virtual collaboratory were to explore the relative strengths of various QSAR approaches in their ability to develop robust and externally predictive models of this particular toxicity end point. We have endeavored to develop the most statistically robust, validated, and *externally* predictive QSAR models of aquatic toxicity. The members of our collaboratory included scientists from the University of North Carolina at Chapel Hill in the United States (**UNC**); University of Louis Pasteur (**ULP**) in France; University of Insubria (**UI**) in Italy; University of Kalmar (**UK**) in Sweden; Virtual Computational Chemistry Laboratory (**VCCLAB**) in Germany; and the University of British Columbia (**UBC**) in Canada. Each group relied on its own QSAR modeling approaches to develop toxicity models using the same modeling set, and we agreed to evaluate the realistic model performance using the same external validation set(s) (cf. Table 1 for the summary of approaches).

The *T. pyriformis* toxicity dataset used in this study was compiled from several publications of the Schultz group [113,120,121,122,123] as well as from data available at the Tetratox database website of (http://www.vet.utk.edu/TETRATOX/). After deleting duplicates as well as several compounds with conflicting test results and correcting several chemical structures in the original data sources, our final dataset included 983 unique compounds (the structural information is included in the Appendix). The dataset was randomly divided into two parts: 1) the modeling set of 644 compounds; 2) the validation set including 339 compounds. The former set was used for model development by each participating group and the latter set was used to estimate the external prediction power of each model as a universal metric of model performance. In addition, when this project was already well underway, a new dataset had become available from the most recent publication by the Schultz group [124]. It provided us with an additional *external* set to evaluate the predictive power and reliability of all QSAR models. Among compounds reported in [125] 110 were unique, i. e., not present among the original set of 983 compounds; thus, these 110 compounds formed the second independent validation set for our study.

Universal Statistical Figures of Merit for All Models

Different groups have employed different techniques and (sometimes) different statistical parameters to evaluate the performance of models developed independently for the modeling set (described below). To harmonize the results of this study the same standard parameters were chosen to describe each model's performance as applied to the modeling and external test set predictions. Thus, we have employed Q^2_{abs} (squared leave-one-out cross-validation correlation coefficient) for the modeling set, R^2_{abs} (frequently described as coefficient of determination) for the external

QSAR Modeling and QSAR Based Virtual Screening, Complexity and Challenges of Modern, Table 1
Overview of QSAR modeling approaches employed by six cheminformatic groups involved in this study

Group ID	Modeling Techniques	Descriptor Type	Applicability Domain Definition
UNC	kNN, SVM	MolconnZ, Dragon	Euclidean distance threshold between a test compound and compounds in the modeling set
ULP	MLR, SVM, kNN	Fragments (ISIDA), Molecular (CODESSA-Pro)	Euclidean distance threshold between a compound and compounds in the modeling set; bounding box
UI	MLR/OLS	Dragon	Leverage approach
UK	PLS	Dragon	Residual standard deviation and leverage within the PLSR model
VCCLAB	ASNN	E-state indices	Maximal correlation coefficient of the test molecule to the training set molecules in the space of models
UBC	MLR, ANN, SVM, PLS	IND_I	Undefined

validations sets, and MAE (mean absolute error) for the linear correlation between predicted (Y_{pred}) and experimental (Y_{exp}) data (here, $Y = pIGC_{50}$); these parameters are defined as follows:

$$Q^2_{abs} = 1 - \sum_Y (Y_{exp} - Y_{LOO})^2 \Big/ \sum_Y (Y_{exp} - \langle Y \rangle_{exp})^2 \quad (1)$$

$$R^2_{abs} = 1 - \sum_Y (Y_{exp} - Y_{pred})^2 \Big/ \sum_Y (Y_{exp} - \langle Y \rangle_{exp})^2 \quad (2)$$

$$MAE = \sum_Y \left| Y - Y_{pred} \right| \Big/ n \,. \quad (3)$$

Many other statistical characteristics can be used to evaluate model performance; however, we restricted ourselves to these three parameters that provide minimal but sufficient information concerning any model's ability to reproduce both the trends in experimental data for the test sets as well as mean accuracy of predicting all experimental values. The models were considered acceptable if R^2_{abs} exceeded 0.5.

Consensus QSAR Models of Aquatic Toxicity; Comparison Between Methods and Models

The objective of this study from methodological prospective was to explore the suitability of different QSAR modeling tools for the analysis of a dataset with an important toxicological endpoint. Typically, such datasets are analyzed with one (or several) modeling techniques, with a great emphasis on the (high value of) statistical parameters of the training set models. In this study, we went well beyond the modeling studies reported in the original publications [113,126,127,128,129,130] in several respects. First, we have compiled all reported data on chemical toxicity against *T. pyriformis* in a single large dataset and attempted to develop global QSAR models for the entire set. Second, we have employed multiple QSAR modeling techniques thanks to the engagement of six collaborating groups. Third, we have focused on defining model performance criteria not only using training set data but most importantly using external validation sets that were not used in model development in *any* way (unlike any common *cross*-validation procedure) [131]. This focus afforded us the opportunity to evaluate and compare all models using simple and objective universal criteria of *external* predictive accuracy, which in our opinion is the most important single figure of merit for a QSAR model that is of practical significance for experimental toxicologists. Fourth, we have explored the significance of applicability domains and the power of consensus modeling in maximizing the accuracy of external predictivity of our models.

We believe that results of our analysis lend a strong support for our strategy. Indeed, all models performed quite well for the training set (Table 2) with even the lowest Q^2_{abs} among them as high as 0.72. However, there was much greater variation between these models when looking at their (universal and objective) performance criteria as applied to the validation sets I and II (Table 2).

Of 15 QSAR approaches used in this study, nine implemented method-specific applicability domains. Models that did not define the AD showed a reduced predictive accuracy for the validation set II even though they yielded reasonable results for the validation set I. Only CODESSA-MLR (which did not employ any AD) approached in accuracy the lower bound of the models using the AD as measured by $R^2_{abs} = 0.58$ but still had one of the highest MAE of 0.47 (Table 2). On the other hand, among models employing the AD only kNN-MolconnZ had relatively low accuracy of prediction for the validation set II, with R^2_{abs}

QSAR Modeling and QSAR Based Virtual Screening, Complexity and Challenges of Modern, Table 2
Statistical results obtained with all toxicity QSAR models for the modeling and external validation sets

Model	Group ID	Modeling Set ($n = 644$)			Validation Set I ($n = 339$)			Validation Set II ($n = 110$)		
		Q^2_{abs}	MAE	Coverage (%)	R^2_{abs}	MAE	Coverage (%)	R^2_{abs}	MAE	Coverage (%)
kNN-Dragon	UNC	0.92	0.22	100	0.85	0.27	80.2	0.72	0.33	52.7
kNN-MolconnZ	UNC	0.91	0.23	99.8	0.84	0.30	84.3	0.44	0.39	53.6
SVM-Dragon	UNC	0.93	0.21	100	0.81	0.31	80.2	0.83	0.27	52.7
SVM-MolconnZ	UNC	0.89	0.25	100	0.83	0.30	84.3	0.55	0.37	53.6
ISIDA-kNN	ULP	0.77	0.37	100	0.73	0.36	78.5	0.63	0.37	42.7
ISIDA-SVM	ULP	0.95	0.15	100	0.76	0.32	100	0.38	0.50	100
ISIDA-MLR	ULP	0.94	0.20	100	0.81	0.31	95.9	0.65	0.41	51.8
CODESSA-MLR	ULP	0.72	0.42	100	0.71	0.44	100	0.58	0.47	100
OLS	UI	0.86	0.30	92.1	0.77	0.35	97.0	0.59	0.43	98.2
PLS	UK	0.88	0.28	97.7	0.81	0.34	96.1	0.59	0.40	95.5
ASNN	VCCLAB	0.83	0.31	83.9	0.87	0.28	87.4	0.75	0.32	71.8
PLS-IND_I	UBC	0.76	0.39	100	0.74	0.39	99.7	0.45	0.54	100
MLR-IND_I	UBC	0.77	0.39	100	0.75	0.40	99.7	0.46	0.53	100
ANN-IND_I	UBC	0.77	0.39	100	0.76	0.39	99.7	0.46	0.53	100
SVM-IND_I	UBC	0.79	0.31	100	0.79	0.35	99.7	0.53	0.46	100
Consensus Model[a]	–	0.92	0.22	100	0.87	0.27	100	0.70	0.34	100

[a] consensus model: average of the 9 models (kNN-Dragon, kNN-MolconnZ, SVM-Dragon, SVM-MolconnZ, ISIDA-kNN, ISIDA-MLR, OLS, PLS and ASNN) using their individual applicability domains

below 0.5. For all other models the R^2_{abs} ranged between 0.55 and 0.83. On average, the use of applicability domains improved the performance of individual models although the improvement came at the expense of the lower chemistry space coverage (cf. Table 2).

For the most part all models succeeded in achieving reasonable accuracy of external prediction especially when using the AD. It then appeared natural to bring all models together to explore the power of *consensus prediction*. Thus, the *consensus model* was constructed by averaging all available predicted values taking into account the applicability domain of each individual model. In this case we could use only nine of 15 models that had the AD defined. Since each model had its unique way of defining the AD, each external compound could be found within the AD of anywhere between one and nine models so for averaging we only used models covering the compound. The advantage of this data treatment is that the overall coverage of the prediction is still high because it was rare to have an external compound outside of the ADs of all available models. The results (Table 2) showed that the prediction accuracy for both the modeling set (MAE = 0.22) and the validation sets I and II (0.27 and 0.34, respectively) was the best compared to any individual model. The same observation could be made for the correlation coefficient R^2_{abs}. The coverage of this consensus model II was 100% for all three data sets. This observation suggests that consensus models afford both high space coverage and high accuracy of prediction.

In summary, this study presents an example of a fruitful international collaboration between researchers that use different techniques and approaches but share general principles of QSAR model development and validation. Significantly, we did not make any assumptions about the purported mechanisms of aquatic toxicity yet were able to develop statistically significant models for all experimentally tested compounds. In this regard it is relevant to cite an opinion expressed in an earlier publication by Dr. T. Schultz that "models that accurately predict acute toxicity without first identifying toxic mechanisms are highly desirable" [132]. However, the most significant single result of our studies is the demonstrated superior performance of the *consensus modeling* approach when all models are used concurrently and predictions from individual models are averaged. We have shown that both the predictive accuracy and coverage of the final consensus QSAR models were superior as compared to these parameters for individual models. The consensus models appeared robust in terms of being insensitive to both incorporating individual models with low prediction accuracy and the inclusion or exclusion of the AD. Another important result of this study is the power of addressing complex problems in QSAR modeling by forming a virtual collaboratory of independent research groups leading to the formulation and

empirical testing of *best modeling practices*. This latter endeavor is especially critical in light of the growing interest of regulatory agencies to developing most reliable and predictive models for environmental risk assessment [133] and placing such models in the public domain.

Conclusions. Rapid Growth of Publicly Available Databases and Emerging QSAR Research Strategies

With more than 40 years of history behind, QSAR modeling is a well established research field that (as perhaps with any scientific area) has had its ups and downs. There were several recent publications that criticized the current state of the field. Thus, recent editorial published by the leading cheminformatics Journal of Chemical Information and Modeling (JCIM; also reproduced by the Journal of Medicinal Chemistry) introduced severe limitations on the level and quality of QSAR papers to be considered acceptable [73]. Another recent editorial opinion by Dr. Gerry Maggiora [134] outlined limitations and some reasons for failures of QSAR modeling that relate to the so called "activity cliffs". In another recent important paper, Dr. Terry Stouch addressed the question as to why in silico ADME/Tox models fail [135]. These examples naturally lead to an important and perhaps critical question: whether there is any room for further advancement of the field via innovative methodologies and important applications.

Our previous and ongoing research in the area of QSAR suggests that the answer is a resounding 'yes'. We believe strongly that many examples of low impact QSAR research are due to frequent exploration of datasets of limited size with little attention paid to model *external* validation. This limitation leads to models having questionable "mechanistic" explanatory power but perhaps little if any forecasting ability outside of the training sets used for model development. We believe that the latter ability along with the capabilities of QSAR models to explore chemically diverse datasets with complex biological properties should become the chief focus of QSAR studies. This focus requires the re-evaluation of the success criteria for the modeling as well as the development of novel chemical data mining algorithms and model validation approaches. In fact, we think that the most interesting era in QSAR modeling is just beginning with the rapid growth of the experimental SAR data space.

In the past fifteen years, innovative technologies that enable rapid synthesis and high throughput screening of large libraries of compounds have been adopted in almost all major pharmaceutical and biotech companies. As a result, there has been a huge increase in the number of compounds available on a routine basis to quickly screen for novel drug candidates against new targets or pathways. In contrast, such technologies have rarely become available to the academic research community, thus limiting its ability to conduct large scale chemical genetics or chemical genomics research. The NIH Molecular Libraries Roadmap Initiative has changed this situation by forming the national Molecular Library Screening Centers Network (MLSCN) [136] with the results of screening assays made publicly available via PubChem [4]. These efforts have already led to the unprecedented growth of *available* databases of biologically tested compounds (cf. our recent review where we list about 20 available databases of compounds with known bioactivity [10]). This growth creates new challenges for QSAR modeling such as developing novel approaches for the analysis and visualization of large databases of screening data, novel biologically relevant chemical diversity or similarity measures, and novel tools for virtual screening of compound libraries to ensure high expected hit rates. Due to the significant recent increase in publicly available datasets of biologically active compounds and the critical need to improve the hit rate of experimental compound screening there is a strong need in developing widely accessible and reliable computational QSAR modeling techniques and specific endpoint predictors.

Acknowledgments

The studies described in this review were supported in parts by the National Institutes of Health's Cheminformatics Center planning grant P20-RR20751 and the research grants R01GM066940 and R21GM076059.

Bibliography

1. Hansch C, Fujita T (1964) $\rho - \sigma - \pi$ Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. J Amer Chem Soc 86:1616–1626
2. Hansch C, Leo A, Hoekman D (1995) Exploring QSAR: Hydrophobic, Electronic, and Steric Constants. American Chemical Society, Washington, DC
3. Verloop A, Hoogenstraaten W, Tipker J (1976) In: Ariens EJ (ed) Drug Design. Academic Press, New York, pp 165
4. PubChem (2008) http://pubchem.ncbi.nlm.nih.gov/
5. Roth BL, Kroeze WK (2006) Screening the receptorome yields validated molecular targets for drug discovery. Curr Pharm Des 12:1785–1795
6. NCI (2008) http://dtp.nci.nih.gov/docs/3d_database/structural_information/smiles_strings.html
7. FDA (2008) http://www.fda.gov/cder/Offices/OPS_IO/
8. NTP (2008) http://ntp.niehs.nih.gov/ntpweb/
9. DSSTox (2008) http://www.epa.gov/nheerl/dsstox/About.html

10. Oprea T, Tropsha A (2006) Target, Chemical and Bioactivity Databases – Integration is Key. Drug Discov Today 3:357–365
11. Tropsha A (2005) Application of Predictive QSAR Models to Database Mining. In: Oprea T (ed) Cheminformatics in Drug Discovery. Wiley, Darmstadt, pp 437–455
12. Tropsha A (2003) Recent Trends in Quantitative Structure-Activity Relationships. In: Abraham D (ed) Burger's Medicinal Chemistry and Drug Discovery. Wiley, New York, pp 49–77
13. Tropsha A (2006) Predictive QSAR (Quantitative Structure Activity Relationships) Modeling. In: Martin YC (ed) Comprehensive Medicinal Chemistry II. Elsevier, Oxford, pp 113–126
14. Papa E, Villa F, Gramatica P (2005) Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in Pimephales promelas (fathead minnow). J Chem Inf Model 45:1256–1266
15. Tetko IV (2002) Neural network studies 4. Introduction to associative neural networks. J Chem Inf Comput Sci 42:717–728
16. Zupan J, Novic M, Gasteiger J (1995) Neural Networks with Counter-Propagation Learning-Strategy Used for Modeling. Chemom Intell Lab Syst 27:175–187
17. Devillers J (1996) Strengths and Weaknesses of the back propagation neural network in QSAR and QSPR studies. In: Devillers J (ed) Genetic Algorithms in Molecular Modeling. Academic Press, London, pp 1–24
18. Engels MFM, Wouters L, Verbeeck R, Vanhoof G (2002) Outlier mining in high throughput screening experiments. J Biomol Screen 7:341–351
19. Schuurmann G, Aptula AO, Kuhne R, Ebert RU (2003) Stepwise discrimination between four modes of toxic action of phenols in the Tetrahymena pyriformis assay. Chem Res Toxicol 16:974–987
20. Xue Y, Li H, Ung CY, Yap CW, Chen YZ (2006) Classification of a diverse set of Tetrahymena pyriformis toxicity chemical compounds from molecular descriptors by statistical learning methods. Chem Res Toxicol 19:1030–1039
21. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and Regression Trees. Wadsworth, Co, Belmont
22. Deconinck E, Hancock T, Coomans D, Massart DL, Vander Heyden Y (2005) Classification of drugs in absorption classes using the classification and regression trees (CART) methodology. J Pharm Biomed Anal 39:91–103
23. MOE (2008) http://www.chemcomp.com/fdept/prodinfo.htm#Cheminformatics
24. Put R, Perrin C, Questier F, Coomans D, Massart DL, Vander Heyden YV (2003) Classification and regression tree analysis for molecular descriptor selection and retention prediction in chromatographic quantitative structure-retention relationship studies. J Chromatogr A 988:261–276
25. Breiman L (2001) Random forests. Mach Learn 45:5–32
26. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003) Random forest: A classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 43:s1947–1958
27. Put R, Xu QS, Massart DL, Heyden YV (2004) Multivariate adaptive regression splines (MARS) in chromatographic quantitative structure-retention relationship studies. J Chromatogr A 1055:11–19
28. Friedman JH (1991) Multivariate Adaptive Regression Splines. Ann Stat 19:1–67
29. Vapnik VN (1995) The Nature of Statistical Learning Theory. Springer, New York
30. Aires-de-Sousa J, Gasteiger J (2005) Prediction of enantiomeric excess in a combinatorial library of catalytic enantioselective reactions. J Comb Chem 7:298–301
31. Oloff S, Mailman RB, Tropsha A (2005) Application of validated QSAR models of D1 dopaminergic antagonists for database mining. J Med Chem 48:7322–7332
32. Chohan KK, Paine SW, Waters NJ (2006) Quantitative structure activity relationships in drug metabolism. Curr Top Med Chem 6:1569–1578
33. Manallack DT, Ellis DD, Livingstone DJ (1994) Analysis of linear and nonlinear QSAR data using neural networks. J Med Chem 37:3758–3767
34. Mosier PD, Jurs PC (2002) QSAR/QSPR studies using probabilistic neural networks and generalized regression neural networks. J Chem Inf Comput Sci 42:1460–1470
35. Zheng W, Tropsha A (2000) Novel variable selection quantitative structure–property relationship approach based on the k-nearest-neighbor principle. J Chem Inf Comput Sci 40: 185–194
36. Shen M, Beguin C, Golbraikh A, Stables JP, Kohn H, Tropsha A (2004) Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds. J Med Chem 47:2356–2364
37. Cho SJ, Zheng W, Tropsha A (1998) Rational combinatorial library design 2. Rational design of targeted combinatorial peptide libraries using chemical similarity probe and the inverse QSAR approaches. J Chem Inf Comput Sci 38:259–268
38. Oloff S, Zhang S, Sukumar N, Breneman C, Tropsha A (2006) Chemometric analysis of ligand receptor complementarity: identifying Complementary Ligands Based on Receptor Information (CoLiBRI). J Chem Inf Model 46:844–851
39. Kovatcheva A, Golbraikh A, Oloff S, Xiao YD, Zheng W, Wolschann P, Buchbauer G, Tropsha A (2004) Combinatorial QSAR of ambergris fragrance compounds. J Chem Inf Comput Sci 44:582–595
40. Yao XJ, Panaye A, Doucet JP, Zhang RS, Chen HF, Liu MC, Hu ZD, Fan BT (2004) Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression. J Chem Inf Comput Sci 44:1257–1266
41. Zhang S, Golbraikh A, Oloff S, Kohn H, Tropsha AA (2006) Novel Automated Lazy Learning QSAR (ALL-QSAR) Approach: Method Development, Applications, and Virtual Screening of Chemical Databases Using Validated ALL-QSAR Models. J Chem Inf Model 46:1984–1995
42. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley, Weinheim
43. MolconnZ (2008) http://www.edusoft-lc.com/molconn/
44. Cramer RD III, Patterson DE, Bunce JD (1988) Comparative Molecular Field Analysis (CoMFA) 1. Effect of Shape on Binding of Steroids to Carrier Proteins. J Am Chem Soc 110: 5959–5967
45. Klebe G, Abraham U, Mietzner T (1994) Molecular Similarity Indexes in A Comparative-Analysis (Comsia) of Drug Molecules to Correlate and Predict Their Biological-Activity. J Med Chem 37:4130–4146
46. Kubinyi H, Hamprecht FA, Mietzner T (1998) Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices. J Med Chem 41:2553–2564
47. Waller CL (2004) A comparative QSAR study using CoMFA, HQSAR, and FRED/SKEYS paradigms for estrogen receptor

binding affinities of structurally diverse compounds. J Chem Inf Comput Sci 44:758–765
48. Pastor M, Cruciani G, Mclay I, Pickett S, Clementi S (2000) GRid-INdependent descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors. J Med Chem 43:3233–3243
49. Cruciani C, Crivori P, Carrupt PA, Testa B (2000) Molecular fields in quantitative structure-permeation relationships: the VolSurf approach. J Mol Struct-Theochem 503:17–30
50. Kier LB, Hall LH (1986) Molecular Connectivity in Structure-Activity Analysis. Wiley, New York
51. Kier LB (1987) Inclusion of symmetry as a shape attribute in kappa-index analysis. Quant Struct-Activit Relatsh 6:8–12
52. Kier LB, Hall LH (1999) Molecular Structure Description: The Electrotopological State. Academic Press, New York
53. Brown RD, Martin YC (1998) An evaluation of structural descriptors and clustering methods for use in diversity selection. SAR QSAR Environ Res 8:23–39
54. Hoffman B, Cho SJ, Zheng W, Wyrick S, Nichols DE, Mailman RB, Tropsha A (1999) Quantitative structure-activity relationship modeling of dopamine D(1) antagonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and K nearest neighbor methods. J Med Chem 42:3217–3226
55. Zhang Q, Muegge I (2006) Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring. J Med Chem 49:1536–1548
56. Cherkasov A (2005) 'Inductive' Descriptors. 10 Successful Years in QSAR. Curr Comp Aid Drug Des 1:21–42
57. Stinson SC (2001) Chiral pharmaceuticals. Chem Eng News 79:79
58. Cho SJ, Tropsha A (1995) Cross-validated R2-guided region selection for comparative molecular field analysis: a simple method to achieve consistent results. J Med Chem 38:1060–1066
59. Carhart RE, Smith DH, Venkataraghavan R (1985) Atom pairs as molecular features in structure-activity studies: definition and applications. J Chem Inf Comput Sci 25:64–73
60. Schultz HP, Schultz EB, Schultz TP (1995) Topological Organic-Chemistry 9. Graph-Theory and Molecular Topological Indexes of Stereoisomeric Organic-Compounds. J Chem Inf Comput Sci 35:864–870
61. Julian-Ortiz JV, Alapont CD, Rios-Santamarina I, Garcia-Domenech R, Galvez J (1998) Prediction of properties of chiral compounds by molecular topology. J Mol Graphics Model 16:14–18
62. Golbraikh A, Bonchev D, Tropsha A (2001) Novel chirality descriptors derived from molecular topology. J Chem Inf Comput Sci 41:147–158
63. Golbraikh A, Tropsha A (2003) QSAR modeling using chirality descriptors derived from molecular topology. J Chem Inf Comput Sci 43:144–154
64. Golbraikh A, Bonchev D, Tropsha A (2002) Novel ZE-isomerism descriptors derived from molecular topology and their application to QSAR analysis. J Chem Inf Comput Sci 42:769–787
65. Kovatcheva A, Golbraikh A, Oloff S, Feng J, Zheng W, Tropsha A (2005) QSAR modeling of datasets with enantioselective compounds using chirality sensitive molecular descriptors. SAR QSAR Environ Res 16:93–102
66. Gunner OF, Hughes DW, Dumont LM (1991) An integrated approach to three-dimensional information management with MACCS-3D. J Chem Inf Comput Sci 31:408–414
67. Barnard JM, Downs GM (1995) Fingerprint Descriptor Package, 3.1. Barnard Chemical Information Ltd, Scheffield
68. James CA, Weininger D (1995) Daylight Theory Manual. Daylight Chemical Information Systems, Aliso Viejo
69. Durant JL, Leland BA, Henry DR, Nourse JG (2002) Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci 42:1273–1280
70. Deshpande M, Kuramochi M, Karypis J (2002) Frequent Sub-Structure-Based Approaches for Classifying Chemical Compounds. Proc of the 8th International Conference on Knowledge Discovery and Data Mining, Edmonton
71. Golbraikh A, Tropsha A (2002) Beware of q2! J Mol Graph Model 20:269–276
72. Tropsha A, Gramatica P, Gombar VK (2003) The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. Quant Struct Act Relat Comb Sci 22:69–77
73. Jorgensen WL, Tirado-Rives J (2006) QSAR/QSPR and Proprietary Data. J Chem Inf Model 46:937
74. Novellino E, Fattorusso C, Greco G (1995) Use of Comparative Molecular Field Analysis and Cluster Analysis in Series Design. Pharm Acta Helv 70:149–154
75. Norinder U (1996) Single and Domain Made Variable Selection in 3D QSAR applications. J Chemomet 10:95–105
76. Tropsha A, Cho SJ (1998) Cross-Validated R2-Guided Region Selection for CoMFA Studies. In: Kubinyi H, Folkers G, Martin YC (eds) 3D QSAR in Drug Design, 3rd edn. Kluwer, Dordrecht, pp 57–69
77. Golbraikh A, Tropsha A (2002) Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. J Comput Aided Mol Des 16:357–369
78. Golbraikh A, Shen M, Xiao Z, Xiao YD, Lee KH, Tropsha A (2003) Rational selection of training and test sets for the development of validated QSAR models. J Comput Aided Mol Des 17:241–253
79. Deconinck E, Coomans D, Vander HY (2007) Exploration of linear modelling techniques and their combination with multivariate adaptive regression splines to predict gastro-intestinal absorption of drugs. J Pharm Biomed Anal 43:119–130
80. Verma RP, Hansch C (2006) Cytotoxicity of organic compounds against ovarian cancer cells: a quantitative structure-activity relationship study. Mol Pharm 3:441–450
81. Pavan M, Netzeva TI, Worth AP (2006) Validation of a QSAR model for acute toxicity. SAR QSAR Environ Res 17:147–171
82. Vracko M, Bandelj V, Barbieri P, Benfenati E, Chaudhry Q, Cronin M, Devillers J, Gallegos A, Gini G, Gramatica P, Helma C, Mazzatorta P, Neagu D, Netzeva T, Pavan M, Patlewicz G, Randic M, Tsakovska I, Worth A (2006) Validation of counter propagation neural network models for predictive toxicology according to the OECD principles: a case study. SAR QSAR Environ Res 17:265–284
83. Saliner AG, Netzeva TI, Worth AP (2006) Prediction of estrogenicity: validation of a classification model. SAR QSAR Environ Res 17:195–223
84. Roberts DW, Aptula AO, Patlewicz G (2006) Mechanistic applicability domains for non-animal based prediction of toxicological endpoints. QSAR analysis of the schiff base appli-

cability domain for skin sensitization. Chem Res Toxicol 19: 1228–1233
85. Estrada E, Patlewicz G (2004) On the usefulness of graph-theoretic descriptors in predicting theoretical parameters. Phototoxicity of polycyclic aromatic hydrocarbons (PAHs). Croat Chem Acta 77:203–211
86. Moss GP, Cronin MTD (2002) Quantitative structure-permeability relationships for percutaneous absorption: re-analysis of steroid data. Int J Pharm 238:105–109
87. Leo AJ, Hansch C (1999) Role of hydrophobic effects in mechanistic QSAR. Perspect Drug Discov Des 17:1–25
88. Zhang S, Golbraikh A, Tropsha A (2006) Development of quantitative structure-binding affinity relationship models based on novel geometrical chemical descriptors of the protein-ligand interfaces. J Med Chem 49:2713–2724
89. Kovatcheva A, Buchbauer G, Golbraikh A, Wolschann P (2003) QSAR modeling of alpha-campholenic derivatives with sandalwood odor. J Chem Inf Comput Sci 43:259–266
90. Shen M, Xiao Y, Golbraikh A, Gombar VK, Tropsha A (2003) Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates. J Med Chem 46:3013–3020
91. Shen M, LeTiran A, Xiao Y, Golbraikh A, Kohn H, Tropsha A (2002) Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods. J Med Chem 45:2811–2823
92. Golbraikh A, Shen M, Xiao Z, Xiao YD, Lee KH, Tropsha A (2003) Rational selection of training and test sets for the development of validated QSAR models. J Comput Aided Mol Des 17:241–253
93. Mandel J (1982) Use of the Singular Value Decomposition in Regression-Analysis. Am Stat 36:15–24
94. Afantitis A, Melagraki G, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O (2006) A novel QSAR model for predicting induction of apoptosis by 4-aryl-4H-chromenes. Bioorg Med Chem 14:6686–6694
95. Netzeva TI, Gallegos SA, Worth AP (2006) Comparison of the applicability domain of a quantitative structure-activity relationship for estrogenicity with a large chemical inventory. Environ Toxicol Chem 25:1223–1230
96. Tong W, Xie Q, Hong H, Shi L, Fang H, Perkins R (2004) Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. Environ Health Perspect 112:1249–1254
97. Helma C (2006) Lazy structure-activity relationships (lazar) for the prediction of rodent carcinogenicity and Salmonella mutagenicity. Mol Divers 10:147–158
98. de Cerqueira LP, Golbraikh A, Oloff S, Xiao Y, Tropsha A (2006) Combinatorial QSAR Modeling of P-Glycoprotein Substrates. J Chem Inf Model 46:1245–1254
99. Sachs L (1984) Handbook of statistics. Springer
100. Irwin JJ, Shoichet BK (2005) ZINC – a free database of commercially available compounds for virtual screening. J Chem Inf Model 45:177–182
101. Medina-Franco JL, Golbraikh A, Oloff S, Castillo R, Tropsha A (2005) Quantitative structure-activity relationship analysis of pyridinone HIV-1 reverse transcriptase inhibitors using the k nearest neighbor method and QSAR-based database mining. J Comput Aided Mol Des 19:229–242
102. Zhang S, Wei L, Bastow K, Zheng W, Brossi A, Lee KH, Tropsha A (2007) Antitumor Agents 252. Application of validated QSAR models to database mining: discovery of novel tylophorine derivatives as potential anticancer agents. J Comput Aided Mol Des 21:97–112
103. Hsieh JH, Wang XS, Teotico D, Golbraikh A, Tropsha A (2008) Differentiation of AmpC beta-lactamase binders vs. decoys using classification kNN QSAR modeling and application of the QSAR classifier to virtual screening. J Comput Aided Mol Des
104. Tropsha A, Cho SJ, Zheng W (1999) New Tricks for an Old Dog: Development and Application of Novel QSAR Methods for Rational Design of Combinatorial Chemical Libraries and Database Mining. In: Parrill AL, Reddy MR (eds) Rational Drug Design: Novel Methodology and Practical Applications. American Chemical Society, Washington DC, pp 198–211
105. Gussio R, Pattabiraman N, Kellogg GE, Zaharevitz DW (1998) Use of 3D QSAR methodology for data mining the National Cancer Institute Repository of Small Molecules: application to HIV-1 reverse transcriptase inhibition. Methods 14: 255–263
106. Tropsha A, Zheng W (2001) Identification of the descriptor pharmacophores using variable selection QSAR: applications to database mining. Curr Pharm Des 7:599–612
107. Hengstler JG, Foth H, Kahl R, Kramer PJ, Lilienblum W, Schulz T, Schweinfurth H (2006) The REACH concept and its impact on toxicological sciences. Toxicology 220:232–239
108. Richard AM (2006) Future of toxicology–predictive toxicology: An expanded view of: chemical toxicity. Chem Res Toxicol 19:1257–1262
109. Klopman G, Zhu H, Fuller MA, Saiakhov RD (2004) Searching for an enhanced predictive tool for mutagenicity. SAR QSAR Environ Res 15:251–263
110. Richard AM, Benigni R (2002) AI and SAR approaches for predicting chemical carcinogenicity: Survey and status report. SAR QSAR Environ Res 13:1–19
111. Yang C, Benz RD, Cheeseman MA (2006) Landscape of current toxicity databases and database standards. Curr Opin Drug Discov Dev 9:124–133
112. Zhu H, Tropsha A, Fourches D, Varnek A, Papa E, Gramatica P, Oberg T, Dao P, Cherkasov A, Tetko IV (2008) Combinatorial QSAR Modeling of Chemical Toxicants Tested against Tetrahymena pyriformis. J Chem Inf Model 48:766–784
113. Schultz TW, Netzeva TI (2004) Development and evaluation of QSARs for ecotoxic endpoints: The benzene response-surface model for Tetrahymena toxicity. In: Cronin MTD, Livingstone DJ (eds) Modeling Environmental Fate and Toxicity. CRC Press, Boca Raton, pp 265–284
114. Schultz TW (1999) Structure-toxicity relationships for benzenes evaluated with Tetrahymena pyriformis. Chem Res Toxicol 12:1262–1267
115. Netzeva TI, Schultz TW (2005) QSARs for the aquatic toxicity of aromatic aldehydes from Tetrahymena data. Chemosphere 61:1632–1643
116. Schultz TW, Sinks GD, Miller LA (2001) Population growth impairment of sulfur-containing compounds to Tetrahymena pyriformis. Environ Toxicol 16:543–549
117. Schultz TW, Yarbrough JW, Woldemeskel M (2005) Toxicity to Tetrahymena and abiotic thiol reactivity of aromatic isothiocyanates. Cell Biol Toxicol 21:181–189

118. Aptula AO, Roberts DW, Cronin MTD, Schultz TW (2005) Chemistry-toxicity relationships for the effects of Di-and trihydroxybenzenes to Tetrahymena pyriformis. Chem Res in Toxicol 18:844–854
119. Schultz TW, Hewitt M, Netzeva TI, Cronin MTD (2007) Assessing applicability domains of toxicological QSARs: Definition, confidence in predicted values, and the role of mechanisms of action. QSAR Comb Sci 26:238–254
120. Netzeva TI, Schultz TW (2005) QSARs for the aquatic toxicity of aromatic aldehydes from Tetrahymena data. Chemosphere 61:1632–1643
121. Schultz TW, Yarbrough JW, Woldemeskel M (2005) Toxicity to Tetrahymena and abiotic thiol reactivity of aromatic isothiocyanates. Cell Biol Toxicol 21:181–189
122. Schultz TW, Sinks GD, Miller LA (2001) Population growth impairment of sulfur-containing compounds to Tetrahymena pyriformis. Environ Toxicol 16:543–549
123. Aptula AO, Roberts DW, Cronin MTD, Schultz TW (2005) Chemistry-toxicity relationships for the effects of Di-and trihydroxybenzenes to Tetrahymena pyriformis. Chem Res Toxicol 18:844–854
124. Schultz TW, Hewitt M, Netzeva TI, Cronin MTD (2007) Assessing applicability domains of toxicological QSARs: Definition, confidence in predicted values, and the role of mechanisms of action. QSAR Comb Sci 26:238–254
125. Schultz TW, Hewitt M, Netzeva TI, Cronin MTD (2007) Assessing applicability domains of toxicological QSARs: Definition, confidence in predicted values, and the role of mechanisms of action. QSAR Comb Sci 26:238–254
126. Schultz TW, Hewitt M, Netzeva TI, Cronin MTD (2007) Assessing applicability domains of toxicological QSARs: Definition, confidence in predicted values, and the role of mechanisms of action. QSAR Comb Sci 26:238–254
127. Aptula AO, Roberts DW, Cronin MTD, Schultz TW (2005) Chemistry-toxicity relationships for the effects of Di-and trihydroxybenzenes to Tetrahymena pyriformis. Chem Res Toxicol 18:844–854
128. Netzeva TI, Schultz TW (2005) QSARs for the aquatic toxicity of aromatic aldehydes from Tetrahymena data. Chemosphere 61:1632–1643
129. Schultz TW, Yarbrough JW, Woldemeskel M (2005) Toxicity to Tetrahymena and abiotic thiol reactivity of aromatic isothiocyanates. Cell Biol Toxicol 21:181–189
130. Schultz TW, Sinks GD, Miller LA (2001) Population growth impairment of sulfur-containing compounds to Tetrahymena pyriformis. Environ Toxicol 16:543–549
131. Gramatica P (2007) Principles of QSAR models validation: internal and external. QSAR Comb Sci 26:694–701
132. Schultz TW (1999) Structure-toxicity relationships for benzenes evaluated with Tetrahymena pyriformis. Chem Res Toxicol 12:1262–1267
133. Yang C, Richard AM, Cross KP (2006) The Art of Data Mining the Minefields of Toxicity Databases to Link Chemistry to Biology. Curr Comput-Aided Drug Des 2:135–150
134. Maggiora GM (2006) On outliers and activity cliffs–why QSAR often disappoints. J Chem Inf Model 46:1535
135. Stouch TR, Kenyon JR, Johnson SR, Chen XQ, Doweyko A, Li Y (2003) In silico ADME/Tox: why models fail. J Comput Aided Mol Des 17:83–92
136. Austin CP, Brady LS, Insel TR, Collins FS (2004) NIH Molecular Libraries Initiative. Science 306:1138–1139

Quantum Algorithms

Michele Mosca[1,2,3]
[1] Institute for Quantum Computing and Dept. of Combinatorics and Optimization, University of Waterloo, Waterloo, Canada
[2] St. Jerome's University, Waterloo, Canada
[3] Perimeter Institute for Theoretical Physics, Waterloo, Canada

Article Outline

Glossary
Definition of the Subject
Introduction and Overview
Early Quantum Algorithms
Factoring, Discrete Logarithms and the Abelian Hidden Subgroup Problem
Algorithms Based on Amplitude Amplification
Simulation of Quantum Mechanical Systems
Generalizations of the Abelian Hidden Subgroup Problem
Quantum Walk Algorithms
Adiabatic Algorithms
Topological Algorithms
Quantum Algorithms for Quantum Tasks
Future Directions
Bibliography

Glossary

Quantum circuit model One of the standard and most commonly used models of quantum computation which generalizes the classical model of acyclic circuits and closely models most of the proposed physical implementations of quantum computers. When studying algorithms for a problem with an infinite number of possible inputs, one usually restricts attention to *uniform* families of circuits, which are families of circuits in which the circuit for inputs of size n can be generated efficiently as a function of n. For example, one might require that there is a classical Turing machine that can generate the nth circuit in time polynomial in n.

Black box model A model of computation where the input to the problem includes a "black-box" that can be applied (equivalently, an "oracle" that can be "queried"). This is the only way to extract information from the black-box. For example, the black-box could accept inputs $j \in \{0,1\}^n$ and output a value $X_j \in \{0,1\}$. In this particular case, we can think of the black-box as a means for querying the bits of the

string $\mathbf{X} = X_1 X_2 X_3 \ldots X_{2^n}$. In the black-box model, one usually measures complexity in terms of the number of applications of the black-box.

Computational complexity When referring to an algorithm, the computational complexity (often just called the complexity) is a measure of the resources used by the algorithm (which we can also refer to as the *cost* of the algorithm) usually measured as a function of the size of the input to the algorithm. The complexity for input size n is taken to be the cost of the algorithm on a worst-case input of size n to the problem. This is also referred to as *worst-case complexity*. When referring to a problem, the computational complexity is the minimum amount of resources required by any algorithm to solve the problem. See ► Quantum Computational Complexity for an overview.

Query complexity When referring to a black-box algorithm, the query complexity is the number of applications of the black-box or oracle used by the algorithm. When referring to a black-box problem, the query complexity is the minimum number of applications of the black-box required by any algorithm to solve the problem.

Definition of the Subject

The strong Church–Turing thesis states that a probabilistic Turing machine can efficiently simulate any realistic model of computation. By "efficiently", we mean that there is a polynomial p such that the amount of resources used by the Turing machine simulation is not more than $p(M)$ where M is the amount of resources used by the given realistic model of computation.

Since a computer is a physical device, any reasonable model of computation must be cast in a realistic physical framework, hence the condition that the model be "realistic" is very natural. The probabilistic Turing machine is implicitly cast in a classical framework for physics, and it appears to hold as long as the competing model of computation is also cast in a classical framework. However, roughly a century ago, a new framework for physics was developed: quantum mechanics. The impact of this new framework on the theory of computation was not taken very seriously until the early 1970's by Stephen Wiesner [177] for cryptographic purposes (and later by Bennett and Brassard [27]). Benioff [23] proposed using quantum devices in order to implement reversible computation. Feynman [79] noted that a classical computer seems incapable of efficiently simulating the dynamics of rather simple quantum mechanical systems, and proposed that a "quantum" computer, with components evolving according to quantum mechanical rules, should be able to perform such simulations efficiently (see Sect. "Simulation of Quantum Mechanical Systems"). Manin made a similar observation independently [136]. Deutsch [64] worked on proving the original Church–Turing thesis (which was only concerned about effective computability, and not efficient computability) in a quantum mechanical framework, and defined two models of quantum computation; he also gave the first quantum algorithm. One of Deutsch's ideas is that quantum computers could take advantage of the computational power present in many "parallel universes" and thus outperform conventional classical algorithms. While thinking of parallel universes is sometimes a convenient way for researchers to invent quantum algorithms, the algorithms and their successful implementation are independent of any particular interpretation of standard quantum mechanics.

Quantum algorithms are algorithms that run on any realistic model of quantum computation. The most commonly used model of quantum computation is the circuit model (more strictly, the model of uniform families of acyclic quantum circuits), and the *quantum strong* Church–Turing thesis states that the quantum circuit model can efficiently simulate any realistic model of computation. Several other models of quantum computation have been developed, and indeed they can be efficiently simulated by quantum circuits. Quantum circuits closely resemble most of the currently pursued approaches for attempting to construct scalable quantum computers.

The study of quantum algorithms is very important for several reasons. Computationally secure cryptography is widely used in society today, and relies on the believed difficulty of a small number of computational problems. Quantum computation appears to redefine what is a tractable or intractable problem, and one of the first breakthroughs in the development of quantum algorithms was Shor's discovery of efficient algorithms [165] for factoring and finding discrete logarithms. The difficulty of factoring and finding discrete logarithms was (and still is!) at the core of currently-used public-key cryptography, and his results showed that if and when a quantum computer is built, then any messages that had been previously encrypted using our current public-key cryptographic tools could be compromised by anyone who had recorded the ciphertext and public keys. Furthermore the vast public-key infrastructure currently in place would be compromised with no clear alternative to replace it; also, any alternative will take many years to deploy. There was a sudden rush to answer two fundamental questions. Firstly, can we actually build a sufficiently large quantum computer? Perhaps this isn't a reasonable model of computation. Subse-

quent work on quantum fault-tolerant error correction indicates that the answer is "yes", and experimental progress has steadily grown. Secondly, what other interesting and important problems can quantum computers solve more efficiently than the best known classical algorithms? The following sections survey the state of the art on the second question.

As we gain confidence about which problems are still hard in a quantum mechanical framework, we can start to rebuild confidence in a secure public-key cryptographic infrastructure that is robust in the presence of quantum technologies. Although the cryptographic implications are dramatic and of immediate relevance, in the longer term, the most important significance of quantum algorithms will be for a wider range of applications, where important problems cannot be solved because there are no known (or possible) efficient classical algorithms, but there are efficient quantum mechanical solutions. At present, we know of applications such as searching and optimizing (Sect. "Algorithms Based on Amplitude Amplification") and simulating physical systems (see Sect. "Simulation of Quantum Mechanical Systems"), but the full implications are still unknown and very hard to predict. The next few sections give an overview of the current state of quantum algorithmics.

Introduction and Overview

There are several natural models of quantum computation. The most common one is a generalization of the classical circuit model. A detailed description of the circuit model of quantum computation can be found in several textbooks [122,129,149]. One can also define continuous models of computation, where one specifies the Hamiltonian $H(t)$ of the system at time t, where $H(t)$ is a "reasonable" Hamiltonian (e. g. a sum Hamiltonians involving a constant number of nearby subsystems), and one evolves the system for a period of time T. A reasonable measure of the total cost might be $\int_{t=0}^{T} \|H(t)\| \, \mathrm{d}t$.

Most algorithmic work in quantum computing has been developed in discrete models of computation, that is, with a discrete state space and with discrete time steps. In Subsect. "Continuous Time Quantum Walk Algorithms" and Sect. "Adiabatic Algorithms", we discuss algorithms developed in a continuous-time model of computation. Even if, as in the case of classical computers, implementations of scalable fault-tolerant quantum computers have discrete time steps and state spaces, these algorithms are still very useful since there are efficient simulations using any universal discrete model of quantum computation. Note that if no such efficient simulation exists, then either the continuous model of computation in question is physically unrealistic, or the quantum strong Church–Turing thesis is incorrect.

Discrete state spaces can be used to approximate continuous state spaces in order to solve problems normally posed with continuous state spaces. One must choose appropriate discretizations, analyzing errors in the approximation, and quantify the scaling of the algorithm as the overall approximation error gets arbitrarily small. Quantum algorithms for such continuous problems are surveyed in ► Quantum Algorithms and Complexity for Continuous Problems.

Many of the key ideas that led to the development of quantum computation emerged from earlier work on reversible computation [24]. Many facts from the theory of reversible computing are fundamental tools in the development of quantum algorithms. For example, suppose we have a classical algorithm for computing a function $f\colon \{0,1\}^n \to \{0,1\}^m$ (we use binary encoding for convenience). The details of the computing model are not so important, as long as it's a realistic model. For concreteness, suppose we have a reasonable encoding of a circuit of size C, using gates from a finite set, that takes $\mathbf{x} \in \{0,1\}^n$ as input and outputs $\mathbf{y} \in \{0,1\}^m$ (discarding any additional information it might have computed). Then this circuit can be efficiently converted into a circuit, composed only of reversible gates, that maps $|\mathbf{x}\rangle\,|\mathbf{y}\rangle\,|\mathbf{0}\rangle \mapsto |\mathbf{x}\rangle\,|\mathbf{y} \oplus f(\mathbf{x})\rangle\,|\mathbf{0}\rangle$, where $\oplus$ denotes the bitwise XOR (addition modulo 2), and the third register of 0s is ancilla workspace that is reset to all 0s by reversible operations. This new circuit uses $O(C)$ reversible gates from a finite set, so the overhead is modest. A basic introduction to this and other important facts about reversible computing can be find in most standard textbooks on quantum computing. For example, in Sect. "Algorithms Based on Amplitude Amplification" on quantum searching, we use the fact that we can convert any classical heuristic algorithm that successfully guesses a solution with probability p into a reversible quantum algorithm that guesses a solution with probability amplitude $\sqrt{p}$.

Most of the known quantum algorithms can be phrased as black-box algorithms solving black-box problems. A black-box, or oracle, is subroutine or subcircuit that implements some operation or function. It does so in a way that provides no other information other than simply taking an input and giving the prescribed output. One cannot, for example, look at the inner workings of the circuit or device implementing the black-box to extract additional information. For example, Shor's factoring algorithm can be viewed as an algorithm that finds the order of an element in a black-box group (that is, a group for which

the group operations are computed by a black-box), or the period of a black-box function, where the black-box is substituted with a subcircuit that implements exponentiation modulo N. The quantum search algorithm is described as a black-box algorithm, but it is straightforward to substitute in a subcircuit that checks if a given input is a valid certificate for some problem in NP.

If we take a black-box algorithm that uses T applications of the black-box and A other computational steps, and replace each black-box with a subcircuit that uses B elementary gates, then we get an algorithm that uses $TB + A$ gates. Thus if T and A are both polynomial in size, then an efficient black-box algorithm yields an efficient algorithm whenever we replace the black-box with a polynomial time computable function.

Many lower bounds have been found in the black-box model. The query complexity of a black-box problem is the number of applications of the black-box (or queries to the oracle) that a black-box algorithm must make in order to solve the problem. If we try to solve a problem that has query complexity T, where the black-box is implemented by some subcircuit, then we can conclude that any algorithm that treats the subcircuit as a black-box must apply the subcircuit a total of T times and thus use $\Omega(T)$ gates (we use the fact that any implementation of the black-box uses at least one gate; if we had a lower bound on the complexity of implementing the black-box, then we could derive a better lower bound in this case). However, this does not imply that an $\Omega(T)$ lower bound applies to any algorithm that uses the subcircuit, since it might exploit the information within the subcircuit in a way other than just applying the subcircuit. A discussion of the black-box model and its practical relevance can be found in [52,122,173].

In the literature, one can often see a progression from studying basic algorithmic primitives (such as the convergence properties of a generic quantum walk), to the application to solve a black-box problem (such as element distinctness), to solving some concrete computational problem (like factoring an integer) that doesn't involve a black-box.

This survey will include both black-box and non-black-box results. It is infeasible to detail all the known quantum algorithms, so a representative sample is given in this article. For a subset of this sample, there is an explicit definition of the problem, a discussion of what the best known quantum algorithm can do, and a comparison to what can be achieved with classical algorithms. For black-box problems, we attempt to give the number of queries and the number of non-query operations used by the algorithm, as well as the best-known lower bounds on the query complexity. In some cases, all of this information is not readily available in the literature, so there will be some gaps.

As a small technical note, when we refer to a real number r as an input or output to a problem, we are referring to a finite description of a real number from which, for any integer n, one can efficiently compute (in time polynomial in n) an approximation of r with error at most $1/2^n$.

In this article, we start with a brief sketch of the very early quantum algorithms, and then in the subsequent sections the algorithms are grouped according to the kind of problems they solve, or the techniques or algorithmic paradigms used.

Section "Early Quantum Algorithms" summarizes the early quantum algorithms. Section "Factoring, Discrete Logarithms and the Abelian Hidden Subgroup Problem" describes the Abelian hidden subgroup algorithms, including Shor's factoring and discrete logarithm algorithms. Section "Algorithms Based on Amplitude Amplification" describes quantum searching and amplitude amplification and some of the main applications. Section "Simulation of Quantum Mechanical Systems" describes quantum algorithms for simulating quantum mechanical systems, which are another important class of algorithms that appear to offer an exponential speed-up over classical algorithms. Section "Generalizations of the Abelian Hidden Subgroup Problem" describes several non-trivial generalizations of the Abelian hidden subgroup problem, and related techniques. Section "Quantum Walk Algorithms" describes the quantum walk paradigm for quantum algorithms and summarizes some of the most interesting results and applications. Section "Adiabatic Algorithms" describes the paradigm of adiabatic algorithms. Section "Topological Algorithms" describes a family of "topological" algorithms. Section "Quantum Algorithms for Quantum Tasks" describes algorithms for quantum tasks which cannot be done by a classical computer. In Sect. "Future Directions" we conclude with a discussion.

Early Quantum Algorithms

The first explicitly defined quantum algorithm was the one described by David Deutsch in his landmark paper [64] where he defined the model of quantum computation, including a circuit and Turing machine model.

The problem was to decide if a given function $f\colon \{0,1\} \mapsto \{0,1\}$ is constant or "balanced" (the function f is balanced if there are an equal number of 0 and 1 outputs). In other words, output $f(0) \oplus f(1)$, where $\oplus$ denotes addition modulo 2. One is given a circuit that implements $|x\rangle|0\rangle \mapsto |x\rangle|f(x)\rangle$. Deutsch showed that us-

ing one application of the circuit, we can compute

$$\frac{1}{\sqrt{2}}\,|\,0\rangle\,|\,f(0)\rangle + \frac{1}{\sqrt{2}}\,|\,1\rangle\,|\,f(1)\rangle\ .$$

If $f(0) = f(1)$, then the first qubit is in the state $\frac{1}{\sqrt{2}}\,|\,0\rangle + \frac{1}{\sqrt{2}}\,|\,1\rangle$ and thus applying a Hadamard transformation to the first qubit will output $|\,0\rangle$ with certainty. However, if $f(0) \neq f(1)$, then applying a Hadamard gate and measuring will output $|\,0\rangle$ with probability $\frac{1}{2}$ and $|\,1\rangle$ with probability $\frac{1}{2}$. Thus if we obtain $|\,1\rangle$, we are certain that the function is balanced, and thus we would output "balanced" in this case. If we obtain $|\,0\rangle$, we cannot be certain; however, by guessing "constant" with probability $\frac{2}{3}$ in this case, we get an algorithm that guesses correctly with probability $\frac{2}{3}$. This is not possible with a classical algorithm that only evaluates f once.

If one knows, for example, that the unitary transformation also maps $|\,x\rangle\,|\,1\rangle \mapsto |\,x\rangle\,|\,1 \oplus f(x)\rangle$, then one can solve this problem with certainty with only one query [54]. A common tool that is used in this and many other quantum algorithms, is to note that if one applies such an implementation of f on the input $|\,x\rangle\,(1/\sqrt{2}\,|\,0\rangle - 1/\sqrt{2}\,|\,1\rangle)$, then the output can be written as $(-1)^{f(x)}\,|\,x\rangle\,(1/\sqrt{2}\,|\,0\rangle - 1/\sqrt{2}\,|\,1\rangle)$. This technique can be used to replace the 2-step process of first mapping $|\,x\rangle\,|\,0\rangle \mapsto |\,x\rangle\,|\,f(x)\rangle$, then applying a Z gate $|\,x\rangle\,|\,f(x)\rangle \mapsto |\,x\rangle\,(-1)^{f(x)}\,|\,f(x)\rangle = (-1)^{f(x)}\,|\,x\rangle\,|\,f(x)\rangle$, and then applying the function evaluation a second time in order to "uncompute" the value of $f(x)$ from the second register: $(-1)^{f(x)}\,|\,x\rangle\,|\,f(x)\rangle \mapsto (-1)^{f(x)}\,|\,x\rangle\,|\,0\rangle$.

This problem was generalized to the Deutsch–Jozsa problem [65] which asks the same constant versus "balanced" question, but for a function $f\colon \{0,1\}^n \to \{0,1\}$, with the promise that the function is either constant or balanced. The notion of a promise problem appears frequently in the study of quantum algorithms and complexity. One can either think of the condition as a promise, or alternatively, one can accept any answer as correct in the case that the promise is not satisfied.

This problem was solved with two queries with a similar algorithm, and the two queries can be reduced to one if the oracle evaluating f has the form $|\,x\rangle\,|\,b\rangle \mapsto |\,x\rangle\,|\,b \oplus f(x)\rangle$. A classical deterministic algorithm requires $2^{n-1} + 1$ evaluations of the function in order to decide if f is constant or balanced with certainty. However, a classical randomized algorithm can decide if f is constant or balanced with error probability ϵ with only $O(\log \frac{1}{\epsilon})$ queries. Other oracle separations were given in [30,31].

Bernstein and Vazirani [28] defined a special family of functions that are either constant or balanced, in particular, for any $\mathbf{a} \in \{0,1\}^n$, $b \in \{0,1\}$, let $f_{\mathbf{a}}(\mathbf{x}) = \mathbf{a}\cdot\mathbf{x} \oplus b$. They showed how to find $\mathbf{a}$ using only 2 evaluations of a black-box implementing $|\,\mathbf{x}\rangle\,|\,0\rangle \mapsto |\,\mathbf{x}\rangle\,|\,f(\mathbf{x})\rangle$. This can be reduced to 1 evaluation given the standard black-box $U_f\colon\ |\,\mathbf{x}\rangle\,|\,b\rangle \mapsto |\,\mathbf{x}\rangle\,|\,b \oplus f(\mathbf{x})\rangle$. This algorithm can be generalized quite naturally to finding the hidden matrix M in the affine map $\mathbf{x} \mapsto M\mathbf{x} + \mathbf{b}$, where $\mathbf{x} \in \mathbb{Z}_N^n$, $\mathbf{b} \in \mathbb{Z}_N^m$ and M is an $m \times n$ matrix of values in $\mathbb{Z}_N$ (with element-wise addition modulo N), using only m queries [104,143]. In some sense, the quantum algorithm takes a black-box for right-multiplying by M and turns it into a black-box for left-multiplying by M. This allows us to determine M with only m queries instead of $n+1$ queries, which can be advantageous if $m < n$, although no practical application has been developed to date.

Bernstein and Vazirani also give the first instance of a black-box problem (recursive Fourier sampling) where a quantum algorithm gives a exponential improvement over bounded error randomized classical algorithms. Simon [167] gave a problem where there is an exponential gap. The problem is somewhat simpler to state, and is a special case of a broad family of important problems for which efficient quantum algorithms were subsequently found (using similar methods).

Simon's Problem

Input: A black-box U_f implementing $|\,\mathbf{x}\rangle\,|\,0\rangle \mapsto |\,\mathbf{x}\rangle\,|\,f(\mathbf{x})\rangle$ for a function $f\colon \{0,1\}^n \to \{0,1\}^n$ with the property that $f(\mathbf{x}) = f(\mathbf{y})$ if and only if $\mathbf{x} \oplus \mathbf{y} \in K = \{\mathbf{0}, \mathbf{s}\}$ for some $\mathbf{s} \in \{0,1\}^n$.

Problem: Find $\mathbf{s}$.

Note that f is one-to-one if $\mathbf{s} = \mathbf{0}$, the string of 0s, and f is two-to-one if $\mathbf{s} \neq \mathbf{0}$.

Simon's algorithm is very simple and elegant. We sketch it here, since many of the algorithms presented in the upcoming sections follow the same overall structure. We give a modified version that has a definite running time (versus letting it run indefinitely and analyzing the expected running time).

Simon's Algorithm

1. Set $i = 1$.
2. Prepare the $2n$ qubit state $|\,00\ldots0\rangle\,|\,00\ldots0\rangle$.
3. Apply a Hadamard gate to the first n qubits to obtain the state $\frac{1}{\sqrt{2^n}}\,|\,\mathbf{x}\rangle\,|\,\mathbf{0}\rangle$.
4. Apply U_f to create the state $\frac{1}{\sqrt{2^n}}\,|\,\mathbf{x}\rangle\,|\,f(\mathbf{x})\rangle$.
5. Apply a Hadamard gate to the first n qubits.
6. Measure the first n qubits to obtain a string $\mathbf{y}_i \in \{0,1\}^n$.
7. If $i = n+3$, go to the next step. Otherwise, increment i and go to step 2.
8. Let M be the $(n+3) \times n$ matrix whose ith row is the vector $\mathbf{y}_i$. Solve the system $M\mathbf{x} = \mathbf{0}$ (where we treat $\mathbf{x}$

and $\mathbf{0}$ as column vectors). If there is a unique non-zero solution $\mathbf{x} = \mathbf{s}$, then output $\mathbf{s}$. If the only solution is $\mathbf{x} = \mathbf{0}$, then output $\mathbf{0}$. Otherwise, output "FAIL".

Note that in step 3, since we can partition the space $\mathbb{Z}_2^n$ into a disjoint union of cosets $\mathbb{Z}_2^n/K$ of the subgroup $K = \{\mathbf{0}, \mathbf{s}\}$, where $f(\mathbf{x})$ is constant on each coset, we can rewrite the state of the system as

$$\frac{1}{\sqrt{2^n}} |\mathbf{x}\rangle |f(\mathbf{x})\rangle = \frac{1}{\sqrt{2^{n-1}}} \sum_{\{\mathbf{z},\mathbf{z}\oplus\mathbf{s}\}\in\mathbb{Z}_2^n/K} \frac{1}{\sqrt{2}}(|\mathbf{z}\rangle + |\mathbf{z}\oplus\mathbf{s}\rangle)|f(\mathbf{z})\rangle \ .$$

Since the rest of the algorithm ignores the second register, we can assume that in the first register we have a random "coset state" $\frac{1}{\sqrt{2}}(|\mathbf{z}\rangle + |\mathbf{z}\oplus\mathbf{s}\rangle)$ selected uniformly at random from all the coset states.

After applying the Hadamard gates, the coset state gets mapped to a equally weighted superposition of states that are orthogonal to K,

$$\frac{1}{\sqrt{2^{n-1}}} \sum_{\mathbf{y}\in K^\perp} (-1)^{\mathbf{y}\cdot\mathbf{z}} |\mathbf{y}\rangle$$

where $K^\perp = \{\mathbf{y} \mid \mathbf{y} \in \mathbb{Z}_2^n, \mathbf{y}\cdot\mathbf{s} = 0\}$. Thus, with $n + O(1)$ random sampled from $K^\perp$, the samples vectors $\mathbf{y}_i$ will generated $K^\perp$ with high probability. Then, using linear algebra, we can efficiently compute generators for K. In the generalizations of Simon's algorithm that we see in the next few sections, we'll see a more general formulation of $K^\perp$ for a hidden subgroup K.

Shortly after Simon came up with his black-box algorithm, Shor [165] used similar ideas to derive his famous algorithms for factoring and finding discrete logarithms.

Quantum Algorithms for Simon's Problem Simon's algorithm solves this problem with bounded error using $n + O(1)$ applications of U_f and $O(n)$ other elementary quantum operations and $O(n^3)$ elementary classical operations.

Brassard and Høyer [35] combined Simon's algorithm with amplitude amplification order to make the algorithm exact.

Classical Algorithms for Simon's Problem Simon [167] showed a lower bound of $\Omega(2^{\frac{n}{4}})$ queries, and this can be improved to $\Omega(2^{\frac{n}{2}})$.

This was a historical moment, since the field of quantum algorithms progressed from work on black-box algorithms and complexity, to having an algorithm without black-boxes and of broad practical importance. It is important to note the fundamental role of the foundational complexity-theoretic work and black-box algorithms to the development of the practical algorithms.

Factoring, Discrete Logarithms and the Abelian Hidden Subgroup Problem

Factoring, Finding Orders and Periods, and Eigenvalue Estimation

The most well known quantum algorithm is Shor's algorithm [165,166] for the integer factorization problem.

Integer Factorization Problem
Input: An integer N.
Problem: Output positive integers $p_1, p_2, \ldots, p_l, r_1, r_2, \ldots, r_l$ where the p_i are distinct primes and $N = p_1^{r_1} p_2^{r_2} \ldots p_l^{r_l}$.

This problem can be efficiently reduced (that is, in time polynomial in $\log N$), by a probabilistic classical algorithm, to $O(l)$ instances (note that $l \in O(\log N)$) of the problem of finding the multiplicative order of an element modulo N, which can be solved efficiently on a quantum computer.

Order Finding Problem
Input: Positive integers a and N, such that $\mathrm{GCD}(a, N) = 1$ (i. e. a is relatively prime to N).
Problem: Find the order of a modulo N.

Essentially the same quantum algorithm can efficiently find the order of an element a in a finite group G given a black-box for performing the group arithmetic. Shor's algorithm in fact solves the more general problem of finding the period of a periodic function f.

Period Finding Problem
Input: A black-box implementing a periodic function $f\colon \mathbb{Z} \mapsto X$ for some finite set X, where $f(x) = f(y)$ if and only if $r \mid x - y$.
Problem: Find the period r.

Shor described his algorithm for the specific function $f(x) = a^x \mod N$, where N was the integer to be factored; however the algorithm will find the period of any such periodic function (note the assumption that the values of $f(1), f(2), \ldots, f(r)$ are distinct; one can also analyze the case where the values are not entirely distinct [32,145]).

Kitaev later showed that the problem of finding the order of $a \in G$ can alternatively be reduced to the problem of estimating the eigenvalues of the operator that multiplies by a, and he described a efficient quantum algorithm for estimating such eigenvalues [126]. His method is to reduce the problem to phase estimation. Although Kitaev's algo-

rithm was qualitatively different [116], it was shown that using an improved phase estimation algorithm yields an order-finding circuit that is essentially equivalent to Shor's factoring algorithm based on period-finding [54].

Both Shor's and Kitaev's approaches find r by finding good estimates of a random integer multiple of $\frac{1}{r}$ and then applying the continued fractions algorithm to find r.

Sampling Estimates to an Almost Uniformly Random Integer Multiple of $\mathbf{1/r}$
Input: Integers a, and N such that $\mathrm{GCD}(a, N) = 1$. Let r denote the (unknown) order of a.
Problem: Output a number $x \in \{0, 1, 2, \ldots, 2^n - 1\}$ such that for each $k \in \{0, 1, \ldots, r-1\}$ we have

$$\Pr\left(\left|\frac{x}{2^n} - \frac{k}{r}\right| \le \frac{1}{2r^2}\right) \ge c\frac{1}{r}$$

for some constant $c > 0$.

Shor's analysis of the algorithm works by creating a "periodic state", which is done by preparing $\sum_x |x\rangle |a^x \mod N\rangle$ and measuring the second register (in fact, just ignoring or tracing out the second register suffices, since one never uses the value of the measurement outcome). One then applies the quantum Fourier transform, or its inverse, to estimate the period.

Kitaev's algorithm works by estimating a random eigenvalue of the operator U_a that multiplies by a.

Eigenvalue Estimation Problem
Input: A quantum circuit implementing the controlled-U, for some unitary operator U, and an eigenstate $|\psi\rangle$ of U with eigenvalue $e^{2\pi i\omega}$.
Problem: Obtain a good estimate for ω.

He solves the eigenvalue estimation problem by solving a version of the well-studied phase estimation problem, and uses the crucial fact (as did Shor) that one can efficiently implement a circuit that computes $U_a^{2^j} = U_{a^{2^j}}$ using j group multiplications instead of 2^j multiplications. Thus one can efficiently reduce the eigenvalue estimation problem to the following phase estimation problem.

Phase Estimation Problem
Input: The states $\frac{1}{\sqrt{2}}(|0\rangle + e^{2\pi i\omega y}|1\rangle)$, for $y = 1, 2, 4, \ldots, 2^n$, for some $\omega \in [0, 2\pi)$.
Problem: Obtain a good estimate of the phase parameter ω.

Kitaev's phase estimation algorithm used $O(n)$ copies of the states $\frac{1}{\sqrt{2}}(|0\rangle + e^{2\pi i\omega y}|1\rangle)$ for $y = 1, 8, 64, \ldots$, and provides an estimate with error at most $\frac{1}{2^n}$ with high probability. Although there are slightly more efficient phase estimation algorithms, one advantage of his algorithm is that it does not require a quantum Fourier transform, and instead performs some efficient classical post-processing of estimates of $y\omega \mod 1$. This might be advantageous experimentally. The phase estimation problem has been studied for several decades [100,102] recently with some focus on the algorithmic complexity of the optimal or near-optimal estimation procedures.

Quantum Algorithms for Order Finding Finding the order of a random element in $\mathbb{Z}_N^*$

- Quantum complexity is in $O((\log N)^2 \log\log(N) \log\log\log(N))$.

Order finding in a black-box group

- Quantum black-box complexity (for groups with unique encodings of group elements) is $O(\log r)$ black-box multiplications and $O(n + \log^2 r)$ other elementary operations.

Classical Algorithms for Order Finding Finding the order of a random element in $\mathbb{Z}_N^*$

- Best known rigorous probabilistic classical algorithm has complexity in $e^{O(\sqrt{\log N \log\log N})}$.
- Best known heuristic probabilistic classical algorithm has complexity in $e^{O((\log N)^{\frac{1}{3}} (\log\log N)^{\frac{2}{3}})}$.

Order finding in a black-box group

- Classical black-box multiplication complexity is in $\Theta(\sqrt{r})$ [53].

By "heuristic" algorithm, we mean the proof of its running time makes some plausible but unproven assumptions.

Discrete Logarithms

Shor [165,166] also solved the problem of finding discrete logarithms in the multiplicative group of a finite field.

The Discrete Logarithm Problem
Input: Elements a and $b = a^t$ in $\mathbb{Z}_p^*$, where t is an integer from $\{0, 1, 2, \ldots, r-1\}$ and r is the order of a.
Problem: Find t. (The number t is called the discrete logarithm of b with respect to the base a.)

Shor solved this problem by defining $f\colon \mathbb{Z}_r \times \mathbb{Z}_r \mapsto \mathbb{Z}_p^*$ as $f(x, y) = a^x b^y$. Note that $f(x_1, y_1) = f(x_2, y_2)$ if and only if $(x_1 - x_2, y_1 - y_2)$ is in the additive subgroup of $\mathbb{Z}_r \times \mathbb{Z}_r$ generated by $(t, -1)$. Since r is known, there is no need for the continued fractions algorithm. There is also an analogous eigenvalue estimation version of this algorithm.

The algorithm can be defined in general for any group G where we have a black-box for computing the group operation. Given $a, b \in G$, output the smallest positive t such that $b = a^t$. For example, one can apply this algorithm to finding discrete logarithms in the additive group of an elliptic curve over a finite field [121,150], a group widely used in public key cryptography [137]. In this case, the group operation is described as addition, and thus the problem is to find the smallest positive integer t such that $b = ta$, where b and a are points on some elliptic curve.

Quantum Complexities of the Discrete Logarithm Problem

- Finding discrete logarithms in $\mathbb{F}_q^*$
  - Quantum complexity is in $O((\log q)^2 \log\log(q) \log \log\log(q))$.
- Discrete logarithms in a black-box group represented with strings of length n (including elliptic curve groups discussed above)
  - Quantum black-box complexity (for groups with unique encodings of group elements) is $O(\log r)$ black-box multiplications and $O(n + \log^2 r)$ other elementary operations.

Classical Complexities of the Discrete Logarithm Problem

- Finding discrete logarithms in $\mathbb{F}_q^*$
  - Best known rigorous probabilistic classical algorithm has complexity in $e^{O(\sqrt{\log q \log\log q})}$ for certain values of q (including $q = 2^n$ and prime q).
  - Best known heuristic probabilistic classical algorithm has complexity in $e^{O((\log q)^{\frac{1}{3}} (\log\log q)^{\frac{2}{3}})}$.
- Discrete logarithms in a black-box group represented with strings of length n
  - Classical black-box complexity is in $\Theta(\sqrt{r})$. For a large class of elliptic curves, the best known classical algorithms have complexity in $O(\sqrt{r})$ group additions. There are sub-exponential algorithms for special families of curves.

Abelian Hidden Subgroup Problem

Notice how we can rephrase most of the problems we have already discussed, along with some other ones, as a special case of the following problem.

The Abelian Hidden Subgroup Problem Let $f: G \to X$ map an Abelian group G to some finite set X with the property that there exists some subgroup $K \leq G$ such that for any $x, y \in G$, $f(x) = f(y)$ if and only if $x + K = y + K$. In other words f is constant on cosets of K and distinct on different cosets.

Deutsch's problem $G = \mathbb{Z}_2$, $X = \{0, 1\}$, and $K = \{0\}$ if f is balanced, and $K = \{0, 1\}$ if f is constant.

Generalized Simon's problem $G = \mathbb{Z}_2^n$, $X = \{0, 1\}^n$, and K is any subgroup of $\mathbb{Z}_2^n$.

Finding orders $G = \mathbb{Z}$, X is any finite group H, r is the order of $a \in H$. The subgroup $K = r\mathbb{Z}$ is the hidden subgroup of G, and a generator for K reveals r.

Finding the period of a function $G = \mathbb{Z}$, X is any set, r is the period of f. The subgroup $K = r\mathbb{Z}$ is the hidden subgroup of G, and a generator for K reveals the period r.

Discrete logarithms in any group $G = \mathbb{Z}_r \times \mathbb{Z}_r$, X is any group H. Let a be an element of H with $a^r = 1$ and suppose $b = a^k$. Consider the function $f(x_1, x_2) = a^{x_1} b^{x_2}$. We have $f(x_1, x_2) = f(y_1, y_2)$ if and only if $(x_1, x_2) - (y_1, y_2) \in \{(tk, -t), t = 0, 1, \ldots, r-1\}$. The hidden subgroup K is the subgroup generated by $(k, -1)$, where k is the discrete logarithm.

Hidden linear functions [34] $G = \mathbb{Z} \times \mathbb{Z}$. Let g be some permutation of $\mathbb{Z}_N$ for some integer N. Let h be a function from $\mathbb{Z} \times \mathbb{Z}$ to $\mathbb{Z}_N$ defined by $h(x, y) = x + ay \mod N$. Let $f = g \circ h$. The subgroup K is the hidden subgroup generated by $(-a, 1)$, and the generator reveals the hidden linear function h.

Self-shift-equivalent polynomials [83] Given a polynomial P in l variables $X_1, X_2, \ldots, X_l$ over $\mathbb{F}_q$ (the finite field with q elements), the function f which maps $(a_1, a_2, \ldots, a_l) \in \mathbb{F}_q^l$ to $P(X_1 - a_1, X_2 - a_2, \ldots, X_l - a_l)$ is constant on cosets of a subgroup K of $\mathbb{F}_q^l$. This subgroup K is the set of self-shift-equivalences of the polynomial P.

Abelian stabilizer problem [119] Let G be any group acting on a finite set X. That is, each element of G acts as a map from X to X in such a way that for any two elements $a, b \in G$, $a(b(x)) = (ab)(x)$ for all $x \in X$. For a particular element $x \in X$, the set of elements which fix x (that is the elements $a \in G$ such that $a(x) = x$) form a subgroup. This subgroup is called the stabilizer of x in G, denoted $St_G(x)$. Let f_x denote the function from G to X which maps $g \in G$ to $g(x)$. The hidden subgroup of f_x is $St_G(x)$.

If we restrict attention to finite Abelian groups, or more generally, finitely generated Abelian groups, then we can efficiently solve the hidden subgroup problem, by generalizations of the algorithms for factoring, finding discrete logarithms, and Simon's problem.

The Abelian hidden subgroup problem can also be used to decompose a finite Abelian group into a direct sum of cyclic groups if there is a unique representative for each group element [44,143]. For example, the multiplicative group of integers modulo N is an Abelian group, and we

can efficiently perform computations in the group. However, having a decomposition of the group would imply an efficient algorithm for factoring N. The class group of a number field is another Abelian group for which a decomposition is believed to be hard to find on a classical computer. For example, such a decomposition would easily give the size of the class group, which is also known to be as hard as factoring, assuming the Generalized Riemann Hypothesis. Computing the class group of a complex number field is a simple consequence of the algorithm for decomposing an Abelian group into a direct sum of cyclic groups, since there are techniques for computing unique representatives in this case. A unique classical representation is sufficient, but a quantum state could also be used to represent a group element. For example, a uniform superposition over classical representatives of the same group element would also work (this technique was applied by Watrous in the case of solvable groups [175]). Computing the class number of a real number field is not as straightforward, since there is no known way to efficiently compute a unique classical representative for each group element. However Hallgren [95] used the techniques outlined in Subsect. "Lattice and Number Field Problems" to show how to compute quantum states to represent the group elements, assuming the Generalized Riemann Hypothesis, in the case of class group of a real number field of constant degree, and thus is able to compute the class group in these cases as well .

Quantum Algorithms for the Abelian Hidden Subgroup Problem There exists a bounded-error quantum algorithm for finding generators for the hidden subgroup $K \leq G = \mathbb{Z}_{N_1} \times \mathbb{Z}_{N_2} \times \cdots \times \mathbb{Z}_{N_l}$ of f using $O(l)$ evaluations of f and $O(\log^3 N)$ other elementary operations, where $N = N_1 N_2 \ldots N_l = |G|$.

It can be shown that $\Omega(l)$ queries are needed for worst-case K.

Classical Algorithms for the Abelian Hidden Subgroup Problem In the black-box model $\Omega(\sqrt{|G/K|})$ queries are necessary in order to even decide if the hidden subgroup is trivial.

Generalizations are discussed in Sect. "Generalizations of the Abelian Hidden Subgroup Problem".

Algorithms Based on Amplitude Amplification

In 1996 Grover [90] developed a quantum algorithm for solving the search problem that gives a quadratic speed-up over the best possible classical algorithm.

Search Problem
Input: A black-box U_f for computing an unknown function $f\colon \{0,1\}^n \to \{0,1\}$.
Problem: Find a value $\mathbf{x} \in \{0,1\}^n$ satisfying $f(\mathbf{x}) = 1$, if one exists. Otherwise, output "NO SOLUTION".

The *decision version* of this problem is to output 1 if there is a solution, and 0 if there is no solution.

This problem is stated in a very general way, and can be applied to solving a wide range of problems, in particular any problem in *NP*. For example, suppose one wanted to find a solution to an instance Φ of the 3-SAT problem. The Boolean formula Φ is in "3-conjunctive normal form" (3-CNF), which means that it is a conjunction (logical AND) of clauses, each of which is a disjunction (logical OR) of three Boolean variables (or their negations). For example, the following is a 3-CNF formula in the variables $b_1, b_2, \ldots, b_5$:

$$\Phi = (b_1 \vee \bar{b_2} \vee b_5) \wedge (b_1 \vee \bar{b_4} \vee \bar{b_5}) \wedge (b_4 \vee b_2 \vee b_3)\,,$$

where we let $\bar{b}$ denote the logical NOT of b. A "satisfying assignment" of a particular 3-CNF formula Φ is an assignment of 0 or 1 values to each of the n variables such that the formula evaluates to 1. Given a satisfying assignment, it is easy to check if it satisfies the formula. Define f_Φ to be the function that, for any $\mathbf{x} = x_1 x_2 \ldots x_n \in \{0,1\}^n$, maps $\mathbf{x} \mapsto 1$ if the assignment $b_i = x_i$, $i = 1, 2, \ldots, n$ satisfies Φ and $\mathbf{x} \mapsto 0$ otherwise. Solving the search problem for f_Φ is equivalent to finding a satisfying assignment for Φ.

If we only learn about f by applying the black-box U_f, then any algorithm that finds a solution with high probability for any f (even if we just restrict to functions f with at most one solution) requires $\Omega(\sqrt{2^n})$ applications of U_f [25].

The basic intuition as to why a quantum algorithm might provide some speed-up is that a quantum version of an algorithm that guesses the solution with probability $\frac{1}{2^n}$ will in fact produce the correct answer with probability amplitude $1/\sqrt{2^n}$. The hope is that additional guesses increase the amplitude of finding a correction solution by $\Omega(1/\sqrt{2^n})$, and thus $O(\sqrt{2^n})$ guesses and applications of U_f might suffice in order to get the total probability amplitude close to 1. Of course, one needs to find a unitary algorithm for implementing this, and obvious approaches do not permit the amplitudes to add constructively. Lov Grover discovered a quantum algorithm that does permit the amplitudes to add up in such a way. This algorithm was analyzed and generalized to the technique known as "amplitude amplification" [34,35,38]. Further generalizations

of these search algorithms are discussed in Sect. "Quantum Walk Algorithms".

Quantum Algorithms for Searching: There exists a quantum algorithm that finds a solution with probability at least $\frac{2}{3}$ if one exists, otherwise outputs "NO SOLUTION", and uses $O(\sqrt{2^n})$ applications of U_f. The algorithm uses $O(n\sqrt{2^n})$ elementary gates. Note that $\Omega(\sqrt{2^n})$ applications of U_f are necessary in order to find a solution for any f, even if we restrict attention to functions f with at most one solution.

If, as an additional input, we are also given an algorithm A that will find a solution to $f(x) = 1$ with probability p, then amplitude amplification will find a solution with high probability using $O(1/\sqrt{p})$ applications of U_f and of unitary versions of A and A^{-1}.

This more general statement implies that if there are $m \geq 1$ solutions, then there is an algorithm which makes an expected number of queries in $O(\sqrt{2^n/m})$ (since guessing uniformly at random will succeed with probability $\frac{m}{2^n}$). This algorithm works even if the unitary black-box U_f computes f with a small bounded error [108].

Classical Algorithms for the Search Problem: Any classical algorithm must make $\Omega(2^n)$ queries in order to succeed with probability at least $\frac{2}{3}$ on any input f, even if we restrict attention to functions with at most one solution. Exhaustive searching will find a solution using $O(2^n)$ applications of U_f.

If we are also given a black-box A that successfully guesses a solution to $f(x) = 1$ with probability $p \geq \frac{1}{2^n}$, then $\Theta(\frac{1}{p})$ applications of U_f are needed and also sufficient (using random sampling).

If instead of a black-box for U_f, we are given a circuit for f or some other description of f (such as a description of the 3-SAT formula Φ in the case of f_Φ) then there might be more efficient algorithms that uses this additional information to do more than just use it to apply U_f. One can directly use amplitude amplification to get a quadratic speed-up for any such heuristic algorithm that guesses a solution with probability p, and then repeats until a solution is found.

The quantum searching algorithm has often been called a 'database' search algorithm, but this term can be misleading in practice. There is a quadratic speed-up if there is an implicit database defined by an efficiently computable function f. However, if we are actually interested in searching a physical database, then a more careful analysis of the physical set-up and the resources required is necessary. For example, if a database is arranged in a linear array, then to query an N-element database will take time proportional to N. If the data is arranged in a square grid, then a general lookup could be possibly be done in time proportional to $\sqrt{N}$. This limitation is true for both classical and quantum searching; however the hardware assumptions in the two cases are often different, so one needs to be careful when making a comparison. (e. g. [158,180] and Chap. 6 of [149], discuss these issues).

Local searching [2] restricts attention to architectures or models where subsystems can only interact locally, and does not allow access to any memory location in one unit of time regardless of the memory size (which is not truly possible in any case, but can sometimes be appropriate in practice). The algorithms used include quantum walk algorithms (e. g. [182]) and can achieve quadratic or close to quadratic speed-ups depending on details such as the spatial dimension of the database.

It is also important to note that there has been much work studying the quantum query complexity of searching an *ordered* list. It is known that $\Theta(\log N)$ queries are necessary and sufficient, but the constant factor is not yet known [49,98,107].

Other Applications of Amplitude Amplification

Apart from the application to the searching problem, one of the first applications of the quantum searching algorithm was to counting [37], and more generally amplitude amplification can be used to estimate amplitudes [38]. The quantum algorithm for amplitude estimation combines the techniques of the order finding algorithm with amplitude amplification. Bounds on optimal phase estimation translate to bounds on optimal amplitude estimation. It has several applications, including approximate and exact counting [37,38], and approximating the mean (or, in the continuous case ► Quantum Algorithms and Complexity for Continuous Problems, the integral) of a function [172]. These applications offer up to a quadratic speed-up over classical algorithms.

We have already mentioned the straight-forward, and very useful, application of amplitude amplification to searching: providing a quadratic speed-up up for any algorithm that consists of repeating a subroutine that guesses a solution, until a solution is found. However, there are many applications of amplitude amplification that are not so straightforward, and require some additional nontrivial algorithmic insights. Other applications of amplitude amplification include the collision finding problem, which has applications to breaking hash functions in cryptography.

Collision Finding Problem:
Input: A black-box U_f for computing a function

$f\colon \{0,1\}^n \to \{0,1\}^m$. The function f is r to 1, for some positive integer r.
Problem: Find two distinct $\mathbf{x}, \mathbf{y} \in \{0,1\}^n$ with $f(\mathbf{x}) = f(\mathbf{y})$, if $r > 1$. Otherwise output "NO COLLISION".

Quantum Algorithms for Collision Finding There exists a quantum algorithm [36] that uses $O((2^n/r)^{1/3})$ applications of U_f, and $O((2^n/r)^{1/3}\,\mathrm{polylog}\,(2^n/r))$ other elementary operations, and $O((2^n/r)^{1/3})$ space, and outputs a collision with probability at least $\frac{2}{3}$ if $r > 1$, and outputs "NO COLLISION" otherwise.

It can be shown that $\Omega((2^n/r)^{1/3})$ applications of U_f are needed.

Classical Algorithms for Collision Finding There exists a classical probabilistic algorithm that uses $O((2^n/r)^{1/2})$ applications of U_f, and $O((2^n/r)^{1/2}\,\mathrm{polylog}\,(2^n/r))$ other elementary operations, and $O(n)$ space, and outputs a collision with probability at least 2/3 if $r > 1$, and outputs "NO COLLISION" otherwise. It can be shown that $\Omega((2^n/r)^{1/2})$ applications of U_f are needed.

Other applications include finding claws [37], finding the maximum (or minimum) value of a function [106], string matching [153], estimating the median of a function [91,148] and many others [67,84]. There is also a non-trivial application to the element distinctness problem [42], which we define in Sect. "Quantum Walk Algorithms", since there is a better, optimal, quantum walk algorithm. Amplitude amplification also a basic tool in making algorithms exact [35,146].

Simulation of Quantum Mechanical Systems

Feynman's reason for considering a quantum computer was to simulate other quantum mechanical systems [79]. This general idea was described in more detail by Lloyd [131], Wiesner [178] and Zalka [180], and later by many other authors (e. g. [149]). More detailed situations have been studied by numerous authors, including [43,45,119,170].

There are other notions of what one what one might mean by simulating a physical system, such as computing properties of the ground state of the Hamiltonian, or other properties of the spectrum of the Hamiltonian. In this section, we focus on the problem of simulating the evolution of a quantum mechanical system given the initial state (or a description of it) and a description of the Hamiltonian of the system.

For convenience, we will restrict attention to a discrete Hilbert space. In practice, when modelling systems with a continuous state space, the state space will need to be approximated with a discrete state space [180] with enough states so that the accumulated error is below the desired error threshold. This general problem of dealing with continuous problems is discussed in ► Quantum Algorithms and Complexity for Continuous Problems.

We will also restrict attention to time-independent Hamiltonians, since well-behaved time-dependent Hamiltonian evolution can be approximated by a sequence of time-dependent Hamiltonian evolutions.

There are two natural ways that have been studied for representing the Hamiltonian. The first way, is to represent $H = \sum_{k=1}^{M} H_k$, where H_k is a simple enough Hamiltonian that we know how to efficiently simulate its evolution. For convenience, we assume the simulation of the H_k term for a time t takes unit cost and is exact. In practice, if the simulation can be done with error ϵ in time polynomial in $\frac{1}{\epsilon}$, $\tau = ||H_k||t$, and the number of qubits n, then the effect on the overall cost is by a factor that is polynomial in these factors. In many situations, the simulation is polynomial in $\log(\frac{1}{\epsilon})$ and $\log \tau$, and independent of n and ϵ. This leads to the following formulation:

Hamiltonian Simulation Problem 1
Input: An integer n and black-boxes $A_1, A_2, \ldots, A_M$, where A_j takes a non-negative real number r and executes $\mathrm{e}^{iH_j r}$, for a set of Hamiltonians H_j acting on n qubits.
The value of $\|H\|$, the trace norm of $H = \sum_{k=1}^{M} H_k$.
A positive real number t.
A positive real number $\epsilon < 1$.
An n-qubit state $|\psi\rangle$.
Output: A state $|\psi_f\rangle$ satisfying $\big\| |\psi_f\rangle - \mathrm{e}^{iHt}|\psi\rangle \big\| < \epsilon$.

In practice, the input will likely be a classical description of how to prepare the state $|\psi\rangle$. It suffices to have an upper bound on $\|H\|$ (if the bound is within a constant factor, this won't affect the stated complexities). In particular, it suffices that we can efficiently compute or approximate the eigenvalue λ_k of a given an eigenvector $|E_k\rangle$ of H_k. This is the case, for example, when the Hilbert space is a tensor product of n finite sized subsystems and H_k acts non-trivially only on a finite number of subsystems, say c of them. In other words, up to a reordering of the subsystem labels, $H_k = I^{n-c} \otimes \tilde{H}_k$. Since

$$\mathrm{e}^{iH_k t} = I^{n-c} \otimes \mathrm{e}^{i\tilde{H}_k t}\,,$$

in order to simulate the evolution of H_k for a time interval of size t, one only needs to simulate $\tilde{H}_k$ on the relevant c qubits for a time interval of size t. In this case, since we know $\tilde{H}_k$, one can use "brute-force" methods to approximately implement the map $|E_{\lambda_k}\rangle \mapsto \mathrm{e}^{2\pi i \lambda_k t}|E_{\lambda_k}\rangle$, for any time interval t.

An easy example is when the state space is n qubits, and H_k is a tensor product of Pauli operators. This means that

we can easily diagonalize $\tilde{H}_k$ as $\tilde{H}_k = (P_1 \otimes P_2 \otimes \ldots \otimes P_c) Z^c (P_1^\dagger \otimes P_2^\dagger \otimes \ldots \otimes P_c^\dagger)$ for some one-qubit unitary operations $P_1, P_2, \ldots, P_c$. Thus we have

$$e^{i\tilde{H}_k t} = (P_1 \otimes P_2 \otimes \ldots \otimes P_c) e^{iZ^c t} (P_1^\dagger \otimes P_2^\dagger \otimes \ldots \otimes P_c^\dagger) .$$

Since

$$e^{iZ^c t} \,|\, x_1 x_2 \ldots x_c\rangle = e^{itf(x_1 \ldots x_c)} \,|\, x_1 x_2 \ldots x_c\rangle$$

where $f(x_1, \ldots, x_c) = x_1 \oplus \ldots \oplus x_c$, this simulation can be done easily. In fact, as pointed out in [149], in the case that H_k is such a product Hamiltonian, c does not need to be constant.

Another example, [180], is where the eigenvectors of H_k are of the form

$$|\, E_k\rangle = \sum_{j=0}^{2^n-1} e^{2\pi i jk/2^n} \,|\, j\rangle$$

(i. e. "momentum" eigenstates), in which case the inverse quantum Fourier transform will map $|\, E_k\rangle \mapsto |\, k\rangle$, and then one can easily approximate $|\, k\rangle \mapsto e^{i\lambda_k t} \,|\, k\rangle$ for any efficiently computable function λ_k, and then apply the quantum Fourier transform to map $|\, k\rangle \mapsto |\, E_k\rangle$ and thus effectively compute the transformation $|\, E_k\rangle \mapsto e^{i\lambda_k t} \,|\, E_k\rangle$.

Thus we can study any situation where $H = \sum_k H_k$, and we have some means of efficiently simulating the time evolution of H_k. If the H_k commute, then

$$e^{iHt} = e^{iH_1 t} e^{iH_2 t} \ldots e^{iH_n t} ,$$

and the problem is straightforward. For the non-trivial case when the H_k do not commute pairwise, we can take advantage of approximations derived from Trotter formulas, like

$$e^{iHt} = \left(e^{iH_1 t/n} e^{iH_2 t/n} \ldots e^{iH_n t/n} \right)^n + O(t^2/n)$$

and other improved versions. A good estimate of the overall cost of this family of simulations is the number of terms of the form $e^{iH_j r}$ that are used, for any choice of r (we can assume $0 \le r \le t$, and that r is some efficiently computable number).

Quantum Complexity of Hamiltonian Simulation Problem 1 There is a quantum algorithm that simulates e^{iHt} on a given input $|\, \psi\rangle$ with trace distance error at most $\epsilon < 1$ that uses a number of exponential terms N_{exp} satisfying

$$N_{\text{exp}} \in (M\tau + 1) M^{1+o(1)} \tau^{o(1)} \left(\frac{1}{\epsilon}\right)^{O(1/\sqrt{s})} ,$$

where $\tau = \|H\|\, t$ and $\epsilon = 1/2^s$. Since the running time of simulating each exponential term is assumed to be polynomial in n, the overall running time is polynomial in n.

In general, there are Hamiltonians for which $\Omega(\tau)$ time is necessary.

The product $\tau = \|H\|\, t$ is the relevant parameter since one can effectively speed up time by a factor of s by rescaling the Hamiltonian by a multiplicative factor of s, for any $s > 0$ (e. g. one can perform an operation in half the time by doubling all the energies).

It is hard to give an explicit comparison to the best known classical complexity of this general problem. There are classical algorithms and heuristics for special cases, but in the worst-case, the best solutions known require time exponential in n and polynomial in τ (e. g. in $\tau^{1+o(1)}$).

Another natural formulation of the problem [5,29] considers the case of sparse Hamiltonians, where for any basis state $|\, x\rangle$ there are at most d basis states $|\, y\rangle$ such that $\langle\, x|\, H \,|\, y\rangle \neq 0$ and one can efficiently compute this neighborhood for any input x. This leads to the following black-box formulation of the problem.

Hamiltonian Simulation Problem 2
Input: Integers n and d, and a black-box U_H that maps $|\, x, i\rangle\, |\, 0\rangle \mapsto |\, x, i\rangle \,\big|\, y_i, H_{x,y_i}\big\rangle$, where y_i is the index of the ith non-zero entry in column x of the Hermitian matrix H (if there are $d' < d$ non-zero entries, then U_f can output any index for $i > d'$). The values $H_{x,y} = \langle\, x|\, H \,|\, y\rangle$ are the (x, y) entries of H represented in the computational basis.
A positive real number t.
A positive real number $\epsilon < 1$.
An n-qubit state $|\, \psi\rangle$.
Output: A state $\big|\, \psi_f\big\rangle$ satisfying $\big\| \big|\, \psi_f\big\rangle - e^{iHt} \,|\, \psi\rangle \big\| < \epsilon$.

The best known general solution was shown in [29].

Quantum Complexity of Hamiltonian Simulation Problem 2 There is a quantum algorithm that simulates e^{iHt} on a given input $|\, \psi\rangle$ with trace distance error at most $\epsilon < 1$ that uses a number of black-box calls, N_{bb}, satisfying, for any positive integer k,

$$N_{\text{bb}} \in O\left(\tau^{(1+\frac{1}{2k})} d^{(4+\frac{1}{k})} 5^{2k} \left(\frac{1}{\epsilon}\right)^{\frac{1}{2k}} (\log^* n) \right) ,$$

where $\tau = ||H||t$ and $\log^* n$ is the smallest positive integer r such that $\log_2^{(r)} n < 2$, where $\log_2^{(r)}$ refers to iterating the $\log_2$ function r times.

In general, there are Hamiltonians that require $\Omega(\tau)$ black-box evaluations.

For example, setting $k = \sqrt{s}$ where $\epsilon = \frac{1}{2^s}$, we get

$$N_{\mathrm{bb}} \in \tau^{(1+O(\frac{1}{\sqrt{s}}))} d^{(4+O(\frac{1}{\sqrt{s}}))} \left(\frac{1}{\epsilon}\right)^{O\left(\frac{1}{\sqrt{s}}\right)} (\log^* n) .$$

Some applications involve trying to study properties of the ground state of a Hamiltonian by combining simulation of the dynamics with some approach for generating the ground state with non-trivial initial amplitude. For example, if one had a procedure A for generating a ground state $|E_0\rangle$ of H with probability p, then one can combine the algorithm for simulating H with eigenvalue estimation, and then amplitude amplification on the eigenvalue estimates in order to generate the ground state with high fidelity. The algorithm would use $O(1/\sqrt{p})$ applications of A and A^{-1} and $O(1/\sqrt{p})$ eigenvalue estimations. The eigenvalue estimations would have to be precise enough to distinguish $|E_0\rangle$ from the next highest energy eigenstate produced by A. If the gap between these two energies is Δ, then the eigenvalue estimation would involve simulating the evolution of H for a time in $\Omega(\frac{1}{\Delta})$.

Several papers address techniques for generating the ground states for various problems of interest including [3]. Another application for such simulations is to implement algorithms that are designed in the continuous-time model in a discrete-time quantum computation model.

Generalizations of the Abelian Hidden Subgroup Problem

Non-Abelian Hidden Subgroup Problem

One of the most natural ways to generalize the Abelian hidden subgroup problem is to non-Abelian groups. The problem definition is the same, apart from letting G be non-Abelian. One natural problem in this class of problems is the graph automorphism problem.

Graph automorphism problem: Consider $G = S_n$, the symmetric group on n elements, which corresponds to the permutations of $\{1, 2, \ldots, n\}$. Let $\mathbb{G}$ be a graph on n vertices labeled $\{1, 2, \ldots, n\}$. For any permutation $\sigma \in S_n$, let $f_{\mathbb{G}}$ map S_n to the set of n-vertex graphs by mapping $f_{\mathbb{G}}(\sigma) = \sigma(\mathbb{G})$, where $\sigma(\mathbb{G})$ is the graph obtained by permuting the vertex labels of $\mathbb{G}$ according to σ. For the function $f_{\mathbb{G}}$, the hidden subgroup of G is the automorphism group of $\mathbb{G}$ (i. g. the permutations σ such that $\sigma(\mathbb{G}) = \mathbb{G}$).

The graph isomorphism problem (deciding if two graphs $\mathbb{G}_1$ and $\mathbb{G}_2$ are isomorphic) can be reduced to solving the graph automorphism problem. The best known classical algorithm takes time in $e^{O(\sqrt{n \log n})}$ to decide if two graphs are isomorphic, and there is no substantially better quantum algorithm known.

There has been much work attacking the non-Abelian hidden subgroup problem. Ettinger, Høyer and Knill [71] showed the following.

Query Complexity of the Non-Abelian Hidden Subgroup Problem For any finite group G, the non-Abelian hidden subgroup problem can be solved with high probability using $O(\log |G|)$ queries to U_f.

Thus, the main question remaining is whether it is possible for the entire algorithm, including black-box queries and other elementary operations, to be efficient.

Quantum Fourier Transform Approaches One natural approach is to mimic the algorithm for the Abelian HSP, which starts by computing

$$\frac{1}{\sqrt{|G|}} \sum_{x \in G} |x\rangle |f(x)\rangle$$

and noticing that this equals

$$\sum_{y \in f(G)} \frac{1}{\sqrt{|f(G)|}} \left(\frac{1}{\sqrt{|K|}} \sum_{x \in f^{-1}(y)} |x\rangle \right) |y\rangle .$$

For convenience, we suppose we measure the second register (it suffices just to trace out the 2nd register) and get some random outcome y, and thereby project the first register into the state

$$|a + K\rangle = \frac{1}{\sqrt{|K|}} \sum_{x \in K} |a + x\rangle$$

where a is any element of G such that $f(a) = y$. In other words, as in the Abelian HSP algorithm, the first register is an equal superposition of all the elements in some coset of K.

The question is: how do we extract information about K given a random coset state of K? In the Abelian case, a quantum Fourier transform of the coset state allowed us to sample elements that were orthogonal to K, which we illustrated in the case of Simon's algorithm.

A more general way of looking at the quantum Fourier transform of an Abelian group $G = \mathbb{Z}_{N_1} \times \mathbb{Z}_{N_2} \times \cdots \times \mathbb{Z}_{N_l}$ is in terms of representation theory. The Abelian group G can be represented by homomorphisms ρ that maps G to the complex numbers. There are in fact $|G|$ such homomorphisms, in a natural one-to-one correspondence with the elements of G. For each $g = (a_1, a_2, \ldots, a_l) \in G$, define ρ_g to be the homomor-

phism that maps any $h = (b_1, b_2, \ldots, b_l) \in G$ according to

$$\rho_g(h) = \mathrm{e}^{2\pi i \sum_{i=1}^{l} \frac{a_i b_i}{N_i}} .$$

Using these definitions, we derive the following quantum Fourier transform maps:

$$|g\rangle \mapsto \frac{1}{\sqrt{|G|}} \sum_{h \in G} \rho_g(h) \,|\, h\rangle .$$

It is easy to verify that the quantum Fourier transform $\mathrm{QFT}_{N_1} \otimes \mathrm{QFT}_{N_2} \otimes \ldots \otimes \mathrm{QFT}_{N_l}$ maps a coset state of a subgroup K to a superposition of labels h where K is in the kernel of ρ_h, that is, $\rho_h(k) = 1$ for all $k \in K$. Thus, after applying the quantum Fourier transform, one will only measure labels $h = (b_1, b_2, \ldots, b_l)$ such that

$$\rho_g(h) = \mathrm{e}^{2\pi i \sum_{i=1}^{l} \frac{a_i b_i}{N_i}} = 1$$

which generalizes the notion of orthogonality defined in the explanation of Simon's algorithm in a very natural way, since it means $\sum_{i=1}^{l} \frac{a_i b_i}{N_i} = 0 \mod 1$. This gives us a linear equation that must be satisfied by all $h = (b_1, b_2, \ldots, b_l) \in K$, and thus after $l + O(1)$ such random equations are collected, we can efficiently find generators for K by solving a linear system of equations.

One can also define representations for finite non-Abelian groups G, except in order to fully capture the structure of G, we allow homomorphisms ρ to invertible matrices over $\mathbb{C}$ (in the Abelian case, we only need 1×1 matrices). For any finite group G, one can define a finite set of such homomorphisms $\rho_1, \rho_2, \ldots, \rho_k$ that map elements of G to unitary matrices of dimension $d_1 \times d_1$, $d_2 \times d_2, \ldots, d_k \times d_k$, respectively, with the property that $\sum_i d_i^2 = |G|$, and any other such representation is equivalent to a representation that can be factored into a collection of some number of these representations. More precisely, any representation $\rho\colon G \to M_{d\times d}(\mathbb{C})$ has the property that there exists some invertible $P \in M_{d\times d}(\mathbb{C})$, and a list of $\rho_{j_1}, \rho_{j_2}, \ldots$ such that $\sum_i d_i = d$ and for every $g \in G$

$$P\rho(g)P^{-1} = \oplus_i \rho_{j_i}(g)$$

(that is, $P\rho(g)P^{-1}$ is block-diagonal, with the matrices $\rho_{j_i}(g)$ along the diagonals). Furthermore, the representations ρ_j are irreducible in the sense that they cannot be decomposed into the sum of two or more smaller representations in this way. They are also unique up to conjugation.

Since $\sum_i d_i^2 = |G|$, one can define a unitary transformation from G to the vector space spanned by the labels of all the entries $\rho_i(j,k)$ of these irreducible representations ρ_i. In particular, we can map

$$|g\rangle \mapsto \sum_i \sqrt{\frac{d_i}{|G|}} \sum_{0 \le j,k \le d_i} \rho_i(j,k)(g) \,|\, i, j, k\rangle$$

where $\rho_i(j,k)(g)$ is the value of the (j,k) entry in the matrix $\rho_i(g)$. Such a mapping is unitary and is called a quantum Fourier transform for the group G. There is freedom in the specific choice of ρ_i within the set of unitary representations that are conjugate to ρ_i, so there is more than one quantum Fourier transform.

There has been much study of such quantum Fourier transforms for non-Abelian groups, which are sometimes possible to implement efficiently [21,96,140,151,156], but efficient constructions are not known in general. It appears they are of limited use in solving the non-Abelian hidden subgroup, except in special cases [68,87,96,156] such as when K is a normal subgroup of G.

In the next sections we discuss several other lines of attack on the non-Abelian hidden subgroup that have yielded some partial progress on the problem.

"Sieving" Kuperberg [130] introduced a method for attacking the hidden subgroup problem for the dihedral group that leads to a sub-exponential algorithm.

The dihedral group D_N is a non-Abelian group of order $2N$, which corresponds to the set of symmetries of a regular N-gon. It can be defined as the set of elements $\{(x,d) \mid x \in \{0,1,2,\ldots,N-1\}, d \in \{0,1\}\}$, where $\{(x,0) \mid x \in \{0,1,2,\ldots,N-1\}\}$ is the Abelian subgroup of D_N, corresponding to the rotations (satisfying $(x,0)+(y,0) = (x+y \mod N, 0)$), and $\{(0,0),(y,1)\}$ are Abelian subgroups of order 2 corresponding to reflections. In general, $(x,0)+(y,1) = (y-x,1)$, $(y,1)+(x,0) = (x-y,1) = -((x,0)+(y,1))$. If the hidden subgroup is a subgroup of the Abelian subgroup of order N, then finding the hidden subgroup easily reduces to the Abelian hidden subgroup problem. However, there is no known efficient algorithm for finding hidden subgroups of the form $\{(0,0),(y,1)\}$. So we can focus attention to the following restricted version of the dihedral hidden subgroup.

Dihedral Hidden Subgroup Problem (Hard Case)
Input: An integer N, and a black-box implementing $U_f\colon |x,d\rangle\,|0\rangle \mapsto |x,d\rangle\,|f(x,d)\rangle$, where $f(x_1,d_1) = f(x_2,d_1)$ if and only if $(x_1,d_1)-(x_2,d_2) \in \{(0,0),(y,1)\}$ for some y.
Problem: Find y.

As mentioned earlier, the dihedral hidden subgroup problem can be efficiently reduced to the following phase estimation problem [70].

Ettinger-Høyer Phase Estimation Problem for the Dihedral HSP
Input: An integer N, and a black-box O_d that outputs a classical value $k \in \{0, 1, \ldots, N-1\}$ uniformly at random, along with the qubit $1/\sqrt{2}(|0\rangle + e^{i\phi k}|1\rangle)$ where $\phi = (2\pi d)/N$, for some integer $d \in \{0, 1, 2, \ldots, N-1\}$.
Problem: Find d.

Note that if we could sample the values $k = 2^j$ for $j = 1, 2, \ldots, \lceil \log N \rceil$, then the phase estimation problem can be solved directly using the quantum Fourier transform [54].

Regev designed a clever method for generating states of the form $1/\sqrt{2}(|0\rangle + e^{i2^j\phi}|1\rangle)$ using $O(1)$ calls to the black-box O_d, given an oracle that solves the subset sum problem on average. Kuperberg [130] developed a "sieving" method of generating states of the form $1/\sqrt{2}(|0\rangle + e^{i2^j\phi}|1\rangle)$ and the method was refined and improved to use less memory by Regev [154].

Quantum Algorithms for the Dihedral HSP There exists a quantum algorithm that solves the dihedral HSP with running time in $e^{O(\sqrt{\log N})}$ and uses space in $e^{O(\sqrt{\log N})}$. There is also an algorithm with running time in $e^{O(\sqrt{\log N \log\log N})}$ and uses space in $O(\log N)$.

Classical Algorithms for the Dihedral HSP The classical complexity of the dihedral hidden subgroup problem is $\Theta(\sqrt{N})$ evaluations of the function f.

Similar sieving methods were applied [11] to yield a subexponential time algorithm for the HSP over the product groups G^n for a fixed non-Abelian group G. It has also been show that these kinds of quantum sieve algorithms will not give efficient quantum algorithms for graph isomorphism [141].

"Pretty Good Measurements" A natural approach to solving the non-Abelian hidden subgroup problem is to prepare several instances of a random coset state for the hidden subgroup K, and then try to determine what K is. More precisely, after preparing

$$\sum_{x\in G} |x\rangle |f(x)\rangle = \sum_{y+K\in G/K} |y+K\rangle |f(y)\rangle$$

and discarding the second register, we are left with the mixed state

$$\rho_K = \sum_{y+K\in G/K} |y+K\rangle\langle y+K| \ .$$

Thus one could try to implement or approximate the optimal quantum measurement for identifying the mixed states ρ_K, over all possible hidden subgroups $K \le G$. Furthermore, one could sample the state ρ_K a polynomial number of times t, and try to guess K given $\rho_K \otimes \rho_K \otimes \ldots \otimes \rho_K = \rho_K^t$.

Holevo [102] determined the optimal measurement for the following general state distinguishability problem. Given $\rho \in \{\rho_j\}$, output a label m such that the probability that $\rho = \rho_m$ is maximum. Let p_j denote the probability that $\rho = \rho_j$. Holevo proved that the maximum probability of guessing the correct input state ρ is achieved by a POVM with elements $\{G_j\}$ satisfying the following conditions: $\sum_i p_i\rho_i G_i = \sum_i p_i G_i \rho_i$ and $\sum_i p_i \rho_i G_i \ge p_j \rho_j$. However, it is not in general easy to efficiently find and implement such a POVM.

Hausladen and Wootters [99] defined a 'pretty good' measurement for distinguishing quantum states that is not necessarily optimal, but has a simpler form. The measurement used POVM elements $G_j = T^{-1/2}\rho_j T^{-1/2}$ and, if these don't sum to the identity, also $I - \sum_j G_j$, where $T = \sum_i \rho_i$. For the case of the dihedral hidden subgroup problem, it was shown [19] that the pretty good measurement is in fact optimal; however, in this case, it is still not known how to efficiently implement the pretty good measurement. However, it was later shown how to implement the pretty good measurement for the Heisengroup group HSP [19].

For example, in the case of the dihedral hidden subgroup problem for D_{2N}, after a quantum Fourier transform on each of n coset states, one can view the resulting state of n registers as

$$(|0\rangle + e^{2\pi i \frac{k_1 d}{N}}|1\rangle)|k_1\rangle \otimes (|0\rangle + e^{2\pi i \frac{k_2 d}{N}}|1\rangle)|k_2\rangle \otimes \ldots \otimes (|0\rangle + e^{2\pi i \frac{k_n d}{N}}|1\rangle)|k_n\rangle$$

for k_i selected independently and uniformly at random from $\{0, 1, \ldots, N-1\}$.

The $(|0\rangle + e^{2\pi i \frac{k_1 d}{N}}|1\rangle) \otimes (|0\rangle + e^{2\pi i \frac{k_2 d}{N}}|1\rangle) \otimes \ldots \otimes (|0\rangle + e^{2\pi i \frac{k_n d}{N}}|1\rangle)$ part of the state can be rewritten as

$$\sum_r \alpha_r |S_r\rangle$$

where r spans all the possible sums of the form $\sum_i b_i k_i$, $b_i \in \{0, 1\}$ (the 'subset sums'), $|S_r\rangle$ is the uniform superposition of the strings $|b_1 b_2 \ldots b_n\rangle$ that satisfy $\sum_{b_i k_i} = r$, and α_r is the appropriate normalization factor (i. e. $\sqrt{n_r/2^n}$ where n_r is the number of solutions to $\sum_{b_i k_i} = r$).

The optimal measurement [19] (in the restricted case of order two subgroups) can be thought of as mapping $|S_r\rangle \mapsto |r\rangle$, performing an inverse quantum Fourier transform and then measuring a value $\tilde{d}$, which will be the guess of the value d (interestingly, a similar measurement is optimal [61] in the case of trying to optimally estimate an arbitrary phase parameter $\phi \in [0, 2\pi)$ given the state $(|0\rangle + e^{ik_1\phi}|1\rangle) \times (|0\rangle + e^{i\phi k_2}|1\rangle) \times \ldots \times (|0\rangle + e^{i\phi k_n}|1\rangle))$. Note that implementing such a basis change in reverse would solve the subset sum problem (which is NP-hard). In fact, it suffices to solve the subset sum problem on average [154]. A nice discussion of this connection can be found in [19].

For groups that are semidirect products of an Abelian group and a cyclic group, the pretty good measurement corresponds to solving what is referred to [19] as a 'matrix sum problem', which naturally generalizes the subset sum problem. They also show that the pretty good measurements are optimal in these cases, and similarly relate their implementation to solving the matrix sum problem to certain average-case algebraic problems. They show that the pretty good measurement can be implemented for several groups, including semidirect product groups of the form $\mathbb{Z}_p^r \rtimes \mathbb{Z}_p$ for constant r (when $r = 2$, this is the Heisenberg group), and of the form $\mathbb{Z}_N \rtimes \mathbb{Z}_p$ with p prime (which are *metacyclic*) and where the ratio N/p is polynomial in $\log N$.

Other Methods and Results There are also some algorithms for solving other cases of the non-Abelian hidden subgroup problem that don't use any of the above techniques, e. g. [82,110,111]. These results use sophisticated classical and quantum group theoretic techniques to reduce an instance of a non-Abelian HSP to instances of HSP in Abelian groups.

One of the most recent results [112] shows that such a reduction is possible for *nil-2* groups, which are nilpotent groups of *class* 2. The group G is nilpotent of class n if the following holds[1]. Let $A_1 = G$, and let $A_{i+1} = [A_i, G]$, for $i > 0$. A group G is nilpotent if $A_{n+1} = \{1\}$, for some integer n, and the class of a nilpotent group is the smallest positive integer n for which $A_{n+1} = \{1\}$.

One of their techniques is to generalize Abelian HSP to a slightly more general problem, where the hidden subgroup K is hidden by a quantum procedure with the following properties. For every $g_1, g_2, \ldots, g_N \in G$ the algorithm maps

$$|g_1\rangle|g_2\rangle\ldots|g_N\rangle|0\rangle|0\rangle\ldots|0\rangle \mapsto |g_1\rangle|g_2\rangle \ldots |g_N\rangle \left|\psi_{g_1}^1\right\rangle \left|\psi_{g_2}^2\right\rangle \ldots \left|\psi_{g_N}^N\right\rangle$$

where the set of states $\{\left|\psi_g^i\right\rangle \mid g \in G\}$ is a *hiding set* for K, for each $i = 1, 2, \ldots, N$. A set of normalized states $\{\left|\psi_g\right\rangle \mid g \in G\}$ is a *hiding set* for the subgroup K of G if

- If g and h are in the same left coset of K then $|\psi_g\rangle = |\psi_h\rangle$.
- If g and h are in different left cosets of K then $\langle\psi_g|\psi_j\rangle = 0$.

Generators for the subgroup K of G can be found in time polynomial in $\log|G|$ using a quantum hiding procedure with $N \in O(\log|G|)$. They find a series of non-trivial reductions of the standard HSP in nil-2 groups to instances of the Abelian HSP with a quantum hiding function.

Lattice and Number Field Problems

The Abelian hidden subgroup problem also works for finitely generated groups G. We can, thus, define the hidden subgroup problem on $G = \mathbb{Z} \times \mathbb{Z} \times \ldots \times \mathbb{Z} = \mathbb{Z}^n$. The hidden subgroup K will be generated by some n-tuples in $\mathbb{Z}^n$. We can equivalently think of G as a lattice and K as a sublattice. The function $f\colon G \to X$, for some finite set X, that satisfies $f(\mathbf{x}) = f(\mathbf{y})$ if and only if $\mathbf{x} - \mathbf{y} \in K$, can be thought of as hiding the sublattice K.

By generalizing the problem to hiding sublattices of $\mathbb{R}^n$, one can solve some interesting and important number theoretic problems. The solutions in these cases were not a simple extension of the Abelian hidden subgroup algorithm.

Hallgren [93,95] found a quantum algorithm for finding the integer solutions x, y to Pell's equation $x^2 - dy^2 = 1$, for any fixed integer d. He also found an efficient quantum algorithm for the principal ideal problem, and later generalized it to computing the unit group of a number field of constant degree [94]. Solving Pell's equation is known to be at least as hard as factoring integers. We don't have room to introduce all the necessary definitions, but we'll sketch some of the main ideas.

A number field F can be defined as a subfield $Q(\theta)$ of the complex numbers that is generated by the rationals Q together with a root θ of an irreducible polynomial with rational coefficients; the degree of F is the degree of the irreducible polynomial. The "integers" of F are the elements of F that are roots of *monic* polynomials (i. e. polynomials with leading coefficient equal to 1, such as $x^2 + 5x + 1$). The integers of F form a ring, denoted $\mathcal{O}$.

[1] For any two elements g, h of a group, we define their commutator, denoted $[g, h]$ to be $[g, h] = g^{-1}h^{-1}gh$, and for any two subgroups $H, K \leq G$ we define $[H, K]$ to be the (normal) subgroup of G generated by all the commutators $[h, k]$ where $h \in H, k \in K$.

One can define a parameter Δ called the *discriminant* of F (we won't define it here), and an algorithm is considered efficient if its running time is polynomial in $\log \Delta$ and n. The unit group of the ring F, denoted $\mathcal{O}^*$, is the set of elements α in $\mathcal{O}$ that have an inverse $\alpha_{-1} \in \mathcal{O}$. "Computing" the unit group corresponds to finding a description of a system of "fundamental" units, $\epsilon_1, \epsilon_2, \ldots, \epsilon_r$ that generate $\mathcal{O}^*$ in the sense that every unit $\epsilon \in \mathcal{O}^*$ can be written as $\epsilon = \mu\epsilon_1^{k_1}\epsilon_2^{k_2}\ldots\epsilon_r^{k_r}$ for some $k_1, k_2, \ldots, k_r \in \mathbb{Z}$ and some root of unity μ. However, in general, an exact description of a fundamental unit requires an exponential number of bits. There are are some finite precision representations of the elements of the unit group, such as the "Log" embedding into $\mathbb{R}^r$. This representation describes a unit α by an element of $\mathbb{R}^r$ where some finite precision representation of each coordinate suffices. This Log representation of the unit group $\mathcal{O}^*$, corresponds to a sublattice $L = \mathcal{O}^*$ of $\mathbb{R}^r$. Hence, we have a relationship between several important computational number field problems, and lattice problems.

By the above correspondence between $\mathcal{O}^*$ and the lattice $L \subset \mathbb{R}^r$, we can formulate [94,162] the problem of computing the unit group as the problem of finding elements that approximate generators of the sublattice L of $\mathbb{R}^r$. One important non-trivial step is defining a function $f\colon \mathbb{R}^r \to X$ (for some infinite set X) such that $f(x) = f(y)$ if and only if $x - y \in L$ as well as appropriate discrete approximations to this function. The definition of these functions involves substantial knowledge of algebraic number theory, so we will not describe them here.

By designing quantum algorithms for approximating generators of the lattice L, [94,162] one can find polynomial time algorithms for computing the unit group $\mathcal{O}^*$ of an algebraic number field $F = \mathbb{Q}(\theta)$.

A corollary of this result, is a somewhat simpler solution to the principal ideal problem (in a constant degree number field) that had been found earlier by Hallgren [93]. An ideal $\mathcal{I}$ of the ring $\mathcal{O}$ is a subset of elements of $\mathcal{O}$ that is closed under addition, and is also closed under multiplication by elements of $\mathcal{O}$. An ideal $\mathcal{I}$ is *principal* if it can be generated by one element $\alpha \in \mathcal{I}$; in other words $\mathcal{I} = \alpha\mathcal{O} = \{\alpha\beta \mid \beta \in \mathcal{O}\}$. The principal ideal problem is, given generators for $\mathcal{I}$, to decide if $\mathcal{I}$ is principal, and if it is, to find a generator α.

As mentioned in Subsect. "Abelian Hidden Subgroup Problem", the tools developed can also be applied to find unique (quantum) representatives of elements of the class group for constant degree number fields [95] (assuming the generalized Riemann hypothesis), and thus allow for the computation of the class group in these cases.

Hidden Non-linear Structures

Another way to think of the Abelian hidden subgroup problem is as an algorithm for finding a hidden linear structure within a vector space. For simplicity, let's consider the Abelian HSP over the additive group $G = \mathbb{F}_q \times \mathbb{F}_q \times \ldots \times \mathbb{F}_q = \mathbb{F}_q^n$, where $q = p^m$ is a prime power. The elements of G can also be regarded as a vector space over $\mathbb{F}_q$. A hidden subgroup $H \leq G$ corresponds to a subspace of this vector space and its cosets correspond to parallel affine subspaces or *flats*. The function f is constant on these linear structures within the vector space $\mathbb{F}_q^n$.

A natural way to generalize this [51] is to consider functions that are constant on sets that are defined by non-linear equations. One problem they study is the *hidden radius problem*. The circle of radius $r \in \mathbb{F}_q$ centered at $t = (t_1, t_2, \ldots, t_n) \in \mathbb{F}_q^n$ is the set of points $x = (x_1, x_2, \ldots, x_n) \in \mathbb{F}_q^n$ that satisfy $\sum_i (x_i - t_i)^2 = r$. The point x on a circle centered at t will be represented as either $(x, \sigma(s))$, where $s = x - t$ and $\sigma(s)$ is a random permutation of s, or as $(x, \tau(t))$ where $\tau(t)$ is a random permutation of t. We define f_1 and f_{-1} be the functions satisfying $f_1(x, \sigma(s)) = \tau(t)$ and $f_{-1}(x, \tau(t)) = \sigma(s)$.

Hidden Radius Problem
Input: A black box U_{f_1} that implements $|x\rangle\,|\sigma(s)\rangle\,|0\rangle \mapsto |x\rangle\,|\sigma(s)\rangle\,|f_1(x, \sigma(s))\rangle$, and a black box $U_{f_{-1}}$ that implements $|x\rangle\,|\tau(t)\rangle\,|0\rangle \mapsto |x\rangle\,|\tau(t)\rangle\,|f_{-1}(x, \tau(t))\rangle$.
Problem: Find r.

Quantum Algorithms for the Hidden Radius Problem For odd d, there is a bounded error quantum algorithm that makes polylog(q) queries to U_{f_1} and $U_{f_{-1}}$ and finds r. However, there is no known polynomial bound on the non-query operations.

There is a bounded-error quantum algorithm that also makes polylog(q) operations in total, and determines $\chi(r)$, where $\chi(r) = 1$ if r is a quadratic residue (that is, $r = u^2$ for some $u \in \mathbb{F}_q$) and 0 otherwise.

Classical Algorithms for the Hidden Radius Problem It was shown in [51] that the expected number of queries needed to be able to guess any bit of information about r correctly with probability greater than $\frac{1}{2} + \frac{1}{\mathrm{poly}(d \log q)}$ is exponential in $d \log q$.

A number of other black-box problems of this kind were defined in [51] with quantum algorithms that are exponentially more efficient than any classical algorithm, in some cases just in terms of query complexity, and other times in terms of all operations. These problems fit into the frameworks of *shifted subset* problems and *hidden polynomial* problems. They use a variety of non-trivial tech-

niques for these various problems, including the quantum Fourier transform, quantum walks on graphs, and make some non-trivial connections to various *Kloosterman sums*. Further work along these lines has been done in [63].

Hidden Shifts and Translations

There have been a variety of generalizations of the hidden subgroup problem to the problem of finding some sort of hidden shift or translation.

Grigoriev [89] addressed the problem of the shift-equivalence of two polynomials (the self-shift-equivalence problem is a special case of the Abelian hidden subgroup problem). Given two polynomials P_1, P_2 in l variables $X_1, X_2, \ldots, X_l$ over $\mathbb{F}_q$ (the finite field with q elements), does there exist an element $(a_1, a_2, \ldots, a_l) \in \mathbb{F}_q^l$ such that $P_1(X_1 - a_1, X_2 - a_2, \ldots, X_l - a_l) = P_2(X_1, X_2, \ldots, X_l)$. More generally, if there is a group G ($\mathbb{F}_q^l$ in this case) acting on a set X (the set of polynomials in l variables over $\mathbb{F}_q$ in this case), one can ask if two elements $x, y \in X$ are in the same orbit of the action of G on X (that is, if there is a $g \in G$ such that $g(x) = y$). In general, this seems like a hard problem, even for a quantum computer.

The dihedral hidden subgroup problem [70] is a special case of the *hidden translation problem* [82], where there is a finite group G, with unique representation of the elements of G, and two injective functions f_0 and f_1 from G to some finite set X.

Hidden Translation Problem:
Input: Two black boxes U_{f_0} and U_{f_1} that, for any $x \in G$, implement the maps $U_{f_0}\colon |x\rangle|0\rangle \mapsto |x\rangle|f_0(x)\rangle$ and $U_{f_1}\colon |x\rangle|0\rangle \mapsto |x\rangle|f_1(x)\rangle$.
A promise that $f_1(x) = f_0(ux)$, for some $u \in G$.
Problem: Find u.

The same problem expressed with additive group notation has been called the *hidden shift problem*, and instances of this problem were solved efficiently by van Dam, Hallgren and Ip [59]. For example, they find an efficient solution in the case that $f_1(x) = \chi(x + s)$ where $f_0 = \chi$ is a multiplicative character function over a finite field, which implies a method for breaking a class of "algebraically homomorphic" cryptosystems. They also describe a more general *hidden coset problem*.

In [82] it is shown how to solve the hidden translation problem in $G = \mathbb{Z}_p^n$ in polynomial time, and then show how to use this to solve the problem for any group that they call "smoothly solvable". Let us briefly define what this means.

The *derived subgroup* of a group G is $G^{(1)} = [G, G]$. In general, we define $G^{(0)} = G, G^{(n+1)} = [G^{(n)}, G^{(n)}]$, for $n \geq 1$. A group G is solvable if $G^{(n)} = \{1\}$, the trivial group, for some positive integer n, and the series of subgroups is called the *derived series* of G. A group G is called smoothly solvable if m is bounded above by a constant, and if the factor groups $G^{(j+1)}/G^{(j)}$ are isomorphic to a direct product of a group of bounded exponent (the exponent of a group is the smallest positive integer r such that $g^r = 1$ for all g in the group) and a group of size polynomial in $\log|G|$.

The algorithm for smoothly solvable groups works by solving a more general *orbit coset* problem, for which they prove a "self-reducibility" property. In particular, orbit coset problem for a finite group G is reducible to the orbit coset problem in G/N and N, for any solvable normal subgroup N of G.

Quantum Algorithms for the Hidden Translation Problem: For groups $G = \mathbb{Z}_p^n$, the hidden translation problem can be solved with bounded probability using $O(p(n+p)^{p-1})$ queries and $(n+p)^{O(p)}$ other elementary operations.

In general, for *smoothly solvable* groups G, the hidden translation problem can also be solved with a polynomial number of queries and other elementary operations. Another consequence of the tools they developed is a polynomial time solution to the hidden subgroup for such smoothly solvable groups.

Classical Algorithms for the Hidden Translation Problem: In general, including the case $G = \mathbb{Z}_p^n$, the hidden translation problem requires $\Omega(\sqrt{|G|})$ queries on a classical computer.

Another natural generalization of the hidden translation or hidden shift problem and the Abelian hidden subgroup problem is the *generalized hidden shift problem* introduced in [46]. There is a function $f\colon \{0, 1, 2, \ldots, M-1\} \times \mathbb{Z}_N \to X$ for some finite set X, with the property that for a fixed $b \in \{0, 1, \ldots, M-1\}$, the mapping $x \mapsto f(b, x)$ is one-to-one, and there is some hidden value $s \in \mathbb{Z}_N$ such that $f(b, x) = f(b+1, x+s)$ for all $b \in \{0, 1, \ldots, M-2\}$. Note that for $M = 2$, this is equivalent to the dihedral hidden subgroup problem for the group D_N, and for $M = N$, this problem is equivalent to the Abelian hidden subgroup problem for the hidden subgroup $\langle(1, s)\rangle$ of the group $\mathbb{Z}_N \times \mathbb{Z}_N$.

Generalized Hidden Shift Problem:
Input: Positive integers M and N.
A black boxes U_f that maps $|b, x\rangle|0\rangle \mapsto |b, x\rangle|f(b, x)\rangle$ for all $b \in \{0, 1, \ldots, M-1\}$ and $x \in \mathbb{Z}_N$, where f satisfies the properties defined above.
Problem: Find s.

Quantum Algorithms for the Generalized Hidden Shift Problem: There is a quantum algorithm that, for any fixed $\epsilon > 0$, and $M \geq N^{\epsilon}$, solves the generalized hidden shift problem in time polynomial in $\log N$.

The algorithm uses a "pretty good measurement" that involves solving instances of the following matrix sum problem. Given $\mathbf{x} \in \mathbb{Z}_N^k$ and $w \in Z_N$ chosen uniformly at random, find $\mathbf{b} \in \{0, 1, \ldots, M-1\}^k$ such that $\sum b_i x_i = w \mod N$. Note how this generalizes the subset sum problem, which was shown to be related to the dihedral hidden subgroup problem [154]. While there is no efficient solution known for small M (even an average case solution suffices for the dihedral HSP), for $M \geq N^{\epsilon}$, Lenstra's integer programming algorithm allows for an efficient solution to the matrix sum problem.

Classical Algorithms for the Generalized Hidden Shift Problem: Any classical algorithm requires $\Omega(\sqrt{N})$ evaluations of the function f.

Other Related Algorithms

There are a variety of other problems that aren't (as far as we know) generalizations of the hidden subgroup problem, and arguably deserve a separate section. We'll mention them here since various parts of the algorithms for these problems use techniques related to those discussed in one of the other subsections of this section.

Van Dam and Seroussi [56] give an efficient quantum algorithm for estimating Gauss sums. Consider a finite field $\mathbb{F}_{p^r}$ (where p is prime, and r is a positive integer). The multiplicative characters are homomorphisms of the multiplicative group, $\mathbb{F}_{p^r}^*$, to the complex numbers $\mathbb{C}$, and also map $0 \mapsto 0$. Each multiplicative character can be specified by an integer $\alpha \in \{0, 1, \ldots, p^r - 2\}$ by defining $\chi_\alpha(g^j) = \xi^{\alpha j}$, where g is a generator for $\mathbb{F}_{p^r}^*$ and $\xi = e^{2\pi i/(p^r-1)}$.

The additive characters are homomorphisms of the additive group of the field to $\mathbb{C}$, and can be specified by a value $\beta \in \mathbb{F}_{p^r}$ according to $e_\beta(x) = \zeta^{Tr(\beta x)}$, where $Tr(y) = \sum_{j=0}^{r-1} y^{p^j}$ and $\zeta = e^{2\pi i/p}$.

The Gauss sum $G(\mathbb{F}_{p^r}, \chi_\alpha, e_\beta)$ is defined as

$$G(\mathbb{F}_{p^r}, \chi, e_\beta) = \sum_{x \in \mathbb{F}_{p^r}} \chi(x) e_\beta(x) .$$

It is known that the norm of the Gauss sum is $|G(\mathbb{F}_{p^r}, \chi, e_\beta)| = \sqrt{p^r}$, and thus the hard part is determining, or approximating, the parameter γ in the equation $G(\mathbb{F}_{p^r}, \chi, e_\beta) = e^{i\gamma}\sqrt{p^r}$.

Gauss Sum Problem for Finite Fields:
Input: A prime number p, positive integer r and a standard specification of $\mathbb{F}_{p^r}$ (including a generator g).
A positive integer $\alpha \in \{0, 1, \ldots, p^r - 2\}$.
An element $\beta \in \mathbb{F}_{p^r}$.
A parameter ϵ, $0 < \epsilon < 1$.
Problem: Output an approximation, with error at most ϵ, to γ in the equation $G(\mathbb{F}_{p^r}, \chi_\alpha, e_\beta) = e^{i\gamma}\sqrt{p^r}$.

One noteworthy feature of this problem is that it is not a black-box problem.

Quantum Algorithms for the Finite Field Gauss Sum Problem: There is a quantum algorithm running in time $O(\frac{1}{\epsilon}\text{polylog}(p^r))$ that outputs a value $\tilde{\gamma}$ such that $|\gamma - \tilde{\gamma}| < \epsilon$ with probability at least $\frac{2}{3}$.

Classical Complexity of the Finite Field Gauss Sum Problem: It was shown that solving this problem is at least as hard as the discrete logarithm problem in the multiplicative group of $\mathbb{F}_{p^r}$ (see Subsect. "Discrete Logarithms").

Various generalizations of this problem were also studied in [56]. Other examples include [57] which studies the problem of finding solutions to equations of the form $af^x + bg^y = c$, where a, b, c, f, g are elements of a finite field, and x, y are integers.

Quantum Walk Algorithms

Quantum walks, sometimes called quantum random walks, are quantum analogues of (classical) random walks, which have proved to be a very powerful algorithmic tool in classical computer science. The quantum walk paradigm is still being developed. For example, the relationship between the continuous time and discrete time models of quantum walks is still not fully understood. In any case, the best known algorithms for several problems are some type of quantum walk.

Here we restrict attention to walks on discrete state spaces. Because of the quantum strong Church–Turing thesis, we expect that any practical application of a walk on a continuous state space will have an efficient (up to polynomial factors) simulation on a discrete system.

In general, any walk algorithm (classical or quantum), consists of a discrete state space, which is usually finite in size, but sometimes infinite state spaces are also considered when it is convenient to do so. The state space is usually modeled as being the vertices of a graph G, and the edges of the graph denote the allowed transitions. In classical discrete time walks, the system starts in some initial state, v_i. Then at every time step the system moves to a random neighbor w of the current vertex v, according

to some probability distribution $p(v, w)$. Let M denote the matrix where the (v, w) entry is $p(v, w)$. Let $\mathbf{v_0}$ be the column vector with the value p_i in the ith position, where p_i is the probability that the initial vertex is v_i. Then the vector $\mathbf{v_t} = M^t \mathbf{v_0}$ describes the probability distribution of the system after t time steps after starting in a state described by the probability distribution $\mathbf{v_0}$.

The walks are usually analyzed as abstract walks on a graph. In practice, the vertices are representing more sophisticated objects. For example, suppose one wishes to solve a 3-SAT formula Φ on n Boolean variables. One could define a random walk on the 2^n possible assignments of the Boolean variables. So the vertices of the graph would represent the 2^n Boolean strings of length n. One could start the walk on a random vertex (which corresponds to a random assignment of the n-Boolean variables). At every step of the walk, if the current vertex v corresponds to a satisfying assignment, then $p(v, v) = 1$ and the walk should not leave the vertex. Otherwise, a random clause should be picked, and one of the variables in that clause should be picked uniformly at random and flipped. This implicitly defines a probability distribution $p(v, w)$.

In a quantum walk, instead of just having classical probability distributions of the vertices $v_i \in V(G)$, one can have superpositions $\sum_{v_i \in V(G)} \alpha_i \mid v_i\rangle$, and more generally any quantum mixed state of the vertices. If we restrict to unitary transitions, then there is a unitary matrix U that contains the transition amplitudes $\alpha(v, w)$ of going from vertex v to vertex w, and if the systems starts in initial state $\mid \psi_0\rangle$, then after t time steps the state of the system is $U^t \mid \psi_0\rangle$. These unitary walks are not really "random" since the evolution is deterministic. More generally, the transition function could be a completely positive map $\mathcal{E}$, and if the system starts in the initial state $\rho = \mid \psi_0\rangle \langle \psi_0 \mid$, then after t time steps the state of the system will be $\mathcal{E}^t(\rho)$.

One cannot in general define a unitary walk on any graph [164]; however if one explicitly adds a "coin" system of dimension as a large as the maximum degree d of the vertices (i. e. the new state space consists of the states $\mid v_i\rangle \mid c\rangle$, $v_i \in V(G)$ and $c \in \{0, 1, \ldots, d-1\}$) then one can define a unitary walk on the new graph one would derive from the combined graph-coin system. In particular, the state of the coin system indicates to which neighbor of a given vertex the system should evolve. More generally, one can define a unitary walk on states of the form (v_i, v_j), where $\{v_i, v_j\}$ is an edge of G.

A continuous version of quantum walks was introduced by Farhi and Gutmann [72]. The idea is to let the adjacency matrix of the graph be the Hamiltonian driving the evolution of the system. Since the adjacency matrix is Hermitian, the resulting evolution will be unitary. The reason such a unitary is possible even for a graph where there is no unitary discrete time evolution is that in this continuous time Hamiltonian model, for any non-zero time evolution, there is some amplitude with which the walk has taken more than one step.

In classical random walks, one is often concerned with the "mixing time", which is the time it takes for the system to reach its equilibrium distribution. In a purely unitary (and thus reversible) walk, the system never reaches equilibrium, but there are alternative ways of arriving at an effective mixing time (e. g. averaging over time). In general, quantum walks offer at most a quadratically faster mixing. Another property of random walks is the "hitting time", which is the time it takes to reach some vertex of interest. There are examples where quantum walks offer exponentially faster hitting times.

The study of what are essentially quantum walks has been around for decades, and the algorithmic applications have been developed for roughly 10 years. Much of the early algorithmic work developed the paradigm and discovered the properties of quantum walks on abstract graphs, such as the line or circle, and also on general graphs (e. g. [6,17]). There have also been applications to more concrete computational problems, and we will outline some of them here.

Element Distinctness Problem
Input: A black-box U_f that maps $\mid i\rangle \mid b\rangle \mapsto \mid i\rangle \mid b \oplus f(i)\rangle$ for some function $f\colon \{0, 1, \ldots, N-1\} \to \{0, 1, \ldots, M\}$.
Problem: Decide whether there exist inputs i and j, $i \neq j$, such that $f(i) = f(j)$.

Prior to the quantum walk algorithm of Ambainis, the best known quantum algorithm used $O(N^{\frac{3}{4}} \log N)$ queries [42].

Quantum Algorithms for Element Distinctness Problem The quantum walk algorithm in [15] uses $O(N^{2/3})$ evaluations of U_f, $O(N^{2/3}\text{polylog}N)$ non-query operations and $O(N^{2/3}\text{polylog}N)$ space.

Classical Algorithms for Element Distinctness A classical computer requires $N - O(1)$ applications of U_f in order to guess correctly with bounded error for worst-case instances of f.

As is often the case with classical random walk algorithms, the graph is only defined implicitly, and is usually exponentially large in the size of the problem instance. For the element distinctness algorithm, the graph is defined as follows. The vertices are subsets of $\{1, 2, \ldots, N\}$ of size $\lceil N^{2/3} \rceil$. Two vertices are joined if the subsets differ in exactly two elements. A detailed description and analysis of this walk is beyond the scope of this survey.

Szegedy [171] extended the approach of Ambainis to develop a powerful general framework for quantizing classical random walks in order to solve search problems. Suppose we wish to search a solution space of size N and there are ϵN solutions to $f(x) = 1$. Furthermore, suppose there is a classical random walk with transition matrix M, with the property that $p(v, w) = p(w, v)$ (known as a 'symmetric' walk). It can be shown that the matrix M has maximum eigenvalue 1, and suppose the next highest eigenvalue is $1 - \delta$, for $\delta > 0$. The classical theory of random walks implies the existence of a bounded-error classical random walk search algorithm with query complexity in $O(\frac{1}{\delta\epsilon})$. Szegedy developed a "$\sqrt{\delta\epsilon}$-rule" that gives a quantum version of the classical walk with query complexity in $O(1/\sqrt{\delta\epsilon})$. This technique was generalized further in [134] and summarized nicely in [160].

Quantum walk searching has been applied to other problems such as triangle-finding [135], commutativity testing [133], matrix product verification [40], associativity testing when the range is restricted [66], and element k-distinctness [15]. A survey of results in quantum walks can be found in [13,14,123,160].

Continuous Time Quantum Walk Algorithms

In this section we describe two very well-known continuous time quantum walk algorithms. The first algorithm [48] illustrates how a quantum walk algorithm can give an exponential speed-up in the black-box model.

A problem instance of size n corresponds to an oracle O_{G_n} that encodes a graph G_n on $O(2^n)$ vertices in the following way. The graph G_n is the graph formed by taking 2 binary trees of depth n, and then "gluing" the two trees together by adding edges that create a cycle that alternates between the leaves of the first tree (selected at random) and the leaves of the second tree (selected at random). The two root vertices are called the "ENTRANCE" vertex, labeled with some known string, say, the all zeroes string $000\ldots0$ of length $2n$, and "EXIT" vertex, which is labeled with a random string of length $2n$. The remaining vertices of the graph are labeled with distinct random bit strings of length $2n$. The oracle O_{G_n} encodes the graph G_n in the following way. For $|x\rangle$ where $x \in \{0,1\}^{2n}$ encodes a vertex label of G_n, O_{G_n} maps $|x\rangle|00\ldots0\rangle$ to $|x\rangle|n_1(x), n_2(x), n_3(x)\rangle$ where $n_1(x), n_2(x), n_3(x)$ are the labels of the neighbors of x in any order (for the exit and entrance vertex, there will only be two distinct neighbors).

"Glued-Trees" Problem
Input: A black-box implementing O_{G_n} for a graph G_n of the above form.
Problem: Output the label of the EXIT vertex.

Quantum Algorithms for the "Glued-Trees" Problem There is a continuous time quantum walk which starts at the ENTRANCE vertex (in this case $|00\ldots0\rangle$) and evolves according to the Hamiltonian defined by the adjacency matrix of the graph G_n for an amount of time t selected uniformly at random in $[0, n^4/(2\epsilon)]$ where $0 < \epsilon < 1$. Measuring will then yield the EXIT label with probability at least $\frac{(1-\epsilon)}{2n}$.

The authors show how to efficiently simulate this continuous time quantum walk using a universal quantum computer that makes a polynomial number of calls to O_{G_n}.

Classical Algorithms for the "Glued-Trees" Problem Any classical randomized algorithm must evaluate the black-box O_{G_n} an exponential number of times in order to output the correct EXIT vertex label with non-negligible probability. More precisely, any classical algorithm that makes $2^{n/6}$ queries to O_{G_n} can only find the EXIT with probability at most $4 \cdot 2^{-n/6}$.

Another very interesting and recent problem for which a quantum walk algorithm has given the optimal algorithm is the problem of evaluating a NAND-tree (or AND-OR tree). The problem is nicely described by a binary tree of depth n whose leaves are labeled by the integers $i \in \{1, 2, \ldots, 2^n\}$. The input is a black-box $O_{\mathbf{X}}$ that encodes a binary string $\mathbf{X} = X_1X_2\ldots X_N$, where $N = 2^n$. The ith leaf vertex is assigned value X_i, and the parent of any pair of vertices takes on the value which is the NAND of the value of its child vertices (the NAND of two input bits is 0 if both inputs are 1 and 1 if either bit is 0). Thus, given the assignment of values to the leaves of a binary tree, one can compute the values of the remaining vertices in the tree, including the root vertex. The value of the NAND tree for a given assignment is the value of the root vertex.

NAND-Tree Evaluation
Input: A black-box $O_{\mathbf{X}}$ that encodes a binary string $\mathbf{X} = X_1X_2\ldots X_N \in \{0,1\}^N$, $N = 2^n$.
Problem: Output the value of the binary NAND tree whose ith leaf has value X_i.

Classical Algorithms for NAND-Tree Evaluation The best known classical randomized algorithm uses $O(N^{0.753\cdots})$ evaluations of the black-box, and it is also known that $\Omega(N^{0.753\cdots})$ evaluations are required for any classical randomized algorithm.

Until recently, no quantum algorithm worked better.

Quantum Algorithms for NAND-Tree Evaluation Farhi, Goldstone and Gutmann [76] showed a continuous time walk that could solve this it time $O(\sqrt{N})$ using a continuous version of the black-box, and it was subsequently

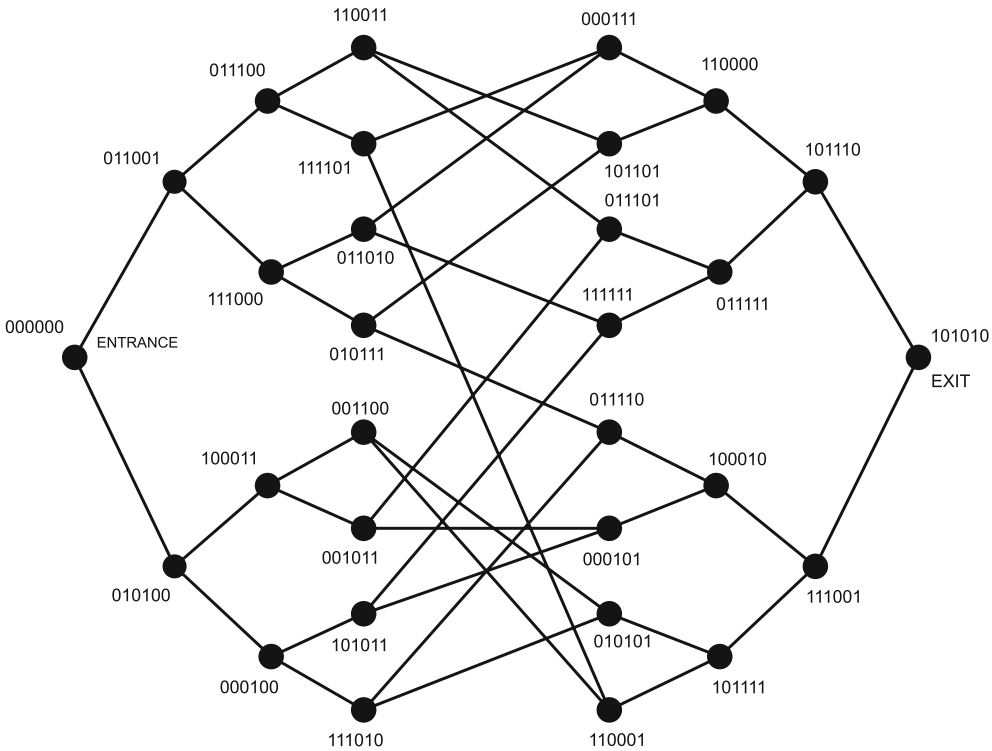

Quantum Algorithms, Figure 1
This is an example of a "glued-trees" graph with random labelings for all vertices (except the "ENTRANCE" vertex). The goal is to find the label of the "EXIT" vertex (in this case, it is 101010), given a black-box that outputs the vertex labels of the neighbors of a given vertex label

showed that $O(N^{1/2+\epsilon})$ queries to the discrete oracle suffice, for any real constant $\epsilon > 0$, and discrete walk versions of the algorithm and other generalizations were developed [18,50]

This was a very interesting breakthrough in solving a fundamental problem that had stumped quantum algorithms experts for a number of years. The general idea is inspired from techniques in particle physics and scattering theory. They consider a graph formed by taking a binary tree and making two additions to it. Firstly, for each leaf vertex where $X_i = 1$, add another vertex and join it to that leaf. Secondly, attach the root vertex to the middle of a long path of length in $\Omega(\sqrt{N})$. Then evolve the system according to the Hamiltonian equal to the adjacency matrix of this graph. Then one should start the system in a superposition of states on the left side of the line graph with phases defined so that if the NAND-tree were not attached, the "wave packet" would move from left to right along the line. If the packet gets reflected with non-negligible amplitude, then the NAND tree has value 1, otherwise, if the packet gets mostly transmitted, then the NAND tree has value 0. Thus one measures the system, and if one obtains a vertex to the left of the NAND-tree, one guesses "1", and if one obtains a vertex to the right of the NAND-tree, one guesses "0". This algorithm outputs the correct answer with high probability.

In a discrete query model, one can carefully simulate the continuous time walk [50] (as discussed in Sect. "Simulation of Quantum Mechanical Systems"), or one can apply the results of Szegedy [171] to define a discrete-time coined walk with the same spectral properties [18] and thus obtain a discrete query complexity of $N^{\frac{1}{2}+\epsilon}$ for any constant $\epsilon > 0$.

This algorithm has also been applied to solve MIN–MAX trees with a similar improvement [55]. The NAND-tree and related problems are related to deciding the winner of two-player games. Another very recent new class of quantum algorithms for evaluating a wider class of formulas, based on "span" programs, was developed in [155].

Adiabatic Algorithms

It is possible to encode the solution to a hard problem into the ground state of an efficiently simulatable Hamiltonian.

For example, in order to try to solve 3-SAT for a formula on n Boolean variables, one could define a Hamiltonian

$$H_1 = \sum_{\mathbf{x}\in\{0,1\}^n} f_\Phi(\mathbf{x}) \,|\, \mathbf{x}\rangle \langle \mathbf{x}|$$

where $f_\Phi(\mathbf{x})$ is the number of clauses of Φ that are violated by the assignment $\mathbf{x}$. Then one could try to define algorithms to find such a ground state, such as quantum analogues of classical annealing or other heuristics (e. g. [62,74,101]).

Adiabatic algorithms (also known as adiabatic optimization algorithms) are a new paradigm for quantum algorithms invented by Farhi, Goldstone, Gutmann and Sipser [74]. The paradigm is based on the fact that, under the right conditions, a system that starts in the ground state of a Hamiltonian $H(0)$, will with high probability remain in the ground state of the Hamiltonian $H(t)$ of the system at a later time t, if the Hamiltonian of the system changes "slowly enough" from $H(0)$ to $H(t)$. This fact is called the adiabatic theorem (see e. g. [115]).

This theorem inspires the following algorithmic paradigm:

- Convert your problem to generating the ground state of some easy-to-simulate Hamiltonian H_1.
- Initialize your quantum computer in an easy-to-prepare state $|\psi_0\rangle$ of an easy-to-simulate Hamiltonian H_0.
- On the quantum computer, simulate a time-dependent Hamiltonian $H(t) = (1 - t/T)H_0 + t/TH_1$, for t going from 0 to T.

An important detail is how slowly to transition from H_0 to H_1 (in other words, how large T should be). This related to two important parameters. Let $\lambda_0(t)$ be the smallest eigenvalue of $H(t)$, and assume that the corresponding eigenspace is non-degenerate. Let $\lambda_1(t)$ be the second smallest eigenvalue of $H(t)$, and define $g(t) = \lambda_1(t) - \lambda_0(t)$ to be the gap between the two lowest eigenvalues. The norm of the Hamiltonian is also relevant. This is to be expected, since one can effectively speed-up time by a factor of s by just multiplying the Hamiltonian of the system by s. In any realistic implementation of the Hamiltonian one pays for such a speed-up by at least a factor of s in some resource. For example, if we simulate the Hamiltonian using quantum circuits (e. g. as described in Sect. "Simulation of Quantum Mechanical Systems"), the overhead is a factor of $s^{1+o(1)}$ in the circuit depth; thus there is no actual speed-up. Furthermore, the norm of the derivatives of the Hamiltonian is also relevant.

There are a variety of theorems and claims in the literature proving, or arguing, that a value of T polynomial in the operator norm of $\mathrm{d}H(t)/\mathrm{d}t$ (or even some higher derivative) and in the inverse of the minimum gap (i. e. the minimum $g(t)$, for $0 \le t \le T$), and in $\frac{1}{\delta}$, is sufficient in order to generate the final ground state with probability at least $1 - \delta$. The general folklore is that with the right assumptions the dependence on the minimum gap $g_{\min}$ is $\Omega(1/g_{\min}^2)$, and one can find examples when this is the case; however more sophisticated descriptions of the dependence on $g(t)$ are known (see e. g. [115]).

Assuming there is exactly one satisfying assignment $\mathbf{w}$, then $|\mathbf{w}\rangle\langle\mathbf{w}|$ is the unique ground state of H_1. This algorithm has been studied numerically and analytically [58,75,157] and variations have been introduced as well to work around various lower bounds that were proved [77]. Unfortunately, it is not known what the worst-case complexity is for these algorithms on such NP-hard problems, since it has proved very hard to provide rigorous or even convincing heuristic bounds on the minimum gap for such problems of interest. It is widely believed that these algorithms will not solve an NP-hard problem in worst-case polynomial time, partly because it is believed that no quantum algorithm can do this.

A slight generalization of adiabatic algorithms, called adiabatic computation (where the final Hamiltonian does not need to be diagonal in the computational basis) was shown to be polynomially equivalent to general quantum computation [7].

Topological Algorithms

The standard models of quantum computation (e. g. quantum Turing machine, quantum acyclic circuits) are known to be equivalent in power (up to a polynomial factor), and the quantum strong Church–Turing thesis states that any realistic model of computation can be efficiently simulated by such a quantum computer. If this were not the case, then one should seek to define a stronger model of computation that encapsulates the full computational power that the laws of physics offer.

Freedman [80] proposed defining a computing model based on topological quantum field theories. The main objective was that such a computer might naturally solve an NP-hard or #P-hard topological problem, in particular, evaluating the Jones polynomial at certain points. A natural family of such topological quantum field theory computers was shown to be equivalent in power to the standard model of quantum computation, thus such a new model of computation would not provide additional computational power, but it was hoped that this new paradigm might inspire new quantum algorithms. In fact, it has been shown that "topological" algorithms can approximate the value of the Jones polynomial at cer-

tain points more efficiently than any known classical algorithm [8,81]. The known approximations are not good enough to solve an NP-hard problem. Several other generalizations and related problems and algorithms have been found recently [10,85,132,179]. We will briefly sketch some of the definitions, results and techniques.

A *knot* is a closed non-intersecting curve embedded in $\mathbf{R}^3$, usually represented via a knot diagram, which is a projection of the knot into the plane with additional information at each cross-over to indicate which strand goes over and which goes under. Two knots are considered equivalent if one can be manipulated into the other by an isotopy (i. e. by a transformations one could make to an actual knot that can be moved and stretched but not broken or passed through itself). A link is a collection of non-intersecting knots embedded in $\mathbf{R}^3$. They can be represented by similar diagrams, and there is a similar notion of equivalence.

The Jones polynomial of a link L is a polynomial $V_L(t)$ that is a link invariant; in other words it has the property that if two links L_1 and L_2 are equivalent, then $V_{L_1}(t) = V_{L_2}(t)$. Computing the Jones polynomial is in general #P-hard for all but a finite number of values of t. In particular, it is #P-hard to evaluate the Jones polynomial exactly at any primitive rth root of unity for any integer $r \geq 5$. However, certain approximations of these values are not known to be #P-hard.

The Jones polynomial is a special case of the Tutte polynomial of a planar graph. For a planar graph $G = (V, E)$, with weights edge weights $\mathbf{v} = \{v_e \mid e \in E\}$ the multivariate Tutte polynomial is defined as

$$Z_G(q; v_{e_1}, v_{e_2}, \ldots) = \sum_{A \subseteq E} q^{k(A)} \prod_{e \in A} v_e$$

where q is another variable and $k(A)$ is the number of connected components in the subgraph (V, A). The standard Tutte polynomial $T_G(x, y)$ is obtained by setting $v_e = v$ for all $e \in E$, $x = 1 + q/v$ and $y = 1 + v$. Connections with physics are discussed, for example, in [120,169,176].

Here we briefly sketch a specific instance of such a problem, and the approach of [10] taken to solve this problem on a quantum computer.

Firstly, for any planar graph G, one can efficiently find its *medial graph* L_G, which is a 4-regular planar graph which can be drawn from a planar embedding of G as follows. Draw a new vertex with weight u_i in the middle of each edge that had label v_i. For each new vertex, on each side of the original edge on which the vertex is placed, draw a new edge going in the clockwise direction joining the new vertex to the next new vertex encountered along the face of the original graph. Do the same in the counter-clockwise direction, and remove all the original edges and vertices.

From this medial graph, one can define another polynomial called the *Kauffman bracket*, denoted $\langle L_G \rangle(d, u_1, u_2, \ldots)$, that satisfies $\langle L_G \rangle(d, u_1, u_2, \ldots) = d^{-|V|} Z_G(d^2; du_1, du_2 \ldots)$. The next section sketches a quantum algorithm that approximates the Kauffman bracket.

Additive Approximation of the Multivariate Tutte Polynomial for a Planar Graph

Input: A description of a planar graph $G = (V, E)$.

Complex valued weights $v_1, v_2, \ldots, v_m$ corresponding to the edges $e_1, e_2, \ldots, e_m \in E$ of G.

A complex number q.

Problem: Output an approximation of $Z_G(q; v_1, v_2, \ldots, v_m)$.

Quantum Algorithms for Approximating the Tutte Polynomial Aharonov et al. [10] give a quantum algorithm that solves the above problem in time polynomial in n with an additive approximate they denote by $\Delta_{\text{alg}}/\text{poly}(m)$, which is described below.

The value Δ_{alg} depends on the embedding of the graph. The results are hard to compare to what his known about classical algorithms for this problem (see [33] for a discussion), but there are special cases that are BQP-hard.

Classical Algorithms for Approximating the Tutte Polynomial It was shown [10] that for certain ranges of parameter choices, the approximations given by the quantum algorithms are BQP-hard, and thus we don't expect a classical algorithm to be able to provide as good of an approximation unless classical computers can efficiently simulate quantum computers.

Sketch of the Structure of the Algorithm

We only have room to give a broad overview of the algorithm. One of the main points is to emphasize that this algorithm looks nothing any of the other algorithms discussed in the previous sections.

At a very high level, the idea is that these medial graphs T_G can be represented by a linear operation Q_G such that $\langle L_G \rangle(d, u_1, u_2, \ldots) = \langle 1 | Q_G | 1 \rangle$. The quantum algorithm approximates the inner product between $|1\rangle$ and $Q_G |1\rangle$, and therefore gives an approximation to the Kauffman bracket for G and thus the generalized Tutte polynomial for G.

The medial graph L_G will be represented as a product of basic elements $\mathcal{T}_i$ from a generalized Temperley–Lieb algebra. These basic elements $\mathcal{T}_i$ will be represented by simple linear transformations on finite dimensional

Hilbert spaces, which can be implemented on a quantum computer. Below we briefly sketch this decomposition and correspondence.

One can easily draw the medial graph L_G in the plane so that one can slice it with horizontal lines so that in between each consecutive horizontal line there is a diagram with only one of the following: a crossing of two lines, a "cap" or a "cup". One can think of the gluing together of these adjacent diagrams $\mathcal{T}_i$ to form the graph L_G as a product operation. We will sketch how to map each $\mathcal{T}_i$ to a linear operation $\rho(\mathcal{T}_i)$ acting on a finite dimensional Hilbert space.

The state space that the operation $\rho(\mathcal{T}_i)$ will act on is the set of finite walks starting at vertex 1 on the infinite graph G with vertices labeled by the non-negative integers and edges $\{i, i+1\}$ for all $i \geq 0$. For example, the walk $1-2-3-2-3$ is a walk of length 4 from 1 to 3. The linear transformation $\rho(\mathcal{T}_i)$ maps a walk w_1 to a linear combination of walks that are "compatible" with $\mathcal{T}_i$ (we won't explain here the details of defining this linear transformation).

In order to apply this technique to a diagram with a crossing, one eliminates the crossing by replacing it with two non-intersecting lines. This replacement can be done in two different ways, and one can represent the diagram with a crossing as a formal linear combination of these two diagrams one gets by replacing a crossing at a vertex (say with label u) with two non-intersecting lines, where one of the two links gets the coefficient u (by a simple rule that we don't explain here). We can then apply the construction to each of these new diagrams, and combine them linearly to get the linear transformation corresponding to the diagram with the crossing.

These linear transformations Q_G are not necessarily unitary (they were unitary in the earlier work on the Jones polynomial, and other related work also constructs unitary representations); however the authors show how one can use ancilla qubits and unitaries to implement non-unitary transformations and approximate the desired inner product using the "Hadamard test" (see Sect. "Quantum Algorithms for Quantum Tasks").

Quantum Algorithms for Quantum Tasks

At present, when we think of quantum algorithms, we usually think of starting with a classical input, running some quantum algorithm, or series of quantum algorithms, with some classical post-processing in order to get a classical output. There might be some quantum sub-routines that have been analyzed, but the main goal in mind is to solve a classical problem.

For example, quantum error correction (see [124] for a recent survey) can be thought of as an algorithm having a quantum input and a quantum output. There are many algorithms for transferring a qubit of information through a network of qubits under some physical constraints. We might develop a quantum cryptographic infrastructure where objects like money and signatures [88] are quantum states that need to be maintained and manipulated as quantum states for long periods of time.

Several of the algorithms described in the previous sections have versions which have a quantum input or a quantum output or both. For example, the amplitude amplification algorithm can be rephrased in terms of producing a quantum state $|\phi\rangle$ given a black-box U_ϕ that recognizes $|\phi\rangle$ by mapping $|\phi\rangle \mapsto -|\phi\rangle$ and acting as the identity on all states orthogonal to $|\phi\rangle$. The end result of such a quantum search is the quantum state $|\phi\rangle$. Amplitude estimation is estimating a transition probability of a unitary operator.

The topological algorithms require as a basic subroutine a quantum algorithm for approximating the inner product of $|0\rangle$ and $U|00\ldots0\rangle$, which the authors call the "Hadamard test". The algorithm consists of using a controlled-U operation to create the state $\frac{1}{\sqrt{2}}|0\rangle|00\ldots0\rangle + \frac{1}{\sqrt{2}}|1\rangle U|00\ldots0\rangle$. Note that if we apply the Hadamard gate to the first qubit, we get

$$\sqrt{\frac{1+\mathrm{Re}\langle 00\ldots0|U|00\ldots0\rangle}{2}}\,|0\rangle|\psi_0\rangle + \sqrt{\frac{1-\mathrm{Re}\langle 00\ldots0|U|00\ldots0\rangle}{2}}\,|1\rangle|\psi_1\rangle$$

for normalized states $|\psi_0\rangle$ and $|\psi_1\rangle$. We can thus estimate the real part of the $\langle 00\ldots0|U|00\ldots0\rangle$ by repeating several times, or applying the quadratically more efficient amplitude estimation algorithm algorithm described earlier (the goal is a superpolynomial speed-up, so a quadratic improvement is not substantial in this case). We can also estimate the complex part similarly.

Another example is the *coset orbit problem* [82] mentioned in Sect. "Generalizations of the Abelian Hidden Subgroup Problem". The input to this problem consists of two quantum states $|\phi_0\rangle$ and $|\phi_1\rangle$ from a set Γ of mutually orthogonal states and black-boxes for implementing the action of a group G on the set Γ. For a state $|\phi\rangle$, let $|u\cdot\phi\rangle$ denote the state resulting from the action of $u \in G$ on $|\phi\rangle$, and let $G_{|\phi\rangle}$ denote the subgroup of G that stabilizes the state $|\phi\rangle$ (i. e. the set of $u \in G$ such that $|u\cdot\phi\rangle = |\phi\rangle$). The question is whether there exists a $u \in G$ such that $|u\cdot\phi_1\rangle = |\phi_0\rangle$. If the answer is "yes", then the set of u satisfying $|u\cdot\phi_1\rangle = |\phi_0\rangle$ is a left coset of

$G_{|\phi_1\rangle}$. Thus, the algorithm should output a coset representative u along with $O(\log n)$ generators of $G_{|\phi_1\rangle}$. The solution to this problem was an important part of the solution to the hidden translation problem. Several other quantum algorithms have such "quantum" sub-routines.

Other examples of quantum tasks include quantum data compression which was known to be information theoretically possible, and efficient quantum circuits for performing it were also developed [26]. Another example is entanglement concentration and distillation.

Researchers have also developed quantum algorithms for implementing some natural basis changes, which have quantum inputs and quantum outputs. For example, the Clebsch-Gordan transformation [20] or other transformations (e. g. [69,103]).

We've only listed a few examples here. In the future, as quantum technologies develop, in particular quantum computers and reliable quantum memories, quantum states and their manipulation will become an end in themselves.

Future Directions

Roughly 10 years ago, many people said that there were essentially only two quantum algorithms. One serious omission was the simulation of quantum mechanical systems, which was in fact Feynman's initial motivation for quantum computers. Apart from this omission, it was true that researchers were developing a better and deeper understanding of the algorithms of Shor and Grover, analyzing them in different ways, generalizing them, and applying them in non-trivial ways. It would have been feasible to write a reasonably sized survey of all the known quantum algorithms with substantial details included. In addition to this important and non-trivial work, researchers were looking hard for "new" approaches, and fresh ideas like quantum walks, and topological algorithms, were being investigated, as well as continued work on the non-Abelian version of the hidden subgroup problem. The whole endeavor of finding new quantum algorithms was very hard and often frustrating. Fortunately, in the last 10 years, there have been many non-trivial developments, enough to make the writing of a full survey of quantum algorithms in a reasonable number of pages impossible. Some directions in which future progress might be made are listed below.

- The complexity of the non-Abelian hidden subgroup problem will hopefully be better understood. This includes addressing the question: Does there exist a quantum polynomial time algorithm for solving the graph isomorphism problem? Of course, a proof that no such quantum algorithm exists would imply that $P \neq PSPACE$. So, more likely, we might develop a strong confidence that no such algorithm exists, in which case this can form the basis of quantum computationally secure cryptography [142].
- There are many examples where amplitude amplification was used to speed-up a classical algorithm in a non-trivial way. That is, in a way that was more than just treating a classical algorithm as a guessing algorithm A and applying amplitude amplification to it. There are likely countless other algorithms which can be improved in non-trivial ways using amplitude amplification.
- The quantum walk paradigm for quantum algorithms emerged roughly 10 years ago, and has recently been applied to find the optimal black-box algorithm for several problems, and has become a standard approach for developing quantum algorithms. Some of these black-box problems are fairly natural, and the black-boxes can be substituted with circuits for actual functions of interest. For example, collision finding can be applied to find collisions in actual hash functions used in cryptography. We will hopefully see more instances where black-box algorithm can be applied to solve an problem without a black-box, or where there is no black-box in the first place.
- In addition to the development of new quantum walk algorithms, we will hopefully have a more elegant and unified general theory of quantum walks that unites continuous and discrete walks, coined and non-coined walks, and quantum and classical walks.
- The adiabatic algorithm paradigm has not reached the level of success of quantum walks, partly because it is hard to analyze the worst case complexity of the algorithms. To date there is no adiabatic algorithm with a proof that it works more than quadratically faster than the best known classical algorithm. Can we do better with an adiabatic algorithm?
  If and when we have large-scale quantum computers, we will be able to just test these algorithms to see if indeed they do have the conjectured running times on instances of interesting size.
- The topological algorithms have received limited attention to date. This is partly because the initial work in this field was largely inaccessible to researchers without substantial familiarity with topological quantum field theory and related areas of mathematics. The more recent work summarized in this paper and other recent papers is a sign that this approach could mature into a fruitful paradigm for developing new important quantum algorithms.

- The paradigm of measurement based computation (see e. g. [118] for an introduction) has been to date mostly focused on its utility as a paradigm for possibly implementing a scalable fault-tolerant quantum computer. We might see the development of algorithms directly in this paradigm. Similarly for globally controlled architectures.
- There is also a growing group of researchers looking at the computational complexity of various computational problems in physics, in particular of simulating certain Hamiltonian systems, often coming from condensed matter physics. Much of the work has been complexity theoretic, such as proving the *QMA*-hardness of computing ground states of certain Hamiltonians (e. g. [9]). Other work has focused on understanding which quantum systems can be simulated efficiently on a classical computer. This work should lead to the definition of some simulation problems that are not known to be in *BPP*, nor believed to be *NP*-hard or *QMA*-hard, and thus might be good candidates for a quantum algorithm. There has been a language and culture barrier between physicists and theoretical computer scientists when it comes to discussing such problems. However, it is slowly breaking down, as more physicists are becoming familiar with algorithms and complexity, and more quantum computer scientists are becoming familiar with language and notions from physics. This will hopefully lead to more quantum algorithms for computational problems in physics, and new algorithmic primitives that can be applied to a wider range of problems.

In summary, as daunting as it is to write a survey of quantum algorithms at this time, it will be a much harder task in another 10 years. Furthermore, in another 10 years we will hopefully have a better idea of when we might expect to see quantum computers large enough to solve problems faster than the best available classical computers.

Bibliography

1. Aaronson S (2003) Algorithms for Boolean function query properties. SIAM J Comput 32:1140–1157
2. Aaronson S, Ambainis A (2003) Quantum search of spatial regions. In: Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS). IEEE Computer Society, Washington, DC, pp 200–209
3. Abrams D, Lloyd S (1999) Quantum algorithm providing exponential speed increase for finding eigenvalues and eigenvectors. Phys Rev Lett 83:5162–5165
4. Aharonov D, Regev O (2005) Lattice problems in NP intersect coNP. J ACM 52:749–765
5. Aharonov D, Ta-Shma A (2007) Adiabatic quantum state generation. SIAM J Comput 37:47–82
6. Aharonov D, Ambainis A, Kempe J, Vazirani U (2001) Quantum walks on graphs. In: Proceedings of ACM Symposium on Theory of Computation (STOC'01). ACM, New York, pp 50–59
7. Aharonov D, van Dam W, Kempe J, Landau Z, Lloyd S, Regev O (2004) Adiabatic auantum computation is equivalent to standard quantum computation. In: Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science (FOCS'04). IEEE Computer Society, Washington, DC, pp 42–51
8. Aharonov D, Jones V, Landau Z (2006) A polynomial quantum algorithm for approximating the Jones polynomial. In: Proceedings of the thirty-eighth annual ACM symposium on Theory of computing (STOC). ACM, New York, pp 427–436
9. Aharonov D, Gottesman D, Irani S, Kempe J (2007) The power of quantum systems on a line. In: Proc. 48th IEEE Symposium on the Foundations of Computer Science (FOCS). IEEE Computer Society, Washington, DC, pp 373–383
10. Aharonov D, Arad I, Eban E, Landau Z Polynomial quantum algorithms for additive approximations of the Potts model and other points of the Tutte plane. quant-ph/0702008
11. Alagic G, Moore C, Russell A (2007) Quantum algorithms for Simon's problem over general groups. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms (SODA). Society for Industrial and Applied Mathematics, Philadelphia, pp 1217–1224
12. Ambainis A (2002) Quantum lower bounds by quantum arguments. J Comput Syst Sci 64:750–767
13. Ambainis A (2003) Quantum walks and their algorithmic applications. Int J Quantum Inf 1:507–518
14. Ambainis A (2004) Quantum search algorithms. SIGACT News. 35(2):22–35
15. Ambainis A (2004) Quantum walk algorithm for element distinctness. In: Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science (FOCS'04). IEEE Computer Society, Washington, DC, pp 22–31
16. Ambainis A, Spalek R (2006) Quantum algorithms for matching and network flows. In: Proceedings of STACS'06. Lecture Notes in Computer Science, vol 3884. Springer, Berlin, pp 172–183
17. Ambainis A, Bach E, Nayak A, Vishwanath A, Watrous J (2001) One-dimensional quantum walks. In: Proceedings of the 33rd ACM Symposium on Theory of Computing. ACM, New York, pp 37–49
18. Ambainis A, Childs A, Reichardt B, Spalek R, Zhang S (2007) Any AND-OR formula of size N can be evaluated in time $N^{1/2+o(1)}$ on a quantum computer. In: 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07). IEEE Computer Society, Washington, DC, pp 363–372
19. Bacon D, Childs A, van Dam W (2005) From optimal measurement to efficient quantum algorithms for the hidden subgroup problem over semidirect product groups. In: Proc 46th IEEE Symposium on Foundations of Computer Science (FOCS 2005). IEEE Computer Society, Washington, DC, pp 469–478
20. Bacon D, Chuang I, Harrow A (2006) Efficient quantum circuits for Schur and Clebsch-Gordan transforms. Phys Rev Lett 97:170502
21. Beals R (1997) Quantum computation of Fourier transforms over symmetric groups. In: Proceedings of the 29th Annual ACM Symposium on Theory of Computing (STOC). ACM, New York, pp 48–53
22. Beals R, Buhrman H, Cleve R, Mosca M, deWolf R (2001) Quantum lower bounds by polynomials. J ACM 48:778–797

23. Benioff P (1982) Quantum mechanical models of Turing machines that dissipate no energy. Phys Rev Lett 48(23): 1581–1585
24. Bennett C (1988) Notes on the history of reversible computation by Charles Bennett. IBM J Res Dev 32(1):16–23
25. Bennett C, Bernstein E, Brassard G, Vazirani U (1997) Strengths and weaknesses of quantum computing. SIAM J Comput 26:1510–1523
26. Bennett C, Harrow A, Lloyd S (2006) Universal quantum data compression via gentle tomography. Phys Rev A 73:032336
27. Bennett CH, Brassard G (1984) Quantum cryptography: Public-key distribution and coin tossing. In: Proceedings of IEEE International Conference on Computers, Systems and Signal Processing, India. IEEE Computer Society, Washington, DC, pp 175–179
28. Bernstein E, Vazirani U (1997) Quantum complexity theory. SIAM J Comput 26:1411–1473
29. Berry DW, Ahokas G, Cleve R, Sanders BC (2007) Efficient quantum algorithms for simulating sparse Hamiltonians. Comm Math Phys 270:359
30. Berthiaume A, Brassard G (1992) The quantum challenge to structural complexity theory. In: Proc 7th Conference Structure Complexity Theory, IEEE Comp Soc Press, pp 132–137
31. Berthiaume A, Brassard G (1994) Oracle quantum computing. J Modern Opt 41(12):2521–2535
32. Boneh D, Lipton R (1995) Quantum cryptanalysis of hidden linear functions (Extended Abstract). In: Proceedings of 15th Annual International Cryptology Conference (CRYPTO'95). Springer, London, pp 424–437
33. Bordewich M, Freedman M, Lovasz L, Welsh DJA (2005) Approximate counting and quantum computation. Comb Probab Comput 14:737–754
34. Boyer M, Brassard G, Høyer P, Tapp A (1998) Tight bounds on quantum searching. Fortschr Phys 56(5-5):493–505
35. Brassard G, Høyer P (1997) An exact quantum polynomial-time algorithm for Simon's problem. In: Proc of Fifth Israeli Symposium on Theory of Computing and Systems (ISTCS'97). IEEE Computer Society, Washington, DC, pp 12–23
36. Brassard G, Høyer P, Tapp A (1997) Cryptology column–quantum algorithm for the collision problem. ACM SIGACT News 28:14–19
37. Brassard G, Høyer P, Tapp A (1998) Quantum Counting. In: Proceedings of the ICALP'98. Lecture notes in computer science. Springer, Berlin, pp 1820–1831
38. Brassard G, Høyer P, Mosca M, Tapp A (2002) Quantum amplitude amplification and estimation. In: Quantum Computation and Quantum Information Science. AMS Contemp Math Ser 35:53–74
39. Brown M (2003) Classical cryptosystems in a quantum setting. Master Thesis, University of Waterloo
40. Buhrman H, Špalek B (2006) Quantum verification of matrix products. In: Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). IEEE Computer Society, Washington, DC, pp 880–889
41. Buhrman H, Fortnow L, Newman I, Röhrig H (2003) Quantum property testing. In: Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, Philadelphia, pp 480–488
42. Buhrman H, Dürr C, Heiligman M, Høyer P, Magniez F, Santha M, de Wolf R (2005) Quantum algorithms for element distinctness. SIAM J Comput 34:1324–1330
43. Byrnes T, Yamamoto Y (2006) Simulating lattice gauge theories on a quantum computer. Phys Rev A 73:022328
44. Cheung K, Mosca M (2001) Decomposing finite Abelian groups. Quantum Inf Comput 1(2):26–32
45. Childs A (2002) Quantum information processing in continuous time. Ph D thesis, MIT
46. Childs A, van Dam W (2007) Quantum algorithm for a generalized hidden shift problem. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). ACM, New York, pp 1225–1232
47. Childs A, Farhi E, Gutmann S (2002) An example of the difference between quantum and classical random walks. Quantum Inf Process 1:35–43
48. Childs A, Cleve R, Deotto E, Farhi E, Gutmann S, Spielman D (2003) Exponential algorithmic speedup by a quantum walk. In: Proceedings of the 35th Annual ACM Symposium on Theory of Computing. IEEE Computer Society, Washington, DC
49. Childs A, Landahl A, Parrilo P (2007) Improved quantum algorithms for the ordered search problem via semidefinite programming. Phys Rev A 75:032335
50. Childs AM, Cleve R, Jordan SP, Yeung D Discrete-query quantum algorithm for NAND trees. quant-ph/0702160
51. Childs A, Schulman L, Vazirani U (2007) Quantum algorithms for hidden nonlinear structures. In: Proceeding of 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2007). IEEE Computer Society, Washington, DC
52. Cleve R (2000) An introduction to quantum complexity theory. In: Macchiavello C, Palma GM, Zeilinger A (eds) Collected Papers on Quantum Computation and Quantum Information Theory. World Scientific, Singapore, pp 103–127
53. Cleve R (2000) The query complexity of order-finding. In: IEEE Conference on Computational Complexity. IEEE Computer Society, Washington, DC, p 54
54. Cleve R, Ekert A, Macchiavello C, Mosca M (1998) Quantum algorithms revisited. Proc Royal Soc Lond A 454:339–354
55. Cleve R, Gavinsky D, Yeung D (2008) Quantum algorithms for evaluating Min-Max trees. In: Proceedings of TQC 2008. Lecture Notes in Computer Science, vol 5106. Springer, Berlin
56. van Dam W, Seroussi G Efficient quantum algorithms for estimating Gauss sums. quant-ph/0207131
57. van Dam W, Shparlinski I (2008) Classical and quantum algorithms for exponential congruences. In: Proceedings of TQC 2008. Lecture Notes in Computer Science, vol 5106. Springer, Berlin, pp 1–10
58. van Dam W, Mosca M, Vazirani U (2001) How powerful is adiabatic quantum computation? In: Proceedings 46th IEEE Symposium on Foundations of Computer Science (FOCS'01). IEEE Computer Society, Washington, DC, pp 279–287
59. van Dam W, Hallgren S, Ip L (2003) Quantum algorithms for some hidden shift problems. In: Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA'03). Society for Industrial and Applied Mathematics, Philadelphia, pp 489–498
60. van Dam W, D'Ariano GM, Ekert A, Macchiavello C, Mosca M (2007) General optimized schemes for phase estimation. Phys Rev Lett 98:090501
61. van Dam W, D'Ariano GM, Ekert A, Macchiavello C, Mosca M (2007) Optimal phase estimation in quantum networks. J Phys A: Math Theor 40:7971–7984
62. Das A, Chakrabarti BK (2008) Quantum annealing and analog quantum computation. Rev Mod Phys 80:1061

63. Decker T, Draisma J, Wocjan P (2009) Efficient quantum algorithm for identifying hidden polynomials. Quantum Inf Comput (to appear)
64. Deutsch D (1985) Quantum theory, the Church–Turing principle and the universal quantum computer. Proc Royal Soc Lond A 400:97–117
65. Deutsch D, Jozsa R (1992) Rapid solutions of problems by quantum computation. Proc Royal Soc Lond A 439:553–558
66. Dörn S, Thierauf T (2007) The quantum query complexity of algebraic properties. In: Proceedings of the 16th International Symposium on Fundamentals of Computation Theory (FCT). Lecture Notes in Computer Science, vol 4639. Springer, Berlin, pp 250–260
67. Durr C, Heiligman M, Høyer P, Mhalla M (2004) Quantum query complexity of some graph problems. In: Proceedings of 31st International Colloquium on Automata, Languages, and Programming (ICALP'04). Lecture Notes in Computer Science, vol 3142. Springer, Berlin pp 481–493
68. Ettinger JM (1998) On noncommutative hidden subgroups. Lecture at AQIP'98. http://www.brics.dk/~salvail/aqip/abstlist.html#upload-paper.10480.1
69. Ettinger JM Quantum time-frequency transforms. quant-ph/0005134
70. Ettinger M, Høyer P (2000) On quantum algorithms for noncommutative hidden subgroups. Adv Appl Math 25(3): 239–251
71. Ettinger M, Høyer P, Knill E (2004) The quantum query complexity of the hidden subgroup problem is polynomial. Inf Process Lett 91:43–48
72. Farhi E, Gutmann S (1998) Quantum computation and decision trees. Phys Rev A 58:915–928
73. Farhi E, Gutmann S An analog analogue of a digital quantum computation. quant-ph/9612026
74. Farhi E, Goldstone J, Gutmann S, Sipser M (2000) Quantum computation by adiabatic evolution. quant-ph/0001106
75. Farhi E, Goldstone J, Gutmann S, Lapan J, Lundgren A, Preda D (2001) A quantum adiabatic evolution algorithm applied to random instances of an NP. Science 20 April:472
76. Farhi E, Goldstone J, Gutmann S A quantum algorithm for the Hamiltonian NAND tree. quant-ph/0702144v2
77. Farhi E, Goldstone J, Gutmann S Quantum adiabatic evolution algorithms with different paths. quant-ph/0208135
78. Fenner S, Zhang Y (2005) Quantum algorithms for a set of group theoretic problems. In: Proceedings of the Ninth ICEATCS Italian Conference on Theoretical Computer Science. Lecture Notes in Computer Science, vol 3701. Springer, Berlin, pp 215–227
79. Feynman R (1982) Simulating physics with computers. Int J Theor Phys 21(6,7):467–488
80. Freedman M (1998) P/NP, and the quantum field computer. Proc Natl Acad Sci 95(1):98–101
81. Freedman M, Kitaev A, Wang Z (2002) Simulation of topological Field theories by quantum computers. Comm Math Phys 227(3):587–603
82. Friedl K, Ivanyos G, Magniez F, Santha M, Sen P (2003) Hidden translation and orbit coset in quantum computing. In: Proceedings of the thirty-fifth annual ACM symposium on Theory of computing (STOC'03). IEEE Computer Society, Washington, DC, pp 1–9
83. Friedl K, Ivanyos G, Santha M (2005) Efficient testing of groups. In: Proceedings of the Thirty-seventh Annual ACM Symposium on Theory of Computing (STOC). IEEE Computer Society, Washington, DC, pp 157–166
84. Furrow B (2008) A panoply of quantum algorithms. Quantum Inf Comput 8(8–9):0834–0859
85. Geraci J, Lidar D (2008) On the exact evaluation of certain instances of the potts Partition function by quantum computers. Comm Math Phys 279:735
86. Gerhardt H, Watrous J (2003) Continuous-time quantum walks on the symmetric group. 6th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems, (APPROX) and 7th International Workshop on Randomization and Approximation Techniques in Computer Science, (RANDOM), pp 290–301
87. Grigni M, Schulman L, Vazirani M, Vazirani U (2001) Quantum mechanical algorithms for the nonabelian hidden subgroup problem. In: Proceedings of the thirty-third annual ACM symposium on Theory of computing (SODA'03), pp 68–74
88. Gottesman D, Chuang I Quantum digital signatures. quant-ph/0105032
89. Grigoriev D (1997) Testing shift-equivalence of polynomials by deterministic, probabilistic and quantum machines. Theor Comput Sci 180:217–228
90. Grover L (1996) A fast quantum mechanical algorithm for database search. In: Proceedings of the 28th Annual ACM Symposium on the Theory of Computing (STOC 1996), pp 212–219
91. Grover L (1998) A framework for fast quantum mechanical algorithms. In: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing (STOC), pp 53–62
92. Hales L, Hallgren S (2000) An Improved Quantum Fourier Transform Algorithm and Applications. FOCS 2000, pp 515–525
93. Hallgren S (2002) Polynomial-time quantum algorithms for Pell's equation and the principal ideal problem. STOC 2002, pp 653–658
94. Hallgren S (2005) Fast quantum algorithms for computing the unit group and class group of a number field. In: Proceedings of the 37th ACM Symposium on Theory of Computing (STOC 2005), pp 468–474
95. Hallgren S (2007) Polynomial-time quantum algorithms for Pell's equation and the principal ideal problem. J ACM 54(1):653–658
96. Hallgren S, Russell A, Ta-Shma A (2000) Normal subgroup reconstruction and quantum computation using group representations. In: Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing, pp 627–635
97. Hallgren S, Moore C, Roetteler M, Russel A, Sen P (2006) Limitations of quantum coset states for graph isomorphism. In: Proceedings of the 38th Annual ACM Symposium on Theory of Computing. IEEE Computer Society, Washington, DC
98. Hassidim A, Ben-Or M (2008) The Bayesian learner is optimal for noisy binary search (and pretty good for quantum as well). In: Proceedings of the 49th Symposium on Foundations of Computer Science. IEEE Computer Society, Washington, DC
99. Hausladen P, Wootters WK (1994) A pretty good measurement for distinguishing quantum. states. J Mod Opt 41:2385
100. Helstrom CW (1976) Quantum detection and estimation theory. Academic Press, New York
101. Hogg T (2000) Quantum search heuristics. Phys Rev A 61:052311

102. Holevo AS (1982) Probabilistic and statistical aspects of quantum theory. North Holland, Amsterdam
103. Høyer P Efficient quantum transforms. quant-ph/9702028
104. Høyer P (1999) Conjugated operators in quantum algorithms. Phys Rev A 59(5):3280–3289
105. Høyer P (2001) Introduction to recent quantum algorithms. In: Proceedings of 26th International Symposium on Mathematical Foundations of Computer Science (MFCS'01). Lecture Notes in Computer Science, vol 2136. Springer, Berlin, pp 62–73
106. Høyer P, Dürr C A quantum algorithm for finding the minimum. quant-ph/9607014
107. Høyer P, Neerbek J, Shi Y (2002) Quantum complexities of ordered searching, sorting, and element distinctness. Algorithmica 34(4):429–448
108. Høyer P, Mosca M, de Wolf R (2003) Quantum search on bounded-error inputs. In: Proceedings of the Thirtieth International Colloquium on Automata, Languages and Programming (ICALP'03), pp 291–299
109. Inui Y, Le Gall F (2008) Quantum property testing of group solvability. In: Proceedings of the 8th Latin American Symposium (LATIN'08: Theoretical Informatics). Lecture Notes in Computer Science, vol 4957. Springer, Berlin, pp 772–783
110. Ivanyos G, Magniez F, Santha M (2003) Efficient quantum algorithms for some instances of the non-Abelian hidden subgroup problem. Int J Found Comput Sci 14(5):723–739
111. Ivanyos G, Sanselme L, Santha M (2007) An efficient quantum algorithm for the Hidden subgroup problem in extraspecial groups. 24th STACS. Lecture Notes in Computer Science, vol 4393. Springer, Berlin, pp 586–597
112. Ivanyos G, Sanselme L, Santha M (2008) An efficient quantum algorithm for the hidden subgroup problem in nil-2 groups. In: Proceedings of the 8th Latin American Symposium (LATIN'08: Theoretical Informatics). Lecture Notes in Computer Science, vol 4957. Springer, Berlin, pp 759–771
113. Ioannou L (2002) Continuous-time quantum algorithms: Searching and adiabatic computation. Master Math Thesis, University of Waterloo
114. Jaeger F, Vertigan DL, Welsh DJA (1990) On the Computational complexity of the Jones and Tutte polynomials. Math Proc Cambridge Phil Soc 108(1):5–53
115. Jansen S, Ruskai M, Seiler R (2007) Bounds for the adiabatic approximation with applications to quantum computation. J Math Phys 48:102111
116. Jozsa R (1998) Quantum algorithms and the Fourier transform. Proc Royal Soc Lond A 454:323–337
117. Jozsa R (2003) Notes on Hallgren's efficient quantum algorithm for solving Pell's equation. quant-ph/0302134
118. Jozsa R (2006) An introduction to measurement based quantum computation. NATO Science Series, III: Computer and Systems Sciences. Quantum Information Processing - From Theory to Experiment, vol 199, pp 137–158
119. Kassal I, Jordan S, Love P, Mohseni M, Aspuru-Guzik A (2008) Quantum algorithms for the simulation of chemical dynamics. arXiv:0801.2986
120. Kauffman L (2001) Knots and Physics. World Scientific, Singapore
121. Kaye P (2005) Optimized quantum implementation of elliptic curve arithmetic over binary fields. Quantum Inf Comput 5(6):474–491
122. Kaye P, Laflamme R, Mosca M (2007) An introduction to quantum computation. Oxford University Press, Oxford
123. Kempe J (2003) Quantum random walks - an introductory overview. Contemp Phys 44(4):307–327
124. Kempe J (2007) Approaches to quantum error correction. Quantum Decoherence, Poincare Seminar 2005. Progress in Mathematical Physics. Birkhäuser, Basel, pp 85–123
125. Kempe J, Shalev A (2005) The hidden subgroup problem and permutation group theory. In: Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms (SODA'05), pp 1118–1125
126. Kitaev A (1995) Quantum measurements and the Abelian stabilizer problem. quant-ph/9511026
127. Kitaev A (1996) Quantum measurements and the Abelian stabilizer problem. Electronic Colloquium on Computational Complexity (ECCC), p 3
128. Kitaev A (1997) Quantum computations: Algorithms and error correction. Russ Math Surv 52(6):1191–1249
129. Kitaev A, Shen A, Vyalvi M (2002) Classical and quantum computation. American Mathematical Society
130. Kuperberg G (2005) A subexponential-time quantum algorithm for the dihedral hidden subgroup problem. SIAM J Comput 35:170–188
131. Lloyd S (1996) Universal quantum simulators. Science 273:1073–1078
132. Lomonaco S, Kauffman L (2006) Topological quantum computing and the Jones polynomial. In: Proc SPIE 6244
133. Magniez F, Nayak A (2007) Quantum complexity of testing group commutativity. Algorithmica 48(3):221–232
134. Magniez F, Nayak A, Roland J, Santha M (2007) Search via quantum walk. 39th ACM Symposium on Theory of Computing (STOC), pp 575–584
135. Magniez F, Santha M, Szegedy M (2007) Quantum algorithms for the triangle problem. SIAM J Comput 37(2):413–424
136. Manin Y (2000) Classical computing, quantum computing, and Shor's factoring algorithm. Séminaire Bourbaki, 41 (1998-1999), 862, pp 375–404. The appendix translates an excerpt about quantum computing from a 1980 paper (in Russian)
137. Menezes A, van Oorschot P, Vanstone S (1996) Handbook of Applied Cryptography. CRC Press, Boca Raton
138. Moore C, Russell A (2007) For distinguishing conjugate hidden dubgroups, the pretty good measurement is as hood as it gets. Quantum Inf Comput 7(8):752–765
139. Moore C, Rockmore D, Russell A, Schulman L (2004) The power of basis selection in fourier sampling: hidden subgroup problems in affine groups. In: Proceedings of the fifteenth annual ACM-SIAM symposium on discrete algorithms (SODA'04), pp 1113–1122
140. Moore C, Rockmore D, Russell A (2006) Generic Quantum Fourier Transforms. ACM Trans Algorithms 2:707–723
141. Moore C, Russell A, Sniady P (2007) On the impossibility of a quantum sieve algorithm for graph isomorphism: unconditional results. In: Proceedings of the thirty-ninth annual ACM symposium on Theory of computing (STOC), pp 536–545
142. Moore C, Russell A, Vazirani U A classical one-way function to confound quantum adversaries. quant-ph/0701115
143. Mosca M (1999) Quantum computer algorithms. Ph D thesis, Oxford
144. Mosca M (2001) Counting by quantum eigenvalue estimation. Theor Comput Sci 264:139–153
145. Mosca M, Ekert A (1998) The hidden subgroup problem and

eigenvalue estimation on a quantum computer. In: Proceedings 1st NASA International Conference on Quantum Computing & Quantum Communications. Lecture Notes in Computer Science, vol 1509. Springer, Berlin, pp 174–188
146. Mosca M, Zalka C (2004) Exact quantum Fourier transforms and discrete logarithm algorithms. Int J Quantum Inf 2(1): 91–100
147. Motwani R, Raghavan P (1995) Randomized Algorithms. Cambridge University Press, Cambridge
148. Nayak A, Wu F (1999) The quantum query complexity of approximating the median and related statistics. In: Proceedings of the Thirty-first Annual ACM Symposium on Theory of Computing (STOC), pp 384–393
149. Nielsen M, Chuang I (2000) Quantum computation and quantum information. Cambridge University Press, Cambridge
150. Proos J, Zalka C (2003) Shor's discrete logarithm quantum algorithm for elliptic curves. Quantum Inf Comput 3:317–344
151. Püschel M, Rötteler M, Beth T (1999) Fast quantum Fourier transforms for a class of non-Abelian groups. In: Proceedings of the 13th International Symposium on Applied Algebra, Algebraic Algorithms and Error-Correcting Codes, pp 148–159
152. Radhakrishnan J, Roetteler M, Sen P (2005) On the power of random bases in Fourier sampling: Hidden subgroup problem in the Heisenberg group. In: Proceedings of the 32nd International Colloquium on Automata, Languages and Programming (ICALP), pp 1399–1411
153. Ramesh H, Vinay V (2003) String matching in $\tilde{O}(\sqrt{n}+\sqrt{m})$ quantum time. J Discret Algorithms 1:103–110
154. Regev O (2004) Quantum computation and lattice problems. SIAM J Comput 33:738–760
155. Reichardt B, Spalek R Span-program-based quantum algorithm for evaluating formulas. In: Proceedings of the fortieth annual ACM symposium on Theory of computing (STOC 2008), pp 103–112
156. Roetteler M, Beth T Polynomial-time solution to the hidden subgroup problem for a class of non-Abelian groups. quant-ph/9812070
157. Roland J, Cerf N (2002) Quantum search by local adiabatic evolution. Phys Rev A 65(4):042308
158. Rudolph T, Grover L (2004) How significant are the known collision and element distinctness quantum algorithms? Quantum Inf Comput 4:201–206
159. Russell A, Shparlinski I (2004) Classical and quantum function reconstruction via character evaluation. J Complex 20: 404–422
160. Santha M (2008) Quantum walk based search algorithms. In: Proceedings of the 5th Annual International Conference on Theory and Applications of Models of Computation (TAMC 2008). Lecture Notes in Computer Science, vol 4978. Springer, Berlin, pp 31–46
161. Sasaki M, Carlini A, Jozsa R (2001) Quantum template matching. Phys Rev A 64:022317
162. Schmidt A, Vollmer U (2005) Polynomial time quantum algorithm for the computation of the unit group of a number field. In: Proceedings of the thirty-seventh annual ACM symposium on Theory of computing, pp 475–480
163. Schumacher B (1995) Quantum coding. Phys Rev A 51:2738–2747
164. Severini S (2003) On the digraph of a unitary matrix. SIAM J Matrix Anal Appl (SIMAX) 25(1):295–300
165. Shor P (1994) Algorithms for quantum computation: discrete logarithms and factoring. In: Proceedings of the 35th Annual Symposium on Foundations of Computer Science, pp 124–134
166. Shor P (1997) Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. SIAM J Comput 26:1484–1509
167. Simon D (1994) On the power of quantum computation. In: Proceedings of the 35th IEEE Symposium on the Foundations of Computer Science (FOCS), pp 116–123
168. Simon D (1997) On the power of quantum computation. SIAM J Comput 26:1474–1483
169. Sokal A (2005) The multivariate Tutte polynomial (alias Potts model) for graphs and matroids. Surveys in Combinatorics. Cambridge University Press, Cambridge, pp 173–226
170. Somma R, Ortiz G, Gubernatis JE, Knill E, Laflamme R (2002) Simulating physical phenomena by quantum networks. Phys Rev A 65:042323
171. Szegedy M (2004) Quantum speed-Up of Markov chain based algorithms. In: Proceedings of the 45th IEEE Symposium on the Foundations of Computer Science (FOCS), pp 32–41
172. Terhal B (1999) Quantum algorithms and quantum entanglement. Ph D thesis, University of Amsterdam
173. Vazirani U (1998) On the power of quantum computation. Phil Trans Royal Soc Lond A 356:1759–1768
174. Vazirani U (1997) Berkeley Lecture Notes. Lecture 8 http://www.cs.berkeley.edu/~vazirani/qc.html
175. Watrous J (2001) Quantum algorithms for solvable groups. In: Proceedings of the thirty-third annual ACM symposium on Theory of computing (STOC), pp 60–67
176. Welsh DJA (1993) Complexity: knots, colourings and counting. Cambridge University Press, Cambridge
177. Wiesner S (1983) Conjugate coding. Sigact News 15(1):78–88
178. Wiesner S (1996) Simulations of many-body quantum systems by a quantum computer. quant-ph/9603028
179. Wocjan P, Yard J (2008) The Jones polynomial: quantum algorithms and applications in quantum complexity theory. Quantum Inf Comput 8(1–2):147–180
180. Zalka C (1998) Efficient simulation of quantum systems by quantum computers. Proc Roy Soc Lond A 454:313–322
181. Zalka C (2000) Using Grover's quantum algorithm for searching actual databases. Phys Rev A 62(5):052305
182. Zhang S (2006) New upper and lower bounds for randomized and quantum local search. In: Proceedings of the thirty-eighth annual ACM symposium on Theory of computing (STOC), pp 634–643

Quantum Algorithms and Complexity for Continuous Problems

Anargyros Papageorgiou, Joseph F. Traub
Department of Computer Science,
Columbia University, New York, USA

Article Outline

Glossary
Definition of the Subject

Introduction
Overview of Quantum Algorithms
Integration
Path Integration
Feynman–Kac Path Integration
Eigenvalue Approximation
Qubit Complexity
Approximation
Elliptic Partial Differential Equations
Ordinary Differential Equations
Gradient Estimation
Simulation of Quantum Systems on Quantum Computers
Future Directions
Acknowledgments
Bibliography

Glossary

Black box model This model assumes we can collect knowledge about an input f through queries without knowing how the answer to the query is computed. A synonym for black box is oracle.

Classical computer A computer which does not use the principles of quantum computing to carry out its computations.

Computational complexity In this article, complexity for brevity. The minimal cost of solving a problem by an algorithm. Some authors use the word complexity when cost would be preferable. An upper bound on the complexity is given by the cost of an algorithm. A lower bound is given by a theorem which states there cannot be an algorithm which does better.

Continuous problem A problem involving real or complex functions of real or complex variables. Examples of continuous problem are integrals, path integrals, and partial differential equations.

Cost of an algorithm The price of executing an algorithm. The cost depends on the model of computation.

Discrete problem A problem whose inputs are from a countable set. Examples of discrete problems are integer factorization, traveling salesman and satisfiability.

ε-Approximation Most real-world continuous problems can only be solved numerically and therefore approximately, that is to within an error threshold ε. The definition of ε-approximation depends on the setting. See worst-case setting, randomized setting, quantum setting.

Information-based complexity The discipline that studies algorithms and complexity of continuous problems.

Model of computation The rules stating what is permitted in a computation and how much it costs. The model of computation is an abstraction of a physical computer. Examples of models are Turing machines, real number model, quantum circuit model.

Optimal algorithm An algorithm whose cost equals the complexity of the problem.

Promise A statement of what is known about a problem a priori before any queries are made. An example in quantum computation is the promise that an unknown 1-bit function is constant or balanced. In information-based complexity a promise is also called global information.

Quantum computing speedup The amount by which a quantum computer can solve a problem faster than a classical computer. To compute the speedup one must know the classical complexity and it is desirable to also know the quantum complexity. Grover proved quadratic speedup for search in an unstructured database. Its only conjectured that Shor's algorithm provides exponential speedup for integer factorization since the classical complexity is unknown.

Query One obtains knowledge about a particular input through queries. For example, if the problem is numerical approximation of $\int_0^1 f(x)\,\mathrm{d}x$ a query might be the evaluation of f at a point. In information-based complexity the same concept is called an information operation.

Quantum setting There are a number of quantum settings. An example is a guarantee of error at most ε with probability greater than 1/2.

Qubit complexity The minimal number of qubits to solve a problem.

Query complexity The minimal number of queries required to solve the problem.

Randomized setting In this setting the expected error with respect to the probability measure generating the random variables is at most ε. The computation is randomized. An important example of a randomized algorithm is the Monte Carlo method.

Worst-case setting In this setting an error of at most ε is guaranteed for all inputs satisfying the promise. The computation is deterministic.

Definition of the Subject

Most continuous mathematical formulations arising in science and engineering can only be solved numerically

and therefore approximately. We shall always assume that we are dealing with a numerical approximation to the solution.

There are two major motivations for studying quantum algorithms and complexity for continuous problems.

1. Are quantum computers more powerful than classical computers for important scientific problems? How much more powerful? This would answer the question posed by Nielsen and Chuang (p. 47 in [48]).
2. Many important scientific and engineering problems have continuous formulations. These problems occur in fields such as physics, chemistry, engineering and finance. The continuous formulations include path integration, partial differential equations (in particular, the Schrödinger equation) and continuous optimization.

To answer the first question we must know the classical computational complexity (for brevity, complexity) of the problem. There have been decades of research on the classical complexity of continuous problems in the field of information-based complexity; see, for example, the monographs [49,57,59,67,68,71,75]. The reason we know the complexity of many continuous problems is that we can use adversary arguments to obtain their query complexity. See, for example, [54] for an exposition. This may be contrasted with the classical complexity of discrete problems where we have only conjectures such as $\mathrm{P} \neq \mathrm{NP}$. Even the classical complexity of the factorization of large integers is unknown. Knowing the classical complexity of a continuous problem we obtain the quantum computation speedup if we know the quantum complexity. If we know an upper bound on the quantum complexity through the cost of a particular quantum algorithm then we can obtain a lower bound on the quantum speedup. To state and prove complexity theorems, the mathematical formulation of the problem, the promise about the class of inputs, and the model of computation must be precisely specified. See [66] for a discussion of the model of computation for continuous problems.

Regarding the second motivation, in this article we will report on high-dimensional integration, path integration, Feynman path integration, the smallest eigenvalue of a differential equation, approximation, partial differential equations, ordinary differential equations and gradient estimation. We will also briefly report on the simulation of quantum systems on a quantum computer.

Introduction

We provide a summary of the contents of the article.

Section "Overview of Quantum Algorithms" We define basic concepts and notation including quantum algorithm, continuous problem, query complexity and qubit complexity.

Section "Integration" High-dimensional integration, often in hundreds or thousands of variables, is one of the most commonly occurring continuous problems in science. In Subsect. "Classical Computer" we report on complexity results on a classical computer. For illustration we begin with a one-dimensional problem and give a simple example of an adversary argument. We then move on the d-dimensional case and indicate that in the worst case the complexity is exponential in d; the problem suffers the curse of dimensionality. The curse can be broken by the Monte Carlo algorithm. In Subsect. "Quantum Computer" we report on the algorithms and complexity results on a quantum computer. Under certain assumptions on the promise the quantum query complexity enjoys exponential speedup over classical worst case query complexity.

A number of the problems we will discuss enjoy exponential quantum query speedup. This does not contradict Beals et al. [6] who prove that speedup can only be polynomial. The reason is that [6] deals with problems concerning total Boolean functions.

For many classes of integrands there is quadratic speedup over the classical randomized query complexity. This is the same speedup as enjoyed by Grover's search algorithm of an unstructured database. To obtain the quantum query complexity one has to give matching upper and lower bounds. As usual the upper bound is given by an algorithm, the lower bound by a theorem. The upper bound is given by the amplitude amplification algorithm of Brassard et al. [12]. We outline a method for approximating high-dimensional integrals using this algorithm. The quantum query complexity lower bounds for integration are based on the lower bounds of Nayak and Wu [47] for computing the mean of a Boolean function.

Section "Path Integration" We define a path integral and provide an example due to Feynman. In Subsect. "Classical Computer" we report on complexity results on a classical computer. If the promise is that the class of integrands has finite smoothness, then path integration is intractable in the worst case. If the promise is that the class of integrands consists of entire functions the query complexity is tractable even in the worst case. For smooth functions intractability is broken by Monte Carlo. In Subsect. "Quantum Computer" we report on the algorithm and complexity on a quantum computer. The quantum query complexity enjoys Grover-type speedup over clas-

sical randomized query complexity. We outline the quantum algorithm.

Section "Feynman–Kac Path Integration" The Feynman–Kac path integral provides the solution to the diffusion equation. In Subsect. "Classical Computer" we report on algorithms and complexity on a classical computer. In the worst case for a d-dimensional Feynman–Kac path integral the problem again suffers the curse of dimensionality which can be broken by Monte Carlo. In Subsect. "Quantum Computer" we indicate the algorithm and query complexity on a quantum computer.

Section "Eigenvalue Approximation" One of the most important problems in physics and chemistry is approximating the ground state energy governed by a differential equation. Typically, numerical algorithms on a classical computer are well known. Our focus is on the Sturm–Liouville eigenvalue (SLE) problem where the first complexity results were recently obtained. The SLE equation is also called the time-independent Schrödinger equation. In Subsect. "Classical Computer" we present an algorithm on a classical computer. The worst case query complexity suffers the curse of dimensionality. We also state a randomized algorithm. The randomized query complexity is unknown for $d > 2$ and is an important open question. In Subsect. "Quantum Computer" we outline an algorithm for a quantum computer. The quantum query complexity is not known when $d > 1$. It has been shown that it is not exponential in d; the problem is tractable on a quantum computer.

Section "Qubit Complexity" So far we've focused on query complexity. For the foreseeable future the number of qubits will be a crucial computational resource. We give a general lower bound on the qubit complexity of continuous problems. We show that because of this lower bound there's a problem that cannot be solved on a quantum computer but that's easy to solve on a classical computer using Monte Carlo.

A definition of a quantum algorithm is given by (1); the queries are deterministic. Woźniakowski [77] introduced the quantum setting with randomized queries for continuous problems. For path integration there is an exponential reduction in the qubit complexity.

Section "Approximation" Approximating functions of d variables is a fundamental and generally hard problem. The complexity is sensitive to the norm, p, on the class of functions and to several other parameters. For example, if $p = \infty$ approximation suffers the curse of dimensionality in the worst and randomized classical cases. The problem remains intractable in the quantum setting.

Section "Partial Differential Equations" Elliptic partial differential equations have many important applications and have been extensively studied. In particular, consider an equation of order $2m$ in d variables. In the classical worst case setting the problem is intractable. The classical randomized and quantum settings were only recently studied. The conclusion is that the quantum may or may not provide a polynomial speedup; it depends on certain parameters.

Section "Ordinary Differential Equations" Consider a system of initial value ordinary equations in d variables. Assume that the right hand sides satisfy a Hölder condition. The problem is tractable even in the worst case with the exponent of ε^{-1} depending on the Hölder class parameters. The complexity of classical randomized and quantum algorithms have only recently been obtained. The quantum setting yields a polynomial speedup.

Section "Gradient Estimation" Jordan [37] showed that approximating the gradient of a function can be done with a single query on a quantum computer although it takes $d + 1$ function evaluations on a classical computer. A simplified version of Jordan's algorithm is presented.

Section "Simulation of Quantum Systems on Quantum Computers" There is a large and varied literature on simulation of quantum systems on quantum computers. The focus in these papers is typically on the cost of particular classical and quantum algorithms without complexity analysis and therefore without speedup results. To give the reader a taste of this area we list some sample papers.

Section "Future Directions" We briefly indicate a number of open questions.

Overview of Quantum Algorithms

A quantum algorithm consists of a sequence of unitary transformations applied to an initial state. The result of the algorithm is obtained by measuring its final state. The quantum model of computation is discussed in detail in [6,7,8,17,27,48] and we summarize it here as it applies to continuous problems.

The initial state $|\psi_0\rangle$ of the algorithm is a unit vector of the Hilbert space $\mathcal{H}_\nu = \mathbb{C}^2 \otimes \cdots \otimes \mathbb{C}^2$, ν times, for some appropriately chosen integer ν, where $\mathbb{C}^2$ is the two dimensional space of complex numbers. The dimension of

$\mathcal{H}_\nu$ is 2^ν. The number ν denotes the number of qubits used by the quantum algorithm.

The final state $|\psi\rangle$ is also a unit vector of $\mathcal{H}_\nu$ and is obtained from the initial state $|\psi_0\rangle$ through a sequence of unitary $2^\nu \times 2^\nu$ matrices, i. e.,

$$|\psi\rangle_f := U_T Q_f U_{T-1} Q_f \cdots U_1 Q_f U_0 |\psi_0\rangle \,. \tag{1}$$

The unitary matrix Q_f is called a quantum query and is used to provide information to the algorithm about a function f. Q_f depends on n function evaluations $f(t_1), \ldots, f(t_n)$, at deterministically chosen points, $n \le 2^\nu$. The $U_0, U_1, \ldots, U_T$ are unitary matrices that do not depend on f. The integer T denotes the number of quantum queries.

For algorithms solving discrete problems, such as Grover's algorithm for the search of an unordered database [26], the input f is considered to be a Boolean function. The most commonly studied quantum query is the *bit* query. For a Boolean function $f\colon \{0, \ldots, 2^m - 1\} \to \{0, 1\}$, the bit query is defined by

$$Q_f |j\rangle |k\rangle = |j\rangle |k \oplus f(j)\rangle \,. \tag{2}$$

Here $\nu = m + 1$, $|j\rangle \in \mathcal{H}_m$, and $|k\rangle \in \mathcal{H}_1$ with $\oplus$ denoting addition modulo 2. For a real function f the query is constructed by taking the most significant bits of the function f evaluated at some points t_j. More precisely, as in [27], the bit query for f has the form

$$Q_f |j\rangle |k\rangle = |j\rangle |k \oplus \beta(f(\tau(j)))\rangle \,, \tag{3}$$

where the number of qubits is now $\nu = m' + m''$ and $|j\rangle \in \mathcal{H}_{m'}$, $|k\rangle \in \mathcal{H}_{m''}$. The functions β and τ are used to discretize the domain $\mathcal{D}$ and the range $\mathcal{R}$ of f, respectively. Therefore, $\beta\colon \mathcal{R} \to \{0, 1, \ldots, 2^{m''} - 1\}$ and $\tau\colon \{0, 1, \ldots, 2^{m'} - 1\} \to \mathcal{D}$. Hence, we compute f at $t_j = \tau(j)$ and then take the m'' most significant bits of $f(t_j)$ by $\beta(f(t_j))$, for the details and the possible use of ancillary qubits see [27].

At the end of the quantum algorithm, the final state $|\psi_f\rangle$ is measured. The measurement produces one of M outcomes, where $M \le 2^\nu$. Outcome $j \in \{0, 1, \ldots, M-1\}$ occurs with probability $p_f(j)$, which depends on j and the input f. Knowing the outcome j, we classically compute the final result $\phi_f(j)$ of the algorithm.

In principle, quantum algorithms may have measurements applied between sequences of unitary transformations of the form presented above. However, any algorithm with multiple measurements can be simulated by a quantum algorithm with only one measurement [8].

Let S be a linear or nonlinear operator such that

$$S\colon \mathcal{F} \to \mathcal{G} \,. \tag{4}$$

Typically, $\mathcal{F}$ is a linear space of real functions of several variables, and $\mathcal{G}$ is a normed linear space. We wish to approximate $S(f)$ to within ε for $f \in \mathcal{F}$. We approximate $S(f)$ using n function evaluations $f(t_1), \ldots, f(t_n)$ at deterministically and a priori chosen sample points. The quantum query Q_f encodes this information, and the quantum algorithm obtains this information from Q_f.

Without loss of generality, we consider algorithms that approximate $S(f)$ with probability $p \ge 3/4$. We can boost the success probability of an algorithm to become arbitrarily close to one by repeating the algorithm a number of times. The success probability becomes at least $1 - \delta$ with a number of repetitions proportional to $\log \delta^{-1}$.

The local error of the quantum algorithm (1) that computes the approximation $\phi_f(j)$, for $f \in \mathcal{F}$ and the outcome $j \in \{0, 1, \ldots, M-1\}$, is defined by

$$e(\phi_f, S) = \min\left\{\alpha\colon \sum_{j\colon \|S(f) - \phi_f(j)\| \le \alpha} p_f(j) \ge \frac{3}{4}\right\} \,, \tag{5}$$

where $p_f(j)$ denotes the probability of obtaining outcome j for the function f. The worst case error of a quantum algorithm ϕ is defined by

$$e^{\text{quant}}(\phi, S) = \sup_{f \in \mathcal{F}} e(\phi_f, S) \,. \tag{6}$$

The query complexity $\text{comp}^{\text{query}}(\varepsilon, S)$ of the problem S is the minimal number of queries necessary for approximating the solution with accuracy ε, i. e.,

$$\text{comp}^{\text{query}}(\varepsilon) = \min\{T\colon \exists\, \phi \text{ such that } e^{\text{quant}}(\phi, S) \le \varepsilon\} \,. \tag{7}$$

Similarly, the qubit complexity of the problem S is the minimal number of qubits necessary for approximating the solution with accuracy ε, i. e.,

$$\text{comp}^{\text{qubit}}(\varepsilon) = \min\{\nu\colon \exists\, \phi \text{ such that } e^{\text{quant}}(\phi, S) \le \varepsilon\} \,. \tag{8}$$

Integration

Integration is one of the most commonly occurring mathematical problems. One reason is that when one seeks the expectation of a continuous process one has to compute an integral. Often the integrals are over hundreds or thousands of variables. Path integrals are taken over an infinite number of variables. See Sect. "Path Integration".

Classical Computer

We begin with a one dimensional example to illustrate some basic concepts before moving to the d-dimen-

sional case (Number of dimensions and number of variables are used interchangeably.) Our simple example is to approximate

$$I(f) = \int_0^1 f(x)\,\mathrm{d}x\,.$$

For most integrands we can not use the fundamental theorem of calculus to compute the integral analytically; we have to approximate numerically (most real world continuous problems have to be approximated numerically). We have to make a promise about f. Assume

$$F_1 = \{f\colon [0,1] \to \mathbb{R} \mid \text{continuous and } |f(x)| \le 1,\ x \in [0,1]\}\,.$$

Use queries to compute

$$y_f = [f(x_1), \ldots, f(x_n)]\,.$$

We show that with this promise one cannot guarantee an ε-approximation on a classical computer. We use a simple adversary argument. Choose arbitrary numbers $x_1, \ldots, x_n \in [0,1]$. The adversary answers all these queries by answering $f(x_i) = 0$, $i = 1, \ldots, n$.

What is f? It could be $f_1 \equiv 0$ and $\int_0^1 f_1(x)\,\mathrm{d}x = 0$ or it could be the f_2 shown in Fig. 1. The value of $\int_0^1 f_2(x)\,\mathrm{d}x$ can be arbitrarily close to 1. Since $f_1(x)$ and $f_2(x)$ are indistinguishable with these query answers and this promise, it is impossible to guarantee an ε-approximation on a classical computer with $\varepsilon < 1/2$. We will return to this example in Sect. "Qubit Complexity".

We move on to the d-dimensional case. Assume that the integration limits are finite. For simplicity we assume we are integrating over the unit cube $[0,1]^d$. So our problem is

$$I(f) = \int_{[0,1]^d} f(x)\,\mathrm{d}x\,.$$

If our promise is that $f \in F_d$, where

$$F_d = \{f\colon [0,1]^d \to \mathbb{R} \mid \text{continuous and } |f(x)| \le 1,\ x \in [0,1]^d\},$$

then we can not compute an ε-approximation regardless of the value of d. Our promise has to be stronger. Assume that integrand class has *smoothness* r. There are various ways to define smoothness r for functions of d variables. See, for example, the definition on p. 25 in [67]. (For other definitions see [71].) With this definition Bakhvalov [4] showed that the query complexity is

$$\mathrm{comp}^{\mathrm{query}}_{\mathrm{clas-wor}}(\varepsilon) = \Theta\left(\varepsilon^{-d/r}\right). \tag{9}$$

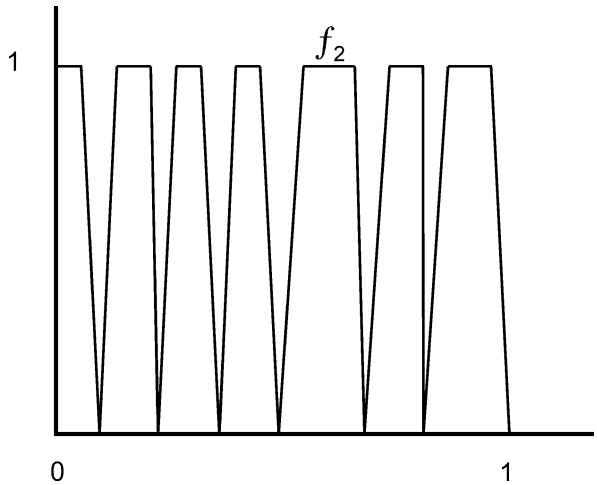

Quantum Algorithms and Complexity for Continuous Problems, Figure 1
All the function evaluations are equal to zero but the integral is close to one

This is the worst case query complexity on a classical computer.

What does this formula imply? If $r = 0$ the promise is that the functions are only continuous but have no smoothness. Then the query complexity is ∞, that is, we cannot guarantee an ε-approximation. If r and ε are fixed the query complexity is an exponential function of d. We say the problem is intractable. Following R. Bellman this is also called the *curse of dimensionality*. In particular, let $r = 1$. Then the promise is that the integrands have one derivative and the query complexity is $\Theta(\varepsilon^{-d})$.

The curse of dimensionality is present for many for continuous problems in the worst case setting. Breaking the curse is one of the central issues of information-based complexity. For high-dimensional integration the curse can be broken for F_d by the Monte Carlo algorithm which is a special case of a randomized algorithm. The Monte Carlo algorithm is defined by

$$\phi^{\mathrm{MC}}(f) = \frac{1}{n}\sum_{i=1}^{n} f(x_i)\,, \tag{10}$$

where the x_i are chosen independently from a uniform distribution. Then the expected error is

$$e^{\mathrm{MC}}(f) = \frac{\sqrt{\mathrm{var}(f)}}{\sqrt{n}}\,,$$

where

$$\mathrm{var}(f) = \int_{[0,1]^d} f^2(x)\,\mathrm{d}x - \left\{\int_{[0,1]^d} f(x)\,\mathrm{d}x\right\}^2$$

is the variance of f. For the promise $f \in F_d$ the query complexity is given by

$$\mathrm{comp}^{\mathrm{query}}_{\mathrm{clas-ran}}(\varepsilon) = \Theta(\varepsilon^{-2}) \,. \tag{11}$$

This is the randomized query complexity on a classical computer.

Thus Monte Carlo breaks the curse of dimensionality for the integration problem; the problem is tractable. Why should picking sample points at random be much better than picking them deterministically in the optimal way? This is possible because we have replaced the guarantee of the worst case setting by the stochastic assurance of the randomized setting. There is no free lunch!

The reader will note that (11) is a complexity result even though it is the cost of a particular algorithm, the Monte Carlo algorithm. The reason is that for integrands satisfying this promise Monte Carlo has been proven optimal. It is known that if the integrands are smoother Monte Carlo is not optimal; see p. 32 in [67].

In the worst case setting (deterministic) the query complexity is infinite if $f \in F_d$. In the randomized setting the query complexity is independent of d if $f \in F_d$. This will provide us guidance when we introduce Monte Carlo sampling into the model of quantum computation in Sect. "Qubit Complexity".

Generally pseudo-random numbers are used in the implementation of Monte Carlo on a classical computer. The quality of a pseudo-random number generator is usually tested statistically; see, for example, Knuth [41]. Will (11) still hold if a pseudo-random number generator is used? An affirmative answer is given by [69] provided some care is taken in the choice of the generator and f is Lipschitz.

Quantum Computer

We have seen that Monte Carlo breaks the curse of dimensionality for high-dimensional integration on a classical computer and that the query complexity is of order ε^{-2}. Can we do better on a quantum computer?

The short answer is yes. Under certain assumptions on the promises, which will be made precise below, the quantum query complexity enjoys exponential speedup over the classical worst case query complexity and quadratic speedup over the classical randomized query complexity. The latter is the same speedup as enjoyed by Grover's search algorithm of an unstructured database [26].

To show that the quantum query complexity is of order ε^{-1} we have to give matching, or close to matching upper and lower bounds. Usually, the upper bound is given by an algorithm, the lower bound by a theorem. The upper bound is given by the amplitude amplification algorithm of Brassard et al. [12] which we describe briefly.

The amplitude amplification algorithm of Brassard et al. computes the mean

$$\mathrm{SUM}(f) = \frac{1}{N}\sum_{i=0}^{N-1} f(i) \,,$$

of a Boolean function $f\colon \{0,1,\ldots,N-1\} \to \{0,1\}$, where $N = 2^k$, with error ε and probability at least $8/\pi^2$, using a number of queries proportional to $\min\{\varepsilon^{-1}, N\}$. Moreover, Nayak and Wu [47] show that the order of magnitude of the number of queries of this algorithm is optimal. Without loss of generality we can assume that $\varepsilon^{-1} \ll N$.

Perhaps the easiest way to understand the algorithm is to consider the operator

$$G = (I - |\psi\rangle\langle\psi|)O_f,$$

that is used in Grover's search algorithm. Here

$$O_f|x\rangle = (-1)^{f(x)}|x\rangle, \quad x \in \{0,1\}^k \,,$$

denotes the query, which is slightly different yet equivalent [48] to the one we defined in (2). Let

$$|\psi\rangle = \sqrt{\frac{1}{N}}\sum_x |x\rangle$$

be the equally weighted superposition of all the states.

If M denotes the number of assignments for which f has the value 1 then

$$\mathrm{SUM}(f) = \frac{M}{N} \,.$$

Without loss of generality $1 \le M \le N-1$. Consider the space $\mathcal{H}$ spanned by the states

$$|\psi_0\rangle = \sqrt{\frac{1}{N-M}}\sum_{x\colon f(x)=0} |x\rangle \quad \text{and}$$

$$|\psi_1\rangle = \sqrt{\frac{1}{M}}\sum_{x\colon f(x)=1} |x\rangle \,.$$

Then

$$|\psi\rangle = \cos(\theta/2)|\psi_0\rangle + \sin(\theta/2)|\psi_1\rangle$$

where $\sin(\theta/2) = \sqrt{M/N}$, and $\theta/2$ is the angle between the states $|\psi\rangle$ and $|\psi_0\rangle$. Thus

$$\sin^2\left(\frac{\theta}{2}\right) = \frac{M}{N} \,.$$

Now consider the operator G restricted to $\mathcal{H}$ which has the matrix representation

$$G = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} .$$

Its eigenvalues are $\lambda_\pm = \mathrm{e}^{\pm i\theta}$, $i = \sqrt{-1}$, and let $|\xi_\pm\rangle$ denote the corresponding eigenvectors.

We can express $|\psi\rangle$ using the $|\xi_\pm\rangle$ to get

$$|\psi\rangle = a|\xi_-\rangle + b|\xi_+\rangle ,$$

with $a, b \in \mathbb{C}$, $|a|^2 + |b|^2 = 1$. This implies that phase estimation [48] with G^p, $p = 2^j$, $j = 0, \ldots, t-1$, and initial state $|0\rangle^{\otimes t}|\psi\rangle$ can be used to approximate either θ or $2\pi - \theta$ with error proportional to 2^{-t}. Note that the first t qubits of the initial state determine the accuracy of phase estimation. Indeed, let $\tilde{\phi}$ be the result of phase estimation. Since $\sin^2(\theta/2) = \sin^2(\pi - \theta/2)$,

$$\left|\sin^2(\pi\tilde{\phi}) - \frac{N}{M}\right| = O\left(2^{-t}\right) ,$$

with probability at least $8/\pi^2$; see [12,48] for the details. Setting $t = \Theta(\log\varepsilon^{-1})$ and observing that phase estimation requires a number of applications of G (or queries O_f) proportional to ε^{-1} yields the result. (The complexity of quantum algorithms for the average case approximation of the Boolean mean has also been studied [35,52]).

For a real function $f\colon \{0, 1, \ldots, N-1\} \to [0, 1]$, we can approximate

$$\mathrm{SUM}(f) = \frac{1}{N}\sum_{i=0}^{N-1} f(i) ,$$

by reducing the problem to the computation of the Boolean mean. One way to derive this reduction is to truncate $f(i)$ to the desired number of significant bits, typically, polynomial in $\log\varepsilon^{-1}$, and then to use the bit representation of the function values to derive a Boolean function whose mean is equal to the mean of the truncated real function, see, e. g. [77]. The truncation of the function values is formally expressed through the mapping β in (3). Variations of this idea have been used in the literature [3,27,50].

Similarly, one discretizes the domain of a function $f\colon [0,1]^d \to [0,1]$ using the function τ in (3) and then uses the amplitude amplification algorithm to compute the average

$$\mathrm{SUM}(f) = \frac{1}{N}\sum_{i=0}^{N-1} f(x_i) , \tag{12}$$

for $x_i \in [0,1]^d$, $i = 0, \ldots, N-1$.

The quantum query complexity of computing the average (12) is of order ε^{-1}. On the other hand, the classical deterministic worst case query complexity is proportional to N (recall that $\varepsilon^{-1} \ll N$), and the classical randomized query complexity is proportional to ε^{-2}.

We now turn to the approximation of high-dimensional integrals and outline an algorithm for solving this problem. Suppose $f\colon [0,1]^d \to [0,1]$ is a function for which we are given some promise, for instance, that f has smoothness r. The algorithm integrating f with accuracy ε has two parts. First, using function evaluations $f(x_i)$, $i = 1, \ldots, n$, at deterministic points, it approximates f classically, by a function $\hat{f}$ with error ε_1, i. e.,

$$\|f - \hat{f}\| \le \varepsilon_1 ,$$

where $\|\cdot\|$ is the L_∞ norm. The complexity of this problem has been extensively studied and there are numerous results [49,67,71] specifying the optimal choice of n and the points x_i that must be used to achieve error ε_1. Thus

$$\int_{[0,1]^d} f(x)\,\mathrm{d}x = \int_{[0,1]^d} \hat{f}(x)\,\mathrm{d}x + \int_{[0,1]^d} g(x)\,\mathrm{d}x ,$$

where $g = f - \hat{f}$. Since $\hat{f}$ is known and depends linearly on the $f(x_i)$ the algorithm proceeds to integrate it exactly. So it suffices to approximate $\int_{[0,1]^d} g(x)\,\mathrm{d}x$ knowing that $\|g\| \le \varepsilon_1$.

The second part of the algorithm approximates the integral of g using the amplitude amplification algorithm to compute

$$\mathrm{SUM}(g) = \frac{1}{N}\sum_{i=0}^{N-1} g(y_i) ,$$

for certain points $y_i \in [0,1]^d$, with error ε_2. Once more, there are many results, see [67,71] for surveys specifying the optimal N and the points y_i, so that $\mathrm{SUM}(g)$ approximates $\int_{[0,1]^d} g(x)\,\mathrm{d}x$ with error ε_2. Finally, the algorithm approximates the original integral by the sum of the results of its two parts. The overall error of the algorithm is proportional to $\varepsilon = \varepsilon_1 + \varepsilon_2$.

Variations of the integration algorithm we described are known to have optimal query complexity for a number of different promises [27,29,34,50]. The quantum query complexity lower bounds for integration are based on the lower bounds of Nayak and Wu [47] for computing the Boolean mean. The quantum algorithms offer an exponential speedup over classical deterministic algorithms and a polynomial speedup over classical randomized algorithms for the query complexity of high-dimensional integration. Tabel 1 summarizes the query complexity (up to polylog factors) of high-dimensional integration in the

Quantum Algorithms and Complexity for Continuous Problems, Table 1
Query complexity of high-dimensional integration

	Worst case	Randomized	Quantum
$F_d^{r,\alpha}$	$\varepsilon^{-d/(r+\alpha)}$	$\varepsilon^{-2d/(2(r+\alpha)+d)}$	$\varepsilon^{-d/(r+\alpha+d)}$
$W_{p,d}^r$, $2 \le p \le \infty$	$\varepsilon^{-d/r}$	$\varepsilon^{-2d/(2r+d)}$	$\varepsilon^{-d/(r+d)}$
$W_{p,d}^r$, $1 \le p \le 2$	$\varepsilon^{-d/r}$	$\varepsilon^{-pd/(rp+pd-d)}$	$\varepsilon^{-d/(r+d)}$
$W_{1,d}^r$	$\varepsilon^{-d/r}$	$\varepsilon^{-d/r}$	$\varepsilon^{-d/(r+d)}$

worst case, randomized and quantum setting for functions belonging to Hölder classes $F_d^{r,\alpha}$ and Sobolev spaces $W_{p,d}^r$. Heinrich obtained most of the quantum query complexity results in a series of papers, which we cited earlier. Heinrich summarized his results in [28] where a corresponding table showing error bounds can be found.

Abrams and Williams [3] were the first to apply the amplitude amplification algorithm to high-dimensional integration. Novak [50] was the first to spell out his promises and thus obtained the first complexity results for high-dimensional integration.

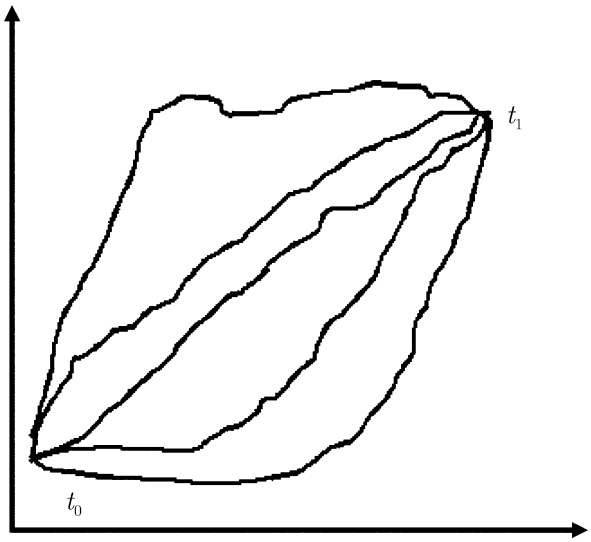

Quantum Algorithms and Complexity for Continuous Problems, Figure 2
Different trajectories of a particle

Path Integration

A path integral is defined as

$$I(f) = \int_X f(x)\,\mu(\mathrm{d}x)\,, \tag{13}$$

where μ is a probability measure on an infinite-dimensional space X. It can be viewed as an infinite-dimensional integral. For illustration we give an example due to R. Feynman. In classical mechanics a particle at a certain position at time t_0 has a unique trajectory to its position at time t_1. Quantum mechanically there are an infinite number of possible trajectories which Feynman called histories, see Fig. 2. Feynman summed over the histories. If one goes to the limit one gets a path integral. Setting $t_0 = 0$, $t_1 = 1$ this integral is

$$I(f) = \int_{C[0,1]} f(x)\,\mu(\mathrm{d}x)\,,$$

which is a special case of (13).

Path integration occurs in numerous applications including quantum mechanics, quantum chemistry, statistical mechanics, and mathematical finance.

Classical Computer

The first complexity analysis of path integration is due to Wasilkowski and Woźniakowski [72]; see also Chap. 5 in [67]. They studied the deterministic worst case and randomized settings and assume that μ is a Gaussian measure; an important special case of a Gaussian measure is a Wiener measure. They make the promise that F, the class of integrands, consists of functions $f\colon X \to \mathbb{R}$ whose rth Fréchet derivatives are continuous and uniformly bounded by unity. If r is finite then path integration is intractable. Curbera [20] showed that the worst case query complexity is of order $\varepsilon^{\varepsilon^{-\beta}}$ where β is a positive number depending on r.

Wasilkowski and Woźniakowski [72] also considered the promise that F consists of entire functions and that μ is the Wiener measure. Then the query complexity is a polynomial in ε^{-1}. More precisely, they provide an algorithm for calculating a worst case ε-approximation with cost of order ε^{-p} and the problem is tractable with this promise. The exponent p depends on the particular Gaussian measure; for the Wiener measure $p = 2/3$.

We return to the promise that the smoothness r is finite. Since this problem is intractable in the worst case, Wasilkowski and Woźniakowski [72] ask whether settling for a stochastic assurance will break the intractability. The obvious approach is to approximate the infinite-dimensional integral by a d-dimensional integral where d may be large (or even huge) since d is polynomial in ε^{-1}. Then Monte Carlo may be used since its speed of convergence does not depend on d. Modulo an assumption that the nth eigenvalue of the covariance operator of μ does not decrease too fast the randomized query complexity is roughly ε^{-2}. Thus Monte Carlo is optimal.

Quantum Computer

Just as with finite dimensional integration Monte Carlo makes path integration tractable on a classical computer and the query complexity is of order ε^{-2}. Again we ask whether we can do better on a quantum computer and again the short answer is yes. Under certain assumptions on the promises, which will be made precise below, the quantum query complexity is of order ε^{-1}. Thus quantum query complexity enjoys exponential speedup over the classical worst case query complexity and quadratic speedup over the classical randomized query complexity. Again the latter is the same speedup as enjoyed by Grover's search algorithm of an unstructured database.

The idea for solving path integration on a quantum computer is fairly simple but the analysis is not so easy. So we outline the algorithm without considering the details. We start with a classical deterministic algorithm that uses an average as in (12) to approximate the path integral $I(f)$ with error ε in the worst case. The number of terms N of the this average is an exponential function of ε^{-1}. Nevertheless, on a quantum computer we can approximate the average, using the amplitude amplification algorithm as we discussed in Sect. "Integration (Quantum Computer)", with cost that depends on $\log N$ and is, therefore, a polynomial in ε^{-1}, Sect. 6 in [70].

A summary of the promises and the results in [70] follows. The measure μ is Gaussian and the eigenvalues of its covariance operator is of order j^{-h}, $h > 1$. For the Wiener measure occurring in many applications we have $h = 2$. The class of integrands consists of functions f whose rth Fréchet derivatives are continuous and uniformly bounded by unity. In particular assume the integrands are at least Lipschitz. Then

- Path integration on a quantum computer is tractable.
- Query complexity on a quantum computer enjoys exponential speedup over the worst case and quadratic speedup over the classical randomized query complexity. More precisely, the number of quantum queries is at most $4.22\varepsilon^{-1}$.

Results on the qubit complexity of path integration will be given in Sect. "Qubit Complexity". Details of an algorithm for computing an ε-approximation to a path integral on a quantum computer may be found in Sect. 6 in [70].

Feynman–Kac Path Integration

An important special case of a path integral is a Feynman–Kac path integral. Assume that X is the space C of continuous functions and that the measure μ is the Wiener measure w. Feynman–Kac path integrals occur in many applications; see [23]. For example consider the diffusion equation

$$\begin{aligned}\frac{\partial z}{\partial t}(u,t) &= \frac{1}{2}\frac{\partial^2 z}{\partial u^2}(u,t) + V(u)\,z(u,t)\\ z(u,0) &= v(u)\,,\end{aligned}$$

where $u \in \mathbb{R}$, $t > 0$, V is a potential function, and v is an initial condition function. Under mild conditions on v and V the solution is given by the Feynman–Kac path integral

$$z(u,t) = \int_C v(x(t)+u)\mathrm{e}^{\int_0^t V(x(s)+u)\,\mathrm{d}s}\, w(\mathrm{d}x)\,. \qquad (14)$$

The problem generalizes to the multivariate case by considering the diffusion equation

$$\begin{aligned}\frac{\partial z}{\partial t}(u,t) &= \frac{1}{2}\Delta z(u,t) + V(u)\,z(u,t)\\ z(u,0) &= v(u)\,,\end{aligned}$$

with $u \in \mathbb{R}^d$, $t > 0$, and $V, v\colon \mathbb{R}^d \to \mathbb{R}$, the potential and the initial value function, respectively. As usual, Δ denotes the Laplacian. The solution is given by the Feynman–Kac path integral

$$z(u,t) = \int_C v(x(t)+u)\mathrm{e}^{\int_0^t V(x(s)+u)\,\mathrm{d}s}\, w(\mathrm{d}x)\,,$$

where C is the set of continuous functions $x\colon \mathbb{R}_+ \to \mathbb{R}^d$ such that $x(0) = 0$.

Note that there are two kinds of dimension here. A Feynman–Kac path integral is infinite dimensional since we are integrating over continuous functions. Furthermore u is a function of d variables.

Classical Computer

We begin with the case when u is a scalar and then move to the case where u is a multivariate function. There have been a number of papers on the numerical solution of (14); see, for example [14].

The usual attack is to solve the problem with a stochastic assurance using randomization. For simplicity we make the promise that $u = 1$ and V is four times continuously differentiable. Then by Chorin's algorithm [16], the total cost is of order $\varepsilon^{-2.5}$.

The first complexity analysis may be found in Plaskota et al. [58] where a new algorithm is defined which enjoys certain optimality properties. They construct an algorithm which computes an ε-approximation at cost of order $\varepsilon^{-.25}$ and show that the worst case complexity is of the same order. Hence the exponent of ε^{-1} is an order of magnitude smaller and with a worst case rather than a stochastic guarantee. However, the new algorithm requires a nu-

merically difficult precomputation which may limit its applicability.

We next report on multivariate Feynman–Kac path integration. First consider the worst case setting with the promise that v and V are r times continuously differentiable with r finite. Kwas and Li [43] proved that the query complexity is of order $\varepsilon^{-d/r}$. Therefore in the worst case setting the problem suffers the curse of dimensionality.

The randomized setting was studied by Kwas [42]. He showed that the curse of dimensionality is broken by using Monte Carlo using a number of queries of order ε^{-2}. The number of queries can be further improved to $\varepsilon^{-2/(1+2r/d)}$, which is the optimal number of queries, by a bit more complicated algorithm. The randomized algorithms require extensive precomputing [42,43].

Quantum Computer

Multivariate multivariate Feynman–Kac path integration on a quantum computer was studied in [42]. With the promise as in the worst and randomized case, Kwas presents an algorithm and complexity analysis. He exhibits a quantum algorithm that uses a number of queries of order ε^{-1} that is based on the Monte Carlo algorithm. He shows that the query complexity is of order $\varepsilon^{-1/(1+r/d)}$ and is achieved by a bit more complicated quantum algorithm. Just as in the randomized case the quantum algorithms require extensive precomputing [42,43].

Eigenvalue Approximation

Eigenvalue problems for differential operators arising in physics and engineering have been extensively studied in the literature; see, e. g. [5,18,19,22,25,40,63,65]. Typically, the mathematical properties of the eigenvalues and the corresponding eigenfunctions are known and so are numerical algorithms approximating them on a classical computer. Nevertheless, the complexity of approximating eigenvalues in the worst, randomized and quantum settings has only recently been addressed for the Sturm–Liouville eigenvalue problem [53,55] (see also [10] for quantum lower bounds with a different kind of query than the one we discuss in this article). In some cases we have sharp complexity estimates but there are important questions that remain open.

Most of our discussion here concerns the complexity of approximating the smallest eigenvalue of a Sturm–Liouville eigenvalue problem. We will conclude this section by briefly addressing quantum algorithms for other eigenvalue problems.

In the physics literature this problem is called the time-independent Schrödinger equation. The smallest eigenvalue is the energy of the ground state. In the mathematics literature it is called the Sturm–Liouville eigenvalue problem.

Let $I_d = [0,1]^d$ and consider the class of functions

$$\mathbf{Q} = \left\{ q\colon I_d \to [0,1] \,\middle|\, q\,,\ D_j q\colon = \frac{\partial q}{\partial x_j} \in C(I_d)\,,\ \|D_j q\|_\infty \le 1\,,\ \|q\|_\infty \le 1 \right\}\,,$$

where $\|\cdot\|_\infty$ denotes the supremum norm. For $q \in \mathbf{Q}$, define $\mathbb{L}_q\colon = -\Delta + q$, where $\Delta = \sum_{j=1}^d \partial^2/\partial x_j^2$ is the Laplacian, and consider the eigenvalue problem

$$\mathbb{L}_q u = \lambda u,\ x \in (0,1)^d\,, \tag{15}$$

$$u(x) \equiv 0,\ x \in \partial I_d\,. \tag{16}$$

In the variational form, the smallest eigenvalue $\lambda = \lambda(q)$ of (15), (16) is given by

$$\lambda(q) = \min_{0 \ne u \in H_0^1} \frac{\int_{I_d} \sum_{j=1}^d [D_j u(x)]^2 + q(x) u^2(x)\,\mathrm{d}x}{\int_{I_d} u^2(x)\,\mathrm{d}x}\,, \tag{17}$$

where H_0^1 is the space of all functions vanishing on the boundary of I_d having square integrable first order partial derivatives. We consider the complexity of classical and quantum algorithms approximating $\lambda(q)$ with error ε.

Classical Computer

In the worst case we discretize the differential operator on a grid of size $h = \Theta(\varepsilon^{-1})$ and obtain a matrix $M_\varepsilon = -\Delta_\varepsilon + B_\varepsilon$, of size proportional to $\varepsilon^{-d} \times \varepsilon^{-d}$. The matrix Δ_ε is the $(2d+1)$-point finite difference discretization of the Laplacian. The matrix B_ε is a diagonal matrix containing evaluations of q at the grid points. The smallest eigenvalue of M_ε approximates $\lambda(q)$ with error $O(\varepsilon)$ [73,74]. We compute the smallest eigenvalue of M_ε using the bisection method (p. 228 in [22]). The resulting algorithm uses a number of queries proportional to ε^{-d}. It turns out that this number of queries is optimal in the worst case, and the problem suffers from the curse of dimensionality.

The query complexity lower bounds are obtained by reducing the eigenvalue problem to high-dimensional integration. For this we use the perturbation formula [53,55]

$$\lambda(q) = \lambda(\bar{q}) + \int_{I_d} \big(q(x) - \bar{q}(x)\big)\, u_{\bar{q}}^2(x)\,\mathrm{d}x + O\left(\|q - \bar{q}\|_\infty^2\right), \tag{18}$$

where $q, \bar{q} \in \mathbf{Q}$ and $u_{\bar{q}}$ is the normalized eigenfunction corresponding to $\lambda(\bar{q})$.

The same formula is used for lower bounds in the randomized setting. Namely, the query complexity is

$$\Omega\left(\varepsilon^{-2d/(d+2)}\right) .$$

Moreover, we can use (18) to derive a randomized algorithm. First we approximate q by a function $\bar{q}$ and then approximate $\lambda(q)$ by approximating the first two terms on the right hand side of (18); see Papageorgiou and Woźniakowski [55] for $d = 1$, and Papageorgiou [53] for general d. However, this algorithm uses

$$O\left(\varepsilon^{-\max(2/3,d/2)}\right) ,$$

queries. So, it is optimal only when $d \leq 2$. Determining the randomized complexity for $d > 2$ and determining if the randomized complexity is an exponential function of d are important open questions.

Quantum Computer

The perturbation formula (18) can be used to show that the quantum query complexity is

$$\Omega\left(\varepsilon^{-d/(d+1)}\right) .$$

As in the randomized case, we can use (18) to derive an algorithm that uses $O(\varepsilon^{-d/2})$ quantum queries. The difference between the quantum algorithm and the randomized algorithm is that the former uses the amplitude amplification algorithm to approximate the integral on the right hand side of (18) instead of Monte Carlo. The algorithm is optimal only when $d = 1$ [53,55].

For general d the query complexity is not known exactly, we only have the upper bound $O(\varepsilon^{-p})$, $p \leq 6$. The best quantum algorithm known is based on phase estimation. In particular, we discretize the problem as in the worst case and apply phase estimation to the unitary matrix

$$e^{i\gamma M_\varepsilon} ,$$

where γ is chosen appropriately so that the resulting phase belongs to $[0, 2\pi)$. We use a splitting formula to approximate the necessary powers of the matrix exponential. The largest eigenvalue of Δ_ε is $O(\varepsilon^{-2})$ and $\|q\|_\infty \leq 1$. This implies that the resulting number of queries does not grow exponentially with d.

Finally, there are a number of papers providing quantum algorithms for eigenvalue approximation without carrying out a complete complexity analysis. Abrams and Lloyd [2] have written an influential paper on eigenvalue approximation of a quantum mechanical system evolving with a given Hamiltonian. They point out that phase estimation, which requires the corresponding eigenvector as part of its initial state, can also be used with an approximate eigenvector. Jaksch and Papageorgiou [36] give a quantum algorithm which computes a *good* approximation of the eigenvector at low cost. Their method can be generally applied to the solution of continuous Hermitian eigenproblems on a discrete grid. It starts with a classically obtained eigenvector for a problem discretized on a coarse grid and constructs an approximate eigenvector on a fine grid.

We describe this algorithm briefly for the case $d = 1$. Suppose $N = 2^k$ and $N_0 = 2^{k_0}$ are the number of points in the fine and coarse grid, respectively. Given the eigenvector for the coarse grid $|U^{(N_0)}\rangle$, we approximate the eigenvector for the fine grid $|U^{(N)}\rangle$ by

$$\left|\tilde{U}^{(N)}\right\rangle = \left|U^{(N_0)}\right\rangle \left(\frac{|0\rangle + |1\rangle}{\sqrt{2}}\right)^{\otimes(k-k_0)} .$$

The effect of this transformation is to replicate the coordinates of $|U^{(N_0)}\rangle 2^{k-k_0}$ times. The resulting approximation is good enough in the sense that the success probability of phase estimation with initial state $|\tilde{U}_h\rangle$ is greater than 1/2.

Szkopek et al. [64] use the algorithm of Jaksch and Papageorgiou in the approximation of low order eigenvalues of a differential operator of order $2s$ in d dimensions. Their paper provides an algorithm with cost polynomial in ε^{-1} and generalizes the results of Abrams and Lloyd [2]. However, [64] does not carry out a complexity analysis but only considers known classical algorithms in the worst case for comparison.

Qubit Complexity

For the foreseeable future the number of qubits will be a crucial computational resource. We give a general lower bound on the qubit complexity of continuous problems.

Recall that in (1) we defined a quantum algorithm as

$$|\psi_f\rangle = U_T Q_f U_{T-1} Q_f \ldots U_1 Q_f U_0 |\psi_0\rangle .$$

where $|\psi_0\rangle$ and $|\psi_f\rangle$ are the initial and final state vectors, respectively. They are column vectors of length 2^ν. The query Q_f, a $2^\nu \times 2^\nu$ unitary matrix, depends on the values of f at $n \leq 2^\nu$ deterministic points. That is

$$Q_f = Q_f(f(x_1), \ldots, f(x_n)) .$$

It is important to note that in the standard model the evolution is completely deterministic. The only probabilistic element is in the measurement of the final state.

For the qubit complexity (8) we have the following lower bound

$$\mathrm{comp}_{\mathrm{std}}^{\mathrm{qubit}}(\varepsilon, S) = \Omega\left(\log \mathrm{comp}_{\mathrm{clas}}^{\mathrm{query}}(2\varepsilon, S)\right) . \tag{19}$$

Here S specifies the problem, see (4), and $\mathrm{comp}_{\mathrm{std}}^{\mathrm{qubit}}(\varepsilon, S)$ is the qubit complexity in the standard quantum setting. On the right hand side $\mathrm{comp}_{\mathrm{clas}}^{\mathrm{query}}(\varepsilon, S)$ is the query complexity on a classical computer in the worst case setting. See [77] for details.

We provide an intuition about (19). Assume 2^ν function evaluations are needed to solve the problem specified by S to within ε. Note that ν qubits are needed to generate the Hilbert space $\mathcal{H}_\nu$ of size 2^ν to store the evaluations.

Equation (19) can be interpreted as a certain limitation of the standard quantum setting, which considers queries using function evaluations at deterministic points. We will show why this is a limitation and show we can do better.

Consider multivariate integration which was discussed in Sect. "Integration". We seek to approximate the solution of

$$S(f) = \int_{0,1]^d} f(x)\,\mathrm{d}x .$$

Assume our promise is $f \in F_0$ where

$$F_d = \Big\{ f\colon [0,1]^d \to \mathbb{R} \mid \text{continuous and } |f(x)| \le 1, x \in [0,1]^d \Big\} .$$

For the moment let $d = 1$. We showed that with this promise we cannot guarantee an ε-approximation on a classical computer with $\varepsilon < 1/2$. That is,

$$\mathrm{comp}_{\mathrm{clas}}^{\mathrm{query}}(\varepsilon, S) = \infty .$$

By (19)

$$\mathrm{comp}_{\mathrm{std}}^{\mathrm{qubit}}(\varepsilon, S) = \infty .$$

If the qubit complexity is infinite even for $d = 1$ its certainly infinite for general d. But we saw that if classical randomization (Monte Carlo) is used

$$\mathrm{comp}_{\mathrm{clas-ran}}^{\mathrm{query}}(\varepsilon, S) = \Theta(\varepsilon^{-2}) .$$

Thus we have identified a problem which is easy to solve on a classical computer and is impossible to solve on a quantum computer using the standard formulation of a quantum algorithm (1). Our example motivates extending the notion of quantum algorithm by permitting *randomized queries*. The quantum setting with randomized queries was introduced by Woźniakowski [77]. The idea of using randomized queries is not entirely new. Shor's algorithm [60] uses a special kind of randomized query, namely,

$$Q_x|x\rangle = |jx \bmod N\rangle ,$$

with $j = 0, \ldots, N-1$ and a random x from the set $\{2, 3, \ldots, N-1\}$.

In our case, the randomization affects only the selection of sample points and the number of queries. It occurs prior to the implementation of the queries and the execution of the quantum algorithm. In this extended setting, we define a quantum algorithm as

$$|\psi_{f,\omega}\rangle = U_{T_\omega} Q_{f,\omega} U_{T_\omega - 1} Q_{f,\omega} \ldots U_1 Q_{f,\omega} U_0 |\psi_0\rangle , \tag{20}$$

where ω is a random variable and

$$Q_{f,\omega} = Q_{f,\omega}(f(x_{1,\omega}), \ldots, f(x_{n,\omega})) ,$$

and the $x_{j,\omega}$ are random points. The number of queries T_ω and the points $x_{j,\omega}$ are chosen at random initially and then remain fixed for the remainder of the computation. Note that (20) is identical to (1) except for the introduction of randomization. Randomized queries require that we modify the criterion (6) by which we measure the error of an algorithm. One possibility is to consider the expected error and another possibility is to consider the probabilistic error with respect to the random variable ω. Both cases have been considered in the literature [77] but we will avoid the details here because they are rather technical.

A test of the new setting is whether it buys us anything. We compare the qubit complexity of the standard and randomized settings for integration and path integration.

Integration

We make the same promise as above, namely, $f \in F_d$.

- Quantum setting with deterministic queries. We remind the reader that

$$\mathrm{comp}_{\mathrm{std}}^{\mathrm{query}}(\varepsilon) = \infty$$
$$\mathrm{comp}_{\mathrm{std}}^{\mathrm{qubit}}(\varepsilon) = \infty .$$

- Quantum setting with randomized queries. Then [77]

$$\mathrm{comp}_{\mathrm{ran}}^{\mathrm{query}}(\varepsilon) = \Theta\left(\varepsilon^{-1}\right)$$
$$\mathrm{comp}_{\mathrm{ran}}^{\mathrm{qubit}}(\varepsilon) = \Theta\left(\log \varepsilon^{-1}\right) .$$

  Therefore, there is infinite improvement in the randomized quantum setting over the standard quantum setting.

Path Integration

- Quantum setting with deterministic queries. With an appropriate promise it was shown [77] that

$$\mathrm{comp}_{\mathrm{std}}^{\mathrm{query}}(\varepsilon) = \Theta\left(\varepsilon^{-1}\right)$$
$$\mathrm{comp}_{\mathrm{std}}^{\mathrm{qubit}}(\varepsilon) = \Theta\left(\varepsilon^{-2}\log\varepsilon^{-1}\right) .$$

  Thus, modulo the log factor, the qubit complexity of path integration is a second degree polynomial in ε^{-1}. That seems pretty good but we probably will not have enough qubits for a long time to do new science, especially with error correction.
- Quantum setting with randomized queries. Then [77]

$$\mathrm{comp}_{\mathrm{ran}}^{\mathrm{query}}(\varepsilon) = \Theta\left(\varepsilon^{-1}\right)$$
$$\mathrm{comp}_{\mathrm{ran}}^{\mathrm{qubit}}(\varepsilon) = \Theta\left(\log\varepsilon^{-1}\right) .$$

  Thus there is an exponential improvement in the randomized quantum setting over the standard quantum setting.

As the analogue of (19) we have the following lower bound on the randomized qubit complexity [77]

$$\mathrm{comp}_{\mathrm{ran}}^{\mathrm{qubit}}(\varepsilon, S) = \Omega\left(\log \mathrm{comp}_{\mathrm{clas-ran}}^{\mathrm{query}}(\varepsilon, S)\right) , \qquad (21)$$

where $\mathrm{comp}_{\mathrm{clas-ran}}^{\mathrm{query}}(\varepsilon, S)$ is the query complexity on a classical computer in the randomized setting.

Approximation

Approximating functions of d variables is a fundamental and generally hard problem. Typically, the literature considers the approximation of functions that belong to the Sobolev space $W_p^r([0,1]^d)$ in the norm of $L_q([0,1]^d)$. The condition $r/d > 1/p$ ensures that functions in $W_p^r([0,1]^d)$ are continuous, which is necessary for function values to be well defined. Thus, when $p = \infty$ the dimension d can be arbitrarily large while the smoothness r can be fixed, which cannot happen when $p < \infty$.

For $p = \infty$ the problem suffers the curse of dimensionality in the worst and the randomized classical cases [49,71]. Recently, Heinrich [30] showed that quantum computers do not offer any advantage relative to classical computers since the problem remains intractable in the quantum setting.

For different values of the parameters p, q, r, d the classical and quantum complexities are also known [30, 49,71]. In some cases quantum computers can provide a roughly quadratic speedup over classical computers, but there are also cases where the classical and quantum complexities coincide. Table 2 summarizes the order of the query complexity (up to polylog factors) of approximation in the worst case, randomized and quantum setting, and is based on a similar table in [30] describing error bounds.

Elliptic Partial Differential Equations

Elliptic partial differential equations have many important applications and have been extensively studied in the literature, see [75] and the references therein. A simple example is the Poisson equation, for which we want to find a function $u\colon \bar{\Omega} \to \mathbb{R}$, that satisfies

$$-\Delta u(x) = f(x) , \quad x \in \Omega$$
$$u(x) = 0 , \quad x \in \partial\Omega ,$$

where $\Omega \subset \mathbb{R}^d$,

More generally we consider elliptic partial differential equations of order $2m$ on a smooth bounded domain $\Omega \subset \mathbb{R}^d$ with smooth coefficients and homogeneous boundary conditions with the right hand side belonging to $C^r(\Omega)$ and the error measured in the L_∞ norm; see [32] for details.

In the worst case the complexity is proportional to $\varepsilon^{-d/r}$ [75] and the problem is intractable. The randomized complexity of this problem was only recently studied along with the quantum complexity by Heinrich [31,32]. In particular, the randomized query complexity (up to polylog factors) is proportional to

$$\varepsilon^{-\max\{d/(r+2m),\, 2d/(2r+d)\}} ,$$

and the quantum query complexity is proportional to

$$\varepsilon^{-\max\{d/(r+2m),\, d/(r+d)\}} .$$

Quantum Algorithms and Complexity for Continuous Problems, Table 2
Query complexity of approximation

	Worst case	Randomized	Quantum
$1 \le p < q \le \infty, \quad r/d \ge 2/p - 2/q$	$\varepsilon^{-dpq/(rpq-d(q-p))}$	$\varepsilon^{-dpq/(rpq-d(q-p))}$	$\varepsilon^{-d/r}$
$1 \le p < q \le \infty, \quad r/d < 2/p - 2/q$	$\varepsilon^{-dpq/(rpq-d(q-p))}$	$\varepsilon^{-dpq/(rpq-d(q-p))}$	$\varepsilon^{-dpq/(2rpq-2d(q-p))}$
$1 \le q \le p \le \infty$	$\varepsilon^{-d/r}$	$\varepsilon^{-d/r}$	$\varepsilon^{-d/r}$

Thus the quantum setting may provide a polynomial speedup over the classical randomized setting but not always. Moreover, for fixed m and r and for $d > 4m$ the problem is intractable in all three settings.

Ordinary Differential Equations

In this section we consider the solution of a system of ordinary differential equations with initial conditions

$$z'(t) = f(z(t)), \quad t \in [a,b], \quad z(a) = \eta\,,$$

where $f\colon \mathbb{R}^d \to \mathbb{R}^d$, $z\colon [a,b] \to \mathbb{R}^d$ and $\eta \in \mathbb{R}^d$ with $f(\eta) \neq 0$. For the right hand side function $f = [f_1, \ldots, f_d]$, where $f_j\colon \mathbb{R}^d \to \mathbb{R}$, we assume that the f_j belong to the Hölder class $F_d^{r,\alpha}$, $r + \alpha \geq 1$. We seek to compute a bounded function on the interval $[a,b]$ that approximates the solution z.

Kacewicz [38] studied the classical worst case complexity of this problem and found it to be proportional to $\varepsilon^{-1/(r+\alpha)}$. Recently he also studied the classical randomized and quantum complexity of the problem and derived algorithms that yield upper bounds that from the lower bounds by only an arbitrarily small positive parameter in the exponent [39]. The resulting randomized and quantum complexity bounds (up to polylog factors) are

$$O\big(\varepsilon^{-1/(r+\alpha+1/2-\gamma)}\big)$$

and

$$O\big(\varepsilon^{-1/(r+\alpha+1-\gamma)}\big)\,,$$

where $\gamma \in (0,1)$ is arbitrarily small, respectively. Observe that the randomized and quantum complexities (up to polylog factors) satisfy

$$\Omega\big(\varepsilon^{-1/(r+\alpha+1/2)}\big)$$

and

$$\Omega\big(\varepsilon^{-1/(r+\alpha+1)}\big)\,,$$

respectively. Even more recently, Heinrich and Milla [33] showed that the upper bound for the randomized complexity holds with $\gamma = 0$, thereby establishing tight upper and lower randomized complexity bounds, up to polylog factors.

Once more, the quantum setting provides a polynomial speedup over the classical setting.

Gradient Estimation

Approximating the gradient of a function $f\colon \mathbb{R}^d \to \mathbb{R}$ with accuracy ε requires a minimum of $d+1$ function evaluations on a classical computer. Jordan [37] shows how this can be done using a single query on a quantum computer.

We present Jordan's algorithm for the special case where the function is a plane passing through the origin, i. e., $f(x_1, \ldots, x_d) = \sum_{j=1}^d a_j x_j$, and is uniformly bounded by 1. Then $\nabla f = (a_1, \ldots, a_d)^T$. Using a single query and *phase kickback* we obtain the state

$$\frac{1}{\sqrt{N^d}} \sum_{j_1=0}^{N-1} \cdots \sum_{j_d=0}^{N-1} \mathrm{e}^{2\pi i f(j_1,\ldots,j_d)} |j_1\rangle \cdots |j_d\rangle\,,$$

where N is a power of 2. Equivalently, we have

$$\frac{1}{\sqrt{N^d}} \sum_{j_1=0}^{N-1} \cdots \sum_{j_d=0}^{N-1} \mathrm{e}^{2\pi i (a_1 j_1 + \cdots + a_d j_d)} |j_1\rangle \cdots |j_d\rangle\,.$$

This is equal to the state

$$\frac{1}{\sqrt{N}} \sum_{j_1=0}^{N-1} \mathrm{e}^{2\pi i a_1 j_1} |j_1\rangle \ldots \frac{1}{\sqrt{N}} \sum_{j_d=0}^{N-1} \mathrm{e}^{2\pi i a_d j_d} |j_d\rangle\,.$$

We apply the Fourier transform to each of the d registers and then measure each register in the computational basis to obtain $m_1, \ldots, m_d$. If a_j can be represented with finitely many bits and N is sufficiently large then $m_j/N = a_j$, $j = 1, \ldots, d$.

For functions with second order partial derivatives not identically equal to zero the analysis is more complicated and we refer the reader to [37] for the details.

Simulation of Quantum Systems on Quantum Computers

So far this article has been devoted to work on algorithms and complexity of problems where the query and qubit complexities are known or have been studied. In a number of cases, the classical complexity of these problems is also known and we know the quantum computing speedup.

The notion that quantum systems could be simulated more efficiently by quantum computers than by classical computers was first mentioned by Manin [44], see also [45], and discussed thoroughly by Feynman [24].

There is a large and varied literature on simulation of quantum systems on quantum computers. The focus in these papers is typically on the cost of particular quantum and classical algorithms without complexity analysis and therefore without speedup results. To give the reader a taste of this area we list some sample papers:

- Berry et al. [9] present an efficient quantum algorithm for simulating the evolution of a sparse Hamiltonian.

- Dawson, Eisert and Osborne [21] introduce a unified formulation of variational methods for simulating ground state properties of quantum many-body systems.
- Morita and Nishimori [46] derive convergence conditions for the quantum annealing algorithm.
- Brown, Clark, Chuang [13] establish limits of quantum simulation when applied to specific problems.
- Chen, Yepez and Cory [15] report on the simulation of Burgers equation as a type-II quantum computation.
- Paredes, Verstraete and Cirac [56] present an algorithm that exploits quantum parallelism to simulate randomness.
- Somma et al. [61] discuss what type of physical problems can be efficiently simulated on a quantum computer which cannot be simulated on a Turing machine.
- Yepez [78] presents an efficient algorithm for the many-body three-dimensional Dirac equation.
- Nielsen and Chuang [48] discuss simulation of a variety of quantum systems.
- Sornborger and Stewart [62] develop higher order methods for simulations.
- Boghosian and Taylor [11] present algorithms for efficiently simulating quantum mechanical systems.
- Zalka [79] shows that the time evolution of the wave function of a quantum mechanical many particle system can be simulated efficiently.
- Abrams and Lloyd [1] provide fast algorithms for simulating many-body Fermi systems.
- Wisner [76] provides two quantum many-body problems whose solution is intractable on a classical computer.

Future Directions

The reason there is so much interest in quantum computers is to solve important problems fast. The belief is that we will be able to solve scientific problems, and in particular quantum mechanical systems, which cannot be solved on a classical computer. That is, that quantum computation will lead to new science.

Research consists of two major parts. The first is identification of important scientific problems with substantial speedup. The second is the construction of machines with sufficient number of qubits and long enough decoherence times to solve the problems identified in the first part. Abrams and Lloyd [2] have argued that with 50 to 100 qubits we can solve interesting classically intractable problems from atomic physics. Of course this does not include qubits needed for fault tolerant computation.

There are numerous important open questions. We will limit ourselves here to some open questions regarding the problems discussed in this article.

1. In Sect. "Path Integration" we reported big wins for the qubit complexity for integration and path integration. Are there big wins for other problems?
2. Are there problems for which we get big wins for query complexity using randomized queries?
3. Are there tradeoffs between the query complexity and the qubit complexity?
4. What are the classical and quantum complexities of approximating the solution of the Schrödinger equation for a many-particle system? How do they it depend on the number of particles? What is the quantum speedup?

Acknowledgments

We are grateful to Erich Novak, University of Jena, and Henryk Woźniakowski, Columbia University and University of Warsaw, for their very helpful comments. We thank Jason Petras, Columbia University, for checking the complexity estimates appearing in the tables.

Bibliography

1. Abrams DS, Lloyd S (1997) Simulation of Many-Body Fermi Systems on a Universal Quantum Computer. Phys Rev Lett 79(13):2586–2589; http://arXiv.org/quant-ph/9703054
2. Abrams DS, Lloyd S (1999) Quantum Algorithm Providing Exponential Speed Increase for Finding Eigenvalues and Eigenvectors. Phys Rev Lett 83:5162–5165
3. Abrams DS, Williams CP (1999) Fast quantum algorithms for numerical integrals and stochastic processes. http://arXiv.org/quant-ph/9908083
4. Bakhvalov NS (1977) Numerical Methods. Mir Publishers, Moscow
5. Babuska I, Osborn J (1991) Eigenvalue Problems. In: Ciarlet PG, Lions JL (eds) Handbook of Numerical Analysis, vol II. North-Holland, Amsterdam, pp 641–787
6. Beals R, Buhrman H, Cleve R, Mosca R, de Wolf R (1998) Quantum lower bounds by polynomials. Proceedings FOCS'98, pp 352–361. http://arXiv.org/quant-ph/9802049
7. Bennett CH, Bernstein E, Brassard G, Vazirani U (1997) Strengths and weaknesses of quantum computing. SIAM J Comput 26(5):1510–1523
8. Bernstein E, Vazirani U (1997) Quantum complexity theory. SIAM J Comput 26(5):1411–1473
9. Berry DW, Ahokas G, Cleve R, Sanders BC (2007) Efficient quantum algorithms for simulating sparse Hamiltonians. Commun Math Phys 270(2):359–371; http://arXiv.org/quant-ph/0508139
10. Bessen AJ (2007) On the complexity of classical and quantum algorithms for numerical problems in quantum mechanics. Ph D thesis. Department of Computer Science, Columbia University

11. Boghosian BM, Taylor W (1998) Simulating quantum mechanics on a quantum computer. Physica D 120:30–42 http://arXiv.org/quant-ph/9701019
12. Brassard G, Hoyer P, Mosca M, Tapp A (2002) Quantum Amplitude Amplification and Estimation. Contemporary Mathematics, vol 305. Am Math Soc, Providence, pp 53–74. http://arXiv.org/quant-ph/0005055
13. Brown KR, Clark RJ, Chuang IL (2006) Limitations of Quantum Simulation Examined by Simulating a Pairing Hamiltonian using Magnetic Resonance. Phys Rev Lett 97(5):050504; http://arXiv.org/quant-ph/0601021
14. Cameron RH (1951) A Simpson's rule for the numerical evaluation of Wiener's integrals in function space. Duke Math J 8:111–130
15. Chen Z, Yepez J, Cory DG (2006) Simulation of the Burgers equation by NMR quantum information processing. Phys Rev A 74:042321; http://arXiv.org/quant-ph/0410198
16. Chorin AJ (1973) Accurate evaluation of Wiener integrals. Math Comp 27:1–15
17. Cleve R, Ekert A, Macchiavello C, Mosca M (1996) Quantum Algorithms Revisited. Proc R Soc Lond A 454(1969):339–354
18. Collatz L (1960) The Numerical Treatment of Differential Equations. Springer, Berlin
19. Courant C, Hilbert D (1989) Methods of Mathematical Physics, vol I. Wiley Classics Library. Wiley-Interscience, New York
20. Curbera F (2000) Delayed curse of dimension for Gaussian integration. J Complex 16(2):474–506
21. Dawson CM, Eisert J, Osborne TJ (2007) Unifying variational methods for simulating quantum many-body systems. http://arxiv.org/abs/0705.3456v1
22. Demmel JW (1997) Applied Numerical Linear Algebra. SIAM, Philadelphia
23. Egorov AD, Sobolevsky PI, Yanovich LA (1993) Functional Integrals: Approximate Evaluation and Applications. Kluwer, Dordrecht
24. Feynman RP (1982) Simulating physics with computers. Int J Theor Phys 21:476
25. Forsythe GE, Wasow WR (2004) Finite-Difference Methods for Partial Differential Equations. Dover, New York
26. Grover L (1997) Quantum mechanics helps in searching for a needle in a haystack. Phys Rev Lett 79(2):325–328; http://arXiv.org/quant-ph/9706033
27. Heinrich S (2002) Quantum Summation with an Application to Integration. J Complex 18(1):1–50; http://arXiv.org/quant-ph/0105116
28. Heinrich S (2003) From Monte Carlo to Quantum Computation. In: Entacher K, Schmid WC, Uhl A (eds) Proceedings of the 3rd IMACS Seminar on Monte Carlo Methods MCM2001, Salzburg. Special Issue of Math Comput Simul 62:219–230
29. Heinrich S (2003) Quantum integration in Sobolev spaces. J Complex 19:19–42
30. Heinrich S (2004) Quantum Approximation II. Sobolev Embeddings. J Complex 20:27–45; http://arXiv.org/quant-ph/0305031
31. Heinrich S (2006) The randomized complexity of elliptic PDE. J Complex 22(2):220–249
32. Heinrich S (2006) The quantum query complexity of elliptic PDE. J Complex 22(5):691–725
33. Heinrich S, Milla B (2007) The randomized complexity of initial value problems. Talk presented at First Joint International Meeting between the American Mathematical Society and the Polish Mathematical Society, Warsaw, Poland
34. Heinrich S, Novak E (2002) Optimal summation by deterministic, randomized and quantum algorithms. In: Fang KT, Hickernell FJ, Niederreiter H (eds) Monte Carlo and Quasi-Monte Carlo Methods 2000. Springer, Berlin
35. Heinrich S, Kwas M, Woźniakowski H (2004) Quantum Boolean Summation with Repetitions in the Worst-Average Setting. In: Niederreiter H (ed) Monte Carlo and Quasi-Monte Carlo Methods, 2002. Springer, New York, pp 27–49
36. Jaksch P, Papageorgiou A (2003) Eigenvector approximation leading to exponential speedup of quantum eigenvalue calculation. Phys Rev Lett 91:257902; http://arXiv.org/quant-ph/0308016
37. Jordan SP (2005) Fast Quantum Algorithm for Numerical Gradient Estimation. Phys Rev Lett 95:050501; http://arXiv.org/quant-ph/0405146
38. Kacewicz BZ (1984) How to increase the order to get minimal-error algorithms for systems of ODEs. Numer Math 45:93–104
39. Kacewicz BZ (2006) Almost optimal solution of initial-value problems by randomized and quantum algorithms. J Complex 22(5):676–690
40. Keller HB (1968) Numerical methods for two-point boundary-value problems. Blaisdell Pub Co, Waltham
41. Knuth DE (1997) The Art of Computer Programming, vol 2: Seminumerical Algorithms, 3rd edn. Addison-Wesley Professional, Cambridge
42. Kwas M (2005) Quantum algorithms and complexity for certain continuous and related discrete problems. Ph D thesis. Department of Computer Science, Columbia University
43. Kwas M, Li Y (2003) Worst case complexity of multivariate Feynman–Kac path integration. J Complex 19:730–743
44. Manin Y (1980) Computable and Uncomputable. Sovetskoye Radio, Moscow (in Russian)
45. Manin YI (1999) Classical computing, quantum computing, and Shor's factoring algorithm. http://arXiv.org/quant-ph/9903008
46. Morita S, Nishimori H (2007) Convergence of Quantum Annealing with Real-Time Schrödinger Dynamics. J Phys Soc Jpn 76(6):064002; http://arXiv.org/quant-ph/0702252
47. Nayak A, Wu F (1999) The quantum query complexity of approximating the median and related statistics. In: Proc STOC 1999, Association for Computing Machinery, New York, pp 384–393. http://arXiv.org/quant-ph/9804066
48. Nielsen MA, Chuang IL (2000) Quantum Computation and Quantum Information. Cambridge University Press, Cambridge
49. Novak E (1988) Deterministic and Stochastic Error Bounds in Numerical Analysis. Lecture Notes in Mathematics, 1349. Springer, Berlin
50. Novak E (2001) Quantum complexity of integration. J Complex 17:2–16; http://arXiv.org/quant-ph/0008124
51. Ortiz G, Gubernatis JE, Knill E, Laflamme R (2001) Quantum algorithms for fermionic simulations. Phys Rev A 64(2):022319; http://arXiv.org/cond-mat/0012334
52. Papageorgiou A (2004) Average case quantum lower bounds for computing the boolean mean. J Complex 20(5):713–731
53. Papageorgiou A (2007) On the complexity of the multivariate Sturm–Liouville eigenvalue problem. J Complex 23(4–6): 802–827

54. Papageorgiou A, Traub JF (2005) Qubit complexity of continuous problems. http://arXiv.org/quant-ph/0512082
55. Papageorgiou A, Woźniakowski H (2005) Classical and Quantum Complexity of the Sturm–Liouville Eigenvalue Problem. Quantum Inf Process 4(2):87–127; http://arXiv.org/quant-ph/0502054
56. Paredes B, Verstraete F, Cirac JI (2005) Exploiting Quantum Parallelism to Simulate Quantum Random Many-Body Systems. Phys Rev Lett 95:140501; http://arXiv.org/cond-mat/0505288
57. Plaskota L (1996) Noisy Information and Computational Complexity. Cambridge University Press, Cambridge
58. Plaskota L, Wasilkowski GW, Woźniakowski H (2000) A new algorithm and worst case complexity for Feynman–Kac path integration. J Comp Phys 164(2):335–353
59. Ritter K (2000) Average-Case Analysis of Numerical Problems. Lecture Notes in Mathematics, 1733. Springer, Berlin
60. Shor PW (1997) Polynomial-time algorithms for prime factorization and discrete logarithm on a quantum computer. SIAM J Comput 26(5):1484–1509
61. Somma R, Ortiz G, Knill E, Gubernatis (2003) Quantum Simulations of Physics Problems. In: Pirich AR, Brant HE (eds) Quantum Information and Computation. Proc SPIE 2003, vol 5105. The International Society for Optical Engineering, Bellingham, pp 96–103. http://arXiv.org/quant-ph/0304063
62. Sornborger AT, Stewart ED (1999) Higher Order Methods for Simulations on Quantum Computers. Phys Rev A 60(3):1956–1965; http://arXiv.org/quant-ph/9903055
63. Strang G, Fix GJ (1973) An Analysis of the Finite Element Method. Prentice-Hall, Englewood Cliffs
64. Szkopek T, Roychowdhury V, Yablonovitch E, Abrams DS (2005) Egenvalue estimation of differential operators with a quantum algorithm. Phys Rev A 72:062318
65. Titschmarsh EC (1958) Eigenfunction Expansions Associated with Second-Order Differential Equations, Part B. Oxford University Press, Oxford
66. Traub JF (1999) A continuous model of computation. Phys Today May:39–43
67. Traub JF, Werschulz AG (1998) Complexity and Information. Cambridge University Press, Cambridge
68. Traub JF, Woźniakowski H (1980) A general theory of optimal algorithms. ACM Monograph Series. Academic Press, New York
69. Traub JF, Woźniakowski H (1992) The Monte Carlo algorithm with a pseudorandom generator. Math Comp 58(197):323–339
70. Traub JF, Woźniakowski H (2002) Path integration on a quantum computer. Quantum Inf Process 1(5):365–388; http://arXiv.org/quant-ph/0109113
71. Traub JF, Wasilkowski GW, Woźniakowski H (1988) Information-Based Complexity. Academic Press, New York
72. Wasilkowski GW, Woźniakowski H (1996) On tractability of path integration. J Math Phys 37(4):2071–2088
73. Weinberger HF (1956) Upper and Lower Bounds for Eigenvalues by Finite Difference Methods. Commun Pure Appl Math IX:613–623
74. Weinberger HF (1958) Lower Bounds for Higher Eigenvalues by Finite Difference Methods. Pacific J Math 8(2):339–368
75. Werschulz AG (1991) The Computational Complexity of Differential and Integral Equations. Oxford University Press, Oxford
76. Wisner S (1996) Simulations of Many-Body Quantum Systems by a Quantum Computer. http://arXiv.org/quant-ph/96
77. Woźniakowski H (2006) The Quantum Setting with Randomized Queries for Continuous Problems. Quantum Inf Process 5(2):83–130
78. Yepez J (2002) An efficient and accurate quantum algorithm for the Dirac equation. http://arXiv.org/quant-ph/0210093
79. Zalka C (1998) Simulating quantum systems on a quantum computer. Proc Royal Soc Lond A 454(1969):313–322; http://arXiv.org/quant-ph/9603026

Quantum Bifurcations

Boris Zhilinskiĭ
Université du Littoral, Dunkerque, France

Article Outline

Glossary
Definition of the Subject
Introduction
Simplest Effective Hamiltonians
Bifurcations and Symmetry
Imperfect Bifurcations
Organization of Bifurcations
Bifurcation Diagrams for Two Degree-of-Freedom Integrable Systems
Bifurcations of "Quantum Bifurcation Diagrams"
Semi-Quantum Limit and Reorganization of Quantum Bands
Multiple Resonances and Quantum State Density
Physical Applications and Generalizations
Future Directions
Bibliography

Glossary

Classical limit The classical limit is the classical mechanical problem which can be constructed from a given quantum problem by some limiting procedure. During such a construction the classical limiting manifold should be defined which plays the role of classical phase space. As soon as quantum mechanics is more general than classical mechanics, going to the classical limit from a quantum problem is much more reasonable than discussing possible quantizations of classical theories [73].

Energy-momentum map In classical mechanics for any problem which allows the existence of several integrals of motion (typically energy and other integrals often named as momenta) the Energy-Momentum (EM) map gives the correspondence between the phase space

of the initial problem and the space of values of all independent integrals of motion. The energy-momentum map introduces the natural foliation of the classical phase space into common levels of values of energy and momenta [13,35]. The image of the EM map is the region of the space of possible values of integrals of motion which includes regular and critical values. The quantum analog of the image of the energy-momentum map is the joint spectrum of mutually commuting quantum observables.

Joint spectrum For each quantum problem a maximal set of mutually commuting observables can be introduced [16]. A set of quantum wave functions which are mutual eigenfunctions of all these operators exists. Each such eigenfunction is characterized by eigenvalues of all mutually commuting operators. The representation of mutual eigenvalues of n commuting operators in the n-dimensional space gives the geometrical visualization of the joint spectrum.

Monodromy In general, the monodromy characterizes the evolution of some object after it makes a close path around something. In classical Hamiltonian dynamics the Hamiltonian monodromy describes for completely integrable systems the evolution of the first homology group of the regular fiber of the energy-momentum map after a close path in the regular part of the base space [13].

For a corresponding quantum problem the quantum monodromy describes the modification of the local structure of the joint spectrum after its propagation along a close path going through a regular region of the lattice.

Quantum bifurcation Qualitative modification of the joint spectrum of the mutually commuting observables under the variation of some external (or internal) parameters and associated in the classical limit with the classical bifurcation is named quantum bifurcation [59]. In other words the quantum bifurcation is the manifestation of the classical bifurcation presented in the classical dynamic system in the quantum version of the same system.

Quantum-classical correspondence Starting from any quantum problem the natural question consists of defining the corresponding classical limit, i. e. the classical dynamic variables forming the classical phase space and the associated symplectic structure. Whereas in simplest quantum problems defined in terms of standard position and momentum operators with commutation relation $[q_i, p_j] = i\hbar\delta_{ij}$, $[q_i, q_j] = [p_i, p_j] = 0$ $(i, j = 1 \ldots n)$ the classical limit phase space is the $2n$-dimensional Euclidean space with standard symplectic structure on it, the topology of the classical limit manifold in many other important for physical applications cases can be rather non-trivial [73,87].

Quantum phase transition Qualitative modifications of the ground state of a quantum system occurring under the variation of some external parameters at zero temperature are named quantum phase transitions [65]. For finite particle systems the quantum phase transition can be considered as a quantum bifurcation [60].

Spontaneous symmetry breaking Qualitative modification of the system of quantum states caused by perturbation which has the same symmetry as the initial problem. Local symmetry of solutions decreases but the number of solutions increases. In the energy spectra of finite particle systems the spontaneous symmetry breaking produces an increase of the "quasidegeneracy", i. e. formation of clusters of quasi-degenerate levels whose multiplicity can be much higher than the dimension of the irreducible representations of the global symmetry group [51].

Symmetry breaking Qualitative changes in the properties (dynamical behavior, and in particular in the joint spectrum) of quantum systems which are due to modifications of the global symmetry of the problem caused by external (less symmetrical than original problem) perturbation can be described as symmetry breaking effects. Typical effects consist of splitting of degenerate energy levels classified initially according to irreducible representation of the initial symmetry group into less degenerate groups classified according to irreducible representation of the subgroup (the symmetry group of the perturbation) [47].

Definition of the Subject

Quantum bifurcations (QB) are qualitative phenomena occurring in quantum systems under the variation of some internal or external parameters. In order to make this definition a little more precise we add the additional requirement: The qualitative modification of the "behavior" of a quantum system can be described as QB if it consists of the manifestation for the quantum system of the classical bifurcation presented in classical dynamic systems which is the classical analog of the initial quantum system. Quantum bifurcations are typical elementary steps leading from the simplest in some way effective Hamiltonian to more complicated ones under the variation of external or internal parameters. As internal parameters one may consider the values of exact or approximate integrals of motion. The construction of an effective Hamiltonian

is typically based on the averaging and/or reduction procedure which results in the appearance of "good" quantum numbers (or approximate integrals of motion). The role of external parameters can be played by forces of external champs, phenomenological constants in the effective Hamiltonians, particle masses, etc. In order to limit the very broad field of qualitative changes and of possible quantum bifurcations in particular, we restrict ourselves mainly to quantum systems whose classical limit is associated with compact phase space and is nearly integrable. This means that for quantum problems the set of mutually commuting observables can be constructed within a reasonable physical approximation almost everywhere at least locally.

Quantum bifurcations are supposed to be universal phenomena which appear in generic families of quantum systems and explain how relatively simple behavior becomes complicated under the variation of some physical parameters. To know these elementary bricks responsible for increasing complexity of quantum systems under control parameter modifications is extremely important in order to make the extrapolation to regimes unaccessible to experimental study.

Introduction

In order to better understand the manifestations of quantum bifurcations and their significance for concrete physical systems we start with the description of several simple model physical problems which exhibit in some sense the simplest (but nevertheless) generic behavior. Let us start with the harmonic oscillator. A one-dimensional harmonic oscillator has an equidistant system of eigenvalues. All eigenvalues can be labeled by consecutive integer quantum numbers which have the natural interpretation in terms of the number of zeros of eigenfunctions. The classical limit manifold (classical phase space) is a standard Euclidean 2-dimensional space with natural variables $\{p, q\}$. The classical Hamiltonian for the harmonic oscillator is an example of a Morse-type function which has only one stationary point $p = q = 0$ and all non-zero energy levels of the Hamiltonian are topological circles. If now we deform slightly the Hamiltonian in such a way that its classical phase portrait remains topologically the same, the spectrum of the quantum problem changes but it can be globally described as a regular sequence of states numbered consecutively by one integer and such description remains valid for any mass parameter value. Note, that for this problem increasing mass means increasing quantum state density and approaching classical behavior (classical limit).

More serious modification of the harmonic oscillator can lead, for example, to creation of new stationary points of the Hamiltonian. In classical theory this phenomenon is known as fold bifurcation or fold catastrophe [3,31]. The phase portrait of the classical problem changes qualitatively. As a function of energy the constant level set of the Hamiltonian has different topological structure (one circle, two circles, figure eight, circle and a point, or simply point). The quantum version of the same problem shows the existence of three different sequences of states which become clearly visible in the limit of the high density of states which can be reached by increasing the mass value parameter [32]. Such qualitative modification of the energy spectrum of the 1D-quantum Hamiltonian gives the simplest example of the phenomenon which can be described as a quantum bifurcation. Figure 1 shows a schematic representation of quantum bifurcations for a model system with one degree-of-freedom in parallel in quantum and classical mechanics.

After looking for one simple example we can formulate a more general question which concerns the appearance in more general quantum systems of qualitative phenomena which can be characterized as quantum bifurcations.

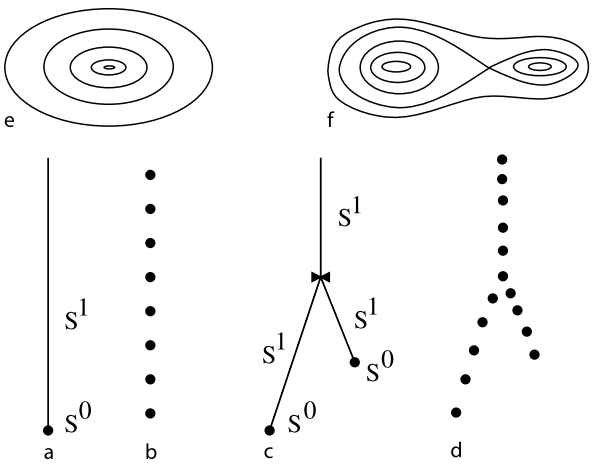

Quantum Bifurcations, Figure 1

Classical and quantum bifurcations for a one degree-of-freedom system. Situations before (a,b,e) and after (c,d,f) the bifurcation are shown. a Energy map for harmonic oscillator-type system. Inverse images of each point are indicated. b Quantum state lattice for harmonic oscillator-type system. c Energy map after the bifurcation. Inverse images of each point are indicated. d Quantum state lattice after bifurcation represented as composed of three regular parts glued together. e Phase portrait for harmonic oscillator-type system. Inverse images are S^1 (generic inverse image) and S^0 (inverse image for minimal energy value). f Phase portrait after bifurcation

Simplest Effective Hamiltonians

We turn now to several models which describe some specific classes of relatively simple real physical quantum systems formed by a finite number of particles (atoms, molecules, ...). Spectra of such quantum objects are studied nowadays with very high accuracy and this allows us to compare the behavior predicted by quantum bifurcations with the precise information about energy level structure found, for example, from high-resolution molecular spectroscopy.

Typically, the intra-molecular dynamics can be split into electronic, vibrational, and rotational ones due to important differences in characteristic energy excitations or in time scales. The most classical is the rotational motion and probably due to that the quantum bifurcations as a counterpart to classical bifurcations were first studied for purely rotational problems [59,61].

Effective rotational Hamiltonians describe the internal structure of rotational multiplets formed by isolated finite particle systems (atoms, molecules, nuclei) [36]. For many molecular systems in the ground electronic state any electronic and vibrational excitations are much more energy consuming as compared with rotational excitations. Thus, to study the molecular rotation the simplest physical assumption is to suppose that all electronic and all vibrational degrees-of-freedom are frozen. This means that a set of quantum numbers is given which have the sense of approximate integrals of motion specifying the character of vibrational and electronic motions in terms of these "good" quantum numbers. At the same time for a free molecule in the absence of any external fields due to isotropy of the space the total angular momentum J and its projection J_z on the laboratory fixed frame are strict integrals of motion. Consequently, to describe the rotational motion for fixed values of J and J_z it is sufficient to analyze the effective problem with only one degree-of-freedom. The dimension of classical phase space in this case equals two and the two classical conjugate variables are: the projection of the total angular momentum on the body fixed frame and conjugate angle variable. The classical phase space is topologically a two-dimensional sphere, S^2. There is a one-to-one correspondence between the points on a sphere and the orientation of the angular momentum in the body-fixed frame. Such a representation gives a clear visualization of a classical rotational Hamiltonian as a function defined over a sphere [36,49].

In quantum mechanics the rotation of molecules is traditionally described in terms of an effective rotational Hamiltonian which is constructed as a series in rotational operators J_x, J_y, J_z, the components of the total angular momentum $\mathbf{J}$. In a suitably chosen molecular fixed frame the effective Hamiltonian has the form

$$H_{\text{eff}} = AJ_x^2 + BJ_y^2 + CJ_z^2 + \sum c_{\alpha\beta\gamma}\, J_x^\alpha\, J_y^\beta\, J_z^\gamma + \cdots\,, \quad (1)$$

where A, B, C and $c_{\alpha\beta\gamma}$ are constants. To relate quantum and classical pictures we note that $\mathbf{J}^2$ and energy are integrals of Euler's equations of motion for dynamic variables J_x, J_y, J_z. The phase space of the classical rotational problem with constant $|\mathbf{J}|$ is S^2, the two-dimensional sphere, and it can be parametrized with spherical angles (θ, ϕ) in such a way that the points on S^2 define the orientation of $\mathbf{J}$, i. e. the position of the axis and the direction of rotation. To get the classical interpretation of the quantum Hamiltonian we introduce the classical analogs of the operators J_x, J_y, J_z

$$\mathbf{J} \longrightarrow \begin{pmatrix} J_x \\ J_y \\ J_z \end{pmatrix} = \begin{pmatrix} \sin\theta\cos\phi \\ \sin\theta\sin\phi \\ \cos\theta \end{pmatrix} \sqrt{J(J+1)} \quad (2)$$

and consider the rotational energy as a function of the dynamical variables (θ, ϕ) and the parameter J.

Thus, for an effective rotational Hamiltonian the corresponding classical symbol is a function $E_J(\theta, \phi)$ defined over S^2 and named usually the rotational energy surface [36].

Taking into account the symmetry imposed by the initial problem and the topology of the phase space the simplest rotational Hamiltonian can be constructed. In classical mechanics the simplest Hamiltonian can be defined (using Morse theory [55,89]) as a Hamiltonian function with the minimal possible number of non-degenerate stationary points compatible with the symmetry group action of the classical phase space. Morse theory in the presence of symmetry (or equivariant Morse theory) implies important restrictions on the number of minima, maxima, and saddle points. In the absence of symmetry the simplest Morse type function on the S^2 phase space has one minimum and one maximum, as a consequence of Morse inequalities. In the presence of non-trivial symmetry group action the minimal number of stationary points on the sphere increases. For example, many asymmetric top molecules (possessing three different moment of inertia of the equilibrium configuration) have D_{2h} symmetry group [47]. This group includes rotations over π around $\{x, y, z\}$ axes, reflections in $\{xy, yz, zx\}$ planes and inversion as symmetry operations. Any D_{2h} invariant function on the sphere has at least six stationary points (two equivalent minima, two equivalent maxima, and two equivalent saddle points). This means that in quantum

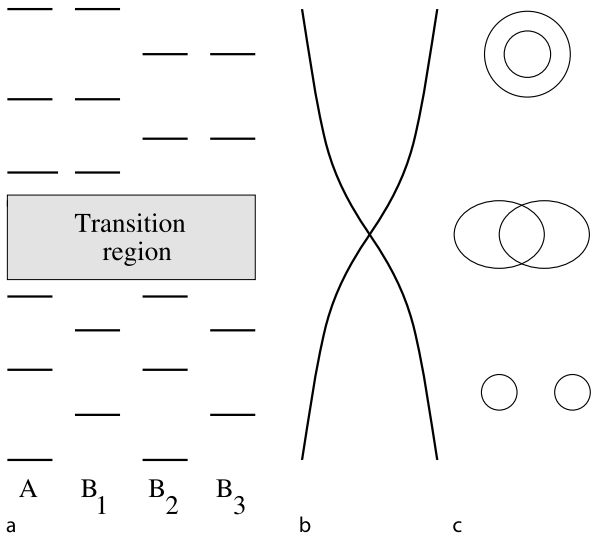

Quantum Bifurcations, Figure 2
a Schematic representation of the energy level structure for asymmetric top molecule. *Vertical axis* corresponds to energy variation. Quantum levels are classified by the symmetry group of the asymmetric top. Two fold clusters at two ends of the rotational multiplet are formed by states with different symmetry. b Foliation of the classical phase space (S^2 sphere) by constant levels of the Hamiltonian given in the form of its Reeb graph. Each *point* corresponds to a connected component of the constant level set of the Hamiltonian (energy). c Geometric representation of the constant energy sections

mechanics the asymmetric top has eigenvalues which form two regular sequences of quasi-degenerate doublets with the transition region between them. The correspondence between the quantum spectrum and the structure of the energy map for the classical problem is shown in Fig. 2. Highly symmetrical molecules which have cubic symmetry, for example, can be described by a simplest Morse-type Hamiltonian with 26 stationary points (6 and 8 minima/maxima and 12 saddle points). As a consequence, the corresponding quantum Hamiltonian shows the presence of six-fold and eight-fold quasi-degenerate clusters of rotational levels.

As soon as the simplest classical Hamiltonian is characterized by the appropriate system of stationary points the whole region of possible classical energy values (in the case of dynamical systems with only one degree-of-freedom the energy-momentum map becomes simply the energy map) appears to be split into different regions corresponding to different dynamical regimes, i. e. to different regions of the phase portrait foliated by topologically non-equivalent systems of classical trajectories. Accordingly, the energy spectrum of the corresponding quantum Hamiltonian can be qualitatively described as formed by regular sequences of states within each region of the classical energy map.

Quantum bifurcations are universal phenomena which lead to a new organization of the energy spectrum into qualitatively different regions in accordance with corresponding qualitative modifications of the classical energy-momentum map under the variation of some control parameter.

Simplest Hamiltonians for Two Degree-of-Freedom Systems

When the quantum system has two or larger number of degrees-of-freedom the simplest dynamical regimes often correspond in classical mechanics to a quasi-regular dynamics which can be reasonably well approximated by an integrable model. The integrable model in classical mechanics can be constructed by normalizing the Hamiltonian and by passing to so-called normal forms [2,49]. The quantum counterpart of normalization is the construction of a mutually commuting set of operators which should not be mistaken with quantization of systems in normal form. Corresponding eigenvalues can be used as "good" quantum numbers to label quantum states. A joint spectrum of mutually commuting operators corresponds to the image of the energy-momentum map for the classical completely integrable dynamical problem. In this context the question about quantum bifurcations first of all leads to the question about qualitative classification of the joint spectra of mutually commuting operators. To answer this question we need to start with the qualitative description of foliations of the total phase space of the classical problem by common levels of integrals of motion which are mutually in involution [2,7]. One needs to distinguish the regular and the singular values of the energy-momentum map. For Hamiltonian systems the inverse images of the regular values are regular tori (one or several) [2]. A lot of different singularities are possible. In classical mechanics different levels of the classifications are studied in detail [7]. The diagram which represents the image of the classical EM map together with its stratification into regular and critical values is named the bifurcation diagram. The origin of such a name is due to the fact that the values of integrals of motion can be considered as control parameters for the phase portraits (inverse images of the EM map) of the reduced systems.

For quantum problems the analog of the classical stratification of the EM map for integrable systems is the splitting of the joint spectrum of several commuting observables into regions formed by regular lattices of joint eigenvalues. Any local simply connected neighborhood of a reg-

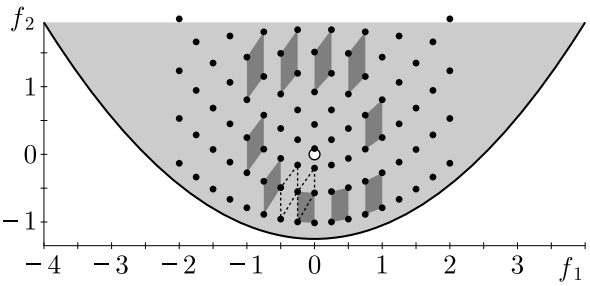

Quantum Bifurcations, Figure 3
Joint spectrum of two commuting operators together with the image of the classical EM map for the resonant 1 : (−1) oscillator given by (3),(4). Quantum monodromy is seen as a result of transportation of the elementary cell of the quantum lattice along a close path through a non simply connected region of the regular part of the image of the EM map. Taken from [58]

ular point of the lattice can be deformed into part of the regular Z^n lattice of integers. This means that local quantum numbers can be consistently introduced to label states of the joint spectrum. If the regular region is not simply connected it still can be characterized locally by a set of "good" quantum numbers. At the same time this is impossible globally. Likewise in classical mechanics the Hamiltonian monodromy is the simplest obstruction to the existence of the global action-angle variables [17,57], in quantum mechanics the analog notion of quantum monodromy [14,34,68,80] characterizes the global non-triviality of the regular part of the lattice of joint eigenvalues. Figure 3 demonstrates the effect of the presence of a classical singularity (isolated focus-focus point) on the global properties of the quantum lattice formed by joint eigenvalues of two commuting operators for a simple problem with two degrees-of-freedom, which is essentially the 1 : (−1) resonant oscillator [58]. Two integrals of motions in this example are chosen as

$$f_1 = \frac{1}{2}\left(p_1^2 + q_1^2\right) - \frac{1}{2}\left(p_2^2 + q_2^2\right), \tag{3}$$

$$f_2 = p_1 q_2 + p_2 q_1 + \frac{1}{4}\left(p_1^2 + q_1^2 + p_2^2 + q_2^2\right)^2. \tag{4}$$

Locally in any simply connected region which does not include the classical singularity of the EM map situated at $f_1 = f_2 = 0$, the joint spectrum can be smoothly deformed to the regular Z^2 lattice [58,90]. Such lattices are shown, for example, in Fig. 4. If somebody wants to use only one chart to label states, it is necessary to take care in respect of the multivaluedness of such a representation. There are two possibilities:

(i) One makes a cut and maps the quantum lattice to a regular Z^2 lattice with an appropriate solid angle removed from it (see Fig. 5 [58,68,90]). Points on the boundary of such a cut should be identified and a spe-

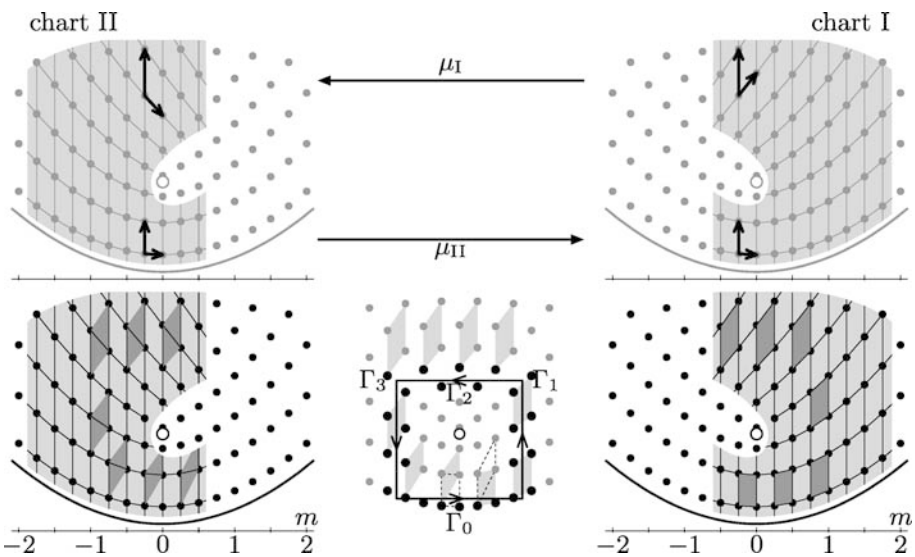

Quantum Bifurcations, Figure 4
Two chart atlas which covers the quantum lattice of the 1 : (−1) resonant oscillator system represented in Fig. 3. *Top plots* show the choice of basis cells and the gluing map between the charts. *Bottom plots* show the transport of the elementary cell (*dark gray quadrangles*) in each chart. *Central bottom panel* shows closed path Γ and its quantum realization (*black dots*) leading to non-trivial monodromy (compare with Fig. 3). Taken from [58]

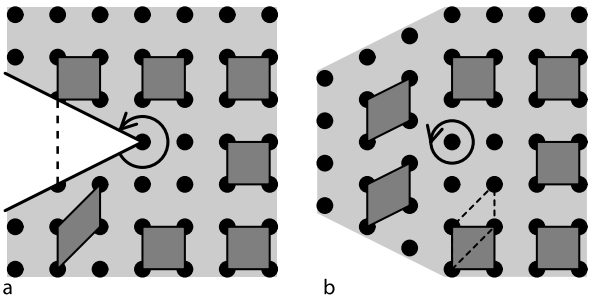

Quantum Bifurcations, Figure 5
Construction of the 1 : (−1) lattice defect starting from the regular Z^2 lattice. The solid angle is removed from the regular Z^2 lattice and points on the so-obtained boundary are identified by vertical shifting. *Dark gray quadrangles* show the evolution of an elementary lattice cell along a closed path around the defect point. Taken from [58]

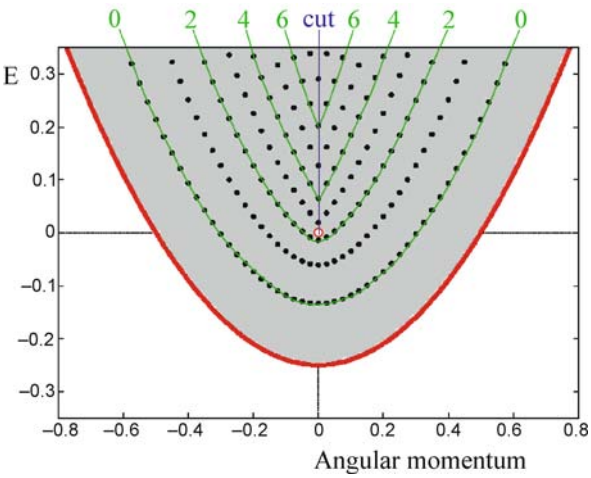

Quantum Bifurcations, Figure 6
Representation of the quantum joint spectrum for the "Mexican hat" potential $V(r) = ar^4 - br^2$ with the "cut" along the eigenray. For such a cut the left and the right limits at the cut give the same values of actions (good quantum numbers) but the lines of constant values of actions exhibit a "kink" at the cut (the discontinuity of the first derivative)

cial matching rule explaining how to cross the path should be introduced. Similar constructions are quite popular in solid state physics in order to represent defects of lattices, like dislocations, disclinations, etc. We just note that the "monodromy defect" introduced in such a way is different from standard construction for dislocation and disclination defects [45,50]. The inverse procedure of the construction of the "monodromy defect" [90] from a regular lattice is represented in Fig. 5. Let us note that the width of the solid angle removed depends on the direction of the cut and the direction of the cut itself can be chosen in an ambiguous way.

(ii) An alternative possibility is to make a cut in such a way that the width of the removed angle becomes equal to zero. For focus-focus singularities one such direction always exists and is named an eigenray by Symington [75]. The same construction is used in some physical papers [10,11,85]. The inconvenience of such a procedure is the appearance of discontinuity of the slope of the constant action (quantum number) line at the cut, whereas the values of actions themselves are continued (see Fig. 6). This gives the wrong impression that this eigenray is associated with some special non-regular behavior of the initial problem, whereas there is no singularity except at one focus-focus point.

Bifurcations and Symmetry

The general mathematical answer about the possible qualitative modifications of a system of stationary points of functions depending on some control parameters can be found in bifurcation (or catastrophe) theory [3,31,33]. It is important that the answer depends on the number of control parameters and on the symmetry. Very simple classification of possible typical bifurcations of stationary points of a one-parameter family of functions under presence of symmetry can be formulated for dynamical systems with one degree-of-freedom. The situation is particularly simple because the phase space is two-dimensional and the complete list of local symmetry groups (which are the stabilizers of stationary points) includes only 2D-point groups [84]. It should be noted that the global symmetry of the problem can be larger than the local symmetry of the bifurcating stationary points. In such a case the bifurcations occur simultaneously for all points forming one orbit of the global symmetry group [51,52]. We describe briefly here (see Table 1) the classification of the bifurcations of stationary points in the presence of symmetry for families of functions depending on one parameter and associated quantum bifurcations [59,61]. Their notation includes the local symmetry group and several additional indexes which specify creation/annihilation of stationary points and the local or non-local character of the bifurcation. The list of possible bifurcations includes:

$C_1^{\pm}$ A non-symmetrical non-local bifurcation resulting in the appearance (+) or disappearance (−) of a stable-unstable pair of stationary points with the trivial local symmetry C_1. In the quantum problem this bifurcation is associated with the appearance or disappearance of a new regular sequence of states glued at its end with

Quantum Bifurcations, Table 1

Bifurcations in the presence of symmetry. *Solid lines* denote stable stationary points. *Dashed lines* denote unstable stationary points. *Numbers in parenthesis* indicate the multiplicity of stationary points

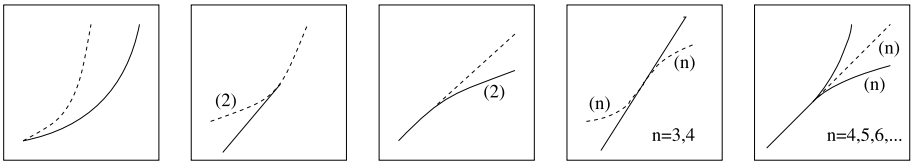

$C_1^{\pm}$ $C_2^{N\pm}$ $C_2^{L\pm}$ $C_n^N,\ n=3,4$ $C_n^{L\pm},\ n\geq 4$

the intermediate part of another regular sequence of quantum states [32,77].

$\boldsymbol{C_2^{L\pm}}$ A local bifurcation with the broken C_2 local symmetry. This bifurcation results either in appearance of a triple of points (two equivalent stable points with C_1 local symmetry and one unstable point with C_2 local symmetry) instead of one stable point with C_2 symmetry, or in inverse transformation. The number of stationary points in this bifurcation increases or decreases by two. For the quantum problem the result is the transformation of a local part of a regular sequence of states into one sequence of quasi-degenerate doublets.

$\boldsymbol{C_2^{N\pm}}$ A non-local bifurcation with the broken C_2 local symmetry. This bifurcation results in appearance (+) or disappearance (−) of two new unstable points with broken C_2 symmetry and simultaneous transformation of the initially stable (for +) or unstable (for −) stationary point into an unstable/stable one. The number of stationary points in this bifurcation increases or decreases by two. For the quantum problem this means the appearance of a new regular sequence of states near the separatrix between two different regular regions.

$\boldsymbol{C_n^N}$ $(n=3,4)$ A non-local bifurcation corresponding to passage of n unstable stationary points through a stable stationary point with C_n local symmetry which is accompanied with the minimum ↔ maximum change for a stable point with the C_n local symmetry. The number of stationary points remains the same. For the quantum problem this bifurcation corresponds to transformation of the increased sequence of energy levels into a decreased sequence.

$\boldsymbol{C_n^{L\pm}}$ $(n\geq 4)$ A local bifurcation which results in appearance (+) or disappearance (−) of n stable and n unstable stationary points with the broken C_n symmetry and a simultaneous minimum ↔ maximum change of a stable point with the C_n local symmetry. The number of stationary points increases or decreases by $2n$. In the quantum problem after bifurcation a new sequence of n-times quasi-degenerate levels appears/disappears.

Universal quantum Hamiltonians which describe the qualitative modification of the quantum energy level system around the bifurcation point are given in [59,61].

The presence of symmetry makes it much easier to observe the manifestation of quantum bifurcations. Modification of the local symmetry of stable stationary points results in the modification of the cluster structure of energy levels, i. e. the number and the symmetry types of energy level forming quasi-degenerate groups of levels. This phenomenon is essentially the spontaneous breaking of symmetry [51]. Several concrete molecular systems which show the presence of quantum bifurcations in rotational structure under rotational excitation are cited in Table 2. Many other examples can be found in [9,23,59,67,71,88,89,92,93] and references therein. In purely vibrational problems breaking dynamical $SU(N)$ symmetry of the isotrope harmonic oscillator till finite symmetry group results in formation of so-called non-linear normal modes [23,54] or quasimodes [1], or local modes [9,25,39,40,43,46,48]. In the case of two degrees-of-freedom the analysis of the vibrational problem can be reduced to the analysis of the problem similar to the rotational one [36,66] and all the results about possible types of bifurcations found for rotational problems remain valid in the case of intra-molecular vibrational dynamics.

Quantum Bifurcations, Table 2

Molecular examples of quantum bifurcations in the rotational structure of individual vibrational components under the variation of the absolute value of angular momentum, J. J_c is the critical value corresponding to bifurcation

Molecule	Component	J_c	Bifurcation type
SiH_4	$\nu_2(+)$	12	C_2^{N+}
SnH_4	$\nu_2(-)$	10	$C_2^{N+}, C_3^N, C_4^N, C_2^{N-}$
CF_4	$\nu_2(+)$	50	C_4^{L+}
H_2Se	$\|0\rangle$	20	C_2^{L+}

Imperfect Bifurcations

According to general results the possible types of bifurcations which are generically present (and persist under small deformations) in a family of dynamical systems strictly depend on the number of control parameters. In the absence of symmetry only one bifurcation of stationary points is present for a one-parameter family of Morse-type functions, namely the formation (annihilation) of two new stationary points. This corresponds to saddle-node bifurcation for one degree-of-freedom Hamiltonian systems. The presence of symmetry increases significantly the number of possible bifurcations even for families with only one parameter [31,33]. From the physical point-of-view it is quite natural to study the effect of symmetry breaking on the symmetry allowed bifurcation. Decreasing symmetry naturally results in the modification of the allowed types of bifurcations but at the same time it is quite clear that at sufficient slight symmetry breaking perturbation the resulting behavior of the system should be rather close to the behavior of the original system with higher symmetry.

In the case of a small violation of symmetry the so-called "imperfect bifurcations" can be observed. Imperfect bifurcations, which are well known in the classical theory of bifurcations [33] consist of the appearance of stationary points in the neighborhood of another stationary point which does not change its stability. In some way one can say that imperfect bifurcation mimics generic bifurcation in the presence of higher symmetry by the special organization of several bifurcations which are generic in the presence of lower symmetry. Naturally quantum bifurcations follow the same behavior under the symmetry breaking as classical ones. Very simple and quite natural examples of imperfect quantum bifurcations were demonstrated on the example of the rotational structure modifications under increasing angular momentum [91]. The idea of appearance of imperfect bifurcations is as follows. Let us suppose that some symmetrical molecule demonstrates under the variation of angular momentum a quantum rotational bifurcation allowed by symmetry. The origin of this bifurcation is due, say, to centrifugal distortion effects which depend strongly on J but are not very sensitive to small variation of masses even in the case of symmetry breaking isotopic substitution. In such a case a slight modification of the masses of one or several equivalent atoms breaks the symmetry and this symmetry violation can be made very weak due to the small ratio $\Delta M/M$ under isotope substitution. In classical theory the effect of symmetry breaking can be easily seen through the variation of the position of stationary points with control parameter. For example, instead of a pitchfork bifurcation which is typical for C_2 local symmetry, we get for the unsymmetrical problem (after slight breaking of C_2 symmetry) a smooth evolution of the position of one stationary point and the appearance of two new stationary points in fold catastrophe (see Fig. 7). In associated quantum bifurcations the most important effect is the splitting of clusters. But one should be careful with this interpretation because in quantum mechanics of finite particle systems the clusters are always split due to quantum mechanical tunneling between different equivalent regions of localization of quantum wave functions. Intercluster splitting increases rapidly approaching the region of classical separatrix. The behavior of quantum tunneling was studied extensively in relation to the quantum breathers problem [6,29]. Systematic application of quasi-classical methods to reproduce quantum energy

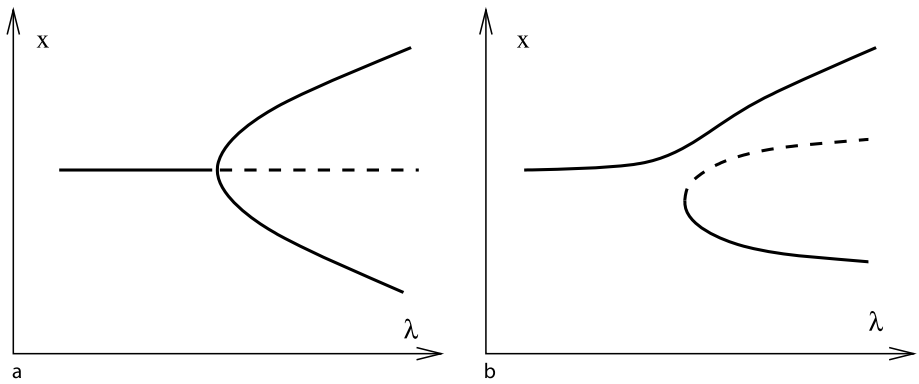

Quantum Bifurcations, Figure 7
Imperfect bifurcations. a Position x of stationary points as a function of control parameter λ during a pitchfork bifurcation in the presence of C_2 local symmetry. b Modifications induced by small symmetry perturbation of lower symmetry. *Solid line*: Stable stationary points. *Dashed lines*: Unstable stationary points

level structure near the singularities of the energy-momentum maps where exponentially small corrections are important is possible but requires special efforts (see for example [12]) and we will not touch upon this problem here.

Organization of Bifurcations

The analysis of the quantum bifurcations in concrete examples of rotating molecules have shown that in some cases the molecule undergoes several consecutive qualitative changes which can be interpreted as a sequence of bifurcations which sometimes cannot even be separated into elementary bifurcations for the real scale of the control parameter [89]. One can imagine in principle that successive bifurcations lead to quantum chaos in analogy with classical dynamical systems where the typical scenario for the transition to chaos is through a sequence of bifurcations. Otherwise, the molecular examples were described with effective Hamiltonians depending only on one degree-of-freedom and the result of the sequence of bifurcations was just the crossover of the rotational multiplets [64]. In some sense such a sequence of bifurcations can be interpreted as an imperfect bifurcation assuming initially higher dynamical symmetry, like the continuous $SO(3)$ group.

Later, a similar crossover phenomenon was found in a quite different quantum problem, like the hydrogen atom in external fields [24,53,72]. The general idea of such organization of bifurcations is based on the existence of two different limiting cases of dynamical regimes for the same physical quantum system (often under presence of the same symmetry group) which are qualitatively different. For example, the number of stationary points, or their stability differs. If H_1 and H_2 are two corresponding effective Hamiltonians, the natural question is: Is it possible to transform H_1 into H_2 by a generic perturbation depending on only one parameter? And if so, what is the minimal number of bifurcations to go through?

The simplest quantum system for which such a question becomes extremely natural is the hydrogen atom in the presence of external static electric (F) and magnetic (G) fields. Two natural limits – the Stark effect in the electric field and Zeeman effect in the magnetic field – show quite different qualitative structure even in the extremely low field limit [15,20,63,72,78]. Keeping a small field one can go from one (Stark) limit to another (Zeeman) and this transformation naturally goes through qualitatively different regimes [24,53]. In spite of the fact that the hydrogen atom (even without spin and relativistic corrections) is only a three degree-of-freedom system, the complete description of qualitatively different regimes in a small field limit is still not done and remains an open problem [24].

An example of clearly seen qualitative modifications of the quantum energy level system of the hydrogen atom under the variation of F/G ratio of the strengths of two parallel electric and magnetic fields is shown in Fig. 8. The calculations are done for a two degree-of-freedom system after the normalization with respect to the global action. In quantum mechanics language this means that only energy levels which belong to the same n-shell of the hydrogen atom are treated and the interaction with other n' shells is taken into account only effectively. The limiting classical phase space for this effective problem is the four-dimensional space $S^2 \times S^2$, which is the direct product of two two-dimensional spheres. In the presence of axial symmetry this problem is completely integrable and the Hamiltonian and the angular momentum provide a complete set of mutually commuting operators. Energies of stationary points of classical Hamiltonian limit are shown on the same Fig. 8 along with quantum levels. When one of the characteristic frequencies goes through zero, the so-called collapse phenomena occurs. Some other non-trivial resonance relations between two frequencies are also indicated. These resonances correspond to special orga-

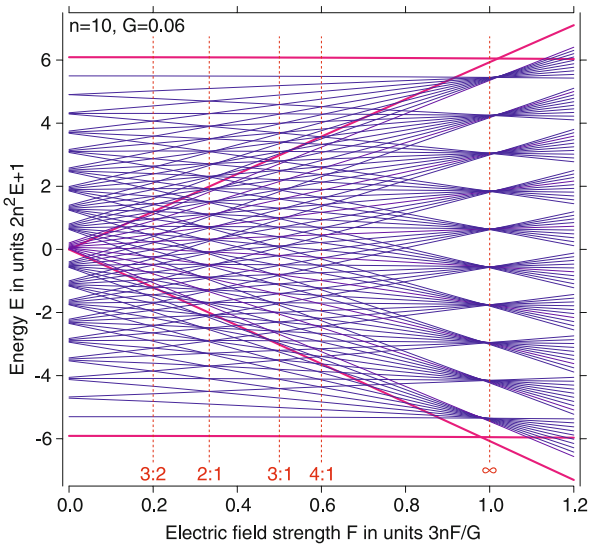

Quantum Bifurcations, Figure 8

Reorganization of the internal structure of the n-multiplet of the hydrogen atom in small parallel electric and magnetic fields. Energies of stationary points of the classical Hamiltonian (*red solid lines*) are shown together with quantum energy levels (*blue solid lines*). The figure is done for $n = 10$ (there are $n^2 = 100$ energy levels forming this multiplet). As the ratio F/G of electric F and magnetic G fields varies this two degree-of-freedom system goes through different zones associated with special resonance relations between two characteristic frequencies (shown by *vertical dashed lines*). Taken from [24]

nization of quantum energy levels. At the same time it is not necessary here to go to joint spectrum representation in order to see the reorganization of stationary points of the Hamiltonian function on $S^2 \times S^2$ phase space under the variation of the external control parameter F/G. A more detailed treatment of qualitative features of the energy level systems for the hydrogen atom in low fields is given in [15,20,24].

Bifurcation Diagrams for Two Degree-of-Freedom Integrable Systems

Let us consider now the two degree-of-freedom integrable system with compact phase space as a bit more complex but still reasonably simple problem. Many examples of such systems possess EM maps with the stratification of the image formed by the regular part surrounded by the singular boundary. The most naturally arising examples of classical phase spaces, like $S^2 \times S^2$, CP^2, are of that type. All internal points on the image of the EM map are regular in these cases. In practice, real physical problems, even after necessary simplifications and approximations lead to more complicated models. Some examples of fragments of images of the EM map with internal singular points are shown in Fig. 9. In classical mechanics the inverse images of critical values are singular tori of different kinds. Some of them are represented in Fig. 10. Inverse images of critical points situated on the boundary of the EM image have lower dimension. They can be one-dimensional tori (S^1-circles), or zero-dimensional (points).

The natural question now is to describe typical generic modifications of the Hamiltonian which lead to qualitative modifications of the EM map image in classical mechanics and to associated modifications of the joint spectrum in quantum mechanics.

The simplest classical bifurcation leading to modification of the image of the EM map is the Hamiltonian Hopf bifurcation [79]. It is associated with the following modification of the image of the EM map. The critical value of the EM map situated on the boundary leaves the boundary and enters the internal domain of regular values (see Fig. 11). As a consequence, the toric fibration over the closed path surrounding an isolated singularity is non-trivial. Its non-triviality can be characterized by the Hamiltonian monodromy which describes the mapping from the fundamental group of the base space into the first homology group of the regular fiber [18]. A typical pattern of the joint spectrum around such a classical singularity is shown in Fig. 3. The joint spectrum manifests the presence of quantum monodromy. Its interpretation in terms of regular lattices is given in Figs. 4 and 5.

Taking into account additional terms of higher order it is possible to distinguish different types of Hamiltonian

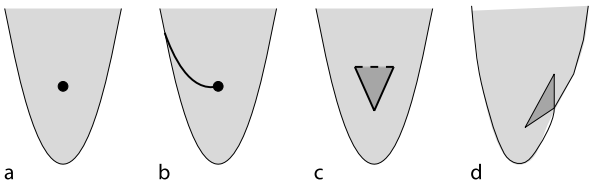

Quantum Bifurcations, Figure 9
Typical images of the energy momentum map for completely integrable Hamiltonian systems with two degrees-of-freedom in the case of: a integer monodromy, b fractional monodromy, c non-local monodromy, and d bidromy. Values in the *light shaded area* lift to single 2-tori; values in the *dark shaded area* lift to two 2-tori. Taken from [69]

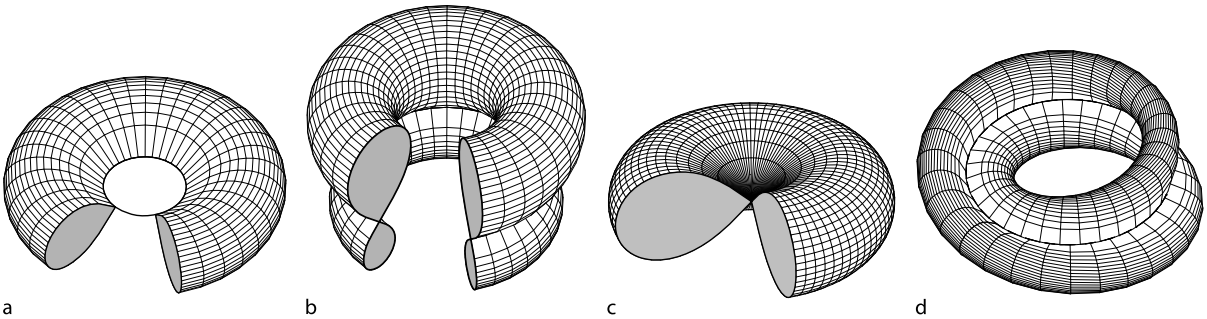

Quantum Bifurcations, Figure 10
Two-dimensional singular fibers in the case of integrable Hamiltonian systems with two degrees-of-freedom (*left* to *right*): singular torus, bitorus, pinched and curled tori. Singular torus corresponds to critical values in Fig. 9c, d (ends of bitoris line). Bitorus corresponds to critical values in Fig. 9c, d, which belong to singular line (fusion of two components). Pinched torus corresponds to isolated focus-focus singularity in Fig. 9a. Curled torus is associated with critical values at singular line in Fig. 9b (fractional monodromy). Taken from [69]

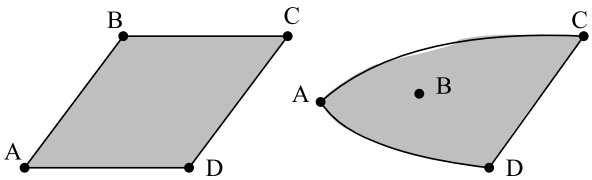

Quantum Bifurcations, Figure 11

Qualitative modification of the image of the EM map due to Hamiltonian Hopf bifurcation. *Left*: Simplest integrable toric fibration over $S^2 \times S^2$ classical phase space. *A*, *B*, *C*, *D*: Critical values corresponding to singular S^0 fibers. Regular points on the boundary correspond to S^1 fibers. Regular internal points: Regular T^2 fibers. *Right*: Appearance of an isolated critical value inside the field of regular values. Critical value *B* corresponds to pinched torus shown in Fig. 10

Hopf bifurcations usually named as subcritical and supercritical [19,79]. New qualitative modification, for example, corresponds to transformation of an isolated singular value of the EM map into an "island", i. e. the region of the EM image filled by points whose inverse images consist of two connected components. Integrable approximation for vibrational motion in the LiCN molecule shows the presence of such an island associated with the non-local quantum monodromy (see Fig. 12) [41]. The monodromy naturally coincides with the quantum monodromy of isolated focus-focus singularity which deforms continuously into the island monodromy. It is interesting to note that in molecule HCN which is rather similar to LiCN, the region with two components in the inverse image of the EM map exists also but the monodromy cannot be defined due to impossibility to go around the island [22].

In the quantum problem the presence of "standard" quantum monodromy in the joint spectrum of two mutually commuting observables can be seen through the mapping of a locally regular part of the joint spectrum lattice to an idealized Z^2 lattice. Existence of local actions for the classical problem which are defined almost everywhere and the multivaluedness of global actions from one side and the quantum-classical correspondence from another side allow the interpretation of the joint spectrum with quantum monodromy as a regular lattice with an isolated defect.

Recently, the generalization of the notion of quantum (and classical) monodromy was suggested [21,58]. For quantum problems the idea is based on the possibility to study instead of the complete lattice formed by the joint spectrum only a sub-lattice of finite index. Such a transformation allows one to eliminate certain "weak line sin-

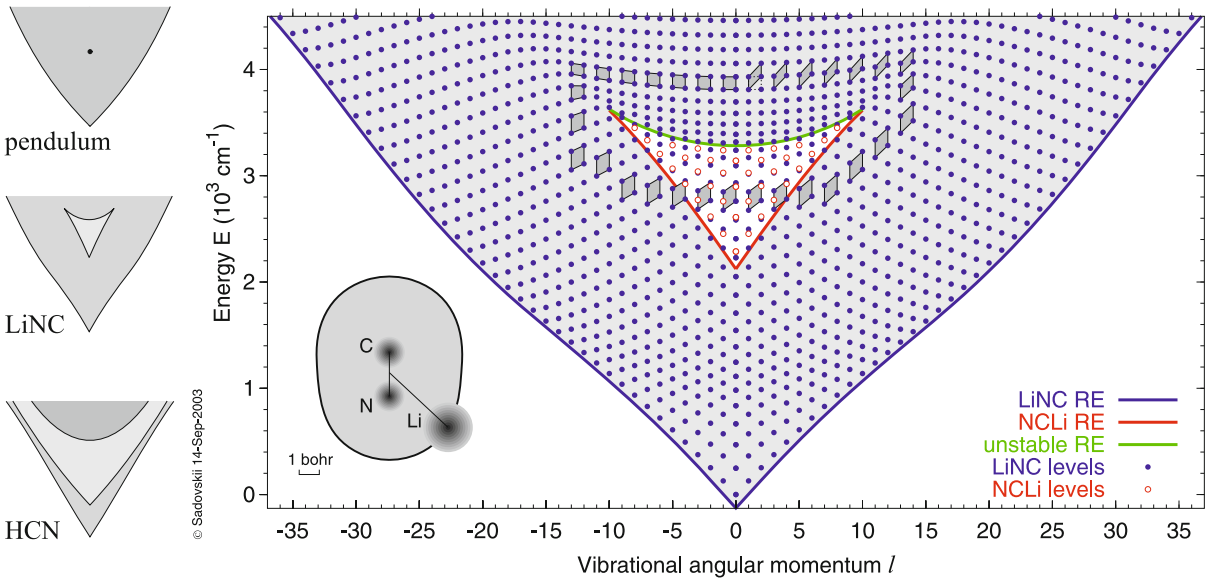

Quantum Bifurcations, Figure 12

Quantum joint spectrum for the quantum model problem with two degrees-of-freedom describing two vibrations in the LiCN molecule. The non-local quantum monodromy is shown by the evolution of the elementary cell of the quantum lattice around the singular line associated with gluing of two regular lattices corresponding in molecular language to two different isomers, LiCN and LiNC. Classical limit (*left*) shows the possible deformation of isolated focus-focus singularity for pendulum to non-local island singularity for LiNC model. In contrast to LiCN, the HCN model has an infinite island which cannot be surrounded by a close path. Taken from [41]

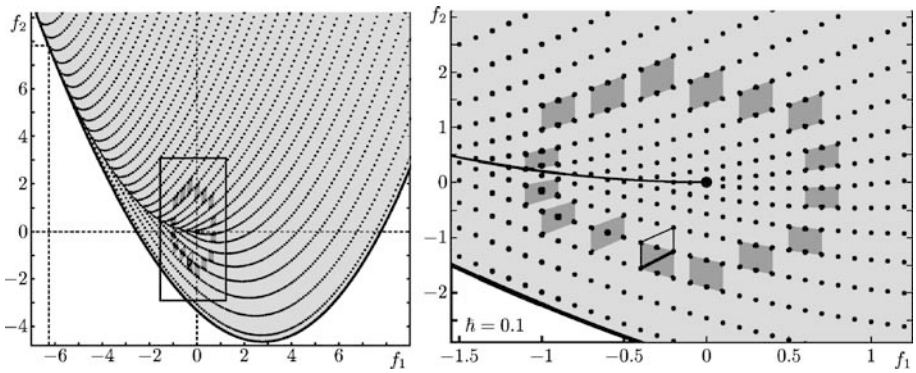

Quantum Bifurcations, Figure 13
Joint quantum spectrum for two-dimensional non-linear 1 : (−2) resonant oscillator (5),(6). The singular line is formed by critical values whose inverse images are curled tori shown in Fig. 10. In order to get the unambiguous result of the propagation of the cell of the quantum lattice along a closed path crossing the singular line, the elementary cell is doubled. Taken from [58]

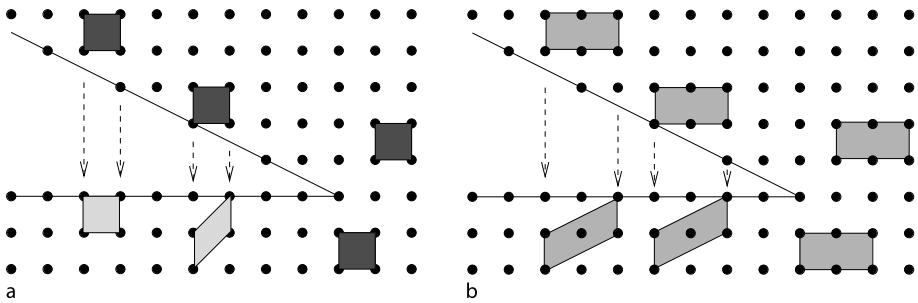

Quantum Bifurcations, Figure 14
Representation of a lattice with 1 : 2 rational defect by cutting and gluing. *Left*: The elementary cell goes through cut in an ambiguous way. The result depends on the place where the cell crosses the cut. *Right*: Double cell crosses the cut in an unambiguous way. Taken from [58]

gularities" presented in the image of the EM map. The resulting monodromy is named "fractional monodromy" because for the elementary cell in the regular region the formal transformation after a propagation along a close path crossing "weak line singularities" turns out to be represented in a form of a matrix with fractional coefficients.

An example of quantum fractional monodromy can be given with a 1 : (−2) resonant oscillator system possessing two integrals of motion f_1, f_2 in involution:

$$f_1 = \frac{\omega}{2}\left(p_1^2 + q_1^2\right) - \frac{2\omega}{2}\left(p_2^2 + q_2^2\right) + R_1(q,p)\,, \qquad (5)$$

$$f_2 = \mathrm{Im}\left[(q_1 + ip_1)^2(q_2 + ip_2)\right] + R_2(q,p)\,. \qquad (6)$$

The corresponding joint spectrum for the quantum problem is shown in Fig. 13. It can be represented as a regular Z^2 lattice with a solid angle removed (see Fig. 14). The main difference with the standard integer monodromy representation is due to the fact that even after gluing two sides of the cut we get the one-dimensional singular stratum which can be neglected only after going to a sub-lattice (to a sub-lattice of index 2 for 1 : 2 fractional singularity).

Another kind of generalization of the monodromy notion is related to the appearance of multi-component inverse images for the EM maps. We have already mentioned such a possibility with the appearance of non-local monodromy and Hamiltonian Hopf bifurcations (see Fig. 12). But in this case two components of the inverse image belong to different regular domains and cannot be joined by a path going only through regular values. Another possibility is suggested in [69,70] and is explained schematically in Fig. 15. This figure shows that the arrangement of fibers can be done in such a way that one connected component can be deformed into another connected component along a path which goes only through regular tori. The existence of a quantum joint spectrum corresponding to such a classical picture was demon-

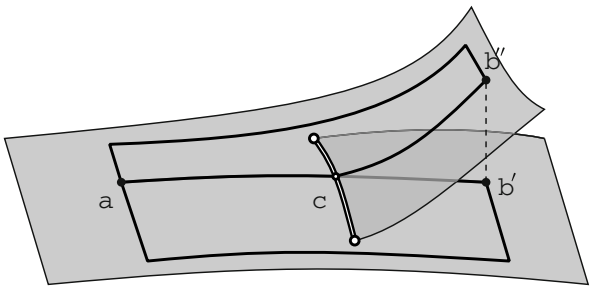

Quantum Bifurcations, Figure 15
Schematic representation of the inverse images for a problem with bidromy in the form of the unfolded surface. Each connected component of the inverse image is represented as a *single point*. The path $b' - a - b''$ starts and ends at the same point of the space of possible values of integrals of motion but it starts at one connected component and ends at another one. At the same time the path goes only through regular tori. Taken from [70]

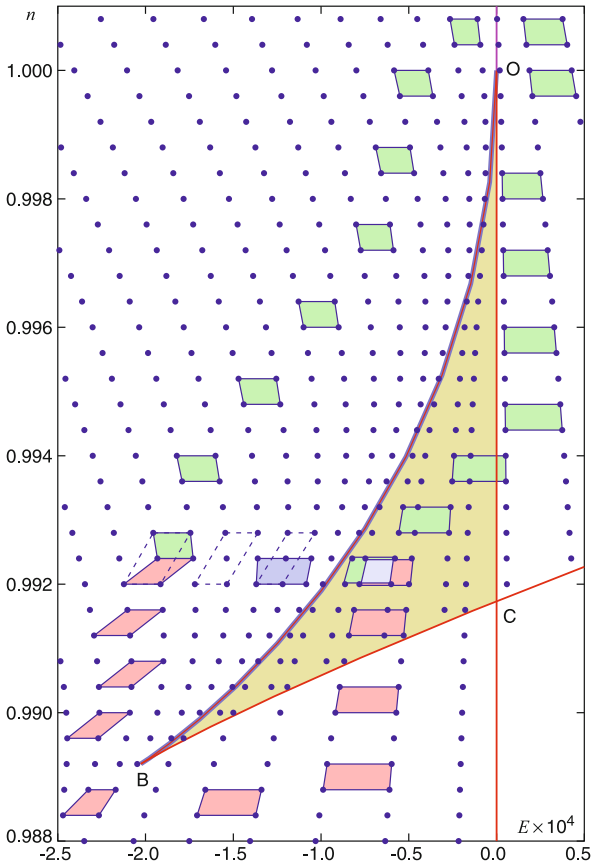

Quantum Bifurcations, Figure 16
Joint quantum spectrum for problem with bidromy. Quantum states are given by two numbers (energy, *E*, and polyad number, *n*) which are the eigenvalues of two mutually commuting operators. Inside the *OAB* curvilinear triangle two regular lattices are clearly seen. One can be continued smoothly through the *OC* boundary whereas another continues through the *BC* boundary. This means that the regular part of the whole lattice can be considered as a one self-overlapping regular lattice. The figure suggests also the possibility to define the propagation of a double cell along a "bipath" through the singular line *BO* which leads to splitting of the cell into two elementary cells fusing at the end into one cell defining in such a way the "bidromy" transformation associated with a bipath. Taken from [70]

strated on the example of a very well-known model problem with three degrees-of-freedom: Three resonant oscillators with 1 : 1 : 2 resonance, axial symmetry and with small detuning between double degenerate and non-degenerate modes [30,70]. The specific behavior of the joint spectrum for this model can be characterized as self-overlapping of a regular lattice. The possibility to propagate the initially chosen cell through a regular lattice from the region of self-overlapping of lattice back to the same region but to another component was named "bidromy". More complicated construction for the same problem allows us to introduce the "bipath" notion. The bipath starts at a regular point of the EM image, and crosses the singular line by splitting itself into two components. Each component belongs to its proper lattice in the self-overlapping region. Two components of the path can go back through the regular region only and fuse together. The behavior of quantum cells along a bipath is shown in Fig. 16. Providing a rigorous mathematical description of such a construction is still an open problem. Although the original problem has three degrees-of-freedom, it is possible to construct a model system with two degrees-of-freedom and with similar properties.

Bifurcations of "Quantum Bifurcation Diagrams"

We want now to stress some differences in the role of internal and external control parameters. From one point-of-view a quantum problem, which corresponds in the classical limit to a multidimensional integrable classical model, possesses a joint spectrum qualitatively described by a "quantum bifurcation diagram". This diagram shows that the joint spectrum is formed from several parts of regular lattices through a cutting and gluing procedure. Going from one regular region to another is possible by crossing singular lines. The parameter defined along such a path can be treated as an internal control parameter. It is essentially a function of values of integrals of motion. To cross the singular line is equivalent to passing the quantum bifurcation for a family of reduced systems with a smaller number of degrees of freedom.

On the other side we can ask the following more general question. What kinds of generic modifications of "bifurcation diagrams" are possible for a family of integrable systems depending on some external parameters? Hamiltonian Hopf bifurcation leading to the appearance of a new isolated singular value and as a consequence appearance of monodromy is just one of the possible effects of this kind. Another possibility is the transformation of an isolated focus-focus singular value into the island associated with the presence of a second connected component of the inverse image of the EM map. It is also possible that such an island is born within the regular region of the EM map. In such a case naturally the monodromy transformation associated with a closed path surrounding the so-obtained island should be trivial (identity).

The boundary of the image of the EM map can also undergo transformation which results in the appearance of the region with two components in the inverse image but, in contrast to the previous example of the appearance of an island, these two components can be smoothly deformed one onto another along a continuous path going only through regular values of the EM map. Examples of all such modifications were studied on simple models inspired by concrete quantum molecular systems like the H atom, CO_2, LiCN molecules and so on [24,30,41].

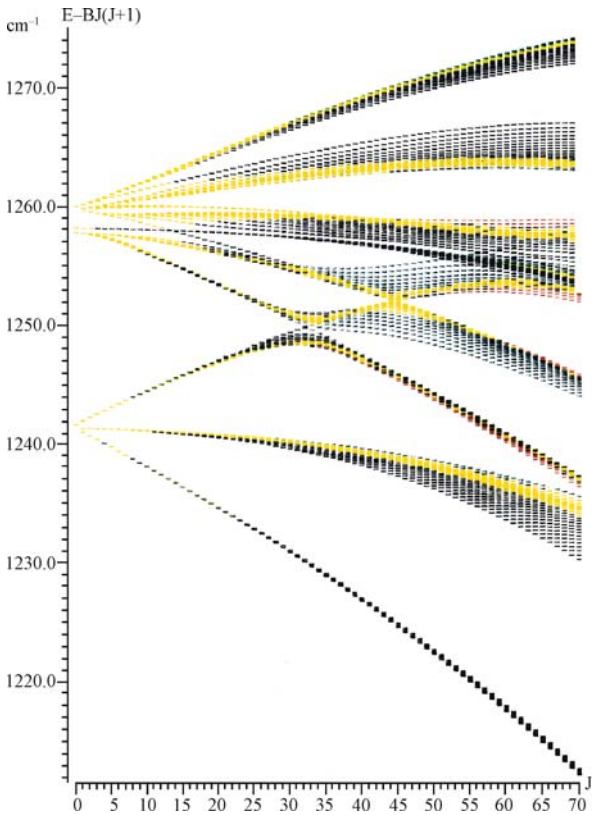

Quantum Bifurcations, Figure 17
System of rovibrational energy levels of $^{13}CF_4$ molecule represented schematically in E, J coordinates. The number of energy levels in each clearly seen band is $2J + 1 + \delta$, where δ is a small integer which remains constant for isolated bands and changes at band intersections. In the semi-quantum model δ is interpreted as the first Chern class, characterizing the non-triviality of the vector bundle formed by eigenfunctions of the "fast" subsystem over the classical phase space of the "slow" subsystem [27]

Semi-Quantum Limit and Reorganization of Quantum Bands

Up to now we have discussed the qualitative modifications of internal structures of certain groups of quantum levels which are typically named bands. Their appearance is physically quite clear in the adiabatic approximation. The existence of fast and slow classical motions manifests itself in quantum mechanics through the formation of so-called energy bands. The big energy difference between energies of different bands correspond to fast classical variables whereas small energy differences between energy levels belonging to the same band correspond to classical slow variables. Typical bands in simple quantum systems correspond to vibrational structure of different electronic states, rotational structure of different vibrational states, etc.

If now we have a quantum problem which shows the presence of bands in its energy spectrum, the natural generalization consists of putting this quantum system in a family, depending on one (or several) control parameters. What are the generic qualitative modifications which can be observed within such a family of systems when control parameters vary? Apart from qualitative modifications of the internal structure of individual bands which can be treated as the earlier discussed quantum bifurcations, another qualitative phenomenon is possible, namely the redistribution of energy levels between bands or more generally, the reorganization of bands under the variation of some control parameters [8,26,28,62,68]. In fact this phenomenon is very often observed in both the numerical simulations and the real experiments with molecular systems exhibiting bands. A typical example of molecular rovibrational energy levels classified according to their energy and angular momentum is shown in Fig. 17. It is important to note that the number of energy levels in bands before and after their "intersection" changes.

The same phenomenon of the redistribution of energy levels between energy bands can be understood by the example of a much simpler quantum system of two coupled angular momenta, say orbital angular momentum

and spin in the presence of a magnetic field interacting only with spin [62,68].

$$H = \frac{1-\gamma}{\mathbf{S}} S_z + \frac{\gamma}{\mathbf{NS}} (\mathbf{N} \cdot \mathbf{S}) , \quad 0 \le \gamma \le 1 . \tag{7}$$

The Hamiltonian for such a system can be represented in the form of a one-parameter family (7) having two natural limits corresponding to uncoupled and coupled angular momenta. The interpolation of eigenvalues between these two limits is shown in Fig. 18 for different values of spin quantum number, $S = 1/2, 1, 3/2$. The quantum number of orbital momentum is taken to be $N = 4$. Although this value is not much larger than the S values, the existence of bands and their reorganization under the variation of the external parameter γ is clearly seen in the figure.

Although the detailed description of this reorganization of bands will take us rather far away from the principal subject it is important to note that in the simplest situations there exists a very close relation between the redistribution phenomenon and the Hamiltonian Hopf bifurcations leading to the appearance of Hamiltonian monodromy [81]. In the semi-quantum limit when part of the dynamical variables are treated as purely classical and all the rest as quantum, the description of the complete system naturally leads to a fiber bundle construction [27]. The role of the base space is taken by the classical phase space for classical variables. A set of quantum wave-functions associated with one point of the base space forms a complex fiber. As a whole the so-obtained vector bundle with complex fibers can be topologically characterized by its rank and Chern classes [56]. Chern classes are related to the number of quantum states in bands formed due to quantum character of the total problem with respect to "classical" variables. Modification of the number of states in bands can occur only at band contact and is associated with the modification of Chern classes of the corresponding fiber bundle [26]. The simplest situation takes place when the number of degrees of freedom associated with classical variables is one. In this case only one topological invariant – the first Chern class is sufficient to characterize the non-triviality of the fiber bundle and the difference in Chern classes is equal to the number of energy levels redistributed between corresponding bands. Moreover, in the generic situation (in the absence of symmetry) the typical behavior consists of the redistribution of only one energy level between two bands. The generic phenomena become more complicated with increasing the number of degrees of freedom for the classical part of variables. The model problem with two slow degrees of freedom (described in classical limit by the CP^2 phase space) and three quantum states was studied in [28]. A new qualitative phenomenon was found, namely, the modification of the number of bands due to formation of topologically coupled bands. Figure 19 shows the evolution of the system of energy levels along with the variation of control parameter λ. Three quantum bands (at $\lambda = 0$) transform into two bands (in the $\lambda = 1$) limit. One of these bands has rank one, i. e. it is associated with one quantum state. Another has rank two. It is associated with two quantum states. Both bands have non-trivial topology (non-trivial Chern classes). Moreover, it is quite important that the newly formed topologically coupled band of rank two can

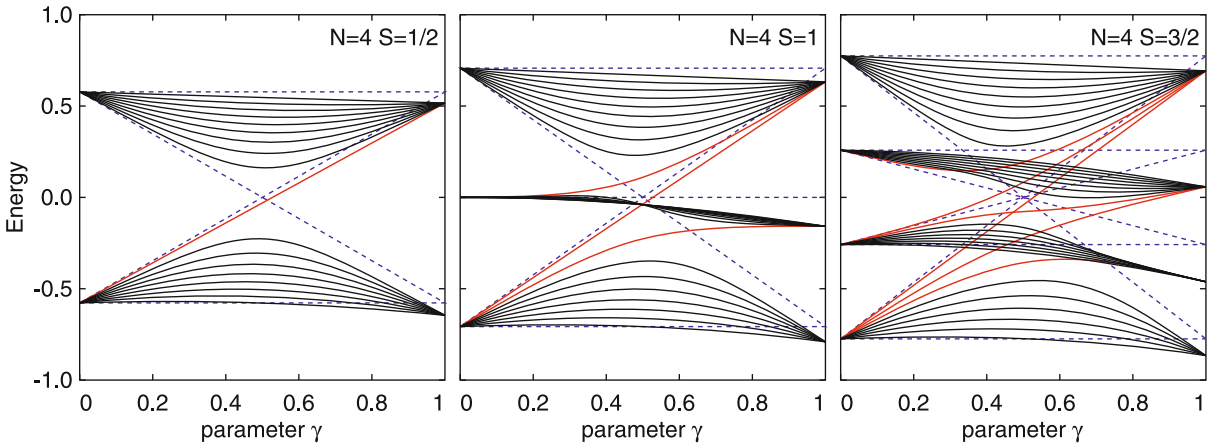

Quantum Bifurcations, Figure 18

Rearrangement of energy levels between bands for model Hamiltonian (7) with two, three, or four states for "fast" variable. Quantum energy levels are shown by *solid lines*. Classical energies of stationary points for energy surfaces are shown by *dashed lines*. Taken from [68]

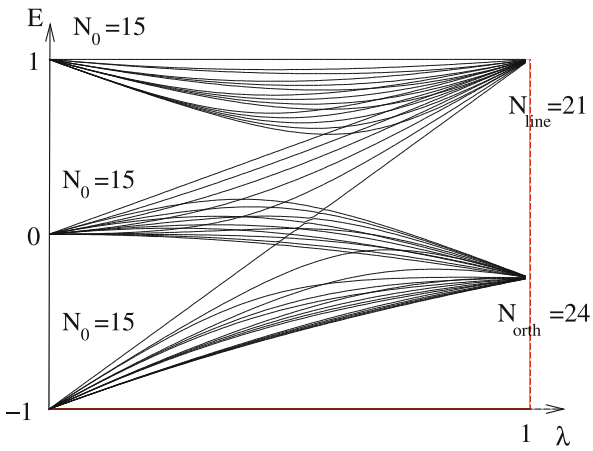

Quantum Bifurcations, Figure 19
Rearrangement of three bands into two topologically non-trivially coupled bands. Example of a model with three electronic states and vibrational structure of polyads formed by three quasi-degenerate modes. At $\lambda = 0$ three bands have each the same number of states, namely 15. In the classical limit each initial band has rank one and trivial topology. At $\lambda = 1$ there are only two bands. One of them has rank 2 and non-trivial first and second Chern classes. Taken from [28]

be split into two bands of rank one only if a coupling with the third band is introduced.

The corresponding qualitative modifications of quantum spectra can be considered as natural generalizations of quantum bifurcations and probably should be treated as topological bifurcations. Thus, the description of possible "elementary" rearrangements of energy bands is a direct consequence of topological restrictions imposed by a fiber bundle structure of the studied problem.

It is interesting to mention here the general mathematical problem of finding proper equivalence or better to say correspondence between some construction made over real numbers and their generalizations to complex numbers and quaternions. This paradigm of complexification and quaternization was discussed by Arnold [4,5] on many different examples. The closest to the present subject is the example of complexification of the Wigner–Neumann non-crossing rule resulting in a quantum Hall effect (in physical terms). In fact, the mathematical basis of the quantum Hall effect is exactly the same fiber bundle construction which explains the redistribution of energy levels between bands in the above-mentioned simple quantum mechanical model.

Multiple Resonances and Quantum State Density

Rearrangement of quantum energy states between bands is presented in the previous section as an example of a generic qualitative phenomenon occurring under variation of a control parameter. One possible realization of bands is the sequence of vibrational polyads formed by a system of resonant vibrational modes indexed by the polyad quantum number. In the classical picture this construction corresponds to the system of oscillators reduced with respect to the global action. The reduced classical phase space is in such a case the weighted projective space. In the case of particular $1:1:\ldots:1$ resonance the corresponding reduced phase space is a normal complex projective space CP^n. The specific resonance conditions impose for a quantum problem specific conditions on the numbers of quantum states in polyads. In the simplest case of harmonic oscillators with $n_1 : n_2 : \ldots : n_k$ resonance the numbers of states in polyads are given by the generating function

$$g = \frac{1}{(1-t^{n_1})(1-t^{n_2})\cdots(1-t^{n_k})} = \sum_N C_N\, t^N\,, \quad (8)$$

where N is the polyad quantum number. Numbers C_N are integers for integer N values, but they can be extended to arbitrary N values and represented in the form of a quasi-polynomial, i.e, a polynomial in N with coefficients being a periodic function whose period equals the least common multiplier of n_i, $i = 1, \ldots, k$. Moreover, the coefficients of the polynomial can be expressed in terms of so-called Todd polynomials which indicates the possibility of topological interpretation of such information [52,89].

Physical Applications and Generalizations

The most clearly seen physical applications of quantum bifurcations is the qualitative modification of the rotational multiplet structure under rotational excitation, i. e. under the variation of the absolute value of the angular momentum. This is related first of all with the experimental possibility to study high J multiplets (which are quite close to the classical limit but nevertheless manifest their quantum structure) and to the possibility to use symmetry arguments, which allow one to distinguish clusters of states before and after bifurcation just by counting the number of states in the cluster, which depends on the order of group of stabilizer. Nuclear rotation is another natural example of quantum rotational bifurcations [60]. Again the interest in corresponding qualitative modifications is due to the fact that rotational bands are extremely well studied up to very high J values. But in contrast to molecular physics examples, in nuclear physics it mostly happens that only ground states (for each value of J) are known. Thus, one speaks more often about qualitative changes of

the ground state (in the absence of temperature) named quantum phase transitions [65].

Internal structure of vibrational polyads is less evident for experimental verifications of quantum bifurcations, but it gives many topologically non-trivial examples of classical phase spaces on which the families of Hamiltonians depending on parameters are defined [25,30,39,42,44,46,66,76,77,86]. The main difficulty here is the small number of quantum states in polyads accessible to experimental observations. But this problem is extremely interesting from the point-of-view of extrapolation of theoretical results to the region of higher energy (or higher polyad quantum numbers) which is responsible as a rule for many chemical intra-molecular processes. Certain molecules, like CO_2, or acetylene (C_2H_2) are extremely well studied and a lot of highly accurate data exist. At the same time the qualitative understanding of the organization of excited states even in these molecules is not yet completed and new qualitative phenomena are just starting to be discovered.

Among other physically interesting systems it is necessary to mention model problems suggested to study the behavior of Bose condensates or quantum qubits [37,38,74,83,84]. These models have a mathematical form which is quite similar to rotational and vibrational models. At the same time their physical origin and the interpretation of results is quite different. This is not an exception. For example, the model Hamiltonian corresponding in the classical limit to a Hamiltonian function defined over S^2 classical phase space is relevant to rotational dynamics, description of internal structure of vibrational polyads formed by two (quasi)degenerate modes, in particular to so-called local-normal mode transition in molecules, interaction of electromagnetic field with a two-level system, the Lipkin–Meshkov–Glick model in nuclear physics, entanglement of qubits, etc.

Future Directions

To date many new qualitative phenomena have been suggested and observed in experimental and numerical studies due to intensive collaboration between mathematicians working in dynamical system theory, classical mechanics, complex geometry, topology, etc., and molecular physicists using qualitative mathematical tools to classify behavior of quantum systems and to extrapolate this behavior from relatively simple (low energy regions) to more complicated ones (high energy regions). Up to now the main accent was placed on the study of the qualitative features of isolated time-independent molecular systems. Specific patterns formed by energy eigenvalues and by common eigenvalues of several mutually commuting observables were the principal subject of study. Existence of qualitatively different dynamical regimes for time-independent problems at different values of exact or approximate integrals of motion were clearly demonstrated. Many of these new qualitative features and phenomena are supposed to be generic and universal although their rigorous mathematical formulation and description is still absent.

On the other side, the analysis of the time-dependent processes should be developed. This step is essential in order to realize at the level of quantum micro-systems the transformations associated with the qualitative modifications of dynamical regimes and to control such time-dependent processes as elementary reactions, information data storage, and so on. From this global perspective the main problem of the future development is to support the adequate mathematical formulation of qualitative methods and to improve our understanding of qualitative modifications occurring in quantum micro-systems in order to use them as real micro-devices.

Bibliography

1. Arnold VI (1972) Modes and quasimodes. Funct Anal Appl 6:94–101
2. Arnold VI (1989) Mathematical methods of classical mechanics. Springer, New York
3. Arnold VI (1992) Catastrophe theory. Springer, Berlin
4. Arnold VI (1995) Remarks on eigenvalues and eigenvectors of Hermitian matrices, Berry phase, adiabatic connections and quantum Hall effect. Selecta Math New Ser 1:1–19
5. Arnold VI (2005) Arnold's problems. Springer, Berlin
6. Aubry S, Flach S, Kladko K, Olbrich E (1996) Manifestation of classical bifurcation in the spectrum of the integrable quantum dimer. Phys Rev Lett 76:1607–1610
7. Bolsinov AV, Fomenko AT (2004) Integrable Hamiltonian systems. Geometry topology classifications. Chapman and Hall/CRC, London
8. Brodersen S, Zhilinskií BI (1995) Transfer of clusters between the vibrational components of CF_4. J Mol Spectrosc 169:1–17
9. Child MS (2000) In: Jensen P, Bunker PR (eds) Computational molecular spectroscopy, chapter 18. Wiley Interscience, Chichester
10. Child MS (2001) Quantum level structure and nonlinear classical dynamics. J Mol Spectrosc 210:157–165
11. Child MS, Weston T, Tennyson J (1999) Quantum monodromy in the spectrum of H_2O and other systems. Mol Phys 96: 371–379
12. Colin de Verdier Y, Vũ Ngoc S (2003) Singular Bohr–Sommerfeld rules for 2D integrable systems. Ann Ec Norm Sup 36:1–55
13. Cushman RS, Bates L (1997) Global aspects of classical integrable systems. Birkhäuser, Basel
14. Cushman RH, Duistermaat JJ (1988) The quantum mechanical spherical pendulum. Bull Am Math Soc 19:475–479
15. Cushman RH, Sadovskií DA (2000) Monodromy in the hydrogen atom in crossed fields. Physica D 142:166–196

16. Dirac PAM (1982) The principles of quantum mechanics. Oxford University Press, Oxford
17. Duistermaat JJ (1980) On global action angle coordinates. Comm Pure Appl Math 33:687–706
18. Duistermaat JJ (1998) The monodromy in the Hamiltonian Hopf bifurcation. Angew Z Math Phys 49:156–161
19. Efstathiou K (2004) Metamorphoses of Hamiltonian systems with symmetry. Lecture Notes in Mathematics, vol 1864. Springer, Heidelberg
20. Efstathiou K, Cushman RH, Sadovskiĭ DA (2004) Hamiltonian Hopf bifurcation of the hydrogen atom in crossed fields. Physica D 194:250–274
21. Efstathiou K, Cushman RH, Sadovskiĭ DA (2007) Fractional monodromy in the $1:-2$ resonance. Adv Math 209:241–273
22. Efstathiou K, Joyeux M, Sadovskiĭ DA (2004) Global bending quantum number and the absence of monodromy in the HCN↔CNH molecule. Phys Rev A 69(3):032504-1–15
23. Efstathiou K, Sadovskiĭ DA, Zhilinskiĭ BI (2004) Analysis of rotation-vibration relative equilibria on the example of a tetrahedral four atom molecule. SIAM J Dyn Syst 3:261–351
24. Efstathiou K, Sadovskiĭ DA, Zhilinskiĭ BI (2007) Classification of perturbations of the hydrogen atom by small static electric and magnetic fields. Proc Roy Soc Lond A 463:1771–1790
25. Ezra GS (1996) Periodic orbit analysis of molecular vibrational spectra: Spectral patterns and dynamical bifurcations in Fermi resonant systems. J Chem Phys 104:26–35
26. Faure F, Zhilinskiĭ BI (2000) Topological Chern indices in molecular spectra. Phys Rev Lett 85:960–963
27. Faure F, Zhilinskiĭ BI (2001) Topological properties of the Born–Oppenheimer approximation and implications for the exact spectrum. Lett Math Phys 55:219–238
28. Faure F, Zhilinskiĭ BI (2002) Topologically coupled energy bands in molecules. Phys Lett A 302:242–252
29. Flach S, Willis CR (1998) Discrete breathers. Phys Rep 295: 181–264
30. Giacobbe A, Cushman RH, Sadovskiĭ DA, Zhilinskiĭ BI (2004) Monodromy of the quantum $1:1:2$ resonant swing spring. J Math Phys 45:5076–5100
31. Gilmore R (1981) Catastrophe theory for scientists and engineers. Wiley, New York
32. Gilmore R, Kais S, Levine RD (1986) Quantum cusp. Phys Rev A 34:2442–2452
33. Golubitsky M, Schaeffer DG (1984) Singularities and groups in bifurcation theory, vol 1. Springer, Berlin
34. Grondin L, Sadovskiĭ DA, Zhilinskiĭ BI (2002) Monodromy in systems with coupled angular momenta and rearrangement of bands in quantum spectra. Phys Rev A 142:012105-1–15
35. Guillemin V (1994) Moment maps and combinatorial invariants of Hamiltonian T^n-spaces. Birkhäuser, Basel
36. Harter W (1988) Computer graphical and semiclassical approaches to molecular rotations and vibrations. Comput Phys Rep 8:319–394
37. Hines AP, McKenzie RH, Milburn GJ (2005) Quantum entanglement and fixed-point bifurcations. Phys Rev A 71:042303-1–9
38. Hou X-W, Chen J-H, Hu B (2005) Entanglement and bifurcation in the integrable dimer. Phys Rev A 71:034302-1–4
39. Joyeux M, Farantos SC, Schinke R (2002) Highly excited motion in molecules: Saddle-node bifurcations and their fingerprints in vibrational spectra. J Phys Chem A 106:5407–5421
40. Joyeux M, Grebenschikov S, Bredenbeck J, Schinke R, Farantos SC (2005) Intramolecular dynamics along isomerization and dissociation pathways. In: Toda M, Komatsuzaki T, Konishi T, Berry RS, Rice SA (eds) Geometric Structures of Phase Space in Multidimensional Chaos: A Special Volume of Advances in Chemical Physics, part A, vol 130. Wiley, pp 267–303
41. Joyeux M, Sadovskiĭ DA, Tennyson J (2003) Monodromy of the LiNC/NCLi molecule. Chem Phys Lett 382:439–442
42. Joyeux M, Sugny D, Tyng V, Kellman ME, Ishikawa H, Field RW, Beck C, Schinke R (2000) Semiclassical study of the isomerization states of HCP. J Chem Phys 112:4162–4172
43. Kellman ME (1995) Algebraic models in spectroscopy. Annu Rev Phys Chem 46:395–422
44. Kellman ME, Lynch ED (1986) Fermi resonance phase space structure from experimental spectra. J Chem Phys 85: 7216–7223
45. Kleman M (1983) Points, lines and walls. Wiley, Chichester
46. Kozin IN, Sadovskiĭ DA, Zhilinskiĭ BI (2005) Assigning vibrational polyads using relative equilibria: Application to ozone. Spectrochim Acta A 61:2867–2885
47. Landau L, Lifschits EM (1981) Quantum mechanics, nonrelativistic theory. Elsevier, Amsterdam
48. Lu Z-M, Kellman ME (1997) Phase space structure of triatomic molecules. J Chem Phys 107:1–15
49. Marsden JE, Ratiu TS (1994) Introduction to mechanics and symmetry. Springer, New York
50. Mermin ND (1979) The topological theory of defects in ordered media. Rev Mod Phys 51:591–648
51. Michel L (1980) Symmetry defects and broken symmetry, configurations, hidden symmetry. Rev Mod Phys 52:617–651
52. Michel L, Zhilinskiĭ BI (2001) Symmetry, invariants, topology, vol I. Basic tools. Phys Rep 341:11–84
53. Michel L, Zhilinskiĭ BI (2001) Symmetry, invariants, topology, vol III. Rydberg states of atoms and molecules. Basic group theoretical and topological analysis. Phys Rep 341:173–264
54. Montaldi J, Roberts R, Stewart I (1988) Periodic solutions near equilibria of symmetric Hamiltonian systems. Philos Trans Roy Soc Lond A 325:237–293
55. Morse M (1925) Relation between the critical points of a real function of n independent variables. Trans Am Math Soc 27:345–396
56. Nakahara M (1990) Geometry, topology and physics. IOP Publishing, Bristol
57. Nekhoroshev NN (1972) Action-angle variables and their generalizations. Trans Moscow Math Soc 26:180–198
58. Nekhoroshev NN, Sadovskiĭ DA, Zhilinskiĭ BI (2006) Fractional Hamiltonian monodromy. Ann Henri Poincaré 7:1099–1211
59. Pavlichenkov I (1993) Bifurcations in quantum rotational spectra. Phys Rep 226:173–279
60. Pavlichenkov I (2006) Quantum bifurcations and quantum phase transitions in rotational spectra. Phys At Nucl 69: 1008–1013
61. Pavlichenkov I, Zhilinskiĭ BI (1988) Critical phenomena in rotational spectra. Ann Phys NY 184:1–32
62. Pavlov-Verevkin VB, Sadovskiĭ DA, Zhilinskiĭ BI (1988) On the dynamical meaning of the diabolic points. Europhys Lett 6:573–578
63. Peters AD, Jaffe C, Gao J, Delos JB (1997) Quantum manifestations of bifurcations of closed orbits in the photodetachment cross section of H^- in parallel fields. Phys Rev A 56:345–355
64. Pierre G, Sadovskiĭ DA, Zhilinskiĭ BI (1989) Organization of quantum bifurcations: Crossover of rovibrational bands in spherical top molecules. Europhys Lett 10:409–414

65. Sachdev S (1999) Quantum phase transitions. Cambridge University Press, Cambridge
66. Sadovskiĭ DA, Fulton NG, Henderson JR, Tennyson J, Zhilinskiĭ BI (1993) Nonlinear normal modes and local bending vibrations of H3+ and D3+. J Chem Phys 99(2):906–918
67. Sadovskiĭ DA, Zhilinskiĭ BI (1993) Group theoretical and topological analysis of localized vibration-rotation states. Phys Rev A 47(4):2653–2671
68. Sadovskiĭ DA, Zhilinskiĭ BI (1999) Monodromy, diabolic points, and angular momentum coupling. Phys Lett A 256:235–244
69. Sadovskiĭ DA, Zhilinskiĭ BI (2006) Quantum monodromy, its generalizations and molecular manifestations. Mol Phys 104:2595–2615
70. Sadovskiĭ DA, Zhilinskiĭ BI (2007) Hamiltonian systems with detuned 1 : 1 : 2 resonance, manifestations of bidromy. Ann Phys NY 322:164–200
71. Sadovskiĭ DA, Zhilinskiĭ BI, Champion JP, Pierre G (1990) Manifestation of bifurcations and diabolic points in molecular energy spectra. J Chem Phys 92:1523–1537
72. Sadovskiĭ DA, Zhilinskiĭ BI, Michel L (1996) Collapse of the Zeeman structure of the hydrogen atom in an external electric field. Phys Rev A 53:4064–4047
73. Simon B (1980) The classical limit of quantum partition functions. Commun Math Phys 71:247–276
74. Somma R, Ortiz G, Barnum H, Knill E, Viola L (2004) Nature and measure of entanglement in quantum phase transitions. Phys Rev A 70:042311-1–21
75. Symington M (2003) Four dimensions from two in symplectic topology. In: Athens GA (ed) Topology and geometry of manifolds. Proc Symp Pure Math, vol 71. AMS, Providence, pp 153–208
76. Tyng V, Kellman ME (2006) Bending dynamics of acetylene: New modes born in bifurcations of normal modes. J Phys Chem B 119:18859–18871
77. Uwano Y (1999) A quantum saddle-node bifurcation in a resonant perturbed oscillator with four parameters. Rep Math Phys 44:267–274
78. Uzer T (1990) Zeeman effect as an asymmetric top. Phys Rev A 42:5787–5790
79. Van der Meer JC (1985) The Hamiltonian Hopf bifurcation. Lect Notes Math, vol 1160. Springer, New York
80. Vũ Ngoc S (1999) Quantum monodromy in integrable systems. Comm Math Phys 203:465–479
81. Vũ Ngoc S (2007) Moment polytopes for symplectic manifolds with monodromy. Adv Math 208:909–934
82. Waalkens H, Dullin HR (2001) Quantum monodromy in prolate ellipsoidal billiards. Ann Phys NY 295:81–112
83. Wang J, Kais S (2004) Scaling of entanglement at a quantum phase transition for a two-dimensional array of quantum dots. Phys Rev A 70:022301-1-4
84. Weyl H (1952) Symmetry. Princeton University Press, Princeton
85. Winnewisser M, Winnewisser B, Medvedev I, De Lucia FC, Ross SC, Bates LM (2006) The hidden kernel of molecular quasi-linearity: Quantum monodromy. J Mol Struct V 798:1–26
86. Xiao L, Kellman ME (1990) Catastrophe map classification of the generalized normal-local transition in Fermi resonance spectra. J Chem Phys 93:5805–5820
87. Zhang W-M, Feng DH, Gilmore R (1990) Coherent states: Theory and some applications. Rev Mod Phys 62:867–927
88. Zhilinskiĭ BI (1996) Topological and symmetry features of intramolecular dynamics through high resolution molecular spectroscopy. Spectrochim Acta A 52:881–900
89. Zhilinskiĭ BI (2001) Symmetry, invariants, and topology, vol II. Symmetry, invariants, and topology in molecular models. Phys Rep 341:85–171
90. Zhilinskiĭ BI (2006) Hamiltonian monodromy as lattice defect. In: Monastyrsky M (ed) Topology in condensed matter. Springer series in solid state sciences, vol 150. Springer, Berlin, pp 165–186
91. Zhilinskiĭ BI, Kozin I, Petrov S (1999) Correlation between asymmetric and spherical top: Imperfect quantum bifurcations. Spectrochim Acta A 55:1471–1484
92. Zhilinskiĭ BI, Pavlichenkov IM (1988) Critical phenomenon in the rotational spectra of water molecule. Opt Spectrosc 64:688–690
93. Zhilinskiĭ BI, Petrov SV (1996) Nonlocal bifurcation in the rotational dynamics of an isotope-substituted $A_2A_2^*$ molecule. Opt Spectrosc 81:672–676

Quantum Cellular Automata

Karoline Wiesner[1,2]
[1] School of Mathematics, University of Bristol, Bristol, UK
[2] Centre for Complexity Sciences, University of Bristol, Bristol, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Cellular Automata
Early Proposals
Models of QCA
Computationally Universal QCA
Modeling Physical Systems
Implementations
Future Directions
Bibliography

Glossary

Configuration The state of all cells at a given point in time.

Neighborhood All cells with respect to a given cell that can affect this cell's state at the next time step. A neighborhood always contains a finite number of cells.

Space-homogeneous The transition function / update table is the same for each cell.

Time-homogeneous The transition function / update table is time-independent.

Update table Takes the current state of a cell and its neighborhood as an argument and returns the cell's state at the next time step.

Schrödinger picture Time evolution is represented by a quantum state evolving in time according to a time-independent unitary operator acting on it.

Heisenberg picture Time evolution is represented by observables (elements of an operator algebra) evolving in time according to a unitary operator acting on them.

BQP complexity class *Bounded error, quantum probabilistic*, the class of decision problems solvable by a quantum computer in polynomial time with an error probability of at most 1/3.

QMA complexity class *Quantum Merlin–Arthur*, the class of decision problems such that a "yes" answer can be verified by a 1-message quantum interactive proof (verifiable in BQP).

Quantum Turing machine A quantum version of a Turing machine – an abstract computational model able to compute any computable sequence.

Swap operation The one-qubit unitary gate $U = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

Hadamard gate The one-qubit unitary gate

$$U = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

Phase gate The one-qubit unitary gate $U = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\phi} \end{pmatrix}$

Pauli operator The three Pauli operators are

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

Qubit 2-state quantum system, representable as vector $a\,|0\rangle + b\,|1\rangle$ in complex space with $a^2 + b^2 = 1$.

Definition of the Subject

Quantum cellular automata (QCA) are a generalization of (classical) cellular automata (CA) and in particular of reversible CA. The latter are reviewed shortly. An overview is given over early attempts by various authors to define one-dimensional QCA. These turned out to have serious shortcomings which are discussed as well. Various proposals subsequently put forward by a number of authors for a general definition of one- and higher-dimensional QCA are reviewed and their properties such as universality and reversibility are discussed.

Quantum cellular automata (QCA) are a quantization of classical cellular automata (CA), d-dimensional arrays of cells with a finite-dimensional state space and a local, spatially-homogeneous, discrete-time update rule. For QCA each cell is a finite-dimensional quantum system and the update rule is unitary. CA as well as some versions of QCA have been shown to be computationally universal. Apart from a theoretical interest in a quantized version of CA, QCA are a natural framework for what is most likely going to be the first application of quantum computers – the simulation of quantum physical systems. In particular, QCA are capable of simulating quantum dynamical systems whose dynamics are uncomputable by classical means. QCA are now considered one of the standard models of quantum computation next to quantum circuits and various types of measurement-based quantum computational models[1]. Unlike their classical counterpart, an axiomatic, all-encompassing definition of (higher-dimensional) QCA is still missing.

Introduction

Automata theory is the study of abstract computing devices and the class of functions they can perform on their inputs. The original concept of cellular automata is most strongly associated with John von Neumann (*1903, †1957), a Hungarian mathematician who made major contributions to a vast range of fields including quantum mechanics, computer science, functional analysis and many others. According to Burks, an assistant of von Neumann, [45] von Neumann had posed the fundamental questions: "What kind of logical organization is sufficient for an automaton to reproduce itself?". It was Stanislav Ulam who suggested to use the framework of cellular automata to answer this question. In 1966 von Neumann presented a detailed analysis of the above question in his book *Theory of Self-Reproducing Automata* [45].

Thus, von Neumann initiated the field of cellular automata. He also made central contributions to the mathematical foundations of quantum mechanics and, in a sense von Neumann's quantum logic ideas were an early attempt at defining a computational model of physics. But he did not pursue this, and did not go in the directions that have led to modern ideas of quantum computing in general or quantum cellular automata in particular.

The idea of quantum computation is generally attributed to Feynman who, in his now famous lecture in 1981, proposed a computational scheme based on quantum mechanical laws [19]. A contemporary paper by Benioff contains the first proposal of a quantum Turing machine [6]. The general idea was to devise a computa-

[1] For details on these and other aspects of quantum computation see the article by Kendon in this Encyclopedia.

tional device based on and exploiting quantum phenomena that would outperform any classical computational device. These first proposals were sequentially operating quantum mechanical machines imitating the logical operations of classical digital computation. The idea of parallelizing the operations was found in classical cellular automata. However, how to translate cellular automata into a quantum mechanical framework turned out not to be trivial. And to a certain extent how to do this in general remains an open question until today.

The study of quantum cellular automata (QCA) started with the work of Grössing and Zeilinger who coined the term QCA and provided a first definition [21]. Watrous developed a different model of QCA [46]. His work lead to further studies by several groups [16,17,18]. Independently of this, Margolus developed a parallelizable quantum computational architecture building on Feynman's original ideas [26]. For various reasons to be discussed below, none of these early proposals turned out to be physical. The study of QCA gained new momentum with the work by Richter, Schumacher, and Werner [36,37] and others [3,4,34] who avoided unphysical behavior allowed by the early proposals [4,37]. It is important to notice that in spite of the over two-decade long history of QCA there is no single agreed-upon definition of QCA, in particular of higher-dimensional QCA. Nevertheless, many useful properties have been shown for the various models. Most importantly, quite a few models were shown to be computationally universal, i. e. they can simulate any quantum Turing machine and any quantum circuit efficiently [16,34,35,38,46]. Very recently, their ability to generate and transport entanglement has been illustrated [14].

A comment is in order on a class of models which is often labeled as QCA but in fact are classical cellular automata implemented in quantum mechanical structures. They do not exploit quantum effects for the actual computation. To make this distinction clear they are now called *quantum-dot QCA*. These types of QCA will not be discussed here.

Cellular Automata

Definition (Cellular Automata) A *cellular automaton* (CA) is a 4-tuple $(\mathrm{L}, \Sigma, \mathcal{N}, f)$ consisting of (1) a d-dimensional lattice of cells L indexed $i \in \mathbb{Z}^d$, (2) a finite set of states Σ, (3) a finite neighborhood scheme $\mathcal{N} \subset \mathbb{Z}^d$, and (4) a local transition function $f\colon \Sigma^{\mathcal{N}} \to \Sigma$.

A CA is discrete in time and space. It is *space and time homogeneous* if at each time step the same transition function, or *update rule*, is applied simultaneously to all cells. The update rule is *local* if for a given lattice L and lattice site x, $f(x)$ is localized in $x+\mathcal{N} = \{x+n | x \in \mathrm{L}, n \in \mathcal{N}\}$, where $\mathcal{N}$ is the *neighborhood scheme* of the CA. In addition to the locality constraint the local transition function f must generate a unique global transition function mapping a lattice *configuration* $C_t \in \Sigma^{\mathrm{L}}$ at time t to a new configuration C_{t+1} at time $t+1$: $F\colon \Sigma^{\mathrm{L}} \to \Sigma^{\mathrm{L}}$. Most CA are defined on infinite lattices or, alternatively, on finite lattices with periodic boundary conditions. For finite CA only a finite number of cells is not in a *quiescent* state, i. e. a state that is not effected by the update.

Quantum Cellular Automata, Table 1
Update table for CA rule '110' (the second row is the decimal number '110' in binary notation)

$M^{110} =$	111	110	101	100	011	010	001	000
	0	1	1	0	1	1	1	0

The most studied CA are the so-called *elementary CA* – 1-dimensional lattices with a set of two states and a neighborhood scheme of *radius* 1 (*nearest-neighbor interaction*). i. e. the state of a cell at point x at time $t+1$ only depends on the state of cells $x-1, x$, and $x+1$ at time t. There are 256 such elementary CA, easily enumerated using a scheme invented by Wolfram [48]. As an example and for later reference, the update table of *rule 110* is given in Table 1. CA with update rule '110' have been shown to be computationally universal, i. e. they can simulate any Turing machine in polynomial time [15].

A possible approach to constructing a QCA would be to simply "quantize" a CA by rendering the update rule unitary. There are two problems with this approach. One is, that applying the same unitary to each cell does not yield a well-defined global transition function nor necessarily a unitary one. The second problem is the synchronous update of all cells. "In practice", the synchronous update of, say, an elementary CA can be achieved by storing the current configuration in a temporary register, then update all cells with odd index in the original CA, update all cells with even index in the register and finally splice the updated cells together to obtain the original CA at the next time step. Quantum states, however, cannot be copied in general due to the so-called *no-cloning theorem* [49]. Thus, parallel update of a QCA in this way is not possible. Sequential update on the other hand leads to either an infinite number of time steps for each update or inconsistencies at the boundaries. One solution is a partitioning scheme as it is used in the construction of reversible CA.

Reversible Cellular Automata

Definition (Reversible CA) A CA is said to be *reversible* if for every current configuration there is exactly one previous configuration.

The global transition function F of a reversible CA is *bijective*. In general, CA are not reversible. Only 16 out of the 256 elementary CA rules are reversible. However, one can construct a reversible CA using a partitioning scheme developed by Toffoli and Margolus for 2-dimensional CA [40].

Consider a 2-dimensional CA with nearest neighborhood scheme $\mathcal{N} = \{x \in \mathbb{Z}^2 | \forall |x_i| \leq 1\}$. In the *partitioning scheme* introduced by Toffoli and Margolus each block of 2×2 cells forms a unit cube $\square$ such that the even translates $\square + 2x$ with $x \in \mathbb{Z}^2$ and the odd translates $\square + \mathbf{1} + 2x$, respectively, form a *partition* of the lattice, see Fig. 1. The update rule of a partitioned CA takes as input an entire block of cells and outputs the updated state of the entire block. The rule is then applied alternatingly to the even and to the odd translates. The Margolus partitioning scheme is easily extended to d-dimensional lattices. A *generalized Margolus scheme* was introduced by Schumacher and Werner [37]. It allows for different cell sizes in the intermediate step.

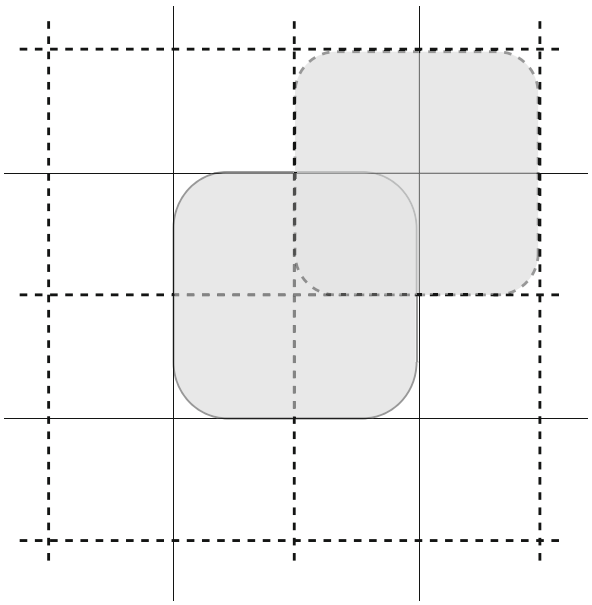

Quantum Cellular Automata, Figure 1
Even (*solid lines*) and odd (*dashed lines*) of a Margolus partitioning scheme in $d = 2$ dimensions using blocks of size 2×2. For each partition one block is shown shaded. Update rules are applied alternatingly to the *solid* and *dashed* partition

A *partitioned CA* is then a CA with a partitioning scheme such that the set of cells are partitioned in some periodic way: Every cell belongs to exactly one block, and any two blocks are connected by a lattice translation. Such a CA is neither time homogeneous nor space homogeneous anymore, but periodic in time and space. As long as the rule for evolving each block is reversible, the entire automaton will be reversible.

Early Proposals

Grössing and Zeilinger were the first to coin the term and formalize a QCA [21]. In the Schrödinger picture of quantum mechanics the state of a system at some time t is described by a state vector $|\psi_t\rangle$ in Hilbert space $\mathcal{H}$. The state vector evolves unitarily,

$$|\psi_{t+1}\rangle = U|\psi_t\rangle \,. \tag{1}$$

U is a unitary operator, i. e. $UU^\dagger = \mathbf{1}$, with the complex conjugate $U^\dagger$ and the identity matrix $\mathbf{1}$. If $\{|\phi_i\rangle\}$ is a computational basis of the Hilbert space $\mathcal{H}$ any state $|\psi\rangle \in \mathcal{H}$ can be written as a superposition $\sum_{|\phi_i\rangle} c_i |\phi_i\rangle$, with coefficients $c_i \in \mathbb{C}$ and $\sum_i c_i c_i^* = 1$. The QCA constructed by Grössing and Zeilinger is an infinite 1-dimensional lattice where at time t lattice site i is assigned the complex amplitude c_i of state $|\psi_t\rangle$. The update rule is given by unitary operator U.

Definition (Grössing–Zeilinger QCA) A *Grössing–Zeilinger QCA* is a 3-tuple $(\mathrm{L}, \mathcal{H}, U)$ which consists of (1) an infinite 1-dimensional lattice $\mathrm{L} \subseteq \mathbb{Z}$ representing basis states of (2) a Hilbert space $\mathcal{H}$ with basis set $\{|\phi_i\rangle\}$, and (3) a band-diagonal unitary operator U.

Band-diagonality of U corresponds to a locality condition. It turns out that there is no Grössing–Zeilinger QCA with nearest-neighbor interaction and nontrivial dynamics. In fact, later on, Meyer showed more generally that "*in one dimension there exists no nontrivial homogeneous, local, scalar QCA. More explicitly, every band r-diagonal unitary matrix U which commutes with the one-step translation matrix T is also a translation matrix T^k for some $k \in \mathbb{Z}$, times a phase*" [27].

Grössing and Zeilinger also introduced QCA where the unitarity constraint is relaxed to only approximate unitarity. After each update the configuration can be normalized which effectively causes non-local interactions.

The properties of Grössing–Zeilinger QCA were studied by Grössing and co-workers in some more detail in following years, see [20] and references therein. This pioneering definition of QCA, however, was not studied much further, mostly because the "non-local" behavior

renders the Grössing–Zeilinger definition non-physical. In addition, it has little in common with the concepts developed in quantum computation later on. The Grössing-Zeilinger definition really concerns what one would call today a quantum random walk (for further details see the review by Kempe [23]).

The first model of QCA researched in depth was that introduced by Watrous [46], whose ideas where further explored by van Dam [16], Dürr, LêThanh, and Santha [17,18], and Arrighi [2]. A Watrous-QCA is defined over an infinite 1-dimensional lattice, a finite set of states including a quiescent state. The transition function maps a neighborhood of cells to a single quantum state instantaneously and simultaneously.

Definition (Watrous-QCA) A *Watrous-QCA* is a 4-tuple $(\mathrm{L}, \Sigma, \mathcal{N}, f)$ which consists of (1) a 1-dimensional lattice $\mathrm{L} \subseteq \mathbb{Z}$, (2) a finite set of cell states Σ including a quiescent state ε, (3) a finite neighborhood scheme $\mathcal{N}$, and (4) a local transition function $f\colon \Sigma^{\mathcal{N}} \to \mathcal{H}_{\Sigma}$.

Here, $\mathcal{H}_{\Sigma}$ denotes the Hilbert space spanned by the cell states Σ. This model can be viewed as a direct quantization of a CA where the set of possible configurations of the CA is extended to include all linear superpositions of the classical cell configurations, and the local transition function now maps the cell configurations of a given neighborhood to a quantum state. One cell is labeled "accept" cell. The quiescent state is used to allow only a finite number of states to be active and renders the lattice effectively finite. This is crucial to avoid an infinite product of unitaries and, thus, to obtain a well-defined QCA.

The Watrous QCA, however, allows for non-physical dynamics. It is possible to define transition functions that do not represent unitary evolution of the configuration, either by producing superpositions of configurations which do not preserve the norm, or by inducing a global transition function which is not unitary. This leads to non-physical properties such as super-luminal signaling [37]. The set of Watrous QCA is not closed under composition and inverse [37].

Watrous defined a restricted class of QCA by introducing a partitioning scheme.

Definition (Partitioned Watrous QCA) A *partitioned Watrous QCA* is a Watrous QCA with $\Sigma = \Sigma_l \times \Sigma_c \times \Sigma_r$ for finite sets Σ_l, Σ_c, and Σ_r, and matrix Λ of size $\Sigma \times \Sigma$. For any state $s = (s_l, s_c, s_r) \in \Sigma$ define transition function f as

$$f(s_1, s_2, s_3, s) = \Lambda_{(s_{l_3}, s_{m_2}, s_{r_1}, s)}\,, \tag{2}$$

with matrix element Λ_{s_i, s_j}.

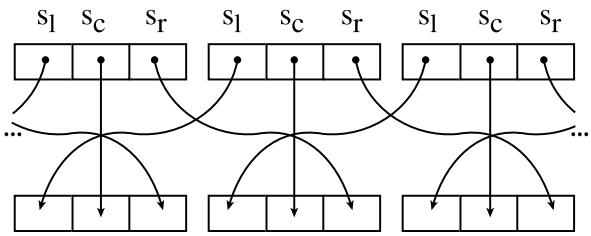

Quantum Cellular Automata, Figure 2
Each cell is divided into three sub-cells labeled l, c, and r for *left*, *center*, and a *right*, respectively. The update rule consists of swapping *left* and *right* sub-cells of neighboring cells and then updating each cell internally using a unitary operation acting on the *left*, *center*, and *right* part of each cell

In a partitioned Watrous QCA each cell is divided into three sub-cells – left, center, and right. The neighborhood scheme is then a nearest-neighbor interaction confined to each cell. The transition function consists of a unitary acting on each partitioned cell and swap operations among sub-cells of different cells. Figure 2 illustrates the swap operation between neighboring cells.

For the class of partitioned Watrous QCA Watrous provides the first proof of computational universality of a QCA by showing that any quantum Turing machine can be efficiently simulated by a partitioned Watrous-QCA with constant slowdown and that any partitioned Watrous-QCA can be simulated by a quantum Turing machine with linear slowdown.

Theorem ([46]) *Given any quantum Turing machine M_{TM}, there exists a partitioned Watrous QCA M_{CA} which simulates M_{TM} with constant slowdown.*

Theorem ([46]) *Given any partitioned Watrous QCA M_{CA}, there exists a quantum Turing machine M_{TM} which simulates M_{CA} with linear slowdown.*

Watrous' model was further developed by van Dam [16], who defined a QCA as an assignment of a product vector to every basis state in the computational basis. Here the quiescent state is eliminated and thus the QCA is made explicitly finite. Van Dam showed that the finite version is also computationally universal. Efficient algorithms to decide whether a given 1-dimensional QCA is unitary was presented by Dürr, LeTanh, and Santha [17,18]. Due to substantial shortcomings such as non-physical behavior, these early proposals were replaced by a second wave of proposals to be discussed below.

Today, there is not a generally accepted QCA model that has all the attributes of the CA model: unique definition, simple to describe, and computationally powerful. In particular, there is no axiomatic definition, contrary to its

classical counterpart, that yields an immediate way of constructing/enumerating all of the instances of this model. Rather, each set of authors defines QCA in their own particular fashion.

The states $s \in \Sigma$ are basis states spanning a finite-dimensional Hilbert space. At each point in time a cell represents a finite-dimensional quantum system in a superposition of basis states. The unitary operators represent the discrete-time evolution of strictly finite propagation speed.

Models of QCA

Reversible QCA

Schumacher and Werner used the Heisenberg picture rather than the Schrödinger picture in their model [37]. Thus, instead of associating a d-level quantum system with each cell they associated an observable algebra with each cell. Taking a *quasi-local* algebra as the tensor product of observable algebras over a finite subset of cells, a QCA is then a homomorphism of the quasi-local algebra, which commutes with lattice translations and satisfies locality on the neighborhood.

The observable-based approach was first used in Ref. [36] with focus on the irreversible case. However, this definition left questions open such as whether the composition of two QCA will again form a QCA. The following definition does avoid this uncertainty.

Consider an infinite d-dimensional lattice $\mathrm{L} \subset \mathbb{Z}^d$ of cells $x \in \mathbb{Z}^d$, where each cell is associated with the observable algebra $\mathcal{A}_x$ and each of these algebras is an isomorphic copy of the algebra of complex $d \times d$-matrices. When $\Lambda \subset \mathbb{Z}^d$ is a finite subset of cells, denote by $\mathcal{A}(\Lambda)$ the algebra of observables belonging to all cells in Λ, i. e. the tensor product $\otimes_{x\in\Lambda}\mathcal{A}_x$. The completion of this algebra is called a *quasi-local* algebra and will be denoted by $\mathcal{A}(\mathbb{Z}^d)$.

Definition (Reversible QCA) A *quantum cellular automaton* with neighborhood scheme $\mathcal{N} \subset \mathbb{Z}^d$ is a homomorphism $T\colon \mathcal{A}(\mathbb{Z}^d) \to \mathcal{A}(\mathbb{Z}^d)$ of the quasi-local algebra, which commutes with lattice translations, and satisfies the locality condition $T(\mathcal{A}(\Lambda)) \subset T(\mathcal{A}(\Lambda + \mathcal{N}))$ for every finite set $\Lambda \subset \mathbb{Z}^d$. The local transition rule of a cellular automaton is the homomorphism $T_0\colon \mathcal{A}_0 \to \mathcal{A}(\mathcal{N})$.

Schumacher and Werner presented and proved the following theorem on one-dimensional QCA.

Theorem (Structure Theorem [37]) *Let T be the global transition homomorphism of a one-dimensional a nearest-neighbor QCA on the lattice $\mathbb{Z}^d$ with single-cell algebra $\mathcal{A}_0 = \mathcal{M}_d$. Then T can be represented in the generalized Margolus partitioning scheme, i. e. T restricts to an isomorphism*

$$T\colon \mathcal{A}(\square) \to \bigotimes_{s\in\Sigma} \mathcal{B}_s\,, \tag{3}$$

where for each quadrant vector $q \in Q$, the subalgebra $\mathcal{B}_q \subset \mathcal{A}(\square + q)$ is a full matrix algebra, $\mathcal{B}_q\mathcal{M}_{n(q)}$. These algebras and the matrix dimensions $n(q)$ are uniquely determined by T.

Theorem (Structure Theorem [37]) does not hold in higher dimensions [47]. A central result obtained in this framework is that almost any [47] 1-dimensional QCA can be represented using a set of local unitary operators and a generalized Margolus partitioning [37], as illustrated in Fig. 3. Furthermore, if the local implementation allows local ancillas, then any QCA, in any lattice dimension can be built from local unitaries [37,47]. In addition, they proved the following Corollary.

Corollary ([37]) *The inverse of a nearest-neighbor QCA exists, and is a nearest-neighbor QCA.*

The latter result is not true for CA. A similar result for finite configurations was obtained in [4]. Here evidence is presented that the result does not hold for two dimensional QCA. The work by Schumacher and Werner can be considered the first general definition for 1-dimensional QCA. A similar result for many-dimensional QCA does not exist.

Local Unitary QCA

Péres-Delgado and Cheung proposed a *local unitary QCA* [34].

Definition (Local-unitary QCA) A *local-unitary QCA* is a 5-tuple $\{(\mathrm{L}, \Sigma, \mathcal{N}, U_0, V_0)\}$ consisting of (1) a d-dimensional lattice of cells indexed by integer tuples $\mathrm{L} \subset \mathbb{Z}^d$, (2) a finite set of orthogonal basis states Σ, (3) a finite

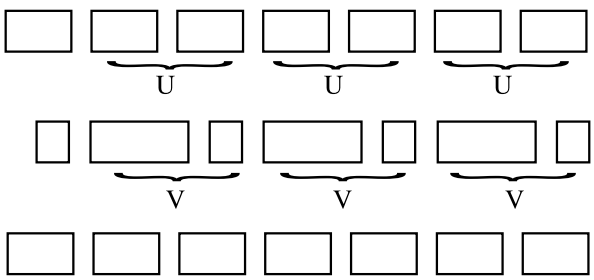

Quantum Cellular Automata, Figure 3
Generalized Margolus partitioning scheme in 1 dimension using two unitary operations *U* and *V*

neighborhood scheme $\mathcal{N} \subseteq \mathbb{Z}^d$, (4) a local read function $U_0\colon (\mathcal{H}_\Sigma)^{\otimes\mathcal{N}} \to (\mathcal{H}_\Sigma)^{\otimes\mathcal{N}}$, and (5) a local update function $V_0\colon \mathcal{H}_\Sigma \to \mathcal{H}_\Sigma$. The read operation carries the further restriction that any two lattice translations U_x and U_y must commute for all $x, y \in \mathrm{L}$.

The product VU is a valid local, unitary quantum operation. The resulting global update rule is well defined and space homogeneous. The set of states includes a *quiescent* state as well as an "ancillary" set of states/subspace which can store the result of the "read" operation. The initial state of a local-unitary QCA consists of identical k^d blocks of cells initialized in the same state. Local-unitary QCA are universal in the sense that for any arbitrary quantum circuit there is a local-unitary QCA which can simulate it. In addition any local-unitary QCA can be simulated efficiently using a family of quantum circuits [34]. Adding an additional memory register to each cell allows this class of QCA to model any reversible QCA of the Schumacher/Werner type discussed above.

Block-Partitioned and Nonunitary QCA

Brennen and Williams introduced a model of QCA which allows for unitary and nonunitary rules [14].

Definition (Block-partitioned QCA) A *Block-partitioned QCA* is a 4-tuple $\{\mathrm{L}, \Sigma, \mathcal{N}, M\}$ consisting of (1) a 1-dimensional lattice of n cells indexed $\mathrm{L} = 0, \ldots, n-1$, (2) a 2-dimensional state space Σ, (3) a neighborhood scheme $\mathcal{N}$, and (4) an update rule M applied over $\mathcal{N}$.

Given a system with nearest-neighbor interactions, the simplest unitary QCA rule has radius $r = 1$ describing a unitary operator applied over a three-cell neighborhood $j-1, j, j+1$:

$$M(u_{00}, u_{01}, u_{10}, u_{11}) = |00\rangle\langle 00| \otimes u_{00} + |01\rangle\langle 01| \otimes u_{01} + |10\rangle\langle 10| \otimes u_{10} + |11\rangle\langle 11| \otimes u_{11}\,, \quad (4)$$

where $|ab\rangle\langle ab| \otimes u_{ab}$ means update the qubit at site j with the unitary u_{ab} if the qubit at the site $j-1$ is in state $|a\rangle$ and the qubit at site $j+1$ is in state $|b\rangle$. M commutes with its own 2-site translation. Thus, a partitioning is introduced by updating simultaneously all even qubits with rule M before updating all odd qubits with rule M. Periodic boundaries are assumed. However, by addressability of the end qubits simulation of a block-partitioned QCA by a QCA with boundaries can be achieved.

Nonunitary update rules correspond to completely positive maps on the quantum states where the neighboring states act as the environment. Take a nearest-neighbor 1-dimensional Block-partitioned QCA. In the density operator formalism each quantum system ρ is given by the probability distribution $\rho = \sum_i p_i |\psi\rangle\langle\psi|$ over outer products of quantum states $|\psi\rangle$. A completely positive map $S(\rho)$ applied to state ρ is represented by a set of Krauss operators F_μ, which are positive operators that sum up to the identity $\sum_\mu F_\mu^\dagger F_\mu = \mathbf{1}$. The map $S_j^{ab}(\rho)$ acting on cell j conditioned on state a of the left neighbor and state b of the right neighbor can then be written as

$$S_j^{ab}(\rho) = |ab\rangle\langle ab| \otimes \sum_\mu F_\mu^{ab} \rho F_\mu^{ab\dagger} \otimes |ab\rangle\langle ab|\,. \quad (5)$$

As an example, the CA rule '110' can now be translated into an update rule for cell j in a block-partitioned nonunitary QCA:

$$F_1^j = |00\rangle\langle 00| \otimes \mathbf{1}^j + |10\rangle\langle 10| \otimes \mathbf{1}^j + |11\rangle\langle 11| \otimes \sigma_x^j + |01\rangle\langle 01| \otimes |1\rangle_{jj}\langle 1| \quad (6)$$

$$F_2^j = |01\rangle\langle 01| \otimes |1\rangle_{jj}\langle 0|\,, \quad (7)$$

where σ_x is the Pauli operator.

The implementation of such a block-partitioned nonunitary QCA is proposed in form of a lattice of even order constructed with an alternating array of two distinguishable species $ABABABAB\ldots$ that are globally addressable and interact via the Ising interaction. Update rules that generate and distribute entanglement were studied in this framework [14].

Continuous-Time QCA

Vollbrecht and Cirac initiated the study of continuous-time QCA [44]. They show that the computability of the ground state energy of a translationally invariant n-neighbor Hamiltonian was QMA-hard. Their QCA model is taken up by Nagaj and Wocjam [31] who used the term *Hamiltonian QCA*.

Definition (Hamiltonian QCA) A *Hamiltonian QCA* is a tuple $\{L, \Sigma = \Sigma_p \times \Sigma_d\}$ consisting of (1) a 1-dimensional lattice of length L, (2) a finite set of orthogonal basis states $\Sigma = \Sigma_p \times \Sigma_d$ containing (2a) a data register Σ_d, and (2b) a program register Σ_p.

The initial state encodes both the program and the data, stored in separate subspaces of the state space:

$$|\phi\rangle = \bigotimes_{j=1}^{L} \left(|p_j\rangle \otimes |d_j\rangle\right)_j \quad (8)$$

The computation is carried out autonomously. Nagaj and Wocjam showed that, if the system is left alone for

a period of time $t = O(L \log L)$, polynomially in the length of the chain, the result of the computation is obtained with probability $p \geq 5/6 - O(1/\log L)$. Hamiltonian QCA are computationally universal, more precisely they are in the complexity class BQP. Two constructions for Hamiltonian QCA are given in [31], one using a 10-dimensional state space, and the resulting system can be thought of as the diffusion of a system of free fermions. The second construction given uses a 20-dimensional state space and can be thought of as a quantum walk on a line.

Examples of QCA

Raussendorf proved an explicit construction of QCA and proves its computational universality [35]. The QCA lives on a torus with a 2×2 Margolus partitioning. The update rule is given by a single 4-qubit unitary acting on 2×2 blocks of qubits. The four-qubit unitary operation consists of swap operations, the Hadamard transformation, and a phase gate. The initial state of the QCA is prepared such that columns encode alternatingly data and program. When the QCA is running the data travel in one direction while the program (encoding classical information in orthogonal states) travels in the opposite direction. Where the two cross the computation is carried out through nearest-neighbor interaction. After a fixed number of steps the computation is done and the result can be read out of a dedicated "data" column. This QCA is computationally universal, more precisely, it is within a constant as efficient as a quantum logic network with local and nearest-neighbor gates.

Shepherd, Franz, and Werner compared *classically controlled* QCA to autonomous QCA [38]. The former is controlled by a classical compiler that selects a sequence of operations acting on the QCA at each time step. The latter operates autonomously, performing the same unitary operation at each time step. The only step that is classically controlled is the measurement (and initialization). They show the computational equivalence of the two models. Their result implies that a particular quantum simulator may be as powerful as a general one.

Computationally Universal QCA

Quite a few models have been shown to be computationally universal, i. e. they can simulate any quantum Turing machine and any quantum circuit efficiently. A Watrous-QCA simulates any quantum Turing machine with constant slowdown [46]. The QCA defined by Van Dam is a finite version of a Watrous QCA and is computationally universal as well [16]. Local-unitary QCA can simulate any quantum circuit and thus are computationally universal [34]. Block-partitioned QCA can simulate a quantum computer with linear overhead in time and space [14]. Continuous-time QCA are in complexity class BQP and thus computationally universal [44]. The explicit constructions of 2-dimensional QCA by Raussendorf is computationally universal, more precisely, it is within a constant as efficient as a quantum logic network with local and nearest-neighbor gates [35]. Shepherd, Franz, and Werner provided an explicit construction of a 12-state 1-dimensional QCA which is in complexity class BQP. It is universally programmable in the sense that it simulates any quantum-gate circuit with polynomial overhead [38]. Arrighi and Fargetton proposed a 1-dimensional QCA capable of simulating any other 1-dimensional QCA with linear overhead [3].

Implementations of computationally universal QCA have been suggested by Lloyd [24] and Benjamin [8].

Modeling Physical Systems

One of the goals in developing QCA is to create a useful modeling tool for physical systems. Physical systems that can be simulated with QCA include Ising and Heisenberg interaction spin chains, solid state NMR, and quantum lattice gases. Spin chains are perhaps the most obvious systems to model with QCA. The simple cases of such 1-dimensional lattices of spins are Hamiltonians which commute with their own lattice translations. Vollbrecht and Cirac showed that the computability of the ground state energy of a translationally invariant n-neighbor Hamiltonian is in complexity class QMA [44]. For simulating non-commuting Hamiltonians a block-wise update such as the Margolus partitioning has to be used (see Sect. "Reversible Cellular Automata"). Here the fact is used that any Hamiltonian can be expressed as the sum of two Hamiltonians, $H = H_a + H_b$. H_a and H_b can then, to a good approximation, be applied sequentially to yield the original Hamiltonian H, even if these do not commute. It has been shown that such 1-dimensional spin chains can be simulated efficiently on a classical computer [43]. It is not known, however, whether higher dimensional spin systems can be simulated efficiently classically.

Quantum Lattice Gas Automata

Any numerical evolution of a discretized partial differential equation can be interpreted as the evolution of some CA, using the framework of *lattice gas automata*. In the continuous time and space limit such a CA mimics the behavior of the partial differential equation. In quantum mechanical lattice gas automata (QLGA) the continuous limit on a set of so called quantum lattice Boltzman equation

recovers the Schrödinger equation [39]. The first formulation of a linear unitary CA was given in Ref. [10]. Meyer coined the term *quantum lattice gas automata* (QLGA) and demonstrated the equivalence of a QLGA and the evolution of a set of quantum lattice Boltzman equations [27,28]. Meyer [29], Boghosian and Taylor [12], and Love and Boghosian [25] explored the idea of using QLGA as a model for simulating physical systems. Algorithms for implementing QLGA on a quantum computer have been presented in [13,30,32].

Implementations

A large effort is being made in many laboratories around the world to implement a model of a quantum computer. So far all of them are confined to a very finite number of elements and are no way near to a quantum Turing machine (which in itself is a purely theoretical construct but can be approximated by a very large number of computational elements). One existing experimental set-up that is very promising for quantum information processing and that does not suffer from this "finiteness" are optical lattices (for a review, see [11]). They possess a translation symmetry which makes QCA a very suitable framework in which to study their computational power. Optical lattices are artificial crystals of light and consist of hundreds of thousands of microtraps. One or more neutral atoms can be trapped in each of the potential minima. If the potential minima are deep enough any tunneling between the traps is suppressed and each site contains the same amount of atoms. A quantum register – here in form of a so-called Mott insulator – has been created. The biggest challenge at the moment is to find a way to address the registers individually to implement quantum gates. For a QCA all that is needed is implementing the unitary operation(s) acting on the entire lattice simultaneously. The internal structure of the QCA guarantees the locality of the operations. This is a huge simplification compared to individual manipulation of the registers. Optical lattices are created routinely by superimposing two or three orthogonal standing waves generated from laser beams of a certain frequency. They are used to study Fermionic and Bosonic quantum gases, nonlinear quantum dynamics, strongly correlated quantum phases, to name a few.

A type of locally addressed architecture by global control was put forward by Lloyd [24]. In this scheme a 1-dimensional array is built out of three atomic species, periodically arranged as $\mathcal{ABCABCABC}$. Each species encodes a qubit and can be manipulated without affecting the other species. The operations on any species can be controlled by the states of the neighboring cells. The end-cells are used for readout, since they are the only individually addressable components. Lloyd showed that such a quantum architecture is universal. Benjamin investigated the minimum physical requirements for such a many-species implementation and found a similar architecture using only two types of species, again arranged periodically $\mathcal{ABABAB}$ [7,8,9]. By giving explicit sequences of operations implementing one-qubit and two-qubit (CNOT) operations Benjamin showed computational universality. But the reduction in spin resources comes with an increase in logical encoding into four spin sites with a buffer space of at least four empty spin sites between each logical qubit.

A continuation of this multi-species QCA architecture is found in the work by Twamley [42]. Twamley constructed a proposal for a QCA architecture based on Fullerene (C_{60}) molecules doped with atomic species ^{15}N and ^{31}P, respectively, arranged alternatingly in a one-dimensional array. Instead of electron spins which would be too sensitive to stray electric charges the quantum information is encoded in the nuclear spins. Twamley constructed sequences of pulses implementing Benjamin's scheme for one- and two-qubit operations. The weakest point of the proposal is the readout operation which is not well-defined.

A different scheme for implementing QCA was suggested by Tóth and Lent [41]. Their scheme is based on the technique of quantum-dot CA. The term quantum-dot CA is usually used for CA implementations in quantum dots (for classical computation). The authors, therefore, called their model a *coherent* quantum-dot CA. They illustrated the usage of an array of N quantum dots as an N-qubit quantum register. However, the set-up and the allowed operations allow for individual control of each cell. This coherent quantum-dot CA is more a hybrid of a quantum circuit with individual qubit control and a QCA with constant nearest-neighbor interaction. The main property of a QCA, operating under global control only, is not taken advantage of.

Future Directions

The field of QCA is developing rapidly. New definitions have appeared very recently. Since QCA are now considered to be one of the standard measurement-based models of quantum computation, further work on a consistent and sufficient definition of higher-dimensional QCA is to be expected. One proposal for such a "final" definition has been put forward in [4,5].

In the search for robust and easily implementable quantum computational architectures QCA are of considerable interest. The main strength of QCA is global con-

trol without the need to address cells individually (with the possible exception of the read-out operation). It has become clear that the global update of a QCA would be a way around practical issues related to the implementation of quantum registers and the difficulty of their individual manipulation.

More concretely, QCA provide a natural framework for describing quantum dynamical evolution of optical lattices, a field in which the experimental physics community has made huge progress in the last decade.

The main focus so far has been on reversible QCA. Irreversible QCA are closely related to measurement-based computation and remain to be explored further.

Bibliography

Primary Literature

1. Aoun B, Tarifi M (2004) Introduction to quantum cellular automata. http://arxiv.org/abs/quant-ph/0401123
2. Arrighi P (2006) Algebraic characterizations of unitary linear quantum cellular automata. In: Mathematical Foundations of Computer Science 2006. Lecture Notes in Computer Science, vol 4162. Springer, Berlin, pp 122–133
3. Arrighi P, Fargetton R (2007) Intrinsically universal one-dimensional quantum cellular automata. 0704.3961. http://arxiv.org/abs/0704.3961
4. Arrighi P, Nesme V, Werner R (2007) One-dimensional quantum cellular automata over finite, unbounded configurations. 0711.3517v1. http://arxiv.org/abs/0711.3517
5. Arrighi P, Nesme V, Werner R (2007) N-dimensional quantum cellular automata. 0711.3975v1. http://arxiv.org/abs/arXiv:0711.3975
6. Benioff P (1980) The computer as a physical system: A microscopic quantum mechanical hamiltonian model of computers as represented by turing machines. J Stat Phys 22:563–591
7. Benjamin SC (2000) Schemes for parallel quantum computation without local control of qubits. Phys Rev A 61:020301–4
8. Benjamin SC (2001) Quantum computing without local control of qubit-qubit interactions. Phys Rev Lett 88(1):017904
9. Benjamin SC, Bose S (2004) Quantum computing in arrays coupled by "always-on" interactions. Phys Rev A 70:032314
10. Bialynicki-Birula I (1994) Weyl, Dirac, and Maxwell equations on a lattice as unitary cellular automata. Phys Rev D 49:6920
11. Bloch I (2005) Ultracold quantum gases in optical lattices. Nature Phys 1:23–30
12. Boghosian BM, Taylor W (1998) Quantum lattice-gas model for the many-particle Schrödinger equation in d dimensions. Phys Rev E 57:54
13. Boghosian BM, Taylor W (1998) Simulating quantum mechanics on a quantum computer. Physica D: Nonlinear Phenomena 120:30–42
14. Brennen GK, Williams JE (2003) Entanglement dynamics in one-dimensional quantum cellular automata. Phys Rev A 68:042311
15. Cook M (2004) Universality in elementary cellular automata. Complex Syst 15:1
16. van Dam W (1996) Quantum cellular automata. Master's thesis, University of Nijmegen
17. Dürr C, Santha M (2002) A decision procedure for unitary linear quantum cellular automata. SIAM J Comput 31:1076–1089
18. Dürr C, LêThanh H, Santha M (1997) A decision procedure for well-formed linear quantum cellular automata. Random Struct Algorithms 11:381–394
19. Feynman R (1982) Simulating physics with computers. Int J Theor Phys 21:467–488
20. Fussy S, Grössing G, Schwabl H, Scrinzi A (1993) Nonlocal computation in quantum cellular automata. Phys Rev A 48:3470
21. Grössing G, Zeilinger A (1988) Quantum cellular automata. Complex Syst 2:197–208
22. Gruska J (1999) Quantum Computing. Osborne/McGraw-Hill. QCA are treated in Section 4.3
23. Kempe J (2003) Quantum random walks: an introductory overview. Contemp Phys 44:307
24. Lloyd S (1993) A potentially realizable quantum computer. Science 261:1569–1571
25. Love P, Boghosian B (2005) From Dirac to diffusion: Decoherence in quantum lattice gases. Quantum Inf Process 4:335–354
26. Margolus N (1991) Parallel quantum computation. In: Zurek WH (ed) Complexity, Entropy, and the Physics of Information, Santa Fe Institute Series. Addison Wesley, Redwood City, pp 273–288
27. Meyer DA (1996) From quantum cellular automata to quantum lattice gases. J Stat Phys 85:551–574
28. Meyer DA (1996) On the absence of homogeneous scalar unitary cellular automata. Phys Lett A 223:337–340
29. Meyer DA (1997) Quantum mechanics of lattice gas automata: One-particle plane waves and potentials. Phys Rev E 55:5261
30. Meyer DA (2002) Quantum computing classical physics. Philos Trans Royal Soc A 360:395–405
31. Nagaj D, Wocjan P (2008) Hamiltonian quantum cellular automata in 1d. 0802.0886. http://arxiv.org/abs/0802.0886
32. Ortiz G, Gubernatis JE, Knill E, Laflamme R (2001) Quantum algorithms for fermionic simulations. Phys Rev A 64:022319
33. Perez-Delgado CA, Cheung D (2005) Models of quantum cellular automata. http://arxiv.org/abs/quant-ph/0508164
34. Perez-Delgado CA, Cheung D (2007) Local unitary quantum cellular automata. Phys Rev A (Atomic, Molecular, and Optical Physics) 76:032320–15
35. Raussendorf R (2005) Quantum cellular automaton for universal quantum computation. Phys Rev A (Atomic, Molecular, and Optical Physics) 72:022301–4
36. Richter W (1996) Ergodicity of quantum cellular automata. J Stat Phys 82:963–998
37. Schumacher B, Werner RF (2004) Reversible quantum cellular automata. quant-ph/0405174. http://arxiv.org/abs/quant-ph/0405174
38. Shepherd DJ, Franz T, Werner RF (2006) Universally programmable quantum cellular automaton. Phys Rev Lett 97:020502–4
39. Succi S, Benzi R (1993) Lattice Boltzmann equation for quantum mechanics. Physica D: Nonlinear Phenomena 69:327–332
40. Toffoli T, Margolus NH (1990) Invertible cellular automata: A review. Physica D: Nonlinear Phenomena 45:229–253
41. Tóth G, Lent CS (2001) Quantum computing with quantum-dot cellular automata. Phys Rev A 63:052315
42. Twamley J (2003) Quantum-cellular-automata quantum computing with endohedral fullerenes. Phys Rev A 67:052318

43. Vidal G (2004) Efficient simulation of one-dimensional quantum many-body systems. Phys Rev Lett 93(4):040502
44. Vollbrecht KGH, Cirac JI (2008) Quantum simulators, continuous-time automata, and translationally invariant systems. Phys Rev Lett 100:010501
45. von Neumann J (1966) Theory of Self-Reproducing Automata. University of Illinois Press, Champaign
46. Watrous J (1995) On one-dimensional quantum cellular automata. In: Proceedings of the 36th Annual Symposium on Foundations of Computer Science, Milwaukee, pp 528–537
47. Werner R Private communication
48. Wolfram S (1983) Statistical mechanics of cellular automata. Rev Mod Phys 55:601
49. Wootters WK, Zurek WH (1982) A single quantum cannot be cloned. Nature 299:802–803

Books and Reviews

Summaries of the topic of QCA can be found in chapter 4.3 of Gruska [21], and in Refs. [1,32].

Quantum Chaos

Giulio Casati[1], Tomaž Prosen[2]
[1] Universita dell'Insubria, Como, Italy
[2] Univerza v Ljubljani, Ljubljana, Slovenia

Article Outline

Glossary
Definition of the Subject
Introduction
Quantum Chaos – Stationary Aspects
Quantum Chaos – Dynamical Aspects
Applications of Quantum Chaos
Future Directions
Bibliography

Glossary

Ergodicity The property of a dynamical system, according to which a single trajectory, starting from almost any initial condition, explores (densely covers) the entire available phase space of physical states.

Integrability A classical Hamiltonian dynamical system of N degrees of freedom is said to be integrable (according to Liouville) if there exist N independent conserved quantities. Integrability implies explicit quasi-periodic solution of the equations of motion.

Random matrix theory The statistical theory which allows to describe the fluctuation properties of quantum systems in terms of the sets (ensembles) of random Hermitian matrices with appropriate invariant measures.

Wigner surmise Nearest neighbor energy level spacing distribution based on the simplest 2×2 Gaussian Hermitian random matrix models, accurately approximating spacing distributions in complex quantum systems.

Periodic orbit theory or trace formula A relationship between certain properties of energy spectrum of a quantized chaotic system, and the set of unstable periodic orbits of the corresponding classical chaotic system.

Quantum Loschmidt echo or fidelity A measure of stability of quantum time evolution. It is computed as a Hilbert space inner product of two slightly different quantum time evolutions starting from the same initial state.

Definition of the Subject

As it is now widely recognized, classical dynamical chaos has been one of the major scientific breakthroughs of the past century. Quantum chaos, sometimes called Quantum chaology, studies the manifestations of chaotic motion and related dynamical phenomena in quantum mechanics [1,2].

More abstractly, one may define as quantum chaos those phenomena of *simple* quantum systems which can be described statistically and exhibit some universal (i. e. system independent) features. By the term *simple* we mean here that the system can be specified by a finite set of parameters or, generally, can be described by a finite amount of information. So we can fundamentally distinguish the phenomena of quantum chaos from similar dynamical phenomena in *disordered* systems – specified in terms of random parameters and which therefore contain infinite amount of information in an appropriate (say thermodynamic) limit.

The universal statistical properties of quantum chaotic systems which have been widely studied include statistics of energy level spectra, statistical and semiclassical structures of the wave-functions and statistical distributions of transition matrix elements (matrix elements of certain physical observables in the energy eigenbasis). These properties are of key importance for understanding quantum state transitions, dissipation, ionization and related phenomena. Traditionally quantum chaos has been intimately connected to problems in atomic physics, nuclear physics, mesoscopic solid state physics, and more recently also to the emerging field of quantum information.

The subject of quantum chaos is at the core of the fundamental and general understanding of the correspondence principle according to which classical mechanics

emerges as a limit of quantum mechanics when an 'effective' value of the Planck's constant goes to zero. However, changing perspective, and treating quantum mechanics as a fundamental theory and classical mechanics as its convenient approximation, one can ask what happens to the phenomena of quantum chaos in systems which lack the classical limit, such as systems of spin 1/2 or quantum bits (qubits). Still, many of the successful statistical tools developed or widely used in the field of quantum chaos can be applied to understand the dynamical and statistical properties of interacting qubits.

Introduction

Classical Hamiltonian dynamical systems display a rich variety of behaviors [3]. At one end, we have *strongly chaotic* dynamical systems, which are distinguished by such properties as *algorithmic complexity*, *exponential sensitivity to initial conditions*, *continuous spectrum* and consequent decay of temporal correlations ('loss of memory'), relaxation to equilibrium, and *ergodicity*. Without entering a discussion of the above properties, we notice that rigorous examples of such strongly chaotic systems are billiard balls bouncing inside a table with inward curved boundaries, or point mass particles moving freely on any surfaces of constant *negative curvature*.

At the opposite end, there are *completely integrable* or *regular* dynamical systems, which, according to Liouville, are characterized by the existence of as many independent smooth constants of motion as there are degrees of freedom, and which are distinguished by analytic predictability of time evolutions and absence of algorithmic complexity. Consequently, integrable systems have a discrete spectrum of time evolution, do not display relaxation to equilibrium and, due to existence of non-trivial constants of motion, they are not ergodic.

Nowadays we know few examples of completely integrable systems, such as an arbitrary system of coupled linear (harmonic) oscillators, the general problem of two moving bodies interacting with a centrally symmetric force, and even many-body models such as the famous Toda lattice which is a system of equal point masses in one dimension interacting with exponentially distance dependent force. On the other hand we also know that a generic, or typical Hamiltonian system is not integrable, neither it is strongly chaotic in a rigorous sense. Instead we find a variety of intermediate behaviors between completely integrable and strongly chaotic.

A famous Kolmogorov–Arnold–Moser (KAM) theorem states that a small generic (smooth) perturbation of a completely integrable Hamiltonian systems preserves most of the features of regular motion, such as quasi-periodicity or discrete spectrum, for most of initial conditions. However, in the vicinity of the so-called resonant tori (i. e. for initial conditions for which the motion of the unperturbed system has commensurate frequencies) the motion becomes locally (weakly) chaotic even for arbitrary weak perturbations. Still, the relative overall phase space volume occupied by chaotic trajectories decreases to zero faster than any power of the perturbation strength.

However, KAM theory does not describe the only scenario of integrability breaking in Hamiltonian systems. Other types of perturbations, which do not obey the conditions of KAM theorem are possible which yield physically interesting behavior. One class of such behaviors is the motion in generic polygonal billiards [4] (namely billiard tables with the shape of a generic polygon, say a triangle), or the motion of any number of elastically colliding point masses in one dimension. Such systems are neither integrable nor chaotic in the sense of exponential sensitivity or algorithmic complexity.

The fundamental problem of quantum chaos concerns the manifestations in the quantum world of the various degrees of complexity of classical dynamics, as described by the above hierarchy. Primarily we are interested in *simple*, closed (isolated) and bounded systems for which, as it is known, the spectrum of the Hamilton operator (the generator of the Schrödinger equation) is always discrete. Notice that in classical ergodic theory, discrete spectrum is associated to integrability which is just the opposite of chaos which requires continuous spectrum. In such a situation quantum mechanics, at most, can follow chaotic classical dynamics only up to the so called *breaking* time scale t^* after which the quantum motion becomes fundamentally different from the classical one. Precise understanding of this transition phenomena, scaling of the time scale t^* with various physical properties, is at the heart of quantum chaos and shall be briefly discussed in the subsequent sections.

Quantum Chaos – Stationary Aspects

The problems of quantum chaos can be viewed from two different but closely connected view-points, namely the stationary aspects in the energy domain and the dynamical aspects in the time domain. Let us start by discussing the stationary aspects.

It has been recognized as early as in 1950's by Wigner, and later by Dyson, that many statistical features of energy levels of complex nuclei, or long lived resonance states, can be adequately described by a simple statistical model of random Hermitian matrices. Wigner's idea was that for

a sufficiently complicated quantum system with many degrees of freedom like a heavy nucleus, one assumes that the matrix elements of the Hamiltonian in a *typical* basis can be treated as independent Gaussian random numbers. Such random matrix model has essentially no free parameters and is invariant under almost arbitrary basis changes, and therefore one can obtain many explicit analytical results.

The simplest and perhaps the main result of random matrix theory [5] predicts that the statistical distribution of spacings between adjacent energy levels, denoted by $S_n = E_{n+1} - E_n$, properly scaled or normalized such that the average spacing S equals one, obeys universal distributions which only depend on certain symmetry properties. Roughly speaking, they depend on the existence of time-reversal invariance of the Hamiltonian, and are, to a high level of accuracy, given by a the so-called Wigner surmise

$$P_\beta(S) = AS^\beta \exp(-BS^2)$$

where the constants A and B are determined from normalization conditions, $\int_0^\infty P(S)\,\mathrm{d}S = 1$, $\int_0^\infty SP(S)\,\mathrm{d}S = 1$. The integer β is known as universality index and equals $\beta = 1$ for real Gaussian random matrices applying in the time-reversal invariant case, and $\beta = 2$ for complex Gaussian random matrices applying in the case where time-reversal invariance is broken. Note that β has also an interesting interpretation as a *level repulsion parameter* since $\beta > 0$ implies repulsive correlations between adjacent energy levels.

Of course, random matrix theory would only remain an interesting mathematical model if it would not have been so immensely successful in describing spectral correlations of complex quantum systems. As already mentioned it started with nuclear spectra for which one is ready to believe that due to immense complexity of interactions random matrix description is intuitively adequate. However in beginning of 1980's numerical evidence started to accumulate [6,7] that even spectra of very simple, but yet non-integrable and classically chaotic systems exhibit universal level fluctuations described by random matrix theory.

This observation, namely *that short range spectral correlations of quantum systems which are strongly chaotic in the classical limit obey universal fluctuation laws which are given by ensembles of random matrices without free parameters*, has been known as *quantum chaos conjecture* and, despite of still not being rigorously proven, remains one of the defining and most important results in the field.

One of the standard paradigmatic models where quantum chaos conjecture has been most often and convincingly demonstrated are quantum billiards. The stationary Schrödinger problem for a billiard of a point particle of mass m moving freely inside a planar domain (billiard table) $\mathcal{D}$ and colliding elastically off its boundary is simply the well known Helmholtz (amplitude) equation for the particle wavefunction

$$\nabla^2\Psi(x, y) + k^2\Psi(x, y) = 0\,,$$

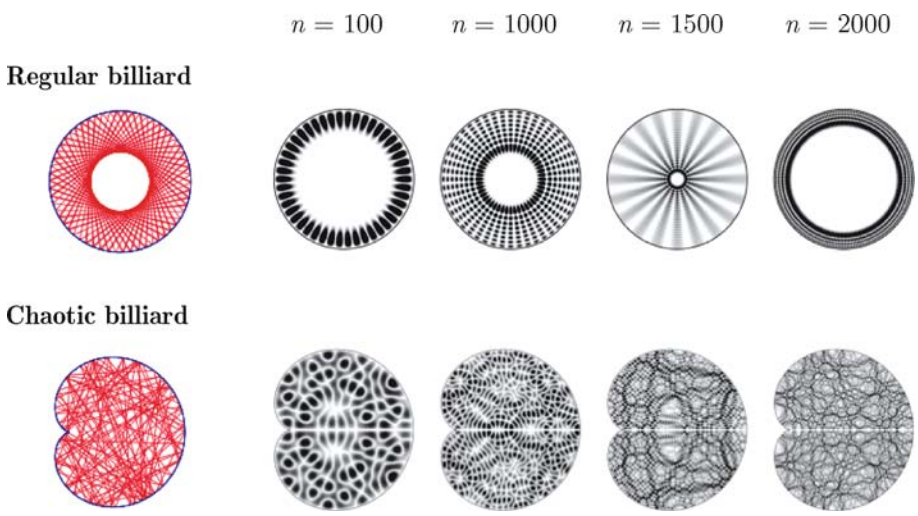

Quantum Chaos, Figure 1

Typical examples of a regular billiard (circular billiard, *above*) and a fully chaotic billiard (cardioid billiard, *below*). In the *left* side of the figure we show two typical trajectories in the billiard table, whereas to the *right* we depict few examples of wave-functions of eigenstates at different sequential level numbers *n*. Courtesy of Bäcker [9]

with Dirichlet boundary condition $\Psi|_{\partial\mathcal{D}} = 0$, having a discrete set of solutions $\{k_n, \Psi_n; n = 1, 2, \dots\}$ with a dispersion relation to the eigenenergies $E_n = \hbar^2 k_n^2/(2m)$. In Fig. 1 we show two examples of a typical integrable billiard (circular billiard), and a typical fully chaotic and ergodic billiard (cardioid billiard proposed by Robnik [8]) – plotting typical trajectories for each and a sequence of few typical chaotic and regular eigenstates.

Solving the Dirichlet Helmholtz problem for a finite domain is one of the most standard problems in linear wave physics and emerges in equivalent forms in gas acoustics, flat electromagnetic (microwave) resonators, transverse waves in optical fibers, etc, the only difference from the stationary quantum problem being the dispersion relation which connects the frequency of stationary waves to the eigenvalues of the wavenumber k_n.

In Fig. 2 one can observe that spectra of chaotic quantum billiard, hydrogen atom in strong magnetic field, excitation spectrum of NO_2 molecule, microwave electromagnetic spectrum of a three dimensional chaotic cavity and even spectra of vibrating elastic solids all exhibit spectral correlations which are given by a Gaussian ensemble of real random matrices.

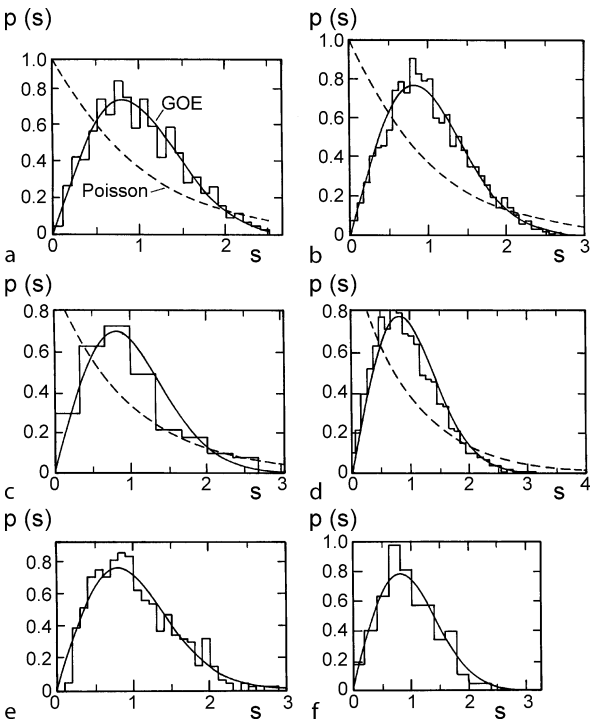

Quantum Chaos, Figure 2

Level spacing distributions for a the fully chaotic Sinai billiard [7], b a Hydrogen atom in a strong magnetic field, c an NO_2 molecule, d a vibrating quartz block shaped like a three dimensional sinai billiard, e the microwave spectrum of a three-dimensional chaotic cavity, f a vibrating elastic disc shaped like a quarter stadium. Courtesy of Stöckmann [2]

The universality of quantum statistics of classically chaotic systems applies also to other properties, such as the distribution of a chaotic wave-function amplitudes, which is Gaussian and very well modeled by the so-called random plane wave superposition [10], the distribution of matrix elements of typical observables in the eigenbasis of chaotic Hamiltonians [11], which is again Gaussian, or even distribution of chaotic Wigner functions (quantum analogues of phase space densities) which is found again to be Gaussian [12].

An interesting related mathematical question is the problem of *quantum ergodicity*. For a generic physical observable A the question is whether its expectation value in a given eigenstate approaches, with increasing quantum level number, the microcanonical average of the corresponding classical observable evaluated at the corresponding energy, namely whether the sequence

$$\{\langle\Psi_n|A|\Psi_n\rangle - A_{\text{classical}}(E_n); n = 1, 2, 3, \dots\}$$

converges to zero? Mathematicians often relax the above condition to hold for a *subsequence of density 1* meaning that the above property holds for a typical eigenstate, whereas the sequence of exceptions (with semiclassically vanishing statistical weight) correspond to the famous *scar states* discovered by Heller [13]. For quantum billiards, the statement of quantum ergodicity is equivalent to the statement of uniform equidistribution of probability density for eigenstates (see e. g. lower panels of Fig. 1). It has been proven (the proof has been announced by Shnirelman [14], and later worked out by Zelditch [15] and Colin de Verrdiere [16]) that strongly chaotic billiards, and some other rigorous examples of strongly chaotic systems, are quantum ergodic. However, the minimal classical ergodic properties (like ergodicity, weak mixing, etc) sufficient for quantum ergodicity are still under debate.

One can now ask a similar questions for the other extreme of ergodic hierarchy, namely for classically regular systems: Are there some universal features of spectral fluctuations of quantum systems whose classical limit is completely regular? The answer is only partly affirmative, in the sense of an argument which has been originally given by Berry and Tabor [17]. The starting point of this argument is that the eigenenergies of a completely integrable system with d degrees of freedom can be labeled by a d-duple of quantum numbers – integers n_j – namely $E_{n_1 n_2 \dots n_d}$. Berry and Tabor argued that level sub-sequences, where $d-1$ quantum numbers are fixed and only one of them

is being varied, are *mutually independent*. As the overall spectrum of an integrable system is a superposition of many such almost uncorrelated level sequences, one concludes that any short range level correlations should be absent. For example, the level spacing distribution for an integrable system should be the same as for a Poissonian distribution of uncorrelated events, namely

$$P_{\text{int}}(S) = \exp(-S) \,.$$

One has to note that for integrable systems one should not really expect a universality in the same sense as for chaotic systems. Berry and Tabor's argument cannot in general be turned into a rigorous proof, and for any particular integrable system one can in principle always find statistical measures which deviate from Poissonian predictions [18]. For example, a one dimensional harmonic oscillator has an equidistant (the so-called "picket fence") spectrum so its level spacing distribution can be written in terms of a Dirac's delta distribution $P_{\text{harm.osc.}} = \delta(S-1)$. In other words, as quantum statistical properties are concerned, there is no such thing as *a typical integrable system*. This can be viewed as another manifestation of quantum non-ergodicity of integrable systems.

Even if spectral fluctuations of integrable systems are relatively universal – in the sense as explained above – this is not at all true for other statistical properties of quantized integrable systems, such as wave-function *amplitudes*, sizes and numbers of *nodal domains* of wave functions, *matrix elements* of typical physical observables, etc. There we find system specific features which can hardly be described by some general statistical rules.

The general case of typical quantum systems, say those which can be understood as quantizations of systems with classically mixed phase space with coexisting regular and chaotic orbits, is the most difficult to describe. It is fair to say that quantum chaos of generic systems is still in its infancy. Among the few results which can be mentioned is the semiclassical theory of Berry and Robnik [19] describing level fluctuations of mixed systems as statistical superposition of independent level subsequences corresponding to each invariant classical phase space component: level subsequences corresponding to areas of chaotic motion are modeled by an appropriate ensemble of random matrices with the relative level density which is given by the classical volume of the chaotic component, and a (single) subsequence corresponding to all regular trajectories is modeled by a Poissonian level sequence of the corresponding overall level density. However, in realistic quantum systems *tunneling* between chaotic and regular states has to be taken into account and, in spite of few attempts, we are still lacking a general statistical theory of tunneling amplitudes in mixed phase space systems.

It should be noted that statistical properties of energy level fluctuations can be related to classical recurrent phase space structures such as periodic orbits. In particular, for chaotic dynamical systems, one finds a beautiful relationship between semiclassical approximation to the energy spectrum and isolated unstable classical periodic orbits, which is known as Gutzwiller's *trace formula* [20], and is based on stationary phase approximation to Feynman path integral representation of Green's function of the Schrödinger equation. A similar trace formula has been proposed by Berry and Tabor also for classically regular systems [21]. Related periodic or closed orbit theories have been later developed for describing semiclassical properties of other quantities, for example scars in chaotic wavefunctions [22,23]. The trace formula has been believed to be the theoretical tool to attack the proof of quantum chaos conjecture. Considerable recent progress has been achieved by Müller et al. [24], building upon an idea of Sieber [25], who have shown that the correct resummation of a trace formula indeed yields short time expansion of the universal *spectral form factor* (Fourier transformation of the two-point spectral correlation function) which is identical to the one obtained from random matrix theory.

Quantum Chaos – Dynamical Aspects

Classical chaos is a property of time evolution and, as we have seen, requires continuous spectrum of the motion. However, the spectrum of bounded closed quantum systems is always discrete, therefore genuine quantum chaos which would be a property of asymptotic time evolution is not possible, namely there is an ultimate breaking time scale t^* which is determined by the density of states $\rho = \mathrm{d}E_n/\mathrm{d}n$. Indeed, writing a completely general solution of the time dependent Schrödinger equation as

$$\Psi(t) = \sum_n c_n \exp(-\mathrm{i}E_n t/\hbar)\Psi_n$$

one sees that the quasi-periodic nature of time evolution shows up, on average, when the difference of two adjacent phases grows to 2π, namely

$$t^* = 2\pi\hbar\rho \,.$$

This breaking time scale is also sometimes referred to as Heisenberg time, and after t^* the time evolution in quantum mechanics is dominated by quantum fluctuations.

Thus genuine quantum chaos is possible only within a time scale $t < t^*$, where quantum phenomena with semiclassical description are possible, such as relaxation,

exponential sensitivity etc. However, not all chaotic phenomena can be observed in quantum dynamics up to t^*, some are much more short lived. For example, exponential sensitivity to initial conditions can be mimicked by quantum motion only up to the so-called Ehrenfest time

$$t_{\mathrm{E}} = \frac{\ln(A_0/\hbar)}{\lambda}$$

where A_0 is the phase space volume explored by a classical chaotic trajectory, and λ is the classical Lyapunov exponent measuring the exponential rate of divergence $\delta x(t) \sim \exp(\lambda t)\delta x(0)$ of nearby trajectories in classical dynamics. Up to time t_{E} Gaussian wave-packets centered on classical trajectories can be used for semi-classical description of quantum motion [26]. Ehrenfest time t_{E} and Heisenberg time t^* are two essential time-scales of quantum chaos.

A simple but fundamental manifestation of quantum chaos is the destruction of quantum interferences by quantum dephasing as a results of chaotic classical dynamics. To illustrate this, let us consider a time-dependent double slit experiment, where the source is closed inside a two dimensional wave resonator in the shape of a chaotic billiard [27]. The setting is depicted in Fig. 3. We take an initial Gaussian wave-packet with average energy corresponding to about the 1600th excited state of the closed billiard and direct it towards two narrow openings – slits – which are a few (about 3) De Broglie wavelengths apart. The wave-packet is taken to be as sharply as possible localized in momentum space, so that, due to Heisenberg uncertainty principle, its spreading in position space is of the order of the diameter of the billiard. Then we solve the time dependent Schrödinger equation and observe the radiation which is transmitted through the two slits to an infinite plane below. We then record the time integrated probability current density $I(x)$ as a function of the horizontal position x on the screen.

The results of such numerical experiment are shown in Fig. 4. We observe almost no sign of quantum interference if the shape of the resonator is chaotic. This is a manifestation of chaotic dynamics of classical rays and consequently phase randomization of multiply reflected waves impacting onto the slits. For comparison we show an analogous result for a regular geometry of the resonator which is represented by an integrable billiard. Here we find the well known interference fringes whose location and visibility is a well controlled function of the initial conditions of the wave-packet. In Fig. 5 we show instant snapshots of the probability density at around half the Heisenberg time for chaotic and regular geometry. The crucial difference is that the jets of probability coming out from the two slits have time-dependent direction in the chaotic case whereas these directions are frozen in time for the regular geometry.

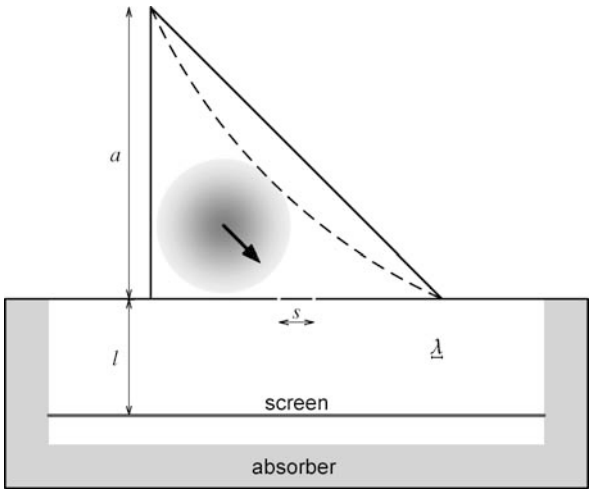

Quantum Chaos, Figure 3
The geometry of the numerical double-slit experiment. All scales are in proper proportions. The two slits are placed at a distance *s* on the lower side of the billiard

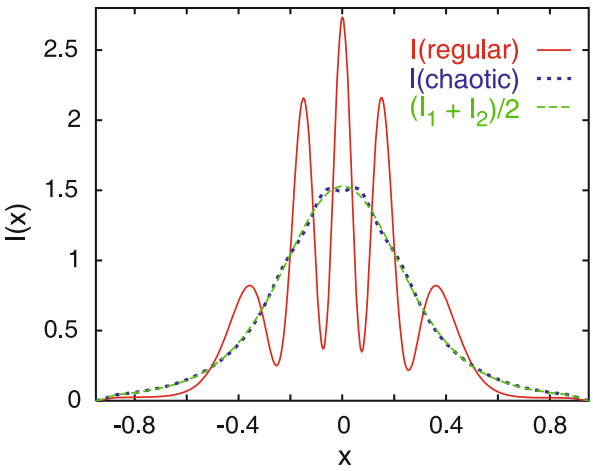

Quantum Chaos, Figure 4
The total intensity after the double-slit experiment as a function of the position on the screen. *I*(*x*) is obtained as the perpendicular component of the probability current, integrated in time. The *red full curve* indicates the case of regular billiard, while the *blue dotted curve* indicates the case of chaotic one. The *green dashed curve* indicates the averaged intensity over two 1-slit experiments (where one of the slits is closed), with either the regular or chaotic billiard (with results being practically the same)

We note that such a simple numerical experiment has a well defined analogue (and explanation) in the stationary quantum chaos. Namely the intensity $I(x)$, in the far field

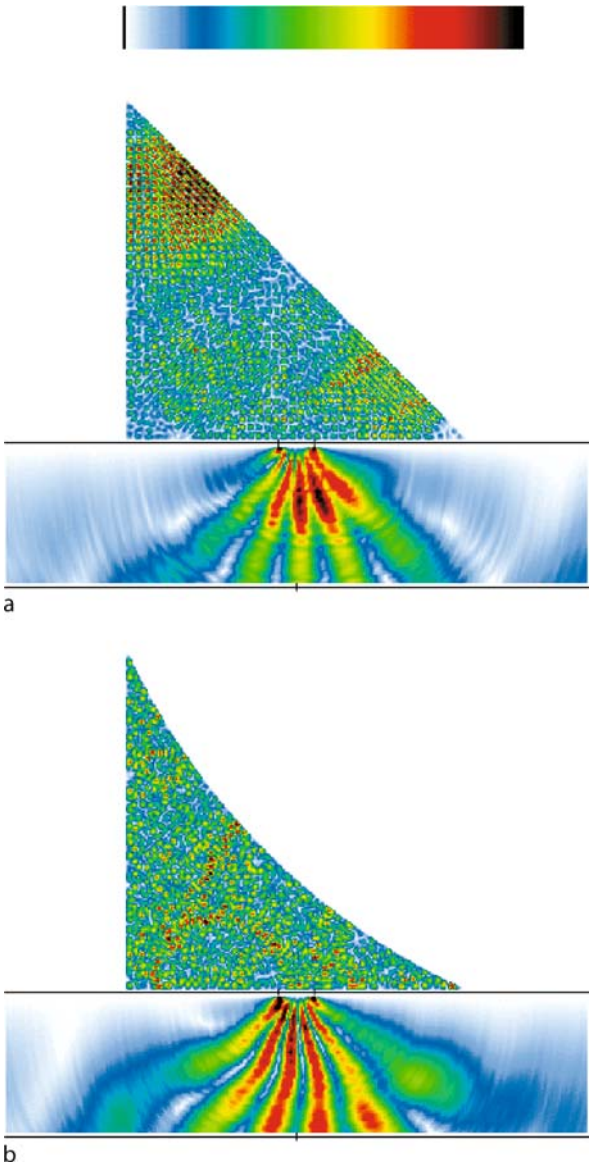

Quantum Chaos, Figure 5

Typical snapshots of the wave-function (plotted is the probability density) for the two cases: a for the regular billiard at $t = 0.325$, and b for the chaotic billiard at $t = 0.275$ (both cases correspond to about half the Heisenberg time). The probability density is normalized separately in both parts of each plot, namely the probability density, in absolute units, in the radiating region is typically less than 1% of the probability density in the billiard domain. The screen, its center, and the positions of the slits are indicated with *thin black lines*. Please note that the color code on the top of the figure is proportional to the square root of probability density

and narrow slit approximations, can be expressed in terms of spatial autocorrelation of the eigenfunctions of the billiard at the two positions of the slit. It is known [10] that chaotic eigenfunctions are characterized by decaying spatial correlations – hence destruction of interference pattern – whereas eigenfunctions of regular systems have long range spatial correlations.

Another important aspect of time-dependent quantum chaos is the study of the so-called Loschmidt echoes or fidelity decay [28]. This quantity has been proposed by Peres [29] as a natural analogy in quantum mechanics of Lyapunov exponents and sensitive dependence to initial conditions. Let $U_0(t)$ represent some unitary quantum mechanical evolution operator, e. g. $U_0(t) = \exp(-\mathrm{i}Ht/\hbar)$ where H is the Hamiltonian, and let $U_\varepsilon(t)$ represent a perturbed evolution, where ε is some small perturbation strength parameter. Quantum Loschmidt echo is defined in terms of fidelity (square modulus of Hilbert space inner product) of a state of perturbed time evolution $|\Psi_\varepsilon(t)\rangle = U_\varepsilon(t)|\Psi(0)\rangle$ with respect to time evolved state of the unperturbed evolution $|\Psi_0(t)\rangle = U_0(t)|\Psi(0)\rangle$, namely

$$F(t) = |\langle\Psi_0(t)|\Psi_\varepsilon(t)\rangle|^2 = |\langle\Psi(0)|U_0(-t)U_\varepsilon(t)|\Psi(0)\rangle|^2 .$$

The first expression (fidelity) can be interpreted as the probability that two nearby quantum evolutions end up in the same state, whereas the second expression (Loschmidt echo) is the probability that the state after forward perturbed time evolution composed with time reversal operation and (backward) unperturbed evolution for the same amount of time, end up in the same (initial) state. The equivalence of the two expressions is a simple consequence of the unitarity of quantum dynamics. The fidelity is a measure of stability of quantum motion. Note that the formalism of Loschmidt echoes can be closely connected to the theory of decoherence in open quantum systems [28]. For example, in many interesting specific situations the Loschmidt echo can be used to bound or estimate certain standard measures of decoherence.

For a semi-classical understanding of Loschmidt echoes it is very useful to write the expression of quantum fidelity in terms of the Wigner phase space function

$$W_\varepsilon(q,p,t) = (2\pi\hbar)^{-d} \cdot \int \mathrm{d}^d r\, \langle q - r/2|\Psi_\varepsilon(t)\rangle\langle\Psi_\varepsilon(t)|q + r/2\rangle \exp(\mathrm{i}p\cdot r/\hbar)$$

corresponding to state $|\Psi_\varepsilon(t)\rangle$, namely

$$F(t) = (2\pi\hbar)^{-d} \int \mathrm{d}^d q\, \mathrm{d}^d p\, W_0(q,p,t) W_\varepsilon(q,p,t) .$$

It is known that the Wigner function of an initial Gaus-

sian wave-packet follows the classical Liouville equation for times below the Ehrenfest time. It can also be easily shown that, for short times, the corresponding *classical fidelity* computed in terms of classical phase space densities decays with the the same rate with which nearby-orbits diverge, namely with the Lyapunov exponent [28]. Therefore, for $t < t_E$, quantum fidelity decays with the perturbation independent rate given by the classical Lyapunov exponent λ, $F_{lyap}(t) \sim \exp(-\lambda t)$.

For times between Ehrenfest and Heisenberg time, $t_E < t < t^*$, and correspondingly small perturbation strength so that the fidelity is appreciable, the decay of Loschmidt echo can be computed using time-dependent perturbation theory, or Fermi golden rule, and one obtains $F_{fgr}(t) \sim \exp(-\varepsilon^2 \sigma t)$ where the coefficient σ is equal to average square of near diagonal matrix elements of the generator of the perturbation. The quantity σ can be semiclassically computed in terms of integrated time-correlation function of the classical perturbation [28].

For even larger times, $t > t^*$, the quantum dynamics is dominated by fluctuations, and fidelity decay, provided ε is small enough, can be computed by static quantum perturbation theory which leads to the Gaussian decay $F_p(t) \sim \exp(-2\varepsilon^2 \sigma t^2/t^*)$.

On the other hand, fidelity decay for classically integrable (regular) systems is somewhat simpler but less universal – namely it depends on the structure of invariant tori of the integrable system and on the initial state. For initial Gaussian wave-packets, the decay of fidelity is typically faster than for chaotic systems, which is a manifestation of *ballistic* nature of regular dynamics as opposed to *diffusive* nature of chaotic dynamics. In Fig. 6 we demonstrate the decay of fidelity for Haake's kicked top model [1] and compare the regimes of regular and chaotic dynamics for the same initial coherent (Gaussian) state and the same value of perturbation parameter. Both cases can be theoretically described and understood [28]. For illustration we plot the Wigner functions of the time evolution as well as the Wigner function of the echo dynamics – perturbed forward evolution composed with unperturbed backward evolution. Note that the Wigner function of a quantum top (spin) is defined over a surface of the sphere.

Quantum chaos has been studied also in unbounded systems with infinite classical phase space, such as the kicked rotator [30], where classical chaos – through decaying temporal correlations – gives rise to deterministic diffusion. It has been shown [30,31,32] that the role of classical deterministic chaos in kicked one-dimensional systems with infinite momentum space is analogous and sometimes even formally equivalent to the role of disorder in one-dimensional tight binding model of a solid. In fact

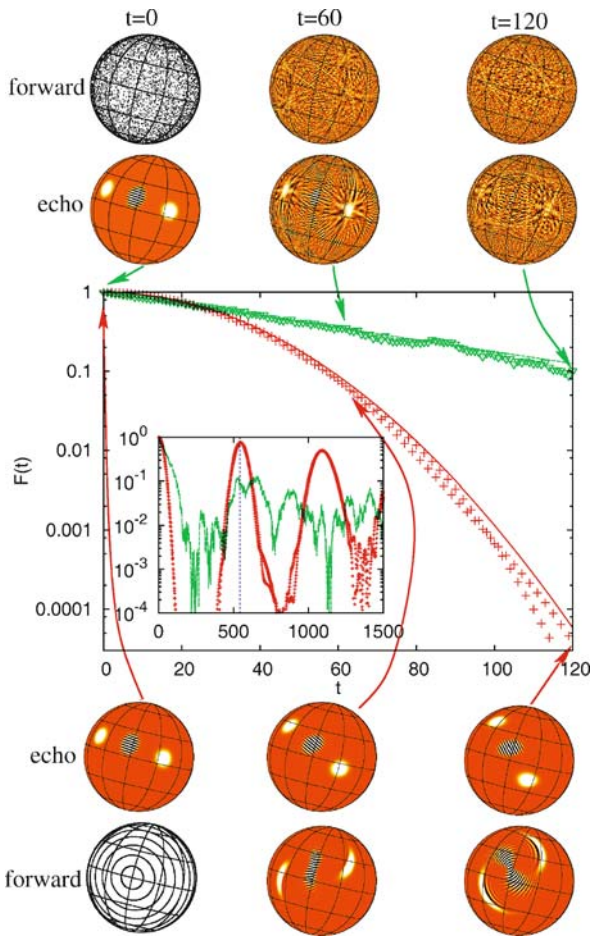

Quantum Chaos, Figure 6
Fidelity decay for chaotic (*top curve* and *pictures*) and regular (*bottom curve* and *pictures*) kicked top. Initial conditions and the perturbation are the same in both case and theoretical formulae, with the only input given in terms of classical dynamics, are shown with *full curves* (see [28] for details). Wigner functions after forward and echo evolution are shown for shown for illustration of the ballistic versus diffusive mechanisms

the phenomenon of *dynamical localization* has been discovered [31] in full analogy with Anderson localization in disordered one-dimensional solids.

Applications of Quantum Chaos

Theoretical phenomena of quantum chaos have been applied or experimentally observed in a variety of fields. Far from being complete we just mention a few.

Atomic Physics

One of the first successful applications of the ideas of quantum chaos has been a theoretical explanation of

multi-photon ionization experiments with hydrogen in microwave field which were performed by Bayfield and Koch [33]. Single hydrogen atoms prepared in very elongated states with high principal quantum number were injected into microwave cavity and the ionization rate was measured. Even though the microwave frequency was well bellow the ionization energy, in fact even lower than the transition energy to the next excited energy level, it was found very surprisingly that very efficient ionization occurred when the electric field intensity exceeded a certain threshold value. The theoretical analysis (described e. g. in [30]) explained the threshold intensities as critical values for the onset of chaotic diffusion in phase space. Quite interestingly, it has been shown that classical mechanics alone accurately describes the results of experiment, namely in the situation above the delocalization border (which has been predicted theoretically [34]) and in the semiclassical regime of high principal quantum number, where an effective value of the Planck constant is sufficiently small. However, in [34] it has been shown that dynamical localization can take place in hydrogen atom and this has been experimentally observed in [35,36].

A second notable experimental achievement has been the observation of dynamical localization, in fact a realization of quantum mechanical kicked rotor in terms of cold Cesium atoms in a standing wave of kicked electric field by the Raizen's group [37].

Mesoscopic Solid State Physics

The ideas of quantum chaos have been extensively investigated in mesoscopic solid state systems, in particular in the studies of quantum transport in the ballistic regime, where the mean free path of the electrons is much larger than the device.

Perhaps it is worth mentioning the discovery of universal conductance fluctuations [38] which correspond to Ericson fluctuations in nuclear physics. It has been shown – by measuring the so-called magneto-resistance in quantum dots (2d conducting structure of a size of the order of a micrometer) – that the conductance fluctuations have some universal statistical features if the shape of the quantum dot represents a chaotic billiard. Universal conductance fluctuations have been later extensively discussed theoretically and explained in terms of random matrix theory (see e. g. review [39]).

Quantum Information

Chaotic quantum dynamics is effective in producing entanglement, which is a key resource for quantum information processing [40,41]. This has been demonstrated theoretically with many chaotic toy models, see e. g. [42]. Perhaps it deserves to be mentioned that two of these models, namely the quantized Baker map and the quantized sawtooth map have been realized in a real world quantum computer [43]. Dynamical chaos can also be explored to engineer robust and accurate decoupling schemes for processing quantum information in the presence of static noise [28].

Wave Chaos

There are many applications of the ideas and mechanisms of quantum chaos outside of quantum mechanics. For example, essentially all the stationary aspects of quantum chaos, and sometimes even some dynamical aspects provided the change in dispersion relations is taken into account, are being beautifully demonstrated in the planar electromagnetic microwave resonators since early 1990's by the groups of Stöckmann (see e. g. [44]), Richter (see e. g. [45]) and Sridhar (see e. g. [46]). Very impressive "quantum chaos" experimental studies were conducted in acoustics, namely with blocks of vibrating solids, see e. g. [47,48]. It turns out that experimentally feasible quality (Q) factors in elastodynamics resonators can be much larger than in microwave billiards, however the complication here being that the underlying amplitude wave equation is of 4th order and cannot be interpreted as a Schrödinger equation. Still, using the universality of quantum chaos one can argue (and find) that statistical properties of resonance frequencies are described by the same statistical random matrix theory [49]. We may also mention a recent application of quantum chaos to optical fibers [50]. Another impressive extension of quantum chaos is to non-linear wave-optics, namely to engineering of directed stimulated emission of micron-size and chaotic-billiard-shaped semiconductor lasers [51].

Future Directions

By now quantum chaos of a single – or few particles – systems is relatively well understood. Among open future problems we should perhaps mention the case of systems with mixed phase space being neither completely integrable nor fully chaotic. In such systems, one of the most important problems is to derive a quantitative theory of tunneling rates between classically disjoint phase space components.

Most of the results so far obtained in the field of quantum chaos are based on numerical evidence and we are still lacking of rigorous proofs. For example, we are missing the proof of quantum chaos conjecture, the precise conditions (ergodic properties of the underlying classical system) un-

der which it holds, precise statements about quantum ergodicity in general Hamiltonian systems, etc.

However, one of the key challenges for future is to understand and apply dynamical chaos mechanisms to quantum systems of many interacting particles. In particular it seems that complete integrability of many body systems implies anomalous non-equilibrium statistical behaviors, such as e. g. ballistic transport, and it seems necessary to operate in the regime of quantum chaos in order to validate diffusive statistical laws in many body transport. Furthermore, it seems that quantum chaos in many body systems is intimately connected to the impossibility of efficient simulation of such systems on classical computers. See e. g. [52] for a recent review of these topics.

At the general level we recall that, while in classical mechanics there is a very well developed ergodic theory which is important for understanding equilibrium and non-equilibrium properties of classical physical systems, in quantum mechanics there is not, so far, a well established theory. A theory would be required which could explain the asymptotic relaxation process (in the correct order of limits, namely letting the time to infinity at last) in the presence of a purely discrete spectrum and in the absence of exponential instability.

We close on a more speculative note. In quantum mechanics we are always facing with the problem of the measurement device which should be treated as a macroscopic classical system. As for such, classical chaos and exponential instability are present. This is indeed even necessary since, by its purpose, a measurement device must be unstable because a microscopic intervention must produce a macroscopic effect. The importance of chaos in the quantum measurement is that it destroys the coherence of the initial pure state to be measured converting it to an incoherent mixture. In the existing theories of quantum measurement this is described as the effect of external noise. Chaos theory allows to get rid of this unsatisfactory assumption and to develop a purely dynamical theory of loss of quantum coherence. Still, a fundamental problem remains open, namely the redistribution of probability according to the result of the measurement: the so-called collapse of the wave-function which remains to be understood.

Bibliography

1. Haake F (2001) Quantum Signatures of Chaos, 2nd edn. Springer, Heidelberg
2. Stöckmann HJ (1999) Quantum Chaos: An Introduction. Cambridge University Press, Cambridge
3. Cornfeld IP, Fomin SV, Sinai YG (1982) Ergodic theory. Springer, New York
4. Casati G, Prosen T (1999) Phys Rev Lett 83:4279; Casati G, Prosen T (2000) Phys Rev Lett 85:4261
5. Mehta ML (1991) Random matrices. Academic Press, New York
6. Casati G, Guarneri I, Valz-Gris (1980) Lett Nuovo Cimento 28:279
7. Bohigas O, Giannoni MJ, Schmit C (1984) Phys Rev Lett 52:1
8. Robnik M (1983) J Phys A 16:3971
9. Bäcker A (2007) Comput Sci Eng 9:60
10. Berry MV (1977) J Phys A 10:2083
11. Feingold M, Peres A (1986) Phys Rev A 34:591
12. Horvat M, Prosen T (2003) J Phys A 36:4015
13. Heller EJ (1984) Phys Rev Lett 53:1515
14. Shnirelman AI (1974) Usp Math Nauk 29:181
15. Zelditch S (1987) Duke Math J 55:919
16. Colin de Verdiere Y (1985) Comm Math Phys 102:111
17. Berry MV, Tabor M (1977) Proc R Soc A 356:375
18. Casati G, Chirikov BV (1985) Phys Rev Lett 54:1350
19. Berry MV, Robnik M (1984) J Phys A 17:2413
20. Gutzwiller MC (1991) Chaos in Classical and Quantum Mechanics. Springer, New York
21. Berry MV, Tabor M (1977) J Phys A 10:371
22. Bogomolny EB (1988) Physica D 31:169
23. Berry MV (1989) Proc Royal Soc A 423:219
24. Müller S, Heusler S, Braun P, Haake F, Altland A (2004) Phys Rev Lett 93:014 103; Müller S, Heusler S, Braun P, Haake F, Altland A (2005) Phys Rev E 72:046 207
25. Sieber M (2002) J Phys A 35:L613
26. Heller EJ (1991) Chaos and Quantum Physics. In: Giannoni MJ, Voros A, Zinn-Justin J (eds) Les Houches, Session LII, 1989. North Holland, Amsterdam, p 547
27. Casati G, Prosen T (2005) Phys Rev A 72:032111
28. Gorin T, Prosen T, Seligman TH, Žnidarič M (2006) Phys Rep 435:33
29. Peres A (1984) Phys Rev A 30:1610
30. Casati G, Chirikov BV (1995) In: Quantum chaos: between order and disorder. Cambridge University Press, Cambridge, p 3
31. Casati G, Chirikov BV, Izrailev FM, Ford J (1979) In: Casati FG, Ford J (eds) Stochastic Behavior in Classical and Quantum Hamiltonian Systems. Lecture Notes in Physics, vol 93. Springer, Berlin
32. Fishman S, Grempel DR, Prange RE (1982) Phys Rev Lett 49:509
33. Bayfield JE, Koch PM (1974) Phys Rev Lett 33:258
34. Casati G, Chirikov BV, Shepelyanski D (1984) Phys Rev Lett 53:2525
35. Bayfield JE, Casati G, Guarneri I, Sokol DW (1989) Phys Rev Lett 63:364
36. Galvez EJ, Sauer BE, Moorman L, Koch PM, Richards D (1988) Phys Rev Lett 61:2011
37. Klappauf BG, Oskay WH, Steck DA, Raizen MG (1999) Physica D 131:78
38. Marcus CM, Rimberg AJ, Westervelt RM, Hopkins PF, Gossard AC (1992) Phys Rev Lett 69:506
39. Beenakker CWJ (1997) Rev Mod Phys 69:731
40. Nielsen MA, Chuang IL (2000) Quantum Computation and Quantum Information. Cambridge University Press, Cambridge
41. Benenti G, Casati G, Strini G (2004) Principles of Quantum Computation and Information, vol I: Basic concepts. World Scientific, Singapore; Benenti G, Casati G, Strini G (2007) Principles of Quantum Computation and Information, vol II: Basic tools and special topics. World Scientific, Singapore

42. Žnidarič M, Prosen T (2003) J Phys A: Math Gen 36:2463
43. Weinstein YA, Lloyd S, Emerson J, Cory DG (2002) Phys Rev Lett 89:157 902; Henry MK, Emerson J, Martinez R, Cory DG (2006) Phys Rev A 74:032 617
44. Stein J, Stöckmann HJ (1992) Phys Rev Lett 68:2867
45. Richter A (2001) Phys Scr T90:212
46. Sridhar S (1991) Phys Rev Lett 67:785
47. Ellegaard C, Schaadt K, Bertelsen P (2001) Phys Scr T90:223
48. Kuhl U, Stöckmann HJ, Weaver R (2005) J Phys A 38:10 433
49. Fyodorov YV, Savin DV, Sommers HJ (2005) J Phys A 38:10 731
50. Doya V, Legrand O, Mortessagne F, Miniatura C (2001) Phys Rev Lett 88:014 102
51. Harayama T, Fukushima T, Sunada S, Ikeda KS (2003) Phys Rev Lett 91:073 903
52. Prosen T (2007) J Phys A: Math Theor 40:7881

Quantum Computational Complexity

John Watrous
Institute for Quantum Computing and School of Computer Science, University of Waterloo, Waterloo, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
The Quantum Circuit Model
Polynomial-Time Quantum Computations
Quantum Proofs
Quantum Interactive Proof Systems
Other Selected Notions in Quantum Complexity
Future Directions
Bibliography

Glossary

Quantum circuit A quantum circuit is an acyclic network of quantum gates connected by wires: the gates represent quantum operations and the wires represent the qubits on which these operations are performed. The quantum circuit model is the most commonly studied model of quantum computation.

Quantum complexity class A quantum complexity class is a collection of computational problems that are solvable by a chosen quantum computational model that obeys certain resource constraints. For example, BQP is the quantum complexity class of all decision problems that can be solved in polynomial time by a quantum computer.

Quantum proof A quantum proof is a quantum state that plays the role of a witness or certificate to a quantum computer that runs a verification procedure. The quantum complexity class QMA is defined by this notion: it includes all decision problems whose yes-instances are efficiently verifiable by means of quantum proofs.

Quantum interactive proof system A quantum interactive proof system is an interaction between a verifier and one or more provers, involving the processing and exchange of quantum information, whereby the provers attempt to convince the verifier of the answer to some computational problem.

Definition of the Subject

The inherent difficulty, or *hardness*, of computational problems is a fundamental concept in computational complexity theory. Hardness is typically formalized in terms of the resources required by different models of computation to solve problems, such as the number of steps of a deterministic Turing machine. A variety of models and resources are often considered, including deterministic, nondeterministic and probabilistic models; time and space constraints; and interactions among models of differing abilities. Many interesting relationships among these different models and resource constraints are known.

One common feature of the most commonly studied computational models and resource constraints is that they are *physically motivated*. This is quite natural, given that computers are physical devices, and to a significant extent it is their study that motivates and directs research on computational complexity. The predominant example is the class of polynomial-time computable functions, which ultimately derives its relevance from physical considerations; for it is a mathematical abstraction of the class of functions that can be efficiently computed without error by physical computing devices.

In light of its close connection to the physical world, it seems only natural that modern physical theories should be considered in the context of computational complexity. In particular, *quantum mechanics* is a clear candidate for a physical theory to have the potential for implications, if not to computational complexity then at least to computation more generally. Given the steady decrease in the size of computing components, it is inevitable that quantum mechanics will become increasingly relevant to the construction of computers—for quantum mechanics provides a remarkably accurate description of extremely small physical systems (on the scale of atoms) where classical physical theories have failed completely. Indeed, an

extrapolation of Moore's Law predicts sub-atomic computing components within the next two decades [74,78]; a possibility inconsistent with quantum mechanics as it is currently understood.

That quantum mechanics should have implications to computational complexity theory, however, is much less clear. It is only through the remarkable discoveries and ideas of several researchers, including Richard Feynman [49], David Deutsch [40,41], Ethan Bernstein and Umesh Vazirani [29,30], and Peter Shor [88,89], that this potential has become evident. In particular, Shor's polynomial-time quantum factoring and discrete-logarithm algorithms [89] give strong support to the conjecture that quantum and classical computers yield differing notions of computational hardness. Other quantum complexity-theoretic concepts, such as the efficient verification of quantum proofs, suggest a wider extent to which quantum mechanics influences computational complexity.

It may be said that the principal aim of quantum computational complexity theory is to understand the implications of quantum physics to computational complexity theory. To this end, it considers the hardness of computational problems with respect to models of quantum computation, classifications of problems based on these models, and their relationships to classical models and complexity classes.

Introduction

This article surveys quantum computational complexity, with a focus on three fundamental notions: polynomial-time quantum computations, the efficient verification of quantum proofs, and quantum interactive proof systems. Based on these notions one defines quantum complexity classes, such as BQP, QMA, and QIP, that contain computational problems of varying hardness. Properties of these complexity classes, and the relationships among these classes and classical complexity classes, are presented. As these notions and complexity classes are typically defined within the quantum circuit model, this article includes a section that focuses on basic properties of quantum circuits that are important in the setting of quantum complexity. A selection of other topics in quantum complexity, including quantum advice, space-bounded quantum computation, and bounded-depth quantum circuits, is also presented.

Two different but closely related areas of study are not discussed in this article: *quantum query complexity* and *quantum communication complexity*. Readers interested in learning more about these interesting and active areas of research may find the surveys of Brassard [33], Cleve [34], and de Wolf [39] to be helpful starting points.

It is appropriate that brief discussions of computational complexity theory and quantum information precede the main technical portion of the article. These discussions are intended only to highlight the aspects of these topics that are non-standard, require clarification, or are of particular importance in quantum computational complexity. In the subsequent sections of this article, the reader is assumed to have basic familiarity with both topics, which are covered in depth by several text books [13,43,60,67,79,82].

Computational Complexity

Throughout this article the binary alphabet $\{0,1\}$ is denoted Σ, and all computational problems are assumed to be encoded over this alphabet. As usual, a function $f\colon \Sigma^* \to \Sigma^*$ is said to be *polynomial-time computable* if there exists a polynomial-time deterministic Turing machine that outputs $f(x)$ for every input $x \in \Sigma^*$. Two related points on the terminology used throughout this article are as follows.

1. A function of the form $p\colon \mathbb{N} \to \mathbb{N}$ (where $\mathbb{N} = \{0,1,2,\ldots\}$) is said to be a *polynomial-bounded function* if and only if there exists a polynomial-time deterministic Turing machine that outputs $1^{f(n)}$ on input 1^n for every $n \in \mathbb{N}$. Such functions are upper-bounded by some polynomial, and are efficiently computable.
2. A function of the particular form $a\colon \mathbb{N} \to [0,1]$ is said to be *polynomial-time computable* if and only if there exists a polynomial-time deterministic Turing machine that outputs a binary representation of $a(n)$ on input 1^n for each $n \in \mathbb{N}$. References to functions of this form in this article typically concern bounds on probabilities that are functions of the length of an input string to some problem.

The notion of *promise problems* [44,52] is central to quantum computational complexity. These are decision problems for which the input is assumed to be drawn from some subset of all possible input strings. More formally, a promise problem is a pair $A = (A_{\text{yes}}, A_{\text{no}})$, where $A_{\text{yes}}, A_{\text{no}} \subseteq \Sigma^*$ are sets of strings satisfying $A_{\text{yes}} \cap A_{\text{no}} = \varnothing$. *Languages* may be viewed as promise problems that obey the additional constraint $A_{\text{yes}} \cup A_{\text{no}} = \Sigma^*$. Although complexity theory has traditionally focused on languages rather than promise problems, little is lost and much is gained in shifting one's focus to promise problems. Karp reductions and the notion of completeness are defined for promise problems in the same way as for languages.

Several classical complexity classes are referred to in this article, and compared with quantum complexity classes when relations are known. The following classical complexity classes, which should hereafter be understood to be classes of promise problems and not just languages, are among those discussed.

P A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in P if and only if there exists a polynomial-time deterministic Turing machine M that accepts every string $x \in A_{\text{yes}}$ and rejects every string $x \in A_{\text{no}}$.

NP A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in NP if and only if there exists a polynomial-bounded function p and a polynomial-time deterministic Turing machine M with the following properties. For every string $x \in A_{\text{yes}}$, it holds that M accepts (x, y) for some string $y \in \Sigma^{p(|x|)}$, and for every string $x \in A_{\text{no}}$, it holds that M rejects (x, y) for all strings $y \in \Sigma^{p(|x|)}$.

BPP A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in BPP if and only if there exists a polynomial-time probabilistic Turing machine M that accepts every string $x \in A_{\text{yes}}$ with probability at least 2/3, and accepts every string $x \in A_{\text{no}}$ with probability at most 1/3.

PP A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in PP if and only if there exists a polynomial-time probabilistic Turing machine M that accepts every string $x \in A_{\text{yes}}$ with probability strictly greater than 1/2, and accepts every string $x \in A_{\text{no}}$ with probability at most 1/2.

MA A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in MA if and only if there exists a polynomial-bounded function p and a probabilistic polynomial-time Turing machine M with the following properties. For every string $x \in A_{\text{yes}}$, it holds that $\Pr[M \text{ accepts } (x, y)] \geq \frac{2}{3}$ for some string $y \in \Sigma^{p(|x|)}$; and for every string $x \in A_{\text{no}}$, it holds that $\Pr[M \text{ accepts } (x, y)] \leq \frac{1}{3}$ for all strings $y \in \Sigma^{p(|x|)}$.

AM A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in AM if and only if there exist polynomial-bounded functions p and q and a polynomial-time deterministic Turing machine M with the following properties. For every string $x \in A_{\text{yes}}$, and at least 2/3 of all strings $y \in \Sigma^{p(|x|)}$, there exists a string $z \in \Sigma^{q(|x|)}$ such that M accepts (x, y, z); and for every string $x \in A_{\text{no}}$, and at least 2/3 of all strings $y \in \Sigma^{p(|x|)}$, there are no strings $z \in \Sigma^{q(|x|)}$ such that M accepts (x, y, z).

SZK A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in SZK if and only if it has a statistical zero-knowledge interactive proof system.

PSPACE A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in PSPACE if and only if there exists a deterministic Turing machine M running in polynomial space that accepts every string $x \in A_{\text{yes}}$ and rejects every string $x \in A_{\text{no}}$.

EXP A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in EXP if and only if there exists a deterministic Turing machine M running in exponential time (meaning time bounded by 2^p, for some polynomial-bounded function p), that accepts every string $x \in A_{\text{yes}}$ and rejects every string $x \in A_{\text{no}}$.

NEXP A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in NEXP if and only if there exists an exponential-time non-deterministic Turing machine N for A.

PL A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in PL if and only if there exists a probabilistic Turing machine M running in polynomial time and logarithmic space that accepts every string $x \in A_{\text{yes}}$ with probability strictly greater than 1/2 and accepts every string $x \in A_{\text{no}}$ with probability at most 1/2.

NC A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in NC if and only if there exists a logarithmic-space generated family $C = \{C_n : n \in \mathbb{N}\}$ of poly-logarithmic depth Boolean circuits such that $C(x) = 1$ for all $x \in A_{\text{yes}}$ and $C(x) = 0$ for all $x \in A_{\text{no}}$.

For most of the complexity classes listed above, there is a standard way to attach an *oracle* to the machine model that defines the class, which provides a subroutine for solving instances of a chosen problem $B = (B_{\text{yes}}, B_{\text{no}})$. One then denotes the existence of such an oracle with a superscriptfor example, P^B is the class of promise problems that can be solved in polynomial time by a deterministic Turing machine equipped with an oracle that solves instances of B (at unit cost). When classes of problems appear as superscripts, one takes the union, as in the following example:

$$\mathrm{P}^{\mathrm{NP}} = \bigcup_{B \in \mathrm{NP}} \mathrm{P}^B .$$

Quantum Information

The standard general description of quantum information is used in this article: mixed states of systems are represented by density matrices and operations are represented by completely positive trace-preserving linear maps. The choice to use this description of quantum information is deserving of a brief discussion, for it will likely be less familiar to many non-experts than the simplified picture of quantum information where states are represented by unit vectors and operations are represented by unitary matrices. This simplified picture is indeed commonly used in

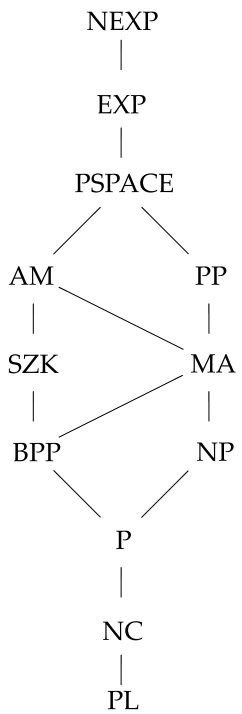

Quantum Computational Complexity, Figure 1
A diagram illustrating known inclusions among most of the classical complexity classes discussed in this paper. Lines indicate containments going upward; for example, AM is contained in PSPACE

the study of both quantum algorithms and quantum complexity theory; and it is often adequate. However, the general picture has major advantages: it unifies quantum information with classical probability theory, brings with it powerful mathematical tools, and allows for simple and intuitive descriptions in many situations where this is not possible with the simplified picture.

Classical simulations of quantum computations, which are discussed below in Subsect. "Classical Upper Bounds on BQP", may be better understood through a fairly straightforward representation of quantum operations by matrices. This representation begins with a representation of density matrices as vectors based on the function defined as $\mathrm{vec}(|x\rangle\langle y|) = |x\rangle|y\rangle$ for each choice of $n \in \mathbb{N}$ and $x, y \in \Sigma^n$, and extended by linearity to all matrices indexed by Σ^n. The effect of this mapping is to form a column vector by reading the entries of a matrix in rows from left to right, starting at the top. For example,

$$\mathrm{vec}\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{pmatrix} .$$

Now, the effect of a general quantum operation Φ, represented in the typical Kraus form as

$$\Phi(\rho) = \sum_{j=1}^{k} A_j \rho A_j^* ,$$

is expressed as a matrix by means of the equality

$$\mathrm{vec}(\Phi(\rho)) = \left(\sum_{j=1}^{k} A_j \otimes \overline{A_j} \right) \mathrm{vec}(\rho) .$$

The matrix

$$K_\Phi = \sum_{j=1}^{k} A_j \otimes \overline{A_j}$$

is sometimes called the *natural representation* (or *linear representation*) of the operation Φ. Although this matrix could have negative or complex entries, one can reasonably view it as being analogous to a stochastic matrix that describes a probabilistic computation.

For example, the complete phase-damping channel for a single qubit can be written

$$D(\rho) = |0\rangle\,\langle 0|\rho|0\rangle\,\langle 0| + |1\rangle\,\langle 1|\rho|1\rangle\,\langle 1| .$$

The effect of this mapping is to zero-out the off-diagonal entries of a density matrix

$$D\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} = \begin{pmatrix} \alpha & 0 \\ 0 & \delta \end{pmatrix} .$$

The natural representation of this operation is easily computed

$$K_D = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} .$$

The natural matrix representation of quantum operations is well-suited to performing computations. For the purposes of this article, the most important observation about this representation is that a composition of operations corresponds simply to matrix multiplication.

The Quantum Circuit Model

General Quantum Circuits

The term *quantum circuit* refers to an acyclic network of *quantum gates* connected by *wires*. The quantum gates represent general quantum operations, involving some

constant number of qubits, while the wires represent the qubits on which the gates act. An example of a quantum circuit having four input qubits and three output qubits is pictured in Fig. 2. In general a quantum circuit may have n input qubits and m output qubits for any choice of integers $n, m \geq 0$. Such a circuit induces some quantum operation from n qubits to m qubits, determined by composing the actions of the individual gates in the appropriate way. The *size* of a quantum circuit is the total number of gates plus the total number of wires in the circuit.

A *unitary quantum circuit* is a quantum circuit in which all of the gates correspond to unitary quantum operations. Naturally this requires that every gate, and hence the circuit itself, has an equal number of input and output qubits. It is common in the study of quantum computing that one works entirely with unitary quantum circuits. The unitary model and general model are closely related, as will soon be explained.

A Finite Universal Gate Set

Restrictions must be placed on the gates from which quantum circuits may be composed if the quantum circuit model is to be used for complexity theory—for without such restrictions it cannot be argued that each quantum gate corresponds to an operation with unit-cost. The usual way in which this is done is simply to fix a suitable finite set of allowable gates. For the remainder of this article, quantum circuits will be assumed to be composed of gates from the following list:

1. *Toffoli gates*. Toffoli gates are three-qubit unitary gates defined by the following action on standard basis states:

$$T\colon |a\rangle|b\rangle|c\rangle \mapsto |a\rangle|b\rangle|c \oplus ab\rangle \,.$$

2. *Hadamard gates*. Hadamard gates are single-qubit unitary gates defined by the following action on standard basis states:

$$H\colon |a\rangle \mapsto \frac{1}{\sqrt{2}}|0\rangle + \frac{(-1)^a}{\sqrt{2}}|1\rangle \,.$$

3. *Phase-shift gates*. Phase-shift gates are single-qubit unitary gates defined by the following action on standard basis states:

$$P\colon |a\rangle \mapsto i^a|a\rangle \,.$$

4. *Ancillary gates*. Ancillary gates are non-unitary gates that take no input and produce a single qubit in the state $|0\rangle$ as output.
5. *Erasure gates*. Erasure gates are non-unitary gates that take a single qubit as input and produce no output. Their effect is represented by the partial trace on the space corresponding to the qubit they take as input.

The symbols used to denote these gates in quantum circuit diagrams are shown in Fig. 3. For the sake of simplifying some of the diagrams that appear later in this article, some additional gates are illustrated in Fig. 4 along with their realizations as circuits with gates from the chosen basis set.

The above gate set is *universal* in a strong sense: every quantum operation can be approximated to within any desired degree of accuracy by some quantum circuit. Moreover, the size of the approximating circuit scales well with respect to the desired accuracy. Theorem 1, stated below, expresses this fact in more precise terms, but requires a brief discussion of a specific sense in which one quantum operation approximates another.

A natural and operationally meaningful way to measure the distance between two given quantum operations Φ and Ψ is given by

$$\delta(\Phi, \Psi) = \frac{1}{2} \| \Phi - \Psi \|_\diamond \,,$$

where $\| \cdot \|_\diamond$ denotes a norm usually known as the *diamond norm* [64,67]. A technical discussion of this norm is

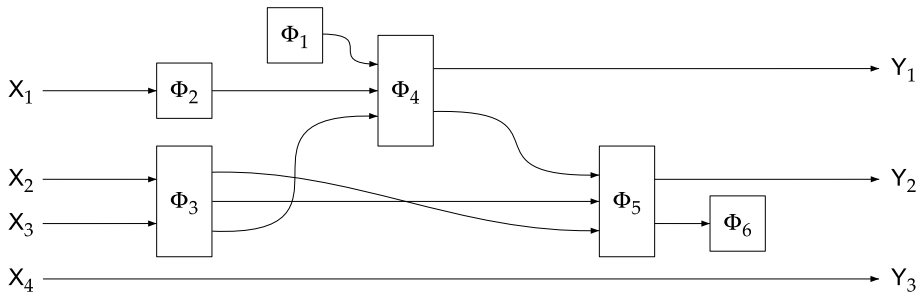

Quantum Computational Complexity, Figure 2
An example of a quantum circuit. The input qubits are labeled $X_1, \ldots, X_4$, the output qubits are labeled $Y_1, \ldots, Y_3$, and the gates are labeled by (hypothetical) quantum operations $\Phi_1, \ldots, \Phi_6$

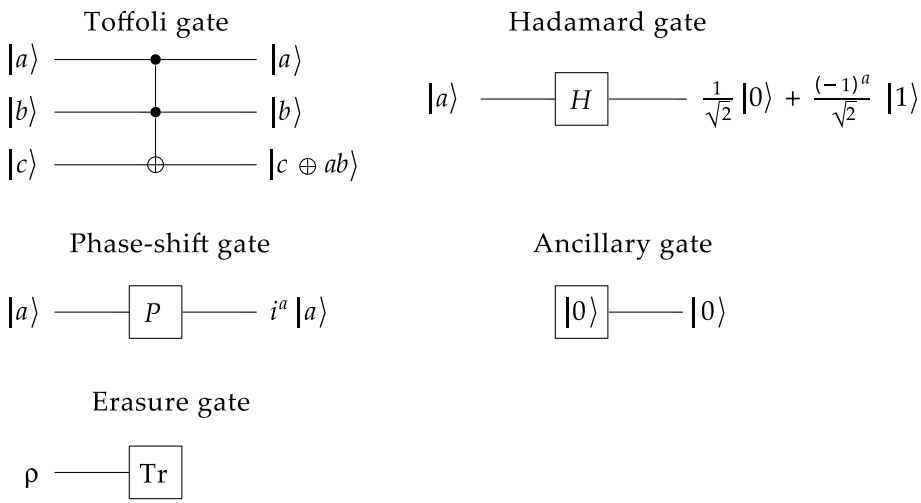

Quantum Computational Complexity, Figure 3
A universal collection of quantum gates: Toffoli, Hadamard, phase-shift, ancillary, and erasure gates

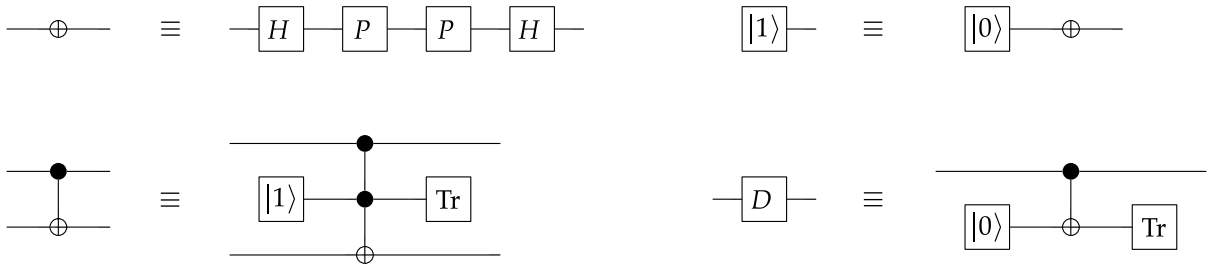

Quantum Computational Complexity, Figure 4
Four additional quantum gates, together with their implementations as quantum circuits. *Top left*: a NOT gate. *Top right*: a constant $|1\rangle$ ancillary gate. *Bottom left*: a controlled-NOT gate. *Bottom right*: a phase-damping (or decoherence) gate

not necessary for the purposes of this article and is beyond its scope. Instead, an intuitive description of the distance measure $\delta(\Phi, \Psi)$ will suffice.

When considering the distance between quantum operations Φ and Ψ, it must naturally be assumed that these operations agree on their numbers of input qubits and output qubits; so assume that Φ and Ψ both map n qubits to m qubits for nonnegative integers n and m. Now, suppose that an arbitrary quantum state on n or more qubits is prepared, one of the two operations Φ or Ψ is applied to the first n of these qubits, and then a general measurement of all of the resulting qubits takes place (including the m output qubits and the qubits that were not among the inputs to the chosen quantum operation). Two possible probability distributions on measurement outcomes arise: one corresponding to Φ and the other corresponding to Ψ. The quantity $\delta(\Phi, \Psi)$ is simply the maximum possible total variation distance between these distributions, ranging over all possible initial states and general measurements. This is a number between 0 and 1 that intuitively represents the *observable difference* between quantum operations.

In the special case that Φ and Ψ take no inputs, the quantity $\delta(\Phi, \Psi)$ is simply one-half the trace norm of the difference between the two output states; a common and useful ways to measure the distance between quantum states.

Now the Universality Theorem, which represents an amalgamation of several results that suits the needs of this article, may be stated. In particular, it incorporates the Solovay–Kitaev Theorem, which provides a bound on the size of an approximating circuit as a function of the accuracy.

Theorem 1 (Universality Theorem) *Let Φ be an arbitrary quantum operation from n qubits to m qubits. Then for every $\varepsilon > 0$ there exists a quantum circuit Q with n input qubits and m output qubits such that $\delta(\Phi, Q) < \varepsilon$.*

Moreover, for fixed n and m, the circuit Q may be taken to satisfy $size(Q) = \text{poly}(\log(1/\varepsilon))$.

Note that it is inevitable that the size of Q is exponential in n and m in the worst case [68]. Further details on the facts comprising this theorem can be found in both Nielsen and Chuang [79] and Kitaev, Shen, and Vyalyi [67].

Unitary Purifications of Quantum Circuits

The connection between general and unitary quantum circuits can be understood through the notion of a *unitary purification* of a general quantum circuit. This may be thought of as a very specific manifestation of the *Stinespring Dilation Theorem* [90], which implies that general quantum operations can be represented by unitary operations on larger systems. It was first applied to the quantum circuit model by Aharonov, Kitaev, and Nisan [9], who gave several arguments in favor of the general quantum circuit model over the unitary model. The term *purification* is borrowed from the notion of a purification of a mixed quantum state, as the process of unitary purification for circuits is similar in spirit. The universal gate set, described in the previous section has the effect of making the notion of a unitary purification of a general quantum circuit nearly trivial at a technical level.

Suppose that Q is a quantum circuit taking input qubits $(X_1, \ldots, X_n)$ and producing output qubits $(Y_1, \ldots, Y_m)$, and assume there are k ancillary gates and l erasure gates among the gates of Q to be labeled in an arbitrary order as $G_1, \ldots, G_k$ and $K_1, \ldots, K_l$, respectively. A new quantum circuit R may then be formed by removing the gates labeled $G_1, \ldots, G_k$ and $K_1, \ldots, K_l$; and to account for the removal of these gates the circuit R takes k additional input qubits $(Z_1, \ldots, Z_k)$ and produces l additional output qubits $(W_1, \ldots, W_l)$. Figure 5 illustrates this process. The circuit R is said to be a unitary purification of Q. It is obvious that R is equivalent to Q, provided the qubits $(Z_1, \ldots, Z_k)$ are initially set to the $|0\rangle$ state and the qubits $(W_1, \ldots, W_l)$ are traced-out, or simply ignored, after the circuit is run—for this is precisely the meaning of the removed gates.

Despite the simplicity of this process, it is often useful to consider the properties of unitary purifications of general quantum circuits.

Oracles in the Quantum Circuit Model

Oracles play an important, and yet uncertain, role in computational complexity theory; and the situation is no different in the quantum setting. Several interesting oracle-related results, offering some insight into the power of quantum computation, will be discussed in this article.

Oracle queries are represented in the quantum circuit model by an infinite family

$$\{R_n : n \in \mathbb{N}\}$$

of quantum gates, one for each possible query length. Each gate R_n is a unitary gate acting on $n + 1$ qubits, with the effect on computational basis states given by

$$R_n|x, a\rangle = |x, a \oplus A(x)\rangle \tag{1}$$

for all $x \in \Sigma^n$ and $a \in \Sigma$, where A is some predicate that represents the particular oracle under consideration. When quantum computations relative to such an oracle are to be studied, quantum circuits composed of ordinary quantum gates as well as the oracle gates $\{R_n\}$ are considered; the interpretation being that each instance of R_n in such a circuit represents one oracle query.

It is critical to many results concerning quantum oracles, as well as most results in the area of quantum query complexity, that the above definition (1) takes each R_n to be unitary, thereby allowing these gates to make queries "in superposition". In support of this seemingly strong definition is the fact, discussed in the next section, that any efficient algorithm (quantum or classical) can be converted to a quantum circuit that closely approximates the action of these gates.

Finally, one may consider a more general situation in which the predicate A is replaced by a function that outputs multiple bits. The definition of each gate R_n is adapted

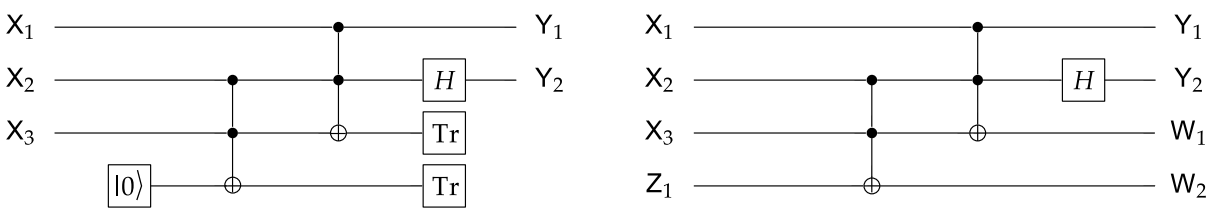

Quantum Computational Complexity, Figure 5
A general quantum circuit *(left)* and its unitary purification *(right)*

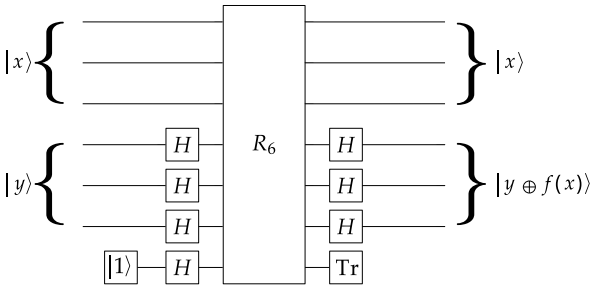

Quantum Computational Complexity, Figure 6
The Bernstein–Vazirani algorithm allows a multiple-bit query to be simulated by a single-bit query. In the example pictured, $f\colon \Sigma^3 \to \Sigma^3$ is a given function. To simulate a query to this function, the gate R_6 is taken to be a standard oracle gate implementing the predicate $A(x, z) = \langle f(x), z\rangle$, for $\langle\cdot,\cdot\rangle$ denoting the modulo 2 inner product

appropriately. Alternately, one may restrict their attention to single-bit queries as discussed above, and use the Bernstein–Vazirani algorithm [30] to simulate one multiple-bit query with one single-bit query as illustrated in Fig. 6.

Polynomial-Time Quantum Computations

This section focuses on *polynomial-time quantum computations*. These are the computations that are viewed, in an abstract and idealized sense, to be efficiently implementable by the means of a quantum computer. In particular, the complexity class BQP (short for *bounded-error quantum polynomial time*) is defined. This is the most fundamentally important of all quantum complexity classes, as it represents the collection of decision problems that can be efficiently solved by quantum computers.

Polynomial-Time Generated Circuit Families and BQP

To define the class BQP using the quantum circuit model, it is necessary to briefly discuss encodings of circuits and the notion of a polynomial-time generated circuit family.

It is clear that any quantum circuit formed from the gates described in the previous section could be *encoded* as a binary string using any number of different encoding schemes. Such an encoding scheme must be chosen, but its specifics are not important so long as following simple restrictions are satisfied:

1. The encoding is sensible: every quantum circuit is encoded by at least one binary string, and every binary string encodes at most one quantum circuit.
2. The encoding is efficient: there is a fixed polynomial-bounded function p such that every circuit of size N has an encoding with length at most $p(N)$. Specific information about the structure of a circuit must be computable in polynomial time from an encoding of the circuit.
3. The encoding disallows compression: it is not possible to work with encoding schemes that allow for extremely short (e. g., polylogarithmic-length) encodings of circuits; so for simplicity it is assumed that the length of every encoding of a quantum circuit is at least the size of the circuit.

Now, as any quantum circuit represents a finite computation with some fixed number of input and output qubits, quantum algorithms are modeled by *families* of quantum circuits. The typical assumption is that a quantum circuit family that describes an algorithm contains one circuit for each possible input length. Precisely the same situation arises here as in the classical setting, which is that it should be possible to efficiently generate the circuits in a given family in order for that family to represent an efficient, finitely specified algorithm. The following definition formalizes this notion.

Definition 2 Let $S \subseteq \Sigma^*$ be any set of strings. Then a collection $\{Q_x : x \in S\}$ of quantum circuits is said to be *polynomial-time generated* if there exists a polynomial-time deterministic Turing machine that, on every input $x \in S$, outputs an encoding of Q_x.

This definition is slightly more general than what is needed to define BQP, but is convenient for other purposes. For instance, it allows one to easily consider the situation in which the input, or some part of the input, for some problem is hard-coded into a collection of circuits; or where a computation for some input may be divided among several circuits. In the most typical case that a polynomial-time generated family of the form $\{Q_n : n \in \mathbb{N}\}$ is referred to, it should be interpreted that this is a shorthand for $\{Q_{1^n} : n \in \mathbb{N}\}$. Notice that every polynomial-time generated family $\{Q_x : x \in S\}$ has the property that each circuit Q_x has size polynomial in $|x|$. Intuitively speaking, the number of quantum and classical computation steps required to implement such a computation is polynomial; and so operations induced by the circuits in such a family are viewed as representing *polynomial-time quantum computations*.

The complexity class BQP, which contains those promise problems abstractly viewed to be efficiently solvable using a quantum computer, may now be defined. More precisely, BQP is the class of promise problems that can be solved by polynomial-time quantum computations that may have some small probability to make an error. For decision problems, the notion of a polynomial-time

quantum computation is equated with the computation of a polynomial-time generated quantum circuit family $Q = \{Q_n : n \in \mathbb{N}\}$, where each circuit Q_n takes n input qubits, and produces one output qubit. The computation on a given input string $x \in \Sigma$ is obtained by first applying the circuit $Q_{|x|}$ to the state $|x\rangle\langle x|$, and then measuring the output qubit with respect to the standard basis. The measurement results 0 and 1 are interpreted as *yes* and *no* (or *accept* and *reject*), respectively. The events that Q accepts x and Q rejects x are understood to have associated probabilities determined in this way.

BQP Let $A = (A_{\text{yes}}, A_{\text{no}})$ be a promise problem and let $a, b : \mathbb{N} \to [0, 1]$ be functions. Then $A \in \text{BQP}(a, b)$ if and only if there exists a polynomial-time generated family of quantum circuits $Q = \{Q_n : n \in \mathbb{N}\}$, where each circuit Q_n takes n input qubits and produces one output qubit, that satisfies the following properties:

1. if $x \in A_{\text{yes}}$ then $\Pr[Q \text{ accepts } x] \geq a(|x|)$, and
2. if $x \in A_{\text{no}}$ then $\Pr[Q \text{ accepts } x] \leq b(|x|)$.

The class BQP is defined as $\text{BQP} = \text{BQP}(2/3, 1/3)$.

Similar to BPP, there is nothing special about the particular choice of error probability 1/3, other than that it is a constant strictly smaller than 1/2. This is made clear in the next section.

There are several problems known to be in BQP but not known (and generally not believed) to be in BPP. Decision-problem variants of the integer factoring and discrete logarithm problems, shown to be in BQP by Shor [89], are at present the most important and well-known examples.

Error Reduction for BQP

When one speaks of the flexibility, or *robustness*, of BQP with respect to error bounds, it is meant that the class $\text{BQP}(a, b)$ is invariant under a wide range of "reasonable" choices of the functions a and b. The following proposition states this more precisely.

Proposition 3 (Error reduction for BQP) *Suppose that $a, b : \mathbb{N} \to [0, 1]$ are polynomial-time computable functions and $p : \mathbb{N} \to \mathbb{N}$ is a polynomial-bounded function such that $a(n) - b(n) \geq 1/p(n)$ for all but finitely many $n \in \mathbb{N}$. Then for every choice of a polynomial-bounded function $q : \mathbb{N} \to \mathbb{N}$ satisfying $q(n) \geq 2$ for all but finitely many $n \in \mathbb{N}$, it holds that*

$$\text{BQP}(a, b) = \text{BQP} = \text{BQP}\left(1 - 2^{-q}, 2^{-q}\right) .$$

The above proposition may be proved in the same standard way that similar statements are proved for classical probabilistic computations: by repeating a given computation some large (but still polynomial) number of times, overwhelming statistical evidence is obtained so as to give the correct answer with an extremely small probability of error. It is straightforward to represent this sort of repeated computation within the quantum circuit model in such a way that the requirements of the definition of BQP are satisfied.

Simulating Classical Computations with Quantum Circuits

It should not be surprising that quantum computers can efficiently simulate classical computers—for quantum information generalizes classical information, and it would be absurd if there were a loss of computational power in moving to a more general model. This intuition may be confirmed by observing the containment $\text{BPP} \subseteq \text{BQP}$. Here an advantage of working with the general quantum circuit model arises: for if one truly believes the Universality Theorem, there is almost nothing to prove.

Observe first that the complexity class P may be defined in terms of Boolean circuit families in a similar manner to BQP. In particular, a given promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in P if and only if there exists a polynomial-time generated family $C = \{C_n : n \in \mathbb{N}\}$ of Boolean circuits, where each circuit C_n takes n input bits and outputs 1 bit, such that

1. $C(x) = 1$ for all $x \in A_{\text{yes}}$, and
2. $C(x) = 0$ for all $x \in A_{\text{no}}$.

A Boolean circuit-based definition of BPP may be given along similar lines: a given promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in BPP if and only if there exists a polynomial-bounded function r and a polynomial-time generated family $C = \{C_n : n \in \mathbb{N}\}$ of Boolean circuits, where each circuit C_n takes $n + r(n)$ input bits and outputs 1 bit, such that

1. $\Pr[C(x, y) = 1] \geq 2/3$ for all $x \in A_{\text{yes}}$, and
2. $\Pr[C(x, y) = 1] \leq 1/3$ for all $x \in A_{\text{no}}$,

where $y \in \Sigma^{r(|x|)}$ is chosen uniformly at random in both cases.

In both definitions, the circuit family C includes circuits composed of constant-size Boolean logic gates—which for the sake of brevity may be assumed to be composed of NAND gates and FANOUT gates. (FANOUT operations must be modeled as gates for the sake of the simulation.) For the randomized case, it may be viewed that the random bits $y \in \Sigma^{r(|x|)}$ are produced by gates that take no input and output a single uniform

random bit. As NAND gates, FANOUT gates, and random bits are easily implemented with quantum gates, as illustrated in Fig. 7, the circuit family C can be simulated gate-by-gate to obtain a quantum circuit family $Q = \{Q_n : n \in \mathbb{N}\}$ for A that satisfies the definition of BQP. It follows that BPP $\subseteq$ BQP.

The BQP Subroutine Theorem

There is an important issue regarding the above definition of BQP, which is that it is not an inherently "clean" definition with respect to the modularization of algorithms. The *BQP subroutine theorem* of Bennett, Brassard, Bernstein and Vazirani [27] addresses this issue.

Suppose that it is established that a particular promise problem A is in BQP, which by definition means that there must exist an efficient quantum algorithm (represented by a family of quantum circuits) for A. It is then natural to consider the use of that algorithm as a subroutine in other quantum algorithms for more complicated problems, and one would like to be able to do this without worrying about the specifics of the original algorithm. Ideally, the algorithm for A should function as an oracle for A, as defined in Subsect. "Oracles in the Quantum Circuit Model".

A problem arises, however, when queries to an algorithm for A are made in superposition. Whereas it is quite common and useful to consider quantum algorithms that query oracles in superposition, a given BQP algorithm for A is only guaranteed to work correctly on classical inputs. It could be, for instance, that some algorithm for A begins by applying phase-damping gates to all of its input qubits, or perhaps this happens inadvertently as a result of the computation. Perhaps it is too much to ask that the existence of a BQP algorithm for A admits a subroutine having the characteristics of an oracle for A?

The BQP subroutine theorem establishes that, up to exponentially small error, this is not too much to ask: the existence of an arbitrary BQP algorithm for A implies the existence of a "clean" subroutine for A with the characteristics of an oracle. A precise statement of the theorem follows.

Theorem 4 (BQP subroutine theorem) *Suppose $A = (A_{\text{yes}}, A_{\text{no}})$ is a promise problem in BQP. Then for any choice of a polynomial-bounded function p there exists a polynomial-bounded function q and a polynomial-time generated family of unitary quantum circuits $\{R_n : n \in \mathbb{N}\}$ with the following properties:*

1. *Each circuit R_n implements a unitary operation U_n on $n + q(n) + 1$ qubits.*
2. *For every $x \in A_{\text{yes}}$ and $a \in \Sigma$ it holds that*
$$\langle x, 0^m, a \oplus 1 | U_n | x, 0^m, a \rangle \geq 1 - 2^{-p(n)}$$
for $n = |x|$ and $m = q(n)$.
3. *For every $x \in A_{\text{no}}$ and $a \in \Sigma$ it holds that*
$$\langle x, 0^m, a | U_n | x, 0^m, a \rangle \geq 1 - 2^{-p(n)}$$
for $n = |x|$ and $m = q(n)$.

The proof of this theorem is remarkably simple: given a BQP algorithm for A, one first uses Proposition 3 to obtain a circuit family Q having exponentially small error for A. The circuit illustrated in Fig. 8 then implements a unitary operation with the desired properties. This is essentially a bounded-error quantum adaptation of a classic construction that allows arbitrary deterministic computations to be performed reversibly [26,92].

The following corollary expresses the main implication of the BQP subroutine theorem in complexity-theoretic terms.

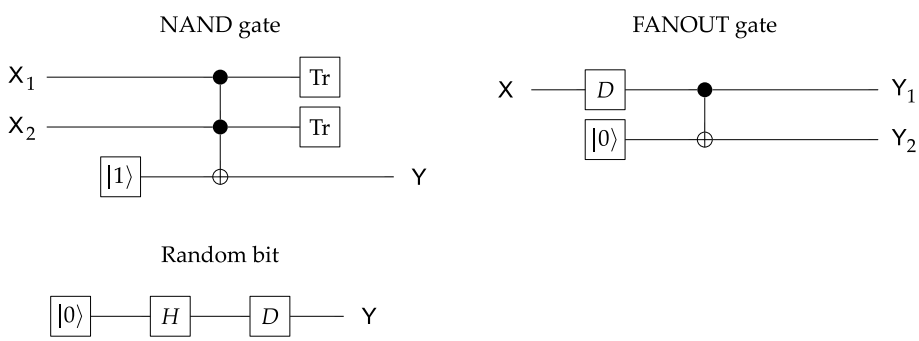

Quantum Computational Complexity, Figure 7
Quantum circuit implementations of a NAND gate, a FANOUT gate, and a random bit. The phase-damping gates, denoted by *D*, are only included for aesthetic reasons: they force the purely classical behavior that would be expected of classical gates, but are not required for the quantum simulation of BPP

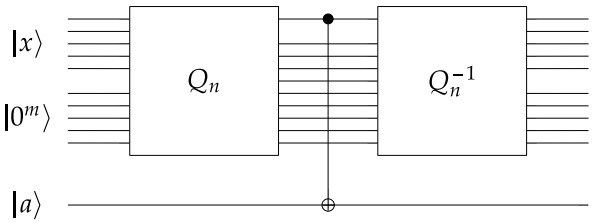

Quantum Computational Complexity, Figure 8
A unitary quantum circuit approximating an oracle-gate implementation of a BQP computation. Here Q_n is a unitary purification of a circuit having exponentially small error for some problem in BQP

Corollary 5 $\mathrm{BQP}^{\mathrm{BQP}} = \mathrm{BQP}$.

Classical Upper Bounds on BQP

There is no known way to efficiently simulate quantum computers with classical computers—and there would be little point in seeking to build quantum computers if there were. Nevertheless, some insight into the limitations of quantum computers may be gained by establishing containments of BQP in the smallest classical complexity classes where this is possible.

The strongest containment known at this time is given by *counting complexity*. Counting complexity began with Valiant's work [94] on the complexity of computing the permanent, and was further developed and applied in many papers (including [22,47,91], among many others).

The basic notion of counting complexity that is relevant to this article is as follows. Given a polynomial-time nondeterministic Turing machine M and input string $x \in \Sigma^*$, one denotes by $\#M(x)$ the number of *accepting* computation paths of M on x, and by $\#\overline{M}(x)$ the number of *rejecting* computation paths of M on x. A function $f\colon \Sigma^* \to \mathbb{Z}$ is then said to be a *GapP function* if there exists a polynomial-time nondeterministic Turing machine M such that $f(x) = \#M(x) - \#\overline{M}(x)$ for all $x \in \Sigma^*$.

A variety of complexity classes can be specified in terms of GapP functions. For example, a promise problem $A = (A_{\mathrm{yes}}, A_{\mathrm{no}})$ is in PP if and only if there exists a function $f \in \mathrm{GapP}$ such that $f(x) > 0$ for all $x \in A_{\mathrm{yes}}$ and $f(x) \leq 0$ for all $x \in A_{\mathrm{no}}$. The remarkable closure properties of GapP functions allows many interesting facts to be proved about such classes. Fortnow's survey [50] on counting complexity explains many of these properties and gives several applications of the theory of GapP functions. The following closure property is used below.

GapP–multiplication of matrices. Let $p, q\colon \mathbb{N} \to \mathbb{N}$ be polynomial-bounded functions. Suppose that for each $n \in \mathbb{N}$ a sequence of $p(n)$ complex-valued matrices is given

$$A_{n,1},\ A_{n,2},\ \ldots,\ A_{n,p(n)}\ ,$$

each having rows and columns indexed by strings in $\Sigma^{q(n)}$. Suppose further that there exist functions $f, g \in \mathrm{GapP}$ such that

$$f(1^n, 1^k, x, y) = \mathrm{Re}\left(A_{n,k}[x, y]\right)$$
$$g(1^n, 1^k, x, y) = \mathrm{Im}\left(A_{n,k}[x, y]\right)$$

for all $n \in \mathbb{N}$, $k \in \{1, \ldots, p(n)\}$, and $x, y \in \Sigma^{q(n)}$. Then there exist functions $F, G \in \mathrm{GapP}$ such that

$$F(1^n, x, y) = \mathrm{Re}\left((A_{n,p(n)} \cdots A_{n,1})[x, y]\right)$$
$$G(1^n, x, y) = \mathrm{Im}\left((A_{n,p(n)} \cdots A_{n,1})[x, y]\right)$$

for all $n \in \mathbb{N}$ and $x, y \in \Sigma^{q(n)}$. In other words, if there exist two GapP functions describing the real and imaginary parts of the entries of a polynomial sequence of matrices, then the same is true of the product of these matrices.

Now, suppose that $Q = \{Q_n : n \in \mathbb{N}\}$ is a polynomial-time generated family of quantum circuits. For each $n \in \mathbb{N}$, let us assume the quantum circuit Q_n consists of gates $G_{n,1}, \ldots, G_{n,p(n)}$ for some polynomial bounded function p, labeled in an order that respects the topology of the circuit. By tensoring these gates with the identity operation on qubits they do not touch, and using the natural matrix representation of quantum operations, it is possible to obtain matrices $M_{n,1}, \ldots, M_{n,p(n)}$ with the property that

$$M_{n,p(n)} \cdots M_{n,1}\mathrm{vec}(\rho) = \mathrm{vec}(Q_n(\rho))$$

for every possible input density matrix ρ to the circuit. The probability that Q_n accepts a given input x is then

$$\langle 1, 1|\left(M_{n,p(n)} \cdots M_{n,1}\right)|x, x\rangle\ ,$$

which is a single entry in the product of the matrices.

By padding matrices with rows and columns of zeroes, it may be assumed that each matrix $M_{n,k}$ has rows and columns indexed by strings of length $q(n)$. The assumption that the family Q is polynomial-time generated then allows one to easily conclude that there exist GapP functions f and g so that

$$f(1^n, 1^k, x, y) = \frac{1}{2}\mathrm{Re}\left(A_{n,k}[x, y]\right)$$
$$g(1^n, 1^k, x, y) = \frac{1}{2}\mathrm{Im}\left(A_{n,k}[x, y]\right)$$

for all $n \in \mathbb{N}$, $k \in \{1, \ldots, p(n)\}$, and $x, y \in \Sigma^{q(n)}$. (Note that his fact makes use of the observation that the natural

matrix representation of all of the gates listed in Subsect. "A Finite Universal Gate Set" have entries whose real and imaginary parts come from the set $\{-1, -1/2, 0, 1/2, 1\}$. The numbers $\pm 1/2$ are only needed for the Hadamard gates.) By the property of GapP functions above, it follows that there exists a GapP function F such that

$$\Pr[Q \text{ accepts } x] = \frac{F(x)}{2^{p(|x|)}} .$$

The containment BQP $\subseteq$ PP follows easily. This containment was first proved by Adleman, DeMarrais, and Huang [7] using a different method, and was first argued using counting complexity along similar lines to the above proof by Fortnow and Rogers [51]. There are two known ways that this upper bound can be improved. Using the fact that BQP algorithms have bounded error, Fortnow and Rogers [51] proved that BQP is contained in a (somewhat obscure) counting complexity class called AWPP, which implies the following theorem.

Theorem 6 $\mathrm{PP}^{\mathrm{BQP}} = \mathrm{PP}$.

Another improvement comes from the observation that the above proof that BQP $\subseteq$ PP makes no use of the bounded-error assumption of BQP. It follows that an unbounded error variant of BQP is equal to PP.

PQP Let $A = (A_{\mathrm{yes}}, A_{\mathrm{no}})$ be a promise problem. Then $A \in \mathrm{PQP}$ if and only if there exists a polynomial-time generated family of quantum circuits $Q = \{Q_n : n \in \mathbb{N}\}$, where each circuit Q_n takes n input qubits and produces one output qubit, that satisfies the following properties. If $x \in A_{\mathrm{yes}}$ then $\Pr[Q \text{ accepts } x] > 1/2$; and if $x \in A_{\mathrm{no}}$ then $\Pr[Q \text{ accepts } x] \leq 1/2$.

Theorem 7 PQP = PP.

Oracle Results Involving BQP

It is difficult to prove separations among quantum complexity classes for apparently the same reasons that this is so for classical complexity classes. For instance, one clearly cannot hope to prove BPP $\neq$ BQP when major collapses such as NC = PP or P = PSPACE are still not disproved. In some cases, however, separations among quantum complexity classes can be proved in relativized settings, meaning that the separation holds in the presence of some cleverly defined oracle. The following oracle results are among several that are known:

1. There exists an oracle A such that $\mathrm{BQP}^A \not\subseteq \mathrm{MA}^A$ [18,30,96]. Such an oracle A intuitively encodes a problem that is solvable using a quantum computer but is not even efficiently verifiable with a classical computer. As the containment BPP $\subseteq$ BQP holds relative to every oracle, this implies that $\mathrm{BPP}^A \subsetneq \mathrm{BQP}^A$ for this particular choice of A.
2. There is an oracle A such that $\mathrm{NP}^A \not\subseteq \mathrm{BQP}^A$ [27]. In less formal terms: a quantum computer cannot find a needle in an exponentially large haystack in polynomial time. This result formalizes a critically important idea, which is that a quantum computer can only solve a given search problem efficiently if it is able to exploit that problem's structure. It is easy to define an oracle search problem represented by A that has no structure whatsoever, which allows the conclusion $\mathrm{NP}^A \not\subseteq \mathrm{BQP}^A$ to be drawn. It is not currently known whether the NP-complete problems, in the absence of an oracle, have enough structure to be solved efficiently by quantum computers; but there is little hope and no indication whatsoever that this should be so. It is therefore a widely believed conjecture that NP $\not\subseteq$ BQP.
3. There is an oracle A such that $\mathrm{SZK}^A \not\subseteq \mathrm{BQP}^A$ [1,5]. This result is similar in spirit to the previous one, but is technically more difficult and rules out the existence of quantum algorithms for unstructured collision detection problems. The graph isomorphism problem and various problems that arise in cryptography are examples of collision detection problems. Once again, it follows that quantum algorithms can only solve such problems if their structure is exploited. It is a major open question in quantum computing whether the graph isomorphism problem is in BQP.

Quantum Proofs

There are many quantum complexity classes of interest beyond BQP. This section concerns one such class, which is a quantum computational analogue of NP. The class is known as QMA, short for *quantum Merlin–Arthur*, and is based on the notion of a *quantum proof*: a quantum state that plays the role of a certificate or witness to a quantum computer that functions as a verification procedure. Interest in both the class QMA and the general notion of quantum proofs is primarily based on the fundamental importance of *efficient verification* in computational complexity. The notion of a quantum proof was first proposed by Knill [69] and consider more formally by Kitaev (presented at a talk in 1999 [65] and later published in [67]).

Definition of QMA

The definition of QMA is inspired by the standard definition of NP included in Subsect. "Computational Complex-

ity" of this article. This definition is of course equivalent to the other well-known definition of NP based on nondeterministic Turing machines, but is much better-suited to consideration in the quantum setting—for nondeterminism is arguably a non-physical notion that does not naturally extend to quantum computing. In the definition of NP from Subsect. "Computational Complexity", the machine M functions as a *verification procedure* that treats each possible string $y \in \Sigma^{p(|x|)}$ as a potential *proof* that $x \in A_{\text{yes}}$. The conditions on M are known as the *completeness* and *soundness* conditions, which derive their names from logic: completeness refers to the condition that true statements have proofs, while soundness refers to the condition that false statements do not.

To define QMA, the set of possible proofs is extended to include quantum states, which of course presumes that the verification procedure is quantum. As quantum computations are inherently probabilistic, a bounded probability of error is allowed in the completeness and soundness conditions. (This is why the class is called QMA rather than QNP, as it is really MA and not NP that is the classical analogue of QMA.)

QMA Let $A = (A_{\text{yes}}, A_{\text{no}})$ be a promise problem, let p be a polynomial-bounded function, and let $a, b\colon \mathbb{N} \to [0,1]$ be functions. Then $A \in \text{QMA}_p(a,b)$ if and only if there exists a polynomial-time generated family of circuits $Q = \{Q_n : n \in \mathbb{N}\}$, where each circuit Q_n takes $n + p(n)$ input qubits and produces one output qubit, with the following properties:

1. *Completeness.* For all $x \in A_{\text{yes}}$, there exists a $p(|x|)$-qubit quantum state ρ such that $\Pr[Q \text{ accepts } (x, \rho)] \geq a(|x|)$.
2. *Soundness.* For all $x \in A_{\text{no}}$ and all $p(|x|)$-qubit quantum states ρ it holds that $\Pr[Q \text{ accepts } (x, \rho)] \leq b(|x|)$.

Also define $\text{QMA} = \bigcup_p \text{QMA}_p(2/3, 1/3)$, where the union is over all polynomial-bounded functions p.

Problems in QMA

Before discussing the general properties of QMA that are known, it is appropriate to mention some examples of problems in this class. Of course it is clear that thousands of interesting combinatorial problems are in QMA, as the containment $\text{NP} \subseteq \text{QMA}$ is trivial. What is more interesting is the identification of problems in QMA that are not known to be in NP, for these are examples that provide insight into the power of quantum proofs. There are presently just a handful of known problems in QMA that are not known to be in NP, but the list is growing.

The Local Hamiltonian Problem Historically speaking, the first problem identified to be complete for QMA was the *local Hamiltonian problem* [65,67], which can be seen as a quantum analogue of the MAX-k-SAT problem. Its proof of completeness can roughly be seen as a quantum analogue of the proof of the Cook–Levin Theorem [38,72]. The proof has subsequently been strengthened to achieve better parameters [61].

Suppose that M is a Hermitian matrix whose rows and columns are indexed by strings of length n for some integer $n \geq 1$. Then M is said to be *k-local* if and only if it can be expressed as

$$M = P_\pi (A \otimes I) P_\pi^{-1}$$

for an arbitrary matrix A indexed by Σ^k, P_π a permutation matrix defined by

$$P_\pi |x_1 \cdots x_n\rangle = |x_{\pi(1)} \cdots x_{\pi(n)}\rangle$$

for some permutation $\pi \in S_n$, and I denoting the identity matrix indexed by Σ^{n-k}. In less formal terms, M is a matrix that arises from a "gate" on k qubits, but where the gate is described by a $2^k \times 2^k$ Hermitian matrix A rather than a unitary matrix. It is possible to express such a matrix compactly by specifying A along with the bit-positions on which A acts.

Intuitively, a k-local matrix assigns a real number (representing the *energy*) to any quantum state on n qubits. This number depends only on the reduced state of the k qubits where M acts nontrivially, and a limit on this value can be thought of as a local constraint on a given quantum state. Loosely speaking, the k-local Hamiltonian problem asks whether there exists a quantum state that satisfies a collection of such constraints.

The k-local Hamiltonian problem

Input: A collection $H_1, \ldots, H_m$ of k-local Hermitian matrices indexed by strings of length n and satisfying $\| H_j \| \leq 1$ for $j = 1, \ldots, m$.

Yes: There exists an n-qubit quantum state $|\psi\rangle$ such that $\psi | H_1 + \cdots + H_m | \psi \leq -1$.

No: For every n-qubit quantum state $|\psi\rangle$ it holds that $\psi | H_1 + \cdots + H_m | \psi \geq 1$.

Theorem 8 *The 2-local Hamiltonian problem is complete for QMA with respect to Karp reductions.*

The completeness of this problem has been shown to be closely related to the universality of the so-called *adiabatic* model of quantum computation [11].

Other Problems Complete for QMA Aside from the local Hamiltonian problem, there are other promise prob-

lems known to be complete for QMA, including the following:

1. Restricted versions of the local Hamiltonian problem. For example, the 2-local Hamiltonian problem remains QMA-complete when the local Hamiltonians are restricted to nearest-neighbor interactions on a two-dimensional array of qubits [81]. The hardness of the local Hamiltonian problem with nearest-neighbor interactions on one-dimensional systems is known to be QMA-complete for 12 dimensional particles in place of qubits [10], but is open for smaller systems including qubits.
2. The *density matrix consistency problem*. Here, the input is a collection of density matrices representing the reduced states of a hypothetical n-qubit state. The input is a yes-instance of the problem if there exists a state of the n-qubit system that is consistent with the given reduced density matrices, and is a no-instance if every state of the n-qubit system has reduced states that have significant disagreement from one or more of the input density matrices. This problem is known to be complete for QMA with respect to Cook reductions [73], but is not known to be complete for Karp reductions.
3. The *quantum clique problem*. Beigi and Shor [23] have proved that a promise version of the following problem is QMA-complete for any constant $k \geq 2$. Given a quantum operation Φ, are there k different inputs to Φ that are perfectly distinguishable after passing through Φ? They name this the quantum clique problem because there is a close connection between computing the zero-error capacity of a classical channel and computing the size of the largest clique in the complement of a graph representing the channel; and this is a quantum variant of that problem.
4. Several problems about quantum circuits. All of the problems mentioned above are proved to be QMA-hard through reductions from the local Hamiltonian problem. Other problems concerning properties of quantum circuits can be proved QMA-complete by more direct means, particularly when the problem itself concerns the existence of a quantum state that causes an input circuit to exhibit a certain behavior. An example in this category is the *non-identity check* problem [59] that asks whether a given unitary circuit implements an operation that is close to some scalar multiple of the identity.

The Group Non-Membership Problem Another example of a problem in QMA that is not known to be in NP is the *group non-membership problem* [17,96]. This problem is quite different from the QMA-complete problems mentioned above in two main respects. First, the problem corresponds to a language, meaning that the promise is vacuous: every possible input can be classified as a yes-instance or no-instance of the problem. (In all of the above problems it is not possible to do this without invalidating the known proofs that the problems are in QMA.) Second, the problem is not known to be complete for QMA; and indeed it would be quite surprising if QMA were shown to have a complete problem having a vacuous promise.

In the group non-membership problem the input is a subgroup H of some finite group G, along with an element $g \in G$. The yes-instances of the problem are those for which $g \notin H$, while the no-instances are those for which $g \in H$.

The group non-membership problem

Input: Group elements $h_1, \ldots, h_k$ and g from some finite group G. Let $H = \langle h_1, \ldots, h_k \rangle$ be the subgroup generated by $h_1, \ldots, h_k$.

Yes: $g \notin H$.

No: $g \in H$.

Like most group-theoretic computational problems, there are many specific variants of the group non-membership problem that differ in the way that group elements are represented. For example, group elements could be represented by permutations in cycle notation, invertible matrices over a finite field, or any number of other possibilities. The difficulty of the problem clearly varies depending the representation. The framework of *black-box groups*, put forth by Babai and Szemerédi [20], simplifies this issue by assuming that group elements are represented by meaningless labels that can only be multiplied or inverted by means of a *group oracle*. Any algorithm or protocol that solves a problem in this framework can then be adapted to any specific group provided that an efficient implementation of the group operations exists that can replace the group oracle.

Theorem 9 *The group non-membership problem is in QMA for every choice of a group oracle.*

Error Reduction for QMA

The class QMA is robust with error bounds, which are reflected by the completeness and soundness probabilities, in much the same way that the same is true of BQP. Two specific error reduction methods are presented in this section: *weak error reduction* and *strong error reduction*. The two methods differ with respect to the length of the quantum proof that is needed to obtain a given error bound, which is a unique consideration that arises when working with quantum proofs.

Suppose that $A = (A_{\text{yes}}, A_{\text{no}})$ is a given promise problem for which a QMA-type quantum verification procedure exists. To describe both error reduction procedures, it is convenient to assume that the input to the problem is hard-coded into each circuit in the family representing the verification procedure. The verification procedure therefore takes the form $Q = \{Q_x : x \in \Sigma^*\}$, where each circuit Q_x takes a $p(|x|)$-qubit quantum state as input, for some polynomial-bounded function p representing the quantum proof length, and produces one output qubit. The completeness and soundness probabilities are assumed to be given by polynomial-time computable functions a and b as usual.

The natural approach to error reduction is repetition: for a given input x the verification circuit Q_x is evaluated multiple times, and a decision to accept or reject based on the frequency of accepts and rejects among the outcomes of the repetitions is made. Assuming $a(|x|)$ and $b(|x|)$ are not too close together, this results in a decrease in error that is exponential in the number of repetitions [67]. The problem that arises, however, is that each repetition of the verification procedure apparently destroys the quantum proof it verifies, which necessitates the composite verification procedure receiving $p(|x|)$ qubits for *each* repetition of the original procedure as illustrated in Fig. 9. The form of error reduction implemented by this procedure is called *weak error reduction*, given that the length of the quantum proof must grow as the error decreases. (Of course one may view that it also has a strength, which is that it does not require a significant increase in circuit depth over the original procedure.)

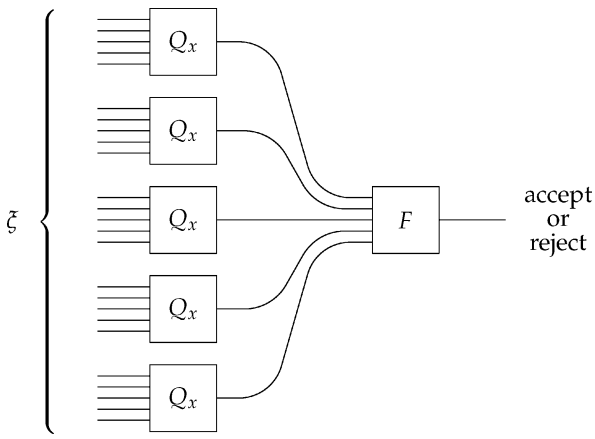

Quantum Computational Complexity, Figure 9
An illustration of the weak error reduction procedure for QMA. The circuits Q_x represent the original verification procedure, and the circuit labeled F outputs acceptance or rejection based on the frequency of accepts among its inputs (which can be adjusted depending on a and b). It cannot be assumed that the input state ξ takes the form of a product state $\rho \otimes \cdots \otimes \rho$; but it is not difficult to prove that there will always be at least one such input state among the states maximizing the acceptance probability of the procedure

The second error reduction procedure for QMA is due to Marriott and Watrous [76]. Like weak error reduction it gives an exponential reduction in error for roughly a linear increase in circuit size, but has the advantage that it does not require any increase in the length of the quantum proof. This form of error reduction is called *strong error reduction* because of this advantage, which turns out to be quite handy in some situations. The procedure is illustrated in Fig. 10. The following theorem follows from an analysis of this procedure.

Theorem 10 *Suppose that $a, b : \mathbb{N} \to [0, 1]$ are polynomial-time computable functions and $q : \mathbb{N} \to \mathbb{N}$ is a polynomial-bounded function such that $a(n) - b(n) \geq 1/q(n)$ for all but finitely many $n \in \mathbb{N}$. Then for every choice of polynomial-bounded functions $p, r : \mathbb{N} \to \mathbb{N}$ such that $r(n) \geq 2$ for all but finitely many n, it holds that* $\mathrm{QMA}_p(a, b) = \mathrm{QMA}_p(1 - 2^{-r}, 2^{-r})$.

Containment of QMA in PP

By means of strong error reduction for QMA, it can be proved that QMA is contained in PP as follows. Suppose that some promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is contained in QMA. Then by Theorem 10 it holds that

$$A \in \mathrm{QMA}_p\left(1 - 2^{-(p+2)}, 2^{-(p+2)}\right) \tag{2}$$

for some polynomial-bounded function p. What is important here is that the soundness probability is smaller than the reciprocal of the dimension of the space corresponding to the quantum proof, which is where strong error reduction is essential.

Now, consider an algorithm that does not receive any quantum proof, but instead just randomly guesses a quantum proof on p qubits and feeds this proof into a verification procedure having completeness and soundness probabilities consistent with the above inclusion (2). To be more precise, the quantum proof is substituted by the totally mixed state on p qubits. A simple analysis shows that this algorithm accepts every string $x \in A_{\text{yes}}$ with probability at least $2^{-(p(|x|)+1)}$ and accepts every string $x \in A_{\text{no}}$ with probability at most $2^{-(p(|x|)+2)}$. The gap between these two probabilities is enough to establish that $A \in \mathrm{PQP} = \mathrm{PP}$.

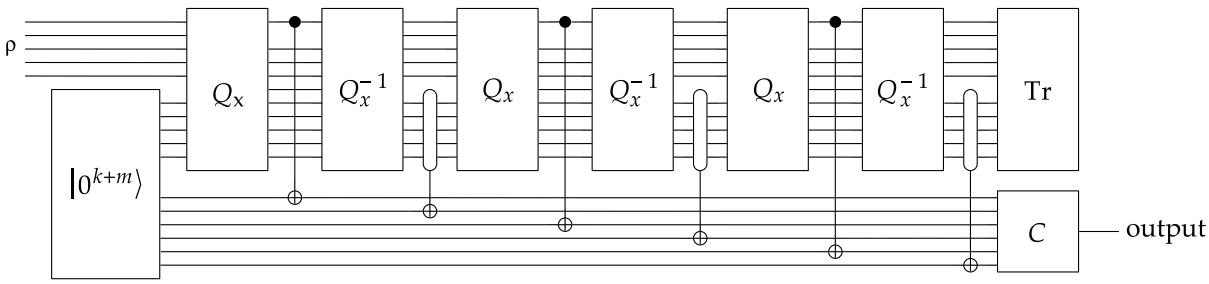

Quantum Computational Complexity, Figure 10
Illustration of the strong error reduction procedure for QMA. The circuit Q_x represents a unitary purification of a given QMA verification procedure on an input x, while the circuit C determines whether to accept or reject based on the number of alternations of its input qubits. The quantum proof is denoted by ρ

Classical Proofs for Quantum Verification Procedures

One may also consider quantum verification procedures that receive classical proofs rather than quantum proofs. Aharonov and Naveh [8] defined a complexity class MQA accordingly.

MQA Let $A = (A_{\text{yes}}, A_{\text{no}})$ be a promise problem. Then $A \in \text{MQA}$ if and only if there exists a polynomial-bounded function p and a polynomial-time generated family of circuits $Q = \{Q_n : n \in \mathbb{N}\}$, where each circuit Q_n takes $n + p(n)$ input qubits and produces one output qubit, with the following properties. For all $x \in A_{\text{yes}}$, there exists a string $y \in \Sigma^{p(|x|)}$ such that $\Pr[Q \text{ accepts } (x, y)] \geq 2/3$; and for all $x \in A_{\text{no}}$ and all strings $y \in \Sigma^{p(|x|)}$ it holds that $\Pr[Q \text{ accepts } (x, y)] \leq 1/3$.

(This class was originally named QCMA, and is more commonly known by that name—but it is impossible to resist the urge to give this interesting class a more sensible name.)

When considering the power of quantum proofs, it is the question of whether MQA is properly contained in QMA that arguably cuts to the heart of the issue. Aaronson and Kuperberg [4] studied this question, and based on a reasonable group-theoretic conjecture argued that the group non-membership problem is likely to be in MQA.

Are Two Quantum Proofs Better than One?

An unusual, yet intriguing question about quantum proofs was asked by Kobayashi, Matsumoto, and Yamakami [71]: *are two quantum proofs better than one?* The implicit setting in this question is similar to that of QMA, except that the verification procedure receives *two* quantum proofs that are *guaranteed to be unentangled*. The complexity class QMA(2) is defined to be the class of promise problems having such two-proof systems with completeness and soundness probabilities 2/3 and 1/3, respectively. (It is not known to what extent this type of proof system is robust with respect to error bounds, so it is conceivable that other reasonable choices of error bounds could give distinct complexity classes.)

The restriction on entanglement that is present in the definition of QMA(2) may seem artificial from a physical perspective, as there is no obvious mechanism that would prevent two entities with otherwise unlimited computational power from sharing entanglement. Nevertheless, the question of the power of QMA(2) is interesting in that it turns around the familiar question in quantum computing about the computational power of entanglement. Rather than asking if computational power is limited by a constraint on entanglement, here the question is whether a constraint on entanglement enhances computational power. It is clear that two quantum proofs are no less powerful than one, as a verification procedure may simply ignore one of the proofs—which means that QMA $\subseteq$ QMA(2). Good upper bounds on QMA(2), however, are not known: the best upper bound presently known is QMA(2) $\subseteq$ NEXP, which follows easily from the fact that a nondeterministic exponential-time algorithm can guess explicit descriptions of the quantum proofs and then perform a deterministic exponential-time simulation of the verification procedure.

Quantum Interactive Proof Systems

The notion of efficient proof verification is generalized by means of the *interactive proof system* model, which has fundamental importance in complexity theory and theoretical cryptography. Interactive proof systems, or *interactive proofs* for short, were first introduced by Babai [17,19] and Goldwasser, Micali, and Rackoff [55,56], and have led

to remarkable discoveries in complexity such as the *PCP Theorem* [14,15,42]. In the most commonly studied variant of interactive proof systems, a polynomial-time *verifier* interacts with a computationally unbounded *prover* that tries to convince the verifier of the truth of some statement. The prover is not trustworthy, however, and so the verifier must be specified in such a way that it is not convinced that false statements are true.

Quantum interactive proof systems [66,99] are interactive proof systems in which the prover and verifier may exchange and process quantum information. This ability of both the prover and verifier to exchange and process quantum information endows quantum interactive proofs with interesting properties that distinguish them from classical interactive proofs, and illustrate unique features of quantum information.

Definition of Quantum Interactive Proofs and The Class QIP

A quantum interactive proof system involves an interaction between a prover and verifier as suggested by Fig. 11.

The verifier is described by a polynomial-time generated family

$$V = \left\{V_j^n : n \in \mathbb{N},\ \ j \in \{1, \ldots, p(n)\}\right\}$$

of quantum circuits, for some polynomial-bounded function p. On an input string x having length n, the sequence of circuits

$$V_1^n, \ldots, V_{p(n)}^n$$

determine the actions of the verifier over the course of the interaction, with the value $p(n)$ representing the number of *turns* the verifier takes. For instance, $p(n) = 4$ in the example depicted in Fig. 11. The inputs and outputs of the verifier's circuits are divided into two categories: private *memory* qubits and *message* qubits. The message qubits are sent to, or received from, the prover, while the memory qubits are retained by the verifier as illustrated in the figure. It is always assumed that the verifier *receives* the last message; so if the number of messages is to be $m = m(n)$, then it must be the case that $p(n) = \lfloor m/2 \rfloor + 1$.

The prover is defined in a similar way to the verifier, although no computational assumptions are made: the prover is a family of arbitrary quantum operations

$$P = \{P_j^n : n \in \mathbb{N},\ \ j \in \{1, \ldots, q(n)\}\}$$

that interface with a given verifier in the natural way. Again, this is as suggested by Fig. 11.

Now, on a given input string x, the prover P and verifier V have an interaction by composing their circuits as described above, after which the verifier measures an output qubit to determine acceptance or rejection. In direct analogy to the classes NP, MA, and QMA, one defines quantum complexity classes based on the completeness and soundness properties of such interactions.

QIP Let $A = (A_{\text{yes}}, A_{\text{no}})$ be a promise problem, let m be a polynomial-bounded function, and let $a, b\colon \mathbb{N} \to [0, 1]$ be polynomial-time computable functions. Then $A \in \text{QIP}(m, a, b)$ if and only if there exists an m-message quantum verifier V with the following properties:

1. *Completeness.* For all $x \in A_{\text{yes}}$, there exists a quantum prover P that causes V to accept x with probability at least $a(|x|)$.
2. *Soundness.* For all $x \in A_{\text{no}}$, every quantum prover P causes V to accept x with probability at most $b(|x|)$.

Also define $\text{QIP}(m) = \text{QIP}(m, 2/3, 1/3)$ for each polynomial-bounded function m and define $\text{QIP} = \bigcup_m \text{QIP}(m)$, where the union is over all polynomial-bounded functions m.

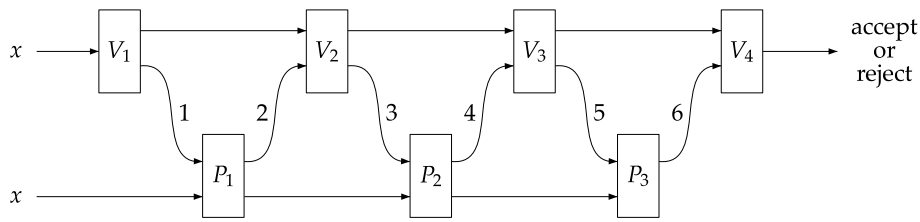

Quantum Computational Complexity, Figure 11
A quantum interactive proof system. There are six messages in this example, labeled 1, . . . , 6. (There may be polynomially many messages in general.) The arrows each represent a collection of qubits, rather than single qubits as in previous figures. The superscript *n* is omitted in the names of the prover and verifier circuits, which can safely be done when the input length *n* is determined by context

Properties of Quantum Interactive Proofs

Classical interactive proofs can trivially be simulated by quantum interactive proofs, and so the containment PSPACE $\subseteq$ QIP follows directly from IP = PSPACE [75, 87]. Unlike classical interactive proofs, quantum interactive proofs are not known to be simulatable in PSPACE. Simulation is possible in EXP through the use of semidefinite programming [66].

Theorem 11 QIP $\subseteq$ EXP.

Quantum interactive proofs, like ordinary classical interactive proofs, are quite robust with respect to the choice of completeness and soundness probabilities. In particular, the following facts hold [66].

1. Every quantum interactive proof can be transformed into an equivalent quantum interactive proof with *perfect completeness*: the completeness probability is 1 and the soundness probability is bounded away from 1. The precise bound obtained for the soundness probability depends on the completeness and soundness probabilities of the original proof system. This transformation to a perfect-completeness proof system comes at the cost of one additional round (i. e., two messages) of communication.
2. Parallel repetition of quantum interactive proofs with perfect completeness gives an exponential reduction in soundness error. An exponential reduction in completeness and soundness error is also possible for quantum interactive proofs not having perfect completeness, but the procedure is more complicated than for the perfect completeness case.

One of the major differences between quantum and classical interactive proof systems is that quantum interactive proofs can be *parallelized*: every quantum interactive proof system, possibly having polynomially many rounds of communication, can be transformed into an equivalent quantum interactive proof system with *three messages* [66]. This transformation comes at the cost of weakening the error bounds. However, it preserves perfect completeness, and the soundness error can subsequently be reduced by parallel repetition without increasing the number of messages beyond three. If classical interactive proof systems could be parallelized in this way, then it would follow that AM = PSPACE; a collapse that would surprise most complexity theorists and have major implications to the theory.

The following theorem summarizes the implications of the facts just discussed to the complexity classes QIP(m, a, b) defined above.

Theorem 12 *Let $a, b\colon \mathbb{N} \to [0,1]$ be polynomial-time computable functions and let p be a polynomial-bounded function such that $a(n) - b(n) \geq 1/p(n)$ for all but finitely many $n \in \mathbb{N}$. Also let m and r be polynomial-bounded functions. Then*

$$\mathrm{QIP}(m, a, b) \subseteq \mathrm{QIP}(3, 1, 2^{-r}) .$$

For a wide range of completeness and soundness probabilities this leaves just four complexity classes among those defined above: QIP(0) = BQP, QIP(1) = QMA, QIP(2), and QIP(3) = QIP.

The final property of quantum interactive proofs that will be discussed in this section is the existence of an interesting complete promise problem [85]. The problem is to determine whether the operations induced by two quantum circuits are significantly different, or are approximately the same, with respect to the same metric on quantum operations discussed in Subsect. "A Finite Universal Gate Set".

THE QUANTUM CIRCUIT DISTINGUISHABILITY PROBLEM

Input: Quantum circuits Q_0 and Q_1, both taking n input qubits and producing m output qubits.
Yes: $\delta(Q_0, Q_1) \geq 2/3$.
No: $\delta(Q_0, Q_1) \leq 1/3$.

Theorem 13 *The quantum circuit distinguishability problem is QIP-complete with respect to Karp reductions.*

Zero-Knowledge Quantum Interactive Proofs

Interactive proof systems can sometimes be made to have a cryptographically motivated property known as the *zero-knowledge* property [56]. Informally speaking, an interactive proof system is zero-knowledge if a verifier "learns nothing" from an interaction with the prover on an input $x \in A_{\mathrm{yes}}$, beyond the fact that it is indeed the case that $x \in A_{\mathrm{yes}}$. This should hold even if the verifier deviates from the actions prescribed to it by the interactive proof being considered. At first this notion seems paradoxical, but nevertheless there are many interesting examples of such proof systems. For an introductory survey on zero-knowledge, see [93].

Several variants of zero-knowledge are often studied in the classical setting that differ in the particular way that the notion of "learning nothing" is formalized. This article will only consider *statistical* zero-knowledge, which is the variant of zero-knowledge that is most easily adapted to the quantum setting.

Suppose that $A = (A_{\mathrm{yes}}, A_{\mathrm{no}})$ is a promise problem, and (V, P) is a quantum interactive proof system for A.

By this it is meant that V is the *honest verifier* and P is the *honest prover* that behave precisely as the proof system specifies. Whether or not this proof system possesses the zero-knowledge property depends on the characteristics of interactions between a *cheating verifier* V' with the honest prover P on inputs $x \in A_{\text{yes}}$. Figure 12 illustrates such an interaction, wherein it is viewed that the cheating verifier V' is using the interaction with P on input x to compute some quantum operation Φ_x. Informally speaking, the input to this operation represents the verifier's state of knowledge before the protocol is run, while the output represents the verifier's state of knowledge after.

Now, the proof system (V, P) is said to be *quantum statistical zero-knowledge* if, for any choice of a polynomial-time cheating verifier V', the quantum operation Φ_x can be efficiently approximated for all $x \in A_{\text{yes}}$. More precisely, the assumption that V' is described by polynomial-time generated quantum circuits must imply that there exists a polynomial-time generated family $\{Q_x : x \in \Sigma^*\}$ of quantum circuits for which $\delta(\Phi_x, Q_x)$ is negligible for every $x \in A_{\text{yes}}$. Intuitively this definition captures the notion of learning nothing—for anything that the cheating verifier could compute in polynomial time with the help of the prover could equally well have been computed in polynomial time without the prover's help.

The class of promise problems having statistical zero-knowledge quantum interactive proofs is denoted QSZK.

QSZK A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in QSZK if and only if it has a statistical zero-knowledge quantum interactive proof system.

Although it has not been proved, it is reasonable to conjecture that QSZK is properly contained in QIP; for the zero-knowledge property seems to be quite restrictive. Indeed, it was only recently established that there exist nontrivial quantum interactive proof systems that are statistical zero-knowledge [100]. The following facts [97,100] are among those known about QSZK.

1. Statistical zero-knowledge quantum interactive proof systems can be parallelized to two messages. It follows that QSZK $\subseteq$ QIP(2).
2. The class QSZK is closed under complementation: a given promise problem A has a quantum statistical zero-knowledge proof if and only if the same is true for the problem obtained by exchanging the yes- and no-instances of A.
3. Statistical zero-knowledge quantum interactive proof systems can be simulated in polynomial space: QSZK $\subseteq$ PSPACE.

Classical analogues to the first and second facts in this list were shown first [86]. (The classical analogue to the third fact is SZK $\subseteq$ PSPACE, which follows trivially from IP $\subseteq$ PSPACE.) A key step toward proving the above properties (which is also similar to the classical case) is to establish that the following promise problem is complete for QSZK. The problem is a restricted version of the QIP-complete quantum circuit distinguishability problem.

THE QUANTUM STATE DISTINGUISHABILITY PROBLEM

Input: Quantum circuits Q_0 and Q_1, both taking no input qubits and producing m output qubits. Let ρ_0 and ρ_1 be the density matrices corresponding to the outputs of these circuits.

Yes: $\delta(\rho_0, \rho_1) \geq 2/3$.

No: $\delta(\rho_0, \rho_1) \leq 1/3$.

Theorem 14 *The quantum state distinguishability problem is QSZK-complete with respect to Karp reductions.*

A quantum analogue of a different SZK-complete problem known as the *entropy difference problem* [53] has recently been shown to be complete for QSZK [16].

Multiple-Prover Quantum Interactive Proofs

Multiple-prover interactive proof systems are variants of interactive proofs where a verifier interacts with two or more provers that are not able to communicate with one another during the course of the interaction. Classical

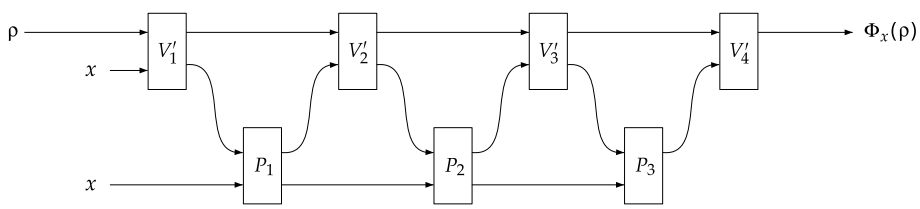

Quantum Computational Complexity, Figure 12
A cheating verifier V' performs a quantum operation Φ_x with the unintentional help of the honest prover

multiple-prover interactive proofs are extremely powerful: the class of promise problems having multiple-prover interactive proofs (denoted MIP) coincides with NEXP [21]. This is true even for two-prover interactive proofs wherein the verifier exchanges just one round of communication with each prover in parallel [46]. The key to the power of multiple-prover interactive proofs is the inability of the provers to communicate during the execution of the proof system. Similar to a detective interrogating two suspects in separate rooms in a police station, the verifier can ask questions of the provers that require strongly correlated answers, limiting the ability of cheating provers to convince the verifier of a false statement.

The identification of MIP with NEXP assumes a completely classical description of the provers. Quantum information, however, delivers a surprising twist for this model: even when the verifier is classical, an entangled quantum state shared between two provers can allow for non-classical correlations between their answers to the verifier's questions. This phenomenon is better known as a Bell-inequality violation [24] in the quantum physics literature. Indeed, there exist two-prover interactive proof systems that are sound against classical provers, but can be cheated by entangled quantum provers [36].

MIP* A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in MIP* if and only if there exists a multiple-prover interactive proof system for A wherein the verifier is classical and the provers may share an arbitrary entangled state.

One may also consider fully quantum variants of multiple-prover interactive proofs, which were first studied by Kobayashi and Matsumoto [70].

QMIP A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in QMIP if and only if there exists a multiple-prover quantum interactive proof system for A.

Various refinements on these classes have been studied, where parameters such as the number of provers, number of rounds of communication, completeness and soundness probabilities, and bounds on the amount of entanglement shared between provers are taken into account.

The results proved in both [36] and [70] support the claim that it is entanglement shared between provers that is the key issue in multiple-prover quantum interactive proofs. Both models are equivalent in power to MIP when provers are forbidden to share entanglement before the proof system is executed.

At the time of the writing of this article, research into these complexity classes and general properties of multiple-prover quantum interactive proofs is highly active and has led to several interesting results (such as [37,62,63], among others). Despite this effort, little can be said at this time about the relationship among the above classes and other known complexity classes. For instance, only the trivial lower bounds PSPACE $\subseteq$ MIP* and QIP $\subseteq$ QMIP, and no good upper bounds, are known. It has not even been ruled out that non-recursive languages could have multiple-prover quantum interactive proof systems. These difficulties seem to stem from two issues: (i) no bounds are known for the size of entangled states needed for provers to perform optimally, or nearly optimally, in an interactive proof, and (ii) the possible correlations that can be induced with entanglement are not well-understood.

One highly restricted variant of MIP* that has been studied, along with its unentangled counterpart, is as follows.

⊕MIP* A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in ⊕MIP* if and only if there exists a one-round two-prover interactive proof system for A wherein the provers each send a single bit to the verifier, and the verifier's decision to accept or reject is determined by the questions asked along with the XOR of these bits. The verifier is classical and the provers are quantum and share an arbitrary entangled state.

⊕ MIP This class is similar to ⊕MIP*, except that the provers may not share entanglement (and therefore can be assumed to be classical without loss of generality).

It holds that ⊕MIP = NEXP for some choices of completeness and soundness probabilities [25,58]. On the other hand, it has been proved [101] that ⊕MIP* $\subseteq$ QIP(2) and therefore ⊕MIP* $\subseteq$ EXP.

Other Variants of Quantum Interactive Proofs

Other variants of quantum interactive proof systems have been studied, including *public-coin* quantum interactive proofs and quantum interactive proofs with *competing provers*.

Public-Coin Quantum Interactive Proofs Public-coin interactive proof systems are a variant of interactive proofs wherein the verifier performs no computations until all messages with the prover have been exchanged. In place of the verifier's messages are sequences of random bits, visible to both the prover and verifier. Such interactive proof

systems are typically called *Arthur–Merlin games* [17,19], and the verifier and prover are called Arthur and Merlin, respectively, in this setting.

Quantum variants of such proof systems and their corresponding classes are easily defined. For instance, QAM is defined as follows [76].

QAM A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in QAM if and only if it has a quantum interactive proof system of the following restricted form. Arthur uniformly chooses some polynomial-bounded number of classical random bits, and sends a copy of these bits to Merlin. Merlin responds with a quantum state on a polynomial-bounded number of qubits. Arthur then performs a polynomial-time quantum computation on the input, the random bits, and the state sent by Merlin to determine acceptance or rejection.

It is clear that QAM $\subseteq$ QIP(2), but unlike the classical case [54] equality is not known in the quantum setting. It is straightforward to prove the upper bound QAM $\subseteq$ PSPACE, while containment QIP(2) $\subseteq$ PSPACE is not known.

The class QMAM may be defined through a similar analogy with classical Arthur–Merlin games. Here, Merlin sends the first message, Arthur responds with a selection of random bits, and then Merlin sends a second message.

QMAM A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in QMAM if and only if it has a quantum interactive proof system of the following restricted form. Merlin sends a quantum state on a polynomial-bounded number of qubits to Arthur. Without performing any computations, Arthur uniformly chooses some polynomial-bounded number of classical random bits, and sends a copy of these bits to Merlin. Merlin responds with a second quantum state. Arthur then performs a polynomial-time quantum computation on the input, the random bits, and the two states sent by Merlin in order to determine acceptance or rejection.

Even when Arthur is restricted to a single random bit, this class has the full power of QIP [76].

Theorem 15 QMAM = QIP.

Quantum Interactive Proofs with Competing Provers Another variant of interactive proof systems that has been studied is one where a verifier interacts with two *competing* provers, sometimes called the *yes-prover* and the *no-prover*. Unlike ordinary two-prover interactive proof systems, it is now assumed that the provers have conflicting goals: the yes-prover wants to convince the verifier that a given input string is a yes-instance of the problem being considered, while the no-prover wants to convince the verifier that the input is a no-instance. The verifier is sometimes called the *referee* in this setting, given that interactive proof systems of this form are naturally modeled as competitive games between the two provers. Two complexity classes based on interactive proofs with competing provers are the following.

RG A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in RG (short for *refereed games*) if and only if it has a classical interactive proof system with two competing provers. The completeness and soundness conditions for such a proof system are replaced by the following conditions:

1. For every $x \in A_{\text{yes}}$, there exists a yes-prover P_{yes} that convinces the referee to *accept* with probability at least 2/3, regardless of the strategy employed by the no-prover P_{no}.
2. For every $x \in A_{\text{no}}$, there exists a no-prover P_{no} that convinces the referee to *reject* with probability at least 2/3, regardless of the strategy employed by the yes-prover P_{yes}.

QRG A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in QRG (*quantum refereed games*) if and only if it has a quantum interactive proof system with two competing provers. The completeness and soundness conditions for such a proof system are analogous to RG.

Classical refereed games have the computational power of deterministic exponential time [45], and the same is true in the quantum setting [57]: RG = QRG = EXP. The containment EXP $\subseteq$ RG represents an application of the arithmetization technique, while QRG $\subseteq$ EXP exemplifies the power of semidefinite programming.

Other Selected Notions in Quantum Complexity

In this section of this article, a few other topics in quantum computational complexity theory are surveyed that do not fall under the headings of the previous sections. While incomplete, this selection should provide the reader with some sense for the topics that have been studied in quantum complexity.

Quantum Advice

Quantum advice is a formal abstraction that addresses this question: how powerful is quantum software? More precisely, let us suppose that some polynomial-time generated

family of quantum circuits $Q = \{Q_n : n \in \mathbb{N}\}$ is given. Rather than assuming that each circuit Q_n takes n input qubits as for the class BQP, however, it is now assumed that Q_n takes $n + p(n)$ qubits for p some polynomial-bounded function: in addition to a given input $x \in \Sigma^n$, the circuit Q_n will take an *advice state* ρ_n on $p(n)$ qubits. This advice state may be viewed as pre-loaded quantum software for a quantum computer. The advice state may depend on the input length n, but not on the particular input $x \in \Sigma^n$. Similar to a quantum proof, the difficulty of preparing this state is ignored; but unlike a quantum proof the advice state is completely trusted. The quantum complexity class BQP/qpoly is now defined to be the class of promise problems that are solved by polynomial-time quantum algorithms with quantum advice in the natural way. The first published paper on quantum advice was [80], while the definitions and most of the results discussed in this section are due to Aaronson [2].

BQP/qpoly A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in BQP/qpoly if and only if there exists a polynomial-bounded function p, a collection of quantum states $\{\rho_n : n \in \mathbb{N}\}$ where each ρ_n is a $p(n)$-qubit state, and a polynomial-time generated family of quantum circuits $Q = \{Q_n : n \in \mathbb{N}\}$ with the following properties. For all $x \in A_{\text{yes}}$, Q accepts $(x, \rho_{|x|})$ with probability at least 2/3; and for all $x \in A_{\text{no}}$, Q accepts $(x, \rho_{|x|})$ with probability at most 1/3.

Similar to BQP without advice, it is straightforward to shown that the constants 2/3 and 1/3 in this definition can be replaced by a wide range of functions $a, b : \mathbb{N} \to [0, 1]$.

As the notation suggests, BQP/qpoly is a quantum analogue of the class P/poly. This analogy may be used as the basis for several relevant points about BQP/qpoly, quantum advice, and their relationship to classical complexity classes and notions.

1. As is well-known, P/poly may be defined either in a manner similar to the above definition for BQP/qpoly, or more simply as the class of promise problems solvable by polynomial-size Boolean circuit families with no uniformity restrictions. Based on similar ideas, the class BQP/qpoly does not change if the circuit family $\{Q_n : n \in \mathbb{N}\}$ is taken to be an arbitrary polynomial-size circuit family without uniformity constraints.
2. There is a good argument to be made that quantum advice is better viewed as a quantum analogue of *randomized advice* rather than *deterministic advice*. That is, BQP/qpoly can equally well be viewed as a quantum analogue of the (suitably defined) complexity class BPP/rpoly. It happens to be the case, however, that BPP/rpoly = P/poly. (The situation is rather different for *logarithmic-length advice*, where randomized advice is strictly more powerful than ordinary deterministic advice.)
3. Any combination of quantum, randomized, or deterministic advice with quantum, randomized, or deterministic circuits can be considered. This leads to classes such as BQP/rpoly, BQP/poly, P/qpoly, P/rpoly, and so on. (The only reasonable interpretation of BPP/qpoly and P/qpoly is that classical circuits effectively measure quantum states in the standard basis the instant that they touch them.)
   At most three distinct classes among these possibilities arise: BQP/qpoly, BQP/poly, and P/poly. This is because BQP/rpoly = BQP/poly and

$$\begin{aligned}\text{BPP/qpoly} &= \text{BPP/rpoly} = \text{BPP/poly} \\ &= \text{P/qpoly} = \text{P/rpoly} = \text{P/poly}\,.\end{aligned}$$

   The principle behind these equalities is that nonuniformity is stronger than randomness [6].

The following theorem places an upper-bound on the power of polynomial-time quantum algorithms with quantum advice [2].

Theorem 16 BQP/qpoly $\subseteq$ PP/poly.

Although in all likelihood the class PP/poly is enormous, containing many interesting problems that one can have little hope of being able to solve in practice, the upper-bound represented by this theorem is far from obvious. The most important thing to notice is that the power of quantum advice (to a BQP machine) is simulated by deterministic advice (to a PP machine). This means that no matter how complex, a polynomial-size quantum advice state can never encode more information accessible to a polynomial-time quantum computer than a polynomial-length string can, albeit to an unbounded error probabilistic machine.

Quantum advice has also been considered for other quantum complexity classes such as QMA and QIP. For instance, the following bound is known on the power of QMA with quantum advice [3].

Theorem 17 QMA/qpoly $\subseteq$ PSPACE/poly.

Generally speaking, the study of both quantum and randomized advice is reasonably described as a sticky business when unbounded error machines or interactive protocols are involved. In these cases, complexity classes defined by both quantum and randomized advice can be

highly non-intuitive and dependent on specific interpretations of models. For example, Raz [83] has shown that both QIP/qpoly and IP/rpoly contain all languages, provided that the advice is not accessible to the prover. Likewise, Aaronson [2] has observed that both PP/rpoly and PQP/qpoly contain all languages. These results are quite peculiar, given that significantly more powerful models can become strictly less powerful in the presence of quantum or randomized advice. For instance, even a Turing machine running in exponential space cannot decide all languages with bounded error given randomized advice.

Space-Bounded Quantum Computation

The quantum complexity classes that have been discussed thus far in this article are based on the abstraction that efficient quantum computations are those that can be performed in polynomial time. Quantum complexity classes may also be defined by bounds on space rather than time, but here a different computational model is required to reasonably compare with classical models of space-bounded computation. One simple choice of a suitable model is a hybrid between quantum circuits and classical Turing machines—and although this model is different from the variants of quantum Turing machines that were originally studied in the theory of quantum computing [30,40], the term *quantum Turing machine* (QTM for short) is nevertheless appropriate. There should be no confusion given that no other models of quantum Turing machines are discussed in this article.

Figure 13 illustrates a quantum Turing machine. A quantum Turing machine has a read-only classical input tape, a classical work tape, and a quantum tape consisting of an infinite sequence of qubits each initialized to the zero-state. Three tape heads scan the quantum tape, allowing quantum operations to be performed on the corresponding qubits. (It would be sufficient to have just two, but allowing three tape heads parallels the choice of a universal gate set that allows three-qubit operations.) A single step of a QTM's computation may involve ordinary moves by the classical parts of the machine and quantum operations on the quantum tape: Toffoli gates, Hadamard gates, phase-shift gates, or single-qubit measurements in the computational basis.

The running time of a quantum Turing machine is defined as for ordinary Turing machines. The space used by such a machine is the number of squares on the classical work tape plus the number of qubits on the quantum tape that are ever visited by one of the tape heads. Similar to the classical case, the input tape does not contribute to the space used by the machine because it is a read-only tape.

The following complexity classes are examples of classes that can be defined using this model.

BQL A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in BQL (bounded-error quantum logarithmic space) if and only if there exists a quantum Turing machine M running in polynomial time and logarithmic space that accepts every string $x \in A_{\text{yes}}$ with probability at least 2/3 and accepts every string $x \in A_{\text{no}}$ with probability at most 1/3.

PQL A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in PQL (unbounded-error quantum logarithmic space) if and only if there exists a quantum Turing machine M running in polynomial time and logarithmic space that accepts every string $x \in A_{\text{yes}}$ with probability strictly greater than 1/2 and accepts every string $x \in A_{\text{no}}$ with probability at most 1/2.

BQPSPACE A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in BQPSPACE (bounded-error quantum polynomial space) if and only if there exists a quantum Turing machine M running in polynomial space that accepts every string $x \in A_{\text{yes}}$ with probability at least 2/3 and accepts every string $x \in A_{\text{no}}$ with probability at most 1/3.

PQPSPACE A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in PQPSPACE (unbounded-error quantum polynomial space) if and only if there exists a quantum Turing machine M running in polynomial space that accepts every string $x \in A_{\text{yes}}$ with probability strictly greater than 1/2 and accepts every string $x \in A_{\text{no}}$ with probability at most 1/2.

Unlike polynomial-time computations, it is known that quantum information does not give a significant increase in computational power in the space-bounded case [95,98].

Theorem 18 *The following relationships hold.*

1. BQL $\subseteq$ PQL $=$ PL.
2. BQPSPACE $=$ PQPSPACE $=$ PSPACE.

The key relationship in the above theorem, from the perspective of quantum complexity, is PQL $\subseteq$ PL, which can be shown using space-bounded counting complexity. In particular, the proof relies on a theory of GapL functions [12] that parallels the theory of GapP functions, and allows for a variety of matrix computations to be performed in PL.

The above theorem, together with the containment PL $\subseteq$ NC [32], implies that both BQL and PQL are con-

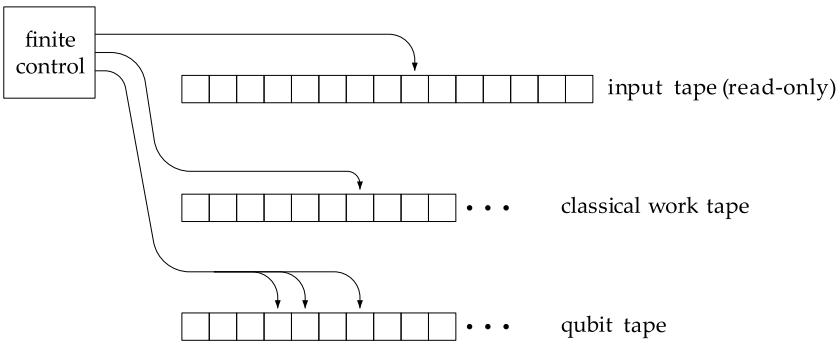

Quantum Computational Complexity, Figure 13
A quantum Turing machine. This variant of quantum Turing machine is classical with the exception of a quantum work tape, each square of which contains a qubit. Quantum operations and measurements can be performed on the qubits scanned by the quantum tape heads

tained in NC. An interpretation of this fact is that logarithmic-space quantum computations can be very efficiently simulated in parallel.

Bounded-Depth Quantum Circuits

The *depth* of a classical or quantum circuit is the maximum number of gates encountered on any path from an input bit or qubit to an output bit or qubit in the circuit. One may reasonably think of circuit depth as the parallel running time, or the number of time units needed to apply a circuit when operations may be parallelized in any way that respects the topology of the circuit.

The following complexity class, first defined by Moore and Nilsson [77], represent bounded-error quantum variants of the class NC.

BQNC A promise problem $A = (A_{\text{yes}}, A_{\text{no}})$ is in BQNC (bounded-error quantum NC) if and only if there exists a logarithmic-space generated family $\{Q_n : n \in \mathbb{N}\}$ of poly-logarithmic depth quantum circuits, where each circuit Q_n takes n input qubits and produces one output qubit, such that $\Pr[Q \text{ accepts } x] \geq 2/3$ for all $x \in A_{\text{yes}}$, and $\Pr[Q \text{ accepts } x] \leq 1/3$ for all $x \in A_{\text{no}}$.

Many other complexity classes based on bounded-depth quantum circuits have been studied as well. The survey of Bera, Green, and Homer [28] discusses several examples.

In the classical case there is a very close relationship between space-bounded and depth-bounded computation [31,32]. This close relationship is based on two main ideas: the first is that space-bounded computations can be simulated efficiently by bounded-depth circuits using parallel algorithms for matrix computations, and the second is that bounded-depth Boolean circuits can be efficiently simulated by space-bounded computations via depth-first traversals of the circuit to be simulated.

For quantum computation this close relationship is not known to exist. One direction indeed holds, as was discussed in the previous subsection: space-bounded quantum computations can be efficiently simulated by depth-bounded circuits. The other direction, which is an efficient space-bounded simulation of bounded-depth quantum circuits, is not known to hold and is arguably quite unlikely. Informally speaking, bounded-depth quantum circuits are computationally powerful, whereas space-bounded quantum Turing machines are not. Three facts that support this claim are as follows.

1. Computing the acceptance probability for even constant-depth quantum circuits is as hard as computing acceptance probabilities for arbitrary polynomial-size quantum circuits [48].
2. Shor's factoring algorithm can be implemented by quantum circuits having logarithmic-depth, along with classical pre- and post-processing [35].
3. The quantum circuit distinguishability problem remains complete for QIP when restricted to logarithmic-depth quantum circuits [84].

It is reasonable to conjecture that BQNC is incomparable with BPP and properly contained in BQP.

Future Directions

There are many possible future directions for research in quantum computational complexity theory. The following list suggests just a few of many possibilities.

1. The power of multiple-prover quantum interactive proofs, for both quantum and classical verifiers, is

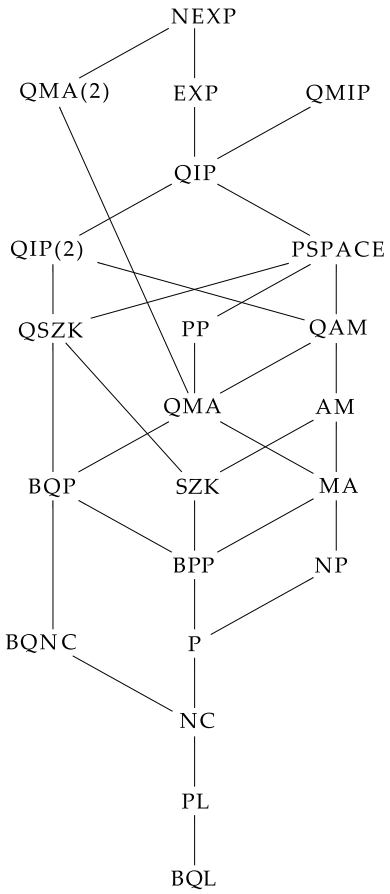

Quantum Computational Complexity, Figure 14
A diagram of inclusions among most of the complexity classes discussed in this article

very poorly understood. In particular: (i) no interesting upper-bounds are known for either MIP* or QMIP, and (ii) neither the containment NEXP $\subseteq$ MIP* nor NEXP $\subseteq$ QMIP is known to hold.

2. The containment NP $\subseteq$ BQP would be a very powerful incentive to build a quantum computer, to say the least. While there is little reason to hope that this containment holds, there is at the same time little evidence against it aside from the fact that it fails relative to an oracle [27]. A better understanding of the relationship between BQP and NP, including possible consequences of one being contained in the other, is an important direction for further research in quantum complexity.
3. Along similar lines to the previous item, an understanding of the relationship between BQP and the polynomial-time hierarchy has remained elusive. It is not known whether BQP is contained in the polynomial-time hierarchy, whether there is an oracle relative to which this is false, or even whether there is an oracle relative to which BQP is not contained in AM.
4. Interest in complexity classes is ultimately derived from the problems that they contain. An important future direction in quantum complexity theory is to prove the inclusion of interesting computational problems in the quantum complexity classes for which this is possible. Of the many problems that have been considered, one of the most perplexing from the perspective of quantum complexity is the graph isomorphism problem. It is a long-standing open problem whether the graph isomorphism problem is in BQP. A seemingly easier task than showing the inclusion of this problem in BQP is proving it is in co-QMA; but even this problem has remained unresolved.

Bibliography

1. Aaronson S (2002) Quantum lower bound for the collision problem. In: Proceedings of the 35th Annual ACM Symposium on Theory of Computing. ACM Press, New York
2. Aaronson S (2005) Limitations of quantum advice and one-way communication. Theory Comput 1:1–28
3. Aaronson S (2006) QMA/qpoly is contained in PSPACE/poly: de-Merlinizing quantum protocols. In: Proceedings of the 21st Annual IEEE Conference on Computational Complexity. IEEE Ceomputer Society Press, Los Alamitos pp 261–273
4. Aaronson S, Kuperberg G (2007) Quantum versus classical proofs and advice. Theory Comput 3:129–157
5. Aaronson S, Shi Y (2004) Quantum lower bounds for the collision and the element distinctness problems. J ACM 51(4):595–605
6. Adleman L (1978) Two theorems on random polynomial time. In: Proceeding of the 19th Annual IEEE Symposium on Foundations of Computer Science. IEEE Ceomputer Society Press, Los Alamitos pp 75–83
7. Adleman L, DeMarrais J, Huang M (1997) Quantum computability. SIAM J Comput 26(5):1524–1540
8. Aharonov D, Naveh T (2002) Quantum NP – a survey. Available as arXiv.org e-Print quant-ph/0210077
9. Aharonov D, Kitaev A, Nisan N (1998) Quantum circuits with mixed states. In: Proceedings of the 30th Annual ACM Symposium on Theory of Computing. ACM Press, New York, pp 20–30
10. Aharonov D, Gottesman D, Irani S, Kempe J (2007) The power of quantum systems on a line. In: Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science. IEEE Ceomputer Society Press, Los Alamitos pp 373–383
11. Aharonov D, van Dam W, Kempe J, Landau Z, Lloyd S, Regev O (2007) Adiabatic quantum computation is equivalent to standard quantum computation. SIAM J Comput 37(1): 166–194
12. Allender E, Ogihara M (1996) Relationships among PL, #L, and the determinant. RAIRO – Theor Inf Appl 30:1–21
13. Arora S, Barak B (2006) Complexity Theory: A Modern Approach. Web draft available at http://www.cs.princeton.edu/theory/complexity/

14. Arora S, Safra S (1998) Probabilistic checking of proofs: a new characterization of NP. J ACM 45(1):70–122
15. Arora S, Lund C, Motwani R, Sudan M, Szegedy M (1998) Proof verification and the hardness of approximation problems. J ACM 45(3):501–555
16. Aroya A-B, Shma A-T (2007) Quantum expanders and the quantum entropy difference problem. Available as arXiv.org e-print quant-ph/0702129
17. Babai L (1985) Trading group theory for randomness. In: Proceedings of the Seventeenth Annual ACM Symposium on Theory of Computing. ACM Press, New York pp 421–429
18. Babai L (1992) Bounded round interactive proofs in finite groups. SIAM J Discret Math 5(1):88–111
19. Babai L, Moran S (1988) Arthur-Merlin games: a randomized proof system, and a hierarchy of complexity classes. J Comput Syst Sci 36(2):254–276
20. Babai L, Szemerédi E (1984) On the complexity of matrix group problems I. In: Proceedings of the 25th Annual IEEE Symposium on Foundations of Computer Science. IEEE Ceomputer Society Press, Los Alamitos pp 229–240
21. Babai L, Fortnow L, Lund C (1991) Non-deterministic exponential time has two-prover interactive protocols. Comput Complex 1(1):3–40
22. Beigel R, Reingold N, Spielman D (1995) PP is closed under intersection. J Comput Syst Sci 50(2):191–202
23. Beigi S, Shor P (2007) On the complexity of computing zero-error and Holevo capacity of quantum channels. Available as arXiv.org e-Print 0709.2090
24. Bell J (1964) On the Einstein-Podolsky-Rosen paradox. Phys 1(3):195–200
25. Bellare M, Goldreich O, Sudan M (1998) Free bits, PCPs, and non-approximability —towards tight results. SIAM J Comput 27(3):804–915
26. Bennett CH (1973) Logical reversibility of computation. IBM J Res Dev 17:525–532
27. Bennett CH, Bernstein E, Brassard G, Vazirani U (1997) Strengths and weaknesses of quantum computing. SIAM J Comput 26(5):1510–1523
28. Bera D, Green F, Homer S (2007) Small depth quantum circuits. ACM SIGACT News 38(2):35–50
29. Bernstein E, Vazirani U (1993) Quantum complexity theory (preliminary abstract). In: Proceedings of the 25th Annual ACM Symposium on Theory of Computing. ACM Press, New York pp 11–20
30. Bernstein E, Vazirani U (1997) Quantum complexity theory. SIAM J Comput 26(5):1411–1473
31. Borodin A (1977) On relating time and space to size and depth. SIAM J Comput 6:733–744
32. Borodin A, Cook S, Pippenger N (1983) Parallel computation for well-endowed rings and space-bounded probabilistic machines. Inf Control 58:113–136
33. Brassard G (2003) Quantum communication complexity. Found Phys 33(11):1593–1616
34. Cleve R (2000) An introduction to quantum complexity theory. In: Macchiavello C, Palma GM, Zeilinger A (eds) Collected Papers on Quantum Computation and Quantum Information Theory. World Scientific. Singapore pp 103–127
35. Cleve R, Watrous J (2000) Fast parallel circuits for the quantum Fourier transform. In: Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science. pp 526–536
36. Cleve R, Høyer P, Toner B, Watrous J (2004) Consequences and limits of nonlocal strategies. In: Proceedings of the 19th Annual IEEE Conference on Computational Complexity. pp 236–249
37. Cleve R, Slofstra W, Unger F, Upadhyay S (2007) Perfect parallel repetition theorem for quantum XOR proof systems. In: Proceedings of the 22nd Annual IEEE Conference on Computational Complexity. pp 109–114
38. Cook S (1972) The complexity of theorem proving procedures. In: Proceedings of the Third Annual ACM Symposium on Theory of Computing. ACM Press, New York pp 151–158
39. de Wolf R (2002) Quantum communication and complexity. Theor Comput Sci 287(1):337–353
40. Deutsch D (1985) Quantum theory, the Church–Turing principle and the universal quantum computer. Proc Roy Soc Lond A 400:97–117
41. Deutsch D (1989) Quantum computational networks. Proc Roy Soc Lond A 425:73–90
42. Dinur I (2007) The PCP theorem by gap amplification. J ACM 54(3)
43. Du D-Z, Ko K-I (2000) Theory of Computational Complexity. Wiley, New York
44. Even S, Selman A, Yacobi Y (1984) The complexity of promise problems with applications to public-key cryptography. Inf Control 61:159–173
45. Feige U, Kilian J (1997) Making games short. In: Proceedings of the 29th Annual ACM Symposium on Theory of Computing. ACM Press, New York pp 506–516
46. Feige U, Lovász L (1992) Two-prover one-round proof systems: their power and their problems. In: Proceedings of the 24th Annual ACM Symposium on Theory of Computing. ACM Press, New York pp 733–744
47. Fenner S, Fortnow L, Kurtz S (1994) Gap-definable counting classes. J Comput Syst Sci 48:116–148
48. Fenner S, Green F Homer S, Zhang Y (2005) Bounds on the power of constant-depth quantum circuits. In: Proceedings of the 15th International Symposium on Fundamentals of Computation Theory. Lect Notes Comput Sci 3623:44–55
49. Feynman R (1983) Simulating physics with computers. Int J Theor Phys 21(6/7):467–488
50. Fortnow L (1997) Counting complexity. In: Hemaspaandra L, Selman A (eds) Complexity Theory Retrospective II. Springer, New York pp 81–107
51. Fortnow L, Rogers J (1999) Complexity limitations on quantum computation. J Comput Syst Sci 59(2):240–252
52. Goldreich O (2005) On promise problems (a survey in memory of Shimon Even [1935–2004]). Electronic Colloquium on Computational Complexity; Report TR05-018
53. Goldreich O, Vadhan S (1999) Comparing entropies in statistical zero-knowledge with applications to the structure of SZK. In: Proceedings of the 14th Annual IEEE Conference on Computational Complexity. pp 54–73
54. Goldwasser S, Sipser M (1989) Private coins versus public coins in interactive proof systems. In: Micali S (ed) Randomness and Computation, vol 5 of Advances in Computing Research. JAI Press, Greenwich, Conn pp 73–90
55. Goldwasser S, Micali S, Rackoff C (1985) The knowledge complexity of interactive proof systems. In: Proceedings of the Seventeenth Annual ACM Symposium on Theory of Computing. ACM Press, New York pp 291–304
56. Goldwasser S, Micali S, Rackoff C (1989) The knowledge

complexity of interactive proof systems. SIAM J Comput 18(1):186–208
57. Gutoski G, Watrous J (2007) Toward a general theory of quantum games. In: Proceedings of the 39th ACM Symposium on Theory of Computing. ACM Press, New York pp 565–574
58. Håstad J (2001) Some optimal inapproximability results. J ACM 48(4):798–859
59. Janzing D, Wocjan P, Beth T (2005) Non-identity-check is QMA-complete. Int J Quantum Inf 3(2):463–473
60. Kaye P, Laflamme R, Mosca M (2007) An introduction to quantum computing. Oxford University Press, Oxford
61. Kempe J, Kitaev A, Regev O (2006) The complexity of the local Hamiltonian problem. SIAM J Comput 35(5):1070–1097
62. Kempe J, Kobayashi H, Matsumoto K, Toner B, Vidick T (2007) Entangled games are hard to approximate. Proceedings ot the 49th Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Society Press, Los Alamitos 2008
63. Kempe J, Regev O, Toner B (2007) The unique games conjecture with entangled provers is false. Proceedings ot the 49th Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Society Press, Los Alamitos 2008
64. Kitaev A (1997) Quantum computations: algorithms and error correction. Russ Math Surv 52(6):1191–1249
65. Kitaev A (1999) Quantum NP. Talk at AQIP'99: Second Workshop on Algorithms in Quantum Information Processing. DePaul University, Chicago
66. Kitaev A, Watrous J (2000) Parallelization, amplification, and exponential time simulation of quantum interactive proof system. In: Proceedings of the 32nd ACM Symposium on Theory of Computing. ACM Press, New York pp 608–617
67. Kitaev A, Shen A, Vyalyi M (2002) Classical and Quantum Computation. Graduate Studies in Mathematics, vol 47. American Mathematical Society, Providence
68. Knill E (1995) Approximation by quantum circuits. Technical Report LAUR-95-2225, Los Alamos National Laboratory. Available as arXiv.org e-Print quant-ph/9508006
69. Knill E (1996) Quantum randomness and nondeterminism. Technical Report LAUR-96-2186, Los Alamos National Laboratory. Available as arXiv.org e-Print quant-ph/9610012
70. Kobayashi H, Matsumoto K (2003) Quantum multi-prover interactive proof systems with limited prior entanglement. J Comput Syst Sci 66(3):429–450
71. Kobayashi H, Matsumoto K, Yamakami T (2003) Quantum Merlin–Arthur proof systems: Are multiple Merlins more helpful to Arthur? In: Proceedings of the 14th Annual International Symposium on Algorithms and Computation. Lecture Notes in Computer Science, vol 2906. Springer, Berlin
72. Levin L (1973) Universal search problems. Probl Inf Transm 9(3):265–266 (English translation)
73. Liu Y-K (2006) Consistency of local density matrices is QMA-complete. In: Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques. Springer, Berlin; Lect Notes Comput Sci 4110:438–449
74. Lloyd S (2000) Ultimate physical limits to computation. Nature 406:1047–1054
75. Lund C, Fortnow L, Karloff H, Nisan N (1992) Algebraic methods for interactive proof systems. J ACM 39(4):859–868
76. Marriott C, Watrous J (2005) Quantum Arthur-Merlin games. Comput Complex 14(2):122–152
77. Moore C, Nilsson M (2002) Parallel quantum computation and quantum codes. SIAM J Comput 31(3):799–815
78. Moore G (1965) Cramming more components onto integrated circuits. Electron 38(8):82–85
79. Nielsen MA, Chuang IL (2000) Quantum Computation and Quantum Information. Cambridge University Press, Cambride
80. Nishimura H, Yamakami T (2004) Polynomial time quantum computation with advice. Inf Process Lett 90(4):195–204
81. Oliveira R, Terhal B (2005) The complexity of quantum spin systems on a two-dimensional square lattice. Available as arXiv.org e-Print quant-ph/0504050
82. Papadimitriou C (1994) Computational Complexity. Addison-Wesley, Readig, Mass
83. Raz R (2005) Quantum information and the PCP theorem. In: 46th Annual IEEE Symposium on Foundations of Computer Science. pp 459–468
84. Rosgen B (2008) Distinguishing short quantum computations. Proceedings of the 25th Annual Symposium on theoretical Aspects of Computer Science, IBFI, Schloss Dagstuhl, pp 597–608
85. Rosgen B, Watrous J (2005) On the hardness of distinguishing mixed-state quantum computations. In: Proceedings of the 20th Annual Conference on Computational. pp 344–354
86. Sahai A, Vadhan S (2003) A complete promise problem for statistical zero-knowledge. J ACM 50(2):196–249
87. Shamir A (1992) IP = PSPACE. J ACM 39(4):869–877
88. Shor P (1994) Algorithms for quantum computation: discrete logarithms ande factoring. In: Proceedings of the 35th Annual IEEE Symposium on Foundations of Computer Science. pp 124–134
89. Shor P (1997) Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. SIAM J Comput 26(5):1484–1509
90. Stinespring WF (1955) Positive functions on C^*-algebras. Proc Am Math Soc 6(2):211–216
91. Toda S (1991) PP is as hard as the polynomial-time hierarchy. SIAM J Comput 20(5):865–887
92. Toffoli T (1980) Reversible computing. Technical Report MIT/LCS/TM-151, Laboratory for Computer Science, Massachusetts Institute of Technology
93. Vadhan S (2007) The complexity of zero knowledge. In: 27th International Conference on Foundations of Software Technology and Theoretical Computer Science. Lect Notes Comput Sci 4855:52–70
94. Valiant L (1979) The complexity of computing the permanent. Theor Comput Sci 8:189–201
95. Watrous J (1999) Space-bounded quantum complexity. J Comput Syst Sci 59(2):281–326
96. Watrous J (2000) Succinct quantum proofs for properties of finite groups. In: Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science. pp 537–546
97. Watrous J (2002) Limits on the power of quantum statistical zero-knowledge. In: Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science. pp 459–468
98. Watrous J (2003) On the complexity of simulating space-bounded quantum computations. Comput Complex 12: 48–84
99. Watrous J (2003) PSPACE has constant-round quantum interactive proof systems. Theor Comput Sci 292(3):575–588
100. Watrous J (2006) Zero-knowledge against quantum attacks.

In: Proceedings of the 38th ACM Symposium on Theory of Computing. ACM Press, New York pp 296–305
101. Wehner S (2006) Entanglement in interactive proof systems with binary answers. In: Proceedings of the 23rd Annual Symposium on Theoretical Aspects of Computer Science. Lect Notes Comput Sci 3884:162–171

Quantum Computing

Viv Kendon
School of Physics and Astronomy, University of Leeds, Leeds, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Digital Quantum Computing
Unconventional Extensions
Future Directions
Bibliography

Glossary

Bit A two state classical system used to represent a binary digit, zero or one.

Bose–Einstein particles Integer spin quantum particles like to be together: any number can occupy the same quantum state.

Classical computing What we can compute within the laws of classical physics.

Error correction In a quantum context, fixing errors without disturbing the quantum superposition.

Entanglement Quantum states can be more highly correlated than classical systems: the extra correlations are known as entanglement.

Fermi–Dirac particles Half-integer spin quantum particles like to be alone: only one such particle can occupy each quantum state.

Quantum communications Using quantum mechanics can gain an advantage when transmitting information.

Quantum dense coding Classical bits can be encoded two for one into qubits for communications purposes.

Quantum key distribution Quantum mechanics allows for secure key distribution in the presence of eavesdroppers and noisy environments. The keys can then be used for encrypted communication.

Quantum teleportation A method to transmit an unknown quantum state using only classical communications plus shared entanglement.

Quantum computing Computation based on the laws of quantum mechanics for the allowed logical operations.

Qubit A two state quantum system such as the spin of an electron or the polarization of a photon. More complex quantum particles (such as atoms) can be used as qubits if just two of their available states are chosen to represent the qubit.

Qubus A quantum version of a computer bus, the fast communications linking memory and processing registers. Can be implemented using a coherent light source (such as a laser).

Scalable A computer architecture designed from modular units that can be efficiently expanded to an arbitrary size.

Squeezing With a pair of complementary quantum observables, making one uncertain so that the other can be measured more precisely.

Threshold result In the context of quantum computation, this results says that error correction can work if the error rate is low enough.

Tunneling Quantum particles can get through barriers that classical particles remain stuck behind. If the barrier is not infinitely high, there is a some probability for the quantum particle to be the other side of it even though it doesn't on average have enough energy to jump over the top.

Unitary operations How to control quantum systems while preserving quantum properties. Quantum systems evolving without any influence from environmental disturbance follow unitary dynamical evolution.

Definition of the Subject

Quantum computing is computing that follows the logic of quantum mechanics. The first part of this subject is well-specified after the hundred-odd years of development of quantum theory. Definitions of computing are fuzzier, and argued over by philosophers: for the purposes of this article we will leave such nuances aside in favor of exploring what is possible when the constraints of classical logic are put aside, but the limitations of the physical world are kept firmly in hand.

Quantum computing is not simply computing that involves quantum effects, that would be too broad to be useful. Since transitors and lasers exploit quantum properties of matter and light, most classical computers would thus be included. Yet it is worth remembering that definitions are never as clearcut as we would like. Figuring out which quantum systems can be simulated efficiently by classical computers is an active area of current research, and pin-

ning down the boundary between quantum and classical is the object of ongoing experimental investigation.

Introduction

This article on *quantum computing* is in the Unconventional Computing section of the encyclopedia. It is thus a somewhat distinctive view of quantum computing, adapted to the context in which you are likely to be reading it. It is self-contained, in that it introduces the concepts you will need to understand the contents, but those desiring more details and a broader perspective on the subject can refer to the introductory article ► Quantum Information Science, Introduction to and referenced articles therein.

To many, quantum computing is already unconventional computing, depending for its conception on the elusive properties of matter at the scale of atoms rather than the simple everyday experience of counting and deduction that underpins classical computation. Yet it is already an established field in which there is a standard approach, and the possibility to be unconventional within a quantum computing context. This article is thus divided into two parts, the first forming a lightening introduction to the standard digital version of quantum computing, and the second expanding into less well-charted territory beside and beyond.

In order to write about such an interdisciplinary subject for a broad audience, I have tried to keep everything at a level a non-specialist can follow, while not oversimplifying to the point of inaccuracy. Please be forgiving about the parts you could have written better yourself, and inquisitive about the parts that may have something new for you.

Digital Quantum Computing

Digital quantum computing developed a more or less parallel structure to classical digital computing, with a quantum version of each component. There are good reasons for this: many of the optimal features of classical digital computing carry over to quantum computing with little or no change. In order to explain why there is a fundamental advantage to using quantum logic for computation, this section will explain the components step by step and then provide examples of how the parts combine into the whole.

What Is a Qubit? (Quantum Mechanics for Dummies)

Quantum computing exploits the logic of quantum mechanics, which is notoriously tricky to understand. The mathematical structure is simple and linear, so treating it as "maths we happen to be able to build (we think)", is one option for getting to grips with it. Here I've chosen the pictorial intuition approach, with some cartoon qubits to help explain how they behave. Those already familiar with quantum mechanics can safely skip over this section. Those still bemused by the end of it may like to sample from the wide range of available textbooks to find an approach which suits them best.

Classical computers use *bits* as their basic unit, any classical system with two states can represent a single bit. Tossing a coin, with the resulting "heads" or "tails" is an example of a two state classical system. In computers, the two states are usually labeled "0" (zero) and "1" (one). Quantum computers use quantum bits, usually called *qubits*, as their basic unit, which are two state quantum systems, and the two states are also usually labeled zero and one. Examples of physical systems with two quantum states include electron spin, photon polarization and phosphorus nuclear spin. These have all been used in experiments aimed towards building a quantum computer. Those who know a bit of quantum mechanics should note that in a quantum computer we usually localize or otherwise separate individual qubits so they are distinguishable from each other, and there are thus no Fermi or Bose statistics to complicate the picture. We will revisit this aspect of quantum mechanics briefly in Subsect. "Topological Quantum Computing".

Here are a couple of cartoon qubits to help explain their quantum properties: . They are wriggly, fidgety little things, they move so fast you can't tell which way up they are just by looking. If you reach down to grab hold of one the only way you can get a grip is by an arm or a leg. First we have to choose our basis states, labeled $|0\rangle$ and $|1\rangle$. We write the zero and one in those funny brackets ("kets") to remind us they are quantum states rather than classical bits. To simplify notation, we will sometimes write the state of several qubits together inside one ket thus, $|0\rangle|0\rangle \equiv |00\rangle$. In terms of the cartoon qubits, this means we decided to label "grabbed by the arm" as the zero state and "grabbed by the leg" as the one state. The key quantum property is that qubits can fidget around anywhere in between the zero and one states, in what is called a superposition state, for example, $(|0\rangle + |1\rangle)/\sqrt{2}$ or $(|0\rangle - |1\rangle)/\sqrt{2}$. If you reach down and grab a qubit in one of these superposition states, you are equally like to grab an arm or a leg, indicat-

ing zero or one in a basis state. So you can't tell the qubit was in a superposition state by measuring it, the outcome is always a basis state. What's more, grabbing it by the arm hauls it onto its feet, and grabbing it by the leg shoves it into a handstand. So after measuring, what you found is what you now have, regardless of what it was before. Luckily, it is possible to do other things to qubits besides measuring them.

The most general state of one qubit can be written $\alpha|0\rangle + \beta|1\rangle$ with $|\alpha|^2 + |\beta|^2 = 1$, and α, β are complex numbers. The probability of measuring zero is given by $|\alpha|^2$ and the probability of measuring one is given by $|\beta|^2$. (Why complex numbers? Because it works conveniently. If you don't like complex numbers, there are ways to formulate quantum mechanics that avoid using them). Cartoon qubits with limbs don't quite capture all the subtle quantum effects, but they are a useful image to keep in mind if you are new to these concepts. This account also ignores all of the more fundamental issues to do with the interpretation of quantum mechanics, especially those generally referred to as "the quantum measurement problem".

Now consider two qubits (traditionally belonging to Alice and Bob) in the state $(|0\rangle_A|0\rangle_B + |1\rangle_A|1\rangle_B)/\sqrt{2}$. This is an example of an entangled state: the two qubits are completely correlated. Suppose Alice measures her qubit: if she finds $|0\rangle_A$ then she knows Bob will find $|0\rangle_B$ when he measures his qubit. If she finds $|1\rangle_A$ then she knows Bob will find $|1\rangle_B$. Alice can't communicate anything to Bob this way because she can't control whether she will find 0 or 1 when she measures. After they have both measured their qubits, they then share one random bit, which is a resource they can use for other tasks.

The purely quantum feature of entanglement that cannot be reproduced with suitably correlated classical systems is that Alice and Bob can use any basis states they like and the perfect correlation still appears. For example, $(|0\rangle + |1\rangle)/\sqrt{2}$ and $(|0\rangle - |1\rangle)/\sqrt{2}$ are another possible pair of basis states: the rule is they must be orthogonal to each other. This is the equivalent of changing the orientation of your classical coordinate system. However, Alice and Bob do both have to use the same basis: for a thorough discussion of the consequences of this and other reference frames in quantum mechanics, see Bartlett et al. [1].

Now that we have qubits to make a quantum register for our quantum computer, and measurements to read out the state of the quantum register, we need some quantum gates to perform our calculation with. There are two ways to change a quantum state: measurements, which are what happen when you (or something else) observes the state of a quantum system, and unitary dynamics, which is what quantum systems do when no-one is looking.

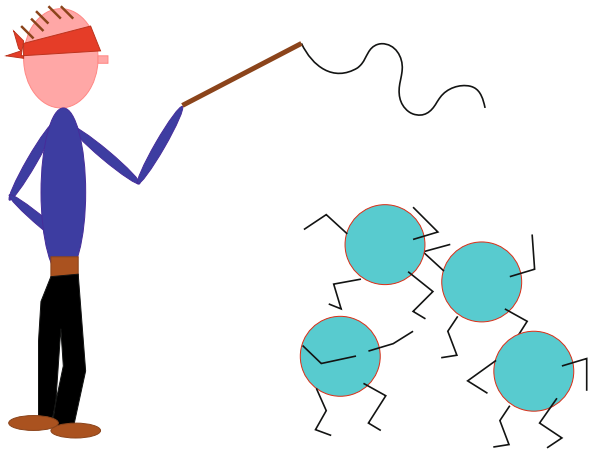

Quantum Computing, Figure 1
Unitary evolution: controlling qubits without looking at them

Happily, it is possible to steer a quantum system without looking at it (see Fig. 1), so we can make it evolve the way we choose through applying *unitary operations*. Unitary operations are reversible, the simplest example, acting on a single qubit, is the Hadamard gate H. We can define what it does by the effect on basis states:

$$\begin{aligned} H|0\rangle &= (|0\rangle + |1\rangle)/\sqrt{2} \\ H|1\rangle &= (|0\rangle - |1\rangle)/\sqrt{2}\,, \end{aligned} \tag{1}$$

and of course it can also act on superposition states,

$$H(\alpha|0\rangle + \beta|1\rangle) = \frac{1}{\sqrt{2}}\{(\alpha+\beta)|0\rangle + (\alpha-\beta)|1\rangle\}\,. \tag{2}$$

The Hadamard gate rotates the qubit through an angle of $\pi/2$. Notice also that applying the Hadamard gate twice gets you back where you started, $H(H|\psi\rangle) = |\psi\rangle$, the Hadamard gate is its own inverse.

Rotation gates can rotate through any other angle, for example,

$$\begin{aligned} R_\theta|0\rangle &= \cos(\theta/2)|0\rangle + \sin(\theta/2)|1\rangle \\ R_\theta|1\rangle &= \sin(\theta/2)|0\rangle - \cos(\theta/2)|1\rangle. \end{aligned} \tag{3}$$

To carry out computations, we also need gates that act on two qubits at once, for example, the familiar CNOT gate:

$$\begin{aligned} C|0\rangle_c|0\rangle_t &= |0\rangle_c|0\rangle_t \\ C|0\rangle_c|1\rangle_t &= |0\rangle_c|1\rangle_t \\ C|1\rangle_c|0\rangle_t &= |1\rangle_c|1\rangle_t \\ C|1\rangle_c|1\rangle_t &= |1\rangle_c|0\rangle_t \end{aligned} \tag{4}$$

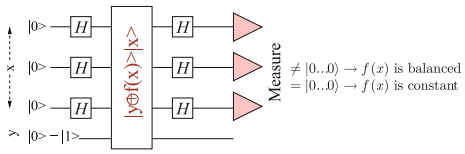

Quantum Computing, Figure 2
Circuit diagram for the Deutsch–Jozsa algorithm for $N < 8$. See text for details. Normalization factors have been omitted from the qubit states

where the first qubit (labeled c) is the control and remains unchanged, while the second qubit (labeled t) is the target and is flipped if the control qubit is in state $|1\rangle$.

Just as with classical computation, small sets of universal quantum gates exist from which any computation can be constructed efficiently [2].

A particular set is usually chosen to suit the physical components of the quantum computer, there are many possible choices, the gates described here are just a few of the one and two qubit gates available.

How Do Qubits Give Us Cool Computing?

Now that we have qubits to form a quantum register, quantum gates to perform our computation, and measurements to read out the result, the next step is to put them all together. We will begin with an example, which will illuminate the basic idea much more clearly than a general description. The Deutsch–Jozsa Algorithm was one of the earliest quantum algorithms, it solves the following promise problem. We are promised that the function $f(x) : x \mapsto \{0, 1\}$ is either

- balanced i. e., the number of times $f(x)$ outputs zero is equal to the number of times $f(x)$ outputs one, compared over all possible input values x (with $0 < x \leq N$), or,
- $f(x)$ is constant, i. e., the output is always either zero or one (but we don't know which) for any input x.

The cost of this computation is measured in "queries" or evaluations of the function $f(x)$. The circuit diagram in Fig. 2 shows how to solve this problem with a quantum computer.

There is one register of $n = \lceil \log_2 N \rceil$ qubits to hold the input x, and one extra single qubit $|y\rangle$. We start with all the qubits in the $|0\rangle$ state. The first trick is the set of Hadamards applied to the $|x\rangle$ register: $H|0\ldots0\rangle = (|0\ldots0\rangle + |0\ldots1\rangle + |0\ldots10\rangle + \cdots + |1\ldots1\rangle)/2^{-n/2}$ which is a superposition of all numbers $0\ldots2^n - 1$. We then make our one query to the oracle that computes $f(x)$ for us, using the superposition of all possible inputs now stored in $|x\rangle$. We get back a superposition of all the answers, which we store by adding it to the $|y\rangle$ qubit.

Remember we cannot detect a superposition by measurement alone, so the superposition of all the possible answers doesn't help us yet. The second trick allows us to detect what sort of superposition of answers we got back. Before adding the answer to $|y\rangle$, we prepare this qubit in the state $(|0\rangle - |1\rangle)/\sqrt{2}$. This can be done by flipping it from $0\rangle$ to $|1\rangle$ then applying a Hadamard gate. Then, when we add the answers to $|y\rangle$, there are three possibilities:

1. the answers are all $|0\rangle$: $|y\rangle|x\rangle = (|0\rangle - |1\rangle)|x\rangle/\sqrt{2} =$ unchanged.
2. the answers are all $|1\rangle$: $|y\rangle|x\rangle = (|1\rangle - |0\rangle)|x\rangle/\sqrt{2} =$ minus what it was before.
3. the answers are half and half: $|y\rangle|x\rangle =$ something else.

The second sequence of Hadamard gates then attempts to undo the first set and convert $|x\rangle$ back to the all zero state. In the first two cases, this succeeds (we can't detect the minus sign when we measure), but in the third case the Hadamards don't get us back where we started so there will be some ones in the $|x\rangle$ qubits which we find when we measure them. So, if we measure all zeros for $|x\rangle$ then $f(x)$ is constant, while if we measure any ones then $f(x)$ is balanced.

To solve this problem classically you have to examine the value of $f(x)$ for one value of x at a time and compare them. As soon as you accumulate both a zero and a one in the set of answers, you know (because of the promise) that $f(x)$ is balanced. But you can't be sure it is constant until you accumulate more than half the answers and see that they are all the same. So the best possible classical algorithm must take at least two queries, and could require up to $N/2 + 1$. Thus the quantum algorithm is fundamentally more efficient, requiring only the one evaluation of $f(x)$.

How Quantum Computing Got Started (A Little Bit of History)

The idea that quantum mechanics could give us fundamentally faster computers first occurred to Feynman [3] and independently to Deutsch [4]. They both observed that quantum systems are like a parallel computer (c.f. Feynman paths or many worlds respectively). It remained a cute idea for the next decade, until a key result showing *error correction* is theoretically possible implied it might be feasible to build one that really works [5,6,7]. Around the same time, Shor [8] found an algorithm for factoring that could in theory break crypto schemes based on products of large primes being hard to factorize. At this point the

funders sat up and took notice. Not just because factoring could become easier with a quantum computer, there are public key crypto schemes based on other "one-way" functions that are hard to compute given the public key, but easy to compute from the private key. But by shifting the boundary between what's easy and what's hard, quantum computing threatens to break other classical one-way functions too.

At the same time, *quantum communications* protocols were being developed, in particular, *quantum key distribution* solves the crypto problem created by Shor's algorithm [9]. Quantum resources can double the channel capacity, a technique known as *quantum dense coding* [10], and *quantum teleportation* can transmit a quantum state through a classical channel [11] with the help of some shared entanglement. Commercial quantum communications devices are available now for key distribution. They work over commercial fibre optics, they work well enough to reach from earth to satellites, and handheld devices under development [12,13]. It remains a niche market, but the successful technology in a closely related quantum information field has helped to keep the funding and optimism for quantum computers buoyant.

Error Correction for Quantum Computers

The most important result that allowed quantum computing to move from being an obscure piece of theory to a future technology is that error correction can work to protect the delicate quantum coherences (those signs, or phases, between different components of the superposition). Usually, the ubiquitous environmental disturbances randomize them very rapidly. Quantum systems are very sensitive to noise, which is why we inhabit a largely classical world.

The basic idea is simple: correct the errors faster than they accumulate. The main method for doing this is also conceptually simple: logical qubits are encoded in a larger physical quantum system, and the extra degrees of freedom used for error correction. A very simple example: for every qubit in your computation, use three identical qubits. If a single qubit error occurs, the remaining two will give you the correct answer. This is illustrated in Fig. 3. Two errors hitting the three qubits and the majority outcome will be wrong, though, so this code is only good if the error rates are very low.

What is not at all obvious is that this has any chance of working. Adding extra degrees of freedom (more qubits, for example) adds more places where errors can occur. Adding yet more qubits to correct these extra errors brings in yet more errors, and so on. Nonetheless, if the error rate

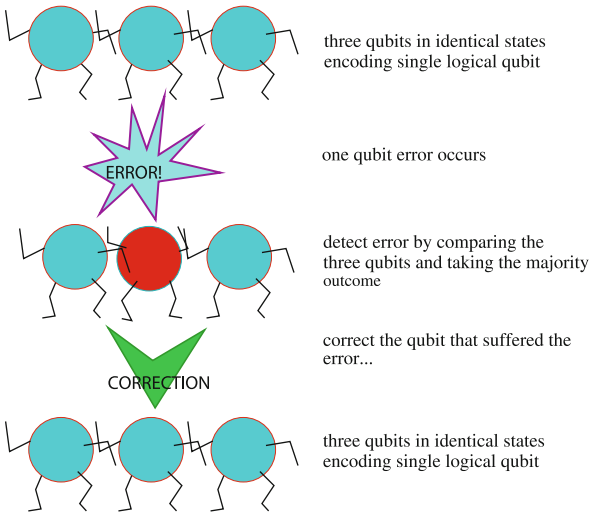

Quantum Computing, Figure 3
Error correction using three qubits to encode one

is low enough, even the errors arising from the error correction can be corrected fast enough [5,6]. This is known as the *threshold result*. The details of exactly how to accomplish this are considerable, and employ many techniques from classical error correcting codes [7,8]. For those wanting to know more, there are lots of tutorials available, e. g., [14].

The crucial piece of information is thus what the actual threshold error rate is. It can only be estimated, and current estimates suggest it is around 10^{-3} to 10^{-4} errors per qubit per quantum operation. This is rather smaller than the error rates in most physical systems that have been proposed for quantum computers, but not so small it is clearly unattainable as technology is refined. Various additional methods for avoiding the effects of decoherence have since been developed, such as decoherence-free subspaces [15,16], and using the quantum Zeno effect [17,18] to keep the system evolving in the direction you want.

Programming a Quantum Computer

The general structure of many quantum algorithms follows the same pattern we already saw in the Deutsch–Jozsa algorithm. This is illustrated in Fig. 4.
Each algorithm needs to have its own trick to make it work, so progress on expanding the number of quantum algorithms has been slow. Sometimes the same trick will work for a family of similar problems, but the number of distinct types remains limited. We have already seen an example of a promise problem in the Deutsch–Jozsa algorithm. Shor's algorithm and the many variants on this

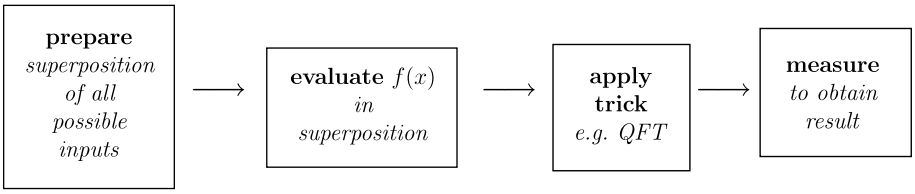

Quantum Computing, Figure 4
General structure of many quantum algorithms

method have essentially similar structure, while Grover's search and quantum walks have a distinctly different way of iterating towards the desired result. Quantum simulation, the earliest suggested application for quantum computation, is in a different category altogether and we will return to it near the end, once we have all the tools necessary to understand how it works. There are also quantum versions of almost anything else computational you can think of, such as quantum game theory, and quantum neural networks. Most of these other types of quantum information processing fall into the category of communication protocols, where the advantage is gained through quantum entanglement shared between more than one person or location. To consolidate our understanding of the functioning of quantum algorithms, we will briefly outline Shor's algorithm and then give an overview of quantum walks.

Shor's Algorithm

The task is to find a factor of a number $N = pq$ where p and q are both prime. First choose a co-prime number $a < N$. If a turns out not to be co-prime then the task is done. This can be checked efficiently using Euclid's algorithm, but only happens in a very small number of cases. Then run the quantum computation shown in Fig. 5 to find r, the periodicity of $a^x \pmod N$. The first half of the computation creates a superposition of $a^x \pmod N$ for all possible inputs x up to N^2. This upper limit is not essential, it determines the precision of the output and thus the number of times the computation has to be repeated on average to succeed. Shor's first insight was that this modular exponentiation step can be done efficiently. The answer is then contained in the entanglement between the qubits in the upper and lower registers [19]. Shor's second insight was that a quantum version of the familiar discrete Fourier transform can be used to rearrange the state so that measuring the upper register allows one to calculate the value of r with high probability. Once we have found r, then $a^{r/2} \pm 1$ gives a factor of N (with high probability). If the first attempt doesn't find a factor, repeating the computation a few times with different values of a rapidly increases the chances of success.

The classical difficulty of factoring is not known, but the best classical algorithms we have are sub-exponential (exponential to some slower than linear function of N), whereas Shor's algorithm is polynomial in N. This is roughly speaking an exponential improvement, and it promises to bring factoring into the set of "easy to solve" problems (i. e. polynomial resources) from the "hard to solve" set (exponential resources). There is a whole family of problems using same method, all based around identifying a hidden subgroup of an Abelian group. Some extensions have been made beyond Abelian Groups but this is in general much harder (for a review, see [20]).

Quantum Walks

Classical random walks underpin many of the best classical algorithms, so finding a faster quantum version of a random walk is a good bet for a new type of quantum algorithm. Quantum random walks that are faster than classical random walks have been found in a variety of guises, but turning them into algorithms is harder. They have been applied most successfully to problems related to searching, such as subset finding [21] and element distinctness [22].

A simple recipe for a quantum walk on a line goes as follows:

1. Start at the origin
2. Toss a qubit (quantum coin)

$$\mathbf{H}|0\rangle \longrightarrow (|0\rangle + |1\rangle)/\sqrt{2}$$
$$\mathbf{H}|1\rangle \longrightarrow (|0\rangle - |1\rangle)/\sqrt{2}$$

3. Move *left* and *right* according to qubit state

$$\mathbf{S}|x, 0\rangle \longrightarrow |x-1, 0\rangle$$
$$\mathbf{S}|x, 1\rangle \longrightarrow |x+1, 1\rangle$$

4. Repeat steps 2. and 3. T times
5. measure position of walker, $-T \leq x \leq T$

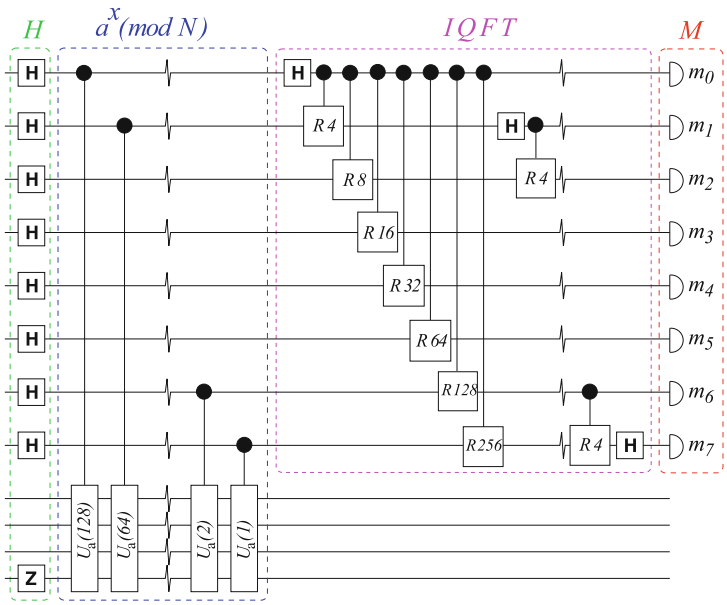

Quantum Computing, Figure 5

Circuit diagram for Shor's algorithm for factoring shown for the example of factoring 15. *H* is a Hadamard gate, *R*n is a rotation by π/n, and the *U* gates perform the modular exponentiation (see text). IQFT is inverse quantum Fourier transform, and *M* is the final measurement

If you repeat steps 1. to 5. many times, you get a probability distribution $P(x, T)$, which turns out to spread quadratically faster than a classical random walk.

If we add a little decoherence (measure the quantum walker with prob p at each step) then we get a top hat distribution for just the right amount of noise [23], see Fig. 6. For related results of mixing times on cycles and torii, see Refs. [24,25].

Quantum walks can be used to find a marked item in an unsorted database, the equivalent of finding a person's name from their phone number by searching the telephone directory. Classical solutions have to check on average $N/2$ entries, and in the worst case need to check them all. This problem was investigated by Grover [26,27], who found a quantum algorithm that is quadratically faster than classical searching. His algorithm doesn't use quantum walks, but Shenvi et al. [28] showed that quantum walks can perform the task just as well. The recipe for this is:

1. Start in a uniform distribution over graph with N nodes.
2. Use Grover coin everywhere except the marked node.
3. Run for approx $\frac{\pi}{2}\sqrt{N/2}$ steps.
4. Particle will now be at the marked node with high probability.

This is the inverse of starting at the origin and trying to get uniform (top hat) distribution. A quadratic speed up is the best that can be achieved by any algorithm tackling this problem [29].

The quantum walk version of Grover's search has been generalized to find more than one item. Magniez et al. [30,31] show how to detect triangles in graphs, Ambainis [22] applies quantum walks to deciding element distinctness, and Childs and Eisenberg [21] generalize to finding subsets. These generally obtain a polynomial improvement over classical algorithms, for example, reducing the running time to $O(N^{2/3})$ compared to $O(N)$ for classical methods.

There is one known problem for which a quantum walk can provide a more impressive speed up. The "glued trees" problem [32] is also about finding a marked state, but this time the starting point is also specified, and an oracle is available to give information about the possible paths between the start and finish. This achieves an exponential speed up, but has not been generalized to other problems. For a short review of quantum walk algorithms, see [33].

How Powerful Is Quantum Computing?

What we've seen so far is quantum computers providing a speed up over classical computation, i. e., new *complexity classes* to add to the zoo [34], but nothing outside of

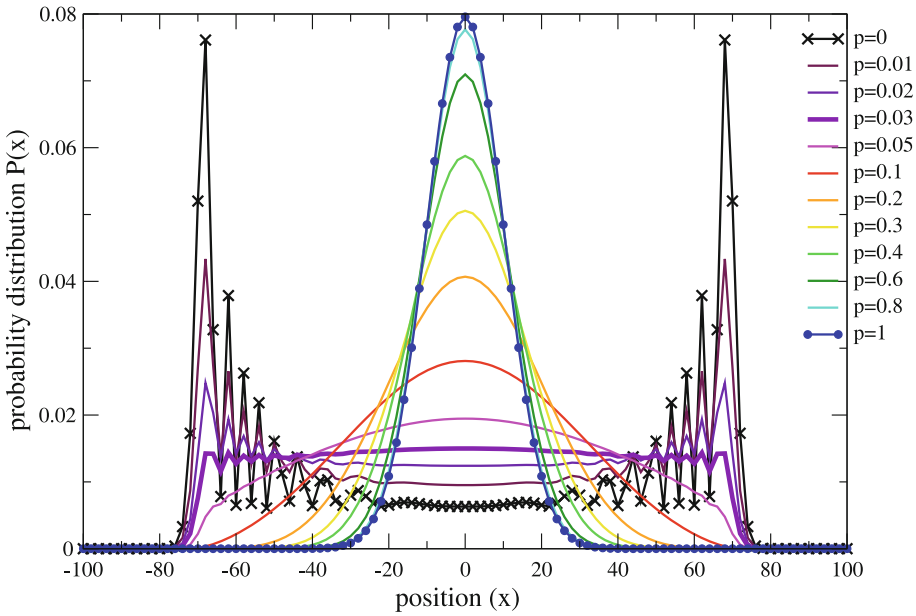

Quantum Computing, Figure 6
Comparison of a quantum walk of 100 steps with the corresponding classical random walk. Progressively increasing the decoherence produces a "top-hat" distribution for the right choice of decoherence rate

the classical computability limits. The mainstream view is that quantum computing achieves the same *computability* as classical computing, and the quantum advantage may be characterized completely by complexity comparisons. The intuitive way to see that this is likely to be correct is to note that, given the mathematical definition of quantum mechanics, we can always simulate quantum dynamics using classical digital computers, but we generally can't do it efficiently. The chink in this argument is that quantum mechanics is not digital. Qubits can be in an arbitrary superposition $\alpha|0\rangle + \beta|1\rangle$, where $|\alpha|$ can take any real value between zero and one. Thus we can only simulate the quantum dynamics to within some precision dictated by the amount of digital resources we employ. However, we cannot measure α, we can only infer an approximate value from the various measurements we do make. And classical computers also have continuous values for quantities such as voltages that we choose to interpret as binary zeros and ones. So it isn't clear there could be any fundamental difference between classical and quantum in this respect that could separate their computational ability.

So will quantum computers give us a practical advantage over classical computers? How many qubits do we need to make a *useful* quantum computer? This depends on what we want to do:

Simulating a quantum system: for example, $N \times$ two-state particles $\longrightarrow 2^N$ possible different states. The state of the system could be in superposition of all of these 2^N possible states. Classical simulations thus require one complex number per state: $2^{N+1} \times$ size-of-double $\longrightarrow$ 1 Gbyte can store the state of just $N = 26$ 2-state systems. The current record is $N = 36$ in 1 Terabyte of storage [35] – each additional particle *doubles* the memory required. Thus more than 40 or so qubits is beyond current classical resources.

In the cases where we don't need to keep track of all the possible superpositions, e. g., if only nearest neighbor interactions are involved, larger classical simulations can be performed, for example, see [36].

Shor's factoring algorithm: the best classical factoring to date is 200 digit numbers (RSA-200) which is approximately 665 bits. Shor's quantum algorithm needs $2n$ qubits in the QFT register plus $5n$ qubits for the modular exponentiation (lower register plus ancilla qubits), a total of $7n$ logical qubits. A 665 bit number therefore needs 4655 logical qubits. We now need to take account of error correction. Again, this depends on the physical quantum computer and the error rates that have to be corrected. If the error rate is close to the threshold of 10^{-3} to 10^{-4}, then more error correction is needed. For low error rates, maybe 20–200 physical qubits per logical qubit

are required. For high error rates the numbers blow up quickly to maybe 10^5 physical qubits per logical qubit. This suggest that factoring won't be the first useful application of quantum computing, despite being the most famous, we may need Tera-qubit quantum computers to produce useful results here. Although the scaling favors quantum computing, the crossover point is very high.

Ultimate Physical Limits to Computation

This is a subject that is revisited regularly. The version by Lloyd [37] is the best known of these calculations that has an explicitly quantum flavor. Physical limits can be deduced from these fundamental constants:

The speed of light: $c = 3 \times 10^8 \mathrm{m\,s^{-1}}$
Boltzmann's constant: $k = 1.38 \times 10^{-23} \mathrm{J\,K^{-1}}$
The gravitational constant:
$G = 6.67 \times 10^{-11} \mathrm{m^3 kg^{-1}\,s^{-2}}$
Planck's constant: $\hbar = 1.055 \times 10^{-34} \mathrm{J\,s}$

Consider 1 kg of matter occupying volume of 1 liter $= 10^{-3}\,\mathrm{m}^3$ – a little smaller than today's laptops. How fast can this matter compute? Quantum mechanics tells us that the speed of computation is limited by its average energy E to $2E/\pi\hbar = 5 \times 10^{50}$ operations per second with $E = mc^2$ given by the mass of 1 kg. This is because the time-energy Heisenberg uncertainty principle, $\Delta E \Delta t \geq \hbar$, says that a state with a spread in energy of ΔE takes a time Δt to evolve into a different state. This was extended by Margolus and Levitin [38,39] to show that a state with average energy E takes a time $\Delta t \geq \pi\hbar/2E$ to evolve to a different state.

The number of bits it can store is limited by its entropy, $S = k \ln W$ where W is the number of states the system can be in. For m bits, $W = 2^m$, so $S = km$, and the system can store at most m bits. Our estimate of m is thus dependent on what our laptop is made of. It also depends on the energy E, since not all of the W possible states have the same energy, so in practice we don't get the maximum possible $W = 2^m$ number of possible states, since we decided our laptop had a mass-energy of 1 kg. In practice this restriction doesn't change the numbers very much. There are various ways to estimate the number of states, one way is to assume the laptop is made up of stuff much like the early universe: high energy photons (the leftover ones being the cosmic microwave background radiation). This gives around 2×10^{31} bits as the maximum possible memory.

Having noted that entropy S is a function of energy E, we can deduce that the number of operations per second per bit is given by $2Ek/\pi\hbar S \propto kT/\hbar$, where $T = \partial S/\partial E$ is the effective temperature of the laptop. For our 1 kg laptop this means the number of operations per bit is about 10^{19}, and the temperature is about 5×10^8 degrees, a very hot plasma. This also implies the computer must have a parallel type of operation since the bit flip time is far smaller than the time it takes for light to travel from one side to the other. If you compress the computer into a smaller volume to reduce the parallelism it reaches serial operation at the black hole density! The evaporation rate for 1 kg black hole is 10^{-19} seconds, after 10^{32} ops on 10^{16} bits. This is very fast, but similar in size to conventional computers. Prototype quantum computers already operate at these limits, only with most of their energy locked up as mass.

The salutary thing about these calculations is that once you scale back to a computer in which the mass isn't counted as part of the available energy, we are actually within some sort of spitting distance of the physical limits. Computational power has been increasing exponentially for so long we have become used to it doubling every few years, and easily forget that it also requires increasing resources to give us this increased computational capacity. Based on current physics, there isn't some vast reservoir of untapped computational power out there waiting for us to harness as our technology advances, and most of us will see a transition to a significantly different regime in our lifetimes. Which, one may argue, highlights the importance of unconventional computation, where further increases in speed and efficiency may still be available after the standard model has run out of steam.

Can We Build a Quantum Computer?

The short answer to this question is "yes", and people already have built small ones, but we don't know if we can build one big enough to be useful for something we can't already calculate more easily with our classical computers. In 2000, DiVincenzo [40] provided a checklist of five criteria to assess whether a given architecture was capable of being used to build a scalable quantum computer, i. e., one that can be made large enough to be useful:

1. *scalable* qubits – to allow quantum computers to be built large enough to solve real problems
2. prepare initial state – e. g. all qubits in state $|0\rangle$
3. decoherence times far longer than gate times – to allow many quantum operations before the quantum coherence is degraded
4. universal set of quantum gates – a suitable two qubit gate is sufficient
5. measurement of single qubits to read out result – can be hardest part and can be much slower than other operations.

The notion of *scalability* is crucial in the quest to build a useful quantum computer. We need designs that can be tested at small sizes, registers of a few qubits, then scaled up without fundamentally changing the operations we tested at small scales. The most feasible architecture for achieving this is using small, repeatable, connectable units [41].

One of the most important models for scalable quantum computers is the combination of stationary and flying qubits. Flying qubits are usually photons while stationary qubits can be almost any other type that can interact with photons, such as atoms or quantum dots. The motivation for this model is that it can be hard to do everything in one type of physical system, for example, one-qubit gates are easy to apply to photons while two-qubit gates are hard. Also, stationary qubits are limited to nearest neighbor interactions, which then involves many swap operations to allow qubits further apart to be gated together, but this can be overcome using flying intermediaries to connect distant qubits. Atoms in traps combined with photons are currently seen as the best scalable architecture, but this is by no means the only player in the ring.

What Have We Got so Far?

Single qubits – with enough control to initialize them in a chosen state and apply single qubit gates to them – have been demonstrated in a wide variety of systems: this is only a partial list:

- photons – using polarization or path to represent the qubit
- atoms and ions in linear traps – qubit registers
- atoms loaded into optical lattices one per site
- quantum dots – with *single* spins (electrons or holes)
- electrons floating on liquid helium – electron spin
- phosphorus nuclei embedded in silicon – Kane model
- superconducting devices – using either currents or magnetic flux
- nuclear magnetic resonance (NMR) – using nuclear spins as qubits.

Some of these systems have also been demonstrated performing two qubit gates, sometimes even more than one gate at a time maintaining coherence. Progress is slow but steady and this list will likely be out of date by the time you read it, though not likely by more than a few small numbers. For more details of the main types of architectures, [42] is a good starting point. The current capabilities are limited to

- a few (less than 10) qubits entangled to order
- a few (less than 20) quantum gate operations
- NMR can factor 15 (using 7 qubits)
- NMR qubit record is 7 (running an algorithm), and 12 (benchmarking)
- atoms: 6 in controlled superposition.

The vision is driving beautiful experiments: to find out up to date information on current progress, visit the websites of the major experimental collaborations, for example,

- NIST: http://qubit.nist.gov/
- MIT/Quanta: http://www.media.mit.edu/quanta/
- University Vienna: http://www.quantum.at/research/quantum-computation.html
- EU Road map: http://qist.ect.it/Reports/reports.htm

and look for recent articles in Nature and Science.

Unconventional Extensions

Quantum computing as just described appears to mimic classical digital computing quite closely in terms of bits (qubits), gates, error correction, scalable architectures and so on. In many ways this is just because this part of the picture is easiest to relate to those familiar with classical computing. In reality, neither the historical development, nor the creativity with which theorists and experimentalists are tackling the difficult task of trying to build a working quantum computer justify this impression. In this section I will attempt to set the record straight, and hint at some of the wilder territory beyond.

What About Analogue?

Classical analogue computation was once commonplace, yet was swept aside by the advance of digital computation. To understand why this happened we need to review what it means to binary encode our data in a digital computer. Binary encoding has profound implications for the resources required for computation, as this table explains.

Unary vs. Binary Coding

Number	Unary	Binary
0		0
1	•	1
2	• •	10
3	• • •	11
4	• • • •	100
...	...	...
N	$N \times$ •	$\log_2 N$ bits
Read out:	distinguish between measurements with N outcomes	$\log_2 N$ measurements with 2 outcomes each
Accuracy:	errors scale linear in N	errors scale $\propto \log N$

Binary encoding provides an exponential advantage (reduction) in the amount of memory required to store the data compared to unary encoding. It also means the precision costs are exponentially smaller. To double the accuracy of a unary representation requires double the resources, while a binary representation achieves this with a single extra bit. This is the main reason why digital computing works so well. It is so obvious it is easy to forget that there was once analogue computing in which the data is represented directly as the magnitude of a voltage or water level displacement, or, in the slide rule, with the logarithms etched onto the stick. Those of you too young to have owned a slide rule should now play with Derek's virtual slide rules here: http://www.antiquark.com/sliderule/sim/ to get a feel for what one version of analogue computation is like.

It does not, of course, have to be binary encoding, numbers in base three or base ten, for example, gain the exponential advantage just as well. One of the first to discuss the implications of binary encoding in a quantum computing context was Jozsa [43], and this was refined in Blume–Kohout et al. [44] and Greentree et al. [45]. Binary encoding matters for quantum computing too in the standard digital model. It gains the same exponential advantage from encoding the data into qubits as classical digital computers gain from using bits.

Continuous Variable Quantum Computing

Although many quantum systems naturally have a fixed number of discrete states (polarization of a photon or electronic spin for example), there are also variables like position x and momentum p that can take any real value. These can be used to represent data in the same way as voltages or other continuous quantities are used in classical analogue computers. This quantum version of analogue computation is usually known as continuous variable, or CV computation.

The important difference between quantum and classical continuous variables is that a conjugate pair, such as the position and momentum, are not independent of each other. Instead, x and p are related by the commutator, $[x, p] = i\hbar$ Knowing x exactly implies p is unknowable and vice versa, this is called *squeezing*. The most commonly described continuous variable quantum system uses electromagnetic fields, but rather than specializing to the case of single photons as one does for digital quantum computing, single modes of the field are employed. Braunstein and van Loock [46] provide a comprehensive review of CV quantum information processing for those interested in further details.

Lloyd and Braunstein [47] showed that CV computation is universal, i. e., it can compute anything that a digital quantum computer can compute. They did this by showing that it is possible to obtain any polynomial in x and p efficiently using operations that are known to be available in physical systems. The universal set of operations has to include a nonlinear operation applied to one of the variables, which means in this context higher than quadratic in x and p. For electromagnetic fields, an example of a suitable nonlinear operation is the Kerr effect, which produces a quartic term in x and p. This has the effect of an intensity dependent refractive index. Unfortunately, although it is observable, in real physical systems the Kerr effect is too weak to be of practical use in CV computation. Without the nonlinear term, the operations available are Gaussian preserving, which means they can be simulated efficiently with a classical computer [48]. Some operations are easier than in digital quantum computers, the Fourier transform is a single operation, rather than a long sequence of controlled rotations and Hadamard gates.

Continuous variable information processing is especially suitable for quantum communications because of the practicality of working with bright beams of light instead of single photons. Most quantum communications tasks only require Gaussian operations, which are easily achieved with standard optical elements. Quantum repeaters [49] to extend quantum communications channels over arbitrary distances are a good example. However, a useful CV quantum computer isn't a practical option using current experimental capabilities. The precision that can be obtained in a single mode is limited by the degree of squeezing that can be performed on the mode, currently about 10 dB, which is equivalent to 3.7 bits. An extra bit of precision doubles the required squeezing so CV quantum computation suffers from the same precision problems as classical analogue computing.

Hybrid Quantum Computing Architectures

Having noted above that continuous variable quantum systems are especially suited to quantum communications tasks, the obvious way to use them for quantum computing is to make a high speed quantum "bus" (qubus) to communicate between stationary qubit registers. Although light and matter interact in simple easy ways, creating a suitable gate to transmit the quantum information between qubits is a far from trivial task. Coherent light as used for continuous variable quantum information is also often employed to measure the final state of the matter qubits at the end of the computation.

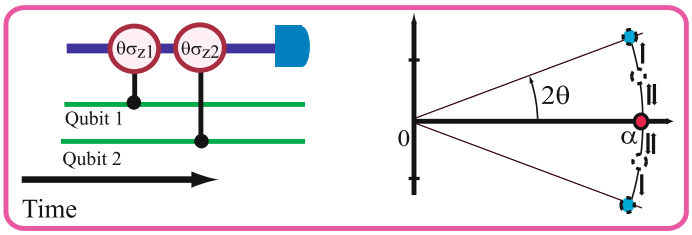

Quantum Computing, Figure 7
Qubus parity gate operation

Unless the required operations are done in a way that reveals only trivial information about the quantum state while the computation is in process, the quantum superpositions will instead be destroyed. The solution is built around an elegant implementation of a parity gate, illustrated in Fig. 7. The left half shows a coherent light beam (blue) interacting with two qubits in turn (green) by means of an interaction that causes a phase shift in the light that depends on the state of the qubit. The right half shows the phase of the light after the interactions. If the two qubits are in the same state, the light will have a phase in one of the two blue positions. If the two qubits are in different states, the two phase shifts will cancel leaving the phase unaltered (red). Provided the angle of the shift, θ is large enough, the two cases can be distinguished by measuring only the coordinate of the horizontal axis, not the vertical distance away from it. For a detailed exposition illustrated in a superconducting qubit setting, see [50].

Cluster State Quantum Computing

First proposed by Raussendorf and Briegel [51], and called a "one way quantum computer", cluster state quantum computing has no direct classical equivalent, yet it has become so popular it is currently almost the preferred architecture for development of quantum computers. Originally designed for atoms in optical lattices, the idea is to entangle a whole array of qubits in global operations, then perform the computation by measuring the qubits in sequence. This architecture trades space resources for a temporal speed up, requiring roughly n^2 physical qubits where the gate model uses n. The measurement sequence can be compressed to allow parallel operation for all but a few measurements that are determined by previous results, and require the outcomes to be fed forward as the computation proceeds. One obvious prerequisite for this architecture is qubits that can be measured quickly and easily. Measurement is often slower than other gate operations so this does limit in practice which types of qubits can be used.

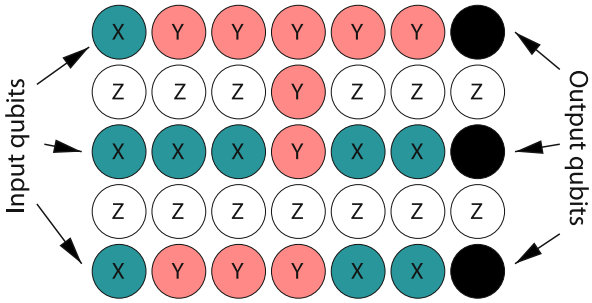

Quantum Computing, Figure 8
Fragment of a cluster state quantum computation. The computation can be thought of as proceeding from left to right, with the qubit register arranged vertically, though all qubits in these gates can be measured at the same time. White qubits do not contribute directly to the computation and are "removed" by measuring in the *Z* basis. Pink qubits are measured in the Y basis and form the gates, at the top a CNOT and at the bottom a Hadamard. Blue qubits are measured in the *X* basis and transmit the logical qubit through the cluster state. The black qubits contain the output state (not yet measured, since their measurements are determined by the next operations). Details in [52]

Since the measurements proceed from one side of the lattice to the other, see Fig. 8, instead of making the whole array in one go, the qubits can be added in rows as they are needed. This also allows the use of a probabilistic method for preparing the qubits, which can be repeated until it succeeds in a separate process to the main computation [53]. This idea grew out of schemes for linear optical quantum computation in which gates are constructed probabilistically and teleported into the main computation once they succeed [54]. Adding rows of qubits "just in time" does remove one of the main advantages of the original proposal, i. e., making the cluster state in a few simple global operations. However, it gains in another respect in that the data is regularly being moved into fresh qubits, so the whole system doesn't have to be kept free of decoherence throughout the computation.

Cluster state quantum computation has been important on a theoretical level also because it made people think

harder about the role of measurement in quantum computing [55]. Instead of striving to obtain perfect unitary quantum evolution throughout the computation, cleverly designed measurements can also be used to push the quantum system along the desired trajectory.

Topological Quantum Computing

In Sect. "Digital Quantum Computing" we learned about qubits: distinguishable quantum two state systems. But there are more exotic quantum species, indeed, quantum particles are indistinguishable when more than one of them is around in the same place. This means if you swap two of them, you can't tell anything happened. And there are precisely two ways to do this for particles living in our familiar three spatial dimensions:

$$|\psi_A\rangle_1|\psi_B\rangle_2 \longrightarrow |\psi_B\rangle_1|\psi_A\rangle_2 \tag{5}$$

corresponding to *Bose–Einstein particles* (photons, Helium atoms), or

$$|\psi_A\rangle_1|\psi_B\rangle_2 \longrightarrow -|\psi_B\rangle_1|\psi_A\rangle_2 \tag{6}$$

corresponding to *Fermi–Dirac particles* (electrons, protons, neutrons). That extra minus sign can't be measured directly, but it does have observable effects, so we do know the world really works this way. One important consequence is that Fermi–Dirac particles cannot occupy the same quantum state, whereas Bose–Einstein particles can, and frequently do. From this we get, among many other things, lasers (photons – Bose–Einstein particles) and semiconductors (electrons – Fermi–Dirac particles), without which modern technology would be very different indeed.

A simple change of sign when swapping particles won't make a quantum computer, but something in between one and minus one can. For this we have to restrict our particles to two spatial dimensions. Then it is possible to have particles, known as *anyons*, that acquire more complicated phase factors when exchanged. Restricting quantum particles to less than three spatial dimensions is not as hard as it sounds. A thin layer of conducting material sandwiched between insulating layers can confine electrons to two-dimensional motion, for example.

The first proposal for a quantum computer using anyons was from Kitaev [56], using an array of interacting spins on a planar surface. The anyons are formed from excitations of the spins – think of this as similar to the way electronic displays such as those found in trains and stations often work by flipping pixels between black and yellow to form characters. This opened up the field to a host of connections with group theory, gauge theory, and braid (knot) theory, and produced a new quantum algorithm to approximate the Jones polynomial [57], a fundamental structure in braid theory. If you already know something about any of the above mathematical subjects, then you can jump into the many accounts of topological quantum computation (also going by the names geometric or holonomic quantum computing). A recent introduction can be found in [58].

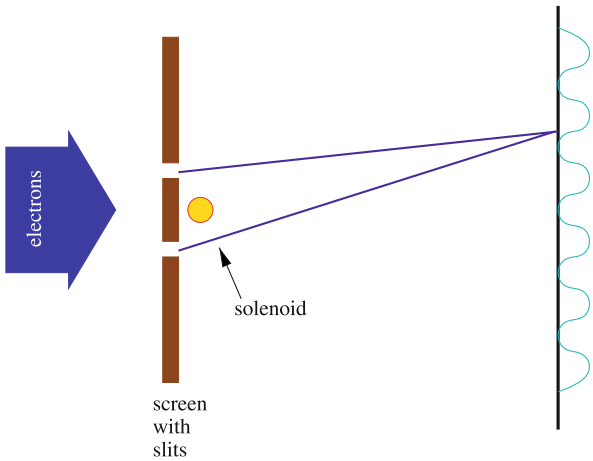

Quantum Computing, Figure 9
Aharonov Bohm affect

Here is a simple example of how quantum phases work, to further illuminate the idea. The Aharonov–Bohm effect [59] was discovered in 1959 and has been observed experimentally. It involves two ingredients. First, electrons can be projected through a double slit much like photons can, and they produce an interference pattern when detected on the far side of the slits. If you aren't familiar with Young's double slit experiment [60], note that you need to think about it in terms of single electrons going through both slits in superposition. This is an example of the quantum behavior of particles that combines both wave and particle properties. The interference pattern arises because the electron can't land on the detector in places where the path length difference is half a wavelength, and prefers to land near where the path difference is a whole number of wavelengths. The second ingredient is a solenoid, a coil of wire with a current running through it. It generates a magnetic field inside the coil that does not leak outside. If you put the solenoid between the two slits, see Fig. 9, the interference pattern shifts, even though there is no magnetic field where the electrons are traveling (they can't pass through the solenoid).

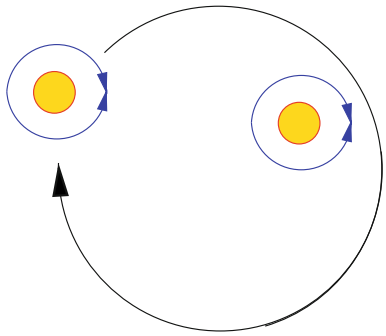

Quantum Computing, Figure 10
Anyons: charge (*blue*) encircling magnetic field (flux, *red+yellow*)

From this we deduce two things: interference patterns are a way to detect the extra quantum phases that are generated in topological quantum effects, and, even though there was no interaction between the electrons and the magnetic field, i. e., no energy exchanged, the magnetic field still affected the quantum properties of the electrons just by being inside the two paths the electrons take through the slits.

Now imagine shrinking that experiment down to a magnetic field with a charge (electron) encircling it. This is one way to think of anyons. Confined to two dimensions, the anyons can be moved round each other, and the charge part of one going round the magnetic field part of the other affects the quantum phase. When you want to measure them, you manipulate them so that they form an interference pattern. The rest of the details of how to make them perform quantum computation comes down to the mathematics of group theory.

Apart from delighting the more mathematically-minded quantum computing theorists, this type of quantum computing may have some practical advantages. The effects are generated just by moving the anyons about, and the exact path isn't critical, only that they go around the right set of other anyons. So there is less chance of errors and disturbance spoiling the quantum computation. However, constructing a physical system in which anyonic excitations can be manipulated to order is not so easy. In fact, it has not yet been demonstrated experimentally at all, though there is no shortage of suggestions for how to do it, see for example, [61].

Adiabatic Quantum Computing

Adiabatic quantum computation is the quantum equivalent of simulated annealing. Simulated annealing is a computational method that finds the minimum solution to a problem by starting from an initial state and evolving towards lower values until a minimum is found. In order to avoid being stuck in a local minimum that is higher than the desired solution, some random moves that increase the value are necessary. This can be thought of as starting in a high temperature distribution and slowly cooling the simulation until the lowest energy state is found. The slowness of the cooling is to allow it to get out of local minima on the way.

Quantum mechanics provides a more effective tool for this problem, thanks to the quantum adiabatic theorem. Quantum states don't get stuck in local minima, they can "tunnel" through barriers to lower-lying states. So there is no need to start in a random initial state, instead, the evolution starts in a minimum energy state that is easy to prepare and transforms from there into the minimum energy state we are interested in. The quantum adiabatic theorem tells you how long it will take to evolve to the desired state without ending up in some higher energy states by mistake.

Suppose you can prepare a system in the grounds state $|\sigma_0\rangle$ of the Hamiltonian H_{simple}, and you want to find the ground state $|\phi_0\rangle$ of the Hamiltonian H_{hard}. Then, provided you can evolve the systems under both Hamiltonians together, you applies the following Hamiltonian $H(t)$ to $|\sigma_0\rangle$,

$$H(t) = \{1 - \alpha(t)\}H_{\text{simple}} + \alpha(t)H_{\text{hard}} \tag{7}$$

where $\alpha(0) = 0$, $\alpha(T) = 1$ and $0 \le \alpha(t) \le 1$. The total time the simulation will take is T, and $\alpha(t)$ controls how fast the Hamiltonians switch over. The key parameter is the difference, or gap Δ, between the ground and first excited energy states. It varies as the state evolves from $|\sigma_0\rangle$ to $|\phi_0\rangle$ and the smaller it is the slower the state must evolve to stay in the ground state. Let Δ_{min} be the smallest gap that occurs during the adiabatic evolution. Then the quantum adiabatic theorem says that provided $T \gg \varepsilon/\Delta^2_{\text{min}}$, the final state will be $|\phi_0\rangle$ with high probability. Here ε is a quantity that relates the rate of change of the Hamiltonian to the ground state and first excited state. It is usually about the same size as the energy of the system. The important parameter is thus Δ_{min}. For hard problems, Δ_{min} will become small and the required time T will be large.

Adiabatic quantum computing was introduced by Farhi et al. [62] as a natural way to solve SAT problems. SAT stands for satisfiability and a typical problem is expressed as a Boolean expression, for example,

$$\begin{aligned} &B(x_1, x_2, x_3, x_4) \\ &= (x_1 \vee x_3 \vee \tilde{x}_4) \wedge (\tilde{x}_2 \vee x_3 \vee x_4) \wedge (\tilde{x}_1 \vee x_2 \vee \tilde{x}_3) \end{aligned} \tag{8}$$

a set of clauses of three variables each combined with "or" that are then combined together with "and". The tilde over a variable indicates the negation of that variable. The problem is to determine whether there is an assignment of the variables x_1, x_2, x_3, x_4 that makes makes the expression true (in the above simple example there are several such assignments, such as $x_1 = 0, x_2 = 0, x_3 = 0, x_4 = 0$).

The Boolean expression is turned into a Hamiltonian by constructing a cost function for each clause. If the clause is satisfied, the cost is zero, if not, the cost is one unit of energy. The Hamiltonian is the sum of all these cost functions. If the ground state of the Hamiltonian is zero the expression can be satisfied, but if it is non-zero then some clauses remain unsatisfied whatever the values of the variables.

The adiabatic quantum computation only works efficiently if it runs in a reasonably short amount of time, which in turn requires the gap, $\Delta_{\min}$ to remain large enough throughout the evolution. Figuring out whether this is so is not easy in general, and has led to some controversy about the power of quantum computation. The 3SAT problem (clauses with three variables as above) is in the general case a hard problem that requires exponential resources to solve classically. Clearly adiabatic quantum computing can also solve it, but the question is how long it takes. The best indications are that the gap will become exponentially small for hard problems, and thus the adiabatic quantum computation will require exponential time to solve it. For dissenting views see, for example, [63] discussing the traveling salesman problem, another standard "hard" problem. One of the difficulties in analyzing this issue is that adiabatic quantum computation uses the continuous nature of quantum mechanics, and care must be taken when assessing the necessary resources to correctly account for the precision requirements.

On the other hand, we do know that adiabatic quantum computing is at least equivalent to digital computation (i. e., it can solve all the same problems using equivalent resources) [57,64,65]. There are also indications that it can be more resilient to some types of errors [66]. It has even been implemented in a toy system using nuclear magnetic resonance (NMR) quantum computation [67].

Quantum Simulation

We have already noted that quantum simulation was the original application that inspired Feynman [3] and Deutsch [4] to propose the idea of quantum computation. We have also explained why quantum systems cannot generally be simulated efficiently by classical digital computers, because of the need to keep track of the enormous number of superpositions involved. In 1996 Lloyd [68] proved that a quantum system can simulate another quantum system efficiently, making the suggestions from Feynman and Deutsch concrete.

The first step in quantum simulation is trivially simple, the quantum system to be simulated is mapped onto the quantum simulator by setting up a one-to-one correspondence between their state spaces (which are called Hilbert spaces for quantum systems). The second step is non-trivial, Lloyd proved that the unitary evolution of the quantum system being studied can be efficiently approximated in the quantum simulator by breaking it down into a sequence of small steps that use standard operations. Mathematically this can be written

$$\exp\{iHt\} \simeq \big(\exp\{iH_1 t/n\}\exp\{iH_2 t/n\} \ldots \exp\{iH_j t/n\}\big)^n + O(t^2/n)[H_j, H_k]$$

where $\exp\{iHt\}$ is the unitary evolution of the quantum system (H is called the Hamiltonian) and $H_1 \ldots H_j$ are a sequence of standard Hamiltonians available for the quantum simulator. The last term says that the errors are small provided n, the step size, is chosen to be small enough. There are variations and improvements on this that use higher order corrections to this approximation, see for example, [69]. Quantum simulation has even been demonstrated experimentally [70] in a small test system using an nuclear magnetic resonance (NMR) quantum computer.

Just as with the quantum algorithms discussed in the first section, evolving the quantum simulation is only half of the job. Extracting the pertinent information from the simulation requires some tricks, depending on what it is you want to find out. A common quantity of interest is the spectral gap, the energy difference between the ground and first excited states, the quantity Δ discussed in Subsect. "Adiabatic Quantum Computing". One way to obtain this is to create a superposition of the two energy states and evolve it for a period of time. The phase difference between the ground and excited state will now be proportional to the spectral gap. To measure the phase difference requires a Fourier transform, either quantum or classical. Either way, the simulation has to be repeated or extended to provide enough data points to obtain the spectral gap to the desired precision. Even creating the initial state in superposition of ground and excited states is non-trivial. It has been shown that this can be done for some systems of interest by using quasi-adiabatic evolution starting from the ground state. By running the adiabatic evolution a bit too fast, the state of the system doesn't remain perfectly in

the ground state and the required superposition in generated [71,72].

However, because there is no binary encoding, unlike classical simulation on a digital computer, accuracy is a problem. In particular, accuracy does not scale efficiently with time needed to run the simulation [73]. Is this going to be a problem in practice? Poor scaling for accuracy didn't stop people building useful analogue computers from electrical circuits or water pipes in the days before digital computers were widely available. As we noted, a quantum simulator becomes useful at around 40 qubits in size, which means it may well be the first real application of quantum computing to calculate something we didn't know already. In the end, the accuracy required depends on the problem we are solving, so we won't know whether it works until we try.

Future Directions

The wide range of creative approaches to quantum computing are largely a consequence of the early stage of development of the field. There is as yet no proven working method to build a useful quantum computer, so new ideas stand a fair chance of being the winning combination. We do not even have any concrete evidence that we will be able to build a quantum computer that actually outperforms classical computers. But we also have no evidence to the contrary, that quantum computers this powerful are not possible for some fundamental physical reason, so the open question, plus the open questions such a quantum computer may be able to solve, keep the field vibrant and optimistic.

Beyond the precision quantum engineering required to construct any of the designs described thus far, there is plenty of discussion of quantum computational processes in natural systems. While all natural systems are made up of particles that obey quantum mechanics at a fundamental level, the extent to which their behavior requires quantum logic to describe it is less ubiquitous. Quantum coherences are more usually dissipated into the surrounding environment than marshaled into co-ordinated activity.

It has been suggested that the brain may exhibit quantum computational capabilities on two different levels. Firstly, Hameroff and Penrose [74] have argued that brain cells amplify quantum effects that feed into the way the brain functions. This is disputable on physical grounds when the time-scales and energies involved are considered in detail. Secondly, a quantum-like logic for brain processes has been proposed, for example, see [75]. This does not require actual quantum systems to provide it, and is thus not dependent on individual brain cells amplifying quantum effects.

Whether any biological computing is exploiting quantum logic is an open question. Since typical biological temperatures and time scales are generally not a hospitable regime for maintaining quantum coherences, we might expect any such quantum effects to be rare at the level of complex computations. Clearly quantum effects are biologically important for of basic processes such as photosynthesis, for example, and transport properties such as the conductance and coherence of single electrons through biological molecules are the subject of much current study. These are perhaps more likely to feature as problems solved by one of the first quantum simulators, than as examples of natural quantum computers.

Bibliography

Primary Literature

1. Bartlett SD, Rudolph T, Spekkens RW (2006) Reference frames, superselection rules, and quantum information. Rev Mod Phys 79:555; ArXiv:quant-ph/0610030
2. Dawson CM, Nielsen MA (2006) The Solovay-Kitaev algorithm. Quantum Inf Comp 6:81–95; The Solovay-Kitaev theorem dates from 1995, but was only partially published in pieces – this Ref. gives a more comprehensive review; ArXiv:quant-ph/0505030
3. Feynman RP (1982) Simulating physics with computers. Int J Theor Phys 21:467
4. Deutsch D (1985) Quantum-theory, the church-turing principle and the universal quantum computer. Proc R Soc Lond A 400(1818):97–117
5. Knill E, Laflamme R, Zurek W (1996) Threshold accuracy for quantum computation. ArXiv:quant-ph/9610011
6. Aharonov D, Ben-Or M (1996) Fault tolerant quantum computation with constant error. In: Proc 29 th ACM STOC. ACM, NY 176–188; ArXiv:quant-ph/9611025
7. Steane A (1996) Multiple particle interference and quantum error correction. Proc Roy Soc Lond A 452:2551; ArXiv:quant-ph/9601029
8. Shor PW (1997) Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. SIAM J Sci Statist Comput 26:1484
9. Bennett CH, Brassard G (1984) Quantum cryptography: public-key distribution and coin tossing. In: IEEE International Conference on Computers, Systems and Signal Processing. IEEE Computer Society Press, Los Alamitos, pp 175–179
10. Bennett CH, Wiesner SJ (1992) Communication via one – and two-particle operators on Einstein–Podolsky–Rosen states. Phys Rev Lett 69(20):2881–2884
11. Bennett CH, Brassard G, Crépeau C, Jozsa R, Peres A, Wootters WK (1993) Teleporting an unknown quantum state via dual classical and Einstein–Podolsky–Rosen channels. Phys Rev Lett 70:1895–1899
12. Graham-Rowe D (2007) Quantum ATM rules out fraudulent web purchases. New Sci 2629:30–31

13. Duligall JL, Godfrey MS, Harrison KA, Munro WJ, Rarity JG (2006) Low cost and compact quantum cryptography. New J Phys 8:249; ArXiv:quant-ph/0608213v2
14. Steane A (2001) Quantum computing and error correction. In: Gonis, Turchi (eds) Decoherence and its implications in quantum computation and information transfer. IOS Press, Amsterdam, pp 284–298
15. Lidar DA, Chuang IL, Whaley KB (1998) Decoherence free subspaces for quantum computation. Phys Rev Lett 81:2594–2598; ArXiv:quant-ph/9807004v2
16. Lidar DA, Whaley KB (2003) Decoherence-free subspaces and subsystems. In: Benatti F, Floreanini R (eds) Irreversible Quantum Dynamics. Lecture Notes in Physics, vol 622. Springer, Berlin, pp 83–120; ArXiv:quant-ph/0301032
17. Misra B, Sudarshan ECG (1977) The Zeno's paradox in quantum theory. J Math Phys 18:756
18. Beige A, Braun D, Tregenna B, Knight PL (2000) Quantum computing using dissipation to remain in a decoherence-free subspace. Phys Rev Lett 85:762–1766; ArXiv:quant-ph/0004043v3
19. Nielsen M, Chuang I (1996) Talk at KITP Workshop: Quantum Coherence and Decoherence, organized by DiVencenzo DP and Zurek W. http://www.kitp.ucsb.edu/activities/conferences/past/. Accessed 2 Sep 2008
20. Lomont C (2004) The hidden subgroup problem-review and open problems. ArXiv:quant-ph/0411037
21. Childs A, Eisenberg JM (2005) Quantum algorithms for subset finding. Quantum Inf Comput 5:593–604; ArXiv:quant-ph/0311038
22. Ambainis A (2004) Quantum walk algorithms for element distinctness. In: 45th Annual IEEE Symposium on Foundations of Computer Science. IEEE Computer Society Press, Los Alamitos, pp 22–31
23. Kendon V, Tregenna B (2003) Decoherence can be useful in quantum walks. Phys Rev A 67:042315; ArXiv:quant-ph/0209005
24. Richter P (2007) Almost uniform sampling in quantum walks. New J Phys 9:72; ArXiv:quant-ph/0606202
25. Richter P (2007) Quantum speedup of classical mixing processes. Phys Rev A 76:042306; ArXiv:quant-ph/0609204
26. Grover LK (1996) A fast quantum mechanical algorithm for database search. In: Proc 28th Annual ACM STOC. ACM, NY, p 212; ArXiv:quant-ph/9605043
27. Grover LK (1997) Quantum mechanics helps in searching for a needle in a haystack. Phys Rev Lett 79:325; ArXiv:quant-ph/9706033
28. Shenvi N, Kempe J, Whaley BK (2003) A quantum random walk search algorithm. Phys Rev A 67:052307; ArXiv:quant-ph/0210064
29. Bennett CH, Bernstein E, Brassard G, Vazirani U (1997) Strengths and weaknesses of quantum computing. SIAM J Comput 26(5):151–152
30. Magniez F, Santha M, Szegedy M (2003) An $o(n^{1.3})$ quantum algorithm for the triangle problem. ArXiv:quant-ph/0310134
31. Magniez F, Santha M, Szegedy M (2005) Quantum algorithms for the triangle problem. In: Proceedings of 16th ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, Philadelphia, pp 1109–1117
32. Childs AM, Cleve R, Deotto E, Farhi E, Gutmann S, Spielman DA (2003) Exponential algorithmic speedup by a quantum walk. In: Proc 35th Annual ACM STOC. ACM, NY, pp 59–68; ArXiv:quant-ph/0209131
33. Ambainis A (2003) Quantum walks and their algorithmic applications. Int J Quantum Information 1(4):507–518; ArXiv:quant-ph/0403120
34. Scott Aaronson (2005) The complexity zoo. http://qwiki.caltech.edu/wiki/Complexity_Zoo. Accessed 2 Sep 2008
35. De Raedt K, Michielsen K, De Raedt H, Trieu B, Arnold G, Richter M, Lippert Th, Watanabe H, Ito N (2007) Massive parallel quantum computer simulator. Comp Phys Comm 176:127–136
36. Cirac JI, Verstraete F, Porras D (2004) Density matrix renormalization group and periodic boundary conditions: A quantum information perspective. Phys Rev Lett 93:227205
37. Lloyd S (2000) Ultimate physical limits to computation. Nature 406:1047–1054; ArXiv:quant-ph/9908043
38. Margolus N, Levitin LB (1996) The maximum speed of dynamical evolution. In: Toffoli T, Biafore M, Liao J (eds) Physcomp96. NECSI, Boston
39. Margolus N, Levitin LB (1998) The maximum speed of dynamical evolution. Physica D 120:188–195; ArXiv:quant-ph/9710043v2
40. DiVincenzo DP (2000) The physical implementation of quantum computation. Fortschritte der Physik 48(9–11):771–783
41. Metodi TS, Thaker DD, Cross AW, Chong FT, Chuang IL (2005) A quantum logic array microarchitecture: Scalable quantum data movement and computation. In: 38th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'05). IEEE Computer Society Press, Los Alamitos, pp 305–318; ArXiv:quant-ph/0509051v1
42. Spiller TP, Munro WJ, Barrett SD, Kok P (2005) An introduction to quantum information processing: applications and realisations. Comptemporary Physics 46:407
43. Jozsa R (1998) Entanglement and quantum computation. In: Huggett SA, Mason LJ, Tod KP, Tsou S, Woodhouse NMJ (eds) The Geometric Universe, Geometry, and the Work of Roger Penrose. Oxford University Press, Oxford, pp 369–379
44. Blume-Kohout R, Caves CM, Deutsch IH (2002) Climbing mount scalable: Physical resource requirements for a scalable quantum computer. Found Phys 32(11):1641–1670; ArXiv:quant-ph/0204157
45. Greentree AD, Schirmer SG, Green F, Lloyd Hollenberg CL, Hamilton AR, Clark RG (2004) Maximizing the hilbert space for a finite number of distinguishable quantum states. Phys Rev Lett 92:097901; ArXiv:quant-ph/0304050
46. Braunstein SL, van Loock P (2005) Quantum information with continuous variables. Rev Mod Phys 77:513–578
47. Lloyd S, Braunstein SL (1999) Quantum computation over continuous variables. Phys Rev Lett 82:1784; ArXiv:quant-ph/9810082v1
48. Bartlett S, Sanders B, Braunstein SL, Nemoto K (2002) Efficient classical simulation of continuous variable quantum information processes. Phys Rev Lett 88:097904; ArXiv:quant-ph/0109047
49. Ladd TD, van Loock P, Nemoto K, Munro WJ, Yamamoto Y (2006) Hybrid quantum repeater based on dispersive cqed interactions between matter qubits and bright coherent light. New J Phys 8:164; ArXiv:quant-ph/0610154v1 doi:10.1088/1367-2630/8/9/184
50. Spiller TP, Nemoto K, Braunstein SL, Munro WJ, van Loock P, Milburn GJ (2006) Quantum computation by communication. New J Phys 8:30; ArXiv:quant-ph/0509202v3
51. Raussendorf R, Briegel HJ (2001) A one-way quantum computer. Phys Rev Lett 86:5188–5191

52. Raussendorf R, Browne DE, Briegel HJ (2003) Measurement-based quantum computation with cluster states. Phys Rev A 68:022312; ArXiv:quant-ph/0301052v2
53. Nielsen MA (2004) Optical quantum computation using cluster states. Phys Rev Lett 93:040503; ArXiv:quant-ph/0402005
54. Yoran N, Reznik B (2003) Deterministic linear optics quantum computation utilizing linked photon circuits. Phys Rev Lett 91:037903; ArXiv:quant-ph/0303008
55. Jozsa R (2005) An introduction to measurement based quantum computation. ArXiv:quant-ph/0508124
56. Kitaev YA (2003) Fault-tolerant quantum computation by anyons. Annals Phys 303:2–30; ArXiv:quant-ph/9707021v1
57. Aharonov D, van Dam W, Kempe J, Landau Z, Lloyd S, Regev O (2004) Adiabatic quantum computation is equivalent to standard quantum computation. ArXiv:quant-ph/0405098
58. Brennen GK, Pachos JK (2007) Why should anyone care about computing with anyons? Proc Roy Soc Lond A 464(2089):1–24; ArXiv:0704.2241v2
59. Aharonov Y, Bohm D (1959) Significance of electromagnetic potentials in quantum theory. Phys Rev 115:485–491
60. Young T (1804) Experimental demonstration of the general law of the interference of light. Phil Trans Royal Soc Lon, London, p 94
61. Jiang L, Brennen GK, Gorshkov AV, Hammerer K, Hafezi M, Demler E, Lukin MD, Zoller P (2007) Anyonic interferometry and protected memories in atomic spin lattices. ArXiv:0711.1365v1
62. Farhi E, Goldstone J, Gutmann S, Sipser M (2000) Quantum computation by adiabatic evolution. ArXiv:quant-ph/0001106
63. Kieu TD (2006) Quantum adiabatic computation and the travelling salesman problem. ArXiv:quant-ph/0601151v2
64. Kempe, Kitaev, Regev (2004) The complexity of the local hamiltonian problem. In: Proc 24th FSTTCS. pp 372–383; ArXiv:quant-ph/0406180
65. Kempe, Kitaev, Regev (2006) The complexity of the local hamiltonian problem. SIAM J Comput 35(5):1070–1097
66. Childs AM, Farhi E, Preskill J (2002) Robustness of adiabatic quantum computation. Phys Rev A 65:012322; ArXiv:quant-ph/0108048
67. van Dam S, Hogg, Breyta, Chuang (2003) Experimental implementation of an adiabatic quantum optimization algorithm. Phys Rev Lett 90(6):067903; ArXiv:quant-ph/0302057
68. Lloyd S (1996) Universal quantum simulators. Science 273:1073–1078
69. Berry DW, Ahokas G, Cleve R, Sanders BC (2007) Efficient quantum algorithms for simulating sparse hamiltonians. Commun Math Phys 270:359; ArXiv:quant-ph/0508139v2
70. Somaroo SS, Tseng CH, Havel TF, Laflamme R, Cory DG (1999) Quantum simulations on a quantum computer. Phys Rev Lett 82:5381–5384; ArXiv:quant-ph/9905045
71. Abrams DS, Lloyd S (1997) Simulation of many-body fermi systems on a universal quantum computer. Phys Rev Lett 79:2586–2589; ArXiv:quant-ph/9703054v1
72. Wu LA, Byrd MS, Lidar DA (2002) Polynomial-time simulation of pairing models on a quantum computer. Phys Rev Lett 89:057904. Due to a proofs mix up there is also an Erratum: 89:139901; ArXiv:quant-ph/0108110v2
73. Brown KR, Clark RJ, Chuang IL (2006) Limitations of quantum simulation examined by simulating a pairing hamiltonian using nuclear magnetic resonance. Phys Rev Lett 97:050504; ArXiv:quant-ph/0601021
74. Hameroff SR, Penrose R (1996) Conscious events as orchestrated spacetime selections. J Conscious Stud 3(1):36–53
75. Khrennikov A (2006) Brain as quantum-like computer. BioSystems 84:225–241; ArXiv:quant-ph/0205092v8

Further Reading

For those who seriously want to learn the quantitative details of quantum computing, this is still the best textbook: Nielsen MA, Chuang IL (2000) Quantum Computation and Quantum Information. CUP, Cambs

For lighter browsing but still with all the technical details, there are several quantum wikis developed by the scientists doing the research:

Quantiki http://www.quantiki.org/wiki/index.php/Main_Page specifically quantum information

Qwiki http://qwiki.stanford.edu/wiki/Main_Page covers wider quantum theory and experiments

For those still struggling with the concepts (which probably means most people without a physics degree or other formal study of quantum theory), there are plenty of popular science books and articles. Please dive in, it's the way the world we all live in works, and there is no reason not dig in deep enough to marvel at the way it fits together and puzzle with the best of us about the bits we can't yet fathom.

Quantum Computing with Trapped Ions

Wolfgang Lange
University of Sussex, Brighton, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Ion Trap Technology
Ions as Carriers of Quantum Information
Laser Cooling and State Initialization
Single-Ion Operations
State Detection of Ionic Qubits
Two-Qubit Interaction and Quantum Gates
Decoherence
Quantum Algorithms
Distributed Quantum Information with Trapped Ions
Future Directions
Bibliography

Glossary

Ion trap Device used for confining charged particles to a small volume of space. There are two types of ion

traps: in a *Paul trap*, the confinement is achieved by using an electric quadrupole radio-frequency field, while a Penning trap uses a static electric quadrupole field and a static magnetic field. For the right field parameters, the charged particles can be stored in the trapping volume permanently.

Lamb–Dicke regime Describes the conditions for strong confinement of ions. In the Lamb–Dicke regime, ions are confined to a region smaller than the wavelength of the optical transition of the ion. Another property of ions in this parameter range is that the ion's recoil energy is less than the energy of a single quantum of vibration in the trap. This suppresses spontaneous sideband transitions. The Lamb–Dicke regime is essential for the reliable operation of laser-induced quantum gates in an ion-trap quantum computer.

Qubit The unit of quantum information processing. A contraction of quantum bit, the term qubit designates quantum system with two quasi-stable states carrying the quantum equivalent of binary information ("0" or "1"). In an ion-trap quantum computer, two long-lived electronic states are employed as quantum memory. Their wavefunctions are represented here by the symbols $|g\rangle$ and $|e\rangle$. The qubit states are manipulated by means of laser pulses.

Quantum register A collection of qubits forms a quantum register. The power of a quantum computer increases exponentially with the size N of the quantum register, since in this case 2^N numbers can be processed in parallel. In ion traps, the largest quantum register implemented so far has size $N = 8$.

Normal modes The motion of ions in a linear string is strongly coupled due to their Coulomb repulsion. The system can still be described by analogy with a set of independent harmonic oscillators, when collective modes of motion are considered, in which all ions oscillate in phase and at the same frequency. These are called normal modes. A linear string of N ions has N normal modes of axial vibration. Each mode has a characteristic distribution of motional amplitudes for the ions. In the lowest frequency mode, all ions oscillate at the same amplitude, so that the string moves like a rigid body.

Phonon bus Since all ions participate in the collective motion, it can be used to transfer quantum information between ions in the string. To this end, each normal mode is considered as a quantum mechanical oscillator, in which quanta of vibrational motion (phonons) can be excited or de-excited. By coupling their excitation to transitions of the ion-qubit, the phonons serve as a quantum data bus to other ions. Reliable transfer requires that the phonons are restricted to a binary system. The state with no vibrational excitation encodes the qubit $|0\rangle$, one quantum of vibration corresponds to $|1\rangle$.

Rabi-oscillations The term describes the change of state of a two-level atom, induced by the coherent excitation with a laser beam close to resonance. After half a period (phase π), the population is completely transferred, after another half period it is returned to the initial distribution again. By choosing suitable phases, amplitudes, detunings and duration of the pulses, the individual qubits may be manipulated in a fully controlled way and quantum gates between different ions may be induced (via the phonon bus).

Coherent superposition According to the laws of quantum mechanics, a qubit can be in an arbitrary superposition of the two basis states with complex amplitudes α and β, written as $\alpha|g\rangle + \beta|e\rangle$. The relative phase is important, for example, $|g\rangle - |e\rangle$ and $|g\rangle + |e\rangle$ are distinct superposition states. As long as a well-defined phase is preserved, the qubit is in a coherent superposition. This is a precondition for quantum information processing.

Decoherence Any loss of coherence of a quantum state is called decoherence. There are many sources of decoherence. In an ion-trap quantum processor, they range from spontaneous decay of the qubit-levels to phase shifts in fluctuating ambient fields. A quantum computation can only proceed reliably while decoherence is negligible. Quantum error correction is a way to counteract decoherence.

Entanglement Entanglement is one of the most important properties of quantum systems and has no classical analogue. A composite system, for example two qubits, is entangled if the states of the individual particles cannot be separated and regarded as independent. Instead, there are strong non-local links between the components, resulting in quantum correlations and state changes of one partner upon measurement of the other. Entangled states have many applications, from spectroscopy to quantum networking. Up to eight ions have been entangled in a deterministic way.

Quantum network A system of either local or distant nodes performing quantum calculations and exchanging results via transport of ions or via photon links. Setting up the latter requires a controlled exchange of quantum data between ions and photons, providing flying qubits. The goal is distributed entanglement, for example to transfer qubits over long distances via teleportation and eventually to perform distributed quantum computation.

Definition of the Subject

In quantum computation, the classical bit as carrier of information is replaced by a two-state quantum system, the quantum bit (*qubit*). The enhanced computing power of qubits is based on two facts: (1) they can simultaneously store arbitrary superpositions of two values 0 and 1 and (2) superpositions in different qubits may be strongly correlated (entangled), even in the absence of any direct physical link. Quantum algorithms exploit these features to efficiently solve problems whose complexity makes any classical method unfeasible.

As a new paradigm of computer science, the concept of quantum computation has generated important results in complexity theory. However, from a practical point of view, the power of quantum information processing can only be unleashed if the necessary quantum hardware is available. What is required is a set of individually accessible binary quantum systems to serve as a quantum register. They must have strong externally controllable interactions between them, but at the same time no coupling to the environment. While there is no system in which these conditions are ideally fulfilled, trapped atomic ions are the closest realization. At present, they provide the most advanced implementation of quantum information processing.

This article summarizes the state-of-the-art of quantum computing with trapped ions. As shown below, all necessary components of a trapped-ion quantum computer have been demonstrated, from quantum memory and fundamental quantum logic gates to simple quantum algorithms. Current experimental efforts are directed towards scaling up the small systems investigated so far and enhancing the fidelity of operations to a level where error correction can be applied efficiently. The first task at which a quantum computer is expected to outperform a classical one is the efficient simulation of quantum systems too complex for classical treatment.

Introduction

The idea of using atomic-scale systems for information processing was first suggested by R. P. Feynman in his lecture *There's plenty of room at the bottom* [1]. In the early 1980s, Feynman and P. Benioff found that by exploiting the dynamics of quantum systems, computations could be performed far more efficiently than using a classical computer [2,3]. They argued that the difficulty of efficiently simulating quantum systems on a classical computer implied the superior power of quantum computing.

The radically novel idea was to not merely use quantum mechanics as a framework for predicting the macroscopic behavior of a system, but to manipulate individual quantum objects directly. While Feynman didn't propose a particular physical implementation, the spectacular progress achieved in atomic and optical physics laboratories around the world in preparing and manipulating single atoms and ions makes these particles an obvious choice as a qubit. Binary quantum information is typically stored in two internal electronic levels.

Owing to their electric charge, ions can be readily trapped by radio-frequency electric fields. However, Coulomb repulsion keeps the trapped ions so far apart, that any direct interaction involving their internal states is negligible. This seems to preclude logic gates for two or more qubits. However, in 1995, I. Cirac and P. Zoller found a way around this problem in a seminal paper which initiated experimental ion-trap quantum computation as an active research field [4]. They proposed to use the long-distance Coulomb repulsion between ions to couple different qubits in a linear string by exchanging single quanta of their vibration.

Trapped ions have other advantages that make them strong contenders as quantum bits. Their internal and external (motional) states can be manipulated using laser light. In atoms or ions, long-lived excited states exist, allowing for the storage and retrieval of quantum superposition states. Since the first experiment demonstrating a simple quantum gate with a single ion, impressive progress has been made in developing or adapting ion trap technologies for quantum information processing.

In this article, the range of techniques used in ion-trap quantum computing is presented, starting with the technological foundations of ion trapping (Sect. "Ion Trap Technology"). The subsequent discussion follows a list of general requirements for a system to serve as a practical device for quantum computation. It was compiled as a general guideline by D. DiVincenzo [5].

(1) The quantum system must provide well-characterized qubits forming a scalable quantum register (Sect. "Ions as Carriers of Quantum Information").
(2) It must be possible to initialize the qubits to a known state (Sect. "Laser Cooling and State Initialization").
(3) A method to efficiently measure individual qubits must exist (Sect. "State Detection of Ionic Qubits").
(4) A universal set of quantum gates is required (Sects. "Single-Ion Operations" and "Two-Qubit Interaction and Quantum Gates").
(5) The time over which quantum states lose coherence should be much longer than gate operation times (Sect. "Decoherence").

At present, trapped ion systems are the only technology, for which all five criteria have been successfully demon-

strated. Major achievements are two-qubit ion gates with a fidelity around 96%, the entanglement of up to 8 ions and a number of simple algorithms (Sect. "Quantum Algorithms").

In his article, DiVincenzo has added two more requirements in case local quantum processors are to exchange quantum information over some distance.

(6) It must be possible to interconvert the stationary qubits of the quantum computer to "flying qubits", suitable for long-distance transmission.
(7) The flying qubits must be faithfully transmitted between specified locations.

These requirements become more important, as schemes for distributed quantum computation are being developed. The obvious choice for flying qubits are photons and there are first results regarding the coupling of ions and photons (Sect. "Distributed Quantum Information with Trapped Ions").

Ion Trap Technology

Ions are particularly well suited for quantum information processing, since due to their charge, they can be confined by electromagnetic fields without affecting their internal electronic levels. Presently, all realizations of quantum information processing with ions use variants of the Paul-trap, in which the ponderomotive force of a time-dependent inhomogeneous electric field leads to stable confinement [6,7]. For a quantum register containing multiple ions, a linear trap geometry is generally chosen.

Linear Paul Trap

The linear Paul trap has its origins in the electric quadrupole mass filter [8], in which transverse confinement of ions with a certain charge to mass ratio is achieved by a radiofrequency quadrupole potential in a plane perpendicular to the axis of the device, assumed to be oriented in the z-direction:

$$\Phi(x, y, t) = (U - V\cos\Omega t)\ \frac{x^2 - y^2}{2r_0^2}\ , \tag{1}$$

where $\pm U/2$ are the dc-voltages and $\pm V/2\cos\Omega t$ the radiofrequency-voltages applied to the quadrupole electrodes at a distance r_0 from the trap center. Equation (1) represents a saddle potential which at any particular time provides confinement in one direction only, as shown in Fig. 1. By using alternating voltages and averaging over a period of the radiofrequency, stable trapping may be achieved in both transverse directions. Figure 2 indicates how the trap parameters V and U must be chosen to obtain confinement. In practice, $U = 0$ and $eV \ll m\Omega^2 r_0^2/2$ are chosen, where m and e denote mass and charge of the ion.

To generate the potential (1), four hyperbolically shaped electrodes are required. Usually these are approximated by simpler structures like rods with circular or triangular cross section, since close to the trap axis the resulting anharmonicity is negligible.

On a timescale long compared to the period of the radiofrequency, the ions move as if they are radially confined in a parabolic pseudopotential given by $\Psi = e^2|\nabla\Phi|^2/4m\Omega^2$. The motion in the radial directions is therefore harmonic, with the radial *secular frequency*

$$\omega_r \approx \frac{eV}{\sqrt{2}m\Omega r_0^2}\ . \tag{2}$$

The pseudopotential Ψ provides confinement in the radial direction only, while the motion along the z-axis is not restricted. For three-dimensional confinement, an additional static potential must be applied in the z-direction

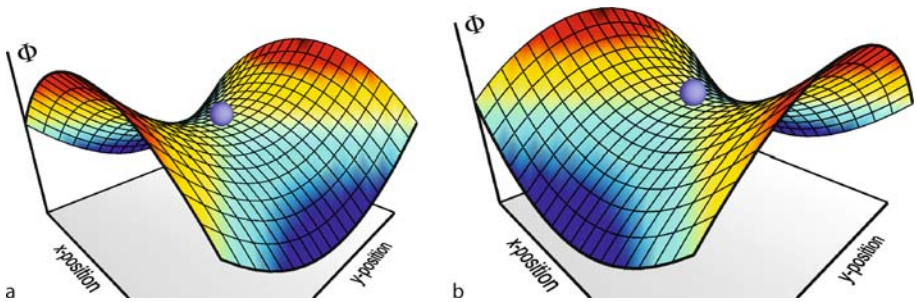

Quantum Computing with Trapped Ions, Figure 1
Two-dimensional saddle potential for a charged particle in the linear Paul trap according to Eq. (1). Due to the quadrupole shape, at any given time there is confinement only along one axis. a Potential at $t = 0$, providing y-confinement only. b Potential at $t = \pi/\Omega$, providing x-confinement only. Stable confinement in both directions is achieved by rapidly alternating the sign of the potential

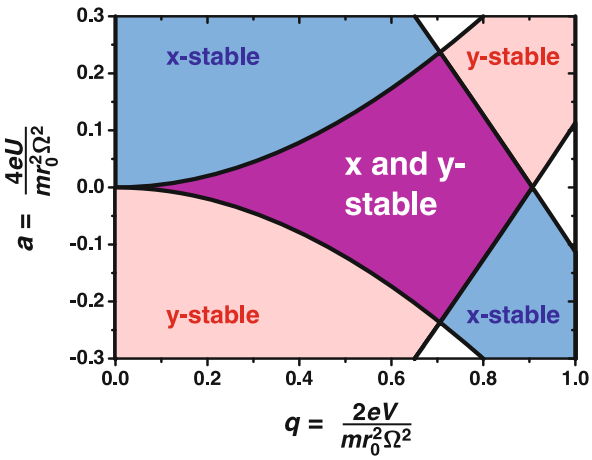

Quantum Computing with Trapped Ions, Figure 2
Stability diagram for the linear Paul trap as a function of radiofrequency amplitude V and dc-amplitude U, scaled to dimensionless parameters q and a, respectively. Confinement in both x- and y-direction is achieved in the purple region at the center, defining the operating range of the trap

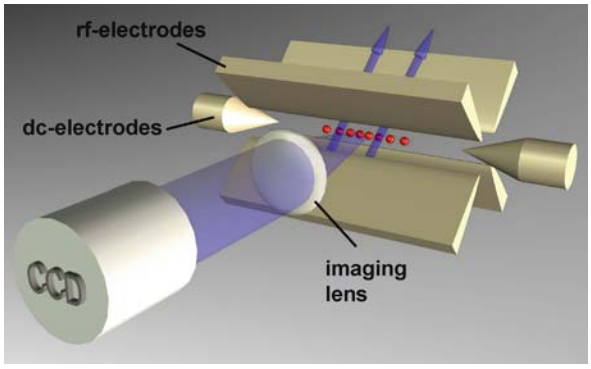

Quantum Computing with Trapped Ions, Figure 3
Schematic drawing of a linear ion trap, as used for quantum information processing. The ions are arranged in a linear string along the trap axis, held by rf- and dc electric fields. The state of the ions is detected by imaging them to a CCD camera. © R. Blatt, University of Innsbruck, Austria

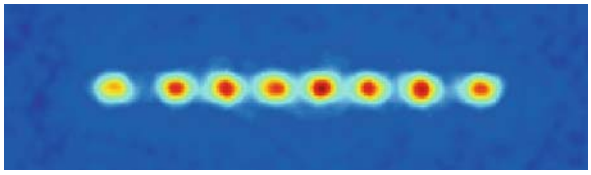

Quantum Computing with Trapped Ions, Figure 4
String of eight ions in a linear Paul trap [10]. Color indicates the intensity of the fluorescent light detected with a CCD camera. The average distance between adjacent ions is about 10 µm. See Refs. [9,11,12] for other ion crystal structures. © R. Blatt, University of Innsbruck, Austria

which is either achieved using additional electrodes at either end of the trap or by segmenting the linear rod electrodes and applying a positive dc-voltage U_z to the outer segments. This provides a static harmonic well along the z-axis which is characterized by the longitudinal trap frequency

$$\omega_z = \sqrt{\frac{2\kappa e U_z}{m z_0^2}} \,. \tag{3}$$

Here, z_0 is half the length between the axially confining electrodes and κ is a geometric factor accounting for the specific electrode configuration. The presence of the axially confining field weakens the radial confinement which is reduced to

$$\omega_r' = \sqrt{\omega_r^2 - \frac{1}{2}\omega_z^2} \,. \tag{4}$$

A sketch of a linear ion trap is shown in Fig. 3. Values for $\omega_r/2\pi$ in ion traps used for quantum information processing range from 3 MHz to 10 MHz, while typical values for $\omega_z/2\pi$ are 1 MHz to 4 MHz.

If multiple ions are confined and cooled to a sufficiently low temperature in a Paul trap (see Sect. "Vibrational Cooling of a Single Ion"), they form ordered structures [9,10,11]. If the radial confinement is strong enough ($\omega_r \gg \omega_z$), ions arrange themselves in a linear pattern along the trap axis at distances determined by the equilibrium of their mutual Coulomb repulsion and the potential providing axial confinement. Figure 4 shows an example of a string of eight $^{40}Ca^+$ ions in a linear trap. The equilibrium spacing of the ions is not uniform and must be determined numerically [12].

The number of ions N that fit on the axis of a linear ion trap is limited by the ratio of radial to axial confinement to approximately $N < 1.82\,(\omega_r/\omega_z)^{1.13}$ [14]. For larger ion numbers, the equilibrium positions no longer coincide with the z-axis of the trap. This must be avoided, since off the axis, the ions undergo *micromotion*, an oscillatory motion around their equilibrium position, driven by the trapping field at frequency Ω. Only on the trap axis, the ions are protected from micromotion, since ideally it corresponds to a node of the confining radiofrequency-field. However, micromotion can still occur if stray electric fields shift the ions off the nodal line of the radiofrequency field. Since this may result in rf-heating of the ions and precludes controlled quantum operations, stray fields must be carefully compensated [15,16].

Spherical Paul Trap

If only a single ion is to be stored, the spherical variant of the Paul trap may be used [17]. Here, the average ponderomotive force of an alternating quadrupole potential is used to provide three-dimensional confinement. The corresponding potential is

$$\Phi(x, y, z, t) = (U - V \cos \Omega t) \frac{x^2 + y^2 - 2z^2}{2r_0^2} . \quad (5)$$

The ideal electrodes are two hyperbolic endcaps along the z-axis and a hyperbolic ring in the $x - y$-plane, but simplified configurations with only two cylindrical endcaps or only one ring-electrode [18,19,20,21] are possible.

In a spherical Paul trap, the micromotion vanishes completely only at the origin, which is why it is suited only for the storage of a single ion. It was used in early studies of quantum information processing in an ion trap [22,23].

In the future, spherical Paul traps may again play an important role as single-ion microtraps, arranged in large, scalable arrays [24]. Coupling of adjacent ions would be accomplished through phase shifts, mediated by state-dependent Coulomb interaction rather than the exchange of vibrational quanta, as will be discussed in Sect. "Vibrational Coupling" for quantum information processing in linear ion strings.

Large-Scale Ion Traps

There is no fundamental limit to the length of a string of trapped ions and hence to the size of a multi-ion quantum register. However, the technical challenges of manipulating many ions in a single trapping zone are substantial. In particular, the large number of vibrational modes of a long ion crystal makes reliable transfer of quantum information difficult (see Sect. "Vibrational Coupling"). So far, quantum operations have been performed with up to eight ions in a single trap.

In order to circumvent the problems associated with long ion chains, trap architectures made up of interconnected individual trap segments have been proposed [25]. In these schemes, processing of quantum information is performed locally in special regions of the trap containing only a small number of ions, which are retrieved from and returned to memory regions. Ion transfer is achieved by changing the axially confining fields with dc-electrodes placed along the trapping zones. This type of trap is also known as a *quantum charge-coupled device* (QCCD). A sketch of a possible electrode layout is shown in Fig. 5.

An important issue for quantum computation in segmented traps is to maintain coherence of internal states during transport of the qubits. This has been demonstrated over a straight distance of 1.2 mm after separating two ions originally held in a single trap [26]. In order to extract arbitrary ions from a quantum register, junctions have to be implemented, for example, a T-junction joining two linear trap sections at an angle of 90°. A trap of this type was realized at the University of Michigan and used to swap the positions of two ions [27].

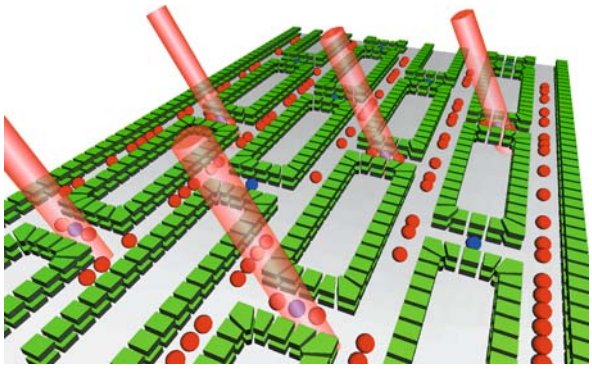

Quantum Computing with Trapped Ions, Figure 5
Sketch of an ion trap with multiple segments, providing memory regions for storing qubits (*red*) and interaction regions for logic operations, including ions to sympathetically cool the vibrational motion (*blue*). Shuttling between different locations is accomplished by suitable dc-voltages applied to the electrodes shown in *green*

Microtraps and Surface-Electrode Traps

The number of electrodes required for multi-zone traps quickly grows as more processing or memory segments are included. These devices require integrated fabrication technologies to replace the manual assembly and alignment of electrodes used in simpler traps. Two approaches have been pursued to make ion traps amenable to microfabrication.

(1) *Three-dimensional microtraps*, fabricated from a monolithic multilayer microchip [28,29]. The trapping geometry is similar to that in a macroscopic trap but minimum features are smaller and no assembly is required. A chip-trap has been fabricated from a doped gallium-arsenide heterostructure [30].
(2) *Surface-electrode traps* further simplify the fabrication of trapping structures. They are derived from three-dimensional Paul traps, but have all their electrodes moved to a single plane [31,32]. The ions are again trapped in a minimum of the pseudopotential, which is typically located several tens of μm above the sur-

face. A planar trap for single ions has been demonstrated at NIST [31].

Using microfabrication techniques for ion traps makes them suitable for miniaturization and scaling. As an additional benefit of chip-traps, microelectronics, e. g., for electrode potential control can be placed directly on the chip [33].

Penning Traps

An alternative way to trap ions in three dimensions is the Penning trap [6]. In contrast to the Paul-trap, there are no alternating fields. Stable confinement of ions is accomplished with a static magnetic field, providing a radial force in the xy-plane and a static electric field exerting a restoring force in the z-direction. While the z-motion in the Penning trap is harmonic, in the radial plane the ions orbit the center in a combination of cyclotron and magnetron motion. This makes a fixed radial arrangement of ions difficult. However, by using a rotating electric field technique, large radial ion crystals rotating at a controlled rate were produced in a Penning trap [34]. Such a collection of ions is equivalent to a quantum hard disk and has been proposed for quantum information processing [35].

Another scheme for quantum computing with Penning traps uses an array of miniature Penning traps each containing only a single ion [36]. Ions could be shuttled between individual traps using suitable arrangements of electric fields. To perform two-bit quantum gates, two ions would be combined in a single trap, forming an axial crystal, which can be manipulated by analogy with to the equivalent structure in a Paul trap. Axial two-ion Coulomb crystals have recently been observed in experiment [37].

An array of Penning traps holding single *electrons* has been proposed as a scalable quantum information processor [38]. Here, gates are implemented by controlling the Coulomb interaction with radio-frequency and microwave techniques. In the remainder of this article, Penning traps will not be further discussed, since to date, all realizations of quantum information processing have been performed in Paul traps.

Ions as Carriers of Quantum Information

In an ion-trap quantum computer, information is encoded in two internal (electronic) states of the ion. Among the large number of levels, only those with a long natural lifetime are suitable for storing quantum information. This excludes levels with electric dipole transitions to lower lying states. For easy manipulation of the qubit-transition, the levels should be coupled either by a two-photon Raman transition, an electric quadrupole or a magnetic dipole transition. Ideally, the states should be insensitive to fluctuations of external magnetic or electric fields to minimize decoherence.

Two classes of states have been used for quantum information processing with ions so far: two hyperfine levels of the ground state of an ion with nuclear spin and a combination of an electronic ground state and a low-lying metastable state.

Hyperfine State Qubits

In any isotope with non-vanishing nuclear spin, the electronic ground state energy levels are split by the interaction of the electrons' angular momentum with the nuclear magnetic moment (and, if applicable, the electric quadrupole moment). These so-called hyperfine states are coupled by magnetic dipole transitions and hence have an extremely long lifetime. Therefore, they are ideally suited for reliably storing quantum information. An example is the $^9Be^+$ ion, which was used in the first implementations of quantum information processing with ions [22]. The two hyperfine levels are distinguished by the quantum numbers $F = 1$ and $F = 2$, differing in the relative orientation of electronic angular momentum and nuclear spin.

The relevant level structure of $^9Be^+$ is shown in Fig. 6. The two basis states of the quantum bit are represented by a magnetic sublevel of each hyperfine state. A common choice is $|F = 2, m_F = 2\rangle$ and $|F = 1, m_F = 1\rangle$. While the outermost levels with $m_F = \pm F$ are most conveniently initialized, they are subject to random shifts by fluctuating magnetic fields, leading to decoherence. For a quantum memory with longer lifetime, a first-order magnetic-field independent transition at a finite magnetic bias field should be used [39].

Metastable State Qubits

Long lifetimes for ionic qubits may also be achieved by employing excited electronic states which are connected to lower lying levels only by higher order electromagnetic multipole transitions. This is the case for ions with a D-state with lower energy than the first excited P-state. It decays to the ground state via an electric quadrupole transition with a rate around $1\mathrm{s}^{-1}$. An example is $^{40}Ca^+$ [40], which is used for ion-trap quantum information processing in a number of laboratories. Here, the states $|^2S_{1/2}, m_j = 1/2\rangle$ and $|^2D_{5/2}, m_j = 1/2\rangle$ are usually chosen as basis states. The level structure of $^{40}Ca^+$ is shown in Fig. 7.

All ions investigated for quantum information processing fall in either of these two categories. An overview

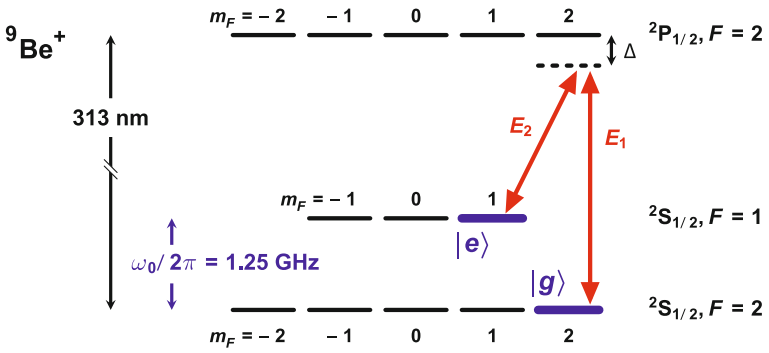

Quantum Computing with Trapped Ions, Figure 6
Qubit-implementation in $^9Be^+$. As indicated, the quantum information is stored in the two magnetic sublevels $|F = 2, m_F = 2\rangle$ and $|F = 1, m_F = 1\rangle$. In order to change the state of the qubit, Raman transitions driven by laser fields E_1 and E_2, off-resonant from the level $P_{1/2}$, are employed

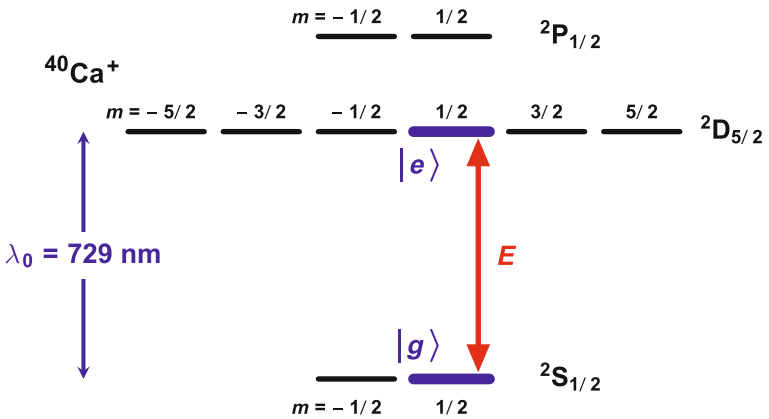

Quantum Computing with Trapped Ions, Figure 7
Qubit-implementation in $^{40}Ca^+$. Quantum information is stored in a magnetic sublevel of the ground state $|^2S_{1/2}, m_j = 1/2\rangle$ and of the metastable state $|^2D_{5/2}, m_j = 1/2\rangle$. The state of the qubit is changed by resonantly driving the quadrupole transition between S and D with a laser field E at 729 nm

of isotopes that have been used is given in Table 1. In the remainder of this article, the two physical basis states of a qubit will be designated by the symbols $|g\rangle$ for the lower (ground-) state and $|e\rangle$ for the higher (excited) state, as indicated in Fig. 6 and 7. Note that in the literature, alternative notations are used, e. g., $|\downarrow\rangle$ and $|\uparrow\rangle$ for hyperfine qubits, $|S\rangle$ and $|D\rangle$ symbolizing the orbital states involved, or the generic $|0\rangle$ and $|1\rangle$. The 2^N basis states of a quantum register with N ions are represented by analogy with single qubit-states by specifying N entries: $|x_1 x_2 x_3 \ldots x_N\rangle$, where x_i is either g or e.

Quantum Computing with Trapped Ions, Table 1
Isotopes used in investigations of quantum information processing with trapped ions

Scheme	Ion Species
Hyperfine qubit	$^9Be^+$ [22,39], $^{43}Ca^+$ [41,42], $^{111}Cd^+$ [43], $^{171}Yb^+$ [44,45], $^{25}Mg^+$ [46]
Metastable qubit	$^{40}Ca^+$ [40,41], $^{88}Sr^+$ [47,48,49,50]

Laser Cooling and State Initialization

Ionization

Before a quantum register can be initialized, the ion trap must be loaded with the desired isotope. This is most efficiently achieved by photo-ionizing an atomic beam [51,52,53]. If one of the excitation steps in multi-photon ionization is resonant, a specific isotope may be loaded, distinguished by its isotope shift [51,54]. If no suitable laser source for photo-ionization is available, electron-impact ionization may be used instead. It is not state-

selective and hence only suited for loading the most abundant species among an ion's isotopes. Due to the long trap lifetimes of many hours to days, the trap has to be loaded only infrequently.

Initialization of the Quantum Register

Any quantum computation must start from a well-defined, i. e. pure state of the quantum register. Therefore, initially all ions must be transferred to a specific internal state, typically the lower basis state, so that the register is initialized to $|ggg\ldots g\rangle$. In ions, this is achieved by the well-established method of optical pumping [55]. A laser with a suitable polarization repeatedly excites the ions to a set of states from which they eventually decay to the desired state $|g\rangle$. This state is either not coupled to the exciting laser or driven to a state whose only decay channel is back to $|g\rangle$ (cycling transition). In any case, all ions reach their initial state after a few excitation-emission cycles. In contrast to the quantum gates discussed below, the initialization of the register is irreversible.

Not all internal states are amenable to initialization by optical pumping with high fidelity. In some cases, it may be necessary to pump to an auxiliary state, which is then coherently transferred to the desired initial state by suitable laser pulses [39].

Vibrational Cooling of a Single Ion

In an ion-trap quantum processor, different ions are coupled by the Coulomb interaction via their motional degrees of freedom (see Sect. "Vibrational Coupling"). In order to transfer quantum information in this way, the ions' harmonic oscillation in the trapping potential must be cooled close to its quantum mechanical ground state. This is achieved by different stages of laser cooling, exploiting the momentum transfer between ions and light.

Initial cooling of the ions is provided by Doppler cooling [56,57,58], using an atomic transition with a natural line width Γ larger than the vibrational frequencies (ω_r or ω_z) of the ions in the trap. If the cooling laser is slightly red-detuned from resonance, the Doppler-shift from the ions' motion ensures that photons are preferably absorbed when an ion is moving towards the laser. The resulting momentum transfer, averaged over many scattering events, reduces the kinetic energy and thus the temperature of the ion. The final temperature which can be reached by this technique is given by the Doppler temperature $T_D = \hbar\Gamma/2k_B$ [59] where k_B denotes Boltzman's constant. T_D is on the order of a few millikelvin, which for typical trap frequencies corresponds to a residual excitation of one to tens of vibrational quanta.

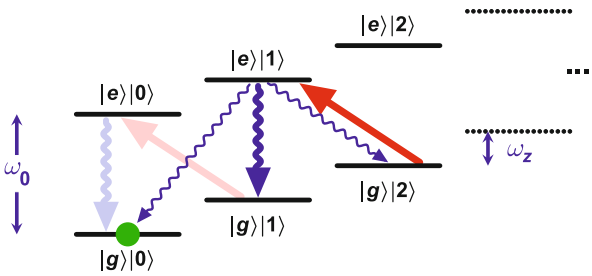

Quantum Computing with Trapped Ions, Figure 8
Resolved sideband cooling in a two-level system (states $|g\rangle$, $|e\rangle$). Associated with each level is a ladder of quantized vibrational states $|n\rangle$, labeled by their vibrational quantum number. Red sideband excitation (*red arrow*), followed predominantly by spontaneous emission on the carrier, reduces the vibrational excitation by one. After a few cycles, the vibrational ground state (*green circle*) is reached, which is not coupled to the excitation

In order to reach the vibrational ground state, an additional cooling stage must be applied. To this end, a transition is used whose Γ is smaller than the frequency of the vibration to be cooled, e. g., $\Gamma \ll \omega_z$. In this case, the ion's absorption spectrum is composed of a line at the transition frequency ω_0 and a series of sidebands at $(\omega_0 \pm n\omega_z)$ with integer n. The strength of these sidebands is given by the vibrational excitation.

Very efficient cooling is obtained by tuning a laser to the first red sideband at $\omega_0 - \omega_z$ (see Fig. 8). If the ion is well localized (Lamb–Dicke regime), subsequent spontaneous emission occurs predominantly at the carrier frequency ω_0, so that there is a net reduction of the ion's kinetic energy by $\hbar\omega_z$. After a few excitation-emission cycles, the ground state of vibration is reached with high probability.

The technique is known as sideband cooling. The ion's mean vibrational quantum number in the trapping potential is reduced to $\langle n\rangle = (\Gamma/2\omega_z)^2 \ll 1$ [60,61]. The technical challenge is to find a transition narrower than the longitudinal trap frequency ω_z. In ions with calcium-like level schemes, this is possible on the $S_{1/2} - D_{5/2}$ quadrupole transition [62]. For hyperfine-qubits, it is necessary to use a two-photon stimulated Raman transition for sideband-cooling [61].

Vibrational Cooling of Ion Strings

In an ion string, the motion of ions is strongly coupled due to their Coulomb repulsion. It is therefore more natural to consider N normal modes of motion of the string in the z-direction, rather than N individual ions. The lowest normal mode has frequency ω_z, the same as that of a sin-

gle ion. Normal modes of ion strings will be discussed in Sect. "Normal Modes".

The cooling of a normal mode proceeds by analogy with the cooling of a single ion. It is not required that all the ions in a string are interacting with the laser. Momentum transfer from the mode can already be achieved through a single ion in the string. This fact provides the possibility of using different ions for cooling and data operations. In this case, one ion species is used for sideband-cooling the collective motion of the entire string, while the remaining ions are used in the subsequent quantum processing. This method is called sympathetic cooling [63,64,65,66]. While only the one vibrational mode to be used as a quantum data-bus must be cooled to the ground state, the remaining $N-1$ vibrational modes should be cold enough not to interfere with quantum operations, ideally also residing in their ground state of motion.

When all ions in the string are in a known internal state and the vibrational modes cooled to (or close to) their ground states of motion, the quantum register is ready for quantum information processing [67].

Single-Ion Operations

After initialization, the state of each qubit may be modified by means of suitable laser pulses applied on the qubit transition. Figure 10 shows the geometry for laser excitation. An important precondition is that the ions may be addressed individually. This requires the laser beam to be focused tighter than the minimum distance between ions, given by [12]

$$s_{\min} \approx \left(\frac{2e^2}{\pi\epsilon_0 m\omega_z^2}\right)^{1/3} N^{-0.56} \tag{6}$$

for a string of N ions with mass m. For typical experimental parameters, $s_{\min}$ ranges from 4 to 10 µm. Focusing must be considerably better to minimize residual laser intensity at the position of adjacent ions and scattered light, which would lead to unwanted excitation of other ions in the string. The laser beam is switched between ions by means of an electro-optic deflector.

Single-qubit operations are best visualized in the Bloch picture. Here, a general pure state of a qubit is parametrized by two angles, a polar angle ϑ $(0 \le \vartheta \le \pi)$ and an azimuthal angle φ $(0 \le \varphi \le 2\pi)$ such that

$$|\psi\rangle = \cos\frac{\vartheta}{2}\ |g\rangle + \sin\frac{\vartheta}{2}\ e^{i\varphi}\ |e\rangle\ . \tag{7}$$

Each qubit state can be represented by a vector pointing in the direction given by (ϑ, φ) (the Bloch-vector), ending on the surface of a sphere of unit radius, the so-called

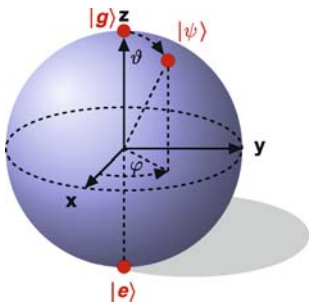

Quantum Computing with Trapped Ions, Figure 9
Bloch-sphere representation of the state of a qubit. Each point on the surface of the sphere represents a quantum state $|\psi\rangle$ of the two-level system, with the poles representing the basis states $|g\rangle$ and $|e\rangle$. A suitable laser pulse rotates the state through a polar angle θ and an azimuthal angle ϕ, so that arbitrary points on the sphere can be reached

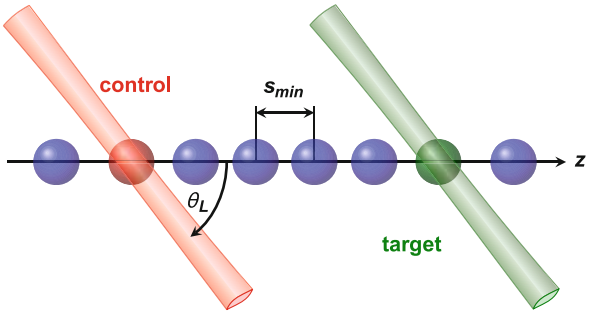

Quantum Computing with Trapped Ions, Figure 10
Geometry of selective excitation of ions in a quantum register. Shown are the two positions of a laser beam, consecutively addressing control- and a target-ion in a two-qubit quantum gate

Bloch-sphere shown in Fig. 9. The poles represent the basis states $|g\rangle$ and $|e\rangle$, while equal weight superpositions of these two states lie on the equator. Note that the phase associated with $|g\rangle$ is arbitrarily chosen to be zero. This is no longer possible in two-qubit operations, when differences between the phases of qubits are important.

In ionic qubits, the states can be manipulated by exciting the ion with electromagnetic pulses. The transition frequencies of qubits encoded in *metastable* atomic states (see Sect. "Metastable State Qubits") are in the optical domain, so that a direct transition driven by a narrowband laser may be used. In the case of $^{40}Ca^+$, this is the S-D quadrupole transition (cf. Fig. 7).

As any two-level atom resonantly excited by electromagnetic radiation, the qubit undergoes Rabi-oscillations, i. e., periodically changes between the states $|g\rangle$ and $|e\rangle$. This corresponds to a rotation of the Bloch-vector. The oscillation frequency Ω_0 (Rabi-frequency for resonant exci-

tation) is given by the time-dependent amplitude $E_0(t)$ of the exciting field and the (quadrupole) coupling strength:

$$\Omega_0(t) = E_0(t)\,\frac{\omega_L Q}{2\hbar c}\,. \tag{8}$$

$\boldsymbol{E}(t) = E_0(t)\,\boldsymbol{\epsilon}\,\cos(\boldsymbol{k}\cdot\boldsymbol{r} - \omega_L t + \phi)$ is the electric field strength of the laser with frequency ω_L and Q a component of the electric quadrupole tensor of the transition, determined by the excitation geometry and polarization $\boldsymbol{\epsilon}$ of the field. The phase ϕ of the laser determines the axis in the xy-plane, around which the Bloch-vector is rotated. For example, $\phi = 0$ leads to a rotation around the x-axis, while $\phi = \pi/2$ rotates around the y-axis. The most important parameter is the rotation angle θ. It is given by the pulse area, obtained by integrating the Rabi-frequency over the pulse duration: $\theta = \int \Omega_0(t)\,\mathrm{d}t$.

After excitation by a pulse characterized by the parameters θ and ϕ, the initial state of a qubit $c_g\,|g\rangle + c_e\,|e\rangle$ is therefore rotated according to the following transformation:

$$\begin{pmatrix} c_g \\ c_e \end{pmatrix} \rightarrow \begin{pmatrix} \cos\theta/2 & -ie^{-i\phi}\,\sin\theta/2 \\ -ie^{i\phi}\,\sin\theta/2 & \cos\theta/2 \end{pmatrix} \cdot \begin{pmatrix} c_g \\ c_e \end{pmatrix}. \tag{9}$$

Table 2 gives an overview of important single-qubit operations and the way they are implemented in an ion-trap quantum computer by choosing the pulse area θ and phase ϕ of laser pulses resonant with the qubit transition. Other single-qubit transformations can be implemented by combining rotations around different axes. An example is the Hadamard transform. The transformation matrix and its realization by X- and Y-pulses in symbolic notation is

$$\begin{pmatrix} c_g \\ c_e \end{pmatrix} \rightarrow \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} c_g \\ c_e \end{pmatrix} \tag{10}$$

Quantum Computing with Trapped Ions, Table 2

Important single qubit operations according to Eq. (9) and their realization with laser pulses. The corresponding transformations on the Bloch-sphere are often referred to as $R(\theta, \phi)$ in the literature. The symbols are used in the quantum circuits discussed below

θ	ϕ	Description	Symbol
4π	any	4π-pulse – **identity**, no change of qubit: $\lvert g\rangle \rightarrow \lvert g\rangle \qquad \lvert e\rangle \rightarrow \lvert e\rangle$	—
2π	any	2π-pulse – **sign change** of qubit-state. $\lvert g\rangle \rightarrow -\lvert g\rangle \qquad \lvert e\rangle \rightarrow -\lvert e\rangle$ This has no effect if applied on the qubit -transition since both levels experience the same phase shift. However, it is one of the most important operations in two-ion gates if applied on a transition with selective coupling to the basis states.	—
π	0	π-pulse – **NOT-gate** = qubit flip (omitting common phase factor $-i$): $\lvert g\rangle \rightarrow \lvert e\rangle \qquad \lvert e\rangle \rightarrow \lvert g\rangle$	$X\,\pi$ = X
π	$\frac{\pi}{2}$	π-pulse, bit-flip and relative sign change: $\lvert g\rangle \rightarrow \lvert e\rangle \qquad \lvert e\rangle \rightarrow -\lvert g\rangle$	$Y\,\pi$ = Y
$\frac{\pi}{2}$	0	$\pi/2$-pulse (basis states are rotated around x-axis to equally weighted superposition): $\lvert g\rangle \rightarrow \frac{\lvert g\rangle - i\lvert e\rangle}{\sqrt{2}} \qquad \lvert e\rangle \rightarrow \frac{\lvert e\rangle - i\lvert g\rangle}{\sqrt{2}}$	$X\,\frac{\pi}{2}$
$\frac{\pi}{2}$	$\frac{\pi}{2}$	$\pi/2$-pulse (basis states are rotated around y-axis to equally weighted superposition): $\lvert g\rangle \rightarrow \frac{\lvert e\rangle + \lvert g\rangle}{\sqrt{2}} \qquad \lvert e\rangle \rightarrow \frac{\lvert e\rangle - \lvert g\rangle}{\sqrt{2}}$	$Y\,\frac{\pi}{2}$
$\frac{\pi}{2}$	$-\frac{\pi}{2}$	$\pi/2$-pulse around negative y-axis – identical to combination of **Hadamard -gate** and **Z-gate** (sign change of $\lvert e\rangle$): $\lvert g\rangle \rightarrow \frac{\lvert g\rangle - \lvert e\rangle}{\sqrt{2}} \qquad \lvert e\rangle \rightarrow \frac{\lvert g\rangle + \lvert e\rangle}{\sqrt{2}}$	$\overline{Y}\,\frac{\pi}{2}$

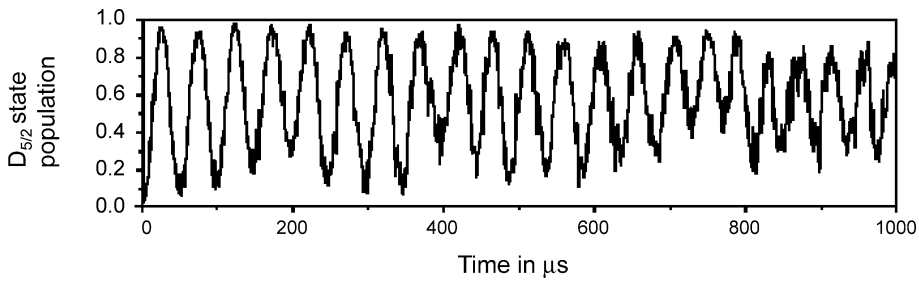

Quantum Computing with Trapped Ions, Figure 11
Rabi-oscillations of a single $^{40}Ca^{+}$-ion on the blue sideband starting in the motional ground state. The Rabi-frequency is given by Eq. (17) instead of Eq. (8), but the flopping between two states is analogous. The decay of contrast is a measure of decoherence in the system (Sect. "Decoherence"). In the example shown, coherence is maintained for up to 1 ms [62]

Pulses around the z-axis, too, are most conveniently excited by a combination of pulses around the x- and y-axes. An example is the realization of a π-rotation around the z-axis by three pulses:

$$\begin{pmatrix} c_g \\ c_e \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} c_g \\ c_e \end{pmatrix} \tag{11}$$

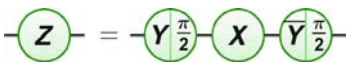

For the second class of qubits, *hyperfine qubits*, transition frequencies between the basis states are several GHz and hence require microwave pulses if driven directly [68]. Due to the large wavelength of microwaves, different ions in a quantum register can only be addressed individually if their transition frequencies are split, for example by the Zeeman-effect when using a strong magnetic field gradient.

An alternative, which has been used in most implementations of hyperfine qubits to date, is to drive the qubit using a two-photon stimulated Raman transition involving a third electronic level, connected to the qubit states through an electric dipole transition. In Fig. 6, the intermediate state is $^2P_{1/2}$. By detuning the driving fields from resonance by an amount Δ which is large compared to the inverse lifetime of the auxiliary state, excitation of this state is largely avoided [61]. The advantage is that individual ions may be addressed with the Raman beams.

The effective Rabi-frequency for a Raman-transition, to be used instead of expression (8), is

$$\Omega_0 = \frac{\Omega_1 \Omega_2}{\Delta}, \qquad \Omega_i = E_{0i} \frac{\boldsymbol{\mu}_i \cdot \boldsymbol{\epsilon}_i}{2\hbar} \tag{12}$$

where Ω_i is the Rabi-frequency for transition i and μ_i the corresponding electric dipole moment. The qubit is resonantly excited if the difference frequency $\omega_{L1} - \omega_{L2}$ of the two lasers (which replaces ω_L in the direct driving scheme) is tuned to the qubit frequency ω_0. The role of the $\boldsymbol{k}$-vector in the case of direct driving is played by the difference $\boldsymbol{k}_1 - \boldsymbol{k}_2$ and the relevant phase determining the single-ion dynamics is given by the difference of the phase constants of the two laser beams, $\phi = \phi_1 - \phi_2$. With these modifications, Eq. (9) applies.

The ion dynamics is more complicated if their vibration in the trapping potential is taken into account. This is the basis of two-bit quantum gates in ion traps discussed in Sect. "Two-Qubit Interaction and Quantum Gates". Generally, excitation of a transition on a vibrational sideband (see Fig. 8) leads to Rabi-oscillations with a modified Rabi-frequency.

By measuring the state of the qubit as a function of the length of the Rabi pulses (and hence the pulse area θ), Rabi-oscillations can be observed experimentally. An example is shown in Fig. 11. The decay of the contrast of the oscillations is an indication of the loss of coherence between the qubit levels. This can be used to determine the decoherence-time of quantum memory (see Sect. "Decoherence").

State Detection of Ionic Qubits

At the end of any data processing, a quantum computer must be subjected to a measurement of its register, in order to determine the result of the calculation. It is important for the success of the quantum computation to achieve a high efficiency readout. One of the greatest advantages of ion-trap quantum processors is that the state of each qubit may be determined with almost 100% efficiency. This is in contrast to other carriers of quantum information such as photons or nuclear spins in a bulk sample.

Quantum Jump Detection

The quantum state of an ionic qubit is measured optically by monitoring the fluorescent light emitted upon selective

laser excitation of one basis state of the qubit on a strong electric dipole transition. The population in the other basis state is either not affected by the probe laser due to selection rules or has previously been transferred (shelved) to another state inaccessible to the probe. The method is therefore also known as *electron shelving* [69].

The crucial point of the procedure is that the fluorescence is emitted on a transition back to the original qubit-level (cycling transition), so that the cycle of excitation and emission of fluorescent light can be repeated indefinitely. In practice, transitions to levels outside the cycle occur, though at a small rate. The number of photons extracted from a single qubit can reach $N_{\mathrm{ph}} \approx 10^6$ [70]. Even for a small photon detection efficiency η_{d}, the number of detected photons is large and can be distinguished from the other, non-fluorescing level with almost 100% certainty [71]. The probability for a false interpretation of a zero-photon signal is only $\exp(-\eta_{\mathrm{d}}\, N_{\mathrm{ph}})$. Originally, the method has been used to detect quantum jumps between two atomic levels by observing the intermittent fluorescence on a transition coupled to only one of the levels [69,72], hence the name *quantum jump detection*. It is now employed as the standard readout procedure in quantum information processing with ions.

Figure 12 shows the cycling transitions used in two different qubit systems. Additional repumping lasers may be required to avoid population trapping in uncoupled levels. An example of typical photon counting statistics in the two qubit states is shown in Fig. 13. By exciting the entire ion string simultaneously and imaging the fluorescent light with a CCD camera (cf. Fig. 3), the state of the entire quantum register can be read out simultaneously.

According to the laws of quantum mechanics, a single measurement cannot retrieve the complete information on the quantum state, ruling out single-shot detection of arbitrary superpositions and entanglement between different qubits. Rather, a projection to the basis states is obtained with a probability given by the magnitude squared of the corresponding coefficient. Correlations between qubits are lost in this way. Alternatively, single qubits might be subjected to a state rotation (see Sect. "Single-Ion Operations") prior to measurement, mapping coherent superpositions of a qubit's states onto the basis states, which then are probed as described above. A useful transformation is achieved with $\pi/2$-pulses.

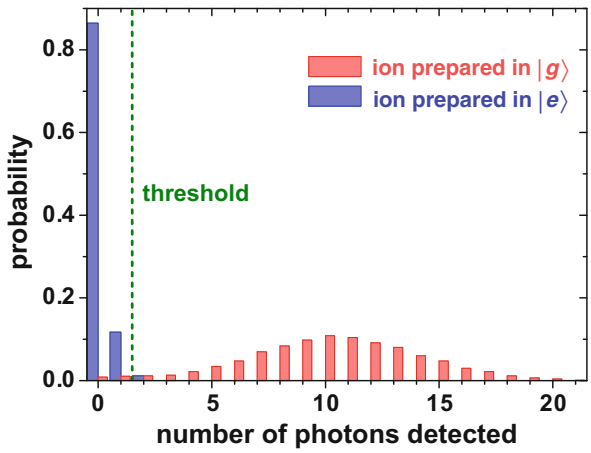

Quantum Computing with Trapped Ions, Figure 13
Histogram of detected photons after an ion is prepared in either the $|g\rangle$ or the $|e\rangle$ state. In correspondence with Fig. 12 $|g\rangle$ designates the fluorescing, while $|e\rangle$ is not coupled to the probe laser and scatters practically no photons. Using the indicated threshold to determine the state of the ion, a high detection fidelity is achieved (97.9% in the case of Yb shown above [45])

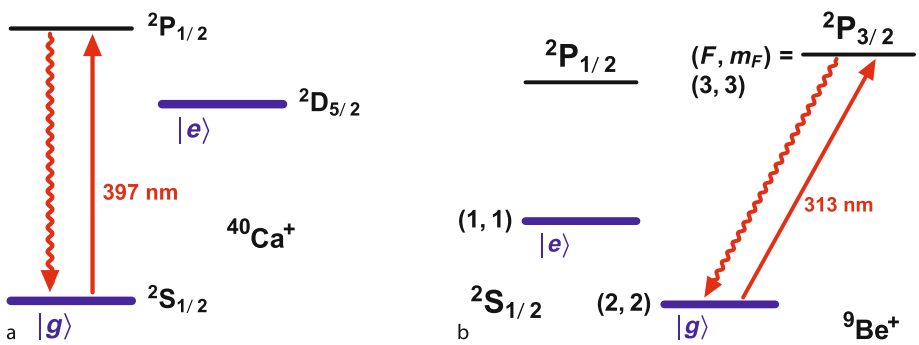

Quantum Computing with Trapped Ions, Figure 12
Transitions used in the electron shelving detection of the state of an ion-qubit. Only when the ion is in state $|g\rangle$, excitation occurs (*red arrow*) and strong fluorescence back to the original state (*wavy line*) is observed. a Relevant levels in $^{40}Ca^+$ b Relevant levels in $^9Be^+$

If correlations between different qubits are of interest, e. g., entanglement of two qubits, the two-ion gates described in Sect. "Two-Qubit Interaction and Quantum Gates" may be used to disentangle the ions first and measure the qubits separately. For a more complete characterization of a quantum operation, methods of quantum state tomography must be used [73]. By preparing a certain quantum result repeatedly and averaging over measurements of 3^N different observables (for an N-ion register), the density matrix of the system may be reconstructed, providing full information of the quantum state [74].

Two-Qubit Interaction and Quantum Gates

The most important requirement for any quantum computer is the implementation of gates, logically connecting the quantum states of two or more qubits in a coherent way. Trapped ions in a linear trap are several μm apart and therefore do not interact directly, for example through their dipole moments. Instead, the interaction is established indirectly, based on two principles: (1) Due to the long-range Coulomb repulsion between ions, their motion is strongly coupled, leading to collective vibrations of the ion-string. (2) By driving the qubit transition with a laser tuned to a motional sideband, the internal state of any ion and the collective vibration of the entire string may be coupled.

Using the collective vibration to logically connect different ions requires introducing a new qubit to the system, stored in the motional quantum state of the ion string. Since all ions participate in the vibration, it constitutes a *quantum data bus* for the entire register.

Normal Modes

The description of the collective vibration of a linear string of ions in the trapping potential is based on the concept of normal modes. In a normal mode, all ions oscillate in phase at the same frequency. A single normal mode can provide the bus for the transfer of quantum information between different ions. A linear string of N ions has N normal modes of collective axial vibration. The two with the lowest vibrational frequencies are of particular significance.

The fundamental mode has a frequency ν equal to the axial frequency of the trap ($\nu = \omega_z$) and is known as the center-of-mass (COM) mode. Here, all ions move synchronously, i. e., in the same direction and by equal amounts, so that the relative positions of the ions stay the same. It has the advantage that all ions couple to the motion with the same strength. The disadvantage is that it also couples strongly to fluctuating ambient electric fields, leading to large heating rates (see Sect. "Decoherence"). Therefore, the next higher normal mode is often preferable. It is known as the stretch- or breathing mode, since the ions on opposite sides of the center move in opposite directions with an amplitude proportional to their distance from the center. Its frequency is $\nu = \sqrt{3}\,\omega_z$. The motion of a string of four ions in the COM and stretch-mode is indicated in Fig. 14.

The higher order modes correspond to a more complicated motion of the ions. Their exact frequencies depend on the number of ions in the string and must be calculated numerically [12]. The vibrational excitation spectrum of an ion string becomes increasingly complex at higher ion number N, since not only the number of normal modes increases, but also oscillation at the sum- and difference-frequencies contributes. This makes quantum operations based on the selective coupling of ions to one particular mode difficult in long strings.

In the limit of low vibrational excitation, each mode must be treated as a quantum mechanical harmonic oscillator, the state of which can only be changed in units of single vibrational quanta of energy $\hbar\nu$ (phonons), where ν is the angular frequency of this particular mode. For quantum computation, one collective normal mode is selected for the exchange of information, while the remaining $N-1$ modes are treated as spectator modes, which should remain unexcited during the calculation.

Vibrational Coupling

In order to use the motion of ions for information exchange, there must be a coupling between the internal state of an ion and the vibration. This is achieved by detuning the exciting laser from the ion's resonance at ω_0 to either the red or the blue vibrational sideband of the desired mode, i. e., to a frequency $\omega_0 \pm \nu$.

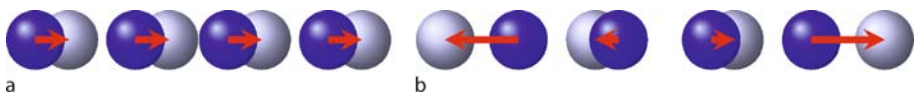

Quantum Computing with Trapped Ions, Figure 14
The two lowest normal modes of a string of four ions. The positions of the ions are shown at two different times. Arrows indicate the motion of the ions. a COM-mode at $\nu = \omega_z$; b stretch-mode at $\nu = \sqrt{3}\,\omega_z$

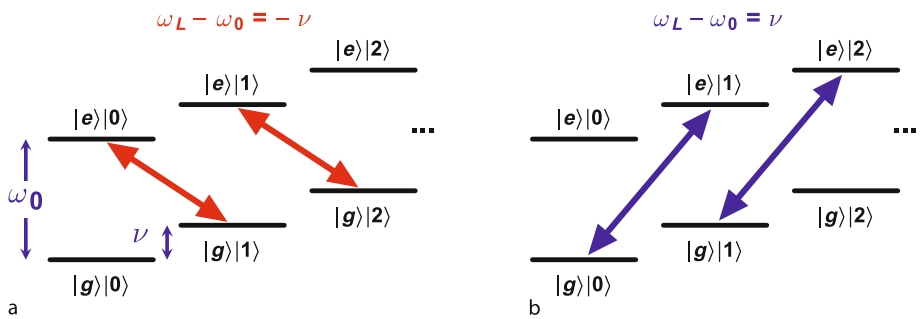

Quantum Computing with Trapped Ions, Figure 15
Vibrational structure of the qubit states and transitions for excitation on a the first red sideband and b the first blue sideband. The three lowest vibrational levels for each qubit state are shown, displaced horizontally according to the number of phonons for clarity

Assuming that the vibrational mode is in a state with phonon number n, red sideband excitation drives the transitions

$$|g\rangle\,|n\rangle \longleftrightarrow |e\rangle\,|n-1\rangle \qquad (n = 1, 2, \ldots)\,, \tag{13}$$

while blue sideband excitation drives the transitions

$$|g\rangle\,|n\rangle \longleftrightarrow |e\rangle\,|n+1\rangle \qquad (n = 0, 1, \ldots)\,. \tag{14}$$

Here, the first symbol represents the ion's state and the second the state of the vibrational mode, with $|n\rangle$ corresponding to a number state with n phonons in the mode. The transitions corresponding to (13) and (14) are depicted in Fig. 15. In each case, a change of the state of the ionic qubit coincides with a change of the phonon number, i. e., the state of the data bus.

By analogy with the resonant case discussed in Sect. "Single-Ion Operations", the evolution of the system for sideband excitation is characterized by an oscillatory change of populations at a modified Rabi-frequency. The probability of sideband transitions scales with a factor of η, the so-called Lamb–Dicke parameter. It is given by

$$\eta = \frac{2\pi\, a_0}{\lambda}\,, \tag{15}$$

where a_0 is the size of an ion's wavepacket and λ the wavelength of the transition. For the stretch-mode of a two-ion crystal, for example, the Lamb–Dicke parameter in the ground-state is given by

$$\eta = \pm\frac{1}{\sqrt[4]{12}}\,\frac{\omega_L}{c}\,\cos\theta_L\,\sqrt{\frac{\hbar}{2m\omega_z}}\,, \tag{16}$$

where θ_L is the angle between the laser beam and the trap axis (cf. Fig. 10). The Rabi-frequency for the first sideband transitions (13) and (14) depends on η and the number of excited vibrational quanta:

$$\begin{array}{lrcl} \text{Red sideband:} & \Omega^-(n) & = & \Omega_0\,\eta\,\sqrt{n} \\ \text{Blue sideband:} & \Omega^+(n) & = & \Omega_0\,\eta\,\sqrt{n+1}\,, \end{array} \tag{17}$$

where Ω_0 is the Rabi-frequency for resonant excitation from Eq. (8) or Eq. (12).

SWAP Gate

A two-qubit gate that illustrates the use of sideband excitation in a system with one ionic and one vibrational qubit is the SWAP-gate. In order to restrict the infinite ladder of vibrational states to a binary system, the qubit is represented by the two lowest vibrational states $|0\rangle$ and $|1\rangle$. The computational space is then spanned by the four basis states

$$|g\rangle\,|0\rangle\,,\,|e\rangle\,|0\rangle\,,\,|g\rangle\,|1\rangle \text{ and } |e\rangle\,|1\rangle\,, \tag{18}$$

i. e. the first two pairs of levels in Fig. 15. Assuming that the vibrational qubit is initially in the state $|0\rangle$ (after sideband cooling), an arbitrary superposition state $\alpha\,|g\rangle + \beta\,|e\rangle$ of the ion can be swapped to the vibrational mode by applying a π-pulse on the first red sideband (Fig. 16). With the state $|g\rangle\,|0\rangle$ not coupled to the red sideband, the mapping obtained is

$$(\alpha|g\rangle + \beta|e\rangle) \otimes |0\rangle \longrightarrow |g\rangle \otimes (\alpha|0\rangle + \beta|1\rangle)\,. \tag{19}$$

This operation is the basis for making the quantum state of an ion in a string accessible to any other ion sharing the collective motion. Two-ion gates may be realized by combining the above SWAP operation with a gate for the vibrational mode and a second ion.

The transformation described by Eq. (19) requires that the vibrational mode isn't excited initially. In particular, if

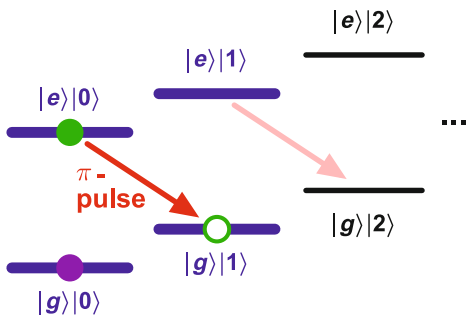

Quantum Computing with Trapped Ions, Figure 16
Mapping of a qubit from an ion to the vibrational mode using a π-pulse on the red sideband

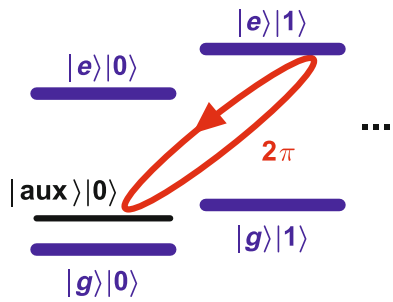

Quantum Computing with Trapped Ions, Figure 17
Controlled Z-gate in the Cirac–Zoller scheme. The laser pulse couples the auxiliary level to the state $|e\rangle\,|1\rangle$, leaving all other levels unaffected. Using a 2π-pulse results in a sign change of the $|e\rangle\,|1\rangle$-component

state $|e\rangle\,|1\rangle$ is exposed to a red sideband pulse, a transition to $|g\rangle\,|2\rangle$ would occur, which is outside the computational subspace. More elaborate pulse sequences are required to keep the ion dynamics within the space spanned by the four states of Eq. (18). A SWAP-gate that preserves the computational basis was realized by pulses of blue sideband excitation [75].

Cirac–Zoller Gate

The coupling of two ions in a string through their collective external motion was first proposed in 1995 in a seminal paper by Cirac and Zoller [4]. At the heart of their proposal is a two-qubit *controlled Z-gate* with one qubit stored in the vibrational state of the COM-mode and the other stored in the internal state of an ion. The controlled Z-gate is the simplest universal quantum gate, from which arbitrary logic gates may be built in conjunction with single qubit operations. The vibrational state controls the transformation of the ion state to which the Z-gate of Eq. (11) is applied if and only if the vibration qubit is in state $|1\rangle$. Alternatively, this can be expressed as the wave function $|\psi\rangle$ of the system acquiring a change of sign if both input qubits are in the excited state ($|1\rangle$ and $|e\rangle$) and is left unchanged in all other cases. It corresponds to the following truth table.

input	output
$\|0\rangle\|g\rangle$	$\|0\rangle\|g\rangle$
$\|0\rangle\|e\rangle$	$\|0\rangle\|e\rangle$
$\|1\rangle\|g\rangle$	$\|1\rangle\|g\rangle$
$\|1\rangle\|e\rangle$	$-\|1\rangle\|e\rangle$

As shown in Tab. 2, a sign change, is achieved by applying a 2π-pulse to the internal states of the ion. In the Cirac–Zoller scheme, the required conditioning on the state $|e\rangle$ of the ion is obtained by using a transition which couples exclusively to the upper internal state of the ion. This requires an auxiliary electronic level, e. g., another Zeeman sublevel of the ground state. Conditioning on the vibrational state $|1\rangle$ is achieved by tuning to the first blue vibrational sideband, which induces a transition to a lower state only if at least one vibrational quantum is present. The scheme of this controlled Z-gate is illustrated in Fig. 17.

Another important gate may be implemented by combining the controlled Z-gate with single qubit rotations. By applying a resonant $\pi/2$-pulse to the ion (changing the electronic states but leaving the vibration unaffected) before the controlled Z-gate and an inverse $\pi/2$-pulse after, a controlled NOT (CNOT) gate is realized, in which the target bit (carried by the electronic state) is flipped depending on the state of the control bit (represented by the vibrational state). Its truth table is

input	output
$\|0\rangle\|g\rangle$	$\|0\rangle\|g\rangle$
$\|0\rangle\|e\rangle$	$\|0\rangle\|e\rangle$
$\|1\rangle\|g\rangle$	$\|1\rangle\|e\rangle$
$\|1\rangle\|e\rangle$	$\|1\rangle\|g\rangle$

The $\pi/2$-pulses effectively change the computational basis in which the controlled Z-gate is performed.

The described CNOT-gate for the internal and the vibrational state of a single ion was demonstrated in 1995 at NIST in Boulder [22] with $^9Be^+$, the first implementation of an ion-trap quantum gate. The auxiliary state $|\text{aux}\rangle = |S_{1/2}, F = 2, m_F = 0\rangle$ was chosen. The experimental results for different combinations of input states are shown in Fig. 18, confirming the realization of the above truth table.

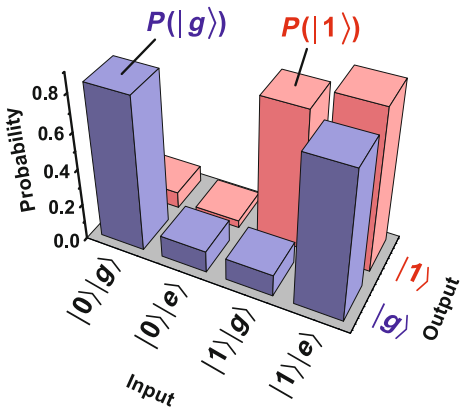

Quantum Computing with Trapped Ions, Figure 18
Experimental verification of the truth table of a CNOT quantum gate, starting from the four combinations of basis states indicated on the input axis [22]. The bars represent the measured probability for being in the lower atomic state $|g\rangle$ (blue) and in the first excited vibrational state $|1\rangle$ (red). The ionic qubit is flipped if the vibration is initially in state $|1\rangle$, while the vibrational state itself is unchanged

In the original scheme of Cirac and Zoller, the qubits are stored in the electronic states of two ions in a string, while the center-of-mass motion of the string is used as a data bus only during the gate operation. Therefore, before applying the controlled Z-gate to the target ion, the state of the first ion (control qubit) must be transferred to the vibrational motion by means of the SWAP operation described in Sect. "SWAP Gate". After completion of the gate, the state of the control ion must be restored by mapping the state of the vibrational mode back to the first ion.

The implementation of the full Cirac–Zoller gate with qubits stored in different ions has been accomplished recently in Innsbruck [76]. In this experiment, a simplified procedure for realizing the controlled Z-gate was used, avoiding the need for an auxiliary level. In principle, a controlled Z-operation can be accomplished by applying a 2π-pulse on the blue sideband of the qubit transition. However, Eq. (17) implies that state $|g\rangle|1\rangle$ has a blue sideband Rabi-frequency $\sqrt{2}$ times higher than $|g\rangle|0\rangle$, so that this state experiences a $2\sqrt{2}\pi$ pulse. A 2π-pulse for all levels (except $|e\rangle|0\rangle$, which does not couple to blue sideband excitation) can be achieved by using a transition composed of four pulses, rotating the state of the target ion with different angles and around different axes [77].

The dynamics of this gate is illustrated in Fig. 19, showing the levels coupled by effective 2π-pulses. Note that there is no excitation of levels outside the computational subspace, since 2π rotations do not change the occupation of the states involved. The experimentally determined truth table is shown in Fig. 20. The corresponding fidelity of the gate operation is between 70% and 80%.

Geometric Gates

The direct coupling of qubits and vibration in the Cirac–Zoller gate requires the bus-mode to be in the quantum mechanical ground state, a condition which is not easy to satisfy for long times in the presence of motional heating. In addition, individual addressing of control and target ion is needed. In order to relax these requirements, a different type of gate was proposed, based on electronic-state-dependent motional displacements [78,79,80,81,82,83]. Being simpler and less error-prone, the technology offers high fidelity gate operation.

In a geometric gate, the state of the harmonic oscillator corresponding to the vibrational bus-mode of the ions is displaced in phase-space (i. e. its position coordinate $\langle z\rangle$ and momentum $\langle p\rangle$ are changed), conditioned on the internal state of the qubits involved. The displacement is applied in such a way that it follows a closed loop in phase space, returning the vibration to its original state. For a detuning $\delta = \omega_L - \omega_0$ between the effective frequency difference ω_L of the Raman excitation and the qubit transition frequency, this occurs after a time τ given by $(\nu - \delta)\tau = 2\pi m$ for integer m. At this time, any entanglement between the bus and the qubits vanishes. Therefore, the only effect of the operation is that the system acquires a phase ϕ (geometric phase), equal to the area in phase space enclosed by the trajectory of the displacement, as shown in Fig. 21 [82]:

$$\phi = \frac{1}{\hbar}\int_{\text{area}} \mathrm{d}z\ \mathrm{d}p = 2\pi m\left(\frac{\eta\Omega}{\nu-\delta}\right)^2 . \tag{20}$$

The phase ϕ has properties which make it particularly suitable for a quantum gate:

(1) It depends on the internal state of the two qubits involved (e. g., ϕ is non-zero if and only if the two qubits are different);
(2) It is independent of the initial state of the bus-mode, as long as excursions stay within the Lamb–Dicke regime (i. e. are smaller than the wavelengths of the transitions involved). Therefore, the phase of the system is much less susceptible to motional heating.

Insensitivity to the bus-mode excitation could also be reached by making the detuning $\nu - \delta$ from single-photon resonance large [79]. However, this leads to an unfavorably slow evolution of the interaction between the qubits.

The selective displacement of the bus-mode for certain qubit states is achieved through optical forces exerted by

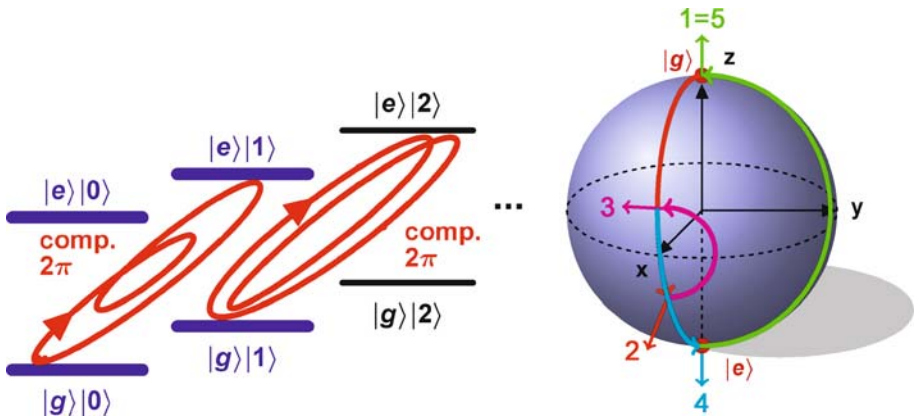

Quantum Computing with Trapped Ions, Figure 19
Controlled Z-gate with composite pulses. Using a sequence of four blue sideband pulses, all ions in $|g\rangle\,|0\rangle$, $|g\rangle\,|1\rangle$ and $|e\rangle\,|1\rangle$ experience a 2π-rotation, while $|e\rangle\,|0\rangle$ is unchanged. This corresponds to a controlled Z-gate, up to an overall minus sign. The drawing on the right shows the motion of the Bloch vector of a target qubit starting and ending in the state $|g\rangle\,|0\rangle$, picking up a phase shift of π

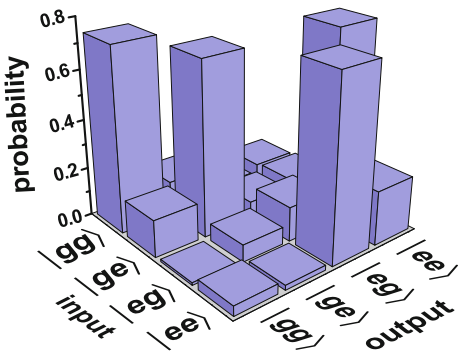

Quantum Computing with Trapped Ions, Figure 20
Experimentally observed truth table of the Cirac–Zoller CNOT derived from measurement of the joint probabilities of the two qubits [76]

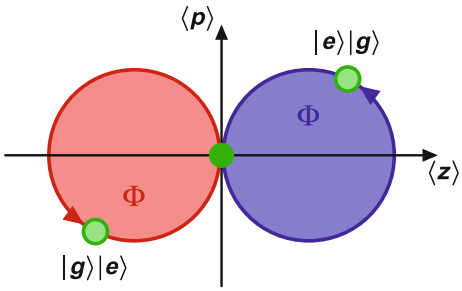

Quantum Computing with Trapped Ions, Figure 21
Phase-space representation of the bus-mode excitation during a geometric gate. The blue and the red circle are the trajectories for internal states $|e\rangle|g\rangle$ and $|g\rangle|e\rangle$, respectively. In each case, a phase ϕ is acquired

suitable Raman pulses. Two different methods have been applied, resulting in different qubit dynamics and therefore different gates.

Mølmer–Sørensen or *xy*-Gate In this case, bichromatic Raman beams are used. They provide two equally strong effective fields, with opposite detunings $\pm\delta$ from the red and the blue vibrational sideband of the qubit-transition. This drives a two-photon transition, corresponding to the two-qubit interaction Hamiltonian

$$H_I = -\frac{\hbar}{2}\frac{(\eta\Omega)^2}{\nu-\delta}(\hat{\sigma}_x \otimes \hat{\sigma}_x)\,, \tag{21}$$

where $\hat{\sigma}_x = (|g\rangle\langle e| + |e\rangle\langle g|)/\sqrt{2}$ is a Pauli spin-matrix. In the Bloch-picture, the Hamiltonian (21) provides a synchronous rotation of the internal states of both qubits around the x-axis (cf. Fig. 9). Alternatively, a rotation around the y-axis could have been chosen, hence the name xy-gate. A gate based on this type of interaction was first implemented in an experiment generating entangled states $(|gg\rangle + i|ee\rangle)/\sqrt{2}$ of two qubits and $(|gggg\rangle - i|eeee\rangle)/\sqrt{2}$ of four qubits, starting from a string of ions initialized in the internal ground state [84,85]. It has recently been used for the entanglement of magnetic-field insensitive qubit-states [86,87].

***z*-Gate** This gate is similar to the xy-gate, except that different basis states are affected by the interaction. Here, the frequency difference ω_L of two Raman beams is chosen to lie close to the vibrational frequency ν. In this case, displacement is provided by the optical dipole force from the

Raman beams, which is different for each qubit-state. If the spacing between the ions, as well as polarization and detuning of the Raman-beams are chosen correctly, there will be a vibrational displacement only if the ions are in different states. Again, for a closed loop of the displacement in phase space, the vibrational state decouples from that of the qubits, and the state of the system acquires a phase shift ϕ. The Hamiltonian corresponding to this interaction is similar to Eq. (21), except that the states of the qubits are rotated around the z-axis of the Bloch-sphere. The transformations of the basis state for the two types of geometric gates discussed are summarized in the following table:

input	z-gate	MS-gate
$\|g\rangle\,\|g\rangle$	$\|g\rangle\,\|g\rangle$	$(\|g\rangle\,\|g\rangle + ie^{-i\phi}\,\|e\rangle\,\|e\rangle)/\sqrt{2}$
$\|g\rangle\,\|e\rangle$	$e^{i\phi}\,\|g\rangle\,\|e\rangle$	$(\|g\rangle\,\|e\rangle + ie^{-i\phi}\,\|e\rangle\,\|g\rangle)/\sqrt{2}$
$\|e\rangle\,\|g\rangle$	$e^{i\phi}\,\|e\rangle\,\|g\rangle$	$(\|e\rangle\,\|g\rangle + ie^{i\phi}\,\|g\rangle\,\|e\rangle)/\sqrt{2}$
$\|e\rangle\,\|e\rangle$	$\|e\rangle\,\|e\rangle$	$(\|e\rangle\,\|e\rangle + ie^{i\phi}\,\|g\rangle\,\|g\rangle)/\sqrt{2}$

A z-gate has been realized experimentally with two $^9Be^+$-ions [88], interacting via the stretch-mode. The entangled state $(|00\rangle\text{-i}|11\rangle)/\sqrt{2}$ was produced with a fidelity of 97%. The high fidelity confirms the benefit of reduced requirements regarding thermally excited motion and addressing of the ions in a geometric gate. The z-gate was used to generate and investigate entanglement between a single $^{40}Ca^+$-ion and its vibrational motion [89], as well as two $^{40}Ca^+$-ions [90].

CNOT-Wavepacket Gate

A different route to a simplified CNOT-gate was proposed by Monroe and coworkers [91]. In contrast to gates of the Cirac–Zoller type, which use sideband transitions for coupling internal and motional states, the gate of Monroe et al. uses pulses tuned to resonance with the qubit transition. The necessary conditioning on the state of the vibrational bus-mode is obtained by taking into account the higher-order dependence of the Rabi-frequency on the Lamb–Dicke parameter η from Eq. (15) . In fact, if the extension of the wavepacket corresponding to the ion is not small compared with the wavelength of the qubit transition, i. e., $\eta \gtrsim 1$, the system is no longer in the Lamb-Dicke regime and the resonant Rabi-frequency becomes

$$\Omega(n) = \Omega_0\, e^{-\eta^2/2}\, L_n(\eta^2)\,, \tag{22}$$

where Ω_0 is the Rabi-frequency in the Lamb–Dicke limit given by either Eq. (8) or Eq. (12) and $L_n(x)$ is the Laguerre polynomial of order n.

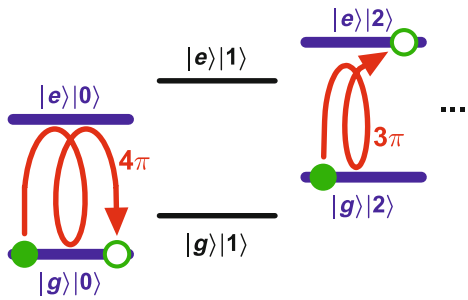

Quantum Computing with Trapped Ions, Figure 22
CNOT-gate with control qubit stored in the motional ground state ($|0\rangle$) and second excited state ($|2\rangle$) of the bus-mode. A laser pulse drives a 4π or a 3π transition, returning the ion to its initial electronic state or toggling its state when the bus is in state 0 or 2, respectively [92]

By adjusting η, the transition can be driven by a single pulse of a suitable length τ in such a way that, for example, the phase change of the system is $\theta = 4\pi$ if the vibration is in state $|0\rangle$ and $\theta = 3\pi$ if it is in state $|2\rangle$ (note that this requires using the basis states $|0\rangle$ and $|2\rangle$ for the bus, replacing the usual $|0\rangle$ and $|1\rangle$). In the first case, there is no change of the system, while in the latter there is a flip of the qubit state as required for a CNOT-gate. This gate was realized experimentally at NIST [92] using $\eta = 0.359$.

Physically the conditional dynamics of the gate relies on the different size of the ions's wavepacket in different vibrational states. The advantages of this type of gate are that only a single pulse is needed, no auxiliary levels are involved and there are no differential Stark-shifts of the levels during the gate operation.

Fast Gates

The two-ion gates discussed so far use one normal mode of the string to couple the qubits. In this case, the gate speed is limited by the mode frequency ν, or, more generally, the trap frequency ω_z, since on a faster time-scale, the vibrational sidebands cannot be resolved for selective excitation. Recently, geometric gates have been proposed which offer processing speeds beyond the vibrational frequencies [93,94,95,96]. This is possible in long ion strings, if all normal modes of vibration are excited simultaneously. By using suitably shaped laser pulses [93,95] or a fast sequence of precisely timed pulses [93,94,95,96], the time required for a two-ion phase-gate may be reduced below $2\pi/\omega_z$. The reason is that with fast pulses local oscillations of the addressed ions may be excited, in which none of the other ions undergo significant motion. This requires the coherent excitation of a large number of normal modes with a minimum of five laser pulses [96].

Quantum Computing with Trapped Ions, Table 3

Important two-qubit gates employed in ion-trap quantum computation. The table lists the matrices representing the gate operators for the joint basis states, along with the symbols used in quantum circuits. See Tab. 2 for single-ion operations

Gate	Description	Quantum Circuit	Ref.
SWAP	Ion state swapped to bus-mode $\begin{array}{c\|cccc} & g0 & g1 & e0 & e1 \\ \hline g0 & 0 & 0 & 0 & 1 \\ g1 & 0 & 1 & 0 & 0 \\ e0 & 0 & 0 & 1 & 0 \\ e1 & 1 & 0 & 0 & 0 \end{array}$ Method: blue sideband composite π-pulse	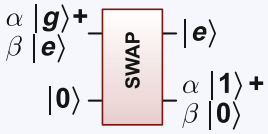	Sect. "SWAP Gate", [75]
CZ	**Controlled Z-gate:** one combination of basis states acquires minus sign (=π phase shift, hence the name *phase gate*) $\begin{array}{c\|cccc} & 0g & 0e & 1g & 1e \\ \hline 0g & 1 & 0 & 0 & 0 \\ 0e & 0 & 1 & 0 & 0 \\ 1g & 0 & 0 & 1 & 0 \\ 1e & 0 & 0 & 0 & -1 \end{array}$ Methods: • 2π-pulse on auxiliary transition • blue sideband composite 2π-pulses	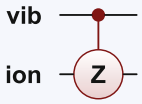	Sect. "Cirac–Zoller Gate", [22], [76]
CNOT	**Controlled NOT gate:** NOT applied to target ion if control qubit (vibration) is 1 $\begin{array}{c\|cccc} & 0g & 0e & 1g & 1e \\ \hline 0g & 1 & 0 & 0 & 0 \\ 0e & 0 & 1 & 0 & 0 \\ 1g & 0 & 0 & 0 & 1 \\ 1e & 0 & 0 & 1 & 0 \end{array}$ Methods: • CZ between $\pi/2$-pulses • wavepacket-gate • two-ion CNOT if combined with SWAP gate	 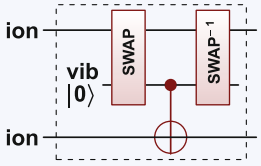	Sect. "Cirac–Zoller Gate" [22], [76], Sect. "CNOT-Wavepacket Gate", [92]

The most important fundamental logic gates realized in ion traps are summarized in Tab. 3, along with their matrix representation and the corresponding notation in quantum circuit diagrams.

Decoherence

The most important property distinguishing a qubit from a classical bit is that its two states have a well-defined relative phase, i. e., they are coherent. Uncontrollable interactions of a qubit with its environment destroy this coherence, a process known as *decoherence*. This includes the radiative decay of qubit states. Quantum memory as well as quantum gates are affected by decoherence and decoherence rates are important figures of merit for any quantum processing device. Trapped ion systems have been demonstrated to reach very low decoherence rates, making them the most successful implementation of quantum information processing to date. One way to measure decoherence is through the loss of contrast of the coherent Rabi-oscillation either on the carrier or on a sideband, as shown in Fig. 11.

Internal State Decoherence

The coherence of internal states is limited by the radiative decay of the qubit-levels. The excited states are either metastable with lifetimes on the order of seconds or hyperfine ground-states with even smaller decay rates.

Quantum Computing with Trapped Ions, Table 3
continued

Gate	Description	Quantum Circuit	Ref.
MS	**Mølmer–Sørensen gate:** multi-ion entanglement $\frac{1}{\sqrt{2}}\begin{array}{c} \begin{array}{cccc} gg & ge & eg & ee \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & ie^{-i\phi} \\ 0 & 1 & ie^{-i\phi} & 0 \\ 0 & ie^{i\phi} & 1 & 0 \\ ie^{i\phi} & 0 & 0 & 1 \end{pmatrix} \end{array} \begin{array}{c} gg \\ ge \\ eg \\ ee \end{array}$ Methods: • bichromatic Raman excitation • closed loop in motional phase-space • scalable to N ions	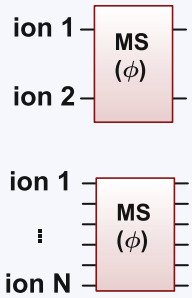	Sect. "Mølmer-Sørensen or *xy*-gate", [79], [84], [97]
$Z\phi$	**Geometric Z-phase gate:** phase shift ϕ for ions in different basis states $\begin{array}{c} \begin{array}{cccc} gg & ge & eg & ee \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & e^{i\phi} & 0 & 0 \\ 0 & 0 & e^{i\phi} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \begin{array}{c} gg \\ ge \\ eg \\ ee \end{array}$ Methods: • optical dipole force • phase shift ϕ from closed trajectory in phase space • controlled Z up to single bit rotations	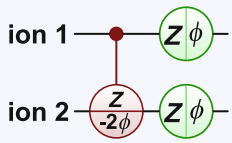	Sect. "*z*-gate", [88], [89]

The dominant contribution to the decoherence of internal states in ions used for quantum information processing are fluctuating ambient magnetic fields [98,99,100]. Through the Zeeman-shift of the levels, they result in random fluctuations of the transition frequency between the qubit states. The sensitivity to magnetic field fluctuations may be reduced by using transitions which are independent of magnetic field to first order. This strategy has been successfully applied in atomic clock systems [101]. An example are states with magnetic quantum number $m = 0$ at zero magnetic field, but similar transitions also exist at finite magnetic field. For a single magnetic field insensitive qubit, memory coherence times greater than 10 s were observed [39], five orders of magnitude larger than in experiments with magnetic field dependent states in $^9Be^+$ [102]. Another source of decoherence are the lasers driving the qubit transition. Via frequency and intensity fluctuations, they contribute to the loss of phase coherence of the qubits.

Coherence is more difficult to maintain as the length of the quantum register increases. Collective dephasing is presently the largest source of decoherence and the limiting factor for the generation of large entangled states (see Sect. "Multiparticle Entanglement"). It is caused, for example, by variations of magnetic, electric field or laser intensity across the size of a typical ion string. Ions in different locations or ions being shuttled through a large-scale trap architecture may be affected by these variations in slightly different ways.

Qubits in Decoherence-free Subspace

Frequently, the environment couples to each qubit of a quantum register in an identical way. In this case, quantum information may be protected by encoding it not in the physical basis states $|g\rangle$ and $|e\rangle$ of the qubits, but in two entangled states of two physical qubits, for example

$$|\Psi_\pm\rangle = \frac{1}{\sqrt{2}}\left(|ge\rangle \pm |eg\rangle\right) . \tag{23}$$

These logical states are invariant against common phase shifts of individual basis states. For example, if both ions

simultaneously acquire a phase shift ϕ in state $|e\rangle$, i.e, undergo the transformation $|e\rangle \rightarrow \exp(i\phi)\,|e\rangle$, $|\Psi_\pm\rangle$ is not affected, while the individual qubits would lose their phase coherence. The space spanned by $|\Psi_\pm\rangle$ is called decoherence-free subspace (DFS) [103,104,105]. It has been demonstrated [85], that a physical qubit can be reversibly encoded in a DFS, achieving coherence times two orders of magnitude longer than for the physical qubit. DFS-encoded states have been used in conjunction with qubits based on field-independent transitions to produce even longer-lived entanglement of the logical qubits [39].

Vibrational State Decoherence

The coupling required for quantum gates in a string of trapped ions is provided by their collective motion. The mode selected as a data-bus must preserve coherent superpositions of the motional ground state $|0\rangle$ and the state $|1\rangle$. Any motional decoherence of the ion chain degrades gate fidelities and must be avoided.

The ions vibrating in a string constitute an oscillating electric dipole and hence their motion couples to fluctuating electric fields in the environment. These may be generated, for example, by patch electric fields on the trap electrodes. The resulting noise in the ions' motion is a prime source of decoherence in ion trap quantum information processing.

The most obvious environmental influence on the vibration of a string is the excitation of vibrational quanta (phonons), corresponding to motional heating of the mode. This has been investigated in a number of experiments [62,65,106]. By measuring the motional excitation as a function of time starting in the ground state, heating rates between 1/ms [106] and 0.005/ms [62] were observed. Factors influencing the heating rate are the distance d between ions and electrodes (scaling approximately as d^{-4} [107]) and coating of trap electrodes with atoms during loading of the trap, resulting in patch-potentials. Therefore, careful shielding of the electrodes from atomic beam exposure is important. Another way to reduce heating rates is cooling the electrodes [107,108].

For a multi-ion string, the rate of motional heating strongly depends on which normal mode is used. The COM-mode has heating rates comparable to that of a single ion, as it can be excited by a homogeneous fluctuating field coupling to each individual ion in an equivalent way. The stretch-mode, on the other hand, can only be excited by a field gradient and is therefore much less sensitive to fluctuating fields emanating from the trap electrodes. Therefore, for high fidelity gate operations, the stretch mode is preferable.

While motional heating sets an upper limit to the coherence time of bus qubits, for reliable quantum operations also dephasing of motional superpositions must be taken into account. This has been investigated at NIST [109], in Innsbruck [110] and in Oxford [111]. Dephasing times found are on the order of 100 ms. Dephasing is faster for larger differences in the quantum numbers of the states involved, making it advantageous to use the lowest vibrational states $|0\rangle$ and $|1\rangle$.

Even though only one normal mode is used for quantum information transfer, heating of the remaining $N-1$ axial vibration modes in an N-ion string (spectator modes) is detrimental, since it affects the Rabi-frequency of the bus-mode. This may be alleviated by cooling all modes of the string, ideally to the quantum mechanical ground state.

With typical gate durations on the order of 1 µs, decoherence times of 1 s would permit 10^6 quantum operations to be performed. Another approach is to consider the probability of error during a gate operation [112]. It should be smaller than 10^{-4} to be able to successfully apply quantum error correction. This corresponds to a fidelity of $F > 0.9999$, which present quantum gates do not yet achieve.

Quantum Algorithms

The two-ion gates described in the previous section are universal, i. e., any logic circuit may be realized by combining them with single qubit operations [113,114]. In this section, examples of algorithms implemented in ion traps are presented. The most important quantum algorithms have been demonstrated in ion-trap systems in their simplest form.

Deutsch–Jozsa Algorithm

The Deutsch–Jozsa algorithm [115] is the simplest example of quantum parallelism. The task it performs is to decide if a function f is constant, i. e., has either always the result $f(x) = 0$ or always $f(x) = 1$, independent of the input x, or if it is balanced, i. e. its output depends on the input ($f(x) = x$ or $f(x) = \mathrm{NOT}x$). With classical means, two function evaluations are necessary, whereas the quantum algorithm performs the task with a single application of the function to a superposition state.

The algorithm has been demonstrated with a single ionic qubit, using its vibrational excitation as an auxiliary qubit [75] (see Fig. 23). The required pulses were applied either resonantly or on the blue sideband, in the latter case using composite pulses in order to avoid excitation of more than one phonon (cf. Sect. "Cirac–Zoller Gate").

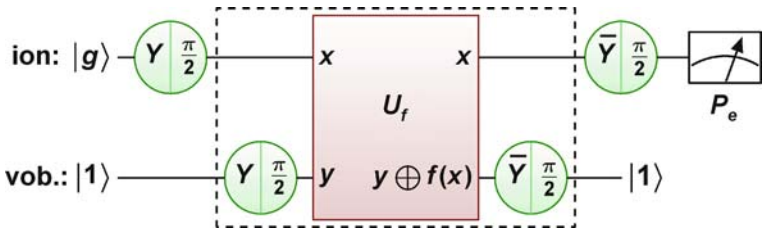

Quantum Computing with Trapped Ions, Figure 23
Quantum circuit implementing the Deutsch–Jozsa algorithm with a trapped ion. The upper line represents the ionic qubit $|x\rangle$, the lower line an auxiliary vibrational qubit. In the central box, one of four possible functions $f(x)$ is implemented by acting on y with the identity, a NOT gate and/or a CNOT-gate controlled by x. A final measurement of the ion in state $|e\rangle$ indicates a balanced function after only a single function evaluation [75]

Quantum Teleportation

A more complex task and one of fundamental importance for quantum computation is quantum teleportation, the transfer of an unknown quantum state from one qubit to another in a distant location [116]. Even though no finite number of measurements is sufficient to obtain a full specification of a quantum state, the transfer is possible with the help of non-local correlations provided by an entangled pair of qubits. While probabilistic teleportation has been demonstrated with photonic qubits, two ion-trap experiments have implemented a fully deterministic quantum teleportation protocol. At the University of Innsbruck, $^{40}Ca^{+}$-ions were used [99], while at NIST the experiment was performed with $^{9}Be^{+}$ [98].

The teleportation procedure requires three ionic qubits. The essential steps are summarized in Fig. 24. Initially, ions B and C are prepared in a maximally entangled state using the gates described in Sect. "Cirac–Zoller Gate" (Innsbruck) and Sect. "Geometric Gates" (NIST). Subsequently, the state to be transferred is prepared in ion A by a single-qubit rotation. The central step of the teleportation protocol is a so-called Bell-state measurement of the states of ions A and B. This is achieved by using another gate to entangle qubits A and B and, after subjecting them to a $\pi/2$-pulse, measuring the state of both qubits. Taken separately, none of the four possible measurement outcomes (gg, ge, eg, or ee), nor the remaining ion C carry any information about the original state. Only when one of four unitary state rotations is applied to ion C, chosen depending on the outcome of the measurement, the state of ion C becomes identical to the original state A.

In their implementation of this protocol, the two groups have employed quite different technologies. The qubits are stored differently and different realizations of the phase gate are employed. The Innsbruck group addresses individual ions with tightly focused lasers, protecting the remaining ions from the target ion's fluorescence by hiding them in another internal state. At NIST, ions are selectively moved to separate zones in a segmented trap, where they can be manipulated individually while maintaining entanglement. In spite of these differences, in both experiments the quantum bit was transferred with a fidelity of around 75%. In a quantum computer, teleportation may be applied to transfer quantum information without physically moving qubits.

Quantum Fourier Transform

One of the most celebrated quantum algorithms was developed by P. Shor for finding the prime factors of a large number [117]. A central element of this algorithm is the *quantum Fourier transform* of a set of qubits. Here, the amplitudes x_k of a superposition of basis states are Fourier-transformed, resulting in a new state with amplitudes y_j:

$$\sum_{k=0}^{N-1} x_k\,|k\rangle \longrightarrow \sum_{j=0}^{N-1} y_j\,|j\rangle \quad \text{with} \quad y_j = \sum_{k=0}^{N-1} x_k e^{i2\pi jk/N}. \tag{24}$$

Here, $|k\rangle$ is related to the corresponding ion-state through its binary representation with $0 \equiv g$ and $1 \equiv e$.

The quantum Fourier transform was experimentally demonstrated in a system of three beryllium ions for states with periods 1,2,4,8 and approximately 3 [118]. The sequence of gates that was applied is shown in Fig. 25a. It comprises Hadamard transforms and controlled Z-gates with different phase factors. In the experiment, these are replaced by equivalent sequences of X- and Y-pulses. A simplification of the scheme is possible by performing the read-out of each qubit immediately after the last Hadamard-gate is applied. Measurement of the individual ions is achieved by separating them in a segmented linear trap. The subsequent quantum gates in Fig. 25a must

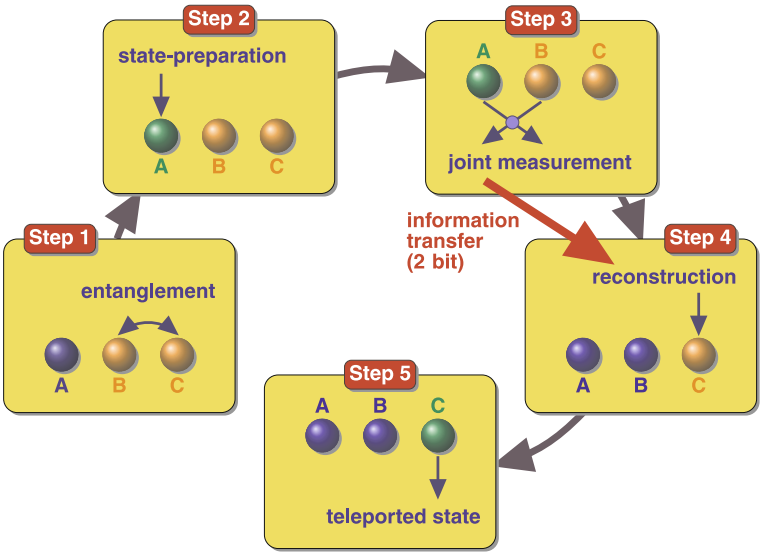

Quantum Computing with Trapped Ions, Figure 24
Schematic representation of quantum teleportation in a chain of three ions: the state of ion A (green) is transferred to ion C with the help of an entangled pair of ions B and C. Blue arrows indicate the manipulation of ions with laser pulses, while the red arrow corresponds to the transmission of two bit of classical information. This protocol is used in two independent experiments [98,99], which differ only in technical details

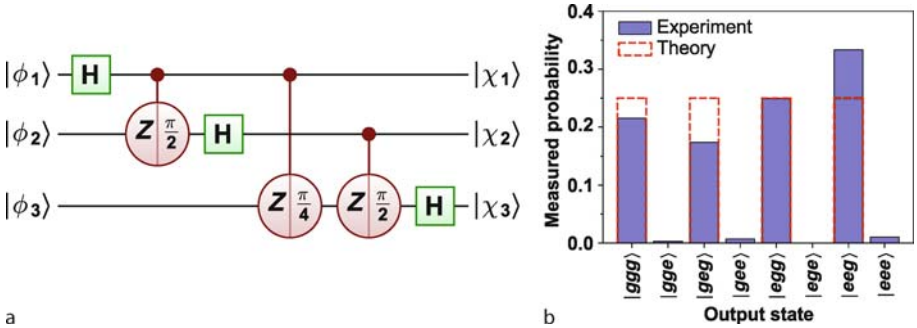

Quantum Computing with Trapped Ions, Figure 25
a Circuit diagram for quantum Fourier transform of three qubits, composed of Hadamard transforms and controlled Z-gates (with rotation angles $\theta = \pi/2$ and $\pi/4$). In the actual experiment, these gates were replaced by equivalent rotation sequences around the x- and y-axis; b experimental output for the quantum Fourier transform of the input state $|gee\rangle + |eee\rangle$ with period 4 [118]

then be replaced by classically controlled phase shifts. Figure 25b shows an example of the outcome of such a measurement. A state with a period 4 in an 8 element space has non-zero Fourier components with a periodicity of 2 (= 8/4). Therefore, only every other state is observed in the output. Depending on the input state, an accuracy between 87% and 99% was reached.

Grover's Quantum Search Algorithm

Grover's search algorithm is another example of a quantum computer outperforming its classical counterpart. It succeeds in finding a marked entry in a database of length n after $\pi\sqrt{n}/4$ queries, rather than n [119]. It requires an operator V (oracle) marking the target state $|a\rangle$ of the search by flipping its sign and an operator W to amplify the contribution of the target state, until, after a few iterations, only the target state has a sizable probability. The position of the target state in the database is then revealed after a measurement. The corresponding quantum circuit is shown in Fig. 26.

In the experiment [120], a four element database was set up with two ionic qubits. Using $\pi/2$-pulses, an equally weighted superposition of all basis states is prepared as

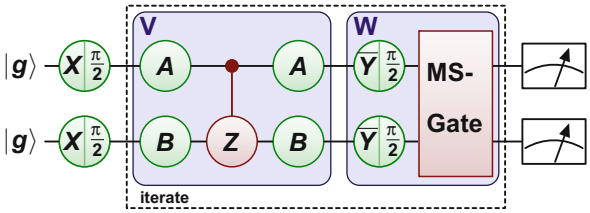

Quantum Computing with Trapped Ions, Figure 26
Scheme of Grover's search algorithm for N qubits encoding $n = 2^N$ entries, illustrated for $N = 2$. Initially, all qubits are prepared in the ground state. $\pi/2$-pulses prepare an equal superposition of all basis states. The oracle V flips the sign of one marked element $|a\rangle$. Which state is marked is determined by the operators A and B, which can be either NOT (X-gate, see Tab. 2) or the identity. Two more $\pi/2$-pulses followed by a Mølmer–Sørensen gate amplify the weight of the marked state. After a number of iterations, the qubits are measured

the input: $|\phi\rangle = \big(|gg\rangle + |ge\rangle + |eg\rangle + |ee\rangle\big)/2$. The oracle V as well as the amplification W were implemented using the Mølmer–Sørensen gate (Sect. "Mølmer-Sørensen or xy-gate") together with single-qubit rotations. The iteration step in Fig. 26 takes the state $|\phi\rangle$ through the following transformation:

$$|\phi\rangle \xrightarrow{V} (|\phi\rangle - |a\rangle) \xrightarrow{W} |a\rangle \ .$$

Therefore, the marked state $|a\rangle$ should be retrieved in only a single query. The measured probability was 60%, surpassing the classical limit of 50%. For a register size larger than $N = 2$, the operators V and W have to be applied iteratively.

Quantum Error Correction

Any quantum computer is subject to noise, resulting in uncontrolled, irreversible changes of the qubits involved. Even using decoherence-free subspaces (see Sect. "Qubits in Decoherence-Free Subspace"), residual noise still reduces the fidelity of quantum memory and quantum gates. The ability to correct these errors is therefore of great importance for the reliable operation of a quantum computer. The challenge is that error correction must be accomplished without gaining any knowledge about the stored information, since this would destroy any quantum superpositions.

An example for an error which might occur during the quantum computation is the flip of a single qubit. The correction of an error of this type has been demonstrated at NIST [100]. To this end, a redundant encoding of a single logical qubit in three physical qubits was applied. A qubit state $\alpha\,|g\rangle + \beta\,|e\rangle$ was encoded as an entangled state $\alpha\,|ggg\rangle + \beta\,|eee\rangle$. At NIST, the two-ion phase gate described in Sect. "Geometric Gates", extended to three ions was used. The logical qubit was then subjected to an artificial bit-flip error of variable rate. After decoding the logical qubit by reversing the three-ion entangling gate, information on which error has occurred was obtained by measuring two of the physical qubits. The outcome determined which correction operation was applied to the remaining qubit to restore the original state. Finally, the qubit was analyzed to determine the efficiency of the method. The protocol is illustrated in Fig. 27.

The data showed that quantum error correction improved the fidelity if bit-flips occurred with a probability larger than 25%. For smaller errors, the imperfections of the gates applied during encoding and decoding of the logical qubit outweighed the benefit of correction and lead to a reduced fidelity. Improved gates are necessary before a fault tolerant operation can be achieved. Since quantum error correction schemes require a large number of physical quantum bits, passive protection against errors, for example through decoherence-free subspaces, is important.

Multiparticle Entanglement

As shown in the previous section, entanglement is an essential resource in many quantum algorithms. In the course of a quantum computation, a large number of qubits must be entangled. It is therefore of great relevance to investigate highly entangled quantum states of many qubits. This is not only an important demonstration of the amount of control over a quantum system, but can also serve as a resource for quantum information processing [121].

Entangled states may be generated with the help of two- or multi-bit quantum gates (Sect. "Two-Qubit Interaction and Quantum Gates"). An early achievement was the entanglement of four ions [84] using the Mølmer–Sørensen gate (Sect. "Mølmer-Sørensen or xy-gate"). Since then, the development of new gates and better control over experimental parameters have led to the creation and verification of entangled states of up to eight ions. Each additional particle greatly increases the difficulty of the experiment.

Two different classes of entangled states have been investigated. The first type is called Greenberger–Horne-Zeilinger (GHZ) or *Schrödinger cat* states and consists of equal superpositions of maximally different quantum

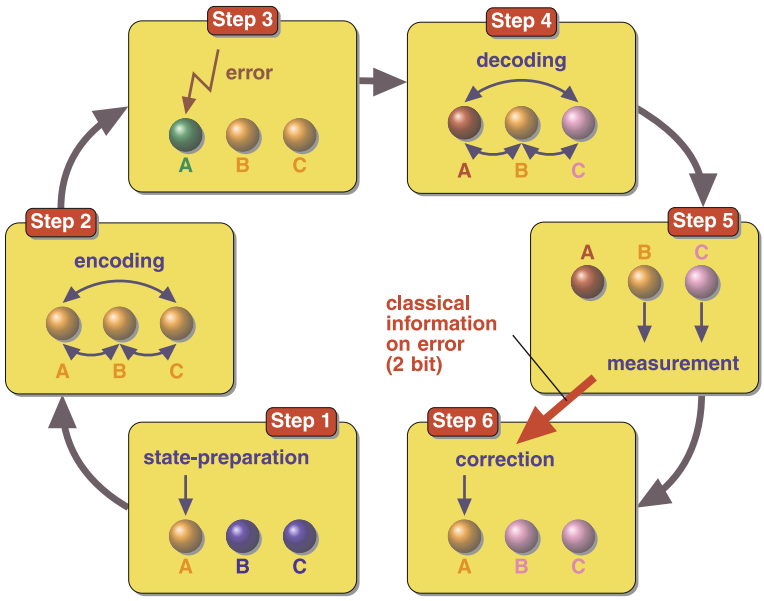

Quantum Computing with Trapped Ions, Figure 27
Schematic representation of quantum error correction of flips of a Single logical quantum bit logical quantum bit, encoded in a chain of three ions. This protocol was demonstrated experimentally [100]

states of N ions.

$$|\mathrm{GHZ}_N\rangle = \frac{1}{\sqrt{2}}\left(\,|\,\underbrace{ggg\ldots g}_{N}\,\rangle + \mathrm{e}^{i\varphi}\,|\,\underbrace{eee\ldots e}_{N}\,\rangle\,\right). \quad (25)$$

Three-ion GHZ-states were generated and controlled in 2004 at NIST ($F = 89\%$) [122] and the University of Innsbruck ($F = 72\%$) [123]. In both experiments, the potential of entanglement for practical applications was explored.

The largest GHZ-state so far was generated at NIST, using a generalization of the phase gate described in Sect. "z-gate" for up to $N = 6$ ions [97]. Together with a common single-bit rotation of all the ions before and after the phase gate, the algorithm involved three steps, irrespective of the number of ions to be entangled, making this a very efficient method for multiparticle entanglement generation. The fidelity obtained for the six-ion entangled state was estimated to be better than 51%.

In order to verify the coherence of the entangled states, experimenters have subjected them to another phase gate, identical to the entangling operation, but in a reference system rotated around the z-axis of the Bloch-sphere by an angle ϕ. In this way, the coherence of the state $|\mathrm{GHZ}_N\rangle$ is converted to a population difference, which can be directly measured. For a perfect N-ion GHZ-state, the population difference and hence the fluorescence signal varies as

$$P_{gg\ldots g} = \tfrac{1}{2}\left[1 - \cos(N\phi)\right]. \quad (26)$$

Results of this method for GHZ-states with four, five, and six ions are shown in Fig. 28. The observed contrast provides a lower bound for the coherence and hence the fidelity of the state.

This method of probing coherences is closely related to Ramsey spectroscopy and, indeed, GHZ-states of multiple ions have been applied in spectroscopy. The N-fold increased oscillation frequency observed in Fig. 28 can be used to improve the phase sensitivity of a Ramsey spectrometer by a factor of $\sqrt{N}$ compared to unentangled atoms [122]. Other applications of GHZ-states include quantum error correction (Sect. "Quantum Error Correction") and the deterministic preparation of Bell states.

A different class of entangled states are the N-particle W-states, consisting of a superposition of N states with exactly one particle in state $|g\rangle$ and all others in $|e\rangle$:

$$|\mathrm{W}_N\rangle = \frac{1}{\sqrt{N}}\left(\mathrm{e}^{i\varphi_1}\,|e\ldots eeeg\rangle + \mathrm{e}^{i\varphi_2}\,|e\ldots eege\rangle + \mathrm{e}^{i\varphi_3}\,|e\ldots egee\rangle + \cdots + \mathrm{e}^{i\varphi_N}\,|gee\ldots ee\rangle\right). \quad (27)$$

These states were generated in Innsbruck with $N = 3$ [123] and later up to $N = 8$ [124]. First, all ions are

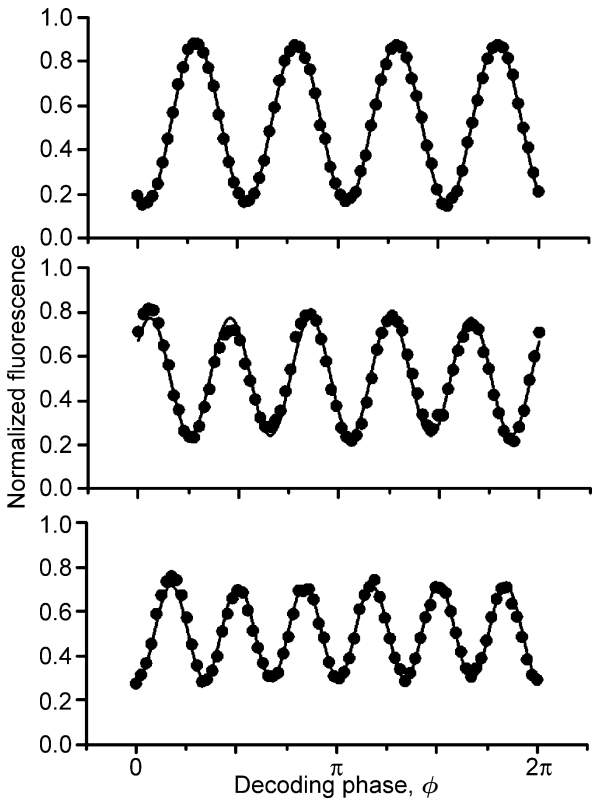

Quantum Computing with Trapped Ions, Figure 28
Fluorescence measurement of decoded GHZ-states as a function of the decoding phase ϕ for $N = 4, 5$, and 6 ions. The sinusoidal dependence on $N\phi$ is a signature of an N-ion GHZ state [97]. The observed contrast is related to the coherence of the state

initialized in state $|e\rangle$ and the vibrational COM-mode in state $|1\rangle$. N pulses on the blue sideband of the qubit transition, are consecutively applied to each ion in the string, exciting the level $|g\rangle$ of that ion with a probability of $1/N$. The entanglement is verified by measuring all N^2 elements of the density matrix, completely characterizing the quantum state. For the state $|W_8\rangle$, a fidelity of $F = 72\%$ was determined.

The entanglement properties of W-states are very different from those of GHZ-states. If the state of one particle is measured to be $|e\rangle$, the remaining particles are still entangled. In addition, W-states are robust against global dephasing. They might serve as a resource for quantum information processing as well as quantum communication.

The six-ion GHZ- and eight-ion W-states define the state-of-the-art in the generation of multiparticle entanglement. The generation of even larger entangled states is a major challenge, due to the increased sensitivity to decoherence and inhomogeneous effects.

Distributed Quantum Information with Trapped Ions

The quantum systems discussed so far were confined to single ion traps. However, many important applications require the distribution of quantum information among different locations. A prime example is quantum communication, which at present is based on photons as carriers of quantum information. Substantial benefits are expected from linking quantum communication with quantum computing. For example, small systems of trapped ions could act as routers or repeaters of quantum information. Teleportation is another application which is relevant mainly over long distances, i. e., between remote ion traps. Even quantum computation itself could gain from being spread among a number of local processing sites (distributed quantum computation), linked by quantum communication channels [125].

It is therefore an important goal to establish reliable quantum networks involving spatially separated traps. A first step is to create entanglement between ions stored in remote locations. There are three methods to achieve this:

(1) An entangled pair of ions is prepared in one location, and one of the particles is subsequently transported to a distant site.
(2) Entanglement is created between an ionic and a photonic qubit, the latter transferring entanglement deterministically to an ion in a distant trap.
(3) Entanglement is created locally in two distant traps between an ion and a photon on each side. Subjecting the emitted photons to a joint measurement projects the ions left behind to an entangled state. This scheme works probabilistically, as it is dependent on a particular measurement outcome.

The first method is suited for entanglement distribution within an extended trap architecture (Sect. "Large-Scale Ion Traps"), while the latter two methods allow long distance entanglement.

Ion–Photon and Remote Ion–Ion Entanglement

A crucial step towards mapping quantum information from an atomic ion to a photon is to generate entanglement between trapped ions and photons, establishing a long-distance communication channel. This was demonstrated at the University of Michigan using a single $^{111}Cd^+$-ion [126]. In the experiment sketched in Fig. 29a, the ion was excited to a state with two channels of spontaneous decay, leading to different hyperfine states of the

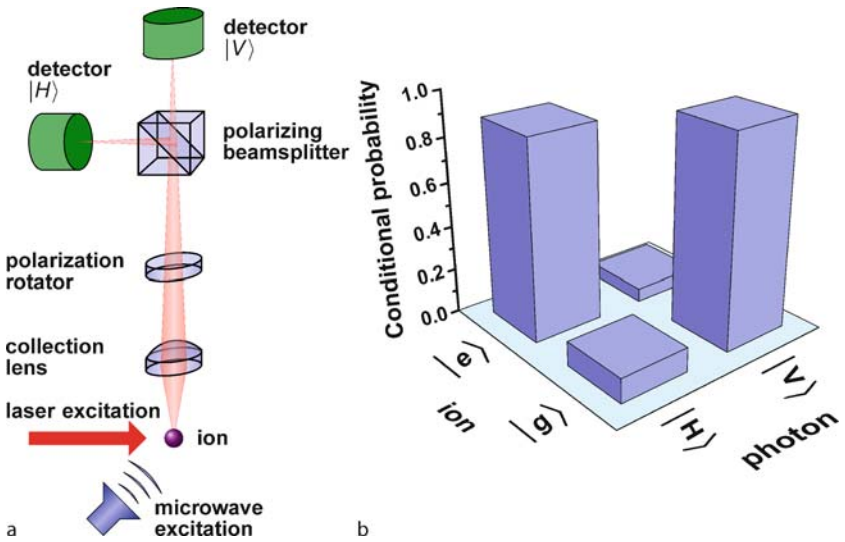

Quantum Computing with Trapped Ions, Figure 29

a Experimental setup for generation of ion-photon entanglement. Photons scattered from the excitation beam are analyzed by a beamsplitter distinguishing V- and H-polarization. The polarization rotator is used to change the measurement basis. The qubit-transition is driven by a microwave field. b Measured conditional probabilities after both atomic and photonic qubits are rotated by a polar angle $\theta = \pi/2$ on the Bloch sphere with equal phase Φ. If the atomic and photonic qubits were not entangled but in a statistically mixed state, all conditional probabilities in the figure would have been 0.5 [126]

ion ($|g\rangle$ or $|e\rangle$). Information on which transition has occurred is contained in the polarization of the emitted photon (horizontal=$|\mathrm{H}\rangle$ or vertical=$|\mathrm{V}\rangle$). Taking into account the relative strength of the transitions, the resulting state is $\sqrt{1/3}\,|\mathrm{H}\rangle\,|e\rangle + \sqrt{2/3}\,|\mathrm{V}\rangle\,|g\rangle$, which is entangled. In the experiment, it was obtained with a fidelity of $F = 97\%$ (see Fig. 29b). Note that this method of generating entanglement is probabilistic, since it relies on the spontaneous emission of a photon.

The generation of entanglement between ion and photon is only an intermediate step towards the goal of entangling the quantum states of ions in different traps. This has been achieved by the same group. Initially, single $^{171}\mathrm{Yb}^{+}$-ions are stored in two identical traps 1 m apart. Using the same principle as described above, an entangled ion-photon state $\left(|e\rangle_i\,|\nu_e\rangle_i - |g\rangle_i\,|\nu_g\rangle_i\right)/\sqrt{2}$ is created in each trap, where the index i identifies the trap. The photon states $|\nu_g\rangle$ and $|\nu_e\rangle$ are distinguished by their frequency. The two photon-states are then brought to interference on a beamsplitter and measured at its two output ports. In case of a coincident detection, the state of the atoms is projected to the entangled state $\left(|e\rangle_1\,|g\rangle_2 - |g\rangle_1\,|e\rangle_2\right)/\sqrt{2}$. An experimental verification yields a fidelity of 63% [127].

Ion-Trap Cavity-QED

One of the drawbacks of entanglement using spontaneously emitted photons is the low success probability, due to the emission of photons into the full solid angle. As a result, only one ion–ion pair was entangled every 8.5 min in the experiment reported in Ref. [127]. A solution to this problem is to place the ion in a high-finesse Fabry–Perot optical cavity resonant with the emitted photons. The cavity enhances the coherent interaction between ion and photon and provides a well-defined mode for the photon, resulting in a predetermined direction of propagation when it eventually escapes through a semi-transparent mirror. Individual $^{40}\mathrm{Ca}^{+}$-ions have already been stored in and coupled to an optical cavity, localized well within the Lamb–Dicke regime [128,129].

Optical cavities have applications beyond enhancing the success rate of probabilistic entanglement schemes. If the cavity mode volume is small enough, the coherent interaction between ions and photons becomes stronger than all spontaneous decay processes and the system dynamics becomes deterministic. This is known as the strong coupling limit of cavity-QED. It has been proposed as a technique to transfer quantum states or distribute entanglement deterministically in a quantum network [130]. According to this scheme, the qubit state of an ion is mapped to a cavity photon with the help of a laser pulse. This photon leaks out of the first cavity and enters a second cavity, in which it is absorbed by another ion. A requirement for this transfer is that the coupling between the second ion and the second cavity is provided by a laser pulse which is time-reversed with respect to the first pulse and

that the photon-wavepacket carrying the quantum bit is time-symmetric.

While the strong coupling limit hasn't been reached with single trapped ions yet, the system has already been used as a source of single photons on demand, emitting photons with a predetermined shape [131]. Cavity-QED techniques have many potential applications in quantum information processing with trapped ions, in particular for the reversible mapping of qubits between photons and ions [125].

Future Directions

As shown in this review, all essential elements of quantum information processing have been demonstrated with trapped ions, making them the most successful technology for quantum computation to date. The biggest challenge now is to scale up the present systems to a size where algorithms of practical interest could be run. At the same time, the fidelity of gate operations must be improved, in order to reach the point where quantum error correction could be successfully applied. The implementation of quantum error correction will further increase the demand for larger quantum registers, due to the memory overhead of logical qubits.

The most promising route to large-scale quantum computation is a segmented trap architecture with a large number of zones, each storing only a few ions [25]. Present activities in laboratories around the world are directed towards building suitable traps using nanofabrication methods. It is still a major challenge to combine all the required techniques in one scalable system. The implementation of a quantum computer with 300 qubits is discussed by A.M. Steane [132].

Before a universal quantum computer will be available, trapped ion systems are expected to find applications in the analog simulation of other quantum systems, as was originally proposed by Feynman [2]. The idea is to use the evolution of a system of trapped ions to mimic the behavior of other, less accessible systems with equivalent Hamiltonians. An example are quantum phase transitions in spin systems, which could be investigated with trapped ions [133]. A recent experiment has simulated the transition from paramagnetic to ferromagnetic order in a quantum system of two ions [134]. A quantum simulation that was successfully performed at NIST demonstrated the increased sensitivity of 2nd- and 3rd-order nonlinear beam splitters [135]. Trapped ions are even predicted to simulate cosmological particle creation in the early universe [136].

Beyond quantum computation or simulation, the techniques developed for ion trap quantum computers have already started to find applications in other fields. One example is frequency standards. Here, the use of N entangled ions has been demonstrated to increase the sensitivity by a factor of $\sqrt{N}$ [122]. Methods from ion-trap quantum information processing have also been applied to spectroscopic investigation of $^{27}Al^+$ by using $^{9}Be^+$ to cool, initialize and detect the Al-ion, which lacks a suitable level structure [137]. The rapid progress of ion-trap technology not only makes harnessing the power of quantum computation a realistic possibility in the near future, but already provides new tools for fundamental and applied research.

Bibliography

Primary Literature

1. Feynman RP (1960) There's plenty of room at the bottom. Eng Sci 23:22–36
2. Feynman RP (1982) Simulating physics with computers. Int J Theor Phys 21:467–488
3. Benioff P (1982) Quantum-mechanical models of turing-machines that dissipate no energy. Phys Rev Lett 48:1581–1585
4. Cirac JI, Zoller P (1995) Quantum computations with cold trapped ions. Phys Rev Lett 74:4091–4094
5. DiVincenzo DP (2000) The physical implementation of quantum computation. Fortschritte Phys-Prog Phys 48:771–783
6. Ghosh PK (1995) Ion Traps. Clarendon, Oxford
7. Paul W (1990) Electromagnetic traps for charged and neutral particles. Rev Mod Phys 62:531–540
8. Paul W, Steinwedel H (1953) Ein neues Massenspektrometer ohne Magnetfeld. Z Naturforsch 8:448–450
9. Waki I, Kassner S, Birkl G et al (1992) Observation of ordered structures of laser-cooled ions in a quadrupole storage ring. Phys Rev Lett 68:2007–2010
10. Nagerl HC, Bechter W, Eschner J et al (1998) Ion strings for quantum gates. Appl Phys B-Lasers Opt 66:603–608
11. Raizen MG, Gilligan JM, Bergquist JC et al (1992) Ionic-crystals in a linear paul trap. Phys Rev A 45:6493–6501
12. Drewsen M, Brodersen C, Hornekaer L et al (1998) Large ion crystals in a linear Paul trap. Phys Rev Lett 81:2827–2830
13. James DFV (1998) Quantum dynamics of cold trapped ions with application to quantum computation. Appl Phys B-Lasers Opt 66:181–190
14. Hughes RJ, James DFV, Gomez JJ et al (1998) The Los Alamos trapped ion quantum computer experiment. Fortschritte Phys-Prog Phys 46:329–361
15. Berkeland DJ, Miller JD, Bergquist JC et al (1998) Minimization of ion micromotion in a Paul trap. J Appl Phys 83:5025–5033
16. Hoffges JT, Baldauf HW, Eichler T et al (1997) Heterodyne measurement of the fluorescent radiation of a single trapped ion. Opt Commun 133:170–174
17. Dehmelt H (1967) Radiofrequency Spectroscopy of Stored Ions. Adv At Mol Phys 3:53
18. Straubel H (1955) Zum Öltröpfchenversuch von Millikan. Naturwissenschaften 42:506–507
19. Brewer RG, Devoe RG, Kallenbach R (1992) Planar ion microtraps. Phys Rev A 46:R6781–R6784
20. Schrama CA, Peik E, Smith WW et al (1993) Novel miniature ion traps. Opt Commun 101:32–36

21. Jefferts SR, Monroe C, Bell EW et al (1995) Coaxial-resonator-driven rf (paul) trap for strong confinement. Phys Rev A 51:3112–3116
22. Monroe C, Meekhof DM, King BE et al (1995) Demonstration of a fundamental quantum logic gate. Phys Rev Lett 75:4714–4717
23. Turchette QA, Wood CS, King BE et al (1998) Deterministic entanglement of two trapped ions. Phys Rev Lett 81:3631–3634
24. Cirac JL, Zoller P (2000) A scalable quantum computer with ions in an array of microtraps. Nature 404:579–581
25. Kielpinski D, Monroe C, Wineland DJ (2002) Architecture for a large-scale ion-trap quantum computer. Nature 417:709–711
26. Rowe MA, Ben-Kish A, Demarco B et al (2002) Transport of quantum states and separation of ions in a dual RF ion trap. Quantum Inform Comput 2:257–271
27. Hensinger WK, Olmschenk S, Stick D et al (2006) T-junction ion trap array for two-dimensional ion shuttling, storage, and manipulation. Appl Phys Lett 88:034101
28. Madsen MJ, Hensinger WK, Stick D et al (2004) Planar ion trap geometry for microfabrication. Appl Phys B-Lasers Opt 78:639–651
29. Brownnutt M, Wilpers G, Gill P et al (2006) Monolithic microfabricated ion trap chip design for scaleable quantum processors. New J Phys 8:232
30. Stick D, Hensinger WK, Olmschenk S et al (2006) Ion trap in a semiconductor chip. Nat Phys 2:36–39
31. Seidelin S, Chiaverini J, Reichle R et al (2006) Microfabricated surface-electrode ion trap for scalable quantum information processing. Phys Rev Lett 96:253003
32. Chiaverini J, Blakestad RB, Britton J et al (2005) Surface-electrode architecture for ion-trap quantum information processing. Quantum Inform Comput 5:419–439
33. Kim J, Pau S, Ma Z et al (2005) System design for large-scale ion trap quantum information processor. Quantum Inform Comput 5:515–537
34. Mitchell TB, Bollinger JJ, Dubin DHE et al (1998) Direct observations of structural phase transitions in planar crystallized ion plasmas. Science 282:1290–1293
35. Porras D, Cirac JI (2006) Phonon superfluids in sets of trapped ions. Found Phys 36:465–476
36. Castrejon-Pita JR, Ohadi H, Crick DR et al (2007) Novel designs for Penning ion traps. J Mod Opt 54:1581–1594
37. Crick DR, Ohadi H, Bhatti I et al (2008) Two-ion Coulomb crystals of Ca+ in a Penning trap. Opt Express 16:2351–2362
38. Ciaramicoli G, Marzoli I, Tombesi P (2003) Scalable quantum processor with trapped electrons. Phys Rev Lett 91:017901
39. Langer C, Ozeri R, Jost JD et al (2005) Long-lived qubit memory using atomic ions. Phys Rev Lett 95:060502
40. Nagerl HC, Roos C, Rohde H et al (2000) Addressing and cooling of single ions in Paul traps. Fortschritte Phys-Prog Phys 48:623–636
41. Lucas DM, Donald CJS, Home JP et al (2003) Oxford ion-trap quantum computing project. Philos Trans R Soc Lond Ser A-Math Phys Eng Sci 361:1401–1408
42. Benhelm J, Kirchmair G, Rapol U et al (2007) Measurement of the hyperfine structure of the S-1/2-D-5/2 transition in Ca-43$^+$. Phys Rev A 75:032506
43. Deslauriers L, Haijan PC, Lee PJ et al (2004) Zero-point cooling and low heating of trapped Cd-111$^+$ ions. Phys Rev A 70:043408
44. Balzer C, Braun A, Hannemann T et al (2006) Electrodynamically trapped Yb+ ions for quantum information processing. Phys Rev A 73:041407
45. Olmschenk S, Younge KC, Moehring DL et al (2007) Manipulation and detection of a trapped Yb+ hyperfine qubit. Phys Rev A 76:052314
46. Schaetz T, Friedenauer A, Schmitz H et al (2007) Towards (scalable) quantum simulations in ion traps. J Mod Opt 54:2317–2325
47. Berkeland DJ (2002) Linear Paul trap for strontium ions. Rev Sci Instrum 73:2856–2860
48. Brown KR, Clark RJ, Labaziewicz J et al (2007) Loading and characterization of a printed-circuit-board atomic ion trap. Phys Rev A 75:015401
49. Brownnutt M, Letchumanan V, Wilpers G et al (2007) Controlled photoionization loading of Sr-88$^+$ for precision ion-trap experiments. Appl Phys B-Lasers Opt 87:411–415
50. Letchumanan V, Wilpers G, Brownnutt M et al (2007) Zero-point cooling and heating-rate measurements of a single Sr-88$^+$ ion. Phys Rev A 75:063425
51. Kjaergaard N, Hornekaer L, Thommesen AM et al (2000) Isotope selective loading of an ion trap using resonance-enhanced two-photon ionization. Appl Phys B-Lasers Opt 71:207–210
52. Gulde S, Rotter D, Barton P et al (2001) Simple and efficient photo-ionization loading of ions for precision ion-trapping experiments. Appl Phys B-Lasers Opt 73:861–863
53. Deslauriers L, Acton M, Blinov BB et al (2006) Efficient photoionization loading of trapped ions with ultrafast pulses. Phys Rev A 74:063421
54. Lucas DM, Ramos A, Home JP et al (2004) Isotope-selective photoionization for calcium ion trapping. Phys Rev A 69:012711
55. Happer W (1972) Optical-pumping. Rev Mod Phys 44:169
56. Wineland DJ, Drullinger RE, Walls FL (1978) Radiation-pressure cooling of bound resonant absorbers. Phys Rev Lett 40:1639–1642
57. Neuhauser W, Hohenstatt M, Toschek P et al (1978) Optical-sideband cooling of visible atom cloud confined in parabolic well. Phys Rev Lett 41:233–236
58. Wineland DJ, Itano WM (1979) Laser cooling of atoms. Phys Rev A 20:1521–1540
59. Stenholm S (1986) The semiclassical theory of laser cooling. Rev Mod Phys 58:699–739
60. Diedrich F, Bergquist JC, Itano WM et al (1989) Laser cooling to the zero-point energy of motion. Phys Rev Lett 62:403–406
61. Monroe C, Meekhof DM, King BE et al (1995) Resolved-sideband raman cooling of a bound atom to the 3d zero-point energy. Phys Rev Lett 75:4011–4014
62. Roos C, Zeiger T, Rohde H et al (1999) Quantum state engineering on an optical transition and decoherence in a Paul trap. Phys Rev Lett 83:4713–4716
63. Kielpinski D, King BE, Myatt CJ et al (2000) Sympathetic cooling of trapped ions for quantum logic. Phys Rev A 61:032310
64. Schmidt-Kaler F, Roos C, Nagerl HC et al (2000) Ground state cooling, quantum state engineering and study of decoherence of ions in Paul traps. J Mod Opt 47:2573–2582
65. Rohde H, Gulde ST, Roos CF et al (2001) Sympathetic ground-state cooling and coherent manipulation with two-ion crystals. J Opt B-Quantum Semicl Opt 3:S34–S41
66. Barrett MD, DeMarco B, Schaetz T et al (2003) Sympathetic

cooling of Be-9$^+$ and Mg-24$^+$ for quantum logic. Phys Rev A 68:042302
67. King BE, Wood CS, Myatt CJ et al (1998) Cooling the collective motion of trapped ions to initialize a quantum register. Phys Rev Lett 81:1525–1528
68. Mintert F, Wunderlich C (2001) Ion-trap quantum logic using long-wavelength radiation. Phys Rev Lett 87:257904
69. Nagourney W, Sandberg J, Dehmelt H (1986) Shelved optical electron amplifier – observation of quantum jumps. Phys Rev Lett 56:2797–2799
70. Wineland DJ, Bergquist JC, Itano WM et al (1980) Double-resonance and optical-pumping experiments on electromagnetically confined, laser-cooled ions. Opt Lett 5:245–247
71. Blatt R, Zoller P (1988) Quantum jumps in atomic systems. Eur J Phys 9:250–256
72. Bergquist JC, Hulet RG, Itano WM et al (1986) Observation of quantum jumps in a single atom. Phys Rev Lett 57:1699–1702
73. Vogel K, Risken H (1989) Determination of quasiprobability distributions in terms of probability distributions for the rotated quadrature phase. Phys Rev A 40:2847–2849
74. Roos CF, Lancaster GPT, Riebe M et al (2004) Bell states of atoms with ultralong lifetimes and their tomographic state analysis. Phys Rev Lett 92:220402
75. Gulde S, Riebe M, Lancaster GPT et al (2003) Implementation of the Deutsch-Jozsa algorithm on an ion-trap quantum computer. Nature 421:48–50
76. Schmidt-Kaler F, Haffner H, Riebe M et al (2003) Realization of the Cirac-Zoller controlled-NOT quantum gate. Nature 422:408–411
77. Childs AM, Chuang IL (2001) Universal quantum computation with two-level trapped ions. Phys Rev A 6301:012306
78. Milburn GJ (1999) Simulating nonlinear spin models in an ion trap. arXiv:quant-ph/9908037v1
79. Sorensen A, Molmer K (1999) Quantum computation with ions in thermal motion. Phys Rev Lett 82:1971–1974
80. Molmer K, Sorensen A (1999) Multiparticle entanglement of hot trapped ions. Phys Rev Lett 82:1835–1838
81. Solano E, de Matos RL, Zagury N (1999) Deterministic Bell states and measurement of the motional state of two trapped ions. Phys Rev A 59:R2539–R2543
82. Sorensen A, Molmer K (2000) Entanglement and quantum computation with ions in thermal motion. Phys Rev A 6202:022311
83. Milburn GJ, Schneider S, James DFV (2000) Ion trap quantum computing with warm ions. Fortschritte Phys – Prog Phys 48:801–810
84. Sackett CA, Kielpinski D, King BE et al (2000) Experimental entanglement of four particles. Nature 404:256–259
85. Kielpinski D, Meyer V, Rowe MA et al (2001) A decoherence-free quantum memory using trapped ions. Science 291: 1013–1015
86. Haljan PC, Brickman KA, Deslauriers L et al (2005) Spin-dependent forces on trapped ions for phase-stable quantum gates and entangled states of spin and motion. Phys Rev Lett 94:153602
87. Haljan PC, Lee PJ, Brickman KA et al (2005) Entanglement of trapped-ion clock states. Phys Rev A 72:062316
88. Leibfried D, DeMarco B, Meyer V et al (2003) Experimental demonstration of a robust, high-fidelity geometric two ion-qubit phase gate. Nature 422:412–415
89. McDonnell MJ, Home JP, Lucas DM et al (2007) Long-lived mesoscopic entanglement outside the Lamb-Dicke regime. Phys Rev Lett 98:063603
90. Home JP, McDonnell MJ, Lucas DM et al (2006) Deterministic entanglement and tomography of ion-spin qubits. New J Phys 8:188
91. Monroe C, Leibfried D, King BE et al (1997) Simplified quantum logic with trapped ions. Phys Rev A 55:R2489–R2491
92. DeMarco B, Ben-Kish A, Leibfried D et al (2002) Experimental demonstration of a controlled-NOT wave-packet gate. Phys Rev Lett 89:267901
93. Garcia-Ripoll JJ, Zoller P, Cirac JI (2003) Speed optimized two-qubit gates with laser coherent control techniques for ion trap quantum computing. Phys Rev Lett 91:157901
94. Duan LM, Blinov BB, Moehring DL et al (2004) Scalable trapped ion quantum computation with a probabilistic ion-photon mapping. Quantum Inform Comput 4:165–173
95. Garcia-Ripoll JJ, Zoller P, Cirac JI (2005) Quantum information processing with cold atoms and trapped ions. J Phys B – At Mol Opt Phys 38:S567–S578
96. Zhu SL, Monroe C, Duan LM (2006) Trapped ion quantum computation with transverse phonon modes. Phys Rev Lett 97:050505
97. Leibfried D, Knill E, Seidelin S et al (2005) Creation of a six-atom 'Schroedinger cat' state. Nature 438:639–642
98. Barrett MD, Chiaverini J, Schaetz T et al (2004) Deterministic quantum teleportation of atomic qubits. Nature 429:737–739
99. Riebe M, Haffner H, Roos CF et al (2004) Deterministic quantum teleportation with atoms. Nature 429:734–737
100. Chiaverini J, Leibfried D, Schaetz T et al (2004) Realization of quantum error correction. Nature 432:602–605
101. Fisk PTH, Sellars MJ, Lawn MA et al (1995) Very high-q microwave spectroscopy on trapped yb-171$^+$ ions – application as a frequency standard. IEEE Trans Instrum Meas 44:113–116
102. Meekhof DM, Monroe C, King BE et al (1996) Generation of nonclassical motional states of a trapped atom. Phys Rev Lett 76:1796–1799
103. Palma GM, Suominen KA, Ekert AK (1996) Quantum computers and dissipation. Proc R Soc London Ser A-Math Phys Eng Sci 452:567–584
104. Barenco A, Berthiaume A, Deutsch D et al (1997) Stabilization of quantum computations by symmetrization. SIAM J Comput 26:1541–1557
105. Zanardi P, Rasetti M (1997) Noiseless quantum codes. Phys Rev Lett 79:3306–3309
106. Turchette QA, Kielpinski D, King BE et al (2000) Heating of trapped ions from the quantum ground state. Phys Rev A 6106:063418
107. Deslauriers L, Olmschenk S, Stick D et al (2006) Scaling and suppression of anomalous heating in ion traps. Phys Rev Lett 97:103007
108. Labaziewicz J, Richerme P, Brown KR et al (2007) Compact, filtered diode laser system for precision spectroscopy. Opt Lett 32:572–574
109. Turchette QA, Myatt CJ, King BE et al (2000) Decoherence and decay of motional quantum states of a trapped atom coupled to engineered reservoirs. Phys Rev A 62:053807
110. Schmidt-Kaler F, Gulde S, Riebe M et al (2003) The coherence of qubits based on single Ca+ ions. J Phys B-At Mol Opt Phys 36:623–636
111. Lucas DM, Keitch BC, Home JP et al (2007) A long-lived mem-

ory qubit on a low-decoherence quantum bus. arXiv:0710.4421v1 [quant-ph]
112. Steane AM (2002). Overhead and noise threshold of fault-tolerant quantum error correction. quant-ph/0207119
113. Barenco A, Bennett CH, Cleve R et al (1995) Elementary gates for quantum computation. Phys Rev A 52:3457–3467
114. Schmidt-Kaler F, Haffner H, Gulde S et al (2003) How to realize a universal quantum gate with trapped ions. Appl Phys B-Lasers Opt 77:789–796
115. Deutsch D (1985) Quantum-theory, the church-turing principle and the universal quantum computer. Proc R Soc London Ser A-Math Phys Eng Sci 400:97–117
116. Bennett CH, Brassard G, Crepeau C et al (1993) Teleporting an unknown quantum state via dual classical and einstein-podolsky-rosen channels. Phys Rev Lett 70:1895–1899
117. Shor PW (1994) Algorithms for quantum computation – discrete logarithms and factoring. In: Goldwasser S (ed) Proceedings of the 35th Annual Symposium on Foundations of Computer Science, Santa Fe. IEEE Computer Society Press, Washington, pp 124–134
118. Chiaverini J, Britton J, Leibfried D et al (2005) Implementation of the semiclassical quantum Fourier transform in a scalable system. Science 308:997–1000
119. Grover LK (1997) Quantum mechanics helps in searching for a needle in a haystack. Phys Rev Lett 79:325–328
120. Brickman KA, Haljan PC, Lee PJ et al (2005) Implementation of Grover's quantum search algorithm in a scalable system. Phys Rev A 72:050306
121. Bennett CH, DiVincenzo DP (2000) Quantum information and computation. Nature 404:247–255
122. Leibfried D, Barrett MD, Schaetz T et al (2004) Toward Heisenberg-limited spectroscopy with multiparticle entangled states. Science 304:1476–1478
123. Roos CF, Riebe M, Haffner H et al (2004) Control and measurement of three-qubit entangled states. Science 304:1478–1480
124. Haffner H, Hansel W, Roos CF et al (2005) Scalable multiparticle entanglement of trapped ions. Nature 438:643–646
125. Monroe C (2002) Quantum information processing with atoms and photons. Nature 416:238–246
126. Blinov BB, Moehring DL, Duan LM et al (2004) Observation of entanglement between a single trapped atom and a single photon. Nature 428:153–157
127. Moehring DL, Maunz P, Olmschenk S et al (2007) Entanglement of single-atom quantum bits at a distance. Nature 449:68–71
128. Guthöhrlein GR, Keller M, Hayasaka K et al (2001) A single ion as a nanoscopic probe of an optical field. Nature 414:49–51
129. Mundt AB, Kreuter A, Becher C et al (2002) Coupling a single atomic quantum bit to a high finesse optical cavity. Phys Rev Lett 89:103001
130. Cirac JI, Zoller P, Kimble HJ et al (1997) Quantum state transfer and entanglement distribution among distant nodes in a quantum network. Phys Rev Lett 78:3221–3224
131. Keller M, Lange B, Hayasaka K et al (2004) Continuous generation of single photons with controlled waveform in an ion-trap cavity system. Nature 431:1075–1078
132. Steane AM (2007) How to build a 300 bit, 1 giga-operation quantum computer. Quantum Inform Comput 7:171–183
133. Porras D, Cirac JI (2004) Effective quantum spin systems with trapped ions. Phys Rev Lett 92:207901
134. Friedenauer A, Schmitz H, Glueckert JT et al (2008) Simulating a quantum magnet with trapped ions. Nat Phys 4:757-761
135. Leibfried D, DeMarco B, Meyer V et al (2002) Trapped-ion quantum simulator: Experimental application to nonlinear interferometers. Phys Rev Lett 89:247901
136. Schutzhold R, Uhlmann M (2007) Analogue of cosmological particle creation in an ion trap. Phys Rev Lett 99:201301
137. Schmidt PO, Rosenband T, Langer C et al (2005) Spectroscopy using quantum logic. Science 309:749–752

Books and Reviews

Wineland DJ, Monroe C, Itano WM et al (1998) Trapped-ion quantum simulator. Phys Scr T76:147–151
Ekert A, Jozsa R (1996) Quantum computation and Shor's factoring algorithm. Rev Mod Phys 68:733–753
Steane A (1997) The ion trap quantum information processor. Appl Phys B – Lasers Opt 64:623–642
Wineland DJ, Barrett M, Britton J et al (2003) Quantum information processing with trapped ions. Philos Trans R Soc Lond Ser A-Math Phys Eng Sci 361:1349–1361
Leibfried D, Blatt R, Monroe C et al (2003) Quantum dynamics of single trapped ions. Rev Mod Phys 75:281–324
Steane A (1998) Quantum computing. Rep Prog Phys 61:117–173
Wineland DJ, Monroe C, Itano WM et al (1998) Experimental issues in coherent quantum-state manipulation of trapped atomic ions. J Res Natl Inst Stand Technol 103:259–328
Blatt R, Wineland D (2008) Entangled states of trapped atomic ions. Nature 453:1008–1015
Häffner H, Roos CF, Blatt R (2008) Quantum computing with trapped ions. Phys Rep 469:155–204

Quantum Computing Using Optics

Gerard J. Milburn, Andrew G. White
Centre for Quantum Computer Technology,
The University of Queensland, Brisbane, Australia

Article Outline

Glossary
Definition of the Subject
Introduction
Quantum Computation with Single Photons
Conditional Optical Two-Qubit Gates
Cluster State Methods
Experimental Implementations
Reprise: Single Photon States
Future Directions
Bibliography

Glossary

Avalanche photodiode (APD) A device for counting photons that absorbs a single photon and, with some probability, produces a single electrical signal.

Beam splitter A linear optical device that partially transmits, and partially reflects, an incoming beam of light.

Bell inequality The results of local measurements of dichotomic observables on each component of two correlated classical systems satisfy a correlation function that is less than or equal to a universal bound. This bound can be exceeded by correlated quantum systems.

Bell states Four orthogonal, maximally entangled states of two qubits that violate a Bell inequality.

Cluster State A highly entangled state of many qubits that enables quantum computation by sequences of conditional single qubit measurements, each conditioned on the results of previous measurements.

C-SIGN A two-qubit gate that leaves all logical states unchanged unless both qubits take the value one, in which case the state suffers a π phase shift.

Entangled state The state of a multi-component quantum system which cannot be expressed as a convex combination of tensor product states of each subsystem. Entangled states cannot be prepared by local operations on each subsystem, even when supplemented by classical communication.

HOM interference Hong, Ou and Mandel discovered that when indistinguishable single photon pulses arrive simultaneously at each of the two input ports of a fifty-fifty beam splitter, the probability for detecting two photons, co-incidently, at each of the two output ports drops to zero. In other words, the photons are either both reflected or both transmitted.

Mixed state A quantum state that is not completely known and thus has non zero-entropy.

Photon A single quantum excitation of an orthonormal optical mode. A field in such a state has definite intensity and completely random phase, so that the average field amplitude is zero.

Pure state A quantum state that is completely known and thus has zero entropy.

Quantum computation The ability to process information in a physical device using unitary evolution of superpositions of the physical states that encode the logical states.

Qubit The fundamental unit of quantum information in which two orthogonal states encode one bit of information. Unlike a classical bit, the physical system that forms the qubit can be in a superposition of the two logical states simultaneously.

Quantum teleportation A quantum communication protocol based on measurement, feed-forward and shared entanglement.

Shor's algorithm An algorithm for finding the prime factors of large integers by unitarily processing information in a quantum computer.

Tomography A measurement scheme for experimentally determining a quantum state in which a large sequence of physical systems, prepared in the same state, are subjected to measurements of a carefully chosen set of physical observables.

Unitary A transformation of a quantum state that is physically, and thus logically, reversible. Unitary transformations take pure states to pure quantum states.

Definition of the Subject

Quantum computation is a new approach to information processing based on physical devices that operate according to the quantum principles of superposition and unitary evolution [5,9]. This enables more efficient algorithms than are available for a computer operating according to classical principles, the most significant of which is Shor's efficient algorithm for finding the prime factors of a large integer [51]. There is no known efficient algorithm for this task on conventional computing hardware.

Optical implementations of quantum computing have largely focused on encoding quantum information using single photon states of light. For example, a single photon could be excited to one of two carefully defined orthogonal mode functions of the field with different momentum directions. However, as optical photons do not interact with each other directly, physical devices that enable one encoded bit of information to unitarily change another are hard to implement. In principle it can be done using a Kerr nonlinearity as was noted long ago [34,61], but Kerr nonlinear phase shifts are too small to be useful. Knill et al. [24] discovered another way in which the state of one photon could be made to act conditionally on the state of another using a measurement based scheme. We discuss this approach in some detail here as it has led to experiments that have already demonstrated many of the key elements required for quantum computation with optics.

Introduction

The optical fiber communication network is the largest artificial complex system on the planet, enabling the internet, the growth of economies and a massive connectivity of minds presaging a but dimly seen revolution. Information, encoded in pulses of light, courses through the system at more than 10 terabits per second. It is hard to believe that it is barely 20 years since the first transoceanic optical fiber was installed, yet the system is still growing at an astonish-

ing pace. The insatiable appetite for *bandwidth* in modern economies does not look like abating any time soon.

The entire system is held together by thin fibers of glass guiding pulses of light. The huge increase in bandwidth that modern communication networks have enabled follows directly from the very high carrier frequency of optical pulses. A wealth of modulation techniques have been developed to exploit the potential of this bandwidth: it is currently limited only by the speed of the switches in the network required to control the flow of information. Early networks required the switches to be largely electronic and this meant first converting the light pulses to electronic pulses. However increasingly these switches are being replaced with all-optical devices.

The routers and switches in the optical fiber network are essentially small computers processing packets of information at the fastest possible speeds. In the 1980s there was a research program to build computers entirely from optical switches processing information encoded optically. The astonishing progress in silicon technology meant that optical computers could never compete with the shrinking scale of semiconductors. The size of an optical switch is largely determined by the wavelength of light. This is beginning to change with the rise of plasmonics and nano photonics. However this constraint does not pose a problem for an optical communication system that spans the entire planet and much of the early work on all-optical switches found its way into the optical fiber system.

Optical fiber networks operate by the principles of classical physics; the computers they connect operate by the principles of classical logic. The semiconductors used as light sources and detectors function by quantum principles, but these do not influence network operation or logic. This is the physics and logic of Newton, Faraday and Maxwell. The motion of charges in semiconductor circuits is governed by the classical understanding of electric and magnetic fields, while the light pulses coursing down a glass fiber are perfectly well described by Maxwell's theory of electromagnetic radiation. However this situation will inevitably change and the first indications are already upon us. We have known since the early part of the last century that the deepest description of the physical world is not classical but quantum. It is now clear that the quantum world enables new computation and communication tasks that are difficult, if not impossible, in a classical world.

Over the same period that the optical communication system has been built, a small group of visionaries have speculated on the ultimate limits to conventional computation. There are two strands to this question. The first strand takes us down a road of ever decreasing dimensions. The astonishing growth in semiconductor technology is the direct result of acquiring the technical ability to make transistors smaller and smaller so that billions can fit onto a single chip. A very natural question is: how small can it get? Ultimately the answer to this question is the domain of quantum mechanics.

The second strand joins a more abstract path that began with von Neumann and Turing, and took a quantum twist in the mid 1980s when Feynman [9] asked if a physical computer operating by quantum principles would be a more efficient computational machine than a conventional classical computer. We now know that the answer is yes.

The quantum description of light began with Einstein in 1905 with his explanation of the photoelectric effect [7]:

> when a light ray starting from a point is propagated, the energy is not continuously distributed over an ever increasing volume, but it consists of a finite number of energy quanta, localized in space, which move without being divided and which can be absorbed or emitted only as a whole.

In Einstein's explanation, the energy of each quanta, or *photon* to use the modern word, is proportional to its frequency while the intensity of the light determines the number of photons per second passing through some given area. Einstein's insight is now routinely confirmed in a semiconductor device known as the avalanche photodiode (APD). This is a device that produces a current pulse when a single photon is absorbed. If we turn the intensity of the optical source down to very low levels and connect the APD to an audio amplifier we can hear the individual clicks as the photons arrive on the surface of the detector. In this operational sense, a single photon is a detection event, Fig. 1

The weak pulses of light used in conventional optical communication systems contain a huge number of pho-

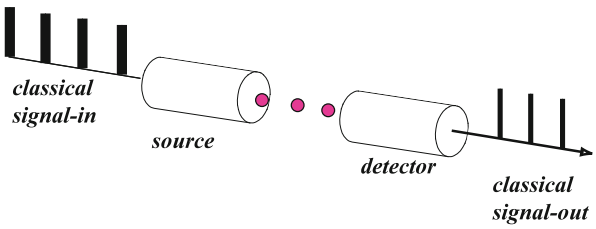

Quantum Computing Using Optics, Figure 1
A source generates a sequence of optical pulses conditional on some classical input, e. g. electrical pulses. A detector registers a sequence of classical electrical pulses that we interpret as due to the propagation of individual photons from source to detector. We might call this the *ballistic* picture of a photon

tons, and furthermore, the number of photons per pulse is not fixed but fluctuates from pulse to pulse. This is a direct consequence of the kind of light source that is used, the *laser*.

The laser is a true quantum device but it necessarily produces pulses of light with an indeterminate number of photons. The reason for this is intimately connected with the coherence properties of laser light. In a laser light source both the amplitude of the light and its phase fluctuate as little as quantum theory will permit. However these fluctuations are related by an uncertainty principle: if we try and produce a sequence of pulses with a well defined number of photons we necessarily make the phase random from pulse to pulse (we shall make this idea more precise in what follows). The random phase would destroy the key feature of first order optical coherence that makes a laser so useful.

Are there any light sources that controllably produce pulses of light each with one and only one photon per pulse? Until very recently, the answer was no. However the discovery of optical communication schemes (quantum key distribution) and computation schemes (as described below) based on quantum principles have given a strong incentive for building such sources. We will now explain how single photons enable quantum computation and postpone our discussion of exactly what a single photon source is to later.

Quantum Computation with Single Photons

Let us now consider the experiment depicted in Fig. 2. A source (labeled D for downward) is producing a sequence of single photons which are directed towards a 50/50 beam splitter. This is a classical optical device which in the wave theory would be described as simply dividing the wave intensity equally between a reflected beam and a transmitted beam. If we had a source of very high intensity, each of the photodetectors in the reflected path and the transmitted path would record on average an equal number of counts per second. As we reduce the intensity to the single photon level however, we see an irreducible uncertainty in which detector will register the photon in each trial. On average it is still the case that over many trials, one half are recorded at the D-detector and one half at the U-detector. If we were to persist in our *ballistic* view of a single photon, we might be tempted to say that each photon is either reflected or transmitted at random; it is a simple coin toss. This picture would adequately capture the experimental facts, *for this experiment*.

The beam splitter experiment has a binary outcome: the photon is either detected at U or D. To encode the result at the output we need a single bit of information. Indeed if we allow another single photon source into the upward going input to the beam splitter, we will need a single binary number to encode the input state as well. The kind of encoding we have just described is called *dual rail*. In more precise terms we have encoded a single bit into one of two perfectly distinguishable momentum states of a single photon.

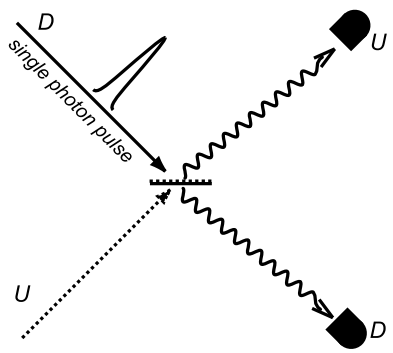

Quantum Computing Using Optics, Figure 2
A single photon pulse incident on a beam splitter in the downward (D) direction. It can be detected in the upward going direction (U) or the downward going direction (D)

At first sight, it would appear that the experiment we have just described is a fanciful coin toss. That this is not the case can be seen if we 'toss the coin again', that is to say, we take the photon after the beam splitter and, instead of running it into a photodetector, we reflect it back onto an identical beam splitter and then ask for the probability for it to be reflected or transmitted (see Fig. 3). It is well known that in this case we can adjust things so that the probability for the photon to be detected at U is certain. Irreducible uncertainty has been turned into certainty simply by changing the experimental conditions for detecting the photon, not by changing the state of the single photon source.

Note that the price we pay for certainty is complete loss of knowledge of what happened to the photon at the left beam splitter. A photon can be detected at either detector in Fig. 3 in two indistinguishable ways corresponding to the two (unknowable) outcomes of reflection or transmission at the left beam splitter. Returning to the simple experiment in Fig. 1, can we say that it is a simple coin toss before the detectors register the photon? Does the output of the beam splitter really encode one bit of information? No. We capture the dual quantum/classical nature of the state of a single photon after a beam splitter by saying that it is a *superposition* of the two mutually exclusive alternatives. In such a case the output is said to encode a single

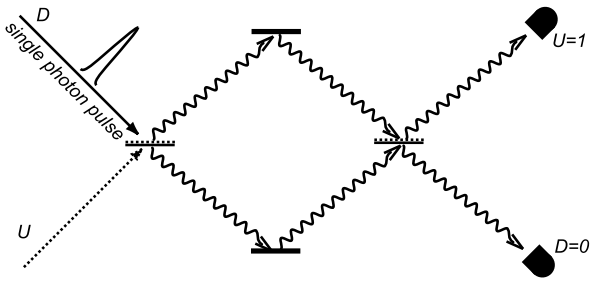

Quantum Computing Using Optics, Figure 3
A single photon pulse incident on a beam splitter in the downward (D) direction. It can be detected in the upward going direction (U) or the downward going direction (D)

quantum bit, or *qubit*. Quantum computation and communication runs on qubits not bits.

The quantum description of the state of the photon after a single beam splitter requires us to give a list of the *probability amplitudes* for the two mutually exclusive possibilities for detection in either the downward or the upward direction. The detection probabilities are then given by the square of the probability amplitudes. For example the input state in Fig. 1 is certain to be in the downward direction, so the ordered pair of probability amplitudes $(1, 0)$ in which the first (second) entry corresponds to detection in the D (U) direction. After a general (not 50/50) beam splitter the state of the photon is depicted by the ordered pair (r, t) where $|r|^2 + |t|^2 = 1$. So in fact the beam splitter has enacted the linear transformation $(1, 0) \to (r, t)$.

One might think that the input state $(0, 1)$ is transformed in exactly the same way. However these two input states are physically distinguishable: in one case the photon is going down and in the other it is going up with certainty. Mathematically this is captured by the fact that the input vectors are orthogonal. The beam splitter is a perfectly reversible device and does not destroy information by making two distinguishable alternatives indistinguishable. The problem is avoided when we note that a full quantum theory of beam splitters will in fact show that $(0, 1) \to (-r, t)$. There is a central lesson here: operations on qubits must not take two physically distinguishable states and make them indistinguishable. We call this sort of transformation *unitary*.

We now make a change to more conventional notation. The input states $(1, 0)$ and $(0, 1)$ are written as $|1, 0\rangle$ and $|0, 1\rangle$ (ordering is preserved) so that the beam splitter transformations are then written as

$$|1, 0\rangle \to r|1, 0\rangle + t|0, 1\rangle \tag{1}$$

$$|0, 1\rangle \to -r|1, 0\rangle + t|0, 1\rangle \,. \tag{2}$$

We now formalize the concept of a qubit by making a distinction between the *logical* state of a qubit and the physical states of the system used to encode it. In the example used here the relation ship is

$$|0\rangle = |1, 0\rangle \tag{3}$$

$$|1\rangle = |0, 1\rangle \,. \tag{4}$$

We then say that we have a qubit code that uses one photon and two optical modes per qubit. The beam splitter transformation on the logical code acts like

$$|0\rangle \to r|0\rangle + t|1\rangle \tag{5}$$

$$|1\rangle \to -r|0\rangle + t|1\rangle \,. \tag{6}$$

This transformation on logical qubits is called a *one-qubit gate*.

If we want to encode two qubits in this dual rail scheme we will need two single photon pulses and four modes, which may be achieved by two independent beam splitters. In a notation that keeps track of the ordering of the beam splitters by an ordering of states from left to right, the logical state $|1, 0\rangle$ would then be represented physically by

$$|1\rangle \otimes |0\rangle = |0, 1\rangle \otimes |1, 0\rangle \,. \tag{7}$$

Note that the number of mutually distinguishable output states from N beam splitters (and N photons) increases exponentially as 2^N. This simply means that for the logical code of N qubits there are 2^N possible logical states.

If we continue in this fashion we are not going to have a very compact notation. We can make things easier by using the fact that our qubits are ordered corresponding to a physical order of the underlying beam splitters. We can then represent a state, for example $|1\rangle \otimes |0\rangle \otimes |1\rangle \otimes |0\rangle$, by regarding it as the binary code for an integer, in this case $|10\rangle$ as $10 = 1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 0 \times 2^0$.

It is then clear that with N physical resources we can code 2^N binary numbers (or their corresponding integers). If we could somehow act on all these numbers simultaneously we might have a very efficient processor. This is precisely the hope of quantum computation. For example, if we pass N photons downward through N independent beam splitters, the state at the output is superposition state

$$|0\rangle \to 2^{-N/2} \sum_{x=0}^{2^N-1} |x\rangle \,, \tag{8}$$

an equal superposition of all integers from zero up to $2^N - 1$. If we could somehow continue to postpone mea-

surement and act unitarily on all these numbers simultaneously, we might gain a considerable advantage in computational efficiency. This was the idea first captured in precise form by David Deutsch [5].

Suppose for example we set aside some number, say N, of photons to encode the input register and some number, say K, to encode an output register. We first prepare the N photons so as to encode the state produced in Eq. (8) and the K to encode the state $|0\rangle$. Then the unitary processing through some, as yet unspecified, physical device could implement the transformation

$$\sum_x |x\rangle|0\rangle \rightarrow \sum_x |x\rangle|f(x)\rangle \,, \tag{9}$$

for some function f (we drop the normalization for simplicity). Of course now when we read out the output register we only get one value of the function with a random distribution. This may not sound very useful if you are in fact interested in a particular value of the function, but in that case why do all computations of the function at once? The primary reason why you might want to evaluate a function on all inputs is when you are not so much interested in any particular value but rather some global property of the function itself – for example is it constant? It is precisely that kind of computation that Deutsch [3] showed could be done more efficiently on a quantum computer.

A far more interesting example was described by Shor [51] who gave an efficient quantum algorithm for finding the period of a modular function. The particular modular function he used is part of an algorithm for finding the prime factors of large integers, so that in effect he gave an efficient quantum algorithm for finding such prime factors. The assumed intractability of this problem for a classical computer is why it is used as a method for public-key encryption. If someone had a quantum computer all these crypto systems would be vulnerable to attack. We describe below a simple optical experiment which captures the kind of methods by which a quantum computer could implement Shor's algorithm.

Of course no one yet has a quantum computer capable of posing a threat to public-key crypto systems, although it begins to look like it might be possible. The problem is to figure out what kind of physical systems would enable the arrow in Eq. (9). Various suggestions have been made including single photon optical systems, which we discuss below. At the level of logical qubits we need two types of elementary transformations. The first we have already seen: It is the single qubit transformation in Eq. (5). In addition all we need is some form of interaction that correlates the state of at least two qubits. One example is the controlled-SIGN or c-SIGN, gate,

$$|x\rangle|y\rangle \rightarrow (-1)^{x\cdot y}|x\rangle|y\rangle \,. \tag{10}$$

For single photon, dual rail encoding this presents a major problem.

If we look at the c-SIGN gate at the level of physical qubits it will require some kind of intensity dependent phase change so that only a two photon component acquires a π phase change. This is possible with a Kerr nonlinearity as was noted long ago [34,61]. Nonlinear optical materials with an intensity dependent refractive index do indeed exist and are generically referred to as Kerr nonlinear materials. The problem is that Kerr nonlinear phase shifts are so small that to expect an intensity as low as two photons to give a π phase change in an optical material of practical size is to expect too much. Fortunately there is another way.

Conditional Optical Two-Qubit Gates

Measurements play quite a different role in quantum mechanics than they do in classical mechanics. In the latter, it is the objective of accurate measurement to reveal the preexisting values of dynamical variables. Of course in reality both the measurement and the preparation of the initial state are subject to noise. However it is assumed that both may be sufficiently refined to reveal the true dynamical state of a system. If we know all there is to know about a classical system, all measurements are dispersion free in principle.

Quantum mechanics is an irreducibly statistical theory. This means that even if a system has been prepared in state about which we have maximal knowledge, there will be at least one measurement the results of which are uncertain. Fortunately there is also at least one physical quantity that may be measured for which the results are certain. If we make a perfect measurement of this quantity we have in effect prepared the system in a pure state of which we have maximal knowledge. Measurement thus plays an essential role in state preparation, where knowledge of the prepared state is heralded by the measurement result.

Can we use the unique role that measurement plays in quantum mechanics to conditionally implement a required state transformation? Consider the situation shown in Fig. 4. In a dual rail scenario, two modes are mixed on a beam splitter. One mode is assumed to be in the vacuum state (a) or a one photon state (b), while the other mode is arbitrary. A single photon counter is placed in the output port of mode-2. What is the conditional state of mode-1 given a count of n photons?

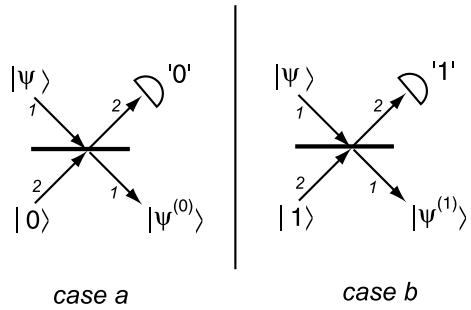

Quantum Computing Using Optics, Figure 4
A conditional state transformation conditioned on photon counting measurements

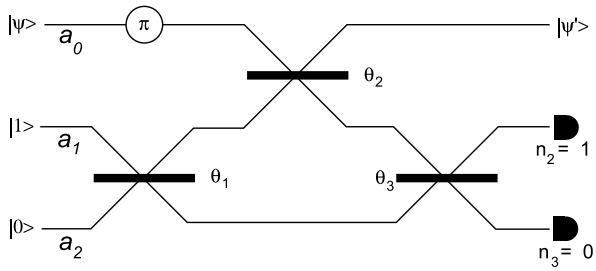

Quantum Computing Using Optics, Figure 5
A conditional state transformation on three optical modes, conditioned on photon counting measurements on the ancilla modes a_1, a_2. The signal mode, a_0 is subjected to a π phase shift

The conditional state of mode a_1 is given by (unnormalized),

$$|\tilde{\psi}^{(i)}\rangle_1 = \hat{\Upsilon}(i)|\psi\rangle_1 \,, \tag{11}$$

where

$$\hat{\Upsilon}(i) = {}_2\langle i|U(\theta)|i\rangle_2 \tag{12}$$

with $i = 1, 0$. Here $U(\theta)$ is the unitary transformation implemented by the beam splitter as we described in the introduction. The probability for the specified photo count event is given by,

$$P(i) = {}_1\langle\psi|\Upsilon^\dagger(i)\hat{\Upsilon}(i)|\psi\rangle_1 \,, \tag{13}$$

which fixes the normalization of the state,

$$|\psi^{(i)}\rangle_1 = \frac{1}{\sqrt{P(i)}}|\tilde{\psi}^{(i)}\rangle_1 \,. \tag{14}$$

It is then possible to show that

$$\hat{\Upsilon}(0) = \sum_{n=0}^{\infty} \frac{(\cos\theta - 1)^n}{n!}(a_1^\dagger)^n a_1^n$$

$$\hat{\Upsilon}(1) = \cos\hat{\Upsilon}^{(0)} - \sin^2\theta\, a_1^\dagger \hat{\Upsilon}^{(0)} a_1 \,.$$

In order to see how we can use these kind of transformations to effect a c-SIGN gate, consider the situation shown in Fig. 5. Three optical modes are mixed on a sequence of three beam splitters with beam splitter parameters θ_i, with corresponding reflection and transmission amplitudes, $r_i = \cos\theta_i$, $t_i = \sin\theta_i$. The *ancilla* modes, a_1, a_2 are restricted to be in the single photon states $|1\rangle_2, |0\rangle_3$ respectively.

We will assume that the *signal* mode, a_0, is restricted to have *at most* two photons, thus

$$|\psi\rangle = \alpha|0\rangle_0 + \beta|1\rangle_0 + \gamma|2\rangle_0 \,. \tag{15}$$

The reason we are only interested in two photon states is that in dual rail encoding a general two qubit state can have at most two photons. We can now chose the beam splitter parameters so that, conditional on the detectors both registering a one, the signal state is transformed as

$$|\psi\rangle \to |\psi'\rangle = \alpha|0\rangle + \beta|1\rangle - \gamma|2\rangle \,, \tag{16}$$

with a probability that is *independent* of the input state $|\psi\rangle$. This last condition is essential as in a quantum computation, the input state to a general two qubit gate is completely unknown. We will call this transformation the NS (for nonlinear sign shift) gate. This can be achieved using [24]: $\theta_1 = -\theta_3 = 22.5\,\text{deg}$ and $\theta_2 = 65.53\,\text{deg}$. The probability of the conditioning event $(n_2 = 1, n_3 = 0)$ is 1/4. Note that we can't be sure in a given trial if the correct transformation will be implemented. Such a gate is called a *nondeterministic* gate. However the key point is that success is heralded by the results on the photon counters (assuming ideal operation).

We can now proceed to a c-SIGN gate in the dual rail basis. Consider the situation depicted in Fig. 6. We first take two dual rail qubits encoding for $|1\rangle|1\rangle$. The single photon components of each qubit are directed towards a 50/50 beam splitter where they overlap perfectly in space and time and produces a state of the form $|0\rangle_2|2\rangle_3 + |2\rangle_2|0\rangle_3$, a effect known as Hong–Ou–Mandel (HOM) interference [15]. We then insert an NS gate into each output arm of the HOM interference. When the conditional gates in each arm work, which occurs with probability 1/16, the state is multiplied by an overall minus sign. Finally we direct these modes towards another HOM interference. The output state is thus seen to be $-|1\rangle|1\rangle$. One easily checks the three other cases for the input logical states to see that this device implements the c-SIGN gate with a probability of 1/16 and successful operation is heralded.

Clearly a sequence of nondeterministic gates is not going to be much use: the probability of success after

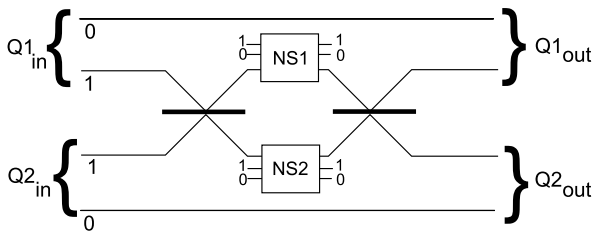

Quantum Computing Using Optics, Figure 6
A conditional state transformation conditioned on photon counting measurements. A c-SIGN gate that works with probability of 1/16. It uses HOM interference and two NS gates

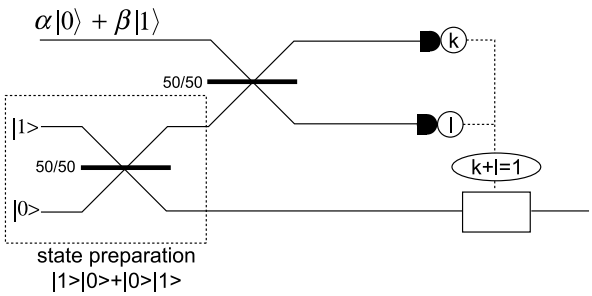

Quantum Computing Using Optics, Figure 7
A partial teleportation system for single photons states using a linear optics

a few steps will be exponentially small. The key idea in using nondeterministic gates for quantum computation is based on the idea of gate teleportation of Gottesmann and Chuang [12]. In quantum teleportation an unknown quantum state can be transferred from A to B provided A and B first share an entangled state. Gottesmann and Chuang realized that it is possible to simultaneously teleport a two qubit quantum state and implement a two qubit gate in the process by first applying the gate to the entangled state that A and B share prior to teleportation.

We use a non deterministic NS gate to prepare the required entangled state, and only complete the teleportation when the this stage is known to work. The teleportation step itself is non deterministic but, as we see below, by using the appropriate entangled resource the teleportation step can be made near deterministic. The near deterministic teleportation protocol requires only photon counting and fast feed-forward. We do not need to make measurements in a Bell basis.

A nondeterministic teleportation measurement is shown in Fig. 7. The client state is a one photon state in mode-0 $\alpha|0\rangle_0 + \beta|1\rangle_0$ and we prepare the entangled ancilla state

$$|t_1\rangle = |01\rangle_{12} + |10\rangle_{12} \tag{17}$$

where mode-1 is held by the sender, A, and mode-2 is held by the receiver, B. For simplicity we omit normalization constants wherever possible. This an ancilla state is easily generated from $|01\rangle_{12}$ by means of a beam splitter.

If the total count is $n_0 + n_1 = 0$ or $n_0 + n_1 = 2$, an effective measurement has been made on the client state and the teleportation has failed. However if $n_0 + n_1 = 1$, which occurs with probability 0.5, teleportation succeeds with the two possible conditional states being

$$\alpha|0\rangle_2 + \beta|1\rangle_2 \quad \text{if } n_0 = 1, n_1 = 0 \tag{18}$$

$$\alpha|0\rangle_2 - \beta|1\rangle_2 \quad \text{if } n_0 = 0, n_1 = 1\,. \tag{19}$$

When successful, this procedure implements a joint measurement on modes 0 and 1. In the conventional deterministic teleportation protocol the joint measurement is a simultaneous measurement of the commuting operators XX and ZZ where X, Z are respectively the Pauli-x and Pauli-z operators. This is a Bell measurement. In the teleportation protocol considered here, we only have a partial Bell measurement. We will refer to it as a non-deterministic teleportation protocol, $\mathbf{T}_{1/2}$. Note that teleportation failure is detected and corresponds to a photon number measurement of the state of the client qubit. Detected number measurements are a very special kind of error and can be easily corrected by a suitable error correction protocol. For further details see [24] and the review [26].

The next step is to use $\mathbf{T}_{1/2}$ to effect a conditional sign flip c-SIGN$_{1/4}$ which succeeds with probability 1/4. Note that to implement c-SIGN on two bosonic qubits in modes 1, 2 and 3, 4 respectively, we can first teleport the first modes of each qubit to two new modes (labeled 6 and 8) and then apply c-SIGN to the new modes. When using $\mathbf{T}_{1/2}$, we may need to apply a sign correction. Since this commutes with c-SIGN, there is nothing preventing us from applying c-SIGN to the prepared state before performing the measurements. The implementation is shown in Fig. 8 and now consists of first trying to prepare two copies of $|t_1\rangle$ with c-SIGN already applied, and then performing two partial Bell measurements. Given the prepared state, the probability of success is $(1/2)^2$. The state can be prepared using c-SIGN$_{1/16}$, which means that the preparation has to be retried an average of 16 times before it is possible to proceed.

The probability of successful teleportation can be boosted to $1 - 1/(n+1)$ using more entangled resource state of the kind

$$|t_n\rangle_{1\ldots 2n} = \sum_{j=0}^{n} |1\rangle^j |0\rangle^{n-j} |0\rangle^j |1\rangle^{n-j}\,. \tag{20}$$

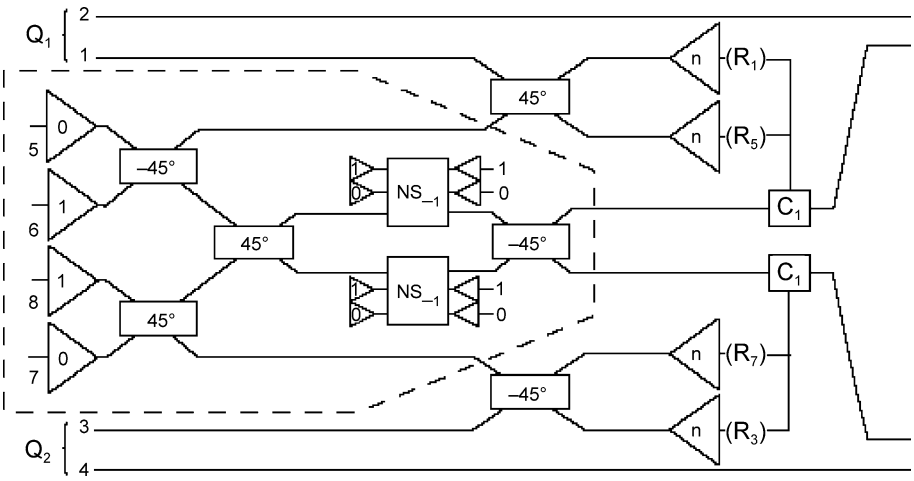

Quantum Computing Using Optics, Figure 8

A c-SIGN two qubit gate with teleportation to increase success probability to 1/4. When using the basic teleportation protocol (T1), we may need to apply a sign correction. Since this commutes with c-SIGN, it is possible to apply c-SIGN to the prepared state before performing the measurements, reducing the implementation of c-SIGN to a state-preparation (outlined) and two teleportations. The two teleportation measurements each succeed with probability 1/2, giving a net success probability of 1/4. The correction operations C1 consist of applying the phase shifter when required by the measurement outcomes

The notation $|a\rangle^j$ means $|a\rangle|a\rangle\ldots, j$ times. The modes are labeled from 1 to $2n$, left to right. Note that the state exists in the space of n bosonic qubits, where the kth qubit is encoded in modes $n+k$ and k (in this order).

We can teleport the state $\alpha|0\rangle_0 + \alpha|1\rangle_0$ using $|t_n\rangle_{1\ldots 2n}$. We first couple the client mode to half of the ancilla modes by applying an $n+1$ point Fourier transform on modes 0 to n. This is defined by the mode transformation

$$a_k \to \frac{1}{\sqrt{n+1}} \sum_{l=0}^{n} \omega^{kl} a_l \,, \tag{21}$$

where $\omega = e^{i2\pi/(n+1)}$. This transformation does not change the total photon number and is implementable with passive linear optics. After applying the Fourier transform, we measure the number of photons in each of the modes 0 to n. If the measurement detects k bosons altogether, it is possible to show [24] that if $0 < k < n+1$, then the teleported state appears in mode $n+k$ and only needs to be corrected by applying a phase shift. The modes $2n-l$ are in state 1 for $0 \le l < (n-k)$ and can be reused in future preparations requiring single bosons. The modes are in state 0 for $n-k < l < n$. If $k=0$ or $k=n+1$ an effective measurement of the client is made, and the teleportation fails. The probability of these two events is $1/(n+1)$, regardless of the input. Note that again failure is detected and corresponds to measurements in the basis $|0\rangle, |1\rangle$ with the outcome known. Note that both the necessary correction and the receiving mode are unknown until after the measurement.

Cluster State Methods

About the same time it was realized that measurements would provide a path to optical single photon computing, Raussendorf and Briegel [48] gave an independent and remarkably novel method by which measurement alone could be used to do quantum information processing. In their approach to quantum computation, an array of qubits is initially prepared in a special entangled state called a cluster state. The computation then proceeds by making a sequence of single qubit measurements. Each measurement is made in a basis that depends on prior measurement outcomes; in other words, the results of past measurements are fed forward to determine the basis for future measurements. Subsequently Popescu showed that the linear optical measurement based scheme of Knill et al. can be interpreted as a Raussendorf and Briegel measurement based quantum computation [44].

Nielsen [37] realized that the LOQC model of [24] could be used to efficiently assemble the cluster using the nondeterministic teleportation t_n. As we saw the failure mode of this gate constituted an accidental measurement of the qubit in the computational basis. The key point is that such an error does not destroy the entire assembled cluster but merely detaches one qubit from the cluster. This enables a protocol to be devised that produces a clus-

ter that grows on average. The LOQC cluster state method dramatically reduces the number of optical elements required to implement the original LOQC scheme. Of course if large single photon Kerr nonlinearities were available, the optical cluster state method could be made deterministic [16].

In its simplest form the Raussendorf and Breigel scheme begins with a two dimensional array of qubits. Each qubit is prepared in a superposition of the computational basis states, $|0\rangle + |1\rangle$. Then an entangling operation is performed between nearest neighbor qubits in the lattice using the two-qubit controlled sign operation

$$|x\rangle|y\rangle \mapsto (-1)^{xy}|x\rangle|y\rangle \,. \tag{22}$$

In the next step one or more qubits are measured in a particular basis and depending on the results of that measurement another basis is chosen for subsequent qubit measurements. Any circuit model of a quantum algorithm can be mapped onto the two dimensional lattice through a sequence of conditional measurements on subsets of qubits.

The key difficulty in doing this with a dual rail optical scheme is that, as we have noted, the transformation in Eq. (22) is very difficult to implement using a deterministic unitary transformation as optical nonlinearities are too small. However a conditional scheme of the kind discussed above might work by conditionally entangling sequences of qubits, provided failure at any point did not destroy the entire developing lattice of entanglement. This is indeed the case because of a key feature of the LOQC model of [24] scheme. If a teleportation gate failure is heralded, it corresponds to an effective measurement of one of the qubits at input. This feature was used by Knill et al. to establish the scalability of the scheme as detected measurements errors can easily be protected by a suitable code.

The importance of this feature for a conditional cluster state assembly is that an accidental measurement of one of the qubits in a developing cluster simply detaches it from the cluster without destroying the remaining entanglement. The picture then emerges of a kind of random cluster assembly in which the cluster grows when a gate succeeds and gets pruned if a gate fails. As long as the probability for gate success is greater than 0.5 there is a chance that the cluster overall will grow. As this does not require a very large teleportation resource, $|t_n\rangle_{1\ldots 2n}$, it can be achieved with quite modest overheads of linear optics and single photons. For this reason optical cluster state methods are preferred over the original LOQC scheme of Knill et al. A number of schemes have been proposed to efficiently assemble a cluster via this probabilistic growth. A recent application of percolation theory is a good example of the kind of optimization that is required [20].

Experimental Implementations

Two Qubit Gates

Considerable progress has been made on demonstrating conditional two-qubit gates with single photon states. After the initial proposal was made by Knill, Laflamme and Milburn [24], there was a flurry of theoretical work to propose linear-optical two-qubit gates that would be easier to realize experimentally. These can be divided into two categories, *internal-ancilla* gates [22,41,42,46], where the ancilla photons are intrinsic to the operation of the logic gates, and *external-ancilla* gates [14,47], where ancilla photons can be used to verify correct gate operation by performing a quantum non-demolition measurement [25] on the gate outputs. There are a variety of configurations of internal-ancilla gates, including simplified [46] and efficient [22] versions of the KLM gate; and gates that gain in efficiency using entangled ancilla photons [41,42].

An entangled internal-ancilla gate was soon realized in Johns Hopkins, where entangling operation was suggested by a $61.5 \pm 7.4\%$ visibility fringe [43]. An unambiguous demonstration of entangling gate operation was performed at the University of Queensland with an external-ancilla gate [39]: all four entangled Bell states were produced as a function of only the logical values of the input qubits, for a single operating condition of the gate. Both these gates filtered on photon-number, i. e. they required the four or two input photons to be detected to signal successful gate operation. An important requirement for LOQC is that it must be possible to detect successful gate operation by measurement of the ancilla photons and then feed-forward this information to the logic photons: this was realized with an external-ancilla gate at the University of Vienna [10].

Linear-optic gates, both internal and external ancilla, have achieved a wide range of firsts and proof-of-principle demonstrations: the first full characterization of a quantum-logic gate, in *any* architecture [38]; production of cluster [58] and graph [31] states, and their use for Grover's algorithm [58]; production of the highest entanglement [27] and fastest gate operation [45] of any physical architecture; and the first demonstration of Shor's algorithm exhibiting the core processes, coherent control, and resultant entangled states required in a full-scale implementation [29,30].

Measuring gate operation is non-trivial. Further, it is desirable to diagnose, and if possible correct, error behaviors introduced by a real gate, such as phase or bit-flip errors, which can induce the wrong amount or type of entanglement, or decoherence. To date, there have been

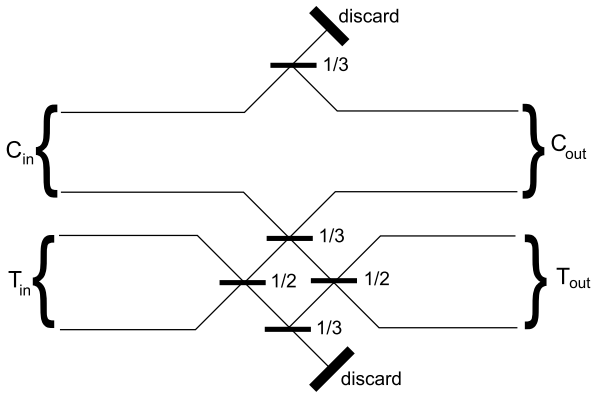

Quantum Computing Using Optics, Figure 9
An external-ancilla CNOT gate. When the control is in the logical-one state, the control and target photons interfere non-classically at the central 1/3 beam splitter which causes a π phase-shift in the upper arm of the central interferometer and the target state qubit is flipped. The qubit value of the control is unchanged. Correct operation has probability 1/9 and occurs when a single photon is detected in each output; this can be done with quantum non-demolition (QND) measurements using external ancilla photons [25] or by filtering on photon number in the final detection. Experimentally, the two modes of each qubit are distinguished by orthogonal polarizations. These may be split into separate spatial modes and sent through normal beamsplitters [38,39] or left in one spatial mode and sent through partially-polarizing beamsplitters [21,27,40]

a wide variety of measures used to gauge the quality of two-qubit gates. A comprehensive comparison of the various measures, and an architecture-independent measurement standard for two-qubit gates, is given in [60]. The key is quantum process tomography, which allows reconstruction of the quantum state transfer function of the gate. Tomography requires sampling the statistics of a fixed number of measurement outcomes, at least 256 for a two-qubit gate. With these statistics data inversion can be devised to reconstruct the gate process. From this we can calculate its overlap, or fidelity, with respect to the ideal quantum-logic gate, e. g. in recent experiments at the University of Queensland we obtained a process fidelity of $F_p = 98.2 \pm 0.3\%$ for a controlled-$\pi/4$ gate [28]. The process fidelity is itself not a metric [11], but forms the basis of several, such as the average gate fidelity, i. e. the fidelity of every gate output state with the ideal, averaged over every possible input state. This is given simply by,

$$\overline{F} = \frac{dF_p + 1}{d + 1} , \tag{23}$$

where d is the dimension of the process matrix, e. g. $d = 16$ for two-qubit gates.

These measures are useful for comparing gate performances, but do not provide an error probability-per-gate to allow direct comparison with fault-tolerance thresholds. A recent publication introduces a semidefinite programming technique to do exactly this, using the measured process matrix [59]. Such fault-tolerant benchmarking identifies the magnitude, but not the source of errors – critical in identifying the critical technology path for improving gate operation. This requires a comprehensive theoretical model of the quantum-logic gate, and its errors. Using such a model, Ref. [59] identifies small amounts of multi-photon emission as the dominant, yet previously unrecognized, source of error in linear-optical quantum computing – higher-order terms of 0.3–0.8% lead to gate errors of ∼16%! Removing multi-photon emission puts photonic quantum computing within striking distance of a recently predicted fault-tolerance threshold [23].

Quantum Optical Algorithms

The simplest useful algorithm for quantum optical information processing is the *quantum repeater*. This is more conventionally referred to as quantum communication protocol rather than an algorithm, but realistic systems will require few qubit quantum computers to do entanglement distillation and purification giving it an algorithmic flavor. Furthermore quantum repeaters could enable distributed quantum computation. We also mention it here as the various quantum optical repeater protocols that have been suggested [6,50,52] will have immediate application in long distance quantum key distribution and thus are pivotal to developing quantum optical networks across the global optical fiber network. This raises many new questions in the area of communication complexity and a great deal of work remains to be done to understand such systems [18,55].

The general idea of a quantum repeater is sketched in Fig. 10. There are three key elements at each node of the network: (i) a pair of entangled qubits, (ii) quantum memories, (iii) a few qubit quantum computer for purification distillation. In the first step two entangled pairs are produced at the origin, one of each pair is held there, while the other member of each pair is sent to distant locations. A measurement is made on the two qubits kept at the origin at step 1, leaving partially entangled qubits at the remote locations, step 2. The remote qubits are then loaded into quantum memory, step 3. The process is repeated, steps 4 and 5 until two qubits are held at the remote locations. A purification algorithm is then run at each remote location so that finally a maximally entangled pair is held at the remote locations separated by twice the distance that

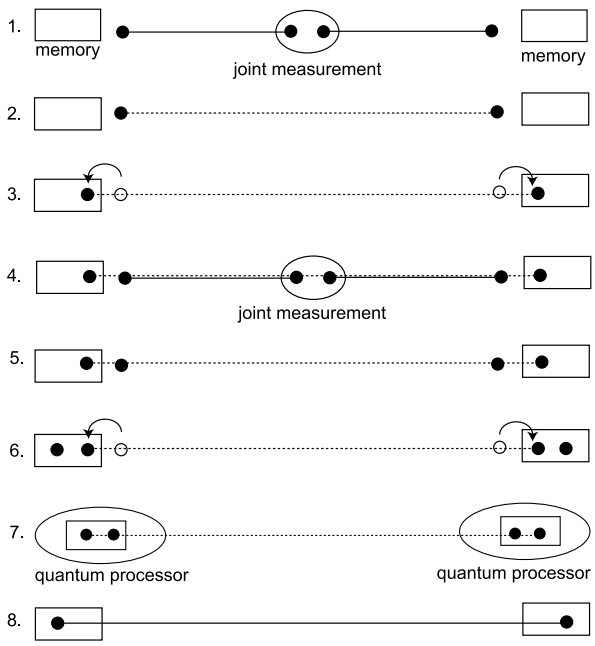

Quantum Computing Using Optics, Figure 10
A simple schematic for a quantum repeater

separated the pairs in step 1. In the quantum optical realization both the generation of entanglement and purification can be done by heralded non deterministic processes.

The previous scheme of course requires a good quantum memory and a great deal of research is underway to develop such systems for photons. One promising approach is to use polarized atomic ensembles [35]. The quantum memory itself may need to have error correction algorithms running to maintain the coherence of the entangled pairs until they required for some future quantum information processing task.

Many current cryptographic protocols rely on the computational difficulty of finding the prime factors of a large number: a small increase in the size of the number leads to an exponential increase in computational resources. Shor's quantum algorithm for factoring composite numbers faces no such limitation [51], and its realization represents a major challenge in quantum computation. Only one step of Shor's algorithm to find the factors of a number N requires a quantum routine. Given a randomly chosen co-prime C (where $1 < C < N$ and the greatest common divisor of C and N is 1), the quantum routine finds the *order* of C modulo N, defined to be the minimum integer r that satisfies $C^r \bmod N = 1$. It is straightforward to find the factors from the order. Consider $N = 15$: if we choose $C = 2$, the quantum routine finds $r = 4$, and the prime factors are given by the non-trivial greatest common divisor of $C^{r/2} \pm 1$ and N, i. e. 3 and 5; similarly if we choose the next possible co-prime, $C = 4$, we find the order $r = 2$, yielding the same factors.

The quantum routine in Shor's algorithm can factor a k-bit number using $72k^3$ elementary quantum gates, e. g. factoring the smallest meaningful number, 15, requires 4608 gates operating on 21 qubits [1], where the gates are one-, two-, or three-qubit logic gates. Recognizing this is well beyond the reach of current technology, Ref. [1] introduced a compiling technique which exploits properties of the number to be factored, allowing exploration of Shor's algorithm with a vastly reduced number of resources. The compiled algorithms are not scalable in themselves, but do allow the characterization of core processes required in a full-scale implementation of Shor's algorithm – including the ability to generate entanglement between qubits by coherent application of a series of quantum gates. In the only previous demonstration of Shor's algorithm, a compiled set of gate operations were implemented in a liquid NMR architecture [56]. Unfortunately, in such a system the qubits are at all times in a highly mixed state [2], and the dynamics can be fully modeled classically [32], so that neither the entanglement nor the coherent control at the core of Shor's algorithm can be implemented or verified.

The quantum order-finding routine consists of three distinct steps: i) *register initialization*, $|0\rangle^{\otimes n}|0\rangle^{\otimes m} \rightarrow (|0\rangle + |1\rangle)^{\otimes n}|0\rangle^{\otimes m-1}|1\rangle = \sum_{x=0}^{2^n-1} |x\rangle|0\rangle^{\otimes m-1}|1\rangle$, where the argument-register is prepared in an equal coherent superposition of all possible arguments (normalization omitted by convention); ii) *modular exponentiation*, which by controlled application of the order-finding function produces the entangled state $\sum_{x=0}^{2^n-1} |x\rangle|C^x \bmod N\rangle$; iii) the *inverse Quantum Fourier Transform* (QFT) followed by measurement of the argument-register in the logical basis, which with high probability extracts the order r after further classical processing. If the routine is standalone, the inverse QFT can be performed without two-qubit gates using an approach based on local measurement and feed-forward.

In recent experiments, Shor's algorithm was performed for $N = 15$ using internal-ancilla logic gates [29, 30]. Co-primes of $C = 4$ and $C = 2$ were realized using circuits of two controlled-SIGN gates, with three and four qubit inputs, respectively. To no-ones surprise, the prime factors of 3 and 5 were found. What was surprising is that was near-ideal algorithm performance – far better than expected given the known errors inherent in the logic gates. This result highlights a subtle point with respect to benchmarking quantum algorithms: *algorithm performance* is not necessarily an accurate indicator of

the underlying *circuit performance*. This is particularly the case in Shor's algorithm, where ideally the algorithm produces mixed states. From algorithm performance alone it is not possible to distinguish between the desired mixture arising from entanglement with the function-register, and the undesired mixture due to environmental decoherence. In [29] this was quantified by using quantum state tomography [17] to measure the *joint* state of the argument and function registers after modular exponentiation. The joint state was both mixed and entangled, only partially overlapping with the expected states, a sure indication of environmental decoherence. Such accurate knowledge of circuit performance is crucial if the circuits are to be incorporated as sub-routines in larger algorithms.

Reprise: Single Photon States

The two key requirements for conditional quantum computing with single photons are; (i) single photon sources and, (ii) highly efficient single photon detectors capable of resolving zero, one or two photon detection events. Of course we will need a lot more than just this. In order to make the scheme scalable some kind of photonic memory, or fast switching, or both, would be desirable. Once a particular entangled multi photon state resource has been prepared it might be necessary to store it for some time until it is required for use. We first discuss in some detail what is required for a single photon source.

It is now time to become much more precise about what is meant by a single photon state. The quantum electromagnetic field is described by an electric field operator [57],

$$\vec{E}(\vec{x},t) = i\sum_{n,\nu}\sqrt{\frac{\hbar\omega_n}{2\epsilon_o V}}\,\vec{e}_{l,\nu}\left[\mathrm{e}^{i(\vec{k}_n.\vec{x}-\omega_n t)}a_{n,\nu} - \mathrm{e}^{-i(\vec{k}_n.\vec{x}-\omega_n t)}a^{\dagger}_{n,\nu}\right], \quad (24)$$

where $\vec{e}_{n,\nu}$ are two orthonormal polarization vectors ($\nu = 1,2$) which satisfy $\vec{k}_n.\vec{e}_{n,\nu} = 0$, as required for a transverse field, and the frequency is given by the dispersion relation $\omega_n = c|\vec{k}_n|$. The positive and negative frequency amplitude operators are respectively $a_{n,\nu}$ and $a^{\dagger}_{m,\nu}$, with bosonic commutation relations

$$[a_{n,\nu}, a_{n',\nu'}] = \delta_{\nu\nu'}\delta_{nn'}\,, \quad (25)$$

with all other commutations relations zero.

Typically we are interested in sources defined by optical cavity modes so that emission is directional and defined by the cavity spatio-temporal mode structure. Much of the difficulty in building single photon sources is in designing the optical cavity to ensure emission into a preferred set of modes. We will assume that the only modes that are excited have the same plane polarization and are all propagating in the same direction, which we take to be the x-direction. The positive frequency components of the quantum electric field for these modes are then

$$E^{(+)}(x,t) = i\sum_{n=0}^{\infty}\left(\frac{\hbar\omega_n}{2\epsilon_0 V}\right)^{1/2} a_n \mathrm{e}^{-i\omega_n(t-x/c)}\,. \quad (26)$$

In ignoring all the other modes, we are implicitly assuming that all our measurements do not respond to the vacuum state, an assumption which is justified by the theory of photon-electron detectors [57]. Let us further assume that all excited modes of this form have frequencies centered on *carrier frequency* of $\Omega \gg 1$. Then we can approximate the positive frequency components by

$$E^{(+)}(x,t) = i\left(\frac{\hbar\Omega_n}{2\epsilon_0 Ac}\right)^{1/2}\sqrt{\frac{c}{L}}\sum_{n=0}^{\infty} a_n \mathrm{e}^{-i\omega_n(t-x/c)} \quad (27)$$

where A is a characteristic transverse area. This operator has dimensions of electric field. In order to simplify the dimensions we now define a field operator that has dimensions of $\mathrm{s}^{-1/2}$. Taking the continuum limit we thus define the positive frequency operator

$$a(x,t) = \mathrm{e}^{-i\Omega(t-x/c)}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathrm{d}\omega' a(\omega')\mathrm{e}^{-i\omega'(t-x/c)}\,, \quad (28)$$

where we have made a change of variable $\omega \mapsto \Omega + \omega'$ and used the fact that $\Omega \gg 1$ to set the lower limit of integration to minus infinity, and

$$[a(\omega_1), a^{\dagger}(\omega_2)] = \delta(\omega_1 - \omega_2)\,. \quad (29)$$

In this form the moment $n(x,t) = \langle a^{\dagger}(x,t)a(x,t)\rangle$ has units of s^{-1}. This moment determines the probability per unit time (the count rate) to count a photon at space-time point (x,t) [57]. The field operators $a(t)$ and $a^{\dagger}(t)$ can be taken to describe the field emitted from the end of an optical cavity, which selects the directionality.

We will contrast single photon states with multimode coherent states defined by a multimode displacement operator acting on the vacuum $D|0\rangle$, defined implicitly by

$$D^{\dagger}a(\omega)D = a(\omega) + \alpha(\omega)\,, \quad (30)$$

where consistent with proceeding assumptions, $\alpha(\omega)$ is peaked at $\omega = 0$ which corresponds to a carrier frequency

$\Omega \gg 1$. The average field amplitude for this state is

$$\langle a(x,t)\rangle = \mathrm{e}^{-i\Omega(t-x/c)}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\alpha(\omega)\mathrm{e}^{-i\omega(t-x/c)} \equiv \alpha(x,t)\mathrm{e}^{-i\Omega(t-x/c)}\,, \tag{31}$$

which implicitly defines the average complex amplitude of the field as the Fourier transform of the frequency dependent displacements $\alpha(\omega)$ We can also calculate the probability per unit time to detect a photon in this state at space-time point (x,t). This is given by $n(x,t) = |\alpha(x,t)|^2$. Note that in this case the second order moment $\langle a^\dagger(x,t)a(x,t)\rangle$ factories, a result characteristic of fields with first order coherence. A coherent state is is closest to our intuitive idea of a classical electromagnetic field.

The multimode single photon state is defined by

$$|1\rangle = \int_{-\infty}^{\infty}\nu(\omega)a^\dagger(\omega)|0\rangle\,. \tag{32}$$

Normalization requires that

$$\int_{-\infty}^{\infty}\mathrm{d}\omega|\nu(\omega)|^2 = 1\,. \tag{33}$$

This last condition implies that the total number of photons, integrated over all modes, is unity,

$$\int_{-\infty}^{\infty}\mathrm{d}\omega\langle a^\dagger(\omega)a(\omega)\rangle = 1\,. \tag{34}$$

This state has zero average field amplitude but

$$n(x,t) = |\nu(t-x/c)|^2\,, \tag{35}$$

where $\nu(t)$ is the Fourier transform of $\nu(\omega)$. So while the state has zero average field amplitude there is apparently some sense in which the coherence implicit in the superposition of Eq. (32) is manifest. In fact comparing this to the case of a coherent state, Eq. (31), we see that the expression for $n(t)$ is also determined by the Fourier transform of a coherent amplitude. For this state the function $\nu(\phi)$ is periodic in the phase $\phi = t - x/c$ and it is not difficult to choose a form with a well defined pulse sequence. However care should be exercised in interpreting these pulses. They do not represent a sequence of pulses each with one photon rather they represent a single photon coherently superposed over all pulses. Once a photon is counted in a particular pulse, the field is returned to the vacuum state. A review of current efforts to produce such a state may be found in [53]. More recent results may be found in [4,8,13,19].

In the absence of true single photon pulse sources, most of the experimental work we have described has been done with a conditional source based on photon pair production in spontaneous parametric down conversion. We now give a model of the kind of state such sources produce.

In the case of a continuous wave pump, the entangled two photon produced by SPDC may be well approximated by

$$|\psi\rangle = \int_0^{\infty}\mathrm{d}\omega_1\int_0^{\infty}\mathrm{d}\omega_2\,\rho(\omega_1,\omega_2)a^\dagger(\omega_1)b^\dagger(\omega_2)|0\rangle\,, \tag{36}$$

where a, b designate distinguishable modes, for example, distinguished either by polarization or wave vector [54]. More precisely, this is the conditional state given that a down conversion process has in fact taken place at all. The probability per unit time for this event is the down conversion efficiency. For spontaneous parametric down conversion with a continuous pump field at frequency 2Ω we can approximate

$$\rho(\omega_1,\omega_2) = \delta(\omega_1+\omega_2-2\Omega)\alpha(\omega_1)\,. \tag{37}$$

We now make the change of variable $\epsilon = \Omega - \omega_1$ and assume that the bandwidth, B, over which $\alpha(\omega_1)$ is significantly different from zero is such that $\Omega \gg B$, then we can write

$$|\psi\rangle = \int_{-\infty}^{\infty}\mathrm{d}\epsilon\beta(\epsilon)a^\dagger(-\epsilon)b^\dagger(\epsilon)|0\rangle\,, \tag{38}$$

where we have defined $\beta(\epsilon) = \alpha(\Omega-\epsilon)$ and $a(-\epsilon) \equiv a(\Omega-\epsilon), b(\epsilon) \equiv b(\Omega+\epsilon)$. We will also assume that $\beta(-\epsilon) = \beta(\epsilon)$. Normalization of $|\psi\rangle$ requires that

$$\int_{-\infty}^{\infty}|\beta(\epsilon)|^2 = 1\,. \tag{39}$$

The probability per unit time to detect a photon from this field with a unit efficiency detector is in fact unity, $n_a(t) = \langle a^\dagger(t)a(t)\rangle = 1$ The probability per unit time to detect a photon from mode-a is thus independent of time. This simply means that photons will be counted at randomly distributed times from each of the fields a, b. This is a reflection of the fact that the state Eq. (38) is invariant under time translations in a Lorentzian frame.

On the other hand let us now compute the coincidence rate,

$$C(t,t') = \langle\psi|a^\dagger(x,t')a(x,t')b^\dagger(x,t)b(x,t)|\psi\rangle\,. \tag{40}$$

This is given by

$$C(t,t') = \left|\int_{-\infty}^{\infty}\beta(\epsilon)\mathrm{e}^{-i\epsilon(t'-t)}\right|^2 \tag{41}$$

$$= |\tilde{\beta}(\tau)|^2 \tag{42}$$

$$\equiv C(\tau)\,, \tag{43}$$

where $\tau = t' - t$. The coincidence rate is thus symmetrical about $t' - t = 0$ and is peaked at $t' = t$. Even though

photons are detected at random, independently from each beam, they are highly correlated in time. In the case of spontaneous SPDC the distributions function $\beta(\epsilon)$ is given approximately [49]

$$\beta(\epsilon) \propto \frac{1}{\kappa^2 + \omega^2} \,. \tag{44}$$

We can now consider a heralded single photon source made by detecting one of the photon pairs and then ask for the kind of single photon state conditionally produced in the other mode. In the case of a CW pump, we first must provide a temporal filter on the detected mode. One can think of this as a time dependent detector that is switched on an off over some time interval. In frequency domain this is simply a filter. If such a detector is placed at the b mode, the conditional state of the a mode is given by Eq. (32) with

$$\nu(\omega) \propto \mathrm{e}^{-\omega^2/\kappa^2} \,, \tag{45}$$

which corresponds to a Gaussian temporal pulse.

If we drive SPDC with a pulsed pump, the pump itself provides a natural temporal filter. In this case the frequency distribution function in Eq. (37) is not delta correlated but takes the form [54]

$$\rho(\omega_1, \omega_2) = \exp\left[-(\omega_1 + \omega_2 - 2\Omega)^2/\sigma_\mathrm{p}^2\right] \alpha(\omega_1) \,, \tag{46}$$

where Ω is the pump carrier frequency and σ_p is the bandwidth of the pump pulse. It is still the case however that no photon down conversion event may take place within the pump pulse window. Thus the source is not a deterministic single photon source. Migdall [33] has proposed a way to over come this by multiplexing many heralded conditional SPDC sources with a conditioning detection on one mode of each of the multiplexed pairs. Another approach has been implemented by the Polzik group [36]. They used a cavity to enhance the parametric down conversion to implement a frequency tunable source of heralded single photons with a narrow bandwidth of 8 MHz. This approach is particularly important as frequency tunability makes the source compatible with atomic quantum memories.

Future Directions

Optical systems are certain to be used for future quantum communication protocols. Indeed the first steps have already been taken with quantum key distribution. In this article we have seen that it is also possible to process quantum information optically using heralded non deterministic schemes of various kinds and simple examples have been implemented experimentally. This greatly enhances the practicability of quantum communication schemes that require some quantum processing, such as quantum repeaters. While this has not yet been realized in practice, we expect that the first demonstrations are not far away.

In the effort to produce single photon sources we are learning new ways to encode and process information in optical pulses. We have seen that coherent communication can be done using single photons despite the fact that the average field amplitude for such states is zero. If information can be encoded and decoded on photon number states, this would represent a major step beyond quantum communication protocols like quantum key distribution.

It is far from clear however if optical schemes will be viable for large scale quantum computation. Currently ion trap schemes and schemes based on super-conducting devices offer the most likely way forward for quantum computation per se. However a number of investigators are turning to the concept of a hybrid quantum computer which combines optical and matter based qubits. Optical qubits with heralded non deterministic processing combined with matter based quantum memories is poised to make significant achievements.

Hybrid schemes offer a path to distributed quantum computation between many nodes each made up of a few hundred qubits. The nodes do not need to be far apart: they could simply be different parts of a single quantum computation device. However, if you will permit us some license, it is not too difficult to imagine a single quantum computation spanning the entire planet, with matter based nodes connected by quantum optical communication channels. Such a system would hold in its web massively entangled quantum states and exhibit a complexity that would make our current optical communication system look rather simple by comparison.

Bibliography

1. Beckman D, Chari AN, Devabhaktuni S, Preskill J (1996) Phys Rev A 54:1034
2. Braunstein SL et al (1999) Phys Rev Lett 83:1054
3. Cleve R, Ekert A, Henderson L, Macchiavello C, Mosca M (1999) On quantum algorithms. LANL quant-ph/9903061
4. Darquie B, Jones MPA, Dingjan J, Beugnon J, Bergamini S, Sortais Y, Messin G, Browaeys A, Grangier P (2005) Science 309:454
5. Deutsch D (1985) Quantum-theory, the Church–Turing principle and the universal quantum computer. Proc R Soc Lond A 400:97–117
6. Duan L-M, Lukin MD, Cirac JI, Zoller P (2001) Nature 414:413
7. Einstein A (1905) On a Heuristic Point of View about the Creation and Conversion of Light? Ann Physik 17:132
8. Eisaman MD, Fleischhauer M, Lukin MD, Zibrov AS (2006) To-

ward quantum control of single photons. Opt Photonics News 17:22–27
9. Feynman RP (1982) Simulating physics with computers. Int J Theor Phys 21:467
10. Gasparoni S et al (2004) Phys Rev Lett 93:020504
11. Gilchrist A, Langford NK, Nielsen MA (2005) Distance measures to compare real and ideal quantum processes. Phys Rev A 71:062310
12. Gottesman D, Chuang IL (1999) Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations. Nature 402:390–393
13. Hennrich M, Legero T, Kuhn A, Rempe G (2004) Photon statistics of a non-stationary periodically driven single-photon source. New J Phys 6:86
14. Hofmann H, Takeuchi S (2002) Phys Rev A 66:024308
15. Hong CK, Ou ZY, Mandel L (1987) Measurement of subpicosecond time intervals between two photons by interference. Phys Rev Lett 59:2044
16. Hutchinson GD, Milburn GJ (2004) Nonlinear quantum optical computing via measurement. J Modern Opt 51:1211–1222
17. James DFV, Kwiat PG, Munro WJ, White AG (2001) Measurement of qubits. Phys Rev A 64:052312
18. Jiang L, Taylor JM, Khaneja N, Lukin MD (2007) Optimal approach to quantum communication using dynamic programming. arXiv:quant-ph/0710.5808
19. Keller M, Lange B, Hayasaka K, Lange W, Walther H (2003) Continuous generation of single photons with controlled waveform in an ion-trap cavity system. Nature 431:1075
20. Kieling K, Rudolph T, Eisert J (2007) Percolation, renormalization, and quantum computing with nondeterministic gates. Phys Rev Lett 99:130501
21. Kiesel N, Schmid C, Weber U, Ursin R, Weinfurter H (2007) Phys Rev Lett 99:250505
22. Knill E (2002) Phys Rev A 66:052306
23. Knill E (2005) Quantum computing with realistically noisy devices. Nature 434:39–44
24. Knill E, Laflamme R, Milburn GJ (2001) Efficient linear optical quantum computation. Nature 409:46
25. Kok P, Lee H, Dowling JP (2002) Phys Rev A 66:063814
26. Kok P, Munro WJ, Nemoto K, Ralph TC, Dowling JP, Milburn GJ (2007) Linear optical quantum computing. Rev Mod Phys 79:135
27. Langford NK, Weinhold TJ, Prevedel R, Gilchrist A, O'Brien JL, Pryde GJ, White AG (2005) Phys Rev Lett 95:210504
28. Lanyon BP, Barbieri M, Almeida MP, Jennewein T, Ralph TC, Resch KJ, Pryde G, O'Brien JL, Gilchrist A, White AG (2007) Quantum computing using shortcuts through higher dimensions. appear in Nat Phys
29. Lanyon BP, Weinhold TJ, Langford NK, Barbieri M, James DFV, Gilchrist A, White AG (2007) Phys Rev Lett 99:250505
30. Lu C-Y, Browne DE, Yang T, Pan J-W (2007) Phys Rev Lett 99:250504
31. Lu C-Y, Zhou X-Q, Gühne O, Gao W-B, Zhang J, Yuan Z-S, Goebel A, Yang T, Pan J-W (2007) Experimental entanglement of six photons in graph states. Nature Phys 3:91
32. Menicucci NC et al (2002) Phys Rev Lett 88:167901
33. Migdall A et al (2002) Phys Rev A 66:053805
34. Milburn GJ (1989) A quantum optical fredkin gate. Phys Rev Lett 62:2124–2127
35. Mishina OS, Kupriyanov DV, Muller JH, Polzik ES (2007) Spectral theory of quantum memory and entanglement via Raman scattering of light by an atomic ensemble. Phys Rev A 75:042326
36. Neergaard-Nielsen JS, Melholt Nielsen B, Takahashi H, Vistnes AI, Polzik ES (2007) High purity bright single photon source. Opt Express 15:7940
37. Nielsen MA (2004) Optical quantum computation using cluster states. Phys Rev Lett 93:040503
38. O'Brien JL et al (2004) Phys Rev Lett 93:080502
39. O'Brien JL, Pryde GJ, White AG, Ralph TC Branning D (2003) Demonstration of an all-optical quantum controlled-NOT gate. Nature 426:264
40. Okamoto R, Hofmann HF, Takeuchi S, Sasaki K (2007) Phys Rev Lett 99:250506
41. Pittman TB, Jacobs BC, Franson JD (2001) Probabilistic quantum logic operations using polarizing beam splitters. Phys Rev A 64:062311
42. Pittman TB, Jacobs BC, Franson JD (2002) Phys Rev Lett 88:257902
43. Pittman TB, Fitch MJ, Jacobs BC, Franson JD (2003) Experimental controlled-NOT logic gate for single photons in the coincidence basis. Phys Rev A 68:032316
44. Popescu S (2006) KLM quantum computation as a measurement based computation. arXiv:quant-ph/0610025
45. Prevedel R et al (2007) Nature 445:65
46. Ralph TC, White AG, Munro WJ, Milburn GJ (2001) Simple scheme for efficient linear optics quantum gates. Phys Rev A 65:012314
47. Ralph TC, Langford NK, Bell TB, White AG (2002) Linear optical controlled-NOT gate in the coincidence basis. Phys Rev A 65:062324
48. Raussendorf R, Briegel HJ (2001) A one-way quantum computer. Phys Rev Lett 86:5188
49. Rhode PP, Ralph TC (2005) Phys Rev A 71:032320
50. Sangouard N, Simon C, Minar J, Zbinden H, de Riedmatten H, Gisin N (2007) Long-distance entanglement distribution with single-photon sources. arXiv:quant-ph/0706.1924v1
51. Shor P (1994) Algorithms for quantum computation: Discrete logarithms and factoring. Proc 35th annual symposium on foundations of computer science. See also LANL preprint quant-ph/9508027
52. Simon C et al (2007) Phys Rev Lett 98:190503
53. Special Issue on Single photon Sources (2004) Single photon Sources 51(9–10)
54. U'Ren AB, Mukamel E, Banaszek K, Walmsley IA (2003) Phil Trans Roy Soc A 361:1471
55. Van Meter R, Ladd TD, Munro WJ, Nemoto K (2007) System Design for a Long-Line Quantum Repeater. arXiv:quant-ph/0705.4128v1
56. Vandersypen LMK et al (2001) Nature 414:883
57. Walls DF, Milburn GJ (2008) Quantum Optics, 2nd edn. Springer, Berlin
58. Walther P, Resch KJ, Rudolph T, Schenck E, Weinfurter H, Vedral V, Aspelmeyer M, Zeilinger A (2005) Experimental one-way quantum computing. Nature 434:169
59. Weinhold TJ, Gilchrist A, Resch KJ, Doherty AC, O'Brien JL, Pryde GJ, White AG (2008) Understanding photonic quantum-logic gates: The road to fault tolerance. arxiv 0808.0794
60. White AG et al (2007) Measuring two-qubit gates. J Opt Soc Am B 24:172–183
61. Yamamoto Y, Kitagawa M, Igeta K (1988) Proc 3rd Asia-Pacific phys conf 779. World Scientific, Singapore

Quantum Cryptography

Hoi-Kwong Lo, Yi Zhao
Center for Quantum Information and Quantum Control, Department of Physics and Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
Quantum Key Distribution: Motivation and Introduction
Security Proofs
Experimental Fundamentals
Experimental Implementation of BB84 Protocol
Other Quantum Key Distribution Protocols
Quantum Hacking
Beyond Quantum Key Distribution
Future Directions
Acknowledgments
Bibliography

Glossary

One-time pad One-time pad is a classical encryption algorithm invented by Gilbert Vernam in 1917. In one-time pad algorithm, the legitimate users share a random key (e. g. a random binary string) that is not known to anyone else. The message is combined with this random key ("pad") which is as long as the message. The key is used only once ("one-time"). The most typical usage is in binary case, where an XOR operation is applied between the message and the key to achieve the ciphertext. Claudé Shannon proved that the one-time pad provides perfect secrecy in 1949. The perfect secrecy is defined that the ciphertext does not give any additional information on the message.

Key distribution problem The key distribution problem originates from the one-time pad encryption. The one-time pad encryption requires that the two parties share a secret random key before the communication. This key is usually generated by one party. The key distribution problem is how to distribute this random key from one party to the other party securely. This problem is non-solvable classically, but is solvable via quantum key distribution.

Quantum key distribution Quantum key distribution (QKD) is a method to distribute a random key between two parties securely. The main idea is to encode the bit value on the quantum state of certain particle (usually photon) and send the particle to the receiver. The quantum no-cloning theorem guaranteed that any eavesdropper cannot duplicate the encoded quantum information perfectly.

BB84 BB84 is the first and so far the most popular quantum cryptography protocol. It was proposed by C. H. Bennet and G. Brassard in 1984 [1]. In the original BB84 proposal, the quantum information is encoded on the polarizations of photons. Later BB84 was extended to the phase coding. Detailed description of BB84 protocol can be found in Sect. "Introduction".

B92 B92 is a quantum cryptography protocol proposed by C. H. Bennet in 1992 [2]. It uses two non-orthogonal states (e. g. the horizontally – and 45° polarized photons) to denote "0" and "1". It is simpler than BB84 protocol in implementation.

E91 E91 is a quantum cryptography protocol proposed by A. Ekert in 1991 [3]. It is based on entangled photon pairs. E91 protocol is often used in free-space quantum key distribution.

Uni-directional QKD Uni-directional QKD is the QKD scheme in which Alice (sender) generates the photon, encodes the quantum information on it, and sends it to Bob (receiver).

Bi-directional QKD, or "Plug & play" QKD Bi-directional QKD, or "Plug & play" QKD is the QKD scheme in which Bob generates strong laser pulses and sends them to Alice. Alice encodes her quantum information on the pulse and attenuates the pulse to single-photon level, and sends it back to Bob through the same channel. This design can automatically compensate the phase and the polarization drifting in the channel.

Single photon source Single photon source is the light source that can generate a *single* photon on demand. A perfect single photon source should have zero probability to generate multi photons once triggered. Single photon source is required in the original BB84 protocol. However, it is no longer under absolute demand due to the discovery and implementation of decoy state QKD.

Fainted laser source Fainted laser source is the light source that has a standard pulsed laser source and a heavy attenuator. The average output photon number is usually set to ~ 0.1 photon per pulse. This low average photon number is to suppress the production of multi photon signals. However, due to the poisson nature of laser source, the probability of multi photon production can never reach zero unless the laser is turned off.

Single photon detector Single photon detector is sensitive to the weakest light signals – signals with single photons. Most single photon detectors are threshold by means that they can only detect the arrival of one or more photons, but cannot count the number of photons within one signal.

Dark count Dark count is the event that the detector reports a detection while no photon actually hits it. It is a key parameter for single photon detectors. Dark count becomes important when the channel loss between the sender and the receiver is high (i. e., when very few photons can reach the receiver).

Qubit Qubit (or quantum bit) is the fundamental unit of quantum information. Whereas a classical bit can take value of either "0" or "1", a qubit can take a value in any superposition of two distinguishable (i. e., orthogonal) states commonly labeled by $|0\rangle$ and $|1\rangle$. In other words, a (pure) qubit state can be written in the form $a|0\rangle + b|1\rangle$ where a and b are complex numbers. The normalization constraint is that $|a|^2 + |b|^2 = 1$. Physically, a qubit can be encoded in any two-level quantum system, such as the two polarization of a single photon or two atomic levels of an atom.

Bit-flip Bit-flip is a typical noise in both classical and quantum communication. In quantum cryptography, a bit-flip in the channel will transform an initial state $|i\rangle = a|0\rangle + b|1\rangle$ into the final state $|f\rangle = a|1\rangle + b|0\rangle$.

Phase-flip Phase-flip is a typical noise that is unique in quantum communication. A phase-flip in the quantum channel will transform an initial state $|i\rangle = a|0\rangle + b|1\rangle$ into the final state $|f\rangle = a|0\rangle - b|1\rangle$.

Definition of the Subject

Quantum cryptography is the synthesis of quantum mechanics with the art of code-making (cryptography). The idea was first conceived in an unpublished manuscript written by Stephen Wiesner around 1970 [4]. However, the subject received little attention until its resurrection by a classic paper published by Bennett and Brassard in 1984 [1]. The goal of quantum cryptography is to perform tasks that are impossible or intractable with conventional cryptography. Quantum cryptography makes use of the subtle properties of quantum mechanics such as the quantum no-cloning theorem and the Heisenberg uncertainty principle. Unlike conventional cryptography, whose security is often based on unproven computational assumptions, quantum cryptography has an important advantage in that its security is often based on the laws of physics. Thus far, proposed applications of quantum cryptography include quantum key distribution (abbreviated QKD), quantum bit commitment and quantum coin tossing. These applications have varying degrees of success. The most successful and important application – QKD – has been proven to be unconditionally secure. Moreover, experimental QKD has now been performed over hundreds of kilometers over both standard commercial telecom optical fibers and open-air. In fact, commercial QKD systems are currently available on the market.

On a wider context, quantum cryptography is a branch of quantum information processing, which includes quantum computing, quantum measurements, and quantum teleportation. Among all branches, quantum cryptography is the branch that is closest to real-life applications. Therefore, it can be a concrete avenue for the demonstrations of concepts in quantum information processing. On a more fundamental level, quantum cryptography is deeply related to the foundations of quantum mechanics, particularly the testing of Bell-inequalities and the detection efficiency loophole. On a technological level, quantum cryptography is related to technologies such as single-photon measurements and detection and single-photon sources.

Introduction

The best-known application of quantum cryptography is quantum key distribution (QKD). The goal of QKD is to allow two distant participants, traditionally called Alice and Bob, to share a long random string of secret (commonly called the key) in the presence of an eavesdropper, traditionally called Eve. The key can subsequently be used to achieve a) perfectly secure communication (via one-time-pad, see below) and b) perfectly secure authentication (via Wigman–Carter authentication scheme), thus achieving two key goals in cryptography.

The best-known protocol for QKD is the Bennett and Brassard protocol (BB84) published in 1984 [1]. The procedure of BB84 is as follows (also shown in Table 1).

1. Quantum communication phase
   (a) In BB84, Alice sends Bob a sequence of photons, each independently chosen from one of the four polarizations – vertical, horizontal, 45-degrees and 135-degrees.
   (b) For each photon, Bob randomly chooses one of the two measurement bases (rectilinear and diagonal) to perform a measurement.
   (c) Bob records his measurement bases and results. Bob publicly acknowledges his receipt of signals.
2. Public discussion phase
   (a) Alice broadcasts her bases of measurements. Bob broadcasts his bases of measurements.

Quantum Cryptography, Table 1
Procedure of BB84 protocol

Alice's bit sequence	0	1	1	1	0	1	0	0	0	1
Alice's basis	+	×	+	+	×	+	×	×	+	×
Alice's photon polarization	↔	⤡	↕	↕	⤢	↕	⤢	⤢	↔	⤡
Bob's basis	+	+	×	+	+	×	×	+	+	×
Bob's measured polarization	↔	↕	⤡	↕	↔	⤢	⤢	↕	↔	⤡
Bob's sifted measured polarization	↔			↕			⤢		↔	⤡
Bob's data sequence	0			1			0		0	1

(b) Alice and Bob discard all events where they use different bases for a signal.
(c) To test for tampering, Alice randomly chooses a fraction, p, of all remaining events as test events. For those test events, she publicly broadcasts their positions and polarizations.
(d) Bob broadcasts the polarizations of the test events.
(e) Alice and Bob compute the error rate of the test events (i. e., the fraction of data for which their value disagree). If the computed error rate is larger than some prescribed threshold value, say 11%, they abort. Otherwise, they proceed to the next step.
(f) Alice and Bob each convert the polarization data of all remaining data into a binary string called a raw key (by, for example, mapping a vertical or 45-degrees photon to "0" and a horizontal or 135-degrees photon to "1"). They can perform classical post-processing such as error correction and privacy amplification to generate a final key.

Notice that it is important for the classical communication channel between Alice and Bob to be authenticated. Otherwise, Eve can easily launch a man-in-the-middle attack by disguising as Alice to Bob and as Bob to Alice. Fortunately, authentication of an m-bit classical message requires only logarithmic in m-bit of an authentication key. Therefore, QKD provides an efficient way to expand a short initial authentication key into a long key. By repeating QKD many times, one can get an arbitrarily long secure key.

This article is organized as follows. In Sect. "Quantum Key Distribution: Motivation and Introduction", we will discuss the importance and foundations of QKD; in Section "Security Proofs", we will discuss the principles of different approaches to prove the unconditional security of QKD; in Section "Experimental Fundamentals", we will introduce the history and some fundamental components of QKD implementations; in Section "Experimental Implementation of BB84 Protocol", we will discuss the implementation of BB84 protocol in detail; in Section "Other Quantum Key Distribution Protocols", we will discuss the proposals and implementations of other QKD protocols; in Section "Quantum Hacking", we will introduce a very fresh and exciting area – quantum hacking – in both theory and experiments. In particular, we provide a catalog of existing eavesdropping attacks; in Section "Beyond Quantum Key Distribution", we will discuss some topics other than QKD, including quantum bit commitment, quantum coin tossing, etc.; in Section "Future Directions", we will wrap up this article with prospectives of quantum cryptography in the future.

Quantum Key Distribution: Motivation and Introduction

Cryptography – the art of code-marking – has a long and distinguished history of military and diplomatic applications, dating back to ancient civilizations in Mesopotamia, Egypt, India and China. Moreover, in recent years cryptography has widespread applications in civilian applications such as electronics commerce and electronics businesses. Each time we go on-line to access our banking or credit card data, we should be deeply concerned with our data security.

Key Distribution Problem and One-Time-Pad

Secure communication is the best-known application of cryptography. The goal of secure communication is to allow two distant participants, traditionally called Alice and Bob, to communicate securely in the presence of an eavesdropper, Eve. See Fig. 1. A simple example of an encryption scheme is the Caesar's cipher. Alice simply shifts each letter in a message alphabetically by three letters. For instance, the word NOW is mapped to QRZ, because N → O → P → Q, O → P → Q → R and W → X → Y → Z. According to legends, Julius Caesar used Caesar's cipher to communicate with his generals. An encryption by an alphabetical shift of a fixed but arbitrary number of po-

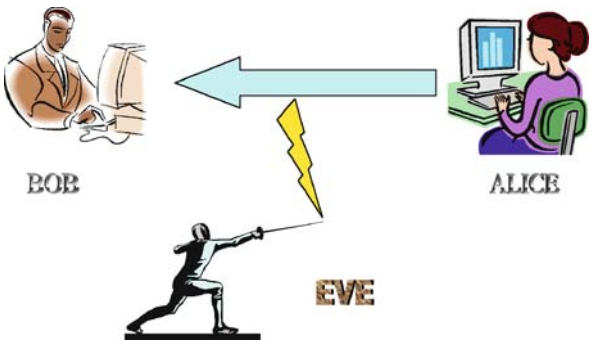

Quantum Cryptography, Figure 1
Communication in presence of an eavesdropper

sitions is also called a Caesar's cipher. Note that Caesar's cipher is not that secure because an eavesdropper can simply exhaustively try all 26 possible combinations of the key to recover the original message.

In conventional cryptography, an unbreakable code does exist. It is called the one-time-pad and was invented by Gilbert Vernam in 1918 [5]. In the one-time-pad method, a message (traditionally called the plain text) is first converted by Alice into a binary form (a string consisting of "0"s and "1"s) by a publicly known method. A key is a binary string of the same length as the message. By combining each bit of the message with the respective bit of the key using XOR (i. e. addition modulo two), Alice converts the plain text into an encrypted form (called the cipher text). i. e. for each bit $c_i = m_i + k_i \mod 2$. Alice then transmits the cipher text to Bob via a broadcast channel. Anyone including an eavesdropper can get a copy of the cipher text. However, without the knowledge of the key, the cipher text is totally random and gives no information whatsoever about the plain text. For decryption, Bob, who shares the same key with Alice, can perform another XOR (i. e. addition modulo two) between each bit of the cipher text with the respective bit of the key to recover the plain text. This is because $c_i + k_i \mod 2 = m_i + 2k_i \mod 2 = m_i \mod 2$.

Notice that it is important not to re-use a key in a one-time-pad scheme. Suppose the same key, k, is used for the encryption of two messages, m_1 and m_2, then the cipher texts are $c_1 = m_1 + k \mod 2$ and $c_2 = m_2 + k \mod 2$. Then, Eve can simply take the XOR of the two cipher texts to obtain $c_1 + c_2 \mod 2 = m_1 + m_2 + 2k \mod 2 = m_1 + m_2 \mod 2$, thus learning non-trivial information, namely the parity of the two messages.

The one-time-pad method is commonly used in top-secret communication. The one-time-pad method is unbreakable, but it has a serious drawback: it supposes that Alice and Bob initially share a random string of secret that is as long as the message. Therefore, the one-time-pad simply shifts the problem of secure communication to the problem of key distribution. This is the key distribution problem. In top-secret communication, the key distribution problem can be solved by trusted couriers. Unfortunately, trusted couriers can be bribed or compromised. Indeed, in conventional cryptography, a key is a classical string consisting of "0" and "1"s. In classical physics, there is no fundamental physical principle that can prevent an eavesdropper from copying a key during the key distribution process.

A possible solution to the key distribution problem is public key cryptography. However, the security of public key cryptography is based on unproven computational assumptions. For example, the security of standard RSA crypto-system invented by Rivest–Shamir–Adleman (RSA) is based on the presumed difficulty of factoring large integers. Therefore, public key distribution is vulnerable to unanticipated advances in hardware and algorithms. In fact, quantum computers – computers that operate on the principles of quantum mechanics – can break standard RSA crypto-system via the celebrated Shor's quantum algorithm for efficient factoring [6].

Quantum No-Cloning Theorem and Quantum Key Distribution (QKD)

Quantum mechanics can provide a solution to the key distribution problem. In quantum key distribution, an encryption key is generated randomly between Alice and Bob by using non-orthogonal quantum states. In contrast to classical physics, in quantum mechanics there is a quantum no-cloning theorem (see below), which states that it is fundamentally impossible for anyone including an eavesdropper to make an additional copy of an unknown quantum state.

A big advantage of quantum cryptography is *forward security*. In conventional cryptography, an eavesdropper Eve has a transcript of all communications. Therefore, she can simply save it for many years and wait for breakthroughs such as the discovery of a new algorithm or new hardware. Indeed, if Eve can factor large integers in 2100, she can decrypt communications sent in 2008. We remark that Canadian census data are kept secret for 92 years on average. Therefore, factoring in the year 2100 may violate the security requirement of our government today! And, no one in his/her sane mind can guarantee the impossibility of efficient factoring in 2100 (except for the fact that he/she may not live that long). In contrast, quantum cryptography guarantees forward security. Thanks to the quan-

tum no-cloning theorem, an eavesdropper does *not* have a transcript of all quantum signals sent by Alice to Bob.

For completeness, we include the statement and the proof of the quantum no-cloning theorem below.

Quantum No-Cloning Theorem *An unknown quantum state cannot be copied.*

(a) The case without ancilla: Given an unknown state $|\alpha\rangle$, show that a quantum copying machine that can map $|\alpha\rangle|0\rangle \to |\alpha\rangle|\alpha\rangle$ does not exist.
(b) The general case: Given an unknown state $|\alpha\rangle$, show that a quantum copying machine that can map $|\alpha\rangle|0\rangle|0\rangle \to |\alpha\rangle|\alpha\rangle|u_\alpha\rangle$ does not exist.

Proof

(a) Suppose the contrary. Then, a quantum cloning machine exists. Consider two orthogonal input states $|0\rangle$ and $|1\rangle$ respectively. We have

$$|0\rangle|0\rangle \to |0\rangle|0\rangle$$

and

$$|1\rangle|0\rangle \to |1\rangle|1\rangle\,.$$

Consider a general input $|\alpha\rangle = a|0\rangle + b|1\rangle$. Since a unitary transformation is linear, by linearity, we have

$$\begin{aligned}|\alpha\rangle|0\rangle &= (a|0\rangle + b|1\rangle)|0\rangle \\ &\to a|0\rangle|0\rangle + b|1\rangle|1\rangle\,.\end{aligned} \tag{1}$$

In contrast, for quantum cloning, we need:

$$\begin{aligned}|\alpha\rangle|0\rangle &\to (a|0\rangle + b|1\rangle)(a|0\rangle + b|1\rangle) \\ &= a^2|0\rangle|0\rangle + ab|0\rangle|1\rangle + ab|1\rangle|0\rangle + b^2|1\rangle|1\rangle\,.\end{aligned} \tag{2}$$

Clearly, if $ab \neq 0$, the two results shown in Eqs. (1) and (2)) are different. Therefore, quantum cloning (without ancilla) is impossible.

(b) similar. □

More generally, for general quantum states, information gain implies disturbance.

Theorem (Information Gain implies disturbance) *Given one state chosen from one of the two distinct non-orthogonal states, $|u\rangle$ and $|v\rangle$ (i.e. $|\langle u|v\rangle| \neq 0$ or 1), any operation that can learn any information about its identity necessarily disturbs the state.*

Proof Given a system initially in state either $|u\rangle$ and $|v\rangle$. Suppose an experimentalist applies some operation on the system. The most general thing that she can try to do is to prepare some ancilla in some standard state $|0\rangle$ and couple it to the system. Therefore, we have:

$$|u\rangle|0\rangle \to |u\rangle|\phi_u\rangle \tag{3}$$

and

$$|v\rangle|0\rangle \to |v\rangle|\phi_v\rangle \tag{4}$$

for some states $|\phi_u\rangle$ and $|\phi_v\rangle$.

In the end, the experimentalist lets go of the system and keeps the ancilla. He/she may then perform a measurement on the ancilla to learn about the initial state of the system.

Recall that quantum evolution is unitary and as such it preserves the inner product. Now, taking the inner product between Eqs. (3) and (4), we get:

$$\begin{aligned}\langle u|v\rangle\langle 0|0\rangle &= \langle u|v\rangle\langle\phi_u|\phi_v\rangle \\ \langle u|v\rangle &= \langle u|v\rangle\langle\phi_u|\phi_v\rangle \\ \langle u|v\rangle(1 - \langle\phi_u|\phi_v\rangle) &= 0 \\ (1 - \langle\phi_u|\phi_v\rangle) &= 0 \\ |\phi_u\rangle &= |\phi_v\rangle\,,\end{aligned} \tag{5}$$

where in the fourth line, we have used the fact that $|\langle u|v\rangle| \neq 0$.

Now, the condition that $|\phi_u\rangle = |\phi_v\rangle$ means that the final state of the ancilla is independent of the initial state of the system. Therefore, a measurement on the ancilla will tell the experimentalist nothing about the initial state of the system. □

Therefore, any attempt by an eavesdropper to learn information about a key in a QKD process will lead to disturbance, which can be detected by Alice and Bob who can, for example, check the bit error rate of a random sample of the raw transmission data.

The standard BB84 protocol for QKD was discussed in Sect. "Introduction". In the BB84 protocol, Alice prepares a sequence of photons each randomly chosen in one of the four polarizations – vertical, horizontal, 45-degrees and 135-degrees. For each photon, Bob chooses one of the two polarization bases (rectilinear or diagonal) to perform a measurement. Intuitively, the security comes from the fact that the two polarization bases, rectilinear and diagonal, are conjugate observables. Just like position and momentum are conjugate observables in the standard Heisenberg uncertainty principle, no measurement by an eavesdropper Eve can determine the value of both observables simultaneously. In mathematics, two conjugate observables are represented by two non-commuting Hermitian matrices. Therefore, they cannot be simultaneously

diagonalized. This impossibility of simultaneous diagonalization implies the impossibility of simultaneous measurements of two conjugate observables.

Example of a Simple Eavesdropper Attack: Intercept-Resend Attack

To illustrate the security of quantum cryptography, let us consider the simple example of an intercept-resend attack by an eavesdropper Eve, who measures each photon in a randomly chosen basis and then resends the resulting state to Bob. For instance, if Eve performs a rectilinear measurement, photons prepared by Alice in the diagonal bases will be disturbed by Eve's measurement and give random answers. When Eve resends rectilinear photons to Bob, if Bob performs a diagonal measurement, then he will get random answers. Since the two bases are chosen randomly by each party, such an intercept-resend attack will give a bit error rate of $0.5 \times 0.5 + 0.5 \times 0 = 25\%$, which is readily detectable by Alice and Bob. Sophisticated attacks against QKD do exist. Fortunately, the security of QKD has now been proven. This subject will be discussed further in Sect. "Security Proofs".

Equivalence Between Phase and Polarization Encoding

Notice that the BB84 protocol can be implemented with any two-level quantum system (qubits). In Sect. "Introduction" and the above discussion, we have described the BB84 protocol in terms of polarization encoding. This is just one of the many possible types of encodings. Indeed, it should be noted that other encoding method, particularly, phase encoding also exists. In phase encoding, a signal consists of a superposition of two time-separated pulses, known as the reference pulse and the signal pulse. See Fig. 2 for an illustration of the phase encoding scheme. The information is encoded in the relative phase between two pulses. i. e., the four possible states used by Alice are $1/\sqrt{2}(|R\rangle + |S\rangle)$, $1/\sqrt{2}(|R\rangle - |S\rangle)$, $1/\sqrt{2}(|R\rangle + i|S\rangle)$, $1/\sqrt{2}(|R\rangle - i|S\rangle)$.

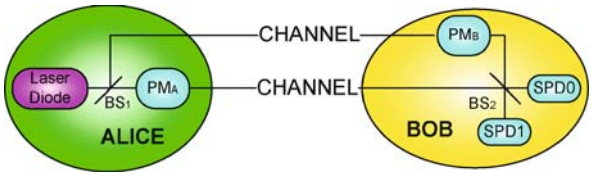

Quantum Cryptography, Figure 2
Conceptual schematic for phase-coding BB84 QKD system. PM: Phase Modulator; BS: Beam Splitter; SPD: Single Photon Detector

Notice that, mathematically the phase encoding scheme is equivalent to the polarization encoding scheme. They are simply different embodiments of the same BB84 protocol.

Security Proofs

"The most important question in quantum cryptography is to determine how secure it really is." (Brassard and Crépeau [7]) Security proofs are very important because a) they provide the foundation of security to a QKD protocol, b) they provide a formula for the key generation rate of a QKD protocol and c) they may even provide a construction for the classical post-processing protocol (for error correction and privacy amplification) that is necessary for the generation of the final key. Without security proofs, a real-life QKD system is incomplete because we can never be sure about how to generate a secure key and how secure the final key really is.

Classification of Eavesdropping Attacks

Before we discuss security proofs, let us first consider eavesdropping attacks. Notice that there are infinitely many eavesdropping strategies that an eavesdropper, Eve, can perform against a QKD protocol. They can be classified as follows:

Individual attacks In an individual attack, Eve performs an attack on each signal independently. The intercept-resend attack discussed in Subsect. "Example of a Simple Eavesdropper Attack: Intercept-Resend Attack" is an example of an individual attack.

Collective attacks A more general class of attacks is collective attack where for each signal, Eve independently couples it with an ancillary quantum system, commonly called an ancilla, and evolves the combined signal/ancilla unitarily. She can send the resulting signals to Bob, but keep all ancillas herself. Unlike the case of individual attacks, Eve postpones her choice of measurement. Only after hearing the public discussion between Alice and Bob, does Eve decide on what measurement to perform on her ancilla to extract information about the final key.

Joint attacks The most general class of attacks is joint attack. In a joint attack, instead of interacting with each signal independently, Eve treats all the signals as a single quantum system. She then couples the signal system with her ancilla and evolves the combined signal and ancilla system unitarily. She hears the public discussion between Alice and Bob before deciding on which measurement to perform on her ancilla.

Proving the security of QKD against the most general attack was a very hard problem. It took more than 10 years, but the unconditional security of QKD was finally established in several papers in the 1990s. One approach by Mayers [8] was to prove the security of the BB84 directly. A simpler approach by Lo and Chau [9], made use of the idea of entanglement distillation by Bennett, DiVincenzo, Smolin and Wootters (BDSW) [10] and quantum privacy amplification by Deutsch et al. [11] to solve the security of an entanglement-based QKD protocol. The two approaches have been unified by the work of Shor and Preskill [12], who provided a simple proof of security of BB84 using entanglement distillation idea. Other early security proofs of QKD include Biham, Boyer, Boykin, Mor, and Roychowdhury [13], and Ben-Or [14].

Approaches to Security Proofs

There are several approaches to security proof. We will discuss them one by one.

1. Entanglement distillation
   Entanglement distillation protocol (EDP) provides a simple approach to security proof [9,11,12]. The basic insight is that entanglement is a sufficient (but not necessary) condition for a secure key. Consider the noiseless case first. Suppose two distant parties, Alice and Bob, share a maximally entangled state of the form $|\phi\rangle_{\mathrm{AB}} = 1/\sqrt{2}(|00\rangle_{\mathrm{AB}} + |11\rangle_{\mathrm{AB}})$. If each of Alice and Bob measure their systems, then they will both get "0"s or "1"s, which is a shared random key. Moreover, if we consider the combined system of the three parties – Alice, Bob and an eavesdropper, Eve, we can use a pure-state description (the "Church of Larger Hilbert space") and consider a pure state $|\psi\rangle_{\mathrm{ABE}}$. In this case, the von Neumann entropy [15] of Eve $S(\rho_{\mathrm{E}}) = S(\rho_{\mathrm{AB}}) = 0$. This means that Eve has absolutely no information on the final key. This is the consequence of the standard Holevo's theorem. See, like, [16].
   In the noisy case, Alice and Bob may share N pairs of qubits, which are a noisy version of N maximally entangled states. Now, using the idea of entanglement distillation protocol (EDP) discussed in BDSW [10], Alice and Bob may apply local operations and classical communications (LOCCs) to distill from the N noisy pairs a smaller number, say M almost perfect pairs i. e., a state close to $|\phi\rangle_{\mathrm{AB}}^{M}$. Once such a EDP has been performed, Alice and Bob can measure their respective system to generate an M-bit final key.
   One may ask: how can Alice and Bob be sure that their EDP will be successful? Whether an EDP will be successful or not depends on the initial state shared by Alice and Bob. In the above, we have skipped the discussion about the verification step. In practice, Alice and Bob can never be sure what initial state they possess. Therefore, it is useful for them to add a verification step. By, for example, randomly testing a fraction of their pairs, they have a pretty good idea about the properties (e. g., the bit-flip and phase error rates) of their remaining pairs and are pretty confident that their EDP will be successful.
   The above description of EDP is for a quantum-computing protocol where we assume that Alice and Bob can perform local quantum computations. In practice, Alice and Bob do not have large-scale quantum computers at their disposal. Shor and Preskill made the important observation that the security proof of the standard BB84 protocol can be reduced to that of an EDP-based QKD protocol [9,11]. The Shor–Preskill proof [12] makes use of the Calderbank–Shor–Steane (CSS) code, which has the advantage of decoupling the quantum error correction procedure into two parts: bit-flip and phase error correction. They can go on to show that bit-flip error correction corresponds to standard error correction and phase error correction corresponds to privacy amplification (by random hashing).
2. Communication complexity/quantum memory
   The communication complexity/quantum memory approach to security proof was proposed by Ben-Or [14] and subsequently by Renner and Koenig [17]. See also [18]. They provide a formula for secure key generation rate in terms of an eavesdropper's quantum knowledge on the raw key: Let Z be a random variable with range $\mathcal{Z}$, let $\boldsymbol{\rho}$ be a random state, and let F be a two-universal function on $\mathcal{Z}$ with range $S = \{0,1\}^s$ which is independent of Z and $\boldsymbol{\rho}$. Then [17]
   $$d(F(Z)|\{F\} \otimes \boldsymbol{\rho}) \leq \tfrac{1}{2} 2^{-\frac{1}{2}(S_2([\{Z\}\otimes\boldsymbol{\rho}]) - S_0([\boldsymbol{\rho}]) - s)} .$$
   Incidentally, the quantum de Finnetti's theorem [19] is often useful for simplifying security proofs of this type.
3. Twisted state approach
   What is a necessary and sufficient condition for secure key generation? From the entanglement distillation approach, we know that entanglement distillation a sufficient condition for secure key generation. For some time, it was hoped that entanglement distillation is also a necessary condition for secure key generation. However, such an idea was proven to be wrong in [20,21], where it was found that a necessary and sufficient condition is the distillation of a private state, rather than a maximally entangled state. A private state is a "twisted" version of a maximally entangled state.

They proved the following theorem in [20]: a state is private in the above sense iff it is of the following form

$$\gamma_m = U|\psi_{2^m}^+\rangle_{\mathrm{AB}}\langle\psi_{2^m}^+| \otimes \varrho_{\mathrm{A'B'}} U^\dagger \tag{6}$$

where $|\psi_d\rangle = \sum_{i=1}^d |ii\rangle$ and $\varrho_{\mathrm{A'B'}}$ is an arbitrary state on A', B'. U is an arbitrary unitary controlled in the computational basis

$$U = \sum_{i,j=1}^{2^m} |ij\rangle_{\mathrm{AB}}\langle ij| \otimes U_{ij}^{A'B'} \,. \tag{7}$$

The operation (7) will be called "twisting" (note that only $U_{ii}^{A'B'}$ matter here, yet it will be useful to consider general twisting later).

Proof. (copied from [20]) The authors of [20] proved for $m = 1$ (for higher m, the proof is analogous). Start with an arbitrary state held by Alice and Bob, $\rho_{\mathrm{AA'BB'}}$, and include its purification to write the total state in the decomposition

$$\begin{aligned}\Psi_{\mathrm{ABA'B',E}} = a|00\rangle_{\mathrm{AB}}|\Psi_{00}\rangle_{\mathrm{A'B'E}} + b|01\rangle_{\mathrm{AB}}|\Psi_{01}\rangle_{\mathrm{A'B'E}} \\ + c|10\rangle_{\mathrm{AB}}|\Psi_{10}\rangle_{\mathrm{A'B'E}} + d|11\rangle_{\mathrm{AB}}|\Psi_{11}\rangle_{\mathrm{A'B'E}}\end{aligned} \tag{8}$$

with the states $|ij\rangle$ on AB and Ψ_{ij} on $A'B'E$. Since the key is unbiased and perfectly correlated, we must have $b = c = 0$ and $|a|^2 = |d|^2 = 1/2$. Depending on whether the key is $|00\rangle$ or $|11\rangle$, Eve will hold the states

$$\varrho_0 = Tr_{\mathrm{A'B'}}|\Psi_{00}\rangle\langle\Psi_{00}|, \quad \varrho_1 = Tr_{\mathrm{A'B'}}|\Psi_{11}\rangle\langle\Psi_{11}|. \tag{9}$$

Perfect security requires $\varrho_0 = \varrho_1$. Thus there exists unitaries U_{00} and U_{11} on $A'B'$ such that

$$\begin{aligned}|\Psi_{00}\rangle &= \sum_i \sqrt{p_i}|U_0\phi_i^{A'B'}\rangle|\varphi_i^E\rangle \\ |\Psi_{11}\rangle &= \sum_i \sqrt{p_i}|U_1\phi_i^{A'B'}\rangle|\varphi_i^E\rangle \,.\end{aligned} \tag{10}$$

After tracing out E, we will thus get a state of the form Eq. (6), where $\varrho_{\mathrm{A'B'}} = \sum_i p_i|\phi_i\rangle\langle\phi_i|\square$.

The main new ingredient of the above theorem is the introduction of a "shield" part to Alice and Bob's system. That is, in addition to the systems A and B used by Alice and Bob for key generation, we assume that Alice and Bob also hold some ancillary systems, A' and B', often called the shield part. Since we assume that Eve has no access to the shield part, Eve is further limited in her ability to eavesdrop. Therefore, Alice and Bob can derive a higher key generation rate than the case when Eve does have access to the shield part.

An upshot is that even a bound entangled state can give a secure key. A bound state is one whose formation (via local operations and classical communications, LOCCs) requires entanglement, but which does not give any distillable entanglement. In other words, even though no entanglement can be distilled from a bound entangled state, private states (a twisted version of entangled states) *can* be distilled from a bound entangled state.

In summary, secure key generation is a more general theory than entanglement distillation.

4. Complementary principle

Another approach to security proof is to use the complementary principle of quantum mechanics. Such an approach is interesting because it shows the deep connection between the foundations of quantum mechanics and the security of QKD. In fact, both Mayers' proof [8] and Biham, Boyer, Boykin, Mor, and Roychowdhury's proof [13] make use of this complementary principle. A clear and rigorous discussion of the complementary principle approach to security proof has recently been achieved by Koashi [22].

The key insight of Koashi's proof is that Alice and Bob's ability to generate a random secure key in the Z-basis (by a measurement of the Pauli spin matrix σ_Z) is equivalent to the ability for Bob to help Alice prepare an eigenstate in the complementary, i. e., X-basis (σ_X), with their help of the shield. The intuition is that an X-basis eigenstate, for example, $|+\rangle_{\mathrm{A}} = 1/\sqrt{2}(|0\rangle_{\mathrm{A}} + |1\rangle_{\mathrm{A}})$, when measured along the Z-basis, gives a random answer.

5. Other ideas for security proofs

Here we discuss two other ideas for security proofs, namely, a) device-independent security proofs and b) security from the causality constraint. Unfortunately, these ideas are still very much under development and so far a complete version of a proof of unconditional security of QKD based on these ideas with a finite key rate is still missing.

Let us start with a) device-independent security proofs So far we have assumed that Alice and Bob know what their devices are doing exactly. In practice, Alice and Bob may not know their devices for sure. Recently, there has been much interest in the idea of device-independent security proofs. In other words, how to prove security when Alice and Bob's devices cannot be trusted. See, for example, [23]. The idea is to look only at the input and output variables. A handwaving argument goes as follows. Using their probability distribution, if one can demonstrate the violation of some Bell inequalities, then one cannot explain the data by a separable system. How to develop such a handwaving argument into a full proof of unconditional security is an important question.

The second idea b) security from the causality constraint is even more ambitious. The question that it tries to address is the following. How can one prove security when even quantum mechanics is wrong? In [24] and references cited therein, it was suggested that perhaps a more general physical principle such as the no-signaling requirement for space-like observables could be used to prove the security of QKD.

Classical Post-processing Protocols

As noted in Sect. "Introduction", after the quantum communication phase, Alice and Bob then proceed with the classical communication phase. In order to generate a secure key, Alice and Bob have to know what classical post-processing protocol to apply to the raw quantum data. This is a highly non-trivial question. Indeed, a priori, given a particular procedure for classical post-processing, it is very hard to know whether it will give a secure key or what secure key will be generated. In fact, it is sometimes said that in QKD, the optical part is easy, the electronics part is harder, but the hardest part is the classical post-processing protocol. Fortunately, security proofs often give Alice and Bob clear ideas on what classical post-processing protocol to use. This highlights the importance for QKD practitioners to study the security proofs of QKD.

Briefly stated, the classical post-processing protocol often consists of a) test for tampering and b) key generation. In a) test for tampering, Alice and Bob may randomly choose a fraction of the signals for testing. For example, by broadcasting the polarizations of those signals, they can work out the bit error rate of the test signals. Since the test signals are randomly chosen, they have high confidence on the bit error rate of the remaining signals. If the bit error rate of the tested signal is higher than a prescribed threshold value, Alice and Bob abort. On the other hand, if the bit error rate is lower than or equal to the prescribed value, they proceed with the key generation step with the remaining signals. They first convert their polarization data into binary strings, the raw keys, in a prescribed manner. For example, they can map a vertical or 45-degrees photon to "0" and a horizontal or 135-degrees photon to "1". As a result, Alice has a binary string x and Bob has a binary string y. However, two problems remain. First, Alice's string may differ from Bob's string. Second, since the bit error rate is non-zero, Eve has some information about Alice's and Bob's string. The key generation step may be divided into the following stages:

1. Classical pre-processing
   This is an optional step. Classical pre-processing has the advantage of achieving a higher key generation rate and tolerating a higher bit error rate [25,26,27].
   Alice and Bob may pre-process their data by either a) some type of error detection algorithm or b) some random process. An example of an error detection algorithm is a B-step [25], where Alice randomly permutes all her bits and broadcasts the parity of each adjacent pair. In other words, starting from $\vec{x} = (x_1, x_2, \ldots, x_{2N-1}, x_{2N})$, Alice broadcasts a string $\vec{x}_1 = (x_{\sigma(1)} + x_{\sigma(2)} \mod 2, x_{\sigma(3)} + x_{\sigma(4)} \mod 2, \ldots, x_{\sigma(2N-1)} + x_{\sigma(2N)} \mod 2)$, where σ is a random permutation chosen by Alice. Moreover, Alice informs Bob which random permutation, σ, she has chosen. Similarly, starting from $\vec{y} = (y_1, y_2, \ldots, y_{2N-1}, y_{2N})$, Bob randomly permutes all his bits using the same σ and broadcasts the parity bit of all adjacent pairs. I. e. Bob broadcasts $\vec{y}_1 = (y_{\sigma(1)} + y_{\sigma(2)} \mod 2, y_{\sigma(3)} + y_{\sigma(4)} \mod 2, \ldots, y_{\sigma(2N-1)} + y_{\sigma(2N)} \mod 2)$. For each pair of bits, Alice and Bob keep the first bit iff their parities of the pair agree. For instance, if $x_{\sigma(2k-1)} + x_{\sigma(2k)} \mod 2 = y_{\sigma(2k-1)} + y_{\sigma(2k)} \mod 2$, then Alice keeps $x_{\sigma(2k-1)}$ and Bob keeps $y_{\sigma(2k-1)}$ as their new key bit. Otherwise, they drop the pair $(x_{\sigma(2k-1)}, x_{\sigma(2k)})$ and $(y_{\sigma(2k-1)} x_{\sigma(2k)})$ completely. Notice that the above protocol is an error detection protocol. To see this, let us regard the case where $x_i \neq y_i$ as an error during the quantum transmission stage. Suppose that for each bit, i, the event $x_i \neq y_i$ occurs with an independent probability p. For each k, the B-step throws away the cases where a single error has occurred for the two locations $\sigma(2k-1)$ and $\sigma(2k)$ and keeps the cases when either no error or two errors has occurred. As a result, the error probability after the B-step is reduced from $O(p)$ to $O(p^2)$. The random permutation of all the bit locations ensures that the error model can be well described by an independent identical distribution (i.i.d.).
   An example of a random process is an adding noise protocol [26] where, for each bit x_i, Alice randomly and independently chooses to keep it unchanged or flip it with probabilities, $1-q$ and q respectively, where the probability q is publicly known.
2. Error correction
   Owing to noises in the quantum channel, Alice and Bob's raw keys, x and y, may be different. Therefore, it is necessary for them to reconcile their keys. One simple way of key reconciliation is forward key reconciliation, whose goal is for Alice to keep the same key x and Bob to change his key from y to x. Forward key reconciliation can be done by either standard error correcting codes such as low-density-parity-check (LDPC) codes

or specialized (one-way or interactive) protocols such as Cascade protocol [28].

3. Privacy amplification
   To remove any residual information Eve may have about the key, Alice and Bob may apply some algorithm to compress their partially secure key into a shorter one that is almost perfectly secure. This is called privacy amplification. Random hashing and a class of two-universal hash functions are often suitable for privacy amplification. See for example [29] and [18] for discussion.

Composability

A key generated in QKD is seldom used in isolation. Indeed, one may concatenate a QKD process many times, using a small part of the key for authentication each time and the remaining key for other purposes such as encryption. It is important to show that using QKD as a sub-routine in a complicated cryptographic process does not create new security problems. This issue is called the composability of QKD and, fortunately, has been solved in [30].

Composability of QKD is not only of academic interest. It allows us to refine our definition of security [30,31] and directly impacts on the parameters used in the classical post-processing protocol.

Security Proofs of Practical QKD Systems

As will be discussed in Sect. "Experimental Fundamentals", practical QKD systems suffer from real-life imperfections. Proving the security of QKD with practical systems is a hard problem. Fortunately, this has been done with semi-realistic models by Inamori, Lütkenhaus and Mayers [32] and in a more general setting by Gottesman, Lo, Lütkenhaus, and Preskill [33].

Experimental Fundamentals

Quantum cryptography can ensure the secure communication between two or more legitimate parties. It is more than a beautiful idea. Conceptually, it is of great importance in the understandings of both information and quantum mechanics. Practically, it can provide an ultimate solution for confidential communications, thus making everyone's life easier.

By implementing the quantum crypto-system in the real life, we can test it, analyze it, understand it, verify it, and even try to break it. Experimental QKD has been performed since about 1989 and great progress has been made. Now, you can even buy QKD systems on the market.

A typical QKD set-up includes three standard parts: a source (Alice), a channel, and a detection system (Bob).

A Brief History

The First Experiment The proposal of BB84 [1] protocol seemed to be simple. However, it took another five years before it was first experimentally demonstrated by Bennett, Bessette, Brassard, Salvail, and Smolin in 1989 [34]. This first demonstration was based on polarization coding. Heavily attenuated laser pulses instead of single photons were used as quantum signals, which were transmitted over 30 cm open air at a repetition rate of 10 Hz.

From Centimeter to Kilometer 30 cm is not that appealing for practical communications. This short distance is largely due to the difficulty of optical alignment in free space. Switching the channel from open air to optical fiber is a natural choice. In 1993, Townsend, Rarity, and Tapster demonstrated the feasibility of phase-coding fiber-based QKD over 10 km telecom fiber [35] and Muller, Breguet, and Gisin demonstrated the feasibility of polarization-coding fiber-based QKD over 1.1 km telecom fiber [36]. (Also, Jacobs and Franson demonstrated both free-space [37] and fiber-based QKD [38].) These are both feasibility demonstrations by means that neither of them applied random basis choosing at Bob's side. Townsend's demonstration seemed to be more promising than Muller's due to the following reasons.

1. The polarization dispersion in fibers is highly unpredictable and unstable. Therefore polarization coding requires much more controlling in the fiber than phase coding. In fiber-based QKD implementations, phase coding is in general more preferred than polarization coding.
2. Townsend et al. used 1310 ns laser as the source, while Muller et al. used 800 ns laser as the source. 1310 nm is the second window wavelength of telecom fibers (the first window wavelength is 1550 nm). The absorption coefficient of standard telecom fiber at 1310 nm is 0.35 dB/km, comparing to 3 dB/km at 800 nm. Therefore the fiber is more transparent to Townsend et al.'s set-up.

P. D. Townsend demonstrated QKD with Bob's random basis selection in 1994 [39]. It was phase-coding and was over 10 km fiber. The source repetition rate was 105 MHz (which is quite high even by today's standard) but the phase modulation rate was 1.05 MHz. This mismatch brought a question mark on its security.

Getting Out of the Lab It is crucial to test QKD technique in the field deployed fiber. Muller, Zbinden, and Gisin successfully demonstrated the first QKD experiment outside the labs with polarization coding in 1995 [40,41]. This demonstration was performed over 23 km installed optical fiber under Lake Geneva. (Being under water, quantum communication in the optical fiber suffered less noise.)

There is less control over the field deployed fiber than fiber in the labs. Therefore its stabilization becomes challenging. To solve this problem, A. Muller et al. designed the "plug & play" structure in 1997 [42]. A first experiment of this scheme was demonstrated by H. Zbinden et al. in the same year [43]. Stucki, Gisin, Guinnard, Robordy, and Zbinden later demonstrated a simplified version of the "plug & play" scheme under Lake Geneva over 67 km telecom fiber in 2002 [44].

With a Coherent Laser Source The original BB84 [1] proposal required a single photon source. However, most QKD implementations are based on faint lasers due to the great challenge to build the perfect single photon sources. In 2000, the security of coherent laser based QKD systems was analyzed first against individual attacks [45]. Finally, the unconditional security of coherent laser based QKD systems was proven in 2001 [32] and in a more general setting in 2002 [33]. Gobby, Yuan, and Shields demonstrated an experiment based on [45] in 2005 [46] (Note that this work was claimed to be unconditionally secure. However, due to the limit of [45], this is only true against individual attacks rather than the most general attack).

The security analysis in [32,33] will severely limit the performance of unconditionally secure QKD systems. Fortunately, since 2003 the decoy state method has been proposed [47,48,49,50,51,52] by Hwang and extensively analyzed by our group at the University of Toronto and by Wang. The first experimental demonstration of decoy state QKD was reported by us in 2006 [53] over 15 km telecom fiber and later over 60 km telecom fiber [54]. Subsequently, decoy state QKD was further demonstrated by several other groups [55,56,57,58,59]. The readers may refer to Subsect. "Decoy State Protocols" for details of decoy state protocols.

Sources

Single Photon Sources are demanded by the original BB84 [1] proposal. Suggested by its name, the single photon sources are expected to generate exactly one photon on demand. The bottom line for a single photon source is that no more than one photon can be generated at one time. It is very hard to build a perfect single photon source (i. e., no multi-photon production). Despite tremendous effort made by many groups, perfect single photon source is still far from practical. Fortunately, the proposal and implementation of decoy state QKD (see Subsect. "Decoy State Protocols") make it unnecessary to use single photon sources in QKD.

Parametric Down-Conversion (PDC) Sources are often used as the entanglement source. Its principle is that a high energy (~ 400 nm) photon propagates through a highly non-linear crystal (usually BBO), producing two entangled photons with frequency halved. PDC sources are usually used for entanglement-based QKD systems (e. g. Ekert [3] protocol).

PDC sources are also used as "triggered single photon sources", in which Alice possesses a PDC source and monitor one arm of its outputs. In case that Alice sees a detection, she knows that there is one photon emitted from the other arm. Experimental demonstration of QKD with PDC sources is reported in [60].

Attenuated Laser Sources are the most commonly used sources in QKD experiments. They are essentially the same as the laser sources used in classical optical communication except for that heavy attenuation is applied on them. They are simple and reliable, and they can reach Gigahertz with little challenge. In BB84 system and differential-phase-shift-keying (DPSK) system (to be discussed in Subsect. "Differential-Phase-Shift-Keying (DPSK) Protocols"), the laser source is usually attenuated to below 1 photon per pulse. In Gaussian-modulated coherent-state (GMCS) system (to be discussed in Subsect. "Gaussian-Modulated Coherent State (GMCS) Protocol"), the laser source is usually attenuated to around 100 photons per pulse.

Attenuated laser sources used to be considered to be non-ideal for BB84 systems as they always have non-negligible probability of emitting multi-photon pulses regardless how heavily they are attenuated. However, the discovery and implementation of decoy method [47,48,49,50,51,52,53,54,55,56,57,58] made coherent laser source much more appealing. With decoy method, it is possible to make BB84 system with laser source secure without significant losses on the performance.

Channels

Standard Optical Single-Mode Fiber (SMF) is the most popular choice for now. It can connect two arbitrary points, and can easily be extended to networks.

Moreover, it is deployed in most developed urban areas.

SMF has two "window wavelengths": one is 1310 nm and the other is 1550 nm. The absorptions at these two wavelengths are particularly low ($\sim$ 0.35 dB/km at 1310 nm, and $\sim$ 0.21 dB/km at 1550 nm). Nowadays most fiber-based QKD implementations use 1550 nm photons as information carriers.

The main disadvantage of optical fiber is its birefringence. The strong polarization dispersion made it hard to implement polarization-coding system. Also it has strong spectral dispersion, which affects the high speed (10+ GHz) QKD systems heavily [61] as the pulses are broadened and overlap with each other. For this reason, the loss in fibers (0.21 dB/km at 1550 nm) puts an limit on the longest distance that a fiber-based QKD system can reach.

Free Space is receiving more and more attention recently. It is ideal for the polarization coding. There is negligible dispersion on the polarization and the frequency. However, the alignment of optical beams can be challenging for long distances, particularly due to the atmospheric turbulence. Notice that open-air QKD requires a direct line of sight between Alice and Bob (unless some forms of mirrors are used). Buildings and mountains are serious obstacles for open-air QKD systems.

The greatest motivation for open-air QKD scheme is the hope for ground-to-satellite [62] and satellite-to-satellite quantum communication. As there is negligible optical absorption in the outer space, we may be able to achieve inter-continental quantum communication with free-space QKD.

Detection Systems

InGaAs-APD Single Photon Detectors are the most popular type of single-photon detectors in fiber-based QKD and they are commercially available. InGaAs-APD Single Photon Detectors utilize the avalanche effect of semiconductor diodes. A strong biased voltage is applied on the InGaAs diode. The incident photon will trigger the avalanche effect, generating a detectable voltage pulse. The narrow band gap of InGaAs made it possible to detect photons at telecom wavelengths (1550 nm or 1310 nm).

InGaAs APD based single photon detectors have simple structure and commercially packaged. They are easy to calibrate and operate. The reliability of InGaAs APD is relatively high. They normally work at $-50°C$ to $-110°C$ to lower the dark count rate. This temperature can be easily achieved by thermal-electric coolers. The detection efficiency of InGaAs-APD based single photon detectors is usually $\sim$ 10% [63].

In single photon detectors, a key parameter (besides detection efficiency) is the dark count rate. The dark count is the event that the detector generates a detection click while no actual photon hits it (i. e. "false alarm"). The dark count rate of InGaAs single photon detector is relatively high (10^{-5} per gate; The concept of gating will be introduced below.) even if it is cooled. The after-pulse effect is that the dark count rate of the detector increases for a time period after a successful detection. This effect is serious for InGaAs single photon detectors. Therefore the blank circuit is often introduced to reduce this effect. The mechanism of the blank circuit is that the detector is set to be deactivated for a time period, which is called the "dead time", after a detection event. The dead time should be set to long enough so that when the detector is re-activated, the after-pulse effect is negligible. The dead time for InGaAs single photon detector is typically in the order of microseconds [63]. The long after pulse effect, together with the large timing jitter limits the InGaAs-APD based single photon detectors to work no faster than several megahertz. Moreover, the blank circuit reduces the detection efficiency of InGaAs-APD based single photon detectors.

An additional method to reduce the dark count rate is to apply the gating mode, i. e., the detectors are only activated when the photons are expected to hit them. Gating mode reduces the dark count rate by several orders and is thus used in most InGaAs-APD single photon detectors. However, it may open up a security loophole [64,65].

There is a trade-off between the detection efficiency and the dark count rate. As the biased voltage on an APD increases, both the detection efficiency and the dark count rate increase.

Recently, it has been reported that, by gating an InGaAs detector in a sinusoidal manner, it is possible to reduce the dead time and operate a QKD system at 500 MHz, see [66]. This result seems to be an important development which could make InGaAs detectors competitive with newer single photon detector technologies such as SSPDs (to be introduced below).

Si-APD Single Photon Detectors are ideal for detection of visible photons (say 800 nm). They have negligible dark count rate and can work at room temperature. They are very compact in size. More importantly, they have high detection efficiency ($>$ 60%) and can work at gigahertz. These detectors are ideal for free space QKD

systems. However, the band gap of silicon is too large to detect photons at telecom wavelength (1550 nm or 1310 nm), and the strong attenuation of telecom fiber on visible wavelengths makes it impractical to use visible photons in long distance fiber-based QKD systems.

Parametric Up-Conversion Single Photon Detectors try to use Si-APD to detect telecom wavelength photons. It uses periodically poled lithium niobate (PPLN) waveguide and a pumping light to up-convert the incoming telecom frequency photons into visible frequency, and uses Si-APD to detect these visible photons. The high speed and low timing-jitter of Si-APDs make it possible to perform GHz QKD on fiber-based system with up-conversion single photon detectors [67,68].

The efficiency of up-conversion detectors is similar to that of InGaAs APD single photon detectors. There is also a trade-off between the detection efficiency and the dark count rate. When increasing the power of the pumping light, the conversion efficiency will increase, improving the detection efficiency. Meanwhile, more pumping photons and up-converted pumping photons (with frequency doubled) will pass through the filter and enter the Si-APD, thus increasing the dark count rate [68].

Transiting-Edge Sensor (TES) is based on critical state superconductor rather than semiconductor APDs. It uses squared superconductor (typically tungsten) thin film as "calorimeter" to measure the electron temperature. A biased voltage is applied on the thin film to keep it in critical state. Once one or more photons are absorbed by the sensor, the electron temperature will change, leading to a change of the current. This current change can be detected by a superconductive quantum-interference device (SQUID) array [69].

The TES single photon detectors can achieve very high detection efficiency (up to 89%) at telecom wavelength. The dark count rate is negligible. Moreover, TES detectors can resolve photon numbers. This is because the electron temperature change is proportional to the number of photons that have been absorbed.

The thermal nature of TES detectors limits their counting rates. Once some photons were absorbed by the sensor, it would take a few microseconds before the heat is dissipated to the substrate. This long relaxation time limits the counting rate of TES detector to no more than a few megahertz [69]. This is a major drawback of TES.

The bandwidth of TES detector is extremely wide. The detector is sensitive to all the wavelengths. Even the black body radiation from the fiber or the environment can trigger the detection event, thus increasing the dark count rate. To reduce the dark counts caused by other wavelengths, a spectral filter is necessary. However, this will increase the internal loss and thus reduce the detection efficiency.

One of the greatest disadvantage of TES detector is its working temperature: 100 mK. This temperature probably requires complicated cooling devices [69].

Superconductive Single Photon Detectors (SSPDs) also use superconductor thin film to detect incoming photons. However, instead of using a piece of plain thin film, a pattern of zigzag superconductor (typically NbN) wire is formed. The superconductive wire is set to critical state by applying critical current through it. Once a photon hits the wire, it heats a spot on the wire and makes the spot over-critical (i. e., non-superconductive). As the current is the same as before, the current density in the areas around this hot-spot increases, thus making these areas non-superconductive. As a result, a section of the wire becomes non-superconductive, and a voltage spike can be observed as the current is kept constant [61].

The SSPD can achieve very high (up to 10 GHz) counting rate. This is because the superconductor wire used in SSPD can dissipate the heat in tens of picoseconds. It also has very low dark count rate (around 10 Hz) due to the superconductive nature. The SSPDs should be able to resolve the incident photon number in principle. However, photon number resolving SSPDs have never been reported yet.

The efficiency of SSPD is lower than that of TES. This is because only part ($\sim$ 50%) of the sensing area is covered by the wire. The fabrication of such complicated zigzag superconductor wire with smooth edge is also very challenging.

The working temperature of SSPD ($\sim$3K) is significantly higher than that of TES. SSPDs can work in closed-cycle refrigerator. Moreover, this relatively high working temperature significantly reduces the relaxation time [61].

Homodyne Detectors are used to count the *photon number* of a very weak pulse ($\sim$ 100 photons). The principle is to use a very strong pulse (often called the local oscillator) to interfere with the weak pulse. Then use two photo diodes to convert the two resulting optical pulses into electrical signals, and make a substraction between the two electrical signals.

The homodyne detectors are in general very efficient as there is always some detection output given some input signals. However, the noise of the detectors as well as in the electronics is very significant. Moreover,

the two photo diodes in the homodyne detector have to be identical, which is hard to meet in practice.

The homodyne detectors are commonly used in GMCS QKD systems, and they are so far the only choice for GMCS QKD systems. Recently, homodyne detectors have also been used to implement the BB84 protocol.

Truly Quantum Random Number Generators

An important but often under-appreciated requirement for QKD is a high data-rate truly quantum random number generator (RNG). An RNG is needed because most QKD protocols (with the exception of a passive choice of bases in an entanglement-based QKD protocol) require Alice to choose actively random bases/signals. Given the high repetition rate of QKD, such a RNG must have a high data-rate. To achieve unconditional security, a standard software-based pseudo-random number generator cannot be used because it is actually deterministic. So, a high data rate quantum RNG is a natural choice. Incidentally, some firms such as id Quantique do offer commercial quantum RNGs. Unfortunately, it is very hard to generate RNG by quantum means at high-speed. In practice, some imperfections/bias in the numbers generated by a quantum RNG are inevitable. The theoretical foundation of QKD is at risk because existing security proofs all assume the existence of perfect RNGs and do not apply to imperfect RNGs.

Experimental Implementation of BB84 Protocol

In this section, we will focus mainly on the optical layer. We will skip several important layers. In practice, the control/electronics layer is equally important. Moreover, it is extremely challenging to implement the classical post-processing layer in real-time, if one chooses block sizes of codes to be long enough to achieve unconditional security.

Polarization Coding

Polarization coding usually uses four laser sources generating the four polarization states of BB84 [1] protocol. A conceptual schematic is shown in Fig. 3. Note that due to polarization dispersion of a fiber, usually people need some compensating like the waveplates.

The polarization compensation should be implemented dynamically as the polarization dispersion in the fiber changes frequently. This is solved by introducing the electrical polarization controller in [58].

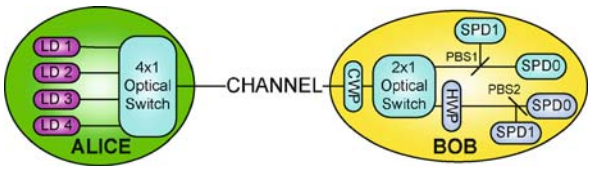

Quantum Cryptography, Figure 3
Conceptual schematic for polarization-coding BB84 QKD system. LD: Laser Diode; CWP: Compensating Wave Plate; HWP: Half Wave Plate; PBS: Polarizing Beam Splitter; SPD: Single Photon Detector

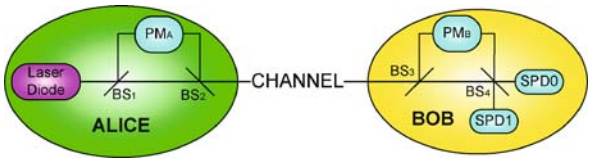

Quantum Cryptography, Figure 4
Conceptual schematic for double Mach–Zehnder interferometer phase-coding BB84 QKD system. PM: Phase Modulator; BS: Beam Splitter; SPD: Single Photon Detector

Phase Coding

Original Scheme is basically a big interferometer as shown in Fig. 2. However it is not practical as the stability of such a huge interferometer is extremely poor.

Double Mach–Zehnder Interferometer Scheme is an improved version of the original proposal. It has two interferometers and there is only one channel connecting Alice and Bob (comparing to the two channels in the original proposal). A conceptual set-up is shown in Fig. 4. We can see that the two signals travel through the same channel. They only propagate through different paths locally in the two Mach–Zehnder interferometers. Therefore people only need to compensate the phase drift of the local interferometers (the polarization drift in the channel still needs to be compensated). This is a great improvement over the original proposal. However, the local compensation has to be implemented in real time. This is quite challenging. An example set-up that implemented the real time compensation of both polarization and phase drifting is reported in [70].

Faraday–Michelson Scheme is an improved version of the double Mach–Zehnder interferometer scheme. It still has two Mach–Zehnder interferometers but each interferometer has only one beam splitter. The light propagates through the same section of fiber twice due to the Faraday mirror. The schematic is shown in Fig. 5. We can see that the polarization drift is self-compensated. This is a great advance in uni-direc-

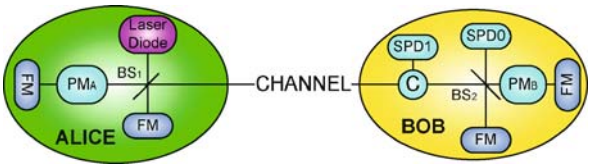

Quantum Cryptography, Figure 5
Conceptual schematic for Faraday–Michelson phase-coding BB84 QKD system. FM: Faraday Mirror, PM: Phase Modulator; BS: Beam Splitter; C: Circulator; SPD: Single Photon Detector

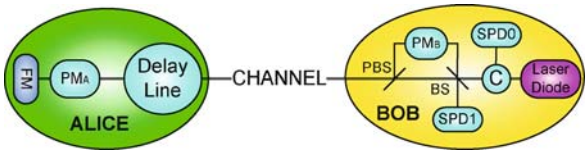

Quantum Cryptography, Figure 6
Conceptual schematic for "Plug & Play" phase-coding BB84 QKD system. FM: Faraday Mirror, PM: Phase Modulator; BS: Beam Splitter; PBS: Polarizing Beam Splitter; C: Circulator; SPD: Single Photon Detector

tional QKD implementation and is first proposed and implemented in [71].

Nonetheless, the phase drift of local interferometers still needs compensation. Due to the fast fluctuation of phase drift (a drift of 2π usually takes a few seconds), this compensation should be done in real-time. A Faraday–Michelson decoy state QKD implementation over 123.6 km has been reported in [59].

Plug & Play Scheme is another improved version of the double Mach–Zehnder interferometer scheme. It has only one Mach–Zehnder interferometer and the light propagates through the same channel and interferometer twice due to the faraday mirror on Alice's side. A conceptual set-up is shown in Fig. 6. We can see that both the polarization drift and the phase drift are automatically compensated. A "Plug & Play" scheme based decoy state QKD implementation over 60 km has been reported in [54].

Nonetheless, the bi-directional design brings complications to security as Eve can make sophisticated operations on the bright pulses sent from Bob to Alice. This is often called the "Trojan horse" attack [72]. Recently, the security of "Plug and Play" QKD system has been proven in [73].

Sagnac Loop Scheme Another bi-directional optical layer design is to use a Sagnac loop where the quantum signal is encoded in the relative phase between the clockwise and counter-clockwise pulses that go through the loop. The typical schematic is shown in Fig. 7.

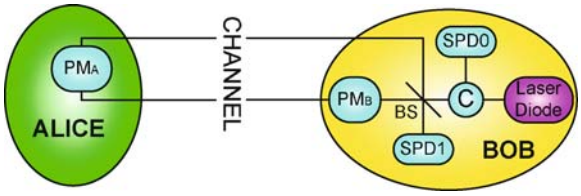

Quantum Cryptography, Figure 7
Conceptual schematic for Sagnac loop phase-coding BB84 QKD system. PM: Phase Modulator; BS: Beam Splitter; C: Circulator; SPD: Single Photon Detector

Sagnac loop QKD is simple to set up and can be easily used in a network setting with a loop topology. However, its security analysis is highly non-trivial.

Other Quantum Key Distribution Protocols

Given the popularity of the BB84 protocol, why should people be interested in other protocols? There are at least three answers to this question. First, to better understand the foundations of QKD and its generality, it is useful to have more than one protocol. Second, different QKD protocols may have different advantages and disadvantages. They may require different technologies to implement. Having different protocols allows us to compare and contrast them. Third, while it is possible to implement standard BB84 protocol with attenuated laser pulses, its performance in terms of key generation rate and maximum transmission distance is somewhat limited. Therefore, we have to study other protocols. Since, from a practical stand point, the third reason is the most important one, we will elaborate on it in the following paragraph.

The original BB84 [1] proposal requires a single photon source. However, most QKD implementations are based on faint lasers due to the great challenge to build perfect single photon sources. Faint laser pulses are weak coherent states that follow Poisson distribution for the photon number. The existence of multi-photon signals opens up new attacks such as photon-number-splitting attack. The basic idea of a photon-number-splitting attack is that Eve can introduce a photon-number-dependent transmittance. In other words, she can selectively suppress single-photon signals and transmit multi-photon signals to Bob. Notice that, for each multi-photon signal, Eve can beamsplit it and keep one copy for herself, thus allowing her to gain a lot of information about the raw key.

The security of coherent laser based QKD systems was analyzed first against individual attacks in 2000 [45], then eventually for a general attack in 2001 [32] and 2002 [33]. Unfortunately, unconditionally secure QKD

based on conventional BB84 protocol [32,33] will severely limit the performance of QKD systems. Basically, Alice has to attenuate her source so that the expected number μ of photon per pulse is of the same order as the transmittance, η. As a result, the key generation rate will scale only quadratically with the transmittance of the channel.

Some of the protocols discussed in the following subsections may dramatically improve the performance of QKD over standard BB84 protocol. For instance, the decoy state protocol has been proven to provide a key generation rate that scales linearly with the transmittance of the channel and has been successfully implemented in experiments.

We conclude with some simple alternative QKD protocols. In 1992, Bennett proposed a protocol (B92) that makes use of only two non-orthogonal states [2]. A six-state QKD protocol was first noted by Bennett and co-workers [74] and some years later by Bruss [75]. It has an advantage of being symmetric. Even QKD protocols with orthogonal states have been proposed [76]. Efficient BB84 and six-state QKD protocols have been proposed and proven to be secure by Lo, Chau and Ardehali [77]. A Singaporean protocol has also been proposed. Recently, Gisin and co-workers proposed a one-way coherent QKD scheme [78].

Entanglement-Based Protocols

Proposals In 1991, Ekert proposed the first entanglement based QKD protocol, commonly called E91 [3]. The basic idea is to test the security of QKD by using the violation of Bell's inequality. Note that one can also implement the BB84 protocol by using an entanglement source. Imagine Eve prepares an entangled state of a pair of qubits and sends one qubit to Alice and the second qubit to Bob. Each of Alice and Bob randomly chooses one of the two conjugate bases to perform a measurement.

Implementations The key part of entanglement-based quantum cryptography is to distribute an entangled pair (usually EPR pair) to two distant parties, Alice and Bob.

Polarization entanglement is preferred in QKD as it is easy to measure the polarization (typically via polarizing beam splitter). The air has negligible birefringence and thus is the perfect channel for polarization-entanglement QKD.

In free-space QKD, atomspherical turbulence may shift the light beam. Therefore the collection of incident photon is challenging. Usually large diameter optical telescope is needed to increase the collection efficiency.

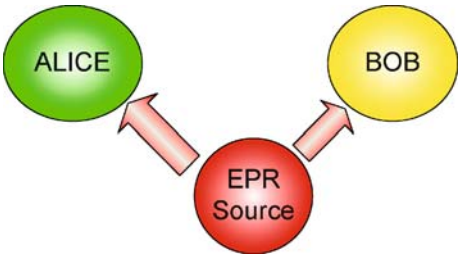

Quantum Cryptography, Figure 8
Conceptual schematic entanglement-based QKD system with the source in the middle of Alice and Bob

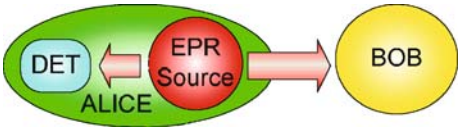

Quantum Cryptography, Figure 9
Conceptual schematic entanglement-based QKD system with the source at Alice's side. DET: Alice's detection system

A standard approach is to put the entanglement source right in the middle of Alice and Bob, see Fig. 8. Once an entangled pair is generated, the two particles are directed to different destinations. Alice and Bob measure the particles locally, and keep the result as the bit value. This approach has potential in the ground-satellite intercontinental entanglement distribution, in which the entanglement source is carried by the satellite and the entangled photons are sent to two distant ground stations. A recent source-in-the-middle entanglement-based quantum communication work over 13 km and is reported in [79].

A simpler version is to include the entanglement source in Alice's side locally, see Fig. 9. Once Alice generates an entangled pair, she keeps one particle and send the other to Bob. Both Alice and Bob measure the particle locally and keep the result as the bit value. This approach is significant simpler than the above design because only Bob needs the telescope and compensating parts. A recent experiment of source-in-Alice entanglement-based quantum communication over 144 km open air is reported in [80].

Decoy State Protocols

Proposals Recall that BB84 implemented with weak coherent state has a key generation rate that scales only quadratically with the transmittance. The decoy state protocol can dramatically increase the key generation rate so that it scales linearly with the transmittance. In a decoy state protocol, Alice prepares some decoy states in addition to signal states. The decoy states are the same as the

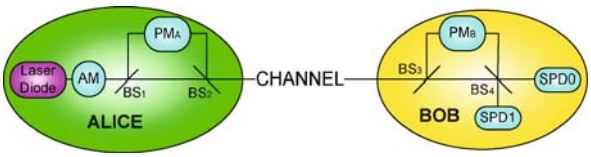

Quantum Cryptography, Figure 10
Conceptual schematic for decoy state BB84 QKD system (double Mach–Zehnder interferometer phase-coding) with amplitude modulator. PM: Phase Modulator; AM: Amplitude Modulator; BS: Beam Splitter; SPD: Single Photon Detector

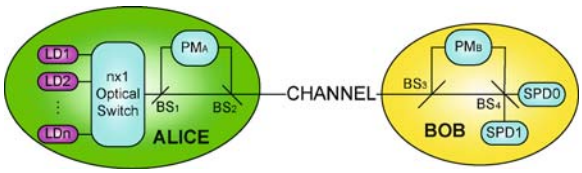

Quantum Cryptography, Figure 11
Conceptual schematic for decoy state BB84 QKD system (double Mach–Zehnder interferometer phase-coding) with multiple laser diodes. LDx: Laser Diodes at different intensities; PM: Phase Modulator; BS: Beam Splitter; SPD: Single Photon Detector

signal state, except for the expected photon number. For instance, if the signal state has an average photon number μ of order 1 (e. g. 0.5), the decoy states have an average photon number ν_1, ν_2, etc. The decoy state idea was first proposed by Hwang [47], who suggested using a large ν (e. g. 2) as a decoy state. Our group provided a rigorous proof of security to decoy state QKD[48,49]. Our numerical simulations showed clearly that decoy states provide a dramatic improvement over non-decoy protocols. In the limit of infinitely many decoy states, Alice and Bob can effectively limit Eve's attack to a simple beam-splitting attack. Moreover, we proposed practical protocols. Instead of using a large ν as a decoy state, we proposed using small ν's as decoy states [48]. For instance, we proposed using a vacuum state as the decoy state to test the background and a weak ν to test the single-photon contribution. We and Wang analyzed the performance of practical protocols in detail [50,51,52].

Notice that the decoy state is a rather general idea that can be applied to other QKD sources. For instance, decoy state protocols have recently been proposed in [81,82,83] for parametric down conversion sources. For a comparison of those protocols, see [84].

Implementations The first experimental demonstration of decoy state QKD was reported by our group in 2006 first over 15 km telecom [53] fiber and later over 60 km telecom fiber [54]. Subsequently, the decoy state QKD was further demonstrated experimentally by several groups worldwide [55,56,57,58].

The implementation of decoy state QKD is straightforward. The key part is to prepare signals with different intensities. A simple solution is to use an amplitude modulator to modulate the intensities of each signal to the desired level, see Fig. 10. Decoy state QKD implementations using amplitude modulator to prepare different states are reported in [53,54,55,56,57].

The amplitude modulator has the disadvantage that the preparation of vacuum state is quite challenging. An alternative solution is to use laser diodes of different intensities to generate different states, see Fig. 11. This solution requires multiple laser diodes and high-speed optical switch, and is thus more complicated than the amplitude modulator solution. It is also challenging to guarantee that all the laser diodes are identical so that Eve cannot tell which source generates some specific pulse. Nonetheless, perfect vacuum states can be easily prepared in this way. Decoy state QKD implementations using multiple laser diodes are reported in [58,59]. Decoy state with a parametric down conversion source has been experimentally implemented in [85].

Strong Reference Pulse Protocols

Proposals The proposal of strong reference pulse QKD dated back to Bennett's 1992 paper [2]. The idea is to add a strong reference pulse, in addition to the signal pulse. The quantum state is encoded in the relative pulse between the reference pulse and the signal pulse. Bob decodes by splitting a part of the strong reference pulse and interfering it with the signal pulse. The strong reference pulse implementation can counter the photon number splitting attack by Eve because it removes the *neutral* signal in the QKD system. Recall that in the photon number splitting attack, Eve suppresses single photon signals by sending a vacuum. This works because the vacuum is a neutral signal that leads to no detection. In contrast, in a strong reference pulse implementation of QKD, a vacuum signal is not a neutral signal. Indeed, if Eve replaces the signal pulse by a vacuum and keeps the strong reference pulse unchanged, then the interference experiment by Bob will give *non-zero* detection probability and a random outcome of "0" or "1". On the other hand, if Eve removes both the signal and the reference pulses, then Bob may detect Eve's attack by monitering the intensity of the reference pulse, which is supposed to be strong.

In some recent papers, the unconditional security of B92 QKD with strong reference pulse has been rigorously proven. However, those proofs require Bob's system

to have certain properties and do not apply to standard threshold detectors.

Implementations The B92 [2] protocol is simpler to implement than BB84 [1] protocol. However, its weakness in security limits people's interest on its implementation. A recent implementation of B92 protocol over 200 m fiber is reported in [86].

Gaussian-Modulated Coherent State (GMCS) Protocol

Proposals Instead of using discrete qubit states as in the BB84 protocol, one may also use continuous variables for QKD. Early proposals of continuous variables QKD use squeezed states, which are experimentally challenging. More recently, gaussian-modulated coherent states have also been proposed for QKD. Since a laser naturally emits a coherent state, compared to a squeezed state QKD proposal, a GMCS QKD protocol is experimentally more feasible. In GMCS QKD, Alice sends Bob a sequence of coherent state signals. For each signal, Alice draws two random numbers X_A and P_A from a set of Gaussian random number with a mean of zero and a variance of $V_A N_0$ and sends a coherent state $|X_A + iP_A\rangle$ to Bob. Bob randomly chooses to measure either the X quadrature of the P quadrature with a phase modulator and a homodyne detector. After performing his measurement, Bob informs Alice which quadrature he has performed for each pulse, through an authenticated public classical channel. Alice drops the irrelevant data and keeps only the quadrature that Bob has measured. Alice and Bob now share a set of correlated Gaussian variables which they regard as the raw key. Alice and Bob randomly select a subset of their signals and publicly broadcast their data to evaluate the excess noise and the transmission efficiency of the quantum channel. If the excess noise is higher than some prescribed level, they abort. Otherwise, Alice and Bob perform key generation by some prescribed protocol.

An advantage of a GMCS QKD is that every signal can be used to generate a key, whereas in qubit-based QKD such as the BB84 protocol losses can substantially reduce the key generation rate. Therefore, it is commonly believed that for short-distance (say < 15 km) applications, GMCS QKD may give a higher key generation rate.

GMCS QKD has been proven to be secure only against individual attacks. The security of GMCS QKD against the most general type of attack – joint attack – remains an open question.

Implementations GMCS protocol has significant advantage over the BB84 [1] protocol at short distances. It was first implemented by F. Grosshans et al. in 2003 [87]. It was shown to be working with channel loss up to 3.1 dB, which is equivalent to the loss of 15 km telecom fiber. Nonetheless, the strong spectral and polarization dispersion of telecom fiber made it challenging to build up a fiber-based GMCS system. Lodewyck, Debuisschert, Tualle-Brouri, and Grangier built the first fiber-based GMCS system in 2005 [88] but only over a few meters. This distance was largely extended to 14 km by Legré, Zbinden, and Gisin in 2006 [89] with the introduction of the "plug & play" design, which brought questions on its security. The uni-directional GMCS QKD has been later implemented over 5 km optical fiber by Qi, Huang, Qian, and Lo in 2007 [90] and over 25 km optical fiber by J. Lodewyck et al. in 2007 [91].

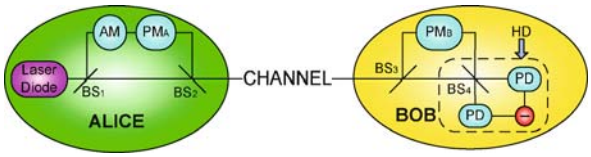

Quantum Cryptography, Figure 12
Conceptual schematic for Gaussian-modulated Coherent State QKD system. PM: Phase Modulator; AM: Amplitude Modulator; BS: Beam Splitter; PD: Photo Diode; HD: Homodyne Detector (inside *dashed box*)

GMCS QKD requires dual-encoding on both amplitude quadrature and phase quadrature, and homodyne detection for decoding, see Fig. 12. Its implementation is in general more challenging than that of BB84 protocol.

Differential-Phase-Shift-Keying (DPSK) Protocols

Proposals In DPSK protocol, a sequence of weak coherent state pulses is sent from Alice to Bob. The key bit is encoded in the relative phase of the adjacent pulses. Therefore, each pulse belongs to two signals. DPSK protocol also defeats the photon number splitting attack by removing the neutral signal. Eve may attack a finite train of signals by measuring its total photon number and then splitting off one photon, whenever the photon number is larger than one. But, since each pulse belongs to two signals, the pulses in the boundary of the train will interfere with the pulses immediately outside the boundary. Therefore, Eve's attack does not allow her to gain full information about all the bit values associated with the train. Moreover, by splitting the signal, Eve has reduced the amplitude of the pulses at the boundary of the train. Therefore, Bob will detect Eve's presence by the higher bit error rates for the bit values between the pulses at the boundary and those just outside the boundary.

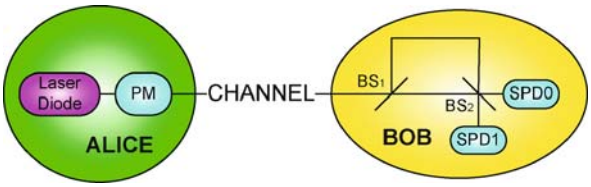

Quantum Cryptography, Figure 13
Conceptual schematic for differential phase shift keying QKD system. PM: Phase Modulator; BS: Beam Splitter; SPD: Single Photon Detector

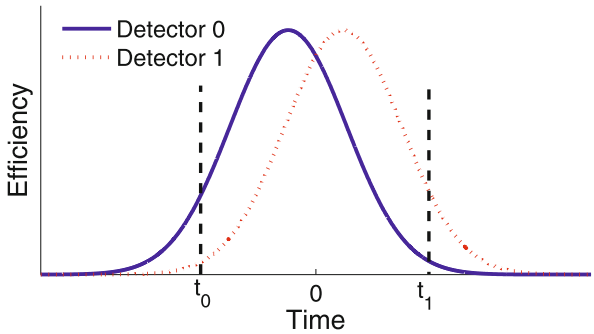

Quantum Cryptography, Figure 14
Conceptual schematic for detection efficiency mismatch in time domain. *X axis*: Time, *Y axis*: Detector Efficiency

While DPSK protocol is simpler to implement than BB84, a proof of its unconditional security is still missing. Therefore, it is hard to quantify its secure key generation rate and perform a fair comparison with, for example, decoy state BB84 protocol. Attacks against DPSK has been studied in, for example, [92,93].

Implementations DPSK protocol is simpler in hardware design than the BB84 [1] protocol as it requires only one Mach–Zehnder interferometer, see Fig. 13. It also has the potential in high-speed applications. Honjo, Inoue, and Takahashi experimentally demonstrated this protocol with a planar light-wave circuit over 20 km fiber in 2004 [94]. This distance was soon extended to 105 km [95] in 2005. In 2007, DPSK scored both the longest and the fastest records in QKD implementations: H. Takesue et al. reported an experimental demonstration of DPSK-QKD over 200 km optical fiber at 10 GHz [61]. However, since a proof of unconditional security is still missing (see last two paragraphs), it is unclear whether the existing experiments generate any secure key. Indeed, the attacks described in [92,93] showed that with only one-way classical post-processing, all existing DPSK experiments are insecure.

Quantum Hacking

Since practical QKD systems exist and commercial QKD systems are on the market, it is important to understand how secure they really are. We remark that there is still a big gap between the theory and practice of QKD. Even though the unconditional security of practical QKD systems with semi-realistic models have been proven [32,33], practical QKD systems may still contain fatal security loopholes. From a historical standpoint, Bennett and Brassard mentioned that the first QKD system [34] was unconditionally secure to any eavesdropper who happened to be deaf! This was because the system made different sounds depending on whether the source was sending a "0" or a "1". Just by listening the sounds, an eavesdropper can learn the value of the final key. This example highlights the existence of side channels in QKD and how easy an eavesdropper might be able to break the security of a QKD system, despite the existence of security proofs.

In this section, we will sketch a few cleverly proposed quantum hacking strategies that are outside standard security proofs and their experimental implementations. We will conclude counter-measures and future outlook. Notice that we will skip eavesdropping attacks that have already been covered by standard security proofs.

Attacks

1) Large Pulse Attack. In a large pulse attack, Eve sends in a strong pulse of laser signal to, for example, Alice laboratory to try to read off Alice's phase modulator setting from a reflected pulse, see [96]. As a result, Eve may learn which BB84 state Alice is sending to Bob.

A simple counter-measure to the large pulse attack is to install an isolator in Alice's system.

2) Faked State Attack. Standard InGaAs detectors suffer from detection efficiency mismatch. More concretely, as noted in Subsect. "Detection Systems", InGaAs detectors are often operated in a gated mode. Therefore, the detection efficiency of each detector is time-dependent. Refer to Fig. 14 for a schematic diagram of the detection efficiencies of two detectors (one for "0" and one for "1") as functions of time. At the expected arrival time, the detection efficiency of the two detectors are similar. However, if the signal is chosen to arrive at some unexpected times, it is possible that the detector efficiencies of the two detectors differ greatly.

The faked state attack proposed by Makarov and coworkers [97] is an intercept-resend attack. In a faked state attack, for each signal, Eve randomly chooses one of the

two BB84 bases to perform a measurement. Eve then sends to Bob a *wrong* bit in the *wrong* basis at a time when the detector for the wrong bit has a low detection efficiency. For instance, if Eve has chosen the rectilinear basis and has found a "0" in the bit value, she then prepares a state "1" in the diagonal basis and sends it to Bob at the arrival time where detector efficiency of the detector for "0" is much higher than that of detector for "1".

Now, should Eve have chosen the wrong basis in her measurement, notice that the detection probability by Bob is greatly suppressed. For instance, in our example, if the correct basis is the diagonal basis, let us consider when happens when Bob measures the signal in the correct basis. Since the bit value resent by Eve is "1" in the diagonal basis and Bob's detector for "1" has a low detection efficiency, most likely Bob will not detect any signal. On the other hand, should Eve have chosen the correct basis in her measurement, Bob has a significant detection efficiency. For instance, in our example, if the correct basis is in fact the rectilinear basis, let us consider what happens when Bob measures in the correct basis. In this case, a bit "1" in the diagonal basis sent by Eve can be re-written as a superposition of a bit "0" and a bit "1" in the rectilinear basis. Since the detector for "0" has a much higher detection efficiency than the detector for "1", most likely Bob will detect a "0". Since "0" was exactly what was originally sent by Alice, Bob will find a rather low bit error rate, despite Eve's intercept-resend attack.

The faked state attack, while conceptually interesting, is hard to implement in practice. This is because it is an intercept-resend attack and as such involves finite detection efficiency in Eve's detectors and precise synchronization between Eve and Alice-Bob's system. For this reason, the faked state attack has never been implemented in practice.

3) Time-shift Attack. The time-shift attack was proposed by Qi, Fung, Lo, and Ma [64]. It also utilizes the detection efficiency mismatch in the time domain, but is much easier to implement than the faked states attack.

As we mentioned in the above section, typical InGaAs-APD detectors usually operate in a gated mode. That is, if the photon hits the detector at unexpected time, the two detectors may have substantially different efficiency. Therefore, Eve can simply shift the arrival time of each signal, creating large efficiency mismatch between "0"s and "1"s.

Let's take a specific example to illustrate this attack: suppose detector 0 has higher efficiency than detector 1 if the signal arrives *earlier* than the expected time, and lower efficiency than detector 1 if the signal arrives *later* than expected. Eve can simply shift the arrival time of each bit by sending it through a longer path or a shorter one. Consider the case in which Eve sends bit i through a shorter path. In this case bit i will hit the detector *earlier* than expected, thus detector 0 has much higher efficiency. If Bob reports a detection event for the ith bit, Eve can make a guess that this bit is a "0" with high probability of success.

Furthermore, Eve can carefully set how many bits should be shifted forward and how many should be shifted backward so that Bob gets similar counts of "0"s and "1"s. In this way, Bob cannot observe a mismatch between the numbers of "0"s and "1"s.

Note that the time-shift attack does not make any measurement on the qubits. Therefore, quantum information is not destroyed. That is, Eve does not change the polarization, the phase, or the frequency of any bit. This means the time-shift attack will not increase the bit error rate of the system in principle. Moreover, since Eve does not need to make any measurement or state preparation, the time-shift attack is practically feasible even with current technology.

The time-shift attack will introduce some loss as the overall detection efficiency is lower if the photon hits the detector at an unexpected time. Nonetheless, Eve can compensate this loss by making the channel more transparent. Notice that, since the quantum channel between Alice and Bob may contain many lossy components such as splices and couplers, it may not be too hard for Eve to make a channel more transparent.

The time-shift attack has been successfully implemented on a commercial QKD system by Zhao, Fung, Qi, Chen, and Lo [65] in 2007. This is the first and so far the only experimentally successful demonstration of quantum hacking on commercial QKD system. It is shown that the system has no-negligible probability to be vulnerable to the time-shift attack. Quantitative analysis shows that the final key shared by Alice and Bob (after the error correction and the privacy amplification of the most general security analysis) has been compromised by Eve.

The success of the time-shift attack in [65] is rather surprising as QKD has been widely believed to be unconditionally secure. The experimental success in quantum hacking highlighted the limit of the whole research program of device-independent security proofs [23] by showing that device-independent security proofs, even if they are found to exist in future, do not apply to a practical QKD system. The success of time-shift attack is not due to some technical imperfection. It is deeply connected with the detection efficiency loophole in the verification of Bell-inequalities. So far the InGaAs detectors have only $\sim 10\%$ detection efficiencies, and the channel connecting Alice and Bob usually has quite large attenuation for long distance communication. The low overall detection efficiency fails the device-independent security proof.

Notice that even non-gated detectors have dead times and generally suffer from detection efficiency loophole. The detection efficiency mismatch is also discussed in [98].

4) Phase remapping attack. In a bi-directional implementation of QKD such as the "Plug and Play" set-up, Eve may attempt to tamper with Alice's preparation process so that Alice prepares four wrong states, instead of the four standard BB84 state. This is called the phase remapping attack and was proposed in [99].

In a "Plug and Play" QKD system, Alice receives a strong pulse from Bob and she then attenuates it to a single-photon level and encodes one of the BB84 state on it. For instance, Alice may encode her state by using a phase modulator. In a phase modulator, the encoded phase is proportional to the voltage applied. In practice, a phase modulator has a finite rise time. For each BB84 setting, one may thus model the applied voltage (and thus the encoded phase) as a trapezium. Ideally, Bob's strong pulse should arrive at the plateau region of the phase modulation, thus getting maximal phase modulation by Alice's phase modulator. Now, imagine that Eve applies a time-shift to Bob's strong pulse so that it arrives in the rise region of the phase modulation graph instead of the plateau region. In this case, Alice has wrongly encoded her phase only partially.

If we assume that the four settings of Alice's encoding (for the four BB84 states) have the same rise region, then by time-shifting Bob's strong pulse, Eve can force Alice to prepare the four state with phase $0, a, 2a, 3a$, rather than $0, \pi/2, \pi, 3\pi/2$. In general, these four states are more distinguishable than the standard BB84 state. Therefore, Eve may subsequently apply an intercept-resend attack to the signal sent out by Alice.

It was proven in [99] that in principle Eve can break the security of the QKD system, without alerting Alice and Bob.

5) Attack by passive listening to side channels. The attack by listening to the sounds made by the source in the first QKD experiment is an example of an attack by passive listening to side channels. Another example is [100]. A counter-measure is to carefully locate all possible side channels and to eliminate them one by one.

6) Saturation Attack. In a recent preprint [101], Makarov studied experimentally how by sending a moderately bright pulse, Eve can blind Bob's InGaAs detector. A simple counter-measure would be for Bob to measure the intensity of the incoming signal.

7) High Power Damage Attack. In Makarov's thesis, it was proposed that Eve may try to make controlled changes in Alice's and Bob's system by using high power laser damage through sending a very strong laser pulse. Again, a simple counter-measure would be for Alice and Bob to measure the intensity of the incoming signals and monitor the properties of various components from time to time to ensure that they perform properly.

Counter-Measures

Once an attack is known, there are often simple counter-measures. For instance, for the large pulse attack, a simple counter-measure would be to add a circulator in Alice's laboratory. As for the faked state attack and time-shift attack, a simple counter-measure would be for Bob to use a four-state setting in his phase modulator. Other counter-measures include Bob applying a random time-shift to his received signals. However, the most dangerous attacks are the *unanticipated* ones.

Notice that it is not enough to say that a counter-measure to an attack exists. It is necessary to actually implement a counter-measure experimentally in order to see how effective and convenient it really is. This will allow Alice and Bob to select a useful counter-measure. Moreover, notice that the implementation of a counter-measure may itself open up new loopholes. For instance, if Bob implements a four-state setting as a counter-measure to a time-shift attack, Eve may still combine a large pulse attack with the time-shift attack to break a QKD system.

Importance of Quantum Hacking

As noted in Sect. "Security Proofs", there has been a lot of theoretical interest on the connection between the security of QKD and fundamental physical principles such as the violation of Bell's inequality. An ultimate goal of such investigations, which has not been realized yet, is to construct a device-independent security proof [23]. Even if such a goal is achieved in future, would any of these theoretical security proofs applies to a quantum key distribution system in *practice*? Unfortunately, the answer is no. As is well-known, the experimental testing of Bell-inequalities often suffers from the detection efficiency loophole [23]. The low detection efficiency of practical detectors not only nullifies security proofs based on Bell-inequality violation, but also gives an eavesdropper a powerful handle to break the security of a practical QKD system. Therefore, the detection efficiency loophole is of both theoretical and practical interest.

A practical QKD system often consists of two or more detectors. In practice, it is very hard to construct detectors of identical characteristics. As a result, two detectors can generally exhibit different detector efficiencies as functions of either one or a combination of variables in the time, frequency, polarization or spatial domains. Now, if an eavesdropper could manipulate a signal in these variables, then

she could effectively exploit the detection efficiency loophole to break the security of a QKD system. In fact, she could even violate a Bell-inequality with only a classical source. In time-shift attack, one consider an eavesdropper's manipulation of the time variable. However, the generality of detection efficiency loophole and detector efficiency mismatch should not be lost.

We should remark that, for eavesdropping attacks, the sky is the limit. The more imaginative one is, the more new attacks one comes up. Indeed, what people have done so far are just scratching the surface of the subject. Much more work needs to be done in the battle-testing of QKD systems and security proofs with testable assumptions. See Sect. "Future Directions".

Beyond Quantum Key Distribution

Besides QKD, many other applications of quantum cryptography have been proposed. Consider, for instance, the millionaires' problem. Two millionaires, Alice and Bob, would like to determine who is richer without disclosing the actual amount of money each has to each other. More generally, in a secure two-party computation, two distant parties, Alice and Bob, with private inputs, x and y respectively, would like to compute a prescribed function $f(x, y)$ in such a way that at the end, they learn the outcome $f(x, y)$, but nothing about the other party's input, other than what can be logically be deduced from the value of $f(x, y)$ and his/her input. There are many possible functions $f(x, y)$. Instead of implementing them one by one, it is useful to construct some cryptographic primitives, which if available, can be used to implement the secure computation of *any* function $f(x, y)$. In classical cryptography, to implement secure two-party computation of a general function will require making additional assumptions such as a trusted third party or computational assumptions. The question is whether we can do unconditionally secure *quantum* secure two-party computations.

Two important cryptographic primitives are namely quantum bit commitment (QBC) and one-out-of-two quantum oblivious transfer (QOT). In particular, it was shown by Kilian [102] that in classical cryptography, oblivious transfer can be used to implement a general two-party secure computation of any function $f(x, y)$. Moreover, in quantum cryptography, it was proven by Yao [103] that a secure QBC scheme can be used to implement QOT securely. For a long time back in the early 1990s, there was high hope that QBC and QOT could be done with unconditional security. In fact, in a paper [104] it was claimed that QBC can be made unconditionally secure. The sky fell around 1996 when Mayers [105] and subsequently, Lo and Chau [106], proved that, contrary to widespread belief at that time, unconditionally secure QBC is, in fact, impossible. Subsequently, Lo [107] proved explicitly that unconditionally secure one-out-of-two QOT is also impossible. Mayers and Lo-Chau's result was a big step backwards and thus a big disappointment for quantum security.

After the fall of QBC and QOT, people turned their attention to quantum coin tossing (QCT). Suppose Alice and Bob are having a divorce and they would like to determine by a coin toss who is going to keep their kid. They do not trust each other. However, they live far away from each other and have to do a coin toss remotely. How can they do so without trusting each other? Classically, coin tossing will require either a trusted third party or making computational assumptions. As shown by Lo and Chau, ideal quantum coin tossing is impossible [108]. Even for the non-ideal case, Kitaev has proven that a strong version of QCT (called *strong* QCT) cannot be unconditionally secure. However, despite numerous papers on the subject (see, for example, [109] and references therein), whether non-ideal *weak* QCT is possible remains an open question.

Other QKD protocols are also of interest. For instance, the sharing of quantum secrets has been proposed in [110]. It is an important primitive for building other protocols such as secure multi-party quantum computation [111]. There are also protocols for quantum digital signatures [112], quantum fingerprinting and unclonable encryption. Incidentally, quantum mechanics can also be used for the quantum sharing of classical secrets [113], conference key agreement and third-man cryptography.

For QKD, so far we have only discussed a point-to-point configuration. In real-life applications, it will be interesting to study QKD in a network setting [114]. Note that the multiplexing of several QKD channels in the single fiber has been successful performed. So has the multiplexing of a classical channel together with a QKD channel. However, much work remains to be done on the design of both the key management structure and the optical layer of a QKD network.

Future Directions

The subject of quantum cryptography is still in a state of flux. We will conclude with a few examples of future directions.

Quantum Repeaters

Losses in quantum channels greatly limit the distance and key generation rate of QKD. To achieve secure QKD over long distances without trusting the intermediate nodes, it is highly desirable to have quantum repeaters. Briefly

stated, quantum repeaters are primitive quantum computers can be perform some form of quantum error correction, thus preserving the quantum signals used in QKD. In more detail, quantum repeaters often rely on the concept of entanglement distillation, whose goal is, given a large number M of noisy entangled states, two parties, Alice and Bob, perform local operations and classical communications to distill out a smaller number (say N) but less noisy entangled states.

The experimental development of a quantum repeater will probably involve the development of quantum memories together with the interface between flying qubits and qubits in a quantum memory.

Ground to Satellite QKD

Another method to extend the distance of QKD is to perform QKD between a satellite and a ground station. If one trusts a satellite, one can even build a global QKD network via a satellite relay. Basically, a satellite can perform QKD with Alice first, when it has a line of sight with Alice. Afterwards, it moves in orbit until it has a line of sight with Bob. Then, the satellite performs a separate QKD with Bob. By broadcasting the XOR of the two keys, Alice and Bob will share the same key. Satellite to ground QKD appears to be feasible with current or near-future technology, for a discussion, see, for example, [62].

With an untrusted satellite, one can still achieve secure QKD between two ground stations by putting an entangled source at the satellite and sending one half of each entangled pair to each of Alice and Bob.

Calculation of the Quantum Key Capacity

Given a specific theoretical model, so far it is not known how to calculate the actual secure key generation rate in a noisy channel. All is known is how to calculate some upper bounds and lower bounds. This is a highly unsatisfactory situation because we do not really know the actual fundamental limit of the system. Our ignorance can be highlighted by a simple open question: what is the highest tolerable bit error rate of BB84 that will still allow the generation of a secure key?

While lower bounds are known [25,26,27] and 25 percent is an upper bound set by a simple intercept-resend attack, we do not know the answer to this simple question.

Notice that this question is of both fundamental and practical interests. Without knowing the fundamental limit of the key generation rate, we do not know what the most efficient procedure for generating a key in a practical setting is.

Multi-party Quantum Key Distribution and Entanglement

Besides its technological interest, QKD is of fundamental interest because it is deeply related to the theory of entanglement, which is the essence of quantum mechanics. So far there have been limited studies on multi-party QKD. Notice that there are many deep unresolved problems in *multi*-party entanglement. It would be interesting to study more deeply multi-party QKD and understand better its connection to multi-party entanglement. Hopefully, this will shed some light on the mysterious nature of multi-party entanglement.

Security Proofs with Testable Assumptions

The surprising success of quantum hacking highlights the big gap between the theory and practice of QKD. In our opinion, it is important to work on security proofs with *testable* assumptions. Every assumption in a security proof should be written down and experimentally verified. This is a long-term research program.

Battle-Testing QKD Systems

Only through battle-testing can we gain confidence about the security of a real-life QKD system. Traditionally, breaking a cryptographic systems is as important as building one. Therefore, we need to re-double our efforts on the study of eavesdropping attacks and their counter-measures.

As stated before, quantum cryptography enjoys forward security. Thanks to the quantum no-cloning theorem, an eavesdropper Eve does not have a transcript of all quantum signals sent by Alice to Bob. Therefore, once a QKD process has been performed, the information is gone and it will be too late for Eve to go back to eavesdrop. Therefore, for Eve to break a real-life QKD system today, it is imperative for Eve to invest in technologies for eavesdropping now, rather than in future.

Acknowledgments

We thank various funding agencies including NSERC, CRC program, QuantumWorks, CIFAR, MITACS, CIPI, PREA, CFI, and OIT for their financial support.

Bibliography

Primary Literature

1. Bennett CH, Brassard G (1984) In: Proceedings of IEEE International Conference on Computers, Systems, and Signal Processing. IEEE, New York, pp 175–179

2. Bennett CH (1992) Phys Rev Lett 68:3121
3. Ekert AK (1991) Phys Rev Lett 67:661
4. Wiesner S (1983) Sigact News 15:78
5. Vernam G (1926) J Am Inst Electr Eng 45:109
6. Shor PW (1997) SIAM J Sci Statist Comput 26:1484
7. Brassard and Crépeau (1996) ACM SIGACT News 27:13
8. Mayers D (2001) J ACM 48:351
9. Lo H-K, Chau HF (1999) Science 283:2050
10. Bennett CH, DiVincenzo DP, Smolin JA, Wootters WK (1996) Phys Rev A 54:3824
11. Deutsch D, Ekert A, Jozsa R, Macchiavello C, Popescu S, Sanpera A (1996) Phys Rev Lett 77:2818
12. Shor P, Preskill J (2000) Phys Rev Lett 85:441
13. Biham E, Boyer M, Boykin PO, Mor T, Roychowdhury V (2000) In: STOC '00: Proceedings of the thirty-second annual ACM symposium on Theory of computing. ACM Press, New York, pp 715–724
14. Ben-Or M (2002) http://www.msri.org/publications/ln/msri/2002/qip/ben-or/1/index.html. Accessed 03 Nov 2008
15. Given a density matrix, ρ, define the von Neumann entropy of ρ as $S(\rho) = -\sum_i \lambda_i \log_2 \lambda_i = -\mathrm{tr}\rho \log_2 \rho$ where λ_i's are eigenvalues of the density matrix ρ
16. http://en.wikipedia.org/wiki/Holevo's_theorem. Accessed 03 Nov 2008
17. Renner R, Koenig R (2005) In: TCC 2005. LNCS, vol 3378. Springer, Berlin, (eprint) quant-ph/0403133
18. Renner R (2005) (eprint) quant-ph/0512258
19. Renner R (2007) Nat Phys 3:645
20. Horodecki K, Horodecki M, Horodecki P, Oppenheim J (2005) Phys Rev Lett 94:160502
21. Horodecki K, Horodecki M, Horodecki P, Leung D, Oppenheim J (2008) IEEE Trans Inf Theory 54(6):2604–2620
22. Koashi M (2007) arXiv:0704.3661
23. Acin A, Brunner N, Gisin N, Massar S, Pironio S, Scarani V (2007) Phys Rev Lett 98:230501
24. Masanes L, Winter A (2006) (eprint) quant-ph/0606049. Accessed 03 Nov 2008
25. Gottesman D, Lo H-K (2003) IEEE Trans Inf Theory 49:457
26. Renner R, Gisin N, Kraus B (2005) Phys Rev A 72:012332
27. Chau HF (2002) Phys Rev A 66:060302
28. Brassard G, Salvail L (1994) Lecture Notes in Computer Science. Springer, vol 765, pp 410–423
29. Bennett CH, Brassard G, Crepeau C, Maurer UM (1995) IEEE Trans Info Theory 41:1915
30. Ben-Or M, Horodecki M, Leung DW, Mayers D, Oppenheim J (2005) In: Kilian J (ed) Theory of Cryptography: Second Theory of Cryptography Conference. TCC 2005. Lecture Notes in Computer Science, vol 3378. Springer, Berlin, pp 386–406
31. Koenig R, Renner R, Bariska A, Maurer U (2007) Phys Rev Lett 98:140502
32. Inamori H, Lütkenhaus N, Mayers D (2007) Eur Phys J D 41:599
33. Gottesman D, Lo H-K, Lütkenhaus N, Preskill J (2004) Quant Info Compu 4:325
34. Bennett CH, Bessette F, Brassard G, Salvail L, Smolin J (1992) J Cryptogr 5:3
35. Townsend PD, Rarity JG, Tapster PR (1993) Electron Lett 29:634
36. Muller A, Breguet J, Gisin N (1993) Europhys Lett 23:383
37. Jacobs BC, Franson JD (1996) Opt Lett 21:1854
38. Franson JD, Lives H (1994) Appl Opt 33:2949
39. Townsend PD (1994) Electron Lett 30:809
40. Muller A, Zbinden H, Gisin N (1995) Nature 378:449
41. Muller A, Zbinden H, Gisin N (1996) Europhys Lett 33:335
42. Muller A, Herzog T, Hutter B, Tittel W, Zbinden H, Gisin N (1997) Appl Phys Lett 70:793
43. Zbinden H, Gautier J-D, Gisin N, Hutter B, Muller A, Tittel W (1997) Electron Lett 33:586
44. Stucki D, Gisin N, Guinnard O, Robordy G, Zbinden H (2002) New J Phys 4:41
45. Lütkenhaus N (2000) Phys Rev A 61:052304
46. Gobby C, Yuan ZL, Shields AJ (2004) Electron Lett 40:1603
47. Hwang WY (2003) Phys Rev Lett 91:057901
48. Lo H-K (2004) In: Proceedings of IEEE International Symposium on Information Theory IEEE, New York, p 137
49. Lo H-K, Ma X, Chen K (2005) Phys Rev Lett 94:230504
50. Ma X, Qi B, Zhao Y, Lo H-K (2005) Phys Rev A 72:012326
51. Wang X-B (2005) Phys Rev Lett 94:230503
52. Wang X-B (2005) Phys Rev A 72:012322
53. Zhao Y, Qi B, Ma X, Lo H-K, Qian L (2006a) Phys Rev Lett 96:070502
54. Zhao Y, Qi B, Ma X, Lo H-K, Qian L (2006b) In: Proceedings of IEEE International Symposium of Information Theory, IEEE, New York, pp 2094–2098
55. Schmitt-Manderbach T et al (2007) Phys Rev Lett 98:010504
56. Rosenberg D et al (2007) Phys Rev Lett 98:010503
57. Yuan ZL, Sharpe AW, Shields AJ (2007) Appl Phys Lett 90:011118
58. Peng C-Z et al (2007) Phys Rev Lett 98:010505
59. Yin Z-Q, Han Z-F, Chen W, Xu F-X, Wu Q-L, Guo G-C (2008) Chin Phys Lett 25:3547
60. Wang Q, Karlsson A (2007) Phys Rev A 76:014309
61. Takesue H et al (2007) Nat Photonics 1:343
62. Villoresi P et al (2004) arXiv:quant-ph/0408067v1
63. Stucki D, Ribordy G, Stefanov A, Zbinden H, Rarity JG, Wall T (2001) J Mod Opt 48:1967
64. Qi B, Fung C-HF, Lo H-K, Ma X (2007a) Quant Info Compu 7:73
65. Zhao Y, Fung C-HF, Qi B, Chen C, Lo H-K (2008) Phys Rev A 78:042333
66. Namekata N, Fujii G, Inoue S (2007) Appl Phys Lett 91:011112
67. Thew RT et al (2006) New J of Phys 8:32
68. Diamanti E, Takesue H, Langrock C, Fejer MM, Yamamoto Y (2006) Opt Express 14:13073, (eprint) quant-ph/0608110
69. Rosenberg D et al (2006) Appl Phys Lett 88:021108
70. Dynes JF, Yuan ZL, Sharpe AW, Shields AJ (2007) Opt Express 15:8465
71. Mo XF, Zhu B, Han ZF, Gui YZ, Guo GC (2005) Opt Lett 30:2632
72. Gisin N, Fasel S, Kraus B, Zbinden H, Ribordy G (2006) Phys Rev A 73:022320
73. Zhao Y, Qi B, Lo H-K (2008) Phys Rev A 77:052327
74. Bennett CH, Brassard G, Breidbart S, Wiesner S (1984) IBM Technical Disclosure Bulletin 26:4363
75. Bruss D (1998) Phys Rev Lett 81:3018
76. Goldenberg L, Vaidman L (1995) Phys Rev Lett 75:1239
77. Lo H-K, Chau HF, Ardehali M (2005) J Cryptology 18:133
78. Gisin N, Ribordy G, Zbinden H, Stucki D, Brunner N, Scarani V (2004) arXiv:quant-ph/0411022v1
79. Peng C-Z et al (2005) Phys Rev Lett 94:150501
80. Ursin R et al (2007) Nat Phys 3:481
81. Mauerer W, Silberhorn C (2007) Phys Rev A 75:050305
82. Wang Q, Wang X-B, Guo G-C (2007) Phys Rev A 75:012312
83. Adachi Y, Yamamoto T, Koashi M, Imoto N (2007) Phys Rev Lett 99:180503

84. Ma X, Lo H-K (2008) New J Phys 10:073018
85. Wang Q et al (2008) Phys Rev Lett 100:090501
86. Mendonça F, de Brito DB, Silva JBR, Thé GAP, Ramos RV (2008) Microw Opt Tech Lett 50:236
87. Crosshans F, Assche GV, Wenger J, Brourl R, Cerf NJ, Grangier P (2003) Nature 421:238
88. Lodewyck J, Debuisschert T, Tualle-Brouri R, Grangier P (2005) Phys Rev A 72:050303
89. Legré M, Zbinden H, Gisin N (2006) Quant Info Compu 6:326
90. Qi B, Huang L-L, Qian L, Lo H-K (2007) Phys Rev A 76:052323. arXiv:0709.3666
91. Lodewyck J et al (2007) Phys Rev A 76:042305
92. Curty M, Zhang LLX, Lo H-K, Lütkenhaus N (2007) Quant Info Compu 7:665
93. Tsurumaru T (2007) Phys Rev A 75:062319
94. Honjo T, Inoue K, Takahashi H (2004) Opt Lett 29:2797
95. Takesue H et al (2005) New J Phys 7:232
96. Gisin N, Fasel S, Kraus B, Zbinden H, Ribordy G (2006) Phys Rev A 73:022320
97. Makarov V, Anisimov A, Skaar J (2006) Phys Rev A 74:022313
98. Larsson J-Å (2002) Quant Info Compu 2:434
99. Fung C-HF, Qi B, Tamaki K, Lo H-K (2007) Phys Rev A 75:032314
100. Lamas-Linares A, Kurtsiefer C (2007) Opt Express 15:9388
101. Makarov V (2007) arXiv:0707.3987v1
102. Kilian J (1988) In: Proceedings of the 20th Annual ACM Symposium on Theory of Computing. ACM, New York, pp 20–31
103. Yao AC-C (1995) In: Proceedings of the 26th Annual ACM Symposium on the Theory of Computing. ACM, New York, p 67
104. Brassard RJG, Crépeau C, Langlois D (1993) In: Proceedings of the 34th Annual IEEE Symposium on the Foundations of Computer Science. IEEE, New York, p 362
105. Mayers D (1997) Phys Rev Lett 78:3414
106. Lo H-K, Chau HF (1997) Phys Rev Lett 78:3410
107. Lo H-K (1997) Phys Rev A 56:1154
108. Lo H-K, Chau HF (1997) Physica D 120:177
109. Mochon C (2005) Phys Rev A 72:022341
110. Cleve R, Gottesman D, Lo H-K (1999) Phys Rev Lett 83:648
111. Ben-Or M, Crépeau C, Gottesman D, Hassidim A, Smith A (2006) In: Proceedings of 47th Annual IEEE Symposium on the Foundations of Computer Science (FOCS'06). IEEE, New York, pp 249–260
112. Gottesman D, Chuang I (2001) arXiv:quant-ph/0105032v2
113. Hillery M, Bužek V, Berthiaume A (1999) Phys Rev A 59:1829
114. Alleaume R et al (2007) quant-ph/0701168. Accessed 03 Nov 2008

Books and Reviews

Brassard G (1994) A Bibliography of Quantum Cryptography. http://www.cs.mcgill.ca/~crepeau/CRYPTO/Biblio-QC.html. Accessed 03 Nov 2008

Gisin N, Thew R (2007) Quantum Communication. Nat Photon 1(3):165–171. On-line available at http://arxiv.org/abs/quant-ph/0703255. Accessed 03 Nov 2008

Gisin N, Ribordy G, Tittel W, Zbinden H (2002) Quantum Cryptography. Rev Mod Phys 74:145–195. On-line Available at http://arxiv.org/abs/quant-ph/0101098. Accessed 03 Nov 2008

Gottesman D, Lo H-K (2000) From Quantum Cheating to Quantum Security. Physics Today, Nov 2000, p 22

Lo H-K, Lütkenhaus N (2007) Quantum Cryptography: from theory to practice Invited paper for Physics In Canada, Sept-Dec 2007. On-line available at http://arxiv.org/abs/quant-ph/0702202. Accessed 03 Nov 2008

Scarani V, Bechmann H-Pasquinucci, Cerf NJ, Dusek M, Lütkenhaus N, Peev M (2008) The security of practical quantum key distribution. arXiv:0802.4155v2

Wikipedia (2008) Quantum Cryptography, http://en.wikipedia.org/wiki/Quantum_cryptography. Accessed 03 Nov 2008

Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring

Arkady M. Satanin[1], Eric R. Hedin[2], Yong S. Joe[2]
[1] Institute for Physics of Microstructures, Russian Academy of Sciences, Nizhny Novgorod, Russia
[2] Center for Computational Nanoscience, Department of Physics and Astronomy, Ball State University, Muncie, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Fano Resonance in a Two Dimensional Electron Waveguide with a Quantum Dot
Model for the Quantum Dots in the Ring: Theory of Waveguides Approach
Effects of Interference Interactions of Fano Resonances on an AB Ring
Fano Resonance Induced by a Quantum Dot in an Open Three-Terminal Interferometer
Future Directions
Acknowledgments
Bibliography

Glossary

Nanostructures Modern nanotechnology achieves the arrangement of atoms into structures that are only a few nanometers in size. The following kinds of artificial structures (semiconductor or metallic) may be formed: two- dimensional or layered heterostructures (superlattices); one-dimensional nanowires, or zero-dimensional quantum dots. The characteristic size of systems is between 0.1 and 100 nm.

Quantum dots (QDs) A quantum dot is a semiconductor nanostructure that confines the motion of electrons in all thee spatial directions. The confinement can be due

to electrostatic potentials (generated by external electrodes, doping, strain, or impurities). A quantum dot has discrete allowed energy levels for which the corresponding wavefunctions are spatially localized within the quantum dot.

Wave interference Wave interference is the phenomenon which occurs when two waves meet while traveling in the same medium. If the waves have the same phases, constructive interference takes place, and destructive occurs for waves with opposite phases.

Aharonov–Bohm (AB) effect The Aharonov–Bohm effect is a quantum mechanical phenomenon by which a charged particle is affected by an electromagnetic field which only exists in regions apart from the particle. If there are two propagation paths by which a particle can move without entering the region of space where the magnetic field exists, then a charged quantum particle can still show an observable phase shift in its interference pattern.

Breit–Wigner (BW) resonances The Breit–Wigner resonances arise due to the constructive interference of two counter-propagating waves in the same scattering channel (similar to resonances of the Fabry–Perot interferometer in optics). For instance, the Breit–Wigner resonances occur when an electron moves in the space between two semiconductor layers (barriers) for some particular electron energies (quasi-bound energies). For this case, the Breit–Wigner resonances appear in the transmission. In addition, the scattering amplitude as a function of electron energy possesses a pole for each quasi-bound state in the complex-energy plane. The real part of a pole can be interpreted as the energy of a quasi-bound state and the imaginary part can be connected with the lifetime of this state.

Fano resonances The Fano resonance is a manifestation of the interference between a localized state and the continuum states in the transmission of multi-channel systems. In AB rings, the Fano resonances arise from quantum mechanical interference between the discrete states of the QD in one arm of the interferometer and the continuum in the other arm, characterized by complete transmission and complete reflection. The profile of the Fano asymmetric line-shape in the transmission depends on the strength of the coupling between discrete and continuum states, and on the phase difference between the paths.

Definition of the Subject

Resonance phenomena are a major subject of theoretical and experimental investigations, and the concept of resonances is ubiquitous in physics. The search for new effects related to wave interference and different kinds of resonances in various physical systems continues to be of interest. Interference of a localized wave with propagating states and the resulting Fano resonances in atomic and solid states structures have attracted much attention recently. Now, it is clear that Fano interference is a universal phenomenon because the manifestation of the interference does not depend on the configuration of the material. The natural question then arises: Why are Fano-interference phenomena so interesting in different fields of physics? From the practical point of view, for instance, the resonances can be considered as quantum *probes* that provide important information on the geometric configuration and internal potential fields of low-dimensional structures. Fano interference may potentially be used for the design of new types of quantum electronic or spintronic devices such as Fano-transistors, spin transistors, and Fano-filters for polarized electrons. In addition, Fano phenomena can also be used for lasing without population inversion. The investigation of these resonances is of great importance in the search for optimal working parameters of new devices, such as the resonant diode and the resonant transistor.

Introduction

At present, nanotechnology provides various solid state systems such as Aharonov–Bohm (AB) rings, two-dimensional (2D) electronic waveguides, nanotubes etc., where alternative electronic paths may be realized. In modern laboratories, various AB rings with embedded quantum dots in the arms have been fabricated [28,29,31,32,42,46,47]. If a quantum ring supports coherent transmission, the wave amplitudes through the two arms interfere. The scattering of the waves on the embedded dots, and the interference of waves in the arms results in resonances in the conductance (or transmission) of the ring. At the same time, the electron scattering on the dots produces an additional phase shift. Thus, the interference of the waves coming from the two arms contains information about the amplitude and phase of the waves. In the other words, an AB electronic interferometer operates similarly to an optical interferometer providing illumination of the phase-shift.

It is well known that resonant-transmission phenomena are related to the quasi-bound states of the systems. According to the Breit–Wigner (BW) formalism in nuclear systems, the scattering amplitude as a function of energy possesses a pole for each quasi-bound state in the complex-energy plane [8]. The real part of a pole can be interpreted as the energy of a quasi-bound state and the

imaginary part can be connected with the life time of this state [15]. Thus, BW resonances arise due to the interference of two counter-propagating waves in the same scattering channel (similar to resonances of the Fabry–Perot interferometer in optics).

In contrast to multi-barrier resonant-tunneling structures, quantum nanostructures such as AB rings and 2D-electronic waveguides, where alternative electronic paths may be realized, possess both transmission zeros and resonance poles. This characteristic of a zero-pole pair, called a Fano resonance [14], has been specifically predicted and observed in a hybrid system of an AB ring and a QD both theoretically and experimentally [2,3,4,5,6,13,16,17,18,28,29,30,31,32,34,36,42,43,45,46,47]. In order to fit the experimental data near an asymmetric resonance, the authors of experimental works have used in expression for the conductance G in the form

$$G = G_{\mathrm{b}} \frac{|E - E_0|^2}{(E - E_{\mathrm{R}})^2 + \Gamma^2} , \tag{1}$$

where G_{b} is the non-resonant conductance, E_{R} is the energy of the resonance, Γ is the width, and E_0 is a complex parameter (in general) – the resonance zero. It has been demonstrated that the pole is an immediate corollary of the electron transition from the bound state to the continuum [45]. Although the nature of the zero is connected with the bound state it also depends on the detail of the interference with the continuum. Notice, that for systems with time reversal symmetry, E_0 always is placed on real axis of energy.

This Fano effect arises from quantum mechanical interference between the discrete state of the QD in one arm and the continuum in other arm. The profile of the Fano asymmetric line-shape in the transmission depends on the strength of the coupling between discrete and continuum states, and on the phase difference between the paths. Here, the scattering amplitude near the zero-pole pair behaves like a dipole, where the pole plays the role of a particle and the zero plays the role of a hole (antiparticle) [16,17,30,34,36,43,45]. The collapse of the particle and hole has been studied in a quasi-one-dimensional constriction with an attractive and finite-size impurity, by modulating the parameters of the system [17,30].

When more than one resonant quasi-bound state is present in a one-channel system, for instance, in a three-barrier system, the resonance levels interact each other and result in the overlapping of resonances [17,19,20,22,23,24,25,26,30,38,39,40,41]. In this situation, therefore, the single BW formula is no longer valid due to the overlapping of resonances. Hence, the interference effects of resonances have been studied by examining the formation of double poles in the transmission amplitude and the effect of the collision of the two poles (or merging of two resonances) associated with the quasi-bound states [19,20,22,23,24,25,26,38,39,40,41].

It was demonstrated recently that the Fano resonance structure can be controlled by changing the confinement parameters of the QD. Transmission through a QD embedded in an AB-ring remains phase-coherent, as indicated by the visibility of the AB-oscillations [28,29,31,32,42,46,47]. An expression for the transmission amplitude through the QD is $t = \sqrt{T_{\mathrm{QD}}} \mathrm{e}^{\mathrm{i}\alpha_{\mathrm{QD}}}$, which is a complex quantity. The phase, α_{QD}, is significant in its effect upon the AB oscillations for a QD embedded in an AB-ring, and has experimentally been seen to exhibit interesting phase-jumps of π as the transmission passes through a resonance. Theoretical analysis of these systems has provided some explanation for the phase behavior seen in experiments [2,3,4,5,6,10,12,13,19,20,21,22,23,24,25,26,27,37,38,39,40,41]. In a two-terminal device, the Onsager relations [35] of time-reversal symmetry and current conservation (unitarity) constrain the transmission phase to values of 0 or π. However, if the two-terminal AB-ring is "opened" by allowing current to flow out through additional terminals, the unitarity condition is broken and it becomes possible to extract meaningful phase information about the QD. Experiments with open rings demonstrate a gradual phase change across the transmission resonances. In particular, Schuster et al. produced a 4-terminal interferometer in a GaAs–AlGaAs heterostructure which showed smooth phase transitions [42].

In this review we shall present two typical approaches for investigation of Fano resonances in AB rings: i) The approach developed in quantum waveguides theory when the arms of the ring are considered as waveguides for electrons and dots are treated as a potential walls embedded in the waveguide; ii) Tight binding theory approach in which the dots are represented by sites which are connected by one-dimensional tight binding chains with different topology. Because both approaches give the same physical effects we shall focus mostly on physical meaning of obtained results. In Sect. "Fano Resonance in a Two Dimensional Electron Waveguide with a Quantum Dot" we give an introduction to Fano phenomena using a simple model of a waveguide with a short range potential well (QD). Applying waveguide theory to a quantum dot in AB rings in Sect. "Model for the Quantum Dots in the Ring: Theory of Waveguides Approach" we will demonstrate the manifestation of the Fano resonance in this system. In Sect. "Effects of Interference Interactions of Fano Resonances on an AB Ring" some novel results regarding Fano resonances in strong overlapping regime are presented.

It is shown that when the overlapping of two Fano resonances takes place in the transmission, two Fano dipoles in the complex-energy plane form a quasi-particle, which behaves as a coupled object – a *Fano quadrupole*. We show a periodic motion of the resonance pole and transmission zero in the complex-energy plane as a magnetic field through the AB ring is changed. In Sect. "Fano Resonance Induced by a Quantum Dot in an Open Three-terminal Interferometer", the properties of Fano resonances in an open AB ring are investigated in the framework of the tight binding model. We shall discuss in this section the phase measurements by using the modes of an open AB interferometer. Section "Future Directions" summarizes the properties of Fano phenomena and its manifestation in AB rings along with future directions.

Fano Resonance in a Two Dimensional Electron Waveguide with a Quantum Dot

In order to see a main feature of the Fano phenomenon associated with the propagating and evanescent waves in a quantum system, we study the propagation of the electron waves in an electronic 2D waveguide of width W arranged along the x-axis [24,30]. The waveguide geometry is schematically depicted in Fig. 1, showing a potential region and an attractive quantum dot (gray-colored area) in the waveguide. Here, the confining potential in the transverse direction is characterized by the function $V_c(y)$ and the attractive potential (dot) by the function $V(x, y)$. There is a complete basis of functions describing the transverse motion $\phi_n(y)$ of an electron with energies, $E_n = \frac{\hbar^2\pi^2 n^2}{2mW^2}$ (with the effective mass m). The electron waves in the perfect waveguide stretched to infinity are described by a combination of a plane wave along the longitudinal direction and confined wave functions in the transverse direction such as $e^{\pm ik_n x}\phi_n(y)$, where the wave vector along the x-direction is $k_n = \sqrt{2m(E - E_n)}/\hbar$, and n is the number of the transverse state. These propagating states can be considered as open channels in the waveguide.

In order to find the wave function of an electron in a waveguide with the dot, we solve the 2D Schrödinger equation

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\Psi(x, y) + V_c(y)\Psi(x, y) + V(x, y)\Psi(x, y) = E\Psi(x, y) \quad (2)$$

with plane wave boundary conditions in the leads ($x \to \pm\infty$). It is convenient to expand the wave function in the complete basis of functions describing the transverse motion:

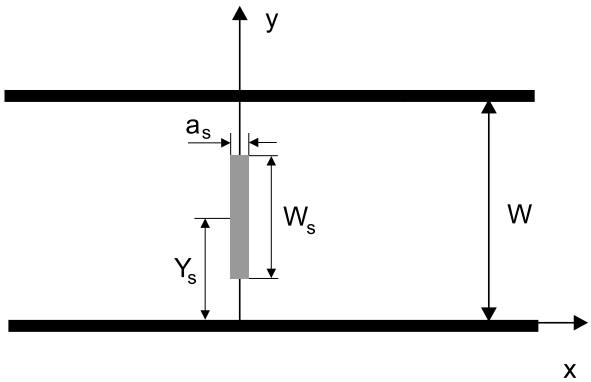

Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 1
Schematic illustration of the electron waveguide with embedded quantum dot (*gray-colored area*), where the attractive potential well is centered at $x = 0$ and $y = 0$ and the electron motion is not limited horizontally, $-\infty < x < \infty$, but is confined vertically, $0 < y < W$

$$\Psi(x, y) = \sum_{n=1}^{\infty} \psi_n(x)\phi_n(y) . \quad (3)$$

Substituting Eq. (3) into Eq. (2), we obtain the coupled-channel equations for an electron in the form

$$[b] - \frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\psi_n(x) + \sum_{n=1}^{\infty} V_{nn'}(x)\psi_{n'}(x) = (E - E_n)\psi_n(x) , \quad (4)$$

where the coupling matrix elements of the dot's potential (which still acts on the x-coordinate) are defined to be

$$V_{nn'}(x) = \int \phi_n(y)V(x, y)\phi_{n'}(y)\,\mathrm{d}y . \quad (5)$$

Since Eq. (4), which is equivalent to the 2D Schrödinger equation, cannot be solved in general, we use some simplification that allow us to use a resonant perturbation theory [15] in the system under the investigation.

We model the scattering potential as a thin rectangular potential-well by assuming that the longitudinal size of the potential well is much smaller than the characteristic wavelengths of the electron. Then, the matrix elements of the potential can be written as

$$V_{nn'}(x) = -\frac{\hbar^2}{m}v_{nn'}\delta(x) , \quad (6)$$

where the parameters $v_{nn'}$ of the dot are expressed in an explicit form [30]. It can be shown that the short range

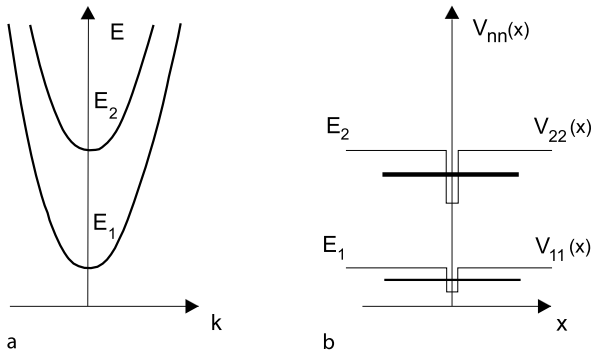

Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 2

a Energy dispersion relation for an electron in a perfect waveguide and b the diagrams for a bound level (near the first subband) and quasi-bound level (near the second subband) in the dot's effective potential

potential provides the following boundary conditions to be imposed on the multi-component wavefunctions at $x = 0$:

$$\psi_n(0+) = \psi_n(0-)\,,$$
$$\psi_n'(0+) - \psi_n'(0-) = -2\sum_{n'=1}^{\infty} v_{nn'}\psi_{n'}(0\pm)\,. \tag{7}$$

Here, we consider the situation when the energy of incoming electron is placed in the interval $E_1 \langle E \langle E_2$ (the first energy window), as shown schematically in Fig. 2. If the characteristic value of matrix element V_{12}, describing the coupling between two nearest channels, is small compared to the subband distance $(E_n - E_{n-1})$, then we only need to consider two coupled equations in the first energy window to understand the main physical features of the interference. It is well known that the remaining modes in the waveguide with the attractive impurity only alter the width and position of the resonances and hence play a minor role in Fano phenomenon. Without much difficulties our formulation can be extended for a multi-band approximation.

The wave function in the first channel, obtained from the solutions of the Schrödinger equation, can be written as

$$\psi_1(x) = \begin{cases} a_1 e^{ik_1 x} + b_1 e^{-ik_1 x}\,, & x < 0\,, \\ c_1 e^{ik_1 x}\,, & x > 0\,, \end{cases} \tag{8}$$

where $k_1 = \sqrt{2m(E - E_1)}/\hbar$ is a wave vector in the first channel. Similarly, the wavefunction in the second channel as

$$\psi_2(x) = \begin{cases} b_2 e^{|k_2|x}\,, & x < 0\,, \\ c_2 e^{-|k_2|x}\,, & x > 0\,, \end{cases} \tag{9}$$

where $|k_2| = \sqrt{2m(E_2 - E)}/\hbar$. Notice that the wave function, ψ_2, in the second channel is an evanescent wave. These two waves interfere in the waveguide and the quantum dot plays a role of a mixer of two different types of waves. The undetermined amplitudes appearing in Eqs. (8) and (9) are specified by applying the matching conditions given in Eq. (7). Consequently, we obtain

$$\begin{aligned} &(ik_1 + v_{11})c_1 + v_{12}c_2 = ik_1 a_1\,, \\ &v_{12}c_1 + (-|k_2| + v_{22})c_2 = 0\,, \end{aligned} \tag{10}$$

which give

$$c_1 = \frac{ik_1(-|k_2| + v_{22})}{(ik_1 + v_{11})(-|k_2| + v_{22}) - v_{12}^2} a_1\,, \tag{11}$$

$$c_2 = -\frac{ik_1 v_{12}}{(ik_1 + v_{11})(-|k_2| + v_{22}) - v_{12}^2} a_1\,. \tag{12}$$

From Eq. (11) the transmission and reflection amplitudes in the first channel are obtained as

$$t_{11} = \frac{c_1}{a_1} = \frac{ik_1(-|k_2| + v_{22})}{(ik_1 + v_{11})(-|k_2| + v_{22}) - v_{12}^2}\,, \tag{13}$$

and

$$r_{11} = \frac{b_1}{a_1} = \frac{-v_{11}(-|k_2| + v_{22}) + v_{12}^2}{(ik_1 + v_{11})(-|k_2| + v_{22}) - v_{12}^2}\,, \tag{14}$$

respectively. As it follows from Eq. (13), the transmission amplitude may vanish if $-|k_2| + v_{22} = 0$. When this happens, the reflection amplitude r_{11} is -1 and the energy at which the transmission becomes zero is determined to be

$$E_0 = E_2 - \frac{\hbar^2 v_{22}^2}{2m}\,. \tag{15}$$

There is a full reflection of the electron wave from a quantum dot when the electron energy is equal to the zero-energy E_0. The wavefunctions in the first and second channels are schematically depicted in Fig. 3 at the zero-energy. Like the classical system, we notice that the position of the amplitude-zero depends on the number of the channels. For instance, if we take into account another closed channel, $n = 3$, by a perturbation, the zero-energy given by Eq. (15) is shifted on the real axis of energy. In the meantime, there is a full transmission of the electron wave through the quantum dot when the reflection amplitude $r_{11} = 0$. If we impose $r_{11} = 0$ in Eq. (14), we get the condition for the reflection-zero, $v_{11}(-|k_2| + v_{22}) - v_{12}^2 = 0$. A real solution to this condition exists at the energy $E_{\max}$

$$E_{\max} = E_2 - \frac{\hbar^2}{2m}\left(v_{22} - \frac{v_{12}^2}{v_{11}}\right)^2\,. \tag{16}$$

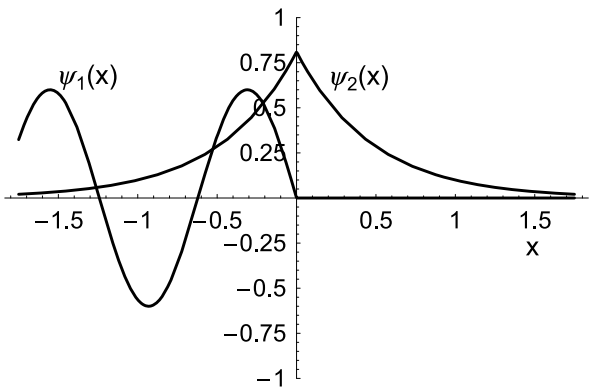

Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 3

The wave functions of the first and second channels when the electron energy matches with the zero-energy. Note that in this case a full reflection occurs

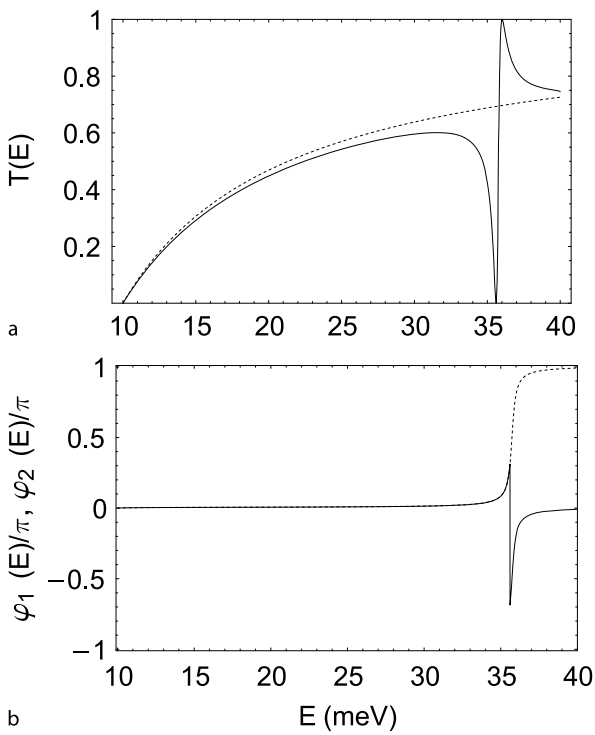

Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 4

a Fano resonance in the transmission for a quantum waveguide with a short-range attractive potential (*solid line*) and a background transmission (*dotted line*). b The phase shift of the amplitude for the electron wave in the first propagating channel (*solid line*) and the second evanescent channel (*dotted line*)

Note that the above expressions (Eqs. (13) and (14)) for the transmission-zero and the reflection-zero energies are exact in the framework of the two-channel approximation.

We have performed a numerical calculation of the transmission using the following parameters of the waveguide and a quantum dot. The width of the waveguide is set to $W = 23.7\,\text{nm}$ and the GaAs effective mass is used as $m = 0.067 m_0$. This gives $E_1 = 10\,\text{meV}$ and $E_2 = 40\,\text{meV}$ for the first two energy levels due to transverse confinement in the waveguide. The parameters of the quantum dot are as follows: $Y_s = 0.55\,\text{W}$ (the position of the dot in the waveguide), $W_s = 0.5\,\text{W}$ (W_s is the transverse width of the dot), and the scattering parameters $a_s V_s = 0.1\,\text{eVnm}$, where $V_s = 100\,\text{meV}$ (V_s is the depth of the attractive potential well) and $a_s = 1\,\text{nm}$ (a_s is the thickness of the potential well). The computed transmission of the system, $T(E) = |t_{11}(E)|^2$, is plotted in Fig. 4a for the chosen energy window, where for numerical purposes the following characteristic energies for the matrix elements of the potential are used: $\hbar^2 v_{11}^2/2m \cong 11.33\,\text{meV}$, $\hbar^2 v_{22}^2/2m \cong 4.40\,\text{meV}$, and $\hbar^2 v_{12}^2/2m \cong 0.34\,\text{meV}$. The pronounced Fano resonant structure (solid line) is clearly shown, i. e. the combined anti-resonance at $E_0 = 35.60\,\text{meV}$ and the nearby resonance peak at $E_\text{R} = 36.01\,\text{meV}$ where the width of resonance line is $\Gamma = 0.19\,\text{meV}$. Notice that if we put our quantum dot at the center of waveguide ($v_{12} = 0$), then the interference vanishes and the potential scattering takes place. In this case, only so-called the background profile in the transmission may be seen. This background transmission is also plotted for comparison in Fig. 4a as a dotted line. In Fig. 4b we also show the phase shift of the transmitted electron wave with respect to the incoming wave as a function of the electron energy for the first propagating channel (solid line) and the second evanescent channel (dotted line). One can see that the phase φ_1 in the propagating channel changes by π abruptly at the zero-energy and that it jumps up around the resonance peak, thus gaining essentially no net phase shift after passing through the zero-pole structure. On the other hand, the phase φ_2 of the evanescent channel changes by π rather smoothly over the anti-resonance and resonance structure.

To obtain a simple expression for the transmission amplitude near a zero-pole region, we consider the system in the weak coupling regime (i. e. v_{12} is assumed to be small in Eq. (13)). Expanding the numerator and denominator of Eq. (13) around the zero and the pole, respectively, one can write the transmission amplitude t_{11} in the desired form

$$t_{11}(E) \sim \frac{E - E_0}{E - E_\text{R} + i\Gamma}\,, \tag{17}$$

where E_R and Γ are the peak position and the width of

the resonance, and E_0 is the zero-energy of the resonance. After performing a perturbation approximation we find that the real part of the resonance pole can be written as $E_R = E_0 + \delta$, where $\delta = \hbar^2 v_{12}^2 v_{11} v_{22}/m(k_1^2 + v_{11}^2)$ and the width $\Gamma = \hbar^2 v_{12}^2 k_1 v_{22}/m(k_1^2 + v_{11}^2)$. (The energy appearing in k_1 is taken at $E = E_R$.) Here, we note that one can neglect the difference between E_{max} and E_R in the weak coupling limit. Furthermore, the expression for the transmission, Eq. (13), can be cast into the canonical Fano form of Eq. (1). The coupling parameter q ($q = v_{11}/k_1$ in our perturbation approximation) measures an asymmetry degree of Fano resonance line shape between the localized sates and the continuum states.

Model for the Quantum Dots in the Ring: Theory of Waveguides Approach

As mentioned in the Introduction, the AB ring plays the role of an interferometer for probing electron states on QD's. For this purpose, the QD's are placed in an arm of the interferometer, whereas the second arm is used as a reference path for the electron waves.

We shall start with a simple model where a QD is embedded in one of its arms, as schematically shown in Fig. 5. Here, the QDs can be formed by two electrodes which play the role of barriers for electrons in the lower arm. We consider both arms of the ring and the leads as perfect waveguides and adopt a single-propagating channel in the quasi-one-dimensional approximation (see the previous section). Propagating waves in the leads and perfect regions of the arms are assumed to be in the form $\psi(x, y) \propto e^{\pm ikx}\varphi(y)$, where x is the local coordinate along the waveguide, y is the transverse coordinate (transverse wave function is $\varphi(y)$), and wave vector $k = \sqrt{2mE}/\hbar$ of an electron with energy E in the open channel (we shift the origin of energy to the band edge of the first subband).

In order to connect incoming and outgoing waves at the junctions of the ring and the leads, we employ a simple junction model [11] where a scattering matrix describes the splitting of the electron wave functions at the junction. Using the amplitudes of electron waves in the ring where the relevant parameters are defined in Fig. 5, the electron transmission from left to right leads through both the upper and lower arms can be expressed by

$$\begin{pmatrix} \bar{A}_u \\ \bar{B}_u \end{pmatrix} = e^{-i\theta/2} M_u \begin{pmatrix} A_u \\ B_u \end{pmatrix} , \qquad \begin{pmatrix} \bar{A}_l \\ \bar{B}_l \end{pmatrix} = e^{-i\theta/2} M_l \begin{pmatrix} A_l \\ B_l \end{pmatrix} . \tag{18}$$

Here, $\theta = 2\pi\Phi/\Phi_0$ describes the phase shift introduced by the magnetic flux Φ threading the AB ring ($\Phi_0 = h/e$

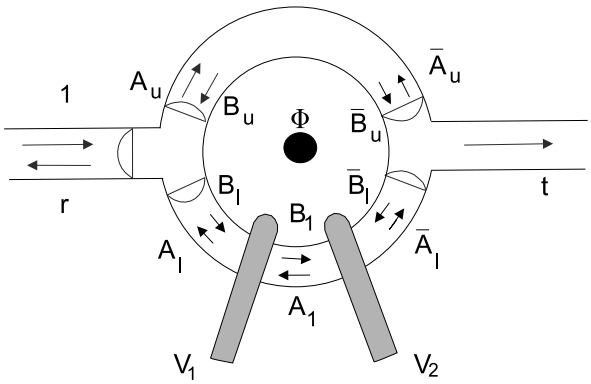

Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 5

Geometry of the AB ring with a single dot. Two electrodes in the lower arm produce the depletion regions in the two-dimensional electron gas and play the role of the barriers for electrons with potentials V_1 and V_2

is the elementary flux quantum), and M_u and M_l are the transfer matrices through the upper and lower arms, respectively. In order to find the transfer matrix M_l in the lower arm, we consider the QD that is formed by two short-range potential barriers (V_j, $j = 1, 2$). Then, the transfer matrix of each barrier has form

$$M_j = \begin{pmatrix} 1 - iu_j & -iu_j \\ iu_j & 1 + iu_j \end{pmatrix} , \tag{19}$$

where $u_j = mV_j/\hbar k$ with $j = 1, 2$. The dimensionless matrix element of the potential u_j describes the strength of the repulsive potential barrier. Therefore, the transfer matrix M_l for the lower arm with two QDs can be expressed by

$$M_l = X(L_l - L_2)M_2 X(L_2 - L_1)M_1 X(L_1) , \tag{20}$$

where $X(x) = \mathrm{diag}(e^{ikx}, e^{-ikx})$, L_l is the lower arm length, and the L_1 and L_2 are distances from the left junction to the first and second electrodes, respectively. On the other hand, the transfer matrix M_u for the upper (reference) arm has the simple form, $M_u = X(L_u)$ where L_u is the length of the upper arm.

Using the transfer-matrix and the junction-matrices [11], we can write connections between the amplitudes:

$$t = \sqrt{\varepsilon}(\bar{A}_u + \bar{A}_l) , \tag{21}$$

$$A_u = \sqrt{\varepsilon} + aB_u + bB_l , \qquad \bar{B}_u = a\bar{A}_u + b\bar{A}_l ,$$
$$A_l = \sqrt{\varepsilon} + bB_u + aB_l , \qquad \bar{B}_l = b\bar{A}_u + a\bar{A}_l . \tag{22}$$

Here, ε plays the role of a coupling parameter between the leads and ring, and the coefficients a and b are expressed

as a function of ε:

$$a = \frac{1}{2}\left(\sqrt{1-2\varepsilon}+1\right)$$

and

$$b = \frac{1}{2}\left(\sqrt{1-2\varepsilon}-1\right).$$

After some matrix manipulations, we obtain the full transmission amplitude $t(E,\Phi)$ analytically as

$$t(E,\Phi) = \frac{\mathrm{i}\varepsilon \mathrm{e}^{\mathrm{i}\theta} N(E,\Phi)}{D(E,\Phi)}. \tag{23}$$

Here, the numerator of Eq. (23) can be written as

$$N(E,\Phi) = \mathrm{e}^{\mathrm{i}(\gamma+2(\delta+\eta))} N_0(E,\Phi), \tag{24}$$

where

$$N_0(E,\Phi) = 4\mathrm{i}\{\sin\gamma + \mathrm{e}^{\mathrm{i}\theta/2}[4u\sin\delta(\cos 2\eta\sin\delta + (\cos\delta + u\sin\delta)\sin 2\eta) + \sin 2(\delta+\eta)]\}$$

with $u \equiv u_1 = u_2$, $\delta = kL_1$, $\eta = k(L_2 - L_1)$, $\gamma = kL_\mathrm{u}$. On the other hand, the denominator of Eq. (23) can be expressed by

$$\begin{aligned} D(E,\Phi) \equiv D_0(E,\Phi) &= \mathrm{e}^{\mathrm{i}(\gamma+2(\delta+\eta))}(1+e^{i\theta}) \\ &+ \mathrm{e}^{\mathrm{i}\theta/2}u^2(\mathrm{e}^{4\mathrm{i}\delta} + 4\mathrm{e}^{4\mathrm{i}\eta} + \mathrm{e}^{2\mathrm{i}(\gamma+2\eta)}) \\ &- 4\mathrm{e}^{2\mathrm{i}\delta}u(u-i) + 4(u-\mathrm{i})^2 \\ &- \mathrm{e}^{\mathrm{i}\theta/2}[\mathrm{e}^{2\mathrm{i}\gamma}(u-\mathrm{i})^2 + 4\mathrm{e}^{2\mathrm{i}(\delta+2\eta)}u(u+\mathrm{i}) \\ &\quad - \mathrm{e}^{4\mathrm{i}(\delta+\eta)}(u+\mathrm{i})^2]. \end{aligned}$$

From Eq. (23) we can determine the positions of transmission zeros and resonance poles by setting $N(E,\Phi) = 0$ and $D(E,\Phi) = 0$. We solve these two transcendental equations numerically by using standard routines. In addition, the total transmission probability through the ring as a function of the electron energy and magnetic flux is given by $T = |t(E,\Phi)|^2$ and the conductance is defined by $G = \frac{2e^2}{h}T$ according to the Landauer–Büttiker formalism.

In our calculations, we present results for the following parameters of the ring and dots: the effective electron mass is $m = 0.067m_0$ for GaAs and the geometrical parameters of the ring and positions of gates are chosen to be $L_\mathrm{u} = L_\mathrm{l} = 80\,\mathrm{nm}$, $L_1 = 10\,\mathrm{nm}$ and $L_2 = 70\,\mathrm{nm}$. The scattering parameters describing the QD in the lower arm are used for $a_s V_j = 0.2\,\mathrm{eVnm}$ ($j = 1, 2$), where the width and heights of the barriers are set to $a_s = 4\,\mathrm{nm}$ and $V_1 = V_3 = 50\,\mathrm{meV}$, respectively. Note, that the maximum coupling between the ring and the leads is used with $\varepsilon = \frac{1}{2}$ so that the coefficients of Eq. (23) become $a = -b = \frac{1}{2}$.

The calculated Fano structure is presented in Fig. 6. We show both the Fano resonance in the transmission as a function of electron energy (a) and a contour plot of transmission in the complex-energy plane (b) to illuminate zero-pole structure of the Fano resonance. We have investigated the energy interval near the second quasi-bound state in the two-barrier structure (two electrodes structure). For the chosen parameters of the system, the odd quasi-bound state (second level) has an energy of approximately $E \approx 5.2\,\mathrm{meV}$. In single channel approximation (for instance, when a one-dimensional system with a two-barrier structure is considered) the symmetric Breit–Wigner resonance will be observed in the transmission. Fano resonances in the transmission appear in Fig. 6a due to the quantum interference between the continuum states in the upper arm and the discrete states formed by the QD's states in the lower arm. The contribution of this isolated resonance is described by Eq. (1), with

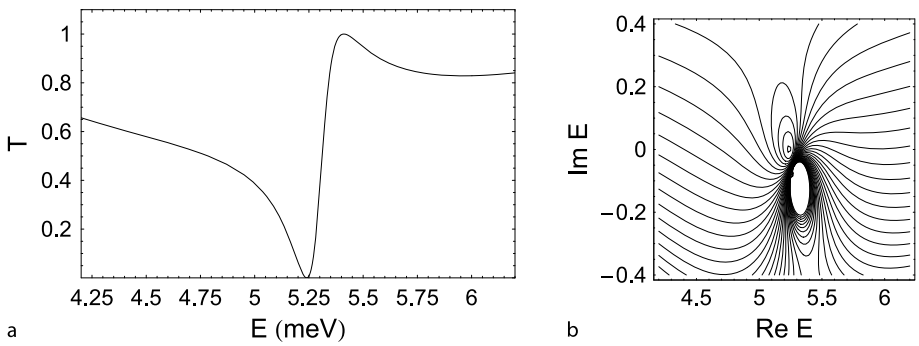

Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 6
The total transmission of a nanoscale AB ring as a function of electron energy (a) and contour plot of the transmission in the complex-energy plane (b). The Fano resonance is shown in a with pole $E_\mathrm{R} = (5.32 - \mathrm{i}0.14)$ meV, and corresponding zero $E_0 = 5.21$ meV

the corresponding parameters of the zero and the pole: $E_R = (5.32 - i0.14)$ meV and zero: $E_0 = 5.21$ meV.

Thus, the single dot produces a Fano resonance in the transmission when the position of the resonance is associated with the level of the dot. The profile depends on the coupling parameters of the dots with the ring.

Effects of Interference Interactions of Fano Resonances on an AB Ring

Next we investigate the effects of the interaction of Fano resonances in the transmission through an AB ring with coupled double QDs. When the overlapping of two Fano resonances takes place in the transmission, two Fano dipoles in the complex-energy plane form a quasi-particle, which behaves as a coupled object – a *Fano quadrupole*. In the regime of strong overlapping resonances, which can be tuned by the interaction parameter between two QDs, the collision of transmission zeros occurs and these zeros leave the real-energy axis and move away in opposite directions in the complex-energy plane. We also obtain an analytical expression of the transmission zeros, and show that these zeros are generally complex when two quasi-bound states lie close together in energy. To explain these effects, an analogous two level system was introduced to demonstrate the feature of interference attraction and repulsion of the zeros. Finally, we show a periodic motion of the resonance pole and transmission zero in the complex-energy plane as a magnetic field through the AB ring is changed.

The model we study is an AB ring where a coupled double QD is embedded in one of its arms, as schematically shown in Fig. 7. Here, double QDs can be formed by three electrodes which play the role of barriers for electrons in the lower arm, and the coupling between two QDs is controlled by the middle electrode.

As shown in the previous section, in order to find the transfer matrix M_l in the lower arm, we consider the coupled QDs that are formed by three short-range potential barriers (V_j, $j = 1, 2, 3$). Therefore, the transfer matrix M_l for the lower arm with two QDs can be expressed by

$$M_l = X(L_l - L_3)M_3X(L_3 - L_2)M_2X(L_2 - L_1)M_1X(L_1). \tag{25}$$

After simple calculation we again arrive at Eq. (23), where we now have an additional term, connected with the dots in the form

$$N(E, \Phi) = e^{i(\gamma+2(\delta+\eta))}(N_0(E, \Phi) + \xi N_1(E, \Phi)), \tag{26}$$

and

$$N_1(E, \Phi) = 8iue^{i\theta/2}(\cos\eta\sin\delta + (\cos\delta + 2u\sin\delta)\sin\eta),$$

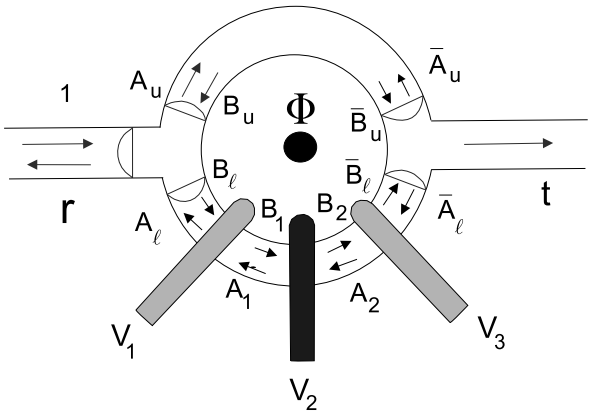

Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 7

Geometry of the AB ring with coupled QDs. Three electrodes in the lower arm produce the depletion regions in the two-dimensional electron gas and play the role of the barriers for electrons with potentials V_1, V_2 and V_3. The interaction between two QDs is controlled by the middle electrode marked by *black color*

and the interaction parameter between two QDs is $\xi(\equiv u_2/u_1 = V_2/V_1)$. The denominator of Eq. (23) can be expressed for two dots by

$$D(E, \Phi) = D_0(E, \Phi) + \xi D_1(E, \Phi), \tag{27}$$

where

$$\begin{aligned} D_1(E, \Phi) = {} & ie^{i(\theta/2+2\gamma)}u^3(e^{2i\eta} - 1)^2 \\ & - ie^{i(\theta/2-2(\delta+\eta))} \\ & \times \{\sin\delta[3i\cos\eta + \sin\eta(1 + 6iu) \\ & - \cos\delta(\cos\eta + (2u - i3)\sin\eta)^2]\}. \end{aligned}$$

First, we study the interaction of Fano resonances in the transmission through the AB ring in the absence of a magnetic field by investigating the behavior of the transmission amplitude for energy near zero-pole pairs. We consider two quasi-bound states in the double QDs, where even and odd quasi-bound states have the energies E_b and E_a, respectively. In Fig. 8, we show both overlapping of the Fano resonances in the transmission as a function of electron energy and contour plots of the transmission in the complex-energy plane for different values of the interaction parameter ξ. For weak coupling between two QDs ($\xi = 2.5$), two distinct Fano resonances in the transmission appear in Fig. 8a due to the quantum interference between the continuum states in the upper arm and two discrete states from the coupled QDs in the lower arm. The contribution of each separate resonance is described by Eq. (1) with the corresponding parameters of the zero

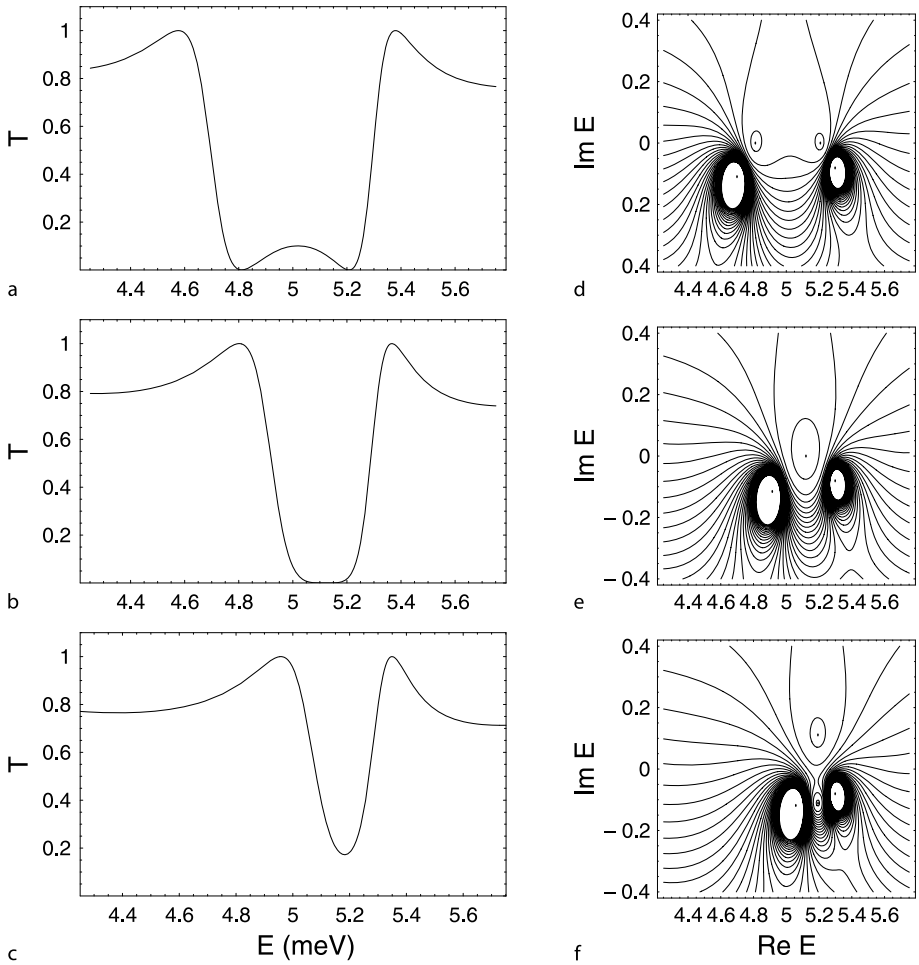

Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 8

The total transmission of a nanoscale AB ring as a function of electron energy (the *left column*) and contour plots of the transmission in the complex-energy plane (the *column*) with increasing of the interaction parameter ξ. The two distinct Fano resonances are shown in a and d for the weak overlapping regime ($\xi = 2.5$) with poles $E_b^R = (4.70 - i0.11)$ meV and $E_a^R = (5.30 - i0.08)$ meV (shown by *circles*), and corresponding zeros $E_b^0 = 4.81$ meV and $E_a^0 = 5.21$ meV (both of them are placed on real axis of energy. The collision of Fano resonances and merging of transmission zeros appear in b and e when $\xi = \xi_c = 3.77$, when the zeros have position $E_b^0 = E_a^0 = 5.12$ meV. The pole corresponding with the bounding level approaches to second pole with the increasing of ξ : $E_b^R = (4.91 - i0.12)$ meV and $E_a^R = (5.30 - i0.08)$ meV for $\xi = \xi_c$. In the strong overlapping regime of Fano resonances ($\xi = 5.5$), the transmission zeros move away in opposite directions from the real-energy axis in c and f: $E_+^0 = (5.19 + i0.11)$ meV and $E_-^0 = (5.19 - i0.11)$ meV (the poles are placed $E_b^R = (5.06 - i0.12)$ meV and $E_a^R = (5.30 - i0.08)$ meV)

and the pole. Notice that the width Γ_a of the Fano resonance through the odd quasi-bound state at $E \approx 5.3$ meV is less than the width Γ_b of Fano resonance through the even quasi-bound state at $E \approx 4.7$ meV ($\Gamma_a < \Gamma_b$). The two transmission zeros (E_a^0 and E_b^0) and two resonance poles ($\tilde{E}_a^R = E_a - i\Gamma_a$ and $\tilde{E}_b^R = E_b - i\Gamma_b$) in the complex-energy plane are seen in Fig. 8d, where the two zeros are separately on the real-energy axis. As ξ becomes a critical value $\xi_c = 3.77$, the two Fano resonances in the transmission merge (Fig. 8b) and the transmission zeros move toward each other and collide on the real-energy axis (Fig. 8e). When $\xi = 5.5 > \xi_c$, the minimum of the Fano resonance in the transmission does not reach to zero (Fig. 8c) and the transmission zeros leave the real-energy axis and move away in opposite directions in the complex-energy plane (Fig. 8f).

In order to see the behavior of transmission zeros and resonance poles in detail, we calculate the trajectories of the zeros and poles with the explicit expressions from the equations $N(E, \Phi) = 0$ and $D(E, \Phi) = 0$. In Fig. 9, we

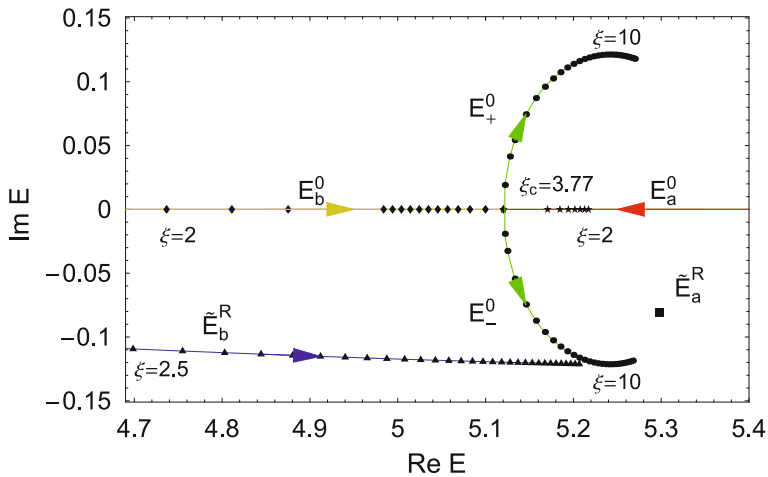

Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 9

The trajectories of resonance poles ($\tilde{E}_a^R$, $\tilde{E}_b^R$) and transmission zeros (E_a^0, E_b^0) in the complex-energy plane with increasing of the coupling parameter ξ ($2 < \xi < 10$). When ξ increases, two zeros E_a^0 (*red arrow*) and E_b^0 (*yellow arrow*) move toward each other and the collision and merging take place at $\xi_c = 3.77$. After merging, two zeros move away from the real-energy axis in opposite directions as complex conjugate pairs (denoted by E_+^0 and E_-^0 with *green arrows*). The resonance pole $\tilde{E}_b^R$ associated with even quasi-bound state moves to the higher energy (*blue arrow*), and $\tilde{E}_a^R$ arising from odd quasi-bound state is nearly motionless ($\tilde{E}_a^R \approx (5.30 - i0.08)$ meV)

present the trajectories of the zeros and poles in the complex-energy plane for a variation of the interaction parameter ξ. As ξ increases, one of Fano resonance zeros E_b^0 (shown as diamonds), arising from the even quasi-bound state in the QDs, moves to higher energy and another zero E_a^0 (shown as stars), arising from the odd quasi-bound state in the QDs, moves to lower energy. When $\xi = \xi_c$, the collision and merging of E_b^0 and E_a^0 takes place at $E_a^0 = E_b^0 = 5.12$ meV. In the strong overlapping regime of Fano resonances ($\xi > \xi_c$), the transmission zeros, denoted by E_+^0 and E_-^0 (shown as circles), move away from the real-energy axis in opposite directions as complex conjugate pairs.

It is interesting to note that the behavior of the resonance poles ($\tilde{E}_b^R$ and $\tilde{E}_a^R$) is different from that of the transmission zeros (E_b^0 and E_a^0). The resonance pole $\tilde{E}_b^R$ arising from the even quasi-bound state in the QDs is shifted to higher energy as ξ increases. However, the movement of $\tilde{E}_b^R$ to higher energy is hindered by one of the transmission zeros E_-^0 at $E \approx 5.24$ meV which prevents the collision of the resonance poles $\tilde{E}_b^R$ and $\tilde{E}_a^R$. On the other hand, the resonance pole $\tilde{E}_a^R$ arising from the odd quasi-bound state in the QDs is nearly motionless because the odd quasi-bound state of the double QDs is not strongly affected by ξ, due to short-range potentials in the lower arm of the ring.

In order to confirm the movement of the zeros of the transmission amplitude away from the real axis in the case

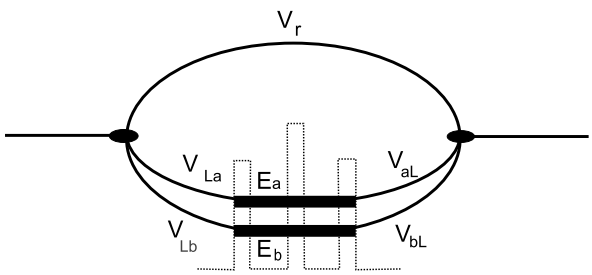

Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 10

A schematic diagram for even and odd quasi-bound states with energies E_b and E_a, connecting with the junction states of the ring by the appropriate matrix elements V_{Lb}, V_{bL}, V_{aL} and V_{La}. The matrix element of the junctions through the upper arm is denoted by V_r.

of overlapping Fano resonances, we calculate the transmission zeros analytically by employing a simple model shown in Fig. 10. We consider two nearest quasi-bound states in the dots which are even and odd levels with energies E_b and E_a, and bare eigenfunctions ψ_b^0 and ψ_a^0, respectively. These two orthogonal states of the dots in the lower arm can be connected with the junction states of the ring by the appropriate matrix elements V_{Lb}, V_{bL}, V_{aL} and V_{La}. On the other hand, V_r is defined as the matrix element of the junctions through the reference arm. Notice that these matrix elements are generally complex num-

bers and they are dependent on the phase difference between localized wave functions in the dots and propagating waves in the leads. If the energy of an incoming electron from the lead is near the resonant energies of the dots, then the matrix elements that couple an even (odd) quasi-bound state to the left and of the dots are real and have same (opposite) signs: $V_{\mathrm{Lb}} > 0$ and $V_{\mathrm{bL}} > 0$ ($V_{\mathrm{La}} > 0$ and $V_{\mathrm{aL}} < 0$) (this property has been noted in [7] where the electron tunneling through a semiconductor barrier with complex dispersion was investigated). We shall consider what happens with the two level system (see Fig. 10) when we couple it with the leads by effective matrix elements

$$U_{\mathrm{aa}} = \frac{V_{\mathrm{aL}} V_{\mathrm{La}}}{V_{\mathrm{r}}}\,, \quad U_{\mathrm{bb}} = \frac{V_{\mathrm{bL}} V_{\mathrm{Lb}}}{V_{\mathrm{r}}} \quad \text{and} \qquad U_{\mathrm{ab}} = U_{\mathrm{ba}} = \frac{V_{\mathrm{aL}} V_{\mathrm{Lb}}}{V_{\mathrm{r}}}\,. \tag{28}$$

Using the Schrödinger equation, it may be shown that the zeros are solutions of the eigenvalue problem

$$\begin{pmatrix} E_{\mathrm{a}} + U_{\mathrm{aa}} & U_{\mathrm{ab}} \\ U_{\mathrm{ba}} & E_{\mathrm{b}} + U_{\mathrm{bb}} \end{pmatrix} \begin{pmatrix} \psi_{\mathrm{a}} \\ \psi_{\mathrm{b}} \end{pmatrix} = E \begin{pmatrix} \psi_{\mathrm{a}} \\ \psi_{\mathrm{b}} \end{pmatrix}\,. \tag{29}$$

In the other words, the problem of finding the zeros from an exact expression for the scattering amplitude is equivalent to the problem of finding zeros from (29). This correspondence takes place because the interaction matrix obeys the property $U_{\mathrm{aa}} U_{\mathrm{bb}} - U_{\mathrm{ab}} U_{\mathrm{ba}} = 0$. The determinant of the system (29) is defined by

$$N(E) = (E - E_{\mathrm{b}})(E - E_{\mathrm{a}}) - (E - E_{\mathrm{b}}) U_{\mathrm{aa}} - (E - E_{\mathrm{a}}) U_{\mathrm{bb}}\,. \tag{30}$$

Then, the transmission zeros (E_{b}^0 and E_{a}^0) can be exactly obtained from the equation $N(E) = 0$, which gives

$$E_{a,b}^0 = \frac{1}{2}(\bar{E}_{\mathrm{a}} + \bar{E}_{\mathrm{b}}) \pm \sqrt{(\bar{E}_{\mathrm{a}} - \bar{E}_{\mathrm{b}})^2 + 4 U_{\mathrm{aa}} U_{\mathrm{bb}}}\,, \tag{31}$$

where $\bar{E}_{\mathrm{a}} = E_{\mathrm{a}} + U_{\mathrm{aa}}$, $\bar{E}_{\mathrm{b}} = E_{\mathrm{b}} + U_{\mathrm{bb}}$. Because $U_{\mathrm{bb}} > 0$ and $U_{\mathrm{aa}} < 0$ there is an effective interference attraction between levels ($\bar{E}_{\mathrm{a}} = E_{\mathrm{a}} - |U_{\mathrm{aa}}|$, $\bar{E}_{\mathrm{b}} = E_{\mathrm{b}} + U_{\mathrm{bb}}$). An examination of Eq. (31) indicates that when the distance between levels $\bar{E}_{\mathrm{a}} - \bar{E}_{\mathrm{b}}$ is less than the effective interaction, $2\sqrt{|U_{\mathrm{aa}}| U_{\mathrm{bb}}}$, in the other words when

$$E_{\mathrm{a}} - E_{\mathrm{b}} < U_{\mathrm{bb}} + |U_{\mathrm{aa}}| + 2\sqrt{|U_{\mathrm{aa}}| U_{\mathrm{bb}}}\,. \tag{32}$$

E_{b}^0 and E_{a}^0 are off from the real axis of energy. This implies that when even and odd quasi-bound states lie close together in energy, the transmission zeros move off from the real axis of energy. Using the location of these zeros and poles that are generally complex numbers, the transmission amplitude of Eq. (23) near the resonances in Fano overlapping regime can be expressed as

$$t_{\mathrm{F}}(E) \propto \frac{(E - E_{\mathrm{a}}^0)(E - E_{\mathrm{b}}^0)}{(E - E_{\mathrm{b}}^+ \mathrm{i}\Gamma_{\mathrm{b}})(E - E_{\mathrm{a}}^+ \mathrm{i}\Gamma_{\mathrm{a}})}\,, \tag{33}$$

which characterizes the transmission line shape in the vicinity of the transmission zeros and resonance poles.

Since the coupled Fano resonances in the transmission can be tuned by the magnetic flux Φ threading the AB ring, we study the magnetic field dependence of the transmission zeros and resonance poles in the strong Fano overlapping regime. The trajectories of the zeros and poles in the complex-energy plane as a function of magnetic flux are shown in Fig. 11 for a given interaction parameter $\xi = 3.9$. As Φ increases, the two zeros at $E_+^0 = (5.19 + \mathrm{i}0.11)$ meV and $E_-^0 = (5.19 - \mathrm{i}0.11)$ meV for $\Phi = 0$ start to move to the higher (green arrow with circles) and lower (red arrow with stars) energies, respectively until they reach to the real axis. When $\Phi = \frac{1}{2}\Phi_0$, these complex conjugate pairs become real numbers ($E_+^0 = 5.45$ meV and $E_-^0 = 4.95$ meV) which indicates full reflection of the transmission. When $\frac{1}{2}\Phi_0 < \Phi < \Phi_0$, these zeros continue to move on the other half of the trajectories to complete a stadium-like orbit. Then, the zero E_+^0 comes to initial position of E_-^0 and vice versa for $\Phi = \Phi_0$. In other words, the upper zero E_+^0 circumscribes half of the orbit and then the lower zero E_-^0 replaces the same part of the orbit after the flux is changed by one period. On the other hand, the behavior of the resonance poles is quite different. As Φ increases, each pole at $E_{\mathrm{b}}^{\mathrm{R}} = (5.06 - \mathrm{i}0.12)$ meV (shown as triangles) and $E_{\mathrm{a}}^{\mathrm{R}} = (5.30 - \mathrm{i}0.08)$ meV (shown as diamonds) for $\Phi = 0$ moves a short nearly-straight line periodically. (If $\Phi = \frac{1}{2}\Phi_0$, these two poles are located at the opposite end of the starting points along these lines.)

It is worthwhile noting here that the transmission through the ring is a periodical function of magnetic flux, which changes the interference between the parts of the wave function in the arms. Hence, the transmission zeros are more sensitive to the magnetic flux because they are defined by wave interference. However, the resonance poles are defined by quasi-bound energy and move slowly in a magnetic field.

Finally, we present an analytical expression for zero trajectories in the presence of magnetic field using the same model shown in Fig. 10. If a magnetic flux is applied to the loop in the perpendicular direction, each wave function acquires a phase shift on the links. The phase shift on the links connecting the dots with the leads is $\phi/2$ so that the matrix elements between the dots and the leads are

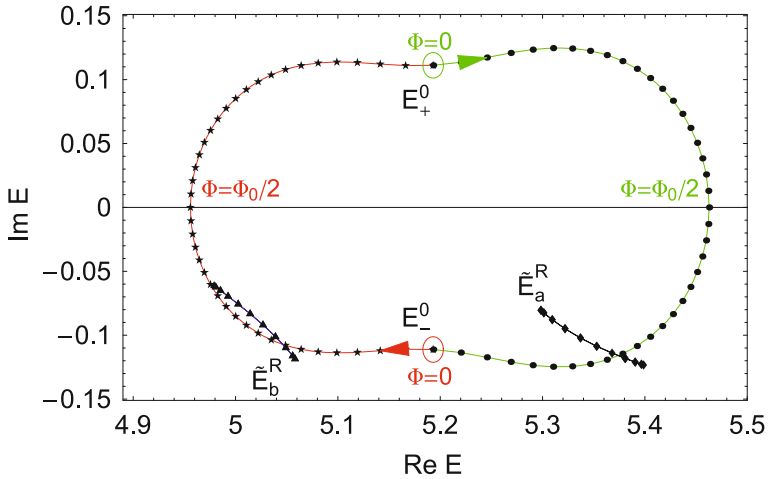

Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 11

The trajectories of zeros and poles as a function of a magnetic flux for a fixed interaction parameter $\xi = 3.9$. The two zeros for $\phi = 0$ at $E_+^0 = (5.19 + \mathrm{i}0.11)$ meV and $E_-^0 = (5.19 - \mathrm{i}0.11)$ meV start to move to the (*green arrow*) and left (*red arrow*), respectively, and they reach to the real axis for that is a full reflection of the transmission. These zeros continue to move on the other half of the trajectories and complete a figure-of-eight orbit for $\frac{1}{2}\Phi_0 < \Phi < \Phi_0$. The resonance poles at $E_b^R = (5.06 - \mathrm{i}0.12)$ meV and $E_a^R = (5.30 - \mathrm{i}0.08)$ meV for $\Phi = 0$ move a short nearly-straight line periodically

replaced by $V_{n,m} \to V_{n,m}\mathrm{e}^{\pm \mathrm{i}\phi/2}$, where the sign depends on the propagation direction of an electron. On the other hand, the phase shift through the reference arm is ϕ and so we can set $V_\mathrm{r} \to V_\mathrm{r}\mathrm{e}^{-\mathrm{i}\phi}$. Then, the effective matrix elements of the two level system in a magnetic field, Eq. (29), can be written as $U_{\mathrm{aa}} \to U_{\mathrm{aa}}\mathrm{e}^{\mathrm{i}2\phi}$ and $U_{\mathrm{bb}} \to U_{\mathrm{bb}}\mathrm{e}^{\mathrm{i}2\phi}$. The analytical expression for the transmission zeros in the presence of a magnetic field is

$$[b]E_{\mathrm{a,b}}^0 = \frac{1}{2}\Big(E_\mathrm{a} + E_\mathrm{b} + (U_{\mathrm{aa}} + U_{\mathrm{bb}})\mathrm{e}^{\mathrm{i}2\phi} \pm \sqrt{(E_\mathrm{a} - E_\mathrm{b} + (U_{\mathrm{aa}} - U_{\mathrm{bb}})\mathrm{e}^{\mathrm{i}2\phi})^2 + 4U_{\mathrm{aa}}U_{\mathrm{bb}}\mathrm{e}^{\mathrm{i}4\phi}}\Big). \tag{34}$$

This expression indicates the periodic trajectories for the transmission zeros as Φ changes, which is shown in Fig. 11. The analysis of the two level model gives a simple explanation for why the zeros return back to the real axis when the flux equals $\Phi = \frac{1}{2}\Phi_0$. Since for this value of flux the signs of matrix elements are reversed: $U_{\mathrm{bb}} < 0$ and $U_{\mathrm{aa}} > 0$, the levels are defined now by $\bar{E}_\mathrm{a} = E_\mathrm{a} + |U_{\mathrm{aa}}|$, and $\bar{E}_\mathrm{b} = E_\mathrm{b} - U_{\mathrm{bb}}$. This means that the phase shift results in an effective repulsion of the levels and we obtain an inequality $2\sqrt{|U_{\mathrm{aa}}|U_{\mathrm{bb}}} < \bar{E}_\mathrm{a} - \bar{E}_\mathrm{b}$, which is opposite to (32) (notice that the parameter $2\sqrt{|U_{\mathrm{aa}}|U_{\mathrm{bb}}}$ has not changed).

We note that for the non-overlapping regime when Fano dipoles exist, the zeros move in separate circular orbits near their associated poles (not shown here). In the strong overlapping regime, however, the zeros act as a coupled object and move in a common *figure-of-eight* orbit around the two coupled poles. This means that two overlapped Fano resonances can be considered as a combined object called a *Fano quadrupole*.

Fano Resonance Induced by a Quantum Dot in an Open Three-Terminal Interferometer

The properties of Fano resonances may be easily demonstrated by the use of simple lattice models in which the dots are represented by sites which are connected by one-dimensional tight-binding chains with different topology. In fact, most of properties of the resonances in the AB ring are the same as in waveguide model. We shall discuss here the phase measurements by using the tight binding model [2,3,4,5,6,13,19,22,41,46].

In this section, we analyze a three-terminal interferometer with an embedded QD in one arm of the AB ring. The general structure studied in our calculations is a three-terminal AB ring with magnetic flux through the ring and an embedded QD, sketched schematically in Fig. 12 where relevant parameters are defined. By discretizing the system spatially with lattice constant a, and denoting the wave function on site n by ψ_n, the Schrödinger equation in the tight-binding approximation can be written as

$$-\sum V_{n,m}\psi_m + \varepsilon_n\psi_n = E\psi_n\,. \tag{35}$$

Here, the sum runs over the nearest neighbors of n, E is the electron energy, and ε_n is the site energy. (In our cal-

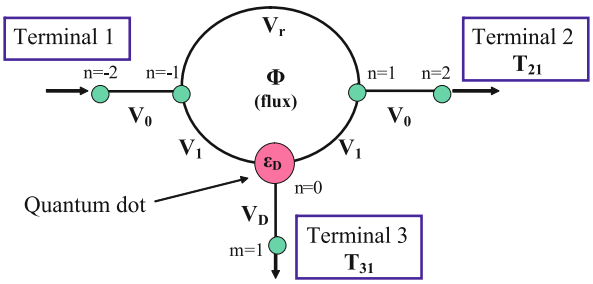

Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 12

Schematic of the three-terminal interferometer with a QD embedded in one of the arms. In addition to the magnetic flux Φ threading the AB ring, the relevant coupling parameters between sites are defined: the confinement V_1 of the QD, coupling V_D to the third terminal, and coupling V_r through the reference arm of the ring.

culations, the site energies ε_n are set to zero for all sites except for the QD at $n = 0$ which has site energy ε_D.) The parameters $V_{n,m}$ are overlap integrals (or coupling parameters) involving the overlap of the single site, atomic-like wave functions from sites m and n with the single-site potential of site n. In the homogeneous leads, the coupling parameters are all set to $V_0 = 1.0$, which we use throughout the discussion as a unit of energy. In the presence of the magnetic flux Φ, a phase difference between the path through the QD and the path through the reference arm is produced [1]. Therefore, we choose a gauge in which the coupling parameter for each segment of the lower arm is modified as $V_1 \to V_1 e^{\pm i\varphi}$, and the reference arm coupling parameter becomes $V_r e^{\pm 2i\varphi}$ ("+" for counter-clockwise transits around the ring and "−" for clockwise transits). The phase φ is related to the magnetic flux Φ by $2\varphi = \pi\Phi/\Phi_0$.

Let us consider an incoming wave function only from terminal 1, with transmitted waves through the ring into the other terminals:

$$\begin{aligned} \psi_n &= e^{in\theta} + r_{11} e^{-in\theta}\,, & n &\le -1\,,\\ \psi_n &= t_{21} e^{in\theta}\,, & n &\ge 1\,,\\ \psi_m &= t_{31} e^{im\theta}\,, & m &\ge 1\,, \end{aligned} \tag{36}$$

with $\theta = ka$. Here, k is the wave vector that is connected with the energy by the dispersion relation: $E = -2V_0 \cos ka$, t_{21} and t_{31} are the transmission amplitudes from terminal 1 into terminal 2 and 3 respectively; and r_{11} is the reflection amplitude back to terminal 1. Applying the Schrödinger equation to the three sites around the AB ring and also to site $m = 1$ of the third terminal, we obtain the following matrix equation for the complex transmission amplitudes:

$$\begin{pmatrix} V_0 & -V_r e^{-i(2\varphi-\theta)} & -V_1 V_0 e^{i\varphi}/V_D \\ -V_r e^{i(2\varphi+\theta)} & V_0 & -V_1 V_0 e^{-i\varphi}/V_D \\ -V_1 e^{-i(\varphi-\theta)} & -V_1 e^{i(\varphi+\theta)} & -(V_D e^{i\theta} + (E-\varepsilon_D)V_0/V_D) \end{pmatrix} \times \begin{pmatrix} r_{11} \\ t_{21} \\ t_{31} \end{pmatrix} = \begin{pmatrix} -V_0 \\ V_r e^{i(2\varphi-\theta)} \\ V_1 e^{-i(\varphi+\theta)} \end{pmatrix}. \tag{37}$$

Inverting the matrix on the left side of Eq. (37), we can find the unknown reflection and transmission amplitudes: r_{11}, t_{21} and t_{31}. In the same way, we can find the other elements of the scattering matrix: r_{22}, t_{12}, t_{32} and r_{33}, t_{13}, t_{23} when incoming waves are chosen from terminals 2 and 3, respectively. The transmission amplitudes for an electron from terminal (channel) j into terminal (channel) i may be written in the form

$$t_{ij}(\Phi) = \frac{N_{ij}(\Phi)}{D(\Phi)}\,, \tag{38}$$

where we have

$$N_{21}(\Phi) = 2iV_0 \sin\theta e^{2i\varphi}\big[V_r\big(V_0(E-\varepsilon_D) + e^{i\theta}V_D^2\big) - e^{-4i\varphi}V_0 V_1^2\big]\,, \tag{39}$$

$$N_{31}(\Phi) = -2iV_0 V_1 V_D \sin\theta e^{-i\varphi}\big[V_0 + e^{i(4\varphi+\theta)}V_r\big]\,, \tag{40}$$

$$\begin{aligned} D(\Phi) =& e^{i\theta}V_0^2\big(2V_1^2 + V_D^2\big) - e^{3i\theta}V_r^2 V_D^2 \\ &+ V_0\big(V_0^2 - e^{2i\theta}V_r^2\big)(E-\varepsilon_D) \\ &+ 2e^{2i\theta}V_0 V_r V \cos(4\varphi)\,, \end{aligned} \tag{41}$$

with symmetry conditions

$$\begin{aligned} &D(\Phi) = D(-\Phi)\,, \quad N_{ij}(\Phi) = N_{ji}(-\Phi)\,, \quad \text{and} \\ &N_{13}(\Phi) = N_{23}(\Phi)\,. \end{aligned} \tag{42}$$

Thus, we have all the transmission coefficients $T_{ij}(\Phi) = |t_{ji}(\Phi)|^2$, which obey the property

$$T_{ij}(\Phi) = T_{ji}(-\Phi)\,. \tag{43}$$

In order to find the non-local conductance of the open ring, we use the Büttiker equations [9]

$$I_i = \frac{2e}{h}\Big[(1-R_{ii})\mu_i - \sum_{j\neq i} T_{ij}\mu_j\Big]\,, \quad (i,j = 1,2,3)\,, \tag{44}$$

where $R_{ij}(\Phi) = |r_{ji}(\Phi)|^2$ are the reflection coefficients (which may be eliminated from the set of equations by using current conservation: $1 - R_{ii} = \sum_{j\neq i} T_{ij}$), and μ_i are

the chemical potentials of the reservoirs (terminals). The factor of two in Eq. (44) stems from the identical contribution of both electron spin states. Here, it is noteworthy that we consider a typical situation in which two terminals (1 and 2) are used for injection of current and measurements of the conductance $G_{12,12}$ (see Büttiker's notations in Ref. [9]), whereas the potential drop (which is characterized by the resistance $R_{12,13}$) is measured only between terminals 1 and 3. For our purpose, we set the current between terminals 1 and 2 as: $I \equiv I_1 = -I_2$. Thus, terminal 3 represents an ideal probe that draws no net current ($I_3 = 0$). Solving the set of Eq. (44), we find the coefficient

$$G_{12,12} = \frac{2e^2}{h}\left(T_{21} + \frac{T_{23}T_{31}}{T_{31}+T_{32}}\right) \tag{45}$$

between the current I and the bias $U_{12} = (\mu_1 - \mu_2)/e$. The potential drop U_{13} between the terminals 1 and 3 is defined by the resistance $R_{12,13} : U_{13} = R_{12,13}I$, where

$$R_{12,13} = \frac{h}{2e^2}\left(\frac{T_{32}}{T_{21}T_{31} + T_{23}T_{31} + T_{32}T_{21}}\right) . \tag{46}$$

Here, we present results for the following parameters of the system: $V_1 = 0.3$ (QD confinement), $V_r = 0.3$ (the coupling through the reference arm of the AB ring), and $\varepsilon_D = 0$ (the site energy of the QD which positions the resonance in the center of the allowed energy band).

Now, we study the effect of coupling to the third output terminal on the transmission $T_{21} = |t_{21}|^2$ through the AB ring in the absence of the magnetic flux. Here, the transmission amplitute t_{21} can be calculated from Eq. (38) as

$$t_{21} = \frac{2\mathrm{i}\sin\theta\left[V_1^2 - \mathrm{e}^{4\mathrm{i}\varphi}V_r\left(E - \varepsilon_D + \mathrm{e}^{\mathrm{i}\theta}V_D^2/V_0\right)\right]}{\mathrm{e}^{2\mathrm{i}\theta}V_r V_1^2\left(\mathrm{e}^{6\mathrm{i}\varphi} + \mathrm{e}^{-2\mathrm{i}\varphi}\right)/V_0 + \mathrm{e}^{2\mathrm{i}\varphi}\left[\left(2V_1^2 + V_D^2\right)\mathrm{e}^{\mathrm{i}\theta} - \mathrm{e}^{3\mathrm{i}\theta}V_r^2V_D^2/V_0^2 + (E-\varepsilon_D)\left(V_0 - \mathrm{e}^{2\mathrm{i}\theta}V_r^2/V_0\right)\right]} \tag{47}$$

The behavior of the transmission zero and phase in a three-terminal interferometer as a function of energy for various values of coupling to the third terminal ($V_D = 0.1$ (solid), $V_D = 0.3$ (dotted), and $V_D = 0.6$ (dashed)) is shown in Fig. 13. Unlike a two-terminal closed AB ring with QD's ($V_D = 0$) [22,41], it is seen from Fig. 13a that in an open ring ($V_D \neq 0$) the Fano resonance peak does not reach unity because of energy loss due to the outgoing electrons into the third terminal. In addition, as V_D increases, the Fano zero (obtained from $N_{21} = 0$ in Eq. (39)) shifts progressively further off the real-energy axis into the complex-energy half-plane. The Fano zero can be returned to the real energy axis at discrete values of magnetic flux (see below) as long as V_D is less than a critical value. This flexible control over the transmission resonance features is unavailable in a closed, two-terminal interferometer.

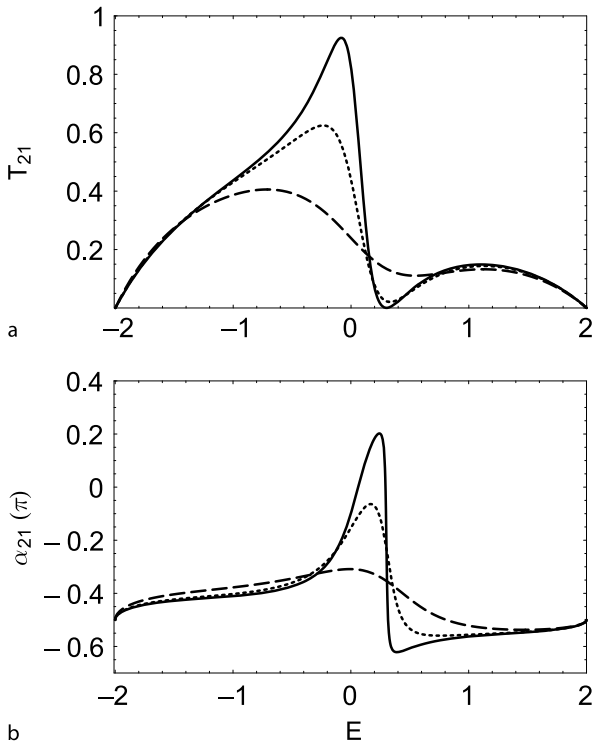

Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 13

Transmission T_{21} and transmission phase α_{21} as a function of energy for different values of $V_D = 0.1$ (*solid curve*), $V_D = 0.3$ (*dotted curve*), and $V_D = 0.6$ (*dashed curve*). a As V_D increases, the Fano resonance no longer reaches unity and the Fano zero lifts off the real-energy axis. b The phase jump of π at the transmission resonance diminishes and softens as the ring is opened with coupling to the third terminal

In Fig. 13b, we show the transmission phase as a function of energy for different values of V_D. The transmission phase α_{21}, which can be calculated from Eq. (47) as $\alpha_{21} = \tan^{-1}[\mathrm{Im}(t_{21})/\mathrm{Re}(t_{21})]$, no longer changes abruptly by π at the resonance, as for non-zero V_D, but is shown to progressively soften and to smoothly change by less than π as V_D increases. We attribute this smearing of the abrupt phase jump of π to the fact that current, which flows to the third terminal, breaks unitarity and disrupts the interference effects due to repeated reflections of the electrons from the junctions and back through the ring.

In order to illustrate the transmission phase changing from an abrupt jump to a smooth transition of less than π

in a three-terminal AB ring, we calculate the phase using a simple analytical model which is based on the properties of the Fano resonance. In the vicinity of the Fano resonance in the transmission versus electron energy for an AB ring with an embedded QD, the transmission amplitude has the form [17]

$$t_{21} \cong t_{\mathrm{bg}} \left(\frac{E - \tilde{E}_0}{E - E_{\mathrm{R}} + i\Gamma} \right) . \tag{48}$$

Here, $\tilde{E}_0$ is the position of the transmission zero and E_{R} gives the energy of the pole. The width of the resonance Γ indicates how far the pole is off the real-energy axis, and t_{bg} represents any background contribution to the amplitude. Because the Fano transmission zero lies off the real-energy axis for an open ring, we can write $\tilde{E}_0$ in a complex form: $\tilde{E}_0 = E_0 - \mathrm{i}\gamma$. The transmission amplitude can now be written as the product of two complex terms:

$$t_{21} \cong t_{\mathrm{bg}} \frac{(E - E_0 + \mathrm{i}\gamma)(E - E_{\mathrm{R}} - \mathrm{i}\Gamma)}{(E - E_{\mathrm{R}})^2 + \Gamma^2} = |t_{21}|^2 \mathrm{e}^{\mathrm{i}\theta_{\mathrm{tot}}}, \tag{49}$$

where θ_{tot} is the combined transmission phase from the two complex terms. By separating Eq. (49) into its real and imaginary parts, we obtain an expression for θ_{tot} as

$$\theta_{\mathrm{tot}} = \arctan \left[\frac{(E - E_{\mathrm{R}})\gamma - (E - E_0)\Gamma}{(E - E_{\mathrm{R}})(E - E_0) + \gamma\Gamma} \right] . \tag{50}$$

In Eq. (50) for θ_{tot}, there is no sharp phase jump at a particular value of energy, as exists at $E = E_{\mathrm{R}}$ for a two-terminal closed ring in which the Fano zero is on the real-energy axis ($\gamma = 0$). In Fig. 14, we show plots of θ_{tot} versus E for various coupling parameters V_{D} to the third terminal. It is clearly seen that as V_{D} is increased from zero, the abrupt phase jump of π at the resonance softens and diminishes in magnitude. This indicates that the Onsager relations (unitarity and time reversal-symmetry) are not valid for an open AB interferometer with an embedded QD [2,13].

Since the Fano zero and resonance pole in the transmission can be tuned by the magnetic flux Φ threading the AB ring, we investigate the magnetic flux dependence of transmission T_{21} for a fixed V_{D}. In Fig. 15, the total transmission as a function of electron energy (the left column) and contour plots of the transmission amplitude in the complex-energy plane (the column) with fixed $V_{\mathrm{D}} = 0.3$ are shown for different magnetic flux values: $\Phi/\Phi_0 = 0.0, 0.25, 0.548, 0.75$, and 0.952 (top to bottom). As the magnetic flux is increased from $\Phi = 0$, the Fano zero begins to move on a clockwise orbit around the Fano pole. At $\Phi/\Phi_0 = 0.25$, the zero is positioned directly below the Fano pole in the complex-energy plane.

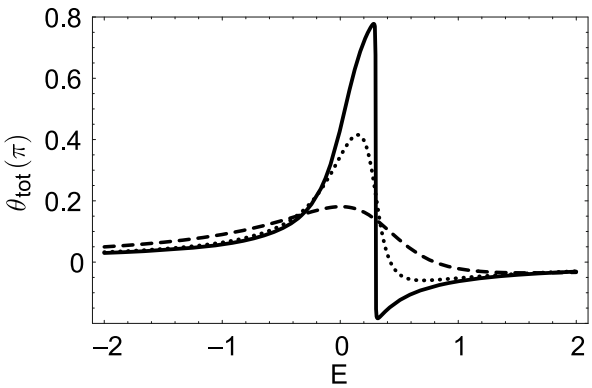

Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 14

Modeled transmission phase θ_{tot} versus energy E for a standard Fano resonance (*solid curve*: $E_0 = 0.3$, $\gamma = 0.0005$, $E_{\mathrm{R}} = 0.04$, $\Gamma = 0.192$), and for modified Fano resonances (*dotted curve*: $E_0 = 0.3$, $\gamma = 0.1$, $E_{\mathrm{R}} = 0.05$, $\Gamma = 0.3$; *dashed curve*: $E_0 = 0.38$, $\gamma = 0.45$, $E_{\mathrm{R}} = 0.1$, $\Gamma = 0.75$) based on the approximate positions of the Fano zeros and poles for $V_{\mathrm{D}} = 0.01, 0.3, 0.6$, respectively. The phase jump of π at the transmission resonance diminishes and softens as V_{D} increases

As Φ continues to increase, the Fano zero moves back up towards the real-energy axis and *crosses* the axis at $\Phi/\Phi_0 = 0.548$. When $\Phi/\Phi_0 = 0.75$, the Fano zero arrives directly above the Fano pole, attaining its most positive imaginary value. As the flux is further increased towards $\Phi/\Phi_0 = 1.0$, the Fano zero again *crosses* the real-energy axis at $\Phi/\Phi_0 = 0.952$ on the way back to its position from which it started at $\Phi/\Phi_0 = 0$.

It is interesting to note from Fig. 15 that for a fixed value of V_{D}, there exist two values of magnetic flux for which the Fano zero *crosses* the real-energy axis. By setting $V_{\mathrm{r}}(V_0(E - \varepsilon_{\mathrm{D}}) + \mathrm{e}^{\mathrm{i}\theta} V_{\mathrm{D}}^2) - \mathrm{e}^{-4\mathrm{i}\varphi} V_0 V_{\mathrm{I}}^2 = 0$ from Eq. (39), the analytical expression for the energy values of the Fano zeros (E_0) and the corresponding normalized magnetic flux values (Φ/Φ_0) in terms of the coupling parameter V_{D} can be obtained as

$$E_0 = \pm \sqrt{\frac{\left(V_{\mathrm{I}}^2 V_0/V_{\mathrm{r}}\right)^2 - V_{\mathrm{D}}^4}{V_0^2 - V_{\mathrm{D}}^2}} \quad \text{and}$$

$$\cos\left[2\pi\left(1 - \frac{\Phi}{\Phi_0}\right)\right] = \pm \sqrt{\frac{1 - \left(V_{\mathrm{r}} V_{\mathrm{D}}^2/V_0 V_{\mathrm{I}}^2\right)^2}{1 - V_{\mathrm{D}}^4/\left(2V_0^2 - V_{\mathrm{D}}^2\right)^2}} . \tag{51}$$

Here, it is required that both the real and imaginary parts of N_{21} from Eq. (39) be zero. For the parameters used in Fig. 15 ($V_0 = 1.0$, $V_{\mathrm{I}} = V_{\mathrm{r}} = 0.3$, and $V_{\mathrm{D}} = 0.3$), the two Fano zeros are at $E_0 = -0.30$ and 0.30 when

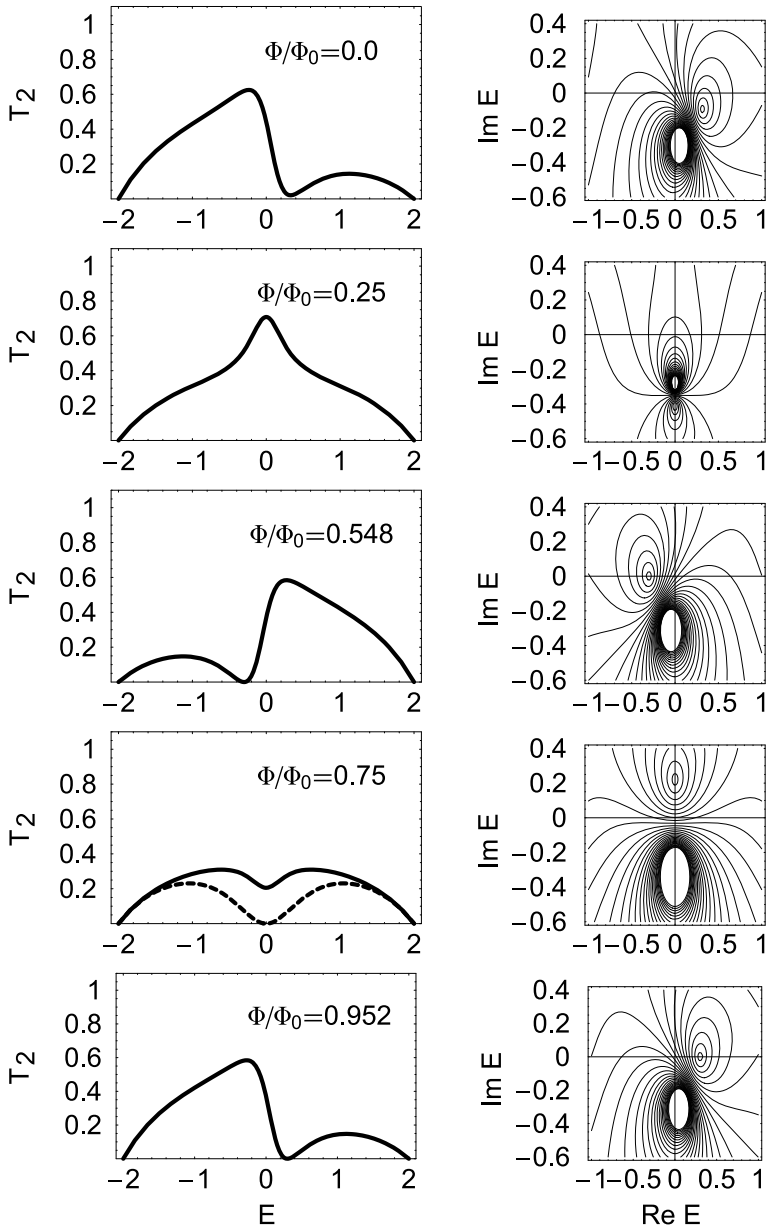

Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 15

The total transmission as a function of electron energy (the *left column*) and contour plots of the transmission amplitude in the complex-energy plane (the *column*) with fixed $V_D = 0.3$, for different magnetic flux $\Phi/\Phi_0 = 0.0, 0.25, 0.548, 0.75$, and 0.952 (*top to bottom*). The Fano zero moves directly downward and *crosses* the real-energy axis at $V_D = V_D^{crit}$, shown in the *dashed curve* for $\Phi/\Phi_0 = 0.75$

$\Phi/\Phi_0 = 0.548$ and 0.952, respectively. Notice, however, that there is a critical value of V_D^{crit} [$V_D^{crit} = V_1\sqrt{V_0/V_r}$, obtained from requiring E_0 to be real in Eq. (51)], which is the *maximum* value of V_D for which there is the possibility of placing the Fano zero on the real-energy axis at any value of flux. We show in Fig. 15 that the Fano zero moves directly downward and *crosses* the real-energy axis at V_D^{crit} (see the dashed transmission curve for $\Phi/\Phi_0 = 0.75$). When $V_D > V_D^{crit}$, the Fano zero passes into the negative complex-energy half-plane and there is no value of flux which can bring the Fano zero back to the real-energy axis.

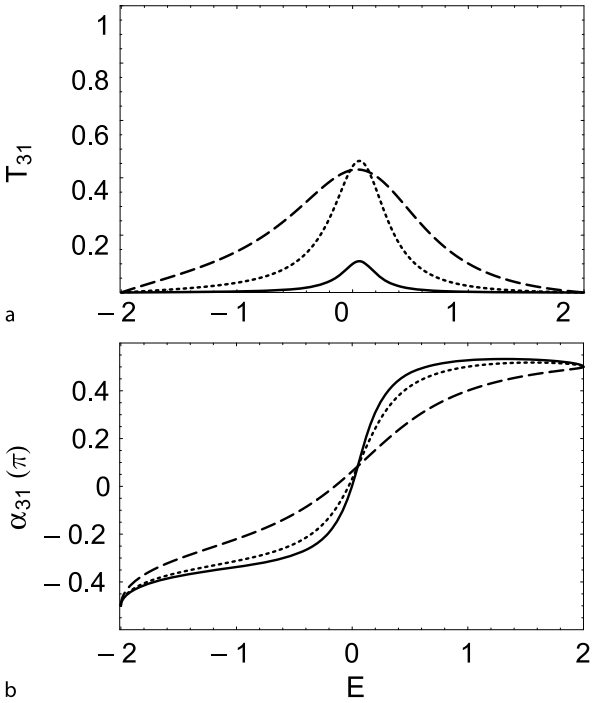

Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 16

Transmission T_{31} and transmission phase α_{31} as a function of energy for different values of $V_D = 0.1$ (*solid curve*), $V_D = 0.3$ (*dotted curve*), and $V_D = 0.6$ (*dashed curve*). a The BW resonances, which arise from the fact that the amplitude t_{31} does not have zeros in the energy plane, are seen for a variation of V_D. b The phase change near the BW resonance softens as V_D increases

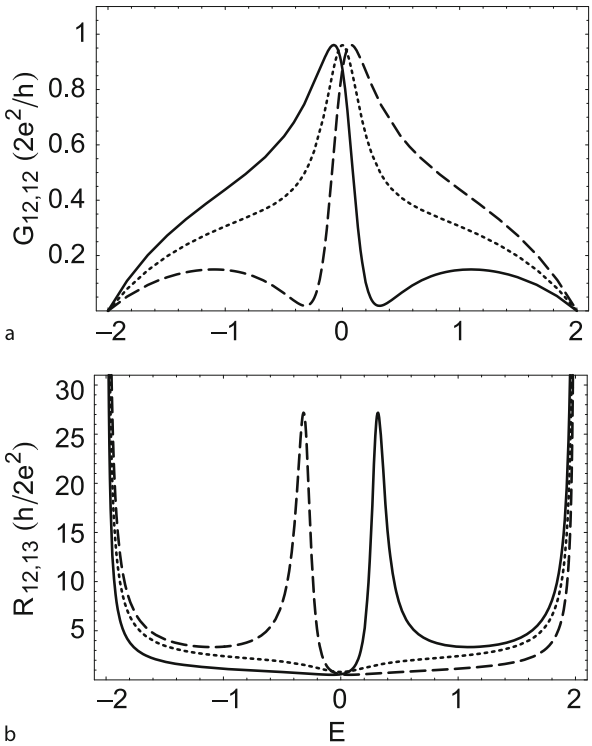

Quantum Dots: Fano Resonances in an Aharonov–Bohm Ring, Figure 17

a The conductance $G_{12,12}$ and b the resistance $R_{12,13}$ are depicted as a function of electron energy with a fixed $V_D = 0.1$ for different magnetic flux $\Phi/\Phi_0 = 0.0$ (*solid curve*), 0.25 (*dotted curve*), and 0.5 (*dashed curve*). As Φ increases, the swing from Fano to BW resonance (or vice versa) appears in the conductance $G_{12,12}$ and the resistance $R_{12,13}$ increases dramatically near the zero of the Fano resonance

In contrast to amplitude t_{21}, the cross amplitudes t_{31} and t_{32} do not have zeros in the actual region of energy plane (see Eq. (40)) and hence, the behavior of the amplitudes near the pole is expected to be similar to that of the amplitudes near a BW resonance. The transmission T_{31} and the transmission phase α_{31} in the absence of a magnetic flux for different coupling parameters, $V_D = 0.1, 0.3$, and 0.6, are shown in Fig. 16 as solid, dotted, and dashed curves, respectively. A simple BW resonance peak in T_{31}, which is less than unity, can clearly be seen in Fig. 16a, and a corresponding phase change α_{31} near the BW resonance is depicted in Fig. 16b. Notice here that α_{31} at the resonance monotonically softens as V_D increases, but the BW peak near $E \approx 0$ has a maximum amplitude at $V_D = 0.5$.

Finally, we investigate the magnetic flux dependence of the conductance $G_{12,12}$ and the resistance $R_{12,13}$ for an open ring with a fixed V_D. In Fig. 17a, the conductance $G_{12,12}$ as a function of electron energy E with a fixed $V_D = 0.1$ is shown for different values of magnetic flux $\Phi/\Phi_0 = 0.0$ (solid), 0.25 (dotted), and 0.5 (dashed). As Φ increases, a transition from Fano resonance (asymmetry parameter $q < 0$, peak $\rightarrow$ dip) to BW resonance and then back to Fano resonance ($q > 0$, dip $\rightarrow$ peak) in $G_{12,12}$ can be observed as a sequence. The Fano resonance produces a very strong influence on the resistance $R_{12,13}$. As shown in Fig. 17b, the resistance increases dramatically when the electron energy approaches the zero energy E_0 of the Fano resonance. The appearance of the peak in the resistance near the zero of the Fano resonance is connected with almost full reflection of the electron waves traveling from terminal 1 to terminal 2. This indicates that at this Fermi energy there is an additional interference resistance in the circuit region between terminals 1 and 3.

Future Directions

Electron transmission through quantum dots and Aharonov–Bohm rings has shown a rich resonance structure,

which includes Fano resonances when the QD is embedded in one arm of the AB ring. The Fano resonance is a manifestation of interference between the localized quasi-bound states of the QD in one arm and the continuum states in the other arm, characterized by both complete transmission and complete reflection. This characteristic of resonance structure (a zero-pole pair) can be controlled by changing the confinement parameters of the QD. Transmission through a QD embedded in an AB ring remains phase-coherent, as indicated by the visibility of the AB oscillations. The intrinsic phase of the QD is significant in its relation to the AB oscillations when the QD is embedded in the AB-ring, and it has experimentally been seen to exhibit interesting phase-jumps of π in a *two-terminal* system when the conductance of the AB ring reaches a peak. In a two-terminal device, the Onsager relations of time-reversal symmetry and current conservation (unitarity) constrain the transmission phase to values of 0 or π. However, if the two-terminal AB-ring is "opened" by allowing current to flow out through additional terminals, the unitarity condition is broken and it becomes possible to extract meaningful phase information about the QD.

We examine new effects resulting from the interaction of Fano resonances in the transmission of an Aharonov–Bohm ring with two embedded quantum dots. As the interaction parameter between two quantum dots is modulated, two Fano dipoles (a resonance zero-pole pair) in the complex-energy plane form a new quasi-particle, which behaves as a coupled object called a *Fano quadrupole*. In the strong overlapping regime of the Fano resonances, the collision and merging of resonance zeros takes place and these zeros move off from the real axis of energy in complex conjugate pairs. A simple two-level model demonstrating the interference attraction and repulsion of the zeros is introduced. The periodic motion of both transmission zeros and resonance poles as a function of a magnetic field is discussed. Although we have investigated the collision of two resonances, our approach may be applied to *multi-resonant* cases in which Fano-complexes in the transmission may be demonstrated.

These predicted effects may be observed in nanoscale devices with geometrical dimensions smaller than the elastic mean free paths. In these nanoscopic systems, electron transport is ballistic and phase coherence can be preserved. Because of this merit of the nanosystem, there have been many studies on electron transmission characteristics in semiconductor structures. Especially, interference of electron waves and Fano resonances in nanostructures have been extensively studied both experimentally and theoretically [2,3,4,5,6,12,13,19,20,21,22,23,24,25,26,27,28,29,31,32,37,38,39,40,41,42,46,47]. For instance, a quantum interference experiment for a quantum dot embedded in an AB ring fabricated in a two-dimensional AlGaAs/GaAs heterostructure was recently performed by Kobayashi et al. [31]. They have studied unique properties of the Fano effect on the phase and coherence of electrons traversing the AB ring. They have also reported in this tunable Fano experiment that the Fano line-shape in the electron transmission through the AB ring is tunable by an external control such as a magnetic field. In other words, the relative phase between a discrete level in the QD and the continuum state changes the Fano-type line shapes which is characterized by a complex number for the asymmetric parameter. Notice that as a result of their small size, the investigated dots are operated in Coulomb blockade regime. Meanwhile, because only a few electrons in the channels were effective in the transport through the dot, the Fano resonances have been observed in the transmission [28,29,31,32,42,46,47]. In these references, it was pointed out that single electron interference effects may be responsible for the formation of resonances.

AB rings with embedded quantum dots may be used for conductance control of quantum interference devices. Fano interference may potentially be used for the design of new types of quantum electronic or spintronic devices such as Fano-transistors [18], spin transistors, and Fano-filters for polarized electrons [44]. In addition, Fano phenomena can also be used for lasing without population inversion [33]. The developed Fano interference theory for electrons in AB rings will allow us to gain an essential understanding of the operation of novel quantum devices, and will open new opportunities for applications.

Acknowledgments

We thank G. Klimeck and C.S. Kim for helpful conversations and collaboration. The work of AMS was supported in part by the Russian Basic Research Foundation Grant No. 05-02-16762.

Bibliography

Primary Literature

1. Aharonov Y, Bohm D (1959) Significance of electromagnetic potentials in the quantum theory. Phys Rev 115:485–491
2. Aharony A, Entin-Wohlman O, Imry Y (2003) Measuring the transmission phase of a quantum dot in a closed interferometer. Phys Rev Lett 90:156802–156805
3. Aharony A, Entin-Wohlman O, Imry Y (2003) Phase measurements in Aharonov–Bohm interferometers. Turk J Phys 27:299–312
4. Aharony A, Entin-Wohlman O, Imry Y (2005) Phase measurements in open and closed Aharonov–Bohm interferometers. Physica E 29:283–288

5. Aharony A, Entin-Wohlman O, Halperin BI, Imry Y (2002) Phase measurement in the mesoscopic Aharonov–Bohm interferometer. Phys Rev B 66:115311–115318
6. Aharony A, Entin-Wohlman O, Otsuka T, Katsumoto S, Aikawa H, Kobayashi K (2006) Breakdown of phase rigidity and variations of the Fano effect in closed Aharonov–Bohm interferometers. Phys Rev B 73:195329, 1–8
7. Bowen RC, Frensley WR, Klimeck G, Lake RK (1995) Transmission resonances and zeros in multiband models. Phys Rev B 52:2754–2765
8. Breit G, Wigner E (1936) Capture of slow neutrons. Phys Rev 49:519–531
9. Büttiker M (1986) Four-terminal phase–coherent conductance, Phys Rev Lett 57:1761–1764
10. Clerk AA, Waintal X, Brouwer PW (2001) Fano resonances as a probe of phase coherence in quantum dots. Phys Rev Lett 86:4636–4639
11. Datta S (1995) Electronic transport in mesoscopic systems. Cambridge University Press, Cambridge
12. Deo PS, Jayannavar AM (1996) Phase of Aharonov–Bohm oscillations in conductance of mesoscopic systems. Mod Phys Lett B 10:787–792
13. Entin-Wohlman O, Aharony A, Imry Y, Levinson Y, Schiller A (2002) Broken unitarity and phase measurements in Aharonov–Bohm interferometers. Phys Rev Lett 88:166801–166805
14. Fano U (1961) Effects of configuration interaction on intensities and phase shifts. Phys Rev 124:1866–1878
15. Goldberg M, Watson K (1964) Collision theory. Wiley, New York
16. Gurvitz SA, Levinson YB (1993) Resonant reflection and transmission in a conducting channel with a single impurity. Phys Rev B 47:10578–10587
17. Kim CS, Satanin AM, Joe YS, Cosby RM (1999) Resonant tunneling in a quantum waveguide: effect of a finite-size attractive impurity. Phys Rev B 60:10962–10970
18. Göres J, Goldhaber-Gordon D, Heemeyer S, Kastner MA, Shtrikman H, Mahalu D, Meirav U (2000) Fano resonances in electronic transport through a single-electron transistor. Phys Rev B 62:2188–2194
19. Hedin E, Joe YS, Satanin AM (2007) Control of Fano resonances and phase of a multi-terminal Aharanov–Bohm ring with three embedded quantum dots, J Comput Electron 6:323–327
20. Hedin E, Cosby RM, Satanin AM, Joe YS (2005) Electron wave interferometry through an asymmetric Aharonov–Bohm ring. J Appl Phys 97:063712–063716
21. Hofstetter W, König J, Schoeller H (2001) Kondo correlations and the Fano effect in closed Aharonov–Bohm interferometers. Phys Rev Lett 87:156803–156807
22. Joe YS, Hedin E, Satanin AM (2007) Manipulation of resonances in an open three-terminal interferometer with an embedded quantum dot. Phys Rev B 76:085419, 1–6
23. Joe YS, Kim J, Satanin AM (2006) Resonance characteristics through double quantum dots embedded in series in an Aharonov–Bohm ring. J Phys D: Appl Phys 39:1766–17667
24. Joe YS, Satanin AM, Kim CS (2006) Classical analogy of Fano resonance. Phys Scr 74:259–266
25. Joe YS, Satanin AM, Klimeck G (2005) Interactions of Fano resonances in the transmission for an Aharonov–Bohm ring with two embedded quantum dots in the presence of a magnetic field. Phys Rev B 72:115310–115316
26. Joe YS, Kim JS, Hedin E, Cosby RM, Satanin AM (2005) Fano resonance through quantum dots in tunable Aharonov–Bohm Rings. J Comput Electron 4:129–133
27. Kang K (1999) Phase evolution of the transmission coefficient in an Aharonov–Bohm ring with Fano resonance. Phys Rev B 59:4608–4611
28. Katsumoto S, Kobayashi K, Aikawa H, Sano A, Iye Y (2003) Quantum coherence in quantum dot-Aharonov–Bohm ring hybrid systems. Superlattices Microstruct 24:151–157
29. Keyser UF, Borck S, Haug RJ, Bichler M, Abstreiter G, Wegscheider W (2002) Aharonov–Bohm oscillations of a tunable quantum ring. Semiconduct Sci Technol 17:L22–L24
30. Kim CS, Roznova ON, Satanin AM, Shtenberg VB (2002) Interference of quantum states in electronic waveguides with impurities. JETP 94:992–1007
31. Kobayashi K, Aikawa H, Katsumoto S, Iye Y (2002) Tuning of the Fano effect through a quantum dot in an Aharonov–Bohm interferometer. Phys Rev Lett 88:256806–256809
32. Kobayashi K, Aikawa H, Sano A, Katsumoto S, Iye Y (2004) Fano resonance in a quantum wire with a side-coupled quantum dot. Phys Rev B 70:035319–035325
33. Nikonov DE, Imamoglu A, Scully MO (1999) Fano interference of collective excitations in semiconductor quantum wells and lasing without inversion. Phys Rev B 59:12212–12215
34. Nöckel JU, Stone AD (1994) Resonance line shapes in quasi-one-dimensional scattering. Phys Rev B 50:17415–17432
35. Onsager L (1931) Reciprocal relations in irreversible processes. Phys Rev 38:2265–2279
36. Porod W, Shao ZA, Lent CS (1993) Resonance-antiresonance line shape for transmission in quantum waveguides with resonantly coupled cavities. Phys Rev B 48:8495–8498
37. Ryu CM, Cho SY (1998) Phase evolution of the transmission coefficient in an Aharonov–Bohm ring with Fano resonance. Phys Rev B 58:3572–3575
38. Satanin AM, Joe YS (2005) Fano-interference and resonances in open systems. Phys Rev B 71:205417, 1–12
39. Satanin AM, Joe YS (2005) Manipulating of resonances in conductance of an electron waveguide with antidotes. J Comput Electron 4:149–153
40. Satanin AM, Joe YS (2007) Fano-interference of decaying states with continuum in nanostructures. J Superconduct Novel Magn 20:179–182
41. Satanin AM, Hedin E, Joe YS (2006) Collision of Fano resonances: an exact solvable model. Phys Lett A 349:45–52
42. Schuster R, Buks E, Heiblum M, Mahalu D, Umansky V, Shtrikman H (1997) Phase measurement in a quantum dot via a double-slit interference experiment. Nature 385:417–421
43. Shao Z, Porod W, Lent CS (1994) Transmission resonances and zeros in quantum waveguide systems with attached resonators. Phys Rev B 49:7453–7465
44. Song JF, Ochiai Y, Bird JP (2003) Fano resonances in open quantum dots and their application as spin filters. Appl Phys Lett 82:4561–4563
45. Tekman E, Bagwell PF (1993) Fano resonances in quasi-one-dimensional electron waveguides. Phys Rev B 48:2553–2559
46. Yacoby A, Schuster R, Heiblum M (1996) Phase rigidity and $h/2e$ oscillations in a single-ring Aharonov–Bohm experiment. Phys Rev B 53:9583–9586
47. Yacoby A, Heiblum M, Mahalu D, Shtrikman H (1995) Coherence and phase sensitive measurements in a quantum dot. Phys Rev Lett 74:4047–4050

Books and Reviews

Adair RK, Bockelman CK, Peterson RE (1949) Experimental corroboration of the theory of neutron resonance scattering. Phys Rev 76:308

Cerdeira F, Fjeldly TA, Cardona M (1973) Effect of free carriers on zone-center vibrational modes in heavily doped p-type Si. II Optical modest. Phys Rev B 8:4734–4745

Dittes F-M (2000) The decay of quantum systems with a small number of open channels. Phys Rept 339:215–311

Fano U, Rau ARP (1986) Atomic collisions and spectra. Academic, Orlando

Feshbach H (1958) Unified theory of nuclear reactions. Ann Phys (NY) 5:357–390

Quantum Dot Spin Transistors, Self-consistent Simulation of

Dmitriy V. Melnikov, Jean-Pierre Leburton
Department of Electrical and Computer Engineering & Beckman Institute for Advanced Science and Technology, University of Illinois, Urbana, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Vertical QD Device Structure
Self-consistent Approach
QD Ground State Charging in Magnetic Fields
Exact Diagonalization of the Many-Electron Schrödinger Equation
Quantum Transport Model
Tunneling Spectroscopy of a Few-Electron QD in the Non-Linear Transport Regime
Conclusion
Future Directions
Acknowledgments
Bibliography

Glossary

Quantum dot (QD) A man-made nanostructure, generally made with semiconductor materials, in which the motion of particles (such as conduction band electrons) is bound in all three spatial directions. As a result of this 3D spatial confinement, quantum dots exhibit a discreet energy spectrum with particle wave functions localized within the quantum dot.

Coulomb blockade The increased resistance experienced at small bias voltages by an electronic device comprising of at least two tunnel junctions (terminals), which is manifested in the electrostatic blockade of the current by the charge accumulation on a small conducting island between the terminals.

Single electron transistor An electronic device characterized by two tunnel junctions ("source" and "drain") between which a conducting "island" (QD) is located. The electrostatic potential of the QD island is controlled by a third electrode (the "base"). By changing the controlling voltage on the base electrode, transport channels for electron current through the QD can be opened or closed. When the channel is open, an electron can tunnel into the island on an available energy level from the source contact and then subsequently tunnels out to the drain electrode (single-electron transport).

Spin blockade The suppression (partial or total) of the current through a QD device populated by electrons and caused by the Pauli exclusion principle.

Definition of the Subject

The electronic and transport properties of quantum dot spin transistors are presented with emphasis on single-electron tunneling and shell structure. A comprehensive modeling approach based on two methods is developed: (1) quantum dot electronic structure and confinement potential are determined from the self-consistent solution of the Poisson and Schrödinger equations within the spin-density-functional theory in magnetic fields and (2) a quantum transport model based on numerically exact diagonalization of the many-body Schrödinger equation is used to describe transport properties of quantum dots. In the linear transport regime characterized by a small applied source-drain voltage, single-electron tunneling through the quantum dot reveals the existence of a shell structure in the ground state electron addition spectra which magnetic field dependence is determined by competition among the many-body interaction effects, confinement potential strength and magnetic field induced localization of electrons. In the non-linear transport regime, the information about excitation states of quantum dots can be extracted from the transport spectra. The current in this case is dominated by the asymmetry in the tunneling barriers, spin selection rules and the overlap between various many-body states. The very good agreement between calculated and measured currents allows one to put in correspondence numerous features in transport and energy spectra of quantum dots.

Introduction

In recent years few-electron semiconductor quantum dots (QDs) have attracted considerable attention because these nanostructures exhibit many features characteristic of real atomic systems. QDs are also promising candidates for future device applications [1]. In these man-made systems conduction electrons are held together in a finite region of space by the confinement potential that is usually created by the hetero-structure barriers and/or the electrostatic potential of remote dopant charge distributions modulated by external gate voltages.

Among several types of QDs, we consider the so-called "gated" quantum dots in which controlled single-electron charging and shell structure was first observed [2]. In vertical QDs, a small, quasi-two-dimensional (quasi-2D) electron island (dot) is formed in a mesa-structure pillar (Fig. 1a) between two heterostructure barriers (see Fig. 4). The size of the island (or the electron occupation of the QD) can be changed electrostatically by applying a voltage to the gate wrapped around the vertical structure. The current flows vertically (hence the term "vertical QD") through the heterostructure in response to a bias applied between the source and drain contacts on top and bottom of the mesa-structure. The fact that this device has three terminals is reminiscent of a field-effect transistor where the base terminal corresponds to the side gate in the QD device and controls the single-electron current through the heterostructure.

Another method frequently utilized to fabricate QDs is the lithographic patterning of gates, i. e., the deposition of metal electrodes on a heterostructure surface (Fig. 1b). By biasing the top gate electrodes, the two-dimensional (2D) electron gas formed at the hetero-interface between different materials such as AlGaAs and GaAs is depleted underneath them, thereby creating an island of non-zero electron density that can be further fine-tuned by changing voltages on nearby gates [3]. This device is called lateral or planar QD because the current flows in the plane of the 2D electron gas.

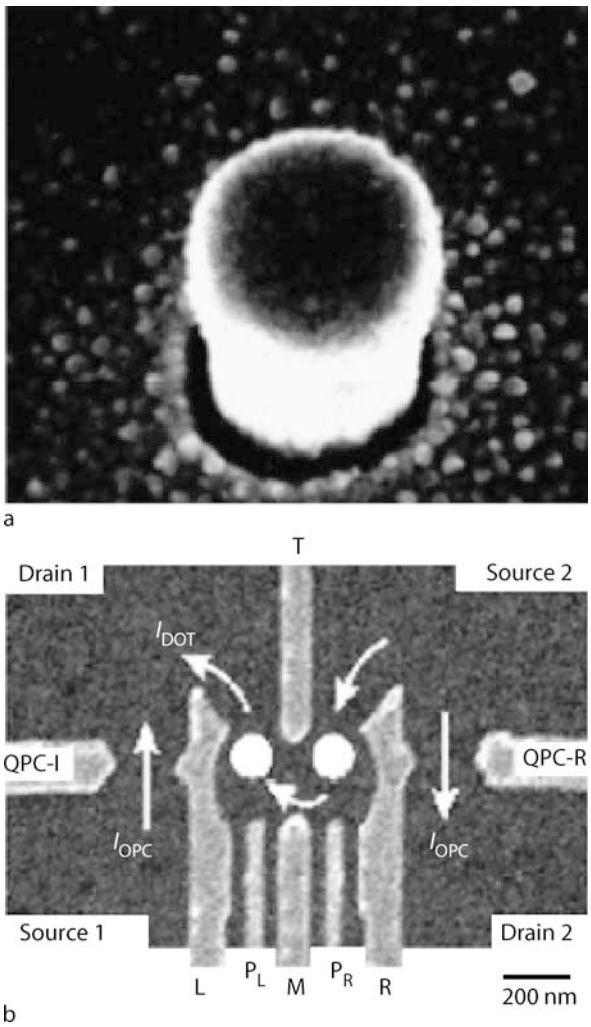

Quantum Dot Spin Transistors, Self-consistent Simulation of, Figure 1

Scanning-electron micrograph of a the vertical QD structure (analogous to the one used in [2]) and b planar system of two laterally coupled QDs [3]

Single-Electron Tunneling

Single-electron effects have been known for a long time. In his famous 1911 experiments, Millikan [4] measured the value of the electron charge observing the falling rate of charged oil drops. Systematically single-electron charging effects in electron tunneling were first studied by transport measurements on thin films of small metallic grains [5,6,7]. In 1975, Kulik and Shekter [8] showed that in a two-terminal system, the current through a small grain at low bias voltages is blocked by the accumulated charge, whereas the differential conductance varies periodically when a larger source-drain bias is applied. This so-called Coulomb blockade and the Coulomb "staircase" were later observed for the first time in granular systems [9] and thin-film tunnel junctions [10].

In principle, QDs are ideal systems for the investigation of single-electron tunneling which can be accomplished by connecting the dot island to surrounding reservoirs [11,12,13]. Figure 2a schematically shows a QD electron island connected to the leads through hetero-barriers as in the vertical QD structure (it can also be electrostatic barriers as in planar QD structures), and a side gate biased

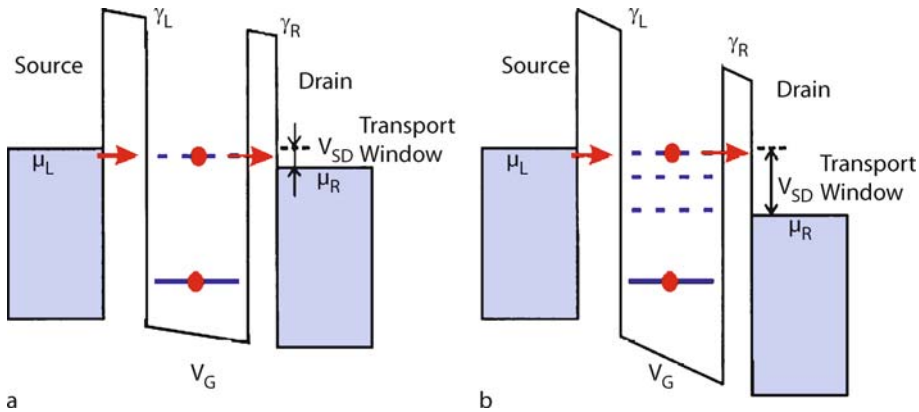

Quantum Dot Spin Transistors, Self-consistent Simulation of, Figure 2
Single-electron transport through the QD ($N = 2$) in a the linear regime ($V_{SD} \to 0$) and b non-linear regime (V_{SD} is large). In both cases the QD contains one electron while one channel (three channels in b) open for the tunneling of the second electron. The left and right barrier permeabilities are γ_L and γ_R respectively

with a controlling voltage V_G that changes positions of the energy levels in the QD with respect to the chemical potentials of the left (L) and right (R) leads μ_L and μ_R. The electro-chemical potential of the dot filled with N electrons (the solid line in Fig. 2), is given by

$$\mu_{QD}(N) = E(N) - E(N-1)\,, \tag{1}$$

where $E(N)$ is the ground-state energy (at zero temperature) of the N-electron QD. As the leads (the source and the drain contacts) are weakly coupled to the QD, the electrochemical potentials μ_L and μ_R become different when a constant source-drain bias V_{SD} is applied between the leads, and a transport window of width $\mu_L - \mu_R = eV_{SD}$ opens up.

In the linear transport regime, the transport window eV_{SD} is much smaller than the average spacing among the quantum states, so that only the ground state of the QD contribute to the conductance. By changing the voltage V_G on the side gate, one can achieve alignment of $\mu_{QD}(N)$ with the transport window so that electrons can subsequently tunnel in and out of the QD; this situation corresponds to a conductance maximum when the number of electrons in the QD cycles between N and $N+1$. Otherwise the current is blocked; this scenario corresponds to zero conductance (and current) when the number of electrons N in the dot is fixed, and it increases to $N+1$ each time a conductance maximum is crossed. This mechanism of discrete charging and discharging of the QD leads to Coulomb blockade oscillations in the conductance as a function of the gate voltage. The distance between neighboring Coulomb peaks [14] equals the difference in the electrochemical potentials of a QD containing $N+1$ and N electrons:

$$\Delta_2(N) = \mu_{QD}(N+1) - \mu_{QD}(N)\,. \tag{2}$$

In the non-linear transport regime, i. e. for a larger transport window eV_{SD} and at a fixed V_G, additional peaks in differential conductance traces occur. They are due to the excitations in the N-electron QD system as electrons can tunnel in and out of the QD via the ground as well as low lying excited states (Fig. 2b). By plotting the positions of these peaks as a function of V_{SD} and V_G, a characteristic diamond-shaped structure is observed, which provides information about the ground and excited states of the QD [15,16,17,18,19]. Note that unlike optical absorption (radiation) processes in real atoms where electron excitations are created by photons, the excitation mechanism in artificial QD atoms is electron-only, e. g., excitations can be created when one electron tunnels out from the ground state, while another one tunnels in the excited state. However, in general, the analysis of the various features observed in the tunneling current through the QD in terms of the excitation energies in the non-linear transport regime is a challenging task that we address in Sect. "Tunneling Spectroscopy of a Few-Electron QD in the Non-Linear Transport Regime".

Shell Structure

In the three-dimensionally (3D) confined systems such as QDs energy quantization and interaction among the confined particles leads to the existence of a "shell structure". It can be illustrated on the following example of a simple

two-dimensional harmonic-oscillator confinement potential:

$$V_{\text{conf}}(x, y) = \tfrac{1}{2} m^* \omega^2 \left(x^2 + y^2\right) , \tag{3}$$

which can be considered as an empirical mean-field potential in which N particles with effective mass m^* move independently and where ω determines the confinement strength. The corresponding single-particle energy levels are obviously

$$E_{n_x, n_y} = \hbar\omega \left(n_x + n_y + 1\right) , \tag{4}$$

where one easily recognizes the $(N_0 + 1)$-fold state degeneracy with respect to the quantum number $N_0 = n_x + n_y = 0, 1, 2, \ldots$ By filling the states with non-interacting electrons and by accounting for the Pauli principle, one can obtained closed shells for a sequence of $N = 2, 6, 12, 20, \ldots$ particles. For these configurations, a particular stability is achieved as the degeneracy of the shell is resolved. Adding one more electron to a closed shell results in the single occupation of an orbital belonging to the next higher shell which makes this configuration less stable since as a larger amount of energy than for the closed shell configurations needs to be supplied to the system. Despite its simplicity, this example illustrates some of the basic features of few-electron QD systems such as the existence of electron shells and the stability of closed-shell configurations.

The periodic table of elements (Mendeleev table), with its groups of elements characterized by similar chemical properties, is the most well-known example of shell structure in nature. Atomic shells are clearly seen in the pronounced maxima of the ionization energies of neutral atoms for atomic numbers $Z(= N) = 2, 10, 18, \ldots$ corresponding to the noble gases He, Ne, Ar, etc. (the 3D spherical symmetry of the rigid confinement caused by the strong $1/r$ Coulomb potential of the nucleus results in closed shells forming at different values N than in the above 2D example). The shells are populated according to first Hund's rule stating that the spin is maximized for half-filled orbitals due to the Pauli principle and the repulsive Coulomb interaction [20].

In a clear analogy with atoms, nuclei [21] or atomic clusters [22], shell structure was observed in the transport spectra of semiconductor QDs for the first time by Tarucha et al. [2] in vertical structures (see Figs. 1a and 4). The electron number N in the QD was varied one-by-one starting from $N = 0$ by increasing the negative voltage V_G applied to the side gate (at $V_G = 0$ the QD was already populated by a large number of electrons). A current I

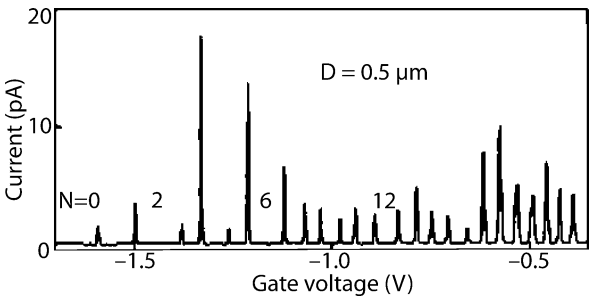

Quantum Dot Spin Transistors, Self-consistent Simulation of, Figure 3
Current oscillations due to the Coulomb blockade in the linear transport regime measured in the vertical QD system [2]

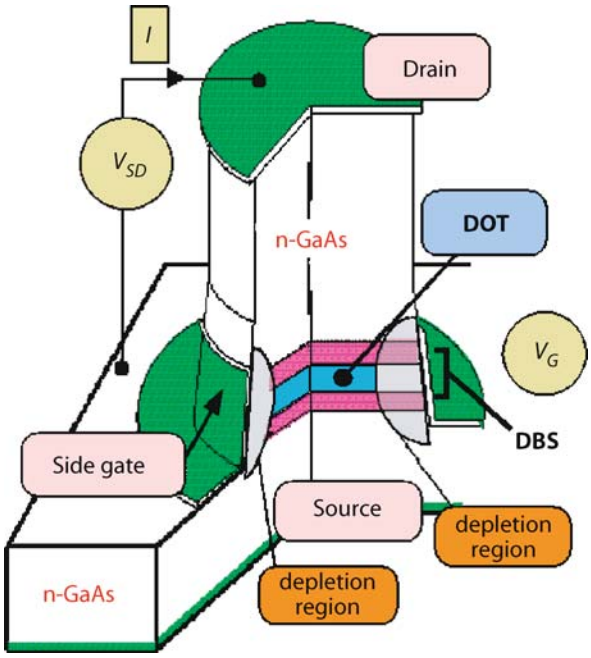

Quantum Dot Spin Transistors, Self-consistent Simulation of, Figure 4
Schematic representation of the vertical QD mesa-structure [2]

flowed through the QD sandwiched between two heterostructure barriers in response to a very small bias V_{SD} (linear transport regime) between the source and drain contacts. By measuring the current as a function of the side gate voltage, current peaks corresponding to single-electron tunneling events were observed. The separation between consecutive peaks is proportional to the difference in energy needed to add an electron to the dot already confining N particles (Eq. (2)). The current-voltage characteristics plotted in Fig. 3 has large inter-peak separations at electron numbers $N^* = 2$, 6, and 12 which correspond to

closed shells for the 2D harmonic oscillator potential (3). The smaller inter-peak separations at the midshell regions for $N = 4, 9, \ldots$ is a consequence of electron spin alignment due to first Hund's rule (for details, see Sect. "QD Ground State Charging in Magnetic Fields").

Experimental Techniques

In semiconductor QDs, numerous experimental techniques for probing single-electron charging effects on single dots or arrays of dots have been implemented which allow a detailed spectroscopic study of the ground and excited states of individual artificial atoms. Originally, far-infrared (FIR) and optical spectroscopy were applied to arrays [23,24] as well as to individual quantum dots [25] because of their success in atomic and molecular physics. However, the parabolic dependence of the confinement potential and the long wave-length of the radiation suppresses most transitions in the excited states by dipole selection rules [26]. In many cases, it was found that the FIR absorption spectra are indicative of a non-interacting electron system [23,24], in which case Kohn's theorem [27] indicates that the effects of electron-electron interactions in a QD can be observed only if the anharmonicity of the confinement is sufficiently strong [28].

On the other hand, transport-based methods are expected to be largely free of the optical spectroscopy limitations which makes them more attractive for studying QDs. The first QD capacitance measurements were reported by Smith et al. [29] and single-electron capacitance spectroscopy has been used for both arrays [30,31,32] and individual quantum dots [33]. In these experiments, electron tunneling into the QD was observed upon increasing the positive bias on the top plate. The spacing between conductance peaks was approximately constant reflecting conventional Coulomb blockade effect similar to earlier experiments [34] in larger, lateral quantum dots. For smaller biases, however, the distance between consecutive peaks increased, and the spacing between them became nonuniform [33,35]. These deviations from the equidistant Coulomb blockade spectra were attributed to the QD energy spectrum quantization which will be explained in details below (see Sect. "QD Ground State Charging in Magnetic Fields").

Many-Body Calculations of the QD Electronic Structure

In general, electrons in GaAs-based QD mesa-structures originate from remote ionized shallow impurities in the (doped) leads and are confined around the lowest energy minimum of the semiconductor conduction band. The electron density in the leads is low with the mean electron-electron distance ~ 10 nm validating the effective-mass approximation: the conduction electrons have an effective mass m^* and their Coulomb interaction is screened with the static dielectric constant ε of the semiconductor in question.

From a theoretical point of view, interacting electrons confined in a quasi-2D QD form a seemingly simple many-body problem . The electron-electron correlations give rise to numerous intriguing QD properties [36]. Kumar et al. [37] calculated the effective single-particle confinement for a QD created by square-shaped metallic gates using a self-consistent approach, where the electrostatic confinement potential was obtained from a self-consistent solution of the combined Hartree and Poisson equations for the whole 3D QD mesa-structure. In their approach, electrons in the QD region were treated fully quantum-mechanically via the solution of Schrödinger equation (neglecting image charge effects as well as electron correlations) while the charge density in the leads was described semi-classically. They found that in the limit of small particle numbers confined in the QD, the effective confinement has a symmetry close to circular, even if the dot region was defined by a square-shaped gate pattern. On the basis of their work, the simple isotropic harmonic oscillator (see Eq. (3) above) was adopted as the "standard" 2D QD *model* potential for future numerous QD electronic structure calculations. Macucci, Hess, and Iafrate [38] extended the work [37] by incorporating the exchange and correlation contributions in electron-electron interaction within the density-functional theory (DFT) and found a shell-like structure for the electro-chemical potentials $\mu_{\mathrm{QD}}(N)$.

However, it turns out that the addition energy calculations based on a model 2D confinement potential and the experimental data start to deviate when the electron number N increases [39]. In this respect, analysis of a series of experimental addition energy spectra for 14 different structures with diameters between 0.44 and 0.6 μm, which are similar to the QD device used by Tarucha et al. [2], finds [40] strong device-to-device variations: While all structures show the first shell at $N = 2$, only 71% of them show shells at both $N = 2$ and 6, 64% at $N = 2$, 6, and 12, and 21% at $N = 2$, 6, 12, and 20. These observations seem to indicate that each QD has "a mind of its own," and one should be cautious with a quantitative comparison between experiment and theoretical calculations based on a fixed model confinement potential. Possible reasons for the disagreement are either non-parabolicity of the confining potential or the unavoidable inaccuracies in device fabrication that disturb the perfect circular symmetry of the QD.

These deviations, at least in part, can be accounted for by performing a full scale 3D analysis of the QD mesa-structure without any a priori assumptions about the shape or strength of the confinement potential (as it was done in the above [40]). In this multi-scale approach spin-density-function theory (SDFT) is utilized to treat the many body interactions among the electrons in the QD fully quantum mechanically including correlation and spin effects by solving Kohn–Sham (KS) equations. On the other hand, a semi-classical Thomas–Fermi approach is used to describe the influence of the charge distribution outside the QD region. As a result, the QD confinement potential and the quantized energy spectrum are obtained directly from the self-consistent solution of the non-linear Poisson and Kohn–Sham (KS) equations with device boundary conditions.

DFT calculations were also performed [41] for studying the electron structure and statistical properties of the level spacings in dots containing ~ 100 electrons. However, in this approach the 3D KS equations were separated into 2D and 1D equations by taking into account quasi-2D nature of the QD confinement. Self-consistent procedure [42] was also used for solving full 3D Poisson and Schrödinger equations for a few-electron cylindrical QDs. Here the N-electron Schrödinger equation was solved by the unrestricted Hartree–Fock method and the ground state charging energies (chemical potentials) of the QD in magnetic fields were computed [43].

In the following chapters a theoretical analysis of the QD electronic and transport properties is given. The discussion is based on the calculations performed for the vertical QD structure described in Sect. "Vertical QD Device Structure". An approach suitable to the simulation of the QD ground state properties which is based on the self-consistent solution of the Poisson and Schrödinger equations within the SDFT is presented in Sect. "Self-consistent Approach". The ground state addition energy spectra and shell structure in magnetic fields are computed and compared with experimental data in Sect. "QD Ground State Charging in Magnetic Fields". Next a description of a numerically exact diagonalization of the many-particle Schrödinger equation (Sect. "Exact Diagonalization of the Many-Electron Schrödinger Equation") and the quantum transport model (Sect. "Quantum Transport Model") is provided followed by Sect. "Tunneling Spectroscopy of a Few-Electron QD in the Non-Linear Transport Regime," where the relationship between the computed current and energy spectra in the non-linear transport regime (excited state spectroscopy) is established and comparison with measured data is given. Finally, Sect. "Conclusion" contains concluding remarks.

Vertical QD Device Structure

Single-electron tunneling controlled by external gates is best illustrated on the example of a gated vertical QD structure schematically shown in Fig. 4 which is similar to the QD used in the pioneering experiments of Tarucha et al. [2]. The mesa-structure was fabricated from a double-barrier heterostructure (DBS) by etching techniques [44], and the electron "puddle" (QD) was located in the quantum well between two hetero-barriers that separated it from the outside environment. A metal Schottky gate (side gate) was wrapped around the base of the circular pillar with a diameter of 0.5 μm. The energy gap between conductance and valence bands was also reduced by including 5% In in the 12 nm GaAs quantum well sandwiched between the two $Al_{0.22}Ga_{0.78}As$ barriers which were nominally 7.5 and 9.0 nm thick. The presence of In lowered the bottom of the conduction band below that of the n-doped GaAs leads (source and drain contacts) so that the lowest energy level in the QD was below the Fermi level of the contacts, i. e., electrons could accumulate in the dot even if no source-drain bias voltage V_{SD} was applied. This made it possible to study electron transport at very small source-drain bias voltages. The side gate voltage V_G changed the effective diameter of the island, i. e., it controlled the strength of the effective confinement potential, thereby allowing one-by-one change in the QD electron population (single-electron QD charging). As the effective diameter of the QD is much larger than its thickness in the vertical direction perpendicular to the hetero-barriers, the motion of the electrons in the vertical z-direction is "frozen" so that only the ground state in that direction is occupied, and it is the energy quantization in lateral x–y-plane which is affected by the side gate voltage.

Self-consistent Approach

We use the SDFT [45] to describe the ground state properties of electrons confined in the QD, in which the charge density $\rho(\mathbf{r})$ is calculated after solving KS equations for electrons with spin up ($\uparrow$) and down ($\downarrow$):

$$\hat{H}^{\uparrow(\downarrow)}\psi^{\uparrow(\downarrow)}(\mathbf{r}) = \varepsilon^{\uparrow(\downarrow)}\psi^{\uparrow(\downarrow)}(\mathbf{r}) , \tag{5}$$

where the single-particle Hamiltonian $\hat{H}^{\uparrow(\downarrow)}$ is given as

$$\hat{H}^{\uparrow(\downarrow)} = \hat{T} + \phi(\mathbf{r}) + \Delta E_c(\mathbf{r}) + v_{xc}^{\uparrow(\downarrow)}(\mathbf{r}) \tag{6}$$

with $\hat{T}$ being the kinetic energy operator which in the presence of an external magnetic field $\mathbf{B}$ reads as:

$$\hat{T} = \frac{1}{2}\left(-i\hbar\nabla + \frac{e}{c}\mathbf{A}\right)\frac{1}{m^*(\mathbf{r})}\left(-i\hbar\nabla + \frac{e}{c}\mathbf{A}\right) . \tag{7}$$

Here $m^*(\mathbf{r})$ is the position dependent electron effective mass, and $\mathbf{A} = (B/2)(y, -x)$ is the vector potential in symmetric gage for the the magnetic field oriented along the z-direction perpendicular to the QD plane (Zeeman splitting is neglected for clarity). In Eq. (6), $\Delta E_c(\mathbf{r})$ stands for the conduction band offset between the different materials. For the vertical QD structure described in Sect. "Vertical QD Device Structure", its values are fixed at 180 and −40 meV for $Al_{0.22}Ga_{0.78}As/GaAs$ and $In_{0.05}Ga_{0.95}As/GaAs$ interfaces, respectively (note that for the real structures with non-zero doping in the GaAs contact regions, the value of $\Delta E_c(\mathbf{r})$ for $In_{0.05}Ga_{0.95}As/GaAs$ is difficult to determine precisely). In general, this method is not limited to single vertical QD structures but can be quite straightforwardly applied to modeling any layered semiconductor structures such as those used for the planar (lateral) QDs [46] as well as systems of double [47] and triple vertical QDs [48].

The exchange-correlation potential $v_{xc}^{\uparrow(\downarrow)}(\mathbf{r})$ in Eq. (5) is computed within the local spin density approximation (LSDA) [49] and does not explicitly depend on magnetic field. Comparison of DFT results with calculations using current-spin density functional theory [50] for 2D systems showed that this approximation is reliable over a wide range of magnetic field, although at higher fields, effects of paramagnetic currents in $v_{xc}^{\uparrow(\downarrow)}(\mathbf{r})$ should become more important [51].

The potential $\phi(\mathbf{r}) = \phi_{ext}(\mathbf{r}) + \phi_{ion}(\mathbf{r}) + \phi_H(\mathbf{r})$ in Eq. (6) is the sum of the external potential $\phi_{ext}(\mathbf{r})$ due to the applied voltage, screening potential $\phi_{ion}(\mathbf{r})$ arising from the ionized impurities in the structure, and Hartree potential $\phi_H(\mathbf{r})$ accounting for the repulsive electron-electron interactions. It is obtained from the solution of the Poisson equation:

$$\nabla\epsilon(\mathbf{r})\nabla\phi(\mathbf{r}) = 4\pi\rho(\mathbf{r}) , \tag{8}$$

where $\epsilon(\mathbf{r})$ is the dielectric constant of the medium, and $\rho(\mathbf{r})$ is the charge density which inside the QD region is equal to

$$\rho(\mathbf{r}) = -e\left(\sum_{\text{occup}}\left|\psi^{\uparrow}(\mathbf{r})\right|^2 + \sum_{\text{occup}}\left|\psi^{\downarrow}(\mathbf{r})\right|^2\right) , \tag{9}$$

with the summations spanning occupied states for electrons with spin up and down (the number of those states is, in general, different). Outside the QD region, charge distribution is determined from electron $n(\mathbf{r})$ and hole $p(\mathbf{r})$ densities calculated within the semi-classical Thomas–Fermi approximation [52]:

$$n(\mathbf{r}) = \frac{4}{\sqrt{\pi}}\left(\frac{2\pi m_e(\mathbf{r})T}{h^2}\right)^{3/2} F_{1/2}\left[\frac{-e\phi(\mathbf{r}) + \Delta E_c(\mathbf{r})}{T}\right] , \tag{10}$$

$$p(\mathbf{r}) = \frac{4}{\sqrt{\pi}}\left(\frac{2\pi m_h(\mathbf{r})T}{h^2}\right)^{3/2} F_{1/2}\left[\frac{e\phi(\mathbf{r}) - E_G(\mathbf{r}) - \Delta E_c(\mathbf{r})}{T}\right] , \tag{11}$$

$$F_{1/2}[\eta] = \int_0^\infty \frac{\mathrm{d}x\, x^{1/2}}{1 + \exp(x - \eta)} , \tag{12}$$

where $E_G(\mathbf{r})$ and $m_h(\mathbf{r})$ are the band gap and the hole effective mass of the constituent materials, and $T = 0.1$ K is the temperature. These densities are screened by the (completely) ionized donors and acceptors $N_D^+(\mathbf{r})$ and $N_A^-(\mathbf{r})$ distributed in the QD leads:

$$\rho(\mathbf{r}) = -e\left[N_D^+(\mathbf{r}) - N_A^-(\mathbf{r}) + p(\mathbf{r}) - n(\mathbf{r})\right] . \tag{13}$$

Since QD regions are usually much smaller than the physical dimensions of the device, the KS wavefunctions actually vanish long before reaching the device boundaries. This allows us to embed a local region in the global mesh for solving the KS equations. This local region is chosen to be large enough to ensure vanishing wavefunctions on its boundaries. For the Poisson equation (8), zero electric field on the lower part of the structure buried in the substrate and on the top contact plane are used as a boundary condition. On other surfaces, not covered by the side gate, the potential ϕ is set equal to the Schottky barrier value $V_S = 0.9$ eV. Boundary values of the potential on the side gate are equal to the Schottky barrier value modified by the applied gate bias, $V_S - V_G$. The system of KS and Poisson equations (5), (8) is solved iteratively until a self-consistent solution for the KS orbitals $\psi^{\uparrow(\downarrow)}(\mathbf{r})$ and eigenvalues $\varepsilon^{\uparrow(\downarrow)}$ is obtained.

The calculations are performed on a parallel platform by means of the finite element method (FEM) with trilinear polynomials on a variable size grid [53]. The advantages of FEM utilization are the ability to systematically improve the accuracy by expanding the basis set and its inherent variational nature. Due to the last reason, calculated energy differences between two similar configurations are usually more accurate than total energy computed for each system separately since each energy is already an upper bound to the exact value.

In our approach Poisson's equation (8) is solved by means of the damped Newton–Raphson method while the

generalized eigenvalue problem obtained after discretization of the KS equation (5) is tackled by means of a subspace iteration method based on a Rayleigh–Ritz analysis [53]. The small number of required eigenpairs (~ 10) made this approach sufficient. A parallel conjugate-gradient method preconditioned with block Jacobi with an incomplete LU factorization on the blocks is utilized for solving the resulting matrix equations. In the presence of a magnetic field, the matrix obtained from the KS equation is Hermitian and the hermitian-conjugate method with the same preconditioner as above had to be used in solving the eigenvalue problem. Compared to the ordinary conjugate gradient method with a Jacobi preconditioner, this approach gives rise to at least an order of magnitude increase in performance especially when working with Hermitian matrices [54].

After the eigenvalues are determined, the ground state electron charging diagram can be calculated by directly comparing the total energy $E(N)$ of the N-electron system with the system containing $N-1$ electrons, the difference of which gives the required value of the chemical potential $\mu_{\mathrm{QD}}(N)$ (see Eq. (1)). One can also use Slater's rule (transition state technique) in order to calculate $\mu_{\mathrm{QD}}(N)$ [55]:

$$\mu_{\mathrm{QD}}(N) = E(N) - E(N-1) \approx \varepsilon(1/2)\,, \tag{14}$$

where $\varepsilon(1/2)$ is the eigenvalue of the state with half occupation in the system with $N-1/2$ electrons (the transition state). By varying the side gate voltage V_{G}, both $E(N)$ and $\mu_{\mathrm{QD}}(N)$ can be changed. If $\mu_{\mathrm{QD}}(N) < 0$, then the N-electron configuration is stable, otherwise the number of electrons in the QD is $N-1$. The side gate bias $V_{\mathrm{G}}(N)$ at which $\mu_{\mathrm{QD}}(N) = 0$ gives the charging voltage for the Nth electron (or equivalently the boundary between the QD stable configurations with N and $N-1$ electrons).

An external magnetic field induces crossings among the states with different total spins, i. e., the spin configuration of the system changes while the number of electrons in the QD remains constant. In this regard, one should note that Slater's rule (14) can only be used when there is a change in the occupation of a single eigenlevel. If the change in the electron number is accompanied by spin rotations of individual electrons, i. e., the occupation of several KS levels is simultaneously changed (this happens, e. g., in $N = 2$ system with $S = 0$ at $B \sim 5$ T in Fig. 6 when the addition of a third electron results in $S = 3/2$ state), Slater's rule should be invoked several times. However, such situations involving several transition states are infrequent so that overall, the utilization of Slater's rule significantly reduces the amount of computational time to calculate the full charging diagram of a QD in magnetic fields [56].

From the $\mu_{\mathrm{QD}}(N)$ value, the ground state electron addition energies can be determined by computing the QD capacitive energy [57] given by Eq. (2). This quantity is evaluated for side gate voltage $V_{\mathrm{G}}(N+1)$ corresponding to the addition of the $(N+1)$th electron to the system so that $\Delta_2(N) = -\mu_{\mathrm{QD}}(N)$. In the following chapter, we discuss in detail is the variation of the charging voltage $V_{\mathrm{G}}(N)$ and addition energy as functions of the electron number N in a QD and magnetic fields.

QD Ground State Charging in Magnetic Fields

The computed electron addition energy spectra $\Delta_2(N)$ of the circular QD in the absence of a magnetic field are shown in Fig. 5 together with the corresponding experimental data from [58]. The spectrum exhibits pronounced maxima for two and six electrons due to the first and second shell closures characteristic of QDs with parabolic 2D circular confinement. The peaks and valleys are a consequence of the interplay between confinement and many-body effects. For $N = 2$, the lowest single-particle state ε_1 is fully occupied. The third electron populates the next available eigenstate with $\varepsilon_2 > \varepsilon_1$, thus making the addition of the electron energetically more costly than for the case of the second electron. The same situation is repeated for $N = 6$ when the second shell is closed. A smaller peak at $N = 4$ is due to the fulfillment of Hund's first rule: The total spin of the N-electron system is equal to zero (singlet) for $N = 2$ and 6 and to one (triplet) for $N = 4$. The

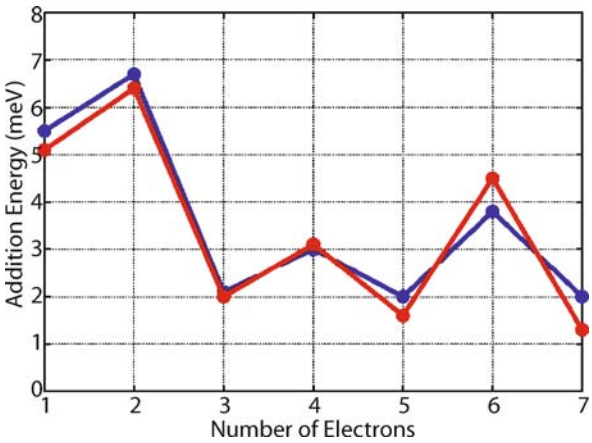

Quantum Dot Spin Transistors, Self-consistent Simulation of, Figure 5

Addition energy spectra for the circular QD at zero magnetic field. *Red line* represents the results of calculations while *blue line* stands for the experimental data

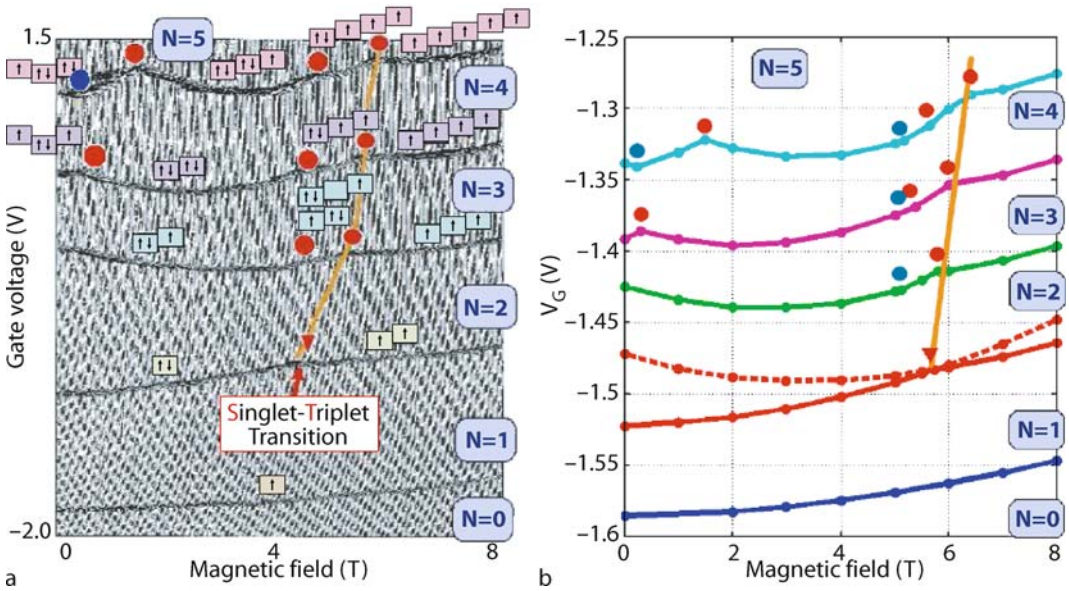

Quantum Dot Spin Transistors, Self-consistent Simulation of, Figure 6

a Experimental [15] and b calculated ground state charging diagrams in magnetic fields. Dashed curves show the lowest excited state for $N = 2$ (see also Fig. 8). *Red circles* define boundaries of regions with different total spin and angular momentum in the ground state of the N-electron system. *Blue circles* depict features in the the N-electron ground state due to reconstruction of the $(N - 1)$-electron system. *Orange curves* show the magnetic field above which the N-electron system is fully spin-polarized

agreement between the calculated and experimental spectra is very good, for both peaks and valleys. From the energy separation between the two lowest adjacent eigenvalues, the confinement strength $\hbar\omega \sim 6$ meV in the empty QD can also be deduced.

In Fig. 6 the charging voltage $V_G(N)$ for $N = 1$–5 electrons as a function of magnetic field (charging diagram) is shown. Comparison between experimental (Fig. 6a) and calculated (Fig. 6b) data shows that overall agreement is very good, albeit the confinement potential at zero magnetic field in our model structure is somewhat weaker, i. e., the curves are more closely spaced. The charging voltages generally (in the considered range) increase with magnetic field since the effective confinement becomes stronger. The curves corresponding to the charging of $N \geq 2$ electrons exhibit "cusps" due to various magnetic field induced spin and angular momentum transitions. For $N = 2$, ∇ marks the magnetic field of the singlet-triplet transition (see also Fig. 8), below which the singlet is the ground state while above the triplet is the ground state. For the $N = 3$ charging curve arising from the addition of the third electron to the two-electron system, the cusp near 5 T is due to a change in the ground state configuration of the two-electron system at $B \approx 5.6$ T (the singlet-triplet transition) which affects the addition energy of the third electron. The shift from 5.6 T to 5.0 T is due to a screening effect and is consistent with the experiment (Fig. 6a). The cusp near 6 T reflects an increase in the total spin of the three-electron system from $S = 1/2$ to $S = 3/2$, namely below this point two electrons are spin-up, and one is spin-down while above this point all three electrons in the QD are spin-up and form a spin-polarized system.

In general, the rightmost cusp in charging diagram (Fig. 6) always corresponds to complete spin polarization of the electron system. The magnetic field at which the formation of this state occurs increases with the number of electrons since a stronger field is required to overcome the large kinetic energy accompanying single occupation of consecutive orbitals. Cusps in the N-electron curves in the vicinity of $B = 5$ T are either due to changes in the $(N - 1)$-electron configuration or due to single spin rotations in the N-electron system. The two cusps around $B \approx 0.25$ T in the $N = 4$ and $N = 5$ curves mark the breakdown of Hund's first rule filling in the four-electron system and respective change in the addition energy of the fifth electron due to the decrease of exchange energy in the four electron system, similar to the $N = 3$ curve. The cusp at $B \approx 1.5$ T in the $N = 5$ curve can be understood in terms of the Fock–Darwin spectrum [59]. Around that point, an electron undergoes a transition moving to a state with higher angular momentum while keeping the total spin value constant.

Exact Diagonalization of the Many-Electron Schrödinger Equation

In order to gain insight in the single-electron tunneling characteristics in the non-linear transport regime at large source-drain bias values, it is necessary first to calculate the energy spectrum of a N-electron system. Since the above described self-consistent SDFT-based approach is insufficient for this purpose as the DFT is predominantly the ground state theory [45], we utilize the numerically exact diagonalization of the corresponding many-electron Hamiltonian:

$$\hat{H} = \sum_{i=1}^{N} \left[-\frac{\hbar^2}{2m^*} \left(\nabla_i - \frac{ie}{\hbar c} \mathbf{A}_i \right)^2 + V_{\text{conf}}(\mathbf{r}_i) \pm \frac{1}{2} g \mu_{\text{B}} B \right] + \sum_{i<j} \frac{e^2}{\epsilon |\mathbf{r_i} - \mathbf{r_j}|}, \quad (15)$$

where m^* and ϵ stand for the electron effective mass and dielectric constant in $In_{0.05}Ga_{0.95}As$ respectively. $\mathbf{A} = (1/2)(Bx, -By, 0)$ is the vector potential in the symmetric gage for the magnetic field B oriented along the z-direction. The term $\pm\frac{1}{2} g \mu_{\text{B}} B$ accounts for Zeeman splitting with $g = -0.44$ being the electron g-factor.

In this equation the confinement potential $V_{\text{conf}}(\mathbf{r})$ is assumed for simplicity to be two-dimensional as the width of the vertical QD is much smaller than the lateral extension of the electron "puddle" (see, e. g., Sect. "Vertical QD Device Structure" and [60]). Due to lithographic and nature imperfections [15,61], in the following calculations it is also assumed to be slightly elliptic, i. e.

$$V_{\text{conf}}(\mathbf{r}) = \tfrac{1}{2} m^* \left[\omega_x^2 x^2 + \omega_y^2 y^2 \right] \quad (16)$$

where the confinement energies in the x- and y-directions are taken to be $\hbar\omega_x = 5.3$ and $\hbar\omega_y = 5.65\,\text{meV}$, respectively, for comparison with experimental structures [15,40,61]. Note that it is also feasible to perform hybrid calculations [60] where the confinement potential is first obtained for an empty QD ($N = 0$) by the self-consistent method of Sect. "Self-consistent Approach" and then the electronic structure of the QD populated with N electrons is computed by numerically diagonalizing the corresponding many-particle Hamiltonian (15) with thus obtained potential. In a few-electron vertical QD structure, this separation works well because the electron system in the QD is sufficiently well isolated from the environment so that interaction between the electrons in the QD and the charges in the outside regions does not affect electron confinement strongly, and it becomes possible to keep the confinement potential frozen and independent on the electron number [60,62,63].

We diagonalize the above Hamiltonian (15) by expanding the N-electron wave function for the αth state in terms of $N \times N$ Slater determinants [64]:

$$\Psi_\alpha(N) = \sum_{i\ldots n} c^\alpha_{i\ldots n} \begin{vmatrix} \varphi_i(\mathbf{r}_1, s_{z1}) & \ldots & \varphi_n(\mathbf{r}_1, s_{z1}) \\ \vdots & & \vdots \\ \varphi_i(\mathbf{r}_N, s_{zN}) & \ldots & \varphi_n(\mathbf{r}_N, s_{zN}) \end{vmatrix}. \quad (17)$$

Here the basis wave function $\varphi_i(\mathbf{r}_j, s_{zj})$ is the product of a 2D anisotropic harmonic oscillator eigenfunction (with frequencies being adjustable parameters) and a spin wave function. Each quantum number $i = (n_x, n_y, s)$ corresponds to the set of the 2D harmonic oscillator (n_x, n_y) and spin s quantum numbers. The summation is carried over all permutations $\{i \ldots n\}$ available for the particular spin state, and the coefficients $c^\alpha_{i\ldots n}$ are determined after the minimization of $\langle \Psi_\alpha | \hat{H} | \Psi_\alpha \rangle$ for a given state which leads to the generalized eigenvalue problem with dense Hermitian matrices. We found that the harmonic oscillator frequencies smaller than the confinement strengths work best due to the fact that the Coulomb interaction tends to flatten out the effective potential [65]. In case of a 2D circular confinement, the Coulomb matrix elements can be evaluated analytically yielding four-fold series. For the anisotropic 2D potential (or in 3D case), the matrix elements are also expressed through the fourfold (sixfold in 3D) series but the auxiliary one-dimensional exponential integral evaluated numerically by means of the Gauss–Kronrod quadrature has to be used to compute their final numerical values [66].

The above method of numerically exact diagonalization of the many-electron Hamiltonian (15) also known as full configuration-interaction (CI) approach yields very accurate values for the eigenenergies and eigenfunctions provided that the single-particle basis set (17) is chosen to be large enough [67]. The downside of this method resides in the size and complexity of the Hamiltonian matrix that grow very quickly with the number of electrons ($\propto M^N$ with M being the number of the single-particle basis states). As a result, it is usually used for QDs with highly symmetric confinement potential containing a small number of electrons ($\lesssim 10$). The most crucial step in this approach is the construction of the many-particle wave function which can be achieved by using either

(1) The single-particle wave functions of the corresponding Hamiltonian or

(2) Analytic wave functions of an elementary confinement potential such as a 2D (3D) harmonic oscillator (as used in this work) or Fock–Darwin states [16,65,68] or
(3) Multi-center expansion of the single-particle wave functions [69].

Computations using method (1) are usually performed fully numerically on a grid [70] and can in principle be applied to various complex confinement potentials [71]. The disadvantage of this approach is the absence of a systematic way to build up an adequate basis set that ensures fast convergence in the computed energy values and the inherent loss of accuracy and dramatic increase in the computational time required for the numerical evaluation of the Coulomb integrals as the basis set gets larger. A hybrid approach in which the single-particle contributions to the Hamiltonian with the real confinement potential (obtained by the method described in Sect. "Self-consistent Approach" for $N = 0$) were computed numerically with the harmonic oscillator basis set while the Coulomb integrals were evaluated analytically has been recently used to simulate double QD systems [72].

Quantum Transport Model

Since the tunneling hetero-barriers in the vertical QD mesa-structure are large (with tunneling times ~ 1 ns [73]), only sequential tunneling of electrons is of interest [11,12,13]. In case of small tunneling barriers such as in lateral (planar) QD systems [74], the probability of electron co-tunneling [75] through the QD increases and should also be taken into account in the current calculations [76,77]. This indicates that in order to describe transport properties of the vertical QD, once the eigenspectrum of the N-electron system is obtained, we need to compute only the sequential current I [13]:

$$I = -e \sum_{\alpha\beta} \Gamma_{\alpha\beta} \left[P_\alpha(N) + P_\beta(N-1)\right] \left[f_{\mathrm{L}}(\mu_{\alpha\beta} - \mu_{\mathrm{L}}) - f_{\mathrm{R}}(\mu_{\alpha\beta} - \mu_{\mathrm{R}})\right] . \quad (18)$$

In this equation $\mu_{\alpha\beta} = E_\alpha(N) - E_\beta(N-1)$ is the energy difference between the αth N-electron and βth $(N-1)$-electron energies (the electro-chemical potential, cf. Eq. (1)) while the Fermi distributions $f_{\mathrm{L(R)}}$ determine the energy level occupation for electrons tunneling in the QD from the left L (right R) lead (or from the source and drain contact, respectively) with chemical potential μ_{L} (μ_{R}). $\Gamma_{\alpha\beta}$ is the effective tunneling rate, which in the random phase approximation (neglecting phase correlations between electrons in the leads and the QD) can be written as a product of the electron "bare" single-particle tunneling rate $\gamma_{\mathrm{L(R)}}$ due to the electron permeating the barriers separating the QD from the leads (see Fig. 2) and the overlap matrix element [78,79]:

$$\sum_i \left| \left\langle \Psi_\alpha(N) \left| a_i^\dagger \right| \Psi_\beta(N-1) \right\rangle 25 \right|^2 , \quad (19)$$

where the operator $a_i^\dagger$ is responsible for the creation of an electron in the QD in the ith single-particle state. The overlap matrix element is equal to unity in the case of the non-interacting particles but, in general, correlations among electrons (due to the Coulomb interaction and/or spin effects) can reduce its value down to zero.

The state occupation factors $P_\alpha(N)$ in Eq. (18) can be determined from the steady state solution of the coupled master (or rate) equations [11,12,80,81]:

$$\frac{\mathrm{d}P_\alpha(N)}{\mathrm{d}t} = -\sum_\beta \left[R_{(\alpha,N)\to(\beta,N\pm1)} P_\alpha(N) - R_{(\beta,N\pm1)\to(\alpha,N)} P_\beta(N\pm1)\right] , \quad (20)$$

where the transition rates are ($\kappa = L, R$)

$$R_{(\alpha,N)\to(\beta,N\pm1)} = \sum_{\kappa\in(L,R)} \Gamma_{\alpha\beta} f(\mu_{\alpha\beta} - \mu_\kappa) , \quad (21)$$

$$R_{(\beta,N\pm1)\to(\alpha,N)} = \sum_{\kappa\in(L,R)} \Gamma_{\alpha\beta} \left[1 - f(\mu_{\alpha\beta} - \mu_\kappa)\right] . \quad (22)$$

In the following chapter, calculations of the transport spectra for the QD with the number of electrons $N \leq 3$ and for $V_{\mathrm{SD}} = |\Delta\mu| = |\mu_{\mathrm{L}} - \mu_{\mathrm{R}}| = 5$ meV are presented. In accordance with parameters of the experimental QD structure (Sect. "Vertical QD Device Structure"), the emitter and collector barriers are assumed to be highly asymmetrical with a ratio of permeabealities $\gamma_{\mathrm{L}}/\gamma_{\mathrm{R}} \approx 30$ [11]. A large number of eigenstates (up to 24) is included in solution of the master equation (20) to describe properly degeneracy effects in the addition energies arising when $\mu_{\alpha\beta}$ become equal for specific combinations of states α and β.

Tunneling Spectroscopy of a Few-Electron QD in the Non-Linear Transport Regime

In Fig. 7 we show the energy diagram for $N = 1, 2, 3$ electrons. Only the levels with energy within 5 meV of the ground state are plotted. The diagram becomes more complex with increasing N as the interplay between confinement, electron-electron interaction, and magnetic field smears out the electronic spectra. Of special interest is the behavior of the lowest singlet $S = 0$ and triplet $S = 1$ energies for $N = 2$ as their energy difference, the exchange

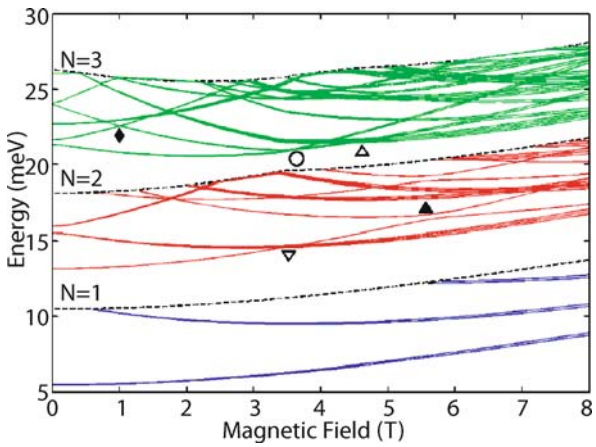

Quantum Dot Spin Transistors, Self-consistent Simulation of, Figure 7

Energy spectrum $E(N) - E_{GS}(N-1)$ for $N = 1, 2, 3$ electrons in magnetic fields. Here $E_{GS}(N)$ is the ground state energy of the N-electron system, $E_{GS}(N = 0) = 0$. *Dashed lines* are the upper boundaries of the transport windows $E_{GS}(N) + |V_{SD}|$, $|V_{SD}| = 5$ meV. Zeeman splitting of the energy levels is visible at higher magnetic fields as multiple closely separated lines. ▲ shows the magnetic field at which the exchange energy magnitude starts to decrease. All other symbols are discussed in the text

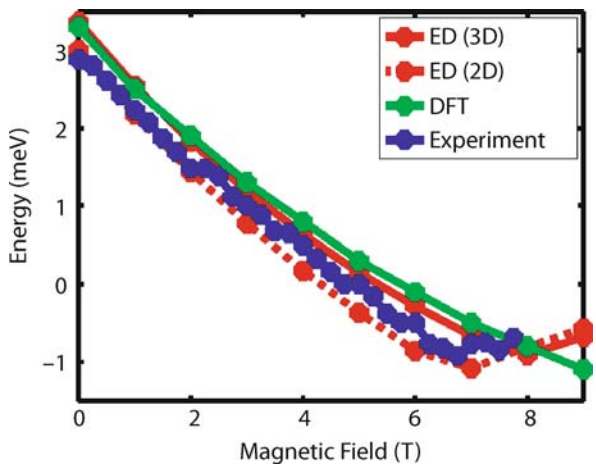

Quantum Dot Spin Transistors, Self-consistent Simulation of, Figure 8

Exchange energy J in magnetic fields. *Dashed red curve* is obtained from the ED calculations including 3D effects, *solid red curve* is the results of the 2D ED while *green curve* is for the results of the DFT calculations (Sect. "Self-consistent Approach"). *Blue curve* is the experimental data [84]

energy J, is of interest for quantum computation schemes involving double QDs [71,72,82]. The exchange energy J as a function of the applied magnetic field computed by both the numerically exact digonalization of the two-electron Hamiltonian (Sect. "Exact Diagonalization of the Many-Electron Schrödinger Equation") and the self-consistent approach based on the spin-density-functional theory (Sect. "Self-consistent Approach") is shown in Fig. 8. One can see that all three computed J-curves agree very well with the experimental data. However, the ED calculations that account for the QD finite width in the vertical dimension (3D ED) have a slight lead over the pure 2D ED results and the self-consistent SDFT-based calculations. The latter method fares progressively worse with increasing magnetic fields, probably because of the lack of the proper accounting of the magnetic field effects in the exchange-correlation potential. In particular, it fails to reproduce a "kink" in the exchange energy at ~ 7 T visible in both experimental data and ED calculations and which finds its origin in the crossing between the two singlet states with different angular momenta [83].

When magnetic fields get larger, density of the electronic states also increases due to the level compression in the lowest Landau band [39]. In the three-electron system, the transition between the two lowest doublets ($S = 1/2$) in the three-electron system (○) around 3.5 T is apparent in the energy spectrum (Fig. 7) while above ~ 4.5 T the system becomes fully spin-polarized (△) and forms a spin quartet ($S = 3/2$). Note also the transition between two doublet levels with different values of the angular momentum (♦) at about 1 T; its relationship with the current traces shown in Fig. 9 will be discussed below. In general, all states shown in Fig. 7 may be involved in the transport process; however, a host of physical reasons make most of them very difficult to distinguish in the measured transport spectra ("dark transport states").

The computed tunneling current in magnetic fields is shown in Fig. 9 for forward ($V_{SD} > 0$, Fig. 9a) and reverse ($V_{SD} < 0$, Fig. 9b) source-drain biases V_{SD}. In the former case, electrons injected in the QD through the thin (forward bias) rather than the thick (reverse bias) emitter barrier. The regions with finite current (the so-called current stripes) of width $|V_{SD}| = 5$ meV, in which the number of electrons in the QD cycles between N and $N-1$ (see Sect. "Introduction" for the general description of the sequential transport through the QD), are separated by black regions with zero current. Inside each stripe, the current variations are reflected by different colors depending on the current increase (color shifts towards the red) or decrease (color shift towards the blue). One can immediately see the striking differences between the forward and reverse bias spectra: the stripes in the latter one have a much more complex structure with a larger number of visible current features. Indeed, since in the single-elec-

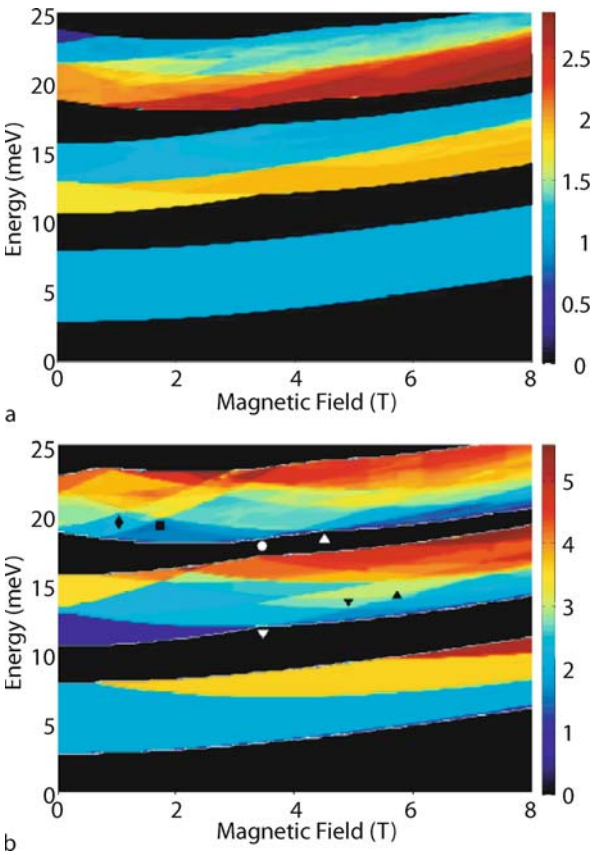

Quantum Dot Spin Transistors, Self-consistent Simulation of, Figure 9

Tunneling current (in arb. units) as a function of the gate voltage potential eV_G (vertical axis) and magnetic field B (horizontal axis) for a $V_{SD} > 0$, and b for $V_{SD} < 0$. ▲ and ▼ mark the magnetic fields at which the lowest region with the constant current and the magnitude of the exchange energy, respectively, start to decrease. The symbols ♦ and are discussed in the text while all other symbols are the same as in Fig. 7

tron tunneling process, the QD electronic configuration fluctuates between N- and $(N-1)$-electron systems, the carrier transmission depends on the detailed occupation of the many-body energy states [11,85]. Hence, in a structure with asymmetric barriers, electrons injected in the QD through the thick rather than the thin emitter barrier give rise to different electron configurations, as in the former case the dot will be in the $(N-1)$-electron state most of the time (because of the thin collector barrier allowing easy escape of electrons from the QD), while in the latter case it will be predominantly occupied by N electrons [11].

This behavior is clearly visible in the first stripe for $V_{SD} > 0$ where the current remains unchanged even though there are several single-particle levels present within the bias window (cf. Fig. 7). However, these levels becomes clearly discernable for $V_{SD} < 0$. In the limiting case of the barrier permeabealities $\gamma_L \gg \gamma_R$ with $f_{L(R)} = 0(1)$, one can find approximate solutions of the master equation (20) as $P_\alpha(1) \approx 0$, $P(0) \approx 1$ for $V_{SD} < 0$ and as $P_\alpha(1) \approx 1/M_{N=1}$, $P(0) \approx 0$ for $V_{SD} > 0$ ($M_{N=1}$ is the number of "active" states in the transport window at the given V_G and B, $\alpha = 1, \ldots, M_{N=1}$), i. e., all one-electron levels are occupied with the same probability. Substituting these occupation factors in the equation for the current (18), one can indeed see that in the former case ($V_{SD} < 0$) the current exhibits step-like increase (broadened by the temperature) when a new energy level dips below the Fermi energy as each new level represents an additional transport channel. On the other hand, when $V_{SD} > 0$, the current remains independent on the number of the levels in the transport window since the sum $\sum_{\alpha=1}^{M_{N=1}} P_\alpha$ (see Eq. (18)) remains largely unchanged.

Similar interpretation can be given for other N, although in these cases the various values of the overlap matrix elements (19) complicate the matter. In the second stripe for $V_{SD} > 0$ (Fig. 9a), the dominant feature is the sharp decrease of the current around the middle of the bias window (the sharp transition from yellow to blue). The large drop in the current is due to the non-zero occupation of the first excited $N = 1$ state which induces a redistribution of the electrons among available $N = 2$ energy levels, i. e., the occupation factors $P_\alpha(N)$, $P_\beta(N-1)$ change, so that the tunneling process associated with this state leaves a visible footprint in the transport spectrum. Physically this corresponds to the situation in which an electron tunnels out of the $N = 2$ ground state within the stripe and leaves the system with the other electron in the excited state [86]. Such process has no corresponding counterpart in the N-electron energy spectrum computed with respect to the ground $(N-1)$-electron state (Fig. 7). In the background of this large current drop, smaller changes in the current due to other processes are also visible (Fig. 9a). In particular, spin effects manifest themselves as a total spin blockade [87] of the current through specific levels when $S(N) - S(N-1) > 1/2$ or as a partial spin blocking of the transport channels for $|S_z(N) - S_z(N-1)| > 1/2$ which lead to current suppression when the corresponding state become available, i. e., the current decreases when the number of channels in transport window increases. These effects become especially clear for smaller values of the source drain bias ($V_{SD} \sim 1$ mV) when the contributions from individual transport channels become discernible in both experimental and calculated current stripes [72].

For reverse source-drain bias ($V_{SD} < 0$, Fig. 9b), the current in the second stripe tends to increase with V_G, though this increase is neither gradual nor monotonic. In

this case, the current variation due to the excited $N = 1$ state is barely noticeable, and in its absence transport processes involving other $N = 2$ energy levels become more apparent. Note that this drastic difference between the current in the two stripes (for $V_{SD} > 0$ and $V_{SD} < 0$) occurs mostly because of the large difference in the emitter and collector barrier permeabealities $\gamma_{L(R)}$. When $\gamma_L \sim \gamma_R$ the current spectrum becomes even more complex due to the mixture of these two distinct tunneling processes.

In principle, important information about the QD electronic structure – such as the behavior of the exchange energy in magnetic fields – can be extracted from the transport spectra, but caution should be exerted. For instance, if one naively assumes that the current in the Nth stripe may only change when a N-electron level enters into the transport window (Fig. 7), then the exchange energy could be simply determined by tracing the boundaries of the lowest region in the $N = 2$ stripe characterized by a constant current. This is because the exchange energy is given by the difference between the two lowest $N = 2$ electron states with different spins which are bound to be occupied first upon increasing gate voltage. By inspecting the current spectrum of Fig. 9b, one can easily see that this region starts to shrink above 5 T (▼), which, according to the above argument, should indicate the decrease in the exchange energy (in magnitude) above this magnetic field. In reality, however, this energy difference continues to increase up to 7 T (Fig. 8) and only then decreases (▲): It is due to contributions from excited $N = 1$ states at ~ 5 T for $V_{SD} = 5$.

Another example of inconsistency between the current and energy spectra can be observed in the third stripe, where a prominent kink at about 2 T (marked by ■ in Fig. 9b) has no counterpart in the energy spectrum shown in Fig. 7. Analysis of the electrochemical potentials shows that the crossing between two $\mu_{\alpha\beta}$ curves originating from the lowest $N = 2$ triplet/the $N = 3$ quartet and the lowest $N = 2$ triplet/$N = 3$ doublet at $B \sim 2$ T is responsible for this kink. This feature was also observed experimentally (Fig. 10) but was interpreted simply in terms of the crossing between the two $N = 3$ doublet levels [15]; however, the corresponding kink in the energy diagram at ~ 1 T (marked as ♦ in Figs. 7 and 9) is barely noticeable in the current spectrum.

The above examples indicate that simplistic interpretation of the transport spectra (assignment of features in the measured current) in terms of the many-particle energies can be inadequate and, in general, for the accurate analysis of the QD transport processes in the non-linear regime at finite V_{SD} [84] the quantum-mechanical overlap between various many-body states [13,78], their non-equilibrium occupancies [11,12,13] and the inherent asymmetry of the double barrier hetero-structure [11] should all be considered simultaneously.

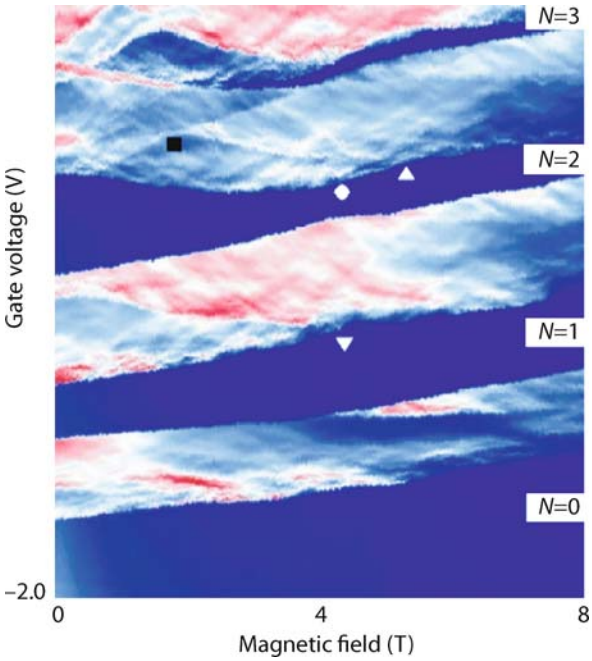

Quantum Dot Spin Transistors, Self-consistent Simulation of, Figure 10

Experimental current stripes [15]. All symbols are the same as in Figs. 7 and 9b. Overall, these early measurements do not show any noticeable features above ∼ 5 T and suffer from the extreme variations in the current due to the random dopant distribution in the leads [15]

Conclusion

While QD artificial atoms offer high potential for quantum electronics and continue to be a fast-growing area of research, they provide exciting opportunities for investigating the basic physical properties of interacting many-body system in vertical structures. Single-electron transport spectroscopy within the Coulomb blockade regime bears signatures of individual tunneling events through the QD which provide invaluable tools to investigate the interaction between incident electrons and resident particles. In the few-electron regime, a comparison between experimental electron addition energies and theoretical calculations confirms the existence of electronic shells occupied according to Hund's rules, as in real atoms. Closed shells are particularly stable as the electron addition energy to access the next shell is large, implying the existence of a "noble-gas" structure for specific numbers of electrons.

In contrast to real atoms, the confinement potential in vertical QDs is relatively weak and long-range so that the electronic structure of QDs can be easily probed by relatively small magnetic fields that induce transitions between the various electronic states as the number of particles and the strength of field are varied. In particular, transitions from a low-spin state to a completely spin-polarized system have been observed experimentally in single QDs and computed theoretically within a self-consistent approach based on SDFT.

Non-linear transport spectroscopy in QDs has been used to probe the excited states of a few-electron system. In this case excitations in the QD electronic structure are electron-only (Auger-like), unlike in real atoms where the excitations are the result of electron-photon interaction. The theoretical description of non-linear transport processes in QDs requires the simultaneous knowledge of the energy spectrum and the overlaps between various many-body states. Here, the significant increase in computational effort pays off by a good agreement between experiments and theory as well as by new insights in the electronic and transport properties of QDs.

Future Directions

Recently a novel type of artificial atoms based on electron confinement in semiconductor nanowires (quantum wires) has emerged. The principal difference from the quasi-2D QDs is the 3D nature of the electron confinement potential that bounds the particles wave functions in all three spatial dimensions. Similarly to 2D QDs, electric gates are used to confine the electron motion along the nanowire direction leading to the formation of a quantum dot in the quantum wire (QDQW). Accurate control over the electron population and the Coulomb diamonds characterizing single electron charging have been recently demonstrated in these systems [88]. From the simulation point of view, modeling QDQWs is more challenging as the geometry do not easily permit omission of the third dimension and, consequently, a simplified 2D model is inadequate. In this respect, the 3D self-consistent approach discussed in this work is more suitable as it does not rely on 2D approximation.

Within traditional 2D QD devices, more complicated systems of multiple dots, such as triple QDs, are now being investigated. They represent natural steps in creating QD networks with potential applications in single-electron logic circuits. In these devices a very large parameter space (sizes of individual QDs, number of electric gates and their positions, applied biases, etc) makes both simulation and design of the triple dot system very difficult, thereby leading to a wealth of possible electronic structure configurations such as spin density-waves and double charging of electrons [48], usually not present in single QD devices. Our preliminary results [48] also showed that in order to ensure a sizable coupling among the constituent QDs, a very careful optimization of the mesa structure physical dimensions should be performed; this can be easily achieved with the methods discussed in this article.

Acknowledgments

We are grateful to R. M. Martin and D. G. Austing for helpful discussions. This work was supported by DARPA-QuIST program, MCC through the NSF, and NCSA.

Bibliography

1. Likharev KK (1999) Proc IEEE 87:606
2. Tarucha S, Austing DG, Honda T, van der Hage RJ, Kouwenhoven LP (1996) Phys Rev Lett 77:3613
3. Elzerman JM, Hanson R, Greidanus JS, van Beveren LHW, De Franceschi S, Vandersypen LMK, Tarucha S, Kouwenhoven LP (2003) Phys Rev B 67:161308(R)
4. Millikan RA (1911) Phys Rev 32:349
5. Gorter CJ (1951) Physica (Amsterdam) 17:777
6. Giaever I, Zeller HR (1968) Phys Rev Lett 20:1504; Zeller HR, Giaever I (1969) Phys Rev 181:789
7. Lambe J, Jaklevic RC (1969) Phys Rev Lett 22:1371
8. Kulik IO, Shekhter RI (1975) Zh Eksp Teor Fiz 68:623; Sov Phys JETP 41:308
9. Kuzmin LS, Likharev KK (1987) Pis'ma Zh Eksp Teor Fiz 45:250; Lett JETP 45:495
10. Fulton TA, Dolan GJ (1987) Phys Rev Lett 59:109
11. Averin DV, Korotkov AN, Likharev KK (1991) Phys Rev B 44:6199; Su B, Goldman VJ, Cunnigham JE (1992) ibid 46:7644
12. Beenakker CW (1991) Phys Rev B 44:1646
13. Kinaret JM, Meir Y, Wingreen NS, Lee PA, Wen X-G (1992) Phys Rev B 46:4681
14. Kastner MA (1993) Phys Today 46(1):24
15. Kouwenhoven LP, Oosterkamp TH, Danoesastro MWS, Eto M, Austing DG, Honda T, Tarucha S (1997) Science 278:1788
16. Kyriakidis J, Ladriere PM, Ciorga M, Sachrajda AS, Hawrylak P (2002) Phys Rev B 66:035320
17. Hanson R, Vandersypen LMK, Willems LH, van Beveren Elzerman JM, Vink IT, Kouwenhoven LP (2004) Phys Rev B 70:241304(R)
18. Zumbühl DM, Marcus CM, Hanson MP, Gossard AC (2004) Phys Rev Lett 93:256801
19. Ellenberger C, Ihn T, Yannouleas C, Landmann U, Ensslin K, Driscoll D, Gossard AC (2006) Phys Rev Lett 96:126806
20. Landau LD, Lifshitz E (1977) Quantum mechanics: non-relativistic theory. Pergamon Press, Oxford
21. Goeppert-Mayer M (1949) Phys Rev 75:1969
22. Ekardt W (1999) Metal clusters. Wiley, New York
23. Sikorski C, Merkt U (1989) Phys Rev Lett 62:2164
24. Meurer B, Heitmann D, Ploog K (1992) Phys Rev Lett 68:1371
25. Brunner K, Bockelmann U, Abstreiter G, Walther M, Böhm G, Gänkle T, Weimann G (1992) Phys Rev Lett 69:3216
26. Maksym PA, Chakraborty T (1990) Phys Rev Lett 65:108

27. Kohn W (1959) Phys Rev 115:1160; (1961) ibid 123:1242
28. Gudmundsson V, Gerhardts RR (1991) Phys Rev B 43:12098
29. Smith TP III, Lee KY, Knoedler CM, Hong JM, Kern DP (1988) Phys Rev B 38:2172
30. Hansen W, Smith TP III, Lee KY, Brum JA, Knoedler CM, Kern DP (1989) Phys Rev Lett 62:2168
31. Silsbee RH, Ashoori RC (1990) Phys Rev Lett 64:1991
32. Ashoori RC, Silsbee RH, Pfeiffer LN, West KW (1992) Nanostructures and mesoscopic systems. In: Reed M, Kirk W (eds) Proceedings of the int. symp., Santa Fe, May, 1991. Academic, San Diego
33. Ashoori RC, Störmer HL, Weiner JS, Pfeiffer LN, Pearton SJ, Baldwin KW, West KW (1992) Phys Rev Lett 68:3088
34. Meirav U, Kastner MA, Wind SJ (1990) Phys Rev Lett 65:771
35. Ashoori RC, Störmer HL, Weiner JS, Pfeiffer LN, Baldwin KW, West KW (1993) Phys Rev Lett 71:613
36. Bryant GW (1987) Phys Rev Lett 59:1140
37. Kumar A, Laux SE, Stern F (1990) Phys Rev B 42:5166
38. Macucci M, Hess K, Iafrate GJ (1993) Phys Rev B 48:17354; Macucci M, Hess K, Iafrate GJ (1995) Appl Phys J 77:3267
39. Reimann SM, Koskinen M, Kolehmainen J, Manninen M, Austing DG, Tarucha S (1999) Eur Phys J D 9:105
40. Matagne P, Leburton J-P, Austing DG, Tarucha S (2001) Phys Rev B 65:085325
41. Stopa M (1996) Phys Rev B 54:13767
42. Bednarek S, Szafran B, Adamowski J (2001) Phys Rev B 64:195303
43. Szafran B, Bednarek S, Adamowski J (2001) Phys Rev B 65:035316
44. Austing DG, Honda T, Tarucha S (1996) Semicond Sci Technol 11:388
45. Jones RO, Gunnarsson O (1989) Rev Mod Phys 61:689
46. Zhang L-X, Matagne P, Leburton J-P, Hanson R, Kouwenhoven LP (2004) Phys Rev B 69:245301
47. Ravishankar R, Matagne P, Leburton J-P, Martin RM, Tarucha S (2004) Phys Rev B 69:035326
48. Kim J, Melnikov DV, Leburton J-P, Austing GD, Tarucha S (2006) Phys Rev B 74:035307
49. Perdew JP, Wang Y (1992) Phys Rev B 45:13244
50. Ancilotto F, Austing DG, Barranco M, Mayol R, Muraki K, Pi M, Sasaki S, Tarucha S (2003) Phys Rev B 67:205311
51. Vignale G, Rasolt M (1987) Phys Rev Lett 59:2360
52. Sze SM (1981) Physics of semiconductor devices, 2nd edn. Wiley Interscience, New York
53. Bathe K-J (1982) Finite element procedures in engineering analysis. Prentice Hall, New Jersey
54. Fettig J, Melnikov DV, Sobh N, Leburton J-P (unpublished)
55. Slater JC (1972) Adv Quantum Chem 6:1
56. Fonseca LRC, Jimenez JL, Leburton J-P, Martin RM (1998) Phys Rev B57:4017
57. Lee I-H, Ahn K-H, Kim Y-H, Martin RM, Leburton J-P (1999) Phys Rev B 60:13720
58. Austing DG, Sasaki S, Tarucha S, Reimann SM, Koskinen M, Manninen M (1999) Phys Rev B 60:11514
59. Reimann S, Manninen MM (2002) Rev Mod Phys 74:1283
60. Melnikov DV, Leburton J-P (2006) Phys Rev B 73:155301
61. Kouwenhoven LP, Austing DG, Tarucha S (2001) Rep Progr Phys 64:701
62. Bruce NA, Maksym PA (2000) Phys Rev B 61:4718
63. Szafran B, Bednarek S, Adamowski J (2003) Phys Rev B 67:115323
64. Mikhailov SA (2002) Phys Rev B 65:115312
65. Reimann SM, Koskinen M, Manninen M (2000) Phys Rev B 62:8108
66. Drouvelis PS, Schmelcher P, Diakonos FK (2004) Phys Rev B 69:155312
67. Chakraborti T (1999) Quantum dots: Survey A of the properties of artificial atoms. North Holland, Amsterdam
68. Saarikoski H, Harju A (2005) Phys Rev Lett 94:246803
69. Szafran B, Peeters FM, Bednarek S, Adamowski J (2004) Phys Rev B 69:125344
70. Bellucci D, Rontani M, Troiani F, Goldoni G, Molinari E (2004) Phys Rev B 69:201308(R)
71. Stopa M, Marcus CM (2008) Nano Lett 8:1778
72. Melnikov DV, Leburton J-P, Taha A, Sobh N (2006) Phys Rev B 74:041309(R)
73. Fujisawa T, Austing DG, Tokura Y, Hirayama Y, Tarucha S (2002) Nature 419:278
74. De Franceschi S, Sasaki S, Elzerman JM, van der Wiel WG, Tarucha S, Kouwenhoven LP (2001) Phys Rev Lett 86:878
75. Averin DV, Odintsov AA (1989) Phys Lett A 140:251
76. Sukhorukov EV, Burkard G, Loss D (2001) Phys Rev B 63:125315
77. Thielmann A, Hettler MH, König J, Schön G (2005) Phys Rev Lett 95:146805
78. Pfannkuche D, Ulloa SE (1995) Phys Rev Lett 74:1194
79. Palacios JJ, Martín-Moreno L, Chiappe G, Louis E, Tejedor C (1994) Phys Rev B50:R5760
80. Bonet E, Deshmukh MM, Ralph DC (2002) Phys Rev B 65:045317
81. Muralidharan B, Ghosh AW, Datta S (2006) Phys Rev B 73:155410
82. Petta JR, Johnson AC, Taylor JM, Laird EA, Yacoby A, Lukin MD, Marcus CM, Hanson MP, Gossard AC (2005) Science 309:2180
83. Wagner M, Merkt U, Chaplik AV (1992) Phys Rev B 45:R1951
84. Melnikov DV, Fujisawa T, Austing DG, Tarucha S, Leburton J-P (2008) Phys Rev B 77:165340
85. Grabert H, Devoret MH (eds) (1992) Single electron tunneling, Series ASINATO, Series B, vol 294. Plenum Press, New York
86. Agam O, Wingreen NS, Altshuller BA, Ralph DC, Tinkham M (1997) Phys Rev Lett 78:1956
87. Weinmann D, Häusler W, Kramer B (1995) Phys Rev Lett 74:984
88. Björk MT, Thelander C, Hansen AE, Jensen LE, Larsson MW, Wallenberg LR, Samuelson L (2004) Nano Lett 4:1621

Quantum Error Correction and Fault Tolerant Quantum Computing

Markus Grassl[1], Martin Rötteler[2]

[1] Institute for Quantum Optics and Quantum Information, Austrian Academy of Sciences, Innsbruck, Austria

[2] NEC Laboratories America, Inc., Princeton, USA

Article Outline

Glossary
Definition of the Subject
Introduction

Basics of Classical Error Correction
Basic Ideas of Quantum Error Correction
A Simple Example and Shor's Nine-Qubit Code
Conditions for Quantum Error Correction
Quantum Codes from Classical Codes
Techniques for Fault-Tolerant Quantum Computing
The Threshold Theorem
Further Aspects
Bibliography

Glossary

Ancilla qubits Refers to qubits that are used to facilitate quantum computations, but serve neither as input nor as output of the computation. The two main uses of ancilla qubits are a) for the purpose of a scratch pad during a quantum computation, i. e., states that are initialized in a known state and can be used during a computation such that they are returned to the initial state at the end, and b) for the purpose of quantum error correction where they serve as space for holding the error syndrome.

Bit-flip error An error that affects a single qubit and interchanges the basis states of that qubit.

CSS code A special class of quantum codes named after their inventors Calderbank, Steane, and Shor. CSS codes allow to construct quantum codes from certain classical error-correcting codes.

CNOT gate The controlled-NOT gate is an example of a quantum gate. It is a two qubit gate and – together with single qubit gates – can be shown to be a universal gate for quantum computation.

Error syndrome A classical bit-vector that describes a quantum error that has affected a quantum code. Contrary to the classical case where there is a one-to-one correspondence between syndromes and errors, in the quantum case a syndrome does in general not uniquely identify a quantum error.

Fault-tolerant quantum computing The discipline that studies how to perform reliable quantum computations with imperfect hardware.

Phase-flip error An error that affects a single qubit and gives a relative phase of -1 to the basis states of that qubit. Contrary to the bit-flip error, this error has no classical analog.

Quantum channel A term describing a general physical operation that can affect the state of a quantum-mechanical system. Another name for quantum channels are completely-positive trace preserving maps.

Quantum error-correcting code (QECC) A method to introduce redundancy to a quantum mechanical system in such a way that certain errors that are affecting the system can be detected or corrected.

Quantum circuits Operations that a quantum computer can carry out. Quantum circuits are composed of quantum gates and, when acting on an n qubit system, correspond to unitary matrices of size $2^n \times 2^n$.

Quantum gate A basic operation that a quantum computer can carry out. Common choices for quantum gates are certain unitary matrices of size 2×2 (single qubit gates) or 4×4 (two qubit gates).

Qubit Shortened form of *quantum bit*. A physical object that can support states in a two-dimensional complex Hilbert space. Contrary to the classical case where a bit refers to both, the physical object and the states it can hold, a qubit refers to the physical object only. Qubits are the basic units of a quantum computer's memory.

Stabilizer code A class of quantum codes that is defined as the joint eigenspace of a group of commuting operators. Stabilizer codes give rise to many examples of quantum codes, comprise the class of CSS codes as a special case, and are equivalent to additive codes over $GF(4)$ that are self-orthogonal with respect to the Hermitian inner product.

Threshold theorem An important result of the theory of fault-tolerant quantum computing, stating that there is a threshold value for this noise level such that arbitrarily long quantum computations become possible if the gates have a noise level that is under the threshold.

Transversal gates Special class of operations acting on encoded quantum data. Transversal gates have the important property to exhibit benign behavior in case errors happen during the application of the gate.

Definition of the Subject

Quantum error correction offers a solution to the problem of protecting quantum systems against noise induced by interactions with the environment or caused by imperfect control of the system. The need for error correction arises not only in communication, when quantum information is sent over some distance, but also in locally, when storing and processing quantum information. Fault-tolerant quantum computing builds on quantum error correction and denotes techniques that allow computations to be performed on a quantum system with faulty gates as well as storage errors. Without mechanisms for quantum error correction and fault-tolerance, quantum computing would be impossible even for moderate error rates.

The idea of quantum error correction was first conceived in a paper [64] by Shor in 1995 in which a particular quantum code was given that encodes one quan-

tum bit (qubit) into nine quantum bits, while being able to correct against one arbitrary error on one of these nine qubits. Before long, Bennett et al. [11] developed a theory of error correction describing a quantum code that encodes one quantum bit into five while still being able to correct against arbitrary single qubit errors. Around the same time, Calderbank, Shor [15] and Steane [66] constructed quantum codes from suitable pairs of classical codes. Named after their inventors, this important class of error correcting codes is called CSS codes.

Other noteworthy work in the early days of quantum error correction were the contributions by Calderbank et al. [14], which related quantum codes to classical codes over the finite field $GF(4)$ that are additively closed and self-orthogonal with respect to the Hermitian inner product, and by Gottesman [25,26], who developed the theory of stabilizer codes. These two constructions are equivalent.

As far as fault-tolerant quantum computing is concerned, again the first method was given by Shor [65] who introduced the idea of performing a universal set of quantum gates fault-tolerantly, including syndrome measurements required for quantum error correction. Around the same time, ideas based on concatenation of quantum codes were used by Aharonov and Ben-Or [2], Kitaev [42], Gottesman [26], Knill, Laflamme and Zurek [45], and Preskill [58], to derive fault-tolerant schemes that work even if only faulty gates are available. The characteristic feature of all these constructions is that there is a certain value for the error rate such that if the actual error rate is below this value, then arbitrarily long quantum computations can be performed with any desired accuracy, whereas if the actual error rate is above this value, then quantum computing is impossible with the same scheme. This watershed type behavior is characterized in the threshold theorem which is a central result in fault-tolerant quantum computing. As the task to determine the exact value of the threshold is daunting, much of work has been done to get upper and lower bounds as well as to perform numerical simulations that give indications of the threshold value.

Introduction

In the early days of quantum computing, Haroche and Raimond asked the poignant question whether the dream of quantum computing could ever be realized in a real physical system or if "the large-scale quantum machine ... is the experimenter's nightmare" [39]. At the time the article was written, the first quantum error-correcting code had just been proposed [64]. However, Haroche and Raimond argued that "the implementation of error-correcting codes will become exceedingly difficult" given any detection efficiency less than 100%. It was only later that it was shown that even with imperfect quantum memory and imperfect quantum operations it is possible to implement an arbitrarily long quantum computation, provided that the failure probability of each element is below a certain threshold [2,45].

Here we provide an overview of the ingredients leading to fault-tolerant quantum computation (FTQC). In the first part, we present the theory of quantum error-correcting codes (QECCs) and, in particular, two important classes of QECCs: CSS codes and stabilizer codes. Both are related to classical error-correcting codes, so we start with some basics from this area. In the second part of the article, we present a high-level view of the main ideas of FTQC and the threshold theorem.

For the background of quantum computing in general, we refer the reader to the related articles in this volume as well as the book by Nielsen and Chuang [56].

Basics of Classical Error Correction

Block Codes

Before discussing the principles of quantum error correction, we briefly summarize some results from classical error correction. For more information see MacWilliams and Sloane [53]. Error correction is part of information theory whose foundations were laid by Claude Shannon in his landmark paper "A Mathematical Theory of Communication" [63]. In that paper, Shannon introduced the basic mathematical concepts for communication systems that are used to send messages from one point in space or time to another point:

> "*The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.*"

Hence we can model the set of all possible messages, for example, by a set of binary strings of fixed length.

Definition 1 (block code) A *block code* B of length n is a subset of all possible sequences of length n over an al-

phabet $\mathcal{A}$, i. e., $B \subseteq \mathcal{A}^n$. The *rate*

$$R = \frac{\log |B|}{\log |\mathcal{A}^n|} = \frac{\log |B|}{n \log |\mathcal{A}|}$$

of the code is a measure of the amount of information that is transmitted per symbol in the code.

While the rate of the code should be as high as possible, using only a proper subset of all possible messages allows us to correct some errors. On the highest level of abstraction, one distinguishes only whether a symbol is transmitted correctly or not. This yields to the following measure for the distance between two sequences of length n.

Definition 2 (Hamming distance) The *Hamming distance* between two sequences $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ equals the number of positions where $\mathbf{x}$ and $\mathbf{y}$ differ, that is,

$$d_{\mathrm{H}}(\mathbf{x}, \mathbf{y}) := |\{i \colon 1 \leq i \leq n \mid x_i \neq y_i\}| \,.$$

If the alphabet contains a symbol zero, the *Hamming weight* of a sequence is defined as the number of non-zero elements of the sequence.

If we send two copies of every symbol to the receiver, we get a *repetition code* of length $n = 2$ and rate $R = 1/2$. The Hamming distance between any two codewords is two. In this situation, the receiver can detect an error if at most one of the symbols has been corrupted during transmission. If we repeat every symbol three times ($n = 3$, $R = 1/3$), the receiver can even deduce the correct message from the majority of the symbols, assuming that no more than one symbol is erroneous. The general situation is as follows:

Theorem 1 *Let B be a block with minimum distance d, that is, any two codewords differ in at least d positions. Then one can either detect errors that affect less than d positions or correct errors that affect strictly less than $d/2$ positions.*

Proof As the minimum distance of the code is d, changing up to $d - 1$ positions of a codeword does not yield another codeword. So in order to detect up to $d - 1$ errors, it is sufficient to check whether the received word is a codeword or not. In order to correct errors, note that the spheres of radius $\lfloor (d-1)/2 \rfloor$ around any codeword are disjoint. In the error correction process, every word in that sphere will be mapped to the corresponding codeword. □

In this setting, finding a good error-correcting code, by which we mean a code with both high rate and high minimum distance, amounts to packing spheres of words of length n with radius $\lfloor (d-1)/2 \rfloor$ with respect to Hamming distance. In general, the resulting code will be a set of codewords without further structure so that to succeed we would have to store the list of all codewords.

Linear Block Codes

Developing more efficient ways of describing all codewords and testing whether a given word lies in the code requires that the code has some additional structure. First, we assume that the alphabet of the code is a finite field $GF(q)$ with q elements. Second, we require that any linear combination of two codewords is again a codeword. The resulting code is a linear block code, denoted by $C = [n, k, d]$ (for more details see [53]). Here k is the dimension of the code as a subspace of the vector space $GF(q)^n$, and d denotes the minimum distance of the code. Instead of listing all q^k codewords, it is sufficient to specify a basis with k linearly independent vectors of the linear space. Alternatively, a subspace of dimension k is uniquely described as the space of solutions of $n - k$ linearly independent homogeneous linear equations.

Definition 3 (generator matrix/parity check matrix) A *generator matrix* of a linear block code $C = [n, k, d]$ is a matrix G with k rows and n columns of full rank such that the row span of G equals the code C. A *parity check matrix* is a matrix H with $n - k$ rows and n columns of full rank such that the row-nullspace of H equals the code C.

The generator matrix G provides both a compact description of a linear block code and an efficient way of encoding a message which can be represented by a vector $\mathbf{i}$ of length k. The corresponding codeword is obtained by the linear mapping $\mathbf{i} \mapsto \mathbf{c} := \mathbf{i}G$. The parity check matrix provides an efficient way to detect errors.

Theorem 2 (error syndrome) *Let H be a parity check matrix of the linear block code $C = [n, k, d]$. Then a vector $\mathbf{v}$ is a codeword if and only if the error syndrome $\mathbf{s} := \mathbf{v}H^t$ is zero. Furthermore, for an erroneous codeword $\mathbf{v} = \mathbf{c} + \mathbf{e}$ the error syndrome depends only on the error $\mathbf{e}$, but not on the codeword $\mathbf{c}$.*

Proof By definition, the code C equals the row-nullspace of the parity check matrix H. Furthermore, any erroneous codeword can be written as $\mathbf{v} = \mathbf{c} + \mathbf{e}$. Then we have

$$\mathbf{s} = \mathbf{v}H^t = (\mathbf{c} + \mathbf{e})H^t = \mathbf{c}H^t + \mathbf{e}H^t = \mathbf{e}H^t \,.$$

□

Later we will need the concept of the *dual code* which also motivates the definition of the parity check matrix with n columns and $n - k$ rows.

Proposition 1 (dual code) *Let $C = [n, k, d]$ be a linear block code. Then the (Euclidean) dual code $C^\perp$ is the space of all vectors that are orthogonal to all codewords with respect to the Euclidean inner product $\mathbf{v} \cdot \mathbf{w} := \sum_{i=1}^n v_i w_i$, i. e.*

$$C^\perp = \{\mathbf{v}\colon \mathbf{v} \in GF(q)^n \mid \mathbf{v} \cdot \mathbf{c} = 0 \quad \textit{for all} \ \ \mathbf{c} \in C\} .$$

If G is a generator matrix and H is a parity check matrix for C, then G is a parity check matrix and H is a generator matrix for $C^\perp$, i. e., the role of G and H is interchanged.

The parity check matrix can also be used to compute the minimum distance of a linear block code.

Theorem 3 *If any $d-1$ columns of the parity check matrix H of a linear block code C are linearly independent, then the minimum distance of the code is at least d.*

Proof First note that for a linear code, the minimum distance equals the minimum Hamming weight of a nonzero codeword as $d_H(\mathbf{x}, \mathbf{y}) = d_H(\mathbf{x} - \mathbf{y}, \mathbf{0})$. Assume that $\mathbf{c}$ is a codeword of Hamming weight $d-1$, that is, there are $d-1$ indices $i_1, \ldots, i_{d-1}$ such that $c_{i_j} \neq 0$. The syndrome is computed as $0 = \mathbf{c}H^t = c_{i_1} h^{(i_1)} + \cdots + c_{i_{d-1}} h^{(i_{d-1})}$, where $h^{(i)}$ denotes the ith column of H. Hence we have a non-trivial linear combination of $d-1$ columns of H that is zero, contradicting the fact that any $d-1$ columns of H are linearly independent. □

Hamming Codes

The previous theorem can be used to construct codes with a prescribed minimum distance d. For a code that can correct a single error, the minimum distance d must be at least three, that is, any two columns of the parity check matrix must be linearly independent. For the simplest case in which the field has two elements $GF(2) = \{0, 1\}$, that is, all operations are modulo two, it is sufficient that all columns of H are distinct and nonzero. This yields the following family of single error-correcting codes (see [38,53]).

Proposition 2 (binary Hamming code) *The rth binary Hamming code is a linear binary block code of length $n = 2^r - 1$, dimension $k = 2^r - r - 1$, and minimum distance $d = 3$. A parity check matrix H of the Hamming code is a matrix whose $2^r - 1$ columns are all nonzero binary vectors of length r.*

The columns of H can be arranged such that the ith column equals the binary expansion of i. Then for an error $\mathbf{e}$ of weight one, the syndrome $\mathbf{e}H^t$ equals the binary expansion of the position of the error. For $r = 3$, we get the following parity check matrix

$$H = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix} .$$

For the error $\mathbf{e} = (0, 0, 1, 0, 0, 0, 0)$, the syndrome is $\mathbf{e}H^t = (0, 1, 1)$, that is, the binary expansion of three.

In general it is difficult to deduce the error of smallest Hamming weight—which is often the most likely error—from the error syndrome. More precisely, given a binary parity check matrix H, an error syndrome $\mathbf{s}$, and a positive integer w it is NP complete to decide whether there is an error vector $\mathbf{e}$ whose weight does not exceed w such that $\mathbf{e}H^t = \mathbf{s}$ [12]. Nonetheless, for some classes of codes, such as BCH codes or Reed–Solomon codes, there exist efficient algorithms for the correction of all errors up to a certain weight (see [53]).

Basic Ideas of Quantum Error Correction

In the classical setting of error correction, information is represented, for example, by a binary string of length n. The extremal case is that we have only one bit that is either zero or one. In the context of quantum information, the simplest quantum system is modeled by a two-dimensional complex vector space. A *quantum bit* (or *qubit*, for short) corresponds to a normalized vector in this space (for more details see the book by Nielsen and Chuang [56]). The qubit is given by

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle , \quad \text{where} \ \ |\alpha|^2 + |\beta|^2 = 1 , \ \ \alpha, \beta \in \mathbb{C} .$$

Here $|0\rangle$ and $|1\rangle$ denote two orthogonal vectors in the two-dimensional vector space. While these basis states correspond to the two classical values 0 and 1 of a bit, a qubit can be in a *superposition* of both $|0\rangle$ and $|1\rangle$. When the qubit is measured with respect to the basis $\{|0\rangle, |1\rangle\}$, the result is either "0" or "1" with probability $|\alpha|^2$ and $|\beta|^2$, respectively.

As we have seen in Subsect. "Block Codes", a simple classical one-error correcting code can be obtained by sending the information three times and taking a majority decision at the receiver's end. However, this does not work in the context of quantum information. First, it is not possible for the sender to compute copies of an unknown quantum state (see [69]). The main idea of this so-called *no-cloning theorem* is as follows.

Theorem 4 *There is no quantum operation that maps an arbitrary quantum state $|\psi\rangle$ and a fixed state $|\phi_0\rangle$ to two independent copies $|\psi\rangle|\psi\rangle$.*

Classical codes
Partition of the set of all words of length n over an alphabet of size 2.

Quantum codes
Orthogonal decomposition of the vector space $\mathcal{H}^{\otimes n}$, where $\mathcal{H} \cong \mathbb{C}^2$.

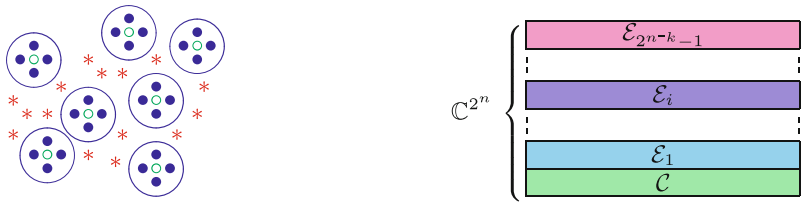

○ codewords
● errors of bounded weight
∗ other errors

$$\mathcal{H}^{\otimes n} = \mathcal{C} \oplus \mathcal{E}_1 \oplus \ldots \oplus \mathcal{E}_{2^{n-k}-1}$$

Quantum Error Correction and Fault Tolerant Quantum Computing, Figure 1
Similarities between classical and quantum error-correcting codes

Proof Let $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$. Two independent copies of $|\psi\rangle$ are given by

$$\begin{aligned}|\psi\rangle|\psi\rangle &= \big(\alpha|0\rangle + \beta|1\rangle\big) \otimes \big(\alpha|0\rangle + \beta|1\rangle\big) \\ &= \alpha^2|00\rangle + \alpha\beta|01\rangle + \alpha\beta|10\rangle + \beta^2|11\rangle\,.\end{aligned}$$

If $|\psi\rangle$ is one of the basis states, we have $|0\rangle|\phi_0\rangle \mapsto |00\rangle$ and $|1\rangle|\phi_0\rangle \mapsto |11\rangle$. By the linearity of quantum mechanics, starting with a superposition we get $\big(\alpha|0\rangle + \beta|1\rangle\big)|\phi_0\rangle \mapsto \alpha|00\rangle + \beta|11\rangle$. This equals $|\psi\rangle|\psi\rangle$ only when $\alpha = 0$ or $\beta = 0$. □

Second, even if independent copies of a quantum state $|\psi\rangle$ are sent (for example, if the sender knows how to prepare the state $|\psi\rangle$), the direct quantum mechanical analogue of a majority decision is not possible for the receiver. Instead, the receiver may, for example, check whether the joint quantum state of all received copies is invariant under permutation of the copies. This allows us to detect and correct some errors [9].

Although the classical repetition code does not have a direct quantum mechanical analogue, we will see that quantum error correction can be related to classical error correction codes. For this we recall that, in general, a classical code is given by a proper subset of the finite set of all possible messages of fixed length. In contrast, quantum information is represented by an arbitrary normalized vector in a complex vector space. In order to achieve the possibility of correcting quantum errors, a quantum error-correcting code must be a proper subspace of a larger vector space. Similar to the problem of packing spheres for classical codes, the subspace has to be chosen in such a way that the spaces corresponding to the quantum errors do not overlap (see Fig. 1). Before addressing this question in more detail in Sect. "Conditions for Quantum Error Correction", we present a simple example.

A Simple Example and Shor's Nine-Qubit Code

The Three-Qubit Code

The shortest classical code that encodes one bit and can correct one error is the repetition code of length three which coincides with the second binary Hamming code. Hence a parity check matrix H and a generator matrix G are given by

$$H = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix} \quad \text{and} \quad G = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}\,.$$

We use the codewords (000) and (111) of the binary code to define the basis states of a three-qubit code, that is, the encoding operation is given by

$$\begin{aligned} \mathsf{C}\colon\ \mathbb{C}^2 &\to (\mathbb{C}^2)^{\otimes 3} \\ |0\rangle &\mapsto |000\rangle \\ |1\rangle &\mapsto |111\rangle\,. \end{aligned}$$

Hence a superposition $\alpha|0\rangle + \beta|1\rangle$ is encoded as $\mathsf{C}(\alpha|0\rangle + \beta|1\rangle) = \alpha|000\rangle + \beta|111\rangle$. The encoding transformation C can be implemented using two controlled-NOT (CNOT) gates (see [56]). The CNOT gate is given by

$$\text{CNOT}\colon |x\rangle|y\rangle \mapsto |x\rangle|x \oplus y\rangle\,,$$

where $x \oplus y$ denotes addition modulo two (XOR). Hence the second (target) qubit is flipped if and only if the first (control) qubit is one. The states $|000\rangle$ and $|111\rangle$ span the

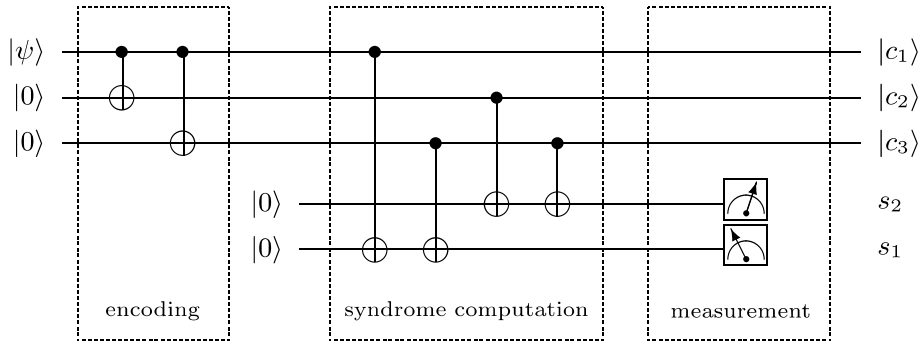

Quantum Error Correction and Fault Tolerant Quantum Computing, Figure 2
Quantum circuit for the three-qubit code. Two ancilla qubits are used for the encoding, and another two for the error syndrome. The measurement yields two classical syndrome bits

quantum code C which is a two-dimensional subspace of $\mathbb{C}^2 \otimes \mathbb{C}^2 \otimes \mathbb{C}^2$. A classical error flips a bit, interchanging 0 and 1. The quantum mechanical analogue is given by the matrix

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

interchanging the basis states $|0\rangle$ and $|1\rangle$. Applying σ_x to at most one of the three subsystems we obtain the following states:

error	state	subspace
no error	$\alpha\|000\rangle + \beta\|111\rangle$	$(I \otimes I \otimes I)\, C =: C_0$
1st position	$\alpha\|100\rangle + \beta\|011\rangle$	$(\sigma_x \otimes I \otimes I)\, C =: C_1$
2nd position	$\alpha\|010\rangle + \beta\|101\rangle$	$(I \otimes \sigma_x \otimes I)\, C =: C_2$
3rd position	$\alpha\|001\rangle + \beta\|110\rangle$	$(I \otimes I \otimes \sigma_x)\, C =: C_3$

(1)

The four different cases yield four mutually orthogonal subspaces, that is, the Hilbert space of three qubits can be decomposed as follows:

$$\mathbb{C}^2 \otimes \mathbb{C}^2 \otimes \mathbb{C}^2 = (I \otimes I \otimes I)\, C \oplus (\sigma_x \otimes I \otimes I)\, C \oplus (I \otimes \sigma_x \otimes I)\, C \oplus (I \otimes I \otimes \sigma_x)\, C\,.$$

In principle it is possible to construct a quantum mechanical observable whose eigenspaces are the four two-dimensional spaces C_i in (1). The corresponding projective measurement projects onto one of these spaces and provides information about the error, but preserves the superposition within the spaces. Alternatively, we can compute information about the error using two auxiliary qubits (ancillae). Recall that for the binary Hamming code, the error syndrome $\mathbf{s} = \mathbf{e}H^t$ equals the binary expansion of the position of the error, provided that there is at most one error. The computation of the error syndrome can also be implemented using CNOT gates (see Fig. 2). The gates in the box labeled "syndrome computation" implement the transformation

$$|\mathbf{c}\rangle|00\rangle = |c_1c_2c_3\rangle|0\rangle|0\rangle \mapsto |c_1c_2c_3\rangle|c_2 \oplus c_3\rangle|c_1 \oplus c_3\rangle = |\mathbf{c}\rangle|\mathbf{c}H^t\rangle\,. \quad (2)$$

Measuring the two ancilla qubits yields a two-bit syndrome $\mathbf{s} = (s_2s_1)$ encoding the position of the error. This classical information can be used to correct the error. Instead of measuring the syndrome qubits, one can use controlled quantum operations to correct the errors as illustrated in Fig. 3.

While this three-qubit code allows us to correct the quantum mechanical analogue of a single bit-flip error, it cannot correct an arbitrary single qubit error. For instance, measuring any of the three qubits with respect to the standard basis $\{|0\rangle, |1\rangle\}$ for the encoded state $|\Phi\rangle = \alpha|000\rangle + \beta|111\rangle$ has the same effect as measuring the unencoded qubit $\alpha|0\rangle + \beta|1\rangle$. In order to turn the three-qubit code into a code that can correct the effect of measuring one qubit, we first use the following identities for projection onto the basis states:

$$|0\rangle\langle 0| = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \frac{1}{2}(I + \sigma_z) \quad \text{and}$$
$$|1\rangle\langle 1| = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \frac{1}{2}(I - \sigma_z)\,, \quad (3)$$

where $\sigma_z = \left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$. The relation between σ_x and σ_z is given by $\sigma_z = H\sigma_x H$, where $H = 1/\sqrt{2}\left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$. Hence the Hadamard matrix H turns bit-flip errors σ_x into phase-flip errors σ_z. Since the three-qubit code C is able to correct a single bit-flip error, the code $C^{\text{phase}} = (H \otimes H \otimes H)\, C$ can correct a single phase-flip error. By linearity, this code

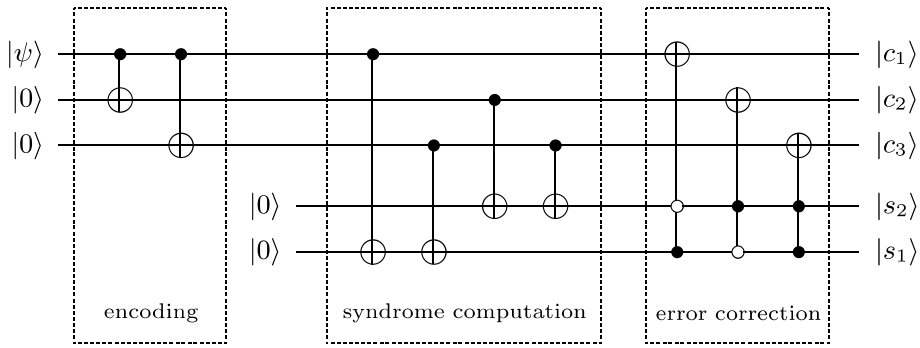

Quantum Error Correction and Fault Tolerant Quantum Computing, Figure 3
Quantum circuit for the three-qubit code. In contrast to Fig. 2, the error is corrected without measuring the syndrome qubits

can also correct the effect of measuring one subsystem. Assume that we measured the first qubit of the encoded state $|\Phi\rangle$ and the result was 0. Using (3), for the state after the measurement we compute

$$\big(|0\rangle\langle 0| \otimes I \otimes I\big)|\Phi\rangle = \frac{1}{2}|\Phi\rangle + \frac{1}{2}\big(\sigma_z \otimes I \otimes I\big)|\Phi\rangle\,,$$

that is, the state is a superposition of the states corresponding to "no error" and "phase error at the first position". Measuring the error syndrome projects either onto the error free state $|\Phi\rangle$ or onto the state with a single phase-flip error that can be corrected. The measurement result 1, as well as measuring one of the other qubits, can be treated similarly.

Shor's Nine-Qubit Code

The three-qubit code C and its Hadamard transformed version C^{phase} of the previous section can correct a single bit-flip error or a single phase-flip error, respectively. However, none of the codes can correct both types of errors. For an encoded state $|\Psi\rangle = \alpha|000\rangle + \beta|111\rangle$ of the three-qubit code, a single phase-flip error σ_z results in the state $(\sigma_z \otimes I \otimes I)|\Psi\rangle = \alpha|000\rangle - \beta|111\rangle$. In terms of the *encoded* or *logical* basis states $|0\rangle_L = |000\rangle$ and $|1\rangle_L = |111\rangle$ corresponding to the encoding of $|0\rangle$ and $|1\rangle$, respectively, a single phase-flip has the effect of an *encoded* σ_z operation. Note that the operations $\sigma_z \otimes I \otimes I$, $I \otimes \sigma_z \otimes I$, and $I \otimes I \otimes \sigma_z$ all have the same effect on the code. Hence, in order to correct also for phase-flip errors, we can add another layer of encoding using the code C^{phase} with the encoded basis states

$$\begin{aligned} |0\rangle &\mapsto \frac{1}{2}(|000\rangle + |011\rangle + |101\rangle + |110\rangle) \\ |1\rangle &\mapsto \frac{1}{2}(|001\rangle + |010\rangle + |100\rangle + |111\rangle)\,. \end{aligned} \tag{4}$$

Each qubit in (4) is replaced by its encoded version with respect to the three-qubit code C, that is, we get the encoding

$$\begin{aligned} |0\rangle &\mapsto \frac{1}{2}(|000000000\rangle + |000111111\rangle \\ &\qquad + |111000111\rangle + |111111000\rangle) \\ |1\rangle &\mapsto \frac{1}{2}(|000000111\rangle + |000111000\rangle \\ &\qquad + |111000000\rangle + |111111111\rangle)\,. \end{aligned} \tag{5}$$

This nine-qubit code has been constructed by Shor [64]. It turns out it can correct an arbitrary single-qubit error. For this, first note that the errors σ_x and σ_z can be corrected independently on the two levels of encoding, respectively. Therefore, the code can correct not only these errors, but also their combination $\sigma_x\sigma_z$. Together with identity we have the matrices

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \\ \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \text{and} \quad \sigma_x\sigma_z = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{6}$$

Any 2×2 matrix can be written as linear combination of the four matrices in (6). Similar to the arguments following (3), it can be shown that in order to correct an arbitrary single-qubit error it is indeed sufficient to correct the errors corresponding to the matrices in (6). The quantum error-correcting code given by (5) is denoted by $C = [\![9, 1, 3]\!]$, indicating that one qubit is encoded into nine qubits and that the minimum distance of the code is three.

Before presenting further constructions for quantum codes, we will discuss necessary and sufficient conditions for quantum error correction.

Conditions for Quantum Error Correction

Quantum Channels

Errors in quantum systems are due to interaction of the system with its environment or imperfection in the control of quantum operations. The latter can be modeled by a perfect operation that additionally depends on the state of the environment. The Hilbert space of the combined system is given by $\mathcal{H}_{\text{sys/env}} = \mathcal{H}_{\text{system}} \otimes \mathcal{H}_{\text{environment}}$. If the dimension of both spaces is sufficiently large, we can assume that the initial state is a pure state. Additionally, we assume that there are no initial correlations between the system and its environment, that is, the initial state is a product state $|\phi\rangle_{\text{sys}}|\varepsilon\rangle_{\text{env}}$. Again, for sufficiently large dimensions of the Hilbert spaces we can assume that the dynamics of the joint system are given by a unitary transformation $U_{\text{sys/env}}$. The resulting state of the system is obtained by tracing out the environment:

$$\rho_{\text{out}} = \text{Tr}_{\text{env}}\left(U_{\text{sys/env}}(|\phi\rangle\langle\phi| \otimes |\varepsilon\rangle\langle\varepsilon|)U^{\dagger}_{\text{sys/env}}\right) . \quad (7)$$

The output state in (7) can equivalently be expressed as a function of the input state $\rho_{\text{in}} = |\phi\rangle\langle\phi|$ in the form

$$\rho_{\text{out}} = \sum_i E_i \, \rho_{\text{in}} \, E_i^{\dagger} ,$$

where the operators E_i are the so-called *error operators* or *Kraus operators* [47]. They depend on both the initial state $|\varepsilon\rangle$ of the environment and the unitary interaction $U_{\text{sys/env}}$. In the following, some important special cases of quantum channels are presented.

Example (depolarizing channel) The *depolarizing channel* on the Hilbert space $\mathcal{H}$ with error parameter p, $0 \le p \le 1$, is given by

$$\rho \mapsto (1-p)\cdot\rho + p \cdot I/\dim\mathcal{H} .$$

The input ρ is transmitted faithfully with probability $1-p$. With probability p, the state is replaced by the completely random state $I/\dim\mathcal{H}$. Note that even in this case, the probability of measuring a particular pure state $|\psi\rangle$ is $1/\dim\mathcal{H} \neq 0$.

While the depolarizing channel treats all input states uniformly, the next quantum channel is basis-dependent.

Example (dephasing channel) The *dephasing channel* on the Hilbert space $\mathcal{H}$ with orthonormal basis $\mathcal{B} = \{|\mathbf{b}_i\rangle : i \in \mathcal{I}\}$ and error parameter p, $0 \le p \le 1$, is given by

$$\rho \mapsto (1-p)\cdot\rho + p\sum_{i\in\mathcal{I}} |\mathbf{b}_i\rangle\langle\mathbf{b}_i|\rho|\mathbf{b}_i\rangle\langle\mathbf{b}_i| .$$

With probability p, the channel performs a projective measurement with respect to the basis $\mathcal{B}$. is derived from the fact that this is equivalent to randomizing the phases of the basis states. The dephasing channel allows us to perfectly transmit classical information by encoding the information as basis states. Coherent superpositions of basis states, however, are changed into classical mixtures.

The final example is a channel that provides the side-information that an error has occurred.

Example (quantum erasure channel [35]*)* The *quantum erasure channel* on the Hilbert space $\mathcal{H}$ with error parameter p, $0 \le p \le 1$, is given by

$$\rho \mapsto (1-p)\cdot\rho + p\cdot|\epsilon\rangle\langle\epsilon| ,$$

where $|\epsilon\rangle$ is a quantum state in the Hilbert space $\mathcal{H}' \supset \mathcal{H}$ that is orthogonal to all states in $\mathcal{H}$. The input ρ is transmitted faithfully with probability $1-p$. With probability p, the state is replaced by the state $|\epsilon\rangle\langle\epsilon|$. As $|\epsilon\rangle$ is orthogonal to all states in $\mathcal{H}$, the receiver can perform a measurement which detects that an error has occurred.

The quantum mechanical analogue of a memoryless channel is a *product channel* which is defined for a quantum system with n subsystems of, say, equal dimension, that is, on $\mathcal{H} = \mathcal{H}_0^{\otimes n}$. The product channel is given by n uses of a channel on $\mathcal{H}_0$ which acts independently on each of the n subsystems. If the channel on $\mathcal{H}_0$ is given by the error operators $\mathcal{E}_0 = \{E_i : i \in \mathcal{I}\}$, the error operators of the product channel on $\mathcal{H}$ are

$$\mathcal{E} = \mathcal{E}_0^{\otimes n} := \{E_{i_1} \otimes E_{i_2} \otimes \ldots \otimes E_{i_n} : (i_1, i_2, \ldots, i_n) \in \mathcal{I}^n\}.$$

Characterization of Quantum Codes

As we have seen in Fig. 1, a quantum error-correcting code is a subspace C of the Hilbert space $\mathcal{H}$. What is more, the full Hilbert space $\mathcal{H}$ can be decomposed into mutually orthogonal unitary images of C, corresponding to different error events (see Eq. (1)). In general, the error operators E_i describing the quantum channel need not be unitary. This leads to the question whether a subspace C of $\mathcal{H}$ is a quantum error-correcting code (QECC) for a given quantum channel.

Theorem 5 (QECC characterization [44]) *Let Q be a quantum channel on $\mathcal{H}$ with error operators $\{E_i : i \in \mathcal{I}_Q\}$. A subspace $C \le \mathcal{H}$ with orthonormal basis $\{|\mathbf{c}_i\rangle : i \in \mathcal{I}_C\}$ is a quantum error-correcting code for Q if and only if the following conditions hold:*

$$\forall k, \ell \in \mathcal{I}_Q \forall i \neq j \in \mathcal{I}_C : \langle\mathbf{c}_i|E_k^{\dagger}E_\ell|\mathbf{c}_j\rangle = 0 \quad (8a)$$

$$\forall k,\ell \in \mathcal{I}_Q \forall i,j \in \mathcal{I}_C\colon \langle \mathbf{c}_i|E_k^\dagger E_\ell|\mathbf{c}_i\rangle = \langle \mathbf{c}_j|E_k^\dagger E_\ell|\mathbf{c}_j\rangle =: \alpha_{k\ell} \in \mathbb{C} \quad (8b)$$

Denoting by $P_C := \sum_{i\in\mathcal{I}_C} |\mathbf{c}_i\rangle\langle\mathbf{c}_i|$ *the projection onto the code* C*, we obtain the following equivalent condition which is independent of the basis of the code:*

$$\forall k,\ell \in \mathcal{I}_Q\colon P_C E_k^\dagger E_\ell P_C = \alpha_{k\ell} P_C\,.$$

In principle, the proof of Theorem 5 implies an algorithm that allows the correction of errors (see [31]). However, as error correction is NP hard in the classical case we cannot expect to have an efficient algorithm for the more general situation of the quantum case.

If the conditions (8) are fulfilled, then the errors corresponding to the operators E_i can be corrected. It must be stressed that this implies that one can correct any operator that is a linear combination of the error operators E_i. For this, we show that the conditions (8) are linear in the error operators. Consider the new error operators

$$A := \sum_k \lambda_k E_k \quad \text{and} \quad B := \sum_l \mu_l E_l$$

which are arbitrary linear combinations of the E_i. Using (8) we compute

$$\begin{aligned}\langle c_i|A^\dagger B|c_j\rangle &= \sum_{k,l} \overline{\lambda_k}\mu_l \langle c_i|E_k^\dagger E_l|c_j\rangle \\ &= \sum_{k,l} \overline{\lambda_k}\mu_l \delta_{i,j}\alpha_{k,l} \\ &= \delta_{i,j}\cdot\alpha'(A,B)\,,\end{aligned}$$

where $\alpha'(A,B) \in \mathbb{C}$ is some constant depending on the operators A and B only. From this it also follows that it is sufficient to check the conditions (8) for a vector space basis of $\mathcal{E}$ and hence for a finite set of errors. For qubit systems, the Pauli matrices

$$X = \sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad Y = \sigma_y = \begin{pmatrix} 0 & -\mathbf{i} \\ \mathbf{i} & 0 \end{pmatrix}, \quad \text{and} \quad Z = \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad (9)$$

together with identity form a vector space basis of all operators in $\mathbb{C}^{2\times 2}$. For a quantum code using n qubits, we consider the tensor product of Pauli matrices and identity as the so-called *error basis*. The number of tensor factors different from identity is referred to as the *number of errors* or the *weight of an error*.

Quantum Codes from Classical Codes

CSS Codes

The three-qubit code of Sect. "The Three-Qubit Code" is based on the classical triple repetition code. Both codes can correct a single bit-flip error. The Hadamard transformation yields the code C^{phase} which can correct a single phase-flip error. In the following we present a similar construction of quantum codes based on linear binary block codes, but the resulting codes will allow both the correction of bit-flip errors and phase-flip errors. For the construction we need the following lemma.

Lemma 1 *Let* $C \le GF(2)^n$ *denote a* k*-dimensional linear subspace of* $GF(2)^n$ *and let* $\mathbf{a},\mathbf{b} \in GF(2)^n$ *be two arbitrary binary vectors. Furthermore, by* $H_{2^n} := H^{\otimes n}$*, where* $H = 1/\sqrt{2}\left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$*, we denote the Hadamard transformation on* n *qubits. Then the state*

$$|\psi\rangle := \frac{1}{\sqrt{|C|}} \sum_{\mathbf{c}\in C} (-1)^{\mathbf{a}\cdot\mathbf{c}} |\mathbf{c}+\mathbf{b}\rangle$$

is mapped by the Hadamard transformation to

$$H_{2^n}|\psi\rangle = \frac{(-1)^{\mathbf{a}\cdot\mathbf{b}}}{\sqrt{|C^\perp|}} \sum_{\mathbf{d}\in C^\perp} (-1)^{\mathbf{b}\cdot\mathbf{d}} |\mathbf{d}+\mathbf{a}\rangle\,.$$

Proof The Hadamard transformation on n qubits can be written as

$$H_{2^n} = \frac{1}{\sqrt{2^n}} \sum_{\mathbf{x},\mathbf{y}\in GF(2)^n} (-1)^{\mathbf{x}\cdot\mathbf{y}} |\mathbf{x}\rangle\langle\mathbf{y}|\,,$$

where $\mathbf{x}\cdot\mathbf{y}$ denotes the inner product of the binary vectors $\mathbf{x}$ and $\mathbf{y}$. Then

$$\begin{aligned} H_{2^n}|\psi\rangle &= \frac{1}{\sqrt{2^n|C|}} \sum_{\mathbf{x},\mathbf{y}\in GF(2)^n} (-1)^{\mathbf{x}\cdot\mathbf{y}} |\mathbf{x}\rangle\langle\mathbf{y}| \sum_{\mathbf{c}\in C} (-1)^{\mathbf{a}\cdot\mathbf{c}} |\mathbf{c}+\mathbf{b}\rangle \\ &= \frac{1}{\sqrt{2^n|C|}} \sum_{\mathbf{x},\mathbf{y}\in GF(2)^n} \sum_{\mathbf{c}\in C} (-1)^{\mathbf{x}\cdot\mathbf{y}+\mathbf{a}\cdot\mathbf{c}} |\mathbf{x}\rangle\langle\mathbf{y}|\mathbf{c}+\mathbf{b}\rangle \\ &= \frac{1}{\sqrt{2^n|C|}} \sum_{\mathbf{x}\in GF(2)^n} \sum_{\mathbf{c}\in C} (-1)^{\mathbf{x}\cdot(\mathbf{c}+\mathbf{b})+\mathbf{a}\cdot\mathbf{c}} |\mathbf{x}\rangle \\ &= \frac{1}{\sqrt{2^n|C|}} \sum_{\mathbf{x}\in GF(2)^n} (-1)^{\mathbf{b}\cdot\mathbf{x}} |\mathbf{x}\rangle \sum_{\mathbf{c}\in C} (-1)^{(\mathbf{x}+\mathbf{a})\cdot\mathbf{c}} \\ &\overset{(*)}{=} \frac{|C|}{\sqrt{2^n|C|}} \sum_{\mathbf{x}\in C^\perp+\mathbf{a}} (-1)^{\mathbf{b}\cdot\mathbf{x}} |\mathbf{x}\rangle \\ &= \frac{(-1)^{\mathbf{a}\cdot\mathbf{b}}}{\sqrt{|C^\perp|}} \sum_{\mathbf{d}\in C^\perp} (-1)^{\mathbf{b}\cdot\mathbf{d}} |\mathbf{d}+\mathbf{a}\rangle\,. \end{aligned}$$

In $(*)$ we have used that the sum $\sum_{\mathbf{c}\in C}(-1)^{\mathbf{x}\cdot\mathbf{c}}$ vanishes if and only if $\mathbf{x} \neq C^{\perp}$. □

Lemma 1 shows that the Hadamard transformation not only changes phase-flip errors into bit-flip errors, it also maps superpositions of all codewords of the linear binary code C to superpositions of all codewords of the dual code $C^{\perp}$ (cf. Proposition 1).

Example (seven-qubit code) The 3rd binary Hamming code $C = [7, 4, 3]$ (cf. Proposition 2) contains its dual code $C^{\perp} = [7, 3, 4]$, that is, $C^{\perp} \subset C$. Hence we can partition the codewords of C into two cosets of $C^{\perp}$ as follows:

$$C = \left(C^{\perp} + \mathbf{x}_0\right) \dot{\cup} \left(C^{\perp} + \mathbf{x}_1\right) ,$$
$$\text{where } \mathbf{x}_0 = (0000000) \quad \text{and } \mathbf{x}_1 = (1111111) .$$

The Hamming weight of all codewords of $C^{\perp}$ is even, while the weight of all vectors in the coset $C^{\perp} + \mathbf{x}_1$ is odd. Based on this decomposition, we define the following encoding:

$$|0\rangle \mapsto |0\rangle_L = \frac{1}{\sqrt{|C^{\perp}|}} \sum_{\mathbf{c}\in C^{\perp}} |\mathbf{c}+\mathbf{x}_0\rangle = \frac{1}{\sqrt{|C^{\perp}|}} \sum_{\mathbf{c}\in C^{\perp}} |\mathbf{c}\rangle \tag{10a}$$

$$|1\rangle \mapsto |1\rangle_L = \frac{1}{\sqrt{|C^{\perp}|}} \sum_{\mathbf{c}\in C^{\perp}} |\mathbf{c}+\mathbf{x}_1\rangle . \tag{10b}$$

Hadamard transformation of these states yields

$$H_{2^7}|0\rangle_L = \frac{1}{\sqrt{|C|}} \sum_{\mathbf{c}\in C} (-1)^{\mathbf{c}\cdot\mathbf{x}_0}|\mathbf{c}\rangle = \frac{1}{\sqrt{|C|}} \sum_{\mathbf{c}\in C} |\mathbf{c}\rangle = \frac{1}{\sqrt{2}} \left(|0\rangle_L + |1\rangle_L\right) \tag{11a}$$

$$H_{2^7}|1\rangle_L = \frac{1}{\sqrt{|C|}} \sum_{\mathbf{c}\in C} (-1)^{\mathbf{c}\cdot\mathbf{x}_1}|\mathbf{c}\rangle = \frac{1}{\sqrt{2}} \left(|0\rangle_L - |1\rangle_L\right) . \tag{11b}$$

A superposition $|\psi\rangle = \alpha|0\rangle_L + \beta|1\rangle_L$ of the logical qubits is a superposition of words of the Hamming code $C = [7, 4, 3]$. Similar to the transformation (2) it is possible to compute an error syndrome for the bit-flip errors using a parity check matrix of the Hamming code. Measuring the error syndrome provides information about the position of a single bit-flip, allowing us to correct this error. From (11) it can be seen that the Hadamard transformation of the state $|\psi\rangle$ is again a superposition of words of the Hamming code, so a single phase-flip error can be corrected as well. Similar to Shor's nine-qubit code, for this seven-qubit code $C = [\![7, 1, 3]\!]$ given by (10), bit-flips and phase-flips can be corrected independently (for more details see [31]).

Equation (11) additionally shows that applying the Hadamard transformation to all seven qubits corresponds to a Hadamard transformation of the logical qubits $|0\rangle_L$ and $|1\rangle_L$. Similarly, applying σ_x or σ_z to all qubits transforms the encoded qubits like encoded versions of σ_x and σ_z, respectively (see also Sect. "Transversal Gates").

The generalization of this construction principle yields so-called CSS codes. This class of codes was independently derived by Calderbank and Shor [15] and Steane [66,67].

Theorem 6 (CSS code) *Let $C_1 = [n, k_1, d_1]$ and $C_2 = [n, k_2, d_2]$ be linear binary codes of length n, dimension k_1 and k_2, respectively, and minimum distance d_1 and d_2, respectively, with $C_2^{\perp} \subseteq C_1$. Furthermore, let $\mathcal{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_K\} \subset GF(2)^n$ be a system of representatives of the cosets of $C_2^{\perp}$ in C_1.*

The $K = 2^{k_1-(n-k_2)}$ mutually orthogonal states

$$|\psi_i\rangle = \frac{1}{\sqrt{|C_2^{\perp}|}} \sum_{\mathbf{c}\in C_2^{\perp}} |\mathbf{c}+\mathbf{w}_i\rangle \tag{12}$$

span a quantum error-correcting code $C = [\![n, k, d]\!]$ with $k := k_1 + k_2 - n$. The code corrects at least $\lfloor (d_1 - 1)/2 \rfloor$ bit-flip errors and simultaneously at least $\lfloor (d_2 - 1)/2 \rfloor$ phase-flip errors. Its minimum distance is $d \geq \min\{d_1, d_2\}$.

Proof An arbitrary combination of bit-flip and phase-flip errors can be written as

$$\mathbf{e} := \left(\sigma_x^{e_{x,1}}\sigma_z^{e_{z,1}}\right) \otimes \ldots \otimes \left(\sigma_x^{e_{x,n}}\sigma_z^{e_{z,n}}\right) , \tag{13}$$

where the binary vectors $\mathbf{e}_x$ and $\mathbf{e}_z$ indicate the positions with bit-flip and sign-flip errors, respectively. An arbitrary state of the CSS code is a superposition of the encoded basis states $|\psi_i\rangle$ in (12). Rewriting the superposition shows that the state is a superposition of codewords of the binary code C_1:

$$|\psi\rangle = \sum_{i=1}^{K} \alpha_i|\psi_i\rangle = \sum_{i=1}^{K} \alpha_i' \sum_{\mathbf{c}\in C_2^{\perp}} |\mathbf{c}+\mathbf{w}_i\rangle = \sum_{\mathbf{c}\in C_1} \beta_{\mathbf{c}}|\mathbf{c}\rangle . \tag{14}$$

Combining (13) and (14) the erroneous state reads

$$\mathbf{e}|\psi\rangle = \sum_{\mathbf{c}\in C_1} \beta_{\mathbf{c}}(-1)^{\mathbf{c}\cdot\mathbf{e}_z}|\mathbf{c}+\mathbf{e}_x\rangle . \tag{15}$$

Using a parity check matrix H_1 for the linear binary code C_1 we can implement the mapping

$$S_1 : |\mathbf{x}\rangle|\mathbf{y}\rangle \mapsto |\mathbf{x}\rangle|\mathbf{x}H_1^t + \mathbf{y}\rangle .$$

Applying S_1 to the state (15) and $n-k_1$ ancilla qubits in the state $|0\rangle$, we get

$$\sum_{\mathbf{c}\in C_1} \beta_{\mathbf{c}}(-1)^{\mathbf{c}\cdot\mathbf{e}_z}|\mathbf{c}+\mathbf{e}_x\rangle|(\mathbf{c}+\mathbf{e}_x)H_1^t\rangle$$
$$= \Big(\sum_{\mathbf{c}\in C_1} \beta_{\mathbf{c}}(-1)^{\mathbf{c}\cdot\mathbf{e}_z}|\mathbf{c}+\mathbf{e}_x\rangle\Big) \otimes |\mathbf{e}_x H_1^t\rangle$$
$$= \big(\mathbf{e}|\psi\rangle\big) \otimes |\mathbf{e}_x H_1^t\rangle \,. \quad (16)$$

As the syndrome $\mathbf{s}_x = \mathbf{e}_x H_1^t$ depend only on the error $\mathbf{e}_x$, not on the codeword $\mathbf{c}$, the state in (16) is a tensor product. Hence we can measure the syndrome of the bit-flip errors without disturbing the state $\mathbf{e}|\psi\rangle$. A classical algorithm for the decoding of the code C_1 can then be used to deduce the error vector $\mathbf{e}_x$ from the syndrome $\mathbf{s}_x$. In order to correct phase-flip errors, we recall that the Hadamard transformation interchanges σ_x and σ_z. Applying the Hadamard transformation to the erroneous state (15) we get

$$H_{2^n}\mathbf{e}|\psi\rangle = (H_{2^n}\mathbf{e}H_{2^n})H_{2^n}\left(\sum_{i=1}^{K}\alpha_i' \sum_{\mathbf{c}\in C_2^\perp}|\mathbf{c}+\mathbf{w}_i\rangle\right)$$
$$= \sum_{i=1}^{K}\alpha_i'' \sum_{\mathbf{c}\in C_2}(-1)^{\mathbf{c}\cdot\mathbf{e}_x}(-1)^{\mathbf{c}\cdot\mathbf{w}_i}|\mathbf{c}+\mathbf{e}_z\rangle \,.$$

The last equation follows using Lemma 1. Similar to (16) we can use a parity check matrix H_2 of the binary code C_2 to define the mapping $S_2\colon |\mathbf{x}\rangle|\mathbf{y}\rangle \mapsto |\mathbf{x}\rangle|\mathbf{x}H_2^t+\mathbf{y}\rangle$. Using $n-k_2$ additional ancilla qubits, we obtain the state

$$\big(H_{2^n}\mathbf{e}|\psi\rangle\big) \otimes |\mathbf{e}_x H_1^t\rangle \otimes |\mathbf{e}_z H_2^t\rangle \,.$$

Undoing the Hadamard transformation on the first n qubits we finally get

$$\big(\mathbf{e}|\psi\rangle\big) \otimes |\mathbf{e}_x H_1^t\rangle \otimes |\mathbf{e}_z H_2^t\rangle$$
$$= \Big(\sum_{\mathbf{c}\in C_1} \beta_{\mathbf{c}}(-1)^{\mathbf{c}\cdot\mathbf{e}_z}|\mathbf{c}+\mathbf{e}_x\rangle\Big) \otimes |\mathbf{e}_x H_1^t\rangle \otimes |\mathbf{e}_z H_2^t\rangle \,.$$

Again we can measure the syndrome $\mathbf{s}_z = \mathbf{e}_z H_2^t$ and use a decoding algorithm for the linear binary code C_2 to find the error vector $\mathbf{e}_z$. □

Stabilizer Codes

The construction of CSS codes uses two binary codes which define a decomposition of the set of all binary strings of length n. Using the binary strings as labels of quantum states, one obtains a decomposition of the complex vector space $(\mathbb{C}^2)^{\otimes n}$ (cf. Fig. 1). Such a decomposition can also be defined via eigenspaces of operators.

Let $\mathcal{P}_n$ denote the group which is generated by tensor products of n Pauli matrices (cf. (9)) and identity. Every element $g \in \mathcal{P}_n$ has a unique representation of the form

$$g = i^c \cdot \sigma_x^{g_{x,1}}\sigma_z^{g_{z,1}} \otimes \cdots \otimes \sigma_x^{g_{x,n}}\sigma_z^{g_{z,n}} \,,$$

where $c \in \{0,1,2,3\}$ and $\mathbf{g}_x$ and $\mathbf{g}_z$ are binary vectors of length n. Two elements g and h of $\mathcal{P}_n$ either commute or anti-commute, that is, $gh = \pm hg$.

Let S be an Abelian subgroup of $\mathcal{P}_n$, that is, any two elements of S commute. Furthermore, we assume that S does not contain $-I$. Then S has 2^{n-k} elements and there are $n-k$ generators. The spectrum of every generator g_i is $\{+1,-1\}$. Furthermore, there are 2^{n-k} common eigenspaces, each of dimension 2^k, which can be labeled by the eigenvalues of the generators. We choose one of these eigenspaces as our quantum code C, e. g., the eigenspace with all eigenvalues $+1$. For the analysis of the error-correcting properties of this code, recall that any two elements in $\mathcal{P}_n$ either commute or anti-commute. Assume that $E \in \mathcal{P}_n$ anti-commutes with some generator g_i of S. Then for every state $|\psi\rangle \in C$ we have

$$g_i(E|\psi\rangle) = -Eg_i|\psi\rangle = -E|\psi\rangle \,, \quad (17)$$

that is, $E|\psi\rangle$ lies in the eigenspace of g_i with eigenvalue -1 which is orthogonal to the code. Hence every error E that anti-commutes with at least one of the generators g_i of S can be detected. From the definition of S it follows that $E|\psi\rangle = |\psi\rangle$ for all $E \in S$ and every state $|\psi\rangle \in C$, that is, these "errors" E have no effect on the code. However, if $E \notin S$ commutes with all elements in S, then E preserves the code space, but acts non-trivially on it. When designing a code, we want the weight of these undetectable errors to be large, that is, we want the number of tensor factors of E that are different from identity to be high.

This construction of so-called *stabilizer codes* has been found independently by Gottesman [25] and Calderbank et al. [14]. In the latter paper, a connection to additive codes over the finite field $GF(4)$ has been established which allows us to use results from classical coding theory for the construction of good quantum codes.

Theorem 7 (stabilizer codes) *Let S be an Abelian subgroup of the n-qubit Pauli group $\mathcal{P}_n$ that does not contain $-I$. The stabilizer code C with stabilizer S is the common eigenspace with eigenvalue + 1 of all operators in S. The dimension of C is 2^k if $|S| = 2^{n-k}$. The normalizer $\mathcal{N}(S)$ is defined as*

$$\mathcal{N}(S) = \{x\colon x \in \mathcal{P}_n \mid x^{-1}Sx = S\}$$
$$= \{x\colon x \in \mathcal{P}_n \mid xg = gx \ \textit{for all}\ g \in S\} \,.$$

The normalizer has 2^{n+k} elements and contains the stabilizer S. If the weight of all elements of the set $\mathcal{N}(S) \setminus S$ is at

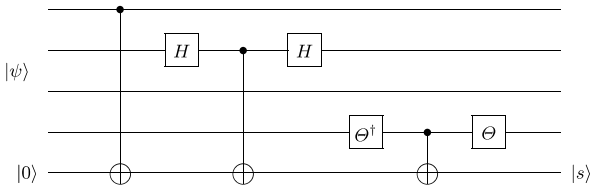

Quantum Error Correction and Fault Tolerant Quantum Computing, Figure 4
Quantum circuit to measure the eigenvalue of $g = \sigma_z \otimes \sigma_x \otimes I \otimes \sigma_y$

least d, then all errors of weight strictly less than d can either be detected or have no effect on the code. Equivalently, one can correct all errors of weight up to $\lfloor (d-1)/2 \rfloor$. The code is denoted by $C = [\![n, k, d]\!]$.

In a manner similar to error correction for CSS codes, error correction for stabilizer codes is based on measuring a syndrome of the error. Recall that an error can be detected if it anti-commutes with one of the generators g_i of the stabilizer S. Then the erroneous state lies in the eigenspace of g_i with eigenvalue -1 (see (17)). Hence measuring the eigenvalues of the generators g_i provides information about the error E. Since the Pauli matrices are both unitary and Hermitian, the generators define a quantum mechanical observable. However, if the weight of the generator g_i is m, one would have to measure an m-qubit observable. Alternatively, one can use one ancilla qubit per generator g_i to compute an error syndrome. Assume that we want to measure the eigenvalue of $g = \sigma_z \otimes \sigma_x \otimes I \otimes \sigma_y$. For this, we can use the quantum circuit shown in Fig. 4.

The eigenstates of σ_z are $|0\rangle$ and $|1\rangle$ with eigenvalues $+1$ and -1, respectively. Hence the first CNOT gate flips the state of the ancilla qubit if and only if the first qubit is in the eigenspace of σ_z with eigenvalue -1. The control qubit of the next CNOT gate is conjugated with the Hadamard matrix which interchanges σ_x and σ_z. Hence the ancilla qubit is flipped if and only if the second qubit is in the eigenspace of σ_x with eigenvalue -1. In order to measure the eigenvalue of σ_y, the control of the final CNOT gate is conjugated with the matrix $\Theta = 1/\sqrt{2}\left(\begin{smallmatrix} 1 & -\mathbf{i} \\ -\mathbf{i} & 1 \end{smallmatrix}\right)$ that maps σ_y to σ_z. If measuring the ancilla qubit $|s\rangle$ yields the result 0 or 1, the state of the first four qubits is projected onto the eigenspace of g with eigenvalue $+1$ or -1, respectively.

On the one hand, the elements of the normalizer $\mathcal{N}(S)$ which do not lie in the stabilizer S correspond to errors that cannot be detected. On the other hand, these elements can be used to perform operations on the code, as we will see in the following example.

Example (five-qubit code) The shortest quantum code that can correct an arbitrary single-qubit error is the code $C = [\![5, 1, 3]\!]$ which is a stabilizer code. The stabilizer S of this code is generated by

$$\begin{aligned} g_1 &= \sigma_z \otimes \sigma_x \otimes \sigma_x \otimes \sigma_z \otimes I = ZXXZI\,, \\ g_2 &= I \otimes \sigma_z \otimes \sigma_x \otimes \sigma_x \otimes \sigma_z = IZXXZ\,, \\ g_3 &= \sigma_z \otimes I \otimes \sigma_z \otimes \sigma_x \otimes \sigma_x = ZIZXX\,, \\ g_4 &= \sigma_x \otimes \sigma_z \otimes I \otimes \sigma_z \otimes \sigma_x = XZIZX\,. \end{aligned}$$

The normalizer of S is generated by $g_1, \ldots, g_4$ together with

$$\begin{aligned} h_1 &= \sigma_x \otimes \sigma_z \otimes \sigma_x \otimes I \otimes I = XZXII\,, \\ h_2 &= \sigma_z \otimes \sigma_z \otimes I \otimes \sigma_x \otimes I = ZZIXI\,. \end{aligned}$$

Both h_1 and h_2 commute with all g_i, and h_1 and h_2 anti-commute with each other. Hence h_1 and h_2 act on the code as encoded versions of σ_x and σ_z.

We close this section by noting that CSS codes are also stabilizer codes. The stabilizer of the seven-qubit code of Example 4 is generated by

$$\begin{aligned} g_1 &= XIXIXIX\,, & g_4 &= ZIZIZIZ\,, \\ g_2 &= IXXIIXX\,, & g_5 &= IZZIIZZ\,, \\ g_3 &= IIIXXXX\,, & g_6 &= IIIZZZZ\,. \end{aligned}$$

The additional generators of the normalizer are given by

$$h_1 = XXXXXXX\,, \qquad h_2 = ZZZZZZZ\,.$$

As already mentioned at the end of Example 4, h_1 and h_2 correspond to the encoded version of σ_x and σ_z, respectively.

Techniques for Fault-Tolerant Quantum Computing

The theory of quantum codes as described in the previous sections is a prerequisite to be able to achieve a larger goal, namely, reliable quantum computations performed with imperfect hardware. Imperfections arise from the fact that every physical device is subject to noise. This includes memory elements as well as gate elements. Not only is memory affected by errors, but the operations by which this memory is modified can also be faulty, so the task of devising reliable quantum computations is very delicate. The idea behind fault-tolerant quantum computing (FTQC) is to achieve this very goal, provided that the noise level introduced by the memory and the gates is not too high.

It should be noted that FTQC is quite different from the techniques used in classical fault-tolerance and dependability of systems (see [7] for an overview). In general, the supply of techniques that can be drawn from in the

quantum case is less rich than in the classical case, where fault-tolerance techniques are known on basically all levels of abstraction, for hardware as well as for software. The reason for this difference is that in the quantum case there are strong restrictions on how quantum information can be maintained. One example might be the "no cloning theorem" that was already mention as Theorem 4. Not having the ability to copy quantum information rules out many classical fault-tolerance techniques such as error-detection for rollback recovery [23].

The results shown in FTQC have the flavor of some of the results shown in the early days of gate level fault-tolerance in the classical case, such as the celebrated threshold result for faulty NAND gates due to J. von Neumann [55]. An important result of the theory of FTQC is that also in the quantum case there is a threshold value for this noise level such that arbitrarily long quantum computations become possible if the gates have a noise level that is under the threshold. We will briefly discuss the current state of the art of estimates of this threshold value from below and above in Sect. "The Threshold Theorem".

FTQC comprises methods to perform the following tasks on encoded quantum states in such a way that errors do not accumulate. This means that during the application of a single gate, errors do not spread out in an uncontrolled fashion over the whole circuit. Instead, errors stay locally confined so that they can be taken care of by error correction routines that are applied periodically to the system:

- Preparation of initial states and ancilla states
- Algorithms for quantum error-correction
- Measurement of quantum states
- Universal set of quantum gates.

Throughout, in FTQC it is assumed that quantum data is never given in a form where it is unencoded. Instead, it is assumed that whenever a qubit consisting of two (logical) basis states $|0\rangle_L$, $|1\rangle_L$ is given, that these states are encoded into some higher-dimensional Hilbert space and that the need never arises to decode this quantum information from its encoded physical representation to its logical representation.

Transversal Gates

FTQC requires careful design of how quantum information is encoded. A first indication of the inherent difficulties is given by the observation that errors have the tendency to be propagated through quantum gates. The CNOT gate might serve as a simple example. A simple calculation reveals that it propagates (even when working perfectly) a single X-error on the control qubit to a double X-error on both control and target qubit. Conversely, the same CNOT gate propagates a single Z-error on the target qubit to a double Z-error on control and target qubit. A basic requirement for all quantum gates that are used to operate on encoded quantum data is that they should not propagate errors, that is, errors should stay locally confined to a small set of qubits (see Fig. 5).

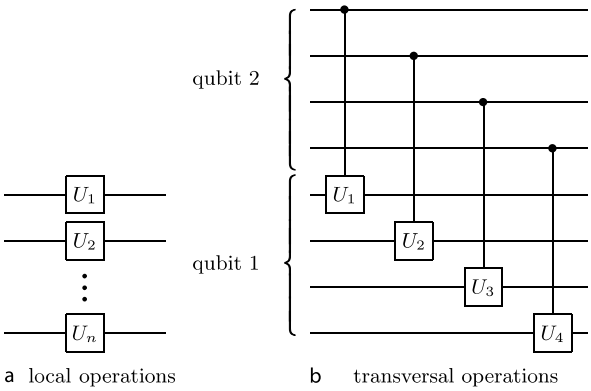

Quantum Error Correction and Fault Tolerant Quantum Computing, Figure 5

Fault-tolerant quantum gates. Shown are two examples, namely *a* local operations that operate on encoded qubits without entangling them, thereby avoiding any error propagation between them, and *b* transversal gates which in this case are given by a sequence of controlled gates. An error in one of these gates will have only a local effect to at most two of the encoded qubits and therefore can be corrected by one round of quantum error-correction provided the qubits are encoded using a quantum code that can correct at least one error. In both *a* and *b* the gates $U_1, U_2, \ldots$, are arbitrary unitaries

It was shown in [27] how to perform encoded Clifford operations transversally on any stabilizer code. Clifford operations are those quantum operations that leave the group of tensor products of Pauli matrices unchanged under conjugation. Clifford operations are not universal for quantum operations. For certain inputs, they can even be efficiently simulated classically [1,5]. However, it was also shown in [27] that, using transversal gates together with local measurements, a universal set of quantum gates can be implemented on any stabilizer code.

Teleporting Gates

While in the beginnings of FTQC universal gate sets were realized using ad hoc constructions involving measurements [65], later it was realized that teleportation [10] also can be used to implement universal sets of quantum gates. This has the following unique advantage: suppose that a particular gate U has to be applied during the com-

putation. Then as shown in [30], under some conditions (which characterize a set of gates more general than Clifford gates, hence a universal set of gates), such a quantum gate can be "precomputed" and stored into a quantum state $|\psi_U\rangle$. Once the gate U has to be applied, this state $|\psi_U\rangle$ is used in a generalized teleportation scheme which can be shown to have the same effect as applying U. The advantage of this method of *teleporting the gate U* is that $|\psi_U\rangle$ can be prepared and verified offline. Then only very high fidelity specimens of $|\psi_U\rangle$ are actually used in the computation.

Fault-Tolerant Error Correction

Several fault-tolerant methods are known to measure the syndrome of a quantum code. We first describe Shor's method which relies on generalized cat states and which can be applied to any stabilizer code. Next, we describe Steane's method which is applicable whenever the code is a CSS code. Finally, we describe Knill's method which is based on teleportation and post-selection.

Shor's Method In Fig. 4 we have seen an example of how the eigenvalue of a generator of the stabilizer can be measured. The principle underlying this example is to compute a controlled operation for each of the non-trivial Pauli operators contained in the generator and to compute the output of this controlled gate into a fresh qubit which is then subsequently measured. Unfortunately, this method is not fault-tolerant as the CNOT has the mentioned property of propagating X-errors from control qubits to target qubits and Z-errors from target qubits to control qubits. Therefore, a different method is needed to implement a fault-tolerant measurement of stabilizer generators. As shown by Shor [65], such a method of fault-tolerant syndrome measurements exists, provided that a supply of cat states, that is, quantum states of the form $1/\sqrt{2}(|0\ldots0\rangle + |1\ldots1\rangle)$, is available. Such a cat state is first transformed using the Hadamard transform applied to all qubits, yielding the equal superposition of all even weight binary words. Next, the same controlled operations as in Fig. 4 are applied but instead of having the same target qubit, they are controlled to individual qubits of the cat state. Afterwards, all ancilla qubits are measured in the standard basis, and the parity of the measurement results is computed classically.

The remaining problem is how to ensure the supply of cat states. This is done in a separate, offline procedure which first performs an encoding of a cat state followed by suitable tests to verify that, indeed, the correct state has been produced.

Steane's Method A different method can be applied if the given quantum code is a CSS code. In this so-called Steane method [68], first an X-error correction is performed, followed by a Z-error correction. Here, these two error correction routines are performed in a very special way. To correct X-errors, an ancilla is prepared in an encoded state $|+\rangle_L = \frac{1}{\sqrt{2}}(|0\rangle_L + |1\rangle_L)$ that corresponds to the superposition of all logical codewords. Then transversal CNOTs from the encoded codeword (control) into the ancillas (target) are applied. Afterwards, the ancilla is measured in the standard basis and, if necessary, a correction is applied. For the Z-errors, the corresponding operations in the Hadamard basis are applied.

A fault-tolerant measurement for a CSS code derived from $C_2^\perp \subseteq C_1$ is defined with respect to the orthogonal projectors onto the states $|\psi_i\rangle$ as in Eq. (6):

$$|i\rangle_L = |\psi_i\rangle = \frac{1}{\sqrt{|C_2^\perp|}} \sum_{\mathbf{c}\in C_2^\perp} |\mathbf{c} + \mathbf{w}_i\rangle \,.$$

First, we describe a way of measuring $|i\rangle_L$ that is fault-tolerant but does not implement the orthogonal projectors given by the CSS basis states (6). Measuring all qubits of $|\psi_i\rangle$ in the computational basis yields a random codeword of the classical code C_1, possibly with some error added. Using a classical error-correction strategy for C_1, this error can be corrected, resulting in a random element of a coset $C_2^\perp + \mathbf{w}_i$. Then, computing the error-syndrome with respect to the classical code $C_2^\perp$, it is possible to infer what i was. In order to obtain a measurement that also implements the orthogonal projection operators, we use ancilla qubits which can be prepared in the state $|0\rangle_L$ and tested offline [68]. Next, the quantum information is transferred from the encoded qubits to the ancilla using transversal CNOTs. Finally, the ancillas are measured in the computational basis and the information about i is computed as described above.

Knill's Method Knill [43] suggested a scheme for fault-tolerant quantum computation based on error detection and post-selection. Knill's scheme involves a "sieving" step in which ancillas are prepared in a suitable state that is later used to teleport a gate into a computation. The gates involved in this sieving phase are protected using a concatenated code that is error-detecting. If an error occurs in this phase, the preparation of the ancilla state is aborted and a new attempt is started. A rigorous threshold for Knill's scheme was given in [4] where it was shown that $\delta \geq 1.04 \times 10^{-3}$ is a lower bound on the threshold. It should be noted that Knill's numerical simulations [43] indicate a much higher value of the threshold of about

$\delta \approx 3 \times 10^{-2}$, albeit at the price of a significant (but additive) hardware overhead.

Both Knill's and Steane's schemes have in common that they are based on post-selection, a feature that also underlies the proposed implementation of a quantum computer based on linear optical elements [46].

The Threshold Theorem

In its simplest form, FTQC is achieved by recursively encoding qubits using a fixed quantum code C. This is the idea underlying several papers in which accuracy thresholds for quantum computing are proven. We briefly describe the reasoning as to why a threshold for FTQC exists and conclude with some historical remarks and pointers to further reading.

Intuition Behind the Threshold Theorem

Let C be a quantum code that encodes only one qubit into several qubits, that is, it is an $[\![n, 1, d]\!]$ which can correct up to $t = \lfloor \frac{d-1}{2} \rfloor$ errors. Encoding one qubit with respect to this code will lead to an improved error rate of p^t over the given error rate of p. Repeating this process h times results in a quantum code C^h with parameters $[\![n^h, 1, d^h]\!]$. Although the code C^h has a very poor rate, it leads to a dramatic reduction of the error probability with a relatively small price in terms of overhead: for h levels of concatenation the error probability becomes $p_*(p/p_*)^{2^h}$, where p_* denotes the threshold probability. This threshold is given by $p_* = 1/c_*$, where c_* roughly denotes the number of events that can cause a logical error under the constraining assumptions made about the physical error model.

It should be noted that, in FTQC, usually one parameter is given, namely the probability p of any gate failing during the computation. This probability p is actually a property of a given, fixed set of universal quantum gates. The quantity $1 - p$ is given as the minimum of the probability of projecting the state obtained from the faulty gate onto the correct quantum state, where the minimum is taken over all gates in the universal set and all states of the Hilbert space.

History of the Threshold Theorem

The first method for FTQC was given by Shor [65], who introduced the idea of alternating error-correction and actual quantum gates in order to do long quantum computations that can tolerate more noise than unencoded operations. He introduced and used universal fault-tolerant gates and showed how to measure the syndrome of a stabilizer code fault-tolerantly by using cat states.

Around the same time, ideas based on concatenation of quantum codes were used by Aharonov and Ben-Or [2], Kitaev [42], Gottesman [26], Knill, Laflamme and Zurek [45], and Preskill [58], to derive FTQC schemes that can perform arbitrarily long computations provided the given error rate lies below a certain value, namely the threshold for FTQC. In these early works, the QECCs used had to be at least two-error correcting. This requirement was later relaxed and it was shown by Reichardt [62] and Aliferis, Gottesman, and Preskill [3] that also one-error correcting quantum codes yield a positive threshold. This allowed us to obtain much higher estimates of the threshold.

In the following years, several attempts have been made to obtain good bounds on the true value of the threshold. Starting with [70] there have been attempts to compute numerical approximations to the true value of the threshold via Monte-Carlo simulations. For this, a particular quantum error-correcting code and a particular error-correction strategy are fixed. Then, in a classical simulation, errors are introduced randomly and it is checked to see how many of them cause uncorrectable errors. This result can be then used to obtain an estimate of the threshold. This technique was later criticized [3] as being prone to overestimating the true threshold value. A further disadvantage is that more advanced error models such as local non-Markovian noise cannot be captured by this method.

On the other hand, in FTQC, there has been a tradition of giving rigorous, analytical proofs for bounds on the threshold by analyzing the recursion. Examples for this method include [2,45]. Typically, this gives a fairly low bound for the true value of the threshold because all error effects are assumed to be worst-case. A recently-studied intermediate model of proving bounds on the threshold makes educated guesses that certain effects are negligible, then calculates the threshold based on that. An advantage of this method is that it never underestimates the true value of the threshold while giving significantly higher values.

Similar to the DiVincenzo criteria for building a quantum computer [21], there is a list of requirements and desiderata for universal FTQC attributable to Gottesman [29] which we repeat here. The point of this list is that any quantum computer that performs its computations in a fault-tolerant fashion must have the following features:

- The gate error rates must be low.
- The architecture must support the ability to perform operations in parallel.

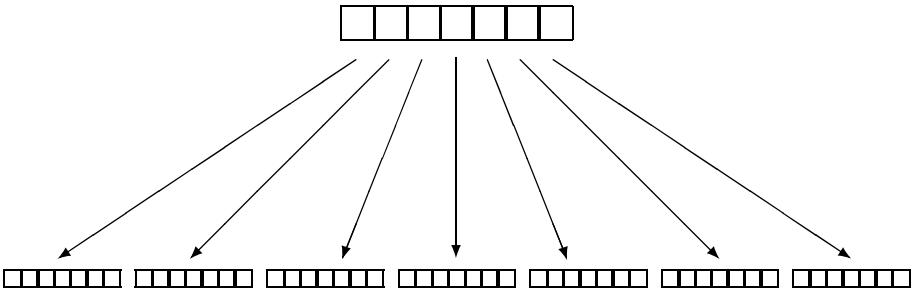

Quantum Error Correction and Fault Tolerant Quantum Computing, Figure 6
Concatenated quantum codes. Shown is one level of concatenation for the seven qubit $[\![7,1,3]\!]$ code. Each of the seven physical qubits in the first layer is replaced by seven qubits encoded again into a $[\![7,1,3]\!]$ quantum code. If the initial error rate is given by ε, then iterating this construction h times yields a quantum code with parameters $[\![7^h,1,3^h]\!]$ and a resulting error rate of roughly $c_*^{2^h-1}\varepsilon^{2^h}$

- There must be a way of remaining in, or of returning to, the computational Hilbert space, thereby preventing leakage errors.
- There must be a source of fresh initialized qubits during the computation.
- The error scaling must be benign, that is, the error rates must not increase as the computer gets larger. Also, there must not be any large-scale correlated errors during the computation.

Further Reading

For an overview of results on FTQC including noise thresholds for models based on post-selection and a detailed treatment of many FTQC constructions, we recommend Reichardt's Ph D thesis [61]. That reference also gives a discussion of alternative techniques to achieve FTQC such as topological quantum computing [20]. For an overview on developments on non-Markovian models, extended rectangles and gadgets used for fault-tolerant error correction, the reader is referred to [3]. Complementing the lower bound results on the threshold are results which give upper bounds, that is, results which determine noise levels above which no quantum computation is possible. Results in this direction were first obtained by Razborov [60] and have been further improved [40]. Finally, for a comparison between different methods for FTQC based on short block codes combined with concatenation, see the recent paper [19].

Further Aspects

In this article we have considered only qubit systems. Both quantum error-correction and fault-tolerant quantum computation can be generalized to quantum systems that are composed of higher dimensional systems (see [6,28,36]). The stabilizer formalism has also been extended from block codes to quantum convolutional codes [17,24,57]. This class of codes allows us to encode and decode a stream of quantum information without partitioning it into blocks of fixed size. For both quantum block codes and quantum convolutional codes there are efficient ways for encoding quantum information [18,33,37].

Building on the connection between stabilizer codes and classical additive codes, various constructions of QECCs have been proposed (for an overview see [41]). Among those are cyclic codes, for which efficient encoding and sub-optimal decoding algorithms are known [32], as well as families of good QECCs based on classical codes from algebraic geometry [54]. More recently, the use of classical LDPC codes for quantum error correction has been proposed [16,52]. Since classical LDPC codes achieve very good performance, it is hoped that their quantum versions result in a high threshold for FTQC. Allowing non-additive classical codes in the construction of QECCs can lead to QECCs whose dimension is larger than those of stabilizer codes [34,59].

Instead of using active methods to obtain information about an error and subsequently correct the error, some physical systems allow for passive error-protection. If the interaction of the quantum system with its environment possesses some symmetry, the quantum information can be stored in a so-called *decoherence free subspace* on which the environment has no effect [22,51,71]. A generalization of this concept is given by *decoherence free subsystems* (more details can be found in the review article [50]). Combining the ideas of active quantum error correction and encoding information into subsystems yields to *operator quantum error correction* [48,49]. Within this framework it is possible to derive simplified active error-cor-

recting schemes where the error is only partially corrected while the residual error affects only the state of a gauge subsystem [8]. These codes yield to a higher threshold for FTQC.

In a communication scenario, quantum information can be sent via teleportation. For this one needs entangled quantum states shared by the sender and receiver. It is possible to distill these entangled states from noisy entangled states [11]. More recently, the construction of QECCs which combine active error correction with pre-shared entanglement has been proposed [13].

Bibliography

1. Aaronson S, Gottesman D (2004) Improved simulation of stabilizer circuits. Phys Rev A 70(5):052328
2. Aharonov D, Ben-Or M (1997) Fault-tolerant quantum computation with constant error. In: Proceedings of 29th ACM symposium on theory of computing (STOC'97). ACM, El Paso, pp 176–188
3. Aliferis P, Gottesman D, Preskill J (2006) Quantum accuracy threshold for concatenated distance-3 codes. Quantum Inf Comput 6(2):97–165
4. Aliferis P, Gottesman D, Preskill J (2008) Accuracy threshold for postselected quantum computation. Quantum Inf Comput 8:181–244
5. Anders S, Briegel HJ (2006) Fast simulation of stabilizer circuits using a graph-state representation. Phys Rev A 73:022334
6. Ashikhmin A, Knill E (2001) Nonbinary quantum stabilizer codes. IEEE Trans Inf Theory 47(7):3065–3072
7. Avizienis A, Laprie JC, Randell B, Landwehr C (2004) Basic concepts and taxonomy of dependable and secure computing. IEEE Trans Dependable Secur Comput 1(1):11–33
8. Bacon D (2006) Operator quantum error-correcting subsystems for self-correcting quantum memories. Phys Rev A 73:012340
9. Barenco A, Berthiaume A, Deutsch D, Ekert A, Macciavello C (1997) Stabilization of quantum computations by symmetrization. SIAM J Comput 26(5):1541–1557
10. Bennett CH, Brassard G, Crepeau C, Jozsa R, Peres A, Wootters W (1993) Teleporting an unknown quantum state via dual classical and Einstein–Podolsky–Rosen channels. Phys Rev Lett 70(13):1895–1899
11. Bennett CH, DiVincenzo DP, Smolin JA, Wootters WK (1996) Mixed-state entanglement and quantum error correction. Phys Rev A 54(5):3824–3851
12. Berlekamp ER, McEliece RJ, van Tilborg HCA (1978) On the inherent intractability of certain coding problems. IEEE Trans Inf Theory 24(3):384–386
13. Brun T, Devetak I, Hsieh MH (2006) Correcting quantum errors entanglement. Science 314(5798):436–439
14. Calderbank AR, Rains EM, Shor PW, Sloane NJA (1998) Quantum error correction via codes over GF(4). IEEE Trans Inf Theory 44(4):1369–1387
15. Calderbank AR, Shor PW (1996) Good quantum error-correcting codes exist. Phys Rev A 54(2):1098–1105
16. Camara T, Ollivier H, Tillich JP (2005) Constructions and performance of classes of quantum LDPC codes. Preprint quant-ph/0502086
17. Chau HF (1998) Quantum convolutional codes. Phys Rev A 58(2):905–909
18. Cleve R, Gottesman D (1997) Efficient computations of encodings for quantum error correction. Phys Rev A 56(1):76–82
19. Cross AW, DiVincenzo DP, Terhal BM (2007) A comparative code study for quantum fault tolerance. Preprint arXiv:0711.1556 [quant-ph]
20. Dennis E, Kitaev A, Landahl A, Preskill J (2002) Topological quantum memory. J Math Phys 43:4452
21. DiVincenzo DP (2000) The physical implementation of quantum computation. Fortschr Phys 48(9–11):771–783
22. Duan LM, Guo CC (1997) Preserving coherence in quantum computation by pairing quantum bits. Phys Rev Lett 79(10):1953–1956
23. Elnozahy EN, Alvisi L, Wang YM, Johnson DB (2002) A survey of rollback-recovery protocols in message-passing systems. ACM Comput Surv 34(3):375–408
24. Forney GD Jr, Grassl M, Guha S (2007) Convolutional and tail-biting quantum error-correcting codes. IEEE Trans Inf Theory 53(3):865–880
25. Gottesman D (1996) A class of quantum error-correcting codes saturating the quantum hamming bound. Phys Rev A 54(3):1862–1868
26. Gottesman D (1997) Stabilizer codes and quantum error correction. Ph D thesis, California Institute of Technology, Pasadena
27. Gottesman D (1998) A theory of fault-tolerant quantum computation. Phys Rev A 57(1):127–137
28. Gottesman D (1999) Fault-tolerant quantum computation with higher-dimensional systems. Chaos, Solitons & Fractals 10(10):1749–1758
29. Gottesman D (2005) Requirements and desiderata for fault-tolerant quantum computing: Beyond the DiVincenzo criteria. http://www.perimeterinstitute.ca/personal/dgottesman/FTreqs.ppt
30. Gottesman D, Chuang IL (1999) Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations. Nature 402:390–393
31. Grassl M (2002) Algorithmic aspects of quantum error-correcting codes. In: Brylinski RK, Chen G (eds) Mathematics of quantum computation. CRC, Boca Raton, pp 223–252
32. Grassl M, Beth T (2000) Cyclic quantum error-correcting codes and quantum shift registers. Proc Royal Soc London A 456(2003):2689–2706
33. Grassl M, Rötteler M (2006) Non-catastrophic encoders and encoder inverses for quantum convolutional codes. In: Proceedings 2006 IEEE International Symposium on Information Theory (ISIT 2006), Seattle, pp 1109–1113
34. Grassl M, Rötteler M (2008) Quantum Goethals-Preparata codes. In: Proceedings 2008 IEEE International Symposium on Information Theory (ISIT 2008), Toronto, pp 300–304
35. Grassl M, Beth T, Pellizzari T (1997) Codes for the quantum erasure channel. Phys Rev A 56(1):33–38
36. Grassl M, Beth T, Rötteler M (2004) On optimal quantum codes. Int J Quantum Inf 2(1):55–64
37. Grassl M, Rötteler M, Beth T (2003) Efficient quantum circuits for non-qubit quantum error-correcting codes. Int J Found Comput Sci 14(5):757–775
38. Hamming RW (1986) Coding and information theory. Prentice-Hall, Englewood Cliffs

39. Haroche S, Raimond JM (1996) Quantum computing: Dream or nightmare? Phys Today 49(8):51–52
40. Kempe J, Regev O, Unger F, de Wolf R (2008) Upper bounds on the noise threshold for fault-tolerant quantum computing. Preprint arXiv:0802.1464 [quant-ph]
41. Ketkar A, Klappenecker A, Kumar S, Sarvepalli PK (2006) Nonbinary stabilizer codes over finite fields. IEEE Trans Inf Theory 52(11):4892–4914
42. Kitaev A (1997) Quantum computations: Algorithms and error correction. Russ Math Surv 52(6):1191–1249
43. Knill E (2005) Quantum computing with realistically noisy devices. Nature 434:39–44
44. Knill E, Laflamme R (1997) Theory of quantum error-correcting codes. Phys Rev A 55(2):900–911
45. Knill E, Laflamme R, Zurek WH (1998) Resilient quantum computation: Error models and thresholds. Proc Royal Soc London Series A, 454:365–384. Preprint quant-ph/9702058
46. Knill E, Laflamme R, Milburn G (2001) A scheme for efficient quantum computation with linear optics. Nature 409:46–52
47. Kraus K (1983) States, effects, and operations. Lecture notes in physics, vol 190. Springer, Berlin
48. Kribs DW, Laflamme R, Poulin D (2005) Unified and generalized approach to quantum error correction. Phys Rev Lett 94(18):180501
49. Kribs DW, Laflamme R, Poulin D, Lesosky M (2006) Operator quantum error correction. Quantum Inf Comput 6(3–4): 382–399
50. Lidar DA, Whaley KB (2003) Decoherence-free subspaces and subsystems. In: Benatti F, Floreanini R (eds) Irreversible quantum dynamics, Lecture Notes in Physics, vol 622. Springer, Berlin, pp 83–120
51. Lidar DA, Chuang IL, Whaley KB (1998) Decoherence-free subspaces for quantum computation. Phys Rev Lett 81(12): 2594–2597
52. MacKay DJC, Mitchison G, McFadden PL (2004) Quantum computations: Algorithms and error correction. IEEE Trans Inf Theory 50(10):2315–2330
53. MacWilliams FJ, Sloane NJA (1977) The theory of error-correcting codes. North-Holland, Amsterdam
54. Matsumoto R (2002) Improvement of Ashikhmin–Litsyn–Tsfasman bound for quantum codes. IEEE Trans Inf Theory 48(7):2122–2124
55. von Neumann J (1956) Probabilistic logic and the synthesis of reliable organisms from unreliable components. In: Shannon CE, McCarthy J (eds) Automata studies. Princeton University Press, Princeton, pp 43–98
56. Nielsen MA, Chuang IL (2000) Quantum computation and quantum information. Cambridge University Press, Cambridge
57. Ollivier H, Tillich JP (2003) Description of a quantum convolutional code. Phys Rev Lett 91(17):177902
58. Preskill J (1998) Reliable quantum computers. Proc Royal Soc London A, 454:385–410. Preprint quant-ph/9705031
59. Rains EM, Hardin RH, Shor PW, Sloane NJA (1997) Nonadditive quantum code. Phys Rev Lett 79(5):953–954
60. Razborov A (2004) An upper bound on the threshold quantum decoherence rate. Quantum Inf Comput 4(3):222–228
61. Reichardt BW (2006) Error-detection-based quantum fault tolerance against discrete Pauli noise. Ph D thesis, University of California, Berkeley
62. Reichardt BW (2006) Fault-tolerance threshold for a distance-three quantum code. In: Proceedings of the 2006 International Colloquium on Automata, Languages and Programming (ICALP'06). Lecture Notes in Computer Science, vol 4051. Springer, Berlin, pp 50–61
63. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423, 623–656, http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html
64. Shor PW (1995) Scheme for reducing decoherence in quantum computer memory. Phys Rev A 52(4):R2493–R2496
65. Shor PW (1996) Fault-tolerant quantum computation. In: Proceedings of 35th Annual Symposium on Fundamentals of Computer Science (FOCS'96). IEEE Press, Burlington, pp 56–65. Preprint quant-ph/9605011
66. Steane AM (1996) Error correcting codes in quantum theory. Phys Rev Lett 77(5):793–797
67. Steane AM (1996) Simple quantum error correcting codes. Phys Rev A 54(6):4741–4751
68. Steane AM (1997) Active stabilization, quantum computation and quantum state synthesis. Phys Rev Lett 78(11):2252–2255
69. Wootters WK, Zurek WH (1982) A single quantum cannot be cloned. Nature 299(5886):802–803
70. Zalka C (1996) Threshold estimate for fault tolerant quantum computation. Preprint quant-ph/9612028
71. Zanardi P, Rasetti M (1997) Noiseless quantum codes. Phys Rev Lett 79(17):3306–3309

Quantum Impurity Physics in Coupled Quantum Dots

Rok Žitko[1], Janez Bonča[1,2]
[1] Jožef Stefan Institute, Ljubljana, Slovenia
[2] Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia

Article Outline

Glossary
Definition of the Subject
Introduction
Quantum Dots as Impurity Systems
Theoretical Tools
Quantum Transport and Kondo Physics
Competing Physical Effects
Universal Behavior Versus Complex Particularities
Future Directions
Bibliography

Glossary

Quantum dot device nanoscopic electronic device resembling a transistor which incorporates a quantum dot as the central active element; sometimes also called single electron transistor. A quantum dot is an extremely small puddle of electrons which can be con-

sidered as an artificial atom since the confinement of electrons leads to quantized energy levels: the electrons form orbitals much like the electrons in orbit around an atomic nucleus. Gate-defined semiconductor quantum dots provide precisely tunable physical realizations of quantum impurity models.

Quantum impurity system system of a localized magnetic impurity in interaction with itinerant free electrons from a conduction band of an otherwise clean metal. It can be described using an idealized quantum impurity model such as Kondo or Anderson model.

Tunneling transmission of electrons from one electrode to another through classically-forbidden potential barriers such as thin insulators or empty space. Tunneling is a characteristic quantum phenomenon that is commonly at play on the nanoscopic scale.

Kondo effect Kondo effect is a many-particle effect which occurs in quantum impurity systems due to increased spin-flip scattering of the conduction band electrons on the magnetic impurity at low temperatures. It leads to various anomalies in thermodynamic and dynamic properties. In the context of the electronic transport through a quantum dot, the Kondo effect is reflected in enhanced conductance (zero-bias anomaly) at reduced temperatures.

Channel In the context of impurity physics, a channel is a set of energy levels in the conduction band which are coupled to the impurity. Several independent channels may be coupled to a single impurity. In the context of quantum dots, the relevant channels may be identified with the conduction channels in the leads attached to the nanostructure, however the number of channels in the effective impurity problem may be lower than the number of physical conduction channels.

Quantum phase transition A quantum phase transition is a zero-temperature phase transition triggered by tuning system parameters. While thermal phase transitions occur due to thermal fluctuations, quantum phase transitions emerge from zero-point quantum fluctuations in the ground state.

Particle-hole symmetry Idealized impurity models exhibit particle-hole symmetry if the model remains unchanged when all occupied levels are mapped into unoccupied levels and vice versa. This occurs for half-filled systems, when precisely one electron occupies each impurity on the average.

Definition of the Subject

Advances in the field of nanoscience and nanotechnology empower us with new tools for probing electronic systems of increasingly small sizes. Nowadays one can, for example, measure electrical conduction of semiconductor quantum dots with lateral extent of a few 10 nm, single molecules, and even individual atoms trapped between two electrodes. Nanodevices of practical interest typically consist of an active element (such as a quantum dot [1,2,3]) weakly coupled to two conducting leads by tunneling junctions so that electric current can flow through the device. The active element confines a small number of electrons. Particularly interesting is the case where this number is an odd integer; the excess single electron is then unpaired and carries magnetic moment. It has been recently demonstrated that a quantum dot of this type behaves as an artificial magnetic atom which can be experimentally tuned using electrodes [4,5,6]. The advantage of performing experiments on such artificial atoms is that various effects that depend on the number of electrons can be studied simply by changing voltages applied on gate electrodes, rather than performing experiments on different chemical elements.

It is now possible to produce nanodevices consisting of a small number of quantum dots which are coupled by tunneling junction between each other and to external electrodes [7,8,9,10,11,12]. Multiple quantum dot systems can be used to study various magnetic effects, such as antiferromagnetic and ferromagnetic ordering, Kondo screening, and other phenomena in which the role of electron-electron interactions is essential. This provides insight into the behavior of similar macroscopic magnetic systems and reveals how magnetic behavior scales from the atomic size. Furthermore, these devices are interesting in their own right as candidates for quantum information storage and processing. They represent the ultimate degree of miniaturization of electronic devices and they are likely to evolve into the building blocks of the circuitry of tomorrow.

Introduction

Real metal is never an ideally clean and homogeneous material. Instead, any metal sample invariably contains a finite concentration of various impurities. Impurities affect resistivity of the metal particularly at low temperatures when electron scattering of thermal origin is suppressed and the residual scattering on static impurities determines at which value the resistivity ultimately saturates. It was remarked very early that in some samples the resistance after the initial decrease unexpectedly increases at the lowest temperatures: this behavior constitutes the "problem of the resistance minimum" [13]. Further experimental work indicated that such anomalies are due to the pres-

ence of magnetic impurities, such as iron, cobalt or manganese, in a non-magnetic metal host. Theoretical understanding was lacking until the work of J. Kondo in 1964 who had shown that the rate of scattering events in which the magnetic moment of the impurity is changed (magnetic or spin-flip scattering) surprisingly increases as the temperature is lowered [14]; this behavior of impurity systems became known as the Kondo effect. More detailed understanding became possible with the development of advanced theoretical tools based on the idea of the renormalization group by P. W. Anderson, K. G. Wilson and others [15].

Impurity models and the Kondo effect are widely studied for several reasons. The Kondo effect is one of the very few non-trivial many-particle effects where an intensive theoretical effort eventually resulted in a very good and detailed understanding. In fact, the Kondo problem was historically the primary motivation for the development of many widely applicable theoretical approaches and has driven the progress in the field of the many-particle physics for many decades. More generally, the impurity models have attracted the condensed matter community due to their unexpectedly complex and rich behavior. On a more practical level, Kondo physics plays an important role in many complex materials which may have practical applications. The Kondo screening of local moments namely competes with magnetic ordering; the result of this competition determines the magnetic properties of materials at low temperatures. Fermi and non-Fermi liquid behaviors, ferromagnetic and antiferromagnetic correlations, and diverse behavior of heavy fermion systems [16] are the outcome of the competition between the Kondo effect and magnetic exchange interaction [8].

More recently, the Kondo problem became popular due to the advances in the field of nanoscience and nanotechnology. It is now possible to perform electron transport measurements on very small systems, such as quantum dots [17], segments of carbon nanotubes [18], single molecules with an embedded magnetic ion [19] and in the extreme case even single magnetic atoms deposited on the surface of a normal metal [20,21,22]. The Kondo effect was predicted to occur in quantum dots in late 1980s [23,24] and experimentally observed a decade later [4,5]. Again, it was found that the Kondo effect leads to transport anomalies at low temperatures. By studying the Kondo physics in systems where parameters can be continuously tuned, we better understand systems where such control is not possible, as in the case of bulk materials. Quantum dot systems are thus a laboratory for studying various effects driven by strong electron correlations.

Quantum Dots as Impurity Systems

Semiconductor Quantum Dots

Particularly interesting devices are quantum dots patterned in high-quality semiconductor heterostructures. In heterostructures a subsurface layer of high-electron-mobility two-dimensional electron gas is confined near the interface between gallium arsenide (GaAs) and aluminum gallium arsenide (AlGaAs) [25]. Such crystal structures may be grown very accurately one atomic layer at a time by molecular beam epitaxy [26]. Lateral quantum dots [27] are then defined by patterning metallic gates on the sample surface (Fig. 1). Using a sufficiently negative gate voltage, the two-dimensional electron gas is depleted in the region below the electrode and a barrier is formed: quantum dot is said to be electrostatically defined. This "split-gate"' technique is also used to build quantum point contacts, quantum wires and similar devices [28]. By changing the voltage on the pinch electrodes, the strength of the coupling of the dot with the electron gas in the leads is controlled. By applying voltage on the gate electrode near the quantum dot region, the number of electrons confined in the dot can be accurately tuned. Nowadays it is possible to fabricate few-electron quantum dots and exactly control the number of electrons starting from zero.

Quantum Impurity Models

An idealized quantum impurity model describes a single point-like impurity (a zero-dimensional defect) in an otherwise homogeneous host environment composed of a gas of particles that form a continuum of extended states. The impurity is assumed to have internal degrees of free-

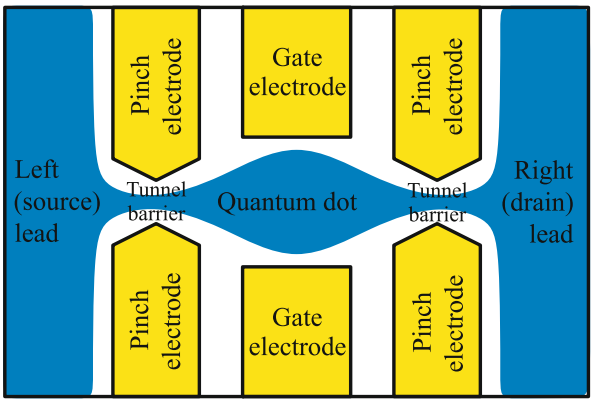

Quantum Impurity Physics in Coupled Quantum Dots, Figure 1
Schematic representation of a semiconductor quantum dot electrostatically defined by the voltages applied on surface metal gate electrodes

dom (such as intrinsic angular momentum, or "spin") and interacts with the continuum particles. A paradigmatic quantum impurity model is the Kondo model for a magnetic impurity atom, such as cobalt, embedded in a host metal which is non-magnetic, such as copper; the magnetic impurity interacts with the conduction band electrons via anti-ferromagnetic exchange interaction. Generalized quantum impurity models may involve several impurities or more complex, non-homogeneous environment. The theoretical significance of the quantum impurity models stems from their ubiquitous applicability to a vast array of physical systems such as bulk Kondo systems, heavy-fermion compounds and other strongly correlated systems, dissipative two-level systems, single magnetic impurities and quantum dots.

In nanoscopic electronic devices the electron-electron interactions are particularly strong and they induce interesting many-particle effects, among them the Kondo effect which appears to be a relatively generic feature of nanodevices [4,19,29,30]. As in bulk systems, the Kondo effect gives rise to various anomalies in the thermodynamic and transport properties, in particular to increased conductance through nanostructures. The conductance through a quantum dot in the Kondo regime is in agreement with theoretical predictions that such dots behave rather universally as single magnetic impurities [17] and can be modelled using single impurity Anderson and Kondo models [17,31]. Quantum dots thus serve as tunable realizations of the quantum impurity models.

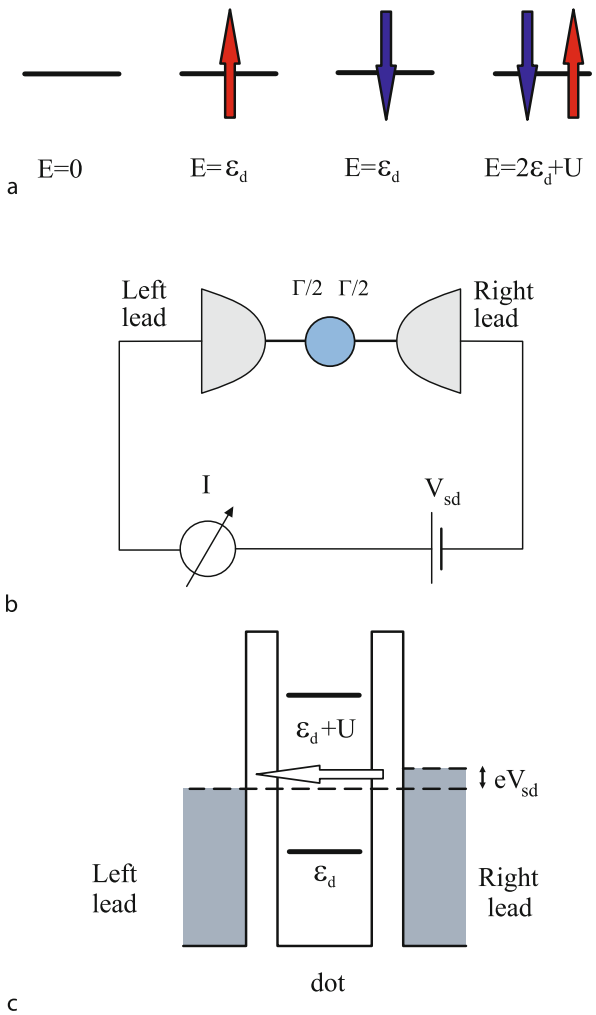

Quantum Impurity Physics in Coupled Quantum Dots, Figure 2
Representations of the single impurity Anderson model for a quantum dot

Anderson Impurity Model

Due to electron confinement, the quantum-mechanical energy levels in a quantum dot form a series of discrete quantized levels. We focus on the electrons in the levels closest to the Fermi level in the leads, i. e. in the last occupied and first unoccupied orbital states of the dot. In the simplest case, a single electron level is relevant and a quantum dot with an odd number of confined electrons is expected to behave as a spin-1/2 magnetic impurity, similar to magnetic ions [32].

In the formalism of the second quantization, the Hamiltonian for interacting electrons in the quantum dot is

$$H_{\text{dot}} = \epsilon_{\text{d}}(n_\uparrow + n_\downarrow) + U n_\uparrow n_\downarrow \,. \tag{1}$$

ϵ_{d} is the energy of the electron orbital in the quantum dot (also named "on-site energy"), U is the strength of the effective electron-electron repulsion between two electrons in the same orbital, and the number operator n_μ is defined as $n_\mu = d_\mu^\dagger d_\mu$, where $d_\mu^\dagger$ and d_μ are the creation and annihilation operators; the spin index μ takes values $\mu = \pm 1/2$ or, equivalently, $\mu = \uparrow, \downarrow$. The on-site energy ϵ_{d} can be regulated using gate voltages which allows the charge state (occupancy) on the dots to be tuned. We may rewrite the Hamiltonian in an equivalent but more symmetric manner as

$$H_{\text{dot}} = \delta n + \frac{U}{2}(n-1)^2 \,, \tag{2}$$

where $n = n_\uparrow + n_\downarrow$ and $\delta = \epsilon_{\text{d}} + U/2$. For $\delta = 0$, the model is particle-hole symmetric and the level is occupied by a single electron on the average. The four possible configurations of the Anderson model and their energies are represented in Fig. 2a.

The conduction bands in the leads are described as

$$H_{\text{band}} = \sum_{k\mu\alpha} \epsilon_k c^{\dagger}_{k\mu\alpha} c_{k\mu\alpha} \,. \tag{3}$$

ϵ_k is the energy of an electron with wave-vector k in left ($\alpha = L$) or right ($\alpha = R$) lead described by the creation/annihilation operator pair $c^{\dagger}_{k\mu\alpha}$ and $c_{k\mu\alpha}$. A conduction band behaves as a sea of electrons: all states below some energy (Fermi level) are occupied, while all other high-energy states are empty. When a source-drain bias voltage V_{sd} is applied on the leads (Fig. 2b), the Fermi levels are displaced and the electrons in an energy interval of width eV_{sd} will attempt to flow from the lead with higher Fermi level through the quantum dot to the other lead, Fig. 2c. Tunneling of electrons through the junctions is described by the Hamiltonian

$$H_{\text{coupling}} = \sum_{k\mu\alpha} V_{k\alpha} (c^{\dagger}_{k\mu\alpha} d_{\mu} + d^{\dagger}_{\mu} c_{k\mu\alpha}) \,, \tag{4}$$

where $V_{k\alpha}$ are the amplitudes for electron tunneling from lead α to the dot. The Anderson impurity model is then given by the sum $H = H_{\text{dot}} + H_{\text{band}} + H_{\text{coupling}}$.

Assuming that the coupling to the left and right electrode is equal, only symmetric combinations $c^{\dagger}_{k\mu\text{L}} + c^{\dagger}_{k\mu\text{R}}$ of conduction band electrons play a role at small bias voltage V_{sd}, while antisymmetric combinations $c^{\dagger}_{k\mu\text{L}} - c^{\dagger}_{k\mu\text{R}}$ are decoupled [23]. The use of a single channel Anderson model is then justified and the index α is unnecessary. This simplification occurs only for simple systems; in general, systems of coupled quantum dots are true multichannel quantum impurity problems.

Often the approximation of taking a constant hopping $V_k \equiv V$ is taken. Further simplification consists of considering the conduction band to have a constant density of states $\rho = 1/(2D)$, where $2D$ is the bandwidth. The hybridization strength Γ which characterizes how strongly the impurity is coupled to the conduction band is then also a constant, $\Gamma = \pi\rho V^2$.

Validity of the approximation of describing the electron in the highest occupied electron level in the quantum dot using the Anderson model has been experimentally well tested: the temperature, magnetic field and gate and bias voltage dependence of the conductance through quantum dots may be described by the simple Anderson model, however the agreement is qualitative, not quantitative [17].

In some parameter regimes, Anderson model reduces to a simpler Kondo model. Kondo model consists of a single spin in interaction with the conduction band:

$$H = \sum_{k\mu} \epsilon_k c^{\dagger}_{k\mu} c_{k\mu} + J\boldsymbol{S}\cdot\boldsymbol{s}(0) \,, \tag{5}$$

where $J \approx 8V^2/U$ is the effective Kondo antiferromagnetic exchange constant, $\boldsymbol{S}$ is the impurity spin-1/2 operator and $\boldsymbol{s}(0)$ is the spin density of the conduction band electrons at the position of the impurity. Despite their seeming simplicity, Anderson and Kondo models are both difficult many-particle problems.

Multi-Impurity Models

Several quantum dots (artificial atoms) can be interconnected to form an "artificial molecule" [33,34]. Systems of multiple coupled impurities are realizations of generalized Kondo models where more exotic types of the Kondo effect may occur. The research in this field has recently intensified due to a multitude of new experimental results; the multi-impurity magnetic nanostructures under study are not only systems of multiple quantum dots [8,9,10,35,36], but also clusters of magnetic adsorbates on surfaces of noble metals (such as Ni dimers [37], Ce trimers [38] and molecular complexes [39]). Systems of two impurities are the simplest systems where one can study a number of very interesting effects, such as the effects of inter-dot electron hopping, formation of ionic or covalent inter-dot bonds, and the competition between magnetic ordering and Kondo screening, leading to quantum phase transitions [40,41]. Recently, few-electron triple quantum dot structures have also been fabricated [11,12] and even more complex multi-dot nanostructures can in principle also be assembled.

Systems of multiple quantum dots can be modeled by suitably generalizing the Anderson model. The properties depend in an essential way on the coupling topology, i. e. on how the various impurities are inter-connected, as represented in the examples in Fig. 3. The system can be modeled using discrete lattice models as an impurity cluster in contact with two conduction leads. Each dot (indexed by the subscript i) is described using a Hamiltonian

$$H_{\text{dot},i} = \delta_i n_i + \frac{U_i}{2}(n_i - 1)^2 \,. \tag{6}$$

Junctions between the dots are described by "hopping Hamiltonian"

$$H_{\text{hopping}} = \sum_{\langle i,j\rangle,\mu} t_{i,j} \left(d^{\dagger}_{i,\mu} d_{j,\mu} + d^{\dagger}_{j,\mu} d_{i,\mu} \right) \,, \tag{7}$$

and junctions between the dots and the conduction leads by suitable generalizations of Eq. (4). The impurity Hamil-

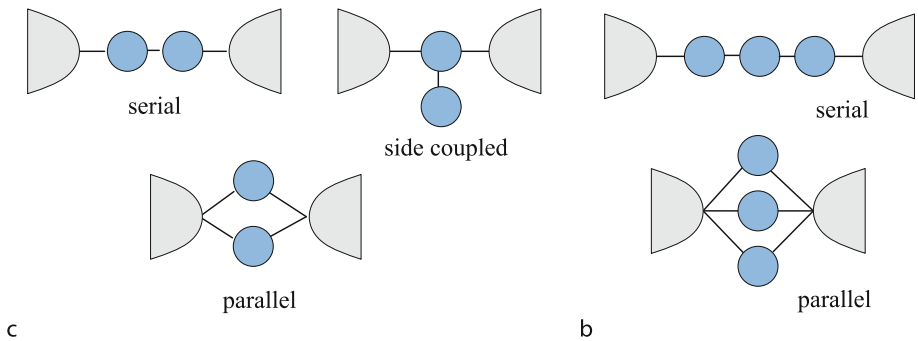

Quantum Impurity Physics in Coupled Quantum Dots, Figure 3
Representations of various multiple-impurity generalizations of the Anderson model

tonian is thus similar in form to the Hubbard model for correlated systems.

Theoretical Tools

Most quantum impurity models are non-perturbative: the commonly used technique of expanding a problem in terms of a small perturbation around an exactly solvable non-interacting model cannot be applied in all parameter regimes due to divergences [13]. The difficulties occur in particular at low temperatures where the systems have anomalous properties. New techniques have been developed to tackle this problem: large-N expansion [42], Bethe-Ansatz [43,44], bosonization-refemionization [45], conformal field theory [46,47], variational methods [48,49] and various renormalization group techniques [15,50,51,52]. Large-N techniques (such as slave-boson mean-field-theory and various improvements) allow in many cases to obtain results in closed form and often provide a qualitatively correct description; in the case of multi-impurity models, however, these methods may fail or they become impractical. Bethe-Ansatz approach provides an exact solution to the thermodynamics of the Kondo model; furthermore, very recently a method to calculate non-equilibrium dynamics was developed [53]. It seems, however, difficult to expand this approach to general multi-impurity models. Bosonization-refermionization technique has been instrumental in providing additional exact results at some special points in the parameter space, however they are less useful for exploring generic problems. Conformal field theory approach based on the non-Abelian bosonization has provided important conceptual insights into the nature of the Kondo effect: the impurity degrees of freedom are engulfed by the continuum and the only residual effect are the modified boundary conditions for continuum electron scattering at the impurity site. The actual implementation of this approach depends from case to case and has not yet been performed for complex multi-impurity problems. Variational methods were the first approach that allowed to study dynamics of quantum impurity models and has been recently generalized to multi-impurity models [54,55,56]. The difficulty in this approach is to correctly describe physics at very low energy scales. Since quantum impurity models become strongly renormalized at low temperatures, the development of renormalization group methods was essential in building correct understanding of the nature of the low-temperature behavior. These methods range from simple scaling of model parameters [52], to mapping to a particle gas [50,51], and finally to Wilson's numerical renormalization group [15,57].

Renormalization

The renormalization is a way of describing and understanding the relation between the different ways a physical system behaves at different energy scales [58]. To study a system at low energies, the irrelevant high-energy degrees of freedom are eliminated from the problem ("integrated out") to obtain an effective description in terms of modified, "renormalized" coupling constants g_i which specify the strengths of various interaction terms in the Hamiltonian, Figs. 4 and 5. The renormalization process can be described using scaling equations, which typically take the form of a system of partial differential equations

$$\partial g_j/\partial l = \beta_j(\{g_i\}) \,, \tag{8}$$

where l is a "running parameter" which runs towards $-\infty$ as the energy scale is decreased and β_j are "beta" functions. Negative beta function corresponds to a relevant coupling constant which grows at low energies, i. e. to an interaction which becomes important at low temperatures. Positive beta function corresponds to an irrelevant coupling

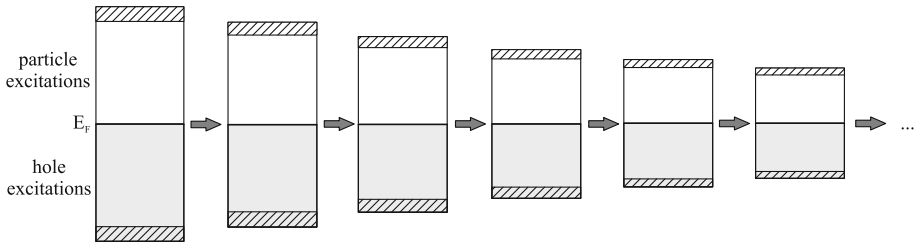

Quantum Impurity Physics in Coupled Quantum Dots, Figure 4
Cutoff renormalization: the particle and hole excitations from the hatched regions at the top and bottom of the conduction band are integrated out to obtain an effective Hamiltonian at lower energy scale

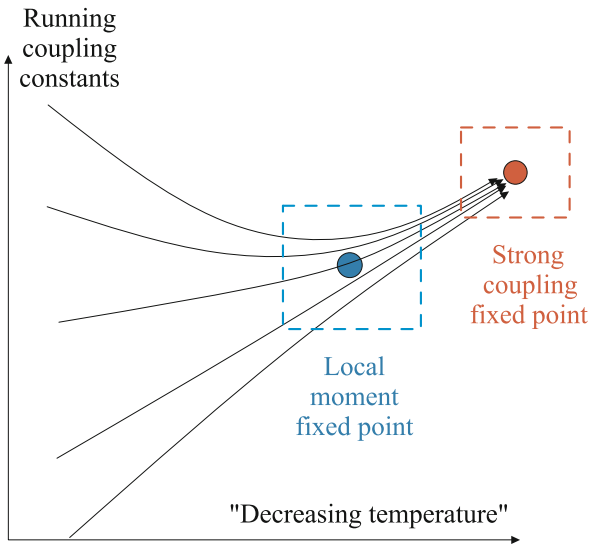

Quantum Impurity Physics in Coupled Quantum Dots, Figure 5
Schematic representation of the renormalization flow in the Anderson model. The *horizontal direction* represents the direction of decreasing energy scale (temperature), while the *vertical direction* represents the multi-dimensional space of the effective Hamiltonians (which can be considered to be parametrized by some large set of coupling constants). When the system is near a fixed point *(dashed boxes)*, its properties can be described by a perturbative expansion around the fixed-point Hamiltonian. The diagram also illustrates the idea of universality: even for widely different original microscopic Hamiltonians, the low-temperature behavior of the systems in the same universality class is essentially the same

constant, which diminishes at low energies. If the coupling constants change only little as the renormalization procedure is performed, the system is said to be near a "fixed point". As the temperature is reduced, the system typically crosses over several times between different fixed points which correspond to particular kinds of system's behavior at different temperature scales, until it ultimately ends up in a stable fixed point which describes the essence of the low energy physics [58].

Generally a simple effective Hamiltonian arises from more complicated ones. A set of model Hamiltonians with the same low energy behavior constitute a universality class, see also Fig. 5. Renormalization is thus an essential ingredient in model building in many-particle theory.

Numerical Renormalization Group

The numerical renormalization group (NRG) was developed in 1970s by K. G. Wilson as a way of numerically exactly solving the Kondo problem [15]. It was later successfully extended to Anderson model and other quantum impurity models [59]. The NRG makes possible to compute the spectrum of excitations of the system, thermodynamic quantities such as magnetic and charge susceptibilities, entropy, and specific heat, dynamic quantities such as spectral functions, dynamical charge and spin susceptibilities, and expectation values of operators such as impurity occupancy, charge fluctuations and spin-spin correlations. The NRG is a non-perturbative method and as such does not suffer from various divergencies as other techniques do. It provides information about the behavior on all temperature scales, from the high-temperature perturbative regime to the low-temperature strong-coupling regime. It can be applied to multi-impurity and multi-channel problems; the complexity of the problem that is still manageable depends on the skillful use of the symmetries present in the problem and ultimately on the available computational resources.

The NRG consists of several steps (Fig. 6):

- Reduction of the quantum impurity problem to an effective one-dimensional problem. Since the impurity is by assumption a zero-dimensional point-like object, it always effectively couples to a continuum of states which can be parametrized by a single variable (or a finite number of such continua that we denote as "channels").
- The one-dimensional continuum of states is discretized into bins (intervals) of geometrically decreasing widths

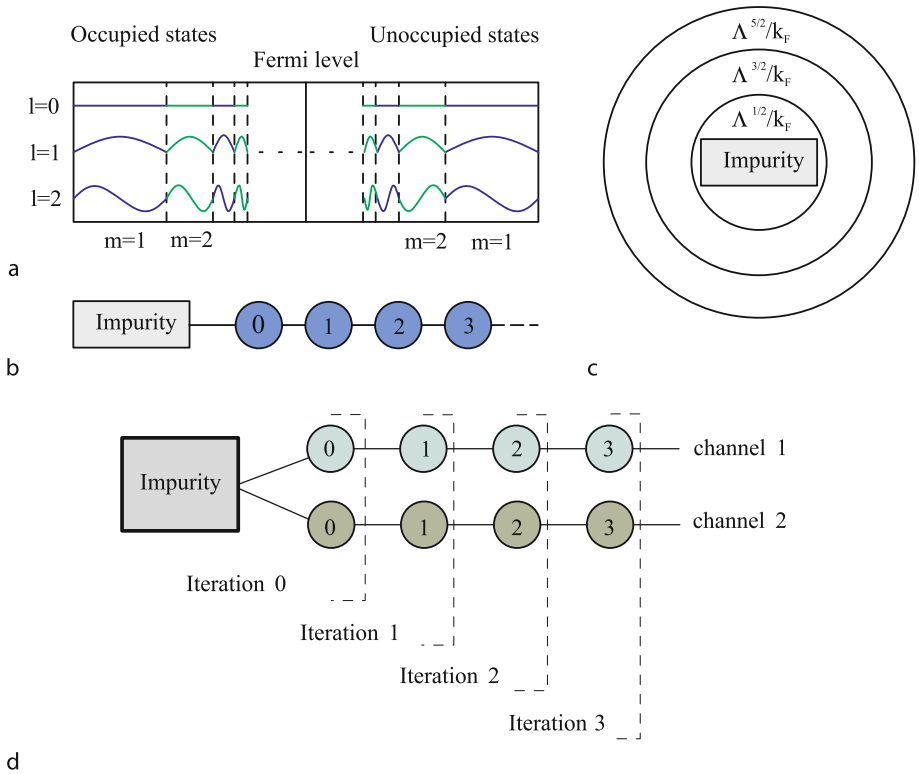

Quantum Impurity Physics in Coupled Quantum Dots, Figure 6

Numerical renormalization group. a Logarithmic discretization. b Chain Hamiltonian (one-channel case). c Onion-shell representation of Wannier orbitals around the impurity. d Chain Hamiltonians and the successive iterations in the NRG procedure: one site from each channel is added during each renormalization group transformation

proportional to Λ^{-m}, where the parameter Λ controls the fineness of the discretization and m is the bin index, Fig. 6a. The continuum limit is recovered for $\Lambda = 1$, while in practical calculations $\Lambda \gtrsim 2$ is used. In each interval, a spectral Fourier decomposition is performed (index l in Fig. 6a). In practical calculations, only the lowest $l = 0$ Fourier mode is retained in each interval, i. e. an interval of states is represented by the energy-averaged state. This procedure is named logarithmic discretization since the continuum degrees of freedom near the Fermi level are described with a logarithmic accuracy. A further transformation allows the problem to be formulated as an impurity attached to a one-dimensional chain of sites with exponentially decreasing hopping parameters, Fig. 6b. The sites in this chain can be interpreted as forming "onion shell"-like orbitals encircling the impurity, Fig. 6c.

- Iterative diagonalization of the chain Hamiltonian is performed, Fig. 6d. The first step consists of an exact diagonalization of the initial cluster, typically composed of the impurity sites and one chain site for each continuum channel in the problem. Additional sites are then added consecutively, one from each channel in every iteration: a new Hamiltonian is constructed and diagonalized exactly. In NRG, this procedure represents the renormalization group transformation and the iteration corresponds to the renormalization flow.
- The problem of the exponential growth of the size of the Hilbert space with the number of sites in the chain is alleviated by truncating the number of states retained at each iteration to a predefined small number of the order thousand. This turns out to be a good approximation for quantum impurity problems since there is little mixing between low-energy and high-energy excitations as the chain sites are added at each step (this property is known as the energy-scale separation).

Each iteration corresponds to the behavior of the system on a temperature scale $T_N \propto \Lambda^{-N/2}$, where N is the iteration number. The full description of the system at step N consists of the eigenstates and irreducible matrix elements for creation operators $f^{\dagger}_{N\mu\alpha}$ in the chain. This description

is clearly much more complex compared to that in the simple renormalization approach where a small set of running coupling constants is used; the advantage is that the NRG is unbiased and, in some sense, essentially exact.

Quantum Transport and Kondo Physics

Experiments

Experiments probing fundamental properties of semiconductor quantum dots are typically performed in helium-3 dilution refrigerators at extremely low temperatures in the range of 100 mK or even less. At low temperatures, electrons occupy distinct energy levels and the Coulomb energy plays a crucial role [26]. Performing experiments at the lowest attainable temperatures is important since the energy resolution of spectroscopic techniques used is limited solely by the sample temperature [26]. Systems are characterized by performing gate-voltage and bias-voltage sweeps (gated transport spectroscopy), or by magnetic spectroscopy. This allows to obtain information about the energy levels, number of confined electrons, and electron–electron repulsion. Furthermore, finite-bias current can be approximately related to the impurity spectral function at finite frequencies.

Three elements affect the transport properties of coupled quantum dots in a characteristic manner: quantum coherence, discrete nature of the electric charge and strong electron–electron interactions.

In nanodevices made of very clean semiconductors the coherence length of electrons at low temperatures exceeds the size of the device through which the electric current flows; electrons then travel coherently through the system and behave in a wave-like manner so that quantum mechanical interference effects can occur. As the electrons scatter only off the boundaries (walls) of the device, rather than on the defects or phonons, the transport is said to be ballistic.

The conductance through nanoscopic constrictions is often found to be quantized in terms of the conductance quantum,

$$G_0 = 2e^2/h = e^2/\pi\hbar \approx 12.9\,\mathrm{k}^{\cdot -1}\,. \tag{9}$$

This is the conductance of a fully transmitting single-mode conduction channel taking into account both spin orientations and is experimentally measured in quantum point contacts and quantum wires. In lateral quantum dots the tunnel barriers from the 2DEG to the quantum dot are obtained by successively pinching off the propagating channels using the gate electrodes. When the last channel is pinched off, the Coulomb blockade regime develops. In this regime, only one channel from each lead is coupled to the dot.

Coulomb Blockade and Cotunneling

According to the analogy between a quantum dot and an atom, we expect that removing an electron from the dot (or adding it) takes energy as this is similar to the ionization of an atom. The transfer of an electron electrically charges the dot and increases the electrostatic energy by $E_C = e^2/2C$ where C is the effective capacitance between the dot and the surrounding electrodes. If the available energy is lower than the charging energy E_C (i. e. for small voltage drop across the system and for low temperature), the conductance is suppressed. This is the Coulomb blockade effect. Unless the energies of quantum dot configurations with N and $N+1$ confined electrons happen to be aligned by suitably tuning the gate voltages (Fig. 7a), the current can flow only by cotunneling (high-order processes in hybridization strength Γ) through the virtual state with excess energy $\sim E_C$ [26,60], Fig. 7b.

Cotunneling is an electric conduction process whereby an electron makes a virtual transition to a high-energy excited intermediate state in the quantum dot to travel from source to destination electrode in a single quantum step. It is to be opposed to a sequential tunneling process, where the electron makes a real transition to an energetically accessible state inside the device and the tunneling proceeds in two steps. Cotunneling is a characteristically quantum phenomenon related to the Heisenberg's uncertainty principle and becomes relevant at low temperatures. Occupation of the virtual state is allowed for a short time, $\sim h/E$, where h is the Planck constant and E the energy cost involved.

In spin-flip co-tunneling process the impurity spin is effectively flipped from spin up to spin down, or vice versa: electron with a given spin orientation tunnels in, while another electron with the opposite spin orientation tunnels out, Fig. 7c. Processes of this type are responsible for the emergence of the Kondo effect.

Conductance Formulas

Particularly important transport quantity is the conductance in the limit of zero source-drain bias voltage $G = \lim_{V_{\mathrm{sd}}\to 0} I/V_{\mathrm{sd}}$, i. e. the linear response of the system to an imposed bias. Linear conductance is an equilibrium property of the system which can be reliably calculated for impurity models. Finite-bias problems require non-equilibrium techniques which are not yet developed to a comparable degree.

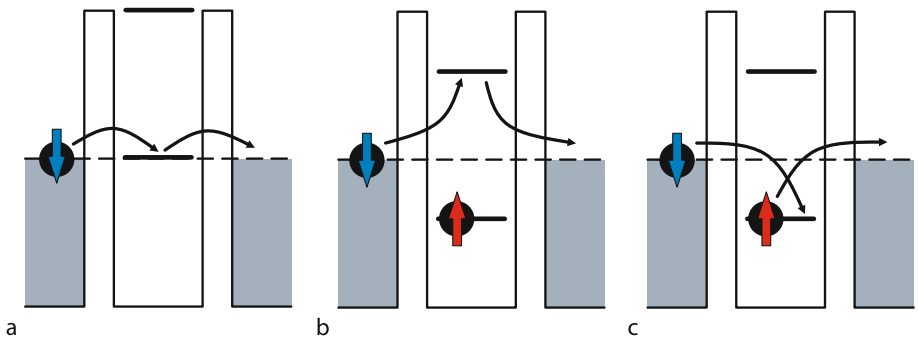

Quantum Impurity Physics in Coupled Quantum Dots, Figure 7
a First-order tunneling, b cotunneling, c cotunneling with a spin-flip

For the purpose of theoretical modeling, an electronic nanodevice may be idealized and considered as a very small scatterer embedded between two metallic contacts. According to R. Landauer, the conductance of a coherent mesoscopic device is related to the transmission probability for incident electrons [61], i. e. to the scattering properties of the impurity region. At zero temperature, the conductance is simply proportional to the transmission probability [62,63]

$$G(T=0) = G_0 |S_{\mathrm{RL}}|^2 \,, \tag{10}$$

where S_{RL} is the right-left component of the scattering matrix, i. e. the amplitude for the scattering of an electron from right to left lead.

In the vast majority of the quantum impurity problems the system behaves at low temperatures as a local Fermi liquid even if it is strongly renormalized. Fermi liquid systems are described in terms of weakly interacting quasiparticles and are fully characterized by the quasiparticle scattering phase shifts which quantify how the quasiparticles scatter in the quantum dot structure [17,64,65]. In the absence of the magnetic field, a single phase shift δ_{qp} per channel is required. Matrix element S_{RL} can be expressed in terms of the phase shifts and an additional angle parameter θ which depends on the symmetry of the problem, yielding the following conductance formula [66]:

$$G = G_0 \sin^2(2\theta) \sin^2(\delta^a_{\mathrm{qp}} - \delta^b_{\mathrm{qp}}) \,. \tag{11}$$

For left-right symmetric problems, it is found that $\theta = \pi/4$ and that the relevant channels are formed by the symmetric (even) and antisymmetric (odd parity) linear combinations of the conduction electron states from left and right lead, so that

$$G = G_0 \sin^2(\delta^{\mathrm{even}}_{\mathrm{qp}} - \delta^{\mathrm{odd}}_{\mathrm{qp}}) \,. \tag{12}$$

Conductance can also be computed from the impurity spectral functions using the Meir–Wingreen formula [67]:

$$G = G_0 \left(-\mathrm{ImTr}\left[\boldsymbol{\Gamma}\mathbf{G}^r\right]\right) \,, \tag{13}$$

where $\boldsymbol{\Gamma}$ is a coupling matrix and $\mathbf{G}^r$ is the matrix of retarded Green's functions of the impurity region. Green's function is essentially the Fourier transform of the probability amplitude for adding an electron to the impurity and extracting it at a later time. The imaginary part of the Green's function is the impurity spectral function. Peaks in the spectral function correspond to electronic excitations of the quantum dot and the value of the spectral function at the Fermi level is directly related to the conductance at zero temperature and can be related to the quasiparticle phase shifts in Fermi liquid systems. Meir–Wingreen formula is actually more general and it is in particular valid also for systems which are not Fermi liquids.

The Kondo Effect

The Kondo effect arises due to strongly enhanced spin-flip scattering of the conduction band electrons on the impurity at low temperatures. In conventional bulk Kondo systems this leads to increased resistivity since electrons scatter isotropically in all directions which impedes the flow of the current. As a consequence, the temperature dependence of the resistivity is non-monotonic and exhibits a minimum at small temperatures, which was the first experimentally observed manifestation of the Kondo effect [14]. Curiously, in quantum dot systems the increased scattering leads to the opposite behavior: at very low temperatures the conductance increases up to the theoretical limit of one conductance quantum, G_0 [68],. The origin of this seeming discrepancy lies in the reduced dimensionality of the problem. In quantum dot problems, scattering is effectively one-dimensional: backwards (reflection back to

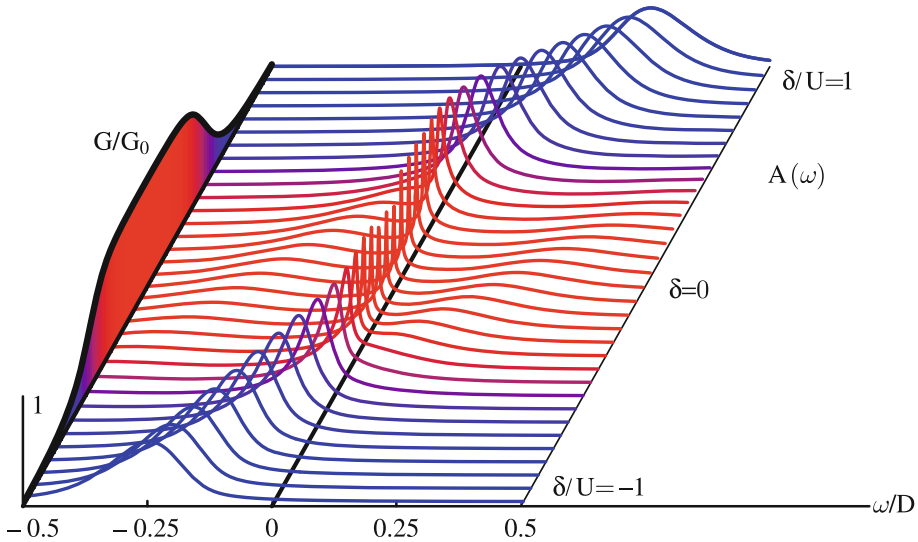

Quantum Impurity Physics in Coupled Quantum Dots, Figure 8
Spectral functions $A(\omega)$ and conductance through a quantum dot described by the single-impurity Anderson model for a range of parameters δ with $U/D = 0.5$, $\Gamma/U = 0.08$. Color of each spectral function corresponds to the value of the conductance

the lead) or forwards (transmission to the other lead). The scattering increases in forward direction, which in this case corresponds to increased electric current.

The temperature scale where the scattering increases is called the Kondo temperature, T_K. The Kondo effect is not a phase transition, but rather a cross-over, therefore the change in conductance is a very smooth function of the temperature. The Kondo temperature is a non-analytic function of model parameters, $T_K \propto \exp(-1/\rho J)$, where ρ is the density of states in the conduction band at the Fermi level and J is the effective Kondo exchange constant. It may be noted that the exponential dependence of T_K reflects the non-perturbative nature of this problem.

At temperatures much below T_K, the impurity spin is screened by the conduction band electrons and the system as a whole is non-magnetic. Properties below T_K are universal and can be described using functions of argument T/T_K; a single parameter T_K fully characterizes the system instead of the microscopic ϵ_d, U, Γ, etc.

The Kondo exchange scattering processes generate an additional resonance in the impurity spectral function at the Fermi level. This "Kondo resonance" is of many-particle origin: the correlated behavior of a large number of electrons is required to produce it. Since properties at low temperatures are predominantly determined by the electron states near the Fermi level, the Kondo resonance significantly modifies the behavior of the system. In particular it leads to the predicted increase of the conductance through the quantum dot, Fig. 8.

The Kondo effect is clearly a magnetic effect related to electron spin. As such, it is strongly perturbed when an external magnetic field is applied. The Kondo resonance splits, the zero-bias voltage is reduced and there are two conductance peaks at finite bias. The characteristic magnetic-field dependence of the transport properties is an ultimate proof that Kondo physics is at play in a nanostructure.

Competing Physical Effects

Physical systems tend to reduce their entropy as the temperature is lowered. In the context of systems of coupled quantum impurities with spin degrees of freedom, this is most often achieved either by Kondo screening, or by magnetic ordering of some kind [69]. Both mechanism of relieving the entropy can be in competition which leads to interesting behavior [69].

Inter-Impurity Magnetic Interactions

There are several possible origins of the inter-impurity exchange interaction in quantum impurity models. One is the super-exchange mechanism which is mediated by the direct inter-impurity electron hopping (tunneling). Virtual excursions of an electron from one impurity to another modify the energy: it is reduced if the spins are anti-aligned, so that the effective exchange interaction is antiferromagnetic. It may be represented by an interaction term $J_{\text{eff}}\boldsymbol{S}_1 \cdot \boldsymbol{S}_2$, where $\boldsymbol{S}_i$ is the impurity spin operator on

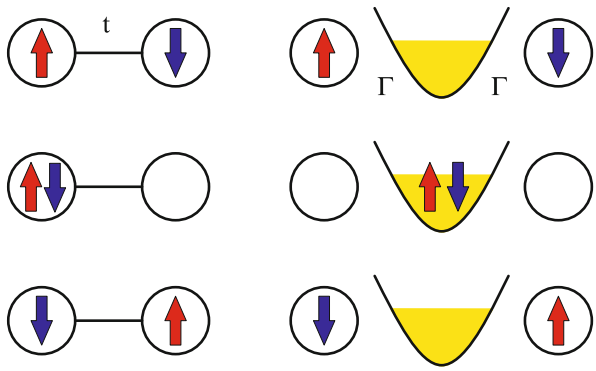

Quantum Impurity Physics in Coupled Quantum Dots, Figure 9
Processes leading to effective inter-impurity exchange interaction. a Superexchange interaction due to electron tunneling between the dots. b RKKY interaction mediated by the conduction band

dot i. For two dots decoupled from the leads, the exchange constant is given by the expression

$$J_{\text{eff}} = \frac{t}{2}\left(\sqrt{\left(\frac{U}{t}\right)^2 + 16} - \frac{U}{t}\right) \approx \frac{4t^2}{U}\,, \tag{14}$$

where t is the inter-impurity hopping parameter (tunneling amplitude).

Another important mechanism is the Ruderman–Kittel–Kasuya–Yosida (RKKY) effective exchange interaction. One impurity polarizes the conduction band electrons in its vicinity; these electrons in turn interact with the other impurity. The intensity and the sign of the resulting exchange interaction depends on the inter-impurity separation. It is ferromagnetic at very short distances, it oscillates with a period proportional to $1/k_F$ where k_F is the Fermi momentum, and decays as $(k_F R)^{-D}$ where D is the effective dimensionality of the conduction band electron gas.

Effects of the Dot-Lead and Inter-Dot Coupling Topology

In multiple dot systems, the coupling of the quantum dots between each other and to the conduction leads affects the conductance in a non-trivial way since the entire system behaves in a quantum coherent way and no part of the system may be considered separately from other parts. Simple circuit theory is *not* applicable. The effects of the coupling topology can be conveniently studied in the case of double quantum dot [70]. As an illustration, we will consider the significantly different behavior of serial and side-coupled double quantum dot [70,71].

Serial Double Quantum Dot The zero-temperature conductance of two quantum dots coupled in series between two electrodes is shown in Fig. 10a for a range of the inter-dot hopping parameters t. The conductance is calculated from the scattering phase shifts obtained in NRG calculations. For large t, the coupling between the dots is strong and the system behaves as an artificial molecule composed of two atoms. When the system is occupied by an odd number of electrons, only the unpaired electron plays an important role and the system may be mapped to an effective single-impurity Anderson model where the role of the in purity orbital is played by the bonding (symmetric) or anti-bonding (anti-symmetric) "molecular orbital". Thus for $t/U = 0.5$, the conductance is high for $|\delta/U| = 0.5, \ldots, 1.2$ when the system is occupied by 1 or 3 electrons. For $|\delta/U| < 0.5$ the conductance is low.

As t decreases, the system starts to behave as two localized magnetic moments and may be approximately described by the two-impurity Kondo model. It is found that the conductance at the particle-hole symmetric point attains the theoretical limit of G_0 at the point where the inter-impurity exchange interaction J_{eff} is comparable to the scale of the Kondo temperature for one impurity coupled to a single conduction lead, T_K, i. e. for $J_{\text{eff}} \sim T_K$ (which corresponds to $t/U \sim 0.05$, see Fig. 10a). In the true two-impurity Kondo model, this point in the parameter space corresponds to a quantum phase transition between a regime of local antiferromagnetic singlet and a regime of separate Kondo screening of each impurity moment. In the double quantum dot system, however, this quantum phase transition is replaced by a smooth crossover due to charge transfer between the two conduction leads [72].

Finally, for very small t, the conductance tends to zero for all values of δ as the system becomes separated in two parts which no longer communicate.

Side-Coupled Double Quantum Dot In the side-coupled configuration, the first quantum dot is embedded between source and drain electrodes while the second dot is coupled to the first through a tunneling junction; there is no direct coupling of the second dot to the leads. By changing the gate voltage and the inter-dot tunneling rate, this system can be tuned to one of the following low-temperature regimes: i) a non-conducting local spin-singlet state, ii) the conventional Kondo regime with odd number of electrons occupying the dots, iii) the two-stage Kondo regime with two confined electrons, or iv) a valence-fluctuating state [71]. In addition, at finite temperatures a Fano resonance appears in the conductance; its origin lies in the sudden filling of the side-

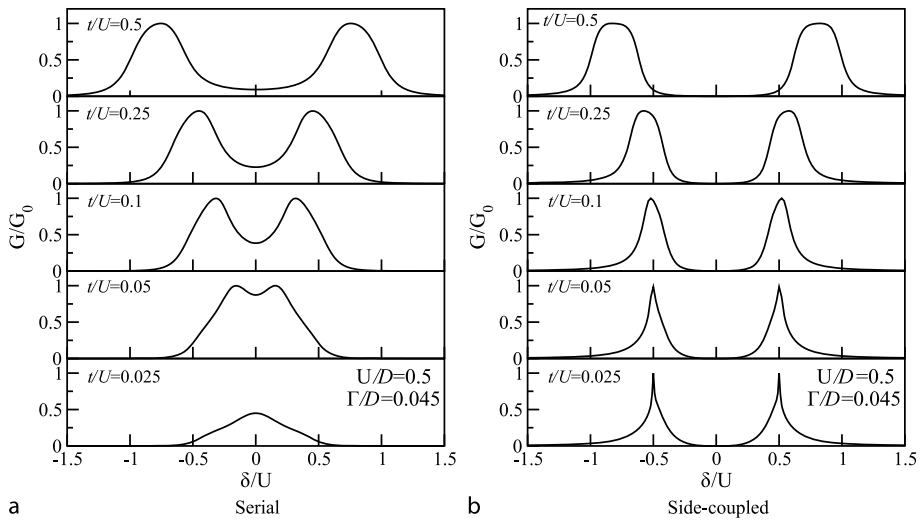

Quantum Impurity Physics in Coupled Quantum Dots, Figure 10

Zero-temperature conductance G/G_0 of the double quantum dot system in a serial configuration and b side-coupled configuration as a function of the gate voltage for a range of the inter-dot hopping parameters t

coupled dot when its on-site energy crosses the Fermi level [71].

For large inter-dot tunneling coupling t, there are two wide regimes where the conductance is enhanced due to the conventional Kondo effect, for example in the ranges $|\delta/U| = 0.5, \ldots, 1.5$ for $t/U = 0.5$ (Fig. 10b) when the dot is occupied by 1 or 3 electrons. These regimes are separated by a low-conductance regime where the localized spins of two electrons are antiferromagnetically coupled for $|\delta/U| \lesssim 0.5$. For large t, the side-coupled and serial configurations of quantum dots thus have qualitatively similar properties.

For small t, in the two stage Kondo regime the two local moments are screened at different Kondo temperatures [71,73,74,75]. The *two-stage Kondo effect* is a generic name for successive Kondo screening of the impurity local moments at different temperatures [66,71,73,74,76,77,78,79]. This term has been used in two different (but closely related) contexts: 1) two-step screening of a $S = 1$ spin in the presence of two channels [76], 2) two step screening of two local moments in the single-channel case [73,77,78]. In the first case, the first-stage Kondo screening is an underscreened spin-1 Kondo effect which reduces the spin to 1/2, while the second-stage Kondo screening is a perfect-screening spin-1/2 Kondo effect which leads to a spin singlet ground state [77,80]. This first case is relevant when the lowest-energy impurity configuration is a spin triplet. In the second case, at a higher Kondo temperature $T_K^{(1)}$ the Kondo effect occurs on the more strongly coupled impurity; the Fermi liquid quasiparticles associated with the Kondo effect on the first impurity participate in the Kondo screening of the second impurity on an exponentially reduced Kondo temperature scale $T_K^{(2)}$ [71,73,74]. This case occurs when the lowest-energy configuration is a singlet, but there is a nearby excited triplet state [78].

In the double quantum dot system, the two-stage Kondo effect occurs when the effective exchange interaction between the dots is such that $J_{\text{eff}} < T_K$, where $T_K = T_K^{(1)}$ is the Kondo temperature of the single-impurity Anderson model that describes impurity 1 (without impurity 2) [71,74]. The second Kondo crossover then occurs at

$$T_K^{(2)} = c_2 T_K^{(1)} \exp(-c_1 T_K^{(1)}/J_{\text{eff}}) \,. \tag{15}$$

Constants c_1 and c_2 are of the order of 1 and they are problem-dependent. The spectral function $A_1(\omega)$ of impurity 1 increases at $\omega \sim T_K^{(1)}$, but then drops at $\omega \sim T_K^{(2)}$, i. e. there is a gap in the spectral function and the system is non-conducting at zero temperature.

The conductance increases to G_0 on the temperature scale of $T_K^{(1)}$, then drops to zero on the scale of $T_K^{(2)}$. The conductance can be high at finite temperatures even in the vicinity of the particle-hole symmetric point, $\delta = 0$, if $T_K^{(2)} < T < T_K^{(1)}$.

Capacitive Coupling and Charge Ordering

The effect of the inter-impurity electron repulsion (induced by capacitive coupling between two quantum dots)

may be modeled using the following Hamiltonian term:

$$H_{\text{dots}} = \sum_{i=1}^{2} H_{\text{dot},i} + U_{12}(n_1 - 1)(n_2 - 1)\,. \tag{16}$$

The inter-impurity repulsion is not an important perturbation as long as $U_{12} < U$; finite U_{12} only modifies the Kondo temperature, while the behavior of the system remains qualitatively unchanged [81]. For $U_{12} > U$ the electrons can lower their energy by forming on-site singlets and the system enters the *charge-ordering regime* [82]. The system behaves in a peculiar way at the transition point $U_{12} = U$, when an intermediate temperature fixed point with a six-fold symmetry of states appears. In serial dots, this leads to an exotic SU(4) Kondo effect [56,82]. For parallel dots, however, the coupling of impurities to the leads breaks the orbital symmetry and conventional Kondo screening occurs [81].

Universal Behavior Versus Complex Particularities

Near the particle-hole symmetric point (or, equivalently, at half filling when one electron occupies each quantum dot on the average), systems consisting of even or odd number of quantum dots have radically different behavior due to the distinct properties of integer and half-integer spin states. The half-integer spin states are always degenerate and quantum dot systems with such impurity configuration tend to exhibit some form of the Kondo effect for any coupling strength; the zero-temperature conductance of systems of an odd number of dots will tend to be high. In systems with an even number of quantum dots, however, the range of half filling is generally associated with Mott–Hubbard insulating behavior [65]: the conductance for a half-filled system decreases exponentially with electron–electron repulsion U [83]. Actual behavior also crucially depends on the coupling topology. The cases of serial and parallel dots will be considered in the following.

Linear Chains of Quantum Dots

The simplest non-trivial system with an odd number of quantum dots consists of three quantum dots coupled in series between two conduction leads. This triple quantum dot system is usually modelled as a three-site Hubbard chain. The special feature of this system is the presence of two equivalent screening channels combined with two-stage Kondo screening and/or magnetic ordering. Triple quantum dot structures have been manufactured in recent years and the analysis of their stability diagrams demonstrates that a description in terms of a Hubbard-like model is in deed a good approximation [11,12].

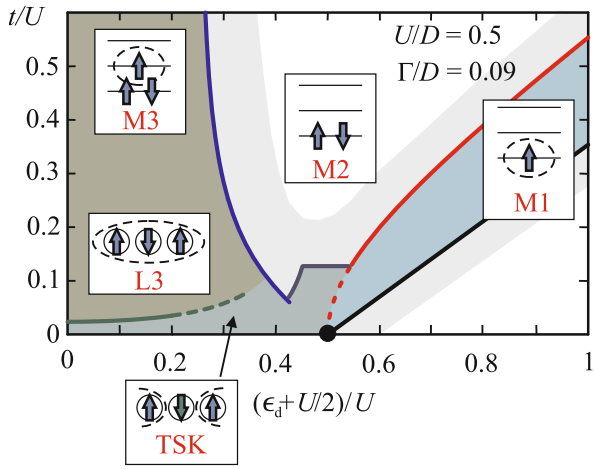

Quantum Impurity Physics in Coupled Quantum Dots, Figure 11
M1, M3: molecular-orbital Kondo regime with $\langle n\rangle \sim 1, 3$. M2: non-conductive even-occupancy state. L3: local Kondo regime with $\langle n\rangle \sim 3$. TSK: two-stage Kondo regime. Due to the particle-hole symmetry of the problem, the diagram is mirror-symmetric with respect to the $\delta = \epsilon_d + U/2 = 0$ axis; for negative $\delta < 0$ we thus find M4 non-conductive regime and M5 molecular-orbital Kondo regime

The behavior of the system depends strongly on the values of the on-site energies and on the inter-impurity hopping. Based on extensive calculations using several complementary methods, a phase diagram has been established, Fig. 11 [55]. It indicates the parameter ranges where the zero-temperature conductance is high.

For strong inter-impurity coupling t, the system may be mapped to an effective single-impurity Anderson model where the role of the impurity orbital is played by the bonding, non-bonding, or anti-bonding "molecular orbital". In this regime, the conductance is high when the occupancy is odd, and it is nearly zero when the occupancy is even, see Fig. 12a for $t/U = 1.0$, 0.5 and 0.2. For smaller inter-impurity coupling ($t/U \lesssim 0.1$), the molecular orbital description becomes inappropriate as the local behavior of the spins becomes important. The system then behaves as a necklace of magnetic atoms, rather than as a strongly-bound molecule.

When there are three electrons in the dots (i. e. for $\delta = 0$) and the coupling t is gradually decreased, the system crosses over from the molecular orbital M3 regime ($t \gtrsim U$) to the antiferromagnetic spin-chain L3 regime ($J_{\text{eff}} \sim t$), and finally to the two-stage Kondo (TSK) regime ($J_{\text{eff}} < T_{\text{K}}^{(1)}$), see Fig. 11. In the spin-chain regime, the three spins lock at $T \sim J_{\text{eff}}$ into a rigid spin-1/2 spin-chain state; at lower temperature, this collective spin is screened by the conventional spin-1/2 Kondo effect. In

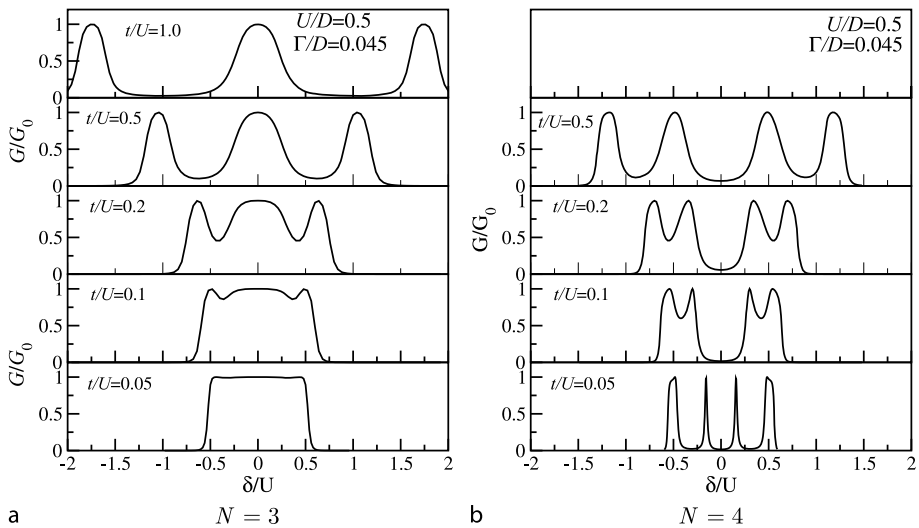

Quantum Impurity Physics in Coupled Quantum Dots, Figure 12
Conductance G/G_0 of the triple and quadruple quantum dot systems as a function of the gate voltage for a range of inter-dot hopping parameters t

the two-stage Kondo regime, the spins on the first and third sites are screened at higher Kondo temperature $T_K^{(1)}$, then the spin on the central site is screened at an exponentially reduced second Kondo temperature $T_K^{(2)} \propto T_K^{(1)} \exp(-cT_K^{(1)}/J_{\text{eff}})$, where $J_{\text{eff}} \approx 4t^2/U$.

Antiferromagnetic and two-stage Kondo regimes are separated by a cross-over region with unusual properties at finite temperatures where the system approaches the so-called *two-channel Kondo model* non-Fermi liquid fixed point [75]. The non-Fermi liquid behavior emerges as the temperature is decreased below the Kondo temperature (T_{scr} screening temperature in Fig. 13) and disappears below the temperature T_Δ below which the behavior again corresponds to that of a Fermi liquid system.

Non-Fermi liquid behavior can be experimentally detected by measuring the differential conductance in a three-terminal configuration (see the insets in Fig. 14). The qualitative temperature dependence of the zero-bias conductance through the system can be approximately inferred from the frequency dependence of the spectral functions. The conductance through the system (from left to right conduction lead) is given by $G_{\text{serial}}/G_0 \approx 4(\pi\Gamma A_{13})^2$ [84] and the conductance through a side dot in the three-terminal configuration by $G_{\text{side}}/G_0 \approx \pi\Gamma A_1$ [67]. The appropriately normalized spectral densities are shown in Fig. 14 for the cases of cross-over regime with a non-Fermi liquid region and the antiferromagnetic regime with no discernible non-Fermi liquid behavior. When non-Fermi liquid fixed point is approached (for $t/D = 0.005$ and $T_\Delta \lesssim T \lesssim T_{\text{scr}}$), the conductance $G_{\text{side}} \sim G_0/2$, while $G_{\text{serial}} \sim 0$. The increase of the conductance G_{serial} through the system at $T \lesssim T_\Delta$ is concomitant with the cross-over from the non-Fermi liquid to Fermi liquid fixed point [72]. In the antiferromagnetic regime with no non-Fermi liquid region, both conductances increase below the same temperature scale, i. e. at $T \lesssim T_{\text{scr}}$.

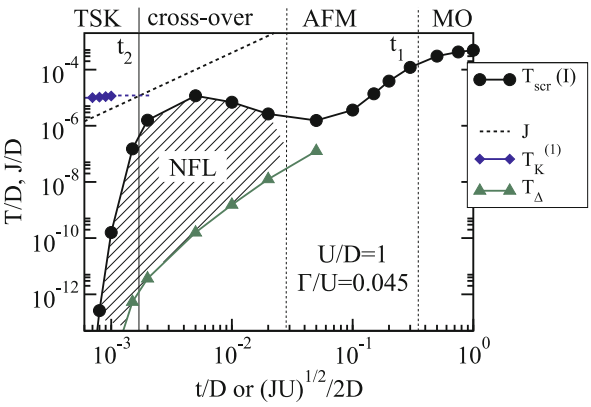

Quantum Impurity Physics in Coupled Quantum Dots, Figure 13
Cross-over scales of triple quantum dot as function of the inter-dot coupling. The magnetic screening temperature T_{scr} is defined by $T_{\text{scr}}\chi(T_{\text{scr}})/(g\mu_B)^2 = 0.07$; it is equal to the Kondo temperature when screening is due to the single-channel Kondo effect. T_Δ is defined through $s_{\text{imp}}(T_\Delta)/k_B = \ln 2/4$, where $s_{\text{imp}}(T)$ is the impurity contribution to the total entropy at temperature T. Here ln 2/4 is half the impurity entropy in the non-Fermi liquid fixed point

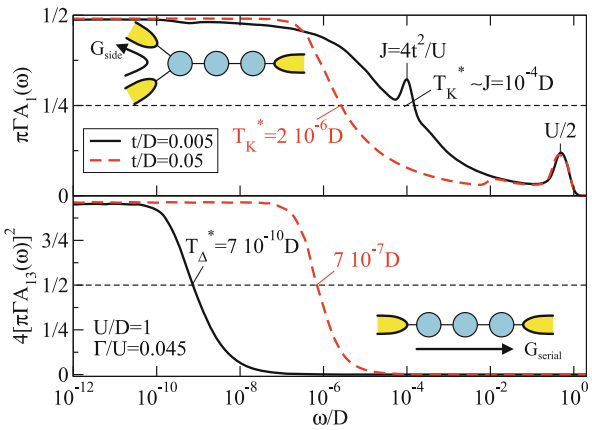

Quantum Impurity Physics in Coupled Quantum Dots, Figure 14
Dynamic properties of triple quantum dot in the antiferromagnetic *(dashed lines)* and in the cross-over regime *(full lines)*. *Upper panel*: on-site spectral function $A_1(\omega)$ of the left dot. *Lower panel*: out-of-diagonal spectral function $A_{13}(\omega)$ squared. Temperature T_Δ^* is of order T_Δ, T_K^* is of order T_{scr}

The conductance for four quantum dots coupled in series is shown in Fig. 12b. For large t the description in terms of molecular orbitals is again appropriate and we observe four conductance peaks [65]. There is a wide region of low conductance around the particle-hole symmetric point which corresponds to the Hubbard gap, and two pairs of conductance peaks which correspond to the Hubbard sub-bands. As t is reduced, the two inner peaks become rapidly extremely sharp, while the outer peaks centered at $|\delta/U| = 1/2$ narrow down more progressively. For very small t, the system is fully insulating at zero temperature for (almost) all values of δ.

On one hand, at zero temperature short chains of even and odd number of dots in a chain are seen to have widely different properties. On the other hand, for a very large number of dots, i. e. in the limit of a macroscopic system, insulating behavior is expected at half-filling irrespective of the even or odd parity of N. These two contrasting predictions can be reconciled by considering the order of taking the $T \to 0$ and $N \to \infty$ limits [65]. Taking the $T \to 0$ limit first, the even/odd alternation is obtained. If the $N \to \infty$ limit is taken first, which is the physically correct procedure, the conductance will vanish since the Kondo temperature (at which the conductance would increase for odd N) decreases with N [65].

Parallel Quantum Dots

Systems of N parallel quantum dots (see Fig. 3 for $N = 2$ and $N = 3$ cases) can be described by the multi-impurity single-channel Anderson model [81]. This model is defined by $H = H_{\text{band}} + \sum_{i=1}^{N} H_i$, where H_{band} describes the conduction band and

$$H_i = \delta n_i + \frac{U}{2}(n_i - 1)^2 + V \sum_{k\mu} (c_{k\mu}^\dagger d_{i\mu} + d_{i\mu}^\dagger c_{k\mu}), \quad (17)$$

describe the N quantum dots.

It is assumed that all impurities hybridize with the same left-right symmetric combinations of states from both leads with a constant hybridization function $\Gamma = \pi\rho V^2$ [81]. This model may equally be applicable to other system where the RKKY interaction is ferromagnetic, for example to clusters of neighboring magnetic adatoms on metallic surfaces [38,85,86].

It is assumed the inter-dot tunneling coupling t and capacitive coupling (inter-dot charge repulsion) U_{12} are negligible, so that all dots are equivalent. At low temperature, the multi-impurity Anderson model maps to a multi-impurity Kondo model. At the particle-hole symmetric point, $\delta = 0$, the conduction-band-mediated RKKY exchange interaction is ferromagnetic, $J_{\text{RKKY}} \sim U(\rho J_K)^2 = (64/\pi^2)(\Gamma^2/U)$, therefore the impurity spins order and the system effectively behaves as a single-impurity spin-$N/2$ Kondo model which undergoes spin-$N/2$ Kondo effect. This behavior is named the strong coupling (SC) regime. The Kondo temperature is approximately the same irrespective of the number of the impurities N. The residual spin at zero-temperature is $N/2 - 1/2$ if there is no coupling to additional screening channels. The ferromagnetically ordered regime and the ensuing spin-$N/2$ Kondo effect are fairly robust against various perturbations. Very strong perturbations lead, however, to quantum phase transitions of different kinds [81].

For very large δ/U, the impurities are unoccupied and the system is in the so-called frozen-impurity (FI) fixed-point with no residual spin. In the single-impurity ($N = 1$) case, the SC and FI fixed points are closely related: they belong to the same class of fixed points which differ in the strength of the potential scattering of the conduction band electrons on the impurity [87]. For multiple impurities ($N \geq 2$), however, the SC and FI lines of fixed points are qualitatively different (each corresponding to a different residual spin) and must be separated by at least one quantum phase transition [88].

At $\delta = 0$, the systems are fully conducting for any N and there is a wide plateau of high conductance associated with the spin-$N/2$ Kondo effect, Fig. 15 [81]. While the $N = 1$ system smoothly crosses over from the highly-conducting Kondo regime to the non-conducting FI regime, in the multi-impurity case we observe sharp

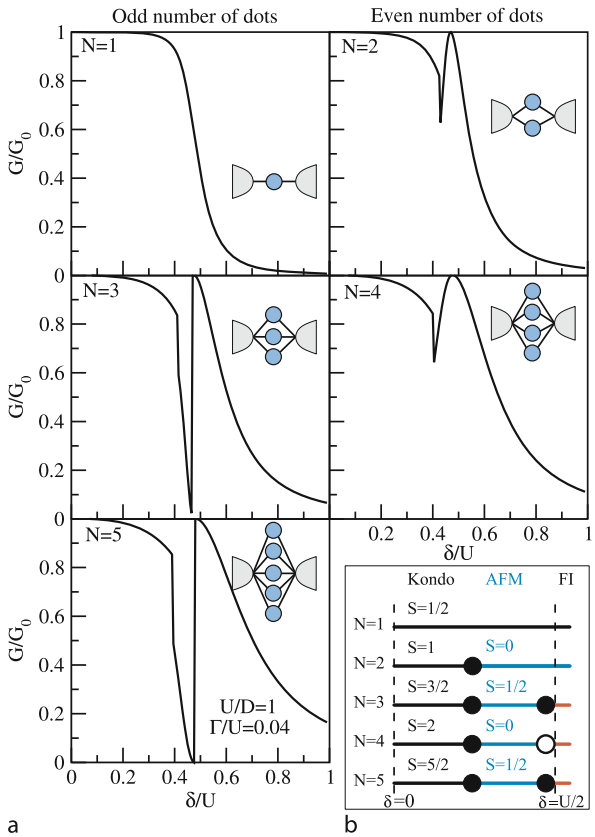

Quantum Impurity Physics in Coupled Quantum Dots, Figure 15
a Zero-temperature conductance through systems of N parallel quantum dots as a function of the gate voltage. Only $\delta > 0$ is shown due to the symmetry of the problem. b Zero-temperature phase diagram delimiting the different regimes as a function of the gate voltage. Filled circles (●) correspond to quantum phase transitions visible in the conductance curves, while the empty circle (○) denotes the phase transition with no associated conductance discontinuity

conductance discontinuities: one discontinuity for N even and two discontinuities for N odd. The conductance culminates in a unitary peak slightly below $\epsilon = 0$ (i. e. below $\delta/U = 1/2$) for all $N \geq 2$. The origin of this peak is simply potential scattering. The magnetic field B has a strong effect on the Kondo plateau: the conductance is significantly reduced as soon as B is of the order of the Kondo temperature T_K. The potential scattering peak, however, is only affected by extremely high fields of the order of U.

Conductance discontinuities find their counterparts in the jumps of the total electron occupancy and spin-spin correlation functions $\langle \boldsymbol{S}_i \cdot \boldsymbol{S}_j \rangle$, $i \neq j$; a new feature, however, is the existence of two points of discontinuity for $N = 4$ even though the conductance only exhibits one. In the Kondo regime for $\delta < \delta_{c1}$, the systems are nearly half-filled and spins are ferromagnetically ordered [81]. As we cross δ_{c1}, the occupancy slightly decreases and the spin correlations turn from ferromagnetic to antiferromagnetic. For $N \geq 3$, a second discontinuity occurs at somewhat higher δ_{c2}; its characteristic property is that the occupancy changes by almost exactly $N-2$, from $N-1$ to 1. According to the Friedel sum rule, a change in the average total impurity occupancy by n is mirrored in a change of the scattering phase shift by $\Delta\delta_{\text{q.p.}} = n\pi/2$. This explains the conductance jump from $G = G_0 \sin^2[(N-1)\pi/2] = 0$ to $G = G_0 \sin^2(\pi/2) = G_0$ in the case of odd $N \geq 3$ and the absence of the second conductance discontinuity for even $N \geq 4$, as 1 and $N-1$ are both odd integers. It is remarkable that the second quantum phase transition occurs precisely at the point where the conductance is extremal.

For $N \geq 3$, the N-impurity Anderson model thus undergoes two phase transitions. The first transition separates the ferromagnetically ordered regime and associated spin-$N/2$ Kondo screening from the antiferromagnetically ordered regime and (for odd N) Kondo screening of the spin-1/2 moment [89]. The second transition reflects the instability of the phases with the occupancy in the interval $1 < \langle n_{\text{tot}} \rangle < N-1$. Furthermore, for odd N the system abruptly switches from being fully conducting to zero conductance; this would facilitate the experimental observation of similar effects in quantum dot systems.

Future Directions

Further research in quantum transport theory will likely be centered at non-equilibrium, time-dependent and finite-temperature properties of interacting impurity systems. This is required to better understand transport at finite applied dc source-drain bias or for ac bias in real nanodevices. While the basic transport formalism is well developed, calculations of non-equilibrium properties of correlated system is still a formidable problem. The recently developed time-dependent NRG technique [90] appears very promising in this respect.

In recent years, the interest in quantum impurity physics has intensified once again due to an observation that extended lattice models of correlated electron systems may be exactly mapped in the limit of infinite lattice connectivity to effective impurity models subject to self-consistency conditions. This forms the foundation of a rapidly developing technique which has become known as the dynamical mean-field theory [91]. Results obtained in this way are believed to be a good approximation to the true behavior of such systems.

Experimental research is making progress toward creating artificial materials consisting of a large number of interconnected quantum dots. Studies of such systems would shed light on the behavior of extended bulk correlated materials. Furthermore, one could study how to tune material parameters to obtain desirable properties.

Bibliography

1. McEuen PL (1996) Artificial atoms: New boxes for electrons. Science 278:1729
2. Kouwenhoven LP, Marcus CM, McEuen PL, Tarucha S, Westervelt RM, Wingreen NS (1997) Electron transport in quantum dots. In: Sohn LL, Kouwenhoven LP, Schön G (eds) Mesoscopic electron transport. E, vol 345. NATO ASI, Kluwer, Dordrecht, pp 105–214
3. Kouwenhoven L, Marcus C (1998) Quantum dots. Phys World June 1998
4. Goldhaber-Gordon D, Shtrikman H, Mahalu D, Abusch-Magder D, Meirav U, Kastner MA (1998) Kondo effect in a single-electron transistor. Nature 391:156
5. Cronenwett SM, Oosterkamp TH, Kouwenhoven LP (1998) A tunable kondo effect in quantum dots. Science 281:540
6. Kouwenhoven L, Glazman L (2001) Revival of the kondo effect. Phys World Jan
7. Waugh FR, Berry MJ, Mar DJ, Westervelt RM, Campman KL, Gossard AC (1995) Single-electron charging in double and triple quantum dots with tunable coupling. Phys Rev Lett 75:705
8. Jeong H, Chang AM, Melloch MR (2001) The kondo effect in an artificial quantum dot molecules. Science 293:2221
9. Chen JC, Chang AM, Melloch MR (2004) Transition between quantum states in a parallel-coupled double quantum dot. Phys Rev Lett 92:176801
10. Craig NJ, Taylor JM, Lester EA, Marcus CM, Hanson MP, Gossard AC (2004) Tunable nonlocal spin control in a coupled-quantum dot system. Science 304:565
11. Gaudreau L, Studenikin SA, Sachrajda AS, Zawadzki P, Kam A, Lapointe J, Korkusinski M, Hawrylak P (2006) Stability diagram of a few-electron triple dot. Phys Rev Lett 97:036807
12. Korkusinski M, Gimenez IP, Hawrylak P, Gaudreau L, Studenikin SA, Sachrajda AS (2007) Topological hund's rules and the electronic properties of a triple lateral quantum dot molecule. Phys Rev B 75:115301 2007
13. Hewson AC (1993) The Kondo Problem to Heavy-Fermions. Cambridge University Press, Cambridge
14. Kondo J (1964) Resistance minimum in dilute magnetic alloys. Prog Theor Phys 32:37
15. Wilson KG (1975) The renormalization group: Critical phenomena and the kondo problem. Rev Mod Phys 47:773
16. Stewart GR (1984) Heavy-fermion systems. Rev Mod Phys 56:755
17. Pustilnik M, Glazman L (2004) Kondo effect in quantum dots. J Phys: Condens Matter 16:R513
18. Nygard J, Cobden DH, Lindelof PE (2000) Kondo physics in carbon nanotubes. Nature 408:342
19. Liang W, Shores MP, Bockrath M, Long JR, Park K (2002) Kondo resonance in a single-molecule transistor. Nature 417:725
20. Madhavan V, Chen W, Jamneala T, Crommie M, Wingreen NS (1998) Tunneling into a single magnetic atom: Spectroscopic evidence of the kondo resonance. Science 280:567
21. Li J, Schneider W-D, Berndt R, Delley B (1998) Kondo scattering observed at a single magnetic impurity. Phys Rev Lett 80:2893
22. Manoharan HC, Lutz CP, Eigler DM (2000) Quantum mirages formed by coherent projection of electronic structure. Nature 403:512
23. Glazman LI, Raikh ME (1988) Resonant kondo transparency of a barrier with quasilocal impurity states. JETP Lett 47:452
24. Ng TK, Lee PA (1988) On-site coulomb repulsion and resonant tunneling. Phys Rev Lett 61:1768
25. Kouwenhoven LP, Oosterkamp TH, Danoesastro MWS, Eto M, Austing DG, Honda T, Tarucha S (1997) Excitation spectra of circular, few-electron quantum dots. Science 278:1788
26. Ashoori RC (1996) Electrons in artificial atoms. Science 379:413
27. Kastner MA (1992) The single-electron transistor. Rev Mod Phys 64:849
28. Thornton TJ, Pepper M, Ahmed H, Andrews D, Davies GJ (1986) One-dimensional conduction in the 2d electron gas of a gaas-algaas heterojunction. Phys Rev Lett 56:1198
29. Park J, Pasupathy AN, Goldsmith JI, Chang C, Yaish Y, Petta JR, Rinkoski M, Sethna JP, Abruna HD, McEuen PL, Ralph DC (2002) Coulomb blockade and the kondo effect in single-atom transistors. Nature 417:722
30. Yu LH, Natelson D (2004) The kondo effect in c_{60} single-molecule transistors. Nanoletters 4:79
31. Schmid J, Weis J, Eberl K, v Klitzing K (2000) Absence of odd-even parity behavior for kondo resonance in quantum dots. Phys Rev Lett 84:5824
32. Anderson PW (1961) Localized magnetic states in metals. Phys Rev 124:41
33. Schedelbeck G, Wegscheider W, Bichler M, Abstreiter G (1997) Coupled quantum dots fabricated by cleaved edge overgrowth: From artificial atoms to molecules. Science 278:1792
34. Oosterkamp TH, Fujisawa T, van der Wiel WG, Ishibashi K, Hijman RV, Tarucha S, Kouwenhoven LP (1998) Microwave spectroscopy of a quantum-dot molecule. Nature 395:873
35. Holleitner AW, Blick RH, Hüttel AK, Eberl K, Kotthaus JP (2002) Probing and controlling the bonds of an artificial molecule. Science 297:70
36. van der Wiel WG, De Franceschi S, Elzerman JM, Fujisawa T, Tarucha S, Kouwenhoven LP (2003) Electron transport through double quantum dots. Rev Mod Phys 75:1
37. Madhavan V, Jamneala T, Nagaoka K, Chen W, Li J-L, Louie SG, Crommie MF (2002) Observation of spectral evolution during the formation of ni_2 kondo molecule. Phys Rev B 66:212411
38. Jamneala T, Madhavan V, Crommie MF (2001) Kondo response of a single antiferromagnetic chromium trimer. Phys Rev Lett 87:256804
39. Wahl P, Diekhoner L, Wittich G, Vitali L, Schneider MA, Kern K (2005) Kondo effect of molecular complexes at surfaces: ligand control of the local spin coupling. Phys Rev Lett 95:166601
40. Jones BA, Varma CM, Wilkins JW (1988) Low-temperature properties of the two-impurity kondo hamiltonian. Phys Rev Lett 61:125
41. Affleck I, Ludwig AWW, Jones BA (1995) Conformal-field-theory approach to the two-impurity kondo problem: Comparison with numerical renormalization-group results. Phys Rev B 52:9528
42. Bickers NE (1987) Review of techniques in the large-*n* expansion for dilute magnetic alloys. Rev Mod Phys 59:845
43. Andrei N, Furuya K, Lowenstein JH (1983) Solution of the kondo problem. Rev Mod Phys 55:331

44. Tsvelick AM, Wiegmann PB (1983) Exact results in the theory of magnetic alloys. Adv Phys 32:453
45. Gogolin AO, Nersesyan AA, Tsvelik AM (1999) Bosonization and strongly correlated systems. Cambridge University Press, Cambridge
46. Affleck I (1990) A current algebra approach to the kondo effect. Nucl Phys B 336:517
47. Affleck I, Ludwig AWW (1991) The kondo effect, conformal field theory and fusion rules. Nucl Phys B 352:849
48. Varma CM, Yafet Y (1976) Magnetic susceptibility of mixed-valence rare-earth compounds. Phys Rev B 13:2950
49. Gunnarsson O, Schönhammer K (1983) Electron spectroscopies for ce compounds in the impurity model. Phys Rev B 28:4315
50. Anderson PW, Yuval G (1996) Exact results in the kondo problem: Equivalence to a classical one-dimensional coulomb gas. Phys Rev Lett 23:89
51. Anderson PW, Yuval G, Hamann DR (1970) Exact results in the kondo problem ii: Scaling theory, qualitatively correct solution and some new results on one-dimensional classical statistical models. Phys Rev B 1:4464
52. Anderson PW (1970) A poor man's derivation of scaling laws for the kondo problem. J Phys C: Solid St Phys 3:2436
53. Mehta P, Andrei N (2006) Nonequilibrium transport in quantum impurity models: The bethe ansatz for open systems. Phys Rev Lett 96:216802
54. Rejec T, Ramšak A (2003) Formulas for zero-temperature conductance through a region with interaction. Phys Rev B 68:035342
55. Žitko R, Bonča J, Ramšak A, Rejec T (2006) Kondo effect in triple quantum dots. Phys Rev B 73:153307
56. Mravlje J, Ramšak A, Rejec T (2006) Kondo effect in double quantum dots with interdot repulsion. Phys Rev B 73:241305(R)
57. Bulla R, Costi T, Pruschke T (2007) The numerical renormalization group method for quantum impurity models. Rev Mod Phys 80:395; cond-mat/0701106
58. Coleman P (2002) Local moment physics in heavy electron systems. cond-mat/0206003
59. Krishna-murthy HR, Wilkins JW, Wilson KG (1980) Renormalization-group approach to the anderson model of dilute magnetic alloys. i. static properties for the symmetric case. Phys Rev B 21:1003
60. Averin DV, Nazarov YV (1990) Virtual electron diffusion during quantum tunneling of the electric charge. Phys Rev Lett 65:2446
61. Imry Y, Landauer R (1999) Conductance viewed as transmission. Rev Mod Phys 71:S306
62. Imry Y (2002) Introduction to mesoscopic physics, 2nd edn. Oxford University Press, Oxford
63. Datta S (1997) Electronic Transport in Mesoscopic Systems. Cambridge University Press, Cambridge
64. Oguri A, Nisikawa Y, Hewson AC (2005) Determination of the phase shifts for interacting electrons connected to reservoirs. J Phys Soc Jpn 74:2554
65. Nisikawa Y, Oguri A (2006) Numerical renormalization group approach to a quartet quantum-dot array connected to reservoirs: Gate-voltage dependence of the conductance. Phys Rev B 73:125108
66. Pustilnik M, Glazman LI (2001) Kondo effect in real quantum dots. Phys Rev Lett 87:216601
67. Meir Y, Wingreen NS (1992) Landauer formula for the current through an interacting electron region. Phys Rev Lett 68: 2512
68. van der Wiel WG, De Franceschi S, Fujisawa T, Elzerman JM, Tarucha S, Kouwenhoven LP (2000) The kodno effect in the unitary limit. Science 289:2105
69. Fabrizio M (2006) Clusters of anderson impurities and beyond: How kondo effect dies and what can we learn about the mott transition. Lecture notes, XI Training Course in the Physics of Strongly Correlated Systems
70. Ramšak A, Mravlje J, Žitko R, Bonča J (2006) Spin qubits in double quantum dots: Entanglement versus the kondo effect. Phys Rev B 74:241305(R)
71. Žitko R, Bonča J (2006) Enhanced conductance through side-coupled double quantum dots. Phys Rev B 73:035332
72. Zaránd G, Chung C-H, Simon P, Vojta M (2006) Quantum criticality in a double quantum-dot system. Phys Rev Lett 97:166802
73. Vojta M, Bulla R, Hofstetter W (2002) Quantum phase transitions in models of coupled magnetic impurities. Phys Rev B 65:140405(R)
74. Cornaglia PS, Grempel DR (2005) Strongly correlated regimes in a double quantum dot device. Phys Rev B 71:075305
75. Žitko R, Bonča J (2007) Fermi-liquid versus non-fermi-liquid behavior in triple quantum dots. Phys Rev Lett 98:047203
76. Jayaprakash C, Krishna-murthy HR, Wilkins JW (1981) Two-impurity kondo problem. Phys Rev Lett 47:737
77. van der Wiel WG, De Franceschi S, Elzerman JM, Tarucha S, Kouwenhoven LP, Motohisa J, Nakajima F, Fukui T (2002) Two-stage kondo effect in a quantum dot at a high magnetic field. Phys Rev Lett 88:126803
78. Hofstetter W, Schoeller H (2002) Quantum phase transition in a multilevel dot. Phys Rev Lett 88:016803
79. Hofstetter W, Zarand G (2004) Singlet-triplet transition in lateral quantum dots: A renormalization group study. Phys Rev B 69:235301
80. Hofstetter W (2000) Generalized numerical renormalization group for dynamical quantities. Phys Rev Lett 85:1508
81. Žitko R, Bonča J (2006) Multi-impurity anderson model for quantum dots coupled in parallel. Phys Rev B 74:045312
82. Galpin MR, Logan DE, Krishnamurthy HR (2005) Quantum phase transition in capacitively coupled double quantum dot. Phys Rev Lett 94:186406
83. Oguri A, Hewson AC (2005) Nrg approach to the transport through a finite hubbard chain connected to reservoirs. J Phys Soc Jpn 74:988
84. Caroli C, Combescot R, Nozieres P, Saint-James D (1971) Calculation of the tunneling current. J Phys C 4:916
85. Aligia AA (2006) Effective kondo model for a trimer on a metallic surface. Phys Rev Lett 96:096804
86. Wahl P, Simon P, Diekhöner L, Stepanyuk VS, Bruno P, Schneider MA, Kern K (2007) Exchange interaction between single magnetic adatoms. Phys Rev Lett 98:056601
87. Krishna-murthy HR, Wilkins JW, Wilson KG (1980) Renormalization-group approach to the anderson model of dilute magnetic alloys. ii. static properties for the asymmetric case. Phys Rev B 21:1044
88. Vojta M (2006) Impurity quantum phase transitions. Phil Mag 86:1807
89. Žitko R, Bonča J (2007) Quantum phase transitions in systems of parallel quantum dots. Phys Rev B 76:241305(R)

90. Anders FB, Schiller A (2005) Real-time dynamics in quantum impurity systems: A time-dependent numerical renormalization group approach. Phys Rev B 74:245113
91. Kotliar G (2005) Quantum impurity models as reference systems for strongly correlated materials: the road from the kondo impurity model to first principles electronic structure calculations with dynamical mean-field theory. J Phys Soc Jpn 74:147

Quantum Information Processing

Seth Lloyd
W.M. Keck Center for Extreme Quantum Information Processing (xQIT), MIT, Cambridge, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Quantum Mechanics
Quantum Computation
Noise and Errors
Quantum Communication
Implications and Conclusions
Bibliography

Glossary

Algorithm A systematic procedure for solving a problem, frequently implemented as a computer program.

Bit The fundamental unit of information, representing the distinction between two possible states, conventionally called 0 and 1. The word 'bit' is also used to refer to a physical system that registers a bit of information.

Boolean algebra The mathematics of manipulating bits using simple operations such as *AND*, *OR*, *NOT*, and *COPY*.

Communication channel A physical system that allows information to be transmitted from one place to another.

Computer A device for processing information. A digital computer uses Boolean algebra (q. v.) to processes information in the form of bits.

Cryptography The science and technique of encoding information in a secret form. The process of encoding is called encryption, and a system for encoding and decoding is called a cipher. A key is a piece of information used for encoding or decoding. Public-key cryptography operates using a public key by which information is encrypted, and a separate private key by which the encrypted message is decoded.

Decoherence A peculiarly quantum form of noise that has no classical analog. Decoherence destroys quantum superpositions and is the most important and ubiquitous form of noise in quantum computers and quantum communication channels.

Error-correcting code A technique for encoding information in a form that is resistant to errors. The syndrome is the part of the code that allows the error to be detected and that specifies how it should be corrected.

Entanglement A peculiarly quantum form of correlation that is responsible for many types of quantum weirdness. Entanglement arises when two or more quantum systems exist in a superposition of correlated states.

Entropy Information registered by the microscopic motion of atoms and molecules. The second law of thermodynamics (q. v.) states that entropy does not decrease over time.

Fault-tolerant computation Computation that uses error-correcting codes to perform algorithms faithfully in the presence of noise and errors. If the rate of errors falls below a certain threshold, then computations of any desired length can be performed in a fault-tolerant fashion. Also known as robust computation.

Information When used in a broad sense, information is data, messages, meaning, knowledge, etc. Used in the more specific sense of information theory, information is a quantity that can be measured in bits.

Logic gate A physical system that performs the operations of Boolean algebra (q. v.) such as *AND*, *OR*, *NOT*, and *COPY*, on bits.

Moore's law The observation, first made by Gordon Moore, that the power of computers increases by a factor of two every year and a half or so.

Quantum algorithm An algorithm designed specifically to be performed by a quantum computer using quantum logic. Quantum algorithms exploit the phenomena of superposition and entanglement to solve problems more rapidly than classical computer algorithms can. Examples of quantum algorithms include Shor's algorithm for factoring large numbers and breaking public-key cryptosystems, Grover's algorithm for searching databases, quantum simulation, the adiabatic algorithm, etc.

Quantum bit A bit registered by a quantum-mechanical system such as an atom, photon, or nuclear spin. A quantum bit, or 'qubit', has the property that it can exist in a quantum superposition of the states 0 and 1.

Qubit A quantum bit.

Quantum communication channel A communication channel that transmits quantum bits. The most common communication channel is the bosonic channel, which transmits information using light, sound, or other substances whose elementary excitations consist of bosons (photons for light, phonons for sound).

Quantum computer A computer that operates on quantum bits to perform quantum algorithms. Quantum computers have the feature that they can preserve quantum superpositions and entanglement.

Quantum cryptography A cryptographic technique that encodes information on quantum bits. Quantum cryptography uses the fact that measuring quantum systems typically disturbs them to implement cryptosystems whose security is guaranteed by the laws of physics. Quantum key distribution (QKD) is a quantum cryptographic technique for distributing secret keys.

Quantum error-correcting code An error-correcting code that corrects for the effects of noise on quantum bits. Quantum error-correcting codes can correct for the effect of decoherence (q. v.) as well as for conventional bit-flip errors.

Quantum information Information that is stored on qubits rather than on classical bits.

Quantum mechanics The branch of physics that describes how matter and energy behave at their most fundamental scales. Quantum mechanics is famously weird and counterintuitive.

Quantum weirdness A catch-all term for the strange and counterintuitive aspects of quantum mechanics. Well-known instances of quantum weirdness include Schrödinger's cat (q. v.), the Einstein–Podolsky–Rosen thought experiment, violations of Bell's inequalities, and the Greenberger–Horne–Zeilinger experiment.

Reversible logic Logical operations that do not discard information. Quantum computers operate using reversible logic.

Schrödinger's cat A famous example of quantum weirdness. A thought experiment proposed by Erwin Schrödinger, in which a cat is put in a quantum superposition of being alive and being dead. Not sanctioned by the Society for Prevention of Cruelty to Animals.

Second law of thermodynamics The second law of thermodynamics states that entropy does not increase. An alternative formulation of the second law states that it is not possible to build an eternal motion machine.

Superposition The defining feature of quantum mechanics which allows particles such as electrons to exist in two or more places at once. Quantum bits can exist in superpositions of 0 and 1 simultaneously.

Teleportation A form of quantum communication that uses pre-existing entanglement and classical communication to send quantum bits from one place to another.

Definition of the Subject

Quantum mechanics is the branch of physics that describes how systems behave at their most fundamental level. The theory of information processing studies how information can be transferred and transformed. Quantum information science, then, is the theory of communication and computation at the most fundamental physical level. Quantum computers store and process information at the level of individual atoms. Quantum communication systems transmit information on individual photons.

Over the past half century, the wires and logic gates in computers have halved in size every year and a half, a phenomenon known as Moore's law. If this exponential rate of miniaturization continues, then the components of computers should reach the atomic scale within a few decades. Even at current (2008) length scales of a little larger than one hundred nanometers, quantum mechanics plays a crucial role in governing the behavior of these wires and gates. As the sizes of computer components press down toward the atomic scale, the theory of quantum information processing becomes increasingly important for characterizing how computers operate. Similarly, as communication systems become more powerful and efficient, the quantum mechanics of information transmission becomes the key element in determining the limits of their power.

Miniaturization and the consequences of Moore's law are not the primary reason for studying quantum information, however. Quantum mechanics is weird: electrons, photons, and atoms behave in strange and counterintuitive ways. A single electron can exist in two places simultaneously. Photons and atoms can exhibit a bizarre form of correlation called entanglement, a phenomenon that Einstein characterized as *spukhafte Fernwirkung*, or 'spooky action at a distance'. Quantum weirdness extends to information processing. Quantum bits can take on the values of 0 and 1 simultaneously. Entangled photons can be used to teleport the states of matter from one place to another. The essential goal of quantum information science is to determine how quantum weirdness can be used to enhance the capabilities of computers and communication systems. For example, even a moderately sized quantum computer, containing a few tens of thousands of bits,

would be able to factor large numbers and thereby break cryptographic systems that have until now resisted the attacks of even the largest classical supercomputers [1]. Quantum computers could search databases faster than classical computers. Quantum communication systems allow information to be transmitted in a manner whose security against eavesdropping is guaranteed by the laws of physics.

Prototype quantum computers that store bits on individual atoms and quantum communication systems that transmit information using individual photons have been built and operated. These prototypes have been used to confirm the predictions of quantum information theory and to explore the behavior of information processing at the most microscopic scales. If larger, more powerful versions of quantum computers and communication systems become readily available, they will offer considerable enhancements over existing computers and communication systems. In the meanwhile, the field of quantum information processing is constructing a unified theory of how information can be registered and transformed at the fundamental limits imposed by physical law.

The remainder of this article is organized as follows:

- Section "Introduction"
  A review of the history of ideas of information, computation, and the role of information in quantum mechanics is presented.
- Section "Quantum Mechanics"
  The formalism of quantum mechanics is introduced and applied to the idea of quantum information.
- Section "Quantum Computation"
  Quantum computers are defined and their properties presented.
- Section "Noise and Errors"
  The effects of noise and errors are explored.
- Section "Quantum Communication"
  The role of quantum mechanics in setting limits to the capacity of communication channels is delineated. Quantum cryptography is explained.
- Section "Implications and Conclusions"
  Implications are discussed.

This review of quantum information theory is mathematically self-contained in the sense that all the necessary mathematics for understanding the quantum effects treated in detail here are contained in the introductory section on quantum mechanics. By necessity, not all topics in quantum information theory can be treated in detail within the confines of this article. We have chosen to treat a few key subjects in more detail: in the case of other topics we supply references to more complete treatments. The standard reference on quantum information theory is the text by Nielsen and Chuang [1], to which the reader may turn for in depth treatments of most of the topics covered here. One topic that is left largely uncovered is the broad field of quantum technologies and techniques for actually building quantum computers and quantum communication systems. Quantum technologies are rapidly changing, and no brief review like the one given here could adequately cover both the theoretical and the experimental aspects of quantum information processing.

Introduction

Information

Quantum information processing as a distinct, widely recognized field of scientific inquiry has arisen only recently, since the early 1990s. The mathematical theory of information and information processing dates to the mid-twentieth century. Ideas of quantum mechanics, information, and the relationships between them, however, date back more than a century. Indeed, the basic formulae of information theory were discovered in the second half of the nineteenth century, by James Clerk Maxwell, Ludwig Boltzmann, and J. Willard Gibbs [2]. These statistical mechanicians were searching for the proper mathematical characterization of the physical quantity known as entropy. Prior to Maxwell, Boltzmann, and Gibbs, entropy was known as a somewhat mysterious quantity that reduced the amount of work that steam engines could perform. After their work established the proper formula for entropy, it became clear that entropy was in fact a form of information — the information required to specify the actual microscopic state of the atoms in a substance such as a gas. If a system has W possible states, then it takes $\log_2 W$ bits to specify one state. Equivalently, any system with distinct states can be thought of as registering information, and a system that can exist in one out of W equally likely states can register $\log_2 W$ bits of information. The formula, $S = k \log W$, engraved on Boltzmann's tomb, means that entropy S is proportional to the number of bits of information registered by the microscopic state of a system such as a gas. (Ironically, this formula was first written down not by Boltzmann, but by Max Planck [3], who also gave the first numerical value $1.38\,10^{-23}$ J/K for the constant k. Consequently, k is called Planck's constant in early works on statistical mechanics [2]. As the fundamental constant of quantum mechanics, $h = 6.63\,10^{-34}$ joule seconds, on which more below, is also called Planck's constant, k was renamed Boltzmann's constant and is now typically written k_B.)

Although the beginning of the information processing revolution was still half a century away, Maxwell, Boltzmann, Gibbs, and their fellow statistical mechanicians were well aware of the connection between information and entropy. These researchers established that if the probability of the ith microscopic state of some system is p_i, then the entropy of the system is $S = k_B(-\sum_i p_i \ln p_i)$. The quantity $\sum_i p_i \ln p_i$ was first introduced by Boltzmann, who called it H. Boltzmann's famous H-theorem declares that H never increases [2]. The H-theorem is an expression of the second law of thermodynamics, which declares that $S = -k_B H$ never decreases. Note that this formula for S reduces to that on Boltzmann's tomb when all the states are equally likely, so that $p_i = 1/W$.

Since the probabilities for the microscopic state of a physical system depend on the knowledge possessed about the system, it is clear that entropy is related to information. The more certain one is about the state of a system – the more information one possesses about the system – the lower its entropy. As early as 1867, Maxwell introduced his famous 'demon' as a hypothetical being that could obtain information about the actual state of a system such as a gas, thereby reducing the number of states W compatible with the information obtained, and so decreasing the entropy [4]. Maxwell's demon therefore apparently contradicts the second law of thermodynamics. The full resolution of the Maxwell's demon paradox was not obtained until the end of the twentieth century, when the theory of the physics of information processing described in this review had been fully developed.

Quantum Mechanics

For the entropy, S, to be finite, a system can only possess a finite number W of possible states. In the context of classical mechanics, this feature is problematic, as even the simplest of classical systems, such as a particle moving along a line, possesses an infinite number of possible states. The continuous nature of classical mechanics frustrated attempts to use the formula for entropy to calculate many physical quantities such as the amount of energy and entropy in the radiation emitted by hot objects, the so-called 'black body radiation'. Calculations based on classical mechanics suggested the amount of energy and entropy emitted by such objects should be infinite, as the number of possible states of a classical oscillator such as a mode of the electromagnetic field was infinite. This problem is known as 'the ultraviolet catastrophe'. In 1901, Planck obtained a resolution to this problem by suggesting that such oscillators could only possess discrete energy levels [3]: the energy of an oscillator that vibrates with frequency ν can only come in multiples of $h\nu$, where h is Planck's constant defined above. Energy is *quantized*. In that same paper, as noted above, Planck first wrote down the formula $S = k \log W$, where W referred to the number of discrete energy states of a collection of oscillators. In other words, the very first paper on quantum mechanics was about information. By introducing quantum mechanics, Planck made information/entropy finite. Quantum information as a distinct field of inquiry may be young, but its origins are old: the origin of quantum information coincides with the origin of quantum mechanics.

Quantum mechanics implies that nature is, at bottom, discrete. Nature is digital. After Planck's advance, Einstein was able to explain the photo-electric effect using quantum mechanics [5]. When light hits the surface of a metal, it kicks off electrons. The energy of the electrons kicked off depends only on the frequency ν of the light, and not on its intensity. Following Planck, Einstein's interpretation of this phenomenon was that the energy in the light comes in chunks, or *quanta*, each of which possesses energy $h\nu$. These quanta, or particles of light, were subsequently termed photons. Following Planck and Einstein, Niels Bohr used quantum mechanics to derive the spectrum of the hydrogen atom [6].

In the mid nineteen-twenties, Erwin Schrödinger and Werner Heisenberg put quantum mechanics on a sound mathematical footing [7,8]. Schrödinger derived a wave equation – the Schrödinger equation – that described the behavior of particles. Heisenberg derived a formulation of quantum mechanics in terms of matrices, matrix mechanics, which was subsequently realized to be equivalent to Schrödinger's formulation. With the precise formulation of quantum mechanics in place, the implications of the theory could now be explored in detail.

It had always been clear that quantum mechanics was strange and counterintuitive: Bohr formulated the phrase 'wave-particle duality' to capture the strange way in which waves, like light, appeared to be made of particles, like photons. Similarly, particles, like electrons, appeared to be associated with waves, which were solutions to Schrödinger's equation. Now that the mathematical underpinnings of quantum mechanics were in place, however, it became clear that quantum mechanics was downright weird. In 1935, Einstein, together with his collaborators Boris Podolsky and Nathan Rosen, came up with a thought experiment (now called the EPR experiment after its originators) involving two photons that are correlated in such a way that a measurement made on one photon appears instantaneously to affect the state of the other photon [9]. Schrödinger called this form of corre-

lation 'entanglement'. Einstein, as noted above, referred to it as 'spooky action at a distance'. Although it became clear that entanglement could not be used to transmit information faster than the speed of light, the implications of the EPR thought experiment were so apparently bizarre that Einstein felt that it demonstrated that quantum mechanics was fundamentally incorrect. The EPR experiment will be discussed in detail below. Unfortunately for Einstein, when the EPR experiment was eventually performed, it confirmed the counterintuitive predictions of quantum mechanics. Indeed, every experiment ever performed so far to test the predictions of quantum mechanics has confirmed them, suggesting that, despite its counterintuitive nature, quantum mechanics is fundamentally correct.

At this point, it is worth noting a curious historical phenomenon, which persists to the present day, in which a famous scientist who received his or her Nobel prize for work in quantum mechanics, publicly expresses distrust or disbelief in quantum mechanics. Einstein is the best known example of this phenomenon, but more recent examples exist, as well. The origin of this phenomenon can be traced to the profoundly counterintuitive nature of quantum mechanics. Human infants, by the age of a few months, are aware that objects – at least, large, classical objects like toys or parents – cannot be in two places simultaneously. Yet in quantum mechanics, this intuition is violated repeatedly. Nobel laureates typically possess a powerful sense of intuition: if Einstein is not allowed to trust his intuition, then who is? Nonetheless, quantum mechanics contradicts their intuition just as it does everyone else's. Einstein's intuition told him that quantum mechanics was wrong, and he trusted that intuition. Meanwhile, scientists who are accustomed to their intuitions being proved wrong may accept quantum mechanics more readily. One of the accomplishments of quantum information processing is that it allows quantum weirdness such as that found in the EPR experiment to be expressed and investigated in precise mathematical terms, so we can discover exactly how and where our intuition goes wrong.

In the 1950's and 60's, physicists such as David Bohm, John Bell, and Yakir Aharonov, among others, investigated the counterintuitive aspects of quantum mechanics and proposed further thought experiments that threw those aspects in high relief [10,11,12]. Whenever those thought experiments have been turned into actual physical experiments, as in the well-known Aspect experiment that realized Bell's version of the EPR experiment [13], the predictions of quantum mechanics have been confirmed. Quantum mechanics is weird and we just have to live with it.

As will be seen below, quantum information processing allows us not only to express the counterintuitive aspects of quantum mechanics in precise terms, it allows us to exploit those strange phenomena to compute and to communicate in ways that our classical intuitions would tell us are impossible. Quantum weirdness is not a bug, but a feature.

Computation

Although rudimentary mechanical calculators had been constructed by Pascal and Leibnitz, amongst others, the first attempts to build a full-blown digital computer also lie in the nineteenth century. In 1822, Charles Babbage conceived the first of a series of mechanical computers, beginning with the fifteen ton Difference Engine, intended to calculate and print out polynomial functions, including logarithmic tables. Despite considerable government funding, Babbage never succeeded in building a working difference. He followed up with a series of designs for an Analytical Engine, which was to have been powered by a steam engine and programmed by punch cards. Had it been constructed, the analytical engine would have been the first modern digital computer. The mathematician Ada Lovelace is frequently credited with writing the first computer program, a set of instructions for the analytical engine to compute Bernoulli numbers.

In 1854, George Boole's *An investigation into the laws of thought* laid the conceptual basis for binary computation. Boole established that any logical relation, no matter how complicated, could be built up out of the repeated application of simple logical operations such as *AND*, *OR*, *NOT*, and *COPY*. The resulting 'Boolean logic' is the basis for the contemporary theory of computation.

While Schrödinger and Heisenberg were working out the modern theory of quantum mechanics, the modern theory of information was coming into being. In 1928, Ralph Hartley published an article, 'The Transmission of Information', in the Bell System Technical Journal [14]. In this article he defined the amount of information in a sequence of n symbols to be $n \log S$, where S is the number of symbols. As the number of such sequences is S^n, this definition clearly coincides with the Planck–Boltzmann formula for entropy, taking $W = S^n$.

At the same time as Einstein, Podolsky, and Rosen were exploring quantum weirdness, the theory of computation was coming into being. In 1936, in his paper "On Computable Numbers, with an Application to the *Entscheidungsproblem*", Alan Turing extended the earlier work of Kurt Gödel on mathematical logic, and introduced

the concept of a Turing machine, an idealized digital computer [15]. Claude Shannon, in his 1937 master's thesis, "A Symbolic Analysis of Relay and Switching Circuits", showed how digital computers could be constructed out of electronic components [16]. (Howard Gardner called this work, "possibly the most important, and also the most famous, master's thesis of the century".)

The Second World War provided impetus for the development of electronic digital computers. Konrad Zuse's Z3, built in 1941, was the first digital computer capable of performing the same computational tasks as a Turing machine. The Z3 was followed by the British Colossus, the Harvard Mark I, and the ENIAC. By the end of the 1940s, computers had begun to be built with a stored program or 'von Neumann' architecture (named after the pioneer of quantum mechanics and computer science John von Neumann), in which the set of instructions – or program – for the computer were stored in the computer's memory and executed by a central processing unit.

In 1948, Shannon published his groundbreaking article, "A Mathematical Theory of Communication", in the Bell Systems Journal [17]. In this article, perhaps the most influential work of applied mathematics of the twentieth century (following the tradition of his master's thesis), Shannon provided the full mathematical characterization of information. He introduced his colleague, John Tukey's word, 'bit', a contraction of 'binary digit', to describe the fundamental unit of information, a distinction between two possibilities, True or False, Yes or No, 0 or 1. He showed that the amount of information associated with a set of possible states i, each with probability p_i, was uniquely given by formula $-\sum_i p_i \log_2 p_i$. When Shannon asked von Neumann what he should call this quantity, von Neumann is said to have replied that he should call it H, 'because that's what Boltzmann called it'. (Recalling the Boltzmann's original definition of H, given above, we see that von Neumann had evidently forgotten the minus sign.)

It is interesting that von Neumann, who was one of the pioneers both of quantum mechanics and of information processing, apparently did not consider the idea of processing information in a uniquely quantum-mechanical fashion. Von Neumann had many things on his mind, however – game theory, bomb building, the workings of the brain, etc. – and can be forgiven for failing to make the connection. Another reason that von Neumann may not have thought of quantum computation was that, in his research into computational devices, or 'organs', as he called them, he had evidently reached the impression that computation intrinsically involved dissipation, a process that is inimical to quantum information processing [18]. This impression, if von Neumann indeed had it, is false, as will now be seen.

Reversible Computation

The date of Shannon's paper is usually taken to be the beginning of the study of information theory as a distinct field of inquiry. The second half of the twentieth century saw a huge explosion in the study of information, computation, and communication. The next step towards quantum information processing took place in the early 1960s. Until that point, there was an impression, fostered by von Neumann amongst others, that computation was intrinsically irreversible: according to this view, information was necessarily lost or discarded in the course of computation. For example, a logic gate such as an *AND* gate takes in two bits of information as input, and returns only one bit as output: the output of an *AND* gate is 1 if and only if both inputs are 1, otherwise the output is 0. Because the two input bits cannot be reconstructed from the output bits, an *AND* gate is irreversible. Since computations are typically constructed from *AND*, *OR*, and *NOT* gates (or related irreversible gates such as *NAND*, the combination of an *AND* gate and a *NOT* gate), computations were thought to be intrinsically irreversible, discarding bits as they progress.

In 1960, Rolf Landauer showed that because of the intrinsic connection between information and entropy, when information is discarded in the course of a computation, entropy must be created [19]. That is, when an irreversible logic gate such as an *AND* gate is applied, energy must be dissipated. So far, it seems that von Neumann could be correct. In 1963, however, Yves Lecerf showed that Turing Machines could be constructed in such a way that all their operations were logically reversible [20]. The trick for making computation reversible is record-keeping: one sets up logic circuits in such a way that the values of all bits are recorded and kept. To make an *AND* gate reversible, for example, one adds extra circuitry to keep track of the values of the input to the *AND* gate. In 1973, Charles Bennett, unaware of Lecerf's result, rederived it, and, most importantly, constructed physical models of reversible computation based on molecular systems such as DNA [21]. Ed Fredkin, Tommaso Toffoli, Norman Margolus, and Frank Merkle subsequently made significant contributions to the study of reversible computation [22].

Reversible computation is important for quantum information processing because the laws of physics themselves are reversible. It's this underlying reversibility that is responsible for Landauer's principle: whenever a logically irreversible process such as an *AND* gate takes place, the

information that is discarded by the computation has to go somewhere. In the case of an conventional, transistor-based *AND* gate, the lost information goes into entropy: to operate such an *AND* gate, electrical energy must be dissipated and turned into heat. That is, once the *AND* gate has been performed, then even if the logical circuits of the computer no longer record the values of the inputs to the gate, the microscopic motion of atoms and electrons in the circuit effectively 'remember' what the inputs were. If one wants to perform computation in a uniquely quantum-mechanical fashion, it is important to avoid such dissipation: to be effective, quantum computation should be reversible.

Quantum Computation

In 1980, Paul Benioff showed that quantum mechanical systems such as arrays of spins or atoms could perform reversible computation in principle [23]. Benioff mapped the operation of a reversible Turing machine onto the a quantum system and thus exhibited the first quantum-mechanical model of computation. Benioff's quantum computer was no more computationally powerful than a conventional classical Turing machine, however: it did not exploit quantum weirdness. In 1982, Richard Feynman proposed the first non-trivial application of quantum information processing [24]. Noting that quantum weirdness made it hard for conventional, classical digital computers to simulate quantum systems, Feynman proposed a 'universal quantum simulator' that could efficiently simulate other quantum systems. Feynman's device was not a quantum Turing machine, but a sort of quantum analog computer, whose dynamics could be tuned to match the dynamics of the system to be simulated.

The first model of quantum computation truly to embrace and take advantage of quantum weirdness was David Deutsch's quantum Turing machine of 1985 [25]. Deutsch pointed out that a quantum Turing machine could be designed in such a way as to use the strange and counterintuitive aspects of quantum mechanics to perform computations in ways that classical Turing machines or computers could not. In particular, just as in quantum mechanics it is acceptable (and in many circumstances, mandatory) for an electron to be in two places at once, so in a quantum computer, a quantum bit can take on the values 0 and 1 simultaneously. One possible role for a bit in a computer is as part a program, so that 0 instructs the computer to 'do this' and 1 instructs the computer to 'do that'. If a quantum bit that takes on the values 0 and 1 at the same time is fed into the quantum computer as part of a program, then the quantum computer will 'do this' and 'do that' simultaneously, an effect that Deutsch termed 'quantum parallelism'. Although it would be years before applications of quantum parallelism would be presented, Deutsch's paper marks the beginning of the formal theory of quantum computation.

For almost a decade after the work of Benioff, Feynman, and Deutsch, quantum computers remained a curiosity. Despite the development of a few simple algorithms (described in greater detail below) that took advantage of quantum parallelism, no compelling application of quantum computation had been discovered. In addition, the original models of quantum computation were highly abstract: as Feynman noted [24], no one had the slightest notion of how to build a quantum computer. Absent a 'killer ap', and a physical implementation, the field of quantum computation languished.

That languor dissipated rapidly with Peter Shor's discovery in 1994 that quantum computers could be used to factor large numbers [26]. That is, given the product r of two large prime numbers, a quantum computer could find the factors p and q such that $pq = r$. While it might not appear so instantaneously, solving this problem is indeed a 'killer ap'. Solving the factoring problem is the key to breaking 'public-key' cryptosystems. Public-key cryptosystems are a widely used method for secure communication. Suppose that you wish to buy something from me over the internet, for example. I openly send you a public key consisting of the number r. The public key is not a secret: anyone may know it. You use the public key to encrypt your credit card information, and send me that encrypted information. To decrypt that information, I need to employ the 'private keys' p and q. The security of public-key cryptography thus depends on the factoring problem being hard: to obtain the private keys p and q from the public key r, one must factor the public key.

If quantum computers could be built, then public-key cryptography was no longer secure. This fact excited considerable interest among code breakers, and some consternation within organizations, such as security agencies, whose job it is to keep secrets. Compounding this interest and consternation was the fact that the year before, in 1993, Lloyd had shown how quantum computers could be built using techniques of electromagnetic resonance together with 'off-the shelf' components such as atoms, quantum dots, and lasers [27]. In 1994, Ignacio Cirac and Peter Zoller proposed a technique for building quantum computers using ion traps [28]. These designs for quantum computers quickly resulted in small prototype quantum computers and quantum logic gates being constructed by David Wineland [29], and Jeff Kimble [30]. In 1996, Lov Grover discovered that quantum comput-

ers could search databases significantly faster than classical computers, another potentially highly useful application [31]. By 1997, simple quantum algorithms had been performed using nuclear magnetic resonance based quantum information processing [32,33,34]. The field of quantum computation was off and running.

Since 1994, the field of quantum computation has expanded dramatically. The decade between the discovery of quantum computation and the development of the first applications and implementations saw only a dozen or so papers published in the field of quantum computation. As of the date of publication of this article, it is not uncommon for a dozen papers on quantum computation to be posted on the Los Alamos preprint archive (ArXiv) every day.

Quantum Communication

While the idea of quantum computation was not introduced until 1980, and not fully exploited until the mid-1990s, quantum communication has exhibited a longer and steadier advance. By the beginning of the 1960s, J.P. Gordon [35] and Lev Levitin [36] had begun to apply quantum mechanics to the analysis of the capacity of communication channels. In 1973, Alexander Holevo derived the capacity for quantum mechanical channels to transmit classical information [37] (the Holevo–Schumacher–Westmoreland theorem [38,39]). Because of its many practical applications, the so-called 'bosonic' channel has received a great deal of attention over the years [40]. Bosonic channels are quantum communication channels in which the medium of information exchange consists of bosonic quantum particles, such as photons or phonons. That is, bosonic channels include communication channels that use electromagnetic radiation, from radio waves to light, or sound.

Despite many attempts, it was not until 1993 that Horace Yuen and Masanao Ozawa derived the capacity of the bosonic channel, and their result holds only in the absence of noise and loss [41]. The capacity of the bosonic channel in the presence of loss alone was not derived until 2004 [42], and the capacity of this most important of channels in the presence of noise and loss is still unknown [43].

A second use of quantum channels is to transmit *quantum* information, rather than classical information. The requirements for transmitting quantum information are more stringent than those for transmitting classical information. To transmit a classical bit, one must end up sending a 0 or a 1. To transmit a quantum bit, by contrast, one must also faithfully transmit states in which the quantum bit registers 0 and 1 simultaneously. The quantity which governs the capacity of a channel to transmit quantum information is called the coherent information [44,45]. A particularly intriguing method of transmitting quantum information is *teleportation* [46]. Quantum teleportation closely resembles the teleportation process from the television series Star Trek. In Star Trek, entities to be teleported enter a special booth, where they are measured and dematerialized. Information about the composition of the entities is then sent to a distant location, where the entities rematerialize.

Quantum mechanics at first seems to forbid Trekkian teleportation, for the simple reason that it is not possible to make a measurement that reveals an arbitrary unknown quantum state. Worse yet, any attempt to reveal that state is likely to destroy it. Nonetheless, if one adds just one ingredient to the protocol, quantum teleportation is indeed possible. That necessary ingredient is entanglement.

In quantum teleportation, an entity such as a quantum bit is to be teleported from Alice at point A to Bob at point B. For historical reasons, in communication protocols the sender of information is called Alice and the receiver is called Bob; an eavesdropper on the communication process is called Eve. Alice and Bob possess prior entanglement in the form of a pair of Einstein–Podolsky–Rosen particles. Alice performs a suitable measurement (described in detail below) on the qubit to be teleported together with her EPR particle. This measurement destroys the state of the particle to be teleported ('dematerializing' it), and yields two classical bits of information, which Alice sends to Bob over a conventional communication channel. Bob then performs a transformation on his EPR particle. The transformation Bob performs is a function of the information he receives from Alice: there are four possible transformations, one for each of the four possible values of the two bits he has received. After the Bob has performed his transformation of the EPR particle, the state of this particle is now guaranteed to be the same as that of the original qubit that was to be teleported.

Quantum teleportation forms an integral part of quantum communication and of quantum computation. Experimental demonstrations of quantum teleportation have been performed with photons and atoms as the systems whose quantum states are to be teleported [47,48]. At the time of the writing of this article, teleportation of larger entities such as molecules, bacteria, or human beings remains out of reach of current quantum technology.

Quantum Cryptography

A particularly useful application of the counterintuitive features of quantum mechanics is quantum cryptogra-

phy [49,50,51]. Above, it was noted that Shor's algorithm would allow quantum computers to crack public-key cryptosystems. In the context of code breaking, then, quantum information processing is a disruptive technology. Fortunately, however, if quantum computing represents a cryptographic disease, then quantum communication represents a cryptographic cure. The feature of quantum mechanics that no measurement can determine an unknown state, and that almost any measurement will disturb such a state, can be turned into a protocol for performing quantum cryptography, a method of secret communication whose security is guaranteed by the laws of physics.

In the 1970s, Stephen Wiesner developed the concept of quantum conjugate coding, in which information can be stored on two conjugate quantum variables, such as position and momentum, or linear or helical polarization [49]. In 1984, Charles Bennett and Gilles Brassard turned Wiesner's quantum coding concept into a protocol for quantum cryptography [50]: by sending suitable states of light over a quantum communication channel, Alice and Bob can build up a shared secret key. Since any attempt of Eve to listen in on their communication must inevitably disturb the states sent, Alice and Bob can determine whether Eve is listening in, and if so, how much information she has obtained. By suitable privacy amplification protocols, Alice and Bob can distill out secret key that they alone share and which the laws of physics guarantee is shared by no one else. In 1990 Artur Ekert, unaware of Wiesner, Bennett, and Brassard's work, independently derived a protocol for quantum cryptography based on entanglement [51].

Commercial quantum cryptographic systems are now available for purchase by those who desire secrecy based on the laws of physics, rather than on how hard it is to factor large numbers. Such systems represent the application of quantum information processing that is closest to every day use.

The Future

Quantum information processing is currently a thriving scientific field, with many open questions and potential applications. Key open questions include,

- Just what can quantum computers do better than classical computers? They can apparently factor large numbers, search databases, and simulate quantum systems better than classical computers. That list is quite short, however. What is the full list of problems for which quantum computers offer a speed up?
- How can we build large scale quantum computers? Lots of small scale quantum computers, with up to a dozen bits, have been built and operated. Building large scale quantum computers will require substantial technological advances in precision construction and control of complex quantum systems. While advances in this field have been steady, we're still far away from building a quantum computer that could break existing public-key cryptosystems.
- What are the ultimate physical limits to communication channels? Despite many decades of effort, fundamental questions concerning the capacity of quantum communication channels remain unresolved.

Quantum information processing is a rich stream with many tributaries in the fields of engineering, physics, and applied mathematics. Quantum information processing investigates the physical limits of computation and communication, and it devises methods for reaching closer to those limits, and someday perhaps to attain them.

Quantum Mechanics

In order to understand quantum information processing in any non-trivial way, some math is required. As Feynman said, "… it is impossible to explain honestly the beauties of the laws of nature in a way that people can feel, without their having some deep understanding of mathematics. I am sorry, but this seems to be the case" [52]. The counterintuitive character of quantum mechanics makes it even more imperative to use mathematics to understand the subject. The strange consequences of quantum mechanics arise directly out of the underlying mathematical structure of the theory. It is important to note that every bizarre and weird prediction of quantum mechanics that has been experimentally tested has turned out to be true. The mathematics of quantum mechanics is one of the most trustworthy pieces of science we possess.

Luckily, this mathematics is also quite simple. To understand quantum information processing requires only a basic knowledge of linear algebra, that is, of vectors and matrices. No calculus is required. In this section a brief review of the mathematics of quantum mechanics is presented, along with some of its more straightforward consequences. The reader who is familiar with this mathematics can safely skip to the following sections on quantum information. Readers who desire further detail are invited to consult reference [1].

Qubits

The states of a quantum system correspond to vectors. In a quantum bit, the quantum logic state 0 corresponds to a two-dimensional vector, $\binom{1}{0}$, and the quantum logic

state 1 corresponds to the vector $\binom{0}{1}$. It is customary to write these vectors in the so-called 'Dirac bracket' notation:

$$|0\rangle \equiv \begin{pmatrix}1\\0\end{pmatrix}, \quad |1\rangle \equiv \begin{pmatrix}0\\1\end{pmatrix}. \tag{1}$$

A general state for a qubit, $|\psi\rangle$, corresponds to a vector $\binom{\alpha}{\beta} = \alpha|0\rangle + \beta|1\rangle$, where α and β are complex numbers such that $|\alpha|^2 + |\beta|^2 = 1$. The requirement that the amplitude squared of the components of a vector sum to one is called 'normalization'. Normalization arises because amplitudes squared in quantum mechanics are related to probabilities. In particular, suppose that one prepares a qubit in the state $|\psi\rangle$, and then performs a measurement whose purpose is to determine whether the qubit takes on the value 0 or 1 (such measurements will be discussed in greater detail below). Such a measurement will give the outcome 0 with probability $|\alpha|^2$, and will give the outcome 1 with probability $|\beta|^2$. These probabilities must sum to one.

The vectors $|0\rangle, |1\rangle, |\psi\rangle$ are column vectors: we can also define the corresponding row vectors,

$$\langle 0| \equiv \begin{pmatrix}1 & 0\end{pmatrix}, \quad \langle 1| \equiv \begin{pmatrix}0 & 1\end{pmatrix}, \quad \langle\psi| \equiv \begin{pmatrix}\bar{\alpha} & \bar{\beta}\end{pmatrix}. \tag{2}$$

Note that creating the row vector $\langle\psi|$ involves both transposing the vector and taking the complex conjugate of its entries. This process is called Hermitian conjugation, and is denoted by the superscript †, so that $\langle\psi| = |\psi\rangle^\dagger$.

The two-dimensional, complex vector space for a qubit is denoted C^2. The reason for introducing Dirac bracket notation is that this vector space, like all the vector spaces of quantum mechanics, possesses a natural inner product, defined in the usual way by the product of row vectors and column vectors. Suppose $|\psi\rangle = \begin{pmatrix}\alpha\\\beta\end{pmatrix}$ and $|\phi\rangle = \begin{pmatrix}\gamma\\\delta\end{pmatrix}$, so that $\langle\phi| = \begin{pmatrix}\bar{\gamma} & \bar{\delta}\end{pmatrix}$ The row vector $\langle\phi|$ is called a 'bra' vector, and the column vector $|\psi\rangle$ is called a 'ket' vector. Multiplied together, these vectors form the inner product, or 'bracket',

$$\langle\phi|\psi\rangle \equiv \begin{pmatrix}\bar{\gamma} & \bar{\delta}\end{pmatrix}\begin{pmatrix}\alpha\\\beta\end{pmatrix} = \alpha\bar{\gamma} + \beta\bar{\delta}. \tag{3}$$

Note that $\langle\psi|\psi\rangle = |\alpha|^2 + |\beta^2| = 1$. The definition of the inner product (3) turns the vector space for qubits C^2 into a 'Hilbert space', a complete vector space with inner product. (Completeness means that any convergent sequence of vectors in the space attains a limit that itself lies in the space. Completeness is only an issue for infinite-dimensional Hilbert spaces and will be discussed no further here.)

We can now express probabilities in terms of brackets: $|\langle 0|\psi\rangle|^2 = |\alpha|^2 \equiv p_0$ is the probability that a measurement that distinguishes 0 and 1, made on the state $|\psi\rangle$, yields the output 0. Similarly, $|\langle 1|\psi\rangle|^2 = |\beta|^2 \equiv p_1$ is the probability that the same measurement yields the output 1. Another way to write these probabilities is to define the two 'projectors'

$$\begin{aligned} P_0 &= \begin{pmatrix}1 & 0\\0 & 0\end{pmatrix} = \begin{pmatrix}1\\0\end{pmatrix}\begin{pmatrix}1 & 0\end{pmatrix} = |0\rangle\langle 0| \\ P_1 &= \begin{pmatrix}0 & 0\\0 & 1\end{pmatrix} = \begin{pmatrix}0\\1\end{pmatrix}\begin{pmatrix}0 & 1\end{pmatrix} = |1\rangle\langle 1|. \end{aligned} \tag{4}$$

Note that

$$P_0^2 = |0\rangle\langle 0|0\rangle\langle 0| = |0\rangle\langle 0| = P_0. \tag{5}$$

Similarly, $P_1^2 = P_1$. A projection operator or projector P is defined by the condition $P^2 = P$. Written in terms of these projectors, the probabilities p_0, p_1 can be defined as

$$p_0 = \langle\psi|P_0|\psi\rangle, \quad p_1 = \langle\psi|P_1|\psi\rangle. \tag{6}$$

Note that $\langle 0|1\rangle = \langle 1|0\rangle = 0$: the two states $|0\rangle$ and $|1\rangle$ are orthogonal. Since any vector $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ can be written as a linear combination, or superposition, of $|0\rangle$ and $|1\rangle$, $\{|0\rangle, |1\rangle\}$ make up an orthonormal basis for the Hilbert space C^2. From the probabilistic interpretation of brackets, we see that orthogonality implies that a measurement that distinguishes between 0 and 1, made on the state $|0\rangle$, will yield the output 0 with probability 1 ($p_0 = 1$), and will never yield the output 1 ($p_1 = 0$). In quantum mechanics, orthogonal states are reliably distinguishable.

Higher Dimensions

The discussion above applied to qubits. More complicated quantum systems lie in higher dimensional vector spaces. For example, a 'qutrit' is a quantum system with three distinguishable states $|0\rangle, |1\rangle, |2\rangle$ that live in the three-dimensional complex vector space C^3. All the mechanisms of measurement and definitions of brackets extend to higher dimensional systems as well. For example, the distinguishability of the three states of the qutrit implies $\langle i|j\rangle = \delta_{ij}$. Many of the familiar systems of quantum mechanics, such as a free particle or a harmonic oscillator, have states that live in *infinite* dimensional Hilbert spaces. For example, the state of a free particle corresponds to a complex valued *function* $\psi(x)$ such that $\int_{-\infty}^{\infty} \bar{\psi}(x)\psi(x)\,\mathrm{d}x = 1$. The probability of finding the particle in the interval between $x = a$ and $x = b$ is then $\int_a^b \bar{\psi}(x)\psi(x)\,\mathrm{d}x$. Infinite dimensional Hilbert spaces

involve subtleties that, fortunately, rarely impinge upon quantum information processing except in the use of bosonic systems as in quantum optics [40].

Matrices

Quantum mechanics is an intrinsically *linear* theory: transformations of states are represented by matrix multiplication. (Nonlinear theories of quantum mechanics can be constructed, but there is no experimental evidence for any intrinsic nonlinearity in quantum mechanics.) Consider the set of matrices U such that $U^\dagger U = Id$, where Id is the identity matrix. Such a matrix is said to be 'unitary'. (For matrices on infinite-dimensional Hilbert spaces, i. e., for linear operators, unitarity also requires $UU^\dagger = Id$.) If we take a normalized vector $|\psi\rangle$, $\langle\psi|\psi\rangle = 1$, and transform it by multiplying it by U, so that $|\psi'\rangle = U|\psi\rangle$, then we have

$$\langle\psi'|\psi\rangle = \langle\psi|U^\dagger U|\psi\rangle = \langle\psi|\psi\rangle = 1 \,. \tag{7}$$

That is, unitary transformations U preserve the normalization of vectors. Equation (7) can also be used to show that any U that preserves the normalization of all vectors $|\psi\rangle$ is unitary. Since to be given a physical interpretation in terms of probabilities, the vectors of quantum mechanics must be normalized, the set of unitary transformations represents the set of 'legal' transformations of vectors in Hilbert space. (Below, we'll see that when one adds an environment with which qubits can interact, then the set of legal transformations can be extended.) Unitary transformations on a single qubit make up the set of two-by-two unitary matrices $U(2)$.

Spin and Other Observables

A familiar quantum system whose state space is represented by a qubit is the spin 1/2 particle, such as an electron or proton. The spin of such a particle along a given axis can take on only two discrete values, 'spin up', with angular momentum $\hbar/2$ about that axis, or 'spin down', with angular momentum $-\hbar/2$. Here, $\hbar$ is Planck's reduced constant: $\hbar \equiv h/2\pi = 1.05457\,10^{-34}$ joule-sec. It is conventional to identify the state $|\uparrow\rangle$, spin up along the z-axis, with $|0\rangle$, and the state $|\downarrow\rangle$, spin up along the z-axis, with $|1\rangle$. In this way, the spin of an electron or proton can be taken to register a qubit.

Now that we have introduced the notion of spin, we can introduce an operator or matrix that corresponds to the measurement of spin. Let $P_\uparrow = |\uparrow\rangle\langle\uparrow|$ be the projector onto the state $|\uparrow\rangle$, and let $P_\downarrow = |\downarrow\rangle\langle\downarrow|$ be the projector onto the state $|\downarrow\rangle$. The matrix, or 'operator' corresponding to spin 1/2 along the z-axis is then

$$I_z = \frac{\hbar}{2}(P_\uparrow - P_\downarrow) = \frac{\hbar}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \frac{\hbar}{2}\sigma_z \,, \tag{8}$$

where $\sigma_z \equiv \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ is called the z Pauli matrix. In what way does I_z correspond to spin along the z-axis? Suppose that one starts out in the state $|\psi\rangle = \alpha|\uparrow\rangle + \beta|\downarrow\rangle$ and then measures spin along the z-axis. Just as in the case of measuring 0 or 1, with probability $p_\uparrow = |\alpha|^2$ one obtains the result $\uparrow$, and with probability $p_\downarrow = |\beta|^2$ one obtains the result $\downarrow$. The *expectation value* for the angular momentum along the z-axis is then

$$\langle I_z\rangle = p_\uparrow(\hbar/2) + p_\downarrow(-\hbar/2) = \langle\psi|I_z|\psi\rangle \,. \tag{9}$$

That is, the expectation value of the observable quantity corresponding to spin along the z-axis is given by taking the bracket of the state $|\psi\rangle$ with the operator I_z corresponding to that observable.

In quantum mechanics, every observable quantity corresponds to an operator. The operator corresponding to an observable with possible outcome values $\{a\}$ is $A = \sum_a a|a\rangle\langle a| = \sum_a aP_a$, where $|a\rangle$ is the state with value a and $P_a = |a\rangle\langle a|$ is the projection operator corresponding to the outcome a. Note that since the outcomes of measurements are real numbers, $A^\dagger = A$: the operators corresponding to observables are Hermitian. The states $\{|a\rangle\}$ are, by definition, distinguishable and so make up an orthonormal set. From the definition of A one sees that $A|a\rangle = a|a\rangle$. That is, the different possible outcomes of the measurement are eigenvalues of A, and the different possible outcome states of the measurement are eigenvectors of A.

If more than one state $|a\rangle_i$ corresponds to the outcome a, then $A = \sum_a aP_a$, where $P_a = \sum_i |a\rangle_i\langle a|$ is the projection operator onto the eigenspace corresponding to the 'degenerate' eigenvalue a. Taking, for the moment, the case of non-degenerate eigenvalues, then the expectation value of an observable A in a particular state $|\chi\rangle = \sum_a \chi_a|a\rangle$ is obtained by bracketing the state about the corresponding operator:

$$\langle A\rangle \equiv \langle\chi|A|\chi\rangle = \sum_a |\chi_a|^2 a = \sum_a p_a a \,, \tag{10}$$

where $p_a = |\chi_a|^2$ is the probability that the measurement yields the outcome a.

Above, we saw that the operator corresponding to spin along the z-axis was $I_z = (\hbar/2)\sigma_z$. What then are the operators corresponding to spin along the x- and y-axes? They

are given by $I_x = (\hbar/2)\sigma_x$ and $I_y = (\hbar/2)\sigma_y$, where σ_x and σ_y are the two remaining Pauli spin matrices out of the trio:

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} . \tag{11}$$

By the prescription for obtaining expectation values (10), for an initial state $|\chi\rangle$ the expectation values of spin along the x-axis and spin along the y-axis are

$$\langle I_x \rangle = \langle \chi | I_x | \chi \rangle , \quad \langle I_y \rangle = \langle \chi | I_y | \chi \rangle . \tag{12}$$

The eigenvectors of I_x, σ_x and I_y, σ_y are also easily described. The eigenvector of I_x, σ_x corresponding to spin up along the x-axis is $|\rightarrow\rangle = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$, while the eigenvector of I_x, σ_x corresponding to spin down along the x-axis is $|\leftarrow\rangle = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$. Note that these eigenvectors are orthogonal and normalized – they make up an orthonormal set. It's easy to verify that, $\sigma_x|\uparrow\rangle = +1|\uparrow\rangle$, and $\sigma_x|\downarrow\rangle = -1|\downarrow\rangle$, so the eigenvalues of σ_x are ± 1. The eigenvalues of $I_x = (\hbar/2)\sigma_x$ are $\pm\hbar/2$, the two different possible values of angular momentum corresponding to spin up or spin down along the x-axis. Similarly, the eigenvector of I_y, σ_y corresponding to spin up along the y-axis is $|\otimes\rangle = \begin{pmatrix} 1/\sqrt{2} \\ i/\sqrt{2} \end{pmatrix}$, while the eigenvector of I_y, σ_y corresponding to spin down along the y-axis is $|\odot\rangle = \begin{pmatrix} i/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$. (Here, in deference to the right-handed coordinate system that we are implicitly adopting, $\otimes$ corresponds to an arrow heading away from the viewer, and $\odot$ corresponds to an arrow heading towards the viewer.)

Rotations and SU(2)

The Pauli matrices $\sigma_x, \sigma_y, \sigma_z$ play a crucial role not only in characterizing the measurement of spin, but in generating rotations as well. Because of their central role in describing qubits in general, and spin in particular, several more of their properties are elaborated here. Clearly, $\sigma_i = \sigma_i^\dagger$: Pauli matrices are Hermitian. Next, note that

$$\sigma_x^2 = \sigma_y^2 = \sigma_z^2 = Id = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} . \tag{13}$$

Since $\sigma_i = \sigma_i^\dagger$, and $\sigma_i^2 = Id$, it's also the case that $\sigma_i^\dagger \sigma_i = Id$: that is, the Pauli matrices are unitary. Next, defining the commutator of two matrices A and B to be $[A, B] = AB - BA$, it is easy to verify that $[\sigma_x, \sigma_y] = 2i\sigma_z$. Cyclic permutations of this identity also hold, e. g., $[\sigma_z, \sigma_x] = 2i\sigma_y$.

Now introduce the concept of a rotation. The operator $e^{-i(\theta/2)\sigma_x}$ corresponds to a rotation by an angle θ about the x-axis. The analogous operators with x replaced by y or z are expressions for rotations about the y- or z-axes. Exponentiating matrices may look at first strange, but exponentiating Pauli matrices is significantly simpler. Using the Taylor expansion for the matrix exponential, $e^A = Id + A + A^2/2! + A^3/3! + \ldots$, and employing the fact that $\sigma_j^2 = Id$, one obtains

$$e^{-i(\theta/2)\sigma_j} = \cos(\theta/2)Id - i\sin(\theta/2)\sigma_j . \tag{14}$$

It is useful to verify that the expression for rotations (14) makes sense for the states we have defined. For example, rotation by π about the x-axis should take the state $|\uparrow\rangle$, spin z up, to the state $|\downarrow\rangle$, spin z down. Inserting $\theta = \pi$ and $j = x$ in (14), we find that the operator corresponding to this rotation is the matrix $-i\sigma_x$. Multiplying $|\uparrow\rangle$ by this matrix, we obtain

$$-i\sigma_x|\uparrow\rangle = -i\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = -i\begin{pmatrix} 0 \\ 1 \end{pmatrix} = -i|\downarrow\rangle . \tag{15}$$

The rotation does indeed take $|\uparrow\rangle$ to $|\downarrow\rangle$, but it also introduces an overall phase of $-i$.

What does this overall phase do? The answer is Nothing! Or, at least, nothing observable. Overall phases cannot change the expectation value of any observable. Suppose that we compare expectation values for the state $|\chi\rangle$ and for the state $|\chi'\rangle = e^{i\phi}|\chi\rangle$ for some observable corresponding to an operator A. We have

$$\langle \chi | A | \chi \rangle = \langle \chi | e^{-i\phi} A e^{i\phi} | \chi \rangle = \langle \chi' | A | \chi' \rangle . \tag{16}$$

Overall phases are undetectable. Keeping the undetectability of overall phases in mind, it is a useful exercise to verify that other rotations perform as expected. For example, a rotation by $\pi/2$ about the x-axis takes $|\otimes\rangle$, spin up along the y-axis, to $|\uparrow\rangle$, together with an overall phase.

Once rotation about the x, $>y$, and z axes have been defined, it is straightforward to construct rotations about any axis. Let $\hat{\iota} = (\iota_x, \iota_y, \iota_z)$, $\iota_x^2 + \iota_y^2 + \iota_z^2 = 1$, be a unit vector along the $\hat{\iota}$ direction in ordinary three-dimensional space. Define $\sigma_{\hat{\iota}} = \iota_x\sigma_x + \iota_y\sigma_y + \iota_z\sigma_z$ to be the *generalized Pauli matrix* associated with the unit vector $\hat{\iota}$. It is easy to verify that $\sigma_{\hat{\iota}}$ behaves like a Pauli matrix, e. g., $\sigma_{\hat{\iota}}^2 = Id$. Rotation by θ about the $\hat{\iota}$ axis then corresponds to an operator $e^{-i(\theta/2)\sigma_{\hat{\iota}}} = \cos(\theta/2)Id - i\sin(\theta/2)\sigma_{\hat{\iota}}$. Once again, it is a useful exercises to verify that such rotations behave as expected. For example, a rotation by π about the

$(1/\sqrt{2}, 0, 1/\sqrt{2})$ axis should 'swap' $|\uparrow\rangle$ and $|\rightarrow\rangle$, up to some phase.

The set of rotations of the form $e^{-i\theta/2\sigma_i}$ forms the *group* $SU(2)$, the set of complex 2 by 2 unitary matrices with determinant equal to 1. It is instructive to compare this group with the 'conventional' group of rotations in three dimensions, $SO(3)$. $SO(3)$ is the set of real 3 by 3 matrices with orthonormal rows/columns and determinant 1. In $SO(3)$, when one rotates a vector by 2π, the vector returns to its original state: a rotation by 2π corresponds to the 3 by 3 identity matrix. In $SU(2)$, rotating a vector by 2π corresponds to the transformation $-Id$: in rotating by 2π, the vector acquires an overall phase of -1. As will be seen below, the phase of -1, while unobservable for single qubit rotations, can be, and has been observed in two-qubit operations. To return to the original state, with no phase, one must rotate by 4π. A macroscopic, classical version of this fact manifests itself when one grasps a glass of water firmly in the palm of one's hand and rotates one's arm and shoulder to rotate the glass without spilling it. A little experimentation with this problem shows that one must rotate glass and hand around twice to return them to their initial orientation.

Why Quantum Mechanics?

Why is the fundamental theory of nature, quantum mechanics, a theory of complex vector spaces? No one knows for sure. One of the most convincing explanations came from Aage Bohr, the son of Niels Bohr and a Nobel laureate in quantum mechanics in his own right [53]. Aage Bohr pointed out that the basic mathematical representation of *symmetry* consists of complex vector spaces. For example, while the apparent symmetry group of rotations in three dimensional space is the real group $SO(3)$, the actual underlying symmetry group of space, as evidenced by rotations of quantum-mechanical spins, is $SU(2)$: to return to the same state, one has to go around not once, but twice. It is a general feature of complex, continuous groups, called 'Lie groups' after Sophus Lie, that their fundamental representations are complex. If quantum mechanics is a manifestation of deep, underlying symmetries of nature, then it should come as no surprise that quantum mechanics is a theory of transformations on complex vector spaces.

Density Matrices

The review of quantum mechanics is almost done. Before moving on to quantum information processing proper, two topics need to be covered. The first topic is how to deal with uncertainty about the underlying state of a quantum system. The second topic is how to treat two or more quantum systems together. These topics turn out to possess a strong connection which is the source of most counterintuitive quantum effects.

Suppose that don't know exactly what state a quantum system is in. Say, for example, it could be in the state $|0\rangle$ with probability p_0 or in the state $|1\rangle$ with probability p_1. Note that this state is not the same as a quantum superposition, $\sqrt{p_0}|0\rangle + \sqrt{p_1}|1\rangle$, which is a definite state with spin oriented in the $x - z$ plane. The expectation value of an operator A when the underlying state possesses the uncertainty described is

$$\langle A\rangle = p_0\langle 0|A|0\rangle + p_1\langle 1|A|1\rangle = \mathrm{tr}\rho A\,, \tag{17}$$

where $\rho = p_0|0\rangle\langle 0| + p_1|1\rangle\langle 1|$ is the *density matrix* corresponding to the uncertain state. The density matrix can be thought of as the quantum mechanical analogue of a probability distribution.

Density matrices were developed to provide a quantum mechanical treatment of statistical mechanics. A famous density matrix is that for the canonical ensemble. Here, the energy state of a system is uncertain, and each energy state $|E_i\rangle$ is weighted by a probability $p_i = e^{-E_i/k_B T}/Z$, where $Z = \sum_i e^{-E_i/k_B T}$ is the partition function. Z is needed to normalize the probabilities $\{p_i\}$ so that $\sum_i p_i = 1$. The density matrix for the canonical ensemble is then $\rho_C = (1/Z)\sum_i e^{-E_i/k_B T}|E_i\rangle\langle E_i|$. The expectation value of any operator, e. g., the energy operator H (for 'Hamiltonian') is then given by $\langle H\rangle = \mathrm{tr}\rho_C H$.

Multiple Systems and Tensor Products

To describe two or more systems requires a formalism called the tensor product. The Hilbert space for two qubits is the space $C^2 \otimes C^2$, where $\otimes$ is the tensor product. $C^2 \otimes C^2$ is a four-dimensional space spanned by the vectors $|0\rangle \otimes |0\rangle, |0\rangle \otimes |1\rangle, |1\rangle \otimes |0\rangle, |1\rangle \otimes |1\rangle$. (To save space these vectors are sometimes written $|0\rangle|0\rangle, |0\rangle|1\rangle, |1\rangle|0\rangle, |1\rangle|1\rangle$, or even more compactly, $|00\rangle, |01\rangle, |10\rangle, |11\rangle$. Care must be taken, however, to make sure that this notation is unambiguous in a particular situation.) The tensor product is *multilinear*: in performing the tensor product, the distributive law holds. That is, if $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, and $|\phi\rangle = \gamma|0\rangle + \delta|1\rangle$, then

$$\begin{aligned}|\psi\rangle \otimes |\phi\rangle &= (\alpha|0\rangle + \beta|1\rangle) \otimes (\gamma|0\rangle + \delta|1\rangle)\\ &= \alpha\gamma|0\rangle \otimes |0\rangle + \alpha\delta|0\rangle \otimes |1\rangle\\ &\quad + \beta\gamma|1\rangle \otimes |0\rangle + \beta\delta|1\rangle \otimes |1\rangle\,.\end{aligned} \tag{18}$$

A tensor is a thing with slots: the key point to keep track of in tensor analysis is which operator or vector acts on which slot. It is often useful to label the slots, e. g., $|\psi\rangle_1 \otimes |\phi\rangle_2$ is

a tensor product vector in which $|\psi\rangle$ occupies slot 1 and $|\phi\rangle$ occupies slot 2.

One can also define the tensor product of operators or matrices. For example, $\sigma_x^1 \otimes \sigma_z^2$ is a tensor product operator with σ_x in slot 1 and σ_z in slot 2. When this operator acts on a tensor product vector such as $|\psi\rangle_1 \otimes |\phi\rangle_2$, the operator in slot 1 acts on the vector in that slot, and the operator in slot 2 acts on the vector in that slot:

$$(\sigma_x^1 \otimes \sigma_z^2)(|\psi\rangle_1 \otimes |\phi\rangle_2) = (\sigma_x^1|\psi\rangle_1) \otimes (\sigma_z^2|\phi\rangle_2) \,. \quad (19)$$

The No-Cloning Theorem

Now that tensor products have been introduced, one of the most famous theorems of quantum information – the no-cloning theorem – can immediately be proved [54]. Classical information has the property that it can be copied, so that $0 \to 00$ and $1 \to 11$. How about quantum information? Does there exist a procedure that allows one to take an arbitrary, unknown state $|\psi\rangle$ to $|\psi\rangle \otimes |\psi\rangle$? Can you clone a quantum? As the title to this section indicates, the answer to this question is No.

Suppose that you could clone a quantum. Then there would exist a unitary operator U_C that would take the state

$$|\psi\rangle \otimes |0\rangle \to U_C|\psi\rangle \otimes |0\rangle = |\psi\rangle \otimes |\psi\rangle \,, \quad (20)$$

for any initial state $|\psi\rangle$. Consider another state $|\phi\rangle$. Since U_C is supposed to clone any state, we have then we would also have $U_C|\phi\rangle \otimes |0\rangle = |\phi\rangle \otimes |\phi\rangle$. If U_C exists, then, the following holds for any states $|\psi\rangle$, $|\phi\rangle$:

$$\begin{aligned}
\langle\phi|\psi\rangle &= ({}_1\langle\phi| \otimes {}_2\langle 0|)(|\psi\rangle_1 \otimes |0\rangle_2) \\
&= ({}_1\langle\phi| \otimes {}_2\langle 0|)(U_C^\dagger U_C)(|\psi\rangle_1 \otimes |0\rangle_2) \\
&= ({}_1\langle\phi| \otimes {}_2\langle 0|U_C^\dagger)(U_C|\psi\rangle_1 \otimes |0\rangle_2) \\
&= ({}_1\langle\phi| \otimes {}_2\langle\phi|)(|\psi\rangle_1 \otimes |\psi\rangle_2) \\
&\quad \cdot ({}_1\langle\phi|\psi\rangle_1)({}_2\langle\phi|\psi\rangle_2) \\
&= \langle\phi|\psi\rangle^2 \,,
\end{aligned} \quad (21)$$

where we have used the fact that U_C is unitary so that $U_C^\dagger U_C = Id$. So if cloning is possible, then $\langle\phi|\psi\rangle = \langle\phi|\psi\rangle^2$ for *any* two vectors $|\psi\rangle$ and $|\phi\rangle$. But this is impossible, as it implies that $\langle\phi|\psi\rangle$ equals either 0 or 1 for all $|\psi\rangle$, $|\phi\rangle$, which is certainly not true. You can't clone a quantum.

The no-cloning theorem has widespread consequences. It is responsible for the efficacy of quantum cryptography, which will be discussed in greater detail below. Suppose that Alice sends a state $|\psi\rangle$ to Bob. Eve wants to discover what state this is, without Alice or Bob uncovering her eavesdropping. That is, she would like to make a copy of $|\psi\rangle$ and send the original state $|\psi\rangle$ to Bob. The no-cloning theorem prevents her from doing so: any attempt to copy $|\psi\rangle$ will necessarily perturb the state. An 'optimal cloner' is a transformation that does the best possible job of cloning, given that cloning is impossible [55].

Reduced Density Matrices

Suppose that one makes a measurement corresponding to an observable A_1 on the state in slot 1. What operator do we take the bracket of to get the expectation value? The answer is $A_1 \otimes Id_2$: we have to put the identity in slot 2. The expectation value for this measurement for the state $|\psi\rangle_1 \otimes |\phi\rangle_2$ is then

$$\begin{aligned}
&{}_1\langle\psi| \otimes {}_2\langle\phi|A_1 \otimes Id_2|\psi\rangle_1 \otimes |\phi\rangle_2 = \\
&\quad {}_1\langle\psi|A_1|\psi\rangle_1 \otimes {}_2\langle\phi|Id_2|\phi\rangle_2 = {}_1\langle\psi|A_1|\psi\rangle_1 \,.
\end{aligned} \quad (22)$$

Here we have used the rule that operators in slot 1 act on vectors in slot 1. Similarly, the operators in slot 2 act on vectors in slot 2. As always, the key to performing tensor manipulations is to keep track of what is in which slot. (Note that the tensor product of two numbers is simply the product of those numbers.)

In ordinary probability theory, the probabilities for two sets of events labeled by i and j is given by a joint probability distribution $p(ij)$. The probabilities for the first set of events on their own is obtained by averaging over the second set: $p(i) = \sum_j p(ij)$ is the marginal distribution for the first set of events labeled by i. In quantum mechanics, the analog of a probability distribution is density matrix. Two systems 1 and 2 are described by a joint density matrix ρ_{12}, and system 1 on its own is described by a 'reduced' density matrix ρ_1.

Suppose that systems 1 and 2 are in a state described by a density matrix

$$\rho_{12} = \sum_{ii'jj'} \rho_{ii'jj'}|i\rangle_1\langle i'| \otimes |j\rangle_2\langle j'| \,, \quad (23)$$

where $\{|i\rangle_1\}$ and $\{|j\rangle_2\}$ are orthonormal bases for systems 1 and 2 respectively. As in the previous paragraph, the expectation value of a measurement made on ρ_{12} alone is given by $\mathrm{tr}\rho_{12}(A_1 \otimes Id_2)$. Another way to write such expectation values is to define the *reduced density matrix*,

$$\begin{aligned}
\rho_1 = \mathrm{tr}_2\rho_{12} &\equiv \sum_{ii'jj'} \rho_{ii'jj'}|i\rangle_1\langle i'| \otimes_2 \langle j'|j\rangle_2 \\
&= \sum_{ii'j} \rho_{ii'jj}|i\rangle_1\langle i'| \,.
\end{aligned} \quad (24)$$

Equation (24) describes the partial trace tr_2 over system 2. In other words, if ρ_{12} has components, $\{\rho_{ii'jj'}\}$,

then reduced density matrix $\rho_1 = \mathrm{tr}_2\rho_{12}$ has components $\{\sum_j \rho_{ii'jj}\}$. The expectation value of a measurement A made on the first system alone is then simply $\langle A\rangle = \mathrm{tr}\rho_1 A$. Just as in ordinary probability theory, where the marginal distribution for system 1 is obtained by averaging over the state of system 2, so in quantum mechanics the reduced density matrix that describes system 1 is obtained by tracing over the state of system 2.

Entanglement

One of the central features of quantum information processing is entanglement. Entanglement is a peculiarly quantum-mechanical form of correlation between quantum systems, that has no classical analogue. Entanglement lies at the heart of the various speedups and enhancements that quantum information processing offers over classical information processing.

A pure state $|\psi\rangle_{12}$ for two systems 1 and 2 is entangled if the reduced density matrix for either system taken on its own has non-zero entropy. In particular, the reduced density matrix for system 1 is $\rho_1 = \mathrm{tr}_2\rho_{12}$, where $\rho_{12} = |\psi\rangle_{12}\langle\psi|$. The entropy of this density matrix is $S(\rho_1) = -\mathrm{tr}\rho_1 \log_2 \rho_1$. For pure states, the entropy of ρ_1 is equal to the entropy of ρ_2 and is a good measure of the degree of entanglement between the two systems. $S(\rho_1) = S(\rho_2)$ measures the number of 'e-bits' of entanglement between systems 1 and 2.

A mixed state ρ_{12} for 1 and 2 is entangled if it is not separable. A density matrix is separable if it can be written $\rho_{12} = \sum_j p_j \rho_1^j \otimes \rho_2^j$. In other words, a separable state is one that can be written as a classical mixture of uncorrelated states. The correlations in a separable state are purely classical.

Entanglement can take a variety of forms and manifestations. The key to understanding those forms is the notion of Local Operations and Classical Communication (LOCC) [56]. Local operations such as unitary transformations and measurement, combined with classical communication, can not, on their own, create entanglement. If one state can be transformed into another via local operations and classical communication, then the first state is 'at least as entangled' as the second. LOCC can then be used to categorize the different forms of entanglement.

Distillable entanglement is a form of entanglement that can be transformed into pure-state entanglement [57]. Systems 1 and 2 posses d qubits worth of distillable entanglement if local operations and classical communication can transform their state into a pure state that contains d e-bits (possibly with some leftover 'junk' in a separate quantum register). Systems that are non-separable, but that possess no distillable entanglement are said to possess bound entanglement [58].

The entanglement of formation for a state ρ_{12} is equal to the minimum number of e-bits of pure-state entanglement that are required to create ρ_{12} using only local operations and classical control [59]. The entanglement of formation of ρ_{12} is greater than or equal to ρ_{12}'s distillable entanglement. A variety of entanglement measures exist. Each one is useful for different purposes. Squashed entanglement, for example, plays an important role in quantum cryptography [60]. (Squashed entanglement is a notion of entanglement based on conditional information.)

One of the most interesting open questions in quantum information theory is the definition of entanglement for marti-partite systems consisting of more than two subsystems. Here, even in the case of pure states, no unique definition of entanglement exists.

Entanglement plays a key role in quantum computation and quantum communication. Before turning to those fields, however, it is worth while investigating the strange and counterintuitive features of entanglement.

Quantum Weirdness

Entanglement is the primary source of what for lack of a better term may be called 'quantum weirdness'. Consider the two-qubit state

$$|\psi\rangle_{12} = \frac{1}{\sqrt{2}}(|0\rangle_1 \otimes |1\rangle_2 - |1\rangle_1|0\rangle_2)\,. \tag{25}$$

This state is called the 'singlet' state: if the two qubits correspond to two spin 1/2 particles, as described above, so that $|0\rangle$ is the spin z up state and $|1\rangle$ is the spin z down state, then the singlet state is the state with zero angular momentum. Indeed, rewriting $|\psi\rangle_{12}$ in terms of spin as

$$|\psi\rangle_{12} = \frac{1}{\sqrt{2}}(|\uparrow\rangle_1 \otimes |\downarrow\rangle_2 - |\downarrow\rangle_1|\uparrow\rangle_2)\,. \tag{26}$$

one sees that if one makes a measurement of spin z, then if the first spin has spin z up, then the second spin has spin z down, and vice versa.

If one decomposes the state in terms of spin along the x-axis, $|\rightarrow\rangle = (1/\sqrt{2})(|\uparrow\rangle + |\downarrow\rangle)$, $|\leftarrow\rangle = (1/\sqrt{2})\cdot(|\uparrow\rangle - |\downarrow\rangle)$, then $|\psi\rangle_{12}$ can be rewritten

$$|\psi\rangle_{12} = \frac{1}{\sqrt{2}}(|\rightarrow\rangle_1 \otimes |\leftarrow\rangle_2 - |\leftarrow\rangle_1 \otimes |\rightarrow\rangle_2)\,. \tag{27}$$

Similarly, rewriting in terms of spin along the y-axis, we obtain

$$|\psi\rangle_{12} = \frac{1}{\sqrt{2}}(|\otimes\rangle_1|\odot\rangle_2 - |\odot\rangle_1|\otimes\rangle_2)\,, \tag{28}$$

where $|\otimes\rangle$ is the state with spin up along the y-axis and $|\odot\rangle$ is the state with spin down along the y-axis. No matter what axis one decomposes the spin about, if the first spin has spin up along that axis then the second spin has spin down along that axis, and vice versa. The singlet state has angular momentum zero about every axis.

So far, this doesn't sound too strange. The singlet simply behaves the way a state with zero angular momentum should: it is not hard to see that it is the unique two-spin state with zero angular momentum about every axis. In fact, the singlet state exhibits lots of quantum weirdness. Look at the reduced density matrix for spin 1:

$$\rho_1 = \mathrm{tr}_2\rho_{12} = \mathrm{tr}_2|\psi\rangle_{12}\langle\psi| = \frac{1}{2}(|\uparrow\rangle_1\langle\uparrow| + |\downarrow\rangle_1\langle\downarrow| = Id/2\,. \tag{29}$$

That is, the density matrix for spin 1 is in a completely indefinite, or 'mixed' state: nothing is known about whether it is spin up or spin down along any axis. Similarly, spin 2 is in a completely mixed state. This is already a little strange. The two spins together are in a definite, 'pure' state, the singlet state. Classically, if two systems are in a definite state, then each of the systems on its own is in a definite state: the only way to have uncertainty about one of the parts is to have uncertainty about the whole. In quantum mechanics this is not the case: two systems can be in a definite, pure state taken together, while each of the systems on its own is in an indefinite, mixed state. Such systems are said to be *entangled* with each other.

Entanglement is a peculiarly quantum form of correlation. Two spins in a singlet state are highly correlated (or, more precisely, anticorrelated): no matter what axis one measures spin along, one spin will be found to have the opposite spin of the other. In itself, that doesn't sound so bad, but when one makes a measurement on one spin, something funny seems to happen. Both spins start out in a completely indefinite state. Now one chooses to make a measurement of spin 1 along the z-axis. Suppose that one gets the result, spin up. As a result of the measurement, spin 2 is now in a definite state, spin down along the z-axis. If one had chosen to make a measurement of spin 1 along the x-axis, then spin 2 would also be put in a definite state along the x-axis. Somehow, it seems as if one can affect the state of spin 2 by making a measurement of spin 1 on its own. This is what Einstein called 'spooky action at a distance'.

In fact, such measurements involve no real action at a distance, spooky or otherwise. If one could really act on spin 2 by making a measurement on spin 1, thereby changing spin 2's state, then one could send information instantaneously from spin 1 to spin 2 by measuring spin 1 alone. Such instantaneous transmission of information would violate special relativity and give rise to all sorts of paradoxical capabilities, such as the ability to travel backwards in time. Luckily, it is easy to see that it is impossible to send information superluminally using entanglement: no matter what one does to spin 1, the outcomes of measurements on spin 2 are unaffected by that action. In particular, operations on spin 1 correspond to operators of the form $A_1 \otimes Id_2$, while operations on spin 2 correspond to operators of the form $Id_1 \otimes B_2$. The commutator between such operators is

$$[A_1 \otimes Id_2, Id_1 \otimes B_2] = A_1 \otimes B_2 - A_1 \otimes B_2 = 0\,. \tag{30}$$

Since they commute, it doesn't matter if one does something to spin 1 first, and then measures spin 2, or if one measures spin 2 first and then does something to spin 1: the results of the measurement will be the same. That is, nothing one does to spin 1 on its own can effect spin 2.

Nonetheless, entanglement is counterintuitive. One's classical intuition would like to believe that before the measurement, the system to be measured is in a definite state, even if that definite state is unknown. Such a definite state would constitute a 'hidden variable', an unknown, classical value for the measured variable. Entanglement implies that such hidden variables can't exist in any remotely satisfactory form. The spin version of the EPR effect described above is due to David Bohm [61]. Subsequently, John Bell proposed a set of relations, the 'Bell inequalities', that a hidden variable theory should obey [62]. Bell's inequalities are expressed in terms of the probabilities for the outcomes of measurements made on the two spins along different axes.

Suppose that each particle indeed has a particular value of spin along each axis before it is measured. Designate a particle that has spin up along the x-axis, spin down along the y-axis, and spin up along the z-axis by $(x+, y-, z+)$. Designate other possible orientations similarly. In a collection of particles, let $N(x+, y-, z+)$ be the number of particles with orientations $(x+, y-, z+)$. Clearly, $N(x+, y-) = N(x+, y-, z+) + N(x+, y-, z-)$. Now, in a collection of measurements made on pairs of particles, originally in a singlet state, let $\#(x_1+, y_2-)$ be the number of measurements that give the result spin up along the x-axis for particle 1, and spin down along the y-axis for particle 2. Bell showed that for classical particles that actually possess definite values of spin along different axes before measurement, $\#(x_1+, y_2+) \leq \#(x_1+, z_2+) + \#(y_1-, z_2-)$, together with inequalities that are obtained by permuting axes and signs.

Quantum mechanics decisively violates these Bell inequalities: in entangled states like the singlet state, particles simply do not possess definite, but unknown, values of spin before they are measured. Bell's inequalities have been verified experimentally on numerous occasions [13], although not all exotic forms of hidden variables have been eliminated. Those that are consistent with experiment are not very aesthetically appealing however (depending, of course, on one's aesthetic ideals). A stronger set of inequalities than Bell's are the CHSH inequalities (Clauser–Horne–Shimony–Holt), which have also been tested in numerous venues, with the predictions of quantum mechanics confirmed each time [63]. One of weirdest violation of classical intuition can be found in the so-called GHZ experiment, named after Daniel Greenberger, Michael Horne, and Anton Zeilinger [64].

To demonstrate the GHZ paradox, begin with the three-qubit state

$$|\chi\rangle = (1/\sqrt{2})(|\uparrow\uparrow\uparrow\rangle - |\downarrow\downarrow\downarrow\rangle) \tag{31}$$

(note that in writing this state we have suppressed the tensor product $\otimes$ signs, as mentioned above). Prepare this state four separate times, and make four distinct measurements. In the first measurement measure σ_x on the first qubit, σ_y on the second qubit, and σ_y on the third qubit. Assign the value $+1$ to the result, spin up along the axis measured, and -1 to spin down. Multiply the outcomes together. Quantum mechanics predicts that the result of this multiplication will always be $+1$, as can be verified by taking the expectation value $\langle\chi|\sigma_x^1 \otimes \sigma_y^2 \otimes \sigma_y^3|\chi\rangle$ of the operator $\sigma_x^1 \otimes \sigma_y^2 \otimes \sigma_y^3$ that corresponds to making the three individual spin measurements and multiplying their results together.

In the second measurement measure σ_y on the first qubit, σ_x on the second qubit, and σ_y on the third qubit. Multiply the results together. Once again, quantum mechanics predicts that the result will be $+1$. Similarly, in the third measurement measure σ_y on the first qubit, σ_y on the second qubit, and σ_x on the third qubit. Multiply the results together to obtain the predicted result $+1$. Finally, in the fourth measurement measure σ_x on all three qubits and multiply the results together. Quantum mechanics predicts that this measurement will give the result $\langle\chi|\sigma_x^1 \otimes \sigma_x^2 \otimes \sigma_x^3|\chi\rangle = -1$.

So far, these predictions may not seem strange. A moment's reflection, however, will reveal that the results of the four GHZ experiments are completely incompatible with any underlying assignment of values of ± 1 to the spin along the x- and y-axes before the measurement. Suppose that such pre-measurement values existed, and that these are the values revealed by the measurements. Looking at the four measurements, each consisting of three individual spin measurements, one sees that each possible spin measurement appears twice in the full sequence of twelve individual spin measurements. For example, measurement of spin 1 along the x-axis occurs in the first of the four three-fold measurements, and in the last one. Similarly, measurement of spin 3 along the 3-axis occurs in the first and second three-fold measurements. The classical consequence of each individual measurement occurring twice is that the product of all twelve measurements should be $+1$. That is, if measurement of σ_x^1 in the first measurement yields the result -1, it should also yield the result -1 in the fourth measurement. The product of the outcomes for σ_x^1 then gives $(-1) \times (-1) = +1$; similarly, if σ_x^1 takes on the value $+1$ in both measurements, it also contributes $(+1) \times (+1) = +1$ to the overall product. So if each spin possesses a definite value before the measurement, classical mechanics unambiguously predicts that the product of all twelve individual measurements should be $+1$.

Quantum mechanics, by contrast, unambiguously predicts that the product of all twelve individual measurements should be -1. The GHZ experiment has been performed in a variety of different quantum-mechanical systems, ranging from nuclear spins to photons [65,66]. The result: the predictions of classical mechanics are wrong and those of quantum mechanics are correct. Quantum weirdness triumphs.

Quantum Computation

Quantum mechanics has now been treated in sufficient detail to allow us to approach the most startling consequence of quantum weirdness: quantum computation. The central counterintuitive feature of quantum mechanics is quantum superposition: unlike a classical bit, which either takes on the value 0 or the value 1, a quantum bit in the superposition state $\alpha|0\rangle + \beta|1\rangle$ takes on the values 0 and 1 simultaneously. A quantum computer is a device that takes advantage of quantum superposition to process information in ways that classical computers can't. A key feature of any quantum computation is the way in which the computation puts entanglement to use: just as entanglement plays a central role in the quantum paradoxes discussed above, it also lies at the heart of quantum computation.

A classical digital computer is a machine that can perform arbitrarily complex logical operations. When you play a computer game, or operate a spread sheet, all that is going on is that your computer takes in the information from your joy stick or keyboard, encodes that information as a sequence of zeros and ones, and then performs

sequences of simple logical operations one that information. Since the work of George Boole in the first half of the nineteenth century, it is known that any logical expression, no matter how involved, can be broken down into sequences of elementary logical operations such as *NOT*, *AND*, *OR* and *COPY*. In the context of computation, these operations are called 'logic gates': a logic gates takes as input one or more bits of information, and produces as output one or more bits of information. The output bits are a function of the input bits. A *NOT* gate, for example, takes as input a single bit, X, and returns as output the flipped bit, *NOT* X, so that $0 \to 1$ and $1 \to 0$. Similarly, an *AND* gate takes in two bits X, Y as input, and returns the output X *AND* Y. X *AND* Y is equal to 1 when both X and Y are equal to 1; otherwise it is equal to 0. That is, an *AND* gate takes $00 \to 0$, $01 \to 0$, $10 \to 0$, and $11 \to 1$. An *OR* gate takes X, Y to 1 if either X or Y is 1, and to 0 if both X and Y are 0, so that $00 \to 0$, $01 \to 1$, $10 \to 1$, and $11 \to 1$. A *COPY* gate takes a single input, X, and returns as output two bits X that are copies of the input bit, so that $0 \to 00$ and $1 \to 11$.

All elementary logical operations can be built up from *NOT, AND, OR* and *COPY*. For example, implication can be written $A \to B \equiv A\ OR\ (NOT\ B)$, since $A \to B$ is false if and only if A is true and B is false. Consequently, any logical expression, e. g.,

$$\Big(\big(A\ AND\ (NOT\ B)\big)\ OR\ \big(C\ AND\ (NOT\ A)\big)\ AND\ \big(NOT(C\ OR\ B)\big)\Big)\,, \quad (32)$$

can be evaluated using *NOT*, *AND*, *OR*, and *COPY* gates, where *COPY* gates are used to supply the different copies of A, B and C that occur in different places in the expression. Accordingly, {*NOT*, *AND*, *OR*, *COPY*} is said to form a 'computationally universal' set of logic gates. Simpler computationally universal sets of logic gates also exist, e. g. {*NAND*, *COPY*}, where $X\ NAND\ Y = NOT(X\ AND\ Y)$.

Reversible Logic

A logic gate is said to be *reversible* if its outputs are a one-to-one function of its inputs. *NOT* is reversible, for example: since $X = NOT(NOT\ X)$, *NOT* is its own inverse. *AND* and *OR* are not reversible, as the value of their two input bits cannot be inferred from their single output. *COPY* is reversible, as its input can be inferred from either of its outputs.

Logical reversibility is important because the laws of physics, at bottom, are reversible. Above, we saw that the time evolution of a closed quantum system (i. e., one that is not interacting with any environment) is given by a unitary transformation: $|\psi\rangle \to |\psi'\rangle = U|\psi\rangle$. All unitary transformations are invertible: $U^{-1} = U^\dagger$, so that $|\psi\rangle = U^\dagger U|\psi\rangle = U^\dagger|\psi'\rangle$. The input to a unitary transformation can always be obtained from its output: the time evolution of quantum mechanical systems is one-to-one. As noted in the introduction, in 1961, Rolf Landauer showed that the underlying reversibility of quantum (and also of classical) mechanics implied that logically irreversible operations such as *AND* necessarily required physical dissipation [19]: any physical device that performs an *AND* operation must possess additional degrees of freedom (i. e., an environment) which retain the information about the actual values of the inputs of the *AND* gate after the irreversible logical operation has discarded those values. In a conventional electronic computer, those additional degrees of freedom consist of the microscopic motions of electrons, which, as Maxwell and Boltzmann told us, register large amounts of information.

Logic circuits in contemporary electronic circuits consist of field effect transistors, or FETs, wired together to perform *NOT*, *AND*, *OR* and *COPY* operations. Bits are registered by voltages: a FET that is charged at higher voltage registers a 1, and an uncharged FET at Ground voltage registers a 0. Bits are erased by connecting the FET to Ground, discharging them and restoring them to the state 0. When such an erasure or resetting operation occurs, the underlying reversibility of the laws of physics insure that the microscopic motions of the electrons in the Ground still retain the information about whether the FET was charged or not, i. e., whether the bit before the erasure operation registered 1 or 0. In particular, if the bit registered 1 initially, the electrons in Ground will be slightly more energetic than if it registered 0. Landauer argued that any such operation that erased a bit required dissipation of energy $k_B T \ln 2$ to an environment at temperature T, corresponding to an increase in the environment's entropy of $k_B \ln 2$.

Landauer's principle can be seen to be a straightforward consequence of the microscopic reversibility of the laws of physics, together with the fact that entropy is a form of information – information about the microscopic motions of atoms and molecules. Because the laws of physics are reversible, any information that resides in the logical degrees of freedom of a computer at the beginning of a computation (i. e., in the charges and voltages of FETs) must still be present at the end of the computation in some degrees of freedom, either logical or microscopic. Note that physical reversibility also implies that if information can flow from logical degrees of freedom to microscopic degrees of freedom, then it can also flow back again:

the microscopic motions of electrons cause voltage fluctuations in FETs which can give rise to logical errors. Noise is necessary.

Because *AND*, *OR*, *NAND* are not logically reversible, Landauer initially concluded that computation was necessarily dissipative: entropy had to increase. As is often true in the application of the second law of thermodynamics, however, the appearance of irreversibility does not always imply the actual fact of irreversibility. In 1963, Lecerf showed that digital computation could always be performed in a reversible fashion [20]. Unaware of Lecerf's work, in 1973 Bennett rederived the possibility of reversible computation [21]. Most important, because Bennett was Landauer's colleague at IBM Watson laboratories, he realized the physical significance of embedding computation in a logically reversible context. As will be seen, logical reversibility is essential for quantum computation.

A simple derivation of logically reversible computation is due to Fredkin, Toffoli, and Margolus [22]. Unaware of Bennett's work, Fredkin constructed three-input, three-output reversible logic gates that could perform *NOT*, *AND*, *OR*, and *COPY* operations. The best-known example of such a gate is the Toffoli gate. The Toffoli gate takes in three inputs, X, Y, and Z, and returns three outputs, X', Y' and Z'. The first two inputs go through unchanged, so that $X' = X$, $Y' = Y$. The third output is equal to the third input, unless both X and Y are equal to 1, in which case the third output is the *NOT* of the third input. That is, when either X or Y is 0, $Z' = Z$, and when both X and Y are 1, $Z' = NOT\ Z$. (Another way of saying the same thing is to say that $Z' = Z\ XOR\ (X\ AND\ Y)$, where *XOR* is the exclusive *OR* operation whose output is 1 when either one of its inputs is 1, but not both. That is, *XOR* takes $00 \to 0$, $01 \to 1$, $10 \to 1$, $11 \to 0$.) Because it performs a *NOT* operation on Z controlled on whether both X and Y are 1, a Toffoli gate is often called a controlled-controlled-*NOT* (*CCNOT*) gate.

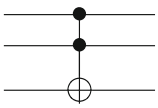

Quantum Information Processing, Figure 1
A Toffoli gate

To see that *CCNOT* gates can be wired together to perform *NOT*, *AND*, *OR*, and *COPY* operations, note that when one sets the first two inputs X and Y both to the value 1, and allows the input Z to vary, one obtains $Z' = NOT\ Z$. That is, supplying additional inputs allows a *CCNOT* to perform a *NOT* operation. Similarly, setting the input Z to 0 and allowing X and Y to vary yields $Z' = X\ AND\ Y$. *OR* and *COPY* (not to mention *NAND*) can be obtained by similar methods. So the ability to set inputs to predetermined values, together with ability to apply *CCNOT* gates allows one to perform any desired digital computation.

Because reversible computation is intrinsically less dissipative than conventional, irreversible computation, it has been proposed as a paradigm for constructing low power electronic logic circuits, and such low power circuits have been built and demonstrated [67]. Because additional inputs and wires are required to perform computation reversibly, however, such circuits are not yet used for commercial application. As the miniaturization of the components of electronic computers proceeds according to Moore's law, however, dissipation becomes an increasingly hard problem to solve, and reversible logic may become commercially viable.

Quantum Computation

In 1980, Benioff proposed a quantum-mechanical implementation of reversible computation [23]. In Benioff's model, bits corresponded to spins, and the time evolution of those spins was given by a unitary transformation that performed reversible logic operations. (In 1986, Feynman embedded such computation in a local, Hamiltonian dynamics, corresponding to interactions between groups of spins [68].) Benioff's model did not take into account the possibility of putting quantum bits into superpositions as an integral part of the computation, however. In 1985, Deutsch proposed that the ordinary logic gates of reversible computation should be supplemented with intrinsically quantum-mechanical single qubit operations [25]. Suppose that one is using a quantum-mechanical system to implement reversible computations using *CCNOT* gates. Now add to the ability to prepare qubits in desired states, and to perform *CCNOT* gates, the ability to perform single-qubit rotations of the form $e^{-i\theta\vec{\sigma}/2}$ as described above. Deutsch showed that the resulting set of operations allowed *universal quantum computation*. Not only could such a computer perform any desired classical logical transformation on its quantum bits; it could perform any desired unitary transformation U whatsoever.

Deutsch pointed out that a computer endowed with the ability to put quantum bits into superpositions and to perform reversible logic on those superpositions could compute in ways that classical computers could not. In particular, a classical reversible computer can evaluate any desired function of its input bits: $(x_1 \ldots x_n, 0 \ldots 0) \to (x_1 \ldots x_n, f(x_1 \ldots x_n))$, where x_i represents the logical value, 0 or 1, of the ith bit, and f is the desired function. In

order to preserve reversibility, the computer has been supplied with an 'answer' register, initially in the state $00\ldots0$, into which to place the answer $f(x_1\ldots x_n)$. In a quantum computer, the input bits to any transformation can be in a quantum superposition. For example, if each input bit is in an equal superposition of 0 and 1, $(1/\sqrt{2})(|0\rangle+|1\rangle)$, then all n qubits taken together are in the superposition

$$\frac{1}{2^{n/2}}(|00\ldots0\rangle+|00\ldots1\rangle+\cdots+|11\ldots1\rangle = \frac{1}{2^{n/2}}\sum_{x_1,\ldots,x_n=0,1}|x_1\ldots x_n\rangle\,. \quad (33)$$

If such a superposition is supplied to a quantum computer that performs the transformation $x_1\ldots x_n \to f(x_1\ldots x_n)$, then the net effect is to take the superposition

$$\frac{1}{2^{n/2}}\sum_{x_1,\ldots,x_n=0,1}|x_1\ldots x_n\rangle|00\ldots0\rangle \to \frac{1}{2^{n/2}}\sum_{x_1,\ldots,x_n=0,1}|x_1\ldots x_n\rangle|f(x_1\ldots x_n)\rangle\,. \quad (34)$$

That is, even though the quantum computer evaluates the function f only once, it evaluates it on every term in the superposition of inputs simultaneously, an effect which Deutsch termed 'quantum parallelism'.

At first, quantum parallelism might seem to be spectacularly powerful: with only one function 'call', one performs the function on 2^n different inputs. The power of quantum parallelism is not so easy to tease out, however. For example, suppose one makes a measurement on the output state in (33) in the $\{|0\rangle,|1\rangle\}$ basis. The result is a *randomly selected* input-output pair, $(x_1\ldots x_n, f(x_1\ldots x_n))$. One could have just as easily obtained such a pair by feeding a random input string into a classical computer that evaluates f. As will now be seen, the secret to orchestrating quantum computations that are more powerful than classical computations lies in arranging quantum interference between the different states in the superposition of equation (34).

The word 'orchestration' in the previous sentence was used for a reason. In quantum mechanics, states of physical systems correspond to waves. For example, the state of an electron is associated with a wave that is the solution of the Schrödinger equation for that electron. Similarly, in a quantum computer, a state such as $|x_1\ldots x_n\rangle|f(x_1\ldots x_n)\rangle$ is associated with a wave that is the solution of the Schrödinger equation for the underlying quantum degrees of freedom (e. g., electron spins or photon polarizations) that make up the computers quantum bits. The waves of quantum mechanics, like waves of water, light, or sound, can be superposed on each other to construct composite waves. A quantum computer that performs a conventional reversible computation, in which its qubits only take on the values 0 or 1 and are never in superpositions $\alpha|0\rangle+\beta|1\rangle$, can be thought of as an analogue of a piece of music like a Gregorian chant, in which a single, unaccompanied voice follows a prescribed set of notes. A quantum computer that performs many computations in quantum parallel is analogous to a symphony, in which many lines or voices are superposed to create chords, counterpoint, and harmony. The quantum computer programmer is the composer who writes and orchestrates this quantum symphony: her job is to make that counterpoint reveal meaning that is not there in each of the voices taken separately.

Deutsch–Jozsa Algorithm

Let's examine a simple example, due to David Deutsch and Richard Jozsa, in which the several 'voices' of a quantum computer can be orchestrated to solve a problem more rapidly than a classical computer [69]. Consider the set of functions f that take one bit of input and produce one bit of output. There are four such functions:

$$f(x)=0,\quad f(x)=1,\quad f(x)=x,\quad f(x)=NOT\,x\,. \quad (35)$$

The first two of these functions are constant functions; the second two are 'balanced' in the sense that half of their inputs yield 0 as output, while the other half yield 1. Suppose that one is presented with a 'black box' that implements one of these functions. The problem is to query this black box and to determine whether the function the box contains is constant or balanced.

Classically, it clearly takes exactly two queries to determine whether the function in the box is constant or balanced. Using quantum information processing, however, one query suffices. The following quantum circuit shows how this is accomplished.

Deutsch–Jozsa Circuit

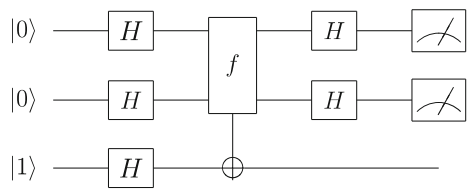

Quantum Information Processing, Figure 2
2-Qubit Deutsch–Jozsa circuit

Quantum circuit diagrams are similar in character to their classical counterparts: qubits enter on the left, un-

dergo a series of transformations effected by quantum logic gates, and exit at the right, where they are measured. In the circuit above, the first gate, represented by H is called a Hadamard gate. The Hadamard gate is a single-qubit quantum logic gate that effects the transformation

$$\begin{aligned} |0\rangle &\to (1/\sqrt{2})(|0\rangle + |1\rangle)\,, \\ |1\rangle &\to (1/\sqrt{2})(|0\rangle - |1\rangle)\,. \end{aligned} \tag{36}$$

In other words, the Hadamard performs a unitary transformation $U_H = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$ on its single-qubit input. Note that the Hadamard transformation is its own inverse: $U_H^2 = Id$.

The second logic gate implements the unknown, black-box function f. It takes two binary inputs, x, y, and gives two binary outputs. The gate leaves the first input unchanged, and adds $f(x)$ to the second input (modulo 2), so that $x \to x$ and $y \to y + f(x)(\bmod 2)$. Such gates can be implemented using the controlled-*NOT* operation introduced above. Recall that the controlled-*NOT* or *CNOT* leaves its first input bit unchanged, and flips the second if and only if the first input is 1. In the symbol for a controlled-*NOT* operation, the $\bullet$ part represents the control bit and the $\oplus$ part represents the bit that can be flipped. The circuits required to implement the four different functions from one bit to one bit are as follows:

$$f(x) = 0: \quad f(x) = 1: \quad f(x) = x: \quad f(x) = NOT\, x: \tag{37}$$

The black box in the Deutsch–Jozsa algorithm contains one of these circuits. Note that the black-box circuits are 'classical' in the sense that they map input combinations of 0's and 1's to output combinations of 0's and 1's: the circuits of (37) make sense as classical circuits as well as quantum circuits.

Any classical circuit that can determine whether f is constant or balanced requires at least two uses of the f gate. By contrast, the Deutsch–Jozsa circuit above requires only one use of the f gate. Going through the quantum logic circuit, one finds that a constant function yields the output $|0\rangle$ on the first output line, while a balanced function yields the output $|1\rangle$ (up to an overall, unobservable phase). That is, only a single function call is required to reveal whether f is constant or balanced.

Several comments on the Deutsch–Jozsa algorithm are in order. The first is that, when comparing quantum algorithms to classical algorithms, it is important to compare apples to apples: that is, the gates used in the quantum algorithm to implement the black-box circuits should be the same as those used in any classical algorithms. The difference, of course, is that the quantum gates preserve quantum coherence, a concept which is meaningless in the classical context. This requirement has been respected in the Deutsch–Jozsa circuit above.

The second comment is that the Deutsch–Jozsa algorithm is decidedly odd and counterintuitive. The f gates and the controlled-*NOT* gates from which they are constructed both have the property that the first input passes through unchanged $|0\rangle \to |0\rangle$ and $|1\rangle \to |1\rangle$. Yet somehow, when the algorithm is identifying balanced functions, the first bit flips. How can this be? This is the part where quantum weirdness enters. Even though the f and controlled-*NOT* gates leave their first input unchanged in the logical basis $\{|0\rangle, |1\rangle\}$, the same property does not hold in other bases. For example, let $|+\rangle = (1/\sqrt{2})(|0\rangle + |1\rangle) = U_H|0\rangle$, and let $|1\rangle = (1/\sqrt{2})(|0\rangle - |1\rangle) = U_H|1\rangle$. Straitforward calculation shows that, when acting on the basis $\{|+\rangle, |-\rangle\}$, the CNOT still behaves like a CNOT, but with roles of its imputs reversed: now the second qubit passes through unchanged, while the first qubit gets flipped.

That is, when acting on the basis $\{|+\rangle, |-\rangle\}$, the *CNOT* still behaves like a *CNOT*, but with the roles of its inputs reversed: now the second qubit passes through unchanged, while the first qubit gets flipped. It is this quantum role reversal that underlies the efficacy of the Deutsch–Jozsa algorithm. Pretty weird.

It is important to note that the Deutsch–Jozsa algorithm is not just a theoretical point. The algorithm has been implemented using techniques from nuclear magnetic resonance (NMR) [33]. The results are exactly as predicted by quantum mechanics: a single function call suffices to determine whether that function is constant or balanced.

The two-qubit algorithm was first described by David Deutsch. Later, with Richard Jozsa, he extended this algorithm to a multi-qubit algorithm. Now consider functions f from n qubits to a single qubit. Once again, the problem is to determine whether or not f is constant or balanced. That is, the function f in the black box is either constant: $f(x) = 0$ for all n-bit inputs x, or $f(x) = 1$ for all x, or balanced: $f(x) = 0$ for exactly half of its 2^n possible input strings, and $f(x) = 1$ for the other half. (If this problem statement seems somewhat artificial, note that the algorithm works equally well for distinguishing between constant functions and 'typical' functions, which are *approximately* balanced.)

On average, a classical algorithm takes a little more than two function calls to distinguish between a constant or a balanced function. However, in the worst case, it takes $2^{n-1} + 1$ calls, as more than half the inputs have to

be sampled. As before, the quantum algorithm takes but a single function call, as the following circuit shows.

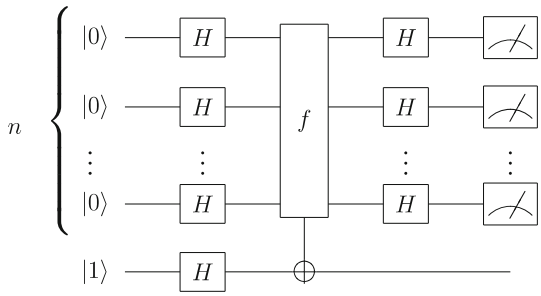

Quantum Information Processing, Figure 3
Caption: *n*-Qubit Deutsch–Jozsa circuit

To determine whether the f is constant or balanced, one measures the first n output bits: if they are all 0, then the function is constant; if one or more is 1, then the function is balanced.

Other Algorithms: The Quantum Fourier Transform

While it conclusively demonstrates that quantum computers are strictly more powerful than classical computers for certain problems, the Deutsch–Jozsa algorithm does not solve a problem of burning interest to applied computer scientists. Once it was clear that quantum computers could offer a speedup over classical algorithms, however, other algorithms began to be developed. Simon's algorithm [70], for example, determines whether a function f from n bits to n bits is (a) one-to-one, or (b) two-to-one with a large period s, so that $f(x+s) = f(x)$ for all x. (In Simon's algorithm the addition is bitwise modulo 2, with no carry bits.)

Simon's algorithm has a similar 'flavor' to the Deutsch–Jozsa algorithm: it is intriguing but does not obviously admit wide application. A giant step towards constructing more useful algorithms was Coppersmith's introduction [71] of the Quantum Fourier Transform (QFT). The fast Fourier transform maps a function of n bits to its discrete Fourier transform function:

$$f(x) \to g(y) = \sum_{x=0}^{2^n-1} \mathrm{e}^{2\pi i x y/2^n} f(x) \,. \tag{38}$$

The fast Fourier transform takes $O(n2^n)$ steps. The quantum Fourier transform takes a *wave function* over n qubits to a Fourier transformed wave function:

$$\sum_{x=0}^{2^n-1} f(x)|x\rangle \to 2^{-n/2} \sum_{x,y=0}^{2^n-1} \mathrm{e}^{2\pi i x y/2^n} f(x)|y\rangle \,. \tag{39}$$

It is not difficult to show that the quantum Fourier transform is a unitary.

To obtain a quantum logic circuit that accomplishes the QFT, it is convenient to express states in a binary representation. In the equations above, x and y are n-bit numbers. Write x as $x_n \ldots x_1$, where $x_n, \ldots x_1$ are the bits of x. This is just a more concise way of saying that $x = x_1 2^0 + \cdots + x_n 2^{n-1}$. Similarly, the expression $0.y_1 \ldots y_m$ represents the number $y_1/2 + \ldots y_m/2^{-m}$. Using this binary notation, it is not hard to show that the quantum Fourier transform can be written:

$$\begin{aligned} |x_1 \ldots x_n\rangle \to & 2^{-n/2}(|0\rangle + \mathrm{e}^{2\pi i 0.x_1}|1\rangle)(|0\rangle \\ & + \mathrm{e}^{2\pi i 0.x_2 x_1}|1\rangle) \ldots (|0\rangle + \mathrm{e}^{2\pi i 0.x_n \ldots x_1}|1\rangle) \,. \end{aligned} \tag{40}$$

When the quantum Fourier transform is written in this form, it is straightforward to construct a circuit that implements it (see Fig. 4).

Note that the QFT circuit for wave functions over n qubits takes $O(n^2)$ steps: it is exponentially faster than the FFT for functions over n bits, which takes $O(n2^n)$ steps. This exponential speedup of the quantum Fourier transform is what guarantees the efficacy of many quantum algorithms.

The quantum Fourier transform is a potentially powerful tool for obtaining exponential speedups for quantum computers over classical computers. The key is to find a way of posing the problem to be solved in terms of finding periodic structure in a wave function. This step is the essence of the best known quantum algorithm, Shor's algorithm for factoring large numbers [26].

Shor's Algorithm

The factoring problem can be stated as follows: Given $N = pq$, where p, q are prime, find p and q. For large p and q, this problem is apparently hard for classical computers. The fastest known algorithm (the 'number sieve') takes $O(N^{1/3})$ steps. The apparent difficulty of the factoring problem for classical computers is important for cryptography. The commonly used RSA public-key cryptosystem relies on the difficulty of factoring to guarantee security. Public-key cryptography addresses the following societally important situation. Alice wants to send Bob some secure information (e. g., a credit card number). Bob sends Alice the number N, but does not reveal the identity of p or q. Alice then uses N to construct an encrypted version of the message she wishes to send. Anyone who wishes to decrypt this message must know what p and q are. That is,

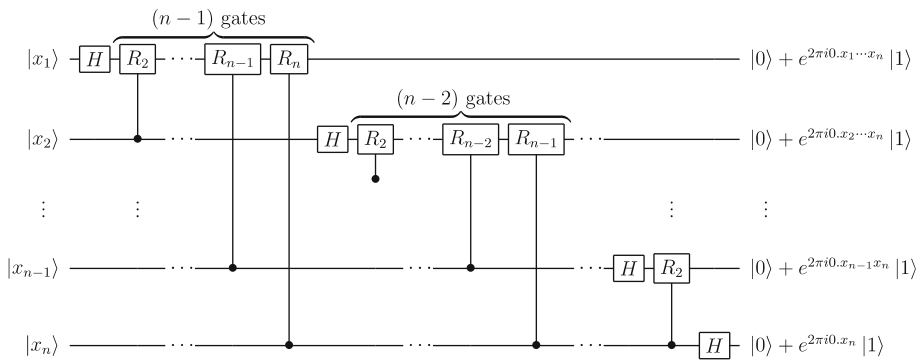

Quantum Information Processing, Figure 4
Quantum Fourier Transform

encryption can be performed using the public key N, but decryption requires the private key p, q.

In 1994, Peter Shor showed that quantum computers could be used to factor large numbers and so crack public-key cryptosystems that whose security rests on the difficulty of factoring [26]. The algorithm operates by solving the so-called 'discrete logarithm' problem. This problem is, given N and some number x, find the smallest r such that $x^r \equiv 1(\bmod N)$. Solving the discrete logarithm allows N to be factored by the following procedure. First, pick $x < N$ at random. Use Euclid's algorithm to check that the greatest common divisor of x and N is 1. (Euclid's algorithm is to divide N by x; take the remainder r_1 and divide x by r_1; take the remainder of that division, r_2 and divide r_1 by that, etc. The final remainder in this procedure is the greatest common divisor, or g.c.d., of x and N.) If the g.c.d. of x and N is not 1, then it is either p or q and we are done.

If the greatest common divisor of x and N is 1, suppose that we can solve the discrete logarithm problem to find the smallest r such that $x^r \equiv 1(\bmod N)$. As will be seen, if r is even, we will be able to find the factors of N easily. If r turns out to be odd, just pick a new x and start again: continue until you obtain an even r (since this occurs half the time, you have to repeat this step no more than twice on average). Once an even r has been found, we have $(x^{r/2}-1)(x^{r/2}+1) \equiv 1(\bmod N)$. In other words, $(x^{r/2}-1)(x^{r/2}+1) = bN = bpq$ for some b. Finding the greatest common divisor of $x^{r/2}-1$, $x^{r/2}+1$ and N now reveals p and q. The goal of the quantum algorithm, then, is to solve the discrete logarithm problem to find the smallest r such that $x^r \equiv 1 \bmod N$. If r can be found, then N can be factored.

In its discrete logarithm guise, factoring possesses a periodic structure that the quantum Fourier transform can reveal. First, find an x whose g.c.d. with N is 1, as above, and pick n so that $N^2 < 2^n < 2N^2$. The quantum algorithm uses two n-qubit registers. Begin by constructing a uniform superposition $2^{-n/2}\sum_{k=0}^{2^n-1}|k\rangle|0\rangle$. Next, perform exponentiation modulo N to construct the state,

$$2^{-n/2}\sum_{k=0}^{2^n-1}|k\rangle|x^k \bmod N\rangle\,. \tag{41}$$

This modular exponentiation step takes $O(n^3)$ operations (note that $x^{2^k} \bmod N$ can be evaluated by first constructing $x^2 \bmod N$, then constructing $(x^2)^2 \bmod N$, etc.). The periodic structure in (41) arises because if $x^k \equiv a \bmod N$, for some a, then $x^{k+r} \equiv a \bmod N$, $x^{k+2r} \equiv a \bmod N$, $\ldots$, $x^{k+mr} \equiv a \bmod N$, up to the largest m such that $k + mr < N^2$. The same periodicity holds for any a. That is, the wave function (41) is periodic with with period r. So if we apply the quantum Fourier transform to this wave function, we can reveal that period and find r, thereby solving the discrete logarithm and factoring problems.

To reveal the hidden period and find r apply the QFT to the *first* register in the state (41). The result is

$$2^{-n}\sum_{jk=0}^{2^n-1} e^{2\pi i jk/2^n}|j\rangle|x^k \bmod N\rangle\,. \tag{42}$$

Because of the periodic structure, positive interference takes place when $j(k+\ell r)$ is close to a multiple of 2^n. That is, measuring the first register now yields a number j such that $jr/2^n$ is close to an integer: only for such j does the necessary positive interference take place. In other words, the algorithm reveals a j such that $j/2^n = s/r$ for some integer s. That is, to find r, we need to find fractions s/r that approximate $j/2^n$. Such fractions can be obtained us-

ing a continued fraction approximation. With reasonably high probability, the result of the continued fraction approximation can be shown to yield the desired answer r. (More precisely, repetition of this procedure $O(2\log N)$ times suffices to identify r.) Once r is known, then the factors p and q of N can be recovered by the reduction of factoring to discrete logarithm given above.

The details of Shor's algorithm reveal considerable subtlety, but the basic idea is straightforward. In its reduction to discrete logarithm, factoring possesses a hidden periodic structure. This periodic structure can be revealed using a quantum Fourier transform, and the period itself in turn reveals the desired factors.

More recent algorithms also put the quantum Fourier transform to use to extract hidden periodicities. Notably, the QFT can be used to find solutions to Pell's equation ($x^2 - ny^2 = 1$, for non-square n) [72]. Generalizations of the QFT to transforms over groups (the dihedral group and the permutation group on n objects S_n) have been applied to other problems such as the shortest vector on a lattice [73] (dihedral group, with some success) and the graph isomorphism problem (S_n, without much success [74]).

The Phase-Estimation Algorithm

One of the most useful applications of the quantum Fourier transform is finding the eigenvectors and eigenvalues of unitary transformations. The resulting algorithm is called the 'phase-estimation' algorithm: its original form is due to Kitaev [75]. Suppose that we have the ability to apply a 'black box' unitary transformation U. U can be written $U = \sum_j e^{i\phi_j}|j\rangle\langle j|$, where $|j\rangle$ are the eigenvectors of U and $e^{i\phi_j}$ are the corresponding eigenvalues. The goal of the algorithm is to estimate the $e^{i\phi_j}$ and the $|j\rangle$. (The goal of the original Kitaev algorithm was only to estimate the eigenvalues $e^{i\phi_j}$. However, Abrams and Lloyd showed that the algorithm could also be used to construct and estimate the eigenvectors $|j\rangle$, as well [76]. The steps of the phase estimation algorithm are as follows.

(0) Begin with the initial state $|0\rangle|\psi\rangle$, where $|0\rangle$ is the n-qubit state $00\ldots0\rangle$, and $|\psi\rangle$ is the state that one wishes to decompose into eigenstates: $|\psi\rangle = \sum_j \psi_j|j\rangle$.

(1) Using Hadamards or a QFT, put the first register into a uniform superposition of all possible states:

$$\to 2^{-n/2}\sum_{k=0}^{2^n-1}|k\rangle|\psi\rangle\,.$$

(2) In the kth component of the superposition, apply U^k to $|\psi\rangle$:

$$\begin{aligned}\to\ & 2^{-n/2}\sum_{k=0}^{2^n-1}|k\rangle U^k|\psi\rangle\\ =\ & 2^{-n/2}\sum_{j,k=0}^{2^n-1}|k\rangle U^k\psi_j|j\rangle\\ =\ & 2^{-n/2}\sum_{j,k=0}^{2^n-1}\psi_k e^{ik\phi_j}|k\rangle|j\rangle\,.\end{aligned}$$

(3) Apply inverse QFT to first register:

$$\to 2^{-n}\sum_{j,k,l=0}^{2^n-1}\psi_k e^{ik\phi_j}e^{-2\pi ikl/2^n}|l\rangle|j\rangle\,.$$

(4) Measure the registers. The second register contains the eigenvector $|j\rangle$. The first register contains $|l\rangle$ where $2\pi l/2^n \approx \phi_j$. That is, the first register contains an n-bit approximation to ϕ_j.

By repeating the phase-estimation algorithm many times, one samples the eigenvectors and eigenvalues of U. Note that to obtain n-bits of accuracy, one must possess the ability to apply U 2^n times. This feature limits the applicability of the phase-estimation algorithm to a relatively small number of bits of accuracy, or to the estimation of eigenvalues of Us that can easily be applied an exponentially large number of times. We've already seen such an example of a process in modular exponentiation. Indeed, Kitaev originally identified the phase estimation algorithm as an alternative method for factoring.

Even when only a relatively small number of applications of U can be performed, however, the phase-estimation algorithm can provide an exponential improvement over classical algorithms for problems such as estimating the ground state of some physical Hamiltonian [76,77], as will now be seen.

Quantum Simulation

One of the earliest uses for a quantum computer was suggested by Richard Feynman [24]. Feynman noted that simulating quantum systems on a classical computer was hard: computer simulations of systems such as lattice gauge theories take up a substantial fraction of all supercomputer time, and, even then, are often far less effective than their programmers could wish them to be. The reason why it's hard to simulate a quantum system on a classical computer is straightforward: in the absence of any sneaky

tricks, the only known way to simulate a quantum system's time evolution is to construct a representation of the full state of the system, and to evolve that state forward using the system's quantum equation of motion. To represent the state of a quantum system on a classical computer is typically exponentially hard, however: an n-spin system requires 2^n complex numbers to represent its state. Evolving that state forward is even harder: it requires exponentiation of a 2^n by 2^n matrix. Even for a small quantum system, for example, one containing fifty spins, this task lies beyond the reach of existing classical supercomputers. True, supercomputers are also improving exponentially in time (Moore's law). No matter how powerful they become, however, they will not be able to simulate more than 300 spins directly, for the simple reason that to record the 2^{300} numbers that characterize the state of the spins would require the use of all 2^{300} particles in the universe within the particle horizon.

Feynman noted that if one used qubits instead of classical bits, the state of an n-spin system can be represented using just n qubits. Feynman proposed a class of systems called 'universal quantum simulators' that could be programmed to simulate any other quantum system. A universal quantum simulator has to possess a flexible dynamics that can be altered at will to mimic the dynamics of the system to be simulated. That is, the dynamics of the universal quantum simulator form an *analog* to the dynamics of the simulated system. Accordingly, one might also call quantum simulators, 'quantum analog computers'.

In 1996, Lloyd showed how Feynman's proposal could be turned into a quantum algorithm [78]. For each degree of freedom of the system to be simulated, allocate a quantum register containing a sufficient number of qubits to approximate the state of that degree of freedom to some desired accuracy. If one wishes to simulate the system's interaction with the environment, a number of registers should also be allocated to simulate the environment (for a d-dimensional system, up to d^2 registers are required to simulate the environment). Now write the Hamiltonian of the system an environment as $H = \sum_{\ell=1}^{m} H_\ell$, where each H_ℓ operates on only a few degrees of freedom. The Trotter formula implies that

$$\begin{aligned} e^{-iH\delta t} =& e^{-iH_1\Delta t} \dots e^{-iH_m\Delta t} - \frac{1}{2}\sum_{jk}[H_j, H_k]\Delta t^2 \\ &+ O\left(\Delta t^3\right) . \end{aligned} \tag{43}$$

Each $e^{-iH_\ell\Delta t}$ can be simulated using quantum logic operations on the quantum bits in the registers corresponding to the degrees of freedom on which H_ℓ acts. To simulate the time evolution of the system over time $t = n\Delta t$, we simply apply $e^{-iH\Delta t}$ n times, yielding

$$\begin{aligned} e^{-iHt} =& \left(e^{-iH\Delta t}\right)^n \\ =& \left(\Pi_\ell e^{-iH_\ell\Delta t}\right)^n - \frac{n}{2}\sum_{jk}[H_j, H_k]\Delta t^2 \\ &+ O\left(\Delta t^3\right) . \end{aligned} \tag{44}$$

The quantum simulation takes $O(mn)$ steps, and reproduces the original time evolution to an accuracy $h^2t^2m^2/n$, where h is the average size of $\|[H_j, H_k]\|$ (note that for simulating systems with local interactions, most of these terms are zero, because most of the local interactions commute with each other).

A second algorithm for quantum simulation takes advantage of the quantum Fourier transform [79,80]. Suppose that one wishes to simulate the time evolution of a quantum particle whose Hamiltonian is of the form $H = P^2/2m + V(X)$, where $P = -i\partial/\partial x$ is the momentum operator for the particle, and $V(X)$ is the potential energy operator for the particle expressed as a function of the position operator X. Using an n-bit discretization for the state we identify the x eigenstates with $|x\rangle = |x_n \dots x_1\rangle$. The momentum eigenstates are then just the quantum Fourier transform of the position eigenstates: $|p\rangle = 2^{-n/2}\sum_{x=0}^{2^n-1} e^{2\pi ixp/2^n}|x\rangle$. That is, $P = U_{QFT}XU_{QFT}^\dagger$.

By the Trotter formula, the infinitesimal time evolution operator is

$$e^{-iH\Delta t} = e^{-iP^2\Delta t/2m}e^{-iV(X)\Delta t} + O\left(\delta t^2\right) . \tag{45}$$

To enact this time evolution operator one proceeds as above. Write the state of the particle in the x-basis: $|\psi\rangle = \sum_x \psi_x|x\rangle$. First apply the infinitesimal $e^{-iV(X)\Delta t}$ operator:

$$\sum_x \psi_x|x\rangle \to \sum_x \psi_x e^{-iV(x)}|x\rangle . \tag{46}$$

To apply the infinitesimal $e^{-iP^2\delta t/2m}$ operator, first apply an inverse quantum Fourier transform on the state, then apply the unitary transformation $|x\rangle \to e^{-ix^2\Delta t/2m}|x\rangle$, and finally apply the regular QFT. Because X and P are related by the quantum Fourier transform, these three steps effectively apply the transformation $e^{-iP^2\Delta t/2m}$. Applying first $e^{-iV(X)\Delta t}$ then $e^{-iP^2\Delta t/2m}$ yields the full infinitesimal time evolution (45). The full time evolution operator e^{-iHt} can then be built up by repeating the infinitesimal operator $t/\Delta t$ times. As before, the accuracy of the quantum simulation can be enhanced by slicing time ever more finely.

Quantum simulation represents one of the most powerful uses of quantum computers. It is probably the application of quantum computers that will first give an advantage over classical supercomputers, as only one hundred qubits or fewer are required to simulate, e. g., molecular orbitals or chemical reactions, more accurately than the most powerful classical supercomputer. Indeed, special purpose quantum simulators have already been constructed using nuclear magnetic resonance techniques [81]. These quantum analog computers involve interactions between many hundreds of nuclear spins, and so are already performing computations that could not be performed by any classical computer, even one the size of the entire universe.

Quantum Search

The algorithms described above afford an exponential speedup over the best classical algorithms currently known. Such exponential speedups via quantum computation are hard to find, and are currently limited to a few special problems. There exists a large class of quantum algorithms afford a *polynomial* speedup over the best possible classical algorithms, however. These algorithms are based on Grover's quantum search algorithm.

Grover's algorithm [31] allows a quantum computer to search an unstructured database. Suppose that this database contains N items, one of which is 'marked', and the remainder of which are unmarked. Call the marked item w, for 'winner'. Such a database can be represented by a function $f(x)$ on the items in the database, such that f of the marked item is 1, and f of any unmarked item is 0. That is, $f(w) = 1$, and $f(x \neq w) = 0$. A classical search for the marked item must take $N/2$ database calls, on average. By contrast, a quantum search for the marked item takes $O(\sqrt{N})$ calls, as will now be shown.

Unstructured database search is an 'oracle' problem. In computer science, an oracle is a 'black box' function: one can supply the black box with an input x, and the black box then provides an output $f(x)$, but one has no access to the mechanism inside the box that computes $f(x)$ from x. For the quantum case, the oracle is represented by a function on two registers, one containing x, and the other containing a single qubit. The oracle takes $|x\rangle|y\rangle \to |x\rangle|y + f(x)\rangle$, where the addition takes place modulo 2.

Grover originally phrased his algorithm in terms of a 'phase' oracle U_w, where $|x\rangle U_w|x\rangle = (-1)^{f(x)}|x\rangle$. In other words, the 'winner' state acquires a phase of -1: $|w\rangle \to -|w\rangle$, while the other states remain unchanged: $|x \neq w\rangle \to |x\rangle$. Such a phase oracle can be constructed from the original oracle in several ways. The first way involves two oracle calls. Begin with the state $|x\rangle|0\rangle$ and call the oracle once to construct the state $|x\rangle|f(x)\rangle$. Now apply a σ_z transformation to the second register. The effect of this is to take the state to $(-1)^{f(x)}|x\rangle|f(x)\rangle$. Applying the oracle for a second time yields the desired phase-oracle state $(-1)^{f(x)}|x\rangle|0\rangle$. A second, sneakier way to construct a phase oracle is to initialize the second qubit in the state $(1/\sqrt{2})(|0\rangle - |1\rangle)$. A *single* call of the original oracle on the state $|x\rangle(1/\sqrt{2})(|0\rangle - |1\rangle)$ then transforms this state into $(-1)^{f(x)}|x\rangle((1/\sqrt{2})(|0\rangle - |1\rangle))$. In this way a phase oracle can be constructed from a single application of the original oracle.

Two more ingredients are needed to perform Grover's algorithm. Let's assume that $N = 2^n$ for some n, so that the different states $|j\rangle$ can be written in binary form. Let U_0 be the unitary transformation that takes $|0\ldots0\rangle \to -|0\ldots0\rangle$, that takes $|j\rangle \to |j\rangle$ for $j \neq 0$. That is, U_0 acts in the same way as U_w, but applies a phase of -1 to $|0\ldots0\rangle$ rather than to $|w\rangle$. In addition, let H be the transformation that performs Hadamard transformations on all of the qubits individually.

Grover's algorithm is performed as follows. Prepare all qubits in the state $|0\rangle$ and apply the global Hadamard transformation H to create the state $|\psi\rangle = (1/\sqrt{N})\sum_{j=0}^{N-1}|j\rangle$. Apply, in succession, U_w, then H, then U_0, then H again. These four transformations make up the composite transformation $U_G = HU_0HU_w$. Now apply U_G again, and repeat for a total of $\approx (\pi/4)\sqrt{N}$ times (that is, the total number of times U_G is applied is equal to the integer closest to $(\pi/4)\sqrt{N}$). The system is now, with high probability, in the state $|w\rangle$. That is, $U_G^{\sqrt{N}}|0\ldots0\rangle \approx |w\rangle$. Since each application of U_G contains a single call to the phase oracle U_w, the winner state $|w\rangle$ has now been identified with $O(\sqrt{N})$ oracle calls, as promised.

The quantum algorithm succeeds because the transformation U_G acts as a *rotation* in the two-dimensional subspace defined by the states $|\psi\rangle$ and $|w\rangle$. The angle of the rotation effected by each application of U_G can be shown to be given by $\sin\theta = 2/\sqrt{N}$. Note that $|\psi\rangle$ and $|w\rangle$ are approximately orthogonal, $\langle\psi|w\rangle = 1/\sqrt{N}$, and that after the initial Hadamard transformation the system begins in the state $|\psi\rangle$. Each successive application of U_G moves it an angle θ closer to $|w\rangle$. Finally, after $\approx (\pi/4)\sqrt{N}$ iterations, the state has rotated the full $\approx \pi/2$ distance to $|w\rangle$.

Grover's algorithm can be shown to be optimal [82]: no black-box algorithm can find $|w\rangle$ with fewer than $O(\sqrt{N})$ iterations of the oracle. The algorithm also works for oracles where there are M winners, so that

$f(x) = 1$ for M distinct inputs. In this case, the angle of rotation for each iteration of U_G is given by $\sin\theta = (2/N)\sqrt{M(N-M)}$, and the algorithm takes $\approx (\pi/4)\sqrt{N/M}$ steps to identify a winner.

The Adiabatic Algorithm

Many classically hard problems take the form of optimization problems. In the well-known traveling salesman problem, for example, one aims to find the shortest route connecting a set of cities. Such optimization problems can be mapped onto a physical system, in which the function to be optimized is mapped onto the energy function of the system. The ground state of the physical system then represents a solution to the optimization problem. A common classical technique for solving such problems is simulated annealing: one simulates the process of gradually cooling the system in order to find its ground state [83]. Simulated annealing is bedeviled by the problem of local minima, states of the system that are close to the optimal states in terms of energy, but very far away in terms of the particular configuration of the degrees of freedom of the state. To avoid getting stuck in such local minima, one must slow the cooling process to a glacial pace in order to insure that the true ground state is reached in the end.

Quantum computing provides a method for getting around the problem of local minima. Rather than trying to reach the ground state of the system by cooling, one uses a purely quantum-mechanical technique for finding the state [84]. One starts the system with a Hamiltonian dynamics whose ground state is simple to prepare (e. g., 'all spins sideways'). Then one gradually deforms the Hamiltonian from the simple dynamics to the more complex dynamics whose ground state encodes the answer to the problem in question. If the deformation is sufficiently gradual, then the adiabatic theorem of quantum mechanics guarantees that the system remains in its ground state throughout the deformation process. When the adiabatic deformation is complete, then the state of the system can be measured to reveal the answer.

Adiabatic quantum computation (also called 'quantum annealing') represents a purely quantum way to find the answer to hard problems. How powerful is adiabatic quantum computation? The answer is, 'nobody knows for sure'. The key question is, what is 'sufficiently gradual' deformation? That is, how slowly does the deformation have to be to guarantee that the transformation is adiabatic? The answer to this question lies deep in the heart of quantum matter. As one performs the transformation from simple to complex dynamics, the adiabatic quantum computer goes through a quantum phase transition. The maximum speed at which the computation can be performed is governed by the size of the minimum energy gap of this quantum phase transition. The smaller the gap, the slower the computation. The scaling of gaps during phase transitions ('Gapology') is one of the key disciplines in the study of quantum matter [85]. While the scaling of the gap has been calculated for many familiar quantum systems such as Ising spin glasses, calculating the gap for adiabatic quantum computers that are solving hard optimization problems seems to be just about as hard as solving the problem itself.

While few quantum computer scientists believe that adiabatic quantum computation can solve the traveling salesman problem, there is good reason to believe that adiabatic quantum computation can outperform simulated annealing on a wide variety of hard optimization problems. In addition, it is known that adiabatic quantum computation is neither more nor less powerful than quantum computation itself: a quantum computer can simulate a physical system undergoing adiabatic time evolution using the quantum simulation techniques described above; in addition, it is possible to construct devices that perform conventional quantum computation in an adiabatic fashion [86].

Quantum Walks

A final, 'physics-based', type of algorithm is the quantum walk [87,88,89,90]. Quantum walks are coherent versions of classical random walks. A classical random walk is a stochastic Markov process, the random walker steps between different states, labeled by j, with a probability w_{ij} for making the transition from state j to state i. Here w_{ij} is a stochastic matrix, $w_{ij} \geq 0$ and $\sum_j w_{ij} = 1$. In a quantum walk, the stochastic, classical process is replaced by a coherent, quantum process: the states $|j\rangle$ are quantum states, and the transition matrix U_{ij} is unitary.

By exploiting quantum coherence, quantum walks can be shown typically to give a square root speed up over classical random walks. For example, in propagation along a line, a classical random walk is purely diffusive, with the expectation value of displacement along the line going as the square root of the number of steps in the walk. By contrast, a quantum walk can be set up as a coherent, propagating wave, so that the expectation value of the displacement is proportional to the number of steps [88]. A particularly elegant example of a square root speed up in a quantum walk is the evaluation of a *NAND* tree [90]. A *NAND* tree is a binary tree containing a *NAND* gate at each vertex. Given inputs on the leaves of the tree, the problem is to evaluate the outcome at the root of the tree: is it zero

or one? *NAND* trees are ubiquitous in, e. g., game theory: the question of who wins at chess, checkers, or Go, is determined by evaluating a suitable *NAND* tree. Classically, a *NAND* tree can be evaluated with a minimum of steps. A quantum walk, by contrast, can evaluate a *NAND* tree using only $2^{n/2}$ steps.

For some specially designed problems, such as propagation along a random tree, quantum walks can give exponential speedups over classical walks [89]. The question of what problems can be evaluated more rapidly using quantum walks than classical walks remains open.

The Future of Quantum Algorithms

The quantum algorithms described above are potentially powerful, and, if large-scale quantum computers can be constructed, could be used to solve a number of important problems for which no efficient classical algorithms exist. Many questions concerning quantum algorithms remain open. While the majority of quantum computer scientists would agree that quantum algorithms are unlikely to provide solutions to NP-complete problems, it is not known whether or not quantum algorithms could provide solutions to such problems as graph isomorphism or shortest vector on a lattice. Such questions are an active field of research in quantum computer science.

Noise and Errors

The picture of quantum computation given in the previous section is an idealized picture that does not take into account the problems that arise when quantum computers are built in practice. Quantum computers can be built using nuclear magnetic resonance, ion traps, trapped atoms in cavities, linear optics with feedback of nonlinear measurements, superconducting systems, quantum dots, electrons on the surface of liquid helium, and a variety of other standard and exotic techniques. Any system that can be controlled in a coherent fashion is a candidate for quantum computation. Whether a coherently controllable system can actually be made to computer depends primarily on whether it is possible to deal effectively with the noise intrinsic to that system. Noise induces errors in computation. Every type of quantum information processor is subject to its own particular form of noise.

A detailed discussion of the various technologies for building quantum computers lies beyond the scope of this article. While the types of noise differ from quantum technology to quantum technology, however, the methods for dealing with that noise are common between technologies. This section presents a general formalism for characterizing noise and errors, and discusses the use of quantum error-correcting codes and other techniques for coping with those errors.

Open-System Operations

The time evolution of a closed quantum-mechanical system is given by unitary transformation: $\rho \to U\rho U^\dagger$, where U is unitary, $U^\dagger = U^{-1}$. For discussing quantum communications, it is necessary to look at the time evolution of open quantum systems that can exchange quantum information with their environment. The discussion of open quantum systems is straightforward: simply adjoin the system's environment, and consider the coupled system and environment as a closed quantum system. If the joint density matrix for system and environment is

$$\rho_{\mathrm{SE}}(0) \to \rho_{\mathrm{SE}}(t) = U_{\mathrm{SE}}\rho_{\mathrm{SE}}(0)U_{\mathrm{SE}}^\dagger \,. \tag{47}$$

The state of the system on its own is obtained by taking the partial trace over the environment, as described above: $\rho_{\mathrm{S}}(t) = \mathrm{tr}_{\mathrm{E}}\rho_{\mathrm{SE}}(t)$.

A particularly useful case of system and environmental interaction is one in which the system and environment are initially uncorrelated, so that $\rho_{\mathrm{SE}}(0) = \rho_{\mathrm{S}}(0) \otimes \rho_{\mathrm{E}}(0)$. In this case, the time evolution of the system on its own can always be written as $\rho_{\mathrm{S}}(t) = \sum_k A_k\rho_{\mathrm{S}}(0)A_k^\dagger$. Here the A_k are operators that satisfy the equation $\sum_k A_k^\dagger A_k = Id$: the A_k are called Kraus operators, or effects. Such a time evolution for the system on its own is called a completely positive map. A simple example of such a completely positive map for a qubit is $A_0 = Id/\sqrt{2}$, $A_1 = \sigma_x/\sqrt{2}$. $\{A_0, A_1\}$ can easily be seen to obey $A_0^\dagger A_0 + A_1^\dagger A_1 = Id$. This completely positive map for the qubit corresponds to a time evolution in which the qubit has a 50% chance of being flipped about the x-axis (the effect A_1), and a 50% chance of remaining unchanged (the effect A_0).

The infinitesimal version of any completely positive map can be obtained by taking $\rho_{\mathrm{SE}}(0) = \rho_{\mathrm{S}}(0) \otimes \rho_{\mathrm{E}}(0)$, and by expanding (46) to second order in t. The result is the Lindblad master equation:

$$\frac{\partial\rho_{\mathrm{S}}}{\partial t} = -i[\tilde{H}_{\mathrm{S}}, \rho_{\mathrm{S}}] - \sum_k (L_k^\dagger L_k\rho_{\mathrm{S}} - 2L_k\rho_{\mathrm{S}}L_k + \rho_{\mathrm{S}}L_k^\dagger L_k)\,. \tag{48}$$

Here $\tilde{H}_{\mathrm{S}}$ is the effective system Hamiltonian: it is equal to the Hamiltonian H_{S} for the system on its own, plus a perturbation induced by the interaction with the environment (the so-called 'Lamb shift'). The L_k correspond to open system effects such as noise and errors.

Quantum Error-Correcting Codes

One of the primary effects of the environment on quantum information is to cause errors. Such errors can be corrected using quantum error-correcting codes. Quantum error-correcting codes are quantum analogs of classical error-correcting codes such as Hamming codes or Reed–Solomon codes [91]. Like classical error-correcting codes, quantum error-correcting codes involve first encoding quantum information in a redundant fashion; the redundant quantum information is then subjected to noise and errors; then the code is decoded, at which point the information needed to correct the errors lie in the code's syndrome.

More bad things can happen to quantum information than to classical information. The only error that can occur to a classical bit is a bit-flip. By contrast, a quantum bit can either be flipped about the x-axis (the effect σ_x), flipped about the y-axis (the effect σ_y), flipped about the z-axis (the effect σ_z), or some combination of these effects. Indeed, an error on a quantum bit could take the form of a rotation by an unknown angle θ about an unknown axis. Since specifying that angle and axis precisely could take an infinite number of bits of information, it might at first seem impossible to detect and correct such an error.

In 1996, however, Peter Shor [92] and Andrew Steane [93] independently realized that if an error correcting code could detect and correct bit-flip errors (σ_x) and phase-flip errors (σ_z), then such a code would in fact correct any single-qubit error. The reasoning is as follows. First, since $\sigma_y = i\sigma_x\sigma_z$, a code that detects and corrects first σ_x errors, then σ_z errors will also correct σ_y errors. Second, since any single-qubit rotation can be written as a combination of σ_x, σ_y and σ_z rotations, the code will correct arbitrary single qubit errors. The generalization of such quantum error-correcting codes to multiple qubit errors are called Calderbank–Shor–Steane (CSS) codes [94]. A powerful technique for identifying and characterizing quantum codes is Gottesman's stabilizer formalism [95].

Concatenation is a useful method for constructing codes, both classical and quantum. Concatenation combines two codes, with the second code acting on bits that have been encoded using the first code. Quantum error-correcting codes can be combined with quantum computation to perform fault-tolerant quantum computation. Fault-tolerant quantum computation allows quantum computation to be performed accurately even in the presence of noise and errors, as long as those errors occur at a rate below some threshold [96,97,98]. For restricted error models [99], this rate can be as high as $1\% - 3\%$. For realistic error models, however, the rate is closer to $10^{-3} - 10^{-4}$.

Re-focusing

Quantum error-correcting codes are not the only technique available for dealing with noise. If, as is frequently the case, environmentally induced noise possesses some identifiable structure in terms of correlations in space and time, or obeys some set of symmetries, then powerful techniques come into play for coping with noise.

First of all, suppose that noise is correlated in time. The simplest such correlation is a static imperfection: the Hamiltonian of the system is supposed to be H, but the actual Hamiltonian is $H + \Delta H$, where ΔH is some unknown perturbation. For example, an electron spin could have the Hamiltonian $H = -(\hbar/2)(\omega + \Delta\omega)\sigma_z$, where $\Delta\omega$ is an unknown frequency shift. If not attended to, such a frequency shift will introduce unknown phases in a quantum computation, which will in turn cause errors.

Such an unknown perturbation can be dealt with quite effectively simply by flipping the electron back and forth. Let the electron evolve for time T; flip it about the x-axis; let it evolve for time T; finally, flip it back about the x-axis. The total time-evolution operator for the system is then

$$\sigma_x e^{i(\omega+\Delta\omega)T\sigma_z}\sigma_x e^{i(\omega+\Delta\omega)T\sigma_z} = Id\,. \tag{49}$$

That is, this simple refocusing technique cancels out the effect of the unknown frequency shift, along with the time evolution of the unperturbed Hamiltonian.

Even if the environmental perturbation varies in time, refocusing can be used significantly to reduce the effects of such noise. For time-varying noise, refocusing effectively acts as a filter, suppressing the effects of noise with a correlation time longer than the refocusing timescale T. More elaborate refocusing techniques can be used to cope with the effect of couplings between qubits. Refocusing requires no additional qubits or syndromes, and so is a simpler (and typically much more effective) technique for dealing with errors than quantum error-correcting codes. For existing experimental systems, refocusing typically makes up the 'first line of defence' against environmental noise. Once refocusing has dealt with time-correlated noise, quantum error correction can then be used to deal with any residual noise and errors.

Decoherence-free Subspaces and Noiseless Subsystems

If the noise has correlations in space, then quantum information can often be encoded in such a way as to be

resistant to the noise even in the absence of active error correction. A common version of such spatial correlation occurs when each qubit is subjected to the *same* error. For example, suppose that two qubits are subjected to noise of the form of a fluctuating Hamiltonian $H(t) = (\hbar/2)\gamma(t)(\sigma_z^1 + \sigma_z^2)$. This Hamiltonian introduces a time-varying phase $\gamma(t)$ between the states $|\uparrow\rangle_i, |\downarrow\rangle_i$. The key point to note here is that this phase is the *same* for both qubits. A simple way to compensate for such a phase is to encode the logical state $|0\rangle$ as the two-qubit state $|\uparrow\rangle_1|\downarrow\rangle_2$, and the logical state $|1\rangle$ as the two-qubit state $|\downarrow\rangle_1|\uparrow\rangle_2$. It is simple to verify that the two-qubit encoded states are now invariant under the action of the noise: any phase acquired by the first qubit is canceled out by the equal and opposite phase acquired by the second qubit. The subspace spanned by the two-qubit states $|0\rangle, |1\rangle$ is called a decoherence-free subspace: it is invariant under the action of the noise.

Decoherence-free subspaces were first discovered by Zanardi [100] and later popularized by Lidar [101]. Such subspaces can be found essentially whenever the generators of the noise possess some symmetry. The general form that decoherence-free subspaces take arises from the following observation concerning the relationship between noise and symmetry.

Let $\{E_k\}$ be the effects that generate the noise, so that the noise takes $\rho \to \sum_k E_k \rho E_k^\dagger$, and let $\mathcal{E}$ be the algebra generated by the $\{E_k\}$. Let G be a symmetry of this algebra, so that $[g, E] = 0$ for all $g \in G, E \in \mathcal{E}$. The Hilbert space for the system then decomposes into irreducible representation of $\mathcal{E}$ and G in the following well-known way:

$$\mathcal{H} = \sum_j \mathcal{H}_{\mathrm{E}}^j \otimes \mathcal{H}_G^j \,, \tag{50}$$

where $\mathcal{H}_{\mathrm{E}}^j$ are the irreducible representations of $\mathcal{E}$, and $\mathcal{H}_G^j$ are the irreducible representations of G.

The decomposition (49) immediately suggests a simple way of encoding quantum information in a way that is immune to the effects of the noise. Look at the effect of the noise on states of the form $|\phi\rangle_j \otimes |\psi\rangle_j$ where $|\phi\rangle_j \in \mathcal{H}_{\mathrm{E}}^j$, and $|\psi\rangle_j \in \mathcal{H}_G^j$ for some j. The effect E_k acts on this state as $(E_k^j|\phi\rangle_j) \otimes |\psi\rangle_j$, where $E_k{}^j$ is the effect corresponding to E_k within the representation $\mathcal{H}_{\mathrm{E}}^j$. In other words, if we encode quantum information in the state $|\psi\rangle_j$, then the noise has *no effect* on $|\psi\rangle_j$. A decoherence-free subspace corresponds to an $\mathcal{H}_G^j$ where the corresponding representation of $\mathcal{E}$, $\mathcal{H}_{\mathrm{E}}^j$, is one-dimensional. The case where $\mathcal{H}_{\mathrm{E}}^j$, is higher dimensional is called a noiseless subsystem [102].

Decoherence-free subspaces and noiseless subsystems represent highly effective methods for dealing with the presence of noise. Like refocusing, these methods exploit symmetry to encode quantum information in a form that is immune to noise that possesses that symmetry. Where refocusing exploits temporal symmetry, decoherence-free subspaces and noiseless subsystems exploit spatial symmetry. All such symmetry-based techniques have the advantage that no error-correcting process is required. Like refocusing, therefore, decoherence-free subspaces and noiseless subsystems form the first line of defense against noise and errors.

The tensor product decomposition of irreducible representations in (49) lies behind all known error-correcting codes [103]. A general quantum-error correcting code begins with a state $|00\ldots0\rangle_{\mathrm{A}}|\psi\rangle$, where $|00\ldots0\rangle_{\mathrm{A}}$ is the initial state of the ancilla. An encoding transformation U_{en} is then applied; an error E_k occurs; finally a decoding transformation U_{de} is applied to obtain the state

$$|\mathrm{e_k}\rangle_{\mathrm{A}}|\psi\rangle = U_{\mathrm{de}} E_k U_{\mathrm{en}} |00\ldots0\rangle_{\mathrm{A}}|\psi\rangle \,. \tag{51}$$

Here, $|e_k\rangle_{\mathrm{A}}$ is the state of the ancilla that tells us that the error corresponding to the effect E_k has occurred. Equation. (50) shows that an error-correcting code is just a noiseless subsystem for the 'dressed errors' $\{U_{\mathrm{de}} E_k U_{\mathrm{en}}\}$. At bottom, all quantum error-correcting codes are based on symmetry.

Topological Quantum Computing

A particularly interesting form of quantum error correction arises when the underlying symmetry is a topological one. Kitaev [104] has shown how quantum computation can be embedded in a topological context. Two-dimensional systems with the proper symmetries exhibit topological excitations called anyons. The name, 'anyon', comes from the properties of these excitations under exchange. Bosons, when exchanged, obtain a phase of 1; fermions, when exchanged, obtain a phase of -1. Anyons, by contrast, when exchanged, can obtain an arbitrary phase $e^{i\phi}$. For example, the anyons that underlie the fractional quantum Hall effect obtain a phase $e^{2\pi i/3}$ when exchanged. Fractional quantum Hall anyons can be used for quantum computation in a way that makes two-qubit quantum logic gates intrinsically resistant to noise [105].

The most interesting topological effects in quantum computation arise when one employs non-abelian anyons [104]. Non-abelian anyons are topological excitations that possess internal degrees of freedom. When two non-abelian anyons are exchanged, those internal degrees of freedom are subjected not merely to an additional

phase, but to a general unitary transformation U. Kitaev has shown how in systems with the proper symmetries, quantum computation can be effected simply by exchanging anyons. The actual computation takes place by dragging anyons around each other in the two-dimensional space. The resulting transformation can be visualized as a braid in two dimensional space plus the additional dimension of time.

Topological quantum computation is intrinsically fault tolerant. The topological excitations that carry quantum information are impervious to locally occurring noise: only a global transformation that changes the topology of the full system can create a error. Because of their potential for fault tolerance, two-dimensional systems that possess the exotic symmetries required for topological quantum computation are being actively sought out.

Quantum Communication

Quantum mechanics provides the fundamental limits to information processing. Above, quantum limits to computation were investigated. Quantum mechanics also provides the fundamental limits to communication. This section discusses those limits. The session closes with a section on quantum cryptography, a set of techniques by which quantum mechanics guarantees the privacy and security of cryptographic protocols.

Multiple Uses of Channels

Each quantum communication channel is characterized by its own open-system dynamics. Quantum communication channels can possess memory, or be memoryless, depending on their interaction with their environment. Quantum channels with memory are a difficult topic, which will be discussed briefly below. Most of the discussion that follows concerns the memoryless quantum channel. A single use of such a channel corresponds to a completely positive map, $\rho \to \sum_k A_k \rho A_k^\dagger$, and n uses of the channel corresponds to a transformation

$$\begin{aligned}\rho_{1\ldots n} &\to \sum_{k_1\ldots k_n} A_{k_n} \otimes \cdots \otimes A_{k_1} \rho_{1\ldots n} A_{k_1}^\dagger \otimes \cdots \otimes A_{k_n}^\dagger \\ &\equiv \sum_K A_K \rho_{1\ldots n} A_K^\dagger \,,\end{aligned} \tag{52}$$

where we have used the capital letter K to indicate the n uses of the channel $k_1 \ldots k_n$. In general, the input state $\rho_{1\ldots n}$ may be entangled from use to use of the channel. Many outstanding questions in quantum communication theory remain unsolved, including, for example, the question of whether entangling inputs of the channel helps for communicating classical information.

Sending Quantum Information

Let's begin with using quantum channels to send quantum information. That is, we wish to send some quantum state $|\psi\rangle$ from the input of the channel to the output. To do this, we encode the state as some state of n inputs to the channel, send the encoded state down the channel, and then apply a decoding procedure at the output to the channel. It is immediately seen that such a procedure is equivalent to employing a quantum error-correcting code.

The general formula for the capacity of such quantum channels is known [44,45]. Take some input or 'signal' state $\rho_{1\ldots n}$ for the channel. First, construct a purification of this state. A purification of a density matrix ρ for the signal is a pure state $|\psi\rangle_{\mathrm{AS}}$ for the signal together with an ancilla, such that the state $\rho_{\mathrm{S}} = \mathrm{tr}_{\mathrm{A}} |\psi\rangle_{\mathrm{AS}}\langle\psi|$ is equal to the original density matrix ρ. There are many different ways to purify a state: a simple, explicit way is to write $\rho = \sum_j p_j |j\rangle\langle j|$ in diagonal form, where $\{|j\rangle\}$ is the eigenbasis for ρ. The state $|\psi\rangle_{\mathrm{AS}} = \sum_j \sqrt{p_j}|j\rangle_{\mathrm{A}}|j\rangle_{\mathrm{S}}$, where $\{|j\rangle_{\mathrm{A}}\}$ is an orthonormal set of states for the ancilla, then yields a purification of ρ.

To obtain the capacity of the channel for sending quantum information, proceed as follows. Construct a purification for the signal $\rho_{1\ldots n}$: $|\psi_n\rangle = \sum_J \sqrt{p_J}|J\rangle_{\mathrm{A}}^n |J\rangle_{\mathrm{S}}^n$, where we have used an index J instead of j to indicate that these states are summed over n uses of the channel. Now send the signal state down the channel, yielding the state

$$\rho_{\mathrm{AS}} = \sum_{JJ'} \sqrt{p_J p_{J'}} |J\rangle_{\mathrm{A}}^n \langle J'| \otimes \sum_K A_K |J\rangle_{\mathrm{S}}^n \langle J'| A_K^\dagger \,, \tag{53}$$

where as above $K = k_1 \ldots k_n$ indicates k uses of the channel. ρ_{AS} is the state of output signal state together with the ancilla. Similarly, $\rho_{\mathrm{S}} = \mathrm{tr}_{\mathrm{A}} \rho_{\mathrm{AS}}$ is the state of the output signal state on its own.

Let $I(AS) = -\mathrm{tr}\rho_{\mathrm{AS}} \log_2 \rho_{\mathrm{AS}}$ be the entropy of ρ_{AS}, measured in bits. Similarly, let $I(S) = -\mathrm{tr}\rho_{\mathrm{S}} \log_2 \rho_{\mathrm{S}}$ be the entropy of the output state ρ_{S}, taken on its own. Define $I(S/A) \equiv I(S) - I(AS)$ if this quantity is positive, and $I(S/A) \equiv 0$ otherwise. The quantity $I(S/A)$ is a measure of the capacity of the channel to send quantum information if the signals being sent down the channel are described by the density matrix $\rho_{1\ldots n}$. It can be shown using either CSS codes [106] or random codes [45,107] that encodings exist that allow quantum information to be sent down the channel and properly decoded at the output at a rate of $I(S/A)/n$ qubits per use.

$I(S/A)$ is a function only of the properties of the channel and the input signal state $\rho_{1\ldots n}$. The bigger $I(S/A)$ is, the less coherence the channel has managed to destroy. For example, if the channel is just a unitary transformation of the input, which destroys no quantum information, then $I(AS) = 0$ and $I(S/A) = I(S)$: the state of the signal and ancilla after the signal has passed through the channel is pure, and all quantum information passes down the channel unscathed. By contrast, a completely decohering channel takes an the input $\sum_j \sqrt{p_j}|j\rangle_A|j\rangle_S$ to the output $\sum_j p_j|j\rangle_A\langle j| \otimes |j\rangle_S\langle j|$. In this case, $I(AS) = I(S)$ and $I(S/A) = 0$: the channel has completely destroyed all quantum information sent down the channel.

In order to find the absolute capacity of the channel to transmit quantum information, we must maximize the quantity $I(S/A)/n$ over all n-state inputs $\rho_{1\ldots n}$ to the channel and take the limit as $n \to \infty$. More precisely, define

$$I_C = \lim_{n\to\infty} \min \sup I(S/A)/n \,, \tag{54}$$

where the supremum (sup) is taken over all n-state inputs $\rho_{1\ldots n}$. I_C is called the coherent information [44,45]: it is the capacity of the channel to transmit quantum information reliably.

Because the coherent information is defined only in the limit that the length of the input state goes to infinity, it has been calculated exactly in only a few cases. One might hope, in analogue to Shannon's theory of classical communication, that for memoryless channels one need only optimize over single inputs. That hope is mistaken, however: entangling the input states typically increases the quantum channel capacity even for memoryless channels [108].

Capacity of Quantum Channels to Transmit Classical Information

One of the most important questions in quantum communications is the capacity of quantum channels to transmit classical information. All of our classical communication channels – voice, free space electromagnetic, fiber optic, etc. – are at bottom quantum mechanical, and their capacities are set using the laws of quantum mechanics. If quantum information theory can discover those limits, and devise ways of attaining them, it will have done humanity a considerable service.

The general picture of classical communication using quantum channels is as follows. The conventional discussion of communication channels, both quantum and classical, designates the sender of information as Alice, and the receiver of information as Bob. Alice selects an ensemble of input states ρ_J over n uses of the channel, and send the Jth input ρ_J with probability p_J. The channel takes the n-state input ρ_J to the output $\tilde{\rho}_J = \sum_K A_K \rho_J A_K^\dagger$. Bob then performs a generalized measurement $\{B_\ell\}$ with outcomes $\{\ell\}$ to try to reveal which state Alice sent. A generalized measurement is simply a specific form an open-system transformation. The $\{B_\ell\}$ are effects for a completely positive map: $\sum_\ell B_\ell^\dagger B_\ell = Id$. After making the generalized measurement on an output state $\tilde{\rho}_J$, Bob obtains the outcome ℓ with probability $p_{\ell|J} = \mathrm{tr} B_\ell \tilde{\rho}_J B_\ell^\dagger$, and the system is left in the state $(1/p_{\ell|J}) B_\ell \tilde{\rho}_J B_\ell^\dagger$.

Once Alice has chosen a particular ensemble of signal states $\{\rho_J, p_J\}$, and Bob has chosen a particular generalized measurement, then the amount of information that can be sent along the channel is determined by the input probabilities p_J and the output probabilities $p_{\ell|J}$ and $p_\ell = \sum_J p_J p_{\ell|J}$. In particular, the rate at which information can be sent through the channel and reliably decoded at the output is given by the mutual information $I(\text{in} : \text{out}) = I(\text{out}) - I(\text{out}|\text{in})$, where $I(\text{out}) = -\sum_\ell p_\ell \log_2 p_\ell$ is the entropy of the output and $I(\text{out}|\text{in}) = \sum_J p_J(-\sum_\ell p_{\ell|J} \log_2 p_{\ell|J})$ is the average entropy of the output conditioned on the state of the input.

To maximize the amount of information that can be sent down the channel, Alice and Bob need to maximize over both input states and over Bob's measurement at the output. The Schumacher–Holevo–Westmoreland theorem, however, considerably simplifies the problem of maximizing the information transmission rate of the channel by obviating the need to maximize over Bob's measurements [37,38,39]. Define the quantity

$$\chi = S\left(\sum_J p_J \tilde{\rho}_J\right) - \sum_J p_J S(\tilde{\rho}_J) \,, \tag{55}$$

where $S(\rho) \equiv -\mathrm{tr}\rho \log_2 \rho$. χ is the difference between the entropy of the average output state and the average entropy of the output states. The Schumacher–Holevo–Westmoreland theorem then states that the capacity of the quantum channel for transmitting classical information is given by the limit as $\lim_{n\to\infty} \min \sup \chi/n$, where the supremum is taken over all possible ensembles of input states $\{\rho_J, p_J\}$ over n uses of the channel.

For Bob to attain the channel capacity given by χ, he must in general make entangling measurements over the channel outputs, even when the channel is memoryless and when Alice does not entangle her inputs. (An entangling measurement is one the leaves the outputs in an entangled state after the measurement is made.) It would simplify the process of finding the channel capacity still further if the optimization over input states could be performed over a single use of the channel for memoryless

channels, as is the case for classical communication channels, rather than having to take the limit as the number of inputs goes to infinity. If this were the case, then the channel capacity for memoryless channels would be attained for Alice sending unentangled states down the channel. Whether or not one is allowed to optimize over a single use for memoryless channels was for many years one of the primary unsolved conjectures of quantum information theory.

Let's state this conjecture precisely. Let χ_n be the maximum of χ over n uses of a memoryless channel. We then have the

Channel Additivity Conjecture: $\chi_n = n\chi_1$ Shor showed that the channel additivity conjecture is equivalent to two other additivity conjectures, the additivity of minimum output entropy and the additivity of entanglement of formation [109]. Entanglement of formation was discussed in the section on entanglement above. The minimum output entropy for n uses of a memoryless channel is simply the minimum over input states ρ_n, for n uses of the channel, of $S(\tilde{\rho}_n)$, where $\tilde{\rho}_n$ is the output state arising from the input ρ_n. We then have the

Minimum Output Entropy Additivity Conjecture: The minimum over ρ_n of $S(\tilde{\rho}_n)$ is equal to n times the minimum over ρ_1 of $S(\tilde{\rho}_1)$.

Shor's result shows that the channel additivity conjecture and the minimum output entropy additivity conjecture are equivalent: each one implies the other. If these additivity conjectures could have been proved to be true, that would have resolved some of the primary outstanding problems in quantum channel capacity theory. Remarkably, however, Hastings recently showed that the minimum output entropy conjecture is *false*, by exhibiting a channel whose minimum output entropy for multiple uses is achieved for entangled inputs [110]. As a result, the question of just how much classical information can be sent down a quantum channel, and just which quantum channels are additive and which are not, remains wide open.

Bosonic Channels

The most commonly used quantum communication channel is the so-called bosonic channel with Gaussian noise and loss [40]. Bosonic channels are ones that use bosons such as photons or phonons to communicate. Gaussian noise and loss is the most common type of noise and loss for such channels, it includes the effect of thermal noise, noise from linear amplification, and leakage of photons or phonons out of the channel. It has been shown that the capacity for bosonic channels with loss alone is attained by sending coherent states down the channel [42]. Coherent states are the sort of states produced by lasers and are the states that are currently used in most bosonic channels.

It has been conjectured that coherent states also maximize the capacity of quantum communication channels with Gaussian noise as well as loss [43]. This conjecture, if true, would establish the quantum-mechanical equivalent of Shannon's theorem for the capacity of classical channels with Gaussian noise and loss. The resolution of this conjecture can be shown to be equivalent to the following, simpler conjecture:

Gaussian Minimum Output Entropy Conjecture: Coherent states minimize the output entropy of bosonic channels with Gaussian noise and no loss.

The Gaussian minimum output entropy is intuitively appealing: an equivalent statement is that the *vacuum* input state minimizes the output entropy for a channel with Gaussian noise. In other words, to minimize the output entropy of the channel, send *nothing*.

Despite its intuitive appeal, the Gaussian minimum output entropy conjecture has steadfastly resisted proof for decades. This conjecture is related to the additivity conjectures above: in particular, if the additivity conjectures can be shown to be true, then the Gaussian minimum output entropy conjecture is also true [110]. It is not known whether the converse implication also holds. Proving the additivity conjectures and the Gaussian minimum output entropy conjecture is one of the primary goals of quantum information theory.

Entanglement Assisted Capacity

Just as quantum bits possess greater mathematical structure than classical bits, so quantum channels possess greater variety than their classical counterparts. A classical channel has but a single capacity. A quantum channel has one capacity for transmitting quantum information (the coherent information), and another capacity for transmitting classical information (the Holevo quantity χ). We can also ask about the capacity of a quantum channel in the presence of prior entanglement.

The entanglement assisted capacity of a channel arises in the following situation. Suppose that Alice and Bob have used their quantum channel to build up a supply of entangled qubits, where Alice possesses half of the entangled pairs of qubits, and Bob possesses the other half of the pairs. Now Alice sends Bob some qubits over the channel. How much classical information can these qubits convey?

At first one might think that the existence of shared prior entanglement should have no effect on the amount of information that Alice can send to Bob. After all, entanglement is a form of correlation, and the existence of prior correlation between Alice and Bob in a classical setting has no effect on the amount of information sent. In the quantum setting, however, the situation is different.

Consider, for example, the case where Alice and Bob have a perfect, noiseless channel. When Alice and Bob share no prior entanglement, then a single qubit sent down the channel conveys exactly one bit of classical information. When Alice and Bob share prior entanglement, however, a single quantum bit can convey more than one bit of classical information. Suppose that Alice and Bob share an entangled pair in the singlet state $(1/\sqrt{2})(|0\rangle_A|1\rangle_B - |1\rangle_A|0\rangle_B)$. Alice then performs one of four actions on her qubit: either she does nothing (performs the identity Id on the qubit), or she flips the qubit around the x-axis (performs σ_x), or she flips the qubit around the y-axis (performs σ_y), she flips the qubit around the z-axis (performs σ_z).

Now Alice sends her qubit to Bob. Bob now possesses one of the four orthogonal states, $(1/\sqrt{2}) \cdot (|0\rangle_A|1\rangle_B - |1\rangle_A|0\rangle_B)$, $(1/\sqrt{2})(|1\rangle_A|1\rangle_B - |0\rangle_A|0\rangle_B)$, $(i/\sqrt{2}) \cdot (|1\rangle_A|1\rangle_B + |0\rangle_A|0\rangle_B)$, $(1/\sqrt{2})(|0\rangle_A|1\rangle_B + |1\rangle_A|0\rangle_B)$. By measuring which of these states he possesses, Bob can determine which of the four actions Alice performed. That is, when Alice and Bob share prior entanglement, Alice can send *two* classical bits for each quantum bit she sends. This phenomenon is known as superdense coding [111].

In general, the quantum channel connecting Alice to Bob is noisy. We can then ask, given the form of the quantum channel, how much does the existence of prior entanglement help Alice in sending classical information to Bob? The answer to this question is given by the following theorem, due to Shor et al. The entanglement assisted capacity of a quantum channel is equal to the maximum of the quantum mutual information between the input and output of the channel [112]. The quantum mutual information is defined as follows. Prepare a purification $|\psi\rangle_{AS}$ of an input state ρ and send the signal state S down the channel, resulting the state ρ_{AS} as in (52) above. Defining $\rho_S = \mathrm{tr}_A \rho_{AS}$, $\rho_A = \mathrm{tr}_S \rho_{AS}$, as before, the quantum mutual information is defined to be $I_Q(A:S) = S(\rho_A) + S(\rho_S) - S(\rho_{AS})$. The entanglement assisted capacity of the channel is obtained by maximizing the quantum mutual information $I_Q(A:S)$ over input states ρ.

The entanglement assisted capacity of a quantum channel is greater than or equal to the channel's Holevo quantity, which is in turn greater than or equal to the channel's coherent information. Unlike the coherent information, which is known not to be additive over many uses of the channel, or the Holevo quantity, which is suspected to be additive but which has not been proved to be so, the entanglement assisted capacity is known to be additive and so can readily be calculated for memoryless channels.

Teleportation

As mentioned in the introduction, one of the most strange and useful effects in quantum computation is teleportation [46]. The traditional, science fiction picture of teleportation works as follows.

An object such as an atom or a human being is placed in a device called a teleporter. The teleporter makes complete measurements of the physical state of the object, destroying it in the process. The detailed information about that physical state is sent to a distant location, where a second teleporter uses that information to reconstruct an exact copy of the original object.

At first, quantum mechanics would seem to make teleportation impossible. Quantum measurements tend to disturb the object measured. Many identical copies of the object are required to obtain even a rough picture of the underlying quantum state of the object. In the presence of shared, prior entanglement, however, teleportation is in fact possible in principle, and simple instances of teleportation have been demonstrated experimentally.

A hint to the possibility of teleportation comes from the phenomenon of superdense coding described in the previous section. If one qubit can be used to convey two classical bits using prior entanglement, then maybe two classical bits might be used to convey one qubit. This hope turns out to be true. Suppose that Alice and Bob each possess one qubit out of an entangled pair of qubits (that is, they mutually possess one 'e-bit'). Alice desires to teleport the state $|\psi\rangle$ of another qubit. The teleportation protocol goes as follows.

First, Alice makes a Bell-state measurement on the qubit to be teleported together with her half of the entangled pair. A Bell-state measurement on two qubits is one that determines whether the two qubits are in one of the four states $|\phi_{00}\rangle = (1/\sqrt{2})(|01\rangle - |10\rangle)$, $|\phi_{01}\rangle = (1/\sqrt{2})(|00\rangle - |11\rangle)$, $|\phi_{10}\rangle = (1/\sqrt{2})(|00\rangle + |11\rangle)$, or $|\phi_{11}\rangle = (1/\sqrt{2})(|01\rangle + |10\rangle)$. Alice obtains two classical bits of information as a result of her measurement, depending on which $|\phi_{ij}\rangle$ the measurement revealed. She sends these two bits to Bob. Bob now performs a unitary transformation on his half of the entangled qubit pair. If he receives 00, then he does nothing. If he receives 01, then he applies σ_x to flip his bit about the x-axis. If he receives

10, then he applies σ_y to flip his bit about the y-axis. If he receives 11, then he applies σ_z to flip his bit about the z-axis. The result? After Bob has performed his transformation conditioned on the two classical bits he received from Alice, his qubit is now in the state $|\psi\rangle$, up to an overall phase. Alice's state has been teleported to Bob.

It might seem at first somewhat mysterious how this sequence of operations can teleport Alice's state to Bob. The mechanism of teleportation can be elucidated as follows. Write $|\psi\rangle = \alpha|0\rangle_i + \beta|1\rangle_i$. Alice and Bob's entangled pair is originally in the state $|\phi_{00}\rangle_{\mathrm{AB}} = (1/\sqrt{2})(|0\rangle_{\mathrm{A}}|1\rangle_{\mathrm{B}} - |1\rangle_{\mathrm{A}}|0\rangle_{\mathrm{B}})$. The full initial state of qubit to be teleported together with the entangled pair can then be written as

$$\begin{aligned}
|\psi\rangle|\phi_{00}\rangle_{\mathrm{AB}} &= (\alpha|0\rangle_i + \beta|1\rangle_i)\frac{1}{\sqrt{2}}(|0\rangle_{\mathrm{A}}|1\rangle_{\mathrm{B}} - |1\rangle_{\mathrm{A}}|0\rangle_{\mathrm{B}}) \\
&= \frac{1}{2\sqrt{2}}(|0\rangle_i|1\rangle_{\mathrm{A}} - |1\rangle_i|0\rangle_{\mathrm{A}}) \otimes (\alpha|0\rangle_{\mathrm{B}} + \beta|1\rangle_{\mathrm{B}}) \\
&\quad + \frac{1}{2\sqrt{2}}(|0\rangle_i|0\rangle_{\mathrm{A}} - |1\rangle_i|1\rangle_{\mathrm{A}}) \otimes (\alpha|1\rangle_{\mathrm{B}} + \beta|0\rangle_{\mathrm{B}}) \\
&\quad + \frac{1}{2\sqrt{2}}(|0\rangle_i|0\rangle_{\mathrm{A}} + |1\rangle_i|1\rangle_{\mathrm{A}}) \otimes (\alpha|1\rangle_{\mathrm{B}} - \beta|0\rangle_{\mathrm{B}}) \\
&\quad + \frac{1}{2\sqrt{2}}(|0\rangle_i|1\rangle_{\mathrm{A}} + |1\rangle_i|0\rangle_{\mathrm{A}}) \otimes (\alpha|0\rangle_{\mathrm{B}} - \beta|1\rangle_{\mathrm{B}}) \\
&= \frac{1}{2}(|\phi_{00}\rangle_{i\mathrm{A}} \otimes |\psi\rangle_{\mathrm{B}} + |\phi_{01}\rangle_{i\mathrm{A}} \otimes \sigma_x|\psi\rangle_{\mathrm{B}} \\
&\quad + |\phi_{10}\rangle_{i\mathrm{A}} \otimes i\sigma_y|\psi\rangle_{\mathrm{B}} + |\phi_{11}\rangle_{i\mathrm{A}} \otimes \sigma_z|\psi\rangle_{\mathrm{B}}) \,.
\end{aligned} \tag{56}$$

When the initial state is written in this form, one sees immediately how the protocol works: the measurement that Alice makes contains exactly the right information that Bob needs to reproduce the state $|\psi\rangle$ by performing the appropriate transformation on his qubit.

Teleportation is a highly useful protocol that lies at the center of quantum communication and fault tolerant quantum computation. There are several interesting features to note. The two bits of information that Alice obtains are completely random: 00, 01, 10, 11 all occur with equal probability. These bits contain no information about $|\psi\rangle$ taken on their own: it is only when combined with Bob's qubit that those bits suffice to recreate $|\psi\rangle$. During the teleportation process, it is difficult to say just where the state $|\psi\rangle$ 'exists'. After Alice has made her measurement, the state $|\psi\rangle$ is in some sense 'spread out' between her two classical bits and Bob's qubit. The proliferation of quotation marks in this paragraph is a symptom of quantum weirdness: classical ways of describing things are inadequate to capture the behavior of quantum things. The only way to see what happens to a quantum system during a process like teleportation is to apply the mathematical rules of quantum mechanics.

Quantum Cryptography

A common problem in communication is security. Suppose that Alice and Bob wish to communicate with each other with the secure knowledge that no eavesdropper (Eve) is listening in. The study of secure communication is commonly called cryptography, since to attain security Alice must encrypt her messages and Bob must decrypt them. The no-cloning theorem together with the fact that if one measures a quantum system, one typically disturbs it, implies that quantum mechanics can play a unique role in constructing cryptosystems. There are a wide variety of quantum cryptographic protocols [49,50,51]. The most common of these fall under the heading of quantum key distribution (QKD).

The most secure form of classical cryptographic protocols is the one-time pad. Here, Alice and Bob each possess a random string of bits. This string is called the key. If no one else possesses the key, then Alice and Bob can send messages securely as follows. Suppose that Alice's message has been encoded in bits in some conventional way (e. g., mapping characters to ASCII bit strings). Alice encrypts the message by adding the bits of the key to the bits of her message one by one, modulo 2 (i. e., without carrying). Quantum information theory is a rich and fundamental field. Its origins lie with the origins of quantum mechanics itself a century ago. The field has expanded dramatically since the mid 1990s, due to the discovery of practical applications of quantum information processing such as factoring and quantum cryptography, and because of the rapid development of technologies for manipulating systems in a way that preserves quantum coherence.

As an example of the rapid pace of development in the field of quantum information, while this article was in proof, a new algorithm for solving linear sets of equations was discovered [116]. Based on the quantum phase algorithm, this algorithm solves the following problem: given a sparse matrix A and a vector $\vec{b}$, find a vector $\vec{x}$ such that $A\vec{x} = \vec{b}$. That is, construct $\vec{x} = A^{-1}\vec{b}$. If A is an n by n matrix, the best classical algorithms for solving this problem run in time $O(n)$. Remarkably, the quantum matrix inversion algorithm runs in time $O(\log n)$, an exponential improvement: a problem that could take 10^{12} – 10^{15} operations to solve on a classical computer could be solved on a quantum computer in fewer than one hundred steps.

When they were developed in the mid twentieth century, the fields of classical computation and communica-

tion provided unifying methods and themes for all of engineering and science. So at the beginning of the twenty first century, quantum information is providing unifying concepts such as entanglement, and unifying techniques such as coherent information processing and quantum error correction, that have the potential to transform and bind together currently disparate fields in science and engineering.

The idea of quantum cryptography was proposed, in embryonic form, by Stephen Wiesner in [49]. The first quantum cryptographic protocol was proposed by Bennett and Brassard in 1984 and is commonly called BB84 [50]. The BB84 protocol together with its variants is the one most commonly used by existing quantum cryptosystems.

In BB84, Alice sends Bob a sequence of qubits. The protocol is most commonly described in terms of qubits encoded on photon polarization. Here, we will describe the qubits in terms of spin, so that we can use the notation developed in Sect. "Quantum Mechanics". Spin 1/2 is isomorphic to photon polarization and so the quantum mechanics of the protocol remains the same.

Alice chooses a sequence of qubits from the set $\{|\uparrow\rangle, |\downarrow\rangle, |\leftarrow\rangle, |\rightarrow\rangle\}$ at random, and sends that sequence to Bob. As he receives each qubit in turn, Bob picks at random either the z-axis or the x-axis and measures the received qubit along that axis. Half of the time, on average, Bob measures the qubit along the same axis along which it was prepared by Alice.

Alice and Bob now check to see if Eve is listening in. Eve can intercept the qubits Alice sends, make a measurement on them, and then send them on to Bob. Because she does not know the axis along which any individual qubit has been prepared, however, here measurement will inevitably disturb the qubits. Alice and Bob can then detect Eve's intervention by the following protocol.

Using an ordinary, insecure form of transmission, e. g., the telephone, Alice reveals to Bob the state of some of the qubits that she sent. On half of those qubits, on average, Bob measured them along the same axis along which they were sent. Bob then checks to see if he measured those qubits to be in the same state that Alice sent them. If he finds them all to be in the proper state, then he and Alice can be sure that Eve is not listening in. If Bob finds that some fraction of the qubits are not in their proper state, then he and Alice know that either the qubits have been corrupted by the environment in transit, or Eve is listening in. The degree of corruption is related to the amount of information that Eve can have obtained: the greater the corruption, the more information Eve may have. From monitoring the degree of corruption of the received qubits, Alice and Bob can determine just how many bits of information Eve has obtained about their transmission.

Alice now reveals to Bob the axis along which she prepared the remainder of her qubits. On half of those, on average, Bob measured using the same axis. If Eve is not listening, those qubits on which Bob measured using the same axis along which Eve prepared them now constitute a string of random bits that is shared by Alice and Bob and by them only. This shared random string can then be used as a key for a one-time pad.

If Eve is listening in, then from their checking stage, Alice and Bob know just how many bits out of their shared random string are also known by Eve. Alice and Bob can now perform classical privacy amplification protocols [113] to turn their somewhat insecure string of shared bits into a shorter string of shared bits that is more secure. Once privacy amplification has been performed, Alice and Bob now share a key whose secrecy is guaranteed by the laws of quantum mechanics.

Eve could, of course, intercept *all* the bits sent, measure them, and send them on. Such a 'denial of service' attack prevents Alice and Bob from establishing a shared secret key. No cryptographic system, not even a quantum one, is immune to denial of service attacks: if Alice and Bob can exchange no information then they can exchange no secret information! If Eve lets enough information through, however, then Alice and Bob can always establish a secret key.

A variety of quantum key distribution schemes have been proposed [50,51]. Ekert suggested using entangled photons to distribute keys to Alice and Bob. In practical quantum key distribution schemes, the states sent are attenuated coherent states, consisting of mostly vacuum with a small amplitude of single photon states, and an even smaller amplitude of states with more than one photon. It is also possible to use continuous quantum variables such as the amplitudes of the electric and magnetic fields to distribute quantum keys [114,115]. To guarantee the full security of a quantum key distribution scheme requires a careful examination of all possible attacks given the actual physical implementation of the scheme.

Implications and Conclusions

Quantum information theory is a rich and fundamental field. Its origins lie with the origins of quantum mechanics itself a century ago. The field has expanded dramatically since the mid 1990s, due to the discovery of practical applications of quantum information processing such as factoring and quantum cryptography, and because of the rapid development of technologies for manipulating systems in

a way that preserves quantum coherence. When they were developed in the mid twentieth century, the fields of classical computation and communication provided unifying methods and themes for all of engineering and science. So at the beginning of the twenty first century, quantum information is providing unifying concepts such as entanglement, and unifying techniques such as coherent information processing and quantum error correction, that have the potential to transform and bind together currently disparate fields in science and engineering.

Indeed, quantum information theory has perhaps even a greater potential to transform the world than classical information theory. Classical information theory finds its greatest application in the man-made systems such as electronic computers. Quantum information theory applies not only to man-made systems, but to all physical systems at their most fundamental level. For example, entanglement is a characteristic of virtually all physical systems at their most microscopic levels. Quantum coherence and the relationship between symmetries and the conservation and protection of information underlie not only quantum information, but the behavior of elementary particles, atoms, and molecules.

When or whether techniques of quantum information processing will become tools of mainstream technology is an open question. The technologies of precision measurement are already fully quantum mechanical: for example, the atomic clocks that lie at the heart of the global positioning system (GPS) rely fundamentally on quantum coherence. Ubiquitous devices such as the laser and the transistor have their roots in quantum mechanics. Quantum coherence is relatively fragile, however: until such a time as we can construct robust, easily manufactured coherent systems, quantum information processing may have its greatest implications at the extremes of physics and technology.

Quantum information processing analyzes the universe in terms of information: at bottom, the universe is composed not just of photons, electrons, neutrinos and quarks, but of quantum bits or qubits. Many aspects of the behavior of those elemental qubits are independent of the particular physical system that registers them. By understanding how information behaves at the quantum mechanical level, we understand the fundamental behavior of the universe itself.

Bibliography

1. Nielsen MA, Chuang IL (2000) Quantum Computation and Quantum Information. Cambridge University Press, Cambridge
2. Ehrenfest P, Ehrenfest T (1912) The Conceptual Foundations of the Statistical Approach in Mechanics. Cornell University Press, Ithaca, NY
3. Planck M (1901) Ann Phys 4:553
4. Maxwell JC (1871) Theory of Heat. Appleton, London
5. Einstein A (1905) Ann Phys 17:132
6. Bohr N (1913) Phil Mag 26:1–25, 476–502, 857–875
7. Schrödinger E (1926) Ann Phys 79:361–376, 489–527; (1926) Ann Phys 80:437–490; (1926) Ann Phys 81:109–139
8. Heisenberg W (1925) Z Phys 33:879–893; (1925) Z Phys 34:858–888; (1925) Z Phys 35:557–615
9. Einstein A, Podolsky B, Rosen N (1935) Phys Rev 47:777
10. Bohm D (1952) Phys Rev 85:166–179, 180–193
11. Bell JS (1964) Physics 1:195
12. Aharonov Y, Bohm D (1959) Phys Rev 115:485–491; (1961) Phys Rev 123:1511–1524
13. Aspect A, Grangier P, Roger G (1982) Phys Rev Lett 49:91–94; Aspect A, Dalibard J, Roger G (1982) Phys Rev Lett 49:1804–1807 ; Aspect A (1999) Nature 398:189
14. Hartley RVL (1928) Bell Syst Tech J 535–563
15. Turing AM (1936) Proc Lond Math Soc 42:230–265
16. Shannon CE (1937) Symbolic Analysis of Relay and Switching Circuits. Master's Thesis, MIT
17. Shannon CE (1948) A Math theory of commun. Bell Syst Tech J 27:379–423, 623–656
18. von Neumann J (1966) In: Burks AW (ed) Theory of Self-Reproducing Automata. University of Illinois Press, Urbana
19. Landauer R (1961) IBM J Res Develop 5:183–191
20. Lecerf Y (1963) Compte Rendus 257:2597–2600
21. Bennett CH (1973) IBM J Res Develop 17:525–532; (1982) Int J Theor Phys 21:905–940
22. Fredkin E, Toffoli T (1982) Int J Theor Phys 21:219–253
23. Benioff P (1980) J Stat Phys 22:563–591; (1982) Phys Rev Lett 48:1581–1585; (1982) J Stat Phys 29:515–546; Ann NY (1986) Acad Sci 480:475–486
24. Feynman RP (1982) Int J Th Ph 21:467
25. Deutsch D (1985) Proc Roy Soc Lond A 400:97–117; (1989) Proc Roy Soc Lond A 425:73–90
26. Shor P (1994) In: Goldwasser S (ed) Proceedings of the 35th Annual Symposium on Foundations of Computer Science. IEEE Computer Society, Los Alamitos, pp 124–134
27. Lloyd S (1993) Science 261:1569; (1994) Science 263:695
28. Cirac JI, Zoller P (1995) Phys Rev Lett 74:4091
29. Monroe C, Meekhof DM, King BE, Itano WM, Wineland DJ (1995) Phys Rev Lett 75:4714
30. Turchette QA, Hood CJ, Lange W, Mabuchi H, Kimble HJ (1995) Phys Rev Lett 75:4710
31. Grover LK (1996) In: Proceedings, 28th Annual ACM Symposium on the Theory of Computing, p 212
32. Cory D, Fahmy AF, Havel TF (1997) Proc Nat Acad Sci USA 94:1634–1639
33. Chuang IL, Vandersypen LMK, Zhou X, Leung DW, Lloyd S (1998) Nature 393:143–146
34. Chuang IL, Gershenfeld N, Kubinec M (1998) Phys Rev Lett 80:3408–3411
35. Gordon JP (1962) Proc IRE 50:1898–1908
36. Lebedev DS, Levitin LB (1963) Sov Phys Dok 8:377
37. Holevo AS (1973) Prob Per Inf 9:3; Prob Inf Trans (USSR) 9:110
38. Schumacher B, Westmoreland M (1997) Phys Rev A 55:2738
39. Holevo AS (1998) IEEE Trans Inf Th 44:69

40. Caves CM, Drummond PD (1994) Rev Mod Phys 66:481–537
41. Yuen HP, Ozawa M (1993) Phys Rev Lett 70:363–366
42. Giovannetti V, Guha S, Lloyd S, Maccone L, Shapiro JH, Yuen HP (2004) Phys Rev Lett 92:027902
43. Giovannetti V, Guha S, Lloyd S, Maccone L, Shapiro JH (2004) Phys Rev A 70:032315
44. Nielsen M, Schumacher B (1996) Phys Rev A 54:2629–2635
45. Lloyd S (1997) Phys Rev A 55:1613–1622
46. Bennett CH, Brassard G, Cripeau C, Jozsa R, Peres A, Wootters WK (1993) Phys Rev Lett 70:1895–1899
47. Pan JW, Daniell M, Gasparoni S, Weihs G, Zeilinger A (2001) Phys Rev Lett 86:4435–4438
48. Zhang TC, Goh KW, Chou CW, Lodahl P, Kimble HJ (2003) Phys Rev A 67:033802
49. Wiesner S (1983) SIGACT News 15:78–88
50. Bennett CH, Brassard G (1984) Proc IEEE Int Conf Comput 10–12:175–179
51. Ekert A (1991) Phys Rev Lett 67:661–663
52. Feynman RP (1965) The Character of Physical Law. MIT Press, Cambridge
53. Bohr A, Ulfbeck O (1995) Rev Mod Phys 67:1–35
54. Wootters WK, Zurek WH (1982) Nature 299:802–803
55. Werner RF (1998) Phys Rev A 58:1827–1832
56. Bennett CH, Bernstein HJ, Popescu S, Schumacher B (1996) Phys Rev A 53:2046–2052
57. Horodecki M, Horodecki P, Horodecki R (1998) Phys Rev Lett 80:5239–5242
58. Horodecki M, Horodecki P, Horodecki R (2000) Phys Rev Lett 84:2014–2017
59. Wootters WK (1998) Phys Rev Lett 80:2245–2248
60. Christandl M, Winter A (2004) J Math Phys 45:829–840
61. Bohm D (1952) Am J Phys 20:522–523
62. Bell JS (2004) Aspect A (ed) Speakable and Unspeakable in Quantum Mechanics. Cambridge University Press, Cambridge
63. Clauser JF, Horne MA, Shimony A, Holt RA (1969) Phys Rev Lett 23:880–884
64. Greenberger DM, Horne MA, Zeilinger A (1989) In: Kafatos M (ed) Bell's Theorem, Quantum Theory, and Conceptions of the Universe. Kluwer, Dordrecht
65. Pan JW, Daniell M, Weinfurter H, Zeilinger A (1999) Phys Rev Lett 82:1345–1349
66. Nelson RJ, Cory DG, Lloyd S (2000) Phys Rev A 61:022106
67. Frank MP, Knight TF (1998) Nanotechnology 9:162–176
68. Feynman RP (1985) Opt News 11:11; (1986) Found Phys 16:507
69. Deutsch D, Jozsa R (1992) Proc R Soc London A 439:553
70. Simon DR (1994) In: Goldwasser S (ed) Proceedings of the 35th Annual Symposium on Foundations of Computer Science. IEEE Computer Society, Los Alamitos, pp 116–123
71. Coppersmith D (1994) IBM Research Report RC19642
72. Hallgren S (2007) J ACM 54:1
73. Kuperberg G (2003) arXiv:quant-ph/0302112
74. Hallgren S, Moore C, Rötteler M, Russell A, Sen P (2006) In: Proc 38th ann ACM symposium theory comput, pp 604–617
75. Kitaev AY, Shen A, Vyalyi MN (2002) Classical and Quantum Computation. American Mathematical Society, Providence
76. Abrams DS, Lloyd S (1997) Phys Rev Lett 79:2586–2589
77. Aspuru-Guzik A, Dutoi AD, Love PJ, Head-Gordon M (2005) Science 309:1704–1707
78. Lloyd S (1996) Science 273:1073–8
79. Wiesner S (1996) arXiv:quant-ph/9603028
80. Zalka C (1998) Proc Roy Soc Lond A 454:313–322
81. Ramanathan C, Sinha S, Baugh J, Havel TF, Cory DG (2005) Phys Rev A 71:020303(R)
82. Boyer M, Brassard G, Hoyer P, Tapp A (1998) Fortsch Phys 46:493–506
83. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Science 220: 671–680
84. Farhi E, Goldstone J, Gutmann S (2001) Science 292:472–476
85. Sachdev S (1999) Quantum Phase Transitions. Cambridge University Press, Cambridge
86. Aharonov D, van Dam W, Kempe J, Landau Z, Lloyd S, Regev O (2004) In: Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2004), pp 42–51. quant-ph/0405098
87. Farhi E, Gutmann S (1998) Phys Rev A 58:915–928
88. Aharonov D, Ambainis A, Kempe J, Vazirani U (2001) In: Proc 33th ACM Symposium on Theory of Computing (STOC 2001), pp 50–59
89. Childs AM, Cleve R, Deotto E, Farhi E, Gutmann S, Spielman DA (2003) Proc 35th ACM Symposium on Theory of Computing (STOC 2003), pp 59–68
90. Farhi E, Goldstone J, Gutmann S (2007) A quantum algorithm for the Hamiltonian NAND tree. arXiv:quant-ph/0702144
91. Hamming R (1980) Coding and Information Theory. Prentice Hall, Upper Saddle River
92. Shor PW (1995) Physical Review A, 52:R2493–R2496
93. Steane AM (1996) Phys Rev Lett 77:793–797
94. Calderbank AR, Shor PW (1996) Phys Rev A 54:1098–1106
95. Gottesman D (1996) Phys Rev A 54:1862
96. Aharonov D, Ben-Or M (1997) In: Proc 29th ann ACM symposium theory comput, pp 176–188
97. Knill E, Laflamme R, Zurek W (1996) arXiv:quant-ph/9610011
98. Gottesman D (1998) Phys Rev A 57:127–137
99. Knill E (2005) Nature 434:39–44
100. Zanardi P, Rasetti M (1997) Phys Rev Lett 79:3306–3309
101. Lidar DA, Chuang IL, Whaley KB (1998) Phys Rev Lett 81: 2594–2597
102. Knill E, Laflamme R, Viola L (2000) Phys Rev Lett 84:2525–2528
103. Zanardi P, Lloyd S (2003) Phys Rev Lett 90:067902
104. Kitaev AY (2003) Ann Phys 30:2–30
105. Lloyd S (2004) Quant Inf Proc 1:15–18
106. Devetak I (2005) IEEE Trans Inf Th 51:44–55
107. Shor PW (2002) Lecture Notes, MSRI Workshop on Quantum Computation. Mathematical Sciences Research Institute, Berkley
108. DiVincenzo DP, Shor PW, Smolin JA (1998) Phys Rev A 57: 830–839
109. Shor PW (2004) Comm Math Phys 246:453–472
110. Hastings MB (2008) A counterexample to additivity of minimum output entropy. arXiv:0809.3972
111. Bennett CH, Wiesner SJ (1992) Phys Rev Lett 69:2881
112. Bennett CH, Shor PW, Smolin JA, Thapliyal AV (1999) Phys Rev Lett 83:3081–3084
113. Bennett CH, Brassard G, Robert JM (1988) SIAM Journal on Computing 17:210–229
114. Ralph TC (1999) Phys Rev A 61:010303
115. Braunstein S, Pati AK (2003) Quantum Information With Continuous Variables. Springer, New York
116. Harrow AW, Hassidim A, Lloyd S (2008) Quantum algorithm for solving linear sets of equations. arXiv:0811.3171

Quantum Information Science, Introduction to

Joseph F. Traub
Computer Science Department, Columbia University, New York, USA

Quantum information science is a hugely exciting fairly new field that has attracted many researchers from diverse fields such as physics, chemistry, computer science, and mathematics. A driving force is that Moore's law will soon end due to a number of factors. These include tunneling, the heat problem, approaching atomic size, difficulty of design, and cost of fabrication facilities. Multicores and manycores may provide a near term fix. But if we are to continue Moore's Law exponential trajectory long term entirely new technologies are needed. Possible future technologies include molecular, biological, photonic, or quantum computers. Quantum computers are based on the laws of quantum mechanics.

We've seen progress on quantum algorithms and complexity. But can quantum computers be built? Two of the impediments include the small number of qubits to date and the short decoherence times. Will there be enough qubits to do new science especially with the requirements of error correction and fault tolerant computing? Will there be enough time to do new science before decoherence sets in? In contrast to quantum computation some quantum communication applications have already been realized.

This section begins with a broad overview of quantum information science. The articles that follow discuss quantum algorithms, quantum algorithms and complexity for continuous problems, quantum computational complexity, quantum error correction and fault tolerant computing, quantum computing with trapped ions, quantum computing using optics, and finally cryptography.

Lloyd (see ► Quantum Information Processing) argues that the most important reason for studying quantum information science is to construct a unified theory of how information can be registered and transformed at the fundamental limits imposed by physical law. He introduces the formalism of quantum mechanics and applies it to the idea of quantum information. He then identifies quantum superposition and entanglement as being at the heart of quantum computation. He discusses the Deutsch–Josza algorithm which solves a problem faster than could be done by any classical computer. This is an artificial example. Decidably not artificial is Shor's algorithm for factoring large integers in polynomial time. The fastest classical algorithm known is exponential in the number of bits. Shor's algorithm is important for cryptography since the commonly used RSA public-key system relies on the difficulty of factoring to guarantee security. The importance of the quantum Fourier transform for algorithm design is discussed.

Lloyd does not describe the various technologies that can be used to build quantum computers. Every one of them is subject to its own particular form of noise. He does present a general formalism for characterizing noise and errors and discusses techniques for coping with them.

Mosca (see ► Quantum Algorithms) defines quantum algorithms as algorithms that run on any realistic model of quantum computation. The most commonly used model of quantum computation is the circuit model (more strictly, the model of uniform families of acyclic quantum circuits), and the quantum Strong Church–Turing thesis states that the quantum circuit model can efficiently simulate any realistic model of computation. Several other models of quantum computation have been developed, and indeed they can be efficiently simulated by quantum circuits. Quantum circuits closely resemble most of the currently pursued approaches for attempting to construct scalable quantum computers.

Papageorgiou and Traub (see ► Quantum Algorithms and Complexity for Continuous Problems) point out that most continuous mathematical formulations arising in science and engineering can only be solved numerically and therefore approximately. There are two major motivations for studying quantum algorithms and complexity for continuous problems.

1. Are quantum computers more powerful than classical computers for important scientific problems? How much more powerful?
2. Many important scientific and engineering problems have continuous formulations. These problems occur in fields such as physics, chemistry, engineering and finance.

The continuous formulations include path integration, partial differential equations (in particular, the Schrödinger equation) and continuous optimization.

To answer the first question the classical computational complexity of the problem must be known. There have been decades of research on the classical complexity of continuous problems in the field of information-based complexity. The reason the complexity of many continuous problems is known is that adversary arguments can be used to obtain their query complexity. Regarding the second motivation, in this article they report on high-dimensional integration, path integration, Feynman path in-

tegration, the smallest eigenvalue of a differential equation, approximation, partial differential equations, ordinary differential equations and gradient estimation. They also briefly report on the simulation of quantum systems on a quantum computer.

Watrous (see ► Quantum Computational Complexity) provides a survey of quantum computational complexity, with a focus on three fundamental notions: polynomial-time quantum computations, the efficient verification of quantum proofs, and quantum interactive proof systems. Based on these notions he defines quantum complexity classes that contain computational problems of varying hardness. Properties of these complexity classes, and the relationships among these classes and classical complexity classes, are presented. As these notions and complexity classes are typically defined within the quantum circuit model, this article includes a section that focuses on basic properties of quantum circuits that are important in the setting of quantum complexity. A selection of other topics in quantum complexity, including quantum advice, space-bounded quantum computation, and bounded-depth quantum circuits, is also presented.

Grassl and Rotteler (see ► Quantum Error Correction and Fault Tolerant Quantum Computing) point out that it has been shown that even with imperfect quantum memory and imperfect quantum operations it is possible to implement arbitrary long quantum computation, provided that the failure probability of each element is below a certain threshold. They provide an overview of the ingredients leading to fault tolerant quantum computation (FTQC). In the first part, they present the theory of quantum error-correcting codes (QECCs) and in particular two important classes of QECCs, namely the so-called CSS codes and stabilizer codes. Both are related to classical error-correcting codes, so they start with some basics from this area. In the second part of the article, they present a high-level view of the main ideas of FTQC and the threshold theorem.

Lange (see ► Quantum Computing with Trapped Ions) summarizes the state-of-the-art of quantum computing with trapped ions. All the necessary components of a trapped ion quantum computer have been demonstrated, from quantum memory and fundamental quantum logic gates to simple quantum algorithms. Current experimental efforts are directed towards scaling up the small systems investigated so far and enhancing the fidelity of operations to a level where error correction can be applied efficiently. The first task at which a quantum computer is expected to outperform a classical one is the efficient simulation of quantum systems too complex for classical treatment.

Milburn and White (see ► Quantum Computing Using Optics) point out that optical implementations of quantum computing have largely focused on encoding quantum information using single photon states of light. For example, a single photon could be excited to one of two carefully defined orthogonal mode functions of the field with different momentum directions. However, as optical photons do not interact with each other directly, physical devices that enable one encoded bit of information to unitarily change another are hard to implement. In principle it can be done using a Kerr nonlinearity, but Kerr nonlinear phase shifts are too small to be useful. Knill et al. discovered another way in which the state of one photon could be made to act conditionally on the state of another using a measurement based scheme. They discuss this approach in some detail as it has led to experiments that have already demonstrated many of the key elements required for quantum computation with optics.

Lo and Zhao (see ► Quantum Cryptography) point out that the goal of quantum cryptography is to perform tasks that are impossible or intractable with conventional cryptography. Quantum cryptography makes use of the subtle properties of quantum mechanics such as the quantum no-cloning theorem and the Heisenberg uncertainty principle. Unlike conventional cryptography, whose security is often based on unproven computational assumptions, quantum cryptography has an important advantage in that its security is often based on the laws of physics. Thus far, proposed applications of quantum cryptography include quantum key distribution (abbreviated QKD), quantum bit commitment and quantum coin tossing. These applications have varying degrees of success. The most successful and important application (QKD) has been proven to be unconditionally secure. Moreover, experimental QKD has now been performed over hundreds of kilometers over both standard commercial telecom optical fibers and open-air. In fact, commercial QKD systems are currently available on the market.

Quantum Phenomena in Semiconductor Nanostructures

Umberto Ravaioli
University of Illinois at Urbana-Champaign,
Urbana, USA

Article Outline

Glossary
Definition of the Subject

Introduction
Quantum Effects in Semiconductors
Size Quantization
Inclusion of Realistic Band Structure
Brief Survey of Quantum Transport Models
Quantum Corrections in Particle Simulation
Online Resources
Future Directions
Acknowledgments
Bibliography

Glossary

Ballistic transport A regime of conduction where charge carriers do not exhibit appreciable scattering while moving in a certain material region.

Band structure The effective energy-momentum dispersion followed by valence and conduction electrons in a crystal, when represented as quasi-particles moving in a uniform medium equivalent to the actual crystal environment characterized by a periodic distribution of atomic potential wells.

Hot carriers Electrons or holes in a semiconductor, with energy and velocity appreciably exceeding the average values for thermal equilibrium with the lattice, typically in regions of high electric fields.

Quantum correction Addition of terms to a semi-classical model to cause particles obeying classical motion laws to follow collectively quantum behavior, for the purpose to simulate correctly quantum transport effects with considerably cheaper computations.

Scattering Collisions experienced by charge carriers moving through a semiconductor with vibrational modes of the crystal lattice (phonon scattering), with charged or neutral impurities (impurity scattering), with other charge carriers in proximity (short-range charge-charge scattering) or with collective charge vibrational modes of other carriers over a long-range (plasmon scattering).

Quantum sub-band Discrete projection of the band structure in the directions of unrestricted motion, corresponding to a specific discrete state created by size quantization.

Size quantization Separation of electronic states into discrete energy levels, in a region of restricted dimensionality, typically at interfaces between different materials which create a quantum potential well. In semiconductor devices, size quantization is normally present in the transverse cross-section of narrow conduction channels delimited by heterojunction interfaces.

Definition of the Subject

Semiconductor devices are the building blocks of integrated circuits at the heart of computer chips and other electronics systems which are ubiquitous in modern society. While quantum physics is essential to describe the properties of semiconductor materials, transport of electronic carriers can be treated with classical laws of motion, as long as particles experience random scattering events with the crystal lattice along their path. However, in many practical devices, carriers are confined in narrow channels, so that size quantization effects should be included in the direction normal to the conduction path to model accurately particle density. As the device size is reduced to pack more and more functionality in integrated circuits, quantum coherence may eventually become important in the direction of propagation if there is a sufficient decrease of scattering events approaching quasi-ballistic transport conditions. If the device structure presents potential barriers along the direction of propagation, coherence may also manifest itself via quantum tunneling. It is essential to have a detailed understanding of quantum effects in electronic transport to design effectively devices at the nanoscale, sustain the miniaturization trends of integrated circuits, and create new engineered nanostructures.

Introduction

In the second half of the 20th century, semiconductor devices have progressively made their way into nearly every aspect of technology and everyday life. The introduction of integrated circuits and microprocessors ushered the present era of computers, internet, and mobile communications. This unprecedented growth has been made possible by the ability to scale down the size of devices and to increase their speed of operation, so that the complexity of integrated circuits has steadily climbed, leading to new and more powerful applications. The pace of innovation is exemplified by Moore's law [49], an empirical observation which has predicted the doubling of transistors on a chip every two years. This has meant that over the last thirty five years the number of transistors on a commercial microprocessor has increased from slightly over two thousand to nearly one billion.

The most common transistor found in integrated circuits is the metal-oxide-semiconductor field-effect transistor (MOSFET) [50,67] based on a thin layer of silicon dioxide (SiO_2) sandwiched between a silicon conducting channel and a gate electrode, which forms a capacitive structure controlling the flow of charge in the channel when one varies the voltage applied to the gate. A schematic diagram is shown in Fig. 1. In modern com-

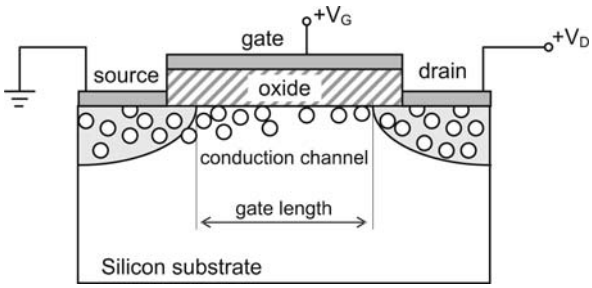

Quantum Phenomena in Semiconductor Nanostructures, Figure 1
Schematic diagram of the metal-oxide-semiconductor field-effect transistor (MOSFET)

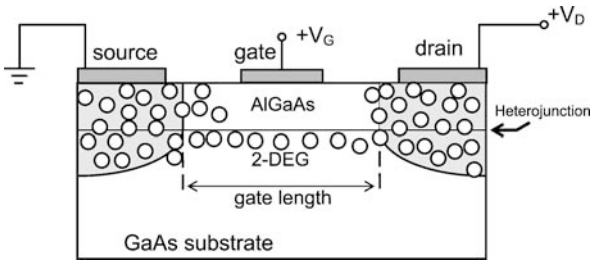

Quantum Phenomena in Semiconductor Nanostructures, Figure 2
Schematic diagram of a high electron mobility transistor (HEMT)

mercial devices, the effective length of the channel has shrunk to about 50 nm and the thickness of the oxide layer can be of the order of just 1 nm.

The use of silicon MOSFETs has become so widespread for a number of reasons. Silicon is a readily available material and it is very easy to grow layers of oxide on a silicon surface. When compared to other semiconductor materials, however, silicon does not seem to possess particularly outstanding properties. Electrons move somewhat slowly under the application of an electric field, because of a relatively heavy effective mass and the high rate of scattering events with the vibrations (phonons) of the atoms forming the crystal structure. Other III–V semiconductor materials like GaAs have a smaller effective mass, so that electrons can move much faster in a certain range of electric fields with a higher drift velocity. It was widely thought at some point that for high-performance applications it would be better to adopt such semiconductor materials with an intrinsically faster mechanism for transistor switching. Also, if a heterojunction is formed between materials like GaAs and the AlGaAs alloy, electron mobility could be increased by orders of magnitudes for motion on the interface plane, leading to high electron mobility transistors (HEMT) [11,67] shown schematically in Fig. 2.

Despite their potential, III–V compound materials have not been widely employed for computing chips and have been mainly relegated to niche applications in microwaves and optoelectronics. Silicon has had of course the advantage of several more decades of development and the corresponding lower production costs and higher reliability, but in many regards the perceived weaknesses of silicon have been the reason for its success in integrated circuits applications. For miniaturization, it is very important for a device to retain its behavior essentially unchanged when scaled, so that fabrication approaches and system architectures do not need to change radically from one generation to the next. The carrier transport behavior in GaAs and other III–V materials is strongly dependent on the actual electric fields established inside the conduction channel. While electron mobility is very high at low fields, the saturation velocity, at fields established in a practical device, is actually slightly lower than in silicon, due to the intervalley scattering mechanisms to upper conduction valleys which have a higher effective mass and lower mobility [26].

To realize even higher mobility, devices need to be sufficiently small so that they can operate in the so-called overshoot regime, where electrons traverse the channel before intervalley transfer events can take place and never reach the saturation velocity in a bulk material at a comparable electric field. However, by scaling further the device, a regime of quasi-ballistic transport can be established more easily where quantum effects and coherence of the electron wave need to be dealt with. Silicon, with its heavier effective mass, has essentially retained until today most of its classical behavior as far as transport along the channel axis is concerned. Size quantization effects in the cross-section of the channel are becoming more influential but the nature of electronic transport has not changed substantially even in the nanometer size range. Generations of device designers have been able to gradually adapt their approaches to make new devices behave acceptably as they have been scaled, but the process has been an evolutionary rather than a revolutionary one. From the point of view of circuit realization, the broad adoption of silicon MOSFETs has been aided by the availability of the C-MOS architecture [50,61] based on a basic inverter structure which only conducts current when the switching between two logic states takes place. If millions of transistors are packed in the same integrated circuit it is crucial to minimize power consumption to avoid the possibility of thermal failure. The ready availability of SiO_2 oxide for the gate insulator has been of fundamental importance, since realization of comparable gate oxides in alternative material systems is in general costly, impractical or not possi-

ble at all. To-date, there is simply no other technology that can approach, even remotely, what can be achieved with present C-MOS silicon technology for large-scale device integration.

This contribution deals with quantum models and simulation techniques used for the analysis of semiconductor structures within the effective mass approximation. These are applicable to devices with standard topology which are scaled down to the nanometer range. At even smaller scale, one has to deal with small clusters of materials and atomistic models are necessary to capture the full physics of the problem, since a bulk description is no longer sufficient to characterize the properties which may be strongly dependent on the cluster size. A detailed treatment of atomistic approaches is beyond the scope of this work.

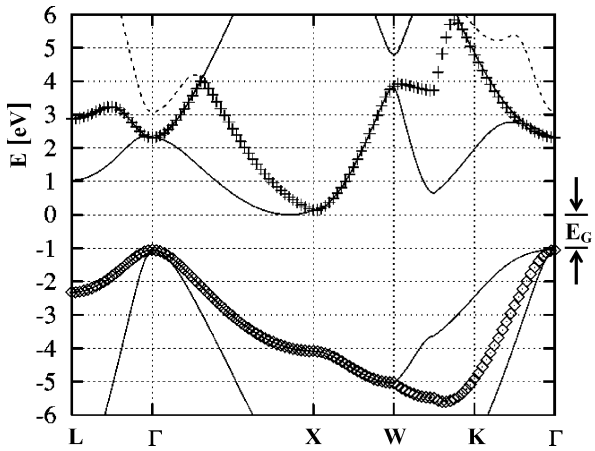

Quantum Phenomena in Semiconductor Nanostructures, Figure 3

Simplified band structure of silicon in the standard textbook representation, drawn along crystal momentum directions connecting the main symmetry points, with the Γ point corresponding to zero crystal momentum. Calculations were performed with an empirical pseudo-potential approach. The band gap E_G between the top of the valence band and the bottom of the conduction band is slightly larger than 1 eV in silicon

Quantum Effects in Semiconductors

Quantum effects manifest themselves in a number of ways in semiconductor structures. First of all, the underlying model of transport in bulk material is based on the semiconductor band structure, which is derived from a quantum mechanical description of the crystal lattice [30]. The band structure describes the extended valence and conduction band states originating from superposition of the individual atomic states of the lattice nodes. The motion of actual electrons can then be studied in terms of quasi-particles obeying the energy-momentum dispersion relation expressed by the band structure. Negative electron quasi-particles are considered for the excited states in the conduction band and positive hole quasi-particles (vacancies of electrons) are considered in the valence band. Semiconductors normally have a band of forbidden energies (band gap) separating conduction and valence band states, as one can see in Fig. 3 showing the band structure of silicon.

Although the realistic band structure for a 3-D bulk semiconductor has in general a rather complicated shape and multiple eigenvalue branches, for simple considerations a two-band model can be used to describe hole motion at the top of the valence band and electron motion at the bottom of the conduction band, adopting an approximately parabolic energy-momentum dispersion relation. The curvatures of these parabolic dispersions define effective masses to express the kinetic energy of the quasi-particles. The effective masses for electrons and holes in a semiconductor normally differ significantly from the actual mass of an electron in vacuum. The essential quantum effects of the crystal are encapsulated in the effective mass, which is adequate to study particle transport at least in conditions not very far from equilibrium. In a bulk region, classical laws of motions can be used by applying the effective mass to relate momentum and energy. In conditions of reduced dimensionality at interfaces, the wave nature of quasi-particles may have to be considered and a simple quantum model also adopts the effective mass, for instance to express the kinetic energy term in the Schrödinger equation. The effective mass parabolic band model is widely used in a quantum model because of its simplicity. When the system is driven sufficiently out of equilibrium, a more complete band structure should be included but the complexity of the model increases rapidly resulting in considerable numerical complications for the solution of transport equations. Often, an intermediate model is acceptable for practical applications, where a nonparabolic relation is introduced to express the deviation from parabolic behavior of the dispersion relation at higher kinetic energies, still retaining the effective mass of the parabolic model as a parameter that calibrates the model.

Significant quantum effects are first encountered when an interface between dissimilar materials (heterojunction) is formed. The two materials will have in general different band gap and different positioning of conduction and valence band edges with respect to a common energy reference. As a result, at the interface there is a potential energy discontinuity between the bands. For the Si/SiO_2 interface of MOS systems, the typical discontinuity of the conduction band is on the order of 3.0 eV because the oxide is

a dielectric medium with a large band gap, while heterojunctions between usual III–V materials is typically only a fraction of eV. Carriers may be concentrated electrostatically at the heterojunction interface, creating a thin potential well, where electronic states are quantized. In the MOS system, the space charge in the semiconductor side is created by applying an appropriate potential bias on a gate electrode placed on top of the oxide layer [50,61,67]. In III–V compound systems, the technique known as modulation doping is used instead, where doping atoms are placed in the layer of semiconductor with a wider band gap [10,67]. Considering for illustration the case of *n*-type doping, the excess electrons provided by the doping migrate to the semiconductor layer with a smaller band gap where the conduction band is at lower potential energy. Thus, a layer of negative charge is formed at the interface, leaving behind a positive layer due to the ionized donor atoms in the wide band gap semiconductor. A metallic gate is also used in this case to control the flow of carriers in the channel and shut the device off.

The mobile charge at the interface forms a thin sheet, confined by the self-consistent potential well which has a nearly triangular profile close to the interface. Geometric confinement may also be added, by placing a second interface underneath, so that the narrow band gap material layer is sandwiched between layers of oxide or wider band gap semiconductor. The mobile charge sheet is usually called a two-dimensional electron gas (2-DEG). Because carriers reside in the potential well, transport normal to the interface is restricted, but they are free to move on a 2-D plane parallel to the interface. The Hamiltonian of the carriers may be decomposed into normal and parallel components. The confined quantum energy states in the potential well define sub-bands [11], for transport in the parallel plane, which are 2-D projections of the band structure. Each transverse quantum level becomes the reference zero of kinetic energy in the corresponding 2-DEG sub-band. When an electric field parallel to the interface is applied, motion of the carriers is two-dimensional and scattering is also restricted to the plane, so that the momentum of the final states cannot acquire a component normal to the interface. MOSFET and HEMT devices are realized by placing a source and a drain electrode at the two ends of a conduction channel realized with a 2-DEG. The gate electrode in the middle regulates the flow of carriers and can be biased with a repulsive potential to shut the device off.

For illustration, Fig. 4 shows an example of self-consistent equilibrium calculations of electron energy states in MOS structure at the Si/SO_2 interface aligned on the [100] direction of the crystal. Silicon has six degenerate

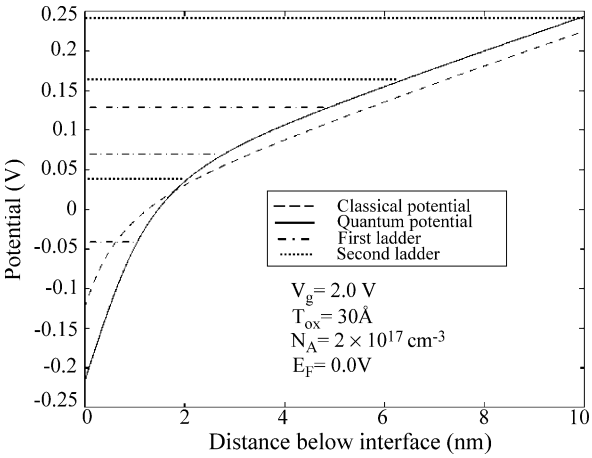

Quantum Phenomena in Semiconductor Nanostructures, Figure 4

Example of potential distribution and energy states obtained on the silicon layer of an MOS capacitor structure. The reference zero energy is the Fermi level in the semiconductor

equivalent valleys, close to the X point of the band structure which, at relatively low energy values, are associated with ellipsoidal iso-energy surfaces aligned two by two along the three principal axes of symmetry perpendicular to equivalent <100> planes. On the interface, two ellipsoids project as degenerate energy circles, corresponding to a first set of quantum state solutions (first ladder). The other four ellipsoids project sideways as ellipses, corresponding to a second set of solution at higher eigenenergies (second ladder). Figure 4 shows the first three energy solutions for each ladder and the corresponding self-consistent potential energy profile obtained from the quantum calculation as well as the profile obtained with a classical model neglecting the quantum states. The corresponding quantum and classical charge distributions are shown in Fig. 5. One can see that the classical model predicts a maximum of charge density exactly at the interface. The quantum density, obtained from the detail of energy states and wave functions, shows nearly zero density at the interface, while the maximum of charge has shifted to a depth of almost 1.0 nm.

As a further step, motion can be restricted in one of the parallel plane directions, either by geometrical confinement, for instance by etching the material, electrical confinement by placing additional electrodes or both. One could obtain in this way a one-dimensional electron gas, confining the carriers to a quantum wire [54]. A 1-D subband structure is associated with the motion along the channel direction corresponding to the only remaining degree of freedom for kinetic energy. In such a structure,

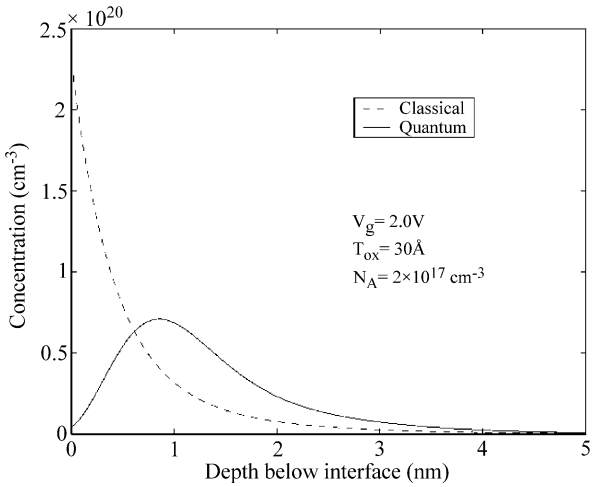

Quantum Phenomena in Semiconductor Nanostructures, Figure 5

Example of inversion electron charge concentration under the gate of an MOS structure as a function of depth below the oxide interface

scattering events can only result in a final state with momentum pointing forward or backwards, with reference to the axis of the conduction channel. Motion in a section of quantum wire can also be further restricted by placing geometric or electrostatic barriers, to create a 3-D cavity called a quantum dot, where states are fully quantized [58]. There are no sub-bands associated with these discrete quantum states because motion is completely restricted. Coupling of the states with the environment outside the cavity may take place, for instance by penetration of evanescent waves through potential barriers delimiting the cavity.

The quantum effects discussed above are the result of size quantization. Restriction of motion in a specific direction causes the energy to be quantized in that direction, in terms of discrete eigenvalues, but quantum effects may also appear along a direction of unrestricted motion where energy belongs to a continuum distribution, if there are sufficiently abrupt potential or geometric discontinuities. A step in potential energy, for instance, creates quantum mechanical reflection, while motion against a thin potential barrier may result in tunneling with evanescent propagation inside the barrier region. Interesting phenomena of mode coupling take place in a quantum wire undergoing changes in the geometric cross-section, because of the relative shifts of the transverse energy states [44].

A much-studied quantum problem in semiconductor devices is resonant tunneling [22,36,56,64]. Here, two narrowly spaced layers of a wide band gap material are embedded in a semiconducting medium of lower band gap, creating two potential barriers. The region between the barriers is a potential well with quantum states which are resonantly coupled with the continuum states on the two sides of the structure. Tunneling is mostly favorable at incident energies corresponding to these resonant states, because the multiple reflections at the interface discontinuities create a constructive interference for transmission and destructive interference for reflection. The current-voltage characteristics peak in correspondence of the resonances, therefore exhibiting negative differential resistance behavior. One can extend the concept and create a large number of wide band gap material layers, thus creating an artificial material called a superlattice [72]. In the direction normal to the interfaces, the coupling between multiple quantum reflections creates ranges of energies with favorable transmission (minibands) and forbidden energy gaps that create an effective artificial band structure.

Size Quantization

In the size quantization problem, discrete stationary states are associated with mobile carriers, obeying the time-independent Schrödinger equation for the coordinates restricted to motion with a general form

$$\hat{H}\psi_j = E_j\psi_j \,, \tag{1}$$

where $\hat{H}$ is the Hamiltonian, E_j the energy eigenvalues of the confined states and ψ_j the corresponding wave functions. For simplicity, the following consideration will be for electrons in the conduction band. Similar considerations hold in principle for holes in the valence band, although the details of the band structure may require some more involved formalism.

For 1-D confinement, where the motion in the quantum channel is delimited by a planar interface, e. g. a heterojunction, the electron states can be characterized by an envelope wave function

$$\psi(\vec{r}) = \xi_j(z)\exp(i\vec{k}\cdot\vec{r}) \,, \tag{2}$$

where z is the direction perpendicular to the heterojunction, and $\vec{r}$ and $\vec{k}$ are the 2-D position vector and wave vector for motion parallel to the interface. In the plane parallel to the interface the wave function component is simply a plane wave, while confinement is described by the function $\xi_j(z)$ which corresponds to the jth confined sub-band and satisfies the 1-D time-independent Schrödinger equation in the z direction

$$-\frac{\hbar^2}{2m^*}\frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{1}{m^*}\frac{\mathrm{d}\xi_j(z)}{\mathrm{d}z}\right) + U(z)\xi_j(z) = E_j\xi_j(z) \,. \tag{3}$$

The potential energy $U(z)$ is given by

$$U(z) = -e\phi(z) + V_h(z) + V_{\mathrm{ex}}(z)\,, \tag{4}$$

where $\phi(z)$ is the electrostatic potential, $V_h(z)$ is a step function describing the interface barrier, and $V_{\mathrm{ex}}(z)$ is the local exchange-correlation potential. Assuming the case of an electron channel obtained on a p-type substrate, the electrostatic potential is obtained from a solution of Poisson's equation

$$\frac{\mathrm{d}}{\mathrm{d}z}\left(\varepsilon_d \frac{\mathrm{d}\phi(z)}{\mathrm{d}z}\right) = -e\left[\sum\nolimits_j N_j \left|\xi_j(z)\right|^2 - p\,(z,\phi) + N_{\mathrm{A}}^-(z) - N_{\mathrm{D}}^+(z)\right] = -\rho\,. \tag{5}$$

$N_{\mathrm{A}}(z)$ and $N_{\mathrm{D}}(z)$ are the concentrations of ionized acceptor and donor dopants, N_j represents the sub-band electron concentration, p is the hole concentration in the substrate, ε_d is the space dependent permittivity and ρ represents the net charge density. The hole concentration may be approximated as a classical equilibrium distribution with a space profile that is nonlinearly dependent on the electrostatic potential. The formulation for N_j involves the integral of the Fermi function which, in the case of a 2-DEG, can be integrated analytically, giving [11]

$$N_j = \frac{D}{e} g_{\mathrm{v}} k_{\mathrm{B}} T \ln\left[1 + \exp\left(\frac{E_F - E_j}{k_{\mathrm{B}} T}\right)\right]\,. \tag{6}$$

Here, g_{v} is the valley degeneracy for the specific eigenvalue ladder, related to the band structure symmetry and the projection of the conduction band valleys on the interface plane. The 2-DEG density of states is a constant as a function of energy and it is given by

$$D = \frac{e m^*}{\pi \hbar^2}\,. \tag{7}$$

If a qualitative approximation is sought, starting from a fixed potential profile, solutions of the Schrödinger equation for the wave functions may just be available analytically in some cases. For instance, for a potential approximated by an infinite barrier at a heterojunction interface ($z = 0$) and by a triangular distribution $V(z) = eF_{\mathrm{s}}z$ for $z > 0$ in the semiconductor (where F_{s} is an effective electric field assumed to be constant throughout the layer) solutions for wave functions and energy eigenvalues can be expressed in terms of Airy functions [1]

$$\xi_j(z) = Ai\left\{\left(\frac{2m^* eF_{\mathrm{s}}}{\hbar^2}\right)^{\frac{1}{3}}\left[z - \left(\frac{E_j}{eF_{\mathrm{s}}}\right)\right]\right\}\,, \tag{8}$$

$$E_j(z) = \left(\frac{\hbar^2}{2m^*}\right)^{\frac{1}{3}}\left[\frac{3}{2}\pi e F_{\mathrm{s}}\left(j + \frac{3}{4}\right)\right]^{\frac{2}{3}}\,. \tag{9}$$

For double confinement, the mathematical formulation of the problem is very similar with more general differential operators [38]

$$-\frac{\hbar^2}{2m^*}\nabla\cdot\left(\frac{1}{m^*}\nabla\psi_j\right) + U(z)\psi_j = E_j\psi_j\,, \tag{10}$$

$$\nabla\cdot(\varepsilon_d\nabla\phi) = -e\left[\sum\nolimits_j N_j\left|\psi_j\right|^2 - p\,(\phi) + N_{\mathrm{A}}^- - N_{\mathrm{D}}^+\right]. \tag{11}$$

The main formal difference is in the expression for the sub-band carrier concentration since, except for the 2-DEG case, the Fermi function cannot be integrated exactly. This quantity is now expressed in terms of the Fermi integral of order $-1/2$

$$N_j = \frac{g_v}{\pi}\left(\frac{2m^* k_{\mathrm{B}} T}{\hbar^2}\right)^{\frac{1}{2}} F_{\frac{-1}{2}}\left(\frac{E_F - E_j}{k_{\mathrm{B}} T}\right)\,, \tag{12}$$

which needs to be evaluated numerically. An efficient and very accurate approach is based on rational function approximations [2]. For practical applications, the most interesting quantity is the quantum electron density, which is obtained from combined information given by the wave function and the energy eigenstates as

$$n_{\mathrm{q}}(\phi) = \frac{g_v}{\pi}\left(\frac{2m^* k_{\mathrm{B}} T}{\hbar^2}\right)^{\frac{1}{2}}\sum_j \psi_j^2 F_{\frac{-1}{2}}\left(\frac{E_F - E_j}{k_{\mathrm{B}} T}\right). \tag{13}$$

The time-independent Schrödinger equation specifies an eigenvalue problem and the Poisson equation is an elliptical partial differential equation. Because of the different mathematical nature, the two equations cannot be solved simultaneously and an iterative procedure is necessary instead. A simple iteration by itself does not converge and usually an underrelaxation in the electron density is applied with a relaxation parameter $\omega^{(m)}$ which may have to be adaptively modified as the iteration progresses. The algorithm flow of this simple approach can be summarized as follows:

(1) Solve the nonlinear Poisson equation using the quantum electron density from the last iteration

$$\nabla\cdot\left(\varepsilon\nabla\phi^{(m+1)}\right) = -\rho\left[n_{\mathrm{q}}^{(m)}, \phi^{(m+1)}\right]\,.$$

(2) Solve Schrödinger equation using the old $n_{\mathrm{q}}^{(m)}$

$$\hat{H}\left[\phi^{(m+1)}, V_{\mathrm{xc}}\left(n_{\mathrm{q}}^{(m)}\right)\right]\psi_i^{(m+1)} = E_i^{(m+1)}\psi_i^{(m+1)}\,.$$

(3) Calculate an intermediate new quantum electron density $n_{\mathrm{q,int}}^{(m+1)}$

$$n_{\mathrm{q,int}}^{(m+1)} = n_{\mathrm{q}}\left[E_i^{(m+1)}, \psi_i^{(m+1)}\right] .$$

(4) Underrelax in n_{q} to achieve convergence

$$n_{\mathrm{q}}^{(m+1)} = \omega^{(m+1)} n_{\mathrm{q,int}}^{(m+1)} + \left(1-\omega^{(m+1)}\right) n_{\mathrm{q}}^{(m)} .$$

(5) Repeat the iteration until n_{q} becomes stationary

$$\left\| n_{\mathrm{q}}^{(m+1)} - n_{\mathrm{q}}^{(m)} \right\|_2 \leq \text{tolerance} .$$

The problem with this method is the inherent instability of the outer iteration which is controlled by the underrelaxation procedure only. The necessary relaxation parameter $\omega^{(m)}$ is not known in advance and needs to be readjusted during the iteration. If $\omega^{(m)}$ is too large, the integral of the quantized charge oscillates without converging or if it is too small, convergence may be unbearably slow.

The algorithm may be modified to address these shortcomings, in a way that partially decouples the differential equations and dampens the electric charge oscillation. If one knew the exact dependence of the quantum electron density n_{q} on the electrostatic potential, it would be sufficient to solve the nonlinear Poisson equation, without the need for coupling with the Schrödinger equation by an outer iteration, but what one can do practically is to find a suitable approximate expression for the quantum electron density $\tilde{n}_{\mathrm{q}}(\phi)$ and use such an expression in a predictor-corrector type of approach [71]. This procedure would approximately decouple both equations and move most of the nonlinearities in a Poisson equation of the type

$$\nabla \cdot (\varepsilon \nabla \phi) = -\rho\left[\tilde{n}_{\mathrm{q}}(\phi), \phi\right] . \tag{14}$$

The predicted result for n_{q} and ϕ from this equation would then be corrected in an outer iteration step by an exact solution of Schrödinger equation. The electrostatic potential enters the quantum electron density $n_{\mathrm{q}}(\phi)$ through the potential dependence of sub-band energy levels and wave functions, following the form of Eq. (13)

$$n_{\mathrm{q}}(\phi) = \frac{g_v}{\pi}\left(\frac{2m^* k_{\mathrm{B}} T}{\hbar^2}\right)^{\frac{1}{2}} \cdot \sum_j \psi_j^2(\phi) F_{\frac{-1}{2}}\left(\frac{E_F - E_j(\phi)}{k_{\mathrm{B}} T}\right) . \tag{15}$$

By using the derivative property of the Fermi–Dirac integral,

$$\frac{\mathrm{d}}{\mathrm{d}x} F_k(x) = F_{k-1}(x) , \tag{16}$$

one can show [71] that, under a perturbation $\delta\phi$ of the potential, the corresponding variation in quantum density can be approximated as

$$\delta\tilde{n}_{\mathrm{q}}(\phi, \delta\phi) = \frac{g_v}{\pi}\left(\frac{2m^* k_{\mathrm{B}} T}{\hbar^2}\right)^{\frac{1}{2}} \cdot \sum_j \psi_j^2(\phi) F_{\frac{-3}{2}}\left(\frac{E_F - E_j(\phi)}{k_{\mathrm{B}} T}\right) \frac{q\delta\phi}{k_{\mathrm{B}} T} . \tag{17}$$

Since for a perturbation one can write

$$\tilde{n}_{\mathrm{q}}(\phi + \delta\phi) = \tilde{n}_{\mathrm{q}}(\phi) + \delta\tilde{n}_{\mathrm{q}}(\phi, \delta\phi) , \tag{18}$$

using once more the derivative properties of the Fermi–Dirac integrals, one obtains

$$\tilde{n}_{\mathrm{q}}(\phi + \delta\phi) = \frac{g_v}{\pi}\left(\frac{2m^* k_{\mathrm{B}} T}{\hbar^2}\right)^{\frac{1}{2}} \cdot \sum_j \psi_j^2(\phi) F_{\frac{-1}{2}}\left(\frac{E_F - E_j(\phi) + q\delta\phi}{k_{\mathrm{B}} T}\right) . \tag{19}$$

Compared with the expression for $n_{\mathrm{q}}(\phi)$ one can see that the effect of the potential perturbation is simply translated into a variation of the energy levels by a corresponding potential amount

$$E_n(\phi) \to E_n(\phi) - q\delta\phi . \tag{20}$$

Using this result, the predictor-corrector procedure starts with the solution of a nonlinear Poisson equation

$$\nabla \cdot (\varepsilon \nabla \phi) = -q\left[\tilde{n}_{\mathrm{q}}(\phi) - p(\phi) - N_{\mathrm{D}}^{+}(\phi) + N_{\mathrm{A}}^{-}(\phi)\right] , \tag{21}$$

where we use the potential dependent solution for the quantum electron density (predictor)

$$\tilde{n}_{\mathrm{q}}(\phi) = \frac{g_v}{\pi}\left(\frac{2m^* k_{\mathrm{B}} T}{\hbar^2}\right)^{\frac{1}{2}} \cdot \sum_j \left(\psi_j^{(m)}\right)^2 F_{\frac{-1}{2}}\left(\frac{E_F - E_n^{(m)} + q\left(\phi - \phi^{(m)}\right)}{k_{\mathrm{B}} T}\right) . \tag{22}$$

The superscripts (m) denote quantities obtained at the previous outer iteration step. Solution of the nonlinear Poisson equation is conveniently accomplished with a Newton–Raphson procedure, since terms of the Jacobian matrix are easily obtained in terms of $F_{-3/2}$. The electrostatic

potential $\phi^{(m+1)}$ obtained from solving the Poisson equation is then used within the Schrödinger equation (corrector) formulated as

$$-\frac{\hbar^2}{2}\nabla\cdot\left(\frac{1}{m^*}\nabla\psi_j^{(m+1)}\right)+\left[V_h-q\phi^{(m+1)}+V_{\mathrm{xc}}\left(\tilde{n}_{\mathrm{q}}^{(m+1)}\right)\right]\psi_j^{(m+1)}=E_j^{(m+1)}\psi_j^{(m+1)}\,, \quad (23)$$

to calculate the update of the corrected quantum density

$$n_{\mathrm{q}}^{(m+1)}=\frac{g_v}{\pi}\left(\frac{2m^*k_{\mathrm{B}}T}{\hbar^2}\right)^{\frac{1}{2}}\cdot\sum_j\left(\psi_j^{(m+1)}\right)^2F_{-\frac{1}{2}}\left(\frac{E_F-E_j^{(m+1)}}{k_{\mathrm{B}}T}\right). \quad (24)$$

The predictor-corrector algorithm steps can be summarized as follows:

(1) Solve the nonlinear Poisson equation using the approximation $\tilde{n}_{\mathrm{q}}(\phi)$.
(2) Solve Schrödinger equation using the latest predictor value $\tilde{n}_{\mathrm{q}}^{(m+1)}$.
(3) Calculate the exact quantum electron density $n_{\mathrm{q}}^{(m+1)}$.
(4) Repeat iteration until the electron density becomes stationary

$$\left\|n_{\mathrm{q}}^{(m+1)}-n_{\mathrm{q}}^{(m)}\right\|_2\leq\text{tolerance}\,.$$

Note that at step (2) the latest predictor value $\tilde{n}_{\mathrm{q}}^{(m+1)}$ is also used to evaluate the exchange correlation term in the Schrödinger equation, since practical observation suggests that convergence would be much worse if the previous value $\tilde{n}_{\mathrm{q}}^{(m)}$ was used. A range of numerical tests have also shown that the predictor-corrector algorithm does not need an additional underrelaxation and that the simple outer iteration is inherently stable and convergent.

Size quantization is important in narrow channel devices. As the channel is shortened, the bulk region under the interface makes it difficult to shut the device off completely because stray carriers may transfer from source to drain causing a parasitic current. The situation is in principle improved by using a double gate MOSFET structure [21], as shown in Fig. 6, which reduces the silicon region to a thin slab which may emphasize the effects of size quantization since two interfaces are now present. Realization of a double gate structure with horizontal interfaces is technologically difficult, because a layer of silicon would need to be grown on oxide.

An approximate practical realization is made with a fin structure with oxide and gate contact wrapped around as

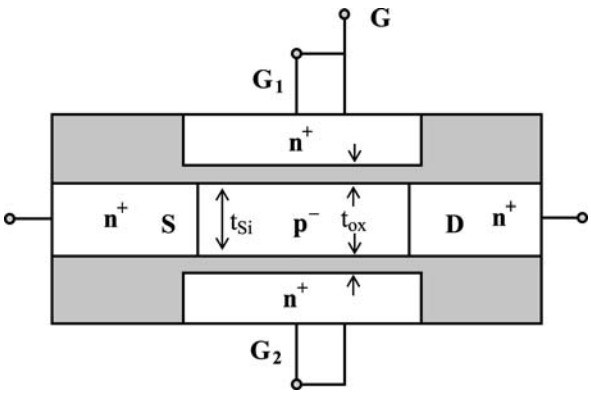

Quantum Phenomena in Semiconductor Nanostructures, Figure 6
Schematic structure of a double gate MOSFET

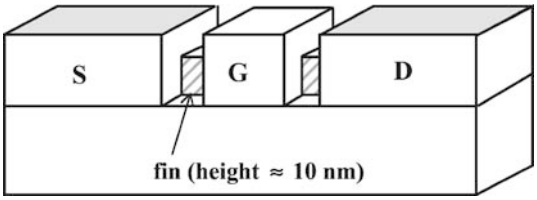

Quantum Phenomena in Semiconductor Nanostructures, Figure 7
Structure of a finFET to realize approximately a double gate MOSFET from the two vertical oxide interfaces with the silicon fin

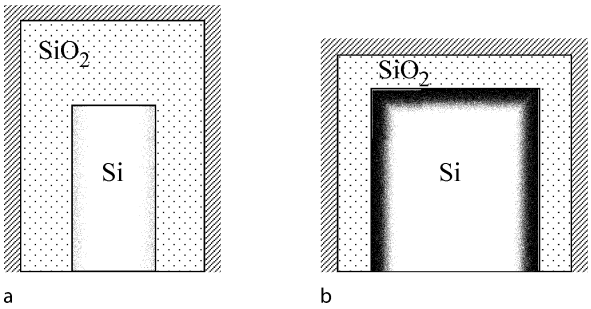

Quantum Phenomena in Semiconductor Nanostructures, Figure 8
Cross-section of conduction channels in finFET (a) and tri-gate MOSFET (b)

in the finFET of Fig. 7 [12,37,52], so that on the cross-section of the channel in Fig. 8a the electronic charge is concentrated at the two vertical interfaces where the oxide is thinner. To increase the total charge in the channel, the channel cross-section may be widened as in Fig. 8b to realized the so-called tri-gate MOSFET [69] where charge accumulates at the three interfaces controlled by the wrap-

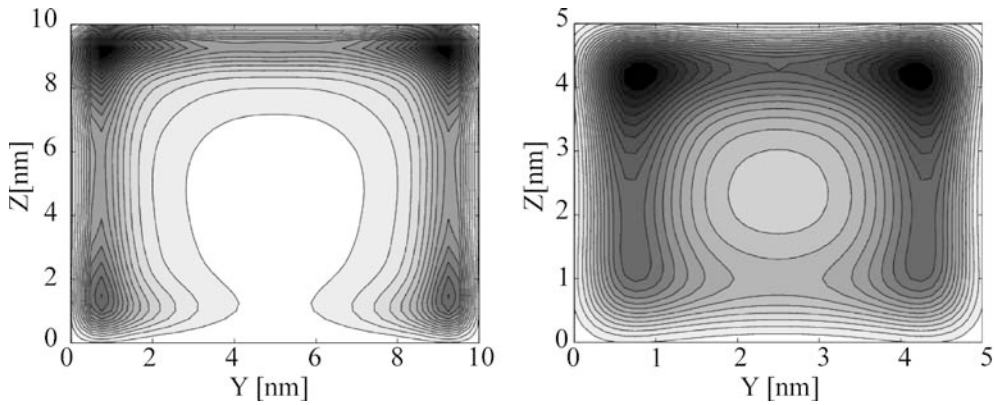

Quantum Phenomena in Semiconductor Nanostructures, Figure 9

Example of self-consistent calculation of quantum electron distribution on the square cross-section of tri-gate MOS channels with sides 10 nm (*left*) and 5 nm (*right*). One may notice the concentration peaks at the *top* corners of the structure

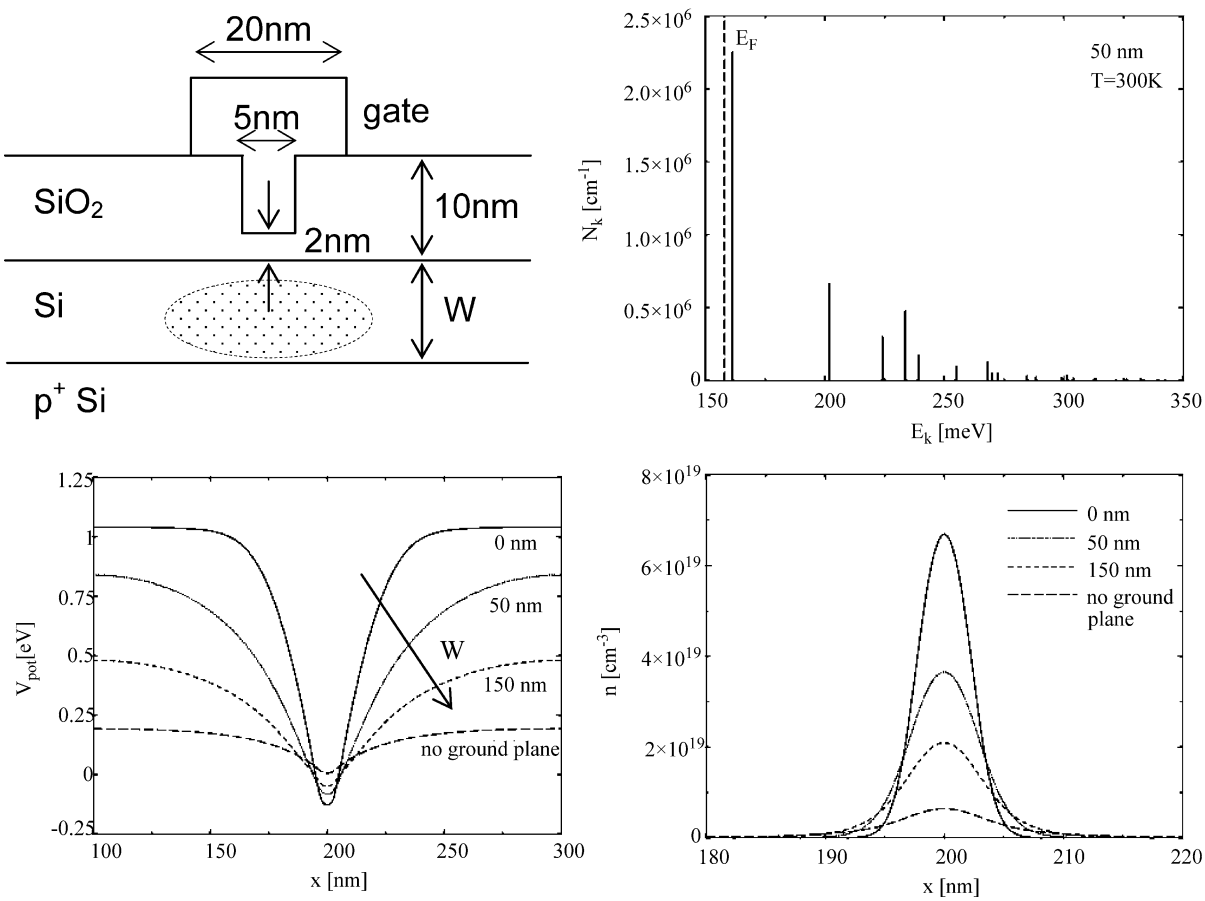

Quantum Phenomena in Semiconductor Nanostructures, Figure 10

***Top left*: quantum wire formation under a trench in the gate oxide for a structure formed with a layer of undoped silicon on top of a highly doped ground plane; *top right*: energy states in the structure with thickness of the undoped silicon layer W = 50 nm; *bottom left*: trench confining potential as function of layer of thickness W; *bottom right*: corresponding total charge densities in the quantum well. Gate voltages are 1.6 V, 1.0 V, 0.7 V and 0.3 V for increasing W. The metal gate is assumed to be aluminum**

around gate structure. Examples of charge distribution obtained on the cross-section by 2-D self-consistent solution of the Schrödinger/Poisson problem are shown in Fig. 9, for tri-gate square cross-sections with sides of 10 nm and 5 nm. Peak concentration of charge close to the top corners of the structure is noticeable.

A simpler planar structure may be realized instead with a gate contact embedded in a trench of the oxide (Fig. 10) to form a quasi-1-D quantum wire, as studied in [24]. This structure may be advantageous because of reduced interface scattering and simpler fabrication, but carriers are confined in the channel by an electrostatic potential, rather than physical interface walls, so that there is less channel isolation. Figure 10 shows several examples of the potential well formed by the trench structure of the gate for various thicknesses W of the undoped silicon layer placed above a highly p-doped ground plane.

Inclusion of Realistic Band Structure

For high carrier concentrations or high confinement barriers in a conduction channel, the range of energies necessary to account for size quantization may be much above the conduction band edge, so that a simple parabolic description of the band structure is questionable. The next approximation level is to describe the deviation from the simple behavior with the addition of non-parabolic terms. There are various possible approaches which can be followed to improve the description of the band structure. A general expression provides the energy as a series expansion

$$E = a_0 + a_2 k^2 + a_4 k^4 + a_6 k^6 + \cdots = \sum_{i=0}^{\infty} a_{2i} k^{2i} \,. \quad (25)$$

Examples in the literature simply add the fourth order term to include nonparabolic effects in the Schrödinger equation [15,53]. Other attempts have tried to define an energy-dependent effective mass, but such a formulation is in general not a sound approach and may lead to erroneous results [53]. Another commonly followed approach is to represent the wave vector instead as a series expansion of energy

$$\frac{\hbar^2 k^2}{2m^*} = E\left(1 + \alpha E + \beta E^2 + \gamma E^3 + \cdots\right) \,. \quad (26)$$

Most commonly, the expression is truncated to include only up to the quadratic energy term, leading to the standard nonparabolic approximation widely used in the literature

$$\frac{\hbar^2 k^2}{2m^*} = E\,(1 + \alpha E) \,, \quad (27)$$

where α is the coefficient of nonparabolicity with dimensions of an inverse energy. Application of this simple nonparabolic form into the Schrödinger equation for the case of a 2-DEG system was presented in [43].

The nonparabolic dispersion relation can be easily inverted by solving the quadratic equation for energy and then applying a series to the resulting square root

$$\begin{aligned} E &= \frac{1}{2\alpha}\left[\sqrt{1 + 4\alpha\frac{\hbar^2 k^2}{2m^*}} - 1\right] \\ &= \frac{1}{2\alpha}\sum_{n=1}^{\infty}\binom{\frac{1}{2}}{n}\left[4\alpha\frac{\hbar^2 k^2}{2m^*}\right]^2 \,. \end{aligned} \quad (28)$$

If we then substitute the crystal momentum k by the momentum operator $i\nabla$ we may write the Schrödinger equation in the form [23,43]

$$\frac{1}{2\alpha}\sum_{n=1}^{\infty}\binom{\frac{1}{2}}{n}\left[-4\alpha\frac{\hbar^2}{2}\right]^n\left[\frac{\nabla^2}{m^*}\right]\psi(\vec{r}) + U(\vec{r})\psi(\vec{r}) = \varepsilon\psi(\vec{r}) \,. \quad (29)$$

One should realize that if the dispersion relation is not strictly parabolic, the effective mass does not contain complete information on the lattice periodic potential effects and that the eigenfunctions are not strictly approximated by plane waves. However, rather than modifying the eigenfunctions, this procedure chooses to modify the form of the Schrödinger equation. In order to proceed, an ansatz is necessary on the form of the wave function, by associating plane waves to the coordinates of unrestricted motion. For instance, if a 1-D confining potential is present along the z-direction, one can write explicitly

$$\psi(\vec{r}) = \mathrm{e}^{i(k_x x + k_y y)}\xi(z) \,, \quad (30)$$

or for double confinement on the (y, z) plane

$$\psi(\vec{r}) = \mathrm{e}^{i\,k_x x}\zeta(y)\xi(z) \,. \quad (31)$$

The forms used for the wave function contain an implicit assumption that the symmetry axes of the effective-mass tensor are along the Cartesian coordinate axes. The energy can be split into components that individually still verify the nonparabolic dispersion relation along the directions of unrestricted motion. For single confinement one can use $\varepsilon = \varepsilon_z + \varepsilon_\parallel$ with

$$\varepsilon_\parallel(1 + \alpha\varepsilon_\parallel) = \frac{\hbar^2}{2}\left[\frac{k_x^2}{m_x} + \frac{k_y^2}{m_y}\right] \,, \quad (32)$$

leading, after several manipulations, to the modified Schrödinger equation [43]

$$\frac{1+2\alpha\varepsilon_{\parallel}}{2\alpha}\sum_{l=1}^{\infty}\binom{\frac{1}{2}}{l}\left(-4\alpha\frac{\hbar^2}{2m_z\left(1+2\alpha\varepsilon_{\parallel}\right)^2}\right)^l \cdot\frac{\partial^{2l}\xi(z)}{\partial z^{2l}}+U(z)\xi(z)=\varepsilon_z\xi(z)\,. \quad (33)$$

For a quantum wire system which is doubly confined, there is no straightforward way to develop a general model for nonparabolicity. An approximate solution is to extend the procedure above for 2-DEG, by decomposing the solution space into two transverse directions and solving two 1-D quantization problems. Following a similar procedure as for 1-D confinement, one obtains the analogous modified equation [23]

$$\frac{1+2\alpha\varepsilon_{\parallel}}{2\alpha}\sum_{l=1}^{\infty}\binom{\frac{1}{2}}{l}\left(-4\alpha\frac{\hbar^2}{2\left(1+2\alpha\varepsilon_x\right)^2}\right)^l \cdot\sum_{q=0}^{l}\binom{l}{q}\left[\frac{1}{m_y^{l-q}}\frac{\partial^{2(l-q)}\zeta(y)}{\partial y^{2(l-q)}}\frac{1}{m_z^q}\frac{\partial^{2q}\xi(z)}{\partial z^{2q}}\right] +U(y,z)\zeta(y)\xi(z)=\varepsilon_{yz}\zeta(y)\xi(z)\,. \quad (34)$$

Both equations recover the usual parabolic form in the limit $\alpha\to 0$. It is of course very difficult to use directly the modified forms of the Schrödinger equation because they involve infinite series. However, one can derive a dispersion relationship if appropriate test functions are formulated for given boundary conditions. For example, considering double confinement, we have for $\varepsilon_{yz}>U(y,z)$ (oscillatory condition for electron energy above the conduction band edge)

$$\zeta(y)=Ae^{ik_yy}+Be^{-ik_yy}\,,\quad \xi(z)=Ce^{ik_zz}+De^{-ik_zz}\,, \quad (35)$$

and for $\varepsilon_{yz}<U(y,z)$ (evanescent condition with decaying behavior)

$$\zeta(y)=Ae^{k_yy}+Be^{-k_yy}\,,\quad \xi(z)=Ce^{k_zz}+De^{-k_zz}\,, \quad (36)$$

with dispersion relationship (choosing "+" for oscillatory and "−" for evanescent)

$$\pm\left[\varepsilon_{yz}-U(y,z)\right]\left\{1+2\alpha\varepsilon_x\pm\alpha\left[\varepsilon_{yz}-U(y,z)\right]\right\} =\frac{\hbar^2}{2}\left(\frac{k_y^2}{m_y}+\frac{k_z^2}{m_z}\right)\,. \quad (37)$$

In order to get a numerical solution, one can use a shooting method. First, a value of the transverse energy ε_{yz} is selected as a guess. Then the dispersion relation is applied to test whether the corresponding wave function satisfies the necessary conditions as number of zeros and derivative sign, within a specified accuracy for satisfying the boundary conditions considering continuity and probability current of the envelope wave function.

The full numerical band structure may be included in a 3-D treatment of the Schrödinger equation [70]. This method is useful for relatively large structures for which an atomistic model is impractical while at the same time a simple analytical band approximation is inadequate. The kinetic energy of carriers can be described by a function $E(\vec{p})$ which may not be available in analytical form but only as an interpolated table, as is common for the output of band structure calculations. The usual quantization rules still apply and one may replace the momentum $\vec{p}$ by the momentum operator $\hat{p}$ to obtain the Hamiltonian. The Schrödinger equation for envelope wave functions, describing carrier motion under the influence of an external potential $V(\vec{x})$, can be written as

$$i\hbar\frac{\partial\psi(\vec{x},t)}{\partial t}=E(\hat{p})\psi(\vec{x},t)+V(\vec{x})\psi(\vec{x},t) \qquad \hat{p}=\frac{\hbar}{i}\nabla\,. \quad (38)$$

This model is reasonable as long as $V(\vec{x})$ does not vary too rapidly within a lattice constant a and the dimensions of the device structure are much larger than a as well.

It is not straightforward to solve an equation of this kind, because the kinetic energy operator may contain high order powers of the momentum. Since $\hat{p}$ is proportional to the gradient vector in the position representation, this means that one has to solve an equally high-order differential equation. Numerical solutions based on finite differences or finite elements require a Taylor expansion of $E(\vec{p})$, for instance, at least up to an order $n+1$ in $\vec{p}$ to compute $E(\vec{p})$ to order n. While it is numerically possible to implement such a high order scheme, the approach becomes inefficient for large n, and depending on the functional form of $E(\vec{p})$ the convergence radius of its Taylor series may be too small for polynomial approximations.

Since the kinetic energy operator $E(\vec{p})$ is diagonal in momentum space, a spectral solution approach using Fourier transforms seems a natural choice for this problem. Starting from the equation in position space, we may insert Fourier transforms around the kinetic energy operator, which then reduces to simple multiplication in momentum space. The Schrödinger equation has the form

$$i\hbar\frac{\partial\psi(\vec{x},t)}{\partial t}=\mathrm{FT}^{-1}\left[E(\vec{p})\mathrm{FT}\psi(\vec{x},t)\right]+V(\vec{x})\psi(\vec{x},t), \quad (39)$$

where one could view

$$E(\hat{p})\psi(\vec{x},t) = \mathrm{FT}^{-1}\left[E(\vec{p})\mathrm{FT}\psi(\vec{x},t)\right] , \tag{40}$$

as the definition for the operator $E(\hat{p})$ in position space. An alternative approach could be to express the Schrödinger equation in momentum space by inserting Fourier transforms around the potential energy as

$$i\hbar\frac{\partial\psi(\vec{p},t)}{\partial t} = E(\vec{p})\psi(\vec{p},t) + \mathrm{FT}\left[V(\vec{x})\mathrm{FT}^{-1}\psi(\vec{p},t)\right] , \tag{41}$$

which provides a Hamiltonian similar to those employed in density functional calculation. In both position and momentum formulation the numerically difficult problem of applying a high-order kinetic energy operator on a wave function has now been reduced to a more manageable calculation of Fourier transforms.

Electronic structure calculations are usually formulated in momentum space and device simulation in position space. It is not clear which of the two formulations might lead to a numerically superior algorithm for the problem at hand, but position space seems to be a more natural choice. In any case, fast Fourier transforms must be used for an efficient numerical implementation, involving only $O(N\log N)$ computational steps on a grid with N nodes, as compared to $O(N^2)$ for other methods.

The use of fast Fourier transforms imposes restrictions on the choice of computational grid. Grid lines should be equidistant in each coordinate direction and for most available fast Fourier transforms the number of grid lines should be a power of 2 in each direction as well. If the grid is too coarse, only parts of momentum space are sampled providing insufficient spatial resolution, as one would readily expect. However, also an excessively fine grid creates problems as well, since one would obtain a computational first Brillouin zone larger than the lattice one, opening the question of how to deal numerically with the states at high momentum values outside the lattice Brillouin zone, which would create spurious solutions if mapped inside the zone. Although this approach allows one, in principle, to account for general band structures, if a parabolic band is acceptable in the quantum problem at hand, it is always more efficient to use a traditional finite difference solver for large problems. Model calculations using this generalized band approach in 3D silicon quantum cavity structures were demonstrated in [70] using a numerical tabulation of a nonparabolic ellipsoidal valley.

Brief Survey of Quantum Transport Models

While size quantization models have been well developed, the treatment of transport remains a complicated quantum problem even in the assumption of envelope wave function under the effective mass approximation. In the most elementary applications, the single particle time-dependent Schrödinger equation as given in Eq. (38) can be used to probe ballistic transport in a nanostructure, assuming a parabolic band [7,25]. Coupled with appropriate conditions to simulate open contacts, one can study transients and switching between ON and OFF states in a quantum structure [59]. The standard approaches that can be used apply finite differences with a direct discretization in time [7,25,41,45,47,59,62] or the split-operator technique [79] based on a spectral method requiring Fast Fourier Transform at each step. The main limitation of the time-dependent Schrödinger equation is that it can treat a single energy at a time and that is naturally suitable for ballistic transport since scattering cannot be easily included, unless a set of coupled equations is set up. Simulation of contacts for an open system requires the numerical development of absorbing boundary conditions, which is not a trivial problem [41,45,47,59,62,68,79].

A more sophisticated transport model that includes a set of states in the formulation is based on the density matrix [48], the evolution of which is governed by the Liouville–von Neumann equation. For a simple parabolic band single-particle Hamiltonian, the evolution equation has the form

$$i\hbar\frac{\partial\rho}{\partial t} = -\frac{\hbar^2}{2m^*}\left(\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial x'^2}\right)\rho + \left[U(x) - U(x')\right]\rho , \tag{42}$$

with the density matrix $\rho(x,x')$ defined as a summation of the real-valued probabilities p_i over a complete set of states i in the system to be studied

$$\rho\left(x,x'\right) = \sum_j p_j \left\langle x|j\right\rangle \left\langle j|x'\right\rangle . \tag{43}$$

Note that, for transport through an open system, a reasonable normalization of the density matrix is not to equate the trace to one, but to relate it to the actual particle density in the system, giving also an intuitive picture of the density matrix. Another approach uses instead the Wigner distribution function [5,46] which is a mathematical transform (Wigner–Weyl transformation) of the density matrix. Starting from the density matrix the Wigner function

is defined as

$$f_W(r,p) = \int_{-\infty}^{\infty} \mathrm{d}r' \rho\left(\frac{r+r'}{2}, \frac{r-r'}{2}\right) \exp\left(\frac{-ipr'}{\hbar}\right), \tag{44}$$

where the coordinate transformations $r = (x + x')/2$ and $r' = x - x'$. The corresponding transport equation is obtained by applying the same transformation to the Liouville equation, obtaining the Wigner transport equation

$$\frac{\partial f_W}{\partial t} = \frac{p}{m^*}\frac{\partial f_W}{\partial r} - \frac{1}{\hbar}\int_{-\infty}^{\infty} \frac{\mathrm{d}p'}{2\pi\hbar} V_W(r, p-p')\, f_W(r', p'). \tag{45}$$

The equation above has similarities with the semi-classical Boltzmann transport equation and the Wigner distribution approaches the classical distribution function as the system tends to a classical state at slowly varying potentials or sufficiently high temperatures. The kernel of the potential operator is provided by the relation

$$V_W(r, p-p') = 2\int_{-\infty}^{\infty} \mathrm{d}r' \sin\left(\frac{pr'}{\hbar}\right)\left[V\left(\frac{r+r'}{2}\right) - V\left(\frac{r-r'}{2}\right)\right]. \tag{46}$$

Numerical implementation of the Wigner equation presents a number of challenges. First of all, it is a partial differential equation with hyperbolic behavior and it requires special solution techniques. Because of the wave nature of the equation, spurious numerical dispersion is difficult to contain since a range of momentum components are solved for simultaneously. Second, both position and momentum and coordinates are included in the Wigner function model, as is the case for the semi-classical Boltzmann equation, so the numerical cost increases quite rapidly for a multi-dimensional problem. On the other hand, fairly detailed scattering models can be implemented [27]. In recent applications, the Wigner function approach has also been combined with particle simulation for resonant tunneling structures [63].

An approach which has gained popularity in recent times is the nonequilibrium Green's function (NEGF) which is governed by Dyson's equation. This approach looks at a nano-system with a unified picture of a quantum channel for charge transport connecting two reservoirs held at different electro-chemical potentials [9,60,74]. The availability of numerical approaches to study a variety of systems, from nanoscale MOSFETs to molecular devices, has greatly contributed to the increase in interest in NEGF. Even an introductory treatment of NEGF would be rather cumbersome and beyond the scope of this work. The interested readers are directed to consult the relevant literature and the material available on line, as detailed in a section below. The main issue with a Green's function approach is that practical applications are possible under carefully defined assumptions. While the approach is potentially the most general and powerful to treat dissipative transport, the formalism and the computational cost may quickly become unmanageable in general conditions.

For a nanoscale system relatively close to equilibrium, a ballistic Green's function model is equivalent to the Schrödinger equation and it is quite useful to readily obtain the transmission and reflection properties of quantum wire and electron wave guide structures requiring 2-D or 3-D simulation [44,65,66]. For a nonrectilinear system with single input and output leads the formulation is very flexible, since it can be split into an equivalent 1-D Green's function problem for the transport and a size quantization problem to resolve transverse states in the cross-section of linear channel elements or cavity states for more general nonrectilinear elements. Another advantage is that application of open-boundary conditions in a Green's function problem is very straightforward, because it is very easy to define the Green's function of a semi-infinite lead. Inclusion of detailed scattering models is possible, but typically complications increase very rapidly, particularly if self-consistent treatment is applied. Reasonable simplified models are possible, for instance the approximate Büttiker probe treatment of scattering [6], which makes a NEGF formulation much less costly.

Quantum Corrections in Particle Simulation

For practical device simulation, it is still desirable to retain the mature and computationally inexpensive semi-classical techniques, extending their validity by the introduction of quantum corrections, to account for size quantization and tunneling effects. In many applications this is a reasonable alternative, since quantum coherence along the transport path is not expected to become significant or dominant for room temperature operation in MOS devices, until channel lengths are scaled down to the sub-10 nm range. The particle Monte Carlo method solves stochastically the Boltzmann Transport Equation [16,29,33,34], by a computer experiment where particle motion is explicitly simulated as sequences of random free flights interrupted by scattering events. The complete details of the band structure may be included in tabular form and the scattering model is very detailed,

so that the description of semi-classical transport provided by a Monte Carlo simulation can be very accurate [13,19,29,42].

Physically based models, such as particle simulation based on Monte Carlo methods, are a useful starting point to add quantum corrections because they adequately describe the nonequilibrium and hot carrier phenomena which dominate present devices, and can be used to calibrate simpler models in the simulation hierarchy, such as drift-diffusion. The computationally expensive alternative to add quantum effects in Monte Carlo simulation is to resolve in detail the transverse sub-bands in a conduction channel and describe particle dynamics within the sub-bands. This requires a detailed model for intra-sub-band and inter-sub-band scattering rates [20,80], as well as 2-D to 3-D scattering mechanisms if a continuum of states is used at high energies [55]. When a complete device simulation is assembled, with contacts characterized by continuum states, one also needs to implement a model that resolves the discontinuity between quantum and continuum description at the channel entrance. Because of the considerable cost associated with a detailed sub-band Monte Carlo model, applications have been limited mainly to uniform 2-DEG simulation under constant electric field conditions [20,80]. In device simulation typical approaches use the global information from the details of quantum states to formulate a potential correction that congregates otherwise semi-classical particles in a configuration that mimics the quantum distributions. The goal is not to simulate individual particles to behave as quantum ones, but to obtain a particle distribution that globally approximates the quantum charge flow, resolving noncoherent quantum effects, like size quantization and non-resonant tunneling [17,73,76].

The motivation to pursue quantum corrections in semi-classical simulation is that full quantum transport remains impractical for many cases. Also, in most realistic device structures one may identify regions where either quantum or classical features of the transport dominate, therefore, a quantum-corrected semi-classical model is useful to treat the whole structure in a unified way. In addition, while a complete quantum simulation may be extremely expensive, there is little computational cost added by quantum corrections to the standard semi-classical Monte Carlo simulators. Quantum corrections to semi-classical simulation may follow various approaches by coupling different quantum formalisms: Feynman effective potential, Wigner transport equation, Bohm potential, Schrödinger equation.

The simplest approach to quantum correction follows the effective potential idea introduced originally by Feynman for statistical mechanics [18]. In this model, particles feel the nearby potential due to quantum fluctuations around the classical path of least action, via a nonlocal function effective potential V_{eff} obtained by convoluting a Gaussian function with the electrostatic potential

$$V_{\text{eff}}(x) = \int V(x) \mathrm{e}^{-\frac{(x-x')^2}{2a^2}} \, \mathrm{d}x' \,, \quad a = \frac{\hbar^2}{12 m k_{\mathrm{B}} T} \,. \quad (47)$$

The effective potential is very simple to implement and computationally inexpensive. In addition it is not sensitive to the intrinsic statistical noise of Monte Carlo simulation [17]. This approach works best for smooth, symmetric potentials. The parameter a describes the effective "size" of the particle and can be treated as a fitting parameter. The detailed solution next to a heterointerface is typically incorrect. Fitting may be applied to obtain the correct average displacement of the charge from an abrupt interface to correspond to the maximum of quantum charge, but the space distribution is not very accurate, because the Gaussian weight function is not suitable to resolve asymmetries. A more general effective potential is possible, which depends explicitly on the wave vector k and does not contain fitting parameters [39]. The Feynman effective potential is a particular case of this general potential correction which has the form

$$V^{\mathrm{Q}}(x,k) = \frac{1}{(2\pi)^3} \int \frac{2m^*}{\beta \hbar^2 k \cdot \xi} \sinh\left(\frac{\beta \hbar^2 k \cdot \xi}{2m^*}\right) \times \exp\left(-\frac{\beta \hbar^2}{8m^*} |\xi|^2\right) V\left(y\right) \mathrm{e}^{i\xi \cdot (x-y)} \, \mathrm{d}y \, \mathrm{d}\xi \,. \quad (48)$$

The cost of applying this correction is much higher if done self-consistently. For specific structures like the double gate MOSFET, one can show that much better agreement is obtained by using a Pearson distribution instead of a Gaussian one for the formulation of the effective potential [35].

The Wigner method was also developed as a correction to statistical dynamics [75], by writing the Wigner transport equation for the distribution function in a form that casts it as the standard Boltzmann transport equation plus a correction containing the quantum terms [73]

$$\frac{\partial f}{\partial t} + \vec{v} \cdot \nabla_{\vec{r}} f - \frac{1}{\hbar} \nabla_{\vec{r}} U \cdot \nabla_{\vec{k}} f + \sum_{\alpha=1}^{\infty} \frac{(-1)^{\alpha+1}}{\hbar 4^{\alpha} (2\alpha+1)!} \left(\nabla_{\vec{r}} U \cdot \nabla_{\vec{k}} f\right)^{2\alpha+1} = \left(\frac{\partial f}{\partial t}\right)_c \,. \quad (49)$$

At first order, one simply stops the first summation term for $\alpha = 1$ and the formula can be rewritten in a form that

closely resembles the standard Boltzmann equation [73]

$$\frac{\partial f}{\partial t} + \vec{v} \cdot \nabla_{\vec{r}} f - \frac{1}{\hbar} F_{\mathrm{qc}} \cdot \nabla_{\vec{k}} f = \left(\frac{\partial f}{\partial t}\right)_c , \tag{50}$$

where the term F_{qc} is the quantum corrected force. The particles move as if under the influence of a classical potential but following approximately equivalent quantum trajectories. By making an assumption on the distribution function, like a displaced Maxwellian, and a band structure, like the simple parabolic dispersion, one can formulate expressions for the quantum corrected force that depend explicitly on momentum. These correction forces still have some problems near sharp interfaces, where quantum effects are prominent. To alleviate this, a smooth potential approximation can be obtained by integrating the displaced Maxwellian distribution with momentum [73]. Simplified versions of the force can be obtained by assuming a thermal energy to specify the momentum. Although this formulation has solid physical foundations, the result for the corrected force depends on the mobile charge density as $\nabla^2 \ln(n)$. Since in a particle simulation the density is recovered from a temporal average of particle occupation in space, convergence is slow, particularly in small structures with few particles and the overall procedure is sensitive to simulation noise. Corrections of this type work instead very well in continuum models like drift-diffusion, where noise is not an issue. From simulation experiments it was found that for a MOS interface no fitting is required in the semiconductor region, but at the oxide interface one has to specify a finite value of the carrier concentration which becomes the fitting parameter in the procedure.

An approximate correction can also be based on Bohm's potential, deriving an effective conduction band edge [78]. If we express the wave function ψ and the eigenenergy E as

$$\begin{aligned}\psi &= R \cdot \exp\left(\frac{iS}{\hbar}\right) E = V\left(\vec{r}\right) - \frac{\hbar^2}{4m}\left[\frac{\nabla^2 P}{P} - \frac{(\nabla P)^2}{2P^2}\right] \\ &= V\left(\vec{r}\right) - \frac{\hbar^2}{2m}\frac{\nabla^2 R}{R} ,\end{aligned} \tag{51}$$

the effective conduction band edge is defined as

$$V^*\left(\vec{r}\right) \approx V\left(\vec{r}\right) - \frac{\hbar^2}{2m}\frac{\nabla^2\sqrt{n}}{\sqrt{n}} = V\left(\vec{r}\right) + V_{\mathrm{B}}^{Q}\left(\vec{r}\right) . \tag{52}$$

Using the approximate relation for carrier density $n \propto \exp\left(-V^*/k_{\mathrm{B}}T\right)$ one can obtain a 1-D effective conduction band equation, valid for conditions close to equilibrium

$$\frac{\mathrm{d}^2 V^*}{\mathrm{d}y^2} - \frac{1}{2k_{\mathrm{B}}T}\left(\frac{\mathrm{d}V^*}{\mathrm{d}y}\right)^2 = \frac{4m^* k_{\mathrm{B}}T}{\hbar^2}\left(V^* - V(y)\right) . \tag{53}$$

The equation represents a first-order quantum correction to the semi-classical BTE by taking into account the effect of carriers only occupying the quantized ground state. This is reasonable in many situations since usually, where the quantum correction is needed, density is not too high and the Fermi level is below the first excited state.

Another technique for quantum correction of particle simulation is based on a direct application of the Schrödinger equation [76]. This technique is strictly suitable for size quantization in the direction perpendicular to transport, while the other techniques can be used in the direction of transport to simulate barrier lowering in tunneling, for instance. The Schrödinger correction, however, does not have the drawbacks of the other approximations, since it is very accurate, requires no fitting parameters and is not sensitive to particle simulation noise. The procedure requires explicit solutions of 1-D or 2-D Schrödinger equations along the channel, but this can be done very efficiently on the regular grids used for particle simulation and it does not affect appreciably the overall computational time. What is lost in cost per iteration is gained in excellent convergence properties.

When the Schrödinger equation is solved at different channel locations, the self-consistent Monte Carlo potential, already available from the overall simulation, is the input while the output is a quantum density, n_{q}, obtained from the detailed sub-band structure and an assumption of quasi-equilibrium distribution on the transverse cross section. If one assumes a simple Maxwellian behavior for the quantum carrier distribution

$$n_{\mathrm{q}}(z) \propto \exp\left[\frac{-\{Vp(z) + Vqc(z)\}}{kT}\right] , \tag{54}$$

the potential quantum correction has a shape as

$$Vqc(z) = -kT\log\left(n_{\mathrm{q}}(z)\right) - V_p(z) + V_0 , \tag{55}$$

where V_0 is a reference potential where the quantum charge is zero. When the Monte Carlo simulation runs, application of the potential correction forces the shape of the quantum density onto the semi-classical particles. The actual simulated density does not need to be evaluated to formulate the correction, since only the potential solution from solving the Poisson equation is needed. The potential is normally very smooth and therefore the iteration is not affected appreciably by numerical noise.

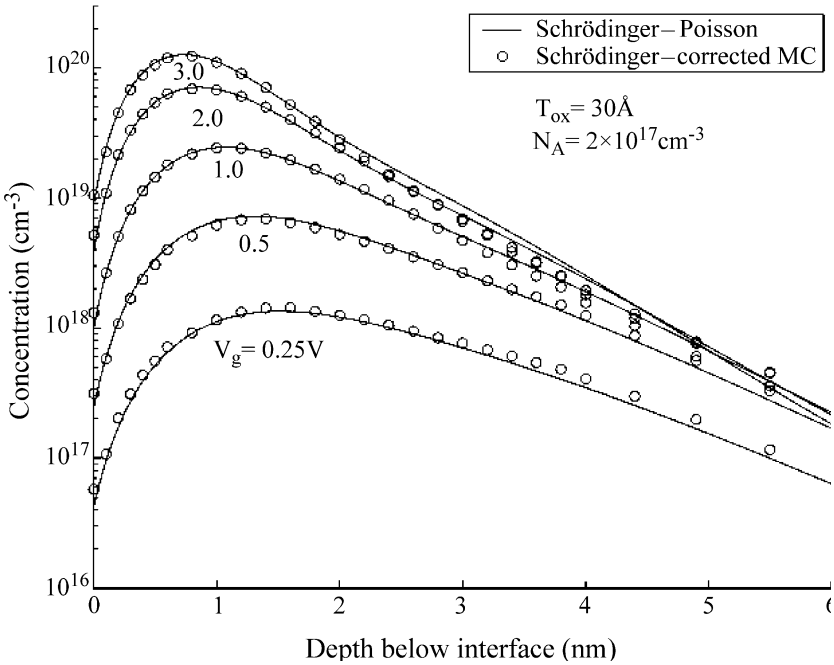

Quantum Phenomena in Semiconductor Nanostructures, Figure 11
Electron concentration in an inverted MOS capacitor calculated with the Schrödinger-corrected Monte Carlo and self-consistent Schrödinger–Poisson methods over a range of gate bias

Monte Carlo simulations of MOS capacitor structures match very closely the result of the self-consistent 1-D Schrödinger/Poisson solution, without the need for any calibration or tuning parameter, proving the overall robustness of the procedure. In the case of device simulation transport occurs in the direction of the channel, therefore, energy is acquired by particles. In this case, the actual lattice temperature should not be used to evaluate the shape of the quantum density profile on the transverse section. One can define an effective parameter with dimension of temperature which has been dubbed transverse temperature, to account for heating of the distribution. The transverse temperature is in general smaller than the temperature obtained from the average carrier energy along the channel. It should be stressed that temperature is really a tensorial quantity and that the parameter called transverse temperature cannot be identified with a well-defined physical quantity. In a structure without a clearly defined substrate reference for the potential, like in a double gate structure, the definition of transverse temperature may become ambiguous. A more detailed analysis can be carried out invoking the stress tensor along each transverse direction, to account for the variation of the electron temperature along the longitudinal direction [77]. Comparisons have been carried out with a NEGF simulator for double gate MOSFET structures. Results are in good agreement as long as the thickness of the channel layer is larger than 3 nm. For thinner layers, a quantum correction does not seem to be adequate any longer and more sub-band details should be included [57].

Monte Carlo simulations were conducted with the Schrödinger correction on MOS capacitor structures to verify the quality of the approach. Figure 11 shows a set of simulations conducted at several gate voltages, showing a consistently good agreement. With no adjustable parameter involved, the interface concentration is always accurately resolved. Figure 12 shows a set of simulation for double gate structures with separation between the oxide interfaces from 5 to 40 nm, again with excellent agreement.

Finally, representative results from a complete 3-D finFET simulation, including a 2-D quantum correction approach, are shown for a channel cross-section in Fig. 13. Potential and carrier distribution are shown both for a classical transport model (left) and with the addition of quantum corrections (right). In this particular simulation both gate and drain bias are 1 V, the side oxide thickness is 1 nm and the background acceptor concentration is 10^{16} cm^{-3}. Figure 14 shows how the quantum correction potential is added to the electrostatic potential to cause charge repulsion from the interfaces.

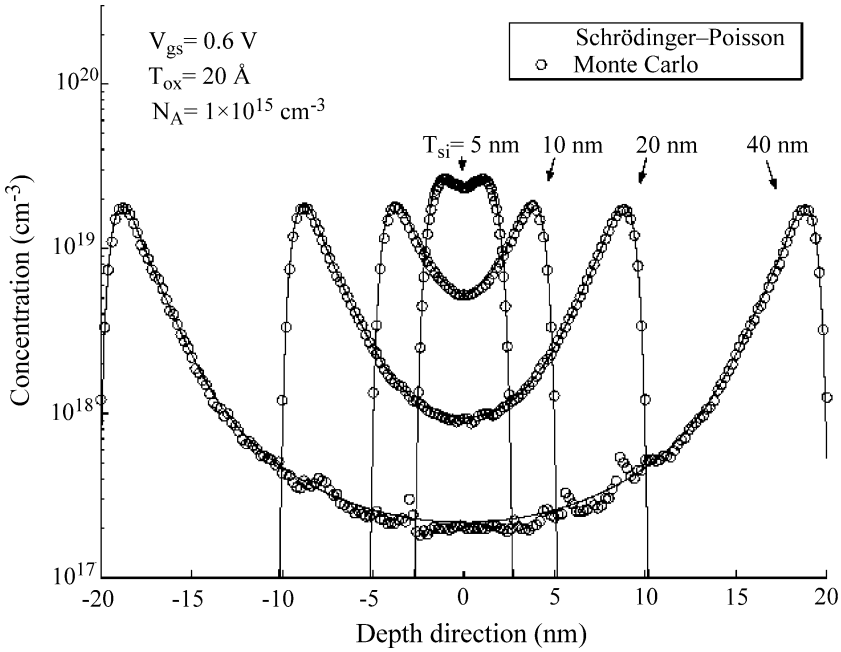

Quantum Phenomena in Semiconductor Nanostructures, Figure 12

Electron concentration in a double-gate MOS capacitor calculated with the Schrödinger-corrected Monte Carlo and self-consistent Schrödinger–Poisson methods over a range of body thicknesses

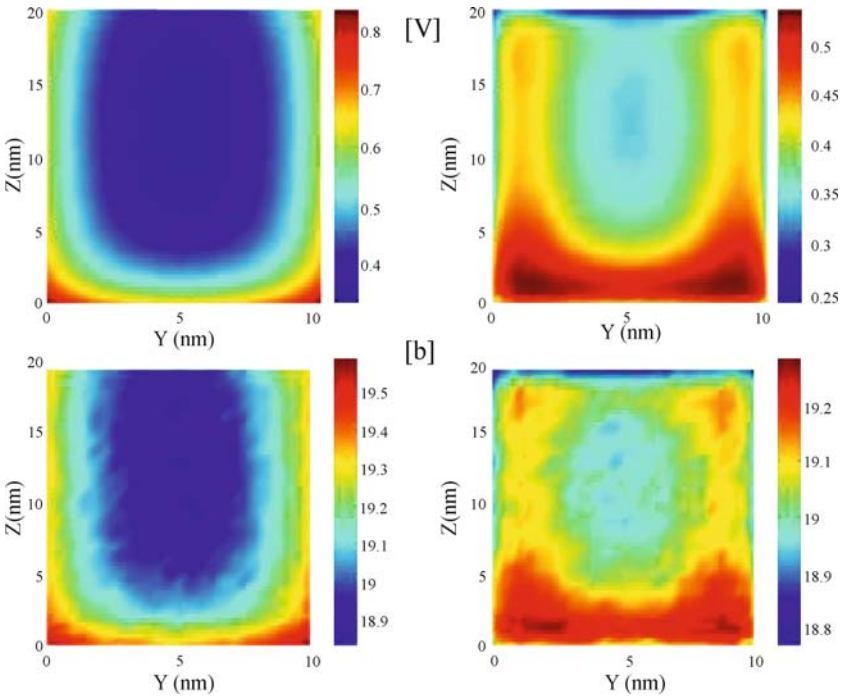

Quantum Phenomena in Semiconductor Nanostructures, Figure 13

Potential (*top row*) and electron concentration (*bottom row*) in the cross-section of a finFET as simulated by 3-D Monte Carlo simulation. Results on the left correspond to a classical simulation without quantum correction. The application of the Schrödinger quantum correction, in the results on the *right*, causes a shift of the charge concentration maximum away from the interfaces. Note that the figures are oriented upside down so that the top oxide interface corresponds to $z = 0$

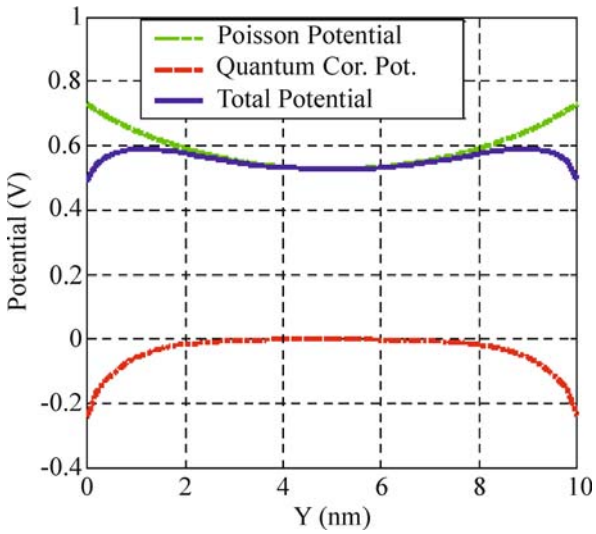

Quantum Phenomena in Semiconductor Nanostructures, Figure 14

The quantum correction potential (*red curve*) added to the electrostatic potential solution from the Poisson equation resulted in a quantum corrected potential that repels electrons from the interface

Online Resources

Over the last several years, the availability of advanced material on internet services has increased dramatically. The readers of this article will be particularly interested in exploring the many relevant resources accessible at http://www.nanohub.org, the portal of the Network for Computational Nanotechnology, an initiative funded by the National Science Foundation. The nanohub provides an ever increasing portfolio of online simulations that allow users to access fully functional research codes through comprehensive interactive interfaces. Many of the available applications in the nanoelectronics section of the nanohub relate directly to topics outlined in this work, including solvers for band structure, size quantization, quantum-corrected Monte Carlo, NEGF applied to nanodevice structures. Besides nanoelectronics, the site provides tools for nano-electro-mechanical system and nanobiology simulation. The unique approach of this organization has been to embed mature and working research codes, virtually unmodified, within a powerful infrastructure which provides at the same time a complete graphical user interface with intuitive selection menus and controls to initialize and start a simulation, access to large computational resources on efficient computer clusters, visualization of the results in the interface window with a range of facilities for data exploration. Because of this arrangement, deployment time of simulation codes has been drastically shortened, leading to availability of a vast array of applications which is expected to keep growing considerably in the future. With integration of all simulation aspects under one single interface, the nanohub services are appealing not only to computational experts in the field, but also to experimentalists and students. Beyond online computation, the nanohub is also the depository of a large number of tutorials, lectures and seminars which are regularly augmented with new contributions. The identification of specific material of interest is facilitated by flexible search tools.

Future Directions

The saturation in our ability to scale further down the dimensions of standard silicon devices for integrated circuits, which is expected to occur in the near future, opens many questions which make the formulation of future directions quite difficult. The main issue is that a replacement for standard silicon MOSFET technology has not been clearly identified, despite a flurry of activities that have explored many possibilities. Molecular systems have been studied as possible candidates for new nanotechnologies. One can find in nature a variety of stable and robust molecules that could be produced or harvested with nearly perfect uniformity of shapes and properties and which exhibit device-like behavior under external stimuli. In certain areas of electronics, organic materials have already made promising inroads, particularly for the realization of optical devices and displays [28]. Conduction through organic molecules is being studied with the goal to realize elementary computational or memory elements while a great deal of attention is also being paid to carbon nanotubes, which yield solid-state structures dominated in their behavior by the specific molecular footprint achieved during growth and which may allow a rich variety of new applications [3,40]. There is also interest in biomolecular systems. DNA molecules, which have long wire shapes, have been suggested as possible structures for computing applications [51]. Even biological ionic channels [31], that act as natural nanoscale devices in every cell and display a range of behaviors, have been examined for possible applications as switches, sensors and actuators [8]. Finally, semiconductor materials with magnetic properties are investigated to provide the ability to select the spin quantum number of carriers injected into a device [4]. Since the time decay constant of a given spin state can be much longer than characteristic device operation times, the possibility to encode information in spin states provides a new computational paradigm full of intriguing possibilities. Fun-

damental research on quantum computation in general is currently receiving considerable attention [32] but practical applications are still a matter of considerable speculation and the main limitation today is due to the fact that extremely low temperatures are still necessary to demonstrate working quantum computing applications of any kind.

While molecular systems are attractive to attempt the realization of more advanced nanoscale device elements, realistically one cannot expect a rapid technological development that may quickly provide an alternative to present semiconductor devices. One should expect that even when intrinsic scaling limits are reached, silicon technology would continue to be refined, still playing a fundamental role in any electronics application. Even if totally new devices become practical, we are likely to see for a long period of time the emergence of hybrid technologies where new systems, made of nanoscale building blocks that realize specialized tasks, will be implanted onto host systems based on traditional semiconductor technology approaches. Systems completely realized with molecular technologies are still difficult to envision. Many efforts are focused on finding a one-to-one replacement for the MOSFET but, arguably, completely new architecture and computational paradigms are necessary to realize the promise of new potential molecular approaches. Biological or artificial (biomimetic) nano-channels may find uses as single molecule detectors in novel sensing applications, possibly operating as a combination of analog and digital computing elements.

This preamble provides some grounds to venture into a prediction of future directions. It is unquestionable that miniaturization will continue and that elementary devices will continue to seek the true ultimate physical limit, which is arguably on the order of the size of an atom. Quantum and first principle calculations will be indispensable to provide the necessary predictive power for the design of reliable systems. Today, computer aided design (CAD) is already an essential stage in the design and fabrication loop. Design of large systems would be inconceivable without the use of CAD tools for circuit simulation [14]. While it was possible in the past to fine tune a device structure by trial-and-error in the laboratory, at the nanoscale regime of today one has to rely more and more on process and device simulators to optimize and calibrate design and fabrication stages to achieve the necessary performance and reliability. As the scale of device elements decreases, so is our ability to observe and measure directly anything related to device behavior. In the future, the exploration of new device concepts will have to rely even more on new generations of multi-scale and multi-physics computational approaches, while development teams will be by necessity increasingly interdisciplinary. One may envision that the computational approaches described earlier will continue to be used and improved, but increasingly in the context of coupled physical effects, where electrical, mechanical and thermal behavior are intimately connected to determine the behavior of a complex nanoscale system. For all of these aspects, quantum mechanics will play a pivotal role, but one should expect that the standard envelope-function effective-mass picture, adopted so successfully in most semiconductor device applications until now, will have to be complemented often by an atomistic approach requiring new practical simulation tools able to perform first principle calculations at various levels of complexity, from electronic structure to molecular dynamics.

This outlook opens up more general issues for which the science community needs to be prepared. Quantum mechanics is not going to be any longer the domain of a limited number of specialists, but it will permeate a much broader range of disciplines and applications. At the same time, computational requirements will be much greater and the complexity of simulations will make it impractical for single individuals to manage complete simulation codes alone. The computational community is already actively addressing the issues of Peta-scale computing technology which is expected to debut in the next five years. This unprecedented computing power will create new difficulties in data management, visualization, and numerical strategies for efficient use of the resources, while new educational challenges will have to be addressed in order to prepare professionals who can deal with the complexity of the simulation models and of the computational environments. This is why this author firmly believes that large cyber-infrastructure initiatives in nanotechnology will be crucial to create virtual collaborative communities and to develop the necessary educational and computational resources that an academic institution or industrial laboratory may not be able to provide individually to solve the electronics problems of the future.

Acknowledgments

This work was supported by the National Science Foundation through the Network for Computational Nanotechnology grant EEC-0228390.

Bibliography

Primary Literature

1. Ando T, Fowler AB, Stern F (1982) Electronic properties of two-dimensional systems. Rev Mod Phys 54:437–669

2. Antia HM (1993) Rational function approximations for Fermi–Dirac integrals. Astrophys J Suppl 84:101–106
3. Appenzeller J, Knoch J, Martel R, Derycke V, Wind SJ, Avouris P (2002) Carbon nanotube electronics. IEEE Trans Nanotech 1:184–189
4. Awschalom DD, Samarth N, Loss D (eds) (2002) Semiconductor spintronics and quantum computation. Springer, Berlin
5. Bertoni A, Bordone P, Brunetti R, Jacoboni C (1999) The Wigner function for electron transport in mesoscopic systems. J Phys Cond Matt 11:5999–6012
6. Büttiker M (1986) Four-terminal phase-coherent conductance. Phys Rev Lett 57:1761–1764
7. Cahay M, Kreskovsky JP, Grubin HL (1989) Electron diffraction through an aperture in a potential wall. Solid-State Electron 32:1185–1989
8. Cornell BA, Braach-Maksvytis, King LG, Osman PDJ, Raguse B, Wieczorek L, Pace RJ (1997) A biosensor that uses ion-channel switches. Nature 387:580–583
9. Datta S (2000) Nanoscale device simulation: the Green's function formalism. Superlatt Microstruc 28:253–278
10. Delagebeaudeuf D, Linh NT (1982) Metal-(n) AlGaAs-GaAs two-dimensional electron gas FET. IEEE Trans Electron Dev 29: 955–960
11. Dingle R, Störmer HL, Gossard AC, Wiegmann W (1978) Electron mobilities in modulation-doped semiconductor heterojunction superlattices. Appl Phys Lett 33:665–667
12. Doyle BS, Datta S, Doczy M, Hareland S, Jin B, Kavalieros J, Linton T, Murthy A, Rios R, Chau R (2003) High performance fully-depleted tri-gate CMOS transistors. IEEE Trans Electron Dev 24:263–265
13. Duncan A, Ravaioli U, Jakumeit J (1998) Full-band Monte Carlo investigation of hot carrier trends in the scaling of metal-oxide-semiconductor field-effect transistors. IEEE Trans Electron Dev 45:867–876
14. Dutton RW, Yu Z (1993) Technology CAD: computer simulation of IC processes and devices. Kluwer Academic Publishers, Norwell
15. Ekenberg U (1989) Nonparabolicity effects in a quantum well: sublevel shift, parallel mass and Landau levels. Phys Rev B 40:7714–7726
16. Fawcett W, Boardman AD, Swain S (1970) Monte Carlo determination of electron transport properties in gallium arsenide. J Phys Chem Sol 31:1963–1990
17. Ferry DK (2000) The onset of quantization in ultra-submicron semiconductor devices. Superlatt Microstruct 27:61–66
18. Feynmann R, Kleinert H (1986) Effective classical partition functions. Phys Rev A 34:5080–5084
19. Fischetti MV, Laux S (1988) Monte Carlo analysis of electron transport in small semiconductor devices including band-structure and space-charge effects. Phys Rev B 38: 9721–9745
20. Fischetti MV, Laux SE (1993) Monte Carlo study of electron transport in silicon inversion layers. Phys Rev B 48: 2244–2274
21. Frank DJ, Laux SE, Fischetti MV (1992) Monte Carlo simulation of 30 nm dual-gate MOSFET: how short can Si go? IEDM Tech Dig 1992:553–556
22. Frensley WR (1987) Wigner function model of a resonant-tunneling semiconductor device. Phys Rev B 36:1570–1580
23. Godoy A, Yang Z, Ravaioli U, Gamiz F (2005) Effects of non-parabolic bands in quantum wires. J Appl Phys 98:013702
24. Godoy A, Ruiz-Gallardo A, Sampedro C, Gámiz F (2007) Quantum-mechanical effects in multiple-gate MOSFETs. J Comput Electron 6:145–148
25. Goldberg A, Schey HM, Schwartz JL (1966) Computer-generated motion pictures of one-dimensional quantum-mechanical transmission and reflection phenomena. Am J Phys 35: 177–186
26. Gunn JB (1963) Microwave oscillations of current in II–V semiconductors. Solid-State Commun 1:88–91
27. Han Z, Goldsman N, Lin CK (2005) Incorporation of quantum correction to semiclassical two-dimensional device modeling with the Wigner–Boltzmann equation. Solid-State Electron 49:145–154
28. Heeger AJ (1998) Light emission from semiconducting polymers: light-emitting diodes, light-emitting electrochemical cells, lasers and white light for the future. Solid State Communic 107:673–679
29. Hess K (1991) Monte Carlo Device Simulation: full band and beyond. Kluwer Academic Publishers, Norwell
30. Hess K (2000) Advanced theory of semiconductors devices. IEEE Press, Piscataway
31. Hille B (1992) Ionic channels of excitable membranes. Sinauer Associates Inc, Sunderland
32. Hirvensalo M (2004) Quantum computing. Springer, New York
33. Hockney R, Eastwood J (1981) Computer simulation using particles. McGraw-Hill, New York
34. Jacoboni C, Lugli P (1989) The Monte Carlo method for semiconductor device simulation. Springer, Wien
35. Jaud MA, Barraud S, Dollfus P, Jaouen H, Le Carval G (2007) Pearson versus Gaussian effective potentials for quantum-corrected Monte-Carlo simulation. J Comput Electron 6:19–22
36. Jensen KL, Buot FA (1989) Numerical simulation of transient response and resonant-tunneling characteristics of double-barrier semiconductor structures as a function of experimental parameters. J Appl Phys 65:5248–5250
37. Kathawala GA, Ravaioli U (2003) 3-D Monte Carlo simulations of FinFETs. IEDM Tech Dig 2003:29.2.1–29.2.4
38. Kerkhoven T, Galick AT, Ravaioli U, Arends JH, Saad Y (1990) Efficient numerical simulation of electron states in quantum wires. J Appl Phys 74:1199–1204
39. Knezevic I, Vasileska DZ, Ferry DK (2002) Impact of strong quantum confinement on the performance of a highly asymmetric device structure: Monte Carlo particle-based simulation of a focused-ion-beam MOSFET. IEEE Trans Electron Dev 49:1019–1026
40. Kong J, Franklin NR, Zhou C, Chapline MG, Peng S, Cho K, Dai H (2000) Nanotube molecular wires as chemical sensors. Science 287:662–625
41. Kuska JP (1992) Absorbing boundary conditions for the Schrödinger equation. Phys Rev B 46:5000–5003
42. Laux SE, Fischetti MV, Frank DJ (1990) Monte Carlo analysis of semiconductor devices: the DAMOCLES program. IBM J Res Develop 34:466–494
43. Lopez-Villanueva JA, Melchor I, Cartujo P, Carceller JE (1993) Phys Rev B 48:1626–1631
44. Macucci M, Galick AT, Ravaioli U (1995) Quasi-three-dimensional Green's function simulation of coupled electron waveguides. Phys Rev B 52:5210–5220
45. Mains RK, Haddad GI (1988) Time-dependent modeling of resonant-tunneling diodes from a direct solution of the Schrödinger equation. J Appl Phys 64:3564–3569

46. Mains RK, Haddad GI (1988) Wigner function modeling of resonant tunneling diodes with high peak-to-valley ratios. J Appl Phys 64:5041–5044
47. Mains RK, Haddad GI (1990) Improved boundary conditions for the time-dependent Schrödinger equation. J Appl Phys 67:591–593
48. Mizuta H, Goodings CJ (1991) Transient quantum transport simulation based on the statistical density matrix. J Phys Cond Matt 3:3739–3756
49. Moore GE (1965) Cramming more components onto integrated circuits. Electronics 38:116–119
50. Nicollian EH, Brews JR (1982) MOS (Metal Oxide Semiconductor) physics and technology. Wiley, New York
51. Paun G, Rozenberg G, Salomaa A (1998) DNA computing: new computing paradigms. Springer, Berlin
52. Pei G, Kedzierski J, Oldiges P, Ieong M, Kan ECC (2002) FinFET design considerations based on 3-D simulation and analytical modeling. IEEE Trans Electron Dev 49:1411–1419
53. Persson A, Cohen RM (1988) Reformulated Hamiltonian for nonparabolic bands in semiconductor quantum wells. Phys Rev B 38:5568–5575
54. Petroff PM, Gossard AC, Logan RA, Wiegmann W (1982) Toward quantum well wires: fabrication and optical properties. Appl Phys Lett 41:635–638
55. Ravaioli U, Ferry DK (1986) MODFET ensemble Monte Carlo model including the quasi-two-dimensional electron gas. IEEE Trans Electron Dev 33:677–681
56. Ravaioli U, Osman MA, Pötz W, Kluksdahl N, Ferry DK (1985) Investigation of ballistic transport through resonant-tunneling quantum wells using Wigner function approach. Physica B 134:36–40
57. Ravishankar R, Kathawala G, Ravaioli U, Hasan S, Lundstrom M (2005) Comparison of Monte Carlo and NEGF simulations of double gate MOSFETs. J Comput Electron 4:39–43
58. Reed MA, Randall JN, Aggarwal RJ, Matyi RJ, Moore TM, Wetsel AE (1988) Observation of discrete electronic states in a zero-dimensional semiconductor nanostructure. Phys Rev Lett 60:535–537
59. Register LF, Ravaioli U, Hess K (1991) Numerical simulation of mesoscopic systems with open boundaries using the multidimensional time-dependent Schrödinger equation. J Appl Phys 69:7153–7158
60. Ren Z, Venugopal R, Goasguen S, Datta S, Lundstrom MS (2003) nanoMOS 2.5: a two-dimensional simulator for quantum transport in double-gate MOSFET's. IEEE Trans Electron Dev 50:1853–1864
61. Sah CT (1991) Fundamentals of solid state electronics. World Scientific
62. Shibata T (1991) Absorbing boundary conditions for the finite-difference time-domain calculation of the one-dimensional Schrödinger equation. Phys Rev B 42:6760–6763
63. Shifren L, Ringhofer C, Ferry DK (2003) A Wigner function-based quantum ensemble Monte Carlo study of a resonant tunneling diode. IEEE Trans Electron Dev 50:769–773
64. Sollner TCLG, Goodhue WD, Tannenwald PE, Parker CD, Peck DD (1983) Resonant tunneling through quantum wells at frequencies up to 2.5 THz. Appl Phys Lett 43:588–590
65. Sols F, Macucci M, Ravaioli U, Hess K (1989) On the possibility of transistor action based on quantum interference. Appl Phys Lett 54:350–352
66. Sols F, Macucci M, Ravaioli U, Hess K (1989) Theory for a quantum modulated transistor. J of Appl Phys 66:3892–3906
67. Sze SM (1981) Physics of semiconductor devices. Wiley, New York
68. Talebian MA, Pötz MA (1996) Open boundary conditions for a time-dependent analysis of the resonant tunneling structure. Appl Phys Lett 69:1148–1150
69. Trellakis A, Ravaioli U (1999) Lateral scalability limits of silicon conduction channels. J Appl Phys 86:3911–3916
70. Trellakis A, Ravaioli U (2002) Three-dimensional spectral solution of the Schrödinger equation for arbitrary band structures. J Appl Phys 92:3711–3716
71. Trellakis A, Galick AT, Pacelli A, Ravaioli U (1997) Iteration scheme for the solution of the two-dimensional Schrödinger–Poisson equation in quantum structures. J Appl Phys 81:7880–7884
72. Tsu R, Esaki L (1973) Tunneling in a finite superlattice. Appl Phys Lett 22:563–564
73. Tsuchiya H, Ravaioli U (2001) Particle Monte Carlo simulation of quantum phenomena in semiconductor nanostructures. J Appl Phys 89:4023–4029
74. Venugopal R, Ren Z, Datta S, Lundstrom MS, Jovanovic D (2000) Simulating quantum transport in nanoscale transistors: real versus mode-space approaches. J Appl Phys 93: 3730–3739
75. Wigner E (1932) On the quantum correction for thermodynamic equilibrium. Phys Rev 40:749–759
76. Winstead B, Ravaioli U (2003) A quantum correction based on Schrödinger equation applied to Monte Carlo device simulation. IEEE Trans Electron Dev 50:440–446
77. Wu B, Tang TW (2004) The effective conduction band edge method of quantum correction to the Monte Carlo device simulation. J Comput Electron 3:347–350
78. Wu B, Tang TW, Nam J, Tsai JH (2003) Monte Carlo simulation of symmetric and asymmetric double-gate MOSFETs using Bohm-based quantum correction. IEEE Trans Nanotech 2:291–294
79. Yalabik MC, Ecemiş MI (1995) Numerical implementation of absorbing and injecting boundary conditions for the time-dependent Schrödinger equation. Phys Rev B 51:2082–2086
80. Yamakawa S, Ueno H, Taniguchi K, Hamaguchi C, Miyatsuji K, Masaki K, Ravaioli U (1996) Study of interface roughness dependence of electron mobility in Si inversion layers using the Monte Carlo method. J Appl Phys 79:911–916

Books and Reviews

Datta S (1995) Electronic transport in mesoscopic systems. Cambridge University Press, New York
Ferry DK, Goodnick SM (1997) Transport in nanostructures. Cambridge University Press, New York
Frensley WR (1990) Boundary conditions for open quantum systems driven far from equilibrium. Rev Mod Phys 62:745–791
Fu Y, Willander M (1999) Physical models of semiconductor quantum devices. Kluwer Academic Publishers, Norwell
Hamaguchi C (2001) Basic semiconductor physics. Springer, Berlin
Lundstrom M (2000) Fundamentals of carrier transport. Addison-Wesley, Reading
Nag BR (2000) Physics of quantum well devices. Springer, Berlin
Paiella R (2006) Intersubband transitions in quantum structures. Mc-Graw Hill, New York

Quantum Similarity and Quantum Quantitative Structure-Properties Relationships (QQSPR)

Ramon Carbó-Dorca, Ana Gallegos
Institut de Química Computacional,
Universitat de Girona, Girona, Spain

Article Outline

Glossary
Definition of the Subject
Introduction
Mathematical Background of Quantum Similarity
Quantum Similarity
Linear Quantum QSPR Fundamental Equation
Non-Linear Terms and Extended Wave Functions
QQSPR Operators, Quantum Similarity Measures and the Fundamental QQSPR Equation
Future Trends
Bibliography

Glossary

There is a brief description of the terms used herein. The defined items appear in alphabetical order. When in a definition a term already defined in the glossary is mentioned, it is written in **bold face**; then, the reader has to refer to the corresponding glossary item, where more information is given.

Carbó (similarity) index The Carbó (Similarity) Index is a **QS** index, which corresponds to the cosine of the angle subtended by the density function (**DF**) tags of any pair of quantum objects. The values of the Carbó index lie within the interval $(0, 1]$. The lower bound corresponds to a complete dissimilarity, while the unit value is encountered when comparing a quantum object with itself. *See also:* **Euclidian Distance (Similarity) Index**.

Core set A **QOS** with the additional information of a known and well-defined property value for every **QO**.

(First order) density function (DF) The first order DF is a quantum theory concept, associated to a known electronic submicroscopic system. To be understood by this term is a **wave function** squared module or the **full DF**, reduced by integration over the space and spin coordinates to a function of the space coordinates of a unique electron. As it is customary in the literature, the name will be shortened to DF. Such a function can be also obtained via direct computation within Density Functional Theory (DFT). DF can be employed as tags associated to well-defined **quantum objects**. DF are non-negative functions. According to the usual quantum mechanical interpretation, within the DF collection there is contained all the information one can extract from the system. This last statement is the basic postulate which appears in first place when developing **QQSPR** theory.

(Full) density function A quantum theoretical concept, associated to a known electronic (or other particle sets) system. It corresponds to the system's **wave function** squared module.

Density function (DF) tags The tag set, (see: **tagged set**) collecting the quantum mechanical density functions used as descriptors in a **QOS**.

DQOS *Discrete Quantum Object Set*. A **QOS** whose tags are finite dimensional vectors, whose elements, in turn, are computed as **quantum similarity measures**.

Euclidean distance (similarity) index Like the **Carbó Index** the Euclidean Distance Index is a **quantum similarity index**, involving the Euclidean distance between two **density functions**. The range of this index is $[0, \infty)$, thus their minimal values correspond to two identical **quantum objects**, while the index grows in relation to the difference in the compared **quantum objects**.

FQQSPR equation *Fundamental Quantum Quantitative Structure-Properties Relationships equation*. The subject of study and analysis of the present contribution. A non-empirical equation, which can be deduced by quantum mechanical theoretical means, serving as the basic tool to obtain **QQSAR** or *Quantum QSAR*. Here, only the linear **FQQSPR** equations are deeply discussed, as they are the main source of **QQSPR** studies, but the **FQQSPR** equation can easily be extended to any order.

Molecular descriptors Parameters of varied origin: empirical, theoretical or experimental, or functions obtained from quantum mechanical manipulations like the **Density Function** or the Electrostatic Molecular Potential, or by empirical considerations, associated to a given molecule, which represent the molecular structure environment and can be employed to obtain **QSPR**. Any set of parameters or functions, which can be used as tags in a **tagged set**, whose objects are molecules.

QO *Quantum Object*. An element of a **QOS**, a **tagged set** where every element is constructed as a composite of a submicroscopic system description (the object) and a density function (the tag). The possibility to represent all the information contained within any quan-

tum system by the full or reduced **DF**, permits one to describe such an entity formed by the structure of the system and its **DF** together; this can be called a **QO**. Plural: **QO's**.

QOS *Quantum Object Set*. A **tagged set** made of **QO's**.

QS *Quantum Similarity*: A discipline dealing with similarity measures between submicroscopic systems. Such kinds of quantum similarity measures (**QSM**) can be computed using the quantum theoretical description of such kinds of objects. According to quantum theory all the information one can obtain about a quantum system is contained in the state **DF** and the set of possible reduced forms, hence **quantum similarity** is part of this information contained in the **DF**. Usually, for computational convenience, the (first order reduced) **DF** has been employed as a universal descriptor for comparison purposes and it is the employed tag of a **QO**. Similarity comparisons become possible by means of comparing the **QO DF** tags.

QSAC *Quantum Similarity Aufbau Condition*: A condition that the **QS Matrix** has to comply with in order to be positive definite and admit Choleski decomposition.

QSAR & QSTR QSAR are **QSPR** employed to estimate biological molecular activity values via **molecular descriptors**. When **QSPR** are employed to estimate molecular toxicity, they can be called QSTR.

Quantum similarity matrices (QS matrices) [Depending on the context so as not to cause confusion with the term **QS Measures** the abbreviation **QSM** can also be used to denote **QS Matrices**]: Any matrix which contains computed **QSM** results involving several **QO**. Usually the **QS Matrix** is symmetric and square when the **QSM** on the involved elements of a unique **QOS** are ordered in pairs. In this case the **QSAC** must hold. The **QS Matrix** can be rectangular when computing and ordering into a matrix the **QSM** of two different **QOS**. **QS Hypermatrices** appear when higher order **QSM** associated to more than two **DF** are involved.

Quantum similarity measures (QSM) and indices (QSI) A positive definite integral, computed employing the **QO DF tags** as integrands and, if necessary, including a positive definite operator which can be supposed to act as a weight. Possessing the structure of a measure, such an integral can be interpreted as a generalized volume. **QSM** between two or several **QO** correspond to integrals, which can be constructed primarily with integrands made with the product of the density tags of the compared **quantum objects** plus a positive definite operator. Such a definition ensures the positive definite nature of the **QS** integrals. A quantum *self-similarity* measure corresponds to the **QSM** computed with the **density function tags** of a unique **QO**. **QSI** are manipulations of the **QSM** in order to obtain **QS** comparisons within an adequate range.

QSPR *(Classical) Quantitative Structure-Properties Relationships*. This term refers to any empirical relationship permitting the connection between molecular structure, represented by a set of parameters (**molecular descriptors**) of any origin and molecular properties. Usually, by a classical QSPR can be understood a non-causal multilinear relationship, obtained via statistical reasoning and procedures. This is the generic name given to any functional (usually linear) connecting the properties of a molecule with the attached **molecular descriptors**. The QSPR functionals are empirically obtained by statistical analysis, usually employing (non-)linear regression techniques or any variant of it.

QQSPR *Quantum Quantitative Structure-Properties Relationships*. These are *non-empirical* functionals derived from the structure of a **FQQSPR equation**. Thus, this kind of relationship, if it exists, to some extent can be considered *causal*. A QQSPR permits us to compute the molecular properties of **U-molecules**, just employing non-empirical considerations and parameters or **molecular descriptors** of quantum mechanical origin. In this sense, these relationships can be considered *universal*, applicable to any molecular structure. However, due to the quantum mechanical nature of the QQSPR, one can obtain structure-properties relationships between **QO's** of any kind: nuclei, atoms, molecules. By QQSPR are here understood **QSPR** obtained by means of the application of canonical quantum theoretical methods to obtain expectation values of a Hermitian operator, which within QQSPR theory is described by functionals of the **density function tags** of a given **QOS**, acting as a basis set. QQSPR are more general than classical **QSPR** as they can be obtained for any **QO**, incorporating information difficult to take into account with classical **molecular descriptors**. The descriptors in the QQSPR framework are the **DF tags** of the **QO**.

QS matrix *Quantum Similarity Matrix*. The matrix possessing positive definite elements, constructed by **quantum similarity measures** using the **DF tags** of the elements of a **QOS**. *See:* **Quantum Similarity Matrices**.

Tagged set A set constructed as the Cartesian product of two separate sets. One of them, the *object set* is composed of well-defined elements of any nature, called

objects. The other set in the tagged set is the *tag set*, made of some *descriptors* attached to every object, which is supposed to contain information about the objects; tags can be mathematical elements made by bit strings, vectors, matrices, functions…

U-m *Unknown (properties) molecule*. A molecule which can be described as a **QO**, but lacks the property information associated to the molecules belonging to the **Core Set**. U-m properties act as the unknowns to be evaluated in the **QQSPR** theory developed here.

U-molecule A **U-m**.

U-m set A **QOS** containing **U-m's** as elements.

Wave function A by-product of solving the Schrödinger equation. When studying stationary submicroscopic systems by means of classical quantum mechanics, the pair energy – wave function correspond to the eigenvalue – eigenfunction pairs of the Hermitian Hamiltonian operator constructed for the system, which substitutes the classical Hamilton function. The squared module of the stationary wave functions for each system state is customarily interpreted, since Born times, as the probability density function to find the system as a whole in some space infinitesimal element of volume.

Definition of the Subject

The concept of Quantum Similarity (QS) was introduced for the first time in 1980 in a paper by Carbó et al. [1] entitled: *How Similar is a Molecule to Another?* There the basic aspects of the theory were set up. The backbone of QS was constituted by the conceptual support of the QSM.

QSM. A QSM between two quantum objects (QO): associated to the density function (DF) tags: $\{\rho_A, \rho_B\}$ was defined as the density overlap integral:

$$Z_{AB} = \int_D \rho_A(\mathbf{r})\rho_B(\mathbf{r})\,\mathrm{d}\mathbf{r} = \langle \rho_A \rho_B \rangle = \langle \rho_B \rho_A \rangle = Z_{BA}\,,$$

which is always a positive real number, because the involved DF are non-negative definite real functions. Self-similarity measure integrals were defined in the same manner:

$$I = A, B\colon Z_{II} = \int_D \rho_I(\mathbf{r})\rho_I(\mathbf{r})\,\mathrm{d}\mathbf{r} = \left\langle \rho_I^2 \right\rangle .$$

QS Matrix. A simple example, which remains formally valid for larger QO sets, can illustrate the basic procedures of QS. For a set of two QO a symmetric QS matrix can be set up:

$$\mathbf{Z} = \begin{pmatrix} Z_{AA} & Z_{AB} \\ Z_{BA} & Z_{BB} \end{pmatrix} = \mathbf{Z}^{\mathrm{T}} \leftarrow Z_{AB} = Z_{BA}\,.$$

The columns (or rows) of the QS matrix:

$$|z_A\rangle = \begin{pmatrix} Z_{AA} \\ Z_{BA} \end{pmatrix} \wedge |z_B\rangle = \begin{pmatrix} Z_{AB} \\ Z_{BB} \end{pmatrix}$$

can be interpreted here as the construction of a two dimensional discrete representation of the DF tags pair, quantum mechanically associated to the two involved QO. In this way one can write the association:

$$\forall I = A, B\colon \rho_I \leftrightarrow \|z_I\rangle\,.$$

QSI. From the QS matrix elements several kinds of QS indices (QSI) can also be described. Two indices were described in the seminal paper:

A) Carbó Index; defined as a cosine of the angle subtended by the pair of DF tags $\{\rho_A, \rho_B\}$:

$$R_{AB} = Z_{AB}(Z_{AA}Z_{BB})^{-\frac{1}{2}} \in [0;1] \rightarrow R_{AB} = R_{BA}\,.$$

B) Euclidean Distance Index; constructed with the well-known form:

$$\begin{aligned} D_{AB}^2 &= \int_D |\rho_A(\mathbf{r}) - \rho_B(\mathbf{r})|^2\,\mathrm{d}\mathbf{r} \\ &= Z_{AA} + Z_{BB} - 2Z_{AB} \in [0;+\infty] \\ &\rightarrow D_{AB} = D_{BA}\,. \end{aligned}$$

Ordering a QOS and Mendeleyev Conjecture. QSM or QSI, computed as previously described on the elements of a Quantum Object Set (QOS), are sufficient to permit ordering the elements of the set; thus, opening the way to construct non-artisan periodic tables of molecular sets, for instance.

The possibility to order QOS by means of their DF tags opened the way to estimate unknown properties of some QO; provided a set of QO with known properties was previously ordered. This has led to the path towards Mendeleyev conjecture, described by Carbó and Besalú [2] in 1996.

QQSPR. Mendeleyev conjecture opened the way to employ QS techniques in order to obtain QSPR (Quantitative Structure-Properties Relationships), providing the necessary background for the description of quantum QSPR (QQSPR); a new kind of QSPR functionals that possess the properties to be: (1) *Universal*: as it can be employed to study any QOS. (2) *Unbiased*: as the user cannot choose any other QO descriptor than the DF tag. (3) *Causal*: as the QQSPR functionals are based on a non-empirical equation, derived from the application of quantum theoretical methodology.

Aims. This contribution pretends to provide a mathematical basis for the understanding of the quantum QSPR problem, which tries to find out how to construct universal, unbiased and causal QSPR models. In turn these QSPR models can be employed to predict complex (in the sense of complicated observables) molecular properties. The ultimate purpose of such a theoretical framework is aimed at overcoming the fact that in previous applications molecular quantum similarity numerical values, ordered in the form of similarity matrices, have been employed just like molecular descriptors within a classical QSPR computational way. The future of molecular quantum similarity must be foreseen within the description and further development of an autonomous QQSPR set of computational procedures.

Introduction

Quantum Object Sets and Core Sets. QSPR studies are based on some set of molecules: M, attached to a collection of descriptors and properties; the whole is the *core set*, symbolized as: C. For all the elements of the set M, via the Schrödinger equation for every molecule in the *core set* a wave function can be computed, providing in turn a set of density functions: $\mathrm{P} = \rho_I$, which can be considered as unique continuous molecular descriptors, by means of the quantum mechanical interpretation [3,4].

According to this it can be considered that:

$$\forall m_I \in \mathrm{M} \rightarrow \exists \rho_I \in \mathrm{P} \wedge \forall I\colon m_I \leftrightarrow \rho_I \,.$$

Therefore, quantum similarity theory [5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22], permits us to perform a Cartesian product of the M and P sets, which is used to build up a *tagged set* [23] $\mathrm{Q} = \mathrm{M} \times \mathrm{P}$, named a *quantum object set* (QOS). In a QOS the molecules constitute the *object set* and the density functions the *tag set* [23,24,25]. The elements of a QOS are *quantum objects* (QO). QO ordered pairs, are constructed in the following way:

$$\forall m_I \in \mathrm{M} \wedge \forall \rho_I \in \mathrm{P} \rightarrow \forall \omega_I \in \mathrm{Q}\colon \omega_I = (m_I; \rho_I)\,. \quad (1)$$

Then the *core set*, C is a well-defined QOS. Because of the QOS definition in Eq. (1), C can have the form:

$$\forall \omega_I \in \mathrm{Q} \wedge \exists \pi_I \in \Pi$$
$$\rightarrow \forall c_I \in \mathrm{C} = \mathrm{Q} \times \Pi : c_I = (\omega_I; \pi_I) \equiv (m_I; \rho_I; \pi_I),$$

Π contains the properties of some elements of M. Hence, C elements are triples made of molecular structures, density functions and properties: $\mathrm{C} = \mathrm{M} \times \mathrm{P} \times \Pi$.

In classical QSPR density functions are replaced by finite dimensional vectors, whose elements are the so-called *molecular descriptors*. Construction of the *core set* with discrete vector spaces, substituting P, will also appear within the QQSPR. The substitution of the continuous density tags by discrete vectors in QQSPR has a mathematical-theoretical meaning, while in empirical QSPR this remains arbitrary. The elements of the *core set* C are core molecules, C-molecules or briefly C-m.

QQSPR Operators, Quantum Similarity Measures and the Fundamental QQSPR Equation. Correspondence principle provides the rules to construct Hermitian operators, with expectation values associated to the experimental outcomes of submicroscopic systems observables [3,4]. For some complex (complicated) observables, like biological activities, the correspondence principle cannot be applied, as Hermitian operators are unavailable or difficult to be obtained. The QQSPR operators and the attached fundamental QQSPR equation, create an approximate quantum mechanical computational environment in order to estimate the expectation values of complicated observables.

QQSPR Operators. The fundamental QQSPR equation arises when density function tags: $\{\rho_I(\mathbf{r})\}$ of some QOS are used to construct a QQSPR operator as:

$$\Omega(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \ldots) = x_0 \Theta_0(\mathbf{r}_1) + \sum_I x_I \rho_I(\mathbf{r}_2) \Theta_1(\mathbf{r}_1, \mathbf{r}_2)$$
$$+ \sum_I \sum_J x_I x_J \rho_I(\mathbf{r}_2) \rho_J(\mathbf{r}_3) \Theta_2(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) + O(3) \quad (2)$$

in Eq. (2), x_0 is an arbitrary constant; $\{\Theta_\omega(\mathbf{R}) | \omega = 0, 1, 2, \ldots\}$ is a known positive definite operator set, acting as a weight set; and $\{x_I\}$ is a set of parameters, determined through the fundamental QQSPR equation.

The structure of a QQSPR operator (2) has to be seen as a first step algorithm permitting us to define approximate quantum mechanical operators. The QQSPR operators can be employed afterwards to evaluate their quantum mechanical expectation values.

Expectation Values of the QQSPR Operator. To determine the parameter set $\{x_I\}$, it is necessary to compute the set of expectation values over the elements, the core molecules or C-m, of the core set C. Besides a well-defined structure and a known density function, as members of a QOS, the C-m possess a known property value of the set: $\Pi = \{x_I\}$, attached to each one.

Then, every known property of the C-m elements can be expressed as an expectation value of a QQSPR operator:

$$\forall m_K \in \mathrm{C}\colon \pi_K \approx \langle \Omega \rho_K \rangle = x_0 \langle \Theta_0 \rho_K \rangle + \sum_I x_I \langle \rho_I \Theta_1 \rho_K \rangle + \sum_I x_I x_J \langle \rho_I \rho_J \Theta_2 \rho_K \rangle + O(3) \,. \tag{3}$$

Zero-th Order Term. In the expectation values (3) of the elements of C, the Zero-th order term is:

$$\theta_K[\Theta_0] = x_0 \langle \Theta_0 \rho_K \rangle = x_0 \int_D \Theta_0(\mathbf{r}_1) \rho_K(\mathbf{r}_1)\, \mathrm{d}\mathbf{r}_1$$

being a constant for each C-m, the Zero-th order term: $x_0 \Theta_0(\mathbf{r})$ acts as an origin shift. Choosing: $\Theta_0(\mathbf{r}) = I$, this term becomes proportional to the number of electrons of the C-m considered:

$$\theta_K[I] = x_0 \langle \rho_K \rangle = x_0 \int_D \rho_K(\mathbf{r}_1)\, \mathrm{d}\mathbf{r}_1 = x_0 N_K \,.$$

The Zero-th order term can be omitted if it is no longer necessary to shift the property values of the C-m.

Quantum Similarity Measures in First and Second-Order Expectation Value Terms. The first-order term in Eq. (3) contains QSM integrals between pairs of C-m density function tags, long time known [15]:

$$z_{IK}[\Theta_1] = \langle \rho_I \Theta_1 \rho_K \rangle = \int_D \int_D \rho_I(\mathbf{r}_2) \Theta_1(\mathbf{r}_1, \mathbf{r}_2) \rho_K(\mathbf{r}_1)\, \mathrm{d}\mathbf{r}_1\, \mathrm{d}\mathbf{r}_2 \,,$$

and the second-order term is made of triple-density quantum similarity measures [16]

$$z_{IJK}[\Theta_2] = \langle \rho_I \rho_J \Theta_2 \rho_K \rangle = \int_D \int_D \int_D \rho_I(\mathbf{r}_2) \rho_J(\mathbf{r}_3) \Theta_2(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) \rho_K(\mathbf{r}_1)\, \mathrm{d}\mathbf{r}_1\, \mathrm{d}\mathbf{r}_2\, \mathrm{d}\mathbf{r}_3 \,.$$

The matrix symbol $\mathbf{Z}$ will be used to represent any collection of QSM: $\{z_{IJ}\}$, independently of the nature of the weighting operator Θ_1.

Quantum Similarity Matrices (QSM) in the Construction of First-Order QSPR Operators and the Definition of Discrete QOS (DQOS). The first-order approach of the QSPR operator [13,14,15,16,17], applied to the *core set* with the known property set: $\Pi = \{\pi_I\}$, produces the equation collection:

$$\forall I\colon p_I = \pi_I - \langle \Theta_0[\rho_I] \rangle \approx \sum_J x_J \langle \rho_J \Theta_1[\rho_I] \rangle = \sum_J x_J z_{JI} \,. \tag{4}$$

If $\Theta_1 = I$ is used, the first-order integrals (4) are:

$$\left\{ \langle \rho_J[\rho_I] \rangle = \int_D \rho_J \rho_I\, \mathrm{d}V = z_{JI} = z_{IJ} = \int_D \rho_I \rho_J\, \mathrm{d}V = \langle \rho_I[\rho_J] \rangle \right\} ,$$

and can be ordered into a $(n \times n)$ symmetric array, constructing in this way the so-called *quantum similarity matrix*: $\mathbf{Z} = \{z_{IJ}\}$ (QS Matrix) [18]. The property set form a column vector: $|\mathbf{p}\rangle = \{p_I\}$. Equations (4) are a linear system, which can be used to evaluate unknown molecular properties for some QOS members of the *core set*.

Every column of the QSM [19]: $\mathbf{Z} = \{|\mathbf{z}_I\rangle = \{z_{JI}\}\}$, is a discrete representation of each QO density function in: $P = \{\rho_I\}$. A one-to-one correspondence exists between the density tag set and the QSM column submatrices:

$$\forall m_I \in M\colon \rho_I \leftrightarrow |\mathbf{z}_I\rangle \Rightarrow \mathrm{P} \Leftrightarrow \mathbf{Z} \,.$$

The QSM column set can be used as a n-dimensional vector tag set, attached to the molecular set, building up a tagged set, called *discrete quantum object set* (DQOS) [19,20,21,22,26]:

$$\mathrm{Q}_Z = \mathrm{M} \times \mathbf{Z} \,. \tag{5}$$

In DQOS, the density function tags of the original QOS, Q, belonging to the tag set P, are substituted by the columns of the QSM. There also exists a one-to-one correspondence between both QOS: $Q \leftrightarrow Q_Z$.

Fundamental QQSPR Equation Setup. Expectation values of the QQSPR operator (3) can be collected in a column vector, providing the fundamental QQSPR equation:

$$|\mathbf{p}\rangle \approx \mathbf{Z}_1 |\mathbf{x}\rangle + \langle \mathbf{x} | \mathbf{Z}_2 | \mathbf{x} \rangle + O(3) \,. \tag{6}$$

In (6), $|\mathbf{p}\rangle = \{p_K\}$ is the shifted C-m properties vector: $|\mathbf{p}\rangle = |\pi\rangle - |\theta\rangle$, where $|\pi\rangle = \{\pi_I\}$ is the original property vector and $|\theta\rangle = \{\theta_K\}$ is the completely determined Zero-th order origin shift vector, $\{\mathbf{Z}_\omega | \omega = 1, 2, \ldots\}$ is a matrix set containing the quantum similarity measures, for instance: $Z_1 = \{z_{IK}\}; Z_2 = \{z_{IJK}\}; \ldots$, and: $|\mathbf{x}\rangle = \{x_I\}$ is a column vector bearing the unknown coefficients, which define explicitly the QQSPR operator (2).

The unknown coefficients $|\mathbf{x}\rangle = \{x_I\}$ can be obtained solving the linear equation contained in the fundamental QQSPR Eqs. (4):

$$|\pi\rangle - |\theta\rangle = |\mathbf{p}\rangle = \mathbf{Z}_1 |\mathbf{x}\rangle \rightarrow |\mathbf{x}\rangle = (\mathbf{Z}_1)^{-1} |\mathbf{p}\rangle \,. \tag{7}$$

Equation (7) has no predictive power. This is so because the first-order similarity matrix $\mathbf{Z}_1$ has to be chosen positive definite by construction. By predictive power it is understood the possible computation of the property value

for an also known quantum object, a U-m, which as such possesses a well-defined structure and density function, but belongs to the U set.

In the last years, since the description of QS measures for the first time [1], the predictive power of the information contained in the QS matrices set has been manipulated within the classical QSPR. This is the same as considering the similarity matrices as a source of molecular parameters to construct empirical QSPR. See references [5] and [27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46].

First-Order Fundamental QQSPR Equation. The study of QQSPR predictive potential starts with the first-order fundamental QQSPR equation, involving the *core set*, containing the associated DQOS molecules, possessing known values of some complex property.

The first-order QQSPR fundamental equation in a compact matrix form [60] is written as:

$$\mathbf{Z}|\mathbf{x}\rangle = |\mathbf{p}\rangle\,; \tag{8}$$

Where the matrix $\mathbf{Z}$ is the already described symmetric QSM, $|\mathbf{p}\rangle$ is the known *core set* shifted property vector and $|\mathbf{x}\rangle$ is a $(n \times 1)$ vector, whose coefficients have to be evaluated.

The predictive power of such an equation is a priori null. This is so because the QSM: $\mathbf{Z}$, is by construction non-singular, then exists a QSM inverse $\mathbf{Z}^{-1}$, with the relationships: $\mathbf{Z}^{-1}\mathbf{Z} = \mathbf{Z}\mathbf{Z}^{-1} = \mathbf{I}$. This leads to the trivial result, defining the unknown coefficient vector:

$$|\mathbf{x}\rangle = \mathbf{Z}^{-1}|\mathbf{p}\rangle\,. \tag{9}$$

And exact property values for any molecule of the *core set*, can be reproduced just choosing the scalar products:

$$\forall I\colon p_I = \langle \mathbf{z}_I|\mathbf{x}\rangle\,. \tag{10}$$

QSM for varied *core sets* have been used in a set of prediction studies [27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46], employing up to date statistical tools, typical of classical QSPR studies, [47,48,49,50,51,52,53,54,55,56,57,58]. The use of the fundamental QQSPR equation to construct algorithms, which can be utilized as predictive tools independently of classical QSPR algorithms, has been previously attempted [59], but it has not been continued in practice until recently [60,61,62]. Here will be discussed in several places not only the QQSPR problem itself, but various points of view and a future perspective as well.

Symmetrical Similarity Matrices. The fundamental QQSAR equation has been presented within the particular case where the basis and probe molecular QOS coincide, forming a square symmetric QS Matrix. This choice has the drawback that the fundamental QQSPR linear system becomes well defined, with a unique solution, whenever the similarity matrix is non-singular as no QO coincides with another within the QOS C. Even then, there is quite a wide range of solutions to overcome this apparent limitation. Among other procedures, one can use the C symmetric QSM, $\mathbf{Z}$, as a source of molecular descriptors and afterwards employ them in classical statistical treatments. This choice, as has been already commented, has been studied in many publications of our laboratory with success [27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46]. Other possible QSM can be constructed bearing rectangular structure, for example using known molecular structures acting as a basis set, which in turn can be compared with the *core set*.

Origin of Hansch QSAR Models. An interesting possibility of the symmetric square representation for the QS Matrices corresponds to its potential to unveil the origin of one parameter classical QSAR models, such as those Hansch [63] described some years ago. A fundamental QQSPR linear equation can be associated to a set of ν equations with the same number of unknowns, and can be rewritten as:

$$\forall J = 1,\nu\colon p_J \simeq \sum_{I=1}^{\nu} x_I z_{IJ} = x_J z_{JJ} + \sum_{I\neq J}^{\nu} x_I z_{IJ}\,. \tag{11}$$

There is no need to attach the QSM elements to any specific QOS, as all of them are computed over a unique basis of density function tags. Considering (11), two terms can be seen. The first one is attached to a self-similarity measure z_{JJ}, while the second term in cases of some not so strongly varying QOS, can be considered almost a constant, that is using:

$$\forall J\colon \alpha = x_I \wedge \beta \simeq \sum_{I\neq J}^{\nu} x_I z_{IJ}\,. \tag{12}$$

Equation (11) takes the final form:

$$\forall J = 1,\nu\colon p_J \simeq \alpha z_{JJ} + \beta \tag{13}$$

which has the required appearance to be considered as possessing a Hansch-like structure.

Equation (13) proves self-similarities can be substitutes of the classical Hansch analysis parameters [63]. They constitute, for co-generic QOS, molecular descriptors with the property to be directly attached to a tri-dimensional molecular structure. Quantum self-similarity

measures vary slowly with conformational changes [10, 64], so their values for the optimal molecular geometry can be safely used.

Mathematical Background of Quantum Similarity

Inward Matrix Product (IMP)

IMP Definition. An essential piece of QSM theory is the matrix operation called the *inward matrix product* (IMP) [65,66,67,68,69], which is based on the structure of the *Hadamard product* [70][1]. Such an operation is an internal composition law, which can be defined within a matrix (or hypermatrix) vector space $M_{(m\times n)}(K)$ of arbitrary dimension $(m \times n)$ and defined over a field K, producing a matrix whose elements are products made in turn by the elements of the matrices appearing in the IMP itself, according to the straightforward algorithm:

$$\forall \mathbf{A} = \{a_{ij}\}, \mathbf{B} = \{b_{ij}\} \in M(K)\colon \mathbf{P} = \mathbf{A} * \mathbf{B} \to \mathbf{P} = \{p_{ij}\} \in M(K) \wedge \forall i, j\colon p_{ij} = a_{ij} b_{ij} \,. \quad (14)$$

IMP is an operation, which can be applied not only to matrix spaces but over a wide variety of mathematical objects, producing another mathematical object of the same kind as the ones involved in the operation.

IMP Properties. On the other hand, IMP is equivalent to a feature involving arrays, present in high-level computer languages such as Fortran 95 [71], so practical programming of the IMP properties and characteristics is straightforward. IMP is commutative, associative, and distributive with respect to the matrix sum. Moreover, it has a multiplicative neutral element, the *unity matrix*, which has been customarily represented by a bold real unit symbol and formally defined as: $\mathbf{1} = \{1_{ij} = 1\}$.

IMP Powers and Functions. By an IMP power over a matrix $\mathbf{A} = \{a_{ij}\}$, noted as: $\mathbf{A}^{[p]}$ is understood the matrix whose elements are the corresponding powers of the elements of $\mathbf{A}$, that is: $\mathbf{A}^{[p]} = \{a_{ij}^{p}\}$. In the same manner, an IMP function of a matrix, noted as: $f[\mathbf{A}]$, is defined as the matrix whose elements are the functions of the original matrix: $f[\mathbf{A}] = \{f(a_{ij})\}$.

Scalar Product as an IMP Composite Operation. A useful application example of IMP is associated to the *total sum* of the elements of an arbitrary matrix, $\mathbf{A} = \{a_{ij}\} \in M$, by means of the symbol:

$$\langle \mathbf{A} \rangle = \sum_i \sum_j a_{ij} \,. \quad (15)$$

Connecting this definition with IMP, one can easily write:

$$\langle \mathbf{A} * \mathbf{B} \rangle = \langle \mathbf{A} \rangle * \langle \mathbf{B} \rangle \,.$$

Then, it is simple to construct the scalar product of two matrices of the same dimension, symbolized here as: $\langle \mathbf{A}|\mathbf{B} \rangle$, by means of the IMP structure:

$$\langle \mathbf{A}|\mathbf{B} \rangle = \sum_i \sum_j a_{ij} b_{ij} = \langle \mathbf{A} * \mathbf{B} \rangle \,. \quad (16)$$

In this way, the definition of distances and cosines of the angle between two matrices can be also outlined. For instance, the cosine of the angle subtended by two matrices can be written, according to Eq. (16), as:

$$\cos(\alpha) = (\langle \mathbf{A} * \mathbf{A} \rangle \langle \mathbf{B} * \mathbf{B} \rangle)^{-\frac{1}{2}} \langle \mathbf{A} * \mathbf{B} \rangle \,.$$

Vector Semispaces (VSS)

A *vector semispace* (VSS) [23,24,25,72,73] is a vector space, where the additive group has been substituted by an additive *semigroup*. An additive semigroup [74] is an additive group without reciprocal elements, which is the same as to consider negative elements not present in VSS. A matrix VSS will be made by matrices whose elements are positive definite or semi-definite. QS matrix structures belong to positive definite VSS. This is the same as to consider the matrix elements forming a VSS constructed by positive definite real numbers, extracted in turn from the $\mathbf{R}^+$ half-line. All the elements of a matrix VSS are non-singular matrices from the IMP point of view, while any matrix possessing a zero element will be non-existent in a VSS, if this strict sense is adopted. A functional VSS can be constructed by positive definite functions over a given domain and lacking of the null function in order to comply with the strict VSS characteristics.

Minkowski Norms in VSS. Because of the positive definite structure of the components of a VSS, the easiest way to define a norm within such a mathematical configuration is Minkowski's. In a matrix VSS one can write:

$$\forall \mathbf{A} \in M(\mathbf{R}^+) \to \langle \mathbf{A} \rangle \in \mathbf{R}^+ \,. \quad (17)$$

[1] The *Hadamard product* (sometimes also called *Schur or Kronecker product*) is related to the multiplication result of two sums and constructed by the sum of the resultant diagonal cross-terms only. In this way, the inward (or Hadamard) product of two sums can be specified by the following algorithm:

$$\left(\sum_I^N a_I\right) * \left(\sum_I^N b_I\right) = \sum_I^N a_I b_I \,.$$

Both sums shall have the same number of terms N, for the IMP being feasible.

Meanwhile, in any general functional VSS, an equivalent form can also be defined:

$$\forall \rho(\mathbf{r}) \in F(\mathbf{R}^+) \to \langle \rho \rangle = \int_D \rho(\mathbf{r})\, \mathrm{d}\mathbf{r} \in \mathbf{R}^+ . \tag{18}$$

As final information one can see that a Minkowski norm in $M(\mathbf{R}^+)$, and, thus, the complete sum of a matrix elements, can be considered as a linear operator, that is:

$$\langle \alpha \mathbf{x} + \beta \mathbf{y} \rangle = \alpha \langle \mathbf{x} \rangle + \beta \langle \mathbf{y} \rangle .$$

σ-Shell Structure in VSS. Minkowski-like norms classify the VSS elements in subsets, the *σ-shells*, $S(\sigma)$, whose elements are defined by the value of such a norm:

$$\forall x \in S(\sigma) \subset V(\mathbf{R}^+) \to \langle x \rangle = \sigma \in \mathbf{R}^+ . \tag{19}$$

The *unit shell* $S(1)$ is a VSS subset, which can generate all the other VSS shells. The existence of this property can be easily constructed as follows:

$$\forall z \in S(\sigma) \to \exists x \in S(1) \colon z = \sigma x . \tag{20}$$

Convex Conditions. The idea underlying *convex conditions*, associated to a numerical set, a vector, a matrix, or a function, has been described since the initial work on VSS and the related questions [23,24,25,72,73]. By the *convex conditions symbol*: $K(\mathbf{x})$ is meant that the conditions:

$$\langle \mathbf{x} \rangle = 1 \wedge \mathbf{x} \in V(\mathbf{R}^+) ,$$

hold simultaneously for a given mathematical object $\mathbf{x}$. Convex conditions become the same as considering that the object belongs to the unit shell of some VSS. Then, for such kinds of elements:

$$K(\mathbf{x}) = \{\langle \mathbf{x} \rangle = 1 \wedge \mathbf{x} \in V(\mathbf{R}^+)\} \equiv \{\mathbf{x} \in S(1)\} .$$

Conversely, the following property holds over any element of the unit shell: $\forall \mathbf{x} \in S(1) \to K(\mathbf{x})$.

Convex Linear Combinations Within a σ-Shell. Given an arbitrary σ-shell: $S(\sigma) \subset V(\mathbf{R}^+)$, of some VSS, then convex linear combinations of the elements of the $\sigma-$shell produce a new element of $S(\sigma)$. That is, suppose that the convex conditions:

$$K(\{\gamma_I\}) = \left\{ \sum_I \gamma_I = 1 \wedge \forall I \colon \gamma_I \in \mathbf{R}^+ \right\} , \tag{21}$$

hold on a known set of scalars $\{\gamma_I\}$. Then, the following property will be fulfilled for any arbitrary subset of elements belonging to the chosen $\sigma-$-shell:

$$\{\mathbf{x}_I\} \subset S(\sigma) \wedge K(\{\gamma_I\}) \to \mathbf{x} = \sum_I \gamma_I \mathbf{x}_I \in S(\sigma) ,$$

owing to the fact that the summation symbol, associated here to a Minkowski norm, is a linear operator, thus:

$$\begin{aligned} \langle \mathbf{x} \rangle &= \left\langle \sum_I \gamma_I \mathbf{x}_I \right\rangle = \sum_I \gamma_I \langle \mathbf{x}_I \rangle \\ &= \sum_I \gamma_I \sigma = \sigma \sum_I \gamma_I = \sigma \to \mathbf{x} \in S(\sigma) . \end{aligned}$$

Such a property is related to the possibility of constructing approximate atomic and molecular DF, by means of convex linear combinations, using a basis set of structurally simpler functions, which shall belong to the same VSS σ-shell. One of the possible technical options has been described in a series of papers, where the choice of the simplified functions, in atomic electronic density fitting, was made by sets of 1s GTO. The approach was termed *atomic shell approximation* (ASA) [75,76,77,78,79,80,81,82,83] and has been successfully employed, among other possibilities, to make the integral computation and molecular superposition inherent in *molecular quantum similarity measures* (MQSM) easier.

Generating Vector Spaces

Generating Symbols. Any VSS σ-shell structure can be supposedly generated by a conventional *vector space* (VS). Such VS can be defined over the complex or real fields. It can be additionally provided by convenience with a positive definite metric structure. Indeed, suppose such a VS, defined for the sake of generality over the complex field: $V(\mathbf{C})$. Then, from a very general point of view, the following algorithm can be envisaged:

$$\begin{aligned} &\forall v \in V(\mathbf{C}) \wedge v \neq \mathbf{0} \to \langle v | v \rangle = \sigma \in \mathbf{R}^+ \\ &\qquad \Rightarrow \exists x = v^* * v \in S(\sigma) \subset V(\mathbf{R}^+) . \end{aligned} \tag{22}$$

Where the IMP: $x = v^* * v$ has been used to construct the VSS vector. Then the following sequence:

$$\langle x \rangle = \langle v^* * v \rangle = \langle v | v \rangle = \sigma$$

holds and has been employed to set the form of Eq. (22).

The quantum mechanical image of the density function construction appears as a particular case of the definition attached to Eq. (22). In addition, from a complementary point of view, a symbol to briefly summarize Eq. (22) could be described. One can say compactly that the vector *generates* a VSS vector x, using a *generating symbol*: $R(v \to x)$, whenever the sequence of relationships in Eq. (22) holds [23,24,25,72,73].

Probability Density Distributions. From the quantum mechanical point of view, when a wave function $\Psi(\mathbf{r})$ is

known, then the attached DF $\rho(\mathbf{r})$ is generated in the following way: $R(\Psi \to \rho)$ [25].

Another interesting point to be noted is that any probability distribution, discrete or continuous, belongs to some VSS unit shell. Probability distributions can be generated by the conveniently normalized elements of some normed or metric space, in order that the resultant VSS element belongs to the unit shell, $S(1)$. Potentially, in this way they can belong to any other σ-shell: $\forall \rho(\mathbf{r}) \in S(1) \wedge \sigma \in \mathbf{R}^+ \to p(\mathbf{r}) = \sigma\rho(\mathbf{r}) \in S(\sigma)$.

Thus, in VSS one can consider that the unit shell resumes every other σ-shell. In this manner, probability distributions of any kind can be transformed into any other element of the associated VSS. One can also say that any VSS σ-shell, $S(\sigma)$, can be considered like a homothetic construct of the unit shell, $S(1)$ [72]. Because the elements of the unit shell comply with the adequate form of a probability distribution, then they also fulfil the convenient convex conditions. That is: $\forall x \in S(1) \to K(x)$. In other words, any probability distribution can be considered as an element of the unit shell forming part of a VSS.

Scalar Products and Measures in VSS. Because of these possible connections attached to probability vectors, scalar products of two distinct compatible probability distributions are always positive definite, as one has:

$$\forall x, y \in S(1) \to \langle x|y\rangle = \langle x * y\rangle \in \mathbf{R}^+ . \tag{23}$$

Therefore, the cosine of the angle subtended by two probability distributions has to be contained in the open interval: $(0, 1]$, due to Eq. (23). This is so, because the cosine of the subtended angle between two elements can be computed as:

$$\forall x, y \in S(1)\colon\ \cos(\alpha) = (\langle x{*}x\rangle\langle y{*}y\rangle)^{-\frac{1}{2}}\langle x{*}y\rangle \in (0, 1].$$

Furthermore, Eq. (23) shows that the scalar product between two, or more, elements of the unit shell of a given metric VSS, constitute in any case a *measure*. In this way, such a scalar product can be considered a generalized volume.

Quantum Similarity

QSM over the Unit Shell. A similarity measure over the unit shell of any VSS can be defined through the description of the mathematical elements described so far. *Quantum similarity measures* (QSM) were described a long time ago [1] and have been constantly used since then [84,85,86,87,88,89,90,91,92,93]. In the most simple and at the same time general way, within the easier formalism available, a QSM can be defined knowing the appropriate DF of two quantum systems: $\{\rho_A; \rho_B\}$, adapted to the unit shell of the corresponding VSS, and using as weight some positive definite operator: $\{\Omega\}$, then the integral measure:

$$\begin{aligned} z_{AB}(\Omega) &= \iint_D \rho_A(\mathbf{r}_1)\Omega(\mathbf{r}_1;\mathbf{r}_2)\rho_B(\mathbf{r}_2)\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2 \\ &= \langle \rho_A \Omega \rho_B\rangle \in \mathbf{R}^+ \end{aligned} \tag{24}$$

will correspond to a weighted scalar product, defined over the unit shell elements, made in turn by the compatible quantum DF. The QSM (24) can be associated to a property very comparable to the one encountered in Eq. (24) and in any instance has to possess a positive definite nature.

Quantum Object Sets. Suppose a set of quantum systems: $S = \{s_I\}$, in a well-defined set of states. Suppose that to every quantum system there is attached a known state DF, forming the set: $P = \{\rho_I\}$, belonging to the unit shell of some functional VSS. A tagged set [23,24,25] can be constructed, using the Cartesian product: $T = S \times P$, where each element, $\tau_I \in T$, is constructed by the ordered pair composition rule: $\tau_I = (s_I; \rho_I)$, forming in this way a *quantum object*. The tagged set T constitutes a *quantum object set*, that is: $T = \{\tau_I\}$. The QSM earlier defined in Eq. (24), can be interpreted, in turn, as a tensor product of the tag part of the QOS.

Quantum Object Sets and Core Sets. The usual problem in all QSPR studies is customarily based on the previous knowledge of some molecular set: M, of cardinality n, such that the structures and properties of the set elements are known beforehand. From now on, one can refer to this collection of molecules, molecular descriptors and properties as the *core set* and name it as: C. With the molecular structures of the elements of the set M known, one can solve the Schrödinger equation associated to, in principle, the ground state of every molecule in the set and compute an attached set of density functions: $\mathrm{P} = \{\rho_I\}$, which can be connected to sole continuous molecular descriptors, according to the quantum mechanical usual custom; such that:

$$\forall m_I \in \mathrm{M} \to \exists \rho_I \in \mathrm{P} \wedge \forall I\colon m_I \leftrightarrow \rho_I .$$

In terms of the theoretical settings related with quantum similarity, the Cartesian product of the molecular and the density function sets is used to construct a *tagged set*: $\mathrm{Q} = \mathrm{M} \times \mathrm{P}$. Such a tagged set has been formerly named a *quantum object set*, where the molecules constitute the *object set* and the density functions act as the *tag set*. Thus,

from now on, one can consider the *core set* as a well-defined QOS. One can name the elements of a QOS as *quantum objects*. Therefore, ordered pairs, constructed in the following way, define any QO:

$$\forall m_I \in \mathrm{M} \wedge \forall \rho_I \in \mathrm{P} \rightarrow \forall \omega_I \in \mathrm{Q}\colon \omega_I = (m_I; \rho_I)\,. \quad (25)$$

However, the *core set* shall be structured in an even extended manner. Starting from the QOS definition, the *core set* C has also to be associated with the following characteristic:

$$\begin{aligned}&\forall \omega_I \in \mathrm{Q} \wedge \exists \pi_I \in \Pi\\ &\rightarrow \forall c_I \in \mathrm{C} = \mathrm{Q} \times \Pi\colon c_I = (\omega_I \pi_I) \equiv (m_I; \rho_I; \pi_I)\,,\end{aligned}$$

Where the set Π contains all the properties of the elements of the molecular set M. Hence, *core set* elements are well-defined triples consisting of molecular structures, density functions, and properties: $\mathrm{C} = \mathrm{M} \times \mathrm{P} \times \Pi$.

In classical QSPR, based on the empirical description of the set M, the set of density functions is replaced by a set of vectors, belonging to a finite dimensional space, whose elements are the chosen *molecular descriptors*. The possible construction of the *core set* within discrete vector spaces, substituting the density tag set P is a characteristic, which will also appear within the QQSPR theoretical development, as will be explained below. It must be said that the QQSPR substitution of the continuous density tags by discrete vectors has a quite well-structured mathematical-theoretical meaning, while in empirical QSPR remains arbitrarily chosen.

Similarity Matrices. Collecting all the QSM computed between the element pairs of a given QOS, a so-called *Quantum Similarity Matrix* (QSM) is obtained, and constructed according to the definition (24) by means of: $\mathbf{Z} = \{z_{ij}\}$. Because of the QS matrix elements structure, the matrix itself can be considered as an element of some VSS of the appropriate dimension. The QS matrix $\mathbf{Z}$ is a symmetric matrix with positive definite elements, whose columns $\{\mathbf{z}_I\}$ (or rows) are also elements of some N-dimensional VSS. As such, there exists a real symmetric matrix, $\mathbf{X}$, such that, in general $R(\mathbf{X} \rightarrow \mathbf{Z})$, that is:

$$\mathbf{Z} = \mathbf{X} * \mathbf{X} = \mathbf{X}^{[2]} \vee \mathbf{X} = \mathbf{Z}^{\left[\frac{1}{2}\right]}\,. \quad (26)$$

As a consequence, any QS matrix belongs to a precise σ-shell of some VSS. That is:

$$\forall \mathbf{Z}\colon \langle \mathbf{Z} \rangle = \sum_i \sum_j z_{ij} = \sigma \rightarrow \mathbf{Z} \in S(\sigma) \subseteq M(\mathbf{R}^+)\,.$$

Stochastic Similarity Matrices. Even if the columns or rows of the QS matrix $\mathbf{Z}$ belong to different σ-shells of some VSS, they can be easily brought to the unit shell, by using a set of simple homothetic transformations, involving a product by a diagonal matrix, with elements constructed by the Minkowski norms of the columns (or rows) of the QS matrix. That is, the diagonal matrix:

$$\mathbf{D} = \mathrm{Diag}(\langle \mathbf{z}_1 \rangle; \langle \mathbf{z}_2 \rangle; \ldots; \langle \mathbf{z}_I \rangle; \ldots)\,, \quad (27)$$

can transform the QS matrix $\mathbf{Z}$ into a column (or row) *stochastic matrix*, simply by multiplying on the right (or the left) of $\mathbf{Z}$ by the inverse of $\mathbf{D}$, respectively [59]. For instance, the stochastic column matrix associated to the QS matrix is:

$$\mathbf{S} = \mathbf{Z}\mathbf{D}^{-1}\,. \quad (28)$$

In this way, the columns $\{s_I\}$ of the stochastic matrix,$\mathbf{S}$, belong to the unit shell of the column vector VSS of the appropriate dimension. That is:

$$\begin{aligned}&\mathbf{S} = \{\mathbf{s}_I\}\\ &\rightarrow \forall I\colon \langle \mathbf{s}_I \rangle = \left\langle \langle \mathbf{z}_I \rangle^{-1} \mathbf{z}_I \right\rangle = \langle \mathbf{z}_I \rangle^{-1} \langle \mathbf{z}_I \rangle = 1\\ &\rightarrow \mathbf{s}_I \in S(1)\,.\end{aligned}$$

However, the column stochastic QS matrix (28) appears to be no longer symmetric as his originating QS matrix $\mathbf{Z}$ is.

Quantum Similarity Matrix Aufbau Procedure [94]

Suppose a known given Quantum Object Set formed by N molecules, with density tags described as: $\{\rho_I(\mathbf{r})\}$. Up to now, the usual procedure to construct the QSM has been to maximize each of the integrals of type (26) with respect to the translations and rotations of one of the implied QO in relation to the others. This can be expressed formally, for instance, as:

$$\forall I > J\colon z_{IJ} = \langle \rho_I \rho_J \rangle = \int_D \rho_I(\mathbf{r}) \rho_J(\mathbf{r})\, \mathrm{d}\mathbf{r} = z_{JI}\,, \quad (29)$$

However, as has been well known since the first paper on the subject [1] the set of quantum similarity measures $\{z_{IJ}\}$ depend on the relative position in 3D space of the implied Quantum Objects (QO's). As the QO density function labels are positive definite functions, the integrals of type (29) can be considered as measures; thus, they are positive definite too. Up to now the usual procedure to construct the QSM has been to maximize each of the integrals of type (29) with respect to the translations and rotations of one of the implied QO in relation to the other. This can be expressed formally, for instance, as:

$$\forall I > J\colon z_{IJ} = \max_{\mathbf{t};\boldsymbol{\Omega}} \int_D \rho_I(\mathbf{r}) \rho_J(\mathbf{r}|\mathbf{t}; \boldsymbol{\Omega})\, \mathrm{d}\mathbf{r} \quad (30)$$

where the pairs: $\{\mathbf{t}; \boldsymbol{\Omega}\}$ are translations and rotations respectively, performed on the center of coordinates of the Jth QO. It is irrelevant which one of the QO pair is chosen in order to optimize the integral (29) by means of the algorithm (30), the same result shall be obtained choosing the Ith QO for undertaking translations and rotations.

Apparently, such a procedure, repeated for every non-redundant couple of QO's, shall provide a QSM $\mathbf{Z}$ with appropriate characteristics associated to a metric matrix. The most important one is that the attached QSM has the property to be positive definite; as the density tag set is linearly independent, if the QOS is made of different QO's, then $\mathbf{Z}$ has to be a metric matrix of a pre-Hilbert space [94]. However, in many cases the use of algorithm (30) does not provide a QSM whose whole spectrum is positive definite, but a small amount of the $\mathbf{Z}$ eigenvalues may appear to be negative. This non-definite behavior of the metric matrix $\mathbf{Z}$ can be attributed to the fact that following algorithm (30), when facing the Jth QO to the rest of the QOS elements, then for every distinct QO a different relative position of the Jth QO is found, while reaching the optimal value of the similarity measure (29) for every pair of QO's; that is: the relative position of the Jth QO with respect to the Ith QO, $\forall I\colon J \neq I$, in order to optimize every element z_{IJ}, becomes different, and therefore when optimizing Eq. (30) one will obtain a set of different optimal translations-rotations: $\{\mathbf{t}_I; \boldsymbol{\Omega}_I\} \forall I \neq J$.

When computing any optimal quantum similarity measure by means of algorithm (30), one also must be aware that the final result, can be used to construct the symmetric (2×2) matrix:

$$\mathbf{Z}^{IJ} = \begin{pmatrix} z_{II} & z_{IJ} \\ z_{JI} & z_{JJ} \end{pmatrix} \wedge z_{IJ} = z_{JI}\,, \tag{31}$$

and also has to provide at least a positive definite matrix (31), which is the same as to consider the following property has to be fulfilled:

$$\mathrm{Det}\,|\mathbf{Z}^{IJ}| = z_{II}z_{JJ} - z_{IJ}^2 > 0 \rightarrow z_{II}z_{JJ} > z_{IJ}^2\,. \tag{32}$$

The restriction (32) can also be written as:

$$z_{JJ} > z_{IJ}^2 z_{II}^{-1}\,, \tag{33}$$

and this will provide a form of the (2×2) positive definite restrictions to be easily related to the general analysis which follows. Therefore, the algorithm (30) has to be modified accordingly incorporating the inequality (32) as a restriction:

$$\forall I > J\colon$$
$$z_{IJ} = \max_{\mathbf{t};\boldsymbol{\Omega}} \int_D \rho_I(\mathbf{r})\rho_J(\mathbf{r}|\mathbf{t};\boldsymbol{\Omega})\,\mathrm{d}\mathbf{r} \wedge z_{IJ}^2 < z_{II}z_{JJ} \tag{34}$$

and one can expect that the general QSM $\mathbf{Z}$, can approach in this way the required complete positive definiteness, although this cannot be completely assured. In fact, this (2×2) restriction constitutes an incomplete point of view, as nothing can be said about the positive definiteness of higher dimensional submatrices of the QSM $\mathbf{Z}$. In this sense, the restricted algorithm (34) is more or less similar to the triangle distance relationship coherence, sought by an already published procedure [91].

The Quantum Similarity Matrix Aufbau Recursive Algorithm. Although one can use the Gershgorin theorem to test the positive definiteness of any QSM, a complete QSM calculation algorithm, based on the generalization of property (33) for (2×2) matrices, in order to assure the QSM $\mathbf{Z}$ positive definiteness, shall be based on an Aufbau procedure; that is: starting from any pair of QO, algorithm (34) is put forward. The result will be a positive definite matrix, $\mathbf{Z_0}$ say, with a structure like the matrix (31) defined above. A simple recursive Aufbau algorithm can be described in order to obtain a final positive definite QSM.

Suppose that for some index $P < N$, a $(P \times P)$ positive definite QSM $\mathbf{Z_0}$ has been obtained, using the QO's sequence: $\{I_K; K = 1, P\}$. One can add a new QO to the Aufbau procedure, the Qth QO, say, in such a way that an augmented QSM, $\mathbf{Z_1}$, is obtained possessing the partitioned structure:

$$\mathbf{Z_1} = \begin{pmatrix} \mathbf{Z_0} & |\mathbf{z}\rangle \\ \langle\mathbf{z}| & \theta \end{pmatrix}\,,$$

with the $(1 \times P)$ row vector defined as: $\langle\mathbf{z}| = (z_{I_1Q}; z_{I_2Q}; \ldots; z_{I_PQ})$, and the column vector $|\mathbf{z}\rangle$, being just the transpose of the former; finally, $\theta \equiv z_{QQ}$ is the self-similarity of the added QO.

The sufficient relationship, which can be written here as the set of conditions:

$$\theta > \langle|\mathbf{z}\rangle\rangle \wedge \forall K = 1, P\colon z_{I_KI_K} > \sum_{L \neq K} z_{I_KI_L} + z_{I_KQ}\,, \tag{35}$$

assuring that the augmented matrix $\mathbf{Z_1}$ has a positive definite structure, can be alternatively rewritten via a recursive Cholesky decomposition algorithm, described in several places [17,96].

The necessary and sufficient condition for the positive definiteness of the augmented QSM $\mathbf{Z_1}$ can be stated as:

$$\theta - \langle\mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{z}\rangle > 0$$
$$\rightarrow z_{QQ} > \sum_K \sum_L z_{I_KQ} z_{I_LQ} Z^{(-1)}_{0;I_KI_L}\,. \tag{36}$$

The Cholesky decomposition condition, which can be called here *Quantum Similarity Aufbau Condition*

(QSAC), means that it cannot be reliable to use a pair of QO's every time that a new element of the QSM has to be computed, but that the added QO density function: $\rho_Q(\mathbf{r}|\mathbf{t};\boldsymbol{\Omega})$ has to be translated-rotated with the same values of the pair: $\{\mathbf{t};\boldsymbol{\Omega}\}$, for every computed element of the vector $|\mathbf{z}\rangle$, connecting recursively the QO Q with all the ones previously employed in constructing the QS submatrix $\mathbf{Z_0}$. When the QS submatrix $\mathbf{Z_0}$ has scalar (1×1) dimension as occurs in the submatrix (31) case, then the QSAC (36) becomes the relationship (33). Moreover, the QSAC condition is a stronger positive definiteness condition than the diagonal dominance, as QSAC becomes the necessary and sufficient condition for constructing a positive definite augmented matrix.

The maximal pair condition (30) can be substituted in the general $(P \times P)$ case, for instance, by maximizing the sum of the whole vector $|\mathbf{z}\rangle$, which due to the positive definiteness of its elements is coincident with the search of a maximal Minkowski norm:

$$\max_{\mathbf{t};\boldsymbol{\Omega}}[\langle|\mathbf{z}\rangle\rangle] = \max_{\mathbf{t};\boldsymbol{\Omega}}\left[\sum_{K=1}^{P}\int_D \rho_{I_K}(\mathbf{r})\rho_Q(\mathbf{r}|\mathbf{t};\boldsymbol{\Omega})\,\mathrm{d}\mathbf{r}\right] = \max_{\mathbf{t};\boldsymbol{\Omega}}\left[\sum_{K=1}^{P} z_{I_K Q}\right]. \tag{37}$$

This can be done admitting the same translation-rotation sequence performed on every term of the vector $|\mathbf{z}\rangle$ in Eq. (37), whenever such transformation increases the Minkowski norm. However, while the maximal value of the sum leading to the Gershgorin radius is searched as in the condition (37) of the previous sentence, the QSAC relationship (36) has to be equally tested and if not fulfilled the pair $\{\mathbf{t};\boldsymbol{\Omega}\}$ rejected. Such a procedure will assure the positive definiteness of the QSM $\mathbf{Z}$ at the final step of the recursion and will provide the same relative position in the calculation of the quantum similarity measures for every recursively added QO.

Geometrical Interpretation of the QSAC. Leaving apart the linear algebra concept of diagonal dominance which similarity matrices usually do not fulfill, the alternative Cholesky decomposition condition property leading to the QSAC, assuring in this manner the positive definite structure of the final QSM form and written as in Eq. (36), has a clear geometrical meaning. The positive definite quadratic form: $\langle\mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{z}\rangle \in \mathbf{R}^+$, is nothing else than the Euclidean norm of the vector $|\mathbf{z}\rangle$ in the reciprocal metric space defined by the density tags: $\{\rho_{I_K}(\mathbf{r}); K = 1, P\}$. The QO's tags are employed to form the QSM $\mathbf{Z_0}$, which because of the QSAC construction has been structured positive definite and acts accordingly as a metric matrix of a P-dimensional pre-Hilbert space. Since in the quadratic form appearing in Eq. (36), the inverse of the metric $\mathbf{Z_0}$ appears, the implicit Euclidean norm equivalent to the aforementioned quadratic form is computed in the metric reciprocal space with the matrix $\mathbf{Z}_0^{-1}$, acting as a positive definite metric matrix, because: $Sp[\mathbf{Z_0}] \in \mathbf{R}^+ \rightarrow Sp\left[\mathbf{Z}_0^{-1}\right] \in \mathbf{R}^+$. Accordingly, the QSAC forces this Euclidean norm in the reciprocal P-dimensional pre-Hilbert space to be less than the self-similarity of the recursively added Qth QO.

This permits us to associate the described Quantum Similarity Aufbau procedure as an algorithm maximizing the Minkowski norm of each recursive column $|\mathbf{z}\rangle$ of the QSM, submitted to the QSAC restriction which means that its Euclidean norm, computed in the recursive reciprocal pre-Hilbert space, remains less than the recursive QSM diagonal self-similarity elements.

Finally, the following points must be taken into account:

1. Because it is not necessary to start the recursive QSAC with any a priori chosen QO, the final QSM will certainly depend on the QO recursive order chosen. Thus, there are just $N!$ possible choices, each one producing an equally positive definite QSM. However, the ordering imposed by the self-similarity measures can be chosen as a way to reach a systematic QSM Aufbau. That is, if one calls the QSM diagonal self-similarity measures set computed on the QOS elements: $D(\mathbf{Z}) = \{z_{II}\}$, then the obvious choices are defined by the maximal ordering:

   $$z_{11} = \max_I[D(\mathbf{Z})] \rightarrow z_{22} = \max_I[D(\mathbf{Z}) - z_{11}]\ldots$$

   or by the minimal:

   $$z_{11} = \min_I[D(\mathbf{Z})] \rightarrow z_{22} = \min_I[D(\mathbf{Z}) - z_{11}]\ldots$$

   This ensures that the QO's will be ordered in decreasing or increasing complexity, while providing a generic reproducible way of computing QSM under QSAC premises.
2. When constructing the QSM according to the proposed Aufbau procedure, it is well known that the overlap quantum similarity measures, as defined in Eq. (29), can be substituted by a more general form involving a positive definite operator: $\Omega(\mathbf{r_1};\mathbf{r_2})$; so, in general, the similarity measures can be described as the integral:

   $$z_{IJ}(\Omega) = \iint_D \rho_I(\mathbf{r_1})\Omega(\mathbf{r_1};\mathbf{r_2})\rho_J(\mathbf{r_2})\,\mathrm{d}\mathbf{r_1}\,\mathrm{d}\mathbf{r_2}\,;$$

   while the positive definite operator choice ensures that the QSM, when constructed according to the equiva-

lent QSAC, like the one depicted previously for overlap quantum similarity measures in Eq. (36), is positive definite. One just shall make the substitution: $z_{IJ} \leftarrow z_{IJ}(\Omega)$.
3. The QSAC is also valid for quantum similarity measures involving the off-diagonal terms of the density matrix.

Linear Quantum QSPR Fundamental Equation

Expectation Values

In quantum mechanics, the expectation value of some QO observable, associated in turn to some Hermitian operator W, is measured in the usual statistical way [97,98,99], using the tag part ρ_A of the corresponding QO:

$$\langle \pi_A \rangle = \langle W | \rho_A \rangle = \int_D W \rho_A \, \mathrm{d}V \, . \tag{38}$$

At the same time, in general, the operator W can be decomposed as follows:

$$W = Q\Omega \, .$$

Ω bears a positive definite and known form. On the other hand, the operator: Q, can be approximately expressed in terms of an appropriate linear (or multilinear) combination of a known density function set $\{\rho_I\}$, provided with the adequate variable count, in order to match the one of ρ_A, that is:

$$Q \simeq \sum_I w_I \rho_I \, . \tag{39}$$

The structure associated to an operator like: W, permits us to construct expectation values of entangled or complicated observables of submicroscopic systems. Such entangled observables can be considered connected to experimental outcomes, like biological activity, whose Hermitian operator cannot be completely well defined. So, this way to proceed appears quite appropriate for cases where the complexity of the observed phenomenon does not possess a straightforward association with any known or easily describable Hermitian operator. One shall stress the fact that the set up (39) is not appropriate for well-defined observables like kinetic energy, Coulomb energy, dipole and multipole moments,...

Quantum QSPR Fundamental Equation

Substituting the expression of the operator, described by Eq. (39) into expectation value expression (38), one arrives at the following result, related to the corresponding QO:

$$\langle \pi_A \rangle = \langle Q\Omega | \rho_A \rangle = \langle Q | \Omega \rho_A \rangle \simeq \sum_I w_I \langle \rho_I | \Omega \rho_A \rangle = \sum_I w_I z_{IA}(\Omega)$$

which, after supposing that several QO or the whole elements of a QOS are considered, this result can be brought into the matrix form of a linear equation:

$$\mathbf{Zw} \simeq |\pi\rangle \, . \tag{40}$$

Where the column vector $|\pi\rangle$ contains the collection of expectation values of the considered QOS, $\mathbf{Z} = \{z_{IA}(\Omega)\}$ is a quantum similarity matrix and, finally, the column vector $\mathbf{w}$ collects the coefficients by which the operator Q is approximately expressed by means of Eq. (39).

Interpretation and Characteristics of the Quantum QSPR Fundamental Equation

The interpretation of the linear system (40) can proceed as follows. The vector $|\pi\rangle$ can supposedly contain known values of a well-defined, but arbitrarily complicated observable property of the chosen QOS. The quantum similarity matrix $\mathbf{Z}$ can be computed, once the elements of the QOS are supposedly known. The coefficient vector $\mathbf{w}$ has to be determined.

Put in such terms, Eq. (40) has the same well-known structure as the usual classical QSPR problems. However, this fact constitutes a very important and crucial result: Because, it permits us to interpret the columns (or rows) of the quantum similarity matrix, $\mathbf{Z}$, as being the QSM finite-dimensional, discrete, *descriptors* of every QO used in the study. These considerations are sufficient to allow us to name Eq. (40) the *quantum QSPR* (QQSPR) *fundamental equation*.

Characteristics of the QQSPR Fundamental Equation. Unlike the problems present in classical QSPR models, the QQSPR fundamental equation has several characteristics lacking in the former usual equations, these are:

1. *Universal applicability*, because Eq. (40) can be used to model any kind of QOS: nuclei, atoms, molecules ...
2. *Unbiased background* descriptor structures, because the QSM elements, forming the quantum similarity matrix, $\mathbf{Z}$, appearing in Eq. (40), are not arbitrarily chosen by the user, among those belonging to a given descriptor pool, but appear as a consequence of the theory.
3. *Causal character*, as the QQSPR models obtained are the result of solving a well-defined equation, as shown through the set up of Eq. (40), and are deducible

from the general theoretical structure of quantum mechanics.

Quantum Similarity Matrices (QSM) in the Construction of First-Order QSPR Operators and the Definition of Discrete QOS. The first-order approach of the QSPR operator, for the *core set* known molecular property tag set: $\Pi = \{\pi_I\}$ generates the following equation collection:

$$\forall I = 1, n:$$
$$p_I = \pi_I - \langle \sigma[\rho_I] \rangle \approx \sum_J x_J \langle \rho_J[\rho_I] \rangle = \sum_J x_J z_{JI} \,. \tag{41}$$

The set of integrals:

$$\left\{ \langle \rho_J[\rho_I] \rangle = \int_D \rho_J \rho_I \, \mathrm{d}V = z_{JI} \right.$$
$$\left. = z_{IJ} = \int_D \rho_I \rho_J \, \mathrm{d}V = \langle \rho_I[\rho_J] \rangle \right\} \,,$$

appearing in Eqs. (197) can be ordered into a $(n \times n)$ symmetric array, constructing in this way the so-called *quantum similarity matrix*: $\mathbf{Z} = \{z_{IJ}\}$ (QSM). In turn, the ordered set of shifted properties: $\{p_I\}$ can form a $(n \times 1)$ column vector: $|\mathbf{p}\rangle = \{p_I\}$. Therefore, the equation set (197) is simply a linear system, which will be discussed next, in order to describe its possible use for evaluating U-m unknown molecular properties.

Discrete QOS. Every column of the QSM: $\mathbf{Z} = \{|\mathbf{z}_I\rangle = \{z_{JI}\}\}$, can be interpreted as a discrete matrix representation of each QO density matrix, present within the density function tag set: $\mathrm{P} = \{\rho_I\}$. In this way a one-to-one correspondence can be established between the density tag set and the QSM column submatrices, which can be written as:

$$\forall m_I \in \mathrm{M}: \rho_I \leftrightarrow |\mathbf{z}_I\rangle \Rightarrow \mathrm{P} \Leftrightarrow \mathbf{Z} \,.$$

In other words, the QSM column set can be used as a new n-dimensional vector tag set, attached to the molecular set M, in order to build up a new tagged set, namely a *discrete quantum object set*:

$$\mathrm{Q}_\mathrm{Z} = \mathrm{M} \times \mathbf{Z} \,. \tag{42}$$

In this DQOS, the density function tags of the original QOS, Q, belonging to the tag set P, are substituted by the columns of the QSM. Therefore, there also exists a one-to-one correspondence between both QOS: $\mathrm{Q} \leftrightarrow \mathrm{Q}_\mathrm{Z}$.

The Nature of the QSM Descriptors. In both the quantum similarity matrix $\mathbf{Z}$, or its stochastic column transformation $\mathbf{S}$, the involved columns forming both matrices possess a very special character, besides the fact that they belong to some VSS of the appropriate dimension.

Starting from the QOS, where each QO is defined by the ordered pair of submicroscopic systems and state density functions:

$$\tau_A = (s_A; \rho_A) \in T = S \times P \,,$$

then, when dealing with the construction of the QSM or the stochastic transformation (28), which one can consider expressed through the decomposition of its columns as: $\mathbf{Z} = \{\mathbf{z}_I\}$ or $\mathbf{S} = \{\mathbf{s}_I\}$, it can be deduced that both matrices induce a new possible QOS, made with discrete N-dimensional tags, instead of the infinite dimensional density function ones, namely:

$$\theta_A = (s_A; \mathbf{z}_A) \in \Theta = S \times \mathbf{Z} \,, \tag{43}$$

or one can see the equivalent structure, which can also be considered alternatively as:

$$\sigma_A = (s_A; \mathbf{s}_A) \in \Sigma = S \times \mathbf{S} \,. \tag{44}$$

$$\mathbf{Sw} \simeq \mathbf{p} \,.$$

The *discrete* QOS, represented by the definitions (43) or (44), can be admittedly considered, without doubt, as finite dimensional representations of the original QOS, based on density function tags. This can be so, as both $\mathbf{z}_A$ and $\mathbf{s}_A$ discrete tags, essentially are elements of some VSS. Perhaps the representation (44), with tags belonging to the unit shell, corresponds to the most adequate of such discrete forms and, at the same time, the one which is more connected to the original infinite dimensional unit shell made by the collection of density functions.

Even if the choice to build up the problem is the DQOS, represented by Eq. (44), the fundamental QQSPR Eq. (40) can be transformed conveniently into the new row stochastic system:

$$\mathbf{Sw} \simeq \mathbf{p} \,, \tag{45}$$

simply by multiplying on the left by the inverse of the diagonal matrix (27), and using accordingly the transformed properties vector: $\mathbf{p} = \mathbf{D}^{-1}|\pi\rangle$. The column stochastic transformation can be used straightforwardly too, and so it will not be discussed here anymore, as it has been exhaustively studied within several papers [59,100].

First-Order Fundamental QQSPR (FQQSPR) Equation

The analysis of the QQSPR problem can start with the first order or linear fundamental QQSPR equation, involving the *core set*, formed with the molecules of the associated DQOS, which are also linked with known values of some property.

One can write the first-order QQSPR fundamental equation in a compact matrix form:

$$\mathbf{Z}|\mathbf{x}\rangle = |\mathbf{p}\rangle \; ; \tag{46}$$

Where the matrix $\mathbf{Z}$ is the already described symmetric QSM, $|\mathbf{p}\rangle$ is the known *core set* property vector and $|\mathbf{x}\rangle$ is a $(n \times 1)$ vector, whose coefficients have to be evaluated.

The predictive power of such an equation is a priori null, because being the QSM: $\mathbf{Z}$, by construction non-singular (otherwise two density functions will be exactly the same), then there always can be computed a QSM inverse: $\mathbf{Z}^{-1}$, obeying the usual relationships: $\mathbf{Z}^{-1}\mathbf{Z} = \mathbf{Z}\mathbf{Z}^{-1} = \mathbf{I}$, in such a way that the trivial result, defining the unknown coefficient vector:

$$|\mathbf{x}\rangle = \mathbf{Z}^{-1}|\mathbf{p}\rangle \; , \tag{47}$$

will be always obtained within a *core set* scenario. Furthermore, one can retrieve the exact value of the property for any molecule of the *core set* QOS choosing the scalar products:

$$\forall I \colon p_I = \langle \mathbf{z}_I | \mathbf{x} \rangle \; . \tag{48}$$

The QSM for diverse *core sets* has been used in a quite large set of prediction studies, in every case employing up-to-date statistical tools, the usual procedures currently available in classical QSPR studies. In the present study, the reader can find in the following sections new theoretical developments of the fundamental QQSPR equation prediction ability. However, a reminder of some simple linear algebra for the FQQSPR equation is needed first in order to understand the following arguments; therefore this will be described in the following sections.

Partitioning the FQQSPR Equation and the QSM Inverse. Supposing now one can organize the QSM in the fundamental Eq. (40) in such a way that the last column and row correspond to a *U-m*, then, the unknown property element will be supposedly stored in the last position, the $(n+1)$th, of the vector $|\mathbf{p}\rangle$ and will be symbolized by an a priori undefined parameter: π. With this in mind, one can design a partition of the QSM and the entire FQQSPR Eq. (40) in the following way:

$$\begin{pmatrix} \mathbf{Z_0} & |\mathbf{z}\rangle \\ \langle \mathbf{z}| & \theta \end{pmatrix} \begin{pmatrix} |\mathbf{x_0}\rangle \\ x \end{pmatrix} = \begin{pmatrix} |\mathbf{p_0}\rangle \\ \pi \end{pmatrix} \; . \tag{49}$$

Where, the $(n \times 1)$ column vector $|\mathbf{z}\rangle$ corresponds to the representation of the *U-m* in terms of the density tags of the *core set* and θ is the *U-m* self-similarity measure, which according to the simplified formalism of the expectation values can be defined by means of a simple overlap quantum similarity measure, as the Euclidean norm:

$$\theta = \int_D \rho_U^2(\mathbf{r})\, \mathrm{d}\mathbf{r} \; .$$

One can find the solution of the partitioned linear system by using the following symbols for the partitioned QSM inverse:

$$\mathbf{Z}^{-1} = \begin{pmatrix} \mathbf{Z_0}^{(-1)} & |\mathbf{z}^{(-1)}\rangle \\ \langle \mathbf{z}^{(-1)}| & \theta^{(-1)} \end{pmatrix} \; , \tag{50}$$

and one can evaluate the inverse elements for partitioned QSM matrices in the usual way.

Remarks on the Structure of the Fundamental QQSPR Equation

The following remarks relate to the result given by the fundamental QQSPR equation that was discussed in the previous section. Each of these remarks poses new problems that will be studied separately in subsequent sections.

Symmetrical Similarity Matrices. In the first place, it must be said that the fundamental QQSAR equation has been usually presented in previous literature within the particular case where the basis and probe molecular quantum object tagged sets coincide, providing a square symmetric similarity matrix, and thus the equality: $\mathbf{A} = \mathbf{Z}$, between the involved similarity matrices holds. This choice has the drawback that the fundamental QQSPR linear system becomes well defined, with a unique solution, whenever the similarity matrix is non-singular, which shall be the usual case, as far as no quantum object coincides with another within the quantum object set.

But even then, there is quite a wide range of solutions to overcome this apparent limitation. Among other procedures, one can use the symmetric similarity matrix as a source of molecular descriptors and afterwards employ them in classical statistical treatments. This choice, as was already commented, has been studied in many publications of our laboratory with success. In the same way, the similarity matrix can be transformed into a column or row stochastic matrix and, as a consequence, this form suggests several possibilities, which still are far from being exploited. Some analysis of the stochastic issue will be developed in a forthcoming section of this paper.

Origin of Hansch QSAR Models. An interesting possibility of the symmetric square representation of the quantum similarity matrices corresponds to its potential to unveil the origin of one parameter classical QSAR models, such as those Hansch described some years ago. Indeed, under the equivalence of both the basis B and probe P quantum object sets, the FQQSPR linear equation corresponds to a set of N equations with the same number of unknowns, and can be rewritten as:

$$\forall J = 1, N\colon p_J \simeq \sum_{I=1}^{N} \omega_I z_{IJ} = \omega_J z_{JJ} + \sum_{I \neq J}^{N} \omega_I z_{IJ} \,, \quad (51)$$

where there is no need to attach the similarity matrix elements to any specific quantum object set, as all of them are computed over a unique basis of density function tags. Considering the two terms at the end of the previous equation, it can be seen that the first one, with a diagonal value of the similarity matrix, is attached to a self-similarity measure z_{JJ}, while the second term in cases of a not so strongly varying family of quantum objects, can be considered almost a constant, that is using:

$$\forall J\colon \alpha = \omega_J \wedge \beta \simeq \sum_{I \neq J}^{N} \omega_I z_{IJ} \,, \quad (52)$$

the above equation takes the final form:

$$\forall J = 1, N\colon p_J \simeq \alpha z_{JJ} + \beta \,, \quad (53)$$

which has the required appearance to be considered as possessing a Hansch structure.

Besides this last deduction, it must be said that self-similarity measures of different kinds have been used to test this simple linear equation with quite a large series of quantum objects, yielding usually good results. Self-similarities can be sound substitutes of the classical Hansch analysis parameters. They constitute for co-generic molecular sets molecular descriptors with the property to be directly attached to a tri-dimensional molecular structure. Self-similarity measures vary slowly with conformational changes, so their values for the optimal geometry can be safely used, knowing that the magnitude of the descriptor will differ not very much from the one which is attached to the active conformation associated to the observable property.

Solutions of the Linear QQSPR Fundamental Equation

Equation (40) can be solved choosing the customary methods, as in the usual algorithms employed within the QSPR field [20]. They have been manipulated in this fashion, since its first description on a great deal of cases, as well as for a large variety of problems and subjects [26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,101]. This means that in all these studies the QQSPR fundamental equation has been solved with some algorithm, based on the least squares or similar technique [102,103,104,105,106,107,108,109,110,111,112,113,114,115,116,117,118,119,120,121,122,123,124,125,126,127,128,129,130,131,132,133,134,135,136,137,138,139,140,141,142,143,144,145,146].

However, the characteristic features of the QQSPR fundamental equation, and its definition within the VSS formalism, show that several alternative possibilities can be described, which will be studied next. Such non-classical solutions have in turn provided a collection of many new properties, concepts and application examples related to the tagged sets, VSS and IMP definitions.

Therefore, in order to exploit the QQSPR a plausible alternative to the principal components analysis will be proposed and then, the use of IMP and other techniques to obtain approximate solutions of the QQSPR fundamental equations will also be discussed.

Similarity Matrix Eigenvectors as Basis Sets to Construct the Solutions of the QQSPR Fundamental Equations. Suppose we set the fundamental QQSPR Eq. (40) for a given problem. The secular equation of the involved quantum similarity matrix, $\mathbf{Z}$, can be written as:

$$\mathbf{ZC} = \mathbf{C}\Theta \,, \quad (54)$$

where: $\mathbf{C} = (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_N)$ is the matrix collection of the eigenvectors of the similarity matrix and: $\Theta = \mathrm{Diag}(\theta_1; \theta_2; \ldots; \theta_N)$ is a diagonal matrix made by the ordered eigenvalues in descending order. The eigenvector matrix can be considered orthogonal, that is, the following property holds: $\mathbf{CC}^{\mathrm{T}} = \mathbf{C}^{\mathrm{T}}\mathbf{C} = \mathbf{I}_N$, with the symbol: $\mathbf{C}^{\mathrm{T}}$ indicating matrix transposition. The eigenvector associated to the greater eigenvalue has their entire elements positive definite, according to the Perron–Frobenius theorems [147]. The spectral decomposition:

$$\mathbf{Z} = \mathbf{C}\Theta\mathbf{C}^{\mathrm{T}} = \sum_{I} \theta_I \mathbf{c}_I \mathbf{c}_I^{\mathrm{T}} \quad (55)$$

can be used in Eq. (40), to obtain, after straightforward rearrangements, the equation:

$$\sum_{I} \theta_I (\mathbf{c}_I^{\mathrm{T}} \mathbf{w}) \mathbf{c}_I \simeq |\pi\rangle \,, \quad (56)$$

so, renaming the set of scalars in Eq. (56) as:

$$\gamma_I = \theta_I (\mathbf{c}_I^{\mathrm{T}} \mathbf{w}) \rightarrow |\gamma\rangle = \{\gamma_I\} \,,$$

then, the new equation could be written by means of a linear combination of the similarity matrix eigenvectors:

$$\sum_I \gamma_I \mathbf{c}_I = \mathbf{C}|\gamma\rangle \simeq |\pi\rangle \,. \tag{57}$$

Equation (57) permits us to compute the new coefficients $|\gamma\rangle$ in an obvious way, by using the orthogonal nature of the eigenvector matrix:

$$|\gamma\rangle \simeq \mathbf{C}^{\mathrm{T}}|\pi\rangle \,.$$

In fact, the original fundamental QQSPR equation coefficient vector can be obtained taking into account the spectral decomposition of the similarity matrix inverse, that is:

$$\begin{aligned}\mathbf{w} \simeq \mathbf{C}\Theta^{-1}\mathbf{C}^{\mathrm{T}}|\pi\rangle &= \sum_I \theta_I^{-1}\mathbf{c}_I\mathbf{c}_I^{\mathrm{T}}|\pi\rangle \\ &= \sum_I \left[\theta_I^{-1}\left(\mathbf{c}_I^{\mathrm{T}}|\pi\rangle\right)\right]\mathbf{c}_I = \sum_I \omega_I \mathbf{c}_I \,. \end{aligned} \tag{58}$$

Therefore, Eq. (58) indicates that the coefficient vector, the solution of the previous Eq. (40), may be expressed by a linear combination of the eigenvectors of the similarity matrix too. Then, to every eigenvector, $\mathbf{c}_I$, there is associated a well-defined scalar coefficient, ω_I, which may be used as a reordering rule in order to obtain approximate solutions of the fundamental QQSPR equation. That is, suppose the eigenvectors are now ordered by the decreasing values of the set: $|\omega\rangle = \{\omega_I\}$, then one can write:

$$\begin{aligned}\mathbf{w} &\simeq \sum_I \delta(\omega_I > \varepsilon)\omega_I\mathbf{c}_I + \sum_J \delta(\omega_J \le \varepsilon)\omega_J\mathbf{c}_J \\ &= \mathbf{w}_{\mathrm{a}} + \mathbf{w}_{\mathrm{error}} \,,\end{aligned}$$

where ε is a given threshold splitting the vector construction in an approximate vector, $\mathbf{w}_{\mathrm{a}}$, and the remaining one, $\mathbf{w}_{\mathrm{error}}$, which can be interpreted or used as a residual error vector.

Stochastic Matrix Eigenvectors as Basis Sets to Construct the Solutions of the Fundamental QQSPR Equations

Similar treatments can be designed for the QQSPR fundamental equation of the stochastic matrices, like the one employed for Eq. (50). The problem is that the stochastic matrix **S** is no longer symmetric and the attached eigensystem, apparently appears to be more laboriously solved, than in the case of the symmetric quantum similarity matrix **Z**. The problem has been already discussed in a general manner [59], so here only some simplified discussion will be given.

Suppose the secular equation, attached to the stochastic matrix **S** is written as:

$$\mathbf{SX} = \mathbf{X}\Sigma \tag{59}$$

substituting the matrix **S** by the expression of the row stochastic transformation as in Eq. (50), then:

$$\mathbf{D}^{-1}\mathbf{ZX} = \mathbf{X}\Sigma \,,$$

which can readily be transformed by simple matrix manipulations and by using the square root of the diagonal matrix **D**, into the new secular equation:

$$\mathbf{D}^{-\frac{1}{2}}\mathbf{ZD}^{-\frac{1}{2}}\mathbf{D}^{\frac{1}{2}}\mathbf{X} = \mathbf{X}\Sigma \,,$$

which from here, calling: $\mathbf{A} = \mathbf{D}^{-\frac{1}{2}}\mathbf{ZD}^{-\frac{1}{2}}$ and $\mathbf{Y} = \mathbf{D}^{\frac{1}{2}}\mathbf{X}$, a new equivalent secular equation is readily made:

$$\mathbf{AY} = \mathbf{Y}\Sigma \,. \tag{60}$$

Equation (60) has the advantage that the matrix A is symmetric, hence the eigenvector matrix Y appears to be orthogonal: $\mathbf{Y}^{\mathrm{T}}\mathbf{Y} = \mathbf{YY}^{\mathrm{T}} = \mathbf{I}$. Then, the sought eigenvectors of the stochastic matrix can be obtained by using the relationship between **Y**and **X**, that is:

$$\mathbf{X} = \mathbf{D}^{-\frac{1}{2}}\mathbf{Y} \,.$$

The matrix **D**, in these circumstances acts as a metric with respect to the eigenvectors of **S**, as one can see that the orthogonality relationships:

$$\mathbf{X}^{\mathrm{T}}\mathbf{DX} = \mathbf{XDX}^{\mathrm{T}} = \mathbf{I} \,,$$

hold, due to the orthogonality of the eigenvector matrix **Y**.

This proves finally that, in any case, Eq. (50) can be solved in the same way as previously commented for Eq. (40), simply by using the appropriate spectral decomposition:

$$\mathbf{S} = \mathbf{X}\Sigma\mathbf{X}^{\mathrm{T}} = \sum_I \sigma_I \mathbf{x}_I\mathbf{x}_I^{\mathrm{T}} \,.$$

Similarity Matrix IMP Decomposition in Order to Construct Approximate Solutions of the QQSPR Fundamental Equations

QQSPR Fundamental Equation over VSS. Going back to the QQSPR fundamental Eq. (40), now one can consider the positive definite nature of the elements, which appear to build the quantum similarity matrix. This can be expressed by means of the symbol: $\mathbf{Z}* > 0$[2]. The structure

[2]By the symbol: $\mathbf{A}* > 0$, applied to an arbitrary matrix $\mathbf{A} = \{a_{ij}\}$, is meant that $\forall i, j$: $a_{ij} > 0$.

of the property vectors can also be taken into account. In the case when the following characteristic also holds: $|\pi\rangle * > 0$, for the involved QOS property or activity vector, then Eq. (40) can be associated to some linear transformation occurring on a VSS, for it can be written:

$$\mathbf{Zw} = \sum_I w_I \mathbf{z}_I \simeq |\pi\rangle \,, \tag{61}$$

showing that a linear combination of vectors, belonging to a VSS, has to be brought into another vector belonging to a VSS.

This situation could only be *generally* achieved by using the condition: $\mathbf{w}* > 0$, which means that the linear system solution also shall belong to the VSS. Equ. (61) above becomes a constrained linear system of equations, since one is seeking solutions for which:

$$\mathbf{Zw} \simeq |\pi\rangle \wedge \mathbf{w}* > \mathbf{0} \,.$$

Approximate Restricted Solutions of Fundamental QQSPR Equation in VSS. An approximate solution of the QQSPR fundamental equation can be obtained in the following way. As all the involved columns of the problem belong to a VSS, they can be decomposed by means of an IMP in terms of some IMP square powers of real matrices. If the treatment has to be more general the squared module of some complex matrices can be alternatively employed, but the treatment becomes slightly more difficult and the needed set of symbols heavier, so just real matrices will be supposed in this discussion. Therefore, owing to these considerations one can write:

$$\mathbf{Z} = \mathbf{A} * \mathbf{A} \wedge \mathbf{w} = \mathbf{x} * \mathbf{x} \wedge |\pi\rangle = \mathbf{p} * \mathbf{p} \,.$$

So Eq. (61) can now be rewritten as:

$$(\mathbf{A} * \mathbf{A})(\mathbf{x} * \mathbf{x}) \simeq \mathbf{p} * \mathbf{p} \,, \tag{62}$$

suggesting an alternative approximate equation, which may be written in the following terms:

$$(\mathbf{Ax}) * (\mathbf{Ax}) \simeq \mathbf{p} * \mathbf{p} \,, \tag{63}$$

which has been obtained in turn, simply using the plausible approximation:

$$(\mathbf{A} * \mathbf{A})(\mathbf{x} * \mathbf{x}) \approx (\mathbf{Ax}) * (\mathbf{Ax}) \,.$$

However, the approximate Eq. (63), suggests that the new linear system:

$$\mathbf{Ax} \simeq \mathbf{p} \tag{64}$$

can now be solved, as it does not have to be submitted to any restriction at all, then:

$$\mathbf{x} \simeq \mathbf{A}^{-1}\mathbf{p} \,,$$

and finally, the approximate solutions of the original system can be written as:

$$^{a}\mathbf{w} = \mathbf{x} * \mathbf{x} \simeq (\mathbf{A}^{-1}\mathbf{p}) * (\mathbf{A}^{-1}\mathbf{p}) \,,$$

however this is sufficient to ensure:

$$\mathbf{w} \approx {}^{a}\mathbf{w}* > 0 \,.$$

The only problem, which now arises, is the existence of an inverse of the IMP square root of a non-singular matrix. Since the system (64) furnishes approximate solutions to the original problem (61) has to be found, there will be no major problem then to use approximate solutions in the least squares sense of the Eq. (64), as a way to obtain the approximate solution of the QQSPR fundamental equation, restricted to belonging to a VSS.

Convex Conditions Imposed on the Solution Vector of the QQSPR Fundamental Equations

Generating Vector Considerations. The associated problem, to a form like Eq. (61), can also be solved, for instance, as in the well-known ASA fitting procedure [75,76,77,78,79,80,81,82,83]. That is, by using a convex condition on the solution vector: $K(\mathbf{w})$, with the additional meaning that the solution is now forced to belong not only to a VSS, but also to the unit shell.

If the solution of the linear Eq. (61) has to be found as an element of a VSS, $\mathbf{w}* > 0$, necessarily it has to be expressible as an IMP power of some generating real vector like: $\mathbf{w} = \mathbf{x} * \mathbf{x}$. In choosing the convex conditions over the solution: $K(\mathbf{w})$, then the additional restriction is admitted to hold too:

$$\langle \mathbf{w} \rangle = \langle \mathbf{x} * \mathbf{x} \rangle = 1 \,. \tag{65}$$

However, this becomes the same as to consider that the generating vector $R(\mathbf{x} \to \mathbf{w})$ is normalized. Orthogonal transformations on the generating vector leaves the vector norm invariant, that is: whenever the condition (65) holds, and an orthogonal transformation $\mathbf{U}$ is performed on the generating vector, still the generating rule and the associated convex conditions apply: $R(\mathbf{Ux} \to {}^{U}\mathbf{w}) \wedge K({}^{U}\mathbf{w})$. Such an idea has been applied to obtain the ASA approximate density functions, using *elementary Jacobi rotations* [148] as a source of orthogonal transformations.

Stochastic Transformations. Still more interesting appears the structure of the fundamental QQSPR equation, when the stochastic transform of the similarity matrix is considered. Equation (40) can thus be multiplied by the inverse of the diagonal matrix **D** on the right as defined in (27), providing:

$$\mathbf{Z}\mathbf{D}^{-1}\mathbf{D}\mathbf{w} = |\pi\rangle \,,$$

which can be transformed into:

$$\mathbf{S}\mathbf{v} = |\pi\rangle \,, \tag{66}$$

whenever it is considered that the following equalities hold:

$$\mathbf{S} = \mathbf{Z}\mathbf{D}^{-1} \wedge \mathbf{v} = \mathbf{D}\mathbf{w} \,. \tag{67}$$

Equation (66) can be also written as a linear combination of the columns of the stochastic matrix $\mathbf{S} = \{s_I\}$:

$$\sum_I v_I \mathbf{s}_I = |\pi\rangle \,.$$

Therefore, this is the same considering the σ-shells of the vectors **v** and $|\pi\rangle$ as being almost the same, as:

$$\langle|\pi\rangle\rangle = \left\langle \sum_I v_I \mathbf{s}_I \right\rangle = \sum_I v_I \langle \mathbf{s}_I\rangle = \sum_I v_I = \langle \mathbf{v}\rangle \,.$$

Thus, if the vector $|\pi\rangle$ is transformed so as to become a unit shell element, this will be completely equivalent to applying the same transformation into the transformed unknown vector **v**. Therefore, the following implications are straightforwardly deduced:

$$|\pi\rangle \in S(1) \rightarrow \mathbf{v} \in S(1) \rightarrow K(\mathbf{v}) \,.$$

Demonstrating that in the stochastic QQSPR fundamental Eq. (66) case, the solution contained within a given VSS amounts to the same as obtaining a convex combination of the stochastic matrix columns.

Stochastic QQSPR Least Squares Solution via Jacobi Rotations. After all the previous discussions, there appears to be another possibility which has remained unexplored. Starting from the transformed Eq. (66), with the appropriate definitions (67) in mind, one can seek an approximate solution of the stochastic equation in the least squares sense, defining the quadratic error function by means of the difference vector:

$$|\delta\rangle = \mathbf{S}\mathbf{v} - |\pi\rangle \,,$$

whose Euclidean norm furnishes the quadratic error function, expressible in terms of a scalar product or the inward product sum:

$$\begin{aligned}\varepsilon^{(2)} = \langle\delta|\delta\rangle &= (\mathbf{S}\mathbf{v} - |\pi\rangle)^{\mathrm{T}}(\mathbf{S}\mathbf{v} - |\pi\rangle) \\ &= \langle(\mathbf{S}\mathbf{v} - |\pi\rangle) * (\mathbf{S}\mathbf{v} - |\pi\rangle)\rangle \,.\end{aligned}$$

One easily arrives at the quadratic form:

$$\varepsilon^{(2)}(\mathbf{v}) = \mathbf{v}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{v} - \mathbf{v}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}|\pi\rangle - \langle\pi|\mathbf{S}\mathbf{v} + \langle\pi|\pi\rangle \,, \tag{68}$$

however, the quadratic error function optimization has to be carried out preserving the condition of convexity $K(\mathbf{v})$ on the solution vector, otherwise one risks obtaining solutions that do not belong to the unit shell. In order to obtain an appropriate algorithm to perform this task, the following considerations can be taken into account.

The vector v can be expressed as an IMP of an auxiliary vector a, that is:

$$\mathbf{v} = \mathbf{a} * \mathbf{a} \rightarrow \forall I\colon \mathrm{v}_I = a_I^2 \,, \tag{69}$$

then the belonging of v to the unit shell is equivalent to the Euclidean normalization of a:

$$\langle\mathbf{v}\rangle = \sum_I \mathrm{v}_I = \sum_I a_I^2 = \langle\mathbf{a}|\mathbf{a}\rangle \,.$$

The quadratic error function (68) can be expressed in terms of the auxiliary vector:

$$\varepsilon^{(2)}(\mathbf{a}) = \mathbf{a}^{[2]\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S}\mathbf{a}^{[2]} -^{\mathrm{T}} \mathbf{a}^{[2]\mathrm{T}}\mathbf{S}^{\mathrm{T}}|\pi\rangle - \langle\pi|\mathbf{S}\mathbf{a}^{[2]} + \langle\pi|\pi\rangle \,, \tag{70}$$

where the symbol: $\mathbf{a}^{[2]} = \mathbf{v} = \mathbf{a} * \mathbf{a}$ has been used. Also, employing to simplify the notation the following conventional symbols:

$$\begin{aligned}\mathbf{H} = \mathbf{S}^{\mathrm{T}}\mathbf{S} = \{H_{IJ}\} \\ \wedge\ \mathbf{h} = \mathbf{S}^{\mathrm{T}}|\pi\rangle = \{h_I\} \wedge \mathbf{h}^{\mathrm{T}} = \langle\pi|\mathbf{S} = \{h_I\} \\ \wedge\ \eta = \langle\pi|\pi\rangle \,,\end{aligned}$$

so Eq. (70) can be explicitly written as:

$$\varepsilon^{(2)}(\mathbf{a}) = \sum_I \sum_J H_{IJ} a_I^2 a_J^2 - 2\sum_I h_I a_I^2 + \eta \,. \tag{71}$$

Starting with an approximate normalized auxiliary vector, orthogonal transformations can be performed, preserving the norm, thus keeping the condition $K(\mathbf{v}) = K(\mathbf{a}^{[2]})$ constant along the optimization of Eq. (71). Orthogonal transformations can be chosen as elementary Jacobi rotations [148], which at every application over the vector a, change two chosen elements $\{a_P; a_Q\}$ into a pair of new

rotated ones $\{a_P^R; a_Q^R\}$, according to the well-known algorithm:

$$\begin{aligned} a_P^R &\leftarrow c a_P - s a_Q \\ a_Q^R &\leftarrow s a_P + c a_Q \,, \end{aligned} \tag{72}$$

where $\{c, s\}$ are the cosine and the sine of the rotation, with the additional obvious relationship: $c^2 + s^2 = 1$.

Over the generating vector coefficients in Eq. (69) it is easy to apply the EJR represented by Eq. (72), and then, the variation of the quadratic error $\delta\varepsilon^{(2)}$, with respect to the active pair of elements $\{a_P; a_Q\}$ may be easily expressed.

Taking also into account that the quadratic elements, for example, will transform and yield variations like:

$$\begin{aligned} \delta a_P^2 &\to s^2(a_Q^2 - a_P^2) - 2cs a_P a_Q \\ \delta(a_P a_Q) &\to cs(a_P^2 - a_Q^2) - 2s^2 a_P a_Q \\ \delta a_Q^2 &\to s^2(a_Q^2 - a_P^2) + 2cs a_P a_Q = -\delta a_P^2 \,. \end{aligned}$$

A Jacobi rotation as shown in the expression (72) will produce a variation in the quadratic error (71) in the chosen rotated elements, when taking also into account the symmetric nature of the matrix H, of the form:

$$\begin{aligned} \delta\varepsilon^{(2)}(\mathbf{a}) = {} & H_{PP}(\delta a_P^2)^2 + H_{QQ}(\delta a_Q^2)^2 + 2H_{PQ}\delta a_P^2 \delta a_Q^2 \\ & + 2\sum_{I\neq P,Q} a_I^2 (H_{IP}\delta a_P^2 + H_{IQ}\delta a_Q^2) \end{aligned}$$

so, also needed are the quartic variations of the auxiliary vector elements, which can be easily computed as in the second-order case.

Substituting such variations into the corresponding equation and collecting terms one finally arrives at a quartic polynomial on the rotation sine:

$$\delta\varepsilon^{(2)} = E_{04}s^4 + E_{13}cs^3 + E_{02}s^2 + E_{11}cs \,, \tag{73}$$

where the parameters $\{E_{IJ}\}$, appearing in Eq. (73), are described as follows:

$$\begin{aligned} E_{04} &= \Theta\left[\left(a_P^2 - a_Q^2\right)^2 - 4a_P^2 a_Q^2\right] \\ E_{13} &= 4\Theta\left(a_P^2 - a_Q^2\right) a_P a_Q \\ E_{02} &= 4\Theta a_P^2 a_Q^2 - 2\left(a_P^2 - a_Q^2\right) G \\ E_{11} &= -4 a_P a_Q G \end{aligned}$$

using the following auxiliary terms:

$$\Theta = H_{PP} + H_{QQ} - 2H_{PQ} \,,$$

and

$$\begin{aligned} G = {} & \sum_{I\neq P,Q} a_I^2 (H_{IP} - H_{QI}) + a_P^2 H_{PP} - a_Q^2 H_{QQ} \\ & - \left(a_P^2 - a_Q^2\right) H_{PQ} - h_P + h_Q \,. \end{aligned}$$

The optimal sine can be chosen with the null gradient condition $\mathrm{d}\delta\varepsilon^{(2)}/\mathrm{d}s = 0$, taking into account that: $s/c = t$ and that: $\mathrm{d}c/\mathrm{d}s = -t$, then:

$$\frac{\mathrm{d}\delta\varepsilon^{(2)}}{\mathrm{d}s} = -c\left(T_1 t^2 - 2T_2 t - T_3\right) = 0 \,, \tag{74}$$

holds with the auxiliary definitions:

$$\begin{aligned} T_1 &= E_{13}s^2 + E_{11} \\ T_2 &= 2E_{04}s^2 + E_{02} \\ T_3 &= 3E_{13}s^2 + E_{11} \,. \end{aligned}$$

The best Jacobi rotation angle is found solving the quadratic polynomial equation in the EJR tangent {t}, appearing in expression (74). The optimization can be conducted through an iterative procedure, until the global variation of Jacobi rotation angles or the quadratic error integral function become negligible. The interested reader is conducted to the references [78,79,80] for more details, where a complete account of all the Jacobi rotation techniques can be found. A simplified algorithm can be also used and it will be briefly commented upon here. The procedure is based on the fact that sine and cosine can be written in function of the rotation angle:

$$\begin{aligned} s &= \alpha - \frac{1}{6}\alpha^3 + O(5) \\ c &= 1 - \frac{1}{2}\alpha^2 + O(4) \,, \end{aligned}$$

and for small angles it is only necessary to use, up to second order:

$$s \simeq \alpha \wedge c \simeq 1 - \alpha \,,$$

so Eq. (73) can be transformed into a second-order polynomial in the rotation angle:

$$\delta\varepsilon^{(2)} \simeq (E_{02} - E_{11})\alpha^2 + E_{11}\alpha \,,$$

which submitted to the extremum conditions yields:

$$\alpha \simeq \frac{1}{2}\left(1 - E_{02}E_{11}^{-1}\right)^{-1} \simeq s \,.$$

Non-Linear Terms and Extended Wave Functions

Sobolev Spaces

From early times, quantum mechanics has been emphasizing not only the role of well-behaved wave functions, but also the relevance of their gradients and Laplacian forms. The reason for such requirements, necessarily holding on

the current wave functions, can be simply connected to the presence, in the Schrödinger equation set up, of a second-order derivative, the result of a Laplace operator application, associated to the quantum system kinetic energy term, see for example the references [28,149].

Usually, the adequate quantum mechanical behavior of the wave function is focused, among other simple and obvious mathematical features, to the *compulsive* property that wave functions have to be square summable. In some reference books such a property has been promoted to the category of a postulate [150] and in the very early development times of quantum mechanics [98] has been interpreted by Born as the fact that the square module of the wave function can be associated to a probability density function. It was von Neumann [97], who related such properties, among other crucial quantum mechanical theoretical elements, with the mathematical structures of Hilbert–Banach spaces [94,151,152]. More recently, Landau and Lifshitz [153] described the role of the wave function gradient as a descriptor of infinitesimal translations and rotations. These authors settled as well the use of the square module of the wave functions gradient, in order to obtain an alternative kinetic energy expectation value expression, more likely related to the statistical formalism than the Laplace operator form.

On the other hand no utilization has been reported of the so-called *Sobolev spaces* [154] in applied quantum mechanics, at least to our knowledge. Curiously enough, Sobolev spaces where defined as early as 1938, and apparently have been of practical use in some remotely related theoretical landscape, associated to generalized relativity applications [155]. It was not until recently that Sobolev spaces were proposed by us as a vehicle to take into account the role of the wave function squared module:$|\Psi|^2$, as well as to make *simultaneously relevant* the presence of the wave function gradient squared module:$|\nabla\Psi|^2$ in an extended density function. In all this previous work both terms were also presented as forming part of a new quantum mechanical composite norm. In this way, the classical quantum mechanical Banach space has been transformed into a Sobolev space structure [14,156,157], without losing generality, but gaining flexibility instead.

Sobolev spaces can be defined in several ways, leading all of them to simple forms, ready to be used in reinterpreting the approximate solution of the Schrödinger equation and prone to be included with immediate applications, such as those found among the references [156,157]. They can be constructed in such a way as possessing extended forms even more complex, in order to be used to include arbitrary non-linear terms in the same equation [14]. This can be understood by recognizing the fact that the Banach space can be considered the limiting simplified form of a quite large collection of Sobolev spaces.

Quantum Mechanical Hilbert and Banach Spaces. In order to present the Sobolev spaces step by step, the simplest formalism will be defined first, and other extended possibilities will be described later on. To achieve this objective, suppose a quantum mechanical wave function Hilbert space, which can formally be described as:

$$H_\infty(\mathbf{C}) = \{\Psi(\mathbf{r})|\mathbf{r} \in V_P(\mathbf{R}) \wedge \Psi(\mathbf{r}) \in \mathbf{C}\}\,, \tag{75}$$

where the symbol $\mathbf{r}$, used as the wave function variables, shall be considered as a vector, containing all the necessary particle position coordinates as its components. The number of particles is shortly noted with the dimension P of the coordinates vector space. The wave function elements of the Hilbert space (75), possess as a sine qua non condition, the following well-known property about the existence of a positive definite density function, which is remembered here, just to present the notation that will be hereafter employed:

$$\forall\Psi(\mathbf{r}) \in H_\infty(\mathbf{C}) \rightarrow \exists\rho(\mathbf{r}) = |\Psi(\mathbf{r})|^2 \in H_\infty(\mathbf{R}^+)\,. \tag{76}$$

Besides, the density function attached to every wave function, as proposed in Eq. (76), can be seen as belonging to a Hilbert *semispace*, $H_\infty(\mathbf{R}^+)$, where all function values and coefficients are strictly allowed to be positive real numbers only. That is: in the same way as in the Hilbert space (75), one can write the corresponding formal definition for the density function semispace:

$$H_\infty(\mathbf{R}^+) = \left\{\rho(\mathbf{r})|\mathbf{r} \in V_P(\mathbf{R}) \wedge \rho(\mathbf{r}) \in \mathbf{R}^+\right\}\,, \tag{77}$$

where the dimension of the coordinates vector space, $V_P(\mathbf{R})$, containing the density function variables, has the same meaning as in definition (75). Moreover, the Hilbert space (75) is a Banach space, as all of their elements shall compulsively fulfill the normalization condition:

$$\forall\Psi(\mathbf{r}) \in H_\infty(\mathbf{C}) \rightarrow \int_D |\Psi(\mathbf{r})|^2\,\mathrm{d}\mathbf{r} = 1\,, \tag{78}$$

which obviously amounts to the same as imposing on every element of the Hilbert semispace (77), a *convexity* condition:

$$\forall\rho(\mathbf{r}) \in H_\infty(\mathbf{R}^+) \rightarrow \int_D \rho(\mathbf{r})\,\mathrm{d}\mathbf{r} = 1\,. \tag{79}$$

Gradient of the Wave Function. As was previously commented, the whole quantum mechanical Hilbert space elements shall present other existence properties, mainly related to their derivatives, for instance:

$$\forall\Psi(\mathbf{r}) \in H_\infty(\mathbf{C}) \rightarrow \exists\nabla\Psi(\mathbf{r})\,, \tag{80}$$

where the *nabla* operator ∇ refers to the gradient with respect to the vector coordinates $\mathbf{r}$. That is, formally it can be also written: $\nabla\Psi \equiv \partial\Psi/\partial\mathbf{r}$. The ordering of the resultant gradient vector elements can be somewhat arbitrary; this means that they can be adapted to the structure of the operating mathematical context. In addition, associated to this mentioned component ordering, the resultant gradient vector can be considered to belong to some appropriate Cartesian product of the initial Hilbert space (75), which can be generally defined and noted in a simplified fashion as:

$$H_{\infty}^{(P)}(\mathbf{C}) = \mathop{\times}_{I=1}^{P} H_{\infty}(\mathbf{C}) \,, \tag{81}$$

because in any ordering case, the resultant gradient vectors will depend on the particle number P. At the same time, the gradients of type (80), can also be easily associated to squared gradient modules, which shall belong to some Hilbert semispace, very similar to the one defined in Eq. (77):

$$\forall\nabla\Psi(\mathbf{r}) \in H_{\infty}^{(P)}(\mathbf{C}) \to \exists\kappa(\mathbf{r}) = |\nabla\Psi(\mathbf{r})|^2 \in H_{\infty}(\mathbf{R}^{+}) \,, \tag{82}$$

where the positive definite function $\kappa(\mathbf{r})$, will produce, when integrated, twice the quantum mechanical kinetic energy expectation value $\langle K\rangle$ of the attached system:

$$\int_D \kappa(\mathbf{r})\,\mathrm{d}\mathbf{r} = \int_D |\nabla\Psi(\mathbf{r})|^2\,\mathrm{d}\mathbf{r} = 2\langle K\rangle \,, \tag{83}$$

which as is well known, can be described alternatively like the classical quantum mechanical expectation value of the Laplace operator:

$$2\langle K\rangle = -\int_D \Psi^*(\mathbf{r})\nabla^2\Psi(\mathbf{r})\,\mathrm{d}\mathbf{r} \,, \tag{84}$$

just employing Green's theorem [158].

The interesting thing to be said now consists in proposing some sentences on the nature of the integrals (83) and (84), which are real and positive definite as kinetic energy shall be, either classically speaking or quantum mechanically, thus providing the integral (83) with a well-defined structure, capable of being interpreted as a norm. It should also be noted that the imaginary unit, usually employed in the quantum mechanical definition of linear momentum does not need to be used here in front of the *nabla* operator. The reason can be found in the fact that the imaginary unit has no active role in the above definitions, unless a Hermitian matrix representation is needed for the ∇ operator. Such an **imaginary** factor has to be present in such a Hermitian representation case, because the matrix associated to the bare *nabla* operator is Skew-Hermitian, that is:

$$\int_D \Psi_I^*(\mathbf{r})(\nabla\Psi_J(\mathbf{r}))\,\mathrm{d}\mathbf{r} = -\int_D (\nabla\Psi_I(\mathbf{r}))^*\Psi_J(\mathbf{r})\,\mathrm{d}\mathbf{r} \,,$$

a property which can be easily interpreted as a consequence of the application of Green's theorem again.

The Simplest Sobolev Space. In any case, the existence of the wave function norm (78) and the subsequent convexity condition (79), can both be recognized as the parallel properties holding for the gradient of the wave function, and corresponding to Eq. (83), which proves collaterally the positive definite nature of the quantum mechanical kinetic energy expectation value. Thus, if the sequence of equations from (75) up to (83) must hold simultaneously, just to obtain a coherent mathematical structure within the quantum mechanical framework, it is feasible to consider that both Banach spaces (75) and (82), can be supposedly forming a composite norm, in the way of the following definition present within the equation shown below. Simplifying the wave function notation from the variable dependence, in order to ease the form of the subsequent equations, one can define the following norm:

$$\forall\Psi \in H_{\infty}(\mathbf{C}) \to \\ \exists\|\Psi\|_1^1 = \int_D |\Psi|^2\,\mathrm{d}\mathbf{r} + \int_D |\nabla\Psi|^2\,\mathrm{d}\mathbf{r} = 1 + 2\langle K\rangle \,. \tag{85}$$

Such a composition provides the first definition of the simplest element among the collection of all possible Sobolev spaces, which can be connected to quantum theory. Thus, it can be assumed from now on that the most adequate quantum mechanical wave function space structure is a *Sobolev space*.

Sobolev Spaces. The notation for the norm (85) will be made immediately obvious, by means of defining a general Sobolev space norm as:

$$\forall\Psi \in H_{\infty}(\mathbf{C}) \to \\ \|\Psi\|_{\alpha}^{\beta} = \sum_{a=1}^{\alpha}\int_D \sum_{b=0}^{\beta} |\nabla^b\Psi|^{2a}\,\mathrm{d}\mathbf{r} \,. \tag{86}$$

The usual Hilbert–Banach space norm can be retrieved from definition (86), simply supposing that the null power of the gradient operator can be substituted by the identity: $\nabla^0 \equiv I$, and using afterwards: $\alpha = 1 \wedge \beta = 0$. In addition, the earlier Sobolev space norm, simplified as in

Eq. (85), is also found employing Eq. (86), but choosing: $\alpha = \beta = 1$.

However, although Eq. (86) contains the classical quantum mechanical Sobolev ∇-norm, it implicitly possesses the restriction consisting in that both wave function and gradient norm powers shall be the same in any circumstance. Consequently, they cannot be monitored as independent terms in the norm definition. An appropriate choice to avoid this situation may be described with the more general formulation:

$$\forall \Psi \in H_\infty(\mathbf{C}) \rightarrow$$
$$\|\Psi\|_\alpha^{\beta,\gamma} = \sum_{a=1}^{\alpha} \int_D |\Psi|^{2a}\, \mathrm{d}\mathbf{r} + \sum_{c=1}^{\gamma} \int_D \sum_{b=1}^{\beta} |\nabla^b \Psi|^{2c}\, \mathrm{d}\mathbf{r} \tag{87}$$

Therefore, Eq. (87) will transform into expression (86), whenever: $\alpha = \gamma$. In order to avoid further interpretation problems, the squared module of the *nabla* powers has to be considered a contraction operation; or has to be considered a scalar product of the corresponding matrix elements, represented by the result of the operation $\nabla^b \Psi$:

$$|\nabla^b \Psi|^{2c} \equiv |\langle \nabla^b \Psi | \nabla^b \Psi \rangle|^c \,. \tag{88}$$

Nested Summation Symbols. This last remark, represented by the Eq. (88), can be alternatively written in a very elegant manner employing the definition of an *inward matrix product* (IMP), already discussed.

Taking the IMP definition into account, then in Eqs. (86) and (87) it can be understood that the present square modules are computed over the resultant wave function derivative hypermatrices as:

$$|\nabla^b \Psi|^{2c} = |\langle (\nabla^b \Psi)^* * (\nabla^b \Psi) \rangle|^c \,,$$

where the symbol $\langle\ \rangle$, associated to any matrix, means a sum of the whole matrix elements, for instance:

$$\forall \mathbf{P} = \{p_{ij}\} \rightarrow \langle \mathbf{P} \rangle = \sum_i \sum_j p_{ij} \,, \tag{89}$$

which constitutes a definition possessing obvious generalization possibilities, within any kind of hypermatrix structure.

This generalization power can be easily seen, taking into account the *nested summation symbol* (NSS) formalism, which was developed several years ago, see references [159,160] for example. Then, using NSS, the expression of the total sum of the elements of an arbitrary hypermatrix can be generally written without any further problem. A NSS is a symbolic device, which has a linear operator nature, and in this way resumes the presence of an undefined number of nested sums and corresponds to an easily programmable algorithm, which generalizes in practice an indefinite number of do loops. In turn, a NSS acts over any kind of complex expression, bearing all the involved indices present within the sums, that is:

$$\sum\nolimits_N (\mathbf{i}) \phi(\mathbf{i}) \equiv \sum_{i_1}^{n_1} \sum_{i_2}^{n_2} \ldots \sum_{i_N}^{n_N} \phi(i_1; i_2; \ldots; i_N) \,,$$

where by the index vector $\mathbf{i}$ it is understood: $\mathbf{i} = (i_1; i_2; \ldots; i_N)$. Thus, if by the definition the following subindex structure is assumed:

$$\mathbf{Z} = \{z_{i_1 i_2 \ldots i_N}\} \equiv \{z(\mathbf{i})\} \,,$$

by which is represented any $(n_1 \times n_2 \times \ldots n_N)$-dimensional hypermatrix element, then the symbolic device associated to the total summation of the elements, particularly defined in Eq. (89), can be generally described by means of the compact NSS expression:

$$\langle \mathbf{Z} \rangle = \sum\nolimits_N (\mathbf{i}) z(\mathbf{i}) \,.$$

Extended Wave and Density Functions

Sobolev spaces appear, after the previous discussion, as a very general kind of extended Hilbert–Banach spaces, which within the quantum mechanical framework are able to put into a unique statement the nature of both the wave function and its gradient. Alternatively, they can produce completely general structures, somehow involving usual quantum mechanical operators, attachable to any system observable. It is a matter of straightforward analysis to translate the subjacent Sobolev mathematical structure into the Hilbert space elements themselves, producing a new breed of spaces, which can be obviously called *Hilbert–Sobolev spaces* [3]. As has been done before, within the previous description of Sobolev spaces, the extension of the wave function and the possible application of the resultant formal structure will be here gradually discussed.

Extended Wave Functions. By a ∇-extended wave function $|\Phi\rangle$ has been understood a composite column vector, or alternatively a diagonal matrix, if one prefers, whose elements are the original wave function Ψ and its gradient $\nabla\Psi$. That is:

$$|\Phi\rangle = \begin{pmatrix} \Psi \\ \nabla\Psi \end{pmatrix} \equiv \mathrm{Diag}(\Psi; \nabla\Psi) = \begin{pmatrix} \Psi & 0 \\ 0 & \nabla\Psi \end{pmatrix} \,. \tag{90}$$

[3] In the case where the nature of the operator or the set of operators, active in the Sobolev norm definition, has to be specified, then the notation Hilbert–Sobolev ∇-, Ω- or Υ-spaces, can be obviously employed.

The column vector form, as will be discussed later on, better represents some applications and the mathematical manipulations one can perform over them; while in other cases the diagonal matrix structure produces more elegant expressions and it is easier to deal with. However, both choices provide equivalent results. This is so because the proposed representations constitute the elements of a pair of isomorphic vector spaces.

The previous discussion on Sobolev spaces permits us to define the extended wave function within a general Hermitian operator scheme, simply as:

$$|\Phi\rangle = \begin{pmatrix} \Psi \\ \Omega\Psi \end{pmatrix} \equiv \mathrm{Diag}(\Psi; \Omega\Psi) = \begin{pmatrix} \Psi & 0 \\ 0 & \Omega\Psi \end{pmatrix} . \quad (91)$$

The definition of Eq. (91) can be called an *Ω-extended wave function*. For example, the quantum mechanical complementary definition of the ∇-extended wave function (90) can be easily written employing the position vector $\mathbf{r}$, that is:

$$|\Theta\rangle = \begin{pmatrix} \Psi \\ \mathbf{r}\Psi \end{pmatrix} \equiv \mathrm{Diag}(\Psi; \mathbf{r}\Psi) = \begin{pmatrix} \Psi & 0 \\ 0 & \mathbf{r}\Psi \end{pmatrix} . \quad (92)$$

producing an $\mathbf{r}$-extended wave function accordingly.

By inspection of the adopted structure until now, extended wave functions can be also considered as the result of applying some adequate operator over the original Schrödinger wave function. For instance, within the already-mentioned diagonal formalism of the Ω-extended wave function (91), defining the diagonal operator:

$$\Gamma = \mathrm{Diag}(\mathbf{I}; \Omega) ,$$

where $\mathbf{I}$ is the unit operator, it will be sufficient to see that:

$$\begin{aligned} |\Phi\rangle = \Gamma[\Psi] &= \mathrm{Diag}(\mathbf{I}; \Omega)[\Psi] \\ &= \mathrm{Diag}(\mathbf{I}[\Psi]; \Omega[\Psi]) \\ &= \mathrm{Diag}(\Psi; \Omega\Psi) . \end{aligned} \quad (93)$$

The same can be said if the corresponding vector operator is constructed by means of the vector structure:

$$\Gamma = \begin{pmatrix} \mathbf{I} \\ \Omega \end{pmatrix} ,$$

which, upon application over the original scalar wave function form, permits us to alternatively obtain the isomorphic vector picture of the diagonal expression (93).

Energy Expectation Values. Returning to the ∇-extended wave function in Eq. (90), it is easy to see how the energy expectation value of the associated Schrödinger equation can be expressed, without loosing any information, when writing the final form it takes. For this purpose, the appropriate Hamilton operator, $\mathbf{H}$, can be structured by means of a diagonal form, as:

$$\mathbf{H} = \mathrm{Diag}(\mathbf{U}; \tfrac{1}{2}\mathbf{I}) = \begin{pmatrix} \mathbf{U} & 0 \\ 0 & \frac{1}{2}\mathbf{I} \end{pmatrix} , \quad (94)$$

where the symbol: $\mathbf{U}$, corresponds to the potential energy operator and $\mathbf{I}$ is just a unit operator built to fit the adequate dimensions of the extended wave function (90) gradient part. Using Eq. (94) and the ∇-extended wave function (90) in the appropriate way, it is immediate to write:

$$\begin{aligned} E = \langle \mathbf{H} \rangle = \langle \Phi | \mathbf{H} | \Phi \rangle &= \langle \Psi | \mathbf{U} | \Psi \rangle + \tfrac{1}{2} \langle \nabla\Psi | \nabla\Psi \rangle \\ &= \langle U \rangle + \langle K \rangle . \end{aligned} \quad (95)$$

In the same way, whenever the normalization of the wave function holds: $\langle \Psi | \Psi \rangle = 1$, the ∇-extended wave function (90) can be manipulated in order to obtain a positive definite norm like:

$$\langle \Phi | \Phi \rangle = \langle \Psi | \Psi \rangle + \langle \nabla\Psi | \nabla\Psi \rangle = 1 + 2\langle K \rangle , \quad (96)$$

which corresponds to the same result as the one provided by the norm obtained in the definition of the simplest Sobolev space, as presented first in Eq. (85). Obviously, such a Sobolev space can be interpreted as a composite Hilbert space, whose elements are defined by the ∇-extended wave function (90). That is, employing the Cartesian product of the Hilbert spaces (75) and (81), upon reordering the ordered pair in the form of column vector:

$$H_\infty \times H_\infty^{(P)} = \left\{ |\Phi\rangle = \begin{pmatrix} \Psi \\ \nabla\Psi \end{pmatrix} \right\} . \quad (97)$$

Extended Density Functions. The Hilbert semispace corresponding to the Hilbert space (97), can be supposedly formed by the total density functions and computed by using the trace of the extended wave function tensor product, for instance:

$$\begin{aligned} \tau(\mathbf{r}) = \mathrm{Tr}\, |\Phi\rangle\langle\Phi| = \mathrm{Tr} \begin{pmatrix} |\Psi|^2 & \Psi(\nabla\Psi)^* \\ (\nabla\Psi)\Psi^* & |\nabla\Psi|^2 \end{pmatrix} \\ = \rho(\mathbf{r}) + \kappa(\mathbf{r}) ; \end{aligned} \quad (98)$$

that is: as the superposition of the electronic density $\rho(\mathbf{r})$ and the kinetic energy density $\kappa(\mathbf{r}) = |\nabla\Psi|^2$. Such a result appears to be consistent with the previous definitions, as already expressed in Eqs. (85) and (96). An alternative definition of the total density (98) can be also obtained, employing the already defined IMP, in association with the

total sum of elements of a vector, as previously given in Eq. (89) and generalized afterwards:

$$\tau(\mathbf{r}) = \langle |\Phi\rangle * |\Phi^*\rangle\rangle = \left\langle \begin{pmatrix} \Psi \\ \nabla\Psi \end{pmatrix} * \begin{pmatrix} \Psi^* \\ (\nabla\Psi)^* \end{pmatrix} \right\rangle$$
$$= \left\langle \begin{pmatrix} |\Psi|^2 \\ |\nabla\Psi|^2 \end{pmatrix} \right\rangle = \left\langle \begin{pmatrix} \rho(\mathbf{r}) \\ \kappa(\mathbf{r}) \end{pmatrix} \right\rangle = \rho(\mathbf{r}) + \kappa(\mathbf{r}) \,. \quad (99)$$

Such superposition of density functions, producing the total density function $\tau(\mathbf{r})$, can be obviously named as a ∇-extended density function.

It must be noted now, that the total density function, as defined in Eqs. (98) or (99), possesses the statistical interpretation of observing, within a space infinitesimal volume element, both the position of the associated system of particles *or* the same system within a corresponding related infinitesimal kinetic energy range. The conjunction *or*, is a consequence of the obtained statistical expressions, where both densities are summed up, and it is in agreement with *Heisenberg's uncertainty principle*, which will forbid the practical use of the product of both distributions, position and momentum not being simultaneously observable.

Quantum Self-Similarity Measures and Non-linear Schrödinger Equation. The previous experience, crystallized in definitions (90) and (91), about the extended wave functions and the details of their subsequent use in the energy definition (95), as well as the construction of the extended density function in Eqs. (98) or (99) formalisms, shows a plausible way to generalize the presented wave function extensions. This effort can be employed as a way to try, afterwards, to obtain information on the possible utility of such extended general wave function forms.

In the same way as the generalization of Sobolev spaces, starting from the simplest form (85), the Ω-extended expression (91) of the wave function can be generalized accordingly.

In order to do so, a systematic exposition will be followed, in the same way as has been previously done. Thus, the first new breed of extended wave functions will be defined by means of the vector, or diagonal, form:

$$|\Phi\rangle = \begin{pmatrix} \Psi \\ |\Psi|^2 \\ \Omega\Psi \end{pmatrix} \equiv \mathrm{Diag}(\Psi; |\Psi|^2; \Omega\Psi) \,. \quad (100)$$

As the second element of the vector (100), is simply the electronic density function, such a vector can be written equivalently as:

$$|\Phi\rangle = \begin{pmatrix} \Psi \\ \rho \\ \Omega\Psi \end{pmatrix} \equiv \mathrm{Diag}(\Psi; \rho; \Omega\Psi) \,; \quad (101)$$

and the corresponding possible energy expression could be written in turn, employing a Hamilton operator and following the previous experience as presented in Eq. (94) in the form:

$$\mathbf{H} = \begin{pmatrix} \mathbf{U} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \alpha\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{1}{2}\mathbf{I} \end{pmatrix} = \mathrm{Diag}(\mathbf{U}; \alpha\mathbf{I}; \tfrac{1}{2}\mathbf{I}) \,, \quad (102)$$

where the first and the last non null elements are the same as shown before in Eq. (94), and besides α is an arbitrary real parameter. Thus, the energy expectation value equivalent to the expression (95), employing in the extended wave function (101) the substitution $\Omega = \nabla$, gives:

$$E = \langle \Phi|\mathbf{H}|\Phi\rangle = \langle U\rangle + \alpha\langle\rho|\rho\rangle + \langle K\rangle \,. \quad (103)$$

So, calling the classical Schrödinger energy (95):

$$E_0 = \langle U\rangle + \langle K\rangle \,,$$

and owing to the fact that the second term in the expectation value (103), can be manipulated as follows:

$$z = \langle\rho|\rho\rangle = \int_D |\Psi|^2|\Psi|^2\,\mathrm{d}\mathbf{r} = \int_D \Psi^*|\Psi|^2\Psi\,\mathrm{d}\mathbf{r}$$
$$= \langle\Psi\|\Psi|^2|\Psi\rangle = \langle\Psi|\rho|\Psi\rangle = \langle\rho\rangle \,, \quad (104)$$

then, one can simply write:

$$E = E_0 + \alpha z \,. \quad (105)$$

The nature of the integral (104) is well known in the theoretical formulation and definitions of *quantum similarity measures* (QSM), corresponding to the so-called *quantum self-similarity measure* (QSSM) associated to the density function ρ. As is evident upon inspecting the QSSM appearing in Eq. (104), the integral also corresponds to a scalar product of the density function by itself. That is: a simple Euclidean norm within the associated Hilbert semispace, containing ρ. In addition, it can be quantum mechanically interpreted as the expectation value of the density function over itself. Finally, the integral is also closely related to a relativistic correction appearing in the definition of the Breit Hamiltonian: the term named *spin-spin contact* [161,162,163], although in the present form (104) the two-electron Dirac δ-function is absent. The characteristic non-trivial features of this kind of QSSM integral, as appear in this particular formalism, and more precisely with respect of the spin part of the wave function, have been deeply analyzed in two separate papers [87,164].

Moreover, Eqs. (101) up to (105), tell that, in fact, the new extended wave function produces an energy expectation value, which can be seen to be in correspondence with the non-linear Schrödinger equation.

Upon inspecting definition (100), or Eq. (101), the corresponding extended density function can be deduced, employing the same technique as the one that was previously used in Eq. (99):

$$\gamma(\mathbf{r}) = \langle |\Phi\rangle * |\Phi^*\rangle\rangle = \left\langle \begin{pmatrix} \Psi \\ \rho \\ \nabla\Psi \end{pmatrix} * \begin{pmatrix} \Psi^* \\ \rho \\ (\nabla\Psi)^* \end{pmatrix} \right\rangle$$
$$= \left\langle \begin{pmatrix} |\Psi|^2 \\ \rho^2 \\ |\nabla\Psi|^2 \end{pmatrix} \right\rangle = \left\langle \begin{pmatrix} \rho(\mathbf{r}) \\ \rho^2(\mathbf{r}) \\ \kappa(\mathbf{r}) \end{pmatrix} \right\rangle = \rho(\mathbf{r}) + \rho^2(\mathbf{r}) + \kappa(\mathbf{r}) \,. \tag{106}$$

So, the total density now can be written as the ∇-extended density function (99), with an extra term made of the squared electronic density:

$$\gamma(\mathbf{r}) = \tau(\mathbf{r}) + \rho^2(\mathbf{r}) \,.$$

Finally, the Sobolev norm of the extended wave function (101) will be easily obtained by integrating the extended density function (106):

$$\langle \Phi|\Phi\rangle = 1 + 2\langle K\rangle + z \,, \tag{107}$$

the QSSM integral z, defined in Eq. (104), can be also considered as a norm, associated by construction, to the elements of a Hilbert semispace, then this fact assures the positive definition of the integral (107). It is not difficult to associate the norm in Eq. (107), with a Sobalev norm of type (87), with the parameters chosen accordingly: $\|\Psi\|_2^{1,1}$.

Thus, a naïve generalization of the idea underlying the ∇-extended wave function definition has revealed itself as a powerful tool, which permits the formal description of the non-linear Schrödinger equation. Such formalism allows producing another kind of Hilbert–Sobolev space, and at the same time, within their integral steps, finally puts into evidence the connection of Sobolev spaces and extended wave functions with the concept of QSSM.

Expectation Values Within Extended Density Functions Framework

Landau and Lifshitz proposed the interpretation of expectation values in a statistical formulation, instead of the usual quantum mechanical form. This was anticipated somewhere in the already quoted volume of reference [153], published within a series dedicated to studying the mechanics of particle systems. The same point of view was also masterly described and adopted, later on, by McWeeny and Sutcliffe in the book of reference [165]. In the present paper, employing the concepts associated to the extended wave functions, it will be shown that a similar possibility as the one mentioned in these previous references can be exactly deduced. The difference with the above-mentioned sources consists of the fact that, in the present way, one only needs to base the arguments on the structure of the already described Hilbert–Sobolev spaces. Some related point of view has been found in the same direction, precluding this property, when the deduction of the energy expectation values has been discussed, as can be noticed when observing Eqs. (95) and (103). According to this, the purpose of this section is to deduce a general Hilbert–Sobolev formalism for the expectation values associated to the extended wave functions and provide an application example.

Statistical Form of Expectation Values in the Extended Wave Function Formalism. One can deduce the general composition of an extended density function, corresponding to the extended function (91). For this purpose, an appropriate operator shall be defined. It has to be able to act over the extended wave function structure. Accordingly, it is sufficient to take into account that a diagonal-like operator can be constructed in the following way:

$$\Theta = \mathrm{Diag}(\Gamma; \Lambda) \,, \tag{108}$$

where, in order to be applied along the appropriate extended wave function elements, the involved operators Γ and Λ themselves have to possess an adequate structure. When, within the global expectation value expression, the Ω-extended wave function is used quantum mechanically over the operator (108), in order to obtain the corresponding equation, then owing to the properties of diagonal matrices, both the wave function and the diagonal operator can be manipulated, in the forthcoming manner, employing obvious notation and symbols, to arrive towards a statistical formulation final form:

$$\begin{aligned}
\langle\Theta\rangle &= \langle\Phi|\Theta|\Phi\rangle \\
&= \langle \mathrm{Tr}(\mathrm{Diag}(\Psi^*; (\Omega\Psi)^*)\,\mathrm{Diag}(\Gamma; \Lambda)\,\mathrm{Diag}(\Psi; \Omega\Psi))\rangle \\
&= \langle \mathrm{Tr}(\mathrm{Diag}(\Gamma; \Lambda)\,\mathrm{Diag}(\Psi^*; (\Omega\Psi)^*)\,\mathrm{Diag}(\Psi; \Omega\Psi))\rangle \\
&= \langle \mathrm{Tr}(\mathrm{Diag}(\Gamma; \Lambda)\,\mathrm{Diag}(|\Psi|^2; |\Omega\Psi|^2))\rangle \\
&= \langle \mathrm{Tr}(\mathrm{Diag}(\Gamma\rho; \Lambda\omega))\rangle = \langle \Gamma\rho + \Lambda\omega\rangle \\
&= \langle \Gamma\rho\rangle + \langle \Lambda\omega\rangle \equiv \langle \Gamma|\rho\rangle + \langle \Lambda|\omega\rangle \,.
\end{aligned} \tag{109}$$

Now it must be taken into account that the external summation symbol employed in the above equation, has to be taken, when appropriate, as an integration procedure.

Energy Expectation Value of a Set of Interacting Quantum Objects. Among the possible uses of the present formalism, it seems worthwhile to consider some theoretical arrangement associated to a previous discussion made by Huzinaga and co-workers [166,167]. This procedure is related to the model potential method, proposed by Bonifacic and Huzinaga [168] in order to study the optimal valence AO, transforming the core electron structure into an electrostatic potential. In the following discussion the structure and final form will be presented, which can take the total energy, when the problem of several interacting quantum objects is studied in a somehow approximate way from the point of view of the wave function. In order to perform such a study, one can suppose known a set of quantum objects [25], whose Hamiltonian operators in this case can be considered constructed with a diagonal structure, similar to the one described in Eq. (102).

However, the second diagonal term has to be transformed necessarily into a new operator and described with the appropriate construction rule, as follows:

$$\mathbf{L} = \mathbf{1} - \mathbf{I} = \{(i \neq j)\} \,. \tag{110}$$

This operator is in some way the reciprocal mirror image of the well-known unit operator:

$$\mathbf{I} = \{\delta(i = j)\} \,. \tag{111}$$

In the last definition (111) as well as in the operator previously defined in Eq. (110), a *logical Kronecker symbol* has been utilized [159,160]. Considering the definition of the unit operator (111), the logical Kronecker symbols appear obviously structured, adopting a self-explanatory description. The definition, for instance, can be made clearer, considering a logical expression Λ taken as the Kronecker symbol argument, and then its resultant value can be generally described by means of the logical content of the possible issues of such an argument, that is:

$$\delta(\Lambda) \in \{\delta(\Lambda \equiv .T.) = 1 \wedge \delta(\Lambda \equiv .F.) = 0\} \,.$$

The operator $\mathbf{1}$, the *unity* operator as used in Eq. (110), means the multiplicative unit of the IMP, that is: $\mathbf{1} = \{\mathbf{1}_{ij} = 1\}$.

Considering α as a parameter to be adjusted, according to the nature of the problem, then the Hamiltonian could be written in this case as:

$$\mathbf{H} = \begin{pmatrix} U\mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \alpha\mathbf{L} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{1}{2}\mathbf{I} \end{pmatrix} = \mathrm{Diag}(U\mathbf{I}; \alpha\mathbf{L}; \tfrac{1}{2}\mathbf{I}) \,, \tag{112}$$

where U is a scalar potential operator, defined by means of the expression:

$$U = \sum_I (V_I + R_I) \,,$$

with the sum encompassing all the involved quantum objects. By the symbol V_I is understood an attractive potential; while by R_I, the repulsion terms can be somehow described. Both operator terms have to be associated in turn to the whole set of particles, constituting the Ith quantum object. For instance: when dealing with atoms and molecules, these operators could be associated to the nuclear attraction operator and to the Coulomb-exchange operator terms, respectively.

Then, in this problem the appropriate extended wave function can be taken as the ∇-extended form of Eq. (101), that is:

$$|\Phi\rangle = \begin{pmatrix} \Psi \\ \rho \\ \nabla\Psi \end{pmatrix} \equiv \mathrm{Diag}(\Psi; \rho; \nabla\Psi) \,, \tag{113}$$

where the wave and the density functions present in the extended function (113) shall be taken as vectors, having as elements the wave and density functions of every quantum object in the considered set, respectively.

Thus, the expectation value of the Hamiltonian (112) under the extended wave function (113), can be easily written, using the technique of Eq. (109) as:

$$\langle\Phi|\mathbf{H}|\Phi\rangle = \sum_I \Big(\sum_J \langle V_J + R_J|\rho_I\rangle + \alpha \sum_{J\neq I} \langle\rho_J|\rho_I\rangle + \langle\kappa_I\rangle \Big), \tag{114}$$

in the above expression the first term corresponds to the potential energy of the objects plus their interactions; the second term can be associated to the expectation value of the projection operator over each quantum object except itself, and the role of this part of the expectation value is intended to prevent the collapsing tendency of the particles, belonging to each separated quantum object, towards a unique system; finally, the third term corresponds to the global kinetic energy obtained in a way such as the quantum objects were non-interacting.

The second element of the expectation value (114) can be also easily interpreted as a sum of the *overlap* QSM [169] between pairs of quantum objects. In this sense, one can observe Huzinaga's treatment as a procedure, taking into account non-linear terms in the approximate solution of the Schrödinger equation.

Quantum Similarity Measures in Extended Hilbert–Sobolev Spaces

QSM in Hilbert semispaces have been studied from the theoretical point of view as well as considering the potential applications of *quantum similarity* over quantum objects. In this paragraph, the structure of QSM over extended wave and density functions will be analyzed. Before proceeding towards such an analysis, it must be said that, as QSM are essentially defined over density functions, they can be constructed even in the Hilbert–Sobolev spaces framework, provided that the extended density function is known. This is so, because, whenever a total density function can be well defined, like the one present in Eq. (98), for instance, then the construction of any similarity measure can also be put forward. Such a general possibility was analyzed in a particular way several years ago [170], when discussing the extension of the QS concepts into partition functions. Statistical mechanics partition functions can be obviously observed as probability distributions and, thus, they can be considered as elements belonging to a characteristic vector set: a Boltzmann semispace, for example. From such a fact they can be used in the general definitions of QSM, as any probability density function can be used for the same purpose.

The most usual way to produce a QSM, corresponds to the integral constructed as:

$$z_{IJ}(\Omega) = \iint_D \rho_I(\mathbf{r}_1)\Omega(\mathbf{r}_1;\mathbf{r}_2)\rho_J(\mathbf{r}_2)\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2\,. \tag{115}$$

Where in Eq. (115), $\{\rho_I(\mathbf{r}_1);\rho_J(\mathbf{r}_2)\}$ is a pair of homogeneous order density functions, and $\Omega(\mathbf{r}_1;\mathbf{r}_2)$ is a positive definite operator. The attached properties of the set of integrands ensure that in any case the values of the QSM, defined such as in Eq. (115), will produce a positive real element. For the present purposes, the integral form (115) is sufficient. The already mentioned overlap QSM, which appears in the building up of energy expectation values within non-linear Schrödinger equations, such as those in expressions (104) and (114), are *overlap-like* QSM, and can be deduced from the QSM Eq. (115), by simply using a Dirac delta function as operator, that is: $\Omega(\mathbf{r}_1;\mathbf{r}_2) = \delta(\mathbf{r}_1 - \mathbf{r}_2)$.

The definition of kinetic energy density and other possible density kinds, deducible from the Ω-extended wave function concepts, as discussed earlier, opens the way to produce QSM using the integral (115), upon substitution of the density function pairs by the appropriate extended density function.

So it seems now clear that the QSM integral form, as described in Eq. (115), can be used as it is for extended density functions, just substituting the usual electronic density by the corresponding expression in terms of the chosen extended density functions. Here, the interesting new feature consists of the emergence of QSM integrals, associated to density functions of different origin. For instance, suppose the ∇-extended density function as defined in Eq. (98): the total density defined there, associated to a pair of quantum objects produces a QSM, which can be written in terms of four hybrid QSM integrals as:

$$\begin{aligned} z_{IJ} &= \iint_D \tau_I(\mathbf{r}_1)\Omega(\mathbf{r}_1;\mathbf{r}_2)\tau_J(\mathbf{r}_2)\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2 \\ &= \iint_D \rho_I(\mathbf{r}_1)\Omega(\mathbf{r}_1;\mathbf{r}_2)\rho_J(\mathbf{r}_2)\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2 \\ &\quad + \iint_D \rho_I(\mathbf{r}_1)\Omega(\mathbf{r}_1;\mathbf{r}_2)\kappa_J(\mathbf{r}_2)\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2 \\ &\quad + \iint_D \kappa_I(\mathbf{r}_1)\Omega(\mathbf{r}_1;\mathbf{r}_2)\rho_J(\mathbf{r}_2)\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2 \\ &\quad + \iint_D \kappa_I(\mathbf{r}_1)\Omega(\mathbf{r}_1;\mathbf{r}_2)\kappa_J(\mathbf{r}_2)\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2\,. \end{aligned}$$

The first term being exactly the one in Eq. (115), and the last one being a QSM over kinetic energy density distributions, the central terms corresponding to hybrid QSM between electronic and kinetic density functions. Total density QSM integrals of any kind still are waiting to be practically employed in a systematic way. It is important to consider them as potentially interesting quantum object descriptors. Because of their flexible generality, the extended density functions may provide new insights and refinements within QSPR models. In the next sections of this work a discussion of several possible uses of the extended densities technique will be discussed.

Fundamental Quantum QSPR (QQSPR) Equation in Sobolev Spaces

From the previous considerations and the *ad hoc* definitions, it can be said, without doubt, that the general structure of Sobolev spaces can be easily associated to the extended Hilbert wave functions, producing a new mathematical structure, which has been named a Hilbert–Sobolev space.

Hilbert–Sobolev spaces have been used too to obtain up to now, with the appropriate definition of the Hamilton operator, adapted to every circumstance, the corresponding energy expectation values. Elementary reasoning permits us to arrive towards the conclusion that the expectation values of a given observable can be obtained in a similar generalized manner and within the formalism, which is

the most valuable finding, perhaps, of the present discussion: the possibility to write the expectation value expressions using a statistical method.

Thus, if some observable **O**, has attached the Hermitian operator Θ, and the associated quantum object extended density function is τ, then the associated quantum mechanical expectation value, $\langle\vartheta\rangle$, can be written, according to the previous considerations as:

$$\langle\vartheta\rangle = \langle\Theta|\tau\rangle \,.$$

In QQSPR reasoning, the expression above can be further arranged in the following way. The Hermitian operator Θ is usually not known, but it can be expressed as the product of a still unknown operator W by a known one Ω, associated at the same time with a positive definite property, that is:

$$\langle\vartheta\rangle = \langle W\Omega|\rho\rangle \,. \tag{116}$$

Because Ω is chosen as having a positive definite form, one is always assured that knowing Θ, then W can be obtained in turn as:

$$\Omega > 0 \rightarrow \exists\Omega^{-1} \rightarrow W = \Theta\Omega^{-1} \,.$$

However, in practice Θ and W are unknown, and in the expectation value formalism there is always the need to obtain W in an approximate way, employing a least squares procedure. This can be easily done whenever a set of compatible density functions, connected to a certain quantum object set as quantum object tags, is already known: $T = \{\tau_I\}$ and can be used in order to express the operator W as a superposition such that:

$$W(\mathbf{r}) = \sum_I w_I \tau_I(\mathbf{r}) \,. \tag{117}$$

Then, the expectation value (116) becomes expressible as:

$$\langle\vartheta\rangle = \sum_I w_I \langle\tau_I|\Omega|\tau\rangle \,,$$

being the resultant QSM integrals, defined in the usual way in Eq. (115). The well-described procedure to obtain the coefficients: $\mathbf{w} = \{w_I\}$ is the least-squares technique, or anyone of the existing variants, as previously described, by means of an alternative method based on IMP reasoning.

The first step is, in any case, to proceed with the construction of a linear system of equations, whose solution is the coefficient vector **w**. To obtain such a linear system it is necessary first to know, for a given quantum object set, a set of properties: $\{\pi_I\}$, which can be associated to the corresponding expectation values: $\{\vartheta_I\}$. Then, employing the quantum objects density functions, which in particular will coincide with the set $T = \{\tau_I\}$, used to construct the unknown part of the operator, although not necessarily both density function sets shall be the same, it is possible to write:

$$\forall K\colon \langle\vartheta_K\rangle \equiv \pi_K = \sum_I w_I \langle\tau_I|\Omega|\tau_K\rangle \,,$$

so collecting in a vector like $|\pi\rangle$, the property values, and defining the integral matrix elements by means of:

$$\mathbf{Z} = \{z_{IK} = \langle\tau_I|\Omega|\tau_K\rangle = \langle\tau_K|\Omega|\tau_I\rangle = z_{KI}\} \,, \tag{118}$$

it is easy to transform the system into a matrix equation in the form of:

$$\mathbf{Zw} = |\pi\rangle \,. \tag{119}$$

The solution of the above system will make known an approximate form of the implied operator and, in this way, provides a possible path to be followed in order to obtain estimates of the property values of any unknown quantum object, just as in the classical QSPR model procedures.

However, the present kind of quantitative structure (represented by the QSM)-property model is completely based on quantum mechanical propositions. More than this, there are no other suppositions than the usual ones, associated to density function algebra and quantum mechanical basic mathematical background. Hence, the results obtained through the linear system of Eqs. (119), do not depend on user choice, but rely directly on theoretical grounds and because of this are statistically unbiased. Equation (119) can be properly called the fundamental QQSPR equation. Moreover, the models obtained in this way can be interpreted in the light of the quantum mechanical expectation value concept. By this simple fact, contrary to the classical QSPR modeling results, they can be associated to a causal relationship relying on quantum object properties and QSM.

Due to the unavoidable presence of the QSM matrix (118) into the fundamental QQSPR equation, the columns of such a matrix play a fundamental role in the discrete representation of quantum objects. Consequently, the columns of the QSM matrix, involving a given quantum object density function, interacting with the whole basis set of density functions employed to represent the unknown operator W, can be safely considered as natural quantum mechanical discrete descriptors of the associated quantum object.

Non-Linear Terms in QQSPR Models

The usual relationships between structure and properties sometimes needs the presence of non-linear terms. Non-linear terms are needed in order to represent accurately the property as a function of the structural descriptors.

The fundamental QQSPR equation can be deduced to introduce in a natural way these terms, if needed. To see this, it is only necessary to think about the easy path, which was used to introduce the non-linear terms in the Schrödinger equation, just by using simple considerations, associated to the structure of the Hilbert–Sobolev spaces.

Suppose that the extended wave function (106) is employed, along with the corresponding extended density functions. In this case, although the Eq. (117), producing the unknown operator expression, can be supposedly set in the same manner as has been proposed in the usual framework, the kinetic energy distribution as well as the non-linear density terms can be employed separately. That is, the unknown operator can be written now as:

$$W = \sum_I w_I \rho_I + \sum_J k_J \kappa_J + \sum_L l_L |\rho_L|^2 \,. \tag{120}$$

Such an approach will produce a set of linear equations, with an extended number of parameters, but also a matrix representation of the unknown operator with added dimensions. The matrix elements, involving squared density functions, are candidates to be interpreted as the quantum representatives of the possible presence of non-linear terms in the fundamental QQSPR equation.

Non-linearity can be introduced in several alternative ways, due to the flexibility promoted by the ideas around the Hilbert–Sobolev concepts. For example, within the QSM matrix definition (118), the operator Ω can be seen as formed by the expression:

$$\begin{aligned}\Omega = \exp(a\rho) &= \sum_{p=0}^{\infty} \frac{a^p}{p!} |\rho|^p \\ &= I + a\rho + \frac{a^2}{2}|\rho|^2 + O(3) \,,\end{aligned} \tag{121}$$

which is assured to be positive definite whenever the exponent a possesses positive values. Convergence can be assured whenever: $a \in (0, 1)$.

In fact, such an expansion can be generalized by using a set of positive definite coefficients $A = \{a_I\}$, such that:

$$\Omega = \sum_{p=0}^{\infty} a_p |\rho|^p = a_0 I + a_1 \rho + a_2 |\rho|^2 + O(3) \,.$$

In such a general case, the QSM matrix elements will be written as a superposition of terms like:

$$\begin{aligned} z_{IK} = \langle \tau_I | \Omega | \tau_K \rangle &= \left\langle \tau_I \middle| \sum_{p=0}^{\infty} a_p \middle| \rho|^p | \tau_K \right\rangle \\ &= \sum_{p=0}^{\infty} a_p \langle \tau_I \| \rho|^p | \tau_K \rangle \\ &= a_0 \langle \tau_I | \tau_K \rangle + a_1 \langle \tau_I | \rho | \tau_K \rangle \\ &\quad + a_2 \langle \tau_I \| \rho|^2 | \tau_K \rangle + O(3) \,. \end{aligned} \tag{122}$$

Where, in the last line of Eq. (122), it is easy to observe the overlap similarity integrals as the zero-th order term, the triple density similarity integrals as the second element constituting the first-order term, and finally, in the second-order term, the quadruple density integrals appear. Such integrals can be readily defined by means of the expression, chosen among other possible definitions, for example, as:

$$\langle \tau_I \| \rho|^2 | \tau_K \rangle = \int_D \tau_I(\mathbf{r}) \tau_K(\mathbf{r}) \rho^2(\mathbf{r}) \, \mathrm{d}\mathbf{r} \,.$$

This result is still more obvious if the following operator structure is employed: upon substituting in Eq. (121) the density function by a convex superposition like the one in Eq. (120), which to obtain simpler expressions will be written as a convex superposition like:

$$\rho = \sum_A \omega_A \rho_A \,,$$

being the coefficients $\{\omega_A\}$ such that: $\forall A\colon \omega_A \in \mathbf{R}^+ \wedge \sum_A \omega_A = 1$.

In this case, the QSM integral (122) will take the following form:

$$\begin{aligned} z_{IK} = \langle \tau_I | \Omega | \tau_K \rangle &= \left\langle \tau_I \middle| \sum_{p=0}^{\infty} \frac{a^p}{p!} \middle| \rho|^p | \tau_K \right\rangle \\ &= \sum_{p=0}^{\infty} \frac{a^p}{p!} \langle \tau_I \| \rho|^p | \tau_K \rangle \\ &= \langle \tau_I | \tau_K \rangle + a \sum_A \omega_A \langle \tau_I | \rho_A | \tau_K \rangle \\ &\quad + \frac{a^2}{2} \sum_A \sum_B \omega_A \omega_B \langle \tau_I | \rho_A \rho_B | \tau_K \rangle \\ &\quad + O(3) \,. \end{aligned}$$

It can be seen that quadratic or higher order terms can naturally appear in the structure of the fundamental QQSPR equation in this way.

Non-Linear Terms and Variational Approach in Quantum QSPR

Fundamental QQSPR Equation in $(N \times M)$ *Similarity Matrix Spaces.* Suppose a quantum object basis set B composed by M quantum systems, whose homogeneous density functions, acting as quantum object tags, are known: $B = \{\rho_I^B | I = 1, M\}$. Suppose also that a probe quantum object set P is well defined and composed by N quantum systems, which have also known density tags: $\{\rho_I^P\}$, and at least is also known a set of property values: $\{p_J\}$ attached to every quantum object of the set; in this manner: $P = \{\rho_J^P \wedge p_J | J = 1, N\}$.

A general operator Ω can be associated to the expectation value computation of the observable property π, in such a way that, knowing the appropriate quantum state density function tag ρ for a given quantum system, such a quantum object observable property can be evaluated in general by using the integral form:

$$\langle \pi \rangle = \langle \Omega | \rho \rangle = \int_D \Omega(\mathbf{r}) \rho(\mathbf{r}) \, \mathrm{d}\mathbf{r} \,, \tag{123}$$

where D is an appropriate integration domain, where the density and operator variables are defined.

Being the operator Ω, in principle, after the adoption of quantum mechanical rules, a Hermitian operator, without loss of generality can be supposedly decomposed into a product of two commutative operators:

$$\Omega(\mathbf{r}) = W(\mathbf{r}) \Theta(\mathbf{r}) \wedge [W(\mathbf{r}); \Theta(\mathbf{r})] = 0 \,, \tag{124}$$

the operator Θ being a known chosen positive definite one, the remnant Hermitian operator is thus defined as:

$$W(\mathbf{r}) = \Omega(\mathbf{r}) \Theta^{-1}(\mathbf{r}) \,. \tag{125}$$

Using Eq. (123) and the operator composition shown in Eq. (124), then it can be formally written:

$$\langle \pi \rangle = \langle W \Theta | \rho \rangle \equiv \langle W | \Theta \rho \rangle = \langle W | \Theta | \rho \rangle \,, \tag{126}$$

suggesting that the operator W could be approximately obtained, even in the case that it is unknown, due to the nature of the observable attached to the property.

In the case, most usual in QQSPR framework, that an approximate construction of the operator W is needed, if an appropriate quantum object set density function tag set, acting as a basis set, B say, is known, as stated at the beginning, that is: $B = \{\rho_I^B | I = 1, M\}$, then the operator W can be written within a first-order linear approach as:

$$W \simeq \sum_{I=1}^{M} \omega_I \rho_I^B \,, \tag{127}$$

so upon substituting this approximate first-order linear expression into the expectation value in Eq. (126), is obtained:

$$\langle \pi \rangle \simeq \sum_{I=1}^{M} \omega_I \langle \rho_I^B | \Theta | \rho \rangle \,, \tag{128}$$

where the integral in Eq. (128), can be interpreted as a quantum similarity measure, that is:

$$\langle \rho_I^B | \Theta | \rho \rangle \equiv \iint_D \rho_I^B(\mathbf{r}_1) \Theta(\mathbf{r}_1; \mathbf{r}_2) \rho(\mathbf{r}_2) \, \mathrm{d}\mathbf{r}_1 \, \mathrm{d}\mathbf{r}_2 \,. \tag{129}$$

The unknown coefficient set in Eq. (128): $|\omega\rangle = \{\omega_I | I = 1, M\}$, which can be collected into an M-dimensional column (or row) vector, will represent the operator W in terms of the known density function basis set B. This situation, clearly represented by Eq. (128), still has a set of undetermined parameters, associated now to the vector $|\omega\rangle$ components, instead of the operator W.

Equation (128) can be used to obtain the vector $|\omega\rangle$. As is usually the case in classical QSPR, it is only necessary to know, a quantum object tag set, associated to some molecular probe set P of cardinality N, $P = \{\rho_J^P \wedge p_J | J = 1, N\}$, where, as previously commented, every quantum object structure in P has also necessarily to be attached to a known value of the involved observable: $|p\rangle = \{p_J | J = 1, N\}$, which can be also collected in form of a N-dimensional column (or row) vector. Then, Eq. (128) can be rewritten for every element in P, employing the known property values instead of the expectation observable values, that is:

$$\forall J = 1, N \colon p_J \simeq \sum_{I=1}^{M} \omega_I \langle \rho_I^B | \Theta | \rho_J^P \rangle \,, \tag{130}$$

in this way the following set of quantum similarity measures is generated:

$$a_{IJ}^{BP}(\Theta) \equiv a_{IJ}^{BP} = \langle \rho_I^B | \Theta | \rho_J^P \rangle \,, \tag{131}$$

which in turn can be considered, after an appropriate rearrangement, as elements of a $(M \times N)$ similarity matrix, involving the basis and probe quantum object molecular sets respectively: $\mathbf{A} = \{a_{IJ}^{BP}\}$.

With this matrix definition in mind, then Eq. (130) can be rewritten as a linear system in matrix form, connecting the already defined vectors in row space form:

$$\langle p | = \langle \omega | \mathbf{A} \,. \tag{132}$$

Such a linear system can be associated to the most common dual problem in column vector space, just defining

the transpose of the similarity matrix, using the usual definition:

$$\mathbf{Z} = \mathbf{A}^{\mathbf{T}} \rightarrow \forall I = 1, N \wedge J = 1, M\colon z_{JI}^{PB} = a_{IJ}^{BP} \ , \quad (133)$$

and in this manner, the fundamental QQSPR equation is set up, writing a column equivalent dual expression of the former row Eq. (132):

$$\mathbf{Z}|\omega\rangle = |p\rangle \ . \quad (134)$$

As in classical QSPR, the solutions of Eq. (134) may provide the knowledge of the coefficient vector $|\omega\rangle$. However, it must again be stressed that Eq. (134) differs from the classical QSPR setup in the sense that such an equation can be deduced from the quantum mechanical statistical structure, associated to expectation value calculations. In this way, the *causal* connection between molecular structure and molecular properties can be deduced from employing quantum mechanical theoretical fundaments, via the ideas of quantum similarity. The interest of such a relationship lies in the fact that fundamental QQSPR equations can be extended to any quantum object structure and properties. So, obviously, these relationships can be applied to molecular systems as well, provided they can be described as quantum objects, making QQSPR *universal* in the sense that it can be applied, under the same conditions, to any sub-microscopic quantum object set.

Non-Linear QQSPR Equations. In a second remark step, which appears to be sufficiently important as to merit a separate section treatment, the approximate operator linear description (127) may be extended with non-linear terms, which can be easily provided by the nature of the involved quantum object density function tags, which can be founded in turn on the theoretical development of extended wave functions.

In this case, Eq. (127), can be written in a more structured manner as a truncated Taylor series, where only the first two terms are kept for simplicity:

$$W \simeq \sum_{I=1}^{M} \omega_I \rho_I^B + \sum_{P=1}^{M} \sum_{Q \geq P}^{M} \omega_{PQ} \rho_P^B \rho_Q^B + O(3) \ , \quad (135)$$

however, with the potential prospect to add terms up to any order. Equation (135) can be perhaps also considered a simplification of a series involving density functions of growing orders, that is:

$$W \simeq \sum_{I=1}^{M} \omega_I^{(1)} \rho_I^{(1)B} + \sum_{P=1}^{M} \omega_P^{(2)} \rho_P^{(2)B} + O(3) \ . \quad (136)$$

The second-order coefficient set $\{\omega_{PQ}\}$ in Eq. (135), can be also substituted as well, in order to retain a minimal number of unknowns, by products of first-order coefficients, in the following way:

$$\forall P, Q\colon \omega_{PQ} \simeq \omega_P \omega_Q \ . \quad (137)$$

Then, just if this is the case, Eq. (130), transforms into a more computationally convenient form:

$$\forall J = 1, N\colon p_J \simeq \sum_{I=1}^{M} \omega_I \langle \rho_I^B | \Theta | \rho_J^P \rangle + \sum_{P=1}^{M} \sum_{Q \geq P}^{M} \omega_P \omega_Q \langle \rho_P^B \rho_Q^B | \Theta | \rho_J^P \rangle + O(3) \ , \quad (138)$$

Triple Density Quantum Similarity Integrals. The integrals included in the second-order terms of Eq. (138) are *triple density similarity measures*, which can have the form chosen, among many other possibilities, in the following way:

$$\langle \rho_P^B \rho_Q^B | \Theta | \rho_J^P \rangle \equiv \iiint_D \rho_P^B(\mathbf{r}_1) \rho_Q^B(\mathbf{r}_2) \Theta(\mathbf{r}_1; \mathbf{r}_2; \mathbf{r}_3) \rho_J^P(\mathbf{r}_3) \, \mathrm{d}\mathbf{r}_1 \, \mathrm{d}\mathbf{r}_2 \, \mathrm{d}\mathbf{r}_3 \ . \quad (139)$$

Moreover, the usual computational form of the triple density measures can be the one, where the operator becomes unit and all the integrand density functions bear the same variable, so the integral in Eq. (139) acquires a simpler structure, like the triple density overlap integral form:

$$\langle \rho_P^B \rho_Q^B \rho_J^P \rangle \equiv \int_D \rho_P^B(\mathbf{r}) \rho_Q^B(\mathbf{r}) \rho_J^P(\mathbf{r}) \, \mathrm{d}\mathbf{r} \ ; \quad (140)$$

while, first-order similarity measures (129) become, under an equivalent simplification, *overlap-like* integrals:

$$\langle \rho_I^B \rho_J^P \rangle \equiv \int_D \rho_I^B(\mathbf{r}) \rho_J^P(\mathbf{r}) \, \mathrm{d}\mathbf{r} \ . \quad (141)$$

Equations (140) and (141), could be obtained defining the respective weighting operators in terms of an integral operator, involving as many products of Dirac's delta functions as density functions appear into the integrand. For instance, in Eq. (139), the operator $\Theta(\mathbf{r}_1; \mathbf{r}_2; \mathbf{r}_3)$ can be sub-

stituted inside the integral in the following manner:

$$\begin{aligned}\langle\rho_P^B\rho_Q^B\rho_J^P\rangle &\equiv \int_D\Bigg[\iiint_D \rho_P^B(\mathbf{r}_1)\rho_Q^B(\mathbf{r}_2)\\ &\quad\cdot(\delta(\mathbf{r}_1-\mathbf{r})\delta(\mathbf{r}_2-\mathbf{r})\delta(\mathbf{r}_3-\mathbf{r}))\\ &\quad\cdot\rho_J^P(\mathbf{r}_3)\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2\,\mathrm{d}\mathbf{r}_3\Bigg]\mathrm{d}\mathbf{r}\\ &= \int_D\Bigg[\int_D \rho_P^B(\mathbf{r}_1)\delta(\mathbf{r}_1-\mathbf{r})\,\mathrm{d}\mathbf{r}_1\\ &\quad\cdot\int_D \rho_Q^B(\mathbf{r}_2)\delta(\mathbf{r}_2\mathbf{r})\,\mathrm{d}\mathbf{r}_2\\ &\quad\cdot\int_D \rho_J^P(\mathbf{r}_3)\delta(\mathbf{r}_3-\mathbf{r})\,\mathrm{d}\mathbf{r}_3\Bigg]\mathrm{d}\mathbf{r}\\ &= \int_D \rho_P^B(\mathbf{r})\rho_Q^B(\mathbf{r})\rho_J^P(\mathbf{r})\,\mathrm{d}\mathbf{r}\end{aligned} \tag{142}$$

It is, then, straightforward to use the same technique to obtain equations possessing a higher number of density function terms, and so it is easily seen how to take into account and to handle them in the same manner, adding higher order terms within non-linear fundamental QQSPR equations of type (138).

Hansch-Type QQSPR Quadratic Models. In the same manner as above in the linear case, the fundamental quadratic QQSPR Eq. (138) can be simplified, so only the diagonal terms of the initial equation remain. First using just a probe set, taking $B = P$ and then supposing that the remnant equation summation terms are constant under the study of some quantum objects, possessing a great deal of homogeneity. In this case one can write:

$$\begin{aligned}&\forall J = 1, N:\\ &\quad p_J \simeq \beta + \alpha\langle\rho_J^P|\Theta|\rho_J^P\rangle + \alpha^2\langle\rho_J^P\rho_J^P|\Theta|\rho_J^P\rangle + O(3)\,,\end{aligned} \tag{143}$$

which constitutes a quadratic extension of the linear Hansch-type relationships.

Quadratic Fundamental QQSPR Equation in Matrix Form. Having set up in the way outlined above the formal structure of the fundamental QQSPR equations, we now need to discuss its matrix implementation, which is an obligatory step when seeking computational algorithms in practical cases. Two possible equivalent modes will be discussed in this section: the first one corresponds to classical matrix product formalism, while a second part will present an equivalent form just employing inward matrix products. The reason for this second formal presentation is the easiness of setting a general framework up to any approximation order.

a) Classical Form

Equation (138) can be easily written in matrix form. For this purpose it is only necessary to define, besides the column vector of the first-order coefficients:

$$|\omega\rangle = \{\omega_I|I = 1, M\}\,; \tag{144}$$

also, for every quantum object within the probe set, the first-order M-dimensional similarity matrix columns:

$$J = 1, N\colon\ \left|\mathbf{z}_{IJ}^{(1)}\right\rangle = \left\{z_{IJ}^{(1)} = \langle\rho_I^B|\Theta|\rho_J^P\rangle|I = 1, M\right\}\,, \tag{145}$$

as well as the second-order $(M \times M)$-dimensional similarity matrices:

$$\begin{aligned}&J = 1, N:\\ &\quad\left|\mathbf{Z}_J^{(2)}\right\rangle = \left\{z_{J;PQ}^{(2)} = \langle\rho_J^P|\Theta|\rho_P^B\rho_Q^B\rangle|P, Q = 1, M\right\}\,,\end{aligned} \tag{146}$$

shall be constructed.

Taking the above-defined similarity matrices into account, Eq. (138) can be written as:

$$J = 1, N\colon\ p_J \simeq \langle\mathbf{z}_J^{(1)}|\omega\rangle + \langle\omega|\mathbf{Z}_J^{(2)}|\omega\rangle + O(3)\,, \tag{147}$$

so, collecting the property observable values into a column vector, as already discussed and then, reordering first- and second-order matrix components in the following way:

$$\mathbf{Z}^{(1)} = \left\{\left|\mathbf{z}_J^{(1)}\right\rangle|J = 1, N\right\}\,, \tag{148}$$

and

$$\mathbf{Z}^{(2)} = \left\{\mathbf{Z}_J^{(2)}|J = 1, N\right\}\,, \tag{149}$$

then the second-order fundamental QQSPR equation becomes a quadratic system of equations in matrix form:

$$|p\rangle \simeq (\mathbf{Z}^{(1)} + [\langle\omega|\mathbf{Z}^{(2)}])|\omega\rangle + O(3)\,. \tag{150}$$

b) Inward Matrix Product Form as a Generalization Device

Alternatively, there is the possibility to express the equations of the previous description by means of *inward matrix products*. The first-order term in Eq. (147) can be expressed within inward product formalism at once, as it is a simple scalar product between the involved vectors, so:

$$J = 1, N\colon\ \langle\mathbf{z}_J^{(1)}|\omega\rangle \equiv \left\langle\left|\mathbf{z}_J^{(1)}\right\rangle * |\omega\rangle\right\rangle\,, \tag{151}$$

while the second-order term may be expressed in inward product form with the aid of the coefficient vector tensor product, forming a square $(N \times N)$ matrix:

$$\mathbf{W} = |\omega\rangle \otimes |\omega\rangle \equiv \{w_{IJ} = \omega_I \omega_J | \forall I, J = 1, N\}, \quad (152)$$

so, one can write then the quadratic term of Eq. (150) as an inward matrix product too:

$$J = 1, N: \\ \langle \omega | \mathbf{Z}_J^{(2)} | \omega \rangle \equiv \langle \mathbf{Z}_J^{(2)} * \mathbf{W} \rangle = \langle \mathbf{Z}_J^{(2)} * (|\omega\rangle \otimes |\omega\rangle) \rangle , \quad (153)$$

and consequently Eq. (147), can be rewritten as:

$$J = 1, N: \\ p_J \simeq \langle |\mathbf{z}_J^{(1)}\rangle * |\omega\rangle \rangle + \langle \mathbf{Z}_J^{(2)} * (|\omega\rangle \otimes |\omega\rangle) \rangle + O(3) . \quad (154)$$

Inward Matrix Product Formalism of Fundamental QQSPR Equation nth Order Terms. Both, classical and inward product, formalisms are equivalent; however, the inward product Eq. (154), permits one to easily imagine any sequence of corrections into the fundamental QQSPR equation, up to any arbitrarily chosen nth order term, just writing:

$$J = 1, N: p_J \simeq \sum_{R=1}^{n} \left\langle \mathbf{Z}_J^{(R)} * \left(\bigotimes_{S=1}^{R} |\omega\rangle \right) \right\rangle + O(n+1), \quad (155)$$

where the leading equation terms are $\{\mathbf{Z}_J^{(R)} | J = 1, N\}$ the Rth order similarity matrices, which can be constructed as:

$$J = 1, N: \mathbf{Z}_J^{(R)} = \{\mathbf{z}_{J;S(\mathbf{i})}^{(R)} = \langle \rho_J^P | \Theta | \rho_{S_1}^B \rho_{S_2}^B \ldots \rho_{S_R}^B \rangle | \forall \alpha \\ = 1, R: L_\alpha \in \{1, 2, \ldots, M\}\} , \quad (156)$$

with the index set: $S(\mathbf{i}) = \{S_1, S_2, \ldots, S_R\}$ formed by any of the M^R combinations with repetition of R elements chosen within the M integers and, finally, the Rth order tensor products of the coefficient vector are noted as: $\otimes_{S=1}^{R} |\omega\rangle$.

Stochastic Transformations. We will now discuss a third remark step, dealing with the stochastic transformation of similarity matrices, because it also merits a separate section. Recently, several studies have dealt with stochastic transformations of the fundamental QQSPR equation in linear symmetric form, that is: using $B = P$.

At the light of the previous manipulation presented in this study, the stochastic structure transformation of the fundamental QQSPR equation has to be performed, at any operator-equation approximation level, using the possibility to compute the sum of the elements of the Rth order similarity matrices as have been previously defined in Eq. (156), that is:

$$\sigma_J^{(R)} = \langle \mathbf{Z}_J^{(R)} \rangle = \sum (\mathbf{i}) z_{J;S(\mathbf{i})}^{(R)} , \quad (157)$$

where a nested summation symbol $\sum(\mathbf{i})$ has been employed in order to indicate the nested sums over the R indices, represented by the index sets: $S(\mathbf{i}) = \{S_1, S_2, \ldots, S_R\}$. Using the sum of the similarity matrix elements (157), then the elements of the new matrices scaled by this sum become scaled in turn as follows:

$$\mathbf{S}_J^{(R)} = (\sigma_J^{(R)})^{-1} \mathbf{Z}_J^{(R)} , \quad (158)$$

and the new Rth order stochastic similarity matrices behave as a discrete probability distribution, as: $\forall S(\mathbf{i})$: $z_{J;S(\mathbf{i})}^{(R)} \in \mathbf{R}^+ \rightarrow s_{J;S(\mathbf{i})}^{(R)} \in \mathbf{R}^+$ and besides:

$$\langle \mathbf{S}_J^{(R)} \rangle = 1 . \quad (159)$$

Both properties can be cast into a unique *convex condition* symbol:

$$K(\mathbf{S}_J^{(R)}) = \{\forall S(\mathbf{i}): s_{J;S(\mathbf{i})}^{(R)} \in \mathbf{R}^+ \wedge \langle \mathbf{S}_J^{(R)} \rangle = 1\} . \quad (160)$$

So, in this way, the stochastic matrix set: $\mathbf{S} = \{\mathbf{S}_J^{(R)} | R = 1, n\}$ can be considered, up to nth order, as a set of M^R-dimensional *unit shell* elements, belonging to some *vector semispace* with the same dimensions. In these circumstances one can consider the fundamental QQSPR Eq. (155) as to be written:

$$J = 1, N: p_J = \sum_{R=1}^{n} \left\langle \mathbf{S}_J^{(R)} * \left(\bigotimes_{S=1}^{R} |\omega\rangle \right) \right\rangle + O(n+1), \quad (161)$$

where everything is the same as in the former Eq. (155), except for the similarity matrix set, which has been substituted by the stochastic matrices (158).

The coefficient vector has been left unchanged, but evidently its character could be no longer the same as in Eq. (155). However, the nature of the coefficient vector can be more precise in this case of the fundamental QQSPR stochastic Eqs. (161). This is due to the characteristic convex condition properties, which possess the semispace unit shell elements obtained transforming the similarity matrices.

In fact, the stochastic similarity matrix set: $\{\mathbf{S}_J^{(R)}|R = 1, n\}$, so naturally obtained from the original similarity matrix set, can be interpreted as a sequential discrete representation of the continuous normalized density function, associated to the involved Jth quantum object. Then, from the quantum mechanical point of view, the whole stochastic matrix set can be viewed as a discrete quantum object tag collection. Thus, in this case, the tensor products of the coefficient vector can be easily considered as arrays of convex sets, that is:

$$\mathbf{W}^{(R)} = \bigotimes_{S=1}^{R} |\omega\rangle = \{w_{S(\mathbf{i})}^{(R)}\} \rightarrow \langle \mathbf{W}^{(R)}\rangle = \sum (\mathbf{i}) w_{S(\mathbf{i})}^{(R)} = 1 \wedge \forall S(\mathbf{i}) \colon w_{S(\mathbf{i})}^{(R)} \in \mathbf{R}^+ \tag{162}$$

because, whenever the generating coefficient vector is a convex vector, that is, fulfilling the convex conditions:

$$K(|\omega\rangle) = \left\{ \forall I \colon \omega_I \in \mathbf{R}^+ \wedge \langle |\omega\rangle\rangle = \sum_I \omega_I = 1 \right\}, \tag{163}$$

then, any tensor product of the convex vector $|\omega\rangle$ fulfils: $K\big(\otimes_{S=1}^{R} |\omega\rangle\big)$. Indeed, if convex conditions (163) hold, then it is easy to see that convex conditions are present within any arbitrary order tensor product of the coefficient vector, as shown in the following deduction:

$$\begin{aligned} \mathbf{W}^{(R)} = \bigotimes_{S=1}^{R} |\omega\rangle &= \{w_{S(\mathbf{i})}^{(R)} = \omega_{S_1}\omega_{S_2}\ldots\omega_{S_R} \in \mathbf{R}^+\} \\ &\wedge \langle \mathbf{W}^{(R)}\rangle = \sum (\mathbf{i}) w_{S(\mathbf{i})}^{(R)} = \left(\sum_I \omega_I\right)^R \\ &= (\langle |\omega_I\rangle\rangle)^R = (1)^R = 1 \\ &\rightarrow K\Big(\bigotimes_{S=1}^{R} |\omega\rangle\Big) \equiv K(\mathbf{W}^{(R)}) . \end{aligned} \tag{164}$$

Variational QQSPR. So far the fundamental QQSPR equation has been solved using the usual strategy associated to classical QSPR. Equations (134), (150) or (161) as in classical terms, can be solved for the coefficient vector $|\omega\rangle$. As has been previously commented this is done, by substituting in the expectation value expression (138) the vector $|\pi\rangle$ by an experimental property vector $|p\rangle$, associated to the probe quantum object set P. The result will be obtained in the same way as in classical QSPR, but using the quantum similarity matrices as molecular descriptors. However, it can be proven that the fundamental QQSPR equation can be solved within the usual quantum variational procedures.

a) Similarity Matrix Unrestricted Variational Treatment
For such a purpose it is sufficient to rewrite the second-order expectation value Eq. (138) as:

$$\forall J = 1, N \colon \langle \pi_J\rangle \simeq \sum_{P=1}^{M} \omega_P z_{J;P}^{(1)} + \sum_{P=1}^{M}\sum_{Q\geq P}^{M} \omega_P\omega_Q z_{J;PQ}^{(2)} + O(3) \tag{165}$$

then, considering every quantum object expectation value as a variational function of the parameters within the coefficient vector $|\omega\rangle$, the resulting expression can be varied, taking into account that the density functions, supposedly obtained by quantum mechanical procedures, no longer need variation. In this way, every Jth quantum object will have to possess a specific coefficient vector $|\omega\rangle$, which can be thus named as $|\omega_J\rangle$. That is:

$$\forall J = 1, N \colon \delta\langle \pi_J\rangle \simeq \sum_{P=1}^{M} \delta\omega_P z_{J;P}^{(1)} + 2\sum_{P=1}^{M}\sum_{Q\geq P}^{M} \omega_P\omega_Q z_{J;PQ}^{(2)} + O(3) , \tag{166}$$

then, using the variation condition for the Jth quantum object:

$$\delta\langle \pi_J\rangle = 0 , \tag{167}$$

is obtained:

$$\forall J = 1, N \colon 0 \simeq \sum_{P=1}^{M} \delta\omega_P z_{J;P}^{(1)} + 2\sum_{P=1}^{M}\sum_{Q\geq P}^{M} \delta\omega_P\omega_Q z_{J;PQ}^{(2)} + O(3) , \tag{168}$$

which can be rewritten as:

$$\forall J = 1, N \wedge P = 1, M \colon 0 \simeq z_{J;P}^{(1)} + 2\sum_{Q=1}^{M} \omega_Q z_{J;PQ}^{(2)} + O(3) . \tag{169}$$

This last equation can be expressed in matrix form, using the appropriate similarity matrices as previously defined in Eqs. (145) and (146):

$$\forall J = 1, N \colon \mathbf{z}_J^{(1)} + 2\mathbf{Z}_J^{(2)}|\omega_J\rangle = 0 , \tag{170}$$

thus, the specific coefficients for each quantum object may be computed as:

$$\forall J = 1, N\colon |\omega_J\rangle = -\frac{1}{2}\left[\mathbf{Z}_J^{(2)}\right]^{-1}\mathbf{z}_J^{(1)} . \tag{171}$$

This is the same as associating a particular operator W to each quantum object, and such a result is not too surprising a feature, as the operator W can be easily supposed to vary from one quantum object to another, in the same way as Hamilton operators do. The variational expectation value for the Jth object could be obtained in this case as:

$$\langle\pi_J\rangle \simeq \left\langle\omega_J|\mathbf{z}_J^{(1)}\right\rangle + \left\langle\omega_J|\mathbf{Z}_J^{(2)}|\omega_J\right\rangle + O(3) . \tag{172}$$

Using Eq. (171) into Eq. (172), the following expectation value final optimal form will result:

$$\langle\pi_J\rangle \simeq -\frac{1}{4}\left\langle\mathbf{z}_J^{(1)}\middle|\left[\mathbf{Z}_J^{(2)}\right]^{-1}\middle|\mathbf{z}_J^{(1)}\right\rangle + O(3) . \tag{173}$$

b) Expectation Versus Experimental Values
Then, the set of stationary expectation values $|\pi\rangle$ can be compared with the experimental value vector $|p\rangle$, in such a way as to have:

$$|p\rangle = a + b|\pi\rangle , \tag{174}$$

$\{a, b\}$ being some origin and scale parameters, respectively. They can be obtained by the usual well-known regression techniques.

c) Algorithm for Unrestricted Variational QQSPR
Once the set of coefficients $\{a, b\}$ is obtained by using Eq. (174) for a given probe quantum object set, the property expectation value $\langle\pi_K\rangle$ of any new quantum object K, say, with known density function ρ_K, can be employed to estimate the experimental value ρ_K of the quantum object studied property, by using the following steps:
1. Compute: $\left\{\mathbf{z}_K^{(1)}; \mathbf{Z}_K^{(2)}\right\}$ using the basis set B.
2. Evaluate: $\langle\pi_K\rangle \simeq -\frac{1}{4}\left\langle\mathbf{z}_K^{(1)}\middle|\left[\mathbf{Z}_K^{(2)}\right]^{-1}\middle|\mathbf{z}_K^{(1)}\right\rangle + O(3)$
3. Obtain the estimated property: $p_K = a + b\langle\pi_K\rangle$.

Stochastic Similarity Matrices Restricted Variational Treatment. Of course, all that has been said up to now in this section remains valid for stochastic similarity matrices: $\left\{\mathbf{s}_K^{(1)}; \mathbf{S}_K^{(2)}\right\}$, they just have to be used instead of the similarity matrix pair: $\left\{\mathbf{z}_K^{(1)}; \mathbf{Z}_K^{(2)}\right\}$ in the above algorithm. However, the stochastic case may be interesting if the coefficient set $|\omega\rangle$ can be obtained obeying convex conditions as a restriction, so that the previous unrestricted variation algorithm may no longer be applicable.

Expectation Value Jacobi Rotations Variational Form. To obtain the desired restricted variation over the coefficient vector involved in expectation value expressions, a similar procedure as the one employed in developing the ASA technique [75,76,77,78,79,80,81,82,83] could be easily set up to perform the variational computation over Eq. (165), but taking into account the additional restriction of obtaining a convex vector, as a result of the optimization process.

a) Preliminary Considerations
When this option as discussed above is chosen, it is only necessary to express the operator W variational coefficients with the aid of a new free normalized auxiliary vector; in order to ensure the convex conditions hold throughout the entire optimization process, that is:

$$\begin{aligned}|\omega\rangle &= |x\rangle * |x\rangle \wedge \langle x|x\rangle = 1 \rightarrow \langle|\omega\rangle\rangle \\ &= \sum_I \omega_I = \sum_I x_I^2 = 1 \wedge \forall I\colon \omega_I = x_I^2 \in \mathbf{R}^+ .\end{aligned} \tag{175}$$

After this consideration, it is only necessary to obtain the variation of Eq. (165), by applying norm conserving, orthogonal elementary Jacobi rotations [148] into the auxiliary vector $|x\rangle$ element pairs, in order to arrive at an expression, depending on the elementary Jacobi rotation angle, which could be easily optimized later on.
An interesting point at this stage is to realize that such a restricted variational procedure can be applied to higher order equations, with orders larger than the ones studied up to now. This is due to the fact that Jacobi rotations over the auxiliary vector just change a couple of the coefficient auxiliary vector elements each time an elementary Jacobi rotation is performed, and the same occurs with the coefficient vector. This knowledge of the coefficient vector variation can be easily brought into the tensor products and worked out up to any tensor order.
The rest becomes a procedure with somehow a growing technical computational complexity, but defined within a well-structured theoretical background algorithm.

b) Elementary Jacobi Rotations Algorithm Scheme
Elementary Jacobi rotations need the cosine, c, and the sine, s, of a rotation angle. These involved trigonometric functions fulfil the usual convex relationship: $c^2 + s^2 = 1$. When acting over a vector, the Jacobi rotations will change two vector components, the Kth

and Lth, say, leaving the remaining components as they are:

$$|x\rangle = \begin{pmatrix} \cdots \\ x_K \\ \cdots \\ x_L \\ \cdots \end{pmatrix} \rightarrow \begin{pmatrix} \cdots \\ cx_K - sx_L \\ \cdots \\ sx_K + cx_L \\ \cdots \end{pmatrix}$$

$$\Rightarrow |w\rangle = \begin{pmatrix} \cdots \\ x_K^2 \\ \cdots \\ x_L^2 \\ \cdots \end{pmatrix} \rightarrow \begin{pmatrix} \cdots \\ (cx_K - sx_L)^2 \\ \cdots \\ (sx_K + cx_L)^2 \\ \cdots \end{pmatrix} . \quad (176)$$

It is easy to obtain the variation in the coefficient vector due to an elementary Jacobi rotation as:

$$|\delta\omega\rangle = v_{KL} \begin{pmatrix} \cdots \\ -1 \\ \cdots \\ +1 \\ \cdots \end{pmatrix} = v_{KL}(|e_L\rangle - |e_K\rangle) , \quad (177)$$

where $\{|e_K\rangle, |e_L\rangle\}$ are the corresponding canonical basis set vectors. The scalar coefficient v_{KL} possesses the form:

$$v_{KL} = s^2\left(x_K^2 - x_L^2\right) + 2csx_Kx_L . \quad (178)$$

Then, employing this result in the equivalent expression of Eq. (147), but written in expectation value matrix form, the following can be deduced:

$$\langle\delta\pi\rangle = \langle\delta\omega|(|\mathbf{z}^{(1)}\rangle + 2\mathbf{Z}^{(2)}|\omega\rangle) + \langle\delta\omega|\mathbf{Z}^{(2)}|\delta\omega\rangle , \quad (179)$$

where the quantum object subindex has been taken out to simplify the notation. Then, upon substituting the coefficient vector variation:

$$\langle\delta\pi\rangle = v_{KL}\left[\left(z_L^{(1)} - z_K^{(1)}\right) + 2\sum_I \omega_I\left(Z_{IL}^{(2)} - Z_{IK}^{(2)}\right)\right] + v_{KL}^2\left(Z_{KK}^{(2)} + Z_{LL}^{(2)} - 2Z_{KL}^{(2)}\right) \quad (180)$$

which, upon equalization to zero and terms rearrangement, can be expressed as a second-order equation on the elementary Jacobi rotation sine and cosine:

$$As^2 + Bsc + \beta = 0 , \quad (181)$$

with the coefficients A and B defined as:

$$\begin{aligned} A &= \alpha(\omega_K - \omega_L) \\ B &= 2\alpha x_K x_L \end{aligned} \quad (182)$$

and, besides, the parameters are constructed by the elements of the similarity matrices in the following way:

$$\begin{aligned} \alpha &= Z_{KK}^{(2)} + Z_{LL}^{(2)} - 2Z_{KL}^{(2)} \\ \beta &= \left(z_L^{(1)} - z_K^{(1)}\right) + 2\sum_I \omega_I\left(Z_{IL}^{(2)} - Z_{IK}^{(2)}\right) . \end{aligned} \quad (183)$$

Higher Order Stochastic Expectation Value Variational Treatment.

a) General Comments

Whenever Eq. (161) is studied, after being conveniently modified for the expectation values form,

$$\forall J = 1, N\colon \langle\pi_J\rangle = \sum_{R=1}^{n} \langle \mathbf{S}_J^{(R)} * \mathbf{W}^{(R)}\rangle + O(n+1) \quad (184)$$

the obvious fact appears that the variation will affect just the Rth order tensor products $\mathbf{W}^{(R)}$ of the coefficient vector. So it can be written, dropping the quantum object subindex J just for convenience, as before:

$$\langle\delta\pi\rangle = \sum_{R=1}^{n} \langle \mathbf{S}^{(R)} * \delta\mathbf{W}^{(R)}\rangle + O(n+1) , \quad (185)$$

so the relevant variation will be associated to the terms $\delta\mathbf{W}^{(R)}$, which can be easily written, using a tensor notation as:

$$\begin{aligned} \delta\mathbf{W}^{(R)} &= \delta\left(\bigotimes_{S=1}^{R} |\omega\rangle\right) \\ &= \sum_{S=1}^{R} \binom{R}{S} \left[\left(\bigotimes_{P=1}^{R-S} |\omega\rangle\right) \otimes \left(\bigotimes_{Q=1}^{S} |\delta\omega\rangle\right)\right] , \end{aligned} \quad (186)$$

but being the definition of the coefficient vector variation, upon Jacobi rotations, well known from Eq. (177), it can be written:

$$\begin{aligned} \delta\mathbf{W}^{(R)} &= \sum_{S=1}^{R} \binom{R}{S} (v_{KL})^S \\ &\cdot \left[\left(\bigotimes_{P=1}^{R-S} |\omega\rangle\right) \otimes \left(\bigotimes_{Q=1}^{S} [(|e_L\rangle - |e_K\rangle)]\right)\right] . \end{aligned} \quad (187)$$

So in this way, the restricted variation of the expectation value QQSPR equations, using elementary Jacobi rotations, is clearly defined up to any order.

b) A Computational Detail Concerning Tensor Products of the Difference of Two Canonical Vectors

The tensor product of the difference between the pair of canonical basis set vectors:

$$|e_L\rangle - |e_K\rangle = \begin{pmatrix} \cdots \\ -1 \\ \cdots \\ +1 \\ \cdots \end{pmatrix} \equiv |L\rangle - |K\rangle \equiv |L-K\rangle\,, \quad (188)$$

which appears in Eq. (187), may be expressed in terms of a nested summation symbol. For example, up to second order the sum of the four tensor terms is readily written as:

$$\begin{aligned} &|L-K\rangle \otimes |L-K\rangle \\ &\quad = |L\otimes L\rangle - |L\otimes K\rangle - |K\otimes L\rangle + |K\otimes K\rangle \end{aligned} \quad (189)$$

with the obvious meaning for the involved tensors:

$$|L\otimes L\rangle = |e_L\rangle \otimes |e_L\rangle = \mathbf{E}_{LL} = \{e_L L; PQ = \delta_{LP}\delta_{LQ}\} \quad (190)$$

and so on.
In general, up to Sth order:

$$\bigotimes_{Q=1}^{S} |L-K\rangle = \sum (\mathbf{i})\sigma(Q(\mathbf{i}))|Q(\mathbf{i})\rangle\,, \quad (191)$$

where $Q(\mathbf{i}) = \{Q_1 \otimes Q_2 \ldots \otimes Q_S\}$ is any of the possible 2^n combinations with repetition of the indices K and L, the symbol $|Q(\mathbf{i})\rangle$ meaning a tensor product of the initial canonical basis set vectors with such an index repetition. That is: an object equivalent to a canonical hypermatrix, whose elements are all zero, except the one with indices associated to those entering the set. Also $\sigma(Q(\mathbf{i}))$ corresponds to the sign, associated to the fact that the index K appears in $Q(\mathbf{i})$ an even, $\sigma(Q(\mathbf{i})) = +1$, or odd, $\sigma(Q(\mathbf{i})) = -1$, number of times.

QQSPR Operators, Quantum Similarity Measures and the Fundamental QQSPR Equation

The correspondence principle in quantum theory furnishes the rules to construct Hermitian operators, whose expectation values can be associated with the experimental outcomes of submicroscopic system observables. However, as has been previously commented, for some observables of complex submicroscopic systems, like some biological activities of pharmaceutical interest, the correspondence principle cannot be applied. The construction of the QQSPR operators and the attached fundamental QQSPR equation provide the possibility to attach an approximate quantum mechanical operator to estimate expectation values for these cases.

The QQSPR Operator. The fundamental QQSPR equation arises when from the known quantum objects, belonging to some quantum object set; one realizes that their density function tags: $\{\rho_I(\mathbf{r})\}$ can be used to construct a QQSPR operator in the form:

$$\begin{aligned} \Omega(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \ldots) &= w_0\Theta_0(\mathbf{r}_1) + \sum_I w_I\rho_I(\mathbf{r}_2)\Theta_1(\mathbf{r}_1, \mathbf{r}_2) \\ &+ \sum_I \sum_J w_I w_J \rho_I(\mathbf{r}_2)\rho_J(\mathbf{r}_3)\Theta_2(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) + O(3)\,, \end{aligned} \quad (192)$$

in the Eq. (192) above, w_0 is an arbitrary constant; $\{\Theta_\omega(\mathbf{R})|\omega = 0, 1, 2, \ldots\}$ is a known positive definite operator set, acting as a weight for each term development; finally, $\{w_I\}$ a set of unknown parameters which shall be determined through the fundamental QQSPR equation as will be explained below.

Thus, the structure of a QQSPR operator like the one defined in Eq. (192) has to be seen as the first step of an algorithm permitting the construction of approximate quantum mechanical operators, associated in turn to some observables of complex submicroscopic systems, whose nature do not permit the application of the correspondence principle to construct Hermitian operators for the evaluation of observable values.

The Expectation Values of the QQSPR Operator. In order to determine the parameter set $\{w_I\}$, defining in this way the QQSPR operator as written in Eq. (192), it is just necessary to compute the set of expectation values over the elements of a quantum object set which belong to the core set C, constituted by the core molecules or C-m. Besides a well-defined structure and a known density function, as members of a quantum object set, the C-m are supposed to possess an element of a known property set $P = \{p_K\}$, attached to each one.

In this way one can express every known property of the C-m elements as the expectation value of some QQSPR operator:

$$\begin{aligned} p_K \approx \langle \Omega\rho_K\rangle &= w_0\langle\Theta_0\rho_K\rangle + \sum_I w_I\langle\rho_I\Theta_1\rho_K\rangle \\ &+ \sum_I w_I w_J\langle\rho_I\rho_J\Theta_2\rho_K\rangle + O(3)\,. \end{aligned} \quad (193)$$

Zero-th Order Term. When describing the expectation values of the C-m as computed in Eq. (193), one can con-

sider first the Zero-th order term:

$$\theta_K[\Theta_0] = w_0\langle\Theta_0\rho_K\rangle = w_0\int_D \Theta_0(\mathbf{r}_1)\rho_K(\mathbf{r}_1)\,\mathrm{d}\mathbf{r}_1 ,$$

as being a constant for each C-m, which can be used as an origin shift of the C-m property tags, thus the Zero-th order term: $w_0\Theta_0(\mathbf{r})$ appearing in the above operator definition acts as a gauge. Choosing the Zero-th order operator as the unit, this term becomes proportional to the number of electrons of the C-m considered:

$$\theta_K[I] = w_0\langle\rho_K\rangle = w_0\int_D \rho_K(\mathbf{r}_1)\,\mathrm{d}\mathbf{r}_1 = w_0 N_K .$$

In case shape functions, defined as:

$$\sigma_K(\mathbf{r}) = N_K^{-1}\rho_K(\mathbf{r}) \rightarrow \int_D \sigma_K(\mathbf{r})\,\mathrm{d}\mathbf{r} = 1 ,$$

are employed in the QQSPR operator definition (192) and in the expectation value expression (193), then the Zero-th order contribution to the expectation values $\theta_K[I]$ is a constant for all C-m.

The Zero-th order term can be omitted if it is no longer necessary to shift the property values of the C-m.

First- and Second-Order Expectation Value Terms. The first-order term of the expectation value Eq. (193) contains quantum similarity measure integrals among pairs of density function tags of the C-m, which have been defined a long time ago as:

$$\begin{aligned} z_{IK}[\Theta_1] &= \langle\rho_I\Theta_1\rho_K\rangle \\ &= \int_D\int_D \rho_I(\mathbf{r}_2)\Theta_1(\mathbf{r}_1,\mathbf{r}_2)\rho_K(\mathbf{r}_1)\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2 , \end{aligned}$$

and in the second-order term the triple density quantum similarity measures appear, defined as well as:

$$\begin{aligned} Z_{IJK}[\Theta_2] = \langle\rho_I\rho_J\Theta_2\rho_K\rangle = \int_D\int_D\int_D \rho_I(\mathbf{r}_2)\rho_J(\mathbf{r}_3) \\ \cdot\,\Theta_2(\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3,)\rho_K(\mathbf{r}_1)\,\mathrm{d}\mathbf{r}_1\,\mathrm{d}\mathbf{r}_2\,\mathrm{d}\mathbf{r}_3 . \end{aligned}$$

Fundamental QQSPR Equation Setup. The expectation values of the QQSPR operator, as described in Eq. (193), can be collected in a column vector providing the fundamental QQSPR equation:

$$|\mathbf{p}\rangle \approx |\theta\rangle + \mathbf{Z}_1|\mathbf{w}\rangle + \langle\mathbf{w}|\mathbf{Z}_2|\mathbf{w}\rangle + O(3) , \qquad (194)$$

where in Eq. (194) the following compact symbols have been used: $|\mathbf{p}\rangle = \{p_K\}$ is the C-m properties vector, $|\theta\rangle = \{\theta_K\}$ is the completely determined gauge shift vector, $\{\mathbf{Z}_\omega|\omega = 1, 2, \ldots\}$ is a matrix set containing the quantum similarity measures, for instance: $\mathbf{Z}_1 = \{z_{IK}\}; \mathbf{Z}_2 = \{z_{IJK}\}; \ldots$, and $|\mathbf{w}\rangle = \{w_I\}$ is a column vector bearing the unknown coefficients, which define explicitly the QQSPR operator.

The easiest way to obtain the unknown coefficients $|\mathbf{w}\rangle = \{w_I\}$ is obviously the linear equation contained in the fundamental QQSPR Eq. (194); that is, they can be evaluated by solving:

$$|\mathbf{p}\rangle = |\theta\rangle + \mathbf{Z}_1|\mathbf{w}\rangle \rightarrow |\mathbf{w}\rangle = (\mathbf{Z}_1)^{-1}(|\mathbf{p}\rangle - |\theta\rangle) . \qquad (195)$$

However, Eq. (195) has no predictive power whatsoever. This is so because the first-order similarity matrix $\mathbf{Z}_1$ has to be chosen positive definite by construction, therefore the coefficient vector has a unique determined form.

By predictive power is meant here the possibility to compute the value of the property, which precisely defines the C-m set, for an also known quantum object, which as such possesses well-defined structure and density function, but belongs to the Unknown property molecular set: U, whose elements are made by unknown property molecules or quantum objects, the U-m.

In the last years, since the description of quantum similarity measures, the predictive power of the information contained in the similarity matrices set has been manipulated in the classical QSPR way. For example, using similarity matrices principal components, and finding with them a QSAR model, usually multilinear. This multilinear model can be employed, afterwards, to estimate U-m properties. This amounts to the same as considering the similarity matrices as a source of molecular parameters to construct empirical QSPR.

However, there is a possible way to use the system (195) for predicting properties of U-m without further considerations than the involved algebraic procedures. The possible QQSAPR prediction algorithms will be developed in a separate section.

Evaluation of Unknown Molecular Properties as Expectation Values

In general, one can choose any molecular structure U, possessing an unknown value of the property needed to build up the *core set* triads. Thereafter, one can call such a QO the *U-molecule* or *U-m*, for sake of simplicity. The *U-molecule* can supposedly be associated to a corresponding density function:ρ_U too. Hence, the *U-m* can be certainly considered as a QO. On the other hand, one must keep in mind that, by construction of the QSPR problem, the property lacking in the information about U has to be

already known for all elements of the *core set*. One can easily express an approximate value of the U-m unknown property through the simplified Minkowski norm:

$$\langle \Omega[\rho_U]\rangle = \int_D \Omega[\rho_U]\,\mathrm{d}V \approx \langle \sigma[\rho_U]\rangle + \sum_{P=1}^{N}\sum {}_P(\mathbf{i})x(\mathbf{i})\langle \rho(\mathbf{i})[\rho_U]\rangle + O(N+1)\,, \tag{196}$$

provided that the set of coefficients $\{x_I\}$ is well-defined.

The Minkowski norm in Eq. (196) can be computed in more sophisticated ways, using a known positive definite operator, W say, as a weight in the expectation value definition: $\langle \Omega W[\rho_U]\rangle$, producing weighted quantum similarity measures of type: $\langle \rho(\mathbf{i})W[\rho_U]\rangle$ in the right part of the expression (196). In order to simplify the formalism, here the convention: $W = I$, has been adopted.

Within the QQSPR problem settings, the set of coefficients: $\{x_I\}$, in Eq. (196), which can be ordered as a column vector: $|\mathbf{x}\rangle = \{x_I\}$, is unknown beforehand, but can be already computed from the first-order approach using the *core set* known property values, as will be discussed below.

Quantum Similarity Matrices in the Construction of First-Order QSPR Operators and the Definition of Discrete QOS. The first-order approach of the QSPR operator for the *core set* known molecular property tag set: $\Pi = \{\pi_I\}$ generates the following equation collection:

$$\forall I = 1, n\colon p_I = \pi_I - \langle \sigma[\rho_I]\rangle \approx \sum_J x_J\langle \rho_J[\rho_I]\rangle = \sum_J x_J z_{JI}\,. \tag{197}$$

The set of integrals:

$$\left\{\langle \rho_J[\rho_I]\rangle = \int_D \rho_J\rho_I\,\mathrm{d}V = z_{JI} = z_{IJ} = \int_D \rho_I\rho_J\,\mathrm{d}V = \langle \rho_I[\rho_J]\rangle\right\}\,,$$

appearing in Eqs. (197) can be ordered into a $(n \times n)$ symmetric array, constructing in this way the *quantum similarity matrix*: $\mathbf{Z} = \{z_{IJ}\}$ (QSM). In turn, the ordered set of shifted properties: $\{p_I\}$ can form a $(n \times 1)$ column vector: $|\mathbf{p}\rangle = \{p_I\}$. Therefore, the equation set (197) is simply a linear system, which will be discussed next, in order to describe its possible use for evaluating U-m unknown molecular properties.

Empirical QSPR. In the empirical QSPR problems, the equivalent matrix to the QSM of the QQSPR framework, as described in Sect. "Quantum Similarity", can be obtained in the following manner. Suppose that every molecular structure of M possesses an arbitrarily chosen empirical descriptor vector, in that way:

$$\forall m_I \in \mathrm{M} \to \exists |\mathbf{d}_I\rangle \in \mathrm{D} \wedge \forall I\colon m_I \leftrightarrow |\mathbf{d}_I\rangle\,,$$

then the descriptor set D acts as a tag set to construct an empirical discrete tagged set:

$$\mathrm{Q_D} = \mathrm{M} \times \mathrm{D}\,,$$

such that:

$$\forall \gamma_I \in \mathrm{Q_D} \to \gamma_I = (m_I; |\mathbf{d}_I\rangle) \wedge m_I \in \mathrm{M}; |\mathbf{d}_I\rangle \in \mathrm{D}\,.$$

The discrete tag set D of molecular descriptors can be considered, in turn, as a linearly independent subset of cardinality n belonging to some real m-dimensional column vector space, that is: $\mathrm{D} \subset V_m(\mathbf{R})$. The linear independence of the set D is strictly necessary to construct a matrix comparable in properties to QSM, and in this way, each molecule becomes independently described from the rest. With this information in mind it is easy to construct, $\mathbf{S}_\mathrm{D}$, a symmetric $(n \times n)$ matrix bearing analogous characteristics as the QSM:

$$\forall \{|\mathbf{d}_I\rangle; |\mathbf{d}_J\rangle\} \in \mathrm{D}\colon \mathbf{S}_\mathrm{D} = \{s_{D;IJ} = \langle \mathbf{d}_I|\mathbf{d}_J\rangle = \mathbf{d}_J|\mathbf{d}_I\rangle = s_{D;JI}\} = \{|\mathbf{s}_{\mathrm{D};I}\rangle = \{s_{D;JI}\}\}\,. \tag{198}$$

In fact, constructing the $(m \times n)$ matrix: $\mathbf{D} = \{|\mathbf{d}_I\rangle\}$, whose columns are the elements of the empirical descriptor set D, then the matrix $\mathbf{S}_\mathrm{D}$ can also be defined as the product:

$$\mathbf{S}_\mathrm{D} = \mathbf{D}^\mathrm{T}\mathbf{D}\,,$$

where $\mathbf{D}^\mathrm{T} = \{\langle \mathbf{d}_I|\}$ is the $(n \times m)$ transpose of matrix $\mathbf{D}$, whose rows are the descriptor vectors ordered in such a way. It is easy to see that matrix $\mathbf{S}_\mathrm{D}$, defined in this manner, is coincident with the Gramian matrix of the tag set D. In order to comply with the same standard properties as the QSM the matrix $\mathbf{S}_\mathrm{D}$ has to fulfil: $\mathrm{Det}\,|\mathbf{S}_\mathrm{D}| > 0$. If this is the case, a discrete *empirical object set* $\mathrm{Q_S}$ can be defined as:

$$\mathrm{Q_S} = \mathrm{M} \times \mathbf{S}_\mathrm{D}\,,$$

in close resemblance to the discrete quantum object set $\mathrm{Q_Z}$ described in Eq. (42).

Finally, one shall comment now that, as a consequence of this definition of the set $\mathrm{Q_S}$, the presentation and discussion about the following procedures, which will be studied

in this paper for QSM, can also be applied to the Gramian matrices, associated to empirical descriptor tag sets and thus to the classical QSPR problem.

However, the different background between quantum and empirical points of view induces the necessary emergence of the following considerations. The QQSPR equations are deductible from the usual quantum theoretical considerations; within the same context, they can be easily generalized to contain higher approximation orders. Therefore, the QQSPR equations of any order can certainly possess in general some causal background; while, except for very particular cases, empirical QSPR equations remain arbitrarily constructed and without a clear causal fundament.

First-Order Fundamental QQSPR (FQQSPR) Equation

The analysis of the QQSPR problem can start with the first order or linear fundamental QQSPR equation, involving the *core set*, formed with the molecules of the associated DQOS, which are also linked with known values of some property, according to the considerations noted above.

One can write Eq. (197) in a compact matrix form:

$$\mathbf{Z}|\mathbf{x}\rangle = |\mathbf{p}\rangle \; ; \tag{199}$$

Where the matrix $\mathbf{Z}$ is the already described symmetric QSM, $|\mathbf{p}\rangle$ is the known *core set* property vector and $|\mathbf{x}\rangle$ is a $(n \times 1)$ vector, whose coefficients have to be evaluated.

The predictive power of such an equation is a priori null, because being the QSM: $\mathbf{Z}$, by construction non-singular (otherwise two density functions will be exactly the same), then there always can be computed a QSM inverse: $\mathbf{Z}^{-1}$, obeying the usual relationships: $\mathbf{Z}^{-1}\mathbf{Z} = \mathbf{Z}\mathbf{Z}^{-1} = \mathbf{I}$, in such a way that the trivial result, defining the unknown coefficient vector:

$$|\mathbf{x}\rangle = \mathbf{Z}^{-1}|\mathbf{p}\rangle \; , \tag{200}$$

will be always obtained within a *core set* scenario. Furthermore, one can retrieve the exact value of the property for any molecule of the *core set* QOS choosing the scalar products:

$$\forall I \colon p_I = \langle \mathbf{z}_I|\mathbf{x}\rangle \; . \tag{201}$$

The QSM for diverse *core sets* has been used in a quite large set of prediction studies [26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46], in every case employing up to date statistical tools, the usual procedures currently available in classical QSPR studies, see for example reference [171]. The use of the first-order fundamental QQSPR equation to construct algorithms, which can be utilized as predictive tools, has been previously attempted [59], but it has not been continued in practice. In the present study, the reader can find in the following sections new theoretical developments of the prediction ability of the fundamental QQSPR equation. However, a reminder of some simple linear algebra relating to the FQQSPR equation is needed first in order to understand the following arguments; therefore it will be described in the forthcoming final section.

Future Trends

Procedure for Adding One Molecular Structure to a Known Core Set

The Partition of the FQQSPR Linear Equation. The general setup described until now, amounts to the same as virtually considering the *U-m* as forming part of the *core set*, but with a parametrized value of the unknown property. One can refer to this extension of the *core set* as the *parametrized core set*.

The following coefficient vector partitioned expression can be easily written in terms of the inverse partitioned QSM matrix elements:

$$\begin{pmatrix} |\mathbf{x_0}\rangle \\ x \end{pmatrix} = \begin{pmatrix} \mathbf{Z_0}^{(-1)}|\mathbf{p_0}\rangle + \pi|\mathbf{z}^{(-1)}\rangle \\ \langle \mathbf{z}^{(-1)}|\mathbf{p_0}\rangle + \pi\theta^{(-1)} \end{pmatrix} \; .$$

However, in order to obtain equivalent expressions possessing less entanglement with the elements of the inverse matrix, the most convenient way is to restart the procedure writing explicitly:

$$\begin{pmatrix} \mathbf{Z_0}|\mathbf{x_0}\rangle + x|\mathbf{z}\rangle \\ \langle \mathbf{z}|\mathbf{x_0}\rangle + x\theta \end{pmatrix} = \begin{pmatrix} |\mathbf{p_0}\rangle \\ \pi \end{pmatrix} \; . \tag{202}$$

From the augmented linear equation first component structure, one can obtain:

$$|\mathbf{x_0}\rangle = \mathbf{Z_0^{-1}}|\mathbf{p_0}\rangle - x\mathbf{Z_0^{-1}}|\mathbf{z}\rangle \; , \tag{203}$$

taking into account that the QSM $\mathbf{Z}_0$, associated to the initial *core set*, is non-singular by construction. Therefore, the first right-hand term is just the solution of the FQQSPR linear equation for the initial *core set*, as shown in Eq. (200). Thus, calling:

$$|\mathbf{q}\rangle = \mathbf{Z_0^{-1}}|\mathbf{p_0}\rangle \wedge |\mathbf{a}\rangle = \mathbf{Z_0^{-1}}|\mathbf{z}\rangle \; , \tag{204}$$

then, Eq. (203) could be rewritten as any of the two following equalities:

$$|\mathbf{x_0}\rangle = |\mathbf{q}\rangle - x|\mathbf{a}\rangle = \mathbf{Z_0^{-1}}(|\mathbf{p_0}\rangle - x|\mathbf{z}\rangle) \; . \tag{205}$$

Taking into account Eq. (205), the second component can be written as:

$$\pi = \langle \mathbf{z}|\mathbf{q}\rangle + (\theta - \langle \mathbf{z}|\mathbf{a}\rangle)x = a_0 + a_1 x \,, \tag{206}$$

where:

$$a_0 = \langle \mathbf{z}|\mathbf{q}\rangle \wedge a_1 = \theta - \langle \mathbf{z}|\mathbf{a}\rangle \,; \tag{207}$$

expression (207) shows the expected trivial result consisting of how the *U-m* property and the coefficient, still not evaluated, obviously are linearly related.

Analysis of the FQQSPR Equation. Only in the case that the *U-m* property can be associated to a concrete numerical value: π_U, say; then the exact linear coefficient: x_U can be written in terms of the quantities appearing in Eq. (206) as:

$$x_U = \frac{\pi_U - a_0}{a_1} = \frac{\pi_U - \langle \mathbf{z}|\mathbf{q}\rangle}{\theta - \langle \mathbf{z}|\mathbf{a}\rangle} = \frac{\pi_U - \langle \mathbf{z}|\mathbf{Z}^{-1}|\mathbf{p}_0\rangle}{\theta - \langle \mathbf{z}|\mathbf{Z}^{-1}|\mathbf{z}\rangle} \,. \tag{208}$$

Equation (208), although it will never hold exactly by construction, tells us about several interesting features.

First, admitting through the previous discussion that the vector $|\mathbf{q}\rangle$ is nothing else than the exact linear coefficient set for the *core set*, then the scalar product: $a_0 = \langle \mathbf{z}|\mathbf{q}\rangle \equiv \pi^{(0)}$, is nothing else than an estimation of the *U-m* property, $\pi^{(0)}$, say, using the discrete representation of the *U-m* with respect to the *core set*. Thus, the numerator of Eq. (208) corresponds to the difference between this rough approximation and the exact property value, if known, of the *U-m*. Obviously enough, if: $\pi_U = \pi^{(0)}$ holds, then the *U-m* coefficient will be null, as the *U-m* property could be solely computed by the descriptor $|\mathbf{z}\rangle$.

Second, the denominator in Eq. (208), tells us about the difference between *U-m* self-similarity and the norm of the vector $|\mathbf{z}\rangle$ computed in the QSM *reciprocal* space, defined as the vector space where the inverse of the QSM acts as a metric matrix. An identity as: $\theta = \langle \mathbf{z}|\mathbf{Z}^{-1}|\mathbf{z}\rangle$, will produce an unacceptable linear algebra result, whichever value the *U-m* property could be. It is plausible to suppose, therefore, that in well-behaved FQQSPR prediction problems, the following inequality shall always hold: $\theta \neq \langle \mathbf{z}|\mathbf{Z}^{-1}|\mathbf{z}\rangle$.

One can conclude, within the settings of the *U-m* prediction problem, that the linear structure of the FQQSPR equation does not permit the evaluation of the *U-m* property in Eq. (206), unless the coefficient x appears defined in some way. The exact coefficient value x_U can be derived, if and only if, a given concrete value of the property is known, but in this case, the prediction problem will not a priori exist as such. Only if the property of the *U-m* appears in a parametrized form, the problem can be handled in an approximate way.

Thus, one arrives at the logical conclusion in that a prediction obstacle is already present in the case of a one-dimensional representation of the QSPR operator Ω, even if the quantum similarity description discrete vector tag: $\begin{pmatrix} |\mathbf{z}\rangle \\ \theta \end{pmatrix}$ is known for the extra added *U-m* structure, but the corresponding property value is not defined, but considered as a parameter.

Analysis of U-m Predicted Property Values. Therefore, the aim of the following discussion will be to find an appropriate way to determine a reasonable optimal approach for the *U-m* coefficient x, by means of manipulating Eq. (205) in order that the unknown property parametrized value π could be estimated using Eq. (206). If some optimal coefficient value $x^{(\mathrm{opt})}$ is found, Eq. (206) can be rewritten as:

$$\pi^{(\mathrm{Estimate})} = \pi^{(0)} + (\theta - \langle \mathbf{z}|\mathbf{a}\rangle)x^{(\mathrm{opt})} \,; \tag{209}$$

in this way the role of the estimated coefficient $x^{(\mathrm{opt})}$ appears with a clear meaning now: it constitutes one of the factors to correct the rough initial estimate of the *U-m* property $\pi^{(0)}$, which can be obtained from the primary information provided by the *core set* by just using Eq. (201). Equation (209) above also enhances the leading role of the *U-m* self-similarity θ for such a property correction-estimation task. In Eq. (209), the *U-m* self-similarity appears shifted, in turn, by the norm of the *U-m* discrete representation vector, $|\mathbf{z}\rangle$, with respect to the *core set*: $\langle \mathbf{z}|\mathbf{a}\rangle = \langle \mathbf{z}|\mathbf{Z}^{-1}|\mathbf{z}\rangle$, computed over *reciprocal* space.

Formulation of the Optimization Problem

In any of both direct and reciprocal space cases, as expected from the linear structure of the fundamental equations used and provided that: $\lambda \in \mathbf{R}$, then it can be written for the unknown sought property:

$$\pi = a + b\lambda \tag{210}$$

also, the equation for the core set unknowns can be written in general as:

$$|\mathbf{u}\rangle = \mathbf{A}(|\pi\rangle - \lambda|a\rangle) \,, \tag{211}$$

where $\mathbf{A}$ is a positive definite matrix.

The unknown property in Eq. (210) will be well-defined whenever, using Eq. (211), one could obtain a well-defined value of the parameter: λ. As the solution of

Eq. (210) corresponds to an infinite collection of real elements, the restricted solution in the case of putting one molecule in, is not unique, as from Eq. (211) one can describe several possible ways to obtain optimal values of the parameter λ. For instance:

a) Defining the difference vector: $|d\rangle = |\pi\rangle - \lambda|a\rangle$, a difference norm can be constructed:

$$\langle d|d\rangle = \langle \pi|\pi\rangle - 2\lambda\langle \pi|a\rangle + \lambda^2\langle a|a\rangle \,, \tag{212}$$

optimizing the expression (212) with respect to the parameter, provides:

$$\lambda^{\text{opt}} = \frac{\langle \pi|a\rangle}{\langle a|a\rangle} \,,$$

besides the optimal value of the difference norm will be a minimum, as the second-order coefficient in Eq. (212) is a Euclidean norm of a non-null vector.

b) One can consider the norm of vector $|\mathbf{u}\rangle$ as defined in Eq. (211) the objective function to be optimized; in this case it can be written:

$$\langle \mathbf{u}|\mathbf{u}\rangle = \langle \pi|\mathbf{A}|\pi\rangle - 2\lambda\langle \pi|\mathbf{A}|a\rangle + \lambda^2\langle a|\mathbf{A}|a\rangle \,,$$

So the optimal value of the parameter is now:

$$\lambda^{\text{opt}} = \frac{\langle \pi|\mathbf{A}|a\rangle}{\langle a|\mathbf{A}|a\rangle} \,,$$

which provides a similar form as in the previous procedure, weighted by the transformation matrix $\mathbf{A}$.

c) The scalar product of the vectors $\{\mathbf{u}\rangle; |\pi\rangle\}$ can be optimized, the objective function is now:

$$\begin{aligned} |\langle \pi|\mathbf{u}\rangle|^2 &= |\langle \pi|\mathbf{A}|\pi\rangle - \lambda\langle \pi|\mathbf{A}|a\rangle|^2 \\ &= |\langle \pi|\mathbf{A}|\pi\rangle|^2 - 2\lambda\langle \pi|\mathbf{A}|\pi\rangle\langle \pi|\mathbf{A}|a\rangle \\ &\quad + \lambda^2|\langle a|\mathbf{T}|a\rangle|^2 \,, \end{aligned}$$

producing:

$$\lambda^{\text{opt}} = \frac{\langle \pi|\mathbf{A}|\pi\rangle}{\langle \pi|\mathbf{A}|a\rangle} \,.$$

d) The scalar product of the vectors $\{\mathbf{u}\rangle; |a\rangle\}$ can be now optimized, in an equivalent way as in the previous procedure; that is, using the objective function:

$$\begin{aligned} |\langle t|\mathbf{u}\rangle|^2 &= |\langle a|\mathbf{A}|\pi\rangle - \lambda\langle a|\mathbf{A}|a\rangle|^2 \\ &= |\langle a|\mathbf{A}|\pi\rangle|^2 - 2\lambda\langle a|\mathbf{A}|\pi\rangle\langle a|\mathbf{A}|a\rangle \\ &\quad + \lambda^2|\langle a|\mathbf{A}|a\rangle|^2 \,, \end{aligned}$$

which permits one to obtain the optimal value:

$$\lambda^{\text{opt}} = \frac{\langle a|\mathbf{A}|\pi\rangle}{\langle a|\mathbf{A}|a\rangle} \,;$$

this result, however, corresponds to the same restriction as the one previously studied in procedure II of reference [60]. Thus, optimizing the norm $\langle \mathbf{u}|\mathbf{u}\rangle$ seems to be equivalent to optimizing the squared module: $|\langle a|\mathbf{u}\rangle|^2$.

A Quadratic Error Restricted First-Order (n+1) Estimation

The easiest procedure to overcome the previously mentioned evaluation impasse for the unknown property of the *U-m*, concretely the approximate evaluation of the coefficient x, appears naturally associated to the possibility to introduce a restriction of some sort into the FQQSPR equation solution. Here follows the description of one among some possible restriction procedures. One will discuss a second option and sketch some alternative procedures as well, within a separate section below.

Setting up the Problem. By inspecting Eq. (205), one can define the difference vector:

$$|\Delta\rangle = |\mathbf{p_0}\rangle - x|\mathbf{z}\rangle \,, \tag{213}$$

and compute with it the following associated quadratic error, which in this case describes a second-order polynomial of the unknown coefficient x:

$$\varepsilon^{(2)} = \langle \Delta|\Delta\rangle = \langle \mathbf{p_0}|\mathbf{p_0}\rangle - 2x\langle \mathbf{z}|\mathbf{p_0}\rangle + x^2\langle \mathbf{z}|\mathbf{z}\rangle \tag{214}$$

which, in turn, using the usual null gradient condition, allows us to obtain an optimal value of the coefficient x, obeying the simple quotient expression:

$$x^{(\text{opt})} = \frac{\langle \mathbf{z}|\mathbf{p}_0\rangle}{\langle \mathbf{z}|\mathbf{z}\rangle} \,. \tag{215}$$

The optimal coefficient value (215) produces a minimum of the quadratic error, being the second-order coefficient of the quadratic error polynomial (214), associated to the positive definite Euclidean norm of the *U-m* discrete representation with respect to the *core set* density tags, that is: $\langle \mathbf{z}|\mathbf{z}\rangle > 0$. Such a quadratic error restriction is equivalent to constructing a difference vector (213) with elements as small as possible. In the case, quite unlikely to occur, where the known property vector $|\mathbf{p_0}\rangle$ and the *U-m* quantum similarity vector $|\mathbf{z}\rangle$ are linearly dependent, the present restriction will construct a difference vector (213),

which will be exactly the null vector at the optimal value of the unknown.

Using $x^{(\mathrm{opt})}$, the optimal value obtained with Eq. (215), the corresponding unknown property for the U-m can be straightforwardly predicted using Eq. (206):

$$\pi^{(\mathrm{opt})} = \langle \mathbf{z}|\mathbf{c}\rangle + \left(\theta - \langle \mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{z}\rangle\right) x^{(\mathrm{opt})}$$
$$= \langle \mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{p}_0\rangle + \left(\theta - \langle \mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{z}\rangle\right) \frac{\langle \mathbf{z}|\mathbf{p}_0\rangle}{\langle \mathbf{z}|\mathbf{z}\rangle} ,$$

providing an expression, which can be easily rearranged by defining the matrix:

$$\mathbf{A} = \mathbf{Z_0^{-1}} + \alpha \mathbf{I} \wedge \alpha = \langle \mathbf{z}|\mathbf{z}\rangle^{-1} \left(\theta - \langle \mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{z}\rangle\right) .$$

Therefore, one can compactly write the optimal property equation for the U-m as:

$$\pi^{(\mathrm{opt})} = \langle \mathbf{z}|\mathbf{A}|\mathbf{p}_0\rangle ,$$

moreover, defining the $(1 \times n)$ row vector: $\langle \mathbf{b}| = \langle \mathbf{zA}$; then, this new description permits us to write the estimated property by means of the following scalar product form:

$$\pi^{(\mathrm{opt})} = \langle \mathbf{b}|\mathbf{p}_0\rangle = \sum_I b_I p_{0;I} . \tag{216}$$

Hence, within the linear FQQSPR equation under the minimal quadratic error restriction, the result (216) shows that the *estimated* optimal unknown property for any U-m, is always expressible as a linear functional of the known molecular properties of the *core set*. Such a result is in agreement with usual classical QSPR treatments.

Additionally, in a very unlikely case, where a linear dependence of the *core set* property vector $|\mathbf{p_0}\rangle$ and the U-m quantum similarity vector: $|\mathbf{z}\rangle$ applies, that is whenever: $|\mathbf{p}_0\rangle = \lambda|\mathbf{z}\rangle$; then, the optimal estimated property value will be expressible as a multiple of the U-m self-similarity:

$$\pi_{\parallel}^{(\mathrm{opt})} = \theta x^{(\mathrm{opt})} = \theta \frac{\langle \mathbf{zp}_0\rangle}{\langle \mathbf{zz}\rangle} = \lambda\theta . \tag{217}$$

It must be finally noted in any case that the gauge operator expectation value: $\langle \sigma[\rho_U]\rangle$, for the U-m, if different from zero, shall be added to the optimal value of the property in Eq. (216) or (217) in order to retrieve the predicted value corresponding to the original property set.

A (2 + 1) Naïve Application Example. In order to illustrate the above procedure one can consider a simple case as follows. Supposing that the *core set* is made of just two molecular structures: $\{A, B\}$ say, with a known shifted property vector:

$$|\mathbf{p}\rangle = \begin{pmatrix} p_A \\ p_B \end{pmatrix} ,$$

and also admitting that the U-m, possessing the unknown parametric property, π, could be labeled as $\{U\}$; then, the QSM of the *core set* and the similarity vector of the U-m can be respectively written as:

$$\mathbf{Z}_0 = \begin{pmatrix} z_{AA} & z_{AB} \\ z_{AB} & z_{BB} \end{pmatrix} \wedge |\mathbf{z}\rangle = \begin{pmatrix} z_{AU} \\ z_{BU} \end{pmatrix} ,$$

with z_{UU} representing the U-m self-similarity measure. One can readily compute the *core set* similarity matrix inverse:

$$\mathbf{Z}_0^{-1} = D^{-1} \begin{pmatrix} z_{BB} & -z_{AB} \\ -z_{AB} & z_{AA} \end{pmatrix} \wedge D = z_{AA} z_{BB} - z_{AB}^2 .$$

It must be now said that when doing this kind of calculation care must be taken with the values of the (2×2) similarity matrix determinant D, because a value approaching zero can render the procedure useless and generate unpredictable computational errors. For all molecular pairs $\{A, B\}$ of the *core set*, the value of the determinant D has to be checked to be significantly greater than a positive definite threshold, that is:

$$\forall \{A, B\}: D \geq \varepsilon > 0 ;$$

failure to comply with this condition for any *core set* molecular pair may well represent a computationally unbalanced QSM triplet $\{A, B, U\}$. This test shall be added to the already described coherent calculation procedures, when accurate QSM have to be computed.

However, the positive definite determinant condition can also be rewritten in a positive definite quantum similarity matrix condition, that is:

$$z_{BB} > z_{AA}^{-1} z_{AB}^2 .$$

A $(N \times N)$ positive definite condition problem, which corresponds in general to the positive definite nature of the quantum similarity matrices, can be shown that it can be readily solved for any core set cardinality, but the nature of this subject, although of capital importance for application purposes, appears to be marginal in the present work and hence will be studied elsewhere.

Thus, one can express the needed vector resulting from the product: $\mathbf{Z}_0^{-1}|\mathbf{z}\rangle$ as:

$$\mathbf{Z}_0^{-1}|\mathbf{z}\rangle = D^{-1} \begin{pmatrix} z_{BB} z_{AU} - z_{AB} z_{BU} \\ z_{AA} z_{BU} - z_{AB} z_{AU} \end{pmatrix} ,$$

and the same can be obtained for the scalar products entering the restricted optimal solution:

$$\langle \mathbf{z}|\mathbf{z}\rangle = z_{AU}^2 + z_{BU}^2 \wedge \langle \mathbf{z}|\mathbf{p}\rangle = z_{AU} p_A + z_{BU} p_B .$$

Finally, one can also write:

$$\langle \mathbf{z}|\mathbf{Z}_0^{-1}|\mathbf{z}\rangle = D^{-1}[z_{AU}(z_{BB}z_{AU} - z_{AB}z_{BU}) + z_{BU}(z_{AA}z_{BU} - z_{AB}z_{AU})]$$

and

$$\langle \mathbf{z}|\mathbf{Z}_0^{-1}|\mathbf{p}\rangle = D^{-1}[p_A(z_{BB}z_{AU} - z_{AB}z_{BU}) + p_B(z_{AA}z_{BU} - z_{AB}z_{AU})] .$$

Therefore, one can obtain the unknown optimal property value, after some trivial manipulation of the previous quantities as:

$$\pi_{AB;U}^{\text{opt}} = \left[D\left(z_{AU}^2 + z_{BU}^2\right)\right]^{-1}[(\alpha z_{AU} - \beta z_{BU})p_A + (\alpha z_{BU} + \beta z_{AU})p_B] , \quad (218)$$

where the following symbols are used:

$$\alpha = Dz_{UU}$$
$$\beta = z_{AU}z_{BU}(z_{AA} - z_{BB}) + z_{AB}(z_{BU}^2 - z_{AU}^2) .$$

The expression (218) for the optimal quadratic error restricted property of the *U-m* self-similarity, constitutes an explicit example involving a very limited number of molecular structures. However, it also corresponds to a general equation involving any triad of molecules, where one of them acts as the *U-m*.

This simple way of estimating a property can be structured into a procedure involving all the: $N = 1/2[n(n-1)]$ possible *core set* distinct molecular pairs. Indeed, given a *core set* and a *U-m*, one can compute all the possible property estimates using Eq. (218). Such a process will produce a set of N values of the *U-m* estimated property:

$$\left\{\pi_{IJ;U}^{\text{Opt}} \middle| \forall(I = 1, n-1; J = I+1, n)\right\} ,$$

which can be finally manipulated in the usual statistical way.

Such an example opens the way to other possible choices using as probe core sets three or another number of QO. In order to leave this study within reasonable limits this possibility will not be further investigated here.

The (n + m) Case Under a Quadratic Error Restriction. One can extend the estimation procedure, outlined in the previous section, in order to include a *U-m set* of arbitrary cardinality, *m* say, so a general quadratic error restricted scheme can be also described in this more general case. One may write the partition of the QSM into the *core set*, bearing the label 0 and the *U-m set*, bearing the label 1, then the FQQSPR equation can be written as:

$$\begin{pmatrix} \mathbf{Z}_{00} & \mathbf{Z}_{01} \\ \mathbf{Z}_{01}^{\mathrm{T}} & \mathbf{Z}_{11} \end{pmatrix} \begin{pmatrix} |\mathbf{x}_0\rangle \\ |\mathbf{x}_1\rangle \end{pmatrix} = \begin{pmatrix} |\mathbf{p}_0\rangle \\ |\mathbf{p}_1\rangle \end{pmatrix} , \quad (219)$$

which produce the two matrix equations, as follows:

$$\begin{aligned} \mathbf{Z}_{00}|\mathbf{x}_0\rangle + \mathbf{Z}_{01}|\mathbf{x}_1\rangle &= |\mathbf{p}_0\rangle \\ \mathbf{Z}_{01}^{\mathrm{T}}|\mathbf{x}_0\rangle + \mathbf{Z}_{11}|\mathbf{x}_1\rangle &= |\mathbf{p}_1\rangle . \end{aligned} \quad (220)$$

So, from the first element of Eq. (220), one can deduce:

$$|\mathbf{x}_0\rangle = \mathbf{Z}_{00}^{-1}[|\mathbf{p}_0\rangle - \mathbf{Z}_{01}|\mathbf{x}_1\rangle]$$

with the possibility to construct a difference vector:

$$|\mathbf{d}\rangle = |\mathbf{p}_0\rangle - \mathbf{Z}_{01}|\mathbf{x}_1\rangle . \quad (221)$$

Then, one may immediately use the difference vector (221) to define a quadratic error function like:

$$\begin{aligned} \varepsilon^{(2)} &= \langle \mathbf{d}|\mathbf{d}\rangle \\ &= \langle \mathbf{p}_0|\mathbf{p}_0\rangle - 2\langle \mathbf{p}_0|\mathbf{Z}_{01}|\mathbf{x}_1\rangle + \langle \mathbf{x}_1|\mathbf{Z}_{01}^{\mathrm{T}}\mathbf{Z}_{01}|\mathbf{x}_1\rangle , \end{aligned} \quad (222)$$

which upon derivation and submitted to the extremum condition of null gradient, produces:

$$\frac{\partial \varepsilon^{(2)}}{\partial |\mathbf{x}_1\rangle} = -2\mathbf{Z}_{01}^{\mathrm{T}}|\mathbf{p}_0\rangle + 2\mathbf{Z}_{01}^{\mathrm{T}}\mathbf{Z}_{01}|\mathbf{x}_1\rangle = |\mathbf{0}\rangle ,$$

so, the *U-m set* unknown coefficients, restricted to minimal quadratic error, can be obtained by means of the matrix expression:

$$|\mathbf{x}_1^{\text{Opt}}\rangle = \left(\mathbf{Z}_{01}^{\mathrm{T}}\mathbf{Z}_{01}\right)^{-1}\mathbf{Z}_{01}^{\mathrm{T}}|\mathbf{p}_0\rangle . \quad (223)$$

The solution in Eq. (223) depends only on the circumstance that the matrix:

$$\mathbf{A}_{11} = \mathbf{Z}_{01}^{\mathrm{T}}\mathbf{Z}_{01}$$

shall be non-singular. In fact, the matrix $\mathbf{A}_{11}$ corresponds to the scalar products of the matrix representations of the *U-m* set with respect to the *core set* elements. This condition, if the computed similarities are not faulty, constitutes a metric matrix of the *U-m* space, subtended by the *U-m* QSM columns. Thus, provided that the *U-m* discrete representations with respect to the *core set* are linearly independent, the inverse of $\mathbf{A}$ is guaranteed to exist, as it will be positive definite; moreover, implying the quadratic error (222) is a minimum at the value $|\mathbf{x}_1^{\text{Opt}}\rangle$ given by Eq. (223).

One can easily estimate the unknown parametrized *U-m* property vector $|\mathbf{p}_1\rangle$, submitted to the quadratic error restriction, after defining the auxiliary matrix:

$$\mathbf{X}_{10} = \mathbf{A}_{11}\mathbf{Z}_{01}^{\mathrm{T}} ,$$

and performing some rearrangements using trivial matrix algebra, it is obtained:

$$|\mathbf{p}_1^{\mathrm{Opt}}\rangle = \left[\mathbf{Z}_{01}^{\mathrm{T}}\mathbf{Z}_{00}^{-1} + \left(\mathbf{Z}_{11} - \mathbf{Z}_{01}^{\mathrm{T}}\mathbf{Z}_{00}^{-1}\mathbf{Z}_{01}\right)\mathbf{X}_{10}\right]|\mathbf{p}_0\rangle . \quad (224)$$

Equation (224) shows that the predicted *U-m set* property vector is a linear transformation of the *core set* known properties, a result consistent with the unique *U-m* case, already described in the previous section and coincident with the usual classical QSPR procedures.

Alternative Restrictions and the Associated Prediction Algorithms

The case analyzed in Sect. "Formulation of the Optimization Problem" is not at all unique. One can describe other possible alternative restrictions, which can be imposed to the FQQSPR equation, as follows in this section.

An Alternative Orthogonality Restriction. Here, choosing one of the possible procedures, a deep discussion will be carried out for a *U-m* set bearing one element only, because the extension to the case of several elements is trivial and similar to the previous Subsect. "A Quadratic Error Restricted First-Order (n+1) Estimation" development, although a brief outline will be given for the sake of completeness. Finally, the remnant plausible restrictions will be only sketched, because the procedure to obtain the application algorithms follows the same trends as in the explicit examples.

a) The $(n+1)$ framework
For the purpose of finding an alternative restriction to the one described in the previous Subsect. "A Quadratic Error Restricted First-Order (n+1) Estimation", it is necessary to recover the first matrix equation of the partition (202), leading to Eq. (203). Then, upon left multiplying both sides by the row vector $\langle\mathbf{z}|$, one can obtain:

$$\langle\mathbf{z}|\mathbf{x_0}\rangle = \langle\mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{p_0}\rangle - x\langle\mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{z}\rangle = \alpha_0 - \alpha_1 x . \quad (225)$$

Then, one can use the resulting Eq. (225) to minimize the scalar product $\langle\mathbf{z}|\mathbf{x_0}\rangle$, appearing on the left-hand side. As in the previous treatment, the right-hand side of Eq. (225) can be considered a difference, which can generate a quadratic error function to be minimized with respect to the unknown parameter, x, which can be evaluated in this manner afterwards. After a trivial manipulation one finds:

$$\varepsilon^{(2)} = |\langle\mathbf{z}|\mathbf{x_0}\rangle|^2 = |\alpha_0 - \alpha_1 x|^2 = \alpha_0^2 - 2\alpha_0\alpha_1 x + \alpha_1^2 x^2 , \quad (226)$$

and in this way the extremum condition imposed upon Eq. (226), becomes a manner to obtain the optimal value of the unknown coefficient:

$$2\alpha_0\alpha_1 - 2\alpha_1^2 x = 0 \rightarrow x_{\perp}^{\mathrm{opt}} = \frac{\alpha_0}{\alpha_1} = \frac{\langle\mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{p_0}\rangle}{\langle\mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{z}\rangle} . \quad (227)$$

The result depicted in the quotient of expression (227), turns out to be equivalent to agreeing that the imposed restriction considers the vectors orthogonal in the scalar product (225), or: $\langle\mathbf{z}|\mathbf{x_0}\rangle = 0$. Admitting that, such a restriction produces a second equation in the system (202), which simplifies to:

$$\pi_{\perp}^{\mathrm{opt}} = \theta x_{\perp}^{\mathrm{opt}} = \theta\frac{\langle\mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{p_0}\rangle}{\langle\mathbf{z}|\mathbf{Z_0^{-1}}|\mathbf{z}\rangle} = \theta\frac{\langle\mathbf{z}|\mathbf{q}\rangle}{\langle\mathbf{z}|\mathbf{a}\rangle} . \quad (228)$$

Expression (228) resembles the limiting case value of the unknown property: $\pi_{\perp}^{\mathrm{opt}}$, for the quadratic error restriction studied in Subsect. "A Quadratic Error Restricted First-Order (n+1) Estimation", as shown into Eq. (217), when the vector of the known properties and the *U-m* discrete representation with respect to the *core set* become linearly dependent. Similar scalar products appear in both expressions. However, in the present orthogonal restriction, they are evaluated using as a metric the inverse of a matrix, which is nothing else than the QSM, associated to the *core set* space. Thus, the scalar products entering Eq. (228) are computed in the *reciprocal* space of the vector space possessing a metric $\mathbf{Z}_0$, as previously commented in Sect. "Linear Quantum QSPR Fundamental Equation". Moreover, the approximate coefficient value in Eq. (227) resembles the exact expression of the *U-m* coefficient, as described in Eq. (208). Rewriting Eq. (228) as:

$$\pi_{\perp}^{\mathrm{opt}} = \left(\frac{\theta}{\langle\mathbf{z}|\mathbf{a}\rangle}\right)\pi^{(0)} = \omega\pi^{(0)} .$$

It can be easily seen how the ratio between the *U-m* self-similarity and the norm of the *U-m* representation with respect to the *core set* in reciprocal space, corrects in this case the rough estimate of the *U-m* property.

b) An alternative restriction case
A variant of the restriction discussed up to now can be easily described. Instead to optimize the scalar product: $\langle\mathbf{z}|\mathbf{x_0}\rangle$, the alternative scalar product: $\langle\mathbf{p_0}|\mathbf{x_0}\rangle$ can

be minimized, so the optimal coefficient will be given by:

$$x^{\text{opt}} = \frac{\langle \mathbf{p}_0 | \mathbf{Z}_0^{-1} | \mathbf{p}_0 \rangle}{\langle \mathbf{p}_0 | \mathbf{Z}_0^{-1} | \mathbf{z} \rangle} \,. \tag{229}$$

c) The simplified $(2+1)$ framework as a naïve application example
The simplified situation $(2+1)$, concerning three molecules, as in the former case of Subsect. "Extended Wave and Density Functions", is simple to solve, for the previous resultant equation in Section 6.1.1, as the involved elements have already been described, when the quadratic error restriction (214) was studied. Under the present orthogonal restriction and using the same notation as the one appearing in the former discussion in Subsect. "Extended Wave and Density Functions", now one can express the estimated property of the U-m as the quotient:

$$\pi^{\text{opt}}_{\perp,AB;U} = z_{UU} \frac{\begin{bmatrix} p_A(z_{BB}z_{AU} - z_{AB}z_{BU}) \\ +p_B(z_{AA}z_{BU} - z_{AB}z_{AU}) \end{bmatrix}}{\begin{bmatrix} z_{AU}(z_{BB}z_{AU} - z_{AB}z_{BU}) \\ +z_{BU}(z_{AA}z_{BU} - z_{AB}z_{AU}) \end{bmatrix}} \,.$$

An equivalent statistical procedure involving the *core set* pairs, as the one described at the end of Subsect. "Extended Wave and Density Functions", can be obviously followed in this case too.

d) The $(n+m)$ framework
The $(n+m)$ algorithm can be also easily set, employing the partition (219) and also Eqs. (220), thus generalizing the algorithm described in Subsect. "Evaluation of Unknown Molecular Properties as Expectation Values". The first equation multiplied by the matrix $\mathbf{Z}_{01}^{\mathrm{T}}$ on the left in both sides of the equality sign, provides:

$$\mathbf{Z}_{01}^{\mathrm{T}} |\mathbf{x}_0\rangle = \mathbf{Z}_{01}^{\mathrm{T}} \mathbf{Z}_{00}^{-1} [|\mathbf{p}_0\rangle - \mathbf{Z}_{01}|\mathbf{x}_1\rangle] \,,$$

and the restriction:

$$\mathbf{Z}_{01}^{\mathrm{T}} |\mathbf{x}_0\rangle = |\mathbf{0}\rangle$$

corresponds to considering the vector $|\mathbf{x}_0\rangle$ as a member of the null space of the matrix $\mathbf{Z}_{01}^{\mathrm{T}}$. Such equality permits, in turn, to write the optimal vector $|\mathbf{x}_1\rangle$ as:

$$|\mathbf{x}^{\text{Opt}}_{\perp;1}\rangle = \left(\mathbf{Z}_{01}^{\mathrm{T}} \mathbf{Z}_{00}^{-1} \mathbf{Z}_{01}\right)^{-1} \left(\mathbf{Z}_{01}^{\mathrm{T}} \mathbf{Z}_{00}^{-1}\right) |\mathbf{p}_0\rangle \,,$$

so, the estimated property vector under the null space restriction simply becomes:

$$|\mathbf{p}^{\text{Opt}}_{\perp;1}\rangle = \mathbf{Z}_{11} |\mathbf{x}^{\text{Opt}}_{\perp;1}\rangle \,,$$

As in the former quadratic error restriction discussed in Subsect. "Expectation Values Within Extended Density Functions Framework", one obtains an expression which shows that this result is nothing else than a linear transformation of the vector of the *core set* properties.

Other Possible Restriction Choices. Besides the previously discussed pair of alternative restriction choices and the one outlined in Section 6.1.2, one can describe other possible procedures to compute the optimal U-m coefficient. They will be briefly explained for a U-m set bearing one element only, as their generalization to $(n+m)$ situations and the $(2+1)$ simplification could be done using the same procedures as before.

a) Quadratic error in reciprocal space vectors: $|\mathbf{x}_0\rangle$ norm restriction
First, one can recall for this purpose Eq. (204), then the *core set* coefficient vector $|\mathbf{x}_0\rangle$ may be expressed with the two basic vectors $\{|\mathbf{p}_0\rangle, |\mathbf{z}_0\rangle\}$ transformed into the reciprocal space: $\{|\mathbf{q}\rangle, |\mathbf{a}\rangle\}$:

$$|\mathbf{x}_0\rangle = |\mathbf{q}\rangle - x|\mathbf{a}\rangle \,. \tag{230}$$

Now, the Euclidean norm of the vector $|\mathbf{x}_0\rangle$ can be optimized, providing the optimal U-m coefficient as:

$$x^{\text{Opt}} = \frac{\langle \mathbf{q}|\mathbf{a}\rangle}{\langle \mathbf{a}|\mathbf{a}\rangle} = \frac{\langle \mathbf{p}_0 | \mathbf{Z}_0^{-2} | \mathbf{z}\rangle}{\langle \mathbf{z} | \mathbf{Z}_0^{-2} | \mathbf{z}\rangle} \,, \tag{231}$$

which is a variant of the form obtained in Eq. (215). The $(n+m)$ case can be easily handled as within the previous discussion on the two described restrictions.

b) Several alternative plausible restrictions within reciprocal space
Finally, another possible restriction set must be described, which can be associated to Eq. (230). Instead of minimizing the norm of the coefficient vector $|\mathbf{x}_0\rangle$ one can minimize either the scalar product $\langle \mathbf{q}|\mathbf{x}_0\rangle$, or work on the alternative form $\langle \mathbf{a}|\mathbf{x}_0\rangle$. The first option provides the optimal U-m coefficient:

$$x^{\text{Opt}} = \frac{\langle \mathbf{q}|\mathbf{q}\rangle}{\langle \mathbf{q}|\mathbf{a}\rangle} \,,$$

which constitutes an expression related to Eq. (229), while the form deduced from the second one is equivalent to Eq. (231).

Some Application Remarks. The first-order FQQSPR equation does not possess immediate predictive power. In order to circumvent this limitation though, one can easily show that two alternative approximate algorithms may

be employed, among other possible similar choices. These procedures can be used to estimate the unknown properties of one or various molecules described as QO.

The present algorithms produce similar formal structures, which can be easily connected with classical QSPR points of view. Such a resemblance can be also simply used to manipulate similar, but empirical, equations in the classical QSPR framework, where the computational formalism appears to be of the same characteristics as in linear QQSPR problems. In order to use the algorithms described here in empirical QSPR cases, there is only need to substitute the QSM, which is the basic matrix in QQSPR procedures, by the Gramian matrix of the molecular descriptor set as defined in Eq. (198), which is the comparable molecular space matrix which can be constructed in classical QSPR. In an indirect manner, therefore, the present study provides an alternative to the widespread QSPR algorithms based on the space descriptor path, a new classical QSPR procedure, which appears, from now on, to be accompanied by a quite diverse toolbox set, common to the linear QQSPR framework, in order to obtain predictions of unknown properties in empirical studies.

Finally, the present results, although exhaustive as far as one can see but without discarding the existence of alternative FQQSPR equation restrictions, from the theoretical point of view they lie on the linear QQSPR framework, they have thus to be considered just as a first step in order to generally solve the prediction problem in QQSPR. This is so as, contrary to classical QSPR procedures, the extension of the FQQSPR equation to higher order terms can be easily described, as well as employed within a set of similar ideas and procedures as these herein discussed.

Extensive numerical results and additional study of high order level problems seems therefore to outline the future research in the open QQSPR area of study.

One Molecule at a Time Linear QQSPR

When constructing the linear QQSPR equation one can choose a system of one core molecule and one U-m, which will constitute the simpler case. The similarity matrix is:

$$\mathbf{Z} = \begin{pmatrix} Z_{II} & Z_{IU} \\ Z_{IU} & Z_{UU} \end{pmatrix} .$$

Where the subindex I stands for any molecule in the core set, that is: a well defined molecular structure with a known property p_I and U for any well-defined molecule with an unknown propertyπ, which has to be estimated. It can be written:

$$\begin{pmatrix} Z_{II} & Z_{IU} \\ Z_{IU} & Z_{UU} \end{pmatrix} \begin{pmatrix} c_I \\ c_U \end{pmatrix} = \begin{pmatrix} p_I \\ \pi \end{pmatrix} \rightarrow \begin{cases} Z_{II}c_I + Z_{IU}c_U = p_I \\ Z_{IU}c_I + Z_{UU}c_U = \pi \end{cases} .$$

Taking the first equation and substituting into the second:

$$c_I = \frac{p - Z_{IU}c_U}{Z_{II}} \rightarrow \pi = Z_{IU}\frac{p_I - Z_{IU}c_U}{Z_{II}} + Z_{UU}c_U = \frac{Z_{IU}}{Z_{II}}p_I + \left(Z_{UU} - \frac{(Z_{IU})^2}{Z_{II}}\right)c_U ,$$

an expression which, after rearrangement, provides a way to estimate the unknown property:

$$\pi = \frac{1}{Z_{II}}\left(Z_{IU}p_I + \left(Z_{II}Z_{UU} - (Z_{IU})^2\right)c_U\right) = \frac{1}{Z_{II}}(Z_{IU}p_I + \Delta c_U) ,$$

where $\Delta = \mathrm{Det}(\mathbf{Z})$.

One can see the undetermined coefficient c_U as equivalent to a parameter λ which in turn can be optimized, thus the unknown property could be rewritten as:

$$\pi = \alpha + \beta\lambda_{\mathrm{opt}} \leftarrow \alpha = \frac{Z_{IU}p_I}{Z_{II}} \wedge \beta = \frac{\Delta}{Z_{II}} .$$

There are several ways to obtain the optimal value of the parameter λ, but all of them are equivalent. For example, one can try to make optimal the coefficient c_I in the first equation:

$$c_I = \frac{1}{Z_{II}}(p_I - Z_{IU}\lambda) \rightarrow \frac{\mathrm{d}}{\mathrm{d}\lambda}\left|\frac{1}{Z_{II}}(p_I - Z_{IU}\lambda)\right|^2 = 0 \rightarrow \lambda_{\mathrm{opt}} = \frac{p_I}{Z_{IU}}$$

so, in this way the optimal U-m property is:

$$\pi_{\mathrm{opt}} = \frac{Z_{IU}p_I}{Z_{II}} + \frac{\Delta p_I}{Z_{II}Z_{IU}} = \frac{(Z_{IU})^2 + \Delta}{Z_{II}Z_{IU}}p_I = \frac{Z_{UU}}{Z_{IU}}p_I .$$

Then the problem consists of obtaining the U-m self-similarity and the similarity between the core molecule and the U-m. So, for every core set C molecular structure one can obtain an estimate of the U-m property, say:

$$\forall I \in C\colon \pi_{U,\mathrm{opt}}[I] = \frac{Z_{UU}}{Z_{IU}}p_I .$$

Then, supposing that the cardinality of the core set is N: $\#C = N$, one can obtain a statistical average of all the core set estimates of the U-m property:

$$\langle \pi_U \rangle \approx N^{-1} \sum_{I=1}^{N} \pi_{\text{opt}}[I] = N^{-1} Z_{UU} \sum_{I=1}^{N} \frac{1}{Z_{IU}} p_I$$
$$= N^{-1} \sum_{I=1}^{N} \omega_{IU} p_I \leftarrow \forall I \in C\colon \omega_{IU} = \frac{Z_{UU}}{Z_{IU}} .$$

An expression which proves that this simple QQSPR formulation arrives at the usual QSPR result, consisting of the fact that the estimated value of the property of the U-m is a weighted sum of the properties of the core set:

$$\langle \pi_U \rangle \approx N^{-1} \sum_{I=1}^{N} \omega_{IU} p_I ,$$

the weights being simply the ratios between the U-m quantum self-similarity and the quantum similarity measure of the U-m with every core set molecular structure.

Practical Considerations. However, in this or other more sophisticated cases, the estimation procedure can be achieved in two steps. First, the elements of the core set can be employed as the U-m ones, one by one in front of the remnant $N-1$ in a kind of Leave One Out procedure. The N optimal estimated values, $\{\langle \pi_I \rangle\}$ say, in this way can be fitted to the experimental property ones, providing in this manner a simple, Hansch-like relationship:

$$p = a \sim \langle \pi \rangle + b ,$$

a relationship which can be further employed to estimate the experimental values of the U-m elements $\{p_U\}$, by using the above-defined equation:

$$p_U = a \langle \pi_U \rangle + b .$$

Moreover, an interesting feature of this procedure is that each estimated value, obtained through solving the fundamental QQSPR equation and irrespective of the fact that the estimation is made over the C or U set elements, can be associated to a mean value, obtained over the set of C-m and also attached to a variance. It is a simple matter of elementary statistical theory application to obtain confidence limits for each estimate, and thus to gather information about, for instance, the outlier nature of some elements and the goodness-of-fit of the whole procedure.

One Molecule at a Time: Quadratic Terms in QQSPR

The operator which can be employed as the source of the fundamental QQSPR equation may be expressed with quadratic and superior terms, within a sequence involving the density elements of the C-m and U-m elements:

$$\Omega(\mathbf{r}) = w_I \rho_I(\mathbf{r}) + w_U \rho_U(\mathbf{r}) + w_I^2 \rho_I^2(\mathbf{r}) + 2 w_I w_U \rho_I(\mathbf{r}) \rho_U(\mathbf{r}) + w_U^2 \rho_U^2(\mathbf{r}) + O(3) .$$

The pair of expectation values of both molecules can be easily written up to third order as:

$$p_I = \langle \Omega \rho_I \rangle = w_I z_{II} + w_U z_{UI} + w_I^2 Z_{III} + 2 w_I w_U Z_{IUI} + w_U^2 Z_{UUI} , \quad (232)$$

and

$$\pi = \langle \Omega \rho_U \rangle = w_I z_{IU} + w_U z_{UU} + w_I^2 Z_{IIU} + 2 w_I w_U Z_{IUU} + w_U^2 Z_{UUU} , \quad (233)$$

where use has been made of the similarity measures like:

$$z_{IU} = \int_D \rho_I(\mathbf{r}) \rho_U(\mathbf{r}) \, \mathrm{d}\mathbf{r} = \int_D \rho_I(\mathbf{r}) \rho_U(\mathbf{r}) \, \mathrm{d}\mathbf{r} = z_{UI} ,$$

and triple similarity measures, for instance:

$$Z_{IUI} = \int_D \rho_I(\mathbf{r}) \rho_U \mathbf{r}) \rho_I(\mathbf{r}) \, \mathrm{d}\mathbf{r} = Z_{IIU} = Z_{UII} = \cdots$$

the properties as expectation values can be rewritten employing the ket-matrix notation:

$$|\mathbf{w}\rangle = \begin{pmatrix} w_I \\ w_U \end{pmatrix} \wedge |\mathbf{z}_I\rangle = \begin{pmatrix} z_{II} \\ z_{UI} \end{pmatrix} \wedge \mathbf{Z}_I = \begin{pmatrix} Z_{III} & Z_{UII} \\ Z_{UII} & Z_{UUI} \end{pmatrix} ,$$

with a similar notation for the ket $|\mathbf{z}_U\rangle$ and the matrix $\mathbf{Z}_U$; the bra notation signifying the corresponding ket transposes. Therefore:

$$\begin{aligned} p_I &= \langle \mathbf{z}_I | \mathbf{w} \rangle + \langle \mathbf{w} | \mathbf{Z}_I | \mathbf{w} \rangle \\ \pi &= \langle \mathbf{z}_U | \mathbf{w} \rangle + \langle \mathbf{w} | \mathbf{Z}_U | \mathbf{w} \rangle . \end{aligned} \quad (234)$$

The QQSPR problem consists of the fact that the coefficient vector is not only unknown $|\mathbf{w}\rangle$ but also the U-m property π. In fact, the quadratic system which corresponds to the quadratic fundamental QQSPR equation in this case, is a set of two different quadratic functions of the same two variables: $|\mathbf{w}\rangle$. The solution may be more complicated than in the linear case, but the procedure can be described in similar terms. That is, first use the C-m equation to express the coefficient w_I in terms of the U-m coefficient and the corresponding similarity measures elements. Then optimize such a coefficient with respect to the U-m one, considered as a parameter. The optimal values of w_U can be employed to evaluate an optimal value of the unknown U-m property π.

A Possible Algorithm. The first fundamental QQSPR equation can be easily transformed into the second-order polynomial root seeking structure.

$$w_I^2 Z_{III} + w_I(z_{II} + 2w_U Z_{IUI}) + (w_U z_{UI} + w_U^2 Z_{UUI}) - p_I = 0\,,$$

which provides:

$$w_I = (2Z_{III})^{-1}\left[-(z_{II} + 2w_U Z_{IUI}) \pm \sqrt{\begin{array}{c}(z_{II} + 2w_U Z_{IUI})^2 - 4Z_{III} \\ \cdot\left((w_U z_{UI} + w_U^2 Z_{UUI}) - p_I\right)\end{array}}\right],$$

and after rearranging the square root part:

$$w_I = (Z_{III})^{-1}\left[-\left(\frac{z_{II}}{2} + w_U Z_{IUI}\right) \pm \sqrt{\begin{array}{c}w_U^2(Z_{IUI}^2 - Z_{III}Z_{UUI}) \\ +w_U(z_{II}Z_{IUI} - z_{UI}Z_{III}) \\ +\left(\left(\frac{z_{II}}{2}\right)^2 + Z_{III}p_I\right)\end{array}}\right],$$

the coefficient w_I appears in terms of w_U and the implied similarity integrals. Also, this expression can be employed in the second fundamental QQSPR equation to obtain the U-m property in terms of only one parameter. As in the linear case, the expression of w_I can be optimized with respect to w_U, which can be considered now as a parameter. The expression to be optimized can be written as:

$$\begin{aligned} w_I &= (Z_{III})^{-1}\left[-(\alpha + w_U\beta) \pm \sqrt{w_U^2\gamma_2 + w_U\gamma_1 + \gamma_0}\right], \\ \alpha &= \frac{z_{II}}{2}\,;\beta = Z_{IUI}\,; \\ \gamma_2 &= Z_{IUI}^2 - Z_{III}Z_{UUI}\,;\gamma_1 = z_{II}Z_{IUI} - z_{UI}Z_{III}\,; \\ \gamma_0 &= \alpha^2 + Z_{III}p_I\,. \end{aligned} \tag{235}$$

Thus, the equation yielding the optimal value of w_U can be easily written as:

$$\begin{aligned} 0 = \frac{\mathrm{d}w_I}{\mathrm{d}w_U} &= -\beta \pm \frac{2w_U\gamma_2 + \gamma_1}{2\sqrt{w_U^2\gamma_2 + w_U\gamma_1 + \gamma_0}} \\ &\to 4\beta^2(w_U^2\gamma_2 + w_U\gamma_1 + \gamma_0) = (2w_U\gamma_2 + \gamma_1)^2 \\ &\to \beta^2(w_U^2\gamma_2 + w_U\gamma_1 + \gamma_0) = \left(w_U\gamma_2 + \frac{\gamma_1}{2}\right)^2 \\ &\qquad = (w_U\gamma_2)^2 + w_U\gamma_2\gamma_1 + \left(\frac{\gamma_1}{2}\right)^2 \\ &\to w_U^2(\beta^2 - \gamma_2)\gamma_2 + w_U(\beta^2 - \gamma_2)\gamma_1 \\ &\qquad + \left(\beta^2\gamma_0 - \left(\frac{\gamma_1}{2}\right)^2\right) = 0\,, \end{aligned}$$

yielding:

$$w_U^{\mathrm{Opt}} = [2\gamma_2]^{-1}\left[-\gamma_1 \pm \sqrt{\gamma_1^2 - 4\gamma_2(\beta^2 - \gamma_2)^{-1}\left(\beta^2\gamma_0 - \left(\frac{\gamma_1}{2}\right)^2\right)}\right], \tag{236}$$

this value permits us to compute w_I^{Opt} by means of Eq. (235) and therefore π^{Opt} can be obtained with Eqs. (235) using the original form (233).

Alternative Unrestricted Variational Algorithm. Starting again from the quadratic Eq. (234), one can vary both parts of the FQQSPR equation:

$$\begin{aligned} p_I &= \langle \mathbf{z}_I|\mathbf{w}\rangle + \langle \mathbf{w}|\mathbf{Z}_I|\mathbf{w}\rangle \to |\mathbf{w}[p_I]\rangle = -\frac{1}{2}\mathbf{Z}_I^{-1}|\mathbf{z}_I\rangle\,, \\ \pi &= \langle \mathbf{z}_U|\mathbf{w}\rangle + \langle \mathbf{w}|\mathbf{Z}_U|\mathbf{w}\rangle \to |\mathbf{w}[\pi]\rangle = -\frac{1}{2}\mathbf{Z}_U^{-1}|\mathbf{z}_U\rangle \end{aligned}$$

so, the optimal estimate values of the C-m and U-m properties will be given by:

$$\begin{aligned} p_I^{\mathrm{est}} &= -\frac{1}{4}\langle \mathbf{z}_I|\mathbf{Z}_I^{-1}|\mathbf{z}_I\rangle\,, \\ \pi^{\mathrm{est}} &= -\frac{1}{4}\langle \mathbf{z}_U|\mathbf{Z}_U^{-1}|\mathbf{z}_U\rangle\,, \end{aligned}$$

which can be associated to minimal values, as the second-order similarity matrices are constructed to be positive definite and thus:

$$\mathrm{Det}\,|\mathbf{Z}_I| = Z_{III}Z_{UUI} - Z_{UII}^2 > 0 \wedge \mathrm{Det}\,|\mathbf{Z}_U| = Z_{UII}Z_{UUU} - Z_{UUI}^2 > 0\,.$$

So, an ultimate procedure could be designed, starting to obtain with every one of the core set elements the following linear equation via a least squares procedure:

$$p = ap^{\mathrm{est}} + b\,,$$

in such a way that the linear equation above provides the possibility to obtain the final estimate of the U-m property value:

$$\pi = a\pi^{\mathrm{est}} + b\,.$$

Bibliography

Primary Literature

1. Carbó R, Arnau J, Leyda L (1980) How similar is a molecule to another? An electron density measure of similarity between two molecular structures. Int J Quantum Chem 17:1185–1189

2. Carbó R, Besalú E (1996) Mendeleev conjecture as a consequence of mendeleev postulates. Afinidad 53:77–79
3. Messiah A (1999) Quantum mechanics. Dover Publications, New York
4. Bohm D (1989) Quantum theory. Dover Publications, New York
5. Carbó-Dorca R, Robert D, Amat L, Girones X, Besalu E (2000) Molecular quantum similarity in QSAR and drug design. In: Lecture notes in chemistry, vol 73. Springer, Berlin
6. Carbó-Dorca R (1995) Molecular similarity and reactivity: From quantum chemical to phenomenological approaches. Kluwer Academic, Dordrecht
7. Carbó-Dorca R, Besalu E (1998) A general survey of molecular quantum similarity. J Mol Struct Theochem 451:11–23
8. Carbó-Dorca R (2000) European congress on computational methods in applied sciences and engineering. Barcelona, pp 1–31
9. Carbó-Dorca R, Besalu E (2000) Quantum theory of QSAR. Contributions Sci 1:399–422
10. Bultinck P, Girones X, Carbó-Dorca R (2005) Molecular quantum similarity: Theory and applications. In: Lipkowitz KB, Larter R, Cundari T (eds) Reviews in computational chemistry, vol 21. Wiley, Hoboken, pp 127–207
11. Carbó-Dorca R, Mezey PG (1998) Advances in molecular similarity. JAI Press, London
12. Carbó-Dorca R (2005) Mathematical elements of quantum electronic density functions. In: Advances in quantum chemistry, vol 49. Elsevier Academic, San Diego, pp 121–208
13. Carbó-Dorca R, Besalu E (2001) Extended Sobolev and Hilbert spaces and approximate stationary solutions for electronic systems within non-linear Schrödinger equation. J Math Chem 29:3–20
14. Carbó-Dorca R, Besalu E (2002) Fundamental quantum QSAR (QQSPR) equation: Extensions, nonlinear terms and generalizations within extended Hilbert–Sobolev spaces. Int J Quantum Chem 88:167–182
15. Johnson MA, Maggiora GM (1990) Concepts and applications of molecular similarity. Wiley, New York
16. Carbó R, Calabuig B, Besalú E, Martínez A (1992) Triple density molecular quantum similarity measures: A general connection between theoretical calculations and experimental results. Mol Eng 2:43–64
17. Carbo R, Calabuig B, Vera L, Besalu E (1994) Molecular quantum similarity: Theoretical framework, ordering principles, and visualization techniques. In: Advances in quantum chemistry, vol 25. Academic Press, San Diego, pp 253–313
18. Carbó-Dorca R (2001) Inward matrix products: Extensions and applications to quantum mechanical foundations of QSAR. J Mol Struct Theochem 537:41–54
19. Carbó-Dorca R, Amat L, Besalu E, Girones X, Robert D (2000) Quantum mechanical origin of QSAR: Theory and applications. J Mol Struct Theochem 504:181–228
20. Carbo R, Besalu E, Amat L, Fradera X (1995) Quantum molecular similarity measures (QMSM) as a natural way leading towards a theoretical foundation of quantitative structure-activity relationships (QSPR). J Math Chem 18:237–246
21. Carbó-Dorca R, Besalu E, Amat L, Fradera X (1996) Quantum molecular similarity measures: Concepts, definitions, and applications to quantitative structure-property relationships. In: Carbó-Dorca R, Mezey PG (eds) Advances in molecular similarity. JAI Press, London, pp 1–42
22. Carbó-Dorca R, Amat L, Besalú E, Gironés X, Robert D (2001) Quantum molecular similarity: Theory and applications to the evaluation of molecular properties, biological activities and toxicity. In: Carbó-Dorca R, Gironés X, Mezey PG (eds) Fundamentals of molecular similarity. Proc of 4th girona seminar on molecular similarity. Kluwer Academic/Plenum Publishers, New York, chap 12, pp. 187–320
23. Carbó-Dorca R (1997) Fuzzy sets and boolean tagged sets. J Math Chem 22:143–147
24. Carbó-Dorca R (1998) Tagged sets, convex sets and QS measures. J Math Chem 23:353–364
25. Carbó-Dorca R (1998) Fuzzy sets and boolean tagged sets; vector semispaces and convex sets; quantum similarity measures and ASA density functions; diagonal vector spaces and quantum chemistry. In: Carbó-Dorca R, Mezey PG (eds) Advances in molecular similarity. JAI Press, London, pp 43–72
26. Robert D, Girones X, Carbó-Dorca R (2000) Quantification of the influence of single point mutations on haloalkane dehalogenase activity: A molecular quantum similarity study. J Chem Inf Comput Sci 40:839–846
27. Robert D, Amat L, Carbó-Dorca R (2000) Quantum similarity QSAR: Study of inhibitors binding to trhombin, trypsin and factor Xa, including a comparision with CoMFA and CoMSIA methods. Int J Quantum Chem 80:265–282
28. Girones X, Gallegos A, Carbó-Dorca R (2000) Modelling antimalarial activity: Application of kinetic energy density quantum similarity measures as descriptors in QSAR. J Chem Inf Comput Sci 40:1400–1407
29. Robert D, Girones X, Carbó-Dorca R (2000) Molecular quantum similarity measures as descriptors for quantum QSAR. Polycycl Aromat Compd 19:51–71
30. Renners I, Ludwig LA, Grauel A, Benfenati E, Pellagatti S, Robert D, Carbó-Dorca R, Girones X (2000) Modeling toxicity with molecular descriptors and similarity measures via B-spline networks. IPMU2000 8th international conference on information processing and management of uncertainty in knowledge based systems, Madrid, Spain, pp 1021–1026
31. Renners I, Carbó-Dorca R, Grauel A, Ludwig LA, Robert D, Girones X (2000) Toxicity prediction by using genetically optimized B-spline networks based on molecular quantum similarity. 2nd international ICSC symposium on neural computation, Berlin, Germany
32. Gallegos A, Robert D, Girones X, Carbó-Dorca R (2001) Structure-toxicity relationships of polycyclic aromatic hydrocarbons using molecular quantum similarity. J Comput Aided Mol Design 15:67–80
33. Besalú E, Gallegos A, Carbó-Dorca R (2001) Topological quantum similarity indices and their use in QSAR: Application to several families of antimalarial compounds. Match-Commun Math Comput Chem 44:41–64
34. Gironés X, Carbó-Dorca R (2001) Sobre les relacions lineals d'energia lliure: Mesures de Semblança Molecular Quàntica sobre Funcions de Densitat Electrònica Modificades. Scientia Gerund 25:5–15
35. Ponec R, Gironés X, Carbó-Dorca R (2002) Molecular basis of LFER. On the nature of inductive effects in aliphatic series. J Chem Inf Comput Sci 42:564–570
36. Gironés X, Gallegos A, Carbó-Dorca R (2001) Antimalarial activity of synthetic 1,2,4-trioxabes and cyclid peroxy ke-

tals, a quantum similarity study. J Comput-Aided Mol Design 15:1053–1063
37. Gironés X, Carbó-Dorca R (2002) Molecular quantum similarity-based QSARs for binding affinities of several steroid sets. J Chem Inf Comput Sci 42:1185–1193
38. Gallegos Saliner AG, Amat L, Carbó-Dorca R, Schultz TW, Cronin MTD (2003) Molecular quantum similarity analysis of estrogenic activity. J Chem Inf Comput Sci 43:1166–1176
39. Amat L, Carbó-Dorca R, Cooper DL, Allan NL, Ponec R (2003) Structure-property relationships and momentum-space quantities: Hammett σ-constants. Mol Phys 101:3159–3162
40. Gironés X, Carbó-Dorca R, Ponec R (2003) Molecular basis of LFER. Modeling of the electronic substituent effect using fragment quantum self-similarity measures. J Chem Inf Comput Sci 43:2033–2038
41. Nino A, Munoz-Caro C, Carbó-Dorca R, Gironés X (2003) Rational modelling of the voltage-dependent K+ channel inactivation by aminopyridines. Biophys Chem 104:417–427
42. Gallegos A, Carbó-Dorca R, Ponec R, Waisser K (2004) Similarity approach to QSAR. Application to antimycobacterial benzoxazines. Int J Pharm 269:51–60
43. Gironés X, Carbó-Dorca R (2004) Molecular similarity and quantitative structure-activity relationships. In: Bultinck P, De Winter H, Langenaecker W, Tollenaere JP (eds) Similarity and quantitative structure-activity relationships in computational medicinal chemistry for drug discovery. Marcel Dekker, New York, pp 365–385
44. Giralt F, Espinosa G, Arenas A, Ferre-Gine J, Amat L, Gironés X, Carbó-Dorca R, Cohen Y (2004) Estimation of infinite dilution activity coefficients of organic compounds in Water with neural classifiers. AIChE J 50:1315–1343
45. Carbó-Dorca R, Gironés X (2005) Foundation of quantum similarity measures and their relationship to QSPR: Density function structure, approximations and application examples. Int J Quantum Chem 101:8–20
46. Gironés X, Carbó-Dorca R (2006) Modelling toxicity using molecular quantum similarity measures. QSAR Comb Sci 25:579–589
47. Myers H (1990) Classical and modern regression with applications. Duxbury Press, California
48. Wold S, Sjöström M, Eriksson L (1994) Partial least squares projections to latent structures (PLS) in chemistry. In: von Ragué Schleyer P, Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer HF III, Screiner PR (eds) Encyclopedia of computational chemistry. Wiley, Chichester, pp 2006–2021
49. Geladi P, Kowalski BR (1986) Partial least-squares regression: A tutorial. Analytica Chimica Acta 185:1–17
50. Wold S, Johansson E, Cocchi M (1993) PLS-partial least-squares projections to latent structures. In: Kubinyi H (ed) 3D QSAR in drug design. ESCOM, Science Publishers BV, Leiden, chap 5
51. Montgomery DC, Peck EA, Vining GG (1992) Introduction to linear regression analysis. Wiley, New York
52. Whitley DC, Ford MG, Livingstone DJ (2000) J Chem Inf Comput Sci 40:1160–1168
53. Wold S (1978) Cross-validatory estimation of the number of components in factor and principal component models. Technometrics 20:397–405
54. Jolliffe IT (1986) Principal component analysis. Springer, New York
55. Cattell RB (1966) The scree test for the number of factors. Multivar Behav Res 1:245–276
56. Wold S, Eriksson L (1995) Statistical validation of QSAR results. In: Van de Waterbeemd H (ed) Chemometric methods in molecular design. VCH, New York, sect IV, pp 309–318
57. Allen DM (1974) The relationship between variable selection and data augmentation and a method for prediction. Technometrics 16:125–127
58. Shao J (1993) Linear-Model selection by cross-validation. J Am Stat Assoc 88:486–494
59. Carbó-Dorca R (2000) Stochastic transformation of quantum similarity matrices and their use in quantum QSAR models. Int J Quantum Chem 79:163–177
60. Carbó-Dorca R (2007) About the prediction of molecular properties using the fundamental quantum QSPR (QQSPR) equation. SAR QSAR Environ Res 18:265–284
61. Carbó-Dorca R, Van Damme S (2007) Theoretical chemistry accounts: Theory, computation, and modeling. Theoretica Chimica Acta 118:673–679
62. Carbó-Dorca R, Van Damme S (2007) Riemann spaces, molecular density function semispaces, quantum similarity measures and quantum quantitative structure-properties relationships (QQSPR). Afinidad 64:147–153
63. Hansch C, Fujita T (1962) J Am Chem Soc 86:10
64. Oliva JM, Carbó-Dorca R, Mestres J (1996) Conformational analysis from the point of view of quantum molecular similarity. In: Carbó-Dorca R, Mezey PG (eds) Advances in molecular similarity. JAI PRESS, London, pp 135–165
65. Carbó-Dorca R (2000) Quantum quantitative structure-activity relationships (QQSAR): A comprehensive discussion based on inward matrix products, employed as a tool to find approximate solutions of strictly positive linear systems and providings QSAR-quantum similarity measures connection. In: Proc of the european congress on computational methods in applied sciences and engineering (ECCOMAS 2000), Barcelona, chap 12
66. Sen K, Carbó-Dorca R (2000) Inward matrix products, generalised density functions and Rayleigh–Schrödinger perturbation Theory. J Mol Struct Theochem 501:173–176
67. Carbó-Dorca R (2001) Inward matrix product algebra and calculus as tools to construct space-time frames of arbitrary dimensions. J Math Chem 30:227–245
68. Carbó-Dorca R (2003) Applications of inward matrix products and matrix wave functions to Hückel MO theory, Slater extended wave functions, spin extended functions and Hartree method. Int J Quantum Chem 91:607–617
69. Carbó-Dorca R (2003) About some questions relative to the arbitrariness of signs: Their possible consequences in matrix signatures definition and quantum chemical applications. J Math Chem 33:227–244
70. Vinogradov IM (1989) Encyclopaedia of mathematics, vol 4. Kluwer Academic, Dordrecht, p 351
71. Lahey Computer Systems (1998) LF 95 Language Reference. Lahey Computer Systems, Incline Village http://www.lahey.com
72. Carbó-Dorca R (2002) Shell partition and metric semispaces: Minkowski norms, root scalar products, distances and cosines of arbitrary order. J Math Chem 32:201–223
73. Bultinck P, Carbó-Dorca R (2004) A mathematical discussion on density and shape functions, vector semispaces and related questions. J Math Chem 36:191–200

74. Vinogradov IM (1992) Encyclopaedia of mathematics, vol 8. Kluwer Academic, Dordrecht, p 249
75. Constans P, Amat LL, Fradera X, Carbó-Dorca R (1996) Quantum molecular similarity measures (QMSM) and the atomic shell approximation (ASA). In: Carbó-Dorca R, Mezey PG (eds) Advances in molecular similarity, vol 1. JAI Press, London, chap 8, pp 187–211
76. Gironés X, Amat L, Carbó-Dorca R (1998) A comparative study of isodensity surfaces using ab initio and ASA density functions. J Mol Graph Model 16:190–196
77. Constants P, Carbó R (1995) Atomic shell approximation: Electron density fitting algorithm restricting coefficients to positive values. J Chem Inf Comput Sci 35:1046–1053
78. Amat LL, Carbó-Dorca R (1997) Quantum similarity measures under atomic shell approximation: First order density fitting using elementary Jacobi rotations. J Comput Chem 18:2023–2039
79. Amat LL, Carbó-Dorca R (1999) Fitted electronic density functions from H to Rn for use in quantum similarity measures: Cis-diamminedichloroplatinum(II) complex as an application example. J Comput Chem 20:911–920
80. Amat LL, Carbó-Dorca R (2000) Molecular electronic density fitting using elementary Jacoby rotations under atomic shell approximation (ASA). J Chem Inf Comput Sci 40:1188–1198
81. Gironés X, Carbó-Dorca R, Mezey PG (2001) Application of promolecular ASA densities to graphical representation of density functions of macromolecular systems. J Mol Graph Model 19:343–348
82. Mestres J, Solà M, Besalú E, Duran M, Carbó R (1995) Electron density approximations for the fast evaluation of quantum molecular similarity measures. In: Carbó R (ed) Molecular similarity and reactivity: From quantum chemical to phenomenological approaches. Kluwer Academic, Dordrecht, pp 77–85
83. Mestres J, Solà M, Duran M, Carbó R (1995) General suggestions and applications of quantum molecular similarity measures from ab initio fitted electron densities. In: Carbó R (ed) Molecular similarity and reactivity: From quantum chemical to phenomenological approaches. Kluwer Academic, Dordrecht, pp 89–111
84. Carbó R, Calabuig B (1990) Molecular similarity and quantum chemistry. In: Johnson MA, Maggiora GM (eds) Concepts and applications of molecular similarity. Wiley, New York, chapt 6, p 147
85. Carbó R, Besalú E, Amat LL, Fradera X (1996) On Quantum Molecular Similarity Measures (QMSM) and Indices (QMSI). J Math Chem 19:47–56
86. Robert D, Carbó-Dorca R (2000) General trends in atomic and nuclear quantum similarity measures. Int J Quantum Chem 77:685–692
87. Besalú E, Carbó-Dorca R, Karwowski J (2001) Generalized one-electron spin functions and Self-Similarity Measures. J Math Chem 29:41–45
88. Amat L, Besalú E, Carbó-Dorca R (2001) Identification of active molecular sites using quantum-self similarity measures. J Chem Inf Comput Sci 41:978–991
89. Besalú E, Gironés X, Amat L, Carbó-Dorca R (2002) Molecular quantum similarity and the fundaments of QSAR. Acc Chem Res 35:289–295
90. Amat L, Carbó-Dorca R, Cooper DL, Allan NL (2003) Classification of reaction pathways via momentum-space and quantum molecular similarity measures. Chem Phys Lett 367:207–213
91. Bultinck P, Carbó-Dorca R (2003) Molecular quantum similarity matrix based clustering of molecules using dendrograms. J Chem Inf Comput Sci 43:170–177
92. Ponec R, Bultinck P, Van Damme S, Carbó-Dorca R, Tantillo DJ (2005) Geometric and electronic similarities between transition structures for electrocyclizations and sigmatropic hydrogen shifts. Theor Chem Acc 113:205–211
93. Bultinck P, Carbó-Dorca R (2005) Molecular quantum similarity using conceptual DFT descriptors. J Chem Sci 117:425–435
94. Carbó-Dorca R (2008) A quantum similarity matrix (QSM) Aufbau procedure. J Math Chem 44:228–234
95. Berberian SK (1961) Introduction to Hilbert space. Oxford University Press, New York
96. Carbó R, Besalú E (1992) Many center AO integral evaluation using cartesian exponential type orbitals (CETO's). Adv Quantum Chem 24:115–237
97. von Neumann J (1955) Mathematical foundations of quantum mechanics. Princeton University Press, Princeton
98. Born M (1945) Atomic physics. Blackie and Son, London
99. Dirac PAM (1983) The principles of quantum mechanics. Clarendon Press, Oxford
100. Carbó-Dorca R (2000) Quantum QSAR and the eignsystems of stochastic quantum similarity matrices. J Math Chem 27:357–376
101. Gironés X, Amat L, Carbó-Dorca R (2000) Use of electron-electron repulsion energy as a molecular description in QSAR or QSPR studies. J Comput Aided Mol Design 14:477–485
102. Kubinyi H (1993) 3D QSAR in drug design. Theory, methods and applications. ESCOM Science Publishers, Leiden
103. Martin YC (1978) Quantitative drug design. Medicinal research series, vol 8. Marcel Dekker, New York
104. Lien EJ (1987) SAR, side effects and drug design. Medicinal research series, vol 11. Marcel Dekker, New York
105. Martin YC, Kutter E, Austel V (1989) Modern drug research. Medicinal research series, vol 12. Marcel Dekker, New York
106. Wermuth CG (1993) Trends in QSAR molecular modelling 92. Escom, Leiden
107. Kubinyi H (1998) Quantitative structure-activity relationships in drug design. In: Schleyer PVR, Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer HF III, Schreiner PR (eds) Encyclopedia of computational chemistry, vol 4. Wiley, Chichester, pp 2309–2319
108. Charton M (1996) Advances in quantitative structure-property relationships, vol 1. JAI Press, London
109. Van de Waterbeemd H (1996) Structure-property correlations in drug research. Academic Press, San Diego
110. Jurs PC (1998) Quantitative structure-property relationships (QSPR). In: Schleyer PVR, Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer HF III, Schreiner PR (eds) Encyclopedia of computational chemistry, vol 4. Wiley, Chichester, pp 2320–2330
111. Boethling RS, Mackay D (2000) Handbook of property estimation methods for chemicals. Environmental and health sciences. Lewis Publishers, London
112. Neter J, Wasserman W, Kutner MH (1990) Applied linear statistical models. RD Irwin, Boston
113. Wagner M, Sadowski J, Gasteiger J (1995) Autocorrelation of molecular surface properties for modeling corticosteroid

binding globulin and cytosolic Ah receptor activity by neural networks. J Am Chem Soc 117:7769–7775
114. Cramer RD, Patterson DE, Bunce JD (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 110:5959–5967
115. Parretti MF, Kroemer RT, Rothman JH, Richards WG (1997) Alignment of molecules by Monte Carlo optimisation and molecular similarity indices. J Comput Chem 18:1344–1353
116. So SS, Karplus M (1997) Three-dimensional quantitative structure-activity relationships from molecular similarity matrices and genetic neural networks. 1. Method and validations. J Med Chem 40:4347–4359
117. Jain AN, Koile K, Chapman D (1994) Compass: Predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. J Med Chem 37:2315–2327
118. Bravi G, Gancia E, Mascagni P, Pegna M, Todeschini R, Zaliani A (1997) MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids. J Comput Aided Mol Design 11:79–92
119. Norinder U (1990) Experimental design based 3D-QSAR analysis of steroid-protein interactions: Application to human CBG complexes. J Comput Aided Mol Design 4:381–389
120. Norinder U (1991) 3D-QSAR analysis of steroid/protein interactions: The use of difference maps. J Comput Aided Mol Design 5:419–426
121. Rum G, Herndon WC (1991) Molecular similarity concepts. 5. Analysis of steroid-protein binding constants. J Am Chem Soc 113:9055–9060
122. Simon Z, Bohl M (1992) Structure-activity relations in gestagenic steroids by the MTD method. The case of hard molecules and soft receptors. Quant Struct Activity Relatsh 11:23–28
123. Waszkowycz B, Clark DE, Frenkel D, Li J, Murray CW, Robson B, Westhead DR (1994) PRO_LIGAND: An approach to de novo molecular design. 2. Design of novel molecules from molecular field analysis (MFA) models and pharmacophores. J Med Chem 37:3994–4002
124. Dunn WJ, Wold S, Edlund U, Hellberg S (1984) Multivariate structure-activity relationships between data from a battery of biological tests and an ensemble of structure descriptors: the PLS method. Quant Struct Activity Relatsh 3: 131–137
125. Good AC, So SS, Richards WG (1993) Structure-activity relationships from molecular similarity matrices. J Med Chem 36:433–438
126. Oprea TI, Ciubotariu D, Sulea TI, Simon Z (1993) Comparison of the minimal steric difference (MTD) and comparative molecular field analysis (CoMFA) methods for analysis of binding of steroids to carrier proteins. Quant Struct Activity Relatsh 12:21–26
127. Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 37:4130–4146
128. Hahn M, Rogers D (1995) Receptor Surface Models. 2. Application to quantitative structure-activity relationships studies. J Med Chem 38:2091–2102
129. Silverman BD, Platt DE (1996) Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. J Med Chem 39:2129–2140
130. Kellogg GE, Kier LB, Gaillard P, Hall LH (1996) E-state fields: Applications to 3D QSAR. J Comput Aided Mol Design 10:513–520
131. Anzali S, Barnickel G, Krug M, Sadowski J, Wagener M, Gasteiger J, Polanski J (1996) The comparison of geometric and electronic properties of molecular surfaces by neural networks: Application to the analysis of corticosteroid-binding globulin activity of steroids. J Comput Aided Mol Design 10:521–534
132. Norinder U (1996) 3D-QSAR Investigation of the tripos benchmark steroids and some protein-tyrosene kinase inhibitors of styrene type using the TDQ approach. J Chemom 10:533–545
133. Schnitker J, Gopalaswamy R, Crippen GM (1997) Objective models for steroid binding sites of human globulins. J Comput Aided Mol Design 11:93–110
134. Turner DB, Willett P, Ferguson AM, Heritage T (1997) Evaluation of a novel infrared range vibration-based descriptor (EVA) for QSAR studies. 1. General application. J Comput Aided Mol Design 11:409–422
135. Cramer RD III, Depriest SA, Patterson DE, Hecht P (1993) The developing practice of comparative molecular field analysis. In: Kubinyi H (ed) 3D QSAR in drug design. ESCOM, Leiden, part III, pp 443–485
136. Good AC (1995) 3D molecular similarity indices and their application in QSAR studies. In: Dean PM (ed) Molecular similarity in drug design. Blackie Academic & Professional (Capman & Hall), London, chap 2
137. Ortiz AR, Pisabarro MT, Gago F, Wade RC (1995) Prediction of drug binding affinities by comparative binding energy analysis: Application to human synovial fluid phospholipase A2 inhibitiors. In: Sanz F, Giraldo J, Manaut F (eds) QSAR and molecular modelling: Concepts, computational tools and biological applications. Prous Pub, Barcelona, sect VII, pp 439–443
138. Oprea TI, Head RD, Marshall GR (1995) The basis of cross-reactivity for a series of steroids binding to monoclonal antibody (DB3) against progesterone. A molecular modeling and QSAR study. In: Sanz F, Giraldo J, Manaut F (eds) QSAR and molecular modelling: Concepts, computational tools and biological applications. Prous Pub, Barcelona, sect VII, pp 461–462
139. Kimura T, Hasegawa K, Funatsu K (1998) GA strategy for variable selection in QSAR studies: GA-based region selection for CoMFA modeling. Chem Inf Comput Sci 38:276–282
140. Mabilia M, Belvisi L, Bravi G, Catalano G, Scolastico C (1995) A PCA/PLS analysis on nonpeptide angiotensin II receptor antagonists. In: Sanz F, Giraldo J, Manaut F (eds) QSAR and molecular modelling: Concepts, computational tools and biological applications. Prous Pub, Barcelona, sect VII, pp 456–460
141. Sulea T, Oprea T, Muresan S, Ling Chan S (1997) A different method for steric field evaluation in CoMFA improves model robustness. Chem Inf Comput Sci 37:1162–1170
142. Palomer A, Giolitti A, García ML, Fos E, Cabré F, Mauleón D, Carganico G (1995) Molecular modeling and CoMFA investigations on LTD4 receptor antagonists. In: Sanz F, Giraldo J, Manaut F (eds) QSAR and molecular modelling: Concepts, computational tools and biological applications. Prous Pub, Barcelona, sect VII, pp 444–450

143. Tetko IV, Villa AEP, Livingstone DJ (1996) Neural network studies. 2. variable selection. Chem Inf Comput Sci 36: 794–803
144. Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 37:4130–4146
145. Richon AB, Young SS (2008) An introduction to QSAR methodology. A web paper of the network science corporation: http://www.netsci.org/Science/Compchem/feature19.html. Accessed 27 Oct 2008
146. So SS, van Helden SP, van Geerestein VJ, Karplus M (2000) Quantitative structure-activity relationship studies of progesterone receptor binding steroids. Chem Inf Comput Sci 40:762–772
147. Gantmacher F R (1966) Théorie des matrices, vol 2. Dunod, Paris
148. Jacobi CGJ (1846) J Reine Angew Math 30:51–94
149. Carbó-Dorca R, Besalú E, Gironés X (2000) Extended density functions. Adv Quantum Chem 38:3–63
150. Eyring H, Walter J, Kimball GE (1944) Quantum chemistry. Wiley, New York
151. Hoffmann K (1988) Banach spaces and analytic functions. Dover Publications, New York
152. Akhiezer NI, Glazman IM (1993) Theory of linear operators in Hilbert space. Dover Publications, New York
153. Landau LD, Lifshitz EM (1967) Mecánica Cuántica No-Relativista. Editorial Reverté, Barcelona
154. Vinogradov IM (1992) Encyclopaedia of mathematics, vol 8. Reidel-Kluwer Academic, Dordrecht, pp 379
155. Stewart J (1991) Advanced General Relativity. Cambridge University Press, Cambridge
156. Bach A, Amat L, Besalú E, Carbó-Dorca R, Ponec R (2000) Quantum chemistry, Sobolev spaces and SCF. J Math Chem 28:59–70
157. Carbó-Dorca R (2002) Density functions and generating wave functions. In: Sen K (ed) Reviews in modern quantum chemistry, vol I, 15, pp 401–412
158. Edwards CH Jr (1994) Advanced calculus of several variables. Dover Publications, New York
159. Carbó R, Besalú E (1996) Applications of nested summation symbols to quantum chemistry: Formalism and programming techniques. In: Ellinger Y, Defranceschi M (eds) Strategies and applications in quantum chemistry. Kluwer Academic, Dordrecht, pp 229
160. Carbó R, Besalú E (1995) Definition and quantum chemical applications of nested summation symbols and logical functions: Pedagogical artificial intelligence devices for formulae writing, sequential programming and automatic parallel implementation. J Math Chem 18:37–72
161. Breit TG; Phys. Revs. 34 (1929) 553–573; Ibid. 36 (1930) 383–397; Ibid. 39 (1932) 616–624
162. Bethe HA, Salpeter EE (1957) Quantum mechanics of one- and two-electron systems. Springer, Berlin
163. Moss RE (1973) Advanced molecular quantum mechanics. Chapman and Hall, London
164. Carbó-Dorca R, Karwowski J (2001) Theoretical and computational aspects of extended wave functions. Int J Quantum Chem 84:331–337
165. McWeeny R, Sutcliffe BT (1969) Methods of molecular quantum mechanics. Academic Press, London
166. Huzinaga S (1991) Theochem 234:51–73
167. Huzinaga S, Katsuki S, Matsuoka O (1992) J Math Chem 10:25–39
168. Bonifacic V, Huzinaga S (1974) J Chem Phys 60:2779–2786
169. Carbó R, Calabuig B (1992) Molecular quantum similarity measures and n-dimensional representation of quantum objects I. theoretical foundations. Int J Quantum Chem 42:1681–1709
170. Carbó-Dorca R, Besalú E (1996) Extending molecular similarity to electronic energy surfaces: Boltzmann similarity measures and indices. J Math Chem 20:247–261
171. Bultinck P, De Winter H, Langenaecker W, Tollenaere JP (2004) Similarity and quantitative structure-activity relationships in computational medicinal chemistry for drug discovery. Marcel Dekker, New York
172. Vinogradov IM (1988) Encyclopaedia of mathematics, vol 1. Reidel-Kluwer Academic, Dordrecht, p 334
173. Korn GA, Korn TM (1967) Manual of mathematics. Mc Graw-Hill, New York
174. Stewart J (1999) Cálculo. International Thompson Editores, México
175. Carbó R, Besalú E (1993) J Math Chem 13:331–342
176. Carbó R, Besalú E (1994) Comput Chem 18:117–126
177. Carbó-Dorca R (2000) A discussion on an apparent MO theory paradox. J Math Chem 27:35–41
178. Fraga S, García de la Vega JM, Fraga ES (1999) The Schrödinger and Riccati equations. Lecture notes in chemistry, vol 70. Springer, Berlin
179. Meyer CD (2000) Matrix analysis and applied linear algebra. SIAM, Philadelphia
180. Amat LL, Carbó-Dorca R, Ponec R (1999) Simple linear QSAR models based on quantum similarity measures. J Med Chem 42:5169–5170
181. Besalú E, Carbó R, Mestres J, Solà M (1995) In: Sen K (ed) Molecular similarity I. Top Curr Chem 173:31
182. Hansch C, Fujita T (1964) J Am Chem Soc 86:5175
183. Amat LL, Carbó-Dorca R, Ponec R (1998) Molecular quantum similarity studies as an alternative to log P values in QSAR studies. J Comput Chem 14:1575–1583
184. Carbó R, Besalú E (1994) Comput Chem 18:117

Books and Reviews

Al-Fahemi J, Cooper DL, Allan NL (2005) The quantitative use of momentum-space descriptors. Chem Phys Lett 416:376–380
Al-Fahemi J, Cooper DL, Allan NL (2005) The use of momentum-space descriptors for predicting octanol-water partition coefficients. J Mol Struct Theochem 727:57–61
Allan NL, Cooper DL (1995) Topics Curr Chem 173:85–111
Amat LL, Carbó-Dorca R (2002) Use of promolecular ASA density functions as a general algorithm to obtain starting MO in SCF calculations. Int J Quantum Chem 87:59–67
Amat LL, Besalú E, Carbó R, Fradera X (1995) Practical applications of quantum molecular similarity measures (QMSM): Programs and examples. Scientia Gerund 21:17–34
Amat LL, Constans P, Carbó R (1996) Descripció d'un algorisme d'optimització global de les Mesures de Semblança Quàntica Molecular. Scientia Gerund 22:109–121
Amat LL, Pradera X, Carbó R (1996) Sobre els mapes de Semblança Quàntica Molecular. Scientia Gerund 22:97–107

Amat LL, Robert D, Besalú E, Carbó–Dorca R (1998) Molecular quantum similarity measures tuned QSAR: An antitumoral family validation study. J Chem Inf Comput Sci 38:624–631
Amovilli C, McWeeny R (1991) J Mol Struct Theochem 227:1–9
Antolín J, Angulo JC (2008) Quantum similarity indices for atomic ionization processes. Eur Phys J D 46:21–26
Atkins PW, Friedman RS (1997) Molecular quantum mechanics. Oxford University Press, Oxford
Bach A, Carbó-Dorca R (1999) Aplicació de la Semblança Molecular Quàntica en la reducció de l'espai configuracional per a l'estat fonamental i primers exitats de l'àtom d'heli. Scientia Gerund 24:183–196
Bader RFW (1990) Atoms in molecules. Clarendon Press, Oxford
Bell JS (1993) Speakable and unspeakable in quantum mechanics. Cambridge University Press, Cambridge
Benigni R, Cotta-Ramusino M, Giorgi F, Gallo G (1995) J Med Chem 38:629–635
Benigni R, Cotta-Rasmusino M, Gallo G, Giorgi F, Giuliani A, Vari MR (2000) J Med Chem 43:3699–3707
Besalú E (2001) Fast computation of cross-validated properties in full linear leave-many-out procedures. J Math Chem 29:191–204
Besalú E, Carbó R (1995) Quantum similarity topological indices: Definition, analysis and comparison with classical molecular topological indices. Scientia Gerund 21:145–152
Besalú E, Amat L, Fradera X, Carbó R (1995) An application of the molecular quantum similarity: Ordering of some properties of the hexanes. In: QSAR and molecular modelling: Concepts, computational tools and biological applications. Proc of the 10th european symposium on structure-activity relationships. Prous Science, Barcelona, pp 396–399
Besalú E, Carbó R, Mestres J, Solà M (1995) Foundations and recent developments of quantum molecular similarity. In: Topics in current chemistry: Molecular similarity I, vol 173. Springer, Berlin, pp 31–62
Besalú E, Carbó R, Duran M, Mestres J (1995) MESSEM: A density-based quantum molecular similarity system of programs. In: Clementi E (ed) Methods and techniques in computational chemistry METECC-95. STEF, Cagliari, pp 491–508
Bethe HA, Jackiw R (1986) Intermediate quantum mechanics. Benjamin, Menlo Park
Bonaccorsi R, Scrocco E, Tomasi J (1970) J Chem Phys 52:5270–5284
Boon G, Van Alsenoy C, De Proft F, Bultinck P, Geerlings P (2005) Molecular quantum similarity of enantiomers of amino acids: a case study. J Mol Struct Theochem 727:49–56
Borgoo A, Godefroid M, Sen KD, De Proft F, Geerlings P (2004) Quantum similarity of atoms: A numerical Hartree–Fock and information theory approach. Chem Phys Lett 399:363–367
Borgoo A, Torrent-Sucarrat M, De Proft F, Geerlings P (2007) Quantum similarity study of atoms: A bridge between hardness and similarity indices. J Chem Phys 126:234104
Born M (1944) Atomic physics. Blackie & Son, London
Botella V, Pacios LF (1998) Analytic atomic electron densities in molecular se lf-similarity measures and electrostatic potentials. J Mol Struct Theochem 426:75–85
Bultinck P, Carbó-Dorca R, Van Alsenoy C (2003) Quality of approximate electron densities and internal consistency of molecular alignment algorithms in molecular quantum similarity. J Chem Inf Comput Sci 43:1208–1217
Bultinck P, Rafat M, Ponec R, Van Gheluwe B, Carbó-Dorca R, Popelier P (2006) Coulomb and overlap self-similarities: A comparative selectivity analysis of structure- electron delocalization and aromaticity in linear polyacenes: Atoms in molecules multicenter delocalization ind. J Phys Chem A 110:7642–7648
Bultinck P, Ponec R, Gallegos A, Fias S, Van Damme S, Carbó-Dorca R (2006) Generalized Polansky index as an aromaticity measure in polycyclic aromatic hydrocarbons. Croatica Chemica Acta 79: 363–371
Bunge M (1979) Causality in modern science. Dover, New York
Burt C, Richards WG, Huxley PH (1990) J Comput Chem 11:1139–1146
Carbó R (1995) Molecular similarity and reactivity: From quantum chemical to phenomenological approaches. Understanding chemical reactivity, vol 14. Kluwer Academic, Dordrecht
Carbó R, Besalú E (1995) Theoretical Foundations of Quantum Molecular Similarity. In: Carbó R (ed) Molecular similarity and reactivity: From quantum chemistry to phenomenological approaches. Kluwer Academic, Dordrecht, pp 3–30
Carbó R, Calabuig B (1989) Molsimil-88: Molecular similarity calculations using a CNDO approximation. Comput Phys Commun 55:117
Carbó R, Calabuig B (1989) Sobre las Medidas de Semejanza Molecular: Una conexión entre Química Quántica e Inteligencia Artificial. Una contribucion a los Anales de la VI Escuela Latinoamericana de Química Teórica, Brasil, vol 1, pp 134
Carbó R, Calabuig B (1992) Molecular quantum similarity measures and n-dimensional representation of quantum objects II. Practical applications. Int J Quantum Chem 42:1695
Carbó R, Calabuig B (1992) Quantum molecular similarity measures and the n-dimensional representation of a molecular set: Phenyldimethylthiazines. J Mol Struct Theochem 254:517–531
Carbó R, Calabuig B (1992) Quantum similarity measures, molecular cloud description and structure-properties relationships. J Chem Inf Comput Sci 32:600–606
Carbó R, Calabuig B (1992) Quantum similarity: Definitions, computational details and applications. In: Fraga S (ed) Computational chemistry: Structure, interactions and reactivity. Elsevier, Amsterdam, pp 300–324
Carbó R, Domingo LL (1987) LCAO MO similarity measures and taxonomy. Int J Quantum Chem 32:517–545
Carbó R, Riera JM (1978) A general SCF theory. Lecture notes in chemistry, vol 5. Springer, Berlin
Carbó R, Martín M, Pons V (1977) Application of quantum mechanical parameters in quantitative structure-activity relationships, Afinidad 34:348–353
Carbó R, Suñé E, Lapeña F, Pérez J (1986) Electrostatic potential comparison and molecular metric spaces. J Biol Phys 14:21
Carbó R, Lapeña F, Suñé E (1986) Similarity measures on electrostatic molecular potentials. Afinidad 43:483
Carbó R, Calabuig B, Martínez A (1991) Semblança molecular: Representació n-dimensional d'un conjunt de molècules. Scientia Gerund 17:133
Carbó R, Molino L, Calabuig B, Besalú E, Martínez A (1992) Medidas de Semejanza Cuántica: Conceptos Clásicos y Nuevas Estrategias. Folia Chimica Theoretica Latina 21–45
Carbó-Dorca R (1998) On the statistical interpretation of density functions: ASA, convex sets, discrete quantum chemical molecular representations, diagonal vector spaces and related problems. J Math Chem 23:365–375
Carbó-Dorca R (2000) Quantum QSAR and the eigensystems

of stochastic quantum similarity matrices. J Math Chem 27:357–376
Carbó-Dorca R (2004) Heisenberg's relations in discrete n-dimensional parameterized metric vector spaces. J Math Chem 36:41–54
Carbó-Dorca R (2004) Infinite-dimensional time vectors as background building blocks of a space-time frame structure. J Math Chem 36:75–81
Carbó-Dorca R (2005) Molecular nuclear fields: A naïve perspective. J Math Chem 38:671–676
Carbó-Dorca R (2004) Non-linear terms & variational approach in quantum QSPR. J Math Chem 36:241–260
Carbó-Dorca R (2006) Descriptors and probability distributions in MO theory: Weighted Mulliken matrices and molecular quantum similarity measures. J Math Chem 39:551–591
Carbó-Dorca R (2008) Mathematical aspects of the LCAO MO first order density function (1): Atomic partition, metric structure and practical applications. J Math Chem 43:1076–1101
Carbó-Dorca R (2008) Mathematical aspects of the LCAO MO first order density function (2): Relationships between density functions. J Math Chem 43:1102–1118
Carbó-Dorca R, Besalú E (2006) Generation of molecular fields, quantum similarity measures and related questions. J Math Chem 39:495–509
Carbó-Dorca R, Bultinck P (2004) A general procedure to obtain quantum mechanical charge and bond order molecular parameters. J Math Chem 36(3):201–210
Carbó-Dorca R, Bultinck P (2004) Quantum mechanical basis for Mulliken population analysis. J Math Chem 36:231–239
Carbó-Dorca R, Bultinck P (2008) Mathematical aspects of the LCAO MO first order density function (3): A general localization procedure. J Math Chem 43:1069–1075
Chaves J, Barroso J, Bultinck P, Carbó-Dorca R (2006) Toward an alternative hardness kernel matrix structure in the electronegativity equalization method (EEM). J Chem Inf Model 46:1657–1665
Cioslowski J, Challacombe M (1991) Int J Quantum Chem S25:81–93
Cioslowski J, Fleishmann ED (1991) J Am Chem Soc 113:64–67
Cioslowski J, Mixon ST (1992) Can J Chem 70:443–449
Cioslowski J, Nanayakkara A (1993) J Am Chem Soc 115:11213–11215
Cioslowski J, Stefanov BB, Constans P (1996) J Comput Chem 17:1352–1358
Constans P, Amat L, Carbó-Dorca R (1997) Towards a global maximization of the molecular similarity function: The superposition of two molecules. J Comput Chem 18:826–846
Cooper DL, Allan NL (1989) J Comput Aided Mol Design 3:253–259
Cooper DL, Allan NL (1992) J Am Chem Soc 114:4773–4776
Cooper DL, Mort KA, Allan NL, Kinchington D, McGuidan CH (1993) J Am Chem Soc 115:12615–12616
Davidson ER (1976) Reduced density matrices in quantum chemistry. Academic Press, New York
Davydov AS (1965) Quantum mechanics. Pergamon Press, New York
Dean PM (1995) Molecular similarity in drug design. Blackie Academic & Professional, London
Dedekind R (1963) Essays on the theory of numbers. Dover, New York
Dunn JF, Nisula BC, Rodbard D (1981) J Clin Endocrinol Metab 53:58–68
Eyring H, Walker J, Kimball GE (1948) Quantum chemistry. Wiley, New York
Ferro N, Gallegos A, Bultinck P, Jacobsen HJ, Carbó-Dorca R, Reinard T (2006) Coulomb and overlap self-similarities: A comparative selectivity analysis of structure-function relationships for Auxin-like molecules. J Chem Inf Model 46:1751–1762
Ferro N, Bultinck P, Gallegos A, Jacobsen HJ, Carbó-Dorca R, Reinard T (2007) Unrevealed structural requrements for auxin-like molecules by theoretical and experimental evidences. Phytochemistry 68:237–250
Fradera X, Amat LL, Besalú E, Carbó-Dorca R (1997) Application of molecular quantum similarity to QSAR. Quant Struct Activity Relatsh 16:25–32
Fratev F, Monev V, Mehlhorn A, Polansky OE (1979) J Mol Struct 56:255–266
Fratev F, Polansky OE, Mehlhorn A, Monev V (1979) J Mol Struct 56:245–253
Gallegos A (2006) Mini-review on chemical similarity and prediction of toxicity. Curr Comput Aided Drug Des 2:105–122
Gallegos A, Girones X (2005) Theoretical Background and QSPR Application. J Chem Inf Model 45:321–326
Gallegos A, Girones X (2005) Topological quantum similarity measures: Applications in QSAR. J Mol Struct Theochem 727:97–106
Gallegos A, Patlewicz G, Worth AP (2005) The use of similarity measures in defining the applicability domain of skin sensitisation SARs. Altex-Alternativen zu Tierexperimenten 22:272
Gallegos A, Netzeva TI, Worth AP (2006) Prediction of estrogenicity: Validation of a classification model. SAR QSAR Environ Res 17:195–223
Gallegos A, Tsakovska I, Pavan M, Patlewicz G, Worth A (2007) Evaluation of SARs for the prediction of skin irritation/corrosion potential-structural inclusion rules in the BfR decision support system. SAR QSAR Environ Res 18:331–342
Gallegos A, Patlewicz G, Worth AP (2008) A review of (Q)SAR models for skin and eye irritation and corrosion. QSAR and Combinatorial Sciences 27:49–59
Geerlings P, De Proft F (2002) Chemical reactivity as described by quantum chemical methods. Int J Mol Sci 3:276–309
Geerlings P, Boon G, Van Alsenoy C, De Proft F (2005) Density functional theory and quantum similarity. Int J Quantum Chem 101:722–732
Gironés X, Amat L, Carbó-Dorca R (1999) Using molecular quantum similarity measures as descriptors in quantitative structure-toxicity relationships. SAR QSAR Environ Res 10:545–556
Gironés X, Robert R, Carbó-Dorca R (2001) TGSA: A molecular superposition program based on topo-geometrical considerations. J Comput Chem 22:255–263
Gironés X, Amat L, Carbó-Dorca R (2002) Modeling large macromolecular structures using promolecular densities. J Chem Inf Comput Sci 42:847–852
Gironés X, Carbó-Dorca R (2004) TGSA-Flex: Extending the capabilities of the topo-geometrical superposition algorithm to handle rotary bonds. J Comput Chem 25:153–159
Goldstein S (1988) Physics Today, March, 42 and April, 38
Good AC (1992) J Mol Graph 10:144–151
Good AC, Richards WG (1993) J Chem Inf Comput Sci 33:112–116
Good AC, Hodgkin EE, Richards WG (1992) J Chem Inf Comput Sci 32:188–191
Greiner W (1997) Relativistic quantum mechanics. Springer, Berlin

Gruber PM, Wills JM (1993) Handbook of convex geometry. North-Holland, Amsterdam
Ho M, Smith VH, Weaver DF, Gatti C (1998) J Chem Phys 108:5469–5475
Hodgkin EE, Richards WG (1987) Int J Quantum Chem 14 :105–110
Huzinaga S, Klobukowski M (1998) J Mol Struct Theochem 167:1–209
Jeffrey A (1995) Handbook of mathematical formulas and integrals. Academic Press, New York
Klein J, Babic D (1997) Chem Inf Comput Sci 37:656
Lee CH, Smithline SH (1994) J Phys Chem 98:1135–1138
Leherte L (2006) Similarity measures based on Gaussian-type promolecular electron density models: Alignment of small rigid molecules. J Comput Chem 27:1800–1816
Lobato M, Besalú E, Carbó R (1996) Relacions Estructura-Propietat per un conjunt d'hidrucarburs a partir de nous descriptors tridimensionals derivats de la Semblança Molecular. Scientia Gerund 22:79–86
Lobato M, Amat L, Besalú E, Carbó-Dorca R (1997) Structure-activity relationships of a steroid family using quantum similarity measures and topological quantum similarity indices. Quant Struct Activity Relatsh 16:465–472
Lobato M, Amat L, Besalú E, Carbó-Dorca R (1998) Estudi QSAR d'una família de quinolones utilitzant indexs de semblança i indexs topològics de semblance. Scientia Gerund 23:17–27
Ludeña EV (1987) In: Fraga S (ed) Química Teórica, Nuevas Tendencias, vol 4. CSIC, Madrid, pp 117–160
Löwdin PO (1955) Quantum theory of many-particle systems. I. Physical interpretations by means of density matrices, natural spin-orbitals, and convergence problems in the method of configuration interaction. Phys Rev 97:1474–1489
Löwdin PO (1955) Quantum theory of many-particle systems. II. Study of the ordinary hartree-fock approximation. Phys Rev 97:1490–1508
Löwdin PO (1955) Quantum theory of many-particle systems. III. Extension of the Hartree–Fock scheme to include degenerate systems and correlation effects. Phys Rev 97:1509–1520
March NH (1992) Electron density theory of atoms and molecules. Academic Press, London
Mc Weeny R (1959) Proc Roy Soc A 253:242–259
Mc Weeny R (1960) Some recent advances in density matrix theory. Rev Mod Phys 32:335–369
Mc Weeny R (1978) Methods of molecular quantum mechanics. Academic Press, London
Mestres J, Solà M, Duran M, Carbó R (1994) On the calculation of ab initio quantum molecular similarities for large systems. J Comput Chem 15:1113–1120
Mestres J, Solà M, Duran M, Carbó R (1994) On the use of ab initio quantum molecular similarities as an interpretative tool for the study of chemical reactions. J Am Chem Soc 116:5909–5915
Mestres J, Solà M, Carbó R (1995) First-order molecular descriptors for molecular steric similarity. Scientia Gerund 21:165–173
Mestres J, Solà M, Carbó R, Luque FJ, Orozco M (1996) Effect of solvation on the charge distribution of a series of anionic, neutral, and cationic species. A quantum molecular similarity study. J Phys Chem 100:606–610
Mezey PG (1993) Shape in chemistry: An introduction to molecular shape and topology. VCH, New York
Mezey PG (1995) Molecular similarity I. In: Sen K (ed) Topics in current chemistry, vol 173. Springer, Berlin, pp 63–83
Mezey PG (1998) Averaged electron densities for averaged conformations. J Comput Chem 19:1337–1344
Mezey PG (2005) Graph representations of molecular similarity measures based on topological resolution. J Comput Methods Sci Eng 5:109–114
Mezey PG, Ponec R, Amat L, Carbó-Dorca R (1999) Quantum similarity approach to the characterization of molecular chirality. Enantiomers 4:371–378
Myers RH (1990) Classical and modern regression with applications. PWS-KENT Publishing company, Boston
Netzeva TI, Gallegos A, Worth AP (2006) Topological quantum similarity indices based on fitted densities: Comparison of the applicability domain of a QSAR for estrogenicity with a large chemical inventory. Environ Toxicol Chem 25:1223–1230
Ortiz JB, Cioslowski J (1991) Chem Phys Lett 185:270–275
Paniagua JC, Alemany P (1999–2000) Química Quàntica. Vols 1 and 2, Barcelona, 1999–2000
Parr RG (1963) The quantum theory of molecular electronic structure. WA Benjamin, New York
Patlewicz G, Jeliazkova N, Gallegos Saliner A, Worth AP (2008) Toxmatch – A new software tool to aid in the development and evaluation of chemically similar groups. SAR QSAR Environ Res 19:397–412
Pauling L, Wilson EB Jr (1985) Introduction to Quantum mechanics. Dover, New York
Petke JD (1993) J Comput Chem 14:928–933
Pierre DA (1969) Optimization theory with applications. Wiley, New York
Pilar FL (1990) Elementary quantum chemistry. McGraw-Hill, Princeton
Pla L (1986) Análisis Multivariado: Método de Componentes Principales Monography no 27. Secretaría General de la OEA, Washington,
Ponec R (1993) J Chem Inf Comput Sci 33:805–811
Ponec R (1995) Overlap determinant method in the theory of pericyclic reactions. Lecture notes in chemistry, vol 65. Springer, Berlin
Ponec R, Strnad M (1990) Collect Czechoslov Chem Commun 55:2583–2589
Ponec R, Strnad M (1990) Collect Czechoslov Chem Commun 55:896–902
Ponec R, Strnad M (1991) J Math Chem 8:103–112
Ponec R, Strnad M (1991) J Phys Organic Chem 4:701–705
Ponec R, Strnad M (1992) Int J Quantum Chem 42:501–508
Ponec R, Amat LL, Carbó-Dorca R (1999) Molecular basis of quantitative structure-properties relationships (QSPR): A quantum similarity approach. J Comput Aided Mol Design 13:259–270
Ponec R, Amat LL, Carbó-Dorca R (1999) Quantum similarity approach to LFER: Substituent and solvent effects on the acidities of carboxylic acids. J Phys Organic Chem 12:447–454
Riera A (1992) J Mol Struct Theochem 259:83–98
Robert D, Carbó-Dorca R (1998) A formal comparison between molecular quantum similarity measures and indices. J Chem Inf Comput Sci 38:469–475
Robert D, Carbó-Dorca R (1998) Analyzing the triple density molecular quantum similarity measures with the INDSCAL model. J Chem Inf Comput Sci 38:620–623
Robert D, Carbó-Dorca R (1998) On the extension of QS to atomic nuclei: Nuclear QS. J Math Chem 23:327–351
Robert D, Carbó-Dorca R (1998) Structure-property relationships in nuclei. Prediction of the binding energy per nucleon

using a quantum similarity approach. Il Nuovo Cimento A111:1311–1321
Robert D, Amat LL, Carbó-Dorca R (1999) Three-dimensional quantitative structure-activity relationships from tuned molecular quantum similarity measures. Prediction of the corticosteroid-binding globulin binding affinity for a steroid family. Chem Inf Comput Sci 39:333–344
Safouhi H (2005) Analytical and numerical development for the two-centre overlap-like quantum similarity integrals over Slater-type functions. J Phys A: Mathematical and General 38:7341–7361
Sen K (1995) Molecular similarity. Topics in current chemistry. vols 173, 174. Springer, Berlin
Shavitt I (1977) In: Schaefer HF III (ed) Methods of electronic structure theory. Modern theoretical chemistry, vol 3. Plenum Press, New York, pp 189–275
Sobolev SL (1938) Math Sb 4:471
Solà M, Mestres J, Duran M, Carbó R (1994) Ab initio quantum molecular similarity measures on metal-substituted carbonic anhydrase (M (II) CA, M=Be, Mg, Mn, Co, Ni, Cu, Zn, and Cd). J Chem Inf Comput Sci 34:1047–1053
Solà M, Mestres J, Carbó R, Duran M (1996) A Comparative analysis by means of Quantum Molecular Similarity Measures of Density Distributions derived from conventional ab initio and Density Functional Methods. J Chem Phys 104:636–647
Solà M, Mestres J, Oliva JM, Duran M, Carbó C (1996) The use of ab initio quantum molecular selfsimilarity measures to analyze electronic charge density distributions. Int J Quantum Chem 58:361–372
Spiegel MR (1968) Mathematical handbook. McGraw-Hill, New York
Stoer J, Witzgall CH (1970) Convexity and optimization in finite dimensions. Die Grundlehren der matematischen Wissenschaften in Einzeldarstellungen, Band 163. Springer, Berlin
Trillas E, Alsina C, Terricabras JM (1995) Introducción a la Lógica Difusa. Ariel Matemática, Barcelona
Tsakovska I, Gallegos A, Netzeva T, Pavan M, Worth AP (2007) Evaluation of SARs for the prediction of eye irritation/corrosion potential-structural inclusion rules in the BfR decision support system. SAR QSAR Environ Res 18:221–235
Vinogradov IM (1989) Encyclopaedia of Mathematics, vol 4. Kluwer Academic, Dordrecht, pp 422–428
Vracko M, Bandelj V, Barbieri P, Benfenati E, Chaudhry Q, Cronin M, Devillers J, Gallegos A, Gini G, Gramatica P, Helma C, Neagu D, Netzeva T, Pavan M, Patlewicz G, Randic M, Tsakovska I, Worth A (2006) Validation of counter propagation neural network models for predictive toxicology according to the OECD principles. A Case Study. SAR QSAR Environ Res 17: 265–284
Wagener M, Sadowski J, Gasteiger J (1995) J Am Chem Soc 117:7769–7775
Whitley DC, Ford MG, Livingstone DJ (2000) Unsupervised forward selection: A method for eliminating redundant variables. Chem Inf Comput Sci 40:1160–1168
Wilkinson JH (1965) The algebraic eigenvalue problem. Clarendon Press, Oxford
Wilkinson JH, Reinsch C (1971) Linear álgebra. Springer, Berlin
Worth AP, Bassan A, de Bruijn J, Gallegos A, Netzeva T, Patlewicz G, Pavan M, Tsakovska I, Eisenreich E (2007) The role of the european chemicals bureau in promoting the regulatory use of (Q)SAR methods. SAR QSAR Environ Res 18:111–125
Zadeh LA (1965) Inform Control 8:338

Quantum Simulations of Ballistic Nanowire Field Effect Transistors

Mincheol Shin
School of Engineering, Information and Communications University, Yuseong, Korea

Article Outline

Glossary
Definition of the Subject
Introduction
Model Systems
Simulation Methods
Simulation Results
Future Directions
Acknowledgment
Bibliography

Glossary

Silicon nanowire field effect transistors A silicon nanowire field effect transistor (SNWFET) has silicon nanowire as the channel, whose cross-sectional area is typically $10 \sim 100\,\mathrm{nm}^2$. The charge transport mainly occurs in one-dimensional subbands which are formed within the channel due to the strong quantum confinement. With enhanced gate control by three-dimensional gates surrounding the channel, SNWFETs can outperform conventional planar metal-oxide-semiconductor field effect transistors (MOSFETs) in their ultimate scaling limit.

Carbon nanotube field effect transistors Carbon nanotube field effect transistors (CNTFETs) resemble SNWFETs except that semiconducting carbon nanotubes are used as the channel material instead of silicon nanowires. Due to intrinsically nano-scale size, good electrical property, and almost ballistic transport nature of carbon nanotubes, CNTFETs exhibit device performance well exceeding conventional MOSFETs. CNTFETs with Schottky-barrier contacts or with doped source/drain can be realized.

Quantum transport in semi-conductor devices As the feature size of semi-conductor devices becomes extremely small, the wave nature of the electrons should prevail in their transport in the devices. The governing equation is the Schrödinger equation and typically single-particle Schrödinger equation with Hartree potential is sufficient for device simulations.

Non-equilibrium green's function The Schrödinger equation with open boundaries can be solved by us-

ing the non-equilibrium Green's Function (NEGF) approach, where the Green's function is defined as the impulse response function of the system Hamiltonian. NEGF approach is formally equivalent to other approaches such as quantum transmitting boundary method, but has advantages in inclusion of individual scattering/interaction terms. A key part of NEGF method is to calculate the self-energies, which contain information about the contacts and scattering or interaction that are considered.

Definition of the Subject

Nanowire FETs such as SNWFETs and CNTFETs are considered as strong candidates for the future nano-electronic devices to replace the today's planar MOSFETs. The devices are intrinsically nano-scale so quantum effects such as the size quantization, barrier tunneling, and interference should prevail. In particular, as the channel length of the nanowire is reduced below 10 nm or so, the source-to-drain tunneling current importantly contributes to the total current. In order to characterize and predict their device performance accurately, quantum device simulations based on the direct solution of the Schrödinger wave equations should be performed. The NEGF method, among others, provides a powerful solution scheme for quantum device simulations. Modeling and simulations are gaining greater importance as computer experiments, particularly in the nano-electronic device area where performance estimation and optimization of newly developed devices are necessary to reduce efforts and costs of real experiments.

Introduction

According to the ITRS 2007 [15], MOSFET devices with sub-10 nm channel length are expected to be fabricated by 2015. In their ultimate scaling regime, performance of conventional planar MOSFETs will be seriously degraded mainly due to the short channel effects [10]. As new device architectures to overcome the problems, nanowire FETs such as SNWFETs and CNTFETs have been recently drawing attention [5,7,16,25,33]. The advantage of SNWFETs mainly lies on the presence of the multiple gates around the channel which can suppress the short channel effects through the enhanced gate control [7,25,33]. In CNTFETs, intrinsically nano-scale carbon nanotubes with excellent electrical properties are used as the channel material so device performance exceeding conventional MOSFETs can be obtained [5,16]. To characterize the new devices and predict their performance accurately in the nano-scale regime, device simulations with solid quantum mechanical treatment are necessary.

For SNWFETs, quantum-mechanical studies have been mostly performed based on the parabolic effective mass theory (PEMT) [6,22,28,30]. PEMT provides a simple Hamiltonian to be dealt with so that efficient simulations employing NEGF method can be performed, especially for SNWFETs with homogeneous cross-sections and ballistic transport. For ultra-thin Si nanowires, however, their bulk property in the transverse direction is not preserved and so their band structure becomes different from that of bulk silicon. In fact, recent studies have revealed that if the cross-sectional area of the silicon nanowire is less than about $3 \times 3\,\mathrm{nm}^2$ non-parabolic and band-edge shift effects becomes so great that PEMT is no longer valid [12,24,31]. In this case, atomistic calculations such as full-band tight binding (TB) or first-principle calculations are necessary to obtain the correct dispersion relationship [12,20,23,24,31]. As practical device simulation approaches, however, they have limitations due to the enormous computational burden. A hybrid approach may give an optimal solution: namely, one continues to use PEMT for ultra-thin Si nanowire channel through appropriate tuning of effective masses and band gap from the atomistic calculations, which has been shown to be valid in recent works [24,31].

For CNTFETs, various approaches ranging from semi-classical to full-band quantum approaches have been implemented for device simulations [2,3,11,13,18,26]. The models yield comparable simulation results in the region above the threshold, but accurate dispersion relationships from atomistic calculations are needed to correctly account for intra-band and inter-band tunneling in the sub-threshold region. For that purpose, the single-band (p_z) TB approach seems to be an optimal one: One may consider more sophisticated Hamiltonians such as full-band or Hückel TB Hamiltonians, but as long as electronic transport is concerned the single-band TB seems to be sufficient because they yield almost the same dispersion relationships [21]. An effective mass approach can be also taken, but the computational complexity is not greatly reduced compared to the TB approach, while loss of accuracy is inevitably incurred due to simplifications involved.

In nanowire FETs of a few tens nanometer in size, diffusive transport due to electron-phonon interaction and surface roughening cannot be entirely ignored and there have recently been considerable efforts to include the scattering effects [14,17,32]. In this article, however, ballistic transport is assumed throughout, since our focus is on the ultimately scaled devices where ballistic transport is expected to prevail and may account for major device characteristics. Simulation based on the ballistic transport alone may also serve as a starting point to build more so-

phisticated simulation tools including scattering and other effects.

In this article, simulation methods employing PEMT and single band TB for SNWFETs and CNTFETs respectively are described in some details as practical approaches for quantum device simulations. Model systems and their Hamiltonians are firstly introduced, followed by NEGF method in the real and mode spaces. Details on the k-space and product-space solutions of cross-sectional Schrödinger equations for SNWFETs are next described. Numerical aspects of three-dimensional Poisson's equations are then addressed. Some simulation results for nanowire FETs are lastly shown, followed by conclusion and future prospects.

Model Systems

Figure 1a shows a schematic diagram of a SNWFET which is a three-dimensional (3D) structure with multiple gates around the silicon nanowire channel. The source (S) and drain (D) regions are either heavily doped (doped S/D device) or metallic such that Schottky barriers exist at the interfaces with the channel (SB device). The channel region is intrinsic or lightly p-doped and a metallic gate with the mid-gap work function of 4.6 eV is assumed here. Possible types of multiple gates around the nanowire channel are Tri, Pi, Omega, Gate-all-around (GAA) [6,29].

In CNTFETs, single-walled (n,0) zigzag nanotubes are used as the channel material as shown in Fig. 1b. For ohmic-contact devices, source and drain regions are assumed to be infinite carbon nanotubes, which are simple extension of the channel nanotube but heavily doped. If the source and drain regions are both n-doped (or p-doped) and channel is intrinsic, the devices resemble MOSFETs and thus called CNT MOSFETs [4,13]. If source-channel-drain is of $p-i-n$ type (or $n-i-p$), the devices operate like tunnel FETs [4]. For Schottky-contact devices, on the other hand, S/D regions are assumed to be metallic zigzag nanotubes such that Schottky barriers of height ϕ_{Bn}, which is given as an input, is formed at the interfaces with channel. Three gate-types, coaxial, semi-cylindrical, and top gates, are considered. The coaxial gate device is in effect two-dimensional (2D) device, with 2D electrostatics and transport, while the other two are truly 3D devices.

Simulation Methods

Self-consistent Method

In our device simulations of nanowire FETs, the quantum transport of electrons is described by the Schrödinger equation with open boundary conditions and electrostatics are given by the 3D Poisson's equation. To correctly describe highly out-of-equilibrium states due to the finite

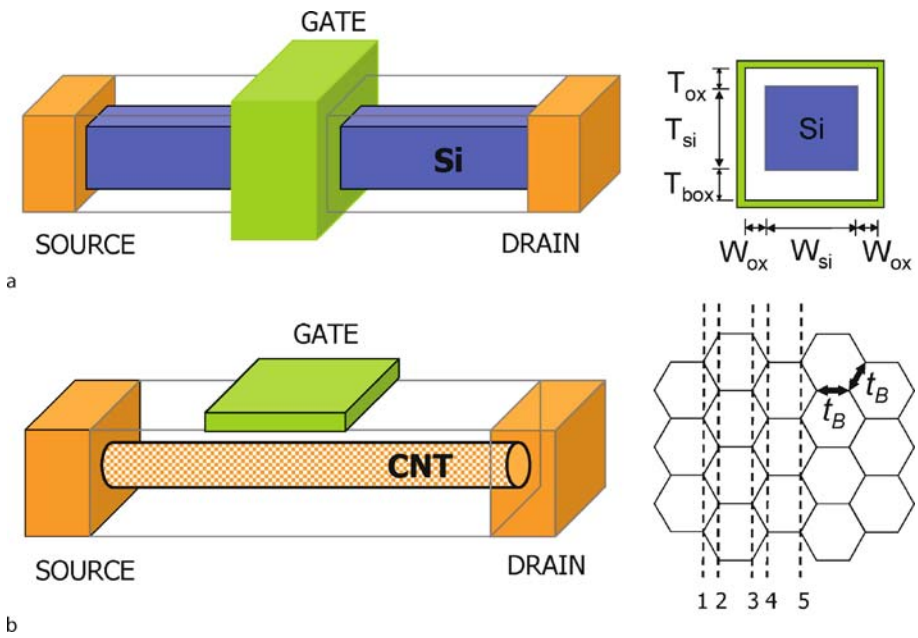

Quantum Simulations of Ballistic Nanowire Field Effect Transistors, Figure 1
Device Schematics: a SNWFET with its cross-section on the right and b CNTFET with zigzag carbon nanotube channel

source-drain voltage, the two equations should be solved together to yield the self-consistent solution. Namely, we iteratively solve the Schrödinger equation

$$H\{\phi\}\psi = E\psi \ , \tag{1}$$

to obtain the electron density $n_{3D} = |\psi|^2$, where H is the system Hamiltonian for a given potential profile $\{\phi\}$ and ψ is the wave function, and the Poisson's equation

$$\nabla^2\phi = -\frac{q_0}{\epsilon}\left(N_d - n_{3D}\right) \ , \tag{2}$$

to obtain the potential profile, where N_d is doping density. Once the self-consistency is reached, the drain current is calculated. Among other approaches to solve the Schrödinger equation with open boundaries, the non-equilibrium Green's function approach [9] is employed and described here.

EMT Hamiltonian for SNWFETs

The effective mass Hamiltonian for the SNWFET in Fig. 1a is given by

$$H = H_\perp + H_\parallel \ , \tag{3}$$

where

$$H_\perp = -\frac{\hbar^2}{2}\frac{\partial}{\partial y}\left(\frac{1}{m_y^*}\frac{\partial}{\partial y}\right) - \frac{\hbar^2}{2}\frac{\partial}{\partial z}\left(\frac{1}{m_z^*}\frac{\partial}{\partial z}\right) + V(x,y,z), \tag{4}$$

$$H_\parallel = -\frac{\hbar^2}{2m_x^*}\frac{\partial^2}{\partial x^2}, \tag{5}$$

where m_x^*, m_y^*, and m_z^* are effective masses in the x, y, and z directions, respectively, and $V(x,y,z)$ is the conduction band-edge. Notice that m_y^* and m_z^* are in general functions of y and z if both silicon and oxide regions are taken into account, while m_x^* is the silicon effective mass in the transport direction. For the effective masses, one may simply use the values of bulk Si; for instance, in the [100] transport direction, $(m_x^*, m_y^*, m_z^*) = (m_t, m_t, m_l)$, (m_t, m_l, m_t), and (m_l, m_t, m_t), for the three valleys respectively, where $m_t = 0.18m_0$ and $m_l = 0.98m_0$, and each valley has the degeneracy of 2. For other transport directions, one can refer to Ref. [27]. Effective masses in the oxide region do not importantly affect the vertical confinement energies nor lateral transport, so one may extend the Si effective masses into the oxide region for computational simplicity.

The Si effective masses can be also extracted from the band diagrams produced by full-band TB or first principle calculations. In the original work of Ko et al. [20], a $sp^3d^5s^*$ TB model was considered and the conduction bands of a Si (100) wire were obtained: Two pairs of valleys out of six valleys with k vectors perpendicular to the wire direction were found to fall onto the Γ point in the wire Brillouin zone, and the electron effective masses along the confinement directions are relatively larger for the states than for the two remaining states that become off-Γ states in the wire. Similar works performed by other groups [24,31] and a recent DFT calculation [12] all indicate that the Si effective masses start to deviate from their bulk values if the wire cross-sectional area becomes smaller than about 10 nm^2.

For a nanowire with circular cross-section, $H_\perp$ can be expressed in the circular polar coordinates, in an approximation,

$$H_\perp = -\frac{\hbar^2}{2m_\perp^*}\left(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial\theta^2}\right) \ , \tag{6}$$

where

$$\frac{1}{m_\perp^*} = \frac{1}{2}\left(\frac{1}{m_x^*} + \frac{1}{m_y^*}\right). \tag{7}$$

In the real-space, if we use the finite difference scheme and slice the nanowire along the transport direction into N_x cross-sections, $H_\perp$ becomes a block diagonal matrix with diagonal elements $H_{\perp,i}$ of size $N_s \times N_s$ which is the 2D Hamiltonian of the ith cross-section, where N_s is the number of grid points of a cross-section, and $H_\parallel$ becomes a block tridiagonal matrix with diagonal elements

$$\alpha = 2t_x \cdot I_{N_s\times N_s} \ , \tag{8}$$

and off-diagonal elements

$$\beta = -t_x \cdot I_{N_s\times N_s} \ , \tag{9}$$

where $I_{N_s\times N_s}$ is an $N_s \times N_s$ identity matrix and $t_x = \hbar^2/2m_x^*(\Delta x)^2$, where Δx is the grid spacing in the transport direction.

TB Hamiltonian for CNTFETs

A $(n,0)$ zigzag nanotube has alternating sublattices, as shown in Fig. 1b. In the nearest-neighborhood single-band (p_z) tight binding description, the coupling matrices between the sublattices are given by t and b, where

$$t = t_B \cdot I_{n\times n} \ , \tag{10}$$

$$b = t_B \begin{bmatrix} 1 & 1 & & & 1 \\ & 1 & 1 & & \\ & & & \ddots & \\ 1 & & & & 1 \end{bmatrix} , \tag{11}$$

where $I_{n\times n}$ is an $n \times n$ identity matrix and t_B is the TB coupling parameter. Then the Hamiltonian is written as

$$H = \begin{bmatrix} a_0 & b^\dagger & & & & \\ b & a_1 & t & & & \\ & t & a_2 & b & & \\ & & b^\dagger & a_3 & t & \\ & & & t & a_4 & \\ & & & & & \ddots \end{bmatrix} \tag{12}$$

where a_i is the coupling matrix within the ith sublattice, which is $n \times n$ diagonal matrix with elements

$$[a_i]_{j,j} = -q_0\,\phi(i,j)\,, \tag{13}$$

where j is the index along the circumferential direction of nanotube ($1 \le j \le n$) and $\phi(i,j)$ is the vacuum level potential at the atom site (i,j) as determined by the Poisson's equation.

NEGF Approach

In the NEGF approach [9], the device Green's function $G(E)$ is given by

$$G(E) = (E - H - \Sigma_{\rm L} - \Sigma_{\rm R})^{-1}\,, \tag{14}$$

where H is the device Hamiltonian and $\Sigma_{\rm L,R}$ are contact self-energies of semi-infinite leads to the left (L) and right (R) of the device region, respectively. The ballistic transport is assumed here as only the contact self-energies are considered, which are formally expressed by

$$\Sigma_{\rm L,R} = \chi^\dagger_{\rm L,R}(E - H_{\rm L,R})^{-1}\chi_{\rm L,R}, \tag{15}$$

where $H_{\rm L,R}$ are Hamiltonians of the left and right leads, respectively, and $\chi_{\rm L,R}$ are coupling matrices between the device and the leads.

Given the Green's function $G(E)$, the charge density at position $\vec{r}$ can be calculated as follows. Let us first define

$$\rho_{\rm L,R}(\vec{r},E) = \left[G(E)\Gamma_{\rm L,R}(E)G^\dagger(E)\right]_{\vec{r},\vec{r}} \tag{16}$$

where

$$\Gamma_{\rm L,R}(E) = i\left(\Sigma_{\rm L,R}(E) - \Sigma^\dagger_{\rm L,R}(E)\right) \tag{17}$$

and $[\cdot]_{\vec{r},\vec{r}}$ denotes the diagonal element of the relevant matrix corresponding to the position $\vec{r}$.

The electron density $n_{\rm 3D}(\vec{r})$ in n-type SNWFETs, where only the electron transport in n-type FETs is considered, is given by

$$n_{\rm 3D}(\vec{r}) = \frac{1}{2\pi\Delta V} \cdot \int_{-\infty}^{\infty} {\rm d}E\left(f^+_{\rm L}\rho_{\rm L}(\vec{r},E) + f^+_{\rm R}\rho_{\rm R}(\vec{r},E)\right)\,, \tag{18}$$

where ΔV is the volume element, whereas the electron density $n_{\rm 2D}(\vec{r})$ and the hole density $p_{\rm 2D}(\vec{r})$ on the CNT surface in CNTFETs are given by

$$n_{\rm 2D}(\vec{r}) = \frac{1}{2\pi\Delta S}\int_{E_{\rm n}(\vec{r})}^{\infty} {\rm d}E\left(f^+_{\rm L}\rho_{\rm L}(\vec{r},E) + f^+_{\rm R}\rho_{\rm R}(\vec{r},E)\right)\,, \tag{19}$$

$$p_{\rm 2D}(\vec{r}) = \frac{1}{2\pi\Delta S}\int_{-\infty}^{E_{\rm n}(\vec{r})} {\rm d}E\left(f^-_{\rm L}\rho_{\rm L}(\vec{r},E) + f^-_{\rm R}\rho_{\rm R}(\vec{r},E)\right)\,, \tag{20}$$

where $E_{\rm n}(\vec{r}) = -q_0\phi(\vec{r})$ is the charge neutrality level and ΔS is the area element. In Eqs. (18)–(20),

$$f^{\pm}_{\rm L,R}(E) = \frac{1}{1 + {\rm e}^{\pm\left(E - E_{\rm F}^{L,R}\right)/k_{\rm B}T}}\,, \tag{21}$$

where $E^L_{\rm F} = E_{\rm F}$ is the Fermi energy at the source and $E^R_{\rm F} = E_{\rm F} - q_0 V_{\rm d}$ is the Fermi energy at the drain, where $V_{\rm d}$ is the drain voltage.

The drain current $I_{\rm d}$ is calculated by using the Landauer–Büttiker formula:

$$I_{\rm d} = \frac{2q}{h}\int_{-\infty}^{\infty} {\rm d}E\,T(E)(f^+_{\rm L}(E) - f^+_{\rm R}(E)) \tag{22}$$

where the transmission probability $T(E)$ is

$$T(E) = {\rm Tr}(\Gamma_{\rm L}G\Gamma_{\rm R}G^\dagger) \tag{23}$$

Mode-Space NEGF

The idea of the mode-space NEGF is basically to perform a unitary transformation from the real space to the "mode" space and to consider only the modes which contribute to the transport [8]. If the unitary matrix is denoted by U, we form

$$\tilde{H} = U^\dagger H U. \tag{24}$$

We can then deal with $\tilde{H}$ instead of the original Hamiltonian and obtain the self-energies, density matrix and current using the transformed Hamiltonian. For the density matrix, however, inverse unitary transformation back to the real-space is needed to obtain the charge density in the real-space.

For SNWFETs described by the effective mass Hamiltonian of Eq. (3), the 2D Hamiltonian $H_{\perp,i}$ of each cross-section is diagonalized by the unitary matrix U_i to yield

$$\tilde{H}_{\perp,i} = U^\dagger_i H_{\perp,i} U_i\,, \tag{25}$$

and the transformation matrix U may be formed as a block diagonal matrix having diagonal elements U_i. On the

other hand, for CNTFETs described by the TB Hamiltonian of Eq. (12), the coupling matrix b is diagonalized as

$$\tilde{b} = U_b^\dagger \, b \, U_b \, , \tag{26}$$

and the block-diagonal unitary matrix U may then be formed to have identical diagonal elements U_b.

The advantage of the mode-space approach is that one may handle with matrices of much smaller size. That is, if the number of modes that effectively contribute to the transport is N_M out of total N_s, the size of H is reduced from $N_x N_s \times N_x N_s$ in the real-space to $N_x N_M \times N_x N_M$ in the mode-space. For instance, for a SNWFET with rectangular cross-section of $5 \times 5\,\mathrm{nm}^2$, N_M is about 10, whereas a moderate real-space meshing would yield $N_s \sim 1000$ so computational efficiency is much greater in the mode-space compared to the real-space.

Uncoupled Mode-space NEGF for SNWFETs

If the coupling between different transport modes in the mode-space is non-existent or can be ignored, the transport problem effectively becomes that of 1D transport modes in parallel [8,13]. For SNWFETs, the so-called uncoupled mode-space NEGF is resulted if we assume that the unitary matrices U_i of Eq. (25) has the following property;

$$U_{i+1}^\dagger U_i \approx I_{N_s \times N_s} \tag{27}$$

i. e., the eigenfunctions of neighboring cross-sectional planes are orthogonal to each other in an approximation. Then

$$\tilde{H} = U^\dagger H U = U^\dagger H_\perp U + U^\dagger H_\parallel U \approx \tilde{H}_\perp + H_\parallel \, , \tag{28}$$

where the block diagonal matrix $\tilde{H}_\perp$ has diagonal elements $\tilde{H}_i$ of Eq. (25). Thus, $\tilde{H}$ becomes a block tridiagonal matrix with elements which are diagonal matrices of size $N_s \times N_s$. By interchanging rows and columns appropriately, $\tilde{H}$ then becomes block-diagonal with elements $\tilde{H}^{(m)}$, where

$$\tilde{H}^{(m)} = \begin{bmatrix} \epsilon_1^{(m)} + 2t_x & -t_x & & & \\ -t_x & \epsilon_2^{(m)} + 2t_x & -t_x & & \\ & & \ddots & & \\ & & & & -t_x \\ & & & -t_x & \epsilon_{N_x}^{(m)} + 2t_x \end{bmatrix}, \tag{29}$$

where $\epsilon_i^{(m)}$ is the mth eigenvalue of the ith cross-sectional plane. $\tilde{H}^{(m)}$ of Eq. (29) corresponds to the 1D Hamiltonian of the mth mode, with "effective" potential $\epsilon^{(m)}(x)$ in the transport direction. The modes are therefore completely uncoupled: With the charge density $n_{1\mathrm{D}}^{(m)}$ of each mode, the 3D charge density is obtained by

$$n_{3\mathrm{D}}(x_i, y_j, z_k) = \sum_m n_{1\mathrm{D}}^{(m)}(x_i) |[U_i]_{jk,m}|^2, \tag{30}$$

and the total drain current is given by

$$I_\mathrm{d} = \sum_m I_{\mathrm{d},m} \, , \tag{31}$$

where $I_{\mathrm{d},m}$ is the current of the mth mode.

In the uncoupled mode-space approach the original 3D problem is therefore split into problems of solving the 2D Schrödinger equations in the cross-sectional planes and the 1D Schrödinger equation in the transport direction. Note that the uncoupled mode-space approach is well suited for the device simulation considered here, because the silicon-on-insulator structure ensures the confinement of the cross-sectional wave functions in the silicon channel and consequently the shape of the 2D wave functions do not change much in the lateral direction.

Uncoupled Mode-Space NEGF for CNTFETs

For CNTFETs with coaxial gates, electrostatic potential $\phi(i)$ at the ith sublattice is same along the circumferential direction. Thus, by the unitary transformation of Eq. (26), a_i's in Eq. (12) remain intact so $\tilde{H}$ becomes block-tridiagonal with elements being $n \times n$ diagonal matrices. As in the case of the SNWFETs, by interchanging rows and columns $\tilde{H}$ becomes block-diagonal with elements $\tilde{H}^{(m)}$, where

$$\tilde{H}^{(m)} = \begin{bmatrix} -q_0\phi(0) & \tilde{b}_m & & & \\ \tilde{b}_m & -q_0\phi(1) & t_\mathrm{B} & & \\ & t_\mathrm{B} & -q_0\phi(2) & \tilde{b}_m & \\ & & \tilde{b}_m & -q_0\phi(3) & t_\mathrm{B} \\ & & & & \ddots \end{bmatrix}. \tag{32}$$

The charge density on the CNT surface and the drain current I_d can be calculated similarly to the case of the SNWFETs. In particular, the electron density

$$n_{2\mathrm{D}}(x_i, y_j) = \sum_m n_{1\mathrm{D}}^{(m)}(x_i) \left|[U_b]_{j,m}\right|^2 = \frac{1}{\Delta y} \sum_m n_{1\mathrm{D}}^{(m)}(x_i) \tag{33}$$

where $\Delta y = 2\pi R_c / n$ where R_c is the radius of the CNT. Notice that the uncoupled mode-space approach for the co-axial CNTFETs is exact [13].

Surface Green's functions

For SNWFETs described by the effective mass Hamiltionian of Eqs. (3)–(5), the contact self-energies $\Sigma_{\rm L}$ and $\Sigma_{\rm R}$ in Eq. (15) have non-zero elements $\beta^\dagger g_{\rm L}\beta$ and $\beta\, g_{\rm R}\, \beta^\dagger$ in the first and last $N_s \times N_s$ blocks, respectively, due to the structure of the coupling matrices $\chi_{\rm L}$ and $\chi_{\rm R}$. The surface Green's functions $g_{\rm L,R}$ can be rather simply obtained as follows. Assuming that the reservoir to the left of the device is simple extension of the left contact, $g_{\rm L}$ can be written as

$$g_{\rm L} = \left(\gamma - \beta^\dagger g_{\rm L}\beta\right)^{-1}, \tag{34}$$

where a $N_s \times N_s$ matrix

$$\gamma = E \cdot I_{N_s\times N_s} - h_0 - \alpha\,, \tag{35}$$

where h_0 is the 2D Hamiltonian at the contact cross-section. If U_0 is the matrix which diagonalizes h_0 and we denote the unitary-transformed matrix $\tilde{A}$ of a matrix A as $\tilde{A} \equiv U_0^\dagger A U_0$, Eq. (34) becomes under the unitary transformation by U_0,

$$\tilde{g}_{\rm L} = \left(\tilde{\gamma} - \beta^\dagger \tilde{g}_{\rm L}\beta\right)^{-1}. \tag{36}$$

The matrices in the equation are all diagonal matrices, thus element kl of the matrix $\tilde{g}_{\rm L}$ is

$$[\tilde{g}_{\rm L}]_{k,l} = \frac{\tilde{\gamma}_k - i\sqrt{4t_x^2 - \tilde{\gamma}_k^2}}{2t_x^2}\delta_{k,l}\,, \tag{37}$$

where $\delta_{k,l}$ is the Kronecker delta. The real-space $g_{\rm L}$ can be finally obtained by inverse unitary transformation. The surface Green's function $g_{\rm R}$ at the drain contact can be calculated likewise.

For CNTFETs described by the TB Hamiltonian of Eq. (12), the mode-space surface Green's function $\tilde{g}_{\rm L}(= U_b^\dagger g_{\rm L} U_b)$ at the source contact can be calculated by solving the following coupled matrix equations:

$$\tilde{g}_{\rm L} = \left(\tilde{c} - \tilde{b}^\dagger \tilde{g}'_{\rm L}\tilde{b}\right)^{-1}, \tag{38}$$

$$\tilde{g}'_{\rm L} = \left(\tilde{c} - t_{\rm B}^2 \tilde{g}_{\rm L}\right)^{-1}, \tag{39}$$

where

$$\tilde{c} = U_b^\dagger\,(E \cdot I_{n\times n} - a_0)\,U_b \tag{40}$$

where a_0 as defined in Eq. (13) is the matrix for the sublattice in contact with the source reservoir. If the potential at the contact is constant along the circumferential direction so that

$$a_0 = -q_0\phi(0)\cdot I_{n\times n} \tag{41}$$

is satisfied, as is true for the case of co-axially-gated CNTFETs, all the matrices in Eqs. (38)–(39) are diagonal. One can then solve for $\tilde{g}_{\rm L}$, similarly to the SNWFET case above, and obtain the real-space surface Green's function via the inverse unitary transformation.

For top-gated CNT MOSFETs, $\tilde{c}$ in Eq. (40) has off-diagonal elements in general due to the fact that the potential at the contact may vary along the circumference of the CNT, and the calculation of the surface Green's function becomes quite involved [11,19]. But if the ideal ohmic contact is assumed, the contact in a simulation can be located arbitrarily far from the gate such that the condition in Eq. (41) is satisfied again. For efficient simulations, one may assume an artificially high dielectric in the immediate vicinity of the contacts, instead of having a very long simulation region, in order to achieve the condition in Eq. (41).

For the Schottky-contact CNTFETs where the source and drain leads are assumed to be metallic zigzag CNTs, $\tilde{g}_{\rm L}$ and $\tilde{g}'_{\rm L}$ in Eqs. (38) and (39) become identical to each other and one has, regardless of the gate types,

$$\tilde{g}_{\rm L} = \left(\tilde{c} - t_{\rm B}^2 \tilde{g}_{\rm L}\right)^{-1}, \tag{42}$$

with $\phi(0)$ in Eq. (41) being a constant built-in potential due to the Schottky contact. In this special case, the real-space surface Green's functions are identical to the mode-space counterparts, because the latter are independent of the modes.

Solution of 2D Schrödinger Equations

In the mode-space approach for SNWFETs, 2D Hamiltonian $H_{\perp,i}$ at each cross-section is diagonalized; i. e., the eigenvalue problem

$$H_{\perp,i}\psi_m(y,z;x_i) = E_m(x)\psi_m(y,z;x_i)\,, \tag{43}$$

where

$$H_{\perp,i} = -\frac{\hbar^2}{2}\frac{\partial}{\partial y}\left(\frac{1}{m_y^*}\frac{\partial}{\partial y}\right) - \frac{\hbar^2}{2}\frac{\partial}{\partial z}\left(\frac{1}{m_z^*}\frac{\partial}{\partial z}\right) + V(y,z;x_i)\,, \tag{44}$$

is solved subject to the boundary condition that the wave functions vanish at the boundaries of the 2D cross-sectional plane. If we write $\psi_m(y,z;x_i)$ in Eq. (43) as

$$\psi_m(y,z;x_i) = \sum_K A_K|K\rangle\,, \tag{45}$$

where $\{|K\rangle\}$ is a basis set and A_K's are expansion coefficients, and insert it in Eq. (43) and multiplying $\langle L|$ to

the both sides of the equation, we obtain an eigenvalue problem

$$\sum_K \mathcal{H}_{LK} A_K = E_m A_L \,, \tag{46}$$

where $\mathcal{H}_{LK} \equiv \langle L|H_{\perp,i}|K\rangle$. In the widely used k-space solution [1], we use

$$|K\rangle = \sqrt{\frac{2}{L_y}}\sqrt{\frac{2}{L_z}}\sin(k_p y)\sin(k_q z) \,, \tag{47}$$

where L_y and L_z are side lengths of the cross-section in the y and z directions, respectively, and

$$k_p = \frac{p\pi}{L_y} \,, \quad (p = 1, \ldots, N_y) \tag{48}$$

$$k_q = \frac{q\pi}{L_z} \,, \quad (q = 1, \ldots, N_z) \tag{49}$$

where N_y and N_z are number of meshes in the y and z directions, respectively. Note that the index K in Eq. (47) is mapped to indices p and q through $K = N_y(p-1) + q$. Specifically, for the Hamiltonian of Eq. (44) with continuous effective masses across the silicon/oxide interfaces, $\mathcal{H}_{LK}$ is given by

$$\mathcal{H}_{LK} = \frac{4}{L_y L_z}\int_0^{L_y}\int_0^{L_z} \mathrm{d}y\,\mathrm{d}z \sin(k_{p'} y)\sin(k_{q'} z) \times \ldots$$
$$\cdots \times \left(\frac{\hbar^2 k_p^2}{2m_y^*} + \frac{\hbar^2 k_q^2}{2m_z^*} + V(y,z;x_i)\right) \times \sin(k_p y)\sin(k_q z), \tag{50}$$

where $k_{p'}$ and $k_{q'}$ are defined similarly to k_p and k_q in Eqs. (48) and (49), respectively, and the index L in the equation is mapped to indices p' and q' through $L = N_y(p'-1) + q'$. Numerically, $\mathcal{H}_{LK}$ in Eq. (50) can be efficiently evaluated using the FFT routines [1].

In the product-space approach [28], on the other hand, we use the basis set

$$|K\rangle \equiv \chi_p(y)\zeta_q(z) \,, \tag{51}$$

where $\chi_p(y) \equiv \chi_p(y;x_i)$ is the pth mode eigenfunction in the y direction with its eigenvalue $\epsilon_p^{(y)} \equiv \epsilon_p^{(y)}(x_i)$, satisfying

$$\left\{-\frac{\hbar^2}{2m_y^*}\frac{d^2}{dy^2} + \bar{V}(y)\right\}\chi_p(y) = \epsilon_p^{(y)}\chi_p(y) \,, \tag{52}$$

where $\bar{V}(y) \equiv \bar{V}(y;x_i)$ is an average potential in the y direction, defined by

$$\bar{V}(y;x_i) = \frac{1}{T_{\mathrm{si}}}\int_{T_{\mathrm{ox}}}^{T_{\mathrm{ox}}+T_{\mathrm{si}}} \mathrm{d}z V(x_i, y, z) \,, \tag{53}$$

where T_{si} is the silicon channel thickness as shown in Fig. 1a. Similarly, we define $\zeta_q(z) \equiv \zeta_q(z;x_i)$ as the qth mode eigenfunction in the z direction with its eigenvalue $\epsilon_q^{(z)} \equiv \epsilon_q^{(z)}(x_i)$, satisfying

$$\left\{-\frac{\hbar^2}{2m_z^*}\frac{d^2}{dz^2} + \bar{V}(z)\right\}\zeta_q(z) = \epsilon_q^{(z)}\zeta_q(z) \,, \tag{54}$$

where $\bar{V}(z) \equiv \bar{V}(z;x_i)$ is an average potential in the z directions, defined by

$$\bar{V}(z;x_i) = \frac{1}{W_{\mathrm{si}}}\int_{W_{\mathrm{ox}}}^{W_{\mathrm{ox}}+W_{\mathrm{si}}} \mathrm{d}y V(x_i, y, z) \,. \tag{55}$$

The 1D eigenvalue problems of Eqs. (52) and (54) can be solved easily in the 1D k-space transformation [1]. If we use a uniform mesh of sizes N_y and N_z in the y and z directions, respectively, K ranges from 1 to $N_y N_z$ and we index it in the order of increasing value of $\epsilon_K \equiv \epsilon_p^{(y)} + \epsilon_q^{(z)}$. Using the relationships of Eqs. (51)–(55), $\mathcal{H}_{LK}$ in the product space becomes

$$\mathcal{H}_{LK} = \epsilon_L \delta_{LK} + \langle L|(V(y,z) - \bar{V}(y) - \bar{V}(z)|K\rangle. \tag{56}$$

An advantage of the product-space approach is that only the first few eigenvalues, which correspond to the subband modes that contribute to the transport, are sufficient to be included in the eigenvalue problem. In other words, if M is the number of subbands that participate in the transport, the size of the matrix $\mathcal{H}$ in Eq. (56) is reduced to M by M. For instance, M is about 10 for the nanowire transistor of cross-sectional area of $5 \times 5\,\mathrm{nm}^2$. One therefore needs to find the first 10 eigenvalues of a 10 by 10 matrix in the product-space solution. As the area of the cross-sectional plane increase, M increases in proportion to the area. The matrix $\mathcal{H}$ is effectively a banded matrix of band size N_B which ranges from 1 to 10 for the cross sectional areas of $5 \times 5\,\mathrm{nm}^2$ to $20 \times 20\,\mathrm{nm}^2$, which also contributes to the efficiency of the numerical calculation.

Solution of 3D Poisson's Equation

In a self-consistent solution of SNWFETs, the following Poisson's equation,

$$\nabla^2\phi^k = -\frac{q}{\epsilon}\left(N_D - n_{3\mathrm{D}}^k\, \mathrm{e}^{(\phi^k - \phi^{k-1})/k_{\mathrm{B}}T}\right) \,, \tag{57}$$

may be used instead of the one in Eq. (2) to achieve faster convergence. In Eq. (57), ϕ^k and $n_{3\mathrm{D}}^k$ are the kth step solutions for the potential and electron density, respectively. For CNTFETs, on the other hand, we solve the Laplace equation in the oxide and interior of CNT,

$$\nabla^2\phi^k = 0 \,, \tag{58}$$

together with the interface condition on the CNT surface,

$$\epsilon_{ox} \left.\frac{\partial \phi^k}{\partial \bar{n}}\right|_{ox} - \epsilon_{cnt} \left.\frac{\partial \phi^k}{\partial \bar{n}}\right|_{cnt} = \ldots$$
$$\ldots q_0 \left(N_D + p_{2D}^k e^{(\phi^{k-1}-\phi^k)/k_B T} - n_{2D}^k e^{(\phi^k-\phi^{k-1})/k_B T} \right) , \tag{59}$$

where $\bar{n}$ denotes the direction normal to the CNT surface, and n_{2D}^k and p_{2D}^k are the kth step solutions for electron and hole densities on the CNT surface given by Eqs. (19) and (20), respectively.

The boundary conditions for the Poisson's equation for ohmic contact devices are such that for the gate contact region,

$$\phi(x, y, z) = V_g + \chi_e - \phi_{mg} \tag{60}$$

where χ_e and ϕ_{mg} are the electron affinity of channel material and the metal gate work function, but for other boundaries including source/drain contacts, the free boundary condition can be used. Especially, the free boundary condition at the source/drain contacts is necessary to achieve the thermal equilibrium conditions in the source/drain [9]. For the Schottky contacts,

$$\phi(x, y, z) = \begin{cases} \phi_{bi} & \text{at the source contact} \\ \phi_{bi} + V_d & \text{at the drain contact} , \end{cases} \tag{61}$$

where the built-in potential $\phi_{bi} \equiv E_c - E_F - \phi_{Bn}$ where E_c is the conduction band edge and ϕ_{Bn} is the Schottky barrier height.

Simulation Results

SNWFETs

Simulation results for some device characteristics of SNWFETs such as transfer characteristics and scaling behaviors are presented in the following. In the simulated devices, the channel silicon is lightly p-doped with $N_A = 10^{15}\,\text{cm}^{-3}$ and it is oriented such that its (100) direction is parallel to the transport direction. The silicon nanowire has square cross-section of $5 \times 5\,\text{nm}^2$, the gate oxide thickness is 1 nm, and the metal gate work function ϕ_{mg} is set at 4.61 eV. In the case of ohmic contact devices, the source/drain are heavily n-doped with the doping concentration of $10^{20}\,\text{cm}^{-3}$ while the Fermi energy of the metal source in the case of Schottky contact devices is assumed to lie 0.5 eV above the (virtual) metal CB edge.

The scaling behavior of GAA SNWFETs with doped S/D with respect to the shrink of the gate length is shown in Fig. 2a. It can be seen from the figure that the device characteristics degrade as the gate length becomes shorter. On-currents do not vary significantly with L, due to the ballistic nature of the transport, but off-current values increase sharply. We note that the SS increases sharply with L_g, especially when L_g is smaller than 10 nm. The behavior of SS with the gate length can be explained in terms of the increasing tunneling-current contribution to the total current, I_{tunn}/I_d, with respect to reduction of gate length, as can be seen in the inset of Fig. 2a. I_{tunn}/I_d becomes increasingly large as L_g becomes shorter, and if $L_g = 5$ nm, the tunneling contribution at $V_g = 0$ is more than 90% of the total current, giving rise to the high off-current. However, if $L_g = 15$ nm, for instance, the tunneling contribution becomes much smaller, below 40% in the off-state. In the on-state, the thermionic components dominate for the two cases. Similar trends were also observed for PI-gate and TRI-gate devices, with almost the same I_{tunn}/I_d irrespectively of the gate types (for the device with the same dimensions and at the same biases). Therefore, we can see that the short channel effects in the SNWFETs arise due to the direct source-to-drain tunneling current.

SB-MOSFETs show similar behavior below threshold. Figure 2b shows $I_d - V_g$ characteristics of SB devices with $\phi_{Bn} = 0.2$ eV when the gate length is gradually reduced from 30 nm down to 5 nm. Due to the increase of the direct source-to-drain tunneling current as in the case of the doped S/D devices, drastic deterioration of the off current behavior is noticed. The degree of degradation is severer for SB-MOSFETs, due to the presence of the SB barriers at the contacts as will be discussed shortly. On the other hand, the on-state currents are dominated by the tunneling current as seen in the inset of Fig. 2b contrary to the case of the doped S/D devices.

In a scaling of ballistic SNWFETs, the general trend is that the device performance improves as the channel length L becomes longer (for the same cross-sectional area) or as the channel cross-sectional area $W \equiv W_{si}$ is decreased (for the same channel length). If L and W are varied simultaneously while the aspect ratio L/W is fixed, the characteristics of SNWFETs are largely determined by their aspect ratio [28]. We show in Fig. 3 the device performance, as measured by its subthreshold swing, of SNWFETs with ohmic and Schottky contacts, respectively, for different aspect ratio L/W and different gate types. Compared to their planar counterparts, SNWFETs with multiple gates show enhanced performance, as expected, but it is noteworthy that the multi-gate effect is much greater for devices with smaller aspect ratio. Due to the tunneling current component in the off-state, SS of SB-MOSFETs are seen to be larger than their doped S/D coun-

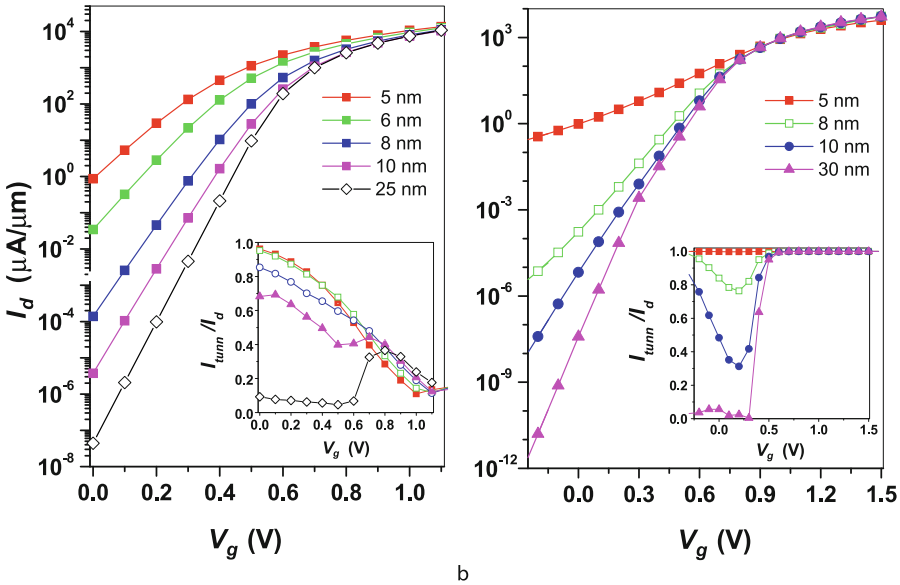

Quantum Simulations of Ballistic Nanowire Field Effect Transistors, Figure 2

$I_d - V_g$ characteristics of GAA SNWFETs with a doped S/D and b Schottky-barrier contacts. In the *insets* are shown the tunneling current contribution I_{tunn}/I_d

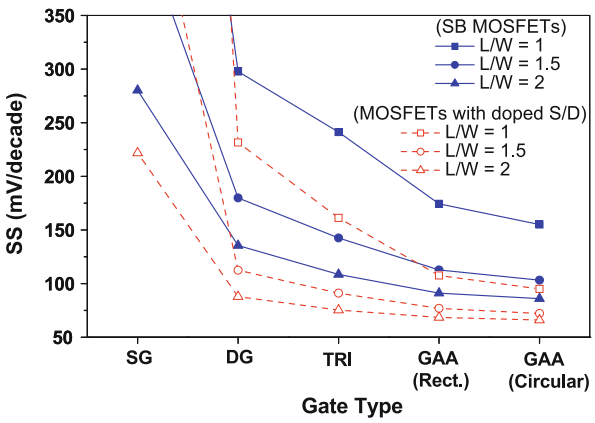

Quantum Simulations of Ballistic Nanowire Field Effect Transistors, Figure 3

Subthreshold slopes of SB-MOSFETs and MOSFETs with doped S/D, respectively, for different gate types and aspect ratios

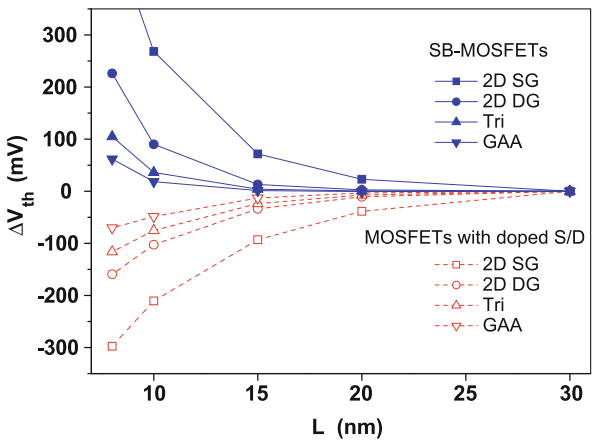

Quantum Simulations of Ballistic Nanowire Field Effect Transistors, Figure 4

Threshold-voltage change ΔV_{th} as a function of channel length L for SB-MOSFETs and MOSFETs with doped S/D. For the SB-MOSFETs, $\phi_{Bn} = 0.2$ eV was assumed

terparts. Notice that SS approaches the theoretical limit of 60 mV/decade for $L/W = 2$ in GAA devices with doped S/D.

Figure 4 shows the change of the threshold voltage ΔV_{th} from 30 nm device for SB and doped S/D devices, respectively, as the channel length is gradually shortened. A sharp difference between the threshold behaviors of SB and doped S/D devices is noticed in the figure: as the channel length is gradually decreased from $L = 30$ nm, ΔV_{th} increases for the SB devices while it decreases for the doped S/D devices. The increasing ΔV_{th} behavior of the SB devices can be explained by the increasing CB bending stiffness: that is, as the channel length becomes shorter, CB bending by the gate voltage becomes less effective, which means that more gate voltage should be applied to enter the on-state and as the consequence, the threshold volt-

age is increased. For the doped S/D devices, CB also becomes stiffer to bending by the gate voltage as the channel length becomes shorter. This however does not lead to increase of the threshold voltage, because the top of the barrier is gradually lowered as the gate voltage is increased in the case of the doped S/D devices (and the off-state top-of-the-barrier is lower for shorter-channel devices, so less gate voltage is needed to be applied to enter the on-state), while it is clamped by the Schottky barriers in the case of the SB devices. An advantage of multiple-gated devices can be noticed in Fig. 4. Compared to the case of 2DSG devices, the threshold-voltage change to the channel length of the GAA devices is considerably suppressed.

Simulation Results of CNTFETs

Some simulation results for CNTFETs such as the device-type and gate-type dependence are presented in the following. CNTFETs with (13,0) zigzag nanotube with diameter of ~1 nm and band gap energy of ~0.83 eV are considered. Thickness and dielectric constant of the gate oxide are 1 nm and 25, respectively. The channel is assumed to be intrinsic, and the midgap Schottky barrier is assumed for SB CNTFETs while the source/drain regions are n-doped with 10^7 cm^{-1} for CNT MOSFETs.

The transfer characteristics of coaxial SB-CNTFETs as the channel length is scaled down from 20 nm to 5 nm are shown in Fig. 5a. For $L \gtrsim 10$nm, $I_d - V_g$ characteristics are almost same regardless of the channel length because the CNT channel is ballistic and the current is dominated, and limited at the same time, by the tunneling current at the source SB contact. As the device is aggressively scaled down below 10 nm in the channel length, subthreshold currents increase sharply, due to the increased source-to-drain tunneling current as in the case of SNWFETs, and for $L < 5$ nm, the on-off current ratio becomes less than 10^2. Note that ambipolar conduction occurs in the SB-CNTFETs owing to the mirror-symmetric conduction and valence band structures and current is at its minimum at $V_g = V_d/2$ for midgap Schottky barrier height.

Similar scaling behavior can be observed in coaxial CNT MOSFETs, as shown in Fig. 5b. The seemingly ambipolar behavior is however due to band-to-band (BTB) tunneling at negatively high gate voltages and the current minima take place when BTB tunneling begins to be initiated. Compared to SB-CNTFETs, on-current level and on-off current ratio are higher for the CNT MOSFETs, because the thermionic currents prevail in the on-state and the minority carrier injection is suppressed unlike the SB devices. Notice also that SS of the electron branch (the right branch in the figure) approaches the theoretical limit of 60 mV/decade for conventional MOSFETs while that of the BTB branch is less than that because of the BTB tunneling. As the channel length is scaled down, however, the

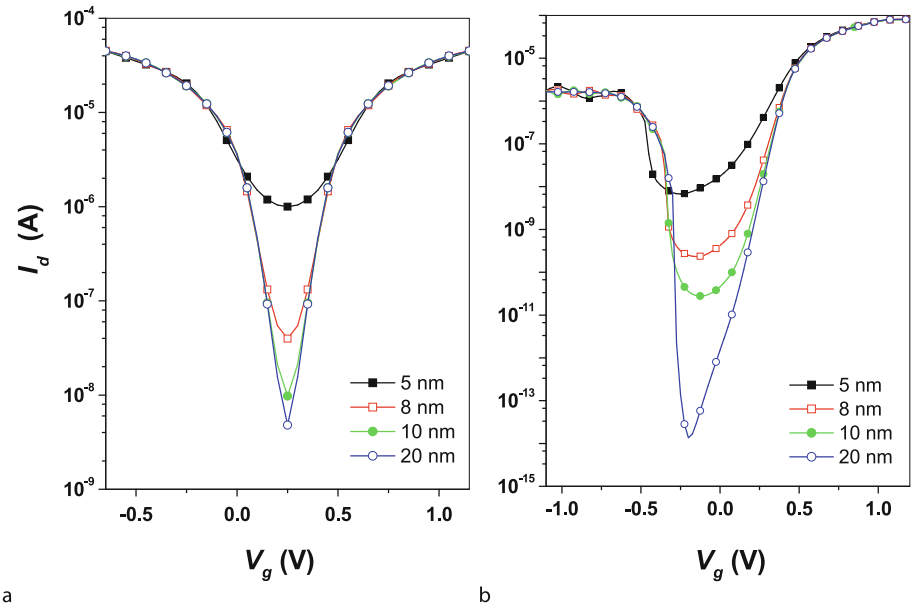

Quantum Simulations of Ballistic Nanowire Field Effect Transistors, Figure 5
$I_d - V_g$ characteristics of coaxial CNTFETs for various channel lengths: a SB-MOSFETs and b CNT MOSFETs

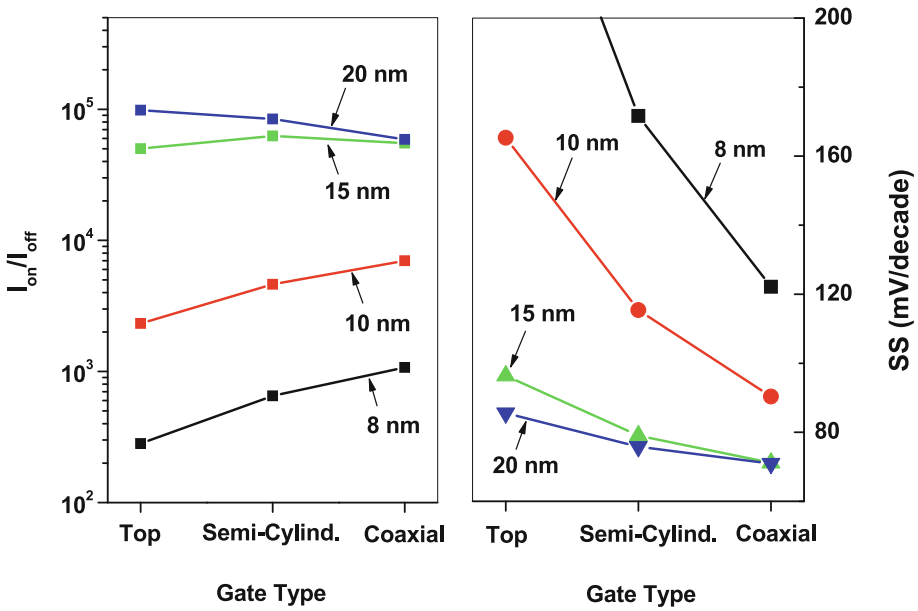

Quantum Simulations of Ballistic Nanowire Field Effect Transistors, Figure 6
I_{on}/I_{off} ratios and subthreshold slopes for top, semi-cylindrical, and coaxial gated CNT MOSFETs with different channel lengths

device performance becomes degraded as can be seen in Fig. 5b but compared to SB devices, the degree of degradation is less severe.

The gate-type dependence of CNTFETs with Schottky barrier contacts is shown in Fig. 6, where SS and I_{on}/I_{off} of the top, semi-cylindrical, and co-axial gated devices, respectively, are shown for different channel lengths. As expected, device performance is improved as the gate number is increased, especially for devices with sub-10 nm channel length.

Future Directions

A future practical device simulator should possess multi-dimensional, multi-scale features. For simulations of silicon MOSFET-type devices, especially, it should support both planar and wire structures in terms of dimensionality and should encompass a few nanometer to a few micrometer in terms of device size. Efforts should be made to develop a simulator that is most efficient in each regime of interest and seamlessly connects different regimes at the same time. Ballistic nanowire transistor simulation as considered in this article covers only a small portion of the whole area: The ballistic transport regime should be expanded and smoothly transit to diffusive transport regime in one direction and simplified approaches based on the effective-mass theory or single-band tight-binding method as adopted here should be connected with fully atomistic or first-principle calculations in the other direction. For the former, mode-space approach the with one-dimensional subbands where transport can be treated via NEGF, Boltzmann equation or Wigner function approaches seems to be quite attractive as recent works reveal. For the latter, hybrid approach that captures essential physics from rigorous calculations and enables efficient computation at the same time should be further elaborated in such a way to provide an integrated platform for a practical device simulator.

Acknowledgment

This research was supported in part by the university collaboration program of KRISS and in part by grant No. R 01-2005-000-10303-0 from the Basic Research Program of the Korea Science and Engineering Foundation.

Bibliography

1. Abramo A, Cardin A, Selmi L, Sangiorgi E (2000) Two-dimensional quantum mechanical simulation of charge distribution in silicon MOSFETs. IEEE Trans Electron Dev 27:1858–1863
2. Ahn C, Shin M (2007) Quantum simulation of coaxially gated CNTFETs by using an effective mass approach. J Korean Phys Soc 50:1887–1893
3. Alam K, Lake RK (2005) Leakage and performance of zero-Schottky-barrier carbon nanotube transistors. J Appl Phys 98:064037-1–064037-8

4. Appenzeller J, Lin YM, Knoch J, Chen Z, Avouris P (2005) Comparing carbon nanotube transistors – the ideal choice: a novel tunneling device design. IEEE Trans Electron Dev 52: 2568–2576
5. Avouris P, Appenzeller J, Martel R, Wind SJ (2003) Carbon nanotube electronics. Proc of the IEEE 91:1772–1784
6. Bescond M, Nehari K, Autran JL, Cavassilas N, Munteanu D, Lannoo M (2004) 3D quantum modeling and simulation of multiple-gate nanowire MOSFETs. IEDM Tech Dig pp 617–620
7. Cui Y, Zhong Z, Wang D, Wang W, Lieber M (2003) High performance silicon nanowire field effect transistors. Nano Lett 3:149–152
8. Damle PS, Ghosh AW, Datta S (2003) Nanoscale device modeling. In: Reed M et al (eds) Molecular nanoelectronics, American Scientific Publishers, Stevenson Ranch, pp 115–135
9. Datta S (2000) Nanoscale device modeling: the Greens function method. Superlatt Microstruct 28:253–278
10. Doris B, Ieong M, Kanarsky T, Zhang Y, Roy RA, Dokumaci O, Ren Z, Jamin FF, Shi L, Natzle W, Huang HJ, Mezzapelle J, Mocuta A, Womack S, Gribelyuk M, Jones EC, Miller RJ, Wong HSP, Haensch W (2002) Extreme scaling with ultra-thin Si channel MOSFETs. IEDM Tech Dig pp 267–270
11. Fiori G, Iannaccone G, Klimeck G (2006) A three-dimensional simulation study of the performance of carbon nanotube field-effect transistors with doped reservoirs and realistic geometry. IEEE Trans Electron Dev 53:1782–1788
12. Gnani E, Reggiani S, Gnudi A, Parruccini P, Colle R, Rudan M, Baccarani G (2007) Band-structure effects in ultrascaled silicon nanowires. IEEE Trans Electron Devices 54:2243–2254
13. Guo J, Datta S, Lundstrom MS, Anantram MP (2004) Towards multi-scale modeling of carbon nanotube transistors. Int J Multiscale Comput Eng 2:257–276
14. Guo J, Lundstrom M (2005) Role of phonon scattering in carbon nanotube field-effect transistors. Appl Phys Lett 86:193103-1–193103-3
15. (2007) International technology roadmap for semiconductor edition. http://public.itrs.net
16. Javey A, Tu R, Farmer DB, Guo J, Gordon RG, Dai H (2005) High performance n-type carbon nanotube field-effect transistors with chemically doped contacts. Nano Lett 5:345–348
17. Jin S, Park YJ, Min HS (2006) A three-dimensional simulation of quantum transport in silicon nanowire transistor in the presence of electron-phonon interactions. J App Phys 99:123719-1–123719-10
18. John DL, Castro LC, Pereira PJS, Pulfrey PL (2004) A Schrödinger-Poisson solver for modeling carbon nanotube FETs. Proc NSTI Nanotech 3:65–68
19. John DL (2006) Simulation Studies of Carbon Nanotube Field Effect Transistors. University of British Columbia, Vancouver
20. Ko YJ, Shin M, Lee S, Park KW (2000) Effects of atomistic defects on coherent electron transmission in Si nanowires: Full band calculations. J Appl Phys 89:374–379
21. Koswatta SO, Neophytou N, Kienle D, Fiori G, Lundstrom MS (2006) Dependence of DC characteristics of CNT MOSFETs on bandstructure models. IEEE Trans Nanotechnol 5:368–372
22. Luisier M, Schenk A, Fichtner W (2006) Quantum transport in two- and three-dimensional nano-scale transistors: Coupled mode effects in the nonequilibrium Greens function formalism. J App Phys 100:043713-1–043713-12
23. Luisier M, Schenk A, Fichtner W (2006) Atomistic simulation of nanowires in the sp3d5s* tight-binding formalism: From boundary conditions to strain calculations. Phys Rev B 74:205323-1–205323-12
24. Nehari K, Cavassilas N, Autran JL, Bescond M, Munteanu D, Lannoo M (2005) Influence of band-structure on electron ballistic transport in silicon nanowire MOSFET's: an atomistic study. Proc 35th European Solid-State Device Research Conf, pp 229–232
25. Park JT, Colinge JP (2002) Multiple-gate SOI MOSFETs: device design guidelines. IEEE Trans Electron Devices 12:2222–2229
26. Radosavljevic M, Heinze S, Tersoff J, Avouris P (2003) Drain voltage scaling in carbon nantoube transistors. Appl Phys Lett 83:2435–2437
27. Rahman A, Lundstrom MS, Ghosh AW (2005) Generalized effective-mass approach for n-type metal-oxide-semiconductor field-effect transistors on arbitrarily oriented wafers. J Appl Phys 97:053702-1–053702-12
28. Shin M (2007) Efficient simulation of silicon nanowire field effect transistors and their scaling behavior. J Appl Phys 101:024510-1–024510-6
29. Shin M (2007) Quantum simulation of device characteristics of silicon nanowire FETs. IEEE Trans Nanotechnol 6:230–237
30. Wang J, Polizzi E, Lundstrom M (2004) A three-dimensional quantum simulation of silicon nanowire transistors with the effective-mass approximation. J Appl Phys 96:2192–2203
31. Wang J, Rahman A, Ghosh A, Klimeck G, Lundstrom M (2005) On the validity of the parabolic effective-mass approximation for the I-V calculation of silicon nanowire transistors. IEEE Trans Electron Devices 52:1589–1595
32. Wang J, Polizzi E, Ghosh A, Datta S, Lundstrom M (2005) Theoretical investigation of surface roughness scattering in silicon nanowire transistors. Appl Phys Lett 87:043101-1–043101-3
33. Wong HSP (2005) Beyond the conventional transistor. Solid State Electron 49:755–762

R

Random Graphs, A Whirlwind Tour of

FAN CHUNG
University of California, San Diego, USA

Article Outline

Glossary
Introduction
Some Basic Graph Theory
Random Graphs in a Nutshell
Classical Random Graphs
Random Power Law Graphs
On-Line Random Graphs
Remarks
Bibliography

Glossary

Random graph a graph which is chosen from a certain probability distribution over the set of all graphs satisfying some given set of constraints. Statements about random graphs are thus statements about "typical" graphs for the given constraints.

Graph diameter the maximum distance between any pair of vertices in a graph. Here the distance between two vertices is the length of any shortest path joining the vertices.

Node degree distribution n_k the number of vertices having degree k.

Power law distribution a node degree distribution which (at least approximately) obeys $n_k \propto k^{-\beta}$ with fixed positive exponent β.

Erdős–Rényi random graph $\mathcal{G}(n, p)$ a graph with n vertices, in which each possible edge between the $\binom{n}{2}$ pairs of vertices is present with probability p. The fact that the edges are chosen independently makes it relatively easy to compute properties of $\mathcal{G}(n, p)$.

Random graph model $\mathcal{G}(\mathbf{w})$ for expected degree sequence $\mathbf{w}$:] a model which generates random graphs which, *on average*, will have a prescribed degree sequence $\mathbf{w} = (w_1, w_2, \ldots, w_n)$.

On-line random graph a graph which grows in size over time, according to given probabilistic rules, starting from a start graph G_0. One can make statements about these graphs in the limit of long time (and hence large vertex number n, as n approaches infinity).

Introduction

Nowadays we are surrounded by assorted large information networks. For example, the phone network has all users as vertices which are interconnected by phone calls from one user to another. The Web can be viewed as a network with webpages as vertices which are then linked to other webpages. There are various biological networks arising from numerous databases, such as the gene network which represents the regulatory effect among genes. Of interest are many social networks expressing various types of social interactions. Some noted examples include the Collaboration graph (denoting coauthorship among mathematicians) and the Hollywood graph (consisting of actors/actresses and their joint appearances in feature films), among others.

How are these networks formed? What are basic structures of such large networks? How do they evolve? What are the underlying principles that dictate their behaviors?

To answer these questions, graph theory comes into play. Random graphs have a similar flavor as these large information networks in a natural way. For example, the phone network is formed by making random phone calls while a random graph results from adding a random edge one at a time. Although the classical random graphs can not directly be used to model real networks and seem to exhibit different 'shapes', the methods and approaches in random graph theory provides useful tools for the modeling and analysis of these information networks.

In this article, we will start with some basic graph theory in Sect. "Some Basic Graph Theory". We then introduce the main themes of random graphs in Sect. "Random

Graphs in a Nutshell". Then we consider the classical random graph theory in Sect. "Classical Random Graphs" before we proceed to describe some general random graph models with given degree distributions, in particular, the power law graphs in Sect. "Random Power Law Graphs". In Sect. "On-Line Random Graphs", we will cover two types of "on-line" graph models, including the model of preferential attachment and the duplication model.

Although random graphs can be used to analyze various aspects of realistic networks, we wish to point out that there is no silver bullet to answer all the difficult problems about these large complex networks. In the last section we will put things in perspective by clarifying what random graphs can and can not do.

Some Basic Graph Theory

All the information networks that we have mentioned can be formulated in terms of graphs. A graph G consists of a vertex set, denoted by $V = V(G)$ (which contains all the objects that we wish to deal with) and an edge set $E = E(G)$ which consists of of specified pairwise relations between vertices. For example, a friendship graph has the vertex set consisting of people of interest and the edge set denoting the pairs of people who are friends. In Table 1 we list a number of graphs associated with various networks.

As an introduction to graph theory, we describe the so-called *party problem*:

> *Among six people in a party, show that there are at least three people who know each other or there are three people who do not know each other.*

This can be said in graph-theoretical terms:

> *Any graph on 6 vertices must contain a triangle or contain three independent vertices with no edge among them.*

Indeed, 6 is the smallest number for this to occur since there is a graph on 5 vertices that contain neither a triangle nor three independent vertices. Such a graph is a cycle on 5 vertices, denoted by C_5, as seen in Fig. 1.

Random Graphs, A Whirlwind Tour of, Table 1
Graph models for several networks

Graph	Vertices	Edges
Flight schedule graph	Cities	Flights
Phone graph	Telephone numbers	Phone calls
Collaboration graph	Authors in Math Review	Coauthorship
Web graph	Webpages	Links
Biological graph	Genes	Regulatory effects

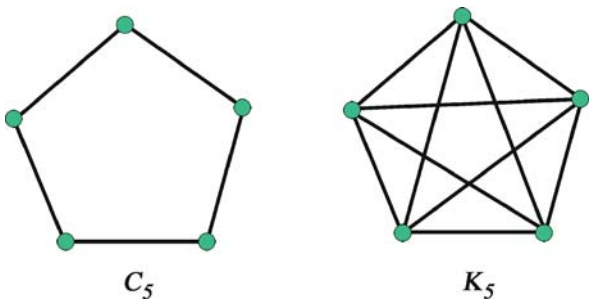

Random Graphs, A Whirlwind Tour of, Figure 1
A five cycle C_5 and a complete graph K_5

Let K_n denote a complete graph on n vertices which has all $\binom{n}{2}$ edges. For example, a triangle is K_3 which turns out is also C_3. The above party problem is a toy case of the so-called *Ramsey theory* which deals with unavoidable patterns in large graphs. In 1930, Ramsey [40] showed the following:

> *For any two positive integers k and l, there is an associated number $R(k, l)$ such that any graph on $n \geq R(k, l)$ vertices must contain either K_k as a subgraph or contain l independent vertices.*

For example, $R(3, 3) = 6$ as stated in the party problem. It is not too difficult to show that $R(4, 4) = 17$. However, the value of $R(5, 5)$ is not yet determined (in spite of the huge computational power we have today). All that is known is $43 \leq R(5, 5) \leq 49$ (see [25] and [36]). Relatively few exact Ramsey numbers $R(k, l)$ are determined. For an extensive survey on this topic, the reader is referred to the dynamic survey in the Electronic Journal of Combinatorics at http://www.combinatorics.org/.

In 1947, Erdős wrote an important paper [21] that helped start two areas including combinatorial probabilistic methods and Ramsey theory. He established the following lower bound for the Ramsey number $R(k, k)$ by proving

$$R(k, k) \geq 2^{\frac{k}{2}}, \tag{1}$$

the argument is quite simple and elegant:

Suppose we wish to find a graph on n vertices that does not contain K_k or an independent subset of k vertices. How large can n be? For a fixed integer n, there are all together $2^{\binom{n}{2}}$ possible graphs on n vertices. We say a graph is *bad* if it contains K_k or an independent subset of k vertices. How many bad graphs can there be? There are $\binom{n}{k}$

ways to choose k out of n vertices. So, there are at most $2\binom{n}{k}2^{\binom{n}{2}-\binom{k}{2}}$ bad graphs. Therefore there is a graph on n vertices that is not bad if

$$2^{\binom{n}{2}} \geq 2\binom{n}{k}2^{\binom{n}{2}-\binom{k}{2}} .$$

So, for $n \geq 2^{k/2}$, there must be a graph on n vertices that is not bad, which implies (1).

We note that for the upper bound there is an inductive proof to show that $R(k,k) \leq \binom{2k-2}{k-2}$ which is about 4^k. In the previous five decades, there have been some improvements only by a factor of a lower order for both the upper and lower bounds [19,42]. It remains unsettled (with Erdős award unclaimed) to determine if $\lim_{k\to\infty}(R(k,k))^{1/k}$ exists or what value it should be.

A basic notion in graph theory is "adjacency". A vertex u is said to be *adjacent* to another vertex v if $\{u,v\}$ is an edge. Or, we say u is a *neighbor* of v. Equivalently, v is a neighbor of u. The *degree* of a vertex u is the number of edges containing u. If we restrict ourselves to *simple* graphs (i. e., at most one edge between any pairs of vertices), then the degree of u is just the number of neighbors that u has. Suppose that in a graph G the vertex v_i has degree d_i for $1 \leq i \leq n$. Then $(d_1, d_2, \ldots, d_n)$ forms a degree sequence for G. Sometimes, we organize the degree sequence so that $d_1 \geq d_2 \geq \cdots \geq d_n$. Here comes a natural question on graph realization: For what values d_i is the sequence $(d_1, d_2, \ldots, d_n)$ a degree sequence of some graph?

To answer this question, first we observe that the sum of all d_i's must be even since that is exactly twice the number of edges. This is the folklore "Handshake Theorem".

In a 1961 paper, Erdős and Gallai [24] answered the above question. They gave a necessary and sufficient condition by showing that a sequence $(d_1, d_2, \ldots, d_n)$, where $d_i \geq d_{i+1}$, is a degree sequence of some graph if and only if the sum of d_i's is even and for each integer $r \leq n-1$,

$$\sum_{i=1}^{r} d_i \leq r(r-1) + \sum_{i=r+1}^{n} \min\{r, d_i\} .$$

Another way to keep track of the degrees of a graph is to consider the degree distribution as follows: Let n_k denote the number of vertices having degree k. Instead of writing down the degree sequence (which consists of n numbers and n can be a very large number), we just use n_k. Therefore, the number of values that we need to keep does not exceed the maximum degree. If all degrees are the same value, we say the graph is regular. In this case, only one of the n_k's is nonzero.

Many real-world graphs have degree distribution satisfying the so-called "*power law*". Namely, the number n_k of vertices of degree k is proportional to $k^{-\beta}$ for some fixed positive value β. For example, the Collaboration graph, as illustrated in Fig. 2, can be approximated by a power law with exponent $\beta = 2.46$. The degree distribution of the Collaboration graph is included in Fig. 3 in log-log scale.

In a graph G, a *path* is a sequence of vertices v_0, $v_1, \ldots, v_k$ such that v_{i-1} is adjacent to v_i for $i = 1, \ldots, k$. The length of a path is the number of edges in the path. For example, the above mentioned path has length k join-

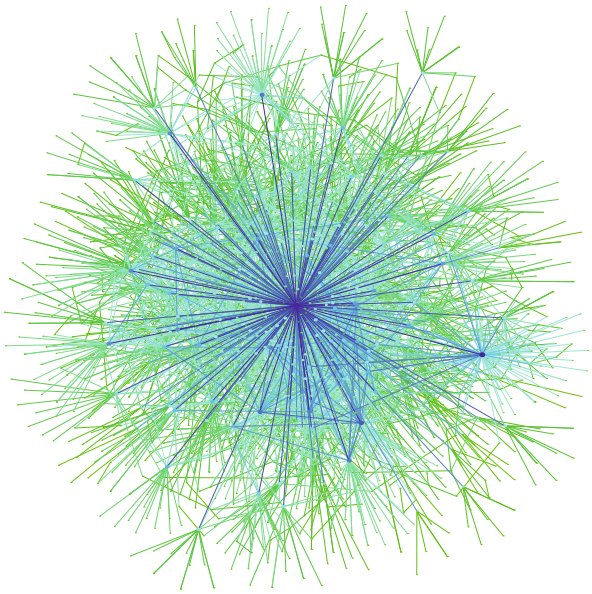

Random Graphs, A Whirlwind Tour of, Figure 2
The collaboration graph

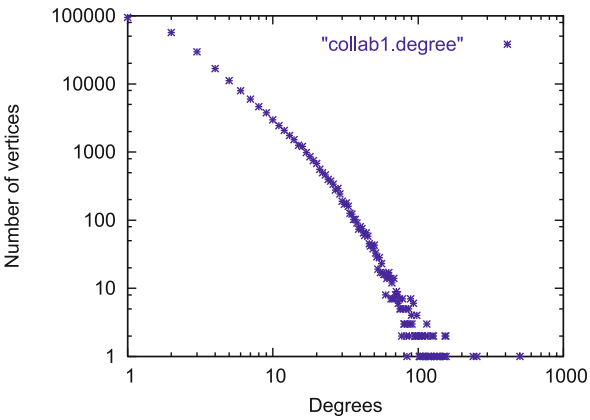

Random Graphs, A Whirlwind Tour of, Figure 3
The number of vertices for each possible degree for the collaboration graph

ing v_0 and v_k. If $v_0 = v_k$, the path is said to be a *cycle*. A graph which contains no cycle is called a *tree*. A graph is *connected* if any two vertices can be joined by a path. For a graph G, a maximum subset of vertices each pair of which can be joined by paths is called a *connected component*. Thus, a graph is connected if there is only one connected component. In a connected graph, the *distance* between two vertices u and v is the length of a shortest path joining u and v. The maximum distance among all pairs of vertices is called the *diameter* of a graph.

In 1967, the psychologist Stanley Milgram [37] conducted a series of experiments which indicated that any two strangers are connected by a chain of intermediate acquaintances with the average chain length about six. Since then, the so-called "*small world phenomenon*" has long been a subject of anecdotal observation and folklore. Recent studies have suggested that the phenomenon is pervasive in numerous networks arising in nature and technology, and in particular, in the structural evolution of the World Wide Web [5,33,44].

In addition to "six degrees of separation", various numbers have emerged with many networks. In 1999, Barabási et al. [5] estimated that any two webpages are at most 19 clicks away from one another (in certain models of the Internet). Broder et al. [11] set up crawlers on a webgraph of 200 million nodes and 1.5 billion links and reported that the average path length is about 16. A mathematician who has written joint papers is likely to have Erdős number at most eight [29] (i. e., with a chain of coauthors with length at most 8 connecting to Erdős). The majority of actors or actresses have the so-called "Kevin Bacon number" two or three.

Before we make sense of these numbers, some clarification is in order: There are in fact two different interpretations of 'short distance' in a network. One notion is the diameter of the graph. Another notion is the *average distance* (which might be closer to what was meant by these experiments). We will discuss the small world phenomenon further in a later section.

Random Graphs in a Nutshell

What does a "random graph" mean? Before proceeding to describe random graphs, some clarification for "random" is in order. According to the Cambridge Dictionary, "random" means "happening, done or chosen by chance rather than according to a plan". Quite contrary to this explanation, our random graphs have precise meanings and can be clearly defined. Using the terminology in probability, a random graph is a random variable defined in a probability space with a probability distribution. In layman's terms, we first put all graphs on n vertices in a lottery box and then the graph we pick out of the box is a random graph. (In this case, all graphs are chosen with equal probability.)

What do we want from our random graphs? Well, we would like to say that a random graph (in some given model) has certain properties (e. g., having small diameter). Such a statement means that with probability close to 1 (as the number n of vertices approaches infinity), the random graph we pick out of the lottery box satisfies the property that we specified. In other words, a random graph has a specified property means that almost all graphs of interest have the desired property. Note that this is quite a strong implication! Any statement about a random graph is really about almost all graphs! The beauty of random graphs lies in being able to use relatively few parameters in the model to capture the behavior of almost all graphs of interest (which can be quite numerous and complex).

In the early days of the subject, Erdős and Rényi introduced two random graph models. The first one is a random graph $\mathcal{F}(n, m)$ defined on all graphs with n vertices and m edges each of which is chosen with equal probability. The second is the celebrated Erdős–Rényi random graph $\mathcal{G}(n, p)$ defined on all graphs on n vertices and each edge is chosen independently with probability p. Consequently, a graph on n vertices and x edges is chosen with probability $p^x(1-p)^{\binom{n}{2}-x}$ in $\mathcal{G}(n, p)$.

The advantage of the Erdős–Rényi model is the independence of choices for the edges (i. e., each pair of vertices has its own dice for determining being chosen as an edge). Since the probability of two independent events is the product of probabilities of two events, we can compute with ease. For example, the probability of a random graph in $\mathcal{G}(n, p)$ containing a fixed triangle is 1/8 for $p = 1/2$. It is possible to compute such a probability for a random graph in $\mathcal{F}(n, m)$, e. g., $\binom{\binom{n}{2}-3}{m-3}/\binom{\binom{n}{2}}{m}$, which is a more complicated expression. For many problems, such as the diameter problem, it can be quite nontrivial for $\mathcal{F}(n, m)$ because the dependency among edges is getting in the way.

To model real graphs, there are some obvious difficulties. For example, the random graph $\mathcal{G}(n, p)$ has all degrees very close to pn if the graph is not so sparse, (i. e., $p \geq \log n/n$). The distribution of the degrees follows the same bell curve for every vertex. As we know, many real-world graphs satisfy the power law which is very different from the degree distribution of $\mathcal{G}(n, p)$. In order to model real-world networks, it is imperative to consider random graphs with general degree distribution and, in particular, the power law distribution.

There are basically two types of random graph models for general degree distributions. The *configuration model* is a take-off from random regular graphs [6]. The way to define random regular graphs $\mathcal{G}_k$ of degree k on n vertices is to consider all possible matchings in a complete graph K_{kn}. Note that a matching is a maximum set of vertex-disjoint edges. Each matching is chosen with equal probability. We then get a random k-regular graph by partitioning the vertices of k into subsets. Each k-subset then is associated with a vertex in a random regular graph $\mathcal{G}_k$. Although such a random regular graph might contain loops (i. e., an edge having both endpoints the same vertex), the probability of such an event is of a lower order and can be controlled. It is then obvious to define random graphs with general degrees. Instead of partitioning the vertex set of the large graph into equal parts, we choose a random matching of a complete graph on $\sum_i d_i$ vertices which are partitioned into subsets of sizes $d_1, d_2, \ldots, d_n$. Then we form the random graph by associating each edge in the matching with an edge between associated vertices.

In the configuration model, there are nontrivial dependencies among the edges. There is another random graph model for general expected degrees as a generalization of the Erdős–Rényi model. Let $\mathbf{w} = (w_1, w_2, \ldots, w_n)$ denote the specified degrees. The $\mathcal{G}(\mathbf{w})$ model yields random graphs with expected degrees $\mathbf{w}$. The edge between v_i and v_j is independently chosen with probability $w_i w_j / W$ where $W = \sum_i w_i$. In other words, each pair of vertices has its own dice with probability assigned so that the expected degree at vertex v_i is exactly w_i. The Erdős–Rényi model is just the case with all w_i's equal to pn. Since the $\mathcal{G}(\mathbf{w})$ model inherits the robustness and independence of the Erdős–Rényi model, many strong properties can be derived. We will discuss some of these further in Sect. "Parameters for Modeling Power Law Graphs", especially when $\mathbf{w}$ satisfies power laws.

All the random graph models mentioned above are *off-line* models. Since real-world graphs are dynamically changing—both in adding and deleting vertices and edges—there are several *on-line random graph models* in which the probability spaces are changing at the tick of the clock. In fact, in the study of complex real-world graphs, the on-line model came to attention first.

There are a large number of research papers, surveys and books on random graphs, mostly about the Erdős–Rényi model $\mathcal{G}(n, p)$. After the year 2000, the study of real-world graphs has led to interesting directions and new methods for analyzing random graphs with general degree distributions. Many on-line models have been proposed and published. Here we will only be able to cover the main ones—the preferential attachment schemes and the duplication model later.

Classical Random Graphs

In early 60's, Erdős and Rényi wrote a series of influential papers on random graphs. Their modeling and analysis are thorough and elegant. Their approaches and methods are powerful and have had enormous impact up to this day. In this section, we will give a brief overview. First we will describe the classical results on the evolution of random graphs $\mathcal{G}(n, p)$ of the Erdős–Rényi model. Then we will discuss the diameter of $\mathcal{G}(n, p)$ as the edge density ranges from 0 to 1.

The Evolution of the Erdős–Rényi Graph

What does a random graph in $\mathcal{G}(n, p)$ look like? Erdős and Rényi [23] gave a full answer for the edge density p ranging from 0 to 1.

At the start, there is no edge and the edge density is 0. We have isolated vertices.

As p increases, the expected number $p\binom{n}{2}$ of edges gets larger. When there are about $\sqrt{n}$ edges, how many connected components are there and what sizes and structures are they? For $0 < p \ll 1/n$, Erdős and Rényi [23] showed that the random graph G is a disjoint union of trees. Furthermore, they gave a beautiful formula. For $p = cn^{-k/(k-1)}$, the probability that j is the number of connected components in G formed by trees on k vertices is $\lambda^j e^{-\lambda}/j!$ where $\lambda = (2c)^{k-1} k^{k-2}/k!$. For example, when we have about $\sqrt{n}$ edges, the probability that the random graph G contains j trees on 3 vertices is close to $2^j/j!$ (if n is large enough).

As we have more edges, cycles start to appear. When the graph has a linear number of edges, i. e., $p = c/n$, with $c < 1$, almost all vertices are in connected components of trees and there are only a small number of cycles. Namely, the expected number of cycles is $1/2 \log 1/(1-c) - c/2 - (c^2)/4$.

When a random graph has edges ranging from slightly below $n/2$ to slightly over $n/2$ edges, i. e., $p = (1 + o(1))/n$, there is an unusual phenomenon, called "*double jumps*". What are double jumps and why is it so unusual? In the study of "threshold function" or, "phase transition" that happens in natural or evolving systems, it is of interest to identify the critical point, below which the behavior is dramatically different from what is above. Erdős and Rényi [23] found that as p is smaller than $1/n$, the largest component in G has size $O(\log n)$ and all components are either trees or unicyclic (i. e., each component contains at most one cycle). If p is $(1 + \mu)/n$ and $\mu > 0$, then the

Random Graphs, A Whirlwind Tour of, Table 2
The diameter of random graphs $\mathcal{G}(n, p)$

Range	diam($\mathcal{G}(n, p)$)	Reference
$\frac{np}{\log n} \to \infty$	Concentrated on at most 2 values	[9]
$\frac{np}{\log n} = c > 8$	Concentrated on at most 2 values	[12]
$8 \geq \frac{np}{\log n} = c > 2$	Concentrated on at most 3 values	[12]
$2 \geq \frac{np}{\log n} = c > 1$	Concentrated on at most 4 values	[10]
$1 \geq \frac{np}{\log n} = c > c_0$	Concentrated on at most $2\lfloor \frac{1}{c_0} \rfloor + 4$ values	[12]
$\log n > np \to \infty$	$\mathrm{diam}(\mathcal{G}(n, p)) = (1 + o(1)) \frac{\log n}{\log(np)}$	[12]
$np \geq c > 1$	The ratio $\mathrm{diam}(\mathcal{G}_{n,p}) / \left(\frac{\log n}{\log(np)}\right)$ is finite (between 1 and $f(c)$)	[12]
$np < 1$	$\mathrm{diam}(\mathcal{G}(n, p))$ equals the diameter of a tree component if $(1 - np)n^{1/3} \to \infty$	[35]

giant component emerges. However, when $p = 1/n$, the largest component is of size $O(n^{2/3})$. There has been detailed analysis examining this tricky transition in details (see [7] and [31]).

When $p = c/n$ for $c > 1$, the random graph G has one giant component and all others are quite small, of size $O(\log n)$. Also, Erdős and Rényi [23] determined the number of vertices in the giant connected component to be $f(c)n$ where

$$f(c) = 1 - \frac{1}{c} \sum_{k=1}^{\infty} \frac{k^{k-1}}{k!} \left(ce^{-c}\right)^k . \tag{2}$$

Finally when $p = c \log n/n$ and $c > 1$, the random graph G is almost always connected. When c gets larger, G is not only connected but gets closer to a regular graph. Namely, all vertices have degrees close to pn.

The Diameter of the Erdős–Rényi Graph

We consider the diameter of a random graph G in $\mathcal{G}(n, p)$ for all ranges of p including the range for which $\mathcal{G}(n, p)$ is not connected. For a disconnected graph G, the diameter of G is defined to be the diameter of its largest connected component.

Roughly speaking, the diameter of a random graph in $\mathcal{G}(n, p)$ is of order $\log n/(\log(np))$ if the expected degree np is at least 1. Note that this is best possible in the following sense. For any graph with degrees at most d, the number of vertices that can be reached within distance k is at most $1 + d + d(d-1) + d(d-1)^2 + \cdots + d(d-1)^{k-1}$. This sum should be at least n if k is the diameter. Therefore we known that the diameter, denoted by $\mathrm{diam}(G)$ is at least $(\log n)/\log(d-1)$.

To be precise, it can be shown the diameter of a random graph G in $\mathcal{G}(n, p)$ is $(1 + o(1))(\log n)/(\log np)$ if the expected degree np goes to infinity as n approaches infinity. When $np \geq c > 1$, the diameter $\mathrm{diam}(G)$ is within a constant factor of $(\log n)/(\log np)$ where the constant depends only on c and is independent of n. When $np = c < 1$, the random graph is surely disconnected and $\mathrm{diam}(G)$ is equal to the diameter of a tree component.

In fact, the diameter of a graph G in $\mathcal{G}(n, p)$ is quite predictable as follows. The values for the diameter of $\mathcal{G}(n, p)$ is almost surely concentrated on at most two values around $(\log n)/(\log np)$ if $(np)/(\log n) = c > 8$. When $(np)/(\log n) = c > 2$, the diameter of $\mathcal{G}(n, p)$ is almost surely concentrated on at most three values. For the range $2 \geq (np)/(\log n) = c > 1$, the diameter of $\mathcal{G}(n, p)$ is almost surely concentrated on at most four values.

It is of particular interest to consider random graphs $\mathcal{G}(n, p)$ for the range of $np > 1$ and $np \leq c \log n$ for some constant c since this range includes the emergence of the unique giant component. Because of a phase transition in connectivity at $p = \log n/n$, the problem of determining the diameter of $\mathcal{G}(n, p)$ and its concentration seems to be difficult for certain ranges of p. If $(np)/(\log n) = c > c_0$ for any (small) constant c and c_0, then the diameter of $\mathcal{G}(n, p)$ is almost surely concentrated on finitely many values, namely, no more than $2\lfloor 1/(c_0) \rfloor + 4$ values.

These facts are summarized in Table 2 with references listed. As we can see from the table, numerous questions remain.

Random Power Law Graphs

Parameters for Modeling Power Law Graphs

A large realistic network usually has a huge number of parameters with complicated descriptions. By "modeling a realistic network", we mean cutting down the number of parameters to relatively few and still capture a good part of the character of the network.

To choose the parameters for modeling a real network, the exponent β of the power law is relatively easy to select.

We can plot the log-degree versus log-frequency table and choose a good approximation of the slope.

In a graph G, suppose that there are y vertices of degree x. Then G is considered to be a power law graph if x and y satisfy (or can be approximated by) the following equation:

$$\log y = \alpha - \beta \log x \,. \tag{3}$$

In other words, we have

$$|\,\{v|\deg(v) = x\}\,| \approx y = \frac{e^{\alpha}}{x^{\beta}} \,.$$

Basically, α is the logarithm of the volume of the graph and β can be regarded as the log-log growth rate of the graph.

To take a closer look of the degree distribution of a typical realistic graph, several impediments obviously exist.

(a) When we fit the power law model, there are discrepancies especially when the degree is very small or very large. There is almost always a heavy tail distribution at the upper range and there seems to be scattering at the lower range. For example, for the collaboration graph, should we or shouldn't we include the data point for isolated vertices (an author with no coauthors)? Should we stay with the largest component or include all small components (including the isolated vertices)?

(b) The power law states that the number of vertices of degree k is proportional to $k^{-\beta}$. We can approximate the number of vertices of degree k by the function $f(k) = ck^{-\beta}$ for some constant c. However, $f(k)$ is usually not an integer. By taking either the ceiling or floor of $f(k)$, some errors are inevitable. In fact, such errors are acute when k or $f(k)$ is small.

(c) The power law model is usually a better fit in the middle range (than at either end). Still, in many examples, there is a visible slight "hump" in the curve instead of the straight line representing the power law in the log-log table.

Among the above three points, (c) is mainly due to first-order approximations. The straight line with slope β is a linear approximation of the actual plotted data. Thus the power law model is an important and necessary step for more complicated real cases. Here, we will first discuss (b) and then (a).

Item (b) concerns rounding errors which can be checked by the following basic calculations about the power law graphs as described by Eq. (3).

(1) The maximum degree of the graph is at most $e^{\alpha/\beta}$. Note that $0 \le \log y = \alpha - \beta \log x$.

(2) The number of vertices n can be computed as follows (under the assumption that the maximum degree is $e^{\alpha/\beta}$). By summing $y(x)$ for x from 1 to $e^{\alpha/\beta}$, we have

$$n = \sum_{x=1}^{e^{\frac{\alpha}{\beta}}} \frac{e^{\alpha}}{x^{\beta}} \approx \begin{cases} \zeta(\beta)e^{\alpha} & \text{if } \beta > 1\,, \\ \alpha e^{\alpha} & \text{if } \beta = 1\,, \\ \frac{e^{\frac{\alpha}{\beta}}}{1-\beta} & \text{if } 0 < \beta < 1\,, \end{cases}$$

where $\zeta(t) = \sum_{n=1}^{\infty} 1/(n^t)$ is the Riemann Zeta function.

(3) The number of edges E can be computed as follows:

$$E = \frac{1}{2}\sum_{x=1}^{e^{\frac{\alpha}{\beta}}} x\frac{e^{\alpha}}{x^{\beta}} \approx \begin{cases} \frac{1}{2}\zeta(\beta-1)e^{\alpha} & \text{if } \beta > 2\,, \\ \frac{1}{4}\alpha e^{\alpha} & \text{if } \beta = 2\,, \\ \frac{1}{2}\frac{e^{\frac{2\alpha}{\beta}}}{2-\beta} & \text{if } 0 < \beta < 2\,. \end{cases}$$

(4) The differences of the real numbers in (1)–(3) and their integer parts can be estimated as follows: For the number n of vertices, the error term is at most $e^{\alpha/\beta}$. For $\beta \ge 1$, it is $o(n)$, which is a lower order term. For $0 < \beta < 1$, the error term for n is relatively large. In this case, we have

$$n \le \frac{e^{\frac{\alpha}{\beta}}}{1-\beta} - e^{\frac{\alpha}{\beta}} = \frac{\beta e^{\frac{\alpha}{\beta}}}{1-\beta} \,.$$

As can be seen, n can have the same magnitude as $(e^{\alpha/\beta})/(1-\beta)$. Therefore the rounding error can be of the same order of magnitude. For the number E of edges, similar situations occur. For $\beta \ge 2$, the rounding error term of E is $o(E)$, a lower order term. For $0 < \beta < 2$, the error of E has the same magnitude as in the formula of item (3). Thus, one is advised to exercise caution when dealing with the case $0 < \beta < 2$.

To deal with the concerns mentioned above in (a), we need additional parameters.

- The average degree w is a useful parameter.
- The second order average degree $\tilde{w} = \sum_i w_i^2 / \sum_i w_i$.
- The maximum degree $m = d_{\max}$ and also the minimum degree $d_{\min}$ denote the range that the power law degree distribution fits (within acceptable approximation). In other words, the maximum degree $m = d_{\max}$ and the minimum degree $d_{\min}$ are meant to be the largest and the least degrees in a power law subgraph of G. Often, $d_{\min}$ is taken to be 1 unless otherwise specified.

With these parameters, we are ready to define a random power law graph. For random graphs with given expected degree sequences satisfying a power law distribution with exponent β, we may assume that the expected degrees

are $w_i = ci^{-\frac{1}{\beta-1}}$ for i satisfying $i_0 \le i < n + i_0$. Here c depends on the average degree and i_0 depends on the maximum degree m, namely, $c = (\beta-2)/(\beta-1)\, wn^{1/(\beta-1)}$, $i_0 = n((w(\beta-2))/(m(\beta-1)))^{\beta-1}$.

The power law graphs with exponent $\beta > 3$ are quite different from those with exponent $\beta < 3$ as evidenced by the value of $\tilde{w}$ (assuming $m \gg w$).

$$\tilde{w} = \begin{cases} (1+o(1))w\frac{(\beta-2)^2}{(\beta-1)(\beta-3)} & \text{if } \beta > 3\,, \\ (1+o(1))\frac{1}{2}w\ln\frac{2m}{w} & \text{if } \beta = 3\,, \\ (1+o(1))d^{\beta-2}\frac{(\beta-2)^{\beta-1}m^{3-\beta}}{(\beta-1)^{\beta-2}(3-\beta)} & \text{if } 2 < \beta < 3\,. \end{cases}$$

The above values of $\tilde{w}$ are quite useful in the study of average distance and diameter of random graphs.

The Evolution of Random Power Law Graphs

A natural question concerning the configuration model is how the random graphs evolve for power law distributions. Can we mimic the classical analysis as in the Erdős–Rényi random graph model?

Here we consider a configuration model with degree distribution as in the (α, β)-graph. As it turns out, the evolution only depends on β and not on α as follows.

1. When $\beta > \beta_0 = 3.47875\ldots$, the random graph almost surely has no giant component where the value $\beta_0 = 3.47875\ldots$ is a solution to

   $$\zeta(\beta-2) - 2\zeta\beta - 1) = 0\,.$$

   When $\beta < \beta_0 = 3.47875\ldots$, there is almost surely a unique giant component.
2. When $2 < \beta < \beta_0 = 3.47875\ldots$, the second largest component is almost surely of size $\Theta(\log n)$. For any $2 \le x < \Theta(\log n)$, there is almost surely a component of size x.
3. When $\beta = 2$, almost surely the second largest component is of size $\Theta(\log n/(\log\log n))$. For any $2 \le x < \Theta(\log n/(\log\log n))$, there is almost surely a component of size x.
4. When $1 < \beta < 2$, the second largest component is almost surely of size $\Theta(1)$. The graph is almost surely not connected.
5. When $0 < \beta < 1$, the graph is almost surely connected.
6. When $\beta = \beta_0 = 3.47875\ldots$, the situation is complicated. It is similar to the double jump of the random graph $\mathcal{G}(n,p)$ with $p = 1/n$. For $\beta = 1$, there is a nontrivial probability for either case that the graph is connected or disconnected.

A useful tool in configuration model is a result of Molloy and Reed [38,39]:

For a random graph with $(\gamma_i + o(1))n$ vertices of degree i, where γ_i are nonnegative values which sum to 1 and n is the number of vertices, the giant component emerges when $Q = \sum_{i\ge1} i(i-2)\gamma_i > 0$, provided that the maximum degree is less than $n^{1/4-\epsilon}$ and some "smoothness" conditions are satisfied. Also, there is almost surely no giant component when $Q = \sum_{i\ge1} i(i-2)\gamma_i < 0$ and the maximum degree is less than $n^{1/8-\epsilon}$.

Let us consider Q for our (α, β)-graphs with $\beta > 3$.

$$\begin{aligned} Q &= \frac{1}{n}\sum_{x=1}^{e^{\frac{\alpha}{\beta}}} x(x-2)\lfloor\frac{e^\alpha}{x^\beta}\rfloor \\ &\approx \frac{1}{\zeta(\beta)}\left(\sum_{x=1}^{e^{\frac{\alpha}{\beta}}}\frac{1}{x^{\beta-2}} - 2\sum_{x=1}^{e^{\frac{\alpha}{\beta}}}\frac{1}{x^{\beta-1}}\right) \\ &\approx \frac{\zeta(\beta-2) - 2\zeta(\beta-1)}{\zeta(\beta)} \end{aligned}$$

Hence, we consider the value $\beta_0 = 3.47875\ldots$, which we recall is a solution to $\zeta(\beta-2) - 2\zeta(\beta-1) = 0$. If $\beta > \beta_0$, we have

$$\sum_{x=1}^{e^{\frac{\alpha}{\beta}}} x(x-2)\lfloor\frac{e^\alpha}{x^\beta}\rfloor < 0\,.$$

We remark that for $\beta > 8$, Molloy and Reed's result immediately implies that almost surely there is no giant component. When $\beta \le 8$, additional analysis is needed to deal with the degree constraints [2].

It can be shown that the second largest component almost surely has size $\Theta(\log n)$. Furthermore, the second largest component has size at least $\Theta(\log n)$.

$\mathcal{G}(\mathbf{w})$ Model for Power Law Graphs

In the Erdős–Rényi model $\mathcal{G}(n,p)$, the threshold function for the phase transition of the giant component is at $p = 1/n$. Namely, when the average degree pn is less than 1, all connected components are small (of size $O(\log n)$) and there is no giant component. When the average degree is more than 1, the giant component emerges in full swing. (There is a "double jump" which takes place when the average degree is close to 1 as discussed in Sect. "Random Graphs in a Nutshell".)

For the random graph model $\mathcal{G}(\mathbf{w})$, with given expected degrees $\mathbf{w}$, it is natural to ask the same question:

> What parameter in which range will trigger the (sudden) emergence of the giant component?

In addition to w, the expected average degree, we have scores of parameters, e. g., $\tilde{w}$ and higher order average degrees. Which parameter w, $\tilde{w}$ or others is critical for the rise of the giant component?

These questions were answered in [13]:
Suppose that G is a random graph in $\mathcal{G}(\mathbf{w})$ with expected degree sequence $\mathbf{w}$. *If the expected average degree w is strictly greater than 1, then the following holds:*

(1) *Almost surely G has a unique giant component. Furthermore, the volume of the giant component is at least* $(1 - \frac{2}{\sqrt{w\,\mathrm{e}}} + o(1))\mathrm{Vol}(G)$ *if* $w \geq 4/\mathrm{e} = 1.4715\ldots$, *and is at least* $(1 - (1 + \log w)/w + o(1))\mathrm{Vol}(G)$ *if* $w < 2$.
(2) *The second largest component almost surely has size at most* $(1 + o(1))\mu(w)\log n$, *where*

$$\mu(w) = \begin{cases} \frac{1}{1+\log w - \log 4} & \text{if } w > \frac{4}{\mathrm{e}}\,; \\ \frac{1}{w-1-\log w} & \text{if } 1 < w < 2\,. \end{cases}$$

Moreover, with probability at least $1 - n^{-k}$, *the second largest component has size at most* $(k + 1 + o(1))\mu(w)\log n$, *for any* $k \geq 1$.

There is a sharp asymptotic estimate for the volume of the giant component for a random graph in $\mathcal{G}(\mathbf{w})$. In [15], it was proved that if the expected average degree is strictly greater than 1, then almost surely the giant component in a graph G in $\mathcal{G}(\mathbf{w})$ has volume $\lambda_0\mathrm{Vol}(G) + O(\sqrt{n}\log^{3.5} n)$, where λ_0 is the unique nonzero root of the following equation:

$$\sum_{i=1}^{n} w_i \mathrm{e}^{-w_i\lambda} = (1-\lambda)\sum_{i=1}^{n} w_i\,. \tag{4}$$

Because of the robustness of the $\mathcal{G}(\mathbf{w})$ model, many properties can be derived for appropriate degree distributions, including power law graphs.

Average Distance and the Diameter A random graph G in $\mathcal{G}(\mathbf{w})$ has average distance almost surely $(1 + o(1)) \cdot (\log n)/(\log \tilde{w})$, if $\mathbf{w}$ satisfies certain conditions (called admissible conditions in [14]). The diameter is almost surely $\Theta((\log n)/(\log \tilde{w}))$. In addition to studying the average distance and diameter, the structure of a random power law graph is very interesting, especially for the range $2 < \beta < 3$ where the power law exponents β for numerous real networks reside. In this range, the power law graph can be roughly described as an "octopus" with a dense subgraph having small diameter $O(\log\log n)$, as the core while the overall diameter is $O(\log n)$ and the average distance is $O(\log\log n)$. When $\beta > 3$ and the average degree w is strictly greater than 1, almost surely the average distance is $(1 + o(1))(\log n)/(\log \tilde{w})$ and the diameter is $\Theta(\log n)$. A phase transition occurs at $\beta = 3$ and then the graph has diameter almost surely $\Theta(\log n)$ and average distance $\Theta(\log n/\log\log n)$.

Eigenvalues Eigenvalues of graphs are useful for controlling many graph properties and consequently have numerous algorithmic applications including clustering algorithms, low rank approximations, information retrieval and computer vision. In the study of the spectra of power law graphs, there are basically two competing approaches. One is to prove analogues of Wigner's semi-circle law (such as for $\mathcal{G}(n, p)$) while the other predicts that the eigenvalues follow a power law distribution [27]. Although the semi-circle law and the power law have nothing in common, both approaches are essentially correct if one considers the appropriate matrices. there are in fact several ways to associate a matrix to a graph. The usual adjacency matrix A associated with a (simple) graph has eigenvalues quite sensitive to the maximum degree (which is a *local* property). The combinatorial Laplacian $D - A$ with D denoting the diagonal degree matrix is a major tool for enumerating spanning trees and has numerous applications. Another matrix associated with a graph is the (normalized) Laplacian $L = I - D^{-1/2}AD^{-1/2}$ which controls the expansion/isoperimetrical properties (which are *global*) and essentially determines the mixing rate of a random walk on the graph. The traditional random matrices and random graphs are regular or almost regular so the spectra of all the above three matrices are basically the same (with possibly a scaling factor or a linear shift). However, for graphs with uneven degrees, the above three matrices can have very different distributions.

Here we state bounds for eigenvalues for random graphs in $\mathcal{G}(\mathbf{w})$ with a general degree distribution from which the results on random power law graphs then follow [18].

1. The largest eigenvalue of the adjacency matrix of a random graph with a given expected degree sequence is determined by m, the maximum degree, and $\tilde{w}$, the weighted average of the squares of the expected degrees. In this case the largest eigenvalue of the adjacency matrix is almost surely $(1 + o(1))\max\{\tilde{w}, \sqrt{m}\}$ provided some minor conditions are satisfied. In addition, if the kth largest expected degree m_k is significantly larger than $\tilde{w}^2$, then the kth largest eigenvalue of the adjacency matrix is almost surely $(1 + o(1))\sqrt{m_k}$.

2. For a random power law graph with exponent $\beta > 2.5$, the largest eigenvalue of a random power law graph is almost surely $(1 + o(1))\sqrt{m}$ where m is the maximum degree. Moreover, the k largest eigenvalues of a random power law graph with exponent β have power law distribution with exponent $2\beta - 1$ if the maximum degree is sufficiently large and k is bounded above by a function depending on β, m and w, the average degree. When $2 < \beta < 2.5$, the largest eigenvalue is heavily concentrated at $cm^{3-\beta}$ for some constant c depending on β and the average degree.
3. The eigenvalues of the Laplacian satisfy the semi-circle law under the condition that the minimum expected degree is relatively large ($\gg$ the square root of the expected average degree). This condition contains the basic case when all degrees are equal (the Erdős–Rényi model). If we weaken the condition on the minimum expected degree, we can still have the following strong bound for the eigenvalues of the Laplacian which implies strong expansion rates for rapid mixing,

$$\max_{i \neq 0} |1 - \lambda_i| \leq (1 + o(1)) \frac{4}{\sqrt{w}} + \frac{g(n) \log^2 n}{w_{\min}}$$

where w is the expected average degree, $w_{\min}$ is the minimum expected degree and $g(n)$ is any slow growing function of n.

On-Line Random Graphs

Preferential Attachment Schemes

The preferential attachment scheme is often attributed to Herbert Simon. In his paper [41] of 1955, he gave a model for word distribution using the preferential attachment scheme and derived *Zipf's law* (i. e., the probability of a word having occurred exactly i times is proportional to $1/i$).

The basic setup for the preferential attachment scheme is a simple *local* growth rule which leads to a *global* consequence—a power law distribution. Since this local growth rule gives preferences to vertices with large degrees, the scheme is often described by "*the rich get richer*". Of interest is to determine the exponent of the power law from the parameters of the local growth rule.

There are two parameters for the preferential attachment model:

- A probability p, where $0 \leq p \leq 1$.
- An initial graph G_0, that we have at time 0.

Usually, G_0 is taken to be the graph formed by one vertex having one loop. (We consider the degree of this vertex to be 1, and in general a loop adds 1 to the degree of a vertex.) Note, in this model multiple edges and loops are allowed.

We also have two operations we can do on a graph:

Vertex-step Add a new vertex v, and add an edge $\{u, v\}$ from v by randomly and independently choosing u in proportion to the degree of u in the current graph.

Edge-step Add a new edge $\{r, s\}$ by independently choosing vertices r and s with probability proportional to their degrees.

Note that for the edge-step, r and s could be the same vertex. Thus loops could be created. However, as the graph gets large, the probability of adding a loop can be well bounded and is quite small.

The random graph model $G(p, G_0)$ is defined as follows:

Begin with the initial graph G_0.
For $t > 0$, at time t, the graph G_t is formed by modifying G_{t-1} as follows:
with probability p, take a vertex-step,
otherwise, take an edge-step.

When G_0 is the graph consisting of a single loop, we will simplify the notation and write $G(p) = G(p, G_0)$.

There were quite a number of papers analyzing the preferential attachment model $G(p)$, usually having similar conclusions of power law degree distribution. However, many of these analyses are heuristics without specifying the ranges for the power law to hold. Heuristics often run into the danger of incorrect deductions and incomplete conclusions. It is quite essential to use rigorous proofs which help specify the appropriate conditions and ranges for the power law. The following statement was proved in [16].

For the preferential attachment model $G(p)$, almost surely the number of vertices with degree k at time t is

$$M_k t + O(2\sqrt{k^3 t \ln(t)}) .$$

where $M_1 = (2p)/(4 - p)$ and $M_k = (2p)/(4 - p)(\Gamma(k)\Gamma(1 + 2/(2 - p)))/(\Gamma(k + 1 + 2/(2 - p))) = O(k^{-(2+p/(2-p))})$, for $k \geq 2$. In other words, almost surely the graphs generated by $G(p)$ have the power law degree distribution with the exponent $\beta = 2 + p/(2 - p)$.

Duplication Models

Networks of interactions are present in all biological systems. The interactions among species in ecosystems, between cells in an organism and among molecules in a cell all lead to complex biological networks. Using current technological advances, extensive data of such interactions has been acquired. To find the underlying structure in

Random Graphs, A Whirlwind Tour of, Table 3
Power law exponents for biological and nonbiological networks

Biological networks	Exponent β	References
Yeast protein -protein net	1.6, 1.7	[20,43]
E. Coli metabolic net	1.7, 2.2	[3,28]
Yeast gene expression net	1.4–1.7	[20]
Gene functional interaction	1.6	[30]
Nonbiological networks		
Internet graph	2.2 (indegree), 2.6 (outdegree)	[4,27,34]
Phone call graph	2.1 –2.3	[1,2]
Collaboration graph	2.4	[29]
Hollywood graph	2.3	[4]

these databases, it is of great importance to understand the basic principles of various genetic and metabolic networks.

It has been observed that many biological networks have power law graphs with exponents β less than 2. The ranges for the exponents of the power law for biological networks are quite different from the ranges for nonbiological networks. Various examples, such as the WWW-graphs, call graphs, and various social networks, among others, are power law graphs with the exponent β between 2 and 3. Table 3 lists the exponents of a variety of biological and nonbiological networks with associated references. As we saw in Sect. "Preferential Attachment Schemes", the preferential attachment model generates graphs with power law degree distribution with exponents β between 2 and 3. Therefore there is a need to consider alternative models for biological networks.

The duplication of the information in the genome – genes and their controlling elements – is a driving force in evolution and a determinative factor of biological networks. The process of duplication is quite different from the preferential attachment process that is regarded by many as the basic growth rule for most nonbiological networks.

Here we consider a duplication model. If we only allow pure duplication, the resulting graph depends heavily on the initial graph and does not satisfy the power law. So we consider a duplication model that allows randomness within the duplication step as defined below. We will see that this duplication model generates power law graphs with exponents in the range including the interval between 1 and 2 and therefore is more suitable for modeling complex biological networks.

There are two basic parameters for the duplication model:

- A selection probability p, where $0 \le p \le 1$.
- An initial graph G_0, that we have at time 0.

Usually, G_0 is taken to be the graph formed by one vertex. However, G_0 can be taken to be any finite simple connected graph. Unlike the preferential attachment model, in this model the generated random graph is always a simple graph.

There is one basic operation:

Duplication step: A *sample* vertex u is selected randomly and uniformly from the current graph. A new vertex v and edge $\{u, v\}$ is added to the graph. For each neighbor w of u, with probability p, $\{v, w\}$ is added as a new edge.

The edge $\{u, v\}$ in the duplication step is called a *basic* edge. The vertex u is said to be the *parent* of v and v is called a *child* of u. We note that a vertex can have several children or no child at all and that each vertex not in the initial graph G_0 has a parent. All basic edges from children to parents form a forest where the vertices in G_0 are roots of component trees. All terms like "leafs" and "descendants", if not defined, refer to this forest.

The duplication step can be further decomposed into two parts—vertex-duplication and edge-duplication as follows:

Vertex-duplication: At time t, randomly select a sample vertex u and add a new vertex v and an edge $\{u, v\}$.

Edge-duplication: At time t, for the vertex v created, its parent u and each neighbor w of u, with probability p, add an edge $\{v, w\}$ to w.

For any vertex v, a *descendant* of v can only be connected to descendants of v's neighbors (including v itself). An edge $\{x, y\}$ is said to be a descendant of an edge $\{u, v\}$, if "x is a descendant of u and y is a descendant of v" or "x is a descendant of v and y is a descendant of u".

We remark that having the basic edges $\{u, v\}$ makes the graph G always connected. This helps avoid degenerate cases such as having mostly isolated vertices.

For the above duplication model, it can be shown [17] that its degree distribution obeys a power law with the exponent β of the power law satisfying the following equation:

$$1 + p = p\beta + p^{\beta-1} . \tag{5}$$

We remark that the solutions for (5) that are illustrated in Fig. 4 consist of two parts. One is the line $\beta = 1$. The other is a curve which is a monotonically decreasing function of p. The two curves intersect at $(x, 1)$ where $x = 0.56714329\ldots$ the solution of $x = -\log x$. One very interesting range for β is when p is near 1/2. To get a power-law with exponent 1.5, for example, one should choose $p = 0.535898\ldots$ Also we see that the second curve intersects zero at $p = (\sqrt{5} - 1)/2$, an intriguing number (the "golden mean"). At $p = 1/2$, one solution for β is 2. Although there are two solutions for each p, the stable

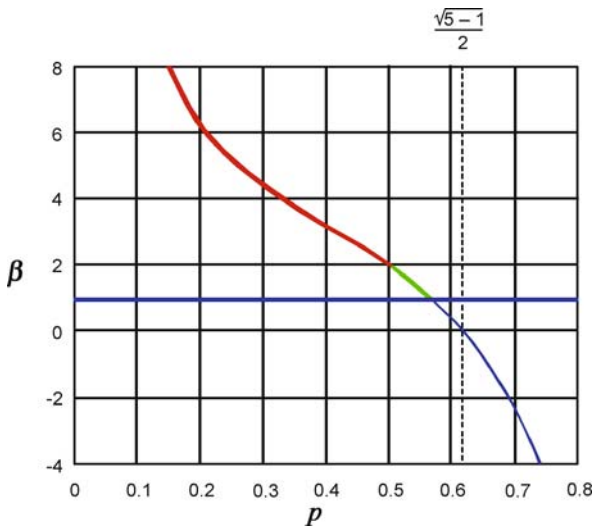

Random Graphs, A Whirlwind Tour of, Figure 4
The value of β as a function of p. The *green* section of the curve shows the range of β values which are typically found for biological networks

solutions are on the curve when $p < 0.56714329\ldots$ and $\beta = 1$ for $p > 0.56714329\ldots$

Remarks

The small world phenomenon, that occurs ubiquitously in numerous existing networks, refers to two similar but different properties:

Small distance Between any pair of nodes, there is a short path.

The clustering effect Two nodes are more likely to be adjacent if they share a common neighbor.

There have been various approaches to model networks that exhibit the small world phenomenon. In particular, the aspect of small distances can be well explained by using random graphs with general degree distributions which include the power law distribution. However, the other feature concerning the clustering effect seems much harder to model.

To model the clustering effect, a typical approach is to add random edges to a grid graph or the like [26,32,44]. Such grid-based models are quite restrictive and far from satisfactory for modeling biological networks or collaboration graphs, for example. On the other hand, random power law graphs are good for modeling small distance, but fail miserably for modeling the clustering effect. In a way, the aspect of small distances is about neighborhood expansion while the aspect of the clustering effect is about neighborhood density. The related graph-theoretical parameters seem to be of an entirely different scale. For example, while the clustering effect is quite sensitive to average degree, the small distance effect is not.

The heart of the problem can be quite simply stated: For a given network, what is its true geometry? How can we capture the geometry of the network (without invoking too many parameters)?

Bibliography

1. Abello J, Buchsbaum A, Westbrook J (1998) A functional approach to external graph algorithms. In: Proc. 6th European Symposium on Algorithms. Springer, Berlin, pp 332–343
2. Aiello W, Chung F, Lu L (2000) A random graph model for massive graphs. In: Proc. of the Thirty-Second Annual ACM Symposium on Theory of Computing. ACM Press, New York, pp 171–180
3. Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
4. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512
5. Barabási A-L, Albert R, Jeong H (2000) Scale-free characteristics of random networks: the topology of the world-wide web. Physica A 281:69–77
6. Bender EA, Canfield ER (1978) The asymptotic number of labeled graphs with given degree sequences. J Comb Theory Ser A 24:296–307
7. Bollobás B (1984) The evolution of random graphs. Trans Amer Math Soc 286:257–274
8. Bollobás B (1981) The diameter of random graphs. Trans Amer Math Soc 267:41–52
9. Bollobás B (1985) Random graphs. Academic Press, New York, pp xvi+447
10. Bollobás B (1984) The evolution of sparse graphs. In: Graph theory and combinatorics. Academic Press, London, New York, pp 35–57, (Cambridge 1983)
11. Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener J (2000) Graph structure in the web. Proc. of the WWW9 Conference, Amsterdam, May 2000, Paper version appeared in Computer Networks 33, pp 309–320
12. Chung F, Lu L (2001) The diameter of sparse random graphs. Adv Appl Math 26:257–279
13. Chung F, Lu L (2002) Connected components in random graphs with given expected degree sequences. Ann Comb 6:125–145
14. Chung F, Lu L (2002) The average distances in random graphs with given expected degrees. Proc Nat Acad Sci 99:15879–15882
15. Chung F, Lu L (2006) The volume of the giant component of a random graph with given expected degrees. SIAM J Discret Math 20:395–411
16. Chung F, Lu L (2006) Complex graphs and networks. In: CBMS Lecture Series, vol 107. AMS Publications, Providence, pp vii+264
17. Chung F, Lu L, Dewey G, Galas DJ (2003) Duplication models for biological networks. J Comput Biol 10(5):677–687
18. Chung F, Lu L, Vu V (2003) The spectra of random graphs with given expected degrees. Proc Nat Acad Sci 100(11):6313–6318

19. Conlon D A new upper bound for diagonal Ramsey numbers. Ann Math (in press)
20. Dorogovtsev SN, Mendes JFF (2001) Scaling properties of scale-free evolving networks: continuous approach. Phys Rev E 63(056125):19
21. Erdős P (1947) Some remarks on the theory of graphs. Bull Amer Math Soc 53:292–294
22. Erdős P, Rényi A (1959) On random graphs, vol I. Publ Math Debrecen 6:290–297
23. Erdős P, Rényi A (1960) On the evolution of random graphs. Magyar Tud Akad Mat Kutató Int Közl 5:17–61
24. Erdős P, Gallai T (1961) Gráfok előírt fokú pontokkal (Graphs with points of prescribed degrees, in Hungarian). Mat Lapok 11:264–274
25. Exoo G (1989) A lower bound for $R(5,5)$. J Graph Theory 13:97–98
26. Fabrikant A, Koutsoupias E, Papadimitriou CH (2002) Heuristically optimized trade-offs: a new paradigm for power laws in the internet. In: Automata, languages and programming. Lecture Notes in Computer Science, vol 2380. Springer, Berlin, pp 110–122
27. Faloutsos M, Faloutsos P, Faloutsos C (1999) On power-law relationships of the Internet topology. In: Proc. of the ACM SIGCOM Conference. ACM Press, New York, pp 251–262
28. Friedman R, Hughes A (2001) Gene duplications and the structure of eukaryotic genomes. Genome Res 11:373–381
29. Grossman J, Ion P, De Castro R (2007) The Erdős number project. http://www.oakland.edu/enp
30. Gu Z, Cavalcanti A, Chen F-C, Bouman P, Li W-H (2002) Extent of gene duplication in the genomes of drosophila, nematode, and yeast. Mol Biol Evol 19:256–262
31. Janson S, Knuth DE, Łuczak T, Pittel B (1993) The birth of the giant component. Rand Struct Algorithms 4:231–358
32. Kleinberg J (2000) The small-world phenomenon: an algorithmic perspective. Proc. 32nd ACM Symposium on Theory of Computing. ACM, New York, pp 163–170
33. Kleinberg J, Kumar R, Raphavan P, Rajagopalan S, Tomkins A (1999) The web as a graph: measurements, models and methods. In: Proc. of the International Conference on Combinatorics and Computing. Lecture Notes in Computer Science, vol 1627. Springer, Berlin, pp 1–17
34. Kumar R, Raghavan P, Rajagopalan S, Tomkins A (1999) Trawling the web for emerging cyber communities. In: Proc. of the 8th World Wide Web Conference. Toronto, 1999
35. Łuczak T (1998) Random trees and random graphs. Rand Struct Algorithms 13:485–500
36. McKay BD, Radziszowski SP (1994) Subgraph counting identities and Ramsey numbers. J Comb Theory (B) 61:125–132
37. Milgram S (1967) The small world problem. Psychol Today 2:60–67
38. Molloy M, Reed B (1995) A critical point for random graphs with a given degree sequence. Rand Struct Algorithms 6:161–179
39. Molloy M, Reed B (1998) The size of the giant component of a random graph with a given degree sequence. Combin Probab Comput 7:295–305
40. Ramsey FP (1930) On a problem of formal logic. Proc London Math Soc 30:264–286
41. Simon HA (1955) On a class of skew distribution functions. Biometrika 42:425–440
42. Spencer J (1975) Ramsey's theorem—a new lower bound. J Comb Theory (A) 18:108–115
43. Stubbs L (2002) Genome comparison techniques. In: Galas D, McCormack S (eds) Genomic technologies: present and future. Caister Academic Press
44. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small world' networks. Nature 393:440–442

Random Matrix Theory

Güler Ergün
Department of Mathematical Sciences,
University of Bath, Bath, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Random Matrix Ensemble Classifications
Construction of Gaussian Ensembles
Eigenvalues of Gaussian Ensembles
Density of States: Semicircle Law
Spacing Distribution: Wigner Surmise
Two-Level Correlation Function
Random Matrix Theory and Complex Systems
Summary
Future Directions
Bibliography

Glossary

Random matrices Large matrices with randomly distributed elements obeying the given probability laws and symmetry classes.

Orthogonal ensembles Real symmetric random matrix ensembles which are invariant under all orthogonal transformations. Majority of practical systems are described by these ensembles.

Unitary ensembles Hermitian random matrix ensembles which are invariant under all unitary transformations.

Symplectic ensembles Hermitian self-dual random matrix ensembles which are invariant under all symplectic transformations.

Definition of the Subject

Random Matrix Theory (RMT) is a method of studying the statistical behavior of large complex systems, by defining an ensemble which considers all possible laws of interactions within the system. The important question addressed by random matrix theory is: Given a random matrix ensemble what are the probability laws which govern its eigenvalues or eigenvectors. This question is pertinent

to many areas in physics and mathematics, for instance statistical behavior of compound nucleus, conductivity in disordered metals, behavior of chaotic systems or zeros of the Riemann zeta function. The success of random matrices lies in the universality regime of the eigenvalue statistics. There is compelling evidence that when the size of the matrix is very large then the eigenvalue distribution tends, in a certain sense, towards a limiting distribution. This only depends on the symmetry properties of the matrix and is independent of the initial probability law imposed on the matrix entries. For example, in many-body systems, the random matrix models are very successful in describing the spectral fluctuation properties of complex atoms, molecules and atomic nuclei, despite the fact that the interactions between the constitutes of these systems are very different, and not at all random: The nuclei are bound by short-range nuclear forces whereas atoms and molecules are governed by the long-range Coulomb forces, but the Hamiltonian of these models only considers the symmetry properties without the detailed knowledge of the system under consideration. The reason for such universality is not so clear but may be an outcome of a law of large numbers operating in the background. In the proceeding sections after giving the history and development of the theory, we will explore the random matrix ensembles, eigenvalue statistics as well as their pertinence to various complex systems.

Introduction

Although, the Random Matrix Theory was initiated in the early 1900s by statisticians in their studies of the product moment distribution [59], it made its progress in the hands of physicists. It was Eugene Wigner who introduced the random matrix idea to the theoretical physics community in the mid 1950s. During the 1950s, due to the development of nuclear weapons and nuclear power stations, physicists of this era were involved in measuring fission resonances in various nuclei. Wigner was predominantly studying [53] the statistical properties of nuclear spectra of heavy nuclei, but the large number of resonances was making his analysis very difficult. It was clear that the resonance spectra were far too complex to be analyzed by existing models like the atomic shell model. This prompted Wigner to look for statistical methods which would deal with such complexity. Later, the realization that the statistical distribution of the nuclear resonance energies shared the same properties as the eigenvalues of random matrices, led him to study a random matrix model [54], in which an infinite real symmetric matrix with identically distributed elements having the values zero in the diagonal and + 1 or − 1 in the off diagonal was used, so that all sign interactions were equally probable. In this model he showed that the distribution of the characteristic values of the matrix i. e., the density of states (DOS) was a semicircle, and it was also obvious that the characteristic values were repelling each other.[1]

During a number of conferences [55,56,57] in 1956–1957, on the basis of eigenvalues of a 2×2 real symmetric matrix, Wigner gave an assumed formula for the level spacing distribution, and he advocated Wishart's work [59], which was giving the joint probability distribution of eigenvalues, to confirm his findings. At about the same time analysis of experimental data by Gurevich and Pevsner [29] was also confirming that the nuclear levels were repelling each other and the distribution of consecutive level spacings was not totally random, but showed some regularities.

The interest in random matrix hypothesis was becoming wide spread, in [44] Porter and Rosenzweig summarized the developments on the statistical properties of atomic and nuclear spectra, by including experimental, theoretical, and numerical findings, which helped to strengthen the applicability of such statistical laws to nuclear reactions. It was time to make these statistical theories more mathematically rigorous; Wigner used the method of moments to prove that the density of eigenvalues for large symmetric matrices distributed as a semi circle [58]. Motivated by this, in 1960 Mehta and Gaudin obtained the same distribution by integrating joint probability density function of eigenvalues, and simultaneously Mehta analytically showed that the Wigner surmise was a very good approximation for the large random matrices. Soon after that, Gaudin gave the exact expression for the spacing distribution [26].

Dyson considered these statistical theories of random matrices as a new kind of Statistical Mechanics [10] such that one ignores the exact knowledge of the nature of the system. That is to say, a compound nucleus can be viewed as a *black box*, and one would need to define an ensemble of systems in which all possible laws of interactions are equally probable. In an attempt to explore this so-called "new kind of statistical mechanics" Dyson presented a series of papers [10,11,12], and in these papers he showed that all random matrix ensembles fall into three universality classes called the *orthogonal ensemble* (OE), *unitary ensemble* (UE) and *symplectic ensemble* (SE). These ensembles account for the symmetry properties of the Hamiltonians as well as the mathematical rigor of all possible interactions being equally likely. For matrices with Gaus-

[1] This was already known to Wigner and von Neuman [51].

sian distributed entries they are referred to as GOE, GUE and GSE, respectively. The random matrix theory (RMT) was developed all through the 1960s by Wigner, Mehta, Gaudin and Dyson. Most of the early work up to 1965 on random matrices is collected in *Statistical Theories of spectra* by Porter.

The RMT took a new dimension in the 1980s when Bohigas, Giannoni and Schmit (BGS) [5] conjectured that it could be applied to describing the spectra of all chaotic systems. This BGS statement not only renewed the interest in random matrices but also started a surge of research in different subject areas such as quantum chaos, quantum graphs, mesoscopic systems etc., and now there exists an overwhelming evidence that the BGS proposition may be true. There are several review articles which summarize the RMT applied research up to 1980s by Brody and up to more recent times by Bohigas, Guhr. A comprehensive book in this subject area is *Random Matrices* by Mehta [38], however it is aimed at the more advanced reader. As for the new comers, the pedagogical derivation of most of the basic principles of RMT may be found in *Quantum Signatures of Chaos* by Haake [30].

Nowadays, RMT is at the center of several areas of mathematics, particularly number theory, combinatorics, diffusion processes, probability and statistics. At the same time it is used as a tool to understand many complex systems from biology, quantum chaos, wireless communication and finance.

In the following sections, first the most commonly used random matrix ensembles, namely the Gaussian ensembles, will be presented. Then some of their central statistical properties such as the density of states and spacing distribution will be discussed.

Random Matrix Ensemble Classifications

Because of the broad use of RMT a variety of random matrix ensembles have been studied. These ensembles can be loosely divided into three types, Gaussian, circular and hyperbolic. In the next sections, although the emphasis will be on the Gaussian ensembles, we also give very brief descriptions of the circular and the hyperbolic ensembles.

Gaussian Ensembles

These ensembles are categorized into three universality classes, which only take into account: (1) the time-reversal invariance (more often referred to as the time-reversal symmetry) and (2) the strong spin-orbit coupling of the Hamiltonian of an underlying quantum system.

In Gaussian ensembles, the type of transformation gives the classifying name to the ensemble. For quick reference these ensembles are grouped in the Table 1 with consideration of the space-time symmetries and the values of the angular momentum.

Circular Ensembles

Introduced by Dyson [10] with an assumption that the system is not characterized by its Hamiltonian but by a unitary matrix S. The matrix is considered to be a scattering matrix (S-Matrix), whose elements give the transition probabilities between various states and its eigenvalues belong to the unit circle in the complex plane. Because of this compactness of the ensembles the universal distribution can be obtained with a uniform probability law. From the point of view of the fundamental symmetries, these ensembles can be categorized into three types as: Circular orthogonal ensembles (COE), circular unitary ensembles (CUE), circular symplectic ensembles (CSE). In the large N limit, the statistical properties of their eigenvalues are identical to those of the Gaussian ensembles.

Hyperbolic Ensembles

These ensembles are appropriately described in terms of Brownian motion in some matrix spaces. A seminal study of such a system is Dyson's Brownian motion model [9], which applies the notion of the RMT to a new type of Coulomb gas model so that the "Coulomb gas" could be interpreted as a dynamical system.

The matrix defined is a transfer matrix (M-matrix), which is a matrix used to give the relationship between the fluxes at one edge of a conductor to fluxes at the other edge (i. e., the number of ingoing electrons minus the outgoing electrons should be the same on each side of a conductor). Unlike the S-matrix, the M-matrix does not belong to a compact group. Also, it is not possible to define a normalizable probability law, but instead a diffusion equation as a function of a fictitious time. The term "time" can be any property of the model such as the length of a conductor.

Construction of Gaussian Ensembles

In fixing the elements of random matrices we are not completely free, since, first of all the matrix elements must obey the class restrictions given in the previous section. In the GOE case the Hamiltonian matrix is considered to be real symmetric. This means that for an $N \times N$ real symmetric matrix, there are $N(N+1)/2$ elements i. e., H_{ij} with $i, j = 1, \ldots, N$ can be chosen independently and the rest of them are determined by the symmetry. The ultimate aim is to define a joint probability density $P(\mathcal{H})$ for the matrix elements satisfying the conditions:

Random Matrix Theory, Table 1

The grouping of the quantum systems by the symmetry properties and the value of angular momentum. β is an index of the classification

Gaussian Ensemble Classification					
Symmetries	Total angular mom.	Hamiltonian	Canonical group	Ensemble	β
Time-reversal	Integer	Real symmetric	Orthogonal	GOE	1
Time-reversal, Rotational	Half-Integer				
Time-reversal, No-rotational	Half-Integer	Quaternion real	Symplectic	GSE	4
None	Any	Hermitian	Unitary	GUE	2

1. The probability $P(\mathcal{H})$ must be invariant under any transformation i. e. $\mathcal{H} \to \mathcal{W}^{-1}\mathcal{H}'\mathcal{W}$ with $\mathcal{W}$ is either orthogonal ($\beta = 1$), unitary ($\beta = 2$), or symplectic ($\beta = 4$) matrix. Then, e. g. for $\beta = 1$, when $\mathcal{W} = \mathcal{O}$ is any real orthogonal matrix then we have

$$P(\mathcal{H})\,\mathrm{d}\mathcal{H} = P(\mathcal{H}')\,\mathrm{d}\mathcal{H}'\,, \qquad \mathcal{O}^{-1} = \mathcal{O}^{\mathrm{T}}, \quad \text{and} \quad \mathcal{O}\mathcal{O}^{\mathrm{T}} = 1\,. \tag{1}$$

2. The matrix elements H_{ij} are statistically independent. Thus, the joint probability density function must be the product of the densities of these elements,

$$P(\mathcal{H}) = \prod_{ij} P_{ij}(H_{ij})\,. \tag{2}$$

The first condition is absolutely essential, $P(\mathcal{H})$ must be invariant under any transformation, because these transforms determine the eigenvalues, but the second one is only chosen to simplify the calculations. However, as will be more apparent later in the derivations, the consideration of these conditions necessitates the distribution of the random matrix elements to be Gaussian.

To find $P(\mathcal{H})$ for the real symmetric Hamiltonian matrices, we will consider a 2×2 real symmetric matrix $\mathcal{H}$ and an orthogonal transformation matrix

$$\mathcal{O} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}, \tag{3}$$

which is the two-dimensional rotation through θ. The rotation $\mathcal{O}$ acts as a transformation $\mathcal{H} = \mathcal{O}^{\mathrm{T}}\mathcal{H}'\mathcal{O}$, from which we get the relations

$$\begin{aligned} H_{11} &= \frac{H'_{11} + H'_{22}}{2} + \frac{H'_{11} - H'_{22}}{2}\cos 2\theta - H'_{12}\sin 2\theta \\ H_{12} &= \frac{H'_{11} - H'_{22}}{2}\sin 2\theta + H'_{12}\cos 2\theta \\ H_{22} &= \frac{H'_{11} + H'_{22}}{2} - \frac{H'_{11} - H'_{22}}{2}\cos 2\theta + H'_{12}\sin 2\theta\,. \end{aligned} \tag{4}$$

The joint probability density function is simply the product of the three densities

$$P(\mathcal{H}) = P_{11}(H_{11})P_{12}(H_{12})P_{22}(H_{22})\,, \tag{5}$$

and the orthogonal invariance of $P(\mathcal{H})$ implies that $\mathrm{d}P(\mathcal{H})\,\mathrm{d}\theta = 0$. Therefore we have

$$\frac{\mathrm{d}P_{11}(H_{11})}{P_{11}(H_{11})}\frac{\mathrm{d}H_{11}}{\mathrm{d}\theta}P(\mathcal{H}) + \frac{\mathrm{d}P_{12}(H_{12})}{P_{12}(H_{12})}\frac{\mathrm{d}H_{12}}{\mathrm{d}\theta}P(\mathcal{H}) + \frac{\mathrm{d}P_{22}(H_{22})}{P_{22}(H_{22})}\frac{\mathrm{d}H_{22}}{\mathrm{d}\theta}P(\mathcal{H}) = 0\,. \tag{6}$$

From the Eqs. (4) we can find equivalent expressions for the three differentials of $\mathrm{d}\mathcal{H}/\mathrm{d}\theta$ as:

$$\frac{\mathrm{d}H_{11}}{\mathrm{d}\theta} = -2H_{12}, \quad \frac{\mathrm{d}H_{12}}{\mathrm{d}\theta} = H_{11} - H_{22}, \quad \frac{\mathrm{d}H_{22}}{\mathrm{d}\theta} = 2H_{12}. \tag{7}$$

Substituting these later relations into Eq. (6), and dividing by $-H_{12}(H_{11} - H_{22})$ we get

$$\frac{\mathrm{d}\ln P_{11}(H_{11})}{\mathrm{d}H_{11}}\frac{2}{H_{11} - H_{22}} - \frac{\mathrm{d}\ln P_{22}(H_{22})}{\mathrm{d}H_{22}} \cdot \frac{2}{H_{11} - H_{22}} = \frac{\mathrm{d}\ln P_{12}(H_{12})}{\mathrm{d}H_{12}}\frac{1}{H_{12}} = -2a\,. \tag{8}$$

The two sides of this equation depend on different variables, which means that each side must be equal to a constant, hence the constant a is introduced. From Eq. (8) we first find a solution for $P_{12}(H_{12})$ by simply integrating

$$\begin{aligned} \mathrm{d}\ln P_{12}(H_{12}) &= -2aH_{12}\mathrm{d}H_{12} \\ P_{12}(H_{12}) &= C_{12}\exp\left(-aH_{12}^2\right)\,, \end{aligned} \tag{9}$$

in which the constant coefficient C_{12} is a consequence of the integration. Then re-writing Eq. (8) as

$$\frac{\mathrm{d}\ln P_{11}(H_{11})}{\mathrm{d}H_{11}} + aH_{11} = \frac{\mathrm{d}\ln P_{22}(H_{22})}{\mathrm{d}H_{22}} + aH_{22} = b \quad (10)$$

where another constant b is introduced as the equation depends on a different variable on each side, and following the same sequence of steps as before, we get solutions of P_{11} and P_{22},

$$\begin{aligned} P_{11}(H_{11}) &= C_{11}\exp\left[(-a/2)H_{11}^2 + bH_{11}\right] \\ P_{22}(H_{22}) &= C_{22}\exp\left[(-a/2)H_{22}^2 + bH_{22}\right] . \end{aligned} \quad (11)$$

These solutions have Gaussian distribution form, and the product of them gives

$$P(\mathcal{H}) = C\exp\left[(-a/2)(H_{11}^2 + H_{22}^2 + 2H_{12}^2) + b(H_{11} + H_{22})\right] . \quad (12)$$

If we fix the mean values of the diagonal elements to be zero i. e., $\langle H_{\mathrm{ii}}\rangle = 0$ then the second constant $b = 0$. The overall constant C is determined by the normalization

$$\int_{-\infty}^{+\infty} \mathrm{d}H_{11}\mathrm{d}H_{22}\mathrm{d}H_{12}P(\mathcal{H}) = 1 . \quad (13)$$

A simple calculation gives $C = 1/2(a/\pi)^{3/2}$. For the convergence of Eq. (13) the constant a must be positive, also it is related to the widths of the distributions of the diagonal and off-diagonal elements. The variances of the diagonal matrix elements

$$\begin{aligned} \langle H_{22}^2\rangle &= \langle H_{11}^2\rangle \\ &= \sqrt{a/2\pi}\int H_{11}^2\exp\left[(-a/2)H_{11}^2\right]\mathrm{d}H_{11} \\ &= 1/a , \end{aligned} \quad (14)$$

whereas for the off-diagonal elements

$$\langle H_{12}^2\rangle = \sqrt{a/\pi}\int H_{12}^2\exp(-aH_{12}^2)\mathrm{d}H_{12} = 1/2a . \quad (15)$$

By taking σ^2 as the variance of the off diagonal matrix elements, we obtain $a = 1/2\sigma^2$ for a 2×2 matrix.

We make a remark here that for large dimensional matrices the off-diagonal elements greatly outnumber the diagonal ones so, for the purpose of numerical simulations it does not matter much if the variances of the diagonal and the off-diagonal elements are not the same.

Although we have only considered the 2×2 Hamiltonian matrix in finding $P(\mathcal{H})$, this can easily be extended to $N\times N$ dimension with a suitable transformation matrix. In this way we find that a real symmetric random matrix having a probability density

$$P(\mathcal{H}) = \left(\frac{1}{4\sigma^2\pi}\right)^{N/2}\left(\frac{1}{2\sigma^2\pi}\right)^{N(N-1)/4} \cdot\exp\left\{-\frac{1}{4\sigma^2}\sum_{i,j}H_{ij}^2\right\} , \quad (16)$$

defines the *Gaussian orthogonal ensemble*.

Analogously, we obtain the *Gaussian unitary ensemble* (GUE) and the *Gaussian symplectic ensemble* (GSE), with the consideration that the joint probability distribution of the matrix elements is invariant under unitary and symplectic transformations, respectively. Thus, the general form of the probability density of a random Hermitian matrix is given by:

$$P(\mathcal{H}) = \left(\frac{1}{4\sigma^2\pi}\right)^{N/2}\left(\frac{1}{2\sigma^2\pi}\right)^{N(N-1)/2} \cdot\exp\left\{-\frac{1}{4\sigma^2}\sum_{i,j}\left[(H_R)_{ij}^2 + (H_I)_{ij}^2\right]\right\} , \quad (17)$$

where $(H_{\mathrm{R}})_{ij}^2$ and $(H_{\mathrm{I}})_{ij}^2$ are the real and the imaginary parts of the off-diagonal matrix elements H_{ij}, respectively. And the probability density of a random quaternion matrix is:

$$P(\mathcal{H}) = \left(\frac{1}{4\sigma^2\pi}\right)^{N/2}\left(\frac{1}{2\sigma^2\pi}\right)^{N(N-1)} \cdot\exp\left\{-\frac{1}{4\sigma^2}\sum_{i,j}\left[(H_0)_{ij}^2 + (H_1)_{ij}^2 + (H_2)_{ij}^2 + (H_3)_{ij}^2)\right]\right\} , \quad (18)$$

where $(H_0)_{ij}$, $(H_1)_{ij}$, $(H_2)_{ij}$ and $(H_3)_{ij}$ are the quaternionic components of $\mathbf{H}_{ij}$.

Without loss of generality, the probability density $P(\mathcal{H})$ with $\mathcal{H}$ of an arbitrary $N\times N$ dimension can be written in a form common to all three ensembles as

$$P_{N\beta}(\mathcal{H}) = C_{N\beta}\,\mathrm{e}^{-\frac{N\beta}{4\sigma^2}\mathrm{Tr}[\mathcal{H}^2]} , \quad (19)$$

where Tr stands for the *matrix trace*: $\mathrm{Tr}\,\mathcal{H} = \sum_i H_{\mathrm{ii}}$, $\beta = 1, 2, 4$ is used as an index for the ensemble classification, with σ^2 being the variance of the off diagonal matrix elements and the factor N ensuring the moments of the eigenvalues λ_i of $\mathcal{H}$ to be finite in the limit $N\to\infty$, while $\langle\lambda_i^2\rangle = 1$.

Eigenvalues of Gaussian Ensembles

Using the joint probability distribution of the matrix entries in Eq. (19) we will write the joint probability distribution of the eigenvalues for all three Gaussian ensembles. Again, we consider the simplest case of a real symmetric 2×2 Hamiltonian matrix H, and the orthogonal transformation matrix $\mathcal{O}$, as given in Eq. (3), such that

$$\mathcal{H} = \mathcal{O}^{\mathrm{T}} \mathcal{H}_{\mathrm{D}} \mathcal{O} \,, \tag{20}$$

where $\mathcal{H}_{\mathrm{D}}$ is the diagonal matrix containing the eigenvalues of $\mathcal{H}$. In this case there are only two eigenvalues, which are given by

$$\lambda_{1,2} = \frac{1}{2}(H_{11} + H_{22}) \pm \sqrt{(H_{11} - H_{22})^2 + 4H_{12}^2} \,. \tag{21}$$

Writing out Eq. (20) explicitly

$$\begin{aligned} H_{11} &= \lambda_1 \cos^2\theta + \lambda_2 \sin^2\theta \\ H_{22} &= \lambda_1 \sin^2\theta + \lambda_2 \cos^2\theta \\ H_{12} &= (\lambda_1 - \lambda_2)\cos\theta \sin\theta \,, \end{aligned} \tag{22}$$

we can see that the elements of $\mathcal{H}$ are linear functions of the eigenvalues and a single angle θ parameter which specifies the set of eigenvectors. The Jacobian of this transformation is simply

$$J = \det\left[\frac{\partial(H_{11}, H_{22}, H_{12})}{\partial(\lambda_1, \lambda_2, \theta)}\right] = (\lambda_1 - \lambda_2) \,, \tag{23}$$

and since $\mathrm{Tr}[\mathcal{H}^2] = (\lambda_1^2 + \lambda_2^2)$, we can just replace the matrix elements H_{ij} with the eigenvalues of $\mathcal{H}$ and integrate over θ to get

$$P(\lambda_1, \lambda_2) = C|\lambda_1 - \lambda_2| \mathrm{e}^{-\frac{N\beta}{4\sigma^2}(\lambda_1^2 + \lambda_2^2)} \,, \tag{24}$$

which is the joint probability density of the eigenvalues for the GOE of 2×2 dimension.

We follow the same procedure as above to find the joint probability density of the eigenvalues for the GUE case. Here we consider a 2×2 complex Hermitian matrix $\mathcal{H}$, and a simple 2×2 unitary transformation matrix $\mathcal{U}$ to diagonalize it, where

$$\mathcal{U} = \begin{pmatrix} \cos\theta & -\mathrm{e}^{-i\phi}\sin\theta \\ \mathrm{e}^{i\phi}\sin\theta & \cos\theta \end{pmatrix} \,. \tag{25}$$

However, we note that this unitary matrix is not the most general 2×2 unitary transformation matrix, but rather an element of the coset space $\mathcal{V}(2)/V(1)V(1)$. This is to ensure a one-to-one correspondence $\mathcal{H} \to (\lambda_1, \lambda_2, \mathcal{U})$ after the ordering of the eigenvalues, which are given by

$$\lambda_{1,2} = \frac{1}{2}(H_{11} + H_{22}) \pm \sqrt{(H_{11} - H_{22})^2 + 4H_{12}^* H_{12}} \,. \tag{26}$$

This expression is essentially the same as the one in Eq. (21) with the only difference being $H_{12}^2 \to |H_{12}|^2$. In complete analogy to Eq. (20) we write

$$\mathcal{H} = \mathcal{U}^{\dagger} \mathcal{H}_{\mathrm{D}} \mathcal{U} \tag{27}$$

to represent the matrix elements of $\mathcal{H}$ in terms of the two eigenvalues $\lambda_{1,2}$ and the two angles θ, ϕ

$$\begin{aligned} H_{11} &= \lambda_1 \cos^2\theta + \lambda_2 \sin^2\theta \\ H_{22} &= \lambda_1 \sin^2\theta + \lambda_2 \cos^2\theta \\ H_{12} &= (\lambda_1 - \lambda_2)[\cos\theta \sin\theta \cos\phi - i\cos\theta \sin\theta \sin\phi] \,, \end{aligned} \tag{28}$$

from which the Jacobian is

$$J = \det\left[\frac{\partial(H_{11}, H_{22}, H_{12}^R, H_{12}^I)}{\partial(\lambda_1, \lambda_2, \theta, \phi)}\right] = (\lambda_1 - \lambda_2)^2 \cos\theta \sin\theta \,. \tag{29}$$

Again, $\mathrm{Tr}[\mathcal{H}^2] = \lambda_1^2 + \lambda_2^2$ thus Eq. (19) can be written for GUE, in terms of the eigenvalues and the two parameters namely angles as

$$P(\lambda_1, \lambda_2, \theta, \phi) = C|\lambda_1 - \lambda_2|^2 \mathrm{e}^{-\frac{N\beta}{4\sigma^2}(\lambda_1^2 + \lambda_2^2)} \cos\theta \sin\theta \,, \tag{30}$$

integrating this over the angles θ and ϕ we get

$$P(\lambda_1, \lambda_2) = C|\lambda_1 - \lambda_2|^2 \mathrm{e}^{-\frac{N\beta}{4\sigma^2}(\lambda_1^2 + \lambda_2^2)} \,. \tag{31}$$

Although the constant C has been modified as a result of the integration, we still call it C to avoid unnecessary complication.

Now we move on to the GSE case, in which the smallest Hamiltonian matrix $\mathcal{H}$ is 4×4 but the procedure for finding the joint probability density of the eigenvalues is completely analogous to the preceding GOE and GUE cases.

The most convenient way to write $\mathcal{H}$ is the quaternion notation, reducing the $2N \times 2N$ matrix to an $N \times N$ by representing each H_{ij} with a block of 2×2, superposition of unity and the quaternion matrices. Then the degenerate eigenvalues of quaternion real $\mathcal{H}$ are

$$\lambda_{1,2} = \frac{1}{2}\left(H_{11}^0 - H_{22}^0\right) \pm \left[\left(H_{11}^0 - H_{22}^0\right)^2 + 4\sum_{\mu=0}^{3}\left(H_{12}^{\mu}\right)^2\right]^{1/2} \,, \tag{32}$$

each of these λs are double, so that we have 4 eigenvalues since the matrix is 4×4. The total number of freely

picked real entries in a $2N \times 2N$ quaternion real matrix $\mathcal{H}$ is given by $N(2N-1)$, thus we have 6 real entries in this case. To diagonalize $\mathcal{H}$, not the most general, but a convenient choice is a 2×2 symplectic matrix

$$S = \begin{pmatrix} e^{-\xi\cdot\tau}\cos\theta & \sin\theta \\ -\sin\theta & e^{\xi\cdot\tau}\cos\theta \end{pmatrix}, \tag{33}$$

where θ, ξ are the four real parameters provided by S. As before using the notation $\mathcal{H} = S^{\dagger}\mathcal{H}_{\mathrm{D}}S$, we write the elements of $\mathcal{H}$ as a function of the eigenvalues and the parameters θ, ξ,

$$\begin{aligned} H^0_{11} &= \lambda_1\cos^2\theta + \lambda_2\sin^2\theta \\ H^0_{22} &= \lambda_1\sin^2\theta + \lambda_2\cos^2\theta \qquad (34) \\ H^0_{12} &= -(\lambda_1-\lambda_2)\cos\theta\sin\theta\cos\xi \\ \mathbf{H}^{\mu}_{12} &= \xi(\lambda_1-\lambda_2)\cos\theta\sin\theta\sin\xi\,. \qquad (35) \end{aligned}$$

The Jacobian of this transformation is

$$J = \det\left[\frac{\partial(H^0_{11}, H^0_{22}, H^{\mu}_{12})}{\partial(\lambda_1,\lambda_2,\theta,\xi)}\right] \propto (\lambda_1-\lambda_2)^4, \tag{36}$$

where the coefficient of the proportionality is not specified since it is independent of the eigenvalues, and will only contribute to the constant C after integration over all four parameters θ and ξ. Thus, the joint probability distribution of the eigenvalues for the smallest GSE is

$$P(\lambda_1,\lambda_2) = C|\lambda_1-\lambda_2|^4 e^{-\frac{N\beta}{4\sigma^2}(\lambda_1^2+\lambda_2^2)}\,. \tag{37}$$

From the general form of the Eqs. (24,31,37), we can guess the generic joint probability distribution of the eigenvalues for all three Gaussian ensembles to be

$$\begin{aligned} &P_{N\beta}(\lambda_1,\ldots,\lambda_N) \\ &\quad = C_{N\beta}\prod_{i<j}|\lambda_i-\lambda_j|^{\beta}\exp\left(-\frac{N\beta}{4\sigma^2}\sum_{k=1}^{N}\lambda_k^2\right), \end{aligned} \tag{38}$$

in which the constant $C_{N\beta}$ is an ensemble dependent constant, and it is chosen such that the $P_{N\beta}$ is normalized to unity.

Density of States: Semicircle Law

Having at our disposal the generalized joint probability density of eigenvalues, now we can derive the famous *semicircle law* for the mean eigenvalue densities (see Fig. 1). The essential step is integrating Eq. (38) in the previous section i. e.,

$$N\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P_{N\beta}(\lambda_1,\lambda_2,\ldots,\lambda_N)\,d\lambda_2,\ldots,d\lambda_N \tag{39}$$

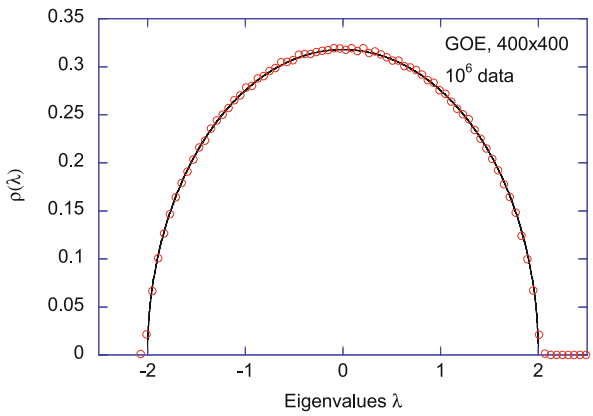

Random Matrix Theory, Figure 1
Averaged density of states (DOS) for GOE (with a suitable scaling, all three ensembles have the same form of DOS distribution). The *solid line* is the theoretical result superimposed over numerical data values (*circles*)

over all variables accept one to determine the ensemble-averaged eigenvalue or level density $\rho(\lambda)$, and the semicircle law arises asymptotically as $N \to \infty$. This integration is not at all trivial, and was first carried out by Mehta and Gaudin [39], using some remarkable techniques. However, there are many ways to derive the semicircle law, but the one that is perhaps the most elegant and up-to-date is the representation of determinants by Gaussian integrals over Grassmaninan variables, we will use this technique in the following derivations of the semicircle law.

General Formalism

Consider an $N \times N$ Hermitian matrix $\mathcal{H}$, which has ordered eigenvalues $\lambda_k (k = 1, \ldots, N)$. The normalized spectrum of $\mathcal{H}$ can be expressed as sums of delta function, which essentially picks out the positions of the eigenvalues,

$$\rho(\lambda) = \frac{1}{N}\sum_{k=1}^{N}\delta(\lambda-\lambda_k)\,. \tag{40}$$

By representing the delta function as a Lorentzian curve with vanishing widths.

$$\delta(\lambda-\lambda_k) = \lim_{\varepsilon\to 0}\frac{\varepsilon}{\pi}\frac{1}{(\lambda-\lambda_k)^2+\varepsilon^2}\,, \tag{41}$$

and substituting this latter relation to the former we get

$$\rho(\lambda) = \frac{1}{N\pi}\mathrm{Im}\,\mathrm{Tr}\left(\frac{1}{\lambda\mathcal{I}-\mathcal{H}}\right), \tag{42}$$

where $\mathcal{I}$ is identity matrix and λ is assumed to have a small imaginary part but we will not write it explicitly for the

sake of less cumbersome expressions. To proceed with the supersymmetric calculations we need to represent this last relation in terms of ratios of the determinants, for which we introduce the following generating function

$$\mathcal{Z}(\lambda,\mu) \equiv \frac{\det(\lambda\mathcal{I}-\mathcal{H})}{\det(\mu\mathcal{I}-\mathcal{H})}\,, \tag{43}$$

where $\mu \equiv \lambda \pm ix/2N$. Thus, differentiating the generating function with respect x, i. e.,

$$\frac{\partial}{\partial x}\mathcal{Z}(\lambda,\mu)|_{x\to 0} = \frac{1}{N}\mathrm{Im}\,\mathrm{Tr}\left(\frac{1}{\lambda\mathcal{I}-\mathcal{H}}\right)\,, \tag{44}$$

and when $x \to 0$ we recover the expression for the density of the eigenvalues we have defined previously. So that our starting point will be

$$\frac{\det(\lambda\mathcal{I}-\mathcal{H})}{\det(\mu\mathcal{I}-\mathcal{H})}\,. \tag{45}$$

We are interested in averaging the expression in Eq. (45) over all realization of $\mathcal{H}$. For which we will use the "supersymmetrization" method. In this method, the denominator of the expression is represented by a general Gaussian integral over a complex N dimensional vector $\mathbf{S} = (S_1, \ldots, S_N)^{\mathrm{T}}$, where T stands for the vector transpose, and S_i are commuting variables

$$(2\pi)\det{}^{-1}(\mu\mathcal{I}-\mathcal{H}) = \frac{1}{i^N}\int \mathrm{d}\mathbf{S}^\dagger\,\mathrm{d}\mathbf{S}\,\mathrm{e}^{\frac{i}{2}\mu\mathbf{S}^\dagger\mathbf{S}-\frac{i}{2}\mathbf{S}^\dagger\mathcal{H}\mathbf{S}}, \tag{46}$$

with the assumption that μ has a small positive imaginary part i. e. we choose $\mu = \lambda + ix/N$ so that the integral is convergent. For the determinant in the numerator, we use Gaussian integrals over anticommuting (Grassmannian) N components vectors χ, $\chi^\dagger$, which gives

$$(2\pi)^{-1}\det(\lambda\mathcal{I}-\mathcal{H}) = \frac{1}{i^N}\int \mathrm{d}\chi^\dagger\,\mathrm{d}\chi\,\mathrm{e}^{\frac{i}{2}\lambda\chi^\dagger\chi-\frac{i}{2}\chi^\dagger\mathcal{H}\chi}\,. \tag{47}$$

Thus, we write the expression in Eq. (45) as

$$\langle\cdots\rangle_{\mathcal{H}} = \int \mathrm{d}^2\chi \cdot \int \mathrm{d}^2\mathbf{S}\,\mathrm{e}^{\frac{i}{2}(\lambda\chi^\dagger\chi+\mu\mathbf{S}^\dagger\mathbf{S})}\left\langle \mathrm{e}^{-\frac{i}{2}(\mathbf{S}^\dagger\mathcal{H}\mathbf{S}+\chi^\dagger\mathcal{H}\chi)}\right\rangle_{\mathcal{H}}\,. \tag{48}$$

The commuting and the anti commuting variables appear symmetrically in this expression, which in fact gives the name "supersymmetry" to the method. To perform the average over the GUE matrix $\mathcal{H}$ we can decompose elements of $\mathcal{H}$ into its real and imaginary parts, $H_{kl} = H^R_{kl} + iH^I_{kl}$ and write

$$\left\langle \mathrm{e}^{-\frac{i}{2}\sum_{kl}H_{kl}A_{kl}}\right\rangle_{\mathcal{H}} = \left\langle \mathrm{e}^{\sum_k -\frac{i}{2}A_{kk}H^R_{kk}}\right\rangle \cdot \left\langle \mathrm{e}^{\sum_{k>l} -\frac{i}{2}(A_{kl}+A_{lk})H^R_{kl}}\right\rangle \cdot \left\langle \mathrm{e}^{\sum_{k>l}\frac{1}{2}(A_{kl}-A_{lk})H^I_{kl}}\right\rangle\,, \tag{49}$$

where $A_{kl} = S^*_k S_l + \chi^*_k\chi_l$. The explicit meaning of angle brackets is the average over the joint probability density for GUE, here we take

$$P(\mathcal{H}) = C_N\mathrm{e}^{-\frac{N}{2}\mathrm{Tr}[\mathcal{H}^2]}\,\mathrm{d}\mathcal{H}\,, \tag{50}$$

where C_N is a constant normalization factor and the probability distribution has zero mean and unity variance. By means of this probability distribution we get

$$\begin{aligned}\left\langle \mathrm{e}^{-\frac{i}{2}A_{kk}H^R_{kk}}\right\rangle &= \exp\left[-\frac{A^2_{kk}}{8N}\right]\,,\\ \left\langle \mathrm{e}^{-\frac{i}{2}(A_{kl}+A_{lk})H^R_{kl}}\right\rangle &= \exp\left[-\frac{(A_{kl}+A_{lk})^2}{16N}\right]\,,\\ \left\langle \mathrm{e}^{\frac{1}{2}(A_{kl}-A_{lk})H^I_{kl}}\right\rangle &= \exp\left[\frac{(A_{kl}-A_{lk})^2}{16N}\right]\,.\end{aligned} \tag{51}$$

Substitution of these equations into (49) yields

$$\left\langle \mathrm{e}^{-\frac{i}{2}\sum_{kl}H_{kl}A_{kl}}\right\rangle_{\mathcal{H}} \propto \mathrm{e}^{-\frac{1}{8N}\sum_{kl}A_{kl}A_{lk}}\,, \tag{52}$$

it is easy to show that $\sum_{kl}H_{kl}A_{lk} = \mathrm{Tr}[\mathcal{H}\mathcal{A}]$ and $\sum_{kl}A_{kl}A_{lk} = \mathrm{Tr}[\mathcal{A}^2]$ then the expression in Eq. (52) is just the identity

$$\left\langle \mathrm{e}^{-\frac{i}{2}\mathrm{Tr}[\mathcal{H}\mathcal{A}]}\right\rangle_{\mathrm{GUE}} \propto \mathrm{e}^{-\frac{1}{8N}\mathrm{Tr}[\mathcal{A}^2]}\,. \tag{53}$$

The far right exponent in Eq. (48) have the relations $\chi^\dagger\mathcal{H}\chi = -\mathrm{Tr}[\mathcal{H}\chi\otimes\chi^\dagger]$ and $\mathbf{S}^\dagger\mathcal{H}\mathbf{S} = \mathrm{Tr}[\mathcal{H}\mathbf{S}\otimes\mathbf{S}^\dagger]$. These relations imply that the right hand side exponent of the identity is:

$$\mathrm{Tr}[\mathcal{A}^2] = \mathrm{Tr}[\mathbf{S}\otimes\mathbf{S}^\dagger - \chi\otimes\chi^\dagger]^2. \tag{54}$$

Expanding the latter we get

$$\mathrm{Tr}[\mathcal{A}^2] = (\mathbf{S}^\dagger\mathbf{S})^2 - 2(\mathbf{S}^\dagger\chi)(\chi^\dagger\mathbf{S}) - (\chi^\dagger\chi)^2. \tag{55}$$

The quartic term in the far right of Eq. (55) needs to be "decoupled" before we can carry on with the integration, we can do so by using the usual Hubbard–Stratonovich transformation

$$\mathrm{e}^{\frac{1}{8N}(\chi^\dagger\chi)^2} = \int_{-\infty}^{\infty}\frac{\mathrm{d}q}{\sqrt{2\pi}}\mathrm{e}^{-\frac{q^2}{2}-\frac{q}{\sqrt{2N}}\chi^\dagger\chi}. \tag{56}$$

Thus, re-writing Eq. (48) with these new relations we have

$$\langle\cdots\rangle_{\mathcal{H}} = \int \mathrm{d}^2\mathbf{S}\exp\left\{-\frac{1}{8N}(\mathbf{S}^\dagger\mathbf{S})^2 + \frac{i}{2}\mu\mathbf{S}^\dagger\mathbf{S}\right\} \cdot \int_{-\infty}^{\infty}\frac{\mathrm{d}q}{\sqrt{2\pi}}\mathrm{e}^{-q^2/2}\int \mathrm{d}\chi^\dagger\,\mathrm{d}\chi \cdot \exp\left\{\chi^\dagger\left[1/2(i\lambda - q/N)\mathcal{I} - \mathbf{S}\otimes\mathbf{S}^\dagger/4N\right]\chi\right\}. \quad (57)$$

The last integral over the Grassmannian variables is the familiar Gaussian integral in Eq. (47), which leads to

$$\langle\cdots\rangle_{\mathcal{H}} = \int \mathrm{d}^2\mathbf{S}\mathrm{e}^{-\frac{1}{8N}(\mathbf{S}^\dagger\mathbf{S})^2+\frac{i}{2}\mu\mathbf{S}^\dagger\mathbf{S}} \cdot \int_{-\infty}^{\infty}\frac{\mathrm{d}q}{\sqrt{2\pi}}\mathrm{e}^{-\frac{q^2}{2}}\det\left[\frac{1}{2}\left(i\lambda - \frac{q}{N}\right)\mathcal{I} - \frac{\mathbf{S}\otimes\mathbf{S}^\dagger}{4N}\right]. \quad (58)$$

Further manipulation can be made by introducing the variable $q_F = i\lambda - q/\sqrt{N}$ and shifting the contour of integration in such a way that, the integral over q_F goes along the real axis, thus, we rewrite the above expression as

$$\langle\cdots\rangle_{\mathcal{H}} \propto \int_{-\infty}^{\infty}\frac{\mathrm{d}q_F}{\sqrt{2\pi}}\mathrm{e}^{-\frac{N}{2}(q_F-i\lambda)^2} \cdot \int \mathrm{d}^2\mathbf{S}\mathrm{e}^{\frac{1}{8N}(\mathbf{S}^\dagger\mathbf{S})^2+\frac{i}{2}\mu\mathbf{S}^\dagger\mathbf{S}}\det\left[q_F\mathcal{I} - \frac{1}{2N}\mathbf{S}\otimes\mathbf{S}^\dagger\right], \quad (59)$$

where we shifted the contour for $q_F \in (-\infty, \infty)$ to be real.

Further simplification can be made by noticing that the $N \times N$ matrix $\mathbf{S}\otimes\mathbf{S}^\dagger$ is of rank unity, i. e., it has $(N-1)$ zero eigenvalues, and only one nonzero eigenvalue equal to $(\mathbf{S}\mathbf{S}^\dagger)$. Then the determinant in the previous expression is

$$\det\left[q_F\mathcal{I} - \frac{1}{2N}\mathbf{S}\otimes\mathbf{S}^\dagger\right] \equiv q_F^{N-1}\left(q_F - \frac{1}{2N}\mathbf{S}^\dagger\mathbf{S}\right). \quad (60)$$

Our next step is to introduce polar coordinates: $\mathbf{S} = r\mathbf{n}$ with $\mathbf{n}^\dagger\mathbf{n} = 1$ and $\int \mathrm{d}^2\mathbf{S} = r^{2N-1}\mathrm{d}r\,\mathrm{d}\mathbf{n}$, where $\int \mathrm{d}\mathbf{n} = \Omega_N$ corresponds to the area of a $2N$ dimensional unit sphere, which only produces a constant factor. Further introducing $p = r^2$ and changing $p \to 2Np$ and then following with the obvious manipulations we get:

$$\left\langle\frac{\det(\lambda\mathcal{I} - \mathcal{H})}{\det(\mu\mathcal{I} - \mathcal{H})}\right\rangle_{\mathcal{H}} = C_N\mathrm{e}^{\frac{N}{2}\lambda^2}\int_{-\infty}^{\infty}\frac{\mathrm{d}q_F}{\sqrt{2\pi}\,q_F} \cdot \exp\left\{-\frac{N}{2}\left(q_F^2 - 2i\lambda q_F - 2\ln q_F\right)\right\} \cdot \int_0^{\infty}\frac{\mathrm{d}p(q_F - p)}{\sqrt{2\pi}\,p}\exp\left\{-\frac{N}{2}\left(p^2 - 2i\mu p - 2\ln p\right)\right\}. \quad (61)$$

In this expression C_N stands for the accumulated constant factors such that when $\mu = \lambda$ the right hand side must yield unity identically.

Up to now all the expressions were valid for finite-size matrices, but for $N \to \infty$ with appropriate scaling we expect the results to be universal. This means that in the asymptotic limit the results will be broadly insensitive to the details of the distribution of random matrices, thus it can be applicable to complex or quantum chaotic systems. More over in such a limit the integrals in Eq. (61) can be evaluated by the saddle-point method, or Laplace's method. To use the latter method we write the terms in the exponent in Eq. (61) as

$$\mathcal{L}(q_F) = [q_F^2 - 2i\lambda q_F - 2\ln q_F] \quad \text{and} \quad \mathcal{L}(p) = [p^2 - 2i\mu p - 2\ln p]. \quad (62)$$

The first derivatives of these equations yield the saddle-points: $q_F^{s.p} = (i\lambda \pm \sqrt{4-\lambda^2})/2$ and $p^{s.p} = (i\mu + \sqrt{4-\mu^2})/2$ since $Re(p) \geq 0$, for $\lambda \leq \sqrt{4}$ the so called *bulk* of the spectrum. The imaginary part of μ here denoted as ix/N does not contribute, when $N \to \infty$ also for consistency we should put $\mu = \lambda$. Because of the presence of the factor $(q_F - p)$ in the integrand, it is easy to see that only the choice $q_F^{s.p} = (i\lambda - \sqrt{4-\lambda^2})/2$ yields the leading-order contribution. Substituting this choice into the integrand in Eq. (61) and evaluating the Gaussian fluctuations around the saddle-point values, finally yields

$$\left\langle\frac{\det(\lambda\mathcal{I} - \mathcal{H})}{\det[(\lambda + ix/N)\mathcal{I} - \mathcal{H}]}\right\rangle_{\mathcal{H}}\Bigg|_{\substack{x>0\\N\to\infty}} = \exp\left\{-\frac{x}{2}[i\lambda + \sqrt{4-\lambda^2}]\right\}, \quad (63)$$

where we re-substituted $\mu = \lambda + ix/N$. To recover averaged density of states $\langle\rho(\lambda)\rangle$ from the expression above we simply need to differentiate both sides with respect to x, and let $x \to 0$, for which we write

$$\frac{\partial}{\partial x}\left\langle\det(\lambda - \mathcal{H})\exp\{-\mathrm{Tr}\ln[(\lambda + ix/N) - \mathcal{H}]\}\right\rangle = \frac{\partial}{\partial x}\exp\left\{-\frac{x}{2}[i\lambda + \sqrt{4-\lambda^2}]\right\}. \quad (64)$$

Performing the differentiation and letting $x \to 0$ we get

$$-i\left\langle \frac{1}{N}\,\mathrm{Tr}\left(\frac{1}{\lambda\mathcal{I}-\mathcal{H}}\right)\right\rangle_{\mathcal{H}}\Bigg|_{\substack{x\to 0\\ N\to\infty}} = -\frac{i\lambda}{2}-\frac{1}{2}\sqrt{4-\lambda^2}\Bigg|_{x\to 0} . \quad (65)$$

Comparing this last relation with Eq. (42) we see that the averaged density of states is given by

$$\langle\rho(\lambda)\rangle = \frac{1}{2\pi}\sqrt{4-\lambda^2} . \quad (66)$$

It is easy to repeat all the calculations for $x < 0$ and find that for any real value of x the result in Eq. (63) can be written as

$$\left\langle \frac{\det(\lambda\mathcal{I}-\mathcal{H})}{\det[(\lambda-ix/N)\mathcal{I}-\mathcal{H}]}\right\rangle_{\mathcal{H}}\Bigg|_{\substack{x<0\\ N\to\infty}} = \exp\left\{-\frac{i\lambda x}{2}-|x|\pi\rho(\lambda)\right\} . \quad (67)$$

Spacing Distribution: Wigner Surmise

The *level repulsion* phenomenon that the small spacings between two neighboring levels are unlikely can be seen from the expression $|\lambda_i-\lambda_j|^\beta$, which appeared as a result of the Jacobian of the transformation in all three cases of the Gaussian ensembles. Wigner in [56] conjectured that the distribution of spacings between adjacent eigenvalues may be given by

$$P(s) \approx Cs^\beta \exp(-As^2) . \quad (68)$$

He simply guessed this probability law, and demonstrated his argument by using a 2×2 real symmetric matrix. Through the joint probability distribution of the eigenvalues, and utilizing many novel techniques Gaudin obtained an exact expression for the spacing distribution. However, the derivations are rather lengthy and beyond the scope of this introductory section, hence we refer the reader to [26]. Here, only the simplest case of $N = 2$ dimensional ensembles of random matrices will be considered by taking the distribution of eigenvalues for $N = 2$. The spacing distribution for a single pair of levels can be written as

$$P(s) = C\int_{-\infty}^{\infty} \mathrm{d}\lambda_1 \cdot \int_{-\infty}^{\infty} \mathrm{d}\lambda_2\,\delta(s-|\lambda_1-\lambda_2|)|\lambda_1-\lambda_2|^\beta \mathrm{e}^{-\frac{A}{2}(\lambda_1^2+\lambda_2^2)} . \quad (69)$$

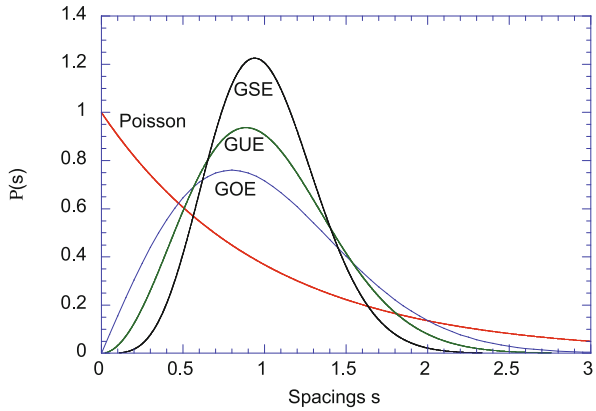

Random Matrix Theory, Figure 2
***Wigner Surmise*; Nearest neighbor spacing distributions for Poissonian random process, the orthogonal, unitary and symplectic ensembles**

Changing the variables as $\lambda_1 = (\sqrt{2}/2)(u+v)$ and $\lambda_2 = (\sqrt{2}/2)(u-v)$, makes the integration easy to perform. The constants A and C are fixed by the following normalization conditions:

$$\int_0^\infty P(s)\mathrm{d}s = 1 \quad \text{and} \quad \langle s\rangle \equiv \int_0^\infty sP(s)\mathrm{d}s = 1 , \quad (70)$$

the latter implies that the units of eigenvalues are such that the mean spacing is unity. Performing the elementary integrations for $\beta = 1, 2$ and 4 gives

$$P(s) = \begin{cases} \frac{\pi}{2}s\mathrm{e}^{-s^2\pi/4} & \beta = 1 \quad \text{(GOE)} \\ \frac{32}{\pi^2}s^2\mathrm{e}^{-s^2 4/\pi} & \beta = 2 \quad \text{(GUE)} \\ \frac{2^{18}}{3^6\pi^3 2}s^4\mathrm{e}^{-s^2 64/9\pi} & \beta = 4 \quad \text{(GSE)} \end{cases} \quad (71)$$

which are referred to as the *Wigner Surmise* (see Fig. 2) in the literature, and these are not very different from the asymptotic i. e., $N \to \infty$ derivations (see Fig. 3).

Two-Level Correlation Function

The two-level correlation function is a highly investigated topic in RMT, because it is very useful for spectral data analysis. A number of spectral correlations such as the *number variance* (for historical reasons also called Δ_3 statistics) are expressed in terms of it. It is defined as the probability density to find two eigenvalues (or levels) λ_1 and λ_2 at two given positions regardless of the positions of all the other eigenvalues. The more general form of this definition is given by the n-level correlation function

$$R_n(\lambda_1, \ldots, \lambda_n) = \frac{N!}{(N-n)!} \cdot \int P_N(\lambda_1, \ldots, \lambda_N)\mathrm{d}\lambda_{n+1} \ldots \mathrm{d}\lambda_N , \quad (72)$$

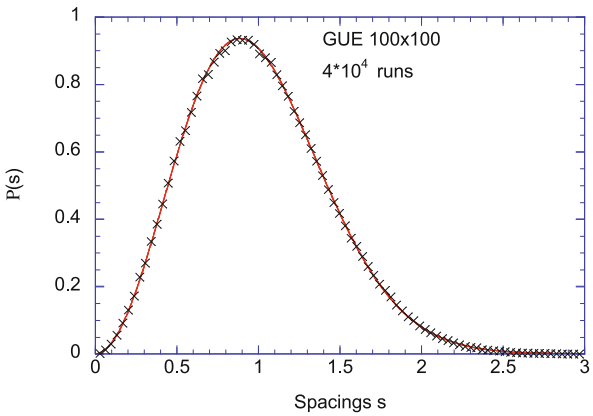

Random Matrix Theory, Figure 3
Comparison of nearest neighbor spacing distributions for 2 × 2 Wigner surmise (*solid line*) and numerical simulation (*crosses*) of 100 × 100 GUE

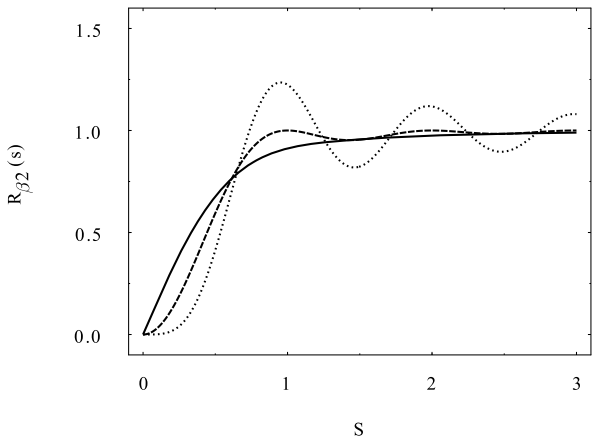

Random Matrix Theory, Figure 4
The two-level correlation functions $R_{\beta 2}(s)$ on the unfolded energy scale. The *solid line* is the GOE, the *dashed line* is the GUE and the *dotted line* is the GSE, which overshoots the value one due to strong oscillations

where the coefficient on the right hand side takes into account all combinations of selecting n eigenvalues out of the total number N. The integration is taken over the joint probability density P_N of eigenvalues. First systematic derivations of n-level correlation function were given by Dyson [12] and since then a number of other methods have been developed, for example the method of orthogonal polynomials and the supersymmetry approach. The detailed derivations of the n-level correlation and cluster functions for all three Gaussian ensembles can be found in Mehta's book, where the essential step of derivations involves the idea of expressing the joint probability density of eigenvalues as a determinant. However, the calculations are too lengthly, in particular the GOE case is rather complicated, and for our purpose it is sufficient to give only the following results of the two-level correlation functions $R_{\beta,2}(s)$ (see Fig. 4) for all Gaussian ensembles:

$$R_{1,2}(s) = 1 - \left[\left(\frac{\sin \pi s}{\pi s} \right)^2 + \frac{\mathrm{d}}{\mathrm{d}s} \left(\frac{\sin \pi s}{\pi s} \right) \int_s^\infty \frac{\sin \pi t}{\pi t} \mathrm{d}t \right],$$

$$R_{2,2}(s) = 1 - \left(\frac{\sin \pi s}{\pi s} \right)^2,$$

$$R_{4,2}(s) = 1 - \left[\left(\frac{\sin 2\pi s}{2\pi s} \right)^2 - \frac{\mathrm{d}}{\mathrm{d}s} \left(\frac{\sin 2\pi s}{2\pi s} \right) \int_0^s \frac{\sin 2\pi t}{2\pi t} \mathrm{d}t \right]. \tag{73}$$

In these summarized results s is measured in the units of mean spacing between neighboring eigenvalues.

Random Matrix Theory and Complex Systems

In this section we will present several examples where RMT is pertinent to complex systems ranging from complex networks to finance.

Spectra of Graphs

Many complex systems whether technological, social or biological have a network structure. The dynamics and the growth of these networks play a major role in our daily life. Spread of disease, information, traffic flow, electricity distribution; all depend on the structure of their related networks. Understandably, the complex network studies are of great interest to researchers in a diverse range of disciplines. From a mathematical point of view complex networks are modeled as graphs [4,7,16,20,21,33] which are sets of many interacting components namely vertices (points) connected by edges (lines).

The eigenvalues of the adjacency matrices of graphs are related to their topological structures and transport properties. Random Matrix Theory is popularly employed to predict the spectral properties of various graphs. Recent numerical studies show that the eigenvalue spectrum of real-world networks, graphs with scale free links distribution, deviates from those of uncorrelated random graphs. In the large N (number of vertices) limit spectral density of random graphs (not sparse [47]) converge to a semi-

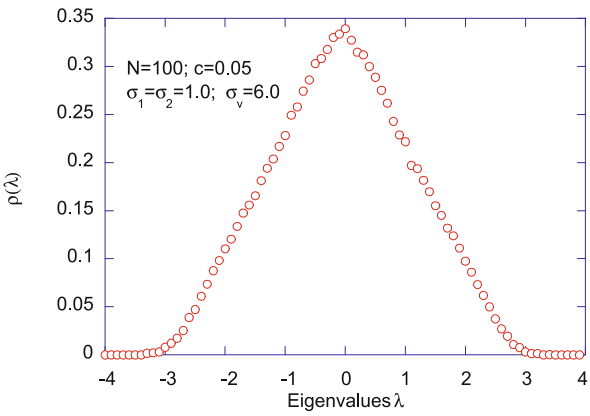

Random Matrix Theory, Figure 5
Two block diagonal random matrices with intra system interaction σ_1 and σ_2 coupled via very sparse (c = 0.05 connections) but strong σ_v couplings strength gives a triangular spectrum for a 100 × 100 system size

circle distribution, while the scale free networks develop a triangle density distribution with a power-law tail, and the small-world networks have a complex spectral density with several sharp peaks [22,23,28]. In a block structured random matrix model introduced by Ergun [17] to study coupled systems or a system with modules, it was shown that introduction of a few strong inter-couplings between sub-systems gives a triangular shape spectrum (see Fig. 5). This model can be thought of as a network with sub-networks with weighted links within and between subnetworks.

Composite Materials

Electrical and optical properties of metal-dielectric mixtures (composites) are frequently investigated by employing randomly connected Resistor-Capacitor (R-C) networks or Inductor (in series with a resistor)-Capacitor (RL-C) networks [1,2,6,8,34]. In such models one considers a binary inhomogeneous media with a random network of metallic bonds (conductance σ_0, concentration p) and dielectric bonds (conductance σ_1, concentration $1-p$). If the reactance network is large enough then there will always be circuits of resonance type with purely imaginary conductance showing poles at some frequencies. These poles are a particular example of eigenvalues of *matrix pencils* $\mathcal{H} - \lambda\mathcal{W}$ where $\mathcal{W}$ is a positive definite and $\mathcal{H}$ is a random real symmetric matrix.

Spectral studies of Kirchhoff or popularly known as Laplacian [50] matrices and random banded matrices [41] can be considered as the theoretical framework of the spectra of such networks.

Protein–Protein Interactions

Protein–protein interactions (PPIs) are essential for every living cell, information about these interactions helps to understand diseases (e. g. cancer), and to develop effective methodologies to treat them. There exists a vast amount of data on human PPIs and more are being mined every day. Currently, the use of network analogies [31,43] on these data suggests that PPIs form an interacting network of networks and have much in common with real world networks such as the World Wide Web and the Internet. In particular, the distribution of PPIs is scaled with power laws; such a scaling is often associated with robustness to random attacks. The underlying principles of power laws in networks with human attributes are identified to be: (1) continuous growth, and (2) preferential attachment processes. These considerations cannot hold at a microscopic level and therefore one must consider some other mechanisms such as spatial constraints and chemical reactions to account for the hierarchical organizations at the cellular level. In an empirical study of the yeast core protein interaction network Luo et al. [35] showed that the Wigner Surmise for the GOE case is a good candidate to see the spectral correlation of the yeast interaction matrix. Their study is essentially confirming that the yeast core protein network is a network of sub networks, which share the spectral properties of block-structured random networks.

To put their observation onto a theoretical footing consider the following: If $\mathcal{M}$ is a block diagonal matrix with $\mathcal{H}_1$ and $\mathcal{H}_2$ being its two blocks of $N \times N$ matrices with the same statistics then the eigenvalues of the sub matrices will form two uncorrelated sub sequences leading to a spacing distribution with $\rho(s) = 1/2$ at $s = 0$.

$$\mathcal{M} = \begin{pmatrix} \mathcal{H}_1 & \oslash \\ \oslash & \mathcal{H}_2 \end{pmatrix} \Longrightarrow \begin{pmatrix} \lambda_1^1 & & & & & \\ & \ddots & & & & \\ & & \lambda_N^1 & & & \\ & & & \lambda_1^2 & & \\ & & & & \ddots & \\ & & & & & \lambda_N^2 \end{pmatrix}.$$

Further increasing the number of non-interacting sub matrices eventually leads to a Poisson spacing distribution where all eigenvalues are uncorrelated. Conversely, if we introduce interactions between $\mathcal{H}_1$ and $\mathcal{H}_2$ via an off diagonal block matrix then the two sub sequences of eigenvalues will merge to a correlated single sequence getting back to Wigner surmise spacing statistics.

Ecological Webs

In general, interactions between ecological species are highly non-linear and may obey a generalized Lotka–Volterra competition relation:

$$\dot{x}_i = x_i \left(1 - \sum_{j=1}^{N} a_{ij} x_j \right) ,$$

where x_i is a function of time giving the density of the ith species, a_{ij} represents the effect of inter-specific (if $i \neq j$), intra-specific (if $i = j$) interactions and a_{ij} forms the $N \times N$ interaction matrix $\mathcal{A}$. The stability of the possible equilibrium or time independent configurations of such systems may be studied by linearization of this equation around equilibrium. The linearized model can be written as

$$\dot{\mathbf{x}} = \mathcal{J}\mathbf{x} ,$$

where $\mathcal{J}$ is the Jacobian matrix having $J_{ij} = x_i^* a_{ij}$ elements with x_i^* equilibrium value of species i, and bold face $\mathbf{x}$ is the column vector of N species. In this formalism an equilibrium is stable if the eigenvalues, of the Jacobian matrix have negative real parts.

In the 1970s Gardner and Ashby [25] introduced a similar linear model

$$\dot{\mathbf{x}} = \mathcal{A}\mathbf{x}$$

to study how the stability of a system would change with its size or if the connections within the system were incomplete, i. e., the elements of the matrix $\mathcal{A}$ are taken from a specified distribution with probability C "connectance" if there is connection (or interaction) between x_i and x_j and set to zero with probability $(1 - C)$ if no-interaction. Their Monte Carlo-type study showed that large complex systems which are assembled randomly would be stable up to a critical level of connectance. May (1972) [36] furthered these investigations by introducing a random matrix model where the matrix

$$\mathcal{A} = \mathcal{B} - \mathcal{I} ,$$

$\mathcal{I}$ is a unit matrix and $\mathcal{B}$ is a random matrix. The elements of $\mathcal{B}$ are picked independently and identically from a probability distribution. From his analysis he linked the stability of a complex system to its spectral radius namely, the spectral radius of a random $N \times N$ matrix $\mathcal{A}$ with connectance C and average interaction strength α is given by

$$\rho(\mathcal{A}) = \alpha (NC)^{1/2} .$$

These investigations started a surge of research interest to employ appropriate random matrices as ecological interaction matrices (sometimes referred to as the community matrices) to study species stability and variability in ecosystems.

Quasi-Species

The macromolecular evolution model was introduced in the 1970s by Manfred Eigen [14] to describe the process of selection in a collection of self-reproducing macromolecular species by employing a set of non-linear equations. Jones et al. [32] showed that in equilibrium, one can determine the solutions of these equations via the spectrum of a real positive matrix which represents the replication rates and mutation probabilities of the macromolecules. The largest real eigenvalue of this matrix gives the average replication rates and mutation probabilities. The corresponding eigenvector, also called Perron vector, describes densities of the quasi-species and hence which can only take real and positive values. This condition limits the number of equilibrium solutions thus a particular interest is to determine what values of a positive matrix guarantee this selective behavior, see for instance the works of Rumschitzki [48] and McCaskill [37].

Financial Correlation Matrices

The study of correlation (or covariance) matrices has a long history in finance dating back to the Markowitz's theory of optimal portfolios. The fundamental question is: Given a set of financial assets characterized by their average return and risk, what is the optimal weight of each asset, such that the overall portfolio provides the best return for a fixed level of risk, or conversely, the smallest risk for a given overall return? This question becomes even more difficult when the data is noisy or incomplete. Potters et al. [46] has suggested using RMT statistics to clean the empirical correlation matrices to minimize possible biases in the estimation of future risk. They also showed that the largest eigenvalue of the correlation matrices has a financial interpretation of being the "market mode".

Summary

We have introduced a number of key concepts used in random matrix theory and given several examples where RMT is utilized. To summarize, this theory was developed by physicists to describe the spectra of complex nuclei. But, nowadays, it is applied to many diverse research areas from quantum chaos, disordered systems to finance.

Below we give a short list of some of the frequently used results which were covered in this article.

The Gaussian ensembles are categorized into three universality classes: (1) Orthogonal ensembles, systems that have time-reversal invariance; the majority of practical systems are described by these ensembles, (2) unitary ensembles, which are used when the time-reversal invariance is violated, and (3) symplectic ensembles apply only to odd-spin systems without rotational symmetry.

The joint probability distribution for all Gaussian ensembles with $N \times N$ random matrix H is given by the general expression

$$P_{N\beta}(H) = C_{N\beta}\, \mathrm{e}^{-\frac{N\beta}{4\sigma^2}\mathrm{Tr}[\mathcal{H}^2]} ,$$

where Tr stands for the *trace*, and β is used as an index for the ensemble classification. Then the eigenvalues of the Gaussian ensembles are distributed according to

$$P_{N\beta}(\lambda_1, \ldots, \lambda_N) = C_{N\beta} \prod_{i<j} |\lambda_i - \lambda_j|^\beta \exp\left(-\frac{N\beta}{4\sigma^2}\sum_{k=1}^{N}\lambda_k^2\right) ,$$

where the constant $C_{N\beta}$ is an ensemble dependent constant, and it is chosen such that the $P_{N\beta}$ is normalized to unity. This expression is common for all Gaussian ensembles and derived from the joint probability distribution given above. Integrating this expression over all eigenvalues except one gives the semicircle law asymptotically,

$$\langle\rho(\lambda)\rangle = \begin{cases} \frac{1}{2\pi\beta\sigma^2}(4N\beta\sigma^2 - \lambda^2)^{1/2} , & |\lambda| < 2\sqrt{\beta\sigma^2 N} \\ 0 , & |\lambda| > 2\sqrt{\beta\sigma^2 N} \end{cases} ,$$

where $\langle\rho(\lambda)\rangle$ is normalized to N, and σ^2 is the variance of the matrix elements.

The three Wigner laws (referred to as the *Wigner Surmise*) for the nearest neighbor spacing distributions for 2×2 matrices are

$$P(s) = \begin{cases} \frac{\pi}{2} s \mathrm{e}^{-s^2\pi/4} & \beta = 1 \quad (\mathrm{GOE}) \\ \frac{32}{\pi^2} s^2 \mathrm{e}^{-s^2 4/\pi} & \beta = 2 \quad (\mathrm{GUE}) \\ \frac{2^{18}}{3^6\pi^3 2} s^4 \mathrm{e}^{-s^2 64/9\pi} & \beta = 4 \quad (\mathrm{GSE}) \end{cases} .$$

They are also found to be very good approximations in the asymptotic limit i. e., $N \to \infty$ derivations.

Future Directions

The use of random matrices in sub-atomic level interactions has led to many great mathematical theorems and shed new light on the understanding of nuclear spectra. Other branches of physical and mathematical sciences equally benefited from the resultant theorems yet biological sciences haven't had the same share despite an important link between random matrices and the field of biological complexity that was established in the 1970s by Robert May. With recent technological advances massive amounts of DNA data sets are being produced, however, the understanding of protein–protein or gene interactions still lags behind. Studies of structured sparse random matrices in the context of DNA data sets may produce fruitful results. Another area where random matrix theory is not fully consulted is the interpretation of the spectra of the adjacency matrices of various graphs and the spectra of the Laplacian of these graphs. There is no doubt that Random Matrix Theory is an effective tool to study a variety of complex systems; one would also hope to see its usage in the theory of machine learning and climate modeling.

Bibliography

Primary Literature

1. Almond DP, Bowen CR (2004) Anomalous power law dispersions in AC conductivity and permittivity shown to be characteristics of microstructural electrical networks. Phy Rev Lett 92:1576011–1576014
2. Almond DP, Vainas B (1999) The dielectric properties of random R-C networks as an explanation of the universal power law dielectric response of solids. J Phys: Condensed Matter 11:9081–9093
3. Bak P, Sneppen K (1993) Punctuated equilibrium and criticality in a simple model of evolution. Phys Rev Lett 71:4083
4. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286:509
5. Bohigas O, Giannoni M-J, Schmit C (1984) Characterization of chaotic quantum spectra and universality of level fluctuation laws. Phys Rev Lett 52:1
6. Clerc P, Giraud G, Laugier JM, Luck JM (1990) The electrical conductivity of binary disordered systems, percolation clusters, fractals and related models. Adv Phys 39:191–309
7. Dorogovtsev SN, Mendes JFF, Samukhin AN (2000) Structure of Growing Networks: Exact Solution of the Barabási–Albert's Model. Phys Rev Lett 85:4633
8. Dykhne AM (1970) Anomalous resistance of a plasma in a strong magnetic field. Zh Eksp Teor Fiz 59:641–647
9. Dyson FJ (1962) A Brownian-motion model for the eigenvalues of a random matrix. J Math Phys 3:1191
10. Dyson FJ (1962) Statistical theory of the energy levels of complex systems, I. J Math Phys 3:140
11. Dyson FJ (1962) Statistical theory of the energy levels of complex systems, II. J Math Phys 3:157
12. Dyson FJ (1962) Statistical theory of the energy levels of complex systems, III. J Math Phys 3:166
13. Efetov KB (1983) Supersymmetry and theory of disordered metals. Adv Phys 32:53

14. Eigen M (1971) Selforganisation of sequence space and tensor products of representation spaces. Naturwissenschaften 58:465–523
15. Erdős P, Rényi A (1959) On random graphs I. Publ Math 6:290
16. Ergün G (2002) Human sexual contact network as a bipartite graph. Physica A 308:483
17. Ergün G, Semicircle to triangular distribution of density of states: Supersymmetric Approach (submitted)
18. Ergün G, Separation of a large eigenvalue from the bulk of the spectrum (submitted)
19. Ergün G, Fyodorov YV (2003) Level Curvature distribution in a model of two uncoupled chaotic subsystems. Phys Rev E 68:046124
20. Ergün G, Rodgers GJ (2002) Growing Random Networks with Fitness. Physica A 303:261
21. Ergün G, Zheng D (2003) Coupled Growing Networks. Adv Comp Syst 6:4
22. Farkas IJ, Derényi I, Barabási A-L, Vicsek T (2001) Spectra of real-world graphs: Beyond the semi-circle law. Phys Rev E 64:026704
23. Farkas IJ, Derényi I, Jeong H, Neda Z, Oltvai ZN, Ravasz E, Schubert A, Barabási A-L, Vicsek T (2002) Networks in life: Scaling properties and eigenvalue spectra. Physica A 314:25
24. Fyodorov YV, Chubykalo OA, Izrailev FM, Casati G (1996) Wigner random banded matrices with sparse structure: Local spectral density of states. Phys Rev Lett 76:1603
25. Gardner MR, Ashby WR (1970) Connectance of large daynamic systems: Critical values for stability. Nature 228:784
26. Gaudin M (1961) Sur la loi limite de l'espacement des valeurs propres d'une matrice aléatoire. Nucl Phys 25:447
27. Gaudin M, Mehta ML (1960) On the density of eigenvalues of a random matrix. Nucl Phys 18:420
28. Goh K-I, Kahng B, Kim D (2001) Spectra and eigenvectors of scale-free networks. Phys Rev E 64:051903
29. Gurevich II, Pevsner MI (1957) Repulsion of nuclear levels. Nucl Phys 2:575
30. Haake F (2000) Quantum Signatures of Chaos, 2nd edn. Springer, Berlin
31. Jeong H, Mason SP, Barabási A-L, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411:41–42
32. Jones BL, Enns RH, Rangnekar SS (1976) On the theory of selection of coupled macromolecular systems. Bull Math Biol 38:15–28
33. Krapivsky PL, Redner S, Leyvraz F (2000) Connectivity of growing random networks. Phys Rev Lett 85:4629
34. Luck JM, Jonckheere T (1998) Dielectric resonances of binary random networks. J Phys A: Math Gen 31:3687–3717
35. Luo F, Zhongb J, Yang Y, Scheuermann RH, Zhou J (2006) Application of random matrix theory to biological networks. Phys Lett A 357:420–423
36. May RL (1972) Will a large complex system be stable? Nature 238:413
37. McCaskill JS (1984) Localisation threshold for macromolecular quasispecies from continuously distributed replication rates. J Chem Phys 80:5194–5202
38. Mehta ML (1991) Random Matrices, 2nd edn. Academic, San Diego
39. Mehta ML, Gaudin M (1960) On the density of eigenvalues of a random matrix. Nuc Phys 18:420
40. Mirlin AD (2000) Statistics of energy levels and eigenfunctions in disordered systems. Phys Rep 326:259
41. Mirlin AD, Fyodorov YV (1991) Universality of level correlation function of sparse random matrics. J Phys A 24:2273–2286
42. Newman MEJ, Strogatz SH, Watts DJ (2001) Random graphs with arbitrary degree distributions and their applications. Phys Rev E 64:026118
43. Pellegrini M, Haynor D, Johnson JM (2004) Protein interaction networks. Expert Rev Proteomics 1:239–249
44. Porter CE, Rosenzweig N (1960) Statistical properties of atomic and nuclear spectra. Suomalaisen Tiedeakatemian Toimituksia (Ann. Acad. Sci. Fennicae) AVI Phys 44:166
45. Porter CE, Thomas RG (1956) Fluctuations of nuclear reaction widths. Phys Rev 104:483
46. Potters M, Bouchaud J-P, Laloux L (2005) Financial applications of random matrix theory: Old laces and new pieces. physics/0512090
47. Rodgers GJ, Bray AJ (1988) Density of states of a sparse random matrix. Phys Rev B 37:3557
48. Rumschitzki DS (1987) Spectral properties of Eigen evolution matrices. J Math Biol 24:667–680
50. Stäring J, Mehlig B, Fyodorov YV, Luck JM (2003) On the random symmetric matrices with a constraint: The spectral density of random impedence networks. Phys Rev E 67:047101
51. von Neuman J, Wigner E (1929). Phys Z 30:467
52. Watts DJ, Strogatz SH (1998) Collective dynamics of "small-world" networks. Nature 393:440
53. Wigner EP (1951) On the statistical distribution of the widths and spacing of nuclear resonance levels. Phil Soc 62:548
54. Wigner EP (1955) Characteristic vectors of bordered matrices with infinite dimensions. Ann Math 62:548
55. Wigner EP (1957) Distribution of neutron resonance levels. In: International conference on the neuron interactions with the nucleus, Columbia University, New York, 9–13 September 1957. Columbia Univ. Rept. CU-175 TID-8547
56. Wigner EP (1957) Results and theory of resonance absorption (Conference on Neutron Physics by Time-of-Flight, Gatlinburg, Tenessee, November 1–2, 1956). Oak Ridge Natl Lab Rept ORNL-2309:59
57. Wigner EP (1957) Statistical Properties of Real Symmetric Matrices with Many dimensions. In: Can. Math. Congr. Proc., 174. Univ of Toronto Press, Toronto
58. Wigner EP (1958) On the distribution of the roots of certain symmetric matrices. Ann Math 67:2
59. Wishart J (1928) The generalised product moment distribution in samples from a normal multivariate population. Biometrica 20A:32

Books and Reviews

Abramowitz M, Stegun I (1972) Handbook of mathematical fucntions, 10th edn. Dover, New York

Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47

Bak P (1996) How Nature Works: The science of self-organised criticality. Copernicus, New York

Barabási A-L (2002) Linked: How everything is connected to everything else and what it means. Perseus, Cambridge

Bellman R (1997) Introduction to Matrix Analysis, 2nd edn. SIAM, Philadelphia

Berezin FA (1987) Introduction to Superanalysis. Reidel, Dordrecht

Bohigas O (1991) Random matrix theories and chaotic dynamics, Session LII, 1989. Chaos and Quantum Physics, Les Houches

Bollobás B (2001) Random Graphs, 2nd edn. Cambridge Univ Press, Cambridge
Brody TA, Flores J, French JB, Mello PA, Pandey A, Wong SSM (1981) Random-matrix physics: spectrum and strength fluctuations. Rev Mod Phys 53:385
Cvetković D, Domb M, Sachs H (1995) Spectra of Graphs: Theory and Applications. Johann Ambrosius Barth, Heidelberg
Cvetković D, Rowlinson P, Simić S (1997) Eigenspaces of Graphs. Cambridge Univ Press, Cambridge
Dorogovtsev SN, Mendes JFF (2002) Evolution of networks. Adv Phys 51:1079
Efetov KB (1997) Supersymmetry in Disorder and Chaos. Cambridge Univ Press, Cambridge
Faloutsos M, Faloutsos P, Faloutsos C (1999) On power-law relationships of the internet topology. Proc ACM SIGCOMM Comp Comm Rev 29:251
Fyodorov YV (1995) Basic features of Efetov's supersymmetry approach. Mesoscopic Quantum Physics, Les Houches, Session LXI, 1994
Guhr T, Müller-Groeling A, Weidenmüller HA (1998) Random matrix theories in quantum physics: common concepts. Phys Rep 299:189
Hinch EJ (1991) Perturbation Methods. Cambridge Univ Press, Cambridge
Lieb EH, Mattis DC (1966) Mathematical Physics in One Dimension. Academic, New York
Markowitz H (1959) Portfolio Selection: Efficient Diversification of Investments. Wiley, New York
Merzbacher E (1970) Quantum Mechanics, 2nd edn. Wiley, London
Porter CE (1965) Statistical Theories of Spectra: fluctuations. Academic, New York
Stöckmann H-J (2000) Quantum Chaos: an introduction. Cambridge Univ Press, Cambridge
Watts DJ (1999) Small Worlds. Princeton Univ Press, Princeton
Wilks S (1972) Mathematical Statistics. Wiley, Japan
Wilson RJ (1996) Graph Theory. Longman, Edingburg

Random Walks in Random Environment

Ofer Zeitouni
Department of Mathematics, University of Minnesota, Minneapolis, USA

Article Outline

Glossary
Definition of the Subject
Introduction
One Dimensional RWRE
Multi Dimensional RWRE – non Perturbative Regime
Multi Dimensional RWRE – the Perturbative Regime
Diffusions in Random Environments
Topics Left Out and Future Directions
Bibliography

Glossary

Aging A stochastic process X_t exhibits aging if for some function $f(x, y)$, the limit of $Ef(X_t, X_{t+s_t})$ as $t \to \infty$ exists and is not trivial for a function $s_t \to_{t\to\infty} \infty$.

Annealed law Average of the quenched law, over all possible environments.

Brownian motion A continuous process $\{X_t\}$ with continuous sample paths, $X_0 = 0$, with increments over non-overlapping intervals independent and normally distributed with zero mean and variance equal to the length of the interval.

Convex A function $f\colon X \to \mathbb{R}$, where X is a linear space, is convex if for $\alpha \in [0, 1]$ and $x, y \in X$, $f(\alpha x + (1-\alpha)y) \le \alpha f(x) + (1-\alpha)f(y)$.

Ergodic law If P is a law on a collection of random variables $\{X - z\}_{z\in\mathbb{Z}^d}$, then it is ergodic if for any event A that is invariant under a transformation $X_z \mapsto X_{z+e}$, with $e \in \mathbb{Z}^d$, $|e| = 1$, it holds that $P(A) \in \{0, 1\}$.

Environment Collection of transition probabilities indexed by the sites of the lattice $\mathbb{Z}^d$, that is a collection of vectors belonging to the $2d$-simplex.

i.i.d. Independent identically distributed.

Large deviations principle (LDP) A sequence of probability measures P_n on a common (topological) space X satisfies the large deviations principle if for some non-negative function $I\colon X \to \mathbb{R}_+$ with closed level sets $\{x \in X\colon I(x) \le a\}$, for any $A \subset X$,

$$-\inf_{x\in A^o} I(x) \le \liminf n^{-1}\log P_n(A) \le \limsup n^{-1}\log P_n(A) \le -\inf_{x\in\bar{A}} I(x)$$

where $\bar{A}$ is the closure of A and A^o is its interior.

Law Probability distribution. When dealing with a stochastic process $\{X_t\}$, this is the collection of distribution functions $P(X_{t_1} \le x_1, \ldots, X_{t_k} \le x_k)$.

Law of large numbers (LLN) A sequence X_n satisfies the LLN if X_n/n converges to a deterministic limit.

Lipschitz norm Given a function on a normed space X with norm $\|\cdot\|$, its Lipschitz norm is defined as $\sup_{x\neq y} |f(x) - f(y)|/\|x - y\|$.

Martingale Let $\mathcal{F}_n$ denote a filtration, that is an increasing sequence of σ-algebras. A process $\{X_n\}$ is a (discrete time) martingale with respect to $\mathcal{F}_n$ if $E(X_{n+1}|\mathcal{F}_n) = X_n$.

Mixing environment An environment whose law satisfies a *mixing condition*, that is if A, B are events that depend on a finite number of sites then $P(A \cap T^z B) \to_{|z|\to\infty} P(A)P(B)$ where T^z is the shift of B by $z \in \mathbb{Z}^d$. For $d = 1$, the environment is strongly mix-

ing if the convergence above is uniform in the choice of A, B of the same local dependence.

P-a.s. A property holds P-a.s. if the probability of the event that the property does not hold (under the law P) vanishes.

Quenched law Law of the random walk, with the environment given.

Variational distance For two probability distributions ν, μ, the variational distance is defined as $\sup_A |\nu(A) - \mu(A)|$.

2d-Simplex Collection of vectors of length $2d$ with nonnegative entries that sum to 1.

Definition of the Subject

The term random walks in random environments (abbreviated RWRE) refers to a class of Markov chains evolving in $\mathbb{Z}^d$, where the transition mechanism itself is random, and forms a stationary random field indexed by $\mathbb{Z}^d$. They first appeared in the literature in the early 1970s, for $d = 1$, modeling in a natural way transport properties of a tagged particle in an inhomogeneous medium, as well as biological systems. They can be naturally used to model complex transport phenomena in any dimension.

The analysis of the case $d = 1$ during the 1970s and early 1980s, revealed that the behavior of such random walks can differ drastically from that of ordinary random walks (that correspond to a homogeneous environment). The extension to higher dimensions ($d \geq 2$) was proposed and studied in [51], but it is only in the last decade, starting with the work [102], that rapid progress has been made, although several basic questions remain open.

Introduction

We recall that random walks and their scaling limits, diffusion processes, provide a simple yet powerful description of random processes, and are fundamental in the description of many fields, from biology through economics, engineering, and statistical mechanics. A large body of work has accumulated concerning the properties of such processes, and very detailed information is available. We refer to [61] and [93] for background on random walks and diffusion processes.

In many situations, the medium in which the process evolves is highly irregular. Without further modeling, this results with spatially inhomogeneous Markov processes, and not much can be said. Things are however different if some degree of homogeneity is assumed on the *law* of the environment. When the underlying state space on which the walk moves with nearest neighbor steps is the lattice $\mathbb{Z}^d$, $d \geq 1$, and the law of the environment is assumed stationary, we call the resulting random walk a *random walks in random environment* (RWRE).

Informally, one starts with an *environment*, which is a configuration of transition probabilities (one transition probability for each site in $\mathbb{Z}^d$). One then starts a random walk that, when at location $z \in \mathbb{Z}^d$, moves at the next time step according to the transition probability associated with z. In the one dimensional case, this model has been thoroughly analyzed, and it is known that the behavior can differ dramatically from the behavior of ordinary random walk (in particular, the standard central limit theorem can fail). In higher dimensions, it is expected that the RWRE behaves more like ordinary random walk, and much of the available analysis points in this direction; however, gaps in the understanding of the model still remain, and the law of large numbers has not yet been proved in full generality.

A precise formulation of the RWRE model is as follows. Let S denote the 2d-dimensional simplex, set $\Omega = S^{Z^d}$, and let $\omega(z, \cdot) = \{\omega(z, z+e)\}_{e \in Z^d, |e|=1}$ denote the coordinate of $\omega \in \Omega$ corresponding to $z \in Z^d$. ω is an "environment" for an inhomogeneous nearest neighbor random walk (RWRE) started at x with *quenched* transition probabilities $P_\omega(X_{n+1} = z+e | X_n = z) = \omega(x, x+e)$ ($e \in \mathbb{Z}^d, |e| = 1$), whose law is denoted P^x_ω. We write E^x_ω for expectations with respect to the law P^x_ω, and write P_ω and E_ω for P^0_ω and E^0_ω. In the RWRE model, the environment is random, of law P, which is always assumed stationary and ergodic. We often assume that the environment is *uniformly elliptic*, that is there exists an $\epsilon > 0$ such that P-a.s., $\omega(x, x+e) \geq \epsilon$ for all $x, e \in \mathbb{Z}^d, |e| = 1$. Finally, we denote by $\mathbb{P}$ the *annealed* law of the RWRE started at 0, that is the law of $\{X_n\}$ under the measure $P \times P^0_\omega$, and again we write $\mathbb{E}$ for expectations with respect to $\mathbb{P}$ and E for expectations with respect to P. Note that under the law P_ω, the RWRE is a (space inhomogeneous) Markov process, whereas under $\mathbb{P}$, its law is space homogeneous but not Markovian, since the knowledge of the path of the RWRE up to time n modifies the a-priori law on the environment.

To get an idea of some of the unusual features of the RWRE model, we begin by discussing the one dimensional case. (A further motivation is that the study of certain reinforced random walks can be reduced to that of one dimensional RWRE, a connection first described in [73]). This model is fairly well understood, and we review the results in Sect. "One Dimensional RWRE", emphasizing the points in which there are differences from walks in a homogeneous environment. We then turn in Sect. "Multi Dimensional RWRE – non Perturbative Regime" to the multidimensional case, while Sect. "Multi Dimensional RWRE – the Perturbative Regime" is devoted to the per-

turbative regime. Sect. "Diffusions in Random Environments" quickly reviews the available results for the related model of (non reversible) diffusions in random environments. In Sect. "Topics Left Out and Future Directions", we collect some information about related models. This article borrows heavily from [101,107] and [108].

One Dimensional RWRE

When $d = 1$, we write $\omega_x = \omega(x, x+1)$, $\rho_x = (1-\omega_x)/\omega_x$, and $u = E\log\rho_0$. Because explicit recursions and computations are possible in the one dimensional case, the understanding of the RWRE is rather complete. We summarize the main features below.

Ergodic Behavior, Limit Laws, and Traps

Recall that for a homogeneous environment (that is, when the stationary measure P has a marginal which charges a single value: $\omega_i = \bar\omega$ for all i), we have $X_n/n \to v_{\bar\omega} := 2\bar\omega - 1$ and $(X_n - nv_{\bar\omega})/\sqrt{4\bar\omega(1-\bar\omega)n}$ converges in distribution to a standard Gaussian. We describe next the corresponding results for the RWRE model, emphasizing the dramatic difference in behavior.

As it turns out, the sign of u determines the direction of escape of the RWRE, while the limiting behavior depends on an explicit function of the law of the environment. The following theorem is essentially due to [92], see also [1,107].

Theorem 1 (Transience, recurrence, limit speed, $d = 1$) *(a) If $u < 0$ then $X_n \to_{n\to\infty} \infty$, $\mathbb{P}$-a.s. If $u > 0$ then $X_n \to -\infty$, $\mathbb{P}$-a.s. Finally, if $u = 0$ then the RWRE oscillates, that is, $\mathbb{P}$-a.s.,*

$$\limsup_{n\to\infty} X_n = \infty\,, \quad \liminf_{n\to\infty} X_n = -\infty\,.$$

Further, there is a deterministic v such that

$$\lim_{n\to\infty} \frac{X_n}{n} = v\,,\ \mathbb{P} - a.s.\,. \tag{1}$$

If P is a product measure, then

$$v = \begin{cases} (1-E(\rho_0))/(1+E(\rho_0))\,, & E(\rho_0) < 1\,, \\ -(1-E(\rho_0^{-1}))/(1+E(\rho_0^{-1}))\,, & E(\rho_0^{-1}) < 1\,, \\ 0\,, & else\,. \end{cases} \tag{2}$$

The statement (1) that X_n/n converges to a deterministic limit (under both the quenched and annealed measures) is referred to as a *law of large numbers* (LLN).

Remark 2 The surprising features of the RWRE model can be appreciated if one notes the following facts, all for a product measure P.

- Suppose $u < 0$, that is $X_n \to \infty$, $\mathbb{P}$-a.s. By Jensen's inequality, $\log E\rho_0 \geq E\log\rho_0$, but it is quite possible that $E\rho_0 > 1$. Thus, it is possible to construct i.i.d. environments in which the RWRE is transient, but the speed v vanishes.
- Suppose $\bar v = 2E\omega_0 - 1$ denotes the speed of a (biased) simple random walk with probability of jump to the right equal, at any site, to $E\omega_0$. It is easy to construct examples with $\bar v > 0$ but $u > 0$, which means that $X_n \to -\infty$ even if the static speed $\bar v$ points to the right. However, by Jensen's inequality, $v < 0$ implies that $\bar v < 0$. Thus, if the static speed $\bar v$ is positive, the RWRE may be transient to the left but if so, only with zero speed. We come back to this point in Subsubsect. "RWRE with Deterministic Components", where we show that the last property is not necessarily true in high dimension.
- Another application of Jensen's inequality reveals that $|v| \leq |\bar v|$, with examples of strict inequality readily available. Thus, the effect of the random environment is to force a *slowdown* with respect to the (averaged, deterministic) environment.

Another aspect in which the RWRE differs from the standard model of random walk is in its fluctuations. Consider product measures P with $u := E(\log\rho_0) < 0$ (i. e., RWRE is transient to $+\infty$). Set $s = \sup\{r\colon E(\rho_0^r) < 1\}$ and note that because $u < 0$, necessarily $s \in (0, \infty]$. When $s > 2$, the behavior is similar to standard random walk, and one has a central limit theorem of the following form. For some deterministic strictly positive constant σ, the random variable $W_n := (X_n - nv)/\sigma\sqrt{n}$ converges, under the annealed law, to a standard Gaussian random variable, that is

$$\mathbb{P}(W_n > x) \to_{n\to\infty} \frac{1}{\sqrt{2\pi}} \int_x^\infty \mathrm{e}^{-y^2/2}\mathrm{d}y\,,$$

see [46,55,75,107] for this statement and a similar one under the quenched law (with random centering). On the other hand, when $s < 2$, due to the existence of localized pockets of environments ("traps") where the walk spends a large time, one gets stable limit laws. In particular, for $s \in (0, 1)$, under mild assumptions on the law of ρ_0, the random variable X_n/n^s converges, under the annealed law, to a stable random variable with parameters (s, b), where b is a deterministic constant, whose value has been identified

in [39]. (A stable law with parameters (s, b) is the distribution of a random variable S with characteristic function

$$E(e^{itS}) = \exp\left(-b|t|^s\left(1 - i\frac{t}{|t|}\tan(\pi s/2)\right)\right), \ t \in \mathbb{R}.)$$

We refer to the regime where $s \in (0, 2]$, in which the fluctuations are not in the CLT scale, [55], as the *sub-diffusive* regime. It is interesting to note that these statements do *not* possess a quenched counterpart, and in fact a quenched limit law is not possible for $s < 2$, see [75,76].

Remark 3 When P is a strongly mixing environment, the parameter s has to be defined differently, by means of the large deviations rate function for the variable $n^{-1}\sum_{i=1}^{n}\log\rho_i$. The Gaussian limit laws apply in such situations when $s > 2$, see [20,75,107]. The stable limit laws are more delicate, and are not known for general ergodic environments with good mixing properties. For a class of Markovian environments, such results are contained in [67].

Limit Laws and Aging, Recurrent RWRE: Sinai's Walk

When $E(\log\rho_0) = 0$, the traps alluded to in the previous section stop being local, and the whole environment becomes a diffused trap. The walk spends most of its time "at the bottom of the trap", and as time evolves it is harder and harder for the RWRE to move. This is the phenomenon of *aging*, captured in the following theorem:

Theorem 4 *There exists a random variable B^n, depending on the environment only, such that for any $\eta > 0$,*

$$\mathbb{P}\left(\left|\frac{X_n}{(\log n)^2} - B^n\right| > \eta\right) \underset{n\to\infty}{\to} 0 .$$

Further, for $h > 1$,

$$\lim_{\eta\to 0}\lim_{n\to\infty}\mathbb{P}\left(\frac{|X_{n^h} - X_n|}{(\log n)^2} < \eta\right) = \frac{1}{h^2}\left[\frac{5}{3} - \frac{2}{3}e^{-(h-1)}\right]. \tag{3}$$

The first part of Theorem 4 is due to Sinai [91], with Kesten [52] providing the evaluation of the limiting law of B^n, see also [47]. It is actually not hard to understand the anomalous scaling $(\log n)^2$: indeed, the time for the particle to overcome a stretch of the environment of length $c_1\log n$ in which the drift points "backwards" is exponential in $c_1\log n$, i. e. an appropriate c_1 can be chosen such that this time is of order n. Hence, the range of the RWRE at time n cannot be larger than the distance in which there exists such a stretch. Due to the scaling properties of random walk, this distance is of order $(\log n)^2$.

The second part of Theorem 4 is implicit in [47], and also follows from the analysis of the time spent by the RWRE at "bottom of traps". We refer to [62] for a detailed study of aging in the Sinai model by renormalization techniques, and to [24,30,107] for rigorous proofs that avoid renormalization arguments. For information concerning the time spent by the walk at the most visited site (which can be of order n in the Sinai model), see [28,49,89], and [43] for the transient case.

Tail Estimates and Large deviations

Another question of interest relates to the probability of seeing a-typical behavior of the RWRE. These probabilities turn out to depend on the measure discussed, that is whether one considers the quenched or annealed measures.

Following Varadhan [105], we say that the sequence of random variables X_n/n satisfies the large deviations principle (LDP) with convex, continuous rate function I on a compact set K, if for any measurable subset of K,

$$\lim_{n\to\infty}\frac{1}{n}\log P(X_n/n \in A) = -\inf_{x\in A} I(x) .$$

Cramér's theorem (Theorem 2.2.3 in [29]) states that rescaled random walk X_n/n in a homogeneous environment with $\omega_i = \bar{\omega}$ for all i satisfies the LDP with a strictly convex rate function $I(x)$ that vanishes only on $v_{\bar{\omega}} = 2\bar{\omega} - 1$. The situation is different for the RWRE. The following theorem combines results from [25,32,48].

Theorem 5 *For P-a.e. realization of the environment ω, the sequence X_n/n satisfies, under P^0_ω, a LDP on $[-1, 1]$ with a deterministic convex rate function $I_P(\cdot)$. If P is an i.i.d. measure, then under the annealed measure $\mathbb{P}$, the same sequence satisfies a LDP with convex rate function*

$$I(x) = \inf_{Q\in\mathcal{M}_1^e}(h(Q|P) + I_Q(x)) , \tag{4}$$

where $h(Q|P)$ is the specific entropy of Q with respect to P and $\mathcal{M}_1^e$ denotes the space of stationary ergodic measures on Ω. Always, $I(x) \le I_P(x)$, and both I and I_P may vanish for $x \in [0, v]$, and only for such x. In particular, neither I nor I_P need be strictly convex.

The rate function for the RWRE thus differs from the case of homogeneous environments in two important aspects: it may vanish on the whole segment $[0, v]$, indicating subexponential behavior for the probability of slowdown, and further the rate function is in general not strictly convex. When $I(x)$ vanishes for $x \in [0, v]$, it means that the probability of seeing an a-typical slowdown of the random walk

decays at a speed less than exponentially. A precise characterization of these is available in [33,44,77,78,107].

The *specific entropy* $h(Q|P)$ (see [29] for definition) appearing in Theorem 5 measures the rate of decay of the probability that a block of the environment will look as if it came from Q, when it was generated by P. Thus, one may interpret Theorem 5 as follows: to create an annealed large deviation, one may first "modify" the environment (at a certain exponential cost measured by the specific entropy h) and then apply the quenched LDP in the new environment.

Multi Dimensional RWRE – non Perturbative Regime

We turn our attention to RWRE in the lattice $\mathbb{Z}^d$ with $d > 1$. Unless stated otherwise explicitly, we only consider in the sequel measures P that are i.i.d. and uniformly elliptic.

Ergodic Properties and a 0–1 Law

A natural starting point for the discussion of ergodic properties of the RWRE (X_n) would have been an analogue of Theorem 1. Unfortunately, obtaining such a statement has been a major challenge since the early 1980's, and is still open. To explain the challenge, we need to digress and introduce a certain conjectured 0–1 law.

Fix $\ell \in S^{d-1}$, i. e. ℓ is a unit vector in $\mathbb{R}^d$. Define the events

$$A_\ell^+ = \left\{ \lim_{n\to\infty} X_n \cdot \ell = \infty \right\} ,$$
$$A_\ell^- = \left\{ \lim_{n\to\infty} X_n \cdot \ell = -\infty \right\} .$$

The following proposition is due to Kalikow [51].

Proposition 6 *Assume P is i.i.d. and elliptic, i. e. $P(\omega(0,e) > 0) = 1$ for all e with $|e| = 1$, and that $\ell \in S^{d-1}$. Then, $\mathbb{P}(A_\ell^+ \cup A_\ell^-) \in \{0,1\}$.*

Note that for $d = 1$, Theorem 1 implies that $\mathbb{P}(A_\ell^+) \in \{0,1\}$. If one ever hopes to obtain a LLN, then one should be able to prove the following.

Conjecture 7 (Kalikow) *Assume P is i.i.d. and uniformly elliptic, and that $\ell \in S^{d-1}$. Then, $\mathbb{P}(A_\ell^+) \in \{0,1\}$.*

Efforts to prove Conjecture 7 are ongoing. The following summarizes its status at the current time, and combines results from [19,111].

Theorem 8

(a) Conjecture 7 holds for $d = 1, 2$ and elliptic i.i.d. environments.

(b) There exist ergodic environments that are elliptic (for $d = 2$) and even uniformly elliptic and mixing (for $d \geq 3$), for which a deterministic direction $\ell \in S^{d-1}$ exists such that $P_\omega^0(A_\ell^+) \in (0,1)$, for P-almost every ω.

As mentioned above, part (a) of Theorem 8 for $d = 1$ is a direct consequence of the LLN, Theorem 1.

As it turns out, the validity of Conjecture 7 is the only obstruction to a LLN, as demonstrated in the following theorem, which combines results from [4,102,106,110].

Theorem 9 *Assume P is i.i.d. and uniformly elliptic.*

(a) Fix $\ell \in S^{d-1}$. Then,

$$\lim_{n\to\infty} \frac{X_n \cdot \ell}{n} = v_+ \mathbf{1}_{A_\ell^+} + v_- \mathbf{1}_{A_\ell^-} , \mathbb{P} - a.s. . \quad (5)$$

In particular, when $d = 2$ the LLN holds true.

(b) $\mathbb{P}$-almost surely, there are at most two possible limit points, denoted v_1, v_2, for the sequence X_n/n. Further, v_1, v_2 are deterministic, and if $v_1 \neq v_2$ then there exists a constant $a \geq 0$ such that $v_2 = -av_1$.

(c) When $d \geq 5$, if $v_1 \neq v_2$ then at least one of v_1 and v_2 equals 0.

Part (a) of the theorem implies that if Conjecture 7 is true, then the LLN holds for P i.i.d. and uniformly elliptic.

The proof of Theorem 9, and of many of the other results in this section, uses the machinery of regeneration times, introduced in [102]. Roughly, a random time k is a regeneration time relative to a direction $\ell \in S^{d-1}$ if $X_k \cdot \ell \geq X_n \cdot \ell$ for all $k \geq n$ but $X_k \cdot \ell < X_n \cdot \ell$ for all $k < n$ (i. e., $X_n \cdot \ell$ sets a record at time k, and never moves backward from that record). It turns out that the sequence of inter-regeneration times and inter-regeneration distances is an i.i.d. sequence under the annealed measure $\mathbb{P}$, if P is i.i.d. Once such an i.i.d. sequence has been identified, ergodic arguments yield the LLN, and the (annealed) CLT involves studying tail behavior of the regeneration times. In some ballistic cases, one may translate an annealed CLT to a quenched one, see [7] and [82]. We note in passing that regeneration techniques have also been useful for certain non-i.i.d. environments. We refer the interested reader to [26,27,79].

So far, there is no known criterion that allows one to decide the question of transience or recurrence for RWRE in dimension $d \geq 2$, although one certainly expect transience as soon as $d \geq 3$.

Ballistic Behavior and Sznitman's Conditions

Lacking an explicit expression for the speed of the RWRE for $d \geq 2$, a natural goal is to identify a large family of

models for which $X_n/n \to v \neq 0$. RWRE's that satisfy such a relation are called *ballistic*. As we saw in Theorem 1, when $d = 1$ and $X_n \to \infty$, and the environment is i.i.d., the RWRE is ballistic if and only if $E\rho_0 < 1$.

Define $d_0 := \sum[\omega(0, e_i) - \omega(0, -e_i)]e_i$ as the drift at the origin. If there exists a direction $\ell \in S^{d-1}$ such that $d_0 \cdot \ell > 0$ for P-a.e. environment, a simple martingale argument shows that $X_n/n \to v$ with $v \cdot \ell > 0$. Following Zerner [109], we call such environments *non-nestling*. We will be mainly interested in *nestling* environments, that is environments in which the origin belongs to the closed convex hull of the support of d_0. (The source for the name lies in the fact that when the walk is nestling, it is possible to construct localized regions, called traps, to which the walk return many times, leading to the mental picture of a bird that keeps returning to a nest. We note that as we show in Remark 18 below, one should not confuse the condition $Ed_0 \neq 0$ with ballistic behavior, as it does not guarantee a limiting non-zero speed.)

Traps tend to slow down the particle. However, unlike $d = 1$, all attempts to build explicit traps that slow down the particle to a sub-diffusive scale quickly fail. One thus suspects that a good control of trapping properties relates to an analysis of the RWRE. With this motivation in mind, Sznitman introduced conditions on the environment that eventually lead to a good understanding of the ballistic regime. Fix a direction $\ell \in S^{d-1}$, and for $b > 0$, define the region $U_{\ell,b,L} = \{x \in \mathbb{Z}^d : x \cdot \ell \in (-bL, L)\}$. Let $T_{\ell,b,L} = \min\{n > 0 : X_n \notin U_{\ell,b,L}\}$.

Definition 10 Let $\gamma \in (0, 1)$ be given. Then, P satisfies condition T_γ relative to ℓ if for all ℓ' in some neighborhood of ℓ, and all $b > 0$,

$$\limsup_{L\to\infty} \frac{1}{L^\gamma} \log \mathbb{P}\left(X_{T_{\ell,b,L}} \cdot \ell < 0\right) < 0 . \tag{6}$$

P satisfies condition T' relative to ℓ if it satisfies condition T_γ relative to ℓ for all $\gamma \in (0, 1)$. It satisfies condition T relative to ℓ if it satisfies condition T_1 relative to ℓ.

In words, condition T relative to ℓ holds if the exit from a slab that is contained between two hyperplanes perpendicular to ℓ, located respectively at distance $+L$ in the ℓ direction and $-bL$ in the opposite direction, occurs through the "backward" direction with probability that is exponentially small in L. Condition T' relaxes the exponential decay to "almost" exponential decay (there is an alternative description of condition T' in terms of regeneration distances, see Proposition 12 below). The power of condition T' is the following.

Theorem 11 (Sznitman) *Assume P is i.i.d. and uniformly elliptic, and that condition T' relative to some direction ℓ holds. Then, the process (X_n) is ballistic, i. e. $X_n/n \to v \neq 0$ for some deterministic v with $v \cdot \ell > 0$, and there is a deterministic $\sigma^2 > 0$ such that, under the annealed measure $\mathbb{P}$, $(X_n - nv)/\sigma\sqrt{n}$ converges in distribution to a standard Gaussian random variable.*

(The convergence in distribution in Theorem 11 actually extends to an invariance principle; The results in [7] yield a quenched CLT as soon as $d \geq 4$.) The key to the usefulness of Condition T' is in the following result from [99], where τ_1 denotes the first regeneration time.

Proposition 12 *Let $d \geq 2$, $\ell \in S^{d-1}$, and $\gamma \in (0, 1]$. The following are equivalent:*

(a) Condition T_γ holds.
(b) $\mathbb{P}(A_\ell^+) = 1$ and, with $X^* := \sup_{0\leq n\leq\tau_1} |X_n|$, there exists a $c > 0$ such that

$$\mathbb{E}\left(\exp(c(X^*)^\gamma)\right) < \infty .$$

The proof is detailed in [99], see also the exposition in [101]. From part (b) of Proposition 12, tail estimates on τ_1 follow. We omit further details.

It can be checked (by a martingale argument) that condition T (and hence T') holds for a certain direction ℓ when the environment is non-nestling. The first example of a nestling environment that satisfies condition T was provided by Sznitman, who showed that the class of environments satisfying the *Kalikow condition* from [51] also satisfy condition T (Kalikow himself had showed that his condition implies the 0–1 law, and it was shown in [102] to imply ballistic behavior). However, the verification of condition T' seems a-priori not obvious. It is thus extraordinary that an effective criterion for checking it exists, see [98]. This criterion is used in [99] to construct an example of a ballistic RWRE that does not satisfy Kalikow's condition but does satisfy T', relative to some ℓ.

Sznitman's Conjecture As noted in Proposition 12, condition T' is equivalent to certain exponential moments on the maximal distance from the origin the RWRE has achieved before time τ_1. In [99], Sznitman actually proves that condition T_γ relative to ℓ with any $\gamma \in (1/2, 1)$ implies condition T' relative to the same ℓ. This led him to the following conjecture, see [98]:

Conjecture 13 (Sznitman) *Assume P is uniformly elliptic and i.i.d. Then, condition T relative to ℓ is implied by condition T_γ relative to ℓ for any $\gamma \in (0, 1)$.*

It is also reasonable to expect ("plausible", in the language of [98]) that in addition, ballistic behavior with speed v implies condition T relative to $\ell = v/|v|$, for $d > 1$.

For $d = 1$, and i.i.d. environment, all the conditions T_γ with respect to the direction $\ell = 1$ are equivalent to $E \log \rho_0 < 0$, see [96]. Hence, Conjecture 13 holds when $d = 1$ (note that this is *not* the case for the conclusions concerning ballistic behavior, which do not hold true for $d = 1$).

In the ballistic situation, some information on the environment viewed from the point of view of the particle can be deduced. We refer to [12] and [79] for details.

Large Deviations, Quenched and Annealed

In dimension $d = 1$, the large deviations for the sequence X_n/n were obtained by considering hitting times. While this approach can be partially extended to obtain quenched LDP's for some RWRE's, see [109], its scope is limited, and in particular it does not apply to all i.i.d. measures P, nor to an annealed LDP.

A different approach was taken by Varadhan [106], who obtained the following.

Theorem 14 *Assume $d \geq 2$.*

(a) Assume P is a uniformly elliptic, ergodic measure. Then, for P-a.e. environment ω, the sequence of variables X_n/n under P^0_ω satisfies the (quenched) LDP (on $[-1, 1]^d$) with speed n and deterministic, convex rate function I.

(b) Assume further that P is i.i.d. Then, the sequence of random variables X_n/n satisfies, under $\mathbb{P}$, the (annealed) LDP with speed n and convex rate function $\mathcal{I}$.

(c) The rate functions I and $\mathcal{I}$ possess the same zero set. Further, this (convex) set is either a single point or a segment of a line.

An alternative description of the quenched rate function, that is more instructive than the sub-additivity argument, has been developed for the related model of diffusions in random environments in [54]. Part (b) of Theorem 14 was extended to certain mixing environments in [80].

As for dimension $d = 1$, both I and $\mathcal{I}$ are in general not strictly convex. The quenched statement is an application of the ergodic subadditive theorem [64]. The annealed LDP is obtained by noting that the process of histories of the walk is a Markov chain, and applying the general large deviations theory for such chains.

Remark 15 In the multi-dimensional case, a formula like (4), with its intuitive description of the way an annealed deviation is obtained, is not available, since the modification of big chunks of the environment has probability which decays exponentially in volume order, i. e. n^d, instead of n.

As for $d = 1$, it is natural to study slowdown estimates in the region where the rate functions vanish, and in particular to study the probability of slowdown. This study is closely related to the analysis of Condition T', and we refer to [96,97] for details.

Non-ballistic Results

The analysis of RWRE for environments that do not exhibit ballistic behavior is still limited. Still, two important classes of models have been identified, for which the analysis could be carried out. We sketch those below.

Balanced Environment Balanced environments satisfy the constraint $\omega(0, e_i) = \omega(0, -e_i)$ for all i, in which case the local drift vanishes everywhere. In that case, X_n itself is a martingale with bounded increments, and thus $X_n/n \to 0$, $\mathbb{P}$-a.s. In fact, an invariance principle also holds.

Theorem 16 *Assume P is stationary and ergodic, balanced, and uniformly elliptic Then $X_n/n \to 0$, $\mathbb{P}$-a.s., and there exists a deterministic $\sigma^2 > 0$ such that $X_n/\sigma\sqrt{n}$ converges in distribution (under the annealed measure $\mathbb{P}$) to a Gaussian random variables. Further, X_n is recurrent if $d = 2$ and transient if $d \geq 3$.*

Theorem 16 is essentially due to [59] (the recurrence statement is due to Kesten, and can be found in [107]). It is one of the few instances where "classical" homogenization can be applied to the study of multi-dimensional RWRE.

RWRE with Deterministic Components A key to the analysis of the ballistic case is the existence of certain regeneration times. Those were used to create an i.i.d. sequence under the measure $\mathbb{P}$.

In the non-ballistic case, regeneration times as defined above do not exist. However, if the dimension of the space is large enough and some of the components are deterministic, an alternative to regeneration times can be found, based on *cut times* for simple random walk. For a simple random walk $\{S_n\}$, a cut time k is such that $\{S_n, n \leq k\} \cap \{S_n, n > k\} = \emptyset$. Such cut times exist for $d \geq 4$ [40]. By considering the cut times induced by the components of the RWRE evolving in the deterministic directions, [14] proved the following.

Theorem 17 *Assume $d = d_1 + d_2$ with $d_1 \geq 5$. Assume P is a uniformly elliptic i.i.d. measure, with $\omega(x, x + e) = q(e)$ for $e = \pm e_i$, $i = 1, \ldots, d_1$ and a deterministic q. Then, there exists a deterministic constant v such that $X_n/n \to v$, $\mathbb{P}$-a.s.. Further, if $d_1 \geq 13$, then the quenched CLT holds, i. e. there exists a deterministic $\sigma^2 > 0$ such that,*

for P-almost every ω, $(X_n - nv)/\sigma\sqrt{n}$ converges in distribution, under P^0_ω to a standard Gaussian variable.

Remark 18

(a) The convergence in distribution in Theorem 17 extends to a full invariance principle.
(b) An amusing consequence of Theorem 17 is that, for $d > 5$, one may construct P i.i.d. and uniformly elliptic such that $E(d_0 \cdot \ell) < 0$ but the resulting RWRE is ballistic with $v \cdot \ell > 0$. Recall that this is impossible in dimension $d = 1$, see Remark 2b). Also, for $d > 6$, one may construct for every $\epsilon > 0$ a P i.i.d. and uniformly elliptic such that $|\omega(x, x+e) - 1/2d| < \epsilon$, $E(d_0) \neq 0$, but $X_n/n \to 0$, $\mathbb{P}$-a.s., or such that $E(d_0) = 0$ but the walk is ballistic. We refer to [14] for the construction.

Multi Dimensional RWRE – the Perturbative Regime

We discuss in this section the perturbative analysis of the RWRE. By P being a small perturbation from a kernel q we mean that $q(\pm e_i) \geq 0$, $\sum_i [q(e_i) + q(-e_i)] = 1$, and for some ϵ small, $|\omega(x, x+e) - q(e)| < \epsilon$ for $e \in \{\pm e_i\}$. When $q(e) = 1/2d$ for $e = \pm e_i$, we say that P is a small perturbation from simple random walk.

We already observed, see Remark 18, that in the perturbative regime for simple random walk, the RWRE can exhibit behavior which is very different from the behavior of simple random walk.

Ballistic Walks

Sznitman's criterion [99] for condition T' to hold together with an renormalization analysis allowed him to give in [99] sufficient conditions for ballistic behavior when ϵ is small. Set $\rho_0(3) = 5/2$ and $\rho_0(d) = 3$ for $d \geq 4$.

Theorem 17 *Let $d \geq 3$ and $\rho < \rho_0(d)$. Then there exists an $\epsilon_0 = \epsilon_0(d, \rho) > 0$ such that if P is i.i.d. and an ϵ perturbation from simple random walk, and $Ed_0 \cdot e_1 > \epsilon^\rho$, then the T' condition relative to e_1 holds.*

Contrasting Theorem 19 with the examples in Remark 18 shows that some condition on the strength of the averaged drift Ed_0 as function of ϵ is necessary for ballistic behavior. Also, $\rho_0(d) > 2$ is used in constructing the examples in [99] that show that Kalikow's condition is strictly included in condition T'. We note that the case $d = 2$ is still open.

In another direction, if one writes $\omega(x, x+e) = q(e) + \epsilon\xi(x, x+e)$ with ξ i.i.d., and either $\sum eq(e) \neq 0$ or $\sum eq(e) = 0$ but $\sum eE\xi(0, e) \neq 0$, then for ϵ small enough, Kalikow's condition holds. Expansions in ϵ of the speed of the RWRE are provided in [83].

Balanced Walks

Recall the balanced walks introduced in Subsect. "Non-Ballistic Results", c.f. Theorem 16. The existence of an invariance measure viewed from the point of view of the particle, and the control achieved on this measure by approximations with periodized environments, allow one to get an expansion of the diffusivity matrix in terms of the strength of the perturbation from simple random walk. We refer the reader to [60] for details.

Isotropic RWRE

The existence of sub-diffusive behavior for the RWRE model in $d = 1$ immediately raises the question as to whether such sub-diffusive behavior is present in higher dimension. As implied by Theorem 11, this is not the case when the environment satisfies condition T'. Since it may be expected that condition T' characterizes ballistic behavior for $d > 1$, it is reasonable to expect (but not proved!) that for P i.i.d. and uniformly elliptic, and $d > 1$, no sub-diffusive behavior is possible when the walk is transient in direction ℓ (and further, in the ballistic regime, when recentering around the limiting velocity v, one expects fluctuations in the diffusive scale).

Outside the ballistic regime, rigorous results are few. Early attempts to address the question of existence of a diffusive regime appeared in [35,41], using a formal renormalization group analysis in the small perturbation regime, with the conclusion that no sub-diffusive behavior exists at $d \geq 3$ in the perturbative regime, and that at most logarithmic corrections to diffusive behavior exist at $d = 2$. While this conclusion certainly conforms with what one would expect, soon after it was pointed out that counter-examples can be constructed (albeit not with i.i.d., or even finite range dependent, environments), see [15,17,18]. Further, some of the examples discussed in this article, and in particular those of Subsect. "Non-Ballistic Results", see Remark 18, do not seem to be consistent with the formal renormalization analysis.

An attempt to put the analysis on a rigorous foundation was made in [22]. Among other things, they introduced the following isotropy condition:

Definition 20 The law P on the environment is isotropic if, for any rotation matrix $\mathcal{O}$ acting on $\mathbb{R}^d$ that fixes $\mathbb{Z}^d$, the laws of $(\omega(0, \mathcal{O}e))_{e:|e|=1}$ and $(\omega(0, e))_{e:|e|=1}$ coincide.

In particular, if P is isotropic then $Ed_0 = 0$. The main result of [22] is the following:

Theorem 21 (Bricmont–Kupiainen) *Assume $d \geq 3$. There exists an $\epsilon_0 = \epsilon_0(d)$ such that if P is i.i.d. and*

isotropic, and an ϵ perturbation of simple random walk with $\epsilon < \epsilon_0$, then for some deterministic $\sigma^2 > 0$ and for P almost every ω, the sequence $X_n/\sigma\sqrt{n}$ converges in distribution, under P^0_ω, to a standard Gaussian random variable.

The approach of [22] is to introduce a (diffusive) rescaling in time and space, and propagate an estimate on both the large scale behavior of the RWRE, as well as about the existence of local traps that have the potential to destroy, at the next level, the diffusivity properties. The restriction to $d \geq 3$ is useful because the underlying simple random walk for $d \geq 3$ is transient, and hence Green function computations can be performed.

Several attempts have recently been made to provide an rescaling argument (alternative to [22]) that is more transparent. The first approach [103], which is closest to Theorem 21, has been in the context of diffusions in random environments. In the remainder of this section, we describe another approach, due to [13], that yields a result concerning the exit measure of (isotropic) RWRE from large balls.

Let $V_L = \{x \in \mathbb{Z}^d : |x| \leq L\}$ be the ball of radius L in $\mathbb{Z}^d$ (where we recall that $|\cdot|$ is the euclidean norm), and let $\partial V_L = \{y \in \mathbb{Z}^d : d(y, V_L) = 1\}$ denote the boundary of V_L. Let $\tau_L = \min\{n : X_n \notin V_L\}$ denote the exit time of the RWRE from V_L, and for $x \in V_L, z \in \partial V_L$, let $\Pi_L(x,z) = P^x_\omega(X_{\tau_L} = z)$ denote the exit measure of the RWRE from V_L, and let $\pi_L(x,z)$ denote the corresponding quantity for simple random walk. Finally, let $\Pi^s_{L,l}(x,z) = \Pi_L \star \pi_{\eta l}$, where $\star$ denotes convolution and η is a random variable with smooth density supported on $(1,2)$. $\Pi^s_{L,l}$ is a smoothed version of Π_L, where the smoothing is at scale l.

One expects that for an isotropic environment that is a small perturbation of simple random walk, the exit measure Π_L approaches that of simple random walk, except for small nonvanishing correction that are due to localized perturbations near the boundary, and that as soon as some additional smoothing is applied, convergence occurs. Under the assumptions of Theorem 21, this is indeed the case. In what follows, for probability measures μ, ν we write $\|\mu - \nu\|$ for the variational distance between μ and ν.

Theorem 22 *Assume $d \geq 3$. There exists a $\delta_0 = \delta_0(d) > 0$ with the following property: for each $\delta < \delta_0$ there exists an $\epsilon_0 = \epsilon_0(d,\delta)$ such that if $\epsilon < \epsilon_0$ and P is an i.i.d. and isotropic law which is an ϵ perturbation of simple random walk, then*

$$\limsup_{L\to\infty} \|\Pi_L(0,\cdot) - \pi_L(0,\cdot)\| \leq \delta\,. \tag{7}$$

Further,

$$\limsup_{L\to\infty} \|\Pi^s_{L,l}(0,\cdot) - \pi_L \star \pi_{\eta l}(0,\cdot)\| \leq c_l \to_{l\to\infty} 0\,. \tag{8}$$

Diffusions in Random Environments

The model of RWRE possesses a natural analogue in the setup of diffusion processes.

One Dimensional Generators

For dimension $d = 1$, the study of analogues of the RWRE model goes back to [23] and [86]. Formally, one looks at solutions to the stochastic differential equation

$$\mathrm{d}X_t = -\frac{1}{2}V'(X_t)\mathrm{d}t + \mathrm{d}\beta_t\,, \quad X_0 = 0\,, \tag{9}$$

where β is a standard Brownian motion and V, the potential, is itself an (independent of β) Brownian motion with constant drift. Of course, (9) does not make sense as written, but one can express the solution to (9) for smooth V in a way that makes sense also when V is replaced by Brownian motion, by saying that conditioned on the environment V, X_t is a diffusion with generator

$$\frac{1}{2}\mathrm{e}^{V(x)}\frac{\mathrm{d}}{\mathrm{d}x}\left(\mathrm{e}^{-V(x)}\frac{\mathrm{d}}{\mathrm{d}x}\right)\,. \tag{10}$$

The diffusion in (9) inherits many of the asymptotic properties of the RWRE model. Additional tools, borrowed from stochastic calculus, are often needed to obtain sharp statements. We refer to [89] for details and additional references.

Multi Dimensional Diffusions: Finite Range Dependence

Like the RWRE in dimension $d = 1$, the model (9) leads to a reversible diffusion. A direct generalization of (9) via the expression (10) for the generator, see for example [68,69], preserves the reversibility of the process, and thus for our purpose does not serve as a true analogue of the RWRE model. Instead, we consider diffusions satisfying the equation in $\mathbb{R}^d$:

$$\mathrm{d}X_t = b(X_t,\omega)\mathrm{d}t + \sigma(X_t,\omega)\mathrm{d}W_t\,, \quad X_0 = 0\,, \tag{11}$$

with generator

$$L = \frac{1}{2}\sum_{i,j=1}^d a_{ij}(x,\omega)\,\partial^2_{ij} + \sum_{i=1}^d b_i(x,\omega)\,\partial_i\,, \tag{12}$$

where $a = \sigma\sigma^T$ is a d-by-d matrix and the coefficients a, b are assumed to satisfy the following:

Assumption 23

(a) The functions $a(\cdot,\omega)$ and $b(\cdot,\omega)$ are uniformly (in ω) bounded by K, with Lipschitz norm bounded by K, and a is uniformly elliptic, i. e. $a(x,\omega)-\kappa I$ is positive definite for some $\kappa>0$ independent of x or ω.
(b) The random field $\big(a(x,\omega),b(x,\omega)\big)_{x\in\mathbb{R}^d}$ is stationary with respect to shifts in $\mathbb{R}^d$.
(c) The collection of random variables $\big(a(x,\cdot),b(x,\cdot)\big)_{x\in A}$ and $\big(a(y,\cdot),b(y,\cdot)\big)_{y\in B}$ are independent when $d(A,B)>R$.

Part (a) of Assumption 23 ensures that (11) possesses a unique strong solution. Part (c) of Assumption 23 is a "finite range dependence" condition. We continue to write P_ω for the quenched law of the trajectories of the diffusion.

Many of the results described in Subsect. "Ergodic Properties and a 0–1 Law" and Subsect. "Ballistic Behavior and Sznitman's Conditions" have been proved also in the context of diffusions, when Assumption 23 holds. We refer to [45,84,85,88] for details.

Isotropic Diffusions in the Perturbative Regime

The analogue of the isotropy condition Assumption 23 (b) in the diffusion context is the following:

Assumption 24 (Isotropy) *For any rotation matrix $\mathcal{O}$ preserving the union of coordinate axes of $\mathbb{R}^d$,*

$$\big(a(\mathcal{O}x,\omega),b(\mathcal{O}x,\omega)\big)_{x\in\mathbb{R}^d} \quad \text{has same law under } P \text{ as} \quad \big(\mathcal{O}a(x,\omega)\mathcal{O}^T,\mathcal{O}b(x,\omega)\big)_{x\in\mathbb{R}^d}\,.$$

The analogue of Theorem 21 is the following theorem, due to [103]. Its proof again uses multi-scale arguments, and is based on controlling the (scaled) Hölder norm of the operator associated with the transition probability of the diffusion.

Theorem 25 *Let Assumptions 23 and 24 hold. Then, there exists a constant $\epsilon_0=\epsilon_0(d,K,R)$ such that if $|a(x,\omega)-I|\le\epsilon_0$ and $|b(x,\omega)|\le\epsilon_0$, for all $x\in\mathbb{R}^d,\omega\in\Omega$, then for some deterministic $\sigma^2>0$, for a.e. ω, the sequence of random variables $X_t/\sigma\sqrt{t}$ converges in distribution to a standard Gaussian random variable.*

(A full quenched invariance principle also holds under the assumptions of Theorem 25.)

Topics Left Out and Future Directions

As pointed out in the text, many open problems remain in the study of RWRE's and motion in random media, with the most important one being the general validity of the law of large numbers, and the existence of diffusive limits for dimensions $d\ge 2$. We briefly mention in the rest of this section several topics that are related to this review but that we have not covered in details.

Random Conductance Model

We have concentrated in this review on RWRE's in i.i.d. environments, which give rise in the multi-dimensional case to non-reversible Markov processes. Although mentioned in several places, we did not discuss in details the reversible case, where homogenization techniques using the environment viewed from the point of view of the particle are very efficient (note that the reversible case is a very particular case of an environment which is not i.i.d. but rather dependent with finite range dependence). The prototype for such reversible models is the "random conductance model", where each edge (x,y) of $\mathbb{Z}^d$ is associated a (random, i.i.d.) conductance $C_{x,y}$, and the transition probability between x and y is $C_{x,y}/(\sum_{z:\,|z-x|=1}C_{x,z})$. Annealed CLT's for the random conductance model are provided in [57,66]. See also [2] for a related model with symmetric transitions. The quenched CLT is obtained in [16] and [90].

One of the motivations to consider the random conductance model is the analysis of random walk on supercritical percolation clusters. The annealed CLT is covered by [66]. Several recent papers discuss the quenched case, first in dimension $d\ge 4$ [90], and then in all dimensions, see [5,70]. In another direction, when one discusses *biased* walks on a percolation cluster, new phenomena occur, for example the lack of monotonicity of the speed of the walk in the strength of the bias, which is again a manifestation of the trapping phenomenon. We refer to [6] and [100] for details.

Brownian Motion in a Field of Random Obstacles

Another closely related (reversible) model is the model of Brownian motion in a field of obstacles in $\mathbb{R}^d$. Here, one defines a potential $V(x,\omega)=\sum_i W(x-x_i)$ where the collection $\{x_i\}$ is a (random) configuration of points in $\mathbb{R}^d$ (usually, taken according to a Poisson law) and W is a fixed nonnegative shape function. Of interest are the properties of Brownian motion $(X_t)_{t\in[0,T]}$, perturbed by the change of measure

$$\Lambda_T=\frac{1}{Z_T(\omega)}\exp\left(-\int_0^T V(X_s,\omega)\mathrm{d}s\right)\,.$$

It is common to distinguish between "soft traps", with W bounded and typically of compact support, and "hard trap", where $W = \infty \mathbf{1}_C$ where C is a given compact set. One is interested in understanding various path properties, as T gets large, or in understanding the quenched partition function $Z_T(\omega)$ and its annealed counterpart EZ_T. Due to reversibility, the problem is closely related to the study of the bottom λ_ω of the spectrum of $-\Delta/2 + V$, and the difficulty is in understanding the structure of those traps that influence λ_ω. A good overview of the model and the techniques developed to analyze it, including the "method of enlargement of obstacles", can be found in [95].

Time Dependent RWRE

An interesting variant of the RWRE model has been proposed in [8]. In this model, the random environment is dynamic, i. e. changes with time, and so we write $\omega(x, x+e, n)$ where we wrote before $\omega(x, x+e)$. In the simplest version, the collection of random vectors $(\omega(x, x+\cdot, n))_{x\in\mathbb{Z}^d, n\in\mathbb{N}}$ is i.i.d. Annealed, the RWRE is then a simple random walk in an averaged environment, but the true interest lies in obtaining quenched statements. Those were obtained in [8,10] by a perturbative approach. An alternative, simpler proof is given by [94]. Another approach to the quenched CLT, that covers other cases of random walk "with a forbidden direction", is developed in [81], based on a general pointwise CLT for additive functionals of Markov chains due to [34].

An interpolation between the RWRE model and the i.i.d. dynamical environment model is when the collection $(\omega(x, x+\cdot, n))_{x\in\mathbb{Z}^d, n\in\mathbb{N}}$ is i.i.d. in x but Markovian in n. This case has been analyzed by perturbative methods in [9], and by regeneration techniques in [3]. In both cases, an annealed CLT holds in any dimension, but the quenched CLT was obtained only in high dimension. Recently, a dynamical approach was developed in [36], that proves the quenched CLT in all dimensions, subject to fast enough mixing of the Markov chain. It is still open to determine whether in the Markovian setup, there are uniformly elliptic, exponentially mixing examples where the quenched CLT fails.

RWRE on Trees and Other Graphs

We have already mentioned the interest in considering random walks on random subgraphs of $\mathbb{Z}^d$, and in particular percolation clusters. Of course, one may consider instead random walk (or biased random walk) on other random graphs. A particularly important class of models treats random walks on random trees, and in particular Galton–Watson trees. We refer to [65] for an excellent overview of the properties and ergodic theory of such random walks, and to [74] for recent results concerning the CLT. See also [50] for slowdown estimates for the analog of the RWRE on the binary tree. We emphasize that these models are all reversible.

Non Nearest Neighbor RWRE

Many of the techniques described in this survey have a natural generalization to non nearest neighbor walks. In particular, the results in [80,106] are already stated in terms of compactly supported transition probabilities, and the development of regeneration times can easily be extended, following the techniques in [26,27], to the non-nearest neighbor, finite range setup. However, to the best of my knowledge, no systematic study of RWRE for non-nearest neighbor RWRE's in dimension $d \geq 2$ has appeared in the literature.

The situation is different in dimension $d = 1$, where the RWRE is not reversible anymore. It was early realized, see [53,63], that ergodic theorems involve the study of certain Lyapounov exponents associated with the product of random matrices. For some recent results, we refer to [11] and [21].

Bibliography

Primary Literature

1. Alili S (1999) Asymptotic behaviour for random walks in random environments. J Appl Probab 36:334–349
2. Anshelevich VV, Khanin KM, Sinai YG (1982) Symmetric random walks in random environments. Commun Math Phys 85:449–470
3. Bandyopadhyay A, Zeitouni O (2006) Random Walk in Dynamic Markovian Random Environment. ALEA 1:205–224
4. Berger N (2006) On the limiting velocity of high-dimensional random walk in random environment. Arxiv: math.PR/0601656
5. Berger N, Biskup M (2007) Quenched invariance principles for simple random walk on percolation clusters. Probab Theory Relat Fields 137:83–120
6. Berger N, Gantert N, Peres Y (2003) The speed of biased random walk on percolation clusters. Probab Theory Relat Fields 126:221–242
7. Berger N, Zeitouni O (2008) A quenched invariance principle for certain ballistic random walks in i.i.d. environments. In: Siduravicius V, Vares ME (eds) In and out of equilibrium, Progress in probability, vol. pp 137–160
8. Boldrighini C, Minlos RA, Pellegrinotti A (1997) Almost-sure central limit theorem for a Markov model of random walk in dynamical random environment. Probab Theory Relat Fields 109:245–273
9. Boldrighini C, Minlos RA, Pellegrinotti A (2000) Random walk in a fluctuating random environment with Markov evolution.

In: On Dobrushin's way. From probability theory to statistical physics. Amer Math Soc Transl Ser 198(2):13–35; Amer Math Soc, Providence
10. Boldrighini C, Minlos RA, Pellegrinotti A (2004) Random walks in quenched i.i.d. space-time random environment are always a.s. diffusive. Probab Theory Relat Fields 129:133–156
11. Bolthausen E, Goldsheid I (2000) Recurrence and transience of random walks in random environments on a strip. Commun Math Phys 214:429–447
12. Bolthausen E, Sznitman AS (2002) On the static and dynamic points of view for certain random walks in random environment. Method Appl Analysis 9:345–375
13. Bolthausen E, Zeitouni O (2007) Multiscale analysis of exit distributions for random walks in random environments. Prob Theor Fields 138:581–645
14. Bolthausen E, Sznitman AS, Zeitouni O (2003) Cut points and diffusive random walks in random environments. Ann Inst H Poincare 39:527–555
15. Bouchaud JP, Georges A, Le Doussal P (1987) Anomalous diffusion in random media: trapping, correlations and central limit theorems. J Phys 48:1855–1860
16. Boivin D, Depauw J (2003) Spectral homogenization of reversible random walks on $\mathbb{Z}^d$ in a random environment. Stoch Proc App 104:29–56
17. Bramson M (1991) Random walk in random environment: a counterexample without potential. J Stat Phys 62:863–875
18. Bramson M, Durrett R (1988) Random walk in random environment: a counterexample? Commun Math Phys 119:199–211
19. Bramson M, Zeitouni O, Zerner MPW (2006) Shortest spanning trees and a counterexample for random walks in random environments. Ann Probab 34:821–856
20. Brémont J (2004) Behavior of random walks on $\mathbb{Z}$ in Gibbsian medium. C R Math Acad Sci Paris 338:895–898
21. Brémont J (2004) Random walks in random medium on $\mathbb{Z}$ and Lyapunov spectrum. Ann Inst H Poincar Probab Stat 40:309–336
22. Bricmont J, Kupiainen A (1991) Random walks in asymmetric random environments. Com mun Math Phys 142:345–420
23. Brox T (1986) A one-dimensional diffusion process in a Wiener medium. Ann Probab 14:1206–1218
24. Cheliotis D (2005) Diffusions in random environments and the renewal theorem. Ann Probab 33:1760–1781
25. Comets F, Gantert N, Zeitouni O (2000) Quenched, annealed and functional large deviations for one dimensional random walk in random environment. Probab Theory Relat Fields 118:65–114
26. Comets F, Zeitouni O (2004) A law of large numbers for random walks in random mixing environments. Ann Probab 32:880–914
27. Comets F, Zeitouni O (2005) Gaussian fluctuations for random walks in random mixing environments. Isr J Math 148:87–114
28. Dembo A, Gantert N, Peres Y, Shi Z (2007) Valleys and the maximum local time for random walk in random environment. Probab Theory Relat Fields 137:443–473
29. Dembo A, Zeitouni O (1998) Large deviations techniques and applications. 2nd edn. Springer, New York
30. Dembo A, Guionnet A, Zeitouni O (2001) Aging properties of Sinai's random walk in random environment. Arxiv:math.PR/0105215
31. Dembo A, Gantert N, Peres Y, Zeitouni O (2002) Large deviations for random walks on Galton–Watson trees: averaging and uncertainty. Probab Theory Relat Fields 122:241–288
32. Dembo A, Gantert N, Zeitouni O (2004) Large deviations for random walk in random environment with holding times. Ann Probab 32:996–1029
33. Dembo A, Peres Y, Zeitouni O (1996) Tail estimates for one-dimensional random walk in random environment. Commun Math Physics 181:667–684
34. Derriennic Y, Lin M (2003) The central limit theorem for Markov chains started at a point. Probab Theory Relat Fields 125:73–76
35. Derrida B, Luck JM (1983) Diffusion on a random lattice: weak-disorder expansion in arbitrary dimension. Phys Rev B 28:7183–7190
36. Dolgopyat D, Keller G, Liverani C (2007) Random walk in markovian environment. ArXiv:math/0702100v1 [math.PR] (preprint)
37. Donsker MD, Varadhan SRS (1983) Asymptotic evaluation of certain Markov process expectations for large time, IV. Commun Pure Appl Math 36:183–212
38. Doyle PG, Snell JL (1984) Random walks and electric networks, Carus Mathematical Monographs, 22. Mathematical Association of America, Washington
39. Enriquez N, Sabot C, Zindy O (2007) Limit laws for transient random walks in random environments on $\mathbb{Z}$. ArXiv:math/0703660v1 [math.PR] (preprint)
40. Erdös P, Taylor SJ (1960) Some intersection properties of random walks paths. Acta Math Acad Sci Hungar 11:231–248
41. Fisher DS (1984) Random walks in random environments. Phys Rev A 30:60–964
42. Gantert N (2002) Subexponential tail asymptotics for a random walk with randomly placed one-way nodes. Ann Inst H Poincaré – Probab Statist 38:1–16
43. Gantert N, Shi Z (2002) Many visits to a single site by a transient random walk in random environment. Stoch Process Appl 99:159–176
44. Gantert N, Zeitouni O (1998) Quenched sub-exponential tail estimates for one-dimensional random walk in random environment. Commun Math Physics 194:177–190
45. Goergen L (2006) Limit velocity and zero-one laws for diffusions in random environment. Ann Appl Probab 16:1086–1123
46. Goldsheid I (2007) Simple transient random walks in one-dimensional random environment: the central limit theorem. Probab Theory Related Fields 139:41–64
47. Golosov AO (1985) On limiting distributions for a random walk in a critical one dimensional random environment. Commun Moscow Math Soc 199:199–200
48. Greven A, den Hollander F (1994) Large deviations for a random walk in random environment. Ann Probab 22:1381–1428
49. Hu Y, Shi Z (2000) The problem of the most visited site in random environment. Probab Theory Relat Fields 116:273–302
50. Hu Y, Shi Z (2007) A subdiffusive behaviour of recurrent random walk in random environment on a regular tree. Probab Theory Related Fields 138:521–549
51. Kalikow SA (1981) Generalized random walks in random environment. Ann Probab 9:753–768
52. Kesten H (1986) The limit distribution of Sinai's random walk in random environment. Physica A 138:299–309

53. Key ES (1984) Recurrence and transience criteria for random walk in a random environment. Ann Probab 12:529–560
54. Kosygina E, Rezakhanlou F, Varadhani RS (2006) Stochastic homogenization of Hamilton-Jacobi-Bellman Equations. Comm Pure Appl Math 59:1489–1521
55. Kesten H, Kozlov MV, Spitzer F (1975) A limit law for random walk in a random environment. Comput Math 30:145–168
56. Kozlov SM (1985) The method of averaging and walks in inhomogeneous environments. Russian Math Surv 40:73–145
57. Kunnemann R (1983) The diffusion limit of reversible jump processes in $\mathbb{Z}^d$ with ergodic random bond conductivities. Commun Math Phys 90:27–68
58. Kuo HJ, Trudinger NS (1990) Linear elliptic difference inequalities with random coefficients. Math Comput 55:37–53
59. Lawler GF (1982) Weak convergence of a random walk in a random environment. Commun Math Phys 87:81–87
60. Lawler GF (1989) Low-density expansion for a two-state random walk in a random environment. J Math Phys 30:145–157
61. Lawler GF (1991) Intersections of Random walks. Birkhauser, Basel
62. Le Doussal P, Monthus C, Fisher D (1999) Random walkers in one-dimensional random environment: exact renormalization group analysis. Phys Rev E 59:4795–4840
63. Ledrappier F (1984) Quelques propriétés des exposants charactéristiques, Lecture Notes in Mathematics, vol 1097. Springer, New York
64. Liggett TM (1985) An improved subadditive ergodic theorem. Ann Probab 13:1279–1285
65. Lyons R, Peres Y Probability on trees and networks. http://mypage.iu.edu/~rdlyons/prbtree/prbtree.html
66. De Masi A, Ferrari PA, Goldstein S, Wick WD (1989) An invariance principle for reversible Markov processes. Applications to random motions in random environments. J Stat Phys 55:787–855
67. Mayer-Wolf E, Roitershtein A, Zeitouni O (2004) Limit theorems for one-dimensional transient random walks in Markov environments. Ann Inst H Poincaré Probab Stat 40:635–659
68. Mathieu P (1994) Zero white noise limit through Dirichlet forms, with applications to diffusions in a random medium. Probab Theory Relat Fields 99:549–580
69. Mathieu P (1995) Limit theorems for diffusions with a random potential. Stoch Proc App 60:103–111
70. Mathieu P, Piatnitski A (2007) Quenched invariance principles for random walks on percolation clusters. Proc R Soc Lond Ser A Math Phys Eng Sci 463:2287–2307
71. Molchanov SA (1994) Lectures on random media. Lecture Notes in Mathematics, vol 1581. Springer, New York
72. Papanicolaou GC, Varadhan SRS (1982) Diffusions with random coefficients. In: Kallianpur G, Krishnaiah PR, Ghosh JK (eds) Statistics and probability: essays in honor of C. R. Rao. North-Holland, Amsterdam, pp 547–552
73. Pemantle R (1988) Phase transition in reinforced random walk and RWRE on trees. Ann Probab 16:1229–1241
74. Peres Y, Zeitouni O (2008) A Quenched CLT for biased random walks on Golton–Watson trees, Prob Theory Rel Fieldsx 140:595–629
75. Peterson J (2008) Ph.D. thesis. University of Minnesota
76. Peterson J, Zeitouni O (2007) Quenched limits for transient, zero speed one-dimensional random walk in random environment. ArXiv:0704.1778v1 [math.PR] (preprint)
77. Pisztora A, Povel T (1999) Large deviation principle for random walk in a quenched random environment in the low speed regime. Ann Probab 27:1389–1413
78. Pisztora A, Povel T, Zeitouni O (1999) Precise large deviation estimates for a one-dimensional random walk in a random environment. Probab Theory Relat Fields 113:191–219
79. Rassoul-Agha F (2003) The point of view of the particle on the law of large numbers for random walks in a mixing random environment. Ann Probab 31:1441–1463
80. Rassoul-Agha F (2004) Large deviations for random walks in a mixing random environment and other (non-Markov) random walks. Commun Pure Appl Math 57:1178–1196
81. Rassoul-Agha F, Seppäläinen T (2005) An almost sure invariance principle for random walks in a space-time random environment. Probab Theory Relat Fields 133:299–314
82. Rassoul-Agha F, Seppäläinen T (2007) Almost sure functional central limit theorem for non-nestling random walk in random environment. ArXiv:0704.1022v1 [math.PR] (preprint)
83. Sabot C (2004) Random walks in random environment at low disorder. Ann Probab 32:2996–3023
84. Schmitz T (2006) Diffusions in random environment and ballistic behavior. Ann Inst H Poincaré Prob Statist 42:683–714
85. Schmitz T (2006) Examples of condition (T) for diffusions in random environment. Electron J Probab 11:540–562
86. Schumacher S (1985) Diffusions with random coefficients. Contemp Math 41:351–356
87. Shen L (2002) Asymptotic properties of certain anisotropic walks in random media. Ann Appl Probab 12:477–510
88. Shen L (2003) On ballistic diffusions in random environment. Ann Inst H Poincare Probab Statist 39:839–876
89. Shi Z (2001) Sinai's walk via stochastic calculus. In: Comets F, Pardoux E (eds) Milieux Aléatoires. Panoramas et Synthèses 12. Soc Math Fr, pp 53–74
90. Sidoravicius V, Sznitman AS (2004) Quenched invariance principles for walks on clusters of percolation or among random conductances. Probab Theory Relat Fields 129:219–244
91. Sinai YG (1982) The limiting behavior of a one-dimensional random walk in random environment. Theor Prob Appl 27:256–268
92. Solomon F (1975) Random walks in random environments. Ann Probab 3:1–31
93. Stroock D, Varadhan SRS (1979) Multidimensional diffusion processes. Springer, New York
94. Stannat W (2004) A remark on the CLT for a random walk in a random environment. Probab Theory Relat Fields 130:377–387
95. Sznitman AS (1998) Brownian motion, obstacles and random media. Springer, New York
96. Sznitman AS (1999) On a class of transient random walks in random environment. Ann Probab 29:724–765
97. Sznitman AS (2000) Slowdown estimates and central limit theorem for random walks in random environment. JEMS 2:93–143
98. Sznitman AS (2002) An effective criterion for ballistic behavior of random walks in random environment. Probab Theory Relat Fields 122:509–544
99. Sznitman AS (2003) On new examples of ballistic random walks in random environment. Ann Probab 31:285–322
100. Sznitman AS (2003) On the anisotropic walk on the supercritical percolation cluster. Commun Math Phys 240.123–148

101. Sznitman AS (2004) Topics in random walks in random environment. School and Conference on Probability Theory. ICTP Lect Notes XVII Abdus Salam, Int Cent Theor Phys Trieste, pp 203–266
102. Sznitman AS, Zerner M (1999) A law of large numbers for random walks in random environment. Ann Probab 27:1851–1869
103. Sznitman AS, Zeitouni O (2006) An invariance principle for isotropic diffusions in random environments. Invent Math 164:455–567
104. Temkin DE (1972) One dimensional random walk in a two-component chain. Soviet Math Dokl 13:1172–1176
105. Varadhan SRS (1966) Asymptotic probabilities and differential equations. Commun Pure Appl Math 9:261–286
106. Varadhan SRS (2003) Large deviations for random walks in a random environment. Commun Pure Appl Math 56:1222–1245
107. Zeitouni O (2004) Random walks in random environment. XXXI Summer School in Probability, St. Flour, 2001. Lect Notes Math, vol 1837. Springer, New York, pp 193–312
108. Zeitouni O (2006) Random walks and diffusions in random environments. J Phys A: Math Gen 39:R433–R464
109. Zerner MPW (1998) Lyapounov exponents and quenched large deviations for multidimensional random walk in random environment. Ann Probab 26:1446–1476
110. Zerner MPW (2002) A non-ballistic law of large numbers for random walks in i.i.d. random environment. Electron Commun Probab 7(19):191–197
111. Zerner MPW, Merkl F (2001) A zero-one law for planar random walks in random environment. Ann Probab 29:1716–1732

Books and Reviews

Bolthausen E, Sznitman AS (2002) Ten lectures on random media DMV Seminar, 32. Birkhäuser, Basel
Hughes BD (1996) Random walks and random environments. Oxford University Press, Oxford
Révész P (2005) Random walk in random and non-random environments, 2nd edn. World Scientific, Hackensack
Revuz D, Yor M (1999) Continuous martingales and Brownian motion, 3rd edn. Springer, New York
Spitzer F (1976) Principles of random walk, 2nd edn. Springer, New York
Zeitouni O (2002) Random Walks in Random Environments. Proceedings of ICM 2002. Documenta Mathematica III:117–127

Rational, Goal-Oriented Agents

Rosaria Conte
LABSS-ISTC, CNR, Rome, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Agents
Concluding Remarks and Future Directions
Bibliography

Glossary

Agent A system that uses its power to bring about a given state in a physical, social, or mental environment.

Autonomy It is a property of a subset of agents, which decide to act on the grounds of their internal criteria (internal representations).

Computing agent An entity that executes the instructions contained in an algorithm running on a computer.

Goal-directed agents are intelligent agents that have an internal representation of the goals they achieve.

Goal-oriented agents are entities designed to achieve a certain state of the world wanted by either the agent itself, which in such a case is also a goal-directed system, or the user/designer.

Intelligent agents are goal-oriented agents using their knowledge to solve problems, including taking decisions and planning actions.

Internal representation Knowledge stored into internal agent structures or distributed over sets of interconnected internal units.

Knowledge Information about the environment, which is stored in agents' memory (archives).

Rational agents are intelligent agents that use their (limited) knowledge to maximize the difference between benefits and costs.

Definition of the Subject

In this article, goal-oriented agents, i. e. agents designed to achieve a given goal, will be treated as a highly general category within agent theory and application, and argued to subsume both rational agents, which maximize the distance between benefits and costs, and goal-directed agents, which have a representation of the goal they achieve. These sub-types of agents exemplify two different but non-exclusive approaches to agent theory, endogenous and exogenous. The endogenous approach is aimed at modeling agents in terms of their internal mechanisms of regulation. The exogenous approach is used to describe them from an external point of view, in terms of the effects they achieve in the world. In this article, these two approaches will be combined with another important classification, weak and strong agents. The resulting specific models and applications will be described. Finally, goal-directed and rational agents will be compared and future directions of research in the agent field will be discussed.

The field of agent theory and technology is growing fast. Suffice it to consider that the community of scientists interested in multi agent systems, which counted at most a few dozen people in the early nineties, is now one of the largest in the area of computational and technological sciences (The 2008 AAMAS (Autonomous Agents and Multi Agent Systems) conference received around 700 submissions.). Furthermore, a growing number of disciplines within the social and behavioral sciences, not to speak of the science of complexity, are turning to agent-based computational modeling in the aim to promote explicit, controllable models of social and behavioral phenomena, and to realize computational laboratories for visualizing such phenomena and conducting repeatable experiments about them.

Despite, or perhaps because of such a fast development, the field of agent theory and application did not receive an adequate scientific re-elaboration, nor commonly shared views and definitions. Agent is still a controversial notion, specified – as somebody observed in [81] – by a growing number of adjectives, for example, autonomous, goal-directed, rational, etc., rather than by explicit properties. Developed by different disciplines – physics, biology, AI, psychology, computer science, economy, decision theory, organization science, etc. – with their own conventions and vocabulary, the notion of agent would necessitate an explicit confrontation, which is underestimated by the different communities. The result is an unpleasant conceptual mix and consequent loss of scientific credibility of the field.

This article is therefore intended to offer a contribution of clarification concerning some properties of agency, in order to compare at least a subset of models and theories currently insufficiently interacting, and help envisage future trends.

Introduction

In the present contribution, rational agents are seen as a subset of intelligent agents (see Fig. 1), and these in turn as a subset of goal-oriented ones. Agents are said to be goal-oriented when they exhibit a finalistic behavior [3,9], i. e. when their behavior brings about a state of the world that achieves somebody's, not necessarily the user's, goals. To be noticed, goal-oriented systems need not be goal-*directed* either. The latter are guided by a representation of the goal they are oriented to [76]: a software agent designed to bring about a given effect is goal-oriented. It becomes goal-directed if, for any reason, the effect needs to be represented as an internal state of the agent.

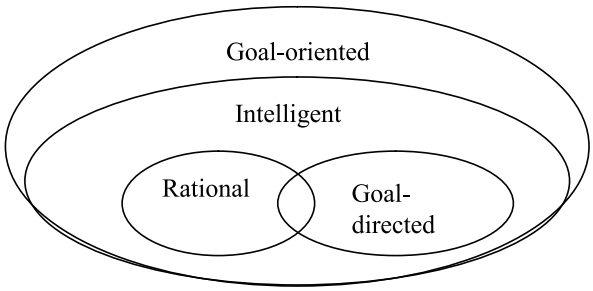

Rational, Goal-Oriented Agents, Figure 1
Set relationships among some types of goal-oriented agents

The distinction between goal-oriented and goal-directed action is crucial, since it allows two main components of computational agents to be distinguished, the design-goals – i. e. the goals of the user/designer that need not be represented into the agent although incorporated in the way it operates and is constructed – and the internal goals, which are represented in the agent that achieves them. Of course, goal-oriented modeling works as a filter for the designer, which enables her to recognize important properties of her domain of interest. It also tells her which internal mechanisms the system ought to possess in order to achieve its design-goal, and whether and which among these goals ought to be internalized.

Within the broad category of goal-oriented agents, we identify intelligent agents, as those entities that use *their own knowledge* to bring about certain effects in the world. Intelligent agents include goal-directed agents, achieving their own goals, and rational agents, maximizing their own utility.

Rational and goal-directed agents are here taken as paradigmatic examples of two complementary approaches in the agent field, endogenous and exogenous: the latter describes agents' behavior as perceived by an external observer; the former models it in terms of the agent's internal mechanisms. Either approach will be characterized separately.

In the final section, rational and goal-oriented agents will be compared, their differences will be outlined and the respective advantages for both theory and application will be pointed out.

Agents

Agents are systems acting on some environment, including their minds. In this case, their action consists of taking decisions, manipulating information, etc. More generally, they act on the physical or social environment producing transitions from one state to another.

They are modeled in more or less formal terms. In general, agent models are abstract, formal, often computational. The latter describe agents as entities carrying out the operations described in an algorithm, accepting inputs from the environment, and performing actions that depend upon the current inputs and produce outputs that modify the environment.

By an abstract model of the agent, it is meant an ideal-type entity endowed with defining properties as specified in the underlying theory. One such entity, which may never be found in nature, facilitates general explanation of real-world phenomena involving natural agents, and provides inspiring models for concrete applications to be designed.

By a formal description, it is meant an abstract description that is also explicit, non-equivocal and sound. It is by no means necessary that such a description be translated into a standard logic language, although this is sometimes the case. Strictly speaking, a formal model of the agent is not necessarily transformed into a computational model either, although this is often the case.

In this contribution, agent *computational* models will mainly be addressed, although the discussion of rational agents will involve mathematical notions, and that of BDI agents – a special type of goal-directed systems – will bring into play *logic* definitions of mental states. In all of these cases, however, formal technicalities will not be provided, and the models will only be reported on in informal terms.

Criteria for Classifying Agent Models

As shall be argued throughout the article, goal-directed and rational agents represent two complementary ways, endogenous and exogenous, to approach intelligent agents, which in turn largely depends on whether agents are instruments or objects of scientific inquiry.

Taken as scientific instruments, agents are modeled and implemented on a computer in order to generate phenomena [33] that are themselves the primary objects of scientific investigation. This is the case with macro-social phenomena *grown* in artificial societies, where agents are specified only to the extent that is required to bring about the (macroscopic) effects of interest. This usually consists of an interesting dynamics at the aggregate level, such as emergent macro-social effects like segregation, public goods provision, opinion dynamics, etc.; or stylized social facts, such as game theoretic scenarios, in order to explore social or collective problems and dilemmas and account for prosocial action, such as cooperation [6], altruism [34], fairness and justice [10], etc.

As targets of scientific interest, agents – in particular autonomous agents [41,96,100] – draw the scientist's attention on the mechanisms of regulation that activate and rule the activity of any system of action. In this sense, agents are studied and described in their internal mechanisms of regulation. Ideally, the language, the concepts, and the formal instruments applied to them, describe agents and their behaviors precisely in the same mechanisms that operate when they behave. Such a description of course corresponds to the real processes as suggested by the current scientific standards.

Which consequences do these different ways of conceptualizing agents bear? The distinction discussed above leads to the emergence of two different approaches to agent modeling, *endogenous* and *exogenous*, which have not as yet been systematically analyzed.

Endogenous vs. Exogenous Approach Generally speaking, endogenous variables are meant as properties originating into the system under study. Exogenous variables instead are properties originating from outside the system, which are not under its control nor are manipulated by the system itself, but can be manipulated by the observer/experimenter.

In computational models, exogenous variables are parameters set to defined values by the computer scientist. It is interesting to observe that an extended parameters' space poses a serious problem of computability to the analyzer (see [99], for a recent review of alternatives and prospects), who must arbitrarily choose which parameters to analyze extensively. Therefore, attention is currently paid to the so-called parameters' "endogeneization" [19].

By endogenous models, it will be here meant to refer to models describing agents and their behaviors in terms of a special type of endogenous variable, i. e. their generative mechanisms. These are supposed to operate *within* the agents when they exhibit any given behavior, including responses to external stimuli. Exogenous models, instead, describe agent's behavior from the outside in terms of its effects or functions (e. g., utility or fitness function). An exogenous model is indifferent to how behavior is actually ruled; it is limited to apply a formal description that accounts for the behavior under study as this is perceived by the external observer.

Notice that the exogenous and endogenous approaches are not necessarily concurrent: they tackle agency from two different perspectives. For example, on the grounds of an exogenous model, one can assume that a certain opinion and the corresponding behavior spread over a population as an effect of the physical space (see [68]). It is interesting to notice that so far, computa-

tional models have adopted exclusively an exogenous approach to the study of opinion dynamics [30,51,71,73,90], modeling the external dynamics of opinions, i. e. their dynamics in society, independent of the endogenous mechanisms allowing agents to accept [84], update [58], revise [12], transmit [72], retreat and finally drop their opinions. However, at least in principle, an endogenous model of opinion dynamics is necessary to unfold the social dynamics: to understand how physical or social distance impacts on opinion dynamics, we ought to understand how opinions *gain ground*, agent by agent, mind by mind. In other words, we ought to model the process by means of which agents influence one other and accept others' opinions.

One might say that these two approaches tackle different aspects of the same phenomena, or different steps of the same processes, the exogenous model being focused on the so called distal causes, the endogenous model addressing the proximal. At a more careful sight, however, this does not seem to be the case. Indeed, it should be noticed that proximal causes are sometimes exogenously modeled. This is the case with game-theoretic models that describe decision-making (proximal cause of behavior) by means of a utility function taking into account wealth distribution and risk sensitivity [42,67]. This can hardly be conceived of as an internal mechanism ruling agents' decision-making. While taking what frequently proves to be an intelligent, adaptive decision, agents haven't the faintest idea that the process they have gone through might well be described by means of a mathematical function. Far from identifying the mechanisms generating the agent's choice from within, this function describes behavior from the outside.

Game-theoretic models have often been charged with the accusation of proceeding from implausible assumptions. This accusation is warranted from the point of view of an endogenous model, not from an exogenous standpoint. That agents maximize their expected utility is probably a truism, at least if utility is not meant to coincide with self-interest [39]. If utility is meant to coincide with own desires or goals [5], the rational assumption is perhaps too obvious, even tautological, but not implausible. What makes it implausible is the description in which it is conveyed: agents do *not* calculate mathematically their expected utility, although many times they *act as if they were doing so*. The utility function is a valid exogenous model, of course, but is an implausible endogenous explanation.

The problem is, what is a good endogenous explanation? What are the internal mechanisms that allow agents to maximize their utility? This contribution does not, nor is intended to provide an answer to this question. However, a good step forward in the direction to "endogenize" factors of regulation, is to turn to current scientific theories that aim explicitly at elaborating an endogenous model.

Analogously, we should refrain from the common fallacy that distal explanations can easily do without endogenous models. Things are not so simple. If one aims at generating the behavioral effect starting from a far removed cause, one needs to construct the whole chain between distal causes and wanted effects. Without an endogenous model, one can get nothing better than a partial and non-generative explanation. This is not equal to saying that this explanation has no right of its own. On the contrary, and somewhat in contrast with latest generativists [33], the present writer's claim is that exogenous models are non-generative statements that may have explanatory value [22].

As Nwana [81] argued in his overview of the field of computational agents, one of the symptoms of the ontological deficiency characterizing this field is the number of adjectives often accompanying the word agent. Rather than adequate definitions of these adjectives, which are easily accessible both in Nwana's or in Wooldridge and Jennings' papers [101], the present contribution attempts at discussing major criteria for categorizing agents and combining them with the epistemological approaches discussed above.

Weak vs. Strong Agents Wooldridge and Jennings [101] distinguished between a weak and a strong form of computing agents. Weak agents are characterized by a number of properties.

Autonomy Computing agents operate without the direct intervention of humans or others, and have some kind of control over their actions and internal state. Agents can be placed on a continuous dimension of autonomy, as several intermediate cases occur between full autonomy and full slavery. Indeed, limited autonomy [23] is the paradigmatic case of humans. Sources of limited autonomy can be found both in resource scarcity and in agents' sharing a common environment. Due to resource scarcity agents may not possess all of the means needed for achieving their goals (limited self-sufficiency). In a common environment, agents interfere with one another positively and negatively. Both factors make them liable to all sorts of mutual influences.

Social ability Agents interact with other agents (and possibly humans) via some kind of agent-communication language [44].

Reactivity Agents perceive and respond to their environment, including other agents.

Rational, Goal-Oriented Agents, Table 1
Set relationships among types of goal-oriented agents

		Goal-Oriented	
		Endogenous approach	*Exogenous approach*
	Sub-Agents	Reflex	Spin glasses Cellular automata
INTELLIGENT	**Weak-Agents**	Learning Evolutionary Neural nets Goal-directed	
	Strong-Agents	Bounded rational	
		Deliberative (BDI, Cognitive)	Rational

Pro-activeness Agents act in response to their environment, but also take the initiative [101].

In addition to these properties, strong agents are computer systems characterized by mentalistic notions, such as knowledge, belief, intention, and obligation [93] as well as emotions [7,8].

What happens if we combine the dimension of weak and strong agency, with the endogenous and exogenous approaches?

In Tab. 1, goal-oriented agents are classified. In the first row, sub-agents appear, i. e. those lacking one or another of the minimal properties for weak agency identified above. In particular, none of the systems classified as sub-agents are proactive. Below, intelligent agents are shown: they include weak and strong agents, endogenously and exogenously modeled.

Overview: Weak Agents

Weak agents are essentially autonomous, interconnected and communicating systems that effectuate state transitions. Unlike simpler and stronger models, no exogenous examples of weak agents are found. They are all described in terms of their internal mechanisms. Verisimilarly, simpler models lend themselves to be described with mathematical formalisms, which are also applied to model some complex capacities, like rational choice (see rational agents). Instead, weak agents are not modeled to *describe* the effects they produce in the world, but to find out algorithms generating them. The interest of these systems is intrinsically computational, even technological, and they gave a strong impulse to endogenous modeling.

Endogenous Weak endogenous agents are described in their internal mechanisms, essentially rules or subsymbolic representations. However simplified, these enable the agents to behave autonomously, react to external stimuli, and show also some kind of proactive behavior.

Evolutionary In their early time, agents were modeled as rigid machines, which might assume a fairly limited set of states on the grounds of a number of fixed rules. This was the case with cellular automata, which at the origin were fixed to given locations on a toroidal grid and could assume either an active or a nonactive state depending on the states of neighbors. Action was taken according to a fixed set of rules [26].

Several factors contributed to promote flexibility, the principal of which probably was the growing interest in evolutionary phenomena and the capacity of agents to modify and adapt to a changing environment. Fitness function was designed as a function measuring how good a given property or trait is in a pool of traits (for example a gene in a population): the agents performing above a given value would survive and reproduce whereas those performing below would extinguish.

Let us look more closely to computational models based on a fitness function. In these models, agents' behavior is the expression of a vector in which given traits are represented. As the metaphor is clearly the evolutionary mechanism, mutations in the vector may occur with a probability that is established offline. Mutations that contribute to the reproductive success of the agent will be transmitted to the future generations after a recombination with the partner's traits.

Genetic algorithms (GAs) are a particular class of algorithms inspired by evolutionary biology, and implementing evolutionary mechanisms and processes, such as inheritance, mutation, selection, and crossover (also called recombination). Starting from a population of randomly generated individuals, at every generation the fitness of every individual in the population is evaluated, the current population genotype is modified (recombined and possibly randomly mutated) and the new population is then used in the next iteration of the algorithm. The fitness function is problem dependent and in some problems, hard to define.

A GA proceeds through iterated executions of mutation, crossover, and selection operators. Whereas selection is an important genetic operator, the importance of crossover versus mutation is more debatable (references in Fogel [36] support the importance of mutation-based search). To be noticed, for specific problems, other optimization algorithms may do better than GAs with equal

speed, for example simulated annealing. (Simulated annealing (SA) is an algorithm for locating a good approximation to the global optimum of a given function in a large search space. It is often used when the search space is discrete (e. g., all tours that visit a given set of cities). In favorable cases, simulated annealing may be more effective than exhaustive enumeration of the search space [59].)

As with other computational models, one main problem is the parameters' space search (mutation probability, recombination probability, etc.) To which value should one set the mutation rate, when no reference is made to evolutionary phenomena known in nature? A very small mutation rate may lead to genetic drift. On the other hand, a high mutation rate may lead to loss of good solutions.

GAs are an invaluable means for optimizing fitness. A fascinating explanation, the so-called building block hypothesis (BBH), has been proposed to explain this effect [46]. Although heavily criticized (see [102]) and poorly supported by experimental evidence, BBH provides an intuitive reason for the mechanism of adaptation, which is said to result from iterated recombination of "building blocks", or low order elements: at any recombination (or generation), past solutions that proved fitter than their own building blocks are recombined to increase the optimality of new solutions. By this means, genetic algorithm achieves step-wisely optimal performance through the iterated recombination of building blocks.

If GAs are meant to optimize the genotype of any given population, genetic programming (GP) is used to optimize ability to perform computational tasks. After some pioneer computer scientists, and especially John Holland in the early 1970s, John R. Koza ([62,63,64,65,66]) has pioneered the application of genetic programming in various contexts. Results obtained by the application of GP include innovations in hardware as well as computer programs.

Learning Learning algorithms used in agent based models include classifier systems, reinforcement learning, Q-learning, and neural nets.

Classifier Systems Classifier systems (CSs) are strongly related to genetic algorithms. Introduced by John Holland [52], CSs consist of sets of rules optimized thanks to a genetic algorithm that operates on reinforcement rather than on a fitness function. There are two main types of CSs. In one case, the genetic algorithm recombines different sets of rules; in the other, the genetic algorithm operates within one set only.

Reinforcement Learning Inspired to behavioral science, reinforcement has been applied to machine learning as a reward mechanism (see [54]). Reinforcement based algorithms have been applied in economics and game theory as a boundedly rational interpretation of how agents endowed with the same learning mechanisms may converge on one equilibrium. Given a set of world-states, a set of actions A, and a set of "rewards", at each time the agent chooses an action and receives a certain reward. Reinforcement will cause agent to maximize the quantity of reward obtained.

Recently, reinforcement learning has been used in cognitive models of human problem solving (e. g., [43,47]) and error-processing [53].

Neural Nets Borrowed from neuroscience, where they are identified as groups of neurons performing a specific physiological function, neural nets currently refer to artificial neural networks [50]. These are made up of interconnecting artificial neurons, and used either to study biological circuits – a simplified view of artificial neural networks may be used to simulate properties of neural networks – or to solve artificial intelligence problems, such as speech recognition, in order to construct software agents or autonomous robots.

Neural networks have been viewed as simplified models of the brain, but the relation between this model and the brain architecture is far from obvious. After all, computers entail sequential processing and are based on explicit instructions, whereas neural networks model biological systems as performing parallel processing on implicit instructions.

An artificial neural network is an adaptive system. Its structure gets modified based on information spreading through the network.

Neural nets are based on sub-symbolic representations. These are physical sets of interconnected units, with variable weighs on connections. The resulting networks are said to take inspiration from the organization and structure of neuronal synapses. As they *stand for* nothing, they hardly fit the notion of "representation" as such. Nonetheless, they effectively play a number of important functionalities of representations, especially pattern-recognition. Given a certain input, a given set of internal units may be gradually "taught" to "recognize" it by contrast with other inputs, thanks to a suitable modification of the weighs of given connections. Subsymbolic representations are

- Quantifiable, whereas symbolic representations vary, and can be compared on, a qualitative base.
- Gracefully degradable, whereas symbolic representations are none-or-all. Hence, unlike these latter, they al-

low for partial and reparable damages of the long-term memory.
- Compatible with non-decided upon and non-controlled activity.

Applications of artificial neural networks include pattern and sequence recognition [87] (radar systems, face identification, object recognition, etc.), sequence recognition (gesture, speech, text recognition), medical diagnosis [39], data mining, etc.

Goal-Directed Agents Goal-directed agents are systems ruled by, and described in terms of, their goals.

In a psychological-behavioral sense, a goal is a motivational factor defined on the grounds of a number of observable features, among which vigorous attainment, persistence in the face of obstacles, resumption after disruption ([69], etc.).

Another notion of goal derives from the theory of systems and cybernetic circuits. Control theory introduced the notion of feedback [15] to control states or outputs of a dynamical system [40]. Its name comes from the information processed in the system: inputs (e. g. a given state of the world) have an effect on the outputs (for example, elicit a given action) that is measured with sensors, and the result of which (the worldstate brought about by the action) is used as input to the process.

On the grounds of such a notion of goal, a theory of behavior as planning activity [78] has been constructed, which has played a foundational role in the development of cognitive science.

Building on control theory, cognitive scientists define a goal as a wanted state of the world that, if discrepant from the currently perceived worldstate, activates the agent and rules its action. A goal is therefore both a trigger and a regulatory state: actions are activated and ruled by their goals [23].

Further analysis allowed several categories of goals to be identified.

End- and sub-goals. A goal can either be the final end of activity – f.i., survival is probably the supreme finality of natural systems – or a condition for its execution, which may not be verified in real matters. In such a case, in terms of AI planning, the initial end-goal gives rise to further sub-goals, i. e. to verify the necessary conditions for action execution.

Innate vs.learned. Whereas end-goals are usually innate (but examples of subgoals that become ends in themselves exist), subgoals are formed and dropped after successful execution.

Side-goals. Goals may guide the action they trigger, or they may rule actions triggered by others goals. The latter are side-goals. A typical example is utility-maximization, which rules a number of activities, especially economic, triggered by other goals.

Achievement and maintenance goals. Goals may be already verified in real matters, i. e. coincide with currently perceived worldstates. In this case, they are usually dropped – as is the case with purchasing a new car, or finding a job. However, in many cases, these fortunate circumstances may suggest a new goal, i. e. to keep one's job or car. In the BDI language (see below), the goal to maintain a current worldstate is a maintenance goal, whereas the goal to realize a given worldstate is an achievement goal [20].

Realizable vs. unrealizable goals. In the logic-based treatment of goals, these must be consistent and realizable. Unrealizable goals are treated as a subset of desires that are never chosen for action.

The first application of goal-directed agents goes back to the AI planning systems [89,97], which soon gave impulse to a Copernican revolution in planning and plan execution, i. e. Distributed Artificial Intelligence (DAI) systems [11]. In the early nineties, the first studies and scientific events [31] on Multi Agent Systems (MAS) took their moves from the necessity to increase effective autonomy of distributed systems and their interactive capacity. Both of these requirements, in turn, necessitated developments of mental capacities, especially in the direction of social intelligence [21], what soon led to the appearance of stronger models of agency.

Overview: Strong Agents

Below, we will review the main approaches to account for complex tasks, like deliberative choice. As was premised above, both the endogenous and the exogenous approaches concurred to this aim. Let us examine them in turn.

Exogenous Here, rational agents are shown to represent the paradigmatic case of exogenous models of strong agents, whereas BDI and cognitive agents examplify strong agents endogenously modeled.

Rational Agents One way of conceptualizing rational action describes it as *reason-based* action. Hence rational agents are sometimes identified with intelligent agents at large [88,100].

In this contribution we go back to the classic notion of rationality as utility-maximization, and define rational

agents as a subset of intelligent agents that act *so as* to maximize their utility [2,80]. In this more specific sense, a rational agent takes actions expected to maximize its chances of success [4]. Its action is said to bring about this effect *no matter how*. In other words, a rational agent model is indifferent to the mechanisms that effectively operate within the agent to ensure the hypothesized effect [1]. At most, what is attributed to a rational agent is (limited) knowledge of its environment, consisting of (a memory of) past experience, current information about the environment (possibly including other agents and their expectations [70]), as well as the estimated benefits and the chances of success of own actions [88].

The classic conceptualization of rational agents derives from the abstract model of *homo oeconomicus* defined by John Stuart Mill [77] as an entity which aims at possessing wealth and is able to evaluate and adopt the means for achieving it. On these assumptions, rational choice took on the specific meaning of *self-interested* action. Thus meant, the term "rational" implies a utility function: independent of the benefits themselves, what is said to be rational is the choice that maximizes the distance between the benefits obtained and the costs sustained to obtain them.

Rational action theory provides a formal model of practical behavior, mainly applied to social and economic domains. We might consider it as the still dominant paradigm in economy, but its hegemony extends to all of the social sciences and to the computational field of agent theory and technology. Essentially, it assumes individuals to choose the best action according to stable utility functions. Rather than describing reality, rational choice theory aims at supporting reasoning (prescriptive use) and modeling social behavior in highly stylized scenarios.

Requirements of Rational Action Let us examine the view of the agent implied by the rationality assumptions. Choosing an action rationally requires individual preferences, a set of options or alternatives for actions, and a set of expectations about their outcomes. Two important assumptions concern preferences:

Completeness All actions are ranked in an order of preference.

Transitivity If any given action a_1 is preferred to action a_2, and the latter is preferred over a third option a_3, then a_1 is preferred to a_3.

Consequently, an individual can always order the available options on the grounds of its preferences, which will always be consistent. This gives rise to a utility function, i. e. an ordinal number assigned to alternative actions – such as: $u(a_i) > u(a_j)$ – with preferences being defined as relations between these assignments. Unrealistic assumptions are often made, such as full or perfect information and the ability and time to weigh every choice against every other choice. Relaxations of these assumptions are included in theories of bounded rationality, which will be examined later.

Although empirical support to rational choice theory is far from satisfactory [48], this paradigm presents two main advantages for the social sciences.

First, precisely because it proceeds from stylized facts and abstract phenomena, it proves a powerful instrument for the scientific study of social action.

Secondly, and consequently, rational action theory in principle allows future actions to be predicted. Compared to stochastic modeling, the predictive power of social scientific explanation is expected to increase. However, it should be observed that the tautological and a posteriori explanation of behavior deriving from the principle of utility maximization strongly reduces the predictive potential of the theory. As preferences and goals are subjective, any choice results in the maximization of utility, whatever its outcome is and can be said to be rational (tautology), once it has been made (*a posteriori*).

Third, the abstract and strong nature of the theory allows testable hypotheses to be formulated and data to be collected. The experimental potential of social scientific explanation is therefore bound to increase.

However, rationality theory has been criticized on empirical grounds. In traditional societies, people's social action, for example the way they exchange goods, follows rules that could not be expected on the grounds of rationality assumptions and which, thanks to the famous work by Mauss [75], seemed to point to a regime of "gift", rather than market, economy. In addition, according to experimental scientists (cf. the wide literature cited in [82]), people are much more cooperative than expected by the rational choice theory. Finally, the rationality assumptions are commonly said to necessitate too much understanding of macroeconomics and economic forecasting for people to take rational decision. We will get back on this argument later in the next section.

Theoretical critiques are no less strong. Among others, it has been observed that the role of intrinsic motivations [29], for example that of altruism, is under-estimated to the advantage of extrinsic, incentive-based, motivations, an objection that is not entirely warranted in consideration of the indetermined nature of preferences: rational agents can be subjectively inclined to altruism. In such a case, even one's extreme sacrifice to the benefit of others' is perfectly "rational".

Rather, the inadequate account of preferences dynamics poses a twofold problem to the social scientist. On one

hand, totally indeterminate preferences reduce the predictive potential of the rationality paradigm. On the other, if intrinsic motivations are acknowledged, little is said about how they can be acquired. As a consequence, the origin of altruistic motivations, like that of any other taste and inclination, and the effects of social influence, training, education, and the like on the utility function are overlooked.

Another consequence of rational agents' poor internal dynamics is the insufficient account of inner conflicts like that between motivations and goals (for example, passing the exam and playing around) or that between types of motivations or finally between individual goals and societal values, which may strongly affect and impair individuals' decision-making.

Substantially, these critiques converge on arguing against a purely exogenous approach, which ignores the acquisition, modification and integration of internal states. To say that agents' choices proceed from their preferences is neither particularly interesting nor informative, but to make the model more realistic does not help much and may render it tautological. The missed point here is a theory of the internal dynamics of preferences. However, this requires an endogenous approach to rationality.

Endogenous Strong models of the agent aiming at modeling internal states and mechanisms include a variant of the rational choice theory, i. e. bounded rationality, as well as theories and architectures of cognitive agents, including but not reduced to BDI agents.

Bounded Rationality Bounded rationality is a variant of rational choice theory that took its move from the obvious consideration that perfect knowledge is never available, and that agents, even investors, deal with uncertainty and risk in business. Far from regretting the special attention paid to economic behavior and the *homo oeconomicus* model, the proposers of bounded rationality questioned the utility of full rationality as a paradigm for the study of economic behavior itself. As Fox and Tversky [37] observed, the very notion of uncertainty was first questioned by theoretical economists. In [61], Knight distinguished between measurable uncertainty or risk, and unmeasurable uncertainty, suggesting that entrepreneurs deal with unmeasurable uncertainty, rather than risk. At the same time, Keynes [59] distinguished between probability of occurrence of a given event and the weight of the evidence supporting it. Nonetheless, game theorists showed little interest in the issue of vagueness of probability [91].

When one considers it carefully, bounded rationality is situated at the borderline between endogenous and exogenous approaches. Prospect Theory [56], one of the main theoretical contributions of bounded rationality, was proposed as a psychologically realistic alternative to rationality theory, in particular to expected utility theory. The authors meant to account for empirical evidence showing that when people evaluate potential losses and gains under risk, e. g. in financial decisions, the probability of occurrence acts differently on the values of alternative options depending on whether agents are gaining utility or avoiding losses. Rather than constructing a model that shows how this happens in the minds of the agents, the authors expressed this evidence in an asymmetric utility function. This shows a bigger impact of losses than of gains (loss aversion) and expresses that people tend to overreact to small probability events, but underreact to medium and large probabilities. The theory, meant as an endogenous development of full rationality, is expressed into an exogenous formalism.

Nonetheless, bounded rationality has done a great job in revising the assumptions of full rational theory. What is more, Herbert Simon [94] did actually argue for an endogenous model, when he claimed that in order to account for rational action, it is necessary to study the mental processes and mechanisms allowing agents to formulate and solve complex problems and process (receive, store, retrieve, transmit) information (see also [98], p. 553, quoting Simon). One important aspect of the mental endowment is represented by heuristics [57,83], i. e. powerful informal methods to solve problems in complex situations. In psychology, heuristics are simple, efficient rules, rules of thumb, which either evolve or are learned and support people in taking decisions, coming to judgments, and solving complex problems under incomplete information. When agents cannot compute the expected utility of every alternative action, they resort to heuristics, which are applied in different contexts and might have been acquired over the course of life, through repeated decision-making processes. These rules work well under most circumstances, but in certain cases lead to systematic cognitive biases, i. e. very basic statistical and memory errors that are common to all human beings.

To be noticed, most models of bounded rationality did not strictly follow Simon's ideas. Some authors preferred to model people's decisions as a sub-optimal version of rationality [32] or to study how people cope with their inability to optimize. For others [45], alternatives to full rationality in decision making frequently lead to better decisions.

Deliberative Agents Let us now turn our attention to fully endogenous models of complex agents, such as BDI architectures and more generally cognitive agents. Both

are goal-directed systems, not incompatible with utility maximization but actually implementing it. However, goal-directed agents do not necessarily imply neither explicit computation of utility maximization, nor heuristics achieving it.

Cognitive systems. These are goal-directed agents endowed with mentalistic notions, such as beliefs, intentions, obligations and the capacity to manipulate mental symbols and accomplish a number of operations on them, including inference, reasoning, problem-solving, planning, etc. Symbolic representations *stand for* the states of the world they refer to in such a manner that these world states can be:

- Compared: a given representation of a worldstate ws1 can be compared with another to check whether they are equal or not. This is fundamental in goal-directed action.
- Manipulated, e. g. asserted or denied, evaluated and possibly selected along common criteria, modified and revised and therefore used as instruments for reasoning, action planning, etc.
- Embedded into one another, thereby giving rise to meta-beliefs. Among other reasons, meta-beliefs are useful for the representation of others' beliefs (the so called "theory of mind"), which is of vital importance in social life [16,85]. Of late, theory of mind has been used to refer to a specific cognitive capacity: the ability to attribute mental states – beliefs, intents, desires, pretending, knowledge, etc. – to oneself and others and to understand that others have beliefs, desires and intentions that are different from one's own [27,49].

BDI agents. A BDI agent is an architecture for a special type of cognitive agent, more specifically for intention-driven agents, endowed with particular mental attitudes, viz: Beliefs, Desires and Intentions (BDI).

The philosophical bases of BDI agents is Bratman's theory of practical reasoning, but has been formally described by Anand Rao and Michael Georgeff's [86], which combine temporal logic with a modal logic enriched with beliefs, desires and intentions.

Beliefs Beliefs represent the agent's representations about the world (including itself and other agents), which are neither necessarily true nor complete. Inference rules generating new beliefs allow to enrich the agent belief base.

Desires Desires represent the motivational state of the agent. Goals are a subset of desires, i. e. consistent and achievable ones.

Intentions Intentions are executable goals, i. e. goals that the agent has chosen for action, to which the agent has to some extent committed (in implemented systems, this means the agent has begun executing a plan).

Aimed to represent the future as an epistemic tree, allowing agents to reason upon alternative courses of action, BDI architectures lend themselves to model the interplay among different mentalistic notions. In particular, an extension of BDI, the BOID architecture, concerns the representation of the interaction among beliefs, obligations, intentions, and desires, and has been used to support reasoning and deliberative choice on legal norms [14].

Mental Representations and Their Dynamics An endogenous approach to the computational model of intelligent agents revolves around the main question as to how agents represent their goals and the information needed to achieve them [28,74,95]. This issue, called knowledge representation, arises at the intersection between cognitive science and Artificial Intelligence: AI scientists have borrowed methods of knowledge representation from cognitive scientists, since it was of primary importance for them to design programs that could store knowledge and re-apply it to analogous problems.

Knowledge Representation techniques – such as frames, scripts, etc. – stem from theories of human information processing. The main goal is to represent knowledge in such a way as to facilitate reasoning, planning, problem solving and the like.

The main questions posed by AI researchers concerning knowledge representation focuses on the format of representations, the nature of knowledge, the preferability of general purpose vs. domain-specific representation schemata, the degree of expressiveness and detail of such schemata, whether it should be declarative or procedural etc.

In AI, many methods of knowledge representations have been tried since the early seventies, e. g. theorem proving and expert systems, with variable degree of success. The major knowledge representation systems go back to the eighties when for example large databases of language information were built, rendering knowledge representations more feasible. Several programming languages were developed, such as Prolog, KL-ONE and XML, facilitating information retrieval and data mining.

The necessity to store and manipulate knowledge in a formal way so that it may be used by mechanisms to accomplish a given task has drawn the attention of knowl-

edge representation scientists. Expert systems, machine translation, and information retrieval are good examples of applications.

Probably the most extensively used structure for knowledge representation from the 1960s, is the knowledge *frame* [35,79], with its own name and set of attributes, or slots that contain values. Frames have been used for expert systems in object-oriented programming, with inheritance of features described by an IS-A link, a type of relation that has posed a number of problems of semantic interpretation [13]. However, in an endogenous perspective, it should probably be observed that frame structures are well-suited for the representation of stereotypes and other socio-cognitive patterns.

A *script* [92] is a frame that describes events, especially behavioral events. The usual example is going to a restaurant, with steps including waiting to be seated, receiving a menu, ordering, etc.

Goals Typologies of goals are not limited to those examined in Sect. "Goal-Directed Agents", but include other sub-categories, which are based on the interplay among goals and other mental notions, such as beliefs and obligations.

Active vs. inactive goals. Maintenance goals are not always active. A goal is active when it rules the system's behavior. When a goal is represented in the agent's mind but does not rule its behavior is inactive. A maintenance goal may be activated when it is likely to be thwarted. An inactive goal may be a maintenance or side-goal, or an achievement goal that becomes (momentarily) inactive because incompatible with a more important one (emergency or conflict).

Executed vs. waiting. An active goal may be chosen for action, or may be interrupted during its execution and kept waiting in a queue, because the conditions for its execution turn out to be false, and must be realized for the goal to be achieved.

Chosen. An active goal may be chosen for action and transformed into an intention when (a) it is not realized, but it is (b) realizable and (c) more important than other active goals, if any. These clauses are conjunctive: in case the goal is realized, or outcompeted, it becomes inactive. In case it is unrealizable, it is dropped (or retrocedes to the status of a simple desire).

Individual vs. social. Goals are individual when mentioning no-one else except the hosting agent. Otherwise, they are weak or strong social goals. Weak social goals mention others as models to imitate or as *sources* of influence – for example their future actions are perceived as either obstacles or favoring events. Strong social goals mention others as *targets* of influence (simple request or persuasion or manipulation or coercion).

Prosocial vs. aggressive. A goal is prosocial when it aims at realizing a worldstate *as long as* and *why* this is wanted by another agent. The process by means of which an agent comes to have another's goal as its own is called goal-adoption. This may be either an end in itself (benevolence) or a sub-goal for an individual goal (e. g., exchange).

An aggressive goal is a goal that aims at thwarting another's goal, as long as and why this is wanted by the latter. Again, an aggressive goal may be an end-goal (hate), or a sub-goal, and in the latter case it may be instrumental to a selfish or a pro-social goal (prevent crime), even to the benefit of the victim (surgery, education, etc.).

Shared vs. collective. Finally, a goal may be shared by a set of agents. It is collective [25] when it mentions a set of agents in which any one is necessary, but none is sufficient to achieve the goal shared by them all (an orchestra or a football team are typical examples of collective goals).

Cognitive Dynamics In a rather generic sense, an autonomous agent is a self-interested agent. In a more specific sense, an autonomous agent is one, which has internal criteria to select among inputs. Inputs might generate beliefs and goals. An autonomous agent is therefore characterized by a "double filter architecture", allowing both beliefs and goals to be selected (cf. [17]). These two filters are sequential, but at the same time they allow for an integrated processing of mental representations.

Filtering beliefs. Thanks to this filter, agents have control over the beliefs they form. This filter is rather complex and implies that a number of tests be executed over a candidate belief against several distinct criteria. These are pragmatic or epistemic criteria. Epistemic criteria include

- Credibility, with agents controlling, among other properties, the coherence of candidate beliefs with previous beliefs, the reliability of the source of the candidate belief: agents accept information from other agents (Gricean principles) provided they have no reasons to doubt their sincerity or competence.
- Non-negotiability, or the Pascal law. To believe or not is a "decision". However, it cannot be made in view of one's pragmatic utility, but only in view of one's epistemic utility. In social interaction, we cannot use threats ("Argumentum ad baculum") or promise to make people believe something. The difference between persuading to do and persuading to believe is crucial. Since beliefs control goals, this represents a further protection of agents' autonomy.

Pragmatic criteria concern the reasons for believing something. Typologies of beliefs generally concern the format of their representation (declarative, procedural); the degrees of certainty; the levels of nesting (one can believe something, without believing that one believes . . .). These typologies are known too well. Perhaps it is less obvious that beliefs may have a different "status" in the mind according to the motives for acceptance. The language provides a rich vocabulary: superstitious belief, creed, faith, credence, doctrine, postulate, axiom, principle, conception, idea, view, opinion, and many others. These beliefs vary on several, often quantitative, dimensions, such as certitude (subjective truth value), retractability (how likely it will be modified), connectivity (how much a belief is connected with other beliefs). One interesting quantitative dimension is the "force" of beliefs (cf. the role of this dimension in the Social Impact Theory, cf. [68]): beliefs vary on how strongly they are held. This is related to certitude, but also to the motives of acceptance, which may lead the agent to ignore the belief's truth-value; for example,

- Self-protection and self-enhancement, agents may be led to accept one among several competing beliefs because of the belief's positive effect on their self-esteem or self-concept.
- Commitment to a given (set of) belief(s) provides one important reason for accepting further, consistent beliefs, despite or independent of incompatible evidence: agents that accept beliefs out of commitment don't check their truth value.
- Hypothetical and counter-factual reasoning, beliefs may be (transitorily) held as means to reason and carry out operations (demonstrations, proofs, experiments). A god example is the priest accepting the atheist's point of view to dismantle it.
- Communication: A psychotherapist may "accept" the delusions of her patient in order to communicate with him and give a clinical sense to his fantasies; here the goal is not to carry out a counterfactual argumentation, but to understand the meaning of delusions.
- Empathy: agents may want to share the views of their close connections.
- Risk-taking or gambling: agents may participate in lotteries, accepting one alternative and investing (money) in it; in such a case, agents will hold an uncertain belief but behave as if it were certain.
- Prudence, agents may accept uncertain information (for example, rumours, gossip, even calumnies) and behave as if they were certain: unlike the preceding situation, a risk-minimization strategy applies.

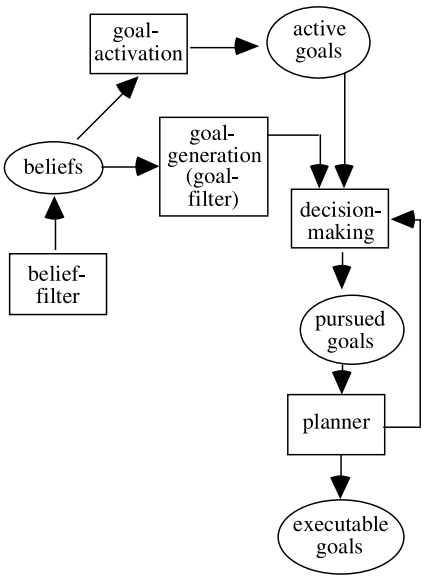

Rational, Goal-Oriented Agents, Figure 2
The "double filter" architecture

Filtering goals. There are at least two fundamental tests which are performed on goals: self-interested goal-generation (an agent is autonomous if whatever new goal it comes to have, there is at least another goal of that agent, for which, in the agent's beliefs, the former is a means), and belief-driven goal-processing (any modification of an autonomous agent's goals can only be allowed by a modification of its beliefs) (cf. Fig. 2).

Both these filters bear interesting social and cultural consequences. First, agents' minds are modified thanks to a process of belief-formation or belief-revision. Secondly, belief-formation and -revision are decision-based and selective processes. With Cavalli-Sforza and Feldman [18], we will speak of beliefs' "acceptance". A cultural process is a process interspersed with decisions taken by the agents involved. But a decision-based process is not necessarily explicit and reflected upon: mental filters do not necessarily operate consciously: agents may not be able to report on them. Thirdly, agents will never accept beliefs under threat, or in order to obtain a benefit in return (non-negotiability). Fourth, agents may accept beliefs for different reasons, and these will affect the probability that such beliefs are held, their strength, and their transmission. The social mechanisms of influence and transmission are strongly intertwined with criteria and motives of acceptance. This leads us to the question of limited autonomy.

Limited autonomy. The model of agent outlined above appears rather abstract and unrealistic. In the real life,

agents appear liable to external influence, prone to accept and transmit prejudices, victim of superstition, prey to false doctrines and creeds. Indeed, autonomy is limited both at the belief and at the goal level: agents are liable to being influenced by external (including social) inputs.

Both at the level of goals and beliefs, autonomy is limited in a very elementary sense: agents are designed to take into account external inputs, if only to discard them later. If an input is received, the filter processing is activated (cf. [17]). At the goal level, agents cannot avoid accept very elementary requests: if somebody asks a passer-by the time, this will not keep on ignoring the request. At most, the passer-by can pretend that she did not perceive it. But if this perception cannot be concealed, an answer whatsoever will be given, if only to say that one has no idea what the time is (minimal level of adoption). Of course, this type of influence is rather superficial and ephemeral. But it paves the way for other more relevant types of influence. Obviously, agents' autonomy is limited because they are not always self-sufficient. They may need the help of other agents to achieve their goals (social dependence), and this causes agents to adopt others' goals and to accept their requests. However, one adoption of others' goals will always be a means for the achievement of one's goals for example through social exchange or cooperation. In turn, these social actions favor the transmission of beliefs, including action plans, techniques, procedures, rules, conventions, social beliefs. Finally, agents' autonomy is limited by norms, which are aimed to regulate agents' behaviors. But agents may accept or reject norms, comply with or violate them, always according to their internal criteria for acceptance.

At the level of beliefs, agents' liability varies depending upon the type of beliefs. For example, social agents are strongly permeable to social evaluations, rumors, gossip, even calumnies (cf. [24]). Rumors and gossip are accepted for prudence, and, as we will see, this favors their spreading. Indeed, these are important phenomena of memetic transmission, which spread social labels, stigmas, and prejudices, but also reputation, social hierarchies and other institutions.

Concluding Remarks and Future Directions

Which are the similarities and different advantages of these different paradigms in agent modeling, especially among rational and goal-directed agents? As initially stated, these should not be viewed as mutually exclusive. Hence, the question arises as to when, for which scientific objectives, and how they could be integrated. But first, let us see what they have in common and where they differ.

Rational and goal-directed agents have three major common properties: autonomy, (limited) knowledge, and strategic action. These attributes characterize the fundamental endowment of intelligent agents, allowing them to act on their own criteria, reasons, and current beliefs.

To be noticed, however, these properties are differently characterized within the two approaches.

Autonomy. Whereas rational agents are self-interested systems – with self-interest being closely related to the notion of adaptation and fitness in a biological sense – goal-directed agents are systems pursuing their own goals, whether these are means for own fitness or not. A suicide is a goal-directed agent, which autonomously decides to take out its life. As this example shows, goal-directed agency is an abstract and general category that can easily be applied to natural systems, which not always pursue their self-interest.

Limited knowledge. Both approaches model agents' beliefs as limited and uncertain. However, the rational approach does not account for the internal dynamics of beliefs, their acquisition, revision, and eventual dropping. Even adaptive views of agents fail to account for the internal dynamics of their motivations, and for the role these have in the management, update, and decay of their beliefs. However, the double filter architecture allows cognitive dynamics of goals and beliefs to be accounted for.

Strategic action. As to this aspect, i. e. the social application of rational action, one should acknowledge a superiority of the rational agents, paradigm, as this is inherently devoted to the solution of social and collective dilemmas. However, in the rationality paradigm, strategic action is meant exclusively in the weak social sense meant above, i. e. as the ability to take into account what other agents know, can do and will do. The strong meaning instead is essentially overlooked, since rational agents are not modeled as goal-directed: they cannot be said to pursue the goal to *modify* others' beliefs, goals, and actions. This is one of the main reasons why it is difficult to integrate social influence, training, education, and the other phenomena of social change in the rationality paradigm.

Despite the aforesaid shortcomings, the rationality paradigm is undeniably influential not only in the social, but also in the computational field. In comparison, the impact of goal-directed agents is far from satisfactory.

There are multiple reasons for such a state of affairs.

First, a persistent cultural hegemony of Homo Oeconomicus should not be ignored. However indeterminate subjective preferences are in principle declared to be, investment and economic business are tacitly assumed to be the master domains of application of rationality.

Secondly, rationality paradigm has all the advantages of an exogenous, in particular a mathematical, approach: (a) quantitative data, (b) theorem-proving, and (c) the reduction of computational complexity, as the rational agent architecture is rather simplified.

Thirdly, and most importantly, the rationality paradigm enjoys the powerful simplicity of abstraction, the invaluable scientific appeal of stylized facts. No other paradigm, so far, has been able to design such simple, highly abstract scenarios, characterized by the same heuristic power, as those represented by the *games* worked out by rational choice theory for the study of strategic action.

Making a difficult but fascinating exercise, let us endeavor to envisage the future of these paradigms. Whilst the rationality paradigm can reasonably be expect to maintain its hegemony in the short term, what can we expect to happen in the middle and long run? The growing success of the neuroscientific approach (think of *mirror neurons*) leads us but to foresee a dominance of the neural nets approach not only within the behavioral sciences (e. g., psychology), but also – and thanks to embodied agents and robotics – within the computational sciences. How and to what extent can this model answer the questions currently answered by rational agents, on one hand, and by goal-directed agents, on the other?

Largely, the answer depends on whether or not the predictable advances of the neuroscientific approach will empower agent technologies, for example, in the direction of social and collective intelligence, acquisition of norms and social values, formation and recognition of institutions etc. However, the answer might also depend on the potential cross-over between the neuroscientific approach and the complexity one. To what extent can embodied agents and neural nets be combined with sociophysical models? The future status of the rationality paradigm depends on the probability of this cross-fertilization. More generally, the future of the agent paradigms depends on the probability of integration between endogenous and exogenous paradigms.

Bibliography

1. Allingham M (2002) Choice theory: a very short introduction. Oxford University Press, Oxford
2. Anand P (1993) Foundations of rational choice under risk. Oxford University Press, Oxford
3. Aristotle (2006) Metaphysics book theta. Oxford University Press, Oxford [translated with an introduction and commentary by Stephen Makin]
4. Arrow KJ (1987) Economic theory and the hypothesis of rationality. In: The new palgrave: a dictionary of economics, vol 2. Macmillan, Houndmills, pp 69–75
5. Aumann RJ (2006) War and peace. Nobel Lect PNAS 103(46):17075–17078. November 14, 2006
6. Axelrod R (1997) The complexity of cooperation. Princeton University Press, Princeton
7. Bates J (1994) The role of emotion in believable agents. Commun ACM 37(7):122–125
8. Bates J, Loyall BA, Reilly SW (1992) An architecture for action, emotion, and social behaviour. Technical Report CMU-CS-92-144. School of Computer Science, Carnegie-Mellon University, Pittsburgh
9. Bigelow J, Rosenblueth A, Wiener N (1943) Behavior, purpose and teleology. Philos Sci 10:18–24
10. Binmore K (1994) Playing fair: game theory and the social contract, vol 1. MIT Press, Cambridge. Vol. 2: Just Playing: Game theory and the Social Contract
11. Bond AH, Gasser L (eds) (1988) Readings in distributed artificial intelligence. Morgan Kaufmann Publishers, San Matteo
12. Boutilier C (1995) Generalized update: belief change in dynamic settings. In: Proc of the 14th international joint conference on artificial intelligence (IJCAI'95). Morgan, Kaufmann, Montreal, pp 1550–1556
13. Brachman RJ (1983) What IS-A is and isn't. An analysis of taxonomic links in semantic networks. IEEE Comput 16(10):30–36
14. Broersen J, Dastani M, Hulstijn J, Huang Z, van der Torre L (2001) The BOID architecture conflicts between beliefs, obligations, intentions and desires. In: Mueller JP, Andre E, Sen S, Frasson C (eds) Proc of the 5th international conference on autonomous agents, Montreal, 28 May–01 June 2001, pp 9–17
15. Bush V (1929) Operational circuit analysis. Wiley, New York
16. Carruthers P, Smith PK (eds) Theories of theories of mind. Cambridge University Press, Cambridge
17. Castelfranchi C (1997) Representation and integration of multiple knowledge sources: issues and questions. In: Cantoni V, Di Gesù V, Setti A, Tegolo D (eds) Human and machine perception: information fusion. Plenum, New York, pp 236–254
18. Cavalli-Sforza LL, Feldman MW (1981) Cultural transmission and evolution: a quantitative approach. Princeton University Press, Princeton
19. Chavalarias D (2007) Endogenous distribution in multi-agents models: the example of endogeneization of ends and time constants, M2M 2007. In: 3rd international Model-to-Model workshop, Marseille, France, 15–16 March 2007
20. Cohen PR, Levesque HJ (1990) Rational interaction as the basis for communication. In: Cohen PR, Morgan J, Pollack ME (eds) Intentions in communication. MIT Press, Cambridge, pp 221–256
21. Conte R (1999) Artificial social intelligence: a necessity for agent systems' developments. Knowl Eng Rev 14:109–118
22. Conte R (2007) From simulation to theory (and backward). In: Squazzoni F (ed) Proc of the epistemological perspectives of simulation, 2nd edn. Springer (in press)
23. Conte R, Castelfranchi C (1995) Cognitive social action. UCL Press, London
24. Conte R, Paolucci M (2002) Reputation in artificial societies: social beliefs for social order. Kluwer, New York
25. Conte R, Turrini P (2006) Argyll-feet giants: a cognitive analysis of collective autonomy. Cogn Syst Res 7:209–219
26. Conway JH (1976) On numbers and games. Academic Press, London

27. Courtin C, Melot A-M (2005) Metacognitive development of deaf children: lessons from the appearance-reality and false belief tasks. J Deaf Stud Deaf Educ 5:266–276
28. Davis R, Shrobe H, Szolovits P (1993) What is a knowledge representation? AI Mag 14(1):17–33
29. Deci EL (1975) Intrinsic motivation. Plenum, New York
30. Deffuant G, Amblard F, Weisbuch G, Faure, T (2002) How can extremism prevail? A study based on the relative agreement interaction model. J Artif Soc Soc Simul 5(4). http://jasss.soc.surrey.ac.uk/5/4/1.html
31. Demazeau Y, Mueller JP (eds) (1990) Decentralized AI. Elsevier, North Holland
32. Elster J (1983) Sour grapes: studies in the subversion of rationality. Cambridge University Press, Cambridge
33. Epstein JM (2007) Generative social science. studies in agent-based computational modelling. Princeton University Press, Princeton
34. Fehr E, Fischbacher U (2003) The nature of human altruism. Nature 425:785–791
35. Fikes RE, Kehler T (1985) The role of frame-based representation in knowledge representation and reasoning. Commun ACM 28(9):904–920
36. Fogel DB (ed) (1998) Evolutionary computation: the fossil record. IEEE Press, New York
37. Fox CR, Tversky A (1995) Ambiguity aversion and comparative ignorance. Q J Econ 110(3):585–603
38. Frank MJ (2005) Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and non-medicated parkinsonism. J Cogn Neurosci 17:51–72
39. Frank RH (1987) If homo economicus could choose his own utility function, would he want one with a conscience? Amer Econ Rev 77(4):593–604
40. Franklin GF, Powell JD, Emami-Naeini A (1994) Feedback control of dynamic systems. Addison-Wesley, Reading
41. Franklin S, Graesser A (1997) Is it an agent, or just a program? A taxonomy for autonomous agents. In: Intelligent agents, vol III. Springer, Berlin, pp 21–35
42. Friedman M, Savage L J (1948) The utility analysis of choices involving risk. J Polit Econ 56(4):279–304
43. Fu W-T, Anderson JR (2006) From recurrent choice to skill learning: a reinforcement-learning model. J Exp Psychol Gen 135(2):184–206
44. Genesereth MR, Ketchpel SP (1994) Software agents. Commun ACM 37(7):48–53
45. Gigerenzer G, Selten R (2002) Bounded rationality. MIT Press, Cambridge. reprint edition
46. Goldberg DE (1989) Genetic algorithms in search, optimization and machine learning. Kluwer, Boston
47. Gray WD, Sims CR, Fu W-T, Schoelles MJ (2006) The soft constraints hypothesis: a rational analysis approach to resource allocation for interactive behavior. Psychol Rev 113(3):461–482
48. Green TP, Shapiro I (1994) Pathologies of rational choice theory: A critique of application in political science. Yale University Press, London
49. Hare B, Call J, Tomasello M (2001) Do chimpanzees know what conspecifics know and do not know? Anim Behav 61:139–151
50. Haykin S (1999) Neural networks: a comprehensive foundation. Prentice Hall, Englewood Cliffs
51. Hegselmann R, Krause U (2002) Opinion dynamics and bounded confidence: models, analysis and simulation. J Artif Soc Soc Simul 5(3). http://jasss.soc.surrey.ac.uk/5/3/2.html
52. Holland J (1989) Using classifier systems to study adaptive nonlinear networks. In: Stein DL (ed) Lectures in the sciences of complexity. Addison Wesley, New York
53. Holroyd CB, Coles MGH (2002) The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychol Rev 109:679–709
54. Kaelbling LP, Littman ML, Moore AW (1996) Reinforcement learning: a survey. J Artif Intell Res 4:237–285
55. Kahneman D (2003) Maps of bounded rationality: psychology for behavioral economics. Amer Econ Rev 93(5):1449–1475
56. Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. Econometrica XLVII:263–291
57. Kahneman D, Tversky A, Slovic P (eds) (1982) Judgement under uncertainty: heuristics and biases. Cambridge University Press, Cambridge
58. Katsuno H, Mendelzon AO (1991) On the difference between updating a knowledge base and revising it. In: Proc of the 2nd international conference on the principles of knowledge representation and reasoning (KR'91). Morgan Kaufmann, San Matteo, pp 387–394
59. Keynes JM (1921) Treatise on probability. MacMillan, London
60. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220(4598):671–680. http://www.cs.virginia.edu/cs432/documents/sa-1983.pdf; http://citeseer.ist.psu.edu/kirkpatrick83optimization.html
61. Knight FH (1921) Risk, uncertainty, and profit. In: Hart, Schaffner and Marx prize essays, vol 31. Houghton Mifflin, Boston, New York
62. Koza JR (1990) Genetic programming: a paradigm for genetically breeding populations of computer programs to solve problems. Stanford University Computer Science Department technical report STAN-CS-90-1314
63. Koza JR (1992) Genetic programming: on the programming of computers by means of natural selection. MIT Press, Cambridge
64. Koza JR (1994) Genetic programming II: automatic discovery of reusable programs. MIT Press, Cambridge
65. Koza JR, Bennett FH, Andre D, Keane MA (1999) Genetic programming III: darwinian invention and problem solving. Morgan Kaufmann, San Mateo
66. Koza JR, Keane MA, Streeter MJ, Mydlowec W, Yu J, Lanza G (2003) Genetic programming IV: routine human-competitive machine intelligence. Kluwer, Norwell
67. Kuznar LA, Frederick WG (2003) Environmental constraints and sigmoid utility: implications for value, risk sensitivity, and social status. Ecol Econ 46(2):293–306. September 2003
68. Latané B, Nowak A, Liu JH (1995) Distance matters: physical space and social impact. Pers Soc Psychol Bull 21:795–805
69. Lewin K (1921) Das Problem der Willensmessung und das Grundgesetz der Assoziation, vol I. Psychol Forsch 1:191–302
70. Lewis DK (1969) Convention: a philosophical study. Harvard University Press, Cambridge
71. Lorenz J (2007) Continuous opinion dynamics of multidimensional allocation problems under bounded confidence: More dimensions lead to better chances for consensus. Eur J Econ Soc Syst EJESS 19:213–227. Beliefs, Norms, and Markets

72. Lynch A (1999) Thought contagion: how belief spreads through society by. Basic Books, New York
73. Malarz K (2006) Truth seekers in opinion dynamics models. Int J Mod Phys C 17(10):1521–1524
74. Markman AB (1998) Knowledge representation. Lawrence Erlbaum Associates, Hillsdale
75. Mauss M (2006) The gift. The form and reason for exchange in archaic societies. Routledge, London
76. McFarland D (1995) Opportunity versus goals in robots, animals and people. In: Roitblat HL, Meyer J-A (eds) Comparative approaches to cognitive science. MIT Press, Cambridge, pp 415–433
77. Mill JS (1874) On the definition of political economy; and on the method of investigation proper to it. In: London and Westminster review. Essays on some unsettled questions of political economy, 2nd edn. Longmans, Green, Reader & Dyer, London. 1874, essay 5, paragraphs 38 and 48
78. Miller GA, Galanter E, Pribram KH (1960) Plans and the structures of behaviour. Holt, New York
79. Minsky M (1975) A framework for representing knowledge. In: Winston PH (ed) The psychology of computer vision. McGraw-Hill, New York
80. Nozick R (1993) The nature of rationality. Princeton University Press, Princeton
81. Nwana HS (1996) Software agents: an overview. Knowl Eng Rev 11(3):1–40. September 1996
82. Ostrom E (2000) Collective action and the evolution of social norms. J Econ Pers 14(3):137–158
83. Pearl J (1983) Heuristics: intelligent search strategies for computer problem solving. Addison-Wesley, New York. p vii
84. Perry J (1980) Belief and acceptance. Midwest Stud Philos V:533–542
85. Premack DG, Woodruff G (1978) Does the chimpanzee have a theory of mind? Behav Brain Sci 1:515–526
86. Rao AS, Georgeff MP (1995) BDI agents from theory to practice. In: Proc of the 1st international conference on multi agent systems ICMAS. AAAI Press, MIT Press, San Francisco
87. Ripley BD (1996) Pattern recognition and neural networks. Cambridge University Press, Cambridge
88. Russell SJ, Norvig P (2003) Artificial intelligence: a modern approach, 2nd edn. Prentice Hall, Upper Saddle River
89. Sacerdoti ED (1977) A structure for plans and behavior. Elsevier, New York
90. Salzarulo L (2006) A continuous opinion dynamics model based on the principle of meta-contrast. J Artif Soc Soc Simul 9(1). http://jasss.soc.surrey.ac.uk/9/1/13.html
91. Savage L J (1954) The foundations of statistics. Wiley, New York
92. Schank RC, Abelson RP (1977) Scripts, plans, goals, and understanding: an inquiry into human knowledge structures. Lawrence Erlbaum, Hillsdale
93. Shoham Y (1993) Agent-oriented programming. Artif Intell 60(1):51–92
94. Simon H (1957) A behavioral model of rational choice. In: Models of man, social and rational: mathematical essays on rational human behavior in a social setting. Wiley, New York
95. Sowa JF (2000) Knowledge representation: logical, philosophical, and computational foundations. Brooks/Cole, New York
96. Sun R (2002) Duality of the mind. Erlbaum, New York
97. Wilensky R (1983) Planning and understanding: a computational approach to human reasoning. Advanced book program. Addison-Wesley, Reading
98. Williamson O (1981) The economies of organization: the transaction cost approach. Amer J Sociol 87:548–577
99. Windrum P, Fagiolo G, Moneta A (2007) Empirical validation of agent based models: alternatives and prospects. JASSS J Artif Soc Soc Simul 10(2):8. http://jasss.soc.surrey.ac.uk/10/2/8.html
100. Wooldridge MJ (2000) Reasoning about rational agents. MIT Press, Cambridge
101. Wooldridge M, Jennings NR (1995) Intelligent agents: theory and practice. Knowl Eng Rev 10(2):115–152
102. Wright AH, Vose MD, Rowe JE (2003) Implicit parallelism. In: Cantu-Paz E et al (eds) Proc of GECCO (the genetic and evolutionary computation conference). Lecture notes in computer science. Springer, Berlin, pp 1003–1014

Reaction-Diffusion Computing

Andrew Adamatzky
University of the West of England, Bristol, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Computational Geometry
Logical Universality
Memory
Programmability
Robot Navigation and Massive Manipulation
Future Directions
Acknowledgments
Bibliography

Glossary

Belousov–Zhabotinsky (BZ) reaction is a chemical reaction where the organic substrate is oxidized by bromate ions in the presence of acid and a one electron transfer redox catalyst. The reaction produces oscillations in well-stirred reactors and traveling waves in thin layers.

Cellular automaton is an array of locally connected finite automata, which update their discrete states in discrete time depending on the states of their neighbors; all automata of the array update their states in parallel.

Collision-based computer is a uniform homogeneous medium which employs mobile compact patterns (particles, wave fragments) which travel in space and

perform computation (e. g. implement logical gates) when they collide with each other. Truth values of logical variables are given by either the absence or presence of a localization or by various types of localizations.

Excitable medium is spatially distributed assembly of coupled excitable systems; spatial distribution and coupling allow for propagation of excitation waves.

Excitable system is a system with a single steady quiescent state that is stable to small perturbations, but responds with an excursion from its quiescent state (excitation event) if the perturbation is above a critical threshold level. After excitation the system enters a refractory period during which time it is insensitive to further excitation before returning to its steady state.

Glider as related to cellular automata, is a compact (neither infinitely expanding nor collapsing) pattern of non-quiescent states that travels along the cellular-automaton lattice.

Image processing is a transformation of an input image to an output image with desirable properties, using manipulation of images to enhance or extract information.

Logical gate is an elementary building block of a digital, or logical, circuit, which represents (mostly) binary logical operations, e. g. AND, OR, XOR, with two input terminals and one output terminal. In Boolean logic terminals are in one of two binary conditions (e. g. low voltage and high voltage) corresponding to TRUE and FALSE values of logical variables.

Logically universal processor is a system which can realize a functionally complete set of logical operations in its development, e. g. conjunction and negation.

Oregonator is a system of three (or two in a modified version) coupled differential equations aimed to simulate oscillatory phenomena in the **Belousov–Zhabotinsky reaction**.

Shortest-path problem is the problem of finding a path between two sites (e. g. vertices of the graph, locations in space) such that the length (sum of the weights of the graph edges or travel distances) of the path is minimized.

Skeleton of a planar contour is a set of centers of bi-tangent circles lying inside the contour.

Subexcitable medium is a medium whose steady state lies between the excitable and the unexcitable domains. In excitable media waves initiated by perturbations of a sufficient size propagate throughout the media. In an unexcitable medium no perturbation is large enough to trigger a wave. In a subexcitable medium wave fragments with open ends are formed.

Voronoi diagram of a planar set **P** of planar points is a partition of the plane into regions, each for any element of **P**, such that a region corresponding to a unique point p contains all those points of the plane that are closer to p than to any other node of **P**.

Wave fragment is an excitation wave formed in a subexcitable medium; this is a segment with free ends, which either expand or contract, depending on their size and the medium's excitability.

Definition of the Subject

A reaction-diffusion computer is a spatially extended chemical system, which processes information using interacting growing patterns, excitable and diffusive waves. In reaction-diffusion processors, both the data and the results of the computation are encoded as concentration profiles of the reagents. The computation is performed via the spreading and interaction of wave fronts.

A reaction-diffusion computer is a thin layer of a reagent mixture which reacts to changes of one reagent's concentration—data configuration—in a predictable way to form a stationary pattern corresponding to the concentration of the reagent—result configuration. A computation in the chemical processor is implemented via the spreading and interaction of diffusive or phase waves [1,7].

The reaction-diffusion computers are parallel because myriads of their micro-volumes update their states simultaneously, and molecules diffuse and react in parallel. Liquid-phase chemical media are wet-analogs of massive-parallel (millions of elementary processors in a small chemical reactor) and locally connected (every micro-volume of the medium changes its state depending on states of its closest neighbors) processors. They have parallel input and outputs, e. g. optical input—control of initial excitation dynamics by illumination masks, output is parallel because the concentration profile representing the results of computation is visualized by indicators. The reaction-diffusion computers show fault-tolerance and are capable of automatic reconfiguration, namely if we remove some quantity of the computing substrate, the topology is restored almost immediately.

Introduction

Reaction-diffusion computers are based on three principles of physics-inspired computing. First, physical action measures amount of information: we exploit active processes in non-linear systems and interpret dynamics of the systems as computation. Second, physical information travels only a finite distance: this means that computation is local and we can assume the non-linear medium is

a spatial arrangement of elementary processing units connected locally, i. e. each unit interacts with closest neighbors. Third, nature is governed by waves and spreading patterns: computation is therefore spatial.

Surface tension, propagating waves, electricity and chemical reactions have been principal "engines" of nature-inspired computers for over two centuries [1]. However, only reaction-diffusion computers utilize all these phenomena at once, to solve large-scale NP-complete problems in a parallel and stable manner. Experimental studies and designs of reaction-diffusion computers could be traced back to the pioneer discovery of Kuhnert [28]. In 1986 he demonstrated that some very basic image transformations can be implemented in a light-sensitive Belousov–Zhabotinsky system [28]. The ideas by Kuhnert, Krinsky and Agladze [29] on image and planar shape transformations in two-dimensional excitable chemical medium were further developed and modified by Rambidi with colleagues, see e. g. [42,43]. At that time, mid and late 1990s, a range of chemical logical gates was experimentally built in the Showalter [52] and Yoshikawa laboratories [30]. Computation of shorter, one of classical optimization problems, has also been implemented in these laboratories using the Belousov–Zhabotinsky medium [7,10,51].

Reaction-diffusion computers give us the best examples of unconventional computers, they feature, following Jonathan Mills' classification of conventional vs. unconventional [34]: wet-ware, non-silicon computing substrate; parallel processing; computation occurring everywhere in substrate space; computation is based on analogies; spatial increase in precision; holistic and spatial programming; visual structure; implicit error correcting. A theory of reaction-diffusion computing was established and a range of practical applications outlined in [1]. Recent discoveries were published in a collective monograph [7].

Herein we will provide an account of achievements in reaction-computing obtained in the latter 1990s and early 2000s. Our present article in no way serves as a substitute for these books but rather an introduction to the field and case study of several characteristic examples. We give a case-study introduction to the novel paradigm of wave-based computing in chemical systems. We show how selected problems and tasks of computational geometry, robotics and logics can be solved by encoding data in configuration of chemical medium's disturbances and programming wave dynamics and interaction. We describe and analyze working reaction-diffusion algorithms for image processing, computational geometry, logical and arithmetical circuits, memory devices, path planning and robot navigation, and control of massive parallel actuators.

Computational Geometry

The Voronoi diagram, or plane subdivision, is our most favorable NP-complete problem, for demonstrating "mechanics" and computational complexity of reaction-diffusion computers.

Let $\mathbf{P}$ be a nonempty finite set of planar points. A planar Voronoi diagram of the set $\mathbf{P}$ is a partition of the plane onto such regions, that for any element of $\mathbf{P}$, a region corresponding to a unique point p contains all those points of the plane which are closer to p than to any other node of $\mathbf{P}$. A unique region $\mathrm{vor}(p) = \{z \in \mathbf{R}^2 : d(p,z) < d(p,m) \quad \forall\ m \in \mathbf{R}^2,\ m \neq z\}$ assigned to point p is called a Voronoi cell of the point p. The boundary of the Voronoi cell of a point p is built of segments of bisectors separating pairs of geographically closest points of the given planar set $\mathbf{P}$. A union of all boundaries of the Voronoi cells determines the *planar Voronoi diagram*: $\mathrm{VD}(\mathbf{P}) = \cup_{p\in\mathbf{P}}\partial\mathrm{vor}(p)$. A variety of Voronoi diagrams and algorithms of their construction can be found in [27].

The basic concept of constructing Voronoi diagrams with reaction-diffusion systems is based on a very simple intuitive technique for detecting the bisector points separating two given points of the set $\mathbf{P}$. If we drop reagents at the two data points the diffusive waves, or phase waves if the computing substrate is active, spread outwards from the drops with the same speed. The waves travel the same distance from the sites of origination before they meet one another. The points, where the waves meet, are the bisector points. This idea of a Voronoi diagram computation was originally implemented in cellular automata models and in experimental parallel chemical processors, see the extensive bibliography in [1,7].

Assuming that the computational space is homogeneous and locally connected, and every site (micro-volume of the chemical medium or cell of the automaton array) is coupled to its closest neighbors by the same diffusive links, we can easily draw a parallel between distance and time, and thus put our wave-based approach into action. In a cellular-automaton representation of the physical reality cell neighborhood u determines that all processes in the cellular automata model are constrained to the discrete metric L_∞. So, when studying automata models we should think rather about the discrete Voronoi diagram than its Euclidean representation. Chemical laboratory prototypes of reaction-diffusion computers do approximate a continuous Voronoi diagram as we will see further.

A discrete Voronoi diagram can be defined on lattices or arrays of cells where, e. g. a two-dimensional lattice $\mathbf{Z}^2$. The distance $d(\cdot,\cdot)$ is calculated not in Euclidean but in one of the discrete metrics, e. g. L_1 and L_∞. A dis-

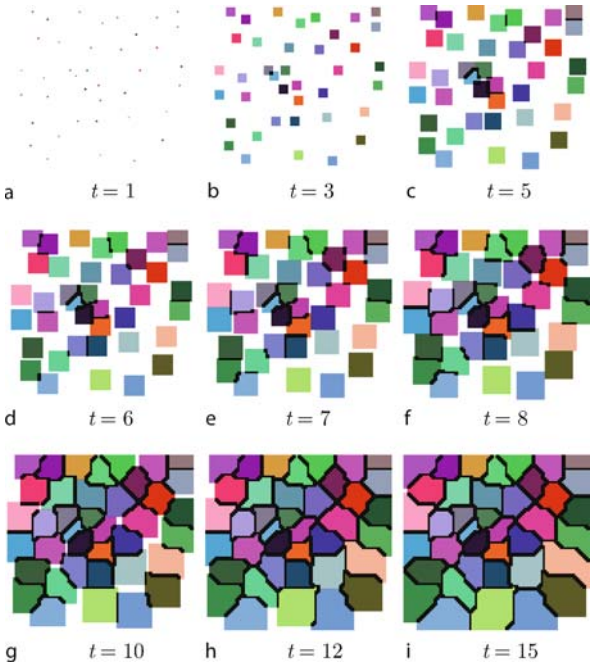

Reaction-Diffusion Computing, Figure 1
Computation of a Voronoi diagram in a cellular-automaton model of a chemical processor with *O*(*n*) reagents. Precipitate is shown in *black*

crete bisector of nodes x and y of $\mathbf{Z}^2$ is determined as $B(x,y) = \{z \in \mathbf{Z}^2 \colon d(x,z) = d(y,z)\}$. However, following such definition we sometimes generate bisectors that fill a quarter of the lattices or produce no bisector at all [1]. If we want the constructed diagrams to be closer to the real-world then we could re-define the discrete bisector as follows $B(x,y) = \{z \in \mathbf{Z}^2 \colon |d(x,z) - d(y,z)| \leq 1\}$. The redefined bisector will comprise edges of Voronoi diagrams constructed in discrete, cellular-automaton models of reaction-diffusion and excitable media.

Now we will discuss several versions of reaction-diffusion wave-based constructions of Voronoi diagrams, from a naïve model, where the number of reagents grow proportionally to the number of data points, to a minimalist implementation with just one reagent and one substrate [1].

Let us start with the $O(n)$-reagent model. In a naïve version of reaction-diffusion computation of a Voronoi diagram one needs two reagents and a precipitate to mark a bisector separating two points. Therefore $n+2$ reagents, including precipitate and substrate, are required to approximate a Voronoi diagram of n points. We place n unique reagents on n points of the given data set $\mathbf{P}$, waves of these reagents spread around the space and interact with each other where they meet. When at least two different reagents meet at the same or adjacent sites of the space, they react and form a precipitate—sites that contain the precipitate represent edges of the Voronoi cell, and therefore constitute the Voronoi diagram. In "chemical reaction" equations the idea looks as follows, α and β are different reagents and # is a precipitate: $\alpha + \beta \rightarrow \#$. This can be converted to a cellular-automaton interpretation as follows:

$$x^{t+1} = \begin{cases} \rho\,, & \text{if} \quad x^t = \bullet \quad \text{and} \quad \Psi(x)^t \subset \{\rho, \bullet\} \\ \#\,, & \text{if} \quad x^t \neq \# \quad \text{and} \quad |\Psi(x)^t/\#| > 1 \\ x^t\,, & \text{otherwise} \end{cases}$$

where $\bullet$ is a resting state (cell in this state does not contain any reagents), $\rho \in \mathbf{R}$ is a reagent from the set $\mathbf{R}$ of n reagents, and $\Psi(x)^t = \{y^t \colon y \in u(x)\}$ characterizes what reagents are present in the local neighborhood $u(x)$ of the cell x at time step t.

The first transition of the above rule symbolizes diffusion. A resting cell takes the state ρ if only this reagent is present in the cell's neighborhood. If there are two different reagents in the cells neighborhood then the cell takes the precipitate state #. Diffusing reagents halt because formation of precipitate reduces the number of "vacant" resting cells. The precipitate does not diffuse. Cell in state # remain in this state indefinitely. An example of a cellular-automaton simulation of $O(n)$-reagent chemical processor is shown in Fig. 3.

The $O(n)$-reagent model is demonstrative, however, computationally inefficient. Clearly we can reduce the number of reagents to four—using map coloring theorems—but pre-processing time will be unfeasibly high. The number of participating reagents can be sufficiently reduced to $O(1)$ when the topology of the spreading waves is taken into account [1].

Now we go from one extreme to another and consider a model with just one reagent, and a substrate. The reagent α diffuses from sites corresponding to points of a planar data set $\mathbf{P}$. When two diffusing wave fronts meet a super-threshold concentration of reagents they do not spread further. A cellular-automaton model represents this as follows.

Every cell has two possible states: resting or substrate state $\bullet$ and reagent state α. If the cell is in state α it remains in this state indefinitely. If the cell is in state $\bullet$ and between one and four of its neighbors are in state α, then the cell takes the state α. Otherwise, the cell remains in the state $\bullet$—this reflects the "super-threshold inhibition" idea. A cell state transition rule is as follows:

$$x^{t+1} = \begin{cases} \alpha\,, & \text{if} \quad x^t = \bullet \quad \text{and} \quad 1 \leq \sigma(x)^t \leq 4 \\ x^t\,, & \text{otherwise} \end{cases}$$

where $\sigma(x)^t = |y \in u(x) \colon y^t = \alpha|$.

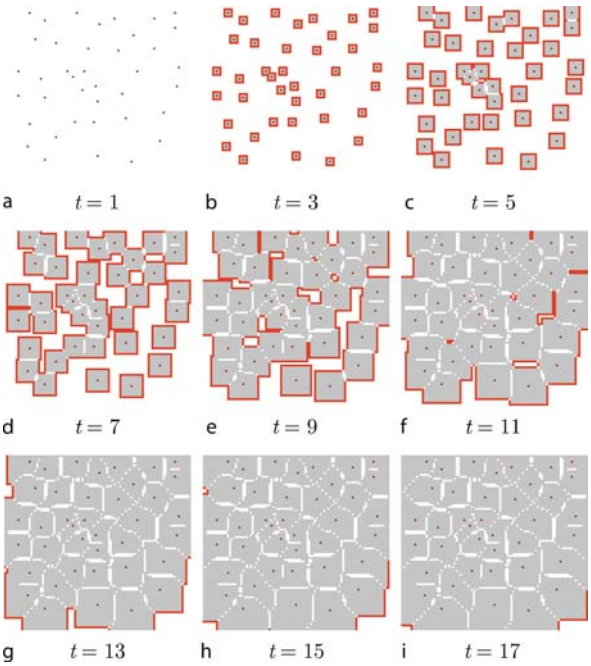

Reaction-Diffusion Computing, Figure 2
An example of Voronoi diagram computing in an automaton model of reaction-diffusion medium with one reagent and one substrate. Reactive parts of wave fronts are shown in *black*. The precipitate is *gray* and the edges of the Voronoi diagram are *white*

Increasing the number of reagents to two (one reagent and one precipitate) would make life easy. A reagent β diffuses on a substrate, from the initial points (drop of reagent) of **P**, and forms a precipitate in the reaction $m\beta \to \alpha$, where $1 \le m \le 4$.

Every cell takes three states: $\bullet$ (resting cell, no reagents), α (e. g. colored precipitate) and β (reagent). The cell updates its states by the rule:

$$x^{t+1} = \begin{cases} \beta\,, & \text{if} \quad x^t = \bullet \quad \text{and} \quad 1 \le \sigma(x)^t \le 4 \\ \alpha\,, & \text{if} \quad x^t = \beta \quad \text{and} \quad 1 \le \sigma(x)^t \le 4 \\ x^t\,, & \text{otherwise} \end{cases}$$

where $\sigma(x)^t = |y \in u(x)\colon y^t = \beta|$.

An example of a Voronoi diagram computed in an automaton model of a reaction-diffusion medium with one reagent and one substrate is shown in Fig. 2.

By increasing the number of cell-state and enlarging cell neighborhood in the cellular automaton model we can produce more realistic—almost perfectly matching outcomes of chemical laboratory experiments—Voronoi diagrams (Fig. 3).

Let us consider the following model. Cells of the automaton take their state from interval $[\rho, \alpha]$, where ρ is a minimum refractory value, and α is a maximum excitation value; $\rho = -2$ and $\alpha = 5$ in our experiments. Cell x's state transitions are strongly determined by normalized local excitation $\sigma_x^t = \sum_{y \in u_x} y^t / \sqrt{(|u_x|)}$. Every cell x updates its state at time $t+1$, depending on its state x^t and state u_x^t of its neighborhood u_x—in experiments we used 15×15 cell neighborhood—as follows:

$$x^{t+1} = \begin{cases} \alpha\,, & \text{if} \quad x^t = 0 \quad \text{and} \quad \sigma_x^t \ge \alpha \\ 0, & \text{if} \quad x^t = 0 \quad \text{and} \quad \sigma_x^t < \alpha \\ x^t + 1\,, & \text{if} \quad x^t < 0 \\ x^t - 1\,, & \text{if} \quad x^t > 1 \\ \rho, & \text{if} \quad x^t = 1\,. \end{cases}$$

This rule represents spreading of "excitation", or simply phase wave-fronts, in computational space, interaction and annihilation of the wave-fronts. To allow the reaction-diffusion computer to "memorize" sites of wave collision we add a precipitate state p_x^t. Concentration p_x^t of precipitate at site x at moment t is calculated as $p_x^{t+1} \sim |\{y \in u_x\colon y^t = \alpha\}|$.

As shown in Fig. 4 the model represents cellular-automaton Voronoi diagrams in "unlike phase" with experimental chemical representation of the diagram. Sites of higher concentration of precipitate in cellular-automaton configurations correspond to sites with lowest precipitate concentration in experimental processors.

Logical Universality

Certain families of thin-layer reaction-diffusion chemical media can implement sensible transformation of initial (data) spatial distribution of chemical species concentrations to a final (result) concentration profile [1,50]. In these reaction-diffusion computers a computation is realized via spreading and interaction of diffusive or phase waves. Specialized, intended to solve a particular problem, experimental chemical processors implement basic operations of image processing [7,29,42,43], computation of optimal paths [7,10,51] and control of mobile robots [7].

A device is called computationally universal if it implements a functionally complete system of logical gates, e. g. a tuple of negation and conjunction, in its space-time dynamics. A number of computationally universal reaction-diffusion devices were implemented, the findings include logical gates [49,52] and diodes [18,30,35] in the Belousov–Zhabotinsky (BZ) medium, and the XOR gate in the palladium processor [4].

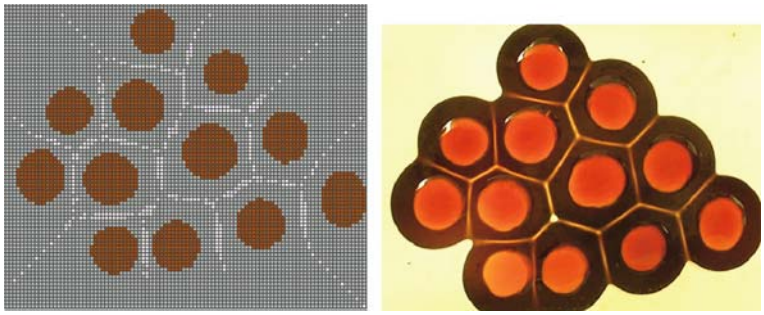

Reaction-Diffusion Computing, Figure 3
Planar Voronoi diagram computed in cellular automaton (*left*) and palladium reaction-diffusion chemical processor (*right*) [7]

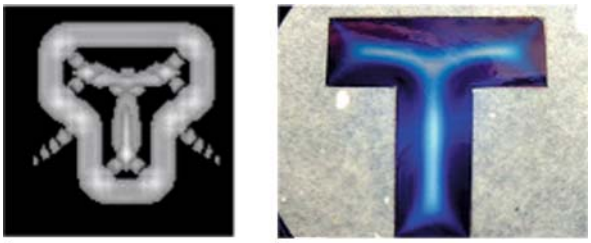

a b

Reaction-Diffusion Computing, Figure 4
Skeleton (internal Voronoi diagram) of planar T-shape constructed in multi-state cellular-automaton model (a) and chemical laboratory Prussian blue reaction-diffusion processor (b) [11]

The most known so far experimental prototypes of reaction-diffusion processors exploit interaction of wave fronts in a geometrically constrained chemical medium, i. e. the computation is based on a stationary architecture of a medium's inhomogeneities. Constrained by stationary wires and gates reaction-diffusion chemical universal processors pose a little computational novelty and no dynamical reconfiguration ability because they simply imitate architectures of conventional silicon computing devices. To appreciate in full massive-parallelism of thin-layer chemical media and to free the chemical processors from limitations of fixed computing architectures we adopt an unconventional paradigm of architecture-less, or collision-based, computing. An architecture-based, or stationary, computation implies that a logical circuit is embedded into the system in such a manner that all elements of the circuit are represented by the system's stationary states. The architecture is static. If there is any kind of "artificial" or "natural" compartmentalization the medium is classified as an architecture-based computing device. Personal computers, living neural networks, cells, and networks of chemical reactors are typical examples of architecture-based computers.

A collision-based, or dynamical, computation employs mobile compact finite patterns, mobile self-localized excitations or simply localizations, in active non-linear medium. Essentials of collision-based computing are as follows. Information values (e. g. truth values of logical variables) are given by either absence or presence of the localizations or other parameters of the localizations. The localizations travel in space and do computation when they collide with each other. There are no predetermined stationary wires, a trajectory of the traveling pattern is a momentary wire. Almost any part of the medium space can be used as a wire. Localizations can collide anywhere within a space sample, there are no fixed positions at which specific operations occur, nor location specified gates with fixed operations. The localizations undergo transformations, form bound states, annihilate or fuse when they interact with other mobile patterns. Information values of localizations are transformed as a result of collision and thus a computation is implemented [2].

The paradigm of collision-based computing originates from the technique of proving computational universality of the Game of Life [14], conservative logic and billiard-ball model [21] and their cellular-automaton implementations [32].

Solitons, defects in tubulin microtubules, excitons in Scheibe aggregates and breather in polymer chains are the most frequently considered candidates for a role of information carrier in nature-inspired collision-based computers, see overview in [1]. It is experimentally difficult to reproduce all these artifacts in natural systems, therefore the existence of mobile localizations in an experiment-friendly chemical media would open new horizons for fabrication of collision-based computers.

The basis for material implementation of collision-based universality of reaction-diffusion chemical media was discovered by Sendiña-Nadal et al. [48]. They experimentally proved the existence of localized excita-

tions—traveling wave fragments which behave like quasi-particles—in a photosensitive sub-excitable Belousov–Zhabotinsky medium.

We show how logical circuits can be fabricated in a sub-excitable BZ medium via collisions between traveling wave fragments. While implementation collision-based logical operations is relatively straightforward [7], more attention should be paid to control of signal propagation in the homogeneous medium. It has been demonstrated that applying light of varying intensity we can control the excitation dynamic in a Belousov–Zhabotinsky medium [13,23,39], wave velocity [46], and pattern formation [55]. Of particular interest are experimental evidences of light-induced back propagating waves, wave-front splitting and phase shifting [63]; we can also manipulate a medium's excitability by varying the intensity of the medium's illumination [16]. On the basis of these facts we show how to control signal-wave fragments by varying geometric configuration of excitatory and inhibitory segments of impurity-reflectors.

We built our model on a two-variable Oregonator equation [19,53] adapted to a light-sensitive BZ reaction with applied illumination [13]:

$$\frac{\partial u}{\partial t} = \frac{1}{\epsilon}\left(u - u^2 - (fv + \phi)\frac{u - q}{u + q}\right) + D_u \nabla^2 u$$
$$\frac{\partial v}{\partial t} = u - v$$

where variables u and v represent local concentrations of bromous acid $HBrO_2$ and the oxidized form of the catalyst ruthenium Ru(III), ϵ sets up a ratio of time scale of variables u and v, q is a scaling parameter depending on reaction rates, f is a stoichiometric coefficient, ϕ is a light-induced bromide production rate proportional to intensity of illumination (an excitability parameter—moderate intensity of light will facilitate the excitation process, a higher intensity will produce excessive quantities of bromide which suppresses the reaction). We assumed that the catalyst is immobilized in a thin-layer of gel, therefore there is no diffusion term for v. To integrate the system we used the Euler method with a five-node Laplasian operator, time step $\Delta\ t = 10^{-3}$ and grid point spacing $\Delta\ x = 0.15$, with the following parameters: $\phi = \phi_0 + A/2$, $A = 0.0011109$, $\phi_0 = 0.0766$, $\epsilon = 0.03$, $f = 1.4$, $q = 0.002$.

The chosen parameters correspond to a region of "higher excitability of the sub-excitability regime" outlined in [48] (see also how to adjust f and q in [41]) that supports propagation of sustained wave fragments (Fig. 5a). These wave fragments are used as quanta of information in our design of collision-based logical circuits. The waves

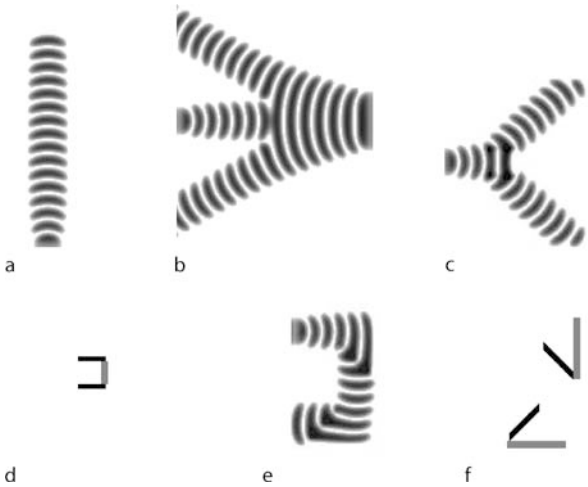

Reaction-Diffusion Computing, Figure 5
Basic operations with signals. Overlay of images taken every 0.5 time units. Exciting domains of impurities are shown in black, inhibiting domains of impurities are shown in gray. a Wave fragment traveling north. b Signal branching without impurities: a wave fragment traveling east splits into two wave fragments (traveling south-east and north-east) when it collides with a smaller wave fragment traveling west. c Signal branching with impurity: wave fragment traveling west is split by impurity (d) into two waves traveling north-west and south-west. e Signal routing (U-turn) with impurities: wave fragment traveling east is routed north and then west by two impurities (f). An impurity-reflector consists of inhibitory (*gray*) and excitatory (*black*) chains of grid sites

were initiated by locally disturbing initial concentrations of species, e. g. ten grid sites in a chain are given the value $u = 1.0$ each, this generated two or more localized wave fragments, similarly to counter-propagating waves induced by temporary illumination in experiments [63]. The traveling wave fragments keep their shape for around $4 \cdot 10^3$–10^4 steps of simulation (4–10 time units), then decrease in size and vanish. The wave's life-time is sufficient however to implement logical gates; this also allows us not to worry about "garbage collection" in the computational medium.

We model signals by traveling wave fragments [13,48]: a relatively stable propagating wave fragment (Fig. 5a) represents a TRUE value of a logical variable corresponding to the wave's trajectory (momentarily wire).

To demonstrate that a physical system is logically universal it is enough to implement negation and conjunction or disjunction in space-time dynamics of the system. To realize a fully functional logical circuit we must also know how to operate input and output signals in the system's dynamics, namely to implement signal branching and routing; delay can be realized via appropriate routing.

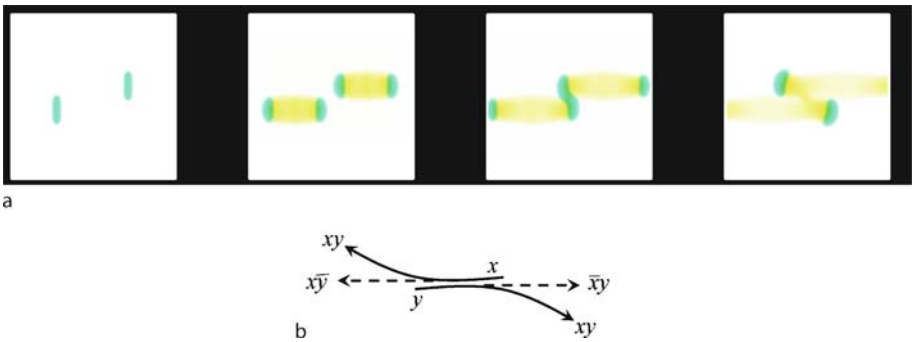

Reaction-Diffusion Computing, Figure 6
Implementation of conservative gate in Belousov–Zhabotinsky system. a Elastic collision of two wave-fragments, one traveling West another East. The fragments change the direction of their motion to North-West and South-East, respectively, as a result of the collision. b Scheme of the gate. In a logical variables are represented as $x = 1$ and $y = 1$

We can branch a signal using two techniques. Firstly, we can collide a smaller auxiliary wave to a wave fragment representing the signal, the signal-wave will split then into two signals (these daughter waves shrink slightly down to stable size and then travel with constant shape further $4 \cdot 10^3$ time steps of the simulation) and the auxiliary wave will annihilate (Fig. 5b).

Secondly, we can temporarily and locally apply illumination impurities on a signal's way to change properties of the medium and thus cause the signal to split (Fig. 5c,d). We must mention, it was already demonstrated in [63], that a wave front influenced by strong illumination (inhibitory segments of the impurity) splits and its ends do not form spirals, as in typical situations of excitable media.

A control impurity, or reflector, consists of a few segments of sites where the illumination level is slightly above or below the overall illumination level of the medium. Combining excitatory and inhibitory segments we can precisely control the wave's trajectory, e. g. realize a U-turn of a signal (Fig. 5e,f).

A typical billiard-ball model interaction gate [21,32] has two inputs—x and y, and four outputs—$x\bar{y}$ (ball x moves undisturbed in the absence of ball y), $\bar{x}y$ (ball y moves undisturbed in the absence of ball x), and twice xy (balls x and y change their trajectories when they collide with each other). Such a conservative interaction gate can be implemented via elastic collision of the wave-fragment, see Fig. 6.

The elastic collision is not particularly common in laboratory prototypes of chemical systems, more often interacting waves either fuse or one of the waves annihilates as a result of the collision with another wave. This leads to a non-conservative version of the interaction gate with two inputs and three outputs, i. e. just one xy output instead of two. Such a collision gate is shown in Fig. 7.

Reaction-Diffusion Computing, Figure 7
Two wave fragments undergo angle collision and implement interaction gate $\langle x, y\rangle \rightarrow \langle x\bar{y}, xy, \bar{x}y\rangle$. a In this example $x = 1$ and $y = 1$, both wave fragments are present initially. Overlay of images taken every 0.5 time units. b Scheme of the gate. In *upper left* and *bottom left* corners of a we see domains of wave generation, two echo wave fragments are also generated, they travel outwards from the gate area and thus do not interfere with computation

The rich dynamic of a sub-excitable Belousov–Zhabotinsky medium allows us also to implement complicated logical operations just in a single interaction event, for details see [7].

Memory

Memory in chemical computers can be represented in several ways as follows. In precipitating systems any site with precipitate is a memory element. However, they are not rewritable. In "classical" excitable chemical systems, like Belousov–Zhabotinsky dynamics one can construct memory as a configuration of sources of spiral or target waves. We used this technique to program movement of a wheeled robot controlled by an on-board chemical reactor with the Belousov–Zhabotinsky system [7]. The method has the same drawback as precipitating memory—as soon as reaction space is divided by spiral or target waves, it is quite

difficult if not impossible to sensibly move the source of the waves. This is only possible with external inhibition or complete reset of the medium.

In geometrically constrained excitable chemical medium, as demonstrated in [36], we can employ old-time techniques of storing information in induction coils and other types of electrical circuits, i. e. dynamical memory. A ring with an input channel is prepared from the reaction substrate. The ring is broken by a small gap and the input is also separated from the ring with a gap of similar width [36], the gaps play the role of one-way gates to prevent excitation from spreading backwards. The waves enter the ring via the input channel and travel along the ring "indefinitely" (as long as the substrate lasts) [36]. The approach aims to split the reaction-diffusion system into many compartments, and thus does not fit our paradigm of computing in a uniform medium.

In our search for real-life chemical systems exhibiting both mobile and stationary localizations we discovered a cellular-automaton model [58] of an abstract activator-inhibitor reaction-diffusion system, which ideally fits the framework of the collision-based computing paradigm and reaction-diffusion computing. The phenomenology of the automaton was discussed in detail in our previous work [58], therefore in the present paper we draw together the computational properties of the reaction-diffusion cellular hexagonal automaton. The automaton imitates spatio-temporal dynamics of the following reaction equations:

$$\begin{array}{rrr} A + 6S \to A & A + I \to I & A + 3I \to I \\ & A + 2I \to S & 2A \to I \\ & 3A \to A & \beta A \to I \\ & & I \to S\,. \end{array}$$

Each cell of the automaton takes three states – substrate S, activator A and inhibitor I. Adopting formalism from [9], we represent the cell-state transition rule as a matrix $\mathbf{M} = (m_{ij})$, where $0 \le i \le j \le 7, 0 \le i + j \le 7$, and $m_{ij} \in \{I, A, S\}$. The output state of each neighborhood is given by the row-index i, the number of neighbors in cell-state I, and column-index j (the number of neighbors in cell-state A). We do not have to count the number of neighbors in cell-state S, because it is given by $7 - (i + j)$. A cell with a neighborhood represented by indexes i and j will update to cell-state M_{ij} which can be read off the matrix. In terms of the cell-state transition function this can be presented as follows: $x^{t+1} = M_{\sigma_2(x)^t \sigma_1(x)^t}$, where $\sigma_i(x)^t$ is a sum of cell x's neighbors in state i, $i = 1, 2$, at time step t. The exact matrix structure, which corresponds to matrix M_3 in [58], is as follows:

$$M = \left\{ \begin{array}{cccccccc} S & A & I & A & I & I & I & I \\ S & I & I & A & I & I & I & \\ S & S & I & A & I & I & & \\ S & I & I & A & I & & & \\ S & S & I & A & & & & \\ S & S & I & & & & & \\ S & S & & & & & & \\ S & & & & & & & \\ & & & & & & & \end{array} \right\}.$$

The cell-state transition rule reflects the nonlinearity of activator-inhibitor interactions for sub-threshold concentrations of the activator. Namely, for a small concentration of the inhibitor and for threshold concentrations, the activator is suppressed by the inhibitor, while for critical concentrations of the inhibitor both inhibitor and activator dissociate producing the substrate. In exact words, $M_{01} = A$ symbolizes the diffusion of activator A, $M_{11} = I$ represents the suppression of activator A by the inhibitor I, $M_{z2} = I$ $(z = 0, \ldots, 5)$ can be interpreted as self-inhibition of the activator in particular concentrations. $M_{z3} = A$ $(z = 0, \ldots, 4)$ means a sustained excitation under particular concentrations of the activator. $M_{z0} = S$ $(z = 1, \ldots, 7)$ means that the inhibitor is dissociated in absence of the activator, and that the activator does not diffuse in sub-threshold concentrations. And, finally, $M_{zp} = I$, $p \ge 4$ is an upper-threshold self-inhibition.

Amongst non-trivial localizations, see full "catalog" in [5], found in the medium we selected eaters gliders G_4 and G_{34}, mobile localizations with activator head and inhibitor tail, and eaters E_6, stationary localizations transforming gliders colliding with them, as components of the memory unit.

The eater E_6 can play the role of a 6-bit flip-flop memory device. The substrate-sites (bit-down) between inhibitor-sites (Fig. 8) can be switched to an inhibitor-

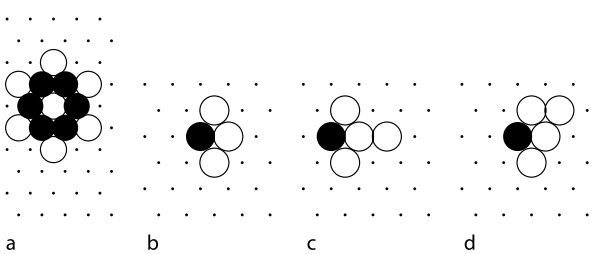

Reaction-Diffusion Computing, Figure 8
Localizations in reaction-diffusion hexagonal cellular-automaton. Cells with inhibitor *I* are empty circles, cells with activator *A* are black disks. a Stationary localization eater E_6, b–c two forms of glider G_{34}, d glider G_4 [5]

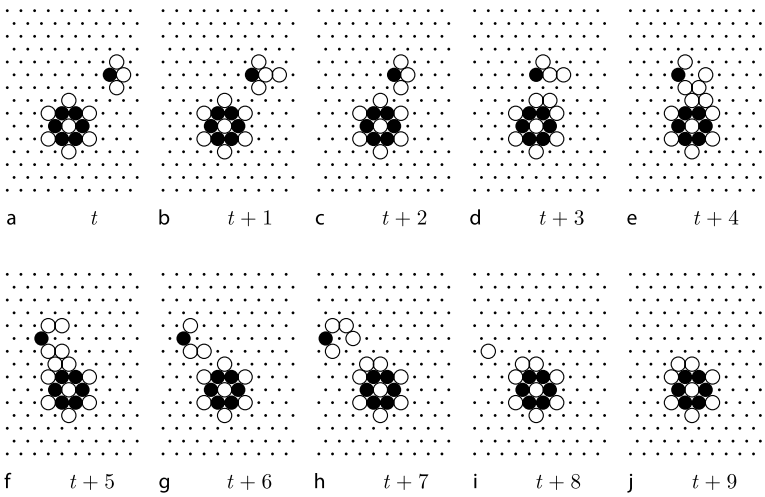

Reaction-Diffusion Computing, Figure 9
Write bit [5]

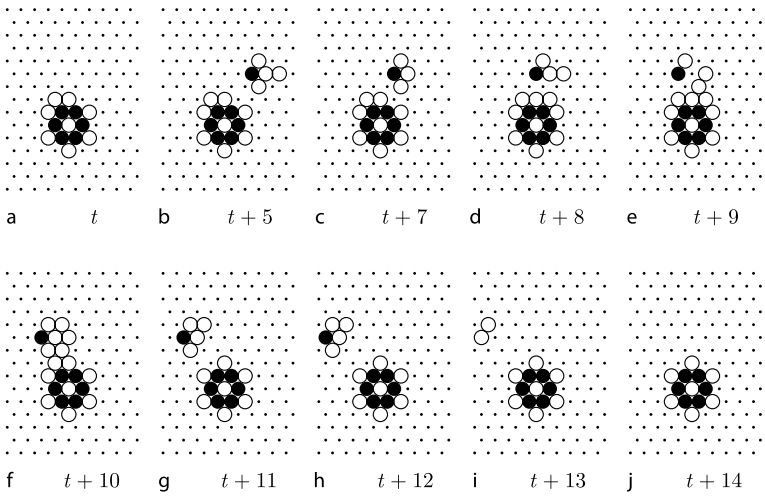

Reaction-Diffusion Computing, Figure 10
Read and erase bit [5]

state (bit-up) by a colliding glider. An example of writing one bit of information in E_6 is shown in Fig. 9. Initially E_6 stores no information. We aim to write one bit in the substrate-site between the northern and north-western inhibitor-sites (Fig. 9a). We generate a glider G_{34} (Fig. 9b,c) traveling West. G_{34} collides with (or brushes past) the North edge of E_6 resulting in G_{34} being transformed to a different type of glider, G_4 (Fig. 9g,h). There is now a record of the collision—evidence that writing was successful. The structure of E_6 now has one site (between the northern and north-western inhibitor-sites) changed to an inhibitor-state (Fig. 9j) —a bit was saved [5].

To read a bit from the E_6 memory device with one bit-up (Fig. 10a), we collide (or brush past) with glider G_{34} (Fig. 10b). Following the collision, the glider G_{34} is transformed into a different type of basic glider, G_{34} (Fig. 10g), and the bit is erased (Fig. 10j).

Programmability

In the chemical laboratory the term programmabilty means controllability. How real chemical systems can be controlled? The majority of the literature, related to theoretical and experimental studies concerning the controllability of the reaction-diffusion medium, deals with application of an electric field. For example, in a thin-layer Belousov–Zhabotinsky reactor stimulated by an electric field the following phenomena are observed: the velocity of ex-

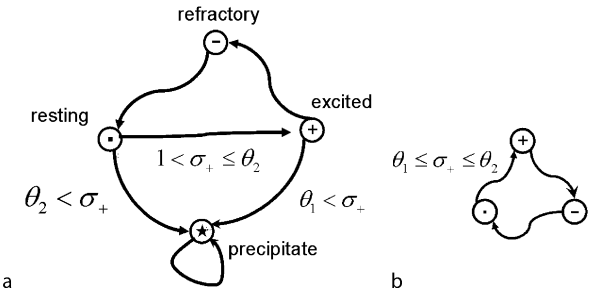

Reaction-Diffusion Computing, Figure 11
Cell state transition diagrams: a model of precipitating reaction-diffusion medium, b model of excitable system

citation waves is increased by a negative and decreased by a positive electric field, a wave is split into two waves that move in opposite directions if a very high electric field is applied across the evolving medium, crescent waves are formed not commonly observed in the field absent evolution of the BZ reaction, stabilization and destabilization of wave fronts, see [7].

Other control parameters may include temperature (to e. g. program transitions between periodic and chaotic oscillations), substrate's structure (controlling formation, annihilation and propagation of waves), and illumination (inputting data and routing signals in light-sensitive chemical systems).

Let us demonstrate a concept of control-based programmability in models of reaction-diffusion processors. Firstly, we show how to adjust reaction rates in chemical medium to make it perform a computation of a Voronoi diagram over a set of given points. Secondly, we show how to switch an excitable system between specialized-processor and universal-processor modes, see [7] for additional examples and details.

Let a cell x of a two-dimensional lattice take four states: resting $\circ$, excited $(+)$, refractory $(-)$ and precipitate $\bigstar$, and update their states in discrete time t depending on the number $\sigma^{t(x)}$ of excited neighbors in its eight-cell neighborhood as follows (Fig. 11a).

A resting cell x becomes excited if $0 < \sigma^t(x) \le \theta_2$ and precipitates if $\theta_2 < \sigma^t(x)$.

An excited cell "precipitates" if $\theta_1 < \sigma^t(x)$ or otherwise becomes refractory.

A refractory cell recovers to the resting state unconditionally, and the precipitate cell does not change its state.

Initially we perturb the medium, excite it in several sites, thus inputting data. Waves of excitation are generated, they grow, collide with each other and annihilate as a result of the collision. They may form a stationary inactive concentration profile of a precipitate, which represents the result of the computation. Thus, we can only be concerned with reactions of precipitation: $+ \rightarrow^{k_1} \star$ and $\circ \boxplus + \rightarrow^{k_2} \star$, where k_1 and k_2 are inversely proportional to θ_1 and θ_2, respectively. Varying θ_1 and θ_2 from 1 to 8, and thus changing precipitation rates from the maximum possible to the minimum, we obtain various kinds of precipitate patterns, as shown in Fig. 12.

Precipitate patterns developed for relatively high ranges of reaction rates ($3 \le \theta_1, \theta_2 \le 4$) represent discrete Voronoi diagrams (a given "planar" set, represented by sites of initial excitation, is visible in pattern $\theta_1 = \theta_2 = 3$ as white dots inside the Voronoi cells) derived from the set of initially excited sites, see Fig. 13a and b. This example demonstrates that by externally controlling precipitation rates we can force the reaction-diffusion medium to compute a Voronoi diagram.

When dealing with excitable media excitability is the key parameter for tuning spatio-temporal dynamics. In [1] we demonstrated that by varying excitability we can force the medium to exhibit almost all possible types of excitation dynamics.

Let each cell of the 2D automaton take three states: resting $(\cdot)$, exciting $(+)$ and refractory $(-)$, and update its state depending on number σ_+ of excited neighbors in its 8-cell neighborhood (Fig. 11a). A cell goes from excited to refractory and from refractory to resting states unconditionally, and a resting cell excites if $\sigma_+ \in [\theta_1, \theta_2]$, $1 \le \theta_1 \le \theta_2 \le 8$. By changing θ_1 and θ_2 we can move the medium dynamics in a domain of "conventional" excitation waves, useful for image processing and robot navigation [7] (Fig. 14a), as well as to make it exhibit mobile localized excitations (Fig. 14b), quasi-particles, discrete analogs of dissipative solitons, employed in collision-based computing [1].

Robot Navigation and Massive Manipulation

As we have seen in previous sections, reaction-diffusion chemical systems can solve complex problems and implement logical circuits. Embedded controllers for non-traditional robotics architectures would be yet another potentially huge field of application of reaction-diffusion computers. The physico-chemical artifacts are well-known to be capable of sensible motion. Most famous are Belousov–Zhabotinsky vesicles [26], self-propulsive chemo-sensitive drops [25,37] and ciliar arrays. Their motion is directional but somewhere lacks sophisticated control mechanisms.

At the present stage of reaction-diffusion computing research it seems to be difficult to provide effective solutions for experimental prototyping of combined sensing, decision-making and actuating. However, as a proof-of-

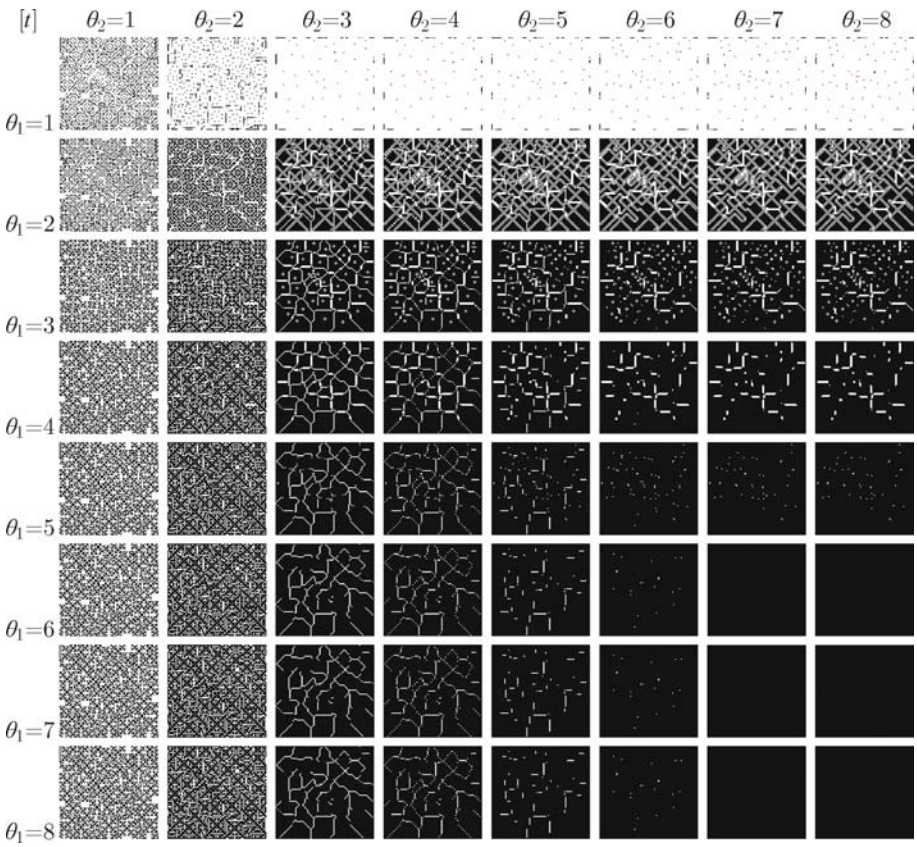

Reaction-Diffusion Computing, Figure 12
Final configurations of reaction-diffusion medium for $1 \leq \theta_1 \leq \theta_2 \leq 2$. Resting sites are *black*, precipitate is *white* [3]

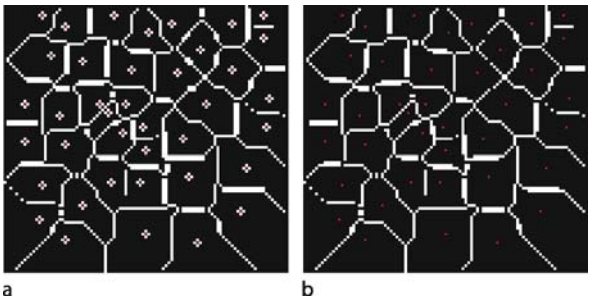

Reaction-Diffusion Computing, Figure 13
Exemplary configurations of reaction-diffusion medium for a $\theta_1 = 3$ and $\theta_2 = 3$, b $\theta_1 = 4$ and $\theta_2 = 3$. Resting sites are *black*, precipitate is *white* [7]

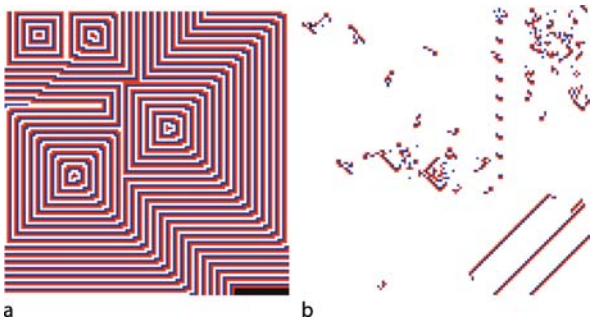

Reaction-Diffusion Computing, Figure 14
Snapshots of space-time excitation dynamics for excitability $\sigma_+ \in [1, 8]$ (a) and $\sigma_+ \in [2, 2]$ (b)

concept we can always consider hybrid "wetware + hardware" systems. For example, to fabricate a chemical controller for a robot (Fig. 15a), we can place a reactor with Belousov–Zhabotinsky solution on-board of a wheeled robot, and allow the robot to observe excitation wave dynamics in the reactor. When the medium is stimulated at one point, target waves are formed. The robot becomes aware of the direction toward source of stimulation from the topology of the wave-fronts [7].

A set of remarkable experiments were undertaken by Hiroshi Yokoi and Ben De Lacy Costello. They built an interface (Fig. 15b) between a robotic hand and Belousov–

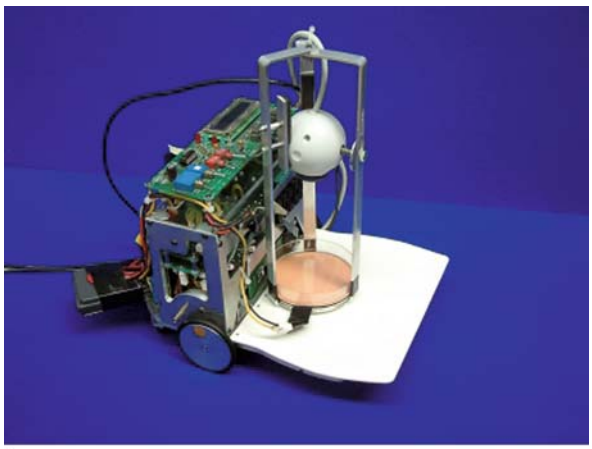

a

b

Reaction-Diffusion Computing, Figure 15
Robots controlled by BZ chemical medium: a mobile robot, b robotic hand (courtesy of Hiroshi Yokoi)

Zhabotinsky chemical reactor [62]. Excitation waves propagating in the reactor were sensed by photo-diodes, which triggered finger motion. Bending fingers touched the chemical medium with their glass nails filled with colloid silver, which triggered circular waves in the medium [7]. Starting from any initial configuration, the chemical-robotic system does always reach a coherent activity mode, where fingers move in regular, somewhat melodic patterns, and few generators of target waves govern dynamics of excitation in the reactor [62].

The chemical processors for navigating the wheeled robot and for controlling, and actively interacting with, a robotic hand are well discussed in our recent monograph [7], therefore we will not go into detail in herein. Instead we will concentrate on rather novel findings on coupling of a reaction-diffusion system with a massive parallel array of virtual actuators.

How a reaction-diffusion medium can manipulate objects? To find out we couple a simulated abstract parallel manipulator with an experimental Belousov–Zhabotinsky (BZ) chemical medium, so the excitation dynamics in the chemical system are reflected in changing the OFF-ON mode of elementary actuating units. In this case, we convert experimental snapshots of the spatially distributed chemical system to a force vector field and then simulate the motion of manipulated objects in the force field, thus achieving reaction-diffusion medium controlled actuation. To build an interface between the recordings of space-time snapshots of the excitation dynamics in the BZ medium and simulated physical objects we calculate force fields generated by mobile excitation patterns and then simulate the behavior of an object in this force field.

The chemical medium to perform actuation is prepared following the typical recipe[1], see [8,20], based on a ferroin catalyzed BZ reaction. A silica gel plate is cut and soaked in a ferroin solution. The gel sheet is placed in a Petri dish and BZ solution added. The dynamics of the chemical system are recorded at 30-second intervals using a digital camera.

The cross-section profile of the BZ wave-front recorded on a digital snapshot shows a steep rise of red color values in the pixels at the wave-front's head and a gradual descent in the pixels along the wave-front's tail. Assuming that excitation waves push the object local force vectors generated at each site—pixel of the digitized image—of the medium should be oriented along local gradients of the red color values. From the digitized snapshot of the BZ medium we extract an array of red components from the snapshot's pixels and then calculate the projection of a virtual vector force at the pixel. Force fields generated by the excitation patterns in a BZ system Fig. 16 result in tangential forces being applied to a manipulated object, thus causing translational and rotational motions of the object [8].

Non-linear medium controlled actuators can be used for sorting and manipulating both small objects, comparable in size to the elementary actuating unit, and larger objects, with lengths of tens or hundreds of actuating units. Therefore, we demonstrate here two types of experiments with BZ-based manipulation: of pixel-sized objects and of planar convex shapes.

Pixel-objects, due to their small size, are subjected to random forces, caused by impurities of the physical medium and imprecision of the actuating units. In this case, no averaging of forces is allowed and the pixel-ob-

[1]Chemical laboratory experiments are undertaken by Dr. Ben De Lacy Costello (UWE, Bristol, UK).

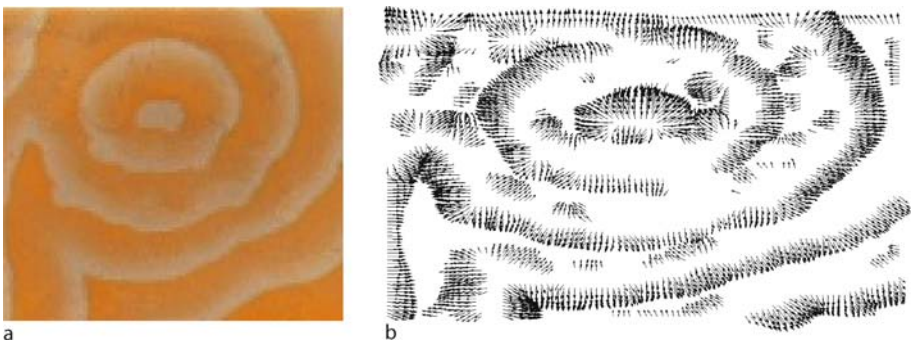

Reaction-Diffusion Computing, Figure 16
Force vector field (b) calculated from BZ medium's image (a) [8]

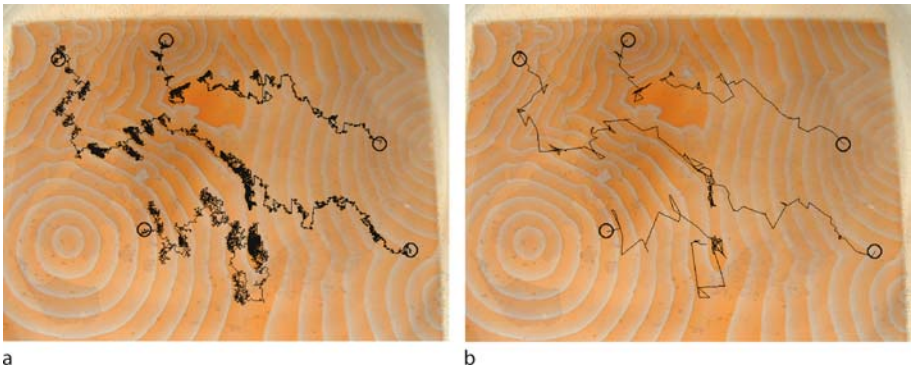

Reaction-Diffusion Computing, Figure 17
Examples of manipulating five pixel-objects using the BZ medium: a trajectories of pixel-objects, b jump-trajectories of pixel-objects recorded every 100th time step. Initial positions of the pixel-objects are shown by *circles* [8]

jects themselves sensitively react to a single force vector. Therefore, we adopt the following model of manipulating a pixel-object: if all force vectors at the 8-pixel neighborhood of the current site of the pixel-object are nil then the pixel-object jumps to a randomly chosen neighboring pixel of its neighborhood, otherwise the pixel-object is translated by the maximum force vector in its neighborhood.

When placed on the simulated manipulating surface, pixel-objects move at random in the domains of the resting medium, however by randomly drifting each pixel-object does eventually encounter a domain of co-aligned vectors (representing the excitation wave front in the BZ medium) and is translated along the vectors. An example of several pixel-objects transported on a "frozen" snapshot of the chemical medium is shown in Fig. 17. Trajectories of pixel-objects (Fig. 17a) show distinctive intermittent modes of random motion separated by modes of directed "jumps" guided by traveling wave fronts. Smoothed trajectories of pixel-objects (Fig. 17b) demonstrate that despite a very strong chaotic component in manipulation, pixel-objects are transported to the sites of the medium where two or more excitation wave-fronts meet.

The overall speed of pixel-object transportation depends on the frequency of wave generations by sources of target waves. As a rule, the higher the frequency the faster the objects are transported. This is because in parts of the medium spanned by low frequency target waves there are lengthy domains of resting system, where no force vectors are formed. Therefore, pixel-sized objects can wander randomly for a long time till climbing the next wave front [8].

To calculate the contribution of each force we partitioned the object into fragments, using a square grid, in which each cell of the grid corresponds to one pixel of the image. We assume that the magnitude of the force applied to each fragment above a given pixel is proportional to the area of the fragment, and is co-directional with a force vector. A momentum of inertia of the whole object with respect to the axis normal to the object and passing through the object's center of mass is calculated from the position

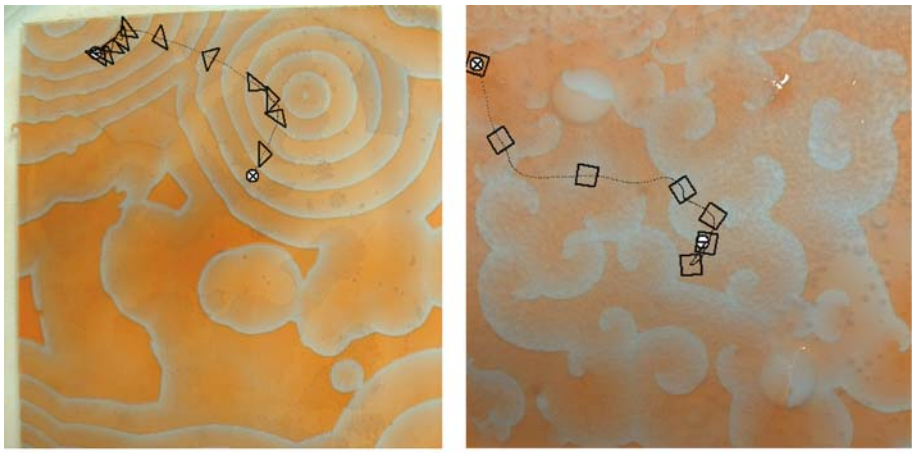

a b

Reaction-Diffusion Computing, Figure 18

Manipulating planar object in BZ medium. a *Right-angled triangle* moved by fronts of target waves. b *Square object* moved by fronts of fragmented waves in sub-excitable BZ medium. Trajectories of center of mass of the square are shown by the *dotted line*. Exact orientation of the objects is displayed every 20 steps. Initial position of the object is shown by ⊖ and the final position by ⊗ [8]

of the center of mass and the mass of every fragment. Since the object's shape and size are constant it is enough to calculate the moment of inertia only at the beginning of simulation. We are also taking into account principal rotational momentum created by forces and angular acceleration of the object around its center of mass. Therefore, object motion in our case can be sufficiently described by the coordinates of its center of mass and its rotation at every moment of time [8].

Spatially extended objects follow the general pattern of motion observed for the pixel-sized objects. However, due to integration of many force vectors the motion of planar objects is smoother and less sensitive to the orientation of any particular force-vector.

Outcome of manipulation depends on the size of the object, with increasing size of the object—due to larger numbers of local vector-forces acting on the object—the objects become more controllable by the excitation wavefronts (Fig. 18).

Future Directions

The field of reaction-diffusion computing started 20 years ago [28] as a sub-field of physics and chemistry dealing with image processing operations in uniform thin-layer excitable chemical media. The basic idea was to apply input data as a two-dimensional profile of heterogeneous illumination, then allow excitation waves to spread and interact with each, and then optically record the result of the computation. The first reaction-diffusion computers were already massively parallel, with parallel optical input and outputs. Later computer engineers entered the field, and started to exploit traditional techniques—wires were implemented by channels where wave-pulses travel, and specifically shaped junctions acted as logical valves. In this manner, the most "famous" chemical computing devices were implemented, including Boolean gates, coincidence detectors, memory units and more. The idea of reaction-diffusion computation was if not ruined then forced into a cul-de-sac of non-classical computation. The breakthrough happened when paradigms and solutions from the field of dynamical, collision-based computing and conservative logic, were mapped onto realms of spatially extended chemical systems. The computers became uniform and homogeneous.

Over several examples we demonstrated that reaction-diffusion chemical systems are capable of solving combinatorial problems with natural parallelism. In spatially distributed chemical processors, the data and the results of the computation are encoded as concentration profiles of the chemical species. The computation per se is performed via the spreading and interaction of wave fronts.

The reaction-diffusion computers are parallel because the chemical medium's micro-volumes update their states simultaneously, and molecules diffuse and react in parallel. During the last decades a wide range of experimental prototypes of reaction-diffusion computing devices have been fabricated and applied to solve various problems of computer science, including: image processing, pattern recognition, path planning, robot navigation, computational ge-

ometry, logical gates in spatially distributed chemical media, arithmetical and memory units.

These important—but scattered across many scientific fields— results convince us that reaction-diffusion systems can do a lot. Are they capable enough to be intelligent? Yes, reaction-diffusion systems are smart—showing a state of readiness to respond, able to cope with difficult situations, capable of determining something by mathematical and logical methods—and endowed with the capacity to reason. Reaction-diffusion computers allow for massive parallel input of data. Equivalently, reaction-diffusion robots would need no dedicated sensors, each micro-volume of the medium, each site of the matrix gel, is sensitive to changes in one or another physical characteristic of the environment. Electric field, temperature and illumination are "sensed" by reaction-diffusion devices, and these are the three principle parameters in controlling and programming reaction-diffusion robots.

Hard computational problems of geometry, image processing and optimization on graphs are resource-efficiently solved in reaction-diffusion media due to intrinsic natural parallelism of the problems [1]. Herein we demonstrated the efficiency of reaction-diffusion computers with the example of constructing Voronoi diagrams. The Voronoi diagram is a subdivision of plane by data planar set. Each point of the data set is represented by a drop of a reagent. The reagent diffuses and produces a color precipitate when reacting with the substrate. When two or more diffusive fronts of the "data" chemical species meet, no precipitate is produced (due to concentration-dependent inhibition). Thus, uncolored domains of the computing medium represent bisectors of the Voronoi diagram. The precipitating chemical processor can also compute a skeleton. The skeleton of a planar shape is computed in a similar manner. A contour of the shape is applied to the computing substrate as a disturbance in reagent concentrations. The contour concentration profile induces diffusive waves. A reagent diffusing from the data-contour reacts with the substrate and the precipitate is formed. The precipitate is not produced at the sites of collision of diffusive waves. The uncolored domains correspond to the skeleton of the data shape. To compute a collision-free shortest path in a space with obstacles, we can couple two reaction-diffusion media. Obstacles are represented by local disturbances of concentration profiles in one of the media. The disturbances induce circular waves traveling in the medium and approximating a scalar distance-to-obstacle field. This field is mapped onto the second medium, which calculates a tree of "many-sources-one-destination" shortest paths by spreading wave-fronts [7].

Just few words of warning—when thinking about chemical algorithms some of you may realize that diffusive and phase waves are pretty slow in physical time. The sluggishness of computation is the only point that may attract criticism to reaction-diffusion chemical computers. There is however a cure—to speed up we are implementing the chemical medium in silicon, micro-processor LSI analogs of reaction-diffusion computers [12]. Further miniaturization of the reaction-diffusion computers can be reached when the system is implemented as a two-dimensional array of single-electron nonlinear oscillators diffusively coupled to each other [38]. Yet another point of developing reaction-diffusion computers—is to design embedded controllers for soft-bodied robots, where usage of conventional silicon materials seem to be inappropriate.

Acknowledgments

Chemical laboratory prototypes of reaction-diffusion computers, discussed in the article, were implemented by Ben De Lacy Costello. I am grateful to Andy Wuensche (hexagonal cellular automata), Hiroshi Yokoi (robotic hand controlled by Belousov–Zhabotinsky reaction), Chris Melhuish (control of robot navigation), Sergey Skachek (massive parallel manipulation), Tetsuya Asai (LSI prototypes of reaction-diffusion computers) and Genaro Martinez (binary-state cellular automata) for their cooperation. Some pictures, where indicated, where adopted from our co-authored publications. My sincere thanks go to Soichiro Tsuda and Tomohiro Shirakawa who ignited my interest in experimenting with plasmodium and were my patient advisers.

Bibliography

Primary Literature

1. Adamatzky A (2001) Computing in nonlinear media and automata collectives. Institute of Physics Publishing, London
2. Adamatzky A (ed) (2003) Collision based computing. Springer, London
3. Adamatzky A (2005) Programming reaction-diffusion computers. In: Unconventional programming paradigms. Springer, New York
4. Adamatzky A, De Lacy Costello BPJ (2002) Experimental logical gates in a reaction-diffusion medium: the XOR gate and beyond. Phys Rev E 66:046112
5. Adamatzky A, Wuensche A (2006) Computing in 'spiral rule' reaction-diffusion hexagonal cellular automaton. Complex Syst 16(4):277–298
6. Adamatzky A, De Lacy Costello BPJ, Melhuish C, Ratcliffe N (2003) Experimental reaction-diffusion chemical processors for robot path planning, J Intell Robt Syst 37:233–249
7. Adamatzky A, De Lacy Costello BPJ, Asai T (2005) Reaction-diffusion computers. Elsevier, Amsterdam

8. Adamatzky A, De Lacy Costello BPJ, Skachek S, Melhuish C (2005) Manipulating objects with chemical waves: open loop case of experimental Belousov–Zhabotinsky medium. Phys Lett A 350(3–4):201–209
9. Adamatzky A, Wuensche A, De Lacy Costello BPJ (2006) Glider-based computation in reaction-diffusion hexagonal cellular automata. Chaos, Solitons Fractals 27:287–295
10. Agladze K, Magome N, Aliev R, Yamaguchi T, Yoshikawa K (1997) Finding the optimal path with the aid of chemical wave. Phys D 106:247–254
11. Asai T, De Lacy Costello BPJ, Adamatzky A (2005) Silicon implementation of a chemical reaction-diffusion processor for computation of Voronoi diagram. Int J Bifurc Chaos 15(1)
12. Asai T, Kanazawa Y, Hirose T, Amemiya Y (2005) Analog reaction-diffusion chip imitating Belousov–Zhabotinsky reaction with hardware oregonator model. Int J Unconv Comput 1:123–147
13. Beato V, Engel H (2003) Pulse propagation in a model for the photosensitive Belousov–Zhabotinsky reaction with external noise. In: Schimansky-Geier L, Abbott D, Neiman A, Van den Broeck C (eds) Noise in complex systems and stochastic dynamics Proc. SPIE, vol 5114. pp 353–362
14. Berlekamp ER, Conway JH, Guy RL (1982) Winning ways for your mathematical plays, vol 2. Academic Press
15. Bode M, Liehr AW, Schenk CP, Purwins H-G (2000) Interaction of dissipative solitons: particle-like behavior of localized structures in a three-component reaction-diffusion system. Phys D 161:45–66
16. Brandtstädter H, Braune M, Schebesch I, Engel H (2000) Experimental study of the dynamics of spiral pairs in light-sensitive Belousov–Zhabotinskii media using an open-gel reactor. Chem Phys Lett 323:145–154
17. Courant R, Robbins H (1941) What is mathematics? Oxford University Press
18. Dupont C, Agladze K, Krinsky V (1998) Excitable medium with left-right symmetry breaking. Phys A 249:47–52
19. Field RJ, Noyes RM (1974) Oscillations in chemical systems: IV. Limit cycle behavior in a model of a real chemical reaction. J Chem Phys 60:1877–1884
20. Field R, Winfree AT (1979) Traveling waves of chemical activity in the Zaikin–Zhabotinsky–Winfree reagent. J Chem Educ 56:754
21. Fredkin F, Toffoli T (1982) Conservative logic. Int J Theor Phys 21:219–253
22. Gerhardt M, Schuster H, Tyson JJ (1990) A cellular excitable media. Phys D 46:392–415
23. Grill S, Zykov VS, Müller SC (1996) Spiral wave dynamics under pulsatory modulation of excitability. J Phys Chem 100:19082–19088
24. Hartman H, Tamayo P (1990) Reversible cellular automata and chemical turbulence. Phys D 45:293–306
25. Kitahata H, Yoshikawa K (2005) Chemo-mechanical energy transduction through interfacial instability. Phys D 205:283–291
26. Kitahata H, Aihara R, Magome N, Yoshikawa K (2002) Cinvective and periodic motion driven by a chemical wave. J Chem Phys 116:5666
27. Klein R (1990) Concrete and abstract voronoi diagrams. Springer, Berlin
28. Kuhnert L (1986) A new photochemical memory device in a light sensitive active medium. Nature 319:393
29. Kuhnert L, Agladze KL, Krinsky VI (1989) Image processing using light-sensitive chemical waves. Nature 337:244–247
30. Kusumi T, Yamaguchi T, Aliev R, Amemiya T, Ohmori T, Hashimoto H, Yoshikawa K (1997) Numerical study on time delay for chemical wave transmission via an inactive gap. Chem Phys Lett 271:355–360
31. Maeda S, Hashimoto S, Yoshida R (2004) Design of chemo-mechanical actuator using self-oscillating gels. In: Proc. IEEE International Conference on Robotics and Biomimetics (ROBIO 2004). p 313
32. Margolus N (1984) Physics-like models of computation. Phys D 0:81–95
33. Markus M, Hess B (1990) Isotropic cellular automata for modeling excitable media. Nature 347:56–58
34. Mills J (2005) The new computer science and its unifying principle: complementarity and unconventional computing. In: The Grand Challenge in Nonclassical Computation, Int. Workshop, Position Papers. York, 18–19 April 2005
35. Motoike IN, Yoshikawa K (2003) Information operations with multiple pulses on an excitable field. Chaos, Solitons Fractals 17:455–461
36. Motoike IN, Yoshikawa K, Iguchi Y, Nakata S (2001) Real-time memory on an excitable field. Phys Rev E 63:036220
37. Nagai K, Sumino Y, Kitahata H, Yoshikawa K (2005) Mode selection in the spontaneous motion of an alcogol droplets. Phys Rev E 71:065301
38. Oya T, Asai T, Fukui T, Amemiya Y (2005) Reaction-diffusion systems consisting of single-electron oscillators. Int J Unconvent Comput 1:179–196
39. Petrov V, Ouyang Q, Swinney HL (1997) Resonant pattern formation in a chemical system. Nature 388:655–657
40. Pour-El MB (1974) Abstract computability and its relation to the general purpose analog computer (some connections between logic, differential equations and analog computers). Trans Am Math Soc 199:1–28
41. Qian H, Murray JD (2001) A simple method of parameter space determination for diffusion-driven instability with three species. Appl Math Lett 14:405–411
42. Rambidi NG (1998) Neural network devices based on reaction-diffusion media: an approach to artificial retina. Supramol Sci 5:765–767
43. Rambidi NG, Shamayaev KR, Peshkov GY (2002) Image processing using light-sensitive chemical waves. Phys Lett A 298:375–382
44. Ramos JJ (2003) Oscillatory dynamics of inviscid planar liquid sheets. Appl Math Comput 143:109–144
45. Saltenis V (1999) Simulation of wet film evolution and the Euclidean Steiner problem. Informatica 10:457–466
46. Schebesch I, Engel H (1998) Wave propagation in heterogeneous excitable media. Phys Rev E 57:3905–3910
47. Schenk CP, Or-Guil M, Bode M, Purwins H-G (1997) Interacting pulses in three-component reaction-diffusion systems on two-dimensional domains. Phys Rev Lett 78:3781–3784
48. Sediña-Nadal I, Mihaliuk E, Wang J, Pérez-Muñuzuri V, Showalter K (2001) Wave propagation in subexcitable media with periodically modulated excitability. Phys Rev Lett 86:1646–1649
49. Sielewiesiuk J, Gorecki J (2001) Logical functions of a cross junction of excitable chemical media. J Phys Chem A 105:8189–8195

50. Sienko T, Adamatzky A, Rambidi N, Conrad M (eds) (2003) Molecular computing. MIT Press
51. Steinbock O, Toth A, Showalter K (1995) Navigating complex labyrinths: optimal paths from chemical waves. Science 267:868–871
52. Tóth A, Showalter K (1995) Logic gates in excitable media. J Chem Phys 103:2058–2066
53. Tyson JJ, Fife PC (1980) Target patterns in a realistic model of the Belousov–Zhabotinskii reaction. J Chem Phys 73:2224–2237
54. Vergis A, Steiglitz K, Dickinson B (1986) The complexity of analog computation. Math Comput Simul 28:91–113
55. Wang J (2001) Light-induced pattern formation in the excitable Belousov–Zhabotinsky medium. Chem Phys Lett 339:357–361
56. Weaire D, Hutzler S, Cox S, Kern N, Alonso MD, Drenckhan W (2003) The fluid dynamics of foams. J Phys Condens Matter 15:S65–S73
57. Wuensche A (2005) Glider dynamics in 3-value hexagonal cellular automata: the beehive rule. Int J Unconvent Comput 1:375–398
58. Wuensche A, Adamatzky A (2006) On spiral glider-guns in hexagonal cellular automata: Activator-inhibitor paradigm. Int J Mod Phys 17:1009–1026
59. Yaguma S, Odagiri K, Takatsuka K (2004) Coupled-cellular-automata study on stochastic and pattern-formation dynamics under spatiotemporal fluctuation of temperature. Phys D 197:34–62
60. Yang X (2004) Pattern formation in enzyme inhibition and cooperativity with parallel cellular automata. Parallel Comput 30:741–751
61. Yang X (2006) Computational modeling of nonlinear calcium waves. Appl Math Model 30:200–208
62. Yokoi H, Adamatzky A, De Lacy Costello B, Melhuish C (2004) Excitable chemical medium controlled for a robotic hand: closed loop experiments. Int J Bifurc Chaos 14(9)3347–3354
63. Yoneyama M (1996) Optical modification of wave dynamics in a surface layer of the Mn-catalyzed Belousov–Zhabotinsky reaction. Chem Phys Lett 254:191–196
64. Young D (1984) A local activator–inhibitor model of vertebrate skin patterns. Math Biosci 72:51

Books and Reviews

Adamatzky A (2001) Computing in nonlinear media and automata collectives. Institute of Physics Publ, London
Adamatzky A (ed) (2003) Collision-Based computing. Springer, London
Adamatzky A, Teuscher C (eds) (2006) From utopian to genuine unconventional computers. Luniver Press, London
Adamatzky A, De Lacy Costello B, Asai T (2005) Reaction-Diffusion computers. Elsevier, Amsterdam
Chua L (1998) CNN: a paradigm for complexity. World Scientific
Gray P, Scott SK (2002) Chemical oscillations and instabilities: nonlinear chemical kinetics. Oxford University Press, Oxford
Scott SK (1994) Oscillations, waves and chaos in chemical kinetics. Oxford University Press, Oxford
Teuscher C, Adamatzky A (eds) (2005) Proceedings of the 2005 workshop on unconventional computing: from cellular automata to wetware. Luniver Press, London
Toffoli T, Margolus N (1987) Cellular automata machines. MIT Press

Reaction Kinetics in Fractals

Ezequiel V. Albano
Facultad de Ciencias Exactas, UNLP, CONICET, Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA) CCT La Plata, La Plata, Argentina

Article Outline

Glossary
Definition of the Subject
Introduction
Fractals and Some of Their Relevant Properties
Random Walks
Diffusion-limited Reactions
Irreversible Phase Transitions in Heterogeneously Catalyzed Reactions
Future Directions
Bibliography

Glossary

Fractals Fractal geometry is a mathematical tool well suited to treating complex systems that exhibit scale invariance or, equivalently, the absence of any characteristic length scale. Scale invariance implies that objects are self-similar: if we take a part of the object and magnify it by the same magnification factor in all directions, the obtained object cannot be distinguished from the original one. Self-similar objects are often characterized by non-integer dimensions, a fact that led B. Mandelbrot, early in the 1980s, to coin the name "fractal dimension". Also, all objects described by fractal dimensions are generically called *fractals*. Within the context of the present work, fractals are the underlying media where the *reaction kinetics* among atoms, molecules, or particles in general, is studied.

Reaction kinetics The description of the time evolution of the concentration of reacting particles (ρ_i, where $i = 1, 2, \ldots, N$ identifies the type of particle) is achieved, far from a stationary regime, by formulating a kinetic rate equation $\dot{\rho}_i(t) = F[\rho_i(t)]$, where F is a function. This description, known in physical chemistry as the *law of mass action*, states that the rate of a chemical reaction is proportional to the concentration of reacting species and was formulated by Waage and Guldberg in 1864. Often, especially when dealing with reactions occurring in homogeneous media, F involves integer powers (also known as the reaction orders) of the concentrations, leading to *classical reaction kinetics*. However, as in most cases treated in this arti-

cle, if a reaction takes place in a fractal, one may also have kinetic rate equations involving non-integer powers of the concentration that lead to *fractal reaction kinetics*. Furthermore, it is usual to find that the slowest step involved in a kinetic reaction determines its rate, leading to a *process-limited reaction*, where e. g. the process could be diffusion or mass transport, adsorption, reaction, etc.

Heterogeneously catalyzed reactions A reaction limited by at least one rate-limiting step could be prohibitively slow for practical purposes when, e. g., it occurs in a homogeneous media. The role of a good solid-state catalyst in contact with the reactants – in the gas or fluid phase – is to obtain an acceptable output rate of the products. Reactions occurring in this way are known as heterogeneously catalyzed. This type of reaction involves at least the following steps: (i) the first one comprises trapping, sticking and adsorption of the reactants on the catalytic surface. Particularly important, from the catalytic point of view, is that molecules that are stable in the homogeneous phase, e. g. H_2, N_2, O_2, etc., frequently undergo dissociation on the catalyst surface. This process is essential in order to speed up the reaction rate. (ii) After adsorption, species may diffuse or remain immobile (chemisorbed) on the surface. The actual reaction step occurs between neighboring adsorbed species of different kinds. The result of the reaction is the formation of products that can either be intermediates of the reaction or its final output. (iii) The final step is the desorption of the products, which is essential not only for the practical purpose of collecting and storing the desired output, but also in order to regenerate the catalytically active surface sites.

Definition of the Subject

Processes involving reaction among atoms, molecules, and particles in general, are ubiquitous both in nature and in laboratories. After the introduction of the concept of fractals by B. Mandelbrot [89] in the 1980s, it has been realized that a wide variety of reactions take place in fractal media, leading to the observation of anomalous behavior, i. e. the so called "fractal reaction kinetics" [33,67,81]. The study of this subject has received considerable theoretical attention [38,43,76,77,103] and a huge numerical effort has been made to improve its understanding [13,17,18,20,34,80,92]. Furthermore, among the practical examples of fractal reaction kinetics one can quote exciton annihilation in composite materials, chemical reactions in membranes pores, charge recombination in colloids and clouds, coagulation, polymerization and growth of dendrites, [15,79,83,94] etc. Another scenery for the study of *reaction kinetics in fractals* is in the field of heterogeneous catalysis. In fact, it is well known that most catalysts are composed of small, catalytically active particles, supported by highly porous (fractal) substrates [21,22,23,30,75,95,106]. So, it is not surprising that this field of research has become particularly active due not only to its practical and technological relevance, but also to the occurrence of quite interesting and challenging phenomena such as irreversible phase transitions, oscillatory behavior, chaos and stochastic resonance, propagation and interference of chemical waves, interface coarsening, metastability and hysteretic effects, etc. [9,10,52,68,70,71,84,85,90,109].

Introduction

The description of the kinetic behavior of reaction processes occurring in homogeneous media can be found in most textbooks on chemical physics. However, care must be taken considering *reaction kinetics in fractals* because conventional textbook equations are based on mean-field approaches that neglect not only the fractal structure of the underlying media, but also fluctuations in the concentration of the reactants, many-particle effects, etc. that may lead to the observation of anomalous behavior [81]. In fact, soon after the recognition of the relevance of the concept of fractals for the description of the structure and properties of physical objects by Mandelbrot [89], more than three decades ago, it was realized that random transport in low-dimensional and disordered media may be anomalous [33,67,105], as are diffusion-limited reactions occurring in those media [13,15,17,18,20,34,43,79,80,81,83,92,94]. Therefore, the classical reaction kinetic approach becomes unsatisfactory in a wide variety of situations, such as when the reactants are spatially constrained by walls, phase boundaries or force fields. Due to these studies, early in the 1980s, it was realized that even elementary reactions (e. g. $A + A \rightarrow$ inert, $A + B \rightarrow$ products, etc.) can no longer be described by classical rate equations with integer exponents (the so-called "reaction order") but instead, fractal orders were identified leading to *fractal reaction kinetics* [89]. The emerging new theory also predicts the occurrence of self-ordering and self-unmixing of reactants [38,76,77,103], as well as time-dependent rate "constants", i. e. rate coefficients with temporal "memories" [89].

On the other hand, after the exhaustive study and characterization of the fractal nature of a wide variety of substrates used as support in most catalysts, due to Avnir et

al. [21,22,23,95], the study of heterogeneously catalyzed reactions in fractal media has also attracted growing attention [1,2,5,7,41,72,86,87]. These studies are focused on the understanding of irreversible phase transitions occurring between an active state with reaction and an inactive state where the reaction ceases irreversibly. This latter state is due to the poisoning of the catalyst, by the reactants and/or their subproducts, and is known as the absorbing state: a system trapped in the absorbing state can never escape from it [9,10,68,85,90].

In order to cover those topics, this article is organized as follows: in Sect. "Fractals and Some of Their Relevant Properties", the basic properties of fractals are presented and discussed. Section "Random Walks" is devoted to the description of the behavior of random walks in fractal media, while in Sect. "Diffusion-limited Reactions" archetypical cases of diffusion limited reactions among random walks are considered. Irreversible phase transitions occurring in heterogeneously catalyzed reactions are addressed in Sect. "Irreversible Phase Transitions in Heterogeneously Catalyzed Reactions". Finally, in Sect. "Future Directions" promising directions for future works are briefly outlined.

Fractals and Some of Their Relevant Properties

Classical Euclidean geometry is useful for describing the properties of regular objects such as circles, spheres, cones, etc. However, disordered objects such as clouds, dielectric breakdown patterns, coastlines, mountains, landscapes, etc., have not been so far satisfactorily described by using the Euclidean geometry. B. Mandelbrot overcome this shortcoming by introducing the fractal geometry as a suitable tool for the treatment of disordered or fractal media [89]. Along this article, fractals will be used as media where reactions of interest take place. A relevant property of fractal media is self-similarity or symmetry under dilatation [27,33,40,61,67]. Here, it is worth noting the main difference between regular Euclidean space and fractal geometry: whereas the former has translation symmetry, this type of symmetry is violated in the latter, which exhibits a new symmetry known as scale invariance or invariance under dilatation. Since these concepts may appear to be obscure, it is convenient to develop an intuitive understanding and outline some basic definitions. Fractals can either be deterministic or non-deterministic (often called random fractals). However, it is worth mentioning that many non-deterministic fractals may be obtained as a result of complex dynamic (non-stochastic) processes. From the large variety of known deterministic fractals, e. g. the Koch curve, the Julia set, Sierpinski gaskets, carpets, and sponges, etc., let us focus our attention on the construction of a Sierpinski Carpet (SC). In order to build up a generic SC(a,c), a square in $d = 2$ dimensions is segmented into a^d subsquares and c of them are then removed. This square, of side a, is known as the generating cell. The segmentation process is then iterated on the remaining squares a number k of segmentation steps. Figure 1 shows the SC(4,4) obtained after $k = 4$ segmentation steps. These kinds of fractals have a lower cutoff length given by the side of the generating cell, but, in principle, they lack an upper cutoff length for $k \to \infty$. However, in practice one performs a finite number of segmentation steps.

Let us now remind the reader about the concept of dimension in regular systems. In this case, the dimension d characterizes the dependence of the mass $M(L)$ as a function of the linear size L of the system. If we now consider a smaller part of the system of size $bL(a < 1)$, then $M(aL)$ will be decreased by a factor of a^d, so that

$$\tilde{M}(a, L) = \mu(a)M(aL) = a^d M(L) \,. \tag{1}$$

The solution of Eq. (1) gives $M(L) = AL^d$, where A is a constant. So, for a wire one has $d = 1$, while for a thin plate $d = 2$ is obtained, etc. Intuitively, the SC(4,4) shown in Fig. 1 seems to be "denser" than a wire but also "sparser" than a thin plate. So, Eq. (1) has to be generalized for fractals, leading to

$$M(aL) = a^{d_f} M(L) \,, \tag{2}$$

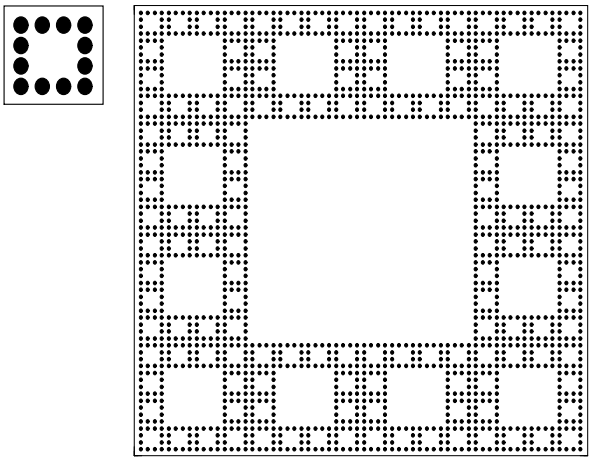

Reaction Kinetics in Fractals, Figure 1
Sierpinski carpet obtained after $k = 4$ iterations on the generating cell shown in the *top-left corner*, i. e. the SC(4,4). The lattice side is $L = 64$ and the fractal dimension of this object is $D_f = log(12)/log(4) \simeq 1.7925$

which yields the following solution

$$M(L) = BL^{d_f} , \tag{3}$$

where B is a constant, and d_f is a non-integer dimension known, after Mandelbrot, as the fractal dimension. By applying Eq. (2) to the SC(4,4) shown in Fig. 1, one has $M(1/4\,L) = 1/12\,M(L)$, so that $d_f = \log(12)/\log(4) \simeq 1.7925$. In general, for an SC one has $M(1/aL) = 1/(a^2 - c)\,M(L)$, which yields $d_f = \log(a^2 - c)/\log(a)$.

The factor a in Eqs. (1) and (3) could be an arbitrary real number, leading to *continuous* scale invariance. However, many fractals exhibit *discrete* scale invariance (DSI) [99], which is a weak kind of scale invariance such that a is no longer an arbitrary real number, but it can only take specific discrete values of the form $a_n = (a_1)^n$, where a_1 is a fundamental scaling ratio. Then, for the case of DSI, the solution of Eq. (1) yields

$$M(L) = L^{d_f} F\left(\frac{\log(L)}{\log(a_1)}\right) , \tag{4}$$

where F is a periodic function of period one. The measurement of soft oscillations in spatial domain [99] is a signature of spatial DSI. For instance, the $SC(4,4)$ shown in Fig. 1 exhibits DSI with $a_1 = 4$.

Deterministic fractals can be finitely ramified or infinitely ramified objects. A fractal is classified as finitely ramified if any bounded set of the fractal can be separated from the whole structure just by removing a finite number of bonds between the individual constituting entities. This is not possible for the case of the SC(4,4) shown in Fig. 1, so that it is an infinitely ramified fractal. However, other fractals such as the Sierpinski Gasket and the Koch curve are finitely ramified. One advantage of finitely ramified clusters is that various physical properties, such as the conductivity and the vibrational excitations, can be calculated exactly helping to understand the anomalous behavior of some observables. However, it is worth mentioning that classical critical phenomena, e. g. the Ising model, are not observed in finitely ramified fractals [58,59,60]. This is not the case of irreversible critical phenomena (e. g. the contact process [68]), criticality occurring at a trivial value of the control parameter, e. g. the coarsening without surface tension observed in the voter model at $T = 0$ [48], which is still present in finitely ramified fractals [102], etc.

For a careful description of the building rules and relevant properties of various deterministic fractals see e. g. [40].

In the large variety of non-deterministic fractals, there are many examples of random fractals such as diffusion-limited aggregates [27,40,107], cluster-cluster aggregates [101], the incipient percolation cluster [101], etc. So,

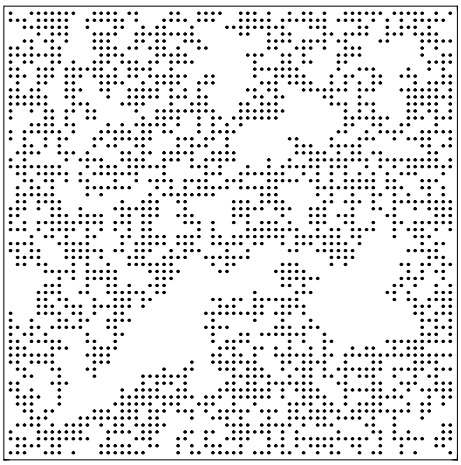

Reaction Kinetics in Fractals, Figure 2
Percolating cluster obtained at the critical threshold. The lattice side is $L = 64$ and the cluster contains 2100 particles. The fractal dimension of the infinite percolation cluster is $D_f \simeq 1.89$

in order to acquaint the reader with this subject, let us briefly describe the percolation model in $d = 2$ dimensions. Considering the square lattice, one has that each site can be occupied randomly with probability p or left empty with probability $(1 - p)$. This model mimics, for example, the deposition of conducting metallic particles in an isolating substrate, so that p is the surface density of deposited particles. At lower densities, one has isolated clusters of particles and the films do not conduct electrical current between the edges of the sample. However, by properly increasing the concentration, the onset of electrical conductivity is abruptly observed at a certain critical value p_c, known as the percolation threshold [101]. So, at criticality one has a spanning cluster that connects opposite edges of the sample, which is known as the Incipient Percolation Cluster (IPC), see Fig. 2. Equations (2) and (3) also hold for random fractals, but the mass M has to be averaged over different realizations of the fractal in order to achieve reliable statistics. The IPC is composed of several fractal substructures such as the backbone, blobs, red bonds, dangling ends, etc., so that one needs to define different fractal dimensions in order to achieve a better description.

Several methods have been developed in order to measure the fractal dimension of non-deterministic fractals, the "sandbox" and the "box counting" methods being among the most extensively used, for further details see e. g., [40].

Random Walks

The properties of random walks are useful for the understanding of a great variety of phenomena in virtually all

sciences, e. g. in Physics, Chemistry, Astronomy, Biology, Ecology, and even in Economics [105].

Throughout this article, we will focus our attention on simple random walks on lattices, either regular or fractal, as a model for diffusion. The simple discrete random walk is a stochastic process, such that the walk advances one step in unit time. The walker steps from its present position to another site of the lattice according to a specified random rule. Since the rule is independent of the history of the walk, the process is Markovian. Along this article, we will consider cases where the step is performed, with the same probability, to one of the nearest neighbor sites of the lattice. While that choice is always possible in regular lattices, it is no longer the case for random walks in fractals, because all nearest neighbor sites may not belong to the substrate. In this situation one often considers steps performed with equal probability to any of the nearest neighbor sites belonging to the substrate.

Let us now consider the displacement of the random walk. After n steps (since one has discrete time scale, $n = t$ holds) the net displacement ($R(t)$) is given by [33,67,105]

$$R(t) = \sum_{j=1}^{n} \mathbf{u}_j \,, \tag{5}$$

where $\mathbf{u}_j$ is a unitary vector pointing to a nearest-neighbor site, so that it represents the jth step of the walk. Due to the fact that $\langle \mathbf{u}_j \rangle = 0$, the displacement averaged over a large number of realizations of the walk vanishes, i. e. $\langle R(t) \rangle = 0$. So, a more interesting and useful kinetic observable of random walks is the rms displacement from the origin (R^2), given by

$$\langle R^2(t) \rangle = \left\langle \left(\sum_{j=1}^{n} \mathbf{u}_j \right)^2 \right\rangle = t + 2 \sum_{j>i}^{n} \langle \mathbf{u}_j . \mathbf{u}_i \rangle = t \,, \tag{6}$$

since $\langle \mathbf{u}_j . \mathbf{u}_j \rangle = 1$ and $\langle \mathbf{u}_j . \mathbf{u}_i \rangle = 0$ for $i \neq j$ because the steps are independent. So, the classical result for Euclidean space is obtained.

A distinctive feature of transport in fractal media is that the linear time dependence of the rms displacement of the walk given by Eq. (6) has to be replaced by

$$R^2 \propto t^{2/d_\mathrm{w}} \,, \tag{7}$$

where d_w is the anomalous diffusion exponent. In most studied models the fractal dimension d_w exceeds 2, due to the fact that disorder tends, on average, to slow down the diffusion of the walk moving on those media.

Another interesting observable of a walk is the average number of distinct sites visited by a single random walk after N steps (S_N), which is also known as the exploration space of the walk [17,19,67,105] and references therein. Assuming that $N \propto t$, one has [96]

$$S_N \propto t^{d_\mathrm{s}/2} \,, \quad t \to \infty \,, \tag{8}$$

where d_s is the spectral dimension and Eq. (8) holds for $d_\mathrm{s} < 2$ [14,96], which corresponds to low-dimensional and fractal media, leading to the so-called "anomalous diffusion" behavior [19,33,67]. Furthermore, $d = 2$ is the upper critical dimension, such as for a d-dimensional regular space one has $d_\mathrm{s} = d > 2$ and Eq. (8) becomes $S_N \propto t$, leading to classical diffusion (for further details see also Eq. (17) below).

Diffusion, or random walks, are also related to the density of states for harmonic excitations of the media [78] ($h(\epsilon)$) through the probability that the random walk returns to the origin. So, one has

$$h(\epsilon) \propto \epsilon^{\frac{d_\mathrm{f}}{d_\mathrm{w}} - 1} \equiv \epsilon^{\frac{d_\mathrm{s}}{2} - 1} \,, \tag{9}$$

where the energy and vibrational density of states are related through $h(\epsilon) d\epsilon = g(\omega) d\omega$. The relationship (9) is similar to the well-known density of states in Euclidean space, except that d has to be replaced by

$$d_\mathrm{s} = 2 \frac{d_\mathrm{f}}{d_\mathrm{w}} \,, \tag{10}$$

which is also known as the fracton dimensionality, since the vibration modes are called fractons instead of phonons.

Very recently we have found evidence of discrete scale invariance in the *time* domain by measuring the relaxation of the magnetization in the Ising model on Sierpinski Carpets [25,26]. Subsequently, we have conjectured that physical processes characterized by an observable $O(t)$, occurring in fractal media with DSI, and that develop a monotonically increasing time-dependent characteristic length $\xi(t)$, may also exhibit *time* DSI. In fact, by assuming $\xi \propto t^{1/z}$, where z is a dynamic exponent, it can be shown that $O(t)$ has to obey time DSI according to [53]

$$O(t) = C t^{\alpha/z} F\left(\phi + \frac{\log(t)}{\log(a_1^z)} \right) \,, \tag{11}$$

where C and ϕ are constants. So, the conjecture given by Eq. (11) implies the existence of a logarithmic periodic modulation of time observables characterized by a time-scaling ratio τ given by

$$\tau = a_1^z \,, \tag{12}$$

see also Eq. (4). So, in the case of random walks on fractal media with DSI, Eqs. (7) and (8) have to be generalized according to Eq. (11) with $d_\mathrm{w} = z$.

Diffusion-limited Reactions

Reaction kinetics is influenced by the characteristic time of the processes involved. The rate of a heterogeneous reaction is often determined by the adsorption of the reactants from a fluid phase on the catalyst surface. For reactions occurring in a single phase the transport of the reactants and the time of reaction influence the overall reaction rate. Here, we focus on diffusion-limited reaction processes for which the transport time, given by the typical time required for the reactants to meet, is much longer than the reaction time. Since the diffusion of particles in fractals is often anomalous, it is expected that this behavior will affect diffusion-limited reactions. Furthermore, one also needs to account for density fluctuations of the reactants occurring at all length scales. For these reasons, as well as for the occurrence of many particle effects, the study of diffusion-limited reactions is difficult and often eludes straightforward mean-field approaches.

In this section, simple and archetypical reactions occurring in fractal media will be discussed. For a more general overview of diffusion-limited reactions in homogeneous and disordered media, see e. g. [33,67].

One-Species Reactions

So far, the most studied case is the one-species annihilation process, where species A diffuse and annihilate upon encounter [33,67], according to the reaction scheme

$$A + A \rightarrow 0 \,, \tag{13}$$

such that the reaction is instantaneous. This example also includes the case in which the product 0 is some inert particle that does not influence the overall kinetics. A closely related reaction is the one-species coalescence process given by [33,67]

$$A + A \rightarrow A \,. \tag{14}$$

In the mean-field limit, which holds for the reaction-limited case, the rate equation for both reactions (13) and (14) is given by

$$\frac{\mathrm{d}\rho_A(t)}{\mathrm{d}t} = -K\rho_A^X(t) \,, \tag{15}$$

where $\rho_A(t)$ is the concentration of A-particles at time t and K is the reaction constant. Eq. (15), where $X = 2$ is the reaction order, is the typical second-order rate equation that can be solved yielding

$$\rho_A(t) = \frac{1}{kt + \frac{1}{\rho_A(0)}} \,, \tag{16}$$

so that in the long-time limit one has $\rho_A(t) \approx t^{-1}$. On the other hand, for the diffusion-controlled regime, it can be proved rigorously that the concentration decays, for both processes (13) and (14) according to [33,67]

$$\rho_A(t) \propto \frac{1}{t^{d/2}} \,, \quad d < d_\mathrm{c} = 2 \,, \tag{17}$$

where $d_\mathrm{c} = 2$ is the upper critical dimension, such that for $d > d_\mathrm{c}$ the density decays according to the mean-field prediction, given by Eq. (16). So, for $d < 2$ the kinetics of the diffusion-controlled process is anomalous and just at d_c one has logarithmic corrections to Eq. (17).

So, let us focus our attention on heterogeneous chemical reactions occurring in media having a fractal structure with $d_\mathrm{f} < d_\mathrm{c} = 2$. Typical examples for these structures are catalysts, porous glass [12], diffusion-limited aggregates [107], percolation clusters [79], zeolites, etc. Since, as already discussed, diffusion in such media is anomalous [14,96], so are expected to be diffusion-limited chemical reactions [13,15,17,18,45,80,92,103]. Then, the simple textbook elementary reaction given by Eq. (15) has to be modified. A straightforward way to understand the nature of the modifications can be envisioned just by considering that S_N (see Eq. (8)) is the exploration space of the random walk, so that the rate constant K can be written in terms of the visitation efficiency γ [80] as follows

$$K \propto \gamma \equiv \frac{\mathrm{d}S_N}{\mathrm{d}t} \propto t^{d_\mathrm{s}/2-1} \,. \tag{18}$$

Of course, for standard diffusion one has $d_\mathrm{s} = 2$, K is actually a constant and the classical reaction order $X = 2$ is recovered. However, it should be noticed that for anomalous diffusion one has $d_\mathrm{s} < 2$, so that K is no longer a constant but depends on time. In fact, by inserting Eq. (8) into Eq. (18) it follows that

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} \propto -t^{d_\mathrm{s}/2-1}\rho^2 \,, \quad t \rightarrow \infty \,, \tag{19}$$

and the integrated rate equation that is obtained by using Eq. (19) reads

$$\rho^{-1} - \rho_o^{-1} = kt^{d_\mathrm{s}/2} \,, \tag{20}$$

where $\rho_o \equiv \rho(t = 0)$ is the initial density of random walks and, of course, Eqs. (19) and (20) are expected to hold in the low density limit since they are derived from the behavior of a single walker. Now, by comparing Eqs. (8) and (20), it follows that both the asymptotic (low-density) regime of the density of species in the diffusion-limited reaction and the exploration space of the random walk are

governed by the same exponent, given by the spectral dimension of the media where the physical process actually takes place. Furthermore, by replacing Eq. (20) into Eq. (19), it follows that

$$\frac{d\rho}{dt} = -k\rho^{1+2/d_s}\,, \quad t \to \infty, \tag{21}$$

and the reaction order for the elementary reaction given by Eq. (15) can be written as [15,17,45,81,92]

$$X = 1 + \frac{2}{d_s}\,, \tag{22}$$

which yields the textbook result ($X = 2$) for classical diffusion, but one has $X > 2$ for anomalous diffusion that is expected to occur in low-dimensional and fractal media for $d_s < 2$, e. g. for $d = 1$ one has $d_s = 1$ and the reaction order is $X = 3$. This relationship for the reaction order, based on the conjecture on the time dependence of the effective rate constant given by Eq. (18), was later derived rigorously by Clément et al. [16] based on calculations of the pair-correlation function and the macroscopic reaction law for Euclidean dimensions $d \leq 3$ and self-similar fractal structures with spectral dimensions $1 \leq d_s \leq 2$.

Figure 3a shows the time dependence of the decay of the density of random walks as a result of the reaction given by Eq. (20). The results correspond to the Sierpinski Gasket SG(5,10). A power-law decay with exponent $d_s/2 = 0.661 \pm 0.003$ is observed in agreement with Eq. (20). However, a careful inspection of the curve allows us to distinguish soft log-periodic oscillations modulating the power-law behavior. These oscillations become even more evident in Fig. 3b, which was obtained from the original data already shown in Fig. 3a after subtraction and subsequent normalization by the fitted power law. The presence of log-periodic oscillations is the signature of the occurrence of *time* DSI (see Eq. (11)) due to the coupling of the reaction kinetics and the *spatial* DSI of the substrate (see Eq. (4)), as discussed in Sect. "Fractals and Some of Their Relevant Properties" and "Random Walks".

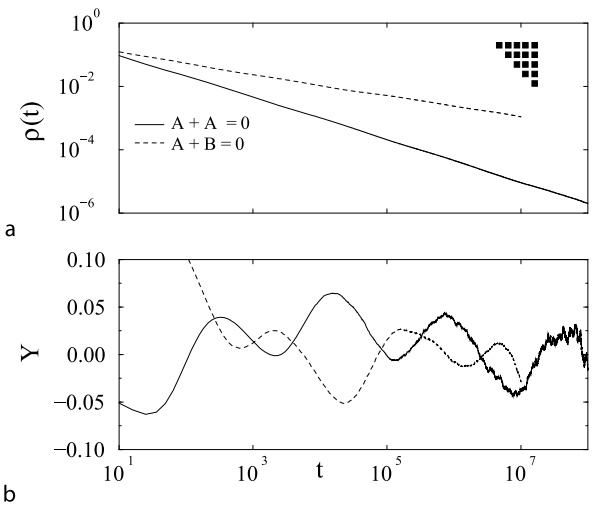

Reaction Kinetics in Fractals, Figure 3
a Log-log plots of the decay of the density of reacting species as obtained by means of Monte Carlo simulations in the SG(5,10) whose generating cell is shown in the *upper-right corner*. b Linear-log plot of the raw data shown in a but after proper subtraction and subsequent normalization by the corresponding power-law behavior fitted in a. In both panels, the *dashed* and *full lines* correspond to the reactions $A + B \to 0$ and $A + A \to 0$, respectively. More details in the text

Two-Species Annihilation

The study of the two-species diffusion-limited reaction of the type

$$A + B \to 0 \tag{23}$$

has received considerable attention due to the fact that, most likely, it is the simplest case where fluctuations in the local concentration of the reactants play a prevailing role in the emerging kinetic behavior [33,38,43,67,76,103]. According to Eq. (23) the concentration difference of both species is conserved, so that $\rho_A(t) - \rho_B(t) = \rho_A(0) - \rho_B(0) \equiv$ const. If the initial concentrations are different, the mean-field approach predicts an exponential decrease of the minority species [33,67]. However, for $\rho_A(0) = \rho_B(0)$ one expects an algebraic decay of the form $\rho \propto 1/t$.

Let us now briefly review the effect caused by spatial fluctuations in the concentrations [33,38,43,67,76,103]. For the sake of simplicity, let us assume that both species have the same diffusion constant (D) and that the initial concentrations are equal. For a random initial distribution of particles, a region of the space of linear size l would initially have $N_A = \rho_A(0)l^d \pm \sqrt{\rho_A(0)l^d}$ particles of type A, and equivalently for B-particles. The second term accounts for the local fluctuation, at a scale of the order of l, in the number of particles. By assuming classical diffusion, one has that for a time t, such as $Dt \simeq l^{1/2}$, all the particles within the considered region would have time enough to react with each other. In the absence of fluctuations, the concentrations of both species will vanish. However, at the end of the reaction process one still has of the order of $\sqrt{\rho(0)l^d}$ particles of the majority species. Then, the concentration of either A- or B-particles at time t is actually

given by

$$\rho_{A,B} \propto \frac{\sqrt{\left(\rho_{A,B}(0) l^d\right)}}{l^d} \propto \frac{\sqrt{\left(\rho_{A,B}(0)\right)}}{D t^{d/4}} , \tag{24}$$

and the decay exponent is twice as small than in the case of the one-species processes (Eq. (17)). According to the above scaling reasoning, one would expect the formation of alternating domains of A- and B-particles. This segregation of the reactants, known as the Toussaint–Wilczek effect [103], causes the slowdown of the kinetics since the reaction actually takes place along the boundary between domains. A careful analysis reveals that for $d \leq 4$, the domains are unstable, so that $d_c = 4$ is the upper critical dimension above which the mean-field approach holds. So, Eq. (24) holds for $d \leq 4$, and segregation has qualitatively been observed, e. g. in $d = 2$, by means of simulations [103] (see also Fig. 3). In order to gain a quantitative insight into the segregation of the reactants, it is useful to evaluate the excess density of either A- or B-particles within spatial domains of size l^d, defined as

$$\Delta\rho_{A,B}(l) = \frac{|N_A - N_B|}{l^d} , \tag{25}$$

where $N_A(N_B)$ is the number of A- (B-) particles within the domain.

Figure 4 shows log-log plots of the time dependence of the decay of the density of particles, as well as that of the excess density, as measured for different values of l. Results are obtained by means of Monte Carlo simulations in $d = 2$ dimensions. Three different regimes can be observed, namely: (i) For the short-time behavior ($t \leq 10^3$ in Fig. 4) one observes deviations from Eq. (24), which are in agreement with the fact that at early times the segregation of the reactants, as evidenced by the measurements of the excess density, is almost negligible. (ii) During an intermediate time regime ($10^3 < t \leq 10^6$ in Fig. 4) the Toussaint–Wilczek effect becomes dominant – the excess densities are maxima in all measured length scales – and the density decays according to Eq. (24) with exponent $d/4 = 1/2$. (iii) Finally, for long times ($t > 10^6$ in Fig. 4) one observes the fast decay of both the density and the excess density due to the finite size of the sample used.

Let us now analyze the expected behavior for the case of reactions occurring in fractals. Here one has that the characteristic scaling time is of the order of $t \propto l^{1/d_w}$, and according to Eq. (24) the residual concentration can be written as

$$\rho_{A,B}(t) \propto \frac{l^{d_f/2}}{l^{d_f}} \propto \frac{t^{d_f/2d_w}}{t^{d_f/d_w}} \propto \frac{1}{t^{d_s/4}} , \tag{26}$$

where Eq. (10) has been used. Therefore, for fractals one has to replace d by the spectral dimension d_s of the fractal in Eq. (24).

Figure 5 shows log-log plots of the time dependence of the decay of the density of particles and the excess density, as in the case of Fig. 4, but obtained by means of Monte Carlo simulations on random-fractal substrates given by percolation clusters at the critical threshold (see

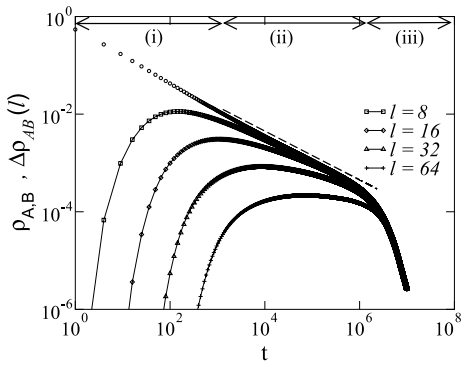

Reaction Kinetics in Fractals, Figure 4

Log-log plots of the density of species (o) and the excess density measured for different length scales as indicated. Results obtained for the $A + B \rightarrow 0$ reaction (Eq. (23)) in $d = 2$ dimensions by using a lattice of side $L = 4096$. Results are averaged over 100 different realizations. The *double-arrow lines* at the *top* indicate the three different time regimes discussed in the text. The length scales used to measure the excess density are identified by the corresponding symbols. The *dashed line*, with slope $d/4 = 1/2$ shows the expected behavior according to Eq. (24)

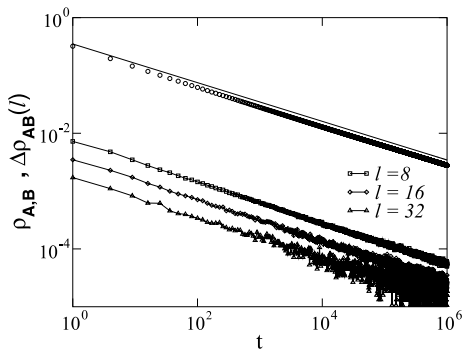

Reaction Kinetics in Fractals, Figure 5

Log-log plots of the density of species (o) and the excess density measured for different length scales as indicated. Results obtained for the $A + B \rightarrow 0$ reaction (Eq. (23)) in incipient percolation clusters in $d = 2$ dimensions by using a lattice of side $L = 4096$. Results are averaged over 100 different realizations. The length scales used to measure the excess density are identified by the corresponding symbols. The *full line* (slightly shifted *up* for the sake of clarity), with slope $d_s/4 = 0.3344$, is the best fit of the data according to Eq. (26)

Fig. 2). In contrast to the $d = 2$ dimensional case discussed within the context of Fig. 4, the disordered nature of the fractal causes the occurrence of density fluctuations at all the spatial scales even for the initial configuration. So, the Toussaint–Wilczek effect starts to play a dominant role from the very beginning of the reaction. Consequently, the density decays according to Eq. (26) along the whole time interval measured. The best fit of the data yields $d_s/4 = 0.3344$, in excellent agreement with reported values [20]. Of course, finite-size effects, which are not observed in Fig. 5, could be expected to occur at later times. It is also worth mentioning that Fig. 5 confirms that the excess density also decays with the same power-law behavior as the density, in agreement with the arguments outlined in order to obtain Eqs. (24) and (26).

Figure 3a also shows the time dependence of the decay of the density of particles for the $A + B \to 0$ reaction given by Eq. (26). These results were obtained in the SG(5,10). The best fit of the data gives $d_s/4 = 0.331 \pm 0.003$, in excellent agreement with the value obtained for the annihilation reaction (Eq. (20)) given by $d_s/2 = 0.663 \pm 0.003$ (see also Fig. 3a). The soft log-periodic oscillation of the density that can be observed in Fig. 3a is enhanced in Fig. 3b. Again, this behavior is the fingerprint of the occurrence of *time* DSI, as discussed in Sect. "Fractals and Some of Their Relevant Properties" and "Random Walks" (see Eqs. (4) and (11), respectively).

Summing up, for the simplest example of a bimolecular irreversible decay of the form $A + B \to 0$ the obtained universal kinetic law depends on the dimensionality, the initial concentration and the particle-conservation law that holds for the system.

So far, the transient $A + B \to 0$ reaction is fundamentally different from the $A + A \to 0$ reaction due to the Toussaint–Wilczek effect. This effect represents a quite delicate balance that may become elusive to be properly identified, as discussed for the case of $d = 2$ dimensions (Fig. 4). Therefore, it is interesting to study its persistence under tiny perturbations, e. g. the steady feeding of reactants. Early simulations by Anacker et al. [93] have revealed that, in a cubic lattice, the Toussaint–Wilczek effect is destroyed by the addition of a steady source of walkers. Furthermore, in contrast to the transient result where the reaction order is $X = 3$, the steady-state $A + B \to 0$ reaction gives $X = 2.00 \pm 0.02$, for steady-state densities as low as 0.001. Since in $d = 2$ and for the transient regime, the occurrence of the effect is confined to a narrow density window (see Fig. 4) careful measurements will be needed in order to test if segregation still remains relevant in $d = 2$ under steady-state conditions. Segregation of the reactants under steady-state conditions has also been observed to occur in percolation clusters [93,100]. Furthermore, dramatic segregation effects are observed in the Sierpinski Gasket under steady-state conditions [93,100]. In this case the value $X = 2.00 \pm 0.2$ has been reported for the reaction order. All these results point out that steady-state segregation cannot simply be due to the Toussaint–Wilczek effect and the understanding of this behavior remains elusive, even after more than 20 years.

Irreversible Phase Transitions in Heterogeneously Catalyzed Reactions

Classical phase transitions, such as those observed in magnets, fluids, alloys, etc., which take place under equilibrium conditions, are reversible [28,35,111]. For example, one can change the phase of a magnet, from the ferromagnetic to the paramagnetic one, just by properly tuning the temperature. In contrast, irreversible phase transitions (IPTs) [9,10,68,85,90] occurring in reaction systems take place between an active state, characterized by a steady reaction among the reactants and the output of the products, and an inactive – also called poisoned or absorbing – state, where the reaction stops irreversibly. The inactive state is absorbing in the sense that after undergoing trapping, the system can never escape from it. Consequently, it is impossible to change from the absorbing state to the active one, just by tuning the control parameter, and the transition is irreversible.

IPTs are typically studied by using lattice-gas reaction models of heterogeneously catalyzed reactions. Within this context IPTs observed upon the catalytic oxidation of CO, H_2 and NO have extensively been studied [42,85,88,108]. Further studies of IPTs comprise the generic monomer-monomer ($A + B \to 0$) reaction [90], as well as a wide variety of reaction processes such as epidemic propagation [84], forest-fire models [8,97], prey-predator systems [51], etc., for reviews see e. g. [68]. Since a wide variety of catalysts are composed by a solid-state fractal support containing catalytically active particles [21,22,23,95], most of the above-mentioned reactions have also been studied in fractal media [1,2,5,7,41,72,86,87]. A brief overview of the state of the art in the characterization of IPTs in fractal substrates will be presented below.

The ZGB Model for the Catalytic Oxidation of Carbon Monoxide

It is well known that the catalytic oxidation of carbon monoxide, namely $2CO + O_2 \to 2CO_2$, which is likely the most studied reaction system, proceeds according to the Langmuir–Hinshelwood mechanism [50,65], i.e with

both reactants adsorbed on the catalyst surface

$$CO(g) + S \rightarrow CO(a)\,, \tag{27}$$

$$O_2(g) + 2S \rightarrow 2O(a)\,, \tag{28}$$

$$CO(a) + O(a) \rightarrow CO_2(g)\,, \tag{29}$$

where S is an empty site on the surface, while (a) and (g) refer to the adsorbed and gas phase, respectively. The reaction takes place with the catalyst, e. g. a transition metal surface such as Pt, in contact with a reservoir of CO and O_2 whose partial pressures are P_{CO} and P_{O_2}, respectively.

Equation (27) describes the irreversible molecular adsorption of CO on a single site of the catalyst surface. It is known that under suitable temperature and pressure reaction conditions, CO molecules diffuse on the surface. Furthermore, there is a small probability of CO desorption that increases as the temperature is raised [71]. Equation (28) corresponds to the irreversible adsorption of O_2 molecules that involves the dissociation of such species and the resulting O atoms occupy two sites of the catalytic surface. Under reaction conditions, both the diffusion and the desorption of oxygen are negligible. Due to the high stability of the O_2 molecule, the whole reaction does not occur in the homogeneous phase due to the lack of O_2 dissociation. So, Eq. (28) dramatically shows the role of the catalyst that makes the rate-limiting step of the reaction feasible. Finally, Eq. (29) describes the formation of the product (CO_2) that desorbs from the catalyst surface. This final step is essential for the regeneration of the catalytically active surface.

For the practical implementation of the Ziff–Gulari–Barshad (ZGB) model [108], the catalyst surface is replaced by a lattice, so a lattice-gas reaction model is actually considered. The catalyst is assumed to be in contact with an infinitely large reservoir containing the reactants in the gas phase. Adsorption events are treated stochastically neglecting energetic interactions. Furthermore, diffusion and desorption of CO are also neglected. Then, on the two-dimensional square lattice the Monte Carlo simulation algorithm for the ZGB model is as follows:

(i) CO or O_2 molecules are selected at random with relative probabilities P_{CO} and P_O, respectively. These probabilities are the relative impingement rates of both species, which are proportional to their partial pressures in the gas phase in contact with the catalyst. Due to the normalization, $P_{CO} + P_O = 1$, the model has a single parameter, i. e. P_{CO}. If the selected species is CO, one surface site is selected at random, and if that site is vacant, CO is adsorbed on it according to Eq. (27). Otherwise, if that site is occupied, the trial ends and a new molecule is selected. If the selected species is O_2, a pair of nearest-neighbor sites is selected at random and the molecule is adsorbed on them only if they are both vacant, as required by Eq. (28).

(ii) After each adsorption event, the nearest-neighbors of the added molecule are examined in order to account for the reaction given by Eq. (29). If more than one [CO(a), O(a)] pair is identified, a single one is selected at random and removed from the surface.

Assuming irreversible adsorption-reaction steps, as in the case of Eqs. (28)–(29), it may be expected that in the limit of large P_{CO} and small P_{O_2} (small P_{CO} and large P_{O_2}) values, the catalyst surface would become saturated by CO- (O_2-) species and the reaction would stop. In fact, the catalyst surface fully covered by a single type of species, where further adsorption of the other species is no longer possible, corresponds to an inactive state of the system. This state is known as 'poisoned', in the sense that adsorbed species on the catalyst surface are the poison that causes the reaction to stop. Physicists used to call such a state (or configuration) 'absorbing' because a system becomes trapped by it forever, with no possibility of escape [49]. These concepts are clearly illustrated in Fig. 6a and b, which show plots of the rate of CO_2 production (R_{CO_2}) and the surface coverage with CO and O_2 (θ_{CO} and θ_O, respectively), versus the partial pressure of CO (P_{CO}), as obtained by simulating the ZGB model [108]. Figure 6a corresponds to simulations performed in homogeneous, two-dimensional lattices. In this case, for $P_{CO} \leq P_{1CO} \simeq 0.38975$ the surface becomes irreversibly poisoned by O species with $\theta_{CO} = 0$, $\theta_O = 1$ and $R_{CO_2} = 0$. In contrast, for $P_{CO} \geq P_{2CO} \simeq 0.525$ the catalyst is irreversibly poisoned by CO molecules with $\theta_{CO} = 1$, $\theta_O = 0$ and $R_{CO_2} = 0$. However, as shown in Fig. 6a, between these absorbing states there is a reaction window, namely for $P_{1CO} < P_{CO} < P_{2CO}$, such that a steady state with sustained production of CO_2 is observed.

As follows from Fig. 6a, when approaching the oxygen absorbing state from the reactive phase, all quantities of interest change smoothly until they adopt the values corresponding to the absorbing state. This behavior typically corresponds to a second-order IPT [9,10,68,85,90]. Remarkably, the behavior of the system is quite different upon approaching the CO absorbing state from the reactive regime (see Fig. 6a). In this case, all quantities of interest exhibit a marked discontinuity close to $P_{2CO} \simeq 0.525$. This is a typical first-order IPT and P_{2CO} is the coexistence point.

Experimental results for the catalytic oxidation of carbon monoxide on Pt(210) and Pt(111) [3,36] are in qualitative agreement with simulation results of the ZGB

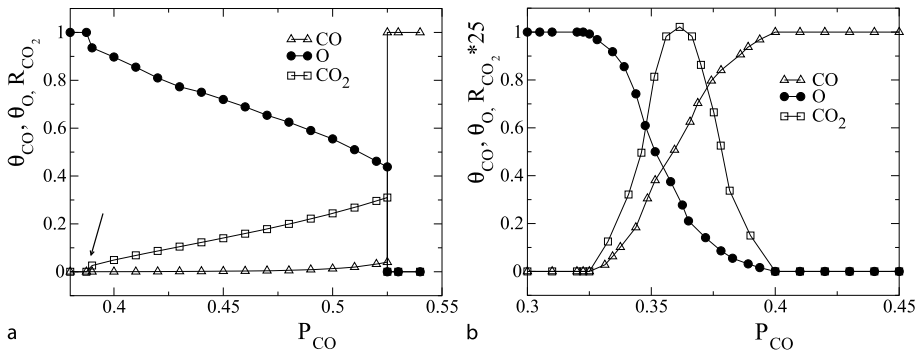

Reaction Kinetics in Fractals, Figure 6

Phase diagrams of the ZGB model showing the dependence of the surface coverage with CO (θ_{CO}, △) and Oxygen (θ_O, •), and the rate of CO_2 production (R_{CO_2}, ⋄) on the partial pressure of CO (P_{CO}) in the gas phase. (a) Results obtained in $d = 2$ homogeneous lattices. The irreversible phase transition (IPT) occurring at $P_{1CO} \simeq 0.38975$ (second-order) is shown by an *arrow*, while at $P_{2CO} \simeq 0.525$ one has a sharp (first-order) IPT. (b) Results obtained by performing simulations in incipient percolation clusters of fractal dimension $d_f \simeq 1.89$ in $d = 2$ dimensions. Second-order IPTs occurring at $P_{1CO} \simeq 0.325$ and $P_{2CO} \simeq 0.400$ can clearly be observed. Notice that the rate of CO_2 production has been amplified by a factor 25 for the sake of clarity

model. A remarkable agreement is the (almost) linear increase in the reaction rate observed when the CO pressure is raised and the abrupt drop of the reactivity when a certain 'critical' pressure is reached. However, two essential differences are worth discussing: (i) the oxygen-poisoned phase exhibited by the ZGB model within the CO low-pressure regime is not observed experimentally. (ii) The CO-rich phase exhibiting low reactivity found experimentally resembles the CO-poisoned state predicted by the ZGB model. However, in the experiments the non-vanishing CO-desorption probability prevents the system from entering into a truly absorbing state and the abrupt, 'first-order-like' transition, observed in the experiments is actually reversible. Of course, these and other disagreements are not surprising since the lattice gas reaction model, with a single parameter, is a simplified approach to the actual catalytic reaction that is far more complex.

In order to observe the influence caused by the fractal nature of the substrate on the phase diagram of the ZGB model, Fig. 6b shows results obtained by means of simulations performed in incipient percolating clusters at the critical threshold in two dimensions (see also Fig. 2). In this case, for $P_{CO} \leq P_{1CO} \simeq 0.325$ the surface becomes irreversibly poisoned by O species, while for $P_{CO} \geq P_{2CO} \simeq 0.400$ the catalyst is irreversibly poisoned by CO molecules [2]. Also, a reaction window between these absorbing states is found. So, by comparing Fig. 6a and b at least three main differences can be identified [2]: (i) The IPT observed at a low CO pressure in homogeneous lattices becomes largely shifted towards lower pressures in the case of the IPC. This effect is due to the constraint imposed by the disordered cluster on the adsorption of oxygen molecules that need two adjacent sites for deposition. (ii) The sharp first-order IPT characteristic of the homogeneous lattice becomes of second-order for the case of the fractal. (iii) The reaction window, which is of the order of $\Delta P_{CO} = P_{2CO} - P_{1CO} \simeq 0.135$ in the homogeneous lattice, becomes narrowed up to $\Delta P_{CO} \simeq 0.085$ for the case of the fractal. These results are in agreement with the findings of Casties et al. [41], who also studied the influence of CO diffusion on the location of the critical points. In particular, it is found that P_{2CO} increases when the rate of CO-diffusion also increases.

On the other hand, simulations of the ZGB model performed in disordered clusters below the critical threshold, i. e. in the so-called 'lattice animals' in percolation [101], show that poisoning can be caused either by CO- or O-species [41,91] and that due to finite-size effects the finite-width reaction window vanishes for lattices of side $L \simeq 20$ [41,91]. These observations are consistent with the fact that the characteristic lengths of the fluctuations in the coverages are larger than the sizes of the animals, playing a dominant role in the catalytic behavior of small aggregates, such as those used in actual catalysts [21,75,106].

The common feature of Monte Carlo simulations of the ZGB model performed in Sierpinski Carpets is that the first-order IPT characteristic of the homogeneous lattice is no longer observed, so that transition becomes of second-order [5,41]. In fact, even for the $SC(5,1)$ with fractal dimension $d_f = \log(24)/\log(25) \simeq 1.975$ one observes a second-order IPT at high CO pressures [5]. Simulations performed in the $SC(3,1)$ with a fractal dimension $d_f \simeq 1.89$, which is very close to that of the IPC ran-

dom fractal, yield $P_{1CO} \simeq 0.397$ and $P_{2CO} \simeq 0.52$ [41], i. e. two critical points that are very close to those measured in $d = 2$ homogeneous media (see e. g. Fig. 6a) but far from the results obtained for the IPC (see e. g. Fig. 6b). These findings point out that the location of the critical points depends on topological properties of the fractals that are not fully accounted for a single fractal dimension. Also, the finite-width reaction window, which lies between both second-order IPTs, becomes narrowed when decreasing d_f [5,41]. It has even been suggested that for Sierpinski Carpets the reaction window may become of zero width close to $d_f \approx 1.6$ [5]. This statement is in agreement with the fact that Casties et al. [41] have reported that the reaction window of the ZGB model in an SC with $d_f = 1.59$ is very narrow, i. e. $\Delta P_{CO} \simeq 0.02$ It is also worth mentioning that, in $d = 1$, the ZGB model lacks a finite-width reaction window.

For second-order IPTs, as in the case of their reversible counterparts, it is possible to define an order parameter, which for the former is given by the concentration of minority species (θ_{CO}, in the case of the second-order IPT of the catalytic oxidation of CO). Furthermore, it is known that θ_{CO} vanishes according to a power law upon approaching the critical point [64], so that

$$\theta_{CO} \propto (P_{CO} - P_{1CO})^{\beta} \ , \tag{30}$$

where β is the order parameter critical exponent and P_{1CO} is the critical point. Careful studies performed in the $d = 2$ homogeneous lattice [63,73,104] have confirmed that this IPT belongs to the universality class of directed percolation [62,82] in $d + 1$ dimensions.

Generic Lattice-Gas Reaction Models of Catalyzed Reactions

The simplest reaction system studied proceeding according to the Langmuir–Hinshelwood mechanism [50,65] is the monomer-monomer (MM) model. In the MM model A- and B-species adsorb, react and desorb according to the following scheme

Adsorption:

$$A(g) + S \to A(a) \ , \tag{31}$$

$$B(g) + S \to B(a) \ , \tag{32}$$

Reaction:

$$A(a) + B(a) \to AB(g) + 2S \ , \tag{33}$$

Desorption:

$$A(a) \to A(g) + S \ , \tag{34}$$

$$B(a) \to B(g) + S \ , \tag{35}$$

where S is an empty site on the surface, while (a) and (g) refer to the adsorbed and gas phase, respectively. Adsorption of A- and B-species takes place with probability p_A and p_B, respectively. So, by taking $p_A + p_B = 1$ one has a single parameter for the adsorption process, given by $p_{ad} \equiv p_A = 1 - p_B$. Also the rates of reaction and desorption of A- and B-species are k_R, P_{dA} and P_{dB}, respectively.

By properly selecting the parameters one can study different regimes of the MM model. Let us first consider the *adsorption-limited* regime that corresponds to $k_R \equiv \infty$ and $P_{dA} = P_{dB} \equiv 0$. Here, for $p_{ad} \neq 1/2$ one has that any sample becomes irreversibly poisoned by the majority species and the final states are always absorbing. However, for $p_{ad} = 1/2$ a slow coarsening process with the formation of solid domains of A- and B-species on the surface of the catalyst is observed. By mapping the MM model into a kinetic Ising model, it can be proved that any finite sample of N_s sites will ultimately become irreversibly poisoned due to fluctuations in the coverages after a time T_P of the order of $T_P \propto N_s \ln(N_s)$ [32]. This exact result has early been predicted by means of numerical simulations by ben-Avraham et al. [4].

On the other hand, by considering desorption of a single species, say $P_{dA} \equiv 0$ and $P_{dB} > 0$, one has that the MM model exhibits a second-order IPT between an active state and a single poisoned state with A-species [6]. The critical points are found to be $P_{dB} = 0.5099 \pm 0.0003$, $P_{dB} = 0.6150 \pm 0.0003$ and $P_{dB} = 0.5562 \pm 0.0003$ for the case of $d = 2$, $d = 1$ homogeneous substrates and the IPC in $d = 2$ [54], respectively.

Interesting results are also obtained by taking $P_d \equiv P_{dA} = P_{dB}$, as proposed by Fichthorn et al. [46]. In fact, by means of numerical simulations, they have shown that for that choice of parameters the MM with desorption exhibits a noise-induced transition from monostability to bistability. This result was subsequently also obtained by means of a mean-field approach [44]. Later on, the occurrence of the transition was proved rigorously [32]. On the other hand, for a fractal lattice and in the limit of $P_d \to 0$ it has been conjectured that any finite sample of N_s sites would become saturated by either A- or B-species when

$$\chi \equiv P_d^{d_s/2} N_s \simeq 1 \tag{36}$$

where d_s is the spectral dimension of the underlying fractal and χ is a reduced parameter [55]. Furthermore, for an intermediate rate of desorption, a simple scaling behavior, which depends on the spectral dimension, is expected. In fact, for $\chi \gg 1$, a segregation regime with a reaction rate

per site $R(P_d)$ given by

$$R(P_d) \simeq P_d^{1-d_s/2} \tag{37}$$

has been predicted [55]. Also, for $\chi \ll 1$, one expects the saturation of any finite cluster due to a fluctuation-dominated regime, and the rate of reaction is expected to depend on the cluster size according to

$$R(P_d) = N_s P_d \,. \tag{38}$$

All these analytical results, namely equations (36), (37) and (38), have been validated by means of numerical simulations by using IPCs in $d = 2$ and $d = 3$ Euclidean dimensions [55] as fractal substrates.

Another interesting generic system is the dimer-dimer (DD) lattice-gas reaction model, which is described by the following Langmuir–Hinshelwood scheme

Adsorption:

$$A_2(g) + 2S \to 2A(a) \,, \tag{39}$$

$$B_2(g) + 2S \to 2B(a) \,, \tag{40}$$

Reaction:

$$A(a) + B(a) \to AB(g) + 2S \,, \tag{41}$$

where S is an empty site on the surface, while (a) and (g) refer to the adsorbed and gas phase, respectively. In this case we are interested in the reaction-controlled regime, so that the adsorption of either A_2- or B_2-species (see e. g. Eqs. (39) and (40)) takes place with the same probability and proceeds instantaneously on every pair of empty sites left behind by the reaction (Eq. (41)) [37]. It is worth mentioning that the DD model can be *exactly* mapped onto the so-called "voter" model [37]. The voter model is a spin system with two possible orientations, namely spin-up and spin-down. In the voter model, a randomly selected spin adopts the orientation of a randomly selected neighbor. Of course, if a spin is surrounded by equally oriented spins, it does not change its orientation. So, the voter dynamics is zero-temperature in nature. In fact, the voter model defines a broad universality class that describes coarsening without surface tension [48]. The voter model is also a simple model for opinion formation [57]: individuals with different opinions, labeled by A and B, respectively, change their opinions according to the voter dynamics. A remarkable feature of the DD model, or equivalently the voter model, is that it is one of the very few models that can be solved exactly in any integer dimension d.

The order parameter of the voter model is the density of interfaces with different states of opinion (C_{AB}), which corresponds to the density of A-B interfaces in the DD model. Starting form a fully disordered configuration that maximizes the interfaces, it has been found that the the density of interface evolves according to [37]

$$C_{AB}(t) = \begin{cases} t^{-\alpha} \,, & d < 2 \\ (\ln(t))^{-1} \,, & d = 2 \\ 1 \,, & d > 2 \end{cases} \tag{42}$$

where $\alpha = 1 - d/2$. So, in the long-time regime on has that $C_{AB} \to 0$ when $d < 2$, and the coarsening of domains is observed. However, for $d > 2$ single-species domains do not appear. The borderline is the two-dimensional case where coarsening proceeds logarithmically, so that $d_c = 2$ is the upper critical dimension of the voter model.

Monte Carlo simulations of the voter model performed in both Sierpinski Gaskets and Carpets with $d_f < 2$ have confirmed the coarsening of the domains. An interesting feature observed in the plots of $C_{AB}(t)$ versus t, as obtained for various fractals, is the presence of soft log-periodic modulations of the power-law decay predicted by Eq. (42) [53,102]. As already discussed in Sect. "Fractals and Some of Their Relevant Properties" and "Random Walks" (see Eqs. (4) and (11), respectively), these oscillations are the signature of *time* discrete scale invariance caused by the coupling of the voter dynamics with the *spatial* discrete scale invariance of the underlying fractal.

Epidemic Simulations for the Characterization of IPTs

The study of second-order IPTs by using standard simulation methods is hindered by the fact that, due to large fluctuations occurring close to critical points, any finite system will ultimately become irreversibly trapped by the absorbing state. So, the measurements are actually performed within metastable states facing two competing constraints: on the one hand, the measurement time has to be long enough in order to allow the system to develop the corresponding correlations and, on the other hand, such time must be short enough to prevent poisoning of the sample. In view of these shortcomings, experience indicates that the best approach to second-order IPTs is to complement the standard approach, consisting in performing finite-size scaling studies of stationary quantities, with epidemic simulations. In fact, the application of the Epidemic Method (EM) to the study of IPTs has become a useful tool for the evaluation of critical points, dynamic critical exponents and eventually for the identification of universality classes [62,82,104]. The idea behind the EM is to initialize the simulation using a configuration very close to the absorbing state. Such a configuration can be achieved by removing some species from the center of an otherwise

fully poisoned sample. Then a small patch of empty sites is left. In the case of the ZGB model, this can be done by filling the whole lattice with O, except for a small patch. Patches consisting of 3–6 neighboring empty sites are frequently employed, but it is known that the asymptotic results are independent of the size of the initial patch. Such patch is the kernel of the subsequent epidemic.

After the generation of the starting configuration, the time evolution of the system is followed, and during this dynamic process the following quantities are recorded: (i) the average number of empty sites ($N(t)$), (ii) the survival probability $P(t)$, which is the probability that the epidemic is still alive at time t, and (iii) the average mean square distance, $R^2(t)$, over which the empty sites have spread. Of course, each single epidemic stops if the sample is trapped in the poisoned state with $N(t) = 0$ and, since these events may happen after very short times (depending on the patch size), results have to be averaged over many different epidemics. It should be noticed that $N(t)$ ($R^2(t)$) is averaged over all (surviving) epidemics.

If the epidemic is performed just at criticality, a power-law behavior (scaling invariance) can be assumed and the following *Ansätze* are expected to hold,

$$N(t) \propto t^{\eta} \,, \tag{43}$$

$$P(t) \propto t^{-\delta} \,, \tag{44}$$

and

$$R^2(t) \propto t^{z_{\mathrm{epi}}} \,, \tag{45}$$

where η, δ and z_{epi} are dynamic critical exponents. Thus, at the critical point log-log plots of $N(t)$, $P(t)$ and $R^2(t)$ will asymptotically show a straight line behavior, while off-critical points will exhibit curvature. This behavior allows the determination of the critical point, and from the slopes of the plots the critical exponents can also be evaluated quite accurately [62].

The validity of Eqs. (43), (44) and (45) for second-order IPTs taking place in substrates of integer dimensions ($1 \leq d \leq 3$) is very well established. Furthermore, the observation of a power-law behavior for second-order IPTs is in agreement with the ideas developed in the study of equilibrium (reversible) phase transitions: scale invariance reflects the existence of a diverging correlation length at criticality. For the ZGB model in $d = 2$ dimensions the best reported values of the exponents are [63]: $\eta = 0.2295 \pm 0.001, \delta = 0.4505 \pm 0.001$ and $z_{\mathrm{epi}} = 1.1325 \pm 0.001$. These values confirm that the IPT belongs to the universality class of directed percolation in $d + 1$ dimensions, where the values reported by Grassberger [82] are $\eta = 0.229 \pm 0.003, \delta = 0.451 \pm 0.003$ and $z_{\mathrm{epi}} = 1.133 \pm 0.002$.

Early epidemic studies of the ZGB model performed in both the IPC and the SC(3,1), with almost the same fractal dimension of $d_{\mathrm{f}} \simeq 1.89$, confirmed that the IPTs are of second-order [7]. The values of the critical exponents obtained for both types of fractals are almost the same, within error bars, and are given by $\eta \approx 0.23$ and $\delta \simeq 0.41$. These values are also in agreement with epidemic calculations of the contact process [84] performed by Jensen [72], who has reported $\eta = 0.235 \pm 0.010$ and $\delta \simeq 0.40 \pm 0.01$. Subsequently, Gao and Yang [110] have studied the influence of the lacunarity of the SC on the epidemic behavior of the ZGB model. The epidemic exponents are very different for lattices of the same fractal dimension but different lacunarity, suggesting that a single fractal dimension is not enough to fully specify the critical points and exponents of the critical process. This finding is in qualitative agreement with the same conclusion obtained by means of stationary measurements and discussed in the previous section.

It is worth mentioning that the better statistics and the longer measurement time achieved by Gao et al. [110], as compared e. g. to references [7,72], allow us to observe a subtle log-periodic modulation of the power-law behavior of some dynamic observables. This effect is particularly clear for the case of the time dependence of $N(t)$ and it is the signature of *time* discrete-scale invariance (DSI). In fact, since the underlying fractals used by Gao et al. [110] exhibit *spatial* DSI with a fundamental scaling ratio $a_1 = 4$ (see also the discussion on Eq. (4)), we expect that the dynamics of any critical process occurring in those fractals would become coupled to the spatial DSI [53] through the growing correlation length $\xi \propto t^{1/z}$, where z is a dynamic exponent. This coupling would ultimately lead to *time* DSI characterized by a soft log-periodic modulation of the power laws, according to Eq. (11). Then the characteristic time-scaling ratio τ is given by Eq. (12) with $z = 2/z_{\mathrm{epi}}$. Of course, the accurate evaluation of τ by fitting the reported epidemic measurements is no longer possible because it will require a careful fit of the data. However, based on the fact that $z_{\mathrm{epi}} \simeq 1$ [110], we predict $\tau \simeq 4^2 \simeq 16$.

The Constant Coverage Ensemble Applied to the Study of First-Order IPTs

It is well known that first-order transitions are characterized by hysteresis effects, long-lived metastabilities, etc., which make difficult their study by means of conventional simulation methods. In order to avoid these shortcomings,

Brosilow and Ziff [39] have proposed the constant coverage (CC) ensemble for the study of first-order IPTs. The method has early been applied to the ZGB model [39], and subsequently the same authors [11] have also performed CC simulations of the model proposed by Yaldran and Khan [88] for the NO + CO catalyzed reaction.

For the case of the ZGB model, the CC method can be implemented by means of the following procedure. First, a stationary configuration of the system is achieved using the standard algorithm, as described in Sect. "The ZGB Model for the Catalytic Oxidation of Carbon Monoxide". For this purpose, one selects a value of the parameter close to the coexistence point, e. g. $P_{\mathrm{CO}} = 0.51$. After achieving the stationary state the simulation is actually switched to the CC method. So, in order to maintain the CO coverage as constant as possible around a prefixed value $\theta_{\mathrm{CO}}^{\mathrm{CC}}$, only oxygen (CO) adsorption attempts take place whenever $\theta_{\mathrm{CO}} > \theta_{\mathrm{CO}}^{\mathrm{CC}}$, ($\theta_{\mathrm{CO}} < \theta_{\mathrm{CO}}^{\mathrm{CC}}$). Let $\mathcal{N}_{\mathrm{CO}}$ and $\mathcal{N}_{\mathrm{OX}}$ be the number of carbon monoxide and oxygen attempts respectively. Then, the value of the "pressure" of CO in the CC ensemble ($P_{\mathrm{CO}}^{\mathrm{CC}}$) is determined just as the ratio $P_{\mathrm{CO}}^{\mathrm{CC}} = \mathcal{N}_{\mathrm{CO}}/(\mathcal{N}_{\mathrm{CO}} + \mathcal{N}_{\mathrm{OX}})$. Subsequently, the coverage is increased by an small amount, say $\Delta\theta_{\mathrm{CO}}^{\mathrm{CC}}$. A transient period τ_P is then disregarded for the proper relaxation of the system to the new $\theta_{\mathrm{CO}}^{\mathrm{CC}}$ value, and finally averages of $P_{\mathrm{CO}}^{\mathrm{CC}}$ are taken over a certain measurement time τ_{M}. The CC ensemble allows to determine spinodal points and investigate hysteresis effects at first-order IPTs [11,29,39], which are expected to be relevant as compared with their counterpart in equilibrium (reversible) conditions [9,10,85].

Future Directions

In spite of the large effort involved in the study of the kinetics of reactions occurring in low-dimensional and fractal media, the subject still poses both theoretical and experimental challenges whose understanding would be of a wide interdisciplinary interest.

It is largely recognized that the development and characterization of new materials and nanostructures is essential in order to boost the technology of the present century. Within this context, one has that reaction kinetics in nanosystems and confined materials exhibits a quite interesting physical behavior that is very different from that of the bulk. The average number of reactants in a nanosystem is typically very small and eventually it may be of the order of the fluctuations. So, mean-field-like treatments are far from being useful and new analytical tools, often guided by numerical simulations, need to be developed. Relevant experiments in this field are the study of the relaxation dynamics of photoexited charge carriers in semiconducting nanotubes and nanoparticles [56]. Also, the presence of catalytically active micro- and nanoparticles could lead to a considerable enhancement of the reaction rate of heterogeneously catalyzed reactions, the exhaustive understanding of this phenomena being an open challenge [30,75,106].

During the 1990s physicists and mathematicians started the study of the properties of complex networks opening a new interdisciplinary branch of science that also involves biology, ecology, economics, social and computer sciences, and others. Early studies in the science of complex networks were aimed to discover general laws governing the creation and growth of such structures. Subsequently, growing interest has been addressed to the understanding of processes taking place in networks, such as e. g. the description of the behavior of random walkers, the characterization of qualitatively new critical phenomena, the study of the kinetics of reaction-diffusion processes, the discovery of the emergency of complex social behavior, etc. Due to the unique properties of the networks, which involve, among others, the combination of the compactness, their inherent inhomogeneity, their complex architectures, etc., reaction-diffusion processes in those media are essentially different from both classical and anomalous results already discussed. In fact, annihilation rates are abnormally high, no segregation is observed in the archetypical $A + B \to 0$ reaction, and depletion zones are absent in the annihilation $A + A \to 0$ reaction [69]. So, we expect that this will be an active field of interdisciplinary research in the study of reaction kinetics in complex media.

The recently formulated conjecture that the coupling between the spatial discrete scale invariance of fractal media and the dynamics and kinetics of physical processes occurring in those media may lead to the observation of time discrete scale invariance [53], opens new theoretical perspectives for the study of diffusion-limited reactions. This kind of study is further stimulated by recent reports suggesting that spatial discrete scale invariance is no longer restricted to deterministic fractals, but it may also emerge as a consequence of different growing mechanisms in some disordered fractals, such as e. g. diffusion-limited aggregates, animals in percolation, etc. [66,74]. In order to achieve some progress in this subject, of course, extensive simulations and careful measurements would be necessary because the phenomenon is quite subtle and the standard averaging methods may destroy it as if was simple noise [66].

The propagation of fronts has attracted considerable attention since its understanding is relevant in many areas of research and technology. Among others, some examples

of front propagation include: material growth and interfaces, forest-fire and epidemic propagation, particle-diffusion fronts, chemical pattern formation, biological invasion problems, displacement of an unstable phase by a stable one close to first-order phase transitions, etc. Within this context, the properties of reaction-diffusion fronts in simple reactions where the reactants are initially segregated, such as $A + B \rightarrow O$, have extensively been studied [31,47]. However, most of the studies addressed reaction-diffusion fronts occurring in homogeneous media while less attention has been devoted to the case of, increasingly important inhomogeneous media [98]. Inhomogeneities can have different origin: i. e. physical, chemical, geometrical, etc. It is expected that the study of front propagation generated by diffusion-limited reactions in fractal media would also be a promising field of future research, where anomalous diffusion may lead to the occurrence of very interesting phenomena. In a related context, theoretical studies of the diffusion of A- and B-species in the bulk of a confined (homogeneous) media, leading to reaction at the surface with desorption of the product, predict the development of an annihilation catastrophe [98]. This phenomenon is due to the self-organized explosive growth of the surface concentration of species leading to an abrupt desorption peak. The influence of anomalous diffusion in fractal media leading to surface reaction will certainly be addressed in the near future, hopefully deserving the observation of interesting physical phenomena.

Bibliography

Primary Literature

1. Albano EV (1990) Finite-size effects in kinetic phase transitions of a model reaction on a fractal surface: Scaling approach and Monte Carlo investigation. Phys Rev B 42:R10818–R10821
2. Albano EV (1990) Monte Carlo simulation of the oxidation of carbon monoxide on fractal surfaces. Surf Sci 235:351–359
3. Albano EV (1991) On the self-poisoning of small particles upon island formation of the reactants in a model for a heterogenoeusly catalyzed reaction. J Chem Phys 94:1499–1504
4. Albano EV (1992) Critical exponents for the irreversible surface reaction $A + B \rightarrow AB$ with B-desorption on homogeneous and fractal media. Phys Rev Lett 69:656–659
5. Albano EV (1992) Irreversible phase transitions in the dimer-monomer surface reaction process on fractal media. Phys Lett A 168:55–58
6. Albano EV (1992) Study of the critical behaviour of an irreversible phase transition in the $A + B \rightarrow AB$ reaction with B-desorption on a fractal surface- Phys Rev A 46:5020–5025
7. Albano EV (1994) Critical behaviour of the irreversible phase transitions of a dimer-monomer process on fractal media. J Phys A (Math Gen) 27:431–436
8. Albano EV (1995) Spreading analysis and finite-size scaling study of the critical behavior of a forest fire model with immune trees. Physica A 216:213–226
9. Albano EV (1996) The Monte Carlo simulation method: A powerful tool for the study of reaction processes. Heter Chem Rev 3:389–418
10. Albano EV (2000) Borówko M (ed) Computational methods in surface and colloid Science. Marcel Dekker Inc, New York, pp 387–435
11. Albano EV (2001) Monte Carlo simulations of teh short-time dynamics of a first-oreder irreversible phase transition. Phys Lett A 288:73–78
12. Albano EV, Mártin HO (1988) Study of recombination reactions of particles adsorbed on fractal and multifractal substrata. Appl Phys A 47:399–407
13. Albano EV, Mártin HO (1988) Temperature-programmed reactions with anomalous diffusion. J Phys Chem 92:3594–3597
14. Alexander S, Orbach R (1982) Density of states on fractals: Fractons. J Physique Lett 43:L625–L631
15. Anacker LW, Kopelman R (1984) Fractal chemical kinetic: Simulations and experiments. J Chem Phys 81:6402–6403
16. Anacker LW, Kopelman R (1987) Steady-state chemical kinetics on fractals: Segregation of reactants. Phys Rev Lett 58:289–291
17. Anacker LW, Kopelman R, Newhouse JS (1984) Fractal chemical kinetics: Reacting random walkers. J Stat Phys 36:591–602
18. Anacker LW, Parson RP, Kopelman R (1985) Diffusion-controlled reaction kinetics on fractal and Euclidian lattices: Transient and steady-state annihilation. J Phys Chem 89:4758–4761
19. Argyrakis P, Kopelman R (1986) Fractal behaviour of correlated random walk on percolation clusters. J Chem Phys 84:1047–1048
20. Argyrakis P, Kopelman R (1992) Diffusion-controlled binary reactions in low dimensions: Refined simulations. Phys Rev A 45:5814–5819
21. Avnir D (1991) Fractal geometry: a new approach to heterogeneous catalysis. Chem Ind 12:912–916
22. Avnir D (1997) Applications of fractal geometry methods in heterogeneous catalysis. In: Ertl G, Knozinger H, Weitkamp J (eds) Handbook of heterogeneous catalysis, vol 2. Wiley-VCH, Weinheim, pp 598–614
23. Avnir D, Farin D, Pfeifer P (1983) Chemistry in noninteger dimensions between two and three. II Fractal surfaces of adsorbents. J Chem Phys 79:3566–3571
24. Bab MA, Fabricius G, Albano EV (2005) Critical behavior of an Ising system on the Sierpinski carpet: A short-time dynamics study. Phys Rev E 71:036139
25. Bab MA, Fabricius G, Albano EV (2006) Discrete scale invariance effects in the nonequilibrium critical behavior of the Ising magnet on a fractal substrate. Phys Rev E 74:041123
26. Bab MA, Fabricius G, Albano EV (2008) On the occurrence of oscillatory modulations in the power-law behavior of dynamic and kinetic processes in fractal media. EPL 81(1)
27. Barabási AL, Stanley HE (1995) Fractal concepts in surface growth. Cambridge University Press, Cambridge
28. Barber MN (1983) Domb C, Lewobitz JL (eds) Phase transitions and Critical Phenomena, vol II. Academic Press, London, p 146

29. Barzykin AV, Tachiya M (2007) Stochastic models of charge carrier dynamics in semiconducting nanosystems. J Phys C Condens Matter 19:065113
30. Bell AT (2003) The impact of nanoscience on heterogeneous catalysis. Science 14:1688–1691
31. Bena I, Droz M, Martens K, Rácz Z (2007) Reaction-diffusion fronts with inhomogeneous initial conditions. J Phys C Condens Matter 19:065103
32. ben-Avraham D, Considine DB, Meakin P, Redner S, Takayasu H (1990) Saturation transition in a monomer-monomer model of heterogeneous catalysis. J Phys A (Math Gen) 23:4297–4312
33. ben-Avraham D, Havlin S (2000) Diffussion and reactions in fractals and disordered systems. Cambridge University Press, Cambridge
34. Berry H (2002) Monte Carlo simulations of enzyme reactions in two dimensions: Fractal kinetics and spatial segregation. Biophys J 83:1891–1901
35. Binder K, Heermann D (2002) Monte Carlo simulations in statistical physics. An introduction, 4th edn. Springer, Berlin
36. Block JH, Ehsasi M, Gorodetskii V (1993) Dynamic Studies of surface reactions with microscopic techniques. Prog Surf Sci 42:143–168
37. Bordogna CM, Albano EV (2007) Statistical methods applied to the study of opinion formation models: a brief overview and results of a numerical study of a model based on the social impact theory. J Phys C Condens Matter 19:065144
38. Bramson M, Lebowitz JL (1988) Asymptotic behavior of densities in difussion-dominated annihilation reactions. Phys Rev Lett 61:2397–2400
39. Brosilow BJ, Ziff RM (1992) Comment on: NO-CO reaction on square and hexagonal surfaces: A Monte Carlo simulation. J Catal 136:275–278
40. Bunde A, Havlin S (1995) A brief introduction to fractal geometry. In: Bunde A, Havlin S (eds) Fractals in science. Springer, Berlin, pp 1–25
41. Casties A, Mai J, von Niessen W (1993) A Monte Carlo study of the CO oxidation on probabilistic fractal. J Chem Phys 99:3082–3091
42. Clar S, Drossel B, Schwabl F (1996) Forest fires and other examples of self-organized criticality. J Phys C Condens Matter 8:6803–6824
43. Clément E, Kopelman R, Sanders LM (1991) The diffusion-limited reaction $A + B \rightarrow 0$ on a fractal lattice. J Stat Phys 65:919–924
44. Clément E, Leroux-Hugon P, Argyrakis P (1994) Catalysis on a fractal lattice: A model for poisoning. Phys Rev E 49:4857–4864
45. Clément E, Sanders LM, Kopelman R (1989) Steady-state diffusion-controlled $A + A \rightarrow 0$ reaction in Euclidean and fractal dimensions: Rate laws and particle self-ordering. Phys Rev A 39:6472–77
46. Considine DB, Redner S, Takayasu H (1989) Comment on: Noise-induced bistability in a Monte Carlo surface-reaction model. Phys Rev Lett 63:2857
47. Cross MC, Hohenberg PC (1993) Pattern formation outside of equilibrium. Rev Mod Phys 65:851–1112
48. Dornic I, Chaté H, Chave J, Hinrichsen H (2001) Critical coarsening without surface tension: The universality class of the Voter model. Phys Rev Lett 87:045701
49. Ehsasi M, Matloch M, Franck O, Block JH, Christmann K, Rys FS, Hirschwald W (1989) Steady and nonsteady rates of reaction in a heterogeneously catalyzed reaction: Oxidation of CO on platinum, experiments and simulations. J Chem Phys 91:4949–4960
50. Engel T (1978) A molecular beam investigation of He, CO, and O_2 scattering from Pd(111). J Chem Phys 69:373–385
51. Engel T, Ertl G (1978) A molecular beam investigation of the catalytic oxidation of CO on Pd(111). J Chem Phys 69: 1267–1281
52. Ertl G (1990) Oscillatory catalytic reactions at single-crystal surfaces. Adv Catal 37:213–277
53. Even U, Rademann K, Jortner J, Manor N, Reisfeld R (1984) Electronic energy transfer on fractals. Phys Rev Lett 42:2164–2167
54. Fichthorn K, Gulari E, Ziff RM (1989) Noise-induced bistability in a Monte Carlo surface-reaction model. Phys Rev Lett 63:1527–1530
55. Frachebourg L, Krapivsky PL (1996) Exact results for kinetics of catalytic reactions. Phys Rev E 53:R3009–R3012
56. Gallos LK, Argyrakis P (2007) Influence of complex network substrate on reaction-diffusion processes. J Phys C Condens Matter 19:065123
57. Gao Z, Yang ZR (1999) Dynamic behavior of the Ziff–Gulari–Barshad model on regular fractal lattices: The influence of lacunarity. Phys Rev E 59:2795–2800
58. Gefen Y, Aharony A, Mandelbrot BB (1983) Phase transitions on fractals. I Quasi-linear lattices. J Phys A (Math Gen) 16:1267–1278
59. Gefen Y, Aharony A, Mandelbrot BB (1984) Phase transitions on fractals. III Infinitely ramified lattices. J Phys A (Math Gen) 17:1277–1289
60. Gefen Y, Mandelbrot BB, Aharony AY (1980) Critical phenomena on fractal lattices. Phys Rev Lett 45:855–858
61. Gouyet JF (1996) Physics and fractal structures. Springer, Paris
62. Grassberger P (1989) Directed percolation in 2+1 dimensions. J Phys A (Math Gen) 22:3673–3679
63. Grassberger P, de la Torre A (1979) Reggeon field theory (Schlögl's first model) on a lattice: Monte Carlo calculations of critical behaviour. Ann Phys (New York) 122:373–396
64. Grinstein G, Lai Z-W, Browne DA (1989) Critical phenomena in a nonequilibrium model of heterogeneous catalysis. Phys Rev A 40:4820–4823
65. Grinstein G, Muñoz MA (1997) Garrido PL, Marro J (eds) Fourth Granada lectures in computational physics. Springer, Berlin, p 223
66. Gálfi L, Rácz Z (1988) Properties of the reaction front in an A+B=C type reaction-diffusion process. Phys Rev A 38:3151–3154
67. Havlin S, ben-Avraham D (1987) Diffusion in disordered media. Adv Phys 36:695–798
68. Hinrichsen H (2000) Non-equilibrium critical phenomena and phase transitions into absorbing states. Adv Phys 49:815–958
69. Huang Y, Ouillon G, Saleur H, Sornette D (1997) Spontaneous generation of discrete scale invariance in growth models. Phys Rev E 55:6433–6447
70. Imbhil R, Ertl G (1995) Oscillatory kinetics in heterogeneous catalysis. Chem Rev 95:697–733
71. Imbihl R (1993) Oscillatory reactions on single crystal surfaces. Prog Surf Sci 44:185–343
72. Jensen I (1991) Non-equilibrium critical behaviour on fractal lattices. J Phys A (Math Gen) 24:L1111–L1117

73. Jensen I, Fogedby H, Dickman R (1990) Critical exponents for an irreversible surface reaction model. Phys Rev A 41:R3411–R3414
74. Johansen A, Sornette D (1998) Evidence of discrete scale invariance in DLA and time-to-failure by canonical averaging. Int J Mod Phys C 9:433–447
75. Johnson BFG (2003) Nanoparticles in catalysis. Topics catal 24:147–159
76. Kang K, Redner S (1984) Scaling approach for the kinetics of recombination processes. Phys Rev Lett 52:955–958
77. Kang K, Redner S (1985) Fluctuation-dominated kinetics in diffusion-controlled reactions. Phys Rev A 32:435–447
78. Kittel C (1986) Introduction to solid state physics, 8th edn. Wiley, New York
79. Kopelman R (1976) In: Fong FK (ed) Radiationless processes in molecules and condensed phases, vol 15. Springer, Berlin, p 297
80. Kopelman R (1986) Rate processes on fractals: Theory, simulations, and experiments. J Stat Phys 42:185–200
81. Kopelman R (1988) Fractal reaction kinetics. Science 241:1620–1626
82. Krapivsky PL (1992) Kinetics of monomer-monomer surface catalytic reactions. Phys Rev A 45:1067–1072
83. Lee Koo Y-E, Kopelman R (1991) Space- and time-resolved diffusion-limited binary reaction kinetics in capillaries: experimental observation of segregation, anomalous exponents, and depletion zone. J Stat Phys 65:893–918
84. Liggett TM (1985) Interacting particle systems. Springer, New York
85. Loscar E, Albano EV (2003) Critical behaviour of irreversible reaction systems. Rep Prog Phys 66:1343–1382
86. Mai J, Casties A, von Niessen W (1992) A model for the catalytic oxidation of CO on fractal lattices. Chem Phys Lett 196:358–362
87. Mai J, Casties A, von Niessen W (1993) A Monte Carlo simulation of the catalytic oxidation of CO on DLA clusters. Chem Phys Lett 211:197–202
88. Maltz A, Albano EV (1992) Kinetic phase transitions in dimer-dimer surface reaction models studied by means of mean-field and Monte Carlo methods. Surf Sci 277:414–428
89. Mandelbrot BB (1983) The fractal geometry of nature. Freeman, San Francisco
90. Marro J, Dickman R (1999) Nonequilibrium phase transitions and critical phenomena. Cambridge Univ Press, Cambridge
91. Meakin P, Scalapino DJ (1987) Simple models for heterogeneous catalysis: Phase transition-like behavior in nonequilibrium systems. J Chem Phys 87:731–741
92. Newhouse JS, Kopelman R (1985) Fractal chemical kinetics: Binary steady-state reaction on a percolating cluster. Phys Rev B 31:1677–1678
93. Newhouse JS, Kopelman R (1988) Steady-state chemical kinetics on surface clusters and islands: Segregation of reactants. J Phys Chem 92:1538–1541
94. Parus SJ, Kopelman R (1989) Self-ordering in diffusion-controlled reactions: Exciton fusion experiments and simulations on naphthalene powder, percolation clusters, and impregnated porous silica. Phys Rev B 39:889–892
95. Pfeifer P, Avnir D (1983) Chemistry in noninteger dimensions between two and three. I Fractal theory of heterogeneous surfaces. J Chem. Phys 79:3558–3565
96. Rammal R, Toulouse G (1983) Random walks on fractal structures and percolation clusters. J Physique Lett 44:L13–L22
97. Rozenfeld AF, Albano EV (2001) Critical and oscillatory behavior of a system of smart preys and predators. Phys Rev E 63:061907
98. Shipilevsky BM (2007) Formation of a finite-time singularity in diffusion-controlled annihilation dynamics. J Phys C Condens Matter 19:065106
99. Sornette D (1998) Discrete scale invariance and complex dimensions. Phys Rep 297:239–270
100. Stanley HE (1971) Introduction to phase transitions and critical phenomena. Oxford University Press, New York
101. Stauffer D, Aharoni A (1992) Introduction to the percolation theory, 2nd edn. Taylor and Francis, London
102. Suchecki K, Hołyst JA (2006) Voter model on Sierpinski fractals. Physica A 362:338–344
103. Toussaint D, Wilczek F (1983) Particle-antiparticle annihilation in diffusive motion. J Chem Phys 78:2642–2647
104. Voigt CA, Ziff RM (1997) Epidemic analysis of the second-order transition in the Ziff–Gulari–Barshad surface-reaction model. Phys Rev E 56:R6241–R6244
105. Weiss GH (1995) A primer of random walkology. In: Bunde A, Havlin S (eds) Fractals in science. Springer, Berlin, p 119
106. Wieckowski A, Savinova ER, Vayenas CG (2003) Catalysis and electrocatalysis at nanoparticle surfaces. Marcel Dekker, New York, pp 1–959
107. Witten TA, Sander LM (1983) Diffusion-limited aggregation. Phys Rev B 27:5686–5697
108. Yaldram K, Khan MA (1991) NO-CO reaction on square and hexagonal surfaces: A Monte Carlo simulation. J Catal 131:369–377
109. Zhdanov VP, Kasemo B (1994) Kinetic phase transitions in simple reactions on solid surfaces. Surf Sci Rep 20:111–189
110. Ziff RM, Brosilow BJ (1992) Investigation of the first-order phase transition in the $A - B_2$ reaction model using a constant-coverage kinetic ensemble. Phys Rev A 46:4630–4633
111. Ziff RM, Gulari E, Barshad Y (1986) Kinetic phase transitions in an irreversible surface-reaction model. Phys Rev Lett 56:2553–2556

Books and Reviews

Binder K (1997) Applications of Monte Carlo methods to statistical physiscs. Rep Prog Phys 60:487–560

Bunde A, Havlin S (1991) Fractals and disordered media. Springer, New York

Cardy J (1997) Goddard P, Yeomas JM (eds) Scaling and renormalization in statistical physics. Cambridge University Press, Cambridge

Christmann K (1991) Introduction to surface physical chemistry. Steinkopff, Darmstadt, p 1

Egelhoff WF Jr (1982) King DA, Woodruff DP (eds) Fundamental studies of heterogeneous catalysis, vol 4. Elsevier, Amsterdam

Lindenberg K, Oshanin G Tachiya M (2007) Chemical kinetics beyond the texbook: Fluctuations, Many-particle effects and anomalous dynamics. Special issue of J Phys C Condens Matter 19

Walgraef D (1997) Spatio-temporal pattern formation. Springer, New York

Record Statistics and Dynamics

Paolo Sibani[1], Henrik, Jeldtoft Jensen[2]
[1] Institut for Fysik og Kemi, SDU, Odense, Denmark
[2] Department of Mathematics, Imperial College London, London, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Complex Dynamics
The Distribution of Records in Stationary Time Series
Aging and Metastability
Marginal Stability
Record Statistics and Intermittency in Glassy Dynamics
Future Directions
Acknowledgments
Bibliography

Glossary

Aging The slow time evolution of a class of complex systems, which are brought into a far-from-equilibrium state by the sudden change of an external parameter, e. g. the temperature.

Complex dynamics The collective or emergent time dependent properties of interacting multi-component and multi agent systems.

Marginal stability A metastable state is marginally stable if it can be destroyed by slight perturbations.

Metastability The ability of a non-equilibrium system to remain in, or close to, the same state for a certain characteristic time.

Record In a time ordered series of random numbers, a record is an entry larger than all preceding entries.

Scale invariant process A scale invariant process looks the same under re-scaling of time and/or space variables.

Stationary process A stationary process is time homogeneous, i. e. it is invariant under time translations.

Definition of the Subject

The term *record statistics* covers the statistical properties of records within an ordered series of numerical data obtained from observations or measurements. A record within such series is simply a value larger (or smaller) than all preceding values. The mathematical properties of records strongly depend on the properties of the series from which they are extracted. These properties have been investigated for many different cases, the simplest cases perhaps being series of independent random numbers drawn from the same (arbitrary) distribution, and series produced by a diffusion process with independent random increments.

The term *record dynamics* covers the rather new idea that records may, in special situations, have measurable dynamical consequences. The approach applies to the aging dynamics of glasses and other systems with multiple metastable states. The basic idea is that record sizes fluctuations of e. g. the energy are able to push the system past some sort of 'edge of stability', inducing irreversible configurational changes, whose statistics then closely follows the statistics of record fluctuations.

Introduction

Floods, droughts and earth-quakes are spectacular, potentially disastrous, and routinely monitored extremal events. Yet, even with no large scale effects involved, record performances of all kinds attract great interest, as witnessed by the famous *Guinness Book of Records*. Aside from their popular appeal, extremes and their properties are an important topic of mathematical statistics [16,28].

Formally, a value larger (or smaller) than all its predecessors within a time ordered series of numerical data constitutes a record. The number of records in a series encodes information on how the data are produced. E. g. measuring every second the velocity of a steadily accelerating object produces a time series where each entry is a record. Conversely, with data generated by a random process, e. g. the coordinates of a drunkard taking a step in any direction with equal probability, considerably fewer records occur. An intermediate case, which is indicative of a climatic warming trend, are temperature records more frequent than expected for fluctuations around a constant [35]. More generally, record-sized events in spatio-temporal series of data, such as seismic shocks, are indicative of correlations and causal relationships [14].

In some contexts, record sized events have important implications: e. g. in population dynamics, mutants surpassing the current standard for highest reproductive success contribute disproportionately to the genetic pool of the next generation, a bias which automatically raises the bar for future improvements. This naturally leads to the expectation that record events be of significance to biological evolution, as perhaps in particular suggested by the catch-phrase 'survival of the fitt*est*'. Darwinism's impact on biology – and human culture at large – can hardly

be exaggerated. In recent years computer optimization methods have appeared, e. g. Genetic Algorithms [17] and Extremal Optimization [5], which are inspired by evolution dynamics and which, implicitly or explicitly, utilize records.

In a computer program, or in a social context, the function, entity or organization in charge of memorizing the existing record and detecting its eventual obliteration is clearly distinct from the process being monitored. By contrast, record keeping in biological evolution is fully integrated in the dynamics, and we will see in examples below that record dynamics can emerge as a collective property of a collection of co-evolving organisms.

In cases where records act as seeds for an amplifying dynamical process, the magnitude of the seed bears no direct relation to the magnitude of the final outcome of the amplification. E. g. a snow avalanche can be triggered by sound vibrations above a certain threshold. However, its impact is more related to the amount of snow on the ground and to the slope and shape of the valley than to the intensity of the sound.

Two important and related characteristic of record-induced dynamics are: *i)* its temporal non-homogeneity, meaning that record fluctuations and any induced effects occur at a decreasing rate and *ii)* the presence of an embedded memory mechanism. Not coincidentally, the same properties are present in a large class of systems known as *glassy*. In the sequel, we focus on the significance of record sized fluctuations for *glassy dynamics*, expanding on the observation that record fluctuations induce transitions between metastable states [43], leaving a so-called intermittent signal as a fingerprint [44].

Complex Dynamics

Systems of many interacting components and with a coupling to an external reservoir of e. g. energy and/or particles, are usually described using Master- or Fokker–Planck equations [57]. These equations combine external and internal deterministic forces with a randomizing influence of the reservoir, and determine the probability flow between a set of allowed configurations or states. A simple yet illustrative example is the Brownian motion of a grain of pollen immersed in a fluid. The collisions with the surrounding molecules exert random forces on the molecules. Furthermore, an external force may be applied to the pollen by means of e. g. an electric field. If the motion of the grain is constrained by the walls of a container, the probability distribution of its position reaches on a certain time scale – called the relaxation time – a stationary form which only depends on the geometry of the container. In this case, all memory of the initial condition is lost. Master equations or other dynamical equations often lead to a stationary state. The latter can be unique, or it can be one out of many, each state draining a basin of attraction in configuration space. An interesting situation arises when the stationarity is only approximate, i. e. when, given enough time, it is possible to escape from a basin of attraction and enter a different one. Each attraction basin then becomes a metastable region of configuration space, which is characterized by a certain *escape time*: the time typically needed for a trajectory to exit the basin.

Gases, liquids and crystalline solids are mostly found in a stationary state where relevant physical observables, e. g. the local density, fluctuate reversibly around fixed values. A stationary fluctuation signal is invariant under time translations, its average is a constant and its autocorrelation function only depends on the difference between its two time arguments. In some cases, e. g. for equilibrium systems fluctuating near a critical temperature, correlations decay algebraically and may lack an associated time-scale. Interestingly, a similar situation arises in certain slowly driven systems, epitomized by the famous 'sand pile' model [4,22], which self-organizes into marginally stable configurations lacking characteristic space and time scales. Friction dominated granular systems, e. g. rice piles, are experimental realizations of this behavior, which is known as Self Organized Critical, or SOC.

In SOC and some other driven dissipative systems the initial state is statistically equivalent to the states subsequently visited. In contrast, complex materials such as glasses, polymers, colloids and disordered magnetic materials, undergo a slow but systematic physical aging following a rapid quench of e. g. the temperature or the particle density. Typically, these materials get trapped in metastable basins, where physical variables appear to fluctuate reversibly on observation time scales shorter than the age. During this pseudo-equilibrium regime, fluctuations have Gaussian distributions with zero averages [44]. On longer time scales, intermittent events add an asymmetric, and typically exponential, tail to the Probability Density Function (PDF) of the fluctuations. The tail is caused by the non-equilibrium drift which, through a sequence of metastable basins of subtly different nature, slowly erases the memory of the initial state and changes physical properties. e. g. the linear response to a small perturbation.

The Distribution of Records in Stationary Time Series

Record-sized entries within a stationary series of independent and identically distributed random numbers form

a sub-series whose interesting statistical properties follow from simple arguments [16,42,45]. E. g., as the magnitude required to qualify as the next record, increases with each new entry, records appear at a decreasing rate. For sufficiently long time series, the number of records found between the first and the tth draw can be shown to have a Poisson distribution, with average $\ln t$. Equivalently, if the kth record appears at times t_k, the ratios $\ln(t_k/t_{k-1})$ are independent random numbers with an exponential distribution.

Consider a sequence of independent random numbers drawn from the same distribution at times $1, 2, 3, \ldots, t$. We exclude distributions supported on a finite set, because these eventually produce a record which cannot be beaten[1]. The first number drawn is by definition a record. Subsequent trials lead to a record if their outcome is larger than the previous record. We seek the probability $P_n(t)$ of finding precisely n records in t successive trials, where $1 \le n \le t$. In the derivation we need the auxiliary function $P_{(1,m_1,\ldots m_{k-1})}(t)$, which is the joint probability that k records happen at times $1 < m_1, \cdots < m_{k-1}$, with $m_{k-1} \le t$. $P_1(t)$ is simply the probability that the first outcome be largest among t. As each outcome has, by symmetry, the same probability of being the largest, it follows that $P_1(t) = 1/t$. In order to obtain two records at times 1 and m, the largest of the first $m-1$ random numbers must be drawn at the very first trial. This happens with probability $1/(m-1)$. Secondly, the mth outcome must be the largest among t. This happens with probability $1/t$, independently of the position of the largest outcome in the first $m-1$ trials. Accordingly,

$$P_{(1,m)}(t) = \frac{1}{(m-1)t} \,. \tag{1}$$

Summing the above over all possible values of m we then obtain

$$P_2(t) = \sum_{m=2}^{t} \frac{1}{(m-1)t} \approx \ln(t)/t \,, \tag{2}$$

and, in the more general case of n events, we similarly obtain

$$P_{(1,m_1,\ldots,m_{n-1})}(t) = \frac{1}{\prod_{i=1}^{n-1}(m_i-1)t} \,. \tag{3}$$

We now take $q_i = m_i - 1$ and sum over all possible values of the q_i's, leading to

$$P_n(t) = \sum_{q_1=1}^{t-n+1} \frac{1}{q_1} \cdots \sum_{q_{n-1}=q_{n-2}+1}^{t-1} \frac{1}{q_{n-1}} \frac{1}{t} \,. \tag{4}$$

[1] Apart from this constraint, the form of the distribution is immaterial.

An approximate closed form expression can now be obtained by replacing the sums by integrals, which is reasonable for $t \gg n \gg 1$. The integrals can then be evaluated, finally yielding

$$P_n(t) = \frac{(\ln t)^{n-1}}{(n-1)!} \frac{1}{t} \,. \tag{5}$$

As anticipated, Eq. 5 is a Poisson distribution, with $\ln t$ replacing the usual time argument t. Clearly, the number of records between extractions t_w and $t > t_\mathrm{w}$ is the difference of two Poisson process, and hence itself a Poisson process with average $\ln(t) - \ln(t_\mathrm{w}) = \ln(t/t_\mathrm{w})$. The discrete series of random numbers can be replaced by stationary random white noise, in which case t becomes a continuous time variable, and the restriction on the asymptotic validity of Eq. 5 for $t \gg n$ can be lifted.

Let $\overline{n(t)}$ and $\sigma_n^2(t)$ be the average and variance of the number of events in time t. As an immediate consequence of Eq. 5 we note that

$$\overline{n(t)} = \sigma_n^2(t) = \ln t \,. \tag{6}$$

Furthermore, the 'current', i. e. the average number of events per unit of time decays as

$$\frac{\mathrm{d}\overline{n}}{\mathrm{d}t} = \frac{1}{t} \,. \tag{7}$$

A third consequence of Eq. 5 is the following: Let $t_1 = 1 < t_2 < \cdots < t_k < \cdots$ be the times at which the record breaking events occur, and let $\tau_1 = \ln t_1 = 0 < \tau_2 \cdots < \tau_k = \ln t_k < \cdots$ be the corresponding natural logarithms. The stochastic variables $\Delta_k = \tau_{k+1} - \tau_k = \ln(t_{k+1}/t_k)$ are independent and identically distributed. Their common distribution is an exponential with unit average. By writing: $\tau_k = \Delta_{k-1} + \Delta_{k-2} + \cdots \Delta_1$ we have that $(\tau_{k+1} - k)/\sqrt{k}$ approaches a standard Gaussian distribution for large k. Hence, the waiting time t_k for the kth event is approximately log-normal, and the average of its logarithm grows linearly in k. By Jensen's inequality [37] we also find

$$\ln(\overline{t_k}) \ge \overline{\ln(t_k)} = k \,. \tag{8}$$

We see that the average waiting time from the kth to the $(k+1)$th event grows at least exponentially in k.

Consider finally the case in which a system is made up of α independent parts. The total number of records is the sum of contributions from each part, and hence remains Poisson distributed. For the most general case where records between times t_w and $t > t_\mathrm{w}$ are of interest, the Poisson process has average

$$\langle n \rangle(t_\mathrm{w}, t) = \alpha \ln(t/t_\mathrm{w}) \,. \tag{9}$$

The cumulative distribution of the 'logarithmic waiting times' $\delta_k = \ln(t_k/t_{k-1})$ is correspondingly

$$P(\delta_k < x) = 1 - \exp(-\alpha x); \quad x > 0; \quad k = 2, 3, \ldots \quad (10)$$

In glassy dynamics, noise fluctuations are drawn from a (quasi-) equilibrium distribution characterizing the current metastable region. This distribution heavily penalizes large excursions from the metastable configuration, making record events rare and uncorrelated.

Importantly, spatially extended systems with short range interactions may break down into a set of weakly interacting subsystems, or domains, which are localized in space and which have independent fluctuations and independent metastable states. E. g. spin glasses are characterized by a thermal correlation length, corresponding to the linear size of the equilibrated domains. The latter grows slowly and remains at all times rather small compared to the linear size of the sample. Each independently thermalized domain evolves in parallel with other domains, from which it is shielded by a backbone of mainly inactive degrees of freedom. In accord with observations discussed below, fluctuation spectra of physical quantities have a Gaussian component with zero average, and a non-Gaussian tail. The former describes the sum of independent equilibrium-like fluctuations, and the latter describes the rare non-equilibrium large events, or *quakes*, which carry the drift. In this situation, the parameter α of Eq. 9 can be identified with the number of domains. Its value enters the width of the Gaussian, which grows as the square root of α, as well as the statistical weight of the tail, which grows linearly with α.

Aging and Metastability

The term glassy dynamics usually refers to the extremely slow relaxation observed in complex systems with many interacting components, where reaching a steady, time independent, state is typically far beyond experimentally accessible time scales. E. g., when melted alloys are cooled down rapidly, they do not enter a crystalline ordered state. Instead the atoms retain the amorphous arrangement characteristic of the liquid phase while the mobility of the molecules decreases by many orders of magnitude. This colossal change in the characteristic dynamical time scales hinders a structural glass at low temperature to reach thermodynamic equilibrium. Nevertheless, the properties of glasses over short observation time scales may appear to be time independent as in thermal equilibrium.

The low temperature out-of-equilibrium behavior of structural glasses and of a host of other complex materials is intriguingly similar. E. g. spin glasses [29] a class of disordered magnetic alloys, polymers [53], colloids and gels [10] and type II superconductors [31,32] all remain out of equilibrium, with their macroscopic physical properties changing algebraically or logarithmically as a function of the time elapsed since the initial quench. Small temperature and field variations applied during the aging process uncover many fascinating 'memory' effects see e. g. [19,23,59], which indicate the presence of a hierarchy of time scales for relaxation in configuration space, a hierarchy which is also explicitly uncovered in numerical studies [49,50].

The topography of complex energy landscapes has been thoroughly investigated, see e. g. [2,30,38,51,52], and various heuristic approaches have been proposed in order to link topography and dynamics. For glassy dynamics, heuristic ideas such as hierarchical mesoscopic models [20,58], and the highly popular trap model [6] have been considered. While many aspects of aging in glassy systems are known and partly understood, important open questions remain. E. g. how are configuration space properties and real space morphology to be combined in a seamless fashion? How should one weigh thermodynamical versus dynamical properties, and, more specifically what is the rôle of equilibrium statistical properties, a rôle initially strongly advocated from different camps [7,15,33], and hotly debated ever since [24,25,34]. In retrospect, knowing that systems with similar aging behavior have completely different equilibrium properties, e. g. vibrated granular matter versus thermalizing polymers, the connection between aging and equilibrium properties appears rather tenuous.

A radically different approach was initiated by Coppersmith and Littlewood who studied memory effects in Charge Density Waves [12] using a simple model with multiple attractors. The focus is there on the mechanism for attractor *selection*, and the suggestion is that the least stable attractors, are those typically selected following a quench, in spite of having little statistical weight. One development of the idea [45] points to fluctuations records, which are likewise statistically insignificant, as key events in aging dynamics.

The focus of both experiments and theory has recently shifted from macroscopic averages to the fluctuations around these averages (Buisson et al. [8]), which are observable in meso- and nanoscaled systems. As mentioned, the fluctuation spectra have a Gaussian part, pertaining to pseudo-equilibrium fluctuations around the 'current' metastable value of the macroscopic average of the quantity of interest, and an asymmetric exponential tail, describing rare and irreversible jumps from one metastable

configuration to the next. Such 'quakes' carry the full macroscopic evolution of the system. Their decreasing rate is indicative of an entrenchment into gradually more stable configurations.

Marginal Stability

Broadly speaking, marginal stability relates to how metastable attractor basins are *selected* during the off-equilibrium evolution of metastable complex systems. Diverse applications include driven systems, e. g. charge density waves [12,45] and the sand-pile model [4] and its conceptual ancestor [3,41,56], thermal aging of e. g. spin-glasses [44] and type II superconductors [32]. The relevance to biological evolution is suggested by a number of model studies, some where a fitness function is explicitly defined and drives the dynamics [46,48] and others based on co-evolutionary interactions without fitness function [1].

Record size fluctuations are irrelevant in stationary processes, where the possibility of triggering permanent changes simply does not exist. They are likewise irrelevant in processes characterized by one or few time scales, e. g. thermal equilibration between two metastable basins: A record has a rank, but lacks an inherent magnitude which could match the scale of the process. For record fluctuations to trigger irreversible changes, many inequivalent meta-stable basins must be accessible to the dynamics. We envisage that each metastable basin is associated to a *finite* characteristic escape time τ, the typical time it takes to leave the basin. For thermal dynamics, the relevant time scale is the Arrhenius time corresponding to a free energy barrier $b = Tk_B \ln(\tau)$, where T is the temperature and k_B is the Boltzmann constant.

Assuming that attractor basins have a broad range of stability, with the more fickle basins out-numbering the more stable ones, a quench from a random initial condition typically leads to very fickle basins. At any stage of the subsequent evolution, all moves leading from the less to the more fickle can be reversed on the time scale at which they first happen. Irreversible moves into more stable attractors, the quakes, are important for the aging dynamics, and typically entail a tiny, or marginal, increase of stability. The marginal increase implies that a random fluctuation slightly larger than the previous largest fluctuation, i. e. a record fluctuation, can trigger the next quake. Defacto irreversibility of the quakes can be strengthened by a thermodynamic mechanism, i. e. the basin change may entail a large decrease in energy, and/or a large increase of entropy.

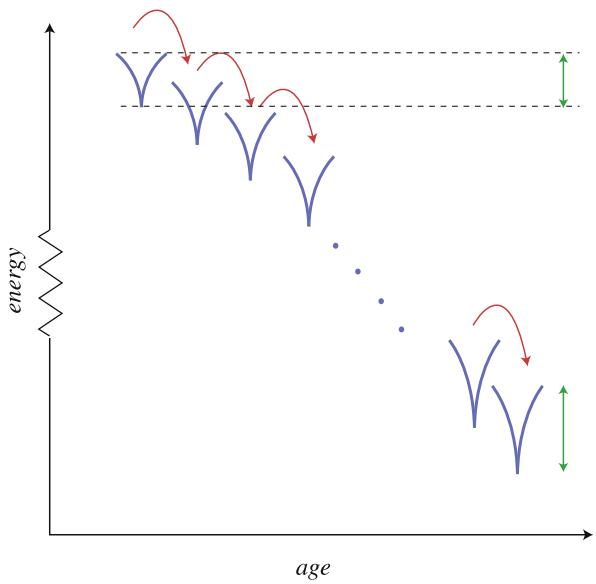

Record Statistics and Dynamics, Figure 1
The cartoon, taken from [43], illustrates the link between marginal stability and record dynamics: the wedges represent metastable region of a complex system, with any reference to their internal hierarchical structure omitted for simplicity. The vertical axis is the energy. Jumps from one metastable region to the next (quakes) are indicated by unidirectional arrows. The thermal stability of each metastable region is represented by the vertical distance from the bottom to the boundary of the wedge, i. e. the energy barrier for the corresponding region. Importantly, the barrier of the current attractor only changes minutely through each quake

The record dynamics scenario sketched above implies that, as the systems age, the metastable basins successively explored develop a nested hierarchical substructure, each layer composed of basins of greater and greater stability. If all metastable regions maintain a similar structure, the noise distribution from which the records are drawn does not itself change during the aging process.

The temporal statistics of record fluctuations and quakes has universal properties which are disconnected from the physical changes associated to the quakes. These effects may e. g. depend on the physical quantity being measured and on the noise strength. Using the *subordination* principle [32,39,47] briefly discussed in the next Section, the changes can be described mathematically in a general, albeit approximate, fashion.

Record Statistics and Intermittency in Glassy Dynamics

The left panel of Fig. 2 (taken from [32]) illustrates important qualitative features of intermittency in complex systems. The figure depicts the time evolution of the number

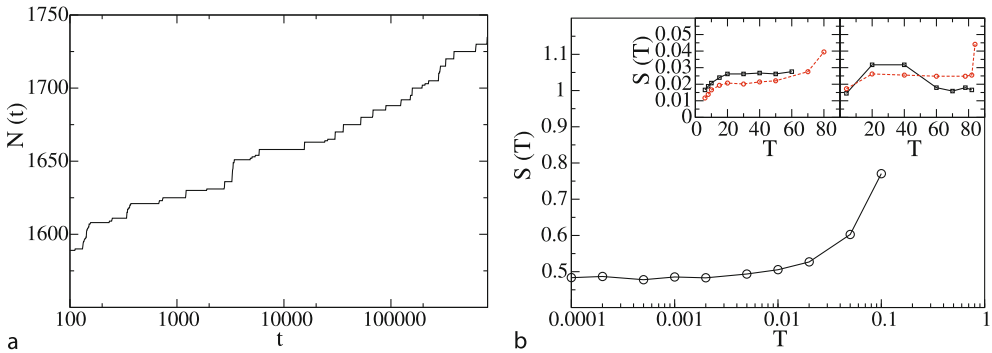

Record Statistics and Dynamics, Figure 2

The figure is taken from [32]. Left panel: The time variation of the total number of vortices $N(t)$ in the system for a single realization of the pinning potential and the thermal noise in a $8 \times 8 \times 8$ lattice for $T = 0.1$. The monotonous step function character of the time series, and the fact that the steps are approximately the same duration on a *logarithmic* time scale indicates the decelerating nature of the dynamics. Right panel: Numerical results for the creep rate versus T for the time interval between $t = 1000$ and $t = 10000$. In the low temperature region the creep rate is constant within our numerical precision for about two orders of magnitude – we observe a nonzero creep rate in the $T \to 0$ limit. *Insets:* experimental results for the creep rate versus T. The *right inset* shows data from Keller et al. [26,27] for melt processed YBCO crystals with the magnetic field applied along the *c*-axis (squares) and *ab* plane (circles). The *left inset* shows data from Civale et al. [11] for un-irradiated (squares) and 3 MeV proton-irradiated (circles) YBCO flux grown crystals with a 1 *T* magnetic Field applied parallel to the *c*-axis

of vortices in the Random Occupancy Model for magnetic flux creep in type II superconductors [31].

The number $n(t)$ of vortices inside the system changes in a step-wise fashion. Each horizontal plateau corresponds to the time spent in a metastable configuration. The plateaus appear to have the same average duration on a *logarithmic* time scale, showing that the dynamics entrenches itself into gradually more stable basins. Each vertical jump corresponds to a quake, leading from one basin to the next. The size of the quake is the height of the corresponding jump.

In a record-dynamics scenario the rate of quakes is independent of the statistical properties of the noise. Specifically for thermally activated dynamics, it is independent of the temperature. The rate of change of a physical observable may have a temperature dependence, which enters through the quake size distribution, i. e. the statistical distribution of the vertical jumps. The right panel of Fig. 2, also taken from [32], shows the striking temperature independence of the creep rate in the ROM model, and the insert shows its experimental counterpart.

Quenching a glassy system from a high temperature invariably leads to a configuration with an energy far above the equilibrium value at the low temperature. It is then of interest to understand how the excess energy leaves the system during the aging process. The release in a spin-glass [44] occurs through sporadic, intermittent quakes, and the average rate of energy flow (nearly) decreases with the inverse of the system age. The p-spin model [54], an Ising spin model with plaquette interactions, has this type of behavior as well. The left panel of Fig. 3 (taken from [39]) shows the Probability Density Function (PDF) of the amount of heat given off by the system over short intervals of length δt, for several values of the system age t_w. The Gaussian part of the PDF, which is nearly independent of t_w, describe reversible energy fluctuations of zero average. The exponential tail describes the out-flow of heat. As t_w increases, all other parameters being constant, the tail becomes less prominent. The right panel shows, for a number of low temperatures, that the rate of energy flow decays as the reciprocal of the age.

Experimental probes of glassy dynamics often involve applying a small perturbing field at a certain age t_w and measuring the linear response, $R(t, t_w)$, for $t > t_w$. The glassy response depends on both t_w and t, while in a stationary situation the only dependence is on the difference $t - t_w$. There is a large and growing body of experimental and numerical results concerning linear response functions in aging systems and their many facets[23,36,47,55,58].

Attempts to describe the off-equilibrium response as analogous to equilibrium response lead to the idea of effective temperature [13]. Being defined in the limits $t_w \to \infty$ and $t \to \infty$, with t/t_w finite, the latter quantity is, in practice, difficult to measure [21]. At a more basic level, one may ask whether a glassy system is mainly responding to the applied field, or rather mainly responding to the initial quench. In the latter case, the field has the auxiliary rôle of biasing the course of the quakes, but no influence on their temporal statistics. Technically, the re-

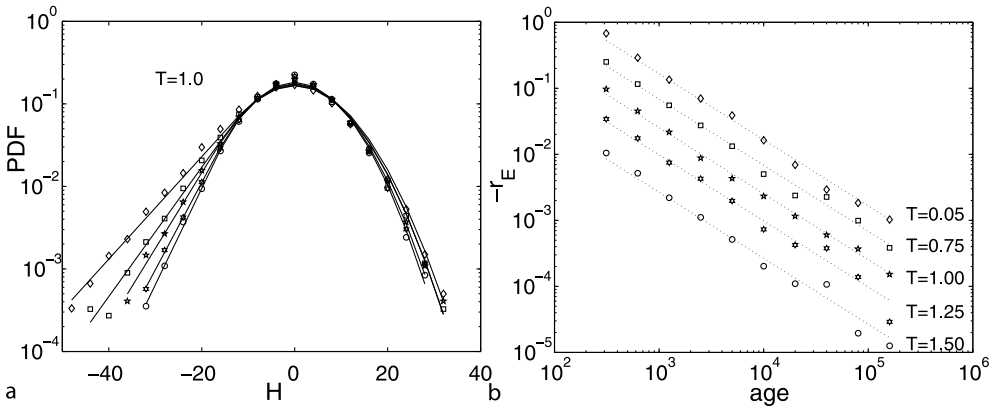

Record Statistics and Dynamics, Figure 3

The figure is taken from [39]. Left panel: The PDF of the heat exchanged between system and thermal bath over a time $\delta t = 5$. Negative values correspond to an energy outflow. The data are based on 200 independent runs, taken at temperature $T = 1$ in the intervals $[t_w, t_w + 500]$ for $t_w = 1000$ (diamonds), $t_w = 2000$ (squares) $t_w = 4000$ (polygons) $t_w = 8000$ (hexagons) and $t_w = 16000$ (circles). Right panel: The average rate of flow of the energy is plotted versus the age for the five temperatures shown. The full line has the form $y = ct_w^{-1}$, with the proportionality constant c estimated as the mean of $t_w r_E$. Since data sets for different temperatures are almost overlapping, they are vertically shifted in the plot for the sake of typographical clarity

sponse is then *subordinated* to the quakes, which are themselves subordinated to record-sized fluctuations. Following this approach [39,47], physical observables are treated as (stochastic) functions of the number of quakes k occurring during the observation interval. Generic eigenvalue expansions in the variable k are available, and the time and age dependence are extracted by averaging them over all possible values of k according to the Poisson distribution with the average given in Eq. 9. The subordination approach is not limited to linear response functions, e. g. it can be applied to a calculation of the configurational autocorrelation function [40] or indeed any other quantity which can be argued to be (mainly) a function of k [32].

The potential relevance of records for the evolution of ecosystems has been investigated using the Tangled Nature model of evolutionary ecology [9,18]. The model deals with a population of individuals undergoing mutation-prone reproduction in a type space. All individuals are equally likely to be removed or killed. Their reproduction, however, depends on the interactions among subsets of individuals: The reproduction probability of an individual of type, say, A will change through interactions with a type B individual which comes into being through a mutation of another existing type. Depending on whether the interaction is cooperative or antagonistic, the reproduction rate of A will increase, respectively decrease.

Metastable ecosystems spontaneously arise as sets of types for which the reproduction and death probability of the individuals are balanced. Such metastable configurations last for periods of varying duration, and are eventually replaced by new metastable configurations. The model does not assume a fitness function, and no optimization procedure is explicitly invoked at the level of the microscopic dynamics. Interestingly, the times spent in consecutive metastable configurations have a statistics in qualitative agreement with the predictions of record dynamics [1]. This illustrates how record dynamics can be identified by a relatively straightforward analysis of time series, and how it can emerge at the level of collective evolution even when the precise nature of the observable undergoing records is not known.

Future Directions

Metastable systems are ubiquitous in nature, and their interaction with an external noisy environment affects their dynamics in significant ways. When the system evolves irreversibly through metastable configurations of marginally increasing stability, fluctuation records determine the course of the dynamics. The temporal statistics of records is an in-road to an approximate analytical description of the time evolution of complex systems, ranging from physical material to biological systems. It is an open challenge to test and develop these idea on the large variety of model and observational data where irreversibility and marginal stability appear in different guises.

Acknowledgments

The authors are indebted to P. Anderson, J. Dall, L. Oliveira and S. Boettcher. P. Sibani did part of this work while visiting the Cherry L. Emerson Center for Scientific Computation at Emory University.

Bibliography

Primary Literature

1. Anderson P, Jensen HJ, Oliveira LP, Sibani P (2004) Evolution in complex systems. Complexity 10:49–56
2. Angelani L, Di Leonardo R, Parisi G, Ruocco G (2003) Topological description of the aging dynamics in simple glasses. Phys Rev Lett 87:055502
3. Bak P (1997) How Nature Works. The science of self-organized criticality. Oxford University Press, Oxford
4. Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality: An explanation of 1/f noise. Phys Rev Lett 59:381–384
5. Boettcher S, Percus A (2001) Optimization with extremal dynamics. Phys Rev Lett 86:5211–5214
6. Bouchaud JP (1992) Weak ergodicity breaking and aging in disordered systems. J Phys I France 2:1705–1713
7. Bray AJ, Moore MA (1987) Chaotic nature of the spin-glass phase. Phys Rev Lett 58:57–60
8. Buisson L, Bellon L, Ciliberto S (2003) Intermittency in aging. J Phys Condens Matter 15:S1163
9. Christensen K, di Collobiano SA, Hall M, Jensen HJ (2002) Tangled Nature: A Model of Evolutionary Ecology. J Theor Biol 216:73–84
10. Cipelletti L, Manley S, Ball RC, Weitz DA. Universal aging features in the restructuring of fractal colloidal gels. Phys Rev Lett 84:2275–2278
11. Civale L, Marwick AD, McElfresh MW, Worthington TK, Malozemoff AP, Holtzberg FH, Thompson JR, Kirk MA (1990) Defect independence of the irreversibility line in proton-irradiated y-ba-cu-o crystals. Phys Rev Lett 65(9):1164–1167
12. Coppersmith SN, Littlewood PB (1987) Pulse-duration memory effect and deformable charge-density waves. Phys Rev B 36:311–317
13. Cugliandolo LF, Kurchan J, Peliti L (1997) Energy flow, partial equilibration, and effective temperature in systems with slow dynamics. Phys Rev E 55:3898–3914
14. Davidsen J, Grassberger P, Paczuski M (2007) Networks of recurrent events, a theory of records, and an application to finding causal signatures in seismicity. cond-mat/0701190
15. Fisher DS, Huse DA (1986) Ordered phase of short-range Ising spin-glasses. Phys Rev Lett 56:1601–1604
16. Glick N (1978) Breaking records and breaking boards. Am Math Mon 85:2–26
17. Goldberg DE (1989) Genetic algorithms in search, optimization and machine learning. Addison-Wesley, Reading
18. Hall M, Christensen K, di Collabiano SA, Jensen HJ (2002) Time dependent extinction rate and species abundance in the Tangled Nature model of biological evolution. Phys Rev E 66:011904
19. Hammann J, Bouchaud J-P, Dupuis V, Vincent E (2001) Separation of time and length scales in spin glasses: Temperature as a microscope. Phys Rev B 65:024439
20. Hoffmann KH, Sibani P (1988) Diffusion in hierarchies. Phys Rev A 38:4261–4270
21. Hérisson D, Ocio M (2002) Fluctuation-dissipation ratio of a spin glass in the aging regime. Phys Rev Lett 88:257202
22. Jensen HJ (1998) Self-Organized Complex Systems. Cambridge University Press, Cambridge
23. Jonason K, Vincent E, Hammann J, Bouchaud JP, Nordblad P (1998) Memory and Chaos Effects in Spin Glasses. Phys Rev Lett 81:3243–3246
24. Jönsson PE, Takayama H, Katori AH, Ito A (2005) Dynamical breakdown of the ising spin-glass order under a magnetic field. Phys Rev B 71:180412(R)
25. Katzgraber HG, Young AP (2005) Probing the Almeida-Thouless line away from the mean-field model. Phys Rev B 72:184416
26. Keller C, Kupfer H, Meier-Hirmer R, Wiech U, Selvamanickam V, Salama K (1990) Irreversible behaviour of oriented grained *yba2cu3ox*. part 2: relaxation phenomena. Cryogenics 30:410–416
27. Keller C, Kupfer H, Meier-Hirmer R, Wiech U, Selvamanickam V, Salama K (1990) Irreversible behaviour of oriented grained YBa2Cu3Ox. part 1: transport and shielding currents. Cryogenics 30:401–409
28. Krug J (2007) Records in a changing world. cond-mat/0702136
29. Lundgren L, Nordblad P, Sandlund L (1986) Memory behaviour of the spin glass relaxation. Europhys Lett 1:529–534
30. Mossa S, Ruocco G, Sciortino F, Tartaglia PP (2002) Quenches and crunches: does the system explore in ageing the same part of the configuration space explored in equilibrium? Phil Mag B 82:695–705
31. Nicodemi M, Jensen HJ (2001) Aging and memory phenomena in magnetic and transport properties of vortex matter: a brief review. J Phys A 34:8425
32. Oliveira LP, Jensen HJ, Nicodemi M, Sibani P (2005) Record dynamics and the observed temperature plateau in the magnetic creep rate of type ii superconductors. Phys Rev B 71: 104526
33. Parisi G (1979) Infinite number of order parameters for spin-glasses. Phys Rev Lett 43:1754–1756
34. Parisi G, Marinari E, Ruiz-Lorenzo JJ (1998) On the phase structure of the 3D Edwards-Anderson spin-glass. Phys Rev B 58:14852
35. Redner S, Petersen MR (2007) Role of global warming on the statistics of record-breaking temperatures. Phys Rev E 74:061114
36. Rodriguez GF, Kenning GG, Orbach R (2003) Full Aging in Spin Glasses. Phys Rev Lett 91:037203
37. Rudin W (1966) Real and complex analysis. McGraw Hill, New York
38. Schön JC, Putz H, Jansen M (1996) Studying the energy hypersurface of continuous systems–the threshold algorithm. J Phys: Condens Matter 8:143–156
39. Sibani P (2006) Aging and intermittency in a p-spin model. Phys Rev E 74:031115
40. Sibani P (2006) Mesoscopic fluctuations and intermittency in aging dynamics. Europhys Lett 73:69–75
41. Sibani P, Andersen CM (2001) Aging and self-organized criticality in driven dissipative systems. Phys Rev E 64:021103
42. Sibani P, Brandt M, Alstrøm P (1998) Evolution and extinction dynamics in rugged fitness landscapes. Int J Modern Phys B 12:361–391
43. Sibani P, Dall J (2003) Log-Poisson statistics and pure aging in glassy systems. Europhys Lett 64:8–14
44. Sibani P, Jensen HJ (2005) Intermittency, aging and extremal fluctuations. Europhys Lett 69:563–569
45. Sibani P, Littlewood PB (1993) Slow Dynamics from Noise Adaptation. Phys Rev Lett 71:1482–1485
46. Sibani P, Pedersen A (1999) Evolution dynamics in terraced NK landscapes. Europhys Lett 48:346–352

47. Sibani P, Rodriguez GF, Kenning GG (2006) Intermittent quakes and record dynamics in the thermoremanent magnetization of a spin-glass. Phys Rev B 74:224407
48. Sibani P, Schmidt M, Alstrøm P (1995) Fitness optimization and decay of the extinction rate through biological evolution. Phys Rev Lett 75:2055–2058
49. Sibani P, Schriver P (1994) Phase-structure and low-temperature dynamics of short range Ising spin glasses. Phys Rev B 49:6667–6671
50. Sibani P, Schön C, Salamon P, Andersson J-O (1993) Emergent hierarchical structures in complex system dynamics. Europhys Lett 22:479–485
51. Stillinger FH, Weber TA (1983) Dynamics of structural transitions in liquids. Phys Rev A 28:2408–2416
52. Stillinger FH, Weber TA (1984) Packing Structures and Transitions in Liquids and Solids. Science 225:983–989
53. Struik LCE (1978) Physical aging in amorphous polymers and other materials. Elsevier, New York
54. Swift MR, Bokil H, Travasso RDM, Bray AJ (2000) Glassy behavior in a ferromagnetic p-spin model. Phys Rev B 62:11494–11498
55. Takayama H, Hukushima K (2002) Numerical Study on Aging Dynamics in 3D Ising Spin-Glass Model. III. Cumulative Memory and 'Chaos' Effects in the Temperature Shift Protocol. J Phys Soc JPN 71:3003–3010
56. Tang C, Wiesenfeld K, Bak P, Coppersmith S, Littlewood P (1987) Phase Organization. Phys Rev Lett 58:1161–1164
57. Van Kampen NG (1992) Stochastic Processes in Physics and Chemistry. North Holland, Amsterdam
58. Vincent E (1991) Slow dynamics in spin glasses and other complex systems. In: Ryan DH (ed) Recent progress in random magnets. McGill University, Montreal, pp 209–246
59. Vincent E, Hammann J, Ocio M, Bouchaud J-P, Cugliandolo LF (1996) Slow dynamics and aging in spin-glasses. SPEC-SACLAY-96/048

Books and Reviews

Tinkham M (1998) Introduction to Superconductivity. Krieger, Melbourne
Drossel B (2001) Biological evolution and statistical physics. Adv Phys 50:209–295

Regional Climate Models: Linking Global Climate Change to Local Impacts

Daniela Jacob
Max-Planck-Institute for Meteorology, Hamburg, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Basic Features and Model Characteristics
Validation
IPCC-Scenarios
Regional Climate Change
Regional Extremes
Future Directions
Acknowledgments
Bibliography

Glossary

Climate models They are mathematical representations of the Earth system, in which physical and biogeochemical processes are described numerically. Climate models can be of a global scale or focus on a sub-region (regional climate model).

Downscaling Dynamical and statistical techniques to interpret global climatic changes in specific regions.

IPCC emission scenario Description of possible developments of the socio-economic system expressed in terms of emissions into the atmosphere.

Projection Simulation of possible climatic changes in the future, dependent on emission scenarios, land-use changes and natural variability in the climate system.

Validation Comparison of observed data against model result for quality assessment of the model.

Definition of the Subject

A variety of observations demonstrates that during the last decades the climate has changed. As reported by the *Intergovernmental Panel on Climate Change* (IPCC, 2001, 2007), a mean increase of temperature by 0.09 K per decade was observed globally from 1951 to 1989. Up to now, 2007, this trend has continued. Europe experienced an extraordinary heat wave in summer 2003, with daily mean temperatures being about 10° warmer locally than the long term mean. The increase of temperature varies depending on the region and season.

The temperature change seems to be accompanied by changes in several meteorological and hydrological quantities, like number and duration of heat waves, frost periods, storminess or monthly mean precipitation. In Germany, for example, winter precipitation has increased in parts by more than 30% within the last four decades. In addition, very intense precipitation was observed in summer 2002 in parts of the Elbe drainage basin, which faced a severe flooding.

It can be expected that extreme weather situations will occur more often in a warming world. Therefore, a growing demand from decision-makers and the general public for detailed information on possible future climate development is evident, worldwide. The quantification of risks

associated with changing climates is a prerequisite for the formulation and implementation of realistic adaptation and mitigation strategies [1].

Global climate models (GCM) have been developed to study the Earth's climate system in the past and future. Unfortunately, even today, global climate models provide information only at a relatively coarse spatial scale, which is often not suitable for regional climate change assessments. To fill this gap, two different principles to transfer the information from a global model to the region of interest have been developed accordingly: statistical downscaling and dynamical downscaling. Statistical downscaling techniques connect the climate change signal provided by the GCM with observations from measurement stations in the region to achieve higher resolved climate change signals.

Dynamical downscaling uses high resolution three-dimensional regional climate models (RCM), which are nested into GCMs. RCMs are similar to numerical weather forecasting models, which are taken into account non-linear processes in the climate system. The results of both downscaling methods depend on both the quality of the global and regional models. In the following, the focus will be on dynamical downscaling, in order to be able to also detect more easily new extremes, which have not been observed so far, and to take into account possible feedback mechanisms, which might appear under climate change conditions, and which influence the extent of regional climatic changes. Regional feedback mechanisms are, for example, snow-albedo/temperature feedbacks or soil moisture-temperature feedbacks. If snow melts the reflectivity of the surface changes from bright (white snow) to dark (vegetation or soil). This enhances the absorption of incoming radiation and leads to warming of the surface, which in turn accelerates the snow melt nearby. Evaporation from soils and vegetation increases with temperature and decreases soil moisture. Dryer soils evaporate less, so that cooling due to evaporation is decreasing, which in turn increases temperatures regionally.

Introduction

The climate on Earth varies from region to region, and is characterized by sequences of weather events. These events are determined by the atmospheric flow, established through a wide range of interacting scales. The interacting scales cover large-scale features of the order of thousands to hundreds of km, mostly determined by the distribution of the continents and oceans, solar radiation and the composition of the atmosphere, regional features of the order of a few hundred km to a few km, forced through complex topography and vegetation distribution, and small scale features, like convection. The description of the Earth's climate needs to consider all scales, respectively. Therefore global climate models (GCMs) have been developed. They are mathematical representations of the Earth system (Fig. 1), in which physical and biogeochemical processes are described numerically to simulate the climate system as realistically as possible. Today GCMs develop into Earth system models (ESM), which are not only coupled atmosphere-ocean general circulation models (AOGCM), but also take into account some biogeochemical feedbacks, like the carbon cycle or dynamical vegetation. They are the most advanced numerical tools for climate modeling and describe changes due to large scale forcing.

Increasing greenhouse gas (GHG) concentration, changing aerosol composition and load as well as land surface changes are influencing the climate of the Earth, globally and regionally. Therefore the demands for fine scale regional climate information were growing and in response to this the development of regional climate models started about 20 years ago. It is obvious that the simulations of regional climate changes requires the simulations of processes from global to local scales, so that very high resolution AOGCMs with grid sizes of about 10 km could be seen as the best solution. However, until today the horizontal resolution is still relatively coarse (100 to 250 km grid size) due to limitations in computer resources. Hence AOGCMs cannot provide regional details.

To overcome the deficiency two different approaches are used: statistical downscaling and dynamical downscaling. Both translate the information from the global model to the region of interest. Statistical downscaling techniques connect the climate change signal provided by the GCM with observations from measurement stations in the region to achieve higher resolved climate change signals. Dynamical downscaling uses high resolution three-dimensional regional climate models (RCM), which are nested into GCMs (Fig. 2). RCMs are similar to numerical weather prediction models, in which non-linear processes in the climate system are taken into account. The results of both methods depend on both the quality of the global and regional models. In the following, the focus will be on dynamical downscaling, to be able to detect more easily new extremes, which have not been observed so far, and to take into account possible feedback mechanisms, which might appear under climate change conditions, like the snow-albedo/temperature feedback or the soil moisture-temperature feedback.

Regional climate models are limited area models. They only cover the area of interest, which can be resolved to

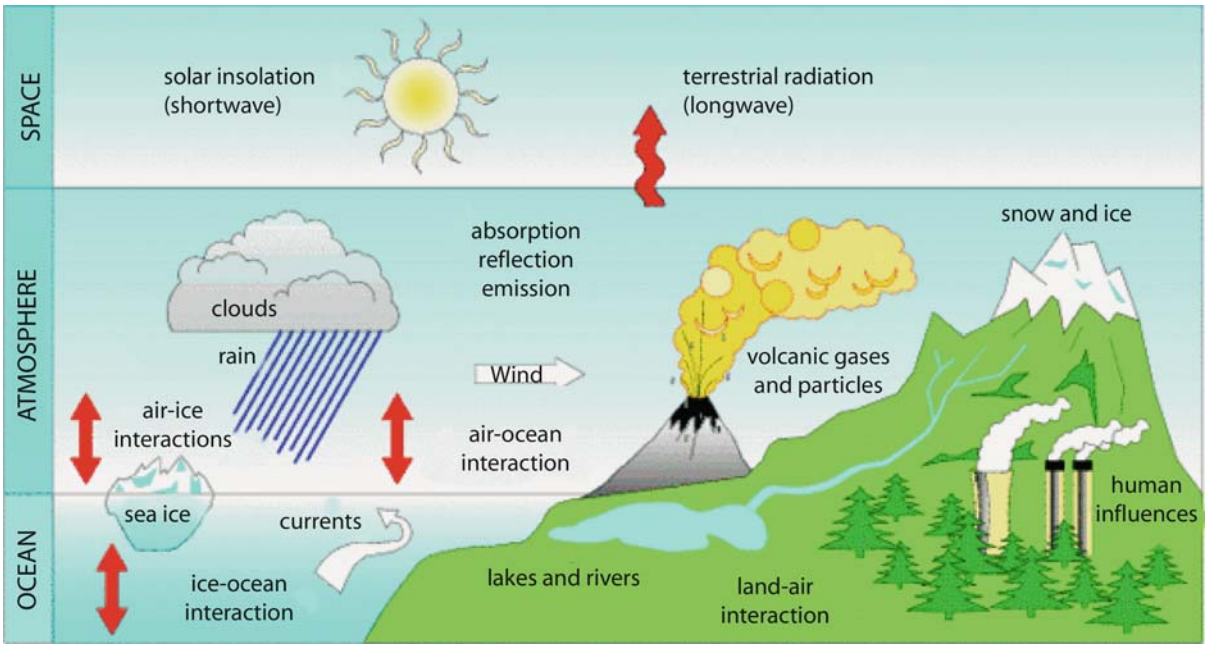

Regional Climate Models: Linking Global Climate Change to Local Impacts, Figure 1
The physical climate system

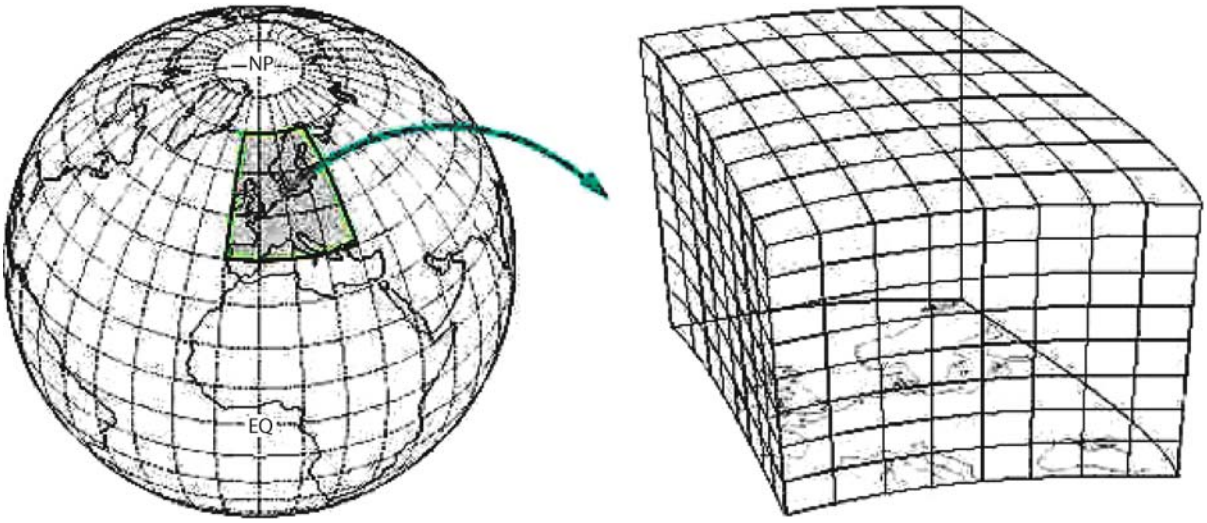

Regional Climate Models: Linking Global Climate Change to Local Impacts, Figure 2
Nesting technique

a much higher degree than GCMs. Therefore RCMs describe the effects of regional and small scale processes within the simulation domain and are connected to the global flow using the nesting technique. This technique was developed to get higher resolution climate information on regional scales, and it is very similar to the nesting procedure in NWP. For initialization and at the lateral boundaries the GCM, in which the RCM is nested, provides information about the state of the atmosphere and the surface conditions. Usually atmospheric fields like wind, pressure, temperature and humidity are provided as well as sea surface temperatures [14,17]. Soil temperature and soil moisture are initialized once, but calculated within the RCMs during the simulation.

The development of regional climate models started in the USA. Filippo Giorgi at NCAR was the first one running the MM4 model in a so-called climate mode [10], which means simulations longer than a few days, as it was common for numerical weather prediction (NWP). For many years RCMs were applied to simulations covering one month. The extension of NWP models to month-long simulations required changes in the formulation of physical processes which were taken into account within the model. In NWP models processes acting on time scales longer than weeks are not important and so not included.

The development of longer-term climate simulations happened very fast and simultaneously in several modeling centers of the world. In the early 1990s, the first multi-year simulations were carried out by Giorgi et al. [11,12], whereas Jones et al. [21,22] and McGregor et al. [27] succeeded in ten-year simulations. Nowadays regional climate simulations stretch from several decades, first achieved by Machenhauer et al. [25], up to more than a century in transient climate change mode [20]. Currently RCMs are widely used for regional climate studies for almost all regions of the world, with horizontal grid spacing ranging from more than 100 km to 10 km. A more detailed overview can be found in Giorgi [9] and for example in [7] focusing on regional climate modeling in the Arctic.

The basic features of RCMs will be explained in Sect. "Basic Features and Model Characteristics" and examples of applications will be presented in Sect. "Validation", climate scenarios in Sect. "IPCC-Scenarios", and examples of applications in Sects. "Regional Climate Change" and "Regional Extremes". A discussion of future perspectives and concluding remarks follow in Sect. "Future Directions".

Basic Features and Model Characteristics

Until today, most RCMs are three-dimensional hydrostatic circulation models, solving the discretized primitive equations of the atmospheric motion. Summaries of model characteristics can be found in many publications, e. g. Jacob et al. [18,19]. As an example for the development and characteristics of many RCMs, the standard set-up of REMO, the regional climate model developed and used at the Max-Planck-Institute for Meteorology is described in more detail below.

The development of REMO started in 1994 utilizing the existing NWP model (EM) of the German Weather Service DWD [26]. Additionally, the physical parametrization package of the general circulation model ECHAM4 [33] has been implemented. During the last decade it could be shown in several applications that the combination of the EM dynamical core plus the ECHAM4 physical parametrization scheme is able to realistically reproduce regional climatic features and therefore became the standard setup.

The atmospheric prognostic variables of REMO are the horizontal wind components, surface pressure, temperature and specific humidity, as well as cloud liquid water. The temporal integration is accomplished by a leap-frog scheme with semi-implicit correction and time filtering after Asselin [2]. REMO is a grid box model, with grid box centers defined on a rotated latitude–longitude coordinate system. For horizontal discretization the model uses a spherical Arakawa-C grid in which all variables except the wind components are defined in the center of the respective grid box. In the vertical, a hybrid vertical coordinate system is applied [35]. Details about the physical parameterizations can be found in Jacob 2001 [16], but will not be explained here in more detail, since they vary slightly from RCM to RCM (see for example [18,19]).

The resolution of the horizontal grids in RCMs varies from about 100 km to 10 km and has increased constantly. For many years 1/2° grid size could be seen as a standard horizontal resolution, which was used in many experiments, even for model inter-comparison studies (e. g. [4,18,25,32]).

REMO uses horizontal grids with 1/12°, 1/6° or 1/2°, corresponding to horizontal resolutions of about 10 km, 18 km and 55 km. In the vertical 20 to 40 levels are applied.

Applying the nesting technique for regional climate models requires large scale atmospheric flow fields to *drive* the RCMs at their lateral boundaries. These fields can be derived from different sources depending on the application. Regional climate simulations require climate change information from AOGCMs, whereas the simulations of the last decades are driven by global analyzes of observations. The analyzes consist of observations, which have been interpolated in space and time using global models. They can be interpreted as the best available representation of the observed atmospheric flow conditions; however, systematic biases cannot be excluded due to the utilization of numerical models for interpolation. In regional climate modeling the use of driving data from analyzes or re-analyzes products are referred to as simulations with perfect boundary conditions (PBC). These experiments have the clear advantage to be directly comparable to observations for the actual time periods and they build the basis for model validation experiments.

In all cases, the relaxation scheme according to Davies [5] is applied in REMO, meaning that the prognostic variables are adjusted towards the large-scale forcing in a lateral sponge zone of 8 grid boxes. Within this zone the

influence of the lateral boundary conditions decreases exponentially towards the inner model domain.

At the lower boundary, RCMs are determined through the interaction with the land surface and, over sea, by the sea surface temperature (SST) and sea ice distribution. The SST can either be interpolated from the large-scale forcing or from observational datasets, or it can be calculated online by a regional ocean model coupled to the RCM, e. g. [23]. The same is true for the sea ice extent, which can as a further option also be diagnosed from the SST. The land surface with its ongoing changes plays a major role in the climate system. Therefore, in all RCMs the exchange between surface and atmosphere is realized by the implementation of a land surface scheme [30]. Generally, one surface grid box can either be covered by water, sea ice or land or can include fractions of land and water areas, all characterized by their own roughness length and albedo. The land fraction of the surface can be covered by bare soil or by vegetation of different type. Depending on the complexity of the land surface scheme the exchange between the atmosphere and the underlying surface is realized through turbulent surface fluxes and the surface radiation flux, which are calculated separately for each fraction and weighted averages of the fluxes are used within the lowest atmospheric model level. Physical properties of the soil and vegetation control the exchange of heat, moisture and momentum over land. In REMO these properties include for example the surface roughness length, the soil field capacity, the water holding capacity of the vegetation, the background albedo, the fractional vegetation cover and the leaf area index (LAI). Some of these parameters strongly depend on the physiological state of the vegetation and are variable between the growing and the dormancy season [31].

There are two options to use the nesting technique. Within the one-way mode a GCM drives a RCM at the lateral boundaries, but no information is given back to the GCM. This method is the standard one used until today in regional climate modeling. It is relatively easy to implement and allows the use of RCMs without running a GCM. The RCM adds information on scales smaller than the driving GCM (e. g. topographical forcing), but is strongly dependent on the superimposed large scale flow. Hence RCMs cannot correct large scale flows originating from GCMs, which might have large errors. However, Giorgi et al. [13], showed that some modulation of the large scale flow is possible within the RCM simulation, stimulated by regional scale forcing.

In the two-way mode, both models GCM and RCM run simultaneously and the RCM feeds back information to the GCM every GCM time step. This method has recently been established for regional climate modeling studies [24]; it has the clear advantage that the atmospheric flow generated within the RCM domain can modulate the large scale flow in areas with strong energetic input from the surface to the atmosphere (e. g. the maritime continent).

If RCM experiments are carried out with high horizontal resolution, like 10 or 20 km, it can be required to use the so-called double nesting technique to avoid mismatch in scales along the lateral boundaries due to the coarse resolution of the driving GCMs. Double nesting means that first a RCM simulation will be carried out with a relatively coarse resolution to generate lateral boundary conditions for further nesting. For REMO, sometimes a sequence of nests is calculated [20].

Finally there are two independent options to run a regional climate model with PBC: the forecast mode and the climate mode. In climate mode the RCM simulation is initialized once from analyzes, and then it is continuously calculated forward in time, driven by regularly up-dated lateral boundaries. In forecast mode, a sequence of short runs (e. g. 30 hours), each initialized every 30 hours from analyzes, is carried out. The forecast mode has the advantage to force the RCM flow to be very close to the observed one, but it has the disadvantage to suppress mesoscale flow features. These mesoscale processes can be excited within the RCM domain by land–sea contrasts or topography and are too small to be taken into account in the GCM.

Validation

The quality of the RCM simulations depends strongly on the performance of the driving model due to the one-way nesting procedure. Therefore, it is extremely important to validate the driving large scale fields before applying RCMs. The model quality, however, can only be judged in comparison with independent observations. Therefore, time periods of the past are simulated and the model results are compared against measurements before the models are used for climate change studies. These comparisons are also part of model development and testing.

As an example, Fig. 3 shows time series of observed and simulated global mean near surface temperatures for the period 1860 to 2000. The simulated results from the global coupled climate model ECHAM5/MPI-OM (Max-Planck-Institute for Meteorology) are in good agreement with ERA40 data, but about 0.5° warmer than the reconstructed observations. The observed increase during the last decades is clearly visible.

As for GCMs, the model quality of RCMs needs to be analyzed. Therefore RCMs are nested into re-analyzed

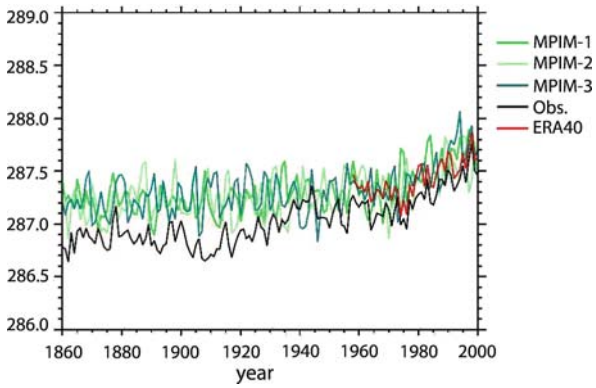

Regional Climate Models: Linking Global Climate Change to Local Impacts, Figure 3
Time series of the global mean near surface air temperature (K): observed (*black*), from reanalyzes (*red*), and from the MPIM global climate model simulations (*green*)

data, which can be seen as close to reality as possible (see above). The results of the RCM simulations of the last decades are compared against independent observations, means as well as extremes are considered. As an example, simulated precipitation climatologies calculated with REMO with two different horizontal grid sizes are compared against observations [8].

The total amount of precipitation and the horizontal pattern are much better resolved using the very high horizontal resolution of about 10 km (Fig. 4). Regional maxima, like the one in the Black Forest, and minima, like in the central valleys of the Alps, are detectable. However, the resolution is still too coarse for climate change studies in individual alpine valleys. The 50 km grid is much finer than standard GCM grids (about 150 to 250 km), but is still insufficient for studying regional details if the regions are too small.

IPCC-Scenarios

The investigation of possible future climate changes requires information about possible changes in the *drivers* of climate change. So-called *drivers* are for example, amount and distribution of aerosols and green house gases (GHG) in the atmosphere, which depend directly on natural and man-made emissions. The IPCC emissions scenarios (Fig. 5) follow so-called story lines, describing possible developments of the socioeconomic system [29].

The emissions are directly used within GCMs and RCMs and they initiate changes in global and regional climates through numerous non-linear feedback mechanisms. As an example, Fig. 6 shows possible developments

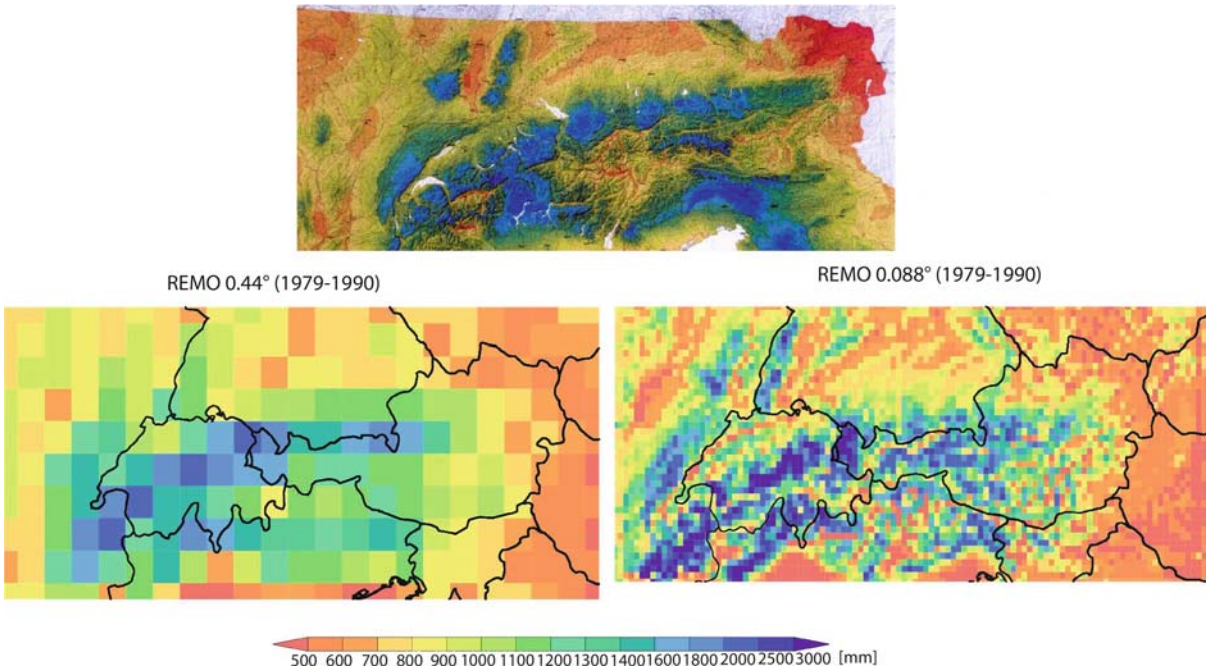

Regional Climate Models: Linking Global Climate Change to Local Impacts, Figure 4
Annual total precipitation (mm), observed (1971–1990, *upper panel*) and simulated with about 50 km grid size (*left*) and 10 km grid size (*right*)

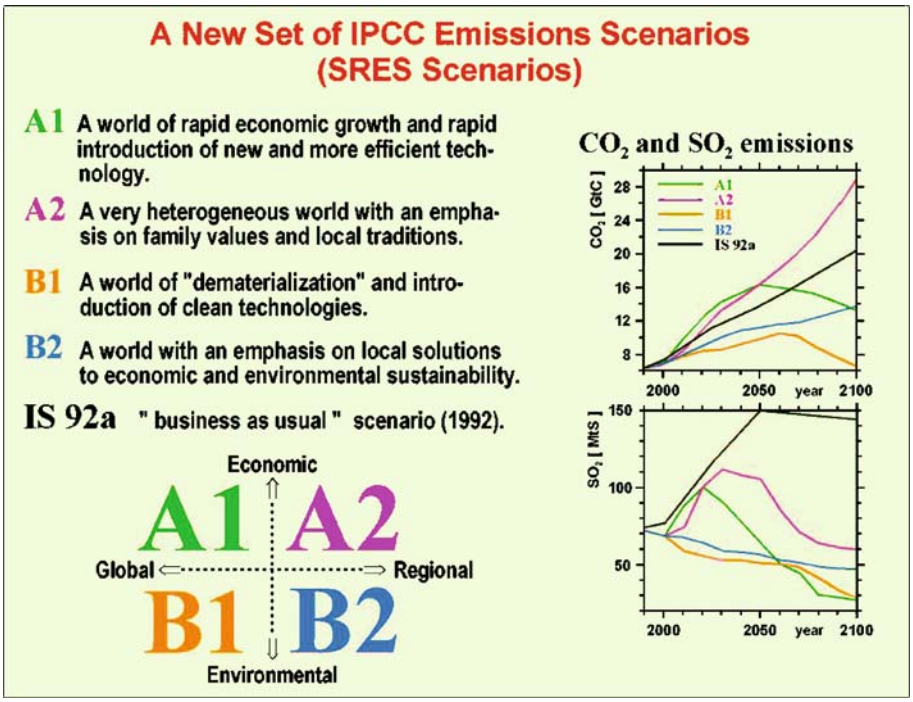

Regional Climate Models: Linking Global Climate Change to Local Impacts, Figure 5
SRES Scenarios, which shows the four major storylines together with the associated developments of CO_2 and SO_2 emissions from 2000 until 2100

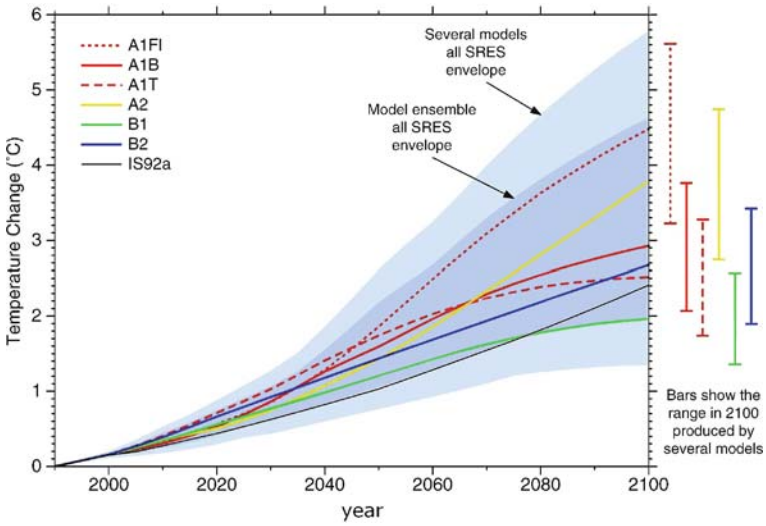

Regional Climate Models: Linking Global Climate Change to Local Impacts, Figure 6
Changes in global mean near surface air temperature as calculated by several GCMs under seven emissions scenarios until 2100

of global mean near surface temperatures calculated by several models for different scenarios.

The global mean changes in near surface temperature until 2050 is about 1.5°C, whereas until the end of the century a wide spread appears from 1.5°C to 5.5°C.

Regional Climate Change

In order to achieve information about the probability, e. g. for the intensification of the hydrological cycle over Europe, several models from different European climate re-

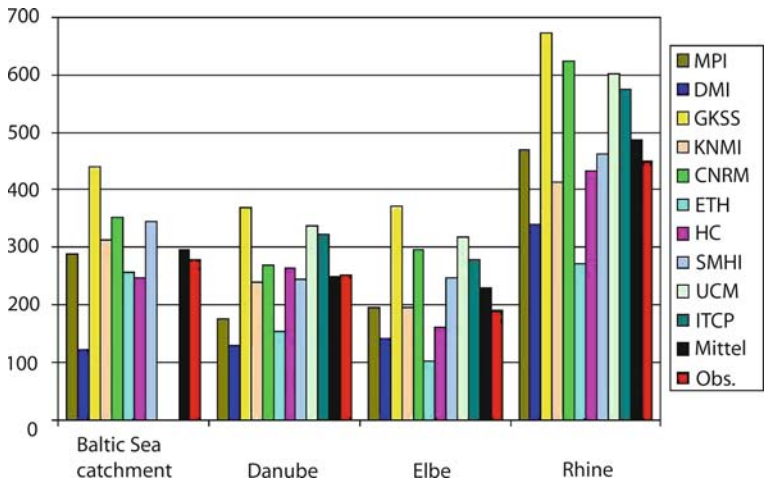

Regional Climate Models: Linking Global Climate Change to Local Impacts, Figure 7
Simulated and observed river run-off (precipitation P – evaporation E) for 1961 to 1990 [15] in the Baltic Sea, Danube, Elbe and Rhine catchments

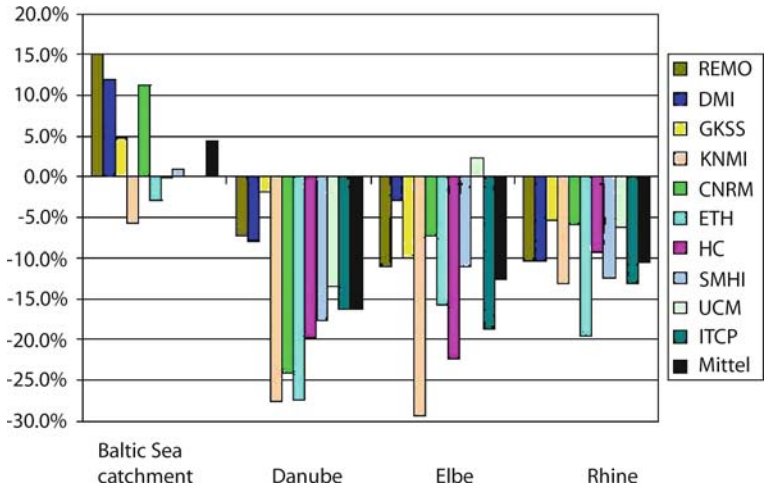

Regional Climate Models: Linking Global Climate Change to Local Impacts, Figure 8
Simulated and observed change in river run-off (precipitation P – evaporation E) for the period 2071 to 2100 compared to 1961 to 1990 [15]. Baltic Sea, Danube, Elbe and Rhine catchments

search institutes are used, as it was done in the EU project PRUDENCE [4].

Following the climate change scenario A2 projecting a strong future increase of greenhouse gases until the year 2100 (IPCC, 2001) and a subsequent global mean temperature increase of about 3.5°, numerous simulations were conducted within PRUDENCE [19]. An analysis of their results for different river catchments [15] shows significant differences between the projected changes over northern and central Europe for the time period 2070–2100 compared to the current climate (1961–1990, Fig. 7).

For the Baltic Sea catchment, a precipitation increase of about +10% for the annual mean is projected, with the largest increase of up to +40% in winter, while a slight reduction of precipitation is calculated for the late summer. Evapotranspiration will increase during the entire year with a maximum increase in winter. These rises in precipitation and evapotranspiration would lead to an increase of river discharge into the Baltic Sea of more than 20% in winter and early spring. Here, the seasonal distribution of discharge is largely influenced by the onset of spring snowmelt.

For the catchments of Rhine, Elbe and Danube, a different change in the water balance components is projected. While the annual mean precipitation will remain almost unchanged, it will increase in late winter (January–

March) and decrease significantly in summer. The evapotranspiration will rise during the entire year, except for the summer, with a maximum increase in winter. These changes lead to a large reduction of 10 to 20% in the annual mean discharge (Fig. 8). Especially for the Danube, the projected summer drying has a strong impact on the discharge that is reduced up to 20% throughout the year except for the late winter (February/March) when the increased winter precipitation causes a discharge increase of about 10%. These projected changes in the mean discharge will have significant impacts on water availability and usability in the affected regions.

Under climate change conditions not only the absolute amounts of precipitation may change but also the precipitation intensities, i. e. the amount of precipitation within a certain time period. The simulation of precipitation intensities or extreme precipitation events requires however a considerably higher resolution than the A2 results presented above so that for example the influence of the topography of the Alps on the formation of precipitation over the Rhine catchment could be adequately calculated. High resolution RCM results show that the global warming until 2050 will lead to an increase of high precipitation events over the Alpine part of the Rhine catchment, especially in summer. This climate change signal becomes clearly visible in the Pre-Alps, but a similar trend is also seen in the high resolution simulations over large parts of Europe.

A major breakthrough was possible with the regional climate change simulations on 10 km grid scale. Within a co-operation with the national environmental agency (UBA), REMO was used for a control simulation from 1950 to 2000 and three transient run for the IPCC SRES scenarios A2, A1B and B1. The simulation domain covers Germany, Austria and Switzerland [20]. As an example the most important results for Germany at the end of this century are summarized as follows:

The simulated annual mean near surface temperature is increasing up to 3.5°C depending on the emission scenario (Fig. 9). The regional pattern of temperature changes shows that the south and southeast warm more than all other areas in the simulation domain. The warming is associated with a decrease of precipitation amount in wide areas of Germany during summer and an increase of precipitation in south and southwest regions during the winter (Fig. 10). The winter precipitation is mostly rain and less precipitation falls as snow.

Regional Extremes

The calculated rapid and strong changes of climate parameters can have severe impacts on humans and the environment. As an example, REMO results for the Rhine basin are presented for a B2 scenario until 2050. Between 1960 and 2050 the near surface temperature will rise by about 3°C and the number of summer days and hot days will increase (Fig. 11). In addition, the number of periods with summer days, this is the period of consecutive days with a daily maximum temperature above 25°C (not shown), will be higher in the future decades. Winter temperature also increases, leading to a decrease in frost and ice days.

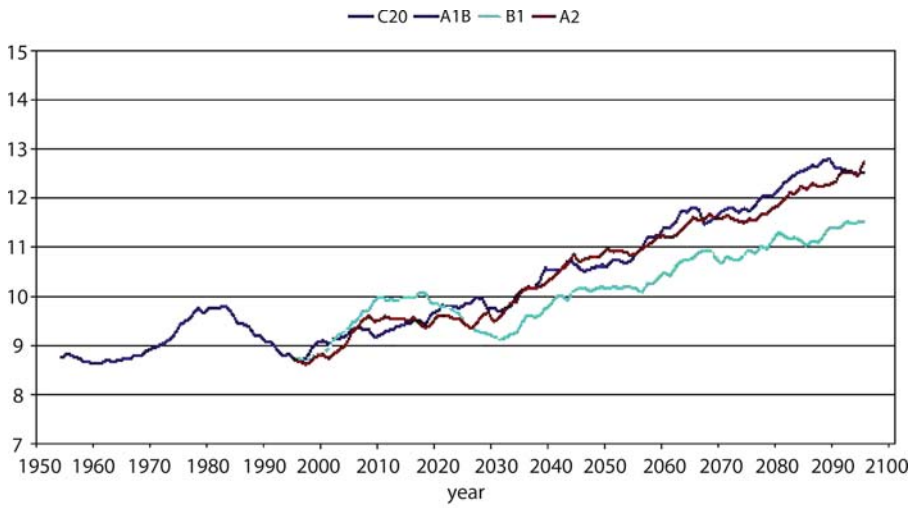

Regional Climate Models: Linking Global Climate Change to Local Impacts, Figure 9
Changes in annual mean near surface temperature (°C) from 1950 to 2100 for three different IPCC SRES scenarios

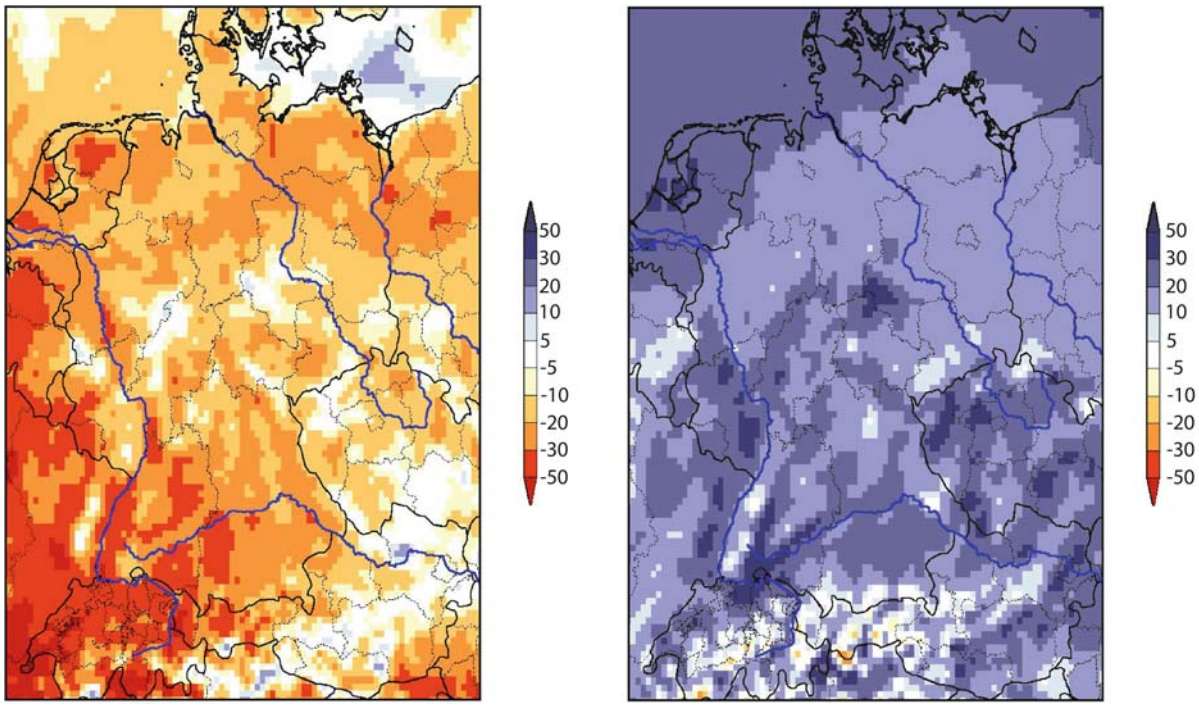

Regional Climate Models: Linking Global Climate Change to Local Impacts, Figure 10
Climate change signals for summer (*left*) and winter (*right*) precipitation (%) in scenario A1B for 2071 to 2100 compared to 1961 to 1990

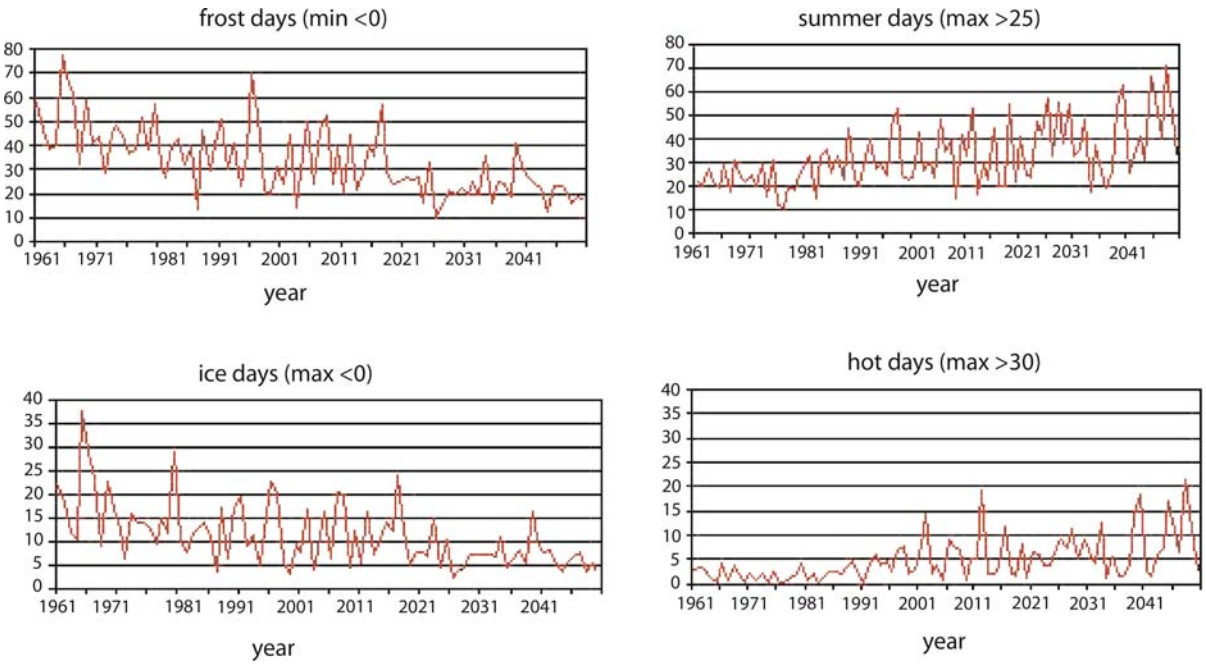

Regional Climate Models: Linking Global Climate Change to Local Impacts, Figure 11
REMO B2 scenario for the Rhine catchment: frost days (*upper left*), ice days (*lower left*), summer days (*upper right*) and hot days (*lower right*)

The investigation of probability distribution functions for temperature and precipitation using the 10 km horizontal resolutions simulations for Germany shows possible monthly mean temperature of more than 30°C for July and >10°C for January appearing in the Rhine valley under the assumption of A1B scenario until the end of this century. In addition, possible increases in monthly mean precipitations are projected for A1B until 2100 in the area of Leipzig (Elbe drainage basin) for January as well as for July (Bülow, Ph D thesis, in preparation).

The projected changes in extremes are coherent with findings of Schär et al. [34] who studied the occurrence of summer heat waves in Europe today and in the future. He showed that the summer 2003 heat wave was extraordinary but can appear much more often in the future.

Another robust finding is the increase of heavy precipitation events in summer, which goes together with a decrease in monthly mean summer precipitation in central Europe [3]. Such short-term, strong convective summer precipitation events have the potential of causing damages, e. g. for agriculture, but also in cities, when sewage systems might be flooded.

Future Directions

Regional climate modeling has made major progress during the last decades and it could be shown that the added value lies mainly in the provision of details in space and time, which is demanded for impact assessments as well as by the public. RCMs are ready for operational use as powerful tools to simulate regional climatic features and their changes in all regions of the world, for time periods from today until the end of this century.

Now, further research on regional climate changes should focus on the reliability in projected regional and local climate change pattern. This can be done by ensemble calculations of GCMs-RCMs, which are also needed for the establishment of probability distribution functions to analyze extreme situations. Here special focus will be on the use of RCMs for regional climate prediction for the next 10 to 30 years, for which it is unclear if weather and climate extremes may change, where it might happen and if regionas will face extreme situations in the future, in which this never happened before. The dissemination of regional climate "change information" can efficiently be done with the establishment of climate information systems in individual regions of the World.

The use of RCMs as intelligent interpolators of observed data in data sparse regions need to be proven, but RCMs have a great potential for this. In addition, the added value of RCMs compared to GCMs needs to be proven by applying the two-way nesting technique. In mountainous regions like the Alps an impact on the large scale flow can be expected.

The extension of regional climate models (RCM) to regional system models (RSM) is a major challenge for the next years. The coupling to regional ocean models, land use and hydrological models has recently started. Along with this the carbon and nitrogen cycles will be implemented on a regional scale. This allows a much better simulation of additional regional feedbacks as mentioned in Sect. "Definition of the Subject", which might have the potential to modulate the regional climate signal projected by GCMs for the region of interest.

Finally, it must be stated that along with improvements of RCMs, the development of GCMs has to be continued. Their performance in individual regions can be quite poor and an improvement is urgently needed. Here RCMs can deliver important information about regional climatic details and help advancing global climate change simulations.

Acknowledgments

I like to thank the REMO-grooup at the Max Planck Institute for Meteorology for their enthusiastic support.

Bibliography

1. Arnell NM (1996) Global warming, river flows and water resources. Wiley, Chichester
2. Asselin R (1972) Frequency filter for time integrations. Mon Weather Rev 100:487–490
3. Christensen JH, Christensen OB (2004) Climate modelling: severe summer flooding in Europe. Nature 421:805–806
4. Christensen JH, Christensen OB (2007) A summary of the PRUDENCE model projections of changes in European climate by the end of this century. Clim Chan 81(1):7–30
5. Davies HC (1976) A latery boundary formulation for multi-level prediction models. Quart J R Meteorol Soc 102:405–418
6. Dethloff K, Abegg C, Rinke A, Hebestad I, Romanov V (2001) Sensitivity of Arctic climate simulations to different boundary layer parameterizations in a regional climate model. Tellus 53(A):1–26
7. Dethloff K, Rinke A, Lynch A, Dorn W, Saha S, Handorf D (2008) Arctic climate change – The ACSYS decade and beyond. Chapter 8: Arctic regional climate models (in press)
8. Frei C, Christensen JH, Deque M, Jacob D, Jones RG, Vidale PL (2003) Daily precipitation statistics in regional climate models: Evaluation and intercomparison for the European Alps. J Geophys Res 108(D3):4124. doi: 10.1029/2002JD002287
9. Giorgi F (2006) Regional climate modeling: Status and perspectives. J Phys IV France 139(2006):101–118. doi: 10.1051/jp4:2006139008
10. Giorgi F, Bates GT (1989) The climatological skill of a regional model over complex terrain. Mon Weather Rev 117:2325–2347

11. Giorgi F, Bates GT, Niemann SJ (1993) The multi-year surface climatology of a regional atmospheric model over the western United States. J Clim 6:75–95
12. Giorgi F, Brodeur CS, Bates GT (1994) Regional climate change scenarios over the United States produced with a nested regional climate model. J Clim 7:375–399
13. Giorgi F, Mearns LO, Shields C, McDaniel L (1998) Regional nested model simulations of present day and 2XCO2 climate over the Central Plains of the US. Clim Chan 40:457–493
14. Giorgi F, Mearns LO (1999) Introduction to special section: Regional climate modelling revisited. J Geophys Res 104:6335–6352
15. Hagemann S, Jacob D (2007) Gradient in the climate change signal of European discharge predicted by a multi-model ensemble. Clim Chan 81(1):309–327
16. Jacob D (2001) A note to the simulation of the annual and inter-annual variability of the water budget over the Baltic Sea drainage basin. Meteorol Atmos Phys 77:61–73
17. Jacob D, Podzun R (1997) Sensitivity studies with the regional climate model REMO. Meteorol Atmos Phys 63:119–129
18. Jacob D, Van den Hurk BJJM, Andrae U, Elgered G, Fortelius C, Graham LP, Jackson SD, Karstens U, Köpken C, Lindau R, Podzun R, Rockel B, Rubel F, Sass BH, Smith RNB, Yang X (2001) A comprehensive model inter-comparison study investigating the water budget during the BALTEX-PIDCAP period. Meteorol Atmos Phys 77:19–43
19. Jacob D, Bärring L, Christensen OB, Christensen JH, Hagemann S, Hirschi M, Kjellström E, Lenderink G, Rockel B, Schär C, Seneviratne SI, Somot S, van Ulden A, van den Hurk B (2007) An inter-comparison of regional climate models for Europe: Design of the experiments and model performance. Clim Chan 81(1):31–52
20. Jacob D, Göttel H, Kotlarski S, Lorenz P, Sieck K (2008) Klimaauswirkungen und Anpassung in Deutschland – Phase 1: Erstellung regionaler Klimaszenarien für Deutschland. Abschlussbericht zum UFOPLAN-Vorhaben 204 41 138, Berichtszeitraum: 1. Oktober 2004 bis 30. September 2007. Max-Planck-Institut für Meteorologie (MPI-M), Hamburg
21. Jones RG, Murphy JM, Noguer M (1995) Simulations of climate change over Europe using a nested regional climate model. I: Assessment of control climate, including sensitivity to location of lateral boundaries. Quart J R Meteorol Soc 121:1413–1449
22. Jones RG, Murphy JM, Noguer M, Keen AB (1997) Simulation of climate change over Europe using a nested regional climate model. II: Comparison of driving and regional model responses to a doubling of carbon dioxide. Quart J R Meteorol Soc 123:265–292
23. Lehmann A, Lorenz P, Jacob D (2004) Modelling the exceptional Baltic Sea flow events in 2002–2003. Geophys Res Lett 31:L21308. doi: 0.1029/2004GL020830
24. Lorenz P, Jacob D (2005) Influence of regional scale information on the global circulation: A two-way nesting climate simulation. Geophys Res Lett 32:L18706. doi: 0.1029/2005GL023351
25. Machenhauer B, Windelband M, Botzet M, Christensen JH, Déqué M, Jones RG, Ruti PM, Visconti G (1998) Validation and analysis of regional present-day climate and climate change simulations over Europe. MPI Report No. 275. MPI, Hamburg
26. Majewski D (1991) The Europa-Modell of the Deutscher Wetterdienst. In: ECMWF seminar on numerical methods in atmospheric models, vol 2. ECMWF, Reading
27. McGregor JL, Katzfey JJ, Nguyen KC (1995) Seasonally varying nested climate simulations over the Australian region. 3rd Int Conf Model Glob Clim Chan Var. Hamburg, Germany, 4–8 Sept 1995
28. McGregor JL, Katzfey JJ, Nguyen KC (1999) Recent regional climate modelling experiments at CISRO. In: Ritchie H (ed) Research activities in atmospheric and oceanic modelling. CAS/JSC Working Group on Numerical Experimentation Report 28. WMO/TD – no. 942. WMO, Geneva, pp 7.37–7.38
29. Nakicenovic N, Alcamo J, Davis G, de Vries B, Fenhann J, Gaffin S, Gregory K, Grübler A, Jung TY, Kram T, La Rovere EL, Michaelis L, Mori S, Morita T, Pepper W, Pitcher H, Price L, Raihi K, Roehrl A, Rogner HH, Sankovski A, Schlesinger M, Shukla P, Smith S, Swart R, van Rooijen S, Victor N, Dadi Z (2000) IPCC special report on emissions scenarios. Cambridge University Press, Cambridge
30. Pitman A (2003) Review: The evolution of, and revolution in, land surface schemes designed for climate models. Int J Climatol 23:479–510
31. Rechid D, Jacob D (2006) Influence of seasonally varying vegetation on the simulated climate in Europe. Meteorol Z 15:99–116
32. Rinke A, Marbaix P, Dethloff K (2004) Internal variability in Arctic regional climate simulations: Case study fort he SHEBA year. Clim Res 27:197–209
33. Roeckner E, Arpe K, Bengtsson L, Christoph M, Claussen M, Dümenil L, Esch M, Giorgetta M, Schlese U, Schulzweida U (1996) The atmospheric general simulation model ECHMA-4: Model description and simulation of present-day climate. Report 218. Max Planck Institute for Meteorology, Hamburg
34. Schär C, Vidale PL, Lüthi D, Frei C, Häberli C, Liniger MA, Appenzeller C (2004) The role of increasing temperature variability in European summer heatwaves. Nature 427:332–336
35. Simmons AJ, Burridge DM (1981) An energy and angular- momentum conserving vertical finite-difference scheme and hybrid vertical coordinate. Mon Weather Rev 109:758–766

Relaxation Oscillations

Johan Grasman
Wageningen University and Research Centre, Wageningen, The Netherlands

Article Outline

Glossary
Definition of the Subject
Introduction
Asymptotic Solution of the Van der Pol Oscillator
Coupled Van der Pol-Type Oscillators
Canards
Dynamical Systems Approach
Future Directions
Acknowledgment
Bibliography

Glossary

Dynamical system At any time t the state of a n-dimensional dynamical system is determined by the state variables forming a vector $x(t)$ in R^n. The change in time of this vector is given by the vector differential equation $\mathrm{d}x/\mathrm{d}t = f(x)$. In the state space R^n solutions of this equation are the trajectories or orbits of the system. If the right-hand side of the equation also depends directly on t, $\mathrm{d}x/\mathrm{d}t = f(t, x)$, then the system is forced and called nonautonomous.

Oscillator An oscillator is a time periodic solution of the system forming a closed orbit in state space. If trajectories, close to a periodic solution, tend to this periodic solution for $t \to \pm\infty$, then the oscillator is called a limit cycle. If a nonautonomous system is periodically forced, $f(t, x) = f(t + T, x)$, then the system may get entrained: Periodic solutions with the same period T or a period kT with some integer larger than 1 (subharmonic solution) may occur.

Singular perturbations If a dynamical system contains a small parameter ε and we let $\varepsilon \to 0$ then the system may become degenerate meaning that it cannot satisfy generic initial- or periodicity conditions. In fact such a system has two time scales: The fast variables vector x and the slow variables vector y: $\varepsilon \mathrm{d}x/\mathrm{d}t = f(x, y)$, $\mathrm{d}y/\mathrm{d}t = g(x, y)$. For $\varepsilon = 0$ initial values outside the manifold $M^{(0)}$: $f(x_0, y_0) = 0$ cannot be satisfied. For small positive ε we see that in the initial phase x changes rapidly until a manifold $M^{(\varepsilon)}$ near $M^{(0)}$ is reached at a part where the points (x_0, y_0) are stable equilibria of $\varepsilon \mathrm{d}x/\mathrm{d}t = f(x, y_0)$. Next the system will enter a quasi-stationary state and changes slowly within $M^{(\varepsilon)}$. With singular perturbations separate approximate solutions are constructed for the two phases. They have the form of power series expansions in ε. Integration constants are determined by a matching procedure.

Relaxation oscillation A relaxation oscillation is a limit cycle of a singularly perturbed dynamical system. Within a cycle at least once the system leaves and returns to the manifold $M^{(\varepsilon)}$.

Canard If within a cycle the relaxation oscillation comes near a part of $M^{(0)}$ with unstable equilibrium points of $\varepsilon \mathrm{d}x/\mathrm{d}t = f(x, y_0)$ then we have a so-called canard type of oscillation. In addition to the parameter ε a second parameter a can be identified that passes a Hopf bifurcation point, where a stable equilibrium changes into a stable relaxation oscillation. Then just before the regular relaxation oscillation arises a canard appears. For very small ε a canard is not easily detected.

Definition of the Subject

A relaxation oscillation is a type of periodic behavior that occurs in physical, chemical and biological processes. To describe it mathematically, a system of coupled nonlinear differential equations is formulated. Such a system is studied with qualitative and quantitative methods of mathematical analysis. The well-known linear pendulum (harmonic oscillator) is not the appropriate system to model real life oscillations. Characteristic for a relaxation oscillator is its nonlinearity and the presence of phases in the cycle with different time scales: A phase of slow change is followed by a short phase of rapid change in which the system jumps to the next stage of slow variation. These oscillations belong to the class of nonlinear systems that give rise to a self-sustained oscillation meaning that the system goes alternately through phases in which energy dissipates and is taken up again.

Van der Pol [66] studied such a type of oscillation in a triode circuit. For small values of a system parameter he found an almost sinusoidal oscillation, while for larger values the system exhibited the type of slow-fast dynamics described above. In the last case the period of the oscillation is almost proportional to that parameter. The name relaxation oscillation, introduced by Van der Pol, refers to this characteristic time constant of the system. In a next publication Van der Pol [67] points out that not only in electronics but in far more fields of science relaxation oscillations may be present. In this respect the physiology of nerve excitation and, more specifically, the heart beat was given special attention by him. In new generations of electronic and electromagnetic systems, such as transistor circuits, Josephson junctions and laser systems, relaxation oscillations are prominently present. Applications are also found in chemistry such as the Zhabotinskii reaction [56], and in geophysics with e. g. the irregular pattern of earthquakes at folds [70]. It is noted that the stick-slip model [71] in the form of a Brownian relaxation oscillator is often brought up in these earthquake studies. In addition to the physiological applications, relaxation oscillations are found in the biology of interacting populations such as in epidemiology [39] and in prey-predator systems, see [12,53,54]. Also in the humanities comparable phenomena are met such as business cycles in economics, see e. g. [31,69].

One way to handle mathematically relaxation oscillations is to exploit the presence of a small number. Dynamical systems with a small parameter ε multiplying the time derivative of some of the state variables degenerate when this parameter is set to zero, as it is not possible anymore to satisfy all initial- or periodicity conditions. In the

theory of singular perturbations the limit process $\varepsilon \to 0$ is followed and the nonuniform convergence of the solution to some limit function is taken in consideration. At the point in time, where this limit solution is discontinuous, a new time scale is introduced by a stretching transformation. Then again the limit $\varepsilon \to 0$ is taken leading to a new (locally valid) limit solution. Integration constants in both limit solutions are found from a matching procedure [15]. In a second approach, based on nonstandard analysis [55], it is not necessary to consider at each step of the computational process this limit procedure. The set of real numbers is extended with infinitesimal numbers being numbers that are nonzero and have an absolute value that is smaller than any real number. Then a solution of a differential equation with an infinitesimally small parameter lies infinitely close to the limit solution in an appropriate function space, see [58,79]. The nonstandard approach of a type of relaxation oscillations known as canards or "French ducks" [6] has given the method an important place in the literature. In a third type of approach the attention fully goes into the analysis of the vector field related to the dynamical system. The onset of a phase of fast change in the period is marked by the passage of a special point in state space (the fold point). By a blow up of the vector field at this point [14] the periodic trajectory can be rigorously described over a time interval containing this point and therefore a full description over the entire period is at hand. It has been worked out for relaxation oscillations [38] and is connected to the dynamical systems theory known as geometrical singular perturbations [19].

Introduction

Periodic processes control our daily life. External influences such as the dynamics of sun, earth and moon play an important role in e. g. the seasons, the tides and our day–night rhythm. However, internally we also have our circadian clock as well as other autonomous periodic processes such as the regulation of our metabolic functions. Periodic electric activities in our brain and heart play a key role in the functioning of these organs. Through the work of Hodgkin and Huxley [34] we have a good understanding of the transport of electric pulses in nerve cells. Looking back we notice that the result of Balthasar van der Pol on periodicity in electric circuits is one of the scientific achievements forming the basis of this breakthrough.

In Fig. 1 Van der Pol's triode circuit is depicted. It produces a self-sustained oscillation due to the nonlinearity in the triode characteristic given by $I_a = V - 1/3V^3$. Replacing V by a scaled potential x and scaling also the time variable we obtain from circuit theory the following differential equation

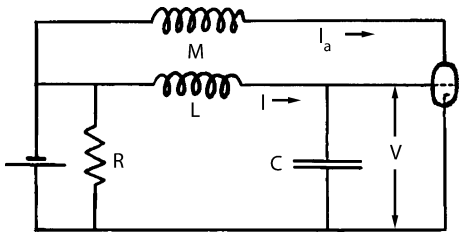

Relaxation Oscillations, Figure 1
Triode circuit giving rise to a self-sustained oscillation. *L* is a self-inductance, *M* a mutual inductance and *R* a resistance. *I* and *V* are, respectively, the current and the grid voltage

$$\frac{d^2x}{d\tau^2} + \mu(x^2-1)\frac{dx}{d\tau} + x = 0, \quad \mu = M/\sqrt{LC} - R\sqrt{C/L}. \tag{1}$$

For μ small a nearly sinusoidal oscillation is found with amplitude close to 2 and a period close to 2π. For $\mu \gg 1$ an almost discontinuous solution appears for which the period is nearly proportional with this parameter, see Fig. 2. The best way to study this last type of oscillation is to rewrite the second-order differential Eq. (1) as a system of two coupled first-order differential equations and to introduce the small parameter $\varepsilon = 1/\mu^2$. Furthermore, a new time scale is introduced $t = \tau/\mu$. It leads to the so-

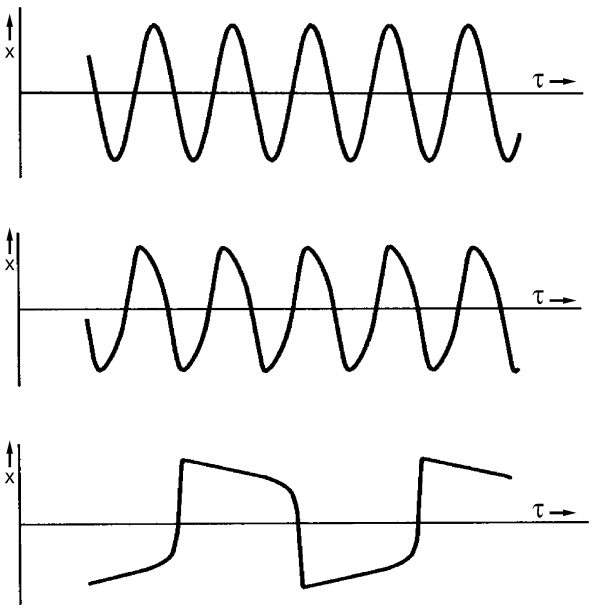

Relaxation Oscillations, Figure 2
Periodic solution of Eq. (1) for a $\mu = 0.1$, b $\mu = 1$, c $\mu = 10$

called Lienard-type of representation of the system [41]:

$$\varepsilon \frac{\mathrm{d}x}{\mathrm{d}t} = y - F(x), \quad F(x) = 1/3\, x^3 - x, \quad 0 < \varepsilon \ll 1, \quad (2)$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -x + a\,, \quad a = 0\,. \quad (3)$$

Discontinuous Limit Solution

In Fig. 3a it is seen how the periodic solution of (2) behaves in the limit $\varepsilon \to 0$. It has the form of a discontinuous solution consisting of two time intervals in which the system follows alternately in state space two branches (DA and BC) of the graph of $y = F(x)$ with $|x| > 1$, see Fig. 3b. These are the stable branches; they are rapidly approached because of a fast changing variable x if $y \neq F(x)$. In between the two branches there is the unstable branch AC at which two trajectories move away from the unstable equilibrium at the origin. At the point where a stable branch becomes unstable the system jumps to the opposite stable branch. Clearly the amplitude of the discontinuous oscillation has the value 2 in the x variable. The period follows from

$$T_0 = 2 \int_{-2/3}^{2/3} \frac{\mathrm{d}t}{\mathrm{d}y} \mathrm{d}y = 2 \int_{-2}^{-1} \frac{1}{-x} F'(x) \mathrm{d}x = 3 - 2 \ln 2\,. \quad (4)$$

This discontinuous solution can be seen as a zero-order asymptotic approximation of the solution with respect to the small parameter ε. In Sect. "Asymptotic Solution of the Van der Pol Oscillator" we will deal with higher-order approximations using the theory of singular perturbations.

Canards

If in formula (2) the parameter a is varied an intriguing type of Hopf bifurcation arises. When passing the values $a = \pm 1$ for decreasing absolute value of a, a stable equilibrium turns unstable and a relaxation oscillation arises. The scenario of a common Hopf bifurcation is that a stable equilibrium changes into an unstable one at the bifurcation point and that a periodic solution branches off with an amplitude that grows in the beginning quadratically as a function of the distance of the parameter to the just passed bifurcation point. However, the above discontinuous periodic solution suggests that directly after the bifurcation point a fully developed relaxation oscillation arises. Using concepts of nonstandard analysis [55] French mathematicians [6] explain how the curious emergence of a relaxation type of oscillation at a Hopf bifurcation can be understood. Also by an intricate asymptotic analysis the phenomenon can be described with singular perturbation theory [16]. In Sect. "Canards" we sketch the result. Finally, as part of a dynamical systems approach (Sect. "Dynamical Systems Approach") a geometrical singular perturbation analysis can be carried out as well [38].

Bonhoeffer–Van der Pol Equation or FitzHugh–Nagumo Equation

When we take in (3) $a = -1 - \delta, 0 < \delta \ll 1$ with δ independent of ε, a stable equilibrium $(\underline{x}, \underline{y})$ arises with $\underline{x} = -1 - \delta$. A small perturbation of the x-component of the system, causing a deviation that stays below 2δ, will damp out quickly. A positive perturbation just above this threshold (dotted line) will trigger a cycle of the system as depicted in Fig. 4. FitzHugh [21] constructed a similar variant of the Van der Pol Equation (2) providing a mathematical model of a nerve excitation as described by Bonhoeffer [3]. Firing of a neuron occurs when an electric stimulation of a dendrite is above some threshold value triggering an electric pulse in the neuron itself which is passed through the axon to other neurons. Also the pres-

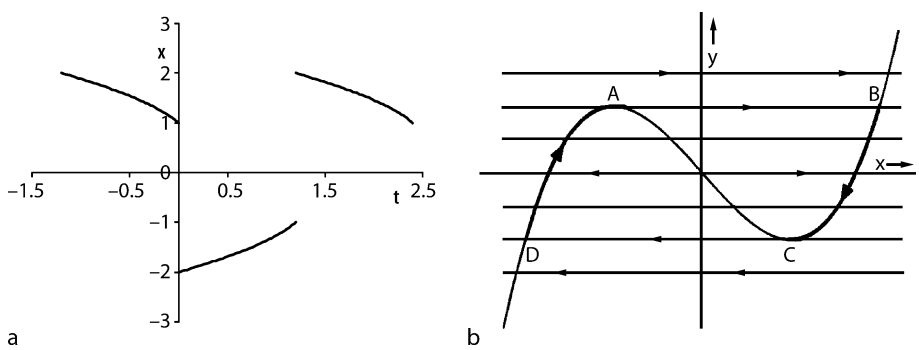

Relaxation Oscillations, Figure 3
The discontinuous approximation of the solution of the Van der Pol equation for $\varepsilon \to 0$. The parts AB and DC of the orbit are taken infinitely fast. a The function $x(t)$ for $\varepsilon \to 0$. b The closed orbit in the x, y-plane

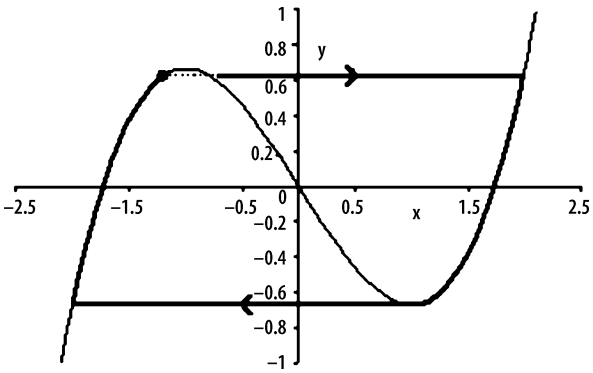

Relaxation Oscillations, Figure 4
A perturbation (*dotted line*) of the equilibrium (●) above threshold triggers a cycle. During the cycle perturbations do not have a large effect (refractory period)

ence of a so-called refractory period can be understood from the above system. The refractory period is the time directly after an above threshold stimulation. During this time the neuron is insensitive to perturbations. Later on more refined models evolved [57]. The introduction of spatial structure and diffusion made it possible to consider traveling pulses [1,34].

Orbital Stability, Entrainment, Chaos, Quenching, and Noise

In applications we typically meet nonlinear oscillations that are orbitally highly stable but rather easily speeded up or slowed down in their cycle. In a biological context organisms are provided in this way with a mechanism that helps them to adapt to external circumstances. Also in physics we meet such behavior in systems with strong energy exchange with the environment (lasers), as opposed to conservative systems without dissipation such as in celestial systems.

In (2) entrainment is found if we let the parameter a depend periodically on time with a period say T. If the period $T^{(0)}$ of the autonomous system with $a = 0$ in (2) is sufficiently close to the forcing period T or if the amplitude of the forcing is sufficiently large, then the system will take over this period T. If $T^{(0)}$ is near a value nT, $n = 2, 3, \ldots$ a subharmonic solution with period nT may arise. It also may occur that two stable subharmonics with different n values co-exist. Then the starting value determines which one is chosen. Van der Pol and Van der Mark [68] and Littlewood [43,44] already concluded on respectively, experimental and theoretical grounds that in such a case also other "strange" (chaotic) solutions may be present in the Van der Pol-type oscillator with periodic forcing, see [4,40,42]. For autonomous systems chaotic dynamics can be found in systems consisting of at least three components [23]. A system with one fast and two slow variables already exhibits a wealth of interesting dynamical features, including chaos. Quenching of an oscillation can be achieved by extending the system with a set of differential equations including a feedback to the original system such that the amplitude of the oscillation is reduced by choosing appropriate values for the parameters. For an application to relaxation oscillations see [72,74].

Relaxation oscillations are met in a wide range of applications. Since in practice the action of random forces mostly cannot be excluded, stochastic oscillation forms an essential part of the theory of periodic phenomena in the description of natural processes as well as in the engineering sciences, see [27,33]. An additional reason to take stochasticity in account comes from the fact that the phase velocity of the relaxation oscillator is easily influenced.

Higher-Order Systems and Coupling of Oscillators

Making a generalization we consider the periodic solution of a system of differential equations of the form

$$\varepsilon \frac{\mathrm{d}x}{\mathrm{d}t} = f(x, y; \varepsilon), \quad \frac{\mathrm{d}y}{\mathrm{d}t} = g(x, y; \varepsilon), \quad 0 < \varepsilon \ll 1, \tag{5}$$

where x and y are respectively k- and l-dimensional vector functions of time. The vector functions f and g remain bounded for $\varepsilon \to 0$. Then the (slow) dynamics is governed by $y'(t) = g(x, y; 0)$ with constraint $0 = f(x, y; 0)$ and for the fast dynamics we have $\varepsilon x'(t) = f(x, y; 0)$ with y constant. An extension of the Van der Pol-type dynamics to such a higher-dimensional system is not difficult to conceive. However, formulating precisely the conditions for the system (5) to have a periodic solution and to work out a full proof for the existence of such a solution is a rather complex task [46]. As we mentioned before, chaotic solutions may occur too.

Another way to arrive at higher-order systems is to couple relaxation oscillators [76]. For relaxation oscillators of type (2) with different autonomous periods we take

$$\begin{aligned} \varepsilon \frac{\mathrm{d}x_i}{\mathrm{d}t} &= y_i - F(x_i)\,, \\ \frac{\mathrm{d}y_i}{\mathrm{d}t} &= -c_i x_i + \sum_{j=0_{j\neq i}}^{n} \delta p_{ij} x_j\,, \quad 0 < \varepsilon \ll \delta \ll 1\,. \end{aligned} \tag{6}$$

For a system of $n = 100$ oscillators with a high degree of coupling, e. g. $p_{ij} \neq 0$ for all $j \neq i$, and widely different frequencies, e. g. $c_i = 0.5 + i/n$, we will observe (after a spin up) a spectrum of frequencies with distinct peaks.

These peaks are related by a simple ratio of their frequencies such as 1:2 or 3:2 and so on, see [17,25]. Considering spatially distributed oscillators with nearest neighbor coupling we meet interesting phenomena such as traveling phase waves. The corresponding fronts may have a preferred direction given by the gradient in the autonomous periods (inherently faster oscillators are likely ahead in phase). Locally, oscillators tend to have an equal actual period (phase velocity) although their intrinsic periods are different. This phenomenon, called plateau behavior, has been investigated by Ermentrout and Kopell [18]. Wave fronts may also move in spirals or travel randomly breaking down when meeting each other as we see in the Zhabotinskii reaction [20] or during fibrillation of the heart [62].

Asymptotic Analysis with Respect to the Small Parameter

By letting $\varepsilon \to 0$ the Van der Pol oscillator (2) reduces to the system:

$$(x_0^2 - 1)\frac{\mathrm{d}x_0}{\mathrm{d}t} + x_0 = 0 \tag{7}$$

having a discontinuous solution

$$t = \ln(x_0) - \tfrac{1}{2}(x_0^2 - 1) \quad \text{for} \quad -\tfrac{1}{2}T_0 < t < 0\,, \tag{8}$$

$$t = \ln(-x_0) - \tfrac{1}{2}(x_0^2 - 1) + \tfrac{1}{2}T_0 \quad \text{for} \quad 0 < t < \tfrac{1}{2}T_0 \tag{9}$$

with T_0 given by (4). This approximation, see Fig. 3a, is improved in Sect. "Asymptotic Solution of the Van der Pol Oscillator" by the construction of higher-order terms with respect to the small parameter ε. Contrary to regular perturbation problems, where a power series expansion with respect to ε holds over the entire period, we have to consider here separate power series expansions for the stable branches. Moreover, we have to exclude a small neighborhood of the point $t = 0$ where the discontinuity occurs. Near $(t, x) = (0, 1)$ a local approximation is made based on fractional powers of ε. It is followed by an internal layer solution approximating the fast change in x from the value 1 to the starting value -2 at the next stable branch. Unknown (integration) constants are determined by matching the local solutions at $t = 0$ to the regular solutions at the two stable branches, see Sect. "Asymptotic Solution of the Van der Pol Oscillator". Matching of local asymptotic solutions typically applies to differential equations with a small parameter multiplying the highest derivative such as in (1) with τ and μ replaced by t and ε, see [15,50]. In the literature on relaxation oscillations the procedure has been carried out in different ways depending on whether one chooses to solve the problem in the x, t-plane [8], the $x, \dot{x}$-plane [13], or the Lienard plane. For a survey of this literature see [25]. For higher-order systems the construction of matched local asymptotic solutions that involve higher-order terms with respect to ε turns out to be rather complicated, see [46].

A different type of approximation, based on power series expansion with respect to the parameter $\mu = 1/\sqrt{\varepsilon}$, was made possible by the use of computerized formula manipulation packages. In this way elements of a periodic solution such as period and amplitude are approximated also for large values of the parameter μ, see [2,10].

Topological Methods, Mappings and the Dynamical Systems Approach

The existence of a periodic solution for (2) has been proved with the Poincaré–Bendixson theorem. This method, based on topological arguments, only applies to two-dimensional systems. If in the x, y-plane we can construct a domain that does not contain equilibrium points and all trajectories crossing the boundary are entering the domain, then the domain contains a limit cycle. If the domain is a narrow annulus enclosing the limit cycle the method also produces an approximation of the solution together with its period and amplitude [52].

Another important tool in the dynamical systems approach is the Poincaré map or transition map. For autonomous n-dimensional systems we may consider starting values at a bounded transversal $(n-1)$-dimensional surface and analyze how the trajectories intersect some other surface in the state space or the same surface (returnmap). In periodically forced systems we may consider the mapping of the full state space into itself after a time interval equal to the forcing period. Such a type of mapping may also apply to the phase of an oscillator or to the phases of a system of coupled oscillators. A periodic solution corresponds with a fixed point of the return map or, in the case of a forced system, with a fixed point of the map of the state space into itself after one or more periods of the forcing term. Chaotic solutions may occur due to the presence of a so-called horse-shoe map [61]. Levi [40] used a map of this type to describe chaotic solutions of a piecewise linear Van der Pol equation with a sinusoidal forcing term.

Presently, relaxation oscillations are also analyzed with geometrical singular perturbation theory [64]. Geometrical singular perturbations deal with the structure of the slow manifold $f(x, y; 0) = 0$ containing trajectories of (5) with $\varepsilon = 0$ that vary slowly in time [19]. A point of the slow manifold is an equilibrium of the fast system in the

time scale $\xi = t/\varepsilon$: $\mathrm{d}x/\mathrm{d}\xi = f(x, y^{(0)}; 0)$. It is assumed that the eigenvalues of this system linearized at the equilibrium have real parts that are bounded away from zero (hyperbolicity condition). For relaxation oscillations this condition is not satisfied [73]. More details are given in Sect. "Dynamical Systems Approach".

Asymptotic Solution of the Van der Pol Oscillator

From the different methods to approximate asymptotically the solution of (2) we choose the approach of Carrier and Lewis [8], who consider the solution in the t, x-plane, see Fig. 5. Writing (2) again as a second-order differential equation

$$\varepsilon \frac{\mathrm{d}^2 x}{\mathrm{d}t^2} + (x^2 - 1)\frac{\mathrm{d}x}{\mathrm{d}t} + x = 0\,, \tag{10}$$

we can expand the solution valid for the stable branch as

$$x(t;\varepsilon) = x_0(t) + \varepsilon x_1(t) + \varepsilon^2 x_2(t) + \cdots\,. \tag{11}$$

Substitution of (11) in (10) gives, after equating terms with equal powers in ε, a recurrent system of differential equations for the coefficients $x_i(t)$ of (11) with $x_0(t)$ given by (7) when taking the branch for $t < 0$. We see from (8) that for $t \uparrow 0$ this first term behaves as $x_0 \approx 1 + \sqrt{-t}$, so the second-order derivative cannot be neglected near the origin. To analyze the local behavior of the solution in more detail we apply a stretching transformation

$$t = \xi\varepsilon^{2/3}\,, \quad x = 1 + \varepsilon^{1/3} v(\xi)\,. \tag{12}$$

Substitution in (10) yields for the leading terms the equation

$$\frac{\mathrm{d}^2 v_0}{\mathrm{d}\xi^2} + 2v_0\frac{\mathrm{d}v_0}{\mathrm{d}\xi} + 1 = 0 \quad \text{or} \quad \frac{\mathrm{d}v_0}{\mathrm{d}\xi} + v_0^2 + \xi = B\,. \tag{13}$$

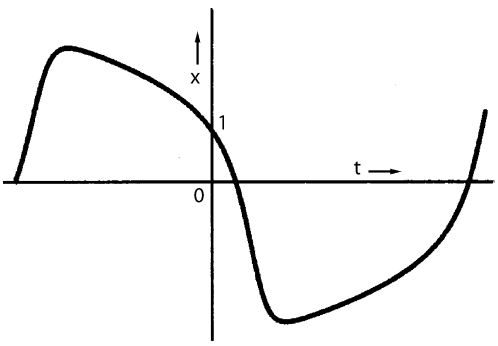

Relaxation Oscillations, Figure 5
The solution in the t, x-plane with the fast change just after $t = 0$ starting with a specific local behavior near the point (0,1)

In order to match x_0 for $\xi \to -\infty$ the solution must satisfy $v_0(\xi) \approx \sqrt{-\xi}$ when taking this limit. The solution

$$v_0(\xi) = -\frac{Ai'(-\xi)}{Ai(-\xi)} \tag{14}$$

with $Ai(z)$ the Airy function complies with this matching condition. As $-\xi$ approaches $\alpha = -2.33811\ldots$, being the first zero of the Airy function, the solution will behave as $v_0(\xi) \approx 1/(\xi - \alpha)$. Consequently, the first two terms of (10) will be leading in the left hand side of the equation and make the solution enter the phase of fast change for which we choose the appropriate time variable η using the transformation

$$t = \alpha\varepsilon^{2/3} + \eta\varepsilon\,, \tag{15}$$

so

$$\frac{\mathrm{d}^2 w_0}{\mathrm{d}\eta^2} + (w_0^2 - 1)\frac{\mathrm{d}w_0}{\mathrm{d}\eta} = 0 \quad \text{or} \quad \frac{\mathrm{d}w_0}{\mathrm{d}\eta} = w_0 - \frac{1}{3}w_0^3 + D\,. \tag{16}$$

Matching with v_0 for $\eta \to -\infty$ yields $D = -2/3 - \alpha\varepsilon^{2/3}$ and integrating Eq. (16) we obtain

$$-\eta = \frac{1}{1 - w_0} + \frac{1}{3}\ln\left(\frac{w_0 + 2 + 1/3\alpha\varepsilon^{2/3}}{1 - w_0}\right)\,. \tag{17}$$

For $\eta \to \infty$ the solution behaves as $w_0 = -2 - 1/3\alpha\varepsilon^{2/3} + O(\exp(-3\eta))$. This must match the regular asymptotic solution at the stable branch for $t > 0$ for which the zero-order approximation satisfies

$$t = \ln(-x_0) - \tfrac{1}{2}(x_0^2 - 1) + E \tag{18}$$

giving $E = 3/2 - \ln(2) + 3/2\alpha\varepsilon^{2/3}$. At $x_0 = -1$ the variable t is at the value $\frac{1}{2}T_0$, so

$$T_0 \approx 3 - 2\ln(2) + 3\alpha\varepsilon^{2/3}\,. \tag{19}$$

The amplitude A follows from the minimal value of w_0 which is approached in (17) for $\eta \to \infty$:

$$A \approx 2 + 1/3\alpha\varepsilon^{2/3}\,. \tag{20}$$

For a higher-order approximation of both the period and the amplitude we refer to Grasman [25]. In Grasman et al. [30] the Lyapunov exponents are approximated in a similar way.

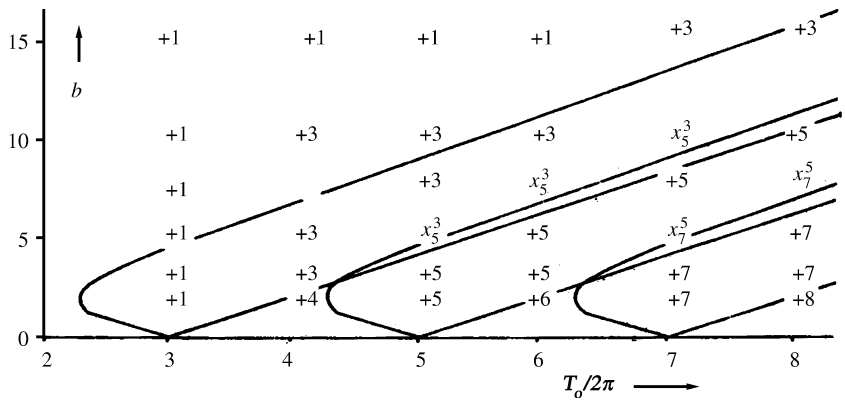

Relaxation Oscillations, Figure 6
Numerical solutions of period $2\pi m$ with m depending on μ and b

Sinusoidal Forcing

We next consider the Van der Pol oscillator with a sinusoidal forcing term

$$\frac{d^2x}{d\tau^2} + \mu(x^2-1)\frac{dx}{d\tau} + x = (b + c\mu)\cos(k\tau)\,, \quad \mu \gg 1\,. \tag{21}$$

A first mathematical investigation of this problem was made by Cartwright and Littlewood [9] followed by Littlewood [43,44]. Subharmonic solutions were constructed. Solutions with different periods may coexist and also solutions that behave chaotically may turn up. Following the method of Carrier and Lewis for this nonautonomous Eq. (21) with $c = 0$ and $k = 1$ we go through the same sequence of asymptotic solutions: The regular approximation of the type (11), the local solution when the trajectory crosses the lines $x = \pm 1$, giving rise to an expression with Airy functions, and a fast changing part directly after with a local approximation similar to (17), see [29]. We may construct various periodic solutions. The most prominent ones have period $2\pi m$ with m an odd number. The value of m depends on μ. In Fig. 6 we depicted the domains in the parameter plane where these solutions are found. Their boundaries follow from the construction of the asymptotic solution. The result is compared with the numerical solution of the system. Notice the overlap of these domains and the coexistence of different numerical solutions. Also chaotic solutions can be approximated asymptotically. They come with the phenomenon of dips and slices which also occurs in the general case for $c \neq 0$ [28]. The main difference with the previous case is that instead of Airy functions parabolic cylinder functions turn up. In Fig. 7 we see how dips and slices arise. They occur when the trajectory stays close to the regular solution

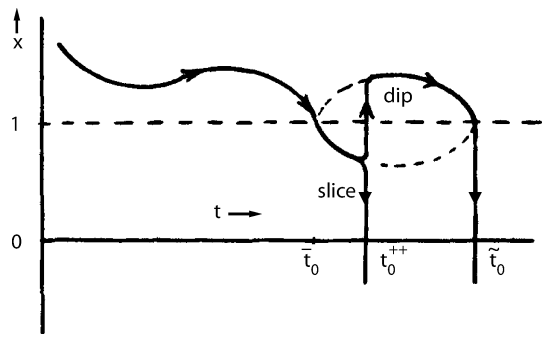

Relaxation Oscillations, Figure 7
A dip and a slice for slightly different initial values chosen such that upon arrival at the line $x = 1$ a point is passed where stable ($x > 1$) and unstable ($x < 1$) regular solutions meet

below the line $x = 1$, while switching from a stable branch ($x > 1$) to an unstable branch ($x < 1$) of this regular solution. It can be seen as a canard type of phenomenon as noted by Smolyan and Wechselberger [63].

Coupled Van der Pol-Type Oscillators

The discontinuous periodic solution of (2) for $\varepsilon \to 0$ is given by (7) and has the form of the vector $\{x_0(t), y_0(t)\}$. We next consider (2) with the parameter a replaced by a small amplitude piecewise continuous periodic function $\delta h(t)$ with period T:

$$\varepsilon\frac{dx}{dt} = y - F(x)\,, \quad \frac{dy}{dt} = -cx + \delta h(t)\,, \quad c = 1\,, \quad 0 < \varepsilon \ll \delta \ll 1\,. \tag{22}$$

For $\varepsilon \to 0$ and δ sufficiently small the solution will take the same orbit as for $\delta = 0$; only the velocity on the limit cycle

is influenced. Consequently, a solution of Eq. (22) can be approximated by

$$\{x_0(\phi(t)), y_0(\phi(t))\} . \tag{23}$$

Substitution in the Eq. (7) yields

$$\frac{\mathrm{d}y_0}{\mathrm{d}\phi}\frac{\mathrm{d}\phi}{\mathrm{d}t} = -x_0(\phi(t)) + \delta h(t) \quad \text{or} \quad \frac{\mathrm{d}\phi}{\mathrm{d}t} = 1 - \frac{\delta h(t)}{x_0(\phi(t))} . \tag{24}$$

Starting with a phase $\phi(0) = \alpha_0$ and integrating over one period T gives a phase shift

$$\delta\psi(\alpha_0) = \int_0^T \frac{\delta h(t)}{x_0(\phi(t))}\mathrm{d}t . \tag{25}$$

Thus, the dynamics of the oscillator is described in a stroboscopic way by the iteration map

$$\alpha_{k+1} = \alpha_k + T + \delta\psi(\alpha_k)(\mathrm{mod})T_0 . \tag{26}$$

Besides a periodic solution of period T different types of subharmonic solutions can be found as well as chaotic solutions depending on the function $\psi(\alpha)$. An entrained solution corresponds with a stationary point α_s of the map (26). Such a point exists if for some $m = 1, 2, \ldots$

$$\{\delta\psi(\alpha)\}_{\min} < mT_0 - T < \{\delta\psi(\alpha)\}_{\max} . \tag{27}$$

Mapping of the phase in the above way is a well-known approach of modeling entrainment by the use of circle maps [51]. It is noted that here we go back further and relate this map to an oscillator given by a differential equation.

If we consider a system of identical oscillators $\{x_0(\phi_i(t)), y_0(\phi_i(t))\}$, $i = 1, 2, \ldots, n$ with a forcing of oscillator i by the others of the form $\delta\Sigma h_{ij}(x_j, y_j)$, we are in the position to analyze different forms of mutual entrainment such as wave phenomena in systems of spatially distributed oscillators with nearest neighbor coupling as described in Sect. "Introduction". A system of spatially distributed oscillators with nearest neighbor coupling can be seen as a discrete representation of a nonlinear diffusion problem with oscillatory behavior. A time delay in the coupling can be handled as well, see [7,26].

It is also allowed that the orbits and the autonomous period of the oscillators are different. As an example we take 25 piecewise linear oscillators with in (22)

$$F(x) = x + 2 \quad \text{for } x < -1 ,$$
$$F(x) = x - 2 \quad \text{for } x > 1 ,$$
$$F(x) = x \quad \text{for } |x| \le 1$$

and $\delta = 0.0025$. For this example an analytic expression for $\psi(\alpha)$ can be derived. For oscillator i the parameter c is in (6) replaced by $c_i(\delta) = 1 - q_i\delta$ with q_i being a random number generated by the normal distribution with expected value zero and standard deviation, so that the autonomous period is $T_0^{(i)} = T_0(1 + \delta q_i)$. As the forcing function of oscillator i we choose $h = \Sigma h_{ij}(x_j) = \Sigma x_j$ with $j \neq i$. The evolution of the dynamics with the oscillators starting in a uniformly distributed random phase has the form of an iteration map of the type (26) and is depicted in Fig. 8. It is observed that by the type of coupling the oscillators are slowed down considerably and that in the almost fully entrained state ($t = 50T_0$) the inherently faster oscillators are slightly ahead in phase.

Canards

We return to the Van der Pol equation in the Lienard-form (2–3) and apply a translation of the state variables x and y as well as the parameter a such that $(x, y; a) = (1, -2/3; 1)$ moves to the origin $(0, 0; 0)$:

$$\varepsilon\frac{\mathrm{d}x}{\mathrm{d}t} = y - F(x), \quad F(x) = 1/3x^3 + x^2, \qquad 0 < \varepsilon \ll 1, \tag{28}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -x + a . \tag{29}$$

For $-2 < a < 0$ with a fixed and $\varepsilon \to 0$ a relaxation oscillation is found as we analyzed in Sect. "Asymptotic Solution of the Van der Pol Oscillator" using the asymptotic method of Carrier and Lewis [8]. The computation was done for $a = -1$, but any value at this interval would have given the same result. For $a < -2$ or $a > 0$ a stable equilibrium turns up. In a small interval near either the point -2 or the point 0 the solution exhibits a drastic change. Eckhaus [16] analyzes that more explicitly near these values critical points exist, where in a small interval the solution rapidly changes from an equilibrium into a full grown relaxation oscillation. Near $a = 0$ we have the critical point

$$\alpha_c(\varepsilon)\} = -1/8\varepsilon + \cdots , \tag{30}$$

where such a change takes place. The following stretching transformation reveals the continuous change of the limit solution as a function of the parameter a:

$$a = \alpha_c(\varepsilon) + \sigma\varepsilon^{3/2}\exp(-k^2/\varepsilon) . \tag{31}$$

In Fig. 9 we sketch two periodic limit solutions with σ having different signs. First notice that for $\sigma = o(1)$ a solution

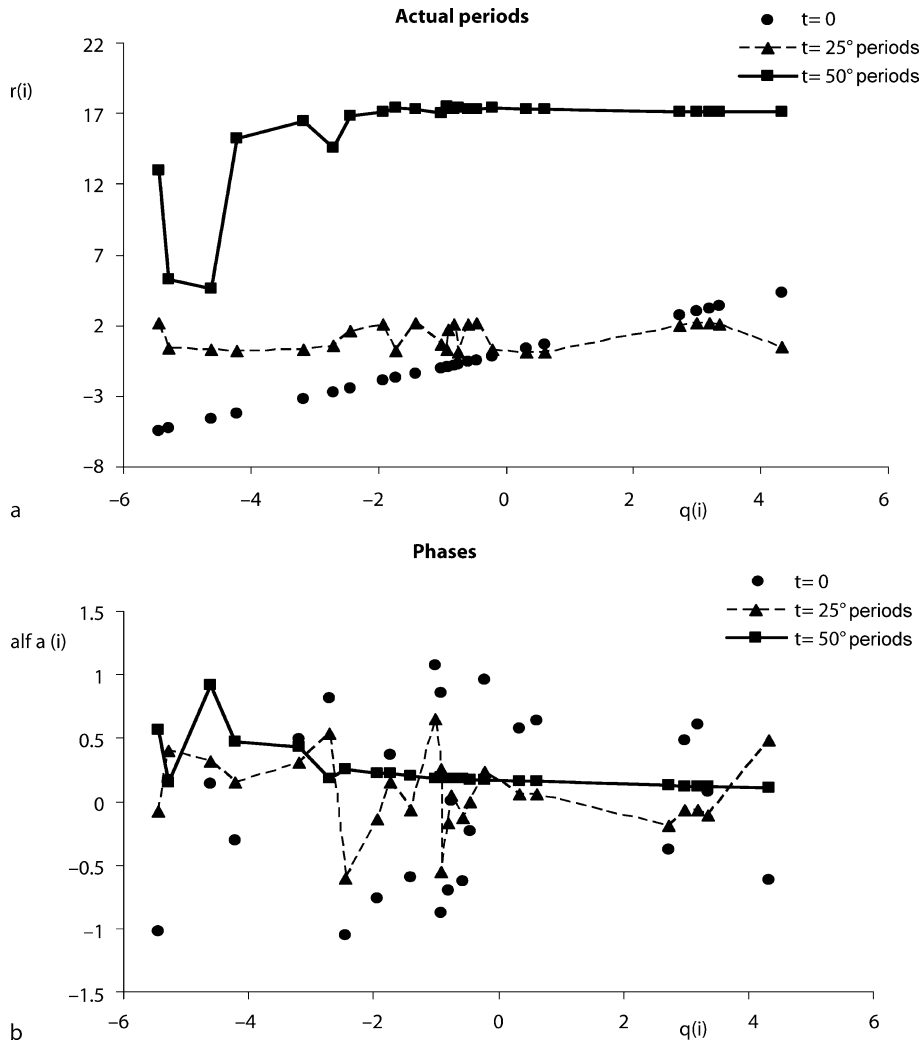

Relaxation Oscillations, Figure 8

A system of coupled piecewise linear Van der Pol relaxation oscillators with autonomous period $T_0^{(i)} = T_0(1 + \delta q_i)$, $i = 1, \ldots, 25$. The phase of each oscillator runs from $-\frac{1}{2}T_0 = -\ln(3)$ to $-\frac{1}{2}T_0$. The actual phase and period at times $t = nT_0$ are depicted. The actual period $T_i(n)$ is given by its deviation from T_0: $r_i = (T_i(n) - T_0)/\delta$. This expression compares with q_i (for $n = 0$, the *points* are situated at the diagonal)

that is on the stable branch ($x > 0$) will keep following the unstable branch until the point where it meets the other stable branch ($x = -2$) and will jump then back to the stable branch where it started. For $\sigma > 0$ we have a periodic solution that in the limit for $\sigma \to \infty$ transforms into an equilibrium solution (Fig. 9a). For $\sigma < 0$ we have a periodic solution having the shape of a duck that in the limit for $\sigma \to -\infty$ takes the form of a fully developed relaxation oscillation (Fig. 9b).

The phenomenon, that is revealed here, is called a "canard explosion". It is basically due to the fact that a trajectory is very close to a stable branch and that at the moment the branch becomes unstable it keeps on going close to the unstable branch because the deviation from this branch does not grow rapidly, see also [24,60]. A similar phenomenon is met in "slow passage through bifurcation" [45].

Dynamical Systems Approach

We start from the system (5) with vector functions $x(t)$ and $y(t)$ and consider the fast dynamics by making a transformation to the fast time scale $\xi = t/\varepsilon$ and keeping $y = y^{(0)}$

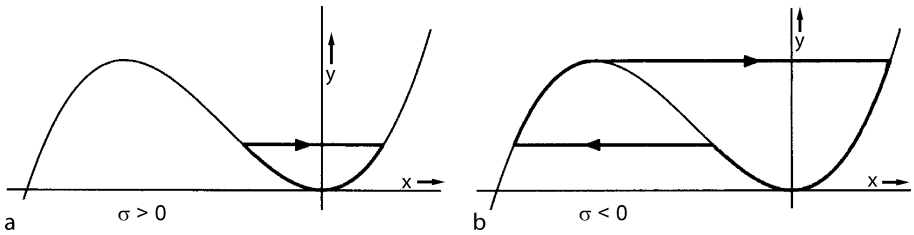

Relaxation Oscillations, Figure 9
Two stages in the canard explosion with parameter values given by (31)

fixed:

$$\frac{dx}{d\xi} = f(x, y^{(0)}; 0) \,. \tag{32}$$

In the dynamical systems approach trajectories of (5) with $\varepsilon = 0$ exist, called fibers having $y = y^{(0)}$ for all t and with x given by (32). Let them move to the point $(x^{(0)}, y^{(0)})$ at the smooth manifold $M^{(0)}$ given by $f(x, y; 0) = 0$. Thus, the point $(x^{(0)}, y^{(0)})$ at $M^{(0)}$ is a stable equilibrium of the fast system (32). The equilibria are the base points of the fibers. Then next at the manifold $M^{(0)}$ the slow dynamics is governed by

$$\frac{dy}{dt} = g(x, y; 0) \quad \text{with constraint} \quad f(x, y; 0) = 0 \,. \tag{33}$$

In the geometrical singular perturbation theory of Fenichel [19] a condition for the equilibrium $(x^{(0)}, y^{(0)})$ of (32) is formulated. Fenichel only considers parts of $M^{(0)}$ for which at all points $(x^{(0)}, y^{(0)})$ the Jacobian $\partial f(x, y; 0)/\partial x$ has eigenvalues with a real part that is bounded away from zero. This is called the hyperbolicity condition. It is noted that eigenvalues with positive real parts are allowed. Then the corresponding eigenvectors span a subspace $V^{(u)}$ for which in the system (32) linearized at $(x^{(0)}, y^{(0)})$ all trajectories move away from the equilibrium. For (32) itself a manifold $W^{(u)}$ is defined with $V^{(u)}$ tangent to this manifold at $(x^{(0)}, y^{(0)})$. The manifold $W^{(u)}$ consists of the set of unstable fibers. Similarly the manifold $W^{(s)}$ is defined containing the set of stable fibers. It is proved that, for ε positive and sufficiently small, near $M^{(0)}$ a slow manifold $M^{(\varepsilon)}$ exists satisfying (5). The fast dynamics is governed by the fibers keeping in mind that we move from one fiber to its neighboring one as described by the slow change of the fiber base point at $M^{(\varepsilon)}$. When approximating asymptotically the solution, a transformation is preferred that differs slightly from the matched asymptotic expansions approach. We illustrate it with the following simple system with scalar functions $x(t)$ and $y(t)$:

$$\varepsilon \frac{dx}{dt} = y - x \,, \quad \frac{dy}{dt} = 1 \,. \tag{34}$$

Then for $\varepsilon = 0$ we have fast fibers with $y =$ constant moving towards the slow manifold $M^{(0)}$: $y = x$, see Fig. 10. For small positive ε the slow manifold is defined by $M^{(\varepsilon)}$: $y = x + \varepsilon$. The motion at this manifold (and the change of the base point of the actual fiber) is governed by $dy/dt = 1$. For the fast motion parallel to the fibers we make the transformation $x = y - \varepsilon + v$, so that the new variable v represents the distance to the base point of the fiber. In the fast time scale we then obtain for the motion parallel to the fibers $dv/d\xi = -v$. For more complex systems the expressions for $M^{(\varepsilon)}$ and for the solutions of the slow and fast systems take the form of power series expansions with respect to ε. For more details we refer to [36,37].

For relaxation oscillations the hyperbolicity condition is not satisfied. This is easily verified from the Van der Pol Eq. (2) where $M^{(0)}$: $y = 1/3x^3 - x$, because at the fold points $x = \pm 1$ the eigenvalue of the Eq. (2) linearized at these x values equals zero. To restore hyperbolicity a blow up technique [14] is used. It applies to the system (2) extended with the differential equation $d\varepsilon/dt = 0$ for which the fold point $(x, y, \varepsilon) = (1, -2/3, 0)$ as well as $(-1, 2/3, 0)$ and their direct neighborhoods are put under a magnifying glass. We follow the exposition by Krupa and Szmolyan [38] and consider the neighborhood of a fold point situated at the origin for the following system in the fast time scale

$$\frac{dx}{d\xi} = -y + x^2 + a(x, y, \varepsilon) \,, \quad \frac{dy}{d\xi} = -\varepsilon + b(x, y, \varepsilon) \,, \quad \frac{d\varepsilon}{d\xi} = 0 \,, \tag{35}$$

where the terms $a(x, y, \varepsilon)$ and $b(x, y, \varepsilon)$ can be neglected if we are sufficiently close to the origin, see Fig. 11a for the dynamics near the fold point. The blow up transformation $(x, y, \varepsilon) \to (\bar{x}, \bar{y}, \bar{\varepsilon}, \bar{r})$ takes the form

$$x = \bar{r}\bar{x} \,, \quad y = \bar{r}^2 \bar{y} \,, \quad \varepsilon = \bar{r}^3 \bar{\varepsilon} \,. \tag{36}$$

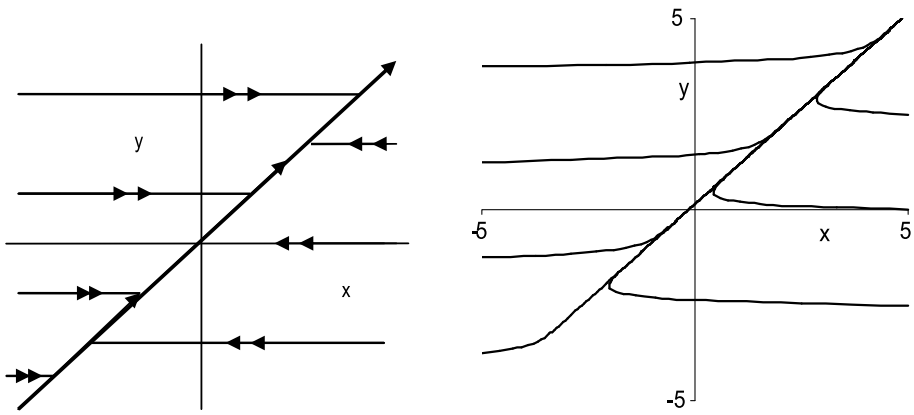

Relaxation Oscillations, Figure 10
The dynamics of (34) for $\varepsilon = 0$ (*left*) with the fast flow represented by the lines y is constant (fibers) and the slow dynamics by the line $y = x$ being the manifold $M^{(0)}$. For small positive ε (*right*) the trajectories cross the fibers (given for $\varepsilon = 0$) until they arrive at an exponentially small neighborhood of the slow manifold $M^{(\varepsilon)}$

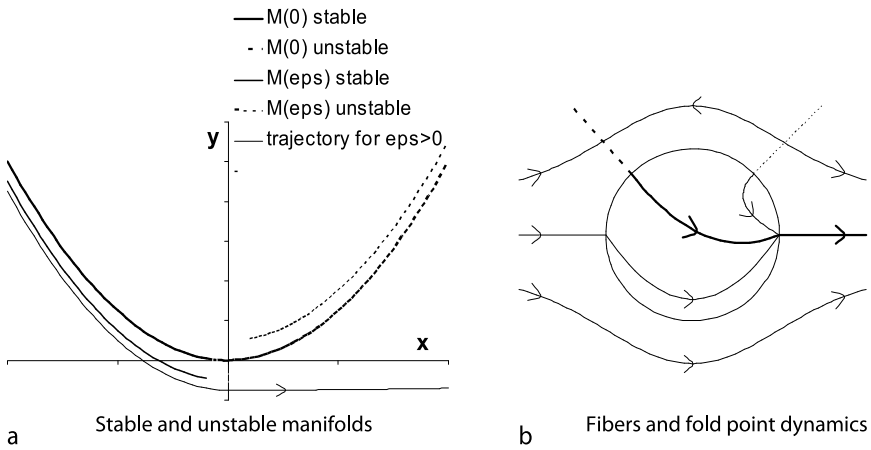

Relaxation Oscillations, Figure 11
Dynamics near and in a fold point, see the text for a detailed description

It is such that in R^3 a ball $B = S^2 \times [0, \rho]$ with ρ sufficiently small is mapped onto R^3. By this transformation the new coordinates $\bar{x}$, $\bar{y}$ and $\bar{\varepsilon}$ indicate a point at a spherical surface S^2. In Fig. 11b a circle is depicted that corresponds with the fold point ($\bar{r} = 0$) in a blow up. Inside the circle, where $\bar{\varepsilon}$ is positive and $\bar{r} = 0$, we are in S^2 above the equator, while at the circle we are at the equator where $\bar{\varepsilon} = 0$. Outside the circle we have $\bar{\varepsilon} = \varepsilon = 0$ but now with $\bar{r} > 0$. There transformed fibers of the limit system near the fold point are drawn. Note that the slow manifold consists of equilibria because of the fast time scale. The bold lines indicate the path of the orbit as it passes the fold point. Away from the equator the dynamics near the fold point is described by taking the fixed value $\bar{\varepsilon} = 1$. We introduce a new time scale $\bar{\xi} = \bar{r}\xi$ and let $\bar{r} \to 0$. In that way we arrive at the system

$$\frac{d\bar{x}}{d\bar{\xi}} = -\bar{y} + \bar{x}^2 , \quad \frac{d\bar{y}}{d\bar{\xi}} = -1 , \tag{37}$$

having the same solution as (13). For more details we refer to Krupa and Szmolyan [38], who proceed with a similar analysis of a canard explosion. Furthermore, the method has been extended to relaxation oscillations in R^3 [64], including canards, see [5,11,63]. In [75] the method is applied to a prey-predator system.

Future Directions

Over more than 75 years studies on relaxation oscillations have appeared in the literature. It is noticed that each

time a new development takes place in physics or mathematics, relaxation oscillations turn out to be an interesting subject to be investigated with such a novel theory. In electronics we find the use of nonlinear devices starting with triodes in electric circuits [66]. In mathematics we see that relaxation oscillation is a rewarding subject to which many new theories can be applied. In chronological order we mention: Rigorous theories for estimating solutions of differential equations [9], singular perturbation theory and matched asymptotic expansions [8], nonstandard analysis [6], catastrophe theory [69], chaos theory [40], and geometrical singular perturbations [38]. The reason for this lies in the fact that a relaxation oscillation is a highly nonlinear phenomenon with internal mechanisms that are close to real life functions like adaptation by entrainment and threshold behavior in case of triggering. Possible new topics in the field of nonlinear oscillation can, in particular, be found in biology. Literature in which models of biological processes are formulated may function as a source of inspiration. In particular we mention Winfree [78] and Murray [47,48]. To be more explicit about new opportunities we bring up the following cases.

The study of the dynamics of oscillations by circle maps or phase equations turned out to be very useful in particular for the description of entrainment of biological oscillators. In Sect. "Coupled Van der Pol-Type Oscillators" we made the connection between a differential equation model for state variables and the phase equations, see (22)–(26), see also [35]. However, the result only holds for the discontinuous solution as we see from (25), where the denominator of the integrand jumps over the value zero. How in the limit for $\varepsilon \to 0$ we arrive at this result has not yet been worked out. We furthermore mention that spatial distribution of oscillators and the way they are coupled may give rise to interesting entrainment phenomena. In the first place there is the nearest neighbor coupling. Taking an appropriate limit in case of diffusive coupling we may model the dynamics by a nonlinear diffusion equation. Although studies in this direction exist, see e. g. [49], still a lot of work can be done. Furthermore, observations of rhythms in neural networks with long distance coupling show special forms of entrainment, see [65]. Mathematical analysis of such a network of coupled relaxation oscillators may help to understand more from these phenomena. Not only the distance can be varied also the type of coupling; we may consider excitatory and inhibitory forcing referring to the effect they have on an oscillator of the Bonhoeffer–Van der Pol-type, see [77]. Besides easy entrainment of coupled relaxation oscillators there is a second form of adaptation which is comparable with that of neural computing. The coupling between two oscillators in a network may increase if they are entrained for a longer period of time and decrease if they are not. In this way the network is structured by the input and if the input changes the network will follow. The neural theory of cell assemblies [32] is based on such a mechanism known as plasticity of the network, see [22,59]. Effects due to a combination of long range coupling and plasticity may occur as well [79].

Studies on systems with complex interactions between many oscillators as we sketched above may take different directions. On one hand large-scale computer simulations may offer a wealth of results, but models that are more conceptual and are investigated (partly) analytically may give more insight into the fundamental properties of such systems. However, it is not easy to see at forehand that a study of the second type will lead to a meaningful result and is for that reason a more risky enterprise.

Acknowledgment

The author is grateful to Ed Veling (Technical University Delft, The Netherlands) for his valuable advice.

Bibliography

Primary Literature

1. Agüera y Arcas B, Fairhall AL, Bialek W (2003) Computation on a single neuron: Hodglin and Huxley revisited. Neural Comput 15:1715–1749
2. Andersen CM, Geer J (1982) Power series expansions for the frequency and period of the limit cycle of the Van der Pol equation. SIAM J Appl Math 42:678–693
3. Bonhoeffer KF (1948) Activation of passive iron as a model for the excitation of nerve. J Gen Physiol 32:69–91
4. Braaksma B, Grasman J (1993) Critical dynamics of the Bonhoeffer–Van der Pol equation and its chaotic response to periodic stimulation. Physica D 68:265–280
5. Brons M, Krupa M, Wechselberger M (2006) Mixed mode oscillations due to the generalized canard phenomenon. Fields Inst Comm 49:39–63
6. Callot JL, Diener F, Diener M (1978) Le problème de la 'chasse au canard'. CR Acad Sc Paris Sér A 286:1059–1061
7. Campbell SR, Wang DL (1998) Relaxation oscillators with time delay coupling. Physica D 111:151–178
8. Carrier GF, Lewis JA (1953) The relaxation oscillations of the Van der Pol oscillator. Adv Appl Mech 3:12–16
9. Cartwright M, Littlewood J (1947) On nonlinear differential equations of the second order II. Ann Math 48:472–494
10. Dadfar MB, Geer JF (1990) Resonances and power series solutions of the forced Van der Pol oscillator. SIAM J Appl Math 50:1496–1506
11. De Maesschalck P, Dumortier F (2005) Canard solutions at generic turning points. In: Dumortier F, Broer HW, Mawhin J, Vanderbauwhede A, Van Duyn Lunel S (eds) Proc Equadiff 2003. World Scientific, Singapore, pp 900–905
12. Deng B (2004) Food chain chaos with canard explosion. Chaos 14:1083–1092

13. Dorodnicyn AA (1947) Asymptotic solution of Van der Pol's Equation. Akad Nauk SSSR Prikl Mat Mech 11:313–328, Russian (1962, Am Math Soc Transl Series 1 4:1–23)
14. Dumortier F (1993) Techniques in the theory of local bifurcations: Blow up, normal forms, nilpotent bifurcations, singular perturbations. In: Schlomiuk D (ed) Structures in bifurcations and periodic orbits of vector fields. Kluwer, Dordrecht, pp 19–73
15. Eckhaus W (1973) Matched asymptotic expansions and singular perturbations. North-Holland Mathematics Studies 6. North Holland, Amsterdam
16. Eckhaus W (1983) Relaxation oscillations including a standard chase on French ducks. In: Verhulst F (ed) Asymptotic analysis II. Lecture Notes in Math, vol 985. Springer, Berlin, pp 449–494
17. Ermentrout GB (1985) Synchronization in a pool of mutually coupled oscillators with random frequencies. J Math Biol 22:1–10
18. Ermentrout GB, Kopell N (1984) Frequency plateaus in a chain of weakly coupled oscillators I. SIAM J Math Anal 15:215–237
19. Fenichel N (1997) Geometrical singular perturbation theory for ordinary differential equations. J Differ Equ 31:53–98
20. Fife PC (1985) Understanding the patterns in the BZ reagent. J Stat Phys 39:687–703
21. FitzHugh R (1955) Mathematical models of threshold phenomena in the nerve membrane. Bull Math Biophys 17:257–278
22. Gerstner W, Kistler WM (2002) Spiking neuron models: Single neurons, populations, plasticity. Cambridge University Press, Cambridge
23. Ginoux JM, Rossetto B (2006) Differential geometry and mechanics applications to chaotic dynamical systems. Int J Bifurc Chaos 16:887–910
24. Gorelov GN, Sobolev VA (1992) Duck-trajectories in a thermal explosion problem. Appl Math Lett 5:3–6
25. Grasman J (1987) Asymptotic methods for relaxation oscillations and applications. Springer, New York
26. Grasman J, Jansen MJW (1979) Mutually synchronized relaxation oscillators as prototypes of oscillating systems in biology. J Math Biol 7:171–197
27. Grasman J, Roerdink JBTM (1989) Stochastic and chaotic relaxation oscillations. J Stat Phys 54:949–970
28. Grasman J, Nijmeijer H, Veling EJM (1984) Singular perturbations and a mapping on an interval for the forced Van der Pol relaxation oscillator. Physica D 13:195–210
29. Grasman J, Veling EJM, Willems GM (1976) Relaxation oscillations governed by a Van der Pol equation with periodic forcing term. SIAM J Appl Math 31:667–676
30. Grasman J, Verhulst F, Shih S-H (2005) The Lyapunov exponents of the Van der Pol oscillator. Math Meth Appl Sci 28:1131–1139
31. Hamburger L (1934) Note on economic cycles and relaxation oscillations. Econometrica 2:112
32. Hebb DO (1949) The organization of behavior: A neuropsychological theory. Wiley, New York
33. Hilborn RC, Erwin RJ (2005) Fokker–Planck analysis of stochastic coherence in models of an excitable neuron with noise in both fast and slow dynamics. Phys Rev E 72:031112
34. Hodgkin A, Huxley A (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. J Physiol 117:500–544
35. Izhikevich EM (2000) Phase equations for relaxation oscillators. SIAM J Appl Math 60:1789–1804
36. Jones CKRT (1994) Geometric singular perturbation theory In: Arnold L (ed) Dynamical systems, montecatini terme, lecture notes in mathematics, vol 1609. Springer, Berlin, pp 44–118
37. Kaper T (1999) An introduction to geometric methods and dynamical systems theory for singular perturbation problems. In: O'Malley RE Jr, Cronin J (eds) Analyzing multiscale phenomena using singular perturbation methods, proceedings of symposia in applied mathematics, vol 56. American Mathematical Society, Providence, pp 85–132
38. Krupa M, Szmolyan P (2001) Extending geometric singular perturbation theory to nonhyperbolic points-fold and canard points in two dimensions. SIAM J Math Anal 33:286–314
39. Lenbury Y, Oucharoen R, Tumrasvin N (2000) Higher-dimensional separation principle for the analysis of relaxation oscillations in nonlinear systems, application to a model of HIV infection. IMA J Math Appl Med Biol 17:243–161
40. Levi M (1981) Qualitative analysis of periodically forced relaxation oscillations. Memoirs of the American Mathematical Society, no 244. American Mathematical Society, Providence
41. Lienard A (1928) Etude des oscillations entretenues. Revue Générale de l'Electricité 23:901–946
42. Lin KK (2006) Entrainment and chaos in a pulse driven Hodgkin–Huxley oscillator. SIAM J Appl Dyn Syst 5:179–204
43. Littlewood JE (1957) On non-linear differential equations of the second order III $y'' - k(1-y^2)y' + y = b\mu k \cos\mu + \alpha$ for large k and its generalisations. Acta Math 97:267–308
44. Littlewood JE (1957) On non-linear differential equations of the second order IV. Acta Math 98:1–110
45. Marée GJM (1996) Slow passage through a pitchfork bifurcation. SIAM J Appl Math 56:889–918
46. Mishchenko EF, Rosov NK (1980) Differential equations with small parameters and relaxation oscillations. Plenum, New York
47. Murray JD (2002) Mathematical biology I: An introduction. Springer, New York
48. Murray JD (2003) Mathematical biology II: Spatial models and biomedical applications. Springer, New York
49. Nishiura Y, Mimura M (1989) Layer oscillations in reaction-diffusion systems. SIAM J Appl Math 49:481–514
50. O'Malley RE Jr (1991) Singular perturbation methods for ordinary differential equations. Springer, New York
51. Osipov CV, Kurths J (2001) Regular and chaotic phase synchronization of coupled circle maps. Phys Rev E 65:016216
52. Ponzo PJ, Wax N (1965) On certain relaxation oscillations: Asymptotic solutions. SIAM J Appl Math 13:740–766
53. Rinaldi S, Muratori S (1992) Slow-fast limit cycles in predator-prey models. Ecol Model 61:287–308
54. Rinaldi S, Muratori S, Kuznetsov YA (1993) Multiple attractors, catastrophes and chaos in seasonally perturbed predator-prey communities. Bull Math Biol 55:15–35
55. Robinson A (1966) Non-standard analysis. North-Holland, Amsterdam
56. Rössler OE, Wegmann K (1978) Chaos in the Zhabotinskii reaction. Nature 271:89–90
57. Rossoreanu C, Georgescu A, Giurgteanu N (2000) The FitzHugh–Nagumo model – bifurcation and dynamics. Kluwer, Dordrecht
58. Sari T (1996) Nonstandard perturbation theory of differential equations, at symposium on nonstandard analysis and its applications. ICMS, Edinburgh

59. Seliger P, Young SC, Tsimring LS (2002) Plasticity and learning in a network of coupled phase oscillators. Phys Rev E 65:041906
60. Shchepakina E, Sobolev V (2001) Integral manifolds, canards and black swans. Nonlinear Anal Ser A: Theory Meth 44:897–908
61. Smale S (1967) Differentiable dynamical systems. Bull Am Math Soc 73:747–817
62. Strumiłło P, Strzelecki M (2006) Application of coupled neural oscillators for image texture segmentation and modeling of biological rhythms. Int J Appl Math Comput Sci 16:513–523
63. Szmolyan P, Wechselberger M (2001) Canards in R^3. J Differential Equ 177:419–453
64. Szmolyan P, Wechselberger M (2004) Relaxation oscillations in R^3. J Differential Equ 200:69–104
65. Traub RD, Whittington MA, Stanford IM, Jefferys JGR (1991) A mechanism for generation of long-range synchronous fast oscillators in the cortex. Nature 383:621–624
66. Van der Pol B (1926) On relaxation oscillations. Phil Mag 2:978–992
67. Van der Pol B (1940) Biological rhythms considered as relaxation oscillations. Acta Med Scand Suppl 108:76–87
68. Van der Pol B, Van der Mark J (1927) Frequency demultiplication. Nature 120:363–364
69. Varian HR (1979) Catastrophe theory and the business cycle. Econ Inq 17:14–28
70. Vasconcelos GL (1996) First-order phase transition in a model for earthquakes. Phys Rev Lett 76:4865–4868
71. Vatta F (1979) On the stick-slip phenomenon. Mech Res Commun 6:203–208
72. Verhulst F (2005) Quenching of self-excited vibrations. J Engin Math 53:349–358
73. Verhulst F (2007) Periodic solutions and slow manifolds. Int J Bifurc Chaos 17:2533–2540
74. Verhulst F, Abadi (2005) Autoparametric resonance of relaxation oscillations. Z Angew Math Mech 85:122–131
75. Verhulst F, Bakri T (2007) The dynamics of slow manifolds. J Indon Math Soc 13:1–10
76. Wang DL (1999) Relaxation oscillators and networks. In: Webster JG (ed) Wiley encyclopedia of electrical and electronics engineering, vol 18. Wiley, Malden, pp 396–405
77. Wang DL, Terman D (1995) Locally excitatory globally inhibitory oscillator networks. IEEE Trans Neural Netw 6:283–286
78. Winfree AT (2000) The geometry of biological time, 2nd edn. Springer, New York
79. Womelsdorf T, Schoffelen JM, Oosterveld R, Singer W, Desimore R, Engel AK, Fries P (2007) Modulation of neural interactions through neural synchronization. Science 316:1609–1612
80. Zvonkin AK, Shubin MA (1984) Nonstandard analysis and singular perturbations of ordinary differential equations. Russ Math Surveys 39:69–131

Books and Reviews

Beuter A, Glass L, Mackey MC, Titcombe MS (eds) (2003) Nonlinear dynamics in physiology and medicine. Springer, New York
Guckenheimer J, Holmes P (1983) Nonlinear oscillations, dynamical systems, and bifurcations of vector fields. Springer, New York
Verhulst F (2005) Methods and applications of singular perturbations: Boundary layers and multiple timescale dynamics. Springer, New York

Repeated Games with Complete Information

Olivier Gossner[1], Tristan Tomala[2]
[1] PSE, UMR CNRS-EHESS-ENPC-ENS 8545, Northwestern University, Paris, France
[2] Economics and Finance Department, HEC Paris, Paris, France

Article Outline

Glossary
Definition of the Subject
Introduction
Games with Observable Actions
Games with Non-observable Actions
Acknowledgments
Bibliography

Glossary

Repeated game A model of repeated interaction between agents.

Behavioral strategy A decision rule that prescribes a randomized choice of actions for every possible history.

Nash equilibrium A strategy profile from which no unilateral deviation is profitable.

Sequential equilibrium A strategy profile and Bayesian beliefs on past histories such that after every history, every agent is acting optimally given his beliefs.

Monitoring structure A description of player's observation of each other's strategic choices. It specifies, for every profile of actions, the probability distribution over the profiles of individual signals received by the agents.

Definition of the Subject

Repeated interactions arise in several domains such as Economics, Computer Science, and Biology.

The theory of repeated games models situations in which a group of agents engage in a strategic interaction over and over. The data of the strategic interaction is fixed over time and is known by all the players. This is in contrast with stochastic games, for which the data of the strategic interaction is controlled by player's choices, and repeated games with incomplete information, where the stage game is not common knowledge among players (the reader is referred to the corresponding chapters of this Encyclopedia). Early studies of repeated games include Luce and Raiffa [48] and Aumann [4]. In the context of production games, Friedman [25] shows that, while the compet-

itive outcome is the only one compatible with individual profit maximization under a static interaction, collusion is sustainable at an equilibrium when the interaction is repeated.

Generally, repeated games provide a framework in which individual utility maximization by selfish agents is compatible with welfare maximization (common good), while this is known to fail for many classes of static interactions.

Introduction

The discussion of an example shows the importance of repeated games and introduces the questions studied.

Consider the following game referred to as the Prisoner's Dilemma:

	C	D
C	4, 4	5, 5
D	5, 0	1, 1

The Prisoner's Dilemma

Player 1 chooses the row, player 2 chooses the column, and the pair of numbers in the corresponding cell are the payoffs to players 1 and 2 respectively.

In a one-shot interaction, the only outcome consistent with game theory predictions is (D, D). In fact, each player is better off playing D whatever the other player does.

On the other hand, if players engage in a repeated Prisoner's Dilemma, if they value sufficiently future payoffs compared to present ones, and if past actions are observable, then (C, C) is a sustainable outcome. Indeed, if each player plays C as long as the other one has always done so in the past, and plays D otherwise, both player have an incentive to always play C, since the short term gain that can be obtained by playing D is more than offset by the future losses entailed by the opponent playing D at all future stages.

Hence, a game theoretical analysis predicts significantly different outcomes from a repeated game than from static interaction. In particular, in the Prisoner's Dilemma, the cooperative outcome (C, C) can be sustained in the repeated game, while only the non-cooperative outcome (D, D) can be sustained in one-shot interactions.

In general, what are the equilibrium payoffs of a repeated game and how can they be computed from the data of the static game? Is there a significant difference between games repeated a finite number of times and infinitely repeated ones? What is the role played by the degree of impatience of players? Do the conclusions obtained for the Prisoner's Dilemma game and for other games rely crucially on the assumption that each player perfectly observes other player's past choices, or would imperfect observation be sufficient? The theory of repeated games aims at answering these questions, and many more.

Games with Observable Actions

This section focuses on repeated games with perfect monitoring in which, after every period of the repeated game, all strategic choices of all the players are publicly revealed.

Data of the Game, Strategies, Payoffs

Data of the Stage Game There is a finite set I of players. A stage game is repeated over and over. Each player i's action set in this stage game is denoted A_i, and $S_i = \Delta(A_i)$ is the set of player i's mixed actions (for any finite set X, $\Delta(X)$ denotes the set of probabilities over X). Every degenerate lottery in S_i (which puts probability 1 to one particular action in A_i) is associated to the corresponding element in A_i. A choice of action for every player i determines an outcome $a \in \prod_i A_i$. The payoff function of the stage game is $g\colon A \to \mathbb{R}^I$. Payoffs are naturally associated to profiles of mixed actions $s \in S = \prod_i S_i$ using the expectation: $g(s) = \mathbf{E}_s g(a)$.

Repeated Game After every repetition of the stage game, the action profile previously chosen by the players is publicly revealed. After the t first repetitions of the game, a player's information consists of the publicly known history at stage t, which is an element of $H_t = A^t$ ($H_0 = \{\emptyset\}$ by convention). A strategy in the repeated game specifies the choice of a mixed action at every stage, as a function of the past observed history. More specifically, a *behavioral* strategy for player i is of the form $\sigma_i\colon \cup_t H_t \to S_i$. When all the strategy choices belong to A_i ($\sigma_i\colon \cup_t H_t \to A_i$), σ_i is called a *pure* strategy.

Other Strategy Specifications A behavioral strategy allows the player to randomize his action depending on past history. If, at the start of the repeated game, the player was to randomize over the set of behavioral strategies, the result would be equivalent to a particular behavioral strategy choice. This result is a consequence of Kuhn's theorem [5,41]. Furthermore, behavioral strategies are also equivalent to randomizations over the set of pure strategies.

Induced Plays Every choice of pure strategies $\sigma = (\sigma_i)_i$ by all the players induces a play $h = (a_1, a_2, \dots) \in A^\infty$ in the repeated game, defined inductively by $a_1 = (\sigma_{i,0}(\emptyset))$ and $a_t = (\sigma_{i,t-1}(a_1, \dots, a_{t-1}))$. A profile of behavioral strategies σ defines a probability distribution P_σ over plays.

Preferences To complete the definition of the repeated game, it remains to define player's preferences over plays. The literature commonly distinguishes infinitely repeated games with or without discounting, and finitely repeated games.

In infinitely repeated games with no discounting, the players care about their long-run stream of stage payoffs. In particular, the payoff in the repeated game associated to a play $h = (a_1, a_2, \dots) \in A^\infty$ coincides with the limit of the Cesaro means of stage payoffs when this limit exists. When this limit does not exist, the most common evaluation of the stream of payoffs is defined through a Banach limit of the Cesaro means (a Banach limit is a linear form on the set of bounded sequences that lies always between the liminf and the limsup).

In infinitely repeated games with discounting, a discount factor $0 < \delta < 1$ characterizes the player's degree of impatience. A payoff of 1 at stage $t+1$ is equivalent to a payoff of δ at stage t. Player i's payoff in the repeated game for the play $h = (a_1, a_2, \dots) \in A^\infty$ is the normalized sum of discounted payoffs: $(1-\delta)\sum_{t\geq 1} \delta^{t-1} g_i(a_t)$.

In finitely repeated games, the game ends after some stage T. Payoffs induced by the play after stage T are irrelevant (and a strategy needs not specify choices after stage T). The payoff for a player is the average of the stage payoffs during stages 1 up to T: $\frac{1}{T}\sum_{t=1}^{T} g_i(a_t)$.

Equilibrium Notions What plays can be expected to be observed in repeated interactions of players who observe each other's choices? Non-cooperative Game Theory focuses mainly on the idea of stable convention, i. e. of strategy profiles from which no player has incentives to deviate, knowing the strategies adopted by the other players.

A strategy profile forms a Nash Equilibrium (Nash [55]) when no player can improve his payoff by choosing an alternative strategy, as long as other players follow the prescribed strategies.

In some cases, the observation of past play may not be consistent with the prescribed strategies. When, for every possible history, each player's strategy maximizes the continuation stream of payoffs, assuming that other players abide with their prescribed strategies at all future stages, the strategy profile forms a subgame perfect equilibrium (Selten [66]).

Perfect equilibrium is a more robust and often considered a more satisfactory solution concept than Nash equilibrium. The construction of perfect equilibria is in general also more demanding than the construction of Nash equilibria.

The main objective of the theory of repeated games is to characterize the set of payoff vectors that can be sustained by some Nash or perfect equilibrium of the repeated game.

Necessary Conditions on Equilibrium Payoffs

Some properties are common to all equilibrium payoffs. First, under the common assumption that all players evaluate the payoff associated to a play in the same way, the resulting payoff vector in the repeated game is a convex combination of stage payoffs. That is, the payoff vector in the repeated game is an element of the convex closure of $g(A)$, called the set of *feasible payoffs* and denoted F.

A notable exception is the work of Lehrer and Pauzner [47] who study repeated games where players have heterogeneous time preferences. The payoff vector resulting from a play does not necessarily belong to F if players have different evaluations of payoff streams. For instance, in a repetition of the Prisoner's Dilemma, if player 1 cares only about the payoff in stage 1 and player 2 cares only about the payoff in stage 2, it is possible for both players to obtain a payoff of 4 in the repeated game.

Now consider a strategy profile σ, and let τ_i be a strategy of player i that plays after every history $(a_1, \dots, a_t)$ a best response to the profile of mixed actions chosen by the other players in the next stage. At any stage of the repeated game, the expected payoff for player i using τ_i is no less than

$$v_i = \min_{s_{-i}\in S_{-i}} \max_{a_i \in A_i} g_i(s_{-i}, a_i) \tag{1}$$

where $s_{-i} = (s_j)_{j\neq i}$ (we use similar notations throughout the paper: for a family of sets $(E_i)_{i\in I}$, e_{-i} denotes an element of $E_{-i} = \Pi_{j\neq i} E_j$ and a profile $e \in \Pi_j E_j$ is denoted $e = (e_i, e_{-i})$ when the ith component is stressed).

The payoff v_i is referred to as player i's min max payoff. A payoff vector that provides each player i with at least [resp. strictly more than] v_i is called *individually rational* [resp. strictly individually rational], and IR [resp. IR^*] denotes the set of such payoff vectors. Since for any strategy profile, there exists a strategy of player i that yields a payoff no less than v_i, all equilibrium payoffs have to be individually rational.

Also note that players $j \neq i$ collectively have a strategy profile in the repeated game that forces player i's payoff down to v_i: they play repeatedly a mixed strategy profile that achieves the minimum in the definition of v_i. Such a strategy profile in the one-shot game is referred to as punishing strategy, or min max strategy against player i.

For the Prisoner's Dilemma game, F is the convex hull of $(1,1)$, $(5,0)$, $(0,5)$ and $(4,4)$. Both player's min max levels are equal to 1. Figure 1 illustrates the set of feasible and individually rational payoff vectors (hatched area):

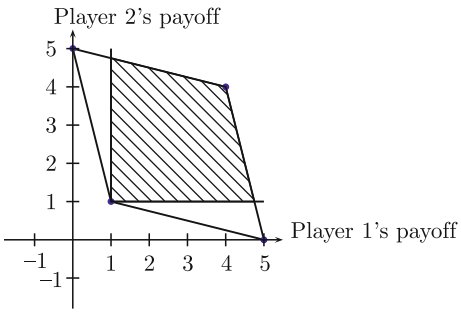

Repeated Games with Complete Information, Figure 1
***F* and *IR* for the Prisoner's Dilemma**

The set of feasible and individually rational payoffs can be directly computed from the stage game data.

Infinitely Patient Players

The following result has been part of the folklore of Game Theory at least since the mid 1960s. Its authorship is obscure (see the introduction of Aumann [7]). For this reason, it is commonly referred to as the "Folk Theorem". By extension, characterization of sets of equilibrium payoffs in repeated games are also referred to as "Folk Theorems".

Theorem 1 *The set of equilibrium payoffs of the repeated game with no discounting coincides with the set of feasible and individually rational payoffs.*

Aumann and Shapley [8,9] and Rubinstein [62,64] show that restricting attention to perfect equilibria does not narrow down the set of equilibrium payoffs. They prove that:

Theorem 2 *The set of perfect equilibrium payoffs of the repeated game with no discounting coincides with the set of feasible and individually rational payoffs.*

We outline a proof of Theorem 2. It is established that any equilibrium payoff is in $F \cap IR$. We need only to prove that every element of $F \cap IR$ is a subgame perfect equilibrium payoff. Let $x \in F \cap IR$, and let $h = a_1, \ldots, a_t, \ldots$ be a play inducing x. Consider the strategies that play a_t in stage t; if player i does not respect this prescription at stage t_0, the other players punish player i for t_0 stages by repeatedly playing the min max strategy profile against player i. After the punishment phase is over, players revert to the play of h, hence playing $a_{2t_0+1}, \ldots$.

Now we explain why these strategies form a subgame perfect equilibrium. Consider a strategy of player i starting after any history. The induced play by this strategy for player i and by other player's prescribed strategies is, up to a subset of stages of null density, defined by the sequence h with interweaved periods of punishment for player i. Hence the induced long-run payoff for player i is a convex combination of his punishment payoff and of the payoff induced by h. The result follows since the payoff for player i induced by h is no worse than the punishment payoff.

Impatient Players

The strategies constructed in the proof of the Folk Theorem for repeated games with infinitely patient players (Theorem 1) do not necessarily constitute a subgame perfect equilibrium if players are impatient. Indeed, during a punishment phase, the punishing players may be receiving low stage payoffs, and these stage payoffs matter in the evaluation of their stream of payoffs. When constructing subgame perfect equilibria of discounted games, one must make sure that after a deviation of player i, players $j \neq i$ have incentives to implement player i's punishment.

Nash Reversion Friedman [25] shows that every feasible payoff that Pareto dominates a Nash equilibrium payoff of the static game is a subgame perfect equilibrium payoff of the repeated game provided that players are patient enough. In Friedman's proof, punishments take the simple form of reversion to the repeated play of the static Nash equilibrium forever. In the Prisoner's Dilemma, (D, D) is the only static Nash equilibrium payoff, and thus $(4, 4)$ is a subgame perfect Nash equilibrium payoff of the repeated game if players are patient enough. Note however that in some games, the set of payoffs that Pareto dominate some equilibrium payoff may be empty. Also, Friedman's result constitutes a partial Folk Theorem only in that it does not characterize the full set of equilibrium payoffs.

The Recursive Structure Repeated games with discounting possess a structure similar to dynamic programming problems. At any stage in time, players choose actions that maximize the sum of the current payoff and the payoff at the subsequent stages. When strategies form a subgame perfect equilibrium, the payoff vector at subsequent stages must be an equilibrium payoff, and players must have incentives to follow the prescribed strategies at the current stage. This implies that subgame perfect equilibrium payoffs have a recursive structure, first studied by Abreu [1]. Subsection "A Recursive Structure" presents the recursive structure in more details for the more general model of games with public monitoring.

The Folk Theorem for Discounted Games Relying on Abreu's recursive results, Fudenberg and Maskin [20]

prove the following Folk Theorem for subgame perfect equilibria with discounting:

Theorem 3 *If the number of players is 2 or if the set feasible payoff vectors has non-empty interior, then any payoff vector that is feasible and strictly individually rational is a subgame perfect equilibrium of the discounted repeated game, provided that players are sufficiently patient.*

Forges, Mertens and Neyman [24] provide an example for which a payoff which is individually rational but not strictly individually rational is not an equilibrium payoff of the discounted game.

Abreu, Dutta and Smith [2] show that the non-empty interior condition of the theorem can be replaced by a weaker condition of "non equivalent utilities": no pair of players have the same preferences over outcomes. Wen [73] and Fudenberg, Levine and Takahashi [22] show that a Folk Theorem still holds when the condition of non equivalent utilities fails if one replaces the minmax level defining individually rational payoffs by some "effective minmax" payoffs.

An alternative representation of impatience to discounted payoffs in infinitely repeated games is the overtaking criterion, introduced by Rubinstein [63]: the play $(a_1, a_2, \dots)$ is strictly preferred by player i to the play $(a'_1, a'_2, \dots)$ if the inferior limit of the difference of the corresponding streams of payoffs is positive, i. e. if $\liminf_T \sum_{t=1}^{T} g_i(a_t) - g_i(a'_t) > 0$. Rubinstein [63] proves a Folk Theorem with the overtaking criterion.

Finitely Repeated Games

Strikingly, equilibrium payoffs in finitely repeated games and in infinitely repeated games can be drastically different. This effect is best exemplified in repetitions of the Prisoner's Dilemma.

The Prisoner's Dilemma Recall that in an infinitely repeated Prisoner's Dilemma, cooperation at all stages is achieved at a subgame perfect equilibrium if players are patient enough. By contrast, at every Nash equilibrium of any finite repetition of the Prisoner's Dilemma, both players play D at every stage with probability 1.

Now we present a short proof of this result. Consider any Nash equilibrium of the Prisoner's Dilemma repeated T times. Let $a_1, \dots, a_T$ be a sequence of action profiles played with positive probability at the Nash equilibrium. Since each player can play D at the last stage of the repetition, and D is a dominating action, $a_T = (D, D)$. We now prove by induction on τ that for any such τ, $a_{T-\tau}, \dots, a_T = (D, D), \dots, (D, D)$. Assume the induction hypothesis valid for $\tau - 1$. Consider a strategy for player i that follows the equilibrium strategy up to stage $T - \tau - 1$, then plays D from stage $T - \tau$ on. This strategy obtains the same payoff as the equilibrium strategy at stages $1, \dots, T - \tau - 1$, and as least as much as the equilibrium strategy at stages $T - \tau + 1, \dots, T - \tau$. Hence, this strategy cannot obtain more than the equilibrium strategy at stage $T - \tau$, and therefore the equilibrium strategy plays D at stage $T - \tau$ with probability 1 as well.

Sorin [68] proves the more general result:

Theorem 4 *Assume that in every Nash equilibrium of G, all players are receiving their individually rational levels. Then, at every Nash equilibrium of any finitely repeated version of G, all players are receiving their individually rational levels.*

The proof of Theorem 4 relies on a backwards induction type of argument, but it is striking that the result applies for *all* Nash equilibria and not only for *subgame perfect* Nash equilibria. This result shows that, unless some additional assumptions are made on the one-shot game, a Folk Theorem cannot obtain for finitely repeated games.

Games with Unique Nash Payoff Using a proof by backwards induction, Benoît and Krishna [10] obtain the following result.

Theorem 5 *Assume that G admits x as unique Nash equilibrium payoff. Then every finite repetition of G admits x as unique subgame perfect equilibrium payoff.*

Theorems 4 and 5 rely on the assumption that the last stage of repetition, T, is common knowledge between players. Neyman [56] shows that a Folk Theorem obtains for the finitely repeated Prisoner's Dilemma (and for other games) if there is lack of common knowledge on the last stage of repetition.

Folk Theorems for Finitely Repeated Games A Folk Theorem can be obtained when there are two Nash equilibrium payoffs for each player. The following result is due to Benoît and Krishna [10] and Gossner [27].

Theorem 6 *Assume that each player has two distinct Nash equilibrium payoffs in G and that the set of feasible payoffs has non-empty interior. Then, the set of subgame perfect equilibrium payoffs of the T times repetition of G converges to the set of feasible and individually rational payoffs as T goes to infinity.*

Hence, with at least two equilibrium payoffs per player, the sets of equilibrium payoffs of finitely repeated games and infinitely repeated games are asymptotically the same.

The condition that each player has two distinct Nash equilibrium payoffs in the stage game can be weakened, see Smith [67]. Assume for simplicity that *one* player has two distinct Nash payoffs. By playing one of the two Nash equilibria in the last stages of the repeated game, it is possible to provide incentives for this player to play actions that are not part of Nash equilibria of the one-shot game in previous stages. If this construction leads to perfect equilibria in which a player $j \neq i$ has distinct payoffs, we can now provide incentives for both players i and j. If successive iterations of this procedure yield distinct subgame perfect equilibrium payoffs for all players, a Folk Theorem applies.

Games with Non-observable Actions

For infinitely repeated games with perfect monitoring, a complete and simple characterization of the set of equilibrium payoffs is obtained: feasible and individually rational payoff vectors. In particular, cooperation can be sustained at equilibrium. How equilibrium payoffs of the repeated game depend on the quality of player's monitoring of each other's actions is the subject of a very active area of research.

Repeated games with imperfect monitoring, in which players observe imperfectly other player's action choices, were first motivated by economic applications. In Stigler [69], two firms are repeatedly engaged in price competition over market shares. Each firm observes its own sales, but not the price set by the rival. While it is in the best interest for both firms to set a collusive price, each firm has incentives to secretly undercut the rival's price. Upon observing plunging sales, should a firm deduce that the rival firm is undercutting prices, and retaliate by setting lower prices, or should lower sales be interpreted as a result of an exogenous shock on market demand? Whether collusive behavior is sustainable or not at equilibrium is one of the motivating questions in the theory of repeated games with imperfect monitoring.

It is interesting to compare repeated games with imperfect monitoring with their perfect monitoring counterparts.

The structure of equilibria used to prove the Folk Theorem with perfect monitoring and no discounting is rather simple: if a player deviates from the prescribed strategies, the deviation is detected and the deviating player is identified, all other players can then punish the deviator. With imperfect monitoring, not all deviations are detectable, and when a deviation is detected, deviators are not necessarily identifiable. The notions of detection and identification allow fairly general Folk Theorems for undiscounted games. We present these results in Subsect. "Detection and Identification".

With discounting, repeated games with perfect monitoring possess a recursive structure that facilitates their study. Recursive methods can also be successfully applied to discounted games with public monitoring. We review the major results of this branch of the literature in Subsect. "Public Equilibria".

Almost-perfect monitoring is the natural framework to study the effect of small departures from the perfect or public monitoring assumptions. We review this literature in Subsect. "Almost Perfect Monitoring".

Little is known about general discounted games with imperfect private monitoring. We present the main known results in Subsect. "General Stochastic Signals".

With perfect monitoring, the worst equilibrium payoff for a player is given by the min max of the one-shot game, where punishing (minimizing) players choose an independent profile of mixed strategies. With imperfect monitoring, correlation past signals for the punishing players may lead to more efficient punishments. We present results on punishment levels in Subsect. "Punishment Levels".

Model

In this section we define repeated games with imperfect monitoring, and describe several classes of monitoring structures of particular interest.

Data of the Game Recall that the one-shot strategic interaction is described by a finite set I of players, a finite action set A_i for each player i, and a payoff function $g\colon A \to \mathbb{R}^I$. Player's observation of each other's actions is described by a *monitoring structure* given by a finite set of signals Y_i for each player i and by a transition probability $Q\colon A \to \Delta(Y)$ (with $A = \prod_{i\in I} A_i$ and $Y = \prod_{i\in I} Y_i$). When the action profile chosen is $a = (a_i)_{i\in I}$, a profile of signals $y = (y_i)_{i\in I}$ is drawn with probability $Q(y|a)$ and y_i is observed by player i.

Perfect Monitoring Perfect monitoring is the particular case in which each player observes the action profile chosen: for each player i, $Y_i = A$ and $Q((y_i)_{i\in I}|a) = \mathbb{1}_{\{\forall i,\, y_i = a\}}$.

Almost Perfect Monitoring The monitoring structure is ε-perfect (see Mailath and Morris [52]) when each player can identify the other player's action with a probability of error less than ε. This is the case if there exist functions $f_i\colon A_i \times Y_i \to A_{-i}$ for all i such that, for all $a \in A$:

$$Q(\forall i, f_i(a_i, y_i) = a_{-i}|a) \geq 1 - \varepsilon\,.$$

Almost-perfect monitoring refers to ε-perfect monitoring for small values of ε.

Canonical Structure The monitoring structure is *canonical* when each player's observation corresponds to an action profile of the opponents, that is, when $Y_i = A_{-i}$.

Public and Almost Public Signals Signals are *public* when all the players observe the same signal, i. e., $Q(\forall i, j, y_i = y_j|a) = 1$, for every action profile a. For instance, in Green and Porter [33], firms compete over quantities, and the public signal is the realization of the price. Firms can then make inferences on rival's quantities based on their own quantity and market price.

The case in which $Q(\forall i, j, y_i = y_j|a)$ is close to 1 for every a is referred to as *almost public monitoring* (see Mailath and Morris [52]).

Private signals refer to the case when these signals are not public.

Deterministic Signals Signals are *deterministic* when the signal profile is uniquely determined by the action profile. When a is played, the signal profile y is given by $y = f(a)$, where f is called the *signaling function*.

Observable Payoffs Payoffs are *observable* when each player i can deduce his payoff from his action and his signal. This is the case if there exists a mapping $\varphi\colon A_i \times Y_i \to \mathbb{R}$ such that for every action profile a, $Q(\forall i, g_i(a) = \varphi(a_i, y_i)|a) = 1$.

The Repeated Game The game is played repeatedly and after every stage t, the profile of signals y_t received by the players is drawn according to the distribution $Q(y_t|a_t)$, where a_t is the profile of action chosen at stage t. A player's information consists of his past actions and signals. We let $H_{i,t} = (A_i \times Y_i)^t$ be the set of *player i's histories* of length t. A strategy for player i now consists of a mapping $\sigma_i\colon \cup_{t\geq 0} H_{i,t} \to S_i$. The set of complete histories of the game after t stages is $H_t = (A \times Y)^t$, it describes chosen actions and received signals for all the players at all past stages. A strategy profile $\sigma = (\sigma_i)_{i\in I}$ induces a probability distribution P_σ on the set of plays $H_\infty = (A \times Y)^\infty$.

Equilibrium Notions

Nash Equilibria Player's preferences over game plays are defined according to the same criteria as for perfect monitoring. We focus on infinitely repeated games, both discounted and undiscounted. A choice of players' preferences defines a set of Nash equilibrium payoffs in the repeated game.

Sequential Equilibria The most commonly used refinement of Nash equilibrium for repeated games with imperfect monitoring is the *sequential equilibrium* concept (Kreps and Wilson, [42]), which we recall here.

A *belief assessment* is a sequence $\mu = (\mu_{i,t})_{t\geq 1, i\in I}$ with $\mu_{i,t}\colon H_{i,t} \to \Delta(H_t)$, i. e., given the private history h_i of player i, $\mu_{i,t}(h_i)$ is the probability distribution representing the belief that player i holds on the full history.

A sequential equilibrium of the repeated game is a pair (σ, μ) where σ is a strategy profile and μ is a belief assessment such that: 1) for each player i and every history h_i, σ_i is a best reply in the continuation game, given the strategies of the other players and the belief that player i holds regarding the past; 2) the beliefs must be consistent in the sense that (σ, μ) is the limit of a sequence (σ^n, μ^n) where for every n, σ^n is a completely mixed strategy (it assigns positive probability to every action after every history) and μ^n is the unique belief derived from Bayes' law under P_{σ^n}.

Sequential equilibria are defined both on the discounted game and the undiscounted versions of the repeated game.

For undiscounted games, the set of Nash equilibrium payoffs and sequential equilibrium payoffs coincide. The two notions also coincide for discounted games when the monitoring has full support (i. e. under every action profile, all signal profiles have positive probability). The results presented in this survey all hold for sequential equilibria, both for discounted and undiscounted games.

Extensions of the Repeated Game When players receive correlated inputs or may communicate between stages of the repeated game, the relevant concepts are correlated and communication equilibria.

Correlated Equilibria A correlated equilibrium (Aumann [6]) of the repeated game is an equilibrium of an extended game in which: at a preliminary stage, a mediator chooses a profile of correlated random inputs and informs each player of his own input; then the repeated game is played. A characterization of the set of correlated equilibrium payoffs for two-player games is obtained by Lehrer [45].

Correlation arises endogenously in repeated games with imperfect monitoring, as the signals received by the players can serve as correlated inputs that influence player's continuation strategies. This phenomenon is called *internal correlation*, and was studied by Lehrer, [44], Gossner and Tomala, [29,30].

Communication Equilibria An (extensive form) communication equilibrium (Myerson [54], Forges [23]) of

a repeated game is an equilibrium of an extension of the repeated game in which after every stage, players send messages to a mediator, and the mediator sends back private outputs to the players. Characterizations of the set of communication equilibrium payoffs are obtained under weak conditions on the monitoring structure, see e. g. Kandori and Matsushima [38], Compte [15], and Renault and Tomala [61].

Detection and Identification

Equivalent Actions A player's deviation is detectable when it induces a different distribution of signals for other players. When two actions induce the same distribution of actions for other players, they are called equivalent (Lehrer [43,44,45,46]):

Definition 1 Two actions a_i and b_i of player i are equivalent, and we note $a_i \sim b_i$, if they induce the same distribution of other players' signals:

$$Q(y_{-i}|a_i, a_{-i}) = Q(y_{-i}|b_i, a_{-i}) , \quad \forall a_{-i} .$$

Example 1 Consider the two-player repeated Prisoner's Dilemma where player 2 receives no information about the actions of player 1 (e. g. Y_2 is a singleton). The two actions of player 1 are thus equivalent. The actions of player 2 are independent of the actions of player 1: player 1 has no impact on the behavior of player 2. Player 2 has no power to threat player 1 and in any equilibrium, player 1 defects at every stage. Player 2 also defects at every stage: since player 1 always defects, he also loses his threatening power. The only equilibrium payoff in this repeated game is thus $(1, 1)$.

Example 1 suggests that between two equivalent actions, a player chooses at equilibrium the one that yields the highest stage payoff. This is indeed the case when the information received by a player does not depend on his own action. Lehrer [43] studies particular monitoring structures satisfying this requirement. Recall from Lehrer [43] the definition of *semi-standard* monitoring structures: each action set A_i is endowed with a partition $\bar{A}_i$, when player i plays a_i, the corresponding partition cell $\bar{a}_i$ is publicly announced. In the semi-standard case, two actions are equivalent if and only if they belong to the same cell: $a_i \sim b_i \iff \bar{a}_i = \bar{b}_i$ and the information received by a player on other player's action does not depend on his own action.

If player i deviates from a_i to b_i, the deviation is undetected if and only if $a_i \sim b_i$. Otherwise it is detected by all other players. A profile of mixed actions is called *immune to undetectable deviations* if no player can profit by a unilateral deviation that maintains the same distribution of other players' signals. The following result, due to Lehrer [43], characterizes equilibrium payoffs for undiscounted games with semi-standard signals:

Theorem 7 *In an undiscounted repeated game with semi-standard signals, the equilibrium payoffs are the individually rational payoffs that belong to the convex hull of payoffs generated by mixed action profiles that are immune to undetectable deviations.*

More Informative Actions When the information of player i depends on his own action, some deviations may be detected in the course of the repeated game even though they are undetectable in the stage game.

Example 2 Consider the following modification of the Prisoner's Dilemma. The action set of player 1 is $A_1 = \{C_1, D_1\} \times \{C_2, D_2\}$ and the action set of player 2 is $\{C_2, D_2\}$. An action for player 1 is thus a pair $a_1 = (\tilde{a}_1, \tilde{a}_2)$. When the action profile $(\tilde{a}_1, \tilde{a}_2, a_2)$ is played, the payoff to player i is $g_i(\tilde{a}_1, a_2)$. We can interpret the component $\tilde{a}_1$ as a *real* action (it impacts payoffs) and the component $\tilde{a}_2$ as a *message* sent to player 2 (it does not impact payoffs). The monitoring structure is as follows:

- player 2 only observes the message component $\tilde{a}_2$ of the action of player 1 and,
- player 1 perfectly observes the action of player 2 if he chooses the cooperative real action ($\tilde{a}_1 = C_1$), and gets no information on player 2's action if he defects ($\tilde{a}_1 = D_1$).

Note that the actions (C_1, C_2) and (D_1, C_2) of player 1 are equivalent, and so are the actions (C_1, D_2) and (D_1, D_2). However, it is possible to construct an equilibrium that implements the cooperative payoff along the following lines:

i) Using his message component, player 1 reports at every stage $t > 1$ the previous action of player 2. Player 1 is punished in case of a non matching report.
ii) Player 2 randomizes between both actions, so that player 1 needs to play the cooperative action in order to report player 2's action accurately. The weight on the defective action of player 2 goes to 0 as t goes to infinity to ensure efficiency.

Player 2 has incentives to play C_2 most of the time, since player 1 can statistically detect if player 2 uses the action D_2 more frequently than prescribed. Player 1 also has incentives to play the real action C_1, as this is the only way to

observe player 2's action, which needs to be reported later on.

The key point in the example above is that the two real actions C_1 and D_1 of player 1 are equivalent but D_1 is *less informative* than C_1 for player 1. For general monitoring structures an action a_i is *more informative* than an action b_i if: *whenever player i plays a_ii, he can reconstitute the signal he would have observed, had he played b_i*. The precise definition of the *more informative* relation relies on Blackwell's ordering of stochastic experiments [11]:

Definition 2 The action a_i of player i is more informative than the action b_i if there exists a transition probability $f\colon Y_i \to \Delta(Y_i)$ such that for every a_{-i} and every profile of signals y,

$$\sum_{y_i} f(y_i'|y_i)Q(y_i, y_{-i}|a_i, a_{-i}) = Q(y_i', y_{-i}|b_i, a_{-i}) .$$

We denote $a_i \succeq b_i$ if $a_i \sim b_i$ and a_i is more informative than b_i.

Assume that prescribed strategies require player i to play b_i at stage t, and let $a_i \succeq b_i$. Consider the following deviation from player i: play a_i at stage t, and reconstruct a signal at stage t that could have arisen from the play of b_i. In all subsequent stages, play as if no deviation took place at stage t, and as if the reconstructed signal had been observed at stage t. Not only such a deviation would be undetectable at stage t, since $a_i \sim b_i$, but it would also be undetectable at all subsequent stages, as it would induce the same probability distribution over plays as under the prescribed strategy. This argument shows that, if an equilibrium strategy specifies that player i plays a_i, there is no $b_i \succeq a_i$ that yields a higher expected stage payoff than a_i.

Definition 3 A distribution of actions profiles $p \in \Delta(A)$ is *immune to undetectable deviations* if for each player i and pair of actions a_i, b_i such that $b_i \succeq a_i$:

$$\sum_{a_{-i}} p(a_i, a_{-i})g_i(a_i, a_{-i}) \geq \sum_{a_{-i}} p(a_i, a_{-i})g_i(b_i, a_{-i})$$

If p is immune to undetectable deviations, and if player i is supposed to play a_i, any alternative action b_i that yields a greater expected payoff can not be such that $b_i \succeq a_i$.

The following proposition gives a necessary condition on equilibrium payoffs that holds both in the discounted and in the undiscounted cases.

Proposition 1 *Every equilibrium payoff of the repeated game is induced by a distribution that is immune to undetectable deviations.*

The condition of Proposition 1 is tight for some specific classes of games, all of them assuming two players and no discounting.

Following Lehrer [45], signals are *non-trivial* if, for each player i, there exist an action a_i for player i and two actions a_j, b_j for i's opponent such that the signal for player i is different under (a_i, a_j) and (a_i, b_j). Lehrer [45] proves:

Theorem 8 *The set of correlated equilibrium payoffs of the undiscounted game with deterministic and non-trivial signals is the set of individually rational payoffs induced by distributions that are immune to undetectable deviations.*

Lehrer [46] assumes that payoffs are observable, and obtains the following result:

Theorem 9 *In a two-player repeated game with no discounting, non-trivial signals and observable payoffs, the set of equilibrium payoffs is the set of individually rational payoffs induced by distributions that are immune to undetectable deviations.*

Finally, Lehrer [44] shows that in some cases, one may dispense with the correlation device of Theorem 8, as all necessary correlation can be generated endogenously through the signals of the repeated game:

Proposition 2 *In two-player games with non-trivial signals such that either the action profile is publicly announced or a blank signal is publicly announced, the set of equilibrium payoffs coincides with the set of correlated equilibrium payoffs.*

Identification of Deviators A deviation is identifiable when every player can infer the identity of the deviating player from his observations. For instance, in a game with public signals, if separate deviations from players i and j induce the same distribution of public signals, these deviations from i or j are not identifiable. In order to be able to punish the deviating player, it is sometimes necessary to know his identity. Detectability and identifiability are two separate issues, as shown by the following example.

Example 3 Consider the following 3-player game where player 1 chooses the row, player 2 chooses the column and player 3 chooses the matrix.

	L	R
T	1, 1, 1	4, 4, 0
B	4, 4, 0	4, 4, 0

W

	L	R
T	0, 3, 0	0, 3, 0
B	0, 3, 0	0, 3, 0

M

	L	R
T	3, 0, 0	3, 0, 0
B	3, 0, 0	3, 0, 0

E

Consider the monitoring structure in which actions are not observable and the payoff vector is publicly announced.

The payoff $(1, 1, 1)$ is feasible and individually rational. The associated action profile (T, L, W) is immune to undetectable deviations since any individual deviation from (T, L, W) changes the payoff.

However, $(1, 1, 1)$ is not an equilibrium payoff. The reason is that, player 3, who has the power to punish either player 1 or player 2, cannot punish both players simultaneously: punishing player 1 rewards player 2 and vice-versa. More precisely, whatever weights player 3 puts on the action M and E, the sum of player 1 and player 2's payoffs is greater than 3. Any equilibrium payoff vector $v = (v_1, v_2, v_3)$ must thus satisfy $v_1 + v_2 \geq 3$. In fact, it is possible to prove that the set of equilibrium payoffs of this repeated game is the set of feasible and individually rational payoffs that satisfy this constraint.

Approachability When the deviating player cannot be identified, it may be necessary to punish a group of suspects altogether. The notion of a payoff that is enforceable under group punishments is captured by the definition of approachable payoffs:

Definition 4 A payoff vector v is *approachable* if there exists a strategy profile σ such that, for every player i and unilateral deviation τ_i of player i, the average payoff of player i under (τ_i, σ_{-i}) is asymptotically less than or equal to v_i.

Blackwell's [12] approachability theorem and its generalization by Kohlberg [40] provide simple geometric characterizations of approachable payoffs. It is straightforward that approachability is a necessary condition on equilibrium payoffs:

Proposition 3 *Every equilibrium payoff of the repeated game is approachable.*

Renault and Tomala [61] show that the conditions of Proposition 1 and 3 are tight for communication equilibria:

Theorem 10 *For every game with imperfect monitoring, the set of communication equilibrium payoffs of the repeated game with no discounting is the set of approachable payoffs induced by distributions which are immune to undetectable deviations.*

Tomala [70] shows that pure strategy equilibrium payoffs of undiscounted repeated games with public signals are also characterized through identifiability and approachability conditions (the approachability definition then uses pure strategies). Tomala [71] provides a similar characterization in mixed strategies for a restricted class of public signals.

Identification Through Endogenous Communication A deviation may be identified in the repeated game even though it cannot be identified in the stage game. In a *network game*, players are located at nodes of a graph, and each player monitors his neighbors' actions. Each player can use his actions as messages that are broadcasted to all the neighbors in the graph. The graph is called *2-connected* if no single node deletion disconnects the graph. Renault and Tomala [60] show that when the graph is 2-connected, there exists a communication protocol among the players that ensures that the identity of any deviating player becomes common knowledge among all players in finite time. In this case, identification takes place through communication over the graph.

Public Equilibria

In a seminal paper, Green and Porter [33] introduce a model in which firms are engaged in a production game and publicly observe market prices, which depend both on quantities produced and on non-observable exogenous market shocks. Can collusion be sustained at equilibrium even if prices convey imperfect information on quantities produced? This motivates the study of *public equilibria* for which sharp characterizations of equilibrium payoffs are obtained.

Signals are public when all sets of signals are identical, i. e. $Y_i = Y_{\text{pub}}$ for each i and $Q(\forall i, j, y_i = y_j | a) = 1$ for every a. A *public history* of length t is a record of t public signals, i. e. an element of $H_{\text{pub},t} = (Y_{\text{pub}})^t$. A strategy σ_i for player i is a *public strategy* if it depends on the public history only: if $h_i = (a_{i,1}, y_1, \ldots, a_{i,t}, y_t)$ and $h'_i = (a'_{i,1}, y'_1, \ldots, a'_{i,t}, y'_t)$ are two histories for player i such that $y_1 = y'_1, \ldots, y_t = y'_t$, then $\sigma_i(h_i) = \sigma_i(h'_i)$.

Definition 5 A *perfect public equilibrium* is a profile of public strategies such that after every public history, each player's continuation strategy is a best reply to the opponents' continuation strategy profile.

The repetition of a Nash equilibrium of the stage game is a perfect public equilibrium, so that perfect public equilibria exist. Every perfect public equilibrium is a sequential equilibrium: any consistent belief assigns probability one to the realized public history and thus correctly forecasts future opponents' choices.

A Recursive Structure A perfect public equilibrium (PPE henceforth) is a profile of public strategies that forms

an equilibrium of the repeated game and such that, after every public history, the continuation strategy profile is also a PPE. The set of PPEs and of induced payoffs therefore possesses a recursive structure, as shown by Abreu, Pearce and Stachetti [3]. The argument is based on a dynamic programming principle. To state the main result, we first introduce some definitions.

Given a mapping $f\colon Y_{\text{pub}} \to \mathbb{R}^I$, $G(\delta, f)$ represents the one-shot game where each player i choose actions in A_i and where payoffs are given by:

$$(1-\delta)g_i(a) + \delta \sum_{y \in Y_{\text{pub}}} Q(y|a) f_i(y) .$$

In $G(\delta, f)$, the stage game is played, and players receive $f(y)$ as an additional payoff if y is the realized public signal. The weights $1-\delta$ and δ are the relative weights of present payoffs versus all future payoffs in the repeated game.

Definition 6 A payoff vector $v \in \mathbb{R}^I$ is decomposable with respect to the set $W \subset \mathbb{R}^I$ if there exists a mapping $f\colon Y_{\text{pub}} \to W$ such that v is a Nash equilibrium payoff of $G(\delta, f)$. $F_\delta(W)$ denotes the set of payoff vectors which are decomposable with respect to W.

Let $E(\delta)$ be the set of perfect public equilibrium payoffs of the repeated game discounted at the rate δ. The following result is due to Abreu et al. [3]:

Theorem 11 *$E(\delta)$ is the largest bounded set W such that $W \subseteq F_\delta(W)$.*

Fudenberg and Levine [19] derive an asymptotic characterization of the set of PPE payoffs when the discount factor goes to 1 as follows. Given a vector $\lambda \in \mathbb{R}^I$, define the *score* in the direction λ as:

$$k(\lambda) = \sup \langle \lambda, v \rangle$$

where the supremum is taken over the set of payoff vectors v that are Nash equilibrium payoffs of $G(\delta, f)$, where f is any mapping such that,

$$\langle \lambda, v \rangle \geq \langle \lambda, f(y) \rangle , \quad \forall y \in Y_{\text{pub}} .$$

Scores are independent of the discount factor. The following theorem is due to Fudenberg and Levine [19]:

Theorem 12 *Let C be the set of vectors v such that for every $\lambda \in \mathbb{R}^I$, $\langle \lambda, v \rangle \leq k(\lambda)$. If the interior of C is non-empty, $E(\delta)$ converges to C (for the Hausdorff topology) as δ goes to 1.*

Fudenberg, Levine and Takahashi [22] relax the non-empty interior assumption. They provide an algorithm for computing the affine hull of $\lim_{\delta \to 1} E(\delta)$ and provide a corresponding characterization of the set C with continuation payoffs belonging to this affine hull.

Folk Theorems for Public Equilibria The recursive structure of Theorem 11 and the asymptotic characterization of PPE payoffs given by Theorem 12 are essential tools for finding sufficient conditions under which every feasible and individually rational payoff is an equilibrium payoff, i. e. conditions under which a Folk Theorem holds.

The two conditions under which a Folk Theorem in PPEs holds are a 1) a condition of detectability of deviations and 2) a condition of identifiability of deviating players.

Definition 7 A profile of mixed actions $s = (s_i, s_{-i})$ has individual full rank if for each player i, the probability vectors (in the vector space $\mathbb{R}^{Y_{\text{pub}}}$)

$$\{Q(\cdot|a_i, s_{-i})\colon a_i \in A_i\}$$

are linearly independent.

If s has individual full rank, no player can change the distribution of his actions without affecting the distribution of public signals. Individual full rank is thus a condition on detectability of deviations.

Definition 8 A profile of mixed actions s has pairwise full rank if for every pair of players $i \neq j$, the family of probability vectors

$$\{Q(\cdot|a_i, s_{-i})\colon a_i \in A_i\} \cup \{Q(\cdot|a_j, s_{-j})\colon a_j \in A_j\}$$

has rank $|A_i| + |A_j| - 1$.

Under the condition of pairwise full rank, deviations from two distinct players induce distinct distributions of public signals. Pairwise full rank is therefore a condition of identifiability of deviating players.

Fudenberg et al. [21] prove the following theorem:

Theorem 13 *Assume the set of feasible and individually rational payoff vectors F has non-empty interior. If every pure action profile has individual full rank and if there exists a mixed action profile with pairwise full rank, then every convex and compact subset of the interior of F is a subset of $E(\delta)$ for δ large enough.*

In particular, under the conditions of the theorem, every feasible and individually rational payoff vector is arbitrarily close to a PPE payoff for large enough discount factors. Variations of this result can be found in [21] and [19].

Extensions

The Public Part of a Signal The definition of perfect public equilibria extends to the case in which each player's signals consists of two components: a public component and

a private component. The public components of all players' signals are the same with probability one. A public strategy is then a strategy that depends only on the public components of past signals, and all the analysis carries through.

Public Communication In the public communication extension of the repeated game, players make public announcements between any two stages of the repeated game. The profile of public announcements then forms a public signal, and recursive methods can be successfully applied. The fact that public communication is a powerful instrument to overcome the difficulties arising from private signals was first observed by Matsushima [49,50]. Ben Porath and Kahneman [14], Kandori and Matsushima [38], and Compte [15] prove Folk Theorems in games with private signals and public communication. Kandori [37] shows that in games with public monitoring, public communication allows to relax the conditions for the Folk Theorem of Fudenberg et al. [21].

Private Strategies in Games with Public Monitoring PPE payoffs do not cover the full set of sequential equilibrium payoffs, even when signals are public, as some equilibria may rely on players using *private strategies*, i. e. strategies that depend on past chosen actions and past private signals. See [53] and [39] for examples. In a *minority game*, there is an odd number of players, each player chooses between actions A and B. Players choosing the least chosen (minority) action get a payoffs of 1, other players get 0. The public signal is the minority action. Renault et al. [58,59] show that, for minority games, a Folk Theorem holds in private strategies but not in public strategies. Only few results are known concerning the set of sequential equilibrium payoffs in privates strategies of games with public monitoring. A monotonicity property is obtained by Kandori [36] who shows that the set of payoffs associated to sequential equilibria in pure strategies is increasing with respect to the quality of the public signal.

Almost Public Monitoring Some PPEs are robust to small perturbations of public signals. Considering strategies with finite memory, Mailath and Morris [52] identify a class of public strategies which are sequential equilibria of the repeated game with imperfect private monitoring, provided that the monitoring structure is close enough to a public one. They derive a Folk Theorem for games with almost public and almost perfect monitoring. Hörner and Olszewski [35] strengthen this result and prove a Folk Theorem for games with almost public monitoring. Under detectability and identifiability conditions, they prove that feasible and individually rational payoffs can be achieved by sequential equilibria with finite memory.

Almost Perfect Monitoring

Monitoring is almost perfect when each player can identify the action profile of his opponents with near certainty. Almost perfect monitoring is the natural framework to study the robustness of the Folk Theorem to small departures from the assumption that actions are perfectly observed.

The first results were obtained for the Prisoner's Dilemma. Sekiguchi [65] shows that the cooperative outcome can be approximated at equilibrium when players are sufficiently patient and monitoring is almost perfect. Under the same assumptions, Bhaskar and Obara [13], Piccione [57] and Ely and Valimaki [16] show that a Folk Theorem obtains.

Piccione [57] and Ely and Valimaki [16] study a particular class of equilibria called *belief free*. Strategies form a *belief free* equilibrium if, whatever player i's belief on the opponent's private history, the action prescribed by i's strategy is a best response to the opponent's continuation strategy.

Ely, Hörner and Olszewski [17] extend the belief free approach to general games. However, they show that, in general, belief free strategies are not enough to reconstruct a Folk Theorem, even when monitoring is almost perfect.

For general games and with any number of players, Hörner and Olszewski [34] prove a Folk Theorem with almost perfect monitoring. The strategies that implement the equilibrium payoffs are defined on successive blocks of a fixed length, and are block-belief-free in the sense that, at the beginning of every block, every player is indifferent between several continuation strategies, independently on his belief as to which continuation strategies are used by the opponents. This result closes the almost perfect monitoring case by showing that equilibrium payoffs in the Folk Theorem are robust to a small amount of imperfect monitoring.

General Stochastic Signals

Besides the case of public (or almost public) monitoring, little is known about equilibrium payoffs of repeated games with discounting and imperfect signals.

The Prisoner's Dilemma game is particularly important for economic applications. In particular, it captures the essential features of collusion with the possibility of secret price cutting, as in Stigler [69].

When signals are imperfect, but independent conditionally on the pair of actions chosen (a condition called conditional independence), Matsushima [51] shows that

the efficient outcome of the repeated Prisoner's Dilemma game is an equilibrium outcome if players are sufficiently patient. In the equilibrium construction, every player's action is constant in every block. The conditional independence assumption is crucial in that it implies that, during every block, a player has no feedback as to what signals the other player has received. The conditional independence assumption is non-generic: it holds for a set of monitoring structures of empty interior.

Fong, Gossner, Hörner, and Sannikov [18] prove that efficiency can be obtained at equilibrium without conditional independence. Their main assumption is that there exists a sufficiently informative signal, but this signal needs not be almost perfectly informative. Their result holds for a family of monitoring structures of non empty interior. It is the first result that establishes cooperation in the Prisoner's Dilemma with impatient players for truly imperfect, private and correlated signals.

Punishment Levels

Individual rationality is a key concept for Folk Theorems and equilibrium payoff characterizations. Given a repeated game, define the individually rational (IR) level of player i as the lowest payoff down to which this player may be punished in the repeated game.

Definition 9 The individual rational level of player i is:

$$\lim_{\delta\to 1} \min_{\sigma_{-i}} \max_{\sigma_i} \mathbf{E}_{\sigma_i,\sigma_{-i}} \left[\sum_t (1-\delta)\delta^{t-1} g_{i,t} \right]$$

where the min runs over profiles of behavior strategies for player $-i$, and the max over behavior strategies of player i.

That is, the individually rational level is the limit (as the discount factor goes to one) of the min max value of the discounted game (other approaches, through undiscounted games or limits of finitely repeated games, yield equivalent definitions, see [29]).

Comparison of the IR Level with the min max With perfect monitoring, the IR level of player i is player i's min max in the one-shot game, as defined by equation (1). With imperfect monitoring, the IR level for player i is never larger than v_i since player i's opponents can force down player i to v_{-i} by repeatedly playing the min max strategy against player i.

With two players, it is a consequence of von-Neumann's min max theorem [72] that v_i is the IR level for player i.

For any number of players, Gossner and Hörner [28] show that v_i is equal to the min max in the one-shot game whenever there exists a garbling of player i's signal such that, conditionally on i's garbled signal, the signals of i's opponents are independent. Furthermore, the condition in [28] is also a necessary condition in normal form games extended by correlation devices (as in Aumann [6]). A continuity result in the IR level also applies for monitoring structure close to those that satisfy the conditional independence condition.

The following example shows that, in general, the IR level can be lower than v_i:

Example 4 Consider the following three-player game. Player 1 chooses the row, player 2 the column and player 3 the matrix. Players 1 and 2 perfectly observe the action profile while player 3 observes player 2's action only. As we deal with the IR level of player 3, we specify the payoff for this player only.

	L	R	L	R
T	0	0	-1	0
B	0	-1	0	0
	W		E	

A simple computation shows that $v_3 = -\frac{1}{4}$ and that the min max strategies of players 1 and 2 are uniform. Consider the following strategies of players 1 and 2 in the repeated game: randomize uniformly at odd stages, play (T, L) or (B, R) depending on player 1's previous action at even stages. Against these strategies, player 3 cannot obtain better than $-\frac{1}{4}$ at odd stages and $-\frac{1}{2}$ at even stages, resulting in an average payoff of $-\frac{3}{8}$.

Entropy Characterizations The exact computation of the IR level in games with imperfect monitoring requires to analyze the optimal trade-off for punishing players between the production of correlated and private signals and the use of these signals for effective punishment. Gossner and Vieille, [31] and Gossner and Tomala [29] develop tools based on information theory to analyze this trade-off. At any stage, the amount of correlation generated (or spent) by the punishing players is measured using the entropy function. Gossner and Tomala [30] derive a characterization of the IR level for some classes of monitoring structures. Gossner, Laraki, and Tomala [32] provide methods explicit computations of the IR level. In particular, for the above example, the IR level computed and is about $-.401$. Explicit computations of IR levels for other games are derived by Goldberg [26].

Acknowledgments

The authors are grateful to Johannes Hörner for insightful comments.

Bibliography

Primary Literature

1. Abreu D (1988) On the theory of infinitely repeated games with discounting. Econometrica 56:383–396
2. Abreu D, Dutta P, Smith L (1994) The folk theorem for repeated games: a NEU condition. Econometrica 62:939–948
3. Abreu D, Pearce D, Stacchetti E (1990) Toward a theory of discounted repeated games with imperfect monitoring. Econometrica 58:1041–1063
4. Aumann RJ (1960) Acceptable points in games of perfect information. Pac J Math 10:381–417
5. Aumann RJ (1964) Mixed and behavior strategies in infinite extensive games. In: Dresder M, Shapley LS, Tucker AW (eds) Advances in Game Theory. Princeton University Press, New Jersey, pp 627–650
6. Aumann RJ (1974) Subjectivity and correlation in randomized strategies. J Math Econ 1:67–95
7. Aumann RJ (1981) Survey of repeated games. In: Aumann RJ (ed) Essays in game theory and mathematical economics in honor of Oskar Morgenstern. Wissenschaftsverlag, Bibliographisches Institut, Mannheim, pp 11–42
8. Aumann RJ, Shapley LS (1976) Long-term competition – a game theoretic analysis. Re-edited in 1994. See [9]
9. Aumann RJ, Shapley LS (1994) Long-term competition – a game theoretic analysis. In: Megiddo N (ed) Essays on game theory. Springer, New York, pp 1–15
10. Benoît JP, Krishna V (1985) Finitely repeated games. Econometrica 53(4):905–922
11. Blackwell D (1951) Comparison of experiments. In: Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley, pp 93–102
12. Blackwell D (1956) An analog of the minimax theorem for vector payoffs. Pac J Math 6:1–8
13. Bhaskar V, Obara I (2002) Belief-based equilibria in the repeated prisoners' dilemma with private monitoring. J Econ Theory 102:40–70
14. Ben EP, Kahneman M (1996) Communication in repeated games with private monitoring. J Econ Theory 70(2):281–297
15. Compte O (1998) Communication in repeated games with imperfect private monitoring. Econometrica 66:597–626
16. Ely JC, Välimäki J (2002) A robust folk theorem for the prisoner's dilemma. J Econ Theory 102:84–106
17. Ely JC, Hörner J, Olszewski W (2005) Belief-free equilibria in repeated games. Econometrica 73:377–415
18. Fong K, Gossner O, Hörner J, Sannikov Y (2007) Efficiency in a repeated prisoner's dilemma with imperfect private monitoring. mimeo
19. Fudenberg D, Levine DK (1994) Efficiency and observability with long-run and short-run players. J Econ Theory 62:103–135
20. Fudenberg D, Maskin E (1986) The folk theorem in repeated games with discounting or with incomplete information. Econometrica 54:533–554
21. Fudenberg D, Levine DK, Maskin E (1994) The folk theorem with imperfect public information. Econometrica 62(5):997–1039
22. Fudenberg D, Levine DK, Takahashi S (2007) Perfect public equilibrium when players are patient. Games Econ Behav 61:27–49
23. Forges F (1986). An approach to communication equilibria. Econometrica 54:1375–1385
24. Forges F, Mertens J-F, Neyman A (1986) A counterexample to the folk theorem with discounting. Econ Lett 20:7–7
25. Friedman J (1971) A noncooperative equilibrium for supergames. Rev Econ Stud 38:1–12
26. Goldberg Y (2007) Secret correlation in repeated games with imperfect monitoring: The need for nonstationary strategies. Math Oper Res 32:425–435
27. Gossner O (1995) The folk theorem for finitely repeated games with mixed strategies. Int J Game Theory 24:95–107
28. Gossner O, Hörner J (2006) When is the individually rational payoff in a repeated game equal to the minmax payoff? DP 1440, CMS-EMS
29. Gossner O, Tomala T (2006) Empirical distributions of beliefs under imperfect observation. Math Oper Res 31(1):13–30
30. Gossner O, Tomala T (2007) Secret correlation in repeated games with signals. Math Oper Res 32:413–424
31. Gossner O, Vieille N (2002) How to play with a biased coin? Games Econ Behav 41:206–226
32. Gossner O, Laraki R, Tomala T (2009) Informationally optimal correlation. Math Programm B 116:147–112
33. Green EJ, Porter RH (1984) Noncooperative collusion under imperfect price information. Econometrica 52:87–100
34. Hörner J, Olszewski W (2006) The folk theorem with private almost-perfect monitoring. Econometrica 74(6):1499–1544
35. Hörner J, Olszewski W (2007) How robust is the folk theorem with imperfect public monitoring? mimeo
36. Kandori M (1992) The use of information in repeated games with imperfect monitoring. Rev Econ Stud 59:581–593
37. Kandori M (2003) Randomization, communication, and efficiency in repeated games with imperfect public monitoring. Econometrica 71:345–353
38. Kandori M, Matsushima H (1998) Private observation, communication and collusion. Rev Econ Stud 66:627–652
39. Kandori M, Obara I (2006) Efficiency in repeated games revisited: The role of private strategies. Econometrica 74:499–519
40. Kohlberg E (1975) Optimal strategies in repeated games with incomplete information. Int J Game Theory 4:7–24
41. Kuhn HW (1953) Extensive games and the problem of information. In: Kuhn HW, Tucker AW (eds) Contributions to the Theory of Games, vol II. Annals of Mathematical Studies, vol 28. Princeton University Press, New Jersey, pp 193–216
42. Kreps DM, Wilson RB (1982) Sequential equilibria. Econometrica 50:863–894
43. Lehrer E (1990) Nash equilibria of *n*-player repeated games with semi-standard information. Int J Game Theory 19:191–217
44. Lehrer E (1991) Internal correlation in repeated games. Int J Game Theory 19:431–456
45. Lehrer E (1992) Correlated equilibria in two-player repeated games with nonobservable actions. Math Oper Res 17:175–199
46. Lehrer E (1992) Two players repeated games with non observable actions and observable payoffs. Math Oper Res 17:200–224

47. Lehrer E, Pauzner A (1999) Repeated games with differential time preferences. Econometrica 67:393–412
48. Luce RD, Raiffa H (1957) Games and Decisions: Introduction and Critical Survey. Wiley, New York
49. Matsushima H (1991) On the theory of repeated games with private information: Part i: Anti-folk theorem without communication. Econ Lett 35:253–256
50. Matsushima H (1991) On the theory of repeated games with private information: Part ii: Revelation through communication. Econ Lett 35:257–261
51. Matsushima H (2004) Repeated games with private monitoring: Two players. Econometrica 72:823–852
52. Mailath G, Morris S (2002) Repeated games with almost-public monitoring. J Econ Theory 102:189–229
53. Mailath GJ, Matthews SA, Sekiguchi T (2002) Private strategies in finitely repeated games with imperfect public monitoring. B E J Theor Econ 2
54. Myerson RB (1982) Optimal coordination mechanisms in generalized principal-agent problems. J Math Econ 10:67–81
55. Nash JF (1951) Noncooperative games. Ann Math 54:289–295
56. Neyman A (1999) Cooperation in repeated games when the number of stages is not commonly known. Econometrica 67:45–64
57. Piccione M (2002) The repeated prisoner's dilemma with imperfect private monitoring. J Econ Theory 102:70–84
58. Renault J, Scarlatti S, Scarsini M (2005) A folk theorem for minority games. Games Econ Behav 53:208–230
59. Renault J, Scarlatti S, Scarsini M (2008) Discounted and finitely repeated minority games with public signals. Math Soc Sci 56:44–74
60. Renault J, Tomala T (1998) Repeated proximity games. Int J Game Theory 27:539–559
61. Renault J, Tomala T (2004) Communication equilibrium payoffs of repeated games with imperfect monitoring. Games Econ Behav 49:313–344
62. Rubinstein A (1977) Equilibrium in supergames. Center for Research in Mathematical Economics and Game Theory. Res Memo 25
63. Rubinstein A (1979) Equilibrium in supergames with the overtaking criterion. J Econ Theory 21:1–9
64. Rubinstein A (1994) Equilibrium in supergames. In: Megiddo N (ed) Essays on game theory. Springer, New York, pp 17–28
65. Sekiguchi T (1997) Efficiency in repeated prisoner's dilemma with private monitoring. J Econ Theory 76:345–361
66. Selten R (1965) Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit. Z Gesamte Staatswiss 12:201–324
67. Smith L (1995) Necessary and sufficient conditions for the perfect finite horizon folk theorem. Econometrica 63:425–430
68. Sorin S (1986) On repeated games with complete information. Math Oper Res 11:147–160
69. Stigler G (1964) A theory of oligopoly. J Political Econ 72:44–61
70. Tomala T (1998) Pure equilibria of repeated games with public observation. Int J Game Theory 27:93–109
71. Tomala T (1999) Nash equilibria of repeated games with observable payoff vector. Games Econ Behav 28:310–324
72. von Neumann J (1928) Zur Theorie der Gesellschaftsspiele. Math Ann 100:295–320
73. Wen Q (1994) The "folk theorem" for repeated games with complete information. Econometrica 62:949–954

Books and Reviews

Aumann RJ (1981) Survey of repeated games. In: Aumann RJ (ed) Essays in game theory and mathematical economics in honor of Oskar Morgenstern. Wissenschaftsverlag, Bibliographisches Institut, Mannheim, pp 11–42

Mailath GJ, Samuelson L (2006) Repeated Games and Reputations: Long-Run Relationships. Oxford University Press, Oxford

Mertens J-F (1986) Repeated games. In: Proceedings of the international congress of Mathematicians. Berkeley, California, pp 1528–1577

Mertens J-F, Sorin S, Zamir S (1994) Repeated games. CORE discussion paper. pp 9420–9422, Univeritsé Catholique de Levain, Louvain-la-neuve

Repeated Games with Incomplete Information

Jérôme Renault
Ceremade, Université Paris Dauphine, Paris, France

Article Outline

Glossary
Definition of the Subject
Strategies, Payoffs, Value and Equilibria
The Standard Model of Aumann and Maschler
Vector Payoffs and Approachability
Zero-Sum Games with Lack of Information on Both Sides
Non Zero-Sum Games with Lack of Information on One Side
Non-observable Actions
Miscellaneous
Future Directions
Acknowledgments
Bibliography

Glossary

Repeated game with incomplete information A situation where several players repeat the same stage game, the players having different knowledge of the stage game which is repeated.

Strategy of a player A rule, or program, describing the action taken by the player in any possible situation which may happen, depending on the information available to this player in that situation.

Strategy profile A vector containing a strategy for each player.

Lack of information on one side Particular case where all the players but one perfectly know the stage game which is repeated.

Zero-sum games 2-player games where the players have opposite payoffs.

Value Solution (or price) of a zero-sum game, in the sense of the fair amount that player 1 should give to player 2 to be entitled to play the game.

Equilibrium Strategy profile where each player's strategy is in best reply against the strategy of the other players.

Completely revealing strategy Strategy of a player which eventually reveals to the other players everything known by this player on the selected state.

Non revealing strategy Strategy of a player which reveals nothing on the selected state.

The simplex of probabilities over a finite set For a finite set S, we denote by $\Delta(S)$ the set of probabilities over S, and we identify $\Delta(S)$ to $\{p = (p_s)_{s\in S} \in \mathbb{R}^S, \forall s \in S\ p_s \geq 0 \text{ and } \sum_{s\in S} p_s = 1\}$. Given s in S, the Dirac measure on s will be denoted by δ_s. For $p = (p_s)_{s\in S}$ and $q = (q_s)_{s\in S}$ in $\mathbb{R}^S$, we will use, unless otherwise specified, $\|p - q\| = \sum_{s\in S} |p_s - q_s|$.

Definition of the Subject

Introduction

In a repeated game with incomplete information, there is a basic interaction called stage game which is repeated over and over by several participants called players. The point is that the players do not perfectly know the stage game which is repeated, but rather have different knowledge about it. As illustrative examples, one may think of the following situations: an oligopolistic competition where firms don't know the production costs of their opponents, a financial market where traders bargain over units of an asset which terminal value is imperfectly known, a cryptographic model where some participants want to transmit some information (e. g., a credit card number) without being understood by other participants, a conflict when a particular side may be able to understand the communications inside the opponent side (or might have a particular type of weapons), ...

Natural questions arising in this context are as follows. What is the optimal behavior of a player with a perfect knowledge of the stage game? Can we determine which part of the information such a player should use? Can we price the value of possessing a particular information? How should one player behave while having only a partial information?

Foundations of games with incomplete information have been studied in [28] and [56]. Repeated games with incomplete information have been introduced in the sixties by Aumann and Maschler [6], and we present here the basic and fundamental results of the domain. Let us start with a few well known elementary examples [6,91].

Basic Examples In each example, there are two players, and the game is zero-sum, i. e. player 2's payoff always is the opposite of player 1's payoff. There are two states a and b, and the possible stage games are given by two real matrices G^a and G^b with identical size. Initially a true state of nature $k \in \{a, b\}$ is selected with even probability between a and b, and k is announced *to player 1 only*. Then the matrix game G^k is repeated over and over: at every stage, simultaneously player 1 chooses a row i, whereas player 2 chooses a column j, the stage payoff for player 1 is then $G^k(i, j)$ but only i and j are publicly announced before proceeding to the next stage. Players are patient and want to maximize their long-run average expected payoffs.

Example 1

$$G^a = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix} \quad \text{and} \quad G^b = \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}.$$

This example is very simple. In order to maximize his payoff, player 1 just has to play, at any stage, the *Top* row if the state is a and the *Bottom* row if the state is b. This corresponds to playing optimally in each possible matrix game.

Example 2

$$G^a = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad G^b = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

A naive strategy for player 1 would be to play at stage 1: *Top* if the state is a, and *Bottom* if the state is b. Such a strategy is called completely revealing, or CR, because it allows player 2 to deduce the selected state from the observation of the actions played by player 1. This strategy of player 1 would be optimal here if a single stage was to be played, but it is a very weak strategy on the long run and does not guarantee more than zero at each stage $t \geq 2$ (because player 2 can play *Left* or *Right* depending on player 1's first action).

On the opposite, player 1 may not use his information and play a non revealing, or NR, strategy, i. e. a strategy which is independent of the selected state. He can consider the average matrix

$$\frac{1}{2}G^a + \frac{1}{2}G^b = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix},$$

and play independently at each stage an optimal mixed action in this matrix, i. e. here the unique mixed action $\frac{1}{2}Top + \frac{1}{2}Bottom$. It will turn out that this is here the optimal behavior for player 1, and the value of the repeated game is the value of the average matrix, i. e. 1/4.

Example 3

$$G^a = \begin{pmatrix} 4 & 0 & 2 \\ 4 & 0 & -2 \end{pmatrix} \quad \text{and} \quad G^b = \begin{pmatrix} 0 & 4 & -2 \\ 0 & 4 & 2 \end{pmatrix}.$$

Playing a CR strategy for player 1 does not guarantee more than zero in the long-run, because player 2 will eventually be able to play *Middle* if the state is a, and *Left* if the state is b. But a NR strategy will not do better, because the average matrix

$$\frac{1}{2}G^a + \frac{1}{2}G^b \text{ is } \begin{pmatrix} 2 & 2 & 0 \\ 2 & 2 & 0 \end{pmatrix},$$

hence has value 0.

We will see later that an optimal strategy for player 1 in this game is to play as follows. Initially, player 1 chooses an element s in $\{T, B\}$ as follows: if $k = a$, then $s = T$ with probability 3/4, and thus $s = B$ with probability 1/4; and if $k = b$, then $s = T$ with probability 1/4, and $s = B$ with probability 3/4. Then at each stage player 1 plays row s, independently of the actions taken by player 2. The conditional probabilities satisfy: $P(k = a|s = T) = 3/4$, and $P(k = a|s = B) = 1/4$. At the end of stage 1, player 2 will have learned, from the action played by his opponent, something about the selected state: his belief on the state will move from $\frac{1}{2}a + \frac{1}{2}b$ to $\frac{3}{4}a + \frac{1}{4}b$ or to $\frac{1}{4}a + \frac{3}{4}b$. But player 2 still does not know perfectly the selected state. Such a strategy of player 1 is called *partially revealing*.

General Definition

Formally, a repeated game with incomplete information is given by the following data. There is a set of players N, and a set of states K. Each player i in N has a set of actions A^i and a set of signals U^i, and we denote by $A = \prod_{i\in N} A^i$ the set of action profiles and by $U = \prod_{i\in N} U^i$ the set of signal profiles. Every player i has a payoff function $g^i : K \times A \longrightarrow \mathbb{R}$. There is a signaling function $q: K \times A \longrightarrow \Delta(U)$, and an initial probability $\pi \in \Delta(K \times U)$. In what follows, we will always assume the sets of players, states, actions and signals to be non empty and finite.

A repeated game with incomplete information can thus be denoted by $\Gamma = (N, K, (A^i)_{i\in N}, (U^i)_{i\in N}, (g^i)_{i\in N}, q, \pi)$. The progress of the game is the following.

- Initially, an element $(k, (u_0^i)_i)$ is selected according to π: k is the realized state of nature and will remain fixed, and each player i learns u_0^i (and nothing more than u_0^i).
- At each integer stage $t \geq 1$, simultaneously every player i chooses an action a_t^i in A^i, and we denote by $a_t = (a_t^i)_i$ the action profile played at stage t. The stage payoff of a player i is then given by $g^i(k, a_t)$. A signal profile $(u_i^t)_i$ is selected according to $q(k, a_t)$, and each player i learns u_t^i (and nothing more than u_i^t) before proceeding to the next stage.

Remark

1. We will always assume that during the play, each player remembers the past actions he has chosen, as well as the past signals he has received. And players will be allowed to select their action independently at random.
2. The players do not necessarily know their stage payoff after each stage (as an illustration, imagine the players bargaining over units of an asset which terminal value will only be known "at the end" of the game). This is without loss of generality, because it is possible to add hypotheses on q so that each player will be able to deduce his stage payoff from his realized stage signal.
3. Repeated games with complete information are a particular case, corresponding to the situation where each initial signal u_0^i reveals the selected state. Such games are studied in the chapter "Repeated games with complete information".
4. Games where the state variable k evolve from stage to stage, according to the actions played, are called stochastic games. These games are not covered here, but in a specific chapter entitled "Stochastic games".
5. The most standard case of signaling function is when each player exactly learns, at the end of each stage t, the whole action profile a_t. Such games are usually called games with "perfect monitoring", "full monitoring", "perfect observation" or with "observable actions".

Strategies, Payoffs, Value and Equilibria

Strategies

A (behavior) strategy for player i is a rule, or program, describing the action taken by this player in any possible case which may happen. These actions may be chosen at random, so a strategy for player i is an element $\sigma^i = (\sigma_t^i)_{t\geq 1}$, where for each t, σ_t^i is a mapping from $U^i \times (U^i \times A^i)^{t-1}$ to $\Delta(A^i)$ giving the lottery played by player i at stage t as a function of the past signals and actions of player i. The set of strategies for player i is denoted by Σ^i.

A history of length t in Γ is a sequence $(k, u_0, a_1, u_1, \ldots, a_t, u_t)$, and the set of such histories is the finite set $K \times U \times (A \times U)^t$. An infinite history is called a play, the set of plays is denoted by $\Omega = K \times U \times (A \times U)^\infty$ and is endowed with the product σ-algebra. A strategy profile $\sigma = (\sigma^i)_i$ naturally induces, together with the initial probability π, a probability distribution over the set of histories of length t. This probability uniquely extends to a probability over plays, and is denoted by $\mathbb{P}_{\pi,\sigma}$.

Payoffs

Given a time horizon T, the average expected payoff of player i, up to stage T, if the strategy profile σ is played

is denoted by:

$$\gamma_T^i(\sigma) = \mathbb{E}_{\mathbb{P}_{\pi,\sigma}}\left(\frac{1}{T}\sum_{t=1}^{T} g^i(k, a_t)\right).$$

The T-stage game is the game Γ_T where simultaneously, each player i chooses a strategy σ^i in Σ^i, then receives the payoff $\gamma_T^i((\sigma^j)_{j\in N})$.

Given a discount factor λ in $(0, 1]$, the λ-discounted payoff of player i is denoted by:

$$\gamma_\lambda^i(\sigma) = \mathbb{E}_{\mathbb{P}_{\pi,\sigma}}\left(\lambda\sum_{t=1}^{\infty}(1-\lambda)^{t-1} g^i(k, a_t)\right).$$

The λ-discounted game is the game Γ_λ where simultaneously, each player i chooses a strategy σ^i in Σ^i, then receives the payoff $\gamma_\lambda^i((\sigma^j)_{j\in N})$.

Remark A strategy for player i is called *pure* if it always plays in a deterministic way. A *mixed strategy* for player i is defined as a probability distribution over the set of pure strategies (endowed with the product σ-algebra). Kuhn's theorem (see [3,38] or [85] for a modern presentation) states that mixed strategies or behavior strategies are equivalent, in the following sense: for each behavior strategy σ^i, there exists a mixed strategy τ^i of the same player such that $\mathbb{P}_{\pi,\sigma^i,\sigma^{-i}} = \mathbb{P}_{\pi,\tau^i,\sigma^{-i}}$ for any strategy profile σ^{-i} of the other players, and *vice-versa* if we exchange the words "behavior" and "mixed". Unless otherwise specified, the word strategy will refer here to a behavior strategy, but we will also sometimes equivalently use mixed strategies, or even mixtures of behavior strategies.

Value of Zero-Sum Games

By definition the game is zero-sum if there are two players, say player 1 and player 2, with opposite payoffs. The T-stage game Γ_T can then be seen as a matrix game, hence by the minmax theorem it has a value $v_T = \sup_{\sigma^1}\inf_{\sigma^2}\gamma_T^1(\sigma^1,\sigma^2) = \inf_{\sigma^2}\sup_{\sigma^1}\gamma_T^1(\sigma^1,\sigma^2)$. Similarly, one can use Sion's theorem [74] to show that the λ-discounted game has a value $v_\lambda = \sup_{\sigma^1}\inf_{\sigma^2}\gamma_\lambda^1(\sigma^1,\sigma^2) = \inf_{\sigma^2}\sup_{\sigma^1}\gamma_\lambda^1(\sigma^1,\sigma^2)$.

To study long term strategic aspects, it is also important to consider the following notion of uniform value. Players are asked to play well uniformly in the time horizon, i. e. simultaneously in all game Γ_T with T sufficiently large (or similarly uniformly in the discount factor, i. e. simultaneously in all game Γ_λ with λ sufficiently low).

Definition 1 Player 1 can guarantee the real number v in the repeated game Γ if: $\forall\varepsilon > 0, \exists\sigma^1 \in \Sigma^1, \exists T_0, \forall T \geq T_0, \forall\sigma^2 \in \Sigma^2, \gamma_T^1(\sigma^1,\sigma^2) \geq v - \varepsilon$. Similarly, Player 2 can guarantee v in Γ if $\forall\varepsilon > 0, \exists\sigma^2 \in \Sigma^2, \exists T_0, \forall T \geq T_0, \forall\sigma^1 \in \Sigma^1, \gamma_T^1(\sigma^1,\sigma^2) \leq v + \varepsilon$. If both player 1 and player 2 can guarantee v, then v is called the uniform value of the repeated game. A strategy σ^1 of player 1 satisfying $\exists T_0, \forall T \geq T_0, \forall\sigma^2 \in \Sigma^2, \gamma_T^1(\sigma^1,\sigma^2) \geq v$ is then called an optimal strategy of player 1 (optimal strategies of player 2 are defined similarly).

The uniform value, whenever it exists, is necessarily unique. Its existence is a strong property, which implies that both v_T, as T goes to infinity, and v_λ, as λ goes to zero, converge to the uniform value.

Equilibria of General-Sum Games

In the general case, the T-stage game Γ_T can be seen as the mixed extension of a finite game, and consequently possesses a Nash equilibrium. Similarly, the discounted game Γ_λ always has, by the Nash Glicksberg theorem, a Nash equilibrium. Concerning uniform notions, couples of optimal strategies are generalized as follows.

Definition 2 A strategy profile $\sigma = (\sigma^i)_{i\in N}$ is a uniform Nash equilibrium of Γ if: 1) $\forall\varepsilon > 0, \sigma$ is an ε-Nash equilibrium in every finitely repeated game sufficiently long, that is: $\exists T_0, \forall T \geq T_0, \forall i \in N, \forall\tau^i \in \Sigma^i, \gamma_T^i(\tau^i,\sigma^{-i}) \leq \gamma_T^i(\sigma) + \varepsilon$, and 2) the sequence of payoffs $\left((\gamma_T^i(\sigma))_{i\in N}\right)_T$ converges to a limit payoff $(\gamma^i(\sigma))_{i\in N}$ in $\mathbb{R}^N$.

Remark The initial probability π will play a great role in the following analyzes, so we will often write $\gamma_T^{i,\pi}(\sigma)$ for $\gamma_T^i(\sigma)$, $v_T(\pi)$ for the value v_T, etc.

The Standard Model of Aumann and Maschler

This famous model has been introduced in the sixties by Aumann and Maschler (see the reedition [6]). It deals with zero-sum games with lack of information on one side and observable actions, as in the basic examples previously presented. There is a finite set of states K, an initial probability $p = (p^k)_{k\in K}$ on K, and a family of matrix games G^k with identical size $I \times J$. Initially, a state k in K is selected according to p, and announced to player 1 (called the informed player) only. Then the matrix game G^k is repeated over and over: at every stage, simultaneously player 1 chooses a row i in I, whereas player 2 chooses a column j in J, the stage payoff for player 1 is then $G^k(i, j)$ but only i and j are publicly announced before proceeding to the next stage. Denote by M the constant $\max_{k,i,j}|G^k(i, j)|$.

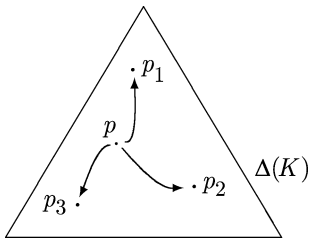

Repeated Games with Incomplete Information, Figure 1
Splitting

Basic Tools: Splitting, Martingale, Concavification, and the Recursive Formula

The following aspects are simple but fundamental. The initial probability $p = (p^k)_{k\in K}$ represents the initial belief, or a priori, of player 2 on the selected state of nature. Assume that player 1 chooses his first action (or more generally a message or signal s from a finite set S) according to a probability distribution depending on the state, i. e. according to a transition probability $x = (x^k)_{k\in K} \in \Delta(S)^K$. For each signal s, the probability that s is chosen is denoted $\lambda(x,s) = \sum_k p^k x^k(s)$, and given s such that $\lambda(x,s) > 0$ the conditional probability on K, or a posteriori of player 2, is $\hat{p}(x,s) = \left((p^k x^k(s))/(\lambda(x,s))\right)_{k\in K}$. We clearly have:

$$p = \sum_{s\in S} \lambda(x,s)\hat{p}(x,s)\,. \tag{1}$$

So the a priori p lies in the convex hull of the a posteriori. The following lemma expresses a reciprocal: player 1 is able to induce any family of a posteriori containing p in its convex hull.

Lemma 1 (Splitting) *Assume that p is written as a convex combination $p = \sum_{s\in S} \lambda_s p_s$ with positive coefficients. There exists a transition probability $x \in \Delta(S)^K$ such that $\forall s \in S, \lambda_s = \lambda(x,s)$ and $p_s = \hat{p}(x,s)$.*

Proof Just put $x^k(s) = \frac{\lambda_s p_s^k}{p^k}$ if $p^k > 0$. □

Equation (1) not only tells that the a posteriori contains p in their convex hull, but also that the expectation of the a posteriori is the a priori. We are here in a repeated context, and for every strategy profile σ one can define the process $(p_t(\sigma))_{t\geq 0}$ of the a posteriori of player 2. We have $p_0 = p$, and $p_t(\sigma)$ is the random variable of player 2's belief on the state after the first t stages. More precisely, we define for any $t \geq 0$, $h_t = (i_1, j_1, \ldots, i_t, j_t) \in (I \times J)^t$ and k in K:

$$p_t^k(\sigma, h_t) = \mathbb{P}_{p,\sigma}(k|h_t) = \frac{p^k \mathbb{P}_{\delta^k,\sigma}(h_t)}{\mathbb{P}_{p,\sigma}(h_t)}\,.$$

$p_t(\sigma, h_t) = (p_t^k(\sigma, h_t))_{k\in K} \in \Delta(K)$ (arbitrarily defined if $\mathbb{P}_{p,\sigma}(h_t) = 0$) is the conditional probability on the state of nature given that σ is played and h_t has occurred in the first t stages. It is easy to see that as soon as $\mathbb{P}_{p,\sigma}(h_t) > 0$, $p_t(\sigma, h_t)$ does not depend on player 2's strategy σ^2, nor on player 2's last action j_t. It is fundamental to notice that:

Lemma 2 (Martingale of a posteriori) *$(p_t(\sigma))_{t\geq 0}$ is a $\mathbb{P}_{p,\sigma}$-martingale with values in $\Delta(K)$.*

This is indeed a general property of Bayesian learning of a fixed unknown parameter: the expectation of what I will know tomorrow is what I know today. This martingale is controlled by the informed player, and the splitting lemma shows that this player can essentially induce any martingale issued from the a priori p. Notice that, to be able to compute the realizations of the martingale, player 2 needs to know the strategy σ^1 used by player 1.

The splitting lemma also easily gives the following concavification result. Let f be a continuous mapping from $\Delta(K)$ to $\mathbb{R}$. The smallest concave function above f is denoted by $\mathrm{cav} f$, and we have:

$$\mathrm{cav} f(p) = \max\left\{ \sum_{s\in S} \lambda_s f(p_s),\, S \text{ finite},\right.$$
$$\left. \forall s \lambda_s \geq 0,\, p_s \in \Delta(K), \sum_{s\in S} \lambda_s = 1, \sum_{s\in S} \lambda_s p_s = p \right\}.$$

Lemma 3 (Concavification) *If for any initial probability p, the informed player can guarantee $f(p)$ in the game $\Gamma(p)$, then for any p this player can also guarantee $\mathrm{cav} f(p)$ in $\Gamma(p)$.*

Non Revealing Games

As soon as player 1 uses a strategy which depends on the selected state, the martingale of a posteriori will move and player 2 will have learned something on the state. This is the dilemma of the informed player: he can not use the information on the state without revealing information. Imagine now that player 1 decides to reveal no information on the selected state, and plays independently of it. Since payoffs are defined via expectations, it is as if the players were repeating the average matrix game $G(p) = \sum_{k\in K} p^k G^k$. Its value is:

$$u(p) = \max_{x\in\Delta(I)} \min_{y\in\Delta(J)} \sum_{i,j} x(i)y(j)G(p)(i,j)$$
$$= \min_{y\in\Delta(J)} \max_{x\in\Delta(I)} \sum_{i,j} x(i)y(j)G(p)(i,j)\,.$$

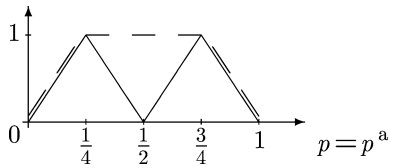

Repeated Games with Incomplete Information, Figure 2
u* and cav*u

u is a Lispchitz function, with constant M, from $\Delta(K)$ to $\mathbb{R}$. Clearly, player 1 can guarantee $u(p)$ in the game $\Gamma(p)$ by playing i. i. d. at each stage an optimal strategy in $G(p)$. By the concavification lemma, we obtain:

Proposition 1 *Player 1 can guarantee* cav$u(p)$ *in the game* $\Gamma(p)$.

Let us come back to the examples. In example 1, we have

$$u(p) = \mathrm{Val}\begin{pmatrix} -(1-p) & 0 \\ 0 & -p \end{pmatrix} = -p(1-p)\,,$$

where $p \in [0,1]$ stands here for the probability of state a. This is a convex function of p, and cav$u(p) = 0$ for all p. In example 2, $u(p) = p(1-p)$ for all p, hence u is already concave and cav$u = u$. Regarding example 3, the following picture show the functions u (regular line), and cav$u$ (dashed line).

Let us consider again the partially revealing strategy previously described. With probability 1/2, the a posteriori will be $3/4a + 1/4b$, and player 1 will play *Top* which is optimal in $3/4G^a + 1/4G^b = \begin{pmatrix} 3 & 1 & 1 \\ 3 & 1 & -1 \end{pmatrix}$. Similarly with probability 1/2, the a posteriori will be $1/4a + 3/4b$ and player 1 will play an optimal strategy in $1/4G^a + 3/4G^b$. Consequently, this strategy guarantees $1/2u(3/4) + 1/2u(1/4) = \mathrm{cav}u(1/2) = 1$ to player 1.

Player 2 can Guarantee the Limit Value

In the infinitely repeated game with initial probability p, player 2 can play as follows: T being fixed, he can play an optimal strategy in the T-stage game $\Gamma_T(p)$, then forget everything and play again an optimal strategy in the T-stage game $\Gamma_T(p)$, etc.... By doing so, he guarantees $v_T(p)$ in the game $\Gamma(p)$. So he can guarantee $\inf_T v_T(p)$ in this game, and this implies that $\limsup_T v_T(p) \le \inf_T v_T(p)$. As a consequence, we obtain:

Proposition 2 *The sequence* $(v_T(p))_T$ *converges to* $\inf_T v_T(p)$, *and this limit can be guaranteed by player 2 in the game* $\Gamma(p)$.

Uniform Value: cav *u* Theorem

We will see here that $\lim_T v_T(p)$ is nothing but cav$u(p)$, and since this quantity can be guaranteed by both players this is the uniform value of the game $\Gamma(p)$. The idea of the proof is the following. The martingale $(p_t(\sigma))_{t\ge 0}$ is bounded, hence will converge almost surely, and we have a bound on its L^1 variation (see Lemma 4 below). This means that after a certain stage the martingale will essentially remain constant, so approximately player 1 will play in a non revealing way, so will not be able to have a stage payoff greater than $u(q)$, where q if a "limit a posteriori". Since the expectation of the a posteriori is the a priori p, player 1 can not guarantee more than $\max\{\sum_{s\in S}\lambda_s u(p_s), S \text{ finite}, \forall s \in S \lambda_s \ge 0, p_s \in \Delta(K), \sum_{s\in S}\lambda_s = 1, \sum_{s\in S}\lambda_s p_s = p\}$, that is more than cav$u(p)$. Let us now proceed to the formal proof.

Fix a strategy σ^1 of player 1, and define the strategy σ^2 of player 2 as follows: play at each stage an optimal strategy in the matrix game $G(p_t)$, where p_t is the current a posteriori in $\Delta(K)$. Assume that $\sigma = (\sigma^1, \sigma^2)$ is played in the repeated game $\Gamma(p)$. To simplify notations, we write $\mathbb{P}$ for $\mathbb{P}_{p,\sigma}$, $p_t(h_t)$ for $p_t(\sigma, h_t)$, etc. We use everywhere norms $\|.\|_1$. To avoid confusion between variables and random variables in the following computations, we will use tildes to denote random variables, e. g. $\tilde{k}$ will denote the random variable of the selected state.

Lemma 4

$$\forall T \ge 1,\quad \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}(\|p_{t+1}-p_t\|) \le \frac{\sum_{k\in K}\sqrt{p^k(1-p^k)}}{\sqrt{T}}.$$

Proof This is a property of martingales with values in $\Delta(K)$ and expectation p. We have for each state k and $t \ge 0$: $\mathbb{E}\left((p^k_{t+1} - p^k_t)^2\right) = \mathbb{E}(\mathbb{E}((p^k_{t+1} - p^k_t)^2|\mathcal{H}_t))$, where $\mathcal{H}_t$ is the σ-algebra on plays generated by the first t action profiles. So $\mathbb{E}\left((p^k_{t+1} - p^k_t)^2\right) = \mathbb{E}(\mathbb{E}((p^k_{t+1})^2 + (p^k_t)^2 - 2p^k_{t+1}p^k_t|\mathcal{H}_t)) = \mathbb{E}((p^k_{t+1})^2) - \mathbb{E}((p^k_t)^2)$. So $\mathbb{E}\left(\sum_{t=0}^{T-1}(p^k_{t+1}-p^k_t)^2\right) = \mathbb{E}\left((p^k_T)^2\right) - (p^k)^2 \le p^k(1-p^k)$. By Cauchy- Schwartz inequality, we also have for each k,

$$\mathbb{E}\left(\frac{1}{T}\sum_{t=0}^{T-1}\left|p^k_{t+1} - p^k_t\right|\right) \le \sqrt{\frac{1}{T}\mathbb{E}\left(\sum_{t=0}^{T-1}(p^k_{t+1} - p^k_t)^2\right)}$$

and the result follows. □

For h_t in $(I \times J)^t$, $\sigma^1_{t+1}(k, h_t)$ is the mixed action in $\Delta(I)$ played by player 1 at stage $t+1$ if the state is k and h_t has previously occurred, and we write $\bar{\sigma}^1_{t+1}(h_t)$ for the law of the action of player 1 of stage $t+1$ after h_t: $\bar{\sigma}^1_{t+1}(h_t) =$

$\sum_{k\in K} p_t^k(h_t)\sigma_{t+1}^1(k,h_t) \in \Delta(I)$. $\bar{\sigma}_{t+1}(h_t)$ can be seen as the average action played by player 1 after h_t, and will be used as a non revealing approximation for $(\sigma_{t+1}^1(k,h_t))_k$. The next lemma precisely links the variation of the martingale $(p_t(\sigma))_{t\geq 0}$, i. e. the information revealed by player 1, and the dependence of player 1's action on the selected state, i. e. the information used by player 1.

Lemma 5

$$\forall t \geq 0, \forall h_t \in (I\times J)^t, \mathbb{E}\left(\|p_{t+1}-p_t\|\,|h_t\right) = \mathbb{E}\left(\left\|\sigma_{t+1}^{\tilde{k}}(h_t)-\bar{\sigma}_{t+1}(h_t)\right\|\,|h_t\right) .$$

Proof Fix $t\geq 0$ and h_t in $(I\times J)^t$ s.t. $\mathbb{P}_{p,\sigma}(h_t)>0$. For (i_{t+1},j_{t+1}) in $I\times J$, one has:

$$\begin{aligned} p_{t+1}^k(h_t,i_{t+1},j_{t+1}) &= \mathbb{P}(\tilde{k}=k|h_t,i_{t+1}) \\ &= \frac{\mathbb{P}(\tilde{k}=k|h_t)\mathbb{P}(i_{t+1}|k,h_t)}{\mathbb{P}(i_{t+1}|h_t)} \\ &= \frac{p_t^k(h_t)\sigma_{t+1}^1(k,h_t)(i_{t+1})}{\bar{\sigma}_{t+1}^1(h_t)(i_{t+1})} . \end{aligned}$$

Consequently,

$$\begin{aligned} &\mathbb{E}\left(\|p_{t+1}-p_t\||h_t\right) \\ &= \sum_{i_{t+1}\in I}\bar{\sigma}_{t+1}^1(h_t)(i_{t+1})\sum_{k\in K}|p_{t+1}^k(h_t,i_{t+1})-p_t^k(h_t)| \\ &= \sum_{i_{t+1}\in I}\sum_{k\in K}|p_t^k(h_t)\sigma_{t+1}^1(k,h_t)(i_{t+1}) \\ &\quad -\bar{\sigma}_{t+1}^1(h_t)(i_{t+1})p_t^k(h_t)| \\ &= \sum_{k\in K}p_t^k(h_t)\|\sigma_{t+1}^1(k,h_t)-\bar{\sigma}_{t+1}^1(h_t)\| \\ &= \mathbb{E}\left(\|\sigma_{t+1}^1(\tilde{k},h_t)-\bar{\sigma}_{t+1}^1(h_t)\||h_t\right) . \end{aligned}$$

□

We can now control payoffs. For $t\geq 0$ and h_t in $(I\times J)^t$:

$$\begin{aligned} &\mathbb{E}\left(G^{\tilde{k}}(\tilde{i}_{t+1},\tilde{j}_{t+1})|h_t\right) \\ &= \sum_{k\in K}p_t^k(h_t)G^k(\sigma_{t+1}^1(k,h_t),\sigma_{t+1}^2(h_t)) \\ &\leq \sum_{k\in K}p_t^k(h_t)G^k(\bar{\sigma}_{t+1}^1(h_t),\sigma_{t+1}^2(h_t)) \\ &\quad + M\sum_{k\in K}p_t^k(h_t)\|\sigma_{t+1}^1(k,h_t)-\bar{\sigma}_{t+1}^1(h_t)\| \\ &\leq u(p_t(h_t)) + M\sum_{k\in K}p_t^k(h_t)\|\sigma_{t+1}^1(k,h_t)-\bar{\sigma}_{t+1}^1(h_t)\|, \end{aligned}$$

where $u(p_t(h_t))$ comes from the definition of σ^2. By Lemma 5, we get:

$$\mathbb{E}\left(G^{\tilde{k}}(\tilde{i}_{t+1},\tilde{j}_{t+1})|h_t\right) \leq u(p_t(h_t)) + M\mathbb{E}\left(\|p_{t+1}-p_t\||h_t\right) .$$

Applying Jensen's inequality yields:

$$\mathbb{E}\left(G^{\tilde{k}}(\tilde{i}_{t+1},\tilde{j}_{t+1})\right) \leq \mathrm{cav}u(p) + M\mathbb{E}\left(\|p_{t+1}-p_t\|\right) .$$

We now apply Lemma 4 and obtain:

$$\begin{aligned} \gamma_T^{1,p}(\sigma^1,\sigma^2) &= \mathbb{E}\left(\frac{1}{T}\sum_{t=0}^{T-1}G^{\tilde{k}}(\tilde{i}_{t+1},\tilde{j}_{t+1})\right) \\ &\leq \mathrm{cav}u(p) + \frac{M}{\sqrt{T}}\sum_{k\in K}\sqrt{p^k(1-p^k)} . \end{aligned}$$

This is true for any strategy σ^1 of player 1. Considering the case of an optimal strategy for player 1 in the T-stage game $\Gamma_T(p)$, we have shown:

Proposition 3 *For p in $\Delta(K)$ and $T\geq 1$,*

$$v_T(p) \leq \mathrm{cav}u(p) + \frac{M\sum_{k\in K}\sqrt{p^k(1-p^k)}}{\sqrt{T}} .$$

It remains to conclude about the existence of the uniform value. We have seen that player 1 can guarantee $\mathrm{cav}u(p)$, that player 2 can guarantee $\lim_T v_T(p)$, and we obtain from proposition 3 that $\lim_T v_T(p) \leq \mathrm{cav}u(p)$. This is enough to deduce Aumann and Maschler's celebrated "cav$u$" theorem.

Theorem 1 (Aumann and Maschler [6]) *The game $\Gamma(p)$ has a uniform value which is* $\mathrm{cav}u(p)$.

T-Stage Values and the Recursive Formula

As the T-stage game is a zero-sum game with incomplete information where player 1 is informed, we can write:

$$\begin{aligned} v_T(p) &= \inf_{\sigma^2\in\Sigma^2}\sup_{\sigma^1\in\Sigma^1}\gamma_T^{1,p}(\sigma) , \\ &= \inf_{\sigma^2\in\Sigma^2}\sup_{\sigma^1\in\Sigma^1}\sum_{k\in K}p^k\gamma_T^{1,\delta_k}(\sigma) , \\ &= \inf_{\sigma^2\in\Sigma^2}\sum_{k\in K}p^k\left(\sup_{\sigma^1\in\Sigma^1}\gamma_T^{1,\delta_k}(\sigma)\right) . \end{aligned}$$

This shows that v_T is the infimum of a family of affine functions of p, hence is a concave function of p. This concavity represents the advantage of player 1 to

possess the information on the selected state. Clearly, we have $v_T(p) \geq u(p)$, hence we get the inequalities: $\forall T \geq 1, \text{cav}u(p) \leq v_T(p) \leq \text{cav}u(p) + (M \sum_{k\in K} \sqrt{p^k(1-p^k)})/\sqrt{T}$.

It is also easy to prove that the T-stage value functions satisfy the following recursive formula:

$$\begin{aligned} v_{T+1}(p) &= \frac{1}{T+1} \max_{x\in\Delta(I)^K} \min_{y\in\Delta(J)} \\ &\quad \cdot \left(G(p,x,y) + T \sum_{i\in I} x(p)(i) v_T(\hat{p}(x,i)) \right) \\ &= \frac{1}{T+1} \min_{y\in\Delta(J)} \max_{x\in\Delta(I)^K} \\ &\quad \cdot \left(G(p,x,y) + T \sum_{i\in I} x(p)(i) v_T(\hat{p}(x,i)) \right), \end{aligned}$$

where $x = (x^k(i))_{i\in I, k\in K}$, with x^k the mixed action used at stage 1 by player 1 if the state is k, $G(p,x,y) = \sum_{k,i,j} p^k G^k(x^k(i), y(j))$ is the expected payoff of stage 1, $x(p)(i) = \sum_{k\in K} p^k x^k(i)$ is the probability that action i is played at stage 1, and $\hat{p}(x,i)$ is the conditional probability on K given i.

The recursive formula simply is a generalization of the dynamic programming principle. The following property interprets easily: the advantage of the informed player can only decrease as the number of stages increases (for a proof, one can show that $v_{T+1} \leq v_T$ by induction on T, using the concavity of v_T).

Lemma 6 *The T-stage value $v_T(p)$ is non increasing in T.*

Optimal Strategies

In order to determine the optimal behavior of the players, it is important to be able to compute optimal strategies.

The recursive formula enables to compute, by induction on T, an optimal strategy for player 1 in the T-stage game $\Gamma_T(p)$. But it can not be used to compute an optimal strategy for player 2 in the finitely repeated games, because such a strategy should not depend on player 1's strategy, and consequently on the a posteriori of player 2. Constructing such an optimal strategy for player 2 can be done via the recursive formula of a *dual* game, see Subsect. "Zero-Sum Games".

These strategies of player 2 may afterwards be used to construct an optimal strategy in the infinitely repeated game (see Subsect. "Player 2 can Guarantee the Limit Value"): define consecutive blocks of stages $B^1, \ldots, B^T, \ldots$ of respective cardinalities $1, \ldots, T$, and play independently at each block B^T an optimal strategy for player 2 in the T-stage game $\Gamma_T(p)$. This strategy guarantees $\limsup v_T(p) = \text{cav}u(p)$ for the uninformed player, hence is optimal for player 2, in $\Gamma(p)$. In the next section we will also see how to directly construct an *explicit* optimal strategy for player 2 in $\Gamma(p)$, taking care simultaneously of all possible states k.

It is much more simpler to construct an optimal strategy for player 1 in the infinitely repeated game: since player 1 has to guarantee $\text{cav}u(p)$, this can be done using the concavification lemma, see proposition 1.

Vector Payoffs and Approachability

The following model has been introduced by D. Blackwell [8] and is, strictly speaking, not part of the general definition given in Sect. "Definition of the Subject". We still have a family of $I \times J$ matrices $(G^k)_{k\in K}$, where K is a finite set of parameters. At each stage t, simultaneously player 1 chooses $i_t \in I$ and player 2 chooses $j_t \in J$, and the stage "payoff" is the full vector $G(i_t, j_t) = (G^k(i_t, j_t))_{k\in K}$ in $\mathbb{R}^K$. Notice that there is no initial probability or true state of nature here, and both players have a symmetric role. We assume here that after each stage both players observe exactly the stage vector payoff (but one can check that assuming that the action profiles are observed wouldn't change the results). A natural question is then to determine the sets C in $\mathbb{R}^K$ such that player 1 (for example) can force the average long term payoff to belong to C? Such sets will be called *approachable* by player 1.

In Sect. "Vector Payoffs and Approachability", we use Euclidean distances and norms. Denote by $F = \{(G^k(i,j))_{k\in K}, i \in I, j \in J\}$ the finite set of possible stage payoffs, and by M a constant such that $\|u\| \leq M$ for each u in F. A strategy for player 1, resp. player 2, is an element $\sigma = (\sigma_t)_{t\geq 1}$, where σ_t maps F^{t-1} into $\Delta(I)$, resp. $\Delta(J)$. Strategy spaces for player 1 and 2 are respectively denoted by Σ and $\mathcal{T}$. A strategy profile (σ, τ) naturally induces a unique probability on $(I \times J \times F)^\infty$ denoted by $\mathbb{P}_{\sigma,\tau}$. Let C be a "target" set, that will always be assumed, without loss of generality, a closed subset of $\mathbb{R}^K$. We denote by g_t the random variable, with value in F, of the payoff of stage t, and we use $\bar{g}_t = \frac{1}{t}\sum_{t'=1}^{t} g_{t'} \in \text{conv}(F)$, and finally $d_t = d(\bar{g}_t, C)$ for the distance from $\bar{g}_t$ to C.

Definition 3 C is approachable by player 1 if: $\forall \varepsilon > 0, \exists \sigma \in \Sigma, \exists T, \forall \tau \in \mathcal{T}, \forall t \geq T, \mathbb{E}_{\sigma,\tau}(d_t) \leq \varepsilon$. C is excludable by player 1 if there exist $\delta > 0$ such that $\{z \in \mathbb{R}^K, d(z, C) \geq \delta\}$ is approachable by player 1.

Approachability and excludability for player 2 are defined similarly. C is approachable by player 1 if for each

$\varepsilon > 0$, this player can force that for t large we have $\mathbb{E}_{\sigma,\tau}(d_t) \leq \varepsilon$, so the average payoff will be ε-close to C with high probability. A set cannot be approachable by a player as well as excludable by the other player. In the usual case where K is a singleton, we are in dimension 1 and the Minmax theorem implies that for each t, the interval $[t, +\infty[$ is either approachable by player 1, or excludable by player 2, depending on the comparison between t and the value $\max_{x\in\Delta(I)} \min_{y\in\Delta(J)} G(x, y) = \min_{y\in\Delta(J)} \max_{x\in\Delta(I)} G(x, y)$.

Necessary and Sufficient Conditions for Approachability

Given a mixed action x in $\Delta(I)$, we write xG for the set of possible vector payoffs when player 1 uses x, i. e. $xG = \{G(x, y), y \in \Delta(J)\} = \text{conv}\{\sum_{i\in I} x_i G(i, j), j \in J\}$. Similarly, we write $Gy = \{G(x, y), x \in \Delta(I)\}$ for y in $\Delta(J)$.

Definition 4 The set C is a B(lackwell)-set for player 1 if for every $z \notin C$, there exists $z' \in C$ and $x \in \Delta(I)$ such that: (i) $\|z' - z\| = d(z, C)$, and (ii) the hyperplane containing z' and orthogonal to $[z, z']$ separates z from xG.

For example, any set xG, with x in $\Delta(I)$, is a B-set for player 1. Given a B-set for player 1, we now construct a strategy σ adapted to C as follows. At each positive stage $t + 1$, player 1 considers the current average payoff $\bar{g}_t$. If $\bar{g}_t \in C$, or if $t = 0$, σ plays arbitrarily at stage $t + 1$. Otherwise, σ plays at stage $t + 1$ a mixed action x satisfying the previous definition for $z = \bar{g}_t$.

Theorem 2 *If C is a B-set for player 1, a strategy σ adapted to C satisfies:*

$$\forall \tau \in \mathcal{T}, \forall t \geq 1,\ \mathbb{E}_{\sigma,\tau}(d_t) \leq \frac{2M}{\sqrt{t}} \quad \textit{and} \quad d_t \longrightarrow_{t\to\infty} 0 \mathbb{P}_{\sigma,\tau}\ a.s.$$

As an illustration, in dimension 1 and for $C = \{0\}$, this theorem implies that a bounded sequence $(x_t)_t$ of reals, such that the product $x_{T+1}\left(\frac{1}{T}\sum_{t=1}^{T} x_T\right)$ is non-positive for each T, Cesaro converges to zero.

Proof Assume that player 1 plays σ adapted to C, whereas player 2 plays some strategy τ. Fix $t \geq 1$, and assume that $\bar{g}_t \notin C$. Consider $z' \in C$ and $x \in \Delta(I)$ satisfying (i) and (ii) of definition 4 for $z = \bar{g}_t$. We have:

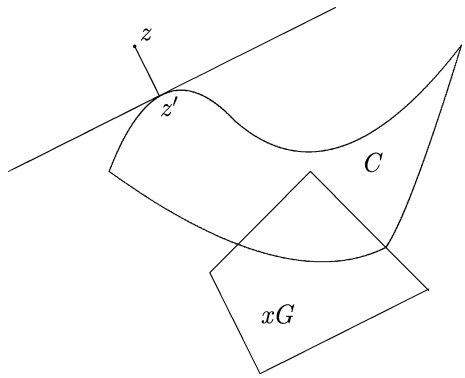

Repeated Games with Incomplete Information, Figure 3
The Blackwell property

$$\begin{aligned} d_{t+1}^2 &= d(\bar{g}_{t+1}, C)^2 \leq \|\bar{g}_{t+1} - z'\|^2 \\ &= \Big\|\frac{1}{t+1}\sum_{l=1}^{t+1} g_l - z'\Big\|^2 \\ &= \Big\|\frac{1}{t+1}(g_{t+1} - z') + \frac{t}{t+1}(\bar{g}_t - z')\Big\|^2 \\ &= \left(\frac{1}{t+1}\right)^2 \|g_{t+1} - z'\|^2 + \left(\frac{t}{t+1}\right)^2 {d_t}^2 \\ &\quad + \frac{2t}{(t+1)^2}\langle g_{t+1} - z', \bar{g}_t - z'\rangle . \end{aligned}$$

By hypothesis, the expectation, given the first t action profiles $h_t \in (I \times J)^t$, of the above scalar product is non-positive, so $\mathbb{E}\left(d_{t+1}^2|h_t\right) \leq (t/(t+1))^2 {d_t}^2 + (1/(t+1))^2 \mathbb{E}\left(\|g_{t+1} - z'\|^2|h_t\right)$. Since $\mathbb{E}\left(\|g_{t+1} - z'\|^2|h_t\right) \leq \mathbb{E}\left(\|g_{t+1} - \bar{g}_t\|^2|h_t\right) \leq (2M)^2$, we have:

$$\mathbb{E}\left(d_{t+1}^2|h_t\right) \leq (\frac{t}{t+1})^2 {d_t}^2 + (\frac{1}{t+1})^2 4M^2 . \tag{2}$$

Taking the expectation, we get, whether $\bar{g}_t \notin C$ or not: $\forall t \geq 1, \mathbb{E}\left(d_{t+1}^2\right) \leq (t/(t+1))^2\mathbb{E}({d_t}^2) + (1/(t+1))^2 4M^2$. By induction, we obtain that for each $t \geq 1, \mathbb{E}(d_t^2) \leq (4M^2)/t$, and $\mathbb{E}(d_t) \leq (2M)/\sqrt{t}$.

Put now, as in Sorin 2002 [85], $e_t = d_t^2 + \sum_{t'>t} \frac{4M^2}{t'^2}$. Inequality (2) gives $\mathbb{E}(e_{t+1}|h_t) \leq e_t$, so (e_t) is a non-negative supermartingale which expectation goes to zero. By a standard probability result (see, e. g., Meyer 1966 [58]), we obtain $e_t \longrightarrow_{t\to\infty} 0\mathbb{P}_{\sigma,\tau}$ a.s., and finally $d_t \longrightarrow_{t\to\infty} 0\mathbb{P}_{\sigma,\tau}$ a.s. □

This theorem implies that any B-set for player 1 is approachable by this player. The converse is true for convex sets.

Theorem 3 *Let C be a closed convex subset of $\mathbb{R}^K$.*

$$\begin{aligned} & (i)\ C \text{ is a B-set for player 1}, \\ \Leftrightarrow\quad & (ii)\ \forall y \in \Delta(J), Gy \cap C \neq \emptyset, \\ \Leftrightarrow\quad & (iii)\ C \text{ is approachable by player 1}, \\ \Leftrightarrow\quad & (iv)\ \forall q \in \mathbb{R}^K, \\ & \max_{x\in\Delta(I)} \min_{y\in\Delta(J)} \sum_{k\in K} q^k G^k(x,y) \geq \inf_{c\in C} \langle q, c\rangle . \end{aligned}$$

Proof The implication $(i) \Longrightarrow (iii)$ comes from Theorem 2. Proof of $(iii) \Longrightarrow (ii)$: assume there exists $y \in \Delta(J)$ such that $Gy \cap C = \emptyset$. Since Gy is approachable by player 2, then C is excludable by player 2 and thus C is not approachable by player 1. Proof of $(ii) \Longrightarrow (i)$: Assume that $Gy \cap C \neq \emptyset \forall y \in \Delta(J)$. Consider $z \notin C$ and define z' as its projection onto C. Define the matrix game where payoffs are projected towards the direction $z' - z$, i. e. the matrix game $\sum_{k\in K}(z'^k - z^k)G^k$. By assumption, one has: $\forall y \in \Delta(J), \exists x \in \Delta(I)$ such that $G(x,y) \in C$, hence such that:

$$\langle z' - z, G(x,y)\rangle \geq \min_{c\in C}\langle z' - z, c\rangle = \langle z' - z, z'\rangle .$$

So $\min_{y\in\Delta(J)} \max_{x\in\Delta(I)} \langle z' - z, G(x,y)\rangle \geq \langle z' - z, z'\rangle$. By the minmax theorem, there exists x in $\Delta(I)$ such that $\forall y \in \Delta(J), \langle z' - z, G(x,y)\rangle \geq \langle z' - z, z'\rangle$, that is $\langle z' - z, z' - G(x,y)\rangle \leq 0$.

(iv) means that any half-space containing C is approachable by player 1. $(iii) \Longrightarrow (iv)$ is thus clear. $(iv) \Longrightarrow (i)$ is similar to $(ii) \Longrightarrow (i)$. □

Up to minor formulation differences, Theorems 2 and 3 are due to Blackwell [8]. More recently, X. Spinat [86] has proved the following characterization.

Theorem 4 *A closed set is approachable for player 1 if and only if it contains a B-set for player 1.*

As a consequence, it shows that adding the condition $d_t \longrightarrow_{t\to\infty} 0 \mathbb{P}_{\sigma,\tau}$ a.s. in the definition of approachability does not modify the notion.

Approachability for Player 1 Versus Excludability for Player 2

As a corollary of Theorem 3, we obtain that: *A closed convex set in $\mathbb{R}^K$ is either approachable by player 1, or excludable by player 2.*

One can show that when K is a singleton, then any set is either approachable by player 1, or excludable by player 2. A simple example of a set which is neither approachable for player 1 nor excludable by player 2 is given in dimension 2 by:

$$G = \begin{pmatrix} (0,0) & (0,0) \\ (1,0) & (1,1) \end{pmatrix}, \quad \text{and}$$
$$C = \{(1/2, v), 0 \leq v \leq 1/4\} \bigcup \{(1, v), 1/4 \leq v \leq 1\}$$

(see [85]).

Weak Approachability

On can weaken the definition of approachability by giving up time uniformity.

Definition 5 C is weakly approachable by player 1 if: $\forall \varepsilon > 0, \exists T, \forall t \geq T, \exists \sigma \in \Sigma, \forall \tau \in \mathcal{T}, \mathbb{E}_{\sigma,\tau}(d_t) \leq \varepsilon$. C is weakly excludable by player 1 if there exists $\delta > 0$ such that $\{z \in \mathbb{R}^K, d(z, C) \geq \delta\}$ is weakly approachable by player 1.

N. Vieille [87] has proved, via the consideration of certain differential games:

Theorem 5 *A subset of $\mathbb{R}^K$ is either weakly approachable by player 1 or weakly excludable by player 2.*

Back to the Standard Model

Let us come back to Aumann and Maschler's model with a finite family of matrices $(G^k)_{k\in K}$, and an initial probability p on $\Delta(K)$. By Theorem 1, the repeated game $\Gamma(p)$ has a uniform value which is cav$u(p)$, and Blackwell approachability will allow for the construction of an *explicit* optimal strategy for the uninformed player. Considering a hyperplane which is tangent to cav$u$ at p, we can find a vector l in $\mathbb{R}^K$ such that

$$\begin{aligned} & \langle l, p\rangle = \text{cav}u(p) \\ & \qquad \text{and} \quad \forall q \in \Delta(K), \langle l, q\rangle \geq \text{cav}u(q) \geq u(q) . \end{aligned}$$

Define now the orthant $C = \{z \in \mathbb{R}^K, z^k \leq l^k \forall k \in K\}$. Recall that player 2 does not know the selected state, and an optimal strategy for him can not depend on player 1' strategy, and consequently on a martingale of a posteriori. He will play in a way such that player 1's long term payoff is, *simultaneously for each k in K*, not greater than l^k if the state is k.

Fix $q = (q^k)_k$ in $\mathbb{R}^K$. If there exists k with $q^k > 0$, we clearly have $\inf_{c\in C}\langle q, c\rangle = -\infty \leq \max_{y\in\Delta(J)} \min_{x\in\Delta(I)} \sum_{k\in K} q^k G^k(x,y)$. Assume now that $q^k \leq 0$ for

each k, with $q \neq 0$. Write $s = \sum_k(-q^k)$.

$$\begin{aligned}\inf_{c\in C}\langle q, c\rangle &= \sum_{k\in K} q^k l^k \\ &= -s\left\langle l, \frac{-q}{s}\right\rangle \\ &\leq -su(\frac{-q}{s}) \\ &\leq -s \max_{x\in\Delta(I)} \min_{y\in\Delta(J)} \sum_{k\in K} \frac{-q^k}{s} G^k(x,y) \\ &= \max_{y\in\Delta(J)} \min_{x\in\Delta(I)} \sum_{k\in K} q^k G^k(x,y)\end{aligned}$$

This is condition (iv) of Theorem 3, adapted to player 2. So C is a B-set for player 2, and a strategy τ adapted to C satisfies by Theorem 2: $\forall\sigma \in \Sigma, \forall k \in K$,

$$\begin{aligned}&\mathbb{E}_{\sigma,\tau}\left(\frac{1}{T}\sum_{t=1}^{T} G^k(\tilde{i}_t, \tilde{j}_t) - l^k\right) \\ &\quad \leq \mathbb{E}_{\sigma,\tau}\left(d\left(\frac{1}{T}\sum_{t=1}^{T} G^k(\tilde{i}_t, \tilde{j}_t), C\right)\right) \leq \frac{2M}{\sqrt{T}},\end{aligned}$$

(where M is here an upper bound for the Euclidean norms of the vectors $(G^k(i,j))_{k\in K}$, with $i \in I$ and $j \in J$.) And this holds as well for any strategy σ of player 1 in the repeated game with incomplete information. So for any such strategy σ,

$$\begin{aligned}\gamma_T^{1,p}(\sigma,\tau) &= \sum_{k\in K} p^k \left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\sigma,\tau}(G^k(\tilde{i}_t, \tilde{j}_t))\right) \\ &\leq \langle p, l\rangle + \frac{2M}{\sqrt{T}} = \mathrm{cav}u(p) + \frac{2M}{\sqrt{T}}.\end{aligned}$$

As shown by Kohlberg [35], the approachability strategy τ is thus an optimal strategy for player 2 in the repeated game $\Gamma(p)$.

Zero-Sum Games with Lack of Information on Both Sides

The following model has also been introduced by Aumann and Maschler [6]. We are still in the context of zero-sum repeated games with observable actions, but it is no longer assumed that one of the players is fully informed. The set of states is here a product $K \times L$ of finite sets, and we have a family of matrices $(G^{k,l})_{(k,l)\in K\times L}$ with size $I \times J$, as well as initial probabilities p on K, and q on L. In the game $\Gamma(p,q)$, a state of nature (k,l) is first selected according to the product probability $p \otimes q$, then k, resp. l, is announced to player 1, resp. player 2 only. Then the matrix game $G^{k,l}$ is repeated over and over: at every stage, simultaneously player 1 chooses a row i in I, whereas player 2 chooses a column j in J, the stage payoff for player 1 is $G^{k,l}(i,j)$ but only i and j are publicly announced before proceeding to the next stage.

The average payoff for player 1 in the T-stage game is written: $\gamma_T^{1,p,q}(\sigma^1,\sigma^2) = \mathbb{E}^{p,q}_{\sigma^1,\sigma^2}\left(\frac{1}{T}\sum_{t=1}^{T} G^{\tilde{k},\tilde{l}}(\tilde{i}_t,\tilde{j}_t)\right)$, and the T-stage value is written $v_T(p,q)$. Similarly, the λ-discounted value of the game will be written $v_\lambda(p,q)$

The non revealing game now corresponds to the case where player 1 plays independently of k and player 2 plays independently of l. Its value is denoted by:

$$u(p,q) = \max_{x\in\Delta(I)} \min_{y\in\Delta(J)} \sum_{k,l} p^k q^l G^{k,l}(x,y). \tag{3}$$

Given a continuous function $f\colon \Delta(K)\times\Delta(L) \longrightarrow \mathbb{R}$, we denote by $\mathrm{cav}_{\mathrm{I}} f$ the concavification of f with respect to the first variable: for each (p,q) in $\Delta(K)\times\Delta(L)$, $\mathrm{cav}_{\mathrm{I}} f(p,q)$ is the value at p of the smallest concave function from $\Delta(K)$ to $\mathbb{R}$ which is above $f(.,q)$. Similarly, we denote by $\mathrm{vex}_{\mathrm{II}} f$ the convexification of f with respect to the second variable. It can be shown that $\mathrm{cav}_{\mathrm{I}} f$ and $\mathrm{vex}_{\mathrm{II}} f$ are continuous, and we can compose $\mathrm{cav}_{\mathrm{I}}\mathrm{vex}_{\mathrm{II}} f$ and $\mathrm{vex}_{\mathrm{II}}\mathrm{cav}_{\mathrm{I}} f$. These functions are both concave in the first variable and convex in the second variable, and they satisfy $\mathrm{cav}_{\mathrm{I}}\mathrm{vex}_{\mathrm{II}} f(p,q) \leq \mathrm{vex}_{\mathrm{II}}\mathrm{cav}_{\mathrm{I}} f(p,q)$.

Maxmin and Minmax of the Repeated Game

Theorem 1 generalizes as follows.

Theorem 6 ([6]) *In the repeated game $\Gamma(p,q)$, the greatest quantity which can be guaranteed by player 1 is $\mathrm{cav}_{\mathrm{I}}\mathrm{vex}_{\mathrm{II}} u(p,q)$, and the smallest quantity which can be guaranteed by player 2 is $\mathrm{vex}_{\mathrm{II}}\mathrm{cav}_{\mathrm{I}} u(p,q)$.*

Aumann, Maschler and Stearns also showed that $\mathrm{cav}_{\mathrm{I}}\mathrm{vex}_{\mathrm{II}} u(p,q)$ can be *defended* by player 2, uniformly in time, i. e. that $\forall\varepsilon > 0, \forall\sigma^1, \exists T_0, \exists\sigma^2, \forall T \geq T_0, \gamma_T^{p,q}(\sigma^1,\sigma^2) \leq \mathrm{cav}_{\mathrm{I}}\mathrm{vex}_{\mathrm{II}} u(p,q)v + \varepsilon$. Similarly, $\mathrm{vex}_{\mathrm{II}}\mathrm{cav}_{\mathrm{I}} u(p,q)$ can be *defended* by player 1.

The proof uses the martingales of a posteriori of each player, and a useful notion is that of the informational content of a strategy: for a strategy σ^1 of the first player, it is defined as: $I(\sigma^1) = \sup_{\sigma^2} \mathbb{E}^{p,q}_{\sigma^1,\sigma^2}\left(\sum_{k\in K}\sum_{t=0}^{\infty} (p^k_{t+1}(\sigma^1) - p^k_t(\sigma^1))^2\right)$, where $p_t(\sigma^1)$ is the a posteriori on K of player 2 after stage t given that player 1 uses σ^1. By linearity of the expectation, the supremum can be re-

stricted to strategies of player 2 which are both pure and independent of l.

Theorem 6 implies that $\mathrm{cav_I vex_{II}} u(p,q) = \sup_{\sigma^1 \in \Sigma^1} \liminf_T \left(\inf_{\sigma^2 \in \Sigma^2} \gamma_T^{1,p,q}(\sigma^1, \sigma^2) \right)$, and $\mathrm{cav_I vex_{II}} u(p,q)$ is called the *maxmin* of the repeated game $\Gamma(p,q)$. Similarly, $\mathrm{vex_{II} cav_I} u(p,q) = \inf_{\sigma^2 \in \Sigma^2} \limsup_T \left(\sup_{\sigma^1 \in \Sigma^1} \gamma_T^1(\sigma^1, \sigma^2) \right)$ is called the *minmax* of $\Gamma(p,q)$. As a corollary, we obtain that the repeated game $\Gamma(p,q)$ has a uniform value if and only if: $\mathrm{cav_I vex_{II}} u(p,q) = \mathrm{vex_{II}}\, \mathrm{cav_I} u(p,q)$. This is not always the case, and there exist counter-examples to the existence of the uniform value.

Example 4 $K = \{a, a'\}$, and $L = \{b, b'\}$, with p and q uniform.

$$G^{a,b} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ -1 & 1 & 1 & -1 \end{pmatrix} \quad G^{a,b'} = \begin{pmatrix} 1 & -1 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$G^{a',b} = \begin{pmatrix} -1 & 1 & -1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad G^{a',b'} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

Mertens and Zamir [52] have shown that here, $\mathrm{cav_I vex_{II}} u(p,q) = -\frac{1}{4} < 0 = \mathrm{vex_{II} cav_I} u(p,q)$.

Limit Values

It is easy to see that for each T and λ, the value functions v_T and v_λ are concave in the first variable, and convex in the second variable. They are all Lipschitz functions, with the same constant $M = \max_{i,j,k,l} |G^{k,l}(i,j)|$, and here also, recursive formula can be given. In the following result, v_T and v_λ are viewed as elements of the set C of continuous mappings from $\Delta(K) \times \Delta(L)$ to $\mathbb{R}$.

Theorem 7 (Mertens and Zamir [52]) *$(v_T)_T$, as T goes to infinity, and $(v_\lambda)_\lambda$, as λ goes to zero, both uniformly converge to the unique solution f of the following system:*

$$\begin{cases} f &= vex_{II} \max\{u, f\} \\ f &= cav_I \min\{u, f\} \end{cases}$$

Besides, the above system can also be fruitfully studied without reference to repeated games (see [39,40,55,81]). For a proof of Theorem 7, one can also see Zamir [91] or Sorin [85]. Mertens and Zamir notably consider responses of a player, to a given strategy of his opponent, which are of the following type: play non revealing up to a particular stopping time, and then start using the information by playing optimally in the remaining subgame.

Remark Let U be the set of all non revealing value functions, i. e. of functions from $\Delta(K) \times \Delta(L)$ to $\mathbb{R}$ satisfying Eq. (3) for some family of matrices $(G^{k,l})_{k,l}$. One can easily show that any mapping in C is a uniform limit of elements in U.

Correlated Initial Information

A more general model can be written, where it is no longer assumed that the initial information of the players are independent. The set of states is now denoted by R (instead of $K \times L$), initially a state r in R is chosen according to a known probability $p = (p^r)_{r \in R}$, and each player receives a deterministic signal depending on r. Equivalently, each player i has a partition R^i of R and observes the element of his partition which contains the selected state.

After the first stage, player 1 will play an action $x = (x^r)_{r \in R}$ which is measurable with respect to R^1, i. e. $(r \longrightarrow x^r)$ is constant on each atom of R^1. After having observed player 1's action at the first stage, the conditional probability on R necessarily belongs to the set:

$$\Pi^{\mathrm{I}}(p) = \left\{ (\alpha^r p^r)_{r \in R}, \forall r \alpha^r \geq 0, \sum_r \alpha^r p^r = 1 \right.$$
$$\left. \text{and } (\alpha^r)_r \text{ is } R^1\text{-measurable} \right\}.$$

$\Pi^{\mathrm{I}}(p)$ contains p, and is a convex compact subset of $\Delta(R)$. A mapping f from $\Delta(R)$ to $\mathbb{R}$ is now said to be I-concave if for each p in $\Delta(R)$, the restriction of f to $\Pi^{\mathrm{I}}(p)$ is concave. And given $g\colon \Delta(R) \longrightarrow \mathbb{R}$ which is bounded from above, we define the concavification $\mathrm{cav_I} g$ as the smallest function above g which is I-concave. Similarly one can define the set $\Pi^{\mathrm{II}}(p)$ and the notions of II-convexity and II-convexification. With these generalized definitions, the results of Theorem 6 and 7 perfectly extend [52].

Non Zero-Sum Games with Lack of Information on One Side

We now consider the generalization of the standard model of Sect. "The Standard Model of Aumann and Maschler" to the non-zero sum case. Hence two players infinitely repeat the same bimatrix game, with player 1 only knowing the bimatrix. Formally, we have a finite set of states K, an initial probability p on K, and families of $I \times J$-payoff matrices $(A^k)_{k \in K}$ and $(B^k)_{k \in K}$. Initially, a state k in K is selected according to p, and announced to player 1 only. Then the bimatrix game (A^k, B^k) is repeated over and over: at every stage, simultaneously player 1 chooses a row i in I, whereas player 2 chooses a column j in J, the stage payoff for player 1 is then $A^k(i,j)$, the stage payoff for player 2 is $B^k(i,j)$, but only i and j are publicly announced before proceeding to the next stage. Without loss of generality, we assume that $p^k > 0$ for each k, and that each player has at least two actions.

Given a strategy pair (σ^1, σ^2), it is here convenient to denote the expected payoffs up to stage T by:

$$\alpha_T^p(\sigma^1, \sigma^2) = \mathbb{E}_{p,\sigma^1,\sigma^2}\left(\frac{1}{T}\sum_{t=1}^{T} A^{\tilde{k}}(\tilde{i}_t, \tilde{j}_t)\right) = \sum_{k\in K} p^k \alpha_T^k(\sigma^1, \sigma^2)\,.$$

$$\beta_T^p(\sigma^1, \sigma^2) = \mathbb{E}_{p,\sigma^1,\sigma^2}\left(\frac{1}{T}\sum_{t=1}^{T} B^{\tilde{k}}(\tilde{i}_t, \tilde{j}_t)\right) = \sum_{k\in K} p^k \beta_T^k(\sigma^1, \sigma^2)\,.$$

Given a probability q on K, we write $A(q) = \sum_k q^k A^k$, $B(q) = \sum_k q^k B^k$, $u(q) = \max_{x\in\Delta(I)} \min_{y\in\Delta(J)} A(q)(x, y)$ and $v(q) = \max_{y\in\Delta(J)} \min_{x\in\Delta(I)} B(q)(x, y)$. If $\gamma = (\gamma(i, j))_{(i,j)\in I\times J} \in \Delta(I \times J)$, we put $A(q)(\gamma) = \sum_{(i,j)\in I\times J} \gamma(i, j)A(q)(i, j)$ and similarly $B(q)(\gamma) = \sum_{(i,j)\in I\times J} \gamma(i, j) B(q)(i, j)$.

Existence of Equilibria

The question of existence of an equilibrium has remained unsolved for long. Sorin [79] proved the existence of an equilibrium for two states of nature, and the general case has been solved by Simon et al. [76].

Exactly as in the zero-sum case, a strategy pair σ induces a sequence of a posteriori $(p_t(\sigma))_{t\geq 0}$ which is a $\mathbb{P}_{p,\sigma}$- martingale with values in $\Delta(K)$. We will concentrate on the cases where this martingale moves only once.

Definition 6 A joint plan is a triple (S, λ, γ), where:

- S is a finite non empty set (of messages),
- $\lambda = (\lambda^k)_{k\in K}$ (signaling strategy) with for each $k, \lambda^k \in \Delta(S)$ and for each $s, \lambda_s =_{\text{def}} \sum_{k\in K} p^k \lambda_s^k > 0$,
- $\gamma = (\gamma_s)_{s\in S}$ (contract) with for each $s, \gamma_s \in \Delta(I \times J)$.

The idea is due to Aumann, Maschler and Stearns. Player 1 observes k, then chooses $s \in S$ according to λ^k and announces s to player 2. Then the players play pure actions corresponding to the frequencies $\gamma_s(i, j)$, for i in I and j in J. Given a joint plan (S, λ, γ), we define:

- $\forall s \in S, p_s = (p_s^k)_{k\in K} \in \Delta(K)$, with $p_s^k = \frac{p^k\lambda_s^k}{\lambda_s}$ for each k. p_s is the a posteriori on K given s.
- $\varphi = (\varphi^k)_{k\in K} \in \mathbb{R}^K$, with for each k, $\varphi^k = \max_{s\in S} A^k(\gamma_s)$.
- $\forall s \in S, \psi_s = B(p_s)(\gamma_s)$ and $\psi = \sum_{k\in K} p^k \sum_{s\in S} \lambda_s^k B^k(\gamma_s) = \sum_{s\in S} \lambda_s\psi_s$.

Definition 7 A joint plan (S, λ, γ) is an equilibrium joint plan if:

(i) $\forall s \in S, \psi_s \geq \text{vex}v(p_s)$,
(ii) $\forall k \in K, \forall s \in S$ s.t. $p_s^k > 0, A^k(\gamma_s) = \varphi^k$, and
(iii) $\forall q \in \Delta(K), \langle\varphi, q\rangle \geq u(q)$.

Condition (ii) can be seen as an incentive condition for player 1 to choose s according to λ^k. Given an equilibrium joint plan (S, λ, γ), one define a strategy pair $(\sigma^{1*}, \sigma^{2*})$ adapted to it. For each message s, first fix a sequence $(i_t^s, j_t^s)_{t\geq 1}$ of elements in $I \times J$ such that for each (i, j), the empirical frequencies converge to the corresponding probability: $\frac{1}{T}\,|\{t, 1 \leq t \leq T, (i_t^s, j_t^s) = (i, j)\}| \longrightarrow_{T\to\infty} \gamma_s(i, j)$. We also fix an injective mapping f from S to I^l, where l is large enough, corresponding to a code between the players to announce an element in S. σ^{1*} is precisely defined as follows. Player 1 observes the selected state k, then chooses s according to λ^k, and announces s to player 2 by playing $f(s)$ at the first l stages. Finally, σ^{1*} plays i_t^s at each stage $t > l$ as long as player 2 plays j_t^s. If at some stage $t > l$ player 2 does not play j_t^s then player 1 punishes his opponent by playing an optimal strategy in the zero-sum game with initial probability p_s and payoffs for player 1 given by $(-B^k)_{k\in K}$. We now define σ^{2*}. Player 2 arbitrarily plays at the beginning of the game, then compute at the end of stage l the message s sent by player 1. Next he plays at each stage $t > l$ the action j_t^s as long as player 1 plays i_t^s. If at some stage $t > l$, player 1 does not play i_t^s, or if the first l actions of player 1 correspond to no message, then player 2 plays a punishing strategy $\bar{\sigma}^2$ such that: $\forall\varepsilon > 0, \exists T_0, \forall T \geq T_0, \forall\sigma^1 \in \Sigma^1, \forall k \in K, \alpha_T^k(\sigma^1, \bar{\sigma}^2) \leq \varphi^k + \varepsilon$. Such a strategy $\bar{\sigma}^2$ exists because of condition (iii): it is an approachability strategy for player 2 of the orthant $\{x \in \mathbb{R}^K, \forall k \in K\, x^k \leq \varphi^k\}$ (see Subsect. "Back to the Standard Model").

Lemma 7 ([79]) *A strategy pair adapted to an equilibrium joint plan is a uniform equilibrium of the repeated game.*

Proof The payoffs induced by $(\sigma^{1*}, \sigma^{2*})$ can be easily computed: $\forall k, \alpha_T^k(\sigma^{1*}, \sigma^{2*}) \longrightarrow_{T\to\infty} \sum_{s\in S} \lambda_s^k A^k(\gamma_s) = \varphi^k$ because of (*ii*), and $\beta_T^p(\sigma^{1*}, \sigma^{2*}) \longrightarrow_{T\to\infty} \sum_{k\in K} p^k \sum_{s\in S} \lambda_s^k B^k(\gamma_s) = \psi$. Assume that player 2 plays σ^{2*}. The existence of $\bar{\sigma}^2$ implies that no detectable deviation of player 1 is profitable, so if the state is k, player 1 will gain no more than $\max_{s'\in S} A^k(\gamma_{s'})$. But this is just φ^k. The proof can be made uniform in σ^1 and we obtain: $\forall\varepsilon > 0\ \exists T_0\ \forall T \geq T_0, \forall k \in K, \forall\sigma^1 \in \Sigma^1, \alpha_T^k(\sigma^1, \sigma^{2*}) \leq \varphi^k + \varepsilon$. Finally assume that player 1 plays σ^{1*}. Condition (*i*) implies that if player 2 uses σ^{2*},

the payoff of this player will be at least $\text{vex}v(p_s)$ if the message is s. Since $\text{vex}\, v(p_s)(= -\text{cav}(-v(p_s)))$ is the value, from the point of view of player 2 with payoffs $(B^k)_k$, of the zero-sum game with initial probability p_s, player 2 fears the punition by player 1, and $\forall\varepsilon > 0, \exists T_0, \forall T \geq T_0, \forall\sigma^2 \in \Sigma^2, \beta_T^p(\sigma^{1*}, \sigma^2) \leq \sum_{s\in S} \lambda_s \psi_s + \varepsilon = \psi + \varepsilon$. □

To prove the existence of equilibria, we then look for equilibrium joint plans. The first idea is to consider, for each probability r on K, the set of payoff vectors φ compatible with r being an a posteriori. This leads to the consideration of the following correspondence (for each $r, \Phi(r)$ is a subset of $\mathbb{R}^K$):

$$\begin{aligned}\Phi\colon \Delta(K) &\rightrightarrows \mathbb{R}^K\\ r &\mapsto \{(A^k(\gamma))_{k\in K}\,, \quad \text{where } \gamma \in \Delta(I\times J)\\ &\qquad \text{satisfies } B(r)(\gamma) \geq \text{vex}\, v(r)\}\end{aligned}$$

It is easy to see that the graph of Φ, i. e. the set $\{(r,\varphi) \in \Delta(K)\times\mathbb{R}^K, \varphi \in \Phi(r)\}$, is compact, that Φ has non empty convex values, and satisfies: $\forall r \in \Delta(K), \forall q \in \Delta(K), \exists\varphi \in \Phi(r), \langle\varphi, q\rangle \geq u(q)$.

Assume now that one can find a finite family $(p_s)_{s\in S}$ of probabilities on K, as well as vectors φ and, for each s, φ_s in $\mathbb{R}^K$ such that: 1) $p \in \text{conv}\{p_s, s \in S\}$, 2) $\langle\varphi, q\rangle \geq u(q)\forall q \in \Delta(K)$, 3) $\forall s \in S, \varphi_s \in \Phi(p_s)$, and 4) $\forall s \in S, \forall k \in K, \varphi_s^k \leq \varphi^k$ with equality if $p_s^k > 0$. It is then easy to construct an equilibrium joint plan. Thus we get interested in proving the following result.

Proposition 4 *Let p be in $\Delta(K)$, $u\colon \Delta(K) \longrightarrow \mathbb{R}$ be a continuous mapping, and $\Phi\colon \Delta(K) \rightrightarrows \mathbb{R}^K$ be a correspondence with compact graph and non empty convex values such that: $\forall r \in \Delta(K)$, $\forall q \in \Delta(K)$, $\exists\varphi \in \Phi(r), \langle\varphi, q\rangle \geq u(q)$. Then there exists a finite family $(p_s)_{s\in S}$ of elements of $\Delta(K)$, as well as vectors φ and, for each s, φ_s in $\mathbb{R}^K$ such that:*

- $p \in \text{conv}\{p_s, s \in S\}$,
- $\langle\varphi, q\rangle \geq u(q)\forall q \in \Delta(K)$,
- $\forall s \in S, \varphi_s \in \Phi(p_s)$,
- $\forall s \in S, \forall k \in K, \varphi_s^k \leq \varphi^k$ *with equality if* $p_s^k > 0$.

The proof of proposition 4 relies, as explained in [62] or [75], on a fixed point theorem of Borsuk–Ulam type proved by Simon, Spież and Toruńczyk [76] via tools from algebraic geometry. A simplified version of this fixed point theorem can be written as follows:

Theorem 8 ([76]) *Let C be a compact subset of an n-dimensional Euclidean space, $x \in C$ and Y be a finite union of affine subspaces of dimension $n-1$ of an Euclidean space. Let F be a correspondence from C to Y with compact graph and non empty convex values. Then there exists $L \subset \partial C$ and $y \in Y$ such that: $\forall l \in L, y \in F(l)$ and $x \in \text{conv}(L)$.*

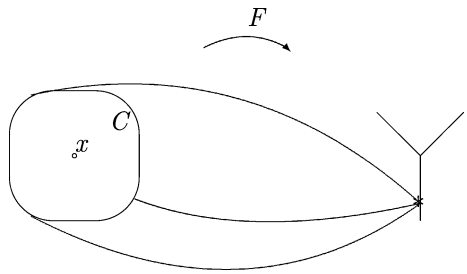

Repeated Games with Incomplete Information, Figure 4
A Borsuk–Ulam type theorem by Simon, Spież and Toruńczyk

Notice that for $n = 1$ (corresponding to 2 states of nature), the image by F of the connected component of C containing x necessarily is a singleton, hence the result is clear. In the general case, one finally obtains:

Theorem 9 ([76]) *There exists an equilibrium joint plan. Thus there exists a uniform equilibrium in the repeated game $\Gamma(p)$.*

Characterization of Equilibrium Payoffs

Characterizing equilibrium payoffs, as the Folk theorem does for repeated games with complete information, has been a challenging problem. We denote here by p_0 the initial probability in the interior of $\Delta(K)$. We are interested in the set of equilibrium payoffs, in the convenient following sense:

Definition 8 A vector (a, b) in $\mathbb{R}^K \times \mathbb{R}$ is called an equilibrium payoff of the repeated game $\Gamma(p_0)$ if there exists a strategy pair $(\sigma^{1*}, \sigma^{2*})$ satisfying:

(i) $\forall\varepsilon > 0\ \exists T_0\ \forall T \geq T_0, \forall k \in K, \forall\sigma^1 \in \Sigma^1, \alpha_T^k(\sigma^1, \sigma^{2*}) \leq \alpha_T^k(\sigma^{1*}, \sigma^{2*}) + \varepsilon$, and $\forall\varepsilon > 0\ \exists T_0 \forall T \geq T_0, \forall\sigma^2 \in \Sigma^2, \beta_T^{p_0}(\sigma^{1*}, \sigma^2) \leq \beta_T^p(\sigma^{1*}, \sigma^{2*}) + \varepsilon$, and
(ii) $(\alpha_T^k(\sigma^{1*}, \sigma^{2*}))_{k,T}$ and $(\beta_T^{p_0}(\sigma^{1*}, \sigma^{2*}))_T$ respectively converge to a and b.

Since p lies in the interior of $\Delta(K)$, the first line of (i) is equivalent to: $\forall\varepsilon > 0\ \exists T_0\ \forall T \geq T_0, \forall\sigma^1 \in \Sigma^1, \alpha_T^p(\sigma^1, \sigma^{2*}) \leq \alpha_T^p(\sigma^{1*}, \sigma^{2*}) + \varepsilon$. The strategy pair $(\sigma^{1*}, \sigma^{2*})$ is thus a uniform equilibrium of the repeated game, with the additional requirement that expected average payoffs of player 1 converge in each state k. In some sense, player 1

is viewed here as $|K|$ different types or players, and we require the existence of the limit payoff of each type. We will only consider such uniform equilibria in the sequel.

Notice that the above definition implies: $\forall k \in K$, $\forall \varepsilon > 0, \exists T_0, \forall T \geq T_0, \forall \sigma^1 \in \Sigma^1, \alpha_T^k(\sigma^1, \sigma^{2*}) \leq a^k + \varepsilon$. So the orthant $\{x \in \mathbb{R}^K, x^k \leq a^k \forall k \in K\}$ is approachable by player 2, and by Theorem 3 and Subsect. "Back to the Standard Model" one can obtain that:

$$\langle a, q \rangle \geq u(q) \quad \forall q \in \Delta(K) \tag{4}$$

Condition (4) is called the individual rationality condition for player 1, and does not depend on the initial probability in the interior of $\Delta(K)$. Regarding player 2, we have: $\forall \varepsilon > 0\ \exists T_0\ \forall T \geq T_0, \forall \sigma^2 \in \Sigma^2, \beta_T^{p_0}(\sigma^{1*}, \sigma^2) \leq \beta + \varepsilon$, so by Theorem 1:

$$\beta \geq \operatorname{vex} v(p_0)\,. \tag{5}$$

Condition (5) is the individual rationality condition for player 2: at equilibrium, this player should have at least the value of the game where player 1's plays in order to minimize player 2's payoffs.

Imagine now that σ^{1*} is a non revealing strategy for player 1, and that the players play actions with empirical frequencies corresponding to a given probability distribution $\pi = (\pi_{i,j})_{(i,j)\in I\times J} \in \Delta(I \times J)$. We will have: $\forall k \in K$, $a^k = \sum_{i,j} \pi_{i,j} A^k(i,j)$ and $\beta = \sum_k p_0^k \sum_{i,j} \pi_{i,j} B^k(i,j)$, and if the individual rationality conditions are satisfied, no detectable deviation of a player can be profitable. This leads to the definition of the following set, where M is the constant $\max\{|A^k(i,j)|, |B^k(i,j)|, (i,j) \in I \times J\}$, and $\mathbb{R}_M = [-M, M]$.

Definition 9 Let G be the set of triples $(a, \beta, p) \in \mathbb{R}_M^K \times \mathbb{R}_M \times \Delta(K)$ satisfying:

1. $\forall q \in \Delta(K), \langle a, q \rangle \geq u(q)$,
2. $\beta \geq \operatorname{vex} v(p)$,
3. $\exists \pi \in \Delta(I \times J)$ s.t. $\beta = \sum_k p^k \sum_{i,j} \pi_{i,j} B^k(i,j)$ and $\forall k \in K, a^k \geq \sum_{i,j} \pi_{i,j} A^k(i,j)$ with equality if $p^k > 0$.

We need to considerate every possible initial probability because the main state variable of the model is, here also, the belief, or a posteriori, of player 2 on the state of nature. $\{(a, \beta), (a, \beta, p_0) \in G\}$ is the set of payoffs of non revealing equilibria of $\Gamma(p_0)$. The importance of the following definition will appear with Theorem 10 below (which unfortunately has not led to a proof of existence of equilibrium payoffs).

Definition 10 G^* is defined as the set of elements $g = (a, \beta, p) \in \mathbb{R}_M^K \times \mathbb{R}_M \times \Delta(K)$ such that there exist a probability space $(\Omega, \mathcal{A}, Q)$, an increasing sequence $(\mathcal{F}_n)_{n\geq 1}$ of finite sub-σ-algebras of $\mathcal{A}$, and a sequence of random variables $(g_n)_{n\geq 1} = (a_n, \beta_n, p_n)_{n\geq 1}$ defined on $(\Omega, \mathcal{A})$ with values in $\mathbb{R}_M^K \times \mathbb{R}_M \times \Delta(K)$ satisfying: (i) $g_1 = g$ a.s., (ii) $(g_n)_{n\geq 1}$ is a martingale adapted to $(\mathcal{F}_n)_{n\geq 1}$, (iii) $\forall n \geq 1, a_{n+1} = a_n$ a.s. or $p_{n+1} = p_n$ a.s., and (iv) $(g_n)_n$ converges a.s. to a random variable g_∞ with values in G.

Let us forget for a while the component of player 2's payoff. A process $(g_n)_n$ satisfying (ii) and (iii) may be called a bi-martingale, it is a martingale such that at every stage, one of the two components remains a.s. constant. So the set G^* can be seen as the set of starting points of converging bi-martingales with limit points in G.

Theorem 10 (Hart [29]) *Let (a, β) be in $\mathbb{R}^K \times \mathbb{R}$.*

$$(a, \beta) \text{ is an equilibrium payoff of } \Gamma(p_0) \iff (a, \beta, p_0) \in G^*\,.$$

Theorem 10 is too elaborate to be proved here, but let us give a few ideas about the proof. First consider the implication $\Longrightarrow$, and fix an equilibrium $\sigma^* = (\sigma^{1*}, \sigma^{2*})$ of $\Gamma(p_0)$ with payoff (a, β). The sequence of a posteriori $(p_t(\sigma^*))_{t\geq 0}$ is a $\mathbb{P}_{p_0,\sigma^*}$- martingale. Modify now slightly the time structure so that at each stage, player 1 plays first, and then player 2 plays without knowing the action chosen by player 1. At each half-stage where player 2 plays, his a posteriori remains constant. At each half-stage where player 1 plays, the "expectation of player 1's future payoff" (which can be properly defined) remains constant. Hence, the heuristic apparition of the bimartingale. And since bounded martingale converge, for large stages everything will be fixed and the players will approximately play a non revealing equilibrium at a "limit a posteriori", so the convergence will be towards elements of G.

Consider now the converse implication $\Longleftarrow$. Let (a, β) be such that $(a, \beta, p_0) \in G^*$, and assume for simplification that the associated bi-martingale (a_n, β_n, p_n) converges in a fixed number N of stages: $\forall n \geq N, (a_n, \beta_n, p_n) = (a_N, \beta_N, p_N) \in G$. One can construct an equilibrium $(\sigma^{1*}, \sigma^{2*})$ of $\Gamma(p_0)$ with payoff (a, β) along the following lines. For each index n, (a_n, β_n) will be an equilibrium payoff of the repeated game with initial probability p_n. Eventually, player 1 will play independently of the state, the a posteriori of player 2 will be p_N, and the players will end up playing a non revealing equilibrium of the repeated game $\Gamma(p_N)$ with payoff (a_N, β_N). What should be played before? Since we are in an undiscounted setup, any finite number of stages can be used for communication without

influencing payoffs. Let $n < N$ be such that $a_{n+1} = a_n$. To move from (a_n, β_n, p_n) to $(a_n, \beta_{n+1}, p_{n+1})$, player 1 can simply use the splitting lemma (Lemma 1) in order to signal part of the state to player 2. Let now $n < N$ be such that $p_{n+1} = p_n$, so that we want to move from (a_n, β_n, p_n) to $(a_{n+1}, \beta_{n+1}, p_n)$. Player 1 will play independently of the state, and both players will act so as to convexify their future payoffs. This convexification is done through procedures called "jointly controlled lotteries" and introduced in the sixties by Aumann and Maschler [6], with the following simple and brilliant idea. Imagine that the players have to decide with even probability whether to play the equilibrium $E1$ with payoff (a^1, β^1) or to play the equilibrium $E2$ with payoff (a^2, β^2). The players may not be indifferent between $E1$ and $E2$, e. g. player 1 may prefer $E1$ whereas player 2 prefers $E2$. They will proceed as follows, with i and i', respectively j and j', denoting two distinct actions of player 1, resp. player 2. Simultaneously and independently, player 1 will select i or i' with probability 1/2, whereas player 2 will behave similarly with j and j'.

$$\begin{array}{cc} & \begin{array}{cc} j & j' \end{array} \\ \begin{array}{c} i \\ i' \end{array} & \begin{pmatrix} \times & \\ & \times \end{pmatrix} \end{array} .$$

Then the equilibrium $E1$ will be played if the diagonal has been reached, i. e. if (i, j) or (i', j') has been played, and otherwise the equilibrium $E2$ will be played. This procedure is robust to unilateral deviations: none of the players can deviate and prevent $E1$ and $E2$ to be chosen with probability 1/2. In general, jointly controlled lotteries are procedures allowing to select an alternative among a finite set according to a given probability (think of binary expansions if necessary), in a way which is robust to deviations by a single player. S. Hart has precisely shown how to combine steps of signaling and jointly controlled lotteries to construct an equilibrium of $\Gamma_\infty(p_0)$ with payoff (a, β).

Biconvexity and Bimartingales

The previous analysis has lead to the introduction and study of biconvexity phenomena. The reference here is [4]. Let X and Y be compact convex subsets of Euclidean spaces, and let $(\Omega, \mathcal{F}, \mathcal{P})$ be an atomless probability space.

Definition 11 A subset B of $X \times Y$ is biconvex if for every x in X and y in Y, the sections $B_{x.} = \{y' \in Y, (x, y') \in B\}$ and $B_{.y} = \{x' \in X, (x', y) \in B\}$ are convex. If B is biconvex, a mapping $f\colon B \longrightarrow \mathbb{R}$ is called biconvex if for each $(x, y) \in X \times Y$, $f(., y)$ and $f(x, .)$ are convex.

As in the usual convexity case, we have that if f is biconvex, then for each α in $\mathbb{R}$, the set $\{(x, y) \in B, f(x, y) \leq \alpha\}$ is biconvex.

Definition 12 A sequence of random variables $Z_n = (X_n, Y_n)_{n\geq 1}$ with values in $X \times Y$ is called a bimartingale if:

(1) there exists an increasing sequence $(\mathcal{F}_n)_{n\geq 1}$ of finite sub-σ-algebra of $\mathcal{F}$ such that $(Z_n)_n$ is a $(\mathcal{F}_n)_{n\geq 1}$-martingale.
(2) $\forall n \geq 1, X_n = X_{n+1}$ a.s. or $Y_n = Y_{n+1}$ a.s.
(3) Z_1 is a.s. constant.

Notice that $(Z_n)_{n\geq 1}$ being a bounded martingale, it converges almost surely to a limit Z_∞.

Definition 13 Let A be a measurable subset of $X \times Y$.

$$\begin{aligned} A^* = \{z \in X \times Y,\ &\text{there exists a bimartingale } (Z_n)_{n\geq 1} \\ &\text{converging to a limit } Z_\infty \text{ such that } Z_\infty \in A \text{ a.s.} \\ &\text{and } Z_1 = z \text{ a.s.}\} . \end{aligned}$$

One can show that any atomless probability space $(\Omega, \mathcal{F}, \mathcal{P})$, or any product of convex compact spaces $X \times Y$ containing A, induce the same set A^*. One can also substitute condition (2) by: $\forall n \geq 1, (X_n = X_{n+1}$ or $Y_n = Y_{n+1})$ a.s. Notice that without condition (2), the set A^* would just be the convex hull of A.

We always have $A \subset A^* \subset \text{conv}(A)$, and these inclusions can be strict. For example, if $X = Y = [0, 1]$ and $A = \{(0, 0), (1, 0), (0, 1)\}$, it is possible to show that $A^* = \{(x, y) \in [0, 1] \times [0, 1], x = 0 \text{ or } y = 0\}$. A^* always is biconvex and thus contains biconv(A), which is defined as the smallest biconvex set which contains A. The inclusion biconv$(A) \subset A^*$ can also be strict, as shown by the following example:

Example 5 Put $X = Y = [0, 1]$, $v_1 = (1/3, 0)$, $v_2 = (0, 2/3)$, $v_3 = (2/3, 1)$, $v_4 = (1, 1/3)$, $w_1 = (1/3, 1/3)$, $w_2 = (1/3, 2/3)$, $w_3 = (2/3, 2/3)$ et $w_4 = (2/3, 1/3)$, and $A = \{v_1, v_2, v_3, v_4\}$.

A is biconvex, so $A = \text{biconv}(A)$. Consider now the following Markov process $(Z_n)_{n\geq 1}$, with $Z_1 = w_1$. If $Z_n \in A$, then $Z_{n+1} = Z_n$. If $Z_n = w_i$ for some i, then $Z_{n+1} = w_{i+1(\text{mod } 4)}$ with probability 1/2, and $Z_{n+1} = v_i$ with probability 1/2. $(Z_n)_n$ is a bimartingale converging a.s. to a point in A, hence $w_1 \in A^* \backslash \text{biconv}(A)$.

We now present a geometric characterization of the set A^*, and assume here that A is closed. For each biconvex subset B of $X \times Y$ containing A, we denote by $nsc(B)$ the set of elements of B which can not be separated from A by

a continuous bounded biconvex function on A. More precisely, $nsc(B) = \{z \in B, \forall f\colon B \longrightarrow \mathbb{R}$ bounded biconvex, and continuous on A, $f(z) \leq \sup\{f(z'), z' \in A\}\}$.

Theorem 11 ([4]) *A^* is the largest biconvex set B containing A such that $nsc(B) = B$.*

Let us now come back to repeated games and to the notations of Subsect. "Approachability for Player 1 Versus Excludability for Player 2". To be precise, we need to add the component of player 2's payoff, and consequently to slightly modify the definitions. G is closed in $\mathbb{R}^K_M \times \mathbb{R}_M \times \Delta(K)$. For $B \subset \mathbb{R}^K_M \times \mathbb{R}_M \times \Delta(K)$, B is biconvex if for each a in $\mathbb{R}^K_M$ and for each p in $\Delta(K)$, the sections $\{(\beta, p'), (a, \beta, p') \in B\}$ and $\{(a', \beta), (a', \beta, p) \in B\}$ are convex. A real function f defined on a biconvex set B is said to be biconvex if $\forall a$, $\forall p$, $f(a, ., .)$ and $f(., ., p)$ are convex.

Theorem 12 ([4]) *G^* is the largest biconvex set B containing G such that: $\forall z \in B, \forall f\colon B \longrightarrow \mathbb{R}$ bounded biconvex, and continuous on A, $f(z) \leq \sup\{f(z'), z' \in G\}$.*

Non-observable Actions

We now consider the case where, as in the general definition of Sect. "Definition of the Subject", there is a signaling function $q\colon K \times A \longrightarrow \Delta(U)$ giving the distributions of the signals received by the players as a function of the state of nature and the action profile just played. The particular case where $q(k, a)$ does not depend on k is called *state independent signaling*. The previous models correspond to the particular case of perfect observation, where the signals received by the players exactly reveal the action profile played.

Theorem 1 has been generalized [6] to the general case of signaling function. We keep the notations of Sect. "The Standard Model of Aumann and Maschler". Given a mixed action $x \in \Delta(I)$, an action j in J and a state k, we denote by $Q(k, x, j)$ the marginal distribution on U^2 of the law $\sum_{i \in I} x(i)q(k, i, j)$, i. e. $Q(k, x, j)$ is the law of the signal received by player 2 if the state is k, player 1 uses x and player 2 plays j. The set of non revealing strategies of player 1 is then defined as:

$$\begin{aligned} NR(p) = \{x = (x^k)_{k \in K} \in \Delta(I)^K, \\ \forall k \in K, \forall k' \in K \text{ s.t. } p^k p^{k'} > 0, \\ \forall j \in J, Q(k, x^k, j) = Q(k', x^{k'}, j)\}\,. \end{aligned}$$

If the initial probability is p and player 1 plays a strategy x in $NR(p)$ (i. e. plays x^k if the state is k), the a posteriori of player 2 will remain a.s. constant: player 2 can deduce no information on the selected state k. The value of the non revealing game becomes:

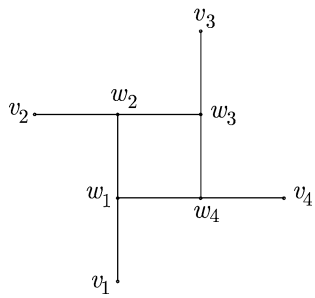

Repeated Games with Incomplete Information, Figure 5
The "four frogs" example of Aumann and Hart: $A^* \neq$ biconv(A)

$$\begin{aligned} u(p) &= \max_{x \in NR(p)} \min_{y \in \Delta(J)} \sum_{k \in K} p^k G^k(x^k, y) \\ &= \min_{y \in \Delta(J)} \max_{x \in NR(p)} \sum_{k \in K} p^k G^k(x^k, y)\,, \end{aligned}$$

where $G^k(x^k, y) = \sum_{i,j} x^k(i) y(j) G^k(i, j)$, and the convention $u(p) = -\infty$ if $NR(p) = \emptyset$. Theorem 1 perfectly extends here: The repeated game with initial probability p has a uniform value given by $\text{cav}u(p)$.

The explicit construction of an optimal strategy of player 2 (see Subsect. "Back to the Standard Model") has also been generalized to the general signaling case (see [35], and part B, p. 234 [57] for random signals).

Regarding zero-sum games with lack of information on both sides, the results of Sect. "Zero-Sum Games with Lack of Information on Both Sides" have been generalized to the case of state independent signaling (see [50,52] and [55]). Attention has been paid to the speed of convergence of the value function $(v_T)_T$, and bounds are identical for both models of lack of information on one side and on both sides, if we assume state independent signaling: this speed is of order $1/T^{1/2}$ for games with perfect observation, and of order $1/T^{1/3}$ for games with signals (these orders are optimal, both for lack of information on one side and lack of information on both sides, see [89,90]). For state dependent signaling and lack of information on one side, it was shown by Mertens [51] that the convergence occurs with worst case error $\sim (\ln n/n)^{1/3}$.

A particular class of zero-sum repeated games with state dependent signaling has been studied (games with no signals, see [54,88] and [82]). In these games, the state k is first selected according to a known probability and is not announced to the players; then after each stage both players receive the same signal which is either "nothing" or "the state is k". It is not possible to obtain here a standard recursive formula with state space $\Delta(K)$, or even

$\Delta(K) \times \Delta(K)$, because even when the strategies are given, during the play none of the players can compute the a posteriori of the other player. It was shown that the maxmin and the minmax may differ, although $\lim_T v_T$ always exists.

In non zero-sum repeated games with lack of information on one side, the existence of "joint plan" equilibria have been generalized to the case of state independent signaling [62], and more generally to the case where "player 1 can send non revealing signals to player 2" [77]. The existence of a uniform equilibrium in the general signaling case is still an open question see [78].

Miscellaneous

Zero-Sum Games

In games with lack of information on one side, it is important that player 1 knows not only the selected state k, but also the a priori p. [84] provides an example of a game with lack of information on "one and a half" side with no uniform value. More precisely, in this example nature first chooses p in $\{p_1, p_2\}$ according to a known probability, and announces p to player 2 only; then k is selected according to p, and announced to player 1 only; finally the matrix game G^k is played.

For games with lack of information on one side, the value function v_T is a concave piecewise linear function of the initial probability p (see [61] for more generality). On the contrary, the discounted value v_λ can be quite a complex function of p: in example 2 of Sect. "Definition of the Subject", Mayberry [49] has proved that for $2/3 < \lambda < 1$, v_λ is, at each rational value of p, non differentiable.

Convergence of the value functions $(v_T)_T$ and $(v_\lambda)_\lambda$ have been widely studied. We have already discussed the speed of convergence in Sect. "Non-observable Actions", but much more can be said.

Example 6 Standard model of lack of information on one side and observable actions.

$$K = \{a, b\}\,,$$
$$G^a = \begin{pmatrix} 3 & -1 \\ -3 & 1 \end{pmatrix} \quad \text{and} \quad G^b = \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix}.$$

One can show [53] that for each $p \in [0, 1]$, viewed as the initial probability of state a, the sequence $\sqrt{T} v_T(p)$ converges to $\varphi(p)$, where $\varphi(p) = 1/\sqrt{2\pi}\,\mathrm{e}^{-x_p^2/2}$, and x_p satisfies $1/\sqrt{2\pi}\int_{-\infty}^{x_p} \mathrm{e}^{-x^2/2}\mathrm{d}x = p$. So the limit of $\sqrt{T} v_T(p)$ is the standard normal density function evaluated at its p-quantile.

The apparition of the normal distribution is by no way an isolated phenomenon, but rather an important property of some repeated games ([12,13,14,15,18], ...).

B. de Meyer introduced the notion of "dual game" (see the previous references and also [17,19,41,69]). Let us now illustrate this on the standard model of Sect. "The Standard Model of Aumann and Maschler".

Let z be a parameter in $\mathbb{R}^K$. In the dual game $\Gamma_T^*(z)$, player 1 first secretly chooses the state k. Then at each stage $t \le T$, the players choose as usual actions i_t and j_t which are announced before proceeding to the next stage. With time horizon T, player 1's payoff finally is $\frac{1}{T}\sum_{t=1}^T G^k(i_t, j_t) - z^k$. This player is thus now able to fix the state equal to k, but has to pay z^k for it. It can be shown that the T-stage dual game $\Gamma_T^*(z)$ has a value $w_T(z)$. w_T is convex, and is linked to the value of the primal game by the conjugate formula:

$$w_T(z) = \max_{p \in \Delta(K)} (v_T(p) - \langle p, z\rangle), \quad \text{and}$$
$$v_T(p) = \inf_{z \in \mathbb{R}^K} (w_T(z) + \langle p, z\rangle)\,.$$

And $(w_T)_T$ satisfies the dual recursive formula:

$$w_{T+1}(z) = \min_{y \in \Delta(J)} \max_{i \in I} \frac{T}{T+1} w_T \cdot \left(\frac{T+1}{T} z - \frac{1}{T} \sum_{j \in J} y_j \left(G^k(i, j) \right)_k \right).$$

There are also strong relations between the optimal strategies of the players in the primal and dual games, and this gives a way to compute recursively optimal strategies of the uninformed player in the finite game (see also [32] on this topic).

Repeated games with incomplete information, as well as stochastic games, can also be studied in a functional analysis setup called the operator approach. This general approach is based on the study of the recursive formula [40,70,85].

In [65], the standard model, as well as the proof of Theorem 1, is generalized to the case where the state is not fixed at the beginning of the game, but evolves according to a Markov chain uniquely observed by player 1 (see also [59] for non observable actions, [48] and [33] for the difficulty of computing the value, [67] for the generalization to a state process controlled and observed by player 1, and [71] for several kinds of stochastic games with lack of information on one side). It is known since [80] that

the uniform value may not exist in general for stochastic games with lack of information on one side (where the stochastic game to be played is first randomly selected and announced to player 1 only).

Blackwell's approachability theorem has been extended to infinite dimensional spaces by Lehrer [43]. Approachability theory has strong links with the existence of no-regret strategies (first studied in [31], see also [10,26,30,44,72] and the recent book [9]), convergence of simple procedures to the set of correlated equilibria [31], and calibration [25,42]. The links between merging, reputation phenomena and repeated games with incomplete information have been studied in [83], where several existing results are unified. Finally, no-regret and approachability have also been studied when the players have bounded computational capacities (finite automata, bounded recall strategies) [45,46].

Let us mention also that de Meyer and Moussa Saley studied the modelization via Brownian motions in financial models [18]. They introduced a marked game based on a repeated game with lack of information on one side, and showed the endogenous apparition of a Brownian motion (see [16] for incomplete information on both sides).

Non Zero-Sum Games

In the setup of Sect. "Non Zero-Sum Games with Lack of Information on One Side", it is interesting to study the number of communication stages which is needed to construct the different equilibria. This number is linked with the convergence of the associated bimartingales (see [4,6,21,24]). Let us mention also that F. Forges [23] gave a similar characterization of equilibrium payoffs, for a larger notion of equilibria called communication equilibria (see also [22] for correlated equilibria). Amitai [2] studied the set of equilibrium payoffs in case of lack of information on both sides. Aumann and Hart [5] characterized the equilibrium payoffs in two player games with lack of information on one side when long, payoff-irrelevant, preplay communication is allowed (see [1] for incomplete information on both sides).

The particular case where each player knows his own payoffs is particularly worthwhile studying (*known own payoffs*). In the two-player case with lack of information on one side, this amounts to say that player 2's payoffs do not depend on the selected state. In this case, Shalev [73] showed that any equilibrium payoff can be obtained as the payoff of an equilibrium which is completely revealing. This result generalizes to the non zero-sum case of lack of information of both sides (see the unpublished manuscript [37]), but unfortunately uniform equilibria may fail to exist even though both players known their own payoffs.

Another model deals with the symmetric case, where the players have an incomplete, but identical, knowledge of the selected state. After each stage they receive the same signal, which may depend on the state. A. Neyman and S. Sorin have proved the existence of equilibrium payoffs in the case of two players (see [60], the zero-sum case being solved in [36] and [20]).

Few papers study the case of more than 2 players. The existence of uniform equilibrium has been studied for 3 players and lack of information on one side [63], and in the case of two states of nature it appears that a completely revealing equilibria, or a joint plan equilibria by one of the informed players, always exists. Concerning n-player repeated games with incomplete information and signals, several papers study how the initial information can be strategically transmitted, independently of the payoffs ([64,66], and [68] for an application to a cryptographic model). As an application, the existence of completely revealing equilibria is obtained in particular cases.

Repeated games with incomplete information have been used to study perturbations of repeated games with complete information (see [27] and [11] for Folk theorem-like results, [7] for enforcing cooperation in games with a Pareto-dominant outcome, and [34] for a perturbation with known own payoffs). The case where the players have different discount factors has also been investigated [11,47].

Future Directions

Several open problems are well formulated and deserve attention. Does a uniform equilibrium always exist in two-player repeated games with lack of information on one side and general signaling, or in n-player repeated games with lack of information on one side? More conceptually, one should look for classes of n-player repeated games with incomplete information which allow for the existence of equilibria, and/or for a tractable description of equilibrium payoffs (or at least of some of these payoffs). Regarding applications, there is certainly a lot of room in the vast fields of financial markets, cryptology, and sequential decision problems.

Acknowledgments

I thank Françoise Forges, Sergiu Hart, Dinah Rosenberg, Robert Simon and Eilon Solan for their comments on a preliminary version of this chapter.

Bibliography

Primary Literature

1. Amitai M (1996) Cheap-Talk with Incomplete Information on Both Sides. Ph D Thesis, The Hebrew University of Jerusalem, http://ratio.huji.ac.il/dp/dp90.pdf
2. Amitai M (1996) Repeated games with incomplete information on Both Sides. Ph D Thesis, The Hebrew University of Jerusalem, http://ratio.huji.ac.il/dp/dp105.pdf
3. Aumann RJ (1964) Mixed and behaviour strategies in infinite extensive games. In: Dresher, Shapley, Tucker (eds) Advances in game theory. Annals of Mathematics Study vol 52, Princeton University Press, Princeton, pp 627–650
4. Aumann RJ, Hart S (1986) Bi-convexity and bi-martingales. Israel J Math 54:159–180
5. Aumann RJ, Hart S (2003) Long cheap talk. Econometrica 71:1619–1660
6. Aumann RJ, Maschler M (1995) Repeated games with incomplete information, with the collaboration of Stearns RE. MIT Press, Cambridge (contains a reedition of chapters of Reports to the US Arms Control and Disarmament Agency ST-80, 116 and 143, Mathematica, 1966–1967–1968)
7. Aumann RJ, Sorin S (1989) Cooperation and bounded recall. Games Econ Behav 1:5–39
8. Blackwell D (1956) An analog of the minmax theorem for vector payoffs. Pac J Math 65:1–8
9. Cesa-Bianchi N, Lugosi G (2006) Prediction, Learning and Games. Cambridge University Press, Cambridge
10. Cesa-Bianchi N, Lugosi G, Stoltz G (2006) Regret minimization under partial monitoring. Math Oper Res 31:562–580
11. Cripps MW, Thomas JP (2003) Some asymptotic results in discounted repeated games of one-sided incomplete information. Math Oper Res 28:433–462
12. de Meyer B (1996) Repeated games and partial differential equations. Math Oper Res 21:209–236
13. de Meyer B (1996) Repeated games, duality and the central limit theorem. Math Oper Res 21:237–251
14. de Meyer B (1998) The maximal variation of a bounded martingale and the central limit theorem. Annales de l'Institut Henri Poincaré, Probabilités et statistiques 34:49–59
15. de Meyer B (1999) From repeated games to Brownian games. Annales de l'Institut Henri Poincaré, Probabilités et statistiques 35:1–48
16. de Meyer B, Marino A (2004) Repeated market games with lack of information on both sides. DP 2004.66, MSE Université Paris I
17. de Meyer B, Marino A (2005) Duality and optimal strategies in the finitely repeated zero-sum games with incomplete information on both sides. DP 2005.27, MSE Université Paris I
18. de Meyer B, Moussa H (2003) Saley On the strategic origin of Brownian motion in finance. Int J Game Theory 31:285–319
19. de Meyer B, Rosenberg D (1999) "*Cavu*" and the dual game. Math Oper Res 24:619–626
20. Forges F (1982) Infinitely repeated games of incomplete information: symmetric case with random signals. Int J Game Theory 11:203–213
21. Forges F (1984) A note on Nash equilibria in repeated games with incomplete information. Int J Game Theory 13:179–187
22. Forges F (1985) Correlated equilibria in a class of repeated games with incomplete information. Int J Game Theory 14:129–149
23. Forges F (1988) Communication equilibria in repeated games with incomplete information. Math Oper Res 13:191–231
24. Forges F (1990) Equilibria with communication in a job market example. Q J Econ 105:375–398
25. Foster D (1999) A Proof of Calibration via Blackwell's Approachability Theorem. Games Eco Behav 29:73–78
26. Foster D, Vohra R (1999) Regret in the On-Line Decision Problem. Games Econ Behav 29:7–35
27. Fudenberg D, Maskin E (1986) The folk theorem in repeated games with discounting or with incomplete information. Econometrica 54:533–554
28. Harsanyi J (1967–68) Games with incomplete information played by 'Bayesian' players, parts I-III. Manag Sci 8:159–182, 320–334, 486–502
29. Hart S (1985) Nonzero-sum two-person repeated games with incomplete information. Math Oper Res 10:117–153
30. Hart S (2005) Adaptative Heuristics. Econometrica 73:1401–1430
31. Hart S, Mas-Colell A (2000) A simple adaptative procedure leading to Correlated Equilibrium. Econometrica 68:1127–1150
32. Heuer M (1992) Optimal strategies for the uninformed player. Int J Game Theory 20:33–51
33. Hörner J, Rosenberg D, Solan E, Vieille N (2006) On Markov Games with Incomplete Information on One Side. Discussion paper 1412, Center for Mathematical Studies in Economics and Management Science, Northwestern University, Evanston
34. Israeli E (1999) Sowing Doubt Optimally in Two-Person Repeated Games. Games Econ Behav 28:203–216
35. Kohlberg E (1975) Optimal strategies in repeated games with incomplete information. Int J Game Theory 4:7–24
36. Kohlberg E, Zamir S (1974) Repeated games of Incomplete information: The Symmetric Case. Ann Stat 2:40–41
37. Koren G (1992) Two-person repeated games where players know their own payoffs. master thesis, Tel-Aviv University, Tel-Aviv http://www.ma.huji.ac.il/hart/papers/koren.pdf
38. Kuhn HW (1953) Extensive games and the problem of information. In: Kuhn, Tucker (eds) Contributions to the Theory of Games, vol II. Annals of Mathematical Studies 28, Princeton University Press, Princeton, pp 193–216
39. Laraki R (2001) Variational inequalities, system of functional equations and incomplete information repeated games. SIAM J Control Optim 40:516–524
40. Laraki R (2001) The splitting game and applications. Int J Game Theory 30:359–376
41. Laraki R (2002) Repeated games with lack of information on one side: the dual differential approach. Math Oper Res 27:419–440
42. Lehrer E (2001) Any inspection is manipulable. Econometrica 69:1333–1347
43. Lehrer E (2003) Approachability in infinite dimensional spaces. Int J Game Theory 31:253–268
44. Lehrer E (2003) A wide range no-regret theorem. Games Econ Behav 42:101–115
45. Lehrer E, Solan E (2003) No regret with bounded computational capacity. DP 1373, Center for Mathematical Studies in Economics and Management Science, Northwestern University, Evanston
46. Lehrer E, Solan E (2006) Excludability and bounded computational capacity. Math Oper Res 31:637–648

47. Lehrer E, Yariv L (1999) Repeated games with lack of information on one side: the case of different discount factors. Math Oper Res 24:204–218
48. Marino A (2005) The value of a particular Markov chain game. Chaps 5 and 6, Ph D thesis, Université Paris I http://alexandre.marino.free.fr/theseMarino.pdf
49. Mayberry J-P (1967) Discounted repeated games with incomplete information. Report of the US Arms control and disarmament agency, ST116, Chap V. Mathematica, Princeton, pp 435–461
50. Mertens J-F (1972) The value of two-person zero-sum repeated games: the extensive case. Int J Game Theory 1:217–227
51. Mertens J-F (1998) The speed of convergence in repeated games with incomplete information on one side. Int J Game Theory 27:343–357
52. Mertens J-F, Zamir S (1971) The value of two-person zero-sum repeated games with lack of information on both sides. Int J Game Theory 1:39–64
53. Mertens J-F, Zamir S (1976) The normal distribution and repeated games. Int J Game Theory 5:187–197
54. Mertens J-F, Zamir S (1976) On a repeated game without a recursive structure. Int J Game Theory 5:173–182
55. Mertens J-F, Zamir S (1977) A duality theorem on a pair of simultaneous functional equations. J Math Anal Appl 60:550–558
56. Mertens J-F, Zamir S (1985) Formulation of Bayesian analysis for games with incomplete information. Int J Game Theory 14:1–29
57. Mertens J-F, Sorin S, Zamir S (1994) Repeated games. CORE discussion paper. Université Catholique de Louvain, Belgium, pp 9420–9422
58. Meyer PA (1966) Probability and Potentials. Blaisdell, New York (in French: Probabilités et potentiel, Hermann, 1966)
59. Neyman A (2008) Existence of Optimal Strategies in Markov Games with Incomplete Information. Int J Game Theory 37:581–596
60. Neyman A, Sorin S (1998) Equilibria in Repeated Games with Incomplete Information: The General Symmetric Case. Int J Game Theory 27:201–210
61. Ponssard JP, Sorin S (1980) The LP formulation of finite zero-sum games with incomplete information. Int J Game Theory 9:99–105
62. Renault J (2000) 2-player repeated games with lack of information on one side and state independent signalling. Math Oper Res 4:552–572
63. Renault J (2001) 3-player repeated games with lack of information on one side. Int J Game Theory 30:221–246
64. Renault J (2001) Learning sets in state dependent signalling game forms: a characterization. Math Oper Res 26:832–850
65. Renault J (2006) The value of Markov chain games with lack of information on one side. Math Oper Res 31:490–512
66. Renault J, Tomala T (2004) Learning the state of nature in repeated games with incomplete information and signals. Games Econ Behav 47:124–156
67. Renault J (2007) The value of Repeated Games with an informed controller. Technical report, Université Paris–Dauphine, Ceremade
68. Renault J, Tomala T (2008) Probabilistic reliability and privacy of communication using multicast in general neighbor networks. J Cryptol 21:250–279
69. Rosenberg D (1998) Duality and Markovian strategies. Int J Game Theory 27:577–597
70. Rosenberg D, Sorin S (2001) An operator approach to zero-sum repeated games. Israel J Math 121:221–246
71. Rosenberg D, Solan E, Vieille N (2004) Stochastic games with a single controller and incomplete information. SIAM J Control Optim 43:86–110
72. Rustichini A (1999) Minimizing Regret: The General Case. Games Econ Behav 29:224–243
73. Shalev J (1994) Nonzero-Sum Two-Person Repeated Games with Incomplete Information and Known-Own Payoffs. Games Econ Behav 7:246–259
74. Sion M (1958) On General Minimax Theorems. Pac J Math 8:171–176
75. Simon RS (2002) Separation of joint plan equilibrium payoffs from the min-max functions. Games Econ Beha 1:79–102
76. Simon RS, Spież S, Toruńczyk H (1995) The existence of equilibria in certain games, separation for families of convex functions and a theorem of Borsuk–Ulam type. Israel J Math 92:1–21
77. Simon RS, Spież S, Toruńczyk H (2002) Equilibrium existence and topology in some repeated games with incomplete information. Trans AMS 354:5005–5026
78. Simon RS, Spież S, Toruńczyk H (2008) Equilibria in a class of games and topological results implying their existence. RACSAM Rev R Acad Cien Serie A Mat 102:161–179
79. Sorin S (1983) Some results on the existence of Nash equilibria for non- zero sum games with incomplete information. Int J Game Theory 12:193–205
80. Sorin S (1984) Big match with lack of information on one side, Part I. Int J Game Theory 13:201–255
81. Sorin S (1984) On a pair of simultaneous functional equations. J Math Anal Appl 98:296–303
82. Sorin S (1989) On recursive games without a recursive structure: existence of lim v_n. Int J Game Theory 18:45–55
83. Sorin S (1997) Merging, reputation, and repeated games with incomplete information. Games Econ Behav 29:274–308
84. Sorin S, Zamir S (1985) A 2-person game with lack of information on 1 and 1/2 sides. Math Oper Res 10:17–23
85. Sorin S (2002) A first course on zero-sum repeated games. Mathématiques et Applications, vol 37. Springer
86. Spinat X (2002) A necessary and sufficient condition for approachability. Math Oper Res 27:31–44
87. Vieille N (1992) Weak approachability. Math Oper Res 17:781–791
88. Waternaux C (1983) Solution for a class of repeated games without recursive structure. Int J Game Theory 12:129–160
89. Zamir S (1971) On the relation between finitely and infinitely repeated games with incomplete information. Int J Game Theory 1:179–198
90. Zamir S (1973) On repeated games with general information function. Int J Game Theory 21:215–229
91. Zamir S (1992) Repeated Games of Incomplete Information: zero-sum. In: Aumann RJ, Hart S (eds) Handbook of Game Theory, I. Elsevier Science Publishers, Amsterdam, pp 109–154

Books and Reviews

Forges F (1992) Repeated Games of Incomplete Information: Nonzero sum. In: Aumann RJ, Hart S (eds) Handbook of Game Theory, vol I. Elsevier Science Publishers, Amsterdam, pp 155–177

Laraki R, Renault J, Tomala T (2006) Théorie des Jeux, Introduction à la théorie des jeux répétés. Editions de l'Ecole Polytechnique, Journées X-UPS, in French (Chapter 3 deals with repeated games with incomplete information). Palaiseau

Mertens J-F (1987) Repeated games. In: Proceedings of the International Congress of Mathematicians 1986. American Mathematical Society, Berkeley, pp 1528–1577

Reputation Effects

George J. Mailath
Department of Economics, University of Pennsylvania, Philadelphia, USA

Article Outline

Glossary
Definition of the Subject
Introduction
A Canonical Model
Two Long-Lived Players
Future Directions
Acknowledgments
Bibliography

Glossary

Action type A type of player who is committed to playing a particular action, also called a *commitment type* or *behavioral type*.

Complete information Characteristics of all players are common knowledge.

Flow payoff Stage game payoff.

Imperfect monitoring Past actions of all players are not public information.

Incomplete information Characteristics of some player are not common knowledge.

Long-lived player Player subject to intertemporal incentives, typically has the same horizon as length of the game.

Myopic optimum An action maximizing stage game payoffs.

Nash equilibrium A strategy profile from which no player has a profitable unilateral deviation (i. e., it is self-enforcing).

Nash reversion In a repeated game, permanent play of a stage game Nash equilibrium.

Normalized discounted value The discounted sum of an infinite sequence $\{a_t\}_{t\geq 0}$, calculated as $(1-\delta)\sum_{t\geq 0}\delta^t a_t$, where $\delta \in (0,1)$ is the discount value.

Perfect monitoring Past actions of all players are public information.

Repeated game The finite or infinite repetition of a stage game.

Reputation bound The lower bound on equilibrium payoffs of a player that the other player(s) believe may be a simple action type (typically the Stackelberg type).

Short-lived player Player not subject to intertemporal incentives, having a one-period horizon and so is myopically optimizing.

Simple action type An action who plays the same (pure or mixed) stage-game action in every period, regardless of history.

Stage game A game played in one period.

Stackelberg action In a stage game, the action a player would commit to, if that player had the chance to do so, i. e., the optimal commitment action.

Stackelberg type A simple action type that plays the Stackelberg action.

Subgame In a repeated game with perfect monitoring, the game following any history.

Subgame perfect equilibrium A strategy profile that induces a Nash equilibrium on every subgame of the original game.

Type The characteristic of a player that is not common knowledge.

Definition of the Subject

Repeated games have many equilibria, including the repetition of stage game Nash equilibria. At the same time, particularly when monitoring is imperfect, certain plausible outcomes are not consistent with equilibrium. *Reputation effects* is the term used for the impact upon the set of equilibria (typically of a repeated game) of perturbing the game by introducing incomplete information of a particular kind. Specifically, the characteristics of a player are not public information, and the other players believe it is possible that the distinguished player is a type that necessarily plays some action (typically the Stackelberg action). Reputation effects fall into two classes: "Plausible" phenomena that are not equilibria of the original repeated game are equilibrium phenomena in the presence of incomplete information, and "implausible" equilibria of the original game are not equilibria of the incomplete information game. As such, reputation effects provide an important qualification to the general indeterminacy of equilibria.

Introduction

Repeating play of a stage game often allows for equilibrium behavior inconsistent with equilibrium of that stage game. If the stage game has multiple Nash equilibrium

	C	D
C	$2, 2$	$-1, 3$
D	$3, -1$	$0, 0$

Reputation Effects, Figure 1
The prisoners' dilemma. The cooperative action is labeled *C*, while defect is labeled *D*

payoffs, a large finite number of repetitions provide sufficient intertemporal incentives for behavior inconsistent with stage-game Nash equilibria to arise in some subgame perfect equilibria. However, many classic games do not have multiple Nash equilibria. For example, mutual defection *DD* is the unique Nash equilibrium of the prisoners' dilemma, illustrated in Fig. 1.

A standard argument shows that the finitely repeated prisoner's dilemma has a unique subgame perfect equilibrium, and in this equilibrium, *DD* is played in every period: In any subgame perfect equilibrium, in the last period, *DD* must be played independently of history, since the stage game has a unique Nash equilibrium. Then, since play in the last period is independent of history, there are no intertemporal incentives in the penultimate period, and so *DD* must again be played independently of history. Proceeding recursively, *DD* must be played in every period independently of history. (In fact, the finitely repeated prisoners' dilemma has a unique Nash equilibrium outcome, given by *DD* in every period.)

This contrasts with intuition, which suggests that if the prisoners' dilemma were repeated a sufficiently large (though finite) number of times, the two players would find a way to play cooperatively (*C*) at least in the initial stages. In response, Kreps, Milgrom, Roberts and Wilson [15] argued that intuition can be rescued in the finitely repeated prisoners' dilemma by introducing incomplete information. In particular, suppose each player assigns some probability to their opponent being a behavioral type who mechanistically plays tit-for-tat (i. e., plays *C* in the first period or if the opponent had played *C* in the previous period, and plays *D* if the opponent had played *D* in the previous period) rather than being a rational player. No matter how small the probability, if the number of repetitions is large enough, the rational players will play *C* in early periods, and the fraction of periods in which *CC* is played is close to one.

This is the first example of a *reputation effect*: a small degree of incomplete information (of the right kind) both rescues the intuitive *CC* for many periods as an equilibrium outcome, and eliminates the unintuitive always *DD* as one. In the same issue of the *Journal of Economic Theory* containing Kreps, Milgrom, Roberts and Wilson [15], Kreps and Wilson[14] and Milgrom and Roberts [18] explored reputation effects in the finite chain store of Selten [22], showing that intuition is again rescued, this time by introducing the possibility that the chain store is a "tough" type who always fights entry.

Reputation effects describe the impact upon the set of equilibria of the introduction of small amounts of incomplete information of a particular form into repeated games (and other dynamic games). Reputation effects fall into two classes: "Plausible" phenomena that are not equilibria of the complete information game are equilibrium phenomena in the presence of incomplete information, and "implausible" equilibria of the complete information game are not equilibria of the incomplete information game.

Reputation effects are distinct from the equilibrium phenomenon in complete information repeated games that are sometimes described as capturing *reputations*. In this latter use, an equilibrium of the complete information repeated game is selected, involving actions along the equilibrium path that are not Nash equilibria of the stage game. As usual, incentives to choose these actions are created by attaching less favorable continuation paths to deviations. Players who choose the equilibrium actions are then interpreted as maintaining a reputation for doing so, with a punishment-triggering deviation interpreted as causing the loss of one's reputation. For example, players who cooperate in the *infinitely* repeated prisoners' dilemma are interpreted as having (or maintaining) a cooperative reputation, with any defection destroying that reputation. In this usage, the link between past behavior and expectations of future behavior is an equilibrium phenomenon, holding in some equilibria, but not in others. The notion of reputation is used to interpret an equilibrium strategy profile, but otherwise adds nothing to the formal analysis.

In contrast, the approach underlying reputation effects begins with the assumption that a player is uncertain about key aspects of her opponent. For example, player 2 may not know player 1's payoffs, or may be uncertain about what constraints player 1 faces on his ability to choose various actions. This incomplete information is a device that introduces an intrinsic connection between past behavior and expectations of future behavior. Since incomplete information about players' characteristics can have dramatic effects on the *set* of equilibrium payoffs, reputations in this approach do not describe certain equilibria, but rather place constraints on the set of possible equilibria.

An Example

While reputation effects were first studied in a symmetric example with two long-lived players, they arise in their

	h	ℓ
H	$2, 3$	$0, 2$
L	$3, 0$	$1, 1$

Reputation Effects, Figure 2
The product-choice game

purest form in infinitely repeated games with one long-lived player playing against a sequence of short-lived players. The chain store game of Selten [22] is a finitely repeated game in which a chain store (the long-lived player) faces a finite sequence of potential entrants in its different markets. Since each entrant only cares about its own decision, it is short-lived.

Consider the "product-choice" game of Fig. 2. The row player (player 1), who is long-lived, is a firm choosing between high (H) and low (L) effort, while the column player (player 2), who is short-lived, is a customer choosing between a high (h) or low (ℓ) priced product. (Mailath and Samuelson [17] illustrate various aspects of repeated games and reputation effects using this example.) Player 2 prefers the high-priced product if the firm has exerted high effort, but prefers the low-priced product if the firm has not. The firm prefers that customers purchase the high-priced product and is willing to commit to high effort to induce that choice by the customer. In a simultaneous move game, however, the firm cannot observably choose effort before the customer chooses the product. Since high effort is costly, the firm prefers low effort, no matter the choice of the customer.

The stage game has a unique Nash equilibrium, in which the firm exerts low effort and the customer purchases the low-priced product. Suppose the game is played infinitely often, with *perfect monitoring* (i. e., the history of play is public information). The firm is *long-lived* and discounts flow profits by the discount factor $\delta \in (0, 1)$, and is *patient* if δ is close to 1. The role of the customer is taken by a succession of *short-lived* players, each of whom plays the game only once (and so myopically optimizes). It is standard to abuse language by treating the collection of short-lived players as a single myopically optimizing player.

When the firm is sufficiently patient, there is an equilibrium outcome in the repeated game in which the firm always exerts high effort and customers always purchase the high-priced product. The firm is deterred from taking the immediate myopically optimal action of low effort by the prospect of future customers then purchasing the low-priced product. Purchasing the high-priced product is a best response for the customer to high effort, so that no incentive issues arise concerning the customer's behavior. In this equilibrium, the long-lived player's payoff is 2 (the firm's payoffs are calculated as the normalized discounted sum, i. e., as the discounted sum of flow payoffs normalized by $(1 - \delta)$, so that payoffs in the infinite horizon game are comparable to flow payoffs). However, there are many other equilibria, including one in which low effort is exerted and low price purchased in every period, leading to a payoff of 1 for the long-lived player. Indeed, for $\delta \geq 1/2$, the set of pure-strategy subgame-perfect-equilibrium player 1 payoffs is given by the entire interval $[1, 2]$.

Reputation effects effectively rule out any payoff less than 2 as an equilibrium payoff for player 1. Suppose customers are not entirely certain of the characteristics of the firm. More specifically, suppose they attach high probability to the firm's being "normal," that is, having the payoffs given above, but they also entertain some (possibly very small) probability that they face a firm who fortuitously has a technology or some other characteristic that ensures high effort. Refer to the latter as the "H-action" type of firm. Since such a type necessarily plays H in every period, it is a type described by behavior (not payoffs), and such a type is often called a *behavioral* or *commitment* type.

This is now a game of *incomplete information*, with the customers uncertain of the firm's type. Since the customers assign high probability to the firm being "normal," the game is in some sense close to the game of complete information. None the less, reputation effects are present: For a sufficiently patient firm, in any Nash equilibrium of the repeated game, the firm's payoff cannot be significantly less than 2. This result holds no matter how unlikely customers think the H-action type to be, though increasing patience is required from the normal firm as the action type becomes less likely.

The intuition behind this result is most easily seen by considering pure strategy Nash equilibria of the incomplete information game where the customers believe the firm is either the normal or the H-action type. In that case, there is no pure strategy Nash equilibrium with a payoff less than 2δ (which is clearly close to 2 for δ close to 1). In the pure strategy Nash equilibrium, either the firm always plays H, (in which case, the customers always play h and the firm's payoff is 2), or there is a first period (say t) in which the firm plays L, revealing to future customers that he is the normal type (since the action type plays H in every period). In such an equilibrium, customers play h before t (since both types of firm are choosing H). After observing H in period t, customers conclude the firm is the H-action type. Consequently, as long as H is always chosen thereafter, customers subsequently play h (since they continue to believe the firm is the H-action type, and

so necessarily plays H). An easy lower bound on the normal firm's equilibrium payoff is then obtained by observing that the normal firm's payoff must be at least the payoff from mimicking the action type in every period. The payoff from such behavior is at least as large as

$$\underbrace{(1-\delta)\sum_{\tau=0}^{t-1}\delta^{\tau}2}_{\substack{\text{payoff in } \tau < t \text{ from pooling}\\ \text{with } H\text{-action type}}} + \underbrace{(1-\delta)\delta^{t}\times 0}_{\substack{\text{payoff in } t \text{ from playing } H \text{ when}\\ L \text{ may be myopically optimal}}}$$

$$+ \underbrace{(1-\delta)\sum_{\tau=t+1}^{\infty}\delta^{\tau}2}_{\substack{\text{payoff in } \tau > t \text{ from playing like}\\ \text{and being treated as the } H\text{-action type}}}$$

$$\begin{aligned} &= (1-\delta^{t})2 + \delta^{t+1}2 \\ &= 2 - 2\delta^{t}(1-\delta) \\ &\geq 2 - 2(1-\delta) = 2\delta\,. \end{aligned}$$

The outcome in which the stage game Nash equilibrium $L\ell$ is played in every period is thus eliminated.

Since reputation effects are motivated by the hypothesis that the short-lived players are uncertain about some aspect of the long-lived player's characteristics, it is important that the results are not sensitive to the precise nature of that uncertainty. In particular, the lower bound on payoffs should not require that the short-lived players *only* assign positive probability to the normal and the H-action type (as in the game just analyzed). And it does not: The customers in the example may assign positive probability to the firm being an action type that plays H on even periods, and L on odd periods, as well as to an action type that plays H in every period before some period t' (that can depend on history), and then always plays L. Yet, as long as the customers assign positive probability to the H-action type, for a sufficiently patient firm, in any Nash equilibrium of the repeated game, the firm's payoff cannot be significantly less than 2.

Reputation effects are more powerful in the presence of imperfect monitoring. Suppose that the firm's choice of H or L is not observed by the customers. Instead, the customers observe a public signal $y \in \{\underline{y}, \bar{y}\}$ at the end of each period, where the signal $\bar{y}$ is realized with probability $p \in (0,1)$ if the firm chose H, and with the smaller probability $q \in (0,p)$ if the firm chose L. Interpret $\bar{y}$ as a good meal: while customers do not observe effort, they do observe a noisy signal (the quality of the meal) of that effort, with high effort leading to a good meal with higher probability. In the game with complete information, the largest equilibrium payoff to the firm is now given by

$$\bar{v}_1 \equiv 2 - \frac{1-p}{p-q}\,, \tag{1}$$

reflecting the imperfect monitoring of the firm's actions (the firm is said to be subject to binding moral hazard, see Sect. 7.6 in [17]). Since deviations from H cannot be detected for sure, there are no equilibria with the deterministic outcome path of Hh in every period. In some periods after some histories, $L\ell$ must be played in order to provide the appropriate intertemporal incentives to the firm.

As under perfect monitoring, as long as customers assign positive probability to the H-action type in the incomplete information game with imperfect monitoring, for a sufficiently patient firm, in any Nash equilibrium of the repeated game, the firm's payoff cannot be significantly less than 2 (in particular, this lower bound exceeds $\bar{v}_1$). Thus, in this case, reputation effects provide an intuitive lower bound on equilibrium payoffs that both rules out "bad" equilibrium payoffs, as well as rescues outcomes in which Hh occurs in most periods.

Proving that a reputation bound holds in the imperfect monitoring case is considerably more involved than in the perfect monitoring case. In perfect-monitoring games, it is only necessary to analyze the evolution of the customers' beliefs when always observing H, the action of the H-action type. In contrast, imperfect monitoring requires consideration of belief evolution on all histories that arise with positive probability.

None the less, the intuition is the same: Consider a putative equilibrium in which the normal firm receives a payoff less than $2-\varepsilon$. Then the normal and action types must be making different choices over the course of the repeated game, since an equilibrium in which they behave identically would induce customers to choose h and would yield a payoff of 2. As in the perfect monitoring case, the normal firm has the option of mimicking the behavior of the H-action type. Suppose the normal firm does so. Since the customers expect the normal type of firm to behave differently from the H-action type, they will more often see signals indicative of the H-action type (rather than the normal type), and so must eventually become convinced that the firm is the H-action type. Hence, in response to this deviation, the customers will eventually play their best response to H of h. While "eventually" may take a while, that time is independent of the equilibrium (indeed of the discount factor), depending only on the imperfection in the monitoring and the prior probability assigned to the H-action type. Then, if the firm is sufficiently patient, the payoff from mimick-

ing the H-action type is arbitrarily close to 2, contradicting the existence of an equilibrium in which the firm's payoff fell short of $2 - \varepsilon$.

At the same time, because monitoring is imperfect, as discussed in Sect. "Temporary Reputation Effects", the reputation effects are necessarily transient. Under general conditions in imperfect-monitoring games, the incomplete information that is at the core of reputation effects is a short-run phenomenon. Player 2 must eventually come to learn player 1's type and continuation play must converge to an equilibrium of the complete information game.

Reputation effects arise for very general specifications of the incomplete information as long as the customers assign strictly positive probability to the H-action type. It is critical, however, that the customers do assign strictly positive probability to the H-action type. For example, in the product-choice game, the set of Nash equilibria of the repeated game is not significantly impacted by the possibility that the firm is either normal or the L-action type only. While reputation effects per se do not arise from the L-action type, it is still of interest to investigate the impact of such uncertainty on behavior using stronger equilibrium notions, such as Markov perfection (see Mailath and Samuelson [16]).

A Canonical Model

The Stage Game

The stage game is a two-player simultaneous-move finite game of public monitoring. Player i has action set A_i, $i = 1, 2$. Pure actions for player i are denoted by $a_i \in A_i$, and mixed actions are denoted by $\alpha_i \in \Delta(A_i)$, where $\Delta(A_i)$ is the set of probability distributions over A_i. Player 2's actions are public, while player 1's are potentially private. The public signal of player 1's action, denoted by y is drawn from a finite set Y, with the probability that y is realized under the pure action profile $a \in A \equiv A_1 \times A_2$ denoted by $\rho(y \mid a)$. Player 1's ex post payoff from the action profile a and signal realization y is $r_1(y, a)$, and so the ex ante (or expected) flow payoff is $u_1(a) \equiv \sum_y r_1(y, a)\rho(y \mid a)$. Player 2's ex post payoff from the action profile a and signal realization y is $r_2(y, a_2)$, and so the ex ante (or expected) flow payoff is $u_2(a) \equiv \sum_y r_2(y, a_2)\rho(y \mid a)$. Since player 2's ex post payoff is independent of player 1's actions, player 1's actions only affect player 2's payoffs through the impact on the distribution of the signals and so on ex ante payoffs. While the ex post payoffs r_i play no explicit role in the analysis, they justify the informational assumptions to be made. In particular, the model requires that histories of signals and past actions are the only information players receive, and so it is important that stage game payoffs u_i are not informative about the action choice (and this is the critical feature delivered by the assumptions that ex ante payoffs are not observable and that payer 2's ex post payoffs do *not* depend on a_1).

Perfect monitoring is the special case where $Y = A_1$ and $\rho(y \mid a) = 1$ if $y = a_1$, and 0 otherwise.

The results in this section hold under significantly weaker monitoring assumptions. In particular, it is not necessary that the actions of player 2 be public. If these are also imperfectly monitored, then the ex post payoff for player 1 is independent of player 2 actions. Since player 2 is short-lived, when player 2's actions are not public, it is then natural to also assume that the period t player 2 does not know earlier player 2's actions.

The Complete Information Repeated Game

The stage game is infinitely repeated. Player 1 is long-lived, with payoffs given by the normalized discounted value $(1 - \delta)\sum_{t=0}^{\infty} \delta^t u_1^t$, where $\delta \in (0, 1)$ is the discount factor and u_1^t is player 1's period t flow payoff. Player 1 is *patient* if δ is close to 1. As in our example, the role of player 2 is taken by a succession of *short-lived* players, each of whom plays the game only once (and so myopically optimizes).

Player 1's set of private histories is $\mathcal{H}_1 \equiv \cup_{t=0}^{\infty}(Y \times A)^t$ and the set of public histories (which coincides with the set of player 2's histories) is $\mathcal{H} \equiv \cup_{t=0}^{\infty}(Y \times A_2)^t$. If the game has perfect monitoring, histories $h = (y^0, a^0; y^1, a^1; \ldots; y^{t-1}, a^{t-1})$ in which $y^\tau \neq a_1^\tau$ for some $\tau \leq t - 1$ arise with zero probability, independently of behavior, and so can be ignored. A strategy σ_1 for player 1 specifies a probability distribution over 1's pure action set for each possible private history, i. e., $\sigma_1 : \mathcal{H}_1 \to \Delta(A_1)$. A strategy σ_2 for player 2 specifies a probability distribution over 2's pure action set for each possible public history, i. e., $\sigma_2 : \mathcal{H} \to \Delta(A_2)$.

Definition 1 The strategy profile (σ_1^*, σ_2^*) is a *Nash equilibrium* if

1. there does not exist a strategy σ_1 yielding a strictly higher payoff for player 1 when player 2 plays σ_2^*, and
2. in all periods t, after any history $h^t \in \mathcal{H}$ arising with positive probability under (σ_1^*, σ_2^*), $\sigma_2^*(h^t)$ maximizes $E[u_2(\sigma_1^*(h_1^t), a_1) \mid h^t]$, where the expectation is taken over the period t-private histories that player 1 may have observed.

The Incomplete Information Repeated Game

In the incomplete information game, the type of player 1 is unknown to player 2. A possible type of player 1 is denoted by $\xi \in \Xi$, where Ξ is a finite or countable set (see Fudenberg and Levine [12] for the uncountable case). Player 2's prior belief about 1's type is given by the distribution μ, with support Ξ. The set of types is partitioned into a set of *payoff types* Ξ_1, and a set of *action types* $\Xi_2 \equiv \Xi \backslash \Xi_1$. Payoff types maximize the average discounted value of payoffs, which depend on their type and which may be nonstationary,

$$u_1 \colon A_1 \times A_2 \times \Xi_1 \times \mathbb{N}_0 \to \mathbb{R} \,.$$

Type $\xi_0 \in \Xi_1$ is the *normal type* of player 1, who happens to have a stationary payoff function, given by the stage game in the benchmark game of complete information,

$$u_1(a, \xi_0, t) = u_1(a) \quad \forall a \in A, \forall t \in \mathbb{N}_0 \,.$$

It is standard to think of the prior probability $\mu(\xi_0)$ as being relatively large, so the games of incomplete information are a seemingly small departure from the underlying game of complete information, though there is no requirement that this be the case.

Action types (also called *commitment* or *behavioral types*) do not have payoffs, and simply play a specified repeated game strategy. For any repeated-game strategy from the complete information game, $\hat{\sigma}_1 \colon \mathcal{H}_1 \to \Delta(A_1)$, denote by $\xi(\hat{\sigma}_1)$ the action type committed to the strategy $\hat{\sigma}_1$. In general, a commitment type of player 1 can be committed to any strategy in the repeated game. If the strategy in question plays the same (pure or mixed) stage-game action in every period, regardless of history, that type is called a *simple action type*. For example, the H-action type in the product-choice game is a simple action type. The (simple action) type that plays the pure action a_1 in every period is denoted by $\xi(a_1)$ and similarly the simple action type committed to $\alpha_1 \in \Delta(A_1)$ is denoted by $\xi(\alpha_1)$. As will be seen soon, allowing for mixed action types is an important generalization from simple pure types.

A strategy for player 1, also denoted by $\sigma_1 \colon \mathcal{H}_1 \times \Xi \to \Delta(A_1)$, specifies for each type $\xi \in \Xi$ a repeated game strategy such that for all $\xi(\hat{\sigma}_1) \in \Xi_2$, the strategy $\hat{\sigma}_1$ is specified. A strategy σ_2 for player 2 is as in the complete information game, i. e., $\sigma_2 \colon \mathcal{H} \to \Delta(A_2)$.

Definition 2 The strategy profile (σ_1^*, σ_2^*) is a *Nash equilibrium* of the incomplete information game if

1. for all $\xi \in \Xi_1$, there does not exist a repeated game strategy σ_1 yielding a strictly higher payoff for payoff type ξ of player 1 when player 2 plays σ_2^*, and
2. in all periods t, after any history $h^t \in \mathcal{H}$ arising with positive probability under (σ_1^*, σ_2^*) and μ, $\sigma_2^*(h^t)$ maximizes $E[u_2(\sigma_1^*(h_1^t, \xi), a_1) \mid h^t]$, where the expectation is taken over both the period t-private histories that player 1 may have observed and player 1's type.

Example 1 Consider the product-choice game (Fig. 2) under perfect monitoring. The firm is willing to commit to H to induce h from customers. This incentive to commit is best illustrated by considering a sequential version of the product-choice game: The firm first publicly commits to an effort, and then the customer chooses between h and ℓ, knowing the firm's choice. In this sequential game, the firm chooses H in the unique subgame perfect equilibrium. Since Stackelberg [23] was the first investigation of such leader-follower interactions, it is traditional to call H the Stackelberg action, and the H-action type of player 1 the Stackelberg type, with associated Stackelberg payoff 2. Suppose $\Xi = \{\xi_0, \xi(H), \xi(L)\}$. For $\delta \geq 1/2$, the grim trigger strategy profile of always playing Hh, with deviations punished by Nash reversion, is a subgame perfect equilibrium of the complete information game. Consider the following adaptation of this profile in the incomplete information game:

$$\sigma_1(h^t, \xi) = \begin{cases} H, & \text{if } \xi = \xi(H), \\ & \quad \text{or } \xi = \xi_0 \text{ and } a^\tau = Hh \\ & \qquad\qquad\qquad \text{for all } \tau < t, \\ L, & \text{otherwise,} \end{cases}$$

and

$$\sigma_2(h^t) = \begin{cases} h, & \text{if } a^\tau = Hh \text{ for all } \tau < t, \\ \ell, & \text{otherwise} \,. \end{cases}$$

In other words, player 2 and the normal type of player 1 follow the strategies from the Nash-reversion equilibrium in the complete information game, and the action types $\xi(H)$ and $\xi(L)$ play their actions.

This is a Nash equilibrium for $\delta \geq 1/2$ and $\mu(\xi(L)) < 1/2$. The restriction on $\mu(\xi(L))$ ensures that player 2 finds h optimal in period 0. Should player 2 ever observe L, then Bayes' rule causes her to place probability 1 on type $\xi(L)$ (if L is observed in the first period) or the normal type (if L is first played in a subsequent period), making her participation in Nash reversion optimal. The restriction on δ ensures that Nash reversion provides sufficient incentive to make H optimal for the normal player 1. After observing $a_1^0 = H$ in period 0, player 2 assigns zero probability to $\xi = \xi(L)$. However, the posterior probability that 2 assigns to the Stackelberg type does not converge to 1. In period 0, the prior probability is $\mu(\xi(H))$.

After one observation of H, the posterior increases to $\mu(\xi^*)/[\mu(\xi^*)+\mu(\xi_0)]$, after which it is constant. By stipulating that an observation of H in a history in which L has previously been observed causes player 2 to place probability one on the normal type of player 1, a specification of player 2's beliefs that is consistent with sequentiality is obtained.

As seen in the introduction, for δ close to 1, $\sigma_1(h^t, \xi_0) = L$ for all h^t is not part of any Nash equilibrium.

The Reputation Bound

Which type would the normal type most like to be treated as? Player 1's *pure-action Stackelberg payoff* is defined as

$$v_1^* = \sup_{a_1 \in A_1} \min_{\alpha_2 \in B(a_1)} u_1(a_1, \alpha_2) , \tag{2}$$

where $B(a_1) = \arg\max_{a_2} u_2(a_1, a_2)$ is the set of player 2 myopic best replies to a_1. If the supremum is achieved by some action a_1^*, that action is an associated Stackelberg action,

$$a_1^* \in \arg\max_{a_1 \in A_1} \min_{\alpha_2 \in B(a_1)} u_1(a_1, \alpha_2) .$$

This is a pure action to which player 1 would commit, if player 1 had the chance to do so (and hence the name "Stackelberg" action, see the discussion in Example 1), given that such a commitment induces a best response from player 2. If there is more than one such action for player 1, the action can be chosen arbitrarily.

However, player 1 would typically prefer to commit to a mixed action. In the product-choice game, for example, a commitment by player 1 to mixing between H and L, with slightly larger probability on H, still induces player 2 to choose h and gives player 1 a larger payoff than a commitment to H. Define the mixed-action Stackelberg payoff as

$$v_1^{**} \equiv \sup_{\alpha_1 \in \Delta(A_1)} \min_{\alpha_2 \in B(\alpha_1)} u_1(\alpha_1, \alpha_2) , \tag{3}$$

where $B(\alpha_1) = \arg\max_{a_2} u_2(\alpha_1, a_2)$ is the set of player 2's best responses to α_1. In the product-choice game, $v_1^* = 2$, while $v_1^{**} = 5/2$. Typically, the supremum is not achieved by any mixed action, and so there is no mixed-action Stackelberg type. However, there are mixed action types that, if player 2 is convinced she is facing such a type, will yield payoffs arbitrarily close to the mixed-action Stackelberg payoff.

As with imperfect monitoring, simple mixed action types under perfect monitoring raise issues of monitoring, since a deviation by the normal type from the distribution α_1 of a mixed action type $\xi(\alpha_1)$, to some action in the support cannot be detected. However, when monitoring of the pure actions is perfect, it is possible to statistically detect deviations, and this will be enough to imply the appropriate reputation lower bound.

When monitoring is imperfect, the public signals are statistically informative about the actions of the long-lived player under the next assumption (Lemma 1).

Assumption 1 *For all $a_2 \in A_2$, the collection of probability distributions $\{\rho(y \mid (a_1, a_2) \colon a_1 \in A_1\}$ is linearly independent.*

This assumption is trivially satisfied in the perfect monitoring case. Reputation effects still exist when this assumption fails, but the bounds are more complicated to calculate (see [12] or Sect. 15.4.1 in [17]).

Fixing an action for player 2, a_2, the mixed action α_1 implies the signal distribution $\sum_{a_1} \rho(y \mid (a_1, a_2))\alpha_1(a_1)$.

Lemma 1 *Suppose ρ satisfies Assumption 1. Then, if for some a_2,*

$$\sum_{a_1} \rho(y \mid (a_1, a_2))\alpha_1(a_1) = \sum_{a_1} \rho(y \mid (a_1, a_2))\alpha_1'(a_1), \quad \forall y , \tag{4}$$

then $\alpha_1 = \alpha_1'$.

Proof Suppose (4) holds for some a_2. Let R denote the $|Y| \times |A_1|$ matrix whose y-a_1 element is given by $\rho(y \mid (a_1, a_2))$ (so that the a_1-column is the probability distribution on Y implied by the action profile $a_1 a_2$). Then, (4) can be written as $R\alpha_1 = R\alpha_1'$, or more simply as $R(\alpha_1 - \alpha_1') = 0$. By Assumption 1, R has full column rank, and so $x = 0$ is the only vector $x \in \mathbb{R}^{|A_1|}$ solving $Rx = 0$. □

Consequently, if player 2 believes that the long-lived player's behavior implies a distribution over the signals close to the distribution implied by some particular action α_1', then player 2 must believe that the long-lived player's action is also close to α_1'. Since A_2 is finite, this then implies that when player 2 is best responding to some belief about the long-lived player's behavior implying a distribution over signals sufficiently close to the distribution implied by α_1', then player 2 is in fact best responding to α_1'.

We are now in a position to state the main reputation bound result. Let $\underline{v}_1(\xi_0, \mu, \delta)$ be the infimum over the set of the normal player 1's payoffs in any (pure or mixed) Nash equilibrium in the incomplete information repeated game, given the distribution μ over types and the discount factor δ.

Proposition 1 (Fudenberg and Levine [11,12]) *Suppose ρ satisfies Assumption 1 and let $\hat{\xi}$ denote the simple action type that always plays $\hat{\alpha}_1 \in \Delta(A_1)$. Suppose $\mu(\xi_0)$, $\mu(\hat{\xi}) > 0$. For every $\eta > 0$, there is a value K such that for all δ,*

$$\underline{v}_1(\xi_0, \mu, \delta) \geq (1-\eta)\delta^K \min_{\alpha_2 \in B(\hat{\alpha}_1)} u_1(\hat{\alpha}_1, \alpha_2) + (1-(1-\eta)\delta^K) \min_{a \in A} u_1(a) . \quad (5)$$

This immediately yields the pure action Stackelberg reputation bound. Fix $\varepsilon > 0$. Taking $\hat{\alpha}_1$ in the proposition as the degenerate mixture that plays the Stackelberg action a_1^* with probability 1, Eq. (5) becomes

$$\begin{aligned}\underline{v}_1(\xi_0, \mu, \delta) &\geq (1-\eta)\delta^K v_1^* + (1-(1-\eta)\delta^K) \min_{a \in A} u_1(a) \\ &\geq v_1^* - (1-(1-\eta)\delta^K) 2M ,\end{aligned}$$

where $M \equiv \max_a |u_1(a)|$. This last expression is at least as large as $v_1^* - \varepsilon$ when $\eta < \varepsilon/(2M)$ and δ is sufficiently close to 1.

The mixed action Stackelberg reputation bound is also covered:

Corollary 1 *Suppose ρ satisfies Assumption 1 and μ assigns positive probability to some sequence of simple types $\{\xi(\alpha_1^k)\}_{k=1}^{\infty}$ with each α_1^k in $\Delta(A_1)$ satisfying*

$$v_1^{**} = \lim_{k \to \infty} \min_{\alpha_2 \in B(\alpha_1^k)} u_1(\alpha_1^k, \alpha_2) .$$

For all $\varepsilon' > 0$, there exists $\underline{\delta} < 1$ such that for all $\delta \in (\underline{\delta}, 1)$,

$$\underline{v}_1(\xi_0, \mu, \delta) \geq v_1^{**} - \varepsilon' .$$

The remainder of this subsection outlines a proof of Proposition 1. Fix a strategy profile (σ_1, σ_2) (which may be Nash, but at this point of the discussion, need not be). The beliefs μ then induce a probability distribution $\mathbf{P}$ on the set of outcomes, which is the set of possible infinite histories (denoted by h^∞) and realized types, $(Y \times A)^\infty \times \Xi \equiv \Omega$. The probability measure $\mathbf{P}$ describes how the short-lived players believe the game will evolve, given their prior beliefs μ about the types of the long-lived player. Let $\hat{\mathbf{P}}$ denote the probability distribution on the set of outcomes induced by (σ_1, σ_2) and the action type $\hat{\xi}$. The probability measure $\hat{\mathbf{P}}$ describes how the short-lived players believe the game will evolve if the long-lived player's type is $\hat{\xi}$. Finally, let $\tilde{\mathbf{P}}$ denote the probability distribution on the set of outcomes induced by (σ_1, σ_2) conditioning on the long-lived player's type not being the action type $\hat{\xi}$. Then, $\mathbf{P} \equiv \hat{\mu}\hat{\mathbf{P}} + (1-\hat{\mu})\tilde{\mathbf{P}}$, where $\hat{\mu} \equiv \mu(\hat{\xi})$.

The discussion after Lemma 1 implies that the optimal behavior of the short-lived player in period t is determined by that player's beliefs over the signal realizations in that period. These beliefs can be viewed as a *one-step ahead prediction* of the signal y that will be realized conditional on the history h^t, $\mathbf{P}(y \mid h^t)$. Let $\hat{\mu}^t(h^t) = \mathbf{P}(\hat{\xi} \mid h^t)$ denote the posterior probability after observing h^t that the short-lived player assigns to the long-lived player having type $\hat{\xi}$. Note also that if the long-lived player is the action type $\hat{\xi}$, then the true probability of the signal y is $\hat{\mathbf{P}}(y \mid h^t) = \rho\big(y \mid (H, \sigma_2(h^t))\big)$. Then,

$$\mathbf{P}(y \mid h^t) = \hat{\mu}^t(h^t)\hat{\mathbf{P}}(y \mid h^t) + (1-\hat{\mu}^t(h^t))\tilde{\mathbf{P}}(y \mid h^t) .$$

The key step in the proof of Proposition 1 is a statistical result on merging. The following lemma essentially says that the short-lived players cannot be surprised too many times. Note first that an infinite public history h^∞ can be thought of as a sequence of ever longer finite public histories h^t. Consider the collection of infinite public histories with the property that player 2 often sees histories h^t that lead to very different one-step ahead predictions about the signals under $\tilde{\mathbf{P}}$ and under $\hat{\mathbf{P}}$ and have a "low" posterior that the long-lived player is $\hat{\xi}$. The lemma asserts that if the long-lived player is in fact the action type $\hat{\xi}$, this collection of infinite public histories has low probability. Seeing the signals more likely under $\hat{\xi}$ leads the short-lived players to increase the posterior probability on $\hat{\xi}$. The posterior probability fails to converge to 1 under $\hat{\mathbf{P}}$ only if the play of the types different from $\hat{\xi}$ leads, on average, to a signal distribution similar to that implied by $\hat{\xi}$. For the purely statistical statement and its proof, see Section 15.4.2 in [17].

Lemma 2 *For all $\eta, \psi > 0$ and $\mu^\dagger \in (0, 1]$, there exists a positive integer K such that for all $\mu(\hat{\xi}) \in [\mu^\dagger, 1)$, for every strategy $\sigma_1 \colon \mathcal{H}_1 \times \Xi \to \Delta(A_1)$ and $\sigma_2 \colon \mathcal{H} \to \Delta(A_2)$,*

$$\hat{\mathbf{P}}\Big(h^\infty \colon \big|\big\{t \geq 1 \colon (1-\hat{\mu}^t(h^t)) \max_y \big|\tilde{\mathbf{P}}(y \mid h^t) - \hat{\mathbf{P}}(y \mid h^t)\big| \geq \psi\big\}\big| \geq K\Big) \leq \eta . \quad (6)$$

Note that the bound K holds for all strategy profiles (σ_1, σ_2) and all prior probabilities $\mu(\hat{\xi}) \in [\mu^\dagger, 1)$. This allows us to bound equilibrium payoffs.

Proof of Proposition 1 Fix $\eta > 0$. From Lemma 1, by choosing ψ sufficiently small in Lemma 2, with $\hat{\mathbf{P}}$-probability at least $1 - \eta$, there are at most K periods in which the short-lived players are not best responding to $\hat{\alpha}_1$.

Since a deviation by the long-lived player to the simple strategy of always playing $\hat{\alpha}_1$ induces the same distribution on public histories as $\hat{\mathbf{P}}$, the long-lived player's expected payoff from such a deviation is bounded below by the right side of (5). □

Temporary Reputation Effects

Under perfect monitoring, there are often pooling equilibria in which the normal and some action type of player 1 behave identically on the equilibrium path (as in Example 1). Deviations on the part of the normal player 1 are deterred by the prospect of the resulting punishment. Under imperfect monitoring, such pooling equilibria do not exist. The normal and action types may play identically for a long period of time, but the normal type always eventually has an incentive to cheat at least a little on the commitment strategy, contradicting player 2's belief that player 1 will exhibit commitment behavior. Player 2 must then eventually learn player 1's type.

In addition to Assumption 1, disappearing reputation effects require full support monitoring.

Assumption 2 *For all $a \in A$, $y \in Y$, $\rho(y \mid a) > 0$.*

This assumption implies that Bayes' rule determines the beliefs of player 2 about the type of player 1 after *all* histories.

Suppose there are only two types of player 1, the normal type ξ_0 and a simple action type $\hat{\xi}$, where $\hat{\xi} = \xi(\hat{\alpha}_1)$ for some $\hat{\alpha}_1 \in \Delta(A_1)$. The analysis is extended to many commitment types in Section 6.1 in Cripps et al. [8]. It is convenient to denote a strategy for player 1 as a pair of functions $\tilde{\sigma}_1$ and $\hat{\sigma}_1$ (so $\hat{\sigma}_1(h_1^t) = \hat{\alpha}_1$ for all $h_1^t \in \mathcal{H}_1$), the former for the normal type and the latter for the action type.

Recall that $\mathbf{P} \in \Delta(\Omega)$ is the unconditional probability measure induced by the prior μ, and the strategy profile $(\hat{\sigma}_1, \tilde{\sigma}_1, \sigma_2)$, while $\hat{\mathbf{P}}$ is the measure induced by conditioning on $\hat{\xi}$. Since $\{\xi_0\} = \Xi \setminus \{\hat{\xi}\}$, $\tilde{\mathbf{P}}$ is the measure induced by conditioning on ξ_0. That is, $\hat{\mathbf{P}}$ is induced by the strategy profile $\hat{\sigma} = (\hat{\sigma}_1, \sigma_2)$ and $\tilde{\mathbf{P}}$ by $\tilde{\sigma} = (\tilde{\sigma}_1, \sigma_2)$, describing how play evolves when player 1 is the commitment and normal type, respectively.

The action of the commitment type satisfies the following assumption.

Assumption 3 *Player 2 has a unique stage-game best response to $\hat{\alpha}_1$ (denoted by $\hat{a}_2$) and $\hat{\alpha} \equiv (\hat{\alpha}_1, \hat{a}_2)$ is not a stage-game Nash equilibrium.*

Let $\hat{\sigma}_2$ denote the strategy of playing the unique best response $\hat{a}_2$ to $\hat{\alpha}_1$ in each period independently of history. Since $\hat{\alpha}$ is not a stage-game Nash equilibrium, $(\hat{\sigma}_1, \hat{\sigma}_2)$ is not a Nash equilibrium of the complete information infinite horizon game.

Proposition 2 (Cripps, Mailath and Samuelson [8]) *Suppose the monitoring distribution ρ satisfies Assumptions 1 and 2, and the commitment action $\hat{\alpha}_1$ satisfies Assumption 3. In any Nash equilibrium of the game with incomplete information, the posterior probability assigned by player 2 to the commitment type, $\hat{\mu}^t$, converges to zero under $\tilde{\mathbf{P}}$, i. e.,*

$$\hat{\mu}^t(h^t) \to 0, \qquad \tilde{\mathbf{P}}\text{-}a.s.$$

The intuition is straightforward: Suppose there is a Nash equilibrium of the incomplete information game in which both the normal and the action type receive positive probability in the limit (on a positive probability set of histories). On this set of histories, player 2 cannot distinguish between signals generated by the two types (otherwise player 2 could ascertain which type she is facing), and hence must believe that the normal and action types are playing the same strategies on average. But then player 2 must play a best response to this strategy, and hence to the action type. Since the action type's behavior is not a best response for the normal type (to this player 2 behavior), player 1 must eventually find it optimal to *not* play the action-type strategy, contradicting player 2's beliefs.

Assumption 3 requires a unique best response to $\hat{\alpha}_1$. For example, in the product-choice game, every action for player 2 is a best response to player 1's mixture α_1' that assigns equal probability to H and L. This indifference can be exploited to construct an equilibrium in which (the normal) player 1 plays α_1' after every history (Section 7.6.2 in [17]). This will still be an equilibrium in the game of incomplete information in which the commitment type plays α_1', with the identical play of the normal and commitment types ensuring that player 2 never learns player 1's type. In contrast, player 2 has a unique best response to any other mixture on the part of player 1. Therefore, if the commitment type is committed to any mixed action other than α_1', player 2 will eventually learn player 1's type.

As in Proposition 1, a key step in the proof of Proposition 2 is a purely statistical result on updating. Either player 2's expectation (given her history) of the strategy played by the normal type ($\tilde{E}[\tilde{\sigma}_1^t \mid h^t]$, where $\tilde{E}$ denotes expectation with respect to $\tilde{\mathbf{P}}$) is in the limit identical to the strategy played by the action type ($\hat{\alpha}_1$), or player 2's posterior probability that player 1 is the action type ($\hat{\mu}^t(h^t)$) converges to zero (given that player 1 is indeed

normal). This is a merging argument and closely related to Lemma 2. If the distributions generating player 2's signals are different for the normal and action type, then these signals provide information that player 2 will use in updating her posterior beliefs about the type she faces. This (converging, since beliefs are a martingale) belief can converge to an interior probability only if the distributions generating the signals are asymptotically uninformative, which requires that they be asymptotically identical.

Lemma 3 *Suppose the monitoring distribution ρ satisfies Assumptions 1 and 2. Then in any Nash equilibrium,*

$$\lim_{t\to\infty} \hat{\mu}^t \max_{a_1} \left|\hat{\alpha}_1(a_1) - \tilde{E}[\,\tilde{\sigma}_1^t(a_1) \mid h^t\,]\right| = 0, \qquad \tilde{\mathbf{P}}\text{-}a.s. \quad (7)$$

Given Proposition 2, it should be expected that continuation play converges to an equilibrium of the complete information game, and this is indeed the case. See Theorem 2 [8] for the formal statement.

Proposition 2 leaves open the possibility that for any period T, there may be equilibria in which uncertainty about player 1's type survives beyond T, even though such uncertainty asymptotically disappears in any equilibrium. This possibility cannot arise. The existence of a sequence of Nash equilibria with uncertainty about player 1's type persisting beyond period $T \to \infty$ would imply the (contradictory) existence of a limiting Nash equilibrium in which uncertainty about player 1's type persists.

Proposition 3 (Cripps, Mailath and Samuelson [9]) *Suppose the monitoring distribution ρ satisfies Assumptions 1 and 2, and the commitment action $\hat{\alpha}_1$ satisfies Assumption 3. For all $\varepsilon > 0$, there exists T such that for any Nash equilibrium of the game with incomplete information,*

$$\tilde{\mathbf{P}}(\hat{\mu}^t < \varepsilon,\ \forall t > T) > 1 - \varepsilon\,.$$

Example 2 Recall that in the product-choice game, the unique player 2 best response to H is to play h, and Hh is not a stage-game Nash equilibrium. Proposition 1 ensures that the normal player 1's expected value in the repeated game of incomplete information with the H-action type is arbitrarily close to 2, when player 1 is very patient. In particular, if the normal player 1 plays H in every period, then player 2 will at least eventually play her best response of h. If the normal player 1 persisted in mimicking the action type by playing H in each period, this behavior would persist indefinitely. It is the feasibility of such a strategy that lies at the heart of the reputation bounds on expected payoffs. However, this strategy is not optimal. Instead, player 1 does even better by attaching some probability to L, occasionally reaping the rewards of his reputation by earning a stage-game payoff even larger than 2. The result of such equilibrium behavior, however, is that player 2 must eventually learn player 1's type. The *continuation* payoff is then bounded below 2 (recall (1)).

Reputation effects arise when player 2 is uncertain about player 1's type, and there may well be a long period of time during which player 2 is sufficiently uncertain of player 1's type (relative to the discount factor), and in which play does not resemble an equilibrium of the complete information game. Eventually, however, such behavior must give way to a regime in which player 2 is (correctly) convinced of player 1's type.

For any prior probability $\hat{\mu}$ that the long-lived player is the commitment type and for any $\varepsilon > 0$, there is a discount factor δ sufficiently large that player 1's expected payoff is close to the commitment-type payoff. This holds no matter how small $\hat{\mu}$. However, for any fixed δ and in any equilibrium, there is a time at which the posterior probability attached to the commitment type has dropped below the corresponding critical value of $\hat{\mu}$, becoming too small (relative to δ) for reputation effects to operate.

A reasonable response to the results on disappearing reputation effects is that a model of *long-run* reputations should incorporate some mechanism by which the uncertainty about types is continually replenished. For example, Holmström [13], Cole, Dow and English [6], Mailath and Samuelson [16], and Phelan [19] assume that the type of the long-lived player is governed by a stochastic process rather than being determined once and for all at the beginning of the game. In such a situation, reputation effects can indeed have long-run implications.

Reputation as a State

The posterior probability that short-lived players assign to player 1 being $\hat{\xi}$ is sometimes interpreted as player 1's reputation, particularly if $\hat{\xi}$ is the Stackelberg type. When Ξ contains only the normal type and $\hat{\xi}$, the posterior belief $\hat{\mu}^t$ is a state variable of the game, and attention is sometimes restricted to Markov strategies (i. e., strategies that only depend on histories through their impact on the posterior beliefs of the short-lived players). An informative example is Benabou and Laroque [2], who study the Markov perfect equilibria of a game in which the uninformed players respond continuously to their beliefs. They show that the informed player eventually reveals his type in any Markov perfect equilibrium. On the other hand, Markov equilibria

need not exist in finitely repeated reputation games (Section 17.3 in [17]).

The literature on reputation effects has typically not restricted attention to Markov strategies, since the results do not require the restriction.

Two Long-Lived Players

The introduction of nontrivial intertemporal incentives for the uninformed player significantly reduces reputation effects. For example, when only simple Stackelberg types are considered, the Stackelberg payoff may not bound equilibrium payoffs. The situation is further complicated by the possibility of non-simple commitment types (i. e., types that follow nonstationary strategies).

Consider applying the logic from Sect. "The Reputation Bound" to obtain the Stackelberg reputation bound when both players are long-lived and player 1's characteristics are unknown, under perfect monitoring. The first step is to demonstrate that, if the normal player 1 persistently plays the Stackelberg action and there exists a type committed to that action, then player 2 must eventually attach high probability to the event that the Stackelberg action is played in the future. This argument, a simple version of Lemma 2, depends only upon the properties of Bayesian belief revision, independently of whether the person holding the beliefs is a long-lived or short-lived player.

When player 2 is short-lived, the next step is to note that if she expects the Stackelberg action, then she will play a best response to this action. If player 2 is instead a long-lived player, she may have an incentive to play something other than a best response to the Stackelberg type.

The key step when working with two long-lived players is thus to establish conditions under which, as player 2 becomes increasingly convinced that the Stackelberg action will appear, player 2 must eventually play a best response to that action. One might begin such an argument by observing that, as long as player 2 discounts, any losses from not playing a current best response must be recouped within a finite length of time. But if player 2 is "very" convinced that the Stackelberg action will be played not only now but for sufficiently many periods to come, there will be no opportunity to accumulate subsequent gains, and hence player 2 might just as well play a stage-game best response.

Once it is shown that player 2 is best responding to the Stackelberg action, the remainder of the argument proceeds as in the case of a short-lived player 2. The normal player 1 must eventually receive very nearly the Stackelberg payoff in each period of the repeated game. By making player 1 sufficiently patient (*relative to player 2*, so that discount factors differ), this consideration dominates player 1's payoffs, putting a lower bound on the latter. Hence, the obvious handling of discount factors is to fix player 2's discount factor δ_2, and to consider the limit as player 1 becomes patient, i. e., δ_1 approaching one.

This intuition misses the following possibility. Player 2 may be choosing something other than a best response to the Stackelberg action out of fear that a current best response may trigger a disastrous future punishment. This punishment would not appear if player 2 faced the Stackelberg type, but player 2 can be made confident only that she faces the Stackelberg *action*, not the Stackelberg type. The fact that the punishment lies off the equilibrium path makes it difficult to assuage player 2's fear of such punishments. Short-lived players in the same situation are similarly uncertain about the future ramifications of best responding, but being short-lived, this uncertainty does not affect their behavior.

Consequently, reputation effects are typically weak with two long-lived players under perfect monitoring: Celentani, Fudenberg, Levine and Pesendorfer [3] and Cripps and Thomas [7], describe examples with only the normal and the Stackelberg types of player 1, in which the future play of the normal player 1 is used to punish player 2 for choosing a best response to the Stackelberg action when she is not supposed to, and player 1's payoff is significantly below the Stackelberg payoff. Moreover, the robustness of reputation effects to additional types beyond the Stackelberg type, a crucial feature of settings with one long-lived player, does not hold with two long-lived players. Schmidt [21] showed that the possibility of a "punishment" type can prevent player 2 best responding to the Stackelberg action, while Evans and Thomas [10] showed that the Stackelberg bound is valid if in addition to the Stackelberg type, there is an action type who punishes player 2 for *not* behaving appropriately (see Sections 16.1 and 16.5 in [17]).

Imperfect monitoring (of *both* players' actions), on the other hand, rescues reputation effects. With a sufficiently rich set of commitment types, player 1 can be assured of at least his Stackelberg payoff. Indeed, player 1 can often be assured of an even higher payoff, in the presence of commitment types who play nonstationary strategies [3]. At the same time, these reputation effects are temporary (Theorem 2 in [9]).

Finally, there is a literature on reputation effects in bargaining games (see [1,4,5,20]), where the issues described above are further complicated by the need to deal with the bargaining model itself.

Future Directions

The detailed structure of equilibria of the incomplete information game is not well understood, even for the canonical game of Sect. "A Canonical Model". A more complete description of the structure of equilibria is needed.

While much of the discussion was phrased in terms of the Stackelberg type, Proposition 1 provides a reputation bound for *any* action type. While in some settings, it is natural that the uninformed players assign strictly positive probability to the Stackelberg type, it is not natural in other settings. A model endogenizing the nature of action types would be an important addition to the reputation literature.

Finally, while the results on reputation effects with two long-lived players are discouraging, there is still the possibility that some modification of the model will rescue reputation effects in this important setting.

Acknowledgments

I thank Eduardo Faingold, KyungMin Kim, Antonio Penta, and Larry Samuelson for helpful comments.

Bibliography

1. Abreu D, Gul F (2000) Bargaining and Reputation. Econometrica 68(1):85–117
2. Benabou R, Laroque G (1992) Using Privileged Information to Manipulate Markets: Insiders, Gurus, and Credibility. Q J Econ 107(3):921–958
3. Celentani M, Fudenberg D, Levine DK, Pesendorfer W (1996) Maintaining a Reputation Against a Long-Lived Opponent. Econometrica 64(3):691–704
4. Chatterjee K, Samuelson L (1987) Bargaining with Two-Sided Incomplete Information: An Infinite Horizon Model with Alternating Offers. Rev Econ Stud 54(2):175–192
5. Chatterjee K, Samuelson L (1988) Bargaining with Two-Sided Incomplete Information: The Unrestricted Offers Case. Oper Res 36(4):605–638
6. Cole HL, Dow J, English WB (1995) Default, Settlement, and Signalling: Lending Resumption in a Reputational Model of Sovereign Debt. Int Econ Rev 36(2):365–385
7. Cripps MW, Thomas JP (1997) Reputation and Perfection in Repeated Common Interest Games. Games Econ Behav 18(2):141–158
8. Cripps MW, Mailath GJ, Samuelson L (2004) Imperfect Monitoring and Impermanent Reputations. Econometrica 72(2):407–432
9. Cripps MW, Mailath GJ, Samuelson L (2007) Disappearing Private Reputations in Long-Run Relationships. J Econ Theory 134(1):287–316
10. Evans R, Thomas JP (1997) Reputation and Experimentation in Repeated Games with Two Long-Run Players. Econometrica 65(5):1153–1173
11. Fudenberg D, Levine DK (1989) Reputation and Equilibrium Selection in Games with a Patient Player. Econometrica 57(4):759–778
12. Fudenberg D, Levine DK (1992) Maintaining a Reputation When Strategies Are Imperfectly Observed. Rev Econ Stud 59(3):561–579
13. Holmström B (1982) Managerial Incentive Problems: A Dynamic Perspective. In: Essays in Economics and Management in Honour of Lars Wahlbeck, pp 209–230. Swedish School of Economics and Business Administration, Helsinki. Published in: Rev Econ Stud 66(1):169–182
14. Kreps D, Wilson R (1982) Reputation and Imperfect Information. J Econ Theory 27:253–279
15. Kreps D, Milgrom PR, Roberts DJ, Wilson R (1982) Rational Cooperation in the Finitely Repeated Prisoner's Dilemma. J Econ Theory 27:245–252
16. Mailath GJ, Samuelson L (2001) Who Wants a Good Reputation? Rev Econ Stud 68(2):415–441
17. Mailath GJ, Samuelson L (2006) Repeated Games and Reputations: Long-Run Relationships. Oxford University Press, New York
18. Milgrom PR, Roberts DJ (1982) Limit pricing and entry under incomplete information: An equilibrium analysis. Econometrica 50:443–459
19. Phelan C (2006) Public Trust and Government Betrayal. J Econ Theory 130(1):27–43
20. Schmidt KM (1993) Commitment Through Incomplete Information in a Simple Repeated Bargaining Game. J Econ Theory 60(1):114–139
21. Schmidt KM (1993) Reputation and Equilibrium Characterization in Repeated Games of Conflicting Interests. Econometrica 61(2):325–351
22. Selten R (1978) Chain-store paradox. Theory Decis 9:127–59
23. Stackelberg HV (1934) Marktform und Gleichgewicht. Springer, Vienna

Resonances in Electronic Transport Through Quantum Wires and Rings

Vassilios Vargiamidis
Department of Physics, Aristotle University, Thessaloniki, Greece

Article Outline

Glossary
Definition of the Subject
Introduction
Electronic Transport Through Quantum Wires
Resonances in the Transmission Probability of a Quantum Wire
Electronic Transport Through Mesoscopic Open Rings
Future Directions
Bibliography

Glossary

Scattering Scattering is a general physical process whereby some forms of radiation, such as light, sound or moving particles, for example, are forced to deviate from a straight trajectory by one or more localized non-uniformities in the medium through which it passes. These non-uniformities are, sometimes, known as scatterers or scattering centers. In quantum transport, a scattering center may be provided by an impurity potential.

Transmission probability The transmission probability, denoted by T, for a quantum mechanical particle to pass through a scattering potential is the ratio between the flux of particles that emerges from the potential and the flux that arrives at it. Equivalently, T is the fraction of incident particles that succeed in passing through the scattering potential.

Transmission resonance Transmission resonance in electronic transport of quantum wires is an abrupt variation of the transmission probability that occurs over a very small interval the incident electron's energy.

Mode In a low-dimensional system where the electrons are confined in one or more directions, the energy eigenstates in the confinement direction(s) represent the modes of the system.

Evanescent mode When the electrons are free to propagate in one direction but are confined in the other directions, a mode whose energy is greater than the Fermi energy cannot propagate. This is called an evanescent mode.

Conductance quantization To measure the conductance of a sample (such as, a constriction) we divide the current through the sample by the electric potential difference between the reservoirs which are connected to the sample. If N_c is the number of occupied subbands, the expression for the conductivity pertaining to the two-terminal measurement is $G = 2e^2 N_c/h$. According to this description, the conductance is quantized such that it increases in steps by an amount equal to the quantum of conductance $2e^2/h$ whenever a new subband opens. This can be achieved either by widening of the constriction (and thus by lowering the subband energies) or by increasing the density of electrons (and thus by raising the Fermi energy).

Definition of the Subject

Since the 1980s advances in the growth techniques and new electronic materials developed therefrom have provided almost defect-free electronic devices, which have dimensions in one or more directions on the quantum scale. New quantum regimes governing such systems of lower dimensionality have led to novel electronic properties with potential applications. Quantum wells, wires, and dots, which have been implemented in the terminology of condensed-matter physics, indicate not only different dimensionality but also exhibit dramatically different electronic properties. Particularly, the electronic transport properties in lower dimensionality have several important features, which have stimulated a great deal of theoretical and experimental research. In electronic transport studies, if the size of the sample (or device) is smaller than the phase breaking length, the transport is phase coherent. In this case, electrons have a well-defined phase throughout the device, even though they may experience elastic scattering. Numerous publications on this type of transport have appeared, thus contributing to a field called mesoscopic physics, a term that indicates a new length scale for physical events between macroscopic and microscopic.

Phase coherent electronic transport through mesoscopic systems has been shown to exhibit resonant behavior as a function of the incident electron energy. This type of resonant behavior is of great interest both as a basis for the creation of new resonant nanoelectronic devices and for revealing a fundamental aspect of quantum mechanics. In various model calculations on electronic transport of microstructures, transmission resonances have been found to exhibit two types of resonance behavior: i) symmetric resonances (Breit–Wigner type), and ii) asymmetric resonances (Fano type). In particular, one often encounters points of vanishing transmission or reflection as a resonance is crossed. The interference effects that give rise to these resonances in the above-mentioned nanostructures are reviewed here, with emphasis given to the more recent interesting effects. This article is intended for those who are familiar with the field but the somewhat detailed derivations and explanations make it also accessible to beginners or researchers from similar disciplines.

Introduction

Since the early 1990s, resonances of the Fano type have been treated theoretically in various condensed matter systems including transport through quantum wires with attractive impurities or quantum dots [4,6,22,23,30,31,35,36,37,38,39,40]. In these systems, the coupling between a bound state of the impurity and the continuum leads to a quasibound (resonant) state. The interference between the direct (nonresonant) transmission pathway and the transmission via the quasibound state gives rise to an asymmetric Fano resonance. The Fano effect is a ubiqui-

tous phenomenon and has been observed in a large variety of experiments including atomic photoionization [13], neutron scattering [1], Raman scattering [8], optical absorption [11], and transport through mesoscopic systems with embedded quantum dots [5,16,19,24,26].

The purpose of this article is to review the resonance phenomena in the electronic transport of non-interacting electrons through: i) infinite rectilinear quantum wires with impurities, and ii) one-dimensional rings with impurities connected to current leads. In Sect. "Electronic Transport Through Quantum Wires", I will review the coupled-channel Feshbach theory, which is particularly suitable for describing the resonant behavior in ballistic electronic transport of quantum wires, and apply it to the case of an impurity with finite-range and smooth profile along the propagation direction. In Sect. "Resonances in the Transmission Probability of a Quantum Wire", it will be shown that varying the strength and size of the impurity causes the transformation of an asymmetric Fano resonance into a symmetric antiresonance (Breit–Wigner dip) and subsequently to an "inverted" Fano resonance. Thermal effects will also be considered. In Sect. "Electronic Transport Through Mesoscopic Open Rings", after reviewing the scattering matrix approach, I will discuss the Fano resonances in electronic transport of one-dimensional (abbreviated, 1D) open rings with a short-range (abbreviated, SR) impurity in one arm and the effects of the ring-lead coupling on the resonance structure. In this Section, an interesting feature of the Fano line shape will also be discussed; namely, the systematic collapse of certain Fano resonances for special impurity positions. The effect of an Aharonov–Bohm (abbreviated, AB) magnetic flux will also briefly be considered as well as the temperature dependence of the Fano effect. Finally, in Sect. "Future Directions", some thoughts on future directions will be given.

Electronic Transport Through Quantum Wires

A relatively new view was established for the quantum theory of scattering since it became clear that the electrical linear response of open conductors could be related to its transmission and reflection properties [2,28]. In addition, experiments performed at that time independently by van Wees et al. [49] and Wharam et al. [44] have been a breakthrough in the field of quantum ballistic transport in a quantum point contact (abbreviated, QPC) in a two-dimensional electron gas (abbreviated, 2D EG). Using high-mobility $GaAs - Al_{1-x}Ga_xAs$ heterojunctions and the split-gate technique, they imposed a small constriction on the sample. A channel was obtained from this constriction by applying a negative bias to the split gate, and thus by causing the depletion of electrons beneath the gate. Thus, the portion of the 2D EG lying below the gap of the split-gate electrode remains conducting. In their experiments the width of the constriction is in the range of the Fermi wavelength λ_F, whereby quantum size effects become relevant.

In the theoretical models [44,49] initially used to explain the electron conduction and the quantization of conductance, the QPC was perceived as a uniform waveguide, and only the events in this waveguide were taken into account. Here, the word uniform refers to the confining potential, which is the same throughout the constriction. The current-transporting states are laterally confined in the waveguide, the width of which is in the range of the Fermi wavelength. Then, the transverse momenta of these states are quantized, resulting in a subband structure. In the following we confine ourselves to such uniform, infinite long waveguide (or quantum wire) at zero temperature.

Feshbach Coupled-Channel Theory

We consider a uniform 2D quantum wire as shown in Fig. 1, in which electrons are confined along the y direction (transverse direction) but are free to move along the x direction (propagation direction). In the presence of a scattering potential the Schrödinger equation describing the electron motion in the wire is

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V_c(y) + V(x,y)\right]\Psi(x,y) = E\Psi(x,y), \quad (1)$$

where $V_c(y)$ is the confining potential, $V(x,y)$ is a scattering potential, and m is the effective mass of the electron. The transverse potential $V_c(y)$, providing confinement of the electron motion along the y direction, gives rise to channel modes $\phi_n(y)$ which satisfy

$$\left[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial y^2} + V_c(y)\right]\phi_n(y) = E_n\phi_n(y)\,, \quad (2)$$

where E_n is the threshold energy for the nth mode. The wave propagation (or transport) in subband n takes place, as long as the threshold E_n (i. e., the bottom of the nth subband) is smaller than the Fermi level E (see Fig. 2). We expand the wave function $\Psi(x,y)$ of Eq. (1) in terms of the channel modes as

$$\Psi(x,y) = \sum_{n=1}^{\infty}\psi_n(x)\phi_n(y)\,. \quad (3)$$

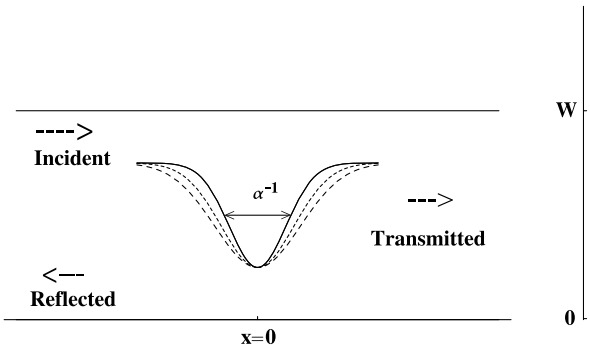

Resonances in Electronic Transport Through Quantum Wires and Rings, Figure 1
A uniform quantum wire of width W. An incident electron wave is partially reflected and transmitted through a Pöschl–Teller attractive impurity $V(x, y) = (\hbar^2\gamma/2m)\,\mathrm{sech}^2(\alpha x)\upsilon(y)$, where α^{-1} is the decay length and $\upsilon(y)$ is arbitrary function. Larger values of α^{-1} represent impurities of larger size, as shown by the *dashed lines*

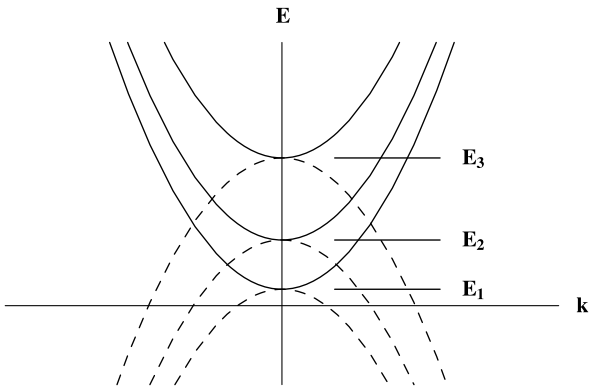

Resonances in Electronic Transport Through Quantum Wires and Rings, Figure 2
Dispersion relation of propagating (*solid*) and evanescent (*dashed*) states in a 2D quantum wire $E = E_n + (\hbar^2 k_n^2/2m)$, where E_n define the threshold energies (bottom of subbands). Modes with energy E_n such that $E_n < E$, where $n = 1, 2, \ldots$, are propagating along the wire while modes with $E_n > E$ (that is, $k_n = i\kappa_n$) are evanescent

Substituting Eq. (3) into Eq. (1) we obtain the following coupled-channel equations for $\psi_n(x)$:

$$(E - E_n - \hat{K})\psi_n(x) = \sum_{l=1}^{\infty} V_{nl}(x)\psi_l(x)\,, \tag{4}$$

where $\hat{K} = -(\hbar^2/2m)d^2/dx^2$ and the coupling matrix elements $V_{nl}(x)$ are given as

$$V_{nl}(x) = \int \mathrm{d}y\phi_n(y)V(x, y)\phi_l(y)\,. \tag{5}$$

We confine ourselves now to a single-subband regime, i. e., $E_1 \le E \le E_2$ where only the first mode $n = 1$ can propagate along the wire while the higher (evanescent) modes can only contribute to electron propagation due to their coupling with the first by the scattering potential. Thus all coupling matrix elements $V_{nl}(x)$ for the higher modes vanish away from the scattering region ($x \to \pm\infty$) and therefore only the first mode $n = 1$ can actually be scattered and, therefore, be found in some scattering state. It can be seen from Eq. (4) that the scattering states for the first mode are given as solutions of the equation

$$\left[\hat{K} + V_{11}(x)\right]\chi_k^{\pm}(x) = (E - E_1)\chi_k^{\pm}(x)\,, \tag{6}$$

where we denoted by $\chi_k^{+}(x)$ and $\chi_k^{-}(x)$ the wave functions that correspond to scattering states for which the incident wave comes from $-\infty$ and $+\infty$, respectively. These wave functions describe the background (nonresonant) scattering, which is the scattering in a hypothetical system in which there is no coupling to a bound state [14,30,37] of the impurity. In Eq. (6) the wave vector of the propagating mode is given as $k = [2m(E - E_1)]^{1/2}/\hbar$ while $V_{11}(x)$ is an effective 1D potential for scattering of the first mode. The scattering wave functions in the asymptotic region can be expressed as

$$\chi_k^{\pm}(x) = \begin{cases} t^{bg}\mathrm{e}^{\pm ikx} & (x \to \pm\infty) \\ \mathrm{e}^{\pm ikx} + r_{\pm}^{bg}\mathrm{e}^{\mp ikx} & (x \to \mp\infty) \end{cases}\,. \tag{7}$$

In Eq. (7) the upper signs correspond to incident wave from $-\infty$, while t^{bg} and $r_{\pm}^{bg}$ correspond to the background transmission and reflection amplitudes in the quantum wire. Let E be close to the energy $\tilde{E}_0$ of the bound state $\Phi_0(x)$ in the potential $V_{22}(x)$ of the uncoupled channel $n = 2$. This bound state is found as the solution of

$$\left[\hat{K} + V_{22}(x)\right]\Phi_0(x) = (\tilde{E}_0 - E_2)\Phi_0(x)\,. \tag{8}$$

Assuming no other channels exhibit bound states close to $\tilde{E}_0$, we can make the approximation of truncating the sum of Eq. (4) at $n = 2$. We then obtain the following system of equations

$$\left[E - E_1 - \hat{K} - V_{11}(x)\right]\psi_1(x) = V_{12}(x)\psi_2(x)\,, \tag{9}$$

$$\left[E - E_2 - \hat{K} - V_{22}(x)\right]\psi_2(x) = V_{21}(x)\psi_1(x)\,. \tag{10}$$

The system of coupled-channel equations [i. e., Eqs. (9) and (10)] is solved using the ansatz [15]

$$\psi_2(x) = A\Phi_0(x)\,. \tag{11}$$

Inserting this ansatz into Eq. (9) we obtain an inhomogeneous equation for $\psi_1(x)$, which can then be solved by employing the retarded Green's operator

$$\hat{G}_1 = (E - E_1 - \hat{K} - V_{11} + i0^+)^{-1} . \tag{12}$$

When $\hat{G}_1$ acts on Eq. (9), the general solution to the inhomogeneous equation can be found as

$$|\psi_1\rangle = |\chi_k^+\rangle + A\hat{G}_1 V_{12}|\Phi_0\rangle . \tag{13}$$

Then, Eq. (10) can be written as

$$A(E - E_2 - \hat{K} - V_{22})|\Phi_0\rangle = V_{21}|\chi_k^+\rangle + AV_{21}\hat{G}_1 V_{12}|\Phi_0\rangle . \tag{14}$$

We can rewrite Eq. (14) with the help of Eq. (8) as

$$A(E - \tilde{E}_0)|\Phi_0\rangle = V_{21}|\chi_k^+\rangle + AV_{21}\hat{G}_1 V_{12}|\Phi_0\rangle . \tag{15}$$

Multiplying Eq. (15) by $\langle\Phi_0|$ allows us to determine A in Eq. (11) as

$$A = \frac{\langle\Phi_0|V_{21}|\chi_k^+\rangle}{E - \tilde{E}_0 - \langle\Phi_0|V_{21}G_1 V_{12}|\Phi_0\rangle} . \tag{16}$$

Using the scattering states $\chi_k^{\pm}(x)$, we can write the explicit form of the retarded Green's function in 1D as

$$G_1(x, x') = \frac{m}{i\hbar^2 k t^{bg}} \times \begin{cases} \chi_k^+(x)\chi_k^-(x') & (x > x') \\ \chi_k^+(x')\chi_k^-(x) & (x < x') \end{cases} \tag{17}$$

We then finally obtain from Eq. (13) the solution for $x \to \infty$ as

$$\psi_1(x) = \chi_k^+(x) + \frac{m}{i\hbar^2 k t^{bg}}\chi_k^+(x) \cdot \frac{\langle(\chi_k^-)^*|V_{12}|\Phi_0\rangle\langle\Phi_0|V_{21}|\chi_k^+\rangle}{E - \tilde{E}_0 - \langle\Phi_0|V_{21}G_1 V_{12}|\Phi_0\rangle} . \tag{18}$$

In Eq. (18) the matrix element $\langle\Phi_0|V_{21}G_1 V_{12}|\Phi_0\rangle$ is the self-energy due to the coupling of the bound state with the continuum and has in general both a real and an imaginary part.

The above formalism applies to a general scattering potential. In the following subsection we will examine a particular potential with finite range and smooth profile; namely, the Pöschl–Teller potential.

Pöschl–Teller Scattering Potential

We consider now a scattering potential of the Pöschl–Teller (abbreviated, PT) type

$$V(x, y) = \frac{\hbar^2\gamma}{2m}\,\mathrm{sech}^2(\alpha x)\upsilon(y) , \tag{19}$$

where $\upsilon(y)$ is an arbitrary function of the coordinate y, α^{-1} is the decay length, and γ sets the magnitude of the potential which is taken to be attractive ($\gamma < 0$). The longitudinal part of this potential has smooth profile as shown in Fig. 1. The matrix elements of Eq. (5) take the form

$$V_{nl}(x) = \frac{\hbar^2\gamma}{2m}\,\mathrm{sech}^2(\alpha x)\upsilon_{nl} , \tag{20}$$

where $\upsilon_{nl} = \langle\phi_n|\upsilon(y)|\phi_l\rangle$. In order to find the scattering states $\chi_k^{\pm}(x)$ we must solve Eq. (6) in the presence of an attractive potential

$$V_{11}(x) = -U_{11}\,\mathrm{sech}^2(\alpha x) , \tag{21}$$

where $U_{11} = -(\hbar^2\gamma/2m)\upsilon_{11}$. The solution proceeds in the same way as in a 1D scattering problem [27] and the asymptotic form of the wave function, as $x \to -\infty$, is written in terms of Gamma functions as

$$\chi_k^+(x) \sim \mathrm{e}^{-ikx}\frac{\Gamma(ik/\alpha)\Gamma(1 - ik/\alpha)}{\Gamma(-\eta)\Gamma(1 + \eta)} + \mathrm{e}^{ikx}\frac{\Gamma(-ik/\alpha)\Gamma(1 - ik/\alpha)}{\Gamma(-ik/\alpha - \eta)\Gamma(-ik/\alpha + \eta + 1)} , \tag{22}$$

where $\eta = (1/2)\left[-1 + \sqrt{1 + (8m^* U_{11}/\alpha^2\hbar^2)}\right]$. In Eq. (22) Γ is the Gamma function. The reflection amplitude r_+^{bg} is the ratio of coefficients of the function $\chi_k^+(x)$ in Eq. (22) and due to symmetry $r_-^{bg} = r_+^{bg}$ holds. Having found the scattering states we now proceed to find the bound state of the 1D potential $V_{22}(x) = -U_{22}\,\mathrm{sech}^2(\alpha x)$ by solving Eq. (8), where $U_{22} = -(\hbar^2\gamma/2m)\upsilon_{22}$. Employing the notation $\epsilon = [-2m(\tilde{E}_0 - E_2)]^{1/2}/\hbar\alpha$ and $s = (1/2)\left[-1 + \sqrt{1 + (8mU_{22}/\alpha^2\hbar^2)}\right]$, we can bring Eq. (8) into a form that has solutions the associated Legendre polynomials $P_s^{\epsilon}(\xi)$, where $\xi = \tanh(\alpha x)$. The energy levels are then obtained from the condition $\epsilon = s - p$ as

$$\tilde{E}_p = E_2 - \frac{\hbar^2\alpha^2}{8m}\left[-(1 + 2p) + \sqrt{1 + \frac{8mU_{22}}{\alpha^2\hbar^2}}\right]^2 , \tag{23}$$

where $p = 0, 1, 2, \ldots$ The number of levels is determined by the condition $\epsilon > 0$ (or $p < s$) and generally depends on the depth U_{22} of the potential as well as on the inverse decay length α. In the following we will assume that U_{22} and α are such that $s \leq 1$, which implies that there is only one bound state with energy $\tilde{E}_0$. When this bound state lies in the continuum of the first subband it is transformed to a quasibound state and acquires a finite width. The normalized bound state wave function is given as $\Phi_0(x) = (\alpha/2)^{1/2}\,\mathrm{sech}(\alpha x)$.

Having found the scattering states $\chi_k^{\pm}(x)$ as well as the bound state, $\tilde{E}_0$ and Φ_0, we now proceed to calculate the matrix elements that occur in Eq. (18) and thus find the transmission amplitude. We then have

$$\begin{aligned}\langle(\chi_k^-)^*|V_{12}|\Phi_0\rangle &= \frac{\hbar^2}{2m}\gamma\upsilon_{12}\sqrt{\frac{\alpha}{2}} \\ &\quad\times\int \mathrm{d}x\left(\mathrm{e}^{-ikx}+r_-^{bg}\mathrm{e}^{ikx}\right)\operatorname{sech}^3(\alpha x) \\ &= \frac{\hbar^2}{2m}\gamma\upsilon_{12}\sqrt{\frac{\alpha}{2}}\left(I_1^*+r_-^{bg}I_1\right),\end{aligned} \tag{24}$$

where

$$\begin{aligned}I_1 &= \int_{-\infty}^{\infty}\mathrm{d}x\mathrm{e}^{ikx}\operatorname{sech}^3(\alpha x), \\ I_1^* &= \int_{-\infty}^{\infty}\mathrm{d}x\mathrm{e}^{-ikx}\operatorname{sech}^3(\alpha x).\end{aligned} \tag{25}$$

It can be verified that $I_1 = I_1^* = [(\alpha^2+k^2)\pi\cdot\operatorname{sech}(k\pi/2\alpha)]/2\alpha^3$. Also,

$$\langle\Phi_0|V_{21}|\chi_k^+\rangle = \frac{\hbar^2}{2m}\gamma\upsilon_{21}\sqrt{\frac{\alpha}{2}}t^{bg}I_1. \tag{26}$$

We calculate now the matrix element that occurs in the denominator of Eq. (18). We have

$$\begin{aligned}&\langle\Phi_0|V_{21}G_1V_{12}|\Phi_0\rangle \\ &= \frac{m}{i\hbar^2kt^{bg}}\int_{-\infty}^{\infty}\mathrm{d}x\int_{-\infty}^{\infty}\mathrm{d}x'\Phi_0(x)V_{12}(x) \\ &\quad\cdot\Phi_0(x')V_{12}(x')\chi_k^-(x)\chi_k^+(x') \\ &\quad+\frac{m}{i\hbar^2kt^{bg}}\int_{-\infty}^{\infty}\mathrm{d}x\int_{-\infty}^{x}\mathrm{d}x'\Phi_0(x)V_{12}(x)\Phi_0(x') \\ &\quad\cdot V_{12}(x')[\chi_k^+(x)\chi_k^-(x')-\chi_k^-(x)\chi_k^+(x')] \\ &= Q_1+Q_2-Q_3.\end{aligned} \tag{27}$$

Inserting the explicit expressions for the bound state Φ_0, the coupling V_{12}, and the scattering states $\chi_k^{\pm}$ into Eq. (27) we obtain

$$Q_1 = \frac{m}{i\hbar^2k}\left(\frac{\hbar^2}{2m}\right)^2\gamma^2\upsilon_{12}^2\left(\frac{\alpha}{2}\right)\left[I_1^*I_1+r_-^{bg}(I_1)^2\right], \tag{28}$$

and

$$Q_2 = \frac{m}{i\hbar^2k}\left(\frac{\hbar^2}{2m}\right)^2\gamma^2\upsilon_{12}^2\left(\frac{\alpha}{2}\right)\left[I_2+r_-^{bg}I_3\right], \tag{29}$$

where

$$I_2 = \int_{-\infty}^{\infty}\mathrm{d}x\mathrm{e}^{ikx}\operatorname{sech}^3(\alpha x)\int_{-\infty}^{x}\mathrm{d}x'\mathrm{e}^{-ikx'}\operatorname{sech}^3(\alpha x'), \tag{30}$$

and

$$I_3 = \int_{-\infty}^{\infty}\mathrm{d}x\mathrm{e}^{ikx}\operatorname{sech}^3(\alpha x)\int_{-\infty}^{x}\mathrm{d}x'\mathrm{e}^{ikx'}\operatorname{sech}^3(\alpha x'). \tag{31}$$

For Q_3 we get the same expression as for Q_2 and therefore $Q_2 - Q_3 = 0$. Thus $\langle\Phi_0|V_{21}G_1V_{12}|\Phi_0\rangle = Q_1$. Using Eq. (28) and the results of the integrals I_1 and I_1^* we can finally write the matrix element as

$$\langle\Phi_0|V_{21}G_1V_{12}|\Phi_0\rangle = \delta - i\Gamma, \tag{32}$$

with

$$\begin{aligned}\delta = &-\frac{\hbar^2\gamma^2\upsilon_{12}^2(\alpha^2+k^2)^2\pi^2\operatorname{sech}^2(k\pi/2\alpha)}{32mk\alpha^5} \\ &\cdot\frac{cos\left[(\pi/2)\sqrt{1+(8mU_{11}/\hbar^2\alpha^2)}\right]\sinh(\pi k/\alpha)}{\sinh^2(\pi k/\alpha)+\cos^2\left[(\pi/2)\sqrt{1+(8mU_{11}/\hbar^2\alpha^2)}\right]},\end{aligned} \tag{33}$$

and

$$\begin{aligned}\Gamma = &\frac{\hbar^2\gamma^2\upsilon_{12}^2(\alpha^2+k^2)^2\pi^2\operatorname{sech}^2(k\pi/2\alpha)}{32mk\alpha^5} \\ &\cdot\frac{\sinh^2(\pi k/\alpha)}{\sinh^2(\pi k/\alpha)+\cos^2\left[(\pi/2)\sqrt{1+(8mU_{11}/\hbar^2\alpha^2)}\right]}.\end{aligned} \tag{34}$$

Equations (33) and (34) for δ and Γ give the shift and width respectively, that the bound state acquires. Furthermore, using Eqs. (24) and (26) we can write the numerator of Eq. (18) as

$$\frac{m}{i\hbar^2kt^{bg}}\langle(\chi_k^-)^*|V_{12}|\Phi_0\rangle\langle\Phi_0|V_{21}|\chi_k^+\rangle = \delta - i\Gamma. \tag{35}$$

Equation (18) can then finally be written for $x\to\infty$ as

$$\psi_1(x) = t^{bg}\mathrm{e}^{ikx}\frac{E-\tilde{E}_0}{E-\tilde{E}_0-\delta+i\Gamma}. \tag{36}$$

From Eq. (36) we can extract the transmission probability of the quantum wire as

$$T = |t^{bg}|^2\frac{\left(E-\tilde{E}_0\right)^2}{(E-E_R)^2+\Gamma^2}. \tag{37}$$

Due to the real part δ of the self-energy acquired by the bound state, a shifted quasibound state energy $E_R = \tilde{E}_0+\delta$

appears in Eq. (37). The background transmission coefficient, which enters Eq. (37), is found as

$$|t^{bg}|^2 = \frac{\sinh^2(\pi k/\alpha)}{\sinh^2(\pi k/\alpha)+\cos^2\left[(\pi/2)\sqrt{1+(8mU_{11}/\hbar^2\alpha^2)}\right]}. \tag{38}$$

Equation (37) can be transformed into the asymmetric Fano function [12] if we define reduced variables $\epsilon = (E - E_R)/\Gamma$ and $q = (E_R - \tilde{E}_0)/\Gamma = \delta/\Gamma$, where $\tilde{E}_0$ is the energy of the transmission zero and q is the asymmetry parameter of the resonance line shape. An interesting feature of Fano resonances in scattering from impurities in quantum wires is that the asymmetry parameter depends only on the characteristics of the impurity potential and not on the coupling of the quasibound state. In addition, the limit $q \to 0$ leads to a symmetric antiresonance (Breit–Wigner dip) while the opposite limit $q \to \infty$ leads to a resonance peak. We now define units as follows. For *GaAs* the value of $\hbar^2/2m \approx 0.57\,\text{eV}\,\text{nm}^2$. We take $\hbar^2/2m$ equal to unity and therefore the unit of the matrix element υ_{nl} can be taken to be 0.1 eV nm which then yields an energy unit of ≈ 17.7 meV and a length unit of ≈ 5.7 nm. Then, we can write the asymmetry parameter with the help of Eqs. (33) and (34) as

$$q = \frac{\delta}{\Gamma} = -\frac{\cos\left[(\pi/2)\sqrt{1+(4U_{11}/\alpha^2)}\right]}{\sinh(\pi k/\alpha)}. \tag{39}$$

The asymmetry parameter is a particularly important quantity in the detailed structure of the line shape and has been investigated experimentally [16,24] and theoretically [9].

Resonances in the Transmission Probability of a Quantum Wire

In this section we will discuss the influence of the fundamental parameters of the impurity discussed in Sect. "Electronic Transport Through Quantum Wires" on the resonance line shape. In the numerical calculations we take the electron mass to be the effective mass for *GaAs* which is 0.067 of the free electron mass.

Inversion of the Resonance Level

An important feature of Eq. (39) is that for certain ranges of values of U_{11}/α^2 the quantity δ (which is the difference between the resonance energy E_R and the energy $\tilde{E}_0$ of the transmission zero) can be negative, positive, or zero. This means that the resonance energy may occur before, after, or be equal to the energy of the transmission zero, which leads to an oscillatory behavior of the asymmetry parameter q. In particular, when $0 < U_{11}/\alpha^2 < 2$ we have $E_R > \tilde{E}_0$ and in this case the peak follows the dip. We call this a $0 \to 1$ type Fano resonance. However, for $2 < U_{11}/\alpha^2 < 6$ we have $E_R < \tilde{E}_0$ and the dip follows the peak. We call this a $1 \to 0$ type Fano resonance and, in this case, the resonance level appears inverted, which means that the the location of the pole is switched with the zero energy. For the interval $6 < U_{11}/\alpha^2 < 12$ we have $E_R > \tilde{E}_0$ and the Fano resonance is again of the $0 \to 1$ type. Exactly at the values of U_{11}/α^2, for which inversion of the resonance level occurs (i. e., at $U_{11}/\alpha^2 = 2, 6, \ldots$), δ and therefore q become zero. The special values, for which $\delta = 0$, are given as

$$\frac{U_{11}}{\alpha^2} = \frac{1}{4}[(2n+1)^2 - 1]\,, \tag{40}$$

where $n = 1, 2, \ldots$ Since $q = 0$ at these special values, the transmission exhibits symmetric Breit–Wigner dips.

These effects are illustrated in Fig. 3 where we show the transmission probability through the attractive potential of Eq. (19) plotted versus the Fermi energy over the first subband, i. e., $E_1 < E < E_2$. In Fig. 3a we have chosen $U_{11} = 1.5$, $U_{22} = 1.4$, $\upsilon_{12} = 0.09$, and $\alpha = 1.6$. The transmission exhibits an asymmetric Fano resonance (solid line) of the $0 \to 1$ type while the dashed line in Fig. 3a represents the direct (nonresonant) transmission, which occurs in the decoupling limit $\upsilon_{12} = 0$ (i. e., in the limit in which there is no coupling of the bound state with the continuum of states of the first subband) for which $\delta = \Gamma = 0$. In this limit there are two scattering mechanisms: a direct scattering from the first subband and a resonant scattering from the quasibound state. When the coupling to the quasibound level becomes nonzero the interference between direct and nonresonant transmission pathways produces the asymmetric Fano line shape [23,30,37,39]. The resonance width is proportional to the square of the coupling, i. e., $\Gamma \sim \upsilon_{12}^2$, as can be seen from Eq. (34).

In Fig. 3b we have chosen, for both solid and dashed lines, $\alpha = 1$, $U_{11} = 2$, and $U_{22} = 1$ but $\upsilon_{12} = 0.095$ has been chosen for plotting the solid line and $\upsilon_{12} = 0.14$ for the dashed line. These give $U_{11}/\alpha^2 = 2$, which is one of the special values of Eq. (40), and it can be seen that the transmission exhibits a completely symmetric antiresonance (Breit–Wigner dip). This is due to the fact that the transmission zero coincides with the energy of the resonance level. Furthermore, as mentioned above, the width of the resonance increases as the intersubband coupling υ_{12} increases, which is shown by the dashed line. It's also

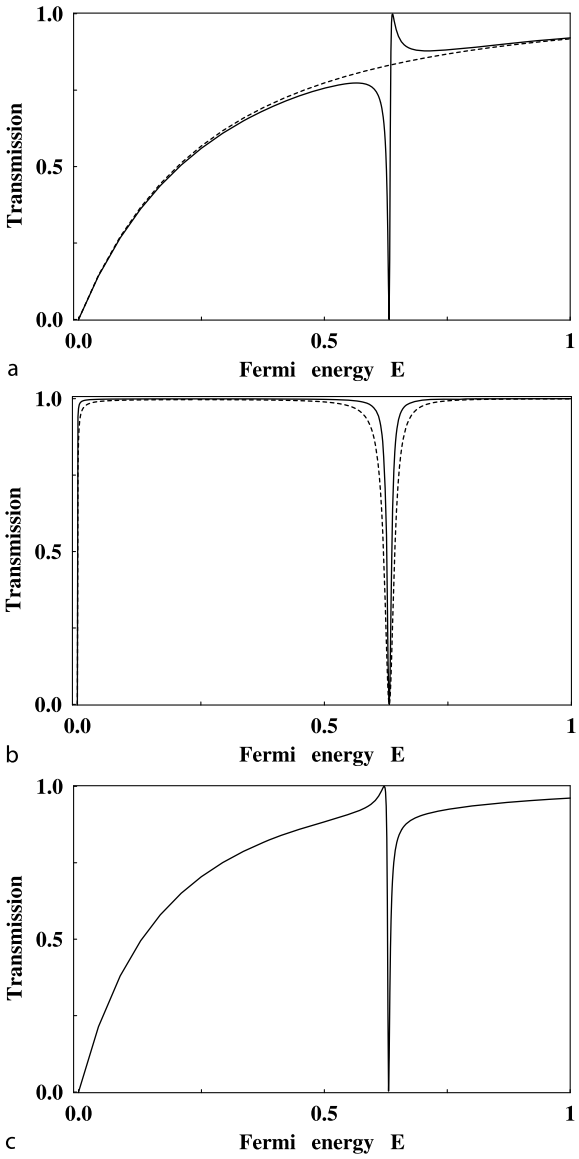

Resonances in Electronic Transport Through Quantum Wires and Rings, Figure 3

Transmission probability *T* versus Fermi energy *E* through an attractive impurity potential of the Pöschl–Teller type, in a 2D uniform quantum wire (see text for the numerical values of the parameters used). a An asymmetric Fano resonance of the 0 → 1 type (*solid*) and the background transmission (*dashed*). b symmetric Breit–Wigner antiresonance for one of the special values given in the text, at which the resonance energy coincides with the energy of the transmission zero. The *dashed line* corresponds to larger coupling parameter which gives rise to larger resonance width. c A 1 → 0 type Fano resonance indicating inversion of the resonance level

interesting to note that there is perfect transmission over the whole energy window, except for the very small energy interval where the antiresonance occurs. This is due to the fact that the scattering potential of Eq. (19) varies smoothly and the direct transmission $|t^{bg}|^2$ becomes unity [see Eq. (38)] whenever U_{11}/α^2 assumes the special values given in Eq. (40). In Fig. 3c, we have chosen $\alpha = 1.4$, $U_{11} = 6.3$, $U_{22} = 1.2$, and $\upsilon_{12} = 0.8$. The resonance line shape is now of the $1 \to 0$ type. The inversion of the resonance level indicates that the roles of destructive and constructive interference between the direct and resonant channels have been reversed.

The resonance level inversion discussed in the context of Fig. 3, has also been predicted in References [23] and [22] where the scattering potential was modeled by a rectangular square well. Increasing the longitudinal size of the impurity, it was found that there are some critical values of the impurity size for which the Fano resonance disappears. This occurs over an energy interval around a critical value E_c where a transformation of the quasi-bound state into a true bound state (for which $\Gamma \to 0$) occurs, giving rise to unity transmission in this interval.

We also mention that the Fano asymmetry parameter q is a particularly important quantity for the detailed structure of the line shape, and has been investigated experimentally and theoretically in the context of various types of mesoscopic systems [9,14,16,24]. In our case, q can be evaluated at the resonance energy from Eq. (39). The calculated values are $q = 0.47$, 0, and -0.43 in Figs. 3a, b, and c, respectively. Furthermore, we note that resonance phenomena in quantum wires with impurities can also lead to a connection between the line shape and the fundamental parameters of the impurity, such as its strength and size.

Temperature Dependence of the Fano Effect

So far we have considered the zero-temperature case. At finite temperatures the resonances are rapidly smeared out. It will be shown that the effect of temperature (due to thermal broadening) becomes gradually weaker with increasing size of the PT impurity given in Eq. (19).

In order to consider thermal effects, we employ the finite-temperature conductance formula

$$G(\mu, T) = G_0 \int \mathrm{d}E \left(\frac{\partial f}{\partial \mu}\right) G(E, 0) \,, \tag{41}$$

where $G(E, 0)$ is the zero-temperature conductance and f is the Fermi distribution function given by

$$f(E) = \frac{1}{\mathrm{e}^{(E-\mu)/k_\mathrm{B}T} + 1} \,. \tag{42}$$

For comparison, we also consider the case of an attractive short-range (SR) impurity. The SR impurity is modeled by the Dirac δ function along the propagation direction

$$V(x, y) = \frac{\hbar^2 \gamma}{2m} \delta(x) \upsilon(y) \,, \tag{43}$$

where, as in the case of the PT impurity, $\upsilon(y)$ is an arbitrary function of the coordinate y, and $\gamma < 0$. For this scattering potential, the bound-state energy is found to be $\tilde{E}_0 = E_2 - (U_{22}^2/4)$. The shift and width that the bound state acquires can also be found analytically as

$$\delta = \frac{U_{11} U_{22} U_{12}^2}{8k^2 + 2U_{11}^2} \,, \tag{44}$$

and

$$\Gamma = \frac{k U_{12}^2 U_{22}}{4k^2 + U_{11}^2} \,. \tag{45}$$

In our energy unit (≈ 17.7 meV) that we employ in this article, $k_B T = 0.001$ corresponds to 205 mK. The temperature dependence of the Fano resonance for the SR impurity is shown in Fig. 4a. The corresponding plot in the case of a PT impurity is similar. In Fig. 4a we have chosen $U_{11} = 1$, $U_{22} = 0.9$, and $\upsilon_{12} = 0.08$. We note that the resonance structure is rapidly smeared out with increasing temperature. That is, the Fano resonance becomes gradually broader while its amplitude, $\Delta G = G_{\max} - G_{\min}$, decreases. However, this amplitude is influenced by the type and size of the impurity in a significant way. The amplitude versus $k_B T$ is plotted in Fig. 4b for both a SR and a PT impurity for two values of α, i. e., for $\alpha = 1.5$ and 1.35. The fastest decrease of the amplitude with increasing temperature occurs for a SR impurity for which $\Delta G \simeq 0.1$ at $k_B T \simeq 0.012$ (or $T \simeq 2.5$ K). However, depending on the size α^{-1} of the PT impurity, the Fano resonance persists for higher temperatures. That is, the effect of temperature becomes progressively weaker as the size of the finite-range PT impurity increases.

The above-mentioned smearing of the Fano resonance can be explained in terms of the thermal broadening, via the smooth peak in $\partial f / \partial \mu$, which obscures the resonance as $k_B T$ becomes comparable to the resonance width. In the case of a SR impurity [solid line of Fig. 4b], the amplitude of the resonance becomes half its zero-temperature value as soon as $k_B T \simeq 0.0018$ (or, equivalently, $T \simeq 380$ mK), which is comparable to the calculated resonance width $\Gamma \simeq 0.0022$. However, as has been shown elsewhere [40], in the case of a PT impurity the resonance width increases with decreasing α (i. e., with increasing size), resulting in a gradually broader resonance. Therefore, for gradually larger PT impurities, the Fano resonance requires higher temperatures in order to diminish [see dashed lines of Fig. 4b]. For a PT impurity with $\alpha = 1.5$, the amplitude becomes half its zero-temperature value at $k_B T \simeq 0.0024$ (which corresponds to $T \simeq 480 mK$), while the resonance width is $\Gamma \simeq 0.0026$. The above-mentioned broadening of the Fano resonance is reflected in an almost linear increase of the resonance width [31].

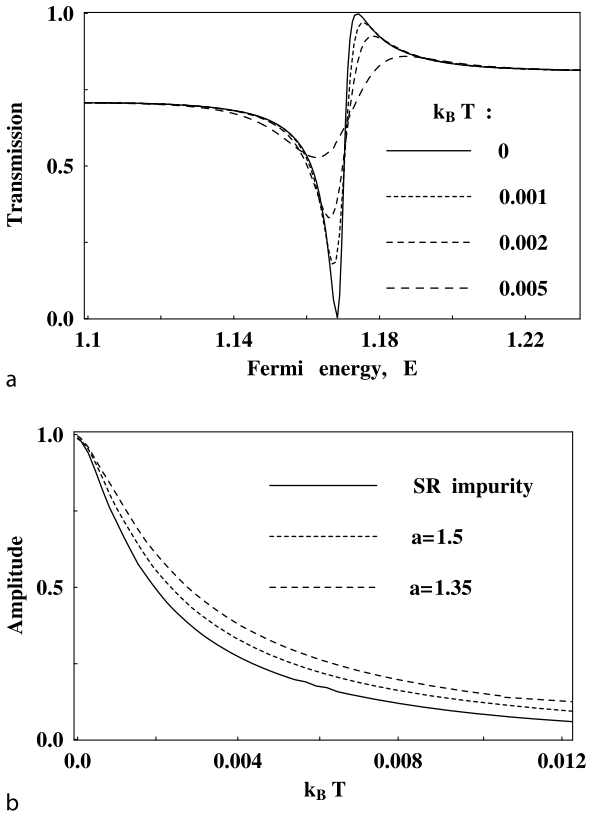

Resonances in Electronic Transport Through Quantum Wires and Rings, Figure 4

The temperature dependence of the transmission probability T versus Fermi energy E through an attractive δ function potential, $V(x, y) = (\hbar^2 \gamma / 2m) \delta(x) \upsilon(y)$, in a uniform 2D quantum wire is shown in a. In our energy unit (see text) $k_B T = 0.001$ corresponds to 205 mK. Note the smearing of the Fano resonance as the temperature increases. b Amplitude, $\Delta G = G_{\max} - G_{\min}$, of the Fano resonance versus $k_B T$ for the δ function (*solid line*) and the PT (*dashed lines*) impurities. Note the progressively weaker effect of temperature for PT impurities with gradually larger size (i. e., gradually smaller α)

Electronic Transport Through Mesoscopic Open Rings

As mentioned above, resonances of the Fano type have been treated theoretically and observed experimentally in

various mesoscopic systems. One particular class of mesoscopic systems in which the Fano effect shows up, is the 1D ring connected to current leads with [41,43] or without [18] an impurity in one of its arms. In such a system, the electron motion between the junctions is treated as purely one-dimensional, i. e., no subband structure is taken into account and therefore no intersubband interaction occurs. Thus, the usual interpretation of the Fano effect as being due to quasibound states splitting off from nonpropagating subbands and the subsequent coupling with the continuum of states, does not apply in this system. Fano resonances have also been studied in various other ring geometries, such as a ring with coupling between the leads [47], and an ideal 1D double ring [48], to name a few.

The presence of an Aharonov–Bohm flux leads to further interesting effects in mesoscopic open rings. The application of an AB magnetic flux between the two electronic paths leads to the interference of the electronic wave functions, giving rise to a circulating (or persistent) current which is periodic in the magnetic flux with period Φ_0, where $\Phi_0 = hc/\mathrm{e}$ is the flux quantum. Great effort has been devoted to the description of this effect in various open ring geometries [7,10,17,18,33,47,48]. The persistent current has also been related to the collapse of Fano resonances in Reference [41].

Most of the investigations on electronic transport in 1D open rings employ the Griffith's boundary condition (conservation of current) and continuity of wave functions at the junctions. Even though useful, the Griffith's boundary condition constitutes a significant limitation on the possible scattering effects at the junctions on the transmission probability. The reason is that all the scattering effects are included in the calculation via a single value of the ring-lead coupling strength, whereas the actual coupling strength may vary, in a continuous manner, from weak to strong. Therefore, the imposition of Griffith's boundary condition is insufficient to fully describe electron scattering at the junctions and important aspects of electronic transport are neglected [41]. After reviewing in the next section the more general S-matrix approach, which is suitable for variable coupling strength, we will illustrate the effect of this coupling on the transmission resonances. We point out that the existence of Fano resonances in 1D mesoscopic rings has been known for a long time [3]. However, recently [41,43] the dependence of the Fano resonance on the system parameters has been studied, in the zero-flux case, and revealed an important feature; namely, the collapse of a Fano resonance when the system parameters are commensurate. Particularly, there are special impurity positions for which certain Fano resonances collapse.

S-Matrix Approach

We consider the electronic transport through a ring connected to current leads in the presence of an AB magnetic flux, as shown in Fig. 5. The leads and the ring are taken to be strictly 1D, i. e., no subband structure is taken into account. We consider the current flow to be from left to right. At a junction of a lead with the ring, the three outgoing waves with amplitudes $(\alpha', \beta', \gamma')$ are related by an S matrix to the three incoming waves (α, β, γ) as

$$\mathbf{O} = S\mathbf{I}\,, \tag{46}$$

where $\mathbf{O}$ represents the outgoing and $\mathbf{I}$ the incoming waves. Current conservation and time-reversal invariance imply that S is unitary and symmetric [3], and given as

$$S = \begin{pmatrix} -(a+b) & \sqrt{\varepsilon} & \sqrt{\varepsilon} \\ \sqrt{\varepsilon} & a & b \\ \sqrt{\varepsilon} & b & a \end{pmatrix}. \tag{47}$$

Due to current conservation,

$$a = \frac{1}{2}\left(\sqrt{1-2\varepsilon} - 1\right)\,, \tag{48}$$

$$b = \frac{1}{2}\left(\sqrt{1-2\varepsilon} + 1\right)\,. \tag{49}$$

Therefore, the coefficients a and b are expressed as functions of the single parameter ε, where $0 \le \varepsilon \le 1/2$. For $\varepsilon = 0$, electrons from the current lead are totally reflected and thus there is no coupling between the current leads and the ring. For $\varepsilon = 1/2$, the junction is completely transparent for incoming electrons and the current lead is strongly coupled to the ring. It can be verified that the Griffith's boundary condition corresponds to $\varepsilon = 4/9$.

In the presence of an impurity and a magnetic flux Φ, the Schrödinger equation that describes the electron motion in the ring is

$$\left[-\frac{\hbar^2}{2m}\left(\frac{\partial}{\partial y} - i\frac{2\pi}{L}\frac{\Phi}{\Phi_0}\right)^2 + V_i(y)\right]\psi(y) = E\psi(y)\,, \tag{50}$$

where y is the coordinate along the ring, $\Phi_0 = hc/\mathrm{e}$ is the flux quantum associated with a single charge e, L is the circumference of the ring, and $V_i(y)$ is the impurity potential in the upper arm (at point D of Fig. 5). The lengths of the two arms, L_1 and L_2, are taken to be equal, i. e., $L_1 = L_2 = L/2$. We consider an impurity potential modeled by a Dirac δ function

$$V_i(y) = \gamma\delta(y - y_i)\,, \tag{51}$$

where y_i is the impurity position in the arm, and γ sets the magnitude of the impurity potential, which is taken to be repulsive (i. e., $\gamma > 0$). The wave functions in both leads are written as

$$\psi_l(x) = \alpha_j e^{ikx} + \alpha'_j e^{-ikx} , \tag{52}$$

where $j = 1(2)$ for the left(right) leads, respectively, x is the coordinate along the leads, and $k = (2mE)^{1/2}/\hbar$ is the wave vector. In the ring the wave functions are expressed as

$$\psi_2(y) = e^{2\pi i(\Phi/\Phi_0)(y/L)} \left(\beta'_1 e^{iky} + \beta_1 e^{-iky}\right) , \tag{53}$$

$$\psi_3(y) = e^{2\pi i(\Phi/\Phi_0)(y/L)} e^{-i\pi(\Phi/\Phi_0)} \left(\beta_2 e^{-ik(L_1-y)} + \beta'_2 e^{ik(L_1-y)}\right) , \tag{54}$$

$$\psi_4(y) = e^{-2\pi i(\Phi/\Phi_0)(y/L)} \left(\gamma'_1 e^{iky} + \gamma_1 e^{-iky}\right) , \tag{55}$$

where $\psi_2(y)$ and $\psi_3(y)$ are the wave functions before and after the impurity, respectively, in the upper arm while $\psi_4(y)$ is the wave function in the lower arm (see Fig. 5). With the help of a transfer matrix, the amplitudes β_2 and β'_2 to the right of the impurity are expressed in terms of the amplitudes β'_1 and β_1 to the left of the impurity. The transfer matrix is found using a standard procedure, i. e., by applying the boundary conditions at the impurity. We then get

$$\begin{pmatrix} \beta_2 \\ \beta'_2 \end{pmatrix} = e^{i\pi(\Phi/\Phi_0)} T_i \begin{pmatrix} \beta'_1 \\ \beta_1 \end{pmatrix} , \tag{56}$$

where the transfer matrix T_i is given as

$$T_i = \begin{pmatrix} (1/t^*)\, e^{iq} & (r/t)\, e^{iq} e^{-2iqs_i} \\ (r^*/t^*)\, e^{-iq} e^{2iqs_i} & (1/t)\, e^{-iq} \end{pmatrix} . \tag{57}$$

In Eq. (57), we have used a dimensionless wave vector defined as $q = kL_1$ and the dimensionless impurity position $s_i = y_i/L_1$. Note that q has also been used previously to denote the Fano asymmetry parameter but its meaning becomes clear from the particular context. With this notation, the transmission and reflection amplitudes of the impurity potential are expressed as

$$t = \frac{2iq}{2iq - V} , \tag{58}$$

and

$$r = \frac{V}{2iq - V} , \tag{59}$$

where we have used the dimensionless parameter $V = 2m\gamma L_1/\hbar^2$ to characterize the impurity strength. We note

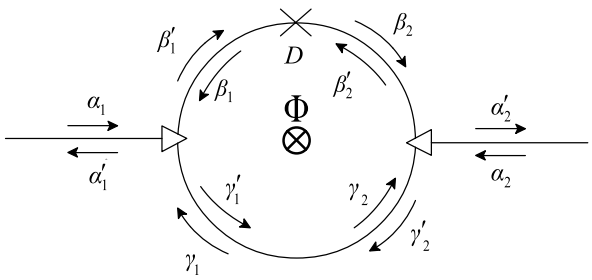

Resonances in Electronic Transport Through Quantum Wires and Rings, Figure 5

A one-dimensional ring connected to current leads with an impurity at point D (in the upper arm) and an Aharonov–Bohm flux Φ. The α's, β's, and γ's represent the amplitudes of the wave functions (see text)

that an impurity with a small spatial extension d and a magnitude of $\bar{\gamma}$ can be approximated by a δ function with $\gamma = d\bar{\gamma}$ and therefore $V = 2m\bar{\gamma}dL_1/\hbar^2$. If we choose an experimentally realizable [20,21,24] arm length $L_1 = 1\,\mu$m, the effective mass for *GaAs* ($m = 0.067m_0$), and $d = 0.01\,\mu$m, the unit of the impurity parameter $V = 1$ corresponds to $\bar{\gamma} = 0.12$ me.

We consider now a wave of unit amplitude $\alpha_1 = 1$, incident from the left. To find the transmission probability $T = |\alpha'_2|^2$ of the ring we have to find the amplitudes β'_1, β'_2, γ'_1, γ'_2 under the condition $\alpha_2 = 0$. Therefore, we employ the S-matrix at the two junctions. Applying Eqs. (46) and (47) at the right junction we can express the γ's in terms of the β's as

$$\begin{pmatrix} \gamma'_2 \\ \gamma_2 \end{pmatrix} = T_j \begin{pmatrix} \beta_2 \\ \beta'_2 \end{pmatrix} , \tag{60}$$

where T_j is a matrix given as

$$T_j = \frac{1}{b} \begin{pmatrix} (b^2 - a^2) & a \\ -a & 1 \end{pmatrix} . \tag{61}$$

Applying Eqs. (46) and (47) at the left junction and using $\alpha_1 = 1$, we are able to express β_1, β'_1 in terms of γ_1, γ'_1 as

$$\begin{pmatrix} \beta'_1 \\ \beta_1 \end{pmatrix} = \frac{\sqrt{\varepsilon}}{b} \begin{pmatrix} b - a \\ -1 \end{pmatrix} + T_j \begin{pmatrix} \gamma_1 \\ \gamma'_1 \end{pmatrix} . \tag{62}$$

In the lower arm we insert $y = L_2$ into $\psi_4(y)$ of Eq. (55) which then yields two terms, the first being γ_2 and the second γ'_2. Eliminating γ_2 and γ'_2 from these two terms by employing Eq. (60) yields a relation between γ_1, γ'_1 and β_2, β'_2 with the help of a matrix T'_j given as

$$T'_j = \frac{1}{b} \begin{pmatrix} (b^2 - a^2)\, e^{iq} & a e^{iq} \\ -a e^{-iq} & e^{-iq} \end{pmatrix} . \tag{63}$$

Using the matrix Π defined as

$$\Pi = T_j e^{i\pi(\Phi/\Phi_0)} T_j' e^{i\pi(\Phi/\Phi_0)} T_i - 1 \,, \tag{64}$$

and employing the relation $b - a = 1$, the transmission amplitude can finally be expressed as

$$\alpha_2' = -\frac{\varepsilon}{b^2} e^{i\pi(\Phi/\Phi_0)} \frac{p}{\det(\Pi)} \,, \tag{65}$$

where

$$p = \det(\Pi)[(\ 1 \quad 0\ ) + (\ 0 \quad 1\ )] T_i \Pi^{-1} \begin{pmatrix} 1 \\ -1 \end{pmatrix} . \tag{66}$$

The poles of the transmission amplitude, i. e., the solutions of $\det(\Pi) = 0$, determine the resonant behavior of the transmission probability. The transmission probability of the ring can then be written as

$$T = |\alpha_2'|^2 = \frac{\varepsilon^2}{b^4} \frac{|p|^2}{|\det(\Pi)|^2} \,. \tag{67}$$

Fano Resonances in the Transmission Probability

In the numerical calculations we take the impurity position at $s_i = 0.42$ and its scattering strength to be $V = 11$. We first consider the zero-flux case.

We show by the solid lines in Fig. 6 the transmission probability T versus the (dimensionless) wave vector q through a ring with an impurity in its upper arm, for three values of the coupling parameter ε. The dashed lines show the transmission probability of a ring with no impurity. In the weak coupling regime of Fig. 6a, we note that for the no impurity case sharp Breit–Wigner peaks occur at $q = n\pi$, where $n = 1, 2, \ldots$ These resonances are the result of constructive interference of the electron waves in the two arms and they become gradually broader with increasing coupling. In the strong coupling regime of Fig. 6c, the ring becomes completely transparent for incoming electrons resulting in unity transmission.

In the presence of an impurity and in the weak coupling regime of Fig. 6a we identify two types of transmission resonances; Breit–Wigner and Fano type. Most of the Breit–Wigner peaks are similar to those of the no impurity case, but note that they occur at different positions on the q axis from those of the no impurity case and their amplitudes can be less than unity. These resonances are attributed to the constructive scatterings from the junctions and the impurity. Thus, in the weak coupling regime, it's possible for electrons entering the ring to spend enhanced periods of time in the ring before being transmit-

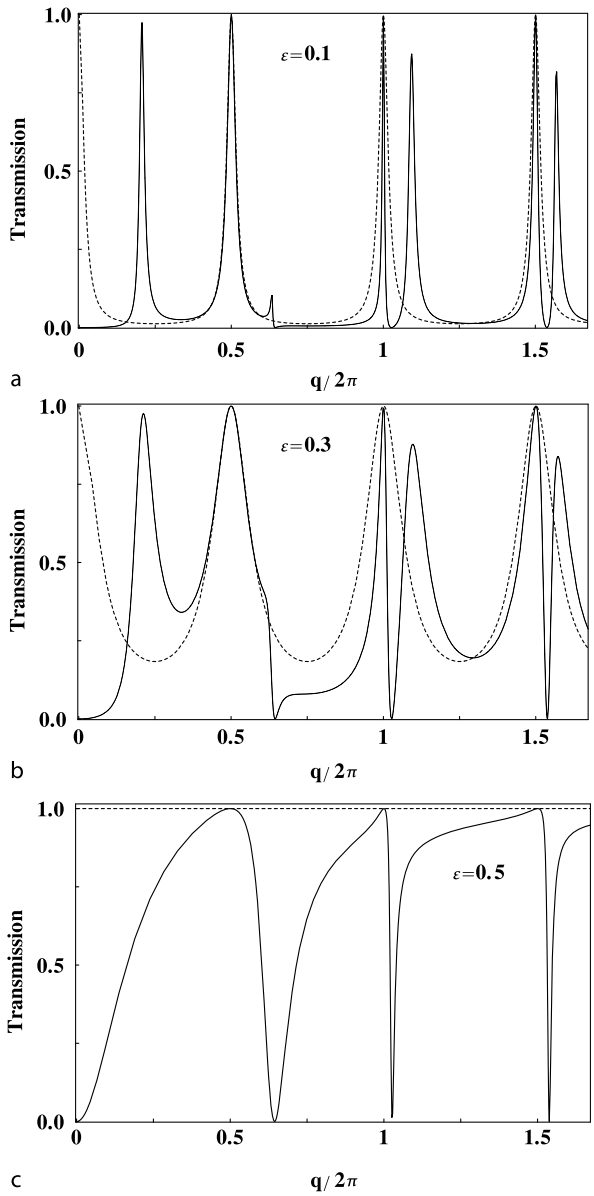

Resonances in Electronic Transport Through Quantum Wires and Rings, Figure 6

Transmission probability T versus dimensionless wave vector $q/2\pi$, where $q = kL_1$, through a one-dimensional ring with equal arms ($L_1 = L_2$) connected to current leads, for three values of the coupling parameter ε. The *dashed lines* represent the case in which there is no impurity. The *solid lines* represent the case in which there is an impurity, $V_i(y) = (\hbar^2 V/2mL_1)\delta(y - y_i)$, in the upper arm (see Fig. 5). We have employed the value $V = 11$ for the impurity strength, where V is defined in the text, and $s_i = 0.42$, where $s_i = y_i/L_1$ and $L_1 = 1\,\mu$m. Depending on the coupling, there may be both Breit–Wigner and Fano line shapes. In the strong coupling regime, $\varepsilon = 0.5$, only the Fano resonances survive. Note that while varying the coupling the positions of the transmission zeros and ones remain fixed

ted. For these electrons the interference is constructive and the Breit–Wigner resonances are narrow.

When the coupling increases to $\varepsilon = 0.3$, scattering effects at the junctions become weaker, resulting in larger transmissivity and larger widths of the Breit–Wigner resonances. The larger resonance widths correspond to larger imaginary parts of the associated poles. In addition, in the limit $\varepsilon \to 0.5$ the poles move away from the real axis resulting in the disappearance of the Breit–Wigner resonances [41], as shown in Fig. 6c.

The Fano resonances occur at $q \approx n\pi$, and in fact their transmission ones occur exactly at $q = n\pi$. These resonances are not much sensitive to the perturbation caused by the junctions, and the transmission probability exhibits Fano line shapes even in the strong coupling regime, as shown in Fig. 6c. Furthermore, increasing the coupling does not affect the positions of the transmission zeros and ones, but note that the asymmetry parameters change. This is reflected in a modification of the resonance width q_i, where q_i is the imaginary part of the associated pole in the complex q plane, with increasing coupling. For example, for the Fano resonance at $q \approx 2\pi$, the difference ξ between the transmission zero q_0 and one q_1, i. e., $\xi = q_0 - q_1$, has the value $\xi = 0.06782$ and remains constant for any value of the coupling parameter ε. However, the values of the associated pole are $q = 6.291 - i0.0182$ for $\varepsilon = 0.1$, whereas for $\varepsilon = 0.3$ the pole location is at $q = 6.312 - i0.039$ and for $\varepsilon = 0.5$ at $q = 6.351 - i0.0302$. To further describe this in the framework of Fano's asymmetric function [12] written in terms of the wave vector q [see Eq. (69)], we note that ξ can be expressed [41] as $\xi = q_i \left(|\tilde{q}| + 1/|\tilde{q}|\right)$, where q_i is the resonance width and $\tilde{q}$ the asymmetry parameter. We see that, as the asymmetry parameter $\tilde{q}$ varies, the width q_i also varies in such a way that ξ remains constant.

Thus, in the absence of impurity, the transmission exhibits Breit–Wigner resonances in the weak coupling regime. However, increasing V from zero in this regime, leads to shifting of the Breit–Wigner resonances on the q axis, but also results in a transmission zero, thus giving rise also to a Fano line shape.

Tuning of the Fano Resonance

Recently [41] am interesting feature of the Fano line shape in 1D open rings was discovered; namely, when the impurity is located at some special positions in the arm, certain Fano resonances collapse. Even though it is independent of the ring-lead coupling, we will illustrate this feature in the strong coupling regime $\varepsilon = 0.54$. In this regime, an analytic expression for the transmission amplitude can be found,

$$\alpha_2' = -\mathrm{e}^{-i\pi\Phi/\Phi_0} \cdot \frac{4iq(\mathrm{e}^{-iq} - \mathrm{e}^{iq}) - V(\mathrm{e}^{iq} + \mathrm{e}^{-iq}) + V(\mathrm{e}^{iD_i q} + \mathrm{e}^{-iD_i q})}{4iq - 4iq\mathrm{e}^{-2iq} - V\mathrm{e}^{-2iQ_i q} - V\mathrm{e}^{-2is_i q} + 2V\mathrm{e}^{-2iq}} , \tag{68}$$

where $D_i = 1 - 2s_i$, $Q_i = 1 - s_i$ and we have also employed the relation $t/r = 2iq/V$. In the vicinity of a Fano resonance, the transmission amplitude α_2' can be expressed in the asymmetric Fano form [30]. Expanding the numerator and denominator of Eq. (68) in a Taylor series around a transmission zero, q_0, and keeping only linear terms we get

$$\alpha_2' = t_d \frac{q - q_0 - \mathrm{Re}(\Sigma) + \delta}{q - q_0 - \Sigma} , \tag{69}$$

where $t_d = \eta_1/\eta_2$, $\delta = \lambda_1/\eta_1 = \widetilde{q} q_i$, and $\Sigma = \lambda_2/\eta_2 = \Delta - iq_i$ with

$$\eta_1 = 8q\cos q_0 + 2V\sin q_0 - 2VD_i \sin(q_0 D_i) , \tag{70}$$

$$\eta_2 = 8q\mathrm{e}^{-2iq_0} - 2iVQ_i\mathrm{e}^{-2iq_0 Q_i} - 2is_i V\mathrm{e}^{-2iq_0 s_i} + 4iV\mathrm{e}^{-2iq_0} , \tag{71}$$

$$\lambda_1 = 8q\sin q_0 - 2V\cos q_0 + 2V\cos(q_0 D_i) , \tag{72}$$

$$\lambda_2 = 4iq - 4iq\mathrm{e}^{-2iq_0} - V\mathrm{e}^{-2iq_0 Q_i} - V\mathrm{e}^{-2iq_0 s_i} + 2V\mathrm{e}^{-2iq_0} . \tag{73}$$

In Eq. (69), δ and Σ can be decomposed as $\delta = \tilde{q} q_i$ and $\Sigma = \Delta - iq_i$. Equation (69) will be used for the description of the collapse of Fano resonances. It will be shown that, for particular impurity positions, the expression for Σ vanishes which further indicates that the resonance width shrinks to zero and the Fano line shape collapses.

Let us consider an impurity located at s_i, where s_i is the ratio of two integers,

$$s_i = \frac{r}{n} , \tag{74}$$

where n is fixed and $r = 1, 2, \ldots, n$. This means that $0 \le s_i \le 1$. For an impurity located at one of the special positions given by Eq. (74), the expression for Σ takes the form

$$\Sigma = \frac{4iq - 4iq\mathrm{e}^{-2iq_0} - V\mathrm{e}^{-2iq_0(1-s_i)} - V\mathrm{e}^{-2iq_0 s_i} + 2V\mathrm{e}^{-2iq_0}}{8q\mathrm{e}^{-2iq_0} - 2iV(1-s_i)\mathrm{e}^{-2iq_0(1-s_i)} - 2is_i V\mathrm{e}^{-2iq_0 s_i} + 4iV\mathrm{e}^{-2iq_0}} . \tag{75}$$

Now let $q_0 \to \mu n\pi$, where n is the same (fixed) integer as that in Eq. (74) and $\mu = 1, 2, \ldots$ For these values of q_0, we note that the numerator of Σ in Eq. (75) vanishes which implies that the widths q_i, of all Fano resonances which occur at integer multiples of $n\pi$, that is at $q \simeq n\pi, 2n\pi, 3n\pi, \ldots$, shrink to zero and therefore collapse. Note that δ also vanishes. In this case, the transmission amplitude α_2' is equal to the nonresonant transmission amplitude t_d and given as $\alpha_2' = t_d = -[1 + (iV/4q)]^{-1}$. To illustrate this behavior, we plot in Fig. 7 the exact transmission probability from Eq. (68) versus the wave vector q, for three values of the impurity position (i. e., for $s_i = 1/2$, 1/3, and 1/4). For comparison, we also plotted the approximate transmission probability with α_2' given by Eq. (69) in Fig. 7c in the vicinity of $q \approx 3\pi$, for the same impurity parameters as those used for plotting the exact expression. We note that there is good agreement with the exact line shape. In Fig. 7 the value of the impurity strength is taken to be $V = 12$.

In agreement with the above discussion related to Eqs. (74) and (75), we note in Fig. 7a that placing the impurity at $s_i = 1/2$, which means $n = 2$ in Eq. (74), causes simultaneous collapse of all Fano resonances at $q \approx 2\pi, 4\pi, 6\pi, \ldots$ Furthermore, as shown in Fig. 7b, an impurity placed at $s_i = 1/3$, which means $n = 3$ in Eq. (74) causes simultaneous collapse of all Fano resonances at $q \approx 3\pi, 6\pi, \ldots$ For $s_i = 1/4$, collapse of all Fano resonances at $q \approx 4\pi, 8\pi, \ldots$ occurs, as shown in Fig. 7c. The above-mentioned collapsing behavior has also been described in terms of the pole structure of the transmission amplitude, in the complex q plane, as a function of the impurity position. It has been shown [41] that as the impurity approaches, for example $s_i = 1/2$, the pole associated with $q = 2\pi$ approaches the real axis and, exactly at $s_i = 1/2$ the pole is located on the real axis of the q plane implying that the resonance width becomes zero.

We also note that we can write Eq. (74) as $s_i = r\pi/n\pi = r\pi/q_n$, where $q_n = n\pi$ are the wave numbers at which the Fano resonances occur. This means that, given the impurity position s_i, enables us to find all Fano profiles that collapse from

$$q = \frac{r\pi}{s_i} \,. \tag{76}$$

For example, at $s_i = 1/3$, all Fano profiles that occur at

$$q = \frac{r\pi}{(\frac{1}{3})} = 3r\pi = 0, 3\pi, 6\pi, \ldots \,, \tag{77}$$

collapse. At the extreme positions $s_i = 0$ and 1, all Fano resonances collapse. This tunability of the Fano reso-

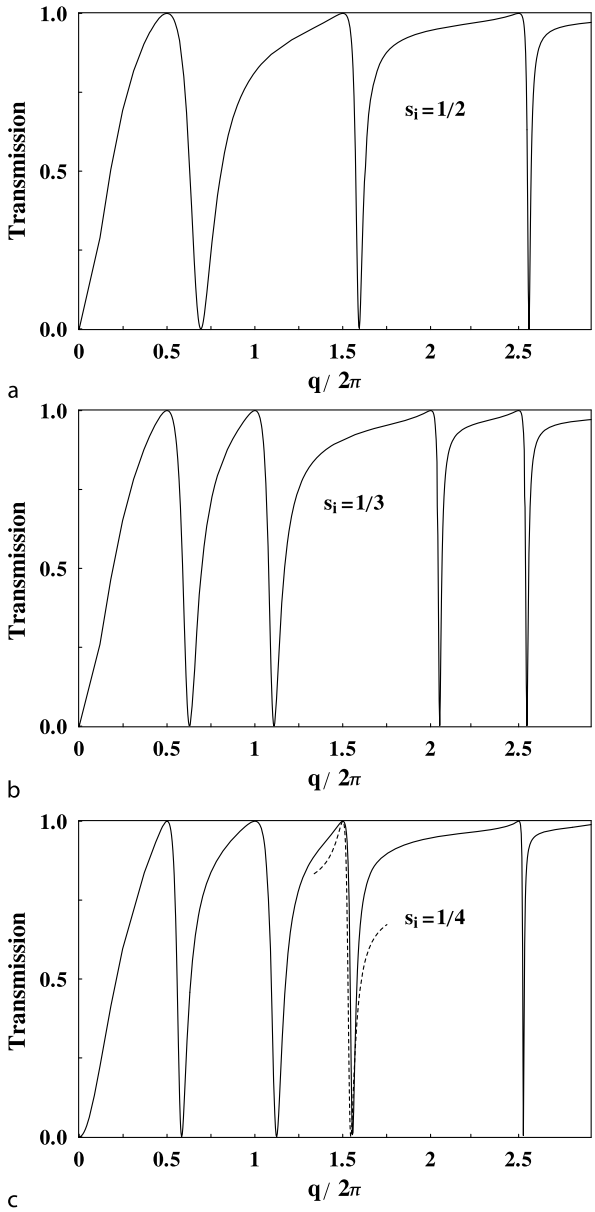

Resonances in Electronic Transport Through Quantum Wires and Rings, Figure 7

Transmission probability T versus dimensionless wave vector $q/2\pi$, where $q = kL_1$, through a 1D ring with equal arms connected to current leads with an impurity of strength $V = 12$ in the upper arm. For the impurity position we have chosen $s_i = 1/2$, 1/3, and 1/4 in parts a, b, and c respectively. In a, all Fano resonances at $q \approx \mu 2\pi = 2\pi, 4\pi, 6\pi, \ldots$, collapse. In b and c, the Fano resonances collapse at $q \approx \mu 3\pi$ and $q \approx \mu 4\pi$, respectively, where $\mu = 1, 2, 3, \ldots$ The *dashed line* in c represents the approximate expression given in Eq. (69)

nances with respect to the impurity position may provide useful means for the design of electronic nanodevices. In this context the impurity potential may be provided by a movable scanning tunneling microscope tip. Note also that this tunability is independent of the scattering strength V of the impurity.

Further interesting issues related to open 1D rings are how the ring-lead coupling, impurity strength and position affect the persistent current [18,32,34,41,45,46,47,48]. Even though we will not go into details here, we would like to mention that the persistent current versus magnetic flux increases dramatically whenever the impurity approaches one of the values given in Eq. (74), i. e., $s_i \to r/n$ and, exactly at these values, the persistent current diverges [41]. This behavior can be related to the motion of the poles of the transmission amplitude in the complex q plane as a function of magnetic flux, taking into account the fact that the persistent current arises near these poles [32,33,34,41,48].

The effect of an AB magnetic flux on the Fano resonance structure is also interesting. It has been shown [41] that increasing the AB flux encircled by the ring results in additional phase changes of the electron waves in the upper and lower arms, leading to modification of the Fano interference pattern at the right junction. Applying gradually the magnetic flux causes weaker destructive and constructive interference and consequently transmission zeros and ones do not exist anymore. As a result of these interference effects, the amplitudes of the resonances usually decrease and become gradually distorted. These effects repeat periodically with period Φ_0, where $\Phi_0 = hc/e$ is the flux quantum.

Temperature Dependence of the Fano Resonances

Until now we discussed the transmission resonances at zero temperature. As in the case of quantum wires discussed in Sect. "Temperature Dependence of the Fano Effect", to consider thermal effects we should employ the finite-temperature conductance formula, Eq. (41). In the integral of this equation we first make a change of variables from energy E to the dimensionless wave vector q as

$$E = E^\star q^2 \,, \tag{78}$$

where $E^\star = \hbar^2/2mL_1^2$. In the following we define a dimensionless parameter $T_d = k_B T/E^\star$ which is a measure of temperature. Then, the unit of this parameter $T_d = 1$ corresponds to $T = 6.59$ mK. After some manipulations, the

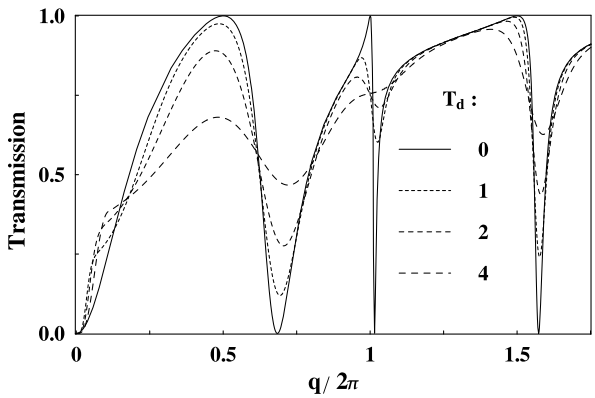

Resonances in Electronic Transport Through Quantum Wires and Rings, Figure 8

Temperature dependence of the transmission probability T versus dimensionless wave vector $q/2\pi$, where $q = kL_1$, through a 1D ring with equal arms connected to current leads. We have chosen $s_i = 0.47$ and $V = 11$ for the impurity position and strength, respectively. We have used the values $T_d = 0, 1, 2, 4$, where $T_d = k_B T/E^*$ and $E^* = \hbar^2/2mL_1^2$. The unit of the parameter $T_d = 1$ corresponds to 6.59 mK. Note the rapid smearing of the Fano resonances (especially of the narrow ones) as the temperature increases

transmission probability of Eq. (41) can be written as

$$\frac{G(q, T_d)}{G_0} = \frac{1}{2T_d} \int q' |\alpha_2'(q')|^2 \cosh^{-2}\left(\frac{q'^2 - q^2}{2T_d}\right) \mathrm{d}q' \,. \tag{79}$$

Restricting ourselves to the strong coupling limit $\varepsilon = 0.5$, we plot in Fig. 8 the transmission probability versus the wave vector q for several values of the parameter T_d, i. e., for $T_d = 0, 1, 2, 4$. In Fig. 8 we have taken the strength of the impurity $V = 11$ and its position $s_i = 0.47$. Note the strong temperature dependence of the narrower Fano resonances, especially of the resonance that occurs at $q \approx 2\pi$. The effect of temperature is relatively weaker for the wider resonances. For high enough temperatures the resonance structure diminishes.

The origin of this rapid smearing of the resonances is again the thermal broadening, via the smooth peak in $\partial f/\partial \mu$, which obscures the Fano resonance as $k_B T$ becomes comparable to the resonance width. For the parameters chosen in the plot of Fig. 8, the amplitude of the Fano resonance at $q \approx 2\pi$ becomes half its zero-temperature value as soon as $T_d = 0.35$, which corresponds to $T \simeq 2.3$ mK while the resonance width is $q_i \simeq 0.11$ (which corresponds to $T \simeq 0.27$ mK). This resonance diminishes at $T_d \approx 3.4$. However, the first resonance at $q \approx \pi$, which is much broader than the sec-

ond, as well as the third one at $q \approx 3\pi$ require higher temperatures in order to diminish. Particularly, the amplitude of the third resonance becomes half its zero-temperature value as soon as $T_d = 1.6$, which corresponds to $T \simeq 10\,\text{mK}$, while the resonance width is $q_i \approx 0.27$ which corresponds to $T \simeq 0.65\,\text{mK}$. The above-mentioned smearing of the Fano resonances has also been discussed in the context of various other mesoscopic systems [16,24,25,26,29].

Future Directions

In the past two decades advances in microfabrication have made possible the confinement of electrons in a conductor with lateral extension of 100 nm and less, resulting in narrow quantum wires, constrictions, and quantum dots. Consequently, it has been possible to engineer device potentials which vary over the length scale such that the electron motion is ballistic or quasiballistic at low temperatures. The ultrasmall size of these structures greatly eliminates electron scattering by defects and therefore one can obtain high mobility conducting channels. In these mesoscopic systems, electronic transport is governed by quantum mechanics rather than classical mechanics. At very low temperatures, the scattering by phonons is suppressed and the phase coherence length of the electron (i. e., the length over which the electron can be considered to be in a pure state) becomes larger than the system dimensions. As a result, the mesoscopic system becomes essentially an electron waveguide where the electronic transport properties are mainly determined by the impurity configuration, the geometrical characteristics of the conductor, and the principles of quantum mechanics. Convincing demonstrations of the quantum transport regime came from experiments on thin-metal or semiconductor films. Given the phase coherence of the electrons throughout the sample, several ideas for active quantum devices (such as, transistors, switches, etc.) have been proposed and are now under investigation. These are the quantum analogs of optical or microwave devices.

My intention in writing this article was to review certain important aspects of resonant electronic transport in mesoscopic systems. The resonant behavior of the transmission probability through such systems has actually been an issue which stimulated a great deal of experimental and theoretical research over the last two decades. The Fano function has been shown to arise as the most general resonance line shape, provided that two scattering channels – a resonant and a nonresonant one interfere. One important aspect of resonant electronic transport, which has actually been a problem common to many branches of physics, is to extract lifetimes of quasibound states from resonances in transmission coefficients or scattering cross sections. This requires knowledge of the resonance widths, which have been discussed in this article for two systems: i) a uniform quantum waveguide with a "smooth" finite size impurity, and ii) a 1D ring connected to current leads with an impurity in one arm. These resonances can also provide us with information about the parameters of the impurity potential. An interesting future experiment would be the determination of the impurity parameters from knowledge of the resonance line shapes.

The effect of temperature is also important since it destroys the Fano resonance structure. This depends on the size of the impurity on the quantum wire and also on the resonance width. The effect of temperature on the Fano resonances becomes stronger for impurities with smaller size while narrow resonances tend to diminish faster with increasing temperature.

An interesting feature is also the systematic collapse of certain Fano line shapes in electronic transport through a 1D ring with a short-range impurity. The collapse occurs for special values of the impurity position in the arm. As mentioned in this article, the tunability of the Fano resonances with respect to the impurity position could possibly prove useful for the design and engineering of electronic nanodevices. Even though it has not been discussed in this article, this tunability could further be useful for the generation of strong persistent currents [41].

Bibliography

Primary Literature

1. Adair RK, Bockelman CK, Peterson RE (1949) Experimental corroboration of the theory of neutron resonance scattering. Phys Rev 76:308–308
2. Büttiker M (1988) Symmetry of electrical-conduction. IBM J Res Dev 32:317–334
3. Büttiker M, Imry Y, Azbel MY (1984) Quantum oscillations in one-dimensional normal-metal rings. Phys Rev A 30:1982–1989
4. Bagwell PF (1990) Evanescent modes and scattering in quasi-one-dimensional wires. Phys Rev B 41:10354–10371
5. Banjamin C, Jayannavar AM (2003) Features in evanescent Aharonov–Bohm interferometry. Phys Rev B 68:085325-1–085325-6
6. Bardarson JH, Magnusdottir I, Gudmundsdottir G, Tang CS, Manolescu A, Gudmundsson (2004) Coherent electronic transport in a multimode quantum channel with Gaussian-type scatterers. Phys Rev B 70:245308-1–245308-10
7. Benjamin C, Jayannavar AM (2001) Current magnification effect in mesoscopic systems at equilibrium. Phys Rev B 64:233406-1–233406-4

8. Cerdeira F, Fjeldly TA, Cardona M (1973) Effect of free carriers on zone-center vibrational modes in heavily doped p-type Si. II. Optical modes. Phys Rev B 8:4734–4745
9. Clerk AA, Waintal X, Brouwer PW (2001) Fano resonances as a probe of phase coherence in quantum dots. Phys Rev Lett 86:4636–4639
10. Deo PS, Moskalets MV (2000) Features of level broadening in a ring-stub system. Phys Rev B 61:10559–10562
11. Faist J, Capasso F, Sirtori C, West KW, Pfeiffer LN (1997) Controlling the sign of quantum interference by tunneling from quantum wells. Nature 390:589–591
12. Fano U (1961) Effects of configuration interaction on intensities and phase shifts. Phys Rev 124:1866–1878
13. Fano U, Cooper JW (1965) Line profiles in the far-uv absorption spectra of the rare gases. Phys Rev 137:A1364–A1379
14. Feshbach H (1958) Unified theory of nuclear reactions. Ann Phys 5:357–390
15. Friedrich H (1990) Theoretical Atomic Physics. Springer, Berlin
16. Göres J, Goldhaber-Gordon D, Heemeyer S, Kastner MA, Shtrikman H, Mahalu D, Meirav U (2000) Fano resonances in electronic transport through a single-electron transistor. Phys Rev B 62:2188–2194
17. Jayannavar AM, Deo PS (1994) Persistent currents and conductance of a metal loop connected to electron reservoirs. Phys Rev B 49:13685–13690
18. Jayannavar AM, Deo PS (1994) Persistent currents in the presence of a transport current. Phys Rev B 51:10175–10178
19. Johnson AC, Marcus CM, Hanson MP, Gossard AC (2004) Coulomb-modified Fano resonance in a one-lead quantum dot. Phys Rev Lett 93:106803-1–106803-4
20. Keyser UF, Borck S, Haug RJ, Bichler M, Abstreiter G, Wegscheider W (2002) Aharonov–Bohm oscillations of a tuneable quantum ring. Semicond Sci Technol 17:L22–L24
21. Keyser UF, Fühner C, Borck S, Haug RJ, Bichler M, Abstreiter G, Wegscheider W (2003) Kondo effect in a few-electron quantum ring. Phys Rev Lett 90:196601-1–196601-4
22. Kim CS, Roznova ON, Satanin AM, Stenberg VB (2002) Interference of quantum states in electronic waveguides with impurities. JETP 94:992–1007
23. Kim CS, Satanin AM, Joe YS, Cosby RM (1999) Resonant tunneling in a quantum waveguide: Effect of a finite-size attractive impurity. Phys Rev B 60:10962–10970
24. Kobayashi K, Aikawa H, Katsumoto S, Iye Y (2002) Tuning of the Fano effect through a quantum dot in an Aharonov–Bohm interferometer. Phys Rev Lett 88:256806-1–256806-4
25. Kobayashi K, Aikawa H, Sano A, Katsumoto S, Iye Y (2003) Mesoscopic Fano effect in a quantum dot embedded in an Aharonov–Bohm ring. Phys Rev B 68:235304-1–235304-8
26. Kobayashi K, Aikawa H, Sano A, Katsumoto S, Iye Y (2004) Fano resonance in a quantum wire with a side-coupled quantum dot. Phys Rev B 70:035319-1–035319-6
27. Landau LD, Lifshitz EM (1965) Quantum Mechanics. Pergamon, New York
28. Landauer R (1987) Electrical transport in open and closed systems. Z Phys 68:217–228
29. Lee M, Bruder C (2006) Spin filter using a semiconductor quantum ring side coupled to a quantum wire. Phys Rev B 73:085315-1–085315-5
30. Nöckel JU, Stone AD (1994) Resonance line shapes in quasi-one-dimensional scattering. Phys Rev B 50:17415–17432
31. Olendski O, Mikhailovska (2003) Fano resonances of a curved waveguide with an embedded quantum dot. Phys Rev B 67:035310-1–035310-13
32. Orellana PA, Pacheco M (2005) Persistent current magnification in a double quantum-ring system. Phys Rev B 71:235330-1–235330-6
33. Pareek TP, Deo PS, Jayannavar AM (1995) Effect of impurities on the current magnification in mesoscopic open rings. Phys Rev B 52:14657–14663
34. Park W, Hong J (2004) Analysis of coherent current flows in the multiply connected open Aharonov–Bohm rings. Phys Rev B 69:035319-1–035319-7
35. Ryu CM, Cho SY (1998) Phase evolution of the transmission coefficient in an Aharonov–Bohm ring with Fano resonance. Phys Rev B 58:3572–3575
36. Satanin AM, Joe YS (2005) Fano interference and resonances in open systems. Phys Rev B 71:205417-1–205417-12
37. Tekman E, Bagwell PF (1993) Fano resonances in quasi-one-dimensional electron waveguides. Phys Rev B 48:2553–2559
38. Tekman E, Ciraci S (1990) Ballistic transport through a quantum point contact: Elastic scattering by impurities. Phys Rev B 42:9098–9103
39. Vargiamidis V, Polatoglou HM (2005) Conductance of a quantum wire with a Gaussian impurity potential and variable cross-sectional shape. Phys Rev B 71:075301-1–075301-16
40. Vargiamidis V, Polatoglou HM (2005) Resonances in electronic transport through a quantum wire with impurities and variable cross-sectional shape. Phys Rev B 72:195333-1–195333-12
41. Vargiamidis V, Polatoglou HM (2006) Fano resonance and persistent current in mesoscopic open rings: Influence of coupling and Aharonov–Bohm flux. Phys Rev B 74:235323-1–235323-14
42. Vargiamidis V, Polatoglou HM (2007) Temperature dependence of the Fano effect in quantum wires with short- and finite-range impurities. Phys Rev B 75:153308-1–153308-4
43. Voo KK, Chu CS (2005) Fano resonance in transport through a mesoscopic two-lead ring. Phys Rev B 72:165307-1–165307-9
44. Wharam DA, Thornton TJ, Newbury R, Pepper M, Ahmed H, Frost JEF, Hasko DG, Peacock DC, Ritchie DA, Jones GAC (1988) One-dimensional transport and the quantization of the ballistic resistance. J Phys C 21:L209–L214
45. Wu HC, Guo Y, Chen XY, Gu BL (2003) Giant persistent current in a quantum ring with multiple arms. Phys Rev B 68:125330-1–125330-5
46. Wunsch B, Chudnovskiy A (2003) Quasistates and their relation to the Dicke effect in a mesoscopic ring coupled to a reservoir. Phys Rev B 68:245317-1–245317-9
47. Xiong YS, Liang XT (2004) Fano resonance and persistent current of a quantum ring. Phys Lett A 330:307–312
48. Yi J, Wei JH, Hong J, Lee SI (2001) Giant persistent currents in the open Aharonov–Bohm rings. Phys Rev B 65:033305-1–033305-4
49. van Wees BJ, van Houten H, Beenakker CWJ, Williamson JG, Kouwenhoven LP, van der Marel D, Foxon CT (1988) Quantized conductance of point contacts in a two-dimensional electron gas. Phys Rev Lett 60:848–850

Books and Reviews

Beenakker CWJ, van Houten H (1991) Quantum transport in semiconductor nanostructures. In: Ehrenreich H, Turnbull D (eds) Solid State Physics, vol 44. Academic Press, New York

Datta S (1995) Electronic Transport in Mesoscopic Systems. Cambridge University Press, Cambridge
Economou EN (1983) Green's Functions in Quantum Physics. Springer, Heidelberg
Imry Y (1986) Physics of mesoscopic systems. In: Grinstein G, Mazenko G (eds) Directions in Condensed Matter Physics. World Scientific Press, Singapore

Reversible Cellular Automata

Kenichi Morita
Hiroshima University, Higashi-Hiroshima, Japan

Article Outline

Glossary
Definition of the Subject
Introduction
Reversible Cellular Automata
Simulating Irreversible Cellular Automata by Reversible Ones
1-D Universal Reversible Cellular Automata
2-D Universal Reversible Cellular Automata
Future Directions
Bibliography

Glossary

Cellular automaton A cellular automaton (CA) is a system consisting of a large (theoretically, infinite) number of finite automata, called cells, which are connected uniformly in a space. Each cell transits its state depending on the states of itself and the cells in its neighborhood. Thus the state transition of a cell is specified by a local function. Applying the local function to all the cells in the space synchronously, the transition of a configuration (i. e., a whole state of the cellular space) is induced. Such a transition function is called a global function. A CA is regarded as a kind of dynamical system that can deal with various kinds of spatio-temporal phenomena.

Cellular automaton with block rules A CA with block rules was proposed by Margolus [18], and it is often called a CA with Margolus-neighborhood. The cellular space is divided into infinitely many blocks of the same size (in the two-dimensional case, e. g., 2×2). A local transition function consisting of "block rules", which is a mapping from a block state to a block state, is applied to all the blocks in parallel. At the next time step, the block division pattern is shifted by some fixed amount (e. g., to the north-east direction by one cell), and the same local function is applied to them. This model of CA is convenient to design a reversible CA. Because, if the local transition function is injective, then the resulting CA is reversible.

Partitioned cellular automaton A partitioned cellular automaton (PCA) is a framework for designing a reversible CA. It is a subclass of a usual CA where each cell is partitioned into several parts, whose number is equal to the neighborhood size. Each part of a cell has its own state set, and can be regarded as an output port to a specified neighboring cell. Depending only on the corresponding parts (not on the entire states) of the neighboring cells, the next state of each cell is determined by a local function. We can see that if the local function is injective, then the resulting PCA is reversible. Hence, a PCA makes it feasible to construct a reversible CA.

Reversible cellular automaton A reversible cellular automaton (RCA) is defined as a one whose global function is injective (i. e., one-to-one). It can be regarded as a kind of a discrete model of reversible physical space. It is in general difficult to construct an RCA with a desired property such as computation-universality. Therefore, the frameworks of a CA with Margolus neighborhood, a partitioned cellular automaton, and others are often used to design RCAs.

Universal cellular automaton A CA is called computationally universal, if it can compute any recursive function by giving an appropriate initial configuration. Equivalently, it is also defined as a CA that can simulate a universal Turing machine. Universality of RCAs can be proved by simulating other systems such as arbitrary (irreversible) CAs, reversible Turing machines, reversible counter machines, and reversible logic elements and circuits, which have already been known to be universal.

Definition of the Subject

Reversible cellular automata (RCAs) are defined as cellular automata (CAs) with an injective global function. Every configuration of an RCA has exactly one previous configuration, and thus RCAs are "backward deterministic" CAs. The notion of reversibility originally comes from physics. It is one of the fundamental microscopic physical laws of Nature. In this sense, an RCA is thought as an abstract model of a physically reversible space as well as a computing model. It is very important to investigate how computation can be carried out efficiently and elegantly in a sys-

tem having reversibility. This is because future computing devices will surely become those of a nanoscale size.

In this article, we mainly discuss on the properties of RCAs from the computational aspects. In spite of the strong constraint of reversibility, RCAs have very rich ability of computing. We can see that even very simple RCAs have universal computing ability. We can also recognize, in some reversible cellular automata, computation is carried out in a very different manner from conventional computing systems, thus they may give new ways and concepts for future computing.

Introduction

Problems related to injectivity and surjectivity of global functions of CAs were first studied by Moore [22] and Myhill [33] in the Garden-of-Eden problem. A Garden-of-Eden configuration is such that it can exist only at time 0. Therefore, if a CA has such a configuration, then its global function is not surjective, and vice versa. They proved the following Garden-of-Eden theorem: A CA has a Garden-of-Eden configuration, if and only if it has an "erasable configuration". After that, many researchers studied on injectivity and surjectivity of global functions more generally [1,19,20,34]. In particular, Richardson [34] showed that if a CA is injective, then it is surjective.

Toffoli [36] first studied reversible (i. e., injective) CAs from the computational viewpoint. He showed that every k-dimensional irreversible CA can be simulated by a $k+1$-dimensional RCA. Hence, a two-dimensional RCA has universal computing ability. Since then, extensive studies on RCAs have been done until now.

After the pioneering work of Bennett [3] on reversible Turing machines, several models of reversible computing were proposed besides RCAs. They are, for example, reversible logic circuits [8,26,37], Billiard Ball Model of computing [8], and reversible counter machines [25]. Most of these models have close relation to physical reversibility. In fact, reversible computing plays an important role when considering inevitable power dissipation in computing [3,4,5,16,38]. It is also one of the basis of quantum computing (see e. g., [9]) because an evolution of a quantum system is a reversible process.

In this article, we discuss how RCAs can have universal computing ability, and how simple they can be. Since reversibility is one of the fundamental microscopic properties of physical systems, it is important to investigate whether we can use such physical mechanisms directly for computation. An RCA is a useful framework to formalize and investigate these problems. Since this article is not an exhaustive survey, many interesting topics related to RCAs, such as complexity of RCA [35], relations to quantum CA (e. g., [41]), etc., are omitted.

An outline of the following sections is as follows. In Sect. "Reversible Cellular Automata", we give basic definitions on RCAs, and design methods for obtaining RCAs. In Sect. "Simulating Irreversible Cellular Automata by Reversible Ones", it is shown how irreversible CAs are simulated by RCAs. In Sect. "1-D Universal Reversible Cellular Automata", two kinds of computation-universal one-dimensional RCAs are shown. In Sect. "2-D Universal Reversible Cellular Automata", several universal two-dimensional RCAs with simple local functions are shown. In Sect. "Future Directions", we discuss future directions and open problems as well as some other problems on RCAs not given in the previous sections.

Reversible Cellular Automata

Formal Definitions

We first give definitions on conventional cellular automata, and then their reversibility.

Definition 1 A *deterministic k-dimensional (k-D) m-neighbor cellular automaton (CA)* is a system defined by

$$A = (\mathbb{Z}^k, Q, (n_1, \ldots, n_m), f, \#) ,$$

where $\mathbb{Z}$ is the set of all integers (hence $\mathbb{Z}^k$ is the set of all k-dimensional points with integer coordinates at which cells are placed), Q is a non-empty finite set of states of each cell, $(n_1, \ldots, n_m)$ is an element of $(\mathbb{Z}^k)^m$ called a neighborhood ($m = 1, 2, \ldots$), $f\colon Q^m \to Q$ is a local function, and $\# \in Q$ is a quiescent state satisfying $f(\#, \ldots, \#) = \#$.

A *k-dimensional configuration* over the set Q is a mapping $\alpha\colon \mathbb{Z}^k \to Q$. Let $\mathrm{Conf}_k(Q)$ denote the set of all k-dimensional configurations over Q, i. e., $\mathrm{Conf}_k(Q) = \{\alpha \mid \alpha\colon \mathbb{Z}^k \to Q\}$. If k is understood, we write it by $\mathrm{Conf}(Q)$. We say that a configuration α is *finite* iff the set $\{x \mid x \in \mathbb{Z}^k \wedge \alpha(x) \neq \#\}$ is finite. Otherwise, it is called *infinite*.

The global function $F\colon \mathrm{Conf}(Q) \to \mathrm{Conf}(Q)$ of A is defined as the one that satisfies the following formula.

$$\forall \alpha \in \mathrm{Conf}(Q),\ x \in \mathbb{Z}^k:$$
$$F(\alpha)(x) = f(\alpha(x + n_1), \ldots, \alpha(x + n_m))$$

Definition 2 Let $A = (\mathbb{Z}^k, Q, (n_1, \ldots, n_m), f, \#)$ be a CA. (1) A is called an *injective CA* iff F is one-to-one. (2) A is called an *invertible CA* iff there is a CA $A' =$

$(\mathbb{Z}^k, Q, N', f', \#)$ that satisfies the following condition:

$$\forall \alpha, \beta \in \mathrm{Conf}(Q)\colon F(\alpha) = \beta \quad \text{iff} \quad F'(\beta) = \alpha\,,$$

where F and F' are the global functions of A and A', respectively.

The following theorem can be derived from the results independently proved by Hedlund [10], and Richardson [34].

Theorem 3 (Hedlund [10] and Richardson [34]) *Let A be a CA. A is injective iff it is invertible.*

By the above theorem, we see the notions of injectivity and invertibility are equivalent. Henceforth, we use the terminology "reversible CA" (RCA) for such a CA, instead of injective CA or invertible CA, because an RCA is regarded as an analog of physically reversible space. (Note that, in some other computing models such as Turing machines, counter machines, and logic circuits, injectivity is trivially equivalent to invertibility, if they are suitably defined. Therefore, for these models, we can directly define reversibility without introducing the notions of injectivity and invertibility.)

How Can We Find RCAs?

The class of RCAs is a special subclass of CAs. Therefore there arises a problem how we can find or construct RCAs with some desired property. It is in general hard to do so if we use the conventional framework of CAs. Because, the following result is shown by Kari [14] for the two-dimensional case (hence it also holds for higher dimensional CAs).

Theorem 4 (Kari [14]) *The problem whether a given two-dimensional CA is reversible is undecidable.*

For the case of one-dimensional CA, Amoroso and Patt [2] showed it is decidable.

Theorem 5 (Amoroso and Patt [2]) *There is an algorithm to test whether a given one-dimensional CA is reversible or not.*

There are also several studies on enumerating all reversible one-dimensional CAs (e. g., [6,23]). But, it is generally difficult to find RCAs with specific properties such as computation-universality, even for the one-dimensional case.

In order to make it feasible to design an RCA, several methods have been proposed until now. They are, for example, CAs with block rules [18,38], partitioned CAs [28], CAs with second order rules [18,38,39], and others (see e. g., [38]). Here, we describe the first two methods in some detail.

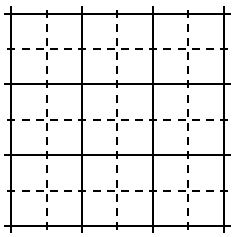

Reversible Cellular Automata, Figure 1
A cellular space with the Margolus neighborhood

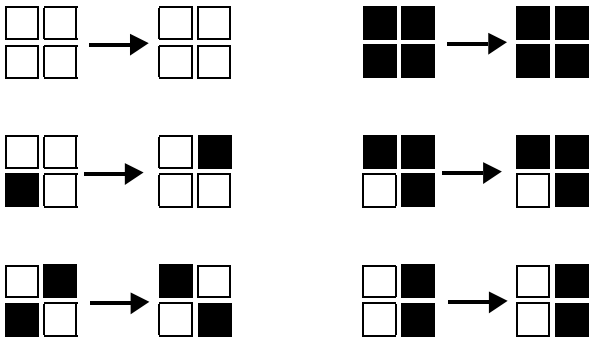

Reversible Cellular Automata, Figure 2
Block rules for the Margolus RCA [18]. (Rotation-symmetry is assumed here. Hence, rules obtained by rotating both sides of a rule are also included.)

Cellular Automata with Block Rules Margolus [18] proposed an interesting variant of a CA, by which he composed a computation-universal two-dimensional two-state reversible CA. In his model, all the cells are grouped into "blocks" of size 2×2 as shown in Fig. 1. A specific example of a transformation specified by "block rules" is shown in Fig. 2. This CA evolves as follows: At time 0 the local transformation is applied to every solid line block, then at time 1 to every dotted line block, and so on, alternately. Since this local transformation is one-to-one, the global function of the CA is also one-to-one. Such a neighborhood is called Margolus neighborhood.

One can obtain reversible CAs, by giving one-to-one block rules. However, CAs with Margolus neighborhood are not conventional CAs, because each cell should know the relative position in a block and the parity of time besides its own state.

Related to this topic, Kari [15] showed that every one- and two-dimensional RCA can be represented by a block permutations and translations.

Partitioned Cellular Automata The method of using partitioned cellular automata (PCA) has some similarity to the one that uses block rules. However, resulting re-

versible CAs are in the framework of conventional CA (in other words, a PCA is a special case of a CA). In addition, flexibility of neighborhood is rather high. Shortcomings of PCA is that, in general, the number of states per cell becomes large.

Definition 6 A *deterministic k-dimensional m-neighbor* partitioned cellular automaton *(PCA)* is a system defined by

$$P = (\mathbb{Z}^k, (Q_1, \ldots, Q_m), (n_1, \ldots, n_m), f, (\#_1, \ldots, \#_m)),$$

where $\mathbb{Z}$ is the set of all integers, Q_i $(i = 1, \ldots, m)$ is a non-empty finite set of states of the i-th part of each cell (thus the state set of each cell is $Q = Q_1 \times \cdots \times Q_m$), $(n_1, \ldots, n_m) \in (\mathbb{Z}^k)^m$ is a neighborhood, $f\colon Q \to Q$ is a local function, and $(\#_1, \ldots, \#_m) \in Q$ is a quiescent state satisfying $f(\#_1, \ldots, \#_m) = (\#_1, \ldots, \#_m)$. (In general, the states $\#_1, \ldots, \#_m$ may be different from each other. However, by renaming the states in each part appropriately, we can identify the states $\#_1, \ldots, \#_m$ as representing the same state $\#$. In what follows, we often assume so, and write the quiescent state by $(\#, \ldots, \#)$.)

The notion of a finite (or infinite) configuration is defined similarly as in CA. Let $p_i\colon Q \to Q_i$ be the projection function such that $p_i(q_1, \ldots, q_m) = q_i$ for all $(q_1, \ldots, q_m) \in Q$. The global function $F\colon \mathrm{Conf}(Q) \to \mathrm{Conf}(Q)$ of P is defined as the one that satisfies the following formula.

$$\forall \alpha \in \mathrm{Conf}(Q), x \in \mathbb{Z}^k:$$
$$F(\alpha)(x) = f(p_1(\alpha(x + n_1)), \ldots, p_m(\alpha(x + n_m)))$$

By the above definition, a one-dimensional radius 1 (3-neighbor) PCA $P_{1\mathrm{d}}$ can be defined as follows.

$$P_{1\mathrm{d}} = (\mathbb{Z}, (L, C, R), (1, 0, -1), f, (\#, \#, \#))$$

Each cell is divided into three parts, i. e., left, center, and right parts, and their state sets are L, C, and R. The next state of a cell is determined by the present states of the left part of the right-neighbor cell, the center part of this cell, and the right part of the left-neighbor cell (not depending on the whole three parts of the three cells). Figure 3 shows its cellular space, and how the local function f works.

Let $(l, c, r), (l', c', r') \in L \times C \times R$. If $f(l, c, r) = (l', c', r')$, then this equation is called a local rule (or simply a rule) of the PCA $P_{1\mathrm{d}}$, and it is sometimes written in a pictorial form as shown in Fig. 4. Note that, in the pictorial representation, the arguments of the lefthand side of $f(l, c, r) = (l', c', r')$ appear in a reverse order.

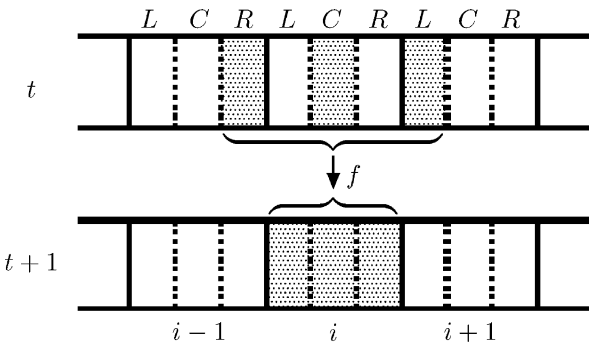

Reversible Cellular Automata, Figure 3
Cellular space of a one-dimensional 3-neighbor PCA $P_{1\mathrm{d}}$, and its local function f

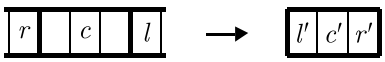

Reversible Cellular Automata, Figure 4
A pictorial representation of a local rule $f(l, c, r) = (l', c', r')$ of a one-dimensional 3-neighbor PCA $P_{1\mathrm{d}}$

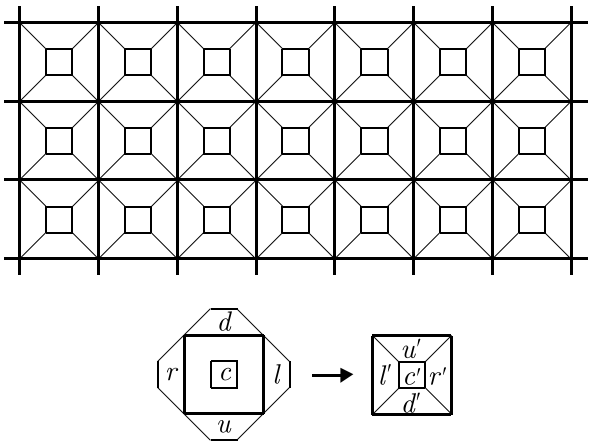

Reversible Cellular Automata, Figure 5
Cellular space of a two-dimensional 5-neighbor PCA $P_{2\mathrm{d}}$, and its local rule

Similarly, a two-dimensional PCA $P_{2\mathrm{d}}$ with Neumann-like neighborhood is defined as follows.

$$P_{2\mathrm{d}} = (\mathbb{Z}^2, (C, U, R, D, L), ((0, 0), (0, -1), (-1, 0), (0, 1), (1, 0)), f, (\#, \#, \#, \#, \#))$$

Figure 5 shows the cellular space of $P_{2\mathrm{d}}$, and a pictorial representation of a rule $f(c, u, r, d, l) = (c', u', r', d', l')$.

Let $P = (\mathbb{Z}^k, (Q_1, \ldots, Q_m), (n_1, \ldots, n_m), f, (\#_1, \ldots, \#_m))$ be a k-dimensional PCA, and F be its global function. It is easy to show the following proposition (a proof for

the one-dimensional case given in [28] can be extended to higher dimensions).

Proposition 7 *The local function f is one-to-one, iff the global function F is one-to-one.*

It is also easy to see that the class of PCAs is a subclass of CAs. More precisely, the following proposition is derived by extending the domain of the local function of P.

Proposition 8 *For any k-dimensional m-neighbor PCA P, we can obtain a k-dimensional m-neighbor CA A whose global function is identical with that of P.*

By above, if we want to construct an RCA, it is sufficient to give a PCA whose local function f is one-to-one. This makes a design of an RCA feasible.

Simulating Irreversible Cellular Automata by Reversible Ones

Toffoli [36] first showed that for every irreversible CA there exists a reversible one that simulates the former by increasing the dimension by one. From this result, computation-universality of two-dimensional RCA is derived, since it is easy to embed a Turing machine in a (irreversible) one-dimensional CA.

Theorem 9 (Toffoli [36]) *For any k-dimensional (irreversible) CA A, we can construct a $k+1$-dimensional RCA A' that simulates A in real time.*

Although Toffoli's proof is rather complex, the idea of the proof is easily implemented by using a PCA. Here we explain it informally. Consider a one-dimensional 3-neighbor irreversible CA A that evolves as in Fig. 6. Then, we can construct a two-dimensional reversible PCA P that simulates A as shown in Fig. 7. The configuration of A is kept in some row of P. A state of each cell of A is stored in left, center, and right parts of a cell in P in triplicate. By this, each cell of P can compute the next state of the corresponding cell of A correctly. At the same time, the previous states of the cell and the left and right neighbor cells (which were used to compute the next state) are put downward as a "garbage" signal to keep P reversible. In other words, the additional dimension is used to record all the past history of the evolution of A. In this way, P can simulate A reversibly.

$t=0$	q_1^0	q_2^0	q_3^0	q_4^0
$t=1$	q_1^1	q_2^1	q_3^1	q_4^1
$t=2$	q_1^2	q_2^2	q_3^2	q_4^2
$t=3$	q_1^3	q_2^3	q_3^3	q_4^3

Reversible Cellular Automata, Figure 6
An example of an evolution in an irreversible one-dimensional CA *A*

As for one-dimensional CA with finite configuration, reversible simulation is possible without increasing the dimension.

Theorem 10 (Morita [24]) *For any one-dimensional (irreversible) CA A with finite configurations, we can construct a one-dimensional RCA A' that simulates A (but not in real time).*

1-D Universal Reversible Cellular Automata

Simulating Reversible Turing Machines by 1-D RCAs

It is possible to prove computation-universality of one-dimensional RCAs by constructing RCAs that can simulate reversible Turing machines directly. Here, we first give definitions on reversible Turing machines, and then show how they can be simulated by RCAs.

Bennett [3] showed a nice construction method of a reversible Turing machine that simulates a given irreversible Turing machine, and never leaves garbage signals on its tape at the end of computation. We now give a definition of a one-tape Turing machine and its reversible version (a multi-tape reversible Turing machine can be also defined similarly). It is convenient to use quadruple formalism [3] of a Turing machine to define a reversible one.

Definition 11 A one-tape *Turing machine* (TM) is defined by

$$T = (Q, S, q_0, q_a, q_r, s_0, \delta)\,,$$

where Q is a non-empty finite set of states, S is a non-empty finite set of symbols, q_0 is an initial state ($q_0 \in Q$), q_a is an accepting state ($q_a \in Q$), q_r is a rejecting state ($q_r \in Q$), s_0 is a special blank symbol ($s_0 \in S$), and δ is a move relation which is a subset of $(Q \times S \times S \times Q) \cup (Q \times \{/\} \times \{-, 0, +\} \times Q)$. Each element of δ is called a quadruple, and either of the form $[q_i, s, s', q_j] \in (Q \times S \times S \times Q)$ or $[q_i, /, d, q_j] \in (Q \times \{/\} \times \{-, 0, +\} \times Q)$. The symbols "$-$", "0", and "$+$" denote left-shift, zero-shift, and right-shift, respectively. $[q_i, s, s, q_j]$ means that if T reads the symbol s in the state q_i, then write s' and go to the state q_j. $[q_i, /, d, q_j]$ means that if T is in the state q_i, then shift the head to the direction d and go to the state q_j.

T is called deterministic iff the following statement holds for any pair of distinct quadruples $[p_1, b_1, c_1, q_1]$

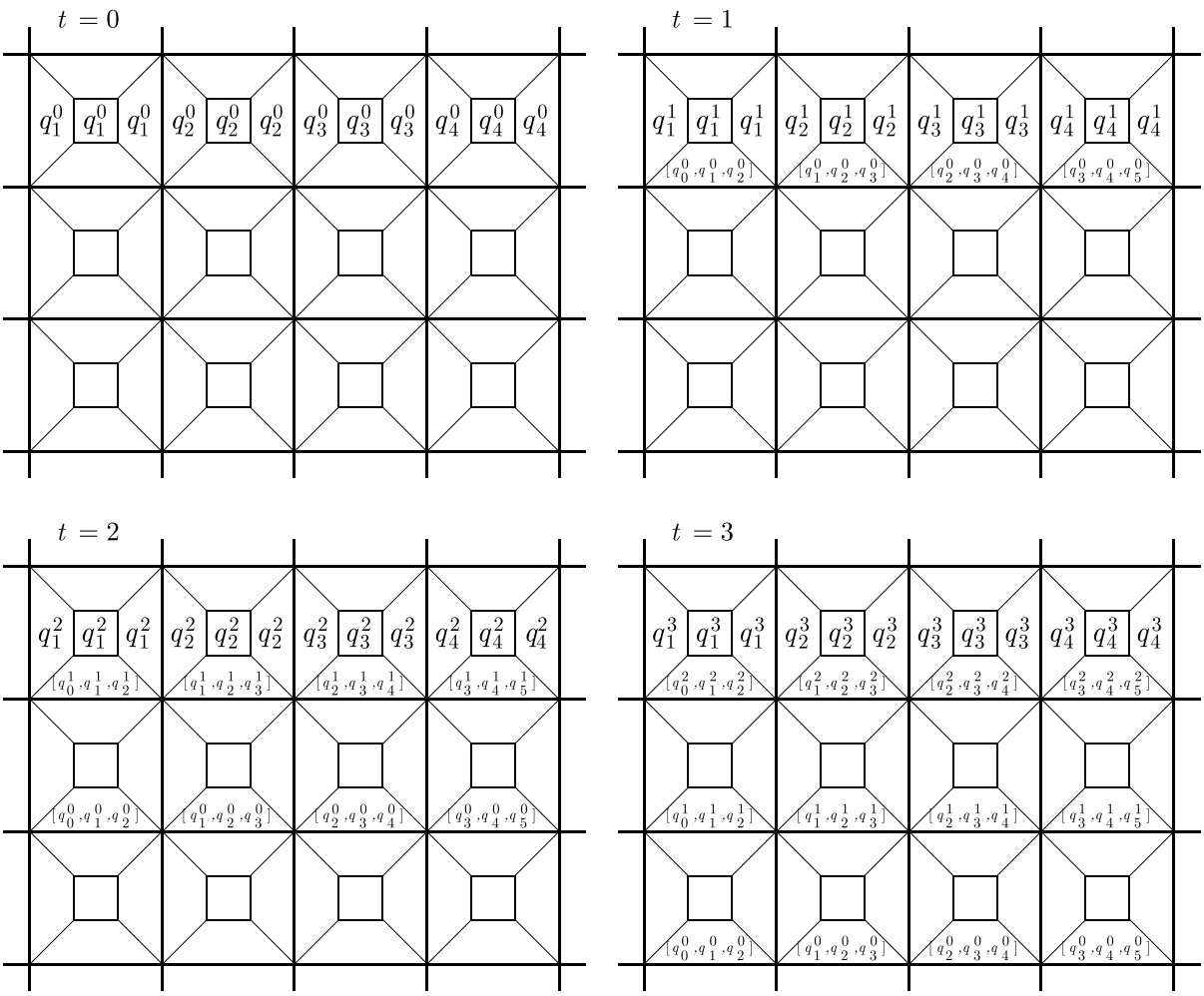

Reversible Cellular Automata, Figure 7
Simulating the irreversible CA *A* in Fig. 6 by a two-dimensional reversible PCA *P*

and $[p_2, b_2, c_2, q_2]$.

$$\text{If } p_1 = p_2\,, \text{ then } b_1 \neq / \wedge b_2 \neq / \wedge b_1 \neq b_2\,.$$

On the other hand, T is called reversible iff the following statement holds for any pair of distinct quadruples $[p_1, b_1, c_1, q_1]$ and $[p_2, b_2, c_2, q_2]$.

$$\text{If } q_1 = q_2\,, \text{ then } b_1 \neq / \wedge b_2 \neq / \wedge c_1 \neq c_2\,.$$

Hereafter, we consider only deterministic Turing machines. The next theorem shows computation-universality of a reversible three-tape Turing machine.

Theorem 12 (Bennett [3]) *For any (irreversible) one-tape Turing machine, there is a reversible three-tape Turing machine which simulates the former.*

It is also shown in [31] that for any irreversible one-tape TM, there is a reversible one-tape two-symbol TM which simulates the former. In fact, to prove computation-universality of a one-dimensional reversible PCA, it is convenient to simulate a reversible one-tape TM.

Theorem 13 (Morita and Harao [28]) *For any reversible one-tape Turing machine T, there is a one-dimensional reversible PCA P that simulates the former.*

We show how P simulates T (the method given below is slightly modified from the one in [28]). Let $T = (Q, S, q_0, q_a, q_r, s_0, \delta)$ be a reversible one-tape TM. We assume that q_0 does not appear as a fourth element in any quintuple in δ (because we can always construct such a reversible TM from an irreversible one [31]). A reversible PCA $P = (\mathbb{Z}, (L, C, R), (1, 0, -1), f, (\#, s_0, \#))$ that simu-

lates T is as follows. The state sets L, C, and R are:

$$L = R = Q \cup \{*, \#\}\,, \quad C = S \cup (Q \times S)\,.$$

The local function f is as below. Note that, in (3)–(6), if $p \in \{q_0\}$ then $x = *$, else $x = \#$. Likewise, if $q \in \{q_a, q_r\}$ then $y = *$, else $y = \#$.

(1) For every $s \in S$,

$$f(\#, s, \#) = (\#, s, \#)\,,$$
$$f(\#, s, *) = (\#, s, *)\,,$$
$$f(*, s, \#) = (*, s, \#)\,.$$

(2) For every $q \in \{q_0, q_a, q_r\}$ and $s \in S$,

$$f(\#, [q, s], \#) = (\#, [q, s], \#)\,.$$

(3) For every $p, q \in Q$ and $s, t \in S$, if $[p, s, t, q] \in \delta$,

$$f(\#, [p, s], x) = (y, [q, t], \#)\,.$$

(4) For every $p, q \in Q$ and $s, t \in S$, if $[p, /, -, q] \in \delta$,

$$f(\#, [p, s], x) = (q, s, \#)\,,$$
$$f(q, t, \#) = (y, [q, t], \#)\,.$$

(5) For every $p, q \in Q$ and $s \in S$, if $[p, /, 0, q] \in \delta$,

$$f(\#, [p, s], x) = (y, [q, s], \#)\,.$$

(6) For every $p, q \in Q$ and $s, t \in S$, if $[p, /, +, q] \in \delta$,

$$f(\#, [p, s], x) = (\#, s, q)\,,$$
$$f(\#, t, q) = (y, [q, t], \#)\,.$$

We can see that the right-hand side of each rule differs from that of any other rule, because T is deterministic and reversible. The rules in (3)–(6) are for simulating T step by step. If the initial computational configuration of T is

$$\cdots s_0 t_1 \cdots q_0 t_i \cdots t_n s_0 \cdots$$

then set P to the following configuration.

$$\ldots, (\#, s_0, \#), (\#, t_1, \#), \ldots, (\#, t_{i-2}, \#), (\#, t_{i-1}, *),$$
$$(\#, [q_0, t_i], \#), (\#, t_{i+1}, \#), \ldots, (\#, t_n, \#), (\#, s_0, \#), \ldots$$

The simulation process starts when a right-moving signal $*$ meets a center state of the form $[q_0, t_i]$. It is easily seen that, from this configuration, P can correctly simulates T by the rules in (3)–(6). If T becomes a halting state q_a or q_r, then a left-moving signal $*$ is emitted, and the final computational configuration of T is kept in all the successive configurations of P. Note that P itself cannot halt (i. e., it cannot keep exactly the same configuration after reaching it from a different configuration), because P is reversible.

Example 14 Consider a reversible TM $T_{\text{parity}} = (Q, \{0, 1\}, q_0, q_a, q_r, 0, \delta)$, where $Q = \{q_0, q_1, q_2, q_3, q_4, q_a, q_r\}$, and δ is as below.

$$\delta = \{[\,q_0, 0, 1, q_1\,], [\,q_1, /, +, q_2\,], [\,q_2, 0, 1, q_a\,], [\,q_2, 1, 0, q_3\,], [\,q_3, /, +, q_4\,], [\,q_4, 0, 1, q_r\,], [\,q_4, 1, 0, q_1\,]\}\,.$$

For a given unary number n on the tape, T_{parity} checks if n is even or odd. If it is even, then T_{parity} halts in the accepting state q_a, otherwise halts in the rejecting state q_r. All the symbols read by T_{parity} are complemented. See Fig. 8. A simulation process of T_{parity} by a reversible PCA P_{parity}, which is constructed by the method described above, is shown in Fig. 9.

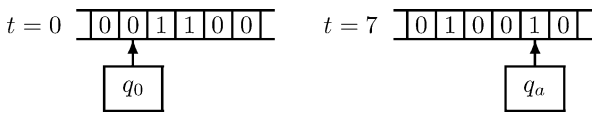

Reversible Cellular Automata, Figure 8
The initial and the final computational configuration of T_{parity} for a given unary input 11

$t = 0$		0	*		q_0 0			1			1			0			0	
$t = 1$		0			q_1 1			1			1			0			0	
$t = 2$		0			1	q_2		1			1			0			0	
$t = 3$		0			1			q_2 1			1			0			0	
$t = 4$		0			1			q_3 0			1			0			0	
$t = 5$		0			1			0	q_4		1			0			0	
$t = 6$		0			1			0			q_4 1			0			0	
$t = 7$		0			1			0			q_1 0			0			0	
$t = 8$		0			1			0			0	q_2		0			0	
$t = 9$		0			1			0			0			q_2 0			0	
$t = 10$		0			1			0			0		*	q_a 1			0	
$t = 11$		0			1			0		*	0			q_a 1			0	
$t = 12$		0			1		*	0			0			q_a 1			0	
$t = 13$		0		*	1			0			0			q_a 1			0	
$t = 14$	*	0			1			0			0			q_a 1			0	

Reversible Cellular Automata, Figure 9
Simulating T_{parity} by a one-dimensional reversible PCA P_{parity}. The state # is indicated by a blank

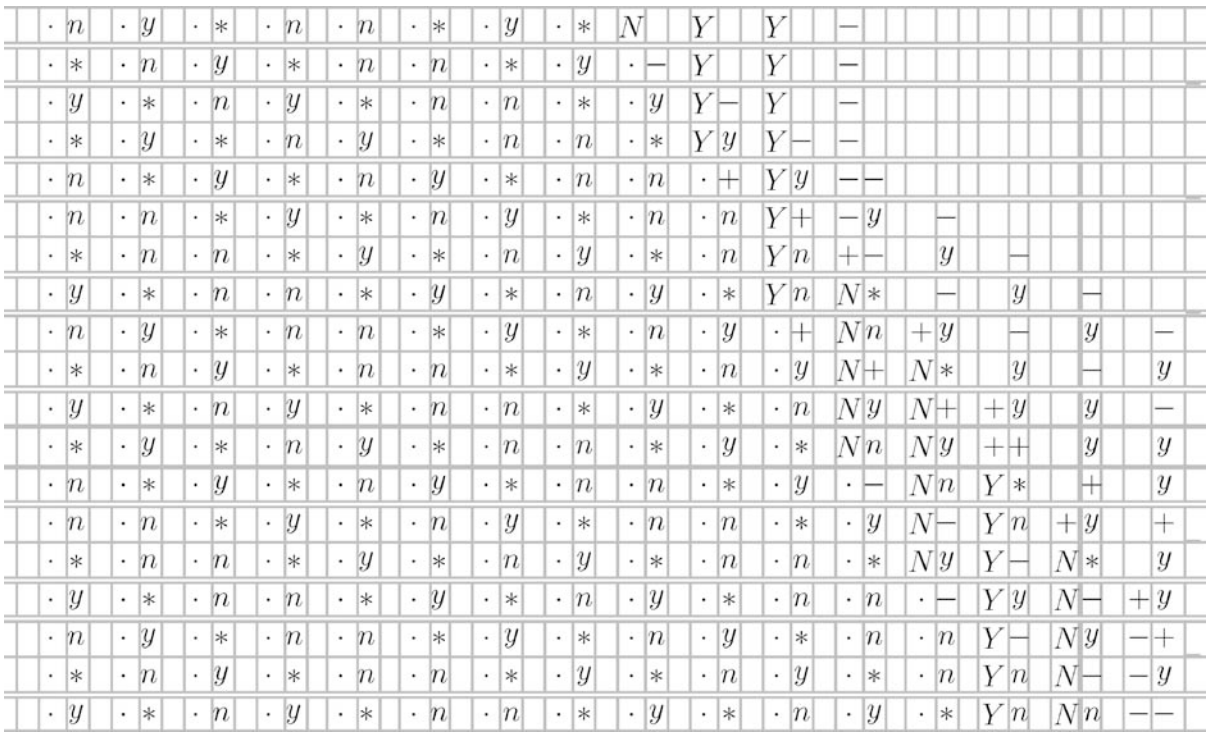

Reversible Cellular Automata, Figure 10
Simulating the CTAG C_0 by the reversible PCA P_{36} [27]

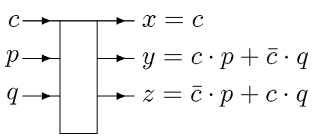

Reversible Cellular Automata, Figure 11
A Fredkin gate [8]

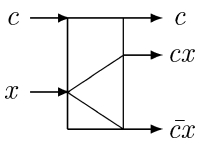

Reversible Cellular Automata, Figure 12
A switch gate [8]

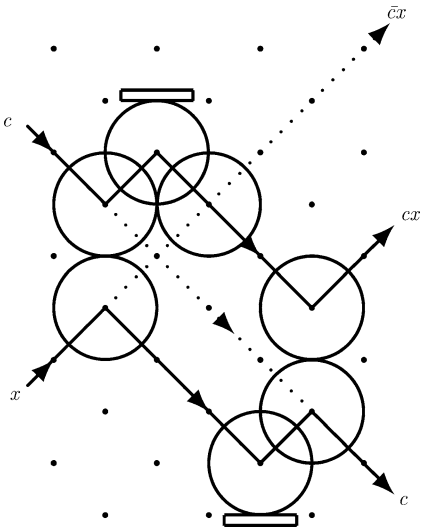

Reversible Cellular Automata, Figure 13
A switch gate realized in the Billiard Ball Model [8]

Simulating Cyclic Tag Systems by 1-D RCAs

From Theorem 12 we can see the existence of a universal reversible TM, and thus from Theorem 13 there is a one-dimensional universal RCA. Then the following problem arises: How simple can it be? To get a universal RCA with a small number (say several dozen) of states, we need another useful framework of a universal system.

A cyclic tag system (CTAG) is proposed by Cook [7] to show universality of the elementary cellular automaton of rule 110. As we shall see, it is also useful for constructing simple universal RCAs.

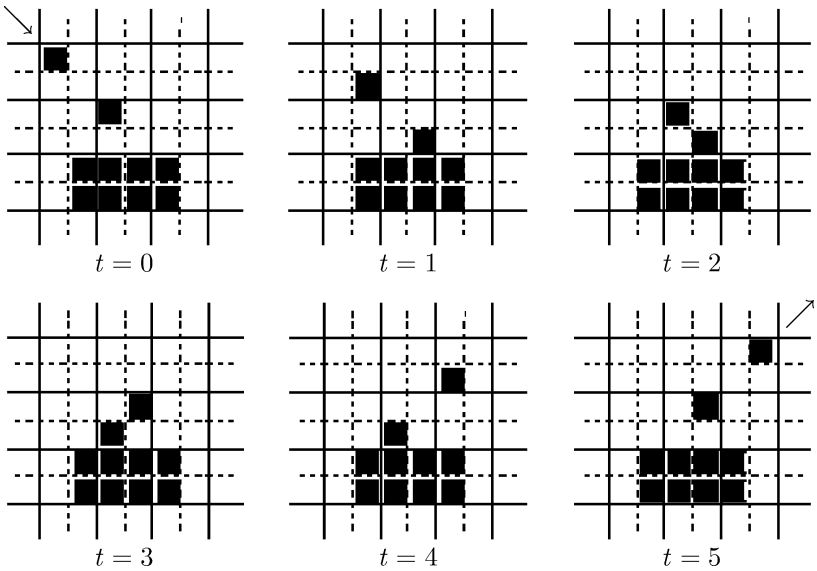

Reversible Cellular Automata, Figure 14
Reflection of a ball by a reflector in the Margolus' RCA [18]

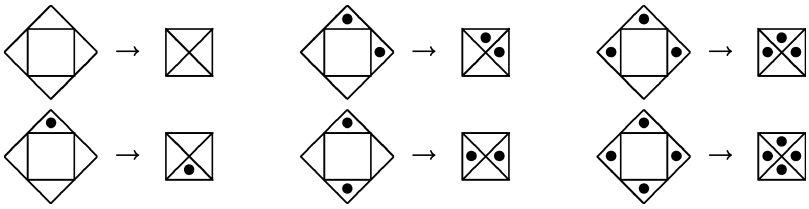

Reversible Cellular Automata, Figure 15
The local function of the 2^4-state rotation-symmetric reversible PCA S_1 [30]

Definition 15 A *cyclic tag system* (CTAG) is defined by $C = (k, \{Y, N\}, (p_0, \ldots, p_{k-1}))$, where k ($k = 1, 2, \ldots$) is the length of a cycle (i. e., period), $\{Y, N\}$ is the (fixed) alphabet, and $(p_0, \ldots, p_{k-1}) \in (\{Y, N\}^*)^k$ is a k-tuple of production rules.

A pair (v, m) is called an *instantaneous description* (ID) of C, where $v \in \{Y, N\}^*$ and $m \in \{0, \ldots, k-1\}$. m is called a *phase* of the ID. A transition relation $\Rightarrow$ on the set of IDs is defined as follows. For any $(v, m), (v', m') \in \{Y, N\}^* \times \{0, \ldots, k-1\}$,

$$\begin{aligned}
&(Yv, m) \Rightarrow (v', m') \\
&\quad \text{iff } [m' = m + 1 \bmod k] \wedge [v' = vp_m], \\
&(Nv, m) \Rightarrow (v', m') \\
&\quad \text{iff } [m' = m + 1 \bmod k] \wedge [v' = v] .
\end{aligned}$$

A sequence of IDs $(v_0, m_0), (v_1, m_1), \ldots$ is called a *computation starting from* $v \in \{Y, N\}^*$ iff $(v_0, m_0) = (v, 0)$ and $(v_i, m_i) \Rightarrow (v_{i+1}, m_{i+1})(i = 0, 1, \ldots)$. (In what follows, we write a computation by $(v_0, m_0) \Rightarrow (v_1, m_1) \Rightarrow \cdots$.)

A CTAG is a variant of a classical tag system (see e. g., [21]), where production rules are applied cyclically. If the first symbol of a host (i. e., rewritten) string is Y, then it is removed and a specified string at that phase is attached to the end of the host string. If it is N, then it is simply removed and no string is attached.

Example 16 Let us consider the following CTAG.

$$C_0 = (3, \{Y, N\}, (Y, NN, YN)) .$$

If we give NYY to C_0 as an initial string, then

$$\begin{aligned}
(NYY, 0) &\Rightarrow (YY, 1) \Rightarrow (YNN, 2) \\
&\Rightarrow (NNYN, 0) \Rightarrow (NYN, 1) \Rightarrow (YN, 2)
\end{aligned}$$

is an initial segment of a computation starting from NYY.

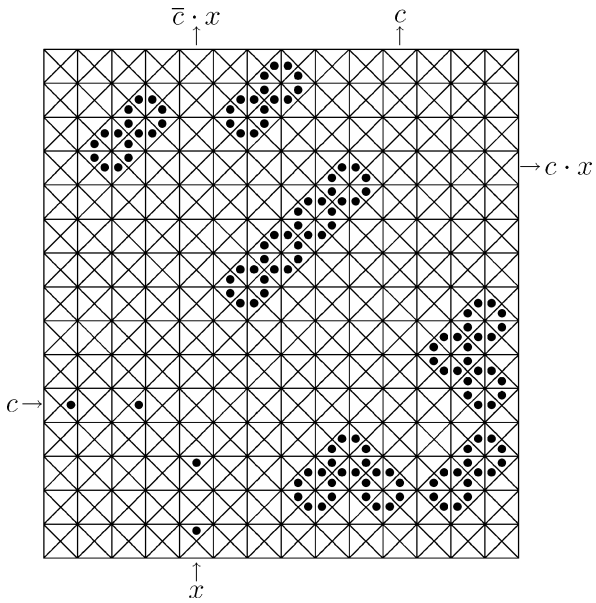

Reversible Cellular Automata, Figure 16
A switch gate realized in the reversible PCA S_1 [30]

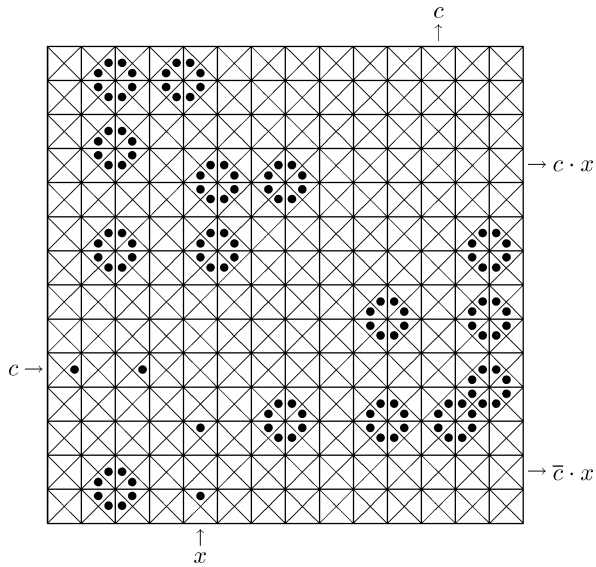

Reversible Cellular Automata, Figure 18
A switch gate realized in the reversible PCA S_1 [30]

Minsky [21] proved that a 2-tag system, which is a special class of classical tag systems, is universal. The following theorem shows the universality of a CTAG.

Theorem 17 (Cook [7]) *For any 2-tag system, there is a CTAG that simulates the former.*

It was shown by Morita [27] that there are universal one-dimensional RCAs that can simulate any CTAG.

Theorem 18 (Morita [27]) *There is a 36-state one-dimensional reversible PCA P_{36} that can simulate any CTAG on infinite (leftward-periodic) configurations.*

Theorem 19 (Morita [27]) *There is a 98-state one-dimensional reversible PCA P_{98} that can simulate any CTAG on finite configurations. (Note: it can also handle halting of a CTAG.)*

The reversible PCA P_{36} in Theorem 18 is given below.

$$P_{36} = (\mathbb{Z}, (\{\#\}, \{\#, Y, N, +, -, \bullet\}, \{\#, y, n, +, -, *\}), (1, 0, -1), f, (\#, \#, \#)),$$

where f is defined as follows.

$$\begin{aligned}
f(\#, c, \#) &= (\#, c, \#) && \text{for } c \in \{\#, Y, N, +, -, \bullet\} \\
f(\#, c, r) &= (\#, c, r) && \text{for } c \in \{\#, Y, N\}, \\
& && \text{and } r \in \{y, n, +, -\} \\
f(\#, -, r) &= (\#, -, r) && \text{for } r \in \{y, n\} \\
f(\#, \bullet, r) &= (\#, \bullet, r) && \text{for } r \in \{y, n, *\} \\
f(\#, Y, *) &= (\#, \bullet, +) \\
f(\#, N, *) &= (\#, \bullet, -) \\
f(\#, c, r) &= (\#, r, c) && \text{for } c, r \in \{+, -\} \\
f(\#, +, y) &= (\#, Y, *) \\
f(\#, +, n) &= (\#, N, *) \\
f(\#, \#, *) &= (\#, +, y)
\end{aligned}$$

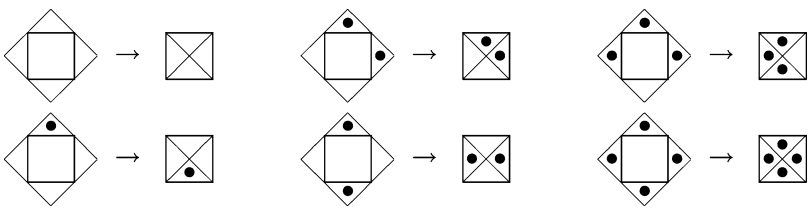

Reversible Cellular Automata, Figure 17
The local function of the 2^4-state rotation-symmetric reversible PCA S_2 [30]

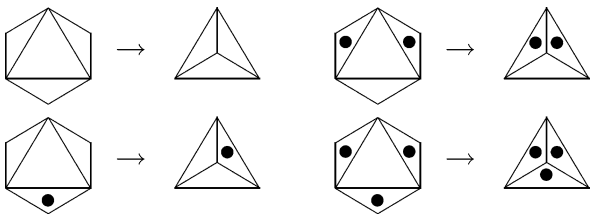

Reversible Cellular Automata, Figure 19
The local function of the 2^3-state rotation-symmetric reversible triangular PCA T_1 [12]

It is easy to see that f is injective. Note that since the state set of the left part of a cell of P_{36} has only one element #, it is actually a two-neighbor PCA.

Consider the CTAG C_0 in Example 16. The computation in C_0 starting from NYY is simulated in P_{36} as shown in Fig. 10. The initial string NYY is placed in the center parts of some consecutive cells of P_{36}. The production rules (Y, NN, YN) in C_0 are given in the reverse order in the right parts of consecutive cells. The state $*$ is used to delimit the production rules. Since the production rules are applied cyclically, infinite copies of the state sequence "$ny * nn * y*$" should be given to the left half of the cellular space of P_{36}. Figure 10 shows an initial segment $(NYY, 0) \Rightarrow (YY, 1) \Rightarrow (YNN, 2) \Rightarrow (NNYN, 0) \Rightarrow (NYN, 1) \Rightarrow (YN, 2)$ of the computation in C_0.

2-D Universal Reversible Cellular Automata

Simulating Reversible Logic Gates by 2-D RCAs

A set of logic elements is called *logically universal*, if any sequential machine (i. e., finite automaton with outputs) can be composed using only elements in the set. Since a finite-state control and tape cells of a Turing machine are in fact sequential machines, we can construct any Turing machine by using these elements. In this section, we show several computation-universal two-dimensional RCAs in which universal reversible logic elements are embedded.

A Fredkin gate [8] is a reversible (i. e., its logical function is one-to-one) and bit-conserving (i. e., the number of 1's is conserved between inputs and outputs) logic gate shown in Fig. 11. It has been known that any combinational logic element (especially, AND, OR, NOT, and fan-out elements) can be realized only with Fredkin gates [8]. Hence, we can construct any sequential machine with Fredkin gate and delay elements.

It is also known that a Fredkin gate can be composed of a much simpler gate called a switch gate and its inverse gate [8]. A switch gate is again a reversible and bit-conserving logic gate (Fig. 12). Furthermore, a switch gate

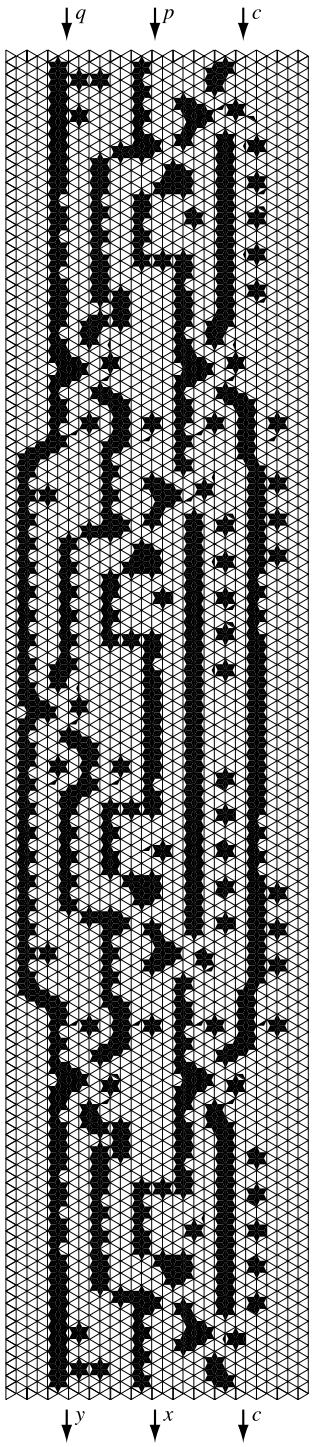

Reversible Cellular Automata, Figure 20
A Fredkin gate realized in the 2^3-state reversible triangular PCA T_1 [12]

is realized by the Billiard Ball Model (BBM) of computation [8]. The BBM is a kind of physical model of computation where a signal "1" is represented by an ideal ball, and logical operations and routing can be performed by their elastic collisions and reflections by reflectors. Figure 13 shows a BBM realization of a switch gate.

The 2-state RCA with Margolus neighborhood The BBM can be realized by the two-dimensional 2-state RCA proposed by Margolus [18], which has a block rules shown in Fig. 2. Figure 14 shows a reflection of a ball by a mirror in the Margolus' CA. Hence, Margolus' CA is computationally universal.

A 16-State Reversible PCA Model S_1 The model S_1 [30] is a 4-neighbor rotation- and reflection-symmetric reversible PCA model. A cell is divided into four parts, and each part has the state set $\{0, 1\}$. Its local transition rules are shown in Fig. 15. Rotated rules are omitted since it is rotation-symmetric (hence each rule can be regarded as a "rule scheme"). The states 0 and 1 are represented by a blank and a dot, respectively. The set of these rules has some similarity with that of Margolus' CA, and in fact, it can simulate the BBM in a similar manner. In S_1, a ball of the BBM is represented by two dots. Figure 16 gives a configuration of a switch gate.

A 16-State Reversible PCA Model S_2 The second model S_2 [30] is also a 4-neighbor reversible PCA having the set of transition rules shown in Fig. 17. It is rotation-symmetric but not reflection-symmetric. In S_2, the shape of a mirror and reflection by it are different from those of S_1, i. e., only a left-turn is possible. Hence a right-turn should be realized by three left-turns. The other features are similar to S_1. Figure 18 shows a configuration of a switch gate.

An 8-State Triangular Reversible PCA Model T_1 The model T_1 [12] is an 8-state reversible PCA on a triangular grid. It is rotation-symmetric but not reflection-symmetric. Its local function is extremely simple as shown in Fig. 19. Signal routing, crossing, and delay are very complex to realize, because a kind of "wall" is necessary to make a signal go straight ahead. Thus the size of the configuration of a Fredkin gate is very large (26×220) as in Fig. 20.

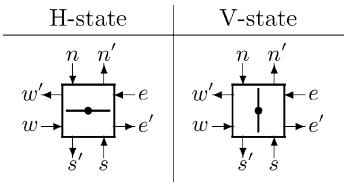

Reversible Cellular Automata, Figure 21
Two states of a rotary element (RE)

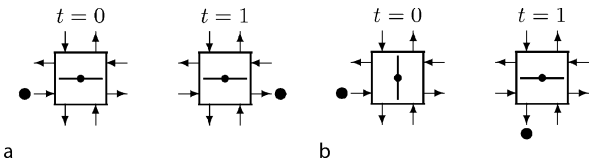

Reversible Cellular Automata, Figure 22
Operations of an RE: a the parallel case, and b the orthogonal case

Simulating Reversible Counter Machines by 2-D RCAs

Besides reversible logic gates like a Fredkin gate, there are also universal reversible logic elements with memory. A rotary element (RE) [26] is a typical one of such elements. An RE has four input lines $\{n, e, s, w\}$ and four output lines $\{n', e', s', w'\}$, and two states called H-state and V-state shown in Fig. 21 (hence it has a 1-bit memory). All the values of inputs and outputs are either 0 or 1. Here, the input (and the output) are restricted as follows: at most one "1" appears as an input (output) at a time. The operation of an RE is undefined for the cases that signal 1s are given to two or more input lines.

We employ the following intuitive interpretation for the operation of an RE. Signals 1 and 0 are interpreted as existence and non-existence of a particle. An RE has a "rotating bar" to control the moving direction of a particle. When no particle exists, nothing happens on the RE. If a particle comes from a direction parallel to the rotating bar, then it goes out from the output line of the opposite side (i. e., it goes straight ahead) without affecting the direction of the bar (Fig. 22a). If a particle comes from a direction orthogonal to the bar, then it makes a right

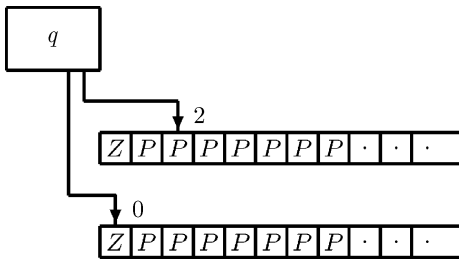

Reversible Cellular Automata, Figure 23
A counter machine with two counters

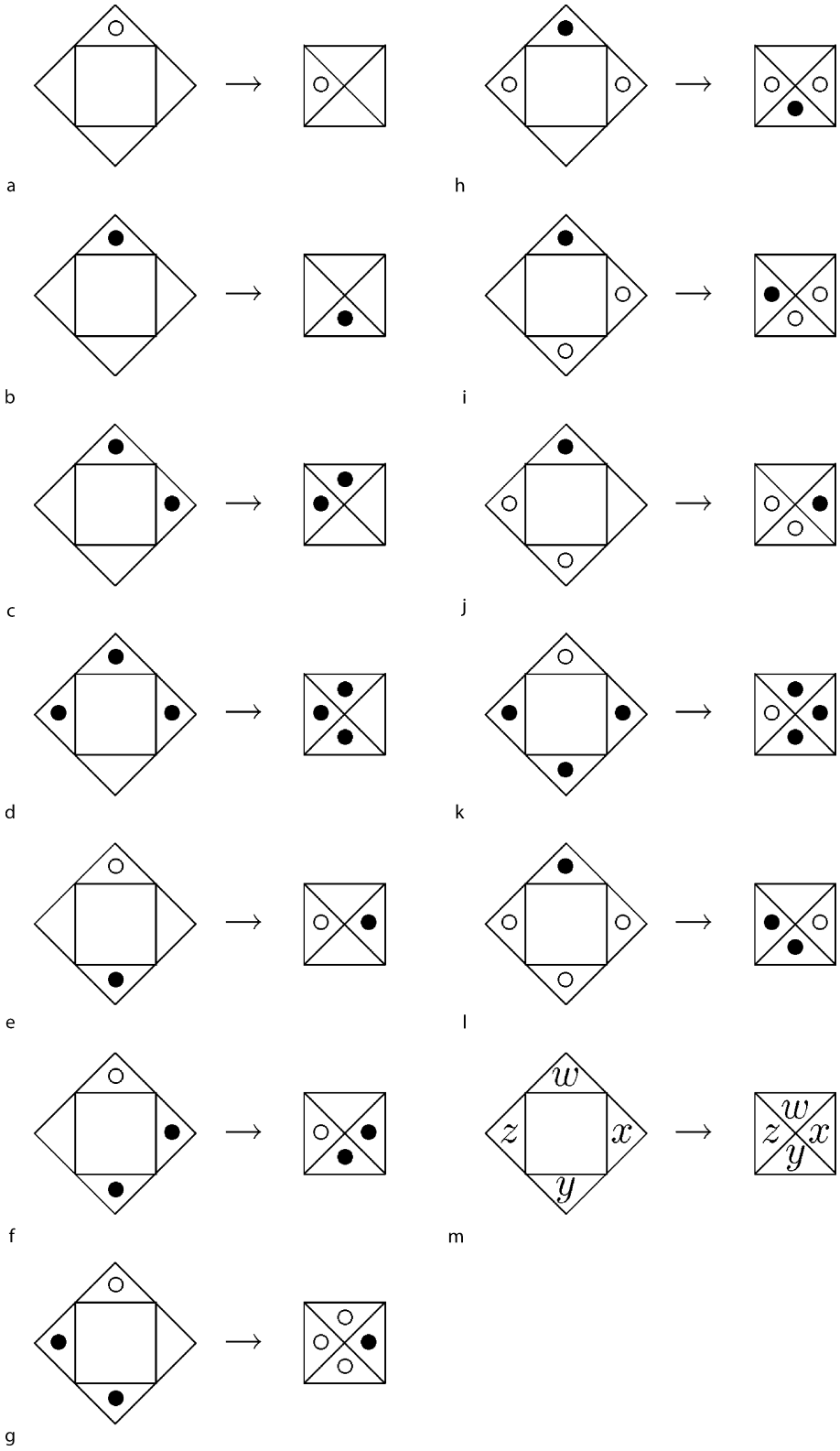

Reversible Cellular Automata, Figure 24

The local function of the 3^4-state rotation-symmetric reversible PCA P_3 [32]. The rule scheme (m) represents 33 rules not specified by (a)–(l), where $w, x, y, z \in \{blank, \circ, \bullet\}$

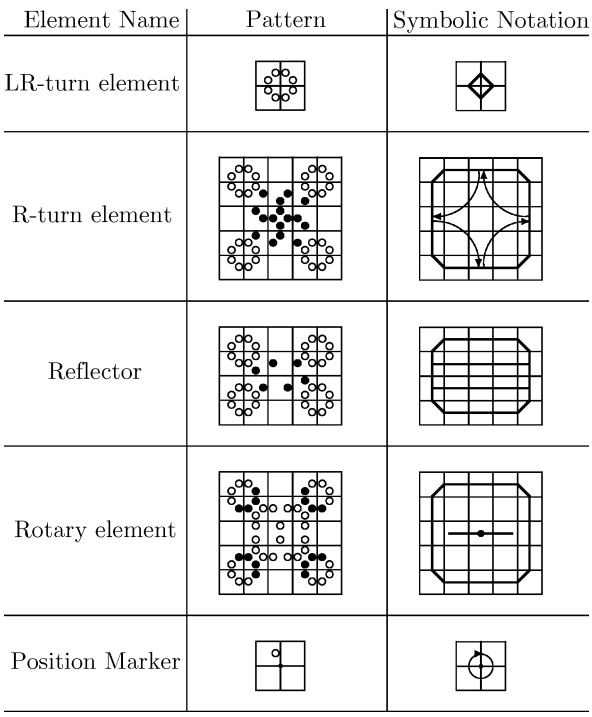

Reversible Cellular Automata, Figure 25
Basic elements realized in the reversible cellular space of P_3 [32]

turn, and rotates the bar by 90 degrees counterclockwise (Fig. 22b). It is clear its operation is reversible.

It has been shown that any reversible two-counter machine can be implemented in a quite simple way by using

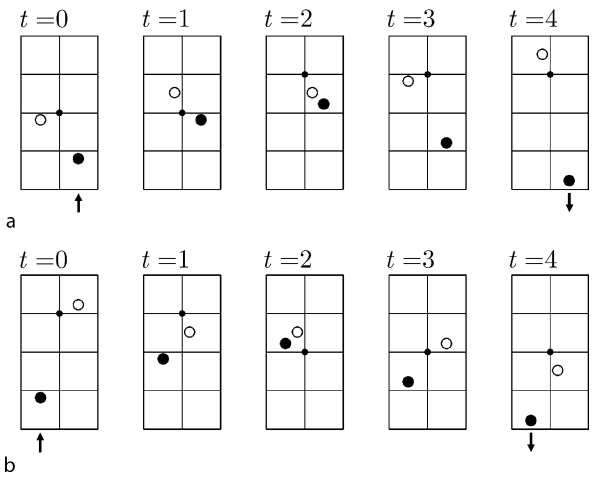

Reversible Cellular Automata, Figure 27
Pushing and pulling operations to a position marker in the reversible PCA P_3

REs and some additional elements [32]. Since a reversible two-counter machine is known to be universal [25], such a reversible PCA is also universal. A counter machine (CM) is a simple computation model consisting of a finite number of counters and a finite-state control [21]. In [25] a CM is defined as a kind of multi-tape Turing machine whose heads are read-only ones and whose tapes are all blank except the leftmost squares as shown in Fig. 23 (P is a blank symbol). This definition is convenient for giving the notion of reversibility on a CM.

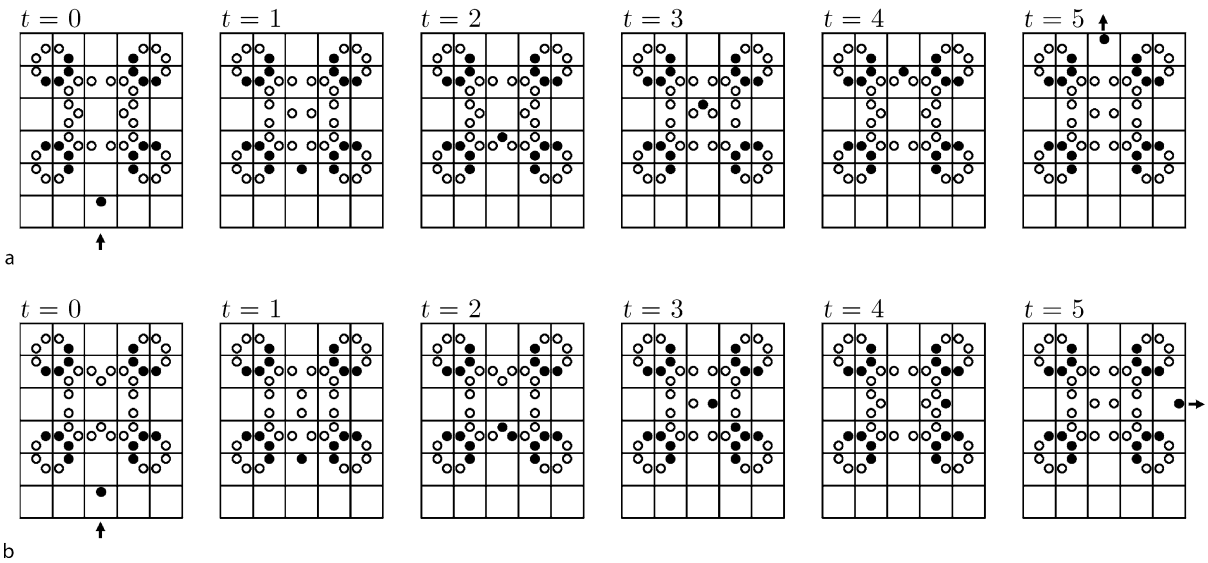

Reversible Cellular Automata, Figure 26
Operations of an RE in the reversible PCA P_3: a the parallel case, and b the orthogonal case

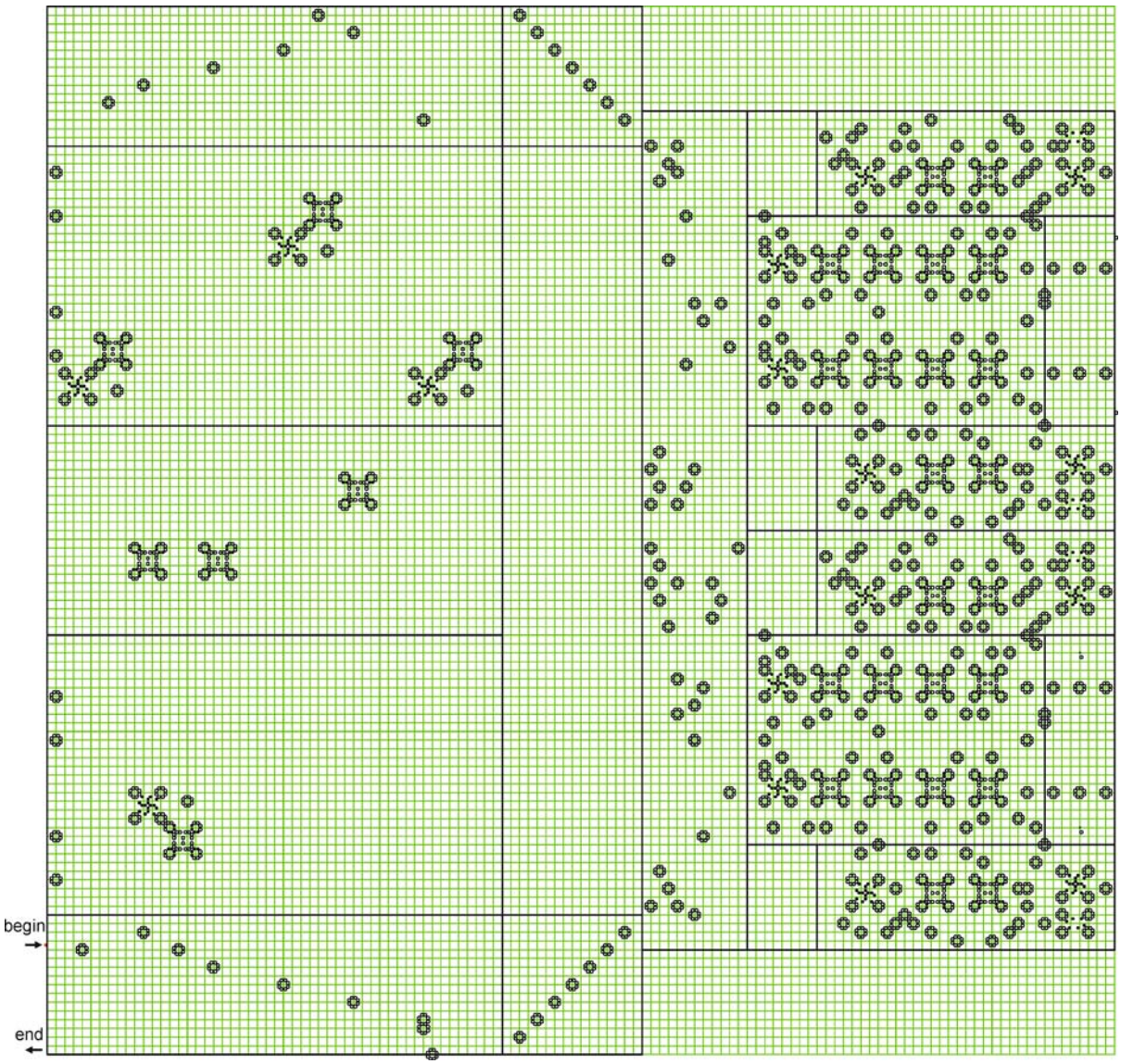

Reversible Cellular Automata, Figure 28
An example of a reversible counter machine, which computes the function $2x + 2$, embedded in the reversible PCA P_3 [32]

It is well known that a CM with two counters is computation-universal [21]. This result also holds even if the reversibility constraint is added.

Theorem 20 (Morita, [25]) *For any Turing machine T, there is a deterministic reversible CM with two counters M that simulates T.*

A 3^4-State Reversible PCA Model P_3 Any reversible CM with two counters is embeddable in the model P_3 with the local function shown in Fig. 24 [32]. In P_3, five kinds of signal processing elements shown in Fig. 25 can be realized. Here, a single ● acts as a signal. An LR-turn element, an R-turn element, and a reflector are used for signal routing. Figure 26 shows the operations of an RE in the P_3 space. A position marker is used to keep a head position of a CM, and realized by a single ○, which rotates clockwise at a certain position by the rule (a) in Fig. 24. Figure 27 shows the pushing and pulling operations of a position marker. Figure 28 shows an example of a whole configuration for a reversible CM with two counters embedded in the P_3 space. In this model, no conventional logic elements

like AND, OR and NOT is used. Computation is simply carried out by a single signal that interacts with REs and position markers.

Future Directions

In this section, we discuss future directions and open problems as well as topics not dealt with in the previous sections.

How Simple Can Universal RCAs Be?

We have seen that there are many kinds of simple RCAs having computation-universality. These RCAs with least number of states known so far are summed up as follows.

- One-dimensional case:
  Finite configuration: 98-state reversible PCA [27].
  Infinite configuration: 36-state reversible PCA [27].
- Two-dimensional case:
  Finite configuration: 81-state reversible PCA [32].
  Infinite configuration: 2-state RCA with Margolus neighborhood [18].
  8-state reversible PCA on triangular grid [12].
  16-state reversible PCA on rectangular grid [30].

We think the number of states for universal RCA can be reduced much more for each case of the above. Although the framework of PCA is useful for designing an RCA of a standard type, the number of states become relatively large because the state set is the direct product of the sets of the states of the parts. Hence we shall need some other technique to find a universal RCA with a small number of states.

How Can We Realize RCAs in Reversible Physical Systems?

This is a very difficult problem. At present there is no good solution. The Billiard Ball Model [8] is an interesting idea, but it is practically impossible to implement it perfectly. Instead of using a mechanical collision of balls, at least some quantum mechanical reversible phenomena should be used.

Furthermore, if we want to implement a CA in a real physical system, the following problem arises. In a CA, both time and space are discrete, and all the cells operate synchronously. On the other hand, in a real system, time and space are continuous, and no synchronizing clock is assumed beforehand. Hence, we need some novel theoretical framework for dealing with such problems.

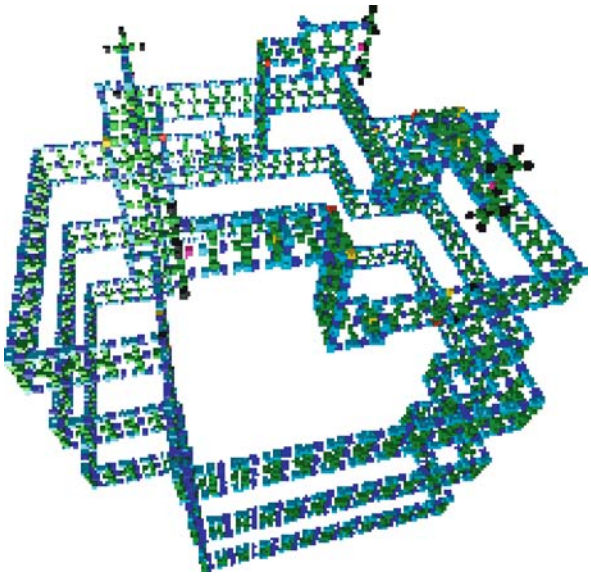

Reversible Cellular Automata, Figure 29
Self-reproduction of a pattern in a three-dimensional RCA [13]

Self-Reproduction in RCAs

Von Neumann first invented a self-reproducing cellular automata by using his famous 29-state CA [40]. In his model, the size of a self-reproducing pattern is quite huge, because the pattern has both computing and self-reproducing abilities. Later, Langton [17] created a very simple self-reproducing CA by removing the condition that the pattern need not have computation-universality.

It was shown that self-reproduction of the Langton's type is possible in two- or three-dimensional reversible PCA [13,29]. Figure 29 shows a self-reproducing pattern in a three-dimensional reversible PCA [13]. But, it is left for the future study to design a simple and elegant RCA in which objects with computation-universality can reproduce themselves.

Firing Squad Synchronization in RCAs

It is also possible to solve the firing squad synchronization problem using RCAs. Imai and Morita [11] gave a 99-state reversible PCA that synchronize an array of n cells in $3n$ time steps. Though it seems possible to give an optimal time solution, i. e., a $(2n-2)$-step solution, its concrete design has not yet done.

Bibliography

Primary Literature

1. Amoroso S, Cooper G (1970) The Garden of Eden theorem for finite configurations. Proc Amer Math Soc 26:158–164

2. Amoroso S, Patt YN (1972) Decision procedures for surjectivity and injectivity of parallel maps for tessellation structures. J Comput Syst Sci 6:448–464
3. Bennett CH (1973) Logical reversibility of computation. IBM J Res Dev 17:525–532
4. Bennett CH (1982) The thermodynamics of computation. Int J Theor Phys 21:905–940
5. Bennett CH, Landauer R (1985) The fundamental physical limits of computation. Sci Am 253:38–46
6. Boykett T (2004) Efficient exhaustive listings of reversible one dimensional cellular automata. Theor Comput Sci 325:215–247
7. Cook M (2004) Universality in elementary cellular automata. Complex Syst 15:1–40
8. Fredkin E, Toffoli T (1982) Conservative logic. Int J Theor Phys 21:219–253
9. Gruska J (1999) Quantum Computing. McGraw-Hill, London
10. Hedlund GA (1969) Endomorphisms and automorphisms of the shift dynamical system. Math Syst Theor 3:320–375
11. Imai K, Morita K (1996) Firing squad synchronization problem in reversible cellular automata. Theor Comput Sci 165:475–482
12. Imai K, Morita K (2000) A computation-universal two-dimensional 8-state triangular reversible cellular automaton. Theor Comput Sci 231:181–191
13. Imai K, Hori T, Morita K (2002) Self-reproduction in three-dimensional reversible cellular space. Artifici Life 8:155–174
14. Kari J (1994) Reversibility and surjectivity problems of cellular automata. J Comput Syst Sci 48:149–182
15. Kari J (1996) Representation of reversible cellular automata with block permutations. Math Syst Theor 29:47–61
16. Landauer R (1961) Irreversibility and heat generation in the computing process. IBM J Res Dev 5:183–191
17. Langton CG (1984) Self-reproduction in cellular automata. Physica 10D:135–144
18. Margolus N (1984) Physics-like model of computation. Physica 10D:81–95
19. Maruoka A, Kimura M (1976) Condition for injectivity of global maps for tessellation automata. Inf Control 32:158–162
20. Maruoka A, Kimura M (1979) Injectivity and surjectivity of parallel maps for cellular automata. J Comput Syst Sci 18:47–64
21. Minsky ML (1967) Computation: Finite and Infinite Machines. Prentice-Hall, Englewood Cliffs
22. Moore EF (1962) Machine models of self-reproduction, Proc Symposia in Applied Mathematics. Am Math Soc 14:17–33
23. Mora JCST, Vergara SVC, Martinez GJ, McIntosh HV (2005) Procedures for calculating reversible one-dimensional cellular automata. Physica D 202:134–141
24. Morita K (1995) Reversible simulation of one-dimensional irreversible cellular automata. Theor Comput Sci 148:157–163
25. Morita K (1996) Universality of a reversible two-counter machine. Theor Comput Sci 168:303–320
26. Morita K (2001) A simple reversible logic element and cellular automata for reversible computing. In: Proc 3rd Int Conf on Machines, Computations, and Universality. LNCS, vol 2055. Springer, Berlin, pp 102–113
27. Morita K (2007) Simple universal one-dimensional reversible cellular automata. J Cell Autom 2:159–165
28. Morita K, Harao M (1989) Computation universality of one-dimensional reversible (injective) cellular automata. Trans IEICE Japan E-72:758–762
29. Morita K, Imai K (1996) Self-reproduction in a reversible cellular space. Theor Comput Sci 168:337–366
30. Morita K, Ueno S (1992) Computation-universal models of two-dimensional 16-state reversible cellular automata. IEICE Trans Inf Syst E75-D:141–147
31. Morita K, Shirasaki A, Gono Y (1989) A 1-tape 2-symbol reversible Turing machine. Trans IEICE Japan E-72:223–228
32. Morita K, Tojima Y, Imai K, Ogiro T (2002) Universal computing in reversible and number-conserving two-dimensional cellular spaces. In: Adamatzky A (ed) Collision-based Computing. Springer, London, pp 161–199
33. Myhill J (1963) The converse of Moore's Garden-of-Eden theorem. Proc Am Math Soc 14:658–686
34. Richardson D (1972) Tessellations with local transformations. J Comput Syst Sci 6:373–388
35. Sutner K (2004) The complexity of reversible cellular automata. Theor Comput Sci 325:317–328
36. Toffoli T (1977) Computation and construction universality of reversible cellular automata. J Comput Syst Sci 15:213–231
37. Toffoli T (1980) Reversible computing, Automata, Languages and Programming. In: de Bakker JW, van Leeuwen J (eds) LNCS, vol 85. Springer, Berlin, pp 632–644
38. Toffoli T, Margolus N (1990) Invertible cellular automata: a review. Physica D 45:229–253
39. Toffoli T, Capobianco S, Mentrasti P (2004) How to turn a second-order cellular automaton into lattice gas: a new inversion scheme. Theor Comput Sci 325:329–344
40. von Neumann J (1966) Theory of Self-reproducing Automata. Burks AW (ed) University of Illinois Press, Urbana
41. Watrous J (1995) On one-dimensional quantum cellular automata. In: Proc 36th Symp on Foundation of Computer Science. IEEE, Los Alamitos, pp 528–537

Books and Reviews

Adamatzky A (ed) (2002) Collision-Based Computing. Springer, London

Bennett CH (1988) Notes on the history of reversible computation. IBM J Res Dev 32:16–23

Burks A (1970) Essays on Cellular Automata. University of Illinois Press, Urbana

Kari J (2005) Theory of cellular automata: A survey. Theor Comput Sci 334:3–33

Morita K (2001) Cellular automata and artificial life–computation and life in reversible cellular automata. In: Goles E, Martinez S (eds) Complex Systems. Kluwer, Dordrecht, pp 151–200

Wolfram S (2001) A New Kind of Science. Wolfram Media, Champaign

Reversible Computing

Kenichi Morita
Hiroshima University, Higashi-Hiroshima, Japan

Article Outline

Glossary
Definition of the Subject
Introduction
Reversible Logic Gates and Circuits

Reversible Logic Elements with Memory
Reversible Turing Machines
Other Models of Reversible Computing
Future Directions
Bibliography

Glossary

Billiard ball model The Billiard Ball Model (BBM) is a physical model of computation proposed by Fredkin and Toffoli [19]. It consists of idealized balls and reflectors. Balls can collide with other balls or reflectors. It is a reversible dynamical system, since it is assumed that collisions are elastic, and there is no friction. Fredkin and Toffoli showed that a reversible logic gate called Fredkin gate, which is known to be logically universal, can be embedded in BBM. Hence, a universal computer can be realized in the space of BBM.

Fredkin gate The Fredkin gate is a typical reversible logic gate with 3 inputs and 3 outputs whose operation is defined by the one-to-one mapping $(c, p, q) \mapsto (c, cp + \bar{c}q, \bar{c}p + cq)$. It is known that any logic function (even if it is not a one-to-one function) can be realized by using only Fredkin gates by allowing constant inputs and garbage outputs (i. e., useless signals). Furthermore, it is shown by Fredkin and Toffoli [19] that garbage signals can be reversibly erased, and hence any logic function is realized by a garbage-less circuit composed only of Fredkin gates.

Reversible cellular automaton A cellular automaton (CA) consists of a large number of finite automata called cells interconnected uniformly, and each cell changes its state depending on its neighboring cells. A reversible cellular automaton (RCA) is a one whose global function (i. e., a transition function from the configurations to the configurations) is one-to-one. RCAs can be thought as spatio-temporal models of reversible physical systems as well as reversible computing models. (See ► Reversible Cellular Automata)

Reversible logic element A reversible logic element is a primitive from which logic circuits can be composed, and whose operation is defined by a one-to-one mapping. There are two kinds of such elements: those without memory, and with memory. Reversible elements without memory are nothing but reversible logic gates. The Fredkin gate and the Toffoli gate are well-known examples of them, which are universal. Reversible elements with 1-bit memory are also useful when constructing reversible computing systems. The rotary element is a typical one of this type, which is also known to be universal.

Reversible Turing machine A reversible Turing machine (RTM) is a "backward deterministic" Turing machine, and is a standard model of a reversible computing system. Bennett [4] showed that for any (irreversible) Turing machine there is an RTM that simulates the former and leaves no garbage information on the tape when it halts. Hence, Turing machines still have computation-universality even if the constraint of reversibility is added. Furthermore, there is a rather small universal reversible Turing machine.

Rotary element A rotary element (RE) is a reversible logic element with 1-bit memory [36]. Conceptually, it is square-shaped, and has a "rotating bar" inside. The direction of the bar is either horizontal or vertical. It has four input ports and four output ports on the four edges. If a "particle" comes from the direction parallel to the bar, then it goes straight ahead and does not affect the direction of the bar. If it comes from the direction orthogonal to the bar, then it turns rightward and the bar rotates by 90 degrees. It is known that an RE is universal, and reversible Turing machines can be built very concisely by using only REs.

Definition of the Subject

A reversible computing system is defined as a "backward deterministic" system, where every computational configuration (state of the whole system) has exactly one previous configuration. Hence, a backward computation can be performed by its "inverse" computing system. Though its definition is rather simple, it is closely related to physical reversibility. The study of reversible computing originated from an investigation of heat generation (or energy dissipation) in computing systems. It deals with problems how computation can be carried out efficiently and elegantly in reversible computing models, and how such systems can be implemented in reversible physical systems. Since reversibility is one of the fundamental microscopic physical laws of Nature, and future computing systems will surely become in a nano-scale size, it is very important to investigate these problems.

In this article, we discuss several models of reversible computing from the viewpoint of computing theory (hence, we do not discuss problems of physical implementation in detail). We deal with several reversible computing models of different levels: from an element level to a system level. They are the Billiard Ball Model – an idealized reversible physical model, reversible logic elements and circuits, reversible Turing machines, and a few others. We see these models are related each other, i. e., reversible Turing machines are composed of reversible logic

elements, and reversible logic elements are implemented in the Billiard Ball Model. We can also see even very simple reversible systems have very rich computing ability in spite of the constraint of reversibility. Such computing models will give new insights for the future computing systems.

Introduction

The definitions of reversible computing systems are rather simple. They are the systems whose computation process can be traced in the reverse direction in a unique way from the end of the computation to the start of it. In other words, they are "backward deterministic" computing models (of course, a more precise definition should be given for each model of reversible computing). But, if we explain reversible computing systems only in this way, it is hard to understand why they are interesting. Their importance lies in the fact that it is closely related to physical reversibility.

In [24], Landauer argued the relation between logical irreversibility and physical irreversibility. He pointed out that any irreversible logical operation, such as a simple erasure of an information from the memory, or a simple merge of two paths of a program, is associated with physical irreversibility, and hence it necessarily causes heat generation. In particular, if 1 bit of information is erased, at least $kT \ln 2$ of energy will be dissipated, where k is the Boltzmann constant, and T is the absolute temperature. This is called "Landauer's principle". (Note that besides simple erasure and simple merge, there are reversible erasure and reversible merge, which are not irreversible operations, and are useful for constructing garbage-less reversible Turing machine as we shall see in Sect. "Universality of RTMs and Garbage-Less Computation"). After the work of Landauer, various studies related to "thermodynamics of computation" appeared [4,5,6,7,9,22]. There is also an argument on computation from the physical viewpoint by Feynman [16], which contains an idea leading to quantum computing (see e. g., [12,15,20]).

Bennett [4] studied reversible Turing machine from the viewpoint of Landauer's principle. He proved that for any irreversible Turing machine there is a reversible Turing machine that simulates the former and leaves no garbage information on the tape. The notion of garbage-less computation is important, because disposal of a garbage information is actually equivalent to erasure of the information, and hence leads to energy dissipation. The result of Bennett suggests that any computation can be performed efficiently in a physically reversible system.

Since then, various kinds of reversible computing models have been proposed and investigated: reversible logic elements and circuits [19,36,47,48], the Billiard Ball Model [19] – an idealized physical model of computation, reversible cellular automata [46,49], reversible counter machines [34], and so on. It has been known that all these models have *computation-universality*, i. e., any (irreversible) Turing machine can be simulated in these frameworks.

In this article, we discuss reversible computing from the theoretical point of view (though there are also studies on hardware implementation issues). We show how they can have computation-universality, and how different they are from conventional models. We can see that in some cases, computation in reversible systems is carried out in a very different fashion from that of traditional irreversible computing models.

An outline of this article is as follows. In Sect. "Reversible Logic Gates and Circuits", reversible logic gates and circuits are discussed. In particular, the Fredkin gate, its logical universality, garbage-less logic circuits, and its realization in the Billiard Ball Model are explained. In Sect. "Reversible Logic Elements with Memory", reversible logic elements with memory are discussed. A rotary element (RE) is introduced, and its logical universality is shown. In Sect. "Reversible Turing Machines", reversible Turing machines (RTMs) are discussed. A converting method to a garbage-less RTM, a simple realization of an RTM by REs, and a universal RTM are shown. In Sect. "Other Models of Reversible Computing", other reversible models of computing, which are not dealt with in the previous sections, are briefly discussed. In Sect. "Future Directions", future directions and open problems are given.

Reversible Logic Gates and Circuits

A reversible logic gate is a many-input many-output gate that realizes a one-to-one logical (i. e., Boolean) function. Early research on such gates is found in the study of Petri [42]. Later, Toffoli [47,48], and Fredkin and Toffoli [19] studied them from the standpoint of reversible computing, i. e., how they can be realized in a reversible physical system, and how they are related to power dissipation in computation. In this section, we discuss issues of logical universality of reversible logic gates, garbage-less logic circuits, and a relation to the Billiard Ball Model of computation, along the line of the reference [19].

Reversible Logic Gates

Definition 1 An *m-input n-output logic gate with a domain D* is defined as an element that realizes a function $D \to \{0,1\}^n$, where $D \subseteq \{0,1\}^m$. When $D = \{0,1\}^m$, we

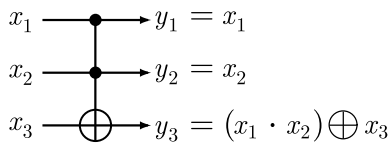

Reversible Computing, Figure 1
The generalized AND/NAND gate of order 3 called the Toffoli gate

simply call it an *m-input n-output* logic gate. It is called a *reversible logic gate*, if the function $D \to \{0,1\}^n$ is one-to-one.

It is easy to see that the traditional 1-input 1-output logic gate NOT is reversible, and the 2-input 1-output logic gates AND and OR are irreversible. An example of an m-input n-output reversible logic gate with a domain $D \neq \{0,1\}^m$ will appear in Sect. "Billiard Ball Model (BBM)".

Definition 2 (Toffoli [47,48]) The *generalized AND/NAND gate* of order n is a reversible logic gate that realizes the function

$$\theta^{(n)}\colon (x_1, \ldots, x_{n-1}, x_n) \mapsto (x_1, \ldots, x_{n-1}, (x_1 \ldots x_{n-1}) \oplus x_n)\,.$$

We can easily verify that $\theta^{(n)}$ is one-to-one, and in fact $(\theta^{(n)})^{-1} = \theta^{(n)}$. The gate that realizes $\theta^{(2)}$ is sometimes called the *controlled NOT (CNOT)*, which is an important gate in quantum computing (see e. g., [20]). In particular, it is shown that all unitary operations are expressed as compositions of CNOT gates and 1-bit quantum gates [3]. The gate for $\theta^{(3)}$ is called the *Toffoli gate* (Fig. 1).

Definition 3 (Fredkin and Toffoli [19]) The *Fredkin gate* is a 3-input 3-output reversible logic gate that realizes the function

$$\varphi\colon (c, p, q) \mapsto (c, c \cdot p + \bar{c} \cdot q, \bar{c} \cdot p + c \cdot q)\,.$$

It is also easy to verify that φ is one-to-one and $\varphi^{-1} = \varphi$. The operations of the Fredkin gate are depicted in Fig. 2.

If $c = 1$ then the inputs p and q are connected to x and y in parallel, while if $c = 0$ then the outputs are exchanged.

In addition to reversibility, the Fredkin gate has bit-conserving property: the total number of 1's in the output lines is the same as that of the input lines (note that the Toffoli gate is not so). This property is an analogue of the conservation law of mass, energy, or momentum in physics. Fredkin and Toffoli [19] proposed a design theory of reversible logic circuits composed of Fredkin gates and unit wires (which have unit-time delay) (Fig. 3). In their theory, a circuit composed of these elements must satisfy the following conditions: every output of a Fredkin gate can be connected to an input of a unit wire (not directly to another Fredkin gate), and every output of a Fredkin gate or a unit wire can be connected to at most one input of another element (i. e., fan-out is not allowed).

Definition 4 A set E of logic elements is called *logically universal*, if any logic function can be realized by a circuit composed only of the elements in E.

It is well known that, e. g., the set {AND, NOT} is logically universal. If we add a delay element (i. e., a memory) to a logically universal set, we can construct any sequential machine, and thus build a universal computer from them (as an infinite circuit).

We can see that AND, NOT, and fan-out can be realized by Fredkin gates [19] as shown in Fig. 4, if we allow to supply constant inputs 0's and 1's, and allow garbage (useless) outputs besides the true inputs and outputs. Hence, the set {Fredkin gate} is logically universal. It is also possible to show logical universality of the set {Toffoli gate} (this is left as an exercise for the reader).

Theorem 1 *The sets {Fredkin gate} and {Toffoli gate} are logically universal.*

On the other hand, Toffoli [47] showed any 2-input 2-output reversible gate is composed only of CNOTs and NOTs. Since it is easy to see that {CNOT, NOT} is not logically universal, we can conclude no 2-input 2-output reversible gate is logically universal.

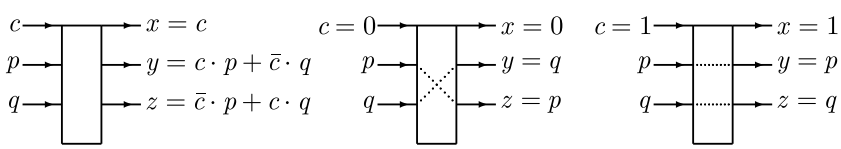

Reversible Computing, Figure 2
The Fredkin gate and its operations

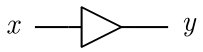

Reversible Computing, Figure 3
A unit wire, which acts as a delay element ($y(t) = x(t-1)$)

Garbage-Less Logic Circuits Composed of Fredkin Gates

As seen from Theorem 1, any logic function $f\colon \{0,1\}^m \to \{0,1\}^n$ (even if it is not one-to-one) can be realized by a circuit Φ composed of Fredkin gates and unit wires, by allowing constant inputs $\boldsymbol{c}$ and garbage outputs $\boldsymbol{g}$ (Fig. 5). Disposing the garbage $\boldsymbol{g}$ outside of the circuit is, very roughly speaking, analogous to heat generation in an actual circuit. Fredkin and Toffoli [19] showed a method to obtain a garbage-less circuit to compute the function f. This method is a logic circuit version of the garbage-less reversible Turing machine by Bennett [4], which will be discussed in Sect. "Universality of RTMs and Garbage–Less Computation".

First, the notion of an *inverse circuit* Φ^{-1} of a given reversible logic circuit Φ composed of Fredkin gates and unit wires is introduced. Φ^{-1} is obtained as follows: first take its mirror image, and then exchange inputs and outputs of each element. Figure 6 shows an example. We assume, Φ has no feedback loop (thus it realizes some combinatorial logic function f), and all the delays between inputs and outputs are the same. Then, Φ^{-1} computes the inverse function f^{-1}, because Φ^{-1} "undoes" the computation performed by Φ (note that the inverse of the Fredkin gate is itself). If we connect Φ and Φ^{-1} in series as in Fig. 7, garbage signals are erased reversibly, and we get again constants, which can be used in the next computation of the logic function. However, the true outputs $\boldsymbol{y}$ are also converted to the inputs $\boldsymbol{x}$. By inserting fan-out (i. e., copying) circuits between Φ and Φ^{-1}, we can obtain the outputs $\boldsymbol{y}$ without producing the garbage signals $\boldsymbol{g}$ (Fig. 8).

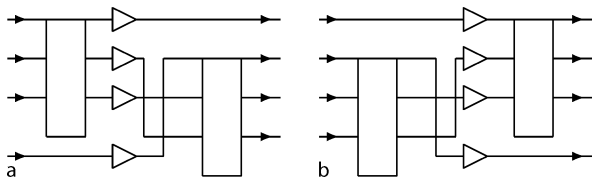

Reversible Computing, Figure 6
a A reversible logic circuit, and b its inverse circuit

Theorem 2 (Fredkin and Toffoli, [19]) *For any logic function $f\colon \boldsymbol{x} \mapsto \boldsymbol{y}$, there is a reversible logic circuit composed of Fredkin gates and unit wires that computes the mapping $f'\colon (\boldsymbol{c}, \boldsymbol{x}, \boldsymbol{0}, \boldsymbol{1}) \mapsto (\boldsymbol{c}, \boldsymbol{x}, \boldsymbol{y}, \bar{\boldsymbol{y}})$, and produces no garbage signal.*

Although the circuit in Fig. 8 does not give garbage signals $\boldsymbol{g}$, it produces $\boldsymbol{x}$ and $\bar{\boldsymbol{y}}$ besides the true output $\boldsymbol{y}$, which may be regarded as a kind of a garbage (though the amount

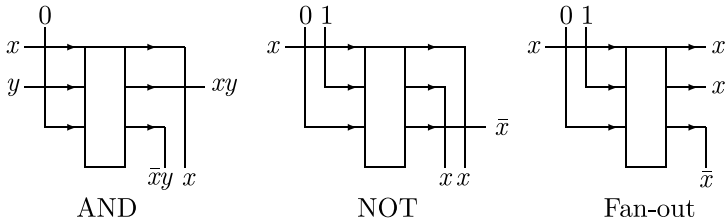

Reversible Computing, Figure 4
Implementing AND, NOT, and fan-out by Fredkin gates

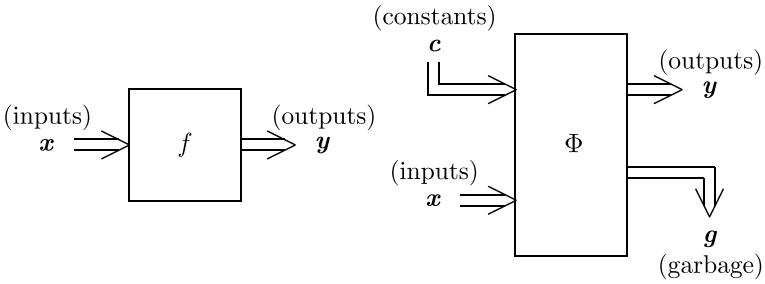

Reversible Computing, Figure 5
Embedding a logic function f in a reversible logic circuit Φ

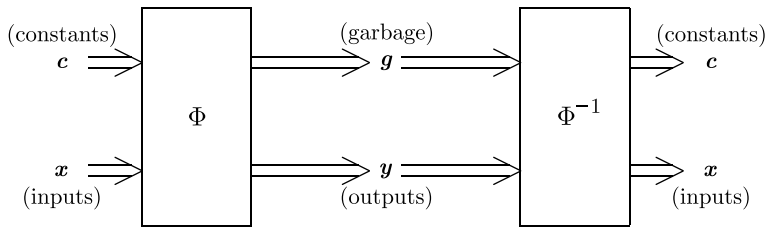

Reversible Computing, Figure 7
Reversibly erasing garbage signals by the inverse circuit Φ^{-1}

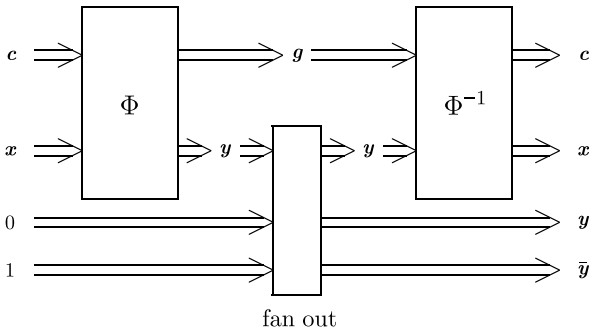

Reversible Computing, Figure 8
A garbage-less reversible logic circuits that computes a logic function

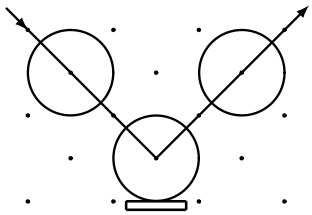

Reversible Computing, Figure 9
The Billiard Ball Model (BBM)

is generally smaller than that of $\boldsymbol{g}$). Especially, when we construct a circuit with memories (i. e., a sequential circuit), this kind of garbage is produced at every time step. In this sense, the above circuit is an "almost" garbage-less one. However, if we consider only "reversible" sequential machines (defined in Sect. "Definitions on Reversible Logic Elements with Memory"), then completely garbage-less circuits can be obtained. This is discussed in Sect. "Garbage-less Construction of RSMs by REs".

Billiard Ball Model (BBM)

Fredkin and Toffoli [19] proposed an interesting reversible physical model of computing called the Billiard Ball Model (BBM). It consists of idealized balls and reflectors (Fig. 9), and balls can collide with other balls or reflectors. We assume collisions are elastic, and there is no friction. Hence, there is no energy dissipation inside of this model.

Fredkin and Toffoli showed that any circuit composed of Fredkin gates and unit wires can be embedded in BBM. Therefore, a universal computer can be constructed in the space of BBM. To show the Fredkin gate is realizable in BBM, we need other reversible logic gates called the *switch gate* and its inverse. The switch gate is a 2-input 3-output gate shown in Fig. 10a, which realizes the one-to one logic function $(c, x) \mapsto (c, cx, \bar{c}x)$. The inverse switch gate is a 3-input 2-output gate whose logical function is defined only on the set $\{(0,0,0),(0,0,1),(1,0,0),(1,1,0)\}$ as shown in Fig. 10b, and is the inverse function of the switch gate. Figure 11 shows how the Fredkin gate can be built from switch gates and inverse switch gates (this circuit is due to R. Feynman and A. Ressler [19]). Finally, the switch gate is realized in BBM as in Fig. 12. Therefore, the Fredkin gate is also implemented in BBM.

Margolus showed that BBM can be realized in a reversible cellular automaton [28]. There are also studies related to BBM and cellular automata [14,35].

Reversible Logic Elements with Memory

In the classical design theory of logic circuits, generally two sorts of logic elements are supposed. They are logic gates (such as AND, OR, NOT, etc.) and a delay element (i. e., a memory like a flip-flop). Its design technique is also divided into two phases: implement Boolean functions as combinatorial logic circuits consisting of gates, and then construct sequential machines from combinatorial logic circuits and delay elements. In this way, the entire process of constructing digital circuits is systematized, and various methods of circuit minimization can be applied (especially for the first phase).

An approach of Sect. "Reversible Logic Gates and Circuits", which uses reversible logic gates and delay elements, is also a method along this line. However, logic el-

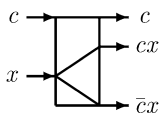

a

c	x	c	cx	$\bar{c}x$
0	0	0	0	0
0	1	0	0	1
1	0	1	0	0
1	1	1	1	0

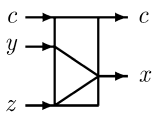

b

c	y	z	c	x
0	0	0	0	0
0	0	1	0	1
1	0	0	1	0
1	1	0	1	1

Reversible Computing, Figure 10
a The switch gate, and b the inverse switch gate

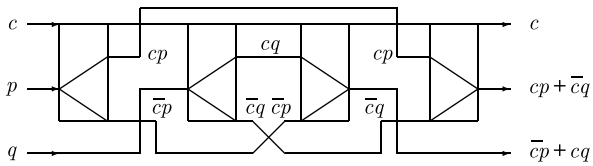

Reversible Computing, Figure 11
The Fredkin gate built from two switch gates and two inverse switch gates

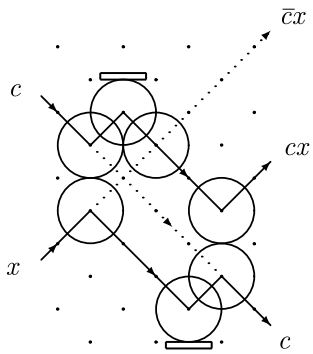

Reversible Computing, Figure 12
A realization of the switch gate in BBM

ements with memory are often more useful in reversible computing. A rotary element (RE) [36] is a typical example of such elements with 1-bit memory. It has 4 input lines and 4 output lines (exactly stated, it has 4 input symbols and 4 output symbols), and its operation is intuitively very simple. In this section, we discuss logical universality of an RE, and also show that any reversible sequential machine can be realized by REs as a completely garbage-less circuit.

In Sect. "Constructing RTMs by REs", we shall also see that by appropriately designing circuits composed only of REs, a clock signal can be eliminated. We should note that a clock is necessary as far as we use "gates", because two or more incoming signals must be synchronized to perform the gate operation. In particular, we shall see any reversible sequential machines and Turing machines can be realized very concisely by such circuits.

Definitions on Reversible Logic Elements with Memory

Since a reversible logic elements with memory is a kind of a reversible sequential machine (RSM), we first give definitions on a sequential machine of Mealy type (i. e., a finite automaton with an output), and its reversibility.

Definition 5 A *sequential machine* (SM) is defined by

$$M = (Q, \Sigma, \Gamma, q_0, \delta)\,,$$

where Q is a finite non-empty set of states, Σ and Γ are finite non-empty sets of input and output symbols, respectively, and $q_0 \in Q$ is an initial state. $\delta\colon Q \times \Sigma \to Q \times \Gamma$ is a mapping called a *move function*. A variation of an SM $M = (Q, \Sigma, \Gamma, \delta)$, where no initial state is specified, is also called an SM for convenience.

M is called a *reversible sequential machine* (RSM) if δ is one-to-one (hence $|\Sigma| \le |\Gamma|$). An RSM is "reversible" in the sense that, from the present state and the output of M, the previous state and the input are determined uniquely.

A reversible logic elements with memory (RLEM) is nothing but an RSM with small numbers of states and input/output symbols, from which reversible computers can be built. However, since there are infinitely many RSMs, we should restrict candidates of RLEMs somehow. Here, we consider only RLEMs with 2 states (i. e., $|Q| = 2$) and with k input/output symbols (i. e., $|\Sigma| = |\Gamma| = k$) for $k = 2, 3, 4$.

We consider, e. g., a 2-state 4-symbol RLEM $M = (Q, \Sigma, \Gamma, \delta)$. Here, we fix the state set as $Q = \{q_0, q_1\}$, and the input and output alphabets as $\Sigma = \{a, b, c, d\}$ and $\Gamma = \{w, x, y, z\}$, respectively. The move function δ is as follows.

$$\delta\colon \{q_0, q_1\} \times \{a, b, c, d\} \to \{q_0, q_1\} \times \{w, x, y, z\}$$

Since δ is one-to-one, it is specified by a permutation from the set

$$\{(q_0, w), (q_0, x), (q_0, y), (q_0, z), \\ (q_1, w), (q_1, x), (q_1, y), (q_1, z)\}\,.$$

Reversible Computing, Table 1
An example of a pair of equivalent 2-state 3-symbol RLEMs

Present state	Input		
	a	b	c
q_0	q_0x	q_1x	q_1y
q_1	q_0y	q_1z	q_0z
(a) RLEM No. 3-61			

Present state	Input		
	a'	b'	c'
q_0'	$q_0'y'$	$q_1'z'$	$q_1'y'$
q_1'	$q_0'x'$	$q_1'x'$	$q_0'z'$
(b) RLEM No. 3-235			

Hence, there are $8! = 40320$ RLEMs for $k = 4$. They are numbered by $0, \ldots, 40\,319$ in the lexicographic order of permutations. 2-state 2-symbol and 3-symbol RLEMs are also numbered in this way [38]. To indicate k-symbol RLEM, the prefix "k-" is attached to the serial number.

Example 1 Table 1 shows two RLEMs No. 3-61 and No. 3-235 specified by the following permutations (in Table 1(b), each symbol has a prime ($'$) for the convenience of the later argument).

$$\text{No. } 3-61:$$
$$((q_0, x), (q_1, x), (q_1, y), (q_0, y), (q_1, z), (q_0, z))$$
$$\text{No. } 3-235:$$
$$((q_0, y), (q_1, z), (q_1, y), (q_0, x), (q_1, x), (q_0, z))$$

There are many kinds of 2-state RLEMs even if we limit $k = 2, 3, 4$, but we can regard two RLEMs are "equivalent" if one can be obtained by renaming the states and the input/output symbols of the other. This is formalized as follows.

Definition 6 Let $M_1 = (Q_1, \Sigma_1, \Gamma_1, \delta_1)$ and $M_2 = (Q_2, \Sigma_2, \Gamma_2, \delta_2)$ be two LEMs. M_1 and M_2 are called *equivalent* (denoted by $M_1 \sim M_2$), if there exist one-to-one onto mappings $f\colon Q_1 \to Q_2$, $g : \Sigma_1 \to \Sigma_2$, and $h\colon \Gamma_1 \to \Gamma_2$ that satisfy

$$\forall\, q \in Q_1,\ \forall\, s \in \Sigma_1\ [\delta_1(q, s) = \psi(\delta_2(f(q), g(s)))]\ ,$$

where $\psi\colon Q_2 \times \Gamma_2 \to Q_1 \times \Gamma_1$ is defined as follows.

$$\forall\, q \in Q_2,\ \forall\, t \in \Gamma_2\ [\psi(q, t) = (f^{-1}(q), h^{-1}(t))]\ .$$

We can see the two RLEMs No. 3-61 and No. 3-235 in Example 1 are equivalent under the following one-to-one onto mappings.

$$f(q_0) = q_1', \quad f(q_1) = q_0'$$
$$g(a) = b', \quad g(b) = a', \quad g(c) = c'$$
$$h(x) = x', \quad h(y) = z', \quad h(z) = y'$$

Reversible Computing, Table 2
The move function δ_{RE} of a rotary element RE

Present state	Input			
	n	e	s	w
H-state:	w'	w'	e'	e'
V-state:	s'	n'	n'	s'

The total numbers of 2-state 2-, 3-, and 4-symbol RLEMs are $4! = 24$, $6! = 720$, and $8! = 40\,320$, respectively. But, the numbers of essentially different RLEMs are relatively few. It has been shown that the numbers of equivalence classes of 2-state 2-, 3-, and 4-symbol RLEMs are 8, 24 and 82, respectively [38].

A Rotary Element (RE) and Its Logical Universality

A *rotary element* (RE) [36] is a 2-state 4-symbol RLEM defined as follows.

$$\mathrm{RE} = (\{\boxed{\bullet\!-}, \boxed{\bullet\!|}\}, \{n, e, s, w\}, \{n', e', s', w'\}, \delta_{\mathrm{RE}})$$

The move function δ_{RE} is shown in Table 2.

An RE can be understood by the following intuitive interpretation. It has two states called H-state ($\boxed{\bullet\!-}$) and V-state ($\boxed{\bullet\!|}$), and four input lines $\{n, e, s, w\}$ and four output lines $\{n', e', s', w'\}$ corresponding to the input and output alphabets (Fig. 13). All the values of input and output lines are either 0 or 1, i. e., $(n, e, s, w), (n', e', s', w') \in \{0, 1\}^4$. However, we restrict the domain of their values as $\{(1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0), (0, 0, 0, 1)\}$, i. e., exactly one "1" appears among the four input (output, respectively) lines at a time, when an input is given (an output is produced). The operation of an RE is left undefined for the cases that signal 1's are given to two or more input lines. Signals 1 and 0 are interpreted as existence and non-existence of a particle. We can interpret that an RE has a "rotating bar" to control the moving direction of a particle. When no particle exists, nothing happens on the RE. If a particle comes from a direction parallel to the rotating bar, then it goes out from the output line of the opposite side (i. e., it goes straight ahead) without affecting the direction of the bar (Fig. 14a). If a particle comes from a direction orthogonal to the bar, then it makes a right turn, and rotates the bar by 90 degrees counterclockwise (Fig. 14b).

Theorem 3 (Morita [36]) *The set {RE} is logically universal.*

This theorem is proved by giving a circuit composed only of REs that simulates a Fredkin gate. Figure 15 shows one such circuit [38]. (A circuit given in [36] uses 16 REs to

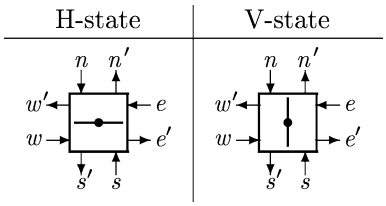

Reversible Computing, Figure 13
Two states of a rotary element (RE)

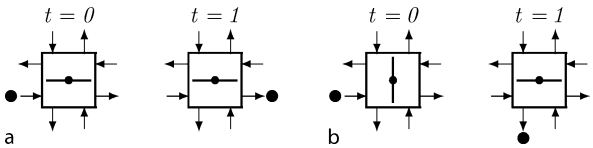

Reversible Computing, Figure 14
Operations of an RE: a the parallel case, and b the orthogonal case

simulate a Fredkin gate.) In Fig. 15, small triangles represent delay elements, where the number written inside of them indicates its delay time. In this circuit, we assume all the elements works synchronously. Note that an RE itself can operate as a delay element.

Garbage-Less Construction of RSMs by REs

We can show any RSM can be realized as a completely garbage-less RE circuit. (We can also construct a circuit of similar property from Fredkin gates [33].)

Theorem 4 (Morita [37]) *For any reversible sequential machine M, there is a garbage-less circuit C that realizes M, and is composed only of REs. C is completely garbage-less in the sense that it has no extra input/output lines other than true input/output lines (hence it is a "closed" circuit except the true input/output).*

Reversible Computing, Table 3
The move function of the RE-column

State x	Input l_i	Input r_i
(marked)	l'_i	l'_i
(unmarked)	r'_i	r'_i

Reversible Computing, Table 4
A move function δ_1 of an RSM M_1 (an example)

State	Input a_1	Input a_2
q_1	q_2b_1	q_3b_2
q_2	q_2b_2	q_1b_1
q_3	q_1b_2	q_3b_1

We show a construction method of a circuit C by an example. First, we consider a circuit composed of $k+1$ REs shown in Fig. 16. It is called an *RE-column*. In a resting state, each RE in the RE-column is in the vertical state except the bottom one indicated by x. If x is in the horizontal state, the RE-column is called in the *marked* state, otherwise *unmarked* state. It can be regarded as if it is a 2-state RSM with $2k$ input symbols $\{l_1, \ldots, l_k, r_1, \ldots, r_k\}$ and $2k$ output symbols $\{l'_1, \ldots, l'_k, r'_1, \ldots, r'_k\}$, though it has many transient states. Its move function is shown in Table 3.

We now consider an example of an RSM $M_1 = (\{q_1, q_2, q_3\}, \{a_1, a_2\}, \{b_1, b_2\}, \delta_1)$, where its move function δ_1 is shown in Table 4. Prepare three RE-columns with $k = 2$, each of which corresponds to each state of M_1, and connect them as shown in Fig. 17. If M_1 is in

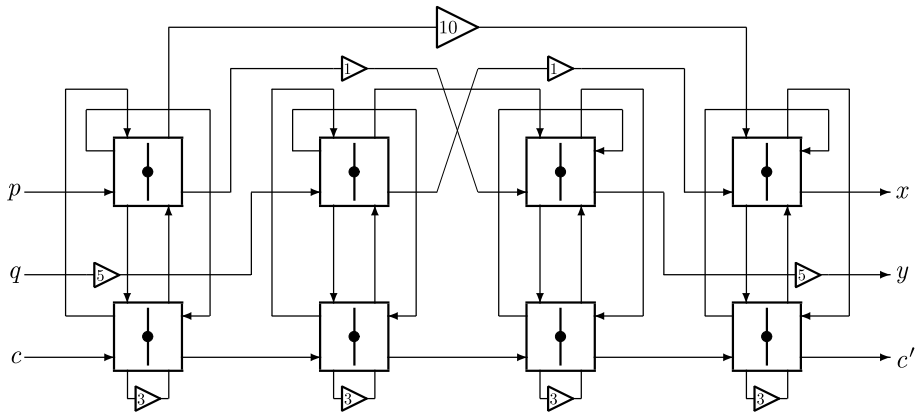

Reversible Computing, Figure 15
A realization of a Fredkin gate out of 8 REs and delay elements

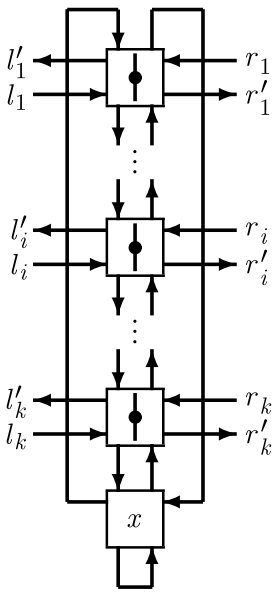

Reversible Computing, Figure 16
An RE-column

the state q_1, then only the first RE-column is set to the marked state. When an input signal comes from, e. g., the input line a_2, then the signal first goes out from the line l'_2 of the first RE-column, setting this column to the unmarked state. Then, this signal enters the third column from the line r_2. This makes the third RE-column marked, and finally goes out from the output line b_2, which realizes $\delta_1(q_1, a_2) = (q_3, b_2)$.

In Fig. 17, the circuit is designed so that the delay between the inputs and outputs is constant. However, if there is no need to do so, delay elements can be removed. In such a case, even if each RE operates asynchronously, the circuit works correctly.

Reversible Turing Machines

A reversible Turing machine (RTM) is a standard model of reversible computing as in the case of traditional (i. e., irreversible) computation theory. The notion of an RTM first appeared in the paper of Lecerf [25], where unsolvability of the halting problem and some related problems on RTMs are shown. Bennett [4] showed that every irreversible Turing machine (TM) can be simulated by an RTM with three tapes without leaving garbage information on the tapes when it halts. He also pointed out the significance of RTMs, since they are closely related to physical reversibility and the problem of energy dissipation in computing process.

In this section, after giving definitions on RTMs, we explain the method of Bennett for converting a given irreversible TM into an equivalent 3-tape RTM. We then discuss the problems how RTMs can be built by reversible logic elements, and how a small universal RTM can be constructed.

Definitions on Reversible Turing Machines (RTMs)

We first introduce a quadruple formulation of RTMs according to Bennett [4]. This formulation makes it easy to define an "inverse" Turing machine for a given RTM. The inverse RTM, which "undoes" the computation performed by the original RTM, plays a key role for garbage-less computation. We also give a quintuple formulation of RTMs

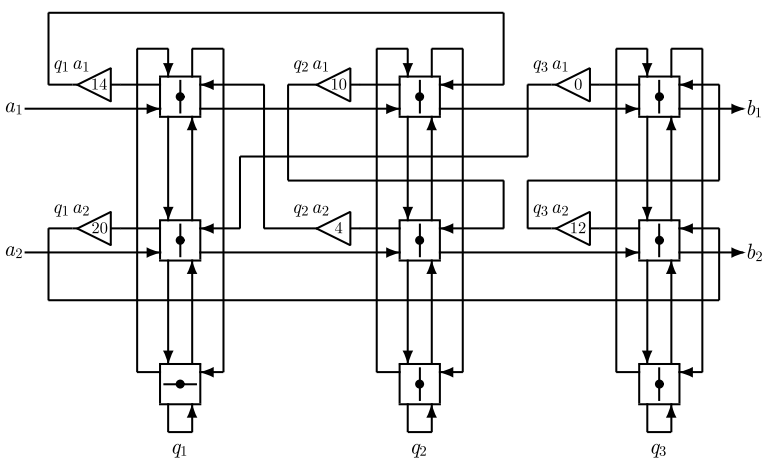

Reversible Computing, Figure 17
A garbage-less circuit made of REs and delay elements that simulates M_1

compatible with the quadruple formulation. This is because most classical TMs are defined in quadruple form, and thus it makes convenient to compare with them.

Definition 7 A *1-tape Turing machine in the quadruple form* is defined by

$$T_4 = (Q, S, q_0, q_f, s_0, \delta)\,,$$

where Q is a non-empty finite set of states, S is a non-empty finite set of symbols, q_0 is an initial state ($q_0 \in Q$), q_f is a final (halting) state ($q_f \in Q$), s_0 is a special blank symbol ($s_0 \in S$). δ is a move relation, which is a subset of $(Q \times S \times S \times Q) \cup (Q \times \{/\} \times \{-, 0, +\} \times Q)$. Each element of δ is a quadruple, and either of the form $[p, s, s', q] \in (Q \times S \times S \times Q)$ or $[p, /, d, q] \in (Q \times \{/\} \times \{-, 0, +\} \times Q)$. The symbols "$-$", "0", and "$+$" stand for "left-shift", "zero-shift", and "right-shift", respectively. $[p, s, s', q]$ means if T_4 reads the symbol s in the state p, then write s' and go to the state q. $[p, /, d, q]$ means if T_4 is in p, then shift the head to the direction d and go to the state q.

T_4 is called a *deterministic Turing machine* iff the following condition holds for any pair of distinct quadruples $[p_1, b_1, c_1, q_1]$ and $[p_2, b_2, c_2, q_2]$ in δ.

$$\text{If } p_1 = p_2\,, \text{ then } b_1 \neq / \wedge b_2 \neq / \wedge b_1 \neq b_2\,.$$

Note that, in what follows, we consider only deterministic Turing machines. So, we omit the word "deterministic" henceforth.

On the other hand, T_4 is called a *reversible Turing machine* (RTM) iff the following condition holds for any pair of distinct quadruples $[p_1, b_1, c_1, q_1]$ and $[p_2, b_2, c_2, q_2]$ in δ.

$$\text{If } q_1 = q_2\,, \text{ then } b_1 \neq / \wedge b_2 \neq / \wedge c_1 \neq c_2\,.$$

It is easy to extend the above definition to multi-tape RTMs. For example, a 2-tape TM is defined by

$$T = (Q, (S_1, S_2), q_0, q_f, (s_{1,0}, s_{2,0}), \delta)\,.$$

A quadruple in δ is either of the form $[p, [s_1, s_2], [s'_1, s'_2], q] \in (Q \times (S_1 \times S_2) \times (S_1 \times S_2) \times Q)$ or of the form $[p, /, [d_1, d_2], q] \in (Q \times \{/\} \times \{-, 0, +\}^2 \times Q)$. Determinism and reversibility are defined similarly as above, namely, T is reversible iff the following condition holds for any pair of distinct quadruples $[p_1, X_1, [b_1, b_2], q_1]$ and $[p_2, X_2, [c_1, c_2], q_2]$ in δ.

$$\text{If } q_1 = q_2\,, \text{ then } X_1 \neq / \wedge X_2 \neq / \wedge [b_1, b_2] \neq [c_1, c_2]\,.$$

Definition 8 A *1-tape Turing machine in the quintuple form* is defined by

$$T_5 = (Q, S, q_0, q_f, s_0, \delta)\,,$$

where Q, S, q_0, q_f, s_0 are the same as in Definition 7. δ is a move relation, which is a subset of $(Q \times S \times S \times \{-, 0, +\} \times Q)$. Each element of δ is a quintuple of the form $[p, s, s', d, q]$. It means if T_5 reads the symbol s in the state p, then write s', shift the head to the direction d, and go to the state q.

We say T_5 is *deterministic* iff the following condition holds for any pair of distinct quintuples $[p_1, s_1, s'_1, d_1, q_1]$ and $[p_2, s_2, s'_2, d_2, q_2]$ in δ.

$$\text{If } p_1 = p_2\,, \text{ then } s_1 \neq s_2\,.$$

We say T_5 is *reversible* iff the following condition holds for any pair of distinct quintuples $[p_1, s_1, s'_1, d_1, q_1]$ and $[p_2, s_2, s'_2, d_2, q_2]$ in δ.

$$\text{If } q_1 = q_2\,, \text{ then } s'_1 \neq s'_2 \wedge d_1 = d_2\,.$$

Proposition 1 *For any RTM T_5 in the quintuple form, there is an RTM T_4 in the quadruple form that simulates each step of the former in two steps.*

Proof Let $T_5 = (Q, S, q_0, q_f, s_0, \delta)$. We define $T_4 = (Q', S, q_0, q_f, s_0, \delta')$ as follows. Let $Q' = Q \cup \{q' \mid q \in Q\}$. The set δ' is given by the next procedure.

First, set the initial value of δ' to the empty set. Next, for each $q \in Q$ do the following operation. Let $[p_1, s_1, s'_1, d_1, q], [p_1, s_2, s'_2, d_2, q], \ldots, [p_m, s_m, s'_m, d_m, q]$ be all the quintuples in δ whose fifth element is q. Note that $d_1 = d_2 = \ldots = d_m$ holds, and $s'_1, s'_2, \ldots, s'_m$ are pair-wise distinct, because T_5 is reversible. Then, include the $m + 1$ quadruples $[p_1, s_1, s'_1, q'], [p_1, s_2, s'_2, q'], \ldots, [p_m, s_m, s'_m, q']$, and $[q', /, d_1, q]$ in δ'.

It is easy to see that T_4 has the required property. □

By Proposition 1, we see the definition of an RTM in the quintuple form is compatible with that in the quadruple form.

The converse of Proposition 1 is also easy to show. It is left for readers as an exercise. (Of course, it is very easy to construct an RTM in the quintuple form so that it simulates each quadruple of a given RTM by a single quintuple. In addition, it is also possible to simulate a consecutive pair of read/write and shift quadruples by one quintuple, and thus we can reduce the numbers of states and quintuples.)

Universality of RTMs and Garbage-Less Computation

We now discuss computation-universality of RTMs. It is actually easy to convert an irreversible TM to an RTM, because we can construct an RTM that simulates the former and records all the information which quadruples were executed by the irreversible TM by using a special history

tape. But, by this method a large amount of garbage information will be left at the end of a computation. An important point is that such a garbage can be reversibly erased, hence a garbage-less computation is possible as shown in the next theorem.

Theorem 5 (Bennett [4]) *For any (irreversible) 1-tape TM, we can construct a 3-tape RTM that simulates the former and leaves only an input string and an output string on the tapes when it halts (hence, it leaves no garbage information).*

Proof Let $T = (Q, S, q_0, q_f, 0, \delta)$ be a given irreversible TM in the quadruple form. We assume T satisfies the following conditions (it is easy to modify a given TM so that it meets the conditions).

(i) T has a rightward-infinite tape whose leftmost square always keeps the blank symbol.
(ii) When T halts, it does so at the leftmost symbol in the state q_f.
(iii) The output is given from the second square of the tape to the right when T halts.
(iv) The output string does not contain blank symbols.
(v) The initial state q_0 does not appear as the fourth element of any quadruple.
(vi) There is only one quadruple in δ whose fourth element is q_f.

Let m be the total number of quadruples contained in δ, and we assume the numbers $1, \ldots, m$ are assigned uniquely to these quadruples.

An RTM T^R that simulates T has three tapes: a working tape (for simulating the tape of T), a history tape (for recording the movement of T at each step), and an output tape (for writing an output string). It is defined as follows:

$$T^R = (Q', (S, \{0, 1, \ldots, m\}, S), q_0, p_0, (0, 0, 0), \delta')$$
$$Q' = \{q_0, q_1, \ldots, q_f\} \cup \{q'_0, q'_1, \ldots, q'_f\} \cup \{c_1, c'_1, c_2, c'_2\} \cup \{p_0, p_1, \ldots, p_f\} \cup \{p'_0, p'_1, \ldots, p'_f\}$$

Computation of T^R has three stages: (1) the forward computation stage, (2) the copying stage, and (3) the backward computation stage. The set δ' of quintuples is defined as the union $\delta_1 \cup \delta_2 \cup \delta_3$ of the sets of quintuples given below. We assume when T^R starts to move, an input string is written in the working tape, and the history and the output tapes contain only blank symbols.

(1) The quadruple set δ_1 for the forward computation stage: When T^R simulates T step by step, it records on the history tape which quadruple of T was used. By this operation T^R can be reversible. This is realized by giving the quadruple set δ_1 as follows.
(i) If the hth quadruple ($h = 1, \ldots, m$) of T is

$$[q_i, s_j, s_k, q_l] \quad (q_i, q_l \in Q, s_j, s_k \in S),$$

then include the following quadruples in δ_1.

$$\begin{array}{llll} [q_i, & /, & [0, +, 0], & q'_i] \\ [q'_i, & [s_j, 0, 0], & [s_k, h, 0], & q_l] \end{array}$$

(ii) If the hth quadruple ($h = 1, \ldots, m$) of T is

$$[q_i, /, d, q_l] \quad (q_i, q_l \in Q, d \in \{-, 0, +\})$$

then include the following quadruples in δ_1.

$$\begin{array}{llll} [q_i, & /, & [d, +, 0], & q'_i] \\ [q'_i, & [x, 0, 0], & [x, h, 0], & q_l] \quad (x \in S) \end{array}$$

(2) The quadruple set δ_2 for the copying stage: If the forward computation stage is completed, and T^R becomes in the state q_f, then copies the output string on the working tape to the output tape. This is realized by giving the quadruple set δ_2 as below. Here let n be the number of the quadruple that contains q_f as the fourth element.

$$\begin{array}{llll} [q_f, & [0, n, 0], & [0, n, 0], & c_1] \\ [c_1, & /, & [+, 0, +], & c_2] \\ [c_2, & [y, n, 0], & [y, n, y], & c_1] \quad (y \in S - \{0\}) \\ [c_2, & [0, n, 0], & [0, n, 0], & c'_2] \\ [c'_1, & [y, n, y], & [y, n, y], & c'_2] \quad (y \in S - \{0\}) \\ [c'_2, & /, & [-, 0, -], & c'_1] \\ [c'_1, & [0, n, 0], & [0, n, 0], & p_f] \end{array}$$

(3) The quadruple set δ_3 for the backward computation stage: After the copying process, the backward computation stage starts in order to reversibly erase garbage information left on the history tape. This is performed by an inverse TM of the forward computing. The quadruple set δ_3 is as follows.
(i) If the hth quadruple ($h = 1, \ldots, m$) of T is

$$[q_i, s_j, s_k, q_l] \quad (q_i, q_l \in Q, s_j, s_k \in S)$$

then include the following quadruples in δ_3.

$$\begin{array}{llll} [p'_i, & /, & [0, -, 0], & p_i] \\ [p_l, & [s_k, h, 0], & [s_j, 0, 0], & p'_i] \end{array}$$

(ii) If the hth quadruple ($h = 1, \ldots, m$) of T is

$$[q_i, /, d, q_l] \quad (q_i, q_l \in Q, d \in \{-, 0, +\})$$

then include the following quadruples in δ_3. Note that if $d = -$ (0, +, respectively), then $d' = +$ (0, −).

$$\begin{array}{llll} [p'_i, & /, & [d', -, 0], & p_i] \\ [p_l, & [x, h, 0], & [x, 0, 0], & p'_i] \quad (x \in S) \end{array}$$

It is easy to see that T^R simulates T correctly. Reversibility of T^R is verified by checking the quadruple set δ'. In particular, the set δ_3 of quadruples is reversible since T is deterministic. □

Example 2 We consider the following irreversible TM T_{erase}, and convert it into an RTM T^R_{erase} by the method of Theorem 5.

$$T_{\text{erase}} = (\{q_0, q_1, q_2, q_3, q_4, q_f\}, \{0, 1, 2\}, q_0, q_f, 0, \delta_{\text{erase}})$$

$$\begin{aligned} \delta_{\text{erase}} = \{ & 1 : [q_0, /, +, q_1], \quad 2 : [q_1, 0, 0, q_4], \\ & 3 : [q_1, 1, 1, q_2], \quad 4 : [q_1, 2, 1, q_2], \\ & 5 : [q_2, /, +, q_1], \quad 6 : [q_3, 0, 0, q_f], \\ & 7 : [q_3, 1, 1, q_4], \quad 8 : [q_4, /, -, q_3]\} . \end{aligned}$$

When given an input string consisting of 1's and 2's, T_{erase} rewrite all the occurrences of 2's into 1's, and halts (see Fig. 18). The set of quadruples of the RTM T^R_{erase} that simulate T_{erase} is given below, and its computation process for the input 122 is shown in Fig. 19.

$$\begin{array}{llllllll}
[q_0, & /, & [+,+,0], & q'_0] & [p'_0, & /, & [-,-,0], & p_0] \\
[q'_0, & [x,0,0], & [x,1,0], & q_1] & [p_1, & [x,1,0], & [x,0,0], & p'_0] \\
[q_1, & /, & [0,+,0], & q'_1] & [p'_1, & /, & [0,-,0], & p_1] \\
[q'_1, & [0,0,0], & [0,2,0], & q_4] & [p_4, & [0,2,0], & [0,0,0], & p'_1] \\
[q'_1, & [1,0,0], & [1,3,0], & q_2] & [p_2, & [1,3,0], & [1,0,0], & p'_1] \\
[q'_1, & [2,0,0], & [1,4,0], & q_2] & [p_2, & [1,4,0], & [2,0,0], & p'_1] \\
[q_2, & /, & [+,+,0], & q'_2] & [p'_2, & /, & [-,-,0], & p_2] \\
[q'_2, & [x,0,0], & [x,5,0], & q_1] & [p_1, & [x,5,0], & [x,0,0], & p'_2] \\
[q_3, & /, & [0,+,0], & q'_3] & [p'_3, & /, & [0,-,0], & p_3] \\
[q'_3, & [0,0,0], & [0,6,0], & q_f] & [p_f, & [0,6,0], & [0,0,0], & p'_3] \\
[q'_3, & [1,0,0], & [1,7,0], & q_4] & [p_4, & [1,7,0], & [1,0,0], & p'_3] \\
[q_4, & /, & [-,+,0], & q'_4] & [p'_4, & /, & [+,-,0], & p_4] \\
[q'_4, & [x,0,0], & [x,8,0], & q_3] & [p_3, & [x,8,0], & [x,0,0], & p'_4]
\end{array}$$

$$\begin{array}{llllllll}
[q_f, & [0,6,0], & [0,6,0], & c_1] & [c'_1, & [0,6,0], & [0,6,0], & p_f] \\
[c_1, & /, & [+,0,+], & c_2] & [c'_2, & /, & [-,0,-], & c'_1] \\
[c_2, & [y,6,0], & [y,6,y], & c_1] & [c'_1, & [y,6,y], & [y,6,y], & c'_2] \\
 & & [c_2, [0,6, & 0], & [0,6, & 0], c'_2] & &
\end{array}$$

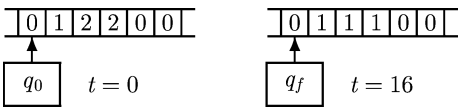

Reversible Computing, Figure 18
The initial and the final computational configuration of T_{erase} for the given input string 122

It is also possible to convert an arbitrary (irreversible) TM into an equivalent 1-tape 2-symbol RTM [31]. There are studies related to computational complexity of RTMs [8,10,21]. Bennett [8] showed the following space-time trade-off theorem for RTMs, where an interesting method of reducing space is given. Note that the RTM in Theorem 5 uses large amount of space, which is proportional to the time that the simulated TM uses.

Theorem 6 (Bennett [8]) *For any $\varepsilon > 0$ and any (irreversible) TM using time T and space S, we can construct an RTM that simulates the former using time $O(T^{1+\varepsilon})$ and space $O(S \log T)$.*

Constructing RTMs by REs

RTMs are related to many other models of reversible computing. For example, for any given RTM, there is a one-dimensional reversible cellular automaton (RCA) that simulates the RTM directly [32]. Hence, a one-dimensional RCA is computation-universal.

It is also possible to construct a garbage-less reversible logic circuit that simulates a given RTM. We have already seen that any reversible sequential machine can be realized as a garbage-less reversible RE circuit (Theorem 4). In [36], it is shown that an RTM can be decomposed into tape cell modules and a finite state control module, each of which is a reversible sequential machine. Therefore, any RTM can be composed of REs as a garbage-less circuit.

Figure 20 shows an example of an RTM composed only of REs [36]. There is a finite state control in the left half of this figure. To the right of it, infinitely many copies of tape cell modules are connected. (Note that, here, some ad hoc techniques are used to reduce the number of REs rather than to use the systematic method employed in Theorem 4.)

Note that this circuit is closed except the Begin and the End ports. This RTM computes the function $f(n) = 2n$ if n is given as a unary number [36]. Setting the initial states of tape cell modules, and then giving a signal from Begin port, a computing process starts. If the RTM halts, a signal goes out from the End port, leaving an answer in the tape cell modules.

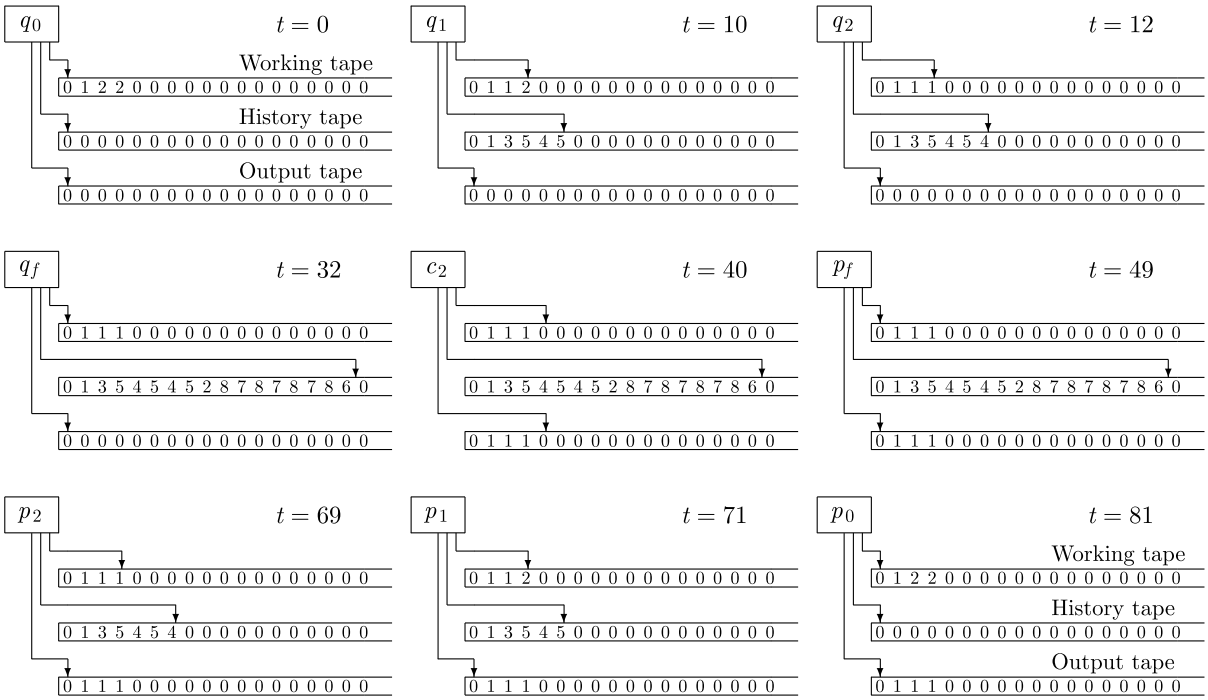

Reversible Computing, Figure 19

The computation process of the RTM T_{erase} for the input string 122: the forward computation stage (from $t = 0$ to $t = 32$), the copying stage (from $t = 33$ to $t = 48$), and the backward computation stage (from $t = 49$ to $t = 81$)

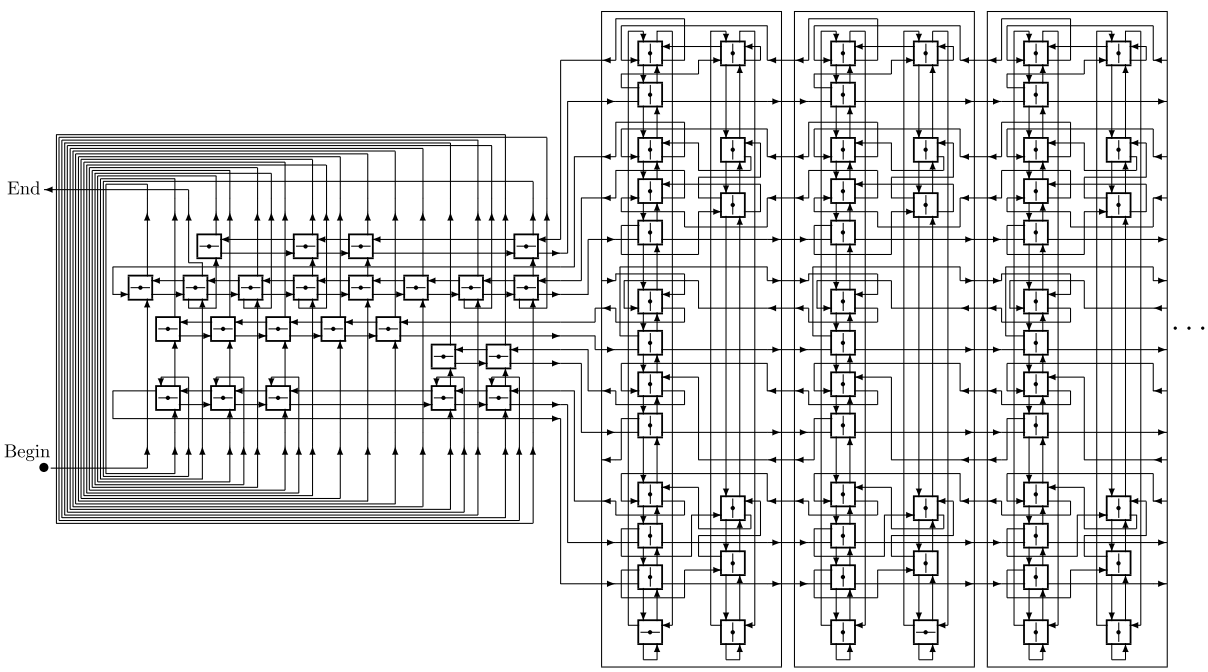

Reversible Computing, Figure 20

A circuit composed only of REs that realizes an RTM [36]

It is also possible to realize RTMs by Fredkin gates and delay elements as a garbage-less circuit using the method of composing RSMs by Fredkin gates shown in [33]. If we do so, however, we need a larger number of Fredkin gates than that of REs. In addition, when designing the circuit, signal delays and timings should be very carefully adjusted, because signals must be synchronized at each gate. On the other hand, the circuit of Fig. 20 works correctly even if each RE operates asynchronously, since only one signal is moving in this circuit. In an RE, a single moving signal interacts with a rotating bar, where the latter is thought as a kind of stationary signal. Hence, its operation may be performed at any time, and there is no need of synchronizing signals at each RE. Of course, if two or more signals are moving in a circuit, some control mechanism for timing is required.

Universal RTMs

From Theorem 5, we can see the existence of a universal RTM (URTM) that can compute any recursive function. However, if we use the technique in the proof of Theorem 5, the numbers of states and symbols of a URTM will become very large.

As for classical (i. e., irreversible) universal Turing machines (UTMs), the following UTMs have been proposed so far, where UTM(m, n) denotes an m-state n-symbol UTM: a UTM(7,4) by Minsky [30], a UTM(24,2), a UTM(10,3), a UTM(7,4), a UTM(5,5), a UTM(4,6), a UTM(3,10) and a UTM(2,18) by Rogozhin [44], a UTM(19,2) by Baiocchi [2], a UTM(3,9) by Kudlek and Rogozhin [23], etc. Most of these small UTMs simulates a 2-tag system [30], which is a simple string rewriting system having computation-universality.

In the following, we describe a 17-state 5-symbol URTM (URTM(17,5)) proposed by Morita and Yamaguchi [39]. It simulates a cyclic tag system (CTAG) given by Cook [11], which is also a very simple string rewriting system with computation-universality. Since the notion of halting was not defined explicitly in the original definition of a CTAG, we use here a modified definition of a CTAG with the halting property, which can simulate a 2-tag system with the halting property [39].

Definition 9 A *cyclic tag system* (CTAG) is defined by

$$C = (k, \{Y, N\}\,,\ \ (\text{halt}, p_1, \ldots, p_{k-1}))\,,$$

where k ($k = 1, 2, \ldots$) is the length of a cycle (i. e., period), $\{Y, N\}$ is the (fixed) alphabet, and $(p_1, \ldots, p_{k-1}) \in (\{Y, N\}^*)^{k-1}$ is a $(k-1)$-tuple of production rules. A pair (v, m) is called an *instantaneous description* (ID) of C, where $v \in \{Y, N\}^*$ and $m \in \{0, \ldots, k-1\}$. m is called a *phase* of the ID. A *halting ID* is an ID of the form $(Yv, 0)(v \in \{Y, N\}^*)$. The transition relation $\Rightarrow$ is defined as follows. For any $(v, m), (v', m') \in \{Y, N\}^* \times \{0, \ldots, k-1\}$,

$$\begin{aligned}(Yv, m) \Rightarrow (v', m') \text{ iff } [m \neq 0] \wedge [m' = m + 1 \mod k] \\ \wedge [v' = vp_m]\,, \\ (Nv, m) \Rightarrow (v', m') \text{ iff } [m' = m + 1 \mod k] \\ \wedge [v' = v]\,.\end{aligned}$$

A sequence of IDs $((v_0, m_0), \ldots, (v_n, m_n))$ is called a *complete computation starting from* $v \in \{Y, N\}^*$ iff $(v_0, m_0) = (v, 0), (v_i, m_i) \Rightarrow (v_{i+1}, m_{i+1})$ $(i = 0, 1, \ldots, n-1)$, and (v_n, m_n) is a halting ID.

Example 3 Consider the CTAG $C_1 = (3, \{Y, N\}, (\text{halt}, YN, YY))$. The complete computation starting from NYY is $(NYY,\ 0) \Rightarrow (YY, 1) \Rightarrow (YYN, 2) \Rightarrow (YNYY, 0)$.

Theorem 7 (Morita and Yamaguchi, [39]) *There is a URTM(17,5).*

A URTM(17,5) $T_{17,5}$ in the quintuple form that simulates any CTAG is as follows [39].

$$T_{17,5} = (\{q_0, \ldots, q_{16}\}\,,\ \ \{b, Y, N, *, \$\}, q_0, b, \delta)\,,$$

where the set δ of quintuples is shown in Table 5. (Note that, in a construction of a UTM, the final state is usually omitted from the state set.) There are 67 quintuples in total. If a CTAG halts with a halting ID, then $T_{17,5}$ halts in the state q_1. If the string becomes an empty string, then it halts in the state q_2. In Table 5, it is indicated by "null".

Figure 21 shows how the CTAG C_1 with the initial string NYY in Example 3 is simulated by the URTM $T_{17,5}$. On the tape of the URTM, the production rules (halt, NY, YY) of C_1 are expressed by the reversal sequence over $\{Y, N, *\}$, i. e., $YY * YN * *$, where $*$ is used as a delimiter between rules, and "halt" is represented by the empty string. Note that in the initial tape of $T_{17,5}$ ($t = 0$), the rightmost $*$ is replaced by b. This indicates that the phase is 0. In general, if the phase is $i - 1$, then the ith $*$ from the right is replaced by b. This symbol b is called a "phase marker". On the other hand, the given initial string for C_1 is placed to the right of the rules, where \$ is used as a delimiter.

The IDs in the complete computation $(NYY, 0) \Rightarrow (YY, 1) \Rightarrow (YYN, 2) \Rightarrow (YNYY, 0)$ of C_1 appear in the computational configurations of $T_{17,5}$ at $t = 0, 6, 59$ and 142, respectively. The symbol \$ in the final string ($t = 148$) should be regarded as Y.

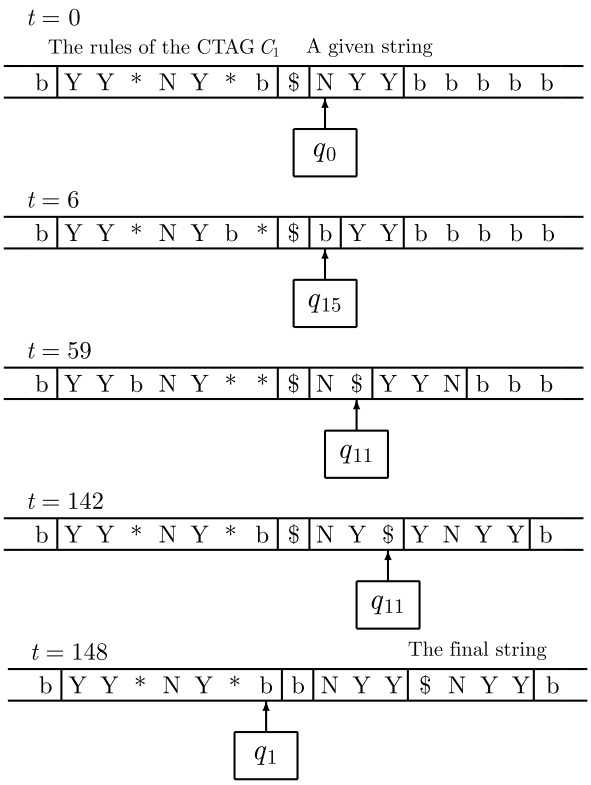

Reversible Computing, Figure 21
Simulating the CTAG C_1 in Example 3 by the URTM $T_{17,5}$

Reversible Computing, Table 5
The set of quintuples of the URTM $T_{17,5}$ [39]

	b	Y	N	$*$	$\$$
q_0	$\$ - q_2$	$\$ - q_1$	$b - q_{13}$		
q_1	halt	$Y - q_1$	$N - q_1$	$* + q_0$	$b - q_1$
q_2	$* - q_3$	$Y - q_2$	$N - q_2$	$* - q_2$	null
q_3	$b + q_{12}$	$b + q_4$	$b + q_7$	$b + q_{10}$	
q_4	$Y + q_5$	$Y + q_4$	$N + q_4$	$* + q_4$	$\$ + q_4$
q_5	$b - q_6$				
q_6	$Y - q_3$	$Y - q_6$	$N - q_6$	$* - q_6$	$\$ - q_6$
q_7	$N + q_8$	$Y + q_7$	$N + q_7$	$* + q_7$	$\$ + q_7$
q_8	$b - q_9$				
q_9	$N - q_3$	$Y - q_9$	$N - q_9$	$* - q_9$	$\$ - q_9$
q_{10}		$Y + q_{10}$	$N + q_{10}$	$* + q_{10}$	$\$ + q_{11}$
q_{11}		$Y + q_{11}$	$N + q_{11}$	$* + q_{11}$	$Y + q_0$
q_{12}		$Y + q_{12}$	$N + q_{12}$	$* + q_{12}$	$\$ - q_3$
q_{13}	$* - q_{14}$	$Y - q_{13}$	$N - q_{13}$	$* - q_{13}$	$\$ - q_{13}$
q_{14}	$b + q_{16}$	$Y - q_{14}$	$N - q_{14}$	$b + q_{15}$	
q_{15}	$N + q_0$	$Y + q_{15}$	$N + q_{15}$	$* + q_{15}$	$\$ + q_{15}$
q_{16}		$Y + q_{16}$	$N + q_{16}$	$* + q_{16}$	$\$ - q_{14}$

It is easy to see that $T_{17,5}$ is reversible by checking the set of quintuples shown in Table 5 according to the definition of an RTM. Intuitively, its reversibility is guaranteed from the fact that no information is erased in the whole simulation process. Furthermore, every branch of the program caused by reading the symbol Y or N is "merged reversibly" by writing the original symbol. For example, the states q_{11} and q_{15} transit to the same state q_0 by writing Y and N, respectively, using the quintuples $[q_{11}, \$, Y, +, q_0]$ and $[q_{15}, b, N, +, q_0]$.

Other Models of Reversible Computing

There are also several models of reversible computing that are not dealt with in the previous sections. Here, we discuss a few of them briefly.

A *reversible cellular automaton* (RCA) is an important model, because it can deal with reversible spatio-temporal phenomena. In fact, it can be thought as an abstract model of a reversible space. So far, a lot of interesting results and properties on RCAs have been shown (see e. g., [35,49]). Some of them are described in ► Reversible Cellular Automata.

A counter machine (CM) is a simple model of a computing consisting of a read-only input tape, a finite number of counters, and a finite state control. It is known that a CM with only two counters has computation-universality [30]. A *reversible counter machine* (RCM) is a backward deterministic version of a CM. An RCM with only two counters is known to be computation-universal [34], though it is a very simple model. This is useful to show universality of other reversible systems.

A *reversible finite automaton* (RFA) is also a backward deterministic version of a finite automaton. Pin [43] studied this model, and characterized its language accepting ability from the formal language theory. Note that an RSM in Sect. "Definitions on Reversible Logic Elements with Memory" is an RFA augmented by an output mechanism.

Future Directions

How Can We Realize Reversible Computers as a Hardware?

Although we have discussed reversible computing mainly from the standpoint of computation theory, it is a very important problem how reversible computers can be realized as a hardware. So far, there have been many interesting attempts from engineering side: e. g., implementing reversible logic as electrically controlled switches [29], c-MOS implementation of reversible logic gates and circuits [13], adiabatic circuits for reversible

computer [17,18], and synthesis of reversible logic circuits [1,45]. However, ultimately, reversible logic elements and computers should be implemented in atomic or molecular level. It is plausible that some nano-scale phenomena can be used as primitives of reversible computing, because the microscopic physical law is reversible. Of course, finding such solution is very difficult, but it is an interesting and challenging problem left for future investigations.

How Simple Can Reversible Computers Be?

To find very simple reversible logic elements with universality is an important problem from both theoretical and practical viewpoints. We may assume, in general, hardware implementation will become easier, if we can find much simpler logical primitives from which reversible computers can be built. As we have discussed in Sect. "Reversible Logic Gates", the Fredkin gate and the Toffoli gate are logically universal gates with minimum number of inputs and outputs. On the other hand, as for reversible logic elements with memory (RLEMs) discussed in Sect. "Reversible Logic Elements with Memory", there are 14 2-state 3-symbol logically universal RLEMS [38,40], which have a less number of symbols than an RE. However, it is an open problem whether there is a single 2-state 2-symbol RLEM that is logically universal (but it is known that there is a set of two 2-state 2-symbol RLEMs that is logically universal [27]). It is also an interesting problem to find much smaller universal reversible Turing machines than $T_{17,5}$ in Sect. "Universal RTMs".

Novel Architectures for Reversible Computers

If we try to construct more realistic computers from reversible logic primitives, we shall need new design theories suited for it. For example, a circuit composed of REs shown in Fig. 20 has a very different feature from traditional logic circuits made of gates. There will surely be many other possibilities of new design techniques that cannot be imagined from the classical theory of logic circuits. To realize efficient reversible computers, development of novel design methods and architectures are necessary.

Asynchronous and Continuous Time Reversible Models

There are still other problems between reversible physical systems and reversible computing systems. The first one is asynchronous systems versus synchronous ones. When building a computing system with a huge number of elements, it is preferable if it is realized as an asynchronous system, because if a clock is eliminated, then power dissipation is reduced, and each element can operate at any time independent to other elements [26,41]. However, it is a difficult problem to define reversibility in asynchronous systems properly. The second problem is continuous time versus discrete time. Natural physical systems are continuous (at least they seem so), while most reversible computing models are defined as discrete systems. To bridge a gap between reversible physical systems and reversible computing models, and to implement the latter in the former systems, we shall need further investigations on these problems.

Bibliography

Primary Literature

1. Al-Rabadi AN (2004) Reversible Logic Synthesis. Springer, Berlin
2. Baiocchi C (2001) Three small universal Turing machines. 3rd Int Conference on Machines, Computations, and Universality. LNCS, vol 2055. Springer, Berlin, pp 1–10
3. Barenco A, Bennett CH, Cleve R, DiVincenzo DP, Margolus N, Shor P, Sleator T, Smolin J, Weinfurter H (1995) Elementary gates for quantum computation. Phys Rev A 52:3457–3467
4. Bennett CH (1973) Logical reversibility of computation. IBM J Res Dev 17:525–532
5. Bennett CH (1982) The thermodynamics of computation. Int J Theoret Phys 21:905–940
6. Bennett CH, Landauer R (1985) The fundamental physical limits of computation. Sci Am 253:38–46
7. Bennett CH (1987) Demons, engines, and the second law. Sci Am 257:108–117
8. Bennett CH (1989) Time/space trade-offs for reversible computation. SIAM J Comput 18:766–776
9. Bennett CH (2003) Notes on Landauer's principle, reversible computation, and Maxwell's Demon. Stud Hist Philos Mod Phys 34:501–510
10. Buhrman H, Tromp J, Vitanyi P (2001) Time and space bounds for reversible simulation. J Phys A 34:6821–6830
11. Cook M (2004) Universality in elementary cellular automata. Complex Syst 15:1–40
12. Deutsch D (1985) Quantum theory, the Church–Turing principle and the universal quantum computer. Proc R Soc Lond A 400:97–117
13. De Vos A, Desoete B, Adamski A, Pietrzak P, Sibinski M, Widerski T (2000) Design of reversible logic circuits by means of control gates. PATMOS 2000. LNCS vol 1918. Springer, Berlin, pp 255–264
14. Durand-Lose J (2002) Computing inside the billiard ball model. In: Adamatzky A (ed) Collision-based Computing. Springer, Londeon, pp 135–160
15. Feynman RP (1982) Simulating physics with computers. Int J Theoret Phys 21:467–488
16. Feynman RP (1996) Hey AJG, Allen RW (eds) Feynman lectures on computation. Perseus Books, Massachusetts
17. Frank M, Vieri C, Ammer MJ, Love N, Margolus NH, Knight TE (1998) A scalable reversible computer in silicon. In: Calude CS,

Casti J, Dinneen MJ (eds) Unconventional Models of Computation. Springer, Singapore, pp 183–200
18. Frank M (1999) Reversibility for Efficient Computing. Ph D Dissertation, MIT, Cambridge
19. Fredkin E, Toffoli T (1982) Conservative logic. Int J Theoret Phys 21:219–253
20. Gruska J (1999) Quantum Computing. McGraw-Hill, London
21. Jacopini G, Mentrasti P, Sontacchi G (1990) Reversible Turing machines and polynomial time reversibly computable functions. SIAM J Disc Math 3:241–254
22. Keys RW, Landauer R (1970) Minimal energy dissipation in logic. IBM J Res Dev 14:152–157
23. Kudlek M, Rogozhin Y (2002) A universal Turing machine with 3 states and 9 symbols. In: Kuich w, Rozenberg G, Salomaa A (eds) Developments in Language Theory (DLT 2001). LNCS vol 2295. Springer, Berlin, pp 311–318
24. Landauer R (1961) Irreversibility and heat generation in the computing process. IBM J Res Dev 5:183–191
25. Lecerf Y (1963) Machines de Turing réversibles – Reursive insolubilité en $n \in N$ de l'équation $u = \theta^n u$, où θ est un isomorphisme de codes. Comptes Rendus Hebd Séances L'acad Sci 257:2597–2600
26. Lee J, Peper F, Adachi S, Mashiko S (2004) On reversible computation in asynchronous systems. In: Hida T, Saito k, Si S (eds) Quantum Information and Complexity. World Scientific, Singapore, pp 296–320
27. Lee J, Peper F, Adachi S, Morita K (2004) An asynchronous cellular automaton implementing 2-state 2-input 2-output reversed-twin reversible elements, Proc Eighth Int Conf on Cellular Automata for Research and Industry, Yokohama, Sep 2008, (to appear)
28. Margolus N (1984) Physics-like model of computation. Physica 10D:81–95
29. Merkle RC (1993) Reversible electronic logic using switches. Nanotechnology 4:20–41
30. Minsky ML (1967) Computation: Finite and Infinite Machines. Prentice-Hall, Englewood Cliffs
31. Morita K, Shirasaki A, Gono Y (1989) A 1-tape 2-symbol reversible Turing machine. Trans IEICE Japan E-72:223–228
32. Morita K, Harao M (1989) Computation universality of one-dimensional reversible (injective) cellular automata. Trans IEICE Japan E-72:758–762
33. Morita K (1990) A simple construction method of a reversible finite automaton out of Fredkin gates, and its related problem. Trans IEICE Japan E-73:978–984
34. Morita K (1996) Universality of a reversible two-counter machine. Theoret Comput Sci 168:303–320
35. Morita K (2001) Cellular automata and artificial life – computation and life in reversible cellular automata. In: Goles E, Martinez S (eds) Complex Systems. Kluwer Academic Publishers, Dordrecht, pp 151–200
36. Morita K (2001) A simple reversible logic element and cellular automata for reversible computing. In: Proc 3rd Int Conf on Machines, Computations, and Universality. LNCS, vol 2055. Springer, Berlin, pp 102–113
37. Morita K (2003) A new universal logic element for reversible computing. In: Martin-Vide C, Mitrana V (eds) Grammars and Automata for String Processing. Taylor & Francis, London, pp 285–294
38. Morita K, Ogiro T, Tanaka K, Kato H (2005) Classification and universality of reversible logic elements with one-bit memory. In: Proc 4th Int Conf on Machines, Computations, and Universality. LNCS, vol 3354. Springer, Berlin, pp 245–256
39. Morita K, Yamaguchi Y (2007) A universal reversible Turing machine. In: Proc 5th Int Conf on Machines, Computations, and Universality. LNCS, vol 4664. Springer, Berlin, pp 90–98
40. Ogiro T, Kanno A, Tanaka K, Kato H, Morita K (2005) Nondegenerate 2-state 3-symbol reversible logic elements are all universal. Int J Unconv Comput 1:47–67
41. Peper F, Lee J, Adachi S, Mashiko S (2003) Laying out circuits on asynchronous cellular arrays: a step towards feasible nanocomputers. Nanotechnology 14:469–485
42. Petri CA (1967) Grundsätzliches zur Beschreibung diskreter Prozesse. In: Proc 3rd Colloquium über Automatentheorie. Birkhäuser, Basel, pp 121–140
43. Pin JE (1992) On reversible automata. In: Proc Latin American Symp on Theoretical Informatics. LNCS vol 583. Springer, Berlin, pp 401–416
44. Rogozhin Y (1996) Small universal Turing machines. Theor Comput Sci 168:215–240
45. Shende VV, Prasad AK, Markov IL, Hayes JP (2003) Synthesis of reversible logic circuits. IEEE Trans Computer-Aided Des Integr Circuits Syst 22:710–722
46. Toffoli T (1977) Computation and construction universality of reversible cellular automata. J Comput Syst Sci 15:213–231
47. Toffoli T (1980) Reversible computing, Automata, Languages and Programming. In: LNCS, vol 85. Springer, Berlin, pp 632–644
48. Toffoli T (1981) Bicontinuous extensions of invertible combinatorial functions. Math Syst Theory 14:12–23
49. Toffoli T, Margolus N (1990) Invertible cellular automata: a review. Physica D 45:229–253

Books and Reviews

Adamatzky A (ed) (2002) Collision-Based Computing. Springer, London
Bennett CH (1988) Notes on the history of reversible computation. IBM J Res Dev 32:16–23
Milburn GJ (1998) The Feynman Processor. Perseus Books, Reading
Vitanyi P (2005) Time, space, and energy in reversible computing. In: Proc 2005 ACM Int Conf on Computing Frontiers, Ischia, pp 435–444

Robotic Networks, Distributed Algorithms for

Francesco Bullo[1], Jorge Cortés[2], Sonia Martínez[2]
[1] Department of Mechanical Engineering, University of California, Santa Barbara, USA
[2] Department of Mechanical and Aerospace Engineering, University of California, San Diego, USA

Article Outline

Glossary
Definition of the Subject

Introduction
Distributed Algorithms on Networks of Processors
Distributed Algorithms for Robotic Networks
Bibliographical Notes
Future Directions
Acknowledgments
Bibliography

Glossary

Cooperative control In recent years, the study of groups of robots and multi-agent systems has received a lot of attention. This interest has been driven by the envisioned applications of these systems in scientific and commercial domains. From a systems and control theoretic perspective, the challenges in cooperative control revolve around the analysis and design of distributed coordination algorithms that integrate the individual capabilities of the agents to achieve a desired coordination task.

Distributed algorithm In a network composed of multiple agents, a coordination algorithm specifies a set of instructions for each agent that prescribe what to sense, what to communicate and to whom, how to process the information received, and how to move and interact with the environment. In order to be scalable, coordination algorithms need to rely as much as possible on local interactions between neighboring agents.

Complexity measures Coordination algorithms are designed to enable networks of agents achieve a desired task. Since different algorithms can be designed to achieve the same task, performance metrics are necessary to classify them. Complexity measures provide a way to characterize the properties of coordination algorithms such as completion time, cost of communication, energy consumption, and memory requirements.

Averaging algorithms Distributed coordination algorithms that perform weighted averages of the information received from neighboring agents are called averaging algorithms. Under suitable connectivity assumptions on the communication topology, averaging algorithms achieve agreement, i. e., the state of all agents approaches the same value. In certain cases, the agreement value can be explicitly determined as a function of the initial state of all agents.

Leader election In leader election problems, the objective of a network of processors is to elect a leader. All processors have a variable "leader" initially set to unknown. The leader-election task is solved when only one processor has set the variable "leader" to `true`, and all other processors have set it to `false`.

LCR algorithm The classic Le Lann–Chang–Roberts (LCR) algorithm solves the leader election task on a static network with the ring communication topology. Initially, each agent transmits its unique identifier to its neighbors. At each communication round, each agent compares the largest identifier received from other agents with its own identifier. If the received identifier is larger than its own, the agent declares itself a non-leader, and transmits it in the next communication round to its neighbors. If the received identifier is smaller than its own, the agent does nothing. Finally, if the received identifier is equal to its own, it declares itself a leader. The LCR algorithm achieves leader election with linear time complexity and quadratic total communication complexity, respectively.

Agree-and-pursue algorithm Coordination algorithms for robotic networks combine the features of distributed algorithms for networks of processors with the sensing and control capabilities of the robots. The agree-and-pursue motion coordination algorithm is an example of this fusion. Multiple robotic agents moving on a circle seek to agree on a common direction of motion while at the same achieving an equally-spaced distribution along the circle. The agree-and-pursue algorithm achieves both tasks combining ideas from leader election on a changing communication topology with basic control primitives such as "follow your closest neighbor in your direction of motion."

Definition of the Subject

The study of distributed algorithms for robotic networks is motivated by the recent emergence of low-power, highly-autonomous devices equipped with sensing, communication, processing, and control capabilities. In the near future, cooperative robotic sensor networks will perform critical tasks in disaster recovery, homeland security, and environmental monitoring. Such networks will require efficient and robust distributed algorithms with guaranteed quality-of-service. In order to design coordination algorithms with these desirable capabilities, it is necessary to develop new frameworks to design and formalize the operation of robotic networks and novel tools to analyze their behavior.

Introduction

Distributed algorithms are a classic subject of study for networks composed of individual processors with communication capabilities. Within the automata-theoretic literature, important research topics on distributed algo-

rithms include the introduction of mathematical models and precise specifications for their behavior, the formal assessment of their correctness, and the characterization of their complexity.

Robotic networks have distinctive features that make them unique when compared with networks of static processors. These features include the operation under ad-hoc dynamically changing communication topologies and the complexity that results from the combination of continuous- and discrete-time dynamics. The spatially-distributed character of robotic networks and their dynamic interaction with the environment make the classic study of distributed algorithms, typically restricted to static networks, not directly applicable.

This chapter brings together distributed algorithms for networks of processors and for robotic networks. The first part of the chapter is devoted to a formal discussion about distributed algorithms for a synchronous network of processors. This treatment serves as a brief introduction to important issues considered in the literature on distributed algorithms such as network evolution, task completion, and complexity notions. To illustrate the ideas, we consider the classic Le Lann–Chang–Roberts (LCR) algorithm, which solves the leader election task on a static network with the ring communication topology. Next, we present a class of distributed algorithms called averaging algorithms, where each processor computes a weighted average of the messages received from its neighbors. These algorithms can be described as linear dynamical systems, and their correctness and complexity analysis has nice connections with the fields of linear algebra and Markov chains.

The second part of the chapter presents a formal model for robotic networks that explicitly takes into account communication, sensing, control, and processing. The notions of time, communication, and space complexity introduced here allow us to characterize the performance of coordination algorithms, and rigorously compare the performance of one algorithm versus another. In general, the computation of these notions is a complex problem that requires a combination of tools from dynamical systems, control theory, linear algebra, and distributed algorithms. We illustrate these concepts in three different scenarios: the agree-and-pursue algorithm for a group of robots moving on a circle, aggregation algorithms that steer robots to a common location, and deployment algorithms that make robots optimally cover a region of interest. In each case, we report results on the complexity associated to the achievement of the desired task. The chapter ends with a discussion about future research directions.

Distributed Algorithms on Networks of Processors

Here we introduce a synchronous network as a group of processors with the ability to exchange messages and perform local computations. What we present is a basic classic model studied extensively in the distributed algorithms literature. Our treatment is directly adopted with minor variations from the texts [57] and [74].

Physical Components and Computational Models

Loosely speaking, a synchronous network is a group of processors, or nodes, that possess a local state, exchange messages among neighbors, and compute an update to their local state based on the received messages. Each processor alternates the two tasks of exchanging messages with its neighboring processors and of performing a computation step.

Let us begin by providing some basic definitions. A *directed graph* [22], in short *digraph*, of order n is a pair $G = (V, E)$ where V is a set with n elements called *vertices* (or sometimes *nodes*) and E is a set of ordered pair of vertices called *edges*. In other words, $E \subseteq V \times V$. We call V and E the *vertex set* and *edge set*, respectively. For $u, v \in V$, the ordered pair (u, v) denotes an edge *from* u *to* v. The vertex u is called an *in-neighbor* of v, and v is called an *out-neighbor* of u. A *directed path* in a digraph is an ordered sequence of vertices such that any two consecutive vertices in the sequence are a directed edge of the digraph. A vertex of a digraph is *globally reachable* if it can be reached from any other vertex by traversing a directed path. A digraph is *strongly connected* if every vertex is globally reachable.

A *cycle* in a digraph is a non-trivial directed path that starts and ends at the same vertex. A digraph is *acyclic* if it contains no cycles. In an acyclic graph, every vertex with no in-neighbors is named *source*, and every vertex with no out-neighbors is named *sink*. A *directed tree* is an acyclic digraph with the following property: there exists a vertex, called the root, such that any other vertex of the digraph can be reached by one and only one path starting at the root. A *directed spanning tree*, or simply a *spanning tree*, of a digraph is a subgraph that is a directed tree and has the same vertex set as the digraph. A *directed chain* is a directed tree with exactly one source and one sink. A *directed ring digraph* is the cycle obtained by adding to the edge set of a chain a new edge from its sink to its source. Figure 1 illustrates these notions.

The physical component of a synchronous network S is a digraph (I, E_{cmm}), where

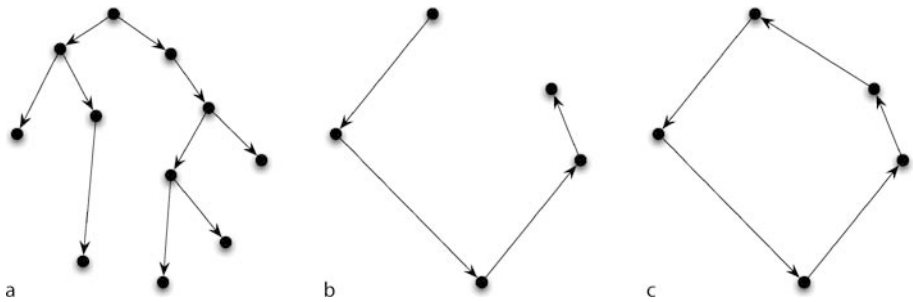

Robotic Networks, Distributed Algorithms for, Figure 1
From left to right, directed tree, chain, and ring digraphs

(i) $I = \{1, \ldots, n\}$ is called the *set of unique identifiers (UIDs)*, and
(ii) E_{cmm} is a set of directed edges over the vertices $\{1, \ldots, n\}$, called the communication links.

The set E_{cmm} models the topology of the communication service among the nodes: for $i, j \in \{1, \ldots, n\}$, processor i can send a message to processor j if the directed edge (i, j) is present in E_{cmm}. Note that, unlike the standard treatments in [57] and [74], we do not assume the digraph to be strongly connected; the required connectivity assumption is specified on a case by case basis.

Next, we discuss the state and the algorithms that each processor possesses and executes, respectively. By convention, we let the superscript $[i]$ denote any quantity associated with the node i. A *distributed algorithm* $\mathcal{DA}$ for a network S consists of the sets:

(i) $\mathbb{A}$, a set containing the `null` element, called the *communication alphabet*; elements of $\mathbb{A}$ are called *messages*;
(ii) $W^{[i]}$, $i \in I$, called the *processor state sets*;
(iii) $W_0^{[i]} \subseteq W^{[i]}$, $i \in I$, sets of *allowable initial values*;

and of the maps:

(i) $\text{msg}^{[i]} : W^{[i]} \times I \to \mathbb{A}$, $i \in I$, called *message-generation functions*;
(ii) $\text{stf}^{[i]} : W^{[i]} \times \mathbb{A}^n \to W^{[i]}$, $i \in I$, called *state-transition functions*.

If $W^{[i]} = W$, $\text{msg}^{[i]} = \text{msg}$, and $\text{stf}^{[i]} = \text{stf}$ for all $i \in I$, then $\mathcal{DA}$ is said to be *uniform* and is described by a tuple $(\mathbb{A}, W, \{W_0^{[i]}\}_{i \in I}, \text{msg}, \text{stf})$.

Now, with all elements in place, we can explain in more detail how a synchronous network executes a distributed algorithm. The *state* of processor i is a variable $w^{[i]} \in W^{[i]}$, initially set equal to an allowable value in $W_0^{[i]}$. At each time instant $\ell \in \mathbb{Z}_{\geq 0}$, processor i sends to each of its out-neighbors j in the communication digraph (I, E_{cmm}) a message (possibly the `null` message) computed by applying the message-generation function $\text{msg}^{[i]}$ to the current values of its state $w^{[i]}$ and to the identity j. Subsequently, but still at time instant $\ell \in \mathbb{Z}_{\geq 0}$, processor i updates the value of its state $w^{[i]}$ by applying the state-transition function $\text{stf}^{[i]}$ to the current value of its state $w^{[i]}$ and to the messages it receives from its in-neighbors. Note that, at each round, the first step is transmission and the second one is computation.

We conclude this section with two sets of remarks. We first discuss some aspects of our communication model that have a large impact on subsequent development. We then collect a few general comments about distributed algorithms on networks.

Remark 1 (Aspects of the communication model)

(i) The network S and the algorithm $\mathcal{DA}$ are referred to as *synchronous* because the communications between all processors takes place at the same time for all processors.
(ii) Communication is modeled as a so-called "point to point" service: a processor can specify different messages for different out-neighbors and knows the processor identity corresponding to any incoming message.
(iii) Information is exchanged between processors as messages, i. e., elements of the alphabet $\mathbb{A}$; the message `null` indicates no communication. Messages might encode logical expressions such as `true` and `false`, or finite-resolution quantized representations of integer and real numbers.
(iv) In some uniform algorithms, the messages between processors are the processors' states. In such cases, the corresponding communication alphabet is $\mathbb{A} = W \cup \{\texttt{null}\}$ and the message generation function $\text{msg}_{\text{std}}(w, j) = w$ is referred to as the *standard message-generation function*.

Remark 2 (Advanced topics: Control structures and failures)

(i) Processors in a network have only partial information about the network topology. In general, each processor only knows its own UID, and the UID of its in- and out-neighbors. Sometimes we assume that the processor knows the network diameter. In some cases [74], actively running networks might depend upon "control structures," i. e., structures that are computed at initial time and are exploited in subsequent algorithms. For example, routing tables might be computed for routing problems, "leader" processors might be elected and tree structures might be computed and represented in a distributed manner for various tasks, e. g., coloring or maximal independent set problems. We present some sample algorithms to compute these structures below.

(ii) A key issue in the study of distributed algorithms is the possible occurrence of failures. A network might experience intermittent or permanent communication failures: along given edges a `null` message or an arbitrary message might be delivered instead of the intended value. Alternatively, a network might experience various types of processor failures: a processor might transmit only `null` messages (i. e., the msg function returns `null` always), a processor might quit updating its state (i. e., the stf function neglects incoming messages and returns the current state value), or a processor might implement arbitrarily modified msg and stf functions. The latter situation, in which completely arbitrary and possibly malicious behavior is adopted by faulty nodes, is referred to as a Byzantine failure in the distributed algorithms literature.

Complexity Notions

Here we begin our analysis of the performance of distributed algorithms. We introduce a notion of algorithm completion and, in turn, we introduce the classic notions of time, space, and communication complexity.

We say that an algorithm *terminates* when only `null` messages are transmitted and all processors states become constants.

Remark 3 (Alternative termination notions)

(i) In the interest of simplicity, we have defined evolutions to be unbounded in time and we do not explicitly require algorithms to actually have termination conditions, i. e., to be able to detect when termination takes place.

(ii) It is also possible to define the termination time as the first instant when a given problem or task is achieved, independently of the fact that the algorithm might continue to transmit data subsequently.

The notion of time complexity measures the time required by a distributed algorithm to terminate. More specifically, the *(worst-case) time complexity* of a distributed algorithm $\mathcal{DA}$ on a network S, denoted $\mathrm{TC}(\mathcal{DA}, S)$, is the maximum number of rounds required by the execution of $\mathcal{DA}$ on S among all allowable initial states until termination.

Next, we quantify memory and communication requirements of distributed algorithms. From an information theory viewpoint [35], the information content of a memory variable or of a message is properly measured in bits. On the other hand, it is convenient to use the alternative notions of "basic memory unit" and "basic message." It is customary [74] to assume that a "basic memory unit" or a "basic message" contains $\log(n)$ bits so that, for example, the information content of a robot identifier $i \in \{1, \ldots, n\}$ is $\log(n)$ bits or, equivalently, one "basic memory unit." Note that elements of the processor state set W or of the alphabet set $\mathbb{A}$ might amount to multiple basic memory units or basic messages; the `null` message has zero cost. Unless specified otherwise, the following definitions and examples are stated in terms of basic memory unit and basic messages.

(i) The *(worst-case) space complexity* of a distributed algorithm $\mathcal{DA}$ on a network S, denoted by $\mathrm{SC}(\mathcal{DA}, S)$, is the maximum number of basic memory units required by a processor executing the $\mathcal{DA}$ on S among all processors and among all allowable initial states until termination.

(ii) The *(worst-case) communication complexity* of a distributed algorithm $\mathcal{DA}$ on a network S, denoted by $\mathrm{CC}(\mathcal{DA}, S)$, is the maximum number of basic messages transmitted over the entire network during the execution of $\mathcal{DA}$ among all allowable initial states until termination.

Remark 4 (Space complexity conventions) By convention, each processor knows its identity, i. e., it requires $\log(n)$ bits to represent its unique identifier in a set with n distinct elements. We do not count this cost in the space complexity of an algorithm.

We conclude this section by discussing ways of quantifying time, space, and communication complexity. The idea, borrowed from combinatorial optimization, is to adopt asymptotic "order of magnitude" measures. Formally, complexity bounds will be expressed with respect

to the Bachman–Laundau symbols O, Ω, and Θ. Let us be more specific. In the following definitions, f denotes a function from $\mathbb{N}$ to $\mathbb{R}$.

(i) We say that an algorithm has time complexity *of order* $\Omega(f(n))$ *over some network* if, for all n, there exists a network of order n and initial processor values such that the time complexity of the algorithm is greater than a constant factor times $f(n)$.
(ii) We say that an algorithm has time complexity *of order* $O(f(n))$ *over arbitrary networks* if, for all n, for all networks of order n and for all initial processor values the time complexity of the algorithm is lower than a constant factor times $f(n)$.
(iii) We say that an algorithm has time complexity *of order* $\Theta(f(n))$ if its time complexity is of order $\Omega(f(n))$ over some network and $O(f(n))$ over arbitrary networks at the same time.

We use similar conventions for space and communication complexity.

In many cases the complexity of an algorithm will typically depend upon the number of nodes of the network. It is therefore useful to present a few simple facts about these functions now. Over arbitrary digraphs $S = (I, E_{\text{cmm}})$ of order n, we have

$$\text{diam}(S) \in \Theta(n), \quad |E_{\text{cmm}}(S)| \in \Theta(n^2) \quad \text{and} \quad \text{radius}(v, S) \in \Theta(\text{diam}(S)) ,$$

where v is any node of S.

Remark 5 Numerous variations of these definitions are possible. Even though we will not pursue them here, let us provide some pointers.

(i) In the definition of lower bound, consider the logic quantifier describing the role of the network. The lower bound statement is "existential" rather than "global," in the sense that the bound does not hold for all graphs. As discussed in [74], it is possible to define also "global" lower bounds, i. e., lower bounds over all graphs, or lower bounds over specified classes of graphs.
(ii) The complexity notions introduced above focus on the worst-case situation. It is also possible to define *expected* or *average* complexity notions, where one might be interested in characterizing, for example, the average number of rounds required or the average number of basic messages transmitted over the entire network during the execution of an algorithm among all allowable initial states until termination.
(iii) It is possible to define complexity notions for problems, rather than algorithms, by considering, for example, the worst-case optimal performance among all algorithms that solve the given problem, or over classes of algorithms or classes of graphs.

Leader Election

We formulate here a classical problem in distributed networks and summarize its complexity measures.

Problem 6 (Leader election) *Assume that all processors of a network have a state variable, say* `leader`*, initially set to* `unknown`*. We say that a leader is elected when one and only one processor has the state variable set to* `true` *and all others have it set to* `false`*. Elect a leader.*

This is a task that is a bit more global in nature. We display here a solution that requires individual processors to know the diameter of the network, denoted by diam(S), or an upper bound on it.

> *[Informal description]* At each communication round, each agent sends to all its neighbors the maximum UID it has received up to that time. This is repeated for diam(S) rounds. At the last round, each agent compares the maximum received UID with its own, and declares itself a leader if they coincide, or a non-leader otherwise.

The algorithm is called FLOODMAX: the maximum UID in the network is transmitted to other agents in an incremental fashion. At the first communication round, agents that are neighbors of the agent with the maximum UID receive the message from it. At the next communication round, the neighbors of these agents receive the message with the maximum UID. This process goes on for diam(S) rounds to ensure that every agent receives the maximum UID. Note that there are networks for which all agents receive the message with the maximum UID in fewer communication rounds than diam(S). The algorithm is formally stated as follows.

`Synchronous Network:` $S = (\{1, \ldots, n\}, E_{\text{cmm}})$

`Distributed Algorithm: FLOODMAX`

`Alphabet:` $\mathbb{A} = \{1, \ldots, n\} \cup \{\texttt{null}\}$

`Processor State:` $w = (\texttt{my-id}, \texttt{max-id}, \texttt{leader}, \texttt{round})$, where

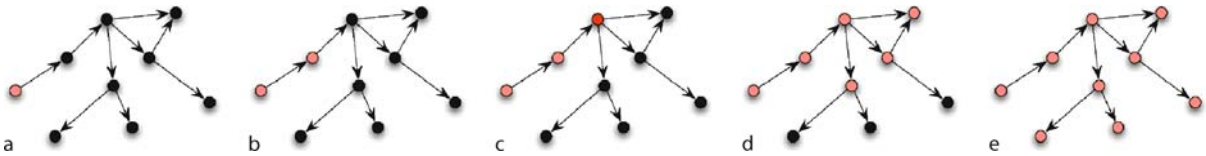

Robotic Networks, Distributed Algorithms for, Figure 2
Execution of the FLOODMAX algorithm. The diameter of the network is 4. In the leftmost frame, the agent with the maximum UID is colored in red. After 4 communication rounds, its message has been received by all agents

my-id $\in \{1, \ldots, n\}$,
initially: my-id$^{[i]} = i$ for all i
max-id $\in \{1, \ldots, n\}$,
initially: max-id$^{[i]} = i$ for all i
leader $\in$ {false, true, unknown},
initially: leader$^{[i]}$ = unknown for all i
round $\in \{0, 1, \ldots, \text{diam}(S)\}$,
initially: round$^{[i]} = 0$ for all i

function msg(w, i)
if {round $<$ diam(S)} **then**
  return max-id
else
  return null

function stf(w, y)
new-id := max{max-id, largest identifier in y}
case
  round $<$ diam(S): new-lead := unknown
  round $=$ diam(S) AND max-id $=$ my-id: new-lead := true
  round $=$ diam(S) AND max-id $>$ my-id: new-lead := false
return (my-id, new-id, new-lead, round + 1)

Figure 2 shows an execution of the FLOODMAX algorithm.

The properties of the algorithm are characterized in the following lemma. A complete analysis of this algorithm, including modifications to improve the communication complexity, is discussed in [Section 4.1 in 1].

Lemma 7 (Complexity upper bounds for the FLOODMAX algorithm) *For a network S containing a spanning tree, the* FLOODMAX *algorithm has communication complexity in $O(\text{diam}(S)|E_{cmm}|)$, time complexity equal to* diam(S), *and space complexity in $\Theta(1)$.*

A simplification of the FLOODMAX algorithm leads to the Le Lann–Chang–Roberts (LCR) algorithm for leader election in rings, see [Chapter 3.3 in 1], that we describe next.[1]. The LCR algorithm runs on a ring digraph and does not require the agents to know the diameter of the network.

[Informal description] At each communication round, if the agent receives from its in-neighbor a UID that is larger than the UIDs received earlier, then the agent records the received UID and forwards it to the out-neighbor during the following communication round. (Agents do not record the number of communication rounds.) When the agent with the maximum UID receives its own UID from a neighbor, it declares itself the leader.

The algorithm is formally stated as follows.

Synchronous Network: ring digraph

Distributed Algorithm: LCR

Alphabet: $\mathbb{A} = \{1, \ldots, n\} \cup \{\text{null}\}$

Processor State: w = (my-id, max-id, leader, snd-flag), where
my-id $\in \{1, \ldots, n\}$,
initially: my-id$^{[i]} = i$ for all i
max-id $\in \{1, \ldots, n\}$,
initially: max-id$^{[i]} = i$ for all i
leader $\in$ {true, false, unknown},
initially: leader$^{[i]}$ = unknown for all i
snd-flag $\in$ {true, false},
initially: snd-flag$^{[i]}$ = true for all i

function msg(w, i)
if {snd-flag = true} **then**
  return max-id
else
  return null

function stf(w, y)
case

[1]Note that the description of the LCR algorithm given here is slightly different from the classic one as presented in [57].

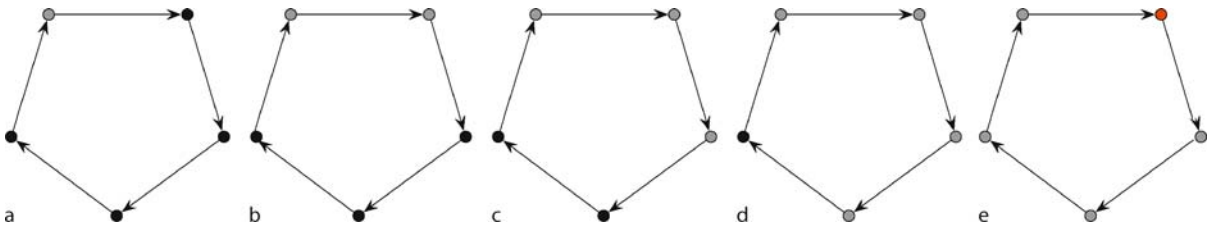

Robotic Networks, Distributed Algorithms for, Figure 3
Execution of the LCR algorithm. In the leftmost frame, the agent with the maximum UID is colored in *red*. After 5 communication rounds, this agent receives its own UID from its in-neighbor and declares itself the leader

```
    (y contains only null msgs) OR
        (largest identifier in y < my-id):
      new-id := max-id
      new-lead := leader
      new-snd-flag := false
      (largest identifier in y = my-id):
      new-id := max-id
      new-lead := true
      new-snd-flag := false
     (largest identifier in y > my-id):
     new-id := largest identifier in y
     new-lead := false
     new-snd-flag := true
  return (my-id, new-id, new-lead,
                    new-snd-flag)
```

Figure 3 shows an execution of the LCR algorithm. The properties of the LCR algorithm can be characterized as follows [57].

Lemma 8 (Complexity upper bounds for the LCR algorithm) *For a ring network S of order n, the* LCR *algorithm has communication complexity in $\Theta(n^2)$, time complexity equal to n, and space complexity in $\Theta(1)$.*

Averaging Algorithms

This section provides a brief introduction to a special class of distributed algorithms called averaging algorithms. The synchronous version of averaging algorithms can be modeled within the framework of synchronous networks. In an averaging algorithm, each processor updates its state by computing a weighted linear combination of the state of its neighbors. Computing linear combinations of the initial states of the processors is one the most basic computations that a network can implement. Averaging algorithms find application in optimization, distributed decision-making, e. g., collective synchronization, and have a long and rich history, see e. g., [29,30,45,65,97,98]. The richness comes from the vivid analogies with physical processes of diffusion, with Markov chain models, and the theory of positive matrices developed by Perron and Frobenius, see e. g., [19,20,44].

Averaging algorithms are defined by stochastic matrices. For completeness, let us recall some basic linear algebra definitions. A matrix $A \in \mathbb{R}^{n\times n}$ with entries a_{ij}, $i, j \in \{1, \dots, n\}$, is

(i) *nonnegative* (resp., *positive*) if all its entries are nonnegative (resp., positive);
(ii) *row-stochastic* (or *stochastic* for brevity) if it is nonnegative and $\sum_{j=1}^{n} a_{ij} = 1$, for all $i \in \{1, \dots, n\}$; in other words, A is row-stochastic if

$$A\mathbf{1}_n = \mathbf{1}_n;$$

where $\mathbf{1}_n = (1, \dots, 1)^T \in \mathbb{R}^n$.
(iii) *doubly stochastic* if it is row-stochastic and column-stochastic, where we say that A is *column-stochastic* if $\mathbf{1}_n^T A = \mathbf{1}_n^T$;
(iv) *irreducible* if, for any nontrivial partition $J \cup K$ of the index set $\{1, \dots, n\}$, there exists $j \in J$ and $k \in K$ such that $a_{jk} \neq 0$.

We are now ready to introduce the class of averaging algorithms. The *averaging algorithm* associated to a sequence of stochastic matrices $\{F(\ell) \mid \ell \in \mathbb{Z}_{\geq 0}\} \subset \mathbb{R}^{n\times n}$ is the discrete-time dynamical system

$$w(\ell + 1) = F(\ell) \cdot w(\ell)\,, \quad \ell \in \mathbb{Z}_{\geq 0}\,. \tag{1}$$

In the literature, averaging algorithms are also often referred to as agreement algorithms or as consensus algorithms.

Averaging algorithms are naturally associated with weighted digraphs, i. e., digraphs whose edges have weights. More precisely, a *weighted digraph* is a triplet $G = (V, E, A)$ where $V = \{v_1, \dots, v_n\}$ and E are a digraph and where $A \in \mathbb{R}_{\geq 0}^{n\times n}$ is a *weighted adjacency matrix* with the following properties: for $i, j \in \{1, \dots, n\}$, the entry $a_{ij} > 0$ if (v_i, v_j) is an edge of G, and $a_{ij} = 0$ otherwise. In other words, the scalars a_{ij}, for all $(v_i, v_j) \in E$,

are a set of weights for the edges of G. Note that edge set is uniquely determined by the weighted adjacency matrix and it can be therefore omitted. The *weighted out-degree matrix* $D_{\text{out}}(G)$ and the *weighted in-degree matrix* $D_{\text{in}}(G)$ are the diagonal matrices defined by

$$D_{\text{out}}(G) = \text{diag}\,(A\mathbf{1}_n)\,, \quad \text{and} \quad D_{\text{in}}(G) = \text{diag}\left(A^T\mathbf{1}_n\right).$$

The weighted digraph G is *weight-balanced* if $D_{\text{out}}(G) = D_{\text{in}}(G)$. Given a nonnegative $n \times n$ matrix A, its *associated weighted digraph* is the weighted digraph with nodes $\{1, \ldots, n\}$, and weighted adjacency matrix A. The unweighted version of this weighted digraph is called the *associated digraph*. The following statements can be proven:

(i) if A is stochastic, then its associated digraph has weighted out-degree matrix equal to I_n;
(ii) if A is doubly stochastic, then its associated weighted digraph is weight-balanced and additionally both in-degree and out-degree matrices are equal to I_n;
(iii) A is irreducible if and only if its associated weighted digraph is strongly connected.

Next, we characterize the convergence properties of averaging algorithms. Let us introduce a useful property of collections of stochastic matrices. Given $\alpha \in]0, 1]$, the set of *non-degenerate matrices with respect to* α consists of all stochastic matrices F with entries f_{ij}, for $i, j \in \{1, \ldots, n\}$, satisfying

$$f_{ii} \in [\alpha, 1]\,, \quad \text{and} \quad f_{ij} \in \{0\} \cup [\alpha, 1] \;\text{ for }\; j \neq i\,.$$

Additionally, the sequence of stochastic matrices $\{F(\ell) \mid \ell \in \mathbb{Z}_{\geq 0}\}$ is *non-degenerate* if these exists $\alpha \in]0, 1]$ such that $F(\ell)$ is non-degenerate with respect to α for all $\ell \in \mathbb{Z}_{\geq 0}$.

We now state the following convergence result from [65].

Theorem 9 (Convergence for time-dependent stochastic matrices) *Let $\{F(\ell) \mid \ell \in \mathbb{Z}_{\geq 0}\} \subset \mathbb{R}^{n\times n}$ be a non-degenerate sequence of stochastic matrices. For $\ell \in \mathbb{Z}_{\geq 0}$, let $G(\ell)$ be the unweighted digraph associated to $F(\ell)$. The following statements are equivalent:*

(i) the set $\text{diag}(\mathbb{R}^n)$ *is uniformly globally attractive for the associated averaging algorithm, that is, every evolution of the averaging algorithm at any time ℓ_0, approaches the set* $\text{diag}(\mathbb{R}^n)$ *in the following time-uniform manner:*

for all $\ell_0 \in \mathbb{Z}_{\geq 0}$, for all $w_0 \in \mathbb{R}^n$, and for all neighborhoods W of $\text{diag}(\mathbb{R}^n)$*, there exists a single $\tau_0 \in \mathbb{Z}_{\geq 0}$ such that the evolution $w : [\ell_0, +\infty[\to \mathbb{R}^n$ defined by $w(\ell_0) = w_0$, takes value in W for all times $\ell \geq \ell_0 + \tau_0$.*

(ii) there exists a duration $\delta \in \mathbb{N}$ such that, for all $\ell \in \mathbb{Z}_{\geq 0}$, the digraph

$$G(\ell + 1) \cup \cdots \cup G(\ell + \delta)$$

contains a globally reachable vertex.

Distributed Algorithms for Robotic Networks

This section describes models and algorithms for groups of robots that process information, sense, communicate, and move. We refer to such systems as robotic networks. In this section we review and survey a few modeling and algorithmic topics in robotic coordination; earlier versions of this material were originally presented in [15,59,60,61].

The section is organized as follows. First, we present the physical components of a network, that is, the mobile robots and the communication service connecting them. We then present the notion of control and communication law, and how a law is executed by a robotic network. We then discuss complexity notions for robotic networks. As an example of these notions, we introduce a simple law, called the agree-and-pursue law, which combines ideas from leader election algorithms and from cyclic pursuit (i. e., a game in which robots chase each other in a circular environment). We then consider in some detail algorithms for two basic motion coordination tasks, namely aggregation and deployment. We briefly formalize these problems and provide some basic algorithms for these two tasks.

Robotic Networks and Complexity

The global behavior of a robotic network arises from the combination of the local actions taken by its members. Each robot in the network can perform a few basic tasks such as sensing, communicating, processing information, and moving according to it. The many ways in which these capabilities can be integrated make a robotic network a versatile and, at the same time, complex system. The following robotic network model provides the framework to formalize, analyze, and compare distinct distributed behaviors.

We consider *uniform networks of robotic agents* defined by a tuple $S = (I, \mathcal{R}, E_{\text{cmm}})$ consisting of a set of unique identifiers $I = \{1, \ldots, n\}$, a collection of control systems $\mathcal{R} = \{R^{[i]}\}_{i\in I}$, with $R^{[i]} = (X, U, X_0, f)$, and a map E_{cmm} from X^n to the subsets of $I \times I$ called the *communication edge map*. Here, (X, U, X_0, f) is a control system with state space $X \subset \mathbb{R}^d$, input space U, set of allowable initial states $X_0 \subset X$, and system dynamics $f : X \times U \to X$. An edge between two identifiers in E_{cmm}

implies the ability of the corresponding two robots to exchange messages. A *control and communication law* for S consists of the sets:

(i) $\mathbb{A}$, called the *communication language*, whose elements are called *messages*;
(ii) W, set of values of some *processor variables* $w^{[i]} \in W$, $i \in I$, and $W_0 \subseteq W$, subset of *allowable initial values*. These sets correspond to the capability of robots to allocate additional variables and store sensor or communication data;

and the maps:

(iii) $\mathrm{msg} : X \times W \times I \to \mathbb{A}$, called *message-generation function*;
(iv) $\mathrm{stf} : X \times W \times \mathbb{A}^n \to W$, called *state-transition function*;
(v) $\mathrm{ctl} : X \times W \times \mathbb{A}^n \to U$, called *control function*.

To implement a control and communication law each robot performs the following sequence or cycle of actions. At each instant $\ell \in \mathbb{Z}_{\geq 0}$, each robot i communicates to each robot j such that (i, j) belongs to $E_{\mathrm{cmm}}(x^{[1]}, \ldots, x^{[n]})$. Each robot i sends a message computed by applying the message-generation function to the current values of $x^{[i]}$ and $w^{[i]}$. After a negligible period of time, robot i resets the value of its logic variables $w^{[i]}$ by applying the state-transition function to the current value of $w^{[i]}$, and to the messages $y^{[i]}(\ell)$ received at ℓ. Between communication instants, i. e., for $t \in [\ell, \ell + 1)$, robot i applies a control action computed by applying the control function to its state at the last sample time $x^{[i]}(\ell)$, the current value of $w^{[i]}$, and to the messages $y^{[i]}(\ell)$ received at ℓ.

Remark 10 (Algorithm properties and congestion models)

(i) In our present definition, all robots are identical and implement the same algorithm; in this sense the control and communication law is called *uniform* (or anonymous). If $W = W_0 = \emptyset$, then the control and communication law is *static* (or memoryless) and no state-transition function is defined. It is also possible for a law to be *time-independent* if the three relevant maps do not depend on time. Finally, let us also remark that this is a synchronous model in which all robots share a common clock.
(ii) Communication and physical congestion affect the performance of robotic networks. These effects can be modeled by characterizing how the network parameters vary as the number of robots becomes larger. For example, in an ad hoc networks with n uniformly randomly placed nodes, it is known [42] that the maximum-throughput communication range $r(n)$ of each node decreases as the density of nodes increases; in d dimensions the appropriate scaling law is $r(n) \in \Theta((\log(n)/n)^{1/d})$. As a second example, it is reasonable to assume that, as the number of robots increase, so should the area available for their motion. An alternative convenient approach is the one taken by [88], where robots' safety zones decrease with decreasing robots' speed. This suggests that, in a fixed environment, individual nodes of a large ensemble have to move at a speed decreasing with n, and in particular, at a speed proportional to $n^{-1/d}$.

Next, we establish the notion of coordination task and of task achievement by a robotic network. Let S be a robotic network and let $\mathcal{W}$ be a set. A *coordination task* for S is a map $\mathcal{T} : X^n \times \mathcal{W}^n \to \{\mathtt{true}, \mathtt{false}\}$. If $\mathcal{W}$ is a singleton, then the coordination task is said to be *static* and can be described by a map $\mathcal{T} : X^n \to \{\mathtt{true}, \mathtt{false}\}$. Additionally, let CC a control and communication law for S.

(i) The law CC is *compatible* with the task $\mathcal{T} : X^n \times \mathcal{W}^n \to \{\mathtt{true}, \mathtt{false}\}$ if its processor state take values in $\mathcal{W}$, that is, if $W^{[i]} = \mathcal{W}$, for all $i \in I$.
(ii) The law CC *achieves* the task $\mathcal{T}$ if it is compatible with it and if, for all initial conditions $x_0^{[i]} \in X_0$ and $w_0^{[i]} \in W_0$, $i \in I$, there exists $T \in \mathbb{Z}_{\geq 0}$ such that the network evolution $\ell \mapsto (x(\ell), w(\ell))$ has the property that $\mathcal{T}(x(\ell), w(\ell)) = \mathtt{true}$ for all $\ell \geq T$.

In control-theoretic terms, achieving a task means establishing a convergence or stability result. Beside this key objective, one might be interested in efficiency as measured by required communication service, required control energy or by speed of completion. We focus on the latter notion.

(i) The *(worst-case) time complexity to achieve* $\mathcal{T}$ *with* CC *from* $(x_0, w_0) \in X_0^n \times W_0^n$ is

$$\mathrm{TC}(\mathcal{T}, CC, x_0, w_0) = \inf\{\ell \mid \mathcal{T}(x(k), w(k)) = \mathtt{true}\,, \text{ for all } k \geq \ell\}\,,$$

where $\ell \mapsto (x(\ell), w(\ell))$ is the evolution of (S, CC) from the initial condition (x_0, w_0);
(ii) The *(worst-case) time complexity to achieve* $\mathcal{T}$ *with* CC is

$$\mathrm{TC}(\mathcal{T}, CC) = \sup\left\{\mathrm{TC}(\mathcal{T}, CC, x_0, w_0) \mid (x_0, w_0) \in X_0^n \times W_0^n\right\}.$$

Some ideas on how to define meaningful notions of space and communication complexity are discussed in [59]. In the following discussion, we describe three coordination algorithms, which have been cast into this modeling framework and whose time complexity properties have been analyzed.

Agree-and-Pursue Algorithm

We begin our list of distributed algorithms with a simple law that is related to leader election algorithms, see Sect. "Leader Election", and to cyclic pursuit algorithms as studied in the control literature. Despite the apparent simplicity, this example is remarkable in that it combines a leader election task (in the processor states) with a uniform robotic deployment task (in the physical state), arguably two of the most basic tasks in distributed algorithms and cooperative control, respectively.

We consider n robots $\{\theta^{[1]}, \dots, \theta^{[n]}\}$ in $\mathbb{S}^1$, moving along on the unit circle with angular velocity equal to the control input. Each robot is described by the tuple $(\mathbb{S}^1, [-u_{\max}, u_{\max}], \mathbb{S}^1, (0, \boldsymbol{e}))$, where $\boldsymbol{e}$ is the vector field on $\mathbb{S}^1$ describing unit-speed counterclockwise rotation. We assume that each robot can sense its own position and can communicate to any other robot within distance r along the circle. These data define the uniform robotic network S_{circle}.

> *[Informal description]* The processor state consists of `dir` (the robot's direction of motion) taking values in $\{\mathtt{c}, \mathtt{cc}\}$ (meaning clockwise and counterclockwise) and `max-id` (the largest UID received by the robot, initially set to the robot's UID) taking values in I. At each communication round, each robot transmits its position and its processor state. Among the messages received from the robots moving towards its position, each robot picks the message with the largest value of `max-id`. If this value is larger than its own value, the agent resets its processor state with the selected message. Between communication rounds, each robot moves in the clockwise or counterclockwise direction depending on whether its processor state `dir` is `c` or `cc`. Each robot moves k_{prop} times the distance to the immediately next neighbor in the chosen direction, or, if no neighbors are detected, k_{prop} times the communication range r.

For this network and this law there are two tasks of interest. First, we define the *direction agreement task* $\mathcal{T}_{\mathtt{dir}} : (\mathbb{S}^1)^n \times W^n \to \{\mathtt{true}, \mathtt{false}\}$ by

$$\mathcal{T}_{\mathtt{dir}}(\theta, w) = \begin{cases} \mathtt{true}, & \text{if } \mathtt{dir}^{[1]} = \cdots = \mathtt{dir}^{[n]}, \\ \mathtt{false}, & \text{otherwise}, \end{cases}$$

where $\theta = (\theta^{[1]}, \dots, \theta^{[n]})$, $w = (w^{[1]}, \dots, w^{[n]})$, and $w^{[i]} = (\mathtt{dir}^{[i]}, \mathtt{max\text{-}id}^{[i]})$, for $i \in I$. Furthermore, for $\varepsilon > 0$, we define the static *equidistance task* $\mathcal{T}_{\varepsilon\text{-eqdstnc}} : (\mathbb{S}^1)^n \to \{\mathtt{true}, \mathtt{false}\}$ to be `true` if and only if

$$\left| \min_{j \neq i} \operatorname{dist}_{\mathtt{c}}\left(\theta^{[i]}, \theta^{[j]}\right) - \min_{j \neq i} \operatorname{dist}_{\mathtt{cc}}\left(\theta^{[i]}, \theta^{[j]}\right) \right| < \varepsilon, \quad \text{for all } i \in I.$$

In other words, $\mathcal{T}_{\varepsilon\text{-eqdstnc}}$ is true when, for every agent, the distance to the closest clockwise neighbor and to the closest counterclockwise neighbor are approximately equal.

An implementation of this control and communication law is shown in Fig. 4. As parameters we select $n = 45$, $r = 2\pi/40$, $u_{\max} = 1/4$ and $k_{\text{prop}} = 7/16$. Along the evolution, all robots agree upon a common direction of motion and, after suitable time, they reach a uniform distribution. A careful analysis based on invariance properties and Lyapunov functions allows us to establish that, under appropriate conditions, indeed both tasks are achieved by the agree-and-pursue law [59].

Theorem 11 (Time complexity of agree-and-pursue law) *For $k_{prop} \in]0, \frac{1}{2}[$, in the limit as $n \to +\infty$ and $\varepsilon \to 0^+$, the network S_{circle} with $u_{max}(n) \geq k_{prop} r(n)$, the law $CC_{\text{AGREE \& PURSUE}}$, and the tasks $\mathcal{T}_{dir}$ and $\mathcal{T}_{\varepsilon\text{-eqdstnc}}$ together satisfy:*

(i) $\mathrm{TC}(\mathcal{T}_{dir}, CC_{\text{AGREE \& PURSUE}}) \in \Theta(r(n)^{-1})$;

(ii) if $\delta(n)$ is lower bounded by a positive constant as $n \to +\infty$, then

$$\mathrm{TC}(\mathcal{T}_{\varepsilon\text{-eqdstnc}}, CC_{\text{AGREE \& PURSUE}}) \in \Omega(n^2 \log(n\varepsilon)^{-1}),$$
$$\mathrm{TC}(\mathcal{T}_{\varepsilon\text{-eqdstnc}}, CC_{\text{AGREE \& PURSUE}}) \in O(n^2 \log(n\varepsilon^{-1})).$$

If $\delta(n)$ is upper bounded by a negative constant, then $CC_{\text{AGREE \& PURSUE}}$ does not achieve $\mathcal{T}_{\varepsilon\text{-eqdstnc}}$ in general.

Finally we compare these results with the complexity result known for the leader election problem.

Remark 12 (Comparison with leader election) Let us compare the agree-and-pursue control and communication law with the classical Le Lann–Chang–Roberts (LCR) algorithm for leader election discussed in Sect. "Leader Election". The leader election task consists of electing a unique agent among all robots in the network; it is therefore different from, but closely related to, the coordination

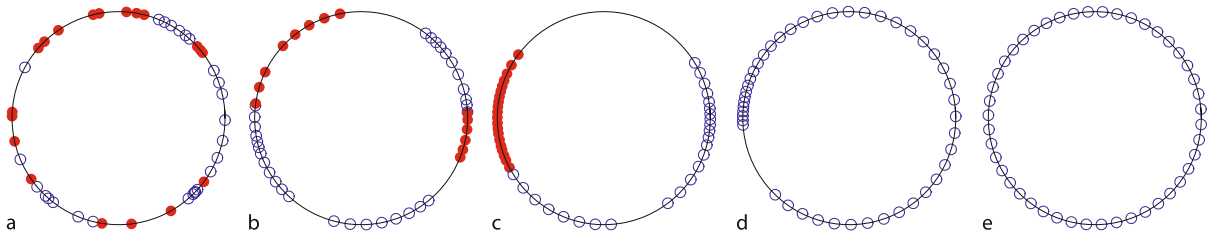

Robotic Networks, Distributed Algorithms for, Figure 4
The AGREE & PURSUE law. *Disks* and *circles* correspond to robots moving counterclockwise and clockwise, respectively. The initial positions and the initial directions of motion are randomly generated. The five pictures depict the network state at times 0, 9, 20, 100, and 800

task $\mathcal{T}_{\mathtt{dir}}$. The LCR algorithm operates on a static network with the ring communication topology, and achieves leader election with time and total communication complexity, respectively, $\Theta(n)$ and $\Theta(n^2)$. The agree-and-pursue law operates on a robotic network with the $r(n)$-disk communication topology, and achieves $\mathcal{T}_{\mathtt{dir}}$ with time and total communication complexity, respectively, $\Theta(r(n)^{-1})$ and $O(n^2 r(n)^{-1})$. If wireless communication congestion is modeled by $r(n)$ of order $1/n$ as in Remark 10, then the two algorithms have identical time complexity and the LCR algorithm has better communication complexity. Note that computations on a possibly disconnected, dynamic network are more complex than on a static ring topology.

Aggregation Algorithms

The rendezvous objective (also referred to as the gathering problem) is to achieve agreement over the location of the robots, that is, to steer each agent to a common location. An early reference on this problem is [2]; more recent references include [26,32,52,53]. We consider two scenarios which differ in the robots' communication capabilities and the environment in which the robots move. First [26], we consider the problem of rendezvous for robots equipped with *range-limited communication* in obstacle-free environments. In this case, each robot is capable of sensing its position in the Euclidean space $\mathbb{R}^d$ and can communicate it to any other robot within a given distance r. This communication service is modeled by the r-disk graph, in which two robots are neighbors if and only if their Euclidean distance is less than or equal to r. Second [36], we consider *visually-guided robots*. Here the robots are assumed to move in a nonconvex simple polygonal environment Q. Each robot can sense, within line of sight, any other robot as well as the distance to the boundary of the environment. The relationship between the robots can be characterized by the so-called visibility graph: two robots are neighbors if and only if they are mutually visible to each other.

In both scenarios, the rendezvous problem cannot be solved with distributed information unless the robots' initial positions form a connected communication graph. Arguably, a good property of any rendezvous algorithm is that of maintaining connectivity between robots. This connectivity-maintenance objective is interesting on its own. It turns out that this objective can be achieved through local constraints on the robots' motion. Motion constraint sets that maintain connectivity are designed in [2,36] by exploiting the geometric properties of disk and visibility graphs.

These discussions lead to the following algorithm that solves the rendezvous problems for both communication scenarios. The robots execute what is known as the *Circumcenter Algorithm*; here is an informal description. Each robot iteratively performs the following tasks:

```
1: acquire neighbors' positions
2: compute connectivity constraint set
3: move toward the circumcenter of
   the point set comprised of its
   neighbors and of itself, while
   remaining inside the connectivity
   constraint set.
```

One can prove that, under technical conditions, the algorithm does achieve the rendezvous task in both scenarios; see [26,36]. Additionally, when $d = 1$, it can be shown that the time complexity of this algorithm is $\Theta(n)$; see [60].

Deployment Algorithms

The problem of deploying a group of robots over a given region of interest can be tackled with the following simple heuristic. Each robot iteratively performs the following tasks:

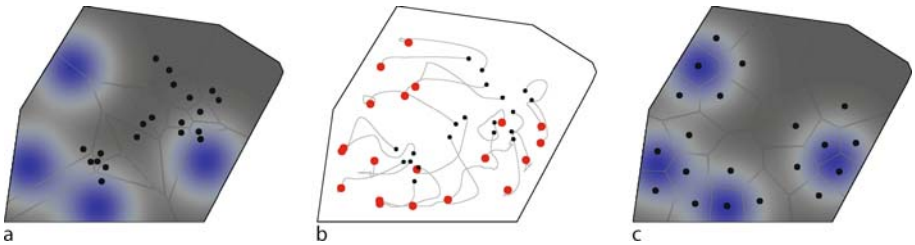

Robotic Networks, Distributed Algorithms for, Figure 5
Deployment algorithm for the area-coverage problem. Each of the 20 robots moves toward the centroid of its Voronoi cell. This strategy corresponds to the network following the gradient of $\mathcal{H}_{\text{ave}}$. Areas of the polygon with greater importance are colored darker. Figures a and c show, respectively, the initial and final locations, with the corresponding Voronoi partitions. Figure b illustrates the gradient descent flow

```
1: acquire neighbors' positions
2: compute own dominance region
3: move towards the center of own
   dominance region
```

This short description can be made accurate by specifying what notions of dominance region and of center are to be adopted. In what follows we mention two examples and refer to [24,25,27,37] for more details.

First, we consider the *area-coverage deployment problem* in a convex environment Q. The objective is to maximize the area within close range of the mobile nodes. This models a scenario in which the nodes are equipped with some sensors that take measurements of some physical quantity in the environment, e. g., temperature or concentration. Assume that certain regions in the environment are more important than others and describe this by a density function ϕ. This problems leads to the coverage performance metric

$$\begin{aligned}\mathcal{H}_{\text{ave}}(p_1,\ldots,p_n) &= \int_Q \min_{i\in\{1,\ldots,n\}} f(\|q-p_i\|)\phi(q)\mathrm{d}q \\ &= \sum_{i=1}^{n}\int_{V_i} f(\|q-p_i\|)\phi(q)\mathrm{d}q\,.\end{aligned}$$

Here p_i is the position of the ith node, f measures the performance of an individual sensor, and $\{V_1,\ldots,V_n\}$ is the Voronoi partition of the environment Q generated by the positions $\{p_1,\ldots,p_n\}$. If we assume that each node obeys a first-order dynamical behavior, then a simple gradient scheme can be easily implemented in a spatially-distributed manner. Figure 5 shows an implementation of this gradient scheme. Following the gradient of $\mathcal{H}_{\text{ave}}$ corresponds, in the algorithm described above, to defining (1) the dominance regions to be the Voronoi cells generated by the robots, and (2) the center of a region to be the centroid of the region (if $f(x)=x^2$). Because the closed-loop system is a gradient flow for the cost function, performance is locally, continuously optimized. As a special case, when the environment is a segment and $\phi=1$, the time complexity of the algorithm can be shown to be $O(n^3\log(n\varepsilon^{-1}))$, where ε is an accuracy threshold below which we consider the task accomplished.

Second, we consider the problem of deploying *to maximize the likelihood of detecting a source*. For example, consider devices equipped with acoustic sensors attempting to detect a sound-source (or, similarly, antennas detecting RF signals, or chemical sensors localizing a pollutant source). For a variety of criteria, when the source emits a known signal and the noise is Gaussian, we know that the optimal detection algorithm involves a matched filter, that detection performance is a function of signal-to-noise-ratio, and, in turn, that signal-to-noise ratio is inversely proportional to the sensor-source distance. In this case, the appropriate cost function is

$$\begin{aligned}\mathcal{H}_{\text{worst}}(p_1,\ldots,p_n) &= \max_{q\in Q}\min_{i\in\{1,\ldots,n\}} f(\|q-p_i\|) \\ &= \max_{q\in V_i} f(\|q-p_i\|)\,,\end{aligned}$$

and a greedy motion coordination algorithm is for each node to move toward the circumcenter of its Voronoi cell. A detailed analysis [24] shows that the detection likelihood is inversely proportional to the circumradius of each node's Voronoi cell, and that, if the nodes follow the algorithm described above, then the detection likelihood increases monotonically as a function of time.

Bibliographical Notes

In this section we present a necessarily incomplete discussion of some relevant literature that we have not yet mentioned in the previous sections.

First, we review some literature on emergent and self-organized swarming behaviors in biological groups with distributed agent-to-agent interactions. Interesting

dynamical systems arise in biological networks at multiple levels of resolution, all the way from interactions among molecules and cells [62] to the behavioral ecology of animal groups [66]. Flocks of birds and schools of fish can travel in formation and act as one unit (see [72]), allowing these animals to defend themselves against predators and protect their territories. Wildebeest and other animals exhibit complex collective behaviors when migrating, such as obstacle avoiding, leader election, and formation keeping (see [41,89]). Certain foraging behaviors include individual animals partitioning their environment into nonoverlapping zones (see [7]). Honey bees [85], gorillas [93], and whitefaced capuchins [13] exhibit synchronized group activities such as initiation of motion and change of travel direction. These remarkable dynamic capabilities are achieved apparently without following a group leader; see [7,13,41,66,72,85,93] for specific examples of animal species and [21,28] for general studies.

With regards to distributed motion coordination algorithms, much progress has been made on pattern formation [9,46,86,94], flocking [67,96], self-assembly [49], swarm aggregation [38], gradient climbing [75], cyclic pursuit [14,58,90], vehicle routing [56,87], and connectivity maintenance problems [83,103].

Much research has been devoted to distributed task allocation problems. The work in [39] proposes a taxonomy of task allocation problems. In papers such as [1,40,63,84], advanced heuristic methods are developed, and their effectiveness is demonstrated through simulation or real world implementation. Distributed auction algorithms are discussed in [18,63] building on the classic works in [10,11]. A distributed MILP solver is proposed in [1]. A spatially distributed receding-horizon scheme is proposed in [33,73]. There has also been prior work on target assignment problems [6,91,102]. Target allocation for vehicles with nonholonomic constraints is studied in [76,81,82].

References with a focus on robotic networks include the survey in [99], the text [5] on behavior-based robotics, and the recent special issue [4] of the IEEE Transaction on Robotics and Automation. An important contribution towards a network model of mobile interacting robots is introduced in [94]. This model consists of a group of identical "distributed anonymous mobile robots" characterized as follows: no explicit communication takes place between them, and at each time instant of an "activation schedule," each robot senses the relative position of all other robots and moves according to a pre-specified algorithm. Communication complexity for control and communication algorithms A related model is presented in [32], where as few capabilities as possible are assumed on the agents, with the objective of understanding the limitations of multi-agent networks. A brief survey of models, algorithms, and the need for appropriate complexity notions is presented in [79]. Recently, a notion of communication complexity for control and communication algorithms in multi-robot systems is analyzed in [48], see also [50].

Finally, with regards to linear distributed algorithms we mention the following references, on top of the ones discussed in Sect. "Averaging Algorithms". Various results are available on continuous-time consensus algorithms [34,54,55,64,70,77], consensus over random networks [43,95,100], consensus algorithms for general functions [8,23], connections with the heat equation and partial difference equation [31], convergence in time-delayed and asynchronous settings [3,12], quantized consensus problems [47,80], applications to distributed signal processing [69,92,101], characterization of the convergence rates and time complexity [16,17,51,71]. Finally, two recent surveys are [68,78].

Future Directions

Robotic networks incorporate numerous subsystems. Their design is challenging because they integrate heterogeneous hardware and software components. Additionally, the operation of robotic networks is subject to computing, energy, cost, and safety constraints. The interaction with the physical world and the uncertain response from other members of the network are also integral parts to consider in the management of robotic networks. Traditional centralized approaches are not valid to satisfy the scalability requirements of these systems. Thus, in order to successfully deploy robotic and communication networks, it is necessary to expand our present knowledge about how to integrate and efficiently control them. The present chapter has offered a glimpse into these problems. We have presented verification and complexity tools to evaluate the cost of cooperative strategies that achieve a variety of tasks.

Networks of robotic agents are an example of the class of complex, networked systems which pervades our world. Understanding how to design robotic swarms requires the development of new fundamental theories that can explain the behavior of general networks evolving with time. Some of the desired properties of these systems are robustness, ease of control, predictability with time, guaranteed performance, and quality of service. Self-organization and distributed management would also allow for the minimal supervision necessary for scalability. However, devising systems that meet all these criteria is not an easy task. A small deviation by an agent or a change in certain parameters may produce a dramatic change in the

overall network behavior. Relevant questions that pertain these aspects are currently being approached from a range of disciplines such as systems and control, operations research, random graph theory, statistical physics, and game theory over networks. Future work will undoubtedly lead to a cross-fertilization of these and other areas that will help design efficient robotic sensor networks.

Acknowledgments

This material is based upon work supported in part by NSF CAREER Award CMS-0643679, NSF CAREER Award ECS-0546871, and AFOSR MURI Award FA9550-07-1-0528.

Bibliography

1. Alighanbari M, How JP (2006) Robust decentralized task assignment for cooperative UAVs. In: AIAA conference on guidance, navigation and control, Keystone, CO, August 2006
2. Ando H, Oasa Y, Suzuki I, Yamashita M (1999) Distributed memoryless point convergence algorithm for mobile robots with limited visibility. Trans IEEE Robotics Autom 15(5):818–828
3. Angeli D, Bliman PA (2006) Stability of leaderless discrete-time multi-agent systems. Math Control Signal Syst 18(4):293–322
4. Arai T, Pagello E, Parker LE (2002) Guest editorial: Advances in multirobot systems. Trans IEEE Robotics Autom 18(5):655–661
5. Arkin RC (1998) Behavior-Based Robotics. MIT Press, Cambridge, MA
6. Arslan G, Marden JR, Shamma JS (2007) Autonomous vehicle-target assignment: A game theoretic formulation. J ASME Dyn Syst Meas Control 129(5):584–596
7. Barlow GW (1974) Hexagonal territories. Anim Behav 22:876–878
8. Bauso D, Giarré L, Pesenti R (2006) Nonlinear protocols for optimal distributed consensus in networks of dynamic agents. Syst Control Lett 55(11):918–928
9. Belta C, Kumar V (2004) Abstraction and control for groups of robots. Trans IEEE Robotics 20(5):865–875
10. Bertsekas DP, Castañón DA (1991) Parallel synchronous and asynchronous implementations of the auction algorithm. Parallel Comput 17:707–732
11. Bertsekas DP, Castañón DA (1993) Parallel primal-dual methods for the minimum cost flow problem. Comput Optim Appl 2(4):317–336
12. Blondel VD, Hendrickx JM, Olshevsky A, Tsitsiklis JN (2005) Convergence in multiagent coordination, consensus, and flocking. In: IEEE conference on decision and control and european control conference, Seville, Spain, December 2005, pp 2996–3000
13. Boinski S, Campbell AF (1995) Use of trill vocalizations to coordinate troop movement among whitefaced capuchins – a 2nd field-test. Behaviour 132:875–901
14. Bruckstein AM, Cohen N, Efrat A (1991) Ants, crickets, and frogs in cyclic pursuit. Technical Report CIS 9105, Center for Intelligent Systems, Technion, Haifa, Israel. http://www.cs.technion.ac.il/tech-reports
15. Bullo F (2006) Notes on multi-agent motion coordination: Models and algorithms. In: Antsaklis PJ, Tabuada P (eds) Network Embedded Sensing and Control, Proceedings of NESC'05 Worskhop. Lecture Notes in Control and Information Sciences, vol 331. Springer, New York, pp 3–8
16. Cao M, Morse AS, Anderson BDO (2008) Reaching a consensus in a dynamically changing environment – convergence rates, measurement delays and asynchronous events. J SIAM Control Optim 47(2):601–623
17. Carli R, Fagnani F, Speranzon A, Zampieri S (2008) Communication constraints in the average consensus problem. Automatica 44(3):671–684
18. Castañón DA, Wu C (2003) Distributed algorithms for dynamic reassignment. In: IEEE conference on decision and control. Maui, HI, December 2003, pp 13–18
19. Chatterjee S, Seneta E (1977) Towards consensus: Some convergence theorems on repeated averaging. J Appl Probab 14(1):89–97
20. Cogburn R (1984) The ergodic theory of Markov chains in random environments. Z Wahrscheinlichkeitstheorie Geb 66(1):109–128
21. Conradt L, Roper TJ (2003) Group decision-making in animals. Nature 421(6919):155–158
22. Cormen TH, Leiserson CE, Rivest RL, Stein C (2001) Introduction to Algorithms, 2nd edn. Press MIT, Cambridge, MA
23. Cortés J (2008) Distributed algorithms for reaching consensus on general functions. Automatica 44(3):726–737
24. Cortés J, Bullo F (2005) Coordination and geometric optimization via distributed dynamical systems. J SIAM Control Optim 44(5):1543–1574
25. Cortés J, Martínez S, Bullo F (2005) Spatially-distributed coverage optimization and control with limited-range interactions. Control ESAIM Optim Calc Var 11:691–719
26. Cortés J, Martínez S, Bullo F (2006) Robust rendezvous for mobile autonomous agents via proximity graphs in arbitrary dimensions. Trans IEEE Autom Control 51(8):1289–1298
27. Cortés J, Martínez S, Karatas T, Bullo F (2004) Coverage control for mobile sensing networks. IEEE Trans Autom Control 20(2):243–255
28. Couzin ID, Krause J, Franks NR, Levin SA (2005) Effective leadership and decision-making in animal groups on the move. Nature 433:513–51
29. Cybenko G (1989) Dynamic load balancing for distributed memory multiprocessors. J Parallel Distrib Comput 7:279–301
30. DeGroot MH (1974) Reaching a consensus. J Am Stat Assoc 69(345):118–121
31. Ferrari-Trecate G, Buffa A, Gati M (2006) Analysis of coordination in multi-agent systems through partial difference equations. IEEE Trans Autom Control 51(6):1058–1063
32. Flocchini P, Prencipe G, Santoro N, Widmayer P (2005) Gathering of asynchronous oblivious robots with limited visibility. Theor Comput Sci 337(1–3):147–168
33. Frazzoli E, Bullo F (2004) Decentralized algorithms for vehicle routing in a stochastic time-varying environment. In: IEEE conference on decision and control. Paradise Island, Bahamas, December 2004. pp 3357–3363
34. Freeman RA, Yang P, Lynch KM (2006) Stability and convergence properties of dynamic average consensus estimators.

In: IEEE conference on decision and control. San Diego, CA, December 2006, pp 398–403
35. Gallager RG (1968) Information Theory and Reliable Communication. Wiley, New York
36. Ganguli A, Cortés J, Bullo F (2005) On rendezvous for visually-guided agents in a nonconvex polygon. In: IEEE conference on decision and control and european control conference, Seville, December 2005, pp 5686–5691
37. Ganguli A, Cortés J, Bullo F (2006) Distributed deployment of asynchronous guards in art galleries. In: American control conference, Minneapolis, June 2006, pp 1416–1421
38. Gazi V, Passino KM (2003) Stability analysis of swarms. IEEE Trans Autom Control 48(4):692–697
39. Gerkey BP, Mataric MJ (2004) A formal analysis and taxonomy of task allocation in multi-robot systems. Int J Robotics Res 23(9):939–954
40. Godwin MF, Spry S, Hedrick JK (2006) Distributed collaboration with limited communication using mission state estimates. In: American control conference. Minneapolis, MN, June 2006, pp 2040–2046
41. Gueron S, Levin SA (1993) Self-organization of front patterns in large wildebeest herds. J Theor Bio 165:541–552
42. Gupta P, Kumar PR (2000) The capacity of wireless networks. Trans IEEE Inf Theory 46(2):388–404
43. Hatano Y, Mesbahi M (2005) Agreement over random networks. Trans IEEE Autom Control 50:1867–1872
44. Horn RA, Johnson CR (1985) Matrix analysis. Cambridge University Press, Cambridge, UK
45. Jadbabaie A, Lin J, Morse AS (2003) Coordination of groups of mobile autonomous agents using nearest neighbor rules. Trans IEEE Autom Control 48(6):988–1001
46. Justh EW, Krishnaprasad PS (2004) Equilibria and steering laws for planar formations. Syst Control Lett 52(1):25–38
47. Kashyap A, Başar T, Srikant R (2007) Quantized consensus. Automatica 43(7):1192–1203
48. Klavins E (2003) Communication complexity of multi-robot systems. In: Boissonnat JD, Burdick JW, Goldberg K, Hutchinson S (eds) Algorithmic foundations of robotics V, vol 7. Springer tracts in advanced robotics. Springer, Berlin
49. Klavins E, Ghrist R, Lipsky D (2006) A grammatical approach to self-organizing robotic systems. Trans IEEE Autom Control 51(6):949–962
50. Klavins E, Murray RM (2004) Distributed algorithms for cooperative control. Pervasive IEEE Comput 3(1):56–65
51. Landau HJ, Odlyzko AM (1981) Bounds for eigenvalues of certain stochastic matrices. Linear Algebra Appl 38:5–15
52. Lin J, Morse AS, Anderson BDO (2004) The multi-agent rendezvous problem: An extended summary. In: Kumar V, Leonard NE, Morse AS (eds) Proceedings of the 2003 block island workshop on cooperative control, ser. Lecture notes in control and information sciences, vol 309. Springer, New York, pp 257–282
53. Lin Z, Broucke M, Francis B (2004) Local control strategies for groups of mobile autonomous agents. IEEE Trans Autom Control 49(4):622–629
54. Lin Z, Francis B, Maggiore M (2005) Necessary and sufficient graphical conditions for formation control of unicycles. Trans IEEE Autom Control 50(1):121–127
55. Lin Z, Francis B, Maggiore M (2007) State agreement for continuous-time coupled nonlinear systems. J SIAM Control Optim 46(1):288–307
56. Lumelsky VJ, Harinarayan KR (1997) Decentralized motion planning for multiple mobile robots: the cocktail party model. Auton Robots 4(1):121–135
57. Lynch NA (1997) Distributed algorithms. Morgan Kaufmann Publishers, San Mateo, CA
58. Marshall JA, Broucke ME, Francis BA (2004) Formations of vehicles in cyclic pursuit. Trans IEEE Autom Control 49(11):1963–1974
59. Martínez S, Bullo F, Cortés J, Frazzoli E (2007) On synchronous robotic networks – Part I: Models, tasks and complexity. Trans IEEE Autom Control 52(12):2199–2213
60. Martinez S, Bullo F, Cortés J, Frazzoli E (2007) On synchronous robotic networks – Part II: Time complexity of rendezvous and deployment algorithms. Trans IEEE Autom Control 52(12):2214–2226
61. Martínez S, Cortés J, Bullo F (2007) Motion coordination with distributed information. Control IEEE Syst Mag 27(4):75–88
62. Miller MB, Bassler BL (2001) Quorum sensing in bacteria. Annu Rev Microbiol 55:165–199
63. Moore BJ, Passino KM (2007) Distributed task assignment for mobile agents. Trans IEEE Autom Control 52(4):749–753
64. Moreau L (2004) Stability of continuous-time distributed consensus algorithms, Preprint. Available at http://xxx.arxiv.org/math.OC/0409010
65. Moreau L (2005) Stability of multiagent systems with time-dependent communication links. Trans IEEE Autom Control 50(2):169–182
66. Okubo A (1986) Dynamical aspects of animal grouping: swarms, schools, flocks and herds. Advances in Biophysics 22:1–94
67. Olfati-Saber R (2006) Flocking for multi-agent dynamic systems: Algorithms and theory. Trans IEEE Autom Control 51(3):401–420
68. Olfati-Saber R, Fax JA, Murray RM (2007) Consensus and cooperation in networked multi-agent systems. Proc IEEE 95(1):215–233
69. Olfati-Saber R, Franco E, Frazzoli E, Shamma JS (2006) Belief consensus and distributed hypothesis testing in sensor networks. In: Antsaklis PJ, Tabuada P (eds) Network embedded sensing and control. Proceedings of NESC'05 worskhop. Lecture notes in control and information sciences, vol 331. Springer, New York, pp 169–182
70. Olfati-Saber R, Murray RM (2004) Consensus problems in networks of agents with switching topology and time-delays. Trans IEEE Autom Control 49(9):1520–1533
71. Olshevsky A, Tsitsiklis JN (2007) Convergence speed in distributed consensus and averaging. J SIAM Control Optim
72. Parrish JK, Viscido SV, Grunbaum D (2002) Self-organized fish schools: an examination of emergent properties. Biol Bull 202:296–305
73. Pavone M, Frazzoli E, Bullo F (2007) Decentralized algorithms for stochastic and dynamic vehicle routing with general target distribution. In: IEEE conference on decision and control, New Orleans, LA, December 2007, pp 4869–4874
74. Peleg D (2000) Distributed computing. A locality-sensitive approach. Monographs on discrete mathematics and applications. SIAM, Philadelphia, PA
75. P Ögren, Fiorelli E, Leonard NE (2004) Cooperative control of mobile sensor networks: Adaptive gradient climbing in a dis-

tributed environment. Trans IEEE Autom Control 49(8):1292–1302
76. Rathinam S, Sengupta R, Darbha S (2007) A resource allocation algorithm for multi-vehicle systems with non holonomic constraints. Trans IEEE Autom Sci Eng 4(1):98–104
77. Ren W, Beard RW (2005) Consensus seeking in multi-agent systems under dynamically changing interaction topologies. Trans IEEE Autom Control 50(5):655–661
78. Ren W, Beard RW, Atkins EM (2007) Information consensus in multivehicle cooperative control: Collective group behavior through local interaction. Control IEEE Syst Mag 27(2):71–82
79. Santoro N (2001) Distributed computations by autonomous mobile robots. In: Pacholski L, Ruzicka P (eds) SOFSEM 2001: Conference on current trends in theory and practice of informatics (Piestany, Slovak Republic). Lecture notes in computer science, vol 2234. Springer, New York, pp 110–115
80. Savkin AV (2004) Coordinated collective motion of groups of autonomous mobile robots: Analysis of Vicsek's model. Trans IEEE Autom Control 49(6):981–982
81. Savla K, Bullo F, Frazzoli E (2007) Traveling Salesperson Problems for a double integrator. Trans IEEE Autom Control, To appear
82. Savla K, Frazzoli E, Bullo F (2008) Traveling Salesperson Problems for the Dubins vehicle. Trans IEEE Autom Control 53(9) To appear
83. Savla K, Notarstefano G, Bullo F (2007) Maintaining limited-range connectivity among second-order agents. J SIAM Control Optim, To appear
84. Schumacher C, Chandler PR, Rasmussen SJ, Walker D (2003) Task allocation for wide area search munitions with variable path length. In: American control conference, Denver, CO, pp 3472–3477
85. Seeley TD, Buhrman SC (1999) Group decision-making in swarms of honey bees. Behav Eco Sociobiol 45:19–31
86. Sepulchre R, Paley D, Leonard NE (2007) Stabilization of planar collective motion: All-to-all communication. Trans IEEE Autom Control 52:811–824
87. Sharma V, Savchenko M, Frazzoli E, Voulgaris P (2005) Time complexity of sensor-based vehicle routing. In: Thrun S, Sukhatme G, Schaal S, Brock O (eds) Robotics: Science and systems. Press MIT, Cambridge, MA, pp 297–304
88. Sharma V, Savchenko M, Frazzoli E, Voulgaris P (2007) Transfer time complexity of conflict-free vehicle routing with no communications. Int J Robotics Res 26(3):255–272
89. Sinclair AR (1977) The african buffalo, study a of resource limitation of population. The University of Chicago Press, Chicago, IL
90. Smith SL, Broucke ME, Francis BA (2005) A hierarchical cyclic pursuit scheme for vehicle networks. Automatica 41(6):1045–1053
91. Smith SL, Bullo F (2007) Monotonic target assignment for robotic networks. Trans IEEE Autom Control, Submitted, To appear
92. Spanos DP, Olfati-Saber R, Murray RM (2005) Approximate distributed Kalman filtering in sensor networks with quantifiable performance. In: Symposium on information processing of sensor networks (IPSN). Los Angeles, CA, April 2005. pp 133–139
93. Stewart KJ, Harcourt AH (1994) Gorillas vocalizations during rest periods – signals of impending departure. Behaviour 130:29–40
94. Suzuki I, Yamashita M (1999) Distributed anonymous mobile robots: Formation of geometric patterns. J SIAM Comput 28(4):1347–1363
95. Tahbaz-Salehi A, Jadbabaie A (2008) Consensus over random networks. Trans IEEE Autom Control 53(3):791–795
96. Tanner HG, Jadbabaie A, Pappas GJ (2007) Flocking in fixed and switching networks. Trans IEEE Autom Control 52(5):863–868
97. Tsitsiklis JN (1984) Problems in decentralized decision making and computation. Dissertation, Laboratory for Information and Decision Systems, MIT, Nov. Technical Report LIDS-TH-1424
98. Tsitsiklis JN, Bertsekas DP, Athans M (1986) Distributed asynchronous deterministic and stochastic gradient optimization algorithms. Trans IEEE Autom Control 31(9):803–812
99. Uny Y Cao, Fukunaga AS, Kahng A (1997) Cooperative mobile robotics: Antecedents and directions. Auton Robots 4(1):7–27
100. Wu CW (2006) Synchronization and convergence of linear dynamics in random directed networks. Trans IEEE Autom Control 51(7):1207–1210
101. Xiao L, Boyd S, Lall S (2005) A scheme for asynchronous distributed sensor fusion based on average consensus. In: Symposium on information processing of sensor networks (IPSN). Los Angeles, CA, April 2005. pp 63–70
102. Zavlanos M, Pappas G (2007) Dynamic assignment in distributed motion planning with local information. In: American control conference. New York, July 2007, pp 1173–1178
103. Zavlanos MM, Pappas GJ (2005) Controlling connectivity of dynamic graphs. In: IEEE conference on decision and control and european control conference. Seville, Spain, December 2005, pp 6388–6393

Rough and Rough-Fuzzy Sets in Design of Information Systems

Theresa Beaubouef[1], Frederick Petry[2]
[1] Southeastern Louisiana University, Hammond, USA
[2] Naval Research Lab, Stennis Space Center, Mississippi, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Rough Sets and Rough-Fuzzy Sets
Rough Relational Database
Information Theory
Entropy and the Rough Relational Database
Rough Fuzzy Relational Database
Rough Set Modeling of Spatial Data
Data Mining in Rough Databases
Future Directions
Acknowledgment
Bibliography

Glossary

Rough sets Rough set theory is a technique for dealing with uncertainty and for identifying cause-effect relationships in databases. It is based on a partitioning of some domain into equivalence classes and the defining of lower and upper approximation regions based on this partitioning to denote certain and possible inclusion in the rough set.

Fuzzy sets Fuzzy set theory is another technique for dealing with uncertainty. It is based on the concept of measuring the degree of inclusion in a set through the use of a membership value. Where elements can either belong or not belong to a regular set, with fuzzy sets elements can belong to the set to a certain degree with zero indicating not an element, one indicating complete membership, and values between zero and one indicating partial or uncertain membership in the set.

Information theory Information theory involves the study of measuring the information content of a signal. In databases information theoretic measures can be used to measure the information content of data. Entropy is one such measure.

Database A collection of data and the application programs that make use of this data for some enterprise is a database.

Information system An information system is a database enhanced with additional tools that can be used by management for planning and decision making.

Data mining Data mining involves the discovery of patterns or rules in a set of data. These patterns generate some knowledge and information from the raw data that can be used for making decisions. There are many approaches to data mining, and uncertainty management techniques play a vital role in knowledge discovery.

Definition of the Subject

Databases and information systems are ubiquitous in this age of information and technology. Computers have revolutionized the way data can be manipulated and stored, allowing for very large databases with sophisticated capabilities. With so much money and manpower invested in the design and daily use of these systems, it is imperative that they be as correct, secure, and as adaptable to the changing needs of the enterprise as possible. Therefore it is important to understand the design and implementation of such systems and to be able to utilize all their capabilities.

Scientists and business executives alike know the value of information. The challenge has been to produce relevant information for an ever changing uncertain world from data and facts stored on computers and archival devices. These data are considered to be exact, certain, factual values. The real world, however, is uncertain, inexact, and fraught with errors. It is a challenge, then, to extract useful and relevant information from ordinary databases. Uncertainty management techniques such as rough and fuzzy sets can help.

Introduction

Databases are recognized for their ability to store and update data in an efficient manner, providing reliability and the elimination of data redundancy. The relational database model, in particular, has well-established mechanisms built into the model for properly designing the database and maintaining data integrity and consistency. Data alone, however, are only facts. What is needed is information. Knowledge discovery attempts to derive information from the pure facts, discovering high level regularities in the data. It is defined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [24,27].

An innovative technique in the field of uncertainty and knowledge discovery is based on rough sets. Rough set theory, introduced and further developed mathematically by [39], provides a framework for the representation of uncertainty. It has been used in various applications such as the rough querying of crisp data [6], uncertainty management in databases [11], the mining of spatial data [8], and improved information retrieval [48]. These techniques may readily be extended for use with object-oriented, spatial, and other complex databases, and may be integrated with additional data mining techniques for a comprehensive knowledge discovery approach.

Rough Sets and Rough-Fuzzy Sets

Rough set theory, introduced by Pawlak [38], is a technique for dealing with uncertainty and for identifying cause-effect relationships in databases. An extensive theory for rough sets and their properties has been developed and they have become a well established approach for the management of uncertainty in a variety of applications. Rough sets involve the following:

$\boldsymbol{U}$ is the *universe*, which cannot be empty,

$\boldsymbol{R}$ is the *indiscernibility relation*, or equivalence relation,

$\boldsymbol{A = (U, R)}$ an ordered pair, is called an *approximation space*,

$[\boldsymbol{x}]_{\boldsymbol{R}}$ denotes the equivalence class of R containing x, for any element x of U,

Elementary sets in A – the equivalence classes of R,

Definable set in A – any finite union of elementary sets in A.

Given an approximation space defined on some universe U that has an equivalence relation R imposed upon it, U is partitioned into equivalence classes called elementary sets that may be used to define other sets in A. A rough set X, where $X \subseteq U$, can be defined in terms of the definable sets in A by the following:

Lower approximation of X in A is the set

$\underline{R}X = \{x \in U \mid [x]_R \subseteq X\}$

Upper approximation of X in A is the set

$\overline{R}X = \{x \in U \mid [x]_R \cap X \neq \emptyset\}$.

$\mathrm{POS}_R(X) = \underline{R}X$ denotes the R-positive region of X, or those elements which certainly belong to the rough set. The R-negative region of X, $\mathrm{NEG}_R(X) = U - \overline{R}X$, contains elements which do not belong to the rough set, and the boundary or R-borderline region of X, $\mathrm{BN}_R(X) = \overline{R}X - \underline{R}X$, contains those elements which may or may not belong to the set. X is R-definable if and only if $\underline{R}X = \overline{R}X$. Otherwise, $\underline{R}X \neq \overline{R}X$ and X is rough with respect to R. A *rough set in A* is the group of subsets of U with the same upper and lower approximations.

Fuzzy set theory [53] is another approach for managing uncertainty. It has been around for a few years longer than rough sets, and also has well developed theory, properties, and applications. Applications involving fuzzy logic are diverse and plentiful, ranging from fuzzy control systems in industry to fuzzy logic in databases.

Because there are advantages to both fuzzy set and rough set theories, several researchers have studied various ways of combining the two theories [22,30,36] Others have investigated the interrelations between the two theories [19,40,52]. A similar approach to our rough-fuzzy set is the fuzzy rough set in [31]. That approach is more in the spirit of functional analysis, however. Fuzzy sets and rough sets are not equivalent, but complementary.

It has been shown in [52] that rough sets can be expressed by a fuzzy membership function $\mu \to \{0, 0.5, 1\}$ to represent the negative, boundary, and positive regions. In this model, all elements of the lower approximation, or positive region, have a membership value of one. Those elements of the boundary region are assigned a membership value of 0.5. Elements not belonging to the rough set have a membership value of zero. Rough set definitions of union and intersection can be modified so that the fuzzy model satisfies all the properties of rough sets [7]. This allows a rough set to be expressed as a fuzzy set.

We integrate fuzziness into the rough set model in order to quantify levels of roughness in boundary region areas through the use of fuzzy membership values. Therefore, we do not require membership values of elements of the boundary region to equal 0.5, but allow them to range from zero to one, noninclusive. Additionally, the union and intersection operators for fuzzy rough sets are comparable to those for ordinary fuzzy sets, where MIN and MAX are used to obtain membership values of redundant elements.

Let U be a *universe*, X a rough set in U.

Definition A *fuzzy rough set* Y in U is a membership function $\mu_Y(x)$ which associates a grade of membership from the interval [0,1] with every element of U where

$$\mu_Y(\underline{R}X) = 1, \quad \mu_Y(U - \overline{R}X) = 0, \text{ and } \quad 0 < \mu_Y(\overline{R}X - \underline{R}X) < 1.$$

Definition The *union* of two fuzzy rough sets A and B is a fuzzy rough set C where

$$C = \{x | x \in A \text{ OR } x \in B\}, \text{ where } \quad \mu_C(x) = \mathrm{MAX}[\mu_A(x), \mu_B(x)].$$

Definition The *intersection* of two fuzzy rough sets A and B is a fuzzy rough set C where

$$C = \{x | x \in A \text{ AND } x \in B\}, \text{ where } \quad \mu_C(x) = \mathrm{MIN}[\mu_A(x), \mu_B(x)].$$

Rough Relational Database

The rough relational database model [13] is an extension of the standard relational database model of Codd [23]. It captures all the essential features of rough sets theory including indiscernibility of elements denoted by equivalence classes and lower and upper approximation regions for defining sets which are indefinable in terms of the indiscernibility.

Every attribute domain is partitioned by some equivalence relation designated by the database designer or user. Within each domain, those values that are considered indiscernible belong to an equivalence class. This information is used by the query mechanism to retrieve information based on equivalence with the class to which the value belongs rather than equality, resulting in less critical wording of queries.

Recall is also improved in the rough relational database because rough relations provide *possible* matches to the query in addition to the *certain* matches which are obtained in the standard relational database. This is accomplished by using set containment in addition to equality of

attributes in the calculation of lower and upper approximation regions of the query result.

The rough relational database has several features in common with the ordinary relational database. Both models represent data as a collection of *relations* containing *tuples*. These relations are sets. The tuples of a relation are its elements, and like elements of sets in general, are unordered and nonduplicated. A tuple $\mathbf{t}_i$ takes the form $(d_{i1}, d_{i2}, \ldots, d_{im})$, where d_{ij} is a *domain value* of a particular *domain set* D_j. In the ordinary relational database, $d_{ij} \in D_j$. In the rough database, however, as in other non-first normal form extensions to the relational model [34,44], $d_{ij} \subseteq D_j$, and although it is not required that d_{ij} be a singleton, $d_{ij} \neq \emptyset$. Let $P(D_i)$ denote the powerset $(D_i) - \emptyset$.

Definition A *rough relation* R is a subset of the set cross product $P(D_1) \times P(D_2) \times \cdots \times P(D_m)$.

A rough tuple $\mathbf{t}$ is any member of R, which implies that it is also a member of $P(D_1) \times P(D_2) \times \cdots \times P(D_\mathrm{m})$. If $\mathbf{t}_\mathrm{i}$ is some arbitrary tuple, then $\mathbf{t}_\mathrm{i} = (d_{i1}, d_{i2}, \ldots, d_{im})$ where $d_{ij} \subseteq D_\mathrm{j}$. A tuple in this model differs from that of ordinary databases in that the tuple components may be sets of domain values rather than single values. The set braces are omitted from singletons for notational simplicity.

Let $[d_{xy}]$ denote the equivalence class to which d_{xy} belongs. When d_{xy} is a set of values, the equivalence class is formed by taking the union of equivalence classes of members of the set; if $d_{xy} = \{c_1, c_2, \ldots, c_n\}$, then $[d_{xy}] = [c_1] \cup [c_2] \cup \cdots \cup [c_n]$.

Definition Tuples $\mathbf{t}_i = (d_{i1}, d_{i2}, \ldots, d_{im})$ and $\mathbf{t}_k = (d_{k1}, d_{k2}, \ldots, d_{km})$ are *redundant* if $[d_{ij}] = [d_{kj}]$ for all $j = 1, \ldots, m$.

In the rough relational database, redundant tuples are removed in the merging process since duplicates are not allowed in sets, the structure upon which the relational model is based.

There are two basic types of relational operators. The first type arises from the fact that relations are considered sets of tuples. Therefore, operations which can be applied to sets also apply to relations. The most useful of these for database purposes are *set difference*, *union*, and *intersection*. Operators which do not come from set theory, but which are useful for retrieval of relational data are *select*, *project*, and *join*.

In the rough relational database, relations are rough sets as opposed to ordinary sets. Therefore, new rough operators (—, $\cup$, $\cap$, σ, π, $\bowtie$), which are comparable to the standard relational operators, must be developed for the rough relational database. Moreover, a mechanism must exist within the database to mark tuples of a rough relation as belonging to the lower or upper approximation of that rough relation. Properties of the rough relational operators can be found in [13].

Information Theory

In communication theory, Shannon [45] introduced the concept of entropy which was used to characterize the information content of signals. Since then, variations of these information theoretic measures have been successfully applied to applications in many diverse fields. In particular, the representation of uncertain information by entropy measures has been applied to all areas of databases, including fuzzy database querying [17], data allocation [25], classification in rule-based systems [42], and measuring uncertainty in rough and fuzzy rough relational databases [12].

In fuzzy set theory the representation of uncertain information measures has been extensively studied [14,20,29]. So this paper relates the concepts of information theory to rough sets and compares these information theoretic measures to established rough set metrics of uncertainty. The measures are then applied to the rough relational database model [13]. Information content of both stored relational schemas and rough relations are expressed as types of rough entropy.

Rough set theory [38] inherently models two types of uncertainty. The first type of uncertainty arises from the indiscernibility relation that is imposed on the universe, partitioning all values into a finite set of equivalence classes. If every equivalence class contains only one value, then there is no loss of information caused by the partitioning. In any coarser partitioning, however, there are fewer classes, and each class will contain a larger number of members. Our knowledge, or information, about a particular value decreases as the granularity of the partitioning becomes coarser.

Uncertainty is also modeled through the approximation regions of rough sets where elements of the lower approximation region have total participation in the rough set and those of the upper approximation region have uncertain participation in the rough set. Equivalently, the lower approximation is the *certain* region and the boundary area of the upper approximation region is the *possible* region.

Pawlak [41] discusses two numerical characterizations of imprecision of a rough set X: *accuracy* and *roughness*. Accuracy, which is simply the ratio of the number of elements in the lower approximation of X, $\underline{R}X$, to the number of elements in the upper approximation of the rough

set X, $\overline{R}X$, measures the degree of completeness of knowledge about the given rough set X. It is defined as a ratio of the two set cardinalities as follows:

$$\alpha_R(X) = \text{card}(\underline{R}X)/\text{card}(\overline{R}X), \text{ where } 0 \leq \alpha_R(X) \leq 1.$$

The second measure, roughness, represents the degree of incompleteness of knowledge about the rough set. It is calculated by subtracting the accuracy from 1: $\rho_R(X) = 1 - \alpha_R(X)$.

These measures require knowledge of the number of elements in each of the approximation regions and are good metrics for uncertainty as it arises from the boundary region, implicitly taking into account equivalence classes as they belong wholly or partially to the set. However, accuracy and roughness measures do not necessarily provide us with information on the uncertainty related to the granularity of the indiscernibility relation for those values that are totally included in the lower approximation region. For example:

Let the rough set X be defined as follows: $X = \{A11, A12, A21, A22, B11, C1\}$ with lower and upper approximation regions defined as

$$\underline{R}X = \{A11, A12, A21, A22\} \text{ and}$$
$$\overline{R}X = \{A11, A12, A21, A22, B11, B12, B13, C1, C2\}.$$

These approximation regions may result from one of several partitionings. Consider, for example, the following indiscernibility relations:

$$A_1 = \{[A11, A12, A21, A22], [B11, B12, B13], [C1, C2]\},$$
$$A_2 = \{[A11, A12], [A21, A22], [B11, B12, B13], [C1, C2]\},$$
$$A_3 = \{[A11], [A12], [A21], [A22], [B11, B12, B13], [C1, C2]\}.$$

All three of the above partitionings result in the same upper and lower approximation regions for the given set X, and hence the same accuracy measure ($4/9 = .444$) since only those classes belonging to the lower approximation region were re-partitioned. It is obvious, however, that there is more uncertainty in A_1 than in A_2, and more uncertainty in A_2 than in A_3. Therefore, a more comprehensive measure of uncertainty is needed.

We derive such a measure from techniques used for measuring entropy in classical information theory. Countless variations of the classical entropy have been developed, each tailored for a particular application domain or for measuring a particular type of uncertainty. Our rough entropy is defined such that we may apply it to rough databases. We define the entropy of a rough set X as follows:

Definition The *rough entropy* $E_r(X)$ of a rough set X is calculated by

$$E_r(X) = -(\rho_R(X))[\Sigma Q_i \log(P_i)] \quad \text{for } i = 1, \ldots n \text{ equivalence classes.}$$

The term $\rho_R(X)$ denotes the roughness of the set X. The second term is the summation of the probabilities for each equivalence class belonging either wholly or in part to the rough set X. There is no ordering associated with individual class members. Therefore the probability of any one value of the class being named is the reciprocal of the number of elements in the class. If c_i is the cardinality of, or number of elements in, equivalence class i and all members of a given equivalence class are equal, $P_i = 1/c_i$ represents the probability of one of the values in class i. Q_i denotes the probability of equivalence class i within the universe. Q_i is computed by taking the number of elements in class i and dividing by the total number of elements in all equivalence classes combined. The entropy of the sample rough set X, $E_r(X)$, is given below for each of the possible indiscernibility relations A_1, A_2, and A_3.

$$\text{Using } A_1: -(5/9)[(4/9)\log(1/4) + (3/9)\log(1/3) + (2/9)\log(1/2)] = .274$$
$$\text{Using } A_2: -(5/9)[(2/9)\log(1/2) + (2/9)\log(1/2) + (3/9)\log(1/3) + (2/9)\log(1/2)] = .20$$
$$\text{Using } A_3: -(5/9)[(1/9)\log(1) + (1/9)\log(1) + (1/9)\log(1) + (1/9)\log(1) + (3/9)\log(1/3) + (2/9)\log(1/2)] = .048$$

From the above calculations it is clear that although each of the partitionings results in identical roughness measures, the entropy decreases as the classes become smaller through finer partitionings.

Entropy and the Rough Relational Database

The basic concepts of rough sets and their information-theoretic measures carries over to the rough relational database model [13]. Recall that in the rough relational database all domains are partitioned into equivalence classes and relations are not restricted to first normal form. We therefore have a type of rough set for each attribute of a relation. This results in a rough relation, since any tuple

having a value for an attribute that belongs to the boundary region of its domain is a tuple belonging to the boundary region of the rough relation.

There are two things to consider when measuring uncertainty in databases: uncertainty or entropy of a rough relation that exists in a database at some given time and the entropy of a relation schema for an existing relation or query result. We must consider both since the approximation regions only come about by set values for attributes in given tuples. Without the extension of a database containing actual values, we only know about indiscernibility of attributes. We cannot consider the approximation regions.

We define the entropy for a rough relation schema as follows:

Definition The *rough schema entropy* for a rough relation schema S is

$$E_s(S) = -\Sigma_j[\Sigma Q_i \log(P_i)] \text{ for } i = 1, \ldots n; j = 1, \ldots, m$$

where there are n equivalence classes of domain j, and m attributes in the schema $R(A_1, A_2, \ldots, A_m)$.

This is similar to the definition of entropy for rough sets without factoring in roughness since there are no elements in the boundary region (lower approximation = upper approximation). However, because a relation is a cross product among the domains, we must take the sum of all these entropies to obtain the entropy of the schema. The schema entropy provides a measure of the uncertainty inherent in the definition of the rough relation schema taking into account the partitioning of the domains on which the attributes of the schema are defined.

We extend the schema entropy $E_s(S)$ to define the entropy of an actual rough relation instance $E_R(R)$ of some database D by multiplying each term in the product by the roughness of the rough set of values for the domain of that given attribute.

Definition The *rough relation entropy* of a particular extension of a schema is

$$E_R(R) = -\Sigma_j D\rho_j(R)[\Sigma DQ_i \log(DP_i)] \quad \text{for } i = 1, \ldots n; j = 1, \ldots, m$$

where $D\rho_j(R)$ represents a type of database roughness for the rough set of values of the domain for attribute j of the relation, m is the number of attributes in the database relation, and n is the number of equivalence classes for a given domain for the database.

We obtain the $D\rho_j(R)$ values by letting the non-singleton domain values represent elements of the boundary region, computing the original rough set accuracy and subtracting it from one to obtain the roughness. DQ_i is the probability of a tuple in the database relation having a value from class i, and DP_i is the probability of a value for class i occurring in the database relation out of all the values which are given.

Information theoretic measures again prove to be a useful metric for quantifying information content. In rough sets and the rough relational database, this is especially useful since in ordinary rough sets Pawlak's measure of roughness does not seem to capture the information content as precisely as our rough entropy measure.

In rough relational databases, knowledge about entropy can either guide the database user toward less uncertain data or act as a measure of the uncertainty of a data set or relation. As rough relations become larger in terms of the number of tuples or attributes, the automatic calculation of some measure of entropy becomes a necessity. Our rough relation entropy measure fulfills this need.

Rough Fuzzy Relational Database

The fuzzy rough relational database, as in the ordinary relational database, represents data as a collection of *relations* containing *tuples*. Because a relation is considered a set having the tuples as its members, the tuples are unordered. In addition, there can be no duplicate tuples in a relation. A tuple $\mathbf{t}_i$ takes the form $(d_{i1}, d_{i2}, \ldots, d_{im}, d_{i\mu})$, where d_{ij} is a *domain value* of a particular *domain set* D_j and $d_{i\mu} \in D_\mu$, where D_μ is the interval [0,1], the domain for fuzzy membership values. In the ordinary relational database, $d_{ij} \in D_j$. In the fuzzy rough relational database, except for the fuzzy membership value, however, $d_{ij} \subseteq D_j$, and although d_{ij} is not restricted to be a singleton, $d_{ij} \neq \emptyset$. Let $P(D_i)$ denote any non-null member of the powerset of D_i.

Definition A *fuzzy rough relation R* is a subset of the set cross product $P(D_1) \times P(D_2) \times \cdots \times P(D_m) \times D_\mu$.

For a specific relation, R, membership is determined semantically. Given that D_1 is the set of names of nuclear/chemical plants, D_2 is the set of locations, and assuming that RIVERB is the only nuclear power plant that is located in VENTRESS,

(RIVERB, VENTRESS, 1)

(RIVERB, OSCAR, .7)

(RIVERB, ADDIS, 1)

(CHEMO, VENTRESS, .3)

are all elements of $P(D_1) \times P(D_2) \times D_\mu$. However, only the element (RIVERB, VENTRESS, 1) of those

listed above is a member of the relation R(PLANT, LOCATION, μ), which associates each plant with the town or community in which it is located. A *fuzzy rough tuple* $\mathbf{t}$ is any member of R. If $\mathbf{t}_i$ is some arbitrary tuple, then $\mathbf{t}_i = (d_{i1}, d_{i2}, \ldots, d_{im}, d_{i\mu})$ where $d_{ij} \subseteq D_j$ and $d_{i\mu} \in D_\mu$.

Definition An *interpretation* $\alpha = (a_1, a_2, \ldots, a_m, a_\mu)$ of a fuzzy rough tuple $\mathbf{t}_i = (d_{i1}, d_{i2}, \ldots, d_{im}, d_{i\mu})$ is any value assignment such that $a_j \in d_{ij}$ for all j.

The interpretation space is the cross product $D_1 \times D_2 \times \cdots \times D_m \times D_\mu$, but is limited for a given relation R to the set of those tuples which are valid according to the underlying semantics of R. In an ordinary relational database, because domain values are atomic, there is only one possible interpretation for each tuple $\mathbf{t}_i$. Moreover, the interpretation of $\mathbf{t}_i$ is equivalent to the tuple $\mathbf{t}_i$. In the fuzzy rough relational database, this is not always the case.

Let $[d_{xy}]$ denote the equivalence class to which d_{xy} belongs. When d_{xy} is a set of values, the equivalence class is formed by taking the union of equivalence classes of members of the set; if $d_{xy} = \{c_1, c_2, \ldots, c_n\}$, then $[d_{xy}] = [c_1] \cup [c_2] \cup \cdots \cup [c_n]$.

Definition Tuples $\mathbf{t}_i = (d_{i1}, d_{i2}, \ldots, d_{in}, d_{i\mu})$ and $\mathbf{t}_k = (d_{k1}, d_{k2}, \ldots, d_{kn}, d_{k\mu})$ are *redundant* if $[d_{ij}] = [d_{kj}]$ for all $j = 1, \ldots, n$.

If a relation contains only those tuples of a lower approximation, i. e., those tuples having a μ value equal to one, the interpretation α of a tuple is unique. This follows immediately from the definition of redundancy. In fuzzy rough relations, there are no redundant tuples. The merging process used in relational database operations removes duplicate tuples since duplicates are not allowed in sets, the structure upon which the relational model is based.

Tuples may be redundant in all values except μ. As in the union of fuzzy rough sets where the maximum membership value of an element is retained, it is the convention of the fuzzy rough relational database to retain the tuple having the higher μ value when removing redundant tuples during merging. If we are supplied with identical data from two sources, one certain and the other uncertain, we would want to retain the data that is certain, avoiding loss of information.

Recall that the rough relational database is in non-first normal form; there are some attribute values that are sets. Another definition, which will be used for upper approximation tuples, is necessary for some of the alternate definitions of operators to be presented. This definition captures redundancy between elements of attribute values that are sets:

Definition Two sub-tuples $X = (d_{x1}, d_{x2}, \ldots, d_{xm})$ and $Y = (d_{y1}, d_{y2}, \ldots, d_{ym})$ are *roughly-redundant*, $\approx_{\mathbf{R}}$, if for some $[p] \subseteq [d_{xj}]$ and $[q] \subseteq [d_{yj}]$, $[p] = [q]$ for all $j = 1, \ldots, m$.

In order for any database to be useful, a mechanism for operating on the basic elements and retrieving specified data must be provided. The concepts of redundancy and merging play a key role in the operations defined.

We must first design our database using some type of semantic model. We use a variation of the entity-relationship diagram that we call a fuzzy-rough E-R diagram. This diagram is similar to the standard E-R diagram in that entity types are depicted in rectangles, relationships with diamonds, and attributes with ovals. However, in the fuzzy-rough model, it is understood that membership values exist for all instances of entity types and relationships. Attributes which allow values where we want to be able to define equivalences are denoted with an asterisk (*) above the oval. These values are defined in the indiscernibility relation, which is not actually part of the database design, but inherent in the fuzzy-rough model.

Our fuzzy-rough E-R model [7] is similar to the second and third levels of fuzziness defined by Zvieli and Chen [54]. However, in our model, all entity and relationship occurrences (second level) are of the fuzzy type so we do not mark an 'f' beside each one. Zvieli and Chen's third level considers attributes that may be fuzzy. They use triangles instead of ovals to represent these attributes. We do not introduce fuzziness at the attribute level of our model in this paper, only roughness, or indiscernibility, and denote those attribute with the '*'. From the Fuzzy-Rough E-R diagram, we design the structure of the fuzzy rough relational database. If we have a priori information about the types of queries that will be involved, we can make intelligent choices that will maximize computer resources.

We next formally define the fuzzy rough relational database operators and discuss issues relating to the real-world problems of data representation and modeling. We may view indiscernibility as being modeled through the use of the indiscernibility relation, imprecision through the use of non-first normal form constructs, and degree of uncertainty and fuzziness through the use of tuple membership values, which are given as the value for the μ attribute in every fuzzy rough relation.

Fuzzy Rough Relational Operators

In [13], we defined several operators for the rough relational algebra. We now define similar operators for the fuzzy rough relational database as in [5]. Recall that for all of these operators the indiscernibility relation is used

for equivalence of attribute values rather than equality of values.

Difference The fuzzy rough relational difference operator is very much like the ordinary difference operator in relational databases and in sets in general. It is a binary operator that returns those elements of the first argument that are not contained in the second argument.

In the fuzzy rough relational database, the difference operator is applied to two fuzzy rough relations and, as in the rough relational database, indiscernibility, rather than equality of attribute values, is used in the elimination of redundant tuples. Hence, the difference operator is somewhat more complex. Let X and Y be two union compatible fuzzy rough relations.

Definition The *fuzzy rough difference*, $\boldsymbol{X}$ - $\boldsymbol{Y}$, between X and Y is a fuzzy rough relation T where

$$T = \{t(d_1, \ldots, d_n, \mu_i) \in X | t(d_1, \ldots, d_n, \mu_i) \notin Y\} \cup \{t(d_1, \ldots, d_n, \mu_i) \in X | t(d_1, \ldots, d_n, \mu_j) \in Y \text{ and } \mu_i > \mu_j\}.$$

The resulting fuzzy rough relation contains all those tuples which are in the lower approximation of X, but not redundant with a tuple in the lower approximation of Y. It also contains those tuples belonging to upper approximation regions of both X and Y, but which have a higher μ value in X than in Y. For example, let X contain the tuple **(MODERN, 1)** and Y contain the tuple **(MODERN, .02)**. It would not be desirable to subtract out certain information with possible information, so X - Y yields **(MODERN, 1)**.

Union Because relations in databases are considered as sets, the union operator can be applied to any two union-compatible relations to result in a third relation which has as its tuples all the tuples contained in either or both of the two original relations. The union operator can be extended to apply to fuzzy rough relations. Let X and Y be two union compatible fuzzy rough relations.

Definition The *fuzzy rough union* of X and Y, $\mathbf{X} \cup \mathbf{Y}$ is a fuzzy rough relation T where

$$T = \{t | t \in X \text{ OR } t \in Y\} \quad \text{and} \quad \mu_T(t) = \text{MAX}[\mu_X(t), \mu_Y(t)].$$

The resulting relation T contains all tuples in either X or Y or both, merged together and having redundant tuples removed. If X contains a tuple that is redundant with a tuple in Y except for the μ value, the merging process will retain only that tuple with the higher μ value.

Intersection The fuzzy rough intersection, another binary operator on fuzzy rough relations, can be defined similarly.

Definition The *fuzzy rough intersection* of X and Y, $\mathbf{X} \cap \mathbf{Y}$ is a fuzzy rough relation T where

$$T = \{t | t \in X \text{ AND } t \in Y\} \quad \text{and} \quad \mu_T(t) = \text{MIN}[\mu_X(t), \mu_Y(t)].$$

In intersection, the MIN operator is used in the merging of equivalent tuples having different μ values and the result contains all tuples that are members of both of the original fuzzy rough relations.

Definition The *fuzzy rough intersection* of X and Y, $\mathbf{X} \cap_{\mathbf{A}} \mathbf{Y}$ is a fuzzy rough relation T where

$$T = \{t | t \in X, \text{ and } \exists s \in Y | t \approx_R s\} \cup \{s | s \in Y, \text{ and } \exists t \in X | s \approx_R t\} \text{ and } \mu_T(t) = \text{MIN}[\mu_X(t), \mu_Y(t)].$$

Select The select operator for the fuzzy rough relational database model, $\boldsymbol{\sigma}$, is a unary operator which takes a fuzzy rough relation X as its argument and returns a fuzzy rough relation containing a subset of the tuples of X, selected on the basis of values for a specified attribute. The operation $\boldsymbol{\sigma}_{\mathbf{A=a}}(X)$, for example, returns those tuples in X where attribute A is equivalent to the class [**a**]. In general, select returns a subset of the tuples that match some selection criteria.

Let R be a relation schema, X a fuzzy rough relation on that schema, A an attribute in R, $\mathbf{a} = \{a_i\}$ and $\mathbf{b} = \{b_j\}$, where $a_i, b_j \in \text{dom}(A)$, and $\cup_x$ is interpreted as "the union over all x".

Definition The *fuzzy rough selection*, $\boldsymbol{\sigma}_{\mathbf{A=a}}(X)$, of tuples from X is a fuzzy rough relation Y having the same schema as X and where

$$Y = \{t \in X | \cup_i [a_i] \subseteq \cup_j [b_j]\},$$

and $a_i \in \mathbf{a}$, $b_j \in t(A)$, and where membership values for tuples are calculated by multiplying the original membership value by

$$\text{card}(\mathbf{a})/\text{card}(\mathbf{b})$$

where card(x) returns the cardinality, or number of elements, in x.

Assume we want to retrieve those elements where **CITY = "ADDIS"** from the following fuzzy rough tuples:

(**ADDIS** **1**)
({**ADDIS, LOTTIE, BRUSLY**} .**7**)
(**OSCAR** **1**)
({**ADDIS, JACKSON**} .**9**).

The result of the selection is the following:

(**ADDIS** **1**)
({**ADDIS, LOTTIE, BRUSLY**} .**23**)
({**ADDIS, JACKSON**} .**45**),

where the μ for the second tuple is the product of the original membership value .7 and 1/3.

Project Project is a unary fuzzy rough relational operator. It returns a relation that contains a subset of the columns of the original relation. Let X be a fuzzy rough relation with schema A, and let B be a subset of A. The fuzzy rough projection of X onto B is a fuzzy rough relation Y obtained by omitting the columns of X which correspond to attributes in A - B, and removing redundant tuples. Recall the definition of redundancy accounts for indiscernibility, which is central to the rough sets theory and that higher μ values have priority over lower ones.

Definition The *fuzzy rough projection* of X onto B, $\boldsymbol{\pi}_{\mathbf{B}}(\mathbf{X})$, is a fuzzy rough relation Y with schema $Y(B)$ where

$$Y(B) = \{t(B) | t \in X\}.$$

Join Join is a binary operator that takes related tuples from two relations and combines them into single tuples of the resulting relation. It uses common attributes to combine the two relations into one, usually larger, relation. Let $X(A_1, A_2, \ldots, A_m)$ and $Y(B_1, B_2, \ldots, B_n)$ be fuzzy rough relations with **m** and **n** attributes, respectively, and $AB = C$, the schema of the resulting fuzzy rough relation T.

Definition The *fuzzy rough join*, $\mathbf{X} \bowtie_{\text{<JOIN CONDITION>}} \mathbf{Y}$, of two relations X and Y, is a relation $T(C_1, C_2, \ldots, C_{m+n})$ where

$$T = \{t | \exists t_X \in X, t_Y \in Y \text{ for } t_X = t(A), t_Y = t(B)\},$$

and where

$$t_X(A \cap B) = t_Y(A \cap B), \mu = 1 \tag{1}$$

$$t_X(A \cap B) \subseteq t_Y(A \cap B) \text{ or } t_Y(A \cap B) \subseteq t_X(A \cap B), \quad \mu = \text{MIN}(\mu_X, \mu_Y) \tag{2}$$

<JOIN CONDITION> is a conjunction of one or more conditions of the form **A** = **B**.

Only those tuples which resulted from the "joining" of tuples that were both in lower approximations in the original relations belong to the lower approximation of the resulting fuzzy rough relation. All other "joined" tuples belong to the upper approximation only (the boundary region), and have membership values less than one. The fuzzy membership value of the resultant tuple is simply calculated as in [18] by taking the minimum of the membership values of the original tuples. Taking the minimum value also follows the logic of [37], where in joins of tuples with different levels of information uncertainty, the resultant tuple can have no greater certainty than that of its least certain component.

Functional Dependencies

A functional dependency can be defined as in [23] through the use of a universal database relation concept. Let $R = \{A_1, A_2, \ldots, A_n\}$ be a universal relation schema describing a database having n attributes. Let X and Y be subsets of R. A functional dependency between the attributes of X and Y is denoted by $X \rightarrow Y$. This dependency specifies the constraint that for any two tuples of an instance r of R, if they agree on the X attribute(s) they must agree on their Y attributes(s): if $t_1[X] = t_2[X]$, then it must be true that $t_1[Y] = t_2[Y]$. Tuples that violate the constraint cannot be inserted into the database. The rough functional dependency is based on the rough relational database model. The classical notion of functional dependency for relational databases does not naturally apply to the rough relational database, since all the "roughness" would be lost. In the rough querying of crisp data [6], however, the data is stored in the standard relational model having ordinary functional dependencies imposed upon it and rough relations result only from querying; they are not a part of the database design in which the designer imposes constraint upon relation schemas. Rough functional dependencies for the rough relational database model are defined as follows [9]:

Definition A *rough functional dependency*, $X \rightarrow Y$, for a relation schema R exists if for all instances $T(R)$,

(1) for any two tuples $t, t' \in \underline{R}T$,

$$redundant(t(X), t'(X)) \Rightarrow redundant(t(Y), t\prime(Y))$$

(2) for any two tuples $s, s' \in \overline{R}T$,

$$\textit{roughly-redundant}(s(X), s'(X)) \Rightarrow \textit{roughly-redundant}(s(Y), s\prime(Y)).$$

Y is roughly functional dependent on X, or X roughly functionally determines Y, whenever the above definition holds. This implies that constraints can be imposed on a rough relational database schema in a rough manner that will aid in integrity maintenance and the reduction of update anomalies without limiting the expressiveness of the inherent rough set concepts.

It is obvious that the classical functional dependency for the standard relational database is a special case of the rough functional dependency; indiscernibility reduces to simple equality and part (2) of the definition is unused since all tuples in relations in the standard relational model belong to the lower approximation region of a similar rough model.

The first part of the definition of rough functional dependency compares with that of fuzzy functional dependencies discussed in [46], where adherence to Armstrong's axioms was proven. The results apply directly in the case of rough functional dependencies when only the lower approximation regions are considered. It is also necessary to show that axioms hold for upper approximations.

Rough Functional Dependencies Satisfy Armstrong's Axioms

Proof

(1) **Reflexive**

If $Y \subseteq X \subseteq U$, then

$$\textit{redundant}(t(X), t'(X)) \Rightarrow \textit{redundant}(t(Y), t'(Y)),$$
$$\text{and } \textit{roughly-redundant}(t(X), t'(X)) \Rightarrow \textit{roughly-redundant}(t(Y), t'(Y)).$$

Hence,

$$X \rightarrow Y.$$

(2) **Augmentation**

If $Z \subseteq U$ and the rough functional dependency $X \rightarrow Y$ holds, then

$$\textit{redundant}(t(XZ), t'(XZ)) \Rightarrow \textit{redundant}(t(YZ), t'(YZ)),$$
$$\text{and } \textit{roughly-redundant}(t(XZ), t'(XZ)) \Rightarrow \textit{roughly-redundant}(t(YZ), t'(YZ)).$$

Hence,

$$XZ \rightarrow YZ.$$

(3) **Transitive**

If the rough functional dependencies $X \rightarrow Y$ and $Y \rightarrow Z$ hold, then

$$\textit{redundant}(t(X), t'(X)) \Rightarrow \textit{redundant}(t(Z), t'(Z)),$$
$$\text{and } \textit{roughly-redundant}(t(X), t'(X)) \Rightarrow \textit{roughly-redundant}(t(Z), t'(Z)).$$

Hence,

$$X \rightarrow Z.$$

□

Hence, rough functional dependencies satisfy Armstrong's axioms. Given a set of rough functional dependencies, the complete set of rough functional dependencies can be derived using Armstrong's axioms. The rough functional dependency, therefore, is an important formalism for design in the rough relational database.

Fuzzy and rough set techniques integrated into the underlying data model result in databases that can more accurately represent real world enterprises since they incorporate uncertainty management directly into the data model itself. This is useful as is for obtaining greater information through the querying of rough and fuzzy databases. Additional benefits may be realized when they are used in the process of data mining.

Rough Set Modeling of Spatial Data

Many of the problems associated with data are prevalent in all types of database systems. Spatial databases and GIS contain descriptive as well as positional data. The various forms of uncertainty occur in both types of data, so many of the issues apply to ordinary databases as well, such as integration of data from multiple sources, time-variant data, uncertain data, imprecision in measurement, inconsistent wording of descriptive data, and "binning" or grouping of data into fixed categories, also are employed in spatial contexts.

First consider an example of the use of rough sets in representing spatially related data. Let U = {tower, stream, creek, river, forest, woodland, pasture, meadow} and let the equivalence relation R be defined as follows:

$$R^* = \{[\text{tower}], [\text{stream, creek, river}], [\text{forest, woodland}], [\text{pasture, meadow}]\}.$$

Given some set X = {tower, stream, creek, river, forest, pasture}, we would like to define it in terms of its lower and

upper approximations:

$$\underline{R}X = \{\text{tower, stream, creek, river}\}, \text{ and}$$
$$\overline{R}X = \{\text{tower, stream, creek, river, forest, woodland, pasture, meadow}\}.$$

A *rough set in A* is the group of subsets of U with the same upper and lower approximations. In the example given, the rough set is

{{tower, stream, creek, river, forest, pasture}
{tower, stream, creek, river, forest, meadow}
{tower, stream, creek, river, woodland, pasture}
{tower, stream, creek, river, woodland, meadow}}.

Often spatial data is associated with a particular grid. The positions are set up in a regular matrix-like structure and data is affiliated with point locations on the grid. This is the case for raster data and for other types of non-vector type data such as topography or sea surface temperature data. There is a tradeoff between the resolution or the scale of the grid and the amount of system resources necessary to store and process the data. Higher resolutions provide more information, but at a cost of memory space and execution time.

If we approach these data issues from a rough set point of view, it can be seen that there is indiscernibility inherent in the process of gridding or rasterizing data. A data item at a particular grid point in essence may represent data near the point as well. This is due to the fact that often point data must be mapped to the grid using techniques such as nearest-neighbor, averaging, or statistics. The rough set indiscernibility relation may be set up so that the entire spatial area is partitioned into equivalence classes where each point on the grid belongs to an equivalence class. If the resolution of the grid changes, then, in fact, this is changing the granularity of the partitioning, resulting in fewer, but larger classes.

The approximation regions of rough sets are beneficial whenever information concerning spatial data regions is accessed. Consider a region such as a forest. One can reasonably conclude that any grid point identified as FOREST that is surrounded on all sides by grid points also identified as FOREST is, in fact, a point represented by the feature FOREST. However, consider points identified as FOREST that are adjacent to points identified as MEADOW. Is it not possible that these points represent meadow area as well as forest area but were identified as FOREST in the classification process? Likewise, points identified as MEADOW but adjacent to FOREST points may represent areas that contain part of the forest. This uncertainty maps naturally to the use of the approximation regions of the rough set theory, where the lower approximation region represents certain data and the boundary region of the upper approximation represents uncertain data. It applies to spatial database querying and spatial database mining operations.

If we force a finer granulation of the partitioning, a smaller boundary region results. This occurs when the resolution is increased. As the partitioning becomes finer and finer, finally a point is reached where the boundary region is non-existent. Then the upper and lower approximation regions are the same and there is no uncertainty in the spatial data as can be determined by the representation of the model.

In [50] Worboys models imprecision in spatial data based on the resolution at which the data is represented, and for issues related to the integration of such data. This approach relies on the issue of indiscernibility – a core concept for rough sets – but does not carry over the entire framework and is just described as "reminiscent of the theory of rough sets" [51]. Ahlqvist and colleagues [2] used a rough set approach to define a rough classification of spatial data and to represent spatial locations. They also proposed a measure for quality of a rough classification compared to a crisp classification and evaluated their technique on actual data from vegetation map layers. They considered the combination of fuzzy and rough set approaches for reclassification as required by the integration of geographic data. Another research group in a mapping and GIS context [49] have developed an approach using a rough raster space for the field representation of a spatial entity and evaluated it on a classification case study for remote sensing images. In [16] Bittner and Stell consider K-labeled partitions, which can represent maps, and then develop their relationship to rough sets to approximate map objects with vague boundaries. Additionally they investigate stratified partitions, which can be used to capture levels of details or granularity such as in consideration of scale transformations in maps, and extend this approach using the concepts of stratified rough sets.

Data Mining in Rough Databases

Association rules capture the idea of certain data items commonly occurring together and have been often considered in the analysis of a "marketbasket" of purchases. For example, a delicatessen retailer might analyze the previous year's sales and observe that of all purchases 30% were of *both* cheese *and* crackers and, for any of the sales that included cheese, 75% also included crackers. Then it is pos-

sible to conclude a rule of the form:

Cheese → Crackers

This rule is said to have a 75% degree of confidence and a 30% degree of support This particular form of data is largely based on the a priori algorithm developed by Agrawal [1]. Let a database of possible data items be [28]

$$D = \{d_1, d_2, \ldots d_n\}$$

and the relevant set of transactions (sales, query results, etc.) are

$$T = \{T_1, T_2, \ldots\}$$

where $T_i \subseteq D$. We are interested in discovering if there is a relationship between two sets of items (called itemsets) X_j, X_k; X_j, $X_k \subseteq D$. For such a relationship to be determined, the entire set of transactions in T must be examined and a count made of the number of transactions containing these sets, where a transaction T_i contains X_m if $X_m \subseteq T_i$. This count, called the support count of X_m, $SC_T(X_m)$, will be appropriately modified in the case of rough sets.

There are then two measures used in determining rules: the percentage of T_i's in T that

1. Contain both X_j and X_k (i. e. $X_j \cup X_k$) – called the *support* s
2. If T_i contains X_j then T_i also contains X_k – called the *confidence* c.

The support and confidence can be interpreted as probabilities:

1. s – Prob $(X_j \cup X_k)$ and
2. c – Prob $(X_k | X_j)$

We assume the system user has provided minimum values for these in order to generate only sufficiently interesting rules. A rule whose support and confidence exceeds these minimums is called a strong rule.

The result of a query is then:

$$T = \{\underline{R}T, \overline{R}T\}$$

and so we must take into account the lower approximation, $\underline{R}T$, and upper approximation, $\overline{R}T$, results of the rough query in developing the association rules.

Recall that in order to generate frequent itemsets, we must count the number of transactions T_j that support an itemset X_j. In the ordinary data mining algorithm one simply counts the occurrence of a value as 1 if in the set, or 0 if not. But now since the query result T is a rough set, we must modify the support count SC_T. So we define the rough support count, RSC_T, for the set X_j, to count differently in the upper and lower approximations:

$$\mathrm{RSC}_T(X_j) = \sum_i W_{T_i}(X_j); X_j \subseteq T_i$$

where

$$W(X_j) = \begin{cases} 1 \text{ if } \; T_i \in \underline{R}T \\ a \text{ if } \; T_i \in \overline{R}T, 0 < a < 1. \end{cases}$$

The value, a, can be a subjective value obtained from the user depending on relative assessment of the roughness of the query result T. For the data mining example of the next section we chose a neutral default value of $a = \frac{1}{2}$. Note that $W(X_j)$ is included in the summation only all of the values of the itemset X_j are included in the transaction, i. e. it is a subset of the transaction.

Finally to produce the association rules from the set of relevant data T retrieved from the spatial database, we must consider the frequent itemsets. For the purposes of generating a rule such as $X_j\, X_k$ we can now extend the approach to rough support and confidence as follows:

$$\mathrm{RS} = \mathrm{RSC}_T(X_j \cup X_k)/|T|$$
$$\mathrm{RC} = \mathrm{RSC}_T(X_j \cup X_k)/\mathrm{RSC}_T(X_j)\,.$$

In the spatial data mining area there have only been a few efforts using rough sets. In the research described in [8] approaches for attribute induction knowledge discovery [43] in rough spatial data are investigated. In [15] Bittner considers rough sets for spatio-temporal data and how to discover characteristic configurations of spatial objects focusing on the use of topological relationships for characterizations. In a survey of uncertainty-based spatial data mining Shi et al. [47] provide a brief general comparison of fuzzy and rough set approaches for spatial data mining.

Future Directions

There are several other approaches to uncertainty representation that may be more suitable for certain applications. Type-2 fuzzy have been of considerable recent interest [35]. In these as opposed to ordinary fuzzy sets in which the underlying membership functions are crisp, here the membership function are themselves fuzzy. Intuitionistic sets introduced by Atanassov [3,4] are another generalization of a fuzzy set. Two characteristic functions are used for capturing both the ordinary idea of degree of membership in the intuitionistic set as well as the degree of

non-membership of elements in the set and can be used in database design [10]. Related to the concepts introduced by rough sets is the idea of granularity for managing complex data by abstraction using information granules as discussed by Lin [32,33]. A granular set approach has also been introduced [30] which is a set and a number of disjoint subsets that constitute a semi-partition. Some prior database research on ordered relations [26], although not presented in the context of uncertainty of data, may provide some approaches to extend our work in this area. A main emphasis for future work is the incorporation of some of these research topics into mainstream database, GIS commercial products and semi-structured data on the semantic web.

Acknowledgment

The authors would like to thank the Naval Research Laboratory's Base Program, Program Element No. 0602435N for sponsoring this research.

Bibliography

Primary Literature

1. Agrawal R, Imielinski T, Swami A (1993) Mining Association Rules between sets of items in large databases. Proc of the 1993 ACM-SIGMOD International Conference on Management of Data. ACM Press, New York, NY, pp 207–216
2. Ahlqvist O, Keukelaar J, Oukbir K (2000) Rough classification and accuracy assessment. Int J Geogr Inf Sci 14:475–496
3. Atanassov K (1986) Intuitionistic Fuzzy Sets. Fuzzy Sets Syst 20:87–96
4. Atanassov K (2000) Intuitionistic Fuzzy Sets; Theory and Applications. Physica Verlag, Heidelberg
5. Beaubouef T, Petry F (1994) Fuzzy Set Quantification of Roughness in a Rough Relational Database Model. Proc. Third IEEE International Conference on Fuzzy Systems, Orlando, Florida, pp 172–177
6. Beaubouef T, Petry F (1994) Rough Querying of Crisp Data in Relational Databases. Proc. Third Int.l Workshop on Rough Sets and Soft Computing (RSSC'94), San Jose, California, pp 368–375
7. Beaubouef T, Petry F (2000) Fuzzy Rough Set Techniques for Uncertainty Processing in a Relational Database. Int J Intell Syst 15:389–424
8. Beaubouef T, Petry F (2002) A Rough Set Foundation for Spatial Data Mining Involving Vague Regions. Proc.FUZZ-IEEE'02, Honolulu, Hawaii, pp 767–772
9. Beaubouef T, Petry F (2004) Rough Functional Dependencies. 2004 Multiconferences: International Conf. On Information and Knowledge Engineering (IKE'04), Las Vegas, pp 175–179
10. Beaubouef T, Petry F (2007) Intuitionistic Rough Sets for Database Applications. In: Peters JF et al (eds) Transactions on Rough Sets VI. LNCS, vol 4374. Springer, Berlin, pp 26–30
11. Beaubouef T, Petry F (2007) Rough Sets: A Versatile Theory for Approaches To Uncertainty Management in Databases. Rough Computing: Theories, Technologies and Applications, Idea Group, Inc.
12. Beaubouef T, Petry F, Arora G (1998) Information-Theoretic Measures of Uncertainty for Rough Sets and Rough Relational Databases. Inf Sci 109:185–195
13. Beaubouef T, Petry F, Buckles B (1995) Extension of the Relational Database and its Algebra with Rough Set Techniques. Comput Intell 11:233–245
14. Bhandari D, Pal NR (1993) Some New Information Measures for Fuzzy Sets. Inf Sci 67:209–228
15. Bittner T (2000) Rough sets in spatio-temporal data mining. Proceedings of International Workshop on Temporal, Spatial and Spatio–Temporal Data Mining. Springer, Berlin-Heidelberg, pp 89–104
16. Bittner T, Stell J (2003) Stratified Rough Sets and Vagueness. In: Kuhn W, Worboys M, Timpf S (eds) Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science International Conference (COSIT'03). Springer, Berlin, pp 286–303
17. Buckles B, Petry F (1983) Information-Theoretical Characterization of Fuzzy Relational Databases. IEEE Trans. Syst Man Cybern 13:74–77
18. Buckles BP, Petry F (1985) Uncertainty models in information and database systems. J Inf Sci 11:77–87
19. Chanas S, Kuchta D (1992) Further remarks on the relation between rough and fuzzy sets. Fuzzy Sets Syst 47:391–394
20. de Luca A, Termini S (1972) A Definition of a Nonprobabilistic Entropy in the Setting of Fuzzy Set Theory. Inf Control 20:301–312
21. Dubois D, Prade H (1987) Twofold Fuzzy Sets and Rough Sets–Some Issues in Knowledge Representation. Fuzzy Sets Syst 23:3–18
22. Dubois D, Prade H (1992) Putting Rough Sets and Fuzzy Sets Together. In: Slowinski R (ed) Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory, Kluwer Academic Publishers, Boston
23. Elmasri R, Navathe S (2004) Fundamentals of Database Systems, 4th edn. Addison Wesley, Boston
24. Frawley W, Piatetsky-Shapiro G, Matheus C (1991) Knowledge Discovery in Databases: An Overview. In: Piatetsky-Shapiro G, Frawley W (eds) Knowledge Discovery in Databases. AAAI, MIT Press, Cambridge, pp 1–27
25. Fung KT, Lam CM (1980) The Database Entropy Concept and its Application to the Data Allocation Problem. INFOR 18(4):354–363
26. Ginsburg S, Hull R (1983) Order Dependency in the Relational Model. Theor Comput Sci 26:146–195
27. Han J, Cai Y, Cercone N (1992) Knowledge Discovery in Databases: An Attribute–Oriented Approach, Proc. 18th VLDB Conf., Vancouver, Brit. Columbia, pp 547–559
28. Han J, Kamber M (2006) Data Mining: Concepts and Techniques, 2nd ed. Morgan Kaufman, San Diego, CA
29. Klir GJ, Folger TA (1988) Fuzzy Sets, Uncertainty, and Information. Prentice Hall, Englewood Cliffs NJ
30. Ligeza A (2002) Granular Sets and Granular Relation. Intelligent Information Systems, Physica Verlag, pp 331–340
31. Lin TY (1992) Topological and Fuzzy Rough Sets. In: Slowinski R (ed) Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory. Kluwer Academic Publishers, Boston, pp 287–304

32. Lin TY (1997) Granular Computing: From Rough Sets and Neighborhood Systems to Information Granulation and Computing in Words, European Congress on Intelligent Techniques and Soft Computing, September 8–12, pp 1602–1606
33. Lin TY (1999) Granular computing: Fuzzy Logic and Rough Sets, In: Zadeh L, Kacprzyk J (eds) Computing with Words in Information/ Intelligent Systems. Physica-Verlag, Heidelberg, pp 183–200
34. Makinouchi A (1977) A Consideration on normal form of not-necessarily normalized relation in the relational data model. Proc. of the 3rd Int. Conf. VLDB, pp 447–453
35. Mendel J, John R (2002) Type-2 Fuzzy Sets Made Simple. IEEE Trans on Fuzzy Sets 10:117–127
36. Nanda S, Majumdar S (1992) Fuzzy rough sets. Fuzzy Sets Syst 45:157–160
37. Ola A, Ozsoyoglu G (1993) Incomplete Relational Database Models Based on Intervals. IEEE Trans Knowl Data Eng 5:293–308
38. Pawlak Z (1982) Rough Sets. Int J Comput Inform Sci 11:341–356
39. Pawlak Z (1984) Rough Sets. Int J Man-Machine Stud 21:127–134
40. Pawlak Z (1985) Rough Sets and Fuzzy Sets. Fuzzy Sets Syst 17:99–102
41. Pawlak Z (1991) Rough Sets: Theoretical Aspects of Reasoning About Data. Kluwer Academic Publishers, Norwell, MA
42. Quinlan JR (1986) Induction of Decision Trees. Mach Learn 1:81–106
43. Raschia G, Mouaddib N (2002) SAINTETIQ:a fuzzy set-based approach to database summarization. Fuzzy Sets Syst 129:137–162
44. Roth M, Korth H, Batory D (1987) SQL/NF: A query language for non-1NF databases. Infor Syst 12:99–114
45. Shannon CL (1948) The Mathematical Theory of Communication. Bell Syst Tech J 27
46. Shenoi S, Melton A, Fan L (1992) Functional Dependencies and Normal Forms in the Fuzzy Relational Database Model. Infor Sci 60:1–28
47. Shi W, Wang S, Li D, Wang X (2003) Uncertainty-based Spatial Data Mining, Proceedings of Asia GIS Association, Wuhan, China, pp 124–35
48. Srinivasan P (1991) The importance of rough approximations for information retrieval. Int J Man-Mach Stud 34:657–671
49. Wang S, Li D, Shi W, Wang X (2002) Rough Spatial Description, International Archives of Photogrammetry and Remote Sensing, XXXII, Commission II, pp 503–510
50. Worboys M (1998) Computation with imprecise geospatial data. Comput Environ Urban Syst 22:85–106
51. Worboys M (1998) Imprecision in Finite Resolution Spatial Data. Geoinformatica 2:257–280
52. Wygralak M (1989) Rough Sets and Fuzzy Sets–Some Remarks on Interrelations. Fuzzy Sets Syst 29:241–243
53. Zadeh L (1965) Fuzzy Sets. Infor Control 8:338–353
54. Zvieli A, Chen P (1986) Entity–Relationship Modeling and Fuzzy Databases. Proc. of International Conference on Data Engineering, pp 320–327

Books and Reviews

Aczel J, Daroczy Z (1975) On Measures of Information and their Characterization. Academic Press, New York

Angryk R, Petry F (2007) Attribute-Oriented Fuzzy Generalization in Proximity and Similarity-based Relation Database Systems. Int J Intell Syst 22:763–781

Arora G, Petry F, Beaubouef T (1997) Information Measure of Type β Under Similarity Relations. Sixth IEEE International Conference on Fuzzy Systems Barcelona, pp 857–862

Arora G, Petry F, Beaubouef T (2001) A Note On New Parametric Measures of Information for Fuzzy Sets. J Comb Infor Syst Sci 26:167–174

Beaubouef T, Petry F (2001) Vague Regions and Spatial Relationships: A Rough Set Approach. Fourth International Conference on Computational Intelligence and Multimedia Applications, Yokosuka City, Japan, pp 313–318

Beaubouef T, Petry F (2001) Vagueness in Spatial Data: Rough Set and Egg-Yolk Approaches, 14th International Conference on Industrial & Engineering Applications of Artificial Intelligence, pp 367–373

Beaubouef T, Petry F (2003) Rough Set Uncertainty in an Object Oriented Data Model, Intelligent Systems for Information Processing. In: Bouchon B, Meunier L, Foulloy, Yager R (eds) From Representation to Applications. Elsevier, Amsterdam, pp 37–46

Beaubouef T, Petry F (2005) Normalization in a Rough Relational Database, International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, pp 257–265

Beaubouef T, Petry F (2005) Representation of Spatial Data in an OODB Using Rough and Fuzzy Set Modeling, Soft Comput J 9:364–373

Beaubouef T, Petry F (2007) An Attribute-Oriented Approach for Knowledge Discovery in Rough Relational Databases, Proc FLAIRS'07, pp 507–508

Beaubouef T, Ladner R, Petry F (2004) Rough Set Spatial Data Modeling for Data Mining. Int J Intell Syst 19:567–584

Beaubouef T, Petry F, Arora G (1998) Information Measures for Rough and Fuzzy Sets and Application to Uncertainty in Relational Databases. In: Pal S, Skowron A (eds) Rough-Fuzzy Hybridization: A New Trend in Decision-Making. Springer, Singapore, pp 200–214

Beaubouef T, Petry F, Ladner R (2007) Spatial Data Methods and Vague Regions: A Rough Set Approach. Appl Soft Comput J 7:425–440

Buckles B, Petry F (1982) Security and Fuzzy Databases, Procs 1982 IEEE Int Conf on Cybernetics and Society, pp 622–625

Codd E (1970) A relational model of data for large shared data banks. Comm ACM 13:377–387

Ebanks B (1983) On Measures of Fuzziness and their Representations. J Math Anal Appl 94:24–37

Grzymala-Busse J (1991) Managing Uncertainty in Expert Systems. Kluwer Academic Publishers, Boston

Han J, Nishio S, Kawano H, Wang W (1998) Generalization-Based Data Mining in Object-Oriented Databases Using an Object-Cube Model. Data Knowl Eng 25:55–97

Havrda J, Charvat F (1967) Quantification Methods of Classification Processes: Concepts of Structural α Entropy. Kybernetica 3:149–172

Kapur J, Kesavan H (1992) Entropy Optimization Principles with Applications. Academic Press, New York

Slowinski R (1992) A Generalization of the Indiscernibility Relation for Rough Sets Analysis of Quantitative Information. 1st Int. Workshop on Rough Sets: State of the Art and Perspectives, Poland, pp 41–48

Rough Set Data Analysis

SHUSAKU TSUMOTO
Department of Medical Informatics, Faculty of Medicine, Shimane University, Shimane, Japan

Article Outline

Introduction
Focusing Mechanism
Definition of Rules
Algorithms for Rule Induction
Experimental Results
What Is Discovered?
Rule Discovery as Knowledge Acquisition and Decision Support
Discussion
Conclusions
Bibliography

Introduction

Rule-induction methods are classified into two categories, induction of deterministic rules and of probabilistic ones [4,5,7,10].

On one hand, deterministic rules are described as *if-then* rules, which can be viewed as propositions. From the set-theoretical point of view, a set of examples supporting the conditional part of a deterministic rule, denoted by C, is a subset of a set whose examples belong to the consequence part, denoted by D. That is, the relation $C \subseteq D$ holds and deterministic rules are supported only by positive examples in a data set.

On the other hand, probabilistic rules are *if-then* rules with probabilistic information [10].

When a classical proposition will not hold for C and D, C is not a subset of D but closely overlapped with D. That is, the relations $C \cap D \neq \phi$ and $|C \cap D|/|C| \geq \delta$ will hold in this case, where the threshold δ is the degree of closeness of overlapping sets, which will be given by domain experts. (For more information, see Sect. "Definition of Rules")

Thus, probabilistic rules are supported by a large number of positive examples and a few negative examples.

The common feature of both deterministic and probabilistic rules is that they deduce their consequence positively if an example satisfies their conditional parts. We call the reasoning by these rules *positive reasoning*.

However, medical experts use not only positive reasoning but also *negative reasoning* for selection of candidates, which is represented as *if-then* rules whose consequences include negative terms.

For example, when a patient who complains of headache does not have a throbbing pain, migraine should not be suspected with a high probability. Thus, negative reasoning also plays an important role in cutting the search space of a differential diagnosis process [10]. Thus, medical reasoning includes both positive and negative reasoning, though conventional rule-induction methods do not reflect this aspect. This is one of the reasons medical experts have difficulty in interpreting induced rules, and interpreting rules for a discovery procedure does not proceed easily. Therefore, negative rules should be induced from databases in order not only to induce rules reflecting experts' decision processes, but also to induce rules that will be easier for domain experts to interpret, both of which are important to enhance the discovery process done by the cooperation of medical experts and computers.

In this chapter, first the characteristics of medical reasoning are discussed and then two kinds of rules, positive rules and negative rules, are introduced as a model of medical reasoning. Interestingly, from the set-theoretic point of view, sets of examples supporting both rules correspond to the lower and upper approximations in rough sets [5]. On the other hand, from the viewpoint of propositional logic, both positive and negative rules are defined as classical propositions or deterministic rules with two probabilistic measures, classification accuracy, and coverage.

Second, two algorithms for induction of positive and negative rules are introduced, defined as search procedures using accuracy and coverage as evaluation indices.

Finally, the proposed method is evaluated on several medical databases. The experimental results show that the induced rules correctly represent experts' knowledge. In addition, several interesting patterns are discovered.

Focusing Mechanism

One of the characteristics in medical reasoning is a focusing mechanism, which is used to select the final diagnosis from many candidates [10,11]. For example, in differential diagnosis of headache, more than 60 diseases will be checked by present history, physical examinations, and laboratory examinations. In diagnostic procedures, a candidate is excluded if a symptom necessary to diagnose is not observed.

This style of reasoning consists of the following two processes: exclusive reasoning and inclusive reasoning. Relations of this diagnostic model with another diagnostic model are discussed in [12]. The diagnostic procedure proceeds as follows (Fig. 2): First, exclusive reasoning ex-

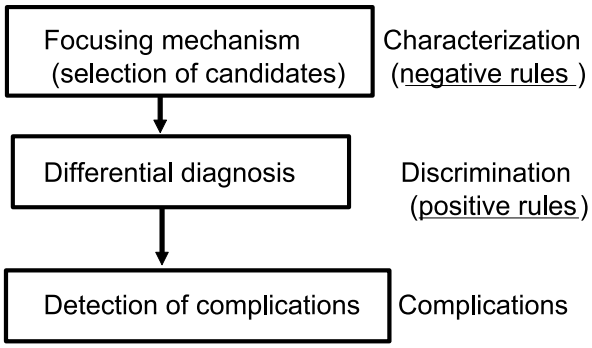

Rough Set Data Analysis, Figure 1
Illustration of focusing mechanism

cludes a disease from candidates when a patient does not have a symptom that is necessary to diagnose that disease.

Second, inclusive reasoning suspects a disease in the output of the exclusive process when a patient has symptoms specific to a disease.

These two steps are modeled as two kinds of rules, negative rules (or exclusive rules) and positive rules; the former corresponds to exclusive reasoning, the latter to inclusive reasoning.

In the next two sections, these two rules are represented as special kinds of probabilistic rules.

Definition of Rules

Rough Sets

In the following sections, we use the following notation introduced by Grzymala–Busse and Skowron [8], based on rough set theory [5].

These notations are illustrated by a small data set shown in Table 1, which includes symptoms exhibited by six patients who complained of headache.

Let U denote a nonempty finite set called the universe and A denote a nonempty, finite set of attributes, i. e., $a: U \to V_a$ for $a \in A$, where V_a is called the domain of a, respectively. Then a decision table is defined as an information system, $A = (U, A \cup \{d\})$. For example, Table 1 is an information system with $U = \{1, 2, 3, 4, 5, 6\}$ and $A = \{\text{age, location, nature, prodrome, nausea}, M1\}$ and $d = \text{class}$. For location $\in A$, V_{location} is defined as {occular, lateral, whole}.

The atomic formulas over $B \subseteq A \cup \{d\}$ and V are expressions of the form $[a = v]$, called descriptors over **B**, where $a \in B$ and $v \in V_a$. The set $F(B, V)$ of formulas over **B** is the least set containing all atomic formulas over B and closed with respect to disjunction, conjunction, and negation. For example, [location = occular] is a descriptor of B.

For each $f \in F(B, V)$, f_A denotes the meaning of f in A, i. e., the set of all objects in U with property f, defined inductively as follows:

1. If f is of the form $[a = v]$, then $f_A = \{s \in U | a(s) = v\}$
2. $(f \wedge g)_A = f_A \cap g_A$; $(f \vee g)_A = f_A \vee g_A$; $(\neg f)_A = U - f_a$.

For example, $f = [\text{location} = \text{whole}]$ and $f_A = \{2, 4, 5, 6\}$. As an example of a conjunctive formula, $g = [\text{location} = \text{whole}] \wedge [\text{nausea} = \text{no}]$ is a descriptor of U and f_A is equal to $g_{\text{location, nausea}} = \{2, 5\}$.

Classification Accuracy and Coverage

Definition of Accuracy and Coverage By use of the preceding framework, classification accuracy and coverage, or true positive rate are defined as follows.

Definition 1 Let R and D denote a formula in $F(B, V)$ and a set of objects that belong to a decision d. Classification accuracy and coverage (true positive rate) for $R \to d$

Rough Set Data Analysis, Table 1
An example of a data set

No.	Age	Location	Nature	Prodrome	Nausea	M1	Class
1	50–59	occular	persistent	no	no	yes	m.c.h.
2	40–49	whole	persistent	no	no	yes	m.c.h.
3	40–49	lateral	throbbing	no	yes	no	migra
4	40–49	whole	throbbing	yes	yes	no	migra
5	40–49	whole	radiating	no	no	yes	m.c.h.
6	50–59	whole	persistent	no	yes	yes	psycho

M1: tenderness of M1; m.c.h.: muscle contraction headache; migra: migraine; psycho: psychological pain

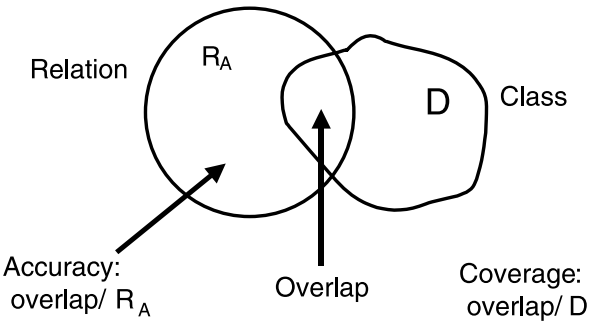

Rough Set Data Analysis, Figure 2
Venn diagram of accuracy and coverage

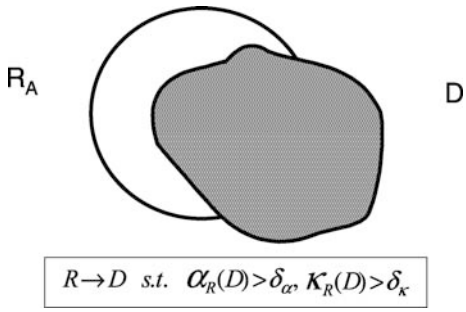

Rough Set Data Analysis, Figure 3
Venn diagram for probabilistic rules

is defined as:

$$\alpha_R(D) = \frac{|R_A \cap D|}{|R_A|} (= P(D|R)) \text{ and}$$
$$\kappa_R(D) = \frac{|R_A \cap D|}{|D|} (= P(R|D)),$$

where $|S|$, $\alpha_R(D)$, $\kappa_R(D)$, and $P(S)$ denote the cardinality of a set S, a classification accuracy of R as to classification of D, and coverage (a true positive rate of R to D), and probability of S, respectively.

Figure 2 depicts the Venn diagram of relations between accuracy and coverage. Accuracy views the overlapped region $|R_A \cap D|$ from the meaning of a relation R. On the other hand, coverage views the overlapped region from the meaning of a concept D.

In the preceding example, when R and D are set to [nau = yes] and [class = migraine], $\alpha_R(D) = 2/3 = 0.67$ and $\kappa_R(D) = 2/2 = 1.0$.

It is notable that $\alpha_R(D)$ measures the degree of the sufficiency of a proposition, $R \to D$, and that $\kappa_R(D)$ measures the degree of its necessity. For example, if $\alpha_R(D)$ is equal to 1.0, then $R \to D$ is true. On the other hand, if $\kappa_R(D)$ is equal to 1.0, then $D \to R$ is true. Thus, if both measures are 1.0, then $R \leftrightarrow D$.

Probabilistic Rules

By use of accuracy and coverage, a probabilistic rule is defined as:

$$R \overset{\alpha,\kappa}{\to} d \quad \text{s.t.} \quad R = \wedge_j [a_j = v_k], \quad \alpha_R(D) \geq \delta_\alpha \quad \text{and} \quad \kappa_R(D) \geq \delta_\kappa .$$

If the thresholds for accuracy and coverage are set to high values, the meaning of the conditional part of probabilistic rules corresponds to the highly overlapped region. Figure 3 depicts the Venn diagram of probabilistic rules with highly overlapped region.

This rule is a kind of probabilistic proposition with two statistical measures, which is an extension of Ziarko's variable precision model (VPRS) [15].[1]

It is also notable that both a positive rule and a negative rule are defined as special cases of this rule, as shown in the next sections.

Positive Rules

A positive rule is defined as a rule supported by only positive examples, the classification accuracy of which is equal to 1.0.

It is notable that the set supporting this rule corresponds to a subset of the lower approximation of a target concept, which is introduced in rough sets [5]. Thus, a positive rule is represented as:

$$R \to d \quad \text{s.t.} \quad R = \wedge_j [a_j = v_k], \quad \alpha_R(D) = 1.0 .$$

Figure 4 shows the Venn diagram of a positive rule. As shown in this figure, the meaning of R is a subset of that of D. This diagram is exactly equivalent to the classic proposition $R \to d$.

In the preceding example, one positive rule of m.c.h. (muscle contraction headache) is:

$$[\text{nausea} = \text{no}] \to \text{m.c.h.} \quad \alpha = 3/3 = 1.0 .$$

This positive rule is often called a deterministic rule. However, we use the term, positive (deterministic) rules, because a deterministic rule supported only by negative examples, called a negative rule, is introduced in the next section.

[1] This probabilistic rule is also a kind of *rough modus ponens* [6]

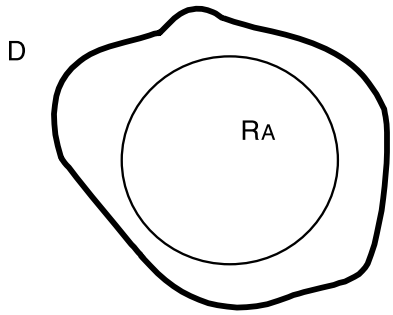

Rough Set Data Analysis, Figure 4
Venn diagram of positive rules

Negative Rules

Before defining a negative rule, let us first introduce an exclusive rule, the contrapositive of a negative rule [10]. An exclusive rule is defined as a rule supported by all the positive examples, the coverage of which is equal to 1.0. That is, an exclusive rule represents the necessity condition of a decision.

It is notable that the set supporting an exclusive rule corresponds to the upper approximation of a target concept, which is introduced in rough sets [5]. Thus, an exclusive rule is represented as:

$$R \rightarrow d \quad \text{s.t.} \qquad R = \vee_j [a_j = v_k]\,, \quad \kappa_R(D) = 1.0\,.$$

Figure 5 shows the Venn diagram of an exclusive rule. As shown in this figure, the meaning of R is a superset of that of D. This diagram is exactly equivalent to the classic proposition $d \rightarrow R$.

In the preceding example, the exclusive rule of m.c.h. is:

$$[M1 = \text{yes}] \vee [\text{nau} = \text{no}] \rightarrow \text{m.c.h.} \quad \kappa = 1.0\,,$$

From the viewpoint of propositional logic, an exclusive rule should be represented as:

$$d \rightarrow \vee_j [a_j = v_k]\,,$$

because the condition of an exclusive rule corresponds to the necessity condition of conclusion d. Thus, it is easy to see that a negative rule is defined as the contrapositive of an exclusive rule:

$$\wedge_j \neg [a_j = v_k] \rightarrow \neg d\,,$$

which means that if a case does not satisfy any attribute value pairs in the condition of a negative rule, then we can exclude a decision d from candidates.

For example, the negative rule of m.c.h. is:

$$\neg[M1 = \text{yes}] \wedge \neg[\text{nausea} = \text{no}] \rightarrow \neg\text{m.c.h.}$$

In summary, a negative rule is defined as:

$$\wedge_j \neg [a_j = v_k] \rightarrow \neg d \quad \text{s.t.} \quad \forall [a_j = v_k] \kappa_{[a_j = v_k]}(D) = 1.0\,,$$

where D denotes a set of samples that belong to a class d. Figure 6 shows the Venn diagram of a negative rule. As shown in this figure, it is notable that this negative region is the "positive region" of "negative concept."

Negative rules should also be included in a category of deterministic rules, because their coverage, a measure of negative concepts, is equal to 1.0. It is also notable that the set supporting a negative rule corresponds to a subset of negative region, which is introduced in rough sets [5].

In summary, positive and negative rules correspond to positive and negative regions defined in rough sets. Figure 7 shows the Venn diagram of those rules.

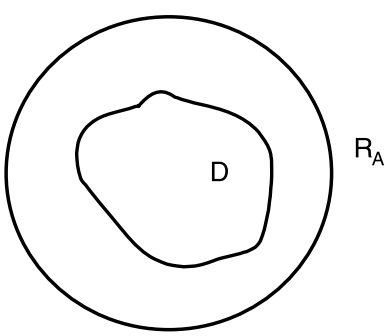

Rough Set Data Analysis, Figure 5
Venn diagram of exclusive rules

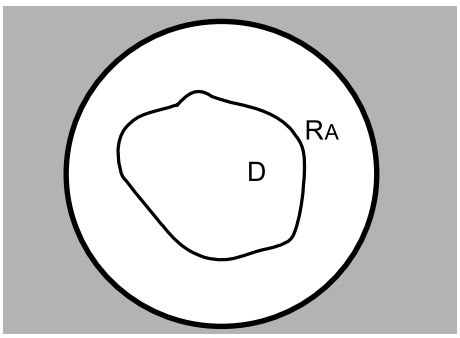

Rough Set Data Analysis, Figure 6
Venn diagram of negative rules

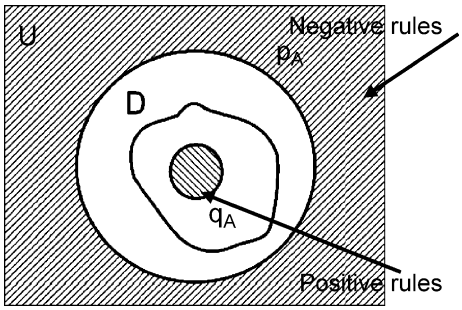

Rough Set Data Analysis, Figure 7
Venn diagram of defined rules

Algorithms for Rule Induction

The contrapositive of a negative rule, an exclusive rule, is induced as an exclusive rule by the modification of the algorithm introduced in PRIMEROSE-REX [10], as shown in Fig. 8. This algorithm works as follows.

(1) First it selects a descriptor $[a_i = v_j]$ from the list of attribute-value pairs, denoted by L.
(2) Then it checks whether this descriptor overlaps with a set of positive examples, denoted by D.
(3) If so, this descriptor is included in a list of candidates for positive rules and the algorithm checks whether its coverage is equal to 1.0. If the coverage is equal to 1.0, then this descriptor is added to R_er, the formula for the conditional part of the exclusive rule of D.
(4) Then $[a_i = v_j]$ is deleted from the list L. This procedure, from (1) to (4), will continue unless L is empty.
(5) Finally, when L is empty, this algorithm generates negative rules by taking the contrapositive of induced exclusive rules.

On the other hand, positive rules are induced as inclusive rules by the algorithm introduced in PRIMEROSE-REX [10], as shown in Fig. 9. For induction of positive rules, the threshold of accuracy and coverage is set to 1.0 and 0.0, respectively.

This algorithm works in the following way.

(1) First it substitutes L_1, which denotes a list of formulas composed of only one descriptor, with the list L_{er} generated by the former algorithm shown in Fig. 9.

```
procedure Exclusive and Negative Rules;
  var
    L : List;
      /* A list of elementary attribute-value pairs */
  begin
    L := P0;
    /* P0: A list of elementary attribute-value pairs given in a database */
    while (L ≠ {}) do
      begin
        Select one pair [ai = vj] from L;
        if ([ai = vj]A ∩ D ≠ ϕ) then do /* D: positive examples of a target class d */
          begin
          Lir := Lir + [ai = vj];    /* Candidates for Positive Rules */
          if (κ[ai=vj](D) = 1.0)
          then Rer := Rer ∧ [ai = vj];
            /* Include [ai = vj] into the formula of Exclusive Rule */
          end
      L := L − [ai = vj];
      end
    Construct Negative Rules:
        Take the contrapositive of Rer.
  end {Exclusive and Negative Rules};
```

Rough Set Data Analysis, Figure 8
Induction of exclusive and negative rules

```
procedure Positive Rules;
  var
    i : integer;    M, L_i : List;
  begin
    L_1 := L_ir;
    /* L_ir: A list of candidates generated by induction of exclusive rules */
    i := 1;    M := {};
    for i := 1 to n do
    /* n: Total number of attributes given
       in a database */
      begin
        while ( L_i ≠ {} ) do
          begin
            Select one pair R = ∧[a_i = v_j] from L_i;
            L_i := L_i − {R};
            if   (α_R(D) > δ_α)
              then  do S_ir := S_ir + {R};
          /* Include R in a list of the Positive Rules */
            else M := M + {R};
          end
        L_{i+1} := (A list of the whole combination of the conjunction formulae in M);
      end
  end {Positive Rules};
```

Rough Set Data Analysis, Figure 9
Induction of positive rules

(2) Then until L_1 becomes empty, the following steps will continue: (a) A formula $[a_i = v_j]$ is removed from L_1. (b) Then the algorithm checks whether $\alpha_R(D)$ is larger than the threshold. (For induction of positive rules, this is equal to checking whether $\alpha_R(D)$ is equal to 1.0.) If so, then this formula is included a list of the conditional parts of positive rules. Otherwise, it will be included in M, which is used for making conjunctions.
(3) When L_1 is empty, the next list L_2 is generated from the list M.

Experimental Results

For experimental evaluation, a new system, called PRIMEROSE-REX2 (Probabilistic Rule Induction Method for Rules of Expert System ver. 2.0), was developed, where the algorithms discussed in Sect. "Algorithms for Rule Induction" were implemented.

PRIMEROSE-REX2 was applied to the following three medical domains:

(1) Headache (RHINOS domain), whose training samples consist of 52,119 samples, 45 classes, and 147 attributes;
(2) Cerebulovasular diseases (CVD), whose training samples consist of 7620 samples, 22 classes, and 285 attributes; and
(3) Meningitis, whose training samples consist of 1211 samples, 4 classes, and 41 attributes (Table 2).

For evaluation, we used the following two types of experiments. One experiment was to evaluate the predictive accuracy using the cross-validation method, which is often used in the machine-learning literature [9]. The other ex-

Rough Set Data Analysis, Table 2
Databases

Domain	Samples	Classes	Attributes
Headache	52,119	45	147
CVD	7620	22	285
Meningitis	1211	4	41

periment was to evaluate the induced rules by medical experts and to check whether these rules lead to a new discovery.

Performance of Rules Obtained

For comparison of performance, the experiments are conducted by the following four procedures. First, rules are acquired manually from experts. Second, the data sets are randomly split into new training samples and new test samples. Third, PRIMEROSE-REX2, conventional rule-induction methods, AQ15 [4] and C4.5 [7] are applied to the new training samples for rule generation. Fourth, the induced rules and rules acquired from experts are tested using new test samples. The second through fourth steps are repeated 100 times, and the average classification accuracy over 100 trials is computed. This process is a variant of repeated two-fold cross-validation, introduced in [10].

Experimental results (performance) are shown in Table 3. The first and second rows show the results obtained using PRIMROSE-REX2; the results in the first row are derived using both positive and negative rules and those in the second row are derived by only positive rules. The third row shows the results derived from medical experts. For comparison, we compare the classification accuracy of C4.5 and AQ-15, which is shown in the fourth and fifth rows.

These results show that the combination of positive and negative rules outperforms positive rules, although it is a little worse than medical experts' rules.

What Is Discovered?

Interesting Rules Are Very Few

One of the most important observations is that there were very few rules interesting or unexpected to medical experts compared to the number of rules extracted from the data sets.

Table 4 shows the mentioned results. The second column denotes the number of positive and negative rules obtained from each data set. The third column denotes the number of positive and negative rules interesting or unexpected for domain experts. For example, the first row shows that the number of induced positive rules in headache is 24,335, but the number of interesting rules are 114, which shows that only 0.47% of rules are interesting for domain experts.

This table shows that the number of interesting rules are very few: even in the case of meningitis, only 6.5% are interesting, which suggests that the interpretation part of domain experts is very hard for the knowledge discovery process.

Next, we show several examples of rules that are interesting to domain experts.

Positive Rules in Differential Diagnosis of Headache

In the domain of differential diagnosis of headache, the following interesting positive rules were found.

$$[\text{Age} < 20] \wedge [\text{History:paroxysmal}] \rightarrow \text{Common Migraine(Coverage:0.75)}$$

This rule is said to be interesting if it is compared with the following rule:

$$[\text{Age} > 40] \wedge [\text{History:paroxysmal}] \rightarrow \text{Classic Migraine(Coverage:0.72)}$$

These two rules include two parts; although the values of the attribute "Age" are different, those of the attribute "History" are the same. This suggests that the attribute "Age" is important for differential diagnosis between common migraine and classic migraine.

Negative Rules in Differential Diagnosis of Headache

In the domain of differential diagnosis of headache, the following interesting negative rule was found.

$$\neg[\text{Nature:Persistent}] \wedge \neg[\text{History:acute}] \wedge \neg[\text{History:paroxysmal}]) \wedge \neg[\text{Neck Stiffness:yes}] \rightarrow \neg\text{Common Migraine}$$

It is notable that it is difficult even for domain experts to interpret the interestingness of this rule if only this rule is shown. The domain experts pointed out that the rule is interesting when compared with the following rules.

$$[\text{Nature:Persistent}] \wedge [\text{History:acute}] \wedge [\text{Neck Stiffness:yes}] \rightarrow \text{Meningitis}[\text{Nature: Persistent}] \wedge [\text{History:chronic}] \rightarrow \text{Brain Tumor}$$

This means that the rule includes information that is very important for differential diagnosis between common migraine and meningitis or brain tumor. Note that these two rules are very similar to the negative rule except for the negative symbols.

Rough Set Data Analysis, Table 3
Experimental results (accuracy: averaged)

Method	Headache	CVD	Meningitis
PRIMEROSE-REX2 (positive+negative)	91.3%	89.3%	92.5%
PRIMEROSE-REX2 (positive)	68.3%	71.3%	74.5%
Experts	95.0%	92.9%	93.2%
PR-REX	88.3%	84.3%	82.5%
C4.5	85.8%	79.7%	81.4%
AQ15	86.2%	78.9%	82.5%

Rough Set Data Analysis, Table 4
Number of extracted rules and interesting rules

	Induced positive rules	Interesting rules
Headache	24,335	114 (0.47%)
CVD	14,715	106 (0.72%)
Meningitis	1,922	40 (2%)

	Induced negative rules	Interesting rules
Headache	12,113	120 (0.99%)
CVD	7,231	155 (2.1%)
Meningitis	77	5 (6.5%)

Positive Rules in Meningitis

In the domain of meningitis, the following positive rules, which medical experts do not expect, are obtained.

$$[\text{WBC} < 12\,000] \wedge [\text{Sex} = \text{Female}] \wedge [\text{Age} < 40] \wedge [\text{CSF_CELL} < 1000] \rightarrow \text{Virus(Coverage:0.91)}$$

$$[\text{Age} \geq 40] \wedge [\text{WBC} \geq 8000] \wedge [\text{Sex} = \text{Male}] \wedge [\text{CSF_CELL} \geq 1000] \rightarrow \text{Bacteria(Coverage:0.64)}$$

The former rule means that if WBC (white blood cell count) is less than 12 000, the gender of a patient is female, the age is less than 40, and CSF_CELL (cell count of cerebulospinal fluid), then the type of meningitis is Viral. The latter means that the age of a patient is less than 40, WBC is larger than 8000, the gender is male, and CSF_CELL is larger than 1000, then the type of meningitis is Bacterial.

The most interesting points are that these rules included information about age and gender, which often seems to be unimportant attributes for differential diagnosis of meningitis. The first discovery was that women did not often suffer from bacterial infection compared with men, because such relationships between gender and meningitis has not been discussed in medical context [1]. By a close examination of the database on meningitis, it was found that most of the patients suffered from chronic diseases, such as DM, LC, and sinusitis, which are the risk factors of bacterial meningitis. The second discovery was that [age < 40] was also an important factor not to suspect viral meningitis, which also matches the fact that most old people suffer from chronic diseases. These results were also reevaluated in medical practice. Recently, the preceding two rules were checked by an additional 21 cases who suffered from meningitis (15 cases viral, 6 cases bacterial meningitis.) Surprisingly, the rules misclassified only three cases (two viral, the other bacterial), that is, the total accuracy was equal to $18/21 = 85.7\%$, and the accuracies for viral and bacterial meningitis were equal to $13/15 = 86.7\%$ and $5/6 = 83.3\%$. The reasons for misclassification are the following: a case of bacterial infection involved a patient who had a severe immunodeficiency, although he is very young. Two cases of viral infection involved patients who suffered from herpes zoster. It is notable that even those misclassified cases could be explained from the viewpoint of the immunodeficiency: that is, it was confirmed that immunodeficiency is a key factor for meningitis.

The validation of these rules is still ongoing; it will be reported in the near future.

Positive and Negative Rules in CVD

Concerning the database on CVD, several interesting rules were derived. The most interesting results were the following positive and negative rules for thalamus hemorrhage:

$$[\text{Sex} = \text{Female}] \wedge [\text{Hemiparesis} = \text{Left}] \wedge [\text{LOC:positive}] \rightarrow \text{Thalamus}$$

$$\neg[\text{Risk:Hypertension}] \wedge \neg[\text{Sensory} = \text{no}] \rightarrow \neg\text{Thalamus}$$

The former rule means that if the gender of a patient is female and he or she suffered from the left hemiparesis ([Hemiparesis=Left]) and loss of consciousness ([LOC:positive]), then the focus of CVD is thalamus. The latter rule means that if he or she suffers neither from hypertension ([Risk: Hypertension]) or sensory disturbance ([Sensory=no]), then the focus of CVD is thalamus.

Interestingly, LOC (loss of consciousness) under the condition of [Gender = Female] ∧ [Hemiparesis = Left] was found to be an important factor to diagnose thalamic damage. In this domain, any strong correlations between these attributes and others, like the database of meningitis, have not been found yet. It will be our future work to find what factor is behind these rules.

Rule Discovery as Knowledge Acquisition and Decision Support

Expert System: RH

Another point of discovery of rules is automated knowledge acquisition from databases. Knowledge acquisition is referred to as a bottleneck problem in development of expert systems [2], which has not fully been solved and is expected to be solved by induction of rules from databases. However, there are few papers that discuss the evaluation of discovered rules from the viewpoint of knowledge acquisition [12].

For this purpose, we have developed an expert system, called RH (rule-based system for headache) using the acquired knowledge. The reason for selecting the domain of headache is that earlier we developed an expert system RHINOS (rule-based headache information organizing system), which makes a differential diagnosis in headache [3]. In this system, it takes about six months to acquire knowledge from domain experts. RH consists of two parts. First, it requires inputs and applies exclusive and negative rules to select candidates (focusing mechanism). Then, it requires additional inputs and applies positive rules for differential diagnosis between selected candidates. Finally, RH outputs diagnostic conclusions.

Evaluation of RH

RH was evaluated in clinical practice with respect to its classification accuracy by using 930 patients who came to the outpatient clinic after the development of this system. Experimental results about classification accuracy are shown in Table 5. The first and second rows show the performance of rules obtained using PRIMROSE-REX2; the results in the first row are derived using both positive and negative rules and those in the second row are derived using only positive rules. The third and fourth rows show the results derived using both positive and negative rules and those by positive rules acquired directly from medical experts. These results show that the combination of positive and negative rules outperforms positive rules and gains almost the same performance as those by experts.

Rough Set Data Analysis, Table 5
Evaluation of RH (accuracy: averaged)

Method	Accuracy
PRIMEROSE-REX2 (positive and negative)	91.4% (851/930)
PRIMEROSE-REX2 (positive)	78.5% (729/930)
RHINOS (positive and negative)	93.5% (864/930)
RHINOS (positive)	82.8% (765/930)

Discussion

Hierarchical Rules for Decision Support

One of the problems with rule induction is that conventional rule-induction methods cannot extract rules that plausibly represent experts' decision processes [12]. The description length of induced rules is too short, compared to the experts' rules. (It may be observed that this length part does not contribute much to the classification performance.)

For example, rule-induction methods introduced in this chapter induced the following common rule for muscle contraction headache from databases on differential diagnosis of headache:

$$\begin{aligned}&[\text{location} = \text{whole}]\\&\quad \wedge [\text{Jolt Headache} = \text{no}]\\&\quad \wedge [\text{Tenderness of M1} = \text{yes}]\\&\rightarrow \text{muscle contraction headache}\end{aligned}$$

This rule is shorter than the following rule given by medical experts:

$$\begin{aligned}&[\text{Jolt Headache} = \text{no}]\\&\quad \wedge ([\text{Tenderness of M0} = \text{yes}]\\&\quad \vee [\text{Tenderness of M1} = \text{yes}]\\&\quad \vee [\text{Tenderness of M2} = \text{yes}])\\&\quad \wedge [\text{Tenderness of B1} = \text{no}]\\&\quad \wedge [\text{Tenderness of B2} = \text{no}]\\&\quad \wedge [\text{Tenderness of B3} = \text{no}]\\&\quad \wedge [\text{Tenderness of C1} = \text{no}]\\&\quad \wedge [\text{Tenderness of C2} = \text{no}]\\&\quad \wedge [\text{Tenderness of C3} = \text{no}]\\&\quad \wedge [\text{Tenderness of C4} = \text{no}]\\&\rightarrow \text{muscle contraction headache}\end{aligned}$$

These results suggest that conventional rule-induction methods do not reflect a mechanism of knowledge acquisition of medical experts.

Typically, rules acquired from medical experts are much longer than those induced from databases, the decision attributes of which are given by the same experts. This is because rule induction methods generally search for shorter rules, compared with decision tree induction. In the case of decision tree induction, the induced trees are sometimes too deep, and in order for the trees to be useful for learning, pruning and examination by experts are required. One of the main reasons rules are short and decision trees are sometimes long is that these patterns are generated by only one criteria, such as high accuracy or high information gain. The comparative study in this section suggests that experts should acquire rules by usage of several measures. Those characteristics of medical experts' rules are fully examined not by comparing between those rules for the same class, but by comparing experts' rules with those for another class. For example, a classification rule for muscle contraction headache is given by:

$$
\begin{aligned}
&[\text{Jolt Headache} = \text{no}] \\
&\quad \wedge ([\text{Tenderness of M0} = \text{yes}] \\
&\quad \vee [\text{Tenderness of M1} = \text{yes}] \\
&\quad \vee [\text{Tenderness of M2} = \text{yes}]) \\
&\quad \wedge [\text{Tenderness of B1} = \text{no}] \\
&\quad \wedge [\text{Tenderness of B2} = \text{no}] \\
&\quad \wedge [\text{Tenderness of B3} = \text{no}] \\
&\quad \wedge [\text{Tenderness of C1} = \text{no}] \\
&\quad \wedge [\text{Tenderness of C2} = \text{no}] \\
&\quad \wedge [\text{Tenderness of C3} = \text{no}] \\
&\quad \wedge [\text{Tenderness of C4} = \text{no}] \\
&\rightarrow \text{muscle contraction headache}
\end{aligned}
$$

This rule is very similar to the following classification rule for disease of cervical spine:

$$
\begin{aligned}
&[\text{Jolt Headache} = \text{no}] \\
&\quad \wedge ([\text{Tenderness of M0} = \text{yes}] \\
&\quad \vee [\text{Tenderness of M1} = \text{yes}] \\
&\quad \vee [\text{Tenderness of M2} = \text{yes}]) \\
&\quad \wedge ([\text{Tenderness of B1} = \text{yes}] \\
&\quad \vee [\text{Tenderness of B2} = \text{yes}] \\
&\quad \vee [\text{Tenderness of B3} = \text{yes}] \\
&\quad \vee [\text{Tenderness of C1} = \text{yes}] \\
&\quad \vee [\text{Tenderness of C2} = \text{yes}] \\
&\quad \vee [\text{Tenderness of C3} = \text{yes}] \\
&\quad \vee [\text{Tenderness of C4} = \text{yes}]) \\
&\rightarrow \text{disease of cervical spine}
\end{aligned}
$$

The differences between these two rules are attribute-value pairs, from tenderness of B1 to C4. Thus, these two rules can be simplified into the following form:

$$
\begin{aligned}
a_1 \wedge A_2 \wedge \neg A_3 &\rightarrow \textit{muscle contraction headache}, \\
a_1 \wedge A_2 \wedge A_3 &\rightarrow \textit{disease of cervical spine}.
\end{aligned}
$$

The first two terms and the third one represent different reasoning. The first and second terms a_1 and A_2 are used to differentiate muscle contraction headache and disease of cervical spine from other diseases. The third term A_3 is used to make a differential diagnosis between these two diseases. Thus, medical experts first select several diagnostic candidates, which are similar to each other, from many diseases and then make a final diagnosis from those candidates.

This problem has been partially solved; Tsumoto introduced a new approach for inducing these rules in [13], as induction of hierarchical decision rules. In that paper, the characteristics of experts' rules are closely examined and a new approach to extract plausible rules is introduced, which consists of the following three procedures. First, the characterization of decision attributes (given classes) is done from databases and the classes are classified into several groups with respect to the characterization. Then two kinds of subrules, characterization rules for each group and discrimination rules for each class in the group, are induced. Finally, those two parts are integrated into one rule for each decision attribute. The proposed method was evaluated on a medical database, the experimental results of which show that induced rules correctly represent experts' decision processes.

This observation also suggests that medical experts implicitly look at the relation between rules for different concepts. Future work should discover the relations between induced rules.

Relations Between Rules

In [14], Tsumoto focuses on the characteristics of medical reasoning(focusing mechanism) and introduces three kinds of rules, positive rules, exclusive rules and total covering rules, as a model of medical reasoning, which is an extended formalization of rules defined in [12].

Interestingly, from the set-theoretic point of view, sets of examples supporting these rules correspond to the lower and upper approximations in rough sets.

Furthermore, from the viewpoint of propositional logic, both inclusive and exclusive rules are defined as classical propositions, or deterministic rules with two probabilistic measures, classification accuracy and coverage. To-

tal covering rules have several interesting relations with inclusive and exclusive rules, which reflects the characteristics of medical reasoning.

Originally, a total covering rule is defined as a set of symptoms that can be observed in at least one case of a target disease. That is, this rule is defined as a collection of attribute-value pairs whose accuracy is larger than 0:

$$R \to d \quad \text{s.t.} \quad R = \vee_j [a_j = v_k], \quad \alpha_R(D) > 0\,.$$

From the definition of accuracy and coverage, this formula can be transformed into:

$$R \to d \quad \text{s.t.} \quad R = \vee_j [a_j = v_k], \quad \kappa_R(D) > 0\,.$$

For each attribute, the attribute-value pairs form a partition of U. Thus, for each attribute, total covering rules include a covering of all the positive examples. According to this property, the preceding formula is redefined as:

$$R \to d \quad \text{s.t.} \quad R = \vee_j R(a_j),$$
$$R(a_j) = \vee_k [a_j = v_k] \quad \text{s.t.} \quad \kappa_R(D) = 1.0\,.$$

It is notable that this definition is an extension of exclusive rules and this total covering rule can be written as:

$$d \to \vee_j \vee_k \, [a_j = v_k] \quad \text{s.t.} \quad \kappa_{[a_j = v_k]}(D) = 1.0\,.$$

Let $S(R)$ denote a set of attribute-value pairs of rule R. For each class d, let $R_{\text{pos}}(d)$, $R_{\text{ex}}(d)$, and $R_{\text{tc}}(d)$ denote the positive exclusive rule and total covering rules, respectively. Then

$$S(R_{\text{ex}}(d)) \subseteq S(R_{\text{tc}}(d))\,,$$

because a total covering rule can be viewed as an upper approximation of exclusive rules. It is also notable that this relation will hold in the relation between the positive rule (inclusive rule) and total covering rule. That is,

$$S(R_{\text{pos}}(d)) \subseteq S(R_{\text{tc}}(d))\,.$$

Thus, the total covering rule can be viewed as an upper approximation of inclusive rules. This relation also holds when $R_{\text{pos}}(d)$ is replaced with a probabilistic rule, which shows that total covering rules are the weakest form of diagnostic rules.

In this way, rules that reflect the diagnostic reasoning of medical experts have sophisticated background from the viewpoint of set theory. Especially, the rough set framework provides a good tool for modeling such focusing mechanisms.

Our future work will investigate the relation between these three types of rules from the viewpoint of rough set theory.

Conclusions

In this chapter, the characteristics of two measures, classification accuracy and coverage, are discussed, which show that both measures are dual and that accuracy and coverage are measures of both positive and negative rules, respectively. Then an algorithm for induction of positive and negative rules is introduced.

The proposed method is evaluated on medical databases. The experimental results have demonstrated that the induced rules are able to correctly represent experts' knowledge. We also demonstrated that the method can discover several interesting patterns.

Bibliography

1. Adams RD, Victor M (1993) Principles of Neurology, 5th edn. McGraw-Hill, New York
2. Buchnan BG, Shortliffe EH (eds) (1984) Rule-Based Expert Systems. Addison-Wesley
3. Matsumura Y, Matsunaga T, Hata Y, Kimura M, Matsumura H (1988) Consultation system for diagnoses of headache and facial pain. Med Inform RHINOS 11:145–157
4. Michalski RS, Mozetic I, Hong J, Lavrac N (1986) The multi-purpose incremental learning system AQ15 and its testing application to three medical domains. Proc 5th National Conference on Artificial Intelligence. AAAI Press, Palo Alto, pp 1041–1045
5. Pawlak Z (1991) Rough Sets. Kluwer, Dordrecht
6. Pawlak Z (1998) Rough Modus Ponens. In: Proc. International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems 98, Paris
7. Quinlan JR (1993) C4.5–Programs for Machine Learning. Morgan Kaufmann, Palo Alto
8. Skowron A, Grzymala-Busse J (1994) From rough set theory to evidence theory. In: Yager R, Fedrizzi M, Kacprzyk J (eds) Advances in the Dempster-Shafer Theory of Evidence. Wiley, New York, pp 193–236
9. Shavlik JW, Dietterich TG (eds) (1990) Readings in Machine Learning. Morgan Kaufmann, Palo Alto
10. Tsumoto S, Tanaka H (1996) Automated discovery of medical expert system rules from clinical databases based on rough sets. In: Proc. 2nd International Conference on Knowledge Discovery and Data Mining 96. AAAI Press, Palo Alto, pp 63–69
11. Tsumoto S (1998) Modelling medical diagnostic rules based on rough sets. In: Polkowski L, Skowron A (eds) Rough Sets and Current Trends in Computing. Lecture Note in Artificial Intelligence, vol 1424. Springer, Berlin
12. Tsumoto S (1998) Automated extraction of medical expert system rules from clinical databases based on rough set theory. J Inf Sci 112:67–84
13. Tsumoto S (1998) Extraction of experts' decision rules from clinical databases using rough set model. Intell Data Anal 2(3)
14. Tsumoto S (2001) Medical diagnostic rules as upper approximation of rough sets. Proc FUZZ-IEEE2001
15. Ziarko W (1993) Variable precision rough set model. J Comput Syst Sci 46:39–59

Rough Sets in Decision Making

Roman Słowiński[1,2], Salvatore Greco[3], Benedetto Matarazzo[3]
[1] Institute of Computing Science, Poznan University of Technology, Poznan, Poland
[2] Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
[3] Faculty of Economics, University of Catania, Catania, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Classical Rough Set Approach to Classification Problems of Taxonomy Type
Dominance-Based Rough Set Approach to Ordinal Classification Problems
DRSA on a Pairwise Comparison Table for Multiple Criteria Choice and Ranking Problems
DRSA for Decision Under Uncertainty
Multiple Criteria Decision Analysis Using Association Rules
Interactive Multiobjective Optimization Using DRSA (IMO-DRSA)
Conclusions
Future Directions
Bibliography

Glossary

Multiple attribute (or multiple criteria) decision support aims at giving the decision maker (DM) a recommendation concerning a set of *objects* A (also called alternatives, actions, acts, solutions, options, candidates, ...) evaluated from multiple points of view called *attributes* (also called features, variables, criteria, objectives, ...).

Main categories of *multiple attribute (or multiple criteria) decision problems*:

- *classification*, when the decision aims at assigning each object to one of predefined classes,
- *choice*, when the decision aims at selecting the best object,
- *ranking*, when the decision aims at ordering objects from the best to the worst.

Two kinds of *classification problems* are distinguished:

- *taxonomy*, when the value sets of attributes and the predefined classes are not preference ordered,
- *ordinal classification* (also known as *multiple criteria sorting*), when the value sets of attributes and the predefined classes are preference ordered.

Two kinds of *choice problems* are distinguished:

- *discrete choice*, when the set of objects is finite and reasonably small to be listed,
- *multiple objective optimization*, when the set of objects is infinite and defined by constraints of a mathematical program.

If value sets of attributes are preference-ordered, they are called *criteria* or *objectives*, otherwise they keep the name of attributes.

Criterion is a real-valued function g_i defined on A, reflecting a worth of objects from a particular point of view, such that in order to compare any two objects $a, b \in A$ from this point of view it is sufficient to compare two values: $g_i(a)$ and $g_i(b)$.

Dominance Object a is non-dominated in set A (Pareto-optimal) if and only if there is no other object b in A such that b is not worse than a on all considered criteria, and strictly better on at least one criterion.

Preference model is a representation of a value system of the decision maker on the set of objects with vector evaluations.

Decision under uncertainty takes into account consequences of decisions that distribute over multiple states of nature with given probability. The preference order, characteristic for data describing *multiple attribute decision problems*, concerns also decision under uncertainty, where the objects correspond to acts, attributes are outcomes (gain or loss) to be obtained with a given probability, and the problem consists in ordinal classification, choice, or ranking of the acts.

Rough set in universe U is an approximation of a set based on available information about objects of U. The rough approximation is composed of two ordinary sets called *lower and upper approximation*. Lower approximation is a maximal subset of objects which, according to the available information, certainly belong to the approximated set, and upper approximation is a minimal subset of objects which, according to the available information, possibly belong to the approximated set. The difference between upper and lower approximation is called *boundary*.

Decision rule is a logical statement of the type "*if* ..., *then* ...", kursiv where the premise (condition part)

specifies values assumed by one or more condition attributes and the conclusion (decision part) specifies an overall judgment.

Definition of the Subject

Scientific analysis of decision problems aims at giving the decision maker (DM) a recommendation concerning a set of *objects* (also called alternatives, solutions, acts, actions, options, candidates, ...) evaluated from multiple points of view considered relevant for the problem at hand and called *attributes* (also called features, variables, criteria, objectives, ...). For example, a decision can concern:

1) diagnosis of pathologies for a set of patients, where patients are objects of the decision, and symptoms and results of medical tests are the attributes,
2) assignment of enterprises to classes of risk, where enterprises are objects of the decision, and financial ratio indices and other economic indicators, such as the market structure, the technology used by the enterprise and the quality of management, are the attributes,
3) selection of a car to be bought from among a given set of cars, where cars are objects of the decision, and maximum speed, acceleration, price, fuel consumption, comfort, color and so on, are the attributes,
4) ordering of students applying for a scholarship, where students are objects of the decision, and scores in different subjects are the attributes.

The following three main categories of decision problems are typically distinguished [44]:

- *classification*, when the decision aims at assigning each object to one of predefined classes,
- *choice*, when the decision aims at selecting the best objects,
- *ranking*, when the decision aims at ordering objects from the best to the worst.

Looking at the above examples, one can say that 1) and 2) are classification problems, 3) is a choice problem and 4) is a ranking problem.

The above categorization can be refined by distinguishing two kinds of classification problems: *taxonomy*, when the value sets of attributes and the predefined classes are not preference ordered, and *ordinal classification* (also known as *multiple criteria sorting*), when the value sets of attributes and the predefined classes are preference ordered [12]. In the above examples, 1) is a taxonomy problem and 2) is an ordinal classification problem. If value sets of attributes are preference ordered, they are called *criteria*, otherwise they keep the name of attributes. For example, in a decision regarding the selection of a car, its price is a criterion because, obviously, a low price is better than a high price. Instead, the color of a car is not a criterion but simply an attribute, because red is not intrinsically better than green. One can imagine, however, that also the color of a car could become a criterion if, for example, a DM would consider red better than green.

Introduction

Scientific support of decisions makes use of a more or less explicit model of the decision problem. The model relates the decision to the characteristics of the objects expressed by the considered attributes. Building such a model requires information about conditions and parameters of the aggregation of multi-attribute characteristics of objects. The nature of this information depends on the adopted methodology: prices and interest rates for cost-benefit analysis, cost coefficients in objectives and technological coefficients in constraints for mathematical programming, a training set of decision examples for neural networks and machine learning, substitution rates for a value function of multi-attribute utility theory, pairwise comparisons of objects in terms of intensity of preference for the analytic hierarchy process, attribute weights and several thresholds for ELECTRE methods, and so on (see the state-of-the-art survey [4]). This information has to be provided by the DM, possibly assisted by an analyst.

Very often this information is not easily definable. For example, this is the case of the price of many immaterial goods and of the interest rates in cost-benefit analysis, or the case of the coefficients of objectives and constraints in mathematical programming models. Even if the required information is easily definable, like a training set of decision examples for neural networks, it is often processed in a way which is not clear for the DM, such that (s)he cannot see what are the exact relations between the provided information and the final recommendation. Consequently, very often the decision model is perceived by the DM as a *black box* whose result has to be accepted because the analyst's authority guarantees that the result is "right". In this context, the aspiration of the DM to find good reasons to make decision is frustrated and rises the need for a more transparent methodology in which the relation between the original information and the final recommendation is clearly shown. Such a transparent methodology searched for has been called *glass box* [32]. Its typical representative is using a learning set of decision examples as the input preference information provided by the DM, and

it is expressing the decision model in terms of a set of "*if* ..., *then* ..." decision rules induced from the input information. From one side, the decision rules are explicitly related to the original information and, from the other side, they give understandable justifications for the decisions to be made.

For example, in case of a medical diagnosis problem, the decision rule approach requires as input information a set of examples of previous diagnoses, from which some diagnostic rules are induced, such as "if there is symptom α and the test result is β, then there is pathology γ". Each one of such rules is directly related to examples of diagnoses in the input information, where there is symptom α, test result β and pathology γ. Moreover, the DM can verify easily that in the input information there is no example of diagnosis where there is symptom α, test result β but no pathology γ.

The rules induced from the input information provided in terms of exemplary decisions represent a decision model which is transparent for the DM and enables his understanding of the reasons of his past decisions. The acceptance of the rules by the DM justifies, in turn, their use for future decisions.

The induction of rules from examples is a typical approach of artificial intelligence. This explains our interest in rough set theory [38,39] which proved to be a useful tool for analysis of vague description of decision situations [41,48]. The rough set analysis aims at explaining the values of some decision attributes, playing the role of "dependent variables", by means of the values of condition attributes, playing the role of "independent variables". For example, in the above diagnostic context, data about the presence of a pathology are given by decision attributes, while data about symptoms and tests are given by condition attributes. An important advantage of the rough set approach is that it can deal with partly inconsistent examples, i. e. cases where the presence of different pathologies is associated with the presence of the same symptoms and test results. Moreover, it provides useful information about the role of particular attributes and their subsets, and prepares the ground for representation of knowledge hidden in the data by means of "*if* ..., *then* ..." decision rules.

Classical rough set approach proposed by Pawlak [38,39] cannot deal, however, with preference order in the value sets of condition and decision attributes. For this reason, the classical rough set approach can deal with only one of four decision problems listed above – classification of taxonomy type. To deal with ordinal classification, choice and ranking, it is necessary to generalize the classical rough set approach, so as to take into account preference orders and monotonic relationships between condition and decision attributes. This generalization, called Dominance-based Rough Set Approach (DRSA), has been proposed by Greco, Matarazzo and Slowinski [12,14,15,18,21,49].

Classical Rough Set Approach to Classification Problems of Taxonomy Type

Information Table and Indiscernibility Relation

Information regarding classification examples is supplied in the form of an *information table*, whose separate rows refer to distinct *objects*, and whose columns refer to different *attributes* considered. This means that each cell of this table indicates an *evaluation* (quantitative or qualitative) of the object placed in the corresponding row by means of the attribute in the corresponding column. Formally, an information table is the 4-tuple $\boldsymbol{S} = \langle U, Q, V, v\rangle$, where U is a finite set of objects, called *universe*, $Q = \{q_1, \ldots, q_n\}$ is a finite set of attributes, V_q is a value set of the attribute q, $V = \bigcup_{q\in Q} V_q$, and $v\colon U \times Q \to V$ is a total function such that $v(x, q) \in V_q$ for each $q \in Q$, $x \in U$, called *information function*.

Therefore, each object x of U is described by a vector (string) $\mathrm{Des}_Q(x) = [v(x, q_1), \ldots, v(x, q_n)]$, called *description* of x in terms of the evaluations on the attributes from Q; it represents the available information about x. Obviously, $x \in U$ can be described in terms of any non-empty subset $P \subseteq Q$.

To every (non-empty) subset of attributes $P \subseteq Q$ there is associated an *indiscernibility relation* on U, denoted by I_P:

$$I_P = \{(x, y) \in U \times U\colon\ v(x, q) = v(y, q)\,,\ \forall q \in P\}\,.$$

If $(x, y) \in I_P$, it is said that the objects x and y are P-indiscernible. Clearly, the indiscernibility relation thus defined is an equivalence relation (reflexive, symmetric and transitive). The family of all the equivalence classes of the relation I_P is denoted by U/I_P, and the equivalence class containing object $x \in U$, by $I_P(x)$. The equivalence classes of the relation I_P are called *P-elementary sets*.

Approximations

Let $\boldsymbol{S}$ be an information table, X a non-empty subset of U and $\emptyset \neq P \subseteq Q$. The *P-lower approximation* and the *P-upper approximation* of X in $\boldsymbol{S}$ are defined, respectively, as:

$$\underline{P}(X) = \{x \in U\colon I_P(x) \subseteq X\}\ ,$$

$$\overline{P}(X) = \{x \in U\colon I_P(x) \cap X \neq \emptyset\}\ .$$

The elements of $\underline{P}(X)$ are all and only those objects $x \in U$ which belong to the equivalence classes generated by the indiscernibility relation I_P, *contained* in X; the elements of $\overline{P}(X)$ are all and only those objects $x \in U$ which belong to the equivalence classes generated by the indiscernibility relation I_P, *containing at least one* object x belonging to X. In other words, $\underline{P}(X)$ is the largest union of the P-elementary sets included in X, while $\overline{P}(X)$ is the smallest union of the P-elementary sets containing X.

The *P-boundary* of X in $\mathbf{S}$, denoted by $Bn_P(X)$, is defined by

$$Bn_P(X) = \overline{P}(X) - \underline{P}(X) \ .$$

The following *inclusion relation* holds:

$$\underline{P}(X) \subseteq X \subseteq \overline{P}(X) \ .$$

Thus, in view of information provided by P, if an object x belongs to $\underline{P}(X)$, then it certainly belongs to set X, while if x belongs to $\overline{P}(X)$, then it possibly belongs to set X. $Bn_P(X)$ constitutes the "doubtful region" of X: nothing can be said with certainty about the membership of its elements to set X, using the subset of attributes P only.

Moreover, the following *complementarity relation* is satisfied:

$$\underline{P}(X) = U - \overline{P}(U - X) \ .$$

If the P-boundary of set X is empty, $Bn_P(X) = \emptyset$, then X is an ordinary (exact) set with respect to P, that is, it may be expressed as union of a certain number of P-elementary sets; otherwise, if $Bn_P(X) \neq \emptyset$, set X is an approximate (rough) set with respect to P and may be characterized by means of the approximations $\underline{P}(X)$ and $\overline{P}(X)$. The family of all sets $X \subseteq U$ having the same P-lower and P-upper approximations is called the *rough set*.

The *quality* of the approximation of set X by means of the attributes from P is defined as

$$\gamma_P(X) = \frac{|\underline{P}(X)|}{|X|} \ ,$$

such that $0 \leq \gamma_P(X) \leq 1$. The quality γ_P(X) represents the relative frequency of the objects correctly classified using the attributes from P.

The definition of approximations of a set $X \subseteq U$ can be extended to a classification, i. e. a partition $Y = \{Y_1, \ldots, Y_m\}$ of U. The P-lower and P-upper approximations of Y in $\mathbf{S}$ are defined by sets $\underline{P}(Y) = \{\underline{P}(Y_1), \ldots, \underline{P}(Y_m)\}$ and $\overline{P}(Y) = \{\overline{P}(Y_1), \ldots, \overline{P}(Y_m)\}$, respectively. The coefficient

$$\gamma_P(Y) = \frac{\sum_{i=1}^{m} |\underline{P}(Y_i)|}{|U|}$$

is called *quality of the approximation of classification* Y by set of attributes P, or in short, *quality of classification*. It expresses the ratio of all P-correctly classified objects to all objects in the system.

Dependence and Reduction of Attributes

An issue of great practical importance is the reduction of "superfluous" attributes in an information table. Superfluous attributes can be eliminated without deteriorating the information contained in the original table.

Let $P \subseteq Q$ and $p \in P$. It is said that attribute p is *superfluous* in P with respect to classification Y if $\underline{P}(Y) = \underline{(P-p)}(Y)$; otherwise, p is *indispensable* in P. The subset of Q containing all the indispensable attributes is known as the *core*.

Given classification Y, any minimal (with respect to inclusion) subset $P \subseteq Q$, such that $\underline{P}(Y) = \underline{Q}(Y)$, is called *reduct*. It specifies a minimal subset P of Q which keeps the quality of classification at the same level as the whole set of attributes, i. e. $\gamma_P(Y) = \gamma_Q(Y)$. In other words, the attributes that do not belong to the reduct are superfluous with respect to the classification Y of objects from U.

More than one reduct may exist in an information table and their intersection gives the *core*.

Decision Table and Decision Rules

In the information table describing examples of classification, the attributes of set Q are divided into *condition* attributes (set $C \neq \emptyset$) and *decision* attributes (set $D \neq \emptyset$), $C \cup D = Q$ and $C \cap D = \emptyset$. Such an information table is called a *decision table*. The decision attributes induce a partition of U deduced from the indiscernibility relation I_D in a way that is independent of the condition attributes. D-elementary sets are called *decision classes*, denoted by Cl_t, $t = 1, \ldots, m$. The partition of U into decision classes is called classification $\mathbf{Cl} = \{Cl_1, \ldots, Cl_m\}$. There is a tendency to reduce the set C while keeping all important relationships between C and D, in order to make decisions on the basis of a smaller amount of information. When the set of condition attributes is replaced by one of its reducts, the quality of approximation of the classification induced by the decision attributes is not deteriorating.

Since the aim is to underline the functional dependencies between condition and decision attributes, a decision table may also be seen as a set of *decision rules*. These are logical statements of the type "*if* ..., *then* ...", where the premise (condition part) specifies values assumed by one or more condition attributes (description of C-elementary sets) and the conclusion (decision part) specifies an assign-

ment to one or more decision classes. Therefore, the syntax of a rule is the following:

> "*if* $v(x, q_1) = r_{q1}$ *and* $v(x, q_2) = r_{q2}$ *and* ... $v(x, q_p) = r_{qp}$, *then* x belongs to decision class Cl_{j1} *or* Cl_{j2} *or* ... Cl_{jk}",

where $\{q_1, q_2, \ldots, q_p\} \subseteq C, (r_{q1}, r_{q2}, \ldots, r_{qp}) \in V_{q1} \times V_{q2} \times \cdots \times V_{qp}$ and $Cl_{j1}, Cl_{j2}, \ldots, Cl_{jk}$ are some decision classes of the considered classification $\boldsymbol{Cl}$. If the consequence is univocal, i. e. $k = 1$, then the rule is *univocal*, otherwise it is *approximate*.

An object $x \in U$ *supports* decision rule r if its description is matching both the condition part and the decision part of the rule. The decision rule r *covers* object x if it matches the condition part of the rule. Each decision rule is characterized by its *strength*, defined as the number of objects supporting the rule. In the case of approximate rules, the strength is calculated for each possible decision class separately.

If a univocal rule is supported by objects from the lower approximation of the corresponding decision class only, then the rule is called *certain* or *deterministic*. If, however, a univocal rule is supported by objects from the upper approximation of the corresponding decision class only, then the rule is called *possible* or *probabilistic*. Approximate rules are supported, in turn, only by objects from the boundaries of the corresponding decision classes.

Procedures for generation of decision rules from a decision table use an inductive learning principle. The objects are considered as examples of classification. In order to induce a decision rule with a univocal and certain conclusion about assignment of an object to decision class X, the examples belonging to the C-lower approximation of X are called *positive* and all the others *negative*. Analogously, in case of a possible rule, the examples belonging to the C-upper approximation of X are positive and all the others negative. Possible rules are characterized by a coefficient, called *confidence*, telling to what extent the rule is consistent, i. e. what is the ratio of the number of positive examples supporting the rule to the number of examples belonging to set X according to decision attributes. Finally, in case of an approximate rule, the examples belonging to the C-boundary of X are positive and all the others negative. A decision rule is called *minimal* if removing any attribute from the condition part gives a rule covering also negative objects.

The existing induction algorithms use one of the following strategies [55]:

(a) Generation of a *minimal* representation, i. e. minimal set of rules covering all objects from a decision table,
(b) Generation of an *exhaustive* representation, i. e. all rules for a given decision table,
(c) Generation of a *characteristic* representation, i. e. a set of rules covering relatively many objects, however, not necessarily all objects from a decision table.

Explanation of the Classical Rough Set Approach by an Example

Suppose that one wants to describe the classification of basic traffic signs to a novice. There are three main classes of traffic signs corresponding to:

- Warning (W),
- Interdiction (I),
- Order (O).

Then, these classes may be distinguished by such attributes as the shape (S) and the principal color (PC) of the sign. Finally, one can consider a few examples of traffic signs, like those shown in Table 1. These are:

a) Sharp right turn,
b) Speed limit of 50 km/h,
c) No parking,
d) Go ahead.

The rough set approach is used here to build a model of classification of traffic signs to classes W, I, O on the basis of attributes S and PC. This is a typical problem of taxonomy.

One can remark that the sets of signs indiscernible by "Class" are:

$$\mathrm{W} = \{a\}_{\mathrm{Class}}, \quad \mathrm{I} = \{b, c\}_{\mathrm{Class}}, \quad \mathrm{O} = \{d\}_{\mathrm{Class}},$$

Rough Sets in Decision Making, Table 1
Examples of traffic signs described by S and PC

Traffic sign	Shape (S)	Primary color (PC)	Class
a)	triangle	yellow	W
b)	circle	white	I
c)	circle	blue	I
d)	circle	blue	O

and the sets of signs indiscernible by S and PC are as follows:

$$\{a\}_{\mathrm{S,PC}}\,, \quad \{b\}_{\mathrm{S,PC}}\,, \quad \{c,d\}_{\mathrm{S,PC}}\,.$$

The above elementary sets are generated, on the one hand, by decision attribute "Class" and, on the other hand, by condition attributes S and PC. The elementary sets of signs indiscernible by "Class" are denoted by $\{\cdot\}_{\mathrm{Class}}$ and those by S and PC are denoted by $\{\cdot\}_{\mathrm{S,PC}}$. Notice that $\mathrm{W} = \{a\}_{\mathrm{Class}}$ is characterized precisely by $\{a\}_{\mathrm{S,PC}}$. In order to characterize $\mathrm{I} = \{b,c\}_{\mathrm{Class}}$ and $\mathrm{O} = \{d\}_{\mathrm{Class}}$, one needs $\{b\}_{\mathrm{S,PC}}$ and $\{c,d\}_{\mathrm{S,PC}}$, however, only $\{b\}_{\mathrm{S,PC}}$ is included in $\mathrm{I} = \{b,c\}_{\mathrm{Class}}$ while $\{c,d\}_{\mathrm{S,PC}}$ has a non-empty intersection with both $\mathrm{I} = \{b,c\}_{\mathrm{Class}}$ and $\mathrm{O} = \{d\}_{\mathrm{Class}}$. It follows, from this characterization, that by using condition attributes S and PC, one can characterize class W precisely, while classes I and O can only be characterized approximately:

- Class W includes sign a certainly and possibly no other sign than a,
- Class I includes sign b certainly and possibly signs b, c and d,
- Class O includes no sign certainly and possibly signs c and d.

The terms "*certainly*" and "*possibly*" refer to the absence or presence of ambiguity between the description of signs by S and PC from the one side, and by "Class", from the other side. In other words, using description of signs by S and PC, one can say that all signs from elementary sets $\{\cdot\}_{\mathrm{S,PC}}$ included in elementary sets $\{\cdot\}_{\mathrm{Class}}$ belong certainly to the corresponding class, while all signs from elementary sets $\{\cdot\}_{\mathrm{S,PC}}$ having a non-empty intersection with elementary sets $\{\cdot\}_{\mathrm{Class}}$ belong to the corresponding class only possibly. The two sets of certain and possible signs are, respectively, the *lower* and *upper approximation* of the corresponding class by attributes S and PC:

$$\begin{aligned}
&\text{lower_approx.}_{\mathrm{S,PC}}(\mathrm{W}) = \{a\}, && \text{upper_approx.}_{\mathrm{S,PC}}(\mathrm{W}) = \{a\},\\
&\text{lower_approx.}_{\mathrm{S,PC}}(\mathrm{I}) = \{b\}, && \text{upper_approx.}_{\mathrm{S,PC}}(\mathrm{I}) = \{b,c,d\},\\
&\text{lower_approx.}_{\mathrm{S,PC}}(\mathrm{O}) = \emptyset, && \text{upper_approx.}_{\mathrm{S,PC}}(\mathrm{O}) = \{c,d\}.
\end{aligned}$$

The *quality of approximation* of the classification by attributes S and PC is equal to the number of all the signs in the lower approximations divided by the number of all the signs in the table, i. e. 1/2.

Rough Sets in Decision Making, Table 2
Examples of traffic signs described by S, PC and SC

Traffic sign	Shape (S)	Primary color (PC)	Secondary color (SC)	Class
a)	triangle	yellow	red	W
b)	circle	white	red	I
c)	circle	blue	red	I
d)	circle	blue	white	O

One way to increase the quality of the approximation is to add a new attribute so as to decrease the ambiguity. Let us introduce the secondary color (SC) as a new condition attribute. The new situation is now shown in Table 2.

As one can see, the sets of signs indiscernible by S, PC and SC, i. e. the elementary sets $\{\cdot\}_{\mathrm{S,PC,SC}}$, are now:

$$\{a\}_{\mathrm{S,PC,SC}}\,, \quad \{b\}_{\mathrm{S,PC,SC}}\,, \quad \{c\}_{\mathrm{S,PC,SC}}\,, \quad \{d\}_{\mathrm{S,PC,SC}}\,.$$

It is worth noting that the elementary sets are finer than before and this enables the ambiguity to be eliminated. Consequently, the quality of approximation of the classification by attributes S, PC and SC is now equal to 1.

A natural question occurring here is to ask if, indeed, all three attributes are necessary to characterize precisely the classes W, I and O. When attribute S or attribute PC is eliminated from the description of the signs, the elementary sets $\{\cdot\}_{\mathrm{PC,SC}}$ or $\{\cdot\}_{\mathrm{S,SC}}$ are defined, respectively, as follows:

$$\{a\}_{\mathrm{PC,SC}}\,, \quad \{b\}_{\mathrm{PC,SC}}\,, \quad \{c\}_{\mathrm{PC,SC}}\,, \quad \{d\}_{\mathrm{PC,SC}}\,,$$

$$\{a\}_{\mathrm{S,SC}}\,, \quad \{b,c\}_{\mathrm{S,SC}}\,, \quad \{d\}_{\mathrm{S,SC}}\,.$$

Using any one of the above elementary sets, it is possible to characterize (approximate) classes W, I and O with the same quality (equal to 1) as it is when using the elementary sets $\{\cdot\}_{\mathrm{S,PC,SC}}$ (i. e. those generated by the complete set of three condition attributes). Thus, the answer to the above question is that the three condition attributes are not all necessary to characterize precisely the classes W, I and O. It is, in fact, sufficient to use either PC and SC or S and SC. The subsets of condition attributes {PC, SC} and {S, SC} are called *reducts* of {S, PC, SC} because they have

this property. Note that the identification of reducts enables us to reduce attributes about the signs from the table to only those which are relevant.

Other useful information can be generated from the identification of reducts by taking their intersection. This is called the *core*. In our example, the core contains attribute SC. This tells us that it is clearly an *indispensable* attribute i. e. it cannot be eliminated from the description of the signs without decreasing the quality of the approximation. Note that other attributes from the reducts (i. e. S and PC) are *exchangeable*. If there happened to be some other attributes which were not included in any reduct, then they would be *superfluous*, i. e. they would not be useful at all in the characterization of the classes W, I and O.

If, however, column S or PC is eliminated from Table 2, then the resulting table is not a minimal representation of knowledge about the classification of the four traffic signs. Note that, in order to characterize class W in Table 2, it is sufficient to use the condition "S = triangle". Moreover, class I is characterized by two conditions ("S = circle" and "SC = red") and class O is characterized by the condition "SC = white". Thus, the minimal representation of this information system requires only four conditions (rather than the eight conditions that are presented in Table 2 with either column S or PC eliminated). This representation corresponds to the following set of *decision rules* which may be seen as classification model discovered in the data set contained in Table 2 (in the braces there are symbols of signs covered by the corresponding rule):

$$\begin{array}{lll} \textit{rule}\ \#1: \textit{if}\ \mathrm{S} = \text{triangle}, & \textit{then}\ \mathrm{Class} = \mathrm{W} & \{a\} \\ \textit{rule}\ \#2: \textit{if}\ \mathrm{S} = \text{circle} & & \\ \qquad \textit{and}\ \ \mathrm{SC} = \text{red}, & \textit{then}\ \mathrm{Class} = \mathrm{I} & \{b, c\} \\ \textit{rule}\ \#3: \textit{if}\ \mathrm{SC} = \text{white}, & \textit{then}\ \mathrm{Class} = \mathrm{O} & \{d\}\,. \end{array}$$

This is not the only representation, because an alternative set of rules is:

$$\begin{array}{lll} \textit{rule}\ \#1': \textit{if}\ \mathrm{PC} = \text{yellow}, & \textit{then}\ \mathrm{Class} = \mathrm{W} & \{a\} \\ \textit{rule}\ \#2': \textit{if}\ \mathrm{PC} = \text{white}, & \textit{then}\ \mathrm{Class} = \mathrm{I} & \{b\} \\ \textit{rule}\ \#3': \textit{if}\ \mathrm{PC} = \text{blue}, & & \\ \qquad \textit{and}\ \ \mathrm{SC} = \text{red}, & \textit{then}\ \mathrm{Class} = \mathrm{I} & \{c\} \\ \textit{rule}\ \#4': \textit{if}\ \mathrm{SC} = \text{white}, & \textit{then}\ \mathrm{Class} = \mathrm{O} & \{d\}\,. \end{array}$$

It is interesting to come back to Table 1 and to ask what decision rules represent this information system. As the description of the four signs by S and PC is not sufficient to characterize exactly all the classes, it is not surprising that not all the rules will have a non-ambiguous decision. Indeed, the following decision rules can be induced:

$$\begin{array}{lll} \textit{rule}\ \#1'': \textit{if}\ \mathrm{S} = \text{triangle}, & \textit{then}\ \mathrm{Class} = \mathrm{W} & \{a\} \\ \textit{rule}\ \#2'': \textit{if}\ \mathrm{PC} = \text{white}, & \textit{then}\ \mathrm{Class} = \mathrm{I} & \{b\} \\ \textit{rule}\ \#3'': \textit{if}\ \mathrm{PC} = \text{blue}, & \textit{then}\ \mathrm{Class} = \mathrm{I\ or\ O} & \{c, d\}\,. \end{array}$$

Note that these rules can be induced from the lower approximations of classes W and I, and from the set called the *boundary* of both I and O. Indeed, for exact rule #1″, the supporting example is in $\text{lower_approx.}_{\mathrm{S,PC}}(\mathrm{W}) = \{a\}$; for exact rule #2″ it is in $\text{lower_approx.}_{\mathrm{S,PC}}(\mathrm{I}) = \{b\}$; and the supporting examples for approximate rule #3″ are in the boundary of classes I and O, defined as:

$$\begin{aligned} &\text{boundary}_{\mathrm{S,PC}}(\mathrm{I}) \\ &\quad = \text{upper_approx.}_{\mathrm{S,PC}}(\mathrm{I}) - \text{lower_approx.}_{\mathrm{S,PC}}(\mathrm{I}) \\ &\quad = \{c, d\}\,, \end{aligned}$$

$$\begin{aligned} &\text{boundary}_{\mathrm{S,PC}}(\mathrm{O}) \\ &\quad = \text{upper_approx.}_{\mathrm{S,PC}}(\mathrm{O}) - \text{lower_approx.}_{\mathrm{S,PC}}(\mathrm{O}) \\ &\quad = \{c, d\}\,. \end{aligned}$$

As a result of the approximate characterization of classes W, I and O by S and PC, an approximate representation in terms of decision rules is obtained. Since the quality of the approximation is 1/2, exact rules (#1″ and #2″) cover one half of the examples and the other half is covered by the approximate rule (#3″). Now, the quality of approximation by S and SC, or by PC and SC, was equal to 1, so all examples were covered by exact rules (#1 to #3, or #1′ to #4′ respectively).

One can see, from this simple example, that the rough set analysis of data included in an information system provides some useful information. In particular, the following results are obtained:

- A characterization of decision classes in terms of chosen condition attributes through lower and upper approximation.
- A measure of the quality of approximation which indicates how good the chosen set of attributes is for approximation of the classification.
- The reducts of condition attributes including all relevant attributes. At the same time, superfluous and exchangeable attributes are also identified.
- The core composed of indispensable attributes.
- A set of decision rules which is induced from the lower and upper approximations of the decision classes. This constitutes a classification model for a given information system.

Dominance-Based Rough Set Approach to Ordinal Classification Problems

Dominance-Based Rough Set Approach (DRSA)

Dominance-based Rough Set Approach (DRSA) has been proposed by the authors to handle background knowledge about ordinal evaluations of objects from a universe, and about monotonic relationships between these evaluations, e. g. "the larger the mass and the smaller the distance, the larger the gravity" or "the greater the debt of a firm, the greater its risk of failure". Such a knowledge is typical for data describing various phenomena. It is also characteristic for data concerning multiple criteria decision or decision under uncertainty, where the order of value sets of condition and decision attributes corresponds to increasing or decreasing preference. In case of multiple criteria decision, the condition attributes are called *criteria*.

Let us consider a decision table including a finite universe of objects (solutions, alternatives, actions) U evaluated on a finite set of criteria $F = \{f_1, \ldots, f_n\}$, and on a single decision attribute d. The set of the indices of criteria is denoted by $I = \{1, \ldots, n\}$. Without loss of generality, $f_i : U \to \Re$ for each $i = 1, \ldots, n$, and, for all objects $x, y \in U$, $f_i(x) \geq f_i(y)$ means that "x is at least as good as y with respect to criterion f_i", which is denoted by $x \succeq_i y$. Therefore, it is supposed that $\succeq_i$ is a complete preorder, i. e. a strongly complete and transitive binary relation, defined on U on the basis of evaluations $f_i(\cdot)$. Furthermore, decision attribute d makes a partition of U into a finite number of decision classes, $\mathbf{Cl} = \{Cl_1, \ldots, Cl_m\}$, such that each $x \in U$ belongs to one and only one class Cl_t, $t = 1, \ldots, m$. It is assumed that the classes are preference ordered, i. e. for all $r, s = 1, \ldots, m$, such that $r > s$, the objects from Cl_r are preferred to the objects from Cl_s. More formally, if $\succeq$ is a *comprehensive weak preference relation* on U, i. e. if for all $x, y \in U$, $x \succeq y$ reads "x is at least as good as y", then it is supposed that

$$[x \in Cl_r,\ y \in Cl_s,\ r > s] \Rightarrow x \succ y\,,$$

where $x \succ y$ means $x \succeq y$ and *not* $y \succeq x$.

The above assumptions are typical for consideration of an ordinal classification (or multiple criteria sorting) problem. Indeed, the decision table characterized above includes examples of ordinal classification which constitute an input *preference information* to be analyzed using DRSA.

The sets to be approximated are called *upward union* and *downward union* of decision classes, respectively:

$$Cl_t^{\geq} = \bigcup_{s \geq t} Cl_s, \quad Cl_t^{\leq} = \bigcup_{s \leq t} Cl_s, \quad t = 1, \ldots, m\,.$$

The statement $x \in Cl_t^{\geq}$ reads "x belongs to at least class Cl_t", while $x \in Cl_t^{\leq}$ reads "x belongs to at most class Cl_t". Let us remark that $Cl_1^{\geq} = Cl_m^{\leq} = U$, $Cl_m^{\geq} = Cl_m$ and $Cl_1^{\leq} = Cl_1$. Furthermore, for $t = 2, \ldots, m$,

$$Cl_{t-1}^{\leq} = U - Cl_t^{\geq} \quad \text{and} \quad Cl_t^{\geq} = U - Cl_{t-1}^{\leq}\,.$$

The key idea of DRSA is representation (approximation) of upward and downward unions of decision classes, by *granules of knowledge* generated by criteria. These granules are *dominance cones* in the criteria values space.

x *dominates* y with respect to set of criteria $P \subseteq I$ (shortly, x *P-dominates* y), denoted by xD_Py, if for every criterion $i \in P$, $f_i(x) \geq f_i(y)$. The relation of P-dominance is reflexive and transitive, i. e. it is a partial preorder.

Given a set of criteria $P \subseteq I$ and $x \in U$, the granules of knowledge used for approximation in DRSA are:

- a set of objects dominating x, called *P-dominating set*, $D_P^+(x) = \{y \in U : yD_Px\}$,
- a set of objects dominated by x, called *P-dominated set*, $D_P^-(x) = \{y \in U : xD_Py\}$.

Let us recall that the dominance principle requires that an object x dominating object y on all considered criteria (i. e. x having evaluations at least as good as y on all considered criteria) should also dominate y on the decision (i. e. x should be assigned to at least as good decision class as y). Objects satisfying the dominance principle are called *consistent*, and those which are violating the dominance principle are called *inconsistent*.

The *P-lower approximation* of $Cl_t^{\geq}$, denoted by $\underline{P}\left(Cl_t^{\geq}\right)$, and the *P-upper approximation* of $Cl_t^{\geq}$, denoted by $\overline{P}\left(Cl_t^{\geq}\right)$, are defined as follows ($t = 2, \ldots, m$):

$$\underline{P}\left(Cl_t^{\geq}\right) = \left\{x \in U : D_P^+(x) \subseteq Cl_t^{\geq}\right\},$$
$$\overline{P}\left(Cl_t^{\geq}\right) = \left\{x \in U : D_P^-(x) \cap Cl_t^{\geq} \neq \emptyset\right\}.$$

Analogously, one can define the *P-lower approximation* and the *P-upper approximation* of $Cl_t^{\leq}$ as follows ($t = 1, \ldots, m-1$):

$$\underline{P}\left(Cl_t^{\leq}\right) = \left\{x \in U : D_P^-(x) \subseteq Cl_t^{\leq}\right\},$$
$$\overline{P}\left(Cl_t^{\leq}\right) = \left\{x \in U : D_P^+(x) \cap Cl_t^{\leq} \neq \emptyset\right\}.$$

The P-lower and P-upper approximations so defined satisfy the following *inclusion properties*, for all $P \subseteq I$:

$$\underline{P}\left(Cl_t^{\geq}\right) \subseteq Cl_t^{\geq} \subseteq \overline{P}\left(Cl_t^{\geq}\right), \quad t = 2, \ldots, m\,,$$

$$\underline{P}\left(Cl_t^{\leq}\right) \subseteq Cl_t^{\leq} \subseteq \overline{P}\left(Cl_t^{\leq}\right), \quad t = 1, \ldots, m-1\,.$$

The P-lower and P-upper approximations of $Cl_t^{\geq}$ and $Clt^{\leq}$ have an important *complementarity property*, according to which,

$$\underline{P}\left(Cl_t^{\geq}\right) = U - \overline{P}\left(Cl_{t-1}^{\leq}\right) \quad \text{and}$$
$$\overline{P}\left(Cl_t^{\geq}\right) = U - \underline{P}\left(Cl_{t-1}^{\leq}\right), \quad t = 2, \ldots, m,$$
$$\underline{P}\left(Cl_t^{\leq}\right) = U - \overline{P}\left(Cl_{t+1}^{\geq}\right) \quad \text{and}$$
$$\overline{P}\left(Cl_t^{\leq}\right) = U - \underline{P}\left(Cl_{t+1}^{\geq}\right), \quad t = 1, \ldots, m-1 .$$

The *P-boundary* of $Cl_t^{\geq}$ and $Cl_t^{\leq}$, denoted by $Bn_P\left(Cl_t^{\geq}\right)$ and $Bn_P\left(Cl_t^{\leq}\right)$, respectively, are defined as follows:

$$Bn_P\left(Cl_t^{\geq}\right) = \overline{P}\left(Cl_t^{\geq}\right) - \underline{P}\left(Cl_t^{\geq}\right), \quad t = 2, \ldots, m,$$
$$Bn_P\left(Cl_t^{\leq}\right) = \overline{P}\left(Cl_t^{\leq}\right) - \underline{P}\left(Cl_t^{\leq}\right), \quad t = 1, \ldots, m-1.$$

Due to the above complementarity property, $Bn_P\left(Cl_t^{\geq}\right) = Bn_P\left(Cl_{t-1}^{\leq}\right)$, for $t = 2, \ldots, m$.

For every $P \subseteq C$, the *quality of approximation* of the ordinal classification $\boldsymbol{Cl}$ by a set of criteria P is defined as the ratio of the number of objects P-consistent with the dominance principle and the number of all the objects in U. Since the P-consistent objects are those which do not belong to any P-boundary $Bn_P(Cl_t^{\geq})$, $t = 2, \ldots, m$, or $Bn_P\left(Cl_t^{\leq}\right)$, $t = 1, \ldots, m-1$, the quality of approximation of the ordinal classification $\boldsymbol{Cl}$ by a set of criteria P, can be written as

$$\gamma_P(\boldsymbol{Cl}) = \frac{\left|U - \left(\bigcup_{t=2,\ldots,m} Bn_P\left(Cl_t^{\geq}\right)\right)\right|}{|U|} = \frac{\left|U - \left(\bigcup_{t=1,\ldots,m-1} Bn_P\left(Cl_t^{\leq}\right)\right)\right|}{|U|} .$$

$\gamma_P(\boldsymbol{Cl})$ can be seen as a degree of consistency of the objects from U, when P is the set of criteria and $\boldsymbol{Cl}$ is the considered ordinal classification.

Each minimal (with respect to inclusion) subset $P \subseteq C$ such that $\gamma_P(\boldsymbol{Cl}) = \gamma_C(\boldsymbol{Cl})$ is called a *reduct* of $\boldsymbol{Cl}$, and is denoted by $\text{RED}_{\boldsymbol{Cl}}$. Let us remark that for a given set U one can have more than one reduct. The intersection of all reducts is called the *core*, and is denoted by $\text{CORE}_{\boldsymbol{Cl}}$. Criteria in $\text{CORE}_{\boldsymbol{Cl}}$ cannot be removed from consideration without deteriorating the quality of approximation. This means that, in set C, there are three categories of criteria:

- *indispensable* criteria included in the core,
- *exchangeable* criteria included in some reducts, but not in the core,
- *redundant* criteria, neither indispensable nor exchangeable, and thus not included in any reduct.

The dominance-based rough approximations of upward and downward unions of decision classes can serve to induce a generalized description of objects in terms of "*if* ..., *then* ..." decision rules. For a given upward or downward union of classes, $Cl_t^{\geq}$ or $Cl_s^{\leq}$, the decision rules induced under a hypothesis that objects belonging to $\underline{P}(Cl_t^{\geq})$ or $\underline{P}\left(Cl_s^{\leq}\right)$ are positive examples, and all the others are negative, suggest a *certain* assignment to "class Cl_t or better", or to "class Cl_s or worse", respectively. On the other hand, the decision rules induced under a hypothesis that objects belonging to $\overline{P}\left(Cl_t^{\geq}\right)$ or $\overline{P}\left(Cl_s^{\leq}\right)$ are positive examples, and all the others are negative, suggest a *possible* assignment to "class Cl_t or better", or to "class Cl_s or worse", respectively. Finally, the decision rules induced under a hypothesis that objects belonging to the intersection $\overline{P}(Cl_s^{\leq}) \cap \overline{P}(Cl_t^{\geq})$ are positive examples, and all the others are negative, suggest an *approximate* assignment to some classes between Cl_s and $Cl_t (s < t)$.

In the case of preference ordered description of objects, set U is composed of examples of ordinal classification. Then, it is meaningful to consider the following five types of decision rules:

1) *certain $D_{\geq}$-decision rules*, providing lower profile descriptions for objects belonging to $\underline{P}(Cl_t^{\geq})$: if $f_{i_1}(x) \geq r_{i_1}$ and ... and $f_{i_p}(x) \geq r_{i_p}$, then $x \in Cl_t^{\geq}, \{i_1, \ldots, i_p\} \subseteq I, t = 2, \ldots, m, r_{i_1}, \ldots, r_{i_p} \in \Re$;
2) *possible $D_{\geq}$-decision rules*, providing lower profile descriptions for objects belonging to $\overline{P}(Cl_t^{\geq})$: if $f_{i_1}(x) \geq r_{i_1}$ and ... and $f_{i_p}(x) \geq r_{i_p}$, then x possibly belongs to $Cl_t^{\geq}, \{i_1, \ldots, i_p\} \subseteq I, t = 2, \ldots, m, r_{i_1}, \ldots, r_{i_p} \in \Re$;
3) *certain $D_{\leq}$-decision rules*, providing upper profile descriptions for objects belonging to $\underline{P}(Cl_t^{\leq})$: if $f_{i_1}(x) \leq r_{i_1}$ and ... and $f_{i_p}(x) \leq r_{i_p}$, then $x \in Cl_t^{\leq}, \{i_1, \ldots, i_p\} \subseteq I, t = 1, \ldots, m-1, r_{i_1}, \ldots, r_{i_p} \in \Re$;
4) *possible $D_{\leq}$-decision rules*, providing upper profile descriptions for objects belonging to $\overline{P}(Cl_t^{\leq})$: if $f_{i_1}(x) \leq r_{i_1}$ and ... and $f_{i_p}(x) \leq r_{i_p}$, then x possibly belongs to $Cl_t^{\leq}, \{i_1, \ldots, i_p\} \subseteq I, t = 1, \ldots, m-1, r_{i_1}, \ldots, r_{i_p} \in \Re$;
5) *approximate $D_{\geq\leq}$-decision rules*, providing simultaneously lower and upper profile descriptions for objects belonging to $Cl_s \cup Cl_{s+1} \cup \ldots \cup Cl_t$, without possibility of discerning to which class: if $f_{i_1}(x) \geq r_{i_1}$ and ... and $f_{i_k}(x) \geq r_{i_k}$ and $f_{i_{k+1}}(x) \leq r_{i_{k+1}}$ and ... and $f_{i_p}(x) \leq r_{i_p}$, then $x \in Cl_s \cup Cl_{s+1} \cup \ldots \cup Cl_t, \{i_1, \ldots, i_p\} \subseteq I, s, t \in \{1, \ldots, m\}, s < t, r_{i_1}, \ldots, r_{i_p} \in \Re$.

In the premise of a $D_{\geq\leq}$-decision rule, there can be "$f_i(x) \geq r_i$" and "$f_i(x) \leq r_i'$", where $r_i \leq r_i'$, for the same

$i \in I$. Moreover, if $r_i = r'_i$, the two conditions boil down to "$f_i(x) = r_i$".

Since a decision rule is a kind of implication, a *minimal* rule is understood as an implication such that there is no other implication with the premise of at least the same weakness (in other words, a rule using a subset of elementary conditions or/and weaker elementary conditions) and the conclusion of at least the same strength (in other words, a $D_{\geq}$- or a $D_{\leq}$-decision rule assigning objects to the same union or sub-union of classes, or a $D_{\geq\leq}$-decision rule assigning objects to the same or smaller set of classes).

The rules of type 1) and 3) represent certain knowledge extracted from data (examples of ordinal classification), while the rules of type 2) and 4) represent possible knowledge; the rules of type 5) represent doubtful knowledge, because they are supported by inconsistent objects only.

Moreover, the rules of type 1) and 3) are *exact* if they do not cover negative examples, and they are *probabilistic* otherwise. In the latter case, each rule is characterized by a *confidence ratio*, representing the probability that an object matching the premise of the rule also matches its conclusion.

Given a certain or possible $D_{\geq}$-decision rule $r \equiv$ "if $f_{i_1}(x) \geq r_{i_1}$ and … and $f_{i_p}(x) \geq r_{i_p}$, then $x \in Cl_t^{\geq}$", an object $y \in U$ *supports* r if $f_{i_1}(y) \geq r_{i_1}$ and … and $f_{i_p}(y) \geq r_{i_p}$. Moreover, object $y \in U$ supporting decision rule r is a *base* of r if $f_{i_1}(y) = r_{i_1}$ and … and $f_{i_p}(y) = r_{i_p}$. Similar definitions hold for certain or possible $D_{\leq}$-decision rules and approximate $D_{\geq\leq}$-decision rules. A decision rule having at least one base is called *robust*. Identification of supporting objects and bases of robust rules is important for interpretation of the rules in multiple criteria decision analysis perspective. The ratio of the number of objects supporting a rule and the number of all considered objects is called *relative support* of a rule. The relative support and the confidence ratio are basic characteristics of a rule, however, some *Bayesian confirmation measures* reflect much better the attractiveness of a rule [25].

A set of decision rules is *complete* if it covers all considered objects (examples of ordinal classification) in such a way that consistent objects are re-assigned to their original classes, and inconsistent objects are assigned to clusters of classes referring to this inconsistency. A set of decision rules is *minimal* if it is complete and non-redundant, i. e. exclusion of any rule from this set makes it incomplete.

Note that the syntax of decision rules induced from rough approximations defined using dominance cones, is using consistently this type of granules. Each condition profile defines a dominance cone in $\frac{p}{\uparrow V(p \leq n)}$-dimensional condition space $\Re^p (p \leq n)$, and each decision profile defines a dominance cone in one-dimensional decision space $\{1, \ldots, m\}$. In both cases, the cones are positive for $D_{\geq}$-rules and negative for $D_{\leq}$-rules.

Let also remark that dominance cones corresponding to condition profiles can originate in any point of $\Re^n$, without the risk of their being too specific. Thus, contrary to traditional granular computing, the condition space $\Re^n$ need not be discretized.

Procedures for rule induction from dominance-based rough approximations have been proposed in [17].

In [10], a new methodology for the induction of monotonic decision trees from dominance-based rough approximations of preference ordered decision classes has been proposed.

Application of DRSA to decision related problems goes far beyond ordinal classification. In [27], DRSA has been used for decision support involving multiple decision makers, and in [28], DRSA has been applied to case-based reasoning. The following sections present applications of DRSA to multiple criteria choice and ranking, to decision under uncertainty and to interactive multiobjective optimization. The surveys [24,26,29,51,52] include other applications of DRSA.

Example Illustrating DRSA in the Context of Ordinal Classification

This subsection presents a didactic example which illustrates the main concepts of DRSA. Let us consider the following ordinal classification problem. Students of a college must obtain an overall evaluation on the basis of their achievements in Mathematics, Physics and Literature. The three subjects are clearly criteria (condition attributes) and the comprehensive evaluation is a decision attribute. For simplicity, the value sets of the criteria and of the decision attribute are the same, and they are composed of three values: bad, medium and good. The preference order of these values is obvious. Thus, there are three preference ordered decision classes, so the problem belongs to the category of ordinal classification. In order to build a preference model of the jury, DRSA is used to analyze a set of exemplary evaluations of students provided by the jury. These examples of ordinal classification constitute an input preference information presented as decision table in Table 3.

Note that the dominance principle obviously applies to the examples of ordinal classification, since an improvement of a student's score on one of three criteria, with other scores unchanged, should not worsen the student's overall evaluation, but rather improve it.

Rough Sets in Decision Making, Table 3
Exemplary evaluations of students (examples of ordinal classification)

Student	Mathematics	Physics	Literature	Overall evaluation
*S*1	good	medium	bad	bad
*S*2	medium	medium	bad	medium
*S*3	medium	medium	medium	medium
*S*4	good	good	medium	good
*S*5	good	medium	good	good
*S*6	good	good	good	good
*S*7	bad	bad	bad	bad
*S*8	bad	bad	medium	bad

Rough Sets in Decision Making, Table 4
Exemplary evaluations of students excluding Literature

Student	Mathematics	Physics	Overall evaluation
*S*1	good	medium	bad
*S*2	medium	medium	medium
*S*3	medium	medium	medium
*S*4	good	good	good
*S*5	good	medium	good
*S*6	good	good	good
*S*7	bad	bad	bad
*S*8	bad	bad	bad

Observe that student *S*1 has not worse evaluations than student *S*2 on all the considered criteria, however, the overall evaluation of *S*1 is worse than the overall evaluation of *S*2. This contradicts the dominance principle, so the two examples of ordinal classification, and only those, are inconsistent. The quality of approximation of the ordinal classification represented by examples in Table 3 is equal to 0.75.

One can observe that in result of reducing the set of considered criteria, i. e. the set of considered subjects, some new inconsistencies can occur. For example, removing from Table 3 the evaluation on Literature, one obtains Table 4, where *S*1 is inconsistent not only with *S*2, but also with *S*3 and *S*5. In fact, student *S*1 has not worse evaluations than students *S*2, *S*3 and *S*5 on all the considered criteria (Mathematics and Physics), however, the overall evaluation of *S*1 is worse than the overall evaluation of *S*2, *S*3 and *S*5.

Observe, moreover, that removing from Table 3 the evaluations on Mathematics, one obtains Table 5, where no new inconsistencies occur, comparing to Table 3.

Similarly, after removing from Table 3 the evaluations

Rough Sets in Decision Making, Table 5
Exemplary evaluations of students excluding Mathematics

Student	Physics	Literature	Overall evaluation
*S*1	medium	bad	bad
*S*2	medium	bad	medium
*S*3	medium	medium	medium
*S*4	good	medium	good
*S*5	medium	good	good
*S*6	good	good	good
*S*7	bad	bad	bad
*S*8	bad	medium	bad

Rough Sets in Decision Making, Table 6
Exemplary evaluations of students excluding Physics

Student	Mathematics	Literature	Overall evaluation
*S*1	good	bad	bad
*S*2	medium	bad	medium
*S*3	medium	medium	medium
*S*4	good	medium	good
*S*5	good	good	good
*S*6	good	good	good
*S*7	bad	bad	bad
*S*8	bad	medium	bad

on Physics, one obtains Table 6, where no new inconsistencies occur, comparing to Table 3.

The fact that no new inconsistency occurs when Mathematics or Physics is removed, means that the subsets of criteria {Physics, Literature} or {Mathematics, Literature} contain sufficient information to represent the overall evaluation of students with the same quality of approximation as using the complete set of three criteria. This is not the case, however, for the subset {Mathematics, Physics}. Observe, moreover, that subsets {Physics, Literature} and {Mathematics, Literature} are minimal, because no other criterion can be removed without new inconsistencies occur. Thus, {Physics, Literature} and {Mathematics, Literature} are the reducts of the complete set of criteria {Mathematics, Physics, Literature}. Since Literature is the only criterion which cannot be removed from any reduct without introducing new inconsistencies, it constitutes the core, i. e. the set of indispensable criteria. The core is, of course, the intersection of all reducts, i. e. in our example:

$$\{\text{Literature}\} = \{\text{Physics, Literature}\} \cap \{\text{Mathematics, Literature}\}.$$

In order to illustrate in a simple way the concept of rough approximation, let us confine our anal-

ysis to the reduct {Mathematics, Literature}. Let us consider student $S4$. His positive dominance cone $D^{+}_{\{\text{Mathematics, Literature}\}}(S4)$ is composed of all the students having evaluations not worse than him on Mathematics and Literature, i. e. of all the students dominating him with respect to Mathematics and Literature. Thus,

$$D^{+}_{\{\text{Mathematics, Literature}\}}(S4) = \{S4, S5, S6\} .$$

On the other hand, the negative dominance cone of student $S4$, $D^{-}_{\{\text{Mathematics, Literature}\}}(S4)$, is composed of all the students having evaluations not better than him on Mathematics and Literature, i. e. of all the students dominated by him with respect to Mathematics and Literature. Thus,

$$D^{-}_{\{\text{Mathematics, Literature}\}}(S4) = \{S1, S2, S3, S4, S7, S8\} .$$

Similar dominance cones can be obtained for all the students from Table 6. For example, for $S2$, the dominance cones are

$$D^{+}_{\{\text{Mathematics, Literature}\}}(S2) = \{S1, S2, S3, S4, S5, S6\}$$

and

$$D^{-}_{\{\text{Mathematics, Literature}\}}(S2) = \{S2, S7\} .$$

The rough approximations can be calculated using dominance cones. Let us consider, for example, the lower approximation of the set of students having a "good" overall evaluation $\underline{P}(Cl^{\geq}_{\text{good}})$, with $P = \{\text{Mathematics, Literature}\}$. Notice that $\underline{P}(Cl^{\geq}_{\text{good}}) = \{S4, S5, S6\}$, because positive dominance cones of students $S4$, $S5$ and $S6$ are all included in the set of students with an overall evaluation "good". In other words, this means that there is no student dominating $S4$ or $S5$ or $S6$ while having an overall evaluation worse than "good". From the viewpoint of decision making, this means that, taking into account the available information about evaluation of students on Mathematics and Literature, the fact that student y dominates $S4$ or $S5$ or $S6$ is a *sufficient* condition to conclude that y is a "good" student.

The upper approximation of the set of students with a "good" overall evaluation is $\overline{P}(Cl^{\geq}_{\text{good}}) = \{S4, S5, S6\}$, because negative dominance cones of students $S4$, $S5$ and $S6$ have a nonempty intersection with the set of students having a "good" overall evaluation. In other words, this means that for each one of the students $S4$, $S5$ and $S6$, there is at least one student dominated by him with an overall evaluation "good". From the point of view of decision making, this means that, taking into account the available information about evaluation of students on Mathematics and Literature, the fact that student y dominates $S4$ or $S5$ or $S6$ is a *possible* condition to conclude that y is a "good" student.

Let us observe that for the set of criteria $P = \{\text{Mathematics, Literature}\}$, the lower and upper approximations of the set of "good" students are the same. This means that examples of ordinal classification concerning this decision class are all consistent. This is not the case, however, for the examples concerning the union of decision classes "at least medium". For this upward union the rough approximations are $\underline{P}(Cl^{\geq}_{\text{medium}}) = \{S3, S4, S5, S6\}$ and $\overline{P}(Cl^{\geq}_{\text{medium}}) = \{S1, S2, S3, S4, S5, S6\}$. The difference between $\overline{P}(Cl^{\geq}_{\text{medium}})$ and $\underline{P}(Cl^{\geq}_{\text{medium}})$, i. e. the boundary $Bn_P(Cl^{\geq}_{\text{medium}}) = \{S1, S2\}$, is composed of students with inconsistent overall evaluations, which has already been noticed above. From the viewpoint of decision making, this means that, taking into account the available information about evaluation of students on Mathematics and Literature, the fact that student y is dominated by $S1$ and dominates $S2$ is a condition to conclude that y can obtain an overall evaluation "at least medium" with some doubts.

Until now, rough approximations of only upward unions of decision classes have been considered. It is interesting, however, to calculate also rough approximations of downward unions of decision classes. Let us consider first the lower approximation of the set of students having "at most medium" overall evaluation $\underline{P}(Cl^{\leq}_{\text{medium}})$. Observe that $\underline{P}(Cl^{\leq}_{\text{medium}}) = \{S1, S2, S3, S7, S8\}$, because the negative dominance cones of students $S1$, $S2$, $S3$, $S7$, and $S8$ are all included in the set of students with overall evaluation "at most medium". In other words, this means that there is no student dominated by $S1$ or $S2$ or $S3$ or $S7$ or $S8$ while having an overall evaluation better than "medium". From the viewpoint of decision making, this means that, taking into account the available information about evaluation of students on Mathematics and Literature, the fact that student y is dominated by $S1$ or $S2$ or $S3$ or $S7$ or $S8$ is a *sufficient* condition to conclude that y is an "at most medium" student.

The upper approximation of the set of students with an "at most medium" overall evaluation is $\overline{P}(Cl^{\leq}_{\text{medium}}) = \{S1, S2, S3, S7, S8\}$, because the positive dominance cones of students $S1$, $S2$, $S3$, $S7$, and $S8$ have a nonempty intersection with the set of students having an "at most medium" overall evaluation. In other words, this means that for each one of the students $S1$, $S2$, $S3$, $S7$, and $S8$, there is at least one student dominating him with an overall evaluation "at most medium". From the viewpoint of decision making, this means that, taking into account

the available information about evaluation of students on Mathematics and Literature, the fact that student y is dominated by $S1$ or $S2$ or $S3$ or $S7$ or $S8$ is a *possible* condition to conclude that y is an "at most medium" student.

Finally, the lower and upper approximations of the set of students having a "bad" overall evaluation are $\underline{P}(Cl^{\leq}_{\text{bad}}) = \{S7, S8\}$ and $\overline{P}(Cl^{\leq}_{\text{bad}}) = \{S1, S2, S7, S8\}$. The difference between $\overline{P}(Cl^{\leq}_{\text{bad}})$ and $\underline{P}(Cl^{\leq}_{\text{bad}})$, i. e. the boundary $Bn_P(Cl^{\leq}_{\text{bad}}) = \{S1, S2\}$ is composed of students with inconsistent overall evaluations, which has already been noticed above. From the point of view of decision making, this means that, taking into account the available information about evaluation of students on Mathematics and Literature, the fact that student y is dominated by $S1$ and dominates $S2$ is a condition to conclude that y can obtain an overall evaluation "bad" with some doubts. Observe, moreover, that $Bn_P(Cl^{\geq}_{\text{medium}}) = Bn_P(Cl^{\leq}_{\text{bad}}) = \{S1, S2\}$.

Given the above rough approximations with respect to the set of criteria $P = \{\text{Mathematics, Literature}\}$, one can induce a set of decision rules representing the preferences of the jury. The idea is that evaluation profiles of students belonging to the lower approximations can serve as a base for some certain rules, while evaluation profiles of students belonging to the boundaries can serve as a base for some approximate rules. The following decision rules have been induced (between parentheses there are id's of students supporting the corresponding rule; the student being a rule base is underlined):

rule 1) if the evaluation on Mathematics is (at least) good, and the evaluation on Literature is at least medium, then the overall evaluation is (at least) good, $\{\underline{S4}, S5, S6\}$,

rule 2) if the evaluation on Mathematics is at least medium, and the evaluation on Literature is at least medium, then the overall evaluation is at least medium, $\{\underline{S3}, S4, S5, S6\}$,

rule 3) if the evaluation on Mathematics is at least medium, and the evaluation on Literature is (at most) bad, then the overall evaluation is bad or medium, $\{S1, \underline{S2}\}$,

rule 4) if the evaluation on Mathematics is at least medium, then the overall evaluation is at least medium, $\{\underline{S2, S3}, S4, S5, S6\}$,

rule 5) if the evaluation on Literature is (at most) bad, then the overall evaluation is at most medium, $\{\underline{S1, S2, S7}\}$,

rule 6) if the evaluation on Mathematics is (at most) bad, then the overall evaluation is (at most) bad, $\{\underline{S7, S8}\}$.

Notice that rules 1)–2), 4)–7) are certain, while rule 3) is an approximate one. These rules represent knowledge discovered from the available information. In the current context, the knowledge is interpreted as a preference model of the jury. A characteristic feature of the syntax of decision rules representing preferences is the use of expressions "at least" or "at most" a value; in case of extreme values ("good" and "bad"), these expressions are put in parentheses because there is no value above "good" and below "bad".

Even if one can represent all the knowledge using only one reduct of the set of criteria (as it is the case using $P = \{\text{Mathematics, Literature}\}$), when considering a larger set of criteria than a reduct, one can obtain a more synthetic representation of knowledge, i. e. the number of decision rules or the number of elementary conditions, or both of them, can get smaller. For example, considering the set of all three criteria, {Mathematics, Physics, Literature}, one can induce a set of decision rules composed of the above rules 1), 2), 3) and 6), plus the following:

rule 7) if the evaluation on Physics is at most medium, and the evaluation on Literature is at most medium, then the overall evaluation is at most medium, $\{S1, S2, \underline{S3}, S7, S8\}$.

Thus, the complete set of decision rules induced from Table 3 is composed of 5 instead of 6 rules.

Once accepted by the DM, these rules represent his/her preference model. Assuming that rules 1)–7) in our example represent the preference model of the jury, it can be used to evaluate new students. For example, student $S9$ who is "medium" in Mathematics and Physics and "good" in Literature, would be evaluated as "medium" because his profile matches the premise of rule 2), having as consequence an overall evaluation at least "medium". The overall evaluation of $S9$ cannot be "good", because his profile does not match any rule having as consequence an overall evaluation "good" (in the considered example, the only rule of this type is rule 1) whose premise is not matched by the profile of $S9$.

DRSA on a Pairwise Comparison Table for Multiple Criteria Choice and Ranking Problems

Multiple Criteria Choice and Ranking Problems

Ordinal classification decisions are based on absolute evaluation of objects on multiple criteria, however, multiple criteria choice and ranking decisions are based on preference relations between objects. Decision table including examples of ordinal classification does not contain information about preference relations between objects, thus, in order to apply DRSA to multiple criteria choice and ranking problems, a different representation of the input preference information is needed.

To handle binary relations within the rough set approach, it has been proposed in [14,15] to operate on, so-called, *pairwise comparison table* (PCT), i. e. a decision table including pairs of objects for which multiple criteria evaluations and a comprehensive preference relation are known. PCT represents preference information provided by the DM in the form of decision examples (pairwise comparisons of objects).

Similarly to ordinal classification, decision examples concerning multiple criteria choice and ranking may be inconsistent with respect to the dominance principle, however, interpretation of the inconsistency is different: It occurs when preferences of a pair of objects, say (a, b), on all considered criteria are not weaker than preferences of another pair of objects, say (c, d), on these criteria, however, the comprehensive preference of object a over object b is weaker than the comprehensive preference of object c over object d.

The Pairwise Comparison Table

Similarly to the ordinal classification, let us consider a finite set of criteria $F = \{f_1, \ldots, f_n\}$, the set of their indices $I = \{1, \ldots, n\}$ and a finite universe of objects (actions, solutions, alternatives) U. For any criterion $i \in I$, let T_i be a finite set of binary relations defined on U on the basis of the evaluations of objects from U with respect to the considered criterion i, such that $\forall (x, y) \in U \times U$ exactly one binary relation $t \in T_i$ is verified; $t \in T_i$ has the meaning of a *preference relation* for a pair of objects on a particular criterion i. More precisely, given value set V_i of $i \in I$, if $v_i', v_i'' \in V_i$ are the respective evaluations of $x, y \in U$ on criterion i, and $(x, y) \in t$, with $t \in T_i$, then for each $w, z \in U$ having the same evaluations v_i', v_i'' on i, $(w, z) \in t$. For interesting applications it should be $|T_q| \geq 2$, $\forall i \in I$. Furthermore, let T_d be a set of binary relations defined on U, such that at most one binary relation $t \in T_d$ is verified $\forall (x, y) \in U \times U$; $t \in T_d$ has the meaning of a *comprehensive preference relation* for a pair of objects (comprehensive pairwise comparison).

The preference information provided by the DM, has the form of pairwise comparisons of some reference objects from $B \subseteq U$. These decision examples are presented in the pairwise comparison table (PCT), defined as information table $\boldsymbol{S}_{\mathrm{PCT}} = \langle \hat{B}, F \cup \{d\}, T_F \cup T_d, g \rangle$, where $\hat{B} \subseteq B \times B$ is a non-empty set of *exemplary pairwise comparisons of reference objects*, $T_F = \bigcup_{i \in I} T_i$, d is a decision corresponding to the comprehensive pairwise comparison (comprehensive preference relation), and $g\colon \hat{B} \times (F \cup \{d\}) \rightarrow T_F \cup T_d$ is a total function such that $g[(x, y), i] \in T_i$, $\forall (x, y) \in U \times U$ and $\forall i \in I$, and $g[(x, y), d] \in T_d$, $\forall (x, y) \in \hat{B}$. It follows that for any pair of reference objects $(x, y) \in \hat{B}$ there is verified one and only one binary relation $t \in T_d$. Thus, T_d induces a partition of $\hat{B}$. In fact, information table $\boldsymbol{S}_{\mathrm{PCT}}$ can be seen as decision table, since the set of considered criteria F and decision d are distinguished.

It is assumed that the exemplary pairwise comparisons provided by the DM can be represented in terms of *graded preference relations* (for example "very weak preference", "weak preference", "strict preference", "strong preference", "very strong preference") $P_i^h\colon \forall i \in I$ and $\forall (x, y) \in U \times U$,

$$T_i = \left\{ P_i^h, h \in H_i \right\},$$

where H_i is a particular subset of the relative integers and

- $xP_i^h y, h > 0$, means that object x is preferred to object y by degree h with respect to criterion i,
- $xP_i^h y, h < 0$, means that object x is not preferred to object y by degree h with respect to criterion i,
- $xP_i^0 y$ means that object x is similar (asymmetrically indifferent) to object y with respect to criterion i.

Of course, $\forall i \in I$ and $\forall (x, y) \in U \times U$, it holds:

$$\left[xP_i^h y, h > 0 \right] \Rightarrow \left[yP_i^k x, k \leq 0 \right].$$

The set of binary relations T_d may be defined in a similar way, but $xP_d^h y$ means that object x is comprehensively preferred to object y by degree h.

Technically, the modeling of the binary relation P_i^h, i. e. the assessment of h, can be organized as follows:

- first, it is observed that criterion i is a function $f_i\colon U \rightarrow \Re$ increasing with respect to the preferences on i, for each $i = 1, \ldots, n$,
- then, for each $i = 1, \ldots, n$, it is possible to define a function $k_i\colon \Re^2 \rightarrow \Re$ which measures the *strength of the preference* (positive or negative) of x over y (e. g. $k_i[f_i(x), f_i(y)] = f_i(x) - f_i(y)$); it should satisfy the following properties $\forall x, y, z \in U$:
  i) $f_i(x) > f_i(y) \Leftrightarrow k_i[f_i(x), f_i(z)] > k_i[f_i(y), f_i(z)]$,
  ii) $f_i(x) > f_i(y) \Leftrightarrow k_i[f_i(z), f_i(x)] < k_i[f_i(z), f_i(y)]$,
  iii) $f_i(x) = f_i(y) \Leftrightarrow k_i[f_i(x), f_i(y)] = 0$,
- next, the domain of k_i can be divided into intervals, using a suitable set of thresholds Δ_i, $\forall i \in I$; these intervals are numbered in such a way that $k_i[f_i(x), f_i(y)] = 0$ belongs to interval no. 0,

- the value of h in the relation $xP_i^h y$ is then equal to the number of the interval including $k_i[f_i(x), f_i(y)]$, for any $(x, y) \in U \times U$.

Actually, property iii) can be relaxed in order to obtain a more general preference model which, for instance, does not satisfy preferential independence.

To simplify the presentation, let us consider a PCT where the set T_d is composed of two binary relations defined on U:

- *x outranks y* (denotation xSy or $(x, y) \in S$), where $(x, y) \in \hat{B}$,
- *x does not outrank y* (denotation $xS^c y$ or $(x, y) \in S^c$), where $(x, y) \in \hat{B}$,

and $S \cup S^c = \hat{B}$, where "x outranks y" means "x is at least as good as y"; observe that the binary relation S is reflexive, but neither necessarily transitive nor complete. In [8], a more general PCT was considered, where the set T_d is composed of multi-graded binary relations defined on U.

Approximation by Means of Graded Dominance Relations

Let $H_P = \bigcap_{i \in P} H_i$, $\forall P \subseteq I$. Given $P \subseteq I$ and $h \in H_P$, $\forall (x, y) \in U \times U$ it is said that *x positively dominates y* by degree h with respect to criteria from P iff $xP_i^{c_i} y$ with $c_i \geq h$, $\forall i \in P$. Analogously, $\forall (x, y) \in U \times U$, *x negatively dominates y* by degree h with respect to criteria from P iff $xP_i^{c_i} y$ with $c_i \leq h$, $\forall i \in P$. Therefore, each $P \subseteq I$ and $h \in H_P$ generate two binary relations (possibly empty) on U, called *positive P-dominance by degree h* (denotation D_{+P}^h) and *negative P-dominance by degree h* (denotation D_{-P}^h), respectively. They satisfy the following conditions:

(P1) if $(x, y) \in D_{+P}^h$, then $(x, y) \in D_{+R}^h$ for each $R \subseteq P$ and $k \leq h$;
(P2) if $(x, y) \in D_{-P}^h$, then $(x, y) \in D_{-R}^h$ for each $R \subseteq P$ and $k \geq h$.

In [15], it has been proposed to approximate the outranking relation S by means of the dominance relation D_{+P}^h. Therefore, S is considered a *rough binary relation*.

The P-lower approximation of S (denotation $\underline{P}(S)$) and the P-upper approximation of S (denotation $\overline{P}(S)$) are defined, respectively, as:

$$\underline{P}(S) = \bigcup_{h \in H_P} \left\{ \left(D_{+P}^h \cap \hat{B} \right) \subseteq S \right\} ,$$
$$\overline{P}(S) = \bigcap_{h \in H_P} \left\{ \left(D_{+P}^h \cap \hat{B} \right) \supseteq S \right\} .$$

$\underline{P}(S)$ may be interpreted as the dominance relation D_{+P}^h having the largest intersection with $\hat{B}$ included in the outranking relation S, and $\overline{P}(S)$ as the dominance relation D_{+P}^h including S and having the smallest intersection with $\hat{B}$.

Analogously, it is possible to approximate the relation S^c by means of the dominance relation D_{-P}^h. Observe that, in general, the definitions of the approximations of S and S^c do not satisfy the condition of complementarity, i. e. it is not true, in general, that $\underline{P}(S)$ is equal to $\hat{B} - \overline{P}(S^c)$ and that $\underline{P}(S^c)$ is equal to $\hat{B} - \overline{P}(S)$. This is because S and S^c are approximated using two different relations, D_{+P}^h and D_{-P}^h, respectively. Nevertheless, the approximations thus obtained constitute a good basis for the generation of simple decision rules.

Decision Rules

It is possible to represent preferences of the DM revealed in terms of exemplary pairwise comparisons contained in a given PCT, using decision rules. Since approximations of S and S^c were made using graded dominance relations, it is possible to induce decision rules being propositions of the following type:

- D_{++}-decision rule: *if* $xD_{+P}^h y$, *then* xSy,
- D_{+-}-decision rule: *if not* $xD_{+P}^h y$, *then* $xS^c y$,
- D_{-+}-decision rule: *if not* $xD_{-P}^h y$, *then* xSy,
- D_{--}-decision rule: *if* $xD_{-P}^h y$, *then* $xS^c y$,

where P is a non-empty subset of I. Therefore, for example, a D_{++}– decision rule is a proposition of the type: "*if x* positively dominates *y* by degree *h* with respect to criteria from *P*, *then x* outranks *y*".

A constructive definition of these rules may be given, being a kind of implication supported by the existence of at least one pair of objects from $\hat{B}$ satisfying one of the four propositions listed above, and by the absence of pairs from $\hat{B}$ contradicting it. Thus, for example, if

- there exists at least one pair $(w, z) \in \hat{B}$ such that $wD_{+P}^h z$ and wSz and
- there does not exist any pair $(v, u) \in \hat{B}$ such that $vD_{+P}^h u$ and $vS^c u$,
- then "*if* $xD_{+P}^h y$, *then* xSy" is accepted as a D_{++}– decision rule.

A D_{++}-decision rule "*if* $xD_{+P}^h y$, *then* xSy" is said to be *minimal* if there does not exist any other rule "*if* $xD_{+R}^k y$, *then* xSy" such that $R \subseteq P$ and $k \leq h$. Analogous definitions hold for the other cases. In other words, a minimal decision rule is a kind of implication for which there is

no other implication whose premise is of at least the same weakness and whose consequence is of at least the same strength.

The following results show connections of the decision rules with the P-lower and P-upper approximations of S and S^c [15]:

- D_{++}-minimal decision rule "*if* $xD^h_{+P}y$, *then* xSy" is supported by pairs of objects belonging to $\underline{P}(S) = D^h_{+P} \cap \hat{B}$,
- D_{--}-minimal decision rule "*if* $xD^h_{-P}y$, *then* $xS^c y$" is supported by pairs of objects belonging to $\underline{P}(S^c) = D^h_{-P} \cap \hat{B}$,
- D_{+-}-minimal decision rule "*if not* $xD^h_{+P}y$, *then* $xS^c y$" is supported by pairs of objects belonging to $\overline{P}(S) = D^h_{+P} \cap \hat{B}$,
- D_{-+}-minimal decision rule "*if not* $xD^h_{-P}y$, *then* xSy" is supported by pairs of objects belonging to $\overline{P}(S^c) = D^h_{-P} \cap \hat{B}$.

Application of the Decision Rules and Final Recommendation

In order to obtain a recommendation in the multiple criteria choice or ranking problems with respect to a set of objects $M \subseteq U$, the decision rules induced from the approximations of S and S^c (defined with respect to reference objects from B) should be applied on set $M \times M$. The application of the rules to any pair of objects $(u, v) \in M \times M$ establishes the presence (uSv) or the absence ($uS^c v$) of outranking with respect to (u, v). More precisely,

- from D_{++}-decision rule "*if* $xD^h_{+P}y$ *then* xSy" and from $uD^h_{+P}v$, one concludes uSv,
- from D_{+-}-decision rule "*if not* $xD^h_{+P}y$ *then* $xS^c y$" and from *not* $uD^h_{+P}v$, one concludes $uS^c y$,
- from D_{-+}-decision rule "*if not* $xD^h_{-P}y$, *then* xSy" and from *not* $uD^h_{-P}v$, one concludes uSv,
- from D_{--}-decision rule "*if* $xD^h_{-P}y$, *then* $xS^c y$" and from $uD^h_{-P}v$, one concludes $uS^c v$.

After the application of the decision rules to each pair of objects $(u, v) \in M \times M$, one of the following four situations may occur:

- uSv and *not* $uS^c v$, that is *true* outranking (denotation $uS^{\mathrm{T}}v$),
- $uS^c v$ and *not* uSv, that is *false* outranking (denotation $uS^{\mathrm{F}}v$),
- uSv and $uS^c v$, that is *contradictory* outranking (denotation $uS^{\mathrm{K}}v$),
- *not* uSv and *not* $uS^c v$, that is *unknown* outranking (denotation $uS^{\mathrm{U}}v$).

The four above situations, which together constitute the so-called four-valued outranking (see [56]), have been introduced to underline the *presence* and the *absence* of *positive* and *negative* reasons for the outranking. Moreover, they make it possible to distinguish contradictory situations from unknown ones.

The following theorem underlines the operational importance of the minimal decision rules [15]: The application of *all* the decision rules obtained for a given $\boldsymbol{S}_{\mathrm{PCT}}$ to a pair of objects $(u, v) \in M \times M$ results in the same outranking relations S and S^c as those obtained from the application of the *minimal* decision rules *only*. Therefore, the set of the minimal decision rules totally characterizes the preferences of the DM contained in $\boldsymbol{S}_{\mathrm{PCT}}$.

A final *recommendation* can be obtained upon a suitable exploitation of the presence and the absence of outranking S and S^c on M. A possible exploitation procedure consists in calculating a specific score, called *Net Flow Score*, for each object $x \in M$:

$$S_{\mathrm{nf}}(x) = S^{++}(x) - S^{+-}(x) + S^{-+}(x) - S^{--}(x)\,,$$

where

$$\begin{aligned}
S^{++}(x) &= |\{y \in M\colon \text{ there is at least one decision rule which affirms } xSy\}|\,,\\
S^{+-}(x) &= |\{y \in M\colon \text{ there is at least one decision rule which affirms } ySx\}|\,,\\
S^{-+}(x) &= |\{y \in M\colon \text{ there is at least one decision rule which affirms } yS^c x\}|\,,\\
S^{--}(x) &= |\{y \in M\colon \text{ there is at least one decision rule which affirms } xS^c y\}|\,.
\end{aligned}$$

The recommendation in multiple criteria ranking problems consists of the total preorder determined by $S_{\mathrm{nf}}(x)$ on M; in multiple criteria choice problems it consists of the object(s) $x^* \in M$ such that $S_{\mathrm{nf}}(x^*) = \max_{x \in M} S_{\mathrm{nf}}(x)$.

The procedure described above has been characterized with reference to a number of desirable properties in [13].

Approximation by Means of Multi-graded Dominance Relations

The graded dominance relation introduced above assumes a common grade of preference for all the considered criteria. While this permits a simple calculation of the approximations and of the resulting decision rules, it is lacking in precision. A dominance relation allowing a different

degree of preference for each considered criterion (multi-graded dominance) gives a far more accurate picture of the preference information contained in the pairwise comparison table $\boldsymbol{S}_{\mathrm{PCT}}$ [14,16,18].

More formally, given $P \subseteq I\,(P \neq \emptyset), (x, y), (w, z) \in U \times U$, (x, y) is said to *dominate* (w, z) with respect to criteria from P (denotation $(x, y)D_P(w, z)$) if x is preferred to y at least as strongly as w is preferred to z with respect to each $i \in P$. Precisely, "at least as strongly as" means "by at least the same degree", i. e. $hi \geq ki$, where $hi, ki \in H_i$, $xP_i^{hi}y$ and $wP_i^{ki}z, \forall i \in P$. Let $D_{\{i\}}$ be the dominance relation confined to the single criterion $i \in P$. The binary relation $D_{\{i\}}$ is reflexive ($(x, y)D_{\{i\}}(x, y), \forall (x, y) \in U \times U$), transitive ($(x, y)D_{\{i\}}(w, z)$ and $(w, z)D_{\{i\}}(u, v)$ imply $(x, y)D_{\{i\}}(u, v), \forall (x, y), (w, z), (u, v) \in U \times U$), and complete ($(x, y)D_{\{i\}}(w, z)$ or $(w, z)D_{\{i\}}(x, y), \forall (x, y), (w, z) \in U \times U$). Therefore, $D_{\{i\}}$ is a complete preorder on $U \times U$. Since the intersection of complete preorders is a partial preorder and $D_P = \bigcap_{i \in P} D_{\{i\}}, P \subseteq I$, then the dominance relation D_P is a partial preorder on $U \times U$.

Let $R \subseteq P \subseteq I$ and $(x, y), (u, v) \in U \times U$; then the following implication holds:

$$(x, y)D_P(u, v) \Rightarrow (x, y)D_R(u, v)\,.$$

Given $P \subseteq I$ and $(x, y) \in U \times U$, let us introduce the positive dominance set (denotation $D_P^+(x, y)$) and the negative dominance set (denotation $D_P^-(x, y)$):

$$D_P^+(x, y)) = \{(w, z) \in U \times U\colon (w, z)D_P(x, y)\}\,,$$

$$D_P^-(x, y)) = \{(w, z) \in U \times U\colon (x, y)D_P(w, z)\}\,.$$

Using the dominance relation D_P, it is possible to define P-lower and P-upper approximations of the outranking relation S with respect to $P \subseteq I$, respectively, as:

$$\underline{P}(S) = \left\{(x, y) \in \hat{B}\colon D_P^+(x, y) \subseteq S\right\}\,,$$

$$\overline{P}(S) = \bigcup_{(x,y)\in S} D_P^+(x, y)\,.$$

Analogously, it is possible to define the approximations of S^c:

$$\underline{P}(S^c) = \left\{(x, y) \in \hat{B}\colon D_P^-(x, y) \subseteq S^c\right\}\,,$$

$$\overline{P}(S^c) = \bigcup_{(x,y)\in S^c} D_P^-(x, y)\,.$$

It may be proved that

$$\underline{P}(S) \subseteq S \subseteq \overline{P}(S)$$

$$\underline{P}(S^c) \subseteq S^c \subseteq \overline{P}(S^c)\,.$$

Furthermore, the following complementarity properties hold:

$$\underline{P}(S) = \hat{B} - \overline{P}(S^c), \quad \overline{P}(S) = \hat{B} - \underline{P}(S^c)\,,$$

$$\underline{P}(S^c) = \hat{B} - \overline{P}(S), \quad \overline{P}(S^c) = \hat{B} - \underline{P}(S)\,.$$

The P-boundaries (P-doubtful regions) of S and S^c are defined as

$$Bn_P(S) = \overline{P}(S) - \underline{P}(S), \quad Bn_P(S^c) = \overline{P}(S^c) - \underline{P}(S^c)\,.$$

It is easy to prove that $Bn_P(S) = Bn_P(S^c)$.

The concepts of quality of approximation, reducts and core can be extended also to the approximation of the outranking relation by multi-graded dominance relations. In particular,

$$\gamma_P = \frac{|\underline{P}(S) \cup \underline{P}(S^c)|}{|\hat{B}|}$$

defines the *quality of approximation* of S and S^c by $P \subseteq I$. It expresses the ratio of all pairs of objects $(x, y) \in \hat{B}$ correctly assigned to S and S^c by the set P of criteria, to all the pairs of objects contained in $\hat{B}$. Each minimal subset $P' \subseteq P$ such that $\gamma_{P'} = \gamma_P$ is called a *reduct* of P (denotation $RED_{PCT}(P)$). Let us remark that $\boldsymbol{S}_{\mathrm{PCT}}$ can have more than one reduct. The intersection of all reducts is called the *core* (denotation $CORE_{PCT}(P)$).

Using the approximations defined above, it is then possible to induce a generalized description of the preference information contained in a given $\boldsymbol{S}_{\mathrm{PCT}}$ in terms of suitable decision rules. The syntax of these rules is based on the concept of upward cumulated preferences (denotation $P_i^{\geq h}$) and downward cumulated preferences (denotation $P_i^{\geq h}$), having the following interpretation:

- $xP_i^{\geq h}y$ means "x is preferred to y with respect to i by at least degree h",
- $xP_i^{\leq h}y$ means "x is preferred to y with respect to i by at most degree h".

Exact definition of the cumulated preferences, for each $(x, y) \in U \times U$, $i \in I$ and $h \in H_i$, is the following:

- $xP_i^{\geq h}y$ if $xP_i^k y$, where $k \in H_i$ and $k \geq h$,
- $xP_i^{\leq h}y$ if $xP_i^k y$, where $k \in H_i$ and $k \leq h$.

Using the above concepts, three types of decision rules can be obtained:

1. $D_{\geq}$ – decision rules, being statements of the type:

$$\textit{if } xP_{i1}^{\geq h(i1)}y \textit{ and } xP_{i2}^{\geq h(i2)}y \textit{ and } \ldots xP_{ip}^{\geq h(ip)}y, \textit{ then } xSy,$$

   where $P = \{i1, \ldots, ip\} \subseteq I$ and $(h(i1), \ldots, h(ip)) \in H_{i1} \times \cdots \times H_{ip}$; these rules are supported by pairs of objects from the P-lower approximation of S only;
2. $D_{\leq}$ – decision rules, being statements of the type:

$$\textit{if } xP_{i1}^{\leq h(i1)}y \textit{ and } xP_{i2}^{\leq h(i2)}y \textit{ and } \ldots xP_{ip}^{\leq h(ip)}y, \textit{ then } xS^c y,$$

   where $P = \{i1, \ldots, ip\} \subseteq I$ and $(h(i1), \ldots, h(ip)) \in H_{i1} \times \cdots \times H_{ip}$; these rules are supported by pairs of objects from the P-lower approximation of S^c only;
3. $D_{\geq\leq}$ – decision rules, being statements of the type:

$$\textit{if } xP_{i1}^{\geq h(i1)}y \textit{ and } xP_{i2}^{\geq h(i2)}y \textit{ and } \ldots xP_{ik}^{\geq h(ik)}y \textit{ and } xP_{ik+1}^{\leq h(ik+1)}y \textit{ and } \ldots xP_{ip}^{\leq h(ip)}y, \textit{ then } xSy \textit{ or } xS^c y,$$

   where $O' = \{i1, \ldots, ik\} \subseteq I$, $O'' = \{ik+1, \ldots, ip\} \subseteq I$, $P = O' \cup O''$, O' and O'' not necessarily disjoint, $(h(i1), \ldots, h(ip)) \in H_{i1} \times H_{i2} \cdots \times H_{ip}$; these rules are supported by objects from the P-boundary of S and S^c only.

Dominance Without Degrees of Preference

The degree of graded preference considered in Subsect. "The Pairwise Comparison Table" is defined on a *quantitative* scale of the strength of preference k_i, $i \in I$. However, in many real world problems, the existence of such a quantitative scale is rather questionable. Roy [45] distinguishes the following cases:

- Preferences expressed on an *ordinal* scale: this is the case where the difference between two evaluations has no clear meaning;
- Preferences expressed on a *quantitative* scale: this is the case where the scale is defined with reference to a unit clearly identified, such that it is meaningful to consider an origin (zero) of the scale and ratios between evaluations (ratio scale);
- Preferences expressed on a *numerical non-quantitative* scale: this is an intermediate case between the previous two; there are two well-known particular cases:
  - *Interval scale*, where it is meaningful to compare ratios between differences of pairs of evaluations,
  - Scale for which a *complete preorder* can be defined on all possible pairs of evaluations.

The preference scale has also been considered within economic theory (e. g. [47]), where cardinal utility is distinguished from ordinal utility: the former deals with a strength of preference, while, for the latter, this concept is meaningless. From this point of view, preferences expressed on an ordinal scale refer to *ordinal utility*, while preferences expressed on a quantitative scale or a numerical non-quantitative scale deal with *cardinal utility*.

The strength of preference k_i and, therefore, the graded preference considered in Subsect. "Approximation by Means of Multi-graded Dominance Relations", is meaningful when the scale is quantitative or numerical non-quantitative. If the information about k_i is non-available, then it is possible to define a rough approximation of S and S^c using a specific dominance relation between pairs of objects from $U \times U$, defined on an ordinal scale represented by evaluations $f_i(x)$ on criterion i, for $x \in U$ [14]. Let us explain this latter case in more details.

Let I^{O} be the set of criteria expressing preferences on an ordinal scale, and I^{N}, the set of criteria expressing preferences on a quantitative scale or a numerical non-quantitative scale, such that $I^{\mathrm{O}} \cup I^{\mathrm{N}} = I$ and $I^{\mathrm{O}} \cap I^{\mathrm{N}} = \emptyset$. Moreover, for each $P \subseteq I$, P^{O} denotes the subset of P composed of criteria expressing preferences on an ordinal scale, i. e. $P^{\mathrm{O}} = P \cap I^{\mathrm{O}}$, and P^{N} the subset of P composed of criteria expressing preferences on a quantitative scale or a numerical non-quantitative scale, i. e. $P^{\mathrm{N}} = P \cap I^{\mathrm{N}}$. Of course, for each $P \subseteq I$, $P = P^{\mathrm{N}} \cup P^{\mathrm{O}}$ and $P^{\mathrm{O}} \cap P^{\mathrm{N}} = \emptyset$.

If $P = P^{\mathrm{N}}$ and $P^{\mathrm{O}} = \emptyset$, then the definition of dominance is the same as in the case of multi-graded dominance (Subsect. "Approximation by Means of Multi-graded Dominance Relations"). If $P = P^{\mathrm{O}}$ and $P^{\mathrm{N}} = \emptyset$, then, given $(x, y), (w, z) \in U \times U$, the pair (x, y) is said to dominate the pair (w, z) with respect to P if, for each $i \in P$, $f_i(x) \geq f_i(w)$ and $f_i(z) \geq f_i(y)$. Let $D_{\{i\}}$ be the dominance relation confined to the single criterion $i \in P^{\mathrm{O}}$. The binary relation $D_{\{i\}}$ is reflexive ($(x, y)D_{\{i\}}(x, y)$, $\forall (x, y) \in U \times U$), transitive ($(x, y)D_{\{i\}}(w, z)$ and $(w, z)D_{\{i\}}(u, v)$ imply $(x, y)D_{\{i\}}(u, v)$, $\forall (x, y), (w, z), (u, v) \in U \times U$), but non-complete (it is possible that *not* $(x, y)D_{\{i\}}(w, z)$ and *not* $(w, z)D_{\{i\}}(x, y)$ for some $(x, y), (w, z) \in U \times U$). Therefore, $D_{\{i\}}$ is a partial preorder. Since the intersection of partial preorders is also a partial preorder and $D_P = \bigcap_{i \in P} D_{\{i\}}$, $P = P^{\mathrm{O}}$, then the dominance relation D_P is also a partial preorder. If some criteria from $P \subseteq I$ express preferences on a quantitative or a numerical non-quantitative scale and others on an ordinal scale, i. e. if $P^{\mathrm{N}} \neq \emptyset$ and $P^{\mathrm{O}} \neq \emptyset$, then, given $(x, y), (w, z) \in U \times U$, the pair (x, y) is said to dominate the pair (w, z) with respect to criteria from P, if (x, y) dominates (w, z) with respect to both P^{N} and P^{O}. Since the dominance relation with respect to P^{N} is a partial preorder on $U \times U$ (because it is a multi-graded dominance) and the dominance with respect to P^{O} is also

a partial preorder on $U \times U$ (as explained above), then also the dominance D_P, being the intersection of these two dominance relations, is a partial preorder. In consequence, all the concepts introduced in the previous subsection can be restored using this specific definition of dominance relation.

Using the approximations of S and S^c based on the dominance relation defined above, it is possible to induce a generalized description of the available preference information in terms of decision rules. These decision rules are of the same type as the rules already introduced in the previous subsection, however, the conditions on criteria from I^{O} are expressed directly in terms of evaluations belonging to value sets of these criteria. Let $F_i = \{f_i(x), x \in U\}$, $i \in I^{\mathrm{O}}$. The decision rules have in this case the following syntax:

1. $D_{\geq}$-decision rule, being a statement of the type:

$$\begin{aligned} &\textit{if } xP_{i1}^{\geq h(i1)}y \textit{ and } \ldots xP_{ie}^{\geq h(ie)}y \textit{ and } f_{ie+1}(x) \geq r_{ie+1} \\ &\textit{and } f_{ie+1}(y) \leq s_{ie+1} \textit{ and } \ldots f_{ip}(x) \geq r_{ip} \\ &\textit{and } f_{ip}(y) \leq s_{qp}, \textit{ then } xSy\,, \end{aligned}$$

where $P = \{i1, \ldots, ip\} \subseteq I$, $P^{\mathrm{N}} = \{i1, \ldots, ie\}$, $P^{\mathrm{O}} = \{ie+1, \ldots, ip\}$, $(h(i1), \ldots, h(ie)) \in H_{i1} \times \cdots \times H_{ie}$ and $(r_{ie+1}, \ldots, r_{ip}), (s_{ie+1}, \ldots, s_{ip}) \in F_{ie+1} \times \cdots \times F_{ip}$; these rules are supported by pairs of objects from the P-lower approximation of S only;

2. $D_{\leq}$-decision rule, being a statement of the type:

$$\begin{aligned} &\textit{if } xP_{i1}^{\leq h(i1)}y \textit{ and } \ldots xP_{ip}^{\leq h(ip)}y \textit{ and } f_{ie+1}(x) \leq r_{ie+1} \\ &\textit{and } f_{ie+1}(y) \geq s_{ie+1} \textit{ and } \ldots f_{ip}(x) \leq r_{ip} \\ &\textit{and } f_{ip}(y) \geq s_{ip}, \textit{ then } xS^c y\,, \end{aligned}$$

where $P = \{i1, \ldots, ip\} \subseteq I$, $P^{\mathrm{N}} = \{i1, \ldots, ie\}$, $P^{\mathrm{O}} = \{ie+1, \ldots, ip\}$, $(h(i1), \ldots, h(ie)) \in H_{i1} \times \cdots \times H_{ie}$ and $(r_{ie+1}, \ldots, r_{ip})$, $(s_{ie+1}, \ldots, s_{ip}) \in F_{ie+1} \times \cdots \times F_{ip}$; these rules are supported by pairs of objects from the P-lower approximation of S^c only;

3. $D_{\geq\leq}$-decision rule, being a statement of the type:

$$\begin{aligned} &\textit{if } xP_{i1}^{\geq h(i1)}y \textit{ and } \ldots xP_{ie}^{\geq h(ie)}y \textit{ and } xP_{ie+1}^{\leq h(ie+1)}y \ldots \\ &xP_{if}^{\leq h(if)}y \textit{ and } f_{if+1}(x) \geq r_{if+1} \\ &\textit{and } f_{if+1}(y) \leq s_{if+1} \textit{ and } \ldots f_{ig}(x) \geq r_{ig} \\ &\textit{and } f_{ig}(y) \leq s_{ig} \textit{ and } f_{ig+1}(x) \leq r_{ig+1} \\ &\textit{and } f_{ig+1}(y) \geq s_{ig+1} \textit{ and } \ldots f_{ip}(x) \leq r_{ip} \\ &\textit{and } f_{ip}(y) \geq s_{ip}, \textit{ then } xSy \textit{ or } xS^c y\,, \end{aligned}$$

where $O' = \{i1, \ldots, ie\} \subseteq I$, $O'' = \{ie+1, \ldots, if\} \subseteq I$, $P^{\mathrm{N}} = O' \cup O''$, O' and O'' not necessarily disjoint, $P^{\mathrm{O}} = \{if+1, \ldots, ip\}$, $(h(i1), \ldots, h(if)) \in H_{i1} \times \cdots \times H_{if}$ and $(r_{if+1}, \ldots, r_{ip}), (s_{if+1}, \ldots, s_{ip}) \in F_{if+1} \times \cdots \times F_{ip}$; these rules are supported by pairs of objects from the P-boundary of S and S^c only.

Rough Sets in Decision Making, Table 7
Information table of the illustrative example

Warehouse	A_1	A_2	A_3	A_4
1	high	good	no	profit
2	medium	medium	no	loss
3	medium	medium	no	profit
4	low	medium	no	loss
5	medium	good	yes	loss
6	high	medium	yes	profit

Example Illustrating DRSA in the Context of Multiple Criteria Choice and Ranking

The following example illustrates DRSA in the context of multiple criteria choice and ranking. Six warehouses have been described by means of four criteria:

- f_1, capacity of the sales staff,
- f_2, perceived quality of goods,
- f_3, high traffic location,
- f_4, warehouse profit or loss.

The components of the information table $\boldsymbol{S}$ are: $U = \{1, 2, 3, 4, 5, 6\}$, $F = \{f_1, f_2, f_3, f_4\}$, $I = \{1, 2, 3, 4\}$, $F_1 = \{\text{high, medium, low}\}$, $F_2 = \{\text{good, medium}\}$, $F_3 = \{\text{no, yes}\}$, $F_4 = \{\text{profit, loss}\}$, the criterion $f_i(x)$, taking values $f_1(1) = \text{high}$, $f_2(1) = \text{good}$, and so on.

It is assumed that the DM accepts to express preferences with respect to criteria f_1, f_2, f_3 on a *numerical non-quantitative scale*, for which a complete preorder can be defined on all possible pairs of evaluations. According to this assumption, in order to build the PCT, as described in Subsect. "Approximation by Means of Multi-graded Dominance Relation", the DM specifies sets of possible degrees of preference; for example, $H_1 = \{-2, -1, 0, 1, 2\}$, $H_2 = \{-1, 0, 1\}$, $H_3 = \{-1, 0, 1\}$. Therefore, with respect to f_1, there are the following preference relations P_1^h:

- xP_1^2y (and $yP_1^{-2}x$), meaning that x is preferred to y with respect to f_1, if $f_1(x) = \text{high}$ and $f_1(y) = \text{low}$,
- xP_1^1y (and $yP_1^{-1}x$), meaning that x is weakly preferred to y with respect to f_1, if $f_1(x) = \text{high}$ and $f_1(y) = \text{medium}$, or $f_1(x) = \text{medium}$ and $f_1(y) = \text{low}$,
- xP_1^0y (and yP_1^0x), meaning that x is indifferent to y with respect to f_1, if $f_1(x) = f_1(y)$.

Analogously, with respect to f_2 and f_3, there are the following preference relations P_2^h and P_3^h:

- xP_2^1y (and $yP_2^{-1}x$), meaning that x is weakly preferred to y with respect to f_2, if $f_2(x) =$ good and $f_2(y) =$ medium,
- xP_2^0y (and yP_2^0x), meaning that x is indifferent to y with respect to f_2, if $f_2(x) = f_2(y)$,
- xP_3^1y (and $yP_3^{-1}x$), meaning that x is weakly preferred to y with respect to f_3, if $f_3(x) =$ yes and $f_3(y) =$ no,
- xP_3^0y (and yP_3^0x), meaning that x is indifferent to y with respect to f_3, if $f_3(x) = f_3(y)$.

As to the comprehensive preference relation, the DM considers that, given two different warehouses $x, y \in U = \{1, 2, 3, 4, 5, 6\}$, if x makes profit and y makes loss, then xSy and yS^cx. Moreover, the DM accepts xSx for each warehouse x. As to warehouses x and y, which both make profit or both make loss, the DM abstains from judging whether xSy or xS^cy. Therefore, the set of exemplary pairwise comparisons supplied by the DM is $\hat{B} = \{(1,1), (1,2), (1,4), (1,5), (2,1), (2,2), (2,3), (2,6), (3,2), (3,3), (3,4), (3,5), (4,1), (4,3), (4,4), (4,6), (5,1), (5,3), (5,5), (5,6), (6,2), (6,4), (6,5), (6,6)\}$.

At this stage, the PCT can be build as shown in Table 8.

The I-lower approximations, the I-upper approximations and the I-boundaries of S and S^c obtained by means of multi-graded dominance relations are as follows:

- $\underline{I}(S) = \{(1,2), (1,4), (1,5), (3,4), (6,2), (6,4), (6,5)\}$,
- $\overline{I}(S) = \{(1,1), (1,2), (1,4), (1,5), (2,2), (2,3), (3,2), (3,3), (3,4), (3,5), (4,4), (5,3), (5,5), (6,2), (6,4), (6,5), (6,6)\}$,
- $\underline{I}(S^c) = \{(2,1), (2,6), (4,1), (4,3), (4,6), (5,1), (5,6)\}$,
- $\overline{I}(S^c) = \{(1,1), (2,1), (2,2), (2,3), (2,6), (3,2), (3,3), (3,5), (4,1), (4,3), (4,4), (4,6), (5,1), (5,3), (5,5), (5,6), (6,6)\}$,
- $Bn_I(S) = Bn_I(S^c) = \{(1,1), (2,2), (2,3), (3,2), (3,3), (3,5), (4,4), (5,3), (5,5), (6,6)\}$.

Therefore, the quality of approximation is equal to 0.58. Moreover, there is only one reduct which is also the core, i. e. $RED_S(I) = CORE_S(I) = \{1\}$.

Finally, the following decision rules can be induced (within parentheses there are the pairs of objects supporting the rule):

- *if* $xP_1^{\geq 1}y$, *then* xSy (or, in words, *if* x is at least weakly preferred to y with respect to f_1, *then* x outranks y), ((1, 2), (1, 4), (1, 5), (3, 4), (6, 2), (6, 4), (6, 5)),
- *if* $xP_1^{\leq -1}y$, *then* xS^cy (or, in words, *if* y is at least weakly preferred to x with respect to f_1, *then* x does not outrank y), ((2, 1), (2, 6), (4, 1), (4, 3), (4, 6), (5, 1), (5, 6)),
- *if* $xP_1^{\geq 0}y$, and $xP_1^{\leq 0}y$, (i. e. *if* xP_1^0y), *then* xSy *or* xS^cy (or, in words, *if* x and y are indifferent with respect to f_1, *then* x outranks y *or* x does not outrank y), ((1, 1), (2, 2), (2, 3), (3, 2), (3, 3), (3, 5), (4, 4), (5, 3), (5, 5), (6, 6)).

Rough Sets in Decision Making, Table 8
Pairwise comparison table

Pairs	P_1^h	P_2^h	P_3^h	Outranking
(1,1)	0	0	0	S
(1,2)	1	1	0	S
(1,4)	2	1	0	S
(1,5)	1	0	−1	S
(2,1)	−1	−1	0	S^c
(2,2)	0	0	0	S
(2,3)	0	0	0	S^c
(2,6)	−1	0	−1	S^c
(3,2)	0	0	0	S
(3,3)	0	0	0	S
(3,4)	1	0	0	S
(3,5)	0	−1	−1	S
(4,1)	−2	−1	0	S^c
(4,3)	−1	0	0	S^c
(4,4)	0	0	0	S
(4,6)	−2	0	−1	S^c
(5,1)	−1	0	1	S^c
(5,3)	0	1	1	S^c
(5,5)	0	0	0	S
(5,6)	−1	1	0	S^c
(6,2)	1	0	1	S
(6,4)	2	0	1	S
(6,5)	1	−1	0	S
(6,6)	0	0	0	S

Let us assume now that the DM accepts to express preferences with respect to criteria f_1, f_2, f_3 on an *ordinal scale* of preference, for which there is only information about a partial preorder on all possible pairs of evaluations. In this case, S and S^c can be approximated in the way described in Subsect. "Dominance Without Degrees of Preference", i. e. without considering degrees of preference. The I-lower approximations, the I-upper approximations and the I-boundaries of S and S^c are as follows:

- $\underline{I}(S) = \{(1,1), (1,2), (1,4), (1,5), (3,4), (4,4), (6,2), (6,4), (6,5), (6,6)\}$,
- $\overline{I}(S) = \{(1,1), (1,2), (1,4), (1,5), (2,2), (2,3), (3,2), (3,3), (3,4), (3,5), (4,4), (5,3), (5,5), (6,2), (6,4), (6,5), (6,6)\}$,

- $\underline{I}(S^c) = \{(2,1), (2,6), (4,1), (4,3), (4,6), (5,1), (5,6)\}$,
- $\overline{I}(S^c) = \{(2,1), (2,2), (2,3), (2,6), (3,2), (3,3), (3,5), (4,1), (4,3), (4,6), (5,1), (5,3), (5,5), (5,6)\}$,
- $Bn_I(S) = Bn_I(S^c) = \{(2,2), (2,3), (3,2), (3,3), (3,5), (5,3), (5,5)\}$.

Let us observe that the pairs $(1,1)$, $(4,4)$ and $(6,6)$ belong now to the I-lower approximation of S and are not contained in the I-boundaries. Therefore, the quality of approximation is equal to 0.71. Moreover, there is still only one reduct which is also the core, i. e. again $RED_S(I) = CORE_S(I) = \{1\}$.

The following decision rules are induced from the above approximations and boundaries (within parentheses there are the pairs of objects supporting the rule):

- *if* $f_1(x)$ is at least high *and* $f_1(y)$ is at most high, *then* xSy, $((1,1), (1,2), (1,4), (1,5), (6,2), (6,4), (6,5), (6,6))$,
- *if* $f_1(x)$ is at least low *and* $f_1(y)$ is at most low, *then* xSy, $((1,4), (3,4), (4,4), (6,4))$,
- *if* $f_1(x)$ is at most medium *and* $f_1(y)$ is at least high, *then* xS^cy, $((2,1), (2,6), (4,1), (4,6)(5,1), (5,6))$,
- *if* $f_1(x)$ is at most low *and* $f_1(y)$ is at least medium, *then* xS^cy, $((4,1), (4,3), (4,6))$
- *if* $f_1(x)$ is at least medium *and* $f_1(y)$ is at most medium *and* $f_1(x)$ is at most medium *and* $f_1(y)$ is at least medium, (i. e. *if* $f_1(x)$ is equal to medium *and* $f_1(y)$ is equal to medium), *then* xSy *or* xS^cy, $((2,2), (2,3), (3,2), (3,3), (3,5), (5,3), (5,5))$.

DRSA for Decision Under Uncertainty

Basic Concepts

To apply the rough set approach to decision under uncertainty, the following basic elements must be considered:

- a set $S = \{s_1, s_2, \ldots, s_u\}$ of states of the world, or simply *states*, which are supposed to be mutually exclusive and collectively exhaustive,
- an a priori probability distribution P over the states of the world: more precisely, the probabilities of states $s_1, s_2, \ldots, s_u$ are $p_1, p_2, \ldots, p_u$, respectively ($p_1 + p_2 + \cdots + p_u = 1, p_i \geq 0, i = 1, \ldots, u$),
- a set $A = \{a_1, a_2, \ldots, a_m\}$ of *acts*,
- a set $X = \{x_1, x_2, \ldots, x_r\}$ of *outcomes* or *consequences* expressed in monetary terms ($X \subseteq \Re$),
- a function $g\colon A \to X$ assigning to each pair act-state $(a_i, s_j) \in A \times S \times S$ an outcome $x \in X$,
- a set of *classes* $\boldsymbol{Cl} = \{Cl_1, Cl_2, \ldots, Cl_p\}$, such that $Cl_1 \cup Cl_2 \cup \ldots \cup Cl_p = A, Cl_r \cap Cl_q = \emptyset$ for each $r, q \in \{1, \ldots, p\}$ with $r \neq q$; the classes of $\boldsymbol{Cl}$ are preference-ordered according to the increasing order of their indices,
- a function $e\colon A \to \boldsymbol{Cl}$ assigning each act $a_i \in A$ to a class $Cl_j \in \boldsymbol{Cl}$.

In this context, two different types of dominance can be considered:

1) *Classical dominance*: given $a_p, a_q \in A$, a_p dominates a_q iff, for each possible state of the world, act a_p gives an outcome at least as good as act a_q; more formally, $g(a_p, s_j) \geq g(a_q, s_j)$, for each $s_j \in S$,
2) *Stochastic dominance*: given $a_p, a_q \in A$, a_p dominates a_q iff, for each outcome $x \in X$, act a_p gives an outcome at least as good as x with a probability at least as great as the probability that act a_q gives the same outcome, i. e. for all $x \in X$,

$$P[S(a_p, x)] \geq P[S(a_q, x)]$$

where, for each $(a_i, x) \in A \times X$, $S(a_i, x) = \{s_j \in S : g(a_i, s_j) \geq x\}$.

In [19], it has been shown how to apply stochastic dominance in this context. On the basis of an a priori probability distribution P, one can assign to each subset of states of the world $W \subseteq S$ ($W \neq \emptyset$) the probability $P(W)$ that one of the states in W is verified, i. e. $P(W) = \sum_{i:s_i \in W} p_i$, and then to build up the set Π of all possible values $P(W)$, i. e.

$$\Pi = \{\pi \in [0,1] : \pi = P(W), W \subseteq S\}\,.$$

Let us define the following functions $z\colon A \times S \to \Pi$ and $z'\colon A \times S \to \Pi$ assigning to each act-state pair $(a_i, s_j) \in A \times S$ a probability $\pi \in \Pi$, as follows:

$$z(a_i, s_j) = \sum_{r:g(a_i,s_r) \geq g(a_i,s_j)} p_r\,,$$

and

$$z'(a_i, s_j) = \sum_{r:g(a_i,s_r) \leq g(a_i,s_j)} p_r\,.$$

Therefore, $z(a_i, s_j)$ represents the probability of obtaining an outcome whose value is at least $g(a_i, s_j)$ by act a_i. Analogously, $z'(a_i, s_j)$ represents the probability of obtaining an outcome whose value is at most $g(a_i, s_j)$ by act a_i. On the basis of function $z(a_i, s_j)$, function $\rho\colon A \times \Pi \to X$ can be defined as follows:

$$\rho(a_i, \pi) = \max_{j:\, z(a_i,s_j) \geq \pi} g(a_i, s_j)\,.$$

Thus, $\rho(a_i, \pi) = x$ means that the outcome got by act a_i is greater than or equal to x with a probability at least p (i. e. a probability p or greater). On the basis of function $z'(a_i, s_j)$, the function $\rho' \colon A \times \Pi \to X$ can be defined as follows:

$$\rho'(a_i, \pi) = \min_{j:\, z(a_i, s_j) \le \pi^{g(a_i, s_j)}} .$$

$\rho'(a_i, \pi) = x$ means that the outcome got by act a_i is smaller than or equal to x with a probability at least p.

Let us observe that information given by $\rho(a_i, \pi)$ and $\rho'(a_i, \pi)$ is related. In fact, if the elements of Π, $0 = \pi_{(0)}, \pi_{(1)}, \pi_{(2)}, \ldots, \pi_{(d)} = 1$ $(d = |\Pi|)$, are reordered in such a way that $0 = \pi_{(0)} \le \pi_{(1)} \le \pi_{(2)} \le \ldots \le \pi_{(d)} = 1$, then

$$\rho\left(a_i, \pi_{(j)}\right) = \rho'\left(a_i, 1 - \pi_{(j-1)}\right) .$$

Therefore, $\rho(a_i, \pi_{(j)}) \le x$ is equivalent to $\rho'(a_i, 1 - \pi_{(j-1)}) \ge x$, $a_i \in A$, $\pi_{(j-1)}, \pi_{(j)} \in \Pi$, $x \in X$. This implies that the analysis of the possible decisions can be equivalently conducted on values of either $\rho(a_i, \pi)$ or $\rho'(a_i, \pi)$. However, from the point of view of representation of results, it is interesting to consider both values $\rho(a_i, \pi)$ and $\rho'(a_i, \pi)$. The reason is that, contrary to intuition, $\rho(a_i, \pi) \le x$ is not equivalent to the statement that by act a_i the outcome is smaller than or equal to x with a probability at least π. The following example clarifies this point. Let us consider a game a with rolling a dice, in which if the result is 1, then the gain is 1, if the result is 2 then the gain is 2, and so on. Suppose, moreover, that the dice is equilibrated and thus each result is equiprobable with probability 1/6. The values of $\rho(a_i, \pi)$ for all possible values of probability are:

$$\rho(a, 1/6) = \$6, \quad \rho(a, 2/6) = \$5, \quad \rho(a, 3/6) = \$4 ,$$

$$\rho(a, 4/6) = \$3, \quad \rho(a, 5/6) = \$2, \quad \rho(a, 6/6) = \$1 .$$

Let us remark that $\rho(a, 5/6) \le \$3$ (indeed, $\rho(a, 5/6) = \$2$, and thus $\rho(a, 5/6) \le \$3$ is true), however, this is not equivalent to the statement that by act a the outcome is smaller than or equal to \$3 with a probability at least 5/6. In fact, this is false because this probability is 3/6 (related to results 1, 2 and 3). Analogously, the values of $\rho'(a, \pi)$ for all possible values of probability are:

$$\rho'(a, 1/6) = \$1, \quad \rho'(a, 2/6) = \$2, \quad \rho'(a, 3/6) = \$3 ,$$

$$\rho'(a, 4/6) = \$4, \quad \rho'(a, 5/6) = \$5, \quad \rho'(a, 6/6) = \$6 .$$

Let us remark that $\rho'(a, 5/6) \ge \$5$ (indeed, $\rho'(a, 5/6) = \$5$, and thus $\rho'(a, 5/6) \ge \$3$ is true), however, this is not equivalent to the statement that by act a the outcome is greater than or equal to 3 with a probability at least 5/6. In fact, this is false because this probability is 4/6 (related to results 3, 4, 5 and 6). Therefore, in the context of stochastic acts, an outcome expressed in positive terms refers to $\rho(a, \pi)$ giving a lower bound of an outcome ("for act a there is a probability π to gain at least $\rho(a, \pi)$"), while an outcome expressed in negative terms refers to $\rho'(a, \pi)$ giving an upper bound of an outcome ("for act a there is a probability π to gain at most $\rho'(a, \pi)$").

Given $a_p, a_q \in A$, a_p *stochastically dominates* a_q if and only if $\rho(a_p, \pi) \ge \rho(a_q, \pi)$ for each $\pi \in \Pi$. This is equivalent to the statement: given $a_p, a_q \in A$, a_p stochastically dominates a_q if and only if $\rho'(a_p, \pi) \ge \rho'(a_q, \pi)$ for each $\pi \in \Pi$.

For example, consider the game a^* with rolling a dice, in which if the result is 1, then the gain is \$7, if the result is 2 then the gain is \$6, and so on until the case in which the result is 6 and the gain is \$2. In this case game a^* stochastically dominates game a because $\rho(a^*, 1/6) = \$7$ is not smaller than $\rho(a, 1/6) = \$6$, $\rho(a^*, 2/6) = \$6$ is not smaller than $\rho(a, 2/6) = \$5$, and so on. Equivalently, game a^* stochastically dominates game a because $\rho'(a^*, 1/6) = \$2$ is not smaller than $\rho'(a, 1/6) = \$1$, $\rho'(a^*, 2/6) = \$3$ is not smaller than $\rho'(a, 2/6) = \$2$, and so on.

DRSA can be applied in the context of decision under uncertainty considering as set of objects U the set of acts A, as set of criteria (condition attributes) I the set Π, as decision attribute $\{d\}$ the classification $\boldsymbol{Cl}$, as value set of all criteria the set X, as information function f a function $\boldsymbol{f}$ such that $\boldsymbol{f}(a_i, \pi) = \rho(a_i, \pi)$ and $\boldsymbol{f}(a_i, cl) = e(a_i)$. Let us observe that due to equivalence $\rho(a_i, \pi_{(j)}) = \rho'(a_i, 1 - \pi_{(j-1)})$, one can also consider information function $\boldsymbol{f}'(a_i, \pi) = \rho'(a_i, \pi)$.

The aim of the rough set approach to preferences under uncertainty is to explain the preferences of the DM represented by the assignments of the acts from A to the classes from $\boldsymbol{Cl}$ in terms of stochastic dominance, expressed by means of function ρ. The syntax of decision rules obtained from this rough set approach is as follows:

1) $D_{\ge}$-decision rules with the following syntax: "if $\rho(a, p_{h_1}) \ge x_{h_1}$ and …, and $\rho(a, p_{h_z}) \ge x_{h_z}$, then $a \in Cl_r^{\ge}$" (i. e. "if by act a the outcome is at least x_{h_1} with probability at least p_{h_1}, and …, and the outcome is at least x_{h_z} with probability at least p_{h_z}, then $a \in Cl_r^{\ge}$") where $p_{h_1}, \ldots, p_{h_z} \in \Pi$, $x_{h_1}, \ldots, x_{h_z} \in X$ and $r \in \{2, \ldots, p\}$;
2) $D_{\le}$-decision rules with the following syntax: "if $\rho'(a, p_{h_1}) \le x_{h_1}$ and …, and $\rho'(a, p_{h_z}) \le x_{h_z}$, then $a \in Cl_r^{\le}$" (i. e. "if by act a the outcome is at most

x_{h_1} with probability at least p_{h_1}, and …, and the outcome is at most x_{h_z} with probability at least p_{h_z}, then $a \in Cl_r^{\leq}$") where $p_{h_1}, \ldots, p_{h_z} \in \Pi, x_{h_1}, \ldots, x_{h_z} \in X$ and $r \in \{1, \ldots, p-1\}$;

3) $D_{\geq\leq}$-decision rules with the following syntax: "if $\rho(a, p_{h_1}) \geq x_{h_1}$ and …, and $\rho(a, p_{h_z}) \geq x_{h_w}$ and $\rho'(a, p_{h_{w+1}}) \leq x_{h_{w+1}}$ and …, and $\rho'(a, p_{h_z}) \leq x_{h_z}$, then $a \in Cl_s \cup Cl_{s+1} \cup \ldots Cl_t$" (i. e. "if by act a the outcome is at least x_{h_1} with probability at least p_{h_1}, and …, and the outcome is at least x_{h_w} with probability at least p_{h_w} and the outcome is at most $x_{h_{w+1}}$ with probability at least $p_{h_{w+1}}$, and …, and the outcome is at most x_{h_z} with probability at least p_{h_z}, then $a \in Cl_s \cup Cl_{s+1} \cup \ldots Cl_t$") where $p_{h_1}, \ldots, p_{h_w}, p_{h_{w+1}}, p_{h_z} \in \Pi$, $x_{h_1}, \ldots, x_{h_z} \in X$ and $s, t \in \{1, \ldots, p\}$, such that $s < t$.

According to the meaning of $\rho(a_i, p)$ and $\rho'(a_i, p)$ discussed above, $D_{\geq}$-decision rules are expressed in terms of $\rho(a_i, p)$, $D_{\leq}$-decision rules are expressed in terms of $\rho'(a_i, p)$, and $D_{\geq\leq}$-decision rules are expressed in terms of both $\rho(a_i, p)$ and $\rho'(a_i, p)$. Let us observe that due to equivalence $\rho(a_i, \pi_{(j)}) = \rho'(a_i, 1 - \pi_{(j-1)})$, all above decision rules can be expressed equivalently in terms of values of $\rho(a_i, p)$ or $\rho'(a_i, p)$. For example, a $D_{\geq}$-decision rule $r_{\geq}(\rho)$ ="if $\rho(a, p_{h_1}) \geq x_{h_1}$ and …, and $\rho(a, p_{h_z}) \geq x_{h_z}$, then $a \in Cl_r^{\geq}$" can be expressed in terms of $\rho'(a_i, p)$ as $r_{\geq}(\rho')$ ="if $\rho'(a, p^*_{h_1}) \geq x_{h_1}$ and …, and $\rho'(a, p^*_{h_z}) \geq x_{h_z}$, then $a \in Cl_r^{\geq}$", where, if $p_{h_r} = \pi(j_r)$, then $p^*_{h_r} = 1 - \pi(j_r - 1)$, with $r = 1, \ldots, z$, and $0 = \pi(0), \pi(1), \pi(2), \ldots, \pi(|P|) = 1$ reordered in such a way that $0 = \pi(0) \leq \pi(1) \leq \pi(2) \leq \cdots \leq \pi(|P|) = 1$. Analogously, a $D_{\leq}$-decision rule $r_{\leq}(\rho')$ ="if $\rho'(a, p_{h_1}) \leq x_{h_1}$ and …, and $\rho'(a, p_{h_z}) \leq x_{h_z}$, then $a \in Cl_r^{\leq}$" can be expressed in terms of $\rho(a_i, p)$ as $r_{\leq}(\rho)$ ="if $\rho(a, p^*_{h_1}) \leq x_{h_1}$ and …, and $\rho(a, p^*_{h_z}) \leq x_{h_z}$, then $a \in Cl_r^{\leq}$, where, if $p_{h_r} = \pi(j_r)$, then $p^*_{h_r} = 1 - \pi(j_r - 1)$, with $r = 1, \ldots, z$, and $0 = \pi(0), \pi(1), \pi(2), \ldots, \pi(|P|) = 1$ reordered in such a way that $0 = \pi(0) \leq \pi(1) \leq \pi(2) \leq \cdots \leq \pi(|P|) = 1$.

Let us observe, however, that $r_{\geq}(\rho)$ is an expression much more natural and meaningful than $r_{\geq}(\rho')$, as well as $r_{\leq}(\rho')$ is an expression much more natural and meaningful than $r_{\leq}(\rho)$. Another useful remark concerns minimality of rules, related to the specific intrinsic structure of the stochastic dominance. Let us consider the following two decision rules:

- $r1 \equiv$ "if by act a the outcome is at least 100 with probability at least 0.25, then a is at least good",
- $r2 \equiv$ "if by act a the outcome is at least 100 with probability at least 0.50, then a is at least good".

$r1$ and $r2$ can be induced from the analysis of the same information table, because they involve different criteria (condition attributes). In fact, $r1$ involves attribute $\rho(a, 0.25)$ (it can be expressed as "if $\rho(a, 0.25) \geq 100$, then a is at least good"), $r2$ involves attribute $\rho(a, 0.50)$ (it can be expressed as "if $\rho(a, 0.50) \geq 100$, then a is at least good"). Considering the structure of the stochastic dominance, the condition part of rule $r1$ is the weakest. In fact, rule $r1$ requires a cumulated outcome to be at least 100 with probability of 0.25, while rule $r2$ requires the same outcome but with a greater probability, 0.5 against 0.25. Since the decision part of these two rules is the same, $r1$ is minimal among these two rules. From a practical point of view, this observation says that, if one induces decision rules using the algorithms designed for DRSA, it is necessary to further filter the obtained results in order to remove rules which are not minimal in the specific context of the DRSA analysis based on stochastic dominance.

Example Illustrating DRSA in the Context of Decision Under Uncertainty

The following example illustrates the approach. Let us consider

- a set $S = \{s_1, s_2, s_3\}$ of states of the world,
- an a priori probability distribution P over the states of the world defined as follows: $p_1 = 0.25, p_2 = 0.35, p_3 = 0.40$,
- a set $A = \{a_1, a_2, a_3, a_4, a_5, a_6\}$ of acts,
- a set $X = \{0, 10, 15, 20, 30\}$ of consequences
- a set of classes $\boldsymbol{Cl} = \{Cl_1, Cl_2, Cl_3\}$, where Cl_1 is the set of bad acts, Cl_2 is the set of medium acts, Cl_3 is the set of good acts,
- a function $g\colon A \to X$ assigning to each act-state pair $(a_i, s_j) \in A \times S \times S$ a consequence $x_h \in X$ and a function $e\colon A \to \boldsymbol{Cl}$ assigning each act $a_i \in A$ to a class $Cl_j \in \boldsymbol{Cl}$ presented in the following Table 9.

Rough Sets in Decision Making, Table 9
Acts, consequences and assignment to classes from *Cl*

	p_j	a_1	a_2	a_3	a_4	a_5	a_6
s_1	0.25	30	0	15	0	20	10
s_2	0.35	10	20	0	15	10	20
s_3	0.40	10	20	20	20	20	20
Cl		good	medium	medium	bad	medium	good

Rough Sets in Decision Making, Table 10
Acts, values of function $\rho(a_i, p)$ and assignment to classes from *Cl*

π	a_1	a_2	a_3	a_4	a_5	a_6
0.25	30	20	20	20	20	20
0.35	10	20	20	20	20	20
0.40	10	20	20	20	20	20
0.60	10	20	15	15	20	20
0.65	10	20	15	15	20	20
0.75	10	20	0	15	10	20
1	10	0	0	0	10	10
Cl	good	medium	medium	bad	medium	good

Table 10 shows the values of function $\rho(a_i, p)$.

Table 10 is the decision table on which the DRSA is applied. Let us give some examples of the interpretation of the values in Table 10. The column of act a_3 can be read as follow:

- the value 20 in the row corresponding to 0.25 means that the outcome is at least 20 with a probability of at least 0.25,
- the value 15 in the row corresponding to 0.65 means that the outcome is at least 15 with a probability of at least 0.65,
- the value 0 in the row corresponding to 0.75 means that the outcome is at least 0 with a probability of at least 0.75.

Analogously, the row corresponding to 0.65, can be read as follows:

- the value 10 relative to a_1, means that by act a_1 the outcome is at least 10 with a probability of at least 0.65,
- the value 20 relative to a_2, means that by act a_2 the outcome is at least 20 with a probability of at least 0.65,
- and so on.

Applying DRSA, the following upward union and downward union of classes are approximated:

- $Cl_2^{\geq} = Cl_2 \cup Cl_3$, i. e. the set of the acts at least medium,
- $Cl_3^{\geq} = Cl_3$, i. e. the set of the acts (at least) good,
- $Cl_1^{\leq} = Cl_1$, i. e. the set of the acts (at most) bad,
- $Cl_2^{\leq} = Cl_1 \cup Cl_2$, i. e. the set of the acts at most medium.

The first result of the DRSA approach was a discovery that the decision table (Table 10) is not consistent. Indeed, Table 10 shows that act a_4 stochastically dominates act a_3, however act a_3 is assigned to a better class (medium) than act a_4 (bad). Therefore, act a_3 cannot be assigned without doubts to the set of the class of the at least medium acts as well as act a_4 cannot be assigned without doubts to the set of the classes of the (at most) bad acts. In consequence, lower approximation and upper approximation of $Cl_2^{\geq}$, $Cl_3^{\geq}$ and $Cl_1^{\leq}$, $Cl_2^{\leq}$ are equal, respectively, to

- $\underline{I}\left(Cl_2^{\geq}\right) = \{a_1, a_2, a_5, a_6\} = Cl_2^{\geq} - \{a_3\}$,
- $\overline{I}\left(Cl_2^{\geq}\right) = \{a_1, a_2, a_3, a_4, a_5, a_6\} = Cl_2^{\geq} \cup \{a_4\}$,
- $\underline{I}\left(Cl_3^{\geq}\right) = \{a_1, a_6\} = Cl_3^{\geq}$,
- $\overline{I}\left(Cl_3^{\geq}\right) = \{a_1, a_6\} = Cl_3^{\geq}$,
- $\underline{I}\left(Cl_1^{\leq}\right) = \emptyset = Cl_1^{\leq} - \{a_4\}$,
- $\overline{I}\left(Cl_1^{\leq}\right) = \{a_3, a_4\} = Cl_1^{\leq} \cup \{a_3\}$,
- $\underline{I}\left(Cl_2^{\leq}\right) = \{a_2, a_3, a_4, a_5\} = Cl_2^{\leq}$,
- $\overline{I}\left(Cl_2^{\leq}\right) = \{a_2, a_3, a_4, a_5\} = Cl_2^{\leq}$.

Since there are two inconsistent acts on a total of six acts (a_3, a_4), then the quality of approximation of the ordinal classification is equal to 4/6. The second discovery was one reduct of criteria (condition attributes) ensuring the same quality of approximation as the whole set Π of probabilities: $\mathrm{RED}_{Cl} = \{0.25, 0.75, 1\}$. This means that the preferences of the DM can be explained using only the probabilities from RED_{Cl}. RED_{Cl} is also the core because no probability value in RED_{Cl} can be removed without deteriorating the quality of approximation. The third discovery was a set of minimal decision rules describing the DM's preferences (within parentheses there is a verbal interpretation of the corresponding decision rule, and the supporting acts):

1) if $\rho(a_i, 0.25) \geq 30$, then $a_i \in Cl_3^{\geq}$ (*if the probability of gaining at least 30 is at least 0.25, then act a_i is at least good*) (a_1),
2) if $\rho(a_i, 0.75) \geq 20$ and $\rho(a_i, 1) \geq 10$, then $a_i \in Cl_3^{\geq}$ (*if the probability of gaining at least 20 is at least 0.75 and the probability of gaining at least 10 is (at least) 1 (i. e. for sure the gaining is at least 10), then act a_i is at least good*) (a_6),
3) if $\rho(a_i, 1) \geq 10$, then $a_i \in Cl_2^{\geq}$ (*if the probability of gaining at least 10 is (at least) 1 (i. e. for sure the gaining is at least 10), then act a_i is at least medium*) (a_1, a_5, a_6),
4) if $\rho(a_i, 0.75) \geq 20$, then $a_i \in Cl_2^{\geq}$ (*if the probability of gaining at least 20 is at least 0.75, then act a_i is at least medium*) (a_2, a_6),
5) if $\rho(a_i, 0.25) \leq 20$ (i. e. $\rho'(a_i, 1) \leq 20$) and $\rho(a_i, 0.75) \leq 15$ (i. e. $\rho'(a_i, 0.35) \leq 15$), then $a_i \in Cl_2^{\leq}$ (*if the probability of gaining at most 20 is (at least) 1 (i. e. for sure you gain at most 20) and the probability to gain at most 15 is at least 0.35, then act a_i is at most medium*) (a_3, a_4, a_5),

6) if $\rho(a_i, 1) \leq 0$ (i. e. $\rho'(a_i, 0.25) \leq 0$), then $a_i \in Cl_2^{\leq}$ (*if the probability of gaining at most 0 is at least 0.25, then act a_i is at most medium*) (a_2, a_3, a_4),
7) if $\rho(a_i, 1) \geq 0$ and $\rho(a_i, 1) \leq 0$ (i. e. $\rho(a_i, 1) = 0$) and $\rho(a_i, 0.75) \leq 15$ ($\rho'(a_i, 0.35) \leq 15$), then $a_i \in Cl_1 \cup Cl_2$ (*if the probability of gaining at least 0 is 1, then act a_i is at most medium*) (a_2, a_3, a_4).

Minimal sets of minimal decision rules represent the most concise and non-redundant knowledge contained in Table 9 (and, consequently, in Table 10). The above minimal set of 7 decision rules uses 3 attributes (probability 0.25, 0.75 and 1) and 11 elementary conditions, i. e. 26% of descriptors from the original data table (Table 10). Of course, this is only a didactic example. Representation in terms of decision rules of larger sets of exemplary acts from real applications are more synthetic in the sense of the percentage of used descriptors from the original decision table.

Multiple Criteria Decision Analysis Using Association Rules

In multiple criteria decision analysis, the DM is often interested in relationships between attainable values of criteria. This information is particularly useful in multiobjective optimization (see [32] and Sect. "Interactive Multiobjective Optimization Using DRSA (IMO-DRSA)"). For instance, in a car selection problem, one can observe that in the set of considered cars, if the maximum speed is at least 200 km/h and the time to reach 100 km/h is at most 7 s, then the price is not less than 40,000 $ and the fuel consumption is not less than 9 liters per 100 km. These relationships are association rules whose general syntax, in case of minimization of criteria f_i, $i \in I$, is:

> "if $f_{i_1}(x) \leq r_{i_1}$ and … and $f_{i_p}(x) \leq r_{i_p}$, then $f_{i_{p+1}}(x) \geq r_{i_{p+1}}$ and … and $f_{i_q}(x) \geq r_{i_q}$", where $\{i_1, \ldots, i_q\} \subseteq I, r_{i_1}, \ldots, r_{i_q} \in \Re$.

If criterion f_i, $i \in I$, should be maximized, the corresponding condition in the association rule should be reversed, i. e. in the premise, the condition becomes $f_i(x) \geq r_i$, and in the conclusion it becomes $f_i(x) \leq r_i$.

Given an association rule $r \equiv$ "if $f_{i_1}(x) \leq r_{i_1}$ and … and $f_{i_p}(x) \leq r_{i_p}$, then $f_{i_{p+1}}(x) \geq r_{i_{p+1}}$ and … and $f_{i_q}(x) \geq r_{i_q}$", an object $y \in U$ *supports* r if $f_{i_1}(y) \leq r_{i_1}$ and … and $f_{i_p}(y) \leq r_{i_p}$ and $f_{i_{p+1}}(y) \geq r_{i_{p+1}}$ and … and $f_{i_q}(y) \geq r_{i_q}$. Moreover, object $y \in U$ supporting decision rule r is a *base* of r if $f_{i_1}(y) = r_{i_1}$ and … and $f_{i_p}(y) = r_{i_p}$ and $f_{i_{p+1}}(y) = r_{i_{p+1}}$ and … and $f_{i_q}(y) = r_{i_q}$. An association rule having at least one base is called *robust*.

An association rule $r \equiv$ "if $f_{i_1}(x) \leq r_{i_1}$ and … and $f_{i_p}(x) \leq r_{i_p}$, then $f_{i_{p+1}}(x) \geq r_{i_{p+1}}$ and … and $f_{i_q}(x) \geq r_{i_q}$" holds in universe U if:

1) there is at least one $y \in U$ supporting r,
2) r is not contradicted in U, i. e. there is no $z \in U$ such that $f_{i_1}(z) \leq r_{i_1}$ and … and $f_{i_p}(z) \leq r_{i_p}$, while *not* $f_{i_{p+1}}(z) \geq r_{i_{p+1}}$ or … or $f_{i_q}(z) \geq r_{i_q}$.

Given the two association rules:

- $r_1 \equiv$"if $f_{i_1}(x) \leq r^1_{i_1}$ and … and $f_{i_p}(x) \leq r^1_{i_p}$, then $f_{i_{p+1}}(x) \geq r^1_{i_{p+1}}$ and … and $f_{i_q}(x) \geq r^1_{i_q}$",
- $r_2 \equiv$"if $f_{j_1}(x) \leq r^2_{j_1}$ and … and $f_{j_s}(x) \leq r^2_{j_s}$, then $f_{j_{s+1}}(x) \geq r^2_{j_{s+1}}$ and … and $f_{j_t}(x) \geq r^2_{j_t}$",

rule r_1 is not weaker than rule r_2, denoted by $r_1 \trianglerighteq r_2$, if:

α) $\{i_1, \ldots, i_p\} \subseteq \{j_1, \ldots, j_s\}$,
β) $r^1_{i_1} \geq r^2_{i_1}, \ldots, r^1_{i_p} \geq r^2_{i_p}$,
γ) $\{i_{p+1}, \ldots, i_q\} \supseteq \{j_{s+1}, \ldots, j_t\}$,
δ) $r^1_{j_{s+1}} \geq r^2_{j_{s+1}}, \ldots, r^1_{j_t} \geq r^2_{j_t}$.

Conditions β and δ are formulated for criteria f_i to be minimized. If criterion f_i should be maximized, the corresponding inequalities should be reversed, i. e. $r^1_i \leq r^2_i$ in condition β as well as in condition δ. Notice that $\trianglerighteq$ is a binary relation on the set of association rules, which is a partial preorder, i. e. it is reflexive (each rule is not weaker than itself) and transitive. The asymmetric part of the relation $\trianglerighteq$ is denoted by $\triangleright$, and $r_1 \triangleright r_2$ reads "r_1 is stronger than r_2".

For example, consider the following association rules:

- $r_1 \equiv$ "if the maximum speed is at least 200 km/h and the time to reach 100 km/h is at most 7 s, then the price is not less than 40,000 $ and the fuel consumption is not less than 9 liters per 100 km",
- $r_2 \equiv$ "if the maximum speed is at least 200 km/h and the time to reach 100 km/h is at most 7 s and the horse power is at least 175 kW, then the price is not less than 40,000 $ and the fuel consumption is not less than 9 liters per 100 km",
- $r_3 \equiv$ "if the maximum speed is at least 220 km/h and the time to reach 100 km/h is at most 7 s, then the price is not less than 40,000 $ and the fuel consumption is not less than 9 liters per 100 km",
- $r_4 \equiv$ "if the maximum speed is at least 200 km/h and the time to reach 100 km/h is at most 7 s, then the price is not less than 40,000 $",
- $r_5 \equiv$ "if the maximum speed is at least 200 km/h and the time to reach 100 km/h is at most 7 s, then the price

is not less than 35,000 $ and the fuel consumption is not less than 9 liters per 100 km",

- $r_6 \equiv$ "if the maximum speed is at least 220 km/h and the time to reach 100 km/h is at most 7 s and the horse power is at least 175 kW, then the price is not less than 35,000 $".

Let us observe that rule r_1 is stronger than each of the other five rules for the following reasons:

- $r_1 \rhd r_2$ for condition α) because, all things equal elsewhere, in the premise of r_2 there is an additional condition: "the horse power is at least 175 kW",
- $r_1 \rhd r_3$ for condition β) because, all things equal elsewhere, in the premise of r_3 there is a condition with a worse threshold value: "the maximum speed is at least 220 km/h" instead of "the maximum speed is at least 200 km/h",
- $r_1 \rhd r_4$ for condition γ) because, all thing equal elsewhere, in the conclusion of r_4 one condition is missing: "the fuel consumption is not less than 9 liters per 100 km",
- $r_1 \rhd r_5$ for condition δ) because, all thing equal elsewhere, in the conclusion of r_5 there is a condition with a worse threshold value: "the price is not less than 35,000 $" instead of "the price is not less than 40,000 $",
- $r_1 \rhd r_6$ for conditions α), β), γ) and δ) because all weak points for which rules r_2, r_3, r_4 and r_5 are weaker than rule r_1 are present in r_6.

An association rule r is *minimal* if there is no other rule stronger than r with respect to $\rhd$. An algorithm for induction of association rules from preference ordered data has been presented in [20].

Interactive Multiobjective Optimization Using DRSA (IMO-DRSA)

This section presents a recently proposed method for Interactive Multiobjective Optimization using Dominance-based Rough Set Approach (IMO-DRSA) [32]. Assuming that objective functions $f_{i,j=1,\ldots,n}$, are to minimized, the method is composed of the following steps.

***Step* 1.** Generate a representative sample of solutions from the currently considered part of the Pareto optimal set.

***Step* 2.** Present the sample to the DM, possibly together with association rules showing relationships between attainable values of objective functions in the Pareto optimal set.

***Step* 3.** If the DM is satisfied with one solution from the sample, then this is the most preferred solution and the procedure stops. Otherwise continue.

***Step* 4.** Ask the DM to indicate a subset of relatively "good" solutions in the sample.

***Step* 5.** Apply DRSA to the current sample of solutions classified into "good" and "others" solutions, in order to induce a set of decision rules with the following syntax "if $f_{j_1}(\mathbf{x}) \le \alpha_{j_1}$ and … and $f_{j_p}(\mathbf{x}) \le \alpha_{j_p}$, then solution $\mathbf{x}$ is good", $\{j_1, \ldots, j_p\} \subseteq \{1, \ldots, n\}$.

***Step* 6.** Present the obtained set of rules to the DM.

***Step* 7.** Ask the DM to select the decision rules most adequate to his/her preferences.

***Step* 8.** Adjoin the constraints $f_{j_1}(\mathbf{x}) \le \alpha_{j_1}, \ldots, f_{j_p}(\mathbf{x}) \le \alpha_{j_p}$ coming from the rules selected in *Step* 7 to the set of constraints imposed on the Pareto optimal set, in order to focus on a part interesting from the point of view of DM's preferences.

***Step* 9.** Go back to *Step* 1.

In a sequence of iterations, the method is exploring the Pareto optimal set of a multiobjective optimization problem or an approximation of this set. In the calculation stage (*Step* 1), any multiobjective optimization method, which finds the Pareto optimal set or its approximation, such as Evolutionary Multiobjective Optimization methods, can be used. In the dialogue stage of the method (*Step* 2 to 7), the DM is asked to select a decision rule induced from his/her preference information, which is equivalent to fixing some upper bounds for the minimized objective functions f_j.

In *Step* 1, the representative sample of solutions from the currently considered part of the Pareto optimal set can be generated using one of existing procedures, such as [33,57,58]. It is recommended to use a fine grained sample of representative solutions to induce association rules, however, the sample of solutions presented to the DM in *Step* 2 should be much smaller (about a dozen) in order to avoid an excessive cognitive effort of the DM. Otherwise, the DM would risk to give non reliable information.

The association rules presented in *Step* 2 help the DM in understanding what (s)he can expect from the optimization problem. More precisely, any association rule

"if $f_{i_1}(x) \le r_{i_1}$ and … and $f_{i_p}(x) \le r_{i_p}$, then $f_{i_{p+1}}(x) \ge r_{i_{p+1}}$ and … and $f_{i_q}(x) \ge r_{i_q}$", where $\{i_1, \ldots, i_q\} \subseteq I, r_{i_1}, \ldots, r_{i_q} \in \Re$

says to the DM that, if (s)he wants attain the values of objective functions $f_{i_1}(x) \le r_{i_1}$ and … and $f_{i_p}(x) \le r_{i_p}$, then (s)he cannot reasonably expect to obtain values of objective functions $f_{i_{p+1}}(x) < r_{i_{p+1}}$ and … and $f_{i_q}(x) < r_{i_q}$.

With respect to the ordinal classification of solutions into the two classes of "good" and "others", observe that "good" means in fact "relatively good", i. e. better than the rest. In case, the DM would refuse to classify as "good" any solution, one can ask the DM to specify some minimal requirements of the type $f_{j_1}(\mathbf{x}) \leq \alpha_{j_1}$ and … and $f_{j_p}(\mathbf{x}) \leq \alpha_{j_p}$ for "good" solutions. These minimal requirements give some constraints that can be used in *Step* 8, in the same way as the analogous constraints coming from selected decisions rules.

The rules considered in *Step* 5 have a syntax corresponding to minimization of objective functions. In case of maximization of an objective function f_j, the condition concerning this objective in the decision rule should have the form $f_j(\mathbf{x}) \geq \alpha_j$.

Remark, moreover, that the Pareto optimal set reduced in *Step* 8 by constraints $f_{j_1}(\mathbf{x}) \leq \alpha_{j_1}, \ldots, f_{j_p}(\mathbf{x}) \leq \alpha_{j_p}$ is certainly not empty if these constraints are coming from one decision rule only. Since robust rules (see the glossary Dominance-Based Rough Set Approach (DRSA)) are considered, the threshold values $\alpha_{j_1}, \ldots, \alpha_{j_p}$ are values of objective functions of some solutions from the Pareto optimal set. If $\{j_1, \ldots, j_p\} = \{1, \ldots, n\}$, i. e. $\{j_1, \ldots, j_p\}$ is the set of all objective functions, then the new reduced part of the Pareto optimal set contains only one solution $\mathbf{x}$ such that $f_1(\mathbf{x}) = \alpha_1, \ldots, f_n(\mathbf{x}) = \alpha_n$. If $\{j_1, \ldots, j_p\} \subset \{1, \ldots, n\}$, i. e. $\{j_1, \ldots, j_p\}$ is a proper subset of the set of all objective functions, then the new reduced part of the Pareto optimal set contains solutions satisfying conditions $f_{j_1}(\mathbf{x}) \leq \alpha_{j_1}$ and … and $f_{j_p}(\mathbf{x}) \leq \alpha_{j_p}$. Since the considered rules are robust, then there is at least one solution $\mathbf{x}$ satisfying these constraints. When the Pareto optimal set is reduced in *Step* 8 by constraints $f_{j_1}(\mathbf{x}) \leq \alpha_{j_1}, \ldots, f_{j_p}(\mathbf{x}) \leq \alpha_{j_p}$ coming from more than one rule, then it is possible that the resulting reduced part of the Pareto optimal set is empty. Thus, before passing to *Step* 9, it is necessary to verify if the reduced Pareto optimal set is not empty. If the reduced Pareto optimal set is empty, then the DM is required to revise his/her selection of rules. The DM can be supported in this task, by information about minimal sets of constraints $f_j(\mathbf{x}) \leq \alpha_j$ coming from the considered decision rules to be removed in order to get a non-empty part of the Pareto optimal set.

The constraints introduced in *Step* 8 are maintained in the following iterations of the procedure, however, they cannot be considered as irreversible. Indeed, the DM can come back to the Pareto optimal set considered in one of previous iterations and continue from this point. This is in the spirit of a learning oriented conception of interactive multiobjective optimization, i. e. it agrees with the idea that the interactive procedure permits the DM to learn about his/her preferences and about the "shape" of the Pareto optimal set.

Example Illustrating IMO-DRSA in the Context of Multiobjective Optimization

To illustrate the interactive multiobjective optimization procedure based on DRSA, a product mix problem is considered. There are three products: A, B, C which are produced in quantities denoted by x_A, x_B, and x_C, respectively. The unit prices of the three products are $p_A = 20$, $p_B = 30, p_C = 25$. The production process involves two machines. The production times of A, B, C on the first machine are equal to $t_{1A} = 5$, $t_{1B} = 8$, $t_{1C} = 10$, and on the second machine they are equal to $t_{2A} = 8, t_{2B} = 6$, $t_{2C} = 2$. Two raw materials are used in the production process. The first raw material has a unit cost of 6 and the quantity required for production of one unit of A, B and C is $r_{1A} = 1, r_{1B} = 2$ and $r_{1C} = 0.75$, respectively. The second raw material has a unit cost of 8 and the quantity required for production of one unit of A, B and C is $r_{2A} = 0.5, r_{2B} = 1$ and $r_{2C} = 0.5$, respectively. Moreover, the market cannot absorb a production greater than 10, 20 and 10 units for A, B and C, respectively. To decide how much of A, B and C should be produced, the following objectives have to be taken into account:

- Profit (to be maximized),
- Time (total production time on two machines – to be minimized),
- Production of A (to be maximized),
- Production of B (to be maximized),
- Production of C (to be maximized),
- Sales (to be maximized).

The above product mix problem can be formulated as the following multiobjective optimization problem:

$$\begin{aligned}
&\text{Maximize} \\
&20x_A + 30x_B + 25x_C - (1x_A + 2x_B + 0.75x_C)6 \\
&\qquad - (0.5x_A + 1x_B + 0.5x_C)8 \quad [\text{Profit}]\,, \\
&\text{Minimize} \quad 5x_A + 8x_B + 10x_C + 8x_A + 6x_B + 2x_C \quad [\text{Time}], \\
&\text{Maximize} \quad x_A \quad [\text{Production of } A]\,, \\
&\text{Maximize} \quad x_B \quad [\text{Production of } B]\,, \\
&\text{Maximize} \quad x_C \quad [\text{Production of } C]\,, \\
&\text{Maximize} \quad 20x_A + 30x_B + 25x_C \quad [\text{Sales}]\,,
\end{aligned}$$

Rough Sets in Decision Making, Table 11
A sample of Pareto optimal solutions proposed in the first iteration

Solution	Profit	Time	Prod. *A*	Prod. *B*	Prod. *C*	Sales	Evaluation
*s*1	165	120	0	0	10	250	*
*s*2	172.69	130	0.769	0	10	265.38	*
*s*3	180.38	140	1.538	0	10	280.77	good
*s*4	141.13	140	3	3	4.92	272.92	good
*s*5	148.38	150	5	2	4.75	278.75	good
*s*6	139.13	150	5	3	3.58	279.58	*
*s*7	188.08	150	2.308	0	10	296.15	*
*s*8	159	150	6	0	6	270	*
*s*9	140.5	150	6	2	3.67	271.67	good
*s*10	209.25	200	6	2	7.83	375.83	*
*s*11	189.38	200	5	5	5.42	385.42	*
*s*12	127.38	130	3	3	4.08	252.08	*
*s*13	113.63	120	3	3	3.25	231.25	*

subject to:

$$x_A \leq 10\,, \quad x_B \leq 20\,, \quad x_C \leq 10$$
[Market absorption limits] ,

$$x_A \geq 0\,, \quad x_B \geq 0\,, \quad x_C \geq 0$$
[Non-negativity constraints] .

A sample of representative Pareto optimal solutions has been calculated and proposed to the DM. Observe that the considered problem is a Multiple Objective Linear Programming (MOLP) problem, and thus representative Pareto optimal solutions can be calculated using classical linear programming looking for the solutions optimizing each one of the considered objectives or fixing all the considered objective functions but one at a satisfying value, and looking for the solution optimizing the remaining objective function. The set of representative Pareto optimal solutions is shown in Table 11. Moreover, a set of potentially interesting association rules have been induced from the sample and presented to the DM. These rules represent strongly supported relationships between attainable values of objective functions. The association rules are the following (between parentheses there are id's of solutions supporting the rule):

1) if Time $\leq$ 140, then Profit $\leq$ 180.38 and Sales $\leq$ 280.77 ($s1, s2, s3, s4, s12, s13$),
2) if Time $\leq$ 150, then Profit $\leq$ 188.08 and Sales $\leq$ 296.15 ($s1, s2, s3, s4, s5, s6, s7, s8, s9, s12, s13$),
3) if $x_B \geq 2$, then Profit $\leq$ 209.25 and $x_A \leq 6$ and $x_C \leq 7.83$ ($s4, s5, s6, s9, s10, s11, s12, s13$),
4) if Time $\leq$ 150, then $x_B \leq 3$ ($s1, s2, s3, s4, s5, s6, s7, s8, s9, s12, s13$),
5) if Profit $\geq$ 148.38 and Time $\leq$ 150, then $x_B \leq 2$ ($s1, s2, s3, s5, s7, s8$),
6) if $x_A \geq 5$, then Time $\geq$ 150 ($s5, s6, s8, s9, s10, s11$),
7) if Profit $\geq$ 127.38 and $x_A \geq 3$, then Time $\geq$ 130 ($s4, s5, s6, s8, s9, s10, s11, s12$),
8) if Time $\leq$ 150 and $x_B \geq 2$, then Profit $\leq$ 148.38 ($s4, s5, s6, s9, s12, s13$),
9) if $x_A \geq 3$ and $x_C \geq 4.08$, then Time $\geq$ 130 ($s4, s5, s8, s10, , s11, s12$),
10) if Sales $\geq$ 265.38, then Time $\geq$ 130 ($s2, s3, s4, s5, s6, s7, s8, s9, s10, s11$).

Then, the DM has been asked if (s)he was satisfied with one of the proposed Pareto optimal solutions. Since his/her answer was negative, (s)he was requested to indicate a subset of relatively "good" solutions which are indicated in the "Evaluation" column of Table 11.

Taking into account the ordinal classification of Pareto optimal solutions into "good" and "others", made by the DM, twelve decision rules have been induced from the lower approximation of "good" solutions. The frequency of the presence of objectives in the premises of the rules gives a first idea of the importance of the considered objectives. These frequencies are the following:

- Profit: $\frac{4}{12}$,
- Time: $\frac{12}{12}$,
- Production of A : $\frac{7}{12}$,
- Production of B : $\frac{4}{12}$,
- Production of C : $\frac{5}{12}$,
- Sales: $\frac{5}{12}$.

The following potentially interesting decision rules were presented to the DM:

1) if Profit $\geq$ 140.5 and Time $\leq$ 150 and $x_B \geq 2$, then product mix is good ($s4$, $s5$, $s9$),
2) if Time $\leq$ 140 and $x_A \geq 1.538$ and $x_C \geq 10$, then product mix is good ($s3$),
3) if Time $\leq$ 150 and $x_B \geq 2$ and $x_C \geq 4.75$, then product mix is good ($s4$, $s5$),
4) if Time $\leq$ 140 and Sales $\geq$ 272.9167, then product mix is good ($s3$, $s4$),
5) if Time $\leq$ 150 and $x_B \geq 2$ and $x_C \geq 3.67$ and Sales $\geq$ 271.67, then product mix is good ($s4$, $s5$, $s9$).

Among these decision rules, the DM has selected rule 1) as the most adequate to his/her preferences. This rule permits to define the following constraints reducing the feasible region of the production mix problem:

- $20x_A+30x_B+25x_C-(x_A+2x_B+0.75x_C)6-(0.5x_A+x_B+0.5x_C)8 \geq 140.5$ [Profit $\geq$ 140.5],
- $5x_A+8x_B+10x_C+8x_A+6x_B+2x_C \leq 150$ [Time $\leq$ 150],
- $x_B \geq 2$ [Production of $B \geq 2$].

These constraints have been considered together with the original constraints for the production mix problem, and a new sample of representative Pareto optimal solutions shown in Table 12 have been calculated and presented to the DM, together with the following potentially interesting association rules:

1′) if Time $\leq$ 140, then Profit $\leq$ 174 and $x_C \leq 9.33$ and Sales $\leq$ 293.33 ($s5'$, $s6'$, $s7'$, $s8'$, $s9'$, $s10'$, $s11'$, $s12'$),
2′) if $x_A \geq 2$, then $x_B \leq 3$ and Sales $\leq$ 300.83 ($s2'$, $s3'$, $s4'$, $s6'$, $s7'$, $s9'$),
3′) if $x_A \geq 2$, then Profit $\leq$ 172 and $x_C \leq 8$ ($s2'$, $s3'$, $s4'$, $s6'$, $s7'$, $s9'$),
4′) if Time $\leq$ 140, then $x_A \leq 2$ and $x_B \leq 3$ ($s5'$, $s6'$, $s7'$, $s8'$, $s9'$, $s10'$, $s11'$, $s12'$),
5′) if Profit $\geq$ 158.25, then $x_A \leq 2$ ($s1'$, $s3'$, $s4'$, $s5'$, $s6'$, $s8'$),
6′) if $x_A \geq 2$, then Time $\geq$ 130 ($s2'$, $s3'$, $s4'$, $s6'$, $s7'$, $s9'$),
7′) if $x_C \geq 7.17$, then $x_A \leq 2$ and $x_B \leq 2$ ($s1'$, $s3'$, $s5'$, $s6'$, $s8'$, $s10'$),
8′) if $x_C \geq 6$, then $x_A \leq 2$ and $x_B \leq 3$ ($s1'$, $s3'$, $s4'$, $s5'$, $s6'$, $s7'$, $s8'$, $s9'$, $s10'$, $s11'$, $s12'$),
9′) if $x_C \geq 7$, then Time $\geq$ 125 and $x_B \leq 2$ ($s1'$, $s3'$, $s5'$, $s6'$, $s8'$, $s10'$, $s11'$),
10′) if Sales $\geq$ 280, then Time $\geq$ 140 and $x_B \leq 3$ ($s1'$, $s2'$, $s3'$, $s4'$, $s5'$, $s7'$),
11′) if Sales $\geq$ 279.17, then Time $\geq$ 140 ($s1'$, $s2'$, $s3'$, $s4'$, $s5'$, $s6'$, $s7'$),
12′) if Sales $\geq$ 272, then Time $\geq$ 130 ($s1'$, $s2'$, $s3'$, $s4'$, $s5'$, $s6'$, $s7'$, $s8'$).

The DM has been asked again if (s)he was satisfied with one of the proposed Pareto optimal solutions. Since his/her answer was negative, (s)he was requested again to indicate a subset of relatively "good" solutions, which are indicated in the "Evaluation" column of Table 12.

Taking into account the ordinal classification of Pareto optimal solutions into "good" and "others", made by the DM, eight decision rules have been induced from the lower approximation of "good" solutions. The frequencies of the presence of objectives in the premises of the rules are the following:

- Profit: $\frac{2}{8}$,
- Time: $\frac{1}{8}$,
- Production of A: $\frac{5}{8}$,
- Production of B: $\frac{3}{8}$,
- Production of C: $\frac{3}{8}$,
- Sales: $\frac{2}{8}$.

The following potentially interesting decision rules were presented to the DM:

1) if Time $\leq$ 125 and $x_A \geq 1$, then product mix is good ($s11'$, $s12'$),
2) if $x_A \geq 1$ and $x_C \geq 7$, then product mix is good ($s3'$, $s6'$, $s11'$),
3) if $x_A \geq 1.5$ and $x_C \geq 6.46$, then product mix is good ($s3'$, $s4'$, $s6'$, $s12'$),
4) if Profit $\geq$ 158.25 and $x_A \geq 2$, then product mix is good ($s3'$, $s4'$, $s6'$),
5) if $x_A \geq 2$ and Sales $\geq$ 300, then product mix is good ($s3'$, $s4'$).

Among these decision rules, the DM has selected rule 4) as the most adequate to his/her preferences. This rule permits to define the following constraints reducing the Pareto optimal set of the production mix problem:

- $20x_A+30x_B+25x_C-(x_A+2x_B+0.75x_C)6-(0.5x_A+x_B+0.5x_C)8 \geq 158.25$ [Profit $\geq$ 158.25],
- $x_A \geq 2$ [Production of $A \geq 2$].

Let us observe that the first constraint is just strengthening an analogous constraint introduced in the first iteration (Profit $\geq$ 140.5).

Considering the new set of constraints, a new sample of representative Pareto optimal solutions shown in Table 13 has been calculated and presented to the DM, together with the following potentially interesting association rules:

Rough Sets in Decision Making, Table 12
A sample of Pareto optimal solutions proposed in the second iteration

Solution	Profit	Time	Prod. *A*	Prod. *B*	Prod. *C*	Sales	Evaluation
$s1'$	186.53	150	0.154	2	10	313.08	*
$s2'$	154.88	150	3	3	5.75	293.75	*
$s3'$	172	150	2	2	8	300	good
$s4'$	162.75	150	2	3	6.83	300.83	good
$s5'$	174	140	0	2	9.33	293.33	*
$s6'$	158.25	140	2	2	7.17	279.17	good
$s7'$	149	140	2	3	6	280	*
$s8'$	160.25	130	0	2	8.5	272	good
$s9'$	144.5	130	2	2	6.33	258.33	*
$s10'$	153.38	125	0	2	8.08	262.08	*
$s11'$	145.5	125	1	2	7	255	good
$s12'$	141.56	125	1.5	2	6.46	251.46	good

1″) if Time $\leq$ 145, then $x_A \leq 2$ and $x_B \leq 2.74$ and Sales ≤ 290.2 ($s2''$, $s3''$, $s4''$),
2″) if $x_C \geq 6.92$, then $x_A \leq 3$ and $x_B \leq 2$ and Sales ≤ 292.92 ($s3''$, $s4''$, $s5''$),
3″) if Time $\leq$ 145, then Profit ≤ 165.13 and $x_A \leq 2$ and $x_C \leq 7.58$ ($s2''$, $s3''$, $s4''$),
4″) if $x_C \geq 6.72$, then $x_B \leq 2.74$ ($s2''$, $s3''$, $s4''$, $s5''$),
5″) if Sales ≥ 289.58, then Profit ≤ 165.13 and Time $\geq$ 145 and $x_C \leq 7.58$ ($s1''$, $s2''$, $s3''$, $s5''$).

The DM has been asked again if (s)he was satisfied with one of the presented Pareto optimal solutions shown in Table 13 and this time (s)he declared that solution $s3''$ is satisfactory for him/her. This ends the interactive procedure.

Characteristics of the IMO-DRSA

The interactive procedure presented in Sect."Interactive Multiobjective Optimization Using DRSA (IMO-DRSA)" can be analyzed from the point of view of input and output information. As to the input, the DM gives preference information by answering easy questions related to ordinal classification of some representative solutions into two classes ("good" and "others"). Very often, in multiple criteria decision analysis, in general, and in interactive multiobjective optimization, in particular, the preference information has to be given in terms of preference model parameters, such as importance weights, substitution rates and various thresholds (see [6] for the Multiple Attribute Utility Theory and [1,5,34,46] for outranking methods; for some well-known interactive multiobjective optimization methods requiring preference model parameters, see the Geoffrion–Dyer–Feinberg method [9], the method of Zionts and Wallenius [60,61] and the Interactive Surrogate Worth Tradeoff method [2,3] requiring information in terms of marginal rates of substitution, the reference point method [58] requiring a reference point and weights to formulate an achievement scalarizing function, the Light Beam Search method [33] requiring information in terms of weights and indifference, preference and veto thresholds, being typical parameters of ELECTRE methods). Eliciting such information requires a significant cognitive effort on the part of the DM. It is generally acknowledged that people often prefer to make exemplary decisions and cannot always explain them in terms of specific parameters. For this reason, the idea of inferring preference models from exemplary decisions provided by the DM is very attractive. The output result of the analysis is the model of preferences in terms of "*if* …, *then* …" decision rules which is used to reduce the Pareto optimal set iteratively, until the DM selects a satisfactory solution. The decision rule preference model is very convenient for decision support, because it gives argumentation for preferences in a logical form, which is intelligible for the DM, and identifies the Pareto optimal solutions supporting each particular decision rule. This is very useful for a critical revision of the original ordinal classification of representative solutions into the two classes of "good" and "others". Indeed, decision rule preference model speaks the same language of the DM without any recourse to technical terms, like utility, tradeoffs, scalarizing functions and so on.

All this implies that IMO-DRSA has a transparent feedback organized in a learning oriented perspective, which permits to consider this procedure as a "glass box", contrary to the "black box" characteristic of many procedures giving final result without any clear explanation. Note that with the proposed procedure, the DM learns about the shape of the Pareto optimal set using the association rules. They represent relationships between attain-

Rough Sets in Decision Making, Table 13
A sample of Pareto optimal solutions proposed in the third iteration

Solution	Profit	Time	Prod. *A*	Prod. *B*	Prod. *C*	Sales	Evaluation
$s1''$	158.25	150	2	3.49	6.27	301.24	*
$s2''$	158.25	145	2	2.74	6.72	290.20	*
$s3''$	165.13	145	2	2	7.58	289.58	selected
$s4''$	158.25	140	2	2	7.17	279.17	*
$s5''$	164.13	150	3	2	6.92	292.93	*
$s6''$	158.25	145.3	3	2	6.56	284.02	*

able values of objective functions on the Pareto optimal set in logical and very natural statements. The information given by association rules is as intelligible as the decision rule preference model, since they speak the language of the DM and permit him/her to identify the Pareto optimal solutions supporting each particular association rule.

Thus, decision rules and association rules give an explanation and a justification of the final decision, that does not result from a mechanical application of a certain technical method, but rather from a mature conclusion of a decision process based on active intervention of the DM.

Observe, finally, that the decision rules representing preferences and the association rules describing the Pareto optimal set are based on ordinal properties of objective functions only. Differently from methods involving some scalarization (almost all existing interactive methods), in any step the proposed procedure does not aggregate the objectives into a single value, avoiding operations (such as averaging, weighted sum, different types of distance, achievement scalarization) which are always arbitrary to some extent. Observe that one could use a method based on a scalarization to generate the representative set of Pareto optimal solutions, nevertheless, the decision rule approach would continue to be based on ordinal properties of objective functions only, because the dialogue stage of the method operates on ordinal comparisons only. In the proposed method, the DM gets clear arguments for his/her decision in terms of "*if* . . . , *then* . . . " decision rules and the verification if a proposed solution satisfies these decision rules is particularly easy. This is not the case of interactive multiobjective optimization methods based on scalarization. For example, in the methods using an achievement scalarization function, it is not evident what does it mean for a solution to be "close" to the reference point. How to justify the choice of the weights used in the achievement function? What is their interpretation? Observe, instead, that the method proposed in this chapter operates on data using ordinal comparisons which would not be affected by any increasing monotonic transformation of scales, and this ensures the meaningfulness of results from the point of view of measurement theory (see e. g. [42]).

With respect to computational aspects of the method, notice that the decision rules can be calculated efficiently in few seconds only using the algorithms presented in [17,20]. When the number of objective functions is not too large to be effectively controlled by the DM (let us say seven plus or minus two, as suggested by Miller [37]), then the decision rules can be calculated in a fraction of one second. In any case, the computational effort grows exponentially with the number of objective functions, but not with respect to the number of considered Pareto optimal solutions, which can increase with no particularly negative consequence on calculation time.

Conclusions

Rough set theory is a mathematical tool for dealing with granularity of information and possible inconsistencies in the description of objects. Considering this description as an input data about a decision problem, the rough set approach permits to structure this description into lower and upper approximations, corresponding to certain and possible knowledge about the problem. Induction algorithms run on these approximations discover, in turn, certain and possible decision rules that facilitate an understanding of the DM's preferences, and that enable a recommendation concordant with these preferences.

The original version of the rough set approach, based on indiscernibility or similarity relation, and typical induction algorithms considered within machine learning, data mining and knowledge discovery, deal with data describing problems of taxonomy-type classification, i. e. problems where neither the attributes describing the objects, nor the classes to which the objects are assigned, are ordered. On the other hand, multiple criteria decision making deals with problems where descriptions (evaluations) of objects by means of attributes (criteria), as well as decisions in classification, choice and ranking problems,

are ordered. Moreover, in data describing multiple criteria decision making, there exist a monotonic relationship between conditions and decisions, like "the bigger the house, the more expensive it is". The generalization of the rough set approach and of the induction algorithms about problems in which order properties and monotonic relationships are important gave birth to the Dominance-based Rough Set Approach (DRSA) which made a breakthrough in scientific decision support.
The main features of DRSA are the following:

- preference information necessary to deal with any multiple criteria decision problem, or with decision under uncertainty, is asked to the DM just in terms of exemplary decisions,
- the rough set analysis of preference information supplies some useful elements of knowledge about the decision situation; these are: the relevance of attributes or criteria, the minimal subsets of attributes or criteria (reducts) conveying the relevant knowledge contained in the exemplary decisions, the set of indispensable attributes or criteria (core),
- DRSA can deal with preference information concerning taxonomy-type classification, ordinal classification, choice, ranking, multiobjective optimization and decision under uncertainty,
- the preference model induced from preference information structured by DRSA is expressed in a natural and comprehensible language of "*if* ..., *then* ..." decision rules,
- suitable procedures have been proposed to exploit the results of application of the decision rule preference model on a set of objects or pairs of objects in order to workout a recommendation,
- no prior discretization of quantitative condition attributes or criteria is necessary,
- heterogeneous information (qualitative and quantitative, ordered and non-ordered, nominal and ordinal, quantitative and numerical non-quantitative scales of preferences) can be processed within DRSA,
- the proposed methodology fulfills some desirable properties for both rough set approach (the approximated sets include lower approximation and are included in upper approximation, and the complementarity property is satisfied), and for multiple criteria decision analysis (the decision rule preference model is formally equivalent to the non-additive, non-transitive and non-complete conjoint measurement model, and to a more general model for preferences defined on all kinds of scales),
- the decision rule preference model resulting from DRSA is more general than all existing models of conjoint measurement, due to its capacity of handling inconsistent preferences (a new model of conjoint measurement is formally equivalent to the decision rule preference model handling inconsistencies),
- the decision rule preference model fulfills the postulate of transparency and interpretability of preference models in decision support; each decision rule can be clearly identified with those parts of the preference information (decision examples) which support the rule; the rules inform the DM in a quasi-natural language about the relationships between conditions and decisions; in this way, the rules permit traceability of the decision support process and give understandable justifications for the decision to be made,
- the proposed methodology is based on elementary concepts and mathematical tools (sets and set operations, binary relations), without recourse to any algebraic or analytical structures.

The decision rules entering the preference model have a special syntax which involves partial evaluation profiles and dominance relations on these profiles. The traditional preference models, which are the utility function and the outranking relation, can be represented in terms of equivalent decision rules. The clarity of the rule representation of preferences enables one to see the limits of these aggregation functions. Several studies [22,23,26,50] presented an axiomatic characterization of all three kinds of preference models in terms of conjoint measurement theory and in terms of a set of decision rules. The given axioms of "cancellation property" type are the weakest possible. In comparison to other studies on the characterization of aggregation functions, these axioms do not require any preliminary assumptions about the scales of criteria. A side-result of these investigations is that the decision rule preference model is the most general among the known aggregation functions.

Future Directions

The article shows that the Dominance-based Rough Set Approach is a very powerful tool for decision analysis and support. Its potential goes, however, beyond the theoretical frame considered in this article. There are many possibilities of applying DRSA to real life problems. The non-exhaustive list of potential applications includes:

Decision support in medicine: In this area there are already many interesting applications (see, e. g., [35,36,40,59]), however, they exploit the classical rough set approach; applications requiring DRSA, which handle

ordered value sets of medical signs, as well as monotonic relationship between the value of signs and the degree of gravity of a disease, are in progress;

Customer satisfaction survey: Theoretical foundations for application of DRSA in this field are available in [30], however, a fully documented application is still missing;

Bankruptcy risk evaluation: This is a field of many potential applications, as can be seen from promising results reported, e. g. in [12,53,54], however, a wider comparative study involving real data sets is needed;

Operational research problems, such as location, routing, scheduling or inventory management: These are problems formulated either in terms of classification of feasible solutions (see, e. g., [11]), or in terms of interactive multiobjective optimization, for which the IMO-DRSA [32] procedure is suitable;

Finance: This is a domain where DRSA for decision under uncertainty has to be combined with interactive multiobjective optimization using IMO-DRSA; some promising results going in this direction have been presented in [31];

Ecology: Assessment of the impact of human activity on the ecosystem is a challenging problem for which the presented methodology is suitable; the up to date applications are based on the classical rough set concept (see, e. g., [7,43]), however, it seems that DRSA handling ordinal data has a greater potential in this field.

Bibliography

1. Brans JP, Mareschal B (2005) PROMETHEE Methods. In: Figueira J, Greco S, Ehrgott M (eds) Multiple criteria decision analysis: State of the art surveys. Springer, Berlin, pp 163–195
2. Chankong V, Haimes YY (1978) The interactive surrogate worth trade-off (ISWT) method for multiobjective decision-making. In: Zionts S (ed) Multiple Criteria Problem Solving. Springer, Berlin, pp 42–67
3. Chankong V, Haimes YY (1983) Multiobjective decision making theory and methodology. Elsiever Science, New York
4. Figueira J, Greco S, Ehrgott M (eds) (2005) Multiple criteria decision analysis: State of the art surveys. Springer, Berlin
5. Figueira J, Mousseau V, Roy B (2005) ELECTRE methods. In: Figueira J, Greco S, Ehrgott M (eds) Multiple criteria decision analysis: State of the art surveys. Springer, Berlin, pp 133–162
6. Fishburn PC (1967) Methods of estimating additive utilities. Manag Sci 13(7):435–453
7. Flinkman M, Michalowski W, Nilsson S, Słowiński R, Susmaga R, Wilk S (2000) Use of rough sets analysis to classify Siberian forest ecosystem according to net primary production of phytomass. INFOR 38:145–161
8. Fortemps P, Greco S, Słowiński R (2008) Multicriteria decision support using rules that represent rough-graded preference relations. Eur J Oper Res 188:206–223
9. Geoffrion A, Dyer J, Feinberg A (1972) An interactive approach for multi-criterion optimization, with an application to the operation of an academic department. Manag Sci 19(4):357–368
10. Giove S, Greco S, Matarazzo B, Słowiński R (2002) Variable consistency monotonic decision trees. In: Alpigini JJ, Peters JF, Skowron A, Zhong N (eds) Rough sets and current trends in computing. LNAI, vol 2475. Springer, Berlin, pp 247–254
11. Gorsevski PV, Jankowski P (2008) Discerning landslide susceptibility using rough sets. Comput Environ Urban Syst 32:5365
12. Greco S, Matarazzo B, Słowiński R (1998) A new rough set approach to evaluation of bankruptcy risk. In: Zopounidis C (ed) Operational tools in the management of financial risks. Kluwer, Dordrecht, pp 121–136
14. Greco S, Matarazzo B, Słowiński R (1999) The use of rough sets and fuzzy sets in MCDM. In: Gal T, Stewart T, Hanne T (eds) Advances in multiple criteria decision making. Kluwer, Boston, pp 14.1–14.59
15. Greco S, Matarazzo B, Słowiński R (1999) Rough approximation of a preference relation by dominance relations. Eur J Oper Res 117:63–83
16. Greco S, Matarazzo B, Słowiński R (2000) Extension of the rough set approach to multicriteria decision support. INFOR 38:161–196
13. Greco S, Matarazzo B, Słowiński R, Tsoukias A (1998) Exploitation of a rough approximation of the outranking relation in multicriteria choice and ranking. In: Stewart TJ, van den Honert RC (eds) Trends in multicriteria decision making. LNEMS, vol 465. Springer, Berlin, pp 45–60
17. Greco S, Matarazzo B, Słowiński R, Stefanowski J (2001) An algorithm for induction of decision rules consistent with dominance principle. In: Ziarko W, Yao Y (eds) Rough sets and current trends in computing. LNAI, vol 2005. Springer, Berlin, pp 304–313
18. Greco S, Matarazzo B, Słowiński R (2001) Rough sets theory for multicriteria decision analysis. Eur J Oper Res 129:1–47
19. Greco S, Matarazzo B, Słowiński R (2001) Rough set approach to decisions under risk. In: Ziarko W, Yao Y (eds) Rough sets and current trends in computing. LNAI, vol 2005. Springer, Berlin, pp 160–169
20. Greco S, Matarazzo B, Słowiński R, Stefanowski J (2002) Mining association rules in preference-ordered data. In: Hacid M-S, Ras ZW, Zighed DA, Kodratoff Y (eds) Foundations of intelligent systems. LNAI, vol 2366. Springer, Berlin, pp 442–450
21. Greco S, Matarazzo B, Slowinski R (2002) Multicriteria classification. In: Kloesgen W, Zytkow J (eds) Handbook of data mining and knowledge discovery. Oxford University Press, New York, pp 318–328
22. Greco S, Matarazzo B, Słowiński R (2002) Preference representation by means of conjoint measurement & decision rule model. In: Bouyssou D, Jacquet-Lagrèze E, Perny P, Słowiński R, Vanderpooten D, Vincke P (eds) Aiding decisions with multiple criteria – essays in honor of Bernard Roy. Kluwer, Dordrecht, pp 263–313
23. Greco S, Matarazzo B, Słowiński R (2004) Axiomatic characterization of a general utility function and its particular cases in terms of conjoint measurement and rough-set decision rules. Eur J Oper Res 158:271–292
24. Greco S, Matarazzo B, Słowiński R (2004) Dominance-based Rough Set Approach to Knowledge Discovery, (I) – General Perspective, (II) – Extensions and Applications. In: Zhong N,

Liu J (eds) Intelligent technologies for information analysis. Springer, Berlin, pp 513–612
25. Greco S, Pawlak Z, Słowiński R (2004) Can Bayesian confirmation measures be useful for rough set decision rules? Eng Appl Artif Intell 17:345–361
26. Greco S, Matarazzo B, Słowiński R (2005) Decision rule approach. In: Figueira J, Greco S, Ehrgott M (eds) Multiple criteria decision analysis: State of the art surveys. Springer, Berlin, pp 507–563
27. Greco S, Matarazzo B, Słowiński R (2006) Dominance-based rough set approach to decision involving multiple decision makers. In: Greco S, Hata Y, Hirano S, Inuiguchi M, Miyamoto S, Nguyen HS, Slowinski R (eds) Rough Sets and Current Trends in Computing, RSCTC. LNAI, vol 4259. Springer, Berlin, pp 306–317
28. Greco S, Matarazzo B, Słowiński R (2006) Dominance-based rough set approach to case-based reasoning. In: Torra V, Narukawa Y, Valls A, Domingo-Ferrer J (eds) Modelling decisions for artificial intelligence. LNAI, vol 3885. Springer, Berlin, pp 7–18
29. Greco S, Matarazzo B, Słowiński R (2007) Dominance-based rough set approach as a proper way of handling graduality in rough set theory. In: Transactions on Rough Sets VII. LNCS, vol 4400. Springer, Berlin, pp 36–52
30. Greco S, Matarazzo B, Słowiński R (2007) Customer satisfaction analysis based on rough set approach. Z Betr 16(3):325–339
31. Greco S, Matarazzo B, Słowiński R (2007) Financial portfolio decision analysis using dominance-based rough set approach. Invited paper at the 22nd European Conference on Operational Research (EURO XXII), Prague, 08–11 July
32. Greco S, Matarazzo B, Słowiński R (2008) Dominance-based rough set approach to interactive multiobjective optimization. In: Branke J, Deb K, Miettinen K, Slowinski R (eds) Multiobjective optimization: Interactive and evolutionary approaches. Springer, Berlin
33. Jaszkiewicz A, Słowiński R (1999) The "Light Beam Search" approach – an overview of methodology and applications. Eur J Oper Res 113:300–314
34. Martel JM, Matarazzo B (2005) Other outranking approaches. In: Figueira J, Greco S, Ehrgott M (eds) Multiple criteria decision analysis: State of the art surveys. Springer, Berlin, pp 197–262
35. Michalowski W, Rubin S, Słowiński R, Wilk S (2003) Mobile clinical support system for pediatric emergencies. J Decis Support Syst 36:161–176
36. Michalowski W, Wilk S, Farion K, Pike J, Rubin S, Słowiński R (2005) Development of a decision algorithm to support emergency triage of scrotal pain and its implementation in the MET system. INFOR 43:287–301
37. Miller GA (1956) The magical number seven, plus or minus two: some limits in our capacity for processing information. Psychol Rev 63:81–97
38. Pawlak Z (1982) Rough Sets. Int J Comput Inf Sci 11:341–356
39. Pawlak Z (1991) Rough Sets. Kluwer, Dordrecht
40. Pawlak Z, Słowiński K, Słowiński R (1986) Rough classification of patients after highly selective vagotomy for duodenal ulcer. Int J Man-Machine Stud 24:413–433
41. Pawlak Z, Słowiński R (1994) Rough set approach to multi-attribute decision analysis. Eur J Oper Res 72:443–459
42. Roberts F (1979) Measurement theory, with applications to decision making, utility and the social sciences. Addison-Wesley, Boston
43. Rossi L, Słowiński R, Susmaga R (1999) Rough set approach to evaluation of stormwater pollution. Int J Environ Pollut 12:232–250
44. Roy B (1996) Multicriteria Methodology for Decision Aiding. Kluwer, Dordrecht
45. Roy B (1999) Decision-aiding today: What should we expect. In: Gal T, Stewart T, Hanne T (eds) Advances in Multiple Criteria Decision Making. Kluwer, Boston, pp 1.1–1.35
46. Roy B, Bouyssou D (1993) Aide Multicritère à la Décision: Méthodes et Cas. Economica, Paris
47. Shoemaker PJH (1982) The expected utility model: its variants, purposes, evidence and limitations. J Econ Lit 20:529–562
48. Słowiński R (1993) Rough set learning of preferential attitude in multi-criteria decision making. In: Komorowski J, Ras ZW (eds) Methodologies for intelligent systems. LNAI, vol 689. Springer, Berlin, pp 642–651
49. Słowiński R, Greco S, Matarazzo B (2002) Rough set analysis of preference-ordered data. In: Alpigini JJ, Peters JF, Skowron A, Zhong N (eds) Rough sets and current trends in computing. LNAI, vol 2475. Springer, Berlin, pp 44–59
50. Słowiński R, Greco S, Matarazzo B (2002) Axiomatization of utility, outranking and decision-rule preference models for multiple-criteria classification problems under partial inconsistency with the dominance principle. Control Cybern 31:1005–1035
51. Słowiński R, Greco S, Matarazzo B (2005) Rough set based decision support, chapter 16. In: Burke EK, Kendall G (eds) Search methodologies: Introductory tutorials in optimization and decision support techniques. Springer, New York, pp 475–527
52. Słowiński R, Greco S, Matarazzo B (2007) Dominance-based rough set approach to reasoning about ordinal data, keynote lecture. In: Kryszkiewicz M, Peters JF, Rybiński H, Skowron A (eds) Rough sets and intelligent systems paradigms. LNAI, vol 4585. Springer, Berlin, pp 5–11
53. Słowiński R, Zopounidis C (1995) Application of the rough set approach to evaluation of bankruptcy risk. Int J Intell Syst Acc Finance Manag 4:27–41
54. Słowiński R, Zopounidis C, Dimitras AI (1997) Prediction of company acquisition in Greece by means of the rough set approach. Eur J Oper Res 100:1–15
55. Stefanowski J (1998) On rough set based approaches to induction of decision rules. In: Polkowski L, Skowron A (eds) Rough sets in data mining and knowledge discovery, vol 1. Physica, Heidelberg, pp 500–529
56. Tsoukias A, Vincke P (1995) A new axiomatic foundation of the partial comparability theory. Theory Decis 39:79–114
57. Steuer RE, Choo E-U (1983) An interactive weighted Tchebycheff procedure for multiple objective programming. Math Program 26:326–344
58. Wierzbicki AP (1980) The use of reference objectives in multiobjective optimization. In: Fandel G, Gal T (eds) Multiple criteria decision making, theory and applications. Springer, Berlin, pp 468–486
59. Wilk S, Słowiński R, Michalowski W, Greco S (2005) Supporting triage of children with abdominal pain in the emergency room. Eur J Oper Res 160:696–709
60. Zionts S, Wallenius J (1976) An interactive programming method for solving the multiple criteria problem. Manag Sci 22:652–663
61. Zionts S, Wallenius J (1983) An interactive multiple objective linear programming method for a class of underlying nonlinear utility functions. Manag Sci 29:519–523

Rough Sets: Foundations and Perspectives

James F. Peters[1], Andrzej Skowron[2], Jarosław Stepaniuk[3]
[1] Computational Intelligence Laboratory, University of Manitoba, Winnipeg, Canada
[2] Institute of Mathematics, Warsaw University, Warsaw, Poland
[3] Computer Science, Białystok University of Technology, Białystok, Poland

Article Outline

Glossary
Definition of the Subject
Introduction
Approximation Spaces
Rough Sets
Dimensionality Reduction
Summary
Future Directions
Acknowledgments
Bibliography

Glossary

Approximation The replacement of mathematical objects by others that resemble them in certain respects [64].

Approximation space An approximation space is denoted by $(\mathcal{O}, \mathcal{F}, \sim_B)$, where $\mathcal{O}$ is a set of perceived objects, $\mathcal{F}$ is a set of probe functions representing object features, and $\sim_B$ is an indiscernibility relation defined relative to $B \subseteq \mathcal{F}$. This approximation space is considered fundamental because it provided a framework for the original rough set theory [37,40]. Several generalizations of this definition of approximation space have been proposed (see, e. g., [40,44,54,55,56,58,59,69]).

Attribute A quality regarded as characteristic or inherent in an object [29]. In rough set theory, an attribute a of an object x is represented by a partial function $f_a(x) = v$, where v is a value in the range of f_a. In rough set theory, the function f_a is often called an attribute [38,45].

Boundary region The B-boundary region of an approximation of a set X is denoted by $\mathrm{Bnd}_B X$ and is defined relative to a set of functions B representing features of objects in X as well as the lower approximation $B_* X$ and the upper approximation $B^* X$, where

$$\mathrm{Bnd}_B X = B^* X \backslash B_* X = \{x \mid x \in B^* X \text{ and } x \notin B_* X\}.$$

Elementary set A B-class in the quotient set $X/\sim_B$.

Equivalence class Given a relation $\sim$, an equivalence class is a set denoted by $[x]$ or $[x]_\sim$ [10] in the quotient set $X/\sim$ (See Glossary item "Quotient set"), where

$$[x] = \{x' \in X \mid x \sim x'\}\,.$$

Equivalence relation A reflexive, symmetric and transitive relation $\sim \subseteq X \times X$. An equivalence relation $\sim$ on a set X defines a partition of X into classes.

Feature Make, form, fashion, shape (of an object) [29]. A characteristic of an object perceived by the senses or knowable by the mind [41,52]. In rough set theory, a feature f of an object x is represented by a function $\phi_f(x) = v$, where v is a value in the range of ϕ_f (e. g., $\phi_{\text{♪}}(x)$ as a measure of the tonality ♪ feature of a Chopin Mazurka x) [41,52]. The function ϕ_f is sometimes also termed an attribute [38,45].

Indiscernibility relation An equivalence relation

$$\sim_B = \{(x, x') \in X \times X | f(x) = f(x') \text{ for any } f \in B\}\,,$$

where X denotes a set of objects, B denotes a set of functions, and $f \in B$ is a function representing a feature of an object $x \in X$. The notation used to denote an equivalence relation in rough set theory has varied widely over time. For example, $\widetilde{B}$ was originally introduced by Zdzisław Pawlak in 1981 [37]. Later, $\mathrm{Ind}(B)$ [18,30,38,66] or IND_B [45] or Ind_B [14] or IND [66] or $I(B)$ [40] or $=_B$ [16] has also been used to denote an equivalence relation on a set X defined relative to attributes of objects. In rough set theory, the equivalence relation $\sim_B$ was introduced by Zdzisław Pawlak [37].

Information granule Information granules are obtained in the process of granulation. Granulation can be viewed as a human way of achieving data compression and it plays a key role in implementing the divide-and-conquer strategy in human problem-solving. An *information granule* represents a set of objects that have descriptions matching the granule [52], e. g. elementary set $[x]_B$, lower approximation $B_* X$, quotient set $X/\sim_B$.

Lower approximation The B-lower approximation of a set X is denoted by $B_* X$ and is defined relative to a set of functions B representing features of objects in X and the quotient set $X/\sim_B$, where

$$B_* X = \bigcup_{x:[x]_B \subseteq X} [x]_B\,.$$

Object Something perceptible to the senses or knowable by the mind [29].

Information Whatever is conveyed or represented by a particular sequence of symbols [29]. In rough set theory, information is derivable either from the patterns in a particular information table or from what can be observed in a particular approximation space [37,40].

Information system A system to represent knowledge [25,36,40]. Syntactic representation of knowledge in table form [25,45].

Partition of a non-empty set X A family of non-empty, pairwise disjoint subsets of X (called classes) such that the union of this family is equal to X.

Quotient set Set of all classes in a partition defined by an equivalence relation $\sim$ on a set X (denoted by $X/\sim$).

Rough set A set X is considered a rough set if, and only if it has a non-empty B-boundary $\mathrm{Bnd}_B X$, i. e., the B-approximation of X has a non-empty boundary.

Upper approximation The B-upper approximation of X is denoted by B^*X and is defined relative to a set of functions B representing features of objects in X and the quotient set $X/\sim_B$, where

$$B^*X = \bigcup_{x:[x]_B \cap X \neq \emptyset} [x]_B \ .$$

Definition of the Subject

Rough set theory was introduced by Zdzisław Pawlak (1926–2006) during the early 1980s. This theory has two distinguishing hallmarks: knowledge description systems introduced by Pawlak in 1973 [36] and set approximation (See Glossary Items "Approximation", "Boundary region", "Lower approximation", "Upper approximation") as a means of classifying a set [37]. Knowledge description systems represent our knowledge about sample objects in tabular form. This knowledge results from identifying a set of attributes A (apparent qualities) of the objects. A total function $\rho : X \times A \longrightarrow V_a$ maps an attribute $a \in A$ of an object $x \in X$ to a value $v \in V_a$. Our knowledge about objects $x_1, \ldots, x_n$ having attributes $a_1, \ldots, a_k$ then can be represented in table form as rows of knowledge descriptions, e. g., description of x_i as a tuple.

$$(\rho(x_i, a_1), \rho(x_i, a_2), \ldots, \rho(x_i, a_k)) \ .$$

At this point, there is a natural transition to the formulation of an indiscernibility relation $\sim_B$ that defines a partition of a set of sample objects relative to set $B \subseteq A$ (see Glossary item "Indiscernibility relation"). In many ways, this relation is both fecund and important. The fecundity (high fertility) of this relation can be seen in several ways. First, $\sim_B$ makes it possible to organize tabular representations of objects into equivalence classes. This, of course, means that now our observations can be made at the level of classes (elementary granules) instead of the more intransigent level of individual objects. Second, the majesty of $\sim_B$ can be seen in Pawlak's discovery of lower and upper descriptions of objects. These descriptions accrue naturally from lower and upper approximations of a set of objects. Third, it is then possible gauge the *roughness* of our knowledge about a set of objects by considering the size of the boundary of an approximation. The roughness of our knowledge about a set of sample objects is directly proportional to the size of the approximation boundary. So, in effect, the indiscernibility relation lead to the discovery of rough sets.

The importance of the indiscernibility relation can be seen in a number of ways, if one considers the ordinary difficulties associated with the complexity of ordinary, perceptual objects. The introduction of this relation led to the creation of an approximation space that provides a framework for perception or observation on the level of classes [33,37]. An important byproduct of the approach to approximation in rough set theory is information granulation [52]. Each selection of features of a set of sample objects leads to granulation of the information associated with the objects. For a set of sample objects X with selected features represented by a set of functions B, a rich harvest of information granules from the partition of X defined by the indiscernibility relation $\sim_B$, e. g., individual classes $[x]_B$, quotient space $X/\sim_B$, lower approximation B_*X, upper approximation B^*X, and boundary $\mathrm{Bnd}_B X$. From the introduction of knowledge description systems and approximation of our knowledge springs various forms of learning, starting with learning from examples exemplified by a number rough set toolsets released during the past decade (see, e. g., [20,50]).

Basic ideas of rough set theory and its extensions, as well as many interesting applications can be found in numerous books [50], issues of the Transactions on Rough Sets [62], special issues of other journals, proceedings of international conferences, tutorials or surveys (see, references and further readings listed in this paper and, e. g., [40] and numerous web pages such as [20,63,65]).

Introduction

This chapter gives a concise overview of some of the features of the foundations and perspectives of rough set theory. The foundations of rough set theory are rich and varied. These foundations include the study of algebras and algebraic structures associated with rough sets [2,8], entropy [7,61], lattice theory [9,23], similarity and indiscernibility [34], rough-satisfiability [14], logic [11,45],

and rough mereology [46]. In general, an overview of the mathematical foundations of rough sets is given by Lech Polkowski [45,46].

In addition to providing a fairly comprehensive inventory of some of the highlights of rough set-based research, this chapter gives a capsule view of several of the hallmarks of rough set theory, namely, approximation, approximation spaces, discovery of rough sets, and feature set reduction (reducts). For many other issues related to rough set theory and its applications, the reader is referred to the bibliography on rough sets.

This chapter has the following organization. The rough set approach to approximation spaces is given in Sect. "Approximation Spaces". This is followed in Sect. "Rough Sets" by an overview of the general notion of rough sets. The rough set approach to dimensionality reduction is briefly presented in Sect. "Dimensionality Reduction". A capsule view of the future of rough sets and their applications is given in Sect. "Future Directions".

Approximation Spaces

In rough set theory, approximations are carried out within the context of an approximation space $(\mathcal{O}, \mathcal{F}, \sim_B)$, where $\mathcal{O}$ is a set of objects, $\mathcal{F}$ is a set of functions representing object features, and $\sim_B$ is an indiscernibility relation defined relative to $B \subseteq \mathcal{F}$. This space is considered fundamental because it provided a framework for the original rough set theory [37]. It has also been observed that an approximation space is the formal counterpart of perception [33]. Approximation starts with the partition $\mathcal{O}/\sim_B$ of $\mathcal{O}$ defined by the relation $\sim_B$. Next, any set $X \subseteq \mathcal{O}$ is approximated by considering the relation between X and the classes $[x]_B \in \mathcal{O}/\sim_B, x \in \mathcal{O}$.

An approximation space $(\mathcal{O}, \mathcal{F}, \sim_B)$ defined relative to a set of perceptual objects $\mathcal{O}$, a set functions $\mathcal{F}$, and relation $\sim_B$, has the following basic framework that facilitates observations concerning sample objects.

Framework for an Approximation Space

$$\begin{aligned}
\mathcal{O} &= \text{Set of perceived objects,}\\
\mathcal{F} &= \text{Set of probe functions objects,}\\
B &\subseteq \mathcal{F},\\
\sim_B &= \{(x, x') \in X \times X \mid f(x) = f(x') \\
&\qquad \text{for any } f \in B\},\\
\mathcal{O}/\sim_B &= \{[x]_B \mid x \in \mathcal{O}, B \subseteq \mathcal{F}\}, B\text{-partition of } \mathcal{O},\\
[x]_B &\in \mathcal{O}/\sim_B,\\
X &\subseteq \mathcal{O}, \text{ set of objects of interest,}\\
B_*X &= \bigcup_{x:[x]_B \subseteq X} [x]_B, B\text{-lower approximation,}\\
B^*X &= \bigcup_{x:[x]_B \cap X \neq \emptyset} [x]_B, B\text{-upper approximation,}\\
\mathrm{Bnd}_B X &= B^*X \setminus B_*X, B\text{-boundary region.}
\end{aligned}$$

Affinities between objects of interest in the set $X \subseteq \mathcal{O}$ and classes in the quotient set $\mathcal{O}/\sim_B$ can be discovered by identifying those classes that have objects in common with X. Approximation of the set X begins by determining which elementary sets $[x]_B \in \mathcal{O}/\sim_B$ are subsets of X. This discovery process leads to the B-lower approximation of $X \subseteq \mathcal{O}$ denoted by B_*X. The B-upper approximation B^*X is a sum of equivalence classes $[x]_B \in \mathcal{O}/\sim_B$, where each class included in B^*X contains at least one object with a description that matches the description of an object in X. The lower and upper approximations of X provide a basis for defining the boundary of an approximation. Notice that B_*X is always a subset of B^*X.

Several generalizations of the classical rough set approach based on approximation spaces defined as pairs of the form $(\mathcal{O}, \mathcal{F}, \sim_B)$, $B \subseteq F$, have been reported in the literature [44,45,54]. Among them it is worthwhile mentioning the rough set approach based on similarity (tolerance) relations and the approach to approximation of vague concepts based on the adaptive extension of approximation spaces [54] and approximations spaces that consider the nearness of objects [44].

Moreover, the approach based on inclusion functions has been generalized to the *rough mereological approach* [32,45,46]. The inclusion relation $x\mu_r y$ with the intended meaning *x is a part of y to a degree at least r* has been taken as the basic notion of the rough mereology being a generalization of the Leśniewski mereology [26]. Research on rough mereology has shown importance of another notion, namely *closeness* of compound objects (e. g., concepts). This can be defined by $x\,\mathrm{cl}_{r,r'}\,y$ if and only if $x\mu_r y$ and $y\mu_{r'}x$.

Rough mereology offers a methodology for synthesis and analysis of objects in a distributed environment of intelligent agents, in particular, for synthesis of objects satisfying a given specification to a satisfactory degree or for control in such a complex environment. Moreover, rough mereology has been used for developing the foundations of the *information granule calculi*, aiming at formalization of the Computing with Words paradigm, recently formulated by Lotfi Zadeh [67,68]. More complex information granules are defined recursively using already defined information granules and their measures of inclusion and closeness. Information granules can have complex structures like classifiers or approximation spaces. Com-

putations on information granules lead to the discovery of relevant information granules, e. g., information granules useful in compound concept approximations, pattern recognition and in machine learning. For example, families of approximation spaces labeled by some parameters are considered. By tuning such parameters according to chosen criteria (e. g., minimal description length), one can search for an optimal approximation space.

Rough Sets

In its original conception by Zdzisław Pawlak during the early 1980s, the discovery of the rough set approach to classification of sets of objects resulted from work on knowledge description systems begun by Pawlak during the early 1970s. In the rough set approach, the roughness of our knowledge of a set of objects is gauged in terms of a new approach to approximation that pivots on object descriptions commonly found in information tables also introduced by Pawlak. Approximation is carried out at the level of sets relative to our knowledge about objects of interest. That is, one approximates one set by another set. This approximation is built on a feature space for individual objects. A feature space for objects is defined relative to functions that map sample objects (pre-images) to measurements (images) that gauge the extent of our knowledge about the appearance of the objects. The basic idea is to select a set X of objects of interest and approximate it with another set that resembles X. This means that approximation is carried out at two different levels in the rough set approach.

1. **Object Level**. For a given set of objects design a set of functions $\phi = \{\phi_1, \ldots, \phi_n\}$ that represent features of sample objects of interest. The description of an object x is approximated by an n-tuple $(\phi_1(x), \ldots, \phi_n(x))$ of feature measurements.

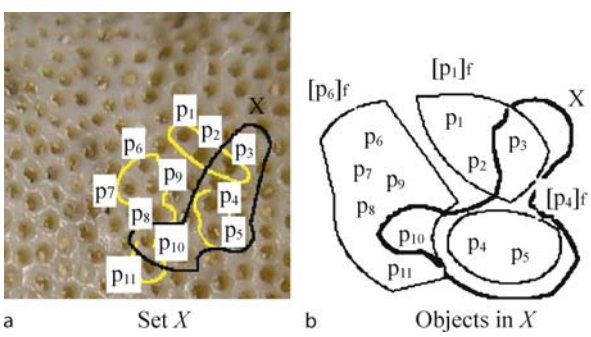

Rough Sets: Foundations and Perspectives, Figure 1
Sample $X \subseteq \mathcal{O}$ to be approximated

2. **Set Level**. Define the partition of objects in the sample space $\mathcal{O}$ using the indiscernibility relation $\sim_B$, where $B \subseteq \phi$. Approximate a set X containing objects of interest relative to the elementary sets in $X/\sim_B$. Approximation of a set X containing objects with features represented by functions ϕ is carried out in terms of the lower approximation B_*X and the upper approximation B^*X.

Example 1 (Sample approximation of a set of coral pores) For this example, consider the set of coral pores in the coral fragment shown in Fig. 1(b). Then sets $\mathcal{O}, \mathcal{X}, \mathcal{B}$ and an approximation of X are defined in the following way:

$$\begin{aligned}
\mathcal{O} &= \{p \mid p \text{ equals a labeled coral pore in Fig. 1}\} \\
&= \{p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8, p_9, p_{10}, p_{11}\}, \\
\mathcal{F} &= \{f\}, \text{ where } f : \mathcal{O} \to \Re, \\
f(p_i) &= \text{ depth (in mm) of a coral pore } p_i \in X, \\
\mathcal{O}/\sim_B &= \{[p_1]_f, [p_4]_f, [p_6]_f\} \text{ shown in Fig. 1a}, \\
X &= \{p \mid p \text{ equals a labeled coral pore in Fig. 1b}\}, \\
&= \{p_3, p_4, p_5, p_{10}\}.
\end{aligned}$$

Hence

$$\begin{aligned}
B_*X &= [p_4]_f = \{p_4, p_5\}\,, \\
B^*X &= [p_1]_f \cup [p_4]_f \cup [p_6]_f\,, \\
\mathrm{Bnd}_B X &= [p_1]_f \cup [p_6]_f\,.
\end{aligned}$$

Then observe that only $[p_4]_f$ is a proper subset of X. In effect, only the class $[p_4]_f$ has a complete affinity with X, since all of the objects $[p_4]_f$ have descriptions that match the descriptions of objects in X. By contrast, classes $[p_1]_f, [p_6]_f$ have only a partial affinity with X, since there are objects in each class that have descriptions that do match the description of any object in X. Observe that the boundary $\mathrm{Bnd}_B X$ is not empty. Hence, X is an example of a rough set.

Information Tables

For computational reasons, a syntactic representation of information systems is usually given in the form of tables. An information system is represented by the pair $(\mathcal{O}, \mathcal{F})$, where $\mathcal{O}$ denotes a set of objects (a universe) and $\mathcal{F}$ denotes a set of functions representing features of objects in $\mathcal{O}$. Let $B \subseteq \mathcal{F}$. Discovering objects in the composition of a class $[x]_B\,, x \in \mathcal{O}$ in the partition $\mathcal{O}/\sim_B$ in the system $(\mathcal{O}, \mathcal{F})$ is accomplished by gathering together inside the class all of those objects that have matching function values. Identifying the classes in $X/\sim_B$ is greatly aided by

a table representation of $(\mathcal{O}, \mathcal{F})$. This is illustrated with the following example.

Example 2 (Sample information table) For this example, assume again that $\mathcal{O}$ is a set of words. For simplicity, we limit $\mathcal{O}$ to words from poems by John Keats (1795–1821). Consider, for example, the words in the following lines in Part 3 of Keats' ode *To Autumn* [24].

TO AUTUMN

3.

Where are the songs of spring? Aye, where are they?

Think not of them, thou has thy music too,—

...

John Keats

19 September 1819

By way of approximation of a set of conceptual objects, consider $X \subseteq \mathcal{O}$ defined as

$$\begin{aligned}\mathcal{O} &= \{x \mid \text{ x is a word in Keats' poetry}\},\\ &= \{Where, are, the, songs, of, spring, \ldots\},\\ \mathcal{A} &= \{f_1, f_2\},\end{aligned}$$

where

$$f_1(x) = \begin{cases} 1, & x \text{ begins with a vowel,} \\ 0, & otherwise. \end{cases}$$

$$f_2(x) = \begin{cases} 1, & x \text{ contains two or more vowels,} \\ 0, & otherwise. \end{cases}$$

The objects in this example are conceptualized as poetic words and the functions appropriately represent attributes of poetic words. On might say that a word gains in poesy by beginning with a vowel sound (taking away the leading vowel sound tends to undermine the poetic character of a word). Similarly, a repetition of vowel sounds in a word is characteristic of poetic words, especially in the poetry of John Keats. For simplicity, consider only the vowels *a, e, i, o, u*, the vowel sound *ha* in the word *has* and omit consideration of other vowel sounds. In addition, $\mathcal{O}$ is restricted to words in Keats' ode *To Autumn*, partially represented in this Example. The function values for each of the words in the fragment from Keat's ode are summarized in Table 1. Let $B = \mathcal{A}$. This leads to the following partition of $\mathcal{O}$:

$$\begin{aligned}[where]_B &= \{where, thou, music, too\}\\ [are]_B &= \{are, Aye\}\\ [the]_B &= \{the, songs, spring, they, think, not,\\ &\qquad them, has, thy\}\\ [of]_B &= \{of\}\end{aligned}$$

Rough Sets: Foundations and Perspectives, Table 1
Information System for Keats' Words

$\mathcal{O}$	f_1	f_2	d
where	0	1	1
are	1	1	1
the	0	0	0
songs	0	0	1
of	1	0	1
spring	0	0	1
Aye	1	1	1
they	0	0	1
think	0	0	1
not	0	0	1
them	0	0	1
thou	0	1	1
has	0	0	0
thy	0	0	1
music	0	1	1
too	0	1	1

Now select, for example,

$$X = \{not, are, Aye, has, of, too\}\ .$$

This choice of X leads to

$$\begin{aligned}B_* X &= [are]_B \cup [of]_B = \{are, Aye, of\}\\ B^* X &= [are]_B \cup [of]_B \cup [the]_B\ .\end{aligned}$$

In effect, the words in $[are]_B \cup [of]_B$ have an affinity to the words in the set X. In particular, notice that each of the words in the class $[where]_B$

$$[where]_B = \{where, thou, music, too\}\ ,$$

have exactly the same f_1, f_2 function values. Similarly, for the composition of the remaining three classes in the partition $X/\sim_B$. Again, for example, notice that the words in class $[the]_A$ have the same function values, namely, $f_1(x) = f(x) = 0$, where x is a word in $[the]_B$. It is a fairly straightforward task to identify the classes extractable from small tables. For large tables, it is necessary to mechanize the class extraction process.

Decision Systems and Decision Rules

Of particular interest is the extension of information systems made possible by including a function d representing what is known as a decision attribute in rough set theory. A decision system is represented by the triple $(\mathcal{O}, \mathcal{F}, d)$, where $\mathcal{O}$ denotes a set of objects (a universe),

$\mathcal{F}$ denotes a set of functions representing features of objects in $\mathcal{O}$, and $d : \mathcal{O} \longrightarrow V_d$, where V_d is a set of values representing decisions. Presenting decision procedures in a tabular form goes back at least to ancient Babylon. Tabular forms for computer programming dates back to the late 1950s. Next, tabular forms became popular in databases (see http://www.catalyst.com/products/logicgem/overview.html). It is clear from the common meaning of the verb *decide* that a decision represents a resolution or determination about something and presumes something knowable by the mind but not necessarily observable by the senses. In more general setting, one can consider a set of decisions D instead of a single decision d.

Example 3 (Sample decision system) For example, consider an extension of the information system $(\mathcal{O}, \mathcal{A})$ with a decision d for Keats' words in Example 3. Recall that $\mathcal{O}$ denotes a set of words and $\mathcal{F}$ denotes a set of features representing attributes of words in $\mathcal{O}$. By way of illustration, let d be defined as follows.

$$d(x) = \begin{cases} 1, & x \in \mathcal{O} \mid x \text{ is a part of an alliteration}, \\ 0, & \textit{otherwise}, \end{cases}$$

where *alliteration* means the occurrence of the same letter or sound at the beginning of adjacent or closely connected words. *Alliteration* is also an example of a concept (i. e., an idea for a class of objects), that provides a basis for decision-making about various words. The common sense view of a concept is the one favored in this introduction to rough sets. Considering *alliteration* to be a concept is consonant with the general meaning of a theoretical concept [1]. That is, a *theoretical concept* is a concept expressed by a theoretical term associated with an axiomatic system underlying logic, mathematics, terms drawn from natural language or theories or that constitute the vocabulary of a particular theory such as poetics or a branch of science such as physics. at odds with the notion of concept in logic, mathematics, scientific theory. Table 1 is a sample decision table.

For example, consider the first line of Part 3 of Keats' *To Autumn*:

Where are the songs of spring? Aye, where are they?

The repetition of the s sound in *songs* and *Spring* is alliterative. Then $d(songs) = d(Spring) = 1$. The word the in the first question in *To Autumn* is not part of an alliteration in that question. Hence, $d(the) = 0$.

Let $V = \bigcup\{V_a \mid a \in C\} \cup \{V_d\}$. Atomic formulae over $B \subseteq C \cup D$ and V are expressions $a = v$ called *descriptors* (*selectors*) over B and V, where $a \in B$ and $v \in V_a$. The set $\mathcal{F}(B, V)$ of formulae over B and V is the least set containing all atomic formulae over B and V and closed with respect to the propositional connectives $\wedge$ (conjunction), $\vee$ (disjunction) and $\neg$ (negation).

By $\|\varphi\|_{\mathcal{A}}$, we denote the meaning of $\varphi \in \mathcal{F}(B, V)$ in the decision table $\mathcal{A}$ which is the set of all objects in $\mathcal{O}$ with the property φ. These sets are defined by $\|a = v\|_{\mathcal{A}} = \{x \in \mathcal{O} \mid a(x) = v\}$, $\|\varphi \wedge \varphi'\|_{\mathcal{A}} = \|\varphi\|_{\mathcal{A}} \cap \|\varphi'\|_{\mathcal{A}}$; $\|\varphi \vee \varphi'\|_{\mathcal{A}} = \|\varphi\|_{\mathcal{A}} \cup \|\varphi'\|_{\mathcal{A}}$; $\|\neg\varphi\|_{\mathcal{A}} = \mathcal{O} - \|\varphi\|_{\mathcal{A}}$ The formulae from $\mathcal{F}(C, V)$, $\mathcal{F}(d, V)$ are called *condition formulae of* $\mathcal{A}$ and *decision formulae of* $\mathcal{A}$, respectively.

Any object $x \in \mathcal{O}$ belongs to the *decision class* $\| \bigwedge_{d \in D} d = d(x)\|_{\mathcal{A}}$ of $\mathcal{A}$. All decision classes of $\mathcal{A}$ create a partition $\mathcal{O}/\mathcal{D}$ of the universe $\mathcal{O}$.

A *decision rule* for $\mathcal{A}$ is any expression of the form $\varphi \Rightarrow \psi$, where $\varphi \in \mathcal{F}(C, V), \psi \in \mathcal{F}(D, V)$, and $\|\varphi\|_{\mathcal{A}} \neq \emptyset$. Formulae φ and ψ are referred to as the *predecessor* and the *successor* of decision rule $\varphi \Rightarrow \psi$. Decision rules are often called "IF ... THEN ..." rules. Such rules are used in machine learning (see, e. g.,[13]).

Decision rule $\varphi \Rightarrow \psi$ is *true* in $\mathcal{A}$ if and only if $\|\varphi\|_{\mathcal{A}} \subseteq \|\psi\|_{\mathcal{A}}$. Otherwise, one can measure its *truth degree* by introducing some inclusion measure of $\|\varphi\|_{\mathcal{A}}$ in $\|\psi\|_{\mathcal{A}}$.

Given two unary predicate formulae $\alpha(x), \beta(x)$ where x runs over a finite set $\mathcal{O}$, Łukasiewicz [28] proposes to assign to $\alpha(x)$ the value $\mathrm{card}(\|\alpha(x)\|)/\mathrm{card}(\mathcal{O})$, where $\|\alpha(x)\| = \{x \in \mathcal{O} : x \text{ satisfies } \alpha\}$. The fractional value assigned to the implication $\alpha(x) \Rightarrow \beta(x)$ is then $\mathrm{card}(\|\alpha(x) \wedge \beta(x)\|)/\mathrm{card}(\|\alpha(x)\|)$ under the assumption that $\|\alpha(x)\| \neq \emptyset$. Proposed by Łukasiewicz, that fractional part was much later adapted by machine learning and data mining literature.

Each object x of a decision system determines a *decision rule*

$$\bigwedge_{a \in C} a = a(x) \Rightarrow \bigwedge_{d \in D} d = d(x) . \tag{1}$$

For any decision table $\mathcal{A} = (\mathcal{O}, C, d)$ one can consider a *generalized decision* function $\partial_A : \mathcal{O} \longrightarrow V_d$ defined by

$$\partial_A(x) = \left\{ i : \exists x' \in \mathcal{O} \left[(x', x) \in I(A) \text{ and } d(x') = i \right] \right\} . \tag{2}$$

$\mathcal{A}$ is called *consistent (deterministic)*, if $\mathrm{card}(\partial_A(x)) = 1$, for any $x \in \mathcal{O}$. Otherwise $\mathcal{A}$ is said to be *inconsistent (nondeterministic)*. Hence, a decision table is inconsistent if it consists of some objects with different decisions but indiscernible with respect to condition attributes. Any set consisting of all objects with the same generalized decision value is called a *generalized decision class*. Now, one can

consider certain (possible) rules (see, e. g. [17]) for decision classes defined by the lower (upper) approximations of such generalized decision classes of $\mathcal{A}$. This approach can be extend, using the relationships of rough sets with the Dempster–Shafer theory (see, e. g., [51]), by considering rules relative to decision classes defined by the lower approximations of unions of decision classes of $\mathcal{A}$.

Numerous methods have been developed for different decision rule generation that the reader can find in the literature on rough sets. Usually, one is searching for decision rules (semi) optimal with respect to some optimization criteria describing quality of decision rules in concept approximations.

In the case of searching for concept approximation in an extension of a given universe of objects (sample), the following steps are typical. When a set of rules has been induced from a decision table containing a set of training examples, they can be inspected to see if they reveal any novel relationships between attributes that are worth pursuing for further research. Furthermore, the rules can be applied to a set of unseen cases in order to estimate their classificatory power.

Dimensionality Reduction

We often face the question whether one or more attributes can be removed and still preserve the basic properties of an information system. Let us express this idea more precisely.

Let $C, D \subseteq A$, be sets of condition and decision attributes respectively. We will say that $C' \subseteq C$ is a *D-reduct* (reduct with *respect* to D) of C, if C' is a minimal subset of C such that

$$\gamma(C, D) = \gamma(C', D) \,. \tag{3}$$

The intersection of all D-reducts is called a *D-core* (core with *respect* to D). Because the core is the intersection of all reducts, it is included in every reduct. Thus, in a sense, the core is the most important subset of attributes, since none of its elements can be removed without affecting the classification power of attributes. Certainly, the geometry of reducts can be more compound. For example, the core can be empty but there can exist a partition of reducts into a few sets with non empty intersection.

Many other kinds of reducts and their approximations are discussed in the literature (see, e. g., [30,60]). For example, if one change the condition (3) to $\partial_A(x) = \partial_B(x)$, then the defined reducts are preserving the generalized decision. Other kinds of reducts are preserving, e. g.,: (i) the distance between attribute value vectors for any two objects, if this distance is greater than a given threshold, (ii) the distance between entropy distributions between any two objects, if this distance exceeds a given threshold [60], or (iii) the so called reducts relative to object used for generation of decision rules [3]. There are some relationships between different kinds of reducts. If B is a reduct preserving the generalized decision, than in B is included a reduct preserving the positive region. For mentioned above reducts based on distances and thresholds one can find analogous dependency between reducts relative to different thresholds. By choosing different kinds of reducts we select different degrees to which information encoded in data is preserved. Reducts are used for building data models. Choosing a particular reduct or a set of reducts has impact on the model size as well as on its quality in describing a given data set. The model size together with the model quality are two basic components tuned in selecting relevant data models. This is known as the minimal length principle (see, e. g., [49]). Selection of relevant kinds of reducts is an important step in building data models. It turns out that the different kinds of reducts can be efficiently computed using heuristics based, e. g., on the Boolean reasoning approach [30,53].

Summary

We are now observing a growing research interest in the foundations of rough sets that include various logical, algebraic, philosophical complexity issues of rough sets. Some relationships have already been established between rough sets and other approaches as well as a wide range of hybrid systems have been developed.

As a result, rough sets are linked with decision system modeling and analysis of complex systems, fuzzy sets, neural networks, evolutionary computing, data mining and knowledge discovery, pattern recognition, machine learning, data mining, and approximate reasoning, multicriteria decision making. In particular, rough sets are used in probabilistic reasoning, granular computing (including information granule calculi based on rough mereology), intelligent control, intelligent agent modeling, identification of autonomous systems, and process specification.

A wide range of applications of methods based on rough set theory alone or in combination with other approaches have been discovered in the following areas: acoustics, bioinformatics, business and finance, chemistry, computer engineering (e. g., data compression, digital image processing, digital signal processing, parallel and distributed computer systems, sensor fusion, fractal engineering), decision analysis and systems, economics, electrical engineering (e. g., control, signal analysis, power systems), environmental studies, digital image process-

ing, informatics, medicine, molecular biology, musicology, neurology, robotics, social science, software engineering, spatial visualization, Web engineering, and Web mining.

For further readings on rough set theory and applications the reader is referred to [40,45] and to books, special issues of journals, issues of Transactions on Rough Sets, and proceedings cited in the bibliography of this article.

Future Directions

One of the main future directions for research based on rough sets in combination with other approaches such as granular computing and other computational intelligence approaches for developing intelligent systems can be found in the framework of Wisdom Technology [21,22], for complex vague concept approximation and approximate reasoning about such concepts by agents or teams of agents searching for solutions of problems in real-life dynamically changing (distributed) environments in which these agents are operating. Such systems consist of autonomous agents operating in highly unpredictable environments and interacting with each other.

In addition, a number of new rough set-based research areas have emerged during the past several years such as various forms of learning, including hierarchical learning or analogy based reasoning [4,6,31,43,55,66], concept approximation [5], multicriteria decision making [15], entropy in information systems [7,61], rough neural computing [32], rough granular computing [52], and a near set approach to image processing [19,27].

Acknowledgments

The research has been supported by the grant from Ministry of Scientific Research and Information Technology of the Republic of Poland and by grant 185986 from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Bibliography

Primary Literature

1. Audi R (1999) The Cambridge Dictionary of Philosophy, 2nd edn. Cambridge University Press, UK
2. Banerjee M, Chakraborty MK (2004) Algebras from rough sets. In: Pal SK, Polkowski L, Skowron A (eds) Rough-Neural Computing. Techniques for Computing with Words. Springer, Berlin, pp 157–184
3. Bazan J (1998) A Comparison of Dynamic and Non-dynamic Rough Set Methods for Extracting Laws from Decision Tables. In: Polkowski L, Skowron A (eds) Rough Sets in Knowledge Discovery 1: Methodology and Applications. Studies in Fuzziness and Soft Computing, vol 18. Physica, Heidelberg, pp 321–365
4. Bazan J (2008) Rough sets and granular computing in behavioral pattern identification and planning In: Pedrycz W, Skowron A, Kreinovich V (eds) Handbook of Granular Computing. Wiley, New York (in press), pp 777–799
5. Bazan J, Skowron A, Swiniarski R (2006) Rough sets and vague concept approximation: From sample approximation to adaptive learning. Transactions on Rough Sets V. Lecture Notes in Computer Science, vol 4100. Springer, Heidelberg, pp 39–62
6. Bazan J, Kruczek P, Bazan-Socha S, Skowron A, Pietrzyk JJ (2006) Automatic planning of treatment of infants with respiratory failure through rough set modeling. In: Greco S, Hato Y, Hirano S, Inuiguchi M, Miyamoto S, Nguyen HS, Słowiński R (eds) Proceedings of the 5th International Conference on Rough Sets and Current Trends in Computing, RSCTC 2006, Kobe, Japan, 6–8 November. Lecture Notes in Artificial Intelligence, vol 4259. Springer, Heidelberg, pp 418–427
7. Bianucci D, Cattaneo G, Ciucci D (2007) Entropies and coentropies of coverings with application to incomplete information systems. Fundam Inform 75:77–105
8. Cattaneo G, Ciucci D (2004) Algebraic structures for rough sets. Transactions on Rough Sets II. LNCS (Lecture Notes in Artificial Intelligence), vol 3135. Springer, Heidelberg, pp 63–101
9. Chakraborty MK, Banerjee M (2007) Rough dialogue and implication lattices. Fundam Inform 75:123–139
10. Devlin J (1993) The Joy of Sets. Fundamentals of Contemporary Set Theory. Springer, Berlin
11. Düntsch I (1997) A logic for rough sets. Theor Comput Sci 179:427–436
12. Frege G (1879) Begriffsschrift: eine der arithmetischen nachgebildete Formelsprache des reinen Denkens. Nebert, Halle
13. Friedman JH, Hastie T, Tibshirani R (2001) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, Heidelberg
14. Gomolińska A (2005) Rough validity, confidence, and coverage of rules in approximation spaces. Transactions on Rough Sets III. LNCS (Lecture Notes in Artificial Intelligence), vol 3400. Springer, Heidelberg, pp 57–81
15. Greco S, Matarazzo B, Słowiński R (2001) Rough sets theory for multicriteria decision analysis. Eur J Oper Res 129:1–47
16. Greco S, Słowiński R, Stefanowski J, Zurawski M (2005) Incremental vs. non-incremental rule induction for multicriteria classification. Transactions on Rough Sets II. LNCS (Lecture Notes in Artificial Intelligence), vol 3135. Springer, Heidelberg, pp 33–53
17. Grzymała-Busse JW (1992) LERS – A system for learning from examples based on rough sets In: Słowiński R (ed) Intelligent Decision Support – Handbook of Applications and Advances of the Rough Sets Theory, System Theory, Knowledge Engineering and Problem Solving, vol 11. Kluwer, Dordrecht, pp 3–18
18. Grzymała-Busse JW, Grzymała-Busse WJ (2007) An experimental comparison of three rough set approaches to missing attribute values. Transactions on Rough Sets VI. LNCS (Lecture Notes in Artificial Intelligence), vol 4374. Springer, Heidelberg, pp 31–50
19. Henry C, Peters JF (2007) Image Pattern Recognition Using Approximation Spaces and Near Sets. In: Proceedings of Eleventh International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. Joint Rough Set Symposium.

Lecture Notes in Artificial Intelligence, vol 4482. Springer, Berlin, pp 475–482
20. Int. Rough Set Society website: www.roughsets.org
21. Jankowski A, Skowron A (2007) Logic for Artificial Intelligence: The Rasiowa – Pawlak School Perspective. In: Ehrenfeucht A, Marek V, Srebrny M (eds) Andrzej Mostowski and Foundational Studies. IOS Press, Amsterdam, pp 106–143
22. Jankowski A, Skowron A (2008) Wisdom Granular Computing. In: Pedrycz W, Skowron A, Kreinovich V (eds) Handbook of Granular Computing. Wiley, New York, pp 329–345
23. Järvinen J (2007) Lattice theory for rough sets. Transactions on Rough Sets VI. LNCS (Lecture Notes in Artificial Intelligence), vol 4374. Springer, Heidelberg, pp 400–496
24. Keats J (1820) Lamia, Isabella, The Eve of St. Agness, and Other Poems. Taylor and Hessey, London
25. Konrad E, Orłowska E, Pawlak Z (1981) Knowledge Representation Systems. Definability of Informations. Prace IPI Pan, ICS PAS Report, vol 433 April. Institute of Computer Science, Polish Academy of Sciences, Warsaw
26. Leśniewski S (1929) Grungzüge eines neuen Systems der Grundlagen der Mathematik. Fundam Math pp 14:1–81
27. Lockery D, Peters JF (2007) Robotic Target Tracking with Approximation Space-based Feedback During Reinforcement Learning, Springer best paper award. In: Proceedings of Eleventh International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. Joint Rough Set Symposium. Lecture Notes in Artificial Intelligence, vol 4482. Springer, Berlin, pp 483–490
28. Łukasiewicz J (1913) Die logischen Grundlagen der Wahrscheinlichkeitsrechnung. In: Borkowski L (ed) Jan Łukasiewicz – Selected Works. North Holland, Amsterdam; (1970) Polish Scientific Publishers, Warsaw, pp 16–63
29. Murray JA, Bradley H, Craigie W, Onions C (1933) The Oxford English Dictionary. Oxford University Press, London
30. Nguyen HS (2006) Approximate boolean reasoning: Foundations and applications in data mining. Transactions on Rough Sets V. LNCS (Lecture Notes in Artificial Intelligence), vol 4100. Springer, Heidelberg, pp 344–523
31. Nguyen SH, Bazan J, Skowron A, Nguyen HS (2004) Layered learning for concept synthesis, Transactions on Rough Sets I. LNCS (Lecture Notes in Artificial Intelligence), vol 3100. Springer, Heidelberg, pp 187–208
32. Pal SK, Polkowski L, Skowron A (2004) Rough-Neural Computing. Techniques for Computing with Words. Springer, Berlin
33. Orłowska E (1985) Semantics of Vague Concepts. In: Dorn G, Weingartner P (eds) Foundations of Logic and Linguistics. Problems and Solutions. Plenum, London, pp 465–482
34. Orłowska E (1985) A logic of indiscernibility relations. Lecture Notes in Computer Science, vol 208. Springer, Berlin, pp 177–186
35. Pal SK, Polkowski L, Skowron A (2004) Rough-Neural Computing: Techniques for Computing with Words. Springer, Berlin
36. Pawlak Z (1973) Mathematical foundations of information retrieval. Proceedings of Symposium of Mathematical Foundations of Computer Science, 3–8 September, High Tartras, pp. 135–136. See also: Mathematical Foundations of Information Retrieval. Computation Center, Polish Academy of Sciences, Research Report CC PAS Report 101, Warsaw
37. Pawlak Z (1981) Classification of objects by means of attributes. Institute for Computer Science, Polish Academy of Sciences. Research Report CC PAS Report 429, Warsaw
38. Pawlak Z (1991) Rough Sets. Theoretical Aspects of Reasoning about Data. Kluwer, Dordrecht
39. Pawlak Z (1995) On rough derivatives, rough integrals, and rough differential equations, ICS Research Report 41/95. Institute of Computer Science, Warsaw
40. Pawlak Z, Skowron A (2007) Rudiments of rough sets. Information Sciences. Int J 177(1): 3–27; Rough sets: Some extensions. Information Sciences. Int J 177(1): 28–40; Rough sets and Boolean reasoning. Information Sciences. Int J 177(1):41–73
41. Peters JF (2008) Classification of perceptual objects by means of features. Int J Inf Technol Intell Comput 3(2):1–35
42. Peters JF, Henry C (2007) Approximation spaces in off-policy Monte Carlo learning. Engineering Applications of Artificial Intelligence. Int J Intell Real-Time Autom 20(5):667–675
43. Peters JF, Henry C, Gunderson DS (2006) Biologically-inspired approximate adaptive learning control strategies: A rough set approach. Inter J Hybrid Intell Syst 3:1–14
44. Peters JF, Skowron A, Stepaniuk J (2007) Nearness of Objects: Extension of Approximation Space Model. Fundam Inform 79(3–4):497–512
45. Polkowski L (2002) Rough Sets: Mathematical Foundations. Advances in Soft Computing. Physica, Heidelberg
46. Polkowski L, Skowron A (1996) Rough mereology: A new paradigm for approximate reasoning. J Approx Reas 15(4):333–365
47. Polkowski L, Skowron A (1998) Rough Sets in Knowledge Discovery 1: Methodology and Applications. Studies in Fuzziness and Soft Computing, vol 18. Physica, Heidelberg
48. Polkowski L, Skowron A (1998) Rough Sets in Knowledge Discovery 2: Applications, Case Studies and Software Systems. Studies in Fuzziness and Soft Computing, vol 19. Physica, Heidelberg
49. Rissanen J (1985) Minimum-description-length Principle. In: Kotz S, Johnson N (eds) Encyclopedia of Statistical Sciences. Wiley, New York, pp 523–527
50. Rough Set Database: http://rsds.univ.rzeszow.pl/
51. Skowron A, Grzymała-Busse JW (1994) From Rough Set Theory to Evidence Theory. In: Yager R, Fedrizzi M, Kacprzyk J (eds) Advances in the Dempster-Shafer Theory of Evidence. Wiley, New York, pp 193–236
52. Skowron A, Peters JF (2008) Rough-granular computing. In: Pedrycz W, Skowron A, Kreinovich V (eds) Handbook of Granular Computing. Wiley, pp 285–328
53. Skowron A, Rauszer C (1992) The Discernibility Matrices and Functions in Information Systems. In: Słowiński R (ed) Intelligent Decision Support – Handbook of Applications and Advances of the Rough Sets Theory, System Theory, Knowledge Engineering and Problem Solving, vol 11. Kluwer, Dordrecht, pp 331–362
54. Skowron A, Stepaniuk J (1996) Tolerance approximation spaces. Fundam Inform 27:245–253
55. Skowron A, Swiniarski R, Synak P (2005) Approximation spaces and information granulation. Transactions on Rough Sets III. LNCS (Lecture Notes in Artificial Intelligence), vol 3400. Springer, Heidelberg, pp 175–189
56. Skowron A, Stepaniuk J, Peters JF, Swiniarski R (2006) Calculi of approximation spaces. Fundam Inform 72(1–3):363–378
57. Stepaniuk J (1998) Approximation spaces, reducts and representatives. In: Polkowski L, Skowron A (eds) Rough Sets in Knowledge Discovery 2: Applications, Case Studies and Soft-

ware Systems, Studies in Fuzziness and Soft Computing, vol 19. Physica, Heidelberg, pp 109–126
58. Stepaniuk J (2000) Knowledge discovery by application of rough set models. In: Polkowski L, Tsumoto S, Lin TY (eds) Rough Set Methods and Applications. New Developments in Knowledge Discovery in Information Systems. Physica, Heidelberg, pp 137–233
59. Stepaniuk J (2007) Approximation Spaces in Multi Relational Knowledge Discovery, Transactions on Rough Sets VI: LNCS (Lecture Notes in Artificial Intelligence), vol 4374. Springer, Heidelberg, pp 351–365
60. Ślęzak D (1996) Approximate reducts in decision tables. Sixth International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU'1996, vol III, Granada, Spain, pp 1159–1164
61. Ślęzak D (2002) Approximate entropy reducts. Fundam Inform 53:365–387
62. Transactions on Rough Sets website: http://www.springer.com/east/home/computer/lncs?SGWID=5-164-6-99627-0, special issues of other journals, proceedings of international conferences, tutorials or surveys (see, references in this paper and, e. g., Pawlak Z, Skowron A (2007) Rudiments of rough sets. Information Sciences. Int J 177(1):3–27; Rough sets: Some extensions. Information Sciences. Int J 177(1):28–40; Rough sets and Boolean reasoning. Information Sciences. Int J 177(1):41–73, and numerous web pages (www.roughsets.org, and logic.mimuw.edu.pl))
63. University of Regina Electronic Bulletin of the Rough Set Community: http://www.cs.uregina.ca/~roughset
64. Vinogradov IM (ed) (1995) Encyclopedia of Mathematics. Kluwer, Dordrecht
65. Warsaw University Logic Group website: logic.mimuw.edu.pl
66. Wojna A (2007) Analogy-based reasoning in classifier construction. Transactions on Rough Sets IV. LNCS (Lecture Notes in Artificial Intelligence), vol 3700. Springer, Heidelberg, pp 277–374
67. Zadeh LA (2001) A new direction in AI: Toward a computational theory of perceptions. AI Mag 22(1):73–84
68. Zadeh LA (2005) Toward a generalized theory of uncertainty (GTU) – An outline. Inform Sci 171:1–40
69. Ziarko W (1993) Variable precision rough set model. J Comput Syst Sci 46:39–59

Further Readings – Books

Demri S, Orłowska E (2002) Incomplete Information: Structure, Inference, Complexity. Monographs in Theoretical Computer Science. Springer, Berlin

Dunin-Kęplicz B, Jankowski A, Skowron A, Szczuka M (eds) (2005) Monitoring, Security, and Rescue Tasks in Multiagent Systems, MSRAS'2004. Advances in Soft Computing. Springer, Heidelberg

Düntsch I, Gediga G (2000) Rough set data analysis: A road to non-invasive knowledge discovery. Methodos, Bangor

Grzymała-Busse JW (1990) Managing Uncertainty in Expert Systems. Kluwer, Norwell

Inuiguchi M, Hirano S, Tsumoto S (eds) (2003) Rough Set Theory and Granular Computing, Studies in Fuzziness and Soft Computing, vol 125. Springer, Heidelberg

Kostek B (2005) Perception-Based Data Processing in Acoustics. Applications to Music Information Retrieval and Psychophysiology of Hearing, Studies in Computational Intelligence, vol 3. Springer, Heidelberg

Lin TY, Cercone N (eds) (1997) Rough Sets and Data Mining – Analysis of Imperfect Data. Kluwer, Boston

Lin TY, Yao YY, Zadeh LA (eds) (2001) Rough Sets, Granular Computing and Data Mining. Studies in Fuzziness and Soft Computing. Physica, Heidelberg

Mitra S, Acharya T (2003) Data mining. Multimedia, Soft Computing, and Bioinformatics. Wiley, New York

Munakata T (ed) (1998) Fundamentals of the New Artificial Intelligence: Beyond Traditional Paradigms. Graduate Texts in Computer Science, vol 10. Springer, New York

Orłowska E (ed) (1997) Incomplete Information: Rough Set Analysis. Studies in Fuzziness and Soft Computing, vol 13. Physica, Heidelberg

Pal SK, Skowron A (eds) (1999) Rough Fuzzy Hybridization: A New Trend in Decision-Making. Springer, Singapore

Polkowski L, Lin TY, Tsumoto S (eds) (2000) Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems, Studies in Fuzziness and Soft Computing, vol 56. Physica, Heidelberg

Słowiński R (ed) (1992) Intelligent Decision Support – Handbook of Applications and Advances of the Rough Sets Theory, System Theory, Knowledge Engineering and Problem Solving, vol 11. Kluwer, Dordrecht

Zhong N, Liu J (eds) (2004) Intelligent Technologies for Information Analysis. Springer, Heidelberg

Further Readings – Transactions on Rough Sets

Peters JF, Skowron A et al (eds) (2004–2008) Transactions on Rough Sets I-VIII. (Lecture Notes in Computer Science), vols 3100, 3135, 3400, 3700, 4100, 4374, 4400, 5084. Springer, Heidelberg

Further Readings – Special Issues of Journals

Cercone N, Skowron A, Zhong N (eds) (2001) Comput Intell Int J 17(3)

Lin TY (ed) (1996) J Intell Autom Soft Comput 2(2)

Pal SK, Pedrycz W, Skowron A, Swiniarski R (eds) (2001) Special volume: Rough-neuro computing. Neurocomputing 36

Peters J, Skowron A (eds) (2001) Special issue: on a rough set approach to reasoning about data. Int J Intell Syst 16(1)

Skowron A, Pal SK (eds) (2003) Special volume: Rough sets, pattern recognition and data mining. Pattern Recognit Lett 24(6)

Słowiński R, Stefanowski J (eds) (1993) Special issue: Proceedings of the First International Workshop on Rough Sets: State of the Art and Perspectives, Kiekrz, Poznań, Poland, 2–4 September 1992. Found Comput Decis Sci 18(3–4)

Ziarko W (ed) (1995) Comput Intell Int J 11(2)

Ziarko W (ed) (1996) Fundam Inform 27(2–3)

Further Readings – Proceedings of International Conferences

Alpigini JJ, Peters JF, Skowron A, Zhong N (eds) (2002) Third International Conference on Rough Sets and Current Trends in Computing, RSCTC'2002, Malvern, PA, 14–16 October. Lecture Notes in Artificial Intelligence, vol 2475. Springer, Heidelberg

An A, Stefanowski J, Ramanna S, Butz CJ, Pedrycz W, Wang G (eds) (2007) Proceedings of the Eleventh International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, RSFDGrC 2007, Toronto, Canada, 14–16 May. Lecture Notes in Artificial Intelligence, vol 4482. Springer, Heidelberg

Greco S, Hata Y, Hirano S, Inuiguchi M, Miyamoto S, Nguyen HS, Słowiński R (eds) (2006) Proceedings of the Fifth International Conference on Rough Sets and Current Trends in Computing, RSCTC 2006, Kobe, Japan, 6–8 November. Lecture Notes in Artificial Intelligence, vol 4259. Springer, Heidelberg

Hirano S, Inuiguchi M, Tsumoto S (eds) (2001) Proceedings of International Workshop on Rough Set Theory and Granular Computing, RSTGC'2001, Matsue, Japan, 20–22 May. Bull Int Rough Set Soc 5(1–2)

Kryszkiewicz M, Peters JF, Rybiński H, Skowron A (eds) (2007) Proceedings of the International Conference on Rough Sets and Intelligent Systems Paradigms, RSEISP 2007, Warsaw, Poland, 28–30 June. Lecture Notes in Artificial Intelligence, vol 4585. Springer, Heidelberg

Lin TY, Wildberger AM (eds) (1995) Soft Computing: Rough Sets, Fuzzy Logic, Neural Networks, Uncertainty Management, Knowledge Discovery. Simulation Councils, USA

Polkowski L, Skowron A (eds) (1998) First International Conference on Rough Sets and Soft Computing, RSCTC'1998. Lecture Notes in Artificial Intelligence, vol 1424. Springer, Poland

Skowron A (ed) (1985) Proceedings of the 5th Symposium on Computation Theory. Zaborów, Poland, 1984. Lecture Notes in Computer Science, vol 208. Springer, Berlin

Skowron A, Szczuka M (eds) (2003) Proceedings of the Workshop on Rough Sets in Knowledge Discovery and Soft Computing at ETAPS, 12–13 April. Electronic Notes in Computer Science, vol 82(4). Elsevier, Netherlands, http://www.elsevier.nl/locate/entcs/volume82.html

Ślęzak D, Wang G, Szczuka M, Düntsch I, Yao Y (eds) (2005) Proceedings of the 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, RSFDGrC'2005, Regina, Canada, 31 August–3 September, Part I. Lecture Notes in Artificial Intelligence, vol 3641. Springer, Heidelberg

Ślęzak D, Yao JT, Peters JF, Ziarko W, Hu X (eds) (2005) Proceedings of the 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, RSFDGrC'2005, Regina, Canada, 31 August–3 September, Part II. Lecture Notes in Artificial Intelligence, vol 3642. Springer, Heidelberg

Terano T, Nishida T, Namatame A, Tsumoto S, Ohsawa Y, Washio T (eds) (2001) New Frontiers in Artificial Intelligence, Joint JSAI'2001. Workshop Post-Proceedings. Lecture Notes in Artificial Intelligence, vol 2253. Springer, Heidelberg

Tsumoto S, Kobayashi S, Yokomori T, Tanaka H, Nakamura A (eds) (1996) Proceedings of the The Fourth Internal Workshop on Rough Sets, Fuzzy Sets and Machine Discovery, University of Tokyo, Japan, 6–8 November. The University of Tokyo, Tokyo

Tsumoto S, Słowiński R, Komorowski J, Grzymała-Busse J (eds) (2004) Proceedings of the 4th International Conference on Rough Sets and Current Trends in Computing, RSCTC'2004, Uppsala, Sweden, 1–5 June. Lecture Notes in Artificial Intelligence, vol 3066. Springer, Heidelberg

Yao JT, Lingras P, Wu WZ, Szczuka M, Cercone NJ, Ślęzak D (eds) (2007) Proceedings of the Second International Conference on Rough Sets and Knowledge Technology, RSKT 2007, Toronto, Canada, 14–16 May. Lecture Notes in Artificial Intelligence, vol 4481. Springer, Heidelberg

Wang G, Liu Q, Yao Y, Skowron A (eds) (2003) Proceedings of the 9th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, RSFDGrC'2003, Chongqing, China, 26–29 May. Lecture Notes in Artificial Intelligence, vol 2639. Springer, Heidelberg

Zhong N, Skowron A, Ohsuga S (eds) (1999) Proceedings of the 7th International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing, RSFDGrC'99, Yamaguchi, 9–11 November. Lecture Notes in Artificial Intelligence, vol 1711. Springer, Heidelberg

Ziarko W (ed) (1994) Rough Sets, Fuzzy Sets and Knowledge Discovery: Proceedings of the Second International Workshop on Rough Sets and Knowledge Discovery, RSKD'93, Banff, Canada, 12–15 October 1993. Workshops in Computing. Springer & British Computer Society, London, Berlin

Ziarko W, Yao Y (2001) Proceedings of the 2nd International Conference on Rough Sets and Current Trends in Computing, RSCTC'2000, Banff, Canada, 16–19 October 2000. Lecture Notes in Artificial Intelligence, vol 2005. Springer, Heidelberg

Rule Induction, Missing Attribute Values and Discretization

Jerzy W. Grzymala-Busse[1,2]

[1] Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, USA

[2] Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

Article Outline

Glossary
Definition of the Subject
Introduction
Discretization
LEM2 Algorithm
Inconsistent Data
Missing Attribute Values
MLEM2
Classification System
Validation
Future Directions
Bibliography

Glossary

Discretization Discretization is a process of converting numerical attributes into symbolic ones by splitting the numerical attribute domain into intervals. Usually discretization is conducted before the main process of rule induction, but in some rule induction algorithms, e. g., in MLEM2 (Modified LEM2), rules are induced concurrently with discretization.

LEM2 algorithm LEM2 (Learning from Examples Module, version 2) is the basic rule induction algorithm of the machine learning/data mining system LERS. LEM2, implemented for the first time in 1990, uses an idea of a local covering to induce a minimal set of minimal rules describing all data concepts.

LERS machine learning/data mining system LERS (Learning from Examples based on Rough Sets) is a rule induction system created at the University of Kansas. Its first implementation was done in Franz Lisp in 1988. This first version of LERS had only one algorithm called LEM1 (Learning form Examples Module, version 1) to induce all rules from input data.

Missing attribute values Missing attribute values frequently affect real-life data. Some attribute values are *lost* (e. g., erased), some are of the type *"do not care" conditions* (such attribute values were irrelevant for classification of the case). In most existing machine learning/data mining systems some method of handling missing attribute values is applied before the main process of rule induction. However, in MLEM2 rule induction and handling missing attribute values are conducted at the same time.

Rule induction Rule induction is understood here as an instance of supervised learning. Rule induction is one of the basic processes of acquiring knowledge (knowledge extraction) in the form of rule sets from raw data. This process is widely used in machine learning (data mining). A data set contains cases (examples) characterized by attribute values and classified as members of concepts by an expert. *Rules* are expressions of the following format:

if condition$_1$ **and** condition$_2$ **and** … **and** condition$_n$ **then** decision.

Definition of the Subject

Rule induction is a process of creating rule sets from raw data called *training data*. Such rules represent hidden and previously unknown knowledge contained in the training data. These rules may be used for successful classification of new cases that were not used for training. One of the possible applications of this methodology is rule-based expert systems. There are many documented examples of successful use of rule induction in medicine (e. g., for decision support for diagnosis), finances, military, etc.

In particular, one of the rule induction systems called LERS has proven its applicability having been used for years by NASA Johnson Space Center (Automation and Robotics Division), as a tool to develop expert systems of the type used in medical decision-making on board the International Space Station. LERS was also used to enhance facility compliance under Sections 311, 312, and 313 of Title III, the Emergency Planning and Community Right to Know. System LERS was used in other areas as well, for example, in the medical field to assess preterm labor risk for pregnant women. Currently used manual methods of assessing preterm birth have a positive predictive value of 17–38%. The data mining methods based on LERS reached positive predictive value of 59–92%. Other applications of LERS to the medical area include diagnosis of melanoma, prediction of behavior under mental retardation and analysis of animal models for prediction of self-injurious behavior. LERS has been used also in nursing, global warming, environmental protection, natural language, and data transmission. LERS can process large datasets and frequently outperforms not only other rule induction systems but also human experts.

Introduction

Data mining uses methods of machine learning to acquire knowledge from raw data. Rule induction is one of the most successful techniques of machine learning. In many application areas, such as medicine, it is essential not only to make an appropriate decision or diagnosis but also to be able to justify or explain the decision. Rules provide such explanations.

We will discuss rule induction using for that purpose one of the very successful algorithms called LEM2 (Learning from Examples Module, version 2), see for example [4]. This algorithm is used in the LERS (Learning from Examples based on Rough Sets) data mining system. LERS can process inconsistent data, i. e., data with conflicting cases, in which values for all attributes are the same yet the decision values are distinct. LERS deals with inconsistent data using a typical rough set approach: it computes lower and upper approximations for all concepts. Then it passes such approximations to LEM2.

Data are frequently incomplete, i. e., some attribute values are missing. Rule sets may be induced from incomplete data using a modified LEM2 algorithm called MLEM2. Again, MLEM2 uses a rough set approach to missing attribute values.

Additionally, many real-life data sets contain numerical attributes, i. e., attributes whose values are integers or real numbers. Numerical data need to be converted into symbolic data by a process called discretization. For a numerical attribute a few intervals are determined, usually these intervals are disjoint and together they cover the entire domain of the numerical attribute. The process of dis-

cretization may be performed before rule induction or, as it is done in MLEM2, during rule induction.

In the entire process of rule learning, handling missing attribute values, and discretization, the applied methodology is based on the same granules called *attribute-value blocks*. Similar granules were discussed in [13,14,15].

Discretization

The input examples are presented in a table, called a decision table. An example of such a table is presented as Table 1. Table 1 represents the input data set in the format of LERS [4]. Table 1 contains four variables: *Age*, *Skill*, *Experience* and *Productivity*. The first three variables are called *attributes*, the last one is called *decision*. Table 1 presents seven cases (or examples) labeled by integers $1, 2, \ldots, 7$.

Any subset of the set of all cases, defined by the same value of the decision is called a *concept*. Table 1 consists of two concepts, defined by (*Productivity*, *low*) and (*Productivity*, *high*). For example, the concept defined by (*Productivity*, *low*) is the set $\{1, 3, 4, 6\}$.

The attribute *Age* from Table 1 is numerical. The numerical attributes should be discretized before rule induction, otherwise rules induced from data with numerical attributes will be too specific. In general, during *discretization* an original range $[a, b]$ of a numerical attribute is partitioned into a set of n intervals

$$\{[a_0, a_1), [a_1, a_2), \ldots, [a_{n-1}, a_n]\},$$

where $a_0 = a$, $a_k < a_{k+1}$ for $k = 0, 1, \ldots, n-1$ and $a_n = b$. The data mining system LERS uses for discretization a number of discretization algorithms [2], including Equal Interval Width, Equal Frequency per Interval, Minimal Class Entropy, and two discretization methods based on Cluster Analysis [5,8]. As an example, let us use one of the simplest discretization methods, discretization based on the Equal Interval Width. Let us start with $n = 2$.

Rule Induction, Missing Attribute Values and Discretization, Table 1
An example of a decision table

Case	Attributes			Decision
	Age	Skill	Experience	Productivity
1	25	low	low	low
2	56	high	high	high
3	36	low	high	low
4	42	high	low	low
5	59	high	low	high
6	25	high	low	low
7	42	high	high	high

Rule Induction, Missing Attribute Values and Discretization, Table 2
An example of an inconsistently discretized decision table

Case	Attributes			Decision
	Age	Skill	Experience	Productivity
1	25..42	low	low	low
2	42..57	high	high	high
3	25..42	low	high	low
4	42..57	high	low	low
5	42..57	high	low	high
6	25..42	high	low	low
7	42..57	high	high	high

Rule Induction, Missing Attribute Values and Discretization, Table 3
An example of a discretized decision table

Case	Attributes			Decision
	Age	Skill	Experience	Productivity
1	25..36.3	low	low	low
2	47.7..59	high	high	high
3	25..36.3	low	high	low
4	36.3..47.7	high	low	low
5	47.7..59	high	low	high
6	25..36.3	high	low	low
7	36.3..47.7	high	high	high

Thus, the range $[25, 59]$ of *Age* is partitioned into two intervals of equal width: $[25, 42)$ and $[42, 59]$. We will denote the first interval by 25..42 and the second interval by 42..59. As a result, the decision table that represents the discretized data set (Table 2) is inconsistent, since cases 4 and 5 are conflicting. Therefore, the number of discretization intervals was too small. Let us try $n = 3$. The new intervals are $[25, 36.3)$, $[36.3..47.7)$, and $[47.7..59]$, denoted, respectively, by 25..36.3, 36.3..47.74, and 47.7..59. The new discretized table is presented in Table 3. This table is consistent and may be considered to be an input data set for rule induction. For more details on discretization methods see [2,6,8,10].

LEM2 Algorithm

LEM2 explores the search space of attribute-value pairs. Its input data set is a lower or upper approximation of a concept, so its input data set is always consistent. In general, LEM2 computes a local covering and then converts it into a rule set. We will quote a few definitions to describe the LEM2 algorithm [1,4,5].

The LEM2 algorithm is based on an idea of an attribute-value pair block. For an attribute-value pair

Rule Induction, Missing Attribute Values and Discretization, Table 4
An example of an inconsistent decision table

Case	Attributes			Decision
	Age	Skill	Experience	Productivity
1	25	low	low	low
2	56	high	high	high
3	36	low	high	low
4	42	high	low	low
5	59	high	low	high
6	25	high	low	low
7	42	high	high	high
8	25	high	low	high

$(a, v) = t$, a *block* of t, denoted by $[t]$, is a set of all cases from U such that for attribute a have value v. Let B be a nonempty lower or upper approximation of a concept represented by a decision-value pair (d, w). Set B *depends* on a set T of attribute-value pairs $t = (a, v)$ if and only if

$$\emptyset \neq [T] = \bigcap_{t \in T} [t] \subseteq B\,.$$

Set T is a *minimal complex* of B if and only if B depends on T and no proper subset T' of T exists such that B depends on T'. Let $\mathcal{T}$ be a nonempty collection of nonempty sets of attribute-value pairs. Then $\mathcal{T}$ is a *local covering* of B if and only if the following conditions are satisfied:

- Each member T of $\mathcal{T}$ is a minimal complex of B,
- $\bigcup_{t \in \mathcal{T}} [T] = B$, and
- $\mathcal{T}$ is minimal, i. e., $\mathcal{T}$ has the smallest possible number of members.

The procedure LEM2, based on rule induction from local coverings, is presented in Fig. 1

For a set X, $|X|$ denotes the cardinality of X.

For Table 3 the following rules are induced by LEM2:

(Age, 25..36.3) → (Productivity, low),
(Age, 36.3..47.7) & (Experience, low) → (Productivity, low),
(Age, 47.7..59) → (Productivity, high),
(Age, 36.3..47.7) & (Experience, high) → (Productivity, high).

Inconsistent Data

LERS handles inconsistencies using rough set theory, introduced by Z. Pawlak in 1982 [16,17]. In rough set theory lower and upper approximations are computed for all concepts involved in conflicts with other concepts.

Procedure LEM2
(**input**: a set B,
output: a single local covering $\mathcal{T}$ of set B);
begin
  $G := B$;
  $\mathcal{T} := \emptyset$;
  while $G \neq \emptyset$
    begin
    $T := \emptyset$;
    $T(G) := \{t | [t] \cap G \neq \emptyset\}$;
    while $T = \emptyset$ **or** $[T] \not\subseteq B$
      begin
        select a pair $t \in T(G)$ such that $|[t] \cap G|$ is maximum; if a tie occurs, select a pair $t \in T(G)$ with the smallest cardinality of $[t]$; if another tie occurs, select first pair;
        $T := T \cup \{t\}$;
        $G := [t] \cap G$;
        $T(G) := \{t | [t] \cap G \neq \emptyset\}$;
        $T(G) := T(G) - T$;
        end {while}
      for each $t \in T$ **do**
        if $[T - \{t\}] \subseteq B$
        then $T := T - \{t\}$;
      $\mathcal{T} := \mathcal{T} \cup \{T\}$;
      $G := B - \cup_{T \in \mathcal{T}} [T]$;
  end {while};
  for each $T \in \mathcal{T}$ **do**
    if $\bigcup_{S \in \mathcal{T} - \{T\}} [S] = B$ **then** $\mathcal{T} := \mathcal{T} - \{T\}$;
end {procedure}.

Rule Induction, Missing Attribute Values and Discretization, Figure 1
Procedure LEM2

Let U denote the set of all examples of the decision table and let P denote a nonempty subset of the set Q of all variables, i. e., attributes and decisions. Let P be a subset of A. An indiscernibility relation ρ on U is defined for all $x, y \in U$ by $x \rho y$ if and only if for both x and y the values for all variables from P are identical. Equivalence classes of ρ are called *elementary sets* of P. An equivalence class of ρ containing x is denoted $[x]_P$. Any finite union of elementary sets of P is called a *definable set* in P. Let X be a concept. In general, X is not a definable set in P. However, set X may be approximated by two definable sets in P, the first one is called a *lower approximation* of X in P,

denoted by $\underline{P}X$ and defined as follows

$$\{x \in U | [x]_P \subseteq X\} .$$

The second set is called an *upper approximation* of X in P, denoted by $\overline{P}X$ and defined as follows

$$\{x \in U | [x]_P \cap X \neq \emptyset\} .$$

The lower approximation of X in A is the greatest definable set in A, contained in X. The upper approximation of X in A is the least definable set in A containing X. A rough set of X is the family of all subsets of U having the same lower and the same upper approximations of X.

Table 4 is an example of inconsistent data. Cases 6 and 8 are conflicting. The lower approximation of the concept $\{1, 3, 4, 6\}$ is the set $\{1, 3, 4\}$ and the upper approximation of the same concept is the set $\{1, 3, 4, 6, 8\}$. The Table 4, discretized using the same Equal Interval Width method that was used before to Table 1, with three intervals, is presented as Table 5. The corresponding rule sets are: the certain rule set:

(Skill, low) → (Productivity, low)
(Age, 36.3..47.7) & (Experience, low) → (Productivity, low)
(Age, 47.7..59) → (Productivity, high),
(Age, 36.3..47.7) & (Experience, high) → (Productivity, high),

and the possible rule set:

(Age, 25..36.3) → (Productivity, low),
(Age, 36.3..47.7) & (Experience, low) → (Productivity, low),
(Skill, high) & (Age, 25..36.3) → (Productivity, high),
(Age, 47.7..59) → (Productivity, high),
(Age, 36.3..47.7) & (Experience, high) → (Productivity, high).

Rule Induction, Missing Attribute Values and Discretization, Table 5
An example of a discretized and inconsistent decision table

Case	Attributes			Decision
	Age	Skill	Experience	Productivity
1	25..36.3	low	low	low
2	47.7..59	high	high	high
3	25..36.3	low	high	low
4	36.3..47.7	high	low	low
5	47.7..59	high	low	high
6	25..36.3	high	low	low
7	36.3..47.7	high	high	high
8	25..36.3	high	low	high

Missing Attribute Values

For incomplete decision tables there are three important and different possibilities to define lower and upper approximations, called singleton, subset, and concept approximations [9]. Singleton lower and upper approximations were studied, for example in [11,12,19,20]. Note that similar definitions of lower and upper approximations, though not for incomplete decision tables, were studied in [15]. Additionally, note that some other rough-set approaches to missing attribute values were presented in [4] as well.

For the rest of the paper we will assume that all decision values are specified, i. e., they are not missing. Also, we will assume that lost values will be denoted by "?" and "do not care" conditions by "$*$". Additionally, we will assume that for each case at least one attribute value is specified.

For incomplete decision tables the definition of a block of an attribute-value pair must be modified in the following way:

- If for an attribute a there exists a case x such that $\rho(x, a) = ?$, i. e., the corresponding value is lost, then the case x should not be included in any blocks $[(a, v)]$ for all values v of attribute a.
- If for an attribute a there exists a case x such that the corresponding value is a "do not care" condition, i. e., $\rho(x, a) = *$, then the case x should be included in blocks $[(a, v)]$ for all specified values v of attribute a.

For Table 6 the blocks of attribute-value pairs are:

$[(\text{Age}, 25)] = \{1, 6\}$,
$[(\text{Age}, 36)] = \{3\}$,
$[(\text{Age}, 42)] = \{4, 7\}$,
$[(\text{Skill}, \text{low})] = \{1, 3, 4, 7\}$,
$[(\text{Skill}, \text{high})] = \{2, 4, 5, 6, 7\}$,

Rule Induction, Missing Attribute Values and Discretization, Table 6
An example of a decision table with missing attribute values

Case	Attributes			Decision
	Age	Skill	Experience	Productivity
1	25	low	*	low
2	?	high	*	high
3	36	low	high	low
4	42	*	low	low
5	?	high	low	high
6	25	high	*	low
7	42	*	high	high

$[(\text{Experience, low})] = \{1, 2, 4, 5, 6\}$,
$[(\text{Experience, high})] = \{1, 2, 3, 6, 7\}$,

For a case $x \in U$ the *characteristic set* $K_B(x)$ is defined as the intersection of the sets $K(x, a)$, for all $a \in B$, where the set $K(x, a)$ is defined in the following way:

- If $\rho(x, a)$ is specified, then $K(x, a)$ is the block $[(a, \rho(x, a)]$ of attribute a and its value $\rho(x, a)$.
- If $\rho(x, a) = ?$ or $\rho(x, a) = *$ then the set $K(x, a) = U$.

For Table 6 and $B = A$,

$$K_A(1) = \{1, 6\} \cap \{1, 3, 4, 7\} = \{1\},$$
$$K_A(2) = \{2, 4, 5, 6, 7\},$$
$$K_A(3) = \{3\} \cap \{1, 3, 4, 7\} \cap \{1, 2, 3, 6, 7\} = \{3\},$$
$$K_A(4) = \{4, 7\} \cap \{1, 2, 4, 5, 6\} = \{4\},$$
$$K_A(5) = \{2, 4, 5, 6, 7\} \cap \{1, 2, 4, 5, 6\} = \{2, 4, 5, 6\},$$
$$K_A(6) = \{1, 6\} \cap \{2, 4, 5, 6, 7\} = \{6\},$$
$$K_A(7) = \{4, 7\} \cap \{1, 2, 3, 6, 7\} = \{7\}.$$

Characteristic set $K_B(x)$ may be interpreted as the set of cases that are indistinguishable from x using all attributes from B and using a given interpretation of missing attribute values. For incomplete decision tables lower and upper approximations may be defined in a few different ways. Here we suggest three different definitions of lower and upper approximations for incomplete decision tables, following [9]. Let B be a subset of the set A of all attributes. Let X be any subset of the set U of all cases. The set X is called a *concept* and is usually defined as the set of all cases defined by a specific value of the decision.

Our first definition uses a similar idea as in the previous articles on incomplete decision tables [11,12,19,20], i. e., lower and upper approximations are sets of singletons from the universe U satisfying some properties. We will call these approximations *singleton*. A singleton B-lower approximation of X is defined as follows:

$$\underline{B}X = \{x \in U | K_B(x) \subseteq X\} .$$

A singleton B-upper approximation of X is

$$\overline{B}X = \{x \in U | K_B(x) \cap X \neq \emptyset\} .$$

In our example of the decision table presented in Table 6 let us say that $B = A$. Then the singleton A-lower and A-upper approximations of the two concepts: $\{1, 3, 4, 6\}$ and $\{2, 5, 7\}$ are:

$$\underline{A}\{1, 3, 4, 6\} = \{1, 3, 4, 6\} ,$$
$$\underline{A}\{2, 5, 7\} = \{7\} ,$$
$$\overline{A}\{1, 3, 4, 6\} = \{1, 2, 3, 4, 5, 6\} ,$$
$$\overline{A}\{2, 5, 7\} = \{2, 4, 5, 6, 7\} .$$

The next possibility is to define lower and upper approximations for incomplete decision tables using characteristic sets instead of singletons. There are two ways to do this. Using the first way, a *subset* B-lower approximation of X is defined as follows:

$$\underline{B}X = \cup\{K_B(x) | x \in U, K_B(x) \subseteq X\} .$$

A *subset* B-upper approximation of X is

$$\overline{B}X = \cup\{K_B(x) | x \in U, K_B(x) \cap X \neq \emptyset\} .$$

For the same decision table, presented in Table 6, the subset A-lower and A-upper approximations are

$$\underline{A}\{1, 3, 4, 6\} = \{1, 3, 4, 6\} ,$$
$$\underline{A}\{2, 5, 7\} = \{7\} ,$$
$$\overline{A}\{1, 3, 4, 6\} = \{1, 2, 3, 4, 6, 7\} = U ,$$
$$\overline{A}\{2, 5, 7\} = \{2, 4, 5, 6, 7\} .$$

The second way is to modify the subset definition of lower and upper approximation by replacing the universe U from the subset definition by a concept X. A *concept* B-lower approximation of the concept X is defined as follows:

$$\underline{B}X = \cup\{K_B(x) | x \in X, K_B(x) \subseteq X\} .$$

A *concept* B-upper approximation of the concept X is defined as follows:

$$\overline{B}X = \cup\{K_B(x) | x \in X, K_B(x) \cap X \neq \emptyset\}$$
$$= \cup\{K_B(x) | x \in X\} .$$

For the decision table presented in Table 6, the concept A-upper approximations are

$$\overline{A}\{1, 3, 4, 6\} = \{1, 3, 4, 6\} ,$$
$$\overline{A}\{2, 5, 7\} = \{2, 4, 5, 6, 7\} .$$

Note that for complete decision tables, all three definitions of lower approximations, singleton, subset and concept, coalesce to the same definition. Also, for complete decision tables, all three definitions of upper approximations coalesce to the same definition. This is not true for incomplete decision tables, as our example shows.

MLEM2

MLEM2, a modified version of LEM2 [7], processes numerical attributes differently than symbolic attributes. For numerical attributes MLEM2 sorts all values of a numerical attribute. Then it computes cutpoints as averages for

any two consecutive values of the sorted list. For each cutpoint q MLEM2 creates two blocks, the first block contains all cases for which values of the numerical attribute are smaller than q, the second block contains remaining cases, i. e., all cases for which values of the numerical attribute are larger than q. The search space of MLEM2 is the set of all blocks computed this way, together with blocks defined by symbolic attributes. Starting from that point, rule induction in MLEM2 is conducted the same way as in LEM2. At the very end MLEM2 simplifies rules by merging appropriate intervals for numerical attributes.

The MLEM algorithm induced the following rules from Table 6 certain rule set:

(Age, 25..39) → (Productivity, low),
(Skill, low) & (Experience, low) → (Productivity, low),
(Age, 39..42) & (Experience, high) → (Productivity, high),

and possible rule set:

(Age, 25..39) → (Productivity, low),
(Skill, low) & (Experience, low) → (Productivity, low),
(Skill, high) → (Productivity, high).

Classification System

Rule sets, induced from data sets, are used mostly to classify new, unseen cases. Such rule sets may be used in rule-based expert systems.

There are a few existing classification systems, for example, associated with rule induction systems LERS or AQ. A classification system used in LERS is a modification of the well-known bucket brigade algorithm [5,18]. In the rule induction system AQ, the classification system is based on a rule estimate of probability. Some classification systems use a decision list, in which rules are ordered, the first rule that matches the case classifies it. In this section we will concentrate on a classification system used in LERS.

The decision as to which concept a case belongs to is made on the basis of three factors: *strength*, *specificity*, and *support*. These factors are defined as follows: *strength* is the total number of cases correctly classified by the rule during training. *Specificity* is the total number of attribute-value pairs on the left-hand side of the rule. The matching rules with a larger number of attribute-value pairs are considered more specific. The third factor, *support*, is defined as the sum of products of strength and specificity for all matching rules indicating the same concept. The concept C for which the support, i. e., the following expression

$$\sum_{\substack{\text{matching rules } r \\ \text{describing } C}} \text{Strength}(r) * \text{Specificity}(r)$$

is the largest is the winner and the case is classified as being a member of C.

In the classification system of LERS, if complete matching is impossible, all partially matching rules are identified. These are rules with at least one attribute-value pair matching the corresponding attribute-value pair of a case. For any partially matching rule r, the additional factor, called *Matching_factor* (r), is computed. Matching_factor (r) is defined as the ratio of the number of matched attribute-value pairs of r with a case to the total number of attribute-value pairs of r. In partial matching, the concept C for which the following expression is the largest

$$\sum_{\substack{\text{partially matching} \\ \text{rules } r \text{ describing } C}} \text{Matching_factor}(r) * \text{Strength}(r) * \text{Specificity}(r)$$

is the winner and the case is classified as being a member of C.

Validation

The most important performance criterion of rule induction methods is the error rate. If the number of cases is less than 100, the *leaving-one-out* method is used to estimate the error rate of the rule set. In leaving-one-out, the number of learn-and-test experiments is equal to the number of cases in the data set. During the ith experiment, the ith case is removed from the data set, a rule set is induced by the rule induction system from the remaining cases, and the classification of the omitted case by rules produced is recorded. The error rate is computed as

$$\frac{\text{total number of misclassifications}}{\text{total number of cases}}.$$

On the other hand, if the number of cases in the data set is greater than or equal to 100, the *tenfold cross-validation* should be used. This technique is similar to leaving-one-out in that it follows the learn-and-test paradigm. In this case, however, all cases are randomly re-ordered, and then a set of all cases is divided into ten mutually disjoint subsets of approximately equal size. For each subset, all remaining cases are used for training, i. e., for rule induction, while the subset is used for testing. This method is used primarily to save time at the negligible expense of accuracy.

Tenfold cross validation is commonly accepted as a standard way of validating rule sets. However, using this method twice, with different preliminary random re-ordering of all cases yields – in general – two different estimates for the error rate [5]. For large data sets (at least 1000

cases) a single application of the train-and-test paradigm may be used. This technique is also known as *holdout*. Two thirds of cases should be used for training, one third for testing.

In yet another way of validation, *resubstitution*, it is assumed that the training data set is identical with the testing data set. In general, an estimate for the error rate is here too optimistic. However, this technique is used in many applications.

Future Directions

Rule induction, a part of supervised learning, is subject to extensive research. Many possible extensions for rule induction include *boosting* or *ensemble learning*, where ensembles of classifiers vote on the final outcome during classification of a new case. *Semi-supervised learning*, where for learning is used not only the case set, pre-classified by an expert, but also a set of not classified cases, is another example.

Bibliography

1. Chan CC, Grzymala-Busse JW (1991) On the attribute redundancy and the learning programs ID3, PRISM, and LEM2. Department of Computer Science, University of Kansas, TR-91-14 20
2. Chmielewski MR, Grzymala-Busse JW (1996) Global discretization of continuous attributes as preprocessing for machine learning. Int J Approx Reason 15:319–331
3. Grzymala-Busse JW (1988) Knowledge acquisition under uncertainty – A rough set approach. J Intell Robot Syst 1:3–16
4. Grzymala-Busse JW (1992) LERS – A system for learning from examples based on rough sets. In: Slowinski R (ed) Intelligent decision support. Handbook of applications and advances of the rough set theory. Kluwer, Dordrecht, pp 3–18
5. Grzymala-Busse JW (1997) A new version of the rule induction system LERS. Fundam Inform 31:27–39
6. Grzymala-Busse JW (2002) Discretization of numerical attributes. In: Klösgen W, Zytkow J (eds) Handbook of data mining and knowledge discovery. Oxford University Press, New York, pp 218–225
7. Grzymala-Busse JW (2002) MLEM2: A new algorithm for rule induction from imperfect data. In: Proceedings of the 9th international conference on information processing and management of uncertainty in knowledge-based systems, IPMU 2002, Annecy, France, pp 243–250
8. Grzymala-Busse JW (2003) A comparison of three strategies to rule induction from data with numerical attributes. In: Proceedings of the international workshop on rough sets in knowledge discovery (RSKD 2003), in conjunction with the European joint conferences on theory and practice of software, Warsaw, pp 132–140
9. Grzymala-Busse JW (2003) Rough set strategies to data with missing attribute values. In: Workshop notes, foundations and new directions of data mining, in conjunction with the 3rd IEEE international conference on data mining, Melbourne, FL, pp 56–63
10. Grzymala-Busse JW (2007) Mining numerical data – A rough set approach. In: Proceedings of the RSEISP'2007, the international conference of rough sets and emerging intelligent systems paradigms, Warsaw, Poland. Lecture Notes in artificial intelligence, vol 4585. Springer, Berlin, pp 12–21
11. Kryszkiewicz M (1995) Rough set approach to incomplete information systems. In: Proceedings of the second annual joint conference on information sciences, pp 194–197
12. Kryszkiewicz M (1999) Rules in incomplete information systems. Inf Sci 113:271–292
13. Lin TY (1989) Chinese Wall security policy – An aggressive model. In: Proceedings of the fifth aerospace computer security application conference, Tucson, AZ, pp 286–293
14. Lin TY (1989) Neighborhood systems and approximation in database and knowledge base systems. In: Proceedings of the ISMIS-89, the fourth international symposium on methodologies of intelligent systems, Charlotte, NC, pp 75–86
15. Lin TY (1992) Topological and fuzzy rough sets. In: Slowinski R (ed) Intelligent decision support. Handbook of applications and advances of the rough set theory. Kluwer, Dordrecht, pp 287–304
16. Pawlak Z (1982) Rough Sets. Int J Comput Inf Sci 11:341–356
17. Pawlak Z (1991) Rough Sets: Theoretical aspects of reasoning about data. Kluwer, Dordrecht
18. Stefanowski J (2001) Algorithms of decision rule induction in data mining. Poznan University of Technology Press, Poznan
19. Stefanowski J, Tsoukias A (1999) On the extension of rough sets under incomplete information. In: Proceedings of the RSFDGrC'1999, 7th international workshop on new directions in rough sets, data mining, and granular-soft computing, pp 73–81
20. Stefanowski J, Tsoukias A (2001) Incomplete information tables and rough classification. Comput Intell 17:545–566

S

Scaling Limits of Large Systems of Non-linear Partial Differential Equations

D. Benedetto, M. Pulvirenti
Dipartimento di Matematica, Università di Roma 'La Sapienza', Roma, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Weak-Coupling Limit for Classical Systems
Weak-Coupling Limit for Quantum Systems
Weak-Coupling Limit in the Bose–Einstein and Fermi–Dirac Statistics
Weak-Coupling Limit for a Single Particle: The Linear Theory
Future Directions
Bibliography

Glossary

Scaling limits A scaling limit denotes a procedure to reduce the degree of complexity of a large particle system. It consists in scaling the space-time variables and, possibly, other quantities (like the interaction potential or the density), in order to obtain a more handable description of the same system. The initial space-time coordinates are called **microscopic**, while the new ones, those well suited for the description of kinetic or hydrodynamical systems, are called **macroscopic**.

Boltzmann equation It is an integro-differential kinetic equation for the one particle distribution function in the classical phase space (see Eq. (64) below). It arises in some physical regimes, namely for a rarefied gas and for a weakly interacting quantum dense gas.

Uehling–Uhlenbeck equation It is a Boltzmann type equation taking into account corrections due to the Bose–Einstein or the Fermi–Dirac statistics (see Eq. (69) below). It holds in the weak-coupling limit.

Fokker–Planck–Landau equation It is a kinetic equation diffusive in velocity (see Eq. (28) below). It arises in the context of a weakly interacting classical dense gas.

Hydrodynamical equations They are evolution equations for macroscopic quantities like density, mean velocity, temperature, and so on.

Low density limit Sometimes called Boltzmann–Grad limit, it is a scaling limit in which the density is vanishing. Applied to classical particle systems it gives the Boltzmann equation.

Weak coupling limit It is a scaling limit in which the density is constant but the interaction vanishes suitably. Applied to a quantum particle systems it gives a Boltzmann equation. Applied to a classical particle systems it gives the Fokker–Planck–Landau equation.

Hydrodynamic limit In this scaling we simply pass from micro to macro variables. We look at the behavior of suitable mean values, which are functions of the space and the time. We expect them to behave in a hydrodynamical way, namely to satisfy a set of hydrodynamical equations.

Wigner transform It is the description of a quantum state as a function in the classical phase space.

Definition of the Subject

Many interesting systems in Physics and Applied Sciences are constituted by a large number of identical components so that they are difficult to analyze from a mathematical point of view. On the other hand, quite often, we are not interested in a detailed description of the system but rather to his global behavior. Therefore it is necessary to look for all procedures leading to simplified models, retaining all the interesting features of the original system, cutting away unnecessary informations. This is exactly the methodology of the Statistical Mechanics and the Kinetic Theory when dealing with large particle systems.

In this contribution we want to approach a problem of this type, namely we try to outline the difficulties in making rigorous the limiting procedure leading the microscopic description of a large particle system (based on the fundamental laws like Newton or Schrödinger equations) to a more handable kinetic or hydrodynamic picture. Such a transition depends on the space-time scale we choose to describe the phenomenon. To fix the ideas we start by considering a system constituted by N identical classical point particles of unitary mass. The Newton equations are:

$$\frac{\mathrm{d}^2}{\mathrm{d}\tau^2} q_i = \sum_{\substack{j=1\ldots N:\\ j\neq i}} F(q_i - q_j)\,, \tag{1}$$

where $\{q_1 \ldots q_N\}$, $q_i \in \mathbb{R}^3$ are the position of the particles and τ is the time. Here $F = -\nabla\phi$ denotes the interparticle (conservative) force and ϕ the two-body interaction potential.

We now introduce a small parameter $\varepsilon > 0$ expressing the ratio between the macroscopic and the microscopic scales. For instance ε could denote the inverse ratio of a typical distance between two molecules of a gas, measured in microscopic unities, and the same distance measured in meters. It can be as well the ratio between typical macroscopic and microscopic times. We now introduce the new variables

$$x = \varepsilon q \qquad t = \varepsilon\tau\,.$$

Since we are interested in the macroscopic properties of the system, namely in the evolution of macroscopic quantities that are those varying on the (x, t) scale and hence almost constant in the (q, τ) scale, it is natural to rescale also Eq. (1). We write

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} x_i = \frac{1}{\varepsilon} \sum_{\substack{j=1\ldots N:\\ j\neq i}} F\left(\frac{x_i - x_j}{\varepsilon}\right). \tag{2}$$

Up to now we did nothing else than an innocent change of variables. However to obtain a non trivial description, N has to diverge when $\varepsilon \to 0$, and we have to specify how. Also additional hypotheses on the strength of the interaction, according to the physical situation at hand, may be necessary. There are various possible scaling limits we shall discuss below.

- *The hydrodynamic limit*
  Here the system is dense, namely $N = O(\varepsilon^{-3})$. For all $x \in \Omega \subset \mathbb{R}^3$, we consider a small box Δ around x. In Δ there are a very large number of particles. We denote the initial density, mean velocity and temperature by $(\rho_0(x), u_0(x), T_0(x))$. At a positive time t the particles starts to interact. For a microscopic time τ the particles in Δ interact practically among themselves only (supposing a short range interaction) and, if an ergodic property would hold, we expect that they reach a thermal equilibrium state. We remind that $t = \varepsilon\tau$ is essentially vanishing, therefore we expect that at macroscopic time $t \approx 0$ the particle are distributed according to the Gibbs measure characterized by the parameters $(\rho_0(x), u_0(x), T_0(x))$, being mass momentum and energy locally conserved. On a larger time scale ($t > 0$) mass momentum and energy are not locally conserved anymore. If a local equilibrium structure is preserved, the parameters of the equilibria $(\rho(x, t), u(x, t), T(x, t))$ will evolve according to five equations which are the exactly the well known Euler equations for compressible gas.
  This complex mechanism is very far to be proven rigorously, even for short times and initial regular data. A heuristic derivation of the Euler equations in terms of particle dynamics was first given by Morrey [35]. We address the reader to [39] and [19] for further references and comments.
- *The low-density (Boltzmann–Grad) limit*
  Now we are dealing with a rarefied gas assuming $N = O(\varepsilon^{-2})$. This is the reason why this scaling is called low-density. We assume also, just for simplicity, that the interaction range of ϕ is one. Consider now a given test particle. If the gas is more or less homogeneous, the number of particles interacting with this particle in a given (macroscopic) time interval is $O(N\varepsilon^2) = O(1)$. Therefore we expect a finite number of collisions per unit time. Moreover each collision is almost instantaneous so that each particle undergoes a jump process in velocity.
  We also assume that initially the particles are independently distributed according to the one-particle distribution on the phase space $f_0(x, v)$ (v is the velocity). Of course the dynamics creates correlations so that, strictly speaking, we expect that the particles, for every positive time t, are not independently distributed anymore. However the probability that a given pair of particles interact is vanishing in the limit $\varepsilon \to 0$. Therefore the statistical independence (called propagation of chaos) is recovered in the same limit. In other words we expect, for the time evolved distribution $f(x, v, t)$, a nonlinear evolution equation, which is the celebrated Boltzmann equation discovered in 1872. This scaling limit, often called the Boltzmann–Grad limit [25], has been proved rigorously by Lanford [31] for a short time

interval (see also [28] for a special case in which the limit holds globally in time). We address the reader to refs. [14,39] and [19] for additional comments and references.

- *The weak-coupling limit*
  Now the gas is dense because we assume $N = O(\varepsilon^{-3})$, however the particles are weakly interacting. This condition is expressed by rescaling the interaction potential by setting

$$\phi \to \sqrt{\varepsilon}\phi \,.$$

After that the equation of motion (in macroscopic variables) become:

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} x_i = -\frac{1}{\sqrt{\varepsilon}} \sum_{\substack{j=1\ldots N:\\ j\neq i}} \nabla\phi\left(\frac{x_i - x_j}{\varepsilon}\right) . \tag{3}$$

To have a vague idea of the behavior of a test particle in this limit, we observe first that the velocity change due to a single collision is

$$\Delta v \approx F\Delta t \approx O(\varepsilon/\sqrt{\varepsilon}) = O(\sqrt{\varepsilon}) \,.$$

Here $\Delta t = O(\varepsilon)$ is the time in which the collision takes place, ϕ is assumed short-range in microscopic variables and then $1/\sqrt{\varepsilon}$ is the order of magnitude of the force F.
The number of collisions per unit time is $\varepsilon^2\varepsilon^{-3} = \varepsilon^{-1}$, so that

$$\sum |\Delta v|^2 = O(1) \,.$$

Therefore, provided that the propagation of chaos is ensured, $v = v(t)$ is expected to be not absolutely continuous in the limit and the one-particle distribution function to satisfy a diffusion equation in velocity, the Landau–Fokker–Planck equation.
The statistical independence at a positive time t is expected because the effect of the interaction between two given particle is going to vanish in the limit $\varepsilon \to 0$. Note that the physical meaning of the propagation of chaos here is quite different from that arising in the contest of the Boltzmann equation. Here two particles can interact but the effect of the collision is small, while in a low-density regime the effect of a collision between two given particles is large but quite unlikely.
The above scalings make sense as well for quantum systems. What we expect in this case? Notice that the transition from micro to macro variables increase the frequency of the quantum oscillations so that we are always in presence of a semiclassical limit as regards the kinetic energy part of the Hamiltonian. However the potential varies on the same scale of the oscillations. As a consequence we face the following scenario. The hydrodynamic limit of quantum systems yields the Euler equations for the equilibria parameters ρ, u, T. The only difference with the classical case is that the relationship between the pressure and the hydrodynamical fields, is that dictated by the Quantum Statistical Mechanics. Indeed the local equilibrium is achieved at a microscopic level and this is the only quantum effect we expect to survive in the limit.
The situation is conceptually similar for the low density limit: here we have the classical Boltzmann equation for the one-particle distribution function. The only quantum macroscopic effect is that the cross-section must be computed in terms of the quantum scattering process (see [6]).

Introduction

The hydrodynamical and the low density limits are relatively more popular then the weak-coupling limit, which will be, for that reason, the object of the present contribution. More precisely we analyze large classical and quantum particle systems in the weak-coupling regime. We show how we expect such systems to be well described by the Landau–Fokker–Planck and the Boltzmann equations, for the classical and the quantum case respectively. We also describe the effects of the Fermi–Dirac and Bose–Einstein statistics.

Unfortunately our arguments we present here are largely formal because the rigorous results we know up to now, are few. A rigorous proof of the validity of the weak-coupling limit, or a rigorous derivation of the corresponding kinetic equations, is a challenging and still open problem.

In Sect. "Weak-Coupling Limit for Classical Systems" we introduce and discuss the problem for classical systems, presenting a formal proof of the validity of the Landau equation.

In Sect. "Weak-Coupling Limit for Quantum Systems" we pass to analyze the case of a quantum system, neglecting the correlations due to the statistics. The case of Bosons and Fermions is discussed in Sect. "Weak-Coupling Limit in the Bose–Einstein and Fermi–Dirac Statistics".

Finally, in Sect. "Weak-Coupling Limit for a Single Particle: the Linear Theory" we briefly introduce and discuss the corresponding linear problems, namely the behavior of a single (classical or quantum) particle in a random distribution of obstacles.

Section "Future Directions" is devoted to concluding remarks.

Weak-Coupling Limit for Classical Systems

We consider a classical system of N identical particles of unitary mass. Rescaled positions and velocities are denoted by $x_1 \dots x_N$ and $v_1 \dots v_N$. After having also rescaled the potential in the following way

$$\phi \to \sqrt{\varepsilon}\phi \,,$$

the Newton equations reads as

$$\frac{\mathrm{d}}{\mathrm{d}t} x_i = v_i \quad \frac{\mathrm{d}}{\mathrm{d}t} v_i = -\frac{1}{\sqrt{\varepsilon}} \sum_{\substack{j=1\dots N:\\ j\neq i}} \nabla\phi\left(\frac{x_i - x_j}{\varepsilon}\right) . \quad (4)$$

We also assume that $N = O(\varepsilon^{-3})$, namely the density is $O(1)$.

We are interested in the statistical behavior of the system so that we introduce a probability distribution on the phase space of the system, $W^N = W^N(X_N, V_N)$, that is the state at time zero. Here (X_N, V_N) denote the set of positions and velocities:

$$X_N = (x_1, \dots, x_N) \quad V_N = (v_1, \dots, v_N) \,.$$

Then from Eq. (4) we obtain the following Liouville equation

$$(\partial_t + V_N \cdot \nabla_N) W^N(X_N, V_N) = \frac{1}{\sqrt{\varepsilon}} \left(T_N^\varepsilon W^N\right)(X_N, V_N) \quad (5)$$

where $V_N \cdot \nabla_N = \sum_{i=1}^N v_i \cdot \nabla_{x_i}$ and $(\partial_t + V_N \cdot \nabla_N)$ is the usual free stream operator. Also, we have introduced the operator

$$\left(T_N^\varepsilon W^N\right)(X_N, V_N) = \sum_{0<k<\ell\leq N} \left(T_{k,\ell}^\varepsilon W^N\right)(X_N, V_N) , \quad (6)$$

with

$$T_{k,\ell}^\varepsilon W^N = \nabla\phi\left(\frac{x_k - x_\ell}{\varepsilon}\right) \cdot (\nabla_{v_k} - \nabla_{v_\ell}) W^N \,. \quad (7)$$

To investigate the limit $\varepsilon \to 0$, $N = \varepsilon^{-3}$, it is convenient to introduce the BBKGY hierarchy (from Bogoliubov, Born, Green, Kirkwood and Yvon, see e. g. [2] and [19]), for the j-particle distributions defined as

$$f_j^N(X_j, V_j) = \int \mathrm{d}x_{j+1} \dots \int \mathrm{d}x_N \int \mathrm{d}v_{j+1} \dots \int \mathrm{d}v_N \, W^N(X_j, x_{j+1} \dots x_N, V_j, v_{j+1} \dots v_N) \quad (8)$$

for $j = 1, \dots, N-1$. Obviously, we set $f_N^N = W^N$.

From now on we shall suppose that, due to the fact that the particles are identical, the function W^N and f_j^N which we have introduced are all symmetric in the exchange of particles.

A partial integration of the Liouville equation (5) and standard manipulations give us the following hierarchy of equations: (for $1 \leq j \leq N$):

$$\left(\partial_t + \sum_{k=1}^j v_k \cdot \nabla_k\right) f_j^N = \frac{1}{\sqrt{\varepsilon}} T_j^\varepsilon f_j^N + \frac{N-j}{\sqrt{\varepsilon}} C_{j+1}^\varepsilon f_{j+1}^N . \quad (9)$$

The operator C_{j+1}^ε is defined as:

$$C_{j+1}^\varepsilon = \sum_{k=1}^j C_{k,j+1}^\varepsilon \,, \quad (10)$$

and

$$\begin{aligned} &C_{k,j+1}^\varepsilon f_{j+1}(x_1 \dots x_j, v_1 \dots v_j) \\ &= -\int \mathrm{d}x_{j+1} \int \mathrm{d}v_{j+1} F\left(\frac{x_k - x_\ell}{\varepsilon}\right) \\ &\cdot \nabla_{v_k} f_{j+1}(x_1, x_2, \dots, x_{j+1}, v_1, \dots, v_{j+1}) \,. \end{aligned} \quad (11)$$

$C_{k,j+1}^\varepsilon$ describes the "collision" of particle k, belonging to the j-particle subsystem, with a particle outside the subsystem, conventionally denoted by the number $j+1$ (this numbering uses the fact that all particles are identical). The total operator C_{j+1}^ε takes into account all such collisions. The dynamics of the j-particle subsystem is governed by three effects: the free-stream operator, the "recollisions", i. e. the collisions "inside" the subsystem, given by the T term, and the "creations", i. e. the collisions with particles "outside" the subsystem, given by the C term.

We finally fix the initial value $\{f_j^0\}_{j=1}^N$ of the solution $\{f_j^N(t)\}_{j=1}^N$ assuming that $\{f_j^0\}_{j=1}^N$ is factorized, that is, for all $j = 1, \dots N$

$$f_j^0 = f_0^{\otimes j} \,, \quad (12)$$

where f_0 is a given one-particle distribution function. This means that any pair of particles are statistically uncorrelated at time zero. Of course such a statistical independence is destroyed at time $t > 0$ and Eq. (9) shows that the time evolution of f_1^N is determined by the knowledge of f_2^N which turns out to be dependent on f_3^N and so on. However, since the interaction between two given particle is going to vanish in the limit $\varepsilon \to 0$, we can hope that such statistical independence, namely the factorization property (12), is recovered in the same limit.

Therefore we expect that in the limit $\varepsilon \to 0$ the one-particle distribution function f_1^N converges to the solution

of a suitable nonlinear kinetic equation f which we are going to investigate.

If we expand $f_j^N(t)$ as a perturbation of the free flow $S(t)$ defined as

$$(S(t)f_j)(X_j, V_j) = f_j(X_j - V_j t, V_j)\,, \tag{13}$$

we find

$$f_j^N(t) = S(t)f_j^0 + \frac{N-j}{\sqrt{\varepsilon}}\int_0^t S(t-t_1)C_{j+1}^{\varepsilon} f_{j+1}^N(t_1)\mathrm{d}t_1 + \frac{1}{\sqrt{\varepsilon}}\int_0^t S(t-t_1)T_j^{\varepsilon} f_j^N(t_1)\mathrm{d}t_1\,. \tag{14}$$

We now try to keep informations on the limit behavior of $f_j^N(t)$. Assuming for the moment that the time evolved j-particle distributions $f_j^N(t)$ are smooth (in the sense that the derivatives are uniformly bounded in ε), then

$$C_{j+1}^{\varepsilon} f_{j+1}^N(X_j, V_j, t_1) = -\varepsilon^3 \sum_{k=1}^{j}\int \mathrm{d}r \int \mathrm{d}v_{j+1} F(r) \cdot \nabla_{v_k} f_{j+1}(X_j, x_k - \varepsilon r, V_j, v_{j+1}, t_1)\,. \tag{15}$$

Assuming now, quite reasonably, that

$$\int \mathrm{d}r F(r) = 0\,, \tag{16}$$

we find that

$$C_{j+1}^{\varepsilon} f_{j+1}^N(X_j, V_j, t_1) = O(\varepsilon^4)$$

provided that $D_v^2 f_{j+1}^N$ is uniformly bounded. Since

$$\frac{N-j}{\sqrt{\varepsilon}} = O(\varepsilon^{\frac{7}{2}})$$

we see that the second term in the right hand side of (14) does not give any contribution in the limit. Moreover

$$\int_0^t S(t-t_1)T_j^{\varepsilon} f_j^N(t_1)\mathrm{d}t_1 = \sum_{i\neq k}\int_0^t \mathrm{d}t_1 \cdot F\left(\frac{(x_i - x_k) - (v_i - v_k)(t-t_1)}{\varepsilon}\right) g(X_j, V_j, t_1)\,, \tag{17}$$

where g is a smooth function.

Obviously the above time integral is $O(\varepsilon)$ so that also the last term in the right hand side of (14) does not give any contribution in the limit. Then we are facing the alternative: either the limit is trivial or the time evolved distributions are not smooth. However we believe that the limit is not trivial (actually we expect to get a diffusion equation, according to the previous discussion) and a rigorous proof of this fact seems problematic.

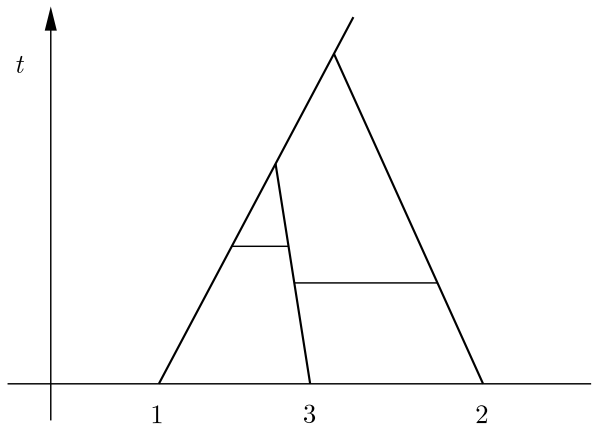

Scaling Limits of Large Systems of Non-linear Partial Differential Equations, Figure 1
The collision sequence (1, 2), (1, 3), (1, 3), (2, 3)

The difficulty in obtaining a-priori estimates induce us to exploit the full series expansion of the solution, namely

$$f_1^N(t) = \sum_{n\geq 0}\sum_{G_n} K(G_n)\int_0^t \mathrm{d}t_1 \int_0^{t_1}\mathrm{d}t_2 \ldots \int_0^{t_{n-1}}\mathrm{d}t_n \cdot \left[S(t-t_1)O_1 S(t_1-t_2)\ldots O_n S(t_n)\right] f_m^0\,. \tag{18}$$

Here O_j is either an operator C or T expressing a creation of a new particle or a recollision between two particles respectively. G_n is a graph namely a sequence of indices

$$(r_1, l_1), (r_2, l_2), \ldots (r_n, l_n)$$

where (r_j, l_j), $r_j < l_j$ are the pair of indices of the particles involved in the interaction at time t_j. The number of particles created in the process is $m-1$. It is convenient to represent the generic graph in the following way, as in Fig. 1.

The legs of the graph denotes the particles and the nodes the creation of new particles (operators C). Recollisions (operators T) are represented by horizontal link. For instance the graph in figure is $(1,2)$, $(1,3)$, $(1,3)$, $(2,3)$, $m = 3$ and the integrand in Eq. (18) is in this case

$$\left[S(t-t_1)C_{1,2}S(t_1-t_2)C_{1,3}S(t_2-t_3)T_{1,3} \cdot S(t_3-t_4)T_{2,3}S(t_4)\right] f_3^0\,. \tag{19}$$

Note that the knowledge of the graph determines completely the sequence of operators in the right hand side of (20).

Finally the factor $K(G_n)$ takes into account the divergences:

$$K(G_n) = O\left(\varepsilon^{-\frac{n}{2}}\varepsilon^{-3(m-1)}\right) . \tag{20}$$

We are not able to analyze the asymptotic behavior of each term of the expansion (18) however we can compute the limit for $\varepsilon \to 0$ of the few terms up to the second order (in time). We have:

$$\begin{aligned} g^N(x_1, v_1, t) &= f^0(x_1 - v_1 t, v_1) + \frac{N-1}{\sqrt{\varepsilon}} \int_0^t S(t-t_1) \\ &\cdot C^{\varepsilon}_{1,2} S(t_1) f^0_2 \mathrm{d}t_1 + \frac{(N-1)}{\varepsilon}\frac{(N-2)}{\varepsilon} \sum_{j=1,2} \int_0^{t_1} \mathrm{d}t_2 \\ &\cdot S(t-t_1) C^{\varepsilon}_{1,2} S(t_1-t_2) C^{\varepsilon}_{j,3} S(t_2) f^0_3 + \frac{N-1}{\varepsilon} \int_0^t \mathrm{d}t_1 \\ &\cdot \int_0^{t_1} \mathrm{d}t_2\, S(t-t_1) C^{\varepsilon}_{1,2} S(t_1-t_2) T^{\varepsilon}_{1,2} S(t_2) f^0_2 . \end{aligned} \tag{21}$$

Here the right hand side of (21) defines g^N.

The second and third term in (21) corresponding to the graphs shown in Fig. 2 are indeed vanishing as follows by the use of the previous arguments.

The most interesting term is the last one, shown in Fig. 3.

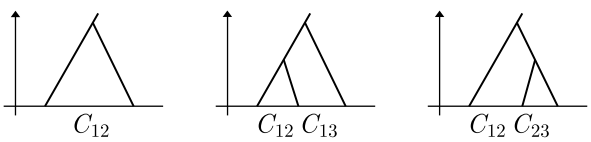

Scaling Limits of Large Systems of Non-linear Partial Differential Equations, Figure 2
The vanishing terms in the expansion of g^N

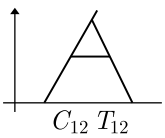

Scaling Limits of Large Systems of Non-linear Partial Differential Equations, Figure 3
The first non vanishing terms in the expansion of g^N

To handle this term we denote by $w = v_1 - v_2$ the relative velocity and note that, for a given function u:

$$\begin{aligned} &S(t_1-t_2) T^{\varepsilon}_{1,2} u(x_1, x_2, v_1, v_2) \\ &= -F\left(\frac{(x_1-x_2) - w(t_1-t_2)}{\varepsilon}\right) \cdot [(\nabla_{v_1} - \nabla_{v_2})u] \\ &\quad \cdot (x_1 - v_1(t_1-t_2), x_2 - v_2(t_1-t_2), v_1, v_2) \\ &= -F\left(\frac{(x_1-x_2) - w(t-t_1)}{\varepsilon}\right) \cdot (\nabla_{v_1} - \nabla_{v_2} + (t_1-t_2) \\ &\quad \cdot (\nabla_{x_1} - \nabla_{x_2})) S(t_1-t_2) u(x_1, x_2, v_1, v_2) . \end{aligned} \tag{22}$$

Therefore the last term in the r.h.s. of (22) is

$$\begin{aligned} &\frac{N-1}{\varepsilon} \int_0^t \mathrm{d}t_1\, S(t-t_1) \int_0^{t_1} \mathrm{d}t_2 \int \mathrm{d}x_2 \int \mathrm{d}v_2 \\ &F\left(\frac{x_1-x_2}{\varepsilon}\right) \cdot \nabla_{v_1} F\left(\frac{(x_1-x_2) - w(t_1-t_2)}{\varepsilon}\right) \\ &\quad \cdot (\nabla_{v_1} - \nabla_{v_2} + (t_1-t_2)(\nabla_{x_1} - \nabla_{x_2})) \\ &\quad \cdot S(t_1) f^0_2(x_1, x_2, v_1, v_2) . \end{aligned} \tag{23}$$

Setting now $r = \frac{x_1-x_2}{\varepsilon}$ and $s = \frac{t_1-t_2}{\varepsilon}$ then

$$\begin{aligned} &g^N(x_1, v_1, t) = (N-1)\varepsilon^3 \int_0^t \mathrm{d}t_1 \int_0^{\frac{t_1}{\varepsilon}} \mathrm{d}s \int \mathrm{d}r \\ &\quad \cdot \int \mathrm{d}v_2\, F(r) \cdot \nabla_{v_1} \\ &F(r - ws) \cdot (\nabla_{v_1} - \nabla_{v_2} + \varepsilon s(\nabla_{x_1} - \nabla_{x_2})) \\ &\quad \cdot S(t_1 - \varepsilon s) f^0_2(x_1, x_2, v_1, v_2) + O(\sqrt{\varepsilon}) . \end{aligned} \tag{24}$$

The formal limit is of (21) is

$$\begin{aligned} g(x_1, v_1, t) = \int_0^t \mathrm{d}t_1 \int \mathrm{d}v_2\, S(t-t_1) \nabla_{v_1} a(v_1 - v_2) \\ \cdot (\nabla_{v_1} - \nabla_{v_2})\, S(t_1) f^0_2 , \end{aligned} \tag{25}$$

where (using $F(r) = -F(-r)$) the matrix a is given by:

$$\begin{aligned} &a(w) = \int \mathrm{d}r \int_0^{+\infty} \mathrm{d}s\, F(r) \otimes F(r - ws) = \frac{1}{2} \int \mathrm{d}r \\ &\cdot \int_{-\infty}^{+\infty} \mathrm{d}s\, F(r) \otimes F(r - ws) \frac{1}{2} \frac{1}{(2\pi)^3} \int_{-\infty}^{+\infty} \mathrm{d}s \\ &\cdot \int \mathrm{d}k\, k \otimes k\, \hat{\phi}(k)^2 \mathrm{e}^{i(w\cdot k)s} = \frac{1}{(8\pi)^2} \int \mathrm{d}k\, k \otimes k\, \hat{\phi}(k)^2 \\ &\cdot \delta(w \cdot k) = \frac{A}{|w|^3}(|w|^2 Id - w \otimes w), \end{aligned} \tag{26}$$

where $\hat{\phi}(k) = \int e^{-il\cdot x}\phi(x)dx$. Here the interaction potential ϕ has been assumed spherically symmetric and:

$$A = \frac{1}{8\pi}\int_0^{+\infty} dr\, r^3\hat{\phi}(r)^2 . \tag{27}$$

Looking at Eq. (25) we are led to introduce the following nonlinear equation:

$$(\partial_t + v\cdot\nabla_x)f = Q_L(f,f) \tag{28}$$

with the collision operator Q_L given by:

$$Q_L(f,f)(v) = \int dv_1 \cdot \nabla_v \Big[a(v-v_1)\,(\nabla_v - \nabla_{v_1})\, f(v)f(v_1)\Big]. \tag{29}$$

Here x plays the role of a parameter and hence his dependence is omitted.

Equation (28) is called the Fokker–Planck–Landau equation (Landau equation in the sequel) and has been introduced by Landau in the study of a dense, weakly interacting gas (see [32]).

From (28) we obtain the following (infinite) hierarchy of equations

$$(\partial_t + V_j\cdot\nabla_{X_j})f_j = C_{j+1}f_{j+1} \tag{30}$$

for the quantities:

$$f_j(t) = f(t)^{\otimes j} \tag{31}$$

where $f(t)$ solves Eq. (28). Accordingly $C_{j+1} = \sum_k C_{k,j+1}$ where

$$\begin{aligned} &C_{k,j+1}f_{j+1}(x_1\ldots x_j, v_1\ldots v_j) \\ &= \prod_{r\neq k} f(x_r,v_r)Q_L(f,f)(x_k,v_k) . \end{aligned} \tag{32}$$

Therefore f has the following series expansion representation

$$\begin{aligned} f(t) = \sum_{n\geq 0}\int_0^t dt_1 \int_0^{t_1} dt_2 \ldots \int_0^{t_{n-1}} dt_n \\ \cdot\big[S(t-t_1)C_2S(t_1-t_2)C_3\ldots C_{n-1}S(t_n)\big]f^0_{n+1} . \end{aligned} \tag{33}$$

The previous calculation shows the formal convergence of g^N to the term with $n = 1$ of the expansion (33), namely we have an agreement between the particle system (18) and the solution to the Landau equation (33) at least up to the first order in time (second order in the potential). Although the above arguments can be made rigorous, under suitable assumption on the initial condition f_0 and the potential ϕ, it seems difficult to show the convergence of the whole series. On the other hand it is clear that the graphs which should contribute in the limit are those formed by a collision-recollision sequence, like that shown in Fig. 4. For those terms it is probably possible to show the convergence. For instance the case in figure has the asymptotics

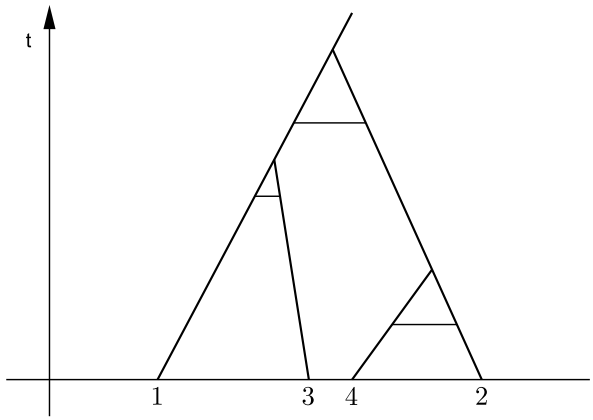

Scaling Limits of Large Systems of Non-linear Partial Differential Equations, Figure 4
The collision-recollision sequence (1, 2), (1, 2), (1, 3), (1, 3), (2, 4), (2, 4)

$$\begin{aligned} \int_0^t dt_1 \int_0^{t_1} dt_2 \int_0^{t_2} dt_3 \big[S(t-t_1)C_{1,2}S(t_1-t_2) \\ \cdot C_{1,3}S(t_2-t_3)C_{2,4}S(t_3)\big]f^0_4 . \end{aligned} \tag{34}$$

However the proof that all other graphs are vanishing in the limit is not easy. Even more difficult is a uniform control of the series expansion (18), even for short times. As we shall see in the next section, something more can be obtained for quantum systems under the same scaling limit.

Some comments are in order.

In the present section we showed how the Landau equation is expected to be derived in the weak-coupling limit from a particle system. In doing this we essentially followed the monograph [2]. This is not however the usual way in which the Landau equation is introduced in the literature. Indeed it is usually recovered from the Boltzmann equation. when the density increases and the grazing collisions become dominant. In particular, the case of the homogeneous Boltzmann equation has been investigated in [1,24,44] (see also [15] for the Coulomb potential case, and see the general survey [43]). In [1] the authors show that, under suitable assumptions on the cross-section, the diffusion Fokker–Planck–Landau equation (28) can in-

deed be derived. The diffusion operator is the form (29) but with a matrix a given by

$$a(w) = \alpha(|w|)(|w|^2 Id - w \otimes w)\,,$$

with α smooth function. Next in [24] and [44] steps forward were performed to arrive to cover the case $\alpha(|w|) \approx 1/|w|^\nu$ for small $|w|$, with $\nu < 1$. We remark that the diffusion coefficient found in this way is different from that derived here in (26).

We conclude by observing that the Landau equation describes a genuine kinetic evolution. Mass, momentum and energy are conserved, while the kinetic entropy is decreasing. The equilibrium is Maxwellian. The difficulties in the validation problem are certainly related to the transition from a reversible (hamiltonian) dynamics to an irreversible one, as for the Boltzmann equation. Here we tried to compare the two series expansions as in the Lanford's validation proof for the Boltzmann equation. This is a sort of Cauchy–Kowalevski brute force argument which works, as well, for negative times. It is clear that, in this way, we cannot go beyond a short time result and even this seems not trivial at all. Other approaches making use of the diffusive nature of the motion of a test particle are, at moment, absent. However the weak-coupling limit of a single particle in a random distribution of scatterers is well understood as we shall discuss later in Sect. "The Weak Coupling Limit for a Single Particle: the Linear Theory".

Weak-Coupling Limit for Quantum Systems

We consider the quantum analog of the system considered in Sect. "Introduction", namely N identical quantum particles with unitary mass in $\mathbb{R}^3$. In the present section the statistical nature of the particles will be ignored.

The interaction between particles is still a two-body potential ϕ so that the total potential energy is taken as

$$U(x_1 \dots x_N) = \sum_{i<j} \phi(x_i - x_j)\,. \tag{35}$$

The associated Schrödinger equation reads

$$i\partial_t \Psi(X_N, t) = -\tfrac{1}{2}\Delta_N \Psi(X_N, t) + U(X_N)\Psi(X_N, t), \tag{36}$$

where $\Delta_N = \sum_{i=1}^N \Delta_i$, Δ_i is the Laplacian with respect to the x_i variables, $X_N = (x_1, \dots, x_N)$ and $\hbar$ is normalized to unity.

As for the classical system considered in Sect." Introduction" we rescale the equation and the potential by

$$x \to \varepsilon x\,, \qquad t \to \varepsilon t\,, \qquad \phi \to \sqrt{\varepsilon}\phi\,. \tag{37}$$

The resulting equation is,

$$i\varepsilon\partial_t \Psi^\varepsilon(X_N, t) = -\frac{\varepsilon^2}{2}\Delta_N \Psi^\varepsilon(X_N, t) + U_\varepsilon(X_N)\,\Psi^\varepsilon(X_N, t), \tag{38}$$

where:

$$U_\varepsilon(x_1 \dots x_N) = \sqrt{\varepsilon} \sum_{i<j} \phi\left(\frac{x_i - x_j}{\varepsilon}\right)\,. \tag{39}$$

We want to analyze the limit $\varepsilon \to 0$ in the above equations, when $N = \varepsilon^{-3}$.

Note that this limit looks, at a first sight, similar to a semiclassical (or high frequency) limit. It is not so: indeed the potential varies on the same scale of the typical oscillations of the wave functions so that the scattering process is a genuine quantum process. Obviously, due to the oscillations, we do not expect that the wave function does converge to something in the limit. The right quantity to look at was introduced by Wigner in 1922 [45] to deal with kinetic problems. It is called the Wigner transform (of Ψ^ε) and is defined as

$$W^N(X_N, V_N) = \left(\frac{1}{2\pi}\right)^{3N} \int \mathrm{d}Y_N\, \mathrm{e}^{iY_N\cdot V_N}\overline{\Psi}^\varepsilon \cdot(X_N + \frac{\varepsilon}{2}Y_N)\,\Psi^\varepsilon(X_N - \frac{\varepsilon}{2}Y_N)\,. \tag{40}$$

The Wigner transform W^N satisfies a transport-like equation, completely equivalent to the Schrödinger equation:

$$(\partial_t + V_N\cdot\nabla_N)W^N(X_N, V_N) = \frac{1}{\sqrt{\varepsilon}}(T_N^\varepsilon W^N)(X_N, V_N). \tag{41}$$

The operator T_N^ε on the right-hand-side of (38) plays the same role of the classical operator denoted with the same symbol in Sect. "Introduction". It is

$$(T_N^\varepsilon W^N)(X_N, V_N) = \sum_{0<k<\ell\le N} (T_{k,\ell}^\varepsilon W^N)(X_N, V_N), \tag{42}$$

where each $T_{k,\ell}^\varepsilon$ describes the interaction of particle k with particle ℓ

$$(T_{k,\ell}^\varepsilon W^N)(X_N, V_N) = \frac{1}{i}\Bigg(\frac{1}{(2\pi)^{3N}}\int \mathrm{d}Y_N \mathrm{d}V_N' \cdot\mathrm{e}^{iY_N\cdot(V_N - V_N')}\left[\phi\left(\frac{x_k - x_\ell}{\varepsilon} - \frac{y_k - y_\ell}{2}\right)\right) - \phi\left(\frac{x_k - x_\ell}{\varepsilon} + \frac{y_k - y_\ell}{2}\right)\Bigg] W^N(X_N, V_N')\,. \tag{43}$$

Equivalently, we may write

$$(T^{\varepsilon}_{k,\ell} W^N)(X_N, V_N) = -i \sum_{\sigma=\pm 1} \sigma \int \frac{\mathrm{d}h}{(2\pi)^3} \hat{\phi}(h) \cdot \mathrm{e}^{i\frac{h}{\varepsilon}(x_k - x_\ell)} W^N(x_1, \ldots, x_N; v_1, \ldots, v_k - \sigma \frac{h}{2}, \ldots, v_\ell + \sigma \frac{h}{2}, \ldots, v_N) . \quad (44)$$

Note that $T^{\varepsilon}_{k,\ell}$ is a pseudodifferential operator which formally converge, at fixed ε, for $\hbar \to 0$ (here $\hbar = 1$) to its classical analog given by Eq. (7). Note also that in (44), "collisions" may take place between *distant* particles ($x_k \neq x_\ell$). However, such distant collisions are penalized by the highly oscillatory factor $\exp(ih(x_k - x_\ell)/\varepsilon)$. These oscillations turn out to play a crucial role throughout the analysis, and they explain why collisions tend to happen when $x_k = x_\ell$ in the limit $\varepsilon \to 0$.

The formalism we have introduced is formally similar to the classical case so that we proceed as before by transforming Eq. (38) into a hierarchy of equations. We introduce the partial traces of the Wigner transform W^N, denoted by f^N_j. They are defined through the following formula, valid for $j = 1, \ldots, N-1$:

$$f^N_j(X_j, V_j) = \int \mathrm{d}x_{j+1} \ldots \int \mathrm{d}x_N \int \mathrm{d}v_{j+1} \ldots \int \mathrm{d}v_N \, W^N(X_j, x_{j+1} \ldots x_N, V_j, v_{j+1} \ldots v_N) . \quad (45)$$

Obviously, we set $f^N_N = W^N$. The function f^N_j is the kinetic object that describes the state of the j particles subsystem at time t.

As before, the wave function Ψ, as well W^N and f^N_j, are assumed to be symmetric in the exchange of particle, a property that is preserved in time.

Proceeding then as in the derivation of the BBKGY hierarchy for classical systems, we readily transform Eq. (38) into the following hierarchy:

$$(\partial_t + \sum_{k=1}^{j} v_k \cdot \nabla_k) f^N_j(X_j, V_j) = \frac{1}{\sqrt{\varepsilon}} T^{\varepsilon}_j f^N_j + \frac{N-j}{\sqrt{\varepsilon}} C^{\varepsilon}_{j+1} f^N_{j+1} , \quad (46)$$

where

$$C^{\varepsilon}_{j+1} = \sum_{k=1}^{j} C^{\varepsilon}_{k,j+1} , \quad (47)$$

and $C^{\varepsilon}_{k,j+1}$ is defined by

$$C^{\varepsilon}_{k,j+1} f^N_{j+1}(X_j, V_j) = -i \sum_{\sigma=\pm 1} \sigma \int \frac{\mathrm{d}h}{(2\pi)^3} \int \mathrm{d}x_{j+1} \cdot \int \mathrm{d}v_{j+1} \, \hat{\phi}(h) \, \mathrm{e}^{i\frac{h}{\varepsilon}(x_k - x_{j+1})} f^N_{j+1}\Big(x_1, x_2, \ldots, x_{j+1}, v_1, \ldots, v_k - \sigma \frac{h}{2}, \ldots, v_{j+1} + \sigma \frac{h}{2}\Big) . \quad (48)$$

As before the initial value $\{f^0_j\}_{j=1}^N$ is assumed completely factorized: for all $j = 1, \ldots, N$, we suppose

$$f^0_j = f_0^{\otimes j} , \quad (49)$$

where f_0 is a one-particle Wigner function, and f^0 is assumed to be a probability distribution.

In the limit $\varepsilon \to 0$, we expect that the j-particle distribution function $f^N_j(t)$, that solves the hierarchy (46) with initial data (49), tends to be factorized for all times: $f^N_j(t) \sim f(t)^{\otimes j}$ (propagation of chaos).

As for the classical case, if f_{j+1} is smooth:

$$C^{\varepsilon}_{k,j+1} f^N_{j+1}(X_j, V_j) = -i\varepsilon^3 \sum_{\sigma=\pm 1} \sigma \int \frac{\mathrm{d}h}{(2\pi)^3} \hat{\phi}(h) \cdot \int \mathrm{d}r \int \mathrm{d}v_{j+1} \; \mathrm{e}^{ih\cdot r} f^N_{j+1}\Big(X_j, x_k - \varepsilon r, \; v_1, \ldots, v_k - \sigma \frac{h}{2}, \ldots, v_{j+1} + \sigma \frac{h}{2}\Big) = O(\varepsilon^4) . \quad (50)$$

Indeed, setting $\varepsilon = 0$ in the integrand, the integration over r produces $\delta(h)$. As a consequence the integrand is independent on σ and the sum vanishes. Therefore the integral is $O(\varepsilon)$. Also

$$\frac{1}{\sqrt{\varepsilon}} \int_0^t \mathrm{d}t_1 \, S(t - t_1) T_{r,k} f^N_j(t_1) = -i \sum_{\sigma=\pm 1} \sigma \int_0^t \mathrm{d}t_1 \frac{\mathrm{d}h}{(2\pi)^3} \hat{\phi}(h) \mathrm{e}^{i\frac{h}{\varepsilon}\cdot(x_r - x_k) - (v_r - v_k)(t - t_1)} \cdot f^N_j(X_j - V_j(t - t_1), V_j, t_1) \quad (51)$$

is weakly vanishing, by a stationary phase argument (see [4]). Therefore, as for the classical case, we analyze

the asymptotics of the collision-recollision term (shown in Fig. 1):

$$-\frac{N-1}{\varepsilon}\int_0^t \mathrm{d}t_1 \int_0^{t_1} \mathrm{d}\tau_1 \cdot S(t-t_1)C_{1,2}S(t_1-\tau_1)T_{1,2}S(\tau_1)f_2^0. \tag{52}$$

Explicitly it looks as:

$$\begin{aligned}
&-\frac{N-1}{\varepsilon}\sum_{\sigma,\sigma'=\pm1}\sigma\sigma'\int_0^t \mathrm{d}t_1\int_0^{t_1}\mathrm{d}\tau_1\int \mathrm{d}x_2\int \mathrm{d}v_2\\
&\cdot\int\frac{\mathrm{d}h}{(2\pi)^3}\int\frac{\mathrm{d}k}{(2\pi)^3}\hat\phi(h)\,\hat\phi(k)\,\mathrm{e}^{i\frac{h}{\varepsilon}\cdot\left(x_1-x_2-v_1(t-t_1)\right)}\\
&\cdot\mathrm{e}^{i\frac{k}{\varepsilon}\cdot\left(x_1-x_2-v_1(t-t_1)-(v_1-v_2-\sigma h)(t_1-\tau_1)\right)}f_2^0\Big(x_1-v_1t\\
&\qquad+\sigma\frac{h}{2}t_1+\sigma'\frac{k}{2}\tau_1,x_2-v_2t_1-\sigma\frac{h}{2}t_1-\sigma'\frac{k}{2}\tau_1;\\
&\qquad\qquad v_1-\sigma\frac{h}{2}-\sigma'\frac{k}{2},v_2+\sigma\frac{h}{2}+\sigma'\frac{k}{2}\Big).
\end{aligned} \tag{53}$$

By the change of variables:

$$t_1-\tau_1=\varepsilon s_1\,,\ (\text{i.e.}\ \tau_1=t_1-\varepsilon s_1)\,,\ \xi=(h+k)/\varepsilon\,, \tag{54}$$

we have

$$\begin{aligned}
(54)=&-(N-1)\,\varepsilon^3\sum_{\sigma,\sigma'=\pm1}\sigma\sigma'\int_0^t\mathrm{d}t_1\int_0^{t_1/\varepsilon}\mathrm{d}s_1\\
&\cdot\int\mathrm{d}x_2\,\mathrm{d}v_2\,\frac{\mathrm{d}\xi}{(2\pi)^3}\,\frac{\mathrm{d}k}{(2\pi)^3}\,\hat\phi(-k+\varepsilon\xi_1)\,\hat\phi(k)\\
&\cdot\mathrm{e}^{i\xi\cdot\left(x_1-x_2-v_1(t-t_1)\right)}\mathrm{e}^{-is_1k\cdot(v_1-v_2-\sigma(-k+\varepsilon\xi))}f_{j+1}^0(\dots)\,,
\end{aligned} \tag{55}$$

In the limit $\varepsilon\to0$, the above formula gives the asymptotics

$$\begin{aligned}
(52)\underset{\varepsilon\to0}{\sim}&-\sum_{\sigma,\sigma'=\pm1}\sigma\sigma'\int_0^t\mathrm{d}t_1\int\mathrm{d}v_2\,\frac{\mathrm{d}k}{(2\pi)^3}\,|\hat\phi(k)|^2\\
&\cdot\left(\int_0^{+\infty}\mathrm{e}^{-is_1k\cdot(v_1-v_2+\sigma k)}\mathrm{d}s_1\right)\\
&\cdot f_2^0\Big(x_1-v_1t-(\sigma-\sigma')\frac{k}{2}t_1,\\
&x_1-v_1(t-t_1)-v_2t_1+(\sigma-\sigma')\frac{k}{2}t_1,v_1+(\sigma-\sigma')\frac{k}{2},\\
&\qquad v_2-(\sigma-\sigma')\frac{k}{2}\Big).
\end{aligned} \tag{56}$$

In [4], we completely justify formula (56) and its forthcoming consequences.

Now, we turn to identifying the limiting value obtained in (56). To do so, we observe that symmetry arguments allow us to replace the integral in s by its real part:

$$\mathrm{Re}\int_0^\infty \mathrm{e}^{-is_1k\cdot(v_1-v_2+\sigma k)}\mathrm{d}s_1=\pi\delta(k\cdot(v_1-v_2+\sigma k)). \tag{57}$$

Using this we realize that the contribution $\sigma=-\sigma'$ in (56) gives rise to the gain term:

$$\begin{aligned}
\int_0^t\mathrm{d}t_1\int\mathrm{d}v_2\int\mathrm{d}\omega\,B(\omega,v_1-v_2)\,f_2^0\big(x_1-v_1(t-t_1)\\
-v_1't_1,x_2-v_2(t-t_1)-v_2't_1,v_1',v_2'\big)\,,
\end{aligned} \tag{58}$$

where the integral in ω is on the surface of the unitary sphere in $\mathbb{R}^3$, and

$$B(\omega,v)=\frac{1}{8\pi^2}|\omega\cdot v|\,|\hat\phi(\omega\,(\omega\cdot v))|^2, \tag{59}$$

and the velocities v_1', v_2' are

$$v_1'=v_1-\omega(\omega\cdot(v_2-v_1)\,,\quad v_2'=v_2+\omega(\omega\cdot(v_2-v_1)\,.$$

Similarly, the term $\sigma=\sigma'$ in (5) yields the loss term:

$$\begin{aligned}
&\int_0^t\mathrm{d}t_1\int\mathrm{d}v_2\int\mathrm{d}\omega\,B(\omega,v_1-v_2)\\
&\cdot f_2^0(x_1-v_1t,x_2-v_2(t-t),v_1,v_2)\,.
\end{aligned} \tag{60}$$

By the same arguments used in the previous section we can conclude that the full series expansion (20) (of course for the present quantum case) agrees, up to the second order in the potential, with

$$S(t)f_0+\int_0^t\mathrm{d}t_1\,S(t-t_1)Q(S(t_1)f_0,S(t_1)f_0) \tag{61}$$

where

$$\begin{aligned}
Q(f,f)=\frac{1}{4\pi^2}\int\mathrm{d}v_1\int\mathrm{d}h\,|\hat\phi(h)|^2\delta((h\cdot(v-v_1+h))\\
\cdot[f(v+h)f(v_1-h)-f(v)f(v_1)]=\int\mathrm{d}v_1\int\mathrm{d}\omega\\
\cdot B(\omega,v-v_1)[f'f_1'-ff_1]\,,
\end{aligned} \tag{62}$$

with

$$\begin{aligned}
&f_1=f(v_1)\,,\ f'=f(v')\,,\ f_1'=f(v_1')\,,\\
&v'=v+h=v-\omega(\omega\cdot(v-v_1))\,,\\
&v_1'=v_1-h=v_1+\omega(\omega\cdot(v-v_1))\,.
\end{aligned} \tag{63}$$

In other words the kinetic equation which comes out is the Boltzmann equation with cross-section B:

$$(\partial_t + v \cdot \nabla_x) f = Q(f, f)\,, \tag{64}$$

where Q is given by Eq. (62). We note once more that the δ function in Eq. (62) expresses the energy conservation, while the momentum conservation is automatically satisfied. Note also that the cross-section B is the only quantum factor in the purely classical expression (62). It retains the quantum features of the elementary "collisions".

An important comment is in order. Why the kinetic equation for quantum systems is of Boltzmann type in contrast with the classical case where we got a diffusion? The answer is related to the asymptotics of a single scattering (see [36,37] and [8]). For quantum systems we have a finite probability of having any angle scattering, while for a classical particle, we surely have a small deviation from the free motion. Therefore a quantum particle, in this asymptotic regime, is going to perform a jump process (in velocity) rather than a diffusion.

From a mathematical view point we observe that [4] proves more than agreement up to second order. We indeed consider the subseries (of the full series expansion expressing $f_j^N(t)$) formed by *all* the collision–recollision terms (as that shown in Fig. 3). In other words, we consider the subseries of $f_j^N(t)$ given by

$$\sum_{n\geq 1} \sum_{r_1,\ldots,r_n,l_1,\ldots,l_n} \varepsilon^{-4n} \int_0^t \mathrm{d}t_1 \int_0^{t_1} \mathrm{d}\tau_1\, S(t-t_1) C^{\varepsilon}_{r_1,l_1} \cdot S(t_1-\tau_1) T^{\varepsilon}_{r_1,l_1} \int_0^{\tau_{n-1}} \mathrm{d}t_n \int_0^{t_n} \mathrm{d}\tau_n\, S(\tau_{n-1}-t_n) C^{\varepsilon}_{r_n,l_n} \cdot S(t_n-\tau_n) T^{\varepsilon}_{r_n,l_n} S(\tau_n) f^0_{j+n+1}\,. \tag{65}$$

Here the sum runs over all possible choices of the particles number r's and l's, namely we sum over the subset of graph of the form in Fig. 4. We establish in [4] that the subseries (65) is indeed absolutely convergent, for short times, uniformly in ε. Moreover, we prove that it approaches the corresponding complete series expansion obtained by solving iteratively the Boltzmann equation with collision operator given by Eq. (62) extending and making rigorous the above argument.

Under reasonable smoothness hypotheses on the potential and on the initial distribution, assuming in addition that $\hat{\phi}(0) = 0$, we also proved in [7], that all other terms than those considered in the subseries (65) are indeed vanishing in the limit. The condition $\hat{\phi}(0) = 0$ is probably only technical: it takes care automatically of some divergences which are difficult to deal with differently.

Unfortunately this result, even under these severe assumptions on the potential, is not yet conclusive because we are not able to show a uniform bound on the full series. Thus a mathematical justification of the quantum Boltzmann equation is a still an open and difficult problem.

It is not surprising that we know more for the quantum case in comparison with the classical problem introduced in Sect. "Introduction". Now the operators involved are differences while, for the classical case, we have to deal with derivatives. Introducing explicitly the $\hbar$ dependence, we easily realize that the results discussed in the present section are not uniform in $\hbar$.

Weak-Coupling Limit in the Bose–Einstein and Fermi–Dirac Statistics

In this section we approach the same problem as in Sect. "Weak-Coupling Limit for Classical Systems", for Bosons or Fermions, namely for particles obeying the Fermi–Dirac or Bose–Einstein statistics. As we shall see, the kinetic equation we expect to hold in the limit, is the so called Uehling and Uhlembeck equation, which is a Boltzmann type equation, with cubic collision operator. We note that the effects of the quantum correction here, enter also in the structure of the operator.

In this case, the starting point is still the rescaled Schrödinger equation (38), or the equivalent hierarchy (46). The only new point is that we cannot take a totally decorrelated initial datum as in (49). Indeed, the Fermi–Dirac or Bose–Einstein statistics yield correlations even at time zero. In this perspective, the most uncorrelated states one can introduce, and that do not violate the Fermi–Dirac or Bose–Einstein statistics, are the so called quasi-free states. They have, in terms of the Wigner formalism, the following form

$$f_j(x_1, v_1, \ldots, x_j, v_j) = \sum_{\pi \in \mathcal{P}_j} \theta^{s(\pi)} f_j^{\pi}(x_1, v_1, \ldots, x_j, v_j)\,, \tag{66}$$

where each f_j^{π} has the value

$$f_j^{\pi}(x_1, v_1, \ldots, x_j, v_j) = \int \mathrm{d}y_1 \ldots \mathrm{d}y_j \int \mathrm{d}w_1 \ldots \mathrm{d}w_j \cdot \mathrm{e}^{i(y_1\cdot v_1 + \cdots + y_j \cdot v_j)} \prod_{k=1}^{j} \mathrm{e}^{-\frac{i}{\varepsilon} w_k \cdot (x_k - x_{\pi(k)})} \cdot \mathrm{e}^{-\frac{i}{2} w_k \cdot (y_k + y_{\pi(k)})} f\left(\frac{x_k + x_{\pi(k)}}{2} + \varepsilon \frac{y_k - y_{\pi(k)}}{4}, w_k\right), \tag{67}$$

and f is a given one-particle Wigner function. Here $\mathcal{P}_j$ denotes the group of all the permutations of j objects and π its generic element.

Note that the uncorrelated case treated in Sect. "Weak-Coupling Limit for Classical Systems", is recovered by the contribution due the permutation $\pi =$ identity.

To verify that (66) and (67) really describe admissible states for the statistics, consider the inverse Wigner transform of (66) and (67). The density matrix obtained in this way does verify the permutation invariance required by the statistics.

Moreover the quasi-free states converge weakly to the completely factorized states as $\varepsilon \to 0$. This is physically obvious because the quantum statistics become irrelevant in the semiclassical limit. However the dynamics take place on the scale ε so that the effects of the statistics are present in the limit. Indeed it is expected that the one-particle distribution function $f_1^N(t)$ converges to the solution of the following cubic Boltzmann equation:

$$(\partial_t + v \cdot \nabla_x) f(x, v, t) = Q_\theta(f)(x, v, t) , \tag{68}$$

$$\begin{aligned} Q_\theta(f)(x, v, t) &= \int \mathrm{d}v_1 \, \mathrm{d}\omega \, B_\theta(\omega, v - v_1) \\ &\cdot \big[f(x, v') f(x, v_1')(1 + 8\pi^3 \theta f(x, v))(1 + 8\pi^3 \theta f(x, v_1)) \\ &- f(x, v) f(x, v_1)(1 + 8\pi^3 \theta f(x, v')(1 + 8\pi^3 \theta f(x, v_1'))\big] , \end{aligned} \tag{69}$$

(for the notations see Eq. (63)). Here $\theta = +1$ or $\theta = -1$, for the Bose–Einstein or the Fermi–Dirac statistics respectively. Finally, B_θ is the symmetrized or antisymmetrized cross-section derived from B (see (59)) in a natural way:

$$\begin{aligned} B_\theta(\omega, v) = \frac{1}{16\pi^2} |\omega \cdot v| \\ \cdot \big[\hat{\phi}(\omega \, (\omega \cdot v)) + \theta \hat{\phi}(v - \omega \, (\omega \cdot v)) \big]^2 . \end{aligned}$$

As we see, the modification of the statistics transforms the quadratic Boltzmann equation of the Maxwell–Boltzmann case, into a cubic one (fourth order terms cancel). Also, the statistics affects the form of the cross-section and B has to be (anti)symmetrized into B_θ.

The collision operator (69) has been introduced by Nordheim in 1928 [38] and by Uehling–Uhlenbeck in 1933 on the basis of purely phenomenological considerations [42].

Plugging in the hierarchy (46) an initial datum satisfying (66), we can follow the same procedure as for the Maxwell–Boltzmann statistics, namely we write the full perturbative series expansion expressing $f_j^N(t)$ in terms of the initial datum and try to analyze its asymptotic behavior.

As we did before, we first restrict our attention to those terms of degree less than two in the potential.

The analysis up to second order is performed in [5]. We actually recover here Eq. (68), (69) with the suitable B_θ. Now the number of terms to control is much larger due to the sum over all permutations that enters the definition (66) of the initial state. Also, the asymptotics is much more delicate. In particular, we stress the fact that the initial datum brings its own highly oscillatory factors in the process, contrary to the Maxwell–Boltzmann case where the initial datum is uniformly smooth, and where the oscillatory factors simply come from the collision operators T and C.

In [5] we consider the graphs of second order in $\hat{\phi}$, show in Fig. 5, which, because of the permutation of initial state, yields various terms: two of them are bilinear in the initial condition f_0 ($C_{12}T_{12}$), and give in the limit the part of Q_θ quadratic in f, twelve are cubic in $f_0(C_{12}C_{13}, C_{12}C_{23})$. Due to a non-stationary phase argument the two terms with permutation $\pi = id = (123)$ vanish, as that the term with $\pi = (321)$ for $C_{12}C_{13}$, and the term with $\pi = (132)$ for $C_{12}C_{23}$. The two terms with $\pi = (321)$ give rise to truly diverging contributions (negative powers of ε), however their sum is seen to cancel. Last, permutations without fixed point ($\pi = (312), (231)$) and $\pi = (132)$ for $C_{12}C_{13}$, and $\pi = (321)$ for $C_{12}C_{23}$, give the part of Q_θ cubic in f. This ends up the analysis of terms up to second order in the potential.

The computation is heavy and hence we address the reader to [5] for the details.

We mention that a similar second order analysis, using commutator expansions in the framework of the second quantization formalism, has been performed in [27] (following [26]) in the case of the van Hove limit for lattice systems (that is the same as the weak-coupling limit, yet without rescaling the distances). For more recent formal results in this direction, but in the context of the weak-coupling limit, we also quote [21].

We finally observe that the initial value problem for Eq. (68) is somehow trivial for Fermions. Indeed we have the a priori bounds $f \leq 1/(8\pi)^3$ making everything easy. For Bosons the situation is much more involved even for the spatially homogeneous case. The statistics favor large value of f and it is not clear whether the equation can explain dynamical condensation. See, for the mathematical side, [33,34].

Under the weak coupling limit we derived two very different equations, the Fokker–Planck–Landau equation in the classical case, and a Boltzmann equation with a quantum cross section in the quantum case. For the original large particle system a classical or a quantum description

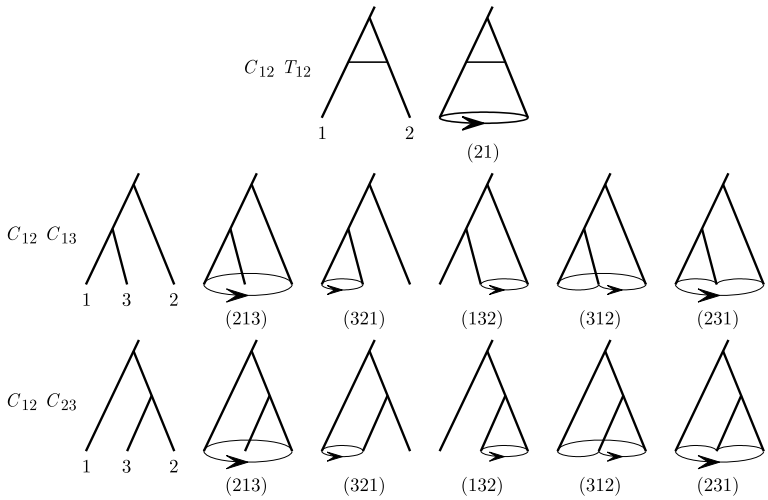

Scaling Limits of Large Systems of Non-linear Partial Differential Equations, Figure 5
The three graphs of second order with the statistical permutations

depends on the value of the physical scales of the macroscopic variables with respect to $\hbar$. The same happen for the two asymptotic equations. The dependence on $\hbar$ of the collision term (69) is given by

$$Q_{UU}(f)(v) = \frac{1}{16\pi^2\hbar^4}\int \mathrm{d}v_1 \int \mathrm{d}\omega \left[\hat{\phi}\left(\frac{\omega\,(\omega\cdot w)}{\hbar}\right) + \theta\hat{\phi}\left(\frac{w-\omega\,(\omega\cdot w)}{\hbar}\right)\right]^2 \{(1\pm(2\pi\hbar)^3 f)(1\pm(2\pi\hbar)^3 f_1) \cdot f' f_1' - (1\pm(2\pi\hbar)^3 f')(1\pm(2\pi\hbar)^3 f_1') f f_1\}\,. \tag{70}$$

The quantum-cross section term $\hat{\phi}(\omega\,(\omega\cdot(v-v_1))\,/\,\hbar)$ make $\omega\cdot(v-v_1) = O(\hbar)$, so that the collision operator concentrates on grazing collisions. In this sense, the classical limit $\hbar \to 0$ for (70) is a natural grazing collision limit. More formally, the operator (70) tends to the operator Q_L given in Eq. (29) (see [9]); nevertheless no results are available for the limit of the solutions of the equations.

We conclude this section with some consideration on the low-density limit.

In the classical case the low-density limit (or the Boltzmann–Grad limit) yields the usual Boltzmann equation for classical systems and this result has been proved for short times [31]. It is natural to investigate what happens, in the same scaling limit, to a quantum system. Here, due to the fact that the density is vanishing, the particles are too rare to make the statistical correlations effective. As a consequence, we expect that the Bose–Einstein, and Fermi–Dirac statistics, as well fully uncorrelated states, all give rise to the same Boltzmann equation along the low-density limit with the true quantum cross-section, given by the sum of the Born series, under smallness assumption for the potential. The analysis of the partial series of the dominant terms (uniform bounds and convergence as for the weak-coupling limit) has been performed in [6].

Weak-Coupling Limit for a Single Particle: The Linear Theory

Consider the time evolution of a single particle, in the low-density or in the weak-coupling regime, under the action of a random configuration of obstacles $\mathbf{c} = \{c_1 \dots c_N\} \subset \mathbb{R}^{3N}$.

More precisely, after the rescaling, the basic equations are:

$$\dot{x}(t) = v(t)\,, \quad \dot{v}(t) = -\sum_j \nabla\phi_\varepsilon(x(t)-c_j) \tag{71}$$

for a classical particle and, for a quantum particle,

$$i\varepsilon\partial_t\psi = -\frac{\varepsilon^2}{2}\Delta\psi + \sum_j \phi_\varepsilon(x(t)-c_j)\psi\,, \tag{72}$$

where $\phi_\varepsilon = \phi(\frac{x}{\varepsilon})$ and $\phi_\varepsilon = \sqrt{\varepsilon}\phi(\frac{x}{\varepsilon})$ for the low-density and weak-coupling limits respectively. Here ϕ denotes a given smooth potential.

We are interested in the behavior of

$$f_\varepsilon(x,v,t) = \mathbb{E}[f_{\mathbf{c}}(x,v,t)] \tag{73}$$

where $f_{\mathbf{c}}(t)$ is the time evolved classical distribution function or the Wigner transform of ψ according to Eq. (71)

or (72) respectively. Finally $\mathbb{E}$ denotes the expectation with respect to the obstacle distribution, for which, a natural choice could be the Poisson distribution of density μ_ε which is also scaled according to

$$\mu_\varepsilon = \varepsilon^{-2}\,, \quad \mu_\varepsilon = \varepsilon^{-3}$$

for the low-density and weak-coupling respectively.

For the low-density scaling (this is the so called Lorentz model), we obtain, for classical systems, a linear Boltzmann equation (see [3,13,17,23,40]). It is also known that the system does not homogenize to a jump process given by a linear Boltzmann equation in case of a periodic distribution of obstacles [10]. For recent results in this direction see [11].

For the weak-coupling limit of a classical particle we obtain a linear Landau equation as it is shown in [29] and [16].

Let us spend a few words on these results. Each solution $(x(t), y(t)) = (x_{\mathbf{c}}(t), v_{\mathbf{c}}(t))$ of Eq. (71) is a sample of a stochastic process. Looking at the behavior of $v_{\mathbf{c}}(t)$ we suddenly realize that it does not enjoy the Markov property because, once an obstacle produces a change in velocity, we know its existence and this creates time correlations with the future behavior of the particle. However if a recollision with the same obstacle is an unlikely event, we could consider the velocity change due to the collisions, as independent and Poisson distributed (as the obstacles) random variables. In this case it is not difficult to prove the convergence to a diffusion process, solution to the following stochastic differential equation

$$\begin{aligned} \mathrm{d}x &= v\mathrm{d}t\,, \\ \mathrm{d}v &= -2A\tfrac{v}{|v|^3}\mathrm{d}t + \sqrt{\tfrac{2A}{|v|}}\left(I - \tfrac{v\otimes v}{|v|^2}\right)\mathrm{d}w\,, \end{aligned} \tag{74}$$

where $\mathrm{d}w$ is a brownian motion, A is the constant given by Eq. (27), and $\sqrt{\frac{2A}{|v|}}\cdot\left(I - v\otimes v/|v|^2\right)$ is the matrix $\sqrt{2a}$, with a given by Eq. (26). Notice that each collision preserves the energy so that the above process lives on the surface of the sphere $|v| = \text{const}$.

Therefore the main point in the proof of the weak-coupling limit is to show that, in the same limit, the process under consideration is stochastically equivalent to a Markov process with the right properties. Moreover, as a consequence, we have also that $f_\varepsilon(t) \to f(t)$ being f solution to the following linear Landau equation

$$(\partial_t + v\cdot\nabla_x)\,f = \mathrm{div}_v[a(v)\nabla_v f]\,, \tag{75}$$

Alternatively we could proceed differently, in the same (Cauchy–Kovalevskii) spirit of our previous analysis for the nonlinear case. Namely we first represent the time evolved distribution $f_{\mathbf{c}}(t)$ in terms of a series expansion. Then, taking the expectation of each term of the series, we exploit its limit by using the same arguments as in Sect. "Introduction". Finally the convergence of the series (uniformly in ε) should also be proven. This is certainly a possible program which is, however, conceptually and technically weaker than that based on the control of the particle trajectory illustrated above. Indeed the absolute control of the series expansion yielding the solution to Eq. (75) does not use the positivity of the diffusion coefficient so that, at best, we could recover the result for a short time only and in spaces of analytic functions, not at all natural in our context.

This remark applies as well to the nonlinear case for which, however, we do not have any result, even for short times.

As regards the corresponding weak-coupling quantum problem, the easiest case is when ϕ is a Gaussian process. The kinetic equation is still a linear Boltzmann equation. The first result, holding for short times, has been obtained in [41] (see also [30]). More recently this result has been extended to arbitrary times [22]. The technique of [22] can be also applied to deal with a Poisson distribution of obstacles [12] Obviously the cross section appearing in the Boltzmann equation is the one computed in the Born approximation. Finally in [20] the low-density case has been successfully approached. The result is a linear Boltzmann equation with the full cross-section.

Future Directions

The program of obtaining macroscopic equations from the microscopic dynamics is an ambitious and difficult problem. It arised in 1900 with the Hilbert's speech at the congress of mathematicians in Paris, as the sixth problem. After more than hundred years, such a program is far to be achieved.

It is probably true that new ideas and techniques are needed. As we said before in discussing the linear problems, it may be quite possible that classical ideas in the field of partial differential equations are not enough to approach this kind of problems successfully. On the other hand a deeper and more detailed analysis of the particle dynamics is technically difficult and philosophically paradoxical. Indeed from the practical side we introduce a partial differential equation to reduce the difficulties of the study of particle evolution, while its justification may require a deeper control of the underlying microscopic dynamics.

Bibliography

Primary Literature

1. Arseniev AA, Buryak OE (1990) On a connection between the solution of the Boltzmann equation and the solution of the Landau–Fokker–Planck equation (Russian). Mat Sb 181(4):435–446 (translation in Math USSR-Sb 69(2):465–478 1991)
2. Balescu R (1975) Equilibrium and Nonequilibrium Statistical Mechanics. Wiley, New York
3. Boldrighini C, Bunimovich LA, Ya Sinai G (1983) On the Boltzmann equation for nthe Lorentz gas. J Stat Phys 32:477–501
4. Benedetto D, Castella F, Esposito R, Pulvirenti M (2004) Some Considerations on the derivation of the nonlinear Quantum Boltzmann Equation. J Stat Phys 116(114):381–410
5. Benedetto D, Castella F, Esposito R, Pulvirenti M (2005) On The Weak–Coupling Limit for Bosons and Fermions. Math Mod Meth Appl Sci 15(12):1–33
6. Benedetto D, Castella F, Esposito R, Pulvirenti M (2006) Some Considerations on the derivation of the nonlinear Quantum Boltzmann Equation II: the low-density regime. J Stat Phys 124(2–4):951–996
7. Benedetto D, Castella F, Esposito R, Pulvirenti M (2008) From the N-body Schrödinger equation to the quantum Boltzmann equation: a term-by-term convergence result in the weak coupling regime. Commun Math Phys 277(1):1–44
8. Benedetto D, Esposito R, Pulvirenti M (2004) Asymptotic analysis of quantum scattering under mesoscopic scaling. Asymptot Anal 40(2):163–187
9. Benedetto D, Pulvirenti M (2007) The classical limit for the Uehling–Uhlenbeck operator. Bull Inst Math Acad Sinica 2(4):907–920
10. Burgain J, Golse F, Wennberg B (1998) On the distribution of free path lenght for the periodic Lorentz gas. Comm Math Phys 190:491–508
11. Caglioti E, Golse F (2003) On the distribution of free path lengths for the periodic Lorentz gas. III Comm Math Phys 236(2):199–221
12. Chen T (2005) Localization lengths and Boltzmann limit for the Anderson model at small disorders in dimension 3. J Stat Phys 120(1–2):279–337
13. Caglioti E, Pulvirenti M, Ricci V (2000) Derivation of a linear Boltzmann equation for a periodic Lorentz gas. Mark Proc Rel Fields 3:265–285
14. Cercignani C, Illner R, Pulvirenti M (1994) The mathematical theory of dilute gases, Applied Mathematical Sciences, vol 106. Springer, New York
15. Degond P, Lucquin–Desreux B (1992) The Fokker–Planck asymptotics of the Boltzmann collision operator in the Coulomb case. Math Models Methods Appl Sci 2(2):167–182
16. Dürr D, Goldstain S, Lebowitz JL (1987) Asymptotic motion of a classical particle in a random potential in two dimension: Landau model. Comm Math Phys 113:209–230
17. Desvillettes L, Pulvirenti M (1999) The linear Boltzmann eqaution for long-range forces: a derivation for nparticle systems. Math Moduls Methods Appl Sci 9:1123–1145
18. Di Perna RJ, Lions PL (1989) On the Cauchy problem for the Boltzmann equatioin. Ann Math 130:321–366
19. Esposito R, Pulvirenti M (2004) From particles to fluids. In: Friedlander S, Serre D (eds) Handbook of Mathematical Fluid Dynamics, vol 3. Elsevier, North Holland, pp 1–83
20. Eng D, Erdös L (2005) The linear Boltzmann equation as the low-density limit of a random Schrödinger equation. Rev Math Phys 17(6):669–743
21. Erdös L, Salmhofer M, Yau HT (2004) On the quantum Boltzmann equation. J Stat Phys 116:367–380
22. Erdös L, Yau HT (2000) Linear Boltzmann equation as a weak-coupling limit of a random Schrödinger equation. Comm Pure Appl Math 53:667–735
23. Gallavotti G (1972) Rigorous theory of the Boltzmann equation in the Lorentz gas in Meccanica Statistica. reprint Quaderni CNR 50:191–204
24. Goudon T (1997) On Boltzmann equations and Fokker–Planck asymptotics: influence of grazing collisions. J Statist Phys 89(3–4):751–776
25. Grad H (1949) On the kinetic Theory of rarefied gases. Comm Pure Appl Math 2:331–407
26. Hugenholtz MN (1983) Derivation of the Boltzmann equation for a Fermi gas. J Stat Phys 32:231–254
27. Ho NT, Landau LJ (1997) Fermi gas in a lattice in the van Hove limit. J Stat Phys 87:821–845
28. Illner R, Pulvirenti M (1986) Global Validity of the Boltzmann equation for a two-dimensional rare gas in the vacuum. Comm Math Phys 105:189–203 (Erratum and improved result, Comm Math Phys 121:143–146)
29. Kesten H, Papanicolaou G (1981) A limit theorem for stochastic acceleration. Comm Math Phys 78:19–31
30. Landau LJ (1994) Observation of quantum particles on a large space-time scale. J Stat Phys 77:259–309
31. Lanford III O (1975) Time evolution of large classical systems. In: Moser EJ (ed) Lecture Notes in Physics 38. Springer, pp 1–111
32. Lifshitz EM, Pitaevskii LP (1981) Course of theoretical physics "Landau–Lifshit", vol 10. Pergamon Press, Oxford-Elmsford
33. Xuguang LU (2005) The Boltzmann equation for Bose–Einstein particles: velocity concentration and convergence to equilibrium. J Stat Phys 119(5–6):1027–1067
34. Xuguang LU (2004) On isotropic distributional solutions to the Boltzmann equation for Bose–Einstein particles. J Stat Phys 116(5–6):1597–1649
35. CB Jr Morrey (1955) On the derivation of the equations of hydrodynamics from statistical mechanics. Comm Pure Appl Math 8:279–326
36. Nier F (1996) A semi-classical picture of quantum scattering. Ann Sci Ec Norm Sup 29(4):149–183
37. Nier F (1995) Asymptotic analysis of a scaled Wigner equation and quantum scattering. Transp Theory Statist Phys 24(4–5):591–628
38. Nordheim LW (1928) On the Kinetic Method in the New Statistics and Its Application in the Electron Theory of Conductivity. Proc Royal Soc Lond Ser A 119(783):689–698
39. Spohn H (1991) Large scale dynamics of interacting particles, Texts and monographs in physics. Springer
40. Spohn H (1978) The Lorentz flight process converges to a random flight process. Comm Math Phys 60:277–290
41. Spohn H (1977) Derivation of the transport equation for electrons moving through random impurities. J Stat Phys 17:385–412
42. Uehling EA, Uhlembeck GE (1933) Transport Phenomena in Einstein–Bose and Fermi–Dirac Gases. Phys Rev 43:552–561
43. Villani C (2002) A review of mathematical topics in collisional kinetic theory. In: Friedlander S, Serre D (eds) Handbook of

Mathematical Fluid Dynamics, vol 1. Elsevier, North Holland, pp 71–307
44. Villani C (1998) On a new class of weak solutions to the spatially homogeneous Boltzmann and Landau equations. Arch Rational Mech Anal 143(3):273–307
45. Wigner E (1932) On the quantum correction for the thermodynamical equilibrium. Phys Rev 40:742–759

Books and Reviews

Balescu R (1975) Equilibrium and Nonequilibrium Statistical Mechanics. Wiley, New York
Cercignani C, Boltzmann L (1998) The man who trusted atoms. Oxford University Press, Oxford
Cercignani C, Illner R, Pulvirenti M (1994) The mathematical theory of dilute gases. Applied Mathematical Sciences, vol 106. Springer, New York
Esposito R, Pulvirenti M (2004) From particles to fluids. In: Friedlander S, Serre D (eds) Handbook of Mathematical Fluid Dynamics, vol 3. Elsevier, North Holland, pp 1–83
Spohn H (1991) Large scale dynamics of interacting particles, Texts and monographs in physics. Springer, Heidelberg
Villani C (2002) A review of mathematical topics in collisional kinetic theory. In: Friedlander S, Serre D (eds) Handbook of Mathematical Fluid Dynamics, vol 1. Elsevier, North Holland, pp 71–307

Scaling Properties, Fractals, and the Renormalization Group Approach to Percolation

Dietrich Stauffer
Institute for Theoretical Physics, Cologne University, Köln, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Methods
Quantities and Exponents
Fractal Dimension; Incipient Infinite Cluster
Simple Renormalization Group
Future Directions
Bibliography

Glossary

Cluster Clusters are sets of occupied neighboring sites.

Critical exponent At a critical point or second-order phase transition, many quantities diverge or vanish with a power law of the distance from this critical point; the critical exponent is the exponent for this power law.

Fractals Fractals have a mass varying with some power of their linear dimension. The exponent of this power law is called the fractal dimension and is smaller than the dimension of the space.

Percolation Each site of a large lattice is randomly occupied or empty.

Renormalization A cell of several sites, atoms, or spins is approximated by one single site etc. At the critical point, these supersites behave like the original sites, and the critical point thus is a fixed point of the renormalization.

Definition of the Subject

Percolation theory mostly deals with large lattices where every site is randomly either occupied or empty. In particular it studies the resulting clusters which are sets of neighboring occupied sites.

Introduction

Paul Flory, who later got the Chemistry Nobel prize, published in 1941 the first percolation theory [1], to describe the vulcanization of rubber [2]. Others later applied and generalized it, in particular by dealing with percolation theory on lattices and by studying it with computers. Most of the theory presented here was known around 1980, though in the case of computer simulation with less accuracy than today. But on the questions of universality, of critical spanning probability and of the uniqueness of infinite clusters, the 1990's have shown some of our earlier opinions to be wrong. And even today it is questioned by some that the critical exponents of percolation theory can be applied to real polymer gelation, the application which Flory had in mind two-thirds of a century ago.

On a large lattice we assume that each site independently and randomly is occupied with probability p and empty with probability $1 - p$. Depending on applications, also other words can be used instead of occupied and empty, e. g. Republican and Democrat for the majority party in an electoral district of the USA. A *cluster* is now defined as set of occupied neighboring sites. Percolation theory deals with the number and structure of these clusters, as a function of their size s, i. e. of the number s of occupied sites in the cluster. In particular it asks whether an infinite cluster spans from one side of the lattice to the opposite side. Alternatively, and more naturally if one wants to describe chemical reactions for rubber vulcanization, this site percolation can be replaced by bond percolation, where every site is occupied but the link between neighboring sites is either present with probability p or absent with probability $1 - p$, again independently and ran-

Scaling Properties, Fractals, and the Renormalization Group Approach to Percolation, Table 1
Site and bond percolation thresholds for one dimension, three two-dimensional, four three-dimensional and four hypercubic lattices in higher dimensions [1,3]

p_c	Site	Bond
$d = 1$ chain	1	1
honeycomb	.697043	$1 - 2\sin(\pi/18)$
square	.592746	1/2
triangular	1/2	$2\sin(\pi/18)$
diamond	.4301	.3893
SC	.311608	.248813
BCC	.245691	.180287
FCC	.199236	.120163
$d = 4$ hypercubic	.196885	.160131
$d = 5$ hypercubic	.140797	.118172
$d = 6$ hypercubic	.109018	.094202
$d = 7$ hypercubic	.088951	.078675

domly for each link. A cluster is now a set of neighboring sites connected by links, and the size s of the cluster can be counted as the number of links, or as the number of sites, in that cluster. Because of this ambiguity we discuss here mainly site percolation; bond percolation is similar in the sense that it belongs to the same universality class (same critical exponents). One may also combine both choices and study site-bond percolation where each site is randomly occupied or empty, and where each bond between neighboring occupied sites is randomly present or absent.

Neither temperature nor quantum effects enter this standard percolation model, which is purely geometrical probability theory. However, to understand why percolation works the way it does it is helpful to understand thermal phase transitions like the vapor-liquid critical point; and for magnetic applications it is useful to know that some spins (atomic magnetic moments) have only two states, up or down, according to quantum mechanics. We will explain these physics aspects later.

For small p, most of the occupied sites are isolated $s = 1$, coexisting with only few pairs $s = 2$ and triplets $s = 3$. For large p, most of the occupied sites form one "infinite" cluster spanning the lattice from left to right, with a few small isolated holes in it. Thus there exists one percolation threshold p_c such that for $p < p_c$ we have no spanning cluster and for $p > p_c$ we have (at least) one spanning cluster. Inspite of decades of research in this seemingly simple problem, no exact solution for p_c is proven or guessed for site percolation on the square lattice with nearest-neighbor bonds; only numerically we know it to be about 0.5927462. For site percolation on the triangular lattice or bond percolation on the square lattice, $p_c = 1/2$ exactly. More thresholds are given in Table 1 [1]. They are valid in the limit of $L \to \infty$ for lattices with about L^d sites in d dimensions. For small L instead of a sharp transition at p_c one has a rounded changeover: with a very low probability one chain of L occupied sites at $p = 1/L^{d-1}$ spans from left to right. In one dimension, a small chain can easily be spanned if p is close to one, but for $L \to \infty$ the threshold approaches $p_c = 1$ since at smaller p a hole will appear about every $1/(1-p)$ sites and prevent any cluster to span.

Methods

This section summarizes some of the methods employed to find percolation properties, first by pencil and paper, and then with the help of computers for which Fortran programs are published e. g. in [4,5]. More details on simulations are reviewed by Ziff in this percolation part of this encyclopedia.

Mean Field Limit

The Bethe lattice or Cayley tree neglects all cyclic links and allows a solution with paper and pencil. We start from one central site, and let z bonds emanate from that. At the end of each bond sits a neighbor. Then from each of these neighbors again z bonds emanate, one back to the central site and $z - 1$ to new sites further outward. They in turn lead again each to $z - 1$ new sites, and so on. None of the newly added sites agrees with one of the already existing sites, and so we can travel along the bonds only outwards or back, but never in a loop. It is quite plausible that an infinite cluster of bond percolation is formed if each site leads to at least one more outward site along an existing bond, that means if $(z-1)p > 1$. This condition also holds for site percolation. Thus

$$p_c = 1/(z-1)\,. \tag{1}$$

In this way Flory calculated the threshold and other percolation properties. Today we call this the "mean field" universality class in analogy with thermal phase transitions. The critical exponents, to be discussed below, are integers or simple fractions. To this universality class belong also the Erdös-Rényi random graphs, where we connect in an assembly of N points each pair with a low probability $\propto 1/N$. And the same universality class is reached if we let the dimension d of the hypercubic lattice go to infinity (or at least take it above 6). A disadvantage of the Bethe lattice is its lack of realism: If the length of the bonds is constant, then the exponential increase of the number of sites and bonds with increasing radius leads to an infinite density.

Small Clusters

The probability of a site to be an isolated $s = 1$ cluster on the square lattice is $n_1 = p(1-p)^4$ since the site must be occupied and all its four neighbors be empty. The formula for pairs is $n_2 = 2p^2(1-p)^6$ since the pair can be oriented horizontally or vertically, resulting in the factor 2. Similar, only more difficult, is the evaluation of n_s with a maximum s usually 10 to 20; the general formula is

$$n_s = \sum_t g_{st}\, p^s\, (1-p)^t \tag{2}$$

where the perimeter t is the number of empty neighbors and g_{st} is the number of configurations (or lattice animals, or polyominoes) of size s and perimeter t. The King's College group in London published these results decades ago. With techniques borrowed from series expansions near thermal critical phenomena, these polynomials allow to estimate not only p_c but also many other quantities (see below) diverging or vanishing near p_c.

Leath Cluster Growth

In the cluster growth method of Leath (1976) one starts with one occupied site in the center of the lattice. Then a cluster is grown by letting each empty neighbor of an already occupied cluster site decide once and for all, whether it is occupied or empty. One needs to keep and to update a perimeter list of undecided neighbors. If that list becomes empty, the cluster growth is finished, and no boundary effects of the lattice influence this cluster. If, on the other hand, the cluster reaches the lattice boundary, one has to stop the simulation and can regard this cluster as spanning (from the center to one of the sides). Repeating many times this growth simulation one can estimate p_c as well as the cluster numbers. More precisely, the cluster statistics obtained in this way is not n_s but $n_s s$ since the original center site belongs with higher probability to a larger than to a smaller cluster.

Hoshen-Kopelman Labelling

To go regularly through a large lattice, which may even be an experimentally observed structure to be analyzed by computer, one could number consecutively each seemingly new cluster, and if no clusters merge later then one has a clear classification: All sites belonging to the first cluster have label 1, all sites of the second cluster have label 2, etc. Unfortunately, this does not work. In the later analysis it may turn out that two clusters which at first seemed separate actually merge and form one cluster:

```
* *     * *        1 1     2 2
  *   * * *          1   3 ? 2
* * * * *          4 ? ? x ?
```

Already in the simple structure shown on the left we have several such label conflicts. The labels to the right come from going through the lattice like a typewriter, from left to right, and after each line to the lower line. When we come to the right neighbor of the 3 we see that 3 is really part of the cluster with label 2. And at the right neighbor of 4 we see that 4 belongs to cluster 1. The stupid method is to go back and to relabel all 3 into 2, and all 4 into 1. If then we come to the site marked with x we see that the whole structure is really one single cluster, and thus all labels 2 have to be relabeled into a 1. This is inefficient for large lattices. Instead, Hoshen and Kopelman (1976) gave each site label $m = 1, 2, 3, \ldots$ another index $n(m)$. This label $n(m)$ of labels equals its argument, $n(m) = m$, if it is still a good "root label", and it equals another number k is the cluster with initial label m later turned out to be part of an earlier cluster k. By iterating the command `m = n(m)` until finally the new m equals $n(m)$ one finds this root label. For the above we make the following assignments and re-assignments to n: $n(1) = 1$, $n(2) = 2$, $n(3) = 3$, $n(3) = 2$, $n(4) = 4$, $n(4) = 1$, $n(2) = 1$. Clusters are now characterized by the same root label for all their labels.

An advantage if this method is that only one line of the square lattice, or one hyperplane of the d-dimensional lattice, needs to be stored at any time, besides the array $n(m)$. And that array can also be reduced in size by regular recycling no longer used labels n, just as beer bottles can be recycled. Lattices with more than 10^{13} sites were simulated, using parallel computers. However, understanding the details of the algorithms and finding errors in them can be very frustrating.

Sometimes one wants to determine the cluster numbers for numerous different p from 0 to 1. Instead of starting a new analysis for each different p one may also fill the lattice with new sites, and make the proper labeling of labels whenever a new site was added [6]. Similarly, one can determine the properties of various lattice sizes L by letting L grow one by one and relabeling the cluster after each growth step [7]. Unfortunately, these two methods came long after most of the percolation properties were already studied quite well by standard Hoshen-Kopelman analysis.

Relation to Ising and Potts Models

The relation between percolation and thermal physics was useful for both sides: Scaling theories for percolation could

follow scaling theories for thermal physics from ten years earlier, and computer simulations for thermal physics could use the Leath and Hoshen-Kopelman algorithms of cluster analysis, leading to the Wolff and Swendsen-Wang methods, respectively, a decade later. A mathematical foundation is given by the Kasteleyn-Fortuin theorem [8] for the partition function Z of the Q-state Potts model at temperature T:

$$Z(Q) = \langle Q^N \rangle \tag{3}$$

where N is the total number $\sum_s n_s$ of clusters for bond percolation at probability $1 = \exp(-2\mathrm{J}/k_\mathrm{B}T)$, $\langle \ldots \rangle$ indicates an average over the configurations at this probability, k_B is Boltzmann's constant and 2J is the energy needed to break a bond between neighboring spins. (Each site i of a Potts lattice carries a variable $S_i = 1, 2, \ldots Q$; the energy of a neighbor pair is $-2\mathrm{J}$ if the two variables agree, and zero otherwise.) For the special Ising case $Q = 2$ see also this author in this encyclopedia, "Opinion dynamics..." and "Phase transitions...".

Q values of 3 and larger are interesting since for increasing Q a second-order phase transition with a continuous order parameter changes into a first-order phase transition with a jumping order parameter, when T increases. The special case $Q = 2$ is the spin 1/2 Ising model (the model is pronounced EEsing, not EYEsing since Ernst Ising was born in Cologne, Germany, and became US citizen Ernest Ising only after publishing his theory in 1925 and surviving Nazi persecution 1933–1944). The limit $Q \to 0$ corresponds to some tree structures (no cyclic links, as in Flory's percolation theory, [9]). Percolation, on the other hand, corresponds to the limit $Q \to 1$, in the following way: The "free energy" in units of $k_\mathrm{B}T$ is in this limit $\ln Z = \ln\langle \exp(N \ln Q)\rangle \simeq \ln\langle \exp[(Q-1)N]\rangle \simeq \ln\langle 1 + (Q-1)N\rangle \simeq (Q-1)N$. Thus for Q near unity this thermal free energy, divided by $Q - 1$, is the number of percolation clusters.

In this way thermal physics and percolation are related, and the cluster numbers N correspond to a free energy. In thermal physics, the negative derivative of the free energy with respect a conjugate field gives the order parameter (e.g magnetic field and magnetization), and the field derivative of the order parameter is called the susceptibility. For liquid-gas equilibria, the order parameter is the volume (or the density), the field is the pressure (or chemical potential), and the analog of the susceptibility is the compressibility. This result Eq. (3), not its derivation, we should keep in mind if we now look at the percolation quantities of interest.

Formally we may define for percolation a free energy F as a generating function of a ghost field h:

$$F(h) = \sum_s n_s \exp(-hs)\,. \tag{4}$$

Then its first h-derivative is $-\sum_s n_s s$, and the second one is $\sum_s n_s s^2$, sums which appear below in the percolation probability P_∞ (the order parameter) and the mean cluster size $S = \sum_s n_s s^2 / \sum_s n_s s$ (the susceptibility).

Quantities and Exponents

The basic quantity is n_s, the number (per site) of clusters containing s sites each, and often is an average over several realizations for the same occupation probability p in the same lattice. Several moments

$$M_k = \sum_s n_s s^k \tag{5}$$

are used to define other quantities of interest; in these sums the infinite (spanning) clusters are omitted. The following proportionalities are valid asymptotically in the limit of large lattice size L and for $p \to p_\mathrm{c}$:

$$F = M_0 \propto |p - p_\mathrm{c}|^{2-\alpha} + \ldots \tag{6a}$$

$$P_\infty = p - M_1 \propto (p - p_\mathrm{c})^\beta \tag{6b}$$

$$S = M_2/M_1 \propto |p - p_\mathrm{c}|^{-\gamma}\,. \tag{6c}$$

Here F is the analog of the thermal free energy, where the three dots represent analytic background terms whose derivatives are all finite. Since every occupied site must belong either to a finite or to an infinite cluster, $P_\infty = p - \sum_s n_s s$ is the fraction of sites belonging to the infinite cluster and gives the probability that from a randomly selected site we can walk to a lattice boundary along a path of occupied sites. It is thus called the percolation probability but needs to be distinguished from the probability p that a single site is occupied and from the probability R, with $R(p < p_\mathrm{c}) = 0,\ R(p > p_\mathrm{c}) = 1$, that there is a spanning cluster in the lattice.

The quantity S is usually called the mean cluster size, and we follow this tradition even though it is very bad. There are many ways to define a mean size, and polymer chemists have the much more precise notation of a number average M_1/M_0, a weight average M_2/M_1 and a z average M_3/M_2 for the cluster size (= degree of polymerization). Physicists arbitrarily call the weight-averaged s the mean cluster size S. Numerically, the exponent γ is determined more easily from the "susceptibility" $\chi = M_2 \propto |p - p_\mathrm{c}|^{-\gamma}$, since the denominator M_1 in Eq. (6c) approaches very slowly its asymptotic limit of 1.

Scaling Properties, Fractals, and the Renormalization Group Approach to Percolation, Table 2

Critical exponents for percolation clusters. The mean field values are valid for six and more dimensions and also apply to Flory's Bethe approximation and to Erdös-Rényi random graphs. The exponents α, δ, σ, τ can be derived from the scaling laws, Eq. (8)

d	β	γ	ν
2	5/36	43/18	4/3
3	0.41	1.796	0.88
≥ 6	1	1	1/2

The radius of a cluster R_s can be defined as the rms distance r_i, $i = 1, 2, \ldots, s$ of cluster sites from the center of mass r_c of the cluster (radius of gyration):

$$R_s^2 = \langle \sum_i (r_i - r_c)^2 / s \rangle \tag{6d}$$

where the $\langle \ldots \rangle$ average over all cluster configurations at probability p. Then the correlation length ξ is related to the z-average cluster radius through

$$\xi^2 = \sum_s R_s^2 n_s s^2 / \sum_s n_s s^2 \propto |p - p_c|^{-\nu} \tag{6e}$$

with another critical exponent ν.

Finally, right at $p = p_c$, the cluster numbers decay as

$$n_s \propto 1/s^{2+1/\delta} \tag{7}$$

where δ must be positive to allow a finite density $\sum_s n_s s = p$.

These five critical exponents are not independent of each other but are related in d dimensions through the scaling laws:

$$2 - \alpha = \gamma + 2\beta = (\delta + 1)\beta = d\nu \tag{8a}$$

as known from thermal phase transitions; the last equation involving d is not valid in mean field theory (large d) but only for $d \leq 6$. Table 2 gives the numerical estimates of the exponents in three dimensions as well as their mean field values for $d \geq 6$ and their exact two-dimensional results [10,11]. Thus, for six and less dimensions, if you know two exponents you know them all; thus far.

These scaling laws (8a) can be derived by assuming

$$n_s = s^{-\tau} f[(p - p_c)s^{\sigma}] \quad (\tau = 2 + 1/\delta,\ 1/\sigma = \beta\delta) \tag{8b}$$

which was first postulated for the thermal Ising model, and then successfully applied to percolation. Here f is a suitable scaling function, which only in the mean-field limit approaches a Gaussian.

For both thermal critical phenomena and percolation, "universality" asserts that these critical exponents are independent of many details and (for the Potts model) depend only on the dimensionality d and the number Q of possible spin states. Since percolation corresponds to $Q \to 1$ this means that the exponents depend only on d. There are exceptions from this universality for thermal phase transitions, but for random percolation thus far it worked. However, the numerical value of the percolation threshold p_c is not a critical exponent, depends on the lattice structure, and is different for site and bond percolation.

This universality is one of the reasons why the investigation of exponents is important: They allow to classify models and materials. Similarly, in biology we have many birds of different colors, and many types of domestic animals. Biology became a systematic science only when it was found that all mammals share certain properties, which birds no not have. Thus there is the universality class of mammals.

(The proportionality factors in Eq. (6) are not universal, but some of their combinations are; for example, the ratio of the proportionality factors for S above to below p_c is universal. In some sense also the probability $R(p = p_c)$ of a lattice to contain one spanning cluster at the threshold is universal: same for bond and site percolation; however, that probability depends on the boundary conditions and the shape of the sample and thus is far less universal that the mentioned ratio for S.)

Unfortunately, there is another exponent which does not follow from the cluster numbers and radii and for which no scaling law is accepted which relates it to the other exponents above. This refers to the electrical conductivity

$$\Sigma \propto (p - p_c)^{\mu} \tag{9}$$

when each occupied site (or bond) conducts electrical current and each empty site (or deleted bond) is an insulator. The numerical values are 1.30, 2.0 and 3 in two, three and at least six dimensions. If bonds are realized by elastic springs with bending forces, the elastic exponent may be $\mu + 2\nu$ if entropy effects are negligible, or $2 - \alpha$ if entropy effects are dominant. Moreover, μ is less universal: the above lattice values do not hold on a continuum (conducting spheres which may overlap). Similarly, the kinetics of the Ising model determine a critical exponent which differs in different variants of the kinetics and may not be related to the static Ising exponents like β and γ.

Fractal Dimension; Incipient Infinite Cluster

Fractal Dimension D

Typical objects of geometry classes in school are one-dimensional lines, two-dimensional squares or circles, and three-dimensional cubes or spheres. They have a length (radius) L and a mass (volume for unit density) M with $M \propto L^d$ for d dimensions. In reality, mother nature produces much more complex objects, like trees, where the mass varies with a power of the tree height below 3:

$$M \propto L^D \quad (D < d,\ L \to \infty)\,. \tag{10a}$$

D is the fractal dimension, and such objects are called fractals, particularly if they also are self-similar in that a small twig looks like a big branch, etc. Finite-size scaling theory then relates D of the largest (spanning?) cluster at $p = p_c$ to the above percolation exponents through

$$D = d - \beta/\nu = (\gamma+\beta)/\nu = 1/(\sigma\nu) = d/(1+1/\delta) \tag{10b}$$

for $d \le 6$. Thus the critical cluster is about 1.9-dimensional in two and 2.5-dimensional in three dimensions, while in the mean field regime for $d \ge 6$ we have $D = 4$. Why is this so? Any quantity X which is supposed to vary near $p = p_c$ as $|p - p_c|^x$ does so only for infinitely large systems. For a finite lattice size L, the transition is rounded, and neither X nor any of its p-derivatives diverges or becomes exactly zero. In particular, the typical cluster radius or correlation length $\xi \propto |p - p_c|^{-\nu}$ cannot become infinite but becomes of order L. Then the relation $X \propto \xi^{-x/\nu}$ is replaced by

$$X(p = p_c) \propto L^{-x/\nu} \tag{11a}$$

at the threshold, and

$$X(p \simeq p_c) = L^{-x/\nu} g[(p - p_c)L^{1/\nu}] \tag{11b}$$

near the threshold, with a suitable scaling function g. In particular, the fraction P_∞ of sites belonging to the largest cluster at $p = p_c$ vanishes as $L^{-\beta/\nu}$, and the total number M of sites in this cluster as

$$M \propto L^{d-\beta/\nu} \quad \text{or} \quad D = d - \beta/\nu \tag{11c}$$

as asserted in Eq. (10b).

Figure 1 shows the second moment $\chi = M_2 = \sum_s n_s s^2$ in small (curve) and large (+) simple cubic lattices, differing only for $p \simeq p_c$. Figure 2 shows right at $p = p_c$ the variation with lattice size of the number M of sites in the largest cluster and of the second moment M_2 (susceptibility).

In a finite lattice, the probability $R(p)$ of a spanning cluster to exist goes from nearly zero to nearly unity in a p-interval proportional to $1/L^{1/\nu}$, according to Eq. (11b) with $x = 0$. The derivative $\mathrm{d}R/\mathrm{d}p$ is the probability that spanning first occurred at probability p. It is plausible that this probability, peaked around p_c, is a Gaussian, i. e. a normal distribution. Unfortunately, the Evil Empire, also known as the Departments of Chemical Engineering, destroyed [12] this beautiful idea: Since for $p \simeq p_c$ and $\xi \sim L$ every part of the lattice is correlated with the rest of the lattice, the central limit theorem does not hold.

(If for $p \ll p_c$ we let the cluster size s go to infinity, which requires a special algorithm, we get into the universality class of lattice animals, Sect. "Small Clusters". Most simply, in the limit $p \to 0$, Eq. (2) simplifies to $n_s/p^s = g_{st}$, that means we look at the distribution of configurations with s sites and perimeter t, where all configurations of a given s are weighted equally, whatever their perimeter t is. An important result for these animals is that in three dimensions their radius R_s varies as $\sqrt{s}$, i,e. their fractal dimension is exactly 2. In two dimensions, only numerical estimates exist with $D \simeq 1.56$. It is highly unusual that a problem has an exact solution in three but not in two dimensions.)

Incipient Infinite Cluster

Right at $p = p_c$ the largest cluster spans the lattice with a pseudo-universal probability $0 < R(p_c) < 1$, and then has a density P_∞ going to zero for L going to infinity. It is also called the incipient infinite cluster IIC. Most of the IIC consists of dangling ends which carry no current if the cluster is interpreted as a random resistor network with conductivity Σ, see Eq. (9) above. The remaining current carrying "backbone" has a fractal dimension 1.643 in two dimensions, 1.7 in three and 2 in at least six dimensions and mostly consists of blobs where current flows along several parallel though connected paths. The few "articulation" sites or bonds, the removal of which cuts the network into two or more parts, are also called "red" since all the current flows through them; they have a fractal dimension of only $1/\nu = 0.75$, 1.14 and 2 in two, three and $\ge$ six dimensions.

How many infinite clusters do we have? The easy answer is: none below, perhaps one at and always one above p_c in an infinite network. Indeed, this is what was claimed mathematically in the 1980's [13]: The number of infinite clusters is zero, one or infinite. Later mathematics excluded the last choice of infinitely many clusters, even though in seven dimensions scaling arguments, confirmed

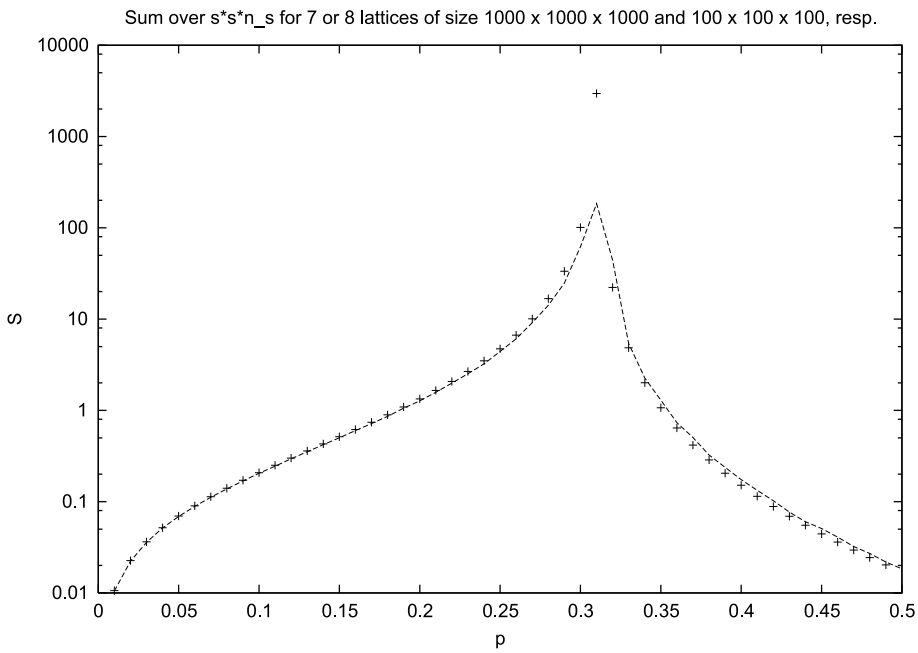

Scaling Properties, Fractals, and the Renormalization Group Approach to Percolation, Figure 1
"Susceptibility" M_2 in simple-cubic lattice. For the smaller size the maximum is reduced appreciably

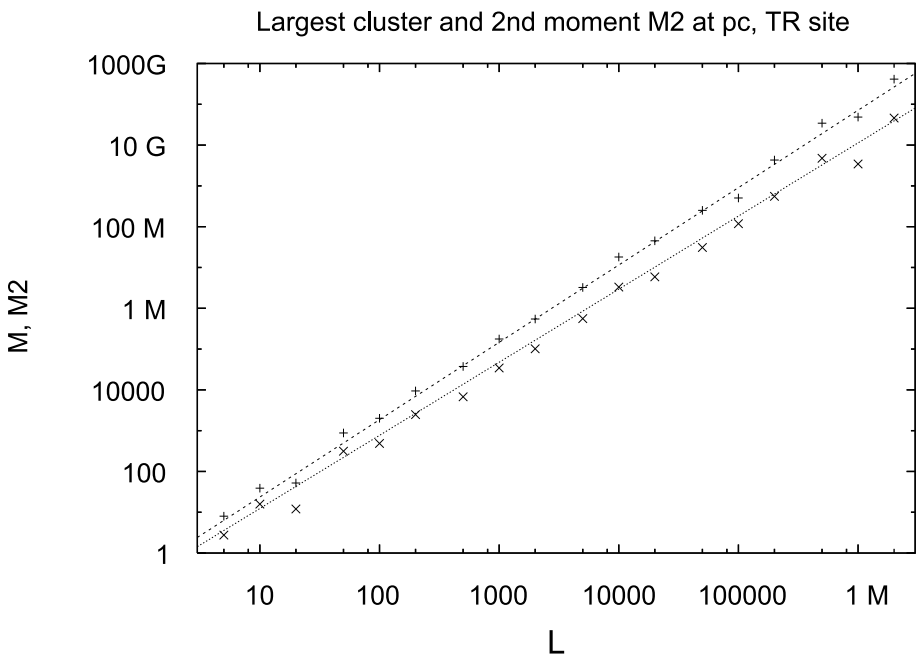

Scaling Properties, Fractals, and the Renormalization Group Approach to Percolation, Figure 2
Number M of sites in largest cluster (+) and susceptibility M_2 (x) at $p = p_c = 1/2$ for triangular site percolation. The two straight lines have the exact slopes $D = 91/48$ and $\gamma/\nu = 43/24$ predicted by finite-size scaling. The largest lattice took about 36 h on a workstation with 2 Gb memory. Tiggemann [7] simulated $L = 7 \times 10^6$, 25024, 1305, 225 for $d = 2, 3, 4, 5$ on a large parallel computer

by numerical studies [14], indicated the number of IIC to go to infinity for increasing L in seven dimensions. Only in 1995 and later Aizenman [15] predicted that in all dimensions one may have several spanning clusters at $p = p_c$, in agreement with simulations [16].

Why were the earlier uniqueness theorems irreproducible at p_c and for very elongated rectangles even above p_c [17]? A clear definition of "infinite" is missing in some of the mathematics, although [13] defined a cluster as infinite if its cardinality (= number of sites in it) is infinite for $L \to \infty$ in a hypercubic lattice of L^d sites. Clear definitions of infinity are, of course, needed for reliable proofs [18]. Measure theory as applied in some theorems may be based on some axioms which are not applicable for a fractal IIC. Very simply, imagine each line of an $L \times L$ square lattice to have one randomly selected site occupied and all others empty. The set of occupied site then has cardinality L which is infinite for infinite lattices, but its density becomes zero. Does your measure theory agree with this? More relevant for percolation, even for $p < p_c$ the largest cluster has a size increasing logarithmically with lattice size and thus can be described as infinite, invalidating the percolation threshold as the onset of infinite clusters. Thus infinite might be defined as increasing with a positive power of L, i. e. having a positive fractal dimension. Then we have infinitely many infinite clusters only at $p = p_c$, though in most cases only the largest of them is a spanning cluster. Using "spanning" as a definition of an infinite cluster seems to cause the smallest problems.

Thus one should not regard a question as settled if some mathematical theorem claims to have answered it. The mathematics may not apply to the same problem one is interested in, or (see bootstrap percolation in this encyclopedia) may apply only for unrealistically large lattices. On the other hand, also computer simulations should be relied upon only if confirmed independently. And in the interpretation of simulation results one should be objective and not try to agree with prevailing theories. For example, [14] might already have seen the multiplicity of infinite clusters in five dimensions, not only in seven, had she not followed her obviously incompetent postdoctoral mentor.

(On a more positive side, mathematicians [19] solved biased diffusion on percolating clusters above p_c only a few years after physicists still had controversies about their simulations.)

Simple Renormalization Group

Why are scaling laws and finite-size scaling so simple? Why is universality valid for the exponents? These question arose for thermal critical phenomena as well as for percolation. The main reason is that the correlation length ξ goes to infinity at the critical point. Thus all approximations which restrict the correlations to some finite lengths eventually become wrong, and instead the scaling ideas become correct. They were explained by Ken Wilson through what he called renormalization group, around 1970, and he got the physics Nobel prize for it in 1982. Basically, since correlations extend over long distances, the single atom or lattice point becomes irrelevant and can be averaged over. In politics, we have a similar effect: Many democracies are based on electoral districts, and the candidate winning most votes within this district represents this district in the national parliament. It is the cooperation of many people within the electoral district, not the single vote, which is decisive.

Returning to an $L \times L$ lattice, we can divide it into many blocks of linear dimension b, and treat a block analogously to an electoral district. Thus in an Ising model, if the majority of block spins point upward, the whole block is represented by a superspin pointing up, analogous to the single representative in politics. These block spins then act like the original spins, one can put $b \times b$ superspins into one superblock, and have just one super-representative following the majority opinion of the representatives within the superblock. This process can be continued: at each stage $b \times b$ lower representatives are normalized into a single higher representative.

Such a renormalization by majority rule works fine with Ising spins, but percolation deals with connections, not with up and down spins. Thus for percolation a $b \times b$ block is normalized into an occupied supersite if and only if there is a spanning cluster within the block; otherwise the superblock is defined empty. In this way, whole blocks are normalized into single sites via connectedness. And the renormalization is reduced to the standard question which was asked already before Wilson's invention: Does a $b \times b$ lattice have a spanning cluster? The supersite is thus occupied if and only if the block spans, which happens with probability $R_b(p)$. If we call p' the probability of the supersite to be occupied, we thus have

$$p' = R_b(p) \,. \tag{12a}$$

If we are at $p = p_c$, then the renormalization should not change anything drastic since ξ is larger than any b; thus if the renormalization would be exact we would have

$$p'_c = R_b(p_c) \,. \tag{12b}$$

Practically we determine a fixed point $p = p^*$ such that

$$p^* = R_b(p^*) \,. \tag{12c}$$

and then find p_c as the limit of p^* for $b \to \infty$, which again is similar to what percolation experts did before this renormalization theory.

A particularly simple example is the triangular site percolation problem with $p_c = 1/2$, if we do not divide the lattice into large $b \times b$ blocks but into small triangles of three sites which are nearest neighbors, as shown on the left:

```
   *              x              x
*     *        x     x        x     .
```

The triangle contains a spanning cluster if either all three sites are occupied (x, central diagram) or two sites are occupied (x) and one site is empty (., right diagram). The first choice appears with probability p^3, the second with probability $p^2(1-p)$. However, this second choice has three possible orientations since each of the three sites can be the single empty site. Thus the total probability of the triangle to have a spanning cluster is

$$p' = p^3 + 3(1-p)p^2 \tag{13a}$$

with three fixed points p^* where $p' = p$:

$$p^* = 0, \quad p^* = 1/2, \quad p^* = 1\,. \tag{13b}$$

The second of these fixed points is the percolation threshold, while the first corresponds to lattice animals (Sect. "Small Clusters" and end of Sect. "Fractal Dimension D") and the third to compact non-fractal clusters. With somewhat more effort one can derive also a good approximation for ν.

This agreement of the fixed point p^* with the true threshold $p_c = 1/2$ is not valid for other lattices or block choices. Nevertheless there was a widespread fixed-point consensus that $R_b(p_c) = p_c$ for sufficiently large b. Regrettably, the Evil Empire [20] again destroyed this beauty and found $R_b(p_c) = 1/2$ for square site percolation where $p_c \simeq 0.593$. In general, $R(p_c)$ is a pseudo-universal quantity depending on boundary conditions and sample shape, while p_c for large samples is independent of these details but is different for site and bond percolation and depends on the size of the neighborhood. Life was much nicer before. Fortunately, if a fixed point is determined by Eq. (12c) and the block size goes to infinity, then the fixed point still approaches p_c.

Future Directions

This review summarized the basic theory, particularly when it was not yet contained in the earlier books [1]. Applications were left to the Sahimi book [1]; even for the very first application [2] there is not yet a complete consensus that the three-dimensional percolation exponents apply to polymer gelation. More recent applications are social percolation [21] for marketing by word-of-mouth, and stock market fluctuations due to herding among traders [22].

Percolation theory, similar to Fortran programming or capitalism, was thought to be finished but seems to be alive and kicking. Nevertheless I think the future is more in its applications.

The manuscript was improved by criticism of A. Aharony.

Bibliography

1. Stauffer D (1979) Phys Rep 54:1; Essam JW (1980) Rep Prog Phys 43:843; Stauffer D, Aharony A (1994) Introduction to Percolation Theory. Taylor and Francis, London (revised second edition); Sahimi M (1994) Applications of Percolation Theory. Taylor and Francis, London; Bunde A, Havlin S (1996) Fractals and Disordered Systems. Springer, Berlin; Grimmett G (1999) Percolation, second edition. Springer, Berlin
2. Flory PJ (1941) J Am Chem Soc 63:3083
3. Grassberger P (2003) Phys Rev E 67:036101
4. Redner S (1982) J Statist Phys 29:309
5. Stauffer D, Jan N (2000) In: Khajehpour MRH, Kolahchi MR, Sahimi M (eds) Annual Reviews of Computational Physics, vol VIII (Zanjan School). World Scientific, Singapore
6. Newman MEJ, Ziff RM (2000) Phys Rev Lett 85:4104
7. Tiggemann D (2006) Int J Mod Phys C 17:1141 and Ph D thesis, Cologne University
8. Kasteleyn PW, Fortuin CM (1969) J Phys Soc Jpn Suppl S 26:11
9. Deng VJ, Garoni TM, Sokal AD (2007) Phys Rev Lett 98:030602
10. Nienhuis B (1982) J Phys A 15:199
11. Smirnov S, Werner W (2001) Math Res Lett 8:729
12. Ziff RM (1994) Phys Rev Lett 72:1942
13. Newman CM, Schulman LS (1981) J Stat Phys 26:613
14. de Arcangelis L (1987) J Phys A 20:3057
15. Aizenman M (1997) Nucl Phys (FS) B 485:551
16. Shchur LN, Rostunov T (2002) JETP Lett 76:475
17. Stauffer D (1999) J Irreproducible Results 44:57
18. Jarai AA (2003) Ann Prob 31:444
19. Berger N, Ganten N, Peres Y (2003) Probab Theory Relat Fields 126:221
20. Ziff RM (1992) Phys Rev Lett 69:2670
21. Weisbuch G, Solomon S (2002) In: Bornholdt S, Schuster HG (eds) Handbook of graphs and networks. Wiley-VCH, Weinheim, p 113
22. Cont R, Bouchaud J-P (2000) Macroeconom Dyn 4:170

Scenario-Driven Planning with System Dynamics

NICHOLAS C. GEORGANTZAS
Fordham University Business Schools, New York, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Environmental Turbulence and Future Uncertainty
SdP with SD: The Modeling Process ≡ Strategic Situation Formulation
Case 1: Cyprus' Environment and Hotel Profitability
Case 2: A Japanese Chemicals Keiretsu (JCK)
Future Directions
Bibliography

Glossary

Mental model how one perceives cause and effect relations in a system, along with its boundary, i. e., exogenous variables, and the time horizon needed to articulate, formulate or frame a decision situation; one's implicit causal map of a system, sometimes linked to the reference performance scenarios it might produce.

Product either a physical good or an intangible service a firm delivers to its clients or customers.

Real option right and obligation to make a business decision, typically a tangible investment. The option to invest, for example, in a firm's store expansion. In contrast to financial 'call' and 'put' options, a *strategic real option* is not tradable. Any time it invests, a firm might be at once acquiring the *strategic real options* of expanding, downsizing or abandoning projects in future. Examples include research and development (abbreviated R&D), merger and acquisition (abbreviated M&A), licensing abroad and media options.

Scenario a postulated sequence or development of events trough time; via Latin *scena 'scene'*, from Greek *σκηνή*, *skēnē* '*tent, stage*'. In contrast to a forecast of what *will* happen in the future, a scenario shows what *might* happen. The term *scenario* must *not* be used loosely to mean situation. *Macro-environmental* as well as *industry-*, *task-* or *transactional-environmental* scenarios are merely inputs to the *strategic objectives* and *real options* a firm must subsequently explore through *strategic scenarios*, *computed* or simulated with an explicit, formal *system dynamics* (abbreviated SD) model of its strategic situation. *Computed strategic scenarios* create the multiple perspectives that strategic thinkers need to defeat the tyranny of dogmatism that often assails firms, governments and other social entities or organizations.

Scenario-driven planning (abbreviated SdP) to attain high performance through strategic flexibility, firms use the SdP *management technology* to create foresight and to anticipate the future with strategic real options, in situations where the business environment accelerates frequently and is highly complex or interdependent, thereby causing uncertainty.

Situation the set of circumstances in which a firm finds itself; its (strategic) state of affairs.

Strategic management process (abbreviated SMP) geared at detecting environmental threats and turning them into opportunities, it *proceeds from* a firm's mission, vision and environmental constraints *to* strategic goals and objectives *to* strategy design or formulation *to* strategy implementation or strategic action *to* evaluation and control *to* learning through feedback (background, Fig. 2).

SMP-1 environmental scanning monitors, evaluates and disseminates knowledge about a firm's internal and external environments to its people. The internal environment contains *s*trengths and *w*eaknesses within the firm; the external shows future *o*pportunities and *t*hreats (abbreviated SWOT).

SMP-2 mission a firm's purpose, *raison d'être* or reason for being.

SMP-3 objectives performance (*P*) goals that SMP often quantifies for some *P* metrics.

SMP-4 policy decision-making guidelines that link strategy design or formulation to action or implementation tactics.

SMP-5 strategy a comprehensive plan that shows how a firm might achieve its mission and objectives. The three strategy levels are: corporate, business and process or functional.

SMP-6 strategy design or formulation the interactive, as opposed to antagonistic, interplay of strategic content and process that creates flexible long-range plans to turn future environmental threats into opportunities; includes internal strengths and weaknesses as well as strategic mission and objectives, and policy guidelines.

SMP-7 strategic action or implementation the process by which strategies and policies are put into action through the development of programs, processes, budgets and procedures.

SMP-8 evaluation and control sub-process that monitors activities and performance, comparing actual results with desired performance.

SMP-9 learning through feedback occurs as knowledge about each SMP element enables improving previous SMP elements (background, Fig. 2).

System an organized group of interrelated components, elements or parts working together for a purpose; parts might be either goal seeking or purposeful.

System dynamics (abbreviated SD) a lucid modeling method born from the need to manage business performance through time. Thanks to Forrester [23], who discovered that all change propagates itself through stock and flow sequences, and user-friendly SD software (*iThink*®, *Vensim*®, etc.), SD models let managers see exactly how and why, like other biological and social organizations, business firms perform the way they do. Unlike other social sciences, SD shows exactly how *feedback loops*, i. e., circular cause and effect chains, each containing at least one time lag or delay, interact within a system to determine its performance through time.

Variable or metric something that changes either though time or among different entities at the same time. An *internal change lever* is a decision or policy variable that a strategy-design modeling, or client, team controls. An *external change trigger* is an environmental or policy variable that a strategy-design modeling team does not control. Both *trigger* and *lever* variables can initiate change and be either endogenous or exogenous to a model of a system.

However certain our expectation, the moment foreseen may be unexpected when it arrives
—T.S. Eliot

Definition of the Subject

Many of us live and work in and about business ecosystems with complex structures and behaviors. Some realize that poor performance often results from our very own past actions or decisions, which come back to haunt us. So business leaders in diverse industries and firms, such as *Airbus, General Motors, Hewlett-Packard, Intel* and *Merck*, use scenario-driven planning (SdP) with system dynamics (SD) to help them identify, design and apply high-leverage, sustainable solutions to dynamically complex strategic-decision situations. One must know, for example, if the effect of an environmental change or strategic action gets magnified through time or is dampened and smoothed away. What may seem insignificant at first might cause major disruption in performance. SdP with SD shows the causal processes behind such dynamics, so firms can respond to mitigate impacts on performance.

Accelerating change and complexity in the global business environment make firms and other social organizations abandon their *inactive, reactive* and *preactive* modes [2]. SdP with SD turns them *proactive*, so they can translate anticipation into action. To properly transform anticipation into action, computed with SD models, 'strategic scenarios' must meet four conditions: *consistency, likelihood, relevance* and *transparency* [37]. Combining SdP with SD for that purpose, with other tools, like actor and stakeholder purposes, morphological methods or probability might help avoid entertainment and explore all possible scenarios. Indeed, SdP with SD

> "does not stand alone… modeling projects are part of a larger effort… modeling works best as a complement to other tools, not as a substitute" (see p. 80 in [75]).

SdP with SD is a systematic approach to a vital top-management job: leading today's firm in the rapidly changing and highly complex global environment. Anticipating a world where product life cycles, technology and the mix of collective- and competitive-strategy patterns change at an unprecedented rate is hard enough. Moving ahead of it might prove larger than the talent and resources now available in leading firms. SdP with SD leads to a decisive integration of strategy design and operations, with the dividing line much lower than at present. As mid-level managers take on more responsibility, senior executives become free to give more time and attention to economic conditions, product innovation and the changes needed to enhance creativity toward strategic flexibility [23].

It is perhaps its capacity to reintegrate strategy content and process that turns SdP with SD into a new paradigm for competitive advantage [42], and simulation modeling in general [28], into a critical fifth tool, in addition to the four tools used in science: observation, logical-mathematical analysis, hypothesis testing and experiment [77]. But full-fledged SD models also allow computing scenarios to assess possible implications of strategic situations. Strategic scenarios are not merely hypothesized plausible futures, but computed by simulating combined changes in strategy and in the business environment [32].

Computed scenarios help managers understand what they do not know, enabling strategy design and implementation through the coalignment of timely tactics to improve long-term performance. Through its judicious use of resources, scenario-driven planning with system dynamics makes the tactics required for implementation clear [27]. And because computed scenarios reveal the required coalignment of tactics through time, SdP with SD helps firms become flexible, dependable and efficient, and save time!

Everyone's mind sees differently, but if there is truth in the adage 'a picture is worth a thousand words', then the complex interrelations that SdP with SD unearth and show must be worth billions. In a world where strategic chitchat dominates, one can only hope that SdP with SD will play

a central role in public and private dialogues about dynamically complex opportunities and threats.

We shape our buildings; thereafter,
our buildings shape us
—Winston Churchill

Introduction

Following on the heels of Ackoff and Emery [3] and Christensen [10], respectively, Gharajedaghi [35] and Raynor [64] show how strategies with the best chances for brilliant success expose firms to debilitating uncertainty. Firms fail as their recipes for success turn bad through time. Gharajedaghi [35] shows, for example, five strategy scenarios that convert success to failure. Each scenario plays a critically different role. Together, however, these scenarios form a dynamically complex system. Through time, as each scenario plays, it enables the context for the next:

1. *Noble ape* or *copycat* strategy imitates and replicates advantage. Also called 'shadow marketing', it lets shadowy copycats instantly *shadow market* product technology, often disruptively.
2. *Patchy* or *sluggish* strategy delays responses to new technology. When this *second* scenario plays, then patching up wastes time, enabling competitors to deliver new technology and to dominate markets. Worse, it causes costs to rise as it drives down product quality.
3. *Satisficing* or *suboptimal* strategy scenarios take many forms. One entails a false assumption: if a policy lever helps produce desired performance, then pulling or pushing on that lever will push performance further.
4. *Gambling* or *changing the game* strategy scenario transforms a strategic situation by playing the game successfully. While dealing with a challenge, firms gradually transform their strategic situation and change the basis for competition, so a whole new game and set of issues emerge. Success marked, for example, the beginning of the *information era*. But competitive advantage has already moved away from having access to information. In our *systems era* [2], creating new knowledge and generating insight is the new game [81].

Lastly, the cumulative effects from these four strategy scenarios trickle down to the:

5. *Archetypal swing* or *paradigm shift* scenario. Both learning and unlearning can cause archetypal swings and paradigm shifts to unfold through time intentionally [76]. These also occur unintentionally when conventional wisdom fails to explain patterns of events that challenge prevailing mental models. The lack of a convincing explanation creates a twilight zone where acceptable ideas are not competent and competent ideas are not acceptable.

Beliefs about the future drive strategies. But the future is unpredictable. Worse, success demands commitments that make it impossible to adapt to a future that turns out surprising. So, strategies with great success potential also bear high failure probabilities. Raynor [64] calls this the *strategy paradox*. Dissolving it requires turning environmental uncertainty into strategic flexibility. To make it so, Raynor urges managers to: anticipate multiple futures with scenarios, formulate optimal strategies for each future, accumulate *strategic real options* [5] and manage the select options portfolio.

SdP with SD helps managers who operate in an uncertain world question their assumptions about how the world works, so they can see it more clearly. To survive, the human mind overestimates small risks and underestimates large risks. Likewise, it is much more sensitive to losses than to gains. So the capability to leverage opportunities and to mitigate risk might have become an economic value driver.

The purpose of computing scenarios is to help managers alter their view of reality, to match it up more closely with reality as is and as it might become. To become a leader, a manager must define reality. The SdP with SD purpose is *not*, however, to paint a more accurate picture of tomorrow, but to improve the quality of decisions about the future. Raynor says that the requisite strategic flexibility, which SdP with SD creates:

> "is not a pastiche of existing approaches. Integrating these tools and grounding them in a validated theory of organizational hierarchy creates something that is quite different from any of these tools on its own, or in mere combination with the others" (see p. 13 in [64]).

Indeed, knowledge of common purposes and the acceptable means of achieving them form and hold together a purposeful hierarchical system. Its members know and share values embedded in their culture, which knits parts into a cohesive whole. And because each part has a lot to say about the whole, consensus is essential to SdP with SD for the co-alignment of diverse interests and purposes.

Ackoff and Emery [3], Gharajedaghi [35] and Nicolis [55] concur that purpose offers the lens one needs to see a firm as a multi-minded social net. A purposeful firm pro-

duces either the same result differently in the same environment or different results in the same or different environments. Choosing among strategic real options is necessary but insufficient for purposefulness. Firms that behave differently but show only one result per environment are goal seeking, not purposeful. Servomechanisms are goal seeking but people are purposeful. As a purposeful system, the firm is part of purposeful sub-systems, such as its *industry value chain* [61] and the society. And firms have purposeful people as members. The result is a dynamically interdependent, i. e., complex, hierarchical purposeful system.

A firm's value chain is, along with its primary and support activities, at once a member of at least one industry value chain and of the society or *macro-environment*. Industry analysis requires looking at value chains independently from the society [61]. But people, the society and firm and industry value chains are so interdependent, so interconnected, that an optimal solution might not exist for any of them independently of the others. SdP with SD helps firms co-align the 'plural rationality' of purposeful stakeholder groups with each other and that of the system as a whole.

Seeing strategic management as a *strategies and tactics net* [27] is in perfect syzygy with the *plural rationality* that SdP with SD accounts for among individuals, groups and organizations. Singer [73,74] contrasts monothematic conventional universes of traditional rationality with the multiverse-directed view of plural rationality. In counterpoint, Morecroft's [52] computed scenarios trace the dysfunctional interactions among sales objectives, overtime and sales force motivation to the intended, i. e., stated, singular rationality that drove action in a large sales organization.

Because their superordinate purpose is neither to compete nor to collaborate, but to develop new wealth-creating capabilities, in unique ways that serve both current and future stakeholder interests, customers and clients included [51], firms can benefit from the multiverse-directed view of strategic management as a net of strategies and tactics. SdP with SD helps firms break free from the tradeoffs tyranny of the mass-production era. Evidently, adherents to tradeoffs-free strategy like *Bell Atlantic*, *Daimler-Benz*, *Hallmark* and *Motorola* "can have it all" [60].

A firm must serve the purposes of its people as well as those of its environment, not as a mindless mechanical system, but as a living, purposeful, *knowledge-bonded* hierarchical system [3,35,55,81]. To clarify, a bike always yields to its rider, for example, regardless of the rider's desire; even if that entails running into a solid brick wall. Ouch! But riding a horse is an entirely different story. Horse and rider form a knowledge-bonded system: the horse must know the rider and the rider must know exactly how to lead the horse.

SdP with SD History: Always Back, Always in Style, Always Practical

Herman Kahn introduced scenarios to planning while at RAND Corporation in the 1950s [45]. Scenarios entered military strategy studies conducted for the US government. In the 1960s, Ozbekhan [58] used urban planning scenarios in Paris, France. Organization theorists and even novelists were quick to catch on. The meaning of scenarios became literary. Imaginative improvisation produced flickering apocalyptic predictions of strikingly optimistic and pessimistic futures. Political and marketing experts use scenarios today to jazz up visions of favorable and unfavorable futures.

Wack [78,79] asserts it was Royal Dutch Shell that came up with the idea of scenarios in the early 1970s. Godet [36] points to the French OTAM team as the first to use scenarios in a futures study by DATAR in 1971. Brauers and Weber [8] claim that Battelle's scenarios method [49] was originally a German approach. In connection with planning, however, most authors see scenario methods as typically American.

Indeed, during the 1970s, US researchers Olaf Helmer and Norman Dalkey developed scenario methods at RAND for eliciting and aggregating group judgments via Delphi and cross-impact matrices [4]. They extended cross impact analysis within statistical decision theory [39]. A synthesis of scenario methods began in the 1970s that draws together multiple views, including those of professional planners, analysts and line managers.

Ansoff [6] and other strategy theorists state that the 1970s witnessed the transformation of global markets. Today, changes in the external sociopolitical environment become pivotal in strategy making. Combined with the geographical expansion of markets, they increase the complexity of managerial work. As environmental challenges move progressively faster, they increase the likelihood of strategic surprises. So, strategic thinkers use scenarios to capture the nonlinearity of turbulent environments. Examples are Hax and Majluf [38] and, more clearly so, Porter [61] and Raynor [64]. They consider scenarios instrumental both in defining uncertainty and in anticipating environmental trends.

Huss and Honton [41] see scenarios emerge as a distinct field of study, a hybrid of a few disciplines. They identify multiple scenarios methods that fall into three major categories:

1. Intuitive logics [78,79], now practiced by *SRI International*,
2. Trend-impact analysis, practiced by the *Futures Group* and
3. Cross-impact analysis, practiced by the *Center for Futures Research* using INTERAX (Interactive Cross-Impact Simulation) and by Battelle using BASICS (BAttelle Scenario Inputs to Corporate Strategies).

Similarly, after joining Ozbekhan to advocate reference scenarios, Ackoff [2] distinguishes between:

1. Reference projections as piecemeal extrapolations of past trends and
2. The overall reference scenario that results from putting them together.

Based on Acar's [1] work under Ackoff, Georgantzas and Acar [32] explore these distinctions with a practical managerial technology: *comprehensive situation mapping* (CSM). CSM is simple enough for MBA students to master in their capstone Business Policy course. With the help of *Vensim® PLE* [18], CSM computes scenarios toward achieving a well-structured process of managing ill-structured strategic situations. In their introduction to SD, Georgantzas and Acar (see Chap. 10 in [32]) draw from the banquet talk that Jay Wright Forrester, Germeshausen Professor Emeritus, MIT, gave at the 1989 *International Conference of the System Dynamics Society*, in Germany, at the University of Stuttgart:

After attending the Engineering College, University of Nebraska, which included control dynamics at its core, Forrester went to MIT. There he worked for Gordon S. Brown, a pioneer in feedback control systems. During World War II, Brown and Forrester worked on servomechanisms for the control of radar antennas and gun mounts. This was research toward an extremely practical end, during which Forrester run literally from mathematical theory to the battlefield, aboard the US carrier *Lexington*.

After the war, Forrester worked on an analog aircraft flight simulator that could do little more than solve its own internal idiosyncrasies. So, Forrester invented *random-access magnetic storage* or *core memory*. His invention went into the heart of *Whirlwind*, a digital computer used for experimental development of military combat systems that eventually became the *semi*automatic *g*round *e*nvironment (SAGE) air defense system for North America.

Alfred P. Sloan, the man who built *General Motors*, founded the *Sloan School of Management* in 1952. Forrester joined the school in 1956. Having spent fifteen years in the science and engineering side of MIT, he took the challenge of exploring what engineering could do for management.

One day, he found himself among students from *General Electric*. Their household appliance plants in Kentucky puzzled them: they would work with three or four shifts for some time and then, a few years later, with half the people laid off. Even if business cycles would explain fluctuating demand, that did not seem to be the entire reason. *GE*'s managers felt something was wrong.

After talking with them about hiring, firing and inventory policies, Forrester did some simulation on a paper pad. He started with columns for inventories, employees and customer orders. Given these metrics and *GE*'s policies, he could tell how many people would be hired or fired a week later. Each decision gave new conditions for employment, inventories and production. It became clear that wholly determined internally, the system had potential for oscillatory dynamics. Even with constant incoming orders, the policies caused employment instability. That longform simulation of *GE*'s inventory and workforce system marked the beginning of system dynamics [23,24,25,26].

SdP with SD Use and Roadmap

Scenarios mostly help forecast alternative futures but, as firms abandon traditional forecasting methods for interactive planning systems, line managers in diverse business areas adopt scenario-driven planning with system dynamics. Realizing that a tradeoffs-free strategy design requires insight about a firm's environment, both business and sociopolitical, to provide intelligence at *all* strategy levels, firms use SdP with SD to design *corporate, business* and *process* or *functional* strategies. SdP with SD is not a panacea and requires discipline, but has been successful in many settings. Its transdisciplinary nature helps multiple applications, namely capital budgeting, career planning, civil litigation [31], competitive analysis, crisis management, decision support systems (DSS), macroeconomic analysis, marketing, portfolio management and product development [65]. SdP with SD is a quest for managers who wish to be leaders, not just conciliators. They recognize that *logical incrementalism*, a piecemeal approach, is inadequate when the environment and their strategy change together.

Top management might see both divisional, i. e., business, and process or functional strategies as ways of implementing corporate strategy. But *active subsidiaries* [43,44] provide both strategic ideas and results to their parent en-

terprise. Drawing too stiff a line between the corporate office and its divisions might be

> "an unhealthy side effect of our collective obsession with generating returns. The frameworks for developing competitive strategy that have emerged over the last thirty years have given us unparalleled insight into how companies can succeed. And competitive strategy remains enormously important, but it should be the preserve of divisional management… corporate strategy should be focused on the management of strategic uncertainty" (see p. 11 in [64]).

Roadmap It is material to disconnect scenarios from unproductive guesswork and to anchor them to sound practices for strategy design. This guided tour through the fascinating but possibly intimidating jungle of scenario definitions shows what the future might hold for SdP with SD. Extensive literature, examples, practical guidelines and two real-life cases show how computed scenarios help manage uncertainty, that necessary disciple of our open market system. Unlike extrapolation techniques, SdP with SD encourages managers to think broadly about the future.

The above sections clarify the required context and provide a glossary. Conceptual confusion leads to language games at best and to operational confusion at worst [15]. SdP with SD helps firms avert both types of confusion. Instead of shifting their focus away from actuality and rationality, managers improve their insight about fundamental assumptions underlying changes in strategy. The mind-set of SdP with SD makes it specific enough to give practical guidance to those managing in the real world, both now and in the future.

The sections below look at *three* SdP with SD facets linked to strategy design and implementation. The *first* facet involves the business environment, the forces behind its texture and future's requisite uncertainty (Sect. "Environmental Turbulence and Future Uncertainty"). The *second* entails unearthing unstated assumptions about changes in the environment and in strategy, and about their potential combined effects on performance. The SdP with SD framework (Sect. "SdP with SD: The Modeling Process ≡ Strategic Situation Formulation") builds on existing scenario methods. Its integrative view delineates processes that enhance institutional learning, bolster productivity and improve performance through strategic flexibility. It shows why interest in computed scenarios is growing.

The *third* facet entails *computing* the combined or mixed effects on performance of changes both in the environment and in strategy. Even in mature economies, no matter how and how frequently said, decision makers often forget how the same action yields different results as the environment changes. The result is often disastrous. Conversely, the tight coupling between computed scenarios and strategic results can create new knowledge. Linking a mixed environmental and decision scenario in a one-to-one correspondence to a strategic result suits the normative inclination of strategic management, placing rationalistic inquiry at par with purely descriptive approaches in strategy research.

The unified treatment of SdP with SD and the strategy-making process grants a practical bonus, accounting for the entry's peculiar nature. It is not only a conceptual or idea contribution, but also an application-oriented entry. Sections "Case 1: Cyprus' Environment and Hotel Profitability" and "Case 2: A Japanese Chemicals Keiretsu (JCK) present two real-life cases of scenario-driven planning with system dynamics. Written with both the concrete and the abstract thinker in mind, the two cases show how firms and organizations build scenarios with a modest investment. SdP with SD provides an effective management technology that serves well those who adopt it. It saves them both time and energy.

Improvements in causal mapping [19,20], and SD modeling and analysis [50,57] contribute to the SdP with SD trend (Sect. "Future Directions"). Behavioral decision theory and cognitive science also help translate the knowledge of managers into SD models. The emphasis remains on small, transparent models of strategic situations and on dialogue between the managers' mental models and the computed scenarios [53].

All prognosticators are bloody fools
—Winston Churchill

Environmental Turbulence and Future Uncertainty

Environmental Turbulence

Abundant frameworks describe the business environment, but the one by Emery and Trist [22], which Duncan [17] abridged, has been guiding many a strategic thinker. It shows four business environments, each more complex and troublesome for the firm than the preceding one (Fig. 1a).

1. *Placid* or *independent-static environment*: infrequent changes are independent and randomly distributed,

i. e., IID. Surprises are rare, but no new major opportunities to exploit either (*cell* 1, Fig. 1a).
2. *Placid-clustered* or *complex-static environment*: patterned changes make forecasting crucial. Comparable to the economist's idea of imperfect competition, this environment lets firms develop distinctive competencies to fit limited opportunities that lead to growth and bureaucracy (*cell* 2, Fig. 1a).
3. *Disturbed-reactive* or *independent-dynamic environment*: firms might influence patterned changes. Comparable to oligopoly in economics, this environment makes changes difficult to predict, so firms increase their operational flexibility through decentralization (*cell* 3, Fig. 1a).
4. *Turbulent field* or *complex-dynamic environment*: most frequent, changes are also complex, i. e., interdependent, originating both from autonomous shifts in the environment and from interdependence among firms and conglomerates. Social values accepted by members guide strategic response (*cell* 4, Fig. 1a).

Ansoff and McDonnell [7] extend the dichotomous environmental uncertainty perceptions by breaking turbulent environments (*cell* 4, Fig. 1a) into *discontinuous* and *surprising*. This is a step in the right direction, but not as helpful as a causal model specific to the system structure of a firm's strategic situation. Assuredly, 2×2 typologies help clarify exposition and are most frequent in the organization theory and strategy literatures. The mystical significance of duality affected even Leibniz, who associated one with God and zero with nothingness in the binary system. The generic solutions that dichotomies provide leave out the specifics that decision makers need. No matter what business they are in (Fig. 1b), managers cannot wait until a better theory comes along; they must act now.

It is worth noting that people often confuse the term 'complex' with 'complicated'. Etymology shows that *complicated* uses the Latin ending *-plic*: *to fold*, but *complex* contains the Greek root *πλέξ-* '*plēx-*': *to weave*. A complicated structure is thereby folded, with hidden facets stuffed into a small space (Fig. 1c). But a complex structure has interwoven parts with mutual interdependencies that cause dynamic complexity [46]. Remember: complex is the opposite of independent or untwined (Fig. 1a) and complicated is the opposite of simple (Fig. 1c).

Daft and Weick's [12] vista on firm *intrusiveness* and *environmental equivocality* is pertinent here. They see many events and trends in the environment as being inherently unclear. Managers discuss such events and trends, and form mental models and visions expressed in a fuzzy language and label system [80]. Within an *enactment* process, equivocality relates to managerial assumptions underlying the *analyzability* of the environment. A firm's *intrusiveness* determines how *active* or *passive* the firm is about environmental scanning. In this context, as the global environment gets turbulent, active firms and their subsidiaries construct SdP with SD models and compute scenarios to improve performance.

Managers of active firms combine knowledge acquisition with interpretations about the environment and their strategic situation. They reduce equivocality by assessing alternative futures through computed scenarios. In frequent meetings and debates, some by videoconferencing, managers use the dialectical inquiry process for *s*trategic *a*ssumption *s*urfacing and *t*esting (SAST), a vital strategic loop. Often ignored, the SAST loop gives active firms a strategic compass [47].

Conversely, passive firms do not actively seek knowledge but reduce equivocality through rules, procedures and regular reports: reams of laser-printed paper with little or no pertinent information. Managers in passive firms use the media to interpret environmental events and trends. They obtain insight from personal contacts with significant others in their environment. Data are personal and informal, obtained as the opportunity arises.

Future Uncertainty

"If we were omnipotent", says Ackoff, then we could get "perfectly accurate forecasts" (see p. 60 in [2]). Thank God the future is unpredictable and we must yet create it. If it were not, then life would have been so boring! Here are some facts about straight forecasting:

1. Forecasts are seldom perfect, in fact, they are always wrong, so a useful forecasting model is one that minimizes error.
2. Forecasts always assume underlying stability in systems.
3. Product family and aggregated forecasts are always more accurate than single product forecasts, so the large numbers law applies.
4. In the short-term, managers can forecast but cannot act because time is too short; in the long term, they can act but cannot forecast.

To offset conundrum #4, SdP with SD juxtaposes the decomposition of performance dynamics into the growth and decline archetypes caused by *balancing* (–) and *reinforcing* (+) recursive causal-link chains or *feedback loops* [33,50]. A thermostat is a typical example of a goal-seeking feedback loop that causes either balancing growth

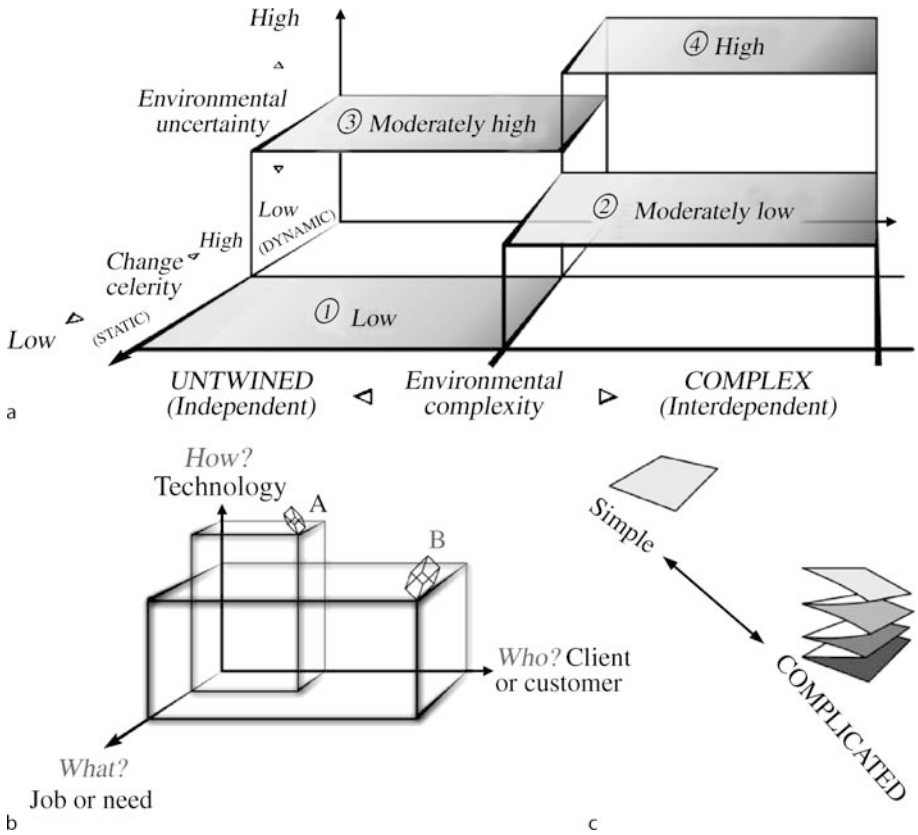

Scenario-Driven Planning with System Dynamics, Figure 1

a Environmental complexity and change celerity dimensions that cause perceived environmental uncertainty (adapted from [32]). b Scenario-driven planning with system dynamics helps with strategy-design fundamentals, such as, for example, defining a business along the requisite client-job-technology three-dimensional grid. c The simple-complicated dimension must *not* be confused with the environmental complexity dimension (adapted from [46])

or decline. The gap between desired and room temperature causes action, which alters temperature with a time lag or delay. Temperature changes in turn close the gap between desired and room temperature.

Conversely, a typical loop that feeds on itself to cause either exponential growth or decline is that of an arms race. One side increases its arms. The other sides increase theirs. The first side then reacts by increasing its arms, and so on. Price wars between stores, promotional competition, shouting matches, one-upmanship and the wildcard interest rates of the late 1970s are good examples too. Escalation might persist until the system explodes or outside intervention occurs or one side quits, surrenders or goes out of business. In the case of wildcard interest rates, outside intervention by a regulatory agency can bring an end to irrationally escalating rates.

We've never been here before
—Peter Senge

SdP with SD: The Modeling Process ≡ Strategic Situation Formulation

The strategic management process (SMP, Fig. 2) starts with environmental scanning, in order to gauge environmental trends, opportunities and threats. Examples include increasing rivalry among existing competitors and Porter's [62] emphasis on the bargaining power of buyers and suppliers as well as on the threats of new entrants and substitutes. Even if some firms reduce environmental scanning to industry analysis in practice, changes in the environment beyond an industry's boundaries can determine what happens within the industry and its entry, exit and inertia barriers. Internal capability analysis comes next. It examines a firm's past actions and internal policy levers that can both propel and limit future actions. The integrative perspective of the SdP with SD framework on Fig. 2 delineates processes that enhance institutional learning, bolster productivity and improve performance through strategic flexibility.

Strategy design begins by identifying variables pertinent to a firm's strategic situation, along with their interrelated causal links. Changes in these variables can have profound effects on performance. Some of the variables belong to a firm's external environment. Examples are emerging new markets, processes and products, government regulations and international interest and currency rates. Changes either in these or their interrelated causal links determine a firm's performance through time.

It is a manager's job to understand the causal links underlying a strategic situation. SdP with SD helps anticipate the effects of future changes triggered in the external environment. Other variables are within a firm's control. Pulling or pushing on these internal levers also affects performance. To evaluate a change in strategy, one must look at potential results along with changes in the environment, matching resource capabilities, stakeholder purposes, and organizational goals and objectives (Fig. 2).

Most variables interact. Often, the entire set of possible outcomes is obscure, difficult to imagine. But if managers oversimplify, then they end up ignoring the combined effects of chain reactions. Even well-intended rationality often leads to oversimplification, which causes cognitive biases (CBs) that mislead decision makers [21,70,72]. Conversely, computing mixed environmental and decision scenarios that link internal and external metrics can reveal unwarranted simplification.

SdP with SD integrates business intelligence with strategy design, not as a narrow specialty, but as an admission of limitations and environmental complexity. It also uses multiperspective dialectics, crucial for strategic assumption surfacing and testing (SAST). Crucial because the language and labels managers use to coordinate strategic real options are imprecise and fuzzy. Fuzzy language is not only adequate *initially* for managing interdependence-induced uncertainty but required [80]. Decision makers rely on it to overcome psychological barriers and Schwenk's [70] groups of CBs.

The best-case scenario for a passive firm is to activate modeling on Fig. 2, sometimes unknowingly. When its managers boot up, for example, electronic spreadsheets that contain inside-out causal models, with assumptions hidden deeply within many a formula. At bootup, only the numbers show. So passive-firm managers use electronic spreadsheets to laser-print matrices with comforting numbers. They

> "twiddle a few numbers and diligently sucker themselves into thinking that they're forecasting the future" [69].

And that is only when rapid changes in the environment force them to stop playing *blame the stakeholder*. They stop fighting the last war for a while, artfully name the situation a crisis, roll up their sleeves, and chat about and argue, but quickly agree on some arbitrary interpretation of the situation to generate strategic face-saving options. Miller and Friesen (see pp. 225–227 in [48]) show how for futile firms, rapid environmental changes lead to crisis-oriented decisions. Conversely, successful firms look far into the future as they counter environmental dynamism through strategy design with real options. Together, their options and interpretation of the environment, through the consensus that SdP with SD facilitates, enable a shared logic to emerge: a shared mental model that filters hidden spreadsheet patterns and heroic assumptions clean and clear.

Managers of active firms enter the SdP with SD loop of Fig. 2 both consciously and conscientiously. They activate strategic intelligence via computed scenarios and the SAST loop. Instead of twiddling spreadsheet numbers, *pro*active firm managers twiddle model assumptions. They stake, through SD model diagrams, their intuition about how they perceive the nature and structure of a strategic situation. Computed scenarios quantitatively assess their perceived implications. Having quantified the implications of shared visions and claims about the structure of the strategic situation, managers of active firms are likely to reduce uncertainty and equivocality. Now they can manage strategic interdependence. Because articulated perception is the starting point of all scenarios, computed scenarios give active firms a fair chance at becoming fast strategic learners.

The design of action or implementation tactics requires detailing how, when and where a strategy goes into action. In addition to assuming the form of *pure communication* (III: 1 and 2, Fig. 2) or *pure action* (III: 3 and 4, Fig. 2), in a pragmatic sense, tactics can be either cooperative or competitive and defensive or offensive. Market location tactics, for example, can be either offensive, trying to rob market share from established competitors, or defensive, preventing competitors from stealing one's market share. An offensive tactic takes the form of frontal assault, flanking maneuver, encirclement, bypass attack or guerilla warfare. A defensive tactic might entail raising structural barriers, increasing expected retaliation or lowering the inducement for future attack. Conversely, cooperative tactics try to gain mutual advantage by working with rather than against others. Cooperative tactics take the form of alliances, joint ventures, licensing agreements, mutual service consortia and value-chain partnerships, the co-location of which often creates industrial districts [29].

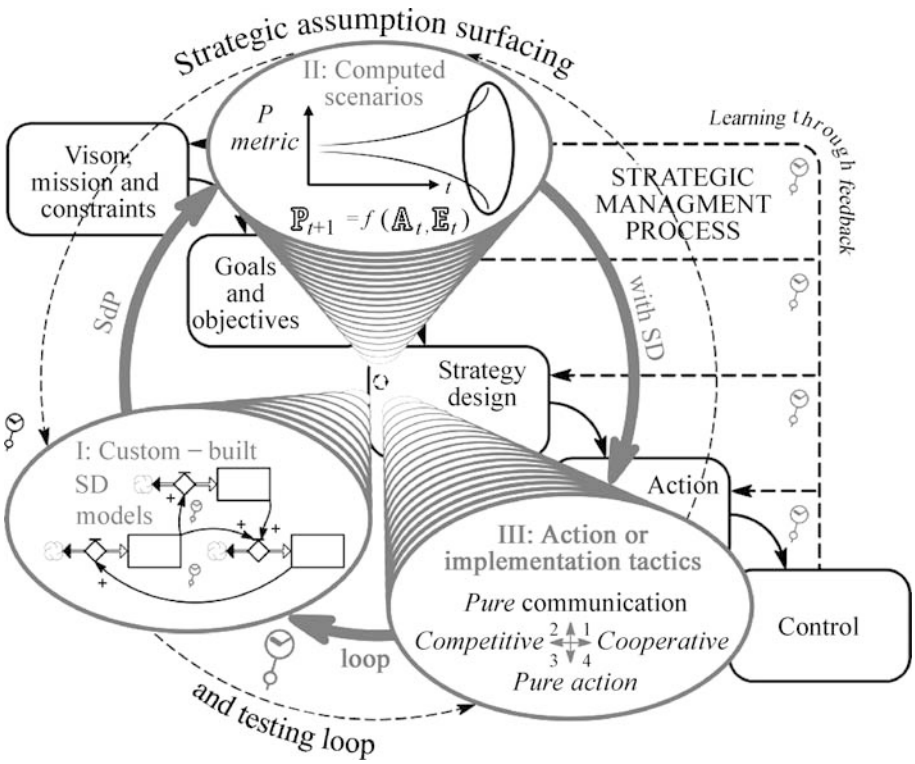

Scenario-Driven Planning with System Dynamics, Figure 2
Cones of resolution show how scenario-driven planning with system dynamics enhances the strategy design component of the strategic management process (SMP; adapted from [32])

The usual copycat strategy retort shows linear thinking at best and clumsy *benchmarking*, also known as shadow marketing, at worst. Its proponents assume performance can improve *incrementally*, with disconnected tactics alone, when strategy design is of primary concern. Piecemeal tactics can undermine strategy, but they are secondary. It might be possible to improve performance through efficient tactics, but is best to design strategies that expel counterproductive tactics. Counterproductive tactics examples are coercive moves that increase rivalry, without a real payoff, either direct or indirect, for the industry incumbent who initiates them. It is atypical of an industry or market leader to initiate such moves.

In strategy, superb action demands superior design. According to the design school, which Ansoff, Channon, McMillan, Porter, Thomas and others lead, logical incrementalism may help implementation, but becomes just another prescription for failure when the environment shifts. Through its judicious use of corporate resources, SdP with SD makes the tactics required for action clear. Also, it reveals their proper coalignment through time, so a firm can build strategic flexibility and save time!

The Modeling Process ≡ Strategic Situation Formulation

SdP with SD (Fig. 2) begins by modeling a business or 'social process' than a business or 'social system'. It is more productive to identify a *social process* first and then seek its causes than to slice a chunk of the real world and ask what dynamics it might generate. Distinguishing between a *social system* and a social process is roughly equivalent to distinguishing between a system's underlying causal structure and its dynamics. Randers (see p. 120 in [63]) defines a social system as a set of cause and effect relations. Its structure is a causal diagram or map of a real-world chunk. A social process is a behavior pattern of events evolving through time. The simulation results of SdP with SD models show such chains of events as they might occur in the real world. An example of a social system (structure) is the set of rules and practices that a firm might enact when dealing with changes in demand, along with the communication channels used for transmitting information and decisions. A corresponding social process (dynamics) might be the stop-and-go pattern of capital in-

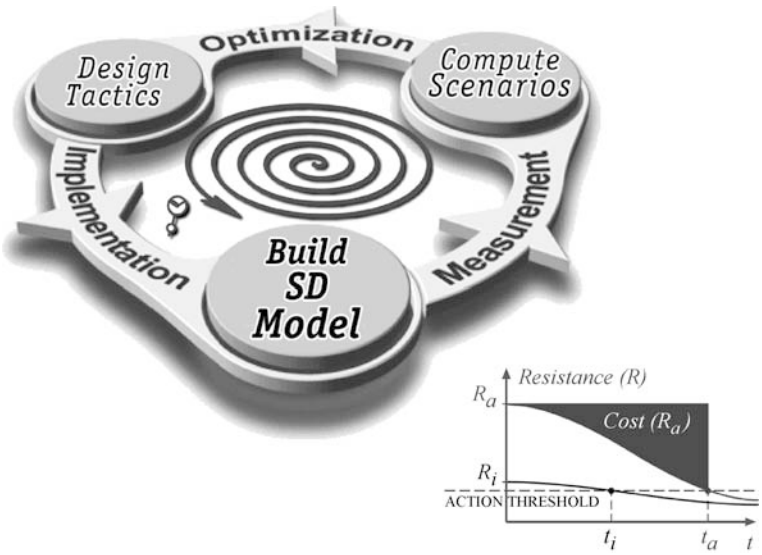

Scenario-Driven Planning with System Dynamics, Figure 3

The recursive nature of the modeling process that scenario-driven planning with system dynamics entails a creates a sustainable, ever-expanding vortex of insight and wisdom, needed in strategic real-options valuation, and b saves both time and money as it renders negligible the cost of resistance (*R*) to change

vestment caused by a conservative bias in a firm's culture.

In his model of a new, fast-growing product line, for example, Forrester [24] incorporates such a facet of corporate culture. Causing sales to stagnate, considerable back orders had to accumulate to justify expansion because the firm's president insisted on personally controlling all capital expenditures.

People often jump into describing system structure, perhaps because of its tangible nature as opposed to the elusive character of dynamics or social process fragments. Also, modelers present model structure first and then behavior. Ultimately, the goal in modeling a strategic situation is to link system structure and behavior. Yet, in the early stages of modeling is best to start with system dynamics and then seek underlying causes. Indeed, SD is particularly keen in understanding system performance, "not structure per se" (see p. 331 in [56]), in lieu of SD's core tenet that structure causes performance.

The modeling process itself is recursive in nature. The path from real-world events, trends and negligible externalities to an effective formal model usually resembles an expanding spiral (Fig. 3a). A useful model requires conceptualization; also focusing the modeling effort by establishing both the time horizon and the perspective from which to frame a decision situation. Typically, strategy-design models require a long-term horizon, over which computed scenarios assess the likely effects of changes both in strategy and in the environment.

Computer simulation is what makes SdP with SD models most useful. Qualitative cause and effect diagrams are too vague, tricky to simulate mentally. Produced through knowledge elicitation, their complexity vastly exceeds the human capacity to see their implications. Casting a chosen perspective into a formal SdP with SD model entails postulating a detailed structure; a diagramming description precise enough to propagate images of alternative futures, i. e., computed scenarios, "though not necessarily accurate" (see p. 118 in [63]). But the modeling process must never downplay the managers' mental database and its knowledge content. Useful models always draw on that mental database [24].

Following Morecroft [53], SdP with SD adopters might strive to replace the notion of modeling an objectively singular world *out there*, with the much softer approach of building formal models to improve managers' mental models. The expanding spiral of Fig. 3a shows that the insight required for decisive action increases as the quantity of information decreases, by orders of magnitude. The required quantification of the relations among variables pertinent to a strategic situation changes the character of the information content as one moves from mental to written to numerical data. Perceptibly, a few data remain, but much more pertinent to the nature and structure of the situation. Thanks to computed scenarios, clarity rules in the end. And, if the modeling process stays *i*nteractive (i), as opposed to *a*ntagonistic (a), then clarity means low resistance to change ($R_i < R_a$, Fig. 3b), which helps reach

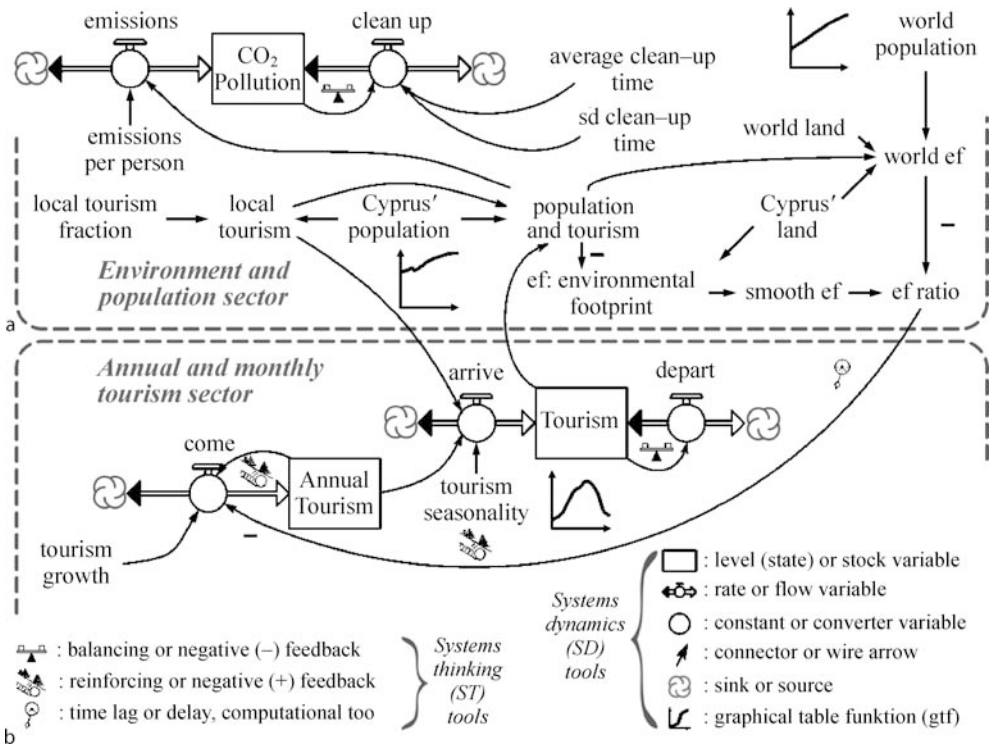

Scenario-Driven Planning with System Dynamics, Figure 4
Cyprus' a environment and population, and b annual and monthly tourism model sectors (adapted from and extending [30])

a firm's action/implementation threshold quickly ($t_i < t_a$, Fig. 3b). This is how SdP with SD users build strategic flexibility while they save both time and money!

Case 1: Cyprus' Environment and Hotel Profitability

Cyprus' Hotel Association wished to test how Cyprus' year 2010 official tourism strategy might affect tourist arrivals, hotel bed capacity and profitability, and the island's environment [30]. Computed with a system dynamics simulation model, four tourism growth scenarios show what might happen to Cyprus' tourism over the next 40 years, along with its potential effects on the sustainability of Cyprus' environment and hotel profitability. Following is a partial description of the system dynamics model that precedes its dynamics.

Model Description (Case 1)

The SD model highlights member interactions along Cyprus' hotel value chain. The model incorporates a generic value-chain management structure that allows modeling customer-supplier value chains in business as well as in physical, biological and other social systems. Although the structure is generic, its situation specific parameters faithfully reproduce the dynamic behavior patterns seen in Cyprus' hotel value-chain processes, business rules and resources.

Cyprus' Environment, Population and Tourism Model Sectors Within Cyprus' environment and population sector (Fig. 4a), the carbon dioxide (CO_2) pollution stock is the accumulation of Cyprus' anthropogenic emissions less the Mediterranean Sea region's self clean-up rate. The clean-up rate that drains Cyprus' CO_2 pollution depends on the level of anthropogenic pollution itself as well as on the average clean-up time and its standard deviation (sd). Emissions that feed CO_2 pollution depend on Cyprus' population and tourism and on emissions per person [9].

In SD models, rectangles represent stocks, i. e., level or state variables that accumulate through time, e. g., the Tourism stock on Fig. 4b. The double-line, pipe-and-valve-like icons that fill and drain the stocks, often emanating from cloud-like sources and ebbing into cloud-like sinks, represent material flows that cause the stocks to change. The arrive rate of Fig. 4b, for example, shows tourists who flow into the tourism stock per month. Single-line arrows represent information flows, while plain text or circular icons depict auxiliary con-

stant or converter variables, i. e., behavioral relations or decision points that convert information into decisions. Changes in the tourism stock, for example, depend on annual tourism, adjusted by tourism seasonality. Both the diagram on Fig. 4a and Table 1 are reproduced from the actual simulation model, first built on the glass of a computer screen using the diagramming interface of *iThink*® [67], and then specifying simple algebraic equations and parameter values. Built-in functions help quantify policy parameters and variables pertinent to Cyprus' tourism situation.

There is a one-to-one correspondence between the model diagram on Fig. 4a and its equations (Table 1). Like the diagram, the equations are the actual output from *iThink*® too. The equations corresponding to Fig. 8b are archived in [30]. Together, Cyprus' population, local tourism and monthly tourism determine the population and tourism sum (Eq. 1.11, Table 1). According to CYSTAT [11], both *Cyprus' Tourism Organization* and its government attach great importance to local tourism. A study on domestic tourism conducted in 1995 revealed that about 46 percent of Cypriots take long holidays. Of these, 61 percent take long holidays exclusively in Cyprus and eight percent in Cyprus and abroad, while 31 percent chose to travel abroad only. These are precisely the percentages in the model (Eq. 1.10, Table 1).

On Fig. 4a, the world land and population data, minus Cyprus' land, population and tourism co-determine the world EF (environmental footprint, Eq. 1.14, Table 1). Compared to Cyprus' smooth EF, i. e., the smooth ratio of the island's free land divided by its total population and tourism, the world EF gives a dynamic measure of Cyprus' relative attractiveness to the rest of the world. The EF ratio (Eq. 1.6, Table 1), i. e., the ratio of Cyprus' smooth EF (Eq. 1.13, Table 1) divided by the world EF (Eq. 1.14, Table 1), assumes that the higher this ratio is, the more attractive the island is to potential tourists, and vice versa. The EF ratio, which depends on Cyprus' total population and tourism, feeds back to the island's annual tourism via the inflow of foreign visitors who come to visit Cyprus every year (Fig. 4b).

The logistic or Verhulst growth model, after François Verhulst who published it in 1838 [66], helps explain Cyprus' actual annual tourism, a quantity that cannot grow forever (Fig. 4b). Every system that initially grows exponentially eventually approaches the carrying capacity of its environment, whether it is food supply for moose, the number of people susceptible to infection or the potential market for a good or a service. As an 'autopoietic' system approaches its limits to growth, it goes through a non-linear transition from a region where positive feedback dominates to a negative feedback dominated regime. S-shaped growth often results: a smooth transition from exponential growth to equilibrium.

The logistic model conforms to the requirements for S-shaped growth and the ecological idea of carrying capacity. The population it models typically grows in a fixed environment, such as Cyprus' foreign annual tourism has done since 1960 up to 2000. Initially dominated by positive feedback, Cyprus' annual tourism might soon reach the island's carrying capacity, with a nonlinear shift to dominance by negative feedback. While accounting for Cyprus' tourism lost to the summer of 1974 Turkish invasion, officially a very long 'military intervention', further depleting annual tourism is the outflow of Cyprus' visitors (not shown here) who might go as the island's free area reaches its Carrying Capacity, estimated at seventy times the number of Cyprus' visitors in 1960 [30].

Cyprus' Hotel Association listed Cyprus tourism seasonality as one of its major concerns. At the time of this investigation, CYSTAT [11] had compiled monthly tourism data for only 30 months. These were used for computing Cyprus' tourism seasonality (Fig. 4b). Incorporating both the foreign annual tourism and the monthly tourism stocks in the model allows both looking at the big picture of annual tourism growth and assessing the potential long-term effects of tourism seasonality on the sustainability of Cyprus' environment and hotel EBITDA, i. e., *e*arnings *b*efore *i*nterest, *t*axes, *d*epreciation and *a*mortization. The publicly available actual annual tourism data allow testing the model's usefulness, i. e., how faithfully it reproduces actual data between 1960 and 2000 [30].

Cyprus' foreign visitors and local tourists arrive at the island's hotels and resorts according to Cyprus' tourism seasonality, thereby feeding Cyprus' monthly tourism stock. About 11.3 days later, according to CYSTAT's [11] estimated average stay days, both foreign visitors and local tourists depart, thereby depleting the monthly tourism stock. By letting tourism growth = 0 and Cyprus' tourism seasonality continue repeating its established pattern, the model computes a zero-growth or *base-run* scenario. Subsequently, however, tourism growth values other than zero initiate different scenarios.

Cyprus' Tourism Growth Scenarios (Case 1)

What can Cyprus' hoteliers expect to see in terms of bottom-line dynamics? According to the four tourism-growth scenarios computed on Fig. 5, seasonal variations notwithstanding, the higher Cyprus' tourism growth is, the lower hotel EBITDA (smooth hE) is, in the short term. In the long term, however, higher tourism growth yields higher profitability in constant year 2000 prices.

Scenario-Driven Planning with System Dynamics, Table 1

Cyprus' environment and population (and local tourism) model sector (Fig. 4a) equations, with variable, constant parameter and unit definitions

***Level or state variables* (stocks)**	***Eq.* #**
CO_2Pollution(t) = CO_2Pollution($t - dt$) + (emissions − clean up) ∗ dt	(1.1)
INIT CO_2 Pollution = emissions (Based on 1995 gridded carbon dioxide anthropogenic emission data; unit: 1000 metric ton C per one degree latitude by one degree longitude grid cell)	(1.1.1)
***Rate variables* (flows)**	
Emissions = emissions per person ∗ population and tourism (unit: 1000 metric tons C/month)	(1.2)
Cleanup = max(0, CO_2Pollution/average clean − up time) (unit: 1000 metric tons C/month)	(1.3)
***Auxiliary variables and constants* (converters)**	
Average clean − up time = 1200 (Med Sea region average self clean-up time = 100 years; unit: months)	(1.4)
Cyprus' land = If (time ≤ 168) then (9251 ∗ 247.1052) else ((9251 − 3355) ∗ 247.1052) (Cyprus' free land area; unit: acres; 1 km^2 = 247.1052 acres)	(1.5)
EF ratio = smooth EF/world EF (unit: unitless)	(1.6)
EF: environmental footprint = Cyprus' land/population and tourism (unit: acres/person)	(1.7)
Emissions per person = 1413.4/702000/12 (unit: anthropogenic emissions/person/month)	(1.8)
Local tourism = local tourism fraction ∗ Cyprus' population (unit: persons/month)	(1.9)
Local tourism fraction = 0.46 ∗ (0.61 + 0.08) (Percentages based on a 1995 study on domestic tourism; unit: unitless)	(1.10)
Population and tourism = Cyprus' population + Tourism − local tourism (Subtracts local tourists already included in Cyprus' population; unit: persons)	(1.11)
Sd clean − up time = 240 (clean-up time standard deviation = 20 years; unit: months)	(1.12)
Smooth EF = SMTH3 (EF: environmental footprint, 36) (Third-order exponential smooth of EF)	(1.13)
World EF = (world land − Cyprus' land)/(world population − population and tourism) (unit: acres/person)	(1.14)
World land = 36677577730.80 (unit: acres)	(1.15)
Cyprus' population = GRAPH(time/12) (Divided by 12 since these are annual data; unit: persons) (0.00, 493984), (1.00, 498898), (2.00, 496570), (3.00, 502001), (4.00, 505622), (5.00, 509329), (6.00, 512950), (7.00, 516743), (8.00, 520968), (9.00, 525364), (10.0, 529847), (11.0, 534330), (12.0, 539934), (13.0, 546486), (14.0, 552348), (15.0, 526313), (16.0, 516054), (17.0, 515881), (18.0, 518123), (19.0, 521657), (20.0, 526744), (21.0, 532692), (22.0, 538210), (23.0, 544675), (24.0, 551659), (25.0, 558038), (26.0, 560366), (27.0, 568469), (28.0, 572622), (29.0, 578394), (30.0, 587392), (31.0, 598217), (32.0, 609751), (33.0, 619658), (34.0, 626534), (35.0, 632082), (36.0, 636790), (37.0, 641169), (38.0, 645560), (39.0, 649759), (40.0, 653786), (41.0, 657686), (42.0, 661502), (43.0, 665246), (44.0, 668928), (45.0, 672554), (46.0, 676147), (47.0, 679730), (48.0, 683305), (49.0, 686870), (50.0, 690425), (51.0, 693975), (52.0, 697524), (53.0, 701056), (54.0, 704547), (55.0, 707970), (56.0, 711305), (57.0, 714535), (58.0, 717646), (59.0, 720613), (60.0, 723415), (61.0, 726032), (62.0, 728442), (63.0, 730629), (64.0, 732578), (65.0, 734280), (66.0, 735730), (67.0, 736928), (68.0, 737887), (69.0, 738627), (70.0, 739172), (71.0, 739540), (72.0, 739743), (73.0, 739792), (74.0, 739697), (75.0, 739472), (76.0, 739123), (77.0, 738658), (78.0, 738083), (79.0, 737406), (80.0, 737406)	(1.16)
World population = GRAPH(time/12) (Divided by 12 since these are annual data; unit: persons) (0.00, 3e+09), (1.00, 3.1e+09), (2.00, 3.1e+09), (3.00, 3.2e+09), (4.00, 3.3e+09), (5.00, 3.3e+09), (6.00, 3.4e+09), (7.00, 3.5e+09), (8.00, 3.6e+09), (9.00, 3.6e+09), (10.0, 3.7e+09), (11.0, 3.8e+09), (12.0, 3.9e+09), (13.0, 3.9e+09), (14.0, 4e+09), (15.0, 4.1e+09), (16.0, 4.2e+09), (17.0, 4.2e+09), (18.0, 4.3e+09), (19.0, 4.4e+09), (20.0, 4.5e+09), (21.0, 4.5e+09), (22.0, 4.6e+09), (23.0, 4.7e+09), (24.0, 4.8e+09), (25.0, 4.9e+09), (26.0, 4.9e+09), (27.0, 5e+09), (28.0, 5.1e+09), (29.0, 5.2e+09), (30.0, 5.3e+09), (31.0, 5.4e+09), (32.0, 5.4e+09), (33.0, 5.5e+09), (34.0, 5.6e+09), (35.0, 5.7e+09), (36.0, 5.8e+09), (37.0, 5.8e+09), (38.0, 5.9e+09), (39.0, 6e+09), (40.0, 6.1e+09), (41.0, 6.2e+09), (42.0, 6.2e+09), (43.0, 6.3e+09), (44.0, 6.4e+09), (45.0, 6.5e+09), (46.0, 6.5e+09), (47.0, 6.6e+09), (48.0, 6.7e+09), (49.0, 6.8e+09), (50.0, 6.8e+09), (51.0, 6.9e+09), (52.0, 7e+09), (53.0, 7e+09), (54.0, 7.1e+09), (55.0, 7.2e+09), (56.0, 7.2e+09), (57.0, 7.3e+09), (58.0, 7.4e+09), (59.0, 7.5e+09), (60.0, 7.5e+09), (61.0, 7.6e+09), (62.0, 7.6e+09), (63.0, 7.7e+09), (64.0, 7.8e+09), (65.0, 7.8e+09), (66.0, 7.9e+09), (67.0, 8e+09), (68.0, 8e+09), (69.0, 8.1e+09), (70.0, 8.1e+09), (71.0, 8.2e+09), (72.0, 8.3e+09), (73.0, 8.3e+09), (74.0, 8.4e+09), (75.0, 8.4e+09), (76.0, 8.5e+09), (77.0, 8.5e+09), (78.0, 8.6e+09), (79.0, 8.6e+09), (80.0, 8.7e+09)	(1.17)

High tourism growth implies accommodating overbooked hotel reservations for tourists who actually show up. Free cruises erode Cyprus' hotel EBITDA. The alternative is, however, angry tourists going off in hotel lobbies. Tourists have gotten angry at hotels before, but hotels have made the problem worse in recent years worldwide [16]. They have tightened check-in rules, doubled their renovations and increased the rate of overbooking by about 30 percent. The results can be explosive if one adds the record flight delays that travelers endure. Anyhow, free cruises to nearby Egypt and Israel sound much better than simply training employ-

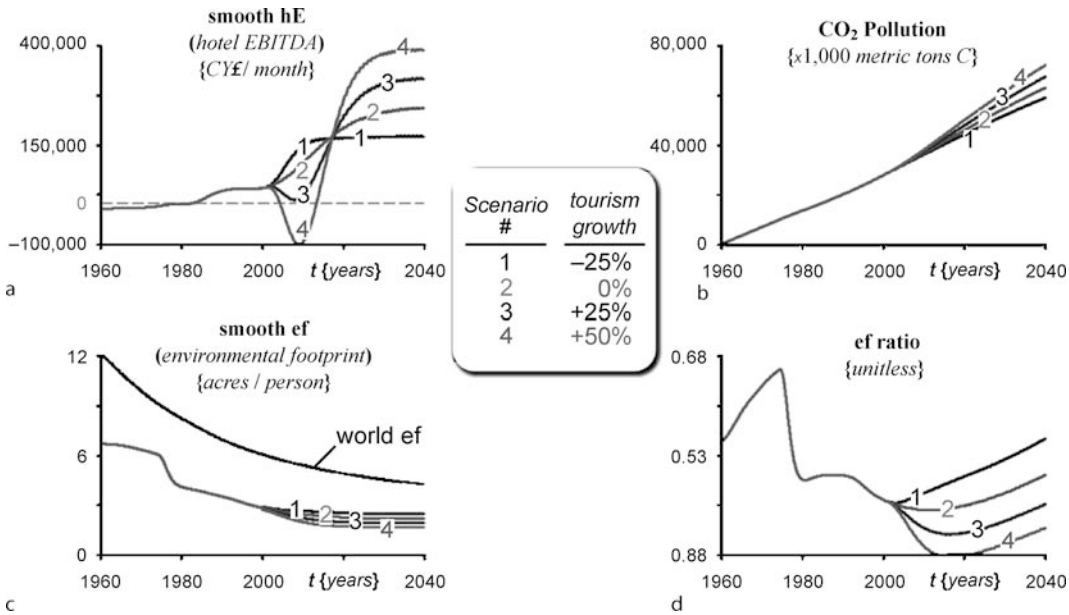

Scenario-Driven Planning with System Dynamics, Figure 5

Four computed scenarios show how tourism growth might affect Cyprus' hotel EBITDA (smooth hE) and the island's environment, with carbon-dioxide (CO_2) pollution (adapted from and extending [30])

ees to handle unhappy guests that scream in hotel lobbies.

Eventually, as Cyprus' bed capacity increases and thereby catches up with tourism demand, there will be less overbooking and a few free cruises to erode Cyprus' hotel EBITDA. Given enough time for an initial bed capacity disequilibrium adjustment, in the long term, high tourism growth increases both hotel EBITDA (Fig. 5a) and cash [30].

In addition to their profound consequences for its hotel value-chain participants, Cyprus' tourism growth might also determine the fate of the island's environment. Depending on the island's population and emissions per person, high tourism growth implies high anthropogenic emissions feeding Cyprus' CO_2 Pollution. Anthropogenic CO_2 emissions attributed to the upward and downward movements of recurring tourist arrivals create much more stress and strain for the island's natural environment than a consistent stream of tourism with low seasonality would. High tourism growth lowers Cyprus' environmental footprint (EF). The summer 1974 Turkish *military intervention* has had a drastic negative effect on Cyprus' relative attractiveness because it reduced the island's free land by 41 percent.

Although qualitatively similar to the world's average EF after the invasion, Cyprus' environmental footprint is lower than the world's average EF (Fig. 5c), rendering the island's free area relatively less attractive as more foreign tourists visit. Manifested in the EF ratio (Fig. 5d), Cyprus becomes relatively less attractive as more visitors choose to vacation on the island's free area.

Qualitatively, Cyprus' CO_2 pollution scenarios (Fig. 5b) look exactly like the A2 scenario family of harmonized anthropogenic CO_2 emissions, which the *Intergovernmental Panel on Climate Change* (IPCC) computed to access the risks of human-induced climate change [54]. Like in the rest of the world, unless drastic changes in policy or technology alter the emissions per person ratio in the next 40 years, CO_2 pollution is expected to grow proportionally with Cyprus' tourism, degrading the island's environment.

Case 2: A Japanese Chemicals Keiretsu (JCK)

Home of NASA's *Johnson Space Center*, the Clear Lake region in Texas boasts strong high technology, biotechnology and specialty chemicals firms. Among them is JCK, whose recent investment helps the Clear Lake region continue its stalwart role in Houston's regional economic expansion [40].

An active member of a famous Japanese giant conglomerate, JCK's history begun in the late 1800s. Despite its long history, however, it has not been easy for JCK to evade the feedback loop that drives Japanese firms to man-

ufacture outside Japan. Since the 1950s, with Japan still recovering from WWII, the better Japanese companies performed, the better their national currency did. But the better Japan's currency did, the harder it became for its firms to export. The higher the yen, the more expensive and, therefore, less competitive Japan's exports become. This simple loop explains JCK's manufacturing lineage from Japan to USA [34].

But the transition process behind this lineage is not that simple. JCK's use of SdP with SD reveals a lot about its strategy design and implementation tactics. The model below shows a tiny fragment of JCK's gigantic effort to reperceive itself. The firm wants to see its keiretsu transform into an agile, virtual enterprise network (VEN) of active agents that collaborate to achieve its transnational business goals. Although still flying low under the media's collective radar screen, VENs receive increased attention by strategic managers [29].

Sterman (see Chap. 17 and 18 in [75]) presents a generic value-chain management structure that can unearth what VENs are about. By becoming a VEN, JCK is poised to bring the necessary people and production processes together to form *autopoietic*, i. e., self-organizing, customer-centric value chains in the specialty chemicals industry. JCK decided to build its own plant in USA because the net present value (NPV) of the anticipated combined cash flow resulting from a merger with other specialty chemicals manufacturers in USA would have been less then the sum of the NPVs of the projected cash flows of the firms acting independently. Moreover, JCK's own technology transfer cost is so low that the internalization cost associated with a merger would far exceed supplier charges plus market transaction costs. To remain competitive [62], JCK will not integrate the activity but offshoot it as a branch of its VEN-becoming keiretsu. The plant will be fully operational in January 2008. In order to maximize the combined *n*et *p*resent *v*alue of *e*arnings *b*efore *i*nterest, *t*axes, *d*epreciation and *a*mortization, i. e., NPV(EBITDA), of its new USA plant and the existing one in Asia, JCK wishes to improve its USA sales revenue before production starts in USA.

JCK's pre-production marketing tactics aim at building a sales force to increase sales in USA. Until the completion of the new plant (Dec. 2007), JCK will keep importing chemicals from its plant in Asia. Once production starts in USA (Jan. 2008), then the flow of goods from Asia to USA stops, the plant in USA supplies the USA market and the flow of goods from USA to Asia begins.

Strategic scenarios are not new to the chemical industry [82]. SdP with SD helps this specialty chemicals producer integrate its business intelligence efforts with strategy design in anticipation of environmental change. Modeling JCK's strategic situation requires a comprehensive inquiry into the environmental causalities and equivocalities that dictate its actions. Computed strategic and tactical scenarios probe the combined consequences of environmental trends, changes in JCK's own strategy, as well as the moves of its current and future competitors. The section below describes briefly how JCK plans to implement its transnational strategy of balanced marketing and production. This takes the form of a system dynamics simulation model, which precedes the interpretation of its computed scenarios.

Model Description (Case 2)

The entire model has multiple sectors, four of which compute financial accounting data. Figure 6a shows the production and sales, and Fig. 6b the total NPV(EBITDA) model sectors. The corresponding algebra is in [34]. While JCK is building its USA factory, its factory in Asia makes and sells all specialty chemicals the USA market cannot yet absorb. This is what the *feed-forward* link from the production in Asia flow to the sales in Asia rate shows. The surplus demand JCK faces in Asia for its fine chemicals accounts for this rather unorthodox model structure. The surplus demand in Asia is the model's enabling *safety valve*, i. e., a major strategic assumption that renders tactical implementation feasible.

With the plant in Asia producing at full capacity before the switch, sales in the USA both depletes the tank in Asian stock and reduces sales in Asia. USA sales depend on JCK's USA sales force. But the size of this decision variable is just one determinant of sales in USA.

Sales productivity depends on many parameters, such as the annual growth before the switch rate of specialty chemicals in USA, the average expected volume a salesperson can sell per month as well as on the diminishing returns that sales people experience after the successful calls they initially make on their industrial customers. B2B or business to business, i. e., industrial marketing, can sometimes be as tough as B2C or business to customer, i. e., selling retail.

Time $t = 30$ months corresponds to January 2008, when the switch time converter cuts off the supply of JCK's chemicals from its plant in Asia. Ready by December 2007, the factory in the USA can supply the entire customer base its USA sales force will have been building for 30 months. As production in the USA begins, the sales in the USA before flow stops draining the tank in Asia and sales in Asia resume to match JCK's surplus demand there. Acting both as a production flow and as a continuous-review

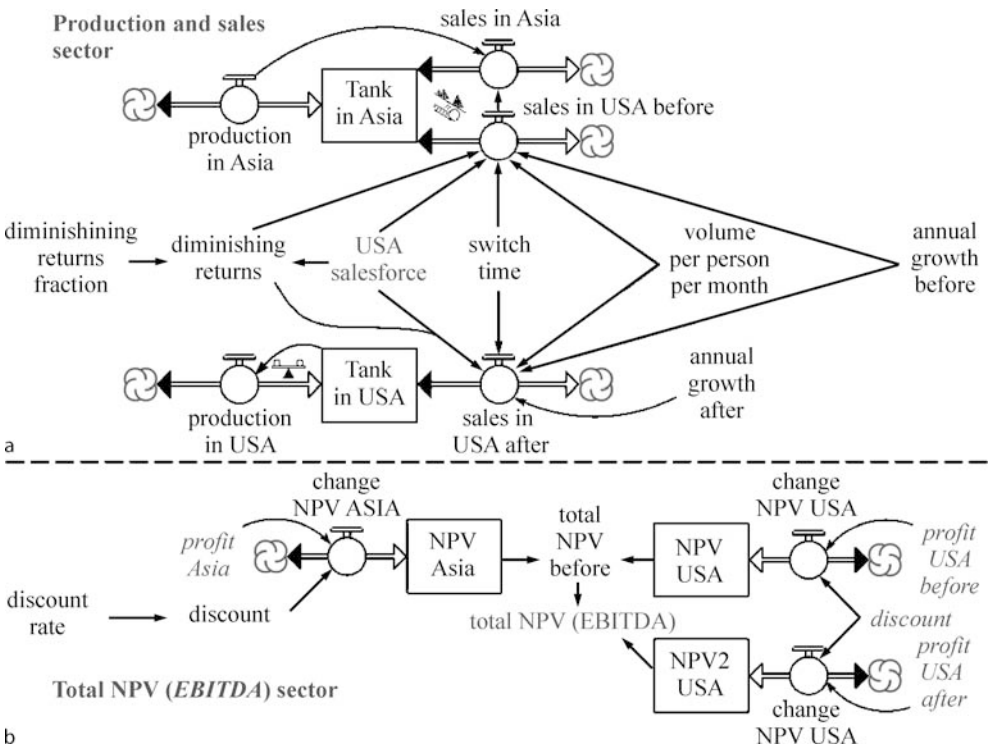

Scenario-Driven Planning with System Dynamics, Figure 6

JCK's a production and sales, and b total NPV(EBITDA) model sectors (adapted from [34]; NPV = net present value, and EBITDA = earnings before interest, taxes, depreciation and amortization)

inventory order point, after January 2008, production in USA feeds the tank in USA stock of the rudimentary value-chain management structure on Fig. 6a.

Value chains entail stock and flow structures for the acquisition, storage and conversion of inputs into outputs, and the decision rules that govern the flows. The jet ski value chain includes, for example, hulls and bows that travel down monorail assembly paths. At each stage in the process, a stock of parts buffers production. This includes the inventory of fiberglass laminate between hull and bow acquisition and usage, the inventory of hulls and bows for the jet ski lower and upper structures, and the inventory of jet skis between dealer acquisition and sales. The decision rules governing the flows entail policies for ordering fiberglass laminate from suppliers, scheduling the spraying of preformed molds with layers of fiberglass laminate before assembly, shipping new jet skis to dealers and customer demand.

A typical firm's or VEN's value chain consists of cascades of supply chains, which often extend beyond a firm's boundaries. Effective value chain models must incorporate different agents and firms, including suppliers, the firm, distribution channels and customers. Scenario-driven planning with system dynamics is well suited for value chain modeling and policy design because value chains involve multiple stock and flow chains, with time lags or delays, and the decision rules governing the flows create multiple feedback loops among VEN members or value- and supply-chain partners (see Chap. 17 and 18 in [75]).

Back to JCK, its tank in the USA feeds information about its level back to production in the USA. Acting first as a decision point, production in the USA compares the tank in the USA level to the tank's capacity. If the tank is not full, then production in the USA places an order to itself and, once the USA factory has the requisite capacity, production in the USA refills the tank in the USA, but only until sales in the USA after the switch drains the tank. Then the cycle begins all over again.

Meanwhile, the profit in Asia, profit in the USA before and profit in the USA after sectors [34] perform all the financial accounting necessary to keep track of the transactions that take place in the value chain production and sales sector (Fig. 6a). As each scenario runs, the profit in Asia, the USA before and the USA after sectors feed the corresponding change in net present value (NPV) flows of

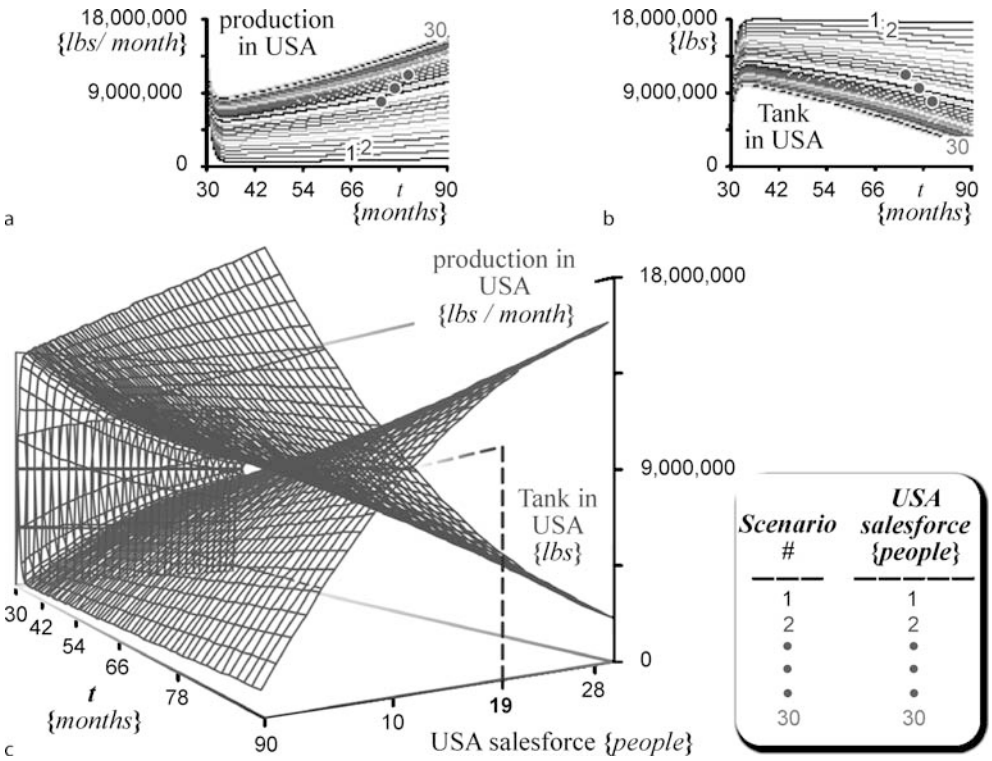

Scenario-Driven Planning with System Dynamics, Figure 7
Thirty computed scenarios show JCK's dual, smooth-switch and profitable purpose in production (adapted from [34])

the model's total NPV(EBITDA) sector (Fig. 6b). By adjusting each profit sector's EBITDA according to the discount rate, the change in NPV flows compute the total NPV(EBITDA) both in Asia and in the USA, both before and after JCK's January 2008 supply switch.

JCK's Computed Scenarios (Case 2)

Recall that the SdP with SD modeling-process spiral enabled our modeling team to crystallize JCK's strategic situation into the cyclical pattern that Fig. 3a shows. Although heavily disguised, the JCK measurement data and econometric sales functions let the system dynamics model compute scenarios to answer that razor-sharp optimization question the JCK executives asked:

> What size a USA sales force must we build in order to get a smooth switch in both sales and production in January 2008, and also to maximize the combined NPV(EBITDA) at our two plants in Asia and USA from now through 2012?

Treating the USA sales force policy parameter in the 'Sensi Specs…' menu item of *iThink*® allowed computing a set of 30 strategic scenarios. The 30 scenarios correspond to JCK's hiring from one to 30 sales people, respectively, to sell specialty chemicals to manufacturing firms in the USA, both before and after the January 2008 switch. Figures 7 and 8 show the 30 computed scenarios.

Figure 7c shows the response surfaces the production in USA rate and tank in the USA stock form after January 2008, in response to the 30 computed scenarios. The computed scenario that corresponds to JCK's building a USA sales force of 19 people achieves a smooth balance between sales in Asia and in the USA. Under this scenario, after January 2008, on the line where the two surfaces cross each other, not only the number of pounds of chemicals made and sold in Asia equals the number of pounds of chemicals made and sold in the USA, but as Fig. 7c shows, production in the USA also equals the tank in USA stock. So hiring 19 sales people now meets JCK's smooth switch in sales and production objective. But how?

How does producing and selling in the USA at rates equal to the corresponding rates in Asia constitute a fair response to JCK's smooth switch objective? The JCK executives seemed to accept this at face value. But our team had to clearly explain the dynamics of JCK's rudimen-

tary USA value chain (Fig. 6a), in order to unearth what the USA member of this transnational VEN-becoming keiretsu might be up to.

It looks simple, but the value chain of the production and sales sector on Fig. 6a can show the same amplification symptoms that much more elaborate value chains show when they fall pray to bullwhip effects. Locally rational policies that create smooth and stable adjustment of individual business units can, through their interaction with other functions and firms, cause oscillation and instability. Figure 8a shows the profound consequences of JCK's switch for its value chain in the USA. Because of the sudden switch in January 2008, the computed scenarios cause 30 sudden step changes. Both variables' adjustment rates increase, but the tank in the USA stock's amplification remains almost constant below 50 percent. As customer demand steps up, so do both metrics' new equilibrium points, but in direct proportion to the step increase in customer demand in the USA.

The 30 computed scenarios confirm Sterman's argument that, while the magnitude of amplification depends on stock adjustment times and delivery lags, its existence does not. No matter how drastically customers and firms downstream in a value chain change an orders' magnitude, they cannot affect supply chain amplification. Value chain managers must never blame customers and downstream firms or their forecasts for bullwhip effects. The production in USA amplification is almost double the tank in USA for a small USA sales force, suggesting that JCK's USA factory faces much larger changes in demand than its sales people do. Although temporary, during its disequilibrium adjustment, the tank in the USA consistently overshoots the new equilibrium points that it seeks after the switch (Fig. 7b), an inevitable consequence of stock and flow structure. Customers are innocent, but JCK's value chain structure is not:

First, the tank in the USA stock adjustment process creates significant amplification of the production in the USA rate. Though the tank in the USA relative amplification is 36.18 percent under the USA sales force = 1 scenario, for example, the relative amplification of production in the USA (Fig. 8a) increases by a maximum of more than 90 percent (the peak production in the USA rate, after t = 30 months, divided by the minimum production in the USA rate $= 11,766,430.01/1,026,107.64 = 91.28$ percent). The *amplification ratio*, i. e., the ratio of maximum change in output to maximum change in input, therefore is $91.28\%/36.18\% = 2.52$. A one-percent increase in demand for JCK's chemicals causes a 2.52 percent surge in demand at JCK's USA plant. While the amplification ratio magnitude depends on the stock adjustment times and delivery lags, its existence does not (see p. 673 in [75]).

Second, amplification is temporary. In the long run, a one-percent increase in sales in the USA after leads to a one-percent increase in production in the USA. After two-adjustment times, i. e., two months, production in the USA gradually drops (Fig. 7a). During the disequilibrium adjustment, however, production in the USA overshoots its new equilibrium, an inevitable consequence of the stock and flow structure of customer-supplier value chains, no matter how tiny or simple they are. The only way the tank in the USA stock can increase is for its inflow production in the USA rate (order rate) to exceed its outflow rate sales in the USA after (Fig. 6a). Within a VEN's or keiretsu's customer-supplier value chain, supply agents face much larger changes in demand than finished-goods inventory, and the surge in demand is temporary.

The computed scenarios show that as the USA sales force increases, the production in the USA's rate of amplification declines because its new long-term equilibrium point is closer to its initial jump in January 2008. Conversely, as the tank in the USA stock's long-term equilibrium point remains consistently high because of the larger USA sales force, its relative amplification begins to rise. Since the two variables' relative amplification moves in opposite directions, eventually, they meet. What a coincidence! They meet above the USA sales force = 19 people. Now, is this not a much better interpretation of the word 'smooth' in fair response to JCK's smooth-switch performance purpose? The answer to JCK is now pertinent to its balancing its value chain in USA. With a USA sales force = 19, JCK's value chain components show equal relative amplification to sudden changes in demand, attaining nothing less than a magnificent amplification ratio = 1. Now that is smooth!

But what of profitability? JCK's polite executives said: "maximize… combined… NPV". In the time domain (Fig. 8b), total NPV(EBITDA) creates intricate dynamics that obscures the USA sales force effect. But the phase plot on Fig. 8c clearly shows a concave down behavior along the USA sales force: USA sales force = 19 maximizes the two plants' combined total NPV(EBITDA).

Future Directions

The above cases show how scenario-driven planning with system dynamics helps control performance by enabling organizational learning and the management of uncertainty. The strategic intelligence system that SdP with SD provides rests on the idea of a collective inquiry, which translates the environmental 'macrocosm' and a firm's

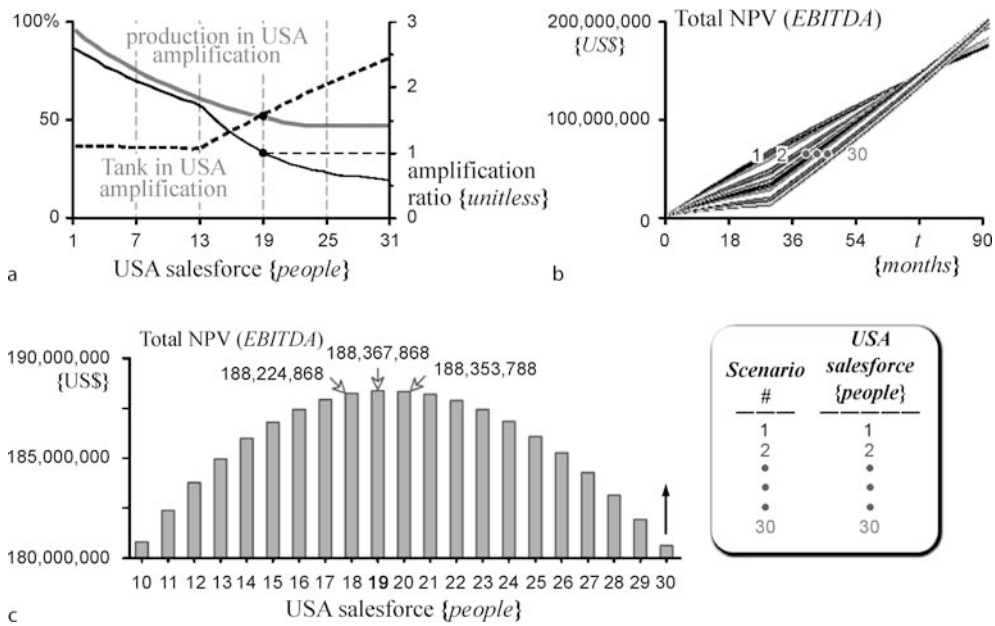

Scenario-Driven Planning with System Dynamics, Figure 8

Thirty computed scenarios show how hiring a sales force of 19 people in the USA might maximize JCK's NPV(EBITDA), and thereby fulfill its dual, smooth-switch and profitable purpose (adapted from [34])

'microcosm' into a shared causal map with computed scenarios. Informed discussion then takes place. Seeing SdP with SD as an inquiry system might help the outcomes of the situation formulation-solution-implementation sequence, each stage built on successive learning.

Strategic situations are complex and uncertain. Because planning is directed toward the future, predictions of changes in the environment are indispensable components of it. Conventional forecasting by itself provides no cohesive way of understanding the effect of changes that might occur in the future. Conversely, SdP with SD and its computed scenarios provide strategic intelligence and a link from traditional forecasting to modern interactive planning systems. In today's quest for managers who are more leaders than conciliators, the strategists' or executives' interest in scenarios must be welcomed. A clearer delineation of SdP with SD might make it a very rich field of application and research.

The SdP with SD inquiry system on Fig. 2 includes several contributions. *First*, by translating the environmental *macrocosm* and the firm *microcosm* into a common context for conceptualization, the requisites of theory building can be addressed. Planning analysts no longer have to operate piecemeal. A theory and a dominant logic typically emerge from shared perceptions about a firm, its environment and stakeholder purposes through model construction.

Second, the outputs of the strategic management process activities build on each other as successive layers. The SAST loop on Fig. 2 follows the counterclockwise direction of multiperspective dialectics [47]. This process allows adjustment of individual and organizational theories and logic, leading to an evolutionary interpretation of the real system that strategic decisions target.

Third, the inquiry system of Fig. 2 enables flexible support for all phases of strategy design. Problem finding or forming, or situation formulation receives equal attention as problem solving.

SdP with SD helps open up the black box of decision makers' mental models, so they can specify the ideas and rules they apply. That in turn helps enrich their language and label system, organizational capability and knowledge, and strategic decision processing system. Computed scenarios bring about transformation rules not previously thought of as well as new variables and interaction paths.

As an entity, each decision maker has a local scope and deals only with specific variables and access paths to other entities. But success factors are not etched in stone. Often, we only observe a representative state of each entity, namely, locally meaningful variables and parts of a scenario. This representativeness changes dynamically in the process of computing scenarios. Beyond the purely technical advantages of computed scenarios, planning becomes interactive, and language and label systems render them-

selves more adequate, effective and precise. Their associated organizational capability develops even more. In addition, the minor and major assumptions in decision makers' mental models surface as computed scenarios specify the conditions under which performance changes.

A line of great immediate concern requires researchers and practitioners alike to explore the modeling process behind SdP with SD. For the sake of realism, to make negotiated perceptions of reality explicit, we need representations where strategic real options and self-interest projections mold the way in which managers incorporate their observations and interpretations into strategy models. This is an unavoidable, most challenging path to tread, if we want to build a dialectical debate into the strategy design process.

Do we really want to? Yes because:

1. The traditional hierarchical organization dogma has been planning, managing and controlling, whereas the new reality of the learning organization incorporates vision, values and mental models. It entails training managers and teams in the IPRD learning cycle conceived by Dewey [14] (cf. Senge and Sterman [71]):

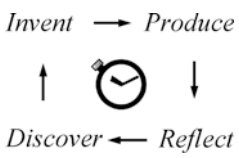

2. In the strategic management process (SMP) evolution, planning is evolving too, from objective-driven to budget-driven to strategy-driven to scenario-driven planning with system dynamics (SdP with SD, see pp. 271–272 in [32]).
3. The inquiry system that mediates the restructuring of organizational *theory in use* [68] determines the quality of organizational learning.

By looking into the dynamics of strategy design and the resulting performance of firms, the SdP with SD framework on Fig. 2 might let managers, planners and business researchers see the tremendous potential of computed strategic scenarios. They might choose to build intelligence systems around SdP with SD to create insight for strategy design. They will be building real knowledge in the process, while developing capability for institutional learning. Both Pascale [59] and de Geus [13] see the capability to speed up institutional learning as a truly sustainable competitive advantage.

Bibliography

Primary Literature

1. Acar W (1983) Toward a theory of problem formulation and the planning of change: Causal mapping and dialectical debate in situation formulation. UMI, Ann Arbor
2. Ackoff RL (1981) Creating the corporate future. Wiley, New York
3. Ackoff RL, Emery FE (1972) On purposeful systems. Aldine-Atherton, Chicago
4. Amara R, Lipinski AJ (1983) Business planning for an uncertain future: Scenarios and strategies. Pergamon, New York
5. Anderson TJ (2000) Real options analysis in strategic decision making: An applied approach in a dual options framework. J Appl Manag Stud 9(2):235–255
6. Ansoff HI (1985) Conceptual underpinnings of systematic strategic management. Eur J Oper Res 19(1):2–19
7. Ansoff HI, McDonnell E (1990) Implanting strategic management, 2nd edn. Prentice-Hall, New York
8. Brauers J, Weber M (1988) A new method of scenario analysis for strategic planning. J Forecast 7(1):31–47
9. Brenkert AL (1998) Carbon dioxide emission estimates from fossil-fuel burning, hydraulic cement production, and gas flaring for 1995 on a one degree grid cell basis. Oak Ridge National Laboratory, Carbon Dioxide Information Analysis Center, Oak Ridge, TN (Database: NDP-058A 2-1998)
10. Christensen CM (1997) The innovator's dilemma: When new technologies cause great firms to fail. Harvard Business School Press, Cambridge
11. CYSTAT (2000) Cyprus key figures: Tourism. The Statistical Service of Cyprus (CYSTAT), Nicosia
12. Daft RL, Weick KE (1984) Toward a model of organizations as interpretation systems. Acad Manag Rev 9:284–295
13. de Geus AP (1992) Modelling to predict or to learn? Eur J Oper Res 59(1):1–5
14. Dewey J (1938) Logic: The theory of inquiry. Holt, Rinehart and Winston, New York
15. Donaldson L (1992) The Weick stuff: Managing beyond games. Organ Sci 3(4):461–466
16. Drucker J, Higgins M (2001) Hotel rage: Losing it in the lobby. Wall Street J (Fri 16 Feb):W1–W7
17. Duncan RB (1972) Characteristics of organizational environments and perceived environmental uncertainty. Adm Sci Q 17:313–327
18. Eberlein RL (2002) Vensim® PLE Software, V 5.2a. Ventana Systems Inc, Harvard
19. Eden C (1994) Cognitive mapping and problem structuring for system dynamics model building. Syst Dyn Rev 10(3):257–276
20. Eden C (2004) Analyzing cognitive maps to help structure issues or problems. Eur J Oper Res 159:673–686
21. Eilon S (1984) The Art of Reckoning: Analysis of performance criteria. Academic Press, London
22. Emery FE, Trist EL (1965) The causal texture of organizational environments. Hum Relat 18:21–32
23. Forrester JW (1958) Industrial dynamics: A major breakthrough for decision makers. Harvard Bus Rev 36(4):37–66
24. Forrester JW (1961) Industrial dynamics. MIT Press, Cambridge
25. Forrester JW (1987) Lessons from system dynamics modeling. Syst Dyn Rev 3(2):136–149

26. Forrester JW (1992) Policies, decisions and information sources for modeling. Eur J Oper Res 59(1):42–63
27. Georgantzas NC (1995) Strategy design tradeoffs-free. Hum Syst Manag 14(2):149–161
28. Georgantzas NC (2001) Simulation modeling. In: Warner M (ed) International encyclopedia of business and management, 2nd edn. Thomson Learning, London, pp 5861–5872
29. Georgantzas NC (2001) Virtual enterprise networks: The fifth element of corporate governance. Hum Syst Manag 20(3):171–188, with a 2003 reprint ICFAI J Corp Gov 2(4):67–91
30. Georgantzas NC (2003) Tourism dynamics: Cyprus' hotel value chain and profitability. Syst Dyn Rev 19(3):175–212
31. Georgantzas NC (2007) Digest® wisdom: Collaborate for win-win human systems. In: Shi Y et al (eds) Advances in multiple criteria decision making and human systems management. IOS Press, Amsterdam, pp 341–371
32. Georgantzas NC, Acar W (1995) Scenario-driven planning: Learning to manage strategic uncertainty. Greenwood-Quorum, Westport
33. Georgantzas NC, Katsamakas E (2007) Disruptive innovation strategy effects on hard-disk maker population: A system dynamics study. Inf Resour Manag J 20(2):90–107
34. Georgantzas NC, Sasai K, Schrömbgens P, Richtenburg K et al (2002) A chemical firm's penetration strategy, balance and profitability. In: Proc of the 20th international system dynamics society conference, 28 Jul–1 Aug, Villa Igiea, Palermo, Italy
35. Gharajedaghi J (1999) Systems thinking: Managing chaos and complexity – A platform for designing business architecture. Butterworth-Heinemann, Boston
36. Godet M (1987) Scenarios and strategic management: Prospective et planification stratégique. Butterworths, London
37. Godet M, Roubelat F (1996) Creating the future: The use and misuse of scenarios. Long Range Plan 29(2):164–171
38. Hax AC, Majluf NS (1996) The strategy concept and process: A pragmatic approach, 2nd edn. Prentice Hall, Upper Saddle River
39. Helmer O (1983) Looking forward. Sage, Beverly Hills
40. Hodgin RF (2001) Clear Lake Area Industry and Projections 2001. Center for Economic Development and Research, University of Houston-Clear Lake, Houston
41. Huss WR, Honton EJ (1987) Scenario planning: What style should you use? Long Range Plan 20(4):21–29
42. Istvan RL (1992) A new productivity paradigm for competitive advantage. Strateg Manag J 13(7):525–537
43. Jarillo JC (1988) On strategic networks. Strateg Manag J 9(1):31–41
44. Jarillo JC, Martínez JI (1990) Different roles for subsidiaries: The case of multinational corporations in Spain. Manag J 11(7):501–512
45. Kahn H, Wiener AJ (1967) The next thirty-three years: A framework for speculation. Daedalus 96(3):705–732
46. Lissack MR, Roos J (1999) The next common sense: The e-manager's guide to mastering complexity. Nicholas Brealey Publishing, London
47. Mason RO, Mitroff II (1981) Challenging strategic planning assumptions. Wiley, New York
48. Miller D, Friesen PH (1983) Strategy making and environment: The third link. Strateg Manag J 4:221–235
49. Millet SM, Randles F (1986) Scenarios for strategic business planning: A case history for aerospace and defence companies. Interfaces. 16(6):64–72
50. Mojtahedzadeh MT, Andersen D, Richardson GP (2004) Using Digest® to implement the pathway participation method for detecting influential system structure. Syst Dyn Rev 20(1):1–20
51. Moore GA (1991) Crossing the chasm. Harper-Collins, New York
52. Morecroft JDW (1985) Rationality in the analysis of behavioral simulation models. Manag Sci 31:900–916
53. Morecroft JDW (1988) System dynamics and microworlds for policymakers. Eur J Oper Res 35:310–320
54. Nakicenovic N, Davidson O, Davis G et al (2000) Summary for policymakers–Emissions scenarios: A special report of working group III of the intergovernmental panel on climate change (IPCC). World Meteorological Organization (WMO) and United Nations Environment Programme (UNEP), New York
55. Nicolis JS (1986) Dynamics of hierarchical systems: An evolutionary approach. Springer, Berlin
56. Oliva R (2004) Model structure analysis through graph theory: Partition heuristics and feedback structure decomposition. Syst Dyn Rev 20(4):313–336
57. Oliva R, Mojtahedzadeh MT (2004) Keep it simple: A dominance assessment of short feedback loops. In: Proc of the 22nd international system dynamics society conference, 25–29 July 2004, Keble College, Oxford University, Oxford UK
58. Ozbekhan H (1977) The future of Paris: A systems study in strategic urban planning. Philos Trans Royal Soc London. 387:523–544
59. Pascale RT (1984) Perspectives on strategy: The real story behind Honda's success. California Manag Rev 26(Spring):47–72
60. Pine BJ-I, Victor B, Boynton AC (1993) Making mass customization work. Harvard Bus Rev 71(5):108–115
61. Porter ME (1985) Competitive advantage: Creating and sustaining superior performance. Free Press, New York
62. Porter ME (1991) Towards a dynamic theory of strategy. Strateg Manag J 12(Winter Special Issue):95–117
63. Randers J (1980) Guidelines for model conceptualization. In: Randers J (ed) Elements of the system dynamics method. MIT Press, Cambridge, pp 117–139
64. Raynor ME (2007) The strategy paradox: Why committing to success leads to failure [And What to Do About It]. Currency-Doubleday, New York
65. Repenning NP (2003) Selling system dynamics to (other) social scientists. Syst Dyn Rev 19(4):303–327
66. Richardson GP (1991) Feedback thought in social science and systems theory. University of Pennsylvania Press, Philadelphia
67. Richmond B et al (2006) iThink® Software V 9.0.2. iSee Systems™, Lebanon
68. Schön D (1983) Organizational learning. In: Morgan G (ed) Beyond method. Sage, London
69. Schrage M (1991) Spreadsheets: Bulking up on data. San Francisco Examiner
70. Schwenk CR (1984) Cognitive simplification processes in strategic decision making. Strateg Manag J 5(2):111–128
71. Senge PM, Sterman JD (1992) Systems thinking and organizational learning: Acting locally and thinking globally in the organization of the future. Eur J Oper Res 59(1):137–150
72. Simon HA (1979) Rational decision making in business organizations. Am Econ Rev 69(4):497–509
73. Singer AE (1992) Strategy as rationality. Hum Syst Manag 11(1):7–21
74. Singer AE (1994) Strategy as moral philosophy. Strateg Manag J 15:191–213

75. Sterman JD (2000) Business dynamics: Systems thinking and modeling for a complex world. Irwin McGraw-Hill, Boston
76. Sterman JD, Wittenberg J (1999) Path dependence, competition and succession in the dynamics of scientific revolution. Organ Sci 10(3, Special issue: Application of complexity theory to organization science, May–Jun):322–341
77. Turner F (1997) Foreword: Chaos and social science. In: Eve RA, Horsfall S, Lee ME (eds) Chaos, complexity and sociology. Sage, Thousand Oaks, pp xi–xxvii
78. Wack P (1985) Scenarios: Uncharted waters ahead. Harvard Bus Rev 63(5):73–89
79. Wack P (1985) Scenarios: Shooting the rapids. Harvard Bus Rev 63(6):139–150
80. Zeleny M (1988) Parallelism, integration, autocoordination and ambiguity in human support systems. In: Gupta MM, Yamakawa T (eds) Fuzzy logic in knowledge-based systems, decision and control. Elsevier Science, North Holland, pp 107–122
81. Zeleny M (2005) Human systems management: Integrating knowledge, management and systems. World Scientific, Hackensack
82. Zentner RD (1987) Scenarios and the chemical industry. Chem Marketing Manag (Spring):21–25

Books and Reviews

Bazerman MH, Watkins MD (2004) Predictable surprises: The disasters you should have seen coming and how to prevent them. Harvard Business School Press, Boston

Bower G, Morrow D (1990) Mental models in narrative comprehension. Science 247(4938):44–48

Mittelstaedt RE (2005) Will your next mistake be fatal? Avoiding the chain of mistakes that can destroy your organization. Wharton School Publishing, Upper Saddle River

Morecroft JDW (2007) Strategic modeling and business dynamics: A feedback systems approach. Wiley, West Sussex

Schnaars SP (1989) Megamistakes: Forecasting and the myth of rapid technological change. The Free Press, New York

Schwartz P (1991) The art of the long view. Doubleday-Currency, New York

Tuchman B (1985) The March of folly: From troy to vietnam. Ballantine Books, New York

Vennix JAM (1996) Group model building: Facilitating team learning using system dynamics. Wiley, Chichester

Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space

Gert Zöller[1], Sebastian Hainzl[2], Yehuda Ben-Zion[3], Matthias Holschneider[1]

[1] Institute of Mathematics and Centre for Dynamics of Complex Systems, University of Potsdam, Potsdam, Germany

[2] GFZ German Research Centre for Geosciences, Potsdam, Germany

[3] Department of Earth Sciences, University of Southern California, Los Angeles, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Modeling Seismicity in Real Fault Regions
Results
Summary and Conclusions
Future Directions
Acknowledgments
Bibliography

Glossary

Bayesian analysis A model estimation technique that accounts for incomplete knowledge. Bayes' theorem is a mathematical formulation of how an a priori estimate of the probability of an event can be updated, if a new information becomes available.

Critical earthquake concept The occurrence of large earthquakes may be described in terms of statistical physics and thermodynamics. In this view, an earthquake can be interpreted as a critical phase transition in a system with many degrees of freedom. The preparatory process is characterized by acceleration of the seismic moment release and growth of the spatial correlation length as in the percolation model. This interpretation of earthquake occurrence is referred to as the critical earthquake process.

Earthquake forecast/prediction The forecast or prediction of an earthquake is a statement about time, hypocenter location, magnitude, and probability of occurrence of an individual future event within reasonable error ranges.

Fault model A fault model calculates the evolution of slip, stress, and related quantities on a fault segment or a fault region. The range of fault models varies from conceptual models of cellular automaton or slider-block type to detailed models for particular faults.

Probability A quantitative measure of the likelihood for an outcome of a random process. In the case of repeating a random experiment a large number of times (e. g. flipping a coin), the probability is the relative frequency of a possible outcome (e. g. head). A different view of probability is used in the → Bayesian analysis.

Seismic hazard The probability that a given magnitude (or peak ground acceleration) is exceeded in a seismic source zone within a pre-defined time interval, e. g. 50 years, is denoted as the seismic hazard.

Self-organized criticality Self-organized criticality (SOC) as introduced by Bak [2] is the ability of a system to organize itself in

the vicinity of a critical point independently of values of physical parameters of the system and initial conditions. Self-organized critical systems are characterized by various power law distributions. Examples include models of sandpiles and forest-fires.

Definition of the Subject

The most fundamental question in earthquake science is whether earthquake prediction is possible. Related issues include the following: Can a prediction of earthquakes solely based on the emergence of seismicity patterns be reliable? In other words, is there a single or several "magic" parameters, which become anomalous prior to a large earthquake? Are pure observational methods without specific physical understanding, like the pattern recognition approach of Keilis–Borok and co-workers [41], sufficient? Taking into account that earthquakes are monitored continuously only for about 100 years and the best available data sets ("earthquake catalogs") cover only a few decades, it seems questionable to forecast earthquakes solely on the basis of observed seismicity patterns. This is because large earthquakes have recurrence periods of decades to centuries; consequently, data sets for most regions include less than ten large events making a reliable statistical testing questionable.

In the studies discussed here, the goal is not to forecast individual earthquakes. Instead, we aim at developing a combined approach based on numerical modeling and data analysis in order to understand seismicity and the emergence of patterns in the occurrence of earthquakes. The discussion and interpretation of seismcity in terms of statistical physics leads to the concept of "critical states", i. e. states in the seismic cycle with an increased probability for abrupt changes involving large earthquakes. A more general goal of this work is to provide perspectives for the understanding of the relevant mechanisms and to give outlines for developments related to time-dependent seismic hazard.

Introduction

Several empirical relationships for the occurrence of seismicity are well-known. The most common one is probably the Gutenberg–Richter law [30] for the relation between frequency and magnitude of earthquakes in a large seismically active region,

$$\log N = a - bM \,, \tag{1}$$

where N is the frequency of earthquakes with magnitude equal to or greater than M; a is a measure of the overall seismicity level in the region and the b value determines the relation between large and small earthquakes. The Gutenberg–Richter law provides an important constraint for the design of physical models and serves as a key ingredient for seismic hazard estimations. Statistical relations for the temporal occurrence of large events are less well known, because the corresponding data records are too short.

Several additional problems exist in the understanding and interpretation of observed seismicity patterns. First, it is important to decide whether an observed pattern has a physical origin or is an artifact, arising for example from inhomogeneous reporting or from man-made seismicity like quarry blasts or explosions [69]. Second, the non-artificial events have to be analyzed with respect to their underlying mechanisms. This leads to an inverse problem with a non-unique solution, which can be illustrated for the most pronounced observed temporal pattern associated with aftershocks. It is empirically known that the earthquake rate $\dot{N}$ after a large event at time t_{M} follows the Omori–Utsu law [49,67]

$$\dot{N} = \frac{K}{(c + t - t_{\mathrm{M}})^p} \,, \tag{2}$$

where t is the time, K and c are constants, and the Omori exponent p is close to unity. In particular, aftershocks are an almost universal phenomenon; that is, they are observed nearly after each mainshock. The underlying mechanisms leading to aftershocks are, however, unknown. Various physical models have been designed to explain aftershock occurrence following Eq. (2). These models include viscoelasticity [32], pore fluid flow [46], damage rheology [9,57], and rate-state friction [24]. The question which mechanism or combination of mechanisms is relevant in a given fault zone remains open. Detailed comparisons of observed and modeled seismicity with respect to the aftershock rate, the duration of aftershock sequences, the dependence on the mainshock size, and other features are necessary to address this problem. Additionally, results from lab experiments on rupture dynamics and satellite observations of deformation provide important information for the design of such models.

Apart from aftershock activity, other seismicity patterns are occasionally associated with observations, including foreshocks [39], seismic quiescence [34,72,78], and accelerating moment release [17,38]. These patterns have been documented in several cases before large earthquakes. They occur, however, far less frequently than aftershocks. For example, foreshocks are known to preceed only 20–30% of large earthquakes [71]. Therefore, their predictive power is questionable. Moreover, it is not clear

whether these patterns can be attributed to physical processes or to random fluctuations in the highly sparse and noisy earthquake catalogs. This problem can be addressed by using fault models which simulate long and complete earthquake sequences over thousands of years. If the models capture the main features of the underlying physics, the occurrence of seismicity patterns can be studied with reasonable statistics. The main ingredients of such models are the geometry of a fault region, empirically known constitutive laws, spatial heterogeneities, and stress and displacement functions following dislocation theory [20,47]. In order to allow for detailed studies of the relations between the imposed mechanisms and the observed seismicity functions, it is important that the number of adjustable parameters is limited.

It is emphasized that these models do not aim to reproduce an observed earthquake catalog in detail. Instead, the main goal is to address questions like: Why is the Parkfield segment of the San Andreas fault characterized by relatively regular occurrence of earthquakes with magnitude $M \approx 6$, while on the San Jacinto fault in California the properties of earthquake occurrence are more irregular? Basic models for seismicity are mainly based on one or more solid blocks, which are driven by a plate over a sliding surface. The plate and the blocks are connected with springs. This model can generate stick-slip events considered to represent earthquakes. The slider-block models can produce a wide range of complexity, beginning with a single block model leading to periodic occurrence of events of uniform size, and progressing to an array of connected blocks [18] leading to complex sequences of events with variable size. In order to reduce the computational effort cellular automata are commonly used [42,48]. Mathematically, these models include maps instead of differential equations; physically, this corresponds to instantaneously occurring slip events, neglecting inertia effects. The main ingredients of slider-block and cellular automaton models are (1) external driving (plate motion), and (2) sudden local change of system parameters (stress), when a critical value (material strength) is reached, followed by an avalanche of block slips (stress drop and coseismic stress transfer during an earthquake). While the first process lasts for years to several hundred years, the second occurs on a time scale of a few seconds. The simplest model including these features has been formulated by Reid [52] and is known as *Reid's elastic rebound theory*; in terms of slider-block models, this corresponds to a single block model with constant plate velocity. Accounting for spatial heterogeneity and fault segmentation, many interacting blocks, or fault segments, have to be considered. This leads to a spatiotemporal stress field instead of a single stress value. In general, the material strength will also become space-dependent. Such a model framework can be treated with the methodology of statistical physics similar to the Ising model or percolation models [43]. In this context, large earthquakes are associated with second-order phase transitions [2,59,64]. The view of earthquakes as phase transitions in a system with many degrees of freedom and an underlying critical point, is hereinafter referred to as the "critical point concept". The period before such a phase transition is expected to be characterized by a preparation process including development of power laws and growing spatial correlation length [14]. However, depending on the parameters of a model, different situations are conceivable: the system trajectory can enter the critical state and the critical point frequently ("supercritical") or it may never becomes critical ("subcritical"). A case of special interest is the class of models [2] showing *self-organized criticality* (SOC), which have their origin in a simple cellular automaton model for a sandpile [3]. In this case the system drives itself permanently to the vicinity of the critical point with almost scale-free characteristics. Consequently, each small event can grow into a large earthquake with some probability [28].

Long simulations of earthquake activity can be used to calculate statistical features like the recurrence time distribution of large earthquakes and the frequency-size distribution with high precision. Despite the scaling behavior (Eq. (1)) in the earthquake magnitudes for small and intermediate earthquakes, which is observed for many sets of model parameters, clear deviations become visible for large magnitudes. Such deviations are known from real catalogs, but their statistical significance is not clear in all cases. The model simulations suggest that deviations from scaling for strong earthquakes can be attributed to physical properties. One important property is the spatial disorder of brittle parameters of the fault. The presence of strong heterogeneities suppresses system-wide events with some probability, whereas such events can evolve more easily on smooth faults. The degree of quenched (time-independent) spatial heterogeneity turns out to be a key parameter for statistical and dynamical properties of seismicity [5,12,80]. This includes the temporal regularity of mainshock occurrence, various properties of the stress and displacement fields, and a spontaneous mode-switching between different dynamical regimes without changing parameters. The degree of heterogeneity can act as a tuning parameter that allows for a continuous change of the model dynamics between the end-member cases of supercritical and subcritical behavior. Such a dependence, which is observed also for other parameters, can be visual-

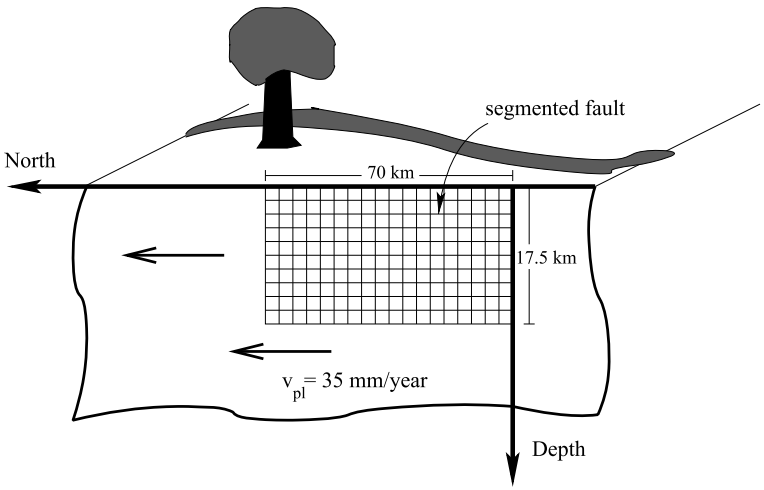

Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 1
Sketch of the fault model framework

ized in phase diagrams similar to the phase diagram for the different aggregate states of water [22,79,80]. For increasing complexity of a model, the number of axes of the phase diagram, representing the relevant model parameters, will increase. The above mentioned question of distinguishing different faults like the Parkfield segment and the San Jacinto fault can be rephrased as the problem of assigning the faults to different regions in such a diagram. An important step in this direction is the physical modeling of observed seismicity patterns, including universal patterns like aftershocks (Eq. (2)), common fluctuations like foreshocks and the acceleration of seismic energy release before large earthquakes. The latter phenomenon which sometimes occurs over large regions including more than one fault, can be interpreted in terms of the approach towards a critical point. This view is supported by an observational study of the growth of the spatial correlation length which is a different aspect of the same underlying physics [73,75,76,77].

The establishment of relationships between model parameters and observational features may be used to tune the model towards a specific fault zone, and use the tuned fault model for practical applications of seismic hazard estimations. Toward this end the recurrence time distribution of large earthquakes is needed. Since observational data records are often short and noisy, the use of Bayesian probability theory is helpful for the estimation of uncertain model parameters, and the incorporation of various types of observational data in seismic hazard estimations. The Parkfield segment, as one of the best monitored seismically-active regions, serves as an excellent natural laboratory for such a case study. A discussed example illustrates how partially known parameters like the stress drop and the seismic hazard can be estimated by combining numerical models and observational data [74].

In Section "Modeling Seismicity in Real Fault Regions", the physical fault model used for the discussed studies is described. Results from numerical simulations are presented in Sect. "Results". A summary is given in Sect. "Summary and Conclusions".

Modeling Seismicity in Real Fault Regions

Numerous frameworks have been used to simulate seismicity (see e. g. [6,9,10,18,32,37,48] and references therein). These include slider-block models, cellular automata, "inherently discrete" fault models where the discreteness is an inherent feature of the imposed physics, and continuum models. In this section we illustrate how a fault model (Fig. 1) can be adjusted in order to simulate seismicity of a real fault region, e. g. the Parkfield segment of the San Andreas fault in California.

Fault Geometry and Model Framework

A first constraint for a specific model is to represent the geometry of the fault segment. As shown in Fig. 2, the region of Parkfield is characterized by a distribution of fault segments, which have generally the same orientation. It is therefore reasonable to map these segments in the model on a plane intersecting the surface at a straight line from SE to NW. The dimensions of the fault segment for modeling (Fig. 1) are chosen to be 70 km in length and 17.5 km in depth. As discussed in [10], this geometry corresponds approximately to the San Andreas fault near Parkfield. The

entire fault is an infinite half-plane, but the brittle processes are calculated on the above rectangular section referred to below as the computational grid. The computational grid is discretized to 128×32 cells of uniform size, where stress and slip are calculated. The size of the cells is not related here to observations; rather it depends on the magnitude range under consideration and the computational effort. The failure of a single cell defines the lowest magnitude. A higher resolution of the grid with same overall dimensions increases the magnitude range, because the magnitude is calculated from the slip of all cells during an earthquake. Following Ben–Zion and Rice [10], the material surrounding the fault is assumed to be a homogeneous elastic half space, which is characterized by elastic parameters and a related Green's function:

1. The elastic properties are expressed by the Lamé constants λ and μ, which connect stress and strain in Hooke's law. For many rocks, these constants are almost equal; therefore we use $\lambda = \mu$ with μ being the rigidity. An elastic solid with this property is called a *Poisson solid*. Because the strain is dimensionless, μ has the same dimension as the stress. In the present study, we use $\mu = 30\,\text{GPa}$.
2. The (static) Green's function $G(\vec{y}_1, \vec{y}_2)$ defines the static response of the half space at a position $\vec{y}_1$ to a displacement at $\vec{y}_2$, which may arise from (coseismic) slip or (aseismic) creep motion. Due to the discretization of the fault plane into computational cells, we use the Green's function for static dislocations on rectangular fault patches of width $\mathrm{d}x$ and height $\mathrm{d}z$, which is given in [20] and [47]. The main difference between this Green's function and the nearest-neighbor interaction of most slider-block models and cellular automata is the infinite-range interaction following a decay according to $1/r^3$, where r is the distance between source cell and receiver point.

Interseismic Processes

The motion of the tectonic plates, indicated in Fig. 2, is responsible for the build-up of stress in the fault zone. Geodetic measurements of surface displacements provide estimates of the velocity of the plates. For the San Andreas fault, a value of $v_{\text{pl}} = 35\,\text{mm/year}$ as a long-term average [55] is widely accepted and is adopted for the model. The displacement $\mathrm{d}u$ in the regions surrounding the grid during a time period Δt is simply $\mathrm{d}u = v_{\text{pl}} \cdot \Delta t$. While the average slip rate $\dot{u}$ is independent of the location of a cell, the stress rate $\dot{\tau}$ depends on space. The assumption that the computational grid is embedded in a half-plane which undergoes constant creep, implies that cells at the boundaries of the grid experience higher load than cells in the center of the grid. The Green's function $G(i, j;\, k, l)$ defines the interaction of points (i, j) and (k, l) in the medium. In particular, the stress response at a position (i, j) to a static change of the displacement field $\mathrm{d}u(k, l)$ is given by

$$\mathrm{d}\tau(i, j) = -\sum_{(k,l)\in\,\text{half space}} G(i, j;\, k, l) \cdot \mathrm{d}u(k, l)\,, \tag{3}$$

where the minus sign stems from the fact that forward (right-lateral) slip of regions around a locked fault segment is equivalent to back (left-lateral) slip of the locked fault segment. Taking into account that

$$\sum_{(k,l)\in\,\text{half space}} G(i, j;\, k, l) = 0\,, \tag{4}$$

Eq. (3) can be written as

$$\tau(i, j;\, t) = -\sum_{(k,l)\in\,\text{half space}} G(i, j;\, k, l)\cdot[u(k, l;\, t) - v_{\text{pl}}t]\,, \tag{5}$$

where $u(k, l;\, t)$ is the total displacement at position (k, l) and time t since the start of the simulation. Because the surrounding regions sustain stable sliding, $u(k, l;\, t) = v_{\text{pl}}t$ for $(k, l) \notin$ grid, the slip deficit outside the fault region vanishes and it is sufficient to perform the summation on the computational grid:

$$\tau(i, j;\, t) = \sum_{(k,l)\in\,\text{grid}} G(i, j;\, k, l)\cdot[v_{\text{pl}}t - u(k, l;\, t)]\,. \tag{6}$$

Equation (6) can be decomposed to a part for the tectonic loading and a residual part for slip on the computational grid. The tectonic loading follows the formula

$$\tau_{\text{load}}(i, j;\, t) = \gamma(i, j) \cdot t \tag{7}$$

with a space-dependent but time-independent loading rate

$$\gamma(i, j) = v_{\text{pl}} \cdot \sum_{(k,l)\in\,\text{grid}} G(i, j;\, k, l)\,. \tag{8}$$

The build-up of stress may be reduced by aseismic creep motion, which is implemented by a local constitutive law corresponding to lab-based dislocation creep [5]:

$$\dot{u}_{\text{creep}}(i, j;\, t) = c(i, j) \cdot \tau^3(i, j;\, t) \tag{9}$$

with space dependent but time-independent creep coefficients $c(i, j)$.

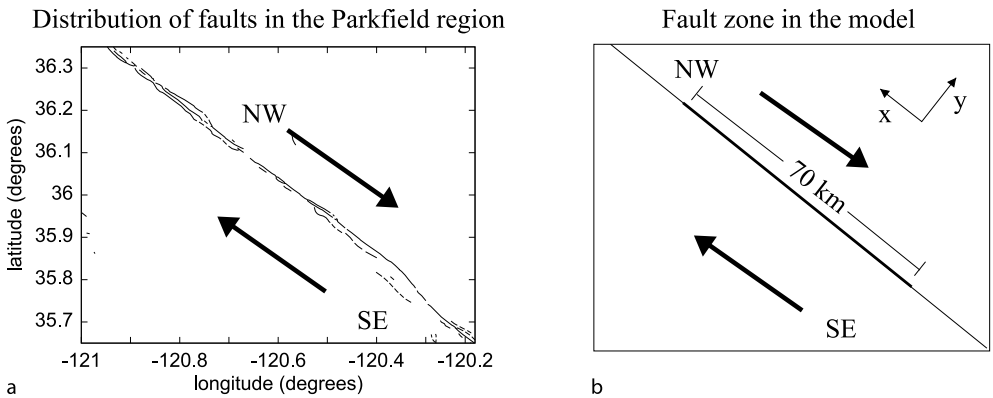

Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 2
a Distribution of faults in the Parkfield (California) region; b fault region in the model

Friction and Coseismic Stress Transfer; Quasidynamic Approach

It is widely accepted that earthquakes on large faults are due to frictional processes on pre-existing structures. The friction is therefore an important empirical ingredient of a fault model [56]. Numerous laboratory experiments have been carried out to characterize frictional behavior of different materials (see e. g. [19]). An important finding is that the friction coefficient defined as the ratio of shear stress τ_{shear} and compressional normal stress τ_{normal}, $\mu_{\text{f}} = \tau_{\text{shear}}/\tau_{\text{normal}}$ at the initiation of slip is approximately constant for many materials; the value of μ_{f} lies between 0.6 and 0.85. This observation, known as *Byerlee's law*, is related to the Coulomb failure criterion [16] for the Coulomb stress CS,

$$CS = \tau_{\text{shear}} - \mu_{\text{f}}\tau_{\text{normal}} \,. \tag{10}$$

The Coulomb stress depends on a plane where shear stress and normal stress are calculated. Neglecting cohesion, the Coulomb criterion for brittle failure is

$$CS \geq \tau_0 \,, \tag{11}$$

which for $CS = 0$ is Byerlee's law.

The North-American plate and the Pacific plate move in opposite directions along the fault plane having strike-slip motion. The absence of normal and thrust faulting reduces the problem to a one-dimensional motion: all parts of the fault move along the fault direction. The stress state of the fault is fully determined by the shear stress τ_{xy} in the coordinates given in Fig. 2b. Slip is initiated if τ_{xy} exceeds $\mu_{\text{f}}\tau_{yy}$. This quantity, which is called the static strength τ_{s} is constant in time if μ_{f} is assumed to be constant. Note that the normal stress on a planar fault in a homogeneous solid does not change [1]. The shear stress τ_{xy} will be denoted simply by τ. In this notation, the failure criterion Eq. (11) reduces to

$$\tau \geq \tau_{\text{s}} \,. \tag{12}$$

When a cell (k, l) fails, the stress drops in this cell to the arrest stress τ_{a}:

$$\tau(k, l) \to \tau_{\text{a}} \,, \tag{13}$$

where the value τ_{a} maybe space-dependent. The stress change produces a corresponding slip

$$\mathrm{d}u(k, l) = \frac{\tau(k, l) - \tau_{\text{a}}}{G(k, l\,;\, k, l)} \tag{14}$$

with the self-stiffness $G(k, l\,;\, k, l)$ of cell (k, l).

The observational effect of dynamic weakening includes also a strength drop from the static strength to a lower dynamic strength:

$$\tau_{\text{s}} \to \tau_{\text{d}} \,. \tag{15}$$

In particular, slipping material becomes weaker during rupture and recovers to the static level at the end of the rupture. This approximation of the strength evolution is known as static-kinetic friction.

The values τ_{s}, τ_{d}, and τ_{a} are connected by the dynamic overshoot coefficient D:

$$D = \frac{\tau_{\text{s}} - \tau_{\text{a}}}{\tau_{\text{s}} - \tau_{\text{d}}} \,, \tag{16}$$

or alternatively by the dynamic weakening coefficient ε:

$$\varepsilon = 1 - \frac{\tau_{\text{d}}}{\tau_{\text{s}}} \,. \tag{17}$$

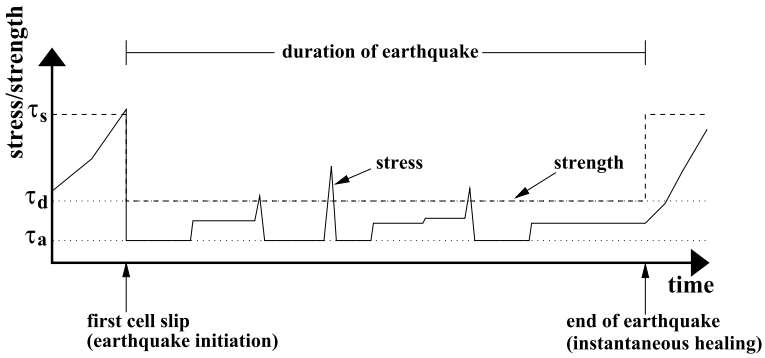

Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 3
Pictorial evolution of stress (*solid line*) and strength (*dashed line*) of a hypocenter cell in the quasidynamic approach

Following [10] we use in most simulations $D = 1.25$.

The redistribution of the stress release $\Delta\tau(k,l) = \tau(k,l) - \tau_a$ from cell (k,l) to a point (i,j) at time t is

$$\mathrm{d}\tau(i,j;\,t) = G(i,j;\,k,l)\cdot\delta\left(t - \frac{r(i,j;\,k,l)}{v_s}\right) \cdot \frac{\Delta\tau(k,l)}{G(k,l\,;\,k,l)}\,, \quad (18)$$

where $\delta(x)$ denotes the δ-function, which is 1 for $x = 0$ and 0 else; v_s is the shear-wave velocity, and $r(i,j;\,k,l)$ is the distance between source cell (k,l) and receiver position (i,j). That is, regions far from the slipping cell receive their stress portion later than regions close to the slipping cell. The value of v_s is assumed to be constant. Each "stress transfer event" associated with Eq. (18) gives a transfer of a stress $\mathrm{d}\tau$ from a source cell (k,l) to a receiver cell (i,j) at time t. This time-dependent stress transfer is called the *quasidynamic* approach in contrast to the *quasistatic* approach used in most similar models.

The evolution of stress and strength in a failing cell is shown schematically in Fig. 3. When the slip is initiated, both the stress and the strength drop. Due to coseismic stress transfer during the event, the cell may slip several times before the earthquake is terminated and instantaneous healing takes place in all cells. The piecewise constant failure envelope (dashed line) indicates static-kinetic friction. A model version with gradual healing was employed by [79]. A review of analytical results associated with the basic model in the context of a large universality class is given in ► Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of.

We note that the Green's function leads to an infinite interaction range. Using open boundary conditions with respect to the computational grid, the stress release from a slipping cell is not conserved on the grid, but on the infinite half plane.

Data

The model produces two types of data, earthquake catalogs and histories of stress and displacement fields. As demonstrated below, all parameters of the model have physical dimensions and can therefore be compared directly with real data, where they are available. This is in contrast to most of the slider-block and cellular automaton models.

Earthquake catalogs include values of the earthquake time, hypocenter, and size. The time of an earthquake is the time of the first slip; the hypocenter is determined by the position of the corresponding cell along strike and depth. The size of an event can be described by different measures: The rupture area A is the total area, which slipped during an earthquake. The potency

$$P = \int u(A)\mathrm{d}A \quad (19)$$

measures [7] the total slip during the event and is related to the seismic moment m_0 by the rigidity: $m_0 = \mu P$. The (moment) magnitude M can be calculated from the potency [10] using

$$M = (2/3)\log_{10}(P) + 3.6\,, \quad (20)$$

where P is given in $\mathrm{cm}\cdot\mathrm{km}^2$.

Results

Numerous simulations of the model described in the previous section have been performed. Firstly, simulations have been examined with respect to the spatiotemporal propagation of stress during single earthquakes ("rupture histories"). Then, long deformation histories have been

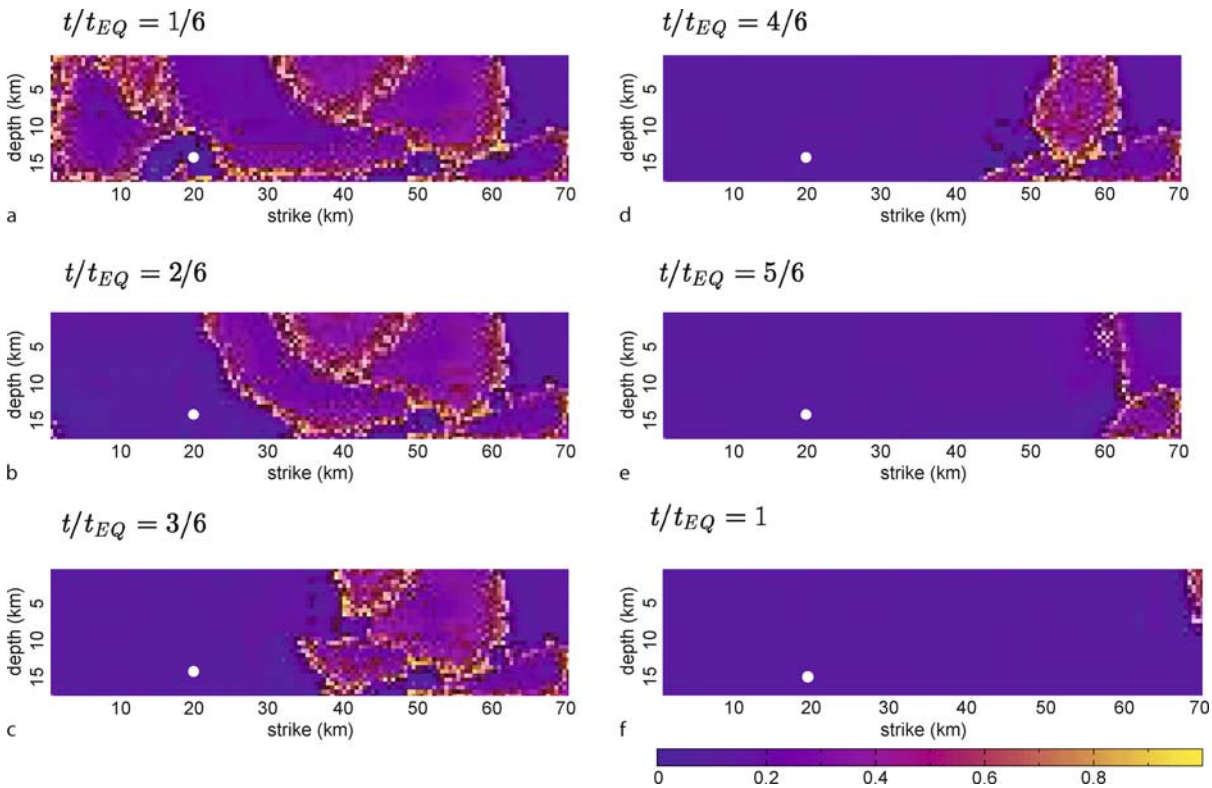

Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 4
Snapshots of rupture evolution for a system-wide event on a smooth fault without creep motion. t denotes the time after the rupture initiation (given in units of the total earthquake duration t_{EQ}). The white circle is the hypocenter of the event. The figure shows the dimensionless stress state $\hat{\tau} = \frac{\tau - \tau_a}{\tau_s - \tau_a}$ of the cells. a $t/t_{EQ} = 1/6$; b $t/t_{EQ} = 2/6$; c $t/t_{EQ} = 3/6$; d $t/t_{EQ} = 4/6$; e $t/t_{EQ} = 5/6$; f $t/t_{EQ} = 1$

simulated in order to search in a large fraction of the parameter space for relationships between input parameters and observed seismicity features. In this section, a selection of key results is presented and discussed in relation to critical states of seismicity.

Rupture Histories

We compare qualitatively rupture histories of large earthquakes for three end-member cases in parameter space:

(1) a smooth fault,
(2) a rough fault, and
(3) a fault without dynamic weakening ($\tau_d = \tau_s$ or $D \to \infty$ in Eq. (16)).

Following [5], we vary the degree of quenched spatial disorder for a particular realization by introducing barriers of high stress drop $\tau_s - \tau_a$ in an environment of low stress drop.

The observation that smooth faults show a more regular earthquake occurrence than rough faults, can be explained by the ability of the stress field to synchronize on certain fault patches. On a disordered fault, this type of synchronization is unlikely. Figure 4a shows the stress field (normalized between 0 and 1) immediately before a large earthquake on a smooth fault. The most striking feature is the emergence of clearly defined patches with highly loaded boundaries. During rupture evolution, these patches rupture almost in series until the fault is nearly unloaded (see Fig. 4b–f). A different situation is shown in Fig. 5, corresponding to a rough fault with creep coefficients $c(i, j)$ (Eq. (9)) that increase with depth leading to a brittle-ductile transition zone as in [5] and [80]. Here the stress field in the brittle regime is irregular without obvious pattern formation. Similar behavior is found in a case where dynamic weakening is switched off ($D \to \infty$); in other words, the material heals instantaneously. Figure 6 shows the stress field in this case. In analytical studies, it has been shown that this corresponds exactly to a critical point in a phase diagram spanned by the stress dissipation and dynamic weakening [22,27]. Observational results indicate [68] that irregular slip histories and power

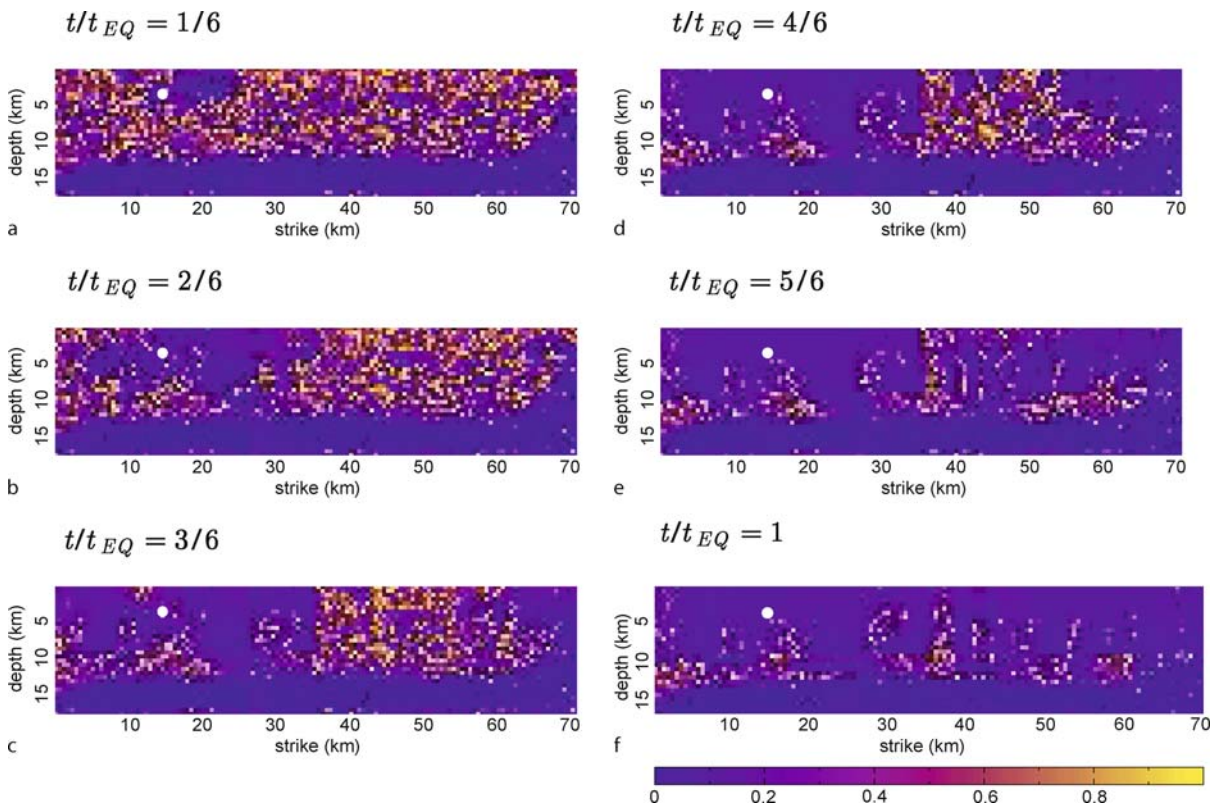

Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 5
Same as Fig. 4 for a strongly disordered fault with a brittle-ductile transition at about 15 km depth. a $t/t_{EQ} = 1/6$; b $t/t_{EQ} = 2/6$; c $t/t_{EQ} = 3/6$; d $t/t_{EQ} = 4/6$; e $t/t_{EQ} = 5/6$; f $t/t_{EQ} = 1$

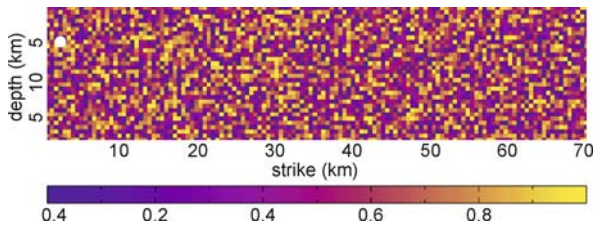

Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 6
Same as Fig. 4a for a fault without dynamic weakening ($\tau_d = \tau_s$) corresponding to a dynamic overshoot coefficient $D \to \infty$ (Eq. (16))

law frequency-size distributions are associated with geometrically disordered fault structures, while characteristic earthquake statistics and overall regular ruptures are found on mature fault with large total displacements.

Although the stress field shows a complex evolution during a simulation, the presence or absence of characteristic length scales indicating the relation to a critical point is easily detected. From an observational point of view, the stress field is not accessible. The evolution of the displacement field may be estimated, e. g. from seismic and geodetic data using slip inversion techniques. Because of the high uncertainties in the calculated slip histories, a quantitative comparison of the simulated data with "natural" data is questionable. However, general features of the quasidynamic ruptures are quite realistic, e. g. the irregular patterns in Fig. 5 resemble the rupture of the Chi–Chi (Taiwan) earthquake on September 21, 1999 ($M_w = 7.6$) [58].

Later we will show that the frequency-size distribution of earthquakes can serve to some extent as a proxy for the degree of disorder of the stress field. Ben-Zion et al. [12] discuss additional seismicity functions that may be used as surrogate variables for the stress.

Frequency–Size Distributions

The frequency-size (FS) distribution is one of the most important characteristics of observed seismicity. For world-

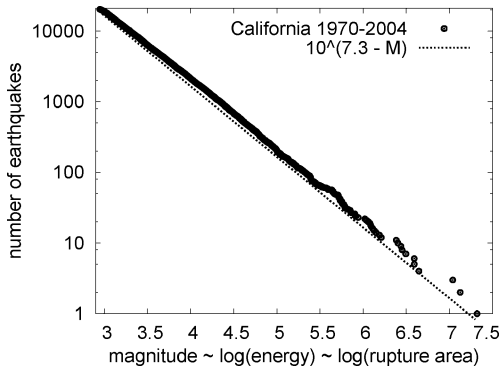

Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 7
Frequency-size distribution for California from 1970 to 2004; the dashed line denotes a power law fit to the data

wide seismicity, this distribution is given by the Gutenberg–Richter law (Eq. (1)). Figure 7 shows the FS distribution of California seismicity from 1970 to 2004. Here we use the non-cumulative version of Eq. (1), where N is the number of earthquakes with magnitude between M and $M + \mathrm{d}M$ with a magnitude bin $\mathrm{d}M = 0.1$.

For individual fault zones the FS distribution can deviate from Eq. (1), especially for high magnitudes. Examples are given in Fig. 8, which shows the FS distribution of the Parkfield segment (Fig. 8a) and for the San Jacinto fault (Fig. 8b) in California for a time span of 45 years. The distribution of the Parkfield segment consists of two parts: A scaling regime for $2.2 \leq M \leq 4.5$ and a "bump" for $4.5 < M \leq 6.0$. For the San Jacinto fault, the scaling range is observed for almost all events ($2.2 \leq M \leq 5.0$). The decrease for $M \approx 2$ in both plots is probably due to catalog incompleteness.

The FS distribution as shown in Fig. 8a is called the characteristic earthquake distribution, because of the increased probability for the occurrence of large ("characteristic") events compared to the prediction of the Gutenberg–Richter relation. The latter is an exponential distribution for the earthquake frequency as a function of magnitude, or a power law distribution for the earthquake frequency as a function of potency (Eqs. (19), (20)), moment, energy, or rupture area, over a broad range of magnitudes [66]. The Gutenberg–Richter relation is "scale-free" because a power law distribution indicates the absence of a characteristic scale of the earthquake size [64]. In terms of critical point processes, the absence of a characteristic length scale indicates that the system is close to the critical point. In this state, earthquakes of all magnitudes can occur, or each small rupture can grow into a large one. Therefore, the frequency-size distribution can serve as a proxy for the current state of a system in relation to a critical point.

The FS distribution in a model can be tuned by varying the mean stress $\langle\tau\rangle$ on the fault, where $\langle\rangle$ denotes the spatial average. This can be achieved, for instance, by varying brittle properties, e. g. in terms of the dynamic overshoot coefficient D (Eq. (16)), or by introducing dissipation [31,79]. Figure 9 shows FS distributions for two different values of D : $D = 5/4$ (Fig. 9a) and a higher value $D = 5/3$ (Fig. 9b). While Fig. 9a follows a characteristic earthquake behavior similar to the Parkfield case (Fig. 8a), Fig. 9b resembles the shape of the FS distribution of the San Jacinto fault (Fig. 8b).

As an outcome, three cases can be distinguished by means of a critical mean stress τ_{crit}:

1. subcritical fault ($\langle\tau\rangle < \tau_{\mathrm{crit}}$): the mean stress on the fault is too small to produce large events. The system is always far from the critical point. The FS distribution is a truncated Gutenberg–Richter law.
2. supercritical fault ($\langle\tau\rangle > \tau_{\mathrm{crit}}$): the mean stress is high and produces frequently large events. After a large earthquake (critical point), the stress level is low (system is far from the critical point) and recovers slowly (approaches the critical point). The FS distribution is a characteristic earthquake distribution.
3. critical fault ($\langle\tau\rangle \approx \tau_{\mathrm{crit}}$): the system is always close to the critical point with scale-free characteristics. The FS distribution is a Gutenberg–Richter law with a scaling range over all magnitudes.

If the FS distribution is plotted as a function of the model parameters, the result can be visualized by a phase diagram [22,31,79,80]. An example is provided in Fig. 10, which shows schematically the phase diagram spanned by the degree of quenched spatial disorder and $1 - \varepsilon$ with the dynamic weakening coefficient ε (Eq. (17)). The phase diagram summarizes results from various studies, which demonstrate that the degree of spatial disorder of the stress drop acts as a tuning parameter for the FS distribution [5,36,80].

If the model is in the transition regime between Gutenberg–Richter statistics and characteristic earthquake behavior, the ability of the stress field to synchronize on parts of the fault can have additional impact on the dynamics of seismicity: for a model with small cells and high stress fluctuations along the cell boundaries arising from a high degree of spatial disorder, the system can undergo a spontaneous transition from an ordered state and characteristic behavior to a disordered state following Gutenberg–Richter statistics. Due to the high fluctuations in the

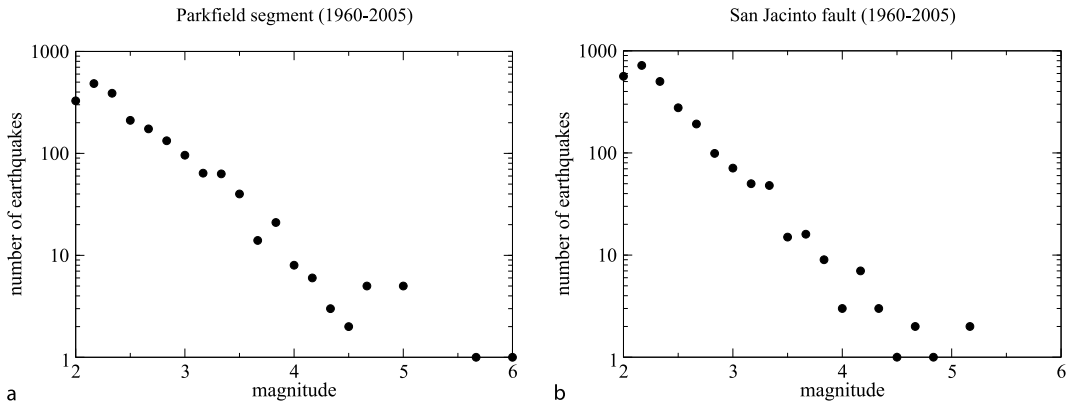

Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 8
Frequency-size distribution for two faults in California: a the Parkfield segment, and b the San Jacinto fault

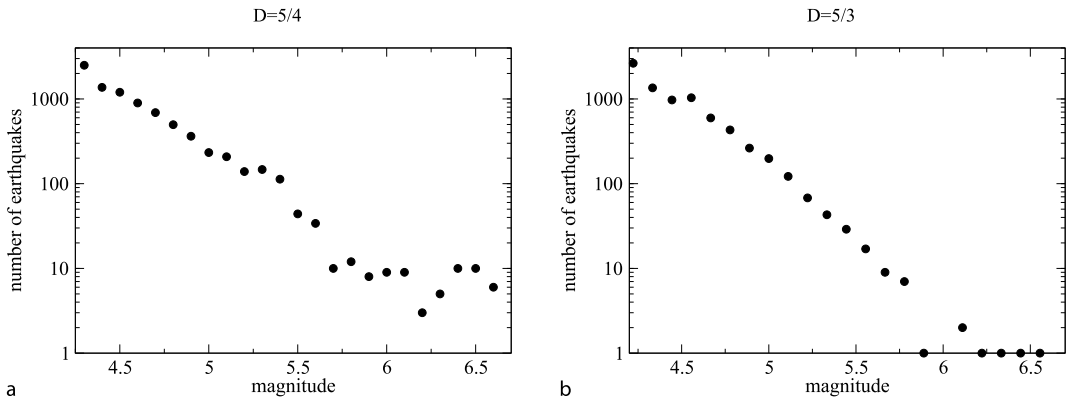

Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 9
Frequency-size distribution for model realizations with different dynamic overshoot coefficients (Eq. (16)): a $D = 5/4$, b $D = 5/3$

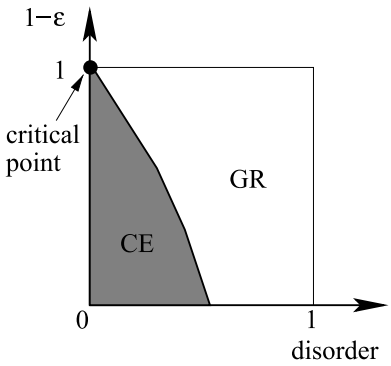

Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 10
Phase diagram for the frequency-size distribution (GR = Gutenberg–Richter distribution, CE=characteristic earthquake distribution) spanned by the degree of quenched spatial disorder and the dynamic weakening represented by ε. The upper left corner corresponds exactly to a critical point [22,27] and results in scale-free characteristics as shown in Fig. 6

stress field, there is some probability that a certain number of cells synchronize by chance, leading to an ordered behavior for some seismic cycles, until the order is destroyed, again resulting from stress fluctuations. This type of mode-switching behavior has been observed earlier in a mean-field model and a damage rheology model [11,22]. Figure 11a gives a corresponding earthquake sequence with spontaneous mode-switching behavior. Figure 11b shows a sequence calculated with a higher grid resolution (128 × 50 cells). The tendency to mode-switching is less pronounced, but still visible. In [79] it is shown that the emergence of such mode-switching depends both on the spatial range of interaction (given as the decay of the Green's function) and the discretization of the computational grid. In the less realistic model of [22], where the stress redistribution is governed by a constant (space-independent) Green's function, analytical expressions for persistence times have been calculated [27]. In [11] some evidence for mode-switching behavior in long seismic

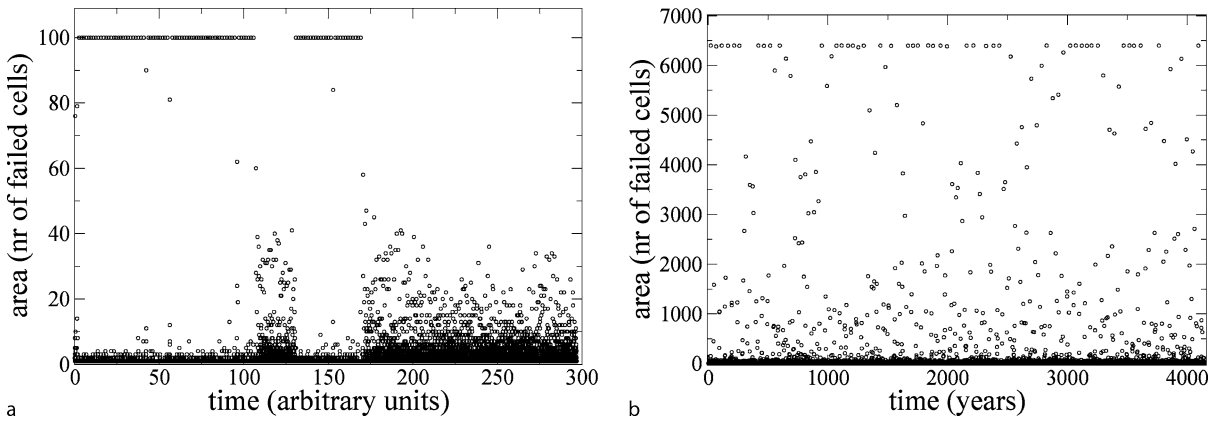

Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 11
Earthquake area (measured as the number of failed cells) as a function of time a in the mean-field model of Dahmen et al. [22] for a fault with 100 cells and b in the elastic model with 128 × 50 cells

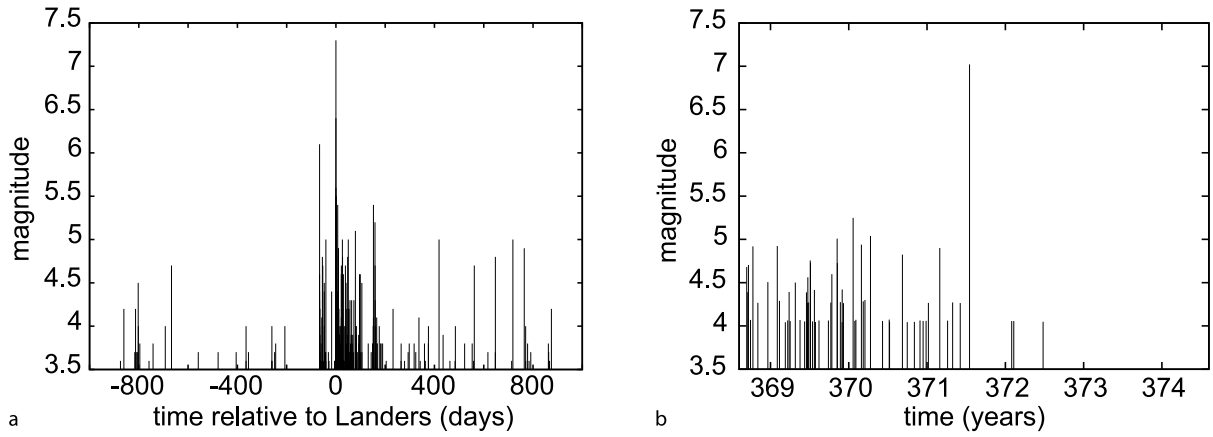

Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 12
Earthquakes before and after a mainshock: a the *M*7.3 Landers (California) earthquake; b *M*7.3 earthquake in the basic version of the model

records based on paleoseismic and geologic data from the Dead Sea fault and other regions are discussed. However, the relevance of mode-switching to natural seismicity remains unclear due to the general lack of very long data records.

Aftershocks and Foreshocks

The most pronounced temporal pattern in observed seismicity is the emergence of strongly clustered aftershock activity following a large earthquake. Apart from the Omori–Utsu law (Eq. (2)), it is widely accepted that aftershocks are characterized by the following properties:

1. The aftershock rate scales with the mainshock size [51].
2. Aftershocks occur predominantly around the edges of the ruptured fault segments [66].
3. Båth's law [4]: The magnitude of the largest aftershock is usually about one unit smaller than the mainshock magnitude.

Deviations from the Omori–Utsu law, especially for rough faults, are discussed in [45]. While aftershocks are observed after almost all large earthquakes, foreshocks occur less frequent [71]. As a consequence, much less is known about the properties of these events. Kagan and Knopoff [40] and Jones and Molnar [39] propose a power law increase of activity according to an "inverse" Omori–Utsu law.

Figure 12a shows an example for the aftershock sequence following the *M*7.3 Landers earthquake in California on June 28, 1992. An earthquake of similar size generated by the model is given in Fig. 12b. The absence of af-

tershocks in the simulation is clearly visible. The reason for the lack of aftershocks is the unloading of the fault resulting from the mainshock: When a large fraction of the fault has ruptured, the stress in this region will be close to the arrest stress after the event. Consequently, the seismicity rate will be almost zero until the stress field has recovered to a moderate level.

Discussions for likely mechanisms of aftershocks are given in [9] and [81]. A common feature is the presence of rapid postseismic stress which generates aftershock activity. In [32], for instance, postseismic stress has been attributed to a viscoelastic relaxation process following the mainshock. In the work discussed here, continuous creep displacement following Eq. (9) is assumed. Additionally, the computational grid is divided by aseismic barriers from the free surface to depth into several seismically active fault segments. As shown in [81], this modification results in a concentration of stress in the aseismic regions during rupture and, subsequently in a release of stress after the event according to the coupled creep process. This stress release triggers aftershock sequences obeying the Omori–Utsu law (Eq. (2)). A typical aftershock sequence after a $M6.8$ event is shown in Fig. 13. In agreement with Båth's law, the strongest aftershock has the magnitude $M = 5.5$ in this sequence. The sequence shows also the effect of secondary aftershocks, namely aftershocks of aftershocks [61]. The stacked earthquake rate as a function of the time after the mainshock is given in Fig. 14. In this case, where the barriers are characterized by creep coefficients higher by a factor of 10^5 than in the other patches, a realistic Omori exponent of $p = 1$ is found.

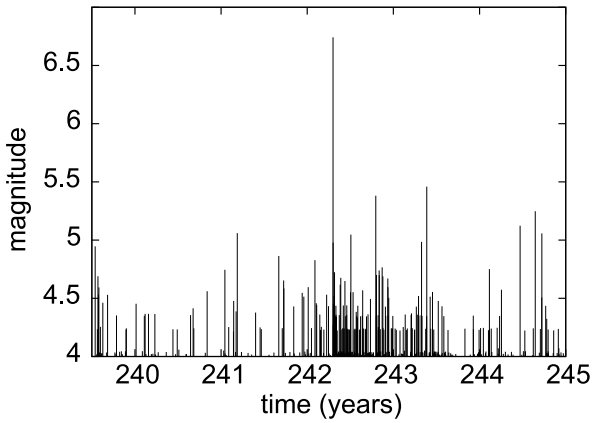

Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 13
Earthquakes before and after a mainshock with $M = 6.8$ in the modified model

Aftershock sequences like in Fig. 13 emerge after all large events in the extended model. In contrast, there is generally no clear foreshock signal visible in single sequences. However, stacking many sequences together unveils a slight increase of the earthquakes rate prior to a mainshock supporting the observation of accelerating foreshock activity. An explanation of these events can be given in the following way: Between two mainshocks the stress field organizes itself towards a critical state, where the next large earthquake can occur. This critical state is characterized by a disordered stress field and the absence of a typical length scale, where earthquakes of all sizes can occur [12]. The mainshock may occur immediately or after some small to moderate events. The latter case can be considered as a single earthquake, which is interrupted in the beginning. This phenomenon of delayed rupture propagation has also provided a successful explanation of foreshocks and aftershocks in a cellular automaton model [33,35].

The hypothesis that foreshocks occur in the critical point and belong, in principle, to the mainshock, can be verified by means of the findings from Subsect. "Frequency–Size Distributions". In particular, the frequency-size distribution in the critical point (or close to the critical point) is expected to show scale-free statistics. If an overall smooth fault model following characteristic earthquake statistics is studied over a long time period, the approach of the critical point should be seen in terms of a change of the frequency-size distribution towards Gutenberg–Richter behavior [12]. This change of frequency-size statistics is indeed observed in the model (Fig. 15) and supports the validity of the critical point concept [82].

Accelerating Moment Release

In the previous section, it has been argued that large earthquakes are associated with a critical point and the preparation process is characterized by increasing disorder of the stress field and increasing tendency for scale-free characteristics in the frequency-size distribution. Further support for critical point dynamics has been provided by the observational finding of [17] that the cumulative Benioff strain $\Sigma\Omega(t)$ follows a power law time-to-failure relation prior to the $M7.1$ Loma Prieta earthquake on October 17, 1989:

$$\Sigma\Omega(t) = \sum_{i=1}^{N(t)} \sqrt{E_i} = A - B(t_f - t)^m \tag{21}$$

Here, E_i is the energy release of earthquake i, $N(t)$ is the number of earthquakes before time t; t_f is the failure time

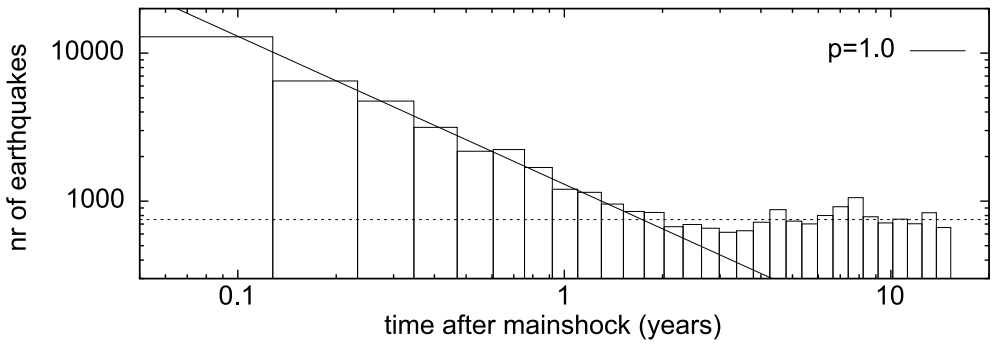

Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 14
Earthquake rate as a function of time for the model with seismic and aseismic regions. The calculation is based on a simulation with 200,000 earthquakes covering about 5000 years; the earthquake rates are averaged over about 300 mainshocks. A fit of the Omori–Utsu law (Eq. (2)) with $p = 1$ is denoted as a solid line. The dashed line gives the estimated background level of seismicity

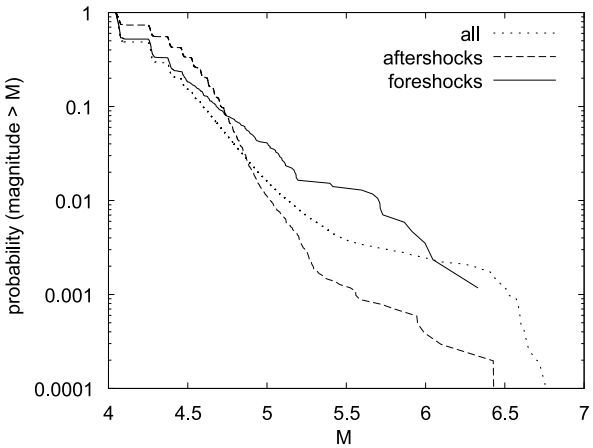

Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 15
Frequency-magnitude distributions of all earthquakes, foreshocks and aftershocks. Foreshocks and aftershocks are defined as earthquakes occurring within one month before and after an earthquake with $M \geq 6$

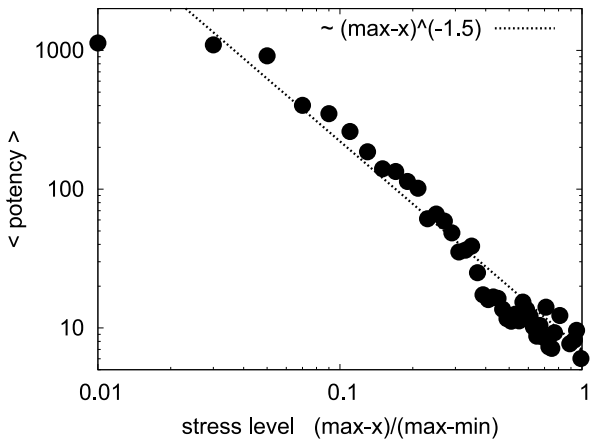

Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 16
Mean potency release (Eq. (19)) as a function of the stress level. The stress level is normalized to the maximum (max) and minimum (min) observed stress

and A, B, and $m > 0$ are constants. Similar studies for numerous seismically active regions followed (see [8,77] and references therein).

The time-to-failure relation Eq. (21) has been explained by [54] and [60] from the viewpoint of renormalization theory and by [8] and [65] in terms of damage rheology. Similar to the findings about foreshocks, the time-to-failure pattern is not universal. Therefore, a stacking procedure is adopted in order to obtain robust results on the validity of Eq. (21) in the model. This is not straightforward, since the interval of accelerating moment release is not known a priori and the duration of a whole seismic cycle, as an upper limit, is not constant. To normalize the time interval for the stacking, the potency release (Eq. (19)) is computed as a function of the (normalized) stress level (Fig. 16). Taking into account that the stress level increases almost linearly during a large fraction of the seismic cycle, the stress level axis in Fig. 16 can effectively be replaced by the time axis leading to a power law dependence of the potency release on time. The best fit is provided with an exponent $s = -1.5$. Transforming the potency release to the cumulative Benioff strain (Eq. (21)), results in an exponent $m = 0.25$ in Eq. (21). This finding is based on a simulation over 5000 years; the exponent is in good agreement with the theoretical work [53], that derives $m = 0.25$ for a spinodal model, and the analytical result of $m = 0.3$ in the damage mechanics model [8]. An observational study of California seismicity finds m between 0.1 and 0.55 [15].

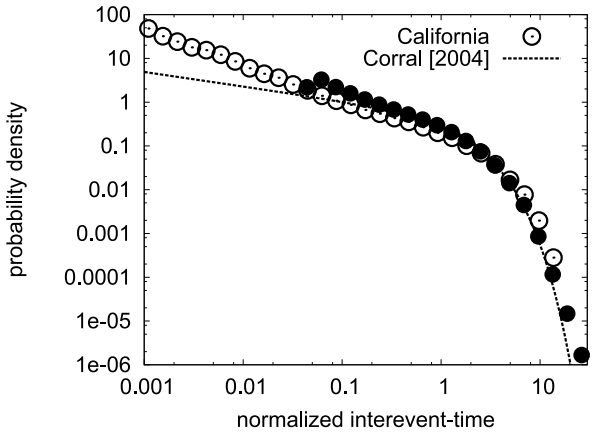

Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 17
The normalized interevent time distribution of the model simulations (black dots) compared with the result of [21] and the distribution of earthquakes in California (ANSS catalog of $M \geq 3$ earthquakes occurred between 1970 and 2004 within 29° and 43° latitude and −113° and −123° longitude)

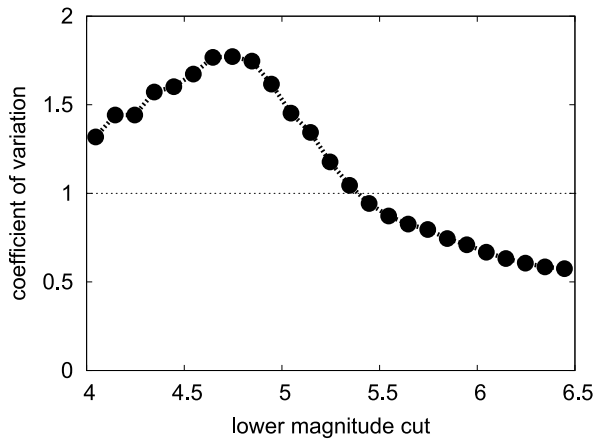

Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 18
The temporal earthquake occurrence quantified by the coefficient of variation as a function of the lower magnitude cutoff. Values larger than 1 indicate clustering, whereas lower values point to quasiperiodic behavior

Interevent Times

In recent studies it has been shown that the distribution of interevent times can be described by a universal law. In particular, the distributions from different tectonic environments, different spatial scales (from worldwide to local seismicity) and different magnitude ranges collapse if the time Δt is rescaled with the rate R_{xy} of seismic occurrence in a region denoted by (x, y) [21]. Such rescaling leads to

$$D_{xy}(\Delta t) = R_{xy} \cdot f(R_{xy}\Delta t)\,, \tag{22}$$

where D_{xy} is the probability density for the interevent time Δt, and f can be expressed by a generalized gamma distribution

$$f(\theta) = C\frac{1}{\theta^{\gamma-1}} \exp\left(-\frac{\theta^{\delta}}{B}\right). \tag{23}$$

The parameters C, γ, δ, and B have been determined by a fit to several observational catalogs [21].

In Fig. 17 we compare $D_{xy}(\Delta t)$ from Eq. (22) with two earthquake catalogs: (1) The ANSS catalog of California (catalog ranges are given in the caption), and (2) the model catalog. Due to the universality of Eq. (22) with respect to different spatial scales, the comparison of the model simulating a single fault of 70 km length with a region of hundreds of kilometers including several faults in California does not require coarse graining the ANSS catalog. In the region where the interevent times are calculated, we find a remarkable agreement of the three curves. For small values of Δt, Eq. (22) deviates from the California data; for high values the model has a slightly better correspondence with the observational data than Eq. (22). Thus the results generally support the recent findings of [21].

The degree of temporal clustering of earthquakes can be estimated by the coefficient of variation CV of the interevent time distribution.

$$CV = \sigma/\mu\,, \tag{24}$$

where σ is the standard deviation and μ the mean value of the interevent time distribution. Values of $CV > 1$ denote clustered activity, while $CV < 1$ represents quasiperiodic occurrence of events. The case $CV = 1$ corresponds to a random Poisson process. The studies of [5] and [80] have found that the clustering properties of the large events depend on the degree of quenched spatial disorder of the fault. Figure 18 shows that CV as a function of the lower magnitude cutoff has a characteristic shape. The values of CV are higher than 1 (clustered) for small and intermediate earthquakes ($M \leq 5.4$) and smaller than 1 (quasiperiodic) for larger earthquakes. This corresponds to the case of a low degree of disorder in [80], because the brittle cells which participate in an earthquake have no significant spatial disorder. We note that this behavior resembles the seismicity on the Parkfield segment of the San Andreas fault, which is characterized by a quasiperiodic occurrence of mainshocks. Based on the analysis of 37 earthquake sequences, an estimation of $CV \approx 0.5$ has been found for multiple tectonic environments [26].

A different behavior is observed on the San Jacinto fault in California, where the largest events occur less reg-

ularly and have overall smaller magnitudes. As discussed in [5] and [80], this can be modeled by imposing higher degrees of disorder leading to a broader range of spatial size scales, e. g. by using a higher number of near-vertical barriers. While barriers provide a simple and physically motivated way to tune the degree of disorder, other types of heterogeneities may work as well, as long as they are able to produce strong enough fluctuations of the stress field. As an example, we mention fractal distributions of the stress drop, which can be tuned easily by changing the fractal dimension [63,80].

Recurrence Times of Large Earthquakes

While interevent times are waiting times between successive earthquakes in a given catalog, recurrence times are defined as waiting times between two successive *large* events, typically in the magnitude range $6 \leq M \leq 9$, depending on the region. For example, on the Parkfield segment of the San Andreas fault seven $\sim M6$ earthquakes occurred between 1857 and 2004 with recurrence times $T_1 = 24$ years, $T_2 = 20$ years, $T_3 = 21$ years, $T_4 = 12$ years, $T_5 = 32$ years, and $T_6 = 38$ years.

The distribution of recurrence times of large earthquakes is crucial for the calculation of seismic hazard. Due to a lack of observational data, this distribution is unknown for real fault systems. Commonly used distributions are based on extreme value statistics and on models for catastrophic failure. These include the lognormal distribution [50], the Brownian passage time distribution [44], and the Gumbel distribution [29]. All distributions are characterized by a maximum for a certain recurrence time followed by an asymptotic decay. The Brownian passage time distribution and the lognormal distribution have been used by the Working Group on California Earthquake Probabilities [70], e. g. for calculating earthquake probabilities in the San Francisco Bay area.

Figure 19a shows the probability density function (pdf) of the recurrence times of earthquakes with magnitude $M > 6.2$ in a realization of the numerical model for the Parkfield region [74]. Since we focus on long time-scales, we use here a minimal model without aseismic creep and strong spatial heterogeneities. This model leads to characteristic earthquake statistics and quasiperiodic occurrence of large events, and can therefore serve as a model framework for large earthquakes on the Parkfield segment. However, quantities which are only poorly known from empirical data, e. g. the stress drop, have to be chosen in order to perform a numerical simulation. Starting with an imposed uniform a priori distribution $P(\Delta\tau)$ of stress drops between a lower bound $\Delta\tau_{\min}$ and an upper bound $\Delta\tau_{\max}$, an a posteriori distribution $P(\Delta\tau|T_1,\ldots,T_N)$ can be estimated using observational recurrence times $T_1,\ldots,T_N$ from Parkfield and Bayes' theorem [13],

$$P(\Delta\tau|T_1,\ldots,T_N) = \frac{P(T_1,\ldots,T_N|\Delta\tau)P(\Delta\tau)}{\sum_{s=\Delta\tau_{\min}}^{\Delta\tau_{\max}} P(T_1,\ldots,T_N|s)P(s)}\,, \tag{25}$$

with the likelihood function

$$P(T_1,\ldots,T_N|\Delta\tau) = \prod_{i=1}^{N} f(T_i|\Delta\tau)\,. \tag{26}$$

The function $f(T_i|\Delta\tau)$ denotes the pdf of recurrence times simulated with a model stress drop $\Delta\tau$. To get an analytic expression of this function, it is fitted by a Gamma distribution $f(t) = \beta^{-1}(\Gamma(\gamma))^{-1}(\frac{t-\mu}{\beta})^{\gamma-1}\exp(-\frac{t-\mu}{\beta})$ with the location parameter μ, the shape parameter $\gamma \equiv 2.0$ and the scale parameter β (with $x \geq \mu\,;\ \gamma,\beta > 0$). For an example see Fig. 19a. In [74] it is shown that the mean value μ_t and the standard deviation σ_t of the fits in this model are related to the average stress drop of a large earthquake $\Delta\tau$ by the simple empirical relations

$$\begin{aligned}\mu_t(\Delta\tau) &= 9.7\cdot\Delta\tau\\ \sigma_t(\Delta\tau) &= 1.8\cdot\Delta\tau^2 - 6.8\cdot\Delta\tau + 11.7\end{aligned} \tag{27}$$

with μ_t, σ_t in years and $\Delta\tau$ in MPa. Using this approximation in combination with six observational recurrence times from $\sim M6$ earthquakes on the Parkfield segment, we find the a posteriori distribution of stress drops shown in Fig. 19b. The position where this distribution reaches the maximum, $\Delta\tau = (3.04 \pm 0.27)$ MPa, is the most representative value of the stress drop of $\sim M6$ Parkfield events.

The cumulative probability density function (cdf) of recurrence times based on Eq. (26) and the observational data can now be calculated by

$$C(t) = \int_0^t \int_{\Delta\tau_{\min}}^{\Delta\tau_{\max}} f(t'|\Delta\tau)P(\Delta\tau|T_1,\ldots,T_N)\mathrm{d}\Delta\tau\mathrm{d}t'\,. \tag{28}$$

The hazard function

$$H(\Delta t|t_0) = \frac{C(t_0+\Delta t) - C(t_0)}{1 - C(t_0)} \tag{29}$$

is the conditional probability that the next large earthquake occurs in the interval $[t_0\,;\ t_0+\Delta t]$ given the time t_0 since the last large event. Results for two choices of observational data (corresponding to two different observa-

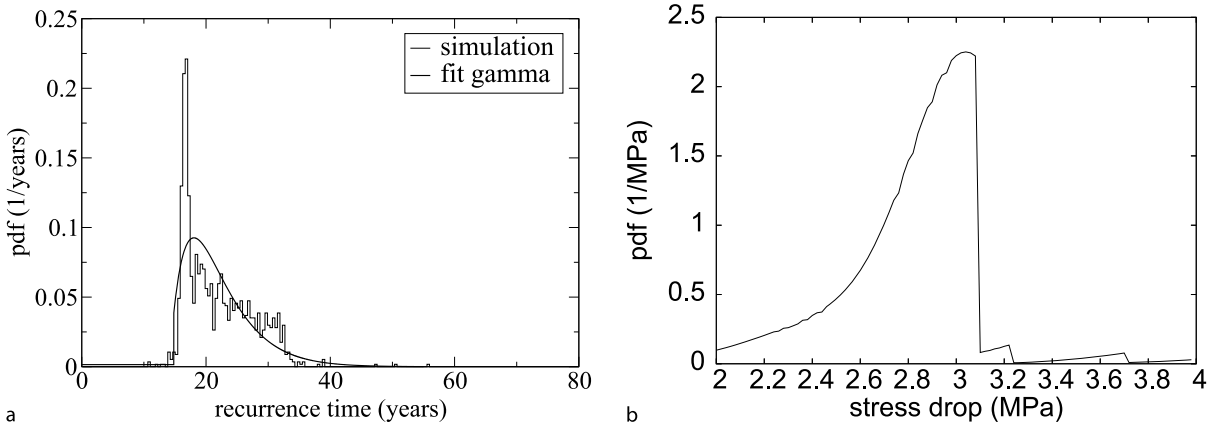

Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 19
a Approximated probability density function of the recurrence time distribution of large earthquakes ($M > 6.2$) for a simulated earthquake catalog and fit with a truncated Gamma distribution; b A posteriori distribution $P(\Delta\tau | T_1, \ldots, T_N)$ of stress drop $\Delta\tau$ calculated with Bayes' theorem (Eq. (25))

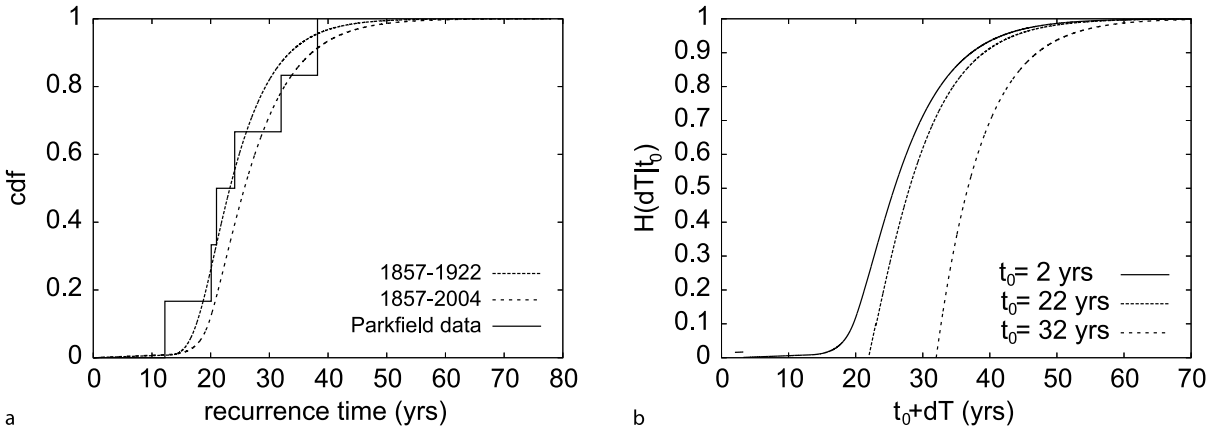

Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space, Figure 20
a Cumulative recurrence time distribution $C(t)$ (Eq. (28)) for (1) the Bayesian approach with three data points: long-dashed line; (2) the Bayesian approach with six data points: short-dashed line. The solid line denotes the cdf of the six Parkfield recurrence times; b Hazard function $H(t_0 | \Delta T)$ (Eq. (29)) based on the six observational recurrence times between 1857 and 2004 a as a function of dT for different values of t_0

tional periods) in comparison to the Parkfield cdf are given in Fig. 20a. The hazard function for three fixed values of t_0 and varying Δt is given in Fig 20b.

This approach enables us to calculate the most likely occurrence time of the next (post 2004) Parkfield earthquake by picking the maximum of the (non-cumulative) recurrence time distribution (inner integral in Eq. (28)) after taking all Parkfield earthquakes (1857–2004) into account. Based on the analysis done so far, we may forecast the next $\sim M6$ Parkfield earthquake to occur in May 2027. The error associated with one standard deviation of the pdf is 7.7 years. We note, however, that the probability for the occurrence of a $\sim M6$ earthquake between May 2026 and May 2028 is only about 14%.

Summary and Conclusions

The present review deals with the analysis, the understanding and the interpretation of seismicity patterns with a special focus on the critical point concept for large earthquakes. Both physical modeling and data analysis are discussed. This study aims at practical applications to model data from real fault zones. A point of particular interest is the detection of phenomena prior to large earthquakes and their relevance for a possible prediction of these events. Despite numerous reports on anomalous precursory seismicity changes [62], there is no precursor in sight which obeys a degree of universality that would make it practically useful. It is, therefore, important to study less fre-

quent precursory phenomena by means of long model simulations.

Toward this goal, we discuss a numerical model which is on one hand reasonably physical, and on the other hand simple enough that it allows to obtain some analytical results and perform long simulations. The basic version of the model consists of a segmented two-dimensional strike-slip fault in a three-dimensional elastic half space and is inherently discrete because of the abrupt transition from static to kinetic friction [10]. This paper and ► Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of summarize a large body of analytical and numerical results associated with the model.

The results of the simulations indicate an overall good agreement of the synthetic seismicity with natural earthquake activity, with respect to frequency-size distributions and various features of earthquake sequences. The degree of spatial heterogeneity on the fault, which is implemented by means of space-dependent rheological properties, has important effects on the resulting catalogs. Smooth faults are associated with the characteristic earthquake statistics, regular occurrence of mainshocks and overall smooth stress fields. On the other hand, rough faults generate scale-free Gutenberg–Richter statistics, irregular mainshock occurrence, and overall rough stress fields. A closer look at the disorder of the stress field shows, however, that even on a smooth fault a gradual roughening takes place when the next large earthquake is approached [12,82]. This is reflected in the frequency-size distribution which evolves towards the Gutenberg–Richter law and other changes of seismicity. The results can be used to establish relations between the proximity of a state on a fault to a critical point, the (unobservable) stress field, and the (observable) seismicity functions. Furthermore, it is demonstrated that the concept of "self-organized criticality" can be folded back to criticality associated with tuning parameters [12,31]. We note that phase diagrams with different dynamic regimes as functions of tuning parameters, in addition to criticality, provide a general and rich description of seismicity. Accelerating seismic release, growing spatial correlation length, changes of frequency-size statistics and evolution of other seismicity parameters may be used to track the approach to criticality [73,75,76,77].

Future Directions

We have demonstrated that numerical fault models are valuable for understanding the underlying mechanisms of observed seismicity patterns, as well as for practical estimates of future seismic hazard. The latter requires model realizations that are tuned to a specific fault zone by assimilating available observational results and their uncertainties. In a case study, the seismic hazard in the Parkfield region has been estimated by combining such a tuned model with few observational data. The use of Bayesian analysis allows us to construct a flexible hazard model for this region which can, in general, incorporate statistical and non-statistical data (e. g. from paleoseismology and geodesy) to improve and update the estimations of the seismic hazard. This approach is particularly promising for less-well monitored regions, and especially for low-seismicity regions like those in central Europe.

Modification of the stress transfer calculations to account for a statistical preference of earthquake propagation direction on a given fault section, e. g. [6,25], can improve the estimates of seismic hazard associated with large faults. It is also possible to extend the discussed framework to other geohazards with even smaller amount of observational data, e. g. the occurrence of landslides. Since the fault model deals with coupled physical processes leading to interacting earthquakes, a challenging future direction will be the design of a more general model for interacting geohazards including earthquakes on different faults as well as landslides triggered by earthquakes, and perhaps tsunamis initiated by (submarine) earthquakes or landslides.

Acknowledgments

We thank Jürgen Kurths, James R. Rice, Frank Scherbaum, Karin Dahmen, Donald L. Turcotte, and many others for useful discussions. GZ and MH acknowledge support from the collaborative research center "Complex Nonlinear Processes" (SFB 555) of the German Research Society (DFG). SH acknowledges support from the DFG-project SCHE280/14 and the EU-project SAFER. YBZ acknowledges support from the National Science Foundation, the United States Geological Survey, and the Southern California Earthquake Center. GZ and YBZ thank the Kavli Institute for Theoretical Physics, UC Santa Barbara, for hospitality during a 2005 program on Friction, Fracture and Earthquake Physics, and partial support based on NSF grant PHY99–0794. We thank James R. Holliday, Vladimir Keilis-Borok and Willie Lee for providing comments on the paper.

Bibliography

1. Aki K, Richards PG (2002) Quantitative seismology. University Science Books, Sansalito
2. Bak P (1996) How Nature Works. The science of self-organised criticality. Springer, New York

3. Bak P, Tang C (1989) Earthquakes as a phenomenon of self-organised criticality. J Geophys Res 94:15635–156637
4. Båth M (1965) Lateral inhomogeneities in the upper mantle. Tectonophysics 2:483–514
5. Ben-Zion Y (1996) Stress, slip, and earthquakes in models of complex single-fault systems incorporating brittle and creep deformations. J Geophys Res 101:5677–5706
6. Ben-Zion Y (2001) Dynamic rupture in recent models of earthquake faults. J Mech Phys Solids 49:2209–2244
7. Ben-Zion Y (2003) Appendix 2, Key Formulas in Earthquake Seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part B. Academic Press, San Diego, pp 1857–1875
8. Ben-Zion Y, Lyakhovsky V (2002) Accelerated seismic release and related aspects of seismicity patterns on earthquake faults. Pure Appl Geophys 159:2385–2412
9. Ben-Zion Y, Lyakhovsky V (2006) Analysis of aftershocks in a lithospheric model with seismogenic zone governed by damage rheology. J Geophys Int 165:197–210; doi:10.1111/j.1365-246X2006.02878.x
10. Ben-Zion Y, Rice JR (1993) Earthquake failure sequences along a cellular fault zone in a three-dimensional elastic solid containing asperity and nonasperity regions. J Geophys Res 98:14109–14131
11. Ben-Zion Y, Dahmen K, Lyakhovsky V, Ertas D, Agnon A (1999) Self-driven mode switching of earthquake activity on a fault system. Earth Plan Sci Lett 172:11–21
12. Ben-Zion Y, Eneva M, Liu Y (2003) Large Earthquake Cycles and Intermittent Criticality On Heterogeneous Faults Due To Evolving Stress and Seismicity. J Geophys Res 108:2307; doi:10.1029/2002JB002121
13. Bernardo JM, Smith AFM (1994) Bayesian Theory. Wiley, Chichester
14. Binney JJ, Dowrick NJ, Fisher AJ, Newman MEJ (1993) The theory of critical phenomena. Oxford University Press, Oxford
15. Bowman DD, Oullion G, Sammis CG, Sornette A, Sornette D (1998) An observational test of the critical earthquake concept. J Geophys Res 103:24359–24372
16. Brace WF (1960) An extension of the Griffith theory of fracture to rocks. J Geophys Res 65:3477–3480
17. Bufe CG, Varnes DJ (1993) Predicitive modeling of the seismic cycle of the greater San Francisco Bay region. J Geophys Res 98:9871–9883
18. Burridge R, Knopoff L (1967) Model and theoretical seismicity. Bull Seim Soc Am 57:341–371
19. Byerlee JD (1978) Friction of rocks. Pure Appl Geophys 116:615–616
20. Chinnery M (1963) The stress changes that accompany strike-slip faulting. Bull Seim Soc Am 53:921–932
21. Corral Á (2004) Long-term clustering, scaling, and universality in the temporal occurrence of earthquakes. Phys Rev Lett 92:108501; doi:10.1103/PhysRevLett.92.108501
22. Dahmen K, Ertas D, Ben-Zion Y (1998) Gutenberg–Richter and characteristic earthquake behavior in simple mean-field models of heterogeneous faults. Phys Rev E 58:1494–1501
23. Daley DJ, Vere-Jones D (1988) An Introduction to the Theory of Point Processes, Springer Series: Probability and its Applications. Springer, Heidelberg
24. Dieterich JH (1994) A constitutive law for earthquake production and its application to earthquake clustering. J Geophys Res 99:2601–2618
25. Dor O, Rockwell TK, Ben-Zion Y (2006) Geologic observations of damage asymmetry in the structure of the San Jacinto, San Andreas and Punchbowl faults in southern California: A possible indicator for preferred rupture propagation direction. Pure Appl Geophys 163:301–349; doi:10.1007/s00024-005-0023-9
26. Ellsworth WL, Matthews MV, Nadeau RM, Nishenko SP, Reasenberg PA, Simpson RW (1999) A physically based earthquake recurrence model for estimation of long-term earthquake probabilities. US Geol Surv Open-File Rept, pp 99–522
27. Fisher DS, Dahmen K, Ramanathan S, Ben-Zion Y (1997) Statistics of earthquakes in simple models of heterogeneous faults. Phys Rev Lett 78:4885–4888
28. Geller RJ, Jackson DD, Kagan YY, Mulargia F (1997) Earthquakes cannot be predicted. Science 275:1616–1617
29. Gumbel EJ (1960) Multivariate Extremal Distributions. Bull Inst Int Stat 37:471–475
30. Gutenberg B, Richter CF (1956) Earthquake magnitude, intensity, energy and acceleration. Bull Seismol Soc Am 46:105–145
31. Hainzl S, Zöller G (2001) The role of disorder and stress concentration in nonconservative fault systems. Phys A 294:67–84
32. Hainzl S, Zöller G, Kurths J (1999) Similar power laws for fore- and aftershock sequences in a spring-block model for earthquakes. J Geophys Res 104:7243–7253
33. Hainzl S, Zöller G, Kurths J (2000) Self-organization of spatio-temporal earthquake clusters. Nonlin Proc Geophys 7:21–29
34. Hainzl S, Zöller G, Kurths J, Zschau J (2000) Seismic quiescence as an indicator for large earthquakes in a system of self-organized criticality. Geophys Res Lett 27:597–600
35. Hainzl S, Zöller G, Scherbaum F (2003) Earthquake clusters resulting from delayed rupture propagation in finit fault segments. J Geophys Res 108:2013; doi:10.1029/2001JB000610
36. Hillers G, Mai PM, Ben-Zion Y, Ampuero JP (2007) Statistical Properties of Seismicity Along Fault Zones at Different Evolutionary Stages. J Geophys Int 169:515–533; doi:10.1111/j.1365-246X2006.03275.x
37. Huang J, Turcotte DL (1990) Are earthquakes an example of deterministic chaos? Geophys Res Lett 17:223–226
38. Jaumé SC, Sykes LR (1999) Evolving towards a critical point: A review of accelerating seismic moment/energy release prior to large and great earthquakes. Pure Appl Geophys 155:279–306
39. Jones LM, Molnar P (1979) Some characteristics of foreshocks and their possible relation to earthquake prediction and premonitory slip on faults. J Geophys Res 84:3596–3608
40. Kagan YY, Knopoff L (1978) Statistical study of the occurrence of shallow earthquakes. J Geophys R Astron Soc 55:67–86
41. Keilis-Borok VI, Soloviev AA (2003) Nonlinear Dynamics of the Lithosphere and Earthquake Prediction, Springer Series in Synergetics. Springer, Heidelberg
42. Lomnitz-Adler J (1999) Automaton models of seismic fracture: constraints imposed by the magnitude-frequency relation. J Geophys Res 98:17745–17756
43. Main IG, O'Brian G, Henderson JR (2000) Statistical physics of earthquakes: Comparison of distribution exponents for source area and potential energy and the dynamic emergence of log-periodic quanta. J Geophys Res 105:6105–6126

44. Matthews MV, Ellsworth WL, Reasenberg PA (2002) A Brownian model for recurrent earthquakes. Bull Seism Soc Am 92:2233–2250
45. Narteau C, Shebalin P, Hainzl S, Zöller G, Holschneider M (2003) Emergence of a band-limited power law in the aftershock decay rate of a slider-block model. Geophys Res Lett 30:1568; doi:10.1029/2003GL017110
46. Nur A, Booker JR (1972) Aftershocks caused by pore fluid flow? Science 175:885–887
47. Okada Y (1992) Internal deformation due to shear and tensile faults in a half space. Bull Seism Soc Am 82:1018–1040
48. Olami Z, Feder HS, Christensen K (1992) Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes. Phys Rev Lett 68:1244–1247
49. Omori F (1894) On the aftershocks of earthquakes. J Coll Sci Imp Univ Tokyo 7:111–200
50. Patel JK, Kapadia CH, Owen DB (1976) Handbook of statistical distributions. Marcel Dekker, New York
51. Reasenberg P (1985) Second-order moment of central California seismicity. J Geophys Res 90:5479–5495
52. Reid HF (1910) The Mechanics of the Earthquake, The California Earthquake of April 18, 1906. Report of the State Investigation Commission, vol 2. Carnegie Institution of Washington, Washington
53. Rundle JB, Klein W, Turcotte DL, Malamud BD (2000) Precursory seismic activation and critical point phenomena. Pure Appl Geophys 157:2165–2182
54. Saleur H, Sammis CG, Sornette D (1996) Discrete scale invariance, complex fractal dimensions, and log-periodic fluctuations in seismicity. J Geophys Res 101:17661–17677
55. Savage JC, Svarc JL, Prescott WH (1999) Geodetic estimates of fault slip rates in the San Francisco Bay area. J Geophys Res 104:4995–5002
56. Scholz CH (1998) Earthquakes and friction laws. Nature 391: 37–42
57. Shcherbakov R, Turcotte DL (2004) A damage mechanics model for aftershocks. Pure Appl Geophys 161:2379; doi:10.1007/s00024-004-2570-x
58. Shin TC, Teng TL (2001) An overview of the 1999, Chichi, Taiwan, earthquake. Bull Seismol Soc Am 91:895–913
59. Sornette D (2004) Self-organization and Disorder: Concepts & Tools, Springer Series in Synergetics. Springer, Heidelberg
60. Sornette D, Sammis CG (1995) Complex critical exponents from renormalization group theory of earthquakes: Implication for earthquake predicitions. J Phys 1(5):607–619
61. Sornette D, Sornette A (1999) Renormalization of earthquake aftershocks. Geophys Res Lett 6:1981–1984
62. Field EH et al. (2007) Special Issue: Regional Earthquake Likelihood Models. Seismol Res Lett 78:1
63. Steacy SJ, McCloskey J, Bean CJ, Ren JW (1996) Heterogeneity in a self-organized critical earthquake model. Geophys Res Lett 23:383–386
64. Turcotte DL (1997) Fractals and chaos in geology and geophysics. Cambridge University Press, New York
65. Turcotte DL, Newman WI, Shcherbakov R (2003) Micro and macroscopic models of rock fracture. J Geophys Int 152:718–728
66. Utsu T (2002) Statistical features of seismicity. In: Int Assoc Seismol & Phys Earth's Interior (ed) International handbook of earthquake and engineering seismology, vol 81A. Academic Press, San Diego, pp 719–732
67. Utsu T, Ogata Y, Matsu'ura RS (1995) The centenary of the Omori formula for a decay law of aftershock activity. J Phys Earth 43:1–33
68. Wesnousky SG (1994) The Gutenberg–Richter or characteristic earthquake distribution, which is it? Bull Seismol Soc Am 90:525–530; 84:1940–1959
69. Wiemer S, Baer M (2000) Mapping and removing quarry blast events from seismic catalogs: Examples from Alaska, the Western United States, and Japan. Bull Seismol Soc Am 90:525–530
70. Working Group on California Earthquake Probabilities (2003) Earthquake probabilities in the San Francisco Bay region. US Geol Survey Open File Report 03–214, US Geological Survey
71. Wyss M (1997) Cannot earthquakes be predicted? Science 278:487
72. Wyss M, Habermann RE (1988) Precursory seismic quiescence. Pure Appl Geophys 126:319–332
73. Zaliapin I, Liu Z, Zöller G, Keilis-Borok V, Turcotte DL (2002) On increase of earthquake correlation length prior to large earthquakes in California. Comp Seismol 33:141–161
74. Zöller G, Ben-Zion Y, Holschneider M, Hainzl S (2007) Estimating recurrence times and seismic hazard of large earthquakes on an individual fault. J Geophys Int 170:1300–1310; doi:10.1111/j.1365-246X200703480.x
75. Zöller G, Hainzl S (2001) Detecting premonitory seismicity patterns based on critical point dynamics. Nat Hazards Earth Syst Sci 1:93–98
76. Zöller G, Hainzl S (2002) A systematic spatiotemporal test of the critical point hypothesis for large earthquakes. Geophys Res Lett 29:1558; doi:10.1029/2002GL014856
77. Zöller G, Hainzl S, Kurths J (2001) Observation of growing correlation length as an indicator for critical point behavior prior to large earthquakes. J Geophys Res 106:2167–2175
78. Zöller G, Hainzl S, Kurths J, Zschau J (2002) A systematic test on precursory seismic quiescence in Armenia. Nat Hazards 26:245–263
79. Zöller G, Holschneider M, Ben-Zion Y (2004) Quasi-static and quasi-dynamic modeling of earthquake failure at intermediate scales. Pure Appl Geophys 161:2103–2118; doi:10.1007/s00024-004-2551-0
80. Zöller G, Holschneider M, Ben-Zion Y (2005) The role of heterogeneities as a tuning parameter of earthquake dynamics. Pure Appl Geophys 162:1027; doi:10.1007/s00024-004-2660-9
81. Zöller G, Hainzl S, Holschneider M, Ben-Zion Y (2005) Aftershocks resulting from creeping sections in a heterogeneous fault. Geophys Res Lett 32:L03308; doi:10.1029/2004GL021871
82. Zöller G, Hainzl S, Ben-Zion Y, Holschneider M (2006) Earthquake activity related to seismic cycles in a model for a heterogeneous strike-slip fault. Tectonophys 423:137–145; doi:10.1016/j.tecto.2006.03.007

Seismicity, Statistical Physics Approaches to

Didier Sornette[1], Maximilian J. Werner[2]

[1] Department of Management, Technology and Economics, ETH Zurich, Switzerland

[2] Swiss Seismological Service, Institute of Geophysics, ETH Zurich, Switzerland

Article Outline

Glossary
Definition of the Subject
Introduction
Concepts and Calculational Tools
Competing Mechanisms and Models
Empirical Studies of Seismicity Inspired by Statistical Physics
Future Directions
Bibliography

Glossary

Chaos Chaos occurs in dynamical systems with two ingredients: (i) nonlinear recurrent re-injection of the dynamics into a finite domain in phase space and (ii) exponential sensitivity of the trajectories in phase space to initial conditions.

Continuous phase transitions If there is a finite discontinuity in the first derivative of the thermodynamic potential, then the phase transition is termed first-order. During such a transition, a system either absorbs or releases a fixed amount of latent heat (e. g. the freezing/melting of water/ice). If the first derivative is continuous but higher derivatives are discontinuous or infinite, then the phase transition is called continuous, of the second kind, or critical. Examples include the critical point of the liquid–gas transition, the Curie point of the ferromagnetic transition, or the superfluid transition [127,235].

Critical exponents Near the critical point, various thermodynamic quantities diverge as power laws with associated critical exponents. In equilibrium systems, there are scaling relations that connect some of the critical exponents of different thermodynamic quantities [32,127,203,216,235].

Critical phenomena Phenomena observed in systems that undergo a continuous phase transition. They are characterized by scale invariance: the statistical properties of a system at one scale are related to those at another scale only through the ratio of the two scales and not through any one of the two scales individually. The scale invariance is a result of fluctuations and correlations at all scales, which prevents the system from being separable in the large scale limit at the critical point [32,203,235].

Declustering In studies of seismicity, declustering traditionally refers to the deterministic identification of fore-, main- and aftershocks in sequences (or clusters) of earthquakes clustered in time and space. Recent, more sophisticated techniques, e. g. stochastic declustering, assign to earthquakes probabilities of being triggered or spontaneous.

Dynamical scaling and exponents Non-equilibrium critical phase transitions are also characterized by scale invariance, scaling functions and critical exponents. Furthermore, some evidence supports the claim that universality classes also exist for non-equilibrium phase transitions (e. g. the directed percolation and the Manna universality class in sandpile models), although a complete classification of classes is lacking and may in fact not exist at all. Much interest has recently focused on directed percolation, which, as the most common universality class of absorbing state phase transitions, is expected to occur in many physical, chemical and biological systems [85,135,203].

Finite size scaling If a thermodynamic or other quantity is investigated at the critical point under a change of the system size, the scaling behavior of the quantity with respect to the system size is known as finite size scaling [32]. The quantity may refer to a thermodynamic quantity such as the free energy or it may refer to an entire probability distribution function. At criticality, the sole length scale in a finite system is the upper cut-off s_c, which diverges in the thermodynamic limit $L \to \infty$. Assuming a lower cut-off $s_0 \ll s_c, s$, a finite size scaling ansatz for the distribution $P(s; s_c)$ of the observable variable s, which depends on the upper cut-off s_c is then given by:

$$P(s; s_c) = a s^{-\tau} G(s/s_c) \quad \text{for} \quad s, s_c \gg s_0 \,, \tag{1}$$

where the parameter a is a non-universal metric factor, τ is a universal (critical) exponent, and G is a universal scaling function that decays sufficiently fast for $s \gg s_c$ [32,36]. Pruessner [163] provides a simple yet instructive and concise introduction to scaling theory and how to find associated exponents. System-specific corrections appear to sub-leading order.

Fractal A deterministic or stochastic mathematical object that is defined by its exact or statistical self-similarity at all scales. Informally, it often refers to a rough or fragmented geometrical shape which can be subdivided into parts which look approximately the same as the original shape. A fractal is too irregular to be described by Euclidean geometry and has a fractal dimension that is larger than its topological dimension but less than the dimension of the space it occupies.

Mean-Field An effective or average interaction field designed to approximately replace the interactions from many bodies by one effective interaction which is constant in time and space, neglecting fluctuations.

Mechanisms for power laws Power laws may be the hallmark of critical phenomena, but there are a host of other mechanisms that can lead to power laws (see Chapter 14 of [203] for a list of power law mechanisms as well as [37,143]). Observations of scale invariant statistics therefore do not necessarily imply SOC, of course. Power laws express the existence of a symmetry (scale invariance) and there are many mechanisms by which a symmetry can be obtained or restored.

Non-equilibrium phase transitions In contrast to systems at equilibrium, non-equilibrium phase transitions involve dynamics, energy input and dissipation. Detailed balance is violated and no known equivalent of the partition function exists, from which all thermodynamic quantities of interest derive in equilibrium. Examples of non-equilibrium phase transitions include absorbing state phase transitions, reaction-diffusion models, and morphological transitions of growing surfaces [85,135].

Phase transitions In (equilibrium) statistical mechanics, a phase transition occurs when there is a singularity in the free energy or one of its derivatives. Examples include the freezing of water, the transition from ferromagnetic to paramagnetic behavior in magnets, and the transition from a normal conductor to a superconductor [127,235].

Renormalization group theory A mathematical theory built on the idea that the critical point can be mapped onto a fixed point of a suitably chosen transformation on the system's Hamiltonian. It provides a foundation for understanding scaling and universality and provides tools for calculating exponents and scaling functions. Renormalization group theory provides the basis for our understanding of critical phenomena [32,216,235]. It has been extended to non-Hamiltonian systems and provides a general framework for constructing theories of the macro-world from the microscopic description.

Self-organized criticality (SOC) Despite two decades of research since its inception by [13] and the ambitious claim by [11] that, as a mechanism for the ubiquitous power laws in Nature, SOC was "How Nature Works", a commonly accepted definition along with necessary and sufficient conditions for SOC is still lacking [93,163,203]. A less rigorous definition may be the following: Self-organized criticality refers to a non-equilibrium, critical and marginally stable steady-state, which is attained spontaneously and without (explicit) tuning of parameters. It is characterized by power law event distributions and fractal geometry (in some cases) and may be expected in slowly driven, interaction-dominated threshold systems [93]. Some authors additionally require that temporal and/or spatial correlations decay algebraically (e. g. [84], but see [163]). Definitions in the literature range from broad (simply the absence of characteristic length scales in non-equilibrium systems) to narrow (the criticality is due to an underlying continuous phase transition with all of its expected properties) (see, e. g., [162] for evidence that precipitation is an instance of the latter definition of SOC in which a non-linear feedback of the order parameter on the control parameter turns a critical phase transition into a self-organized one attracting the dynamics [198]).

Spinodal decomposition In contrast to the slow process of phase separation via nucleation and slow growth of a new phase in a material inside the metasstable region near a first-order phase transition, spinodal decomposition is a non-equilibrium, rapid and critical-like dynamical process of phase separation that occurs quickly and throughout the material. It needs to be induced by rapidly quenching the material to reach a sub-area (sometimes a line) of the unstable region of the phase diagram which is characterized by a negative derivative of the free energy.

Statistical physics is the set of concepts and mathematical techniques allowing one to derive the large-scale laws of a physical system from the specification of the relevant microscopic elements and of their interactions.

Turbulence In fluid mechanics, turbulence refers to a regime in which the dynamics of the flow involves many interacting degrees of freedom, and is very complex with intermittent velocity bursts leading to anomalous scaling laws describing the energy transfer from injection at large scales to dissipation at small scales.

Universality In systems with little or no frozen disorder, equilibrium continuous phase transitions fall into a small set of universality classes that are characterized by the same critical exponents and by certain scaling functions that become identical near the critical point. The class depends only on the dimension of the space and the dimension of the order parameter. For instance, the critical point of the liquid–gas transition falls into the same universality class as the 3D Ising model. Even some phase transitions occurring in high-energy physics are expected to belong to the Ising class. Universality justifies the development and study of extremely simplified models (caricatures) of Nature, since the behavior of the system at the critical point can nevertheless be captured (in some cases exactly). How-

ever, non-universal features remain even at the critical point but are less important, e. g. amplitudes of fluctuations or system-specific corrections to scaling that appear at sub-leading order [32,216,235,239].

Definition of the Subject

A fundamental challenge in many scientific disciplines concerns upscaling, that is, of determining the regularities and laws of evolution at some large scale from those known at a lower scale: biology (from molecules to cells, from cells to organs); neurobiology (from neurons to brain function), psychology (from brain to emotions, from evolution to understanding), ecology (from species to the global web of ecological interactions), condensed matter physics (from atoms and molecules to organized phases such as solid, liquid, gas, and intermediate structures), social sciences (from individual humans to social groups and to society), economics (from producers and consumers to the whole economy), finance (from investors to the global financial markets), Internet (from e-pages to the world wide web 2.0), semantics (from letters and words to sentences and meaning), and so on. Earthquake physics is no exception, with the challenge of understanding the transition from the laboratory scale (or even the microscopic and atomic scale) to the scale of fault networks and large earthquakes.

Statistical physics has had a remarkably successful track record in addressing the upscaling problem in physics. While the macroscopic laws of thermodynamics have been established in the 19th century, their microscopic underpinning were elaborated in the early 20th century by Boltzmann and followers, building the magnificent edifice of statistical physics. Statistical physics can be defined as the set of concepts and mathematical techniques allowing one to derive the large-scale laws of a physical system from the specification of the relevant microscopic elements and of their interactions. Dealing with huge ensembles of elements (atoms, molecules) of the order of the Avogadro number ($\simeq 6 \cdot 10^{23}$), statistical physics uses the mathematical tools of probability theory combined with other relevant fields of physics to calculate the macroscopic properties of large populations.

One of the greatest achievement of statistical physics was the development of the renormalization group analysis, to construct a theory of interacting fields and of critical phase transitions. The renormalization group is a perfect example of how statistical physics addresses the micro-macro upscaling problem. It decomposes a problem of finding the macroscopic behavior of a large number of interacting parts into a succession of simpler problems with a decreasing number of interacting parts, whose effective properties vary with the scale of observation. The renormalization group thus follows the proverb "divide to conquer" by organizing the description of a system scale-by-scale. It is particularly adapted to critical phenomena and to systems close to being scale-invariant. The renormalization group translates into mathematical language the concept that the overall behavior of a system is the aggregation of an ensemble of arbitrarily defined sub-systems, with each sub-system defined by the aggregation of sub-subsystems, and so on [203].

It is important to stress that up to now the term "statistical" has different meanings in statistical physics and in statistical seismology, a field that has developed as a marriage between probability theory, statistics and the part of seismology concerned with empirical patterns of earthquake occurrences [225] (but not with physics). Statistical seismology uses stochastic models of seismicity, which are already effective large-scale representation of the dynamical organization. In contrast, a statistical physics approach to earthquake strives to derive these statistical models or other descriptions from the knowledge of the microscopic laws of friction, damage, rupture, rock-water interactions, mechano-chemistry and so on, at the microscopic scales [200,201]. In other words, what is often missing in statistical seismology is the physics to underpin the stochastic model on physically-based laws, e. g. rate-and-state friction [55].

The previously mentioned successes of statistical physics promote the hope that a similar program can be developed for other fields, including seismology. The successes have been more limited, due to the much more complex interplay between mechanisms, interactions and scales found in these out-of-equilibrium systems. This short essay provides a subjective entry to understand some of the different attempts, underlining the few successes, the problems and open questions. Rather than providing an exhaustive review, we mention what we believe to be important topics and have especially included recent work.

Introduction

Much of the recent interest of the statistical physics community has focused on applying scaling techniques, which are common tools in the study of critical phenomena, to the statistics of inter-event recurrence times or waiting times [14,40,41,42,43,44,46,48,134]. However, the debate over the relevance of critical phenomena to earthquakes stretches back as far as 30 years [7,11,12,47,61,84,93,104,106,109,114,147,151,178,192,193,194,202,203,207,223], ► Jerky Motion in

Slowly Driven Magnetic and Earthquake Fault Systems, Physics of and ► Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space. The current debate on recurrence statistics is thus the latest tack in an evolving string of arguments with a long history. As discussed below, many of the claims made in the recent articles on recurrence statistics have either been challenged, refuted or explained by previously known facts about earthquake statistics [132,133,145,176,177,230]. As will be discussed below, this debate in the literature is important because of the potential consequences for understanding earthquakes, but it needs to be pursued with rigorous scientific arguments accessible to both the seismological and the statistical physics communities.

The debate would almost certainly benefit significantly from testing hypotheses with simulations to establish null hypotheses and benchmarks: seismicity patterns are sufficiently stochastic and earthquake catalogs contain a sufficient amount of observational uncertainties so as to make inference difficult. It is often not straightforward to predict the signal of well-known statistical features such as clustering in new data analysis techniques. Therefore, testing the purported claims by realistic simulations of earthquake catalogs can provide a strong benchmark against which the claims can be evaluated. This view and the corresponding criticism of many studies has been put forward and defended for a long time by Kagan [111].

Such a model-dependent approach may be at odds with the philosophy of a so-called "model-free" analysis, which the community of statistical physicists claim to take in their analysis. For instance, network theory-based approaches, space-time window-based finite size scaling, box-covering methods and other techniques used in the study of critical and fractal phenomena are said to be "model-free" because no assumptions about seismicity are supposedly made at the outset. By using model-free analysis techniques, the often uncertain and sometimes clearly wrong assumptions of flawed models and resulting biased results are meant to be circumvented.

However, as is almost always the case in statistical hypothesis testing, the less assumptions are made about the test, the less powerful the test statistic. More importantly, seismicity is sufficiently stochastic so that well-known features may appear as novel in new analysis methods. Furthermore, to convince the seismological community of new data analysis techniques, the methods need to be tested on established knowledge and show the improvement over traditional methods. These types of initial tests are rarely performed by the statistical physics community.

In the next Sect. "Concepts and Calculational Tools", we present a summary of some of the concepts and calculational tools that have been developed in attempts to apply statistical physics approaches to seismology. Then, Sect. "Competing Mechanisms and Models" summarizes the leading theoretical physical models of the space-time organization of earthquakes. Section "Empirical Studies of Seismicity Inspired by Statistical Physics" presents a general discussion and several examples of the new metrics proposed by statistical physicists, underlining their strengths and weaknesses. Section "Future Directions" briefly outlines expected developments.

Concepts and Calculational Tools

Renormalization, Scaling and the Role of Small Earthquakes in Models of Triggered Seismicity

A common theme in many of the empirical relations in seismology (and in those employed in seismicity models) is the lack of a dominating scale. Many natural phenomena can be approached by the traditional reductionist approach to isolate a process at a particular scale. For example, the waves of an ocean can be described quite accurately by a theory that entirely ignores the fact that the liquid is made out of individual molecules. Indeed, the success of most practical theories in physics depends on isolating a scale [234], although since this recognition, much progress has been made in developing a holistic approach for processes that do not fall into this class. Given current observational evidence, earthquakes seem to belong to the set of processes characterized by a lack of one dominating length scale: fluctuations of many or perhaps a wide continuum of sizes seem to be important and are in no way diminished – even when one is interested solely in large-scale descriptions [206].

The traditional reductionist approach in seismology, which, for instance, attempted to separate large (main) shocks from small (fore- or after-) shocks, is slowly giving way to the holistic approach, in which all earthquakes are created equal and seismicity is characterized by fluctuations of all sizes. This gradual shift is supported, on a conceptual and qualitative level, by the vision of critical phenomena. A particularly strong model of the interactions between earthquakes has emerged in the concept of triggering, which places all earthquakes on the same footing: each earthquake can trigger its own events, which in turn can trigger their own events, and so on, according to the same probability distributions, and the resulting seismicity can be viewed as the superposed cascades of triggered earthquakes that cluster in space and time [77,117,149,150].

From this point of view, it is natural that small earthquakes are important to the overall spatio-temporal patterns of seismicity. Indeed, the scaling of aftershock productivity with mainshock magnitude suggests that small earthquakes are cumulatively as important for the triggered seismicity budget as rarer but larger events [60,76,82]. The importance of small earthquakes has also been documented in, e. g., [73,138,141].

But earthquake catalogs do not contain information (by definition) about the smallest, unobserved events, which we know to exist from acoustic emission experiments and earthquakes recorded in mines. To guarantee a finite seismicity budget, Sornette and Werner argued [209] for the existence of a smallest triggering earthquake, akin to a "ultra-violet cut-off" in quantum field theory, below which earthquakes do not trigger other events. Introducing a formalism which distinguishes between the detection threshold and the smallest triggering earthquake, Sornette and Werner placed constraints on its size by using a simplified version of the popular Epidemic-Type Aftershock Sequence (ETAS) Model [149], a powerful model of triggered seismicity based on empirical statistics, and by using observed aftershock sequences. Sornette and Werner [210] also considered the branching structure of one complete cascade of triggered events, deriving an apparent branching ratio and the apparent number of untriggered events, which are observed when only the structure above the detection threshold is known. As a result of our inability to observe the entire branching structure, inferred clustering parameters are significantly biased and difficult to interpret in geophysical terms. Second, separating triggered from untriggered events, commonly known as declustering, also strongly depends on the threshold, so that it cannot even in theory constitute a physically sound method.

Sornette and Werner [210] also found that a simplified, averaged version of the ETAS model can be renormalized onto itself, with effective clustering parameters, under a change of the threshold. Saichev and Sornette [175] confirmed these results for the stochastic number statistics of the model using a rigorous approach in terms of generating probability functions, but also showed that the temporal statistics could not be renormalized. Furthermore, it can be shown (see Chapter 4 of [229]) that the conditional intensity function of the ETAS model, the mathematical object which uniquely defines the model, cannot be renormalized onto itself under a change of magnitude threshold. It is not a fixed-point of the renormalization process operating via magnitude coarse-graining. The functional form of the model must change under a change in the detection threshold [175]. In other words, if earthquakes occur according to an ETAS model above some cut-off m_0, then earthquakes above m_d cannot be described by the ETAS model in a mathematically exact way. Although in practice, the ETAS model provides an excellent fit.

The issue of how to deal with small earthquakes is thus reminiscent of the decades of efforts that have been invested in physics to deal with the famous ultra-violet cut-off problem, eventually solved by the so-called "renormalization" theory of Feynmann, Schwinger and Tomonaga. In the 1960s and 1970s, this method of renormalization was extended into the "renormalization group" (in fact a semi-group in the strict mathematical sense) for the theory of critical phenomena (see glossary), which we also mention in Sect. "Competing Mechanisms and Models". It is fair to say that there has been limited success in developing a multi-scale description of the physics of earthquakes and, in particular, in addressing the upscaling problem and the impact of the many small earthquakes.

One tantalizing approach, not yet really understood in terms of all its consequences and predictions, is the variant of the ETAS model proposed by Vere-Jones [224], which has the remarkable property of being bi-scale invariant under a transformation involving time and magnitudes. One of the modifications brought in by [224] is to assume that the distribution of the daughter magnitudes is dependent on the mother magnitude m_i through a modification of the Gutenberg–Richter distribution of triggered earthquake magnitudes by a term of the form $\exp(-\delta|m - m_i|)$, where $\delta > 0$ quantifies the distance to the standard Gutenberg–Richter distribution. Remarkably, Saichev and Sornette [174], who studied the Vere-Jones model, found that, due to the superposition of the many magnitude distributions of each earthquake in the cascades of triggered events, the resulting distribution of magnitudes over a stationary catalog is a pure Gutenberg–Richter law. Thus, there might be hidden characteristic scales in the physics of triggering that are not revealed by the standard observable one-point statistical distributions. Simulation and parameter estimation algorithms for the Vere-Jones model are not yet available. If and when these algorithms become available, the study of this bi-scale invariant branching model may be a strong alternative to the ETAS model, as this model is exactly scale invariant with neither ultra-violet nor infra-red cut-offs.

Nevertheless, being empirically based, these stochastic point process models lack a genuine microscopic physical foundation. The underlying physics is not explicitly addressed and only captured effectively by empirical statistics, even at the smallest scales. The physical processes and their renormalization are missing in this approach.

Universality

Universality, as defined in the glossary, justifies the development and study of extremely simplified models (caricatures) of Nature, since the behavior of a studied system at the critical point can nevertheless be captured by toy models (in some cases exactly). For instance, the liquid–gas transition, the ferromagnetic to paramagnetic transition, and the behavior of binary alloys, all apparently different systems, can be described successfully by an extremely simplified picture of Nature (the Ising model) because of universality. The hope that a similar principle holds for earthquakes (and other non-equilbrium systems) underpins many of the models and tools inspired by statistical physics that have been applied to seismicity, to bring about the "much coveted revolution beyond reductionism" [17,67].

Speaking loosely, the appearance of power laws in many toy models is often interpreted as a kind of universal behavior. The strict interpretation of universality classes, however, requires that critical exponents along with scaling functions are identical for different systems. It is interesting to note that slight changes in the sand-pile model already induce new universality classes, so that even within a group of toy models, the promise of universality is not, strictly, fulfilled [13,98,136].

A lively debate in seismology concerns the universality of, on one hand, the frequency-size distribution of earthquake magnitudes (e. g. [47,61,91,108,111,232]), and, on the other hand, the universality of the exponent of the Gutenberg–Richter distribution (e. g. [25,184,212,233]). A spatio-temporally varying critical exponent is not traditionally part of the standard critical phenomena repertoire, although analytical and numerical results based on a simple earthquake model on a fault showed a possible spontaneous switching between Gutenberg–Richter and characteristic earthquake behavior associated with a non-equilibrium phase transition [24,47,61], ► Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of and ► Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space. Another possible mechanism for the coexistence of and intermittent shifts between different regimes (Gutenberg–Richter scaling and characteristic earthquakes) stems from the competition between several interacting faults which give rise to long intervals of activity in some regions followed by similarly long intervals of quiescence [129,213,214]. Both careful empirical investigations of seismicity parameters and theoretical progress on heterogeneous, spatially extended critical phenomena may help elucidate the controversy.

Intermittent Periodicity and Chaos

As part of the conquest of chaos theory in the 70s and 80s [58,139], its concepts and methods were invariably also applied to seismicity. Huang and Turcotte [87,88] modeled the interaction between two faults by two sliding blocks that are driven by a plate through springs, coupled to one another via another spring and endowed with a velocity-weakening friction law. The dynamical evolution of the blocks showed chaotic behavior and period-doubling (the Feigenbaum route to chaos [58]). Huang and Turcotte [87,88] suggested that the interaction of the Parkfield segment with the southern San Andreas fault may be governed by the kind of chaotic behavior they observed in their model: the lighter block slipped quasi-periodically for several times until both slipped together. Their study explained that apparent quasi-periodicity of earthquakes on fault segments may be a result of chaotic interactions between many fault segments, thereby providing a warning that the extrapolations of quasi-periodic models are not to be trusted (e. g. [15,75] and references therein). However, it is doubtful that models with just a few degrees of freedom can go a long way towards providing deeper physical insights, or predictive tools. One needs to turn to models with a large number of degrees of freedom, for which turbulence appears as the leading paradigm of complexity.

Turbulence

A drastically different approach has been favored by Yan Kagan [104,106,109,114], who described seismicity as the "turbulence of solids" – attesting to the far greater problems in earthquake seismology than the theory of critical phenomena promises to solve. While renormalization group methods and scaling theory have contributed immensely to the study of turbulence [63], the problem of turbulence involves significant additional complications. First, loosely speaking, renormalization group theory helps predict global behavior by coarse-graining over degrees of freedom, which is essentially a bottom–up approach. In turbulence, the enstrophy acts bottom–up, but the energy cascades top down. Secondly, there is a significant spatial and topological aspect to turbulence, for instance involving topological defects, such as filament structures, which are crucial to the dynamical evolution of the system. The existence of the two cascades (top–down and bottom–up) as well as the influence of the dissipation scale all the way within the so-called inertial range makes turbulence the most important problem still unsolved in classical Physics. The importance of addressing the issue of the interplay between the top–down and bottom–up cas-

cades in earthquake toy models has been outlined by [65,66,236,237].

The analogous problem for seismicity lies in the complex fault network, which constrains seismicity through its weak structures but also grows and evolves because of earthquakes redistributing stresses and rupturing fresh surfaces. The statistical description of this tensorial and dynamical problem is only at its beginning [64,99,100,101,102,103,107,116,118,119,142,213]. But it is likely to be a key aspect to the dynamical evolution of faults and seismicity. New physics and approaches are required to tackle the tensorial nature of the stress and strain fields and the complex topological structures of defects, from dislocations to joints and faults, as well as the many different physical processes operating from the atomic scale to the macro-scale [200,201].

Self-Organized Criticality

Self-organized criticality (SOC) refers to the spontaneous organization of a system driven from outside into a globally dynamical statistical stationary state, which is characterized by self-similar distributions of event sizes and sometimes fractal geometrical properties. SOC applies to the class of phenomena occurring in driven out-of-equilibrium systems made of many interactive components, which possess the following fundamental properties: 1) a highly non-linear behavior, 2) a very slow driving rate, 3) a globally stationary regime, characterized by stationary statistical properties, and 4) power-law distributions of event sizes and fractal geometrical properties. The crust obeys these four conditions, as first suggested by [12,193], who proposed to understand the spatio-temporal complexity of earthquakes in this perspective.

The appeal of placing the study of earthquakes in the framework of critical phenomena may be summarized as follows. Power law distributions can be understood as a result of an underlying continuous phase transition into which the crust has organized itself [197,211]. Applying the methods of renormalization group theory may help calculate exponents and scaling functions and rationalize the spatio-temporal organization of seismicity along with its highly correlated structures. For instance, Sornette and Virieux [208] provided a theoretical framework which links the scaling laws exhibited by earthquakes at short times and plate tectonic deformations at large times. Perhaps earthquakes fall into a universality class which can be solved exactly and/or investigated in toy models. Moreover, studying the detailed and highly complicated microphysics involved in earthquakes may not lead to insights about the spatio-temporal organization, because, as a critical phenomenon, the traditional approach of separating length scales to describe systems is inadequate. On the other hand, as mentioned above, there is the possibility of a hierarchy of physical processes and scales which are inter-related [156,157], for which the simplifying approach in terms of critical phenomena is likely to be insufficient.

As another reason for the importance of the topic, interesting consequences for the predictability of earthquakes might be derived, for instance by mapping earthquakes to a genuine critical point (the accelerating moment release hypothesis, e. g. [202,207]) or by mapping earthquakes to SOC (e. g. [70,147]). The latter mapping had led some to argue that earthquakes are inherently unpredictable. In the sandpile paradigm [13], there is little difference between small and large avalanches, and this led similarly to the concept that "large earthquakes are small earthquakes that did not stop," hence their supposed lack of predictability. More than ten years after this contentious proposal, a majority of researchers, including most of the authors of this "impossibility claim," recognize that there is some degree of predictability [83,97]. Actually, the clarifications came from investigators of SOC, who recognized that the long-term organization of sandpiles [51,52] and of toy models of earthquakes and fault networks [142,213,214] is characterized by long-range spatial and temporal correlations. Thus, large events may indeed be preceded by subtle long-range organizational structures, an idea at the basis of the accelerating moment release hypothesis. This idea is also underlying the pattern recognition method introduced by Gelfand et al. [69] and developed extensively by V. Keilis-Borok and his collaborators for earthquake predictions [122]. In addition, Huang et al. [89] showed that avalanche dynamics occurring within hierarchical geometric structures are characterized by significant precursory activity before large events; this provides a clear proof of the possible coexistence between critical-like precursors of large events and a long-term self-organized critical dynamical state.

In summary, self-organized criticality provides a general conceptual framework to articulate the search for a physical understanding of the large-scale and long-time statistical properties of the seismogenic process and of the predictability of earthquakes. Beyond this, it is of little help as many different mechanisms have been documented at the origin of SOC (see, e. g., chapter 15 in [203]). SOC is not a theory, it does not provide any specific calculation tools; it is a concept offering a broad classification of the kinds of dynamics that certain systems, including the Earth crust, seem to spontaneously select.

Competing Mechanisms and Models

It should be noted at this point that the statistical physics approach to earthquake science is not limited to SOC. Over the years, several groups have proposed to apply the concepts and tools of statistical physics to understanding the multiscale dynamics of earthquake and fault systems. Various mechanisms drawn conceptually from statistical mechanics but not necessary even limited to critical (phase transition) phenomena have been proposed and are being pursued. Such approaches include the concept of the critical point earthquake related to accelerated moment release, network theory, percolation and fiber models as models for fracture, and many more, some of which can be found in [84,203,220,221].

In this section, we outline some of the major model classes which underpin distinct views on what are the dominating mechanisms to understand earthquakes and their space-time organization.

Roots of Complexity in Seismicity: Dynamics or Heterogeneity?

The 1990s were characterized by vigorous discussions at the frontier between seismology and statistical physics aimed at understanding the origin of the observed complexity of the spatio-temporal patterns of earthquakes. The debate was centered on the question of whether space-time complexity can occur on a single homogeneous fault, solely as a result of the nonlinear dynamics [23,38,39,128,186,187,188,189], associated with the slip and velocity dependent friction law validated empirically in particular by [53,54,55,56]. Or, is the presence of quenched heterogeneity necessary [21,22,126,168]?

The rediscovery of the multi-slider-block-spring model of [31,33] led to a flurry of investigations by physicists [34,128,170], finding an enticing entry to this difficult field, in the hope of capturing the main empirical statistical properties of seismicity. It is now understood that complexity in the stress field, in co-seismic slips and in sequences of earthquakes can emerge purely from the nonlinear laws. However, heterogeneity is probably the most important factor dominating the multi-scale complex nature of earthquakes and faulting [156,157,181,182]. It is also known to control the appearance of self-organized critical behavior in a class of models relevant to the crust [191,214].

Critical Earthquakes

This section gives a brief history of the "critical earthquake" concept.

We trace the ancestor of the critical earthquake concept to Vere-Jones [223], who used a branching model to illustrate that rupture can proceed through a cascade of damage events. Allègre et al. [7] proposed what is in essence a percolation model of damage/rupture describing the state of the crust before an earthquake. They formulated the model using the language of real-space renormalization group, in order to emphasize the multi-scale nature of the underlying physics, and the incipient rupture as the approach to a critical percolation point. Their approach is actually a reformulation in the language of earthquakes of the real-space renormalization group approach to percolation developed by [165]. Chelidze [35] independently developed similar ideas. In the same spirit, Smalley et al. [192] proposed a renormalization group treatment of a multi-slider-block-spring model. Sornette and Sornette [194] took seriously the concept put forward by [7] and proposed to test it empirically by searching for the predicted critical precursors. Voight [227,228] was probably the first author to introduce the idea of a time-to-failure analysis quantified by a second order nonlinear ordinary differential equation. For certain values of the parameters, the solution of [227,228]'s time-to-failure equation takes the form of a finite time singularity (see [180] for a review and [204] for a mechanism based on the ETAS model). He proposed and did use it later to predict volcanic eruptions. The concept that earthquakes are somehow associated with critical phenomena was also underlying the research efforts of a part of the Russian school [120,222].

The empirical seed for the critical earthquake concept were the repeated observations that large earthquakes are sometimes preceded by an increase in the number of intermediate size events [29,59,92,96,121,123,131,144,164,217]. The relation between these intermediate events and the subsequent main event took a long time to be recognized because the precursory events occur over such a large area. Sykes and Jaumé [217] proposed a specific law $\sim \exp[t/\tau]$ quantifying the acceleration of seismicity prior to large earthquakes. Bufe and Varnes [30] proposed that the finite-time singularity law

$$\epsilon_{\text{Benioff}} \sim 1/(t_c - t)^m \tag{2}$$

is a better empirical model than the exponential law. In (2), $\epsilon_{\text{Benioff}}$ is the cumulative Benioff strain, t_c is critical time of the occurrence of the target earthquake and m is a positive exponent. The fit with this law of the empirical Benioff strain calculated by summing the contribution of earthquakes in a given space-time window is supposed to provide the time t_c of the earthquake and thus constitutes

a prediction. This expression (2) was justified by a mechanical model of material damage. It is important to understand that the law (2) can emerge as a consequence of a variety of mechanisms, as reviewed by [180].

One of these mechanisms has been coined the "critical earthquake" concept, first formulated by Sornette and Sammis [207], who proposed to reinterpret the formula (2) proposed by [30] and previous related works by generalizing them within the statistical physics framework. This concept views a large earthquake as a genuine critical point. Using the insight of critical points in rupture phenomena, Sornette and Sammis [207] proposed to enrich Eq. (2), now interpreted as a kind of diverging susceptibility in the sense of critical phenomena, by considering complex exponents (i. e. log-periodic corrections to scaling). These structures accommodate the possible presence of a hierarchy of characteristic scales, coexisting with power laws expressing the scale invariance associated with a critical phenomenon [199]. This was followed by several extensions [89,94,95,178]. Sornette [202] reviewed the concept of critical "ruptures" and earthquakes with application to prediction. Ike and Sornette [90] presented a simple dynamical mechanism to obtain finite-time singularities (in rupture in particular) decorated by complex exponents (log-periodicity). Bowman et al. [28,153,154,242,243] proposed empirical tests of the critical earthquake concept. The early tests of [28] have been criticized by [74], while [226] commented on the lack of a formal statistical basis of the accelerating moment release model. This stresses the need for rigorous tests in the spirit of [166,167]. The debate is wide open, especially in view of the recent developments to improve the determination of the relevant spatio-temporal domain that should be used to perform the analyzes [26,27,124,130] (see [62] for a review).

Spinodal Decomposition

Klein, Rundle and their collaborators have suggested a mean-field approach to the multi-slider-block-spring model justified by the long-range nature of the elastic interactions between faults. This has led them to propose that the fluctuations of the strain and stress field associated with earthquakes are technically those occurring close to a spinodal line of an underlying first-order phase transition (see [125,169,172] and references therein). This conceptual view has inspired them to develop the "Pattern Informatics" technique, an empirical seismicity forecasting method based on the idea that changes in the seismicity rate are proxies for changes in the underlying stress [86,218].

The fluctuations associated with a spinodal line are very similar to those observed in critical phenomena. It is thus very difficult if not impossible in principle to falsify this hypothesis against the critical earthquake hypothesis, since both are expected to present similar if not identical signatures. Perhaps, the appeal of the spinodal decomposition proposal has to be found at the theoretical level, from the fact that first-order phase transitions are more generic and robust than critical phenomena, for systems where heterogeneity and quenched randomness are not too large.

Dynamics, Stress Interaction and Thermal Fluctuation Effects

Fisher et al. [24], Dahmen et al. [47], Ben-Zion et al. [61] and co-workers (see the reviews by ► Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of and ► Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space) have introduced a mean-field model (resulting from a uniform long-range Green function) of a single fault, whose dynamical organization is controlled by two control parameters, ϵ which measures the dynamic stress weakening and c which is the deviation from stress conservation (due for instance to coupling with ductile layers). The point ($\epsilon = 0$; $c = 0$) is critical in the sense of a phase transition in statistical physics, with its associated scale invariant fluctuations described by power laws. Dynamic stress strengthening ($\epsilon < 0$) leads to truncated Gutenberg–Richter power laws. Dynamic stress weakening ($\epsilon > 0$) is either associated with a truncated Gutenberg–Richter power law for $c > 0$ or with characteristic earthquakes decorating a truncated power law for $c < 0$. The coexistence of a characteristic earthquake regime with a power-law regime is particularly interesting as it suggests that they are not exclusive properties but may characterize the same underlying physics under slightly different conditions. This could provide a step towards explaining the variety of empirical observations in seismology [5,110,185,232].

Sornette et al. [214] have obtained similar conclusions using a quasi-static model in which faults grow and self-organize into optimal structures by repeated earthquakes. Depending on the value of dynamical stress drop (controlling the coupling strength between elements) relative to the amplitude of the frozen heterogeneity of the stress thresholds controlling the earthquake nucleation on each fault segment, a characteristic earthquake regime with a truncated power law is found for small heterogeneity or large stress drop while the power law SOC regime is recovered by large heterogeneity or small stress drop. The two

approaches of [24,47,61] and ► Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of on the one hand and of [214] on the other hand can be reconciled conceptually by noting that the dynamic stress weakening of ► Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of controls the dynamical generated stress heterogeneity while the lack of stress conservation c controls the coupling strength. Fundamentally, the relevant control parameter is the degree of coupling between fault elements seen as threshold oscillators of relaxation versus the variance of the disorder in their spontaneous large earthquake recurrence times. Generically, power law statistics are expected to co-exist with synchronized behavior in a general phase diagram in the heterogeneity-coupling strength plane [152,214].

Let us finally mention a promising but challenging theoretical approach, which has the ambition to bridge the small-scale physics controlled by thermal nucleation of rupture to the large-scale organization of earthquakes and faults [155,205]. Partial success has been obtained with a remarkable prediction on the ("multifractal") dependence of the Omori law exponent of aftershocks on the magnitude of the mainshock, verified by careful empirical analyzes on earthquakes in California, Japan and worldwide [158].

Empirical Studies of Seismicity Inspired by Statistical Physics

> *"False facts are highly injurious to the progress of science, for they often endure long; but false views, if supported by some evidence, do little harm, for everyone takes a salutary pleasure in proving their falseness."*
>
> Charles Darwin, in The Origin of Man, Chap. 6.

Early Successes and Subsequent Challenges

A significant benefit of the statistical physics approach to seismology has been the introduction of novel techniques to analyze the available empirical data sets, with the goal of obtaining new insights into the spatio-temporal organization of seismicity and of revealing novel regularities and laws that may guide the theoretical analysis.

A prominent forerunner is the application of the concept of fractals introduced by Mandelbrot [137] and of the measures of fractal dimensions to describe complex sets of earthquake epicenters, hypocenters and fault patterns. The use of fractals has constituted an epistemologic breakthrough in many fields, and not only in seismology. Indeed, before Mandelbrot, when dealing with most complex systems, one used to say: "this is too complicated for a quantitative analysis" and only qualitative descriptions were offered. After Mandelbrot, one could hear: "this is a fractal, with a fractal dimension equal to xxx!" By providing a new geometrical way of thinking about complex systems associated with novel metrics, Mandelbrot and his fractals have extended considerably the reach of quantitative science to many complex systems in all fields.

However, while there have been some attempts to use fractal dimensions as guidelines to infer the underlying organization processes, as for instance in [195,196], most of the initial reports have lost their early appeal [18,19,183,219] since the complexity of seismicity and faulting is much too great to be captured by scaling laws embodying solely a simple scale invariance symmetry. Among others, multifractal and adapted wavelet tools are needed to quantify this complexity, see for instance [68,146,156,157,159]. It should also be noted that few studies of the fractal dimensions of seismicity address the significant issues of errors, biases and incomplete records in earthquake catalogs – a notable exception being [115].

Since the beginning of the 21st century, a renewal of interest and efforts have burgeoned as groups of statistical physicists, interested in earthquakes as a potential instance of self-organized criticality (SOC), have claimed "novel", "universal" and "robust" scaling laws from their analysis of the spatio-temporal organization of seismicity. The authors purport to have discovered universal and hitherto unknown features of earthquakes that give new insights into the dynamics of earthquakes and add to the evidence that earthquakes are self-organized critical. We now discuss a few of these recent studies to illustrate the existence of potential problems in the "statistical physics" approach. In a nutshell, we show that perhaps most of these "novel scaling laws" can be explained entirely by already known statistical seismicity laws. This claim has been defended by other experts of statistical seismology, the most vocal being perhaps Yan Kagan at UCLA [111].

The flurry of interest from physicists comes from their fascination with the self-similar properties exhibited by seismicity (e. g. the Gutenberg–Richter power law of earthquake seismic moments, the Omori–Utsu law of the decay of aftershock rates after large earthquakes, the fractal and multifractal space-time organization of earthquakes and faults, etc.), together with the development of novel concepts and techniques that may provide new insights. But, and this is our main criticism based on several detailed examples discussed below, many of the new approaches and results do not stand close scrutiny. This failure is rooted in two short-comings: (i) the lack of testing of new methods on synthetic catalogs generated by benchmark mod-

els which are based on well-known statistical laws of seismicity, and (ii) the failure to consider earthquake catalog bias, incompleteness and errors. The latter may cause catalog artifacts to appear as genuine characteristics of earthquakes. Testing the results on a variety of catalogs and considering the influence of various catalog errors can help minimize their influence. The former short-coming often leads to the following scenario: authors fail to realize that a simpler null hypothesis could not be rejected, namely that their "discovery" could actually be explained by just a combination of basic statistical laws known to seismologists for decades.

The well-established laws of statistical seismicity that authors should consider before claiming for novelty include the following:

1. The Gutenberg–Richter law for the distribution of earthquake magnitudes with a b-value close to 1 (corresponding to an exponent $\simeq 2/3$ for the probability density function of seismic moments),
2. The Omori–Utsu law for the decay of the rate of aftershocks following a mainshock,
3. The inverse Omori law for foreshocks,
4. The fact that aftershocks also trigger their own aftershocks and so on, and that aftershocks do not seem to exhibit any distinguishable physical properties,
5. The fact that the distribution of distances between mainshocks and aftershocks has a power law tail,
6. The fertility law (the fact that earthquakes of magnitude M trigger of the order of 10^{aM} aftershocks with $a\sim \leq b \simeq 1$,
7. The fractal distribution of faults which are concentration centers for earthquakes.

This above non-exhaustive list selects "laws" which are arguably non-redundant, in the sense that it is likely not possible to derive one of these laws from the others (a possible exception is the inverse Omori law for foreshock, which can be derived from the direct Omori law for aftershocks in the context of the ETAS model [78,81]). Some experts would argue that we should add to this list other claimed regularities, such as "Båth's law" (see e. g. [190] for a recent discussion emphasizing the importance of this law), that states that the differences in magnitudes between mainshocks and their largest aftershocks are approximately constant, independent of the magnitudes of mainshocks [20]. However, Helmstetter and Sornette [79] and Saichev and Sornette [173] have shown that Bath's law can be accurately recovered in ETAS-type models combining the first, second, fourth, and sixth laws stated above, with the assumption that any earthquake can trigger subsequent earthquakes.

Entropy Method for the Distribution of Time Intervals Between Mainshocks

Mega et al. [140] used the "diffusion entropy" method to argue for a power-law distribution of time intervals between a large earthquake (the mainshock of a seismic sequence or cluster) and the next one. Helmstetter and Sornette [80] showed that all the "new" discoveries reported by [140] (including the supposedly new scaling) can be explained solely by Omori's law for intra-cluster times, without correlation between clusters, thus debunking the claim for novelty.

Scaling of the PDF of Waiting Times

Bak et al. [14] analyzed the scaling of the probability density function of waiting times between successive earthquakes in southern California as a function of "box size" or small regions in which subsequent earthquakes are considered. They found an approximate collapse of the pdfs for different seismic moment thresholds S and box sizes L which suggested the following scaling ansatz for the waiting times T:

$$T^{\alpha} P_{S,L}(T) = f(TS^{-b}L^{d_f}) , \tag{3}$$

where $b = 1$ is the Gutenberg–Richter exponent, $d_f \simeq 1.2$ was claimed to be a spatial fractal dimension of seismicity (see [146] and [115] for more in-depth studies), $\alpha = 1$ was identified as the exponent in the Omori law and $f(\cdot)$ is a scaling function which was proposed to be roughly constant up to a constant ("kink") beyond which it quickly decays. The scaling (3) was claimed to be a unified law for earthquakes that revealed a novel feature in the spatio-temporal organization of seismicity in that the Gutenberg–Richter, the Omori law and the spatial distribution of earthquakes were unified into a single picture that made no distinction between fore-, main- and aftershocks. The scaling relations and critical exponents were claimed to be contained in the scaling ansatz. Corral [40,41,42,43] and others broadened the analysis to other regions of the world. Corral [41] proposed a slightly different scaling ansatz for a modified data analysis.

Early criticism came from Lindman et al. [132], who noted that synthetic data generated using a non-homogeneous Poisson process derived from Omori's law was able to reproduce some of the results of [14], indicating a rather trivial origin of the unified scaling law. Molchan [145] showed that, if at least two regions in the data set are independent, then, if a scaling relation were to hold exactly, this scaling function could only be exponential. All other functions could only result in approximate data collapses.

Proponents of the unified scaling law, e. g. [44], argued that indeed all regions were correlated, as expected in systems near a critical point so that the assumption of independence between different regions should not hold. But Molchan [145] also showed that a simple Poisson cluster model (Poissonian mainshocks that trigger Omori-type aftershock sequences) could reproduce the short and long time limits of the observed statistics, indicating that the Omori law, the Gutenberg–Richter relationship and simple clustering were the sole ingredients necessary for the observed short and long time limit, and no spatial correlation was needed.

Saichev and Sornette [176,177] extended Molchan's arguments to show that the approximate data collapse of the waiting times could be explained completely by the Epidemic-Type Aftershock Sequence (ETAS) model of [149]. This provided further evidence that the apparent data collapse was only approximate. Remarkably, the theoretical predictions of the ETAS model seem to fit the observed data better than the phenomenological scaling function proposed by [41] to fit the data. Saichev and Sornette [176,177] thus showed that a benchmark model of seismicity was able to reproduce the apparent unified scaling law and that therefore the distribution of interevent times did not reveal new information beyond what was already known via statistical laws: The combination of the Gutenberg–Richter law, the Omori law, and the concept of clustering suffice to explain the apparent "universal" scaling of the waiting times.

Sornette et al. [215] developed an efficient numerical scheme to solve accurately the set of nonlinear integral equations derived previously in [177] and found a dramatic lack of power for the distribution of inter-event times to distinguish between quite different sets of parameters, casting doubt on the usefulness of this statistics for the specific purpose of identifying the clustering parameter (e. g. [72]).

Scaling of the PDF of Distances Between Subsequent Earthquakes

Davidson and Paczuski [49] claimed evidence contradicting the theory of aftershock zone scaling in favor of scale-free statistics. Aftershock zone scaling refers to the scaling of the mainshock rupture length, along which most aftershocks occur, with the mainshock magnitude [112]. Davidson and Paczuski [49] suggested that the probability density function of spatial distances between successive earthquakes obeys finite size scaling with a novel dynamical scaling exponent, suggesting that the mainshock rupture length scale has no impact on the spatial distribution of aftershocks and that earthquakes are self-organized critical.

Werner and Sonette [230] debunked this claim by showing that (i) the purported power law scaling function is not universal as it breaks down in other regions of the world; (ii) the results obtained by [49] for southern California depend crucially on a single earthquake (the June 28, 1992, M7.3 Landers earthquake): without Landers and its aftershocks, the power law disappears; (iii) a model of clustered seismicity, with aftershock zone scaling explicitly built in, is able reproduce the apparent power law, indicating that an apparent lack of scales in the data does not necessarily contradict aftershock zone scaling and the existence of scales associated with mainshock rupture length scales.

The Network Approach

The recent boom in the statistical mechanics of network analysis has recently extended to applications well beyond physics (for reviews, see [6,16,57,148]). Earthquake seismology is no exception [1,2,3,4,8,9,10,160,161]. The resulting impact has been limited so far for several reasons.

A major concern is the assumption that earthquake catalogs as downloaded from the web are data sets fit for immediate analysis. References [82,113,229] and [231] present modern and complementary assessments of the many issues of incompleteness spoiling even the best catalogs. In particular, we should stress that magnitude determinations are surprisingly inaccurate, leading to large errors in seismic rate estimates [231]. Furthermore, there is no such thing as a complete catalog above a so-called magnitude of completeness, due to the fact that a non-negligible fraction of earthquakes are missed in the aftermath of previous earthquakes [82,113]. One should be concerned that analyses in terms of network metrics could be particularly sensitive to these defects. Nevertheless, Abe and Suzuki [1,2,3,4] applied metrics of network analysis to "raw" catalogs which included events well below the estimated magnitude of completeness. As a result of neglecting to use a (reasonably) homogeneous and trustworthy data set, the results of their analysis may be severely biased, because the reliability of the inferred network structure is probably more sensitive than other metrics to the correct spatio-temporal ordering of the earthquake catalog. No serious study has yet been performed to quantify the usually serious impact of quality issues on the metrics used in network analysis. As a consequence, it is also not clear how to interpret the "success" of [160,161] in reproducing the "features" of Abe and Suzuki's analysis on the synthetic seismicity generated by a spring-block model.

In addition, at best limited attempts have been made to interpret the results of the new network metrics using well-known, established facts in seismology. Many of the claimed novel features are probably very well understood – they are mostly related to scale-invariance and clustering of seismicity, facts documented for decades. The authors should always strive to show that the new metrics that they propose give results that cannot be explained by the standard laws in statistical seismology. Toward this end, there are well-defined benchmark models that incorporate these laws and that can generate synthetic catalogs on which the new metrics can be tested and compared.

A few exceptions are worth mentioning. Motivated by the long-standing and unresolved debate over "aftershock" identification, Baiesi and Paczuski [9,10] and Baiesi [8] provided a new metric for the correlations between earthquakes based on the space-time-magnitude nearest-neighbor distance between earthquakes. The authors compared their results with known statistical laws in seismology and with the predictions of the ETAS model, actually confirming both. While no new law has been unearthed here, such efforts are valuable to validate known laws and continue to test the possible limits. Zaliapin et al. [238] extended their study and investigated the theoretical properties of the metric and its ability to decluster catalogs (i. e., separate mainshocks from aftershocks). They concluded that aftershocks defined from this metric seem to be different from the rest of earthquakes. It will be interesting to see head-to-head comparisons with current state-of-the-art probabilistic declustering techniques that are based on empirical statistical laws and likelihood estimation [105,240,241].

Future Directions

The study of the statistical physics of earthquakes remains wide-open with many significant discoveries to be made. The promise of a holistic approach – one that emphasizes the interactions between earthquakes and faults – is to be able to neglect some of the exceedingly complicated micro-physics when attempting to understand the large scale patterns of seismicity. The marriage between this conceptual approach, based on the successes of statistical physics, and seismology thus remains a highly important domain of research. In particular, statistical seismology needs to evolve into a genuine physically-based statistical physics of earthquakes.

The question of renormalizability of models of earthquake occurrence and the role of small earthquakes in the organization of seismicity is likely to remain an important topic. It connects with the problem of foreshocks and the predictability of large events from small ones and therefore has real and immediate practical applications as well as physical implications.

More detailed and rigorous empirical studies of the frequency-size statistics of earthquake seismic moments and how they relate to seismo-tectonic conditions are needed in order to help settle the controversy over the power-law versus the characteristic event regime, and the role of regime-switching and universality.

Spatially extended, dynamically evolving fault networks and their role in the generation of earthquakes are mostly ignored in the statistical physics approach to seismicity. Akin to the filaments in turbulence, these may provide key insights into the spatio-temporal organization of earthquakes. Novel methods combining information from seismology to faulting will be required (e. g., [71,159,195,196,197]) to build a real understanding of the self-organization of the chicken-and-egg structures that earthquakes-faults constitute. Furthermore, a true physical approach requires understanding the spatio-temporal evolution of stresses, their role in earthquake nucleation via thermally activated processes, in the rupture propagation and in the physics of arrest, both involved in the generation of complex stress fields.

The important debate regarding statistical physics approaches to seismicity would benefit significantly from two points. Firstly, earthquake catalogs contain data uncertainties, biases and subtle incompleteness issues. Investigating their influence on the results of data analyses inspired by statistical physics increases the relevance of the results. Secondly, the authors should make links with the literature on statistical seismology which deals with similar questions. It is their task to show that the new metrics that they propose give results that cannot be explained by the standard laws in statistical seismology. For this, there are well-defined benchmark models that incorporate these laws and that can generate synthetic catalogs on which the new metrics can be tested.

Bibliography

1. Abe S, Suzuki N (2004) Scale-free network of earthquakes. Europhys Lett 65:581–586. doi:10.1209/epl/i2003-10108-1
2. Abe S, Suzuki N (2004) Small-world structure of earthquake network. Physica A: Stat Mech Appl 337:357–362. doi:10.1016/j.physa.2004.01.059
3. Abe S, Suzuki N (2005) Scale-invariant statistics of period in directed earthquake network. Eur Phys J B 44:115–117. doi:10.1140/epjb/e2005-00106-7
4. Abe S, Suzuki N (2006) Complex earthquake networks: Hierarchical organization and assortative mixing. Phys Rev E 74(2):026, 113–+. doi:10.1103/PhysRevE.74.026113

5. Aki K (1995) Earthquake prediction, societal implications. Rev Geophys 33:243–248
6. Albert R, Barabási AL (2002) Statistical mechanics of complex networks, Rev Mod Phys 74(1):47–97. doi:10.1103/RevModPhys.74.47
7. Allègre CJ, Le Mouel JL, Provost A (1982) Scaling rules in rock fracture and possible implications for earthquake prediction. Nature 297:47–49. doi:10.1038/297047a0
8. Baiesi M (2006) Scaling and precursor motifs in earthquake networks. Physica A: Stat Mech Appl 359:775–783. doi:10.1016/j.physa.2005.05.094
9. Baiesi M, Paczuski M (2004) Scale-free networks of earthquakes and aftershocks. Phys Rev E 69(6):066, 106. doi:10.1103/PhysRevE.69.066106
10. Baiesi M, Paczuski M (2005) Complex networks of earthquakes and aftershocks. Nonlin Proc Geophys 12:1–11
11. Bak P (1996) How Nature Works: The Science of Self-Organized Criticality. Springer, New York, p 212
12. Bak P, Tang C (1989) Earthquakes as a self-organized critical phenomena. J Geophys Res 94(B11):15635–15637
13. Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality: An explanation of the 1/*f* noise. Phys Rev Lett 59(4):381–384. doi:10.1103/PhysRevLett.59.381
14. Bak P, Christensen K, Danon L, Scanlon T (2002) Unified scaling law for earthquakes. Phys Rev Lett 88(17):178,501. doi:10.1103/PhysRevLett.88.178501
15. Bakun, WH, Aagaard B, Dost B, Ellsworth WL, Hardebeck JL, Harris RA, Ji C, Johnston MJS, Langbein J, Lienkaemper JJ, Michael AJ, Murray JR, Nadeau RM, Reasenberg PA, Reichle MS, Roeloffs EA, Shakal A, Simpson RW, Waldhauser F (2005) Implications for prediction and hazard assessment from the 2004 Parkfield earthquake. Nature 437:969–974. doi:10.1038/nature04067
16. Barabási AL, Albert R (1999) Emergence of Scaling in Random Networks. Science 286(5439):509–512. doi:10.1126/science.286.5439.509
17. Barabási AL, Albert R, Jeong H (1999) Mean-field theory fore scale-free random networks. Physica A 272:173–187. doi:10.1016/S0378-4371(99)00291-5
18. Barton CC, La Pointe PR (eds) (1995) Fractals in the Earth Sciences. Plenum Press, New York, London
19. Barton CC, La Pointe PR (eds) (1995) Fractals in petroleum geology and earth processes. Plenum Press, New York, London
20. Båth M (1965) Lateral inhomogeneities in the upper mantle. Tectonophysics 2:483–514
21. Ben-Zion Y, Rice JR (1993) Earthquake failure sequences along a cellular fault zone in a 3-dimensional elastic solid containing asperity and nonasperity regions. J Geophys Res 93:14109–14131
22. Ben-Zion Y, Rice JR (1995) Slip patterns and earthquake populations along different classes of faults in elastic solids. J Geophys Res 100:12959–12983
23. Ben-Zion Y, Rice JR (1997) Dynamic simulations of slip on a smooth fault in an elastic solid. J Geophys Res 102:17771–17784
24. Ben-Zion Y, Dahmen K, Lyakhovsky V, Ertas D, Agnon A (1999) Self-driven mode switching of earthquake activity on a fault system. Earth Planet Sci Lett 172:11–21
25. Bird P, Kagan YY (2004) Plate-tectonic analysis of shallow seismicity: Apparent boundary width, beta, corner magnitude, coupled lithosphere thickness, and coupling in seven tectonic settings. Bull Seismol Soc Am 94(6):2380–2399
26. Bowman DD, King GCP (2001) Stress transfer and seismicity changes before large earthquakes. C Royal Acad Sci Paris, Sci Terre Planetes 333:591–599
27. Bowman DD, King GCP (2001) Accelerating seismicity and stress accumulation before large earthquakes. Geophys Res Lett 28:4039–4042
28. Bowman DD, Oullion G, Sammis CG, Sornette A, Sornette D (1998) An observational test of the critical earthquake concept. J Geophys Res 103:24359–24372
29. Brehm DJ, Braile LW (1998) Intermediate-term earthquake prediction using precursory events in the New Madrid Seismic Zone. Bull Seismol Am Soc 88(2):564–580
30. Bufe CG, Varnes DJ (1993) Predictive modeling of the seismic cycle of the greater San Francisco Bay region. J Geophys Res 98:9871–9883
31. Burridge R, Knopoff L (1964) Body force equivalents for seismic dislocation. Seism Soc Am Bull 54:1875–1888
32. Cardy JL (1996) Scaling and Renormalization in Statistical Physics. Cambridge University Press, Cambridge
33. Carlson JM, Langer JS (1989) Properties of earthquakes generated by fault dynamics. Phys Rev Lett 62:2632–2635
34. Carlson JM, Langer JS, Shaw BE (1994) Dynamics of earthquake faults. Rev Mod Phys 66:657–670
35. Chelidze TL (1982) Percolation and fracture. Phys Earth Planet Interiors 28:93–101
36. Christensen K, Farid N, Pruessner G, Stapleton M (2008) On the finite-size scaling of probability density functions. Eur Phys B 62:331–336
37. Clauset A, Shalizi CR, Newman MEJ (2007) Power-law distributions in empirical data. E-print arXiv:0706.1062
38. Cochard A, Madariaga R (1994) Dynamic faulting under rate-dependent friction. Pure Appl Geophys 142:419–445
39. Cochard A, Madariaga R (1996) Complexity of seismicity due to highly rate-dependent friction. J Geophys Res 101:25321–25336
40. Corral A (2003) Local distributions and rate fluctuations in a unified scaling law for earthquakes. Phys Rev E 68(3):035, 102. doi:10.1103/PhysRevE.68.035102
41. Corral A (2004) Universal local versus unified global scaling laws in the statistics of seismicity. Physica A 340:590–597
42. Corral A (2004) Long-term clustering, scaling, and universality in the temporal occurrence of earthquakes. Phys Rev Lett 92:108, 501
43. Corral A (2005) Mixing of rescaled data and bayesian inference for earthquake recurrence times. Nonlin Proc Geophys 12:89–100
44. Corral A (2005) Renormalization-group transformations and correlations of seismicity. Phys Rev Lett 95:028, 501
45. Corral A (2006) Universal earthquake-occurrence jumps, correlations with time, and anomalous diffusion. Phys Rev Lett 97:178, 501
46. Corral A, Christensen K (2006) Comment on "earthquakes descaled: On waiting time distributions and scaling laws". Phys Rev Lett 96:109, 801
47. Dahmen K, Ertaş D, Ben-Zion Y (1998) Gutenberg–Richter and characteristic earthquake behavior in simple mean-field models of heterogeneous faults. Phys Rev E 58:1494–1501. doi:10.1103/PhysRevE.58.1494

48. Davidsen J, Goltz C (2004) Are seismic waiting time distributions universal? Geophys Res Lett 31:L21612. doi:10.1029/2004GL020892
49. Davidsen J, Paczuski M (2005) Analysis of the spatial distribution between successive earthquakes. Phys Rev Lett 94:048, 501. doi:10.1103/PhysRevLett.94.048501
50. Davidsen J, Grassberger P, Paczuski M (2006) Earthquake recurrence as a record breaking process. Geophys Res Lett 33:L11304. doi:10.1029/2006GL026122
51. Dhar D (1990) Self-organized critical state of sandpile automaton models. Phys Rev Lett 64:1613–1616
52. Dhar D (1999) The Abelian sandpile and related models. Physica A 263:4–25
53. Dieterich JH (1987) Nucleation and triggering of earthquake slip; effect of periodic stresses. Tectonophysics 144:127–139
54. Dieterich JH (1992) Earthquake nucleation on faults with rate-dependent and state-dependent strength. Tectonophysics 211:115–134
55. Dieterich J (1994) A constitutive law for rate of earthquake production and its application to earthquake clustering. J Geophys Res 99:2601–2618
56. Dieterich J, Kilgore BD (1994) Direct observation of frictional constacts- New insight for state-dependent properties. Pure Appl Geophys 143:283–302
57. Dorogevtsev SN, Mendes JFF (2003) Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford University Press, New York
58. Eckman JP (1981) Roads to Turbulence in Dissipative Dynamical Systems. Rev Mod Phys 53:643–654
59. Ellsworth WL, Lindh AG, Prescott WH, Herd DJ (1981) The 1906 San Francisco Earthquake and the seismic cycle. Am Geophys Union Maurice Ewing Monogr 4:126–140
60. Felzer KR, Becker TW, Abercrombie RE, Ekstrom G, Rice JR (2002) Triggering of the 1999 Mw 7.1 Hector Mine earthquake by aftershocks of the 1992 Mw 7.3 Landers earthquake. J Geophys Res 107(B09):2190
61. Fisher DS, Dahmen K, Ramanathan S, Ben-Zion Y (1997) Statistics of Earthquakes in Simple Models of Heterogeneous Faults. Phys Rev Lett 78:4885–4888. doi:10.1103/PhysRevLett.78.4885
62. Freund F, Sornette D (2007) Electro-Magnetic Earthquake Bursts and Critical Rupture of Peroxy Bond Networks in Rocks. Tectonophysics 431:33–47
63. Frisch U (1995) Turbulence. The legacy of A.N. Kolmogorov. Cambridge University Press, Cambridge
64. Gabrielov A, Keilis-Borok V, Jackson DD (1996) Geometric Incompatibility in a Fault System. Proc Nat Acad Sci 93:3838–3842
65. Gabrielov A, Keilis-Borok V, Zaliapin I, Newman W (2000) Critical transitions in colliding cascades. Phys Rev E 62:237–249
66. Gabrielov A, Zaliapin I, Newman W, Keilis-Borok V, (2000) Colliding cascades model for earthquake prediction. Geophys J Int 143:427–437
67. Gallagher R, Appenzeller T (1999) Beyond Reductionism. Science 284(5411):79
68. Geilikman MB, Pisarenko VF, Golubeva TV (1990) Multifractal Patterns of Seismicity. Earth Planet Sci Lett 99:127–138
69. Gelfand IM, Guberman SA, Keilis-Borok VI, Knopoff L, Press F, Ranzman EY, Rotwain IM, Sadovsky AM (1976) Pattern recognition applied to earthquake epicenters in California. Phys Earth Planet Interiors 11:227–283
70. Geller RJ, Jackson DD, Kagan YY, Mulargia F (1997) Earthquakes cannot be predicted. Science 275:1616–1617
71. Gorshkov A, Kossobokov V, Soloviev A (2003) Recognition of earthquake-prone areas. In: Keilis-Borok V, Soloviev A (eds) Nonlinear Dynamics of the Lithosphere and Earthquake Prediction. Springer, Heidelberg, pp 239–310 [122]
72. Hainzl S, Scherbaum F, Beauval C (2006) Estimating Background Activity Based on Interevent-Time Distribution. Bull Seismol Soc Am 96(1):313–320. doi:10.1785/0120050053
73. Hanks TC (1992) Small earthquakes, tectonic forces. Science 256:1430–1432
74. Hardebeck JL, Felzer KR, Michael AJ (2008) Improved tests reveal that the accelerating moment release hypothesis is statistically insignificant. J Geophys Res 113:B08310. doi:10.1029/2007JB005410
75. Harris RA, Arrowsmith JR (2006) Introduction to the Special Issue on the 2004 Parkfield Earthquake and the Parkfield Earthquake Prediction Experiment. Bull Seismol Soc Am 96(4B):S1–10. doi:10.1785/0120050831
76. Helmstetter A (2003) Is earthquake triggering driven by small earthquakes? Phys Rev Lett 91(5):058, 501. doi:10.1103/PhysRevLett.91.058501
77. Helmstetter A, Sornette D (2002) Subcritical and supercritical regimes in epidemic models of earthquake aftershocks. J Geophys Res 107(B10):2237. doi:10.1029/2001JB001580
78. Helmstetter A, Sornette D (2003) Foreshocks explained by cascades of triggered seismicity. J Geophys Res (Solid Earth) 108(B10):2457 doi:10.1029/2003JB00240901
79. Helmstetter A, Sornette D (2003) Bath's law Derived from the Gutenberg–Richter law and from Aftershock Properties. Geophys Res Lett 30:2069. doi:10.1029/2003GL018186
80. Helmstetter A, Sornette D (2004) Comment on "Power-Law Time Distribution of Large Earthquakes". Phys Rev Lett 92:129801 (Reply is Phys Rev Lett 92:129802 (2004))
81. Helmstetter A, Sornette D, Grasso J-R (2003) Mainshocks are Aftershocks of Conditional Foreshocks: How do foreshock statistical properties emerge from aftershock laws. J Geophys Res 108(B10):2046. doi:10.1029/2002JB001991
82. Helmstetter A, Kagan YY, Jackson DD (2005) Importance of small earthquakes for stress transfers and earthquake triggering. J Geophys Res 110:B05508. doi:10.1029/2004JB003286
83. Helmstetter A, Kagan Y, Jackson D (2006) Comparison of short-term and long-term earthquake forecast models for Southern California. Bull Seism Soc Am 96:90–106
84. Hergarten S (2002) Self-Organized Criticality in Earth Systems. Springer, Berlin
85. Hinrichsen H (2000) Non-equilibrium critical phenomena and phase transitions into absorbing states. Adv Phys 49:815–958(144)
86. Holliday JR, Rundle JB, Tiampo KF, Klein W, Donnellan A (2006) Systematic procedural and sensitivity analysis of the Pattern Informatics method for forecasting large ($M > 5$) earthquake events in Southern California. Pure Appl Geophys 163(11–12):2433–2454
87. Huang J, Turcotte DL (1990) Evidence for chaotic fault interactions in the seismicity of the San Andreas fault and Nankai trough. Nature 348:234–236
88. Huang J, Turcotte DL (1990) Are earthquakes an example of deterministic chaos? Geophys Rev Lett 17:223–226

89. Huang Y, Saleur H, Sammis CG, Sornette D (1998) Precursors, aftershocks, criticality and self-organized criticality. Europhys Lett 41:43–48
90. Ide K, Sornette D (2002) Oscillatory Finite-Time Singularities in Finance, Population and Rupture. Physica A307(1–2):63–106
91. Jackson DD, Kagan YY (2006) The 2004 Parkfield Earthquake, the 1985 Prediction, and Characteristic Earthquakes: Lessons for the Future. Bull Seismol Soc Am 96(4B):S397–409. doi:10.1785/0120050821
92. Jaumé SC, Sykes LR (1999) Evolving Towards a Critical Point: A Review of Accelerating Seismic Moment/Energy Release Prior to Large and Great Earthquakes. Pure Appl Geophys 155:279–305
93. Jensen HJ (1998) Self-Organized Criticality: Emergent Complex Behavior in Physical and Biological Systems. Cambridge University Press, Cambridge
94. Johansen A, Sornette D, Wakita G, Tsunogai U, Newman WI, Saleur H (1996) Discrete scaling in earthquake precursory phenomena: evidence in the Kobe earthquake, Japan J Phys I France 6:1391–1402
95. Johansen A, Saleur H, Sornette D (2000) New Evidence of Earthquake Precursory Phenomena in the 17 Jan. 1995 Kobe Earthquake, Japan. Eur Phys J B 15:551–555
96. Jones LM (1994) Foreshocks, aftershocks, and earthquake probabilities: accounting for the Landers earthquake. Bull Seismol Soc Am 84:892–899
97. Jordan TH (2006) Earthquake Predictability, Brick by Brick. Seismol Res Lett 77(1):3–6
98. Kadanoff LP, Nagel SR, Wu L, Zhou S-M (1989) Scaling and universality in avalanches. Phys Rev A 39(12):6524–6537. doi:10.1103/PhysRevA.39.6524
99. Kagan YY (1981), Spatial distribution of earthquakes: The three-point moment function. Geophys J R Astron Soc 67:697–717
100. Kagan YY (1981) Spatial distribution of earthquakes: The four-point moment function. Geophys J Roy Astron Soc 67:719–733
101. Kagan YY (1987) Point sources of elastic deformation: Elementary sources, static displacements. Geophys J R Astron Soc 90:1–34
102. Kagan YY (1987) Point sources of elastic deformation: Elementary sources, dynamic displacements. Geophys J R Astron Soc 91:891–912
103. Kagan YY (1988) Multipole expansions of extended sources of elastic deformation. Geophys J R Astron Soc 93:101–114
104. Kagan YY (1989) Earthquakes and fractals. Ann Rev Mater Sci: Fractal Phenom Disordered Syst 19:520–522
105. Kagan YY (1991) Likelihood analysis of earthquake catalogs. Geophys J Int 106:135–148
106. Kagan YY (1992) Seismicity: Turbulence of solids. Nonlinear Sci Today 2:1–13
107. Kagan YY (1992) On the geometry of an earthquake fault system. Phys Earth Planet Interiors 71:15–35
108. Kagan YY (1993) Statistics of characteristic earthquakes. Bull Seismol Soc Am 83(1):7–24
109. Kagan YY (1994) Observational evidence for earthquakes as a nonlinear dynamic process. Physica D 77:160–192
110. Kagan YY (1994) Comment on "The Gutenberg–Richter or char-acteristic earthquake distribution, which is it?" by Wesnousky. Bull Seismol Soc Am 86:274–285
111. Kagan YY (1999) Is earthquake seismology a hard, quantitative science? Pure Appl Geophys 155:33–258
112. Kagan YY (2002) Aftershock Zone Scaling. Bull Seismol Soc Am 92(2):641–655. doi:10.1785/0120010172
113. Kagan YY (2003) Accuracy of modern global earthquake catalogs. Phys Earth Planet Interiors 135:173–209
114. Kagan YY (2006) Why does theoretical physics fail to explain and predict earthquake occurrence? In: Bhattacharyya P, Chakrabarti BK (eds) Modelling Critical and Catastrophic Phenomena in Geoscience: A Statistical Physics Approach. Lecture Notes in Physics, vol 705. Springer, Berlin, pp 303–359
115. Kagan YY (2007) Earthquake spatial distribution: the correlation dimension. Geophys J Int 168:1175–1194. doi:10.1111/j.1365-246X.2006.03251.x
116. Kagan YY, Knopoff L (1980) Spatial distribution of earthquakes: The two-point correlation function. Geophys J R Astron Soc 62:303–320
117. Kagan YY, Knopoff L (1981) Stochastic synthesis of earthquake catalogs. J Geophys Res 86(B4):2853–2862
118. Kagan YY, Knopoff L (1985) The first-order statistical moment of the seismic moment tensor. Geophys J R Astron Soc 81:429–444
119. Kagan YY, Knopoff L (1985) The two-point correlation function of the seismic moment tensor. Geophys J R Astron Soc 83:637–656
120. Keilis-Borok VI (ed) (1990) Intermediate-term earthquake prediction: models, algorithms, worldwide tests. Phys Earth Planet Interiors 61(1–2)
121. Keilis-Borok VI, Malinovskaya LN (1964) One regularity in the occurrence of strong earthquakes. J Geophys Res B 69:3019–3024
122. Keilis-Borok V, Soloviev A (2003) Nonlinear Dynamics of the Lithosphere and Earthquake Prediction. Springer, Heidelberg
123. Keilis-Borok VI, Knopoff L, Rotwain IM, Allen CR (1988) Intermediate-term prediction of occurrence times of strong earthquakes. Nature 335:690–694
124. King GCP, Bowman DD (2003) The evolution of regional seismicity between large earthquakes. J Geophys Res 108(B2):2096. doi:10.1029/2001JB000783
125. Klein W, Rundle JB, Ferguson CD (1997) Scaling and nucleation in models of earthquake faults. Phys Rev Lett 78:3793–3796
126. Knopoff L (1996) The organization of seismicity on fault networks. Proc Nat Acad Sci USA 93:3830–3837
127. Landau LD, Lifshitz EM (1980) Statistical Physics Course on Theoretical Physics, vol 5, 3rd edn. Butterworth-Heinemann, Oxford
128. Langer JS, Carlson JM, Myers CR, Shaw BE (1996) Slip complexity in dynamical models of earthquake faults. Proc Nat Acad Sci USA 93:3825–3829
129. Lee MW, Sornette D, Knopoff L (1999) Persistence and Quiescence of Seismicity on Fault Systems. Phys Rev Lett 83(N20):4219–4222
130. Levin SZ, Sammis CG, Bowman DD (2006) An observational test of the stress accumulation model based on seismicity preceding the 1992 Landers, CA earthquake. Tectonophysics 413:39–52
131. Lindh AG (1990) The seismic cycle pursued. Nature 348:580–581

132. Lindman M, Jonsdottir K, Roberts R, Lund B, Bdvarsson R (2005) Earthquakes descaled: On waiting time distributions and scaling laws. Phys Rev Lett 94:108, 501
133. Lindman M, Jonsdottir K, Roberts R, Lund B, Bdvarsson R (2006) Reply to comment by A. Corral and K. Christensen. Phys Rev Lett 96:109, 802
134. Livina VN, Havlin S, Bunde A (2006) Memory in the occurrence of earthquakes. Phys Rev Lett 95:208, 501
135. Luebeck S (2004) Universal scaling behavior of non-equilibrium phase transitions. Int J Mod Phys B 18:3977
136. Manna S (1991) Critical exponents of the sandpile models in two dimensions. Physica A179(2):249–268
137. Mandelbrot BB (1982) The Fractal Geometry of Nature. W.H. Freeman, San Francisco
138. Marsan D (2005) The role of small earthquakes in redistributing crustal elastic stress. Geophys J Int 163(1):141–151. doi:10.1111/j.1365-246X.2005.02700.x
139. May RM (1976) Simple mathematical models with very complicated dynamics. Nature 261:459–467
140. Mega MS, Allegrini P, Grigolini P, Latora V, Palatella L, Rapisarda A, Vinciguerra S (2003) Power law time distributions of large earthquakes. Phys Rev Lett 90:18850
141. Michael AJ, Jones LM (1998) Seismicity alert probabilities at Parkfield, California, revisited. Bull Seismol Soc Am 88(1):117–130
142. Miltenberger P, Sornette D, Vanneste C (1993) Fault self-organization as optimal random paths selected by critical spatiotemporal dynamics of earthquakes. Phys Rev Lett 71:3604–3607. doi:10.1103/PhysRevLett.71.3604
143. Mitzenmacher M (2004) A Brief History of Generative Models for Power Law and Lognormal Distributions. Internet Math 1(2):226–251
144. Mogi K (1969) Some features of recent seismic activity in and near Japan 2: activity before and after great earthquakes. Bull Eq Res Inst Tokyo Univ 47:395–417
145. Molchan G (2005) Interevent time distribution in seismicity: A theoretical approach. Pure Appl Geophys 162:1135–1150. doi:10.1007/s00024-004-2664-5
146. Molchan G, Kronrod T (2005) On the spatial scaling of seismicity rate. Geophys J Int 162(3):899–909. doi:10.1111/j.1365-246X.2005.02693.x
147. Nature Debates (1999) Nature debates: Is the reliable prediction of individual earthquakes a realistic scientific goal? available from http://www.nature.com/nature/debates/earthquake/equake_frameset.html
148. Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256. doi:10.1137/S003614450342480
149. Ogata Y (1988) Statistical models for earthquake occurrence and residual analysis for point processes. J Am Stat Assoc 83:9–27
150. Ogata Y (1998) Space-time point-process models for earthquake occurrences. Ann Inst Stat Math 5(2):379–402
151. Olami Z, Feder HJS, Christensen K (1992) Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes. Phys Rev Lett 68(8):1244–1247
152. Osorio I, Frei MG, Sornette D, Milton J, Lai Y-C (2007) Seizures and earthquakes: Universality and scaling of critical far from equilibrium systems. submitted to Phys Rev Lett. http://arxiv.org/abs/0712.3929
153. Ouillon G, Sornette D (2000) The critical earthquake concept applied to mine rockbursts with time-to-failure analysis. Geophys J Int 143:454–468
154. Ouillon G, Sornette D (2004) Search for Direct Stress Correlation Signatures of the Critical Earthquake Model. Geophys J Int 157:1233–1246
155. Ouillon G, Sornette D (2005) Magnitude-Dependent Omori Law: Theory and Empirical Study. J Geophys Res 110:B04306. doi:10.1029/2004JB003311
156. Ouillon G, Sornette D, Castaing C (1995) Organization of joints and faults from 1 cm to 100 km scales revealed by Optimized Anisotropic Wavelet Coefficient Method and Multifractal analysis. Nonlinear Process Geophys 2:158–177
157. Ouillon G, Castaing C, Sornette D (1996) Hierarchical scaling of faulting. J Geophys Res 101(B3):5477–5487
158. Ouillon G, Ribeiro E, Sornette D (2007) Multifractal Omori Law for Earthquake Triggering: New Tests on the California, Japan and Worldwide Catalogs. submitted to Geophys J Int. http://arxiv.org/abs/physics/0609179
159. Ouillon G, Ducorbier C, Sornette D (2008) Automatic reconstruction of fault networks from seismicity catalogs: Three-dimensional optimal anisotropic dynamic clustering. J Geophys Res 113:B01306. doi:10.1029/2007JB005032
160. Peixoto TP, Prado CP (2004) Distribution of epicenters in the Olami–Feder–Christensen model. Phys Rev E 69(2):025101. doi:10.1103/PhysRevE.69.025101
161. Peixoto TP, Prado CPC (2006) Network of epicenters of the Olami–Feder–Christensen model of earthquakes. Phys Rev E 74(1):016, 126 doi:10.1103/PhysRevE.74.016126
162. Peters O, Neelin JD (2006) Critical phenomena in atmospheric precipitation. Nature Phys 2:393–396. doi:10.1038/nphys314
163. Pruessner G (2004) Studies in self-organized criticality, Ph D thesis, Imperial College London, available from http://www.ma.imperial.ac.uk/%7Epruess/publications/thesis_final/
164. Raleigh CB, Sieh K, Sykes LR, Anderson DL (1982) Forecasting Southern California Earthquakes. Science 217:1097–1104
165. Reynolds PJ, Klein W, Stanley HE (1977) Renormalization Group for Site and Bond Percolation. J Phys C 10:L167–L172
166. Rhoades DA, Evison FF (2004) Long-range earthquake forecasting with every earthquake a precursor according to scale. Pure Appl Geophys 161:47–72
167. Rhoades DA, Evison FF (2005) Test of the EEPAS forecasting model on the Japan earthquake catalogue. Pure Appl Geophys 162:1271–1290
168. Rice JR (1993) Spatio-temporal complexity of slip on a fault. J Geophys Res 98:9885–9907
169. Rundle JB, Klein W (1993) Scaling and critical phenomena in a cellular automaton slider block model for earthquakes. J Stat Phys 72:405–412
170. Rundle JB, Klein W (1995) New ideas about the physics of earthquakes. Rev Geophys 33:283–286
171. Rundle PB, Rundle JB, Tiampo KF, Sa Martins JS, McGinnis S, Klein W (2001) Nonlinear network dynamics on earthquake fault systems. Phys Rev Lett 87(14):148, 501. doi:10.1103/PhysRevLett.87.148501
172. Rundle JB, Turcotte DL, Shcherbakov R, Klein W, Sammis C (2003) Statistical physics approach to understanding the multiscale dynamics of earthquake fault systems. Rev Geophys 41(4):1019

173. Saichev A, Sornette D (2005) Distribution of the Largest Aftershocks in Branching Models of Triggered Seismicity: Theory of the Universal Bath's law. Phys Rev E 71:056127
174. Saichev A, Sornette D (2005) Vere-Jones' self-similar branching model. Phys Rev E 72:056, 122
175. Saichev A, Sornette D (2006) Renormalization of branching models of triggered seismicity from total to observable seismicity. Eur Phys J B 51:443–459
176. Saichev A, Sornette D (2006) "Universal" distribution of interearthquake times explained. Phys Rev Lett 97:078, 501
177. Saichev A, Sornette D (2007). Theory of earthquake recurrence times. J Geophys Res 112:B04313. doi:10.1029/2006JB004536
178. Saleur H, Sammis CG, Sornette D (1996) Renormalization group theory of earthquakes. Nonlinear Process Geophys 3:102–109
179. Saleur H, Sammis CG, Sornette D (1996) Discrete scale invariance, complex fractal dimensions and log-periodic corrections in earthquakes. J Geophys Res 101:17661–17677
180. Sammis SG, Sornette D (2002) Positive Feedback, Memory and the Predictability of Earthquakes. Proc Nat Acad Sci USA V99:SUPP1:2501–2508
181. Scholz CH (1991) Earthquakes and faulting: Self-organized critical phenomena with a characteristic dimension. In: Riste T, Sherrington D (eds) Spontaneous Formation of Space Time Structure and Criticality. Kluwer, Norwell, pp 41–56
182. Scholz CH (2002) The Mechanics of Earthquakes and Faulting, 2nd edn, Cambridge University Press, Cambridge
183. Scholz CH, Mandelbrot BB (eds) (1989) Fractals in Geophysics. Birkhäuser, Basel
184. Schorlemmer D, Wiemer S, Wyss M (2005) Variations in earthquake-size distribution across different stress regimes. Nature 437:539–542. doi:10.1038/nature04094
185. Schwartz DP, Coppersmith KJ (1984) Fault behavior and characteristic earthquakes: examples from the Wasatch and San Andreas Fault Zones. J Geophys Res 89:5681–5698
186. Shaw BE (1993) Generalized Omori law for aftershocks and foreshocks from a simple dynamics. Geophys Res Lett 20:907–910
187. Shaw BE (1994) Complexity in a spatially uniform continuum fault model. Geophys Res Lett 21:1983–1986
188. Shaw BE (1995) Frictional weakening and slip complexity in earthquake faults. J Geophys Res 102:18239–18251
189. Shaw BE (1997) Model quakes in the two-dimensional wave equation. J Geophys Res 100:27367–27377
190. Shcherbakov R, Turcotte DL (2004) A modified form of Bath's law. Bull Seismol Soc Am 94(5):1968–1975
191. Shnirman MG, Blanter EM (1998) Self-organized criticality in a mixed hierarchical system. Phys Rev Lett 81:5445–5448
192. Smalley RF Jr, Turcotte DL, Solla SA (1985) A renormalization group approach to the stick-slip behavior of faults. J Geophys Res 90:1894–1900
193. Sornette A, Sornette D (1989) Self-organized criticality and earthquakes. Europhys Lett 9:197–202
194. Sornette A, Sornette D (1999) Earthquake rupture as a critical point: Consequences for telluric precursors. Tectonophysics 179:327–334
195. Sornette A, Davy P, Sornette D (1990) Growth of fractal fault patterns. Phys Rev Lett 65:2266–2269
196. Sornette A, Davy P, Sornette D (1990) Fault growth in brittle-ductile experiments and the mechanics of continental collisions. J Geophys Res 98:12111–12139
197. Sornette D (1991) Self-organized criticality in plate tectonics. In: Proceedings of the NATO ASI. vol 349, "Spontaneous formation of space-time structures and criticality" Geilo, Norway 2–12 April 1991. Riste T, Sherrington D (eds) Kluwer, Dordrecht, Boston, pp 57–106
198. Sornette D (1992) Critical phase transitions made self-organized: a dynamical system feedback mechanism for self-organized criticality. J Phys I France 2:2065–2073. doi:10.1051/jp1:1992267
199. Sornette D (1998) Discrete scale invariance and complex dimensions. Phys Rep 297(5):239–270
200. Sornette D (1999) Earthquakes: from chemical alteration to mechanical rupture. Phys Rep 313(5):238–292
201. Sornette D (2000) Mechanochemistry: an hypothesis for shallow earthquakes. In: Teisseyre R, Majewski E (eds) Earthquake Thermodynamics and Phase Transformations in the Earth's Interior. Int Geophys Series, vol 76. Cambridge University Press, Cambridge, pp 329–366, e-print at http://xxx.lanl.gov/abs/cond-mat/9807400
202. Sornette D (2002) Predictability of catastrophic events: material rupture, earthquakes, turbulence, financial crashes and human birth. Proc Nat Acad Sci USA 99:2522–2529
203. Sornette D (2004) Critical Phenomena in Natural Sciences: Chaos, Fractals, Selforganization and Disorder: Concepts and Tools, 2nd edn. Springer, Berlin, p 529
204. Sornette D, Helmstetter A (2002) Occurrence of Finite-Time-Singularity in Epidemic Models of Rupture, Earthquakes and Starquakes. Phys Rev Lett 89(15):158501
205. Sornette D, Ouillon G (2005) Multifractal Scaling of Thermally-Activated Rupture Processes. Phys Rev Lett 94:038501
206. Sornette D, Pisarenko VF (2003) Fractal Plate Tectonics. Geophys Res Lett 30(3):1105. doi:10.1029/2002GL015043
207. Sornette D, Sammis CG (1995) Complex critical exponents from renormalization group theory of earthquakes: Implications for earthquake predictions. J Phys I France 5:607–619
208. Sornette D, Virieux J (1992) A theory linking large time tectonics and short time deformations of the lithosphere. Nature 357:401–403
209. Sornette D, Werner MJ (2005) Constraints on the size of the smallest triggering earthquake from the epidemic-type aftershock sequence model, Båth's law, and observed aftershock sequences. J Geophys Res 110:B08304. doi:10.1029/2004JB003535
210. Sornette D, Werner MJ (2005) Apparent clustering and apparent background earthquakes biased by undetected seismicity. J Geophys Res 110:B09303. doi:10.1029/2005JB003621
211. Sornette D, Davy P, Sornette A (1990) Structuration of the lithosphere in plate tectonics as a self-organized critical phenomenon. J Geophys Res 95:17353–17361
212. Sornette D, Vanneste C, Sornette A (1991) Dispersion of b-values in Gutenberg–Richter law as a consequence of a proposed fractal nature of continental faulting. Geophys Res Lett 18:897–900
213. Sornette D, Miltenberger P, Vanneste C (1994) Statistical physics of fault patterns self-organized by repeated earthquakes. Pure Appl Geophys 142:491–527. doi:10.1007/BF00876052

214. Sornette D, Miltenberger P, Vanneste C (1995) Statistical physics of fault patterns self-organized by repeated earthquakes: synchronization versus self-organized criticality. In: Bouwknegt P, Fendley P, Minahan J, Nemeschansky D, Pilch K, Saleur H, Warner N (eds) Recent Progresses in Statistical Mechanics and Quantum Field Theory. Proceedings of the conference 'Statistical Mechanics and Quantum Field Theory', USC, Los Angeles, May 16–21, 1994. World Scientific, Singapore, pp 313–332
215. Sornette D, Utkin S, Saichev A (2008) Solution of the Nonlinear Theory and Tests of Earthquake Recurrence Times. Phys Rev E 77:066109
216. Stanley HE (1999) Scaling, universality, and renormalization: Three pillars of modern critical phenomena. Rev Mod Phys 71(2):S358–S366. doi:10.1103/RevModPhys.71.S358
217. Sykes LR, Jaumé S (1990) Seismic activity on neighboring faults as a long-term precursor to large earthquakes in the San Francisco Bay Area. Nature 348:595–599
218. Tiampo KF, Rundle JB, Klein W (2006) Stress shadows determined from a phase dynamical measure of historic seismicity. Pure Appl Geophys 163(11–12):2407–2416
219. Turcotte DL (1986) Fractals and fragmentation. J Geophys Res 91:1921–1926
220. Turcotte DL (1997) Fractals and Chaos in Geology and Geophysics, 2nd edn. Cambridge University Press, Cambridge, p 398
221. Turcotte DL, Newman WI, Gabrielov A (2000) A statistical physics approach to earthquakes. In: Rundle JB, Turcotte DL, Klein W (eds) GeoComplexity and the Physics of Earthquake. American Geophysical Union, Washington, pp 83–96
222. Tumarkin AG, Shnirman MG (1992) Computational seismology 25:63–71
223. Vere-Jones D (1977) Statistical theories of crack propagation. Math Geol 9:455–481
224. Vere-Jones D (2005) A class of self-similar random measure. Adv Appl Probab 37(4):908–914
225. Vere-Jones D (2006) The development of statistical seismology: A personal experience. Tectonophysics 413(1–2):5–12
226. Vere-Jones D, Robinson R, Yang W (2001) Remarks on the accelerated moment release model: problems of model formulation, simulation and estimation. Geophys J Int 144:517–531. doi:10.1046/j.1365-246X.2001.01348.x
227. Voight B (1988) A method for prediction of volcanic eruptions. Nature 332:125–130
228. Voight B (1989) A relation to describe rate-dependent material failure. Science 243:200–203
229. Werner MJ (2007) On the fluctuations of seismicity and uncertainties in earthquake catalogs: Implications and methods for hypothesis testing. Ph D thesis, University of California, Los Angeles
230. Werner MJ, Sornette D (2007) Comment on "Analysis of the Spatial Distribution Between Successive Earthquakes" by Davidsen and Paczuski. [Phys Rev Lett 94:048501 (2005)]. Phys Rev Lett 99::179801
231. Werner MJ, Sornette D (2008) Magnitude Uncertainties Impact Seismic Rate Estimates, Forecasts and Predictability Experiments. J Geophys Res 113:B08302. doi:10.1029/2007JB005427
232. Wesnousky SG (1994) The Gutenberg–Richter or characteristic earthquake distribution, which is it? Bull Seismol Soc Am 84(6):1940–1959
233. Wiemer S, Katsumata K (1999) Spatial variability of seismicity parameters in aftershock zones. J Geophys Res 104:13135–13152. doi:10.1029/1999JB900032
234. Wilson K (1979) Problems in physics with many scales of length. Sci Am 241:140–157
235. Yeomans JM (1992) Statistical Mechanics of Phase Transitions. Oxford University Press Inc, New York
236. Zaliapin I, Keilis-Borok V, Ghil M (2003) A Boolean delay equation model of colliding cascades. Part I: Multiple seismic regimes. J Stat Phys 111:815–837
237. Zaliapin I, Keilis-Borok V, Ghil M (2003) A Boolean delay equation model of colliding cascades. Part II: Prediction of critical transitions. J Stat Phys 111:839–861
238. Zaliapin I, Gabrielov A, Keilis-Borok V, Wong H (2008) Clustering analysis of seismicity and aftershock identification. Phys Rev Lett 101:018501. doi:10.1103/PhysRevLett.101.018501
239. Zee A (2003) Quantum Field Theory in a Nutshell. Princeton University Press, Princeton
240. Zhuang J, Ogata Y, Vere-Jones D (2002) Stochastic declustering of space-time earthquake occurrences. J Am Stat Assoc 97:369–380
241. Zhuang J, Ogata Y, Vere-Jones D (2004) Analyzing earthquake clustering features by using stochastic reconstruction. J Geophys Res 109:B05301. doi:10.1029/2003JB002879
242. Zöller G, Hainzl S (2002) A systematic spatiotemporal test of the critical point hypothesis for large earthquakes. Geophys Rev Lett 29:53–1
243. Zöller G, Hainzl S, Kurths J (2001) Observation of growing correlation length as an indicator for critical point behavior prior to large earthquakes. J Geophys Res 106:2167–2176. doi:10.1029/2000JB900379

Seismic Wave Propagation in Media with Complex Geometries, Simulation of

Heiner Igel[1], Martin Käser[1], Marco Stupazzini[2]

[1] Department of Earth and Environmental Sciences, Ludwig-Maximilians-University, Munich, Germany

[2] Department of Structural Engineering, Politecnico di Milano, Milano, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
The Evolution of Numerical Methods and Grids
3D Wave Propagation on Hexahedral Grids: Soil-Structure Interactions
3D Wave Propagation on Tetrahedral Grids: Application to Volcanology
Local Time Stepping: Δt-Adaptation
Discussion and Future Directions
Acknowledgments
Bibliography

Glossary

Numerical methods Processes in nature are often described by partial differential equations. Finding solutions to those equations is at the heart of many studies aiming at the explanation of observed data. Simulations of realistic physical processes requires generally the use of numerical methods – a special branch of applied mathematics – that approximate the partial differential equations and allows solving them on computers. Examples are the finite-difference, finite-element, or finite-volume methods.

Spectral elements The spectral element method is an extension of the finite element method that makes use of specific basis functions describing the solutions inside each element. These basis functions (e. g., Chebyshev or Legendre polynomials) allow the interpolation of functions exactly at certain collocation points. This is often termed spectral accuracy.

Discontinuous Galerkin method The discontinuous Galerkin method is a flavor of the finite-element method that allows discontinuous behavior of the spatial or temporal fields at the element boundaries. The discontinuities – that might be small in the case of continuous physical fields such as seismic waves – then define so-called Riemann problems that can be handled using the concepts from finite-volume techniques. Therefore, the approximate solution is updated via numerical fluxes across the element boundaries.

Parallel algorithms All modern supercomputers make use of parallel architectures. This means that a large number of processors are performing (different) tasks on different data at the same time. Numerical algorithms need to be adapted to these hardware architectures by using specific programming paradigms (e. g., the message passing interface MPI). The computational efficiency of such algorithms strongly depends on the specific parallel nature of the problem to be solved, and the requirement for inter-processor communication.

Grid generation Most numerical methods are based on the calculation of the solutions at a large set of points (grids) that are either static or depend on time (adaptive grids). These grids often need to be adapted to the specific geometrical properties of the objects to be modeled (volcano, reservoir, globe). Grids may be designed to follow domain boundaries and internal surfaces. Before specific numerical solvers are employed the grid points are usually connected to form triangles or rectangles in 2D or hexahedra or tetrahedra in 3D.

Definition of the Subject

Seismology is the science that aims at understanding the Earth's interior and its seismic sources from measurements of vibrations of the solid Earth. The resulting images of the physical properties of internal structures and the spatio-temporal behavior of earthquake rupture processes are prerequisites to understanding the dynamic evolution of our planet and the physics of earthquakes. One of the key ingredients to obtain these images is the calculation of synthetic (or theoretical) seismograms for given earthquake sources and internal structures. These synthetic seismograms can then be compared quantitatively with observations and acceptable models be searched for using the theory of inverse problems. The methodologies to calculate synthetic seismograms have evolved dramatically over the past decades in parallel with the evolution of computational resources and the ever increasing volumes of permanent seismic observations in global and regional seismic networks, volcano monitoring networks, and experimental campaigns. Today it is a tremendous challenge to extract an optimal amount of information from seismograms. The imaging process is still primarily carried out using ray theory or extensions thereof not fully taking into account the complex scattering processes that are occurring in nature.

To model seismic observations in their full complexity we need to be able to simulate wave propagation through 3D structures with constitutive relations that account for anisotropic elasticity, attenuation, porous media as well as complex internal interfaces such as layer boundaries or fault systems. This implies that numerical methods have to be employed that solve the underlying partial differential equations on computational grids. The high-frequency oscillatory nature of seismic wave fields makes this an expensive endeavor as far as computational resources are concerned. As seismic waves are propagating hundreds of wavelengths through scattering media, the required accuracy of the numerical approximations has to be of the highest possible order. Despite the fact that the physics of wave propagation is well understood, only recently computational algorithms are becoming available that allow us to accurately simulate wave propagation on many scales such as reservoirs, volcanoes, sedimentary basins, continents, and whole planets.

In addition to the imaging problem for subsurface structure and earthquake sources, the possibilities for 3D wave simulations have opened a new route to forecasting strong ground motions following large earthquakes in seismically active regions. In the absence of any hope to deterministically predict earthquakes, the calculation

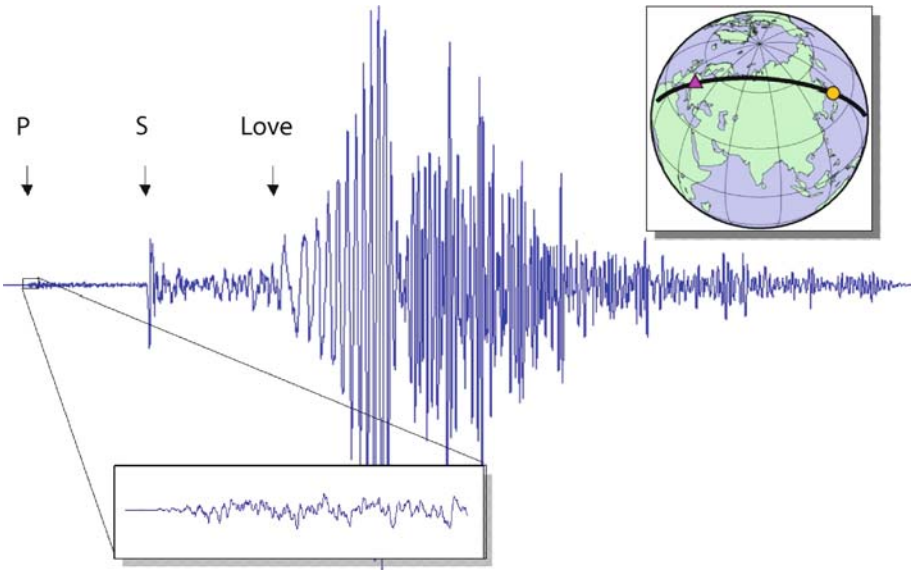

Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 1
Transverse velocity seismogram of the M8.3 Tokachi-Oki earthquake near Hokkaido observed at station WET in Germany with a broadband seismometer. The total seismogram length is one hour. Arrival times of body wave phases (P, S) and the onset of transversely polarized surface (Love) waves are indicated

of earthquake scenarios in regions with sufficiently well known crustal structures and fault locations will play an important role in mitigating damage particularly due to potentially amplifying local velocity structures. However, to be able to employ the advanced 3D simulation technology in an efficient way, and to make use of the fast advance of supercomputing infrastructure, a paradigm shift in the concept of wave simulation software is necessary: The Earth science community has to build soft infrastructures that enable massive use of those simulation tools on the available high-performance computing infrastructure.

In this paper we want to present the state of the art of computational wave propagation and point to necessary developments in the coming years, particularly in connection with finding efficient ways to generate computational grids for models with complex topography, faults, and the combined simulation of soil and structures.

Introduction

We first illustrate the evolution of methodologies to calculate and model aspects of seismic observations for the case of global wave propagation. Seismology can look back at almost 50 years of systematic observations of earthquake induced teleseismic ground motions with the standardized global seismic and regional networks. The digital revolution in the past decades has altered the recording culture such that now seismometers are recording ground motions permanently rather than in trigger-mode, observations are becoming available in near-real time, and – because of the required sampling rates – the daily amount of observations automatically sent to the data centers is gigantic. If we take a qualitative look at a seismic observation (Fig. 1) we can illustrate what it takes to model either part or the whole information contained in such physical measurements.

In Fig. 1 a seismogram observed using a broadband seismometer (station WET in Germany) is shown. Globally observed seismograms following large earthquakes contain frequencies up to 1 Hz (P-wave motions) down to periods of around one hour (eigenmodes of the Earth) in which case modeling is carried out in the frequency domain. Seismograms of the kind shown in Fig. 1 contain many types of information. For large earthquakes the first part of the seismogram (inlet) contains valuable information on the spatio-temporal evolution of the earthquake rupture on a finite-size fault. A model of the fault slip history is a prerequisite to model the complete wave form of seismograms as the whole seismogram is affected by it unless severe low-pass filtering is applied. Information on the global seismic velocity structure is contained in the arrival times of numerous body-wave phases (here only P- and S-wave arrivals are indicated) and in the dispersive behavior of the surface waves (here the onset of the low-frequency Love waves is indicated). Further information is contained in the characteristics of the coda to body wave phases in-

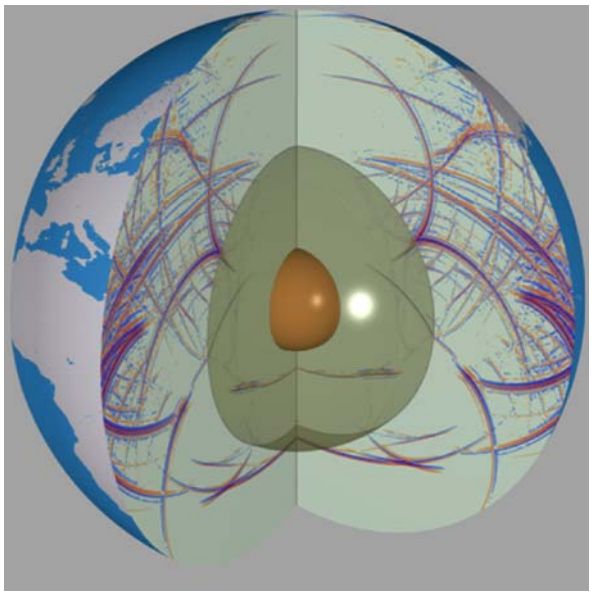

Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 2
Snapshot of wave propagation inside the Earth approx. 25 minutes after an earthquake occurs at the top part of the model. The radial component of motion is shown (blue and red denote positive and negative velocity, resp.). The simulation was carried out using an axi-symmetric approximation to the wave equation [55,58] and high-order finite-differences. Motion is allowed in the radial and horizontal directions. This corresponds to the P-SV case in 2D cartesian calculations. Therefore the wavefield contains both P- and S-waves and phase conversions

dicative of scattering in various parts of the Earth (see [62] for an account of modern observational seismology).

Adding a temporal and spatial scale to the above qualitative discussion reveals some important insight what it takes to simulate wave propagation on a planetary scale using grid-based numerical methods. Given the maximum frequency of around 1 Hz (P-waves) and 0.2 Hz (S-waves) the minimum wavelength in the Earth is expected to be O(km), requiring O(100 m) type grid spacing at least in the crustal part of the Earth leading to $O(10^{12})$ necessary grid points (or volume elements) for accurate numerical simulations. This would lead to memory requirements O(100 TByte) that are today possible on some of the world's largest supercomputers. The message here is that despite the rapid evolution of computational power, the complete modeling of teleseismic observations using approaches such as spectral elements (e. g., [63,64]) requiring tremendous numbers of calculations to constrain structure and sources will remain a grand challenge for some time to come. However, in many cases it is not necessary or not even desirable to simulate or model the whole seismogram, i. e. the complete observed frequency band. If we lower the cutoff frequency to 0.1 Hz (period 10 s), the required memory drops down to O(100 GByte). Such calculations can be done today on PC-clusters that can be inexpensively assembled and run on an institutional level (e. g., [8]). In addition, it means that the massive use of such forward simulations for imaging purposes and phenomenological investigations of wavefield effects is around the corner. This does not only apply to wave propagation or imaging on a planetary scale but in the same way to problems in volcanology, regional seismology, and exploration geophysics.

An illustration of global wave simulations using the finite difference method (e. g., [14,54,55,58,109,110,114]) is shown in Fig. 2 (more details on the methodologies are given in Sect. "The Evolution of Numerical Methods and Grids"). The snapshot of the radial component of motion at a time when the direct P-wave has almost crossed the Earth reveals the tremendous complexity the wave field exhibits even in the case of a spherically symmetric Earth model (PREM, [37]). The wavefield with a dominant period of ca. 15 seconds also highlights the short wavelengths that need to be propagated over very large distances. This is the special requirement for computational wave propagation that is quite different in other fields of computational Earth Sciences. While the theory of linear elastic wave propagation is well understood and most numerical methods have been applied to it in various forms, the accuracy requirements are so high that – particularly when models with complex geometrical features need to be modeled – there are still open questions as to what works best. One of the main goals of this paper is to highlight the need to focus on the grid generation process for various types of computational grid cells (e. g., rectangular, triangular in 2D, and hexahedral and tetrahedral in 3D) and the interface to appropriate highly accurate solvers for wave propagation problems.

As mentioned above computational modeling of strong ground motions following large earthquakes (see Fig. 3 for an illustration) is expected to play an increasingly important role in producing realistic estimates of shaking hazard. There are several problems that are currently unsolved: (1) to achieve frequencies that are interesting for earthquake engineers in connection with structural damage the near surface velocity structure needs to be known and frequencies beyond 5 Hz need to be calculated. In most cases this structure is not well known (on top of the uncertainties of the lower basin structures) and the required frequencies demand extremely large computational models. (2) In addition to structural uncertain-

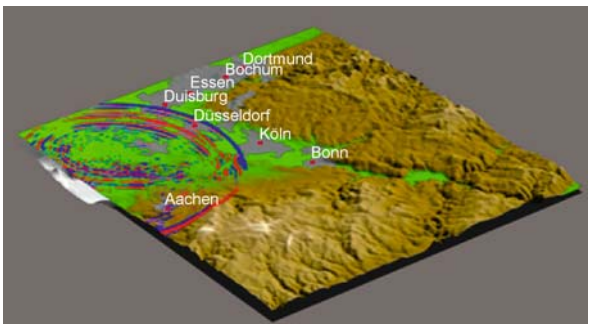

Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 3
Snapshot (horizontal component) for a simulation of the M5.9 Roermond earthquake in the Cologne Basin in 1992 [38]. The 3D sedimentary basin (maximum depth 2 km) leads to strong amplification and prolongation of the shaking duration that correlates well with basin depth. Systematic calculations may help mitigating earthquake induced damage

ties, there are strong dependencies on the particular earthquake rupture process that influence the observed ground motions. This suggests that many 3D calculations should be carried out for any characteristic earthquake of interest, to account for such variations (e. g., Wang et al. 2006). (3) The large velocity variations (e. g., 300 m/s up to 8 km/s) require locally varying grid densities which is difficult to achieve with some of the classical numerical methods in use (e. g. finite differences). Some of the potential routes are developed below.

In summary, computational simulation of 3D wave propagation will be more and more a central tool for seismology with application in imaging problems, earthquake rupture problems, questions of shaking hazard, volcano seismology and planetary seismology. In the following we briefly review the history of the application of numerical methods to wave propagation problems and the evolution of computational grids. The increasing complexity of models in terms of geometrical features and range of physical properties imposes the use of novel methodologies that go far beyond the initial approximations based on finite differences.

The Evolution of Numerical Methods and Grids

In this section we give a brief history of the application of numerical methods to the problem of seismic wave propagation. Such a review can not be complete, certainly gives a limited perspective, and only some key references are given. One of the points we would like to highlight is the evolution of the computational grids that are being employed for wave propagation problems and the consequences on the numerical methods of choice now and in the future.

Why do we need numerical approximations to elastic wave propagation problems at all? It is remarkable what we learned about the Earth without them! In the first decades in seismology, modeling of seismic observations was restricted to the calculation of ray-theoretical travel times in spherically symmetric Earth models (e. g., [13,16]). With the advent of computing machines these approaches could be extended to 2D and 3D media leading to ray-theoretical tomography and the images of the Earth's interior that we know today (e. g., [115]). The analytical solution of wave propagation in spherical coordinates naturally leads to spherical harmonics and the possible quasi-analytical solution of wave propagation problems in spherically symmetric media using normal modes. As this methodology leads to complete waveforms the term "waveform inversion" was coined for fitting the waveforms of surface waves by correcting the phase differences for surface waves at particular frequencies (e. g., [118]). This allowed the recovery of seismic velocity models particularly of crust and upper mantle (surface wave tomography). A similar approach in Cartesian layered geometry led to complete solutions of the wave equation in cylindrical coordinates through the summation of Bessel functions, the reflectivity method [46]. This method was later extended to spherical media through the Earth-flattening transformation [85]. Recently, ray-theory was extended allowing the incorporation of finite-frequency effects (e. g., [84]). The impact on the imaging process is still being debated.

Most of these methods are still today extremely valuable in providing first estimates of 2D or 3D effects and are important for the use in standard seismic processing due to their computational efficiency. Nevertheless, with the tremendous improvements of the quality of seismic observations we strive today to extract much more information on Earth's structure and sources from recorded waveforms. As waveforms are in most places strongly affected by 3D structural variations the application of numerical methods that solve "directly" the partial differential equations descriptive of wave propagation becomes mandatory. This necessity was recognized early on and the developments of numerical wave propagation began in the sixties of the 20th century.

Numerical Methods Applied to Wave Propagation Problems

The finite-difference technique was the first numerical method to be intensively applied to the wave propagation

problem (e. g., [1,6,61,77,82,83,88,89,116,117]). The partial differentials in the wave equation are replaced by finite differences leading to an extrapolation scheme in time that can either be implicit or explicit. The analysis of such simple numerical schemes led to concepts that are central to basically all numerical solutions of wave propagation problems. First, the discretization in space and time introduces a scale into the problem with the consequence that the numerical scheme becomes dispersive. This numerical dispersion – for the originally non-dispersive problem of purely elastic wave propagation – has the consequence that for long propagation distances wave pulses are no longer stable but disperse. The consequence is, that in any simulation one has to ascertain that enough grid points per wavelength are employed so that numerical dispersion is reduced sufficiently. Finding numerical schemes that minimize these effects has been at the heart of any new methodology ever since. Second, the so-called CFL criterion [24] that follows from the same theoretical analysis of the numerical scheme basically relates a "grid velocity" – the ratio between the space and time increments $\mathrm{d}x$ and $\mathrm{d}t$, respectively – to the largest physical velocity c in the model. In order to have a stable calculation, this ratio has to be smaller than a constant ε that depends on the specific scheme and the space dimension

$$c\frac{\mathrm{d}t}{\mathrm{d}x} \leq \varepsilon \,. \tag{1}$$

This simple relationship has important consequences: When the grid spacing $\mathrm{d}x$ must be small, because of model areas with low seismic velocities, then the time step $\mathrm{d}t$ has to be made smaller accordingly leading to an overall increase in the number of time steps and thus overall computational requirements. In addition, the early implementations where based on regular rectangular grids, implying that large parts of the model where carrying out unnecessary calculations. As shown below local time-stepping and local accuracy are important ingredients in efficient modern algorithms.

The fairly inaccurate low order spatial finite-difference schemes were later extended to high-order operators [26,48,49,50,51,56,76,103]. Nevertheless, the required number of grid points per wavelength was still large, particularly for long propagation distances. This has led to the introduction of pseudo-spectral schemes, "pseudo" because only the calculations of the derivatives where done in the spectral domain, but the wave equation was still solved in the space-time domain with a time-extrapolation scheme based on finite differences [10,45,47,67]). The advantage of the calculation of derivatives in the spectral domain is at hand: The Fourier theorem tells us that by multiplying the spectrum with ik, i being the imaginary unit and k the wavenumber, we obtain an *exact* derivative (exact to numerical precision) on a regular set of grid points. This sounds attractive. However, there are always two sides to the coin. The calculation requires FFTs to be carried out extensively and the original "local" scheme becomes a "global" scheme. This implies that the derivative at a particular point in the computational grid becomes dependent on any other point in the grid. This turns out to be computationally inefficient, in particular on parallel hardware. In addition, the Fourier approximations imply periodicity which makes the implementation of boundary conditions (like the free surface, or absorbing boundary conditions) difficult.

By replacing the basis functions (Fourier series) in the classical pseudo-spectral method with Chebyshev polynomials that are defined in a limited domain $(-1,1)$ the problem with the implementation of boundary problems found an elegant solution (e. g., [66,107,108]). However, through the irregular spacing of the Chebyshev collocation points (grid densification at the domain boundaries, see section below) new problems arose with the consequence that this approach was not much further pursued except in combination with a multi-domain approach in which the field variables exchange their values at the domain boundaries (e. g., [108]).

So far, the numerical solutions described are all based on the *strong* form of the wave equation. The finite-element method is another main scheme that found immediate applications to wave propagation problems (e. g., [79]). Finite element schemes are based on solving the *weak* form of the wave equation. This implies that the space- and time-dependent fields are replaced by weighted sums of basis (also called trial) functions defined inside elements. The main advantage of finite element schemes is that elements can have arbitrary shape (e. g., triangular, trapezoidal, hexahedral, tetrahedral, etc.). Depending on the polynomial order chosen inside the elements the spatial accuracy can be as desired. The time-extrapolation schemes are usually based on standard finite differences. There are several reasons why finite-element schemes were less widely used in the field of wave propagation. First, in the process a large system matrix needs to be assembled and must be inverted. Matrix inversion in principle requires global communication and is therefore not optimal on parallel hardware. Second, in comparison with the finite-element method, finite- difference schemes are more easily coded and implemented due to their algorithmic simplicity.

A tremendous step forward was the introduction of basis functions inside the elements that have spectral accuracy, e. g., Chebyshev or Legendre polynomials [11,15,39,40,65,86,90,98]. The so-called spectral element scheme became particularly attractive with the discovery that – by using Legendre polynomials – the matrices that required inversion became diagonal [65]. This implies that the scheme does no longer need global communication, it is a local scheme in which extrapolation to the next time step can be naturally parallelized. With the extension of this scheme to spherical grids using the cubed-sphere discretization [63,64] this scheme is today the method of choice on many scales unless highly complex models need to be initiated.

Most numerical schemes for wave propagation problems were based on regular, regular stretched, or hexahedral grids. The numerical solution to unstructured grids had much less attention, despite the fact that highly complex models with large structural heterogeneities seem to be more readily described with unstructured point clouds. Attempts were made to apply finite volume schemes to this problem [31], and other concepts (like natural neighbor coordinates [7] to find numerical operators that are applicable on unstructured grids [72,73,78]). These approaches were unfortunately not accurate enough to be relevant for 3D problems. Recently, a new flavor of numerical method found application to wave propagation on triangular or tetrahedral grids. This combination of a discontinuous Galerkin method with ideas from finite volume schemes [33,70] allows for the first time arbitrary accuracy in space and time on unstructured grids. While the numerical solution on tetrahedral grids remains computationally slower, there is a tremendous advantage in generating computational grids for complex Earth models. Details on this novel scheme are given below.

Before presenting two schemes (spectral elements and the discontinuous Galerkin method) and some applications in more detail we want to review the evolution of grids used in wave propagation problems.

Grids for Wave Propagation Problems

The history of grid types used for problems in computational wave propagation is tightly linked to the evolution of numerical algorithms and available computational resources. The latter in the sense that – as motivated in the introduction – even today realistic simulations of wave propagation are still computationally expensive. This implies that it is not sufficient to apply stable and simple numerical schemes and just use enough grid points per wavelength and/or extremely fine grids for geometrically complex models. Optimal mathematical algorithms that minimize the computational effort are still sought for as the recent developments show that are outlined in the following sections.

In Fig. 4 a number of different computational grids in two space dimensions is illustrated. The simple-most equally-spaced regular finite-difference grid is only of practical use in situations without strong material discontinuities. With the introduction of the pseudospectral method based on Chebyshev polynomials grids as shown in Fig. 4a grids appeared that are denser near the domain boundaries and coarse in the interior. While this enabled a much more efficient implementation of boundary conditions the ratio between the size of the largest to the smallest cell depends on the overall number of grid points per dimension and can be very large. This leads to very small time steps, that can in some way be compensated by grid stretching [9] but overall the problem remains. An elegant way of allowing grids to be of more practical shape is by stretching the grids using analytical functions (Fig. 4c, this basically corresponds to a coordinate transformation, e. g., [50,107]). By doing this either smooth surface topography or smoothly varying internal interfaces can be followed by the grid allowing a more efficient simulation of geometrical features compared to a blocky representation on standard finite difference grids.

The problem of global wave propagation using spherical coordinates (here in the two-dimensional, axi-symmetric approximation) nicely illustrates the necessity to have spatially varying grid density (e. g., [42,43,53,59,89,109]). The grid shown in Fig. 4b demonstrates that in spherical coordinates a regular discretization leads to grid distances that get smaller and smaller towards the center of the Earth. This is in contrast to what is required to efficiently model the Earth's velocity structure: Velocities are small near the surface (requiring high grid density) and increase towards the center of the Earth (requiring low grid density). One way of adjusting is by re-gridding the mesh every now and then, in this case doubling the grid spacing appropriately. This is possible, yet it requires interpolation at the domain boundaries that slightly degrades the accuracy of the scheme.

The problems with grid density and complex surfaces cry for the use of so-called unstructured grids. Let us define an unstructured grid as an initial set of points (a point cloud), each point characterized by its spatial coordinates. We wish to solve our partial differential equations on this point set. It is clear that – with appropriate grid generation software – it is fairly easy to generate such grids that obey exactly any given geometrical constraints be it in connection with surfaces or velocity mod-

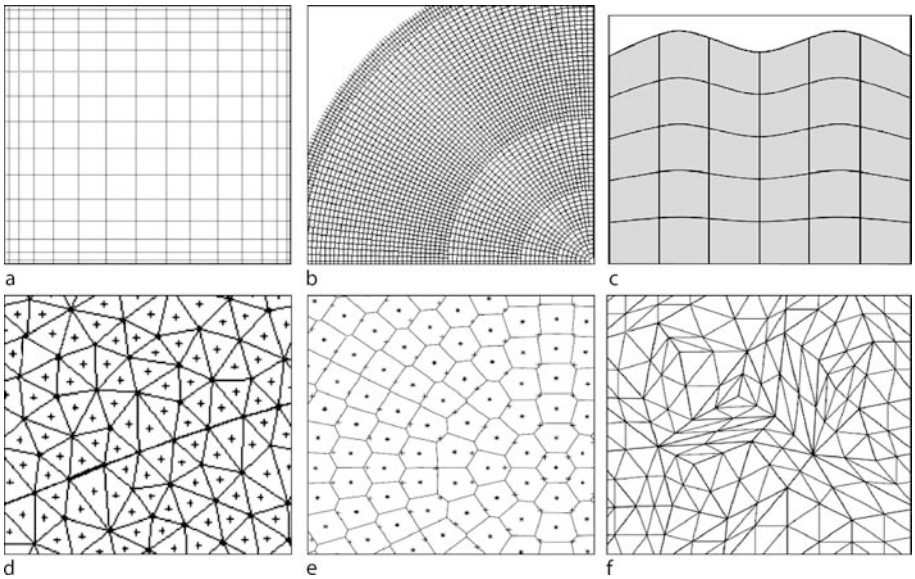

Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 4
Examples of 2D grids used for wave propagation simulations. a Chebyshev grid with grid densification near the domain boundaries. b Multidomain finite-difference grid in regular spherical coordinates. c Stretched regular finite-difference grid that allows following smoothly varying interface or surface boundaries. d Triangular staggered grid following an interface that allows finite-difference type operators. e Unstructured grid with associated Voronoi cells for calculations using the finite-volume method. f Triangular cells for finite-element type calculations. See text for details and references

els (i. e., varying grid density). It is important to note that such point clouds cannot be represented by 2D or 3D matrices as is the case for regular or regular stretched grid types. This has important consequences for the parallelization of numerical schemes. The first step after defining a point set is to use concepts from computational geometry to handle the previously unconnected points. This is done through the idea of Voronoi cells, that uniquely define triangles and their neighbors (Delauney triangulation). In Fig. 4d an example is shown for a triangular grid that follows an internal interface [72]. For finite-difference type operators on triangular grids a grid-staggering makes sense. Therefore, velocities would be defined in the center of triangles and stresses at the triangle vertices. Voronoi cells (Fig. 4e) can be used as volumetric elements for finite volume schemes [31,73]. For finite-element schemes triangular elements (Fig. 4f, e. g., [70]) with appropriate triangular shape functions are quite standard but have not found wide applications in seismology.

If the grid spacing of a regular finite-difference grid scheme in 3D would have to be halved this would result in an overall increase of computation time by a factor of 8 (a factor 2 per space dimension and another factor 2 because of the necessary halving of the time step). This simply means that the accuracy of a specific numerical scheme and the saving in memory or computation time is much more relevant in three dimensions. The evolution of grids in three dimensions is illustrated with examples in Fig. 5. A geometrical feature that needs to be modeled correctly particularly in volcanic environments is the free surface. With standard regular-spaced finite-difference schemes only a stair-step representation of the surface is possible (Fig. 5a, e. g., [87,92]). While the specific numerical implementation is stable and converges to the correct solution a tremendous number of grid points is necessary to achieve high accuracy.

Chebyshev grids and regular grids were applied to the problem of wave propagation in spherical sections (Fig. 5b, e. g., [52,57]). The advantage of solving the problem in spherical coordinates is the natural orthogonal coordinate system that facilitates the implementation of boundary conditions. However, due to the nature of spherical coordinates the physical domain should be close to the equator and geographical models have to be rotated accordingly. A highly successful concept for wave propagation in spherical media was possible through the adoption of the cubed-sphere approach in combination with spectral-elements (Fig. 5c, [63,64]). The cubed-sphere discretization is based on hexahedral grids. Towards the center of the Earth the grid spacing is altered to keep the number of elements per wavelength approximately constant.

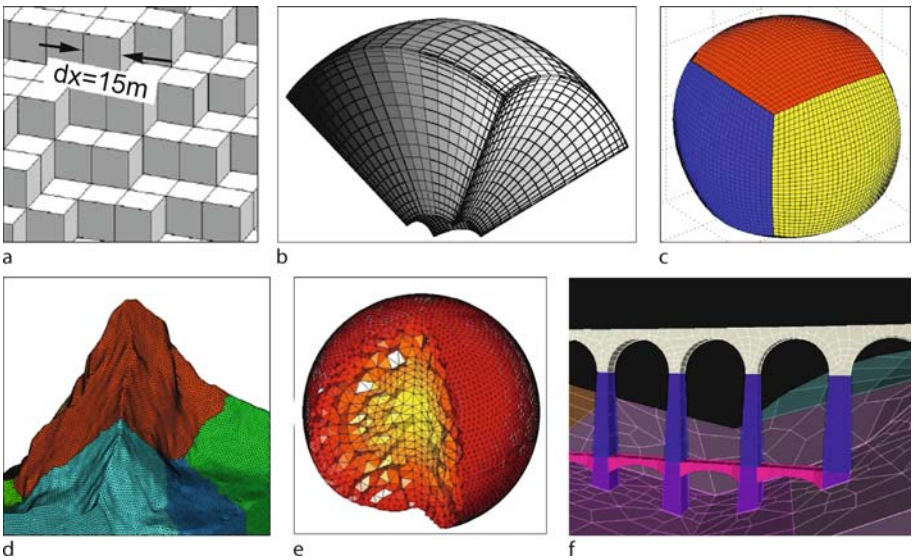

Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 5
Examples of 3D grids. a Stair-step representation of a complex free surface with finite-difference cells. b Chebyshev grid in spherical coordinates for a spherical section. c Cubed sphere grid used for spectral-element and multi-domain Chebyshev calculations. d Tetrahedral grid of the Matterhorn. e Tetrahedral grid of the Earth's interior with the grid density tied to the velocity model. f Hexahedral grid of bridge structure and subsurface structure for spectral-element calculations. See text for details and references

Computational grids for wave propagation based on tetrahedra (Fig. 5d,e) are only recently being used for seismic wave propagation in combination with appropriate numerical algorithms such as finite volumes [34] or discontinuous Galerkin (e. g., [70]). The main advantage is that the grid generation process is greatly facilitated when using tetrahedra compared to hexahedra. Generating point clouds that follow internal velocity structures and connecting them to tetrahedra are straight forward and efficient mathematical computations. However, as described in more detail below, tetrahedral grids require more involved computations and are thus less efficient than hexahedral grids. Complex hexahedral grids – even for combined modeling of structure and soil (Fig. 5f) are possible but – at least at present – require a large amount of manual interaction during the grid generation process. It is likely that the combination of both grid types (tetrahedral in complex regions, hexahedral in less complex regions) will play an important role in future developments.

In the following we would like to present two of the most competitive schemes presently under development, (1) the spectral element method and (2) the discontinuous Galerkin approach combined with finite-volume flux schemes. The aim is to particularly illustrate the role of the grid generation process and the pros and cons of the specific methodologies.

3D Wave Propagation on Hexahedral Grids: Soil-Structure Interactions

We briefly present the spectral element method (SEM) based on Lagrange polynomials, focusing only on its main features and on its implementation for the solution of the elasto-dynamic equations. The SEM can be regarded as a generalization of the finite element method (FEM) based on the use of high order piecewise polynomial functions. The crucial aspect of the method is the capability of providing an arbitrary increase in spatial accuracy simply enhancing the algebraic degree of these functions (the spectral degree SD). On practical ground, this operation is completely transparent to the users, who limit themselves to choosing the spectral degree at runtime, leaving to the computational code the task of building up suitable quadrature points for integration and new degrees of freedom. Obviously, the increasing spectral degree implies raising the required computational effort.

On the other hand, one can also play on the grid refinement to improve the accuracy of the numerical solution, thus following the standard finite element approach. Spectral elements are therefore a so-called "$h-p$" method, where "h" refers to the grid size and "p" to the degree of polynomials. Referring to Faccioli et al. [40], Komatitsch and Vilotte [65], Chaljub et al. [15] for further details, we briefly remind in the sequel the key features of the spectral

element method adopted. We start from the wave equation for the displacement $\boldsymbol{u}$:

$$\rho \frac{\partial \boldsymbol{u}^2}{\partial t^2} = \operatorname{div} \sigma_{ij}(\boldsymbol{u}) + f \,, \quad i, j = 1 \ldots \mathrm{d}(\mathrm{d} = 2, 3) \quad (2)$$

where t is the time, $\rho = \rho(x)$ the material density, $f = f(x, t)$ a known body force distribution and σ_{ij} the stress tensor. Introducing Hooke's law:

$$\sigma_{ij}(\boldsymbol{u}) = \lambda \operatorname{div} \boldsymbol{u} \delta_{ij} + 2\mu \varepsilon_{ij}(\boldsymbol{u}) \,, \quad (3)$$

where

$$\varepsilon_{ij}(\boldsymbol{u}) = \frac{1}{2} \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) \quad (4)$$

is the strain tensor, λ and μ are the Lamé coefficients, and δ_{ij} is the Kronecker symbol, i. e. $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$, otherwise.

As in the FEM approach, the dynamic equilibrium problem for the medium can be stated in the weak, or variational form, through the principle of virtual work [121] and through a suitable discretization procedure that depends on the numerical approach adopted, can be written as an ordinary differential equations system with respect to time:

$$[M]\, \ddot{\boldsymbol{U}}(t) + [K]\, \boldsymbol{U}(t) = \boldsymbol{F}(t) + \boldsymbol{T}(t) \quad (5)$$

with matrices $[M]$ and $[K]$, respectively the mass and stiffness matrices, and vectors $\boldsymbol{F}$ and $\boldsymbol{T}$ representing the contributions of external forces and traction conditions, respectively. In our SE approach, $\boldsymbol{U}$ denotes the displacement vector at the Gauss–Lobatto–Legendre (GLL) nodes, that correspond to the zeroes of the first derivatives of Legendre polynomial of degree N. The advancement of the numerical solution in time is provided by the explicit 2nd order leap-frog scheme. This scheme is conditionally stable and must satisfy the well known and already mentioned Courant–Friedrichs–Levy (CFL) condition. The key features of the SE discretization are described in the following.

Like in the FEM standard technique, the computational domain may be split into quadrilaterals in 2D or hexahedra in 3D, both the local distribution of grid points within the single element and the global mesh of all the grid points in the domain must be assigned. Many of these grid points are shared amongst several spectral elements. Each spectral element is obtained by a mapping of a master element through a suitable transformation and all computations are performed on the master element. Research is in progress regarding the introduction of triangular spectral elements [80]. The nodes within the element where displacements and spatial derivatives are computed, on which volume integrals are evaluated, are not necessarily equally spaced (similar to the Chebyshev approach in pseudospectral methods mentioned above). The interpolation of the solution within the element is done by Lagrange polynomials of suitable degree. The integration in space is done through Gauss–Lobatto–Legendre quadrature formula.

Thanks to this numerical strategy, the exponential accuracy of the method is ensured and the computational effort minimized, since the mass matrix results to be diagonal. The spectral element (SE) approach developed by Faccioli et al. [40] has been recently implemented in the computational code GeoELSE (GeoELasticity by Spectral Elements) [93,102,120] for 2D/3D wave propagation analysis. The most recent version of the code includes: (i) the capability of dealing with fully unstructured computational domains, (ii) the parallel architecture, and (iii) visco-plastic constitutive behavior [30]. The mesh can be created through an external software (e. g., CUBIT [25]) and the mesh partitioning is handled by METIS [81].

Hexahedral Grids

As already mentioned in the SEM here presented the computational domain is decomposed into a family of non overlapping quadrilaterals in 2D or hexahedra in 3D. The grid discretization should be suitable to accurately propagate up to certain frequencies. Obviously, owing to the strong difference of the mechanical properties between soft-soil and stiff-soil (or building construction material) and to the different geometrical details as well, the grid refinement needed in the various parts of the model varies substantially. Therefore, a highly unstructured mesh is needed to minimize the number of elements. While 3D unstructured tetrahedral meshes can be achieved quite easily with commercial or non commercial software, the creation of a 3D non structured hexahedra mesh is still recognized as a challenging problem. In the following paragraph we provide state of the art results concerning the mesh creation.

Grid Generation

Hexahedral grids have more severe restrictions in meshing efficiently. This is basically related to the intrinsic difficulty that arises from the mapping of the computational domain with this particular element. As a consequence automatic procedures have difficulty capturing specific boundaries, create poor quality elements, the assigned size is difficult to be preserved and the generation process is usually much slower compared to the tetrahedral mesh generation algo-

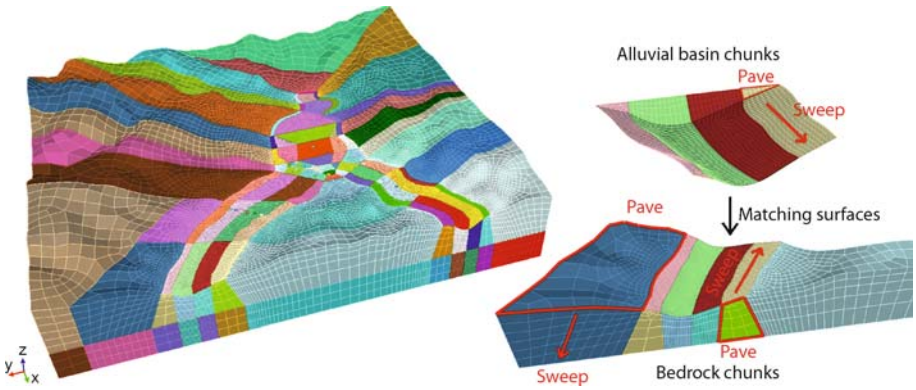

Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 6
3D numerical model used for the simulations of ESG06 "Grenoble Benchmark". "Honoring" technique: The computational domain is subdivided into small chunks and each one is meshed starting from the alluvial basin down to the bedrock. For simplicity only the spectral elements are shown without GLL nodes

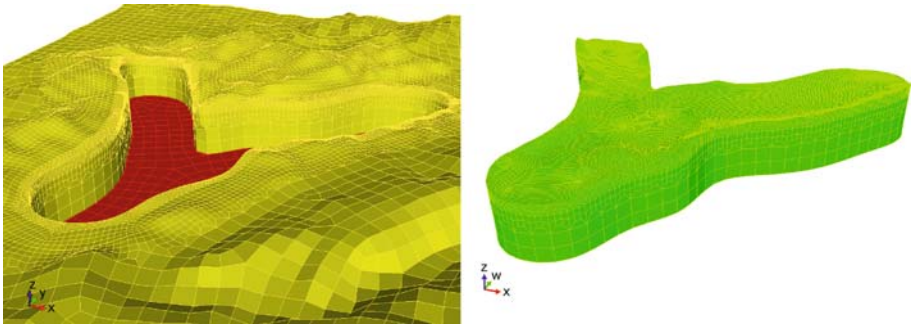

Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 7
3D numerical model used for the simulations of ESG06 "Grenoble Benchmark". "Not Honoring" technique: The computational domain is meshed with a coarse mesh and then refined twice approximately in the area where the alluvial basin is located

rithms. On the other hand the advantages of hexahedral meshes are usually related to the lower computational cost of the wave propagation solutions with respect to the one based on triangular meshes or hexahedral structured grids (like in the finite difference method).

Nevertheless certain problems can be addressed reasonably well with specific solutions. A quite typical case in earthquake seismology is the study of the alluvial basin response under seismic excitation. In handling this problem, a first strategy is to try to "honor" the interface between the sediment (soft soil) and the bedrock (stiff soil). The two materials are divided by a physical interface and the jump in the mechanical properties is strictly preserved. The major drawback of this approach is that usually it requires strong skills from the user to build-up the mesh and a significant amount of working time (Fig. 6). Given that the "honoring approach" is not always feasible in a reasonable time (or with a reasonable effort) a second strategy is worth to be mentioned: The so called "not honoring" procedure. In this second case the mesh is refined in proximity of the area where the soft deposit are localized but the elements do not respect the interface. On a practical ground the mechanical properties are assigned node by node and the sharp jump is smoothed through the Lagrange interpolation polynomial and substituted with smeared interfaces (Fig. 7). At the present time it is still strongly under debate if it is worth to honor or not the physical interfaces.

Finally, we highlight the fact that meshing software (e. g., CUBIT [25]) is available that seems to be extremely promising and potentially very powerful for the creation of geophysical and seismic engineering unstructured hexahedral meshes. Further very interesting mesh generation procedures based on hexahedral are under investigation [99].

Scale Problem with Structure and Soil

In engineering practice one of the most common approaches to design buildings under seismic load is the im-

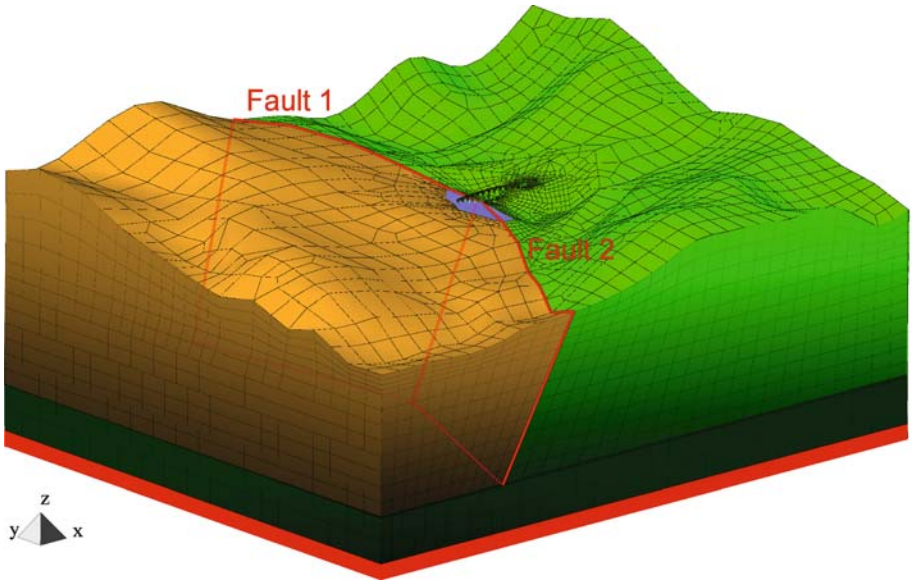

Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 8
3D model of Acquasanta bridge and the surrounding geological configuration. The investigated area is 2 km in length, 1.75 km in width and 0.86 km in depth. The model was designed to propagate waves up to 5 Hz with a SD = 3 (Order 4) and has 38,569 hexahedral elements and 1,075,276 grid points. The contact between calcareous schists (*brown color*) and serpentine rocks (*green color*) is modeled with two sub-vertical faults (*red-line*). *Cyan color* represents the alluvial and weathered deposits

position of an acceleration time history to the structure, basically acting like an external load. An excellent example of this technique can be found in recent publications (e. g., [68,69]) and in the study of the so-called "urban-seismology", recently presented by Fernandez-Ares et al., [44]. In this case the goal is to understand how the presence of an entire city can modify the incident wave-field. Due to the size of the simulation and the number of buildings, the latter are modeled as single degrees of freedom oscillators. The interaction between soil and structure is preserved but the buildings are simplified. For important structure (e. g.: Historical buildings, world heritage buildings, hospitals, schools, theaters, railway and highways) it is worth to provide an ad-hoc analysis capable to take into account the full complexity of the phenomena.

Here we present an example of a fully coupled modeling (Fig. 8): A railway bridge and its geotechnical-topographical surroundings. The Acquasanta bridge on the Genoa-Ovada railway, North Italy, is located in the Genoa district and represents a typical structure the ancestor of which can be traced back to the Roman "Pont du Gard". This structural type did not change significantly along the centuries, thanks to the excellent design achieved no less than 1900 years ago. The Acquasanta bridge structure is remarkable both for the site features and the local geological and geomorphological conditions. The foundations of several of the piers rest on weak rock; moreover, some instability problems have been detected in the past on the valley slope towards Ovada.

Several simulations have been performed with GeoELSE, in order to evaluate the influence of seismic site effects on the dynamic response of the Acquasanta bridge. A fully coupled 3D soil-structure model was designed: The grid is characterized by a "subvertical fault" between calcareous schists and serpentine rocks. This is in accordance with available data, even if further investigations in future should identify more in detail the tectonic structure of the area. The geometry of weathered materials overlaying the calcareous schists on the Ovada side has been assumed according to available information. The dimension of hexahedral elements ranges some tens of centimeters to about 1000 m. With such a model, the problem can be handled in its 3D complexity and we can examine the following aspects that are usually analyzed under restrictive and simplified assumptions: (i) soil-structure interaction, (ii) topographic amplification, (iii) soft soil amplification (caused by the superficial alluvium deposit shown in cyan), (iv) subvertical fault (red line) between the schists, on the Ovada side, and serpentine rock, on the Genoa side. For excitation a shear plane wave (x-direction) was used (Ricker wavelet, fmax = 3 Hz, $t0 = 1.0\,s$. and amplitude = 1 mm) propagating vertically from the bottom (red elements in Fig. 8).

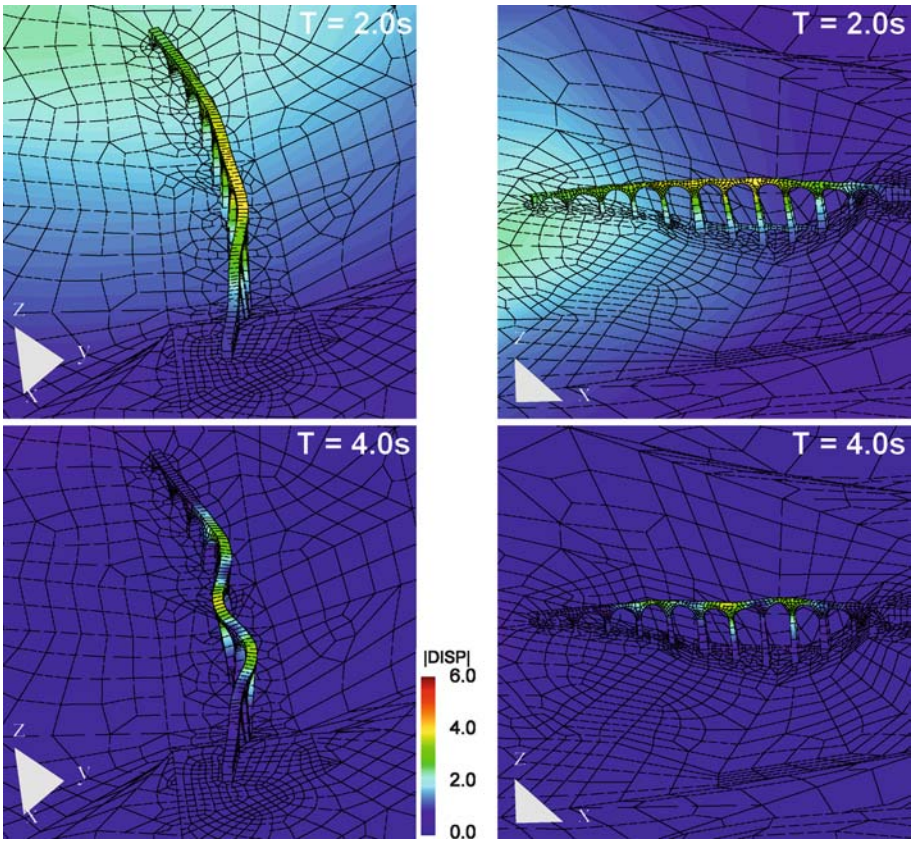

Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 9
Snapshots of the modulus of the displacement vector and the magnified deformed shape of the bridge (in mm)

In Fig. 9 we present some snapshots of the modulus of the displacement vector and the magnified deformed shape of the bridge. It is worth to note that at $T = 2$ s the motion of the bridge is almost in-plane (direction x), while at $T = 4$ s is clearly evident how the coupling between the in-plane and out-plane (y-direction) motion starts to be important.

The study of the soil-structure interaction problem could be easily enhanced (i) improving the input excitation of the model here presented and (ii) taking into account complex constitutive behavior both from the soil and the structure side. The former is already available in GeoELSE thanks to the recent implementation [41,93] of the domain reduction method (DRM), a methodology that divides the original problem into two simpler ones [4,119], to overcome the problem of multiple physical scales that is created by a seismic source located some kilometers away from the structure with typical element size of the order of meters and located over a relatively small area (less than 1 km^2) on soft deposit. The latter still needs to be improved because of the lack of a complete tool capable to handle in 3D non linear soil behavior, non-linear structural behavior and the presence of the water, that play a crucial role in the failure of buildings. Partial response to this problem can be found in the recent work of Bonilla et al. [5] and in the visco-plastic rheology recently introduced in GeoELSE [30].

3D Wave Propagation on Tetrahedral Grids: Application to Volcanology

As indicated above, the simulation of a complete, highly accurate wave field in realistic media with complex geometry is still a great challenge. Therefore, in the last years a new, highly flexible and powerful simulation method has been developed that combines the Discontinuous Galerkin (DG) Method with a time integration method using Arbitrary high order DERivatives (ADER) of the approximation polynomials. The unique property of this numerical scheme is, that it achieves arbitrarily high approximation order for the solution of the governing seismic

wave equation in space and time on structured and unstructured meshes in two and three space dimensions.

Originally, this new ADER-DG approach [32,35] was introduced for general linear hyperbolic equation systems with constant coefficients or for linear systems with variable coefficients in conservative form. Then, the extension to non-conservative systems with variable coefficients and source terms and its particular application to the simulation of seismic waves on unstructured triangular meshes in two space dimensions was presented [70]. And finally, the further extension of this approach to three-dimensional tetrahedral meshes has been achieved [33]. Furthermore, the accurate treatment of viscoelastic attenuation, anisotropy and poroelasticity has been included to handle more complex rheologies [28,29,71]. The governing system of the three-dimensional seismic wave equations is hereby formulated in velocity-stress and leads to the hyperbolic system of the form

$$\frac{\partial \boldsymbol{Q}_p}{\partial t} + A_{pq}\frac{\partial \boldsymbol{Q}_q}{\partial \xi} + B_{pq}\frac{\partial \boldsymbol{Q}_q}{\partial \eta} + C_{pq}\frac{\partial \boldsymbol{Q}_q}{\partial \zeta} = S_p \,, \quad (6)$$

where the vector $\boldsymbol{Q}$ of unknowns contains the six stress and the three velocity components and S is the source term. The Jacobian matrices A, B and C include the material values as explained in detail in [33,70].

The ADER-DG Method: Basic Concepts

The ADER-DG method is based on the combination of the ADER time integration approach [113], originally developed in the finite volume (FV) framework [96,97,111] and the Discontinuous Galerkin finite element method [18,19,20,21,22,23,91]. As described in detail in [33] in the ADER-DG approach the solution is approximated inside each tetrahedron by a linear combination of space-dependent polynomial basis functions and time-dependent degrees of freedom as expressed through

$$(\boldsymbol{Q}_h)_p(\xi,\eta,\zeta,t) = \hat{\boldsymbol{Q}}_{pl}(t)\boldsymbol{\Phi}_l(\xi,\eta,\zeta) \,, \quad (7)$$

where the basis functions Φ_l form an orthogonal basis and are defined on the canonical reference tetrahedron. The unknown solution inside each element is then approximated by a polynomial, whose coefficients – the degrees of freedom $\hat{\boldsymbol{Q}}_{pl}$ – are advanced in time. Hereby, the solution can be discontinuous across the element interfaces, which allows the incorporation of the well-established ideas of numerical flux functions from the finite volume framework [75,112]. To define a suitable flux over the element surfaces, so-called Generalized Riemann Problems (GRP) are solved at the element interfaces. The GRP solution provides simultaneously a numerical flux function as well as a time-integration method. The main idea is a Taylor expansion in time in which all time derivatives are replaced by space derivatives using the so-called Cauchy–Kovalewski procedure which makes recursive use of the governing differential Eq. (6). The numerical solution of Eq. (6) can thus be advanced by one time step without intermediate stages as typical e. g. for classical Runge–Kutta time stepping schemes. Due to the ADER time integration technique the same approximation order in space and time is achieved automatically. Furthermore, the projection of the elements in physical space onto a canonical reference element allows for an efficient implementation, as many computations of three-dimensional integrals can be carried out analytically beforehand. Based on a numerical convergence analysis this new scheme provides arbitrary high order accuracy on unstructured meshes. Moreover, due to the choice of the basis functions in Eq. (7) for the piecewise polynomial approximation [23], the ADER-DG method shows even spectral convergence.

Grid Generation: Unstructured Triangulations and Tetrahedralization

Both tetrahedral and hexahedral elements are effectively used to discretize three-dimensional computational domains and model wave propagation with finite element type methods. Tetrahedra can be the right choice because of the robustness when meshing any general shape. Hexahedra can be the element of choice due to their ability to provide more efficiency and accuracy in the computational process. Furthermore, techniques for automatic mesh generation, gradual mesh refinement and coarsening are generally much more robust for tetrahedral meshes in comparison to hexahedral meshes. Straightforward tetrahedral refinement schemes, based on longest-edge division, as well as the extension to adaptive refinement or coarsening procedures of a refined mesh exist [3,12]. In addition, parallel strategies for refinement and coarsening of tetrahedral meshes have been developed [27].

Less attention has been given to the modification of hexahedral meshes. Methods using iterative octrees have been proposed [74,95], but these methods often result in nonconformal elements that cannot be accommodated by some solvers. Lately also conformal refinement and coarsening strategies for hexahedral meshes have been proposed [2]. Other techniques insert non-hexahedral elements that result in hybrid meshes that need special solvers that can handle different mesh topologies. Commonly, the geometrical problems in geosciences arise through rough surface topography, as shown for the Mer-

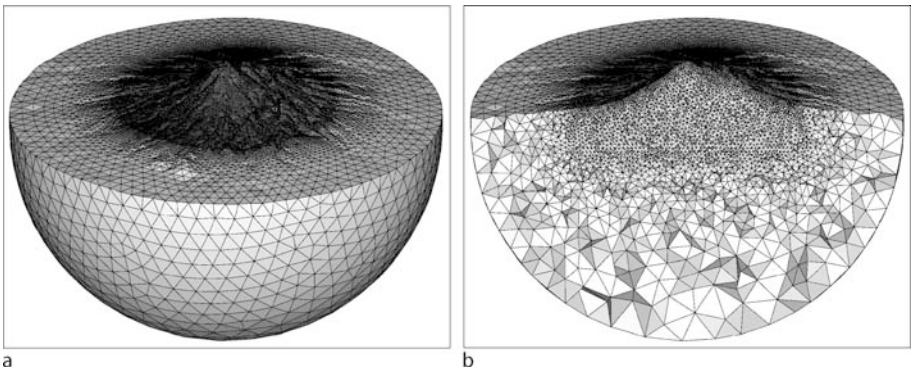

Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 10
Tetrahedral mesh for the model of the volcano Merapi. The zone of interest, such as the free surface topography and the volcano's interior are discretized by a fine mesh, whereas the spatial mesh is gradually coarsened towards the model boundaries

api volcano in Fig. 10, and internal material boundaries of complex shape that lead to wedges and overturned or discontinuous surfaces due to folding and faulting. However, once the geometry of the problem is defined by the help of modern computer aided design (CAD) software, the meshing process using tetrahedral elements is automatic and stable. After the mesh generation process, the mesh vertices, the connectivity matrix and particular information about boundary surfaces are typically imported to a solver.

The computational possibilities and algorithmic flexibility of a particular solver using the ADER-DG approach for tetrahedral meshes are presented in the following.

Local Accuracy: *p*-Adaptation

In many large scale applications the computational domain is much larger than the particular zone of interest. Often such an enlarged domain is chosen to avoid effects from the boundaries that can pollute the seismic wave field with possible, spurious reflections. Therefore, a greater number of elements has to be used to discretize the domain describing the entire model. However, in most cases the high order accuracy is only required in a restricted area of the computational domain and it is desirable to choose the accuracy that locally varies in space. This means, that it must be possible to vary the degree p of the approximation polynomials locally from one element to the other [36]. As the ADER-DG method uses a hierarchical order of the basis functions to construct the approximation polynomials, the corresponding polynomial coefficients, i. e. the degrees of freedom, for a lower order polynomial are always a subset of those of a higher-order one. Therefore, the computation of fluxes between elements of different approximation orders can be carried out by using only the necessary part of the flux matrices.

Furthermore, the direct coupling of the time and space accuracy via the ADER approach automatically leads to a local adaptation also in time accuracy, which often is referred to as p_t-adaptivity. In general, the distribution of the degree p might be connected to the mesh size h, i. e. the radius of the inscribed sphere of a tetrahedral element. In particular, the local degree p can be coupled to the mesh size h via the relations

$$p = p_{\min} + \left(p_{\max} - p_{\min}\right)\left(\frac{h - h_{\min}}{h_{\max} - h_{\min}}\right)^r , \qquad (8)$$

$$p = p_{\max} - \left(p_{\max} - p_{\min}\right)\left(\frac{h - h_{\min}}{h_{\max} - h_{\min}}\right)^r , \qquad (9)$$

where the choice of the power r determines the shape of the p-distribution. Note, that depending on the choice of the first term and the sign the degree p can increase as in Eq. (8) or decrease as in Eq. (9) with increasing h, starting from a minimum degree $p_{\min}$ up to a maximum degree $p_{\max}$. This provides additional flexibility for the distribution of p inside the computational domain. An example of a p-distribution for the volcano Merapi is given in Fig. 11.

Here the idea is to resolve the slowly propagating surface waves with high accuracy, whereas the waves propagating towards the absorbing model boundaries pass through a zone of low spatial resolution. This approach leads to numerical damping due to an amplitude decay that reduces possible boundary reflections. Furthermore, the computational cost is reduced significantly due to the strongly reduced number of total degrees of freedom in the model.

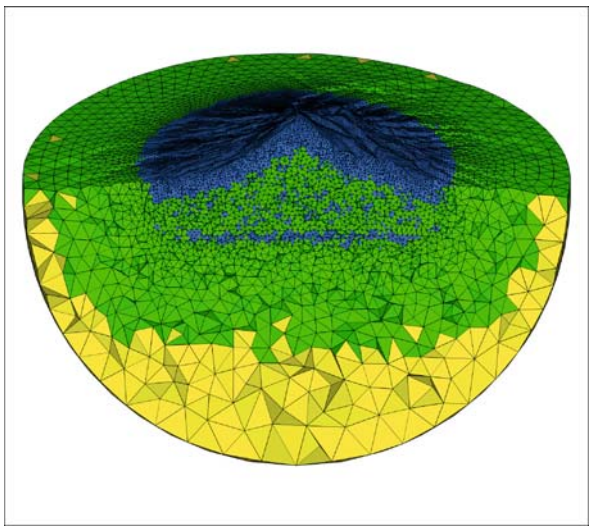

Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 11
The local degree *p* of the approximation polynomial depends on the insphere radius of each tetrahedral element and is given in color code. Close to the surface topography an approximation polynomial of degree $p = 5$ (*blue*) is used, whereas in depth the degree is reduced to $p = 4$ (*green*) and $p = 3$ (*yellow*)

Local Time Stepping: Δt-Adaptation

Geometrically complex computational domains or spatial resolution requirements often lead to meshes with small or even degenerate elements. Therefore, the time step for explicit numerical schemes is restricted by the ratio of the size h of the smallest element and the corresponding maximum wave speed in this element. For global time stepping schemes all elements are updated with this extremely restrictive time step length leading to a large amount of iterations. With the ADER-DG approach, time accurate local time stepping can be used, such that each element is updated by its own, optimal time step [36]. Local time-stepping was used in combination with the finite-difference method [42,106].

An element can be updated to the next time level if its actual time level and its local time step Δt fulfill the following condition with respect to all neighboring tetrahedra n:

$$t + \Delta t \leq \min(t_n + \Delta t_n) \,. \tag{10}$$

Figure 12 is visualizing the evolution of four elements (I, II, III and IV) in time using the suggested local time stepping scheme. A loop cycles over all elements and checks for each element, if condition (10) is fulfilled. At the initial state all elements are at the same time level, however, element II and IV fulfill condition (10) and therefore can be updated. In the next cycle, these elements are already advanced in time (grey shaded) in cycle 1. Now elements I and IV fulfill condition (10) and can be updated to their next local time level in cycle 2. This procedure continues and it is obvious, that the small element IV has to be updated more frequently than the others. A synchronization to a common global time level is only necessary, when data output at a particular time level is required as shown in Fig. 12.

Information exchange between elements across interfaces appears when numerical fluxes are calculated. These fluxes depend on the length of the local time interval over which a flux is integrated and the corresponding element is evolved in time. Therefore, when the update criterion (10) is fulfilled for an element, the flux between the element itself and its neighbor n has to be computed over the local time interval:

$$\tau n = [\max(t, tn) \,, \;\; \min(t + \Delta t, tn + \Delta tn)] \,. \tag{11}$$

As example, the element III fulfills the update criterion (10) in cycle 5 (see Fig. 12). Therefore, when computing the fluxes only the remaining part of the flux given by the intervals in Eq. (11) has to be calculated. The other flux contribution was already computed by the neighbors II and IV during their previous local updates. These flux contributions have been accumulated and were stored into a memory variable and therefore just have to be added.

Note that e. g. element IV reaches the output time after 10 cycles and 9 local updates, which for a global time stepping scheme would require $9 \times 4 = 36$ updates for the all four elements. With the proposed local time stepping scheme only 16 updates are necessary to reach the same output time with all elements as indicated by the final number of grey shaded space time elements in Fig. 12.

Comparing these numbers leads to a speedup factor of 2.25. For strongly heterogeneous models and local time step lengths this factor can become even more pronounced. However, due to the asynchronous update of elements that might be spatially very close to each other the mesh partitioning for parallel computations becomes an important and difficult issue. Achieving a satisfying load balancing is a non-trivial task and still poses some unresolved problems as explained in the following.

Mesh Partitioning and Load Balancing

For large scale applications it is essential to design a parallel code that can be run on massively parallel supercomputing facilities. Therefore, the load balancing is an important issue to use the available computational resources efficiently. For global time stepping schemes with-

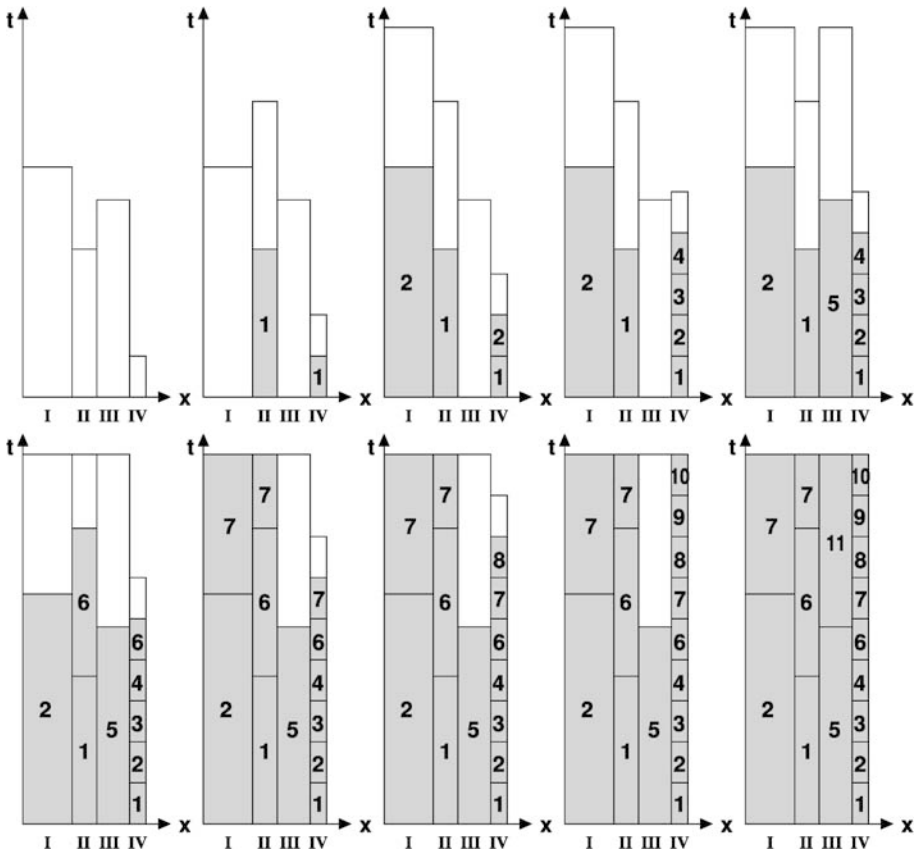

Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 12
Visualization of the local time stepping scheme. The actual local time level *t* is at the top of the *gray shaded* area with numbers indicating the cycle, in which the update was done. *Dotted lines* indicate the local time step length Δt with which an element is updated

out *p*-adaptation standard mesh partitioning as done e. g. by METIS [60] is sufficient to get satisfying load balancing. The unstructured tetrahedral mesh is partitioned into subdomains that contain an equal or at least very similar number of elements as shown in Fig. 13. Therefore, each processor has to carry out a similar amount of calculations. However, if *p*-adaptation is applied, the partitioning is more sophisticated as one subdomain might have many elements of high order polynomials whereas another might have the same number of elements but with lower order polynomials. Therefore, the parallel efficiency is restricted by the processor with the highest work load. However, this problem can usually be solved by weighted partitioning algorithms, e. g. METIS.

In the case of local time stepping, mesh partitioning is becoming a much more difficult task. One solution is to divide the computational domain into a number of zones, that usually contain a geometrical body or a geological zone that typically is meshed individually with a particular mesh spacing *h* and contains a dominant polynomial order. Then each of these zones is partitioned separately into subdomains of approximately equal numbers of elements. Then each processor receives a subdomain of each zone, which requires a similar amount of computational work as shown in Fig. 13. In particular, the equal distribution of tetraheda with different sizes is essential in combination with the local time stepping technique. Only if each processor receives subdomains with a similar amount of small and large elements, the work load is balanced. The large elements have to be updated less frequently than the smaller elements and therefore are computationally cheaper. Note, that the separately partitioned and afterwards merged zones lead to non-connected subdomains for each processor (see Fig. 13). This increases the number of element surfaces between subdomains of different processors and therefore increases the communication required. However, communication is typically low as the degrees of freedom have to be exchanged only once per

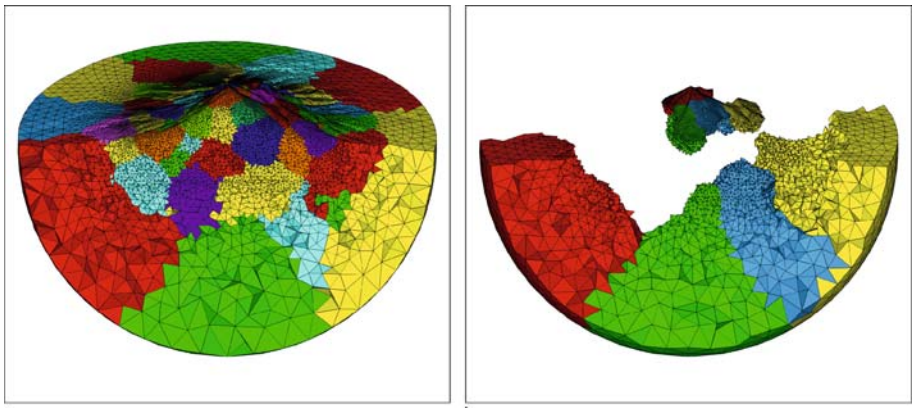

Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 13
Standard partitioning of the computational domain (*left*) and an example of 4 subdomains grouped together for more efficient local time stepping

time step and only for tetrahedra that have an interface at the boundary between subdomains. Therefore, the improvements due to the new load balancing approach are dominant and outweigh the increase in communication.

However, care has to be taken as the distribution of the polynomial degree p or the seismic velocity structure might influence the efficiency of this grouped partitioning technique. A profound and thorough mesh partitioning method is still a pending task as the combination of local time stepping and p-adaptivity requires a new weighting strategy of the computational cost for each tetrahedral element considering also the asynchronous element update. The automatic partitioning of unstructured meshes with such heterogeneous properties together with the constraint of keeping the subdomains as compact as possible to avoid further increase of communication is still subject to future work.

In Fig. 13 an example of a grouped partition of the tetrahedral mesh is shown for 4 processors. Two non-connected subdomains indicated by the same color are assigned to each processor including small – and therefore computationally expensive – tetrahedra that are updated frequently due to their small time step, and much larger elements that typically are cheap due to their large time step. This way, the work load often is balanced sufficiently well over the different processors.

Relevance of High Performance Computing: Application to Merapi Volcano

In recent years the development of the ADER-DG algorithm including the high order numerical approximation in space and time, the mesh generation, mesh adaptation, parameterization, and data visualization created the basis of an efficient and highly accurate seismic simulation tool. Realistic large scale applications and their specific requirements will further guide these developments. On the other hand, the study and incorporation of geophysical processes that govern seismic wave propagation insures, that the simulation technology matches the needs and addresses latest challenges in modern computational seismology. Hereby, the accurate modeling of different source mechanisms as well as the correct treatment of realistic material properties like anelasticity, viscoplasticity, porosity and highly heterogeneous, scattering media will play an important role.

However, only the combination of this state-of-the-art simulation technology with the most powerful supercomputing facilities actually available can provide excellent conditions to achieve scientific progress for realistic, large scale applications. This combination of modern technologies will substantially contribute to resolve current problems, not only in numerical seismology, but will also influence other disciplines. The phenomenon of acoustic, elastic or seismic wave propagation is encountered in many different fields. Beginning with the classical geophysical sciences seismology, oceanography, and volcanology such waves also appear in environmental geophysics, atmospheric physics, fluid dynamics, exploration geophysics, aerospace engineering or even medicine.

With the rapid development of modern computer technology and the development of new highly accurate simulation algorithms computer modeling just started to herald a new era in many applied sciences. The 3D wave propagation simulations in realistic media require a substantial amount of computation time even on large par-

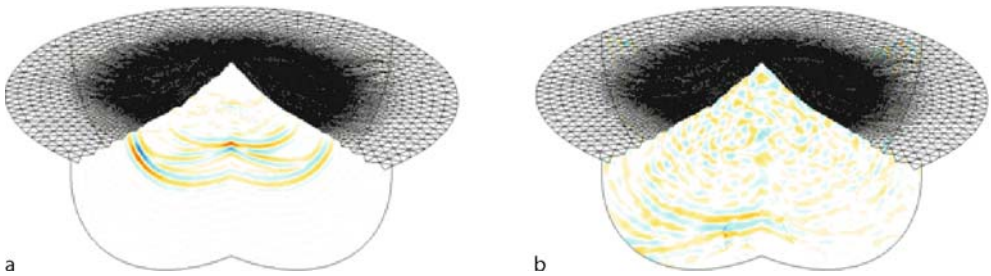

Seismic Wave Propagation in Media with Complex Geometries, Simulation of, Figure 14
Snapshots of the seismic wave field after an explosive event close to the summit of Merapi volcano. The free surface topography introduces strong scattering of the waves making it extremely difficult to invert for the seismic source mechanism or the exact source location

allel computers. Extremely powerful national supercomputers already allow us to run simulations with unrivaled accuracy and resolution. However, using the extremely high accuracy and flexibility of new simulation methods on such massively parallel machines the professional support of experts in supercomputing is absolutely essential. Only professional porting, specific CPU-time and storage optimizations of current software with respect to continuously changing compilers, operating systems, hardware architectures or simply personnel, will ensure the lifetime of new simulation technologies accompanied by ongoing improvements and further developments. Additionally, the expertise and support in the visualization of scientific results using technologies of Virtual Reality for full 3D models not only enhances the value of simulation results but will support data interpretation and awake great interest in the new technology within a wide research community.

As an example, volcano monitoring plays an increasingly important role in hazard estimation in many densely populated areas in the world. Highly accurate computer modeling today is a key issue to understand the processes and driving forces that can lead to dome building, eruptions or pyroclastic flows. However, data of seismic observations at volcanoes are often very difficult to interpret. Inverting for the source mechanism, i. e. seismic moment tensor inversion, or just locating an exact source position is often impossible due to the strongly scattered wave field caused by an extremely heterogeneous material distribution inside the volcano. Furthermore, the rough topography alone can affect the wave field by its strongly scattering properties as shown in Fig. 14.

Therefore, it is fundamental to understand the effects of topography and scattering media and there influence on the seismic wave field. A systematic study of a large number of scenarios computed by highly accurate simulation methods to provide reliable synthetic data sets is necessary to test the capabilities of currently used inversion tools. Slight changes in parameters like the source position, the source mechanism or the elastic and geometric properties of the medium can then reveal the limits of such tools and provide more precise bounds of their applicability in volcano seismology.

Finally, the implementation of the ADER-DG method is still much more expensive than other state-of-the-art implementations of existing methods. However, a fair comparison between accuracy and computational cost is still a pending task. The main reason for the CPU-time difference is the much larger number of tetrahedral elements than hexahedra that have to be used to cover the same volume. Furthermore, due to the choice of the basis functions, the flux computations are expensive, as the matrix-matrix multiplications involved are not sparse.

However, the ADER-DG method is currently implemented on hexahedral meshes to make fair comparisons possible. Preliminary tests show, that the change of mesh topology from tetrahera to hexahedra significantly reduces the computational cost. However, final results are subject to future investigations.

Discussion and Future Directions

As indicated in the introduction and highlighted in the previous sections, computational tools for wave propagation problems are getting increasingly sophisticated to meet the needs of current scientific problems. We are far away from simple finite-difference time schemes that are solving problems on regular grids on serial computers in which case the particular programming approach did not affect dramatically the overall performance. Today, competitive algorithms are results of years of partly highly professional coding. Implementations on high-performance computing hardware requires in-depth knowledge of parallel algorithms, profiling, and many technical aspects of modern computations. To make complex scientific soft-

ware available to other researchers requires implementation and testing on many different (parallel) platforms. This may involve parallelization using different programming paradigms (e. g., the combination of OpenMP and MPI on nodes of shared memory machines), and interoperability on heterogeneous computational GRIDs.

This has dramatic consequences particularly for young researchers in the Earth Sciences who want to use advanced computational tools to model observations. While in the early days a finite-difference type algorithm could be understood, coded, implemented and tested in a few weeks, this is no longer possible. In addition, standard curricula do not offer training in computational methods allowing them to efficiently write and test codes. This suggests that at least for some, well-defined computational problems verified and professionally engineered scientific software solutions should be provided to the community and also professionally extended and maintained in close collaboration with scientists. In seismology we are in a quite fortunate situation. In contrast to many other fields of physical sciences, our constitutive relations (e. g., stress-strain) are fairly well understood, and – as indicated in this paper – numerical solutions for 3D problems and their implementation on parallel hardware are well advanced. Another argument for stable tested "community"-codes for wave propagation is the fact that advancement in many scientific problems (e. g., imaging the Earth's interior, quantifying earthquake-induced shaking hazard) relies on zillions of forward modeling runs with only slight variations of the internal velocity models.

As far as technical developments are concerned, the efficient initialization of complex 3D models on computational grids is still a great challenge. Realistic models may be composed of complex topography, families of overlapping fault surfaces, discontinuous interfaces, and varying rheologies (e. g., elastic, anisotropic, viscoelastic, viscoplastic, porous). This may require the combination of tetrahedral and hexahedral grid in models with strongly varying degree of complexity. Ideally, standards for Earth models (and synthetic data) formats should be established by the communities that allow easy exchange and multiple use of models with different simulation tools (e. g., wave propagation, deformation, earthquake rupture). In addition, the rapid developments towards PetaFlop computing opens new questions about the scalability and efficient parallelization of current and future algorithms.

As the forward problem of wave propagation is at the core of the seismic imaging problem for both source and Earth's structure, in the near future we will see the incorporation of 3D simulation technology into the imaging process. Provided that the background seismic velocity models are fairly well known (e. g., reservoirs, global Earth, sedimentary basins), adjoint methods provide a powerful analytical tool to (1) relate model deficiencies to misfit in observations and (2) quantify the sensitivities to specific aspects of the observations (e. g., [100,104,105]). As the core of the adjoint calculations is the seismic forward problem, the challenge is the actual application to real data and the appropriate parametrizations of the model and the data that optimize the data fitting process.

In summary, while we look back at (and forward to) exciting developments in computational seismology, a paradigm shift in the conception of one of the central tools of seismology – the calculation of 3D synthetic seismograms – is necessary. To extract a maximum amount of information from our high-quality observations scientists should have access to high-quality simulation tools. It is time to accept that "*software is infrastructure*" and provide the means to professionally develop and maintain community codes and model libraries at least for basic Earth science problems and specific focus regions. Developments are one the way along those lines in the SPICE project (Seismic Wave Propagation and Imaging in Complex Media, a European Network [101]), the Southern California Earthquake Center (SCEC [94]) and the CIG Project (Computational infrastructure in geodynamics [17]).

Acknowledgments

We would like to acknowledge partial support towards this research from: The European Human Resources and Mobility Program (SPICE-Network), the German Research Foundation (Emmy Noether-Programme), the Bavarian Government (KOHNWIHR, graduate college THESIS, BaCaTec), and MunichRe. We would also like to thank J. Tromp supporting MS's visit to CalTech. We also thank two anonymous reviewers for constructive comments on the manuscript.

Bibliography

Primary Literature

1. Alterman Z, Karal FC (1968) Propagation of elastic waves in layered media by finite-difference methods. Bull Seism Soc Am 58:367–398
2. Benzley SE, Harris NJ, Scott M, Borden M, Owen SJ (2005) Conformal refinement and coarsening of unstructured hexahedral meshes. J Comput Inf Sci Eng 5:330–337
3. Bey J (1995) Tetrahedral grid refinement. Computing 55:355–378
4. Bielak J, Loukakis K, Hisada Y, Yoshimura C (2003) Domain reduction method for three-dimensional earthquake modeling in localized regions, Part I: Theory. Bull Seism Soc Am 93:817–824

5. Bonilla LF, Archuleta RJ, Lavallée D (2005) Hysteretic and dilatant behavior of cohesionless soils and their effects on nonlinear site response: Field data observations and modelling. Bull Seism Soc Am 95(6):2373–2395
6. Boore D (1972) Finite-difference methods for seismic wave propagation in heterogeneous materials. In: Bolt BA (ed) Methods in Computational Physics, vol 11. Academic Press, New York
7. Braun J, Sambridge MS (1995) A numerical method for solving partial differential equations on highly irregular evolving grids. Nature 376:655–660
8. Bunge HP, Tromp J (2003) Supercomputing moves to universities and makes possible new ways to organize computational research. EOS 84(4):30, 33
9. Carcione JM, Wang J-P (1993) A Chebyshev collocation method for the elastodynamic equation in generalised coordinates. Comp Fluid Dyn 2:269–290
10. Carcione JM, Kosloff D, Kosloff R (1988) Viscoacoustic wave propagation simulation in the earth. Geophysics 53: 769–777
11. Carcione JM, Kosloff D, Behle A, Seriani G (1992) A spectral scheme for wave propagation simulation in 3-D elastic-anisotropic media. Geophysics 57:1593–1607
12. Carey G (1997) Computational grids: Generation, adaptation, and solution strategies. Taylor Francis, New York
13. Cerveny V (2001) Seismic ray theory. Cambridge University Press, Cambridge
14. Chaljub E, Tarantola A (1997) Sensitivity of SS precursors to topography on the upper-mantle 660-km discontinuity. Geophys Res Lett 24(21):2613–2616
15. Chaljub E, Komatitsch D, Vilotte JP, Capdeville Y, Valette B, Festa G (2007) Spectral element analysis in seismology. In: Wu R-S, Maupin V (eds) Advances in wave propagation in heterogeneous media. Advances in Geophysics, vol 48. Elsevier, London, pp 365–419
16. Chapman CH (2004) Fundamentals of seismic wave propagation. Cambridge University Press, Cambridge
17. CIG www.geodynamics.org. Accessed 1 Jul 2008
18. Cockburn B, Shu CW (1989) TVB Runge"Kutta local projection discontinuous Galerkin finite element method for conservation laws II: General framework. Math Comp 52:411–435
19. Cockburn B, Shu CW (1991) The Runge–Kutta local projection P1-Discontinuous Galerkin finite element method for scalar conservation laws. Math Model Numer Anal 25:337–361
20. Cockburn B, Shu CW (1998) The Runge–Kutta discontinuous Galerkin method for conservation laws V: Multidimensional systems. J Comput Phys 141:199–224
21. Cockburn B, Lin SY, Shu CW (1989) TVB Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws III: One dimensional systems. J Comput Phys 84:90–113
22. Cockburn B, Hou S, Shu CW (1990) The Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: The multidimensional case. Math Comp 54:545–581
23. Cockburn B, Karniadakis GE, Shu CW (2000) Discontinuous Galerkin methods, theory, computation and applications. LNCSE, vol 11. Springer, New York
24. Courant R, Friedrichs KO, Lewy H (1928) Über die partiellen Differenzialgleichungen der mathematischen Physik. Mathematische Annalen 100:32–74
25. CUBIT cubit.sandia.gov. Accessed 1 Jul 2008
26. Dablain MA (1986) The application of high-order differencing to the scalar wave equation. Geophysics 51:54–66
27. De Cougny HL, Shephard MS (1999) Parallel refinement and coarsening of tetrahedral meshes. Int J Numer Methods Eng 46:1101–1125
28. de la Puente J, Dumbser M, Käser M, Igel H (2007) Discontinuous Galerkin methods for wave propagation in poroelastic media. to appear in Geophysics
29. de la Puente J, Käser M, Dumbser M, Igel H (2007) An arbitrary high order discontinuous Galerkin method for elastic waves on unstructured meshes IV: Anisotropy. Geophys J Int 169(3):1210–1228
30. di Prisco C, Stupazzini M, Zambelli C (2007) Non-linear SEM numerical analyses of dry dense sand specimens under rapid and dynamic loading. Int J Numer Anal Methods Geomech 31(6):757–788
31. Dormy E, Tarantola A (1995) Numerical simulation of elastic wave propagation using a finite volume method. J Geophys Res 100(B2):2123–2134
32. Dumbser M (2005) Arbitrary high order schemes for the solution of hyperbolic conservation laws in complex domains. Shaker, Aachen
33. Dumbser M, Käser M (2006) An arbitrary high order discontinuous galerkin method for elastic waves on unstructured meshes II: The three-dimensional isotropic case. Geophys J Int 167:319–336
34. Dumbser M, Käser M (2007) Arbitrary high order non-oscillatory finite volume schemes on unstructured meshes for linear hyperbolic systems. J Comput Phys 221:693–723. doi:10.1016/j.jcp.2006.06.043
35. Dumbser M, Munz CD (2005) Arbitrary high order discontinuous Galerkin schemes. In: Cordier S, Goudon T, Gutnic M, Sonnendrucker E (eds) Numerical methods for hyperbolic and kinetic problems. IRMA series in mathematics and theoretical physics. EMS Publishing, Zurich, pp 295–333
36. Dumbser M, Käser M, Toro EF (2007) An arbitrary high-order discontinuous Galerkin method for elastic waves on unstructured meshes – V. Local time stepping and p-adaptivity. Geophys J Int 171:695–717
37. Dziewonski AM, Anderson DL (1981) Preliminary reference earth model. Phys Earth Planet Inter 25:297–356
38. Ewald M, Igel H, Hinzen K-G, Scherbaum F (2006) Basin-related effects on ground motion for earthquake scenarios in the lower rhine embayment. Geophys J Int 166:197–212
39. Faccioli E, Maggio F, Quarteroni A, Tagliani A (1996) Spectral-domain decomposition methods for the solution of acoustic and elastic wave equation. Geophysics 61:1160–1174
40. Faccioli E, Maggio F, Paolucci R, Quarteroni A (1997) 2D and 3D elastic wave propagation by a pseudo-spectral domain decomposition method. J Seismol 1:237–251
41. Faccioli E, Vanini M, Paolucci R, Stupazzini M (2005) Comment on "Domain reduction method for three-dimensional earthquake modeling in localized regions, part I: Theory." by Bielak J, Loukakis K, Hisada Y, Yoshimura C, and "Part II: Verification and Applications." by Yoshimura C, Bielak J, Hisada Y, Fernández A. Bull Seism Soc Am 95:763–769
42. Falk J, Tessmer E, Gajewski D (1996) Efficient finite-difference modelling of seismic waves using locally adjustable time steps. Geophys Prosp 46:603–616

43. Falk J, Tessmer E, Gajewski D (1996) Tube wave modelling by the finite differences method with varying grid spacing. Pure Appl Geoph 148:77–93
44. Fernandez A, Bielak J, Prentice C (2006) Urban seismology; City effects on earthquake ground motion and effects of spatial distribution of ground motion on structural response paper presented at 2006 annual meeting. Seism Res Lett 77(2):305
45. Fornberg B (1996) A practical guide to pseudospectral methods. Cambridge University Press, Cambridge
46. Fuchs K, Müller G (1971) Computation of synthetic seismograms with the reflectivity method and comparison with observations. Geophys J Royal Astronom Soc 23(4):417–33
47. Furumura T, Takenaka H (1996) 2.5-D modeling of elastic waves using the pseudospectral method. Geophys J Int 124: 820–832
48. Geller RJ, Takeuchi N (1998) Optimally accurate second-order time-domain finite difference scheme for the elastic equation of motion: One-dimensional case. Geophys J Int 135:48–62
49. Graves RW (1993) Modeling three-dimensional site response effects in the Marina district basin, San Francisco, California. Bull Seism Soc Am 83:1042–1063
50. Hestholm SO, Ruud BO (1998) 3-D finite-difference elastic wave modeling including surface topography. Geophysics 63:613–622
51. Holberg O (1987) Computational aspects of the choice of operator and sampling interval for numerical differentiation in large-scale simulation of wave phenomena. Geophys Prospect 35:629–655
52. Igel H (1999) Wave propagation through 3-D spherical sections using the Chebyshev spectral method. Geop J Int 136:559–567
53. Igel H, Gudmundsson O (1997) Frequency-dependent effects on travel times and waveforms of long-period S and SS waves. Phys Earth Planet Inter 104:229–246
54. Igel H, Weber M (1995) SH-wave propagation in the whole mantle using high-order finite differences. Geophys Res Lett 22(6):731–734
55. Igel H, Weber M (1996) P-SV wave propagation in the Earth's mantle using finite-differences: Application to heterogeneous lowermost mantle structure. Geophys Res Lett 23:415–418
56. Igel H, Mora P, Riollet B (1995) Anisotropic wave propagation through finite-difference grids. Geophysics 60:1203–1216
57. Igel H, Nissen-Meyer T, Jahnke G (2001) Wave propagation in 3-D spherical sections: Effects of subduction zones. Phys Earth Planet Inter 132:219–234
58. Jahnke G, Igel H, Cochard A, Thorne M (2007) Parallel implementation of axisymmetric SH wave propagation in spherical geometry. Geophys J Int (in print)
59. Jastram C, Tessmer E (1994) Elastic modelling on a grid with vertically varying spacing. Geophys Prosp 42:357–370
60. Karypis G, Kumar V (1998) Multilevel k-way Partitioning Scheme for Irregular Graphs. J Parallel Distrib Comput 48(1):96–129
61. Kelly KR, Ward RW, Treitel S, Alford RM (1976) Synthetic seismograms: A finite-difference approach. Geophysics 41: 2–27
62. Kennett BLN (2002) The seismic wavefield, vol I + II. Cambridge University Press, Cambridge
63. Komatitsch D, Tromp J (2002) Spectral-element simulations of global seismic wave propagation, part I: Validation. Geophys J Int 149:390–412
64. Komatitsch D, Tromp J (2002) Spectral-element simulations of global seismic wave propagation, part II: 3-D models, oceans, rotation, and gravity. Geophys J Int 150:303–318
65. Komatitsch D, Vilotte JP (1998) The spectral-element method: An efficient tool to simulate the seismic response of 2D and 3D geological structures. Bull Seism Soc Am 88:368–392
66. Komatitsch D, Coutel F, Mora P (1996) Tensorial formulation of the wave equation for modelling curved interfaces. Geophys J Int 127(1):156–168
67. Kosloff D, Baysal E (1982) Forward modeling by a fourier method. Geophysics 47(10):1402–1412
68. Krishnan S, Ji C, Komatitsch D, Tromp J (2006) Case studies of damage to tall steel moment-frame buildings in Southern California during large San Andreas earthquakes. Bull Seismol Soc Am 96(4A):1523–1537
69. Krishnan S, Ji C, Komatitsch D, Tromp J (2006) Performance of two 18-story steel moment-frame buildings in Southern California during two large simulated San Andreas earthquakes. Earthq Spectra 22(4):1035–106
70. Käser M, Dumbser M (2006) An arbitrary high order discontinuous Galerkin method for elastic waves on unstructured meshes I: The two-dimensional isotropic case with external source terms. Geophys J Int 166:855–877
71. Käser M, Dumbser M, de la Puente J, Igel H (2007) An arbitrary high order discontinuous Galerkin method for elastic waves on unstructured meshes III: Viscoelastic attenuation. Geophys J Int 168(1):224–242
72. Käser M, Igel H (2001) Numerical simulation of 2D wave propagation on unstructured grids using explicit differential operators. Geophys Prospect 49(5):607–619
73. Käser M, Igel H, Sambridge M, Braun J (2001) A comparative study of explicit differential operators on arbitrary grids. J Comput Acoust 9(3):1111–1125
74. Kwak D-Y, Im Y-T (2002) Remeshing for metal forming simulations – part II: Three dimensional hexahedral mesh generation. Int J Numer Methods Eng 53:2501–2528
75. LeVeque RL (2002) Finite volume methods for hyperbolic problems. Cambridge University Press, Cambridge
76. Levander AR (1988) Fourth-order finite-difference P-SV seismograms. Geophysics 53:1425–1436
77. Madariaga R (1976) Dynamics of an expanding circular fault. Bull Seismol Soc Am 66(3):639–66
78. Magnier S-A, Mora P, Tarantola A (1994) Finite differences on minimal grids. Geophysics 59:1435–1443
79. Marfurt KJ (1984) Accuracy of finite-difference and finite-element modeling of the scalar and elastic wave equations. Geophysics 49:533–549
80. Mercerat ED, Vilotte JP, Sanchez-Sesma FJ (2006) Triangular spectral element simulation of two-dimensional elastic wave propagation using unstructured triangular grids. Geophys J Int 166(2):679–698
81. METIS glaros.dtc.umn.edu/gkhome/views/metis. Accessed 1 Jul 2008
82. Moczo P (1989) Finite-difference techniques for SH-waves in 2-D media using irregular grids – Application to the seismic response problem. Geophys J Int 99:321–329

83. Moczo P, Kristek J, Halada L (2000) 3D 4th-order staggered grid finite-difference schemes: Stability and grid dispersion. Bull Seism Soc Am 90:587–603
84. Montelli R, Nolet G, Dahlen FA, Masters G, Engdahl ER, Hung S (2004) Finite-frequency tomography reveals a variety of plumes in the mantle. Science 303(5656):338–343
85. Müller G (1977) Earth-flattening approximation for body waves derived from geometric ray theory – improvements, corrections and range of applicability. J Geophys 42:429–436
86. Nissen-Meyer T, Fournier A, Dahlen FA (2007) A 2-D spectral-element method for computing spherical-earth seismograms – I. Moment-tensor source. Geophys J Int 168:1067–1092
87. Ohminato T, Chouet BA (1997) A free-surface boundary condition for including 3D topography in the finite-difference method. Bull Seism Soc Am 87:494–515
88. Opršal I, J Zahradník (1999) Elastic finite-difference method for irregular grids. Geophysics 64:240–250
89. Pitarka A (1999) 3D elastic finite-difference modeling of seismic motion using staggered grids with nonuniform spacing. Bull Seism Soc Am 89:54–68
90. Priolo E, Carcione JM, Seriani G (1996) Numerical simulation of interface waves by high-order spectral modeling techniques. J Acoust Soc Am 95:681–693
91. Reed WH, Hill TR (1973) Triangular mesh methods for the neutron transport equation. Technical Report, LA-UR-73-479, Los Alamos Scientific Laboratory
92. Ripperger J, Igel H, Wassermann J (2004) Seismic wave simulation in the presence of real volcano topography. J Volcanol Geotherm Res 128:31–44
93. Scandella L (2007) Numerical evaluation of transient ground strains for the seismic response analyses of underground structures. Ph D Thesis, Milan University of Technology, Milan
94. SCEC www.scec.org. Accessed 1 Jul 2008
95. Schneiders R (2000) Octree-Based Hexahedral Mesh Generation. Int J Comput Geom Appl 10(4):383–398
96. Schwartzkopff T, Munz CD, Toro EF (2002) ADER: A high-order approach for linear hyperbolic systems in 2D. J Sci Comput 17:231–240
97. Schwartzkopff T, Dumbser M, Munz CD (2004) Fast high order ADER schemes for linear hyperbolic equations. J Comput Phys 197:532–539
98. Seriani G, Priolo E, Carcione JM, Padovani E (1992) High-order spectral element method for elastic wave modeling: 62nd Ann. Internat. Mtg., Soc. Expl. Geophys., Expanded Abstracts, 1285–1288
99. Shepherd JF (2007) Topologic and geometric constraint-based hexahedral mesh generation. Ph.D. Thesis on Computer Science, School of Computing The Universiy of Utah, Salt Lake City
100. Sieminski A, Liu Q, Trampert J, Tromp J (2007) Finite-frequency sensitivity of surface waves to anisotropy based upon adjoint methods. Geophys J Int 168:1153–1174
101. SPICE www.spice-rtn.org. Accessed 1 Jul 2008
102. Stupazzini M (2004) A spectral element approach for 3D dynamic soil-structure interaction problems. Ph D Thesis, Milan University of Technology, Milan
103. Takeuchi N, Geller RJ (2000) Optimally accurate second order time-domain finite difference scheme for computing synthetic seismograms in 2-D and 3-D media. Phys Earth Planet Int 119:99–131
104. Tape C, Liu Q, Tromp J (2007) Finite-frequency tomography using adjoint methods: Methodology and examples using membrane surface waves. Geophys J Int 168:1105–1129
105. Tarantola A (1986) A strategy for nonlinear elastic inversion of seismic reflection data. Geophysics 51(10):1893–1903
106. Tessmer E (2000) Seismic finite-difference modeling with spatially varying time teps. Geophysics 65:1290–1293
107. Tessmer K, Kosloff D (1996) 3-D elastic modeling with surface topography by a Chebychev spectral method. Geophysics 59:464–473
108. Tessmer E, Kessler D, Kosloff K, Behle A (1996) Multi-domain Chebyshev–Fourier method for the solution of the equations of motion of dynamic elasticity. J Comput Phys 100:355–363
109. Thomas C, Igel H, Weber M, Scherbaum F (2000) Acoustic simulation of P-wave propagation in a heterogeneous spherical earth: Numerical method and application to precursor energy to PKPdf. Geophys J Int 141:307–320
110. Thorne M, Lay T, Garnero E, Jahnke G, Igel H (2007) 3-D seismic imaging of the D″ region beneath the Cocos Plate. Geophys J Int 170:635–648
111. Titarev VA, Toro EF (2002) ADER: Arbitrary high order Godunov approach. J Sci Comput 17:609–618
112. Toro EF (1999) Riemann solvers and numerical methods for fluid dynamics. Springer, Berlin
113. Toro EF, Millington AC, Nejad LA (2001) Towards very high order Godunov schemes, in Godunov methods; Theory and applications. Kluwer/Plenum, Oxford, pp 907–940
114. Toyokuni G, Takenaka H, Wang Y, Kennett BLN (2005) Quasi-spherical approach for seismic wave modeling in a 2-D slice of a global earth model with lateral heterogeneity. Geophys Res Lett 32:L09305
115. Van der Hilst RD (2004) Changing views on Earth's deep mantle. Science 306:817–818
116. Virieux J (1984) SH-wave propagation in heterogeneous media: Velocity-stress inite-difference method. Geophysics 49:1933–1957
117. Virieux J (1986) P-SV wave propagation in heterogeneous media: Velocity-stress finite-difference method. Geophysics 51:889–901
118. Woodhouse JH, Dziewonski AM (1984) Mapping the upper mantle: Three dimensional modelling of earth structure by inversion of seismic waveforms. J Geophys Res 89:5953–5986
119. Yoshimura C, Bielak J, Hisada Y, Fernández A (2003) Domain reduction method for three-dimensional earthquake modeling in localized regions, part II: Verification and applications. Bull Seism Soc Am 93:825–841
120. Zambelli C (2006) Experimental and theoretical analysis of the mechanical behaviour of cohesionless soils under cyclic-dynamic loading. Ph D Thesis, Milan University of Technology, Milan
121. Zienckiewicz O, Taylor RL (1989) The finite element method, vol 1. McGraw-Hill, London

Books and Reviews

Carcione JM, Herman GC, ten Kroode APE (2002) Seismic modelling. Geophysics 67:1304–1325

Mozco P, Kristek J, Halada L (2004) The finite-difference method for seismologists: An introduction. Comenius University, Bratislava. Available in pdf format at ftp://ftp.nuquake.eu/pub/Papers

Moczo P, Kristek J, Galis M, Pazak P, Balazovjech M (2007) The finite difference and finite-element modelling of seismic wave propagation and earthquake motion. Acta Physica Slovaca, 57(2)177–406

Wu RS, Maupin V (eds) (2006) Advances in wave propagation in heterogeneous earth. In: Dmowska R (ed) Advances in geophysics, vol 48. Academic/Elsevier, London

Seismic Waves in Heterogeneous Earth, Scattering of

Haruo Sato
Department of Geophysics, Graduate School of Science, Tohoku University, Sendai-shi, Miyagi-ken, Japan

Article Outline

Glossary
Definition of the Subject
Introduction
Radiative Transfer Theory for a Scattering Medium
Wave Envelopes in Random Media and Statistical Characterization
Envelope Broadening of a High-Frequency Seismogram
Spatial Variation of Scattering Characteristics
Temporal Change in the Earth Medium Structure
Future Directions
Acknowledgments
Bibliography

Glossary

Attenuation factor Q^{-1} A measure of attenuation characteristics of a medium caused by intrinsic absorption and scattering loss. The former means the transfer of vibration energy into heat and the latter means the transfer of vibration energy from the direct wave to coda waves caused by scattering due to medium heterogeneity.

Coda waves Wave trains that follow the arrival of the direct S-wave phase are called S-coda waves or simply coda waves. Coda waves are interpreted as a superposition of S waves scattered by distributed heterogeneities. Wave trains between direct P and S wave arrivals are called P-coda waves.

Coda attenuation factor Q_C^{-1} This parameter characterizes the amplitude decay of S coda of a local earthquake with the lapse time increasing based on the S-to-S single scattering. The coda duration shortens for a larger coda attenuation factor.

Envelope broadening The source duration time of a microearthquake is short; however, the apparent duration time of the S-wave seismogram increases with the travel distance increasing because of diffraction and scattering by medium heterogeneities. This phenomenon is called envelope broadening.

Radiative transfer theory A phenomenological theory that describes scattering process of wave energy in a scattering medium on the basis of causality, geometrical spreading and the energy conservation. It neglects the interference of waves but focuses on the intensity only. This theory admits various types of scattering patterns. It is often applied to model the energy propagation of high-frequency seismic-waves in heterogeneous Earth media.

Random media A mathematical model for media whose parameters are described by random functions of space coordinates. The stochastic properties of the ensemble of random media are characterized by their autocorrelation function or the power spectral density function.

Scattering coefficient g A measure of the scattering power in a unit solid angle at a certain direction by a unit volume of heterogeneous media for the incidence of unit energy flux density. The average of g over the solid angle gives the total scattering coefficient g_0, of which the reciprocal gives the mean free path. This quantity characterizes the coda excitation and the scattering loss in the heterogeneous media.

Definition of the Subject

The structure of the solid Earth was extensively studied by using seismic waves such as travel time analysis based on Snell's law, dispersion analysis of surface waves, and spectral analysis of free oscillation, where the notion of a horizontally stratified structure or a spherical shell structure prevailed among the geophysical community. This means the acceptance of the dominance of gravity in geodynamic process. Velocity tomography revealed that the solid Earth structure is three-dimensionally inhomogeneous with various ranges of scales; however, the resolution of velocity tomography is much coarser than the wavelength of seismic waves. In 1970s, the existence of distributed inhomogeneities having the order of the wavelength of seismic waves was recognized from the observation of coda waves of local earthquakes, which are long-lasting wave trains following the direct S-wave arrival in high-frequency seismograms. Here, we use "high-frequency" for frequency higher than about 1 Hz. The long duration time of coda waves can be interpreted as a direct evidence

of wide-angle scattering caused by distributed small-scale heterogeneities since the source duration time is generally very short. S-wave seismograms of microearthquakes show broadened envelopes with travel distance increasing. This envelope broadening phenomenon is also an evidence of scattering around the forward direction due to random velocity inhomogeneities.

Since then, focusing on the frequency dependence of seismogram envelopes, geophysicists have extensively studied the scattering process of high-frequency seismic waves in relation to the spectral structure of velocity inhomogeneities, where the statistical characterization of the medium inhomogeneity is inevitable. The radiative transfer theory and the stochastic Markov approximation have been developed as mathematical tools for the analyzes of seismogram envelopes. The strength of scattering and/or the spectral structure of random inhomogeneity have been measured in various regions of the solid Earth. The scattering approach is found to be also useful for detecting temporal changes in the crustal medium associated with earthquake occurrences. Thus, scattering of high-frequency seismic waves in the heterogeneous Earth medium is important for understanding the physical structure and the geodynamic process reflecting the evolution of the solid Earth.

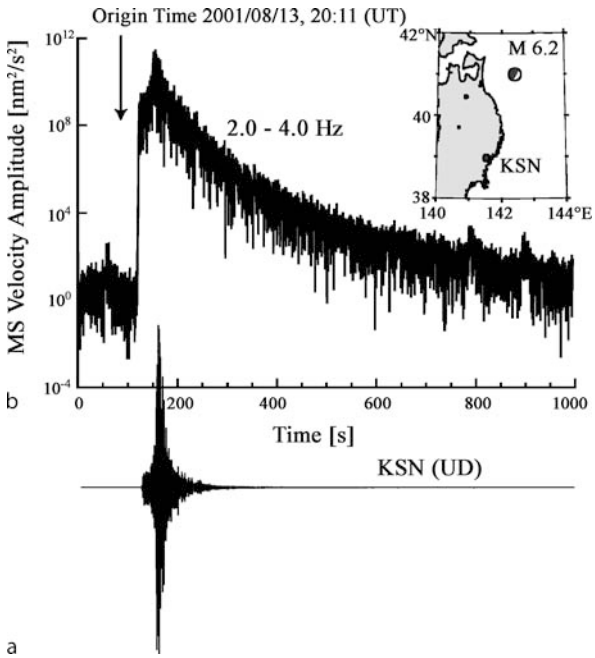

Seismic Waves in Heterogeneous Earth, Scattering of, Figure 1
a Seismogram of a local earthquake of M 6.2 in northeastern Honshu, Japan recorded by F-net, NIED. b Bandpass-filtered MS envelope (Courtesy of T. Maeda)

Introduction

Coda Waves

The high-frequency seismogram of a local earthquake has a long tail after the direct S-coda arrival. The tail portion of seismogram is called "S-coda waves" or simply "coda". As an example, Fig. 1a and b show the raw seismogram and the band-pass filtered mean square (MS) trace of an earthquake of magnitude (M) 6.1, respectively. We note that the mean square wave envelope, which is the running mean of the squared trace with characteristic time of a few times the center period, is proportional to the time trace of the wave energy density. The coda wave oscillation lasts more than several hundreds of seconds. The duration of coda waves measured from the P-wave onset until when the coda amplitude decreases to the microseism's level has been used as a quick measure of the earthquake magnitude from a single station observation since the 1960s. Having a motivation to extract the source spectrum of a large earthquake from clipped seismograms, Aki [1] first studied the characteristics of coda waves as scattered waves. Coda envelopes of a local earthquake have a smoothly decaying common curve with lapse time increasing irrespective of epicentral distances and the source radiation pattern. Aki and Chouet [4] interpreted coda waves as single back-scattered S-waves due to heterogeneities randomly distributed in the lithosphere. Their model based on the radar equation for the same location of a source and a receiver can be written as follows.

Point-like isotropic scatterers characterized by total scattering cross-section σ_0 are randomly and homogeneously distributed with number density n in a 3-D medium with background wave velocity V_0. The scattering power per unit volume is characterized by the total scattering coefficient $g_0 = n\sigma_0$, of which the reciprocal gives the mean free path. When the total wave energy W is impulsively radiated from a point source at time $t = 0$, the wave energy density of singly back-scattered waves at the source location is written as

$$E^{\mathrm{SB}}(t) \approx \frac{W g_0}{2\pi V_0^2 t^2} \mathrm{e}^{-Q_{\mathrm{C}}^{-1} 2\pi f t} \tag{1}$$

since the interference of scattered waves can be neglected because of the random distribution of scatterers. The inverse square of lapse time means geometrical spreading in a 3-D space. Here, an exponential damping term with coda attenuation factor Q_{C}^{-1} is introduced to represent phenomenological attenuation effect. This simple formula has been widely used for measurements of g_0 and Q_{C}^{-1} for S-waves in the world since the 1980s. Reported g_0-values

are of the order of 0.01 km^{-1} for 1–30 Hz and Q_C^{-1} values are about 10^{-2} at 1 Hz and decrease to about 10^{-3} at 20 Hz in the lithosphere [74].

Envelope Broadening of S-Seismogram

There is another evidence of scattering due to random inhomogeneity in high-frequency seismograms. Observed S-seismograms of a microearthquake have broadened envelopes around their peaks after the direct arrivals. As shown by an example in Fig. 2, the apparent duration time of the S-seismogram just after the direct S-arrival increases with increasing travel distance. It is more than ten seconds at distances larger than 100 km, where the source duration time is less than one second for an earthquake of M 4.0. Sato [70] called this phenomenon observed in an island arc as "envelope broadening", and Atkinson [6] reported similar phenomenon in a continent. For P-waves of teleseismic events, broadening of the vertical component envelope [35] and the excitation in the transverse component amplitude [52] have been used as a measure of lithospheric heterogeneity. These phenomena can be interpreted by multiple scattering within a narrow angle around the global ray direction due to random velocity inhomogeneities. When the wavelength is much shorter than the characteristic scale of the random velocity inhomogeneity, the scattering process of waves can be represented by successive ray bending processes, where scattering angles are statistically controlled by the spectrum of random velocity inhomogeneity. At a given distance from the source, a small number of rays with large scattering angles arrive long after the direct ray.

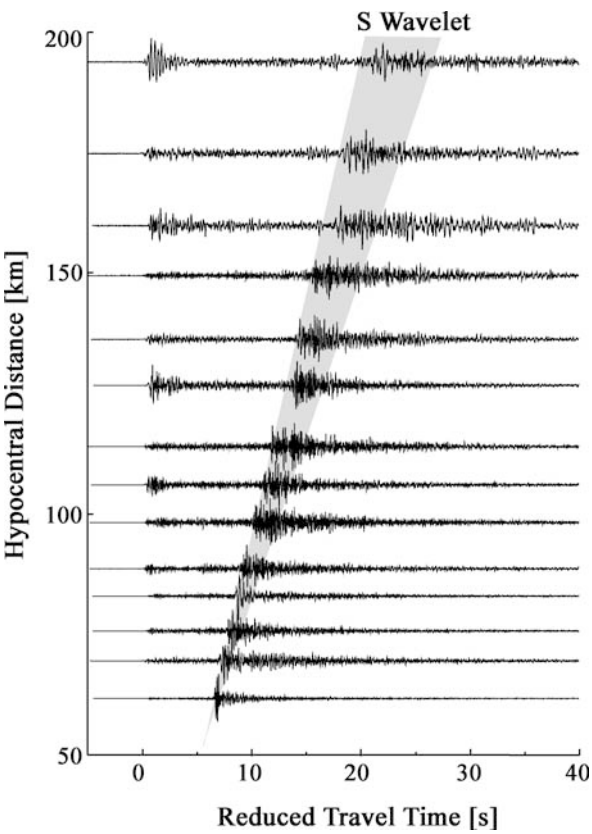

Seismic Waves in Heterogeneous Earth, Scattering of, Figure 2
Envelope broadening shown in horizontal component seismograms of a microearthquake with M 4.0 in Japan recorded by Hi-net, NIED, where the abscissa is reduced travel time with move-out velocity 7 km/s (Courtesy of T. Takahashi)

Radiative Transfer Theory for a Scattering Medium

Radiative Transfer Integral Equation for the Isotropic Scattering Process

Disregarding wave interference and focusing on wave power, the radiative transfer theory [8] treats the propagation of wave energy in a scattering medium. Wu [85] first introduced the radiative transfer theory for the stationary state in the synthesis of seismogram envelopes. The non-stationary multiple isotropic scattering process in 1-D was solved by Hemmer [24] and that in 2-D was solved by Shang and Gao [77]. Later, Zeng et al. [92] formulated the time-dependent multiple isotropic scattering process in 3-D as an extension of the single backscattering model [4] as follows.

In a 3-D scattering medium characterized by background velocity V_0 and total scattering coefficient g_0, when the total wave energy W is impulsively radiated isotropically from a source at the origin, the multiple isotropic scattering process is written by the following integral equation for energy density:

$$E(\mathbf{x}, t) = W G_E(\mathbf{x}, t) + g_0 V_0 \cdot \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G_E(\mathbf{x}-\mathbf{x}', t-t')\, E(\mathbf{x}', t')\, d\mathbf{x}' dt' , \quad (2)$$

where the convolution integral in the second term means the propagation of energy from the last scattering point $\mathbf{x}'$ to a receiver at $\mathbf{x}$. The first term is the ballistic term that means the direct propagation of energy from the source with scattering loss, $G_E(\mathbf{x}, t) = \delta(t - r/V_0)\exp(-g_0 V_0 t)/(4\pi V_0 r^2)$, where $r = |\mathbf{x}|$. The solution

is written as [92]

$$E(\mathbf{x},t) = WG_{\mathrm{E}}(\mathbf{x},t) + \frac{Wg_0 e^{-g_0 V_0 t}}{4\pi r^2}\frac{r}{V_0 t} \cdot \ln\left[\frac{V_0 t + r}{V_0 t - r}\right] H\left(t - \frac{r}{V_0}\right) + Wg_0^2 V_0^2 \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d\omega dk\, e^{-i\omega t - ikr} \cdot \frac{ik}{2\pi r}\frac{\bar{\bar{G}}_{\mathrm{E}}(-k,-i\omega)^3}{1 - g_0 V_0 \bar{\bar{G}}_{\mathrm{E}}(-k,-i\omega)}, \quad (3)$$

where $\bar{\bar{G}}_{\mathrm{E}}(k,s) = (1/kV_0)\tan^{-1} kV_0/(s + g_0 V_0)$ is the Fourier–Laplace transform of G_{E} with respect to coordinate and time, respectively. The second term represents the single scattering process (see Sato [67]), which has a logarithmic divergence at the direct arrival $t = r/V_0$ and decreases according to the inverse square of lapse time at long lapse times as $Wg_0/(2\pi V_0^2 t^2)$. The third term representing multiple scattering converges to a diffusion solution with lapse time increasing as

$$E^{\mathrm{Dif.}}(\mathbf{x},t) = W\left(\frac{3g_0}{4\pi V_0 t}\right)^{3/2} e^{-\frac{3g_0 r^2}{4V_0 t}} H(t), \quad (4)$$

where the factor $V_0/(3g_0)$ is the diffusivity. Later, Paaschens [54] proposed an approximation as

$$E^{\mathrm{Paa.}}(\mathbf{x},t) \approx WG_{\mathrm{E}}(\mathbf{x},t) + \frac{We^{-g_0 V_0 t}}{\left(4\pi V_0 t/(3g_0)\right)^{3/2}} \cdot \left(1 - \frac{r^2}{V_0^2 t^2}\right)^{\frac{1}{8}} M\left(g_0 V_0 t\left(1 - \frac{r^2}{V_0^2 t^2}\right)^{\frac{3}{4}}\right) \cdot H\left(t - \frac{r}{V_0}\right), \quad (5)$$

where $M(x) \approx e^x\sqrt{1 + 2.026/x}$. The error of this approximation is of the order of 2% outside the ballistic peak and its tail for $g_0 V_0 t < 6$ and $2 < g_0 r < 4$.

Figure 3 shows spatiotemporal variations in energy density in a scattering medium theoretically predicted by the approximation solution (5) for instantaneous spherical source radiation at the origin. Scattered energy density is shown by a black curve, where the ballistic term is shown by a vertical gray line. At a small distance from the source compared with the mean free path $1/g_0$, the energy density decreases rapidly after the direct arrival as predicted by the single scattering term; however, the decay rate becomes smaller due to multiple scattering according to the power of lapse time $t^{-3/2}$ at long lapse times.

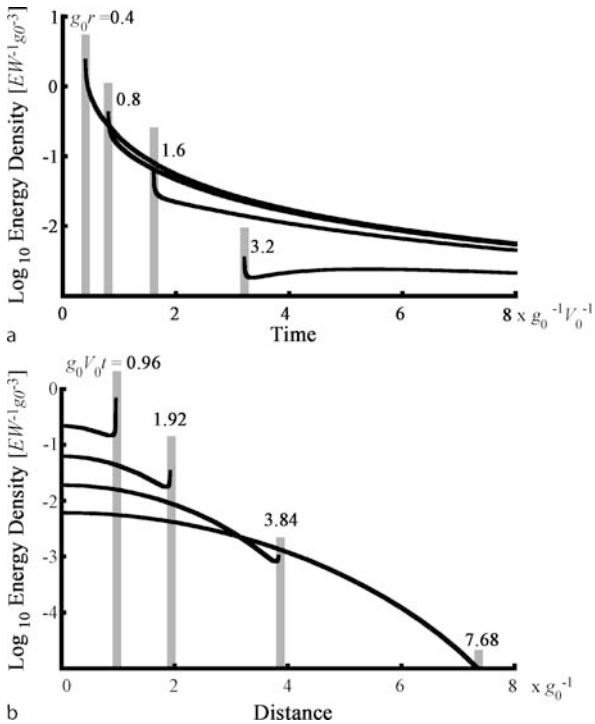

Seismic Waves in Heterogeneous Earth, Scattering of, Figure 3
a Temporal change and b spatial variation of energy density in an isotropic-scattering medium for a point source radiation. Each *black curve* shows the scattering contribution predicted by the Paaschens approximation and each *vertical gray line* shows a ballistic term

At a long distance, for example at $r = 3.2/g_0$, the energy density has an additional diffusion peak. The spatial distribution of scattered energy density is uniform around the source at a short lapse time compared with the mean free time $1/g_0 V_0$; however, it converges to a Gaussian curve at a long lapse time as theoretically predicted by the diffusion solution (4), for example at $t = 7.48/g_0 V_0$. There is no violation of causality since no signal exists beyond the ballistic peak. The smooth spatial distribution of scattered energy density around the source location gives the physical basis of the coda normalization method for measurements of S-wave attenuation and site amplification factors (e. g. [3,56,89]).

Gusev and Abubakirov [21] and Hoshiba [27] numerically solved the radiative transfer equation for the isotropic scattering process by using the Monte Carlo method. Yoshimoto [88] numerically simulated envelopes in scattering media of which the background velocity decreases with depth. He found the concentration of scattered energy near the surface because of seismic ray bending.

Nonisotropic radiation from a source can be easily introduced in the radiative transfer equation. Sato et al. [75] analytically solved the case of double couple source radiation: the energy density theoretically predicted faithfully reflects the source radiation pattern near the direct arrival; however, the azimuthal dependence diminishes with increasing lapse time. It qualitatively agrees with the observed radiation pattern independence of coda amplitudes at long lapse times. Their solution has been used in the envelope inversion of strong motion records for the spatial distribution of high-frequency wave energy radiation from an earthquake fault (e. g. [47]).

Measurements of Total Scattering Coefficient and Attenuation

For the practical application of the radiative transfer theory to observed seismograms, it is necessary to introduce intrinsic absorption Q_{Int}^{-1} by multiplying an exponential temporal decay factor $\exp[-Q_{\mathrm{Int}}^{-1} 2\pi f t]$ to the resultant energy density. By using the solution (3) of the radiative transfer theory for the isotropic scattering model, total scattering coefficient g_0 and intrinsic absorption factor $Q_{\mathrm{Int.}}^{-1}$ of the S-wave have been measured. Reported g_0 values in the lithosphere are of the order of $10^{-2}\,\mathrm{km}^{-1}$ for frequencies from 1 to 30 Hz as plotted in Fig. 4.

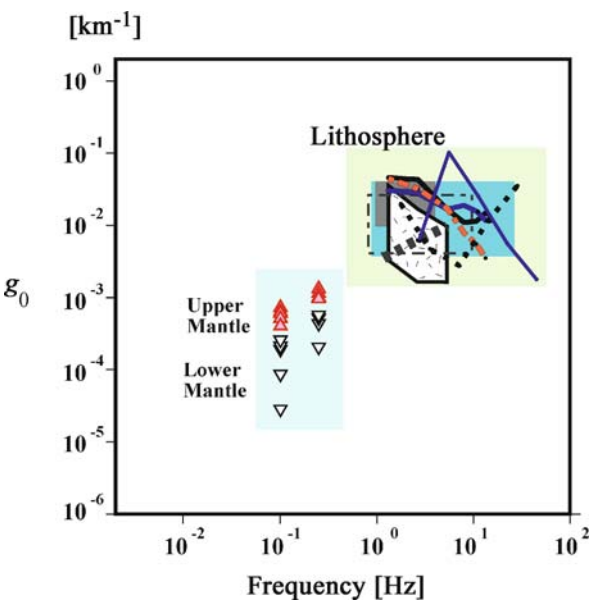

Seismic Waves in Heterogeneous Earth, Scattering of, Figure 4
Total scattering coefficient of S-waves in the Earth. Measurements in the mantle [40] are added to lithospheric inhomogeneity [74]

From the observed lapse time dependence of Q_{C}^{-1}, Gusev [20] quantitatively explained the decrease of g_0 with depth. Lee et al. [40] analyzed coda envelopes of regional earthquakes before and after the ScS arrival around 900s in lapse time from the origin time based on the numerically simulated envelopes for the PREM model, which is characterized by depth-dependent background velocity and total attenuation for S-waves. They reported lower g_0 values in 4s and 10s period bands in the upper and lower mantle compared with those in the lithosphere as illustrated in Fig. 4.

Hoshiba et al. [28] and Fehler et al. [10] developed a method to measure simultaneously g_0 and Q_{Int}^{-1} values for the S-wave from the whole S-envelope analysis based on the synthetic envelope derived from the radiative transfer theory. Their multiple lapse-time window analysis method has been widely used in the world. Estimated scattering loss $g_0 V_0/\omega$ decreases with frequency; however, intrinsic absorption Q_{Int}^{-1} is rather insensitive to frequency. Estimated seismic albedo $B \equiv g_0 V_0 / \left(\omega Q_{\mathrm{Int}}^{-1} + g_0 V_0\right)$, the ratio of scattering loss to the total attenuation, of S-waves in the lithosphere widely distribute from 0.2 to 0.8 for 1–6 Hz, but they are limited between 0.2 and 0.5 for 6–20 Hz.

Total scattering cross-section and number density of scatterers appear jointly as the total scattering coefficient in theoretical models; however, Matsumoto [44] proposed a method to separate them from the temporal variation of the semblance coefficient of coda waves recorded by a seismic array. Analyzing data obtained in the aftershock area of the 2000 western Tottori earthquake, Japan, he estimated $n = 0.03\,\mathrm{km}^{-3}$ and $g_0 = 0.001\,\mathrm{km}^{-1}$ at 20 Hz.

It should be noted that lunar seismograms have coda durations exceeding one hour. Applying the diffusion model for the explanation of spindle-like envelopes of lunar-quakes, Dainty and Toksöz [9] estimated g_0 to be as large as 0.05–$0.5\,\mathrm{km}^{-1}$ at 0.45 Hz.

Wave Envelopes in Random Media and Statistical Characterization

Statistical Characterization of Random Media

As revealed from tomography analyses (e. g. [93]), velocity structure is three-dimensionally inhomogeneous especially in the lithosphere. Well log data show typical samples of the shallow crust. Figure 5a shows well log data of P- and S-wave velocities and mass density obtained in Kyushu, Japan [80], where wave velocities and mass density are measured in a borehole by using ultrasonic waves and gamma rays, respectively. These well log data clearly show random fluctuation with short wavelengths.

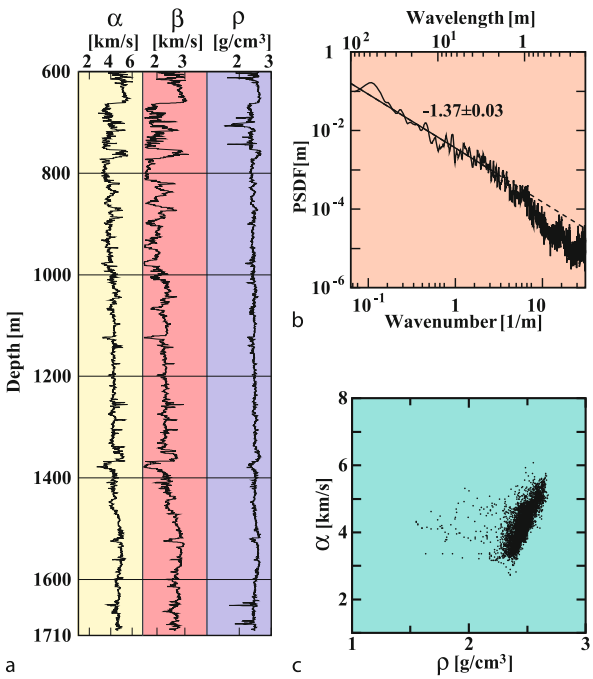

Seismic Waves in Heterogeneous Earth, Scattering of, Figure 5
a Well log data at the YT-2 site, Kyushu, Japan. b Power spectral density function of the fractional fluctuation of P-wave velocity log. c Scattergram of P-wave velocity and mass density. Reproduced from [80]

As a natural consequence, we imagine random inhomogeneities widely distributed in the solid Earth medium.

By using computer power, wave propagation in random media has been numerically studied extensively. Using finite difference simulations of waves in random media, Frankel and Clayton [14] first examined the relation between coda excitation and scattering loss and the spectrum of random inhomogeneity. On the basis of numerical simulations, Frankel and Wennerberg [15] proposed the energy flux model that has a uniform distribution of scattered energy behind the direct waves for the analysis of high-frequency seismogram envelopes. Using a boundary integral method for the simulation of waves in a medium containing many cavities, Yomogida and Benites [87] examined a relation between coda attenuation and the distribution of cavities.

There is an alternative approach to treat statistically randomly inhomogeneous media. The wave-velocity is written as $V(\mathbf{x}) = V_0\{1 + \xi(\mathbf{x})\}$, where V_0 is the average velocity and fractional fluctuation $\xi(\mathbf{x})$ is a homogeneous and isotropic random function of space coordinate $\mathbf{x}$. We imagine an ensemble of random media $\{\xi\}$, which is statistically characterized by the autocorrelation function (ACF) $R(\mathbf{x}) \equiv \langle \xi(\mathbf{x}+\mathbf{x}')\,\xi(\mathbf{x}')\rangle$, where angular brackets mean the ensemble average. The MS fractional fluctuation $\varepsilon^2 \equiv R(0)$ and the correlation distance a are key parameters. The Fourier transform of ACF gives the power spectral density function (PSDF) P. The PSDF of the P-wave velocity fractional fluctuation of well log data shows a power-law characteristic at large wavenumbers as illustrated in Fig. 5b. P-wave velocity and mass density show a good correlation as shown in Fig. 5c. The statistical view is useful for representing geological data, too (e. g. [18,26]).

Scattering Coefficient Based on the Born Approximation

When the medium inhomogeneity is small $|\xi| \ll 1$, scalar wave ϕ is governed by the following wave equation:

$$\left(\Delta - \frac{1}{V_0^2}\partial_t^2\right)\phi + \frac{2}{V_0^2}\xi(\mathbf{x})\partial_t^2\phi = 0\,. \tag{6}$$

Velocity inhomogeneity is supposed to localize in a volume around the origin, of which the dimension is chosen to be much larger than a. For the incidence of a plane wave of unit amplitude at angular frequency ω as $e^{i(k_0\mathbf{e_z}\mathbf{x}-\omega t)}$, where $\mathbf{e_z}$ is the unit vector to the z direction, we calculate the spherically outgoing scattered waves due to a localized inhomogeneity by using the Born approximation as $\phi^1(\mathbf{x},t) = -k_0^2 e^{i(k_0 r-\omega t)}\tilde{\xi}(k_0\mathbf{e_r} - k_0\mathbf{e_z})/(2\pi r)$, where the tilde means the Fourier transform with respect to coordinates in 3-D space and $\mathbf{e_r}$ is a radial unit vector (e. g. [74]). According to Aki and Chouet [4], the scattering coefficient defined as the scattering power in a unit solid angle at certain direction by a unit volume of random inhomogeneous media for the incidence of unit energy flux density is statistically written by using its PSDF as

$$g(\psi;\omega) = \frac{k_0^4}{\pi}P(k_0\mathbf{e_r} - k_0\mathbf{e}_3) = \frac{k_0^4}{\pi}P\left(2k_0\sin\frac{\psi}{2}\right), \tag{7}$$

where ψ is the scattering angle measured from the z direction. This functional form means anisotropic scattering depending on frequency. In general, scattering near around the forward direction becomes larger with increasing wavenumber in random media.

Radiative Transfer Theory with Scattering Coefficients Calculated by the Born Approximation

Extending the above formulation to vector wave propagation in random elastic media, we can define scattering coefficients for different scattering modes as PP, PS, SP,

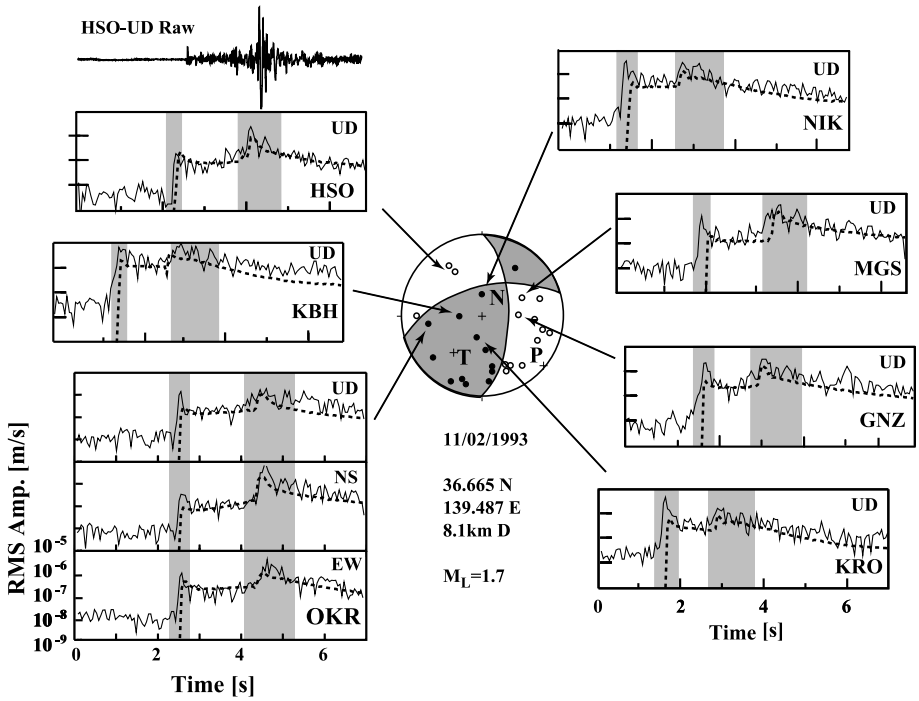

Seismic Waves in Heterogeneous Earth, Scattering of, Figure 6
RMS envelopes of a microearthquake in the shallow crust, Nikko in northern Kanto, Japan. *Fine curves* and *broken curves* are observed and best-fit synthesized envelopes, respectively, where *shades* show time windows used for the estimation of the source radiation energy. The trace on the *top left* shows a raw seismogram. Reproduced from [90]

and SS. For the case of random elastic media characterized by an exponential ACF with $\varepsilon = 10\%$ and $a = 2\,\text{km}$, Sato [68] synthesized three-component seismogram envelopes of a microearthquake of M 3 as a superposition of polarized scattered waves' power at a finite distance from a point shear dislocation source. That is the single scattering approximation of the radiative transfer theory with scattering coefficients calculated by the Born approximation. The SS scattering mode dominates in S coda, and pseudo P and S waves are produced even at a receiver on the null direction of the source radiation. By using a von Kármán-type ACF for describing random elastic media, Sato [71] estimated parameters $\kappa = 0.35, \varepsilon = 8.4\%$ and $a = 2.1\,\text{km}$ from observed frequency dependence of S-wave attenuation and g_0, where the parameter κ controls the role-off of PSDF at large wavenumbers. Extending the above vector wave envelope synthesis to include mode conversions at the free surface, Yoshimoto et al. [90] analyzed three-component seismogram envelopes of microearthquakes in the shallow crust in Nikko, northern Kanto, Japan. A raw seismogram is shown at the top-left of Fig. 6 as an example. Fine curves in Fig. 6 are logarithmic plots of observed root mean square (RMS) envelopes in the 2–16 Hz band. Broken curves are best-fit theoretical envelopes for random elastic media characterized by an exponential ACF with $\varepsilon = 5.7\%$ and $a = 400\,\text{m}$. We find that the fitness is good not only for S coda but also for P coda.

Wave theory in random media predicts that the scattering coefficient has a large lobe in the forward direction in higher frequencies. Gusev and Abubakirov [22] used the Monte Carlo method to simulate envelopes for the multiple nonisotropic scattering process. There have been mathematical developments to derive the radiative transfer equation for multiple nonisotropic scattering from the stochastic averaging of the wave equation in random media (e. g. [13,29,41,64]). Przybilla et al. [58] showed an excellent coincidence of vector envelopes calculated from finite difference simulation in 2-D random elastic media and those synthesized by the radiative transfer theory with scattering amplitudes derived from the Born approximation and the wandering effect of travel time.

Interference of Scattered Waves

The interference effect is neglected in conventional studies of wave scattering in random media; however, it becomes important for a specific case even in random media. When

randomness is strong enough to produce multiple scattering, coda wave intensity at a receiver near a source is enhanced compared to the prediction of conventional radiative transfer theory. Margerin et al. [42] showed that a spot of backscattering enhancement stabilizes in a sphere of radius half a wavelength centered at the source after a transient regime. The enhancement persists in time and should be observable as long as a coda is measurable. From field experiment of seismic waves in a shallow volcanic structure Larose et al. [39] reported the existence of weak localization, where the size of enhancement spot was one wavelength and the estimated mean-free path was 200 m for seismic waves around 20 Hz.

Envelope Broadening of a High-Frequency Seismogram

Markov Approximation for Parabolic Wave Equation

For the study of light propagation through the upper atmosphere and/or acoustic sound propagation through internal waves in oceans, various stochastic methods have been developed in the fields of physics. One of the most attractive methods for explaining the wave envelope around the direct arrival is the Markov approximation for the parabolic wave equation, which is an extension of the phase screen method or the split step Fourier method (e. g. [29,63]). This method is found to be applicable to seismogram envelopes. We imagine an elastic medium composed of a homogeneous half space $z < 0$ and an inhomogeneous half space $z > 0$, where the inhomogeneity is supposed to be small ($\varepsilon^2 \ll 1$) and the randomness is statistically homogeneous and isotropic. When the wavelength is smaller than the correlation distance a, we may neglect conversion scattering between P and S waves. Then, we can describe the principal characteristics of vector wave propagation by using potentials. For the incidence of plane P-wavelet to the z direction from the homogeneous zone, scalar potential is written as a superposition of harmonic waves of angular frequency ω as $\phi = \int_{-\infty}^{\infty} (2\pi i k_0)^{-1} U(\mathbf{x}_\perp, z, \omega)\, \mathrm{e}^{i k_0 z - i\omega t}\, \mathrm{d}\omega$ for $z > 0$, where $\mathbf{x}_\perp = (x, y)$ on the transverse plane. Neglecting the second derivative with respect to z, we have the parabolic-type equation for U as

$$2ik_0 \partial_z U + \left(\partial_x^2 + \partial_y^2\right) U - 2k_0^2 \xi(\mathbf{x})\, U = 0\,. \tag{8}$$

We define the two-frequency mutual coherence function (TFMCF) of field U between two different locations on the transverse plane at a distance z and different angular frequencies at ω' and ω'' as $\Gamma_2(\mathbf{x}_{\perp c}, \mathbf{x}_{\perp d}, z, \omega_c, \omega_d) \equiv \langle U(\mathbf{x}'_\perp, z, \omega') U(\mathbf{x}''_\perp, z, \omega'')^* \rangle$, where ω_c and ω_d are center-of-mass and difference angular frequencies, respectively. In the case of quasi-monochromatic waves $|\omega_d| \ll |\omega_c|$, using causality and neglecting back scattering, we derive the master equation for TFMCF. This derivation is called the Markov approximation. For the ith component, the intensity is defined as the ensemble average of the square of displacement $\langle \partial_i \phi\, \partial_i \phi^* \rangle = 1/(2\pi) \int_{-\infty}^{\infty} \widehat{I}_i^{\mathrm{P}}\, \mathrm{d}\omega_c$. The integrand is the intensity spectral density (ISD) $\widehat{I}_i^{\mathrm{P}}$, which means the time trace of MS amplitude in a band having the central angular frequency ω_c.

Vector Envelopes for a Gaussian ACF

The case of a Gaussian ACF $R(\mathbf{x}) = \mathrm{Exp}\left(-r^2/a^2\right)$ is mathematically tractable. For the initial condition $\widehat{I}_x^{\mathrm{P}} = \widehat{I}_y^{\mathrm{P}} = 0$ and $\widehat{I}_z^{\mathrm{P}} = \delta(t - z/V_0)$ at $z = 0$, ISDs are analytically written as [72]

$$\begin{aligned} \widehat{I}_{x0}^{\mathrm{P}}(z, t, \omega_c) &= \widehat{I}_{y0}^{\mathrm{P}}(z, t, \omega_c) \\ &= 2(V_0/z)(t - z/V_0) \cdot \widehat{I}_0^{\mathrm{R}}(z, t, \omega_c) \\ \widehat{I}_{z0}^{\mathrm{P}}(z, t, \omega_c) &= [1 - 4(V_0/z)(t - z/V_0)] \\ &\quad \cdot \widehat{I}_0^{\mathrm{R}}(z, t, \omega_c)\,, \end{aligned} \tag{9}$$

where subscript "0" means ISD without the wandering effect. The reference ISD is a solution for scalar waves for the initial condition $\Gamma_2(\mathbf{x}_\perp, z = 0) = 1$ [82]:

$$\widehat{I}_0^{\mathrm{R}}(z, t, \omega_c) = \frac{1}{t_{\mathrm{M}}} \frac{\pi}{8} \vartheta_1{}'\left(0, \mathrm{e}^{-\frac{\pi^2}{4}\frac{(t - z/V_0)}{t_{\mathrm{M}}}}\right) \cdot H\left(t - \frac{z}{V_0}\right)\,, \tag{10}$$

where $t_M = \sqrt{\pi}\varepsilon^2 z^2/(2V_0 a)$ is the characteristic time and function ϑ_1' is the derivative of the elliptic theta function of the first kind. Function $\widehat{I}_0^{\mathrm{R}}$ shows a broadened envelope having a delayed peak and a smoothly decaying tail as illustrated by a chained curve in Fig. 7, where solid and broken curves show three-component ISDs $\widehat{I}_{0z}^{\mathrm{P}}$ and $\widehat{I}_{x0}^{\mathrm{P}} (= \widehat{I}_{y0}^{\mathrm{P}})$, respectively, for $V_0 t_{\mathrm{M}}/z = 0.05$ as an example. All the three component envelopes have broadened traces; however, the peak height of the transverse component is smaller than that of the longitudinal component and the peak delay of the transverse component is larger than that of the longitudinal component. When $\varepsilon^2 z/a \ll 1$, the peak height of $\widehat{I}_{z0}^{\mathrm{P}}$ approximately decays according to the square of travel distance and the peak ratio of the trans-

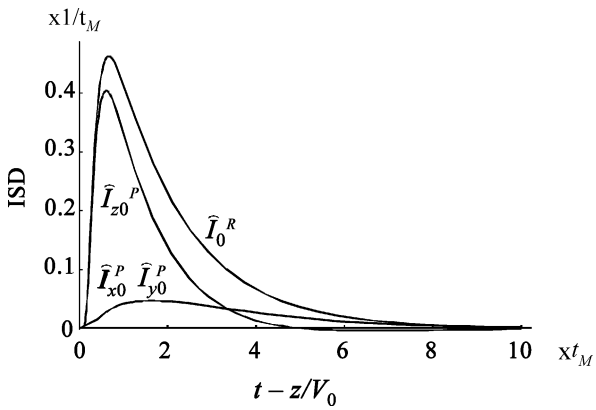

Seismic Waves in Heterogeneous Earth, Scattering of, Figure 7
Chained curve shows the reference ISD without the wandering effect in 3-D random elastic media characterized by a Gaussian ACF for the incidence of a plane P-wavelet. *Solid* and *broken curves* show three-component ISDs without the wandering effect for $V_0 t_M/z = 0.05$. Reproduced from [72]

verse component to longitudinal component is proportional to $\varepsilon^2 z/a$. ISDs $\widehat{I}_x^{\mathrm{P}}$, $\widehat{I}_y^{\mathrm{P}}$, and $\widehat{I}_z^{\mathrm{P}}$ can be calculated by using the convolution of (9) with the travel-time wandering effect $\exp\left[-(V_0 t - z_0)^2/2\sqrt{\pi}\varepsilon^2 az\right] V_0/\sqrt{2\pi\sqrt{\pi}\varepsilon^2 az}$ in time domain. For 2-D cases, the validity of the Markov approximation was numerically confirmed by a comparison with the finite difference simulations [11,36].

The above synthesis can be simply extended to S wave envelopes. Extension from plane wave incidence to impulsive radiation from a point source is also possible [73]. Figure 8 shows simulated three-component RMS envelopes along the z axis for a point source radiation, where random media are characterized by average P and S wave velocities, 6 km/s and 3.46 km/s, respectively, and a Gaussian ACF with $\varepsilon = 5\%$ and $a = 5$ km. We assume that P-wavelet radiation is isotropic and S-wavelet radiation is axially symmetric around the y axis with polarization to the x axis, where the ratio of S to P-wave source energy is chosen to be 23.3. Envelope broadening is common to both P and S waves in the syntheses. Excitation of the transverse component for P-waves and that of the radial component for S-waves are prominent. The appearance of scattered S-waves having long duration in synthesized envelopes at large travel distances qualitatively well explains observed characteristics shown in Fig. 2.

Randomness in the Lithosphere

Applying the envelope broadening model to S-wave seismograms recorded in Kanto, Japan, Sato [70] estimated the ratio $\varepsilon^2/a \approx 10^{-3}\,\mathrm{km}^{-1}$ with the assumption of a Gaussian ACF and S-wave attenuation $Q^{-1} = 0.014 f^{-1}$.

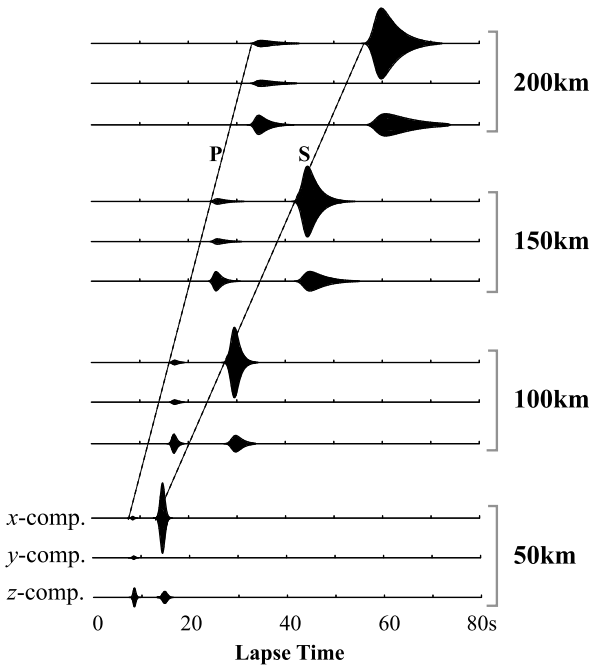

Seismic Waves in Heterogeneous Earth, Scattering of, Figure 8
Synthesized vector envelopes in random media characterized by a Gaussian ACF for radiation of P wavelet and S wavelet with a polarization to the *x*-axis from a point source. Reproduced from [73]

Saito et al. [66] studied the case of a von Kármán-type random media having a power-law spectrum at large wavenumbers, which are more appropriate for the real Earth inhomogeneity. The resultant envelope shows frequency dependence, which is controlled by the roll-off of the PSDF. Analyzing the hypocentral-distance dependence and frequency dependence of S-wave seismogram envelopes in northern Honshu, Japan for 2–32 Hz, Saito et al. [66] estimated parameters of von Kármán-type ACF as $\kappa = 0.6$ and $\varepsilon^{2.2}/a \approx 10^{-3.6}\,\mathrm{km}^{-1}$ with $Q^{-1} = 0.009 f^{-1}$. It means the PSDF decreases as wavenumber to the power of −4.2. Petukhin and Gusev [55] averaged S-wave seismogram envelopes of small earthquakes recorded in Kamchatka and compared the shapes with those numerically calculated for various types of random media. They concluded that random media whose PSDF decreases as the wavenumber to the power of −3.5 to −4 are appropriate.

Spatial Variation of Scattering Characteristics

Scattering Coefficient and Active Faults

Precisely examining coda envelopes of local earthquakes against lapse time measured from the origin time, we find

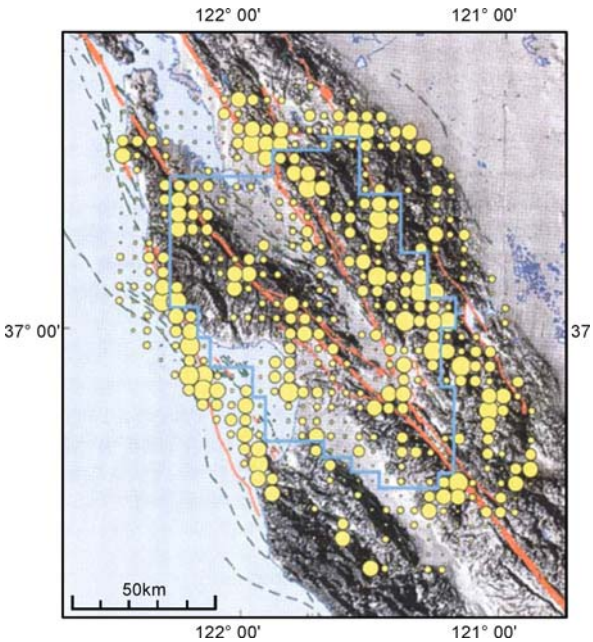

Seismic Waves in Heterogeneous Earth, Scattering of, Figure 9
Distribution of relative scattering coefficient at a depth of 0–5 km in central California revealed from the coda envelope inversion. *Circles* with larger diameter indicate stronger scattering and *solid lines* represent active faults. Reproduced with permission from [49]

temporal fluctuations around the smoothly decaying master curve. We may interpret that swellings and dips around the master curve are caused by stronger and weaker scatterers, respectively, distributed in the subsurface. By using a single isotropic scattering model, Nishigami [48] proposed an inversion scheme from coda envelopes of local earthquakes recorded at multiple stations for estimating the spatial variation of the scattering coefficient. Applying this inversion scheme to coda records obtained in central California, Nishigami [49] mapped the distribution of relative scattering coefficient in the shallow crust as in Fig. 9. A good correlation is found between sub-parallel active faults and relatively stronger scattering zones marked by larger circles, where some large circles are caused by topographic roughness. He also suggested that segment boundaries of the San Andreas Fault are characterized by relatively stronger scattering.

Stacking forward scattered energy in the coda of teleseismic P waves observed by a local seismographic network, Revenaugh [59] proposed a Kirchhoff coda migration method, which puts a focus on small-angle scattering from the forward direction. He made a map of *P*-wave scatterers in the upper mantle beneath southern California. Between depths of 50 km and 200 km, the southern flank of the slab subducting beneath the Transverse Ranges was marked by strong scattering. Using the same method, Revenaugh [60,61] estimated geographic variation of the statistical significance of scattering potential in the upper crust in California, where the scattering potential is a measure of the likelihood that scattering strength locally exceeds the regional mean. In the region surrounding the 1992 Landers earthquake of M 7.3, he found a noticeable tendency for aftershocks to cluster in regions of strong scattering potential.

There were more precise mappings of scattering coefficient. Slant-stacking records of 12 explosions in the Awaji island, Japan registered by a dense seismic array for a 6–10 Hz band, Matsumoto et al. [45] mapped the spatial distribution of PP single scatterers. The resultant distribution of scatterers shows higher strengths beneath the initiation point of the mainshock rupture and in the southwestern part of the fault plane of the 1995 Kobe earthquake (M 7.2). Analyzing precisely aftershock records of the 2000 western Tottori earthquake (M 7.3), Japan registered by a dense seismic network, Asano and Hasegawa [5] found strong scattering along and around the fault zone of 20 km in length.

Coda Attenuation and Deformation Zone

Regional variation of coda attenuation Q_c^{-1} has been measured from the decay gradient of coda amplitude envelopes of small earthquakes in various areas in the world. Jin and Aki [31] made a map of Q_c^{-1} at 1 Hz in China. They reported that Q_c^{-1} is as large as 0.01 in Tibet at the active continental collision. They found that large historical earthquakes took place in large Q_c^{-1} regions. Jin and Aki [33] made precise analysis of Q_c^{-1} for 1–32 Hz in Japan. They found significant spatial variation up to a factor of 3 for the lower frequency bands, as well as its strong frequency dependence. They found conspicuous large Q_c^{-1} for 1–4 Hz in a narrow belt from Niigata towards the south-west to the Biwa lake along the Japan Sea coast, which coincides with a high-deformation rate zone revealed from the GPS observation. For frequency bands 4–16 Hz (2–4 Hz in Kyushu), large Q_c^{-1} areas agree with volcanic and geothermal areas.

Attenuation and Volcanoes

Yoshimoto et al. [91] studied the spatial variation of MS amplitude of *S*-coda at a fixed lapse time across the volcanic front in northeastern Honshu, Japan: *S*-coda energy is uniformly distributed in the fore-arc side, whereas an exponential decrease with horizontal offset to the west from the volcanic front was found in the back-arc side. The

decay rate increases with increasing frequency. They interpreted this variation by a diffusion–absorption model, where the intrinsic absorption factor of S-wave $Q_{\mathrm{Int}}^{-1} = 0.02$ at a frequency of 10 Hz beneath the back-arc side, which is about twice as large as those reported for the fore-arc side.

Scattering and Volcanoes

Medium heterogeneity is strong beneath volcanoes. Applying the diffusion model to seismogram envelopes beneath Merapi volcanoes, Friedrich and Wegler [16] estimated the total scattering coefficient as large as $5\,\mathrm{km}^{-1}$ as shown in Fig. 4. Nishimura et al. [50] applied an envelope inversion method based on the isotropic scattering model for PP and PS scattering to artificial explosion records obtained in Jemez volcanic field, New Mexico. They found that the mid-crust under most of the region is fairly transparent but that the lower crust is heterogeneous. The strongest scattering occurs at shallow depths beneath the center of the caldera, where the medium is highly heterogeneous.

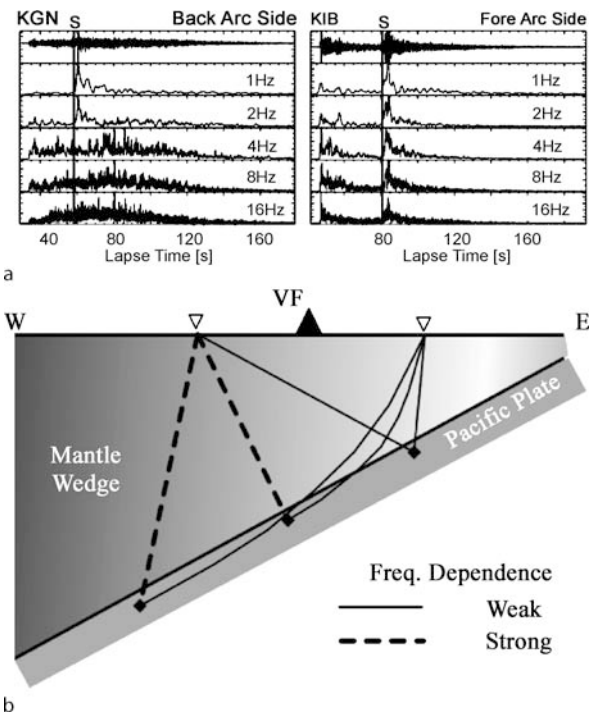

Seismic Waves in Heterogeneous Earth, Scattering of, Figure 10 **a** Seismogram envelopes observed in the back-arc side and fore-arc side in Kanto-Tokai, Japan and **b** a schematic illustration of seismic rays. Reproduced from [53]

Obara and Sato [53] analyzed S-wave envelopes of microearthquakes in Kanto-Tokai, Japan, where the Pacific plate is subducting from east to west, to examine regional differences in their envelope broadening. As shown by examples in Fig. 10a, envelope broadening is typically stronger for higher frequencies in records at stations on the back-arc side of the volcanic front but weaker and frequency independent in records on the fore-arc side. These regional differences in the envelope broadening mean that PSDF of velocity inhomogeneity is rich in short-wavelength components in the mantle wedge on the back-arc side and poor on the fore-arc side as schematically illustrated in Fig. 10b. Takahashi et al. [83] precisely examined how the peak delay from the S-wave onset depends on the ray path in northern Japan. They found that peak delays observed in the back-arc side of the volcanic front are larger for rays which propagate beneath Quaternary volcanoes (see Fig. 11b and d); however, peak delays for rays which propagate between Quaternary volcanoes are as short as those observed in the fore-arc side (see Fig. 11a, c, and e). Large peak delay suggests strong scattering due to medium inhomogeneity. That is, the structure beneath Quaternary volcanoes is not only characterized by low velocity and large intrinsic absorption revealed from tomography studies but also by strong inhomogeneity.

Nonisotropic Random Medium Oceanic Slab

If random media are statistically nonisotropic, scattering contribution depends on the propagation direction. Saito [65] studied the envelope broadening in nonisotropic random media based on the Markov approximation. His simulations show that the envelope of scalar wavelet propagating in parallel to the longer correlation direction has longer duration compared to that with the shorter correlation direction. The effective envelope broadening in the elongated direction shows the wave trap phenomenon of nonisotropic random media.

An intensity anomaly is observed on the eastern seaboard of northern Japan for deep earthquakes. The waveform records in the region of high intensity show a low-frequency ($f < 0.25\,\mathrm{Hz}$) onset for both P and S waves, followed by large-amplitude high-frequency ($f > 2\,\mathrm{Hz}$) later arrivals with long coda. A simple subduction zone model comprising a high-velocity plate with low attenuation cannot explain quantitatively these observed facts. Furumura and Kennett [17] proposed a scattering slab model that the nonisotropic random structure in the Pacific plate works as a wave-guide for high-frequency seismic waves. Their preferred random medium is characterized by a von Kármán-type ACF with elongated cor-

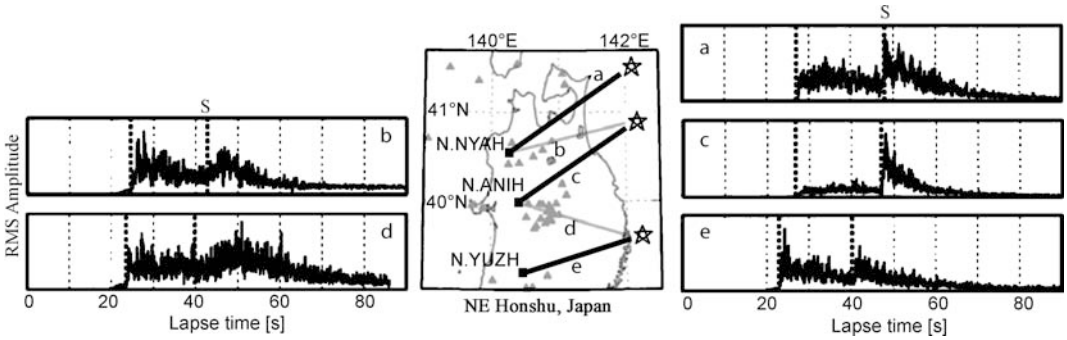

Seismic Waves in Heterogeneous Earth, Scattering of, Figure 11

Path dependence of RMS envelopes (16–32 Hz) in northeastern Honshu, Japan, where *stars* and *triangles* indicate earthquake epicenters and Quaternary volcanoes, respectively. Ray paths b and d travel beneath Quaternary volcanoes (*triangles*), and a, c, and e travel between Quaternary volcanoes. Reproduced from [83]

relation distance of about 10 km parallel to the plate margin and much shorter correlation length of about 0.5 km in thickness and ε of about 2%. They clearly demonstrated the scattering waveguide effects and frequency selectivity for seismic waves traveling through the Pacific plate by using 3-D numerical simulations.

Lateral Variation of Lithospheric Heterogeneity

Korn [34] developed the energy flux model [15] appropriate for the wave front of teleseismic P and P-coda waves propagating through a scattering layer. During the propagation the primary wave loses energy due to scattering and intrinsic absorption, then the scattered energy appears as coda energy behind the wave front. From the analysis of vertical component trace envelopes observed in the world, Korn [35] found strong scattering at island arcs and smaller scattering on stable continental areas like Australia. Nishimura et al. [52] analyzed the energy partition of teleseismic P and P-coda into the transverse component to evaluate the lithospheric heterogeneity in the western Pacific region. They showed the presence of strong heterogeneity in and around the tectonically active regions. Kubanza et al. [37] systematically characterized the medium heterogeneity of the lithosphere by analyzing the partition of P-wave energy into the transverse component for 0.5–4 Hz. They found significant regional differences as shown in Fig. 12. The energy partition to the transverse component is small at stations on stable continents while the partition is large at stations in tectonically active regions such as island arcs or collision zones.

Random Inhomogeneity in the Lithosphere and Mantle

Records of earthquakes registered by arrays of seismographs are useful for the statistical measurement of the Earth inhomogeneity. Aki [2] first analyzed teleseismic P-waves centered on about 0.6 Hz registered by a seismic array in Montana for the quantification of the lithospheric inhomogeneity. Measuring transverse correlation functions of teleseismic P-waves arriving from near vertical incidence, he found a positive correlation between log-amplitude and phase fluctuations as theoretically predicted for a Gaussian ACF. From plots of the ratio of RMS log-amplitude to RMS phase fluctuations against the correlation between log-amplitude and phase fluctuations, he estimated the thickness of the inhomogeneous lithosphere to be 60 km, $a = 10$ km, and $\varepsilon = 4\%$. Flatté and Wu [12] measured the transverse correlation of log-amplitude and phase fluctuations of teleseismic P-wave beams with 2 Hz center frequency recorded at NORSAR. They also introduced the new concept of angular correlation functions, which are based on measurements of two rays with different incident angles. They proposed a model for lithospheric and asthenospheric inhomogeneities that consists of two overlapping layers: the upper layer extending from the surface to about the 200 km depth has a white PSDF, however, the lower layer extending from 15 to 250 km has poor amplitude in the short wavelength spectrum. There are more small-scale inhomogeneities near the surface compared with the deeper portions.

The radiative transfer theory with scattering coefficients calculated from the Born approximation was also used for the study of mantle inhomogeneity. Analyzing stacked P and P coda envelopes of teleseismic (>10°) events at 1 Hz, Shearer and Earle [79] concluded that most scattering occurs in the lithosphere and upper mantle, but that some lower mantle scattering is likely required. They estimated ε to be 3–4% and $a = 4$ km in the upper mantle and 0.5% and 8 km in the lower mantle. Analyzing envelopes of precursors to PKP, Margerin and Nolet [43] found that inhomogeneity can not be restricted to the D″ layer and a small inhomogeneity spread over the whole

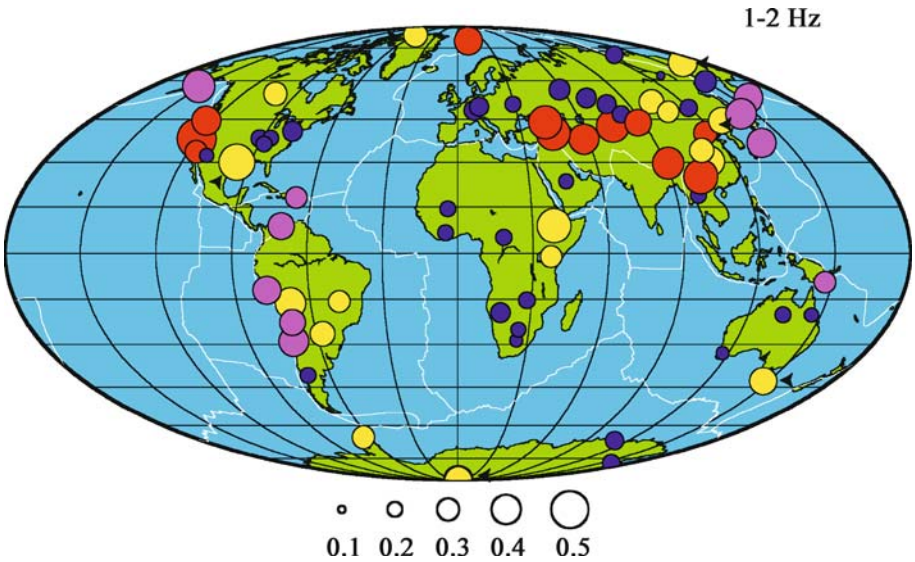

Seismic Waves in Heterogeneous Earth, Scattering of, Figure 12
Plots of the square root of the relative partition of wave energy into the transverse component (1–2 Hz) by a diameter of the circle revealed from the teleseismic P-waves. Reproduced from [37]

lower mantle. They proposed a von Kármán ACF of $\kappa = 0$ for random media, which has a power law PSDF rich in short wavelengths compared with an exponential ACF. They mentioned that ε of 0.1–0.2% in the whole lower mantle is enough to explain the observation even though correlation distance is irresolvable because of the limited range of observations.

Imaging of Subsurface Heterogeneity

There have been developments in deterministic imaging of medium inhomogeneity from the analysis of array records of P-coda waves. When the structures of interest are characterized by laterally variable stratification, the receiver-function technique [38] is useful since it is based on the deconvolution of the horizontal component trace in the radial direction by the vertical component trace for measuring the Ps conversion depth. On the other hand, scattering from localized volume inhomogeneity is most readily treated by using the Born approximation. Analyzing array records of teleseismic P coda in central Oregon by using the Born approximation for both forward scattered waves and backscattered free-surface reflected waves, Rondenay et al. [62] successfully imaged the precise structure of the Cascadia subduction zone, which is consistent with the consequences of prograde metamorphic reactions occurring within the oceanic crust. Analyzing array records of P coda waves of regional earthquakes at Izu-Oshima volcano, which erupted in 1986, Mikada et al. [46] deterministically imaged PP and PS scatterers on the basis of diffraction tomography. They interpreted a cloud of scatterers centered at about 10 km depth beneath the volcano crater as a primary magma reservoir and smaller and shallower patches of high scattering strength with sub-magma reservoirs.

Temporal Change in the Earth Medium Structure

There were reports on the temporal changes in the Earth medium structure revealed from the analyses of scattered waves. One is coda amplitude envelope analysis, which gives information about the change in intrinsic absorption and scattering strength of the crustal heterogeneity. Another is coda phase interferometry, which offers information about the change in background velocity.

Change in Coda Characteristics

Monitoring coda envelopes of local earthquakes, Gusev and Lemzikov [23] reported temporal change in Q_c^{-1} before and after the 1971 Ust-Kamchatsk earthquake (M 7.8), and Jin and Aki [30] reported temporal change in Q_c^{-1} associated with the 1976 Tangshan earthquake (M 7.8) in China. Their observation attracted the interest of geophysicists to the temporal variation of coda characteristics because of a potential for monitoring the stress accumulation process preceding an earthquake occurrence. Analyzing high-frequency seismograms recorded at River-

side, California for 55 years, Jin and Aki [32] found a temporal variability in Q_c^{-1} at about 1.6 Hz having a positive correlation with the seismic b-value calculated for M > 3 earthquakes within a 180 km radius. The seismic b-value is a measure of the ratio between the numbers of small to large earthquakes; smaller b-values mean that there are relatively fewer small earthquakes compared to the number of larger ones. They interpreted these changes in Q_c^{-1} and the b-value by creep fractures in the ductile part of the lithosphere. Hiramatsu [25] precisely examined the temporal variation in Q_c^{-1} and b-value for 10 years before and after the 1995 Hyogo-ken Nanbu earthquake (M 7.2) in Japan. At frequencies between 1.5 and 4.0 Hz the temporal variation in Q_c^{-1} increased after the mainshock occurrence, where the variation in b-value was opposite.

Sato [69] analyzed the relation between coda duration time and earthquake magnitude of small earthquakes before and after an M 6.8 earthquake in central Japan. He found that coda durations were anomalously longer than usual for 16 months before the earthquake occurrence. From 24-year observation of coda at 0.5 Hz in Kamchatka Gusev [19] reported prominent anomalies in coda level residual from the mean coda excitation level at 100s lapse time associated with two M 8 earthquakes and a volcanic eruption.

Sawazaki et al. [76] measured the temporal variation of the spectral ratio of coda waves registered on the ground surface to that at the bottom of a borehole of 100 m depth in Japan, which experienced strong ground motion of several hundred gals. They reported a sudden drop of the site amplification factor caused by earthquake strong motion and gradual recovery for a few years approaching to the original ratio. They suggested crack formation and ground water movement for explaining the site factor weakening observed.

Coda Interferometry

Pairs of earthquakes with almost identical focal mechanisms are called earthquake doublets. The cross-correlation function of earthquake doublet records allow us to detect differences in the background velocity of the Earth medium that took place in between the pair of earthquakes. Applying the phase spectral analysis to coda wave records of earthquake doublets before and after the 1979 Coyote earthquake of M 5.9 in California, Poupinet et al. [57] found that the coda wave arrivals for some stations are progressively delayed for the second earthquake in the doublet. They interpreted systematic variation along the coda as a decrease of background S-wave velocity by 0.2% in an oblong region 5–10 km in radius at the south end of the aftershock zone. Applying the phase spectral analysis to records of repeated artificial explosions, Nishimura et al. [51] found that the average seismic velocity of the crust in the frequency range of 3–6 Hz decreased by about 1% around the focal region of an M 6.1 earthquake at Iwate volcano in northeastern Honshu, Japan in 1998. They interpreted this velocity drop by the dilatation caused by the M 6.1 earthquake with stress sensitivity of the velocity change $(\delta V/V)/\delta\sigma$ of the order of 0.1 MPa^{-1}. From the set of successive artificial explosion experiments, they observed gradual recovery of the seismic velocity towards its original value over the next four years. While interferometry detected a change in velocity of the order of 1%, it was unidentifiable from travel time analysis of first arrivals. Using coda waves is superior to direct waves since coda waves volumetrically sample the Earth medium.

Snieder et al. [81] demonstrated detection of the nonlinear dependence of the seismic velocity in granite on temperature and the associated acoustic emissions from the interference measurement of coda waves in rock samples as a laboratory experiment. They named this method "coda interferometry" and proposed to use it for detecting the presence of temporal changes in the medium, or in diagnostic mode. There is an idea to retrieve the Green function from the stacked cross-correlation function (CCF) of multiple scattered waves or microseisms at a pair of stations on condition that the propagation directions of those waves are randomly isotropic. Stacking CCFs of coda waves at several pairs of stations for regional earthquakes in Mexico, Campillo and Paul [7] estimated the surface wave velocity between each station pair from the peak delay. The idea was extended for monitoring the temporal change in the crustal structure. Wegler and Sens-Schönfelder [84] computed the ACF of microseisms recorded at a site in the vicinity of the source region of the 2004 Mid-Niigata earthquake (M 6.6) in Japan for three months. They detected a sudden decrease of relative seismic velocity in the crust of 0.6% at the occurrence of the earthquake from the temporal variation of stacked ACFs.

Future Directions

In addition to classic parameterization as a layered structure with sharp edges and smooth velocity perturbation, we introduced new approaches using scattered waves that reflect solid Earth heterogeneity. For high-frequency seismograms of earthquakes, envelope characteristics such as the excitation level and the decay gradient of coda envelopes and the envelope broadening of the direct wavelet

are useful for the study of small-scale inhomogeneities. The lithospheric inhomogeneity is phenomenologically well characterized by the scattering coefficient and coda attenuation factor as a function of frequency. Furthermore, the power spectral density function of random velocity inhomogeneity is estimated from the frequency dependence of high-frequency seismogram envelopes of local earthquakes or the array analysis of teleseismic waves. The radiative transfer theory with scattering coefficients calculated from the Born approximation and the Markov approximation for the parabolic wave equation are useful mathematical tools for the analyses.

Scattering characteristics are found to vary spatially reflecting seismotectonic settings. It will be necessary for us to make a classification of seismogram-envelope patterns in various regions in the world under different tectonic conditions. It is interesting to model how such a variation of medium inhomogeneity was created through the geodynamic process. Compared to the lithospheric inhomogeneity, there were insufficient numbers of studies on the mantle inhomogeneity. It will be necessary to map the distribution of inhomogeneities deep in the mantle, which is useful for the study of the evolution of the planet Earth.

For mathematical simplicity; however, most approaches assume homogeneity and isotropy of randomness and a constant background velocity, which are somewhat different from reality. It will be necessary to mathematically develop the envelope synthesis in inhomogeneous media that are a superposition of small-scale random inhomogeneities and a gradually varying background velocity. As revealed from the ray path dependence of S-wave envelope broadening, randomness varies from place to place. It is also necessary to develop the envelope synthesis for random media having spatially varying statistical parameters. In addition, it is important to examine how conversion scattering between P and S waves contributes to form spindle-like envelopes in highly scattering media as shown in high-frequency seismograms observed in volcanoes and on the Moon.

For further understanding, there are monographs that treat the discussed subjects as follows: Sato and Fehler [74] review seismological observation and mathematical models; Shapiro and Hubral [78] put special focus on wave propagation through stratified random media; Goff and Holliger [18] summarize the crustal heterogeneity; Wu and Maupin [86] compile recent developments in mathematical modeling of wave propagation in inhomogeneous media; Chandrasekhar [8] is a classic text for radiative transfer theory; Ishimaru [29] and Rytov et al. [63] offer advanced mathematical tools for the study of wave propagation in random media.

Acknowledgments

The author thanks Michael Korn, Heiner Igel, and an anonymous reviewer. Their comments were helpful for improving the readability. The author is grateful to NIED, Japan for providing digital seismic data.

Bibliography

1. Aki K (1969) Analysis of seismic coda of local earthquakes as scattered waves. J Geophys Res 74:615–631
2. Aki K (1973) Scattering of P waves under the Montana LASA. J Geophys Res 78:1334–1346
3. Aki K (1980) Attenuation of shear-waves in the lithosphere for frequencies from 0.05 to 25 Hz. Phys Earth Planet Inter 21:50–60
4. Aki K, Chouet B (1975) Origin of coda waves: Source, attenuation and scattering effects. J Geophys Res 80:3322–3342
5. Asano Y, Hasegawa A (2004) Imaging the fault zones of the 2000 western Tottori earthquake by a new inversion method to estimate three-dimensional distribution of the scattering coefficient. J Geophys Res 109:B06306. doi:10.1029/2003JB002761
6. Atkinson GM (1993) Notes on ground motion parameters for Eastern North America: Duration and H/V ratio. Bull Seismol Soc Am 83:587–596
7. Campillo M, Paul A (2003) Long-Range Correlations in the Diffuse Seismic Coda. Science 299:547–549. doi:10.1126/science.1078551
8. Chandrasekhar S (1960) Radiative Transfer. Dover, New York
9. Dainty AM, Toksöz MN (1981) Seismic codas on the earth and the moon: A comparison. Phys Earth Planet Inter 26:250–260
10. Fehler M, Hoshiba M, Sato H, Obara K (1992) Separation of scattering and intrinsic attenuation for the Kanto-Tokai region, Japan, using measurements of S-wave energy versus hypocentral distance. Geophys J Int 108:787–800
11. Fehler M, Sato H, Huang LJ (2000) Envelope broadening of outgoing waves in 2-D random media: A comparison between the Markov approximation and numerical simulations. Bull Seismol Soc Amer 90:914–928
12. Flatté SM, Wu RS (1988) Small-scale structure in the lithosphere and asthenosphere deduced from arrival time and amplitude fluctuations at NORSAR. J Geophys Res 93:6601–6614
13. Foldy LL (1945) The multiple scattering of waves- I General theory of isotropic scattering by randomly distributed scatterers. Phys Rev 67:107–119
14. Frankel A, Clayton RW (1986) Finite difference simulations of seismic scattering: Implications for the propagation of short-period seismic waves in the crust and models of crustal heterogeneity. J Geophys Res 91:6465–6489
15. Frankel A, Wennerberg L (1987) Energy-flux model of seismic coda: Separation of scattering and intrinsic attenuation. Bull Seismol Soc Am 77:1223–1251
16. Friedrich C, Wegler U (2005) Localization of seismic coda at Merapi volcano (Indonesia). Geophys Res Lett 32:L14312. doi:10.1029/2005GL023111
17. Furumura T, Kennett BLN (2005) Subduction zone guided waves and the heterogeneity structure of the subducted plate: intensity anomalies in northern Japan. J Geophys Res 110:B10302. doi:10.1029/2004JB003486

18. Goff JA, Holliger K (2002) Heterogeneity in the Crust and Upper Mantle – Nature, Scaling and Seismic Properties. Kluwer Academic/Plenum Publishers, Dordrecht, pp 1–358
19. Gusev AA (1995) Baylike and continuous variations of the relative level of the late coda during 24 years of observation on Kamchatka. J Geophys Res 100:20311–20319
20. Gusev AA (1995) Vertical profile of turbidity and coda Q. Geophys J Int 123:665–672
21. Gusev AA, Abubakirov IR (1987) Monte-Carlo simulation of record envelope of a near earthquake. Phys Earth Planet Inter 49:30–36
22. Gusev AA, Abubakirov IR (1996) Simulated envelopes of non-isotropically scattered body waves as compared to observed ones: Another manifestation of fractal heterogeneity. Geophys J Int 127:49–60
23. Gusev AA, Lemzikov VK (1985) Properties of scattered elastic waves in the lithosphere of Kamchatka: Parameters and temporal variations. Tectonophysics 112:137–153
24. Hemmer PC (1961) On a generalization of Smoluchowski's diffusion equation. Physica A 27:79–82
25. Hiramatsu Y, Hayashi N, Furumoto M (2000) Temporal changes in coda *Q*21 and *b* value due to the static stress change associated with the 1995 Hyogo-ken Nanbu earthquake. J Geophys Res 105:6141–6151
26. Holliger K, Levander A (1992) A stochastic view of lower crustal fabric based on evidence from the Ivrea zone. Geophys Res Lett 19:1153–1156
27. Hoshiba M (1991) Simulation of multiple-scattered coda wave excitation based on the energy conservation law. Phys Earth Planet Inter 67:123–136
28. Hoshiba M, Sato H, Fehler M (1991) Numerical basis of the separation of scattering and intrinsic absorption from full seismogram envelope – A Monte-Carlo simulation of multiple isotropic scattering. Pa Meteorol Geophys, Meteorol Res Inst 42:65–91
29. Ishimaru A (1978) Wave Propagation and Scattering in Random Media, vol 1 and 2. Academic, San Diego
30. Jin A, Aki K (1986) Temporal change in coda *Q* before the Tangshan earthquake of 1976 and the Haicheng earthquake of 1975. J Geophys Res 91:665–673
31. Jin A, Aki K (1988) Spatial and temporal correlation between coda *Q* and seismicity in China. Bull Seismol Soc Am 78:741–769
32. Jin A, Aki K (1989) Spatial and temporal correlation between coda Q^{-1} and seismicity and its physical mechanism. J Geophys Res 94:14041–14059
33. Jin A, Aki K (2005) High-resolution maps of Coda Q in Japan and their interpretation by the brittle-ductile interaction hypothesis. Earth Planets Space 57:403–409
34. Korn M (1990) A modified energy flux model for lithospheric scattering of teleseismic body waves. Geophys J Int 102:165–175
35. Korn M (1993) Determination of site-dependent scattering Q from P-wave coda analysis with an energy-flux model. Geophys J Int 113:54–72
36. Korn M, Sato H (2005) Synthesis of plane vector-wave envelopes in 2-D random elastic media based on the Markov approximation and comparison with finite difference simulations. Geophys J Int 161:839–848
37. Kubanza M, Nishimura T, Sato H (2006) Spatial variation of lithospheric heterogeneity on the globe as revealed from transverse amplitudes of short-period teleseismic P-waves. Earth Planets Space 58:45–e48
38. Langston CA (1979) Structure under Mount Rainer, Washington, inferred from teleseismic body waves. J Geophys Res 84:4749–4762
39. Larose E, Margerin L, van Tiggelen BA, Campillo M (2004) Weak Localization of Seismic Waves. Phys Rev Lett 93:048501-4. doi:10.1103/PhysRevLett.93.048501
40. Lee WS, Sato H, Lee KW (2003) Estimation of S-wave scattering coefficient in the mantle from envelope characteristics before and after the ScS arrival. Geophys Res Lett 30:2248. doi:10.1029/2003GL018413
41. Margerin L (2005) Introduction to radiative transfer of seismic waves. In: Levander A, Nolet G (eds) Seismic Earth: Array Analysis of Broad-band Seismograms, Geophysical Monograph Series, vol 157, chap 14. AGU, Washington, pp 229–252
42. Margerin L, Campillo M, van Tiggelen BA (2001) Coherent backscattering of acoustic waves in the near field. Geophys J Int 145:593–603
43. Margerin L, Nolet G (2003) Multiple scattering of high-frequency seismic waves in the deep Earth: PKP precursor analysis and inversion for mantle granularity. J Geophys Res 108, B11:2514. doi:10.1029/2003JB002455
44. Matsumoto S (2005) Scatterer density estimation in the crust by seismic array processing. Geophys J Int 163:622–628
45. Matsumoto S, Obara K, Hasegawa A(1998) Imaging P-wave scatterer distribution in the focal area of the 1995 M7.2 Hyogo-ken Nanbu (Kobe) Earthquake. Geophys Res Lett 25:1439–1442
46. Mikada H,Watanabe H, Sakashita S (1997) Evidence for subsurface magma bodies beneath Izu-Oshima volcano inferred from a seismic scattering analysis and possible interpretation of the magma plumbing system of the 1986 eruptive activity. Phys Earth Planet Inter 104:257–269
47. Nakahara H, Nishimura T, Sato H, Ohtake M (1998) Seismogram envelope inversion for the spatial distribution of high-frequency energy radiation from the earthquake fault: Application to the 1994 far east off Sanriku earthquake, Japan. J Geophys Res 103:855–867
48. Nishigami K (1991) A new inversion method of coda waveforms to determine spatial distribution of coda scatterers in the crust and uppermost mantle. Geophys Res Lett 18:2225–2228
49. Nishigami K (2000) Deep crustal heterogeneity along and around the San Andreas fault system in central California and its relation to the segmentation. J Geophys Res 105:7983–7998
50. Nishimura T, Fehler M, Baldridge WS, Roberts P, Steck L (1997) Heterogeneous structure around the Jemez Volcanic Field, New Mexico, USA, as inferred from the envelope inversion of active-experiment seismic data. Geophys J Int 131:667–681
51. Nishimura T, Tanaka S, Yamawaki T, Yamamoto H, Sano T, Sato M, Nakahara H, Uchida N, Hori S, Sato H (2005) Temporal changes in seismic velocity of the crust around Iwate volcano, Japan, as inferred from analyses of repeated active seismic experiment data from 1998 to 2003. Earth Planets Space 57:491–505
52. Nishimura T, Yoshimoto K, Ohtaki T, Kanjo K, Purwana I (2002) Spatial distribution of lateral heterogeneity in the upper mantle around the western Pacific region as inferred from analysis of transverse components of teleseismic P-coda. Geophys Res Lett 29:2089–2137. doi:10.1029/2002GL015606

53. Obara K, Sato H (1995) Regional differences of random inhomogeneities around the volcanic front in the Kanto-Tokai area, Japan, revealed form the broadening of S wave seismogram envelopes. J Geophys Res 100:2103–2121
54. Paaschens JCJ (1997) Solution of the time-dependent Boltzmann equation. Phys Rev E 56:1135–1141
55. Petukhin AG, Gusev AA (2003) The Duration-distance Relationship and Average Envelope Shapes of Small Kamchatka Earthquakes. Pure Appl Geophys 160:1717–1743
56. Phillips WS, Aki K (1986) Amplification of coda waves from local earthquakes in Central California. Bull Seismol Soc 76:627–648
57. Poupinet G, Ellsworth WL, Frechet J (1984) Monitoring velocity variations in the crust using earthquake doublets: An application to the Calaveras fault, California. J Geophys Res 89:5719–5731
58. Przybilla J, Korn M, Wegler U (2006) Radiative transfer of elastic waves versus finite difference simulations in two-dimensional random media. J Geophys Res 111:B04305. doi:10.1029/2005JB003952
59. Revenaugh J (1995) A scattered-wave image of subduction beneath the Transverse Ranges. Science 268:1888–1892
60. Revenaugh J (1995) Relationship of the 1992 Landers, California, earthquake sequence to seismic scattering. Science 270:1344–1347
61. Revenaugh J (1999) Geologic Applications of Seismic Scattering. Annu Rev Earth Planet Sci 27:55–73
62. Rondenay S, Bostock MG, Shragge J (2001) Multiparameter two-dimensional inversion of scattered teleseismic body waves 3. Application to the Cascadia 1993 data set. J Geophys Res 106:30795–30807
63. Rytov SM, Kravtsov YA, Tatarskii VI (1987) Principles of Statistical Radio Physics, vol 4, Wave Propagation Through Random Media. Springer, Berlin
64. Ryzhik LV, Papanicolaou GC, Keller JB (1996) Transport equations for elastic and other waves in random media. Wave Motion 24:327–370
65. Saito T (2006) Synthesis of scalar-wave envelopes in two-dimensional weakly anisotropic random media by using the Markov approximation. Geophys J Int 165:501–515. doi:10.1111/j.1365-246X2006.02896.x
66. Saito T, Sato H, Ohtake M (2002) Envelope broadening of spherically outgoing waves in three-dimensional random media having power-law spectra. J Geophys Res 107:2089. doi:10.1029/2001JB000264
67. Sato H (1977) Single isotropic scattering model including wave conversions: Simple theoretical model of the short period body wave propagation. J Phys Earth 25:163–176
68. Sato H (1984) Attenuation and envelope formation of three-component seismograms of small local earthquakes in randomly inhomogeneous lithosphere. J Geophys Res 89:1221–1241
69. Sato H (1987) A precursor-like change in coda excitation before the western Nagano earthquake (Ms = 6.8) of 1984 in central Japan. J Geophys Res 92:1356–1360
70. Sato H (1989) Broadening of seismogram envelopes in the randomly inhomogeneous lithosphere based on the parabolic approximation: Southeastern Honshu, Japan. J Geophys Res 94:17735–17747
71. Sato H (1990) Unified approach to amplitude attenuation and coda excitation in the randomly inhomogeneous lithosphere. Pure Appl Geophys 132:93–121
72. Sato H (2006) Synthesis of vector wave envelopes in three-dimensional random elastic media characterized by a Gaussian autocorrelation function based on the Markov approximation: Plane wave case. J Geophys Res 111:B06306. doi:10.1029/2005JB004036
73. Sato H (2007) Synthesis of vector-wave envelopes in 3-D random elastic media characterized by a Gaussian autocorrelation function based on the Markov approximation: Spherical wave case. J Geophys Res Solid Earth 112:B01301. doi:10.1029/2006JB004437
74. Sato H, Fehler M (1998) Seismic Wave Propagation and Scattering in the Heterogeneous Earth. AIP Press/Springer, New York
75. Sato H, Nakahara H, Ohtake M (1997) Synthesis of scattered energy density for non-spherical radiation from a point shear dislocation source based on the radiative transfer theory. Phys Earth Planet Inter 104:1–281
76. Sawazaki K, Sato H, Nakahara H, Nishimura T (2006) Temporal Change in Site Response Caused by Earthquake Strong Motion as Revealed from Coda Spectral Ratio Measurement. Geophys Res Lett 33:L21303. doi:10.1029/2006GL027938
77. Shang T, Gao L (1988) Transportation theory of multiple scattering and its application to seismic coda waves of impulsive source. Sci Sin 31B:1503–1514
78. Shapiro SA, Hubral P (1999) Elastic Waves in Random Media – Fundamentals of Seismic Stratigraphic Filtering. Springer, Berlin
79. Shearer PM, Earle PS (2004) The global short-period wavefield modeled with a Monte Carlo seismic phonon method. Geophys J Int 158:1103–1117
80. Shiomi K, Sato H, Ohtake M (1997) Broad-band power-law spectra of well-log data in Japan. Geophys J Int 130:57–64
81. Snieder R, Gret A, Douma A, Scales J (2002) Coda wave interferometry for estimating nonlinear behavior in seismic elocity. Science 295:2253–2255
82. Sreenivasiah I, Ishimaru A, Hong ST (1976) Two-frequency mutual coherence function and pulse propagation in a random medium: An analytic solution to the plane wave case. Radio Sci 11:775–778
83. Takahashi T, Sato H, Nishimura T, Obara K (2006) Strong inhomogeneity beneath Quaternary volcanoes revealed from the peak delay analysis of S-wave seismograms of microearthquakes in northeastern, Japan. Geophys J Int 168:90–99. doi:10.1111/j.1365-246X2006.03197.x
84. Wegler U, Sens-Schönfelder C (2007) Fault zone monitoring with passive image interferometry. Geophys J Int 168:1029-1033. doi:10.1111/j.1365-246X2006.03284.x
85. Wu RS (1985) Multiple scattering and energy transfer of seismic waves – separation of scattering effect from intrinsic attenuation – I Theoretical modeling. Geophys J R Astron Soc 82:57–80
86. Wu RS, Maupin V (eds) (2007) Advances in Wave Propagation in Heterogeneous Earth. In: Dmowska R (ed) Advanced in Geophysics, vol 48. Academic Press, San Diego, pp 561–596
87. Yomogida K, Benites R (1995) Relation between direct wave *Q* and coda *Q*: A numerical approach. Geophys J Int 123:471–483
88. Yoshimoto K (2000) Monte-Carlo simulation of seismogram envelope in scattering media. J Geophys Res 105:6153–6161
89. Yoshimoto K, Sato H, Ohtake M (1993) Frequency-dependent attenuation of P and S waves in the Kanto area, Japan, based on the coda-normalization method. Geophys J Int 114:165–174

90. Yoshimoto K, Sato H, Ohtake M (1997) Short-wavelength crustal inhomogeneities in the Nikko area, central Japan, revealed from the three-component seismogram envelope analysis. Phys Earth Planet Inter 104:63–73
91. Yoshimoto K, Wegler U, Korn M (2006) A volcanic front as a boundary of seismic attenuation structures in northeastern Honshu, Japan. Bull Seismol Soc Am 96:637–646
92. Zeng Y, Su F, Aki K (1991) Scattering wave energy propagation in a random isotropic scattering medium I Theory. J Geophys Res 96:607–619
93. Zhao D, Hasegawa A, Horiuchi S (1992) Tomographic imaging of P and S wave velocity structure beneath Northeastern Japan. J Geophys Res 97:19909–19928

Self-assembled Materials

Aatto Laaksonen[1], Lennart Bergström[2]
[1] Division of Physical Chemistry, Dept. of Physical, Inorganic and Structural Chemistry, Arrhenius Laboratory, Stockholm University, Stockholm, Sweden
[2] Materials Chemistry Research Group, Dept. of Physical, Inorganic and Structural Chemistry, Arrhenius Laboratory, Stockholm University, Stockholm, Sweden

Article Outline

Glossary
Definition of the Subject
Introduction
First-Principles Calculations and Simulations
Atomistic Molecular Dynamic Simulations
Monte Carlo Simulations
Kinetic Monte Carlo Simulations
Phase-Field Modeling
Random Sequential Adsorption
Cellular Automata
Multi-Scale Modeling
Future Directions
Acknowledgments
Bibliography

Glossary

AI Ab initio
AM1 Austin model 1
AO Atomic orbital
B3LYP Becke's three parameter hybrid LYP functional
BKL Bortz–Kalos–Lebowitz
BO Born–Oppenheimer
CA Cellular automata
CE Cluster expansion
CC Coupled cluster
CI Configuration interaction
CP Car-Parrinello
DCA Dynamic cellular automata
DFT Density functional theory
DLVO Derjaguin–Landau–Verwey–Overbeek
GGA Generalized gradient approximation
HF 1) Hartree–Fock 2) Harris–Foulkes functional
kMC kinetic Monte Carlo
KS Kohn–Sham
LbL Layer-by-layer
LCAO Linear Combination of Atomic Orbitals
LDA Local density approximation
LYP Lee–Young–Parr
MC Monte Carlo
MD Molecular Dynamics
MNDO Modified neglect of differential overlap
MO Molecular Orbital
MP Møller–Plesset
NDDO Neglect of diatomic differential overlap
NDO Neglect of differential overlap
NPT Isobaric ensemble: constant N, P and T
NVE Microcanonical ensemble: constant N, V and E
NVT Canonical ensemble: constant N, V and T
PBC Periodic boundary conditions
PF Phase-field
PM3 Parametric model 3
PP Pseudo-potential
PW Perdew–Wang
QCA Quantum-dot cellular automata
RSA Random sequential adsorption
SA Self-assembly
SAM Self-assembled monolayers
SCF Self-consistent field
STM Scanning tunneling microscopy
TB Tight-binding
XC Exchange correlation

Definition of the Subject

The concept of self-assembly is today highly popular and frequently used to describe a wide range of phenomena. It is also a concept that has a possibility to change the way we produce various types of materials [1]. Self-assembly can broadly speaking be defined as a process with the following features [2], where: (i) it involves pre-existing components, i. e. the components are not formed by the reaction itself. (ii) the process should be reversible to some extent. (iii) it can be controlled by design.

The supramolecular chemistry approach, pioneered by Jean-Marie Lehn [3], where molecular recognition is used to assemble supramolecular materials is clearly an impor-

tant starting point for much of this work. The pre-existing components are usually molecules, amphiphilic species, or oligomeric or polymeric species. However, self-assembly can also be observed on larger length scales involving e. g. nanoparticles (superlattices) and larger colloidal particles (colloidal crystals). The final self-assembled material should display certain features that can be related to the building blocks in addition to the novel features that are created by the self-assembled material itself. Reversibility is a key feature since self-assembly processes commonly scan a rich energy landscape with many metastable states. This is commonly regulated by the balance between the attractive and repulsive interactions between the self-assembling components, involving e. g. van der Waals, electrostatic, hydrophobic, hydrogen bonding and entropic forces (molecules and mesoscale objects), and magnetic, capillary, and gravity forces (meso- or macroscale objects).

The degree of versatility, simplicity and flexibility of the various self-assembly methods and the ability to introduce specific functions with a high degree of spatial accuracy are important features for self-assembled materials to move beyond academic beauty and be interesting for applications. The preparation of self-assembled coatings can be traced back to the seminal work by Irving Langmuir and Katherine Blodgett prior to the Second World War. The development of a theory for adsorption on surfaces by Langmuir together with the design of a technique that both could measure the pressure-area isotherms of monolayers of amphiphilic molecules assembled at the air-liquid interface, and also be used to transfer these monolayers onto different substrates was a very important discovery that has had a profound effect on future work on self-assembled layers. The observation by Nuzzo and coworkers [4] that alkanethiolates can self-assemble on gold surfaces and result in well-ordered films sparked a large research interest in these so called self-assembled monolayers (SAMs). The relatively non-local nature of the thiol bond to the gold surface allows the much weaker, attractive interaction *between* the amphiphilic chains in the monolayer to control the film structure. The SAMs are a beautiful example of self-assembly as a crucial process in the formation of a dense monolayer.

Self-organization of nanoparticles into two- and three-dimensional superlattices has attracted much interest since the early work on iron oxide "super crystals" [5]. Indeed, understanding and optimizing the structures, at all length scales, of nanocrystal superlattices is an important step towards controlled design of novel nanostructured materials and devices. This work did not really gain momentum until methods to achieve shape and size control with a high fidelity was established in the end of 1980s by e. g. Brus, Steigerwald et al. [6] and Moerup, Thölen and Koch [5] mentioned above. The ability to assemble different nanocrystals with size-tunable optical, electronic and magnetic properties into well-defined structures opens up the possibility to study and develop new materials with tailored couplings between the constituent units [7,8]. Layer-by-layer (LbL) self-assembly of charged polymers has evolved as probably the most versatile technique to create multifunctional coatings on a wide range of substrates [9]. These are but a limited set of examples of self-assembly processes to produce novel materials that has been demonstrated in the last decades [10,11,12,13,14,15,16,17,18,19,20]. While the examples of self-assembly are plentiful, the attempts to model the self-assembly processes and systems is still in its infancy. This review attempts to be a guide and introduction to the field of computer modeling of self-assembly.

Introduction

Computer modelling of structural properties and dynamical processes in matter is based on physical and mathematical models. Ideally, we would prefer to use first-principles methods of quantum mechanics with no other input than atomic nuclei and the electrons spontaneously forming atoms, molecules and more complicated systems through local intermolecular interactions. Unfortunately this is possible only for systems of very modest sizes using rather drastic approximations and simplifications. We can always blame the computers for not been powerful enough but the truth is that the "first-principles" models would grow too complicated and out of hands very rapidly. We therefore use different physical models to describe matter at different length and time scales together with an increasing amount of empirical data as input to the models.

Atomistic computer simulations such as the classical Molecular Dynamics (MD) are currently run for molecular systems consisting of the order of 10^5 atoms, corresponding typically to a system size (length scale) of 5–10 nm. These simulations are typically extended to 10–100 ns (time scale). This may be enough for simulations of isotropic liquids of simple molecules, whereas for many complex systems, the requirements for the length and time scales are much more extensive. To extend the simulations to cover longer scales requires either simplifications of used models or choosing more suitable modelling technique [21]. In the first strategy the resolution of the model is reduced by removing certain degrees of freedom of the studied system which are not important or can be considered as fluctuations while keeping those degrees of freedom which are operational on the longer scales. This is

Self-assembled Materials, Table 1
Time and length scales covered by modelling methods reviewed in this article

Scale	Typical length	Typical time
sub-atomic	< 0.1 nm	∼ 1 fs
atomic	0.1–1 nm	1 fs–1 ns
meso	1 nm–1 μm	1 ns–1 ms
macro	1 μm–	1 ms–

generally called coarse-graining of the model leading to a much fewer interaction sites thereby speeding up the calculations. An alternative approach is to use a hybrid by keeping an accurate description for a selected area of importance and the environment is merely made to a fluctuating medium with more or less specific interactions. These techniques may include combining quantum and classical mechanics [22] or classical mechanics and hydrodynamic/continuum models.

In Table 1 there is a somewhat artificial division of matter on scales from sub-atomic (with nuclei and electrons) to macroscopic. There are modelling and simulation methods designed separately to these scales. Going from one scale to the next in the table requires either coarse-graining or fine-graining of the used model by building a bridge between two techniques. This means taking as much vital information as possible to the next level. Coarse-graining is obviously easier than fine-graining as the latter requires re-activating non-existing degrees of freedom from an under-defined system. Multiscale modelling is currently a hot topic and in a rapid development with the purpose to bridge several scales in a more systematic way.

The aim of this chapter is to describe a box of tools containing some selected modelling methods to study self-assembly. The tools should together cover the scales in the table in Table 1. We will give a short description of each of them together with some examples of applications. At the end of the chapter we show examples of how to combine several techniques to extend length and time scales in modelling. We will work from the "bottom-up" by starting with first-principles calculations on electrons and nuclei. Classical Molecular Dynamics (MD) and Monte Carlo (MC) are mainly used in atomistic simulations with empirically parameterized potential functions. Meso-scale simulations of aggregates and soft particles use coarse-grained models. We will discuss methods equally applicable on both short and long scales such as the kinetic Monte Carlo (kMC) and Phase Field (PF) modelling. All these techniques so far are examples of off-lattice modelling methods. We then discuss two lattice methods with some further development towards more off-lattice character, namely Random Sequential Adsorption (RSA) and Cellular Automata (CA). Finally we will discuss multiscale modeling before the paper ends with a discussion on future directions.

First-Principles Calculations and Simulations

Down in the bottom of the world of atoms and molecules, quantum mechanics should be used to treat the moving nuclei and electrons building up the matter. The electronic structure in atoms and molecules and molecules gives the first level of properties to all matter around us. Theoretical calculations utilize the so-called Born–Oppenheimer (BO) approximations which keeps the much heavier nuclei in fixed positions while the electrons adapt themselves pair wise in binding and non-binding molecular orbitals (MO), constructed as linear combinations of atomic orbitals (LCAO) using suitable basis sets (normally Gaussian functions or plane waves). Computational quantum chemistry is largely based on solving the Schrödinger equations interactively within the self-consistence field (SCF) theory until suitable convergence criteria are reached. The previously commonly used Hartree–Fock (HF) method was made computationally convenient in matrix form by Roothaan [23] in early 50s and allowed accurate calculations for larger and larger molecules at the same time as computers were made more and more powerful [24]. As there is no exact solution to many-body problems (N > 2) all modelling methods use approximations to reduce the problem to independent particle or pair problem [25]. Within the HF approximation the full N-electron problem is reduced to an independent N single-electrons moving in a mean-field of the other electrons. By increasing the size of the basis set a so called HF limit can be reached where the electron-electron repulsions are treated uncorrelated. As the electron correlation effects are very important for many molecular properties a number of post-HF schemes have been developed, for example Møller–Plesset (MPn ($n = 2, 4, \ldots$)), configuration interaction (CI) or coupled cluster (CC) [25].

The most affordable quantum chemistry methods are the semi-empirical MO methods [26,27,28]. Among the most popular are those based on the neglect of diatomic differential overlap (NDDO) approximation, such as Austin Model 1 (AM1) [29], Parametric Model 3 (PM3) [30,31] and modified NDO (MNDO) [32]. All these schemes start from HF-Roothaan equations [25] introducing the frozen-core approximation and using only the valence AOs. Some electron-electron integrals are simply neglected, some calculated approximately and some are replaced by empirical parameters. These methods are very

fast compared to *ab initio* methods but give often unreliable results; the errors tend to be unsystematic and therefore both relative energies and conformational energies become less reliable when calculated using semi-empirical schemes. Also hydrogen bonds are not accurately described. These shortcomings are found to be largely due to an unbalanced treatment of core-electron and electron-electron interactions. Orthogonalization corrections have been worked out to improve the NDDO schemes [33,34] but have given only a slight improvement. This unbalance is corrected in the recently published Extended NDDO scheme [35] where the old NDDO is expressed as zeroth order term. However, it still remains to be parameterized and packaged to a new semi-empirical scheme, which is a major undertaking.

In both materials science and chemistry, the density functional theory (DFT) has recently gained the position as the most popular method [25]. It is basically as fast as the HF method but can incorporate a large part of the electron correlation effects. In DFT the energy is given as a functional of the electron density $\rho(\mathbf{r})$. The contributions to total energy come from the kinetic energy of the electrons, nuclear-electron attraction, electron-electron Coulombic repulsion and finally from combined exchange and correlation energy. The most difficult problem in DFT is to derive the exchange-correlation (XC) term. In materials science for solids and metals the so called local density approximation (LDA) is used as it is suitable in cases of slowly fluctuating electron densities [25]. LDA is not a good choice for chemical problems and so called gradient corrected methods like the generalized gradient approximation (GGA) are used where the XC term is not just functional of the electron density but also derivatives of the density with respect of the coordinates (x, y, z) [25]. The most popular XC functionals are currently those by Lee, Yang and Parr (LYP) [36] and by Perdew and Wang (PW91) [37]. In chemical applications the hybrid methods where the HF exchange energy is mixed the corresponding DFT term have gained popularity. Becke's three parameter hybrid [38] with the LYP functional, the so called B3LYP, is frequently used. Current DFT methods are based on formulation proposed by Kohn and Sham (KS) [39,40] by constructing the total electron density from a KS orbital. This gives conceptually similar equations as in HF method. Car and Parrinello (CP) presented a DFT based Molecular Dynamics method using plane-wave basis functions [41] which is the most popular *ab initio* simulation tool both in chemistry and materials science.

Recently a robust approximate DFT method is introduced where the electron-electron interactions in the total energy density functional in DFT are expanded with respect to a reference density and keeping the first order correction and neglecting the higher order corrections. This is called Harris–Foulkes functional [42,43] and is stable with respect to the used reference density as a small change in the reference density introduces an error only to second order term. Computational schemes based on Harris–Foulkes functionals are normally called *ab initio* tight-binding density functional theory (AB-TBDFT) methods, and were first introduced by Sankey et al. [44]. A highly efficient AB-TBDFT method is recently presented by Tu and Laaksonen [45,46] applicable on simulations of large systems. Various methods to incorporate quantum mechanics into molecular dynamics simulations are reviewed in [22].

First principles electronic structure calculations and simulations have a firm position in materials science. First-principles MD simulations to materials properties are reviewed in [47]. In the following a few examples are chosen from literature. Materials simulations using VASP gives a quantum perspective to materials science according to [48] using pseudo potentials (PP) and a plane wave basis set based on a finite-temperature local-density approximation. Density functional theory meeting statistical physics [49]: from the atomistic to the mesoscopic properties of alloys by combining DFT calculations with the so-called Cluster Expansion (CE) methods and Monte-Carlo (MC) simulations is reported. Determination of solid-state nanostructure from *ab initio* structure calculations of nano-structured materials using diffraction data in combination with distance geometry methods is explained in [50]. Tools, results and perspectives of quantum software interfaced with crystal structure databases are given [51]. Density-functional theory electronic structure calculations of static and elastic properties and *ab initio* molecular dynamics simulations for poly-atomic systems made possible in package reported in [52] and first-principles computation of material properties using the ABINIT software package in [53]. A primer to efficient tight-binding molecular dynamics is given by Colombo [54]. CONQUEST code for large-scale *ab initio* calculations in materials science is presented in [55]. Prediction of materials properties by *ab initio* computer simulations is reported [56].

Atomistic Molecular Dynamic Simulations

Molecular Dynamics (MD) [57,58,59,60,61,62,63,64,65] is, in its simplest formulation, numerical integrations of Newton's second law (acceleration ($\mathbf{a}(t)$) of a particle is equal to the force ($\mathbf{F}(t)$) acting on a particle divided by the mass ($\mathbf{m}$) of the particle) applied on a collection of inter-

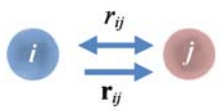

$$U(r_{ij}) = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\ \varepsilon_0\ r_{ij}}$$

$$\sum_{j}^{N} \mathbf{F}_{i<j} = \sum_{j<i}^{N}\left(-\frac{\partial U_{ij}(r_{ij})}{\partial r_{ij}}\frac{\mathbf{r}_{ij}}{|\mathbf{r}_{ij}|}\right) = \quad \mathbf{F}_i = \mathrm{m}_i\mathbf{a}_i = \mathrm{m}_i\frac{d\mathbf{v}_i}{dt} = \mathrm{m}_i\frac{d^2\mathbf{r}_i}{dt^2}$$

$$\mathbf{V}_i = \frac{d\mathbf{r}_i}{dt} \quad \Rightarrow \quad d\mathbf{r}_i = \mathbf{V}_i dt \qquad \mathbf{a}_i = \frac{d\mathbf{v}_i}{dt} \quad \Rightarrow \quad d\mathbf{v}_i = \mathbf{a}_i dt$$

$$\mathbf{r}_i(t+\Delta t) = \mathbf{r}_i(t) + \Delta t\ \mathbf{v}_i\left(t+\tfrac{1}{2}\Delta t\right) \qquad \mathbf{v}_i\left(t+\tfrac{1}{2}\Delta t\right) = \mathbf{v}_i\left(t-\tfrac{1}{2}\Delta t\right) + \Delta t\mathbf{a}_i(t)$$

$$\sum_{i=1}^{N}\frac{1}{2}\mathrm{m}\mathbf{v}_i^2 = \frac{3}{2}Nk_BT \quad \Rightarrow \quad \langle T\rangle = \frac{1}{M}\sum_{n}^{M}\left\langle\frac{1}{3Nk_B}\sum_{i=1}^{N}\mathrm{m}_i\mathbf{v}_{in}^2\right\rangle$$

$$V = L * L * L \qquad \chi = \chi - L * \mathrm{anint}\ (\mathrm{x}*\ 1.0/\mathrm{L})$$

Self-assembled Materials, Figure 1
Molecular Dynamics simulations in a nutshell

acting particles. Integrating the acceleration of each particle gives the velocity ($\mathbf{v}(t)$) and a subsequent integration of the velocity gives the actual position ($\mathbf{r}(t)$) of the particle. Thereby the trajectory of the particle can be determined. The forces momentarily acting on the particles (masses) are calculated as negative spatial derivatives of the potential functions $U(\mathbf{r})$ (force fields) describing the attractive and repulsive interactions between the particles (including the internal degrees of freedom): $\mathbf{F} = -\nabla U(\mathbf{r})$. Normally only the pairwise intermolecular interactions are included as an approximation in classical MD simulations of liquids and solutions. The number of particles (N) in current simulations is from thousands to hundreds of thousands and is a compromise set by the available computing power. Particles are inserted into a cell (box) with a volume V. The shape of the cell is preferably chosen to have translational symmetry so that replicas of the cell make a pseudo-infinite periodic system where particles are synchronously able to leave and enter the cells according to periodic boundary conditions (PBC). This keeps the particle density constant. So called minimum image convention connected to the PBC is applied to avoid artificial condensation effects (for more details see Table 1 and the text in connection to the figure). In Newtonian dynamics the energy (E) is expected to be conserved at equilibrium conditions and the simulation corresponds to a microcanonical (NVE) ensemble. By using suitable Hamiltonians containing thermostats canonical (NVT) ensembles can be constructed and with corresponding barostats isobaric (NPT) ensembles. Several other ensembles can be constructed for MD simulations as well.

Historically, Alder and Wainwright introduced the method of Molecular Dynamics 50 years ago by simulating a system of hard spheres [66,67] observing a phase transition, while Rahman [68] carried out the first simulation for a liquid system using argon atoms in 1964 giving fairly accurate diffusion coefficient. Verlet was able to obtain thermodynamic state functions noble gases [69]. A very important pioneering work for molecular systems was that of Rahman and Stillinger for liquid water in 1974 obtaining radial distribution functions in close agreement with experiments [70], which also started a real boom in computer simulations. MD simulation is a powerful method to study structural, thermodynamic and dynamical properties in gases, liquids and solids. Using MD the motion of molecules can be followed to gain a better understanding of chemical reactions, fluid flow, phase transformations, droplet formation, and other physical phenomena. MD is a statistical mechanical method (both equilibrium and non-equilibrium).

Figure 1 gives all the information needed to write a simple computer program to run MD simulations of N

charged particles in a cubic volume V. *First row*: Pairwise interaction potential consisting of Lennard–Jones (van der Waals) and Coulombic terms. Particle sizes are determined by the σ parameter and strength of the van der Waals attractions is given by ε. Particle charges are given by q in the Coulombic term and pairwise distances are given by r_{ij} (scalar) and the $\mathbf{r}_{ij}$ (vector). *Second row*: Forces acting on each particle are obtained as the negative spatial derivative of the potential function given in first row and are equal to particles mass times its acceleration (Newton's II law). *Third row*: By integrating the fluctuating time-dependent acceleration gives particle's velocity and a subsequent integration of the velocity gives the momentary position of the particle. *Fourth row*: Example of a numerical finite difference algorithm called the "Verlet leap-frog" to carry out the integration. Δt is the time step and is normally of order of a few femto seconds depending of the particle masses and the density of the system. Observe that the velocities in the Verlet leap-frog scheme are obtained as half time steps to increase the accuracy and stability of the algorithm. *Fifth row*: Example of how compute the average temperature from simulation data using the kinetic energy and the equipartition principle (each translational degree of freedom contributes 1/2 kT to the kinetic energy). M is the total number of time steps in the simulation. Other average quantities, for example the total energy (potential+kinetic energy) is calculated similarly. *Sixth row*: For a cubic simulation cell of a volume V (containing the N particles) the side length is L. Periodic boundary conditions (PBC) are easily applied on the positions (to keep the particles inside the cell) by the simple formula where the anint is a function which truncates the floating point number keeping only the integer digit. Example here is given for the x-components and should be done same way for y- and z-components. PBCs are applied on the distances as well according to the minimum distance convention to avoid artificial condensation effects during the simulations. Large scale simulations are carried out using parallel computers with a large number of processors. This requires computer software adapted to parallel architectures [71,72]. A large number of simulation packages are available from the authors developing them. Most of them are publicly available while some (most often those who are made particularly user-friendly by graphical interfaces) require a license to be purchased.

MD is an example of N-particle simulation techniques, a wide-spread method found virtually in all areas dealing with interacting dynamical particles from solvated electrons to protein-lipid interactions and from orientations of liquid crystals to formation of galaxies [58,59,60,61,62,63,64,65]. It is also the main simulation method in materials science from metals to zeolites and from polymers to colloids.

Self-assembled thiol monolayers on Au(111) made hydrophilic with polar end groups are studied with STM and MD simulations [73] to examine the role of the formation of hydrogen bonds between the molecules in the layer and with polar co-adsorbates (water and solvent). MD have been performed [74] to investigate the two-dimensional structure of organosilane self-assembled monolayers (SAMs). Unlike alkanethiol SAMs, the arrangement of molecules in organosilane SAMs is not crystalline. Simulations performed for structures with different bonding networks in the polysiloxane layer shows that the ratio of hydrogen bonds has a profound effect on conformations and strain energies. Water density profile close to spherical and planar hydrophobic objects is studied [75] using MD simulations. A substantial increase of the depletion layer thickness was found with the temperature. High electrostatic surface potential presumably plays an important role in the presence of charged solutes possibly promoting adsorption into the interfacial layer. MD investigations [76] of the morphology and structure of monolayer and multilayer of chiral molecule *N*-stearoy-*L*-glutamic acid (C-18-L-Glu) self-assembled on a mica surface show that hydrogen-bonding effects are the major driving forces in the layer formation proposing a multilayered model for the self-assembling. Static and dynamic properties of 2:1 layered silicates ion exchanged with alkyl-ammonium surfactants are studied using MD simulations [77] providing the structure and dynamics of the intercalated surfactant in agreement with experiments. Structure and properties of self-assembled monolayers (SAMs) of a bi-stable[2]rotaxane on Au(111) surfaces as a function of surface coverage are reported based on atomistic molecular dynamics (MD) studies with a force field optimized from DFT calculations with several experiments validating the predictions [78]. To develop strategies to self-assembly of nanostructured materials MD simulations for telechelic molecules composed of two polyhedral oligomeric silsesquioxane cages connected by one hydrocarbon backbone dissolved in liquid normal hexane were carried out [79]. MD simulations of nanoparticle self-assembly at a liquid-liquid interface are carried out showing in situ formation of clusters and migration of both single particles and clusters from the water phase to the trichloroethylene phase [80].

Among similar modelling works related to self-assembly could be mentioned recent advances of classical density functional theory with emphasis on applications to quantitative modeling of the phase and interfacial behavior of condensed fluids and soft materials, including

colloids, polymer solutions, nanocomposites, liquid crystals, and biological systems are summarized [81]. Molecular mechanical methods are used to calculate strain distribution in self-assembled interfacial misfit dislocation arrays in highly mismatched III–V semiconductor materials in good agreement with experimental data using cross section high-resolution transmission electron micrograph images and also with other theoretical values [82]. Molecular self-assembly at equilibrium is studied using a family of simple directional and short-range van der Waals potentials giving rise to the self-assembly of linear polymeric, random surface, tubular, and hollow icosahedral structures. Dipolar potential with its continuous rotational symmetry about the dipolar axis contributes to chain formation, while higher multipoles lead to the self-assembly of open sheet, nanotube, and hollow icosahedral geometries [83]. To further understand the mechanism behind Layer-by Layer assembly processes requires even theoretical work, including thermodynamics calculations and molecular dynamics simulation as many kinds of physicochemical molecular interactions, including hydrogen bonding, charge transfer interactions, and stereocomplex formation are involved [84].

Monte Carlo Simulations

With modern computers the Monte Carlo technique has evolved from a classical numerical algorithm to solve multi-dimensional definite integrals based on statistical sampling to a very powerful simulation technique to study a large variety of complex systems with interacting particles [59]. The name "Monte Carlo", given by Stanislaw Ulam [85], refers to the capitol of Monaco at Mediterranean, famous for its casinos and roulette tables. The corresponding "roulette" in the MC method is the random number generator supplying a sequence of numbers normally between 0 and 1 from a uniform random distribution (rand[0, 1]). MC methods can be used, for example, to compute equilibrium properties of classical many-particle systems. MC was in fact among the first applications run in the first computers in early 50s. The first liquid simulation using MC was carried out by Metropolis and coworkers [86]. As molecular computer simulation methods MC and MD are very similar until the particles start to move. The simulated systems in both MC and MD are built in the same way by inserting the N particles in a simulation cell of volume V, applying the periodic boundary conditions and choosing a force field to describe the interactions between the particles and the cut-off distances for short and long-range interactions. Ewald summation can be employed if found necessary. By fixing the temperature T we can perform the MC simulation in a canonical (NVT) ensemble which is also the simplest choice in molecular MC simulations. Recall that the NVE (microcanonical) ensemble is the natural choice in the case of Newtonian MD simulations. While the MD simulation method is completely deterministic in the sense that the particles always follow exactly same trajectories if same input is given, the Monte Carlo (MC) method is stochastic and the particles are moved as a random walk. How the particles are moved depends on the studied systems. In the case of small and rigid molecules (as a liquid) the molecules (center-of-mass) is translated (displaced) in x, y and z direction of a small distance one molecule at a time rotating it slightly around its principal axes. For larger and flexible molecules the individual atoms are slightly moved and in addition different types of specific movements are performed mimicking somehow the real motion of the molecule. For example MC simulations of polymers are undergoing residual "flips", "rotations", "crankshaft", "slithering snake" moves and "end-bridging" plus many more. Probably the simplest MC technique to simulate interacting particles at a desired temperature is Metropolis Monte Carlo [86].

The mother of all Monte Carlo methods is the Markovian master equation:

$$\frac{\mathrm{d}P(\vec{r}_i,t)}{\mathrm{d}t} = \sum_{\vec{r}_j} W(\vec{r}_j,\vec{r}_i)P(\vec{r}_j,t) - \sum_{\vec{r}_j} W(\vec{r}_i,\vec{r}_j)P(\vec{r}_i,t) \quad (1)$$

where $P(\vec{r},t)$ is the probability that the system is in state $\vec{r}$ at time t (time-dependent distribution of states/configurations) and $W(\vec{r}_i,\vec{r}_j)$ is the transition probability/rate per unit time for the system to undergo a transition from i-state to j-state. All MC techniques can be seen as methods for solving the master equation. At steady state (stationary or equilibrium):

$$\frac{\mathrm{d}P(\vec{r}_i,t)}{\mathrm{d}t} = 0; \quad t \to \infty \wedge P(\vec{r}_i,t) \to P(\vec{r}_i)$$
$$\Rightarrow \sum_{\vec{r}_j} W(\vec{r}_i,\vec{r}_j)P(\vec{r}_i) = \sum_{\vec{r}_j} W(\vec{r}_j,\vec{r}_i)P(\vec{r}_j)\,. \quad (2)$$

The detailed balance (microscopic reversibility) is valid. The canonical contribution is:

$$\sum_{\vec{r}_j} W(\vec{r}_i,\vec{r}_j)\mathrm{e}^{-\frac{E(\vec{r}_i)}{k_\mathrm{B}T}} = \sum_{\vec{r}_j} W(\vec{r}_j,\vec{r}_i)\mathrm{e}^{-\frac{E(\vec{r}_j)}{k_\mathrm{B}T}} \quad (3)$$

$$W(\vec{r}_i,\vec{r}_t) = \min\left(1, \mathrm{e}^{-\frac{E(\vec{r}_t)-E(\vec{r}_i)}{k_\mathrm{B}T}}\right)\,. \quad (4)$$

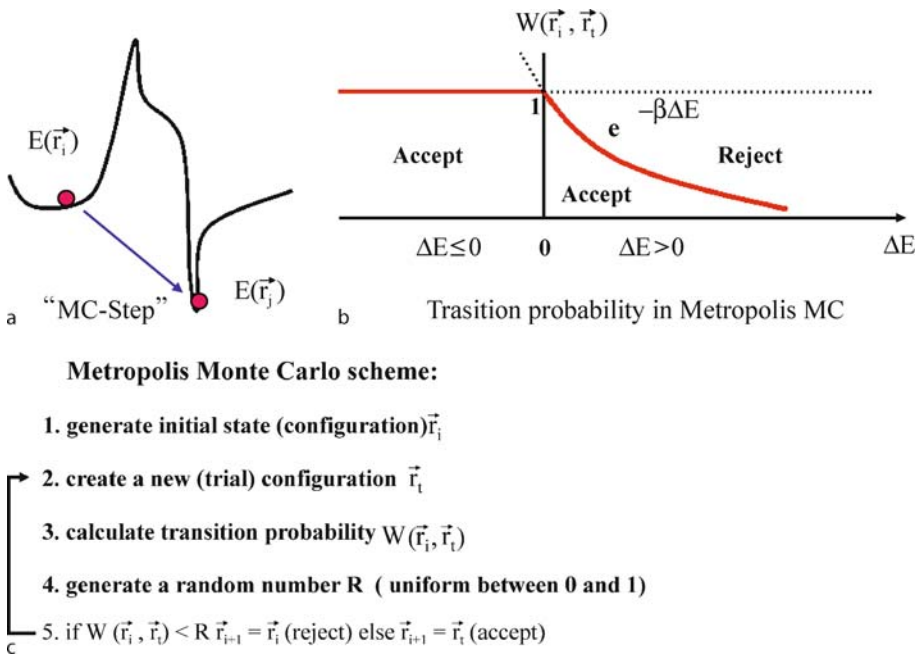

Self-assembled Materials, Figure 2
a Schematic illustration of a Monte Carlo move. b Transition probability *W* in Metropolis Monte Carlo. c Computational scheme for Metropolis Monte Carlo

However, it is not possible to calculate W from this equation. MC methods therefore use different functional forms for W. Transition probabilities in thermodynamical MC (like Metropolis) do not need to have *any* relationship to the dynamics (energy barriers) of the system. To assign real time to MC steps (Fig. 2a) is not possible. Conventional MC methods are only applied to sample systems in (or close to) thermal equilibrium. Functional form of transition probability in Metropolis MC:

In standard Metropolis MC the looping should not be done until accepted. All moves with probability W larger than 1 are accepted (the energy goes down) while if W is less or equal to 1 the move is accepted if the random number is less than the Boltzmann factor in Eq. (4) (see also Figs. 2). A rule of thumb is to choose the size of the displacements in such a way that the acceptance/rejection ration is close to 1/2 which might be considered as the upper limit while the lower limit is roughly 1/3 [87]. Several particles should not be moved simultaneously in Metropolis MC. There is no direct correspondence between a Monte Carlo step and real time but it is possible to use experimental information of for example diffusion to calibrate the MC simulations to obtain a rough estimate of real time.

Some examples of MC works in connection to self-assembly are for example the MC (both multi-scale NVT and grand canonical) simulations carried out to investigate nano patterns an self-assembly of surfactants inside SWCNT systems [88] and Monte Carlo simulations of gold nanocrystals and (111) slabs covered with alkyl thiols are carried out with and without explicit solvent (*n*-hexane) at $T = 300\,\mathrm{K}$. Inclusion of explicit solvent is found important in [89]. Self-organization and chain-forming of large ensembles of nanoparticles is studied by combining DLVO theory and MC simulations [90]. The growth of linear agglomerates is kinetically controlled by a high activation barrier from all of the directions except one at end of the chain. Adsorption of different model amphiphiles in apolar and polar solvents is investigated using MC simulations with a coarse grained model. As coating agents the surfactants with a single hydrocarbon tail or two branches are found to protect better the particle surfaces than amphiphiles with three or more branches [91]. Lattice Monte Carlo simulation is used to investigate the equilibrium between free surfactant molecules in aqueous solution and those adsorbed layers on structured solid. The solid surfaces are composed of hydrophilic and hydrophobic surface regions [92]. Lattice MC simulation of self-assembled ordered hybrid materials is reported and a comparison of structural characterization of the different phases using aggregate size distribution, density profiles, and radial distribution functions [192]. MC simulations are used to study self-assembly of symmetric diblock copolymers

in confined state resulting novel self-assembled strip, circle, core-multishell, and multi-barrel layer structures [94]. Self-assembly and adsorption on hydrophobic surfaces are studied by MC methods in [95,96,97,98].

Kinetic Monte Carlo Simulations

Ideally we would like to have a general computer simulation method operational on atomistic level of matter which we can used for both short and long length and time scales. We wish to use it to simulate general dynamical processes with many different types of transitions between various states the system populates during its evolution towards a more permanent equilibrium state. The method of molecular dynamics (MD) is often the first choice. Using MD we model our system by applying realistic physical conditions. We also have correct real time in our simulation procedure. However, the time step Δt in atomistic MD simulation to numerically solve the equations of motion has to be chosen according to the frequency ($\omega_{\max}$) of the fastest dynamical event in the molecular system this being in practice $\Delta t \ll (\omega_{\max})^{-1}$. This makes atomistic MD simulations very inefficient if the events of interest occur beyond nanosecond time scales. Unfortunately a vast amount of motional modes and dynamical events take place beyond the capability of MD simulations even with the fastest computers around.

We can use Monte Carlo (MC), such as the common Metropolis method discussed above. We can build the molecular system similarly as in MD and impose the desired conditions onto it. We start simulations and pick up the particles randomly and one at a time moving them in small displacements based on the Boltzmann probability. Again, we realize that as the size of the system increases together with the number of motional degrees of freedom our simulation starts to take too long time. Besides, as mentioned in the previous chapter, we do not even know how to add the time into a Metropolis simulation as different events have different characteristic times for their dynamical behavior. MC as we perform it is normally only for sampling ensembles.

We now consider using a method called the kinetic Monte Carlo (kMC). Using kMC dynamical processes can be simulated with real time incorporated and all possible time scales can be covered easily. This sounds simply too good to be true so what is the "catch" here? Before going into details of kMC method (in fact there are many variants of it) we try to trace its origin by going back almost four decades ago and to computer simulations of vapor deposition on two-dimensional lattices [99] and crystal growth with surface diffusion [100]. These two works

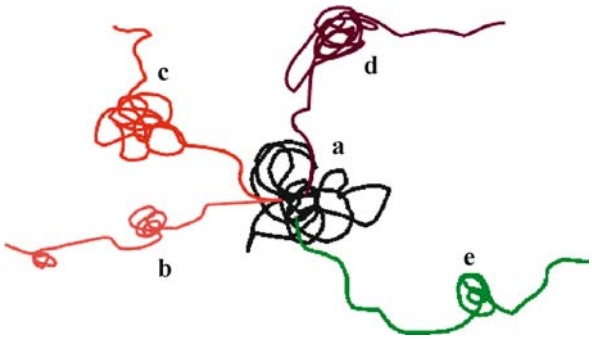

Self-assembled Materials, Figure 3
Accessible stationary states (b–d) from state a

carry much of the basic underlying idea behind several kMC schemes developed later. The paper by Bortz, Kalos and Lebowitz (BKL) presents elegantly the theory and algorithms for a method finding immediately a strong position among physicists [101]. The BKL scheme is the solid framework in many kMC programs used today. In almost parallel and from purely chemical point of view Gillespie presented his stochastic simulation method for coupled chemical reactions [102] with the starting point being that on atomistic level the time evolution of chemical systems such as chemical reactions can be seen as both discrete and stochastic rather than continuous and deterministic as the text book kinetic equations are presented for us. Although the BKL method and the Gillespie method appear to have been developed in separate communities they both lead to same family of kMC methods [103]. The name kinetic Monte Carlo appears first as late as in 1992 [104].

Kinetic Monte Carlo is also based on solving the Markovian master Eq. (1) discussed in the previous Chap-

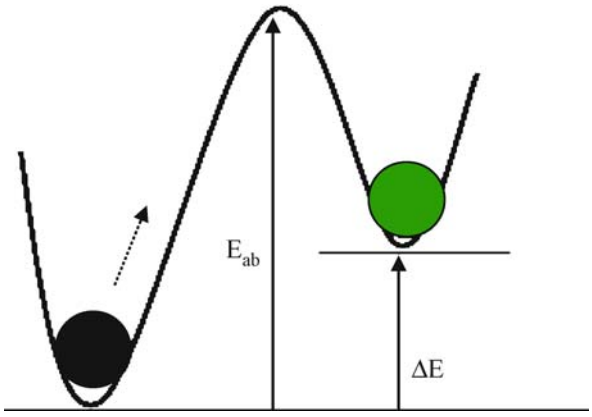

Self-assembled Materials, Figure 4
Transition from state a to state b

ter. In kMC we assume that the system is in a (stationary) state and that a number of other states are accessible from the current state (Fig. 4). A new state is randomly chosen among the neighboring states and a transition is made to this new state. The real time spent in the state before transition is added to the time development and can be seen as the time step in kMC. The time spent in a state (Fig. 4) is related to the transition rate:

$$r_{ab} = \upsilon e^{-\frac{E_{ab}}{k_B T}} \tag{5}$$

– where υ is the "attempt frequency". The time step is then calculated as:

$$\Delta t = -\frac{\ln(\mathrm{rand}[0,1])}{\sum\limits_{i=b,c,d,\ldots} r_{ai}} \tag{6}$$

kMC requires that the simulated events are Poisson processes. There should also be a dynamic hierarchy among the processes. The real time increments (time steps) should be able to be calculated based on Eqs. (5)–(6). In practice the detailed balance (Eq. (2)) is not strictly required as the systems can be far from equilibrium. If the transition probabilities are: (i) independent on previous history, (ii) the same at all times and (iii) a uniform function of time, the the transitions are Poisson processes. A kMC transition should not have any effect on any other transition. Also the time when the next transition will occur should be independent from the occasion of the previous transition. To illustrate kMC we may consider a "wheel of fortune" (Fig. 5) where the sectors correspond to individual Poisson processes (events) and are not necessarily evenly divided. The perimeter of the wheel is the cumulative process (also a Poisson process). The corresponding dynamical hierarchy in Fig. 5 is as:

$$[0 \leqslant r_5/r_{\max} \leqslant r_8/r_{\max} \leqslant r_1/r_{\max} \leqslant \cdots \leqslant r_{\max}/r_{\max} \leqslant 1]\,.$$

The whole spectrum of rates exists and the fastest rate (r_2 in Fig. 5) gives the highest probability. Note, the Metropolis MC does not fullfill dynamic hierarchy as all $\Delta E \leqslant 0$ transitions are given probability 1. After a kMC time step we are in a new state and we have a new "wheel" like the one in Fig. 5 but possibly with a different number of sectors corresponding to the new states with a finite probability to make transitions to. A new ensemble of processes is obtained after each transition. This also means that the time step in kMC changes continuously from one step to another. There are no restrictions how short or long the time step is: it can be for example a few femtoseconds or several hours. It simply follows the evolution of the system during the simulation. Time steps can be shorter in the beginning of simulation due to (initially) fast moving objects. Also, slow moving objects may break to smaller ones moving faster. Kinetic MC sounds like the optimal simulation method which can cover all possible time and length scales. The "catch" here is that the *user* has to supply both the states and the transition rates to the new states after each time steps. The kMC is not a self-going scheme in the same way as the other simulation methods are. This of course limits its use currently. Transitions rates can be obtained for example from quantum chemical calculations, MD simulations, transition state theory and experiments. As all the information of the system and rates is available the method has many attractive features: No thermodynamical equilibration is needed and no simulation of "dead dynamics" like in MD (event/step is always guaranteed). kMC is suitable for "driven systems" (irradiation, high pressure, etc.). kMC has been used in studies surface diffusion, crystal growth, molecular beam epitaxy growth, defect mobility, vacancy diffusion, reactions of surfaces, diffusion in zeolites, clustering dynamics, nucleic acid hybridization, protein folding, etc. The number of new application areas is growing rapidly. kMC can be easily incorporated in multiscale simulations schemes.

Kinetic MC is used to determine the critical layer thickness for misfit dislocations using deposition flux, temperature and a pairwise interaction potential between the particles as parameters [105]. Kinetic MC simulations are used to study the deposition rate dependence of nanopattern formation on periodically strained surfaces. The optimum nanopattern quality depends on surface strain field, temperature and deposition rate and the amount of material deposited [106]. Self-assembly of the

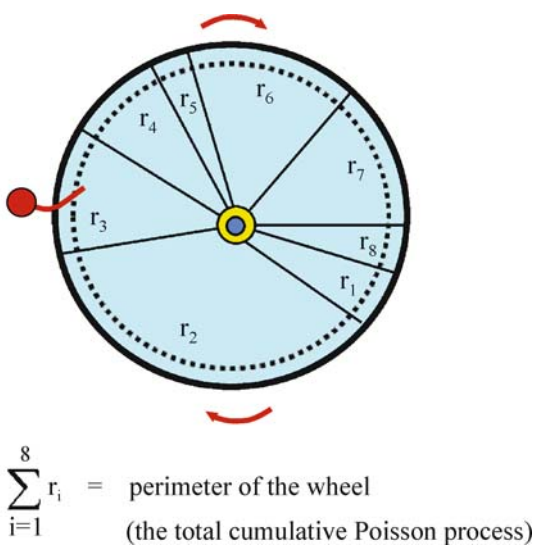

Self-assembled Materials, Figure 5
The Poisson "wheel of fortune" to illustrate the kMC

elementary building blocks of nanophase materials are studied using both a kinetic mean-field model and a mesoscopic approach in which self-assembly is viewed as an encounter-controlled process on a discrete lattice [107]. A 3D kMC model with a Green's function-based long-range strain energy contribution and with an up-down ratio for atoms to jump out of the plane of the surface is developed to simulate the growth of self-assembled quantum dot islands [108].

Phase-Field Modeling

Classical (macroscopic) theories & models assume sharp interfaces (infinitely thin with a specific surface tension) between phases and thermodynamical variables such as T, P, concentration, etc. are used independently in each phase obeying the phenomenological rules (schematically shown in Fig. 6). Phase diagrams are common tools in chemistry text books and traditionally the compositional and structural evolution of the phase regions have been treated mathematically using distinct interfaces with suitable boundary conditions at the interfaces. The idea of a *sharp* interface is old going back to early 19th century and Young, Laplace and Gauss. Gibbs later presented the interface as a dividing surface so that the properties were distinctly different on both sides of the interface (Fig. 7). Subtraction of the upper curve (real system) from the idealized system below gives the corresponding excess quantity (here concentration) and sharp mathematical surface line. The Gibbs dividing surface has zero volume but non-zero excess quantities.

Also the idea of a *diffuse* continuous interface (Fig. 8) goes back to 19th century when it was thought that gas and liquid states were simply distant points of the same condition of matter and it was possible to move from one phase to another along a continuous path. This picture was established both experimentally by Andrews and theoretically by Thompson and van der Waals [109]. Compared to singularly sharp interface model there are many benefits with using a diffuse inteface with a finite interface volume. All field variables and corresponding equations can be defined simultaneously for the whole system. Diffuse interface allows modelling of various types of physical variables, systems and effects in a flexible and efficient way. Diffuse inteface exists on atomic scale as the surface layers are rough on atomic scale rather than mono-molecular. The density profile at the interface becomes larger than the atomic length scale due to the roughness.

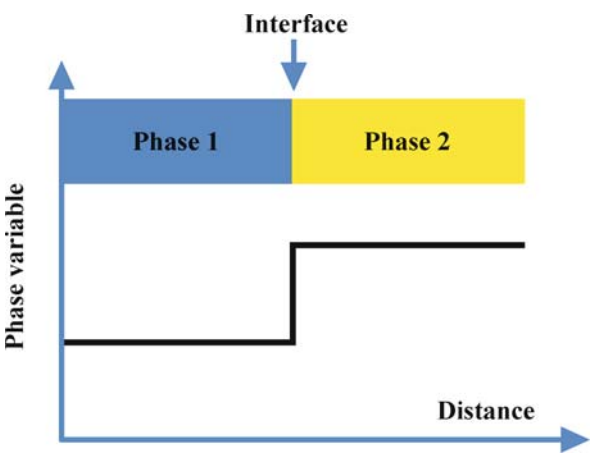

Self-assembled Materials, Figure 6
Sharp limit between two phases

Phase field modelling gathers all categories of diffuse interface models to describe a wide variety of materials phenomena (for excellent reviews, see [110,111,112,113,114,115,116,117]). It can describe easily a large number of phenomena from nucleation to solute trapping as different physical effects can be easily included [115]. It is based on two assumptions: (i) There exists a continuous phase field variable $\phi(x, t)$ which characterizes the phases and interfaces of the system at any point point in space and time. In bulk phases it is given a constant value. (ii) The total (Helmholtz) free energy, $F[\phi, \ldots]$ can be given as a functional of the phase field variable ϕ together with any other thermodynamical variable, for example temperature, concentration etc. Even other variables can be introduced depending on the studied system. A general form of F is given as [113]:

$$F = \int \Big[f(c_1, c_2, \ldots, c_n, \eta_1, \eta_2, \ldots, \eta_p) + \sum_{i=1}^{n} \alpha_i (\nabla c_i)^2 + \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{p} \beta_{ij} \nabla_i \eta_k \nabla_j \eta_k \Big] \mathrm{d}^3 r + \iint G(r - r') \mathrm{d}^3 r \mathrm{d}^3 r' \tag{7}$$

c denotes *conserved* variables (satisfying local conservation conditions)
η denotes *non-conserved* variables.
f is local free-energy density,
α and β are coefficients for energy gradients.

The first integral in Eq. (7) contains the local short-range interactions to the free energy, while the second integral represents the non-local long-range interactions, such as, elastic, electrostatic, polar, etc. Various phase-field models differ in their ways to treat all these contributions

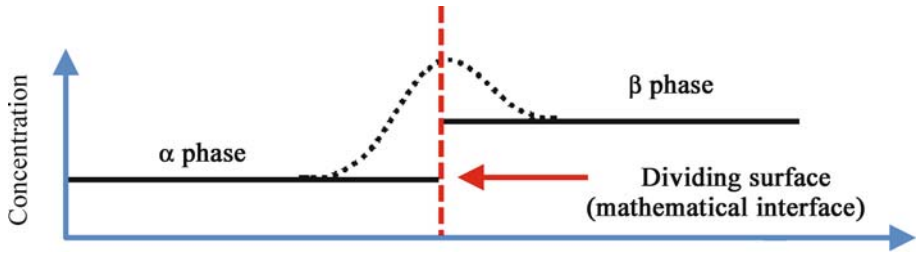

Self-assembled Materials, Figure 7
Gibbs dividing surface

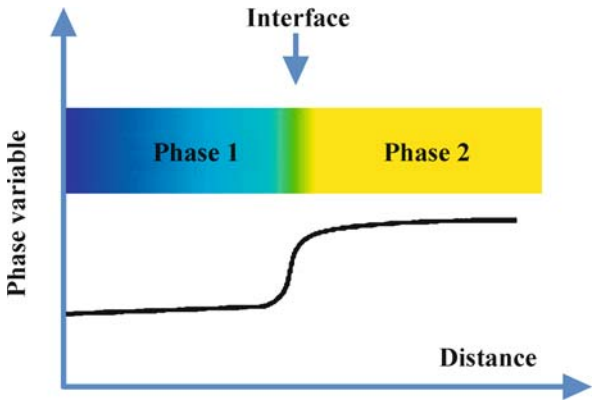

Self-assembled Materials, Figure 8
Diffuse continuous interface between two phases

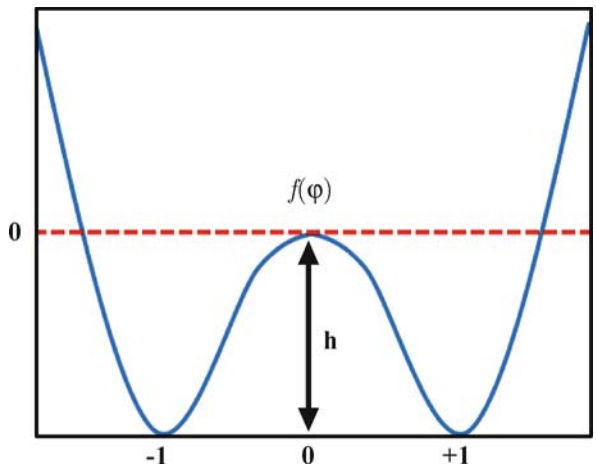

Self-assembled Materials, Figure 9
Simple double-well free energy density function

to the total free energy functional. The local free energy f is the most important term in the phase field models:

$$f = f(c_1, c_2, \ldots, c_n, \eta_1, \eta_2, \ldots, \eta_p) \,.$$

A large variety of these functions have been proposed. In many phase-field models (like for solidification) simple double-well potential functions are used for local free density like the one given in Eq. (8) and in Fig. 9:

$$f(\varphi) = 4h\left(-\tfrac{1}{2}\varphi^2 + \tfrac{1}{4}\varphi^4\right) \,. \tag{8}$$

The variable ϕ in Eq. (8) is normally called order parameter. It varies smoothly from one phase to another, distributing interfacial forces and other interactions over the interface region [8]. We can often use it in a simple double-well potential for a case where the values $\phi = -1$ and $+1$ represent liquid and solid state, respectively, h is the barrier height between the two states. We may also have the phase field $\phi = -1$ and $+1$ representing two different heterogenous mixtures during an isostructural (spinodal) decomposition while 0 is a single homogeneous phase and h is the driving force for the transformation. Other types of similar simple local free energy density functions can also be found in literature.

Some other forms of local free energy profiles are for example the "double-obstacle" potential:

$$f(\varphi) = h(1-\varphi^2) + I(\varphi)$$
$$I(\varphi) = \begin{cases} \infty & \text{if } |\varphi| > 1 \\ 0 & \text{if } |\varphi| \leqslant 1 \,. \end{cases} \tag{9}$$

Or a "crystalline" potential having infinite number of minima:

$$f(\varphi) = h\sin(\pi\varphi) \,. \tag{10}$$

Often additional phase field parameters like temperature are needed, T = temperature, T_m melting point and α is a positive constant:

$$f(\varphi, T) = 4h\left(-\tfrac{1}{2}\varphi^2 + \tfrac{1}{4}\varphi^4\right) + \frac{15\alpha}{8}\left(\varphi - \tfrac{2}{3}\varphi^2 + \tfrac{1}{5}\varphi^4\right)(T - T_\mathrm{m}) \,. \tag{11}$$

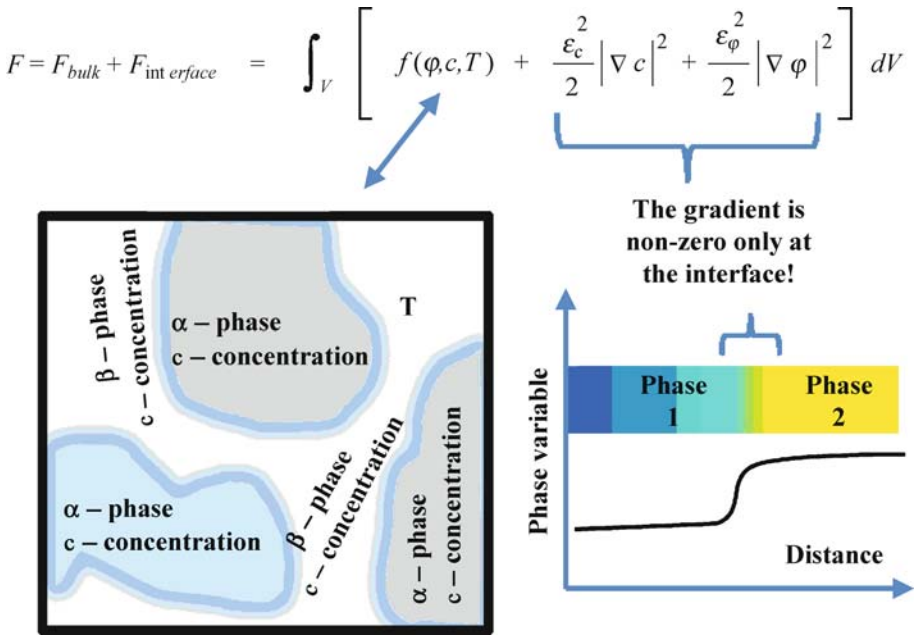

Self-assembled Materials, Figure 10
Two main contributions to free energy: bulk and interface

Often coupling of phase field variables are needed – like in the following function to describe the spatial distribution for grain growth (for grains with different orientations):

$$f(\varphi_1, \varphi_2, \ldots) = 4h\left(-\tfrac{1}{2}\sum_i \varphi_i^2 + \tfrac{1}{4}\sum_i \varphi_i^4\right) + \alpha \sum_i \sum_{j>i} \varphi_i^2 \varphi_j^2 \,. \quad (12)$$

When $\alpha > 2h$ (in Eq. (12)), infinite number of of minima are located at $(1, 0, \ldots)$, $(0, 1, \ldots)$, $(-1, 0, \ldots)$, etc., representing possible orientations of grains in a polycrystal. Functions in Eqs. (9)–(12) are discussed in [113] where several other types local free density functions are found. In general there are a large number of free energy functions like those above, all designed for specific applications and studies of phenomena. For many studies like solid-state transformations the energy functions are made of well-defined physical order parameters and use required symmetry operations of the studied high temperature phases. In general, the interfacial energies should be seen as anisotropic due to the crystalline feature of solid phase [113]. Multiple order parameter models have been proposed using both thermodynamical and geometrical formulations [114].

There are two main contributions to the total free energy; bulk and interface as illustrated in the example in Fig. 10.

Consider the isothermal process from the example in Fig. 10 with a decreasing total free energy F. At equilibrium, if the gradient coefficients are constants, the variational derivatives of F satisfy following equations [114]:

$$\frac{\delta F}{\delta \varphi} = \frac{\delta f}{\delta \varphi} - \varepsilon_\varphi^2 \nabla^2 \varphi = 0 \quad (13)$$

$$\frac{\delta F}{\delta c} = \frac{\delta f}{\delta c} - \varepsilon_c^2 \nabla^2 c = \text{constant} \,. \quad (14)$$

The Eq. (13) is for non-conservative variables while Eq. (14) is for conservative variables. Equation (14) is constant as the amount of solute is constant in the volume (concentration is one of the conserved quantities).

Time evolution of the phase field is assumed to be proportional to variation of the total free energy functional with respect to the order parameter ϕ [110]: $\partial\varphi/\partial t \propto L(\delta F/\delta\varphi)$. Where L = partial differential operator ($L(0) = 0$). The phase field equations are given:

Cahn–Hillard equation [118,119]:

$$\frac{\partial c}{\partial t} = \nabla M \nabla \frac{\delta F}{\delta c} = \nabla M \nabla \left(\frac{\delta f}{\delta c} - \varepsilon_c^2 \nabla^2 c \right) \quad (15)$$

– for conserved variables such as: concentration, molar fraction, mass, etc.

Allen–Cahn equation [120,121]:

$$\frac{\partial \eta}{\partial t} = -L \frac{\delta F}{\delta \eta} = -L \left(\frac{\delta f}{\delta \eta} - \varepsilon_\eta^2 \nabla^2 \eta \right) \quad (16)$$

– for non-conserved variables such as: order parameters, phase fields, etc. This equation is also known as the time-dependent Ginzburg–Landau equation.

The evaluation of the phase field variables can be obtained by solving the kinetic Eqs. (15) and (16) above. Finite difference methods with explicit time-stepping are normally used by constructing a uniform grid. Discussion of numerical methods and stability of the computational schemes are given in [122].

The areas of applications of phase field simulations are increasing rapidly. Solidification and solid-state transformations are examples of the use phase field modelling as well as coarsening, grain growth and dislocation dynamics. A list of several applications can be found in Table 1 of [113].

A further development of phase field theory, called phase-field crystal (PFC) method, was present by Elder and Grant [123] for elastic and plastic deformations, free surfaces and multiple crystal orientations.

Continuous phase field model by Lu and Suo have been found applicable on self-assembly and monolayer growth on surfaces describing all required components: the phase separation, phase coarsening and phase refining processes [112]. Lu and Salac [124] have developed a phase field model to simulate pattern formation due to electric dipole interactions among adsorbate molecules revealing a unique self-assembly behavior. Wang [125] has developed a diffuse interface field approach (DIFA) which is capable to explicitly describe short- and long-range interactions and treat arbitrarily shaped particle packing processes and other related applications.

Molecular dynamics simulations can be combined with phase field modelling. Two basic approaches have been used where MD simulations carried out at the interface region are connected the phase field simulations using finite element methods [126,127] or where (moving region) MD simulations are carried out at the interface region to supply parameters, such as the strain response, kinetic coefficient (velocity vs. under-cooling behavior) and diffusivity, to the phase field model [128]. As the microstructure in many phase-field models is assumed to be coherently uniform, which is, in many cases incorrect, atomistic MD simulations may be of help in supplying anisotropic information about structure and dynamics of interface regions. A very interesting theoretical model for binary crystal nucleation was developed by Gránásy and coworkers [129,130,131,132,133,134,135,136,137,138,139], and incorporated into simulations and applied on Lennard–Jones and ice-water systems [129,130,131], polymeric dendrites [132,134,136], hydrate formation by Gránásy, Kvamme and coworkers [133,135,137,138] and heterogeneous crystal nucleation [139]. Davidchack and Laird [140,141] have used MD simulations with hard spheres to study interfacial planes between crystal and fluid-like domains.

Plapp and Karma have proposed a robust hybrid multi-scale scheme applicable in one, two and three dimensions by combining diffusion Monte Carlo (similar to the one used to solve Schrödinger equation) and phase-field models to perform simulations of dendrictic solidification. The method is applicable for other applications than solidification [142]. General critical discussion of microstructure modelling is given in [143] highlighting phase field and cellular automata methods. More examples of multi-scale methods will be given below in a dedicated chapter.

Random Sequential Adsorption

Random sequential adsorption (RSA) is a family of models for a variety of phenomena in chemistry, physics, biology, ecology, sociology and in many other fields as it is a general and highly interesting mathematical and physical packing problem [144]. It is applicable on all kinds of moving particles from atoms to motor vehicles arriving randomly as a Poisson process selecting a free space based on a distribution of sites or slots. Particles become bound if they find a vacant place (with no overlap with other particles already occupying the actual space), or they leave the system in case of no luck (possibly to try again). The process continues for a certain period of time until the distribution of sites becomes saturated (complete packing) following the random packing scheme, or if no more particles are arriving although there may still be sites available. For excellent reviews see [145,146,147,148,149]. Colloids and Surfaces A, vol. 165 is a special issue containing several articles of RSA. The one-dimensional RSA model is known as the "car parking" problem as it can be seen as an illustrative and familiar analog. Imagine a long street with apartment houses where people and families live. During the day (office hours) there is plenty of space to park but in the evening the parking spots become quickly occupied (adsorption). In the next morning cars one after another are driven away (desorption). The solution of this 1D RSA problem would give us an answer of how many drivers on average (driving similar cars) can park at random positions along a street of a certain length? This can be solved analytically if the cars are equally long and the parking slots are discrete. For continuous space along the street the exact solution was given by a Hungarian mathematician Alfréd Rényi [150]. Higher dimensional problems are complicated and approximate solutions are given

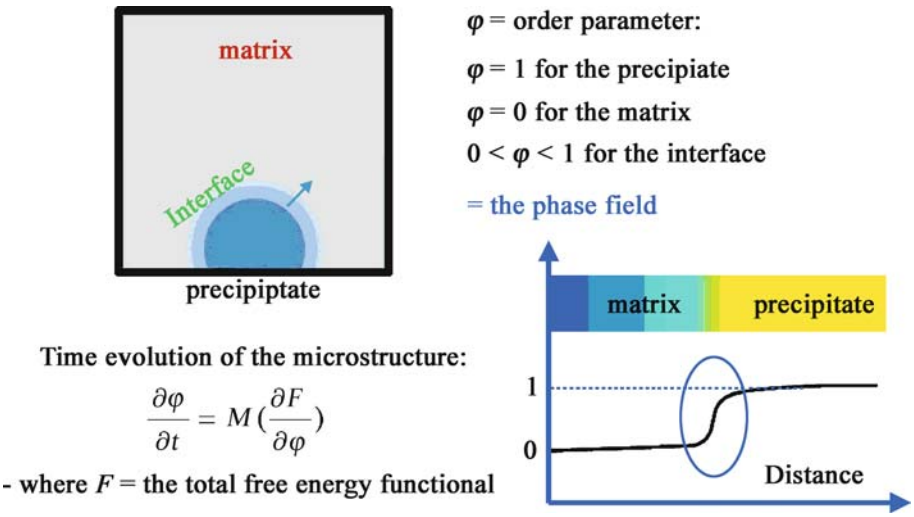

Self-assembled Materials, Figure 11
A phase field model for a growth of a precipitate

based on series expansion, or as a sequential Markovian process solved using Monte Carlo simulations although the detailed balance does not apply for irreversible processes.

The two dimensional RSA is an important and reliable model for molecules adsorbing on a surface reversible or binding irreversibly (with covalent bonds) to specific sites on the surface. Molecules can typically be monomers, dimers (occupying two neighboring sites which are vacant). Polymers can be adsorbed if they find successive empty slots on the surface. Various rules can be imposed on gas molecules approaching already occupied site. The interactions can be both specific and non-specific depending on the studied systems. Orchestration of van der Waals, electrostatic, hydrophilic/phobic interactions as well as hydrogen bonds covalent and ionic bonds leads to adsorption of molecules onto surfaces (consisting of atoms and molecules). Collective sequential adsorption (CSA) is a further generalization of RSA taking into account the local environment around the adsorption site [146]. Even large biomolecular systems (e. g. proteins, DNA, colloids, cells) can be coarse-grained and treated with RSA methods. RSA have been used even for translocation of biopolymers through pores [151] and nucleosomes on a stretched single strand DNA [152]. Generalized RSA methods can used to simulate multilayer build-up in a self-assembled layer-by-layer (LbL) process [153]. Good agreement was obtained between the simulations and experiment for colloids and poly-electrolytes suggesting that the method could be extended to polymers and proteins. Kinetics, jamming limit and structural phase behavior of polydisperse tethered nanoparticles are studied by Gray et al. [154]. Dynamics of self-assembled polyelectrolyte multilayers on glass substrates have been studied by Breit and coworkers [155]. The kinetics could be quantitatively understood using a RSA model for the buildup of a film consisting of polyelectrolyte disks with polydisperse sizes. RSA types of methods in modelling of self-assembled monolayers of charged colloidal particles are reviewed in [156]. Erban and Chapman have presented a diffusion driven random sequential adsorption simulation model operating in real physical time containing features for both reaction kinetics between the molecules and the surface and geometrical constraints of the molecular surface. It is illustrated by an assembly of reactive polymers on a virus surface [157]. Pre-patterning is used as a tool in RSA-based lattice Monte Carlo simulations to improve the self-assembly in nano and micro-scale structure engineering by Cadilhe et al. [158]. RSA model is applied for gelatin self-assembly in binary mixture of water and ethanol [159]. Short-time and long-time kinetics in colloidal adsorption to a monolayer are studied by combining RSA model and Brownian dynamics by Gray and Bonnegaze [160] in the case of self-assembly to meso-structures.

Cellular Automata

The Cellular Automaton (CA) concept was created by John von Neuman and Stanislaw Ulam more than a half century ago to originally solve problems in evolutionary biology [161]. CAs are discrete dynamical systems

where space and time variables and all properties have finite discrete values. CAs start from simple identical (or nearly identical) individual systems interacting locally but evolve to complex systems following simple rules in a synchronous manner. CAs are similar to Petri nets [162] in providing a less mathematical way to differential equations and calculus. Other similar automata are the Turing systems, presented by Alan Turing in 1952 [163] showing how a simple mathematical model can describe spontaneously spreading of reacting chemical specie giving stationary concentration patterns. Turing patterns can also be found as stripes on zebras and spots on cheetahs and elsewhere in nature. Maybe the most widely known application of CAs is the tic-tac-toe like "Game of Life" by John Conway from 1970, a very simple model simulating birth and death of cells in a square lattice interacting according to a set of Boolean conditions. It was easily adapted from paper sheets to computer screen. Dynamic Cellular Automata (DCA) by Wishart and coworkers [164] is a simulation approach allowing a Brownian type of stochastic motion of individual molecules. DCA is more flexible as it allows several cells/molecules to move in a single step. Lattice gas automata are another common use of CA which together with the Boltzmann equation and kinetic theory has led to development of Lattice Boltzmann simulations [165], a powerful method to do fluid dynamics and to solve the Navier–Stokes equations in a from-bottom-up way. The obvious drawback in the current CA methods is that the objects do not have realistic features.

The quantum-dot cellular automata (QCA) approach is proposed as an alternative to molecular electronics. Theoretical behavior of QCA (arrays of Coulomb- coupled quantum-dot cells) is studied in [166] and the state of each QCA cell is determined by its interaction with neighboring cells through the Coulomb interaction and demonstrated experimentally [167] and the field was reviewed [168,169, 170]. A single-molecule implementation of a QCA cell is presented combined with *ab initio* electronic structure calculations [171] using simple prototype molecular systems with a molecule in which charge is localized on specific sites and can tunnel between those sites while the role of the dots is played by redox sites, with tunneling paths provided by bridging ligands. Possibilities to coarse-grain cellular automata, emergence, and the predictability of complex systems are studied by Israeli and Goldenfeld [172]. As molecular self-assembly is driven by local, short-range forces and therefore the dynamics is solely based on local interactions and as atomistic simulations of self-assembly become quickly complicated, a cellular automata based simulation, in which data structures, representing different molecular entities such as water and hydrophilic and hydrophobic monomers, share locally propagated force information on a hexagonal, two-dimensional lattice is proposed, with the purpose of this level of description to gain insight about entropy-driven processes in molecular many-particle systems [173].

Nilsson and Rasmussen present a lattice gas technique for simulating molecular self-assembly of amphiphilic polymers in aqueous environments, where water molecules, hydrocarbons tail-groups and amphiphilic head-groups are explicitly represented on a three dimensional discrete lattice [174]. A Cellular Automata Model of Diffusion in Aqueous Systems show that lipophilic solutes diffuse faster than do polar solutes (in agreement with experiments) [175]. Cellular automaton and MC simulations for the case of CO oxidation on a catalytic surface with simultaneous adsorption, reaction, and diffusion, including the Eley–Rideal step in the reaction mechanism is presented [176]. As properties of diffusing species in microporous materials are strongly influenced by the confining framework, providing the energy landscape for the transport process. Because of the simple topology and the cellular nature of the cages of a zeolite Demontis et al. [177] suggest that it is appropriate to apply to the study of the problem of diffusion in tight confinement a time-space discrete model such as a lattice-gas cellular automaton.

Multi-Scale Modeling

Molecules are moving in a world where "nanometer" could be used as a standard length unit and "femtosecond" is a convenient time unit. For us humans "meter" and "second" are the familiar units to which we all can relate our daily activities. In measuring the time from the big bang and distances between galaxies, both meter and second turn out immediately not to be very convenient to use. Obviously phenomena occur at different length and time scales. Connecting time and length scales has rapidly become a very vital area in from-bottom-up strategies. It will certainly still take quite some time for multi-scale modelling of matter to mature. Below we give examples of both concurrent and hierarchical approaches to multiscale modelling, the two main approaches.

"Concurrent" multiscale modelling can be accomplished by making geometric network of grid points describing the system at a detailed level successively coarser while losing more and more details. For example a crystal could be described with atoms or molecules in crystal sites connected together in 3D net. By applying a scheme with successive coarsening of connecting grid points approaching a continuum description can be coupled to the atomistic scheme by finite element methods [178]. A spe-

cific hand-shake region is created between the both scales. This is used to model cracks and fractures both at atomistic and macroscopic scale.

"Hierarchical" multiscale modelling can be performed by successively reducing the degrees of freedom which are not needed in the next less accurate description of the studied system. This makes the modelling faster and allows longer length and time scales to be covered. A powerful method called the Inverse Monte Carlo (IMC) was presented by us [179] where the "Inverse Problem" was solved completely to obtain interaction potentials (force fields) from radial distribution functions (RDF). The IMC constructs effective potentials consistent with the original RDFs. It can be used for example by starting from *ab initio* MD simulations and construct atomistic interaction potentials to be used in classical all-atom simulations. In all-atom simulations IMC can construct coarse-grained potentials for example by grouping together functional groups in large molecular systems and eliminate the solvent by incorporating it into the coarse-grained potentials. These coarse-grained potentials reproduce the RDFs in the all-atom simulations carried out with explicit solvent. This scheme can be used even beyond this kind of meso-scale description. There are other related methods presented [180,181].

Innovative computational schemes for multiscale modelling are reported continuously based on varying strategies. Examples below are given of studies (i) to couple first-principles simulations and other quantum schemes (ii) to use results from atomistic simulations to carry out subsequent coarse-graining (iii) to connect atomistic schemes with continuum models (iv) to apply kinetic Monte Carlo and other kinetic schemes subsequent to particle simulations, as well as (v) other combinations of various methods discussed earlier in this article are listed below, including a beautiful text-book example (vi) of multi-scale modelling of human hair.

(i) Transparent interface is created between classical molecular dynamics and first-principles molecular dynamics combining first-principles Born–Oppenheimer local spin density molecular dynamics (BO-LSD-MD) with classical molecular dynamics in [182]. Effective quantum mechanical classical mechanical (QM/CM) partitioning method for multi-scale modeling is proposed in [183]. Quantum mechanics at the core of multi-scale simulations with neglect of diatomic differential overlap theory (used in common semi-empirical schemes such as AM1, PM3 and MNDO) is considered in [184]. Potential parameterization based on extended model systems where the force data is calculated from QM methods on a limited range of applications is shown to be essential for a consistent and ultimately, predictive embedding approach to concurrent multi-scale materials simulation in [185].

(ii) Automatic coarse-graining of polymers by deriving effective potentials for multi-atom units or super-atoms from atomistic simulations is reported in [186]. Hierarchical multi-scale modelling of plasticity of submicron thin metal films allowing modelling of thicker films a discrete dislocation model of "diffusional creep" is presented in [187]. Materials modeling platform bridging the molecular characteristics of polymers with macroscopic properties is given in [188]. Linkage between atomistic and meso-scale coarse-grained simulation based on model where the force acting on the CG particles is divided into the mean force (calculated from constrained MD simulations), friction force and random force is discussed in [189,190]. Combined atomistic and meso-scale simulation of grain growth in nanocrystalline thin films with input materials parameters obtained by MD simulation in [191]. Fuzzy clustering approach to hierarchical MD simulation of multi-scale materials phenomena by combining a hierarchy of sub-dynamics: (i) rigid-body cluster dynamics for global conformational changes; (ii) implicit integration of Newton's equations for the coalescence of the clusters; and (iii) normal-mode analysis of fast atomic oscillations is used to facilitate the seamless integration of the multiple levels of abstraction in [192]. A multi-scale simulation of tungsten film delamination from silicon substrate where MD simulations of a single crystal *W* block under tension are used to calibrate a new decohesion model to investigate the effect of specimen size and loading rate on the material properties in [193].

(iii) Coupling of atomistic and continuum models in computational materials science using finite element methods is studied in [194] and finite-element method multi-scale atomistic-continuum modelling of crack propagation in a two-dimensional macroscopic plate by coupling the crack dynamics at the macro scales and nano-scales via an intermediate meso-scale continuum is carried out with molecular dynamics simulation driving the crack tip forward in [195].

(iv) Molecular simulations in zeolitic process design using configurational-biased Monte Carlo simulations combined with transition rate theories kinetic Monte Carlo and molecular dynamics simulations [196].

Multi-scale modeling of hydrogen isotope diffusion in graphite is carried out where molecular dynamics calculations resolve the microscopic length scale and deliver reliable input data for kinetic Monte Carlo calculations (jump frequencies, migration energies, jump step-sizes) from meso-scale up to the macroscopic system length [197]. An MD study of the carbon-catalyst interaction energy for multi-scale modelling of single wall carbon nano-tube growth is presented [198]. A multi-scale approach to radiation-induced segregation at various grain boundaries based on a new rate equation model and MD simulations is reported in [199]. Multi-Scale Computational Framework by integrating a Computational Fluid Dynamics software, a Kinetic Monte Carlo solver and an MD simulator for the self-assembly of atoms into molecular structures is presented in [200]. A novel polycrystalline thin film growth simulator with an atomic level one-dimensional kinetic lattice Monte Carlo model and a real time feature scale two-dimensional facet nucleation and growth model is demonstrated in [201]. A multi-scale atomistic study of the interstitials agglomeration in crystals using a hierarchy of atomistic approaches: the tight-binding molecular dynamics, molecular dynamics based on environment dependent inter-atomic potentials and a lattice kinetic Monte Carlo (LKMC) I reported in [202]. Computer simulations, spanning across different time and length scales, are used to study thin film growth morphology in organic self-assembled monolayers using thiophenes on gold. *Ab initio* calculations created a catalog of the energetics in vacuum and interactions in three orthogonal orientations to a Au (111) surface in [203]. This information was supplied as the input for kMC simulations of dimer and trimer representations of small organic molecules to describe both sub-monolayer and multilayer growth. Finally, MD studies were used to understand the packing structures of stable polymorphs of thiophene SAMs.

(v) Studies of behavior and structure of charged surfaces on different length scales is carried out using a different computational schemes yielding different levels of description of charged surfaces. (a) Coarse grained MC simulations of idealized surfaces incorporate large-length-scale fluctuation and correlation effects in the counter-ion cloud at a charged surface. (b) Brownian dynamics simulations of more realistic and structured surfaces give modified counter-ion distributions, also allowing estimations of mobility of counter-ions at charged surfaces. (c) All-atomistic MD simulations reproduce water structuring effects at surfaces such as hydration and hydrophobic de-wetting. (d) *Ab initio* calculations finally give the effective interactions between oppositely charged groups in vacuum and in solution in [204]. Multi-scale simulations using generalized interpolation material point method and Structured Adaptive Mesh Refinement Application Infrastructure parallel processing as a multiple length scale tool from nanometer to millimeter is presented in [205]. To simulate the agglomeration carbonaceous nanoparticle assembly was studied using a multiscale coarse-graining by starting with an atomistic ensemble of 10 000 nanoparticles (or effectively 2 million total carbon atoms). The coarse-graining was accomplished applying a force-matching procedure. The results show rich and varied clustering behaviors for different particle morphologies [206]. A multiscale coarse-graining to derive coarse-grained models is applied to C60 and to carbonaceous nano-particles produced in combustion environments. The coarse-graining of the inter-particle force field is accomplished applying a force-matching procedure from all-atom MD simulations reproduce accurately the structural properties of the nano-particle systems [207].

(vi) A very impressive multiscale modelling approach to the mechanics of human hair fibres is given by Akkermans in [208] providing a very concrete example of multiscale modelling work-flows at their best where mesoscale models are constructed from atomistic simulations and where meso-scale simulation methods are used as input to finite-element calculations.

There are several excellent reviews discussing a great variety of invented multi-scale modelling schemes in materials science listed below and well worth of consulting. Simulation methods including broad areas of quantum mechanics, molecular dynamics and multiple-scale approaches, based on coupling the atomistic and continuum models are discussed in [209] and continuum/quasi-continuum approaches, the kinetic Monte Carlo technique and accelerated molecular dynamics simulation are gathered in [210]. Recent advances in bridging scale between quantum mechanical and continuum coupling are briefly described in [211] and multiscale hybrid simulation methods for material systems based on tight-binding DFT, MD and continuum models for deformation and diffusion on surfaces are reviewed in [212]. A seamless coupling of quantum to statistical to continuum mechanics involving models for unifying finite elements, molecular dynamics and semi-empirical tight-binding and coarse-grained Molec-

ular Dynamics as an effective model in solid state is explained in [213]. Current trends from atomistic simulation towards multiscale modelling of materials based on quantum mechanical, especially density functional theory for electronic properties linked to atomic/molecular dynamics and kinetic Monte Carlo simulations where coarse graining leads to lattice gas and cellular automata, and also to continuum equations solved by finite-element and finite-difference techniques are discussed in [214]. Vvedensky presents a comprehensive outlook and an excellent review of current multiscale modelling methods for nanostructures across a vast range of length and time scales clearly stating that a complete understanding of the behavior of materials thereby requires theoretical and computational tools that span from the atomic-scale detail of first-principles methods to the more coarse-grained description provided by continuum equations with the ultimate aim to systematically couple the scales from the atomistic to the continuum level in [215].

Multi-scale mechanics of nano-composites including an interface to better understand the phenomenological changes across multiple length and time scales are reviewed in [216] and current multi-scale modeling and simulation of nano-structured materials in [217]. Modelling the nano-scale phenomena in condensed matter physics via computer-based numerical simulations focusing on the adhesive and indentation properties of the solid surfaces in nano-contacts, the nucleation and growth of nanophase metallic and semi-conducting atomic and molecular films on supporting substrates, and the nano- and multiscale crack propagation properties of metallic lattices are discussed in [218]. Heterogeneous multiscale methods, applicable on complex fluids, micro-fluidics, solids, interface problems, stochastic problems, and statistically self-similar problems are reviewed in [219] and the emerging role of multiscale modeling in nano- and micro-mechanics of materials in [220].

Hierarchical modeling of amorphous polymers where a broad spectra of length and time scales governing the behavior of these materials is based on the use of connectivity-altering Monte Carlo algorithms for rapid equilibration of atomistic models of long-chain polymer systems, calculation of their conformational, packing and volumetric properties, and assessment of their entanglement structure, and self-consistent field calculations of morphology development in complex systems containing block copolymers, coupled with rubber elasticity theory for the prediction of the stress-strain behavior of these systems are discussed in [221]. Linking various length scales via materials informatics is reviewed in [222]. Recent progress in the simulations of liquid crystals across a range of length and time scales is reviewed in [223] with three material properties of liquid crystals (the archetypal self-assembled materials in Nature) are discussed in detail: elastic constants, rotational viscosity and helical twisting powers. Computer simulations of surfactant solutions are reviewed by Shelley & Shelley discussing the importance of connecting the inherently disparate length scales treating the systems using a single coherent multiscale simulation [224]. In an excellent review dealing with self-assembly from nano-scale to micro-scale colloids, issues like the effect of the shape and composition on assembly and the role of these factors in the self-organization of particles into ordered assemblies, scale dependent effects on assembly related to interparticle forces, vitrification and gelation and also building block design rules from computer simulations using anisotropic interactions are discussed in [225]. Karakasisis and Charitidis give an excellent review of methods in multiscale computational materials science how to connect electronic structure calculations with classical atomistic simulation using molecular dynamics or Monte Carlo methods at the nano/micro scale and further with kinetic Monte Carlo for larger system/time scales and finite elements for very large scales, presenting both hierarchical and hybrid strategies in [226]. A multiscale modelling scheme is reviewed, starting the initial nucleation processes using MD simulation and inputing the data into a cellular automaton (CA)-based model of the microstructure formation at the micro-scale, as well as to the macroscopic heat flow equation in [227]. Thermodynamics of self-assembly of surfactants in solution through simulations is now being expanded to include phenomena in the fluid dynamic regime a smooth link from MD to mesoscopic and macroscopic length and time scales. Recent trends in this area along with new results based on classical approaches are reviewed in [228].

Future Directions

Self-assembly is a process of pivotal importance that should be very attractive to study using modelling and simulations. The intermolecular interactions that drives the process can be directly introduced into common forms of the interaction potentials and force-fields. Many self-assembly processes can indeed also be rationalized and understood both molecularly and macroscopically by means of thermodynamics.

Self-assembly is typically a relatively slow process that proceeds through several metastable states. The extended time-scale and structure evolution that usually extends of several length scales requires a combination of approaches. The rapid expansion of multi scale modeling approaches

show much promise but it is still a long way to go before the fine details of self assembly processes with durations of hours up to weeks can be captured accurately. In principle the kinetic Monte Carlo method could cover arbitrary long time scales but it requires detailed knowledge about the system, its states and barriers between them.

Another great challenge in modern computational molecular science is the transition from a descriptive to a predictive level of molecular modeling. Despite several examples of successful predictions of molecular properties *in silico*, current molecular modeling remains largely on a descriptive stage: simulations are used to describe already known experimental phenomena, in order to give a better understanding, for testing the theories describing experiment, or for correct interpretation of the experimental data. Ability of simulations to predict, to discover something new is still rather limited. While some properties are predicted, others are not, successful investigations are published while failures are not, etc., so it is still difficult to rely on simulation results and experimental verification of simulation is therefore always needed. One of the reasons to a limited predictability of molecular models is that the currently used force fields were developed originally at the early stages of molecular modeling in the 80s with rather limited computing resources at hands and later only refined in a few occasions. In the case of meso-scale modeling on a coarse-grained level, even a functional form of the interaction potential is not known and therefore many *ad hoc* models are used. In order to substantially improve reliability and predictability of molecular simulations, a new generation of force fields would be necessary, which have to be derived in a more consecutive manner with more emphasis on the fundamental *ab-initio* approaches. To replace experiment with a computer experiment surely lies far in the future. However, if modelling can point the right direction to an experimentalist, much guesswork, effort and resources can be saved.

Acknowledgments

Both authors are connected to the Berzelii centre EXSELENT on porous materials at Stockholm University (http://www.exselent.su.se/).

Bibliography

1. Ozin GA, Arsenault AC (2005) Nanochemistry, A Chemical Approach to Nanomaterials. RSC Publishing, London
2. Whitesides GM, Grzybowski B (2002) Science 295:2418
3. Lehn J-M (1988) Angew Chem Int Ed 27:89
4. Nuzzo RG, Fusco FA, Allara DL (1987) J Am Chem Soc 109:2358
5. Bentzon MD, van Wonterghem J, Moerup S, Tholen A, Koch CJW (1989) Phil Mag B 60:169
6. Steigerwald ML, Alivisatos AP, Gibson JM, Harris TD, Kortan R, Muller AJ, Thayer AM, Duncan TM, Douglass DC, Brus LE (1988) J Am Chem Soc 110:3046
7. Redl FX, Cho K-S, Murray CB, O'Brien S (2003) Nature 423:968
8. Shevchenko EV, Talapin DV, Kotov NA, O'Brien S, Murray CB (2006) Nature 439:55
9. Decher G (1997) Science 277:1232
10. Whitesides GM, Mathias JP, Seto CT (1991) Science 254:1312
11. Fendler JH (1996) Chem Matter 8:1616
12. Lawrence DS, Jiang T, Levett M (1996) Chem Rev 95:2229
13. Zhang S (2003) Nature Biotech 21:1171
14. Tirrell M (2005) AIChE J 51:2386
15. Zhao X, Zhang S (2006) Chem Soc Rev 35:1105
16. Dubois LH, Nuzzo RG (1992) Ann Rev Phys Chem 43:437
17. Ulman A (1996) Chem Rev 96:1533
18. Laibinis PE, Whitesides GM, Allara DL, Tao YT, Parikh AN, Ruzzo RG (1991) J Am Chem Soc 113:7152
19. Tanev PT, Pinnavaia TJ (1995) Science 267:865
20. Zeng FW, Zimmerman SC (1997) Chem Rev 97:1681
21. Lyubartsev AP, Laaksonen A (2003) In: Karttunen M, Vattulainen I, Lukkarinen A (eds) Novel methods in soft matter simulations. Lecture Notes in Physics, vol 640. Springer, Berlin
22. Laaksonen A, Tu Y (1999) In: Balbuena PB, Seminario JM (eds) Molecular Dynamics: from classical to quantum methods. Elsevier, Amsterdam
23. Roothaan CCJ (1951) Rev Mod Phys 23:69
24. Hehre WJ, Radom L, Schleyer PVR, Pople JA (1986) Ab initio molecular orbital theory. Wiley, New York
25. Leach A (2001) Molecular Modelling: Principles and Applications, 2nd edn. Pearson Education, New Jersey
26. Pople JA, Santry DP, Segal GA (1965) J Chem Phys 43:S129
27. Pople JA, Segal GA (1965) J Chem Phys 43:S136
28. Pople JA, Beveridge DL, Dobosh PA (1967) J Chem Phys 47:2026
29. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) J Am Chem Soc 107:3902
30. Stewart JJP (1989) J Comput Chem 10:209
31. Stewart JJP (1989) J Comput Chem 10:221
32. Dewar MJS, Thiel W (1977) J Am Chem Soc 99:4899
33. Kolb M, Thiel W (1993) J Comput Chem 14:775
34. Weber W, Thiel W (2000) Theor Chem Acc 103:495
35. Tu Y, Jacobsson SP, Laaksonen A (2003) Mol Phys 101:3009
36. Lee C, Yang W, Parr RG (1988) Phys Rev B 37:785
37. Burke K, Perdew JP, Wang Y (1998) In: Dobson JF, Vignale G, Das MP (eds) Electronic Density Functional Theory: Recent Progress and New Directions. Plenum Press, New York
38. Becke AD (1988) Phys Rev A 38:3098
39. Hohenberg P, Kohn W (1964) Phys Rev B 136:864
40. Kohn W, Sham LJ (1965) Phys Rev A 140:1133
41. Car RR, Parrinello M (1985) Phys Rev Lett 55:2471
42. Foulkes WMC, Haydock R (1989) Phys Rev B 39:12520
43. Harris J (1985) Phys Rev B 31:1770
44. Sankey OF, Niklewski DJ (1989) Phys Rev B 40:3979
45. Tu Y, Laaksonen A (2005) In: Chipot C, Elber R, Laaksonen A, Leimkuhler B, Mark A, Schlick T, Schuette C, Skeel R (eds) Advances in Algorithms for Macromolecular Simulation. Lecture Notes in Computational Science & Engineering, vol 49. Springer, Heidelberg
46. Tu Y, Jacobsson SP, Laaksonen A (2006) Phys Rev B 74:205104

47. Madden PA, Heaton R, Aguado A, Jahn S (2006) J Mol Struct – Theochem 771:9
48. Hafner J (2007) Comp Phys Commun 177:6
49. Muller S (2006) Surf Interf Anal 38:1158
50. Juhas P, Cherba DM, Duxbury PM, Punch WF, Billinge SJL (2006) Nature 440:655
51. Le Page Y, Rodgers JR (2005) Acta Appl Cryst 38:697
52. Bockstedte M, Kley A, Neugebauer J, Scheffler M (1997) Comp Phys Commun 107:187
53. Gonze X, Beuken JM, Caracas R, Detraux F, Fuchs M, Rignanese GM, Sindic L, Verstraete M, Zerah G, Jollet F, Torrent M, Roy A, Mikami M, Ghosez P, Paty JY, Allan DC (2002) Comp Mater Sci 25:478
54. Colombo L (2005) Riv Nuovo Cim 28:1
55. Bowler DR, Choudhury R, Gillan MJ, Miyazaki T (2006) Phys Stat Solidi B 243:989
56. Kawazoe Y (2003) Bull Mater Sci 26:13
57. Allen MP, Tildesley D (1987) Computer Simulation of Liquids. Clarendon, Oxford
58. Rapaport DC (2004) The Art of Molecular Dynamics Simulation, 2nd edn. Cambridge University Press, Cambridge
59. Frenkel D, Smit B (2002) Understanding Molecular Simulation: from Algorithms to Applications, 2nd edn. Computational Science Series, vol 1. Academic Press, London
60. Schlick T (2002) Molecular Modeling: An Interdisciplinary Guide. Springer, New York
61. Hinchliffe A (2003) Molecular Modelling for Beginners. Wiley, New York
62. Field MF (2007) A Practical Introduction to the Simulation of Molecular Systems. Cambridge University Press, Cambridge
63. Seminario JM, Balbuena P (eds) (1999) Molecular Dynamics (Theoretical and Computational Chemistry). Elsevier Science, Amsterdam
64. Haile JM (1992) Molecular Dynamics Simulation: Elementary Methods. Wiley, New York
65. Leimkuhler B, Reich S (2004) Simulating Hamiltonian Dynamics. Cambridge University Press, Cambridge
66. Alder BJ, Wainwright TE (1957) J Chem Phys 27:1208
67. Alder BJ, Wainwright TE (1959) J Chem Phys 31:459
68. Rahman A (1964) Phys Rev 136:405
69. Verlet L (1967) Phys Rev 159:98
70. Stillinger FH, Rahman A (1974) J Chem Phys 60:1545
71. Hedman F, Laaksonen A (1999) In: Balbuena PB, Seminario JM (eds) Molecular Dynamics: from classical to quantum methods. Theoretical Chemistry Series, vol 7. Elsevier Science BV, Amsterdam
72. Hedmann F (2006) http://www.diva-portal.org/diva/getDocument?urn_nbn_se_su_diva-1008-4__fulltext.pdf
73. Sprik M, Demarchee E, Michel B, Rothlisberger U, Klein ML, Wolf H, Ringsdorf H (1994) Langmuir 10:4116
74. Yamamoto H, Watanabe T, Nishiyama K, Tatsumura K, Ohdomari I (2006) J Phys IV 132:189
75. Mamatkulov SI, Khabibullaev PK, Netz TR (2004) Langmuir 20:4756
76. Yu KQ, Li ZS, Sun JZ (2002) Langmuir 18:1419
77. Hackett E, Manias E, Giannelis EP (1998) J Chem Phys 108:7410
78. Jang SS, Jang YH, Kim Y-H, Goddard III WA, Flood AH, Laursen BW, Tseng H-R, Stoddart JF, Jeppesen JO, Choi JW, Steuerman DW, DeIonno E, Heath JR (2005) J Am Chem Soc 127:1563
79. Striolo A, NcCabe C, Cummings PT (2006) J Chem Phys 125:104904
80. Luo M, Mazyar OA, Zhu Q, Vaughn MW, Hase WL, Dai LL (2006) Langmuir 22:6385
81. Wu J (2006) AIChE J 52:1169
82. Jallipalli A, Balakrishnan G, Huang SH, Khoshakhlagh A, Dawson LR, Huffaker DL (2007) J Cryst Growth 303:449
83. van Workum K, Douglas JF (2006) Phys Rev E 73:031502
84. Ariga K, Hill JP, Ji Q (2007) Phys Chem Chem Phys 9:2319
85. Cooper N (1987) Stanislaw Ulam 1909–1984. Los Alamos Sci, vol 15. Special Issue. See: http://www.fas.org/sgp/othergov/doe/lanl/pubs/number15.htm
86. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AN, Teller E (1953) J Chem Phys 21:1087
87. Jayaram B (1996) J Math Chem 20:395
88. Ni R, Cao D, Wang WC (2007) J Phys Chem C 111:11802
89. Pool R, Schapotschnikow P, Vlugt TJH (2007) J Phys Chem 111:10201
90. Sinyagin AY, Belov A, Tang Z, Kotov NA (2006) J Phys Chem B 110:7500
91. Reimer U, Zehl T, Wahab M, Schiller P, Moegel H-J (2006) Coll Surf A Phys Chem Eng Asp 290:25
92. Reimer U, Wahab M, Schiller P, Moegel HJ (2006) Langmuir 21:1640
93. Patti A, Mackie AD, Siperstein FR (2007) Langmuir 23:6771
94. He XH, Song M, Liang HJ, Pan CY (2001) J Chem Phys 114:10510
95. Wijmans C, Linse P (1995) Langmuir 11:3748
96. Wijmans C, Linse P (1996) J Phys Chem 100:12583
97. Wijmans C, Linse P (1997) J Chem Phys 106:328
98. Linse P (2000) In: Alexandridis P, Lindman B (eds) Amphiphilic Block Copolymers: Self-Assembly and Application. Elsevier Science, Amsterdam
99. Abraham FF, White GM (1970) J Appl Phys 41:1841
100. Gilmer GH, Bennema P (1972) J Appl Phys 43:1347
101. Bortz AB, Kalos MH, Lebowitz JL (1975) J Comput Phys 17:10
102. Gillespie DT (1977) J Phys Chem 81:2340
103. Chatterjee A, Vlachos DG (2007) J Computer-Aided Matter Des 14:253
104. Metiu H, Lu Y-T, Zhang Z (1992) Science 255:1088
105. Much F, Ahr M, Biehl M, Kinzel W (2001) Europhys Lett 56:791
106. Larsson MI (2004) Surf Sci 551:69
107. Kozak JJ, Nicolis C, Nicolis G (2007) J Chem Phys 126:154701
108. Zhu R, Pan E, Chung PW (2007) Phys Rev B 75:205339
109. JS Rowlinson Notes (2003) Rec R Soc Lond 57:143
110. Wheeler AA, Boettinger WJ, McFadden GB (1992) Phys Rev A 45:7424
111. Anderson DM, McFadden GB, Wheeler AA (1998) Ann Rev Fluid Mech 30:139
112. Lu W, Suo Z (2001) J Mech Phys Solids 49:1937
113. Chen LQ (2002) Annu, Rev Mater Res 32:113
114. Boettinger WJ, Warren JA, Beckermann C, Karma A (2002) Ann Rev Mater Res 32:163
115. Thornton K, Ågren J, Voorhees PW (2003) Acta Mater 51:5675
116. Beckermann C, Diepers H-J, Steinbach I, Karma A, Tong X (1999) J Comput Phys 154:468
117. Boettinger WJ, Coriell SR, Greer AL, Karma A, Kurz W, Rappaz M, Trivedi R (2000) Acta Mater 48:43
118. Cahn JW, Hillard JE (1957) J Chem Phys 28:258
119. Cahn JW, Hillard JE (1959) J Chem Phys 31:688
120. Cahn JW (1961) Acta Metall 9:795

121. Allen SM, Cahn JW (1979) Acta Metall Mater 27:1085
122. Chen LQ, Shen J (1998) Comput Phys Commun 108:147
123. Elder KR, Grant M (2004) Phys Rev E 70:051605
124. Lu W, Salac D (2005) Phys Rev Lett 94:146103
125. Wang YU (2007) Acta Mater 55:3835
126. Inoue T (2004) J Phys IV 120:3
127. Inoue T (2004) Mater Sci Res Intl 10:1
128. Hoyt JJ, Sadigh B, Asta M, Foiles SM (1999) Acta Mater 47:3181
129. Gránásy L, Börsönyi T, Pusztai T (2002) Phys Rev Lett 88: 206105
130. Gránásy L, Börsönyi T, Pusztai T (2002) J Cryst Growth 237–239:1813
131. Gránásy L, Pusztai T, James PF (2002) J Chem Phys 117:6157
132. Gránásy L, Pusztai T, Warren JA, Douglas JF, Börsönyi T, Ferrero V (2003) Nat Mater 2:92
133. Kvamme B, Graue A, Aspenes E, Kuznetsova T, Gránásy L, Tóth G, Pusztai T, Tegze G (2004) Phys Chem Chem Phys 6:2327
134. Gránásy L, Pusztai T, Börsönyi T, Warren JA, Douglas JF (2004) Nat Mater 3:645
135. Svandal A, Kvamme B, Gránásy L, Pusztai T (2005) J Phase Equil Diff 26:534
136. Pusztai T, Bortel G, Gránásy L (2005) Mater Sci Eng A 413–414:412
137. Svandal A, Kvamme B, Gránásy L, Pusztai T, Buanes T, Hove J (2006) J Cryst Growth 486:1813
138. Tegze G, Pusztai T, Tóth G, Gránásy L, Svandal A, Buanes T, Kuznetsova T, Kvamme B (2006) J Chem Phys 124:234710
139. Gránásy L, Pusztai T, Saylor D, Warren JA (2007) Phys Rev Lett 98:035703
140. Davidchack RL, Laird BB (1998) J Chem Phys 108:9452
141. Davidchack RL, Laird BB (2000) Phys Rev Lett 85:4751
142. Plapp M, Karma A (2000) J Comput Phys 165:592
143. Raabe D (2002) Adv Mater Progr Rept 14:639
144. Penrose MD (2001) Commun Math Phys 218:153
145. Bartelt MC, Privman V (1991) Intern J Mod Phys B5:2883
146. Evans JW (1993) Rev Modern Phys 65:1281
147. Privman V (2000) Colloids Surf A 165:231
148. Talbot J, Tarjus G, van Tassel PR, Viot P (2000) Colloids Surf A 165:287
149. Senger B, Voegel J-C, Schaaf P (2000) Colloids Surf A 165:255
150. Rényi A (1963) Trans Math Stat Prob 4:205
151. D'Orsogna MR, Chou T, Antal T (2007) J Phys A Math Theor 40:5575
152. Ranjith P, Marko JF (2006) Phys Rev E 74:041602
153. Adamczyk Z, Weroński P, Barbasz J, Kolańsiska M (2007) Appl Surf Sci 253:5776
154. Gray JJ, Klein DH, Korgel BA, Bonnecaze RT (2001) Langmuir 17:2317
155. Breit M, Gao M, von Plessen G, Lemmer U, Feldmann J, Cundiff ST (2002) J Chem Phys 117:3956
156. Yuan Y, Oberholzer MR, Lenhoff AM (2000) Colloids Surf A165:125
157. Erban R, Chapman SJ (2007) Phys Rev E 75:041116
158. Cadilhe A, Aráujo N, Privman V (2007) J Phys Condens Matter 19:065124
159. Bohidar HB, Mohanty B (2004) Phys Rev E 69:021902
160. Gray JJ, Bonnecaze RT (2001) J Chem Phys 114:1366
161. Wolfram S (2002) A new kind of science. Wolfram Media, Champaign
162. Petri CA (1962) Kommunikation mit Automaten. Ph D Thesis, University of Bonn
163. Turing AM (1952) Phil Trans R Soc Lond B237:37
164. Wishart DS, Yang R, Arndt D, Tang P, Cruz J (2005) In Silico Biol 5:139
165. Succi S (2001) The Lattice Boltzmann Equation for Fluid Dynamics and Beyond. Oxford University Press, Oxford
166. Tougaw PD, Lenta CS (1996) J Appl Phys 80:4722
167. Amlani I, Orlov AO, Snider GL, Lent CS, Bernstein GH (1998) Appl Phys Lett 72:2179
168. Orlov AO, Amlani I, Bernstein GH, Lent CS, Snider GL (1997) Science 277:928
169. Snider GL, Orlov AO, Amlani I, Zuo X, Bernstein GH, Lent CS, Merz JL, Porod W (1999) J Appl Phys 85:4283
170. Cole T, Lusth JC (2001) Progr Quant Electr 25:165
171. Lent CS, Isaksen B, Lieberman M (2003) J Am Chem Soc 125:1056
172. Israeli N, Goldenfeld N (2006) Phys Rev E 73:026203
173. Mayer B, Köhler G, Rasmussen S (1997) Phys Rev E 55:4489
174. Nilsson M, Rasmussen S (2003) Discret Math Theor Comput Sci AB(DMCS):31
175. Kier LB, Cheng C-K, Testa B, Carrupt P-A (1997) J Pharmac Sci 86:774
176. Sreekumar P, Jayaraman VK, Kulkarni BD (1998) Ind Eng Chem Res 37:2188
177. Demontis P, Pazzona FG, Suffritti GB (2007) J Chem Phys 126:194709
178. Gumbsch P (1995) J Mater Res 10:2897
179. Lyubartsev AP, Laaksonen A (1995) Phys Rev E 52:3730
180. Muller-Plathe F (2002) Chem Phys Chem 3:754
181. Izvekov S, Parrinello M, Burnham CJ, Voth GA (2004) J Chem Phys 120:10896
182. Du MH, Cheng HP (2003) Int J Quant Chem 93:1
183. Mallik A, Taylor DCE, Runge K, Dufty JW (2004) Int J Quant Chem 100:1019
184. Bartlett RJ, McClellan J, Greer JC, Monaghan S (2006) J Comp Aided Mater Des 13:89
185. Zhu WM, Runge K, Trickey SB (2006) J Comp Aided Mater Des 13:75
186. Faller R (2004) Polymer 45:3869
187. Buehler MJ, Hartmaier A, Gao H (2004) Model Sim Mater Sci 12:S391
188. Doi M (2001) J Comp Appl Math 149:13
189. Kinjo T, Hyodo S (2007) Mol Simul 33:417
190. Hyodo S (2007) Mol Sim 33:279
191. Haslam AJ, Moldovan D, Phillpot SR, Wolf D, Gleiter H (2002) Comp Mater Sci 23:15
192. Nakano A (1997) Comp Phys Commun 105:139
193. Shen LM, Chen Z (2005) Int J Solids Struct 42:5036
194. Curtin WA, Miller RE (2003) Model Simul Mater Sci Eng 11:R33
195. Rafii-Tabar H, Hua L, Cross M (1998) J Phys Cond Matter 10:2375
196. Smit B, Krishna R (2003) Chem Eng Sci 58:557
197. Warrier M, Schneider R, Salonen E, Nordlund K (2004) Contr Plasma Phys 44:307
198. Shibuta Y, Elliott JA (2006) Chem Phys Lett 427:365
199. Sakaguchi N, Watanabe S, Takahashi H, Faulkner RG (2004) J Nucl Mater 329:1166
200. Vasenkov AV, Fedoseyev AI, Kolobov VI, Choi HS, Hong KH, Kim K, Kim J, Lee HS, Shin JK (2006) J Comp Theor Nanosci 3:453

201. Zhang J, Adams JB (2004) Comp Mater Sci 31:317
202. La Magna A, Coffa S, Libertino S, Brambilla L, Alippi P, Colombo L (2001) Nucl Instr Meth Phys Res B 178:154
203. Haran M, Goose JE, Clote NP, Clancy P (2007) Langmuir 23:4897
204. Netz RR (2004) J Phys Cond Mat 16:S2353
205. Ma J, Lu H, Wang B, Roy S, Hornung R, Wissink A, Komanduri R (2005) Comp Model Eng Sci 8:135
206. Izvekov S, Violi A (2006) J Chem Theory Comput 2:504
207. Izvekov S, Violi A, Voth GA (2005) J Phys Chem B Lett 109:17019
208. Akkermans RLC, Warren PB (2004) Phil Trans Royal Soc A 362:1783
209. Liu WK, Karpov EG, Zhang S, Park HS (2004) Comp Meth Appl Mech Eng 193:1529
210. Starrost F, Carter EA (2002) Surf Sci 500:323
211. Lium WK, Park HS, Qian D, Karpov EG, Kadowaki H, Wagner GJ (2006) Comp Meth Appl Mech Eng 195:1407
212. Csanyi G, Albaret T, Moras G, Payne MC, De Vita A (2005) J Phys Cond Matter 17:R691
213. Rudd E, Broughton JQ (2000) Physica, Status Solidi B – Basic Res 217:251
214. Nieminen RM (2002) J Phys Cond Matter 14:2859
215. Vvedensky DD (2004) J Phys Condens Matter 16:R1537
216. Buryachenko VA, Roy A, Lafdi K, Anderson KL, Chellapilla S (2005) Compos Sci Tech 65:2435
217. Gates TS, Odegard GM, Frankland SJV, Glancy TC (2005) Compos Sci Tech 65:2416
218. Rafii-Tabar H (2000) Phys Rept Rev Sect Phys Lett 325:240
219. Weinan E, Engquist B, Li X, Ren WQ, vanden-Eijnden E (2007) Commun Comp Phys 2:367
220. Ghoniem NM, Cho K (2002) Comp Model Eng Sci 3:147
221. Theodorou N (2005) Comp Phys Commun 169:82
222. Liu ZK, Chen LQ, Rajan K (2006) JOM 58:42
223. Wilson M (2005) Intl Rev Phys Chem 24:421
224. Shelley JC, Shelley MY (2000) Curr Opin Coll Interf Sci 5:101
225. Glotzer SC, Solomon MJ, Kotov NA (2004) AIChE J 50:2978
226. Karakasidis TE, Charitidis CA (2007) Mater Sci Eng C 27:1082
227. Rafii-Tabar H, Chirazi A (2002) Phys Rep 365:145
228. Rajagopalan R (2001) Curr Opin Coll Interf Sci 6:357

Self-Organization and the City

Juval Portugali
ESLab (Environmental Simulation Lab), Department of Geography and the Human Environment, Tel Aviv University, Tel Aviv, Israel

Article Outline

Glossary
Definition of the Subject
Introduction
Complexity Theories of Cities – an Overview
Self-Organization and the City
Future Directions
Bibliography

Glossary

Self organization A property of open and complex systems that achieve their order spontaneously, that is, by means of "self-organization".

City A form of settlement that first emerged in the Near East (the core being Mesopotamia) some 5500 years ago. Since its first appearance in Mesopotamia it has diffused in space and time. With colonialism the Western–European form of city has diffused to the entire world, suppressing on the way other forms of cities (e. g. in South America, East Asia, etc.). In the last few decades the city has become the most dominant form of settlement: for the first time in human history more than half of world population lives in cities.

Urbanism The term refers to the totality of life in cities: the interrelations between the social structure, culture, economy, politics, architecture, physical morphology, … associated with life in cities. The first appearance of cities is thus termed *the urban revolution*. Most students of urbanism would agree that 21st century society is undergoing a major urban transformation; some describe it as a new urban revolution.

Planning Planning is, on the one hand, a basic cognitive capability of humans while on the other, a profession and research domain termed interchangeably *city planning*, *urban and regional planning* or *environmental planning*. City planners work and act in the context of planning law and administration the aim of which is to regulate and control life in cities.

SIRN (synergetic inter-representation network) An approach to cognitive mapping and urban dynamics suggesting that cities emerge, maintain their order and change again as a consequence of an on-going interaction between cognitive maps that are constructed in the mind/brain of humans as internal representations and the city as a collective external representation. This on-going interaction gives rise to a network some of whose elements are in the mind/brain, while others in the world.

Information compression, inflation and adaptation A view suggesting that Shannon's notion of information is a property of closed systems and that in complex, self-organizing systems one has to take into consideration the role of *semantic information*. Due to semantic information, the process of self-organization often entails *information compression*; in some cases it entails *information inflation*. The suggestion is that information compression and inflation are two facets of the process of *information adaptation*.

Definition of the Subject

Self-organization is a central property of open and complex systems. While the concept had already appeared in the 1940s, its modern use was pioneered in the 1960s by people such as Haken [50,51] with his theory of *synergetics*, Prigogine with his notion of *dissipative structures* [87,109,110] and others (see review in Chap. 3.1 [95]). Such systems are typically in "a far from equilibrium condition" and exhibit phenomena of chaos, fractal structure and the like. For a long time, the term "self-organization" was used also as an umbrella name for these theories; nowadays it is common to refer to these theories as *complexity theories*.

The notions of self-organization and complexity originated in the sciences, specifically in physics, as a property of natural systems. However, as we shall see below, from the start they were associated with the city – at the beginning the city was used as a metaphor to convey the notion of "self-organization" [87] and at a later stage it was studied as a genuine self-organization system in its own sake [5].

Most theories and methodologies of complexity developed in the last three decades have been applied also to the study of cities with the result that we now have a rich body of research on *fractal cities* [20], *self-organization and the city* [95], *cities and complexity* [16], cellular automata and agent base urban simulation models [22], studies on cities from the perspective of Bak's self-organized criticality [20], studies on cities as networks [16] and much more. This growing body while enriching our understanding of cities and providing sophisticated tools to city planning, also exposes problems that will become the challenges for the next generation of studies on complexity, self-organization and the city.

Introduction

The title "Self-Organization and the City" enfolds *two notions* – 'Self-organization' and 'City', *a thesis* suggesting that cities are complex self-organizing systems, and an *inconsistency* – the view of cities as complex systems that achieve their order spontaneously contradicts the traditional view of cities as symbols of organized order and planning. From the title thus follow four questions: what is self-organization? What is a city? In what sense are cities self-organizing systems? What is the meaning of planning in a self-organizing system? In an *Encyclopedia of Complexity and System Sciences* such as this, there is no need to introduce the term self-organization beyond what has been said about it above; we are thus left with three introductory tasks, namely, to clarify the notion "City", elaborate the thesis that cities are self-organizing systems, and, solve the contradiction between self-organization and planning in the realm of cities.

What is a City?

There are two ways to answer this question: first, by looking at explicit attempts to define a city. Second, by exposing the way the different theories of cities explicitly describe, or implicitly perceive, a city.

Explicit Attempts to Define a City The history of the many attempts to define "a city" is rather confusing: Whenever a definition was proposed, it was always possible to falsify it (in Popper's [90] sense) by putting forward cities that do not comply with the definition. The main reason for the failure to define cities is that the various attempts to do so were always made with reference to what in cognitive science [115] is called *classical categories*. That is, groups composed of entities sharing some necessary and sufficient conditions that define them as a category and distinguish them from other categories. Students of urbanism have implicitly treated cities as classical categories, and yet, cities are not classical categories – they form a category due to what Wittgenstein [136] has termed *family resemblance*. As a consequence, attempts to define them in terms of a classical category ended up with a failure [95].

A family resemblance category becomes a category not when its elements share some common denominators, but when they form a *network* of partial links and similarities. Further research and experiments have found that many family resemblance categories have a *core-periphery* structure, in the sense that some instances of the category are more *prototypical* of the category than others and they thus form its center while the rest of the instances form the category's periphery [74,78,115].

The city is a good example of a family resemblance category with a core periphery structure. On the one hand, there are no common elements between the "first" cities of some 5,500 years ago and the cities of today except for the name. On the other hand, the first cities had space-time links and similarities with subsequent cities, which in turn had common elements with subsequent cities, and so on until the global cities of today. The result of this process is that cities form a huge space-time family resemblance network extending in time and space from the ancient cities of 5,500 years ago to the cities of today. In this network, one can identify space-time moments during which certain cities became more characteristic or proto-

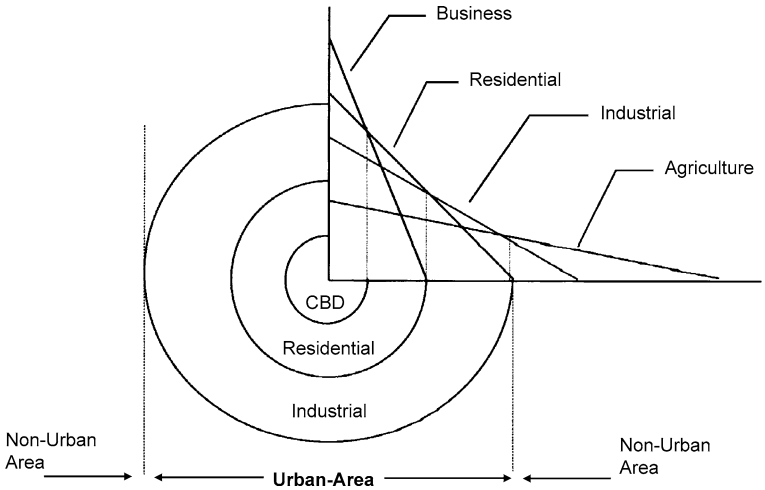

Self-Organization and the City, Figure 1
Thünen's type land-use system as transformed into an urban land-use system by location theory. Businesses are prepared to pay high rent at the center of the city, but are reluctant to "live" far from it. Their spatial demand curves (or rbc – rent bid curves) are thus the highest and steepest. Industrialists, in this exposition, are exactly the opposite and residents are in between: they cannot afford to pay the high prices at the center, but are prepared to live far from it, and so on. Each land use thus occupies a ring were it can pay (bid for) the highest rent. Note that the principle of marginal utility which is implicit in Thünen's landscape, here appears explicitly as the central economic principle

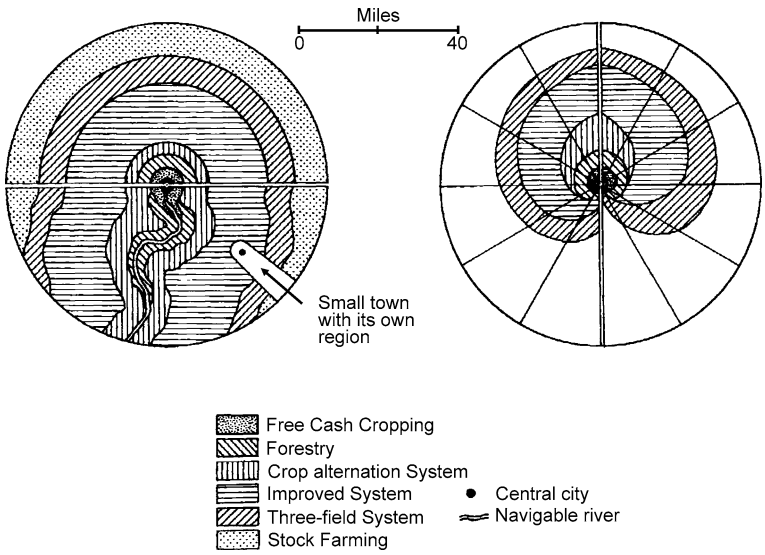

Self-Organization and the City, Figure 2
The diagrams of Thünen's *Isolated State*: *Left, upper part of the diagram* "This shows the Isolated State in the shape it must take from the assumptions made in Section One…". *Left, lower part* "Here we see the Isolated State crossed by a navigable river. Here the ring of crop alternation become very much larger, stretching along the river … The effect of constructing a highway is similar, … " (Par. 385). *Right* "The diagram illustrates the effect of the Town grain price on the extension of the cultivated plain" (Par. 386). Source: [129]

typical of the category than others. Such cities have temporarily captured the center of the category city, pushing to the periphery the rest of the instances, only to be replaced in subsequent space-time moments by other prototypical cities, other centers and other peripheries. How does this huge network evolve in time and space? The answer is: "by means of self-organization" [95].

Images of Cities This section discusses images of cities that are implicit or explicit in several of the urban theories.

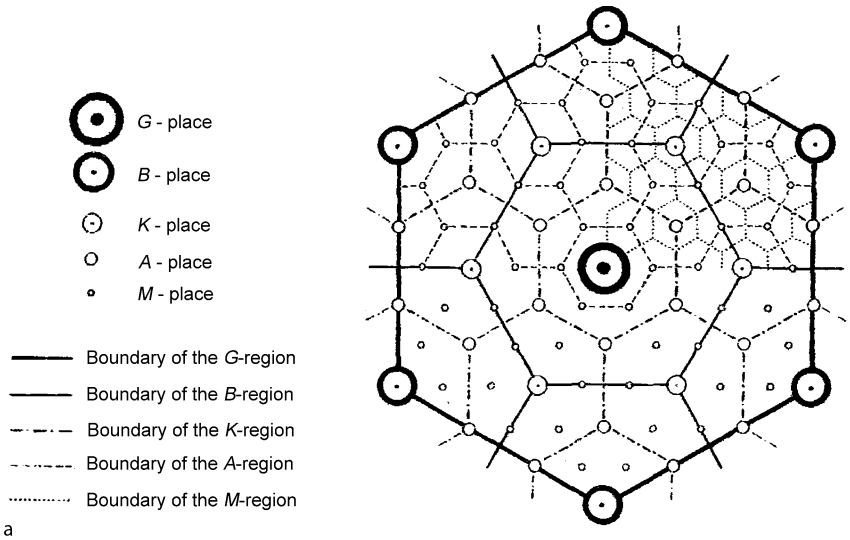
G - place
B - place
K - place
A - place
M - place
Boundary of the G-region
Boundary of the B-region
Boundary of the K-region
Boundary of the A-region
Boundary of the M-region

a

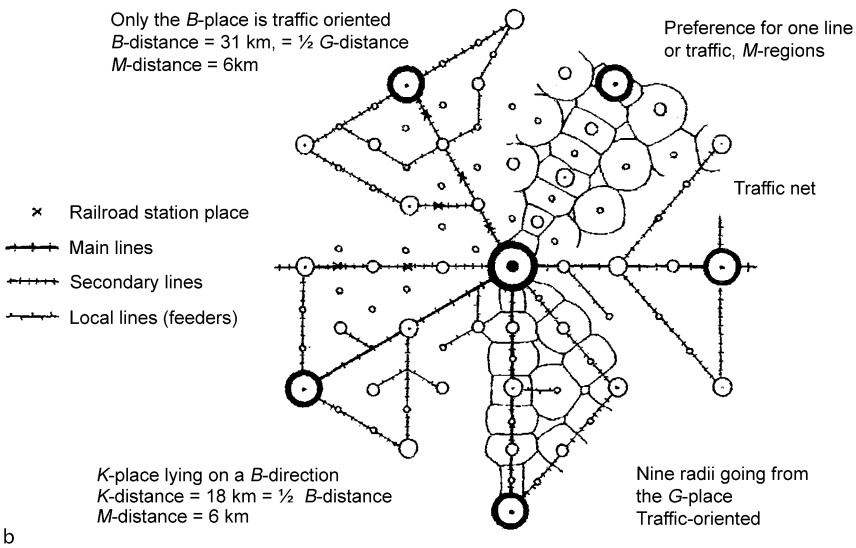
Only the B-place is traffic oriented
B-distance = 31 km, = ½ G-distance
M-distance = 6km
Preference for one line
or traffic, M-regions
Traffic net
Railroad station place
Main lines
Secondary lines
Local lines (feeders)
K-place lying on a B-direction
K-distance = 18 km = ½ B-distance
M-distance = 6 km
Nine radii going from
the G-place
Traffic-oriented

b

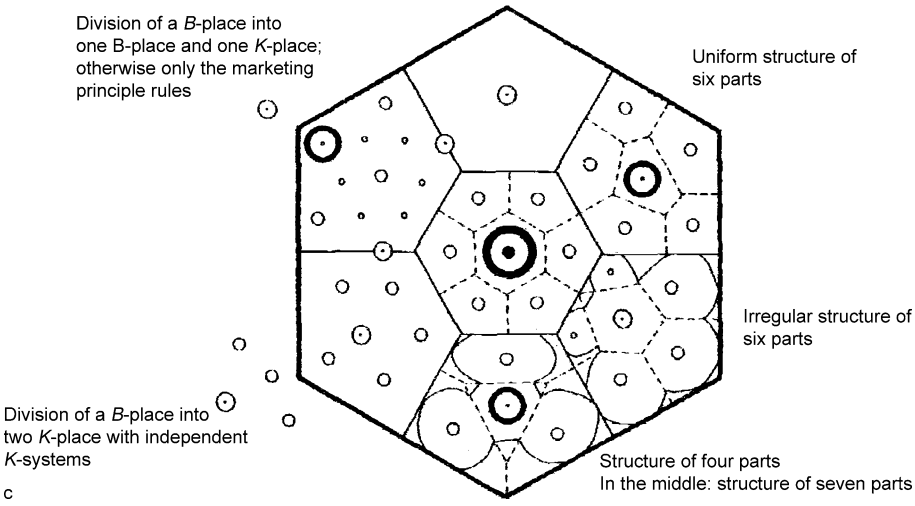
Division of a B-place into
one B-place and one K-place;
otherwise only the marketing
principle rules
Uniform structure of
six parts
Irregular structure of
six parts
Division of a B-place into
two K-place with independent
K-systems
Structure of four parts
In the middle: structure of seven parts

c

◀ **Self-Organization and the City, Figure 3**
Christaller's systems of central places according to the three locational principles. a The marketing regions in a system of central places. b A system of central places developed according to the traffic principle. c A system of central places developed according to the separation principle. (Source: Figs. 2, 4, 6 in [36])

Only theories that facilitate subsequent discussion will be discussed.

The Economic City The city is portrayed as a center-periphery structure with several rings that come into being by an economic competition between land uses that differ in their spatial demand function (Fig. 1). The land uses with the highest and steepest *spatial demand curves* or *rent bid curves* [8,69] capture the central ring – the most accessible area of the city – thus forming the CBD (central business district); the rest of the land uses occupy the peripheral rings. The origin of this city image is von Thünen's [129] *Isolated State* (Fig. 2): Originally formulated as a theory of agricultural land uses it became the founding theory of all location theories including urban land use theories.

The City as a Central Place The city is perceived as a central place that mediates between the city's complementary region and other cities that form a hierarchical network of central places (Fig. 3). The origin of this view is Christaller's [36] *central place theory* that perceived the city as a central place for tertiary activities (a market place, transportation node and administrative center). Losch's [81] central place theory was more ambitious and complicated and portrayed the city as a central place for all production, consumption, transportation and political activities (Fig. 4).

The City as a Node in a System of Cities The city here "looses" its autonomy in the sense that it is perceived as a node in a system of cities – no attention is paid to the city's role or function; the focus of interest is on the system as a whole. This view is due to Auerbach [10] who already at the turn of the 20th century showed that the rank-size distribution of cities obeys the power law. In a famous work from 1949 Zipf showed that this rank-size distribution typifies not only cities (Fig. 5), but a whole range of phenomena [138]. Zipf's work provided a source of inspiration to a long list of subsequent studies on systems of cities [111].

The Ecological City As in the economic city, here too, the city is portrayed as a center-periphery ring structure. However, here the city's structure emerges out of a competition between cultural and socio-economic groups in a way similar to competition between species in natural ecology. The various studies in this domain are termed *urban ecology*; their source of inspiration is the Chicago school of social ecology. Several urban landscapes were suggested the most dominant one is Burgess' [30,31,32] ring structure (Fig. 6).

The City as a Representation of Society The city is here perceived as a spatial representation of society as a whole. This image of the city emerged in the early 1970s as a consequence of a paradigm shift the study of cities underwent – from liberal social and economic theories to more radical ones with Marxism being the most dominant view. Two Marxist interpretations of the city can give the flavor of this approach. The first is Castells' [33] view according to which the city is a spatial representation of the structure of society as perceived by structuralist-Marxist theory (Fig. 7). The second is Harvey's view according to which the city's landscape emerges, as a logical consequence, out of internal contradictions inherent in the capitalist *mode of production* that according to this view, dominates world society of the 20th and 21st centuries. Namely, between forces of spatial agglomeration and processes of spatial dispersion. As illustrated in Fig. 8, this tension can be resolved only by *the urbanization of capital* [47,63].

The City as a Socio-cultural Force The city is here perceived as a force that is shaping the life of the people living in it. In urban societies it implies that the city is in fact shaping society. This view is due to the study of Wirth's [135] *Urbanism as a way of life* and also of Park's study *The City* [88]. In 1970 Lefebvre has published a monograph *La Révolution Urbaine* suggesting from a Marxist point of view that society is reaching a stage of being completely urban so that urbanism is replacing industrialism as the major force of society [79].

The Postmodern City The city of 21st century is described as the postmodern city [95]: Untamed, shrew, capricious, ever-changing; actually it is not a city but a text written by millions of unknown writers, unaware that they are writers, read by millions of readers, each reading his or her own personal and subjective story in this ever-changing chaotic text, thus changing and recreating and further complicating it. Today's urbanism is a big theater at the center of whose stage we see a kaleidoscope of shapes, forms, high-tech science-fiction structures, cultures and sub-cultures, Italians, Chinese, Japanese, Jews, Indians, Gays, Lesbians. Yapese; nothing is stable, nothing

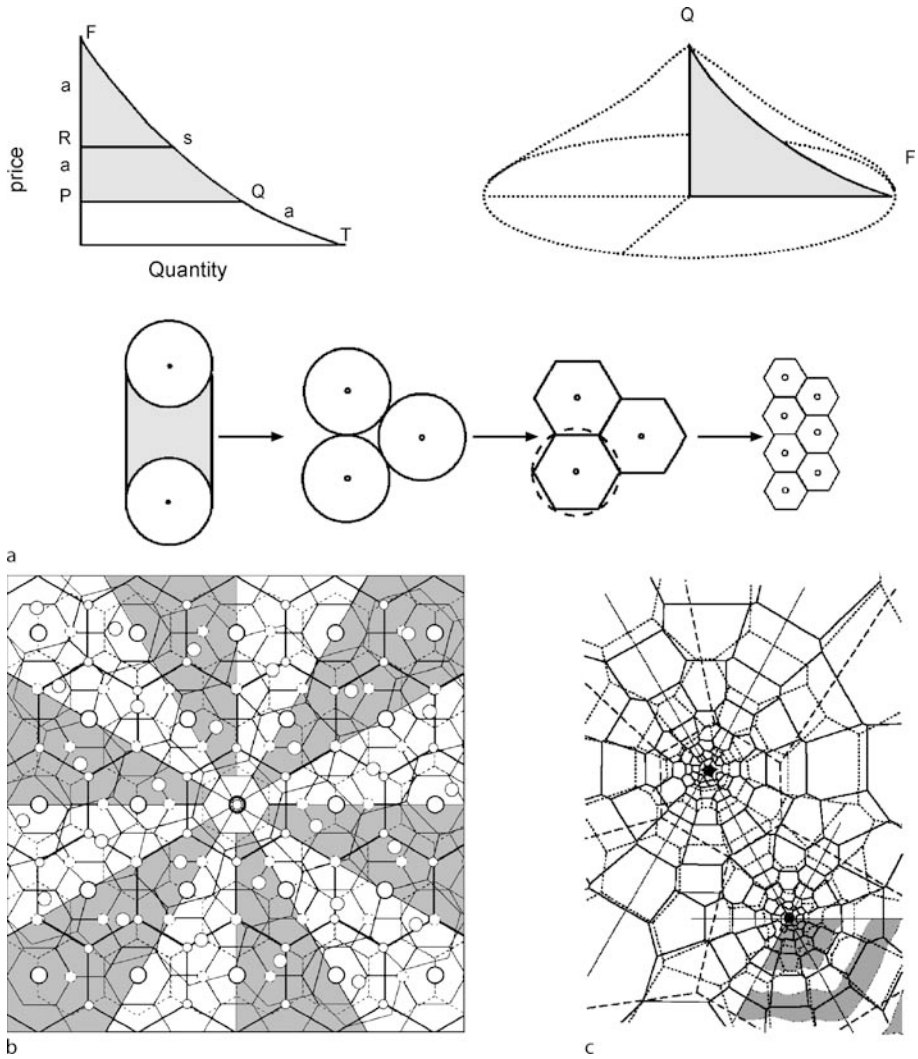

Self-Organization and the City, Figure 4

a The derivation of Lösch's system of central places. *Top* The derivation of a *spatial demand cone* with its market area (*right*) out of an "ordinary" demand curve (*left*). *Bottom* Development of market areas from the large circle to the final small hexagon. Source: Figs. 20–23 in [81]. **b** Lösch's derived system of central places with their market areas, divided into "city-poor", "city-rich" sectors. Source: Fig. 28 in [81]. **c** A Lösch system of central places modified by Isard [69] so as to be consistent with the resulting population distribution

is true nor matter for more than a second, not the Marxist urban categories, nor any other grand theory or truth; all must go, must move, clear the way to the new next whatever it is.

The Self-organizing City Strangely enough, an image of the city similar to the postmodern one is emerging out of complexity studies of cities. This is a seemingly similarity, however; a closer look reveals, first, that theories of complexity made a direct and explicit link to the views of cities as central places, to the studies on systems of cities and to the ecological views on the city. Second, that the notion of self-organizing city has several important resemblances with modern social theory oriented urban studies that perceive the city as the representation of society.

In What Sense Are Cities Self-organized Systems?

Self-organization is a property of systems that are open and complex. No one plan such systems, no one fully controls them and yet they have order, rules, organization and all these emerge spontaneously by means of self orga-

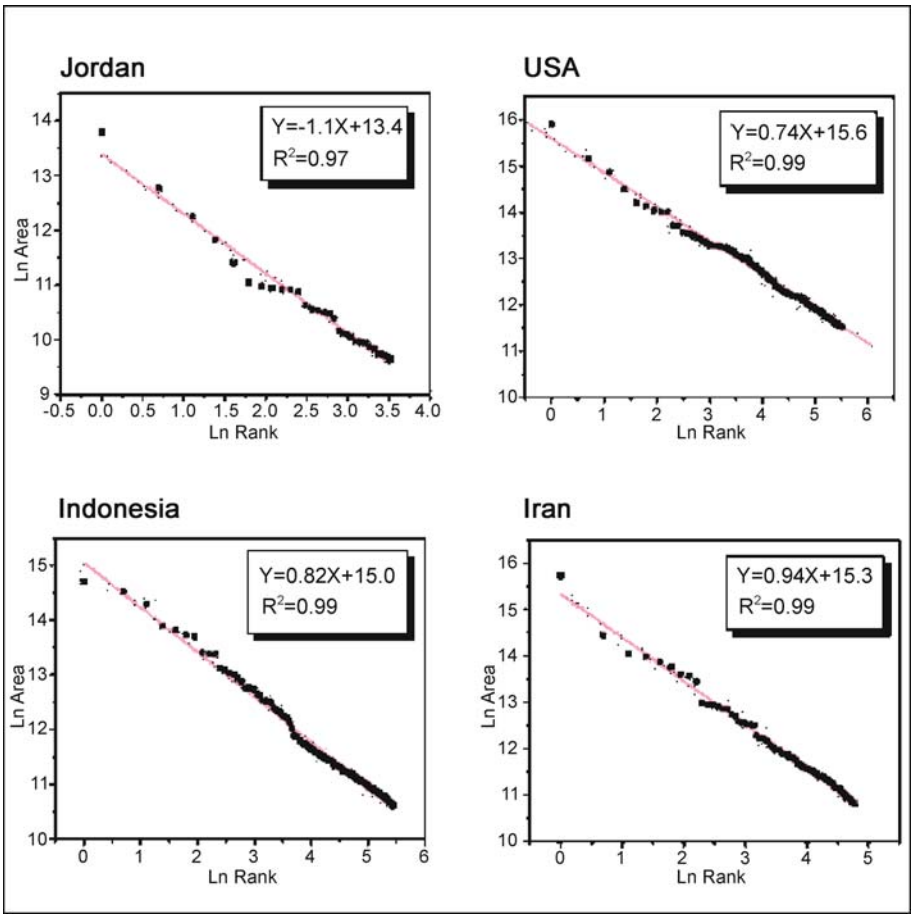

Self-Organization and the City, Figure 5
Rank size distribution of cities with more than 100,000 inhabitants in four different countries. Source: Fig. 2.7 in [27]

nization. A nice example of self-organization is provided by human languages (Chinese, Hebrew, English . . .). Each such language is an open system; each is a complex system; each is a system that emerged out of synergetic interaction between a huge number of people (the "parts" of such systems); no one has ever fully controlled languages; no one has fully planned a language; and yet each of the human languages has order, rules, organization and all these emerged spontaneously by means of self organization.

Similarly to human languages, each city is an artifact; each is an open system; each is a complex system; each city is a system that emerged out of interaction between a huge number of people; no one fully controls it; and yet it has order, rules, organization and all these spontaneously by means of self organization.

But cities are not languages. For one thing, their products are stand-alone objects such as buildings, roads, bridges, etc. that exist and survive independently of their producers. The products of languages are humans' voices and gestures that have no existence independent of their producers. Cities, in this respect, are akin to writing and texts – the external, stand-alone, representations of languages. The appearance of cities, some 5,500 years ago, hand in hand with writing, is, to my mind, not accidental.

A second difference concerns planning: There are no language planners and the attempt to "plan" the international language of Esperanto ended in failure. But there are many city planners – much more than appreciated in conventional planning theory. This is so because planning is a basic human property with the implication that each agent operating in the city (person, family, company) is a planner on a certain level. In certain cases, because of the nonlinearities that typify the complexity of cities, the planned action of a single individual might influence the city more than that of the official planners and their plans. Urban dynamics can thus be seen as an on-going interaction between planners and their plans when none of them

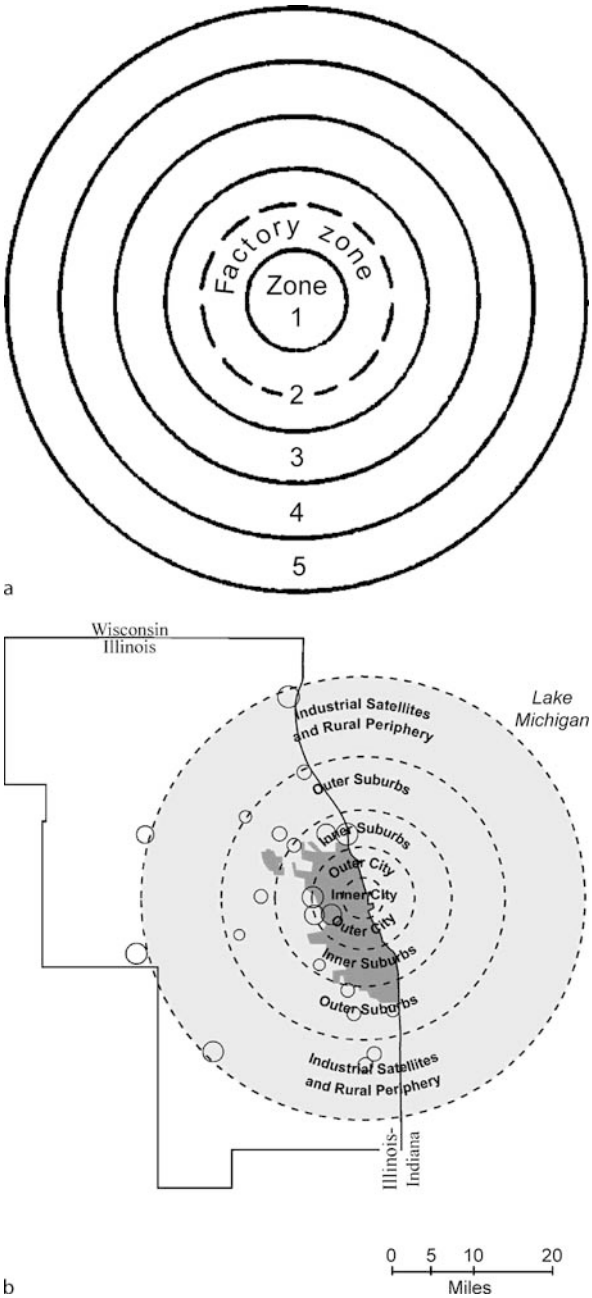

Self-Organization and the City, Figure 6
a Burgess concentric zone model. (1) Central Business District, (2) Zone in transition, (3) Zone of working men's homes, (4) Residential zone, (5) commuters' zone. b Regional differentiation of Chicago as commonly presented in several geographical textbooks. Source: [95]

can fully determine the final form and structure of the city. They are all *participants* in a big city-planning game (see Part III in [95]).

The Inconsistency Between Self-Organization and Planning

All theories of cities were associated with city planning. The basic idea is that the city is an artifact and as such a product of humans' intentions and needs. Planning is needed in order to implement human needs and intentions in a rational way. The notion of complexity suggests that the city is a product of self-organization; if this is so, who needs city planning? The answer to this inconsistency has already been given above: the plans produced by city planners, like those produced by "ordinary" urban agents, are *participants* in a big city-planning game.

Complexity Theories of Cities – an Overview

The discussion in this section proceeds under the titles of eight "cities" that are related to general theories or specific methodologies. It starts with 'dissipative cities' to indicate that Prigogine's was the first complexity theory applied to the study of cities

Dissipative Cities

In their introduction to *Self-Organization in Nonequilibrium Systems* Nicolis and Prigogine [87] use the example of a city as a metaphor to convey to their fellow physicists what they mean by "self-organization."

> "An appropriate illustration would be a town that can only survive as long as it is a center for inflow of food, fuel … and sends out products and wastes."

Peter Allen [5] – Prigogine's student – showed that towns and cities are not just metaphors, but genuine self-organizing systems. He did so by reformulating *central place theory* (above Sect. "Images of Cities") in terms of Prigogine's theory. (Note the resemblance between the hexagonal landscapes of *central place theory* and the hexagonal Bénard cells – one of the canonical experiments of the paradigm of self-organization).

Allen and co-workers' have developed a sequence of several models which elaborated their theoretical treatment of hierarchical landscapes of central places, first with respect to systems of cities in a given region and later at the intra-urban scale in connection with a single city. At a later stage they have also applied their models to real case studies of Brussels and the Belgian provinces [121], see also [109].

A typical model of Allan's starts with an infrastructure of localities in a region, each with its residents and jobs. The actors are individuals who migrate in order to get employment, and employers who offer or take away

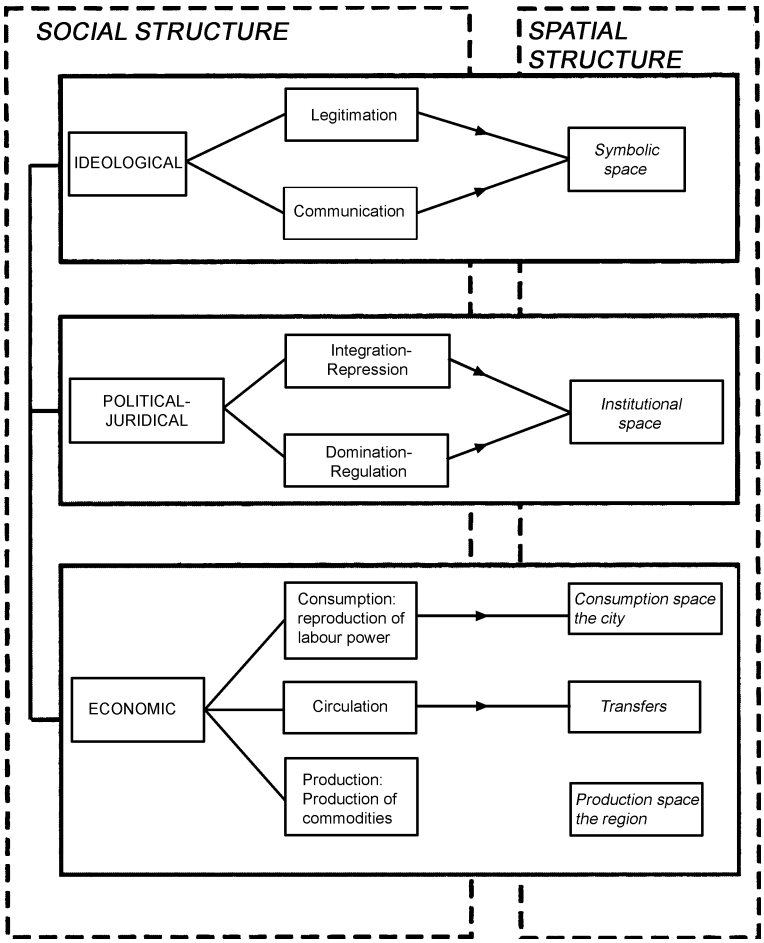

Self-Organization and the City, Figure 7
The Marxist city as a spatial representation of social structure [33]. Source: [47]

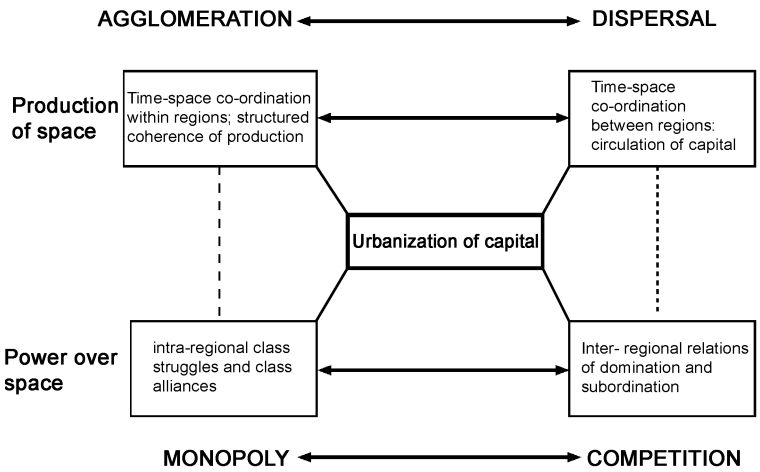

Self-Organization and the City, Figure 8
The Marxist city as an outcome of basic tensions in the landscape of capitalism [63]. Source: [47]

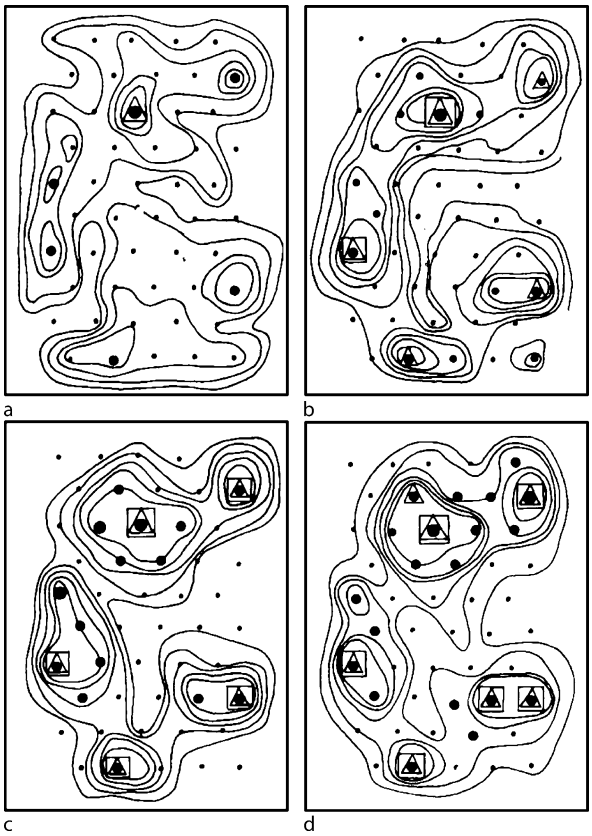

Self-Organization and the City, Figure 9
Allen and Sanglier's simulated evolution of a dissipative system of cities. a at time (*t*) $t = 4$; b at $t = 12$; c at $t = 20$; d at $t = 34$

jobs depending on the market's situation. The migration/interaction between localities and the introduction and extraction of economic activities (i. e. employment opportunities), create for each locality a kind of local "carrying capacity" and for the system as a whole nonlinearities and feedback loops which link population growth and manufacturing activities. An example for a simulated scenario produced by the model is Fig. 9. It starts (Fig. 9a) with a hypothetical region characterized by a rectangular lattice of homogeneous localities. Then, the mere play of chance factors, such as the place and time where different enterprises and migrations start, produce symmetry breakings which entail an uneven distribution of population, employment and so on (Figs. 9b–d). The result is an evolutionary process by which new urban centers emerge, grow, and form the whole of the regional system of central places; as the system evolves, some old localities grow, others decline or even disappear, thus constructing the specific history of this region.

Allen and co-workers' approach exposes the similarity and difference between the "old" static approaches of Christaller and Lösch, and the new treatment by means of self-organization. In both, economic activities and interactions give rise to cities as central places. However, while in the old formulations the landscape reflects an equilibrium state which is the optimized sum of the properties of the various economic forces, the new landscape reflects a far-from-equilibrium situation in which the spatial hierarchical order among the central places is obtained, maintained and then transformed, by means of interplay between interaction, fluctuations and dissipation.

Synergetic Cities

Two main approaches of synergetics have been applied to the study of cities. The first is the master-equation approach that is characteristic of Weidlich and co-workers studies in sociology, economics and urban dynamics [131,132,133,134]. For many years this was the main synergetic approach to cities, and most applications thus far have been within this conceptual frame [48,112,119]. The second, the pattern recognition approach, typifies the synergetics' treatment of pattern formation, cognition, pattern recognition and brain activities, as developed in the last three decades by Haken and co-workers [56]. Since the 1990s this approach has been applied to the study of cities as self-organizing systems [58,95].

Slow Cities and Fast Regions One way to look at Haken's synergetics and its slaving principle is in terms of interplay between slow and fast processes:

> If in a system of nonlinear equations of motion for many variables these variables can be separated into slow ones and fast ones, a few of the slow variables … are predestined to become "order parameters" dominating the dynamics of the whole system on the macro-scale [134].

This perspective stands at the basis of Weidlich's and co-workers studies on sociodynamics and cities [131,132,133,134]. According to this perspective, fast and slow processes are easily identifiable in processes of settlement and urbanism. The fast ones typify the local microlevel of building sites, streets, subways, etc., whereas the slow processes typify the macrolevel of whole regions which are often described as systems of cities. The relations between the slow and the fast processes are described by the slaving principle: on the one hand, the regional system

> serves as the environment and the boundary condition under which each local urban microstructure evolves. On the other hand, the … regional macrostructure is … the global resultant of many local structures [134].

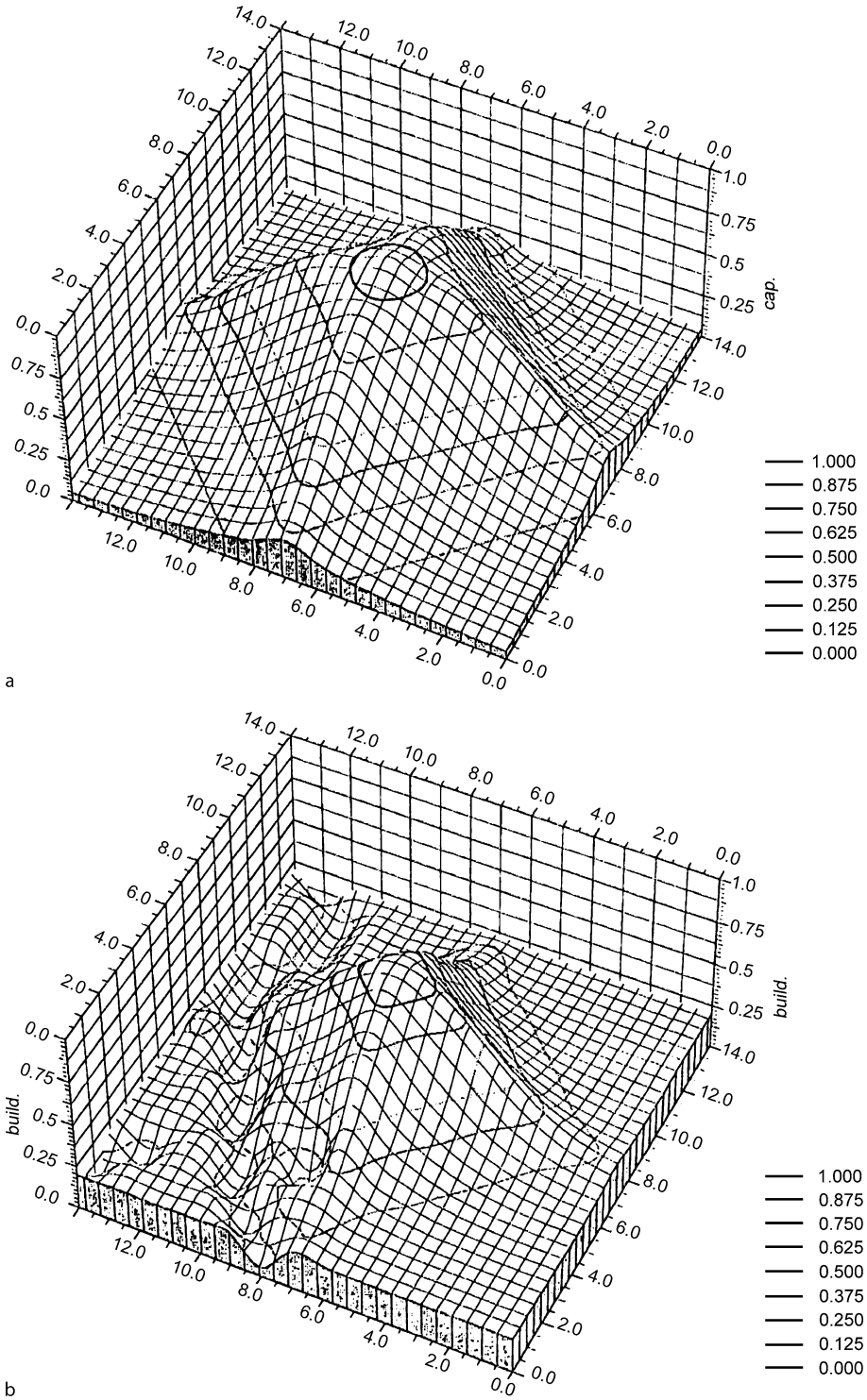

Self-Organization and the City, Figure 10
Building and development under population pressure [134]. *Top*: on a uniform urban plain. *Bottom:* on an urban plain with disturbances

This circular causality between the local and the global, allows one to study global regional systems by assuming that local processes adapt to the slow regional ones, and to study local urban processes by treating the regional context as given, and of course to study the complex interplay between the local and the global. In all three cases Weidlich has recently prescribed a four stages approach: stage 1 concerns the configuration space of the variables; stage 2, measures the utility of each configuration; stage 3, defines transition rates between configurations which are in fact utility differences; stage 4 derives stochastic or quasi-deterministic evolution equations for the system under consideration. The central evolution equation is the master equation which defines the probability that the configuration under examination is realized at a certain time.

The above theoretical procedure has been used to study the role of population pressure in "fast and slow processes in the evolution of urban and regional settlement structures". Figure 10 brings some results from these studies, in which the city capacity for building and development is related to population pressure. Figure 10a shows the evolving city capacity when the urban plain is uniform, and Fig. 10b, when it is disturbed in one of its sites.

Pattern Formation and Pattern Recognition in the City

The paradigm of pattern recognition was derived by an analogy to the material process of pattern formation [54]. Haken and Portugali suggested that the synergetic pattern recognition paradigm is specifically attractive for the study of cities [58]. The latter can be perceived as self-organizing systems which are both physical and cognitive: individuals' cognitive maps determine their location and actions in the city, and thus the physical structure of the city and the latter simultaneously affects individuals' cognitive maps of the city. In their preliminary mathematical model Haken and Portugali construct the city as a hilly landscape which is evolving, changing and moving as a consequence of the movement and actions of individuals (firms etc.). The latter give rise to the order parameters which compete and enslave the individual parts of the system and thus determine the structure of the city. The new feature of this exposition is that the order parameters enslave and thus determine two patterns (Fig. 11): one is the material pattern of the city, and the other is the cognitive pattern of the city – its cognitive map(s).

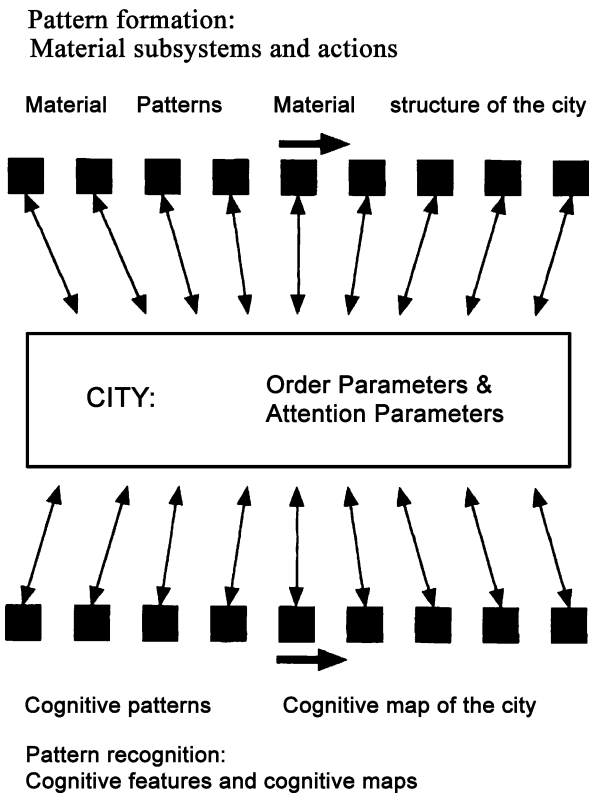

Self-Organization and the City, Figure 11
The city as self-organizing systems which is at the same time both physical and cognitive. Its emerging order- and attention-parameters enslave the city's cognitive and material patterns. Source: [95]

Chaotic Cities

Self-organization is often regarded as a theory about *Order out of Chaos* [110] and yet, with a few exceptions [39,95] chaotic behavior is rarely studied in cities. Most complexity studies of cities perceive cities as ordered structures which the theory of complexity explains just how their order state was created. According to Batty (see p. 29 in [16]) this is "because the required growth rates [for chaotic behavior to appear in cities] are far too large". My view is that this is due, firstly, to the tendency of most students of complexity to focus on the short-term dynamics of Western cities from which perspective cities are indeed structurally stable. However, when the focus of interest turns to the long-term rural urban migration process in a country such as China, or to the archaeological record of the rise and fall of urban cultures [95,105], chaos suddenly appears. Looking at this *longue durée* [26] of cities, their evolution exhibits a very distinct and routinized path: a long period of "steady state", followed by a short period of strong fluctuations or chaos, from which the system re-emerges to a new level of steady state and structural stability, and so on (Fig. 12). As can be seen, the urban system moves from one structurally stable state to another, via bifurca-

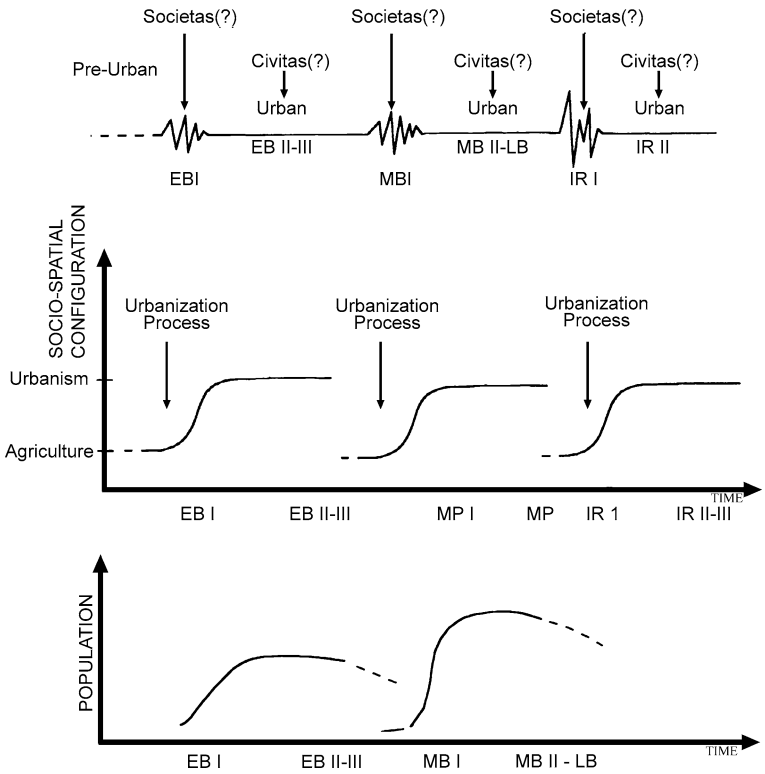

Self-Organization and the City, Figure 12

The evolution of the settlement system in Palestine from the Early Bronze Age period to the Iron Age. Source: [100]. *Top*: A description in terms of chaos and order. *Middle*: A description of the process as a rhythm between agriculture and urbanism, interrupted by global collapses of the urban system. *Bottom*: Calculated population changes in the Early Bronze and Middle Bronze periods. Source: [93,95]

tions, when every evolutionary move is a transition from a microscopic chaotic state to an ordered, macro, steady state.

Secondly, this is due to the fact that phenomena of chaos and their role in cities during their short-term structurally stable periods have not as yet been fully studied. In a preliminary attempt to do so it has been found that often, when the city as a whole evolves stably, a few local unstable chaotic areas are found captive within the otherwise stable city. This phenomenon has been termed *the captivity principle* with the suggestion that it might play a supplementary role to Haken's slaving principle (see Chap. 5.8 in [95], [61]), namely, that these local islands of instability are needed in order to maintain the overall global stability of the city. Figure 15 below provides a hypothetical example simulated by means of cellular automata.

The play between chaos and order might show up also in the daily routines of cities. The movement of cars on the roads, or of pedestrians on pavements, are characterized by shifts between instable and stable motions and might thus be candidates for this kind of interpretation.

Fractal Cities

Mandelbrot's theory of fractals is based on the notions of *self similarity* and the *fractal dimension*, and, on the idea that a rather simple iterative process might produce highly complex geometrical shapes. Using these principles, several scholars have demonstrated, first, that the complex geometries of urban form, growth and evolution, on intra-urban and inter-urban regional scales, can be generated by means of a simple iterative process with a few and simple rules. Second, that many urban structures are self-similar and have fractal structure. The most comprehensive study in this domain is Batty and Longley's [19] *Fractal Cities*, to which one can add studies on urban structure, on the fractal structure of transportation networks [23], on the question "when and where is a city fractal?" [24] and more (For updated survey of studies see [16]). Figure 13 illustrates the evolving fractal structure of the Tel-Aviv metropolitan area from 1935 onwards.

Another important insight implied by fractal cities studies is that a city, or a system of cities, in a steady state

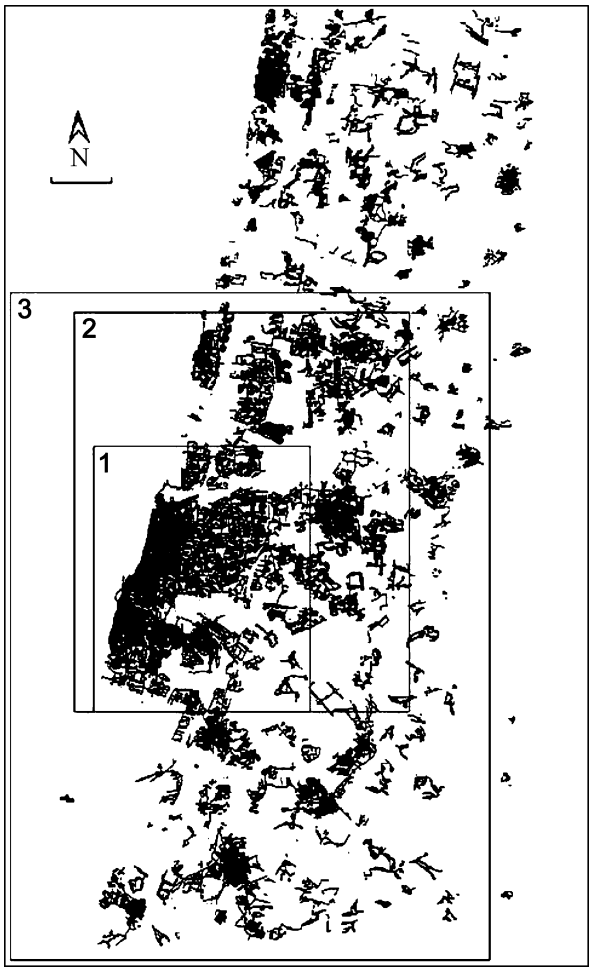

Self-Organization and the City, Figure 13
The evolving fractal structure of the Tel-Aviv metropolitan area from 1935 onwards: The central parts 1 and 2 were fractal during the entire period, while their fractal dimension increased with time. The entire metropolitan area became fractal only after 1985. In 1991 the fractal dimension of the Tel Aviv metropolitan area was found to be 1.667 with error of 0.037. Source: [23]

does not mean equilibrium and stability, as is the case of Christaller's and Lösch's central place theories, for example, but rather a rich and complex evolution and change according to a given ordering principle.

Cellular Automata Cities

The attraction of CA (cellular automata) models to the study of cities is almost self-evident. Real cities are built of discrete spatial units such as houses, lots, city-blocks and the like. CA models are also built of discrete spatial units – the cells. In real cities the properties of local spatial units (e. g. land value) are determined, to a large extent, in relation to their immediate neighbors; so are the

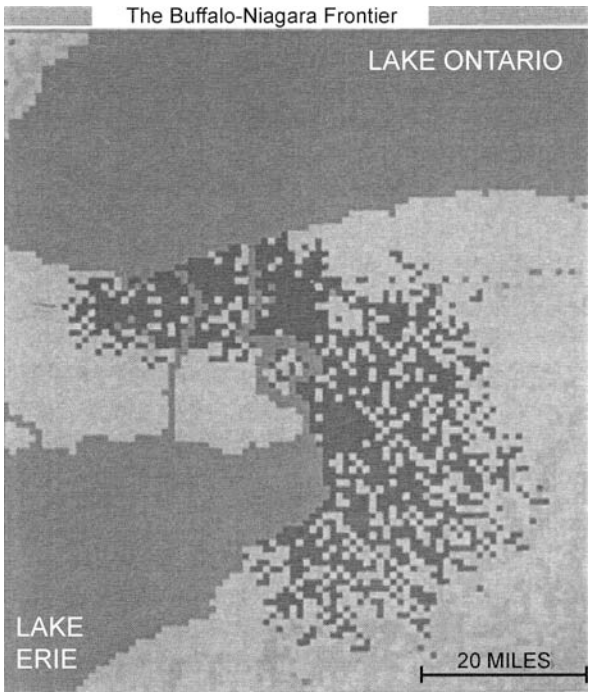

Self-Organization and the City, Figure 14
Cellular automata simulation of the Buffalo–Niagra frontier. Source: [17]

properties of the cells in CA models. These resemblances, together with the mathematical simplicity of CA models, make them natural tools to simulate urban processes. In the last few years CA urban simulation models are among the most popular approach to simulate the dynamic of cities [16,22,95].

One can divide the various models of CA cities into implicit and explicit self-organized CA cities. The first group refers to studies the aim of which is to use the simulation capabilities of the CA city in order to best-fit a certain simulated pattern to an existing one [16]. Figure 14 is an example. The general motivation here is to explain an existing or historical pattern, or alternatively to predict a future one for the purpose of planning. The fact that the model has properties of self-organization, just adds more realism and sophistication to the simulation.

The second group concerns explicit self-organized CA cities. Here the central motivation is to use the model as means to investigate the self-organization properties inherent in cities and urbanism. For example, how micro decisions and behavior of individuals and firms, taken at the local scale, are related to the global behavior and structure of the city. Such models are essentially heuristic and they regard the simulated CA city as essentially a learning device (Fig. 15).

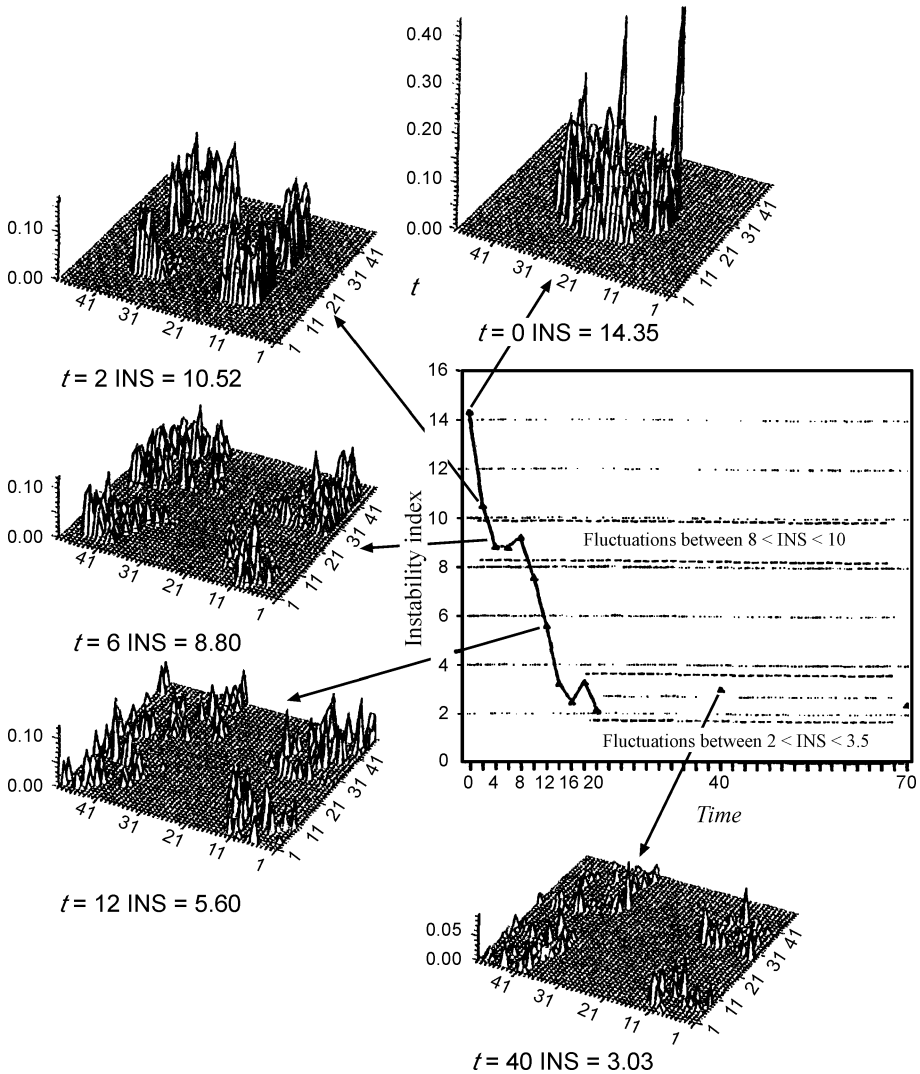

Self-Organization and the City, Figure 15

Time evolution of consecutive stages of SIS (stability-instability surface) in the development of a city with 33% neutral Greens when the rest of the Greens and all the Blues are segregatives. Source: [95]

Because of their iterative structure, CA models can be used as convenient tools to generate fractal structures [16,17] and the insight they add to our understanding of cities is similar: an iterative process guided by a few simple rules, can generate complex structures such as cities [16].

AB and FACS Cities

CA is an efficient tool to model the relations between infrastructure objects of the city. Unlike infrastructure object, urban agents have aims and plans, can learn and move in the city, and see and know beyond their nearest neighbors. *Agent base* (AB) models that are built to imitate such cognitive entities were applied to the city dynamic too. An important source of inspiration here was Schelling's model that demonstrated how local and simple behavior of urban agents can give rise to complex residential segregation in cities – even when their tendency for segregation is minimal [122]. Subsequent agent base studies have supported Schelling's finding [16] and added that a small minority of agents with a tendency for segregation is sufficient to turn the whole city into a segregative structure [95].

Free agents on a cellular space (FACS) models combine CA and AB models [95]. A typical such model is built as a superposition of a CA layer simulating the relationship

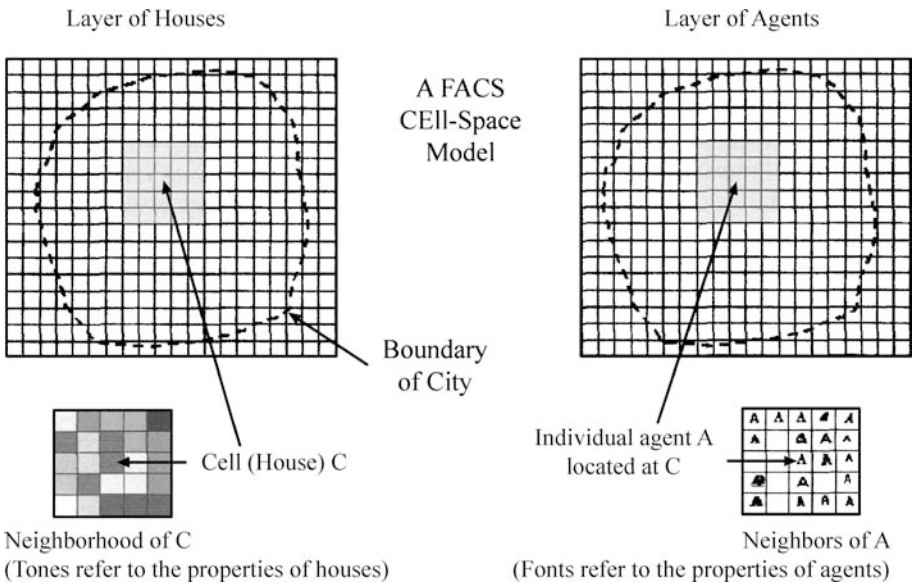

Self-Organization and the City, Figure 16
A typical FACS model is constructed of two-layers: a population layer of human agents describing the migratory and interaction activities of individuals (*right*), superimposed on a CA infrastructure describing the urban landscape (*left*). Source: [95]

between the city's infrastructure objects (buildings, roads …) and AB layer that simulates the activities of the urban agents (Fig. 16). At each model iteration new agent(s) come to the city with a certain intention in mind – say to find a house to live in. The agent then examines the available empty cells/buildings, ranks them according to its set of preferences and picks the best one. Once the agent located itself in a certain cell, the CA dynamics starts: The properties of each cell are determined by reference to the properties of its neighbors; and if the cell is occupied by a certain agent, by some mix between the properties of the agent and its neighbors. Figure 17 presents typical results.

Sandpile Cities

The sandpile, the canonical example of *self-organized criticality* [11,12,16,20], has two incongruous features: the system is unstable in many of its local locations; nevertheless its global state is absolutely robust: The local configurations of the sand change all the time because of the avalanches, while the statistical properties, such as the size distribution of the avalanches, remain essentially the same. Similarly to the sandpiles, cities appear volatile and fast moving at their local scales while at their global scale they appear absolutely robust [16,20]. For example, the size-distribution of many cities and systems of cities remains essentially the same under circumstances such as ongoing population growth (Above Sect. "Image of Cities").

Compared to the "grand" synergetic and dissipative cities, the sandpile city is a kind of a zooming-in to the internal dynamics of self-organized cities in their steady state periods – when they are controlled by what in synergetics is called order parameters. Sandpile cities show how complex and rich is the internal dynamics of a city in steady state (Fig. 18).

Small World Cities

The notion *network* is implicit in all theories of complexity. Recently, Watts and Strogatz [130] showed that complex networks have 'small world' characteristics [84] and Barabasi and Alberst [14] demonstrated that complex networks are scale free thus following the *power law* that according to Barabasi [13] is a mark of self organization.

The link to cities as complex systems was just natural: The view of systems of cities as networks characterized by the power law was indicated above. Single cities too were described as networks. Thus Alexander's classic "a city is not a tree" [1] demonstrated that cities are typified not by a simple *tree network*, but by a complex *semi-lattice network* (Fig. 19). Alexander's view was recently reformulated in terms of the new science of networks [117,118]. Another example is Hillier's *space syntax* that analyzes the morphology of urban spaces in terms of networks [65,66]. Space syntax exposes the way society

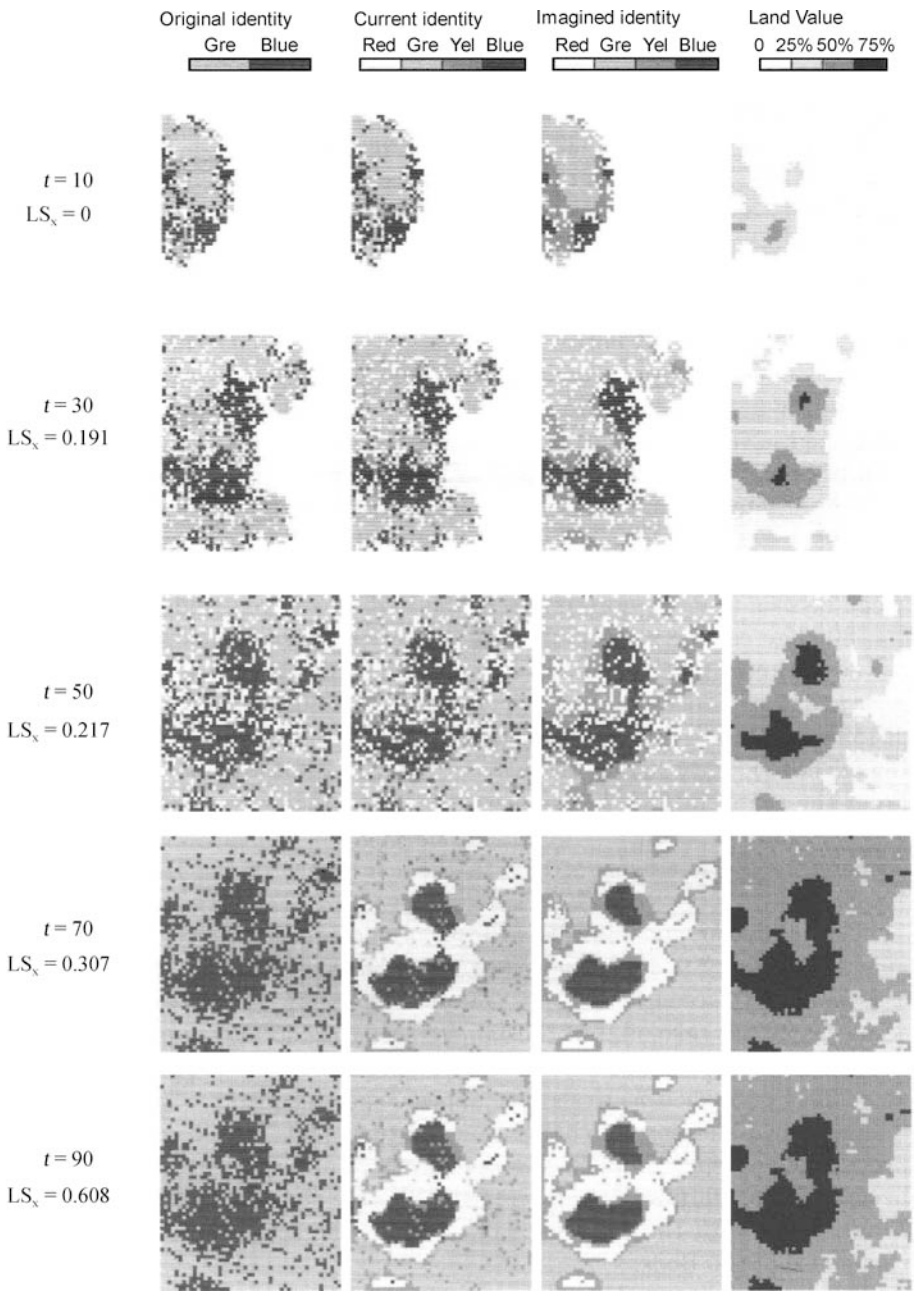

Self-Organization and the City, Figure 17
Several snapshots from an evolving FACS city. Source: [95]

determines the urban morphology and the way the letter feeds back and re-shapes society. The link between space syntax and network analysis has already produced several useful results [38,41,67,89,114].

In the domain of transportation one can mention studies that characterize roads' traffic dynamics in terms scale-free networks (Fig. 20) [68,72], the same was found for the transit system in Beijing [137], pedestrian movement [73] and for the canal networks of Venice [25]. Andersson et al. [9] showed that the market dynamics generates land values that can be represented as a growing scale-free network. Finally, Batty [16] has suggested viewing cities and their dynamics from the combined perspectives of networks, fractals, self-organized criticality and AB.

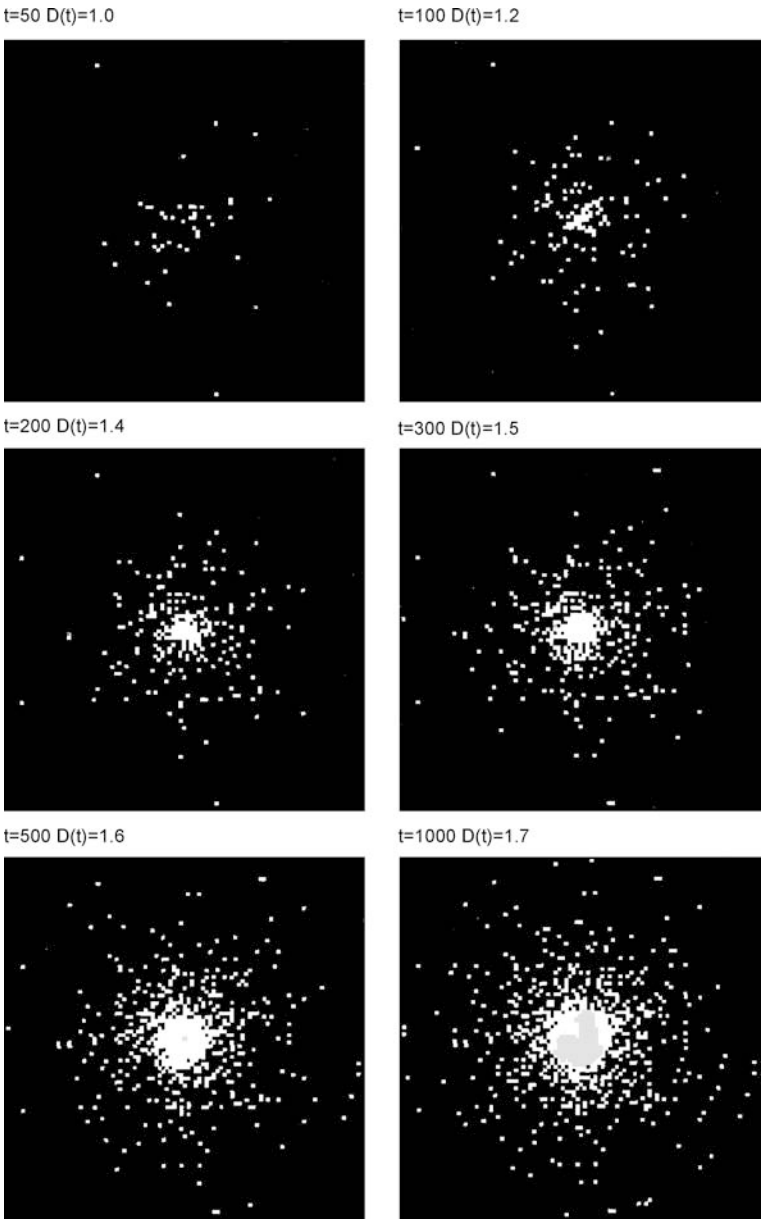

Self-Organization and the City, Figure 18
Self-organized criticality: "Simulation of a hypothetical urban growth pattern in its critical level". Source: [16]

Self-Organization and the City

Self-Organization and the City is an ongoing project that explores the city as a complex system from two interrelated perspectives: Haken's (1983) synergetic theory of complex systems, in particular from the perspective of the pattern recognition paradigm [50,51,55] and IRN – inter-representation nets [94]. The link between the two is termed SIRN – synergetic inter-representation nets.

SIRN – Synergetic Inter-representation Networks

IRN commences with a distinction between cognitively *simple tasks* that can be performed by working memory (e. g. $2 \times 3 = 6$) and *complicated tasks* (e. g. $257 \times 389 = 99\,973$) that are the result of the "magic number seven" that constraints our ability to process information in working memory [85]. One way to overcome this limitation is by means of IRN: We first externalize the task

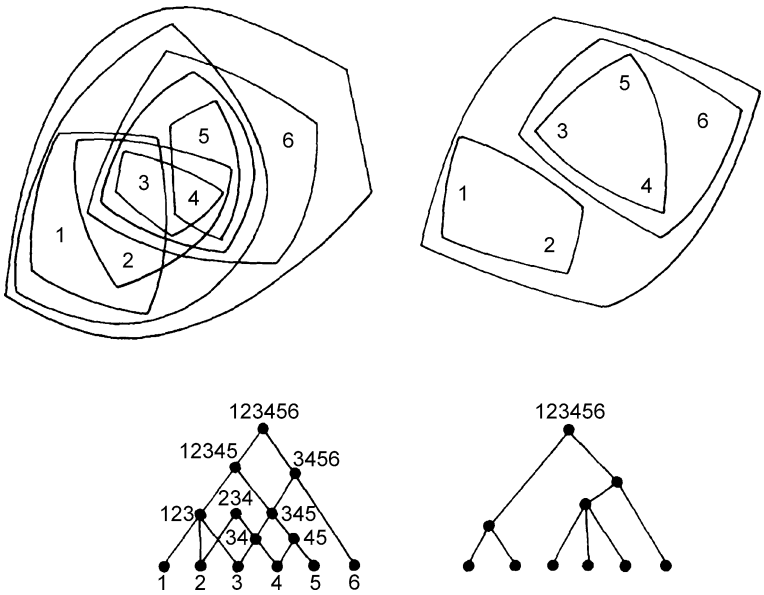

Self-Organization and the City, Figure 19
A tree network (*right*) versus a semi-latice network (*left*). Source: [1]

(write it down on a paper); then we solve part of it internally ($9 \times 7 = 63$); externalize it again and so on in a sequence until the task is completed.

Complex tasks refer to creative cognitive tasks, when a person writes, paints, designs etc. Such a task often starts with a vague idea in mind that the person then externalizes by writing it down or painting … Here too the process proceeds by interplay between internal and external representations, but with one important addition – it involves emerging properties. It is here were synergetics gets in and the process becomes SIRN. More specifically, the process might start with a preliminary internal idea (or external cue that entails internal idea), that the person then externalizes and so on. After a few internal-external iterations an order parameter (in the sense of synergetics) emerges and enslaves subsequent iterations.

The development of the notion of SIRN was inspired by Bartlett's serial reproduction scenarios in his study *Remembering* [15]. A typical such scenario starts when a test person is shown a text or a figure and is asked to reproduce it out of memory (Fig. 21). The result is offered to a second person that is asked to do the same and so on. As shown by Bartlett, at the beginning the reproductions change from person to person, however, at certain stage they stabilize and become a *scheme*. Stadler and co-workers [126] demonstrated that the scenarios proceed as synergetic self-organized process. The focus of interest in the above studies was on the way schemata are created. Haken and Portugali have used the Bartlett scenarios as illustration of the play between internal and external representations [59,94].

It is important to emphasis, first, that external representations are media that enable communication between persons and the emergence of collective SIRNs – e. g. a brain storm. Second, that internal and external representations are *generative* – once produced, they generate new ideas and properties not seen before in previous representations.

The Basic SIRN Model Haken and Portugali [59] have cast the SIRN process into the formalism of synergetics. They started with Haken's [54] 'synergetic computer' (Fig. 22, *Top*), composed as it is of an input layer with model neurons representing the initially given input activity; a middle layer representing the order parameters, and an output layer with neurons representing the final activity of each neuron. The first step is to look at this network from the side, as indicated by the arrow. The result is shown in Fig. 22, *bottom, left*. Adding to the latter external inputs and outputs, we arrive at our basic SIRN model (Fig. 22, *bottom, right*) that has two kinds of inputs, internal and external and two kinds of outputs, again internal and external. The middle node symbolizes the order parameters that emerge out of the interaction between internal and external representations.

The basic SIRN model can be seen as symbolizing a self-organizing active agent that is subject to two flows of information: internal and external (Fig. 23). The first

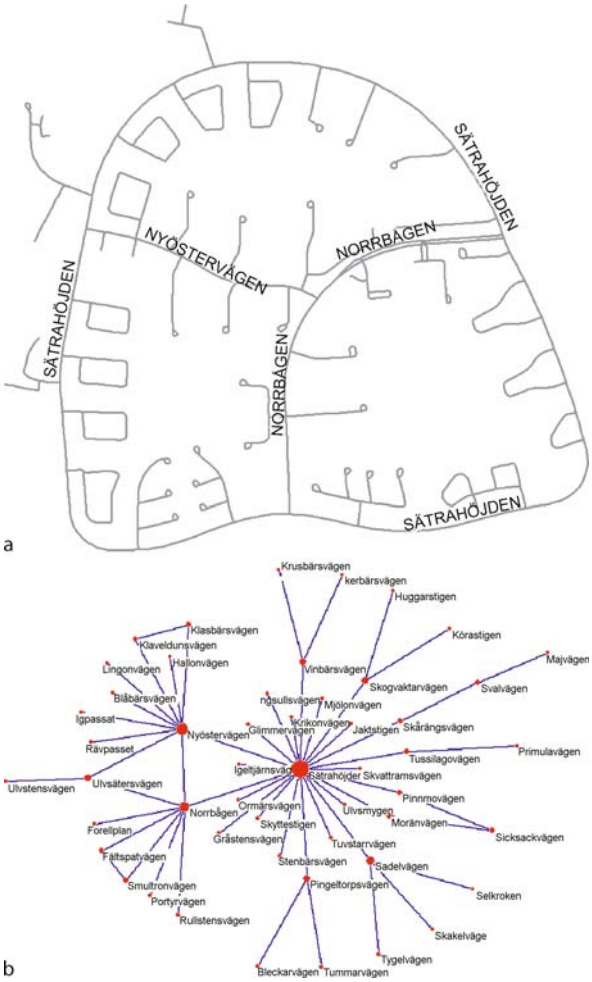

Self-Organization and the City, Figure 20

A small street network (a) and its connectivity graph (b). Every node in b is labeled by the corresponding street name, and the size of nodes shows the degree of connectivity of individual streets. Source: [73]

is coming from the mind/brain, in the form of ideas, fantasies, dreams, thoughts, and the like, while the second from the 'world' – via the senses, the agent's body and/or artifacts. The interaction between these two flows gives rise to an order parameter that governs the agent's action and behavior, as well as the feedback information flow to the agent's mind. 'Action or behavior' may refer to a single individual executing exploratory behavior, reproducing texts or drawing, as well as to several individuals collectively reproducing a large-scale artifact such as a city. In an analogous fashion, the 'feedback information flow' refers to the formation of internal representations, such as images or learned patterns. The order parameters are determined by a competition in line with the synergetics' pattern recognition paradigm noted above. Note that all the above steps (and below), can and have been, performed by a computer so that the approach is entirely operational.

In order to apply the basic SIRN model to specific case studies, Haken and Portugali [59] reformulated it in terms of three prototype sub-models: the *intrapersonal*, the *interpersonal collective*, and the *interpersonal with a common reservoir* sub-models (Fig. 24). The first refers to a solitary agent, the second to a sequential dynamics of several agents whereas the third to a simultaneous interaction. The third sub-model is, in fact, a theory of urban dynamic. The intrapersonal is typical to the way of an artist, for instance, develops her/his work (Fig. 25), whereas the interpersonal to the Bartlett scenario that provided a source of inspiration to IRN (above, Fig. 21).

In the first two sub-models the process depends fully on the biological memories of individuals. In the third sub-model the process depends partly on biological memories, as before, but partly also on externalized non-biological memory that we term a *common reservoir*. This common reservoir of external, artificial and non-biological memory, might take the form of texts, Internet, buildings or whole cities.

Figures 24 (*bottom*) and 26 illustrate graphically this public-collective SIRN sub-model. Each individual agent is subject to internal input constructed by the mind/brain, and external input which is the legible information coming from the common reservoir, that is, the city. The interaction between these two forms of input gives rise to a competition between alternative decision rules that ends up when one or a few decision rules "wins". The winning rule(s) is/are the order parameter(s) that enslave(s) the system. The emerging order parameter governs an external output, which in the case of a city is the agent's behavior and action in the city, and an internal output, which is an information feedback loop back to the agent's mind/brain.

Both the previous sub-model and the present one involve a two-scale self-organization process: an individual-local scale referring to each individual agent as a self-organizing system, and a collective-global scale, referring to the whole city as a self-organizing system. The individual agents by their action and behavior determine the city, which by means of its emerging order parameter(s) enslaves the minds of the individual agents. In the language of synergetics this process is termed *circular causality*. In terms of social theory it is close to notions of socio-spatial *reproduction* and *structuration*. Recent applications show that the common reservoir might be a non-biological externalized memory such as a city[95,96,97,98], a planning

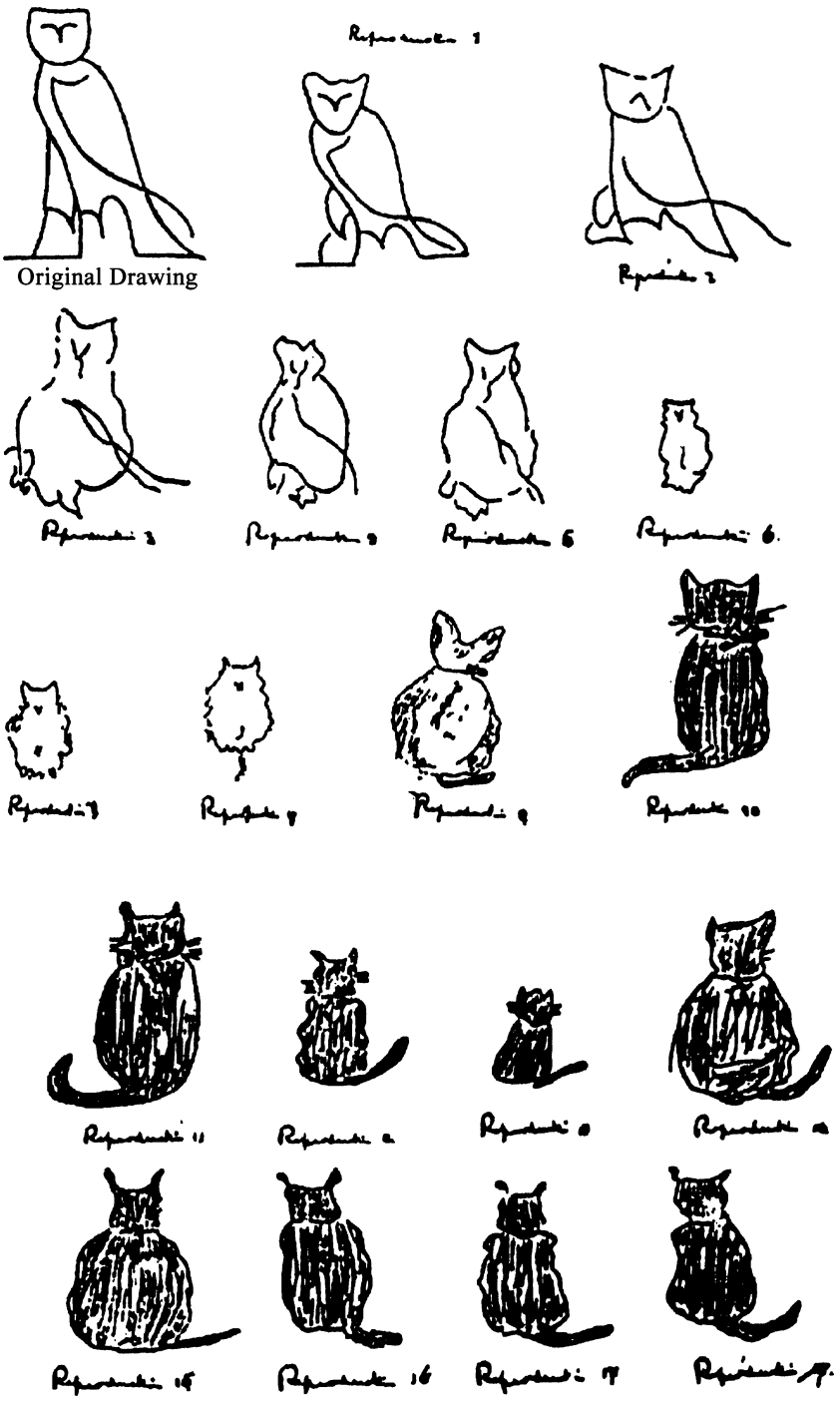

Self-Organization and the City, Figure 21
A Bartlett's scenario of serial reproduction: an Egyptian 'Mulak' (owl) transformed into a cat (see pp. 180–181 of [15])

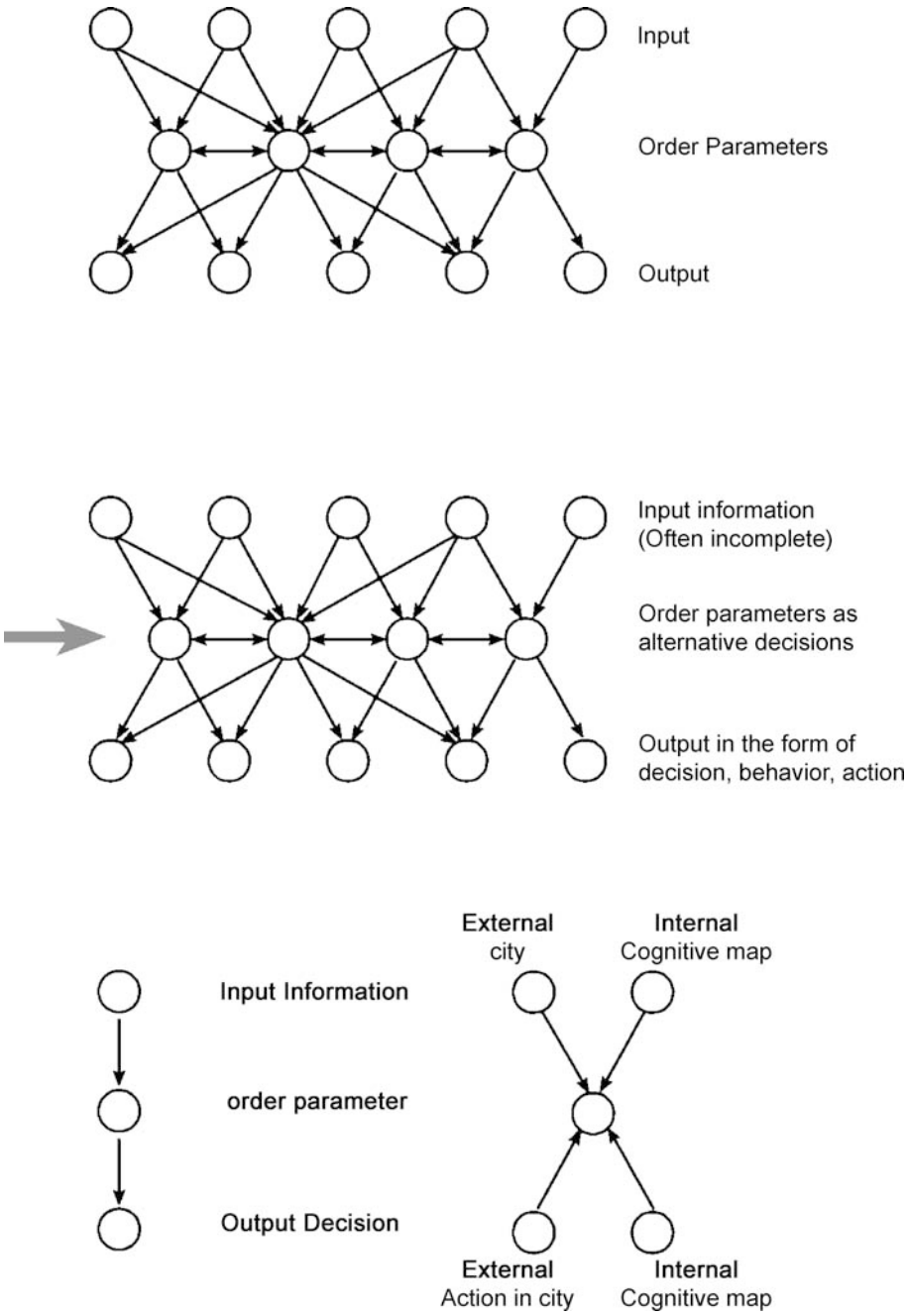

Self-Organization and the City, Figure 22
The derivation of the SIRN model. See text

textual report or an urban planning policy emerging out of a discourse among the members of a planning team [105]. Note that as in the previous model, here too, due to circular causality, as the process evolves the subjective cognitive maps of the individual agents are becoming more similar to each other and an inter-subjective, collective cognitive map emerges. Both private–subjective cognitive maps and public-collective ones are thus *constructions*.

The City Game A simple and effective way to illustrate the SIRN view on the dynamics of cities is by means of a set of experiments termed *city games* [94]. A city game

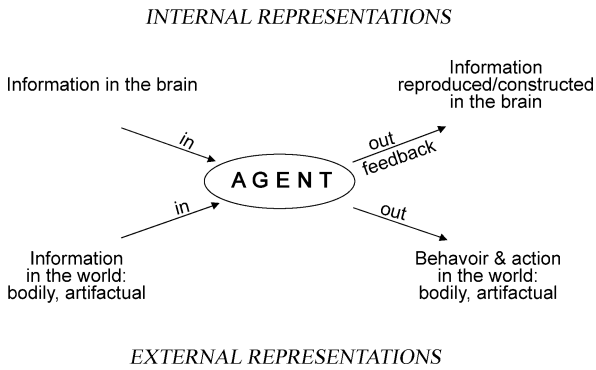

Self-Organization and the City, Figure 23
The SIRN model symbolizes a self-organizing agent that is subject to two forms of information: internal and external, and is actively constructing two forms of information, again internal and external. Graphically, Fig. 23 corresponds to Fig. 22 (*bottom, right*) turned 180° on its NW-SE axis

can be described as a group dynamics that involves some 40 to 70 participants. Their aim is to build a city on a floor, representing the site for a new city. Each player is given a 1:100 mock-up of a building and in his/her turn is asked to locate it in the virtual city on the floor, in what s/he considers as the best location for that building. In a typical game (Fig. 27a and Fig. 27b), the players observe the city as it develops, and in the process also learn the spontaneously emerging order on the ground. It is typical in such games that, after a few initial iterations, an observable urban order emerges. The participants internalize this emerging order and tend to locate their buildings in line with it. As can be seen, the main features of such a game are the main ingredients of SIRN, namely, a sequential interplay between internal and external representations, the emergence of a collective complex city as an artifact and a typical synergetic process of self-organization. Needless to say that the city game is not a 1:1 description of reality, but an illustration of the dynamics of cities as *dual* self-organizing systems.

Cognition and the City

SIRN is at once a theory of cognition, cognitive mapping and urban dynamics. This emphasis on cognition is a direct consequence of complexity theory; a major achievement of complexity theory was to show how local behavior and interaction between urban agents give rise to the global structure of the city. The agent is thus the main and most important actor. Given this, one would assume that practitioners of complexity theory and urban simulation models will have an elaborated theory of agents' perception, behavior, decision making and action; especially so, in light of the fact that a whole body of research on agents' behavior was readily available. I'm referring to studies on spatial cognition, spatial behavior and cognitive mapping that were developed on the interface between cognitive science and urban studies [45,77,92,98,99,103]. And yet, with few exceptions such as SIRN, this body of theoretical and empirical studies is largely overlooked by students of complexity theory of cities. Researchers in this field tend to follow economists by assuming that individuals behave in space as simple "economic persons". The result is that the rather *simple* behavior of agents in the models contradicts the *complex* behavior revealed by studies on cognitive mapping and spatial behavior.

In *The Sciences of the Artificial* Simon [124] suggested that the observed complex behavior of human agents, guided as it is by aims, plans, intentions, needs, policies and so on, misleads us as it is only an external appearance of innately simple behaving entities: Similarly to simple animals, we humans as

> "behaving systems, are quite simple. The apparent complexity of our behavior over time is largely a reflection of the complexity of the environment in which we find ourselves" (ibid. 53).

Most AB/CA urban simulation models are built in line with Simon's logic. They typically start with local interactions between agents having a few simple aim(s) "in mind". This interaction gives rise to an urban system, which from iteration to iteration becomes increasingly complex. Complexity is thus understood as a property of the whole global system, but not of its individual parts.

The efficiency of the *simple cause→complex effect* model is apparent. But there is a catch here: Several empirical studies, of animals' and humans' exploratory behavior, for example, falsify Simon's view [96,97,98]. Furthermore, the property of the city as a *dual* self-organizing system implies that the initial conditions of such complex systems are relatively large numbers of interacting parts, each of which is itself a complex system exhibiting complex behavior. Can there then be a science of cities that is not based on Simon's model? The answer is yes! To see how we shall look at the relations between self organization and information.

Information Compression, Inflation, Adaptation

Complexity is a property of systems that exchange matter and *information* with their environment and that their huge number of parts forms networks characterized by complex feedback and feedforward loops that allow intensive flow of *information* inside the system.

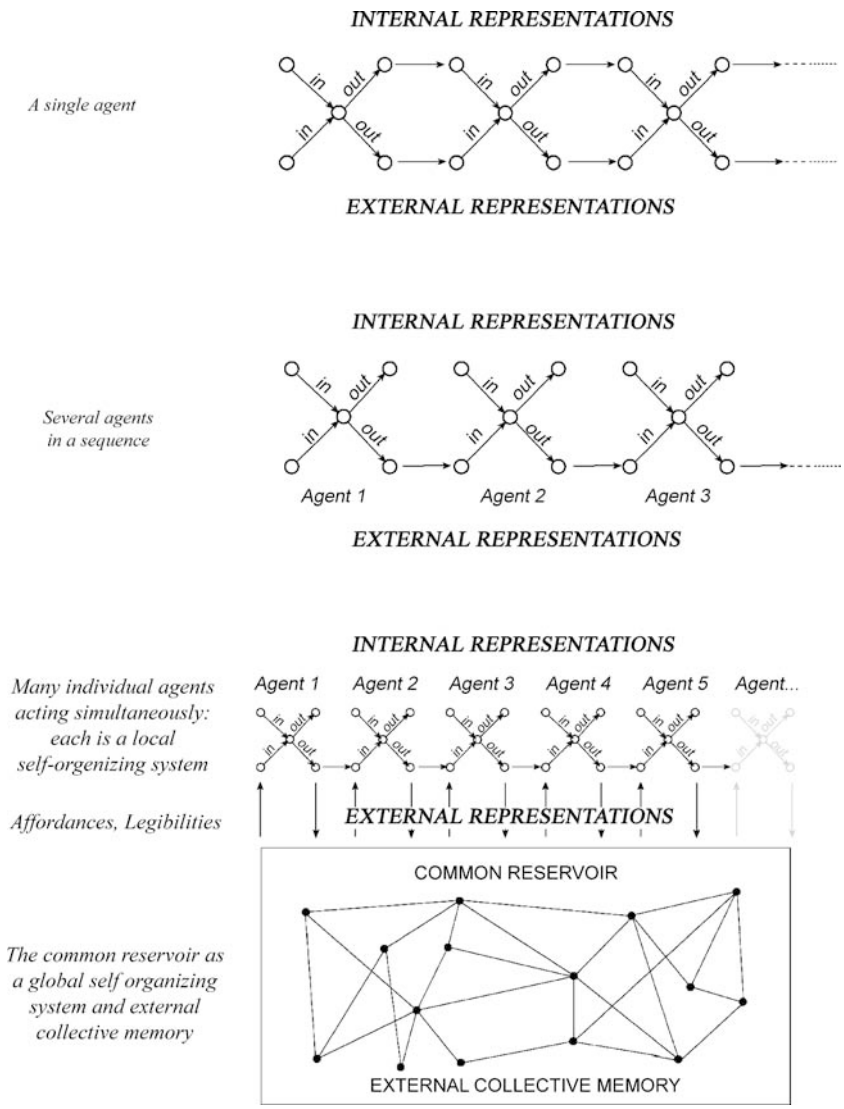

Self-Organization and the City, Figure 24

Top: **The intrapersonal SIRN submodel of a single person.** ***Middle:*** **The interpersonal submodel: serial reproduction of several persons.** ***Bottom:*** **The interpersonal with a common reservoir submodel. Note that in the intrapersonal submodel information is transmitted via external and internal outputs, in the interpersonal via external output only (action and behavior), while in the third submodel information and interaction between the agents are mediated by the common reservoir (e. g. a text, a city, Internet, etc.). Source: [96]**

The notion *Information* is associated with Shannon's theory of information [123] that has played a seminal role in the development of system thinking. In Shannon's theory the notion of information is a pure quantity (usually measured by *bits*) devoid of any meaning. Such a concept of information makes sense only in closed systems where the number of possible states the system can take is finite and a-priori known; hence the link between information and the notion of Entropy, which is a property of closed systems. For example, the information conveyed by throwing a die is 2.5 bits, that is, the logarithm to the base of 2 of the six possible states the system can take. But the complex systems we are dealing with are by definition *open*. So what is the meaning of information in complex systems?

In *Information and Self-Organization* Haken [52] suggested that complex systems 'self-organize', that is, 'interpret', the information that comes from the environment. In other words, the meaning assigned to the message depends on the receiver (the receiving system) and not just on the message itself as in Shannon's theory. Haken (see

Self-Organization and the City, Figure 25
'The Kiss' by Brancusi: from a figurative kiss in 1907, to the geometrical 'Gate of the Kiss' in 1937: An intrapersonal SIRN process in sculpturing

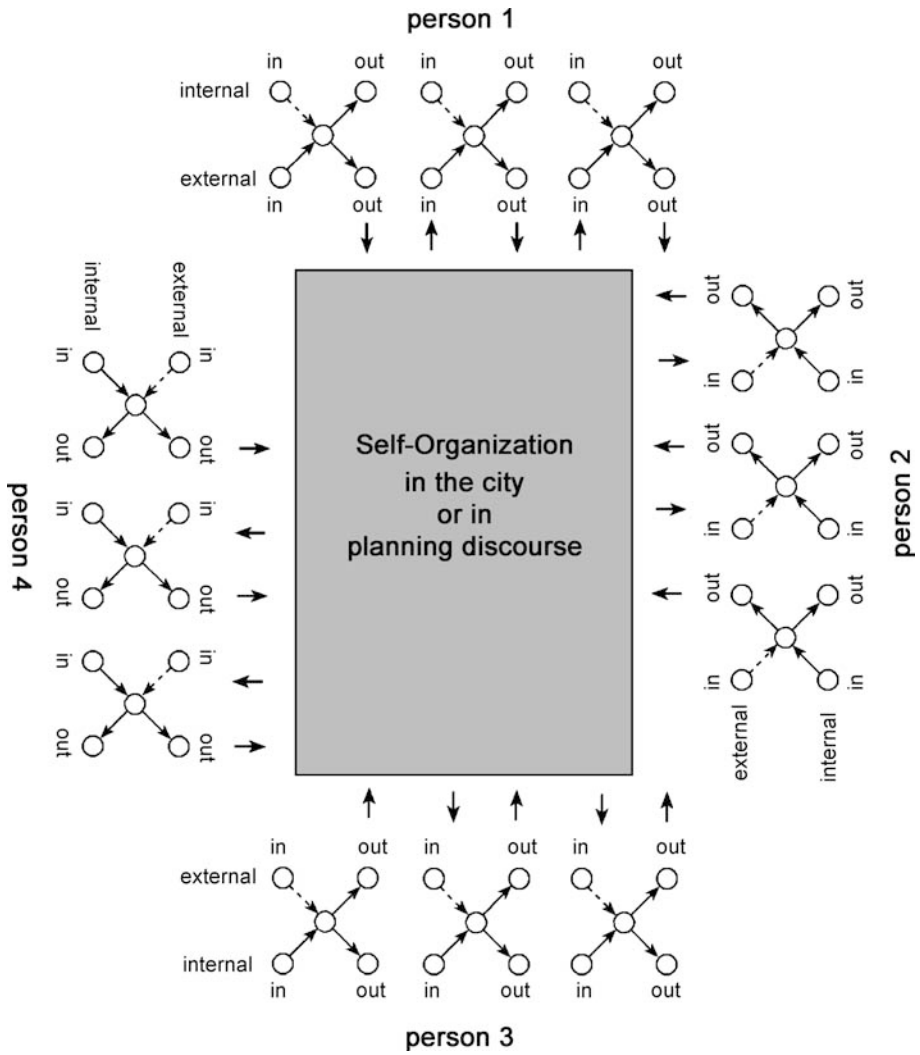

Self-Organization and the City, Figure 26

Another conceptualization of the public-collective SIRN sub-model. Each individual agent is subject to internal input – a cognitive map constructed by the mind/brain and external input –the legible information coming from the common reservoir, that is, the city of a planning team. The interaction between these two forms of input gives rise to a competition between alternative decision rules that ends up when one or a few decision rules "wins". The winning rule(s) is/are the order parameter(s) that enslave(s) the system. The emerging order parameter governs an external output, which in the case of a city is the agent's behavior and action in the city, and an internal output, which is an information feedback loop back to the agent's mind/brain

p. 15 in [53]) has consequently suggested two forms of information: *semantic information* which is information *with* meaning, versus *Shannonian information* which is "information with meaning exorcised". Haken further emphasizes that the process of self-organization implies "an enormous compression of information" (see pp. 25, 35, 151 in [52]).

Haken and Portugali [60] have studied information in the context of the city. They show that different elements of the city transmit different quantities of Shannonian information that can be practically measured by means of information *bits*, for example (Fig. 28). They further show that cognitive processes such as pattern recognition and categorization entail an enormous information compression thus affecting the quantity of the Shannonian information conveyed by the city (ibid) and that information compression is just one facet of the process – the other facet is *information inflation* [62]: In certain urban situations categorization might entail information compression while in others information inflation (Fig. 29). Infor-

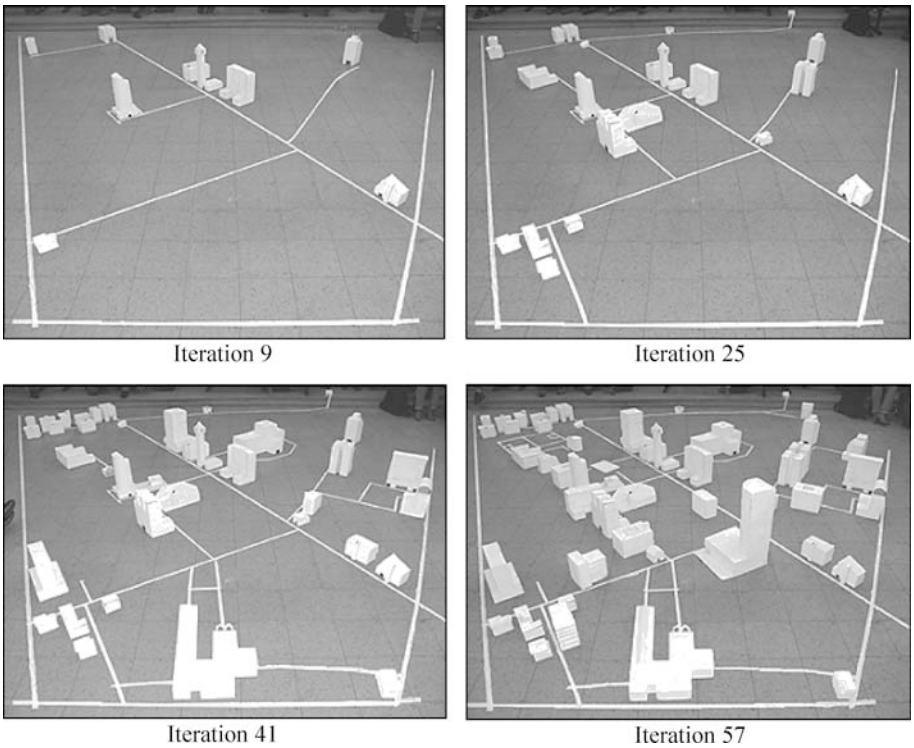

Self-Organization and the City, Figure 27a
Four snapshots from a typical city game (at iterations 1, 15, 35, 57)

mation inflation and compression are thus two aspects of the process of *information adaptation* by which individuals and collectivities shape the city in a way that is adapted to the inhabitants' cognitive capabilities.

The notion information adaptation has far reaching implications to the above discussion about the scientific method and the science of cities: Self-organization as information compression implies a *complex* → *simple* model and thus an alternative Simon's *simple* → *complex* model. The process of information inflation, on the other hand, is in line with Simons' model. These two models are thus two facets of a single process of information adaptation that in some cases requires inflations while in others compression. Complexity theory shows that whatever are the opening conditions (complex or simple) a scientific approach is possible.

CogCity A central property of complex systems is the process of *circular causality* that typifies also the dynamics of cities: Thus in the SIRN model the interaction between the local/micro urban agents gives rise to the global structure of the city, which then feeds back and prescribes the behavior, interaction and action of the agents, and so on. Guided by Simon's *simple* → *complex* model, standard urban simulation models have become excellent tools to simulate the first part of this loop – the way local interactions give rise to a global structure – but they fail to describe the second, feedback part of the loop. CogCity (cognitive city) is a model that attempts to simulate the dynamic of cities as a process of circular causality [98].

CogCity is essentially a FACS model (above Sect. "AB and FACS cities") with several additions that make it an explicit SIRN, cognitive, urban simulation model. It differs from standard AB/CA urban simulation models in that the latter are essentially bottom-up in their structure (Fig. 30, *left*). CogCity, per contra, is characterized by an on-going interaction between top-down and bottom-up. Figure. 30, *right* describes a typical scenario: It starts top-down when an agent arrives to the city with a global cognitive map in mind; compares it to the global structure of the city and selects a local sub-area. Now starts the bottom-up process: the agent selects the empty cells in that local area; evaluates the appropriateness of the cells and their nearest neighbors in light of its needs and then takes a decision and action. In parallel, the properties of every cell are determined according to its relations to its nearest neighbors and so on.

In a regular AB/CA simulation the process ends here: the global outcome is recorded and mapped as the output

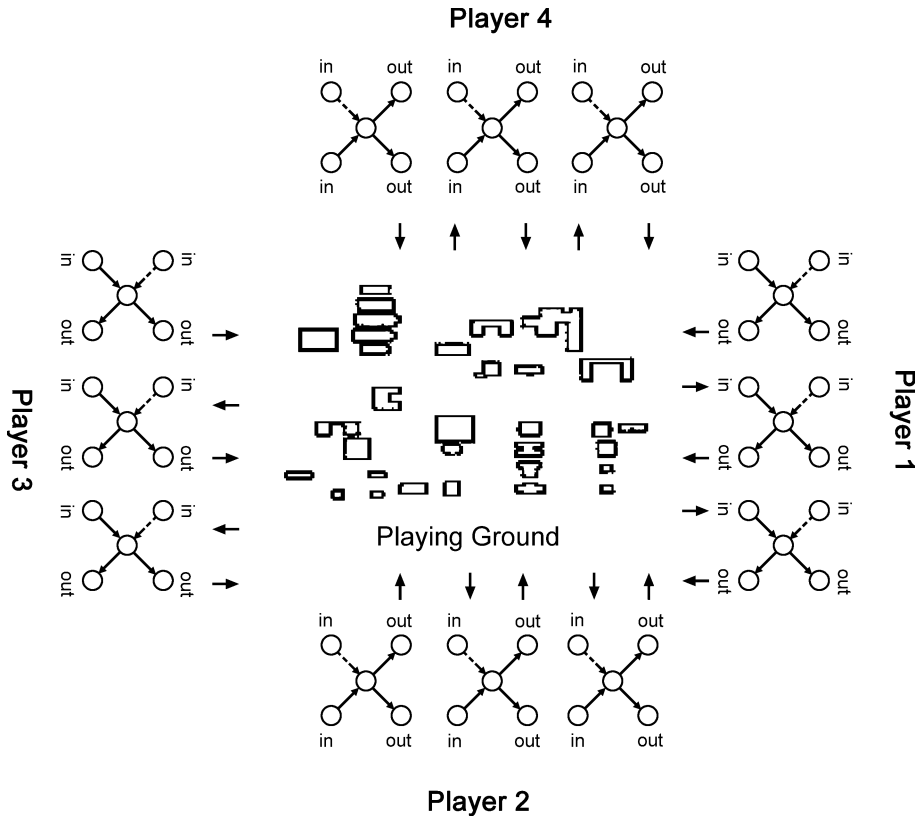

Self-Organization and the City, Figure 27b
A conceptualization of the city game

of this specific iteration and the model is ready for a new iteration. In a SIRN-CogCity model the process continues and feeds back to the global structure of the city that allows the top-down process in the next iteration: Firstly, the state of the various central places is determined. Secondly, peripheries are determined around central places. Thirdly, areas are defined or redefined. Fourthly, subareas are redefined. Fifthly, given areas and subareas, the global state of the city as a whole and its rank-size structure, is defined. The latter changes redefine the local membership state of each cell in the various infrastructure objects and become the externally represented input for a new agent in the next iteration, and so on in circular causality (Figs. 31, 32).

Planning

The link between self-organization and cities is contradictory. Firstly, since cities were always regarded as symbols of planned action – walls, roads, castles, fortresses, indicated a central authority that is capable of planning. Secondly, since planning as means to achieve a controlled order diametrically opposes self-organization as the spontaneous emergence of order. This is a seemingly contradiction, however, since cities are *dual* self-organizing systems with the implication that every urban agent is a planner at a certain scale. This view is supported by psychology and cognitive science that consider planning as one of the basic cognitive capabilities of humans [86].

Cognitive planning, that is, the ability to think, decide and act ahead, must be based on information about the future which by definition is partial and insufficient – a situation that according to Haken typifies also the process of pattern recognition as conceptualize by synergetics. Based on this analogy Haken [57] described decision situation in the context of planning as in Fig. 33.

This decision situation raises the question of 'How do people complement the unknown data?' According to Haken and Portugali, as in pattern recognition tasks here too, the unknown data is being supplied by means of associative memory [57,95], conceptual cognitive maps [98,99] and decision heuristics [95,127,128]. Table 1 specified several decision heuristics and their interpretation in the context of cities.

Self-Organization and the City, Table 1
Seven heuristics and their interpretation in the context of cities. For sources see [95]

Heuristic	Description	Urban interpretation
Similarity	The similarity of two items is expressed as a function of their common and distinctive features.	The recognition of fundamental rules of urban composition, such as a grid layout, built form, or urban fabric, is performed through similarity.
Representativeness	The probability that an object or event belongs to a particular class is judged by the degree to which the description is representative of a stereotype.	Urban categories are identified on the basis of architectural stereotypes, such as church, skyscraper, boulevard, tower, park, arch, etc.
Availability	The probability of an event, or frequency of a class, is assessed by the ease with which instances or occurrences can be brought to mind, or recalled.	Availability would make universal symbols (such as Macdonald's signs, Stop signs, etc.) more easily identified and recalled.
Decision frame	The frame that a decision maker formulates of the problem (gain versus loss, etc.) is influenced by norms, habits, and personal characteristics of the decision maker.	Urban frames for decisions (congested versus free, public versus privet, etc.) depend on the cultural code of each agent (e. g. a tourist, a taxi driver, a policeman, etc.).
Anchoring	The tendency of people to make estimates by starting from an initial base value that is adjusted to yield the final answer.	City's internal representations can contain certain categories such as style, urban violence, town size, etc., which can be 'fired on' early in the process of cognition; once switched on, it stays on and is only eventually reprocessed.
Synergetic I: Collective effects	When facing complex decision situations people tend to rely on what other people are doing or saying.	Drivers, pedestrians, intra- and inter-urban immigrants, tend to 'follow the stream'. That is: to take decisions in line with what others are doing.
Synergetics II: Attention parameters effect	When facing complex decision situations people often employ several heuristics in a sequence. First, the attention parameter calls into use a heuristic. Then, when exhausted, another attention parameter heuristic emerges and so on.	Intra- and inter-urban immigrants, for example, often start with a given location decision heuristic (say synergetic I); if it doesn't work, they switch to an alternative heuristic and so on.

Haken and Portugali [95] have further suggested that complex processes of decision making in the context of city planning evolve according to their SIRN model. This suggestion was further elaborated by Portugali and Alfasi [105] who demonstrated empirically how this SIRN process practically takes place in the reality of *planning discourse* as it evolved among members of a planning team engaged in formulating urban policies concerning the development of the city of Beer Sheva, Israel (Fig. 34).

As a basic cognitive capability planning is intimately associated with the fact that humans are social creatures – people tend to plan together (e. g. families, friends, firms etc). Some planning decisions are thus made solitarily while others collectively. Planning is also a profession that is closely linked to the central authorities of society (municipal, regional, national governments etc.). We thus have three forms of planning – *solitary*, *collective* and *professional*.

The notion that cities are complex self-organizing systems thus implies a novel view on planning the essence of which is, first, that all three forms of planning (solitary, collective and professional) participate in the dynamic of cities. Second, that due to non-linearities that typify cities as complex systems, the act of a single solitary planner might affect the evolution of a city more than the planning act of a professional planning team. (For an example see [101]). Does that mean that due to self-organization there is no need for city planning? Not at all! – It means that we have to adopt a new perception of *plans as participants* in the overall urban dynamics. It also means that we have to adopt a new perception of urban dynamics as a complex interaction between many plans at different scales, or more specifically, between solitary, collective and professional, planning agents, each with its specific plan.

Can there be an administrative planning process that is built in line with the above? The answer suggested by Portugali and Alfasi is positive [105]: In a sequence of studies they have portrayed the principles of such self-planned city [95] and the way it can be applied to the reality of city planning law and structure of Israel [2]. Similarly to current planning systems it is a 3-layers system: the legislative, the judiciary and the executive. It differs in the following: First, its planning law refers to the qualitative relations between the various city objects (Fig. 35) and not to land use plans that assume to determine top-down the urban landscape. Second, it suggests a novel planning judicature

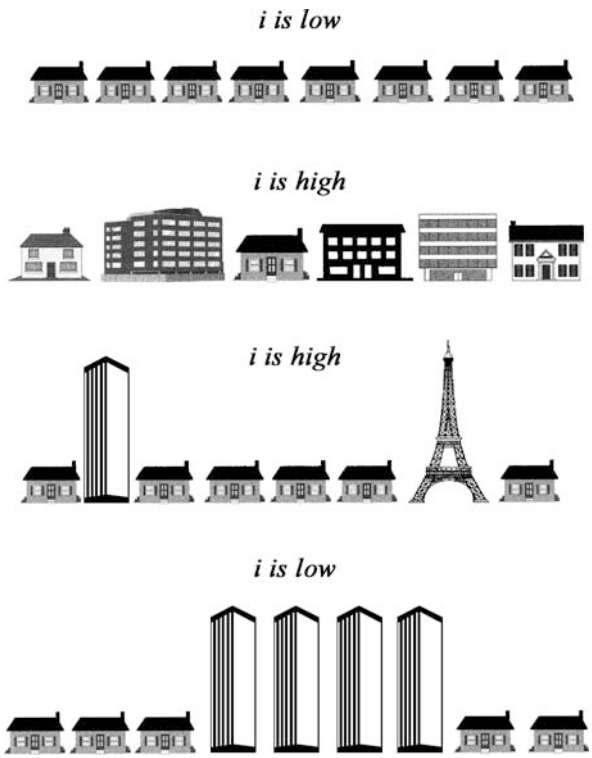

Self-Organization and the City, Figure 28
Different configurations and categorizations of buildings convey different quantities of Shannonian information. When all building are similar, information (*i*) is low. When they are different, *i is high* but difficult to memorize. When landmarks are added *i is high*, provided that they are located apart from each other; otherwise, *i is low*. Source: [60]

composed of spatially distributed "planning courts" conducted by professionals who have specialized in both law and planning. Their aim is to evaluate, accept or reject the plans proposed by all planning agents – solitary, collective or professional. Third, it suggests a separation of authorities that doesn't exist today in standard planning administrations. Fourth, it suggests a process of hermeneutic planning that enables phase transition and adaptation to new situations. Figure 36 describes the structure and operation of this self-organized planning system.

Prediction, Planning, Self-Organization and Cities
The first principle of the above planning system is that its planning laws refer to qualitative relations between the various city objects and not to land use plans. The reason is that land use plans are commonly based on predictions. This is problematic since prediction in the context of complex systems such as cities is associated with four fundamental properties. First, the nonlinearities that typify cities imply that one cannot establish predictive cause-effect relationships between some of the variables. Second, many of the triggers for change in complex systems have the nature of unpredictable mutations [4]; not because of lack of data, but because of their very nature. Third, unlike closed systems, in complex systems, the observer, with his/her predictions, is part of the system – a point made by Jantsch [71] more than two decades ago and largely ignored since then. In such a situation, predictions are essentially feed-forward loops, affecting the system and its future evolution with implications that include self-fulfilling and self-falsifying or self-defeating predictions [101].

From the above follows a dilemma: complex systems are in essence unpredictable and yet, the current practice of planning as well as planning administration and law are based on the ability to predict. In a recent paper (ibid) it was shown that this situation leads to planning paradoxes that are the result of phenomena of self-fulfilling and self-falsifying predictions. It was further shown that these phenomena are the result of the *feed-forward* and *feedback loops* that are typical of complex systems in general and of cities and regions in particular. The existence of such loops is one of the properties that make systems complex. Such loops are responsible to the situation by which a prediction or a plan, once produced, becomes a *participant* in the system's dynamic.

Another way to look at this issue is from the point of view of the distinction between Shannonian and semantic information (above, Sect "Information Compression, Inflation, Adaptation"): Predictions and plans are essentially kinds of information transmission. One can thus speak of Shannonian prediction and semantic prediction. In the first, the outcome of the prediction is independent of the receiver(s) while in the second it depends on the meaning attached to it by a receiver or receivers. A weather forecast is a good example for both: it has no effect on the climatic system, but it might affect the urban system – following the prediction people might behave in different ways that might entail phenomena of self-falsifying and self-fulfilling predictions as described above.

Planning theory has not as yet internalized the implications of complexity theory to city planning. For example, in the planning and decision support systems (PSS, DSS) that are currently discussed and built by proponents of the complexity paradigm, urban simulation models are assumed to function as sophisticated prediction devises [28,29,43]. The result is a discrepancy that to my mind characterizes the domain of urban and regional planning: On the one hand, planning theory, as well as the structure of planning law, practice and administration, are all based on the (usually implicit) assumption that

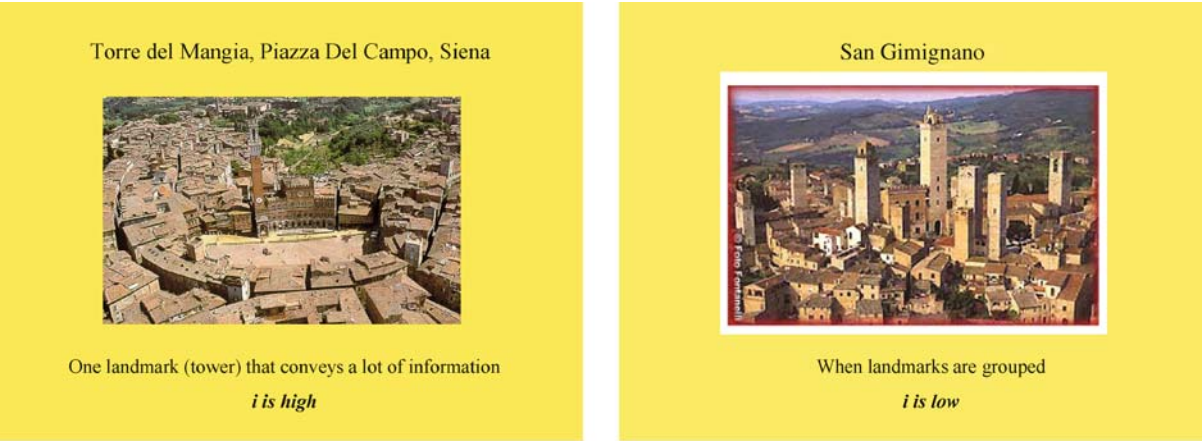

a b

Self-Organization and the City, Figure 28a, b
a An example of a good landmark. b An example of not very effective landmark

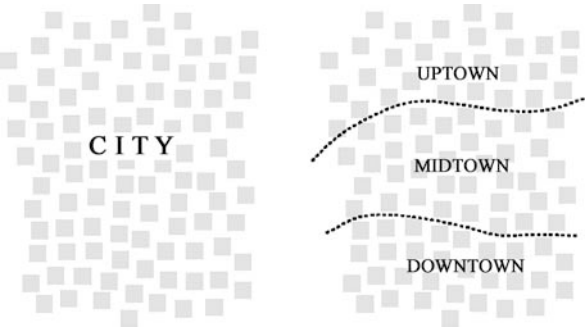

Self-Organization and the City, Figure 29
If all buildings in the city are different from each other, categorization entails information compression; if they are similar – information inflation

cities are essentially predictable entities; that given sufficient data, information and models, their future behavior is in essence predictable. On the other hand, current urban theory suggests that cities are complex, self-organizing and non-linear systems and that as a consequence their future behavior is in essence not predictable; even if sufficient information and data is collected and available [95].

Urbanism

In Sect. "Explicit Attempts to Define a City" we've defined the category "city" in terms of a family resemblance according to which a settlement becomes a "city" not by having some necessary and sufficient properties, but by being a member in a network of cities that has center, periphery etc. This view is in line with cognitive science's approach to concepts and categories. In the latter it is common also to distinguish between *basic level* categories and *super-ordinate* categories [116]. A 'chair', for instance, is a basic level category whereas 'furniture' a super-ordinate. The suggestion here is that a *city* is a basic level category whereas *urbanism* is a super-ordinate one, referring to the totality of cities ranging from their physical structure, architecture, economics, politics, social and cultural composition, and so on. The term 'urban revolution' (coined by Childe, [35]) thus implies a major transformation in society with the basic level category 'city' at its center.

Most complexity studies of cities have traditionally focused on specific aspects of cities – land-use, morphology, transportation, social segregation etc. – but not on the totality of city life which is what urbanism is all about. Why? Because they evolved mainly out of regional sciences' attempt to develop a scientific approach to cities and the consequent tendency to choose research issues that can be analyzed by reference to 'real world' data. The study of urbanism was thus left to the "soft" social theory approaches to cities[33,34,63,64].

This is rather unfortunate because 21st century world society is undergoing a major transformation with urbanism at its center: Massive rural-urban migration and demographic processes entailed a situation by which cities such as Mexico City, Bombay (Mumbai) and Sao Paolo grew from 8.8 million, 6.2 million and 8.3 million respectively in 1970 to over 20 million, over 16 million and again over 18 million today; for the first time in human history, the number of people living in cities is crossing 50% of the world's population and the process is still on. In the last

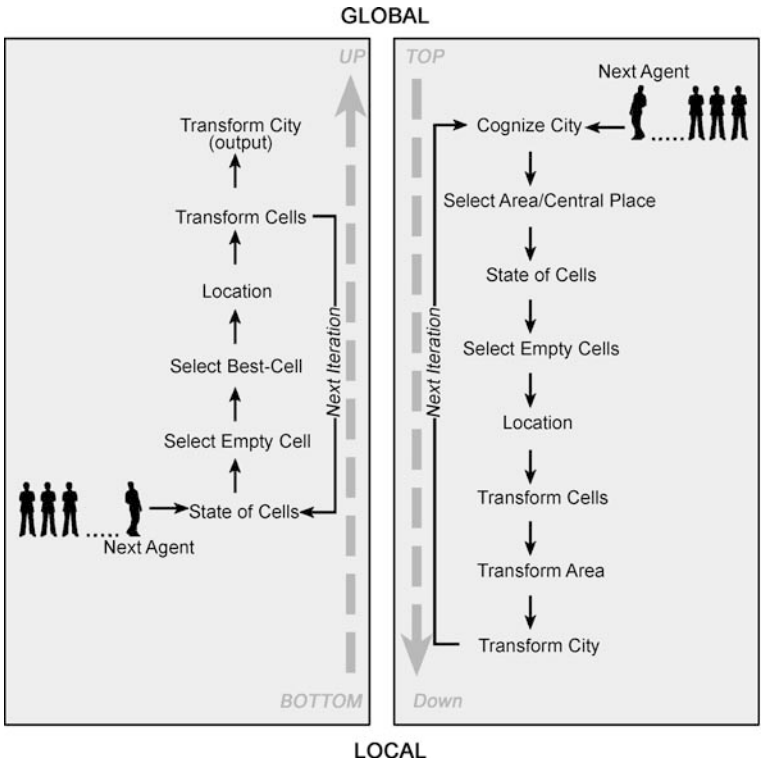

Self-Organization and the City, Figure 30
A cognitive (*right*) vs non-cognitive (*left*) AB/CA urban simulation model. Source: [98]

few decades we've witnessed the emergence of *world cities*, or *global cities*, that form the centers for the globalization process.

These quantitative processes are associated with several qualitative processes: a process of privatization that leads to the decline of the welfare nation-state; the emergence of a civil society that takes over many of the past duties and functions of the nationalist welfare state; the crucial problems of many (post)modern counties are no longer classical national problems (e. g. national self-determination, national boundaries etc.), but rather the problems of cities. The events of September 11 and the ensuing wars in Afganistan and the Middle East are tragic indications to the *urbanization of war*. Finally, the process of globalization is making some world cities more dominant than the states within which they exist thus repressing the nation states.

All the above indicates the more fundamental change: According to Lefebvre [79], its essence is that urbanism is replacing industrialization as the dominant force in society. My view is that the essence of this change is that urbanism is challenging nationalism as the order parameter of modern society [101].

Complexity and Urbanism Complexity studies of cities, with their focus on the short-term dynamics of cities and of national systems of cities are indeed highly advanced in terms of mathematical formalism and data analysis but rather anachronistic in terms of the issues studied; as such they have so far said very little on the dramatic urban phenomena of 21st century. Can they say more about the issue of urbanism? The answer is yes! And for several reasons: To my mind the "deeper messages" of complexity theories is that they have discovered properties in matter hitherto assigned to life, art and society [91]. It is not surprising therefore that complexity theories, particularly synergetics, bear many similarities to social theory and philosophy and, as a consequence, several of the notions that originated in the study of complex systems can be related to similar notions that originated in the domain of social theory [100]:

- Both are essentially systemic and even holistic.
- Both tend to conceptualize 'development' and 'evolution' in terms of abrupt changes rather then a smooth progression. In social theory the common terms for an abrupt change is (social/political/cultural) 'revolution',

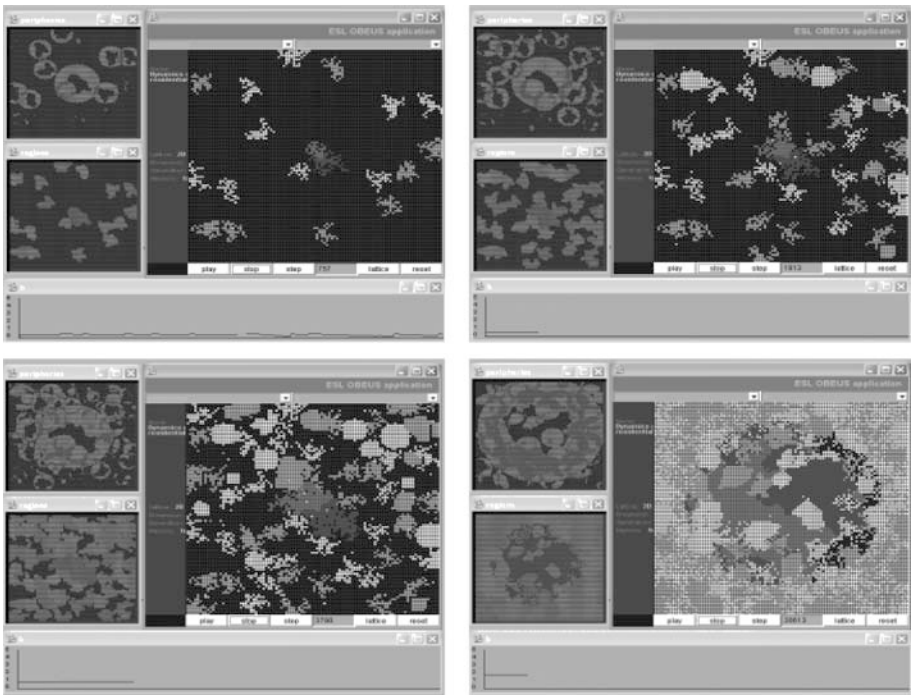

Self-Organization and the City, Figure 31
Preliminary results from an evolving scenario simulated by CogCity: The central screen in each of the four snapshots shows the evolving spatial distribution of various kinds of agents, the top left screen the evolution of centers and sub-centers, while the bottom left, the evolving cognitive maps of agents. Source: [98]

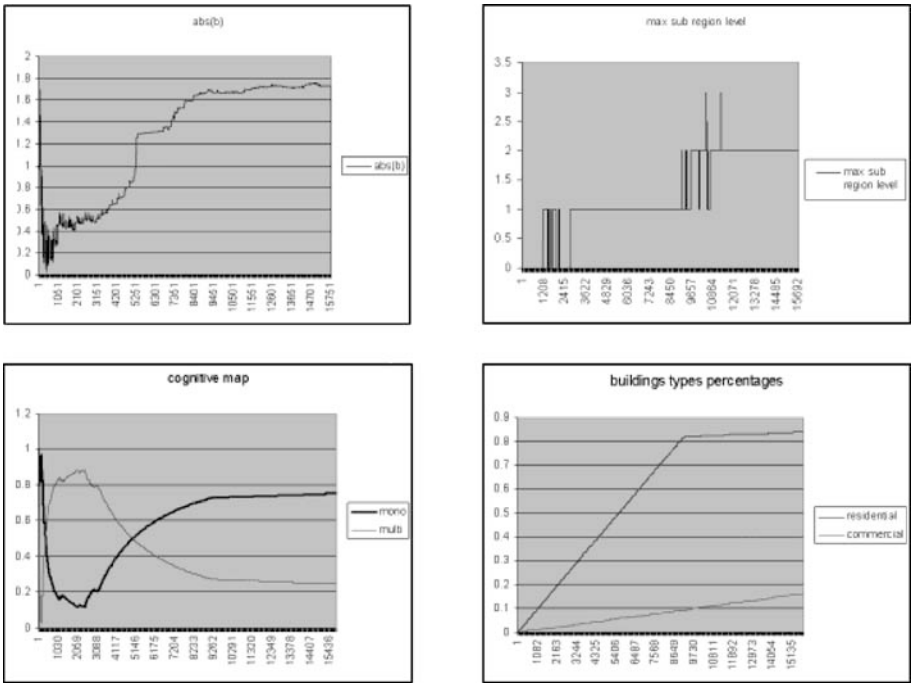

Self-Organization and the City, Figure 32
Preliminary results from an evolving scenario simulated by CogCity – the graphs. Source: [98]

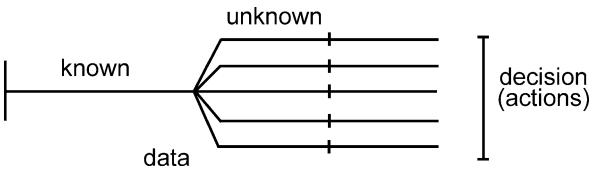

Self-Organization and the City, Figure 33
Decisions in the city are characterized by insufficient data. In such a reality the known data may be complemented in a variety of ways. Each of theses ways might entail a different decision and action. Source: [57,95]

while in the language of complexity 'bifurcations' and 'phase transitions' (that reminds one of Gould's and Eldredge's, [46] *punctuated equilibrium*).

- Synergetics' notion of 'order parameter' is similar to social theory's notion of 'mode of production'.
- Synergetics' notions of 'enslavement' and 'circular causality' are close to social theory's notions of 'social reproduction' and 'socio-spatial reproduction' [44,80].
- Complexity's view of systems in 'a far from equilibrium condition' comes close to postmodernism's recent emphasis on viewing reality as ever changing and transforming; hence the general popularity of notions such as 'chaos' and 'butterfly effect'.

Several writers have already responded to these similarities from the perspective of the sciences, philosophy, media/cultural critics and modern and postmodern social theory [37,75,76,100,113]. A preliminary attempt has also been made to employ synergetics as a complexity theory of urbanism [95,101]. That is, to interpret the current changes in cities and urbanism in terms of synergetics along the following scenario: the combined force of rapid population growth, urban expansion and technological change throughout the 20th century acted as a control parameter. Toward the end of the 20th century and at the beginning of the 21st, we are witnessing a bifurcation and phase transition followed by a competition between the newly emerging *urban order parameter* and the old nationalist one. My personal view is that what we see emerging today out of this competition is not the replacement of nationalism as an order parameter by urbanism as an order parameter, but the *urbanization of nationalism*.

Future Directions

Looking in retrospect at more than two and a half decades of complexity theory studies of cities, one can now appreciate some of its major achievements: First, the link between cities and complexity theory gave urban studies a strong theoretical basis it never had before. The fact that complexity theory was applied to a large number of domains gave urban studies a wide context and many sources of inspiration. The fact that complexity theory comes with a rich and strong mathematical formalism gave urban studies a sound methodological background. The attempt

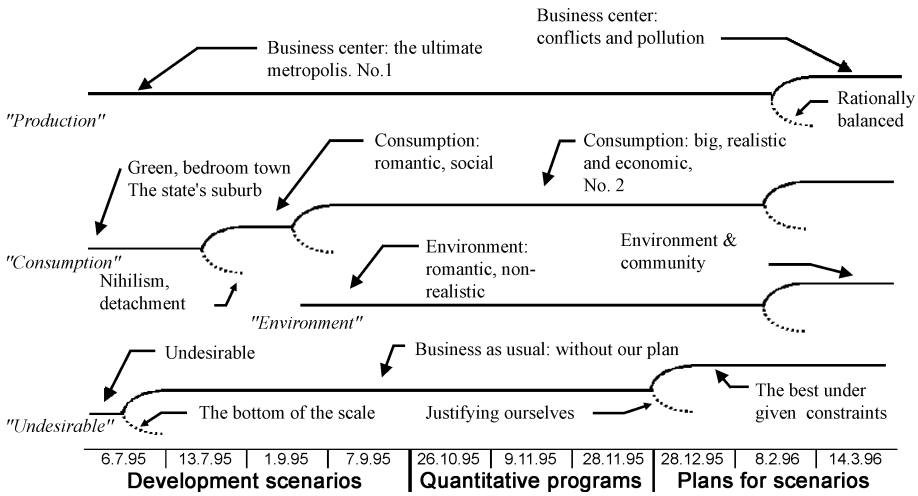

Self-Organization and the City, Figure 34
Bifurcation diagram of the planning discourse. Each alternative is represented with a continuous line along the time axis (the X-axis). A horizontal line represents an order state during which the alternative scenario maintains a certain image and possesses certain attributes. Bifurcation points indicate a shift from one order state to another. The broken lines represent optional order states that were not actualized. Source: [105]

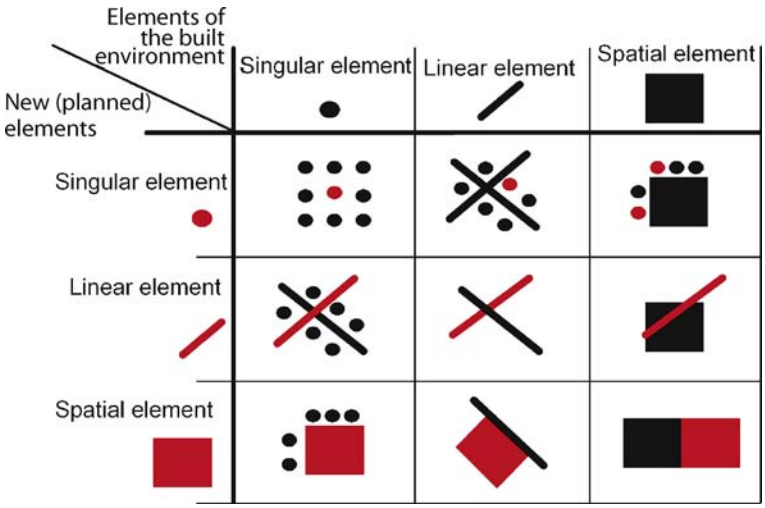

Self-Organization and the City, Figure 35
Proposition for a new planning structure for a self-planned city: The interrelations between urban elements provide the basis for planning law. A singular urban element might be a building, a linear one might be a road, while an example for a spatial urban element is a park

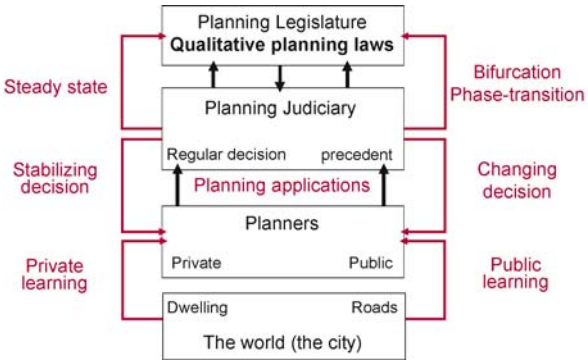

Self-Organization and the City, Figure 36
Proposition for a new planning structure for a self-planned city: The system is built of private planners (the inhabitants of the city) and professional planners. Each of them might submit a plan to the planning judiciary. In the latter the "planning judge" takes decisions according to the planning law as determined by the planning legislature. Unlike the current structure, there is a clear separation of authorities

to transform the study of cities into a science of cities is today closer than ever.

Complexity theory has given us a new insight to our understanding of the dynamics of cities. According to Batty [16] the most important contribution is that complexity studies of cities have verified the intuitive views of Jean Jacobs [70] and Alexander [1], namely, that the complex entity "city", with its variety of different land-uses, socio-spatially and culturally segregated communities, transportation networks and all the rest, is an outcome of "bottom-up" processes: The local interaction between agents at local scale, conducted by very few and simple rules gives rise the complexity we term 'city'.

At the same time, however, it must be admitted that the potential contribution of complexity theories to urbanism, planning and urban design has yet to be realized. The complexity approach has indeed given the bottom-up views on the nature of cities a strong mathematical formalism that can be quantified by real data. But this focus on the local, the bottom-up and the quantifiable was not without price: Cities and urbanism of the 21st century are in the midst of a dramatic transformation, new forms of cities are emerging – world cities, global cities, megacities, and yet the vast majority of complexity studies still focus on the old traditional quantitative urban questions, leaving the qualitative grand urban issues to the "non-scientific" social theory oriented urban studies. Can complexity theories of cities contribute? The answer suggested above is yes! SIRN is one approach in this direction and the field is ripe for others.

The same applies to planning. In the literature on planning theory it is common to make a distinction between *planning theory* versus *theory in planning* [40]. That is, between a theories about how to plan and theories about urban and regional dynamics that planners can use during the planning process. Examining complexity theories of cities from this perspective we see that they are very innovative with respect to theory in planning, but very conservative when it comes to theory of planning: The vast

majority of studies simply ignore the implications of complexity to urban, regional and environmental planning.

Why do we need a complexity theory of planning? The answer is twofold: First, standard planning theory was developed in the 1950s and 1960s hand in hand with what we consider today as anachronistic urban theory: Both are based on the (usually implicit) assumption that cities are in essence simple, mechanistic systems that given sufficient data and advanced technologies, their future behavior is predictable and hence controllable. As we've seen above, complexity theories tell us a different story: Cities are complex self-organized systems that are in essence unpredictable and controllable; even if sufficient data and the most advanced technologies are at hand. From here follow a whole set of new and interesting questions: what is the role of planning in a complex system? Are all parts and components of the system unpredictable? In the above we've suggested some preliminary answers – but new ones must still come.

Finally it is important to mention the issue of extreme events in cities. The rapid processes of urbanization cities underwent in the last few decades and the 'urbanization of war' made cities rather vulnerable areas in cases of extreme events. The question of how cities and their inhabitants behave and respond to extreme events, is a pressing social issue that already started to capture the attention of students of complexity theory of cities and urbanism.

Bibliography

1. Alexander C (1965) A city is not a tree. Architectural Forum April-May, 58–62
2. Alfasi N, Portugali J (2007) Planning rules for a self-planned city. Planning Theory 6(2):164–182
3. Allen PM, Strathern M (2004) Complexity: The integrated framework for integrated models of urban and regional systems. A talk delivered at a conference on The Dynamics of Complex Urban Systems, November 4–6, 2004, Monte Verita
4. Allen PM (1997) Cities and Regions As Self-Organizing Systems: Model of Complexity. Routledge, London
5. Allen PA (1981) The evolutionary paradigm of dissipative structures. In: Jantsch E (ed) The Evolutionary Vision. Westview Press, Boulder, pp 25–71
6. Allen P, Sanglier M (1981) Urban evolution, self-organization and decision making. Environ Planning A 13:169–183
7. Allen P, Sanglier M, Engelen G, Boon F (1985) Towards a new synthesis in the modeling of evolving complex systems. Environ Planning B Planning Des 12:65–84
8. Alonso W (1965) Location and Land Use. Harvard University Press, Cambridge, MA
9. Andersson C, Hellervik A, Lindgren K, Hagson A, Tornberg J (2003) Urban economy as a scale-free network. Phys Rev E 68(3):036124
10. Auerbach F (1913) Das Gesetz der Bevolkerungskoncentration. Petermanns Geogr Mitt 59:74–76
11. Bak P, Chen K, Creutz M (1989) Self-organized criticality in the game of life. Nature 342:780–782
12. Bak P, Chen K (1991) Self-organized criticality. Sci Am 28: 26–33
13. Barabási A-L (2002) Linked: The new science of networks. Perseus, Cambridge
14. Barabási A-L, Réka A (1999) Emergence of scaling in random networks. Science 286:509–512
15. Bartlett FC (1932/1961) Remembering: A Study in Experimental and Social Psychology. Cambridge University Press, Cambridge
16. Batty M (2005) Cities and Complexity: Understanding Cities with Cellular Automata, Agent Based Models and Fractals. MIT Press, Cambridge, Mass
17. Batty M (1997) Cellular automata and urban form: a primer. J Am Planning Assoc 63(2):266–274
18. Batty M (1996) Urban Evolution on the Desktop: Simulations Using Extended Cellular Automata. Unpublished paper, Centre for Academic Spatial Analysis, UCL
19. Batty M, Longley P (1994) Fractal Cities. Academic Press, London
20. Batty M, Xie Y (1999) Self-organized criticality and urban development. In: Portugali J (ed) Population, Environment and Society on the Verge of the 21st Century, pp 109–124
21. Batty M, Xie Y (1994) From cells to cities. Environ Planning B: Planning Des 21:531–548
22. Benenson I, Torrens PM (2004) Geosimulation: Automata based modeling for urban phenomena. Wiley, London
23. Benguigui L (1995) A fractal analysis of the public transportation system of Paris. Environ Planning A 27:1147–1161
24. Benguigui L, Czamanski D, Marinov M, Portugali J (2000) When and where is a city fractal. Environ Planning B 27(4): 507–519
25. Blanchard P, Volchenkov D (2007) Scale-free Segregation in Transport Networks. arXiv:0710.1592
26. Braudel F (1993) A History of Civilizations, translated by Richard Mayne. Penguin Books, New York
27. Blumenfeld-Lieberthal E (2005) Dynamics of Urban Morphology in the Tel-Aviv Metropolitan Area. Ph.D thesis, Technion, Haifa
28. Brail RK (2006) Planning support systems evolving: When the rubber hits the road. In: Portugali J (ed) Complex Artificial Environments, Complexity series. Springer, Heidelberg, pp 307–317
29. Brail RK, Klosterman RE (eds) (2001) Planning Support Systems. ESRI Press, New York
30. Burgess B (1925) The growth of the city: An introduction to a research project. In: Park R, Burgess EW, McKenzie RD (eds) The City. Chicago University Press, Chicago, pp. 47–62
31. Burgess EW (1926/1968) The Urban Community. Greenwood Press, New York
32. Burgess EW (1927) The determination of gradients in the growth of the city. American Sociological Society Publications 21:178–184
33. Castells M (1977) The Urban Question. MIT Press, Cambridge, MA
34. Castells M (1996) The Rise of the Network Society. Blackwell Publishers, Malden, MA
35. Childe VG (1950) The urban revolution. Town Planning Rev 21:3–17

36. Christaller W (1933/1966) Central Places in Southern Germany. Prentice Hall, Englewood Cliffs, NJ
37. Cilliers P (1998) Complexity and Postmodernism. Routledge, London
38. Dalton N, Peponis J, Conroy-Dalton R (2003) To tame a TIGER one has to know its nature: extending weighted angular integration analysis to the description of GIS road-centerline data for large scale urban analysis. In: Hanson J (ed) Proceedings of the 4th International Space Syntax Symposium. University College London, London, pp 65.1–65.10
39. Dendrinos DS, Sonis M (1990) Chaos and Socio–Spatial Dynamics. Springer, New York, Berlin
40. Faludi A (1973) A Reader in Planning Theory. Pergamon Press, Oxford
41. Figueiredo L, Amorim L (2005) Continuity lines in the axial system. In: Van Nes A (ed) 5th International Space Syntax Symposium. TU Delft, Faculty of Architecture, Section of Urban Renewal and Management, Delft, pp 161–174
42. Garling T, Golledge RG (eds) (1993) Behavior and Environment. North-Holland, Amsterdam, London, New York, Tokyo
43. Geertman S, Stillwell J (eds) (2003) Planning Support Systems in Practice. Springer, Heidelberg
44. Giddens A (1984) The Constitution of Society: Outline of the Theory of Structuration. University of California Press, Berkeley
45. Golledge RG, Timmermans H (eds) (1988) Behavioral Modeling in Geography and Planning. Croom Helm, London, New York, Sydney
46. Gould SJ (1980) The Panda's Thumb. Norton, New York
47. Gregory D (1994) Geographical Imaginations. Blackwell, Cambridge, MA
48. Haag G, Muntz M, Pumain D, Saint-Julien T, Sanders L (1992) Interurban migration and the dynamics of a system of cities. Environment and Planning A 24:181–198
49. Haken H (1979) Pattern formation and pattern recognition – an attempt at a synthesis. In: Haken H (ed) Pattern Formation by Dynamical Systems and Pattern Recognition. Springer, Berlin, pp 2–13
50. Haken H (1983) Synergetics, an Introduction, 3rd edn. Springer, Berlin
51. Haken H (1987) Advanced Synergetics: An Introduction (2nd print). Springer, Berlin, New York
52. Haken H (1988/2000) Information and Self-Organization: A Macroscopic Approach to Complex Systems. Springer, Berlin
53. Haken H (1990) Synergetics of Cognition. Springer, Berlin
54. Haken H (1991/2004) Synergetic Computers and Cognition. Springer, Berlin
55. Haken H (1993) Synergetics as a strategy to cope with complex systems. In: Haken H, Mikhailov A (eds) Interdisciplinary Approaches to Non-Linear Complex Systems. Springer, Berlin
56. Haken H (1996) Principles of Brain Functioning: A Synergetic Approach to Brain Activity, Behavior and Cognition. Springer, Berlin
57. Haken H (1998) Decision making and optimization in regional planning. In: Beckmann MJ, Johansson B, Snickars F, Thord R (eds) Knowledge and Networks in a Dynamic Economy. Springer, Berlin
58. Haken H, Portugali J (1995) A synergetic approach to the self-organization of cities. Environ Planning B: Planning Des 22:35–46
59. Haken H, Portugali J (1996) Synergetics, inter-representation networks and cognitive maps. In: Portugali J (ed) The Construction of Cognitive Maps. Kluwer, Dordrecht, pp 45–67
60. Haken H, Portugali J (2003) The face of the city is its information. Environ Psychol 23:385–408
61. Haken H, Portugali J (in preparation) Synergetics and captivity in a city
62. Haken H, Portugali J (in preparation) Information adaptation
63. Harvey D (1985) The Urbanization of Capital. Basil Blackwell, Oxford
64. Harvey D (1996) Justice, Nature and Geography of differences. Basil Blackwell, Oxford
65. Hillier B (1999) Space is the Machine: A Configurational Theory of Architecture. Cambridge University Press, Cambridge
66. Hillier B, Hanson J (1984) The social logic of space. Cambridge University Press, Cambridge
67. Hillier B, Iida S (2005) Network effects and psychological effects: a theory of urban movement. In: Cohn A, Mark D (eds) Spatial Information Theory. Lecture Notes in Computer Science, vol 3603. Springer, Berlin, pp 473–490
68. Hu MB, Wang WX, Jiang R, Wu QS, Wang BH, Wu YH, (2006) Urban Traffic Dynamics: A Scale-Free Network Perspective, http://arxiv.org/abs/physics/0606086v1
69. Isard W (1956) Location and Space Economy. Published jointly by the Technology Press of Massachusetts Institute of Technology and Willey, New York
70. Jacobs J (1961) The Death and Life of Great American Cities. Penguin Books, England
71. Jantsch E (ed) (1981) The Evolutionary Vision. Westview Press, Boulder, CO
72. Jiang B (2007) A topological pattern of urban street networks: Universality and peculiarity. Physica A 384:647–655
73. Jian B (2006) Small world modeling for complex geographic environments. In: Portugali J (ed) Complex Artificial Environments: simulation, cognition and VR in the study and planning of cities. Springer, Berlin
74. Johnson M (1987) The body in the mind: The bodily basis of meaning, imagination, and reason. The University of Chicago Press, Chicago
75. Johnson (2001) media/cultural critics – postmodern complexity
76. Kellert SH (1993) In the Wake of Chaos. University of Chicago Press, Chicago
77. Kitchin R, Freundschuh S (2000) Cognitive Mapping: Past, Present and Future. Outledge, London
78. Lakoff G (1987) Women, Fire and Dangerous Things: What Categories Reveal About the Mind. The University of Chicago Press, Chicago, London
79. Lefebvre H (1970) La Révolution Urbaine. Gallimard, Paris
80. Lefebvre H (1974) The Production of Space. English translation, 1995. Blackwell, Oxford
81. Lösch A (1954) The Economics of Location. Yale University Press, New Haven
82. Lynch K (1960) The Image of the City. MIT Press, Cambridge, MA
83. Mandelbrot BB (1983) The Fractal Geometry of Nature. Freeman, San Francisco
84. Milgram S (1967) The Small World Problem. Psychol Today May:60
85. Miller GA (1956) The magic number seven, plus or minus two:

Some limits on our capacity for processing information. Psychol Rev 63(2):81–97
86. Morris R, Ward G (eds) (2005) The cognitive psychology of planning. Psychology Press, Hove, East Sussex, New York
87. Nicolis G, Prigogine I (1977) Self-Organization in Nonequilibrium Systems: From Dissipative Structures to Order Through Fluctuations. Wiley-Interscience, New York
88. Park RE (1925/1967) The City. Chicago University Press, Chicago
89. Porta S, Crucitti P, Latora V (2005) The Network Analysis of Urban Streets: A Primal Approach, http://arxiv.org/abs/physics/0506009
90. Popper KR (1959) The Logic of Scientific Discovery. Hutchinson, London
91. Portugali J (1985) Parallel currents in the natural and social Sciences. In: Portugali J (ed) Links Between Natural and Social Sciences. A special theme issue. Geoforum 16(2), pp 227–238
92. Portugali J (1990) Preliminary notes on social synergetics, cognitive maps and environmental recognition. In: Haken H, Stadler M (eds) Synergetics of Cognition. Springer, Berlin, pp 379–392
93. Portugali J (1994) Theoretical speculations on the transition from nomadism to monarchy. In: Finkelstein I, Na'aman N (eds) From Nomadism to Monarchy. Yad Izhak Ben-zvi, Jerusalem, pp 203–217
94. Portugali J (1996) Inter-representation networks and cognitive maps. In: Portugali J (ed) The Construction of Cognitive Maps. Kluwer, Dordrecht, Boston, London, pp 11–43
95. Portugali J (2000) Self-Organization and the City. Springer, Heidelberg
96. Portugali J (2002) The Seven Basic Propositions of SIRN (Synergetic Inter-Representation Networks). Nonlinear Phenom Complex Syst 5(4):428–444
97. Portugali J (2003) SIRN (Synergetic Inter-Representation Networks), Artifacts and Snow's two cultures. In: Tschacher W, Dauwalder J-P (eds) Dynamical System Approaches to Embodied Cognition. World Scientific, Singapore, pp 277–294
98. Portugali J (2004) Toward a cognitive approach to urban dynamics. Environ Planning B: Planning Des 31:589–613
99. Portugali J (2005) Cognitive Maps Are over 60. In: Cohn AG, Mark DM (eds) COSIT 2005. LNCS, vol 3693. Springer, Berlin, pp 251–264
100. Portugali J (2006) Complexity theory as a link between space and place. Environ Planning A 38:647–664
101. Portugali J (2006) The scope of complex artificial environments. In: Portugali J (ed) Complex Artificial Environments: Simulation, cognition and VR in the study and planning of cities. Springer, Heidelberg
102. Portugali J (ed) (1992) Geography, Environment and Cognition. A special theme issue. Geoforum 23(2)
103. Portugali J (ed) (1996) The Construction of Cognitive Maps. Kluwer, Dordrecht
104. Portugali J (ed) (2006) Complex Artificial Environments: Simulation, cognition and VR in the study and planning of cities. Springer, Heidelberg
105. Portugali J, Alfasi N (2008) An approach to planning discourse analysis. Urban Studies, Forthcoming
106. Portugali J, Benenson I, Omer I (1994) Socio–spatial residential dynamics: stability and instability within a self-organizing city. Geograph Anal 26(4):321–340
107. Portugali J, Benenson I, Omer I (1997) Spatial cognitive dissonance and sociospatial emergence in a self-organizing city. Environ Planning B: Planning Des 27:263–285
108. Portugali J, Haken H (1992) Synergetics and cognitive maps. In: Portugali J (ed) Geography, Environment and Cognition. A special issue. Geoforum 23(2), pp 111–130
109. Prigogine I (1980) From Being to Becoming. Freeman & Co, San Francisco, CA
110. Prigogine I, Stengers I (1984) Order Out of Chaos. Bantam, New York
111. Pumain D (ed) (2005) Hierarchy in Natural and Social Sciences, Methodos Series (Hardcover). Springer, Heidelberg
112. Pumain D, Saint-Julien T, Sanders L (1987) Application of a dynamic urban model. Geograph Anal 19(2):152–166
113. Rasch W, Wolfe C (2000) Observing Complexity. University of Mennesota Press, Minneapolis
114. Ratti C (2004) Space syntax: some inconsistencies. Environ Planning B Planning Des 31:487–499
115. Rosch E (1999) Reclaiming concepts. J Consciousness Stud 6(11–12):61–77
116. Rosch E, Mervis C, Gray W, Johnson D, Boyes-Braem P (1976) Basic objects in natural categories. Cogn Psychol 8:382–439
117. Salingaros N (2006) A Theory of Architecture. Umbau-Verlag, Solingen, Germany
118. Salingaros N (2005) Principles of Urban Structure. Techne Press, Amsterdam, Holland
119. Sanders L (1992) Systems de Villes et Synergetique. Anthropos, Paris
120. Sanders L, Pumain D, Mathian H, Guerin-Pace F, Bura S (1997) SIMPOP: a multiagent system for the study of urbanism. Environ Planning B Planning Des 24:287–306
121. Sanglier M, Allen P (1989) Evolutionary models of urban systems: an application to the Belgian provinces. Environ Planning A 21:477–498
122. Schelling T (1974) On the ecology of micro-motives. In: Marris R (ed) The Corporate Society. Macmillan, London
123. Shannon CE, Weaver W (1959/1963) The Mathematical Theory of Communication. University of Illinois Perss, Illinois
124. Simon H (1979) The Science of the Artificial. MIT Press, Cambridge, MA
125. Snow CP (1964) The Two Cultures and a Second Look. Cambridge University Press, Cambridge
126. Stadler M, Kruse P (1990) The self organization perspective in cognition research: historical remarks and new experimental approaches. In: Haken H, Stadler M (eds) Synergetics of Cognition. Springer, Berlin, pp 32–52
127. Tversky A, Kahneman D (1974) Judgement under uncertainty: Heuristics and biases. Science 185:1124–1131
128. Tversky A, Kahneman D (1981) The framing of decision and psychology of choice. Science 211:4538
129. von Thünen JH (1826/1966) von Thünen's Isolated State. An English translation (Hall P (ed)) Pergamon, Oxford
130. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442
131. Weidlich W, Haag G (1983) Concepts and Models of a Quantitative Sociology. Springer, Berlin, New York
132. Weidlich W (1987) Synergetics and social science. In: Graham R, Wunderlin A (eds) Lasers and Synergetics. Springer, Berlin, pp 238–256
133. Weidlich W (1994) Synergetic modeling concepts for sociodynamics with application to collective political opinion formation. J Math Sociol 1(4):267–291

134. Weidlich W (1999) From fast to slow processes in the evolution of urban and regional settlement structures: the role of population pressure. In: Portugali J (ed) Population, Environment and Society on the Verge of the 21st Century. A special theme issue, Discrete Dynamics in Nature and Society, pp 137–147
135. Wirth L (1938) Urbanism as a Way of Life. Am J Sociol 64: 1–24
136. Wittgenstein L (1953) Philosophical Investigations. Blackwell, Oxford. Translated by G. E. M. Anscombe
137. Wu J, Gao Z, Sun H, Huang H (2004) Urban Transit System as a Scale-Free Network. Mod Phys Lett B 18(19–20):1043–1049
138. Zipf GK (1949) Human Behaviour and the Principle of Least-Effort. Addison-Wesley, Cambridge

Self-Organization in Clinical Psychology

Günter Schiepek[1], Volker Perlitz[2]
[1] Institute of Synergetics and Psychotherapy Research, Paracelsus Medical University, Salzburg, Austria
[2] Klinik für Psychosomatik, Universitätsklinikum der RWTH Aachen, Aachen, Germany

Article Outline

Definition of the Subject
Introduction
Dynamic Diseases
Self-Organized Synchronization Patterns in Peripheral Physiological Systems
Nonlinear Dynamics in the Communication of Patient and Therapist
Self-Organization in Human Change Processes
The Concept of Self-Organization Promotes New Information Technologies in Clinical Psychology – The Synergetic Navigation System
The Self-Organizing Brain
Future Directions
Bibliography

Definition of the Subject

Clinical psychology is a sub-discipline of psychology engaged in the description, classification, explanation, and treatment of mental disorders. The primary focus is on psychological methods, models, and topics such as behavior, cognition, emotion, and social interaction with substantial overlap with related areas in psychiatry, psychosomatics, or behavioral medicine. Yet, the main stream in clinical psychology views the etiology of mental disorders, their time courses and susceptibility to psychological treatment still through the magnification glass of linear input-output philosophy of human functions. Owing to this paradigm, linear combinations of variables (as inner conflicts, irrational cognitions, or stressors) trigger the development of psychiatric diseases or disorders in genetically predisposed individuals. Therefore, linear multivariate regression models are assumed be able to predict the probability of falling ill or suffering from disorders. As an important field in clinical psychology, psychotherapy research defines randomized controlled trials as the golden standard of outcome research. Here, patients randomly assigned to different treatment modalities are being compared with respect to the outcome of different tests. In this regard, the input (treatment) is thought to determine the outcome (treatment effects).

Contrary, or rather, supplementing this line of research is the scientific paradigm of self-organization, i. e. the functioning of complex nonlinear systems with circular causality at its center. Gestalt psychology, traditionally concerned with patterns ("Gestalts") in perception, human behavior and interaction (e. g., those prevalent in group dynamics, Lewin [25]) focuses on such self-organization processes. Gestalt psychologists like Wolfgang Köhler (e. g. [28]), Wolfgang Metzger, Max Wertheimer, Kurt Lewin and others can be seen as direct predecessors of modern complexity researchers in psychology [67]. Another root of this development is Jean Piaget's equilibration theory of action-cognition patterns (schemata) describing assimilation-accommodation-cycles of these schemata [49]. During these processes, input from the inner and outer environment assumes the role of disturbing stimulation of individual system dynamics. A third important line of thinking in circular causality comes from anthropological medicine. The "Gestaltkreis" integrates feedback loops between sensorial and actional systems on the one side, and individual and environmental systems on the other side (ecosystemic approach) [84].

Introduction

During the past decades in clinical psychology, it was particularly the transdisciplinary approach of *synergetics* [16] which inspired a specific nonlinear and complexity research on cognition [17,79]), social interaction [41,77], etiology and dynamics of mental diseases (e. g., [57,78]), and psychotherapy (for an overview see Haken and Schiepek [21]). Synergetics describes, measures, and explains the autonomous processes of pattern formation and pattern transitions in complex nonlinear systems. Founded on Haken's fundamental discovery that these processes do not depend of the matter of the sys-

tems they occur in, synergetics became one of the most important inspirations to many scientific fields and topics. Especially, Haken early transferred synergetics to brain research (e. g., [3]), since the brain is an outstanding example of a complex, self-organizing system. Today, it is widely accepted that the brain and a serial computer not only differ profoundly, but there is almost nothing they share: No wonder in light of the more than 10^{11} nonlinear interconnected neurons forming a dynamic mega-network of neural networks with essential features like arrays of emerging and submerging synchronizations, its flexibility and ever changing pattern formation, working at the edge of chaos, or realizing combined (activating and inhibiting) feedback mechanisms following the principles set forth by synergetics which describes the laws of self-organizing systems [18,19].

Taking a closer look at most of the phenomena clinical psychology is concerned with it becomes obvious that they appear to be of dynamic nature. Human development processes, human change and learning processes, the dynamics and prognosis of mental disorders, problems manifesting in social systems like couples, families, teams, or the question of how psychotherapy works: Self-organization is a ubiquitous entity.

Dynamic Diseases

Mental disorders are characterized by specific dynamic patterns, mirroring "endogenous" and common features of a disorder (like the repetitive phases of unipolar major depression or the bipolar phases of bipolar disorders, oscillating between mania and depression), as well as the effects of an individual life-style including individual coping and treatment efforts. Mental disorders can be conceived as highly structured and coherent states which enslave and thus impair the individual's mental and social functioning. Following the "enslaving principle", emerging order parameters reduce the degrees of freedom in the behavior of the single parts of a system. There is phenomenal evidence that this is the case in many mental disorders. Obsessive-compulsive disorder patients coerced to repeat unwished thoughts or rituals are just but one most impressive example. On the brain level, such pathological states correspond with abnormal synchronization in specific neural networks impairing brain functions. In obsessive-compulsive disorders, cortico-striato-thalamo-cortical feedback-loops are thought to be at the center of the dysfunctional network [55,59], while abnormal synchronization in highly similar neural populations is the source of Parkinsonian resting tremor [34,43].

At times, transitions between different pathological states or between states of health and disease are linear and balanced, at other times they are discontinuous and abrupt, such as in nonlinear phase transitions accompanied by critical fluctuations described by synergetics to occur in physical systems. Such transitions have been reported for unipolar or bipolar cyclic depression (e. g., [1]), and also for schizophrenia [68]. The usefulness of the concept of attractors in psychopathology is best reflected by the final common pathway of different disorders with similar phenomenology and syndromal patterns. Different initial conditions and different qualities and degrees of stressors and vulnerability factors may result in similar pathological end-states on the one hand. But on the other hand, small fluctuations within the intrapsychic or environmental conditions or small differences of some boundary or threshold conditions may result quite different disorders or may decide between health and disease (for a dynamical simulation of major depression see Schaub and Schiepek [56]). The encouraging message synergetics delivers is that while the structure of a generic system may stay unchanged, small changes in control parameters, threshold conditions, and internal or external fluctuations are able to trigger dramatic changes in the behavior of the system. As a consequence, therapy exerting changes of these parameters is thus able to trigger return of the system to a healthy state.

For illustrative purposes we present results of a computer simulation of different chronic courses of schizophrenia [57]. A qualitative network model of five macroscopic variables was transformed into a set of nonlinear difference equations, with each equation describing and determining the change rate of each variable from t to $t+1$. The empirical references of the simulation model were empirical studies of the chronic course of schizophrenic patterns, mostly mixed psychotic episodes, healthy functioning, and chronic states. For example, Ciompi and Müller [12] report on eight different patterns in the long-term evolution of schizophrenia, most of them reproduced by our model. These patterns result from various combinations of slow vs. acute onset, acute episodes vs. progressive deterioration, and remission vs. chronic end-state.

Variables taken into account were chosen from reviews in psychiatric and psychological schizophrenia research (e. g., [6,7,11]). Selected were (1) degree of cognitive disorders, (2) emotional and interpersonal stress, (3) withdrawal and social isolation, (4) degree of expressed (negative) emotions in the social environment of the patient, and (5) positive symptoms like delusions and hallucinations. The parameters mediating the nonlinear in-

terplay of these macroscopic variables or order parameters were (a) diffuseness of affective-cognitive schemata as a central long-term vulnerability of mental functioning, (b) dopamine and serotonin metabolism, (c) social deficits and lack of competencies, (d) genetic risk for schizophrenia, and (e) some parameters mediating mixed feedback processes, especially the negative feedback responsible for antipsychotic damping effects of the pathology. Results of the simulation runs are indicated in Fig. 1. The simulation reproduces most precisely (a) episodic patterns with prodromal symptoms and acute onset, (b) acute onset, but continuous evolution with chronic end-state, and (c) slow and smooth onset with chronic long-term course (see also [76]).

Other simulation models are effective on the microscopic level of neural networks. Kruse et al. [27] introduced a model focusing on the coupling dynamics between neurons processing brain correlates of social experiences. If unable to learn from cues delivered by the relevant environment, this system will fail to establish adaptive and coherent structures. When inducing fluctuations which promote re-learning and self-healing processes, the neural network causes incoherent and chaotic behavior. Most current models of schizophrenia take into account the neural circuits of relevant brain regions (cortical areas, basal ganglia like striatum and pallidum, thalamic areas, brain stem centers) and particularly the equilibria between different neurotransmitters and neuromodulators (e. g. [9]). The complicated local balances and their (non-) equilibria states are in the focus of the strongly evolving field of computational or systems neuroscience [15,44].

Not only central neuroscience has benefitted from concepts introduced by synergetics, however, but also physiology studying the effects of the autonomic nervous system (ANS) activities on peripheral systems, such as the cardiovascular, the respiratory, or the microcirculation system. Such activities are most prominent as the ANS engages in the mediation of emotions.

Self-Organized Synchronization Patterns in Peripheral Physiological Systems

For decades, the study of the ANS involvement in emotional arousal and its impact on the cardiovascular system has attracted clinical and scientific attention in psychology and psychophysiology (e. g. [65]). Ever since the seminal findings of W. B. Cannon [8] and H. Selye [63] on the general adaptation syndrome, colloquially condensed as stress, the clinical relevance of emotional responses became ultimately clear. This obvious clinical relevance is thwarted by the fact that direct observations of ANS activity in humans are restricted not only because of ethical constraints but also because of the fear to provoke what they strive to detect. Therefore, the study of the ANS in humans had to rely for a long time preferably on indirect measures, such as the power spectral density (PSD), a linear computation method. Based on the fast Fourier transform (FFT) which extracts periodic components in the frequency domain, the PSD was favored by many researchers due to its computational ease to analyze frequencies inherent in the two branches of the ANS, the parasympathetic (PNS) and the sympathetic nervous system (SNS). This allowed to divide the effects of the PNS and SNS activity on variations of the heart rate, the so-called heart rate variability (HRV), into three major variance components, the very low frequency (VLF) band below 0.04 Hz, the low frequency (LF) band between 0.04–0.15 Hz, and the high frequency (HF) band between 0.15–0.45 Hz. While the origin of both VLF and the HF bands is not debated, controversy reigns whether the origin of the LF band is attributable to SNS activity or whether it represents a mixture of SNS and PNS activity. Subsequently, some authors propose calculating the LF to HF ratio assumed to reflect the sympathovagal balance (for an overview see [73]). There has been growing discontent and criticism as to the validity of such drawer style classifications based on the consideration that the PSD, or FFT resp., as a linear routine is only able to detect linear properties, that are to some extent included in most physiological signals [23,86]. That, however, should restrict and limit its use since an increasing body of scientific evidence is demonstrating the obvious: In times of adaptation and rapid changes – a hallmark of life and its living systems – healthy ANS activity exhibits nonlinear dynamics necessary to mediate responses appropriate to those change processes. This is particularly the case for emoting as one of the most volatile change patterns.

However, this is not only true for discrete emotion transitions but also for a process crucial for the maintenance of health, namely psychophysical relaxation. Contrary to the rigid scheme depicted above, Perlitz and coworkers have introduced a relaxation model which takes into account adaptive, self-organizing characteristics of the central and peripheral subsystems involved in the psychophysical relaxation process. They scrutinized the physiological conditions and interactions observed with the emergence of a frequency at ca. 0.15 Hz, which in terms of the classical scheme is attributed to the transition between parasympathetic and sympathetic nervous activity. This frequency prevailed at different amplitudes in HRV, blood pressure and respiration, but foremost in the microcirculation of the forehead skin. Using several nonlinear methods, such as wavelet time frequency distribu-

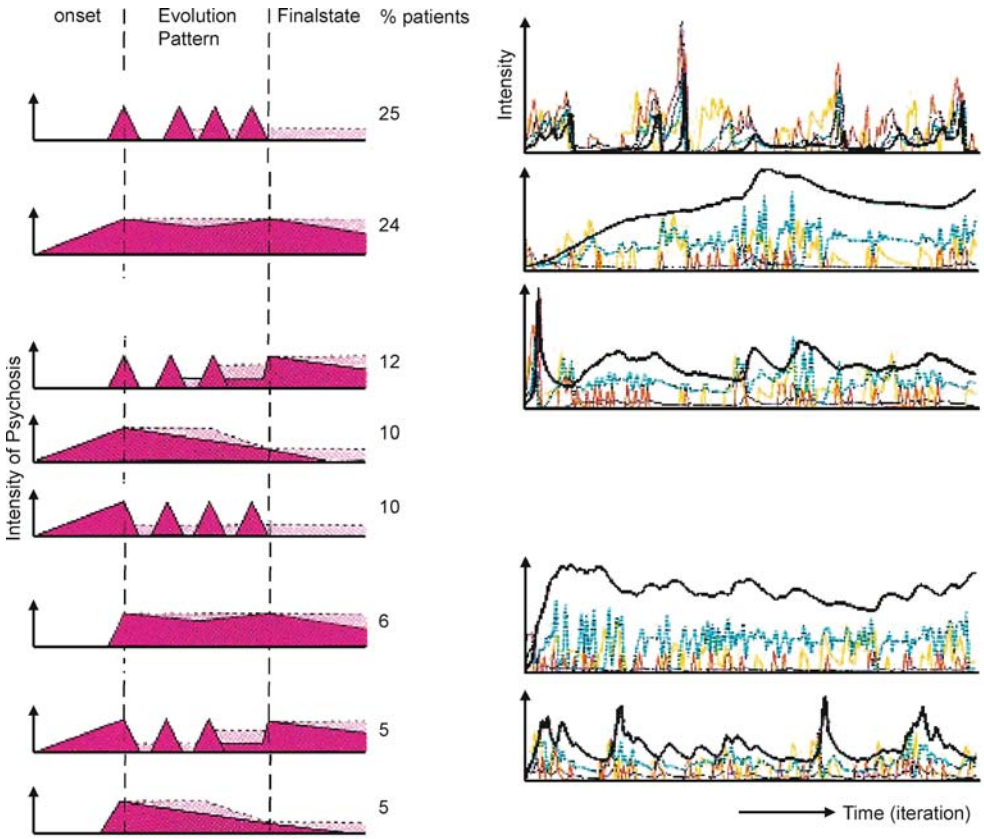

Self-Organization in Clinical Psychology, Figure 1
Different patterns of the long-term evolution of schizophrenia (empirical data from a study by Ciompi and Müller [12]) (*left*) were reproduced by simulations based on a set of five coupled nonlinear difference equations with different parameter values [57] (*right*)

tions (TFD) or post-event-scan (PES) analysis, this 0.15 Hz frequency band (range 0.12–0.18 Hz) emerged or erupted with amplified oscillations and periods of 6–7 s in all time series of subsystems under study. The emergence clearly depended on psychomotor drive reduction which can be either reduced by taking naive relaxation maneuvers (such as closing the eyes), or be enhanced using auto-suggestive means, such as autogenic training. Their zest to elaborate the origin of this frequency was supported by invasive observations with anesthetized dogs made by Lambertz and colleagues who had presented their findings earlier. They found a rhythm at a similar frequency which originated in reticular brainstem neurons of freely breathing dogs when administering narcotics to reduce drive. Followed by the emergence in those unspecific reticular neurons, this frequency also emerged in arterial blood pressure, HRV, and respiration [32,47,48]. This reticular rhythm, termed retR, was unaffected by changes in the frequency of respiration or arterial blood pressure which could both be presumed to exert distinct influences owing to linear models. Rather, in these experiments respiration and HRV were entrained to the 0.15 Hz band at 1:1, 2:1 and 1:2 integer number ratios which are, according to Bethe [5], an outflow of central-peripheral order–to–order transitions. With regard to parallels in frequency and dynamics observed in man and dog, Perlitz and coworkers suggested that also in humans the 0.15 Hz band most likely originates from reticular neurons of the lower brainstem network [46,47,48].

In summary, the findings presented in Fig. 2 underpin the theory of synergetics, since there is reason to regard the ca. 0.15 Hz frequency as an order parameter and the level of mental drive as control parameter. The ca. 0.15 Hz frequency is a prominent example of biological pattern formation lacking external or macroscopic control.

Nonlinear Dynamics in the Communication of Patient and Therapist

As mentioned above, psychotherapy is usually conceptualized as the application of psychological treatments to pa-

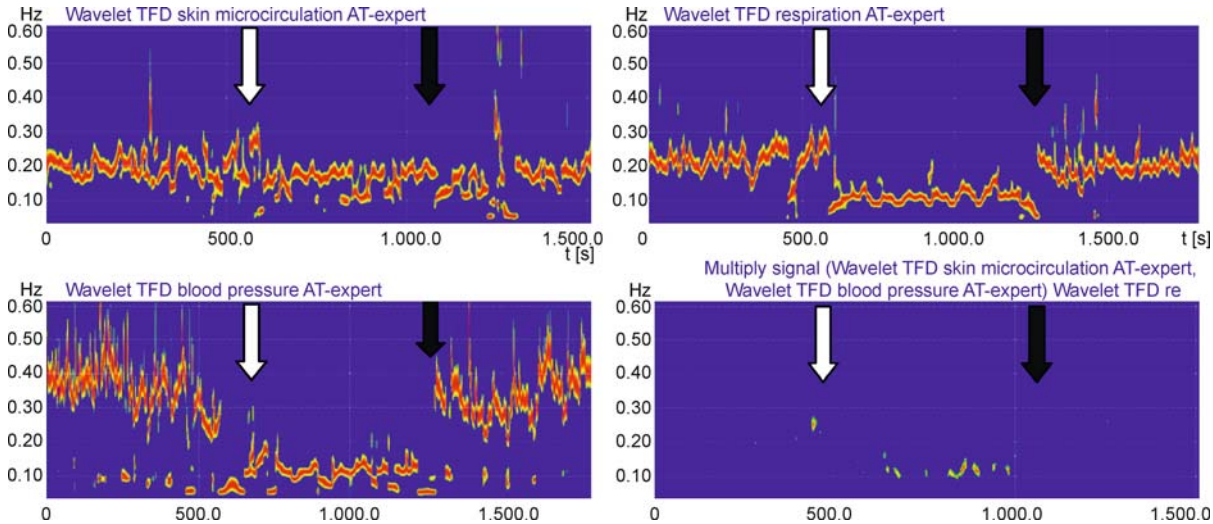

Self-Organization in Clinical Psychology, Figure 2

Wavelet time frequency distributions (TFD) of peripheral noninvasively obtained recordings of a female expert in autogenic training (AT, 56 yrs., healthy, non-smoker, 15 yrs practice AT). *Top left*: TFD of glabella skin microcirculation photoplethysmography; *top right*: TFD of chest respiration related movements; *bottom left*: TFD of peripheral systemic arterial blood pressure; *bottom right*: So-called "joined TFD" of PPG-, respiration- and blood pressure-TFD, a novel method by Besting and colleagues (2005) (multiplying TFDs yielding only frequencies prominent at identical times and identical frequencies, www.Simplana.de) used to compute the intersection of TFD time series. *White arrows* mark the start of AT, *black arrows* mark the end of AT. In the TFD of PPG, the main frequency is at ca. 0.21 Hz prior to the onset of AT and is clearly stabilized at ca. 0.18 Hz with the start of AT, with signs of dissociation when terminating AT. The TFD of respiration supplies ample evidence of an order–order transition triggered by the practice of AT: The main frequency plummets from ca. 0.25 to 0.15 and 0.07 Hz to be maintained at ca. 0.12 Hz. With termination of AT, the main frequency skips back to frequencies shown beforehand. The TFD of systemic arterial blood pressure exhibits an intersection of approx. 90% during the AT section, but also few minor intersections before and after AT (data not shown); the joined TFD intersection shows merely few frequency "spots" at ca. 0.12 Hz during AT

tients in order to change their problem states and diseases. However, as different research programs revealed during the last decades, psychological change processes show all important features of nonlinear systems – like deterministic chaos, nonstationary phase transitions, and nonlinear coupling between patient and therapist. Physiological synchronization appears to be realized at an interpersonal level (between therapist and patient) as well as between different phenomenological levels of the interpersonal system (speech qualities and psycho-physiological variables). In a study of Villmann et al. [83] heart rate, respiratory frequency, muscular tension, and skin conductivity were measured from both, therapist and patient, during 37 therapy sessions. Speech production was analyzed by the Mergenthaler model focusing on emotional feeling and cognitive referential activity/abstraction [38]. Physiological data were analyzed by an artificial neural network approach (growing self-organizing map), which uses a kernel smoothing for improved data density estimation. It was possible to generate an entropy model of psycho-physiological variability detecting emotionally instable phases during the therapy process. The entropy reflecting psycho-physiological and emotional variability was related to the dramatic value of speech analysis according to the cycle model of Mergenthaler.

Empirical evidence exists also for synchronized chaoto-chaotic phase transitions in the brains of therapist and patient during a therapeutic interview, measured by local largest Lyapunov exponents in the EEGs of both interaction partners [52].

Taking into account the importance of the therapeutic relationship for the treatment outcome the attention of a study realized by Schiepek and co-workers focused on the interactional process between therapist and patient [26,58]. The authors used the method of sequential plan analysis, which is a development of the hierarchical plan analysis proposed by Grawe and Caspar (e. g. [10]). Plans in this sense are verbally or non-verbally communicated intentions of self-presentation in a social situation. Patient's and therapist's interactional behavior was analyzed on the basis of video recordings. Two complete therapies (13 and 9 therapy sessions, resp.) were encoded with a sampling rate of 10 s (Fig. 3). The construction of an inclusive hierarchical plan-analysis leads to an

Self-Organization in Clinical Psychology, Table 1

Second-order plans and categories of self-presentations as identified by the hierarchical plan analysis of a complete 13-session psychotherapy. Encoding of therapist and patient. Plans and categories are used as ideographic observation categories for the sequential plan analysis

	Second-order plans	Categories of self-presentation
Therapist	1 show competence 2 encourage a trusting relationship 3 show understanding 4 motivate her	I encourage trust/create a secure atmosphere
	5 encourage her to reflect on her patterns of thinking 6 confront her with her avoidance and problem behavior	II confrontation/exposing to insecurity
	7 activate her 8 show her that she is responsible	III encourage self-responsibility of the patient
	9 guide her focus of attention 10 give her structure	IV activate structuring work
Patient	1 demonstrate strength and competence 2 make it clear that things are or have been difficult 3 be a good patient/create a good relationship to the therapist	I search for sympathy/appreciation/good relationship
	4 show that your suffering is strongly influenced by external causes 5 ask for help from the therapist	II externalization/demonstration of helplessness
	6 show interest and willingness in solving your problems 7 protect yourself from threatening changes	III problem-oriented work (self-relatedness vs. avoidance)

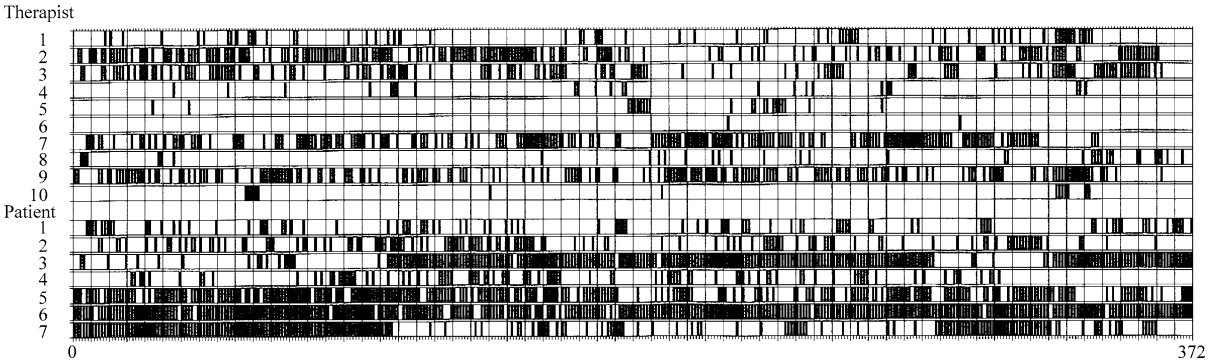

Self-Organization in Clinical Psychology, Figure 3

Nominal sequences of interactional plans of the therapist (*top*) and the patient (*bottom*) during a psychotherapy session. The sampling rate is 10 s. Different plans can be realized simultaneously. The pattern looks like a music score with the plans representing the different instruments of an orchestra. A sonification of the score of plans coded from a 13-session psychotherapy is recorded on a DVD added to the textbook of Haken and Schiepek [21]

ideographic categorical system for the observation of the client–therapist interaction (Table 1).

The first hints of order in the dynamics came from the distribution of simultaneous configurations (on-off-patterns) of plans in the scores. This distribution follows a power law ($1/f^a$) demonstrating a distinct structure order within the data (Fig. 4). Following Bak et al. [2], power law-distributions as demonstrated in Fig. 4 emerge from self-organized criticality within dynamic systems.

Further data analysis was based on the time series of the highest-level categories, the so-called categories of self-presentation (see Table 1). Since in the hierarchical system of the plan analysis the operators at the lowest observation level were quantified by intensity ratings, the plans and the self-presentation categories at the top level integrating the lower level categories were also quantified. The time series were analyzed by methods which are sensitive to the nonlinearity as well as the nonstationarity of the time series [21,26,58,70]. Nonlinearity was proofed by surrogate data tests [51] using random surrogates and FFT-based phase-randomized surrogates [69]. Whereas fractal dimensionalities of the empirical time series (based on

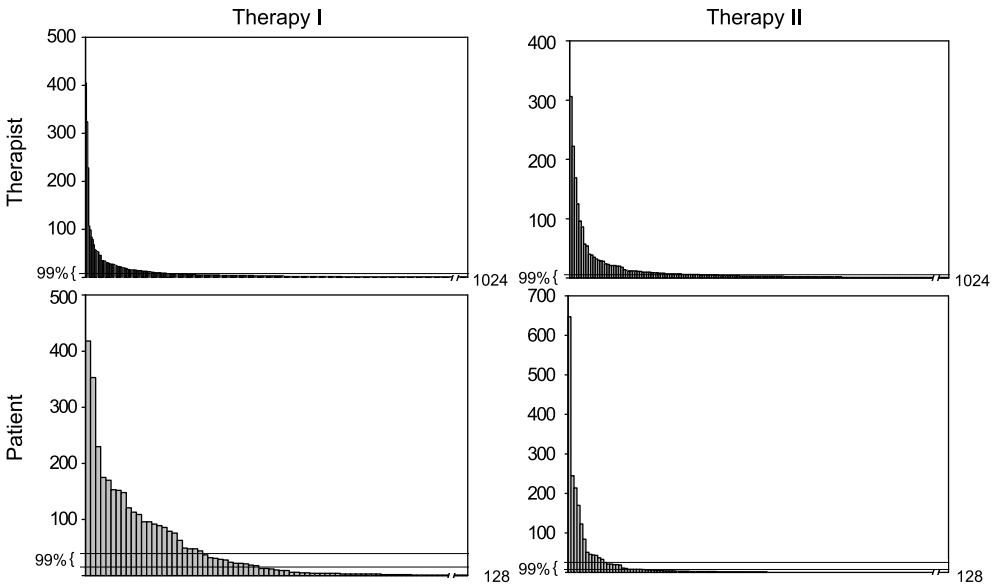

Self-Organization in Clinical Psychology, Figure 4

Empirical frequencies of constellations of interactional plans realized by therapists (10 plans) and patients (7 plans) within two psychotherapies (therapy I: 13 sessions, therapy II: 9 sessions). *X*-axis: Number of all possible configurations of plans (therapist: $2^{10} = 1024$, patient: $2^7 = 128$) ordered by the frequency of their realization. *Y*-axis: Frequencies of plan configurations. The distributions follow a power-law ($1/f^a$) distribution

the correlation dimension D2 as well as mean Pointwise D2 [66]) saturated at finite values (convergence to a fractal dimensionality of about 6), random and FFT-surrogates did not. The methods of PD2 [66] and of the local largest Lyapunov exponents (algorithm from Rosenstein et al. [53]) were used to identify phase-transition like discontinuities. Following the evolution of PD2 dimensionalities, both therapies realized nonstationarities, and both therapies showed periods of strongly synchronized (with correlations from 0.80 to 1.00) and anti-synchronized PD2-processes (with correlations from −0.80 to −1.00) between patient and therapist. Quite similar and even more pronounced dynamical jumps were to be seen in the development of the local largest Lyapunov exponents (Fig. 6), representing changes in the chaoticity of a time signal [26]. An important part of the discontinuities of the LLLE were exactly synchronized between patient and therapist. Obviously both persons create a dynamic self-organizing communication system, which allows for the individual change processes of the patient.

These results get support from nonlinear coupling measures between the time series of the interaction partners. Pointwise transinformation as well as pointwise coupling conditional divergence [33,80] were applied to the data, and both indicate changing and time-dependent coupling strengths between the time series of the interaction partners. There is no priority to the therapist's influence on the patient, which contradicts the classical idea that input from the therapist should determine the client's output. The other way round is also true and both constitute the circular causality of psychotherapeutic self-organization.

In other studies, sequential plan analysis was applied to the microdynamics of group interaction [21]. In a group of five persons a creativity and problem solving task was to be solved within 2,5 h (creation of ideas, rules, and physical handicraft realization of a prototype board game from different materials). Similar to the psychotherapy study the sampling rate was 10 s. The superordinate plans which could be identified for all five persons were (1) spontaneity and emotional engagement vs shyness, restricted behavior, and orientation to social norms, (2) engagement in the group interaction and in positive social climate, (3) task orientation. Length of time series was about 810 coding points (= intervals). D2 as well as mean PD2 estimates saturated at a fractal dimensionality of about 5 for all categories. The embedding of the time series was realized by two ways: (1) The phase space was constituted by the three dimensions of superordinated plans with five trajectories representing the five group members, or (2) the phase space was constituted by the five persons with three trajectories representing the time course of the three plans

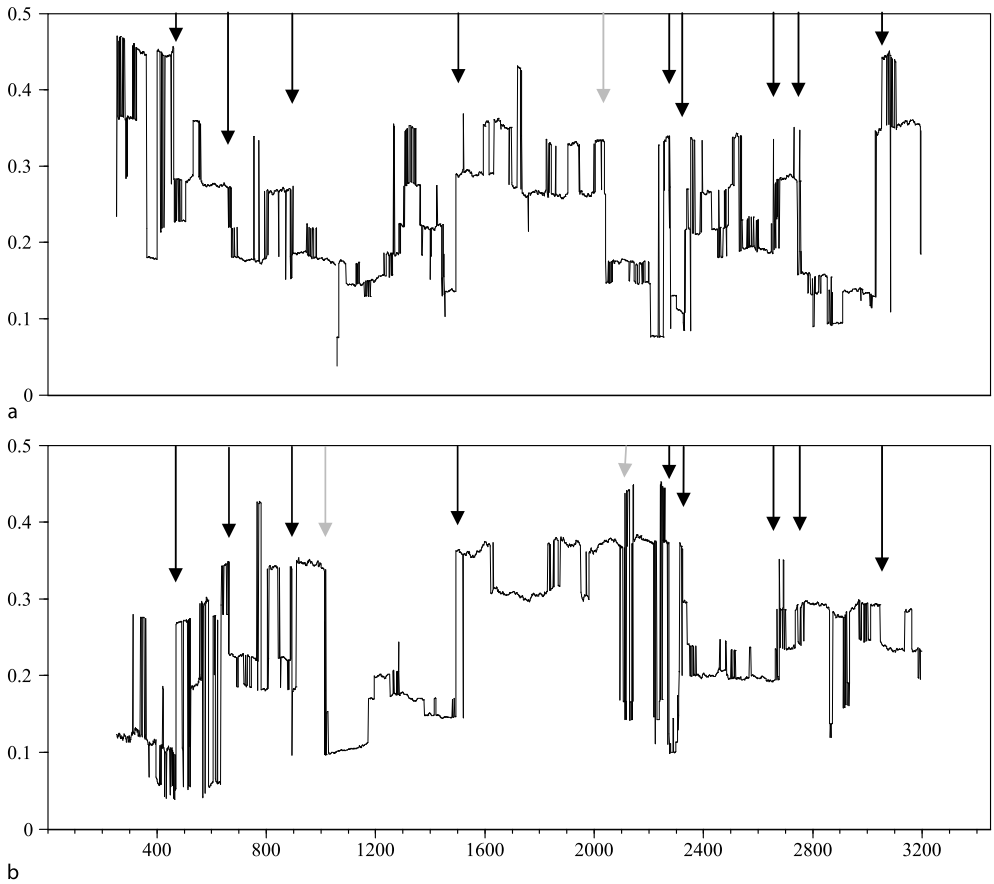

Self-Organization in Clinical Psychology, Figure 5

Synchronized jumps in the dynamics of local largest Lyapunov exponents (*black arrows*). *Grey arrows* indicate not clearly synchronized changes. a Therapist, b Patient

Self-Organization in Clinical Psychology, Table 2

Factors (principal component analysis) of the Therapy Process Questionnaire (TPQ). Factor analysis was based on 94 therapy processes (mean stay = 66 days, daily ratings). Seven first-order factors (*right*) are related to three second-order factors (*left*). Numbers behind the first-order factors indicate factor loadings on second-order factors (for details see [21])

I(2) Change involvement	I Therapeutic progress/confidence in treatment effects/self-efficacy (.571) VI Intensity of therapeutic work/motivation to change (.596) V Opening of perspectives/personal innovations (.649)
II(2) Relationship/Social climate	III Quality of the therapeutic relationship/openness/confidence in the therapist (.705) II Ward atmosphere, social relationship to other inpatients (.692)
III(2) Emotionality	IV Dysphoric emotions/self-relatedness (.732) VII Impairment by symptoms and problems

(additional embedding dimensions result from time delay coordinates). In both cases PD2 results show an evolving pattern of quasi-attractors with changing complexity, and LLLEs (algorithm from Rosenstein et al. [53]) portray chaoto-chaotic phase transitions with clear-cut and interpersonally synchronized jumps – similar to the dyadic interaction of the psychotherapy study.

Self-Organization in Human Change Processes

A quite different approach to human change processes focuses on inpatient treatments at a hospital of psychosomatics. In a study by Schiepek and coworkers (results in [21]) 94 change processes were investigated, realized by 91 inpatients with different diagnoses (depression, anxiety

disorders, posttraumatic stress disorders, eating disorders, somatoform disorders, and others). The time series data was produced by patients' self-ratings which were completed once a day in the evening. For this purpose a 53-item rating sheet was developed (*Therapy Process Questionnaire [TPQ]*, [21]) whose factor analysis resulted in a solution of seven factors defining the subscales of the questionnaire (Table 2). The ratings combined seven-step Likert scales and visual analogue scales especially for ratings of emotions. TPQ measurements reflect important aspects of the patient's experience of progress and goal attainment, emotional involvement, self-efficacy, therapeutic relationship, social relations with other inpatients, and the ward atmosphere.

The inclusive outcome criterion integrated the following measures: Inventory of Interpersonal Problems (IIP), Gießener Beschwerdebogen (GBB), Hospital Anxiety and Depression Scale (HADS), Questionnaire for Social Support (F-SOZU), a life-quality questionnaire (Münchener Lebensqualitäts-Dimensionenliste), a self-efficacy questionnaire (Fragebogen zur Generalisierten Kompetenzerwartung), the Sense-of-Coherence Questionnaire, and an interview-based assessment of personal resources. Additionally, therapists and patients scored the overall treatment effectiveness and treatment quality.

Results confirmed synergetic conceptualizations of how psychotherapy works and corroborated hypotheses drawn from this model. Here therapy is supposed to provide support for the patient's own self-organization processes, which should be characterized by cascades of order-to-order transitions accompanied by critical instabilities of the process. Pathological and restrictive order should be transformed into more flexible and adaptive patterns of behavior, and the synchronization of the different aspects of the patient's experience should undergo some transformations. Exactly this could be observed.

Significant correlations exist between the local maxima of critical fluctuations and the outcome of psychotherapy. The local maxima were defined by the difference between the mean dynamic complexity of the whole psychotherapy process and the maximum of the complexity which was observed during the process. Correlations were -0.455 (second-order factor I: "Change involvement" of the TPQ, $p = 0.002$), -0.431 (second-order factor 2: "Relationship/social climate", $p = 0.003$), and -0.572 (second-order factor 3: "Emotionality", $p = 0.000$) (compare Table 2). Negative correlations result from the fact that increased local maxima of dynamic complexity correspond to reduced problems, symptoms, and impairment.

The *dynamic complexity* combines a fluctuation index with a distribution index. The fluctuation index measures the frequency and amplitude of the change rates of a time series between the reversals of the development within a scanning window gliding over the whole time series. For analysis purpose a window width of seven measurement points (= days) was introduced. The distribution index measures the scattering of realized values within a given scanning window. The more scores are restricted to only narrow intervals of the available scale range, the smaller the distribution index becomes. The score of this index increases as the interval filled by the realized values grows. The algorithm solves the problem of value distribution independently of the scale resolution, the width of the scanning window, and of any combination of these parameters.

In order to answer the question if the observed intensities of dynamic complexity reach critical values, intra-item calibration procedures were used in order to define adequate thresholds fitting to the actual dynamics. The time series of dynamic complexity were standardized by z-transformations, providing significance thresholds of 5% or 1%. Applying this threshold method to all items of the TPQ reduces the quantitative complexity signals of each time series to a three-step signal (not significant, complexity exceeds a 5% threshold, complexity exceeds a 1% threshold). A synopsis of these qualitative signals referring to all items of the TPQ gives an impression of the localization of critical fluctuations during the whole process. Dynamic complexities seem to be synchronized over many items and factors of the TPQ, resulting in the structure of columns of grey ($<5\%$) or black ($<1\%$) dots. In a large part of the investigated therapies such column-like structures could be identified. In an item-by-time synopsis they indicate phases of intensified as well as synchronized fluctuations and entropies of quite different aspects of the process. Consequently, these item-by-time synopses are called *complexity resonance diagrams* (Fig. 6).

In order to confirm the structures found within the complexity-resonance-diagrams, surrogate tests were realized based on random as well as on FFT-based surrogates of the time series. The empirical patterns are impressively different from the surrogate-based patterns (all realized comparisons with $p = 0.000$). Further support for phase-transition like phenomena in the change processes came from recurrence plots representing similarities and dissimilarities of dynamic segments of a whole time series [13,80,85]. This method is based on the embedding of time series into a phase space constructed by time-delay coordinates, a method which is also crucial in the algorithms for the estimation of dimensional complexity or chaoticity (e. g., Kolmogorov–Sinai-Entropy, Lyapunov Exponents). Neighbors in the time-delay phase space rep-

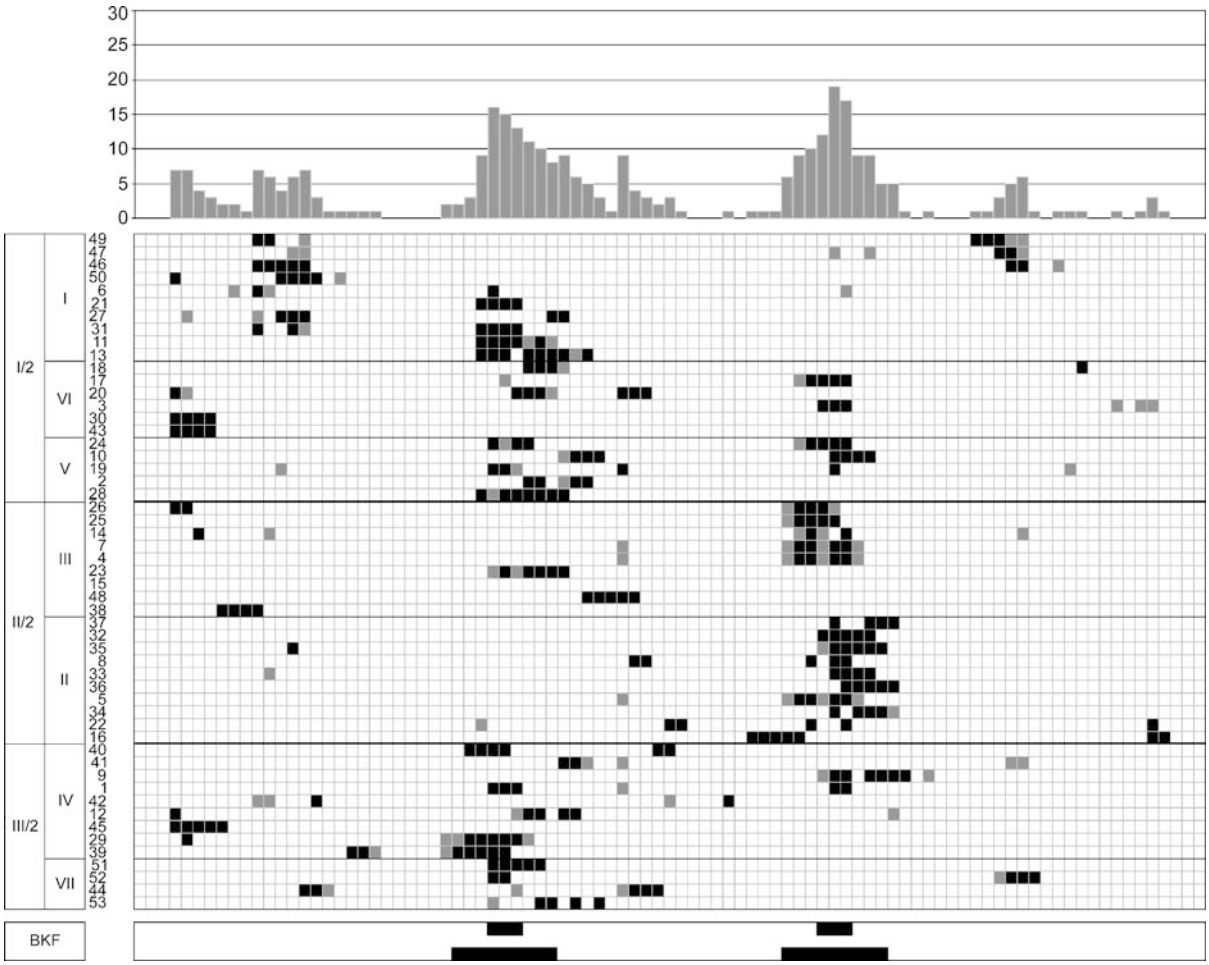

Self-Organization in Clinical Psychology, Figure 6

Complexity resonance diagram of a psychotherapy process. Such diagrams portray the threshold exceeding dynamic complexities of a process encoded by the 53 items of the Therapy Process Questionnaire (TPQ). *Gray dots*: 5% threshold of significance; *black dots*: 1% threshold of significance. *X*-axis: Days of hospital stay, *Y*-axis: Items of the TPQ arranged by the order of the factors as reported in Table 2. Window width for the calculation of dynamic complexities is 7. Column-like structures indicate phases of critical instabilities during the process

resent similar dynamic segments and are plotted by a dot in the recurrence plot. Dissimilarities are represented by empty columns in the recurrence plots, which in many cases exactly correspond to the columns of dots in the complexity-resonance diagrams. The overall correlation is −0.45, if small shifts (lags of + or −3 measurement points at maximum) will be allowed. This means that periods of critical instability correspond to transient dynamics outside of the quasi-attractors established by the self-organizing system under consideration. These different ways to identify critical phase transitions are further validated by the time frequency distribution (TFD) of the time series. The TFD method uses wavelet spectra in order to scan the evolution of the frequency distributions within a signal [33,80]. It is a dynamic counterpart to the static fast Fourier transformation and allows for the identification of pronounced frequency amplitudes or changes in the frequency distributions. In the data set of the referred study these often appear exactly during the phase transitions which can be identified by other methods (see the synoptical representations of different time series analysis methods on the DVD in the textbook of Haken and Schiepek [21]).

An overall result of the study is shown in Fig. 7. It portrays the evidence that in order to bring forth change processes within self-organizing systems at least two conditions should be realized. The first condition: The degree of the control parameter energizing the system and push-

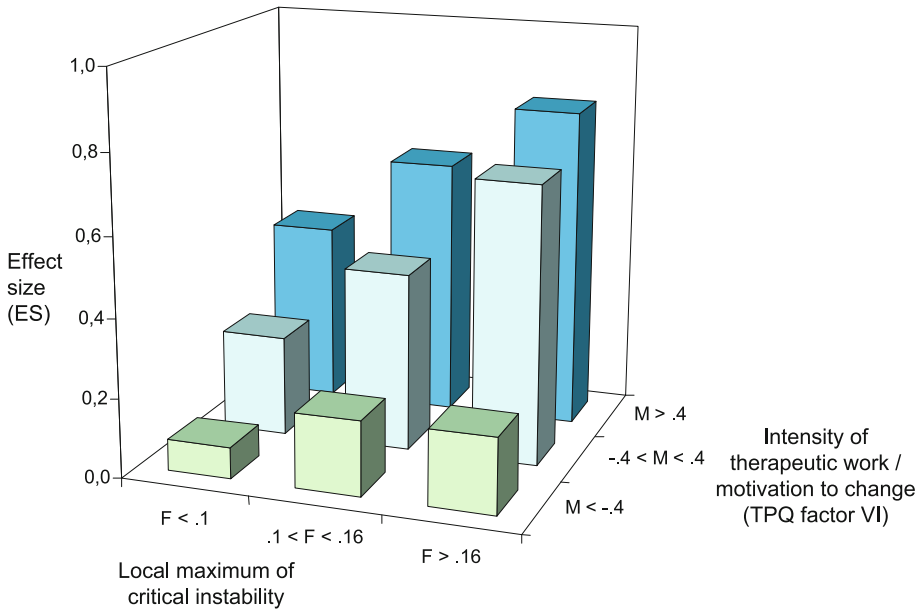

Self-Organization in Clinical Psychology, Figure 7

The effect size (ES) (mean ES of all outcome measures introduced in the study, see text) of inpatient psychotherapy is produced by an interaction between the local maximum of critical fluctuations and the intensity of the control parameter realized during the change process. The local maxima of fluctuations were defined by the difference between the mean dynamic complexity of the whole therapy process and the maximum of the complexity observed during the process. The diagram is based on the mean of the local maxima of all items. The control parameter was defined by the overall mean of the TPQ factor VI: Intensity of therapeutic work/motivation to change

ing it away from its actual equilibrium state should exceed a certain intensity level. With respect to psychotherapies this control parameter could be the patient's motivation to change including his engagement into the therapeutic work. Second condition: The degree of instability the system attains during its change process. This instability during emerging symmetries and symmetry breaking transitions is given by the local maximum of dynamic complexity during the hospital stay. The interaction of both conditions results in treatment effectiveness. A third important condition is not represented in Fig. 7: It is the experienced stability of the outer environment (context at the ward or therapeutic bond) or of the inner environment (as self-esteem, self-confidence, or activated resources). This context of stability is a prerequisite for a system to undergo critical instabilities.

The Concept of Self-Organization Promotes New Information Technologies in Clinical Psychology – The Synergetic Navigation System

Since self-organization and nonlinear dynamics seem to be ubiquitous in human change processes, it should be helpful to go beyond the diagnostics of steady states to an assessment of dynamics. Practitioners should get information on the therapy and its features *during* the ongoing process in order to use this information for an adequate placement of interventions and a control of the dynamics. "Controlling" self-organization processes in psychotherapies means the generation and co-creation (together with the patient) of adequate boundary conditions, the decision to do or to retain certain interventions, and to support the dynamics which the system is creating by itself. The patient takes an active and cooperative role in this understanding of data-based and co-creative change processes. Another important motivation for the development of real-time assessment comes from the evidence that most of the empirically identified specific and non-specific factors driving therapeutic change processes are connected with specific persons (the concrete therapist who meets a concrete patient in a concrete setting) and evolve by its nonlinear interactions in specific systems. These factors are (i) personal features of the patient like his motivation to change, his premorbid adaptation and degree of social functioning, personality integration, ego-strength, or comorbidities, (ii) personal and professional features of the therapist like his own personality integration, social and professional competencies, allegiance to his approach of

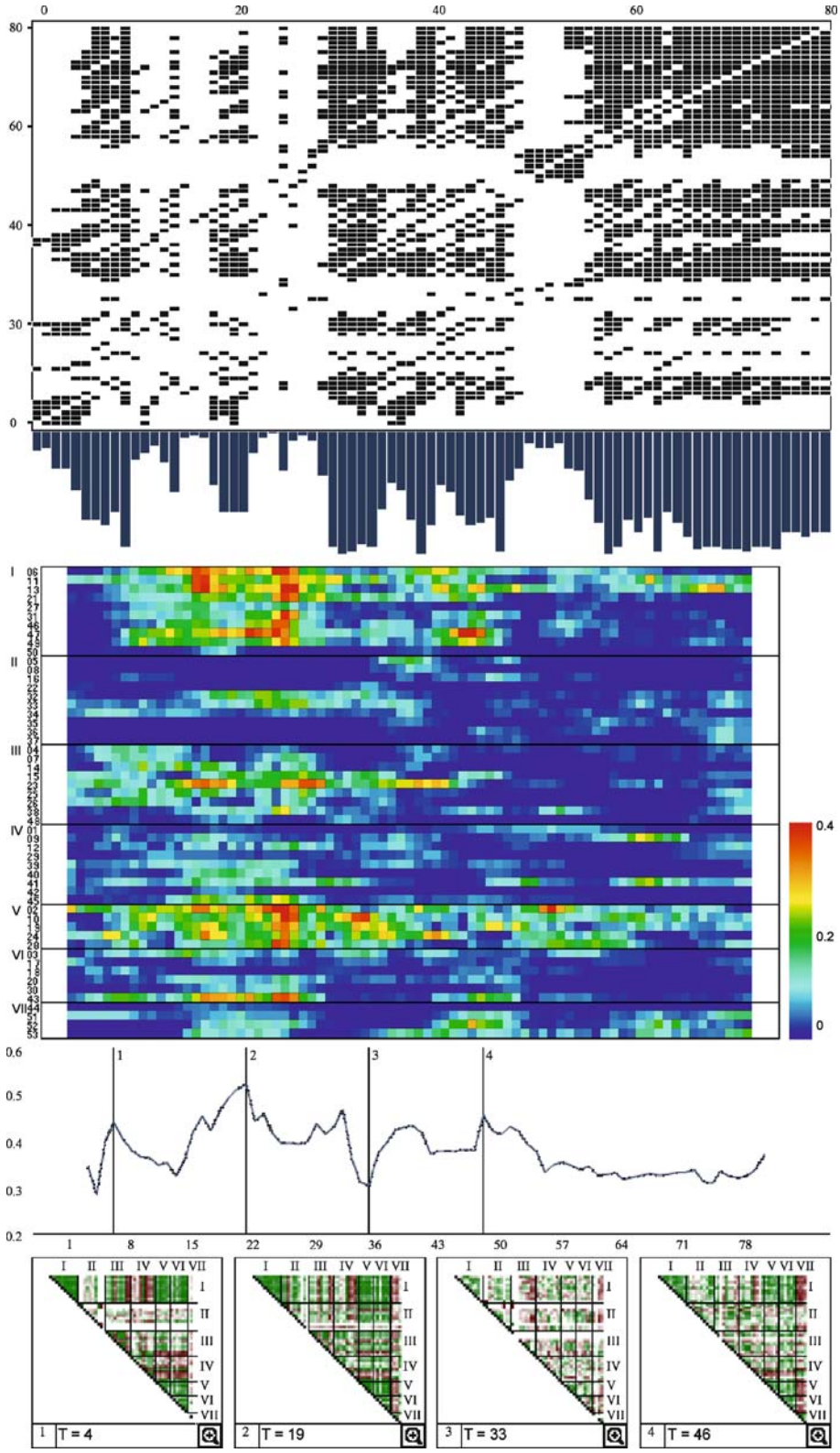
0
20
40
60
80
80
60
40
30
0
I
II
III
IV
V
VI
VII
0.4
0.2
0
0.6
0.5
0.4
0.3
0.2
1
2
3
4
1
8
15
22
29
36
43
50
57
64
71
78
I II III IV V VI VII
1 T = 4
2 T = 19
3 T = 33
4 T = 46

◀ **Self-Organization in Clinical Psychology, Figure 8**
Synopsis of a psychotherapy process as monitored by the Synergetic Navigation System. The time course of the inpatient treatment of a patient with eating disorders portrays a clear cut phase-transition associated with critical instabilities. *Top*: Recurrence plot of the item "Today I was successful to do steps towards my personal goals". *Dots* represent recurrent segments of the time series, empty spaces represent transitions. *Middle*: Complexity resonance diagram of all items of the TPQ. Different from Fig. 6, the intensities of the dynamic complexity of each item is transformed into colors. Items are arranged by the order of the first- and second-order factors of TPQ. *Bottom*: Mean of all inter-item correlations irrespective of the sign (absolute values). This is a measure of the overall synchronisation of the patient's experiences as represented by the items of the TPQ. The correlation structure is shown at four measurement points (days) of the psychotherapy process ($t = 4$, $t = 19$, $t = 33$, $t = 46$). Intensity of green represents positive correlations, intensities of red represents negative correlations

doing therapy, stress-resistance, and so on, and (iii) factors of the professional and social context (see the so-called generic model of psychotherapy [29,42]). In consequence, *evidence-based treatments* should be based on the evidence of concrete data mirroring the ongoing change process and on the professional decisions reflecting this insight.

Real-time monitoring actually uses internet-based presentations (including PDA or cell phone technology) of outcome and process questionnaires. Data are sent to a server, where they are stored and analyzed. Professionals and patients can inspect the results whenever they want. Experiences with real-time feedback to therapists (based on an outcome questionnaire the patient fills out during the therapy sessions in an ambulatory or outpatient context) are encouraging. Lambert and co-workers (e. g., [31]) were able to identify processes on the way of getting difficult or unsuccessful ("not on track" therapies, compared to more promising "on track" therapies), and helped therapists to correct these not-on-track dynamics by specific interventions. By this, threatening drop-outs could be avoided, bad results could be corrected, and on-track processes could be optimized and even shortened.

More sophisticated than the distinction between "on-track" and "not-on-track" courses is the feedback on self-organization features realized by a system based on synergetics [21]. The *Synergetic Navigation System* uses the therapy process questionnaire for daily ratings and applies methods from nonlinear time series analysis in order to identify important qualities of the change process. This are:

- Stability or instability of the dynamics as represented by the subscales (factors) of the TPQ (see Table 2), which is measured by the dynamic complexity
- Recurrence plots indicating transitions or repeating patterns
- Intensity of synchronization and time-dependent synchronization patterns between the items and the factors of the TPQ (realized by the cross-correlations of all items of the TPQ, calculated within a running window).

Figure 8 shows a synopsis of these analysis methods applied to a specific change process. Preceding the inspection of all analysis results the raw data series of the items and the time courses of the factors (z-transformed values) are available. Additionally patients can write an electronic diary after filling out the questionnaire. The diary entries can be presented within a gliding tip-tool running over the time series. By this, corresponding qualitative and quantitative information completes the picture.

The Self-Organizing Brain

The human brain is one of the most outstanding examples of a complex nonlinear system producing self-organized patterns of functioning. Since function corresponds to structure and vice versa, structural changes (changes of intersynaptic coupling strengths and network configurations, (re-)wiring patterns following the synchronized co-activity of neurons) can be explained by functional self-organization of neural populations. Perception, action and transition of action patterns, decision making, and cognitive, behavioral, as well as emotional learning are psychological functions following the principles of self-organization [21]. At a neural level they correspond to and are based on nonlinear brain dynamics. The emergence of order parameters and the occurrence of phase transitions can be described and measured on a psychological as well as on a neural level.

One of the phenomena modeled by synergetics is Gestalt perception – the construction of percepts and the switching of ambiguous visual patterns (e. g., Necker cube or stroboscopic alternative motion). These processes of Gestalt perception constitute the link between Gestalt psychology and actual mathematical modeling in synergetics [17]. The binding of different perceptual features or components to coherent structures or "qualia" seems to be due to synchronization processes of extended brain regions and converging integrative areas [64]. Pattern perception corresponds to pattern formation – as H. Haken puts it into pointed words. Tallon-Baudry et al. [71,72] measured enhanced gamma-band activity (30–50 Hz) in the EEG of the primary and secondary visual cortex while subjects identified a triangle within the offered stimulus material. This could be a fingerprint of correspond-

ing neural synchronization processes. This activity occurred when subjects saw a real object (triangle) as well as a figural illusion of the object (Kanizsa triangle), but not if geometrical components could not be composed to a true Gestalt. The research group of Basar–Eroglu and Stadler [4] measured significant gamma-band activity in EEG during states of perceptual switching triggered by stroboscopic alternative motions. In summary: Perception of multistability is one of the multifold cognitive processes giving rise to 40 Hz enhancement in the cortex, and coherent oscillations reflect an important mechanism of feature linking in the visual cortex which corresponds to the emergence of a neural order parameter. Changing order parameter dynamics during different cognitive activities was shown by Schupp et al. [62]. Mental imagery of an object could be differentiated from its concrete perception. The dimensional complexity of prefrontal EEG was increased during sensory imagery compared to the real perception of the same object (compare [36]).

The well-known movement coordination paradigm modeled by Haken et al. [22] was used to demonstrate neural correlates of instability and symmetry breaking processes in the motor brain. The order parameter in this finger movement experiment is the relative phase of the index fingers of both hands. Metronome-pacing – with movement frequency as the control parameter – triggers the system from parallel (out-of-phase) to mirror (in-phase) movement. Meyer-Lindenberg et al. [39] showed that the emergence of patterns in open, nonequilibrium systems like the brain is governed by their stability in response to small disturbances. Transitions could be elicited by interference at the neural level. Functional neuroimaging (PET) identified premotor (PMA) and supplementary motor (SMA) cortices as having neural activity linked to the degree of behavioral instability, induced by increasing frequency of the finger movement. These regions then were transiently disturbed with graded transcranial magnetic stimulation (TMS), which caused sustained and macroscopic behavioral transitions from the less stable out-of-phase to the stable in-phase movement, whereas the stable pattern could not be affected. Moreover, the strength of the disturbance needed (a measure of neural stability) was linked to the degree of the control parameter (movement frequency) and thereby to the behavioral stability of the system.

Synergetic research in clinical psychology is now reaching the brain level. The aim of an actual fMRI-study [60] is the investigation of phase transitions of brain activity and related subjective experiences of patients during their psychotherapy process. Repeated fMRI scans are related to the degree of stability or instability of the ongoing dynamics (measured by the dynamic complexity of daily TPQ-ratings) as well as to the therapy outcome. Real-time monitoring by the *Synergetic Navigation System* allows for the identification of stable or unstable periods and by this for a decision on the appropriate moments of fMRI acquisitions. Three or four scans are realized during each of the psychotherapy processes of 15 patients. The study includes only patients with obsessive-compulsive disorder (OCD) of the washing/contamination fear subtype (DSM IV: 300.3), without any medication or comorbid psychiatric or somatic diagnoses. Patients are matched to healthy controls. (This research is a multi-center study of the Ludwig-Maximilians-University Munich, Institute of Psychology (Prof. Dr. Günter Schiepek, head of the project), and Clinic of Psychiatry (PD Dr. Oliver Pogarell, Dipl. Psych. Susanne Karch, Dr. Christoph Mulert), Hospital of Psychosomatic Medicine Windach/Ammersee and Day Treatment Centre Munich/Westend (Dr. Igor Tominschek, cand. Psych. Stephan Heinzel, Prof. Dr. Michael Zaudig), University Hospital Vienna/Astria, Clinic of Psychiatry (Prof. Dr. Martin Aigner, Prof. Dr. Gerhard Lenz, cand. med. Markus Dold, Dr. Annemarie Unger), MR Centre of Excellence, Medical University Vienna/Austria (Prof. Dr. Ewald Moser, Dr. Christian Windischberger).

OCD seems to be an appropriate model system for synergetic studies in clinical psychology, since the pathological order parameter is phenomenologically quite evident, the disease has an obvious and quite stable time course, and therapeutic phase transitions – if they do occur at all – are easy to be observed. OCD-specific functional neuroanatomy is partially known: Friedlander and Desrocher [14] report on an executive dysfunction model corresponding to the cortico-striato-thalamo-cortical feedback-loops involved in perseverations and compulsions, and on a modulatory control model involved in the pathological mechanisms of anxiety and distress provoking obsessions.

The visual stimulation paradigm of the study uses symptom provoking, disgust provoking, and neutral pictures. The disgust and the neutral pictures are taken from the International Affective Picture System, whereas the OCD-related pictures are photographed in the home setting of the patients, showing specific and individualized symptom provoking stimuli.

For illustrative purposes we report on the results of a single case. It is a female patient, whose fMRI scans were taken three times during the 59 days of their hospital stay at days 9, 30, and 57. The healthy control was also scanned three times at identical time intervals as the patient. The second acquisition was done after an intensive period of critical instability of the TPQ-based time se-

ries, but just before the flooding was started. (Flooding or response prevention is an essential therapy technique in the treatment of OCD, where patients are confronted with symptom provoking stimuli but abstain from performing compulsive rituals.) The instability of the patient's process was the precursor of an important personal decision to divorce from her husband. (It should be noted that the development of her OCD symptoms was in the context of a long-lasting marital conflict.) This decision was the essential phase transition of the therapy.

Indeed, the most pronounced changes in brain activity occurred from the first to the second fMRI scan, whereas BOLD response differences from the second to the third session were only slight. They perhaps represent the neural correlates of an important personal phase transition related to the resolution of a severe personal conflict. Because these changes occurred before the flooding procedure was started, this can be seen as indicator of an early rapid response in the therapy [29]. Additionally, marked alternations in brain activity were to be observed before or during symptom reduction took place (measured by the Y-BOCS), not afterwards.

Alternations in brain activity involved widespread areas, e. g. the medial frontal brain regions including anterior cingulate cortex, superior and middle frontal gyrus, inferior frontal and precentral gyrus, superior temporal gyrus, superior parietal lobe, cuneus, thalamus and caudate nucleus in both hemispheres, as well as the right fusiform gyrus (see Fig. 9 for a OCD to disgust contrast). Thalamic and basal ganglia activation is part of the dorsolateral-caudate-striatum-thalamus circuitry of OCD. Especially the caudate nucleus takes a role within the executive dysfunction model of compulsions, and its activity has been found to be reduced after treatment [40].

The function of the anterior cingulate cortex is interesting with regard to synergetics. The cingulate cortex comprises various functions like somatosensoric integration, mediation of affective and cognitive processes, control of attention, and processing of painful stimuli. Additionally, it plays an important role as conflict monitoring system: It is sensitive to ambiguous or conflicting information [81,82], is involved in decision making [24,54], and its activation is predictive to treatment outcome in depression (e. g. [37]). This is true especially for the dorsal (cognitive) structures of the ACC. It could be an indicator of symmetry states of brain functioning, which is characterized by two or more dynamic patterns or attractors in competition. In the present case, the ACC activation at the beginning of the therapy could be either part of the pathology or could be indicative for the critical instability of the cognitive-affective system of the patient, preparing her important decision. The second fMRI measure was conducted during a local minimum of critical fluctuations. If the impressive change in cingulate activation could be attributed to a changed critical symmetry state of the neural self-organization before vs. after the phase-transition or to changes in symptom severity cannot be decided within a single case study, but seems to be an interesting question to further research. Perhaps the fact that during the second fMRI measure the Y-BOCS score was nearly on the same level as during the first measure – only 14% reduction, compared to 50% reduction in dynamic complexity – could be a first argument in favor of the instability hypothesis.

The paradigm of self-organization is a very promising approach to clinical as well as other fields of psychology. Its interdisciplinary is due to the fact that the laws and principles of self-organization are true for neural, mental, and behavioral processes (and the corresponding data qualities). Interdisciplinary cooperation is underpinned by the unifying terminology as well as by the unifying formalism and modeling tools of synergetics. This opens new perspectives for basic and applied research, but also for the treatment of mental disorders. New developments in the real-time monitoring of human change processes based on synergetics and nonlinear science have been mentioned. Another field of encouraging developments is deep brain stimulation (DBS), which apply to neurological diseases as Parkinsonian or essential tremor, but also to psychiatric disorders as OCD or mayor depression [75]. The difference between new technologies (applying the mathematical instruments and concepts of synergetics as well as methods from stochastic phase resetting) and classical electrical deep brain stimulation is that normal DBS at high frequencies has a blocking effect on the stimulated target and mimics the effect of tissue lesioning. New technologies are demand-controlled, working with low stimulation frequencies, and avoid the suppression of neurons' firing. Its effect is a desynchronization of pathologically synchronized populations of neurons, using multi-site coordinated reset stimulation [74] or nonlinear delayed feedback stimulation [50]. Both methods counteract abnormal interactions and detune the macroscopic frequency of the collective oscillators – that is the abnormally established order parameters of neural synchronization. Thereby they restore the natural frequencies of the individual oscillatory units. Neurons get in the range of physiological functioning and can engage in changing and varying synchronization patterns. If altered synchronization patterns also change the coupling strength connecting synapses, a rewiring of neural nets could be reached. Changed function triggers the emergence of healthy at-

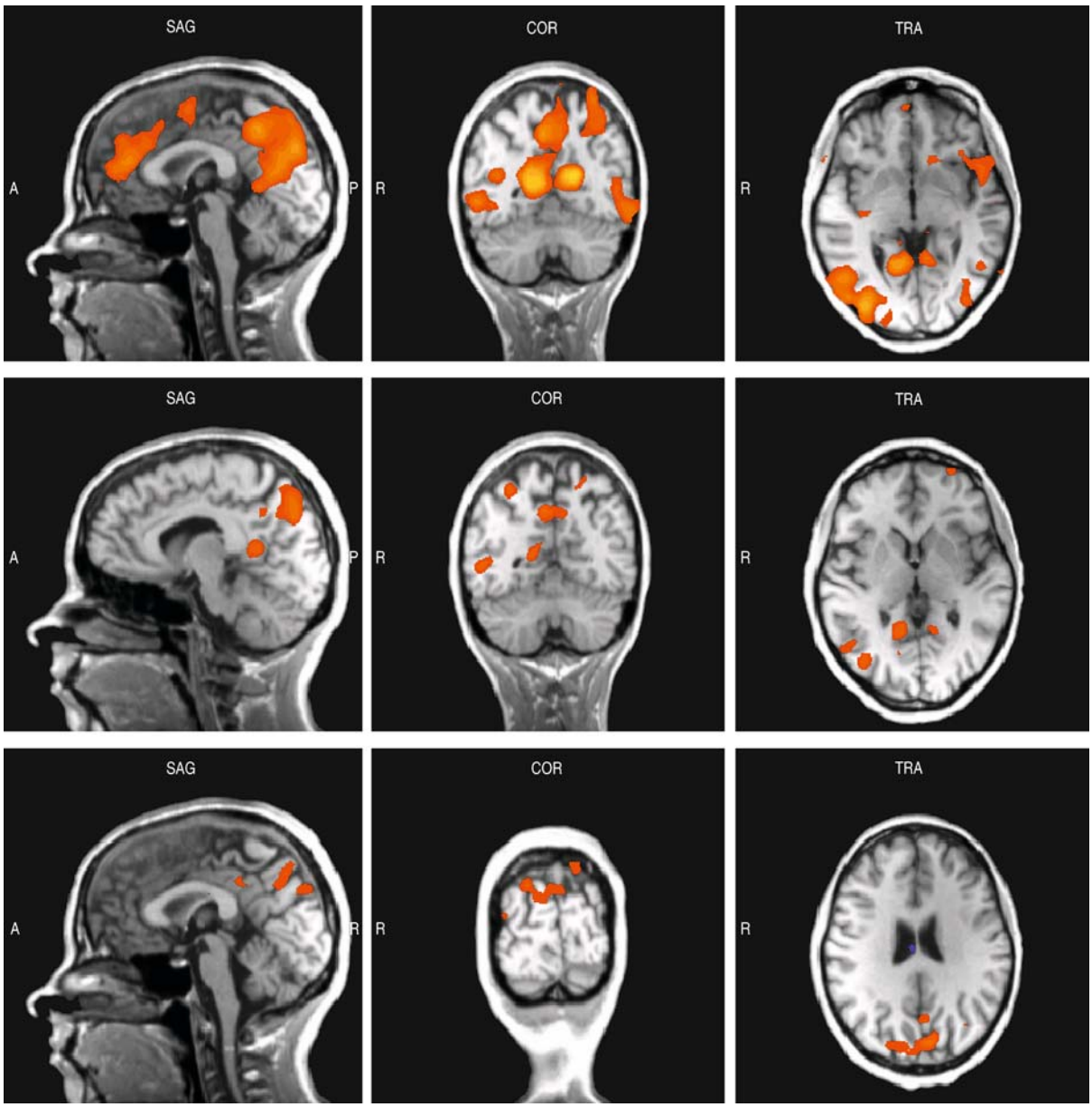

Self-Organization in Clinical Psychology, Figure 9
Brain activation patterns of a patient with OCD during psychotherapy. BOLD signals from a 1.5 Tesla fMRI scanner. *Top*: First scan (9th day of hospital stay; $x = 0$, $y = -55$, $z = -2$; p(uncor) <0.001). *Middle*: Second scan (30th day of hospital stay; $x = 8$, $y = -54$, $z = 5$; p(uncor) < 0.001). *Bottom*: The third scan (57th day of hospital stay; $x = 0$, $y = -85$, $z = 26$; p(uncor) < 0.001). Activations during the presentation of OCD-related pictures compared to activations during the presentation of neutral pictures (OCD > disgust)

tractors and by this changes the structure of neural networks. Perhaps in the future technologies of DBS or even non-invasive brain stimulation could be combined with psychotherapy and psychological navigation instruments developed to optimize self-organizing change processes.

Future Directions

The future developments of self-organization and complexity research in clinical psychology and psychotherapy will be interconnected to its acceptance in practice and

training. Perhaps this sounds paradoxically, since in most other scientific fields the future depends on the investigations to basic research and to new technologies. Of course this holds also for synergetics and its applications to clinical psychology. However, it should be noted that complexity research and nonlinear dynamics are done since more than two decades in European academic psychology with poor impact to mainstream science. So, the future will depend on a greater number of new arriving and highly qualified students in this topic who do not avoid the touch with mathematics. Self-organization and complexity research including its mathematical backgrounds should become part of the training curricula in psychology and psychotherapy. Since the Synergetic Navigation System waits for its broad application in clinical and psychotherapeutic practice, a new decade of practice-based research can be started. But these developments depend on its acceptance by practitioners because of the competencies required for the widespread use of sophisticated methods. This integration of science with practice will open huge sources and new dimensions of data gathering on dynamic systems. An important database for outcome and time series data (including biomarkers) of human change processes is actually prepared.

Another stream of development is concerning the integration of psychological and biological/physiological data. Since human self-organization takes place on synchronized mental, social, and biological system levels, all of them should be taken into account in further research. One research paradigm was suggested in this chapter: The investigation of individual and social processes by the Synergetic Navigation System, and in parallel repeated brain scans using fMRI technology or other methods to get insight into brain dynamics (EEG, gene expression markers [25], immune or endocrine markers [61], or others). Two final remarks: First, future developments of synergetic-based minimal invasive DBS could be combined with psychotherapy and psychological interventions – as pharmacological and psychological treatments are combined nowadays. Second, the nonlinear networks underlying psychological as well as neural self-organization will not be understood without applying appropriate mathematical tools, giving raise to a new systemic psychology and neuroscience.

Bibliography

1. an der Heiden U (1992) Chaos in health and disease – phenomenology and theory. In: Tschacher W, Schiepek G, Brunner EJ (eds) Self-organization and clinical psychology. Springer, Berlin, pp 55–87
2. Bak P, Chen K, Creutz M (1989) Self-organized criticality and the 'Game of Life'. Nature 342:780–782
3. Basar E, Flohr H, Haken H, Mandell AJ (1983) Synergetics of the brain. Springer Series in Synergetics, vol 23. Springer, Berlin
4. Basar-Eroglu C, Strüber D, Kruse P, Basar E, Stadler M (1996) Frontal gamma-band enhancement during multistable visual perception. Int J Psychophysiol 24:113–125
5. Bethe A (1940) Die biologischen Rhythmusvorgänge als selbständige und erzwungene Kippvorgänge betrachtet. Pflügers Arch 244:1–42
6. Böker W, Brenner HD (1996) Stand systemischer Modellvorstellungen zur Schizophrenie und Implikationen für die Therapieforschung. In: Böker W, Brenner HD, Genner RM (eds) Integrative Therapie der Schizophrenie. Huber, Bern, pp 17–32
7. Böker W, Brenner HD (eds) (1989) Schizophrenie als systemische Störung. Huber, Bern
8. Cannon WB (1915) Bodily changes in pain, hunger, fear, and rage: An account of recent researches into the function of emotional excitement. Appleton, New York
9. Carlsson A (2006) The neurochemical circuitry of schizophrenia. Pharmacopsychiatry 39(Suppl 1):S10–S14
10. Caspar F (1996) Beziehungen und Probleme verstehen. Eine Einführung in die psychotherapeutische Plananalyse. Huber, Bern
11. Ciompi L (1989) Zur Dynamik komplexer biologisch-psychosozialer Systeme: Vier fundamentale Mediatoren in der Langzeitentwicklung der Schizophrenie. In: Böker W, Brenner HD (eds) Schizophrenie als systemische Störung. Huber, Bern, pp 27–38
12. Ciompi L, Müller C (1976) Lebensweg und Alter der Schizophrenen. Springer, Berlin
13. Eckmann JP, Oliffson Kamphorst S, Ruelle D (1987) Recurrence plots of dynamical systems. Europhys Lett 4:973–977
14. Friedlander L, Desrocher M (2006) Neuroimaging studies of obsessive-compulsive disorder in adults and children. Clin Psychol Rev 26:32–49
15. Friston KJ, Harrison L, Penny WD (2003) Dynamic causal modelling. NeuroImage 19:1273–1302
16. Haken H (1990) Synergetics – an introduction. Nonequilibrium phase transitions in physics, chemistry, and biology. Springer, Berlin (first edition 1977)
17. Haken H (1990) Synergetics as a tool for the conceptualization and mathematization of cognition and behavior – How far can we go? In: Haken H, Stadler M (eds) Synergetics of cognition. Springer, Berlin, pp 2–31
18. Haken H (1996) Principles of brain functioning. A synergetic approach to brain activity, behavior, and cognition. Springer, Berlin
19. Haken H (2002) Brain dynamics. Springer, Berlin
20. Haken H (2004) Synergetics. Introduction and advanced topics. Springer, Berlin
21. Haken H, Schiepek G (2006) Synergetik in der Psychologie. Selbstorganisation verstehen und gestalten. Hogrefe, Göttingen
22. Haken H, Kelso JAS, Bunz H (1985) A theoretical model of phase transition in human hand movements. Biol Cybern 51:347–356
23. Kettunen J, Keltikangas-Järvinen L (2001) Intraindividual analysis of instantaneous heart rate variability. Psychophysiology 38:659–668

24. King-Casas B, Tomlin D, Anen C, Camerer CF, Quartz SR, Montague PR (2005) Getting to know you: Reputation and trust in a two-person economic exchange. Science 308:78–83
25. Koch JM, Kell S, Hinze-Selch D, Aldenhoff JB (2002) Changes in CREB-phosphorylation during recovery from major depression. J Psychiat Res 36:369–375
26. Kowalik ZJ, Schiepek G, Kumpf K, Roberts LE, Elbert T (1997) Psychotherapy as a chaotic process II: The application of nonlinear analysis methods on quasi time series of the client-therapist-interaction: A nonstationary approach. Psychother Res 7:197–218
27. Kruse P, Carmesin HO, Stadler M (1997) Schizophrenie als Korrespondenzproblem plastischer neuronaler Netze. In: Schiepek G, Tschacher W (eds) Selbstorganisation in Psychologie und Psychiatrie. Vieweg, Braunschweig, pp 171–190
28. Köhler W (1947) Gestalt psychology. Liveright, New York
29. Lambert MJ, Ogles BM (2004) The efficacy and effectiveness of psychotherapy. In: Lambert MJ (ed) Bergin and Garfield's handbook of psychotherapy and behavior change. Wiley, New York, pp 139–193
30. Lambert MJ, Whipple JL, Smart DW, Vermeersch DA, Nielsen SL, Hawkins EJ (2001) The effects of providing therapists with feedback on patient progress during psychotherapy: Are outcomes enhanced? Psychother Res 11:49–68
31. Lambert MJ, Whipple JL, Vermeersch DA, Smart DW, Hawkins EJ, Nielsen SL, Goates M (2002) Enhancing psychotherapy outcomes via providing feedback on client progress: A replication. Clin Psychol Psychother 9:91–103
32. Lambertz M, Vandenhouten R, Grebe R, Langhorst P (2000) Phase transitions in the common brainstem and related systems investigated by nonstationary time series analysis. J Autonom Nerv Syst 78:141–157
33. Lambertz M, Vandenhouten R, Langhorst P (2003) Transiente Kopplungen von Hirnstammneuronen mit Atmung, Herzkreislaufsystem und EEG: Ihre Bedeutung für Ordnungsübergänge in der Psychotherapie. In: Schiepek G (ed) Neurobiologie der Psychotherapie. Schattauer, Stuttgart, pp 302–324
34. Lenz F, Kwan H, Martin R, Tasker R, Dostrovsky J, Lenz Y (1994) Single unit analysis of the human ventral thalamic nuclear group. Tremor-related activity in functionally identified cells. Brain 117:531–543
35. Lewin K (1951) Field theory in social psychology. Harper, New York
36. Lutzenberger W, Elbert T, Birbaumer N, Ray WJ, Schupp H (1992) The scalp distribution of the fractal dimension of the EEG and its variation with mental tasks. Brain Topogr 5:27–34
37. Mayberg HS, Brannan SK, Mahurin RK, Jerabek PA, Brickman JS, Tekell JL, Silva JA, McGinnis S, Glass TG, Martin CC, Fox PT (1997) Cingulate function in depression: A potential predictor of treatment response. Neuroreport 8:1057–1061
38. Mergenthaler E (1998) Cycles of emotion-abstraction patterns: A way of practice oriented process research? Br Psychol Soc – Psychother Sect Newsl 24:16–29
39. Meyer-Lindenberg A, Ziemann U, Hajak G, Cohen L, Faith Berman K (2002) Transitions between dynamical states of differing stability in the human brain. PNAS USA 99:10948–10953
40. Nakao T, Nakagawa A, Yoshiura T, Nakatani E, Nabeyama M, Yoshizato C, Kudoh A, Tada K, Yoshioka K, Kawamoto M, Togao O, Kanba S (2005) Brain activation of patients with obsessive-compulsive disorder during neuropsychological and symptom provocation tasks before and after symptom improvement: A functional magnetic resonance imaging study. Biol Psychiatry 57:901–910
41. Nowak A, Vallacher RR (1998) Dynamical social psychology. Guilford, New York
42. Orlinsky DE, Ronnestad MH, Willutzki U (2004) Fifty years of psychotherapy process-outcome research: Continuity and change. In: Lambert MJ (ed) Bergin and Garfield's handbook of psychotherapy and behavior change. Wiley, New York, pp 307–389
43. Pare D, Curro'Dossi R, Steriade M (1990) Neuronal basis of the parkinsonian resting tremor: A hypothesis and its implications for treatment. Neurosci 35:217–226
44. Penny WD, Stephan KE, Mechelli A, Friston KJ (2004) Modelling functional integration: A comparison of structural equation and dynamic causal models. NeuroImage 23(Suppl 1):264–274
45. Perlitz V, Cotuk B, Haberstock S, Kahn N, Grebe R, Petzold ER, Schmid-Schönbein H (2003) Differentiation of cutaneous haemo- and neurodynamics using multiscale Time-Frequency-Distribution portrays. In: Blazek V, Schultz-Ehrenburg U (eds) Proceedings of the 10th international symposium CNVD 2001 at Aachen. Germany
46. Perlitz V, Cotuk B, Lambertz M, Grebe R, Schiepek G, Petzold ER, Schmid-Schönbein H, Flatten G (2004) Coordination dynamics of circulatory and respiratory rhythms during psychomotor relaxation. Autonom Neurosci 115(1–2):82–93
47. Perlitz V, Cotuk B, Schiepek G, Sen A, Haberstock S, Schmid-Schönbein H, Petzold ER, Flatten G (2004) Synergetik der hypnoiden Relaxation. Psychother Psych Med 54:250–258
48. Perlitz V, Lambertz M, Cotuk B, Grebe R, Vandenhouten R, Flatten G, Petzold ER, Schmid-Schönbein H, Langhorst P (2004) Cardiovascular rhythms in the 0.15 Hz band: Common origin of identical phenomena in man and canine in the reticular formation of the brain stem? Pflügers Arch – Europ J Physiol 448(6):579–592
49. Piaget J (1976) Die Äquilibration der kognitiven Strukturen. Klett-Cotta, Stuttgart
50. Popovych OV, Hauptmann C, Tass PA (2006) Control of neural synchrony by nonlinear delayed feedback. Biol Cybern 95:69–85
51. Rapp PE, Albano ME, Zimmerman ID et al (1994) Phase-randomized surrogates can produce spurious identifications of non-random structure. Phys Lett A 192:27–33
52. Rockstroh B, Watzl H, Kowalik ZJ, Cohen R, Sterr A, Müller M, Elbert T (1997) Dynamical aspects of the EEG in different psychopathological states in an interview situation. A pilot study. Schizophr Res 28:77–85
53. Rosenstein MT, Collins JJ, de Luca CJ (1993) A practical method for calculating Largest Lyapunov Exponents from small data sets. Physica D 65:117–134
54. Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD (2003) The neural basis of economic decision-making in the Ultimatum Game. Science 300:1755–1758
55. Saxena S, Rauch SL (2000) Functional neuroimaging and the neuroanatomy of obsessive-compulsive disorder. Psychiatr Clin North Am 23:563–586
56. Schaub H, Schiepek G (1992) Simulation of psychological processes: Basic issues and an illustration within the etiology of a depressive disorder. In: Tschacher W, Schiepek G, Brunner EJ (eds) Self-organization and clinical psychology. Springer, Berlin, pp 121–149

57. Schiepek G, Schoppek W, Tretter F (1992) Synergetics in psychiatry: Simulation of evolutionary patterns of schizophrenia on the basis of nonlinear difference equations. In: Tschacher W, Schiepek G, Brunner EJ (eds) Self-organization and clinical psychology. Springer, Berlin, pp 163–194
58. Schiepek G, Kowalik ZJ, Schütz A, Köhler M, Richter K, Strunk G, Mühlnickel W, Elbert T (1997) Psychotherapy as a chaotic process I. Coding the client-therapist-interaction by means of sequential plan analysis and the search for chaos: A stationary approach. Psychother Res 7:173–194
59. Schiepek G, Tominschek I, Karch S, Mulert C, Pogarell O (2007) Neurobiologische Korrelate der Zwangsstörungen – Aktuelle Befunde zur funktionellen Bildgebung. Psychother Psych Med 57:379–394
60. Schiepek G, Tominschek I, Karch S, Lutz J, Mulert C, Born C, Pogarell O (2008) A controlled single case study with repeated fMRI measures during the treatment of a patient with obsessive-compulsive disorder: Testing the nonlinear dynamics approach to psychotherapy. World J Biol Psychiatry. doi:10.1080/15622970802311829
61. Schubert C, Schiepek G (2003) Psychoneuroimmunologie und Psychotherapie: Psychosozial induzierte Veränderungen der dynamischen Komplexität von Immunprozessen. In: Schiepek G (ed) Neurobiologie der Psychotherapie. Schattauer, Stuttgart, pp 485–508
62. Schupp HAT, Lutzenberger W, Birbaumer N, Miltner W, Braun C (1994) Neurophysiological differences between perception and imagery. Cog Brain Res 2:77–86
63. Selye HA (1936) Syndrome produced diverse nocuous agents. Nature 138:32
64. Singer W, Gray CM (1995) Visual feature integration and the temporal correlation hypothesis. Ann Rev Neurosci 18:555–586
65. Sinha R, Lovallo WR, Parsons OA (1992) Cardiovascular differentiation of emotions. Psychosom Med 54:422–435
66. Skinner JE, Molnar M, Tomberg C (1994) The point correlation dimension: Performance with nonstationary surrogate data and noise. Int Physiol Behav Sci 29:217–234
67. Stadler M, Kruse P (1990) The self-organization perspective in cognition research. Historical remarks and new experimental approaches. In: Haken H, Stadler M (eds) Synergetics of cognition. Springer, Berlin, pp 32–52
68. Strauss JS (1989) Intermediäre Prozesse in der Schizophrenie: Zu einer neuen dynamisch orientierten Psychiatrie. In: Böker W, Brenner HD (eds) Schizophrenie als systemische Störung. Huber, Bern, pp 39–50
69. Strunk G (2004) Organisierte Komplexität. Mikroprozess-Analysen der Interaktionsdynamik zweier Psychotherapien mit den Methoden der nichtlinearen Zeitreihenanalyse. Dissertation, Universität Bamberg
70. Strunk G, Schiepek G (2006) Systemische Psychologie. Eine Einführung in die komplexen Grundlagen menschlichen Verhaltens. Spektrum Akademischer Verlag, Heidelberg
71. Tallon-Baudry C, Bertrand O (1999) Oscillatory gamma activity in humans and its role in object representation. Trends Cog Sci 3:151–162
72. Tallon-Baudry C, Bertrand O, Wienbruch C, Ross B, Pantev C (1997) Combined EEG and MEG recordings of visual 40 Hz resonses to illusory triangles in human. Neuroreport 8:1103–1107
73. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology (1996) Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. Eur Heart J 17:354–381
74. Tass P, Hauptmann C (2007) Therapeutic modulation of synaptic connectivity with desynchronizing brain stimulation. Int J Psychophysiol 64:53–61
75. Tass PA, Klosterkötter J, Schneider F, Lenartz D, Kouluosakis A, Sturm V (2003) Obsessive-compulsive disorder: Development of demand-controlled deep brain stimulation with methods of stochastic phase resetting. Neuropsychopharmacology 28:S27–S34
76. Tretter F, Scherer J (2006) Schizophrenia, neurobiology, and the methodology of systemic modelling. Pharmacopsychiatry 39(Suppl 1):S26–S35
77. Tschacher W (1997) Prozessgestalten. Hogrefe, Göttingen
78. Tschacher W, Kupper Z (2002) Time series models of symptoms in schizophrenia. Psychiatry Res 113:127–137
79. Tschacher W, Dauwalder JP (1999) Situated cognition, ecological perception, and synergetics: A novel perspective for cognitive psychology? In: Tschacher W, Dauwalder JP (eds) Dynamics, synergerics, autonomous agents. World Scientific, Singapore, pp 83–104
80. Vandenhouten R (1998) Analyse instationärer Zeitreihen komplexer Systeme und Anwendungen in der Physiologie. Shaker, Aachen
81. van Veen V, Carter CC (2002) The anterior cingulate as a conflict monitor: fMRI and ERP studies. Physiol Behav 77:477–482
82. van Veen V, Carter CC (2002) The timing of action-monitoring processes in the anterior cingulate cortex. J Cog Neurosci 14:593–602
83. Villmann T, Liebers C, Bergmann B, Gumz A, Geyer M (2008) Investigation of psycho-physiological interactions between patient and therapist during a psychodynamic therapy and their relation to speech in terms of entropy analysis using a neural network approach. New Ideas in Psychology 26:309–325
84. von Uexküll T, Wesiack W (1996) Wissenschaftstheorie: Ein biopsycho-soziales Modell. In: Adler RH, Herrmann JM, Köhle K, Schonecke OW, von Uexküll T, Wesiack W (eds) Thure von Uexküll. Psychosomatische Medizin. Urban Schwarzenberg, München, pp 13–52
85. Webber CL, Zbilut JP (1994) Dynamical assessment of physiological systems and states using recurrence plot strategies. J Appl Physiol 76:965–973
86. Yuru Z, Jan KM, Ju KH, Chon KH (2006) Quantifying cardiac sympathetic and parasympathetic nervous activities using principal dynamic modes analysis of heart rate variability. Am J Physiol Heart Circ Physiol 291:H1475–H1483

Self-Organization in Magnetohydrodynamic Turbulence

Pierluigi Veltri, Vincenzo Carbone, Fabio Lepreti, Giusy Nigro
Dipartimento di Fisica, Università della Calabria, Arcavacata di Rende, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
MHD Equations and Quadratic Invariants
Self-Organization at Large Scales
Self-Organization at Small Scales and Intermittent Structures in MHD Turbulence
Future Directions
Bibliography

Glossary

Alfvén speed The Alfvén speed c_A is the propagation speed of Alfvén waves and is given by $c_A = B_0/(4\pi\rho)^{1/2}$, where $\mathbf{B}_0$ is mean magnetic field and ρ the plasma mass density. Alfvèn waves are transverse, incompressible magnetohydrodynamic waves that propagate along $\mathbf{B}_0$ and originate from the tension of magnetic field lines.

Elsässer variables Elsässer variables $\mathbf{z}^\sigma$ are defined by $\mathbf{z}^\sigma = \mathbf{v} + \sigma\mathbf{B}/(4\pi\rho)^{1/2}$, with $\sigma = \pm 1$, $\mathbf{v}$ the velocity field, $\mathbf{B}$ the magnetic field and ρ the plasma mass density. The equations of incompressible MHD are often written in terms of these variables in order to describe the propagation of Alfvén waves and the non-linear couplings occurring in MHD turbulence.

Magnetohydrodynamics Magnetohydrodynamics (abbreviated, MHD) represents a one-fluid mathematical model which describes plasma dynamics at low frequencies: The main dynamical variables are the velocity of the fluid and the magnetic field. The vector equations for these variables are the fluid momentum conservation and the induction Maxwell equation.

MHD turbulence MHD turbulence is that turbulence which develops inside plasmas at macroscopic level, when viscosity and resistivity are low. Apart from velocity fluctuations which are also present in ordinary fluids, it is characterized also by the presence of magnetic field fluctuations.

Reverse field pinch Reverse field pinch (abbreviated, RFP) are plasma fusion toroidal devices whose conception is based on the idea that non-linear interactions in plasmas spontaneously give rise to magnetic structures where the Laplace force is zero (force free structures).

Shell models Shell models of turbulence are dynamical systems consisting of a set of ordinary differential equations representing a simplified version of the Navier–Stokes or MHD equations in the wavevector space. These models provide the possibility to investigate turbulence at very high Reynolds number regimes at the cost of neglecting information about spatial structures.

Solar corona The solar corona is the region extending from the solar surface up to one million of kilometers in the space, which can be visible to the naked eye during the eclipses. It is constituted mainly by a hydrogen plasma (proton and electrons) at a temperature of about two million degrees. The corona is highly structured by the magnetic field generated at the sun surface.

Solar wind The solar wind is a stream of plasma mainly composed of protons and electrons (hydrogen plasma), which flows out of the sun, due to the fact that plasma pressure associated to the very high coronal temperature overcomes the sun gravity. The flow velocity ranges from 250 km/s in the equatorial plane to about 900 km/s in the polar regions. Solar wind represents an extremely efficient plasma laboratory where the turbulence associated with the supersonic flow can be studied using space experiment data.

Definition of the Subject

Plasma dynamics at low frequency, i. e. at frequencies lower then the ion cyclotron frequency, can be described using a one-fluid model usually called magnetohydrodynamics (MHD), where the main dynamical variables are represented by fluid velocity and magnetic field, which evolve non-linearly being coupled to each other. This description applies both to laboratory plasma devices (tokamaks, reverse field pinch etc.), devoted to realize controlled nuclear fusion, and to space and astrophysical plasmas (Solar corona, Solar wind). Very often, when viscosity and resistivity are sufficiently small, the plasma behavior is characterized by the presence of a developed turbulence. In the last 20 years huge progress in understanding the properties of such turbulence has been realized, both by the use of high resolution computer simulations and by analysis of space and laboratory data. One of the most fascinating results of these studies concerns the evidence of self organization processes which have been shown to be very effective inside this kind of turbulence. Other aspects of turbulence and dynamical complexity in space plasmas are considered in the review by Chang ► Space Plasmas, Dynamical Complexity in in this Encyclopedia.

Introduction

Dynamical systems, whose time evolution is described by non-linear equations, can often give rise to chaotic behavior, i. e. to a behavior characterized by a strong dependence on initial conditions, which, in all practical cases,

does not allow to predict the final fate of such systems, or at least limits considerably the possibility to make predictions. Nevertheless, inside these systems, sometimes new complex, coherent structures, without any need of external intervention, can develop. This phenomenon is called self-organization.

Self-organization requires decreasing the entropy inside the coherent organized structures and exporting such entropy to the surroundings. For this reason the coherent structures can be called, following Prigogine, *dissipative structures*, in that their organization is always associated with an increased efficiency of dissipative processes around them.

MHD turbulence, as well as ordinary fluid turbulence, is characterized by a phenomenology, first described some centuries ago around the year 1500 by Leonardo da Vinci (Piumati 1894, fo. 74,v), where energy injected at large scales (injection range) is transferred by non-linear interactions (inertial range) towards smaller and smaller scales where it is finally dissipated (dissipative range). The physical phenomenon by which coherent organized structures are produced through an increased dissipation inside MHD turbulence has been evidenced both at large scale and at small scales. In the latter case it is usually called *intermittency*. We will discuss separately the two cases.

The occurrence of self organization at large scale seems to be directly related to the existence in MHD equations of some quadratic invariants (energy, magnetic helicity, cross helicity), which are conserved inside inviscid flows and are dissipated at very different rates in dissipative flows. Actually the existence of some long-living non-trivial states, towards which dissipative flows are attracted, has been evidenced for a long time. Moreover it has been suggested that these states could be derived from a variational principle, i. e. by minimizing an energy integral subject to some constraints on the other quadratic invariants. Woltjer [58] first explored the astrophysical implications of this possibility. The conjecture by Taylor [51] that an MHD system relaxes towards a state where the energy is minimum, subject to the constraint that magnetic helicity is conserved, allowed one to obtain a quite particular solution, usually called *Taylor's vortex*. This solution was extremely useful to explain the large scale behavior of reverse field pinch (RFP) devices [51]. The same conjecture has also been used to estimate the energy release in coronal structures [25,42] in connection to the problem of coronal heating.

The discovery in the solar wind of quite peculiar *Alfvénic fluctuations*, characterized by a high degree of correlation between velocity and magnetic field [3], which was interpreted by Dobrowolny et al. [20] as the result of a self organization of MHD turbulence, has suggested the formulation [10,38,53] of a different minimum principle: Alfvénic solutions can be obtained by minimizing the energy subject to the condition that cross-helicity is conserved.

In any case the energy minimization is dynamically realized through a *selective decay process*, i. e. a process during which some energy is dissipated into heat and the appearance of the coherent large scale structures is associated with redistribution of entropy to the surroundings.

The self organization processes discussed above, and related to the spontaneous creation of large scale coherent structures during the MHD turbulence dynamics, have their counterpart also at small scales. Actually, the nonlinear energy cascade process in fully developed MHD turbulence is characterized by the formation at the smallest dissipative scale of coherent structures. These structures are strictly related to the intermittency phenomenon; that is, to the breakdown of global self-similarity in the turbulent cascade, which represents a typical signature of nonlinear interactions.

In the following sections we first present the MHD equations (Sect. "MHD Equations and Quadratic Invariants") properties and discuss the characteristics of the self organized structures at large scales (Sect. "Self-Organization at Large Scales"). Then in Sect. "Self-Organization at Small Scales and Intermittent Structures in MHD Turbulence" we show that at small scales coherent dissipative structures are spontaneously formed in a physical process usually called *intermittency*. Finally in Sect. "Future Directions" we discuss the possible future developments and perspectives of such kind of studies.

MHD Equations and Quadratic Invariants

The equations describing the time evolution of an incompressible magnetofluid can be written

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} = -\frac{1}{\rho}\nabla p + (\nabla\times\mathbf{b})\times\mathbf{b} + \nu\nabla^2\mathbf{v}\,, \quad (1)$$

$$\frac{\partial \mathbf{b}}{\partial t} = \nabla\times(\mathbf{v}\times\mathbf{b}) + \mu\nabla^2\mathbf{b} \quad (2)$$

where $\mathbf{v}$ represents the fluid velocity and p the fluid pressure, $\mathbf{b} = \mathbf{B}/\sqrt{4\pi\rho}$ ($\mathbf{B}$ being the magnetic field and ρ the mass density), ν is the kinematic viscosity and $\mu = (c^2\eta)/(4\pi)$ the magnetic diffusivity (η being the fluid resistivity). The velocity and magnetic field must also satisfy the incompressibility conditions

$$\nabla\cdot\mathbf{v} = 0 \qquad \nabla\cdot\mathbf{b} = 0\,. \quad (3)$$

Quadratic Invariants

It can be easily shown that during their ideal time evolution, i. e. in absence of viscosity and resistivity, these equations conserve the following quadratic invariants: Total energy (kinetic + magnetic) per mass unit

$$E = \frac{1}{2}\int (\mathbf{v}^2 + \mathbf{b}^2)\mathrm{d}V \tag{4}$$

and cross helicity

$$H_\mathrm{c} = \int \mathbf{v}\cdot\mathbf{b}\mathrm{d}V\,. \tag{5}$$

Moreover Woltjer [58] showed that in the ideal 3D MHD, also magnetic helicity

$$H_\mathrm{m} = \int \mathbf{a}\cdot\mathbf{b}\mathrm{d}V \tag{6}$$

($\mathbf{a}$ being the magnetic vector potential defined through $\mathbf{b} = \nabla\times\mathbf{a}$) remains invariant during the evolution of any closed magnetic flux system. In 2D MHD on the contrary, the third conserved quantity is the square of the magnetic field vector potential [24]

$$A = \int \mathbf{a}^2\mathrm{d}V \tag{7}$$

In Eqs. (4), (5), (6) and (7) the integrals are extended to the whole volume of the magnetofluid.

Elsässer's Variables

In the following we will find it useful to rewrite these equations in terms of the Elsässer [21] variables $\mathbf{z}^\sigma$ defined by $\mathbf{z}^\sigma = \mathbf{v} + \sigma\mathbf{b}$ with $\sigma = \pm 1$. Using such variables the equations governing incompressible MHD are recast to the more compact form

$$\frac{\partial\mathbf{z}^\sigma}{\partial t} + (\mathbf{z}^{-\sigma}\cdot\nabla)\mathbf{z}^\sigma = -\frac{1}{\rho}\nabla(p + \frac{B^2}{8\pi}) + \nu^+\nabla^2\mathbf{z}^\sigma + \nu^-\nabla^2\mathbf{z}^{-\sigma} \tag{8}$$

with $\nu^\sigma = \nu + \sigma\mu$. Obviously also the fields $\mathbf{z}^\sigma$ must satisfy the incompressibility conditions

$$\nabla\cdot\mathbf{z}^\sigma = 0\,. \tag{9}$$

In terms of these variables, the conservation of total energy per mass unit and cross helicity can be written as the conservation of the two pseudo-energies

$$E^\sigma = \frac{1}{2}\int \mathbf{z}^{\sigma^2}\mathrm{d}V\,. \tag{10}$$

MHD equations display and infinite number of ideal non-quadratic invariants, but quadratic invariants (4), (5) and (6) (or in 2D (7)) play a key role. Actually, let us introduce the Fourier transform of the field $\mathbf{z}^\sigma$ in a cubic periodic box of size a

$$\mathbf{z}^\sigma(\mathbf{r}, t) = \sum_\mathbf{k} \mathbf{z}^\sigma(\mathbf{k}, t)\exp(i\mathbf{k}\cdot\mathbf{r})$$

where $\mathbf{k} = 2\pi\mathbf{m}/a$, and $\mathbf{m}$ is vector of integer numbers. In terms of these Fourier transforms (taking into account the incompressibility condition (9) and neglecting dissipative terms), MHD Equations (8) can be written

$$\frac{\partial z_i^\sigma(\mathbf{k}, t)}{\partial t} = \sum_{l,s=1}^{3}\sum_{\mathbf{p},\mathbf{q}} M_{i,l,s}(\mathbf{k},\mathbf{p},\mathbf{q}) z_l^\sigma(\mathbf{p}, t)\, z_s^{-\sigma}(\mathbf{q}, t)\delta_{\mathbf{k},\mathbf{p}+\mathbf{q}} \tag{11}$$

with $M_{i,l,s}(\mathbf{k},\mathbf{p},\mathbf{q}) = -i\left(\delta_{il} - \frac{k_i k_l}{k^2}\right)k_s$ and where $\delta_{\mathbf{k},\mathbf{p}+\mathbf{q}}$ is the usual Kronecker symbol. The peculiar form of these equations shows that non-linear terms are characterized by triad interactions among wave vectors $\mathbf{k}$, $\mathbf{p}$ and $\mathbf{q}$ such that $\mathbf{k} = \mathbf{p} + \mathbf{q}$. Quadratic invariants are also called *rugged* invariants, because they are conserved in any single interaction among a triad of wave vector and for this reason

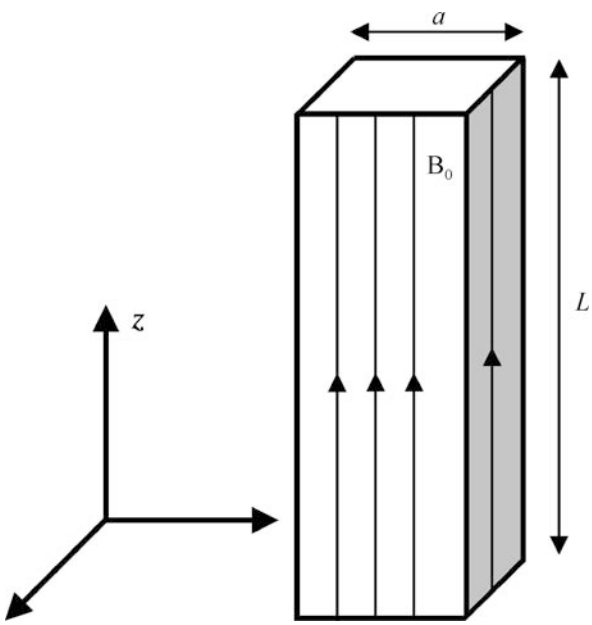

Self-Organization in Magnetohydrodynamic Turbulence, Figure 1
Sketch of the reduced MHD configuration

they survive also when a Galerkin approximation is performed on MHD equations. In other words, even if the r. h. s. of (11) contains an infinite number of non-linear interactions, $E(t)$, $H_c(t)$ and $H_m(t)$ remain invariant also by retaining any finite number of interacting triads in Equation (11).

Reduced Magnetohydrodynamics

In some fusion plasma devices, or in the solar corona, the plasma β parameter (that is the ratio between the kinetic and magnetic pressure) is low $\beta \simeq 10^{-2}$, and MHD equations can be simplified into the so-called Reduced MHD [47,61]. This approximation is valid, for example, for a plasma column with a low aspect ratio $a/L \ll 1$ (a and L being respectively the width and the height of the plasma column), with a strong magnetic field $\mathbf{B}_0$ along the z-direction (Fig. 1).

In this case the plasma dynamical variables reduce to the velocity field $\mathbf{v}_\perp$ and the magnetic field $\mathbf{b}_\perp$ perpendicular to $\mathbf{B}_0$, so that starting from MHD Equs. (8), the following set of RMHD equations [61] is obtained:

$$\frac{\partial \mathbf{z}^\sigma}{\partial t} - \sigma c_A \frac{\partial \mathbf{z}^\sigma}{\partial z} + (\mathbf{z}^{-\sigma} \cdot \nabla_\perp)\mathbf{z}^\sigma = -\frac{1}{\rho}\nabla(p + \frac{B^2}{8\pi}) + \nu^+ \nabla_\perp^2 \mathbf{z}^\sigma + \nu^- \nabla_\perp^2 \mathbf{z}^{-\sigma} \quad (12)$$

$$\nabla_\perp \cdot \mathbf{z}^\sigma = 0 \quad (13)$$

where $\mathbf{z}^\sigma = \mathbf{v}_\perp + \sigma \mathbf{b}_\perp$ (with $\sigma = \pm 1$), $\nabla_\perp$ is the spatial gradient perpendicular to $\mathbf{B}_0$, $c_A = B_0/(4\pi\rho)^{1/2}$ is the Alfvén velocity associated with B_0.

Equation (12) shows that, in this approximation, nonlinear interactions are retained only in the directions perpendicular to $\mathbf{B}_0$, while only propagation at the Alfvén speed takes place parallel to $\mathbf{B}_0$.

Let us now Fourier transform the fields with respect to the perpendicular variable $\mathbf{r}_\perp = (x, y)$, in a 2D periodic box of size a

$$\mathbf{v}_\perp(\mathbf{r}, z, t) = \sum_{\mathbf{k}} v(\mathbf{k}, z, t)\mathbf{e}(\mathbf{k}) \exp(i\mathbf{k} \cdot \mathbf{r}) , \quad (14)$$

$$\mathbf{b}_\perp(\mathbf{r}, z, t) = \sum_{\mathbf{k}} b(\mathbf{k}, z, t)\mathbf{e}(\mathbf{k}) \exp(i\mathbf{k} \cdot \mathbf{r}) , \quad (15)$$

where $\mathbf{k} = 2\pi \mathbf{m}/a$, ($\mathbf{m}$ is a couple of integers) and $\mathbf{e}(\mathbf{k})$ is a unit vector perpendicular to $\mathbf{k}$. After some algebra, it can be shown that in this approximation, for each value of z, the MHD equations reduce to [49]

$$\begin{aligned} \frac{\partial v(\mathbf{k})}{\partial t} &= c_A \frac{\partial b(\mathbf{k})}{\partial z} + \sum_{\mathbf{pq}} c(k, p, q)(p^2 - q^2) \\ &\quad \times \left[v(\mathbf{p})v(\mathbf{q}) - b(\mathbf{p})b(\mathbf{q})\right] \delta_{\mathbf{k},\mathbf{p}+\mathbf{q}} \\ \frac{\partial b(\mathbf{k})}{\partial t} &= c_A \frac{\partial v(\mathbf{k})}{\partial z} + \sum_{\mathbf{pq}} c(k, p, q)k^2 \\ &\quad \times \left[b(\mathbf{p})v(\mathbf{q}) - v(\mathbf{p})b(\mathbf{q})\right] \delta_{\mathbf{k},\mathbf{p}+\mathbf{q}} \end{aligned} \quad (16)$$

(for simplicity we omit the time and z dependence of the Fourier amplitudes) where

$$c(k, p, q) = \frac{p_x q_y - p_y q_x}{2kpq}$$

means that the sum is extended over all wave vectors $\mathbf{p}$ and $\mathbf{q}$ which satisfy the triad-interaction relation $\mathbf{k} = \mathbf{p} + \mathbf{q}$. Let us note that, if we chose a box of size $a = 2\pi$, each wave vector $\mathbf{k}$ turns out to be represented by a couple of integers.

Equations (16) have quadratic invariants namely the total energy

$$E(t) = \sum_{\mathbf{k}} [|v(\mathbf{k}, t)|^2 + |b(\mathbf{k}, t)|^2]$$

and the cross-helicity

$$H_c(t) = \sum_{\mathbf{k}} \mathrm{Re}[v(\mathbf{k}, t)b^*(\mathbf{k}, t)] .$$

When the background magnetic field is set to zero $B_0 = 0$, also the mean-square of the vector potential

$$A(t) = \sum_{\mathbf{k}} |b(\mathbf{k}, t)|^2/k^2$$

is conserved. On the contrary, when $B_0 \neq 0$, the last quantity is almost constant, with relative fluctuations of about $\Delta A/A \simeq 10^{-2}$ [49].

This means that 2D MHD has a much wider field of physical applications then those situations where no explicit dependence on the z components exists. Actually 2D MHD represents a good approximation for the low-beta and low aspect ratio plasmas as for example some laboratory devices (tokamaks, RFP, etc.) or some magnetic structures (loops) in the solar corona. In the next section we will take advantage of this result.

Self-Organization at Large Scales

Relaxation Processes in MHD

In MHD, the quadratic invariants (4, 5, 6) play a crucial role in the problem of predicting the final fate of solutions starting from quite general initial conditions. When

studying the ideal relaxation, the quadratic invariants remain constant during time evolution, and the problem of predicting the final state reduces to calculating the ensemble–averaged equilibrium spectra of these invariants. Using a standard statistical approach it has been shown by Frisch et al. [23] that these spectra are determined by the ensemble-averaged initial values of invariants. Dissipative relaxation processes in MHD are much more complicated, since the values of the ideal invariants change during the evolution. It seems however that some long-living nontrivial states toward which dissipative flows are attracted exist, and that these states could be derived by minimizing an energy integral subject to some constraints.

Dissipative relaxations have been investigated in connection with measurements in laboratory plasmas, and observations in the solar wind turbulence. Taylor [51], inspired by laboratory plasma experiments, conjectured that an MHD system relaxes towards a state where the energy tends to a minimum, subject to the constraint that magnetic helicity is conserved. The relaxation can then be seen as a *selective decay* between the two invariants. From a mathematical point of view, the solution of the problem can be obtained through a variational principle

$$\delta\left\{\int(\mathbf{v}^2+\mathbf{b}^2)\mathrm{d}V-\lambda\int\mathbf{a}\cdot\mathbf{b}\mathrm{d}V\right\}=0\,,$$

where λ is a Lagrange multiplier. By imposing the conservation of magnetic helicity, we get the so-called force-free solution, characterized by $\mathbf{v}=0$ and

$$\nabla\times\mathbf{b}=\alpha\mathbf{b} \tag{17}$$

(α being a constant), i. e. a solution where the velocity field and the Laplace force are both null. This means that the kinetic energy decays to zero and the magnetic energy occupies the largest scale. Taylor's hypothesis allows one to circumvent in an elegant way the complex plasma relaxation dynamics (involving turbulence and reconnection).

When looking at (17) in cylindrical coordinates, the following solution can be obtained

$$B_T=B_0J_0(\alpha r)\quad B_P=B_0J_1(\alpha r) \tag{18}$$

where B_T and B_P are respectively the axial and poloidal components of the magnetic field (B_0 is a constant and J_0 and J_1 are Bessel functions). This solution, which can display a change in the sign of the magnetic field axial component (Fig. 2), was able to explain the puzzling phenomenon of field reversal spontaneously observed in some laboratory experiments, where no current necessary to produce such reversal was present in the initial state. This theory has also been the basis for the development

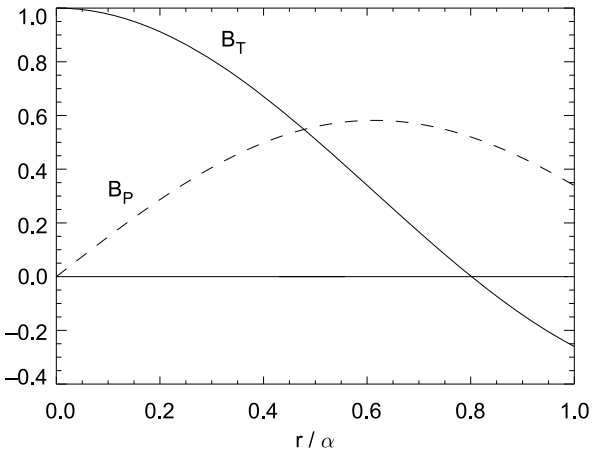

Self-Organization in Magnetohydrodynamic Turbulence, Figure 2

Axial (*solid line*) and poloidal (*dashed line*) components of magnetic field in a force free cylindrical plasma column as function of the radial coordinate normalized to α^{-1} ($B_0=1$)

of RFP devices (Fig. 3), toroidal devices which at a first order approximation can be considered periodic cylinders (Fig. 4). Applications to spheromak and other plasma fusion devices [4,52,59] have been considered more recently.

Belcher and Davis [3] analyzing solar wind data discovered velocity and magnetic field fluctuations were extremely well-correlated (Fig. 5). Dobrowolny et al. [20] suggested that this correlation could be the result of a self organization of MHD turbulence and set up a phenomenological explanation for this self organization, based on the idea that between the two modes of propagation of *Alfvénic* fluctuations in solar wind only that initially dominating survives to the non-linear cascade process towards dissipative scales.

This kind of self organization can also be obtained through a variational principle. Let us impose that energy assumes a minimum value, constrained to the conservation of cross-helicity, we have then to write

$$\delta\left\{\int(\mathbf{v}^2+\mathbf{b}^2)\mathrm{d}V-\lambda\int\mathbf{v}\cdot\mathbf{b}\mathrm{d}V\right\}=0\,,$$

once again λ is a Lagrange multiplier. By imposing the conservation of cross helicity, we finally get the two solutions

$$\mathbf{v}=\sigma\mathbf{b} \tag{19}$$

with $\sigma=\pm1$. These solutions, which can also be written $\mathbf{z}^\sigma=0$ for either $\sigma=1$ or $\sigma=-1$ in terms of the Elsässer [21] variables, correspond to annihilating nonlinear interactions in (1, 8). This phenomenon has been

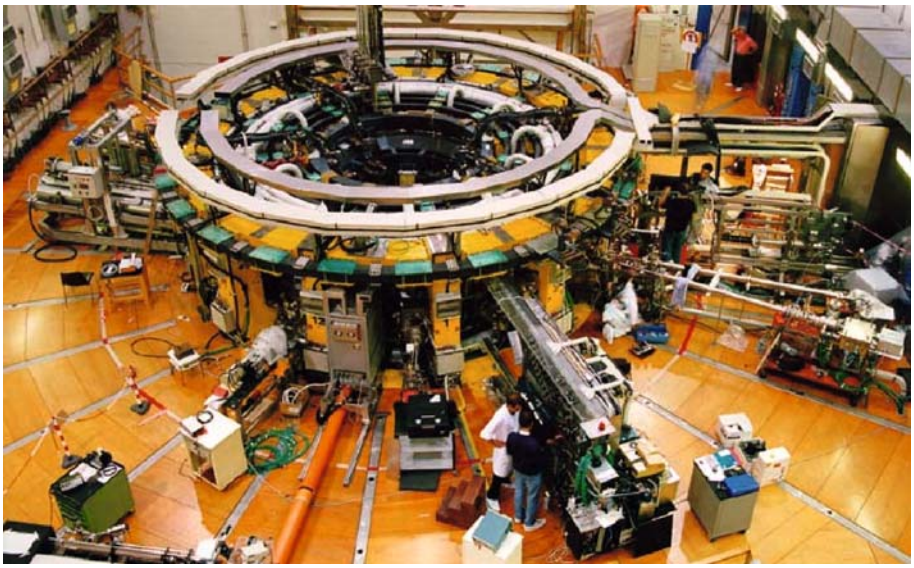

Self-Organization in Magnetohydrodynamic Turbulence, Figure 3
The RFX: A reverse field pinch plasma device in Padua (Italy) (Consorzio RFX is a research organization promoted by CNR, ENEA, Università di Padova, Acciaierie Venete S.p.A. and INFN, within the framework of the Euratom – ENEA Association)

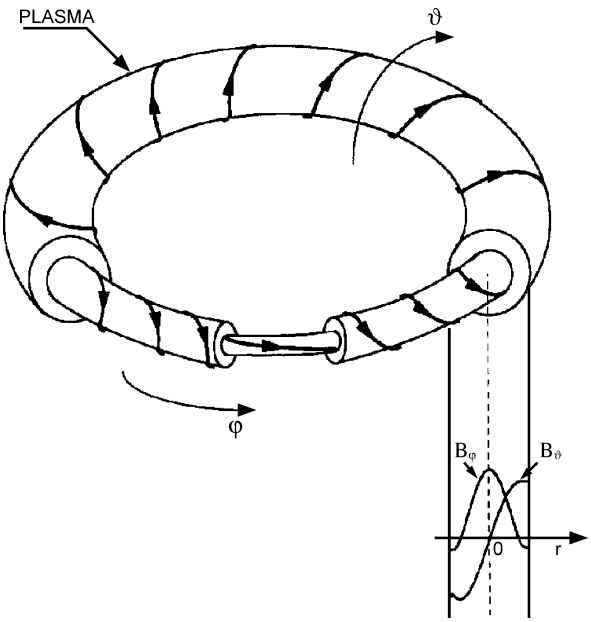

Self-Organization in Magnetohydrodynamic Turbulence, Figure 4
Typical profile of the magnetic field in an RFP device (reprinted from [48])

called *dynamical alignment* in that velocity and magnetic field tend to become aligned to each other.

It has been shown that also in numerical simulations of 2D and 3D MHD equations [17,38,39,46,53] there exists systematic behavior of turbulent flows. In particular, relaxation processes which bring the fluid toward those particular states like *Taylor solutions* or *Alfvénic solutions* have been found to emerge systematically from numerical simulations. In terms of ideal invariants, it has been shown that the Taylor's solution is obtained, when starting with almost zero cross helicity, while the *dynamical alignment* is found when the initial value of the cross helicity is substantially different from zero. In general when both effects are in competition, the situation is much more complicated. A systematic analysis of a wide number of 2D MHD simulations at different resolutions have been performed by Ting et al. [53]. Their results can be summarized by referring to the time behavior of the system projected on the plane (a, h), where $a = A/E$ and $h = 2H_c/E$. In fact, starting from whatever initial conditions, in that plane the system tends to approach the ellipse

$$\frac{1}{a} = 2\left(\frac{k_{\min}}{h}\right)^2 \left[1 \pm \sqrt{1-h^2}\right], \tag{20}$$

where $k_{\min}$ represents the minimum wave vector allowed in the simulations. Of course the points $(a, h) = (1/2, \pm 1)$ represents the *attractors* of dynamical alignment solutions, the point $(a, h) = (1, 0)$ represents the attractor of selective decay, and the point $(a, h) = (0, 0)$ represents the attractor when the system behaves like a fluid, with zero magnetic field. These extreme points can be recovered through the minimization of energy subject to the

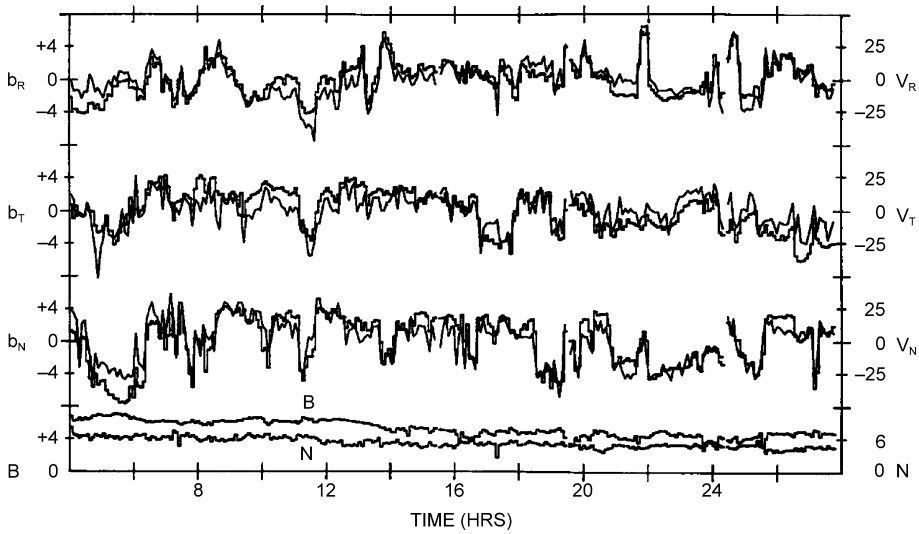

Self-Organization in Magnetohydrodynamic Turbulence, Figure 5

Velocity (v_R, v_T, and v_N) and magnetic field (b_R, b_T, and b_N) fluctuations measured in solar wind in the RTN reference system (reprinted from [3]). The RTN system has the R unity vector along the radial direction, positive from the sun to the spacecraft, the T unity vector perpendicular to the plane formed by the rotation axis Ω of the sun and the radial direction, i. e., T = Ω × R, and the N component resulting from the vector product N = R × T. Particle density *N* and magnetic field intensity *B* are also shown in the lowest part of the plot. Units are 10^{-5} gauss for magnetic fields, km/s for velocities, and cm^{-3} for particle density

conservation of the other invariants. When the magnetic field is zero, the kinetic helicity is invariant and the point $(a, h) = (0, 0)$ can be recovered by allowing that the kinetic energy decays, subject to the constraint that the kinetic helicity remains constant. The entire curve, however, does not represent the locus of the extreme of anything over its entire range of definition.

A Minimal Triad-Interaction Model of Relaxation Processes in 2D MHD

As we have seen, 2D MHD plays a privileged role, in that it describes low-beta and low aspect ratio 3D plasmas. Here we briefly present how the basic non-chaotic triad-interaction works in 2D MHD, and how this represents a minimal basic model for relaxation processes. Choosing only three wave vectors, namely $\mathbf{k}_1 = (1, 1)$, $\mathbf{k}_2 = (2, -1)$, $\mathbf{k}_3 = (3, 0)$, we obtain a set of 12 ordinary differential equations [18] for the complex Fourier modes of both velocity $u_i(t) = u(\mathbf{k}_i, t)/2\sqrt{5}$ and magnetic field $b_i(t) = b(\mathbf{k}_i, t)/2\sqrt{5}$. The system can be further reduced through a projection of equations on a subset of the phase space, that is by considering only real fields. This can be seen by writing the fields in the form $u_j = |u_j|e^{i\alpha_j}$ and $b_j = |b_j|e^{i\beta_j}$, and by defining real fields through $V_j(t) = |u_j| \cos\alpha_j$ and $B_j(t) = |b_j| \cos\beta_j$ (subject to the conditions $\sin\alpha_j = 0$ and $\sin\beta_j = 0$). In this case we found a set of 6 ordinary differential equations for V_j and B_j, namely

$$
\begin{aligned}
(\mathrm{d}/\mathrm{d}t + 2\nu)V_1 &= 4(V_3V_2 - B_3B_2)\\
(\mathrm{d}/\mathrm{d}t + 5\nu)V_2 &= -7(V_3V_1 - B_3B_1)\\
(\mathrm{d}/\mathrm{d}t + 9\nu)V_3 &= 3(V_1V_2 - B_1B_2)\\
(\mathrm{d}/\mathrm{d}t + 2\mu)B_1 &= 2(B_3V_2 - V_3B_2)\\
(\mathrm{d}/\mathrm{d}t + 5\mu)B_2 &= 5(V_3B_1 - B_3V_1)\\
(\mathrm{d}/\mathrm{d}t + 9\mu)B_3 &= 9(V_1B_2 - B_1V_2)\,.
\end{aligned}
$$

For its simplicity the model represents the basic system to investigate the structure of nonlinear interactions in 2D MHD, and to study the role played by the rugged invariants during the dynamical evolution.

Let us introduce the phase space Ω, of dimension $\mathrm{Dim}[\Omega] = 6$, which can be built up by using as coordinates the Fourier amplitudes. A point $\Psi_i(t) \in \Omega$, defined as $\Psi_i(t) := \{[u_i(t), b_i(t)] \in \Omega\}$ represents the system at a certain time, and this point moves in Ω according to the flow $T_\tau[\Psi_i(t)] = \Psi_i(t + \tau)$ which represents the result of the equation of motion for the system. In absence of dissipative terms the phase space volume is conserved

$$
H = \sum_i \frac{\partial}{\partial \Psi_i}\left(\frac{\mathrm{d}\Psi_i}{\mathrm{d}t}\right) = 0\,, \tag{21}
$$

where H represents the rate of change of volumes in phase space.

If we define the ideal flow $T_\tau^{\rm id}$ as the flow T_τ obtained when $\nu = \mu = 0$, the phase space volume conservation can be expressed as $T^{id}_{\tau\to\infty}[\Psi_i(t)] = \Psi_i(t+\tau)$. In this case, for a given set of initial values $\Psi_i(0)$, the point moves on a hyper-surface $S \subset \Omega$ defined by the initial value of the invariants. In presence of dissipative terms the quadratic invariants decay, and the rate of change of the volume in the phase space is $H = -16(\nu+\mu) \le 0$. The condition $H \le 0$ implies that the dissipative flow pushes the system toward the trivial asymptotic state where all the amplitudes of the fields are zero $T^{\rm diss}_{\tau\to\infty}[\phi(t)] = 0$.

The triad-interaction model is able to capture the dynamics of the quiescent states observed in MHD. We solved our system by starting from random initial conditions uniformly distributed $\Psi_i(0) \in [-1, 1]$. We used a fourth-order Runge–Kutta scheme, with a time step $\Delta t = 10^{-3}$ and $\nu = \mu = 0.01$. This value for the dissipative coefficients allows the nonlinear interactions to have sufficient time to set up the dynamical behavior. In Fig. 6 we report the curve ε along with two ensembles of points. The first ensemble (white circles) represents the set of points (a, h) obtained with some different initial conditions $\Psi_i(0)$ randomly chosen in the interval $[-1; 1]$. The second ensemble (black circles) represents the set of points (a, h) at time $t = 80$ (in unit of time steps), calculated from the set of fields $\Psi_i(t)$ which are obtained through the time evolution of the set $\Psi_i(0)$, that is $\Psi_i(t) = T_t^{\rm diss}[\Psi_i(0)]$. As it can be seen that all the initial conditions lead to the final state which belongs to the ellipse ε.

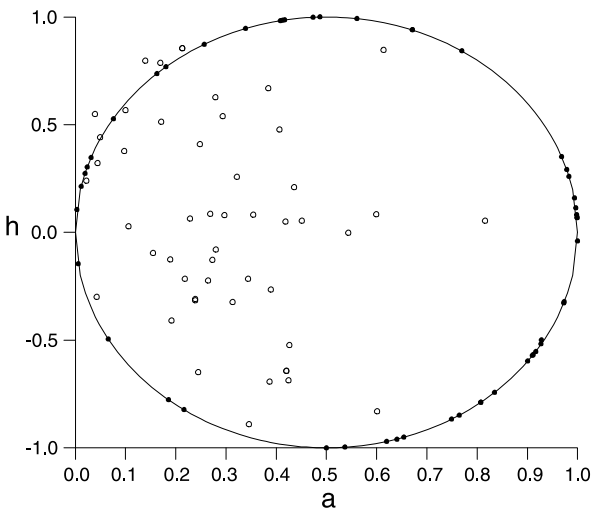

Self-Organization in Magnetohydrodynamic Turbulence, Figure 6

Numerical simulations of the triad-interaction model reported in the plane (a, h). We show two sets of different solutions at two different times. *White circles* refer to a set of 45 initial values $\Psi_i(0)$, *black circles* represent the set of point at a given time $t = 80$ (in unit of time steps) obtained from the time evolution of the above initial values $\Psi_i(t = 80) = T^{\rm diss}_{t=80}[\Psi_i(0)]$. The solutions represented on that plane are such that, after an initial transient, they fall on the ellipse represented as a *full line*

Since nonlinear interactions in the simple triad-interaction model have the same structure as in the 2D MHD equations, the model is able to capture relaxation properties. These properties are the fixed points and some invariant subspaces. The fixed points of the system can be classified as follows:

a) Two Alfvénic fixed points (say $A^\pm$) characterized by $u_i = \pm b_i$;
b) Three fluid fixed points $F(i)$ $(i = 1, 2, 3)$ where all variables but the velocity V_i are zero;
c) Three magnetic fixed points $M(i)$ $(i = 1, 2, 3)$ where all variables but the magnetic field B_i are zero.

A standard analysis of stability can be performed by linearizing the system around each fixed point. We find that $A^\pm$ are always stable, $M(1)$ is the only stable magnetic fixed point, while $F(1)$ and $F(3)$ are stable. The only stable magnetic fixed point $M(1)$ is such that the energy is localized on the minimum wave vector.

Apart from fixed points the system displays some other interesting properties. Looking at the equations, it can be easily shown that there exists some 3D subspaces of the 6D phase space which remain invariant under the ideal flow operator. Let us denote by I^α the αth invariant subspace, which is then characterized by the fact that if $\Psi_i(0) \in I^\alpha$ then $\Psi_i(t) = T_t^{id}[\Psi_i(0)] \in I^\alpha$ for each $t > 0$. In other words an ideal invariant subspace is a portion of the phase space where the system lies for all times. The most useful way to classify these structures is through the initial values the rugged invariants assume on them, since under the ideal flow they remain constant:

a) A fluid subspace F, characterized by $a = A/E = 0$ and $h = 2H_c/E = 0$. This can be recovered by imposing that the magnetic field is always zero, namely

$$\Psi_i[F] = (V_1, V_2, V_3, 0, 0, 0)$$

is the vector which describes this subspace.
b) Two Alfvénic subspaces $A^\pm$, characterized by $a = A/E = 0$ and $h = 2H_c/E = 1$, which can be recovered by imposing that $V_i = \pm B_i$ for each $i = 1, 2, 3$. These subspaces are also fixed points of the system.

c) Three magnetic subspaces H^A, H^B and H^C characterized by $a \neq 0$ and $h = 0$, i. e. a minimal value for the cross-helicity. These subspaces can be recovered by imposing that the cross-helicity is initially zero over all wave vectors, namely either $V_i = 0$ and $B_i \neq 0$ or vice versa; specifically,

$$\begin{aligned}\Psi_i[H^A] &= (0, 0, V_3, B_1, B_2, 0)\\ \Psi_i[H^B] &= (V_1, 0, 0, 0, B_2, B_3)\\ \Psi_i[H^C] &= (0, V_2, 0, B_1, 0, B_3)\,.\end{aligned}$$

The stability properties of each subspace can be investigated numerically [11]. Simulations start by putting the system on a given subspace, and by adding a small perturbation on the complementary manifold. The motion is not limited to the 3D subspace, and the problem of the stability of a particular I^α consists of determining if the perturbed solution remains close to I^α or goes away from it covering all the allowed 6D phase space. For each subspace we can define two pseudo-energies $E^{(\alpha)}_{\text{int}}(t)$ (built up with the fields which belong to the αth subspace) and $E^{(\alpha)}_{\text{ext}}(t)$ (built up with the fields which do not belong to the αth subspace). The external energy represents the distance $\|\Delta^{(\alpha)}[\Psi_i(t)]\|$ between the point $\Psi_i(t)$ and the αth invariant subspace. Since the total energy $E(t) = E^{(\alpha)}_{\text{ext}}(t) + E^{(\alpha)}_{\text{int}}(t)$ must remain constant under the ideal flow, two situations can arise, namely

1) During the time evolution both $E^{(\alpha)}_{\text{ext}}(t) \simeq E^{(\alpha)}_{\text{ext}}(0)$ and $E^{(\alpha)}_{\text{int}}(t) \simeq E^{(\alpha)}_{\text{int}}(0)$, which means that the solution remains trapped near the invariant subspace. In that case, the subspace is ideally stable.
2) During the time evolution energies become comparable $E^{(\alpha)}_{\text{ext}}(t) \simeq E^{(\alpha)}_{\text{int}}(t)$, which means that the solution is repelled from the invariant subspace. In that case the subspace is ideally unstable.

Numerical simulations show that the fluid subspace is always stable. As far as the magnetic subspaces are concerned, we find that H^A is always stable, while H^B and H^C are always unstable. More information about the behavior of the system near the invariant subspaces can be obtained by considering the characteristic of solutions when the dissipative coefficients are set different from zero. In that case, the energies decay in time, but the rate of decay is different, thus indicating a kind of selective dissipation. In particular, looking at the time evolution of their ratios $R^\alpha(t) = E^{(\alpha)}_{\text{ext}}(t)/E^{(\alpha)}_{\text{int}}(t)$, we find that it decays to zero for the ideally stable subspaces, while it settles to a constant value for the unstable subspaces. This means that ideally stable subspaces, namely the Alfvénic, fluid and magnetic H^A represent a kind of attractor for the system, while unstable subspaces, say H^B and H^C, repel solutions.

When we start numerical solutions near the stable subspaces, the system is attracted to the extreme points of the plane (a, h). When we start near the fluid attractor, the system reach the point $(a, h) = (0, 0)$, and when we start near one of the Alfvénic attractors the system is attracted to $(a, h) = (1/2, \pm 1)$. Finally, when we start near the stable magnetic subspace H^A the system is attracted to the extreme point $(a, h) = (1, 0)$. This last point represents a Taylor regime, say the magnetic field lies on the lowest allowed wave vector. On the contrary, when we start near the unstable magnetic attractors H^B or H^C, the system evolves in an erratic way towards any point (a, h) of the curve (20). Each point (a, h) is made by only one mode with wave vector $k_1 = k_{\min} = \sqrt{2}$, so that if $\eta = V_1/B_1$ we have

$$\begin{aligned} a &= \frac{2}{\left(1+\eta^2\right) k_{\min}^2}\\ h &= \frac{\eta}{1+\eta^2}\,. \end{aligned} \tag{22}$$

This last equation is nothing but the parametric equation of the curve (20).

The behavior we have just described can be recovered also when we start the numerical computation with general initial conditions. In that case, we can associate to each initial condition an invariant subspace according to the rule that the αth subspace is associated with the initial condition $\Psi_i(0)$ when the distance $\|\Delta^{(\alpha)}[\Psi_i(0)]\|$ is the minimum one over α. In other words, we associate an initial condition to the nearest subspace. Numerical simulations show that when the initial conditions are associated to the unstable subspaces H^B and H^C, the final point reached by the system will be any point (a, h) on the curve (20). When the initial conditions are associated with one of the stable subspaces, the final state will be one of the extreme points of the curve (20).

Self-Organization at Small Scales and Intermittent Structures in MHD Turbulence

The self organization processes discussed in the previous sections are related to the spontaneous creation of large scale coherent structures during the MHD turbulence dynamics. However, also at small scales the nonlinear energy cascade process in fully developed MHD turbulence is characterized by the presence of self organization and coherent structures. These structures are strictly related to the intermittency phenomenon, that is, the breakdown of global self-similarity in the turbulent cascade.

The turbulence theory proposed by Kolmogorov in 1941 [30] is based on the conjecture that, within the inertial range, the statistical properties of the fluid motions depend only on the mean energy transfer rate ε and on the scale ℓ, and that the nonlinear energy cascade occurring in this range is a self similar (fractal) process. According to this idea, the velocity field increments $\delta \boldsymbol{v}_\ell(\boldsymbol{r}) = \boldsymbol{v}(\boldsymbol{r}+\boldsymbol{\ell}) - \boldsymbol{v}(\boldsymbol{r})$ scale as $\delta v_\ell \sim \ell^{1/3}$. The self similarity of the turbulent cascade implies that the probability density function (PDF) of field increments $P(\delta v_\ell)$ should be invariant under the scale change

$$P(\delta v_\ell) = \ell^{-1/3} F\left(\frac{\delta v_\ell}{\ell^{1/3}}\right) , \tag{23}$$

which means that PDFs of normalized velocity increments at different scales collapse onto the same curve. Another consequence of the global self similarity is that the structure functions $S_p(\ell) = \langle \delta v_\ell^p \rangle$, usually defined using the longitudinal field increments, that is, $\delta v_\ell(\boldsymbol{r}) = \left[\boldsymbol{v}(\boldsymbol{r}+\boldsymbol{\ell}) - \boldsymbol{v}(\boldsymbol{r})\right] \cdot \frac{\boldsymbol{\ell}}{\ell}$, should follow the scaling law $S_p(\ell) \sim \ell^{p/3}$.

When considering incompressible MHD turbulence, the scaling properties of the field increments $\delta \boldsymbol{z}_\ell^\sigma(\boldsymbol{r}) = \boldsymbol{z}^\sigma(\boldsymbol{r}+\boldsymbol{\ell}) - \boldsymbol{z}^\sigma(\boldsymbol{r})$, expressed here in terms of Elsässer variables $\boldsymbol{z}^\sigma = \boldsymbol{v} + \sigma \boldsymbol{B}/\sqrt{4\pi\rho}$ ($\boldsymbol{B}$ being the magnetic field and ρ the mass density) can be modified with respect to the fluid case by the so-called Alfvén effect, consisting of the fact that nonlinear interactions take place between eddies of different σ, which propagate in opposite directions at the Alfvén velocity along the large scale magnetic field reducing the efficiency of the non linear energy cascade [20,29,31].

In the fluid-like case a Kolmogorov-like scaling can be expected for the field increments, that is, $\delta z_\ell^\sigma \sim \ell^{1/3}$. On the other hand, when the Alfvén effect is at work the Kraichnan's scaling $\delta z_\ell^\sigma \sim \ell^{1/4}$ is recovered. However, in both cases the turbulent energy cascade is supposed to be a self similar process, resulting in a linear behavior of the structure function exponents, that is, $S_p(\ell) \sim \ell^{p/m}$, where $m = 3$ and $m = 4$ for the fluid-like and Alfvenic case respectively.

Several studies on the statistical properties of fields in fully developed turbulence have shown that the PDFs of field increments are not self similar, that is, their shape change with the scale ℓ. The PDFs at large scales are nearly Gaussian, while, as the scale decreases, the PDFs show sharper and sharper peaks and, correspondingly, higher and higher tails. As far as MHD turbulence is concerned, this behavior has been found both in the solar wind [36,50] and in laboratory plasmas [13]. This departure from self similarity has also been inferred from structure function analysis (see e. g. [8,12,13,35]) which have shown that for both velocity and magnetic field fluctuations the scaling exponents are nonlinear functions of p.

This behavior indicates thus that the turbulent, nonlinear energy cascade is not a fractal (self similar), but rather a multifractal process. In other words, small scale fluctuations much larger than their RMS are present in some spatial positions and the corresponding spatial fluctuations of the energy transfer rate $\varepsilon(\boldsymbol{r})$ play a primary role in the cascade process. These small scale fluctuations have been interpreted as the signature of localized coherent structures spontaneously produced by the nonlinear dynamics. The phenomenon described above is known as *intermittency* in fully developed turbulence.

Another important aspect of intermittency is the occurrence of impulsive bursts of energy dissipation which are observed in different systems characterized by the presence of MHD turbulence, e. g. the magnetic loops of the solar corona [1], and laboratory plasma devices such as reversed field pinches [5] and tokamaks [56]. This phenomenon is often referred to as *temporal intermittency*, to distinguish it from the *spatial intermittency* described above, as it consists in the observation of strong bursts of activity in the time series of a quantity which traces the energy dissipation in the system (e. g. radiation intensity). These bursts are separated from each other by intervals of low activity, denoted as laminar times or waiting times.

While temporal intermittency occurring in turbulent systems has been described, until the end of 90s, in the framework of theoretical paradigms other than turbulence (e. g. Self organized criticality [2]), the idea that spatially and temporal intermittency in fully developed turbulence are both due to the underlying nonlinear dynamics of the energy cascade, giving rise to small scale coherent structures, has now gained considerable standing. In the MHD turbulence context this picture has been supported by a number of results based on the analysis of solar wind and solar corona observations and laboratory plasma experimental data.

The intermittent, coherent events can be viewed as localized zones of fluid where phase correlations exist. These structures, which dominate the statistics of small scales, occur as isolated events. They continuously appear and disappear, apparently in a random fashion, at some locations in the fluid, and they carry most of the flow energy. Among the different techniques which can be used to identify such structures within a turbulent signal, a very effective one is based on the use of wavelet transforms [22]. The wavelet transform of a real square integrable signal $f(x)$ (where x is usually the time or a spatial coordinate) is de-

fined as

$$W(x,r) = C_\psi^{-1/2} r^{-1/2} \int_{-\infty}^{\infty} \psi\left(\frac{x'-x}{r}\right) f(x')\mathrm{d}x' \,, \quad (24)$$

where r is a scale dilation, x is a position (time) translation, $\psi(x)$ is the so called mother wavelet function, and C_ψ is a normalization constant, which must satisfy the admissibility condition $\int_{-\infty}^{+\infty} |k|^{-1}|\hat{\psi}(k)|^2 \mathrm{d}k < \infty$ where $\hat{\psi}(k)$ is the Fourier transform of $\psi(x)$. The wavelet coefficients $W(x,r)$ give a decomposition of $f(x)$ at the scale r as a function of the position x. For each scale, it is possible to identify the position (or the time occurrence) of strong, intermittent events through a method proposed by Farge [22] based on the so-called Local Intermittency Measure (LIM) $l(x,r)$ defined as

$$l(x,r) = \frac{|W(x,r)|^2}{\langle |W(x,r)|^2 \rangle} \,. \quad (25)$$

The identification method of intermittent structures is based on the idea that large values of $l(x,r)$ represent a signature of large fluctuations with respect to the background level. The wavelet coefficients can thus be classified as "passive" if $l(x,r) < F$ or "intermittent" when $l(x,r) > F$, where F is a threshold which can be chosen by using different suitable criteria [9,41,54].

Another possible method to select intermittent bursts [6] consists of defining bursts as the time intervals during which the condition $f(x) \geq f_{\mathrm{th}}$ is satisfied, where the threshold value f_{th} is calculated through the following iterative method. One starts by defining $f_{\mathrm{th}} = \langle f(t) \rangle + a\sigma$ (a is an arbitrary positive number, usually 2 or 3), where the average and the standard deviation are computed from the whole time series. After excluding the values which exceed f_{th}, the average and the standard deviation are calculated again using the remaining part of the time series. This process is repeated until convergence of f_{th} is reached. This procedure allows one to remove strong events and to evaluate the threshold taking into account only the background contribution.

Once intermittent structures are identified, it becomes possible to study their typical profiles and some relevant statistical properties such as the distributions of event energy and of time intervals between successive bursts. Such investigations allow one to shed light on the physical mechanisms underlying the bursting process. We discuss now the properties of intermittent events observed in three different systems where MHD turbulence occurs, namely the solar wind, reversed field pinches devices, and magnetic loops of the solar corona. Related aspects of intermittency and multifractality are discussed in the review by Chang ► Space Plasmas, Dynamical Complexity in in this Encyclopedia.

Intermittent Structures in Solar Wind MHD Turbulence

Intermittent events in solar wind turbulence have been studied by Veltri and Mangeney [54] and Veltri et al. [55] by applying the LIM technique, with the Haar wavelet basis, to fluid velocity and magnetic field measurements performed during about 1 year in the space experiment ISEE. In this experiment only two components of the fluid velocity, namely, V_x and V_y, were measured, together with all the three components of the magnetic field. The reference frame used was the standard GSE frame and the sample was formed by data at a time resolution of $T = 1$ min, so that the sampling rate was $\Delta = V_{\mathrm{sw}} T \sim 24.000$ km (V_{sw} is the average solar wind velocity).

The classification of wavelet coefficients allows for an identification and a study of the most intermittent events in solar wind turbulence, which occur in those positions where the amplitude of the wavelet coefficients displays the largest values compared to the average. These events, which occur on time scale of the order of few minutes, exhibit a small number of typical profiles, summarized as follows:

a) "Tangential discontinuities": These structures are almost incompressible, pressure balanced one dimensional current sheets. A minimum variance analysis performed on the magnetic field around the singularity shows (Fig. 7) that the component of the magnetic field which varies most changes sign, and this component is perpendicular to the average magnetic field (the magnetic field component along the third axis being almost zero). The magnetic field rotates then in a plane by an angle which is about 120°–130°. There is one more interesting property: When these structures occur during an Alfvénic period (velocity and magnetic field fluctuations highly correlated), the Alfvénic correlations go from 1 to zero during the traversal of the current sheet (Fig. 7, left panel), when, on the contrary, these structures occur during a period of almost no Alfvénic correlation, the correlation increases to about 1 at the current sheet location (Fig. 7, right panel).
b) "Compressive discontinuities": These structures can be either parallel shocks, mainly observed on the radial component of the velocity field, but clearly seen also on the magnetic field intensity, proton temperature and density (Fig. 8, left panel); or slow mode wavetrains, characterized by a value of $\beta \sim 0$, a constantpressure,

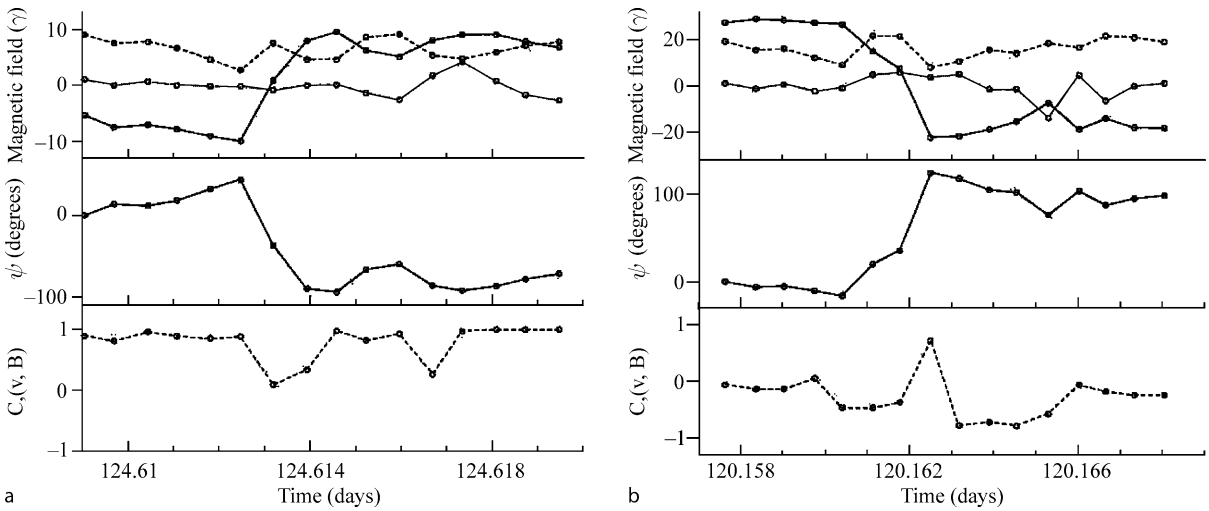

Self-Organization in Magnetohydrodynamic Turbulence, Figure 7

Current sheet intermittent events in solar wind MHD turbulence: The three components of the magnetic field obtained through a minimum variance analysis (*upper panels*); the angle of rotation of the magnetic field in the plane perpendicular to the minimum variance direction (*middle panels*); the coefficient of correlation between velocity and magnetic field fluctuations (*lower panels*)

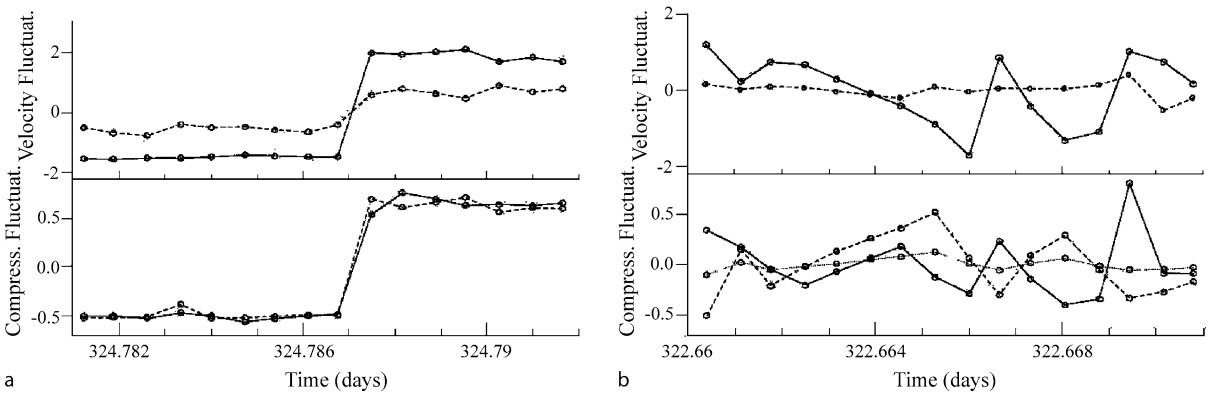

Self-Organization in Magnetohydrodynamic Turbulence, Figure 8

Intermittent events associated with compressive discontinuities in the solar wind. a A parallel shock intermittent event in solar wind, V_x (*full line*) and V_y (*dashed line*) fluctuations normalized to the local average sound velocity (*upper panel*); proton density (*full line*) and sound velocity (*dashed line*) fluctuations normalized to their local average values (*lower panel*). b A slow shock intermittent event in solar wind, velocity fluctuations parallel (*full line*) and perpendicular (*dashed line*) to the local average magnetic field normalized to the local average sound velocity (*upper panel*); proton density (*full line*), sound velocity (*dashed line*) and total pressure (*dotted line*) fluctuations normalized to their local average values (*lower panel*)

anticorrelated density and proton temperature fluctuations and with velocity fluctuations along the average magnetic field (Fig. 8, right panel).

Very interesting statistics can be studied on the time separation Δt (often denoted as *waiting time*) between the occurrence of two consecutive structures. Waiting times of solar wind intermittent structures are distributed according to a power law $P(\Delta t) \sim \Delta t^{-\gamma}$ extended over two decades at least, with an exponent $\gamma \simeq 2.18$ (Fig. 9) [14].

This property is very interesting, because it indicates that the energy cascade process is non-Poissonian. Waiting times occurring between isolated Poissonian events must be distributed according to an exponential function. The power law represents the asymptotic behavior of a Lévy function, which describes self-affine processes and is obtained from the central limit theorem by relaxing the hypothesis that the variance of variables is finite. The power law waiting time PDF is thus a clear evidence that long-range correlations (or in other words "memory") exist in the underlying cascade process.

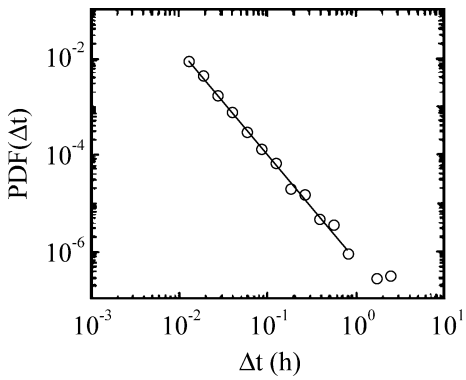

Self-Organization in Magnetohydrodynamic Turbulence, Figure 9

Probability density functions of waiting times between consecutive intermittent structures in solar wind MHD turbulence identified by applying the LIM technique on magnetic field measurements acquired by the HELIOS II spacecraft. The *solid line* represents a power law with an exponent $\gamma = 2.18$

Intermittent Structures in Laboratory MHD Turbulence

The properties of intermittent events in the magnetic turbulence observed at the edge of a toroidal plasma device in reversed field pinch configuration has been studied by Carbone et al. [14] through the analysis of magnetic field fluctuations in the Reversed Field Pinch experiment RFX [48]. Events have been identified by applying the LIM technique on radial and toroidal magnetic field time series. The magnetic field fluctuations for some of the most intermittent events are shown in Fig. 10 by using the minimum variance reference frame.

One of the two components displays a peak when the other component changes sign. This behavior suggests that the intermittent structures identified in the RFX magnetic field are tangential discontinuities similar to those found in the solar wind, although measurements of all the magnetic field components would be needed to confirm this interpretation.

Also in RFP magnetic turbulence, the waiting time PDF of intermittent structures shows a power law behavior $P(\Delta t) \sim \Delta t^{-\gamma}$ extended over two decades as in the case of solar wind MHD turbulence, with an exponent $\beta \simeq 1.5$ (Fig. 11) [14].

Intermittent Energy Release Events in the Solar Corona

Another manifestation of intermittency phenomena in systems where MHD turbulence is present is the occurrence of impulsive energy release events, the so-called solar flares, which take place in the magnetic loops of the solar corona. Energy is released mainly through accelerated particles and emission of electromagnetic radiation in a wide range of wavelengths, from radio to γ-rays, and can vary between 10^{24} and 10^{33} erg. The smallest events, namely those between 10^{24} and 10^{27}, are usually denoted as nanoflares and microflares.

Parker [43] conjectured that nanoflares are produced by the dissipation of small current sheets, associated with tangential discontinuities, forming as a consequence of the continuous shuffling and intermixing of field footpoints in the photospheric convection. In the Parker's picture, flares of all sizes result from the superposition of small dissipation events (nanoflares) which can trigger energy releases in the neighboring discontinuities, originating a fragmented energy release process which can produce both small and large-scale events, depending on the details of the magnetic configuration.

Starting from the idea that MHD turbulence is most probably a fundamental ingredient of coronal loop (Fig. 12) dynamics, it has been proposed [6] that nanoflares and flares can be identified with dissipation events of small-scale current sheets forming as a consequence of the nonlinear turbulent cascade which occurs inside coronal magnetic structures. The cascade would be driven by the energy input due to photospheric footpoint motions. In this framework, current sheets are coherent intermittent small scale structures of MHD turbulence. The energy injected at large scales by photospheric motions is transferred to small scales through the nonlinear cascade, which goes down to the dissipative scales. The intermittent nature of the energy release process is thus associated with the intermittency of energy dissipation in MHD turbulence.

This picture is supported by the statistical properties of solar flares inferred from the analysis of the associated radiation bursts. Nanoflares are detected as extreme ultraviolet (EUV) brightenings, while larger flares are observed through several types of emission (radio, microwaves, $H\alpha$, UV, X-rays) and most often studied analyzing Soft X-ray (SXR) and Hard X-ray (HXR) flare bursts. Probability distributions of the relevant quantities (peak flux, total energy, duration) characterizing the flare bursts have been found to be well represented by power laws $P(x) = Ax^{-\alpha}$ (see Fig. 13 for an example).

Using various SXR and HXR observations it has been found that, for energy and peak flux distributions, $\alpha \simeq 1.6$–2 (see e. g. [15,16,34]). For the EUV nanoflare brightenings the power law exponent is much more uncertain, $\alpha = 1.3$–2.8 (see e. g. [32,44]). The waiting time distribution (WTD) has been studied both for HXR and SXR bursts [6,45,57]. SXR observations acquired by the GOES

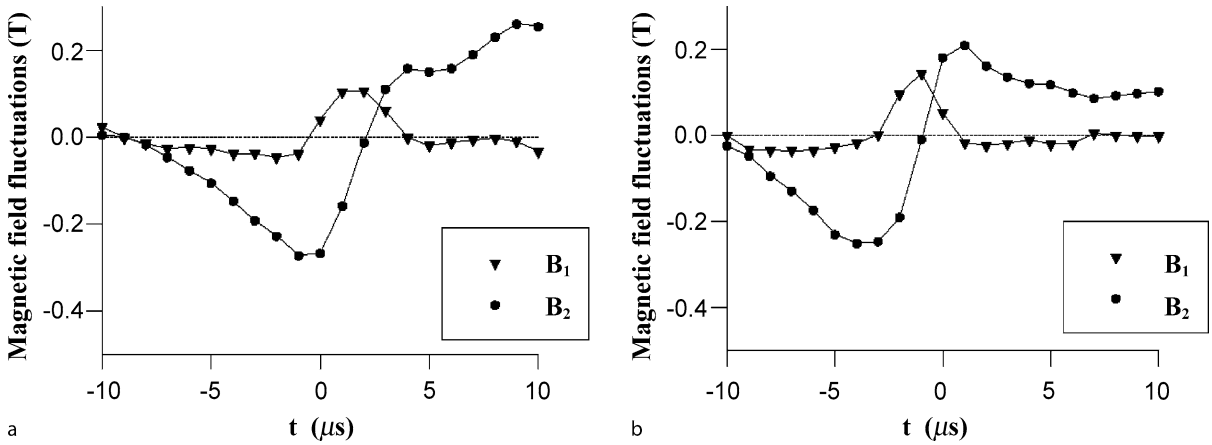

Self-Organization in Magnetohydrodynamic Turbulence, Figure 10
Two components of magnetic field fluctuations (denoted as B_1 and B_2), in the minimum variance reference frame, of two intermittent events observed in the RFX machine

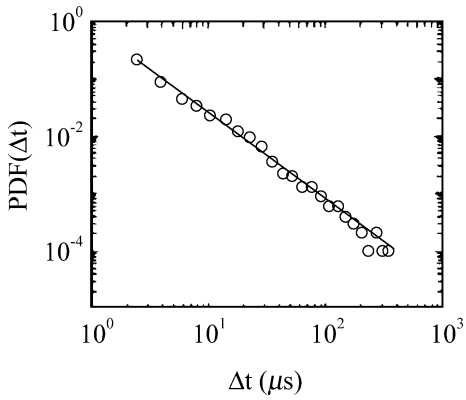

Self-Organization in Magnetohydrodynamic Turbulence, Figure 11
Probability density functions of waiting times between consecutive intermittent structures in RFP magnetic turbulence identified by applying the LIM technique on magnetic field measurements acquired at the RFX machine. The *solid line* represents a power law with an exponent $\gamma = 1.5$

Self-Organization in Magnetohydrodynamic Turbulence, Figure 12
An image of hot coronal loops which span 30 or more times the diameter of planet Earth obtained from Transition Region and Coronal Explore (TRACE) satellite. The image was taken in the Fe IX 171 Å emission line. (Transition Region and Coronal Explorer, TRACE, is a mission of the Stanford–Lockheed Institute for Space Research, and part of the NASA Small Explorer program)

satellites have the advantage to provide a long sequence of bursts (from 1975 to today) with few gaps and allow, thus, to analyze the WTD with a much better statistical accuracy than HXR data. The WTD of GOES SXR flares has been shown to display a clear power law tail $P(\Delta t) \sim \Delta t^{-\gamma}$ in the range $5\,\mathrm{h} \lesssim \Delta t \lesssim 100\,\mathrm{h}$, with an exponent $\gamma \simeq 2.4$ (Fig. 14) [6].

The power law behavior of the solar flare WTD represents an indication of the existence of correlations between successive bursts [6,33]. These correlations can be related to the nonlinear dynamics of the MHD turbulent energy cascade which generates intermittent bursts of chaoticity at small scales.

From the theoretical point of view, describing in a proper way the intermittency of the turbulent energy

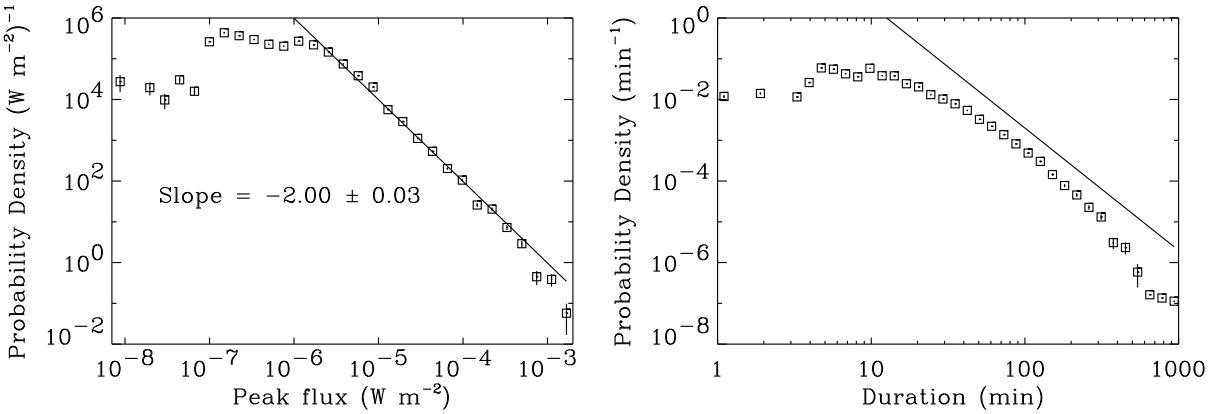

Self-Organization in Magnetohydrodynamic Turbulence, Figure 13

Probability density functions of peak flux (*left panel*) and duration (*right panel*) of Soft *X*-ray solar flare bursts detected by the Geostationary Operational Environmental Satellites (GOES). The *solid line* in the *left panel* represents a power law with an exponent $\alpha = 2$ obtained from a least squares fit. The *solid line* in the *right panel*, shown as a comparison, represents a power law with an exponent $\alpha = 3$

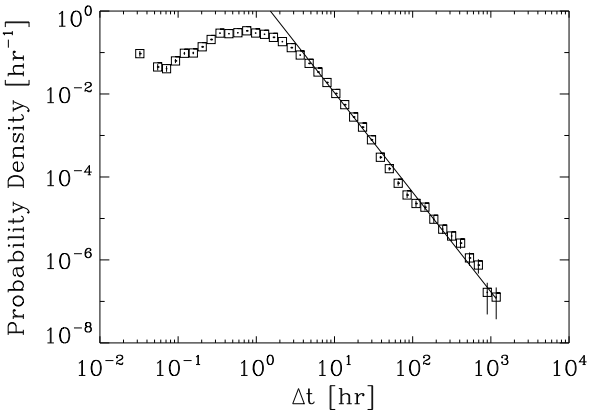

Self-Organization in Magnetohydrodynamic Turbulence, Figure 14

Probability density function of waiting times for SXR solar flare bursts detected by the GOES satellites in the interval from September 1975 to December 2001. The solid line represents a power law with an exponent $\alpha = 2.4$ as found by Boffetta et al. [6]

dissipation process is one of the basic ingredients for the study of intermittent events occurring in the solar corona, as well as in other astrophysical systems such as the solar wind and accretion disks. These astrophysical plasmas are characterized by huge Reynolds numbers. The number $N_{\rm df}$ of degrees of freedom, i.e the number of grid points necessary to resolve the entire inertial range of turbulence in a direct numerical simulation of Navier–Stokes or MHD equations, grows with the Reynolds number Re as $N_{\rm df} \sim {\rm Re}^{9/4}$. Therefore performing direct numerical simulations at Reynolds number regimes of interest for space plasmas is out of the present computational possibilities. The so called shell models, a class of dynamical deterministic models of turbulence in which $N_{\rm df}$ grows logarithmically with Re, can represent an extremely helpful tool for the modeling of such physical systems, as these models are able to simulate the turbulent cascade and the related intermittency of the energy dissipation process in Reynolds number regimes which are not far from the real ones (at least with respect to direct numerical simulations). The statistical properties of intermittent events in MHD turbulence can thus be effectively investigated through the use of shell models [6,14,15].

Shell models [7] are dynamical systems designed in order to represent in a simplified way the spectral Navier–Stokes or MHD equations for turbulence. They were originally proposed by Obukhov [40], Desnyansky and Novikov [19], and Gledzer [26] in hydrodynamic turbulence.

In order to build up the evolution equations for a shell model, the wavevector space is divided into a discrete number of shells of radius $k_n = \rho^n k_0$, with $\rho > 1$, $n = 1, 2, \ldots, N$. Each shell is assigned a scalar dynamic variable (real or complex) $v_n(t)$ which represents the averaged effect of velocity modes with wavenumber between k_n and k_{n+1}. $v_n(t)$ can also be regarded as the velocity increments $|v(x+\ell) - v(x)|$ on an eddy of scale $\ell \sim k_n^{-1}$. For MHD shell models, besides $v_n(t)$, another variable $b_n(t)$, representing magnetic field modes between k_n and k_{n+1} (or, alternatively, magnetic field increments $|b(x+\ell) - b(x)|$), is assigned to each shell. The nonlinear terms of the equations are quadratic combinations of the

dynamic variables and are written under the assumption that the interactions among shells are local in k-space. This means that only nearest and, at most, next-nearest shells are introduced in the non-linear terms. The coupling coefficients are found by imposing conservation of the ideal invariants of Navier–Stokes or MHD equations.

The main advantage of shell models is that they can be studied through numerical simulations at very high Reynolds numbers. Moreover, they provide the possibility to obtain long time series with a fairly small computational effort, allowing thus a robust investigation of the statistical properties of intermittent events The weak point is that they are scalar models, that is, any information about spatial structures, such as vortices, sheets, and filaments, is lost.

Among the various shell models which have been proposed in the literature, we consider here the so called GOY (Gledzer–Ohkitani–Yamada) model, which has been used by several authors both in hydrodynamics and MHD. In the MHD context, the advantage of the GOY model with respect to previous shell models, where only the conservation of two quadratic invariants, i. e. total energy and cross helicity, can be imposed, is that it allows to conserve also magnetic helicity. The GOY model involves nearest and next-nearest interactions and was originally introduced in the framework of hydrodynamic turbulence by Gledzer [26], using real shell variables. The model was extended by Yamada and Ohkitani [60] to the case of complex variables. The MHD generalization of the GOY model has been considered by several authors (see [6,27,28] and references therein). The evolution equations for the dynamic variables $v_n(t)$ and $b_n(t)$ in the GOY shell model can be written as [6]

$$\frac{\mathrm{d}v_n}{\mathrm{d}t} = -\nu k_n^2 v_n + f_n + ik_n \Big\{ (v_{n+1}v_{n+2} - b_{n+1}b_{n+2}) - \frac{1}{4}(v_{n-1}v_{n+1} - b_{n-1}b_{n+1}) - \frac{1}{8}(v_{n-2}v_{n-1} - b_{n-2}b_{n-1}) \Big\}^* , \quad (26)$$

$$\frac{\mathrm{d}b_n}{\mathrm{d}t} = -\mu k_n^2 b_n + g_n + ik_n \frac{1}{6} \Big\{ (v_{n+1}b_{n+2} - b_{n+1}v_{n+2}) + (v_{n-1}b_{n+1} - b_{n-1}v_{n+1}) + (v_{n-2}b_{n-1} - b_{n-2}v_{n-1}) \Big\}^* . \quad (27)$$

f_n and g_n represent external forcing terms, usually acting on low wavenumber shells, while ν and μ appearing in the linear, dissipative terms, are the kinematic viscosity and the resistivity respectively, viscous and resistive terms provide a mechanism for energy dissipation at small scales. In turbulent space plasmas such as solar corona and solar wind dissipation mechanisms are different, involving kinetic plasma processes. Nevertheless, the properties of low frequency MHD turbulent cascade in the inertial range and of the related intermittency do not depend, in general, on the details of the dissipation mechanism, hence the use of viscous and resistive dissipation in the standard incompressible form does not represent a serious limitation in this context.

The GOY model described here has been designed to investigate 3D MHD turbulence. This means that it satisfies the conservation of the ideal 3D MHD quadratic invariants (4), (5) and (6). In terms of shell variables, the invariants read [27,28]

$$E = \frac{1}{2} \sum_{n=1}^{N} (|v_n|^2 + |b_n|^2) , \quad (28)$$

$$H_\mathrm{c} = \frac{1}{4} \sum_{n=1}^{N} \Re(v_n b_n^*) , \quad (29)$$

$$H_\mathrm{m} = \sum_{n=1}^{N} (-1)^n \frac{|b_n|^2}{k_n} . \quad (30)$$

The energy dissipation rate in the GOY MHD shell model can be defined as [6]

$$\varepsilon(t) = \nu \sum_{n=1}^{N} k_n^2 |v_n|^2 + \mu \sum_{n=1}^{N} k_n^2 |b_n|^2 . \quad (31)$$

A statistical analysis of intermittent events in the GOY MHD shell model shows that the PDFs of event energies, duration and waiting times display power law tails (Figs. 15 and 16) [6,14]. This behavior is consistent with the statistics of intermittent events observed in the solar wind, reversed field pinch devices and solar corona.

In closing this section, it can thus be remarked that the study of intermittent events in MHD turbulence, mainly based on wavelet transforms and PDFs of various event parameters, has provided a quite detailed picture of the coherent structures produced by the self-organization of the MHD turbulence cascade at small scales. The analysis of data from solar wind and RFP devices has demonstrated that the most intermittent structures in incompressible MHD are one dimensional current sheets associated with tangential discontinuities. The statistical analysis of intermittent events occurring in solar wind, laboratory plasma and magnetic structures of the solar corona has shown that these events are characterized by power law distributions of event energy and waiting times, which are naturally reproduced by MHD turbulence models. The power law behavior of the waiting time distribution indicates the

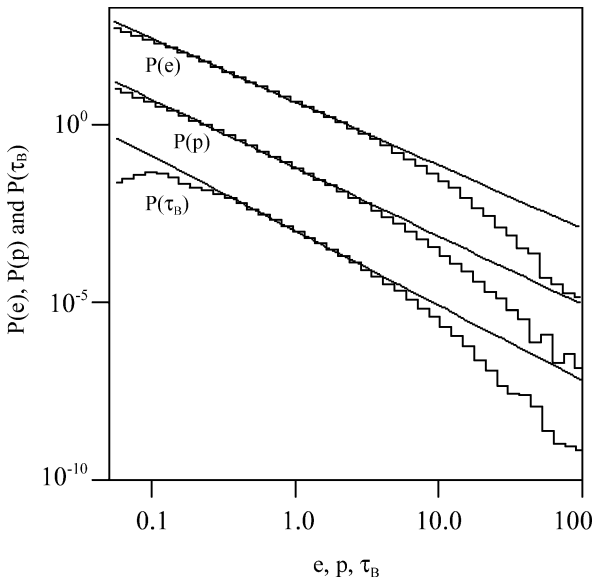

Self-Organization in Magnetohydrodynamic Turbulence, Figure 15

Total energy distribution $P(e)$, peak energy distribution $P(p)$ and burst duration distribution $P(\tau_B)$ for the GOY MHD shell model. The values of $P(e)$ and $P(\tau_B)$ are offset by a factor 10^2 and 10^{-2} respectively. The straight lines represent power law fits, with exponents $\alpha_e = 1.8$, $\alpha_p = 2.05$ and $\alpha_{\tau_B} = 2.2$ respectively

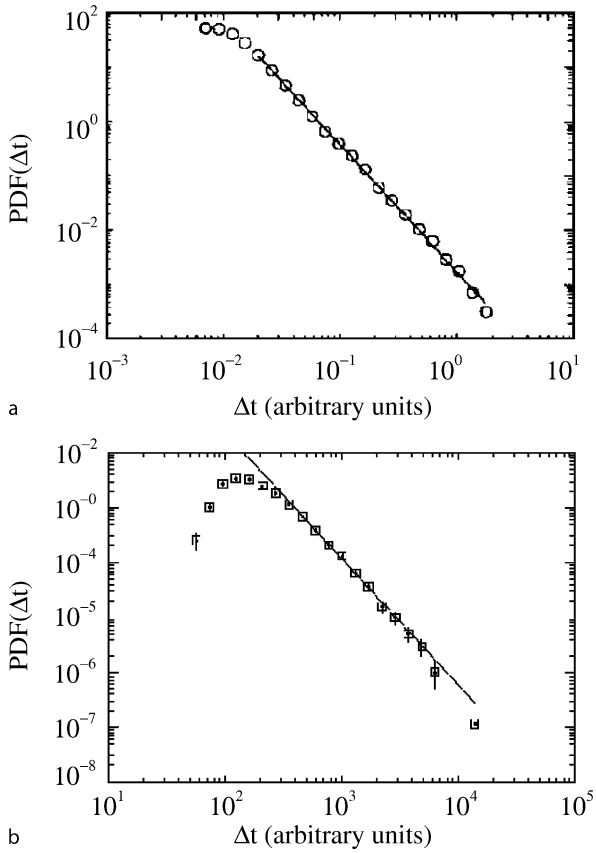

Self-Organization in Magnetohydrodynamic Turbulence, Figure 16

Probability density function of waiting times between successive intermittent events in the GOY MHD shell model. a Waiting time PDF for intermittent events identified in the time series of the magnetic field variables at the largest wavenumber shell. The *solid line* represents a power law with exponent $\beta = 2.13$. b Waiting time PDF for intermittent events identified in the time series of the energy dissipation rate ε (Eq. 31). The solid line represents a power law with exponent $\beta = 2.33$. In both cases, intermittent events were identified through the iterative procedure proposed in [6] and described in the text, using the threshold $\langle\cdot\rangle + 2\sigma$

presence of long range time correlations which can be attributed to the non-linear dynamics of the turbulent energy cascade.

Future Directions

In the previous sections we have discussed how coherent dissipative structures are spontaneously formed by non-linear interactions both in laboratory and in natural plasmas. A further understanding of these processes requires an increase in the performance of computers, which will allow one to simulate 3D MHD turbulence at Reynolds numbers larger than 10^3, which represents the actual limit. On the other hand, an improved effort in the analysis of solar wind space data and laboratory plasma devices measurements should also be necessary. An increased knowledge of these phenomena is crucial in order to try to control the transport processes in fusion plasmas and then to improve the plasma confinement, since the presence of coherent structures inside turbulence greatly affect the plasma diffusion process. Moreover a better comprehension of the self organization in MHD turbulence can greatly help in explaining some basic phenomena in solar corona (flares and coronal heating), which occur in a region where in situ observations cannot be performed.

Bibliography

Primary Literature

1. Aschwanden MJ (2004) Physics of the solar corona: An introduction. Springer, Berlin
2. Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality: An explanation of 1/*f* noise. Phys Rev Lett 59:381–384
3. Belcher JW, Davis L (1971) Large-amplitude Alfvèn waves in the interplanetary medium-II. J Geophys Res 76:3534–3563
4. Bellan PM (2000) Spheromaks: A practical application of magnetohydrodynamic dynamos and plasma self-organization. Imperial College Press, London

5. Biskamp D (1993) Non-linear magnetohydrodynamics. Cambridge University Press, Cambridge
6. Boffetta G, Carbone V, Giuliani P, Veltri P, Vulpiani A (1999) Power laws in solar flares: Self-organized criticality or turbulence? Phys Rev Lett 83:4662–4665
7. Bohr T, Jensen MH, Paladin G, Vulpiani A (1998) Dynamical systems approach to turbulence. Cambridge University Press, Cambridge
8. Burlaga LF (1991) Multifractal structure of the interplanetary magnetic field Voyager 2 observations near 25 AU, 1987 1988. Geophys Res Lett 18:69–72
9. Camussi R, Guj G (1997) Orthonormal wavelet decomposition of turbulent flows: Intermittency and coherent structures. J Fluid Mech 348:177–199
10. Carbone V, Veltri P (1987) A simplified cascade model for M.H.D. turbulence. Astron Astrophys 188:239–250
11. Carbone V, Veltri P (1992) Relaxation processes in MHD: A triad-interaction model. Astron Astrophys 259:359–372
12. Carbone V (1994) Scaling exponents of the velocity structure functions in the interplanetary medium. Ann Geophys 12:585–590
13. Carbone V, Sorriso-Valvo L, Martines E, Antoni V, Veltri P (2000) Intermittency and turbulence in a magnetically confined fusion plasma. Phys Rev E 62:49–56
14. Carbone V, Cavazzana R, Antoni V, Sorriso-Valvo L, Spada E, Regnoli G, Giuliani P, Vianello N, Lepreti F, Bruno R, Martines E, Veltri P (2002) To what extent can dynamical models describe statistical features of turbulent flows? Europhys Lett 58:349–355
15. Carbone V, Lepreti F, Sorriso-Valvo L, Veltri P, Antoni V, Bruno R (2004) Scaling laws in plasma turbulence. Riv Nuovo Cim 27(8-9):1–108
16. Crosby NB, Aschwanden MJ, Dennis BR (1993) Frequency distributions and correlations of solar *X*-ray flare parameters. Solar Phys 143:275–299
17. Dahlburg JP, Montgomery D, Doolen GD, Turner R (1986) Turbulent relaxation to a force-free field-reversed state. Phys Rev Lett 57:428–431
18. De Bartolo R, Carbone V (2006) The role of the basic three-modes interaction during the free decay of magnetohydrodynamic turbulence. Europhys Lett 73:547–552
19. Desnyansky VN, Novikov EA (1974) Evolution of turbulence spectra to self-similar regime. Atmos Oceanic Phys 10:127–136
20. Dobrowolny M, Mangeney A, Veltri P (1980) Fully developed anisotropic hydromagnetic turbulence in interplanetary space. Phys Rev Lett 45:144–147
21. Elsässer WM (1950) The hydromagnetic equations. Phys Rev 79:183–183
22. Farge M (1992) Wavelet transforms and their applications to turbulence. Ann Rev Fluid Mech 24:395–457
23. Frisch U, Pouquet A, Leorat J, Mazure A (1975) Possibility of an inverse cascade of magnetic helicity in magnetohydrodynamic turbulence. J Fluid Mech 68:769–778
24. Fyfe D, Montgomery D (1976) High-beta turbulence in two-dimensional magnetohydrodynamics. J Plasma Phys 16: 181–191
25. Heyvaerts J, Priest ER (1984) Coronal heating by reconnection in DC current systems – A theory based on Taylor's hypothesis. Astronom Astrophys 137:63–78
26. Gledzer EB (1973) System of hydrodynamic type admitting two quadratic integrals of motion. Sov Phys Dokl 18:216–217
27. Giuliani P, Carbone V (1998) A note on shell models for MHD turbulence. Europhys Lett 43:527–532
28. Giuliani P (1999) Non-linear MHD waves and turbulence, edited by Passot T, Sulem PL. Lect Notes Phys 536:331–345
29. Iroshnikov PS (1963) Turbulence of a conducting fluid in a strong magnetic field. Astron Zhur 40:742–750 (in Russian, translated in Sov Astron 7:566–571)
30. Kolmogorov AN (1941) The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers. Dokl Akad Nauk SSSR 30:299–303 (in Russian translated in Proc Royal Soc London 434:9–13)
31. Kraichnan RH (1965) Inertial range spectrum of hydromagnetic turbulence. Phys Fluids 8:1385–1387
32. Krucker S, Benz AO (1998) Energy Distribution of Heating Processes in the Quiet Solar Corona. Astrophys J 501:L213–L216
33. Lepreti F, Carbone V, Veltri P (2001) Solar flare waiting time distribution: Varying rate Poisson or Lévy function? Astrophys J 555:L133–136
34. Lin RP, Schwartz RA, Kane SR, Pelling RM, Hurley KC (1984) Solar hard *X*-ray microflares. Astrophys J 283:421–425
35. Marsch E, Liu S (1993) Structure functions and intermittency of velocity fluctuations in the inner solar wind. Ann Geophys 11:227–238
36. Marsch E, Tu CY (1994) Non-Gaussian probability distributions of solar wind fluctuations. Ann Geophys 12:1127–1138
37. Tu CY, Marsch E (1995) MHD structures, waves and turbulence in the solar wind: Observations and theories. Space Sci Rev 73:1–210
38. Matthaeus WH, Montgomery D (1980) Selective decay hypothesis at high mechanical and magnetic Reynolds numbers. Ann NY Acad Sci 357:203–222
39. Matthaeus WH, Goldstein ML, Montgomery D (1982) Turbulent generation of outward-traveling interplanetary alfvnic fluctuations. Phys Rev Lett 51:1484–1487
40. Obukhov AM (1971) Some general characteristic equations of the dynamics of the atmosphere. Atmos Oceanic Phys 7:41–45
41. Onorato M, Camussi R, Iuso G (2000) Small scale intermittency and bursting in a turbulent channel flow. Phys Rev E 61:1447–1454
42. Parker EN (1983) Magnetic neutral sheets in evolving fields. II. Formation of the solar corona. Astrophys J 264:642–647
43. Parker EN (1988) Nanoflares and the solar *X*-ray corona. Astrophys J 330:474–479
44. Parnell CE, Jupp PE (2000) Statistical Analysis of the Energy Distribution of Nanoflares in the Quiet Sun. Astrophys J 529:554–569
45. Pearce G, Rowe A, Yeung J (1993) A statistical analysis of HRD *X*-ray solar flares. Astrophys Space Sci 208:99–111
46. Pouquet A, Meneguzzi M, Frish U (1986) Growth of correlations in magnetohydrodynamic turbulence. Phys Rev A 33:4266–4276
47. Rosenbluth MN, Monticello DA, Strauss HR (1976) Numerical studies of non-linear evolution of kink modes in tokamaks. Phys Fluids 19:1987–1996
48. Rostagni G (1995) RFX: An expected step in RFP research. Fusion Eng Design 25:301–313
49. Servidio S, Carbone V (2005) Non-linear dynamics of inviscid reduced MHD plasmas: The appearance of quasi-single-helicity states. Phys Rev Lett 95:045001
50. Sorriso-Valvo L, Carbone V, Veltri P, Consolini G, Bruno R (1999) Intermittency in the solar wind turbulence through proba-

bility distribution functions of fluctuations. Geophys Res Lett 26:1801–1804
51. Taylor JB (1974) Relaxation of toroidal plasma and generation of reverse magnetic fields. Phys Rev Lett 33:1139–1141
52. Taylor J (1986) Relaxation and magnetic reconnection in plasmas. Rev Mod Phys 58:741–763
53. Ting AC, Matthaeus WH, Montgomery D (1986) Turbulent relaxation processes in magnetohydrodynamics. Phys Fluids 29:3261–3274
54. Veltri P, Mangeney A (1999) Scaling laws and intermittent structures in solar wind MHD turbulence. In: Habbal SR, Esser R, Hollweg JV, Isenberg PA (eds) Solar wind nine, Proceedings of the ninth international solar wind conference, AIP Conference Proceedings vol 471 pp 543–546
55. Veltri P, Nigro G, Malara F, Carbone V, Mangeney A (2005) Intermittency in MHD turbulence and coronal nanoflares modelling. Non-lin Proc Geophys 12:245–255
56. Wesson J (2003) Tokamaks. Oxford Science, Oxford
57. Wheatland MS, Sturrock PA, McTiernan JM (1998) The waiting-time distribution of solar flare hard *X*-Ray bursts. Astrophys J 509:448–455
58. Woltjer L (1958) A theorem on force-free magnetic fields. Proc Natl Acad Sci 44:489–491
59. Yamada M (1999) Study of magnetic helicity and relaxation phenomena in laboratory plasmas. AGU Geophys Monogr 111:129–140
60. Yamada M, Ohkitani K (1987) Lyapunov spectrum of a chaotic model of three dimensional turbulence. J Phys Soc Japan 56:4210–4213
61. Zank GP, Matthaeus WH (1992) The equations of reduced magnetohydrodynamics. J Plasma Phys 48:85–100.

Books and Reviews

Bruno R, Carbone V (2005) The solar wind as a turbulence laboratory. Living Rev Solar Phys. 2: http://www.livingreviews.org/lrsp-2005-4
Biskamp D (2003) Magnetohydrodynamic turbulence. Cambridge University Press, Cambridge
Frisch U (1995) Turbulence: The legacy of A. N. Kolmogorov. Cambridge University Press, Cambridge
Moffat HK (1978) Magnetic field generation in electrically conducting fluids. Cambridge University Press, Cambridge

Self-organized Criticality and Cellular Automata

Michael Creutz
Physics Department,
Brookhaven National Laboratory, Upton, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Self-Organized Criticality
Cellular Automata
The Sandpile Revisited
Future Directions
Acknowledgments
Bibliography

Glossary

Abelian group A mathematical group wherein all the elements commute.

Avalanche A possibly large disturbance induced in a system by a small perturbation.

Cellular automaton This refers to the dynamics of a collection of cells each of which can be in a finite set of states. The evolution is discrete, with the state of a cell at the next time step being dependent only on its previous state and that of its neighbors.

Chaos The tendency of a system of a few degrees of freedom to exhibit highly erratic behavior characterized by an infinite range of time scales.

Self-organized criticality The tendency of certain discrete and dissipative dynamical systems to evolve to a state where changes occur over all possible length scales.

Definition of the Subject

Self-organized criticality is a concept invoked to explain the frequent occurrence of fractal structures and multiscale phenomena in nature. In contrast with the ideas of chaos, here simple common features appear in systems with many degrees of freedom. For modeling this phenomenon, cellular automata provide an elegant class of dynamical systems which are easily simulated numerically.

Introduction

Cellular automata provide a fascinating class of dynamical systems based on very simple rules of evolution yet capable of displaying highly complex behavior. These include simplified models for many phenomena seen in nature. Among other things, they provide insight into self-organized criticality, wherein dissipative systems naturally drive themselves to a critical state with important phenomena occurring over a wide range of length and time scales.

This article begins with an overview of self-organized criticality. This is followed by a discussion of a few examples of simple cellular automaton systems, some of which may exhibit critical behavior. Finally, some of the fascinating exact mathematical properties of the Bak–Tang–Wiesenfeld sandpile model [1] are discussed.

Self-Organized Criticality

Self-organized criticality refers to the tendency of many dynamical systems to naturally drive themselves to a state displaying fluctuations over a wide range of scales [1]. The concept is invoked as a possible "explanation" of the omnipresent multi-scale structures throughout the natural world, ranging from the fractal structure of mountains, to the power law spectra of earthquake sizes [2]. Recent applications include such diverse topics as punctuated evolution [3] and traffic flow [4]. The concept has even been invoked to explain the unpredictable nature of economic systems; i. e. why you can not beat the stock market [5].

Self-organized criticality nicely compliments the concept of chaos. In the latter, dynamical systems with a few degrees of freedom, say as little as three, can display highly complex behavior, often generating beautiful fractal structures. With self-organized criticality, we start instead with systems of many degrees of freedom, and find a few general common features.

Another attractive feature of both self organized criticality and chaos is the ease with which computer models can be implemented and the elegance of the resulting graphics. Most of the figures in this chapter were produced using my publicly available set of programs "xtoys" [6]. Indeed, much of this presentation is based on my similar article in [7].

The paradigm for the phenomenon is the sandpile. On slowly adding grains of sand to an empty table, a pile will grow until its slope becomes critical and avalanches start spilling over the sides. If the slope becomes too large, a large catastrophic avalanche is likely, and the slope will reduce. If the slope is too small, then the sand will accumulate to make the pile steeper. Ultimately one should obtain avalanches of all sizes, with the prediction of the size for the next avalanche being impossible to determine without actually running the experiment.

The original Bak, Tang, Wiesenfeld paper [1] presented a particularly simple model to mimic the sandpile idea. For this, each site of a two dimensional lattice has a state represented by a positive integer z_i. This integer can be thought of as representing the amount of sand at that location, or in another sense it represents the slope of the sandpile at that point. Neither of these analogies is fully accurate, the model has aspects of each.

The dynamics follows by setting a threshold z_T above which any given z_i is unstable. Without loss of generality, I take this threshold to be $z_T = 3$. Time now proceeds in discrete steps. In one such step each unstable site with $z_i \geq 4$ "tumbles" or "topples," dropping by four and adding one grain to each of its four nearest neighbors. This may produce other unstable sites, and thus an avalanche can ensue. This proceeds for further time steps until all sites are stable. Figure 1 shows a typical configuration on a 198 by 198 lattice after lots of random sand addition followed by relaxation. Figure 2 shows an avalanche proceeding on this lattice, and Fig. 3 shows the final avalanch region after the system reaches stability.

A natural experiment consists of adding a grain of sand to a random site and measuring the number of topplings and the number of time steps for the resulting avalanche.

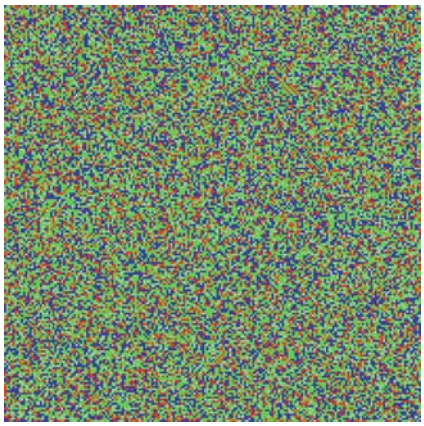

Self-organized Criticality and Cellular Automata, Figure 1
The sandpile model in the final stable state after adding lots of sand to random places. The lattice is 198 cells by 198 cells. The color code is *gray*, *red*, *blue*, and *green* for heights 0, 1, 2, and 3, respectively. Despite the lack of obvious patterns, subtle correlations are present; for example no two adjacent sites have height zero

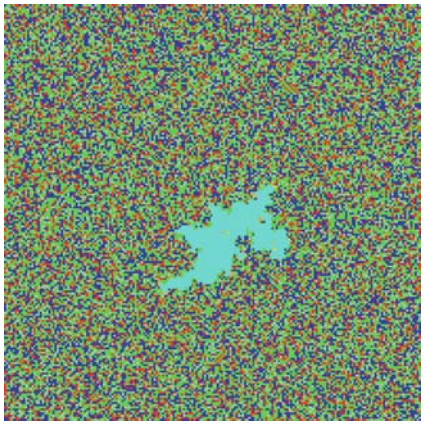

Self-organized Criticality and Cellular Automata, Figure 2
An ongoing avalanche obtained by adding a small amount of sand to the configuration in Fig. 1. Stable sites which have tumbled during the avalanche are distinguished by being colored *light blue*. The still active sites are colored *yellowish brown*

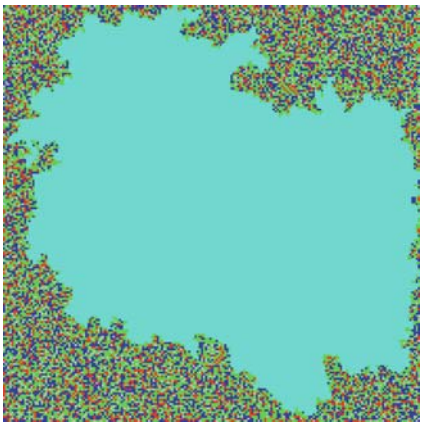

Self-organized Criticality and Cellular Automata, Figure 3
The final state after the avalanch in Fig. 2 has completed. The sites which tumbled during the avalanche are distinguished by being colored *light blue*. Note that the final avalanche region is simply connected. This is a general result proven later in the text

Repeating this many times to gain statistics, the distribution of avalanche sizes and lengths displays a power law behavior, with all sizes appearing. In [8] such experiments showed that the distribution of the number of tumbling events s in an avalanche empirically scales as

$$P(s) \sim s^{-1.07} \tag{1}$$

and the number of time steps τ for avalanches scales as

$$P(\tau) \sim \tau^{-1.14}\,. \tag{2}$$

This model has been extensively studied analytically. While as yet there is no exact calculation of these exponents, a lot is known. In particular, the critical ensemble is well characterized. I will return to these points later.

The extent to which laboratory experiments reproduce these phenomena is somewhat controversial. A study of avalanche dynamics [9] in rice piles showed power laws with long-grain rice, but more ambiguous results followed similar experiments with short-grain rice.

Cellular Automata

The sandpile model is a simple example of a system of cellular automata [10,11]. Each site or "cell" of our lattice follows a prescribed rule evolving in discrete time steps. At each step, the new value for a cell depends only on the current state of itself and its neighbors. These systems are fascinating in that deceptively simple rules can give rise to extremely complex behavior. Furthermore, slight changes in the rules can dramatically change their behavior.

Even though the formulation of a cellular automaton may seem almost trivial, there are a huge number of possible rules. For example, suppose I consider two dimensional models where each cell can take only one of two possible states. These might be referred to as unset or set bits, or more figuratively as "dead" or "alive." Suppose furthermore that I restrict myself to rules where the evolution of a given cell to the next time step depends only on the current values of the cell and each of its eight neighbors. In this case there are $2^9 = 512$ possible arrangements for the cell and its neighbors. A general rule needs to specify the next state of the cell for each of these arrangements. This gives $2^{512} = 1.3 \times 10^{154}$ possible rules. Given that the universe is only of order 4×10^{17} seconds old, clearly only a vanishing fraction of these rules have a chance of being studied in any of our lifetimes.

A simple subset of rules called "totalistic" have the state of the updated cell only depend on the total number of living neighbors. With the eight cell neighborhood, there are nine possible values for this sum, and the new value for the cell requires specification of the new state for each of these as well as the current state of the cell. This gives $2^{18} = 262,144$ rules; still large, but not truly astronomical. If I restrict the rule to depending on the total of only the four nearest neighbors, I then have a modest $2^{10} = 1024$ cases to consider. Other than the sandpile model, most of the following will be restricted to such totalistic rules.

With a discrete set of states, cellular automata have the appealing feature of being easily implementable entirely by logical operations, the natural functions of computer circuitry. Also, the state of several cells can be stored and manipulated within a single computer word. Using such tricks, these models can often be implemented to run extremely fast, leading to hope that such models may supply simulation methods as good as or better than the conventional use of floating point fields on a discrete grid. With this motivation, considerable attention has been paid to cellular automata that may simulate fluid flow. Another advantage of this approach is the ability to work with arbitrary boundary conditions. These topics go beyond the scope of this article. A nice review can be found in [12].

Conway's Life

Perhaps the most famous cellular automaton model is Conway's "Game of Life" [13]. For this there exists a vast literature; so, I will only mention a couple of interesting features. The rule involves the eight cell neighborhood, and if a cell is initially "dead" it becomes alive if and only if it has exactly three live neighbors, or "parents". A living cell dies of loneliness if it has less than two live neighbors,

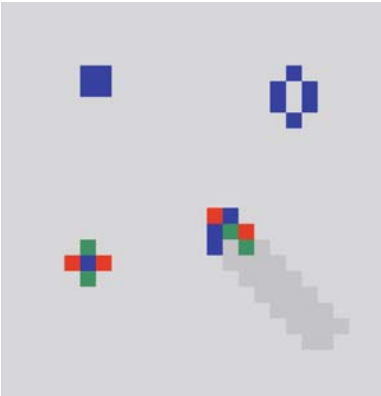

Self-organized Criticality and Cellular Automata, Figure 4
Some living configurations in life. The *top* two are stable patterns. The *lower left* shows a "blinker" or "traffic light" which oscillates with a period of two. On the *lower right* is a glider, which propagates diagonally through the lattice. *Blue* denotes a state that is and just was alive, *red* is newborn, and *green* represents just died. The track of the glider is *darkened* slightly over the remaining *gray background* to show its motion

and of overcrowding if it has more than three live neighbors. Only in the case of exactly two or three live neighbors does it survive.

While simple to state, this model displays fascinating complexity. There are simple isolated sets of live cells that quietly survive, such as a block of four neighboring live cells forming a two by two square. Other configurations oscillate, such as three live cells in a row, which alternate between being vertically and horizontally oriented. A particularly amusing local configuration has five live cells; say starting with coordinates $\{(0,0),(0,1),(0,2),(1,2),(2,1)\}$. After four time steps this configuration returns to its original shape, but displaced by $(-1,1)$. On an otherwise empty board, this "glider" continues to propagate as a single entity. In an on-screen simulation, it appears much as a small insect crawling about. Some elementary configurations are shown in Fig. 4. A large collection of fascinating life configurations can be found in the Wikipedia [14].

Gliders allow information to be propagated over long distances, and it has been proven that with a complicated enough initial configuration, one can construct a computer out of live cells on a life board [13]. Special sub-configurations form the analog of electronic gates, which can control beams of gliders representing bits. Indeed, since life is capable of universal computation, one might imagine a life board programmed to simulate the game of life.

There is some limited evidence that the game of life also displays self-organized criticality [15,16]. One can repeatedly throw down gliders, which collide and create a background of static and oscillating clumps. While oscillators of arbitrarily long period are known to exist, those with period longer than two are extremely rare and almost never created from unorganized initialization. Once the system has settled into a loop, then another glider can be tossed on, giving a disturbance. An avalanche is defined to occur during the period until the system again goes into an oscillating state. Figure 5 shows the effect of such a disturbance. In Fig. 6 I show the distribution of such avalanches as measured on modest lattices. There is a hint of a power law superposed on additional structure from avalanches of only a few time steps, and a rounding at large times possibly due to finite size effects. The criticality of life remains controversial; [17] has looked unsuccessfully for a power

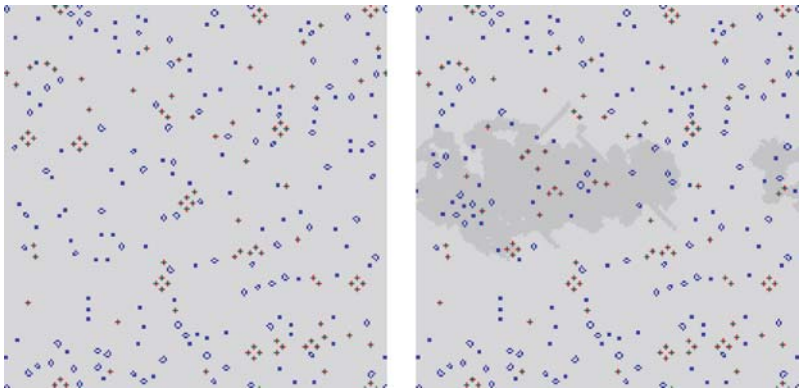

Self-organized Criticality and Cellular Automata, Figure 5
On the *left* is a configuration in life resulting from a random start and evolved until only stable and period two oscillators remain. On the *right* is the state after a small disturbance was introduced in the center and allowed to die out. Note the irregular shape of the disturbed region, which has been tinted a *darker gray*. The lattice here is 198 sites wide by 198 sites high, with periodic boundaries

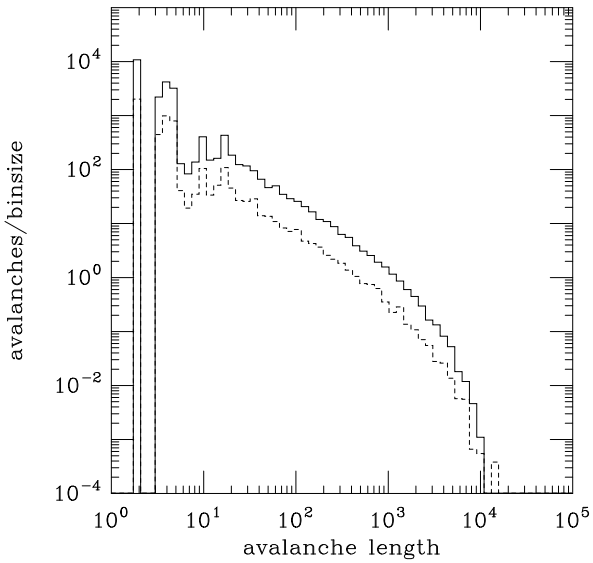

Self-organized Criticality and Cellular Automata, Figure 6
The distribution of avalanches generated by adding gliders randomly to a system in the game of life consisting of stable and period two oscillators. An avalanche occupies the period until the system has relaxed again into such a periodic state. The *solid line* represents 25,000 avalanches on a 512 by 512 lattice, and the *dashed line* is for 6,000 avalanches on a 1024 by 1024 system. This figure is taken from [16]

law distribution of activity as one moves in from a source on the boundary. The relation between these two experiments is unclear.

Fredkin's Modulo-Two Rule

An extremely simple but highly amusing rule takes at each time step the "exclusive or" (XOR) operation between a site and its neighbors. This rule has the remarkable property of self replication [18]. Starting with any given initial pattern, after 2^n time steps copies of the original state occupy positions separated by 2^n spatial sites from the original in every direction as specified in the chosen neighborhood. In Fig. 7 I show an example of this with the four cell neighborhood.

In this rule, the pattern is generally rather complex just before returning to the replicated case, i. e. after $2^n - 1$ steps. Figure 8 shows the pattern obtained from a single set pixel after this rule has been applied for 63 time steps using the four nearest cells as the neighborhood. Note the fractal structure. In one more time step, all but five copies of the original set bit die.

Unlike most cellular automaton rules, this gives a dynamics which in some sense is not really "complex". In most cases the simplest way to predict the evolution of a cellular automaton rule is to actually run it. Here, however, there is an easier way to predict what the final pattern will look like; it is always an XOR operation between several displaced copies of the configuration that appeared 2^n time steps in the past. Despite the lack of complexity, this rule shows rather dramatically that cellular automata are capable of "reproduction".

Reversible Rules

Reversibility is rather elusive among cellular automata. In the game of life, a single isolated cell immediately dies leaving no trace; thus it is impossible from the state at a given time to reconstruct what was there one time step back. A related difficult problem is to construct "garden of Eden" configurations which are impossible to arrive at from any previous state [19].

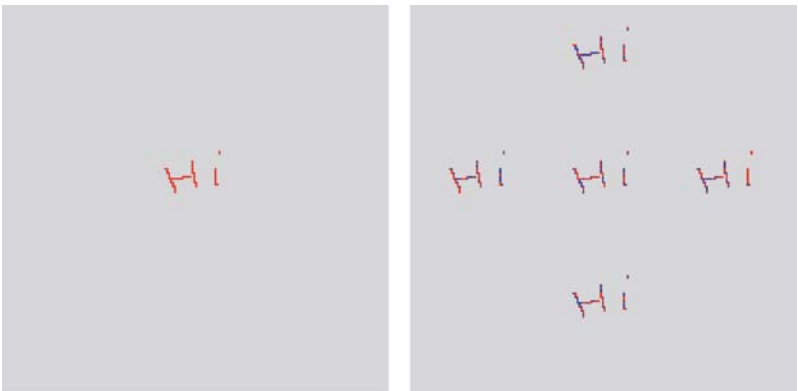

Self-organized Criticality and Cellular Automata, Figure 7
Starting from the initial configuration *on the left*, the modulo two rule is evolved for 64 steps using the four nearest neighbors. At a certain stage, five copies of the original image appear. The *blue pixels* indicate which sites were also alive one step before

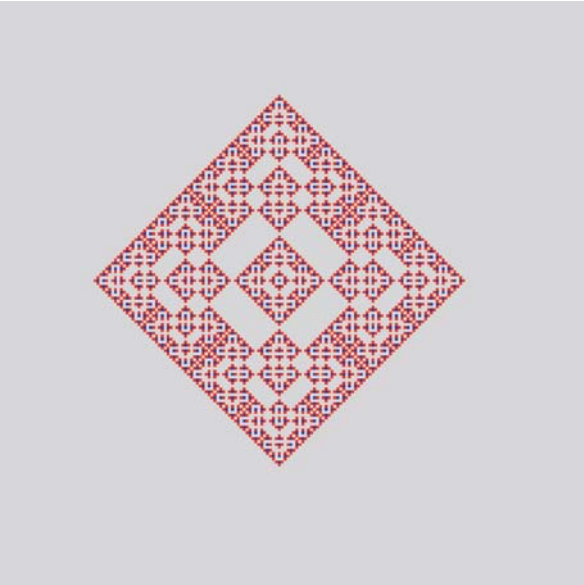

Self-organized Criticality and Cellular Automata, Figure 8
The state after applying 63 steps of the modulo two rule using the four nearest neighbors to an initial state of a single set bit. After the next time step this fractal structure decays into only five remaining live cells

Fredkin pointed out an interesting class of reversible rules based on an analogy with molecular dynamics [11]. In the later one specifies both the position and the velocities of a set of particles and evolves the system under Newton's equations with some given inter-particle force law. Reversal can then be accomplished by merely changing the signs of all the velocities.

In a cellular automaton an analog of velocity requires the value of the cells at two successive time steps. Based on this, Fredkin presented a very simple scheme using the previous state to generate a wide class of reversible rules. He considered taking an arbitrary automaton rule at a given time, and then added an exclusive or (XOR) operation of the result with the state one step back in time. These combined operations could then be reversed by merely interchanging two successive time steps, the analogy of reversing the velocities.

To see this more mathematically, suppose the state at time i is s_i, and the underlying rule begins by taking some arbitrary function $f(s_i)$. Then the full rule takes for the next time step $s_{i+1} = f(s_i)$ XOR s_{i-1}. Here the exclusive or operation is taken site by site over the entire lattice. Elementary properties of the XOR operation then give $s_{i-1} = f(s_i)$ XOR s_{i+1}, which is the identical rule for the time reversed dynamics.

These rules provide a wonderful way to play with the concepts of entropy and reversibility. Indeed, an idealized universe of cellular automata enables experiments which would be impossible to carry out in the real world. In Fig. 9 I show the evolution of a simple image under such a rule. The experiment is a crude simulation of a beer glass shattering after being dropped on the floor. After a few steps it appears quite randomized. Reversal of the momenta of all relevant atoms in the beer glass would allow its reconstruction. In the model this is easily accomplished by swapping two time steps. After reversal, continuing with the same rule reconstructs the original image. At all stages the "information" contained in the system must be constant, even though the image may appear of drastically different complexity.

The reconstruction process is highly sensitive to the reversal being precise. The analog here is to the sensitivity to initial conditions in dynamical systems. In the bottom of Fig. 9c I try to reproduce the beer cup from its shards as in the above experiment, except that now at the time of reversal I modify the state of exactly one pixel. The reversal process recovers the original image only in regions outside the "light cone" for the modified pixel. As the disturbance can only propagate to neighbors in one time step, pixels outside n steps can not know of the change before an equal number of time steps. This use of an XOR operation to generate reversible complex mappings is an integral part of the Data Encryption Standard; see, for example, [20].

Forest Fires

An amusing model of forest fires has three possible states per cell, empty, a tree, or a fire. For the updating step, any empty site can have a tree born with a small probability. At the same time, any existing fire spreads to neighboring trees leaving its own cell empty. The rule here differs from those discussed previously in having a stochastic nature. As the system is made larger, the growth rate for the trees should decrease to just enough to keep the fires going.

If too many trees grow, one obtains a large fire reducing their density, while if there are too few trees, fires die out. On a finite system, one should light a fire somewhere to get the system started. On the other hand, as the system becomes larger, the growth rate for the trees can be reduced without the fire expiring. In a steady state the system has fire fronts continually passing through the system, as illustrated in Fig. 10. Perhaps there is a moral here that one should be careful about extinguishing all fires in the real world, for this may enhance the possibility for a catastrophic uncontrollable fire. It is not entirely clear whether this model is actually critical. What seems to happen on

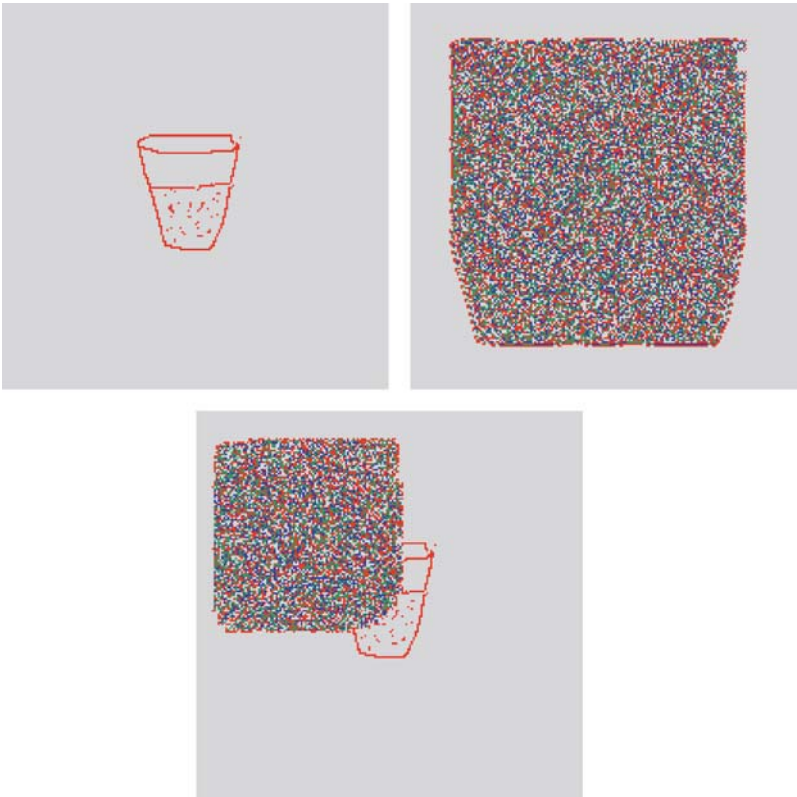

Self-organized Criticality and Cellular Automata, Figure 9
The encryption of a glass of beer. The original rule uses the eight cell neighborhood with births on 1, 3, 5, and 7 neighbors and survivors on exactly 1 neighbors. The rule is modified at each step by XOR'ing the result with the history one time step back. Swapping two adjacent time steps will bring the glass back exactly. The first figure is the starting configuration, the second after 50 steps of evolution. At this point one bit in the *upper left hand quadrant* is flipped, and the dynamics is reversed. The glass is restored in all places beyond 50 steps from the flipped bit. Note the effect of a "speed of light" in the problem

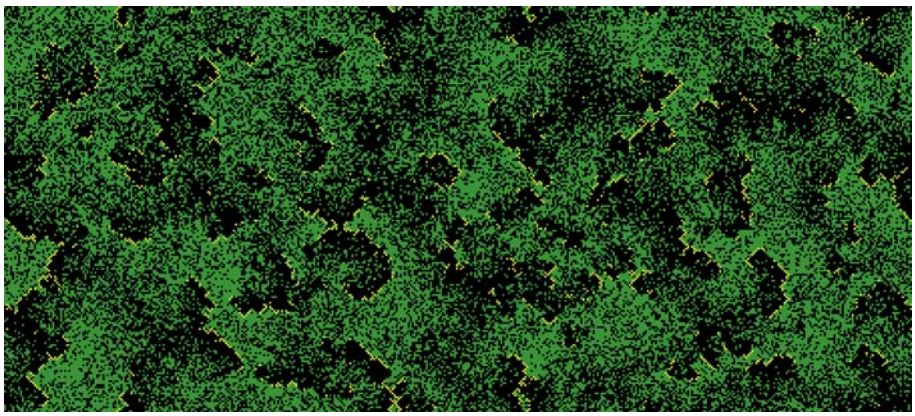

Self-organized Criticality and Cellular Automata, Figure 10
A snapshot of the forest fire model on a 450 by 200 periodic lattice. Trees are continuously burning at a slow rate, while fires burn them down and spread to nearest neighbor trees. Here the four cell neighborhood is used

large systems is that stable spiral structures form and set up a steady rotation. For a review of this and several related models, see [21].

The Sandpile Revisited

Very little is known analytically about general cellular automata. However, in a series of papers, Deepak Dhar and co-workers have shown that the sandpile model has some rather remarkable mathematical properties [22,23,24,25]. In particular, the critical ensemble of the system has been well characterized in terms of an Abelian group. In the following I will generally follow the discussions given in [2,26].

Dhar introduced the useful toppling matrix $\Delta_{i,j}$ with integer elements representing the change in the height, z at site i resulting from a toppling at site j [22]. More precisely, under a toppling at site j, the height at any site i becomes $z_i - \Delta_{i,j}$. For the simple two dimensional sand model the toppling matrix is thus

$$\begin{aligned} \Delta_{i,j} &= 4 && i = j \\ \Delta_{i,j} &= -1 && i, j \text{ nearest neighbors} \\ \Delta_{i,j} &= 0 && \text{otherwise.} \end{aligned} \tag{3}$$

For this discussion there is little special to the specific lattice geometry; indeed, the following results easily generalize to other lattices and dimensions. The analysis requires only that under a toppling of a single site i, that site has its slope decreased ($\Delta_{i,i} > 0$), the slope at any other site is either increased or unchanged ($\Delta_{i,j} \leq 0, j \neq i$), the total amount of sand in the system does not increase ($\sum_j \Delta_{i,j} \geq 0$), and, finally, that each site be connected through toppling events to some location where sand can be lost, such as at a boundary.

For the specific case in Eq. (3), the sum of slopes over all sites is conserved whenever a site away from the lattice edge undergoes a toppling. Only at the lattice boundaries can sand be lost. Thus the details of this model depend crucially on the boundaries, which we take to be open. A toppling at an edge loses one grain of sand and at a corner loses two.

The actual value of the maximum stable height z_T is unimportant to the dynamics. This can be changed by simply adding constants to all the z_i. Thus, as in Sect. "Self-Organized Criticality", I consider $z_\mathrm{T} = 3$. With this convention, if all z_i are initially non-negative they will remain so, and I thus restrict myself to states C belonging to that set. The states where all z_i are non-negative and less than 4 are called stable; a state that has any z_i larger than or equal to 4 is called unstable. One conceptually useful configuration is the minimally stable state C^* which has all the heights at the critical value z_T. By construction, any addition of sand to C^* will give an unstable state leading to a large avalanche.

I now formally define various operators acting on the states C. First, the "sand addition" operator α_i acting on any C yields the state $\alpha_i C$ where $z_i = z_i + 1$ and all other z are unchanged. Next, the toppling operator t_i transforms C into the state with heights z'_j where $z'_j = z_j - \Delta_{i,j}$. The operator U which updates the lattice one time step is now simply the product of t_i over all sites where the slope is unstable,

$$UC = \prod_i t_i^{p_i} C \tag{4}$$

where $p_i = 1$ if $z_i \geq 4$; 0 otherwise. Using U repeatedly gives the relaxation operator R. Applied to any state C this corresponds to repeating U until no more z_i change. Neither U nor R have any effect on stable states. Finally, I define the avalanche operators a_i describing the action of adding a grain of sand followed by relaxation

$$a_i C = R\alpha_i C\,. \tag{5}$$

At this point it is not entirely clear that the operator R exists; in particular, it might be that the updating procedure enters a non-trivial cycle consisting of a never ending avalanche. I now prove that this is impossible. First note that a toppling in the interior of the lattice does not change the total amount of sand. A toppling on the boundary, however, decreases this sum due to sand falling off the edge. Thus, during an avalanche the total sand in the system is a non-increasing quantity. No closed cycle can have toppling at the boundary since this will decrease the sum. Next, the sand on the boundary will monotonically increase if there is any toppling one site further in. This also can not happen in a cycle; thus, there can be no topplings one site away from the edges. By induction there can be no toppling arbitrary distances in from the boundary; thus, there can be no cycle, and the relaxation operator exists. Note that for a general geometry this argument requires that every site be eventually connected to an edge where sand can be lost.

With a system lacking edges, such as under periodic boundaries, no sand would be lost and thus cycles are expected and easily observed. These models might be called "Escher models" after the artist constructing drawings of water flowing perpetually downhill and yet circulating in the system. While little is known about the dynamics of this variation on the sandpile model, some studies have been done under the nomenclature of "chip-firing games" [27]. It has been argued [28] that this lossless sand-

pile model on an appropriate lattice is capable of universal computation.

I now introduce the concept of recursive states. This set, denoted $\mathcal{R}$, includes those stable states which can be reached from any stable state by some addition of sand followed by relaxation. This set is not empty because it contains at least the minimally stable state C^*. Indeed, that state can be obtained from any other by carefully adding just enough sand to each site to make each z_i equal to three. Thus, one might alternatively define $\mathcal{R}$ as the set of states which can be obtained from C^* by acting with some product of the operators a_i.

It is easily shown that there exist non-recursive, transient states; for instance, no recursive state can have two adjacent heights both being zero. If you try to tumble one site to zero height, then it drops a grain of sand on its neighbors. If you then tumble a neighbor to zero, it dumps a grain back on the original site. One can also show that the self-organized critical ensemble, reached under random addition of sand to the system, has equal probability for each state in the recursive set. This is a consequence of the Abelian nature of this system, as discussed below.

The crucial results of [22,23,24,25] are that the operators a_i acting on stable states commute, and they generate an Abelian group when restricted to recursive states. I begin by showing that the operators commute, that is $a_i a_j C = a_j a_i C$ for all C. First I express the a's in terms of toppling and adding operators

$$a_i a_j C = \left(\prod_{k=1}^{n_1} t_{l_k}\right) \alpha_i \left(\prod_{k=n_1+1}^{n} t_{l_k}\right) \alpha_j C \tag{6}$$

where the specific number of topplings n_1 and n depend on i, j, and C. Acting on general states, the operators t and α all commute because they merely linearly add or subtract heights. Therefore I can shift α_i to the right in this expression:

$$a_i a_j C = \left(\prod_{k=1}^{n} t_{l_k}\right) \alpha_i \alpha_j C\,. \tag{7}$$

Now I rearrange the product of topplings. In the nontrivial case that the α-operators render either i or j (or both) unstable, the product must contain toppling operators corresponding to those unstable sites. I shift those operators to the right. Those operators constitute by definition the update operator, U, so I can write

$$a_i a_j C = \left(\prod t_{l_k}\right) U \alpha_i \alpha_j C \tag{8}$$

where the factors within the bracket are the remaining t's. Now, the update operator may leave some sites still unstable, and then the product must include further toppling operators; working on those sites, I can pull out another factor of the update operator. This procedure can be repeated until I have used all the toppling factors and the state is stable. Thus, I can identify the operator within the brackets in Eq. (8) as the relaxation operator R. But $\alpha_i \alpha_j C$ is the same state as $\alpha_j \alpha_i C$, so $a_i a_j C = a_j a_i C$.

A trivial consequence of this argument is that the total number of tumbling events occurring in the operations $a_i a_j C$ and $a_j a_i C$ are the same. Of course, if a particular site k tumbles it can be caused by either addition; the orders of the tumbling events may or may not be altered.

An intuitive argument that sand addition may be commutative uses an analogy with combining many digit numbers under long addition. The tumbling operation is much like carrying, except rather than transferring to the next digit, the overflow spreads to several neighbors. As addition is known to be Abelian, despite the confusing elementary-school rules, I might expect the sandpile addition rule also to be.

I now prove that the avalanche operators have unique inverses when restricted to recursive states; that is, there exists a unique operator a_i^{-1} such that $a_i(a_i^{-1}C) = C$ for all C in $\mathcal{R}$. This implies that the operators a_i acting on the recursive set generate an Abelian group. For any recursive state C I first find another recursive state such that a_i acting on it gives C, and I then show that this construction is unique.

I begin by adding a grain of sand at site i to the state C and then relax the system. This generates a new recursive state $a_i C$. Now since the state C is by assumption recursive, there is some way to add sand to regenerate C from any given state. In particular, there is some product P of addition operators a_j such that

$$C = P a_i C\,. \tag{9}$$

But the a's commute, so I have

$$C = a_i P C \tag{10}$$

and thus PC is a recursive state on which a_i gives C.

I must now show that this state is unique. Consider repeating the above process to find a series of states C_n satisfying

$$(a_i)^n C_n = C\,. \tag{11}$$

Because on a finite system the total number of stable states is finite, the sequence of states C_n must eventually enter a loop. I can run backwards around this loop by adding back the sand repeatedly to the given site. As the original

state C appears in resupplying the sand, C itself must itself belong to the loop. Calling the length of the loop m, I have $(a_i)^m C = C$. I now uniquely define $a_i^{-1} C = a_i^{m-1} C$.

I now have sufficient machinery to count the number of recursive states. As all such can be obtained by adding sand to C^*, I can write any state $C \in \mathcal{R}$ in the form

$$C = \left(\prod_i a_i^{n_i} \right) C^* \,. \tag{12}$$

Here the integers n_i represent the amount of sand to be added at the respective sites. However, in general there are several different ways to reach any given state. In particular, adding four grains of sand to any one site must force a toppling and is equivalent to adding a single grain to each of its neighbors. This can be expressed as the operator statement

$$a_i^4 = \prod_{j \in nn} a_j \tag{13}$$

where the product is over the nearest neighbors to site i. I can rewrite this equation by multiplying by the product of inverse avalanche operators on the nearest neighbors on both sides, thus obtaining for any site i

$$\prod_j a_j^{\Delta_{ij}} = E \tag{14}$$

where E is the identity operator. This allows me to shift the powers appearing in Eq. (12). Define N to be the number of sites in the system. If I label states by the vector $\mathbf{n} = (n_1, n_2, n_3, \ldots, n_N)$ I see that two states are equivalent if the difference of these vectors is of the form $\sum_j \beta_j \Delta_{ij}$ where the coefficients β_j are integers. These are the only constraints; if two states can not be related by toppling they are independent. Thus any vector $\mathbf{n}$ can be translated repeatedly until it lies in an N-dimensional hyper-parallelepiped whose base edges are the vectors Δ_{ji}, $j = 1, \ldots, N$. The vertices of this object have integer coordinates and its volume is the number of integer coordinate points inside it. This volume is just the absolute value of the determinant of Δ. Thus the number of recursive states equals the absolute value of the determinant of the toppling matrix Δ.

For large lattices this determinant can be found easily by Fourier transform. In particular, whereas there are $4N$ stable states, there are only

$$\exp\left(N \int_{(-\pi,-\pi)}^{(\pi,\pi)} \frac{\mathrm{d}^2 q}{(2\pi)^2} \ln(4 - 2q_x - 2q_y) \right) \simeq (3.2102\ldots)^N \tag{15}$$

recursive states. Thus starting from an arbitrary state and adding sand, the system "self-organizes" into an exponentially small subset of states forming the attractor of the dynamics.

An Isomorphism

Following [26], I now look into the consequences of stacking sand piles on top of one another. Given stable configurations C and C' with configurations z_i and z_i', I define the state $C \oplus C'$ to be that obtained by relaxing the configuration with heights $z_i + z_i'$. Clearly, if either C or C' is a recursive state, so is $C \oplus C'$.

Under the operation $\oplus$ the recursive states form an Abelian group isomorphic to the algebra generated by the a_i. First, the addition of a state C with heights z_i is equivalent to operating with a product of a_i raised to z_i, that is

$$B \oplus C = \left(\prod a_i^{z_i} \right) B \tag{16}$$

for any recursive state B. The operation $\oplus$ is associative and Abelian because the operators a_i are.

Since any element of a discrete group raised to the order of the group gives the identity, it follows that $a_i^{|\Delta|} = E$. This implies the simple formula $a_i^{-1} = a_i^{|\Delta|-1}$. The analog of this for the states is the existence of an inverse state, $-C$

$$-C = (|\Delta| - 1) \otimes C \,. \tag{17}$$

Here, $n \otimes C$ means adding n copies of C and relaxing. The state $-C$ has the property that for any state $B \oplus C \oplus (-C) = B$.

The state $I = C \oplus (-C)$ represents the identity and has the property $I \oplus B = B$ for every recursive state B. The state which is isomorphic to the operator a_i is simply $a_i I$. The identity state provides a simple way to check if a state, obtained for instance by a computer simulation, has reached the attractor, i. e. if a given state is a recursive state: A stable state is in $\mathcal{R}$ if and only if $C \oplus I = C$. The proof is simple. By construction, a recursive state has this property. On the other hand, since I is recursive, so is $C \oplus I$.

The identity state can be constructed by taking any recursive state, say C^* and repeatedly adding it to itself to use $|\Delta| \otimes C = I$. However, on any but the smallest lattices, $|\Delta|$ is a very large integer. A more economical scheme is to start with an empty table but use sandy boundary conditions which continually pour sand onto the table. Once it reaches a steady state, switch to open boundary condi-

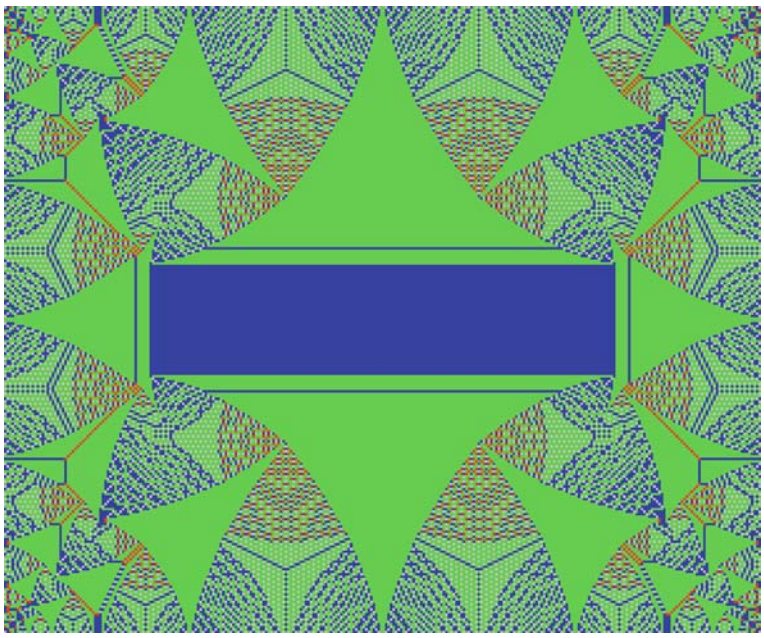

Self-organized Criticality and Cellular Automata, Figure 11
The identity state for the sandpile model on a 302 by 250 lattice. The color code is *gray*, *red*, *blue*, and *green* for heights 0, 1, 2, and 3, respectively

tions and let the sand run back off. This then relaxes to the desired identity. Figure 11 shows the identity state on a 302 by 250 lattice. Note the fractal structure, with features on many length scales.

Majumdar and Dhar [25] have constructed a simple "burning" algorithm to determine if a state belongs to the recursive set. For a given configuration, first add one particle to each of the edge sites and two particles to the corners. This again corresponds to imagining a large source of sand just outside the boundaries, which then tumbles one step onto the system. Then return to open boundaries and update according to the usual rules. If and only if the original state is recursive, this will generate an avalanche under which each site of the system tumbles exactly once. Also, the final state after the avalanche will be identical to the original. However, if the state is not recursive, some untumbled sites will remain. Figure 12 shows such a process underway on the configuration of Fig. 1. Here sites which have already burned are shown in cyan, while the remaining sites in the center have not yet tumbled. The small number of sites shown in orange are the still active sites, which eventually burn the entire remaining lattice.

The burning algorithm provides a simple way to prove that the avalanche regions are simply connected once one is in the critical state. In a burning process, any sub-lattice of the original will have all of its sites tumbled onto from outside. This is the condition for starting a burning on the sub-lattice. Thus, if a configuration is in the critical ensemble for the whole lattice, then any extracted piece of

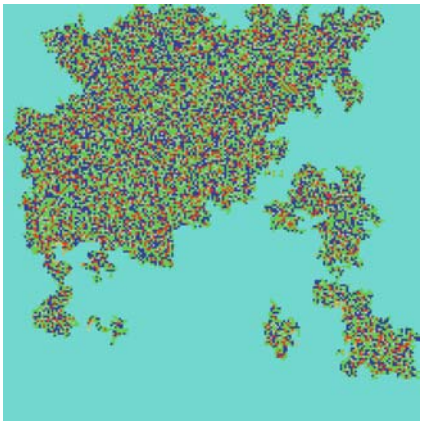

Self-organized Criticality and Cellular Automata, Figure 12
The burning algorithm being applied to the state in Fig. 1. Burnt sites are *cyan*, burning sites are *orange*, and the remaining sites are colored as previously. This avalanche eventually tumbles every site exactly once

this configuration on a subset of the original lattice is also in the critical ensemble of the extracted part. Now suppose that one constructs an avalanche with any initial addition to a state from the critical ensemble. In any subregion enclosed by this avalanche, sand will fall from the tumbling sites on its outside. Since the sub-lattice is itself in its own critical ensemble, this must induce an avalanche which, by the burning algorithm, will tumble all enclosed sites. Thus any avalanche on a state from the critical ensemble cannot leave untumbled any sites in a region isolated from the boundary, i. e. an untumbled island. This result that avalanches must be simply connected does not follow for states outside the recursive set, as can be easily demonstrated by considering a sandpile with a hole of empty sites in the middle.

The burning algorithm has several amusing consequences. One is that any configuration with only height 2 or 3 present is in the critical ensemble as long as the lattice has corners. For example, with all height 2, the burning will start at the four corners of a rectangular lattice and steadily work its way to the center of the system. Another consequence is that in addition to the tumbling region from an avalanche being simply connected, so will the smaller region where the number of tumblings exceed any fixed number; i. e. the region of sites that tumble twice or more is also simply connected.

Future Directions

Simple models as implemented by cellular automata provide a rich area for the study of complex phenomena. Some systems can self organize with physics at many scales, while others provide fascinating demonstrations of thermodynamic laws. I have only touched on a few issues here, leaving out many related topics such as lattice gasses, driven interfaces in random media, growth processes, and evolution. As the ease of programming and the speed of modern computers continue to rush forward, so will the fascination with such models.

Acknowledgments

I am thankful for discussions with many people, but most particularly P. Bak, D. Dar, E. Fredkin, N. Margolis, M. Paczuski, T. Toffoli, F. Van Scoy, and G. Vichniac. This manuscript has been authored under contract number DE-AC02-76CH00016 with the US Department of Energy. Accordingly, the US Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for US Government purposes.

Bibliography

Primary Literature

1. Bak P, Tang C, Wiesenfeld K (1987) Phys Rev Lett 59:381; (1988) Phys Rev A 38:3645
2. Bak P, Creutz M (1994) Fractals and self-organized criticality. In: Bunde A, Havlin S (eds) Fractals in Science. Springer, Berlin, pp 26–47
3. Paczuski M, Maslov S, Bak P (1996) Phys Rev E 53:414
4. Nagel K, Paczuski M (1995) Phys Rev E 51:2909
5. Levy M, Solomon S, Ram G (1996) Int J Mod Phys C 7:65
6. The latest version of the xtoys package is available at http://thy.phy.bnl.gov/www/xtoys/xtoys.html
7. Creutz M (1997) Cellular automata and self organized criticality. In: Bhanot G, Chen S, Seiden P (eds) Some new directions in science on computers. World Scientific, Singapore, pp 147–169
8. Christensen K (1992) Ph D Thesis, University of Aarhus
9. Frette V et al (1996) Nature 379:49
10. Wolfram S (1986) Theory and Applications of Cellular Automata. World Scientific, Singapore
11. Toffoli T, Margolus N (1987) Cellular Automata Machines. MIT Press, Cambridge
12. Bogosian B (1993) Nucl Phys B, Proc Suppl 30:204
13. Berlekamp E, Conway J, Guy R (1982) Winning Ways for your Mathematical Plays, vol 2. Academic Press, New York
14. Wikipedia (2007) Conway's Game of Life. http://en.wikipedia.org/wiki/Conway's_life. Accessed 6 Apr 2007
15. Bak P, Chen K, Creutz M (1989) Nature 342:780
16. Creutz M (1992) Nuclear Phys B, Proc Suppl 26:252
17. Bennett C, Bourzutschy M (1991) Nature 350:468
18. Gardner M (1983) Wheels, Life, and Other Mathematical Amusements. W.H. Freeman, New York
19. Wikipedia (2007) Garden of Eden pattern. http://en.wikipedia.org/wiki/Garden_of_Eden_pattern. Accessed 6 Apr 2007
20. Press W, Teukolsky S, Vetterling W, Flannery B (1988) Numerical Recipes in C. Cambridge University Press, Cambridge
21. Clar S, Drossel B, Schwabl F (1996) Phys J Cond Mat 8:6803
22. Dhar D (1990) Phys Rev Lett 64:1613
23. Dhar D, Ramaswamy R (1989) Phys Rev Lett 63:1659
24. Dhar D, Majumdar SN (1990) Phys J A 23:4333
25. Majumdar SN, Dhar D (1992) Physica A 185:129
26. Creutz M (1991) Comp Phys 5:198
27. Anderson R et al (1989) Amer Math Monthly 96:981; Björner A, Lovász L, Shor P (1991) Europ J Combinatorics 12:283; Eriksson K (1996) SIAM J Discret Math 9:118
28. Goles E, Margenstern M (1996) Int J Mod Phys C 7:113

Books and Reviews

Bak P (1996) How Nature Works: The Science of Self-Organised Criticality. Springer, Berlin

Gore A (1992) Earth in the Balance: Ecology and the Human Spirit. Plume, Boston

Jensen HJ (1998) Self-Organized Criticality: Emergent Complex Behavior in Physical and Biological Systems. Cambridge, Cambridge

Toffoli T, Margolus N (1987) Cellular Automata Machines: A New Environment for Modeling. MIT Press, Cambridge

Wolfram S (1994) Cellular Automata and Complexity: Collected Papers. Westview Press, Boulder

Self-organizing Systems

WOLFGANG BANZHAF
Department of Computer Science, Memorial University of Newfoundland, St. John's, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
History of the Concept of Self-Organization
Examples of Natural Self-Organizing Systems
Examples of Artificial Self-Organizing Systems
Explanatory Concepts of Self-Organization
Modeling Methods
The Role of Self-Organization in Science, the Social Sciences and Engineering
Open Issues and Future Directions
Cross References
Bibliography

Glossary

Attractor A special set of system states approached by a dynamical system after some time has passed when starting from a variety of initial states.

Autopoiesis The process by which systems maintain their identity and organization and regenerate their components in the course of their operation.

Competition and cooperation Types of interaction between two or more elements of a system. Competition refers to each element striving to maximize its use of a finite and/or non-renewable resource. Cooperation refers to the elements engaging in a mutually beneficial exchange.

Complexity Measure of number of elements and way of their interaction (structural c.); measure of variety of behavioral repertoire of a system (functional c.).

Constructive system A system whose later components are generated during the interaction of its earlier components.

Dynamics The quantitative development of a system's state variables over time.

Emergence The appearance of qualitatively new phenomena on higher levels of a hierarchical system.

Evolution A process of structural or qualitative change in some direction.

Instability Inability of a system to keep its state or structure.

Mode Macroscopic behavior of a system caused by the interaction of its microscopic parts via long-range correlations.

Non-equilibrium System state with inflow of matter, energy and/or information causing it to stay away from its most probable state under the hypothetical condition of isolation.

Phase transition A point at which the appearance or behavior, or qualitative nature of the steady state of a system changes suddenly.

Resilience Measure of a system's ability to remain within a domain of stability in response to fluctuations of the system by a perturbation, and the ability of the system to return to that stable domain having once left.

Self-organized criticality The ability of a system to evolve in such a way as to approach a critical point and then maintain itself at that point.

Definition of the Subject

Self-organization is a core concept of Systems Science. It refers to the ability of a class of systems (self-organizing systems (SOS)) to change their internal structure and/or their function in response to external circumstances. Elements of self-organizing systems are able to manipulate or organize other elements of the same system in a way that stabilizes either structure or function of the whole against external fluctuations. The process of self-organization is often achieved by growing the internal space-time complexity of a system and results in layered or hierarchical structures or behaviors. This process is understood not to be instructed from outside the system and is therefore called *self-organized*.

Modern ideas about self-organization start with the foundation of cybernetics in the 1940s. W. Ross Ashby, H. von Foerster and N. Wiener, among others, have contributed to an early understanding. Later, the concept was adopted in physics and nowadays pervades most of natural sciences. Many systems have been identified as possessing aspects of self-organization, though a clear definition is still lacking. As a result of this inaccuracy, the theory of self-organization is still in its infancy. While the concept has found applications in the social sciences and engineering as well, SOSs are an area of active research, with fundamental questions still being explored.

Introduction

Over the last decades a variety of features have been identified as typical for self-organizing systems. Not all of these features are present in all systems able to self-organize. Self-organizing systems are *dynamic*, often *non-determin-*

istic, *open*, exist *far from equilibrium* and sometimes employ *autocatalytic amplification of fluctuations*. Often, they are characterized by *multiple time-scales* of their internal and/or external interactions, they possess a *hierarchy of structural and/or functional levels* and they are able to *react* to external input *in a variety of ways*. Many self-organizing systems are *non-teleological*, i. e. they do not have a specific purpose except their own existence. As a consequence, *self-maintenance* is an important function of many self-organizing systems. Most of these systems are *complex* and use *redundancy* to achieve *resilience* against external pertubation tendencies.

Key aspects of self-organizing systems are:

- Growth of Complexity
- Emergence of new phenomena
- Positive and negative feedback loops of internal regulation.

The process of self-organization has been invoked to explain numerous phenomena in the natural sciences. From non-living systems like galaxies and stars down to nanoparticle aggregates, self-organizing systems have been observed. In the living world cells, organisms and ecosystems provide examples of systems classified as self-organizing. The concept has found applications in man-made systems like communication networks, societies, economies, and has been identified to be at work in the world of ideas in the development of world views, scientific beliefs and norm systems.

History of the Concept of Self-Organization

Early History

The concept of self-organization can be traced back to at least two sources: Western philosophy influenced heavily by Greek thinking; and eastern philosophy, centered around the process thinking of Bhuddism. The ideas derived from both sources resound with the modern way of thinking about self-organization although the word itself had never been used.

On wondering about the origin of the world, Greek atomists from Democritos of Abdera to Epicuros of Samos argued that world order arose from chance collisions of particles. First, the cosmos (from Greek *kosmos* = *the ordered*) did not exist but chaos instead (from Greek *chaos* = *the disordered*). In modern times chaos theory has taken up this topic again, with deep connections to ideas about self-organization and the origin of order in the universe.

In the Christian tradition, St Thomas Aquinas contributed through his interest in logical proofs for the existence of God. One of these proofs considered God to be the ultimate organizer or designer. The argument was that everything had to be organized and this called for an organizer. In turn, the organizer had to be organized and so on back to the original organizer: this was God. Since God is present without cause (otherwise he would have to be organized by another entity), he must have somehow organized himself.

The Bhuddist way of thinking, on the other hand, was fundamentally process-oriented. Things are considered not to be in static existence, but rather are thought to be generated and maintained by proper processes. The emphasis on processes is reminescent of self-organizing systems whose structure is determined by proper processes of internal and external interactions.

The First Use of the Term

Work on General Systems Theory (von Bertalanffy) [1] and Cybernetics (Wiener) [2] paved the way for the idea of self-organization.

The concept of a self-organizing system was introduced by Ashby in 1947 [3]. In the 1950s a self-organizing system was considered to be a system which changes its basic structure as a function of its experience and environment. The term was used by Farley and Clark in 1954 to describe learning and adaptation mechanisms [4]. Ashby [5], in 1960, redefined a self-organizing system to include the environment with the system proper. Von Foerster argued [6], also in 1960, that a self-organizing dynamical system possesses some stable structures (eigenvalues, attractor states) which he later termed eigenbehavior.

Further Developments

This notion was further developed by Haken [7] in 1977 who termed the global cooperation of elements of a dynamical system – resulting in it assuming an attractor state – *self-organization*. Both Haken and Kauffman (1993) [8] argued for a deep connection between self-organization and *selection*. Haken found that modes of collective behavior are competing against each other and considered this process to be Darwinian selection in the non-living world. Kauffman, on the other hand, emphasized the role of constraints on the direction of evolution (mostly of the living), caused by self-organization.

Already in the 1970s, however, ideas branched out into different directions. One branch of the development of the idea deepened the relation to studies of learning and adaptation (Conrad, Kohonen, [9,10]), another branch studied processes of self-organization in systems far from equilibrium (Prigogine, Haken) [11,12]. Chaos

Self-organizing Systems, Table 1
Examples of Self-Organizing Systems

System	Flow	Self -organizing entities	Emergence
Atmosphere	Solar energy	Gas molecules	Patterns of atmospheric circulation
Climate	Energy	Weather conditions (humidity, precipitation, temperature, …)	Distribution patterns
Liquid between plates	Heat	Particle circulation	Movement patterns
Laser	Excitation energy	Phase of light waves	Phase-locked mode
Reaction vessel	Chemicals for BZ reaction	Chemical reactions	Patterns of reaction fronts
Neural networks	Information	Synapses	Connectivity patterns
Living cells	Nutrients	Metabolic reactions	Metabolic pathways/network patterns
Food webs	Organisms of different species	Species rank relation	Networks of species
Highway traffic	Vehicles	Distance of vehicles	Density waves of traffic
City	Goods, information	Human housing density	Settling patterns
Internet	Computer nodes	Connections between nodes	Network connection pattern
Web	Information posted in websites	Links between websites	Patterns of web communities

theory (Thom, Mandelbrot) [13,14] was the line of inquiry into nonlinear systems in mathematics, whereas autopoiesis and self-maintenance where at center stage in biology (Eigen, Rosen) [15,16] neurophysiology (von der Mahlsburg, Linsker [17,18]) and cognitive science (von Foerster, Maturana and Varela) [19,20].

In recent years, self-organizing systems have assumed center stage in the natural sciences [21,22], and the social sciences [23,24,25]. Engineering is beginning to see the usability of the concept [26] in connection with the approach of nano-scale applications and the growing complexity of human artefacts.

Examples of Natural Self-Organizing Systems

Classical examples of natural self-organizing systems are the formation of Benard convection cells in non-equilibrium thermodynamics, the generation of laser light in non-linear optics and the Belousov–Zhabotinsky reaction in chemistry. These are examples from the non-living world, and the complexity of resulting macroscopic space-time patterns is restricted.

Nearly unrestricted complexity through self-organization can be achieved in the living world. For instance, the interaction of species in foodwebs could be looked at from this point of view [22]. Here, we shall briefly look at the self-organization of the Earth's biosphere known as the Gaia hypothesis [27]. This hypothesis states that Earth's living and non-living components self-organize into a single entity called *Gaia*. Gaia can be understood as the whole of the biosphere, that is able to self-stabilize. The model states, in other words, that the biomass of Earth self-regulates to make conditions on the planet habitable for life. In this way, a sort of homeostasis would be sought by the self-organizing geophysical/physiological system of Earth.

In recent years, the Gaia hypothesis has found its place in Earth Systems Science as the realization that there is just one global ecosystem, containing the entirety of resources and all living organisms, all interacting with each other in multiple regulatory cycles. These ideas have been connected to the Darwinian theory of evolution via natural selection [28,29], providing a mechanism by which such a stable state can be assumed to have emerged.

Other examples of natural self-organizing systems can be found in Table 1.

Examples of Artificial Self-Organizing Systems

There are numerous examples of man-made systems or systems which involve man that exhibit self-organization phenomena. Among them are traffic patterns, self-organizing neural networks, celular phone networks or the development of web communities.

The example we shall briefly discuss is that of traffic flow patterns. Macroscopic patterns of traffic jams on highways have been observed and experimentally examined [30]. Their appearance is closely related to traffic density, the model of behavior for drivers and the traffic flow that this allows [31]. Traffic flow is an open system, and it develops waves of traffic jams (solitons) excited by the density of traffic. Transitions between different traffic flow patterns have been considered as phase transitions, typical products of self-organization in the non-living world.

A number of examples of self-organizing systems from different fields is given in Table 1, lower section.

Explanatory Concepts of Self-Organization

Despite half a century of inquiry, the theory of self-organizing systems is still in its infancy. There is no "standard model" of SOS, only various aspects emphasized by different researchers. Here we shall discuss the most important of these.

Non-Equilibrium Thermodynamics

Thermodynamics has been concerned with the notion of order and disorder in physical systems for more than a century. The theory of self-organization has to address fundamental issues of this field. The most important question in this regard is, how order can arise through self-organization.

Classical thermodynamics has focused on closed systems, i. e. systems isolated from external influence in the form of matter and energy flow. This allowed to understand the processes involved when a system evolves undisturbed. A key result of this inquiry is the second law of thermodynamics, originally formulated by Carnot and later refined by Clausius in the 19th century. It states that "any physical or chemical process under way in a system will always degrade the energy". Clausius introduced a quantitative measure of this irreversibility by defining entropy:

$$S \equiv \int \mathrm{d}Q/T \tag{1}$$

with Q the heat energy at a given temperature T. In any process of a closed system, entropy always rises

$$\frac{\mathrm{d}S}{\mathrm{d}t} \geq 0\,. \tag{2}$$

According to Eddington, 1928 [32] this universal increase in entropy "draws the arrow of time" in nature.

Boltzmann had reformulated entropy earlier in terms of the energy microstates of matter. In his notion, entropy is a measure of the number of different combinations of microstates in order to form a specific macrostate.

$$S = k_\mathrm{B}\,\ln(W) \tag{3}$$

with k_B Boltzmann's constant and W the thermodynamic probability of a macrostate. He argued that the macrostate with most microstates (with maximum entropy) would be most probable and would therefore develop in a closed system. This is the central tenet of equilibrium thermodynamics.

More interesting phenomena occur if the restrictions for isolation of a system are removed. Nicolis and Prigogine [11] have examined these systems of non-equilibrium thermodynamics which allow energy and matter to flow across their boundary. Under those conditions, total entropy can be split into two terms, one characterizing internal processes of the system, $\mathrm{d_i}S$ and one characterizing entropy flux across the border $\mathrm{d_e}S$. In a generalization of the second law of thermodynamics, Prigogine and Nicolis postulated the validity of the second law for the internal processes,

$$\frac{\mathrm{d_i}S}{\mathrm{d}t} \geq 0 \tag{4}$$

but explicitly emphasized that nothing can be said about the sign of the entropy flux. Specifically, it could carry a negative sign and it could be larger in size than the internal entropy production. Since the total entropy is the sum of both parts, the sign of the total entropy change of an open system could be negative,

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \frac{\mathrm{d_i}S}{\mathrm{d}t} + \frac{\mathrm{d_e}S}{\mathrm{d}t} < 0 \tag{5}$$

a situation impossible in equilibrium thermodynamics. Thus, increasing order of the system considered would be possible through export of entropy. Self-organization of a system, i. e. the increase of order, would not contradict the second law of thermodynamics. Specifically, the non-equilibrium status of the system could be considered a source of order.

Even in the distance from thermodynamic equilibrium, however, certain stable states will occur, the *stationary* states. These states assume the form of *dissipative structures* if the system is far enough from thermodynamic equilibrium and dominated by non-linear interactions. The preconditions for dissipative structures can be formulated as follows:

1. The system is open.
2. The inner dynamics is mainly non-linear.
3. There are cooperative microscopic processes.
4. A sufficient distance from equilibrium is assumed, e. g. through flows exceeding critical parameter values.
5. Appropriate fluctuations appear.

If those conditions are fullfilled, the classical thermodynamic branch of stationary solutions becomes unstable and dissipative structures become stable system solutions.

Synergetics

Prigogine's description of dissipative structures is formally limited to the neighborhood of equilibrium states. As

Haken pointed out, this is a severe restriction on its application and in particular precludes its formal application to living systems. Instead, Haken proposed *order parameters* and the *slaving principle* as key concepts for systems far from equilibrium. Let the time evolution of a continuous dynamical system is described by

$$\frac{\mathrm{d}\mathbf{q}}{\mathrm{d}t} = \mathbf{N}(\mathbf{q}, \alpha) + \mathbf{F}(t) \tag{6}$$

where $\mathbf{q}(t) = [q_1(t), q_2(t), \ldots, q_N(t)]$ is the system's state vector and $\mathbf{N}$ is the deterministic part of the system's interaction whereas $\mathbf{F}$ represent fluctuating forces, and α are the so-called control parameters. Then the stable and unstable parts of the solution can be separated by linear stability analysis, as can the time dependent and time independent parts. As a result, the solution can be written as

$$\mathbf{q}(t) = \mathbf{q}_0 + \sum_u \xi_u(t)\mathbf{v}_u + \sum_s \xi_s(t)\mathbf{v}_s \tag{7}$$

$\mathbf{v}_u, \mathbf{v}_s$ are the unstable and stable modes, respectively, and $\xi_u(t), \xi_s(t)$ are their amplitudes. These amplitudes obey the following equations

$$\frac{\mathrm{d}\xi_u}{\mathrm{d}t} = \lambda_u \xi_u + N_u(\xi_u, \xi_s) + \mathbf{F}_u(t) \tag{8}$$

$$\frac{\mathrm{d}\xi_s}{\mathrm{d}t} = \lambda_s \xi_s + N_s(\xi_u, \xi_s) + \mathbf{F}_s(t) \tag{9}$$

with λ_u, λ_s characterizing the linear part of the equations and function N summarizing the non-linear deterministic components. The slaving principle formulated by Haken now allows to eliminate the stable mode development by expressing them as a function of unstable modes

$$\xi_s(t) = f_s[\xi_u(t), t]\,. \tag{10}$$

Thus, the unstable modes (order parameters) enslave the stable modes and determine the development of the system's dynamics. This result is useful both to describe *phase transitions* and *pattern formation* in systems far from equilibrium.

Synergetic concepts have been applied in a variety of disciplines [33].

Chaos and Complexity

The treatment of *chaotic systems* has been derived from non-linear system theory. Chaotic systems are usually low-dimensional systems which are unpredictable, despite being deterministic. The phenomenon was originally discovered by the meteorologist E. Lorenz in 1963 [34], although H. Poincare in 1909 was aware of the possibility of certain systems to be sensitive to initial conditions [35]. The reason for the difficulty to predict their behavior stems from the fact that initially infinitesimal differences in trajectories can be amplified by non-linear interactions in the system. These instabilities, together with the lack of methods for solving even one-dimensional non-linear equations analytically, produce the difficulties for predictions. Modern theory of deterministic chaos came into being with the publication of a seminal article by May in 1976 [36].

Complex systems, on the other hand, have many degrees of freedom, mostly interacting in complicated ways, i. e. they are high-dimensional. All the more astonishing is the fact that our world is not totally chaotic in the sense that nothing can be predicted with any degree of certainty. It became apparent, that chaotic behavior is but one of the ways non-linear dynamical systems behave, with other modes being complex attractors of a different kind.

Complexity itself can be measured, notably there exist a number of complexity measures in computer science, but describing or measuring complexity is not enough to understand complex systems.

Self-Organized Criticality

For particular high-dimensional systems, Bak et al. [43] have suggested a dynamic system approach toward the formation of fractal structures, which are found to be widespread both in natural and artificial environments. Their canonical example was a pile of sand. They examined the size and frequency of avalanches under certain well-prepared conditions, notably that grains of sand would fall on the pile one by one. This is an open system with the forces of gravity and friction acting on the possibly small fluctuations that are caused by deviations in the hitting position of each grain of sand. They observed how the grains would increase the slope of the sand pile until more or less catastrophic avalanches developed.

Bak suggested the notion of *self-organized criticality* (SOC) as a key concept which states that large dissipative systems drive themselves to a critical state with a wide range of length and time scales. This idea provided a unifying framework for the large-scale behavior in systems with many degrees of freedom. It has been applied to a diverse set of phenomena, e. g. in economic dynamics and biological evolution. SOC serves as an explanation for many power-law distributions observed in natural, social and technical systems, like earthquakes, forest fires, evolutionary extinction events, and wars. As opposed to the widely studied low-dimensional chaotic systems, SOC systems

have a large number of degrees of freedom, and still exhibit fractal structures as are found in the extended space-time systems in nature.

The Hypercycle

In a series of contributions since 1971, Eigen and Schuster have discussed particular chemical reaction systems responsible for the origin, self-organization and evolution of life [37,38,39,40]. By considering *autocatalytic sets* of reactions they arrived at the most simple form of organization, the *hypercycle*, which is able to explain certain aspects of the origin of life. They have considered a chemical reaction system composed of a variety of self-reproductive macromolecules and energy-rich monomers required to synthesize those macromolecules. The system is open and maintained in a non-equilibrium state by a continuous flux of energy-rich momomers. Under further assumptions they succeeded in deriving Darwinian selection processes at the molecular level. Eigen and Schuster have proposed rate equations to describe the system.

The simplest system realizing the above mentioned conditions can be described by the following rate equations

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = (A_i Q_i - D_i)x_i + \sum_{k \neq i} w_{ik} x_k + \Phi_i(\mathbf{x}) \tag{11}$$

where i enumerates the individual self-reproducing units and x_i measures their respective concentrations. Metabolism is quantified by the formation and decomposition terms $A_i Q_i x_i$ and $D_i x_i$. The ability of the self-reproducing entities to mutate into each other is summarized by the quality factor for reproduction, Q_i, and the term $w_{ik} x_k$ which takes into account all catalytic productions of one sort using the other. A_i, D_i are rate constants for selfreproduction and decay respectively. The flow term Φ_i finally balances the production/destruction in this open system in order to achieve $\sum_k x_k = \text{const}$.

By introducing a new feature called excess production

$$E_i \equiv A_i - D_i \tag{12}$$

and its weighted average

$$\bar{E}(t) = \sum_k E_k x_k / \sum_k x_k \tag{13}$$

and symbolizing the "intrinsic selection value" of a sort i by

$$W_{ii} = A_i Q_i - D_i \tag{14}$$

one arrives at reduced rate equations

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = (W_{ii} - \bar{E})x_i + \sum_{k \neq i} w_{ik} x_k \,. \tag{15}$$

These equations can be solved under certain simplifying assumptions and notably yield the concept of a *quasi-species* and the following extremum principle: A quasi-species y_i is a transformed self-replicating entity with the feature that it can be considered as a cloud of sorts x_i whose average or consensus sequence it is. The extremum principle reads: Darwinian selection in the system of quasi-species will favor that quasi-species which possesses the largest eigen-value of the rate equation system above.

The Origin of Order

In the 1990s Kauffman [41] pointed out one of the weaknesses of Darwinian theory of evolution by natural selection: It cannot explain the '*origin* of species' but rather only their subsequent development. Kauffman instead emphasized the tendency of nature to constrain developments along certain paths, due to restrictions in the type of interaction and the constraints of limited resources available to evolution. In particular he held up the view that processes of spontaneous order formation conspire with the Darwinian selection process to create the diversity and richness of life on Earth.

Previously, Kauffman had formulated and extensively studied [42] the NK fitness landscapes formed by random networks of N Boolean logic elements with K inputs each. Kauffman observed the existence of cyclic attractor states whose emergence depended on the relation between N and K, and the absolute value of K. In the case of large K ($K \approx N$), the landscape is very rugged and behavior of the network appears stochastic. The state sequence is sensitive to minimal disturbances and to slight changes of the network. The attractor length is very large, $\approx N/2$, and there are many attractors. In the other extremal case, $K = 2$, the network is not very sensitive to disturbances. Changes of the network do not have strong and important consequences for the behavior of the system.

Kauffman proposed NK networks as a model of regulatory systems of living cells. He further developed the notion of a *canalizing function* that is a Boolean function in which at least one variable in at least one state can completely determine the output of the function. He proposed that canalizing functions are an essential part of regulatory genetic networks.

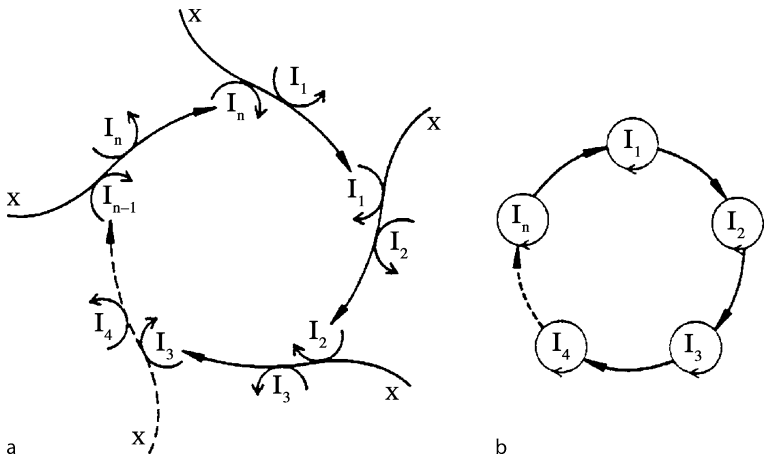

Self-organizing Systems, Figure 1
The Hypercycle (reproduced from [15])

Emergence and Top-Down Causation

The notion of *emergence* has been introduced in complex systems theory in order to explain the appearance of new qualitative features on the level of an entire system that could not be observed at the level of its components. Emergent behavior can be connected to the afore-mentioned complex attractors. It requires switching the level of description of behavior of a system, from local (component-centered) to global (system-centered), or at least to a meso-level (sub-system-centered). Emergent behavior happens when

a) the system shows qualitatively new behavior on a higher level of description which
b) could not have been easily predicted from the interactions of components at the lower level (obeys a non-linear relationship)
c) is the result of a self-organization process.

Emergence is strongly related to self-organization. It is often understood as a pattern formation process. While it essentially has to do with changing the perspective and looking at the system at a different level, it concerns itself with a change in behavior (e. g. the system is getting more organized, shows new coordinated modes of behavior). It has been further conjectured that there is top-down causation, i. e. the structures forming on the higher level of the system are able to affect the lower levels (system components) and influence them in a way that stabilizes the newly emergent behavior. Haken could show in the context of Synergetics that this phenomenon exists. *Top-down causation* is believed to be an important source of complexity, especially in living systems, because it stabilizes patterns.

Self-organization draws heavily from this source of qualitative innovation in complex systems.

Modeling Methods

A formal model is a simplified mathematical or algorithmic representation of a system. Often it has been simplified to the point of a carricature, and this has to be born in mind when making conclusions about the consequences of model predictions. No model can predict beyond the limits of its approximations.

Mean-Field Methods

One of the most important methods used to model complex systems is tied to the notion of dynamical systems. Dynamical systems are systems whose time development is accessible to a description by state changes. It entails the existence of a state space in which these changes can be traced and quantified.

Mean-field methods of description focus on average behavior. They abstract away from the local correlations between a system's elements and describe only long-range changes. For instance, the behavior of a planet could be described as a point on its trajectory around the star it circles. Detailed interactions of its atmosphere would not be part of that description.

Mean-field methods are formulated in the form of time-dependent differential or difference equations which can be solved under certain conditions and predict the behavior of a system in its state space.

Assuming that the state of a system can be subsumed in a vector of state variables **x** whose values are observable

and depend on time, we can generally formulate an equation for continuous time development as:

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}) \,. \tag{16}$$

If time does not develop continuously, a discrete (iterative) equation can be used to describe system behavior:

$$\mathbf{x}_{t+1} = \mathbf{f}(\mathbf{x}_t) \,. \tag{17}$$

The notion is that the state of the system at time t, $\mathbf{x}(t)$ or $\mathbf{x}_t$, is everything that is of interest and can be known about the system. It turned out that with this extreme simplification many systems became treatable that would have otherwise resisted quantitative treatment. The non-linear nature of many of the state development equations, however, and the high-dimensionality of state space vectors often constitute prohibitive hurdles to exact or even approximate mathematical solution of these equations.

As a result, algorithmic approaches for modeling self-organizing systems have become more prevalent in recent years.

Agent-Based Models

A very general class of algorithmic systems is subsumed under the term agent-based models. In these systems, individual entities are modeled that interact with each other. Thus, the approximation of average behavior, and the interest for long-term behavior only is abandoned in favor of a microscopic description of the elements of a system and their interactions. The abstraction of features of a system is achieved through the assumption of rules of behavior of the agents, including their interaction behavior. Agent-based systems must be implemented as computational systems, and run on a computer to obtain results. Agents are assigned states, and transition rules between states, depending on interacting agents, and then these rules are executed in parallel over the set of agents under consideration.

Cellular Automata A particular subset of agent-based models is the class of cellular automata introduced by von Neumann [47] going back to lattice networks of Ulam. The agents of cellular automata are placed on a grid of cells and allowed to assume a finite number of states. Interactions are determined by state transition rules and the definition of a neighborhood, which determines the interactivity of the cellular automaton. Many variants of cellular automata exist, differing in the number of dimensions of the grid, the number of states, the sort and distribution of transition rules and the nature of the neighborhood.

A typical cellular automata model might, e. g. consist of digital cells (allowing only two states, "ON" and "OFF"), homogeneous and deterministic transition rules between states, a one-dimensional grid, and nearest-neighbor interactions. Cellular automata of this type have been thoroughly examined in [44,45] and show a surprising variety and richness in behavior.

In a cellular automaton like LIFE, for instance, one can observe how macroscopic and mesocopic structures appear through self-organization, that is, as a process determined solely by the local interaction of the CA's elements. some structures, e. g. spiral waves, are more resilient against perturbation than others, e. g. glider canons. A moving structure like a glider can be interpreted as an emergent phenomenon as it does not seem to be present on the microscopic scale (single CA cells do not move).

Graphs and Networks A more general class of automata can be formulated if the notion of cellular neighborhood is abandoned. Instead of a rigidly defined grid, a graph or network of automata connected through edges to other automata is introduced. Each node of the graph/network represents an automaton, with interactions allowed via edges.

The notion of a graph is, however, more general, and allows other agent-based systems to be simulated. For instance, the nodes of a graph might represent species of an ecosystem interacting with other species (connected by edges). Each species might be represented by a state counting the number of individuals of that species. Nodes might further hold information on particular features of individuals, and possibly their variants. Explicit simulations of such systems have been considered in the context of "Artificial Chemistries" [46].

In recent years, the structure and dynamics of networks has been a major focus of interest in the scientific community. Network science has become a converging point for different disciplines interested in modelling complex behavior.

Observables

Self-organizing phenomena rest on the appearance of particular sets of behaviors. If ever they are to be understood, a clear notion of observable quantities needs to be developed that allows a proper description of the behavior of such systems. At present, no such canonical set of observables exists, owing to the bewildering variety of systems that show signs of self-organization. However, one can discern a number of different measures and observables that might form the core of such a set [49].

Entropic and Information Theoretic Measures One class of observables can be considered entropic and information-theoretic measures. These measures have in common a statistical root, and seek to describe a self-organizing system in terms of the order (or disorder) that develops over time [48].

Stability Measures Another class of observables can be discussed as stability measures. In this class, systems are sought to be disturbed from their regular behavior in order to obtain a clearer idea of their resilience.

Scaling Measures A further class of observables can be attached to features of scaling. Both theoretical and experimental approaches can be used to vary the number of dimensions, number of equations/agents, number and complexity of interactions, etc, in these systems. Scaling behavior can then be observed for particular quantities and systems classified accordingly.

Patterns and Flows The defining observables of a self-organizing system are patterns. These refer to the collective behavior of the elements of a system, differentiating them from noise. If individual entities would not show such correlations in their behavior, self-organization could not be observed. Patterns can be described in a variety of ways, e. g. as multidimensional vectors, using spatial and temporal coordinates. If patterns change dynamically one can speak of flows.

The central tenet of self-organization is that systems exist whose pattern forming tendencies are determined by themselves, and not by an outside agency.

The Role of Self-Organization in Science, the Social Sciences and Engineering

Self-organization as a concept has assumed center stage in Science. With the advent of nonlinear systems and studies on complex systems in non-equilibrium situations, the explanatory power of self-organization now permeates every branch of scientific enquiry.

From structure formation at the level of super-galactic clusters, even starting from the development of the entire universe, down to microscopic particles and their interaction patterns, self-organizing phenomena have been postulated, theorized, observed and confirmed.

In particular the origin and evolution of life have been studied under the aspect of self-organization. Within Biology, the developmental process of organisms as well as their metabolisms, growth and learning have been identified as self-organizing processes.

In the humanities, the idea of self-organization has taken roots, although the paradigm is far from being fully recognized yet. Since the 1990s the origin and development of languages has been an object of study under the premise of self-organization. In social science the concept of self-organization has been studied since a number of years, due to the obvious fact that interaction between social actors generate a society. Even in psychology, self-organizing principles begin to appear.

Economy and Management Science have taken notice of the concept, and a growing number of enterprise concepts promote the idea of a form as a self-organizing entity.

Finally, Philosophy has embraced the concept of self-organization and connected it to earlier thoughts on the development of the scientific process and epistemology. Whitehead put forward his process philosophy, and Smuts, already in the 1920s, promoted the notion of holism which has strong connections to self-organization. Evolutionary epistemology was formulated as a response to traditional epistemology and emphasizes the aspect of natural selection affecting senses and cognitive abilities.

Engineering is beginning to grasp the ubiquity of self-organization in Nature. Specifically in the area of nanotechnology the concept is used extensively for the purpose of self-assembly of molecular entities. At nanoscales, it is very difficult to directly specify the structuring behavior of entities. As a result, self-organizing properties of matter are used to the advantage of the structural outcome.

Different kinds of infrastructure networks have been recognized as self-organizing, and Engineering begins to make use of the tendency of networked systems to self-organize.

In the area of adaptation, there exists a long tradition of making use of self-organization principles. The self-organizing feature map, introduced by Kohohen, has been a key step forward in the domain of unsupervised learning of artificial neural networks.

Open Issues and Future Directions

So far, there is no unique theory of self-organization. Over the course of many years different approaches have been used, but a coherent picture has not yet emerged.

An important open question in the area of the mathematical basis for self-organization is the formulation of a theory of *constructive (evolutionary) systems*, that is systems which, in the course of their development, generate new elements that subsequently interact with elements already created earlier.

Another question aims at the raison d'etre of *hierarchical systems*. Why do they form, how do they structure

themselves, and what would be possible to apply from these principles in Engineering? Notably, how would one build self-organizing systems such that they do something useful? How could they be controlled?

In Science, the build-up of complexity remains a controversial issue. Is it true that evolution of the universe tends to increase complexity, or is there no tendency of complexity increase at all? What are the mechanisms by which Nature increases complexity, if any? How could we apply this knowledge in planning and managing complexity in the human world?

A wealth of questions remains, and it is anticipated that the 21st century will shed light on at least a few of them.

Cross References

- ► Entropy in Ergodic Theory
- ► Mathematical Basis of Cellular Automata, Introduction to
- ► Swarm Intelligence
- ► Biological Development and Evolution, Complexity and Self-organization in
- ► Rough and Rough-Fuzzy Sets in Design of Information Systems

Bibliography

Primary Literature

1. von Bertalannfy L (1968) General System Theory: Essays on its Foundation and Development. George Braziller, New York
2. Wiener N (1965) Cybernetics: Or Control and Communication in the Animal and the Machine. MIT Press, Cambridge
3. Ashby WR (1947) Principles of the Self-Organizing Dynamic System. J Gen Psychol 37:125–128
4. Farley BG, Clark WA (1954) Simulation of self-organizing systems by digital computer. IRE Trans Inf Theor 4:76–84
5. Ashby WR (1960) Design for a Brain: The Origin of Adaptive Behavior. Chapman and Hall, London
6. von Foerster H (1960) On Self-organizing systems and their environments, In: Yovits MC, Cameron S (eds) Self-Organizing Systems. Pergamon Press, New York, pp 31–50
7. Haken H (1983) Synergetics – An Introduction, 3rd edn. Springer, Berlin
8. Kauffman S (1993) The Origins of Order. Oxford University Press, Oxford
9. Conrad M (1983) Adaptability. Plenum Press, New York
10. Kohonen T (2000) Self-Organizing Maps. Springer Series in Information Sciences, vol 30. Springer, Berlin
11. Nicolis G, Prigogine I (1977) Self-Organization in Nonequilibrium Systems. Wiley, New York
12. Haken H (2000) Information and Self-Organization: A Macroscopic Approach to Complex Systems. Springer Series in Synergetics, 2nd edn. Springer, Berlin
13. Thom R (1989) Structural Stability and Morphogenesis: An Outline of a General Theory of Models. Addison-Wesley, Reading
14. Mandelbrot B (1982) The fractal Geometry of Nature. W.H. Freeman, San Francisco
15. Eigen M, Schuster P (1979) The Hypercycle. Springer, Berlin
16. Rosen R (1991) Life Itself: A Comprehensive Inquiry into the Nature, Origin, and Fabrication of Life. Columbia University Press, New York
17. von der Malsburg C (1973) Self-organization of orientation sensitive cells in the striate cortex. Kybernetik 14:85–100
18. Linsker R (1988) Self-organization in a perceptual network. Computer 21:105–117
19. von Foerster H (2002) Understanding Understanding. Springer, New York
20. Maturana H, Varela F (1979) Autopoiesis and Cognition. Reidel, Dordrecht
21. Camazine S, Deneubourg J, Franks N, Sneyd J, Bonabeau E, Theraulaz G (2000) Self-Organization in Biological Systems. Princeton University Press, Princeton
22. Sole R, Bascompte J (2006) Self-Organization in Complex Ecosystems. Princeton University Press, Princeton
23. Luhmann N (1995) Social Systems. Stanford University Press, Palo Alto
24. Focardi S, Cincotti S, Marchesi M (2002) Self-organization and market crashes. J Econ Behav Organ 49:241–267
25. Portugali J (2000) Self-Organization and the City. Springer, Berlin
26. Bruckner S, di Marzo Serugendo G, Karageorgos A, Nagpal R (eds) (2005) Engineering Self-Organizing Systems. Springer, Berlin
27. Lovelock JE, Margulis L (1974) Atmospheric homoeostasis by and for the biosphere: The Gaia hypothesis. Tellus 26:2–9
28. Lenton TM (1998) Gaia and natural selection. Nature 394:439–447
29. Staley M (2002) Darwinian Selection Leads to Gaia. J Theor Biol 218:35–46
30. Kerner BS (1998) Experimental Features of Self-Organization in Traffic Flow. Phys Rev Lett 81:3797–3800
31. Treiber M, Helbing D (1999) Explanation of Observed Features of Self-Organization in Traffic Flow. Arxiv preprint cond-mat/9901239
32. Eddington AS (1928) The Nature of the Physical World. Gifford Lectures. Cambridge University Press, Cambridge
33. Haken H (2004) Synergetics. Introduction and Advanced Topics. Springer, Berlin
34. Lorenz E (1963) Deterministic Nonperiodic Flow. J Atmospheric Sci 20:130–141
35. Poincare H (1909) Science et Methode. Flammarion, Paris
36. May RM (1976) Simple Mathematical Models with very complicated Dynamics. Nature 261:459–467
37. Eigen M (1971) Selforganization of Matter and the Evolution of Biological Macromolecules. Naturwissenschaften 58:465–523
38. Eigen M, Schuster P (1977) The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle. Naturwissenschaften 64:541–565
39. Eigen M, Schuster P (1978) The hypercycle: a principle of natural self-organization, Part B. Naturwissenschaften 65:7–41
40. Eigen M, Schuster P (1978) The hypercycle: a principle of natural self-organization, Part C. Naturwissenschaften 65:341–369
41. Kauffman SA (1993) The Origins of Order. Oxford University Press, Oxford
42. Kauffman SA (1969) Metabilic Stability and epigenesis in randomly constructed nets. J Theor Biol 22:437–467

43. Bak P, Tang C, Wiesenfeld K (1988) Self-organized criticality. Phys Rev A 38:364–374
44. Wolfram S (1983) Statistical Physics of Cellular Automata. Rev Mod Phys 55:601–644
45. Wolfram S (1984) Cellular Automata as Models of Complexity. Nature 311:419–424
46. Dittrich P, Ziegler J, Banzhaf W (2001) Artificial Chemistries – A Review. Artif Life 7:225–275
47. von Neumann J (1966) Theory of Self-Reproducing Automata. Univ of Illinois Press, Chicago
48. Polani D (2003) Measuring Self-Organization via Observers, In: Banzhaf W et al (eds) Advances in Artificial Life. Lecture Notes in Artificial Intelligence, vol 2801. Springer, Berlin, pp 667–675
49. Shalizi CR, Shalizi KL, Haslinger R (2004) Quantifying Self-Organization with Optimal Predictors. Phys Rev Lett 93:118701

Books and Reviews

Bak P (1996) How Nature Works – The Science of Self-Organized Criticality. Springer, New York
Bar-Yam Y (2003) Dynamics of Complex Systems. Westview Press/Perseus Books, New York
Boccara N (2004) Modelling Complex Systems. Springer, New York
Edelman G (1987) Neural Darwinism. Basic Books, New York
Hemelrijk C (ed) (2005) Self-organization and Evolution of Social Systems. Cambridge University Press, Cambridge
Holland J (2000) Emergence: From Chaos to Order. Oxford University Press, Oxford
Jantsch E (1980) The Self Organizing Universe: Scientific and Human Implications. Pergamon Press, New York
Johnson S (2001) Emergence. Scribner, New York
Kelso S (1995) Dynamic Patterns: The Self-organization of Brain and Behavior. MIT Press, Cambridge
Lovelock JE (1995) The Ages of Gaia. W.W. Norton, New York
Mingers J (1995) Self-Producing Systems. Plenum Press, New York
Morowitz H (2002) The Emergence of Everything. Oxford University Press, Oxford
Rosen R (1999) Essays on Life Itself. Columbia University Press, New York
Sornette D (2003) Critical Phenomena in Natural Sciences. Springer, Berlin

Self-replicating Robotic Systems

Jackrit Suthakorn[1], Matthew Moses[2], Gregory S. Chirikjian[2]
[1] Biomedical and Robotics Technology Lab, Faculty of Engineering, Mahidol University, Salaya, Thailand
[2] Robot and Protein Kinematics Lab, Department of Mechanical Engineering, Johns Hopkins University, Baltimore, USA

Article Outline

Glossary
Introduction
Our Previous Replicating Prototypes
Design and Descriptions of an Autonomous Self-Replicating Robot
Controls and Programming of the Autonomous Self-Replicating Robot
Experiments and Results of the Autonomous Self-Replicating Robot
Our Approach to Self-Replicating Control Circuitry
Towards a Universal Constructor
Discussion
Future Works
Acknowledgments
Bibliography

Glossary

Artificial life Artificial life refers to the study of artificial systems that exhibit lifelike properties, typically involving some form of self-replication and/or evolution. This is a very broad field of study and includes work involving computer simulations, cellular automata, chemistry, robotics, and synthetic biology.

Cellular automata A cellular automaton (CA) is a theoretical construct where a collection of cells are organized into regular grids or lattices. Many arrangements are possible, but typically one-dimensional CAs are composed of square cells arranged in a line, and two-dimensional CAs are composed of square cells arranged in a square grid. Each cell contains a finite state machine. The state of all cells in a CA are typically updated synchronously (at the same time) with each cell changing to a new state that is a function of its previous state and the previous states of its neighbors. The surrounding cells that affect a given cell's state transition are called the cell's neighborhood. CAs can simulate a wide variety of physical processes by designing an appropriate neighborhood and state transition function. CAs are most often studied by implementing them with computer simulations.

Finite state machine A finite state machine (FSM) is a conceptual computing machine with an input, output, and memory. At regular intervals in time the machine transitions (changes) to a new state, which is a function of the current state and the machine's current input. The output is a function of the input and the state. The function of the memory is to store the state information between each transition. FSMs are easily implemented with discrete digital electronic components, microcontrollers, or computer simulations. Implementations of FSMs are widely used to control industrial devices and consumer electronics.

Modular robot A modular robot is composed of distict "modules" which contain motors, mechanisms, electronics, and interconnections. Modules are typically designed so that they can easily be connected and disconnected from each other. In some modular robotic systems, many identical modules are assembled into larger robots. This allows the same set of modules to form robots optimized for different tasks. In other cases, the modules may be specialized for certain tasks, and when assembled they form a robot with additional functionality. Motivation for building modular robots includes increased versatility, ease of replacing damaged components, and potentially lower manufacturing costs.

Self-reconfigurable robot A self-replicating robotic system is a robotic device that exhibits some form of self-replication. This chapter is concerned primarily with directed robotic self-replication, in which a robotic device interprets some form of coded instructions in order to carry out the replication process. In nature deoxyribonucleic acid (DNA) typically encodes replication instructions. In the examples of robotic self-replication presented in this chapter, the instructions may be encoded in a computer program, an arrangement of modular components, or as a pattern of lines that guide the motion of a mobile robot.

Self-replication Self-replication is the process by which an entity creates a duplicate of itself. The most familiar example is the self-replication of living organisms, although other natural processes such as crystal growth can be classified as self-replication.

Universal constructor A universal constructor (UC) is a conceptual machine that reads instructions and executes them to construct an object. A key property of a UC is that it can construct any object which can be described to it via the instructions, including duplicates of itself. UCs have been demonstrated in computer simulations using cellular automata. Modern manufacturing tools, such as assembly robots and computer-controlled machine tools, are similar to UCs but practical limitations make these machines "somewhat less than universal" constructors. The ribosome, a complex molecule present in nearly all biological cells, performs a function very similar to that of a UC, assembling proteins according to instructions encoded in messenger ribonucleic acid (mRNA).

von Neumann universal constructor The mathematician John von Neumann proposed a cellular automata model of a universal constructor capable of self-replication. Many researchers have refined and improved von Neumann's original design since it was first presented in the 1950s. The original design used tens of thousands of cells in a two-dimensional CA with 29 states. The general structure of the model is a movable constructing arm controlled by instructions encoded in a long line of cells resembling a tape, conceptually similar to a computer-controlled robot arm.

Introduction

The concept of an artificial self-replicating system was introduced in the 1950s by John von Neumann [31]. Von Neumann introduced the theory of self-replicating automata and established a quantitative definition of self-replication. His early results on self-replicating machines have become useful in several diverse research areas such as: cellular automata, nanotechnology, macromolecular chemistry, and computing [8,23,26,27]. However, prior to the turn of the millennium, a fully autonomous self-replicating physical robot had never been implemented. In this chapter, a series of prototype designs from our laboratory and their physical implementation are described. We begin by discussing some motivation and history, then go on to describe a remote-controlled replicating robotic system and a semi-autonomous replicating robotic system. We then describe some fully autonomous self-replicating systems, and discuss how manufacturing work cells might be designed so as to reproduce.

Motivation

People have imagined for years a factory that could autonomously replicate itself for multiple generations, requiring neither people nor the monstrous machinery typically associated with a factory. Over recent decades, outer space has been mentioned as one potential application for such self-replicating robotic factories (see e. g., [3,10], and [7]). However, enormous technical barriers must be overcome before these systems can become feasible. The purpose of the current work is to take one small step toward realizing this goal.

In contrast to self-reconfigurable robotics [1,12,14,21,34], self-replication utilizes an original unit to actively assemble an exact copy of itself from passive components. This has the potential to result in exponential growth in the number of robots available to perform a job, thus drastically shortening the original unit's task time.

Descriptions of Self-Replication by Johns von Neumann

According to von Neumann there are four components required for a self-replicating machine: the builder, the

controller, the copier, and the blueprint [31]. The process of self-reproduction begins with the controller commanding the builder to fabricate an exact mechanical system by following the blueprint. Then the controller instructs the copier to replicate the blueprint, input the new blueprint in the replica, and start the new machine. The process described in the previous paragraph was written as a mathematical model called the "Von Neumann Kinematic Beast" by Rolf Pfeifer et al. of the University of Zurich [24], which goes as follows:

1. Let A represent the builder. If a machine G is desired, one can build this machine with the blueprint, *Blue(G)*.

$$A + Blue(G) \rightarrow G\ .$$

   Where '+' indicates that the machine is composed of the left and right components ($A + Blue(G)$) and '→' indicates construction.
2. Let B represent the copier. B would make a copy of the blue print, *Blue(G)*.

$$B + Blue(G) \rightarrow Blue(G)\ .$$

3. Let C be the controller. With the combination of A and B, this would trigger them to follow in the correct order to generate a desired mechanical system G and the new blueprint *Blue(G)*, and then wrap them up together and split them from the original machine

$$A + B + C + Blue(G) \rightarrow G + Blue(G)\ .$$

4. Let G be the machine $A + B + C$ then we get:

$$A{+}B{+}C{+}Blue(A{+}B{+}C) \rightarrow A{+}B{+}C{+}Blue(A{+}B{+}C).$$

Any system that observes this is a self-reproducing machine in Neumann's view. A short time after successfully presenting some of his theoretical work, von Neumann began working on the implementation of his theories. In the 1950's, with the help of Stanislaw Ulam, von Neumann invented the Cellular Automata concept (redrawn and shown in Fig. 1.)

Previous Efforts in Mechanical Self-Replicating System

Von Neumann [31] was the first to seriously study the idea of self-replicating machines from a theoretical perspective. In the late 1950's, Penrose performed the first recognized demonstration of a self-replicating mechanical system [22]. It consisted of passive elements that self-assembled under external agitation. This is similar in many ways to the modern work of Whitesides [32], only at a different length scale. Moore [19] was interested in von Neumann's concepts, but he commented that Neumann's Cellular Automata was only for demonstration, and that applications of self-replicating systems needed to be carried out. Moore described several conceptual designs of artificial living plants which could duplicate themselves not only from off-the-shelf artificial parts, but also from materials from nature. Jacobson [13] constructed a self-replicating machine using parts of toy trains. His replication was done on a round section of toy track.

In the 1980s, NASA established a series of studies on the topic of "Advanced Automation for Space Missions" [10]. These studies investigated the possibility of building a self-replicating factory on the moon. References [5,6,7,9,30] also outlined strategies for space utilization. Recently, research on robots that are capable of designing other machines with little help from humans has also been performed (see [17] and references therein). This uses rapid prototyping technologies.

Our Previous Replicating Prototypes

This section describes in chronological order the robotic prototypes capable of various levels of self-replication that have been designed and built in our laboratory. Our prototypes are constructed from modified LEGO Mindstorm kits with enhanced electrical connections because of their modularity, functionality and ease of use.

Demonstration I: Prototype 1

We built this robot as the first prototype to demonstrate that it is mechanically feasible for one simple robot to produce a copy of itself. This robot depends on external passive fixtures for self-replication. Figure 2 shows fixtures, the original prototype 1, and a set of the replica's subsystems. (Details of Prototype 1 can be found in [27].)

Demonstration II: Prototypes from the Spring 2002 Mechatronics Class at JHU.

From experience gained from Demonstration 1, we introduced the concept of self-replicating robots to students in a hands-on Mechatronics course (taught in the Department of Mechanical Engineering at the Johns Hopkins University) under the supervision of the last author, and TA'ed by the first author. We divided students in this course into eight groups to explore designs and implementations of the concept of self-replicating robotic systems. In order to focus on the mechanical issues involved in the design of the self-replicating systems, the robots

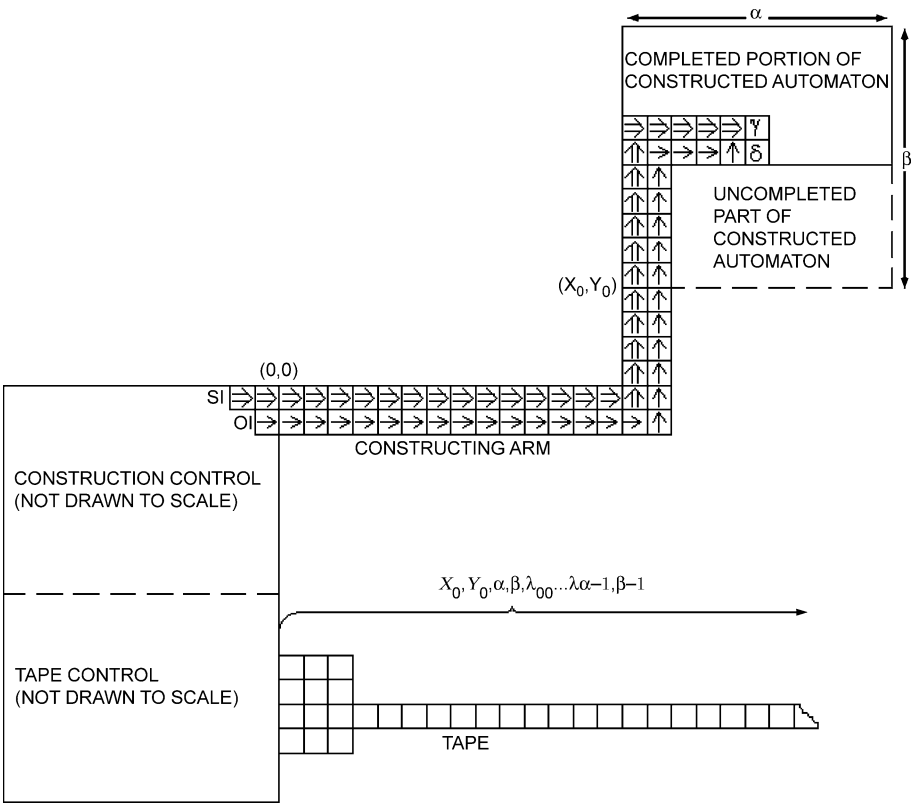

Self-replicating Robotic Systems, Figure 1
Schematic of von Neumann's self-reproducing cellular automaton (Redrawn from [31])

Self-replicating Robotic Systems, Figure 2
Fixtures, the original prototype 1, and a set of the replica's subsystems

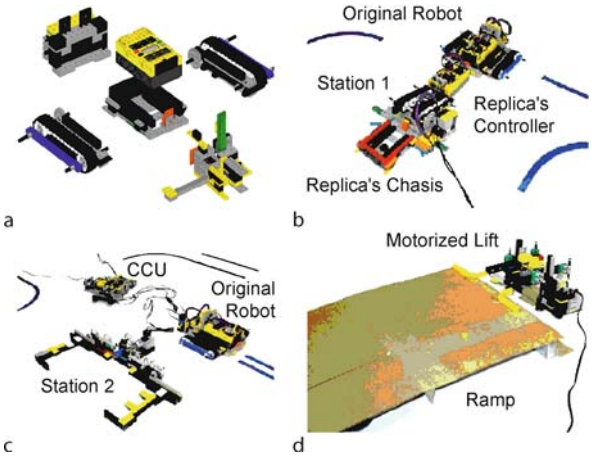

Self-replicating Robotic Systems, Figure 3
a An assembly view of the semi-autonomous robotic system, b Assembling station 1, c Assembling station 2, and d Assembly station 3

were remote-controlled rather than autonomous. (Details of Demonstration II can be found in [27], and [2].)

Demonstration III: A Semi-Autonomous Replicating System

This work builds upon previous results in remote-controlled robotic replication with a new feature that many subtasks in the replication process are now autonomously performed by the robot. The use of feedback sensors was implemented. The robot is unable to directly build

its replicas. Therefore, several work-cells (intermediate robots) are required to assist the original robot in the replicating process. Each work-cell works as a station in this factory-liked replicating system. Figures 3a, b, c, and d illustrate an assembly view of the original robot, stations 1, 2, and 3, respectively. Full details of the semi-replicating robot can be found in [27,28].

Design and Descriptions of an Autonomous Self-Replicating Robot

The robot and its replicas each consist of four subsystems: controller, left tread, right tread, and gripper/sensor subsystems. All subsystems are connected to others using magnets and shape constraints. Figure 4 shows an assembly view of the robot.

The controller subsystem is made up of a LEGO RCX programmable controller fit inside a chassis. The chassis's sides are used to connect to the left and right treads. Each side has a set of magnets, a set of shape-constraining blocks, and a set of electrical connections. The front end of the chassis is designed to attach with the gripper. The front end also has a set of magnets, a set of shape-constraining blocks, and a set of electrical connections, which transfer electrical signals and power from the controller to the gripper's motor and the navigating sensors installed on the gripper subsystem.

The magnets and the shape-constraining blocks are used in collaboration to aid aligning and interlocking subsystems. On each chassis side, the magnets are symmetrically placed in opposite polar directions to each other. This is to protect against incorrect positioning of the subsystems. The concept of using the magnets (with different polarizations) and shape-constraining blocks was influenced by the self-complementary molecules of Rebek [25]. Figure 5 illustrates the concepts of using the polar magnets and shape-constrained blocks to align and interlock subsystems. By design, it is very difficult for these connectors to misalign.

The left and right tread subsystems are designed to be identical to each other, with the purpose of reducing the system's design complexity. A tread subsystem hosts a rubber tread with a driving gear system, a 9 V LEGO DC motor, and a light-reflective pad which helps the original robot's navigation. One side of the tread has a set of magnets, a set of shape-constrained blocks, and a set of electrical connections, all of which correspond to the side of the controller subsystem. On the other side, the tread has a wedge which is fitted to the gripper. The wedge is used during the tread subsystem's transferring and assembling processes. Figure 6 shows how the original robot grasps

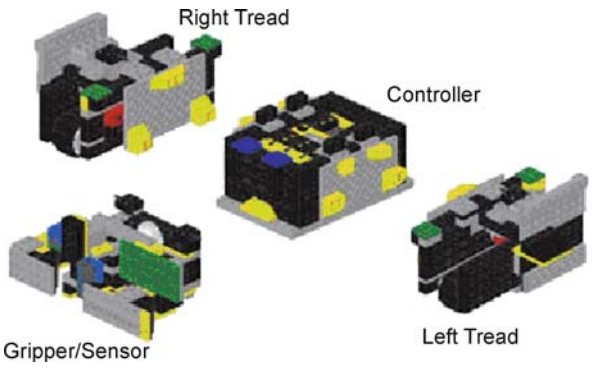

Self-replicating Robotic Systems, Figure 4
An assembly view of the self-replicating robot

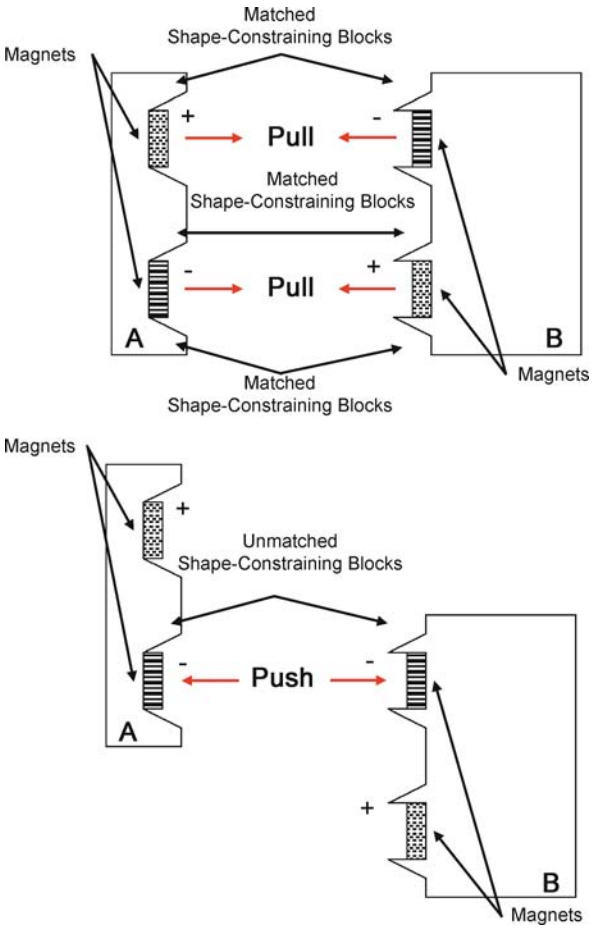

Self-replicating Robotic Systems, Figure 5
This diagram illustrates the concept of using polar magnets and shape-constraining blocks (*top*: correctly aligning, and *bottom*: incorrectly aligning)

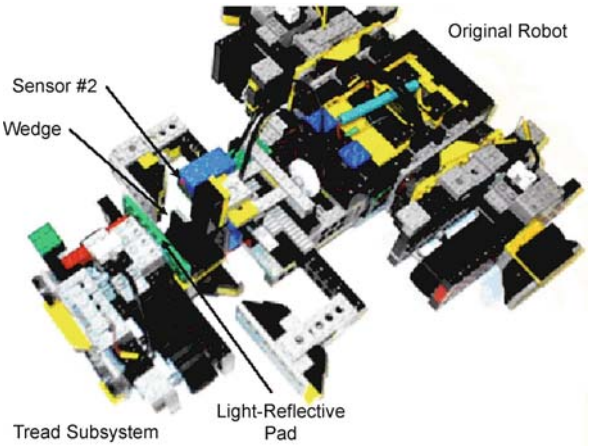

Self-replicating Robotic Systems, Figure 6
The original robot grasps the tread subsystem

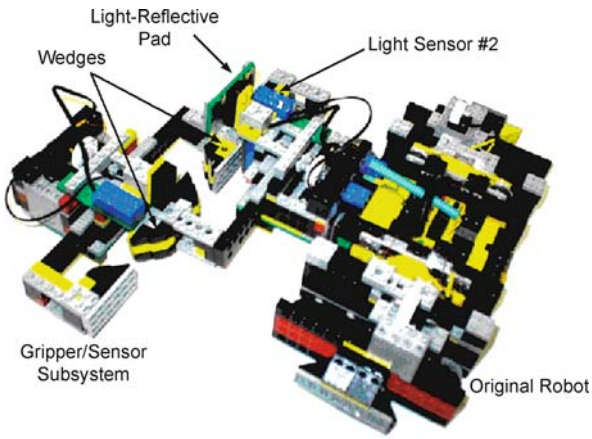

Self-replicating Robotic Systems, Figure 8
The original robot grasps the gripper/sensor subsystem

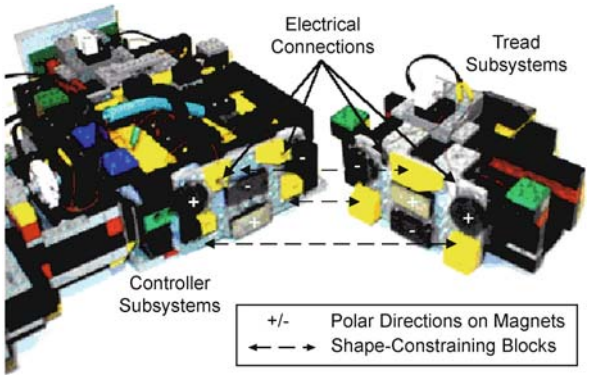

Self-replicating Robotic Systems, Figure 7
The connections located between controller and tread subsystems

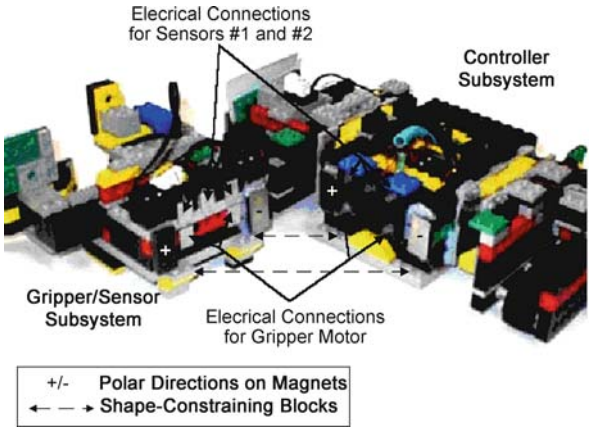

Self-replicating Robotic Systems, Figure 9
The connections located between gripper/sensor and controller subsystems

the tread subsystem, and Fig. 7 shows the connections located between the controller and tread subsystems.

The gripper/sensor subsystem is comprised of a 9 V LEGO DC motor, a set of rack and pinion gears used to drive the left/right fingers of the gripper, a set of magnets, a set of shape-constrained blocks, a set of electrical connections, and two light sensors (one is pointed downward, and the other is pointed forward).

The set of magnets, shape-constraining blocks, and electrical connections are attached to their corresponding part, on the front side of the controller subsystem. The left finger of the gripper is designed in a wedge shape to be fitted with the gripper in any identical robot. This wedge is used in the same manner as in the tread subsystems during assembling processes. Figure 8 shows how the original robot grasps the gripper/sensor subsystem, and Fig. 9 shows the connections located between gripper/sensor and controller subsystems. The two LEGO light sensors are employed in the robot's navigation system. The first light sensor (pointed downward) is used to detect the blue painted lines and silver acrylic spots on the experiment surface. The second light sensor (pointed forward) is used to detect objects (the subsystems of the replica) which the robot runs into.

The experimental area is a 2 m x 3 m area made of white colored paper with lines and spots painted in blue and silver acrylic colors. The original robot starts at the initial position, and the replica's subsystems are at their locations. Figure 10 shows the experimental area with locations of the replica's subsystems and the initial position of the original robot.

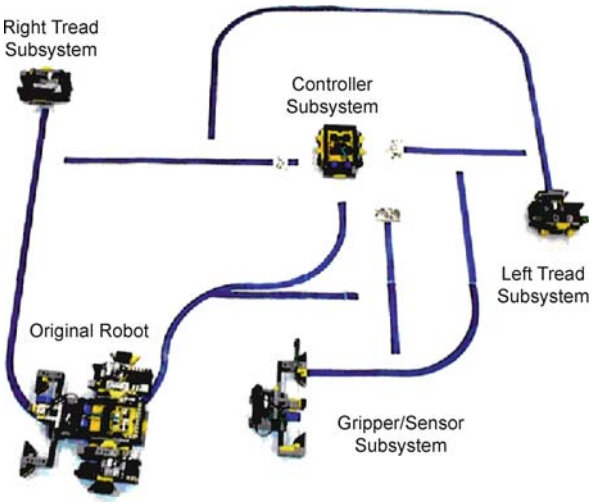

Self-replicating Robotic Systems, Figure 10
A map of the experimental area

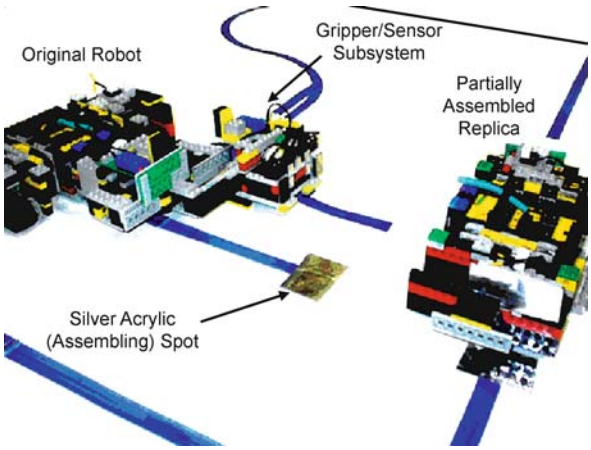

Self-replicating Robotic Systems, Figure 12
The robot is searching for the assembly location while holding the replica's gripper/sensor subsystem

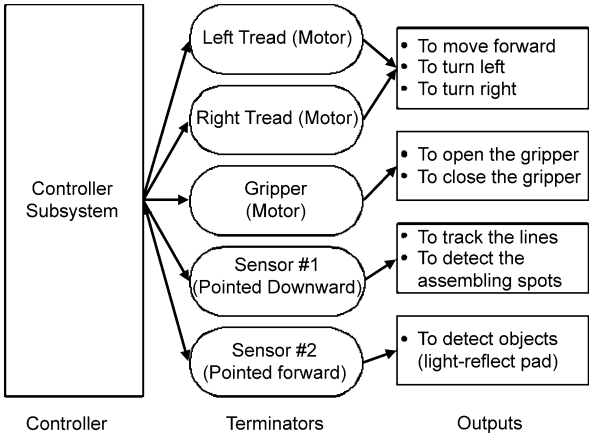

Self-replicating Robotic Systems, Figure 11
The control architecture of the autonomous self-replicating robot

Controls and Programming of the Autonomous Self-Replicating Robot

The prototype robot is a fully autonomous system. Figure 11 shows the conceptual control architecture of the robot. The robot and its replicas, using the LEGO light sensor No. 1 (pointed downward), is capable of tracking the blue lines, and it can recognize the assembling spots, painted in a silver acrylic. The sensor detects and returns different analog values, corresponding to different colors. The robot tracks the painted lines to navigate between positions. Once the robot detects the assembling spot, the robot begins the assembling process. The LEGO light sensor No. 2 (pointed forward) returns an analog value once it detects a light-reflective pad attached to the tread and gripper/sensor subsystems. This notifies the robot to begin grasping the detected subsystem.

The grasping process consists of an aligning push toward the subsystem, and closing the gripper to grasp the subsystem. On the other hand, the assembly process consists of opening the gripper to release the subsystem, and an aligning push forward to snap the subsystem to the controller. Figure 12 shows the original robot grasping a subsystem, and moving toward an assembling spot, in silver acrylic, along the blue line.

The programming of the prototype is described here. The code is programmed on a PC and transferred through a LEGO infrared program-transferring tower. In the order in which events take place in the replication process, the programming is separated into seven stages: 1) replication process is activated, 2) line tracking and searching for a subsystem, 3) grasping the subsystem and changing to a new path which leads to the next step, 4) line tracking and searching for the assembly location, 5) assembling the subsystem to the controller and changing to a new path, 6) the new path leads to the next subsystem, 7) The final step loops back to steps 2 through 6 so the process repeats indefinitely. Figure 13 illustrates the programming flowchart of the self-replicating robot system.

Experiments and Results of the Autonomous Self-Replicating Robot

The following is a step-by-step outline of the procedure that our autonomous self-replicating robot system under-

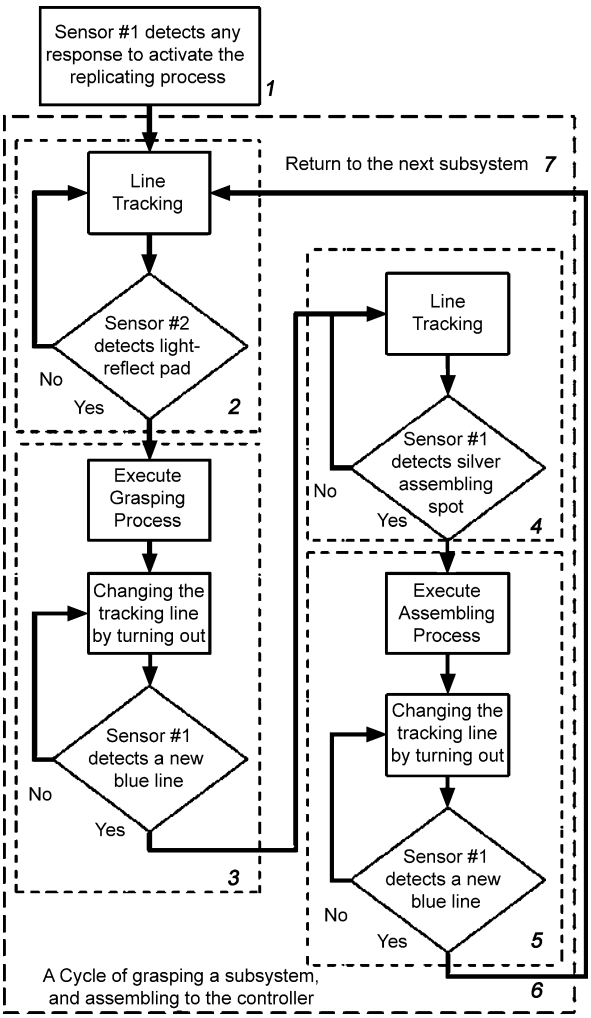

Self-replicating Robotic Systems, Figure 13
The flowchart of the self-replication process

goes (Fig. 14 is a photographic representation of these step but they are not synchronous):

1. The original robot starts following the line from the starting point to the first subsystem using sensor No. 1.
2. Once light sensor No. 2 detects the first subsystem (right tread), the original robot begins the grasping process and grasps the right tread subsystem.
3. After the grippers are closed the original robot turns to the right until it detects a line.
4. The original robot follows the second line until it reaches the assembly location.
5. When light sensor No. 1 on the original robot detects the silver acrylic spot (the assembly location), the robot stops, and begins the attaching process.
6. The original robot opens the grippers, and gives a final push to secure the right tread subsystem to the controller subsystem.
7. The original robot then backs up and turns to the left until it detects a line value on sensor No. 1.
8. The original robot follows the line until it reaches the left tread subsystem.
9. Once light sensor No. 2 detects the second subsystem, the robot will stop, and begin the grasping process by closing its gripper around the left tread's wedge.
10. The original robot turns right until it detects the next line.
11. The original robot will follow the second line until it reaches the assembly location.
12. The original robot opens its gripper to release the left tread subsystem.
13. The original robot gives a final push on the left tread subsystem to help secure it.
14. The original robot then backs up and turns left until it detects the next line, using sensor No. 1.
15. The original robot follows the line to the final subsystem.
16. Once it reaches the gripper/sensor subsystem, it stops, and begins the grasping process.
17. The original robot closes its gripper, and turns right until it detects a line value with sensor No. 1.
18. The gripper/sensor subsystem is now transferred to the assembly location.
19. Once the original robot reaches the assembly location it stops, and opens the gripper.
20. The original robot backs up and turns left until sensor No. 1 is a line value.
21. The original robot then follows the line back to the starting point, and is ready to replicate again.
22. The completed replica self-activates (20 seconds after completion) and begins following the line to the starting point.
23. Once each robot reaches the starting point, it begins the replication procedure again.

The replication process takes two minutes and fifteen seconds per cycle. Although each subsystem is required to be placed in its starting location, errors in initial position and orientation are not very critical. We found slight errors during the grasping process in a few experiments caused by improper placement of the subsystems. Overall, the system is robust and very repeatable.

Our Approach to Self-Replicating Control Circuitry

Ideally, for a robotic system to be truly self-replicating, it would have to demonstrate the ability to assemble all of

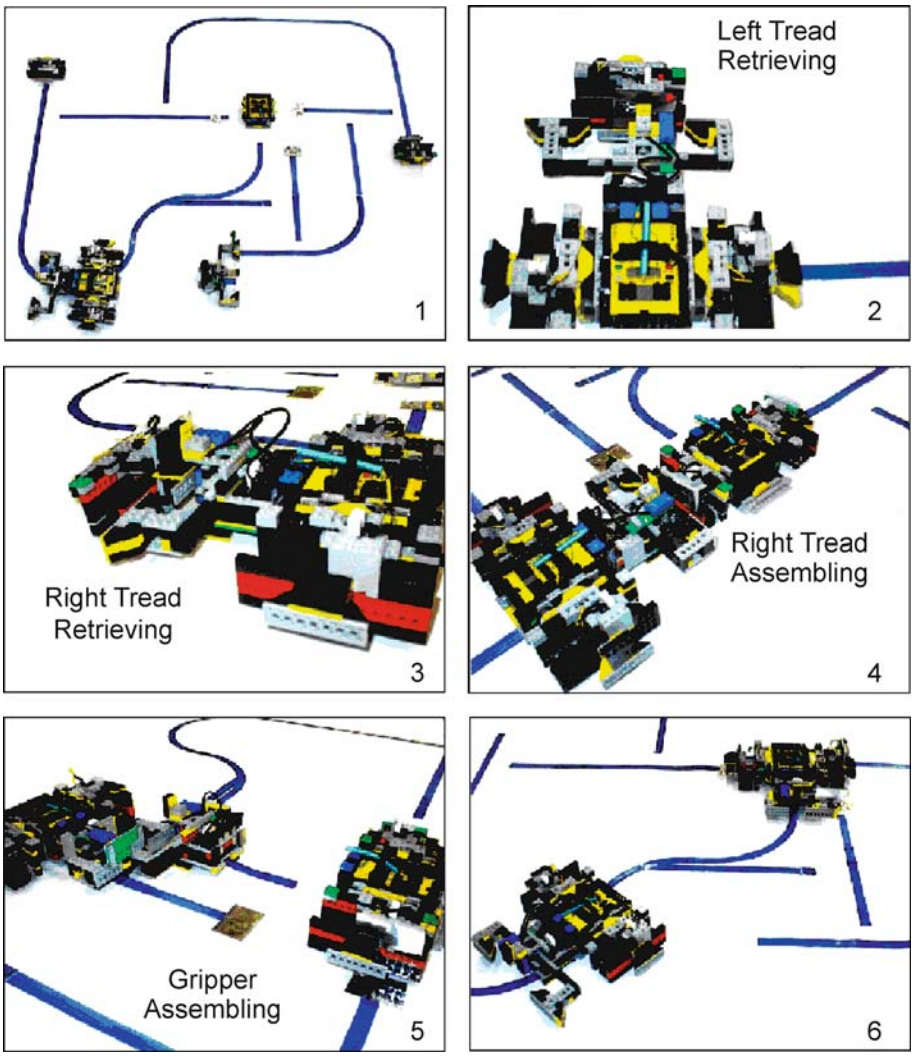

Self-replicating Robotic Systems, Figure 14
Self-replication process: 1) the original robot begins at the initial position with every part placed in their position. 2) The robot detects the right-tread subsystem. 3) The robot grasps the right-tread subsystem, and attaches it to the controller. 4) In the same manner, the robot performs the assembly of the left-tread subsystem. 5) After the robot grasps the gripper/sensor subsystem the robot transfers the subsystem to the next assembly step at the controller. 6) After being fully assembled, the replica is self-activated, and ready to replicate just like the original

its own subsystems from the most fundamental components. In the case of the robot controller, we consider the most fundamental components to be transistors, resistors, capacitors, etc., whereas microcontrollers are too complex to be considered as basic elements.

Our approach is to build a circuit capable of controlling an electro-mechanical system to re-build replicas of the control circuit from the most fundamental electronic components. In the von Neumann universal constructor paradigm, an associated instruction code is also required. In contrast it is possible to replicate a particular system by self-inspection without invoking von Neumann's universal constructor. We illustrate both concepts in hardware designed and constructed by students in a Mechatronics course taught at Johns Hopkins University in 2003. Two prototypes illustrate replication by self-inspection, and one demonstrates the universal constructor. In all three cases, pre-built electro-mechanical systems (called the SRI-builders) use the transistorized circuit as its controller. While in the von Neumann paradigm, the controller follows instructions that are explicitly encoded (and hence must reproduce the code for the overall system to

be self-replicating), in the self-inspection paradigm, actions are taken implicitly as a result of observing the spatial layout of components in the original and feeding that information into the circuit itself. Clever electromechanical design ensures that observations obtained during self-inspection are translated directly into actions without requiring the interpretive step of consulting a long sequence of encoded construction commands.

A von Neumann Universal Constructor Prototype

The prototype is a two-arm gantry-style robot with two degrees of freedom. The first degree of freedom moves along the whole system, includes the feeders and the assembly boards. The second degree of freedom moves vertically to pick and place codes and circuit pieces. A controller circuit is used to control motions of the robot. There are two boards being assembled at a certain amount of time. The first board is the replicated circuit, and the second is the replicated codes. The circuit is pre-wired before it is placed into the carrier. This is similar to a process of placing a chip into a circuit board. In the code part, each code consists of three lines of black and white strip representing 3 bits. A series of codes is set up on a code array where it is fed simultaneously to a reader array. The reader array is made of an array of photo-transistors and infrared LED emitters used to read black (0) and white (1) colors in each code. Infrared LED emitters are used instead of regular LED for reducing problem with ambient light distractions. Figure 15 shows the system in a side view. Once the acquisition part of the system reads the code, the assembly part of the system works by following the instruction code, replicating the code and circuit of that part.

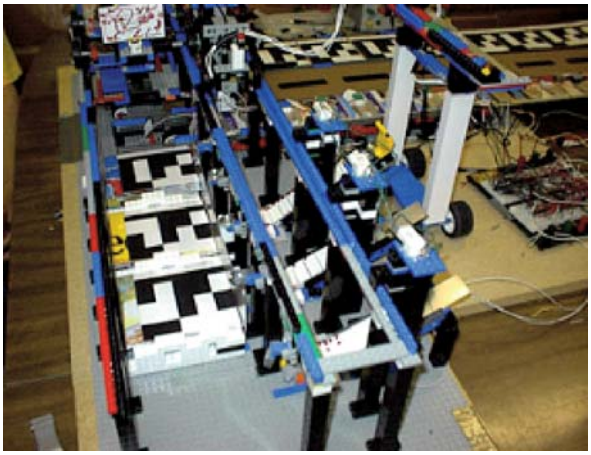

Self-replicating Robotic Systems, Figure 15
Side view of the replicating system (von Neumann universal constructor prototype)

Non-Universal Self-Replication by Self-Inspection (Design 1)

This design is the first design of a non-universal self-replication by self-inspection. The self-replicating control circuit has the ability to identify the proper electronic components required, translate information about its own constituent parts obtained from self-inspection into mechanical tasks that create a replica, and transfer all functions to the replica. There is no list of instructions in the form of a code. Each electronic component has a black-and-white color code. Parts are loaded into feeders, and as a reading head traverses the control circuit, the information about which part of the control circuit is being observed is fed into the circuit itself. This actuates the solenoid in the ap-

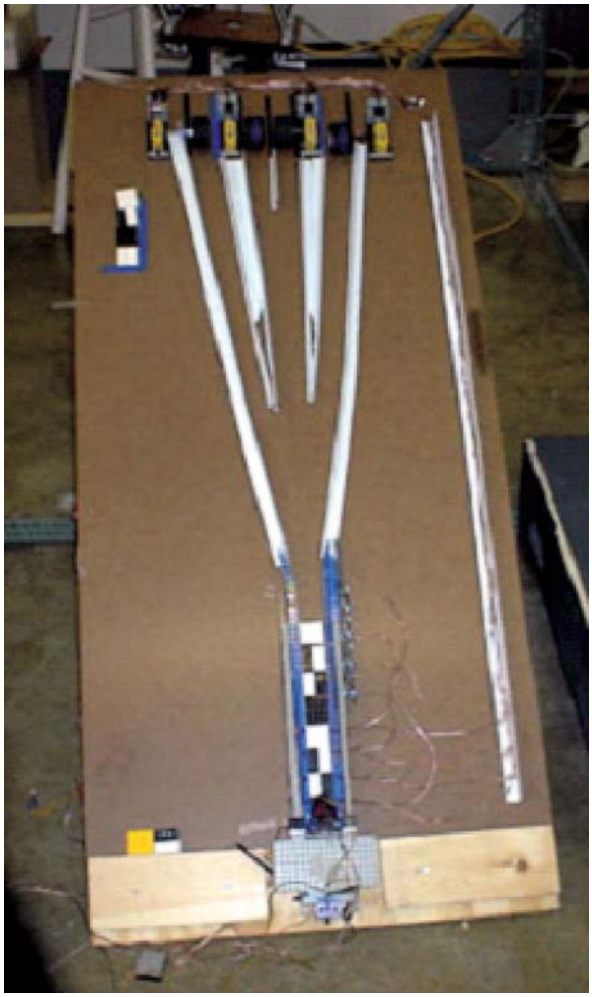

Self-replicating Robotic Systems, Figure 16
Side view of the replicating system (non-universal self-replication by self-inspection – design 1. (See [11] for details))

propriate feeder to release the parts needed to form the replica. Parts then slide down an incline and form an orderly array. The reading head continues to move and creates replicas until resources are completely utilized or its track ends. The design is scalable and the components are modular, allowing many different levels of intelligence to be replicated. This concept is one of many which we are investigating to enable self-replicating robots to perform complex behaviors. See Fig. 16.

Non-Universal Self-Replication by Self-Inspection (Design 2)

This robotic system is an X-Y table constructed from modified LEGO components. A photo-transistor sensor system is attached to the end-effector of the X-Y system in order to inspect the control circuit (the components of which are each assigned a unique black and white code). On the top of the X-Y system, a set of component feeders is installed. The circuit converts the signal from the sensor system to control the component feeders to release the correct component to the parts assembler. The parts assembler then arranges all the components to create a new replica of the control circuit. See Fig. 17.

Merging the Self-Replicating Robot and Self-Replicating Circuit Concepts

In prior prototypes we considered two disjoint cases: (1) self-replicating robots constructed from modules where one module contained a computer; (2) a control circuit that commanded a mechanical device to assemble copies of the circuit. In this section we review very recent work with K. Lee and the JHU authors that merges these two concepts.

Basically, the autonomous self-replicating robot described previously in this paper functions as a finite-state machine in a highly structured environment. This same behavior can be implemented by a robot controlled by a simple circuit without using a microprocessor. Discrete electronics elements of this circuit such as transistors, resistors, etc., can be distributed over the modules from which the robot is constructed. This means that as the original robot assembles these modules to form a replica, the controller for the replica is assembled as the robot is. Prototypes are shown in Figs. 18 and 19.

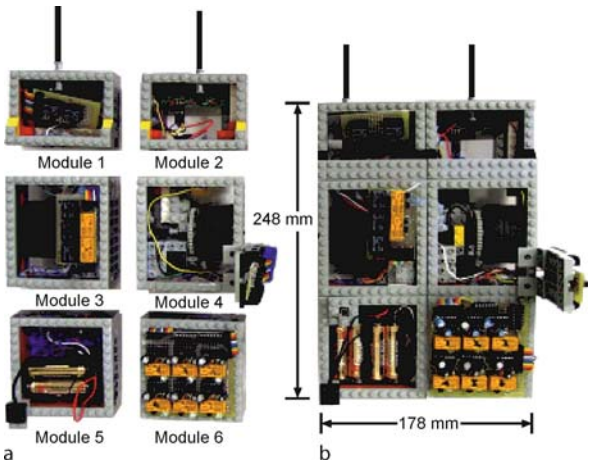

Self-replicating Robotic Systems, Figure 18
A modular self-replicating robot controlled by a finite state circuit (see [15,16])

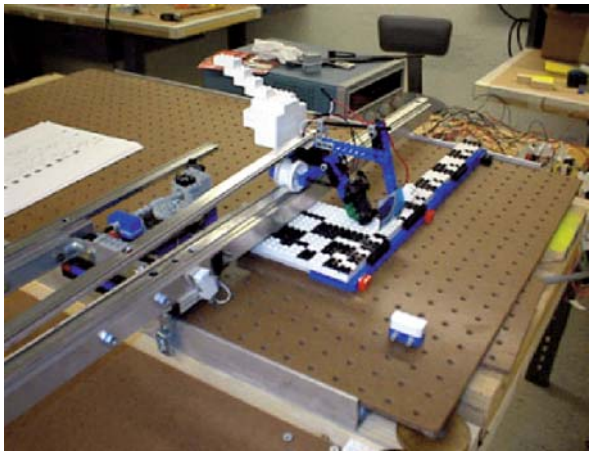

Self-replicating Robotic Systems, Figure 17
Side view of the replicating system (non-universal self-replication by self-inspection – design 2)

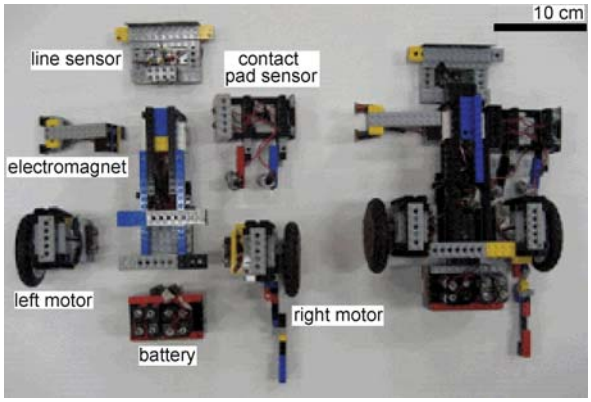

Self-replicating Robotic Systems, Figure 19
Example of another modular self-replicating robot controlled by a finite state machine (see [4,16,18] for details)

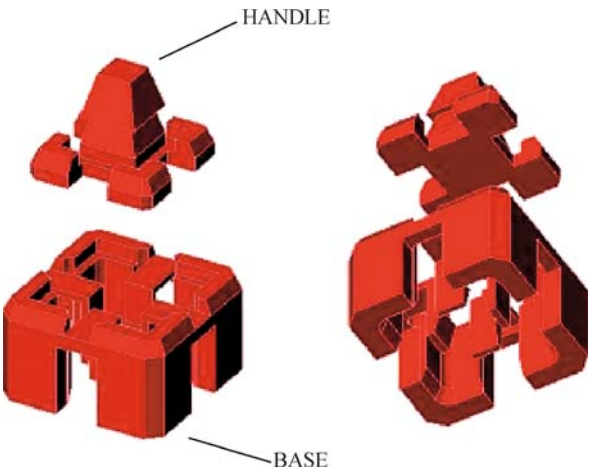

Self-replicating Robotic Systems, Figure 20
Two exploded views of the basic component. The handle and base are individually cast in polyurethane, then bonded together with epoxy

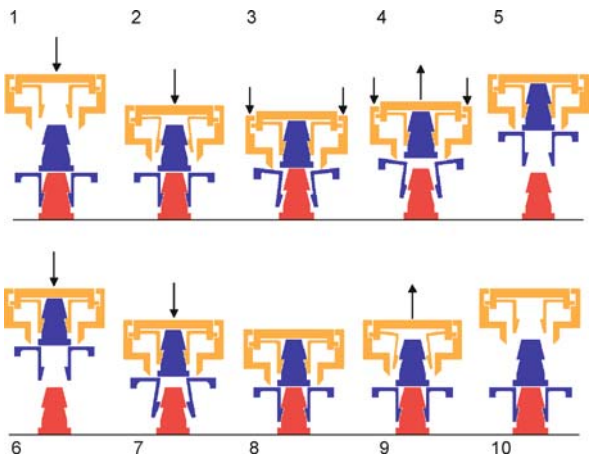

Self-replicating Robotic Systems, Figure 21
Steps in reversible assembly process. The tool (top part) grasps, retrieves, and reconnects one component to another

Towards a Universal Constructor

Whereas other prototypes of self-replicating robotic systems use relatively few subsystems as the initial parts, and connect these subsystems with magnets, the desire to create machines that can reproduce from a large number of basic parts requires thought about mechanical design, manufacturing, and assembly issues. Therefore, in this section the work of the second author in [20] is reviewed. In this work, a novel set of mechanical components that can be assembled into a wide variety of devices is presented. The components are specifically designed to be handled and assembled by devices made of the same type of components. A 3-axis Cartesian manipulator built with these components is presented. It is shown that in principle the manipulator can assemble duplicates of itself in addition to arbitrary devices when provided with properly oriented components and controlled by a human operator.

All components in the component set are variants of the "basic component". Figure 20 shows a cutaway view of two basic components. Each component consists of two parts: an upper "handle" and a lower "base". The base dimensions are 3.8 by 3.8 by 2.3 cm. The handle and base are individually cast from polyurethane resin in a silicone mold and then fixed together with epoxy adhesive. The base contains four compliant snap tangs that are complementary to the four tapered surfaces of the handle. Figure 21 shows how the snap tangs of the base grasp the undercut of the handle of another part. Additionally, the convex tapered surfaces of the base mate with concave tapered surfaces of the handle in order to provide rigidity with respect to shear and torsion between parts. Chamfers on the edges of the parts allow a certain amount of positioning error during assembly. The tangs and handle also contribute to error tolerance, since they are tapered such that two parts need not be exactly positioned before assembly. The slots cut in each side of the base allow the part to be assembled onto other parts that have reinforcing segments. The tangs are L-shaped, and the top part of each tang is exposed by a slot cut in each side of the handle.

The parts have a common handle, so they all can be manipulated by a single grasping tool. A part can be connected to another part by simply lowering it into place such that the tangs engage the handle of the other part (sliding components are slid sideways onto the handle of the other part). Depending on the grasping tool used, part connections can be reversible or irreversible. A section view of the steps in reversible assembly are shown in Fig. 21. The grasping tool (upper part) has thin tangs and a release mechanism. The process is initiated when the grasping tool is lowered onto the part to be disconnected. The release mechanism is engaged, depressing the tangs of the part to be disconnected. The upper part of the tangs deflect, disengaging the tangs of the middle part from the handle of the lower part (step 3). In step 4 the grasping tool and middle part are lifted from the lower part. To re-connect the part, the grasping tool and part are lowered onto the destination part (step 7). The grasping tool is removed simply by lifting it from the newly connected part. The snap tangs in the grasping tool are thinner than normal tangs, so they deflect first and disengage the grasping tangs from the handle on the newly connected part (step 9).

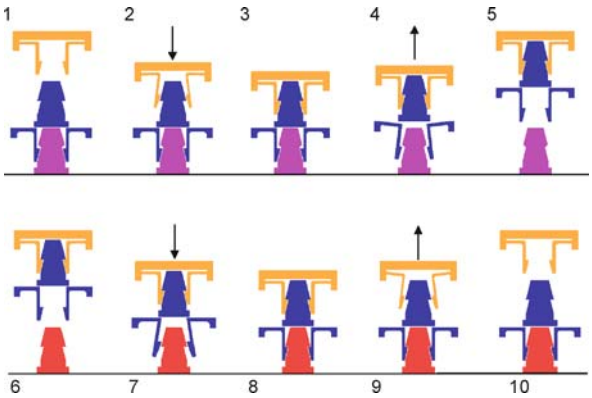

Self-replicating Robotic Systems, Figure 22
Steps in irreversible assembly process. The tool retrieves a part from a special storage site and attaches it to another component

Irreversible assembly is also possible, as shown in Fig. 22. In this case, the grasping tool has thin tangs like the reversible tool, but there is no release mechanism. Parts must be stored at a special "storage site" before assembly. The storage site (lower part in steps 1–5) is a normal part with reduced undercut on the handles. The amount of undercut on the storage site handles and the thickness of the tangs on the tool are carefully chosen so that the force required to disassemble two components is greatest between two normal parts, and least between a normal part and the storage site. This allows the grasping tool to pick a part from a storage site and connect it to another part with two simple up/down vertical movements. The irreversible assembly method, due to its simplicity, is used in the manipulator described below.

These basic components can be used to form a constructing machine capable of assembling a wide variety of devices, including duplicates of itself. By "capable" is meant kinematically capable - the machine can grasp, manipulate, and place all of the components required for constructing a duplicate of its "mechanical self". The machine's "mechanical self" includes the hardware required to manipulate mechanical components, but excludes the portion responsible for control. The machine lacks any type of sensor or onboard control, and so must be controlled either by a human operator or a sophisticated external controller. By "assemble" is meant the retrieval of components from a storage site and subsequent connection of them to an assembly or operating plane. Components must arrive at the storage site through action of another entity – namely a human operator.

A total of 16 different variations on the basic component were built. Most of these are used in the manipulator, and the remainder are useful for constructing other

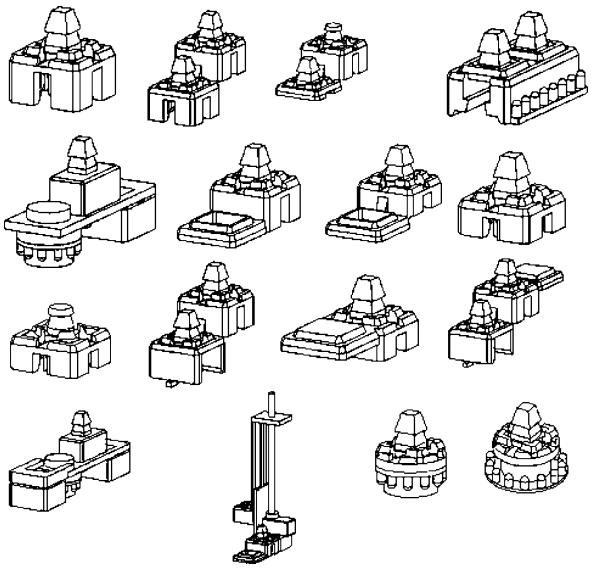

Self-replicating Robotic Systems, Figure 23
Sixteen variations on the basic component

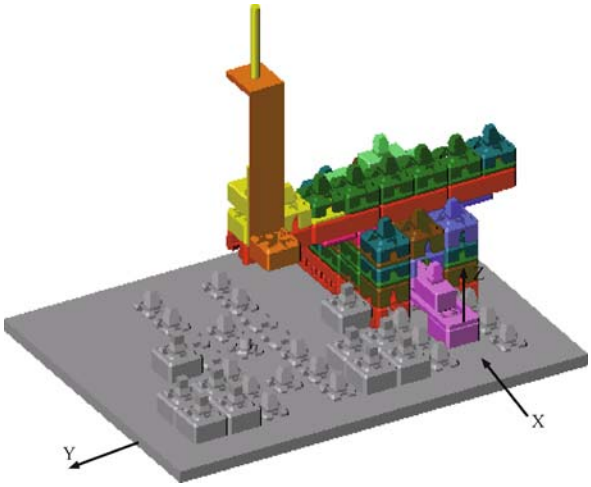

Self-replicating Robotic Systems, Figure 24
A constructing machine made from parts in the component set

types of devices. Figure 23 shows a diagram of these parts. Figure 24 shows a CAD model of a constructing machine made from parts in the component set. It is essentially a 3-axis Cartesian manipulator. The base of the machine is a platform that slides along the x-axis, and is actuated by a DC motor. A motor-driven boom rides on top of the platform, and slides along the y-axis. The end-effector is mounted to a vertically-linear motor on the end of the boom. The reachable space of the end-effector is a rectangular volume of size $3 \times 4 \times 6$ in units of component-widths. Note that the tool does not rotate. This requires

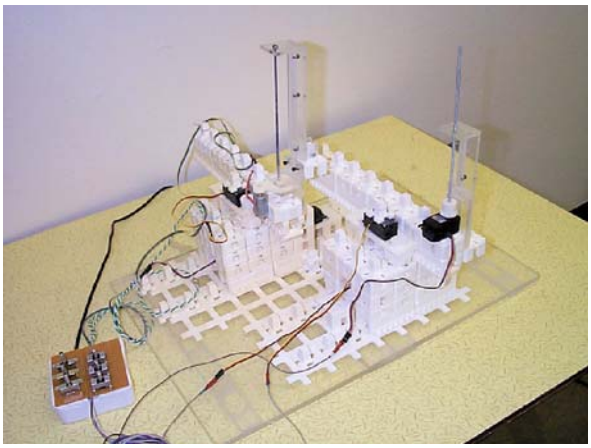

Self-replicating Robotic Systems, Figure 25
An initial physical implementation of a universal constructor [20]

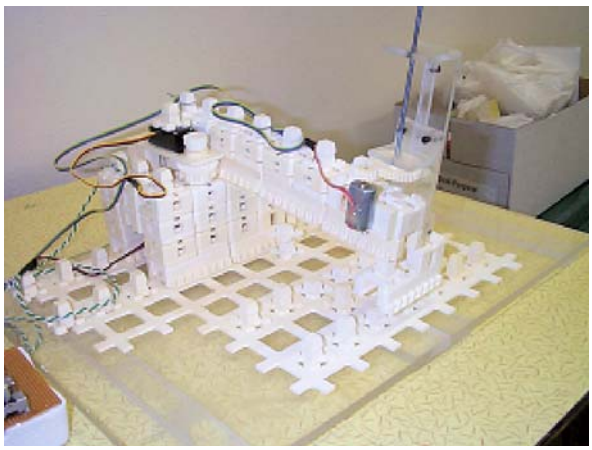

Self-replicating Robotic Systems, Figure 26
Large downward deflection of extended boom

parts to be loaded at the storage site in their proper orientation. The constructor has a total of 45 components of eleven different types, not counting the operating plane. The details of the assembly process of this machine can be found in [20].

Figure 25 shows the experimental setup in its finished state. Both devices are operable. A human operator controls them through a switchbox that turns the various motors on or off. A topic that has been neglected so far is that of running wires to the motors. As seen from the figure, this is done rather haphazardly. This prototype had a number of problems that are worth mentioning which are helping us to design better systems.

The first problems that were encountered dealt with getting the constructor boom to move smoothly. When the boom is extended, it applies high forces to the parts that hold it to the platform. This results in more friction, and requires a higher torque from the motor. In these conditions the motor tended to separate from the platform, and the drive gear would lose contact with the racks on the boom. This problem was eventually "solved" by using a metal screw and adhesive between certain parts to hold them in place.

Most of the other problems occurred during assembly, and centered around the extended boom. Figure 26 shows the large downward deflection occurring in the extended boom. This made it difficult to align parts. A human operator could successfully place parts, using repeated effort and visual feedback, but a simple controller surely could not.

Deflection of the boom in the x-direction presented problems for sliding the far racks into place (Fig. 27). This would cause the sliding platform under the boom to jam.

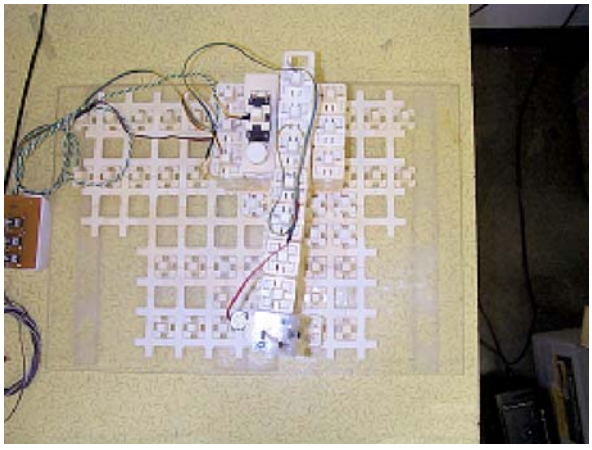

Self-replicating Robotic Systems, Figure 27
Sideways deflection of extended boom caused the sliding platform to jam

In addition to sideways deflection, the boom would also twist along the y-axis (Fig. 28). The constructor was able to place the near racks, but the far racks had to be placed by hand.

The torsion of the boom created problems in later assemblies steps as well. The twisting deflection of the boom would tend to misalign parts so that their snap-tangs would not engage with the parts below them. For this reason, the tool was mounted to the slide in a rotating bracket. This allowed the tool a small, passive amount of rotation about the y-axis. With this feature, the boom could still twist, but the parts would self-align and engage.

In many cases, the constructor could not develop the necessary force to assemble components. This was especially true when assembling components with multiple sets

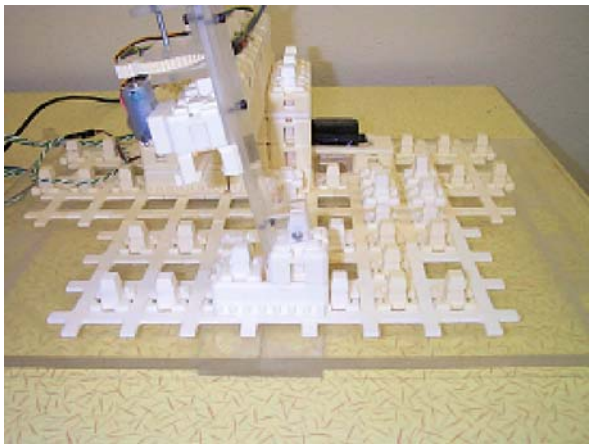

Self-replicating Robotic Systems, Figure 28
Torsion of boom

Self-replicating Robotic Systems, Figure 30
Separation of components within the boom

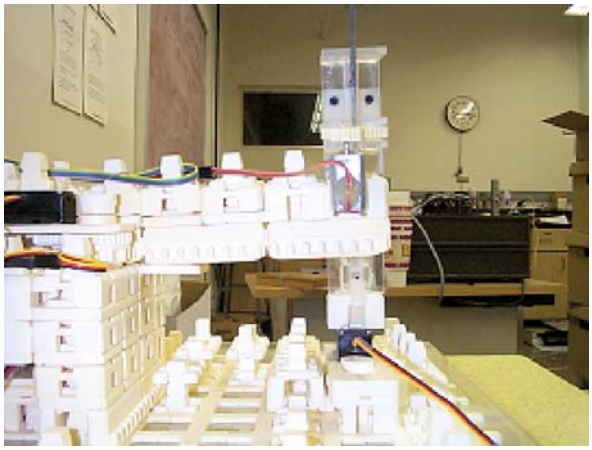

Self-replicating Robotic Systems, Figure 29
Upward deflection of boom

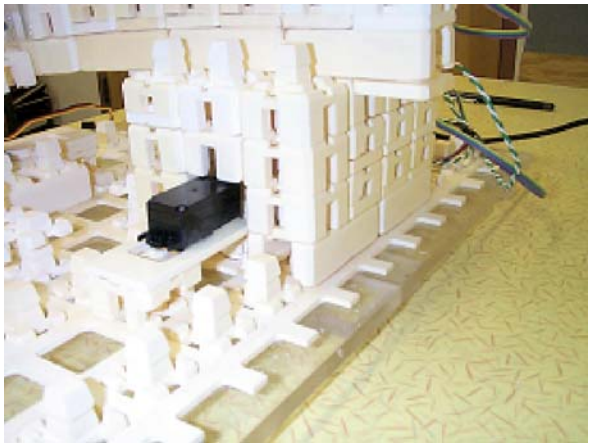

Self-replicating Robotic Systems, Figure 31
Separation of sliding platform and operating plane

of snap tangs, and in parts where the tangs were offset horizontally from the handle. Sometimes this occurred because the vertical linear actuator did not produce enough force, and sometimes it occurred because upward deflection of the boom caused the parts to misalign (Fig. 29). In a few cases, components in the boom would separate (Fig. 30). This only occurred when the boom was prevented from deflecting because it was engaged on the growing assembly.

The extended boom indirectly caused another problem. Recall that two of the handles in the tracks on the operating plane are modified as entry sites for racks. These sites do not help hold the platform to the plane. This, in combination with the high moment exerted on the platform by the extended boom, would cause the platform to separate from the plane and get stuck on adjacent, non-modified handles (Fig. 31).

Some aspects of the constructor worked well, such as the stage-by-stage assembly of the boom and the retrieval of parts from the storage site. The boom does not have to extend to reach the storage site, so deflections are not a problem and part retrieval was easy. The thickness of the tool tangs was found by trial and error – the tangs of a basic part were successively machined until the disassembly force fell within the correct range. Although no measurements were taken, the distribution of assembly and disassembly forces across various parts was fairly tight. The tool could lift all of the parts. Once parts were connected to an assembly, the tool could be separated from the connected parts in all cases.

In short, this initial concept of a desktop universal constructor provided us many valuable lessons about the importance of mechanical design, parts assembly and manufacturability issues that we are incorporating in future designs.

Discussion

Several self-replicating robot prototype have been constructed and tested. An autonomous prototype uses two light sensors in its navigation system to detect objects and also to track lines. Magnets and shape-constraining blocks are used to aid in aligning and interlocking the subsystems of the replica. As a result, the robot is capable of automatically assembling its replicas. All of the replicas are also capable of completing the same replicating process. The autonomous self-replicating system presented here has been recognized by authors of the book, "Kinematic Self-Replicating Machines" [25] that this prototype is the world's first fully functional autonomous self-replicating robot. Self-replicating circuits, as well as self-replicating robots whose controllers can be fully decomposed into basic parts have also been reviewed here.

Whereas von Neumann's architecture for self-replicating kinematic automata is the most widely known approach, it is not the only one. Self-reproduction by self-inspection in which a non-universal constructor "reads" an original device and "writes" a copy by executing a very small set of hardwired commands is an alternative. In our experience observing students attempting to build self-replicating devices, self-replication by self-inspection appears to be a more robust and less complicated alternative to the universal constructor.

Future Works

In previous sections we discussed a series of mechanical replicating prototypes in which programs are preloaded onto control computers which are then treated as one of several subsystems to be assembled. However, our ultimate goal is to develop a self-replicating robotic system capable of autonomously assembling its replicas from simple components using only electro-mechanical intelligence, i. e. a mechanical code, and transistor-based control circuits. This eliminates complicated electronic components, such as programmable micro-controllers, and makes the concept more appropriate for future space systems that can use in-situ resources for self-replication. The prototypes reviewed here are one step in this direction. Other related technologies are discussed in [33].

Acknowledgments

This work was made possible by support from the Royal Thai government, the Mellon Foundation, and the National Science Foundation under Grant ITS 0205466. The results and opinions expressed are solely those of the authors. We thank A. Cushing, Y. Kwon, K. Lee, M. Kutzer and all of the undergraduate students in the Mechatronics courses over the past five years who have designed several of the prototypes shown in this review. The work presented in Sect. "Towards a Universal Constructor" was performed at University of New Mexico with help from Professors G. Starr, H. Tran, and J. Wood.

Bibliography

1. Chirikjian GS, Pamecha A, Ebert-Uphoff I (1996) Evaluating Efficiency of Self-Reconfiguration in a Class of Modular Robots. J Robot Syst 13(5):317–338
2. Chirikjian GS, Suthakorn J (2002) Towards Self-Replicating Robots. In: Proceedings of the Eight International Symposium on Experimental Robotics (ISER), Italy, July 2002
3. Chirikjian GS, Zhou Y, Suthakorn J (2002) Self-Replicating Robots for Lunar Development. IEEE/ASME Trans Mechatron 7(4):462–472, December 2002
4. Eno S, Mace L, Liu J, Benson B, Raman K, Lee K, Moses M, Chirikjian GS (2007) Robotic Self-Replication in a Structured Environment without Computer Control. In: Proceedings of the 2007 IEEE CIRA. Piscataway
5. Freitas RA Jr (1980) A Self-Reproducing Interstellar Probe. J Br Interplanet Soc 33:251–264
6. Freitas RA Jr (1980) Report on the NASA/ASEE Summer Study on Advanced Automation for Space Missions. J Br Interplanet Soc 34:139–142
7. Freitas RA Jr (1983) Terraforming Mars and Venus Using Self-Replicating Systems. J Br Interplanet Soc 36:139–142
8. Freitas RA Jr, Merkle RC (2004) Kinematic Self-Replicating Machines. Landes Bioscience, Georgetown
9. Freitas RA Jr, Valdes F (1980) Comparison of Reproducing and Non-Reproducing Starprobe Strategies for Galactic Exploration. J Br Planet Soc 33:402–408
10. Freitas RA Jr, WP Gilbreath (eds) (1982) Advanced Automation for Space Missions. In: Proceedings of the 1980 NASA/ASEE summer study. Replicating Systems Concepts: Self-Replicating Lunar Factory and Demonstration, NASA, Scientific and Technical Information Branch (Conference Publication 2255), US Government Printing Office, Washington DC, chap 5
11. Hastings WA, Labarre M, Viswanathan A, Lee S, Sparks D, Tran T, Nolin J, Curry R, David M, Huang S, Shuthakorn J, Zhou Y, Chirikjian GS (2004) A minimalist parts manipulation system for a self replicating electromechanical circuit. IMG'04, Genoa, Italy, July 1–2, 2004
12. Hosokawa K, Fujii T, Kaetsu H, Asama H, Kuroda HY, Endo I (1999) Self-organizing collective robots with morphogenesis in a vertical plane. JSME Int J Ser C-Mechanical Syst Mach Elements Manuf 42(1):195–202
13. Jacobson H (1958) On Models of Reproduction. Am Scient 46:255–284

14. Kotay K, Rus D, Vona M, McGray C (1998) The Self-reconfiguring Molecule: Design and Control Algorithms. WAFR'98: Proceedings of the third workshop on the algorithmic foundations of robotics. Publisher A.K. Peters, Ltd. Wellesley
15. Lee K, Chirikjian GS (2007) Robotic Self-Replication. Low Complexity Parts. IEEE Robotics and Automation Magazine 14(4):34–43. Published by IEEE, Piscataway
16. Lee K, Moses M, Chirikjian GS (2008) Robotic Self-Replication in Structured Environments: Physical Demonstrations and Complexity Measures. Int J Robot Res 27(3-4):387–401. doi:10.1177/0278364907084982
17. Lipson H, Pollack B (2000) Automatic design and Manufacture of Robotic Lifeforms. Nature 406:974–978
18. Liu A, Sterling M, Kim D, Pierpont A, Schlothauer A, Moses M, Lee K, Chirikjian GS (2007) A Memoryless Robot that Assembles Seven Subsystems to Copy Itself. In: Proceedings of the 2007 IEEE ISAM, Piscataway
19. Moore EF (1956) Artificial Living Plants. Scient Am 195:118–126
20. Moses M (2001) A Physical Prototype of a Self-Replicating Universal Constructor. M.S. University of New Mexico
21. Murata S, Kurokawa H, Kokaji S (1994) Self-Assembling Machine. In: Proceedings of the 1994 IEEE International Conference on Robotics and Automation. Piscataway, San Diego, CA, pp 441–448
22. Penrose LS (1959) Self-Reproducing Machines. Sci Am 200(6):105–114
23. Pesavento U (1995) An implementation of von Neumann's self-reproducing machine. Artif Life J 2(4):337–354
24. Pfeifer R, Kunz H, Weber MM, Thomas D (2001) Lecture of the Artificial Life. http://www.ifi.unizh.ch/ailab/teaching/AL01/chap7.pdf. Dept of Information Technology, University of Zurich. Accessed 19 June 2008
25. Rebek J Jr (1994) Synthetic Self-Replicating Molecules. Sci Am 271(1):48–55
26. Sipper M (1998) Fifty Years of Research on Self-Replication: An Overview. Artif Life 4(3):237–257
27. Suthakorn J (2004) Paradigm for Service Robotics. Ph D Dissertation. Johns Hopkins University
28. Suthakorn J, Kwon YT, Chirikjian GS (2003) A Semi-Autonomous Replicating Robotic System. In: Proceedings of the 2003 IEEE/ASME International Symposium on Computational Intelligence for Robotics and Automation (CIRA), Kobe, Japan, Piscataway
29. Suthakorn J, Zhou Y, Chirikjian GS (2002) Self-Replicating Robots for Space Utilization. In: Proceedings of the 2002 Robophere workshop on Self Sustaining Robotic Ecologies, NASA Ames Research Center, California, NASA Ames Research Center
30. Tiesenhausen GV, Darbro WA (1980) Self-Replicating Systems – A Systems Engineering Approach. In: Technical Memorandum: NASA TM-78304, Washington DC
31. Von Neumann J, Burks AW (1966) Theory of Self-Reproducing Automata. University of Illinois Press,Champaign
32. Whitesides GM (1995) Self-Assembling Materials. Sci Am 273(3):146–149
33. Yim M, Shen WM, Salemi B, Rus D, Moll M, Lipson H, Klavins E, Chirikjian GS (2007) Modular self-reconfigurable robot systems – Challenges and opportunities for the future. IEEE Robotics Autom Mag 14(1):43–52
34. Yim M, Zhang Y, Lamping J, Mao E (2001) Distributed Control for 3D Metamorphosis. Auton Robots 10:41–56

Self-Replication and Cellular Automata

Gianluca Tempesti[1], Daniel Mange[2], André Stauffer[2]

[1] University of York, York, UK

[2] Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Article Outline

Glossary
Definition of the Subject
Introduction
Von Neumann's Universal Constructor
Self-Replication for Artificial Life
Other Approaches to Self-Replication
Future Directions
Bibliography

Glossary

Cellular automaton A cellular automaton (CA) is a mathematical framework modeling an array of *cells* that interact locally with their neighbors. In this *cellular space*, each cell has a set of *neighbors*, cells have *values* or *states*, all the cells update their values simultaneously at discrete *time steps* or *iterations*, and the new state of a cell is determined by the current state of its neighbors (including itself) according to a local *function* or *rule*, identical for all cells. In the article, the term is extended to account for systems that introduce variations to the basic definition (for example, systems where cells do not update simultaneously or do not have the same set of rules in every cell).

Following the historical pattern, in the article the same term is also used to refer to an object or structure built within the cellular space, i. e., a set of cells in a particular, usually active, state (overlapping with the definition of *Configuration*).

Configuration A set of cells in a given state at a given time. Usually, but not always, the term refers to the state of all the cells in the entire space. The *initial configuration* is the state of the cells at time $t = 0$.

Self-replication The process whereby a cellular automaton configuration creates a copy of itself in the cellular space. Incidentally, you will note that in the article we use the terms self-replication and self-reproduction interchangeably. In reality, the two terms are not really synonyms: self-reproduction is more properly applied to the reproduction of organisms, while self-replica-

tion concerns the cellular level. The more correct term to use in most cases would probably be self-replication, but since von Neumann favored self-reproduction, we will ignore the distinction.

Self-reproduction See *Self-Replication*

Construction The process that occurs when one or more cells, initially in the inactive or *quiescent* state are assigned an active state (in the context of this article, by the self-replicating structure).

Definition of the Subject

Machine self-replication, besides inspiring numerous fictional books and movies, has long been considered a powerful paradigm to allow artifacts, for example, to survive in hostile environments (such as other planets) or to operate more efficiently by creating *populations* of machines working together to achieve a given task. Where the self-replication of *computing* machines is concerned, other motivations can also come into play, related to concepts such as fault tolerance and self-organization.

Cellular automata have traditionally been the framework of choice for the study of self-replicating computing machines, ever since they were used by John von Neumann, who pioneered the field in the 1950s. In this context, self-replication is seen as the process whereby a configuration in the cellular space is capable of creating a copy of itself in a different location.

As a mathematical framework, CA allow researchers to study the mechanisms required to achieve self-replication in a simplified environment, in view of eventually applying this process to real-world systems, either to electronics or, more generally, to computing systems.

Introduction

The self-replication of computing systems is an idea that dates back to the very origins of electronics. One of the pioneers of the field, John von Neumann, was among the first to investigate the possibility of creating machines capable of self-replication [1] with the purpose of achieving reliability through the redundant operation of "populations" of computing machines.

Throughout the more than 50 years since von Neumann's seminal work, research on this topic has gone through several transformations. While interest in applying self-replication to electronic systems waned because of technological hurdles, the field of Artificial Life, starting with the pioneering work of Chris Langton [14], began studying this process in the more general context of achieving life-like properties in artificial systems.

Throughout its long history, cellular automata (CA) have remained one of the environments of choice to study how self-replication can be applied to computing systems. In general, researchers in the domain (including von Neumann) have never regarded CA as the environment in which self-replication would be ultimately applied. Rather, CA have traditionally provided a useful platform to test the complexity of self-replication at an early stage, in view of eventually applying this process to real-world systems, either to electronics or, more generally, to computing systems.

Of course, the concept of self-replication has been applied to artificial systems in contexts other than computing. A classic example is the 1980 NASA study by Robert Freitas Jr. and Ralph Merkle [10] (recently expanded in a remarkable book [11]), where self-replication is used as a paradigm for efficiently exploring other planets. However, this kind of self-replication, applied to physical machines rather then computing systems, does not commonly make use of cellular automata and is beyond the scope of this article.

Following the historical progress of self-replication in cellular automata (derived in part from [40]), we will first examine in some detail von Neumann's seminal work (Sect. "Von Neumann's Universal Constructor"). Then, the use of self-replication as an Artificial Life paradigm will be discussed (Sect. "Self-Replication for Artificial Life") before dealing with some of the latest advances in the field in Sect. "Other Approaches to Self-Replication".

Von Neumann's Universal Constructor

Many of the existing approaches to the self-replication of computing systems are essentially derived from the work of John von Neumann [1], who pioneered this field of research in the 1950s.

Von Neumann, confronted with the lack of reliability of computing systems, turned to nature to find inspiration in the design of fault-tolerant computing machines. Let us remember that the computers von Neumann was familiar with were based on vacuum-tube technology, and that vacuum tubes were much more prone to failure than modern transistors. Moreover, since the writing and the execution of complex programs on such systems represented many hours (if not many days) of work, the failure of a system had important consequences in wasted time and effort.

In particular, Von Neumann investigated self-replication as a way to design and implement digital logic devices. Unfortunately, the state of the art in the fifties restricted von Neumann's investigations to a purely theoretical level, and the work of his successors mirrored this constraint.

Indeed, it is not until fairly recently that some of the technological problems associated with the implementation of such a process in silicon have been resolved with the introduction of new kinds of electronic devices (see Sect. "Other Approaches to Self-Replication").

In this section, we will analyze von Neumann's research on the subject of self-replicating computing machines, and in particular his universal constructor, a self-replicating cellular automaton [44].

Von Neumann's Self-Replicating Machines

Natural systems are among the most reliable complex systems known to man, and their reliability is a consequence not of any particular robustness of the individual cells (or organisms), but rather of their extreme redundancy. The basic natural mechanism which provides such reliability is *self-reproduction*, both at the cellular level (where the survival of a single organism is concerned) and at the organism level (where the survival of the species is concerned).

Thus von Neumann, drawing inspiration from natural systems, attempted to develop an approach to the realization of self-replicating computing machines (which he called *artificial automata*, as opposed to natural automata, that is, biological organisms). In order to achieve his goal, he imagined a series of five distinct models for self-reproduction ([44], pp. 91–99):

1. The *kinematic* model, introduced by von Neumann on the occasion of a series of five lectures given at the University of Illinois in December 1949, is the most general. It involves structural elements such as sensors, muscle-like components, joining and cutting tools, along with logic (switch) and memory elements. Concerning, as it does, physical as well as electronic components, its goal was to define the bases of self-replication, but was not designed to be implemented.
2. In order to find an approach to self-replication more amenable to a rigorous mathematical treatment, von Neumann, following the suggestion of the mathematician S. Ulam, developed a *cellular* model. This model, based on the use of cellular automata as a framework for study, was probably the closest to an actual realization. Even if it was never completed, it was further refined by von Neumann's successors and was the basis for most further research on self-replication.
3. The *excitation-threshold-fatigue* model was based on the cellular model, but each cell of the automaton was replaced by a neuron-like element. Von Neumann never defined the details of the neuron, but through a careful analysis of his work, we can deduce that it would have borne a fairly close relationship to today's simplest artificial neural networks, with the addition of some features which would have both increased the resemblance to biological neurons and introduced the possibility of self-replication.
4. For the *continuous* model, von Neumann planned to use differential equations to describe the process of self-reproduction. Again, we are not aware of the details of this model, but we can assume that von Neumann planned to define systems of differential equations to describe the excitation, threshold and fatigue properties of a neuron. At the implementation level, this would probably correspond to a transition from purely digital to analog circuits.
5. The *probabilistic* model is the least well-defined of all the models. We know that von Neumann intended to introduce a kind of automaton where the transitions between states were probabilistic rather than deterministic. Such an approach would allow the introduction of mechanisms such as mutation and thus of the phenomenon of evolution in artificial automata. Once again, we cannot be sure of how von Neumann would have realized such systems, but we can assume they would have exploited some of the same tools used today by genetic algorithms.

Of all these models, the only one von Neumann developed in some detail was the cellular model. Since it was the basis for the work of his successors, it deserves to be examined more closely.

Von Neumann's Cellular Model

In von Neumann's work, self-reproduction is always presented as a special case of universal construction, that is, the capability of building any machine given its description (Fig. 1). This approach was maintained in the design of his cellular automaton, which is therefore much more than a self-replicating machine. The complexity of its purpose is reflected in the complexity of its structure, based on three separate components:

1. A *memory tape*, containing the description of the machine to be built, in the form of a one-dimensional string of elements. In the special case of self-reproduction, the memory contains a description of the universal constructor itself (Fig. 2). It is interesting to note that the memory of von Neumann's automaton bears a strong resemblance to the biological genome. This resemblance is even more remarkable when considering that the structure of the genome was not discovered until after the death of von Neumann.

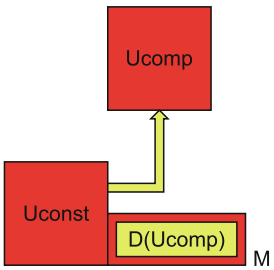

Self-Replication and Cellular Automata, Figure 1
Von Neumann's universal constructor `Uconst` can build a specimen of any machine (e. g., a universal Turing machine `Ucomp`) given its description `D(Ucomp)`

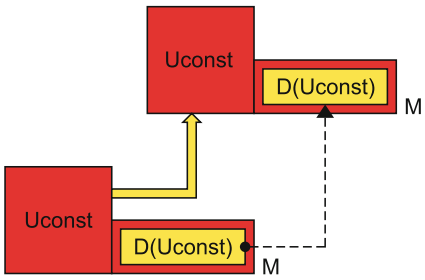

Self-Replication and Cellular Automata, Figure 2
Given its own description `D(Uconst)`, von Neumann's universal constructor is capable of self-replication

2. The *constructor* itself, a very complex machine capable of reading the memory tape and interpreting its contents.
3. A *constructing arm*, directed by the constructor, used to build the offspring (i. e., the machine described in the memory tape). The arm moves across the space and sets the state of the elements of the offspring to the appropriate value.

The implementation as a cellular automaton is no less complex. Each element has 29 possible states, and thus, since the next state of an element depends on its current state and that of its four cardinal neighbors, $29^5 = 20,511,149$ transition rules are required to exhaustively define its behavior.

If we consider that the size of von Neumann's constructor is of the order of several hundred thousand elements, we can easily understand why a hardware realization of such a machine is not really feasible. In fact, as part of our research, we did realize a hardware implementation of a set of elements of von Neumann's automaton [3,35]. By carefully designing the hardware structure of each element, we were able to considerably reduce the amount of memory required to host the transition rules. Nevertheless, our system remains a demonstration unit, as it consists of a few elements only, barely enough to illustrate the behavior of a tiny subset of the entire machine.

It is also worth mentioning that von Neumann went one step further in the design of his universal constructor. If we consider the universal constructor from a biological viewpoint, we can associate the memory tape with the genome, and thus the entire constructor with a single cell (which would imply a parallel between the automaton's elements and molecules). However, the constructor, as we have described it so far, has no functionality outside of self-reproduction. Von Neumann recognized that a self-replicating machine would require some sort of functionality to be interesting from an engineering point of view, and postulated the presence of a *universal computer* (in practice, a universal Turing machine, an automaton capable of performing any computation) alongside the universal constructor (Fig. 3). Von Neumann's constructor can thus be regarded as a unicellular organism, containing a genome stored in the form of a memory tape, read and interpreted by the universal constructor (the mother cell) both to determine its operation and to direct the construction of a complete copy of itself (the daughter cell).

Von Neumann's Successors

The extreme size of von Neumann's universal constructor has so far prevented any kind of physical implementation (apart from the small demonstration unit we mentioned). But further, even the simulation of a cellular automaton of such complexity was far beyond the capability of early computer systems. Today, such a simulation is starting to be conceivable. Umberto Pesavento, a young Italian high school student, developed a simulation of von Neumann's entire universal constructor [27]. The computing power available did not allow him to simulate either the entire self-replication process (the length of the memory tape needed to describe the automaton would have required too large an array) or the Turing machine necessary to implement the universal computer, but he was able to demonstrate the full functionality of the constructor.

Considering the rapid advances in computing power of modern computer systems, we can assume that a complete simulation could be envisaged with today's technology. In fact, an effort is currently under way [4] to implement a complete specimen of the constructor. To achieve this goal, Buckley is revisiting and analyzing in detail the operation of the constructor. To give an idea of the scope of this work, Buckley's results indicate that the interpretercopier (without the tape) is bounded by a region of $751 \times 1048 = 787.048$ cells.

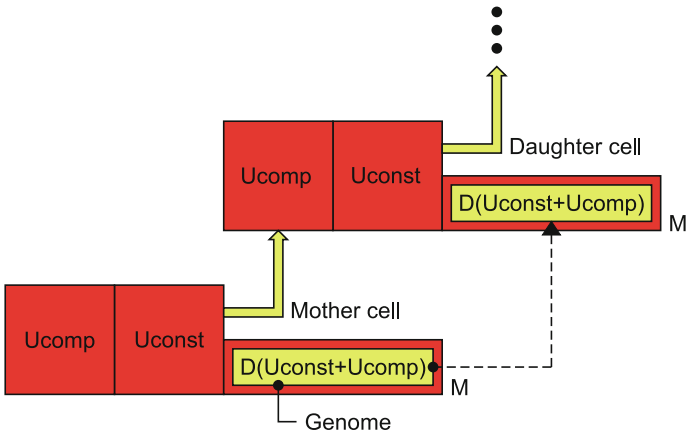

Self-Replication and Cellular Automata, Figure 3
By extension, von Neumann's universal constructor can include a universal computer and still be capable of self-replication

The impossibility of achieving a physical realization did not however deter some researchers from trying to continue and improve von Neumann's work [2,15,22]. Arthur Burks, for example, in addition to editing von Neumann's work on self-replication [5,44], also made several corrections and advances in the implementation of the cellular model. Codd [9], by altering the states and the transition rules, managed to simplify the constructor by a considerable degree. Vitanyi [43] studied the possibility of introducing sexual reproduction in von Neumann's approach. However, without in any way lessening these contributions, we can say that no major theoretical advance in the research on self-reproducing automata occurred until C. Langton, in 1984, opened a second stage in this field of research.

Self-Replication for Artificial Life

While the *implementation* of von Neumann's universal constructor faced insurmountable (at the time) technological hurdles, the same could not be said of the *theoretical* contribution that his approach represented as an attempt to study a biologically-inspired process within the world of computing systems.

In this context, the main drawback of von Neumann's work lay in the inability to achieve self-replication without resorting to an extremely complex simulation of a complete machine. Von Neumann's Universal Constructor was so complex because it tried to implement self-reproduction as a particular case of construction universality, i. e. the capability of constructing any other automaton, given its description. C. Langton approached the problem somewhat differently, by attempting to define the simplest cellular automaton, commonly known as *Langton's loop* [14], capable exclusively of self-reproduction.

Langton's loop had a major impact on research in self-replication by introducing a new way to think about this process in more "abstract" terms as a study of the application of biologically-inspired mechanisms to computing, exemplifying the field known as *Artificial Life*. In this context, rather than the replicating machine, it is the process of self-replication itself that becomes the object of study.

This novel approach generated research that can be considered fundamentally different from that of von Neumann and started discussion on topics such as the analogy with cellular division and with the reproduction of individuals in a population (e.g, in [17], Section V.B), the difference between trivial and non-trivial self-replication in cellular automata (e. g., in automata such as those described in [16]), or the connections between evolution and self-replication (e.g, in the work of Sayama [17]).

Langton's Loop

As a consequence of his approach, Langton's Loop is orders of magnitude simpler than von Neumann's constructor. In fact, it is loosely based on one of the simplest organs (an organ in this context can be seen as a self-supporting structure capable of a single sub-task) in Codd's automaton: the *periodic emitter* (itself derived from von Neumann's periodic pulser), a relatively simple structure capable of generating a repeating string of a given sequence of pulses.

Langton's loop (Fig. 4) is named after the dynamic storage of data inside a square sheath (red in the figure). The data is stored as a sequence of instructions for directing the constructing arm, coded in the form of a set of three states. The data turns counterclockwise in permanence within the sheath, creating a loop.

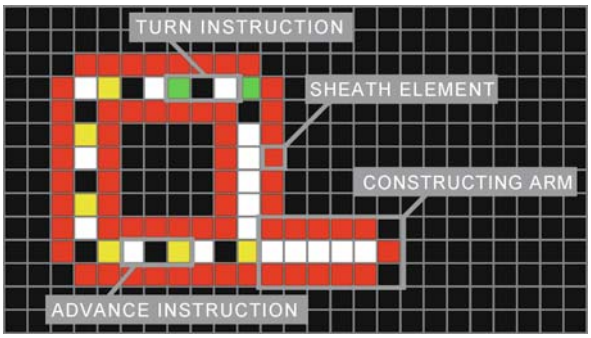

Self-Replication and Cellular Automata, Figure 4
The initial configuration of Langton's Loop (iteration 0)

The two instructions in Langton's loop are extremely simple. One instruction (uniquely identified by the yellow element in the figure) tells the arm to advance by one position (Fig. 5), while the other (green in the figure) directs the arm to turn 90 degrees to the left (Fig. 6). Obviously, after three such turns, the arm has looped back on itself (Fig. 7), at which stage a messenger (the pink element) starts the process of severing the connection between the parent and the offspring, thus concluding the replication process.

Once the copy is over, the parent loop proceeds to construct a second copy of itself in a different direction (to the north in this example), while the offspring itself starts to reproduce (to the east in this example). The sequential nature of the self-reproduction process generates a spiraling pattern in the propagation of the loop through space (Fig. 8): as each loop tries to reproduce in the four cardinal directions, it finds the place already occupied either by its parent or by the offspring of another loop, in which case it dies (the data within the loop is destroyed).

Langton's loop uses 8 states for each of the 86 non-quiescent cells making up its initial configuration, a 5-cell neighborhood, and a few hundred transition rules (the exact number depends on whether default rules are used and whether symmetric rules are included in the count). Further simplifications to Langton's automaton were introduced by Byl [6], who eliminated the internal sheath and reduced the number of states per cell, the number of transition rules, and the number of non-quiescent cells in the initial configuration. Reggia et al. [29] managed to remove also the external sheath, thus designing the smallest self-replicating loop known to date. Given their modest complexity, at least relative to von Neumann's automaton, all of the mentioned automata have been thoroughly simulated.

Langton's loop has been used as the basis for several approaches, mostly aimed at studying the properties of self-replication within a cellular system in the context of artificial life. Sayama [32] introduced structural dissolution (whereby a loop can destroy itself, in addition to replicating) to obtain colonies of loops that are dynamically stable and exhibit a potentially evolvable behavior. Nehaniv [21] extended Langton's approach to asynchronous cellular automata, while Sipper [34] developed a self-replicating loop in a non-uniform CA (i. e., a CA where the transition rules are not necessarily identical in all cells).

Perrier's Loop

In the context of applying self-reproduction to the replication of computing machines, and hence return to von Neumann's original goals, the main weakness of Langton's loop resides in the absence of any functionality beyond self-reproduction itself. To overcome this limitation, Per-

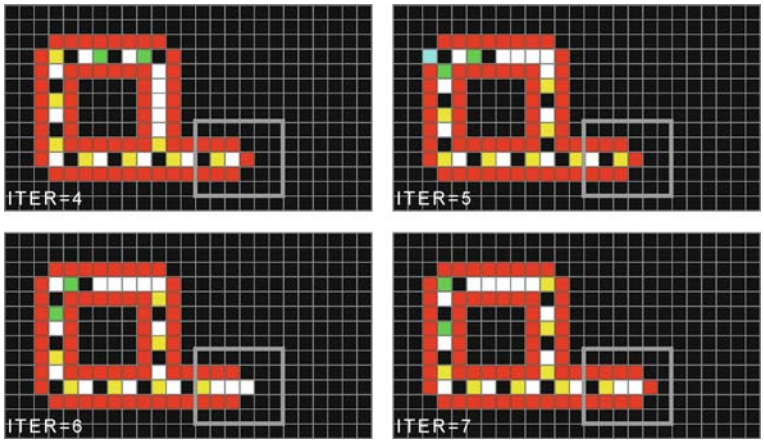

Self-Replication and Cellular Automata, Figure 5
The constructing arm advances by one space

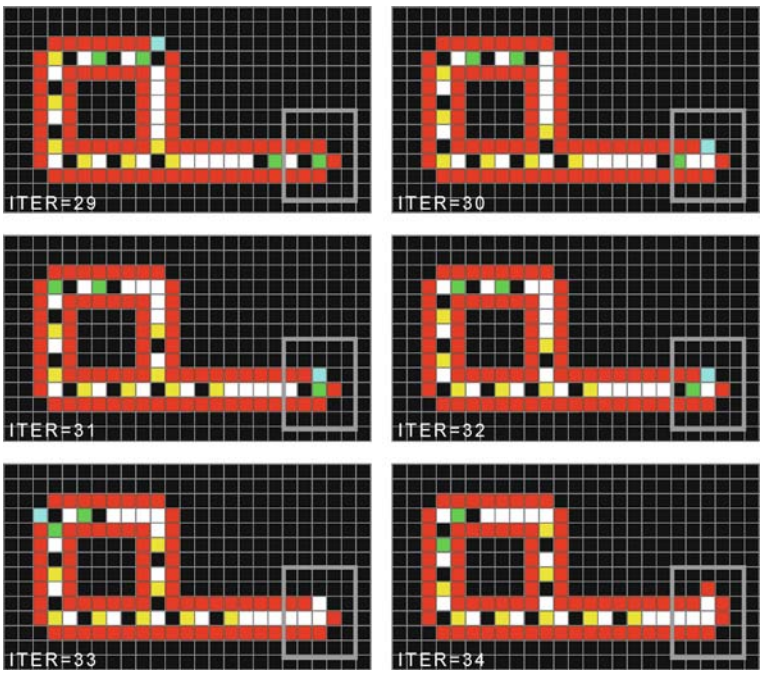

Self-Replication and Cellular Automata, Figure 6
The constructing arm turns 90 degrees to the left

rier and Zahnd developed a relatively complex automaton (Fig. 9) in which a two-tape Turing machine was appended to Langton's loop [26].

This automaton exploits Langton's loop as a sort of "carrier" (Fig. 10): the first operation of Perrier's loop is to allow Langton's loop to build a copy of itself (iteration 128: note that the copy is limited to one dimension, since the second dimension is taken up by the Turing machine). The main function of the offspring is to determine the location of the copy of the Turing machine (iteration 134). Once the new loop is ready, a "messenger" runs back to the parent loop and starts to duplicate the Turing machine (iterations 158 and 194), a process completely disjoint from the operation of the loop. When the copy is finished (iteration 254), the same messenger activates the Turing machine in the parent loop (the machine had to be inert during the replication process in order to obtain a perfect copy). The process is then repeated in each offspring until the space is filled (iteration 720: as the automaton exploits Langton's loop for replication, meeting the boundary of the array causes the last machine to crash).

Perrier's automaton implements a self-replicating Turing machine, a powerful construct which is unfortunately handicapped by its complexity: in order to implement a Turing machine, the automaton requires a very considerable number of additional states (more than 60), as well as an important number of additional transition rules. This kind of complexity, while still relatively minor compared to von Neumann's universal constructor, is nevertheless too important to be really considered for an actual implementation.

Tempesti's Loop

Always in the context of achieving self-reproduction of computing machines, and beside the lack of functionality mentioned in Subsect. "Perrier's Loop", another problem of Langton's loop is that it is not well adapted to finite CA arrays. Its self-reproduction mechanism assumes that there is enough space for a copy of the loop, and the entire loop becomes nonfunctional otherwise (Fig. 8).

To overcome this limitation and move a step closer to the realization of self-replicating machines, we developed a self-replicating loop designed specifically to exist in a finite, but arbitrarily large, space, and at the same time capable, unlike Langton's loop, to have a functionality in addition to self-replication.

In designing our self-replicating automaton [39,40], we did maintain some of the more interesting features of Langton's loop. In particular, we preserved the structure based on a square loop to dynamically store information. Such storage is convenient in CA because of the locality of

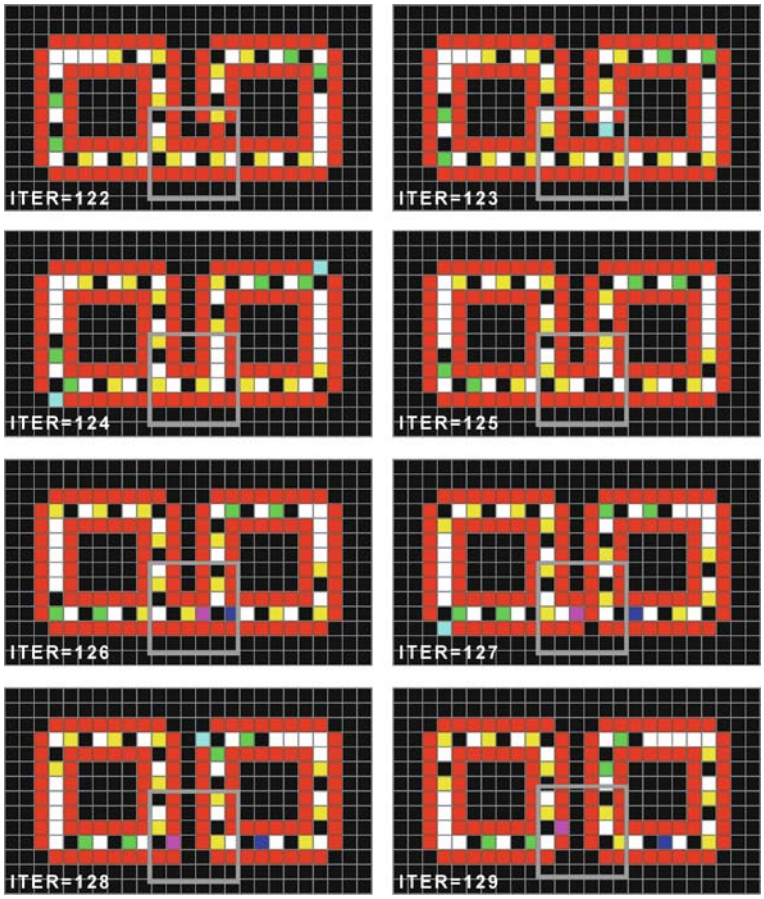

Self-Replication and Cellular Automata, Figure 7
The copy is complete and the connection from parent to offspring is severed

the rules. Also, we maintained the concept of constructing arm, in the tradition of von Neumann and his successors, even if we introduced considerable modifications to its structure and operation. While preserving some of the more interesting features of Langton's loop, we nevertheless introduced some basic structural alterations (Fig. 11):

- As in Byl's version of Langton's loop, we use only one sheath. In addition, four *gate elements* (in the same state as the sheath) at the four corners of the automaton enable or disable the replication process.
- We extend four constructing arms in the four cardinal directions at the same time, and thus create four copies of the original automaton in the four directions in parallel. When the arm meets an obstacle (either the boundary of the array or an existing copy of the loop), it simply retracts and puts the corresponding gate element in the closed position.
- The arm does not immediately construct the entire loop. Rather, it constructs a sheath of the same size as the original. Once the sheath is ready, the data circulating in the loop is duplicated and the copy is sent along the constructing arm to wrap around the new sheath. When the new loop is completed, the constructing arm retracts and closes the gate.
- As a consequence, we use only four of the elements circulating in the loop to control the self-replication process. The others can be used as a "program", i. e., a set of states with their own transition rules which will then be applied alongside the self-reproduction to execute some function.
- Unlike Langton's loop, our loop does not "die" once duplication is achieved, as the circulating data remains untouched by the self-reproduction process. This feature is a requirement for implementing functions which work after the copy has finished.

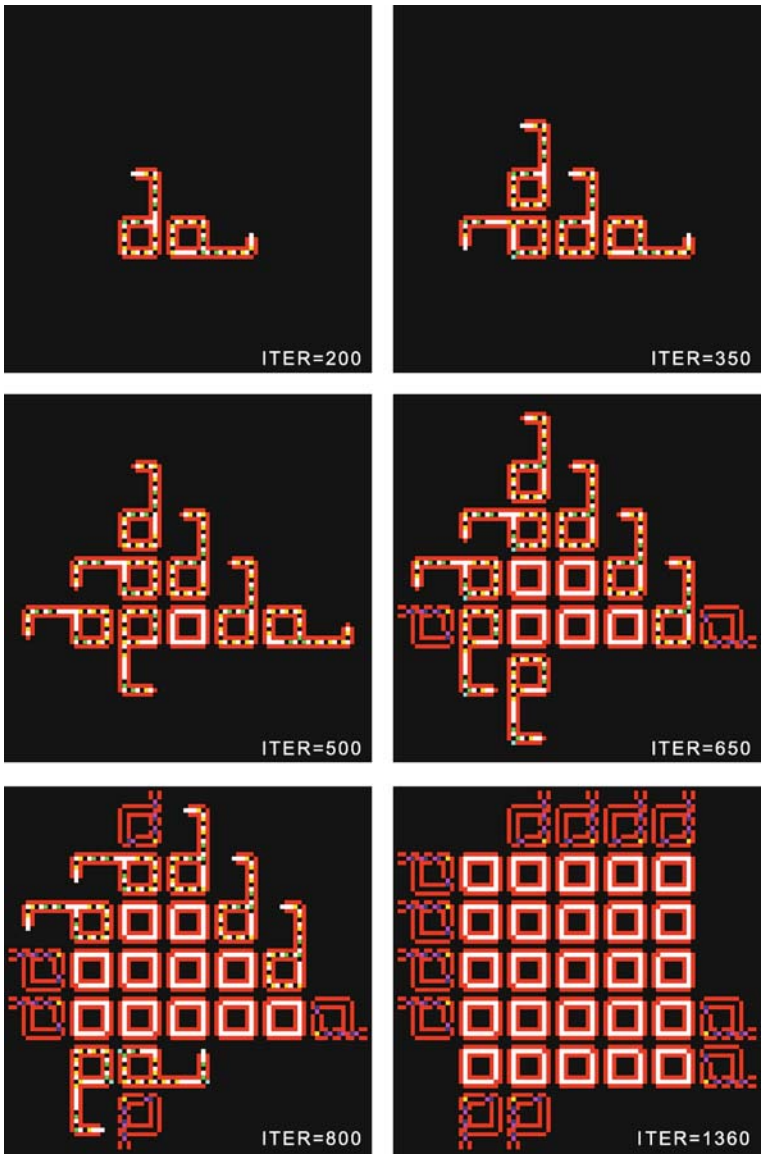

Self-Replication and Cellular Automata, Figure 8
Propagation pattern of Langton's loop

We use a 9-element neighborhood (the element itself plus its 8 neighbors) and the basic configuration of the loop (Fig. 11) requires five states plus at least one data state. State 0 (black) is the *quiescent* state: it represents the inactive background. State 1 (white) is the *sheath* state, that is the state of the elements making up the sheath and the four gates. State 2 (red) is the *activation* state or *control* state. The four gate elements are in state 2, as are the messengers which will be used to command the constructing arm and the tip of the constructing arm itself for the first phase of construction, after which the tip of the arm will switch to state 3 (light blue), the *construction* state. State 3 will construct the sheath that will host the offspring, signal the parent loop that the sheath is ready, and lead the duplicated data to the new loop. State 4 (green), the destruction state, will destroy the constructing arm once the copy is completed. In addition to these states, two additional *data* states (light and dark grey) represent the information stored in the loop. In this example, they are inactive, while the next section describes a loop where they are used to store an executable program.

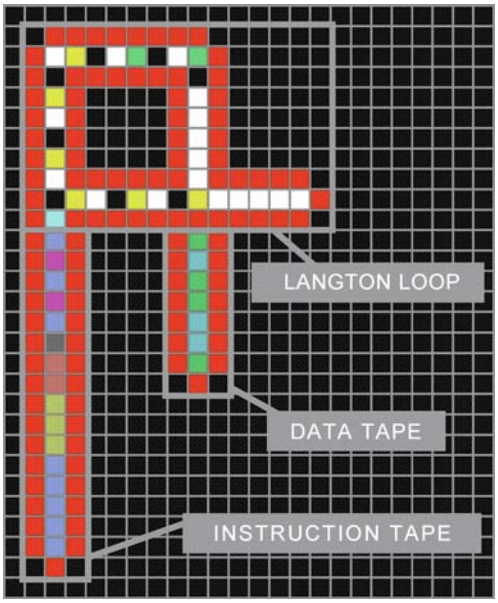

Self-Replication and Cellular Automata, Figure 9
A two-tape Turing machine appended to Langton's loop (iteration 0)

The initial configuration is in the form of a square loop wrapped around a sheath. The size of the loop is variable, and for our example is set to 8×8. Once the iterations begin, the data starts turning counterclockwise around the loop. Nothing happens until the first control element (red) reaches a corner of the loop, where it checks the status of the gate. Since the gate is open, the control element splits into two identical elements: the first continues turning around the loop, while the second starts extending the arm (Fig. 12).

The arm advances automatically by one position every two iterations. Once the arm has started extending, each control element that arrives to a corner will again split and one of the copies will start running along the arm, advancing twice as fast. When the first of these messengers reaches the tip of the arm, the tip, which was until then in state 2, passes to state 3 and continues to advance at the same speed. This transformation tells the arm that it has reached the location of the offspring loop and to start constructing the new sheath. The next three messengers will force the tip of the arm to turn left, while the fourth causes the sheath to close upon itself (Fig. 13). It then runs back along the arm to signal to the original loop that the new sheath is ready. Once the return signal arrives at the corner of the original loop, it causes a copy of the data in the loop to run along the arm and wrap itself around the new sheath. Once the second copy has completed the loop (Fig. 14), it sends a destruction signal (green) back along the arm. The signal will destroy the arm until it reaches the corner of the original loop, where it closes the gate to avoid further copies.

After 121 time periods the gates of the original automaton will be closed and it will enter an inactive state, with the understanding that it will be ready to reproduce itself again should the gates be opened. The main advantage of the new mechanism is that it becomes relatively simple to retract the arm if an obstacle (either the boundary of the array or another loop) is encountered, and therefore our loop is perfectly capable of operating in a finite space.

In Fig. 15, we illustrate an example of how the data states can be used to carry out operations alongside self-reproduction. The operation in question is the construction of three letters, LSL (the acronym of Logic Systems Laboratory, where the research was made), in the empty space inside the loop. Obviously this is not a very useful operation from a computational point of view, but it is a far from trivial construction task which should suffice to demonstrate the capabilities of the automaton.

As should be obvious, while our loop owes to von Neumann the concept of constructing arm and to Langton (and/or Codd) the basic loop structure, it is in fact a very different automaton, endowed with some of the properties of both. We have seen that von Neumann's automaton is extremely complex, while Langton's loop is very simple. The complexity of our automaton is more difficult to estimate, as it depends on the data circulating in the loop. The number of non-quiescent elements making up the initial configuration depends directly on the size of the circulating program. The more complex (i. e. the longer) the program, the larger the automaton (it should be noted, however, that the complexity of the self-reproduction process does not depend on the size of the loop). The number of transition rules is obviously a function of the number of data states: in the basic configuration (i. e., one data state), the automaton needs 692 rules (173 rules rotated in the four directions), assuming that, by default, all elements remain in the same state. The complexity of the basic configuration is therefore in the same order as that of Langton's and Byl's loops, with the proviso that it is likely to increase drastically if the data in the loop is used to implement a complex function.

Other Approaches to Self-Replication

Von Neumann's and Langton's structures represent the main landmarks in the study of self-replication in com-

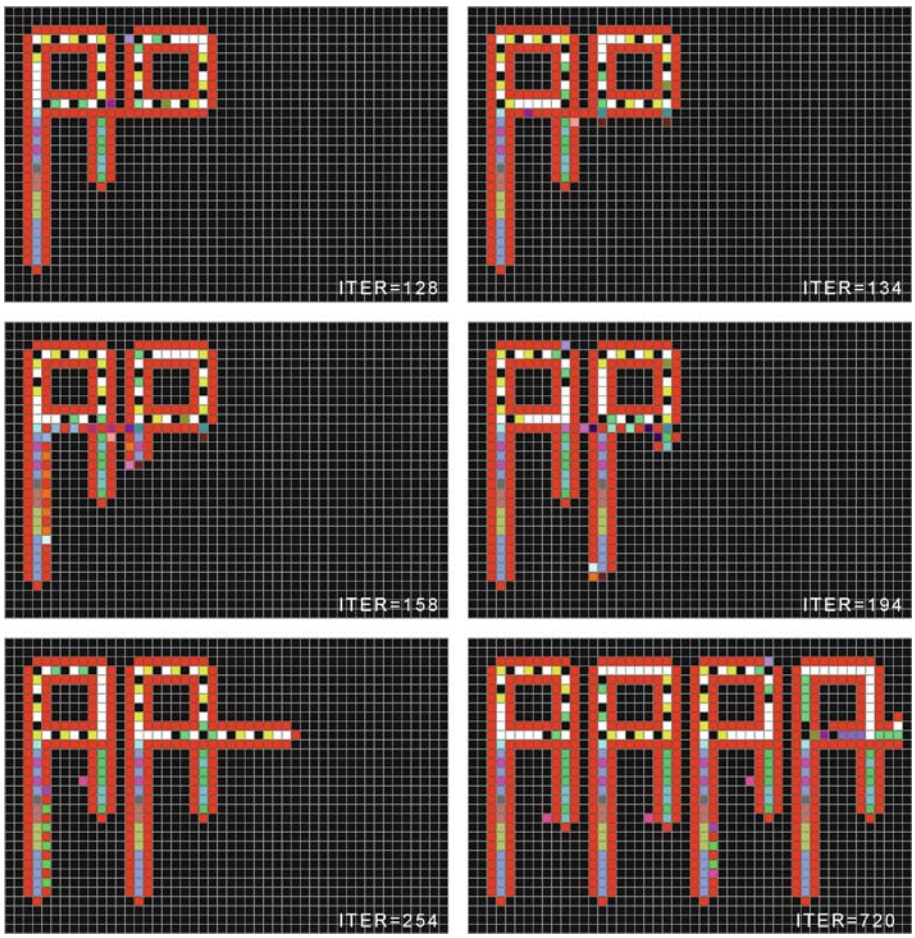

Self-Replication and Cellular Automata, Figure 10
Self-replication of the Turing machine

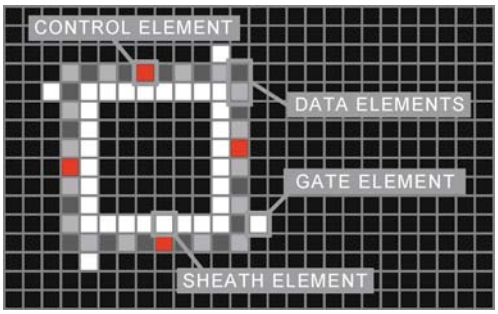

Self-Replication and Cellular Automata, Figure 11
The initial configuration of our loop (iteration 0)

puting machines. It can safely be said that all other approaches refer, directly or indirectly, to these two systems. However, there exist some approaches to self-replication that cannot be easily reduced to simple variations on one of these two themes, either because they specifically take into consideration some issues that are not addressed by Langton and von Neumann, or because they occur in environments that are considerably different from the two original approaches.

In this section, we will deal in depth with one example, the *Tom Thumb algorithm* that, while referring back to von Neumann insofar as its goal is the implementation of self-replicating logic circuits, is specifically designed to operate efficiently in the kind of digital devices that are available today. The algorithm approaches cellular automata from a slightly unconventional angle [37], with the objective of a hardware realization of self-replication within a programmable logic device, or FPGA [41].

In the second part of the section, we will look at a set of approaches to self-replication that represent notable extensions to the approaches of von Neumann and Langton, because of different mechanisms (self-inspection), operat-

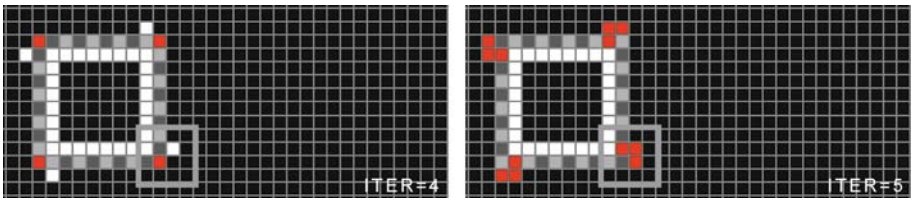

Self-Replication and Cellular Automata, Figure 12
The constructing arm begins to extend

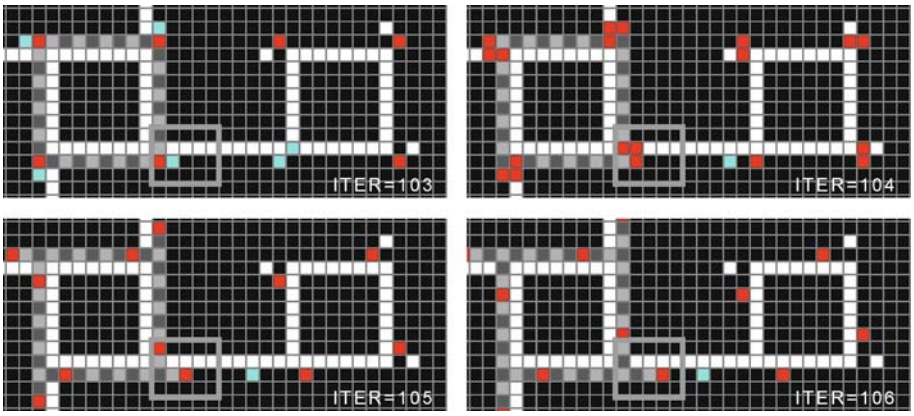

Self-Replication and Cellular Automata, Figure 13
The new sheath has been fully constructed and a copy of the data is sent from the parent to the offspring

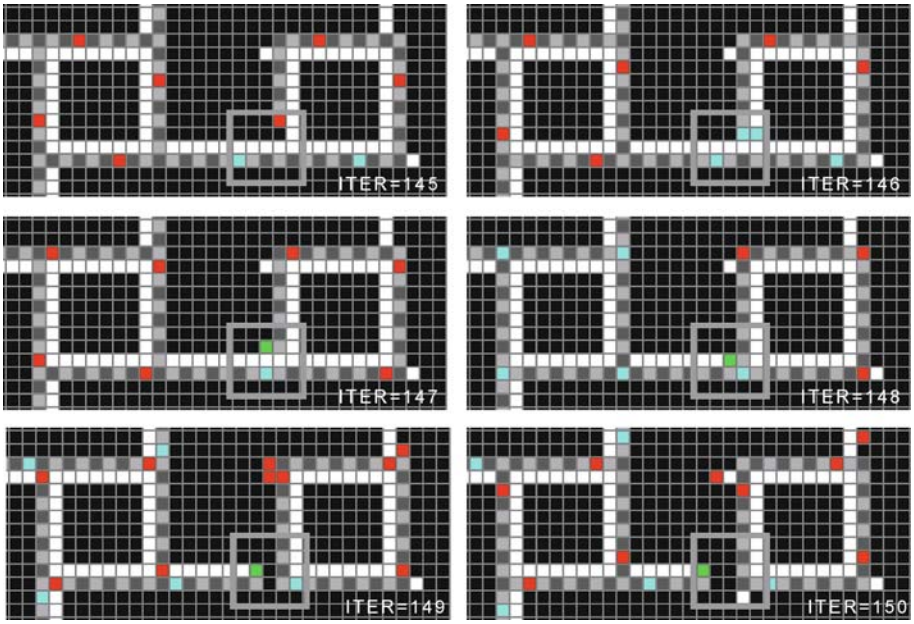

Self-Replication and Cellular Automata, Figure 14
The copy is complete and the constructing arm retracts

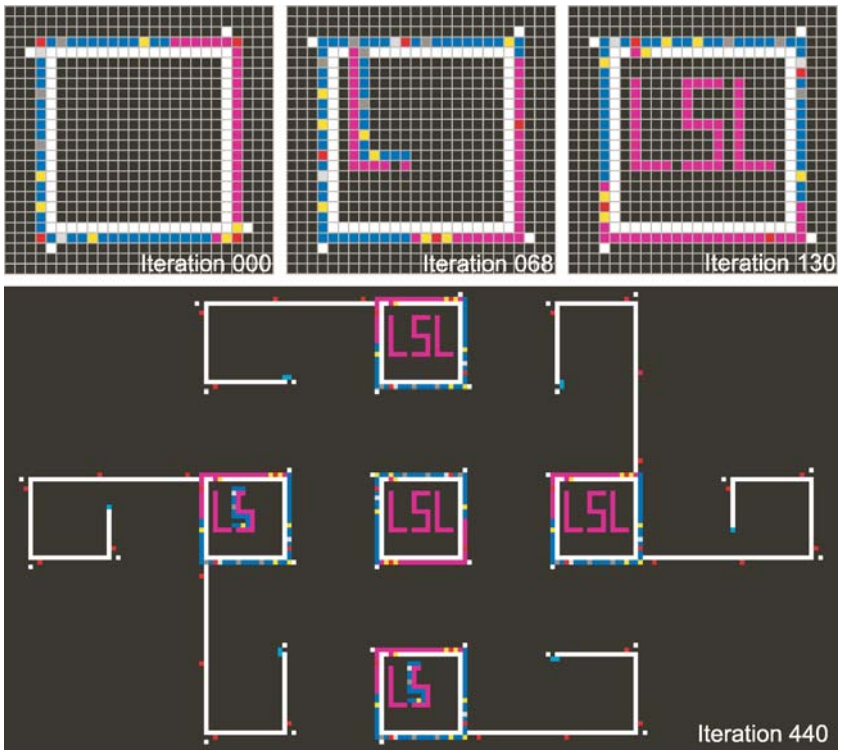

Self-Replication and Cellular Automata, Figure 15
The LSL automaton at different iterations

ing milieus (three-dimensional or self-timed CA), or design rules (evolutionary approaches).

Self-Replication in Hardware: The Tom Thumb Algorithm

In past years, we have devoted considerable effort to research on self-replication, studying this process from the point of view of the design of high-complexity multi-processor systems (with particular attention to next-generation technologies such as nanoelectronics). When considering self-replication in this context, Langton's loop and its successors share several weaknesses. Notably, besides the lack of functionality of Langton's loop (remedied only partially by its successors), which severely limits its usefulness for circuit design, each of these automata is characterized by a very loose utilization of the resources at its disposal: the majority of the elements in the cellular array remain in the quiescent state throughout the entire self-replication process.

A new loop was then developed specifically to address these very practical issues. In fact, the system is targeted to the implementation of self-replication within the context of digital circuits realized with programmable logic devices

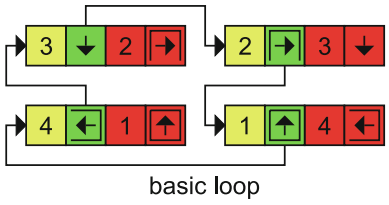

Self-Replication and Cellular Automata, Figure 16
Tom Thumb algorithm: basic loop

(the states of the cellular automaton can then be seen as the configuration bits of the elements of the device). The new loop is based on an original algorithm, the so-called *Tom Thumb algorithm* [18,19].

The minimal loop compatible with this algorithm is made up of four cells, organized as a square of two rows by two columns (Fig. 16). Each cell is able to store in its four memory positions four hexadecimal characters of an artificial *genome* (defined as the information required for the construction of the loop). The whole loop thus embeds 16 such characters.

The original genome for the minimal loop is organized as another loop, the *basic loop*, of eight hexadecimal characters, i. e. half the number of characters in the minimal

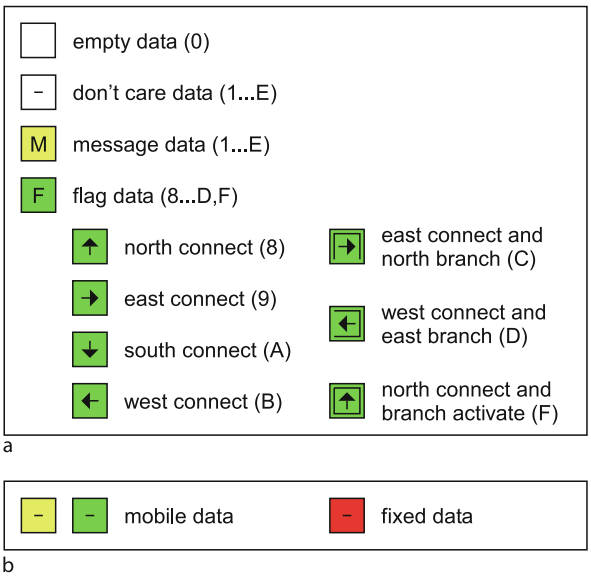

Self-Replication and Cellular Automata, Figure 17
a Graphical and hexadecimal representations of the 15 characters forming the alphabet of the artificial genome. b Graphical representation of the status of each character

loop, moving counterclockwise by one character at each time step.

The 15 hexadecimal characters that compose the alphabet of the artificial genome are detailed in Fig. 17a. They are either *empty data* (0), *message data* ($M = 1 \ldots E$), or *flag data* ($F = 8 \ldots D, F$). Message data will be used to configure our final artificial organism, while flag data are indispensable for constructing the skeleton of the loop. Furthermore, each character is given a status and will eventually be *mobile data*, moving indefinitely around the loop, or *fixed data*, definitely trapped in a memory position of a cell (Fig. 17b). It is important to note that, while in this simple example the message data can take value from 1 to E, the Tom Thumb algorithm is perfectly scalable in this respect, that is, the size of the message data can be increased at will, while the flag data remain constant. This is a crucial observation in view of the exploitation of this algorithm in a programmable logic device, where the message data (the configuration data for the programmable elements of the circuit) are usually much more complex.

At each time step ($t = 1, 2, \ldots$), a character of the *original loop* is sent to the lower leftmost cell (Fig. 18). The construction of the loop, i. e., storing the fixed data and defining the paths for mobile data, depends on two rules:

- If the four, three, or two rightmost memory positions of the cell are empty (blank squares), the characters are shifted by one position to the right (rule #1: shift data).
- If the rightmost memory position is empty, the characters are shifted by one position to the right (rule #2: load data). In this situation, the two rightmost characters are trapped in the cell (fixed data), and a new connection is established from the second leftmost position toward the northern, eastern, southern or western cell, depending on the fixed flag information (in Fig. 18, at

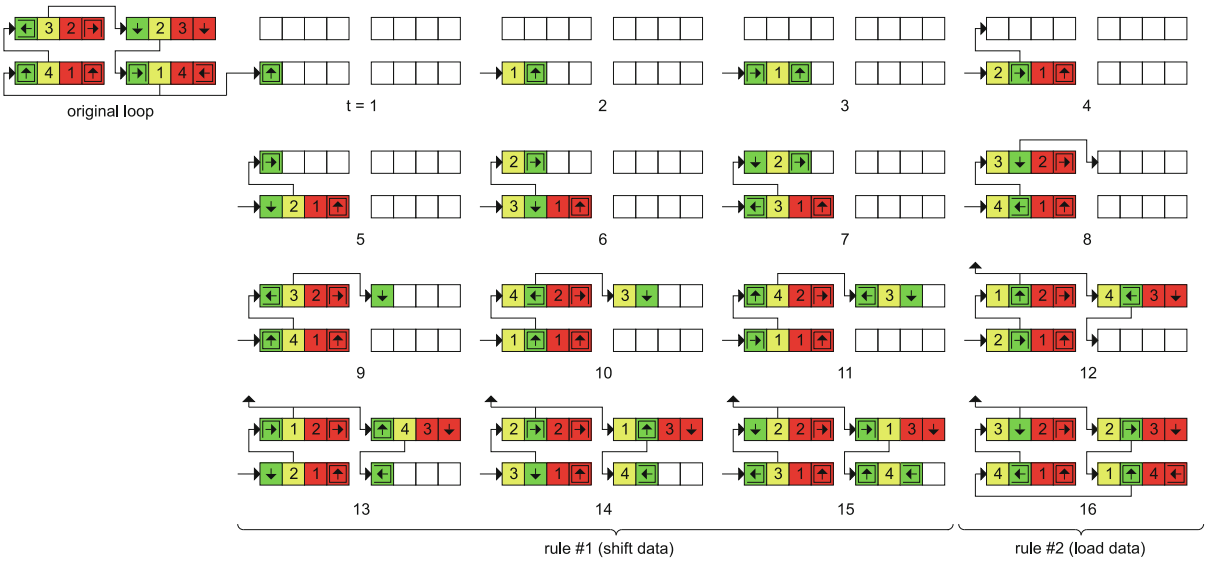

Self-Replication and Cellular Automata, Figure 18
Construction of a first specimen of the loop

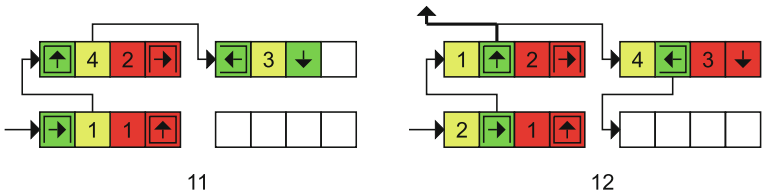

Self-Replication and Cellular Automata, Figure 19
Creation of a new daughter loop to the north (rule #3)

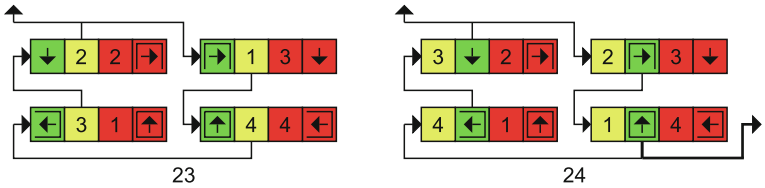

Self-Replication and Cellular Automata, Figure 20
Creation of a new daughter loop to the east (rule #4)

time $t = 4$, the fixed flag F = F determines a northern connection).

At time $t = 16$, 16 characters, i. e., twice the contents of the basic loop, have been stored in the 16 memory positions of the loop (Fig. 18). Eight characters are fixed data, forming the *phenotype* of the final loop, and the eight remaining ones are mobile data, composing a copy of the original genome, i. e., the *genotype*. Both *interpretation* (the construction of the cell) and *copying* (the duplication of the genetic information) have been therefore achieved.

The fixed data trapped in the rightmost memory positions of each cell remind us of the pebbles left by Tom Thumb for memorizing his way in the famous children's story, an analogy that gives our algorithm its name.

In order to grow loops in both horizontal and vertical directions, the mother loop should be able to trigger the construction of two daughter loops, northward and eastward. Two new rules are then necessary:

- At time $t = 11$ (Fig. 19), we observe a pattern of characters which is able to start the construction of the northward daughter loop; the upper leftmost cell is characterized by two specific flags, i. e., a fixed flag in the rightmost position, indicating a north branch (F = C) and the branch activation flag (F = F), in the leftmost position (rule #3: daughter loop to the north). The new path to the northward daughter loop will start from the second leftmost memory position ($t = 12$).
- At time $t = 23$ (Fig. 20), another particular pattern of characters starts the construction of the eastward daughter loop; the lower rightmost cell is characterized by two specific flags, i. e., a fixed flag indicating an east branch (F = D), in the rightmost position, and the branch activation flag (F = F), in the leftmost position (rule #4: daughter loop to the east). The new path to the eastward daughter loop starts from the second leftmost memory position ($t = 24$).

When two or more paths are activated simultaneously, a clear priority should be established between the different paths. Three growth patterns were chosen (Fig. 21), leading to four more rules:

- For loops in the lower row a collision occurs between the closing path, inside the loop, and the path entering the lower leftmost cell. The westward path has priority over the eastward path (rule #5).
- With the exception of the bottom loop, the inner path (i. e. the westward path) has priority over the northward path (rule #6) for the loops in the leftmost column.
- For all other loops, two types of collisions may occur: between the northward and eastward paths (2-signal collision) or between these two paths and a third one, the closing path (3-signal collision). In this case, the northward path will have priority over the eastward path (2-signal collision), and the westward path will have priority over the two other ones (3-signal collision)(rules #7 and #8).

We finally opted the following hierarchy: an east to west path has priority over a south to north path, which has priority over a west to east path.

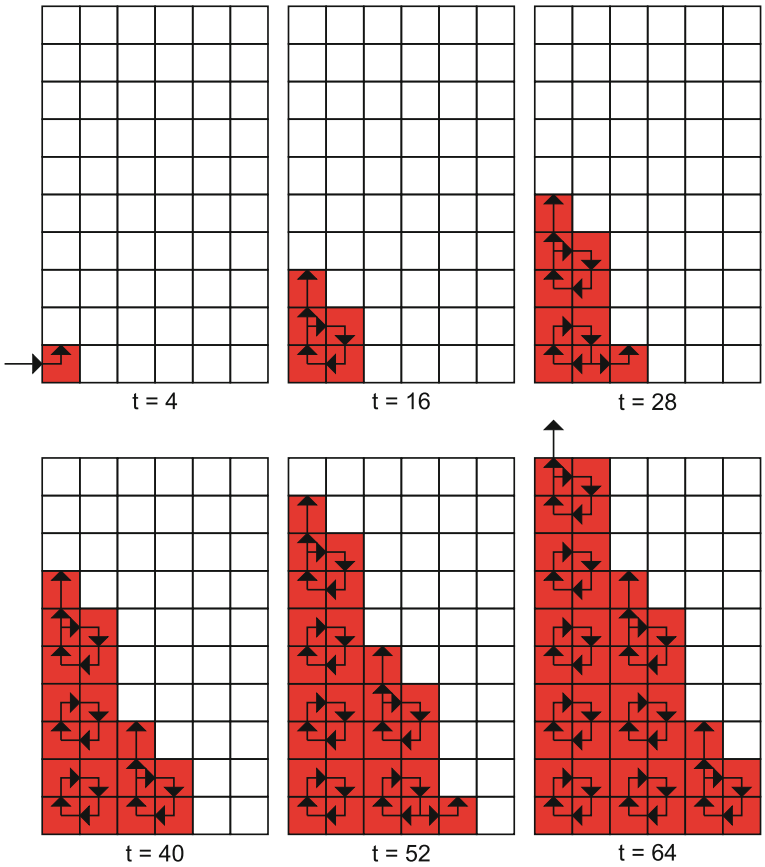

Self-Replication and Cellular Automata, Figure 21
Growth of a colony of minimal loops represented at different time steps

The results of such a choice are as follows (Fig. 21): a closing loop has priority over all other outer paths, which makes the completed loop entirely independent of its neighbors, and the loops will grow bottom-up vertically. This choice is quite arbitrary and may be changed according to other specifications.

Unlike its predecessors, the Tom Thumb loop has been developed with a specific purpose beyond the theoretical study of self-replication. We believe that, in the not-so-distant future, circuits will reach densities such that conventional design techniques will become unwieldy. Should such an hypothesis be confirmed, self-replication could become an invaluable tool, allowing the engineer to design a single processing element, part of an extremely large array that would build itself through the duplication of the original element.

Current technology does not, of course, provide a level of complexity that would render this kind of process necessary. However, programmable logic devices (already among the densest circuits on the market) can be used as a first approximation of the kind of circuits that will become available in the future. Our loop is then targeted to the implementation of self-replication on this kind of device.

To this end, our loop introduces a number of features that are not present in any of the historical self-replicating loops we presented. Most notably, the structure of the loop (that is, the path used by the configuration data) is determined by the sequence of flags in the genome, implying that structures of almost any shape and size can be constructed and replicated using this algorithm, as long as the loop can be closed and that there is space for the daughter organisms. In practice, this implies that, if the Tom Thumb algorithm is used for the configuration logic of a programmable device, any of its configurations, and hence any digital circuit, can be made capable of self-replication.

In addition, particular care was given to develop a self-replication algorithm that is *efficient* (in the sense that it fully exploits the underlying medium, rather than leaving the vast majority of elements inert as past algorithms did), *scalable* (all the interactions between the elements are

purely local, implying that organisms of any size can be implemented), and amenable to a *systematic design process*. These features are important requirements for the design of highly-complex systems based on either silicon or molecular-scale components.

Different Techniques and Environments

Von Neumann's and Langton's automata share a common basic technique to obtain self-replication: the construction of the new machine is directed through the *interpretation* of a description, coded as a sequence of states. In the case of von Neumann, this description (which, in biological terms, is usually identified as the genome of the artificial organism) is stored within the memory tape, which is read and interpreted by the universal constructor to build a copy of the machine. In Langton's case, the description is stored in the mobile data that runs within the sheath of the loop.

This mechanism of interpretation, while standard in many approaches, is not however unique: some examples of self-replicating CA exploit a different mechanism, that of *self-inspection*. In these approaches, instead of reading and interpreting a description, the self-replicating automaton inspects itself and produces a copy of what it finds. While less general than the universal constructor (obviously, the machine can only build an exact copy of itself), the functionality of this approach is similar to that of Langton's loop. Indeed, the most representative example of self-inspection is that of a self-replicating loop [12]. A more recent example is a variation of the Tom Thumb algorithm, where self-inspection was used to self-replicate a small processor within a field-programmable gate array [30].

And while the Tom Thumb algorithm targets in priority silicon-based circuits, other approaches have tried to explore alternative environments that, in some way, might more closely resemble the kind of technologies that will be available in the future. An example is Morita and Imai's study of self-replication in the context of reversible cellular automata [20] (in a reversible CA, every configuration has at most one predecessor), inspired by reversible logic in digital circuits.

Similarly, Peper et al. [24,38] have developed self-replicating structures in Self-Timed Cellular Automata (STCA). This kind of automata do not rely on a global synchronization mechanism to update the states of the cells, but rather the state transitions only occur when triggered by transitions in neighboring cells. The basic assumption in this work is that STCA is a model that might more closely resemble molecular-scale nanoelectronic devices.

A final example in this context is the three-dimensional extension of self-replication, usually based on the assumption that silicon, with its rigidly two-dimensional structure, will one day be superseded by a technology that can exploit all three dimensions. In this context, Imai et al. [13] have extended their reversible approach, with the assumption that reversible logic is more amenable to an extension to three dimensions than conventional logic because of the greatly reduced power dissipation. Stauffer at al. [36] have also shown that the Data and Signal Cellular Automaton (DSCA) approach, designed to simplify the implementation of CA in digital hardware, can be extended to realize self-replication in three dimensions.

The study of self-replicating CA in the context of new technologies holds the promise of one day bringing a major contribution to computation. To determine how self-replication might be useful in this context, some attempts have been made at using self-replicating structures for computation. An example of this approach is the work of Chou and Reggia [8], who use self-replication as a mechanism to obtain massively parallel machines which can potentially be used to solve hard problems (the example used in the paper is the NP-complete problem of satisfiability).

An attempt was also made to perform computation using Tempesti's loops. In alternative to embedding a complex program, this kind of loops are used to perform computation by inter-loop communication. Using the *collision-based computing* paradigm, Petraglio et al. [28] have shown that it is possible to implement arithmetic operations by passing messages from one loop to another after building a network structure through self-replication. This approach, while valuable from a theoretical standpoint, shares however the same weakness of other loop-based computing approaches in that the poor utilization of resources makes a physical realization of such a system highly impractical.

Another problem to be solved for a practical implementation of self-replicating structures is their design: few approaches have attempted to define a precise methodology to define and create self-replicating structures. In this context, several researchers have attempted to use evolutionary techniques to find automatically self-replicating machines. In this context, the work of Sayama et al. has gone through several iterations [31,32,33] in an attempt to define loops that evolve through the self-replication process towards "fitter" individuals. Chou and Reggia [7], on the other hand, use evolution to find novel self-replicating structures within a CA, whereas Pan and Reggia [23] studied the conditions in which self-replicating structures might spontaneously emerge in a cellular space.

Future Directions

Historically, self-replication in cellular automata began as a paradigm to achieve fault tolerance in computing devices. In the following decades, much of the emphasis shifted towards a more "theoretical" approach where self-replication was considered as part of a more general investigation into the application of biologically-inspired mechanisms to computing. And while the latter approach remains an active research topic, the original paradigm was somewhat set aside because technology would not allow a practical implementation of self-replication in digital hardware.

More recently, however, advances in electronic devices (notably with the introduction of programmable devices, or FPGAs), together with emerging technological issues have rekindled interest in self-replication in a context similar to von Neumann's original work. In particular, the drastic device shrinking, low power supply levels, and increasing operating speeds, which accompany the technological evolution of silicon to deeper submicron levels, significantly reduce the noise margins and increase the soft-error rates [42]. In addition, the nascent field of nanoelectronics holds great promise for the future of computing devices, but introduces extremely high fault rates (e. g., [25]) and complex layout issues.

Thus, self-replication is currently attracting a considerable amount of attention for the same reasons that initially pushed von Neumann to investigate it as a possible solution to reliability and layout issues. Fault tolerance and self-organization are thus becoming the focal point of research in the field and the features of molecular-scale electronics seem to imply that self-replication at the device level will be an extremely useful paradigm in next-generation devices.

Bibliography

1. Asprey W (1992) John von Neumann and the Origins of Modern Computing. The MIT Press, Cambridge
2. Banks ER (1970) Universality in Cellular Automata. In: Proc. IEEE 11th Annual Symposium on Switching and Automata Theory, Santa Monica, CA, pp 194-215
3. Beuchat J-L, Haenni J-O (2000) Von Neumann's 29-State Cellular Automaton: A Hardware Implementation. IEEE Trans Education 43(3):300–308
4. Buckley WR, Mukherjee A (2005) Constructibility of Signal-Crossing Solutions in von Neumann 29-State Cellular Automata. Proc. 2005 Int. Conf. on Computational Science. LNCS, vol 3515. Springer, Berlin, pp 395–403
5. Burks A (ed) (1970) Essays on Cellular Automata. University of Illinois Press, Urbana
6. Byl J (1989) Self-Reproduction in Small Cellular Automata. Physica D 34:295–299
7. Chou H-H, Reggia JA (1997) Emergence of self-replicating structures in a cellular automata space. Physica D 110(3-4):252–276
8. Chou H-H, Reggia JA (1998) Problem solving during artificial selection of self-replicating loops. Physica D 115(3-4):293–312
9. Codd EF (1968) Cellular Automata. Academic Press, New York
10. Freitas RA Jr, Gilbreath WP (eds) (1980) Advanced Automation for Space Missions. In: Proc. 1980 NASA/ASEE Summer Study, Scientific and Technical Information Branch (available from U.S. G.P.O.), Washington, DC
11. Freitas RA Jr, Merkle RC (2004) Kinematic Self-Replicating Machines. Landes Bioscience, Georgetown
12. Ibanez J, Anabitarte D, Azpeitia I, Barrera O, Barrutieta A, Blanco H, Echarte F (1995) Self-inspection based reproduction in cellular automata. In: Proc. 3rd European Conf. on Artificial Life (ECAL95). LNCS, vol 929. Springer, Berlin, pp 564–576
13. Imai K, Hori T, Morita K (2002) Self-reproduction in three-dimensional reversible cellular space. Artif Life 8(2):155–174
14. Langton CG (1984) Self-Reproduction in Cellular Automata. Physica D 10:135–144
15. Lee C (1968) Synthesis of a Cellular Computer. In: Applied Automata Theory. Academic Press, London, pp 217–234
16. Lohn JD, Reggia JA (1997) Automatic Discovery of Self-Replicating Structures in Cellular Automata. IEEE Trans Evol Comput 1(3):165–178
17. Mange D, Sipper M, Stauffer A, Tempesti G (2000) Towards Robust Integrated Circuits: The Embryonics Approach. Proc IEEE 88(4):516–541
18. Mange D, Stauffer A, Petraglio E, Tempesti G (2004) Self-replicating loop with universal construction. Physica D 191:178–192
19. Mange D, Stauffer A, Peparolo L, Tempesti G (2004) A Macroscopic View of Self-Replication. Proc IEEE 92(12):1929–1945
20. Morita K, Imai K (1996) Self-reproduction in a reversible cellular space. Theor Comput Sci 168:337–366
21. Nehaniv CL (2002) Self-Reproduction in Asynchronous Cellular Automata. In: Proc. 2002 NASA/DoD Conf. on Evolvable Hardware (EH02). IEEE Computer Society, Washington, DC, pp 201–209
22. Nourai F, Kashef RS (1975) A Universal Four-State Cellular Computer. IEEE Trans Comput 24(8):766–776
23. Pan Z, Reggia J (2005) Evolutionary Discovery of Arbitrary Self-replicating Structures. In: Proc. 5th Int. Conf. on Computational Science (ICCS 2005) - Part II. LNCS, vol 3515. Springer, Berlin, pp 404–411
24. Peper F, Isokawa T, Kouda N, Matsui N (2002) Self-Timed Cellular Automata and their computational ability. Future Gener Comput Syst 18(7):893–904
25. Peper F, Lee J, Abo F, Isokawa T, Adaki S, Matsui N, Mashiko S (2004) Fault-Tolerance in Nanocomputers: a Cellular Array Approach. IEEE Trans Nanotechnol 3(1):187–201
26. Perrier J-Y, Sipper M, Zahnd J (1996) Toward a Viable, Self-Reproducing Universal Computer. Physica D 97:335–352
27. Pesavento U (1995) An Implementation of von Neumann's Self-Reproducing Machine. Artif Life 2(4):337–354
28. Petraglio E, Tempesti G, Henry J-M (2002) Arithmetic Operations with Self-Replicating Loops. In: Adamatsky A (ed) Collision-Based Computing. Springer-Verlag, London, pp 469-490
29. Reggia JA, Armentrout SA, Chou H-H, Peng Y (1993) Simple Systems That Exhibit Self-Directed Replication. Science 259:1282–1287

30. Rossier J, Thoma Y, Mudry PA, Tempesti G (2006) MOVE Processors that Self-Replicate and Differentiate. In: Proc. 2nd Int. Workshop on Biologically-Inspired Approaches to Advanced Information Technology (Bio-ADIT06). LNCS, vol 3853. Springer, Berlin, pp 328–343
31. Salzberg C, Antony A, Sayama H (2004) Evolutionary dynamics of cellular automata-based self-replicators in hostile environments. BioSystems 78:119–134
32. Sayama H (1998) Introduction of Structural Dissolution into Langton's Self-Reproducing Loop. In: Artificial Life VI, Proc. 6th Int. Conf. on Artificial Life, MIT Press, Cambridge, pp 114–122
33. Sayama H (2000) Self-replicating worms that increase structural complexity through gene transmission. In: Artificial Life VII: Proc. 7th Int. Conf. on Artificial Life, MIT Press, Cambridge, pp 21–30
34. Sipper M (1995) Studying artificial life using a simple, general cellular model. Artif Life 2(1):1–35
35. Sipper M, Mange D, Stauffer A (1997) Ontogenetic Hardware. BioSystems 44:193–207
36. Stauffer A, Mange D, Petraglio E, Vannel F (2004) DSCA Implementation of 3D Self-Replicating Structures. In: Proc. 6th Int. Conf. on Cellular Automata for Research and Industry (ACRI2004). LNCS, vol 3305. Springer, Berlin, pp 698–708
37. Stauffer A, Sipper M (2004) The Data-and-Signals Cellular Automaton and Its Application to Growing Structures. Artif Life 10(4):463–477
38. Takada Y, Isokawa T, Peper F, Matsui N (2006) Universal Construction and Self-Reproduction on Self-Timed Cellular Automata. Int J Mod Phys C 17(7):985–1007
39. Tempesti G (1995) A New Self-Reproducing Cellular Automaton Capable of Construction and Computation. In: Proc. 3rd European Conf. on Artificial Life. LNAI, vol 929. Springer, Berlin, pp 555–563
40. Tempesti G (1998) A Self-Repairing Multiplexer-Based FPGA Inspired by Biological Processes. Ph.D. Thesis, Ecole Polytechnique Fédérale de Lausanne (EPFL)
41. Trimberger S (ed) (1994) Field-Programmable Gate Array Technology. Kluwer, Boston
42. Various (1999) A D&T Roundtable: Online Test. IEEE Design Test Comput 16(1):80–86
43. Vitanyi PMB (1973) Sexually reproducing cellular automata. Math Biosci 18:23–54
44. Burks AW (ed) von Neumann J (1966) The Theory of Self-Reproducing Automata. University of Illinois Press, Urbana

Semantic Web

Wendy Hall, Kieron O'Hara
Intelligence, Agents, Multimedia Group, School of Electronics and Computer Science, University of Southampton, Southampton, United Kingdom

Article Outline

Glossary
Definition of the Subject
Introduction
Linking Data
The Layered Model of the Semantic Web
Applications
Controversies
Future Directions
Bibliography

Glossary

Dereferencing Dereferencing a URI is using the URI to identify the object or resource it refers to.

Folksonomy 'Folksonomy' is a neologism applied to structures that emerge from the practice of 'tagging' Web content. In some Web 2.0 applications, users can apply a tag (a descriptive term) to content such as a photograph or video clip. The tags need only be meaningful to the individual tagger, but if a large enough number of users tag content, descriptive structures analogous to more formal ontologies can emerge that are meaningful to wide communities.

GRDDL Gleaning Resource Descriptions from Dialects of Languages (GRDDL, pronounced 'griddle') is a mechanism for helping bootstrap the Semantic Web, by extracting RDF from XML documents, using transformations expressed in XSLT. GRDDL became a W3C recommendation in 2007.

Metadata Metadata is data about data. In the context of the Semantic Web, metadata are also called 'markup' or 'annotations'. Because one aim of the Semantic Web is to support machine processing of information, metadata are helpful in describing the content of data. For example, metadata attached to a series of numerals could explain that it represents a zip code or a height or a population figure.

Ontology An ontology defines, describes and constrains the concepts and relationships that are used in some particular domain of knowledge. Ontologies may therefore have an important role in data sharing, for example by providing a means of expressing which concepts (in the ontology) are referred to by particular terms (in a set of databases, which may use widely differing vocabularies).

OWL The Web Ontology Language (OWL) is a language for describing and sharing ontologies on the Web. It is an extension of RDF, and has three variants. OWL Full is maximally expressive and compatible with RDF (any legal RDF document is a legal OWL Full document), but is undecidable. OWL DL is based on a series of restrictions of OWL Full which support efficient reasoning, at the cost of losing the strong connection with RDF (not all RDF documents are legal OWL DL

documents). OWL Lite is even more restricted in expressivity, but is easier to grasp. OWL became a W3C recommendation in 2004.

RDF The Resource Description Framework (RDF) is a standard framework for representing data on the Web, representing it in a three-place relation, of subject-relation-object form, called a *triple*. It uses URIs to refer not only to the two items related, but also to the relationship asserted between them. In this way it extends the linking structure of the Web by allowing the nature of the link asserted to be described. Its syntax is XML-based, to support syntactic interoperability between the two. RDF became a W3C recommendation in 1999.

RDF(S) RDF Schema (RDF(S) or RDFS) is a language for representing information on the Web. It is an extension of RDF to allow the description of the relationships which can be asserted between resources using RDF. So, for instance, RDF allows the assertion of the relationship 'author' between, say, 'Herman_Melville' and 'Moby_Dick', but RDF(S) is required to assert the properties of the 'author' relationship (e. g. that every document has at least one author, or that the 'author' relationship is the inverse of the 'written_by' relationship). RDF(S) became a W3C recommendation in 2004.

Rules and RIF Rules govern the transformation of, and inference from, data. In particular, when data are being shared, it may be useful to make basic inferences over the data (for example, to determine whether two names refer to the same object or different objects). Rule-based knowledge such as this cannot be expressed in individual data stores, and may be hard to express in an ontology. The Rule Interchange Format (RIF) is a language, not complete at the time of writing, to express the most common or basic types of rule.

SPARQL SPARQL is a special query language designed specifically for querying data stored in RDF ('SPARQL' is a recursive acronym standing for 'SPARQL Protocol And RDF Query Language', and is pronounced 'sparkle'). SPARQL became a W3C recommendation in January 2008.

Triples and triplestores A triple is a statement in RDF, consisting of a subject, an object, and a binary predicate that relates them. Each item of the triple is identified by a URI reference (or can be a string literal, such as a date or a number or name). A knowledge repository which contains RDF triples is called a triplestore. Triplestores may need to contain several millions of triples, and so need to be able to support fast querying at these potentially very large scales.

URI A Uniform Resource Identifier (URI) is a string of characters for identifying an abstract or physical object or resource. There are different URI schemes (i. e. different ways to restrict the syntax of a URI to make it meaningful). A Uniform Resource Locator (URL) is a particular type of URI that identifies its object by means of its access mechanism or 'location' in the network. URIs are important in that they act as a standard way of referring to objects on the Web.

W3C The World Wide Web Consortium (W3C) is an international non-profit consortium which coordinates the development of Web standards, founded in 1994 under the directorship of Sir Tim Berners-Lee. The W3C groups together all the bodies involved in specifying Semantic Web standards into the Semantic Web Activity. A W3C working group works on a standard for a particular formalism, and when the standard is judged by the working group to be in a final form, fit for purpose and properly interoperable with other W3C standards, it ratifies the formalism by recommending it. Hence a W3C recommendation is an important standard.

Web of data The Web of Data is another way of referring to or explaining the vision of the Semantic Web, which emphasises the idea of creating links between data rather than documents (as on the current World Wide Web). Linking data, as with linking documents, enables them to be reused in interesting or unexpected contexts (when the data are interpreted using explicit semantic theories of the sort provided by ontologies). RDF is the anticipated mechanism for creating the links between data; dereferencing one or more of the URIs in an RDF triple will lead to descriptions of the resources referred to, which in turn are likely to contain further triples, which can again lead to dereferencing and so on.

Web science Web Science is the multidisciplinary activity of trying to understand the two-way dynamic relationship between Web technology and wider society, in order to ensure that the technological changes made to the Web (including the Semantic Web) are generally beneficial rather than otherwise.

XML The eXtensible Markup Language (XML) is a language for allowing users to tag or mark up content using tags of their own devising (for instance based on a particular vocabulary used in a small community of practice). As such, XML can serve as a basic data exchange format between applications. It is an advance on other well-known markup languages (such as HTML, a standard language for marking up content for display on the Web) in that it separates instructions

to do with content and document structure from those to do with formatting. It is a basic language for representing and exchanging structured information.

Definition of the Subject

The Semantic Web is a proposed extension to the World Wide Web (WWW) that aims to provide a common framework for sharing and reusing data across applications. The most common interfaces to the World Wide Web present it as a Web of Documents, linked in various ways including hyperlinks. But from the data point of view, each document is a black box – the data are not given independently of their representation in the document. This reduces its power, and also (as most information needs to be extracted from documents by a human agent) inhibits the use of automatic information processing methods on the Web. The Semantic Web is an effort, steered by the World Wide Web Consortium, to develop a set of protocols, formalisms and standards to transform the Web into a *Web of Data*. Links would be between data, and data could be accessed independently of the applications that created them. This would allow both the sharing of data, and the amalgamation of data from different sources, using heterogeneous formats, in new contexts.

Introduction

The idea of the Semantic Web (SW), of exploiting the possibilities for serendipitous reuse of linked data, dates back at least to Sir Tim Berners-Lee's plenary talk at the first International World Wide Web Conference at CERN in Geneva in 1994 [33]. In that talk, Berners-Lee argued that there is too little machine-readable information on the WWW as was currently constituted. "The meaning of the documents is clear to those with a grasp of (normally) English, and the significance of the links is only evident from the context around the anchor. To a computer, then, the Web is a flat, boring world devoid of meaning. This is a pity, as in fact documents on the Web describe real objects and imaginary concepts, and give particular relationships between them. … Adding semantics to the Web involves two things: allowing documents which have information in machine-readable forms, and allowing links to be created with relationship values. Only when we have this extra level of semantics will we be able to use computer power to help us exploit the information to a greater extent than our own reading." Indeed, the original vision of the WWW was intended to support greater machine understanding of people's work and interactions; the 'flat' understanding of the world that the WWW produces in machines is a first step towards the richer vision.

The SW was from the beginning conceived as a set of layered standards and formalisms (see Sect. "The Layered Model of the Semantic Web" for details). The development of the Resource Description Framework (RDF) in the 1990s was a key technology [80]. RDF is an important formalism, as it allows expression not only *of* the link between two objects, but also about the *nature* of the link itself. Hence, one can follow a chain of links not only via the objects linked, but also the types of links involved. In the WWW of documents, links connect documents written in the Hypertext Markup Language HTML; in the SW, links in RDF connect not only documents but arbitrary things (objects and relationships) identified by the Uniform Resource Identifiers (URIs [38]) in the RDF triples representing the data.

The ability to move between data linked in such a way opens up the possibility of exposing data to the Web, and then being able to access such data from any application. As a simple example, consider personal information such as one's bank statements, information-based resources such as digital photographs, and an application such as a calendar or diary. Each of these depends on data which is controlled by the applications that use them. But in a genuine Web of Data, we could link these data in a productive way – something as simple as being able to present one's financial information in one's calendar. The metadata in one's photographs often includes information about the time of their creation; an application able to get at the photograph metadata and the data on one's calendar might be able to suggest where the photo was taken, and its possible location. The ability to use all these data in a constructive way is impossible without a Web of linked data to enable applications to move between the data sources.

In this way, the SW changes our model of the value of information. Currently, it is generally presumed that the value of information stems from its *scarcity* – people and organizations gain value from information they have gathered, and are given monopoly rights to exploit that information via such legal contrivances as copyright, intellectual property rights, licensing, and so on. Even when organizations do not resort to the law, they will make great investments in protecting trade secrets. However, this scarcity-based model seems inadequate for the digital age.

In the first place, as economist William Baumol has argued, the social benefits from unlicensed use of 'protected' knowledge and innovation, were already large in the pre-digital economy, and indeed account for much of our wealth today: "some 80 percent of the benefits [innovation] may – i. e. change round brackets to square, as this is an editorial interpolation by us plausibly have gone to per-

sons who made no direct contribution to innovation. The rather startling implication of all this is that the spillovers of innovation, both direct and indirect, can be estimated to constitute well over half of current [US] GDP – and it can even be argued that this is a very conservative figure" (see p. 135 in [32]). And secondly, the Internet and the Web have made it harder to preserve monopoly rights to information, as copying and distribution reduces the marginal cost to producers to close to zero. Although many media companies have taken rearguard action to protect their intellectual property, so simple is the distribution model on the Web that the basis of the value of information is rapidly switching from scarcity to *abundance*. It is the large quantity of data, that can be placed in novel and unintended contexts with little cost, that makes it increasingly valuable in the age of digital technologies – and it is this abundance that the SW is designed to foster.

One of the major drivers of the SW has been the transformation of science into *e-science*, a computer-enabled, data-heavy view of science as the analysis of the very large quantities of information that improved instrumentation, larger computer power, more prevalent sensor networks and greater memory storage have released. Several disciplines have seized on the opportunity to exploit such data, which are often available only in diverse and heterogeneous datasets. In particular, interdisciplinary research is growing in importance, requiring data developed in different disciplines, using a confusion of vocabularies and methods of collection. Methods for dealing with such large and heterogeneous datasets are required in many areas, including the life sciences, climate research, medicine and epidemiology, and genomics, to name but four, which explains the interest in many of these fields in the SW.

The use of the SW in such large, public projects was perhaps predictable, but much debate and discussion has focused more on the individual's interface to it. We discuss this in more detail below, but one reason for this was that the landmark publication for the public view of the SW, an article in *Scientific American* for 2001, written by Berners-Lee, James Hendler and Ora Lassila [37], developed the idea of a Web of Data with a number of household gadgets interfacing with it. The possibilities envisaged in their scenario included: a telephone that turned down the volume of all local devices with volume controls when it rang; an agent that could plan a program of medical care; and a calendar that could integrate this information to adjust a set of appointments. The point of the article was not the impressive set of agents, but rather the Web of Data that sat underneath them. However, many readers focused on the gadgets, and – given that such gadgets are not at the time of writing very common or effective – have concluded either that the SW has been a failure, or that it was an unrealistic vision from the beginning (see Sect. "Controversies").

In 2006, a further publication by Berners-Lee, together with Nigel Shadbolt and Wendy Hall, appeared in the publication *IEEE Intelligent Systems* de-emphasising the agents and focusing on the idea of the SW as a Web of Data or actionable information [101]. This paper argued that the agents described in 2001 could only flourish when standards for data sharing are well-established. The need for such standards, and for SW technologies in general, was growing, thanks to developments such as e-science, information-based medicine, and e-government.

Since the late 1990s, the World Wide Web Consortium (W3C), under the direction of Berners-Lee, has led the drive to create the standards. Figure 1 shows a diagrammatic representation of the progress of the SW in the development of its layered standards (the hierarchically-arranged layers are marked in Fig. 1 down the left hand column as markup/data/ontology etc.). The development of each layer goes through a long, sometimes tortuous, process of research and discussion. Initial stages of research, sometimes competitive, after some time produce a rough consensus about the general properties of a formalism to implement a layer. At that point, consideration is undertaken about creating a Web standard by the W3C.

To create a standard, a group, representative of stakeholders from academe and private enterprise, is assembled by the W3C, which first creates a *working draft*, which is released under that status for review by the community. Commentary is welcomed, and the draft may be changed dramatically. Once the group responsible for the standard is satisfied that it is fully capable of doing what it required, it is released as a *candidate recommendation*, when it is critiqued in terms of the practicability of its implementation. The next stage is to become a *proposed recommendation*, when it is submitted to the W3C advisory council. Finally, it is released as a W3C recommendation. It generally takes years to negotiate these various stages. Currently, the markup language XML, the data representation and interchange languages RDF and RDF(S), and the ontology language OWL are full recommendations. The next layer up from the ontology layer is that of rules and querying: the query language SPARQL became a W3C recommendation in January 2008 [96], while the rule expression language RIF is at the time of writing at the working draft stage [41].

Figure 1 represents the historical progress of the SW in terms of a 'wave' rolling over 'dry land'. The depth of the 'sea' indicates the extent to which SW standards have been accepted and been deployed widely. The SW wave is traveling from the bottom left of the diagram to the top

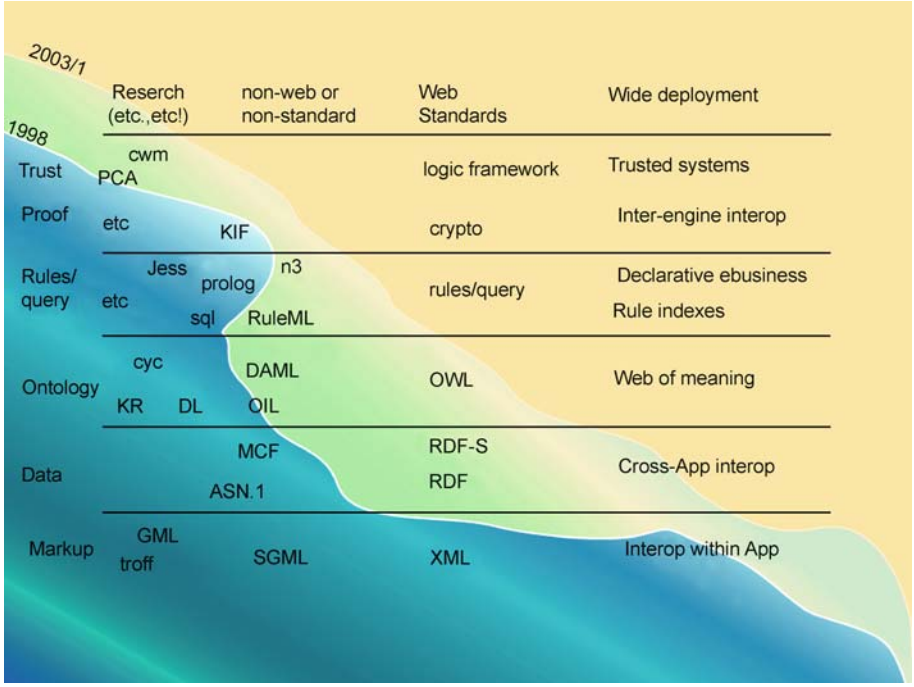

Semantic Web, Figure 1
The wave of development of the Semantic Web [51]

right; the net result of this is that the lower levels of the SW are coming into being and wide acceptance ahead of the higher levels. RDF is, at the time of writing, reasonably widely deployed; SPARQL has recently become a recommendation; trust (the highest level represented here) is still a research issue, and the consensus about what form a standard to represent and promote trust in data is still in the process of formation.

Linking Data

The underlying aim of the SW is to allow data to be explored and queried on the Web, analogously to the way that documents are currently investigated online. One precondition for this is obviously the publishing of data on the Web, but another is to create the links that allow data to be explored. RDF allows representation of data in such a way that anything referred to in the data can be linked to, and from. If URIs are used to name things, common naming schemes are allowed to emerge; one of the most important is the Hypertext Transfer Protocol (HTTP [59]), which affords a straightforward mechanism for people to look up the names. In a properly linked Web of Data, the URI, once looked up, should provide access to useful information about the resource named, as well as useful links out to other data.

Links can be made using various mechanisms, the simplest of which is to use a URI that points to another. For example (taken from [35]), someone might describe some relationships in RDF as follows:

```
<rdf:Description about="#albert"
   <fam:child rdf:Resource="#brian">
   <fam:child rdf:Resource="#carol">
</rdf:Description>
```

This RDF is about three resources given the local identifiers '#albert', '#brian' and '#carol', and might be placed in a file called '<http://example.org/smith>'. The architecture of the Web can use these names to provide a global identifier for the three resources; for instance "http://example.org/smith#albert" refers to #albert, and so on. And now there is a global identifier, links can be made. For instance, a document '<http://example.org/jones>' might contain the following RDF:

```
<rdf:Description about="#denise"
   <fam:child rdf:Resource="#edwin">
   <fam:child rdf:Resource=
                    "http://example.org/smith#carol">
</rdf:Description>
```

Here a series of relationships between resources #denise, #edwin and #carol have been asserted, but the datum

about #carol makes it possible to link to the data in the other file. Someone following the link dereferences the URI, i. e. decomposes 'http://example.org/smith#carol' into two parts: the part before the '#' which gives the name and location of the file; and '#carol' which is the local identifier in that file. Hence the information about #carol in the first file can be accessed thanks to the link included in the second file. The series of links between different resources can be represented, at least on a small scale, graphically, as in Fig. 3. This is the simplest way of linking data, though there are others [35]. And if the URI used for reference is created under a widely-supported system in a community, then the prospects for linking data are that much larger.

One of the main drivers of the SW is the vast quantity of data available in the form of relational databases (RDBs) (RDBs), which often exist in isolation from each other. Each database has its own value, but as argued above, the major source of informational value in the digital age is abundance, the possibilities for serendipitous reuse of data by placing it in fruitful contexts. To that end, a key aim of the SW is to harvest the large amount of data held in RDBs – a much larger quantity than is currently available in the document Web – and to support its amalgamation. The net result will be to facilitate the treatment of all the data as, in effect, sitting in a single queryable database.

Much of the current Web is supported by larger databases that sit below the level of what can be seen on a webpage; these databases are known as the *deep Web*, as opposed to the *shallow Web* on webpages. So, for instance, when one looks at one's bank statement on the shallow Web, the webpage that the bank's site creates uses a much larger quantity of data about one's bank account than is visible at any one time. Hence an alternative way of looking at this role for the SW is as a method for allowing users to query the whole of the deep Web, rather than simply the information released onto the shallow Web by the current set of implemented applications.

An RDB consists of a series of *tables*, which consist of *rows* or *records* of individual data items. Each record consists of a set of values of *fields* or *attributes*. The records (rows) and fields (columns) together can be conceived as a table or matrix (m records and n fields give us an $m \times n$ matrix) [50]. So, for instance, each record may represent an individual person, while a particular field may represent zip codes. The value placed in the zip code field of a person's record is therefore the value of that person's zip code. This tabular structure for RDBs can be mapped straightforwardly into an RDF representation. The record can be seen as the subject of an RDF triple, the field name is a link or property type, while the value of that field is the object of the triple. Hence the person who is represented by the record in the example RDB above would be represented by the first item of the triple; the zip code field would be represented by a zip_code property the second item of the triple; and the value of the zip code would be the third item. In that way, each cell of the RDB matrix is represented by an individual RDF triple, and the total set of triples would represent the entire database [34].

Having said that, there are additional factors about RDBs that are harder to capture in the RDF, and there are many open questions about the export of RDB. For instance, it may be that the database is definitive – that is, the institution holding the data has some responsibility for the data. It may be that the State of Texas holds a database of all Texas vehicle registration numbers; any car not on the database is not, as a matter of definition, registered in Texas. On a smaller scale, some RDBs allow a primary field for a unique identifier for the record, which also holds a significance beyond the particular piece of data. There are various ways of modeling these context-based properties of RDBs, perhaps most likely devolving the representation of such matters to the applications that use the data, via the ontologies, rules or query types that they use. See [34] for a worked example of methods to expose an RDB to the Web.

The process of exposing databases to the Web should not be too prescriptive – the whole point of the Web is as a decentralized collection of linked information, whether in the form of data or documents. The links are entirely democratic, and can be made between any pair of data items, or any pair of documents. This is where the power of the Web's ability to promote serendipitous reuse comes in. Attempting to fix the methods or languages used, or to 'police' the links made, will blur this vision, and create bottlenecks impeding information flow. Indeed, the widespread use of the Web is largely down to its non-prescriptive nature – prescription will simply drive users, who will not want particular information management strategies forced upon them, from the Web.

The result is a particularly untidy situation, unusual in the history of information management. If data, information or knowledge is generated within a single organization or affiliation (the usual situation for information managers before the growth of the Web), then information systems can trade on a number of simplifying assumptions. The size of such repositories can be assumed to be small or medium, and representation schemes would be planned and homogeneous. The quality of information would be likely to be high, and managers' trust in it correspondingly high. But on the Web scale, these assumptions fail. The amount of data available for query may be extremely high, and represented in highly heterogeneous ways – rarely in

the optimal way for the manager's task in hand. Information quality, and trust in that quality, would be very variable. Linking data using common URI schemes will never work perfectly, precisely because there are deliberately no enforcement mechanisms on the Web, and people cannot be forced to use any particular naming convention.

It is primarily for this reason that ontologies have always played a central role in the vision of the SW [57]. It is the ontology that puts the 'semantic' into 'Semantic Web'. Ontologies specify the vocabulary, concepts and relationships of a domain. They must be a rationalization of current practice, and managed and endorsed by a community. They act as a specification of the terms used in discussion. But they should not be too prescriptive – terms change over time, others are in constant dispute. Ontologies need to develop as a discipline or domain develops. Different areas will have different requirements of ontologies; some sciences will have large, publicly-managed ontologies which act as a public vocabulary standard, while others will make do with small, lightweight ontologies that only define the relationships between a few terms. The ontological requirements of any individual application may be quite small; the question for the application developer is whether to develop a small special-purpose ontology for her own individual purposes, or alternatively whether to reuse a larger, better-known ontology that may overspecify vocabulary for her purposes. Much will depend on the usual practices of her wider community. It should be noted that the SW project does *not* require a single overarching ontology, referring to and prescribing everything.

The ontology is key to being able to deal with heterogeneous datasets as described above; the data, and the terms used in it, must be mapped onto other terms held in common. Once this has been done, then databases can be understood in common terms – and, most crucially, the information they hold amalgamated and processed by machines. It is of course obvious to a human user that if one database of people has a field *ZC*, while another has a field *zip_code*, that there is at least a good chance that the two fields refer to the same attribute of people, viz., the zip codes of their addresses; a computer will merely try, and fail, to match the strings identifying each field. But if the computer is referred to an ontology and given mappings from the terms used in the two databases to the ontology's terms, then it can be told about the equivalence, and accordingly its inference space is opened up (for instance, it could make some inferences from the fact that $ZC(X) = zip_code(y)$). Machine processing of this heterogeneous, distributed data is made possible by ontologies.

There are, of course, several issues pertaining to the use of ontologies in the SW, some of which will be discussed in greater detail in the Sect. "Controversies". The development of an ontology for an application is a (potentially large) initial cost to an application developer, and it is likely that ontologies, especially well-known ones, will be reused. To that end, searching for ontologies is likely to be a growth area in the future; there are already specialized search engines, such as Swoogle, dedicated to this task [8,54]. It may also be that one application might use several ontologies (for instance, an interdisciplinary scientific application may well reuse well-known ontologies from each discipline that it crosses). In that case, mappings *between* the ontologies will be important, and such mappings, as opposed to the ontologies themselves, could become the semantic basis for the application [78].

Building an ontology from scratch is always an option, especially if the application requires only relatively lightweight ontological apparatus [89]; again, special-purpose tools, such as Protégé are already available and well-used in the SW community [6,90]. Generating an ontology from an RDB can be done semi-automatically, and then mappings (which will be fairly straightforward, given the method of ontology generation) defined between the database and the ontology [105]. The problem is more complex if the aim is to map a legacy database onto an existing ontology. In particular, the mappings between the database and the ontology can be expected to be quite complex, and therefore very expressive languages will be required to describe them, such as the language R_2O [31].

The Layered Model of the Semantic Web

The Web, as a decentralized construction, cannot be created by fiat or prescription, which would limit its growth and create bottlenecks for information mobility. But to allow the Web of Data to reach fruition at the scale envisaged, several related tasks are required to be performed [101]. As discussed above, the W3C has devoted resources over the last few years to developing formalisms and standards to allow these tasks to be addressed, and the tasks themselves have also been arranged in a series of layers, depicted in a hierarchical diagram. The diagram has evolved with the vision of the SW, but is not dissimilar to its first incarnation, and at the time of writing is seen as in Fig. 2.

As noted above, and in Fig. 1, the development of these layers has been bottom-up, concentrating on the lower levels before work begins on those further up. The lower layers are now often the subject of W3C recommendations, while work on the upper layers remains generally more theoretical, and contains more open research questions.

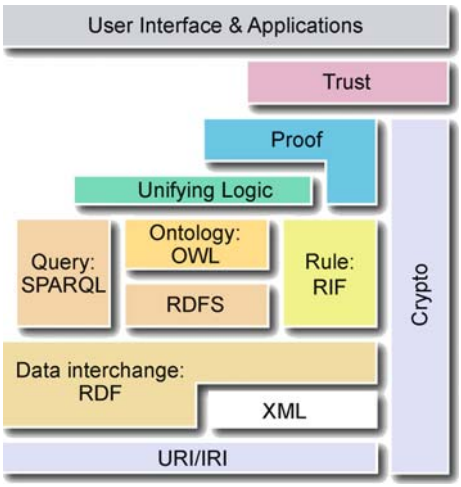

Semantic Web, Figure 2
The layered view of the Semantic Web [39]

At the lowest level we have URIs (and IRIs, Internationalized Resource Identifiers, which are generalizations of URIs allowing non-ASCII characters to be used [56]). URIs identify resources in a global way – in other words, they are interpreted consistently across applications, unlike individual naming conventions, and therefore are central to the vision of a Web of Data [38]. Using a URI to identify a resource (whether that resource be a piece of information, a real-world object, an abstract concept, etc.) allows others to use the same identifier to link to the resource, refer to it, or retrieve a representation of it; this shifts the emphasis online from documents to data, and allows direct machine processing of data. If the URI scheme used is the Hypertext Transfer Protocol (http), then that is particularly helpful as http guides the user as to the location of the resource (although there are several other URI schemes, and indeed users can invent their own).

This ability to refer globally is why the export of data from RDBs to the SW should be facilitated, as noted above, by exporting database objects as first-class objects identified with URIs. A number of SW applications have diverged from this vision by not releasing their data onto the Web, but instead archiving them in inaccessible files (at least sometimes because of privacy concerns); Berners-Lee in particular has complained about this tendency [35]. However, this has happend at least sometimes because of privacy concerns.

The next layer up from names is that of markup and data interchange – the realm of XML and RDF. The eXtensible Markup Language (XML [44]) is a metalanguage for markup – in other words, a way of supporting communication and data interchange within communities by defining specialized vocabularies – commonly used in a number of sectors.

XML, like the Hypertext Markup Language (HTML) that underpins the current Web, is descended from the Standard Generalized Markup Language (SGML), an international standard for defining system-independent methods of representing information, and so has no conceptual connection to the SW. The main language for data interchange on the SW, on the other hand, RDF, is specially designed for the task, by assigning specific URIs to the fields in its triples. Figure 3 shows an RDF graph of nodes and arcs made up of several triples – each triple consists of two labeled nodes, from which is pointing a labeled directed arc. The two nodes are the first and third elements of the triple; the arc is the second (it points *from* the first element *to* the third). The use of URIs to refer to the properties as well as the objects is an important step to providing semantics – it enables us to reason about and link to relationships as well as objects.

Figure 3 shows four triples, all 'about' an individual called Eric Miller, identified by 'http://www.w3.org/People/EM/contact#me'. If we look at these triples clockwise from the right, the first represents a connection between Miller, the property of 'having the name . . . ' (which is referred to by the URI 'http://www.w3.org/2000/10/swap/pim/contact#fullName'), and a character string 'Eric Miller'. The second links Miller, by the property of having a mailbox, to the value of that property, which is his email address given using the common 'mailto:' URI scheme. The third again links Miller with a personal title, the property given as an http URI, and the title as a character string. The fourth triple provides some vocabulary in RDF – it refers to a namespace (an RDF document defining expressive resources which are imported by the RDF graph in Fig. 3) which defines some important relations – and in effect says that Miller (the first object in the triple) is an instance of (a relationship which is the second object in the triple) a person (a class which is the third object in the triple).

Based as it is on triples, RDF is simple yet powerful, exploiting the resources of the common subject/predicate/object structure, and its basis in URIs is very important for the SW. It is a minimalist knowledge representation language for the Web – there are some types of knowledge that cannot be represented in RDF, or only represented with difficulty. For instance, a predicate with more than two arguments has to be represented in a somewhat awkward way as a conjunction of two-argument predicates, while statements about hierarchical class relationships, say, need a further formalism. Furthermore, although the graph structure is quite intuitive, the actual

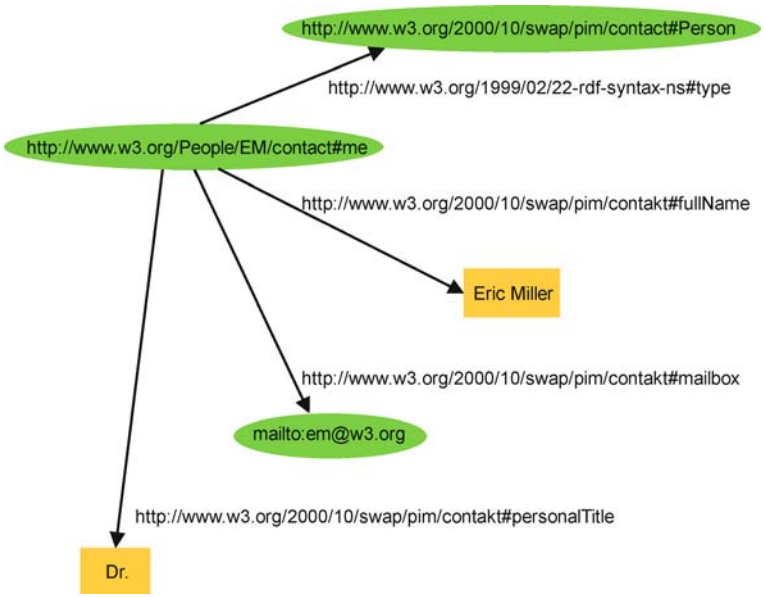

Semantic Web, Figure 3
An RDF graph representing Eric Miller [83]

syntax of RDF is based on XML (it is called RDF/XML) and, although it is well-suited to machine processing, it is not very easy for the human to read (see especially pp. 68–69 in [27]).

The growth in use of RDF has led to the need for special-purpose data stores for holding large quantities of RDF triples (often a set of data will be represented by millions of triples). Such data stores are known as *triplestores*, and need to provide not only storage, but efficient means of reasoning over and retrieving the data that will scale to the large sizes that will be needed. Examples of triplestores include JENA [4], 3store [7,67] and Oracle 11*g* [16].

Greater expressivity is required than is given in RDF, and to that end there another layer upwards which allows the expression of important information about the vocabularies used to express data. As we can see in Fig. 2, the layer here is relatively complex, and contains four boxes, RDFS, Ontology, Rules and Query. These between them provide representation and capabilities that are essential for putting the SW to use.

RDF Schema (RDFS, and sometimes RDF(S) [47]) provides a basic set of tools for producing structured vocabularies that allow different users to agree on particular uses of terms. An extension of RDF, it adds a few modelling primitives with a fixed meaning (such as class, subclass and property relations, and domain and range restriction). It is a basic ontology language that has been adopted fairly widely, and although fairly minimal it can express important constraints on vocabularies.

RDFS is deliberately minimal, and concentrates on expressing subclass and property hierarchies, with various restrictions on these, but the research community, including the Web Ontology Working Group of W3C, identified a number of requirements for greater expressivity for ontologies. As a couple of examples, RDFS allows the stating of subclass relationships, but not, say, that two classes are disjoint; neither does it allow class cardinality restrictions (e. g. a person has *exactly two* parents). As can be seen in Fig. 1, early research efforts into ontology languages led to two leading candidates being developed: DAML (DARPA Agent Markup Language [84]), and OIL (Ontology Interchange Language or Ontology Inference Layer [58]). These two, combined as DAML+OIL [94], became the seed for the W3C ontology language OWL. Unsubstantiated rumor suggests that the fact that it is called 'OWL' and not 'WOL' is an arcane joke: in A. A. Milne's *Winnie-the-Pooh* stories, Owl, who is wise and unusually literate for a forest-dweller, spells his name W-O-L. It is more likely that this is a post hoc rationalization of the naming decision.

The needs of ontology languages are great and potentially problematic. Expressivity is an issue. Even at the level of RDF, its reification mechanism allows the modeler to make statements about statements – an expressive possibility that can lead to logical problems. RDFS has even more powerful modelling primitives, including 'rdfs:Class', the class of all classes. OWL [85] is a strong language for representing concepts and their relations, and its creators needed to wrestle with the inevitable trade-

off between expressivity and efficient reasoning support, with two particular constraints demanded by the Semantic Web. First, there is the strong decentralization and lack of enforcement mechanisms on the Web, so that people cannot be forced to use a language they do not want to. Creating a language this powerful, for all purposes, may well have resulted in it not being used at all. Those who need great expressivity might be inclined to use their own favorite non-standard language, while those who want efficient reasoning might prefer to revert to RDFS. The second constraint is that the layered view of the SW (Fig. 2) makes it desirable that OWL should be an extension of RDFS – it should use the RDF interpretation of classes and properties and add primitives to provide richer expressivity. However, this cannot be the basis for OWL, because the addition of powerful reasoning to the expressive power of RDFS (to define such items as the class of all classes) would result in a language very hard to control.

To pick their way between these various pitfalls, the developers of OWL created three separate languages, OWL Full, OWL DL and OWL Lite [85]. OWL Full is the complete language, a full set of OWL primitives, which can be combined with RDF and RDFS in arbitrary ways. This includes the possibility that an OWL Full ontology could augment or alter the meaning of a pre-defined RDF, RDFS or OWL term (for instance, one could put a cardinality constraint on the size of the class of all classes, thereby limiting the number of possible classes that could be constructed). OWL Full is compatible with RDF, so that any legal RDF document is an OWL Full document. The downside of all this expressivity is that the language is undecidable, which rules out the possibility of complete reasoning support.

OWL DL is intended to be open to computational support, and is a sublanguage of OWL Full (so that any legal OWL DL ontology is a legal OWL Full ontology). OWL DL is so named because it is based on *description logic*, a type of knowledge representation language used to describe the knowledge definitive of a particular application domain [29], and is designed to be complete and decidable (in particular, application of OWL's constructors to each other is restricted). This does mean that full compatibility with RDF is lost: although every legal OWL DL document is a legal RDF document, the inverse is not true.

OWL Lite is a very lightweight language to support users requiring a classification hierarchy with simple constraints. Reasoning support should be relatively efficient, and it is intended to provide a straightforward migration route for bringing thesauri and taxonomies to the SW. A legal OWL Lite ontology is also a legal OWL DL ontology, but the cost of ease of reasoning is a lack of expressivity, so for instance enumerated classes, statements of disjointness and assignment of arbitrary cardinality constraints (i. e. restrictions of class cardinality to any number other than 0 and 1) are disallowed.

Even though the relationship between OWL and RDF is complex, its roots in RDF allow OWL to exploit RDF's linking capabilities to allow ontologies to be distributed across systems. When constructing an ontology in OWL, the developer can refer to terms in other ontologies, which then encourages the sharing of terminology across distributed data sources. Sharing ontologies is not always sufficient when it comes to data sharing – an organization may find that nearly all of an imported ontology is adequate, but it needs extra identifiers and descriptions, and in such a case it should be allowed to add them, rather than build a new ontology from scratch [70]. This ability to assemble distributed ontologies is central to the SW vision.

Expressing ontological relations vital, but having achieved a representation of the domain with semantics, one still needs to make inferences. OWL has some inferential support, such as subsumption and classification, but there are several inferential methods that will be required on the SW. Hence, work is currently ongoing on the Rule Interchange Format (RIF), which is intended to allow a variety of rule-based formalisms, ranging from Horn-clause logics, higher order logics and production systems, to be used [41]. Various insights from Artificial Intelligence (AI) have also been adapted for use for the SW for various purposes, including temporal (time-based) logic, causal logic and probabilistic logics [30,75,101,102].

And given the domain description at its desired level of expressivity and a means of making inferences, then the next important function at this level is the ability to query the data. Query language SPARQL works in effect by constructing a graph of RDF-like triples that may contain variables, which is then matched against the RDF graph to be queried; the query is successful if there is a subgraph of the RDF graph which matches the query when RDF terms are substituted for its variables [96].

Sitting on top of these layers are further layers with a unifying logic, proof systems and trust systems. As can be seen from Fig. 1, these upper layers remain topics of exploratory research.

Trust is perhaps key to widespread application of the SW. If information is being drawn from heterogeneous sources, then it is important that users are able to trust such sources if they are to act on the inferences that result. Trust will of course depend on the criticality of the inferences – trust entails risk, and a risk-averse user will naturally trust fewer sources [42,92]. Measuring trust, however, is a complex problem [62]. A key parameter is that of the

provenance of data, a statement of the conditions under which data were produced (including statements about the methods of production and the organization that carried them out). Methods are appearing to describe provenance [63], but more needs to be discovered about how information spreads across the Web, and therefore how it can be tracked and understood [39].

Related issues include respect for intellectual property, and the privacy of data subjects. In each case the reasoning abilities of the SW can be of value, and initiatives are currently under way to try to exploit them [93]. Protocols that allow users to express their own privacy preferences, and to enable those who wish to reuse information to reason about those preferences, are being created under the program of research into the Policy Aware Web [108]. Creative Commons is an initiative for representing copyright policies and preferences based on RDF to promote reuse where possible (current standard copyright assumptions are deliberately restrictive with respect to reuse) [2]. Cryptography protocols to protect information and privacy will also play an important role at all levels, as shown in Fig. 2.

Applications

The top layer of Fig. 2 is that of a user interface and applications. This recognizes the fact that if the SW cannot be used easily, and integrated into people's workflows in order to add value to their informational transactions, then it will not attract a large user base, without which the network effects already seen in the development of the World Wide Web will not transpire. Network effects are those positive benefits that increase in certain communication systems faster than its user base expands. In the same way as a telephone system is of limited value to a handful of people and enormous value to a large number, a few people exposing data to the SW is unlikely to make much of a difference, whereas if scalable SW technologies were applied to something like the quantity of data to be found currently in the Deep Web, the gains would be immense.

One not entirely frivolous way of expressing the need for the top layer of the SW is to say that its user base needs to grow quickly, and what is needed is a 'killer app', in other words an application that will meet a felt need and create a perception of the technology as 'essential'. Less ambitiously, the SW's spread depends not only on having an impressive set of formalisms, but also the tools to use the linked data [26].

Bootstrapping

One particular user issue is the importance of bootstrapping content for the SW. Even if RDF began to be published routinely, the amount of legacy content on the Web would dwarf new data for some time, and to make this legacy accessible to SW technology some automation of the process of creating RDF from other formats is required. CS AKTive Space [98], discussed in more detail below, amassed a large quantity of information about the state of computer science research in the United Kingdom through a relatively laborious process of harvesting information from the webpages of computer science departments in British universities without necessarily acting as a source, and using natural language processing and an ontology to interpret the data. The application was very successful, but the researchers were SW researchers, and the process is likely to be too onerous to be repeated on a large scale by non-experts. Assumptions can be made about webpage structure (for example, about regular layouts generated from a database by an individual website), and tools have been developed to exploit them [82].

An important development in this field is GRDDL (Gleaning Resource Descriptions from Dialects of Languages) which became a W3C recommendation in September 2007, and which allows the extraction of RDF from XML and XHTML (a further markup language) documents using transformations expressed in XSLT, an extensible stylesheet language based on XML. It is hoped that such extraction could allow bootstrapping of some of the hoped-for SW network effects, given the amount of XML and XHTML data in the Deep Web [52].

Annotating documents and data with metadata about their content, provenance and other useful dimensions (even including the emotional dimension to content – [100]) is also important for the effort to bring more content into the range of SW technologies [64]. Multimedia are a particular focus for research into annotation [106]. Manual annotation is a great burden for information holders, and a major initial cost for the SW, so methods of automating annotation have been investigated by a number of research teams in order to increase the quantity of annotated data available without excessive expenditure of resources [64,65,107].

In addition, as a large quantity of the Web is actually written in natural language, some have seen a role for natural language processing (NLP), and information extraction (IE), for analyzing this text statistically. So large is the Web's store of written language (two thousand billion words) that it can function as a corpus which dwarfs the most ambitious attempts of dedicated corpus builders in computational linguistics of only a few years ago [79]. And given this, and the need for automating or semi-automating annotation, NLP techniques, augmented by ontologies and training with humans, can be used to

extracted machine-readable structured information from plain text [43,48,49,76]. There have also been attempts to build ontologies using NLP techniques, another of the major anticipated bottlenecks for the SW [45]. See also the Subsect. "Commercial and Non-academic Applications" for some examples of SW applications using NLP.

Application Areas

Predicting which particular applications will succeed is unscientific and usually inaccurate. In fact, as it is the rich information contexts that the web of linked data provides that will increase the value of individual pieces of data, one way in which such growth can be encouraged is to focus on small communities with pressing information-processing requirements, and various more-or-less common goals; such communities can be the 'killer apps', or, more accurately, the early adopters of the technology, exactly as the high energy physics discipline played a vital role in the development of the WWW (cf. e. g. [20]). A series of case studies and use cases is maintained at [21].

The most promising of these communities is *e-science*, the data-driven, computationally-intensive pursuit of science in highly distributed computational environments [72]. Large quantities of data are created by analysis and experiments in disciplines such as particle physics, meteorology and the life sciences. Furthermore, in many contexts, different communities of scientists will come together to perform interdisciplinary work, so that data from various fields (e. g. genomics, clinical drug trials and epidemiology) varying not only in vocabulary, but also in the scale of description, need to be integrated. Many scientific disciplines have created large-scale and robust ontologies for this and other purposes. The most well-known of these is the Gene Ontology, a controlled vocabulary to describe gene and gene product attributes in organisms, and related vocabularies developed by Open Biomedical Ontologies. Others include the Protein Ontology, the Cell Cycle Ontology, MeSH (Medical Subject Headings, used to index life science publications), SNOMED (Systematized Nomenclature of Medicine) and AGROVOC (agriculture, forestry, fisheries and food). For more examples and references see [101].

E-government is another important application area, where heterogeneous information of varying quality is deployed widely. Government information varies in provenance, confidentiality and "shelf life" (some information will be good for decades or even centuries, while other information can be out of date within hours), while it can also have been created by various levels of government (national/federal, regional, state, city, parish). Privacy and security are also obviously important factors in this space. Integrating government information in a timely way is clearly an important challenge (see for instance a pilot study for the United Kingdom's Office of Public Sector Information, exploring the use of SW technologies for disseminating, sharing and reusing data held in the public sector [25]).

Academic Applications

Many applications for the SW have been developed with the specific purpose of bringing the SW to maturity. These are often written up in conferences such as the regular World Wide Web Conferences, the International Semantic Web Conferences (ISWC), the European Semantic Web Conferences (ESWC), as well as several one-off conferences and workshops, and can be found in the proceedings (usually online) of any of these. See also the *Journal of Web Semantics* [22].

One initiative of interest here is the Semantic Web Challenge [1], which runs annually alongside the ISWC. This is a good-natured competition to find applications that show SW technology in the best light and which can act as benchmarks for the research community. These applications, therefore, are to some extent an objective list of applications through the years that use semantic technologies to solve real-world problems involving heterogeneous real-world data. The winners of the SW Challenges from its inception, 2003, to the time of writing, 2007, are as follows.

2003: CS AKTive Space (University of Southampton) is an application to explore the UK Computer Science Research domain across multiple dimensions for multiple stakeholders, allowing the tracking of the activities of all agents from funding agencies to individual researchers, using information harvested from the Web, and mediated through an ontology [98].

2004: Flink (Vrije Universiteit Amsterdam) is a 'Who's Who' of the SW which allows the interrogation of information gathered automatically from Web-accessible resources about researchers who have participated in ISWC conferences [86].

2005: CONFOTO (appmosphere web applications, Germany) is a browsing and annotation service for conference photographs [88].

2006: MultimediaN E-Culture Demonstrator (Vrije Universiteit Amsterdam, Centre for Mathematics and

Computer Science, Universiteit van Amsterdam, Digital Heritage Netherlands and Technical University of Eindhoven) searches, navigates and annotates media collections interactively, using digital representations of items from the collections of several well-known museums and art repositories [99].

2007: Revyu.com (Open University) is a reviewing and rating site specifically designed for the SW, allowing reviews of any kind of resource, content or event to be integrated and interlinked with data from other sources (in particular, other reviews, which proliferate on the Web) [69].

A typical SW application will generate a new ontology for its application domain (e. g. art, as with MultimediaN or computer science, as with CS AKTive Space), and use it to interrogate large stores of data, whether legacy data or freshly harvested. This strand of research is tending to confirm the hypothesis that ontologies have an important role in mediating the integration of data from heterogeneous sources.

Commercial and Non-academic Applications

SW applications are generally presented using custom-built interfaces. This suggests a very important area for future research, the development of scalable visualizers capable of navigating the graph of connected information expressed in RDF. As can be seen, the importance of applications and user interfaces was made clear in the layered SW diagram (Fig. 2). However, we shouldn't expect to 'see' the SW in a special browser, in the way that we can see the Web of Documents through browsers such as Internet Explorer, Netscape or Mozilla Firefox. Rather, SW technologies, facilitating the exploration of data, may well work at the back end of websites to improve the user experience. Examples of such sites pointed to by the W3C include Sun's white paper collection site [18], Nokia's developers' discussion forum [11], Oracle's virtual press room [5], and Harper's online magazine [13].

There is an increasing number of applications supporting deeper querying of linked data. The DBpedia [28] is based on the collaborative encyclopedia Wikipedia created by volunteers, and is intended to extract structured information from Wikipedia allowing much more sophisticated querying. Sample queries given on the DPpedia website include a list of people influenced by Friedrich Nietzsche, and the set of images of American guitarists. DBpedia uses RDF, and is also interlinked with other data sources on the Web. When accessed in late 2007, the DBpedia dataset contained 103 million RDF triples. Other examples of linked data applications include the DBLP bibliography of scientific papers [23], and the GeoNames database which represents descriptions of millions of geographical features in RDF [12].

As well as existing organizations using semantic technologies to improve user experience, and applications exploiting linked data, commercial firms are beginning to appear whose business model is based on the possibilities of the SW. Garlik [24] aims to provide individual consumers with more power over their digital data. It reviews what is held about people, harvesting data from the open Web, and represents this in a people-centric structure. Natural Language Processing is used to find occurrences of people's names, sensitive information, and relations to other individuals and organisations (Declaration of interest: Wendy Hall is Chair of the Garlik Advisory Board). Twine [19] aims to facilitate knowledge and information sharing, and to organize that information using various SW technologies (also, like Garlik, using NLP). Twine's developer Nova Spivack coined the term 'knowledge networking' to describe the sharing process, analogous to the Web 2.0 idea of 'social networking'.

Controversies

The SW has been controversial during its history, with several commentators arguing that it is based upon unrealistic expectations, or repeats the mistakes of other initiatives. The arguments against the SW have tended to appear more in the blogosphere rather than the academic world, perhaps because people in the SW world are genuinely enthusiasts while those without confidence in the SW project are doing other things. The pro-SW website GetSemantic supports a wiki page of arguments against the SW, with references and responses [3]. In this section, we will examine three of the most prominent arguments raised against the SW.

The Semantic Web Repeats the Mistakes of "Good Old-Fashioned Artificial Intelligence"

It has been argued that the SW is basically a throwback to the project to program machine intelligence [77] which was jokingly christened by John Haugeland 'GOFAI' (Good Old-Fashioned AI). GOFAI proved impossible: so much of human intelligence is implicit, context-dependent and situated that writing down everything a computer needs to know to produce output that exhibits human-like intelligence is out of the question [68].

One attempt to work around this problem is the Cyc project, set up in 1984, which aims to produce a gigantic ontology that will encode all common-sense knowledge,

in order to support human-like reasoning by machines (i. e. GOFAI) [81]. The project has always aroused controversy, but it is fair to say that over two decades later, Cyc is not nearing completion and has not widely perceived as a solution for the GOFAI problem. The implicit nature of common-sense knowledge arguably makes it impossible to write it all down.

Many commentators have argued that the SW is basically a re-creation of the (misconceived) GOFAI idea, that the aim is to create machine intelligence over the Web, to allow machines to reason about Web content in such a way as to exhibit intelligence [77]. This, however, is a misconception, possibly abetted by the strong focus in the 2001 *Scientific American* article on an agent-based vision of the SW [37], although co-author of that article James Hendler has stated very firmly that he believes that the article was radically misinterpreted, and that no-one "can … say we're advocates of the big AI vision when we explicitly make it clear we're pushing for something else" [71]. The *Scientific American* article states that "Traditional knowledge-representation systems typically have been centralized, requiring everyone to share exactly the same definition of common concepts such as 'parent' or 'vehicle'. But central control is stifling, and increasing the size and scope of such a system rapidly becomes unmanageable" [37].

The SW is not GOFAI reheated, but rather an attempt to facilitate sharing of, and context-based machine reasoning over, content (and therefore the provision of machine-readable data on the Web). The aim is not to bring a single ontology, such as Cyc, to bear on all problems (implicitly defining or anticipating all problems and points of view), but to allow data to be interrogated in ways that were not anticipated by their creators. Different ontologies will be appropriate for different purposes; composite ontologies can be assembled from distributed parts [78,85] and it is frequently very basic ontologies (defining simple terms such as 'customer', 'account number' or 'account balance') that add most value to content. In this respect, the situation in the SW simply mirrors offline life where people from different communities and disciplines can and do interact without making any kind of common *global* ontological commitment [36,39,101]. The engineering challenge, as Berners-Lee et al. argue, is to allow independent consistent data systems to be connected locally without requiring global consistency [40].

Yorick Wilks, accepting that the SW is not an attempt to recreate GOFAI, argues that this is both a gain and a loss: a gain because the knowledge representation structures the SW proposes are computationally tractable, as opposed to the various GOFAI formalisms; a loss because DAML+OIL (and presumably by extension OWL) is less sophisticated than those formalisms, and may not have the representational power for the complexity of the world, whether common-sense or scientific [109]. Equally, as both Wilks and Berners-Lee point out, many in the SW world began their research careers in artificial intelligence, as Shadbolt et al. argue that "it will draw on some key insights, tools and techniques derived from 50 years of AI research" [101].

Ontologies

Ontologies, as we have seen, are vital for the SW vision of a Web of Data, but are perceived by many as expensive to develop and hard to maintain. The ideal conceptual apparatus is relative to the task in hand, and different ontologies are appropriate for different tasks. Classifications are also made relative to some background assumptions, and impose those assumptions onto the resulting ontology. To that extent, the expensive development of ontologies reflects the world view of the ontology builders, not necessarily the users. They are top-down and authoritarian, and therefore opposed to the Web ethos of decentralization and open conversation. They are fixed in advance, and so they don't work very well to represent knowledge in dynamic, situated contexts. [95] argues, for instance, that ontologies do not capture the situated processes of scientific research, the social construction of knowledge or the emergence and evolution of understanding over time, and presents an alternative way of representing this knowledge. [104] implicitly endorses this view, showing how there are issues in biology that OWL DL is not well-equipped to handle.

Other papers have made the point that some types of knowledge are more naturally modeled in ontologies than others, and, while not opposing the use of ontologies, warn against too strong a reliance on them for knowledge representation. [61] argues that ontologies cannot be too ambitious, and attempts to reify the context of an ontology (i. e. to provide context-independent accounts of knowledge) will be undermined by knowledge's situated nature. [91] argues that the social context of knowledge requires application builders to be maximally receptive to diverse types of heterogeneous reasoning, which might use knowledge that is hard to capture in hierarchical structures. See also [46] for a series of short essays debating this point.

A related critical point is that the Web as a decentralized, linked information structure must reflect the pragmatic needs of its large, heterogeneous user base which includes very many people who are naive in their understanding of computing issues. The infrastructure has to be usable widely, which argues for simplicity. The rich link-

ing structure of the Web of Documents, combined with statistically-based search engines such as Google, is much more responsive to the needs of unsophisticated users. The SW, in contrast, demands new information representation, markup and publishing practices, and corporations and information owners need to invest in new technologies. Not only that, but current statistical methods will scale up as the number of users and interactions grows, whereas logic-based methods such as those advocated by the SW, on the other hand, scale less well (cf. e. g. [111]).

The dispute has been fueled by the flowering since 2005 or so of the so-called 'Web 2.0' paradigm (of systems, communities and services facilitating collaboration and information-sharing among users). In particular, it has been argued that the meaningful structures that emerge when sufficiently large numbers of users 'tag' content with key words, structures which have been called 'folksonomies', resulting in a structure of connections and classifications emerging without central control, 'really' express the assumptions of the users, and furthermore in such a way as to respect their familiar patterns of communication and workflow. Meanwhile, ontologies 'really' express the needs of the ontology developers and their sponsors [103].

However, folksonomies are much less expressive than ontologies; they are basically variants on keyword searches. A tag 'SF' may refer to science fiction or San Francisco, even if we make the unrealistic assumption of a monoglot English user community. In a multilingual environment such as the Web, further ambiguity is possible – for instance, 'SF' might refer to the Swiss television station Schweizer Fernsehen. Furthermore, the semantics of Web 2.0 are relatively shallow, with few links and very sparse hierarchies.

When a community is large enough and the benefits clear enough to provide incentives to work together, then a large-scale ontology building and maintenance program is justified. It is true that large fixed costs will tend to skew the effort involved towards authorities who may be unrepresentative [91], but Shadbolt et al. argue explicitly that "the ontologies that will furnish the Semantics for the Semantic Web must be developed, managed, and endorsed by committed practice communities. Whether the subject is meteorology or bank transactions, proteins or engine parts, we need concept definitions we can use" [101].

It is of course an undecided question as to whether this community involvement will transpire, but in a recent note, Berners-Lee argues that such conditions will be perhaps more frequently encountered than sceptics believe. On the broad assumption that the size of an ontology-building team increases on the order of the log of the size of the ontology's user community, and that the resources needed to build an ontology increase on the order of the square of community size, the cost per individual of ontology building will diminish rapidly as community size increases. These assumptions are explicitly intended to be indicative rather than realistic [36].

More to the point, not all ontologies need be of great size and expressive depth. It is certainly not the case that the SW requires a single ontology of all discourse on the model of Cyc. Such an ontology, even if possible, would not scale, and in a decentralized structure like the Web its use could not be enforced. Even in complex scientific domains, [74] argues, using a case study from the field of medical informatics, that ontologies should be firmly based on work practices in the domain. In more mundane applications, we should expect a lot of use of small-scale, *shallow* ontologies defining just a few terms that nevertheless are widely applicable [36].

For example, the machine-readable Friend-of-a-Friend (FOAF) ontology is intended to describe people, their activities and their relations to other people. It is not complex, and publishing a FOAF profile is a fairly simple matter for which there are dedicated tools [15]. The resulting network of people has become very large indeed. A survey performed in 2004 discovered over 1.5 million documents using the FOAF ontology [55].

In any case, ontologies and folksonomies serve different purposes. Folksonomies are based on word tags, whereas the basis for ontology reference is via a URI. One of the main aims of ontology definition is to *remove* ambiguity – not globally, for this may well be impossible, but rather within the particular context of the application. Folksonomies will necessarily inherit the ambiguity from the natural language upon which they are based. Nevertheless, a strong possibility that has been considered is to use cheaply-gathered folksonomies as starting points for ontology development, gradually morphing the Web 2.0 structures into something with greater precision and less ambiguity [73,87].

Symbol Grounding

An important aspect of the SW is that URIs must be interpreted consistently. However, terms and symbols are highly variable in their definitions and use through time and space. The SW project will be boosted by processes whereby URIs are given to objects by communities and individuals, endorsed by the user community, who ensure consistency. Responsible URI 'ownership' is critical to the smooth functioning of the SW [101].

But the process of ensuring a fixed and known link between a symbol and its referent, which has been called *symbol grounding*, is at best hard [66], and at worst impossible [110]. Meanings do not stay fixed, but alter, often imperceptibly. They are delineated not only by logical definitions in terms of necessary and sufficient conditions, but also by procedures, technologies and instrumentation, and alter subtly as practice alters.

Any attempt to fix the reference of URIs is a special case of symbol grounding, and is consequently hard to do globally. Attempting to resist the alteration in community practices and norms, and reformulation of meanings of terms, would be doomed. This is understood by leading developers of the SW, who agree that "communities and practice will change norms, conceptualizations, and terminologies in complex and sociologically subtle ways. We shouldn't be surprised or attempt to resist these reformulations" [101]. But there is an important issue, as the same authors concede. "The issue for a Semantic Web built (in a community-driven way) is to know when parts need revision" [101].

Yorick Wilks has argued that Natural Language Processing techniques are essential for grounding the SW, because of the preponderance of text-based content on the Web. NLP is a vital procedural bridge from texts to knowledge representation, usually via automatic information extraction [109]. Berners-Lee has argued in response to Wilks, at a Web Science Workshop in 2005 that the SW was necessarily based on logic and firm definitions (even if those definitions were imperfect, or highly situated and task-relative), not words, use patterns or statistics. Though meanings are not fully stable, they can be stable *enough* relative to individual applications and in particular contexts to allow the SW approach to work [10]. In the case of large-scale, deep ontologies describing sciences, that perhaps will be where the SW is likely to add most value, the Berners-Lee view is reminiscent of that of Hilary Putnam that scientists are 'guardians' of meaning, who determine the 'true' referent of a word like 'water' [97]. But Berners-Lee agrees that ontologies will need to evolve – some quite quickly, and that such meanings cannot be fixed irrevocably; nevertheless, for the purposes of particular applications, this is unlikely to be a problem in practice [101].

Future Directions

The SW is a work in progress, though Shadbolt et al. argue that the need for shared semantics and a Web of Data have increased, and furthermore that the SW is "attainable" [101]. This final section will sketch some of the anticipated directions of future SW work.

Standards

The most obvious future direction is to continue the research as planned. The development of the SW has been conceived as a tide rolling over a beach, covering some areas fully, enveloping other areas more slowly (Fig. 1). As has been noted, the upper layers of the SW, looking at trust, logic and proof, are relatively underdeveloped, and are the focus for exploratory research at the cutting edge. The lower layers of the SW are in place and deployed widely. The middle layers are more or less in place; OWL and SPARQL are complete, while RIF is progressing, and should become a W3C recommendation in the fullness of time.

The Semantic Grid

Grid computing is a type of distributed computing designed to apply computational power from a number of different distributed, complete computers working in parallel, and in cooperation, on a single problem. For some extremely data-heavy problems requiring a lot of computation (particularly in e-science), grid computing is an important time-saving solution. Particular issues in grid computing include the problems of coordinated resource sharing, distributed problem-solving and the creation of 'virtual organisations' to pool data and share outcomes. The SW, of course, is another distributed computing paradigm where data sharing is a key issue – with the SW, a Web of Data, sharing is the whole point. A third distributed paradigm – software agents – is also a relevant factor.

This synergy has led to a research strand to apply semantic technologies to the problems of grid computing, adding meaning via ontologies and RDF metadata annotations to the grid. Information and services for the grid are thereby given well-defined meaning, which enables the interaction between humans and computers to be better coordinated. In particular, all the components, services and resources are adequately described for machine processing. The use of semantics to describe grid resources is known as the *Semantic Grid*, and research is ongoing [17,53,60,72].

The Policy-Aware Web

As is clear in Fig. 2, trusted systems are very important to the development of the SW. There are two reasons for this. First, if someone is reasoning with heterogeneous data harvested from the Web, then they will need to trust the data they have harvested and are using. As noted above, research is ongoing into methods for specifying the provenance of such data [63]. The second reason is that peo-

ple will not release their data if they thought it would be misused; the importance of data privacy in our digital age is easily underestimated [93]. The *policy-aware Web* is an initiative designed to rectify this problem.

The assumption behind the policy-aware Web is that inflexible and simplistic security systems and access control for the decentralized environment of the Web has hampered its development. Insufficiently sophisticated controls have made people reluctant to share data, particularly with other parties with which they do not have pre-existing information-sharing policies. Furthermore, the Web of Documents is rather coarse-grained for detailed security: the security decision to be made is to grant access to an entire website or page, or not, because policy control mechanisms for access at a finer-grained level aren't available. Thus, despite increasing amounts of useful information residing on the Web in a machine-retrieval form, reluctance to share that information remains.

The aim of policy-aware Web technology is to provide for the publication of access policies in a way that allows significant transparency for sharing among partners without requiring pre-agreement. In addition, greater control over information release can be placed in the hands of the information owner, allowing discretionary (rather than mandatory) access control to flourish. Policies would be another kind of metadata attached to information, and those wishing to use that information would be able to reason about them. For instance, one should be able to specify that the information can only be used by the agent gaining access, and that that agent should not pass the information on. Or it may be specified that the information should be deleted after a certain period of time. Or if it is to be used in a certain manner, then data should be anonymized.

Enforcement of these policies is another matter, but at present the research effort is focused on how to express such policies, and on creating theorem provers to reason about them. The result should be a much more fine-grained security picture, with greater transparency and accountability of information use [108].

Web Science

Although since its inception the Web has revolutionized communication, collaboration and education (particularly within science), relatively little is known about the way it develops. There is a growing feeling among researchers across a number of disciplines that a clear research agenda aimed at understanding the current, evolving and potential Web is needed to assure its continued growth. Such researchers want to model the Web, understand the architectural principles that have provided for its growth, and be as sure as possible that it supports the basic social values of trustworthiness, privacy, and respect for social boundaries, and their solution is to chart out a research agenda that targets the Web as a primary focus of attention [39,40].

This agenda has been dubbed *Web Science*, a combination of analysis of the Web and its dynamics, and synthesis of new languages and protocols. The Web is an engineered space created via formally specified languages but, as humans are the creators of Web pages and links between them, their interactions form emergent patterns in the Web at a macroscopic scale. These human interactions are, in turn, governed by social conventions and laws. Web Science is, therefore, inherently interdisciplinary; its goal is to both understand the growth of the Web and to create approaches allowing new powerful and more beneficial patterns to occur.

Such a research area does not yet exist in a coherent form. Within computer science Web-related research has largely focused on information retrieval algorithms and the algorithms for the routing of information through the underlying Internet. Outside of computing, researchers grow ever more dependent on the Web, but there is no concerted agenda for exploring emerging trends on the Web. Nor are those outside computer science fully engaged with the emerging Web research community to focus more specifically on the needs of science and of society as a whole, while preserving the essential invariants of the Web experience: decentralization to avoid social and technical bottlenecks, openness to the reuse of information in unexpected ways, and freedom and equality of information as it passes across the Web.

Despite excitement about the SW, the majority of the world's data is locked in large data stores and is not published as an open web of inter-referring resources. As a result, the reuse of information has been limited. Substantial research challenges arise in changing this situation. We have already discussed the need for policy controls, and for tools to allow scientists to exploit data when it emerges. But on top of that, releasing data is both a technical and a social problem, and understanding how to free data to the SW is a matter of understanding society in relation to the Web (in social, legal and economic terms) and the Web in relation to society. This is the foundation of the emerging Web Science agenda which it is hoped will inform the development of the SW [101]. The recent foundation of the Web Science Research Initiative (WSRI [9]), a joint venture between the Massachusetts Institute of Technology and the University of Southampton, is intended to drive the agenda on, acting as a focus (e. g. advising in particular on curricula to support it).

Bibliography

Primary Literature

1. http://challenge.semanticweb.org/. Accessed Aug 2008
2. http://creativecommons.org/about/. Accessed Dec 2007
3. http://getsemantic.com/wiki/Arguments_against_the_Semantic_Web. Accessed Dec 2007
4. http://jena.sourceforge.net/. Accessed Dec 2007
5. http://pressroom.oracle.com/. Accessed Dec 2007
6. http://protege.stanford.edu/. Accessed Dec 2007
7. http://sourceforge.net/projects/threestore. Accessed Dec 2007
8. http://swoogle.umbc.edu/. Accessed Dec 2007
9. http://webscience.org/. Accessed Dec 2007
10. http://www.cs.umd.edu/~hendler/2005/WebSciReport.pdf. Accessed Dec 2007
11. http://www.forum.nokia.com/. Accessed Dec 2007
12. http://www.geonames.org/. Accessed Dec 2007
13. http://www.harpers.org/. Accessed Dec 2007
14. http://www.informatik.uni-bremen.de/agki/www/swc/index.html. Accessed Dec 2007
15. http://www.ldodds.com/foaf/foaf-a-matic. Accessed Dec 2007
16. http://www.oracle.com/technology/tech/semantic_technologies/index.html. Accessed Dec 2007
17. http://www.semanticgrid.org/. Accessed Dec 2007
18. http://www.sun.com/servers/wp.jsp. Accessed Dec 2007
19. http://www.twine.com/. Accessed Dec 2007
20. http://www.w3.org/2001/sw/SW-FAQ. Accessed Dec 2007
21. http://www.w3.org/2001/sw/sweo/public/UseCases/. Accessed Dec 2007
22. http://www.websemanticsjournal.org/. Accessed Dec 2007
23. http://www4.wiwiss.fu-berlin.de/dblp/. Accessed Dec 2007
24. https://www.garlik.com/index.php. Accessed Dec 2007
25. Alani H, Dupplaw D, Sheridan J, O'Hara K, Darlington J, Shadbolt N, Tullo C (2007) Unlocking the potential of public sector information with Semantic Web technology. In: Proceedings of the 6th international Semantic Web conference, Busan, 2007. http://iswc2007.semanticweb.org/papers/701.pdf. Accessed Dec 2007
26. Alani H, Kalfoglou Y, O'Hara K, Shadbolt N (2005) Towards a killer app for the Semantic Web. In: Gil Y, Motta E, Benjamins VR, Musen MA (eds) The Semantic Web, proceedings of the international Semantic Web conference, Hiroshima, 2005. Springer, Berlin, pp 829–843
27. Antoniou G, van Harmelen F (2004) A Semantic Web primer. MIT Press, Cambridge
28. Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z (2007) DBpedia: a nucleus for a Web of open data. In: Proceedings of the 6th international Semantic Web conference, Busan, South Korea, 2007. http://iswc2007.semanticweb.org/papers/715.pdf. Accessed Dec 2007
29. Baader F, Calvanese D, McGuinness D, Nardi D, Patel-Schneider P (eds) (2003) The description logic handbook: theory, implementation and applications. Cambridge University Press, Cambridge
30. Baclawski K, Niu T (2005) Ontologies for bioinformatics. MIT Press, Cambridge
31. Barrasa J, Corcho O, Gómez-Pérez A (2004) R_2O, an extensible and semantically based database-to-ontology mapping language. In: 2nd Workshop on Semantic Web and Databases (SWDB2004), Toronto, 2004. http://www.cs.man.ac.uk/~ocorcho/documents/SWDB2004_BarrasaEtAl.pdf. Accessed Dec 2007
32. Baumol WJ (2002) The free-market innovation machine: analyzing the growth miracle of capitalism. Princeton University Press, Princeton
33. Berners-Lee T (1994) Plenary at WWW Geneva 94. http://www.w3.org/Talks/WWW94Tim/. Accessed Dec 2007
34. Berners-Lee T (1998) Relational databases on the Semantic Web. http://www.w3.org/DesignIssues/RDB-RDFhtml. Accessed Dec 2007
35. Berners-Lee T (2006/2007) Linked data. http://www.w3.org/DesignIssues/LinkedData.html. Accessed Dec 2007
36. Berners-Lee T (2007) The fractal nature of the Web. http://www.w3.org/DesignIssues/Fractal.html. Accessed Dec 2007
37. Berners-Lee T, Hendler J, Lassila O (2001) The Semantic Web. Scientific American 284(5):34–43. http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21. Accessed Dec 2007
38. Berners-Lee T, Fielding R, Masinter L (2005) Uniform Resource Identifier (URI): generic syntax. http://gbiv.com/protocols/uri/rfc/rfc3986.html. Accessed Dec 2007
39. Berners-Lee T, Hall W, Hendler JA, O'Hara K, Shadbolt N, Weitzner DJ (2006) A framework for Web Science. Found Trends Web Sci 1(1):1–134 http://www.nowpublishers.com/product.aspx?product=WEB&doi=1800000001. Accessed Dec 2007
40. Berners-Lee T, Hall W, Hendler J, Shadbolt N, Weitzner D (2006) Creating a science of the Web. Science 313(5788):769–771
41. Boley H, Kifer M (2007) RIF basic logic dialect. http://www.w3.org/TR/rif-bld/. Accessed Dec 2007
42. Bonatti PA, Duma C, Fuchs N, Nejdl W, Olmedilla D, Peer J, Shahmehri N (2006) Semantic Web policies – a discussion of requirements and research issues. In: Sure Y, Domingue J (eds) The Semantic Web: research and applications, 3rd European Semantic Web Conference 2006 (ESWC-06), Budva. Springer, Berlin
43. Bontcheva K, Tablan V, Maynard D, Cunningham H (2004) Evolving GATE to meet new challenges in language engineering. Nat Lang Eng 10(3/4):349–373
44. Bray T, Paoli J, Sperberg-McQueen CM, Maler E, Yergeau F (2006) Extensible Markup Language (XML) 1.0 (Fourth Edition). http://www.w3.org/TR/xml/. Accessed Dec 2007
45. Brewster C, Ciravegna F, Wilks Y (2002) User-centred ontology learning for knowledge management. In: Andersson B, Bergholtz M, Johannesson P (eds) Proceedings of the 6th international conference on applications of natural language to information systems. Springer, Berlin, pp 203–207
46. Brewster C, O'Hara K (2004) Knowledge representation with ontologies: the present and future. IEEE Intell Syst 19(1):72–81
47. Brickley D, Guha RV, McBride B (2004) RDF vocabulary description language 1.0: RDF Schema. http://www.w3.org/TR/rdf-schema/. Accessed Dec 2007
48. Ciravegna F, Dingli A, Guthrie L, Wilks Y (2003) Integrating information to bootstrap information extraction from Web sites. In: IJCAI 2003 Workshop on Information Integration on the Web, in conjunction with the 18th International Joint

Conference on Artificial Intelligence (IJCAI 2003), Acapulco, 2003
49. Ciravegna F, Wilks Y (2003) Designing adaptive information extraction for the Semantic Web in Amilcare. In: Handschuh S, Staab S (eds) (2003) Annotation for the Semantic Web. IOS Press, Amsterdam
50. Codd EF (1970) A relational model of data for large shared data banks. Commun ACM 13(6):377–387
51. Connolly D (2003) Semantic Web update: OWL and beyond. http://www.w3.org/2003/Talks/1017-swup/all.htm. Accessed Dec 2007
52. Connolly D (ed) (2007) Gleaning Resource Descriptions from Dialects of Langages (GRDDL). http://www.w3.org/TR/grddl/. Accessed Dec 2007
53. De Roure D, Sure Y (eds) (2006) The Semantic Grid. J Web Semant 4(2):81–123. http://www.websemanticsjournal.org/navigation.html#4. Accessed Dec 2007
54. Ding L, Finin T, Joshi A, Pan R, Cost RS, Peng Y, Reddivari P, Doshi VC, Sachs J (2004) Swoogle: a search and metadata engine for the Semantic Web. In: Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management. ACM Press, pp 652–659. http://ebiquity.umbc.edu/_file_directory_/papers/115.pdf. Accessed Dec 2007
55. Ding L, Zhou L, Finin T, Joshi A (2005) How the Semantic Web is being used: an analysis of FOAF documents. In: Proceedings of the 38th international conference on system sciences. http://ebiquity.umbc.edu/_file_directory_/papers/120.pdf. Accessed Dec 2007
56. Duerst M, Suignard M (2005) Internationalized Resource Identifiers (IRIs). http://www.ietf.org/rfc/rfc3987.txt. Accessed Dec 2007
57. Fensel D (2004) Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce, 2nd edn. Springer, Berlin
58. Fensel D, van Harmelen F, Horrocks I, McGuinness DL, Patel-Schneider PF (2001) OIL: an ontology infrastructure for the Semantic Web. IEEE Intell Syst 16(2):38–45 http://www.cs.man.ac.uk/~horrocks/Publications/download/2001/IEEE-IS01.pdf. Accessed Dec 2007
59. Fielding R, Irvine UC, Gettys J, Mogul JC, Frystyk H, Masinter L, Leach P, Berners-Lee T (1999) Hypertext Transfer Protocol – HTTP/1.1. http://www.w3.org/Protocols/HTTP/1.1/rfc2616.pdf. Accessed Dec 2007
60. Goble C, Kesselman C, Sure Y (eds) (2005) Proceedings of the Dagstuhl seminar on Semantic Grid: the convergence of technologies. http://drops.dagstuhl.de/portals/index.php?semnr=05271. Accessed Dec 2007
61. Goguen JA (2004) Ontology, ontotheology and society. In: International conference on formal ontology in information systems (FOIS 2004). http://charlotte.ucsd.edu/users/goguen/pps/fois04.pdf. Accessed Dec 2007
62. Golbeck J, Hendler J (2004) Accuracy of metrics for inferring trust and reputation in Semantic Web-based social networks. In: Motta E, Shadbolt N, Stutt A, Gibbins N (eds) Engineering Knowledge in the Age of the Semantic Web, Proceedings of 14th International Conference, EKAW 2004, Whittlebury Hall. Springer, Berlin, pp 116–131
63. Groth P, Jiang S, Miles S, Munroe S, Tan V, Tsasakou S, Moreau L (2006) An architecture for provenance systems. http://eprints.ecs.soton.ac.uk/13216/1/provenanceArchitecture10.pdf. Accessed Dec 2007
64. Handschuh S, Staab S (eds) (2003) Annotation for the Semantic Web. IOS Press, Amsterdam
65. Handschuh S, Staab S, Ciravegna F (2002) S-CREAM – Semi-automatic CREAtion of Metadata. In: Gómez-Pérez A, Benjamins VR (eds) Knowledge engineering and knowledge management: ontologies and the Semantic Web, proceedings of 13th international conference, EKAW 2002, Siguënza. Springer, Berlin, pp 358–372
66. Harnad S (1990) The symbol grounding problem. Physica D 42:335–346 http://users.ecs.soton.ac.uk/harnad/Papers/Harnad/harnad90.sgproblem.html. Accessed Dec 2007
67. Harris S, Gibbins N (2003) 3store: efficient bulk RDF storage. In: Proceedings of the 1st International Workshop on Practical and Scalable Systems, Sanibel Island, Florida. http://km.aifb.uni-karlsruhe.de/ws/psss03/proceedings/harris-et-al.pdf. Accessed Dec 2007
68. Haugeland J (1979) Understanding natural language. J Philos 76:619–632
69. Heath T, Motta E (2007) Revyu.com: a reviewing and rating site for the Web of data. In: Proceedings of the 6th international Semantic Web conference 2007, Busan. http://iswc2007.semanticweb.org/papers/889.pdf. Accessed Dec 2007
70. Heflin J (2004) OWL Web Ontology Language use cases and requirements. http://www.w3.org/TR/webont-req/. Accessed Dec 2007
71. Hendler J (2007) Shirkyng my responsibility. http://www.mindswap.org/blog/2007/11/21/shirkyng-my-responsibility. Accessed Dec 2007
72. Hendler J, de Roure D (2004) E-science: the grid and the Semantic Web. IEEE Intell Syst 19(1):65–71
73. Hendler J, Golbeck J (2008) Metcalfe's law, Web 2.0 and the Semantic Web. J Web Semant 6(1):14–20 http://www.websemanticsjournal.org/papers/2007119/MetcalfsLawGolbeckV6I1.pdf. Accessed Dec 2007
74. Hu B, Dasmahapatra S, Dupplaw D, Lewis P, Shadbolt N (2007) Reflections on a medical ontology. Int J Human-comput Stud 65(7):569–582
75. Huang Z, Stuckenschmidt H (2005) Reasoning with multi-version ontologies: a temporal logic approach. In: Proceedings of the 4th International Semantic Web Workshop, http://www.cs.vu.nl/~heiner/public/ISWC05a.pdf. Accessed Dec 2007
76. Iria J, Ciravegna F (2005) Relation extraction for mining the Semantic Web. In: Dagstuhl seminar on machine learning for the Semantic Web. http://tyne.shef.ac.uk/t-rex/pdocs/dagstuhl.pdf. Accessed Dec 2007
77. Jones KS (2004) What's new about the Semantic Web? Some questions. SIGIR forum 38(2) http://www.sigir.org/forum/2004D/sparck_jones_sigirforum_2004d.pdf. Accessed Dec 2007
78. Kalfoglou Y, Schorlemmer M (2003) Ontology mapping: the state of the art. Knowl Eng Rev 18(1):1–31
79. Kilgarrif A, Grefenstette G (2003) Introduction to the special issue on the Web as corpus. Comput Linguist 29(3):333–348 http://www.kilgarriff.co.uk/Publications/2003-KilgGrefenstette-WACIntro.pdf. Accessed Dec 2007
80. Klyne G, Carroll JJ McBride B (2004) Resource Description Framework (RDF): concepts and abstract syntax. http://www.w3.org/TR/rdf-concepts/. Accessed Dec 2007
81. Lenat DB (1995) Cyc: a large-scale investment in knowledge infrastructure. Commun ACM 38(11):33–38

82. Leonard T, Glaser H (2001) Large scale acquisition and maintenance from the Web without source access. In: proceedings of workshop on knowledge markup and semantic annotation, K-CAP2001. http://eprints.ecs.soton.ac.uk/6185/1/Paper.pdf. Accessed Dec 2007
83. Manola F, Miller E, McBride B (2004) RDF primer. http://www.w3.org/TR/rdf-primer/. Accessed Dec 2007
84. McGuinness DL, Fikes R, Stein LA, Hendler J (2003) DAML-ONT: an ontology language for the Semantic Web. In: Fensel D, Hendler J, Lieberman H, Wahlster W (eds) Spinning the Semantic Web: bringing the World Wide Web to its full potential. MIT Press, Cambridge, pp 65–93
85. McGuinness DL, van Harmelen F (2004) OWL Web Ontology Language overview. http://www.w3.org/TR/owl-features/. Accessed Dec 2007
86. Mika P (2005) Flink: Semantic Web technology for the extraction and analysis of social networks. J Web Semant 3(2):211–223 http://www.websemanticsjournal.org/papers/20050719/document7.pdf. Accessed Dec 2007
87. Mika P (2007) Ontologies are us: a unified model of social networks and semantics. J Web Semant 5(1):5–15
88. Nowack B (2005) CONFOTO: A semantic browsing and annotation service for conference photos. In: Gil Y, Motta E, Benjamins VR, Musen MA (eds) The Semantic Web, proceedings of the international Semantic Web conference 2005, Hiroshima. Springer, Berlin, pp 1067–1070
89. Noy NF, McGuinness DL (2001) Ontology development 101: a guide to creating your first ontology. http://smi.stanford.edu/smi-web/reports/SMI-2001-0880.pdf. Accessed Dec 2007
90. Noy NF, Sintek M, Decker S, Crubezy M, Fergerson RW, Musen MA (2001) Creating Semantic Web contents with Protégé-2000. IEEE Intell Syst 16(2):60–71
91. O'Hara K (2004) Ontologies and technologies: knowledge representation or misrepresentation. SIGIR forum 38(2) http://sigir.org/forum/2004D/ohara_sigirforum_2004d.pdf. Accessed Dec 2007
92. O'Hara K, Alani H, Kalfoglou Y, Shadbolt N (2004) Trust strategies for the Semantic Web. In: Workshop on trust, security and reputation on the Semantic Web, 3rd international Semantic Web conference (ISWC 04), Hiroshima, 2004. http://eprints.ecs.soton.ac.uk/10029/. Accessed Dec 2007
93. O'Hara K, Shadbolt N (2008) The spy in the coffee machine: the end of privacy as we know it. Oneworld, Oxford
94. Patel-Schneider P, Horrocks I, van Harmelen F (2002) Reviewing the design of DAML+OIL: an ontology language for the Semantic Web. In: Proceedings of the 18th National Conference on Artificial Intelligence (AAAI02), Edmonton, 2002. http://www.cs.vu.nl/~frankh/postscript/AAAI02.pdf. Accessed Dec 2007
95. Pike W, Gahegan M (2007) Beyond ontologies: toward situated representations of scientific knowledge. Int J Human-comput Stud 65(7):674–688
96. Prud'hommeaux E, Seaborne A (2008) SPARQL query language for RDF. http://www.w3.org/TR/rdf-sparql-query/. Accessed Aug 2008
97. Putnam H (1975) The meaning of 'meaning'. Philosophical papers vol 2: mind, language and reality. Cambridge University Press, Cambridge
98. Schraefel MMC, Shadbolt NR, Gibbins N, Glaser H, Harris S (2004) CS AKTive Space: representing computer science on the Semantic Web. In: Proceedings of WWW 2004, New York, 2004. http://eprints.ecs.soton.ac.uk/9084/. Accessed Dec 2007
99. Schreiber G, Amin A, van Assem M, de Boer V, Hardman L, Hildebrand M, Hollink L, Huang Z, van Kersen J, de Niet M, Omelayenko B, van Ossenbruggen J, Siebes R, Taekema J, Wielemaker J, Wielinga B (2006) MultimediaN e-culture demonstrator. http://www.cs.vu.nl/~guus/papers/Schreiber06a.pdf. Accessed Dec 2007
100. Schröder M, Zovato E, Pirker H, Peter C, Burkhardt F (2007) W3C emotion incubator group report. http://www.w3.org/2005/Incubator/emotion/XGR-emotion/. Accessed Dec 2007
101. Shadbolt N, Hall W, Berners-Lee T (2006) The Semantic Web revisited. IEEE Intell Syst 21(3):96–101
102. Shafer G (1998) 'Causal logic'. In: Proceedings of IJCAI-98. http://www.glennshafer.com/assets/downloads/articles/article62.pdf. Accessed Dec 2007
103. Shirky C (2005) Ontology is overrated: categories, links and tags. http://www.shirky.com/writings/ontology_overrated.html. Accessed Dec 2007
104. Stevens R, Egaña Aranguren M, Wolstencroft K, Sattler U, Drummond N, Horridge M, Rector A (2007) Using OWL to model biological knowledge. Int J Human-comput Stud 65(7):583–594
105. Stojanovic L, Stojanovic N, Volz R (2002) Migrating data-intensive Web sites into the Semantic Web. Symposium on Applied Computing, Madrid
106. Troncy R, van Ossenbruggen J, Pan JZ, Stamou G, Halaschek-Wiener C, Simou N, Tsouvaras V (2007) Image annotation on the Semantic Web. http://www.w3.org/2005/Incubator/mmsem/XGR-image-annotation/. Accessed Dec 2007
107. Vargas-Vera M, Motta E, Domingue J, Lanzoni M, Stutt A, Ciravegna F (2002) MnM: ontology-driven semi-automatic and automatic support for semantic markup. In: Gómez-Pérez A, Benjamins VR (eds) Knowledge engineering and knowledge management: ontologies and the Semantic Web, proceedings of 13th international conference, EKAW 2002, Siguënza, 2002. Springer, Berlin, pp 379–391
108. Weitzner DJ, Hendler J, Berners-Lee T, Connolly D (2005) Creating a policy-aware Web: discretionary, rule-based access for the World Wide Web. In: Ferrari E, Thuraisingham B (eds) Web and information security. Idea Group Inc, Hershey
109. Wilks Y (2008) The Semantic Web as the apotheosis of annotation, but what are its semantics? IEE Intell Syst 23(3):41–49
110. Wittgenstein L (1953) Philosophical investigations. Basil Blackwell, Oxford
111. Zambonini D (2006) The 7 (f)laws of the Semantic Web. http://www.oreillynet.com/xml/blog/2006/06/the_7_flaws_of_the_semantic_we.html [sic]. Accessed Dec 2007

Books and Reviews

Antoniou G, van Harmelen F (2004) A Semantic Web primer. MIT Press, Cambridge

Berners-Lee T (1999) Weaving the Web: the past, present and future of the World Wide Web by its inventor. Texere Publishing, London

Berners-Lee T, Hendler J, Lassila O (2001) The Semantic Web. Scientific American 284(5):34–43. http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21. Accessed Dec 2007

Berners-Lee T, Hall W, Hendler JA, O'Hara K, Shadbolt N, Weitzner DJ (2006) A framework for Web Science. Found Trends Web Sci 1(1):1–134 http://www.nowpublishers.com/product.aspx?product=WEB&doi=1800000001. Accessed Dec 2007
Berners-Lee T, Hall W, Hendler J, Shadbolt N, Weitzner D (2006) Creating a science of the Web. Science 313(5788):769–771
Fensel D (2004) Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce, 2nd edn. Springer, Berlin
Fensel D, Hendler J, Lieberman H, Wahlster W (eds) (2003) Spinning the Semantic Web: bringing the World Wide Web to its full potential. MIT Press, Cambridge
Shadbolt N, Hall W, Berners-Lee T (2006) The Semantic Web revisited. IEEE Intell Syst 21(3):96-101

Semiclassical Spin Transport in Spin-Orbit Coupled Systems

Dimitrie Culcer[1,2]
[1] Advanced Photon Source, Argonne National Laboratory, Argonne, USA
[2] Northern Illinois University, De Kalb, USA

Article Outline

Abstract
Glossary
Definition of the Subject
Introduction
Spin Currents in Electric Fields
Future Directions
Bibliography

Abstract

This article discusses spin transport in systems with spin-orbit interactions and how it can be understood in a semiclassical picture. I will first present a semiclassical wave-packet description of spin transport, which explains how the microscopic motion of carriers gives rise to a spin current. Due to spin non-conservation the definition of the spin current has some arbitrariness. In the second part I will briefly review the physics from a density matrix point of view, which makes clear the relationship between spin transport and spin precession and the important role of scattering.

Glossary

Extrinsic effect An effect which has an explicit dependence on the form or strength of the disorder potential.

Intrinsic effect An effect which does not depend explicitly on the form and strength of the disorder potential.

Semiclassical theory A theory in which a particle's position and momentum are considered simultaneously.

Spin-orbit interaction A relativistic interaction between the spin of a particle and its momentum (which is associated with its orbital motion.)

Steady-state spin current A flow of spins induced by an electric field.

Steady-state spin density A net spin density induced by an electric field.

Definition of the Subject

Spin transport refers to the physical movement of spins across a sample and, if spin were a conserved quantity, one could make a straightforward distinction between spin-up and spin-down charge currents. The recent upsurge of interest in spin transport is, however, motivated by systems in which spin is not conserved due to the presence of spin-orbit interactions, which give rise to spin precession. Here, due to non-conservation of spin the spin current is not well defined [1,2,3]. Spin transport in these cases usually does not involve charge transport as the charge currents in the direction of spin flow cancel out. Finally, in certain materials, spin currents are accompanied by steady-state spin densities. The appearance of a spin density is not a transport phenomenon, but it is a steady-state process and is intimately connected to spin transport.

The word *semiclassical* as used in this work refers to theories which consider the position and momentum of a particle simultaneously. Semiclassical pictures are intuitive and useful in descriptions of transport, particularly in inhomogeneous systems and in spatially dependent fields, which typically vary on length scales much larger than atomic size.

In recent years, steady progress has been made towards realization of convenient semiconducting ferromagnets and spin injection into semiconductors from ferromagnetic metals [4,5,6,7] yet spin injection from a ferromagnetic metal into a semiconductor is hampered by the resistivity mismatch between the two [8]. This is one factor, in addition to basic science, motivating the search for an understanding of the way spins are manipulated electrically. The last few years have seen many experimental advances in spin transport, and spin currents have been measured directly [9,10] and indirectly [11,12,13,14,15].

Introduction

Novel physical phenomena that may lead to improved memory devices and advances in quantum information processing are closely related to spin-orbit interactions [16]. Spin-orbit interactions are present in the band

structure and in potentials due to impurity distributions. Spin-orbit coupling is in principle always present in impurity potentials and gives rise to skew scattering. Band structure spin-orbit coupling may arise from the inversion asymmetry of the underlying crystal lattice [17] (bulk inversion asymmetry), from the inversion asymmetry of the confining potential in two dimensions [18] (structure inversion asymmetry), and may be present also in inversion symmetric systems [19].

Although many observations in this entry are general, the discussion will focus on non-interacting spin-1/2 electron systems, which are pedagogically easier. The Hamiltonian of these systems typically contains a kinetic energy term and a spin-orbit coupling term, $H_{\boldsymbol{k}} = (\hbar^2 k^2)/(2m^*) + H^{\rm so}_{\boldsymbol{k}}$, where m^* is the electron effective mass. In spin-1/2 electron systems, band structure spin-orbit coupling can always be represented as a Zeeman-like interaction of the spin with a wave vector-dependent effective magnetic field $\boldsymbol{\Omega}_{\boldsymbol{k}}$, thus $H^{\rm so}_{\boldsymbol{k}} = (1/2)\boldsymbol{\sigma}\cdot\boldsymbol{\Omega}_{\boldsymbol{k}}$. Common examples of effective fields are the Rashba spin-orbit interaction, [18] which is often dominant in quantum wells with inversion asymmetry, and the Dresselhaus spin-orbit interaction, [17] which is due to bulk inversion asymmetry. The spin operator is given by $s^\sigma = (\hbar/2)\sigma^\sigma$, where σ^σ is a Pauli spin matrix. The spin current operator in these systems will be taken to be $\hat{j}^\sigma_i = (1/2)\{s^\sigma, v^i\}$, where the velocity operator is $v^i = (1/\hbar)\partial H_{\boldsymbol{k}}/\partial k_i$.

An electron spin at wave vector $\mathbf{k}$ precesses about the effective field $\boldsymbol{\Omega}_{\boldsymbol{k}}$ with frequency $\Omega_{\boldsymbol{k}}/\hbar \equiv |\boldsymbol{\Omega}_{\boldsymbol{k}}|/\hbar$ and is scattered to a different wave vector within a characteristic momentum scattering time τ_p. I will assume in this work that $\varepsilon_F\tau_p/\hbar \gg 1$, where ε_F is the Fermi energy, which is equivalent to the assumption that the carrier mean free path is much larger that the de Broglie wavelength. Within this range, the relative magnitude of the spin precession frequency $\Omega_{\boldsymbol{k}}$ and inverse scattering time $1/\tau_p$ define three qualitatively different regimes. In the ballistic (clean) regime no scattering occurs and the temperature tends to absolute zero, so that $\varepsilon_F\tau_p \to \infty$ and $\Omega_{\boldsymbol{k}}\tau_p/\hbar \to \infty$. The weak scattering regime is characterized by fast spin precession and little momentum scattering due to, e. g., a slight increase in temperature, yielding $\varepsilon_F\tau_p/\hbar \gg \Omega_{\boldsymbol{k}}\tau_p/\hbar \gg 1$. In the strong momentum scattering regime $\varepsilon_F\tau_p/\hbar \gg 1 \gg \Omega_{\boldsymbol{k}}\tau_p/\hbar$. I will concentrate on effects originating in the band structure, the observation of which requires the assumption that the materials under study are in the weak momentum scattering regime. Electric fields will be assumed uniform.

The first part of this article will present a semiclassical theory of spin transport, identifying the terms responsible for spin currents in the microscopic dynamics of carriers. Spin non-conservation as a result of spin precession leads to several possible definitions of the spin current, which emerge out of the spin equation of continuity. The second part presents a different point of view, which explains aspects not easily captured in the semiclassical approach. The steady-state density matrix is shown to contain a contribution due to precessing spins and one due to conserved spins. Steady state corrections $\propto \tau_p$ are associated with the *absence* of spin precession and give rise to spin densities in external fields [20,21,22,23,24,25,26,27,28]. Steady state corrections independent of τ_p are associated with spin precession and give rise to spin currents in external fields [1,2,3,9,10,11,12,13,14,15,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46]. Scattering between these two distributions induces significant corrections to steady-state spin currents.

Spin Currents in Electric Fields

Wave-Packet Picture of Spin Transport

This section presents a semiclassical theory of spin transport valid for a general spin-orbit system. The semiclassical method is a suitable approach to the study of transport, because, typically, in the relevant systems the external fields vary smoothly on atomic length scales. All information about the system is taken to be contained in the band structure, thus allowing a description of spin transport which does not make reference to the detailed form of the spin-orbit interaction.

The system under study is regarded as as a collection of carriers, whose semiclassical dynamics in a non-degenerate band i are described by a wave packet [47], with its charge centroid having coordinates $(\mathbf{r}_c, \mathbf{k}_c)$

$$|w_i\rangle = \int \mathrm{d}^3k\, a(\mathbf{k}, t)\mathrm{e}^{i\mathbf{k}\cdot\hat{\mathbf{r}}}|u_i(\mathbf{r}_c, \mathbf{k}, t)\rangle\ . \tag{1}$$

In the above, the function $a(\mathbf{k}, t)$ is a narrow distribution sharply peaked at $\mathbf{k}_c$, the phase of which specifies the center of charge position $\mathbf{r}_c$, while $|u_i(\mathbf{r}_c, \mathbf{k}, t)\rangle$ are lattice-periodic Bloch wave functions. The size of the wave packet in momentum space must be considerably smaller than that of the Brillouin zone. In real space, this implies that the wave packet must stretch over many unit cells.

The external electric field drives the center of the wave packet in k-space according to the semiclassical equations of motion

$$\begin{aligned}\hbar\dot{\mathbf{r}}_c &= \frac{\partial\varepsilon_i}{\partial\mathbf{k}_c} - q\mathbf{E}\times\boldsymbol{\Omega}_i\\ \hbar\dot{\mathbf{k}}_c &= q\mathbf{E}\ ,\end{aligned} \tag{2}$$

with q the charge of the carriers, ε_i the band energy, and the Berry curvature

$$\boldsymbol{\Omega}_i = i\left\langle \frac{\partial u_i}{\partial \mathbf{k}} \right| \times \left| \frac{\partial u_i}{\partial \mathbf{k}} \right\rangle . \tag{3}$$

The electric field also gives rise to an adiabatic correction to the wave functions, which mixes the states making up the wave packet. The wave functions $|u_i\rangle$ therefore have the following form:

$$|u_i\rangle = |\phi_i\rangle - \sum_j \frac{\langle \phi_j | i\hbar \frac{\mathrm{d}}{\mathrm{d}t} | \phi_i \rangle}{\varepsilon_i - \varepsilon_j} |\phi_j\rangle , \tag{4}$$

where the ϕ_i are the unperturbed Bloch eigenstates. The $|u_i\rangle$ form a complete set and retain the Bloch periodicity.

The distribution of carriers is described by a function f. When scattering is present, the distribution function satisfies the following equation:

$$\frac{\partial f}{\partial t} + \dot{\mathbf{r}}_c \cdot \frac{\partial f}{\partial \mathbf{r}_c} + \dot{\mathbf{k}}_c \cdot \frac{\partial f}{\partial \mathbf{k}_c} = \left(\frac{\mathrm{d}f}{\mathrm{d}t}\right)_{\mathrm{coll}} , \tag{5}$$

where $(\frac{\mathrm{d}f}{\mathrm{d}t})_{\mathrm{coll}}$ is the usual collision term. In independent bands, in the relaxation time approximation, the collision term takes the form $(f_0 - f)/\tau_p$, with f_0 the equilibrium distribution and τ_p the momentum relaxation time. In the Boltzmann theory, the change in the distribution function with time arises through the drift terms, which are determined from the semiclassical equations of motion, as well as through scattering with other carriers, with localized impurities or with phonons. For transport in a nondegenerate band, it is consistent to ignore interband scattering effects in the weak scattering limit. In this case the relaxation time is a scalar quantity. The effects of interband coherence due to scattering will be explored in the next section.

In order to obtain expressions for macroscopic quantities of interest, such as densities and currents, one needs to carry out a coarse graining by averaging over microscopic fluctuations. In classical dynamics this coarse graining is performed by means of a sampling function, which is smooth and has a significant magnitude only in a finite range [48]. This range is large compared to atomic dimensions, but small compared to the scale of variation of the distribution function. Moreover, it has a rapidly converging Taylor expansion over distances of atomic dimensions, and its form does not need to be specified. This method has a close analog in wavepacket dynamics, where the sampling function is replaced by a δ-function.

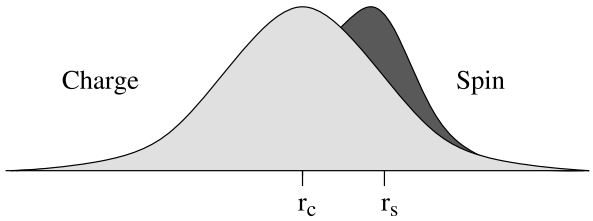

Semiclassical Spin Transport in Spin-Orbit Coupled Systems, Figure 1

For a particle of finite extent the charge and spin distributions in real space in general do not coincide. The same is true of the charge and spin distributions in reciprocal space

It is crucial to recognize that, in general, the center of spin and the center of charge are distinct (Fig. 1), since the wave packet samples a range of wave vectors and the spin is usually a function of $\mathbf{k}$. Following the line of thought outlined above, the spin density is defined to be (henceforth $\mathbf{k}_c$ will be abbreviated to $\mathbf{k}$)

$$S^\sigma(\mathbf{R}, t) = \iint \mathrm{d}^3k\mathrm{d}^3r_c f(\mathbf{r}_c, \mathbf{k}, t) \langle \delta(\mathbf{R} - \hat{\mathbf{r}})\hat{s}^\sigma \rangle , \tag{6}$$

where the bracket indicates quantum mechanical averaging over the wave packet with charge centroid $(\mathbf{r}_c, \mathbf{k})$. As the δ-function has operator arguments, it will be regarded as a *sampling operator*, whose expectation value yields a spatial average, evaluated at position $\mathbf{r}$. To account for the fact that spin is not conserved (Fig. 2), a new quantity is introduced, which will be referred to as the torque density, defined by

$$\mathcal{T}^\sigma(\mathbf{R}, t) = \iint \mathrm{d}^3k\mathrm{d}^3r_c f(\mathbf{r}_c, \mathbf{k}, t) \langle \delta(\mathbf{R} - \hat{\mathbf{r}})\hat{\tau}^\sigma \rangle . \tag{7}$$

$\hat{\tau}^\sigma$ in the above stands for the rate of change of the spin operator, given by $i/\hbar[\hat{H}, \hat{s}^\sigma]$, and symmetrization of products of non-commuting operators has been assumed. Finally, the microscopic spin current density is defined as:

$$\boldsymbol{\mathcal{J}}^\sigma(\mathbf{R}, t) = \iint \mathrm{d}^3k\mathrm{d}^3r_c f(\mathbf{r}_c, \mathbf{k}, t) \langle \delta(\mathbf{R} - \hat{\mathbf{r}})\hat{s}^\sigma \hat{\mathbf{v}} \rangle . \tag{8}$$

We obtain the following continuity equation for the spin density and current:

$$\frac{\partial S^\sigma}{\partial t} + \nabla \cdot \boldsymbol{\mathcal{J}}^\sigma = \mathcal{T}^\sigma + \mathcal{F}^\sigma . \tag{9}$$

The equation of continuity contains a bulk source term, which coincides with the torque density and acts as a mechanism for spin generation. Similar source terms are associated with nonconserved quantities, for example,

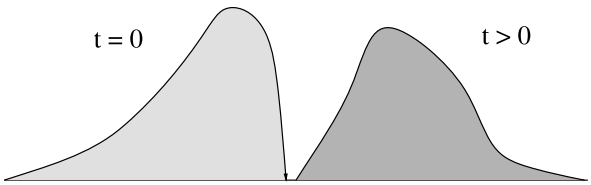

Semiclassical Spin Transport in Spin-Orbit Coupled Systems, Figure 2
In the presence of spin-orbit interactions the spin distribution of a particle changes in time. The horizontal axis may represent position or wave vector

in quantum electrodynamics and in Maxwell's equations. The last term in (9) represents the scattering contribution, which will be discussed further below.

Let us discuss the terms in the equation of continuity, beginning with the spin density. The argument of the sampling operator can be expressed as $[\mathbf{r} - \mathbf{r}_c - (\hat{\mathbf{r}} - \mathbf{r}_c)]$, and, as the second term is of atomic dimensions, the sampling operator can be written as a Taylor expansion about $(\hat{\mathbf{r}} - \mathbf{r}_c)$. The density can therefore be re-expressed, in terms of macroscopic quantities, as

$$S^\sigma(\mathbf{R}, t) = \rho^{s\sigma}(\mathbf{R}, t) - \nabla_{\mathbf{R}} \cdot \mathbf{P}^{s\sigma}(\mathbf{R}, t)\,, \tag{10}$$

where summation over repeated indices has been assumed. In the above, the monopole density is given by

$$\begin{aligned}\rho^{s\sigma}(\mathbf{R}, t) &= \iint \mathrm{d}^3k\mathrm{d}^3r_c f(\mathbf{r}_c, \mathbf{k}, t)\langle\hat{s}^\sigma\rangle\delta(\mathbf{R} - \mathbf{r}_c)\\ &= \int \mathrm{d}^3k f\langle\hat{s}^\sigma\rangle|_{\mathbf{r}_c=\mathbf{R}}\,,\end{aligned} \tag{11}$$

where f in the second line, and henceforth, is to be understood as $f(\mathbf{R}, \mathbf{k}, t)$, and the dipole density is

$$\begin{aligned}\mathbf{P}^{s\sigma}(\mathbf{R}, t) &= \iint \mathrm{d}^3k\mathrm{d}^3r_c f\langle(\hat{\mathbf{r}} - \mathbf{r}_c)\hat{s}^\sigma\rangle\delta(\mathbf{R} - \mathbf{r}_c)\\ &= \int \mathrm{d}^3k f\mathbf{p}^{s\sigma}|_{\mathbf{r}_c=\mathbf{R}}\,.\end{aligned} \tag{12}$$

The average spin of the wave packet has been denoted by $\langle\hat{s}^\sigma\rangle$, and the spin-dipole is defined to be $\mathbf{p}^{s\sigma} = \langle(\hat{\mathbf{r}} - \mathbf{r}_c)\hat{s}^\sigma\rangle|_{\mathbf{r}_c=\mathbf{R}}$. It will be seen that the first term in the density is the average of a monopole density located at $\mathbf{r}_c$, while the dipole term is the average of a point dipole density located at $\mathbf{r}_c$, and similarly for higher orders. The dipole must be understood as the average of the quantum mechanical dipole operator, as an exact analogy with the electric dipole of classical electrodynamics cannot be made. The density can thus be viewed as a collection of point multipoles, located at the centroid of each wave packet. The microscopic distribution of spin is important at the molecular level, but at the macroscopic level the effect of this molecular distribution is replaced by a sum of multipoles. Since the center of spin is different from the center of charge, in principle all multipoles are present.

Following a similar manipulation and using the Boltzmann equation, the torque density is re-expressed as:

$$\mathcal{T}^\sigma(\mathbf{R}, t) = \rho^{\tau\sigma}(\mathbf{R}, t) - \nabla \cdot \mathbf{P}^{\tau\sigma}(\mathbf{R}, t) \tag{13}$$

with the torque monopole density

$$\begin{aligned}\rho^{\tau\sigma}(\mathbf{R}, t) &= \iint \mathrm{d}^3k\mathrm{d}^3r_c f(\mathbf{r}_c, \mathbf{k}, t)\langle\hat{\tau}^\sigma\rangle\delta(\mathbf{R} - \mathbf{r}_c)\\ &= \int \mathrm{d}^3k f\langle\hat{\tau}^\sigma\rangle|_{\mathbf{r}_c=\mathbf{R}}\,,\end{aligned} \tag{14}$$

and the torque dipole density

$$\begin{aligned}\mathbf{P}^{\tau\sigma}(\mathbf{R}, t) &= \iint \mathrm{d}^3k\mathrm{d}^3r_c f(\mathbf{r}_c, \mathbf{k}, t)\langle(\hat{\boldsymbol{R}} - \mathbf{r}_c)\hat{\tau}^\sigma\rangle|_{\mathbf{r}_c=\mathbf{R}}\\ &= \int \mathrm{d}^3k f\mathbf{p}^{\tau\sigma}|_{\mathbf{r}_c=\mathbf{R}}\,.\end{aligned} \tag{15}$$

In analogy with the spin dipole, the torque dipole has been defined as $\mathbf{p}^{\tau\sigma} = \langle(\hat{\mathbf{r}} - \mathbf{r}_c)\hat{\tau}^\sigma\rangle|_{\mathbf{r}_c=\mathbf{R}}$. The torque density is therefore also a sum of multipole moments, that is, the moments of a point spin source located at $\mathbf{r}_c$. Even in the case when the center of $\langle\hat{s}^\sigma\rangle$ coincides with the center of charge, $\langle\hat{\tau}^\sigma\rangle$ may not be centered at $\mathbf{r}_c$, with the result that the higher order terms in the torque density are in general present. The second and higher terms of $\mathcal{T}^\sigma$ cancel exactly the analogous terms in the continuity equation which come from the current.

Since only the gradient of the spin current appears in the equation of continuity, in the expansion of the sampling operator we keep the leading term

$$\mathcal{J}^s(\mathbf{R}, t) = \int \mathrm{d}^3k f\langle\hat{\mathbf{v}}\hat{s}^\sigma\rangle|_{\mathbf{r}_c=\mathbf{R}}\,. \tag{16}$$

Keeping terms to first order in $(\hat{\boldsymbol{r}} - \mathbf{r}_c)$, the current can be decomposed into the following:

$$\mathcal{J}^{s\sigma} = \mathbf{C}^{s\sigma} + \mathbf{D}^{s\sigma} - \mathbf{P}^{\tau\sigma}\,. \tag{17}$$

The convective term $\mathbf{C}^{s\sigma}$ represents the spin being transported along with the wave packet

$$\begin{aligned}\mathbf{C}^{s\sigma}(\mathbf{R}, t) &= \iint \mathrm{d}^3k\mathrm{d}^3r_c f(\mathbf{r}_c, \mathbf{k}, t)\dot{\mathbf{r}}_c\langle\hat{s}^\sigma\rangle\delta(\mathbf{R} - \mathbf{r}_c)\\ &= \int \mathrm{d}^3k f\mathbf{c}^{s\sigma}|_{\mathbf{r}_c=\mathbf{R}}\,,\end{aligned} \tag{18}$$

while $\mathbf{D}^{s\sigma}$ comes from the rate of change of the spin dipole, which has already been introduced. It has the form:

$$\mathbf{D}^{s\sigma}(\mathbf{R}, t) = \int \mathrm{d}^3k f \frac{\mathrm{d}\mathbf{p}^{s\sigma}}{\mathrm{d}t}|_{\mathbf{r}_c=\mathbf{R}} . \tag{19}$$

$\mathbf{P}^{\tau\sigma}$ is the torque dipole introduced above. The corresponding monopole term appears in the source term of the continuity equation, as will emerge below. The presence of the torque dipole here is to be contrasted with the absence of an analogous term in classical electrodynamics. There, an electric dipole arises from the placement of two charges a small distance from each other, but the charge itself is conserved.

Finally, we come to the source term in (9). The first part, composed of the torque density, has already been discussed. The second term, denoted by $\mathcal{F}^\sigma$, becomes, in the relaxation time approximation

$$\mathcal{F}^\sigma = \frac{S_0^\sigma - S^\sigma}{\tau_p} = \frac{1}{\tau_p}\int \mathrm{d}^3k(f_0 - f)\langle \hat{s}^\sigma \rangle , \tag{20}$$

where τ_p is the momentum relaxation time and f_0 the equilibrium distribution, which is usually the Fermi–Dirac distribution function.

Based on the continuity equation alone, there is some flexibility in defining the current and the source. In systems in which spin is conserved, the torque density becomes, to first order in $(\hat{\boldsymbol{r}} - \mathbf{r}_c)$, a pure divergence, which can be incorporated into a redefinition of the spin current. This current, henceforth referred to as the spin transport current, is only due to the convective and spin dipole contributions:

$$\mathbf{J}^{t\sigma}(\mathbf{R}, t) = \mathbf{C}^{s\sigma}(\mathbf{R}, t) + \mathbf{D}^{s\sigma}(\mathbf{R}, t) . \tag{21}$$

With respect to this spin transport current, the continuity equation takes the following form:

$$\frac{\partial S^\sigma}{\partial t} + \nabla \cdot \mathbf{J}^{t\sigma} = \frac{\mathrm{d}}{\mathrm{d}t}\int \mathrm{d}^3k f \langle \hat{s}^\sigma \rangle . \tag{22}$$

In the steady state under a constant electric field, the distribution function is composed of an equilibrium part, independent of the field, and a non-equilibrium part, which is first order in the field. Henceforth, terms in the spin current and source which depend on the equilibrium distribution function will be referred to as intrinsic, whereas the terms depending on the nonequilibrium shift in the distribution will be referred to as extrinsic. For example, the integrand in Eq. (16) can be decomposed into a zero order spin-velocity, $\mathbf{v}\langle \hat{s}^\sigma \rangle$, where $\mathbf{v}$ is the usual group velocity of the band, and a first order correction. Therefore, there will be a contribution to the current from the non-equilibrium part of the distribution and the zero order spin-velocity, which has been discussed extensively in previous work [29,30,31,32,33]. There will also be a contribution from the equilibrium distribution and the first order correction to the spin-velocity, which is referred to as the intrinsic contribution. In the wave packet formalism presented here this effect arises from the change in wave functions induced by the electric field, rather than from the change in distribution functions that is responsible for most conventional transport effects. The intrinsic spin current is calculated from (16) using the equilibrium distribution and the expectation values of the spin and spin dipole operators in a Bloch state perturbed to first order in $\mathbf{E}$.

In its turn, the source in (9) can be decomposed into intrinsic and extrinsic contributions. The present entry considers homogeneous systems, so that all the gradient terms vanish, and the torque density is simply $f\langle \hat{\tau}^\sigma \rangle$. The zeroth order contribution to this term is null, as the Bloch wave functions are eigenstates of the Hamiltonian. Thus, to first order in the electric field, we find that $\langle \hat{\tau}^\sigma \rangle$ is simply given by $(e\mathbf{E}/\hbar) \cdot (\partial\langle \hat{s}^\sigma \rangle/\partial\mathbf{k})$. One is thus justified in replacing f by its equilibrium value f_0, in which case this term is purely intrinsic. The second term in the source, $\mathcal{F}^\sigma$, which depends on the nonequilibrium shift in the distribution function, is entirely extrinsic.

The extrinsic source term takes into account the effect of scattering, and is a term which usually appears in the equation of continuity. The intrinsic source accounts for the effect of all spin-nonconserving terms, and must be present even in a clean system, if the Hamiltonian contains spin-dependent contributions. In general, in addition to the rate of change of spin arising from the spin-dependent terms in the Hamiltonian, scattering processes may alter the orientation of the spin, with the result that any one spin component is not conserved, and the orientation of spins is randomized over a longer time period. For a uniform steady-state system, the current is constant and the intrinsic source term must vanish. However, near the boundary of the system, or at an interface with a different semiconductor with (for example) weaker spin-orbit interactions, the spin current driven by an electric field will vary spatially and $\mathcal{T}$ must reach a non-zero value.

Let us take a closer look at the spin dipole and torque dipole, which are seen to be the main mechanisms responsible for generating the spin current. Because of its narrow distribution in $\mathbf{k}$, the mean spin of the wave packet is $\langle w_i|\hat{s}^\sigma|w_i\rangle = \langle u_i|\hat{s}^\sigma|u_i\rangle$, where it is understood that the wave vector of the Bloch function is set at $\mathbf{k}_c$ and $\hat{s}^\sigma$ is an arbitrary projection of the spin vector operator. The spin

dipole of the wave packet, defined relative to the charge center of the wave packet is given, in terms of Bloch functions, by the expression:

$$\mathbf{p}_i^{s\sigma} = \frac{i}{2}\left[\left\langle u_i|\hat{s}^\sigma|\frac{\partial u_i}{\partial \mathbf{k}}\right\rangle - \left\langle \frac{\partial u_i}{\partial \mathbf{k}}|\hat{s}^\sigma|u_i\right\rangle\right] - \left\langle u_i|i\frac{\partial u_i}{\partial \mathbf{k}}\right\rangle \langle u_i|\hat{s}^\sigma|u_i\rangle \ . \quad (23)$$

Interestingly, the spin dipole is independent of the wave packet width. The expression is also invariant under a local gauge transformation, in the sense that if $|u_i\rangle$ is modified by a phase factor $e^{i\alpha(\mathbf{k})}$ the spin dipole is unchanged.

The torque dipole term has a special interpretation in the case of spin transport. The rate of change of spin is equivalent to a torque, and the torque dipole represents the moment exerted by this torque about the center of the wave packet. The semiclassical expression for the torque moment is

$$\mathbf{p}_i^{\tau\sigma} = \frac{i}{2}\left[\left\langle u_i|\dot{\hat{s}}^\sigma|\frac{\partial u_i}{\partial \mathbf{k}}\right\rangle - \left\langle \frac{\partial u_i}{\partial \mathbf{k}}|\dot{\hat{s}}^\sigma|u_i\right\rangle\right] - \left\langle u_i|i\frac{\partial u_i}{\partial \mathbf{k}}\right\rangle \left\langle u_i|\dot{\hat{s}}^\sigma|u_i\right\rangle \ . \quad (24)$$

The torque moment has the same gauge invariance properties as the spin dipole, and like the spin dipole it also does not depend on the wave packet width.

It is important to note that the spin current, $\boldsymbol{\mathcal{J}}^\sigma$, can be simplified to:

$$\boldsymbol{\mathcal{J}}^\sigma(\mathbf{R}, t) = \int \mathrm{d}^3k f \,\mathrm{tr}\langle u_i|\hat{s}^\sigma \hat{\mathbf{v}}|u_i\rangle \ , \quad (25)$$

which is the semiclassical equivalent of the Kubo formula for spin currents.

Density Matrix Picture of Spin Transport

Semiclassical theory provides a straightforward, intuitive picture of the way spin currents arise in the course of carrier dynamics in an electric field. The theory was developed for independent bands. It turns out that interband coherence arising from scattering is crucial in spin transport, and is difficult to treat semiclassically. Although the semiclassical theory can be generalized to multiple bands [49,50], it is more instructive to examine spin transport from a different point of view that is closer in outlook to the philosophy underlying the Kubo formula (with which the semiclassical theory agrees.) This will shed some light on additional issues, such as the relationships between spin currents and spin precession, between spin currents and spin densities, the complex effect of disorder and the vanishing of spin current in certain systems.

A large, uniform system of non-interacting spin-1/2 electrons is represented by a one-particle density operator $\hat{\rho}$. The expectation value of an observable represented by a Hermitian operator $\hat{O}$ is given by $\mathrm{tr}(\hat{\rho}\hat{O})$ and $\hat{\rho}$ satisfies the quantum Liouville equation

$$\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}t} + \frac{i}{\hbar}[\hat{H} + \hat{U}, \hat{\rho}] = 0 \ . \quad (26)$$

The Liouville equation is projected onto a set of time-independent states of definite wave vector $\{|\mathbf{k}\rangle\mathbf{s}\}$, which are not assumed to be eigenstates of the Hamiltonian $\hat{H}$. The matrix elements of $\hat{\rho}$ in this basis will be written as $\rho_{\mathbf{kk}'} \equiv \rho_{\mathbf{kk}'}^{ss'} = \langle \mathbf{k}s\hat{\rho}|\mathbf{k}'s'\rangle$. Spin indices are not shown explicitly, and $\rho_{\mathbf{kk}'}$ is a matrix in spin space, referred to as the density matrix. In this work we require the expectation values of operators which are diagonal in wave vector, and will thus require the part of the density matrix diagonal in wave vector, $\rho_{\mathbf{kk}} \equiv f_{\mathbf{k}} = n_{\mathbf{k}}\mathbb{1} + S_{\mathbf{k}}$. In the presence of a constant uniform electric field $\mathbf{E}$, $f_{\mathbf{k}} = f_{0\mathbf{k}} + f_{E\mathbf{k}}$, where the equilibrium density matrix $f_{0\mathbf{k}}$ is given by the Fermi–Dirac distribution, and the correction $f_{E\mathbf{k}}$ is due to the $\mathbf{E}$. We subdivide $f_{0\mathbf{k}} = n_{0\mathbf{k}}\mathbb{1} + S_{0\mathbf{k}}$ and $f_{E\mathbf{k}} = n_{E\mathbf{k}}\mathbb{1} + S_{E\mathbf{k}}$. The spin-dependent part of the nonequilibrium correction to the density matrix $S_{E\mathbf{k}}$ is interpreted as the spin density induced by $\mathbf{E}$. The equations governing the time evolution of $n_{E\mathbf{k}}$ and $S_{E\mathbf{k}}$ is [51]

$$\begin{aligned} \frac{\partial n_{E\mathbf{k}}}{\partial t} + \hat{J}_0(n_{E\mathbf{k}}) &= \frac{e\mathbf{E}}{\hbar}\cdot\frac{\partial n_{0\mathbf{k}}}{\partial \mathbf{k}} \\ \frac{\partial S_{E\mathbf{k}}}{\partial t} + \frac{i}{\hbar}[H_{\mathbf{k}}, S_{E\mathbf{k}}] + \hat{J}_0(S_{E\mathbf{k}}) &= \frac{e\mathbf{E}}{\hbar}\cdot\frac{\partial S_{0\mathbf{k}}}{\partial \mathbf{k}} - \hat{J}_s(n_{E\mathbf{k}}) \\ &\equiv \Sigma_{E\mathbf{k}} \ , \end{aligned} \quad (27)$$

where the scalar part of the scattering operator $\hat{J}_0$ and its spin-dependent part $\hat{J}_s$ have been defined in [51]. The equation for $n_{E\mathbf{k}}$ has the well-known solution $n_{E\mathbf{k}} = (e\mathbf{E}\tau_p/\hbar)\cdot(\partial n_{0\mathbf{k}}/\partial \mathbf{k})$, in other words, $n_{E\mathbf{k}}$ describes the shift of the Fermi sphere in the presence of the electric field $\mathbf{E}$, with the momentum relaxation time τ_p. It is seen from Eq. (27) that spin-dependent scattering gives rise to a renormalization of the driving term in the equation for $S_{E\mathbf{k}}$. This renormalization has no analog in charge transport.

We need to find the expectation value of the spin current operator $\hat{\mathcal{J}}_i^\sigma$ defined in the introduction. In the systems under study the spin current operator can be written as $\hat{\mathcal{J}}_i^\sigma = \hbar k_i s^\sigma/m^* + (1/4\hbar)\partial\Omega^\sigma/\partial k_i\mathbb{1}$. We need to determine $S_{E\mathbf{k}}$. To this end we remember that an electron spin at wave vector $\mathbf{k}$ precesses about an effective magnetic field $\boldsymbol{\Omega}_{\mathbf{k}}$. The spin can be resolved into components parallel

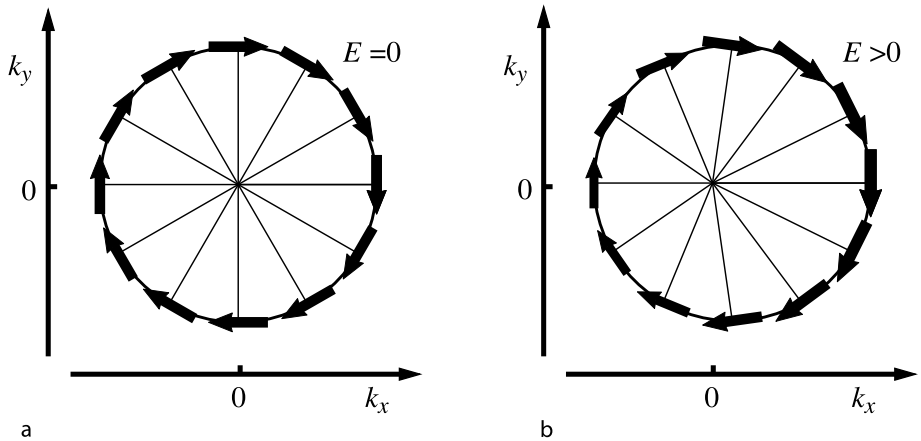

Semiclassical Spin Transport in Spin-Orbit Coupled Systems, Figure 3
Effective field $\boldsymbol{\Omega}_k$ at the Fermi energy in the Rashba model [18] a without ($E = 0$) and b with an external electric field ($E > 0$)

and perpendicular to $\boldsymbol{\Omega}_{\mathbf{k}}$. In the course of spin precession the component of the spin parallel to $\boldsymbol{\Omega}_{\mathbf{k}}$ is conserved, while the perpendicular component is continually changing. Corresponding to this decomposition of the spin is an analogous decomposition of the spin distribution $S_{E\mathbf{k}}$ into a part representing conserved spin and a part representing precessing spin, denoted by $S_{E\mathbf{k}\parallel}$ and $S_{E\mathbf{k}\perp}$ respectively. There is an analogous decomposition of the source on the RHS of Eq. (27) into $\Sigma_{E\mathbf{k}\parallel}$ and $\Sigma_{E\mathbf{k}\perp}$. This decomposition is carried out by introducing projection operators $P_\parallel$ and $P_\perp$ as described in [51], giving for $S_{E\mathbf{k}\parallel}$ and $S_{E\mathbf{k}\perp}$ in the weak momentum scattering limit

$$\frac{\partial S_{E\mathbf{k}\parallel}}{\partial t} + P_\parallel \hat{J}_0(S_{E\mathbf{k}}) = \Sigma_{E\mathbf{k}\parallel}\,, \tag{28a}$$

$$\frac{\partial S_{E\mathbf{k}\perp}}{\partial t} + \frac{i}{\hbar}[H_{\mathbf{k}}, S_{E\mathbf{k}\perp}] = \Sigma_{E\mathbf{k}\perp} - P_\perp \hat{J}_0(S_{E\mathbf{k}})\,. \tag{28b}$$

Equation (28b) shows that scattering mixes the distributions of conserved and precessing spins. This is so because when one spin at wave vector $\mathbf{k}$ and precessing about $\boldsymbol{\Omega}_{\mathbf{k}}$ is scattered to wave vector $\mathbf{k}'$ and precesses about $\boldsymbol{\Omega}_{\mathbf{k}'}$, its conserved component changes, a process which alters the distributions of conserved and precessing spin. Equations (28a,28b) can be solved straightforwardly if one assumes the impurity potential to be short-ranged, obtaining [51]

$$S_{E\mathbf{k}\parallel} = \Sigma_{E\mathbf{k}\parallel}\tau_p + P_\parallel(1-\bar{P}_\parallel)^{-1}\bar{\Sigma}_{E\mathbf{k}\parallel}\tau_p\,, \tag{29a}$$

$$S_{E\mathbf{k}\perp} = \frac{\boldsymbol{\Omega}_{\mathbf{k}} \times (\boldsymbol{\Sigma}_{E\mathbf{k}\perp}\tau_p + P_\perp \bar{\mathbf{S}}_{E\mathbf{k}\parallel}) \cdot \boldsymbol{\sigma}\tau_p}{2\hbar(1+\Omega_{\mathbf{k}}^2\tau_p^2/\hbar^2)} - \frac{(\Sigma_{E\mathbf{k}\perp}\tau_p + P_\perp \bar{S}_{E\mathbf{k}\parallel})}{1+\Omega_{\mathbf{k}}^2\tau_p^2/\hbar^2}\,. \tag{29b}$$

The correction $S_{E\mathbf{k}\parallel}$ does not give rise to a spin current. Inspection of Eq. (29a) shows that integrals of the form $\int \mathrm{d}\theta\, \hat{J}_i^\sigma S_{E\mathbf{k}\parallel}$ contain an odd number of powers of $\mathbf{k}$ and are therefore zero. It can, however, give rise to a nonequilibrium spin density, since integrals of the form $\int \mathrm{d}\theta\, \hat{s}^\sigma S_{E\mathbf{k}\parallel}$ contain an even number of powers of $\mathbf{k}$ and may be nonzero. Similarly $S_{E\mathbf{k}\perp}$ does not lead to a nonequilibrium spin density. The expectation value of the spin operator yields integrals of the form $\int \mathrm{d}\theta\, \hat{s}^\sigma S^{(0)}_{E\mathbf{k}\perp}$, which involve odd numbers of powers of $\mathbf{k}$ and are therefore zero. This term does, however, give rise to nonzero spin currents, since integrals if the form $\int \mathrm{d}\theta\, \hat{J}_i^\sigma S_{E\mathbf{k}\perp}$ contain an even numbers of powers of $\mathbf{k}$ and may be nonzero. Therefore, in the absence of spin-orbit coupling in the scattering potential, nonequilibrium spin currents arise from spin precession (as outlined by Sinova et al. [35]), and nonequilibrium spin densities from the absence of spin precession. The dominant contribution to the nonequilibrium spin density in an electric field exists because in the course of spin precession a component of each individual spin is preserved. For an electron with wave vector $\mathbf{k}$, this spin component is parallel to $\boldsymbol{\Omega}_k$. In equilibrium the average of these conserved components is zero. When an electric field is applied the Fermi surface is shifted and the average of the conserved spin components may be nonzero, as illustrated in Fig. 1. This argument explains why the nonequilibrium spin density $\propto \tau_p^{-1}$ and *requires* scattering to balance the drift of the Fermi surface. Although spin densities in electric fields require band structure spin-orbit interactions and therefore spin precession, the dominant contribution arises as a result of the absence of spin precession.

Systems in which $\boldsymbol{\Omega}_{\boldsymbol{k}}$ is linear in $\mathbf{k}$ are special, in that the spin current as defined in this section vanishes [39,40,41,42,43,44,45,46]. This is because of the renormalization of the driving term on the RHS of Eq. (28b) for $S_{E\mathbf{k}\perp}$, in other words because of scattering from the conserved spin distribution to the precessing spin distribution. In Eq. (29b) it is also clear that if $\Sigma_{E\mathbf{k}\perp}\tau_p + P_\perp \bar{S}_{E\mathbf{k}\|}$ vanishes, then all the contributions to $S_{E\mathbf{k}\perp}$ also vanish. Since $\bar{S}_{E\mathbf{k}\|}$ effectively represents a steady-state spin density, we see that the presence of this spin density tends to diminish the spin current. In systems with energy dispersion linear in $\mathbf{k}$ it cancels the spin current completely.

Future Directions

Whereas the community appears to be in agreement that spin currents exist and are measurable, many questions remain unanswered. Theoretically, intrinsic and extrinsic effects (such as due to skew scattering and side jump) have not been studied on the same footing for an arbitrary form of band structure spin-orbit interactions. The relative magnitude of intrinsic and extrinsic spin currents in such a general system remains to be determined. Also, different definitions of the spin current give results that often differ by a sign [1,2]. The relationship between spin current and spin accumulation at the boundary is not clear, again thanks to the non-conservation of spin. It appears that what happens at the boundary is sensitive to the type of boundary conditions assumed. Thus so far as quantitative interpretation of experimental data is concerned, theory has some way to go.

Despite tremendous progress, experiment is still searching for a reliable way to *measure*, as opposed to *detect*, spin currents directly. Practically, the question of what to do with spin once it has been transported/generated remains. The revolutionary electronic device that harnesses spin currents for a practical purpose remains to be made, and the challenge of its design confronts experimentalists and theorists alike.

The research at Argonne National Laboratory was supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357.

Bibliography

1. Shi J, Zhang P, Xiao D, Niu Q (2006) Proper definition of spin current in spin-orbit coupled systems. Phys Rev Lett 96:076604
2. Sugimoto N, Onoda S, Murakami S, Nagaosa N (2006) Spin hall effect of a conserved current: Conditions for a nonzero spin hall current. Phys Rev B 73:113305
3. Wang Y, Xia K, Su Z-B, Ma Z (2006) Consistency in formulation of spin current and torque associated with a variance of angular momentum. Phys Rev Lett 96:066601
4. Fiederling R et al (1999) Injection and detection of a spin-polarized current in a light-emitting diode. Nature 402:787
5. Ohno Y et al (1999) Electrical spin injection in a ferromagnetic semiconductor heterostructure. Nature 402:790
6. Jiang X et al (2003) Optical detection of hot-electron spin injection into GaAs from a magnetic tunnel transistor source. Phys Rev Lett 90:256603
7. Jonker BT (2003) Progress toward electrical injection of spin-polarized electrons into semiconductors. Proceedings IEEE 91:727
8. Schmidt G et al (2000) Fundamental obstacle for electrical spin injection from a ferromagnetic metal into a diffusive semiconductor. Phys Rev B 62:R4790
9. Valenzuela SO, Tinkham M (2006) Direct electronic measurement of the spin Hall effect. Nature 442:176
10. Liu B, Shi J, Wang W, Zhao H, Li D, Zhang S-C, Xue Q, Chen D (2006) Experimental observation of the inverse spin hall effect at room temperature. cond-mat/0610150
11. Kato YK, Myers RC, Gossard AC, Awschalom DD (2004) Observation of the spin hall effect in semiconductors. Science 306:1910
12. Wunderlich J, Kaestner B, Sinova J, Jungwirth T (2005) Experimental observation of the Spin-Hall Effect in a two-dimensional Spin-Orbit coupled semiconductor system. Phys Rev Lett 94:047204
13. Sih V, Myers RC, Kato YK, Lau WH, Gossard AC, Awschalom DD (2005) Spatial imaging of the spin Hall effect and current-induced polarization in two-dimensional electron gases. Nature Phys 1:31–35
14. Stern NP, Ghosh S, Xiang G, Zhu M, Samarth N, Awschalom DD (2006) Current-induced polarization and the Spin Hall Effect at room temperature. Phys Rev Lett 97:126603
15. Ganichev SD, Ivchenko EL, Danilov SN, Eroms J, Wegscheider W, Weiss D, Prettl W (2001) Conversion of spin into directed electric current in quantum wells. Phys Rev Lett 86:4358
16. Žutić I, Fabian J, Das Sarma S (2004) Spintronics: Fundamentals and applications. Rev Mod Phys 76:323
17. Dresselhaus G (1955) Spin-Orbit coupling effects in zinc blende structures. Phys Rev 100:580
18. Bychkov YA, Rashba EI (1984) Properties of a 2D electron gas with lifted spectral degeneracy. JETP Lett 39:78
19. Luttinger JM (1956) Quantum theory of cyclotron resonance in semiconductors: General theory. Phys Rev 102:1030
20. Ivchenko EL, Pikus GE (1978) New photogalvanic effect in gyrotropic crystals. JETP Lett 27:604
21. Levitov LS, Nazarov YV, Eliashberg GM (1985) Magnetoelectric effects in conductors with mirror isomer symmetry. Sov Phys JETP 61(1):133
22. Edelstein VM (1990) Spin polarization of conduction electrons induced by electric current in two-dimensional asymmetric electron systems. Solid State Comm 73:233
23. Aronov AG, Lyanda-Geller YB, Pikus GE (1991) Spin polarization of electrons by an electric current. Sov Phys JETP 73:537
24. Magarill LI, Chaplik AV, Entin MV (2001) Spin response of 2D electrons to a lateral electric field. Semiconductors 35(9):1081
25. Vorob'ev LE, Ivchenko EL, Pikus GE, Farbstein II, Shalygin VA, Sturbin AV (1979) Optical activity in tellurium induced by a current. JETP Lett 29:441

26. Kato YK, Myers RC, Gossard AC, Awschalom DD (2004) Current-induced spin polarization in strained semiconductors. Phys Rev Lett 93:176601
27. Ganichev SD, Danilov SN, Schneider P, Bel'kov VV, Golub LE, Wegscheider W, Weiss D, Prettl W (2004) Can an electric current orient spins in quantum wells? cond-mat/0403641
28. Silov AY, Blajnov PA, Wolter JH, Hey R, Ploog KH, Averkiev NS (2004) Current-induced spin polarization at a single heterojunction. Appl Phys Lett 85:5929
29. D'yakonov MI, Perel' VI (1971) Spin orientation of electrons associated with the interband absorption of light in semiconductors. Sov Phys JETP 33:1053–1059
30. Hirsch JE (1999) Spin Hall Effect. Phys Rev Lett 83:1834
31. Zhang S (2000) Spin Hall Effect in the presence of spin diffusion. Phys Rev Lett 85:393
32. Qi Y, Zhang S (2002) Crossover from diffusive to ballistic transport properties in magnetic multilayers. Phys Rev B 65:214407 (ibid 67:052407 (2003))
33. Engel H-A, Halperin BI, Rashba EI (2005) Theory of Spin Hall Conductivity in n-doped GaAs. Phys Rev Lett 95:166605
34. Murakami S, Nagaosa N, Zhang S-C (2003) Dissipationless quantum spin current at room temperature. Science 301:1348
35. Sinova J, Culcer D, Niu Q, Sinitsyn NA, Jungwirth T, MacDonald AH (2004) Universal intrinsic Spin Hall Effect. Phys Rev Lett 92:126603
36. Adagideli I, Bauer GEW (2005) Intrinsic Spin Hall edges. Phys Rev Lett 95:256602
37. Rashba EI (2004) Sum rules for Spin Hall conductivity cancellation. Phys Rev B 70:201309
38. Zhang S, Yang Z (2005) Intrinsic spin and orbital angular momentum hall effect. Phys Rev Lett 94:066602
39. Inoue J, Bauer GEW, Molenkamp LW (2004) Suppression of the persistent spin Hall current by defect scattering. Phys Rev B 70:041303
40. Dimitrova OV (2005) Spin-Hall conductivity in a two-dimensional Rashba electron gas. Phys Rev B 71:245327
41. Schwab P, Raimondi R (2002) Magnetoconductance of a two-dimensional metal in the presence of spin-orbit coupling. Eur Phys J B 25:483
42. Mishchenko EG, Shytov AV, Halperin BI (2004) Spin current and polarization in impure two-dimensional electron systems with spin-orbit coupling. Phys Rev Lett 93:226602
43. Khaetskii A (2006) Nonexistence of intrinsic Spin Currents. Phys Rev Lett 96:056602
44. Shytov AV, Mishchenko EG, Engel H-A, Halperin BI (2006) Small-angle impurity scattering and the spin hall conductivity in two-dimensional semiconductor systems. Phys Rev B 73:075316
45. Liu SY, Lei XL, Horing NJM (2006) Vanishing spin-hall current in a diffusive Rashba two-dimensional electron system: A quantum Boltzmann equation approach. Phys Rev B 73:035323
46. Mal'shukov AG, Chao KA (2005) Spin Hall conductivity of a disordered two-dimensional electron gas with Dresselhaus spin-orbit interaction. Phys Rev B 71:121308
47. Sundaram G, Niu Q (1999) Wave-packet dynamics in slowly perturbed crystals: Gradient corrections and Berry-phase effects. Phys Rev B 59:14915
48. Jackson JD (1999) Classical Electrodynamics, 3rd edn. Wiley, New York, section 6.6, pp 248–258
49. Culcer D, Yao Y, Niu Q (2005) Coherent wave-packet evolution in degenerate bands. Phys Rev B 72:085110
50. Culcer D, Niu Q (2006) Geometrical phase effects in the Wigner distribution of Bloch electrons. Phys Rev B 74:035209
51. Culcer D, Winkler R (2007) Steady states of spin distributions in the presence of spin-orbit interactions. Phys Rev B 76:245322

Shallow Water Waves and Solitary Waves

Willy Hereman
Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Completely Integrable Shallow Water Wave Equations
Shallow Water Wave Equations of Geophysical Fluid Dynamics
Computation of Solitary Wave Solutions
Water Wave Experiments and Observations
Future Directions
Acknowledgments
Bibliography

Glossary

Deep water A surface wave is said to be in deep water if its wavelength is much shorter than the local water depth.

Internal wave A internal wave travels within the interior of a fluid. The maximum velocity and maximum amplitude occur within the fluid or at an internal boundary (interface). Internal waves depend on the density-stratification of the fluid.

Shallow water A surface wave is said to be in shallow water if its wavelength is much larger than the local water depth.

Shallow water waves Shallow water waves correspond to the flow at the free surface of a body of shallow water under the force of gravity, or to the flow below a horizontal pressure surface in a fluid.

Shallow water wave equations Shallow water wave equations are a set of partial differential equations that describe shallow water waves.

Solitary wave A solitary wave is a localized gravity wave that maintains its coherence and, hence, its visibility through properties of nonlinear hydrodynamics. Solitary waves have finite amplitude and propagate with constant speed and constant shape.

Soliton Solitons are solitary waves that have an elastic scattering property: they retain their shape and speed after colliding with each other.

Surface wave A surface wave travels at the free surface of a fluid. The maximum velocity of the wave and the maximum displacement of fluid particles occur at the free surface of the fluid.

Tsunami A tsunami is a very long ocean wave caused by an underwater earthquake, a submarine volcanic eruption, or by a landslide.

Wave dispersion Wave dispersion in water waves refers to the property that longer waves have lower frequencies and travel faster.

Definition of the Subject

The most familiar water waves are waves at the beach caused by wind or tides, waves created by throwing a stone in a pond, by the wake of a ship, or by raindrops in a river (see Fig. 1). Despite their familiarity, these are all different types of water waves. This article only addresses shallow water waves, where the depth of the water is much smaller than the wavelength of the disturbance of the free surface. Furthermore, the discussion is focused on gravity waves in which buoyancy acts as the restoring force. Little attention will we paid to capillary effects, and capillary waves for which the primary restoring force is surface tension are not covered.

Although the history of shallow water waves [15,20,22] goes back to French and British mathematicians of the eighteenth and early nineteenth century, Stokes [71] is considered one of the pioneers of hydrodynamics (see [21]). He carefully derived the equations for the motion of incompressible, inviscid fluid, subject to a constant vertical gravitational force, where the fluid is bounded below by an impermeable bottom and above by a free surface. Starting from these fundamental equations and by making further simplifying assumptions, various shallow water wave models can be derived. These shallow water models are widely used in oceanography and atmospheric science.

Shallow Water Waves and Solitary Waves, Figure 1
Capillary surface waves from raindrops. Photograph courtesy of E. Scheller and K. Socha

This article discusses shallow water wave equations commonly used in oceanography and atmospheric science. They fall into two major categories: Shallow water wave models with wave dispersion are discussed in Sect. "Completely Integrable Shallow Water Wave Equations" Most of these are completely integrable equations that admit smooth solitary and cnoidal wave solutions for which computational procedures are outlined in Sect. "Computation of Solitary Wave Solutions". Sect. "Shallow Water Wave Equations of Geophysical Fluid Dynamics" covers classical shallow water wave models without dispersion. Such hyperbolic systems can admit shocks. Sect. "Water Wave Experiments and Observations" addresses a few experiments and observations. The article concludes with future directions in Sect. "Future Directions".

Introduction

The initial observation of a solitary wave in shallow water was made by John Scott Russell, shown in Fig. 2. Russell was a Scottish engineer and naval architect who was conducting experiments for the Union Canal Company to design a more efficient canal boat.

In Russell's [63] own words: "I was observing the motion of a boat which was rapidly drawn along a narrow channel by a pair of horses, when the boat suddenly stopped – not so the mass of water in the channel which it had put in motion; it accumulated round the prow of the vessel in a state of violent agitation, then suddenly leaving it behind, rolled forward with great velocity, assuming the form of a large solitary elevation, a rounded, smooth and well-defined heap of water, which continued its course along the channel apparently without change of form or diminution of speed. I followed it on horseback, and overtook it still rolling on at a rate of some eight or nine miles an hour, preserving its original figure some thirty feet long and a foot to a foot and a half in height. Its height gradually diminished, and after a chase of one or two miles I lost it in the windings of the channel. Such, in the month of August 1834, was my first chance interview with that singular and beautiful phenomenon which I have called the Wave of Translation."

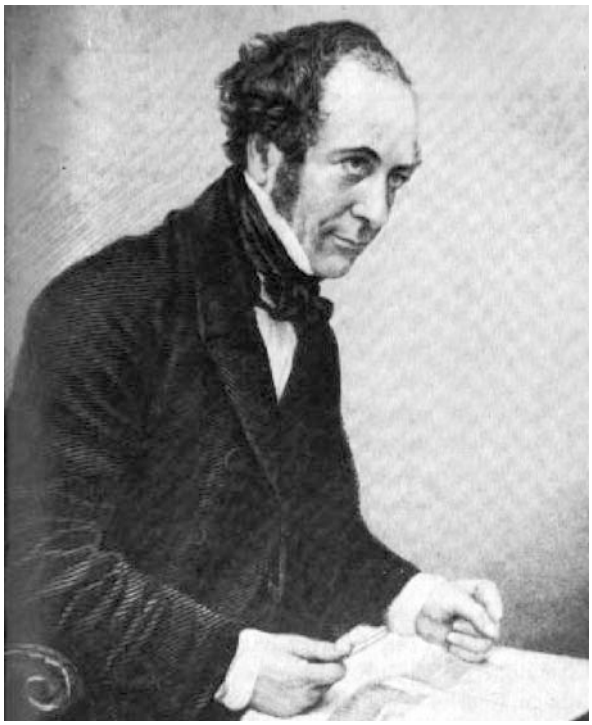

Shallow Water Waves and Solitary Waves, Figure 2
John Scott Russell. Source: [27]. Courtesy of John Murray Publishers

Russell built a water tank to replicate the phenomenon and research the properties of the solitary wave he had observed. Details can be found in a biography of John Scott Russell (1808–1882) by Emmerson [27], and in review articles by Bullough [15], Craik [20], and Darrigol [22], who pay tribute to Russell's research of water waves.

In 1895, the Dutch professor Diederik Korteweg and his doctoral student Gustav de Vries [47] derived a partial differential equation (PDE) which models the solitary wave that Russell had observed. Parenthetically, the equation which now bears their name had already appeared in seminal work on water waves published by Boussinesq [13,14] and Rayleigh [59]. The solitary wave was considered a relatively unimportant curiosity in the field of nonlinear waves. That all changed in 1965, when Zabusky and Kruskal realized that the KdV equation arises as the continuum limit of a one dimensional anharmonic lattice used by Fermi, Pasta, and Ulam [29] to investigate "thermalization" – or how energy is distributed among the many possible oscillations in the lattice. Zabusky and Kruskal [78] simulated the collision of solitary waves in a nonlinear crystal lattice and observed that they retain their shapes and speed after collision. Interacting solitary waves merely experience a phase shift, advancing the faster and retarding the slower. In analogy with colliding particles, they coined the word "solitons" to describe these elastically colliding waves. A narrative of the discovery of solitons can be found in [77].

Shallow Water Waves and Solitary Waves, Figure 3
Recreation of a solitary wave on the Scott Russell Aqueduct on the Union Canal. Photograph courtesy of Heriot–Watt University

Since the 1970s, the KdV equation and other equations that admit solitary wave and soliton solutions have been the subject of intense study (see, e. g., [23,30,60]). Indeed, scientists remain intrigued by the physical properties and elegant mathematical theory of the shallow water wave models. In particular, the so-called completely integrable models which can be solved with the Inverse Inverse scattering transform (IST). For details about the IST method the reader is referred to Ablowitz et al. [3], Ablowitz and Segur [2], and Ablowitz and Clarkson [1]. The completely integrable models discussed in the next section are infinite-dimensional Hamiltonian systems, with infinitely many conservation laws and higher-order symmetries, and admit soliton solutions of any order.

As an aside, in 1995, scientists gathered at Heriot–Watt University for a conference and successfully recreated a solitary wave but of smaller dimensions than the one observed by Russell 161 years earlier (see Fig. 3).

Completely Integrable Shallow Water Wave Equations

Starting from Stokes' [71] governing equations for water waves, completely integrable PDEs arise at various levels of approximation in shallow water wave theory. Four length scales play a crucial role in their derivation. As shown in Fig. 4, the wavelength λ of the wave measures the distance between two successive peaks. The amplitude a measures the height of the wave, which is the varying distance between the undisturbed water to the peak of the wave. The

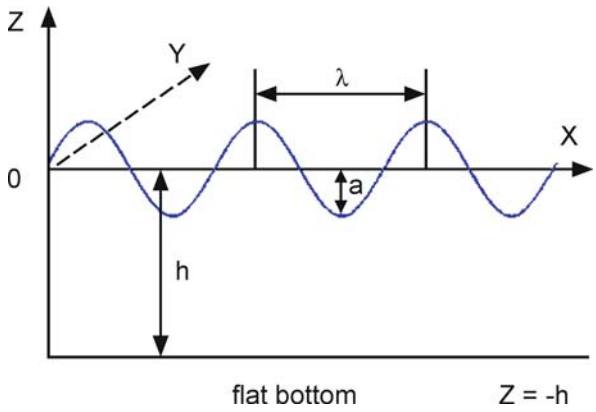

Shallow Water Waves and Solitary Waves, Figure 4
Coordinate frame and periodic wave on the surface of water

depth of the water h is measured from the (flat) bottom of the water up to the quiescent free surface. The fourth length scale is along the Y-axis which is along the crest of the wave and perpendicular to the (X, Z)-plane.

Assuming wave propagation in water of uniform (shallow) depth, i. e. h is constant, and ignoring dissipation, the model equations discussed in this section have a set of common features and limitations which make them mathematically tractable [68]. They describe (i) long waves (or shallow water), i. e. $h \ll \lambda$, (ii) with relatively small amplitude, i. e. $a \ll h$, (iii) traveling in one direction (along the X-axis) or weakly two-dimensional (with a small component in the Y-direction). Furthermore, the small effects must be comparable in size. For example, in the derivation of the KdV and Boussinesq equations one assumes that $\varepsilon = a/h = O(h^2/\lambda^2)$, where ε is a small parameter ($\varepsilon \ll 1$), and O indicates the order of magnitude.

The Korteweg–de Vries Equation

The KdV equation was originally derived to describe shallow water waves of long wavelength and small amplitude. In the derivation, Korteweg and de Vries assumed that all motion is uniform in the Y-direction, along the crest of the wave. In that case, the surface elevation (above the equilibrium level h) of the wave, propagating in the X-direction, is a function only of the horizontal position X (along the canal) and of time T, i. e. $Z = \eta(X, T)$.

In terms of the physical parameters, the KdV equation reads

$$\frac{\partial \eta}{\partial T} + \sqrt{gh}\frac{\partial \eta}{\partial X} + \frac{3}{2}\frac{\sqrt{gh}}{h}\eta\frac{\partial \eta}{\partial X} + \frac{1}{2}h^2\sqrt{gh}\left(\frac{1}{3} - \frac{\mathcal{T}}{\rho g h^2}\right)\frac{\partial^3 \eta}{\partial X^3} = 0\,, \quad (1)$$

where h is the uniform water depth, g is the gravitational acceleration (about 9.81m/sec^2 at sea level), ρ is the density, and $\mathcal{T}$ stands for the surface tension. The dimensionless parameter $\mathcal{T}/\rho g h^2$ is called the *Bond number* which measures the relative strength of surface tension and the gravitational force.

Keeping only the first two terms in (1), the speed of the associated linear (long) wave is $c = \sqrt{gh}$. This is indeed the maximum attainable speed of propagation of gravity-induced water waves of infinitesimal amplitude. The speed of propagation of the small-amplitude solitary waves described by (1) is slightly higher. According to Russell's empirical formula the speed equals $\sqrt{g(h+k)}$, where k is the height of the peak of the solitary wave above the surface of undisturbed water. As Bullough [15] has shown, Russell's approximate speed and the true speed of solitary waves only differ by a term of $O(k^2/h^2)$.

The KdV equation can be recast in dimensionless variables as

$$u_t + \alpha u u_x + u_{xxx} = 0\,, \quad (2)$$

where subscripts denote partial derivatives. The parameter α can be scaled to any real number. Commonly used values are $\alpha = \pm 1$ or $\alpha = \pm 6$.

The term u_t describes the time evolution of the wave propagating in one direction. Therefore, (2) is called an *evolution* equation. The nonlinear term $\alpha u u_x$ accounts for steepening of the wave, and the linear dispersive term u_{xxx} describes spreading of the wave. The linear first-order term $\sqrt{gh}\frac{\partial \eta}{\partial X}$ in (1) can be removed by an elementary transformation. Conversely, a linear term in u_x can be added to (2).

The nonlinear steepening of the water wave can be balanced by dispersion. If so, the result of these counteracting effects is a stable solitary wave with particle-like properties. A solitary wave has a finite amplitude and propagates at constant speed and without change in shape over a fairly long distance. This is in contrast to the concentric group of small-amplitude capillary waves, shown in Fig. 1, which disperse as they propagate.

The closed-form expression of a solitary wave solution is given by

$$u(x,t) = \frac{\omega - 4k^3}{\alpha k} + \frac{12k^2}{\alpha}\operatorname{sech}^2(kx - \omega t + \delta) \quad (3)$$

$$= \frac{\omega + 8k^3}{\alpha k} - \frac{12k^2}{\alpha}\tanh^2(kx - \omega t + \delta)\,, \quad (4)$$

where the wave number k, the angular frequency ω, and δ are arbitrary constants.

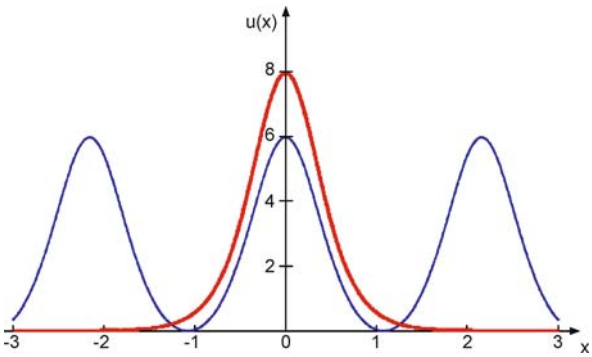

Shallow Water Waves and Solitary Waves, Figure 5
Solitary wave (red) and periodic cnoidal (blue) wave profiles

Requiring that $\lim_{x\to\pm\infty} u(x,t) = 0$ for all time leads to $\omega = 4k^3$. Then (3) and (4) reduce to

$$u(x,t) = \frac{12k^2}{\alpha}\operatorname{sech}^2(kx - 4k^3 t + \delta) = \frac{12k^2}{\alpha}[1 - \tanh^2(kx - 4k^3 t + \delta)] \,. \tag{5}$$

The position of the hump-type wave at $t = 0$ is depicted in Fig. 5 for $\alpha = 6,\ k = 2$, and $\delta = 0$. As time changes, the solitary wave with amplitude $2k^2 = 8$ travels to the right at speed $v = \omega/k = 4k^2 = 16$. The speed is exactly twice the peak amplitude. So, the higher the wave the faster it travels, but it does so without change in shape. The reciprocal of the wavenumber k is a measure of the width of the sech-square pulse.

As shown by Korteweg and de Vries [47], Eq. (2) also has a simple periodic solution,

$$u(x,t) = \frac{\omega - 4k^3(2m-1)}{\alpha k} + \frac{12k^2 m}{\alpha}\operatorname{cn}^2(kx - \omega t + \delta; m), \tag{6}$$

which they called the *cnoidal wave* solution since they involve the Jacobi elliptic cosine function, cn, with modulus m $(0 < m < 1)$. The wavenumber k gives the characteristic width of each oscillation in the "cnoid."

Three cycles of the cnoidal wave are depicted in Fig. 5 at $t = 0$. The graph corresponds to $\alpha = 6, k = 2, m = 3/4, \omega = 16$, and $\delta = 0$. Using the property $\lim_{m\to 1}\operatorname{cn}(\xi; m) = \operatorname{sech}(\xi)$, one readily verifies that (6) reduces to (3) as m tends to 1. Pictorially, the individual oscillations then stretch infinitely far apart leaving a single-pulse solitary wave.

The celebrated KdV equation appears in all books and reviews about soliton theory. In addition, the equation has been featured in, e. g., Miura [52] and Miles [51].

Regularized Long-Wave Equations

A couple of alternatives to the KdV equation have been proposed. A first alternative,

$$u_t + u_x + \alpha u u_x - u_{xxt} = 0 \,, \tag{7}$$

was proposed by Benjamin, Bona, and Mahony [7]. Hence, (7) is referred to as the BBM or regularized long-wave (RLW) equation.

Equation (7), which has a solitary wave solution,

$$u(x,t) = \frac{\omega - k - 4k^2\omega}{\alpha k} + \frac{12k\omega}{\alpha}\operatorname{sech}^2(kx - \omega t + \delta), \tag{8}$$

was also derived by Peregrine [57] to describe the behavior of an undular bore (in water), which comprises a smooth wavefront followed by a train of solitary waves. An undular bore can be interpreted as the dispersive analog of a shock wave in classical non-dispersive, dissipative hydrodynamics [26].

The linear dispersion relation for the KdV equation, $\omega = k(1 - k^2)$, can be obtained by substituting $u(x,t) = \exp[i(kx - \omega t + \delta)]$ into $u_t + u_x + u_{xxx} = 0$. The linear phase velocity, $v_p = \omega/k = 1 - k^2$, becomes negative for $|k| > 1$, thereby contradicting the assumption of unidirectional propagation. Furthermore, the group velocity $v_g = d\omega/dk = 1 - 3k^2$ has no lower bound which implies that there is no limit to the rate at which shorter ripples propagate in the negative x-direction.

The BBM equation, where $\omega = k/(1 + k^2), v_p = 1/(1 + k^2)$, and $v_g = (1 - k^2)/(1 + k^2)^2$, was proposed to get around these objections and to address issues related to proving the existence of solutions of the KdV equation. The dispersion relation of (7) has more desirable properties for high wave numbers, but the group velocity becomes negative for $|k| > 1$. In addition, the KdV and BBM equations are first order in time making it impossible to specify both u and u_t as initial data.

To circumvent these limitations, a second alternative,

$$u_t + u_x + \alpha u u_x + u_{xtt} = 0 \,, \tag{9}$$

was proposed by Joseph and Egri [45] and Jeffrey [43]. It is called the time regularized long-wave (TRLW) equation and its solitary wave solution is given by

$$u(x,t) = \frac{\omega - k - 4k\omega^2}{\alpha k} + \frac{12\omega^2}{\alpha}\operatorname{sech}^2(kx - \omega t + \delta). \tag{10}$$

The TRLW equation shares many of the properties of both the KdV and BBM equations, at the cost of a more complicated dispersion relation, $\omega = (-1 \pm \sqrt{1 + 4k^2})/2k$ with two branches. Only one of these branches is valid because

the derivation of the TRLW equation shows that (9) is uni-directional, despite the two time derivatives in u_{xtt}.

Bona and Chen [8] have shown that the initial value problem for the TRLW equation is well-posed, and that for small-amplitude, long waves, solutions of (9) agree with solutions of either (2) or (7). As a matter of fact, all three equations agree to the neglected order of approximation over a long time scale, provided the initial data is properly imposed (see also [9]).

Fine-tuning the dispersion relation of the KdV equation comes at a cost. In contrast to (2), the RLW and TRLW equations are no longer completely integrable. Perhaps that is why these equations never became as popular as the KdV equation.

The Boussinesq Equation

The classical Boussinesq equation,

$$\eta_{TT} - c^2\eta_{XX} - \frac{3c^2}{h}\left(\eta_X^2 + \eta\eta_{XX}\right) - \frac{c^2h^2}{3}\eta_{XXXX} = 0, \quad (11)$$

was derived by Boussinesq [12] to describe gravity-induced surface waves as they propagate at constant (linear) speed $c = \sqrt{gh}$ in a canal of uniform depth h.

In contrast to the KdV equation, (11) has a second-order time-derivative term. Ignoring all but the first two terms in (11), one obtains the linear *wave* equation, $\eta_{TT} - c^2\eta_{XX} = 0$, which describes both left-running and right-running waves. However, (11) is *not* bi-directional because in the derivation Boussinesq used the constraint $\eta_T = -c\eta_X$, which limits (11) to waves traveling to the right.

In dimensionless form, the Boussinesq equation reads

$$u_{tt} - c^2u_{xx} - \alpha u_x^2 - \alpha u u_{xx} - \beta u_{xxxx} = 0\,. \quad (12)$$

The values of the parameters $c, \alpha > 0$, and β do not matter, but the sign of β matters. Typically, one sets $c = 1, \alpha = 3$, and $\beta = \pm 1$.

A simple solitary wave solution of (12) is given by

$$u(x,t) = \frac{\omega^2 - c^2k^2 - 4\beta k^4}{\alpha k^2} + \frac{12\beta k^2}{\alpha}\operatorname{sech}^2(kx - \omega t + \delta)\,. \quad (13)$$

The equation with $\beta = 1$ is a scaled version of (11) and thus most relevant to shallow water wave theory. Mathematically, (12) with $\beta = 1$ is ill-posed, (even without the nonlinear terms), which means that the initial value problem cannot be solved for arbitrary data. This shortcoming does not happen for (12) with $\beta = -1$, which is therefore nicknamed the "good" Boussinesq equation [50]. Nonetheless, the classical and good Boussinesq equations are completely integrable.

The "improved" or "regularized" Boussinesq equation (see, e. g., [10]) has $\beta = 1$ but u_{xxtt} instead of u_{xxxx}, which improves the properties of the dispersion relation. Like (12), the regularized version describes uni-directional waves. The regularized Boussinesq equation and other alternative equations listed in the literature (see, e. g., Madsen and Schäffer [49]) are not completely integrable.

Bona et al. [10,11] analyzed a family of Boussinesq systems of the form

$$\begin{aligned} w_t + u_x + (uw)_x + \alpha u_{xxx} - \beta w_{xxt} &= 0\,, \\ u_t + w_x + uu_x + \gamma w_{xxx} - \delta u_{xxt} &= 0\,, \end{aligned} \quad (14)$$

which follow from the Euler equations as first-order approximations in the parameters $\varepsilon_1 = a/h \ll 1$, $\varepsilon_2 = h^2/\lambda^2 \ll 1$, where the Stokes number, $S = \varepsilon_1/\varepsilon_2 = a\lambda^2/h^3 \approx 1$.

In (14) $w(x,t)$ is the non-dimensional deviation of the water surface from its undisturbed position; $u(x,t)$ is the non-dimensional horizontal velocity field at a height θh (with $0 \le \theta \le 1$) above the flat bottom of the water. The constant parameters α through δ in (14) satisfy the following consistency conditions: $\alpha + \beta = \frac{1}{2}(\theta^2 - \frac{1}{3})$ and $\gamma + \delta = \frac{1}{2}(1 - \theta^2) \ge 0$. Solitary wave solutions of various special cases of (14) have been computed by Chen [18].

Boussinesq systems arise when modeling the propagation of long-crested waves on large bodies of water (such as large lakes or the ocean). The Boussinesq family (14) encompasses many systems that appeared in the literature. Special cases and properties of well-posedness of (14) are addressed by Bona et al. [10,11].

1D Shallow Water Wave Equation

The so-called one-dimensional (1D) shallow water wave equation,

$$v_{xxt} + \alpha v v_t - v_t - v_x + \beta v_x \int_{\infty}^{x} v_t(y,t)\,dy = 0\,, \quad (15)$$

can be derived from the classical shallow water wave theory (see Sect. "Shallow Water Wave Equations of Geophysical Fluid Dynamics") in the Boussinesq approximation. In that approximation one assumes that vertical variations of the static density, ρ_0, are neglected, except the buoyancy term proportional to $d\rho_0/dz$, which is, in fact, responsible for the existence of solitary waves. The integral term in (15) can be removed by introducing the potential u. Indeed, setting $v = u_x$, Eq. (15) can be written as

$$u_{xxxt} + \alpha u_x u_{xt} - u_{xt} - u_{xx} + \beta u_{xx} u_t = 0\,. \quad (16)$$

The equation is completely integrable and can be solved with the IST if and only if either $\alpha = \beta$ [41] or $\alpha = 2\beta$ [3]. When $\alpha = \beta$, Eq. (16) can be integrated with respect to x and thus replaced by

$$u_{xxt} + \alpha u_x u_t - u_t - u_x = 0 \,. \tag{17}$$

Closed-form solutions of (15), and in particular of (17), have been computed by Clarkson and Mansfield [19].

The Camassa–Holm Equation

The CH equation, named after Camassa and Holm [16,17],

$$u_t + 2\kappa u_x + 3uu_x - \alpha^2 u_{xxt} + \gamma u_{xxx} - 2\alpha^2 u_x u_{xx} - \alpha^2 u u_{xxx} = 0 \,, \tag{18}$$

also models waves in shallow water. In (18), u is the fluid velocity in the x-direction or, equivalently, the height of the water's free surface above a flat bottom, and κ, γ and α are constants. Retaining only the first four terms in (18) gives the BBM Eq. (7). Setting $\alpha = 0$ reduces (18) to the KdV equation.

The CH equation admits solitary wave solutions, but in contrast to the hump-type solutions of the KdV and Boussinesq equations, they are implicit in nature (see, e. g., [44]). In the limit $\kappa \to 0$, Eq. (18) with $\gamma = 0, \alpha = 1$ has a cusp-type solution of the form $u(x,t) = c\exp(-|x - ct - x_0|)$. The solution is called a peakon since it has a peak (or corner) where the first derivatives are discontinuous. The solution travels at speed $c > 0$ which equals the height of the peakon.

The Kadomtsev–Petviashvili Equation

In their 1970 study [46] of the stability of line solitons, Kadomtsev and Petviashvili (KP) derived a 2D-generalization of the KdV equation which now bears their name. In dimensionless variables, the KP equation is

$$(u_t + \alpha u u_x + u_{xxx})_x + \sigma^2 u_{yy} = 0 \,, \tag{19}$$

where y is the transverse direction. In the derivation of the KP equation, one assumes that the scale of variation in the y-direction (along the crest of the wave as shown in Fig. 4) is much longer than the wavelength along the x-direction.

The solitary wave and periodic (cnoidal) solutions of (19) are, respectively, given by

$$u(x,t) = \frac{k\omega - 4k^4 + \sigma^2 l^2}{\alpha k^2} + \frac{12k^2}{\alpha}\operatorname{sech}^2(kx + ly - \omega t + \delta), \tag{20}$$

Shallow Water Waves and Solitary Waves, Figure 6
Periodic plane waves in shallow water, off the coast of Lima, Peru. Photograph courtesy of A. Segur

and

$$u(x,t) = \frac{k\omega - 4k^4(2m-1) - \sigma^2 l^2}{\alpha k^2} + \frac{12k^2 m}{\alpha}\operatorname{cn}^2(kx + ly - \omega t + \delta; m) \,. \tag{21}$$

As shown in Fig. 6, near a flat beach the periodic waves appear as long, quasilinear stripes with a cn-squared cross section. Such waves are typically generated by wind and tides.

The equation with $\sigma^2 = -1$ is referred to as KP1, whereas (19) with $\sigma^2 = 1$ is called KP2, which describes shallow water waves [67]. Both KP1 and KP2 are completely integrable equations but their solution structures are fundamentally different (see, e. g., [65], pp. 489–490).

Shallow Water Wave Equations of Geophysical Fluid Dynamics

The shallow water equations used in geophysical fluid dynamics are based on the assumption $D/L \ll 1$, where D and L are characteristic values for the vertical and horizontal length scales of motion. For example, D could be the average depth of a layer of fluid (or the entire fluid) and L could be the wavelength of the wave.

The geophysical fluid dynamics community (see, e. g., [55,72,73]) uses the following 2D shallow water equations,

$$u_t + uu_x + vu_y + gh_x - 2\Omega v = -gb_x \,, \tag{22}$$

$$v_t + uv_x + vv_y + gh_y + 2\Omega u = -gb_y, \tag{23}$$

$$h_t + (hu)_x + (hv)_y = 0, \tag{24}$$

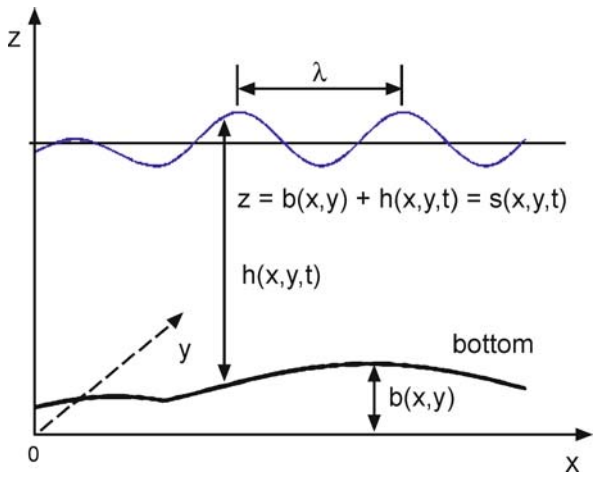

Shallow Water Waves and Solitary Waves, Figure 7
Setup for the geophysical shallow water wave model

to describe water flows with a free surface under the influence of gravity (with gravitational acceleration g) and the Coriolis force due to the earth's rotation (with angular velocity Ω.) As usual, $\mathbf{u} = (u, v)$ denotes the horizontal velocity of the fluid and $h(x, y, t)$ is its depth. As shown in Fig. 7, $h(x, y, t)$ is the distance between the free surface $z = s(x, y, t)$ and the variable bottom $b(x, y)$. Hence, $s(x, y, t) = b(x, y) + h(x, y, t)$. Eqs. (22) and (23) express the horizontal momentum-balance; (24) expresses conservation of mass. Note that the vertical component of the fluid velocity has been eliminated from the dynamics and that the number of independent variables has been reduced by one. Indeed, z no longer explicitly appears in (22)–(24), where u, v, and h only depend on x, y, and t.

A shortcoming of the model is that it does not take into account the density stratification which is present in the atmosphere (as well as in the ocean). Nonetheless, (22)–(24) are commonly used by atmospheric scientists to model flow of air at low speed.

More sophisticated models treat the ocean or atmosphere as a stack of layers with variable thickness. Within each layer, the density is either assumed to be uniform or may vary horizontally due to temperature gradients. For example, Lavoie's rotating shallow water wave equations (see [24]),

$$\boldsymbol{u}_t + (\boldsymbol{u} \cdot \nabla)\boldsymbol{u} + 2\boldsymbol{\Omega} \times \boldsymbol{u} = -\nabla(h\theta) + \tfrac{1}{2}h\nabla\theta \,, \qquad (25)$$

$$h_t + \nabla \cdot (h\boldsymbol{u}) = 0 \,, \qquad (26)$$

$$\theta_t + \boldsymbol{u} \cdot (\nabla\theta) = 0 \,, \qquad (27)$$

consider only one active layer with layer depth $h(x, y, t)$, but take into account the forcing due to a horizontally varying potential temperature field $\theta(x, y, t)$. Vector $\boldsymbol{u} = u(x, y, t)\boldsymbol{i} + v(x, y, t)\boldsymbol{j}$ denotes the fluid velocity and $\boldsymbol{\Omega} = \Omega\boldsymbol{k}$ is the angular velocity vector of the Earth's rotation. $\nabla = \frac{\partial}{\partial x}\boldsymbol{i} + \frac{\partial}{\partial y}\boldsymbol{j}$ is the gradient operator, and $\boldsymbol{i}$, $\boldsymbol{j}$, and $\boldsymbol{k}$ are unit vectors along the x, y, and z-axes.

Lavoie's equations are part of a family of multi-layer models proposed by Ripa [61] to study, for example, the effects of solar heating, fresh water fluxes, and wind stresses on the upper equatorial ocean. A study of the validity of various layered models has been presented by Ripa [62]. The more sophisticated the models become the harder they become to treat with analytic methods so one has to apply numerical methods. Numerical aspects of various shallow water models in atmospheric research and beyond are discussed in e. g. [48,75,76].

Computation of Solitary Wave Solutions

As shown in Sect. "Completely Integrable Shallow Water Wave Equations", solitary wave solutions of the KdV and Boussinesq equations (and like PDEs), can be expressed as polynomials of the hyperbolic secant (sech) or tangent (tanh) functions, whereas their simplest period solutions involve the Jacobi elliptic cosine (cn) function.

There are several methods to compute exact, analytic expressions for solitary and periodic wave solutions of nonlinear PDEs. Two straightforward methods, namely the direct integration method and the tanh-method, will be discussed. Both methods seek traveling wave solutions. By working in a traveling frame of reference the PDE is replaced by an ordinary differential equation (ODE) for which one seeks closed-form solutions in terms of special functions.

In the terminology of dynamical systems, the solitary wave solutions correspond to heteroclinic or homoclinic trajectories in the phase plane of a first-order dynamical system corresponding to the underlying ODE (see, e. g., [6]). The periodic solutions are bounded in the phase plane by these special trajectories, which correspond to the limit of infinite period and modulus one.

Other more powerful methods, such as the aforementioned IST and Hirota's method (see, e. g., [40]) deal with the PDE directly. These methods allow one to compute closed-form expressions of soliton solutions (in particular, solitary wave solutions) addressed elsewhere in the encyclopedia.

Direct Integration Method

Exact expressions for solitary wave solutions can be obtained by direct integration. The steps are illustrated for

the KdV equation given in (2). Assuming that the wave travels to the right at speed $v = \omega/k$, Eq. (2) can be put into a traveling frame of reference with independent variable $\xi = k(x - vt - x_0)$. This reduces (2) to an ODE, $-v\phi' + \alpha\phi\phi' + k^2\phi''' = 0$, for $\phi(\xi) = u(x, t)$. A first integration with respect to ξ yields

$$-v\phi + \frac{\alpha}{2}\phi^2 + k^2\phi'' = A\,, \tag{28}$$

where A is a constant of integration. Multiplication of (28) by ϕ', followed by a second integration with respect to ξ, yields

$$-\frac{v}{2}\phi^2 + \frac{\alpha}{6}\phi^3 + \frac{k^2}{2}\phi'^2 = A\phi + B\,, \tag{29}$$

where B is an integration constant. Separation of variables and integration then leads to

$$\int_{\phi_0}^{\phi} \frac{d\phi}{\sqrt{a\phi^2 - b\phi^3 + \tilde{A}\phi + \tilde{B}}} = \pm \int_{\xi_0}^{\xi} d\xi\,, \tag{30}$$

where $a = v/k^2$, $b = \alpha/3k^2$, $\tilde{A} = 2A/k^2$, and $\tilde{B} = 2B/k^2$.

The evaluation of the elliptic integral in (30) depends on the relationship between the roots of the function $f(\phi) = a\phi^2 - b\phi^3 + \tilde{A}\phi + \tilde{B}$. In turn, the nature of the roots depends on the choice of $\tilde{A}$ and $\tilde{B}$. Two cases lead to physically relevant solutions.

Case 1: If the three roots are real and distinct, then the integral can be expressed in terms of the inverse of the cn function (see, e. g., [25] for details). This leads to the cnoidal wave solution given in (6).

Case 2: If the three roots are real and (only) two of them coincide, then the tanh-squared solution follows. This happens when $\tilde{A} = \tilde{B} = 0$. Integrating both sides of (30) then gives

$$\begin{aligned}\int_{\phi_0}^{\phi} \frac{d\phi}{\phi\sqrt{a - b\phi}} &= -\frac{2}{\sqrt{a}}\,\mathrm{Arctanh}\left[\frac{\sqrt{a - b\phi}}{\sqrt{a}}\right] + C \\ &= \pm(\xi - \xi_0)\,.\end{aligned} \tag{31}$$

where, without loss of generality, C and ξ_0 can be set to zero. Solving (31) for ϕ yields

$$\phi(\xi) = \frac{a}{b}\left(1 - \tanh^2\left(\frac{\sqrt{a}}{2}\xi\right)\right) = \frac{a}{b}\mathrm{sech}^2\left(\frac{\sqrt{a}}{2}\xi\right). \tag{32}$$

Returning to the original variables, one gets

$$\phi(\xi) = \frac{3v}{\alpha}\mathrm{sech}^2\left(\frac{\sqrt{v}}{2k}\xi\right)\,, \tag{33}$$

or

$$u(x, t) = \frac{3v}{\alpha}\mathrm{sech}^2\left(\frac{\sqrt{v}}{2}(x - vt - x_0)\right)\,, \tag{34}$$

where v is arbitrary. Setting $v = \omega/k = 4k^2$, where k is arbitrary, and $\delta = -kx_0$, one can verify that (34) matches (5).

The Tanh Method

If one is only interested in tanh- or sech-type solutions, one can circumvent explicit integration (often involving elliptic integrals) and apply the so-called tanh-method. A detailed description of the method has been given by Baldwin et al. [5]. The method has been fully implemented in *Mathematica*, a popular symbolic manipulation program, and successfully applied to many nonlinear differential equations from soliton theory and beyond.

The tanh-method is based on the following observation: all derivatives of the tanh function can be expressed as polynomials in tanh. Indeed, using the identity $\cosh^2\xi - \sinh^2\xi = 1$ one computes $\tanh'\xi = \mathrm{sech}^2\xi = 1 - \tanh^2\xi$, $\tanh''\xi = -2\tanh\xi + 2\tanh^3\xi$, etc. Therefore, all derivatives of $T(\xi) = \tanh\xi$ are polynomials in T. For example, $T' = 1 - T^2$, $T'' = -2T + 2T^3$, and $T''' = -2T + 8T^2 - 6T^4$.

By applying the chain rule, the PDE in $u(x, t)$ is then transformed into an ODE for $U(T)$ where $T = \tanh\xi = \tanh(kx - \omega t + \delta)$ is the new independent variable. Since all derivatives of T are polynomials of T, the resulting ODE has polynomial coefficients in T. It is therefore natural to seek a polynomial solution of the ODE. The problem thus becomes algebraic. Indeed, after computing the degree of the polynomial solution, one finds its unknown coefficients by solving a nonlinear algebraic system.

The method is illustrated using (2). Applying the chain rule (repeatedly), the terms of (2) become $u_t = -\omega(1 - T^2)U'$, $u_x = k(1 - T^2)U'$, and

$$\begin{aligned}u_{xxx} = k^3(1 - T^2)\big[&-2(1 - 3T^2)U' \\ &- 6T(1 - T^2)U'' + (1 - T^2)^2U'''\big]\,,\end{aligned} \tag{35}$$

where $U(T) = U(\tanh(kx - \omega t + \delta)) = u(x, t)$, $U' = dU/dT$, etc.

Substitution into (2) and cancellation of a common $1 - T^2$ factor yields

$$\begin{aligned}&-\omega U' + \alpha kUU' - 2k^3(1 - 3T^2)U' \\ &\quad - 6k^3T(1 - T^2)U'' + k^3(1 - T^2)^2U''' = 0\,.\end{aligned} \tag{36}$$

This ODE for $U(T)$ has polynomial coefficients in T. One therefore seeks a polynomial solution

$$U(T) = \sum_{n=0}^{N} a_n T^n \,, \tag{37}$$

where the integer exponent N and the coefficients a_n must be computed.

Substituting T^N into (36) and balancing the highest powers in T gives $N = 2$. Then, substituting

$$U(T) = a_0 + a_1 T + a_2 T^2 \tag{38}$$

into (36), and equating to zero the coefficients of the various power terms in T, yields

$$\begin{aligned} a_1(\alpha a_2 + 2k^2) &= 0 \,, \\ \alpha a_2 + 12k^2 &= 0 \,, \\ a_1(\alpha k a_0 - 2k^3 - \omega) &= 0 \,, \\ \alpha k a_1^2 + 2\alpha k a_0 a_2 - 16k^3 a_2 - 2\omega a_2 &= 0 \,. \end{aligned} \tag{39}$$

The unique solution of this nonlinear system for the unknowns a_0, a_1 and a_2 is

$$a_0 = \frac{8k^3 + \omega}{\alpha k}, \;\; a_1 = 0, \;\; a_2 = -\frac{12k^2}{\alpha} \,. \tag{40}$$

Finally, substituting (40) into (38) and using $T = \tanh(kx - \omega t + \delta)$ yields (4).

The solitary wave solutions and cnoidal wave solutions presented in Sect. "Completely Integrable Shallow Water Wave Equations" have been automatically computed with a *Mathematica* package [5] that implements the tanh-method and variants.

A review of numerical methods to compute solitary waves of arbitrary amplitude can be found in Vanden-Broeck [74].

Water Wave Experiments and Observations

Through a series of experiments in a hydrodynamic tank, Hammack investigated the validity of the BBM equation [35] and KdV equation as models for long waves in shallow water [36,37,38] and long internal waves [69]. Their research addressed the question: Would an initial displacement of water, as it propagates forward, eventually evolve in a train of localized solitary waves (solitons) and an oscillatory tail as predicted by the KdV equation? Based on the experimental data, they concluded that (i) the KdV dynamics only occurs if the waves travel over a long distance, (ii) a substantial amount of water must be initially displaced (by a piston) to produce a soliton train, (iii) the water volume of the initial wave determines the shape of the leading wave in the wave train, and (iv) the initial direction of displacement (upward or downward piston motion) determines what happens later. Quickly raising the piston causes a train of solitons to emerge; quickly lowering it causes all wave energy to distribute into the oscillatory tail, as predicted by the theory.

Several other researchers have tested the validity of the KdV equation and variants in laboratory experiments (see, e. g., [39,60]). Bona et al. [9] give an in-depth evaluation of the BBM Eq. (7) with and without dissipative term u_{xx}. Their study includes (i) a numerical scheme with error estimates, (ii) a convergence test of the computer code, (iii) a comparison between the predictions of the theoretical model and the results of laboratory experiments. The authors note that it is important to include dissipative effects to obtain reasonable agreement between the forecast of the model and the empirical results.

Water tank experiments in conjunction with the analysis of actual data, allows researchers to judge whether or not the KdV equation can be used to model the dynamics of tsunamis (see [67]). Tsunami research intensified after the December 2004 tsunami devastated large coastal regions of India, Indonesia, Sri Lanka, and Thailand, and killed nearly 300,000 people.

Apart from shallow water waves near beaches, the KdV equation and its solitary wave solution also apply to internal waves in the ocean. Internal solitary waves in the open ocean are slow waves of large amplitude that travel at the interface of stratified layers of different density. Stratification based on density differences is primarily due to variations in temperature or concentration (e. g. due to salinity gradients). For example, absorption of solar radiation creates a near surface thin layer of warmer water (of lower density) above a thicker layer or colder, denser water. The smaller the density change, the lower the wave frequency, and the slower the propagation speed. If the upper layer is thinner than the lower one, then the internal wave is a wave of depression causing a downward displacement of the fluid interface.

Internal solitary waves are ubiquitous in stratified waters, in particular, whenever strong tidal currents occur near irregular topography. Such waves have been studied since the 1960s. An early, well-documented case deals with internal waves in the Andaman Sea, where Perry and Schimke [58] found groups of internal waves up to 80 m high and 2000 m long on the main thermocline at 500 m in 1500 m deep water. Their measurements were confirmed by Osborne and Burch [53] who showed that internal waves in the Andaman Sea are generated by tidal flows and can travel over hundreds of kilometers.

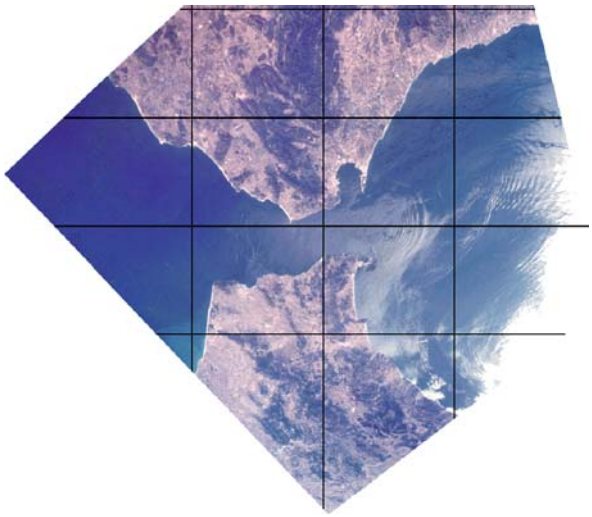

Shallow Water Waves and Solitary Waves, Figure 8
Three solitary wave packets generated by internal waves from sills in the Strait of Gibraltar. Original image STS41G-34-81 courtesy of the Earth Sciences and Image Analysis Laboratory, NASA Johnson Space Center (http://eol.jsc.nasa.gov). Ortho-rectified, color adjusted photograph courtesy of Global Ocean Associates

Strong internal waves can affect biological life and interfere with underwater navigation. Understanding the behavior of internal waves can aid in the design of offshore production facilities for oil and natural gas.

The near-surface current associated with the internal wave locally modulates the height of the water surface. Hence, the internal wave leaves a "signature" or "footprint" at the sea surface in the form of a packet of solitary waves (sometimes called current rips or tide rips). These visual manifestations appear as long, quasilinear stripes in satellite imagery or photographs taken during space flights. Over 50 case studies and hundreds of images of oceanic internal waves can be found in "An Atlas of Internal Solitary-like Waves and Their Properties" [42].

Figure 8 shows a photograph of three solitary waves packets which are the surface signature of internal waves in the Strait of Gibraltar. The photograph was taken from the Space Shuttle on October 11, 1984. Spain is to the North, Morocco to the South. Alternate solitary wave packets move toward the northeast or the southeast. The amplitude of these waves is of the order of 50 m; their wavelength is in the range of 500–2000 m. The separation between the packets is approximately 30 km. Waves of longer wavelengths and higher amplitudes have traveled the furthest. The number of oscillations within each packet increases as time goes on. Solitary wave packets can reach 200 km into the Western Mediterranean sea and live for more than two days before dissipating. A in-depth study of solitary waves in the Strait of Gibraltar can be found in Farmer and Armi [28].

The KdV model is applicable to stratified fluids with two layers and internal solitary waves if (i) the ratio of the amplitude a to the upper layer depth h is small, and (ii) the wavelength λ is long compared with the upper layer depth. More precisely, $a/h = O(h^2/\lambda^2) \ll 1$. A detailed discussion of internal solitary waves and additional references can be found in Garrett and Munk [31], Grimshaw [32,33,34], Helfrich and Melville [39], Apel et al. [4], and Pelinovsky et al. [56]. The last three papers discuss a variety of other theoretical models including the extended KdV equation (also known as Gardner's equation or combined KdV-modified KdV equation) which contains both quadratic and cubic nonlinearities. Solitary wave solutions of the extended KdV equation can be found in Scott ([65], p. 856), Helfrich and Melville [39], and Apel et al. [4]. A review of laboratory experiments with internal solitary waves was published by Ostrovsky and Stepanyants [54].

As discussed in the review paper by Staquet and Sommeria [70], internal gravity waves also occur in the atmosphere, where they are often caused by wind blowing over topography and cumulus convective clouds. Internal gravity waves reveal themselves as unusual cloud patterns, which are the counterpart of the solitary wave packets on the ocean's surface.

Future Directions

For many shallow water wave applications, the full Euler equations are too complex to work with. Instead, various approximate models have been proposed. Arguably, the most famous shallow water wave equations are the KdV and Boussinesq equations.

The KdV equation was originally derived to describe shallow water waves in a rectangular channel. Surprisingly, the equation also models ion-acoustic waves and magnetohydrodynamic waves in plasmas, waves in elastic rods, mid-latitude and equatorial planetary waves, acoustic waves on a crystal lattice, and more (see, e. g., [64,65,66]). The KdV equation has played a pivotal role in the development of the Inverse Scattering Transform and soliton theory, both of which had a lasting impact on twentieth-century mathematical physics.

Historically, the classical Boussinesq equation was derived to describe the propagation of shallow water waves in a canal. Boussinesq systems arise when modeling the propagation of long-crested waves on large bodies of water (e. g. large lakes or the ocean). As Bona et al. [10] point

out, a plethora of formally-equivalent Boussinesq systems can be derived. Yet, such systems may have vastly different mathematical properties. The study of the well-posedness of the nonlinear models is of paramount importance and is the subject of ongoing research.

Shallow water wave theory allows one to adequately model waves in canals, surface waves near beaches, and internal waves in the ocean (see [4]). Due to their widespread occurrence in the ocean (see [42]), solitary waves and "solitary wave packets" (solitons) are of interest to oceanographers and geophysicists. The (periodic) cnoidal wave solutions are used by coastal engineers in studies of sediment movement, erosion of sandy beaches, interaction of waves with piers and other coastal structures.

Apart from their physical relevance, the knowledge of solitary and cnoidal wave solutions of nonlinear PDEs facilitates the testing of numerical solvers and aids in stability analysis.

Shallow water wave models are widely used in atmospheric science as a paradigm for geophysical fluid motions. They model, for example, inertia-gravity waves with fast time scale dynamics, and wave-vortex interactions and Rossby waves associated with slow advective-timescale dynamics.

This article has reviewed commonly used shallow water wave models, with the hope of bridging two research communities: one that focuses on nonlinear equations with dispersive effects; the other on nonlinear hyperbolic equations without dispersive terms. Of common concern are the testing of the theoretical models on measured data and further validation of the equations with numerical simulations and laboratory experiments. A fusion of the expertise of both communities might advance research on water waves and help to answer open questions about wave breaking, instability, vorticity, and turbulence. Of paramount importance is the prevention of natural disasters, ecological ravage, and damage to man-made structures due to a better understanding of the dynamics of tsunamis, steep waves, strong internal waves, rips, tidal currents, and storm surges.

Acknowledgments

This material is based upon work supported by the National Science Foundation (NSF) of the USA under Award No. CCF-0830783. The author is grateful to Harvey Segur, Jerry Bona, Michel Grundland, and Douglas Poole for advise and comments. William Navidi, Benoit Huard, Anna Segur, Katherine Socha, Christopher Jackson, Michel Peyrard, and Christopher Eilbeck are thanked for help with figures and photographs.

Bibliography

Primary Literature

1. Ablowitz MJ, Clarkson PA (1991) Solitons, nonlinear evolution equations and inverse scattering. Cambridge University Press, Cambridge
2. Ablowitz MJ, Segur H (1981) Solitons and the inverse scattering transform. SIAM, Philadelphia, Pennsylvania
3. Ablowitz MJ et al (1974) The inverse scattering transform: Fourier analysis for nonlinear problems. Stud Appl Math 53:249–315
4. Apel JR et al (2007) Internal solitons in the ocean and their effect on underwater sound. J Acoust Soc Am 121:695–722
5. Baldwin D et al (2004) Symbolic computation of exact solutions expressible in hyperbolic and elliptic functions for nonlinear PDEs. J Symb Comp 37:669–705
6. Balmforth NJ (1995) Solitary waves and homoclinic orbits. Annu Rev Fluid Mech 27:335–373
7. Benjamin TB et al (1972) Model equations for long waves in nonlinear dispersive systems. Phil Trans Roy Soc London Ser A 272:47–78
8. Bona JL, Chen H (1999) Comparison of model equations for small-amplitude long waves. Nonl Anal 38:625–647
9. Bona JL et al (1981) An evaluation of a model equation for water waves. Phil Trans Roy Soc London Ser A 302:457–510
10. Bona JL et al (2002) Boussinesq equations and other systems for small-amplitude long waves in nonlinear dispersive media. I: Derivation and linear theory. J Nonl Sci 12:283–318
11. Bona JL et al (2004) Boussinesq equations and other systems for small-amplitude long waves in nonlinear dispersive media. II: The nonlinear theory. Nonlinearity 17:925–952
12. Boussinesq J (1871) Théorie de l'intumescence liquide appelée onde solitaire ou de translation, se propageant dans un canal rectangulaire. C R Acad Sci Paris 72:755–759
13. Boussinesq J (1872) Théorie des ondes et des remous qui se propagent le long d'un canal rectangulaire horizontal, en communiquant au liquide contenu dans ce canal des vitesses sensiblement pareilles de la surface au fond. J Math Pures Appl 17:55–108
14. Boussinesq J (1877) Essai sur la théorie des eaux courantes. Académie des Sciences de l'Institut de France, Mémoires présentés par divers savants (ser 2) 23:1–680
15. Bullough RK (1988) "The wave" "par excellence", the solitary, progressive great wave of equilibrium of the fluid–an early history of the solitary wave. In: Lakshmanan M (ed) Solitons: introduction and applications. Springer, Berlin, pp 7–42
16. Camassa R, Holm D (1993) An integrable shallow water equation with peakon solitons. Phys Rev Lett 71:1661–1664
17. Camassa R, Holm D (1994) An new integrable shallow water equation. Adv Appl Mech 31:1–33
18. Chen M (1998) Exact solutions of various Boussinesq systems. Appl Math Lett 11:45–49
19. Clarkson PA, Mansfield EL (1994) On a shallow water wave equation. Nonlinearity 7:975–1000
20. Craik ADD (2004) The origins of water wave theory. Annu Rev Fluid Mech 36:1–28

21. Craik ADD (2005) George Gabriel Stokes and water wave theory. Annu Rev Fluid Mech 37:23–42
22. Darrigol O (2003) The spirited horse, the engineer, and the mathematician: Water waves in nineteenth-century hydrodynamics. Arch Hist Exact Sci 58:21–95
23. Dauxois T, Peyrard M (2004) Physics of solitons. Cambridge University Press, Cambridge, UK. Transl of Peyrard M, Dauxois T (2004) Physique des solitons. Savoirs Actuels, EDP Sciences, CNRS Editions
24. Dellar P (2003) Common Hamiltonian structure of the shallow water equations with horizontal temperature gradients and magnetic fields. Phys Fluids 15:292–297
25. Drazin PG, Johnson RS (1989) Solitons: an introduction. Cambridge University Press, Cambridge, UK
26. El GA (2007) Korteweg–de Vries equation: solitons and undular bores. In: Grimshaw RHJ (ed) Solitary waves in fluids. WIT Press, Boston, pp 19–53
27. Emmerson GS (1977) John Scott Russell: A great Victorian engineer and naval architect. John Murray Publishers, London
28. Farmer DM, Armi L (1988) The flow of Mediterranean water through the Strait of Gibraltar. Prog Oceanogr 21:1–105
29. Fermi F et al (1955) Studies of nonlinear problems I. Los Alamos Sci Lab Rep LA-1940. Reproduced in: AC Newell (ed) (1974) Nonlinear wave motion. AMS, Providence, RI
30. Filippov AT (2000) The versatile soliton. Birkhäuser-Verlag, Basel
31. Garrett C, Munk W (1979) Internal waves in the ocean. Annu Rev Fluid Mech 11:339–369
32. Grimshaw R (1997) Internal solitary waves. In: Liu PLF (ed) Advances in coastal and ocean engineering, vol III. World Scientific, Singapore, pp 1–30
33. Grimshaw R (2001) Internal solitary waves. In: Grimshaw R (ed) Environmental stratified flows. Kluwer, Boston, pp 1–28
34. Grimshaw R (2005) Korteweg–de Vries equation. In: Grimshaw R (ed) Nonlinear waves in fluids: Recent advances and modern applications. Springer, Berlin, pp 1–28
35. Hammack JL (1973) A note on tsunamis: their generation and propagation in an ocean of uniform depth. J Fluid Mech 60:769–800
36. Hammack JL, Segur H (1974) The Korteweg–de Vries equation and water waves, part 2. Comparison with experiments. J Fluid Mech 65:289–314
37. Hammack JL, Segur H (1978a) The Korteweg–de Vries equation and water waves, part 3. Oscillatory waves. J Fluid Mech 84:337–358
38. Hammack JL, Segur H (1978b) Modelling criteria for long water waves. J Fluid Mech 84:359–373
39. Helfrich KR, Melville WK (2006) Long nonlinear internal waves. Annu Rev Fluid Mech 38:395–425
40. Hirota R (2004) The direct method in soliton theory. Cambridge University Press, Cambridge, UK
41. Hirota R, Satsuma J (1976) N-soliton solutions of model equations for shallow water waves. J Phys Soc Jpn 40:611–612
42. Jackson CR (2004) An atlas of internal solitary-like waves and their properties, 2nd edn. Global Ocean Associates, Alexandria, VA
43. Jeffrey A (1978) Nonlinear wave propagation. Z Ang Math Mech (ZAMM) 58:T38–T56
44. Johnson RS (2003) On solutions of the Camassa–Holm equation. Proc Roy Soc London Ser A 459:1687–1708
45. Joseph RI, Egri R (1977) Another possible model equation for long waves in nonlinear dispersive systems. Phys Lett A 61:429–432
46. Kadomtsev BB, Petviashvili VI (1970) On the stability of solitary waves in weakly dispersive media. Sov Phys Doklady 15:539–541
47. Korteweg DJ, de Vries G (1895) On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves. Philos Mag (Ser 5) 39:422–443
48. LeVeque RJ (2002) Finite volume methods for hyperbolic problems. Cambridge University Press, Cambridge, UK
49. Madsen PA, Schäffer HA (1999) A review of Boussinesq-type equations for surface gravity waves. In: Liu PLF (ed) Advances in coastal and ocean engineering, vol V. World Scientific, Singapore, pp 1–95
50. McKean HP (1981) Boussinesq's equation on the circle. Comm Pure Appl Math 34:599–691
51. Miles JW (1981) The Korteweg–de Vries equation: a historical essay. J Fluid Mech 106:131–147
52. Miura RM (1976) The Korteweg–de Vries equation: a survey of results. SIAM Rev 18:412–559
53. Osborne AR, Burch TL (1980) Internal solitons in the Andaman sea. Science 208:451–460
54. Ostrovsky LA, Stepanyants YA (2005) Internal solitons in laboratory experiments: Comparison with theoretical models. Chaos 15:037111
55. Pedlosky J (1987) Geophysical fluid dynamics, 2nd edn. Springer, Berlin
56. Pelinovsky E et al (2007) In: Grimshaw RHJ (ed) Solitary waves in fluids. WIT Press, Boston, pp 85–110
57. Peregrine DH (1966) Calculations of the development of an undular bore. J Fluid Mech 25:321–330
58. Perry RB, Schimke GR (1965) Large amplitude internal waves observed off the north-west coast of Sumatra. J Geophys Res 70:2319–2324
59. Rayleigh L (1876) On waves. Philos Mag 1:257–279
60. Remoissenet M (1999) Waves called solitons: Concepts and experiments, 3rd edn. Springer, Berlin
61. Ripa P (1993) Conservation laws for primitive equations models with inhomogeneous layers. Geophys Astrophys Fluid Dyn 70:85–111
62. Ripa P (1999) On the validity of layered models of ocean dynamics and thermodynamics with reduced vertical resolution. Dyn Atmos Oceans Fluid Dyn 29:1–40
63. Russell JS (1844) Report on waves, 14th meeting of the British Association for the Advancement of Science. John Murray, London, pp 311–390
64. Scott AC (2003) Nonlinear science: Emergence and dynamics of coherent structures. Oxford University Press, Oxford, UK
65. Scott AC (ed) (2005) Encyclopedia of Nonlinear science. Routledge, New York
66. Scott AC et al (1973) The soliton: a new concept in applied science. Proc IEEE 61:1443–1483
67. Segur H (2007a) Waves in shallow water, with emphasis on the tsunami of 2004. In: Kundu A (ed) Tsunami and nonlinear waves. Springer, Berlin
68. Segur H (2007b) Integrable models of waves in shallow water. In: Pinski M, Birnir B (eds) Probability, geometry and integrable systems. Cambridge University Press, Cambridge, UK
69. Segur H, Hammack JL (1982) Soliton models of long internal waves. J Fluid Mech 118:285–304

70. Staquet C, Sommeria J (2002) Internal gravity waves: From instabilities to turbulence. Annu Rev Fluid Mech 34:559–593
71. Stokes GG (1847) On the theory of oscillatory waves. Trans Camb Phil Soc 8:441–455
72. Toro EF (2001) Shock-capturing methods for free-surface shallow flows. Wiley, New York
73. Vallis GK (2006) Atmospheric and oceanic fluid dynamics: fundamentals and large scale circulation. Cambridge University Press, Cambridge, UK
74. Vanden-Broeck J-M (2007) Solitary waves in water: numerical methods and results. In: Grimshaw RHJ (ed) Solitary waves in fluids. WIT Press, Boston, pp 55–84
75. Vreugdenhil CB (1994) Numerical methods for shallow-Water flow. Springer, Berlin
76. Weiyan T (1992) Shallow water hydrodynamics: mathematical theory and numerical solution for two-dimensional systems of shallow water equations. Elsevier, Amsterdam
77. Zabusky NJ (2005) Fermi–Pasta–Ulam, solitons and the fabric of nonlinear and computational science: History, synergetics, and visiometrics. Chaos 15:015102
78. Zabusky NJ, Kruskal MD (1965) Interaction of 'solitons' in a collisionless plasma and the recurrence of initial states. Phys Rev Lett 15:240–243

Books and Reviews

Boyd JP (1998) Weakly nonlinear solitary waves and beyond-all-order asymtotics. Kluwer, Dordrecht
Calogero F, Degasperis A (1982) Spectral transform and solitons I. North Holland, Amsterdam
Dickey LA (2003) Soliton equations and Hamiltonian systems, 2nd edn. World Scientific, Singapore
Dodd RK et al (1982) Solitons and nonlinear wave equations. Academic Press, London
Eilenberger G (1981) Solitons: Mathematical methods for physicists. Springer, Berlin
Faddeev LD, Takhtajan LA (1987) Hamiltonian methods in the theory of solitons. Springer, Berlin
Fordy A (ed) (1990) Soliton theory: A survey of results. Manchester University Press, Manchester
Grimshaw RHJ (ed) (2007) Solitary waves in fluids. WIT Press, Boston
Infeld E, Rowlands G (2000) Nonlinear waves, solitons, and chaos. Cambridge University Press, New York
Johnson RS (1977) A modern introduction to the mathematical theory of water waves. Cambridge University Press, Cambridge, UK
Lamb GL (1980) Elements of soliton theory. Wiley Interscience, New York
Miles JW (1980) Solitary waves. Annu Rev Fluid Mech 12:11–43
Mei CC et al (2005) Theory and applications of ocean surface waves. World Scientific, Singapore
Mitropol'sky YZ (2001) Dynamics of internal gravity waves in the ocean. Kluwer, Dordrecht
Newell AC (1983) The history of the soliton. J Appl Mech 50:1127–1137
Newell AC (1985) Solitons in mathematics and physics. SIAM, Philadelphia, PA
Makhankov VG (1990) Soliton phenomenology. Kluwer, Dordrecht, The Netherlands
Novikov SP et al (1984) Theory of solitons: The inverse scattering method. Consultants Bureau(Plenum Press), New York
Pedlosky J (2003) Waves in the ocean and atmosphere: introduction to wave dynamics. Springer, Berlin
Russell JS (1885) The wave of translation in the oceans of water, air and ether. Trübner, London
Stoker JJ (1957) Water waves. Wiley Interscience, New York
Whitham GB (1974) Linear and nonlinear waves. Wiley Interscience, New York

Signaling Games

Joel Sobel
Department of Economics, University of California, San Diego, USA

Article Outline

Glossary
Definition of the Subject
Introduction
The Model
Equilibrium
The Basic Model
Cheap Talk
Verifiable Information
Communication About Intentions
Applications
Future Directions
Acknowledgments
Bibliography

Glossary

Babbling equilibrium An equilibrium in which the sender's strategy is independent of type and the receiver's strategy is independent of signal.

Behavior strategy A strategy for an extensive-form game that specifies the probability of taking each action at each information set.

Behavioral type A player in a game who is constrained to follow a given strategy.

Cheap-talk game A signaling game in which players' preferences do not depend directly on signals.

Condition D1 An equilibrium refinement that requires out-of-equilibrium beliefs to be supported on types that have the most to gain from deviating from a fixed equilibrium.

Divinity An equilibrium refinement that requires out-of-equilibrium beliefs to place relatively more weight on types that gain more from deviating from a fixed equilibrium.

Equilibrium outcome The probability distribution over terminal nodes in a game determined by equilibrium strategy.

Handicap principle The idea that animals communicate fitness through observable characteristics that reduce fitness.

Incomplete information game A game in which players lack information about the strategy sets or payoff functions of their opponents.

Intuitive criterion An equilibrium refinement that requires out-of-equilibrium beliefs to place zero weight on types that can never gain from deviating from a fixed equilibrium outcome.

Nash equilibrium A strategy profile in a game in which each player's strategy is a best response to the equilibrium strategies of the other players.

Neologism-proof equilibrium An equilibrium that admits no self-signaling set.

Pooling equilibrium A signaling-game equilibrium in which each all sender types send the same signal with probability one.

Receiver In a signaling game, the uninformed player.

Self-signaling set A set of types C with the property that precisely types in the set C gain from inducing the best response to C relative to a fixed equilibrium.

Sender In a signaling game, the informed agent.

Separating equilibrium A signaling-game equilibrium in which sender types sent signals from disjoint subsets of the set of available signals.

Signaling game A two-player game of incomplete information in which one player is informed and the other in not. The informed player's strategy is a type-contingent message and the uninformed player's strategy is a message-continent action.

Single-crossing condition A condition that guarantees that indifferent curves from a given family of preferences cross at most one.

Spence-Mirrlees condition A differential condition that orders the slopes of level sets of a function.

Standard signaling game A signaling game in which strategy sets and payoff functions satisfy monotonicity properties.

Type In an incomplete information game, a variable that summarizes private information.

Verifiable information game A signaling game with the property that each type has a signal that can only be sent by that type.

Definition of the Subject

Signaling games refer narrowly to a class of two-player games of incomplete information in which one player is informed and the other is not. The informed player's strategy set consists of signals contingent on information and the uninformed player's strategy set consists of actions contingent on signals. More generally, a signaling game includes any strategic setting in which players can use the actions of their opponents to make inferences about hidden information. The earliest work on signaling games was Spence [72]'s model of educational signaling and Zahari [76]'s model of signaling by animals. During the 1980s researchers developed the formal model and identified conditions that permitted the selection of unique equilibrium outcomes in leading models.

Introduction

The framed degree in your doctor's office, the celebrity endorsement of a popular cosmetic, and the telephone message from an old friend are all signals. The signals are potentially valuable because they allow you to infer useful information. These signals are indirect and require interpretation. They may be subject to manipulation. The doctor's diploma tells you something about the doctor's qualifications, but knowing where and when the doctor studied does not prove that she is a good doctor. The endorsement identifies the product with a particular lifestyle, but what works for the celebrity may not work for you. Besides, the celebrity was probably paid to endorse the product and may not even use it. The phone message may tell you how to get in touch with your friend, but is unlikely to contain all of the information you need to find him – or to evaluate whether you'll meet to discuss old times or to be asked a favor. While the examples all involve signaling, the nature of the signaling is different. The doctor faces large penalties for misrepresenting her credentials. She is not required to display all of her diplomas, but it is reasonable to assume that degrees are not forged. The celebrity endorsement is costly – certainly to the manufacturer who pays for the celebrity's services and possibly to the celebrity himself, whose reputation may suffer if the product works badly. It is reasonable to assume that it is easier to obtain an endorsement of a good product, but there are also good reasons to be skeptical about the claims. In contrast, although a dishonest or misleading message may lead to a bad outcome, leaving a message is not expensive and the content of the message is not constrained by your friend's information. The theory of signaling games is a useful way to describe the essential features of all three examples.

Opportunities to send and evaluate signals arise in many common natural and economic settings. In the canonical example (due to Spence [72]), a high-ability worker invests in education to distinguish herself from less skilled workers. The potential employer observers educational attainment, but not innate skill, and infers that

a better educated worker is more highly skilled and pays a higher wage. To make this story work, there must be a reason that low-ability workers do not get the education expected of a more highly skilled worker and hence obtain a higher wage. This property follows from an assumption that the higher the ability the worker, the easier it is for her to produce a higher signal.

The same argument appears in many applications. For example, a risk-averse driver will purchase a lower cost, partial insurance contract, leaving the riskier driver to pay a higher rate for full insurance (Rothschild and Stiglitz [65] or Wilson [75]). A firm that is able to produce high-quality goods signals this by offering a warranty for the goods sold (Grossman [37]) or advertising extensively. A strong deer grows extra large antlers to show that it can survive with this handicap and to signal its fitness to potential mates [76].

Game theory provides a formal language to study how one should send and interpret signals in strategic environments. This article reviews the basic theory of signaling and discusses some applications. It does not discuss related models of screening. Kreps and Sobel [43] and Riley [64] review both signaling and screening.

Section "The Model" describes the basic model. Section "Equilibrium" defines equilibrium for the basic model. Section "The Basic Model" limits attention to a special class of signaling game. I give conditions sufficient for the existence of equilibria in which the informed agent's signal fully reveals her private information and argue that one equilibrium of this kind is prominent. The next three sections study different signaling games. Section "Cheap Talk" discusses models of costless communication. Section "Verifiable Information" discusses the implications of the assumptions that some information is verifiable. Section "Communication about Intentions" briefly discusses the possibility of signaling intentions rather than private information. Section "Applications" describes some applications and extensions of the basic model. Section "Future Directions" speculates on directions for future research.

The Model

This section describes the basic signaling game. There are two players, called S (for sender) and R (for receiver). S knows the value of some random variable t whose support is a given set T. t is called the **type** of S. The prior beliefs of R are given by a probability distribution $\pi(\cdot)$ over T; these beliefs are common knowledge. When T is finite, $\pi(t)$ is the prior probability that the sender's type is t. When T is uncountably infinite, $\pi(\cdot)$ is a density function. Player S learns t and sends to R a signal s drawn from some set M. Player R receives this signal, and then takes an action a drawn from a set A. (It is possible to allow A to depend on s and S to depend on t.) This ends the game: The payoff to i is given by a function $u^i\colon T \times M \times A \to \mathbb{R}$.

This canonical game captures the essential features of the classic applications of market signaling. In the labor-market signaling story due to Spence [72] a worker wishes to signal his ability to a potential employer. The worker has information about ability that the employer lacks. Direct communication about ability is not possible, but the worker can acquire education. The employer can observe the worker's level of education and use this to form a judgment about the worker's true level of ability. In this application, S is a worker; R represents a potential employer (or a competitive labor market); t is the student's productivity; s is her level of education; and a is her wage.

Equilibrium

Defining Nash equilibrium for the basic signaling game is completely straightforward when T, S, and A are finite sets. In this case a behavior strategy for S is a function $\mu\colon T \times M \to [0,1]$ such that $\sum_{s\in M}\mu(t,s) = 1$ for all t. $\mu(t,s)$ is the probability that sender-type t sends the signal s. A behavior strategy for R is a function $\alpha\colon M \times A \to [0,1]$ where $\sum_{a\in A}\alpha(s,a) = 1$ for all s. $\alpha(s,a)$ is the probability that R takes action a following the signal s.

Proposition 1 *Behavior strategies (α^*, μ^*) form a Nash Equilibrium if and only if for all $t \in T$*

$$\mu(t,s) > 0 \ \textit{implies} \ \sum_{a\in A} U^S(t,s,a)\alpha(s,a) = \max_{s'\in S}\sum_{a\in A} U^S(t,s',a)\alpha(s',a) \quad (1)$$

and, for each $s \in S$ such that $\sum_{t\in T}\mu(t,s)\pi(t) > 0$, if $\sum_{t\in T}\mu(t,s)\pi(t) > 0$, then

$$\alpha(s,a) > 0 \ \textit{implies} \ \sum_{t\in T} U^R(t,s,a)\beta(t,a) = \max_{a'\in A}\sum_{t\in T} U^R(t,s,a')\beta(t,a') , \quad (2)$$

where

$$\beta(t,s) = \frac{\mu(t,s)\pi(t)}{\sum_{t'\in T}\mu(t',s)\pi(t')} . \quad (3)$$

Condition (1) states that the S places positive probability only on signals that maximize expected utility. This condition guarantees that S responds optimally to R's strategy.

Condition (2) states that R places positive probability only on actions that maximize expected utility, where is taken with respect to the distribution $\beta(\cdot, s)$ following the signal s. Condition (3) states that $\beta(\cdot, s)$ accurately reflects the pattern of play. It requires that R's beliefs be determined using S's strategy and the prior distribution whenever possible. Equilibrium refinements also require that R has beliefs following signals s that satisfy

$$\sum_{t \in T} \mu(t, s)\pi(t) = 0 \,, \tag{4}$$

that is are sent with probability zero in equilibrium. Specifically, sequential equilibrium permits $\beta(\cdot, m)$ to be an arbitrary distribution when Eq. (4) holds, but requires that Eq. (2) holds even for these values of s. This restriction rules out equilibria in which certain signals are not sent because the receiver responds to the signal with an action that is dominated.

The ability to signal creates the possibility that R will be able to draw inferences about S's type from the signal. Whether he is able to do so is a property of the equilibrium. It is useful to define two extreme cases.

Definition 1 An equilibrium (α^*, μ^*) is called a separating equilibrium if each type t sends different signals. That is, M can be partitioned into sets M_t such that for each t, $\sum_{s \in M_t} \mu(t, s) = 1$. An equilibrium (α^*, μ^*) is called a pooling equilibrium if there is a single signal s^* that is sent by all types with probability one.

In a separating equilibrium, R can infer S's private information completely. In a pooling equilibrium, R learns nothing from the sender's signal. This definition excludes other possible situations. For example, all sender types can randomize uniformly over a set of two or more signals. In this case, the receiver will be able to draw no inference beyond the prior from a signal received in equilibrium. More interesting is the possibility that the equilibrium will be partially revealing, with some, but not all of the sender types sending common signals.

It is not difficult to construct pooling equilibria for the basic signaling game. Take the labor market model and assume S sends the message s^* with probability one and that the receiver responds to s^* with his best response to the prior distribution and to all other messages with the best response to the belief that t is the least skilled agent. Provided that the least skilled agent prefers to send s^* to sending the cheapest alternative signal, this is a Nash Equilibrium outcome.

The Basic Model

The separating equilibrium is a benchmark outcome for signaling games. When a separating equilibrium exists, then it is possible for the sender to share her information fully with the receiver in spite of having a potential conflict of interest.

Existence of separating equilibria typically requires a systematic relationship between types and signals. An appropriate condition, commonly referred to as the single-crossing condition, plays a prominent role in signaling games and in models of asymmetric information more generally.

In this section I limit attention to a special class of signaling game in which there is a monotonic relationship between types and signals. In these models, separating equilibria typically exist.

I begin by stating the assumption in the environment most commonly seen in applications. Assume that the sets T, S, and A are all real intervals.

Definition 2 $U^S(\cdot)$ satisfies the single-crossing condition if $U^S(t, s, a) \leq U^S(t, s', a')$ for $s' > s$ implies that $U^S(t', s, a) < U^S(t', s', a')$ for all $t' > t$.

In a typical application, $U^S(\cdot)$ is strictly decreasing in its second argument (the signal) and increasing in its third argument (R's response) for all types. Consequently indifference curves are well defined in $M \times A$ for all t. The single-crossing condition states that indifference curves of different sender types cross once. If a lower type is indifferent between two signal-action pairs, then a higher type strictly prefers to send the higher signal. In this way, the single-crossing condition links signals to types in such a way as to guarantee that higher types send weakly higher signals in equilibrium.

Note two generalizations of Definition 2. First, the assumption that the domain of $U^S(\cdot)$ is the product of intervals can be replaced by the assumption that these sets are partially ordered. In this case, weak and strict order replace the weak and strict inequalities comparing types and actions in the statement of the definition. Second, it is sometimes necessary to extend the definition to mixed strategies. In this case, the ordering of A induces a partial ordering of distributions of A through first-order stochastic dominance.

When one thinks of the single-crossing condition geometrically, it is apparent that it implies a ranking of the slopes of the indifference curves of the sender. Suppose that $U^S(\cdot)$ is smooth, strictly increasing in actions and strictly decreasing in signals so that indifference curves are well defined for each t. Writing the indifference curve as $\{(s, \bar{a}(s; t))\}$, it must be that $U^S(t, s, \bar{a}(s; t)) \equiv 0$, so that

the slope of the indifference curve of a type t sender is

$$\bar{a}_1(s;t) = -\frac{U_2^S(t,s,a)}{U_3^S(t,s,a)}\,, \tag{5}$$

where $\bar{a}_1(s;t)$ is the partial derivative of $\bar{a}(s;t)$ with respect to the first argument, and $U_k^S(\cdot)$ denotes the partial derivative of $U^S(\cdot)$ with respect to its kth argument. Under these conditions, the single-crossing condition is implied by the requirement that the right-hand side of Eq. (5) is decreasing in t. The differentiable version of the single-crossing condition is often referred to as the Spence–Mirrlees condition. Milgrom and Shannon [57] contains general definitions of the single-crossing and Spence–Mirrlees conditions and Edlin and Shannon [26] provides a precise statement of when the conditions are equivalent.

To provide a simple construction of a separating equilibrium, limit attention to a standard signaling game in which the following conditions hold.

1. $T = \{0, \ldots, K\}$ is finite.
2. A and M are real intervals.
3. Utility functions are continuous in action and signal.
4. $U^S(\cdot)$ is strictly increasing in action and strictly decreasing in signal.
5. The single-crossing property holds.
6. The receiver's best-response function is uniquely defined, independent of the signal, and strictly increasing in t so that it can be written $BR(t)$.
7. There exists $\bar{s} \in S$ such that $U^S(K, \bar{s}, BR(K)) < U^S(K, s_0^*, BR(0))$.

Conditions 1 and 2 simplify exposition, but otherwise are not necessary. It is important that T, M, and A be partially ordered so that some kind of single-crossing condition applies. Conditions 4–6 impose a monotone structure on the problem so that higher types are more able to send high signals, and that higher types induce higher (and uniformly more attractive) actions. These conditions imply that in equilibrium higher types will necessarily send weakly higher signals. Condition 7 is a boundary condition that makes sending high signals unattractive. It states that the highest type of sender would prefer to be treated like the lowest type rather than use the signal $\bar{s}$. These properties hold in many standard applications. Condition 6 would be satisfied if $U^R(t,s,a) = -(a-t)^2$.

Separating Equilibrium

To illustrate these ideas, consider a construction of a separating equilibrium.

Proposition 2 *The standard signaling game has a separating equilibrium.*

One can prove the proposition by constructing a possible equilibrium path and confirming that the path can be part of a separating equilibrium.

Step 1. t_0 selects the signal s_0^* that maximizes $U^S(t_0, s, BR(t_0))$.

Step 2. Suppose that s_i^* have been specified for $i = 0, \ldots, k-1$ and let $U^*(t_i) = U^S(t_i, s_i^*, BR(t_i))$. Define s_k^* to solve:

$$\max U^S(t_k, s, BR(t_k)) \text{ subject to } U^S(t_{k-1}, s, BR(t_k)) \le U^*(t_{k-1})\,.$$

Provided that the optimization problems in Steps 1 and 2 have solutions, the process inductively produces a signaling strategy for the sender and a response rule for the receiver defined on $\{s_0^*, \ldots, s_K^*\}$. When $BR(\cdot)$ is strictly increasing, the single-crossing condition implies that the signaling strategy is strictly increasing. To complete the description of strategies, assume that the receiver takes the action $BR(t_k)$ in response to signals in the interval $[s_k, s_{k+1})$, $BR(t_0)$ for $s < s_0^*$, and $BR(t_K)$ for $s > s_K^*$. By the definition of the best-response function, the receiver is best responding to the sender's strategy. When the boundary condition fails, a fully separating equilibrium need not exist, but when M is compact, one can follow the construction above to obtain an equilibrium in which the lowest types separate and higher types pool at the maximum signal in M (see Cho and Sobel [22] for details).

In the construction, the equilibrium involves inefficient levels of signaling. When $U^S(\cdot)$ is decreasing in the signal, all but the lowest type of sender must make a wasteful expenditure in the signal in order to avoid being treating as having a lower quality. The result that expenditures on signals are greater than the levels optimal in a full-information model continue to hold when $U^S(\cdot)$ is not monotonic in the signal. The sender inevitably does no better in a separating equilibrium than she would do if R had full information about t. Indeed, all but the lowest type will do strictly worse in standard signaling games. On the other hand, the equilibrium constructed above has a constrained efficiency property: Of all separating equilibria, it is Pareto dominant from the standpoint of S. To confirm this claim argue inductively that in any separating equilibrium if t_j sends the signal s_j, then $s_j \ge s_j^*$, with equality only if all types $i < j$ send s_i^* with probability one.

Mailath [49] provides a similar construction when T is a real interval. In this case, the Spence–Mirrlees formulation of the single-crossing condition plays an important role and the equilibrium is a solution to a differential equation.

Multiple Equilibria and Selection

Section "Equilibrium" ended with the construction of a pooling equilibrium. A careful reconsideration of the argument reveals that there typically are many pooling equilibrium outcomes. One can construct a potential pooling outcome by assuming that all sender types send the same signal, that the receiver best responds to this common signal, and responds to all other signals with the least attractive action. Under the standard monotonicity assumptions, this strategy profile will be an equilibrium if the lowest sender type prefers pooling to sending the cheapest available out-of-equilibrium message. Section "Separating Equilibrium" ended with the construction of a separating equilibrium. There are also typically many separating equilibrium outcomes. Assume that types $t = 0, \ldots, r-1$ send signals $s^*(t)$, type r sends $\tilde{s}(k) > s^*(k)$, and subsequent signals $\tilde{s}^*(t)$ for $t > r$ solve:

$$\max U^S(t_k, s, BR(t_k)) \text{ subject to}$$
$$U^S(t_{k-1}, s, BR(t_k)) \leq U(t_{k-1}, \tilde{s}, BR(t_{k-1}))\ .$$

In both of these cases, the multiplicity is typically profound, with a continuum of distinct equilibrium outcomes (when M is an interval). The multiplicity of equilibria means that, without refinement, equilibrium theory provides few clear predictions beyond the observation that the lowest type of sender receives at least $U^*(t_0)$, the payoff it would receive under complete information, and the fact that the equilibrium signaling function is weakly increasing in the sender's type. The first property is a consequence of the monotonicity of S's payoff in a and of R's best response function. The second property is a consequence of the single-crossing condition.

This section describes techniques that refine the set of equilibria. Refinement arguments that guarantee existence and select unique outcomes for standard signaling games rely on the Kohlberg–Mertens [42] notion of strategic stability. The complete theory of strategic stability is only available for finite games. Consequently the literature applies weaker versions of strategic stability that are defined more easily for large games. Banks and Sobel [8], Cho and Kreps [21], and Cho and Sobel [22] introduce these arguments.

Multiple equilibria arise in signaling games because Nash equilibrium does not constrain the receiver's response to signals sent with zero probability in equilibrium. Specifying that R's response to these unsent signals is unattractive leads to the largest set of equilibrium outcomes. (In standard signaling games, S's preferences over actions do not depend on type, so the least attractive action is well defined.) The equilibrium set shrinks if one restricts the meaning of unsent signals. An effective restriction is condition D1, introduced in Cho and Kreps [21]. This condition is less restrictive than the notion of universal divinity introduced by Banks and Sobel [8], which in finite games is less restrictive than Kohlberg and Mertens's notion of strategic stability.

Given an equilibrium (α^*, μ^*), let $U^*(t)$ be the equilibrium expected payoff of a type t sender and let $D(s, t) = \{a\colon u(t, s, a) \geq U^*(t)\}$ be the set of pure-strategy responses to s that lead to payoffs at least as great as the equilibrium payoff for player t. Given a collection of sets, $X(t)$, $t \in T$, $X(t^*)$ is maximal if it not a proper subset of any $X(t)$.

Definition 3 Behavior strategies (α^*, μ^*) together with beliefs β^* satisfy D1 if for any unsent message s, $\beta(\cdot, s)$ is supported on those t for which $D(s, t)$ is maximal.

In standard signaling games, $D(s, t)$ is an interval: all actions greater than or equal to a particular action will be attractive relative to the equilibrium. Hence these sets are nested. If $D(s, t)$ is not maximal, then there is another type t' that is "more likely to deviate" in the sense that there exists out-of-equilibrium responses that are attractive to t' but not t. Condition D1 requires that the receiver place no weight on type t making a deviation in this case. Notice if $D(s, t)$ is empty for all t, then D1 does not restrict beliefs given s (and any choice of action will support the putative equilibrium). Condition D1 is strong. One can imagine weaker restrictions. The intuitive condition (Cho and Kreps [21]) requires that $\beta(t, s) = 0$ when $D(t, s) = \phi$ and at least one other $D(t', s)$ is non empty. Divinity (Banks and Sobel [8]) requires that if $D(t, s)$ is strictly contained in $D(t', s)$, then $\beta(t', s)/\beta(t, s) \geq \pi(t')/\pi(t)$, so that the relative probability of the types more likely to deviate increases.

Proposition 3 *The standard signaling game has a unique separating equilibrium outcome that satisfies Condition D1.*

In standard signaling games, the only equilibrium outcome that satisfies Condition D1 is the separating outcome described in the previous section. Details of the argument appear in Cho and Sobel. The argument relies on two insights. First, types cannot be pooled in equilibrium because slightly higher signals will be interpreted as coming from the highest type in the pool. Second, in any separating equilibrium in which a sender type fails to solve Step 2, deviation to a slightly lower signal will not lower R's beliefs.

The refinement argument is powerful and the separating outcome selected receives prominent attention in the literature. It is worth pointing out that the outcome has

one unreasonable property. The separating outcome described above depends only on the support of types, and not on the details of the distribution. Further, all types but the lowest type must make inefficient (compared to the full-information case) investments in signal in order to distinguish themselves from lower types. The efficient separating equilibrium for a sequence of games in which the probability of the lowest type converges to zero does not converge to the separating equilibrium of the game in which the probability of the lowest type is zero. In the special case of only two types, the (efficient) pooling outcome may be a more plausible outcome when the probability of the lower type shrinks to zero. Grossman and Perry [38] and Mailath, Okuno-Fujiwara, and Postlewaite [50] introduce equilibrium refinements that select the pooling equilibrium in this setting. These concepts share many of the same motivations of the refinements introduced by Banks and Sobel and Cho and Kreps. They are qualitatively different from the intuitive criterion, divinity, and Condition D1, because they are not based on dominance arguments and lack general existence properties.

Cheap Talk

Models in which preferences satisfy the single-crossing property are central in the literature, but the assumption is not appropriate in some interesting settings. This section describes an extreme case in which there is no direct cost of signaling.

In general, a cheap-talk model is a signaling model in which $u^i(t, s, a)$ is independent of s for all (t, a). Two facts about this model are immediate. First, if equilibrium exists, then there always exists an equilibrium in which no information is communicated. To construct this "babbling" equilibrium, assume that $\beta(t, s)$ is equal to the prior independent of the signal s. R's best response will be to take an action that is optimal conditional only on his prior information. Hence R's action can be taken to be constant. In this case, it is also a best response for S to send a signal that is independent of type, which makes $\beta(t, s)$ the appropriate beliefs. Hence, even if the interests of S and R are identical, so that it there are strong incentives to communicate, there is a possibility of complete communication break down.

Second, it is clear that non-trivial communication requires that different types of S have different preferences over R's actions. If it is the case that whenever some type t prefers action a to action a' then so do all other types, then (ruling out indifference), it must be the case that in equilibrium the receiver takes only one action with positive probability. To see this, note that otherwise one type of sender is not selecting a best response. The second observation shows that cheap talk is not effective in games, like the standard labor-market story, in which the sender's preferences are monotonic in the action of the receiver. With cheap communication, the potential employee in the labor market will always select a signal that leads to the higher possible wage and consequently, in equilibrium, all types of workers will receive the same wage.

A Simple Cheap-Talk Game

There are natural settings in which cheap talk is meaningful in equilibrium. To describe examples, I follow the development of Crawford and Sobel [24] (Green and Stokey [35] independently introduced a similar game in an article circulated in 1981). In this paper, A and T are the unit interval and M can be taken to be the unit interval without loss of generality. The sender's private information or type, t, is drawn from a differentiable probability distribution function, $F(\cdot)$, with density $f(\cdot)$, supported on $[0, 1]$. S and R have twice continuously differentiable von Neumann–Morgenstern utility functions $U^i(a, t)$ that are strictly concave in a and have a strictly positive mixed partial derivative. Let $i = R, S$, $a^i(t)$ denotes the unique solution to $\max_a U^i(a, t)$ and further assume that $a^S(t) > a^R(t)$ for all t. (The assumptions on $U^i(\cdot)$ guarantee that $U^i(\cdot)$ is well defined and strictly increasing.)

In this model, the interests of the sender and receiver are partially aligned because both would like to take a higher action with a higher t. The interests are different because S would always like the action to be a bit higher than R's ideal action. In a typical application, t represents the idea action for R, such as the appropriate expenditure on a public project. Both R and S want actual expenditure to be close to the target value, but S has a bias in favor of additional expenditure.

For $0 \le t' < t'' \le 1$, let $\bar{a}(t', t'')$ be the unique solution to $\max_a \int_{t'}^{t''} U^R(a, t)\mathrm{d}F(t)$. By convention, $\bar{a}(t, t) = a^R(t)$.

Without loss of generality, limit attention to pure-strategy equilibria. The concavity assumption guarantees that R's best responses will be unique, so R will not randomize in equilibrium. An equilibrium with strategies (μ^*, α^*) **induces** action a if $\{t\colon \alpha^*(\mu^*(t)) = a\}$ has positive prior probability. Crawford and Sobel [24] characterize equilibrium outcomes.

Proposition 4 *There exists a positive integer N^* such that for every integer N with $1 \le N \le N^*$, there exists at least one equilibrium in which the set of induced actions has cardinality N, and moreover, there is no equilibrium*

which induces more than N^* actions. An equilibrium can be characterized by a partition of the set of types, $t(N) = (t_0(N), \ldots, t_N(N))$ with $0 = t_0(N) < t_1(N) < \ldots < t_N(N) = 1$, and signals m_i, $i = 1, \ldots, N$, such that for all $i = 1, \ldots, N-1$

$$U^S(\bar{a}(t_i, t_{i+1}), t_i)) - U^S(\bar{a}(t_{i-1}, t_i), t_i)) = 0\,, \tag{6}$$

$$\mu(t) = m_i \;\; for \;\; t \in (t_{i-1}, t_i]\,, \tag{7}$$

and

$$\alpha(m_i) = \bar{a}(t_{i-1}, t_i)\,. \tag{8}$$

Furthermore, essentially all equilibrium outcomes can be described in this way.

In an equilibrium, adjacent types pool together and send a common message. Condition (6) states that sender types on the boundary of a partition element are indifferent between pooling with types immediately below or immediately above. Condition (7) states that types in a common element of the partition send the same message. Condition (8) states that R best responds to the information in S's message.

Crawford and Sobel make another monotonicity assumption, which they call condition (M). (M) is satisfied in leading examples and implies that there is a unique equilibrium partition for each $N = 1, \ldots, N^*$, the ex-ante equilibrium expected utility for both S and R is increasing in N, and N^* increases if the preferences of S and R become more aligned. These conclusions provide justification for the view that with fixed preferences "more" communication (in the sense of more actions induced) is better for both players and that the closer are the interests of the players the greater the possibilities for communication.

As in the case of models with costly signaling, there are multiple equilibria in the cheap-talk model. The multiplicity is qualitatively different. Costly signaling models have a continuum of Nash Equilibrium outcomes. Cheap-talk models have only finitely many. Refinements that impose restrictions on off-the-equilibrium path signals work well to identify a single outcome in costly signaling models. These refinements have no cutting power in cheap-talk models because any equilibrium distribution on type-action pairs can arise from signaling strategies in which all messages are sent with positive probability. To prove this claim, observe that if message m' is unused in equilibrium, while message m is used, then one can construct a new equilibrium in which R interprets m' the same way as m and sender types previously sending m randomize equally between m and m'.

In the basic model messages take on meaning only through their use in an equilibrium. Unlike natural language, they have no external meaning. There have been several attempts to formalize the notion that messages have meanings that, if consistent with strategic aspects of the interaction, should be their interpretation inside the game. The first formulation of this idea is due to Farrell [28].

Definition 4 Given an equilibrium (α^*, σ^*) with sender expected payoffs $u^*(\cdot)$, the subset $G \subset T$ is self signaling if $G = \{t\colon U^S(t, BR(G)) > u^*(t)\}$.

That is, G is self signaling if precisely the types in G gain by making a statement that induces the action that is a best response to the information that $t \in G$. (When $BR(t)$ is not single valued it is necessary to refine the definition somewhat and permit the possibility that $U^S(t, BR(G)) = u^*(t)$ for some t. See Matthews, Okuno-Fujiwara, and Postlewaite [51].) Farrell argues that the existence of a self-signaling set would destroy an equilibrium. If a subset G had available a message that meant "my type is in G," then relative to the equilibrium R could infer that if he were to interpret the message literally, then it would be sent only by those types in G (and hence the literal meaning would be accurate). With this motivation, Farrell proposes a refinement.

Definition 5 An equilibrium (α^*, σ^*) is neologism proof if there exist no self-signaling sets relative to the equilibrium.

Rabin [62] argues convincingly that Farrell's definition rules out too many equilibrium outcomes. Indeed, for leading examples of the basic cheap-talk game, there are no neologism-proof equilibria. Specifically, in the Crawford–Sobel model in which S has a bias towards higher actions, there exist self signaling sets of the form $[t, 1]$. On the other hand, Chen, Kartik, and Sobel [20] demonstrate that the N^*-step equilibrium always satisfies the no incentive to separate (NITS) condition:

$$U^S(\alpha^*(\mu^*(0)), 0) \geq U^S(a^R(0), 0)\,, \tag{9}$$

and that under condition (M) this is the only equilibrium that satisfies Condition (9).

NITS states that the lowest type of sender prefers her equilibrium payoff to the payoff she would receive if the receiver knew her type (and responded optimally). [41] introduced and named this condition. The NITS condition can be shown to rule out equilibria that admit if self-signaling sets of the form $[0, t]$. Chen [19] and Kartik [41] show that the condition holds in the limits of perturbed versions of the basic cheap-talk game.

Inequality (9) holds automatically in any perfect Bayesian equilibrium of the standard signaling model. This follows because when R's actions are monotonic in type and S's preferences are monotonic in action, the worst outcome for S is to be viewed as the lowest type. This observation would not be true in Nash Equilibrium, where it is possible for R to respond to an out-of-equilibirum message with an action $a < BR(0)$.

Variations on Cheap Talk

In standard signaling models, there is typically an equilibrium that is fully revealing. This is not the case in the basic cheap-talk model. This leads to the question of whether it is possible to obtain more revelation in different environments.

One possibility is to consider the possibility of signaling over many dimensions. Chakraborty and Harbaugh [18] consider a model in with T and A are multidimensional. A special case of their model is one in which the components of T are independent draws from the same distribution and A involves taking a real-valued action for each component of T. If preferences are additively separable across types and actions, Chakraborty and Harbaugh provide conditions under which categorical information transmission, in which the S transmits the order of the components of T, is credible in equilibrium even when it would not be possible to transmit information if the dimensions were treated in isolation. It may be credible for S to say "$t_1 > t_2$," even if she could not credibly provide information about the absolute value of either component of t.

Communication is non-trivial if some Sender type strictly prefers to induce one equilibrium action over another. Non-trivial communication requires that different types have different preferences over outcomes. In standard signaling models, the heterogeneity arises because different sender types have different costs of sending messages. In cheap-talk models, the heterogeneity arises with one-dimensional actions if different sender types have different ideal actions. With multi-dimensional actions, heterogeneity could come simply from different sender types having different preferences over the relative importance of the different issues. Another simple variation is to assume the existence of more than one sender. In the two-sender game, nature picks t as before, both senders learn t and simultaneously send a message to the receiver, who makes a decision based on the two messages. The second sender has preferences that depend on type and the receiver's action, but not directly on the message sent. In this environment, assume that $M = T$, so that the set of available messages (this is essentially without loss of generality). One can look for equilibria in which the senders report honestly. Denote by $a^*(t, t')$ R's response to the pair of messages (t, t'). If an equilibrium in which both senders report honestly exists, then R's response to identical messages, $a^*(t, t) = a^R(t)$, and it must be the case that there exists a specification of $a(t, t')$ for $t \neq t'$ such that for all $i = 1$ and 2 and $t \neq t'$,

$$U^{S_i}(t, a^*(t, t)) \geq U^{S_i}(t, a^*(t, t')) \,. \tag{10}$$

It is possible to satisfy Condition (10) if the biases of the senders are small relative to the set of possible best responses. Krishna and Morgan [45] studies a one-dimensional model of information transmission with two informed players. Ambrus and Takahashi [1] and Battiglini [9] provide conditions under which full revelation is possible when there are two informed players and possibly multiple dimensions of information.

In many circumstances, enriching the communication structure either by allowing more rounds of communication [2,29], mediation [10], or exogenous uncertainty [16] enlarges the set of equilibrium outcomes.

Verifiable Information

Until now, the focus has been on situations in which the set of signals available does not depend on the true state. There are situations in which this assumption is not appropriate. There may be laws that ban false advertisement. The sender may be able to document details about the value of t. Models of this kind were first studied by Grossman [37] and Milgrom [56]. For example, if t is the sender's skill at playing the piano, then if there is a piano available t could demonstrate that she has skill at least as great as t (by performing at her true ability), but she may not be able to prove that her skill is no more than t (the receiver may think that she deliberately played the piano badly).

To model these possibilities, suppose that the set of possible messages is the set of all subsets of T. In this case, messages have "literal" meanings: When the sender uses the message $s = C \in T$, this can be interpreted as a statement of the form: "my type is in C." If senders cannot lie, then $M(t)$ must be the set of subsets of T that contain t. If type t is verifiable, then $\{t\} \in M(t')$ if and only if $t' = t$. If there are no additional costs of sending signals, this model can be viewed as a variation of cheap talk models in which the message space depends on t. In general, one can treat verifiable information models as a special case of the general signaling game in which the cost of sending certain signals is so large that these signals can be ruled out. Lying

is impossible if $M(t) = \{C \subset 2^T : t \in C\}$. In this setting, it is appropriate to require equilibria to be consistent with the signaling structure.

Definition 6 The equilibrium (σ^*, α^*) is rationalizable if

$$\alpha(C, a) > 0 \text{ implies } \sum_{t \in T} U^R(t, s, a)\beta(t, a) = \max_{a' \in A} \sum_{t \in T} U^R(t, s, a')\beta(t, a') , \quad (11)$$

where $\beta(t, a) = 0$ if $t \notin C$.

Compared to (2), (11) requires that beliefs place positive probability only on types capable of sending the message "my type is an element of C."

Proposition 5 *Suppose that A and T are linearly ordered, that the receiver's best response function is increasing in type, and that all sender types prefer higher actions. If lying is not possible, then in any rationalizable equilibrium (α^*, μ^*), $\alpha^*(s, BR(t)) = 1$ whenever $\mu^*(t, s > 0)$.*

Grossman [37] and Milgrom [56] present versions of this proposition. Seidman and Winter [68] generalize the result.

Provided that the receiver responds to the signal $\{t\}$ with $BR(t)$, each type can guarantee a payoff of $BR(t)$. On the other hand, if any type receives a payoff greater than $BR(t)$, then some higher type must be doing worse. Another way to make the same point is to notice that the highest type $\{\bar{t}\}$ has a weakly dominant strategy to reveal her type by announcing $\{\bar{t}\}$. Once this type is revealed, the next highest type will want to reveal herself and so on. Hence verifiable information will be revealed voluntarily in an environment where cheap talk leads to no revealing and costly signaling will be compatible with full revelation, but at the cost of dissipative signaling.

The full-revelation result depends on the assumption that the sender and receiver share a linear ranking over the quality of information. Giovannini and Seidmann [31] discuss more general settings in which the ability to provide verifiable information need not lead to full revelation.

Communication About Intentions

In a simple signaling game, signals potentially provide information about private information. Another possibility is to add a round of pre-play communication to a given game. Even if the game has complete information, there is the possibility that communication would serve to select equilibria or permit correlation that would otherwise be infeasible. Farrell and Rabin [63]'s review article discusses this literature in more detail.

Aumann [3] argues that one cannot rely on pre-play communication to select a Pareto-efficient equilibrium. He considers a simple two-player game with Pareto-ranked equilibria and argues that no "cheap" pre-play signal would be credible.

Ben-Porath and Dekel [11] show that adding a stage of "money burning" (a signal that reduces all future payoffs by the same amount) when combined with an equilibrium refinement can select equilibria in a complete information game. Although no money is burned in the selected equilibrium outcome, the potential to send costly signals creates dominance relationships that lead to a selection.

Vida [74] synthesizes a literature that compares the set of equilibrium outcomes available when communication possibilities are added to a game to the theoretically larger set available if there is a reliable mediator available to collect information and recommend actions to the players.

Applications

Economic Applications

There is an enormous literature that uses signaling models in applications. Riley's [64] survey contains extended discussion of some of the most important applications. What follows is a brief discussion of some central ideas.

In a simple signaling game, one informed agent sends a single signal to one uninformed decision maker. This setting is reach enough to illustrate many important aspects of signaling, but it plainly limited. Interesting new issues arise if there are many informed agents, if there are many decision makers, and if the interaction is repeated. Several of the models below add some or all of these novel features to the basic model.

Advertising Advertisements are signals. Models similar to the standard model can explain situations in which higher levels of advertisement can lead consumers to believe the quality of the good is higher. In a separating equilibrium, advertising expenditures fully reveal quality. As in all costly signaling models, it is not important that there be a direct relationship between quality and signal, it is only necessary that firms with higher quality have lower marginal costs of advertising. Hence simply "burning money" or sending a signal that lowers utility by an amount independent of quality and response can be informative. The consumer may obtain full information in equilibrium, but someone must pay the cost of advertis-

ing. There are other situations where it is natural for the signal to be linked to the quality of the item. Models of verifiable information are appropriate in this case. When the assumptions of Proposition 5 hold, one would expect consumers to obtain all relevant information through disclosures without wasteful expenditures on signaling. Finally, cheap talk plays a role in some markets. One would expect costless communication to be informative in environments where heterogeneous consumers would like to identify the best product. Cheap talk can create more efficient matching of product to consumer. Here communication is free although in leading models separating equilibria do not exist.

Limit Pricing Signaling models offer one explanation for the phenomenon of limit pricing. An incumbent firm have private information about its cost. Potential entrants use the pricing behavior of the firm to draw inferences about the incumbent's cost, which determines profitability of entry. Milgrom and Roberts [54] construct an equilibrium in which the existence of incomplete information distorts prices: Relative to the full information model, the incumbent charges lower prices in order to signal that the market is relatively unprofitable. This behavior has the flavor of classical models of limit pricing, with one important qualification. In a separating equilibrium the entrant can infer the true cost of the incumbent and therefore the low price charged by the incumbent firm fails to change the entry decision.

Bargaining Several authors have proposed bargaining models with incomplete information to study the existence and duration of strikes [30,71]. If a firm with private information about its profitability makes a take-it-or-leave it offer to a union, then the strategic interaction is a simple signaling model in which the magnitude of the offer may serve as a signal of the firm's profitability. Firms with low profits are better able to make low wage offers to the union because the threat of a strike is less costly to a firm with low profits than one with high profits. Consequently settlement offers may reveal information. Natural extensions of this model permit counter offers. The variation of the model in which the uninformed agent makes offers and the uninformed agent accepts and rejects is formally almost identical to the canonical model of price discrimination by a durable-goods monopolist [4,39].

Finance Simple signaling arguments provide potential explanations for firms' choices of financial structure. Classic arguments due to Modigliani and Miller [58] and imply that firms' profitability should not depend on their choice of capital structure. Hence this theory cannot organize empirical regularities about firm's capital structure. The Modigliani–Miller theorem assumes that the firm's managers, shareholders, and potential shareholders all have access to the same information. An enormous literature assumes instead that the firm's managers have superior information and use corporate structure to signal profitability.

Leland and Pyle [47] assume that insiders are risk averse so they would prefer to diversify their personal holdings rather than maintain large investments in their firm. The value of diversification is greater the lower the quality of the firm. Hence when insiders have superior information than investors, there will be an incentive for the insiders of highly profitable firms to maintain inefficiently large investments in their firm in order to signal profitability to investors.

Dividends are taxed twice under the United States tax code, which raises the question of why firms would issue dividends when capital gains are taxed at a lower rate. A potential explanation for this behavior comes from a model in which investors have imperfect information about the future profitability of the firm and profitable firms are more able than less profitable firms to distribute profits in the form of dividends (see [14]).

Reputation Dynamic models of incomplete information create the opportunity for the receiver to draw inferences about the sender's private information while engaging in an extended interaction. Kreps and Wilson [44] and Milgrom and Roberts [55] provided the original treatments of reputation formation in games of incomplete information. Motivated by the limit pricing, their models examined the interaction of a single long-lived incumbent facing a sequence of potential entrants. The entrants lack information about the willingness of the incumbent to tolerate entry. Pricing decisions of the incumbent provide information to the entrants about the profitability of the market.

In these models, signals have implications for both current and future utility. The current cost is determined by the effect the signal has on current payoffs. In Kreps–Wilson and Milgrom–Roberts, this cost is the decrease in current profits associated with charging a low price. In other models (for example [59,70]) the actual signal is costless, but it has immediate payoff implications because of the response it induces. Signals also have implications for future utility because inferences about the sender's private information will influence the behavior of the opponents in future periods. Adding concern for reputation to a signaling game will influence behavior, but whether it

leads to more or less informative signaling depends on the application.

Signaling in Biology

Signaling is important in biology. In independent and almost contemporaneous work, Zahavi [76] proposed a signaling model that shared the essential features of Spence [72]'s model of labor-market signaling. Zahavi observed that there are many examples in nature of animals apparently excessive physical displays. It takes energy to produce colorful plumage, large antlers, or loud cries. Having a large tail may actually make it harder for peacocks to flea predators. If a baby bird makes a loud sound to get his mother's attention, he may attract a dangerous predator. Zahavi argued that costly signals could play a role in sexual selection. In Zahavi's basic model, the sender is a male and the receiver is a female of the same species. Females who are able to mate with healthier males are more likely to have stronger children, but often the quality of a potential mate cannot be observed directly. Zahavi argued that if healthier males could produce visible displays more cheaply than less healthy males, then females would be induced to use the signals when deciding upon a mate. Displays may impose costs that "handicap" a signaler, but displays would persist when additional reproductive success compensates for their costs. Zahavi identifies a single-crossing condition as a necessary condition for the existence of costly signals.

The development of signaling in biology parallels that in economics, but there are important differences. Biology replaces the assumption of utility maximization and equilibrium with fitness maximization and evolutionary stability. That is, their models do not assume that animals consciously select their signal to maximize a payoff. Instead, the biological models assume that the process of natural selection will lead to strategy profiles in which mutant behavior has lower reproductive fitness than equilibrium behavior. This notion leads to static and dynamic solution concepts similar to Nash Equilibrium and its refinements. Fitness in biological models depends on contributions from both parents. Consequently, a full treatment of signaling must take into account population genetics. Grafen [34] discusses these issues and Grafen [33] and Siller [69] provide further theoretical development of the handicap theory. Finally, one must be careful in interpreting heterogeneous quality in biological models. Natural selection should operate to eliminate the least fit genes in a population. To the extent that this arises, there is pressure for quality variation within a population to decrease over time. The existence of unobserved quality variations needed for signaling may be the result of relatively small variations about a population norm.

While most of the literature on signaling in biology focuses on the use of costly signals, there are also situations in which cheap talk is effective. A leading example is the "Sir Philip Sidney Game," originally developed by John Maynard Smith [53] to illustrate the value of costly communication between a mother and child. The child has private information about its level of hunger and the mother must decide to feed the child or keep the food for itself. Since the players are related, survival of one positively influences the fitness of the other. This creates a common interest needed for cheap-talk communication. There are two ways to model communication in this environment. The first is to assume that signaling is costly, with hungrier babies better able to communicate their hunger. This could be because the sound of a hungry baby is hard for sated babies to imitate or it could be that crying for food increases the risk of predation and that this risk is relatively more dangerous to well fed chicks than to starving ones (because the starving chicks have nothing to lose). This game has multiple equilibria in which signals fully reveal the state of the baby over a range of values (see [46,53]). These papers look at a model in which both mother and child have private information. Alternatively, Bergstrom and Lachmann [13] study a cheap-talk version of the game. Here there may be an equilibrium outcome in which the baby bird credibly signals whether or not he is hungry. Those who signal hunger get fed. The others do not. Well fed baby birds may wish to signal that they are not hungry in order to permit the mother to keep food for herself. Such an equilibrium exists if the fraction of genes that mother and child share is large and the baby is already well fed.

Political Science

Signaling games have played an important role in formal models of political science. Banks [7] reviews models of agenda control, political rhetoric, voting, and electoral competition. Several important models in this area are formally interesting because they violate the standard assumptions frequently satisfied in economic models. I describe two such models in this subsection.

Banks [6] studies a model of agenda setting in which the informed sender proposes a policy to a receiver (decision-maker), who can either accept or reject the proposal. If the proposal is accepted, it becomes the outcome. If not, then the outcome is a fall-back policy. The fall-back policy is known only to the sender. In this environment, the

sender's strategy may convey information to the decision maker. Signaling is costly, but, because the receiver's set of actions in binary, fully revealing equilibria need not exist. Refinements limit the set of predictions in this model to a class of outcomes in which only one proposal is accepted in equilibrium (and that this proposal is accepted with probability one), but there are typically a continuum of possible equilibrium outcomes.

Matthews [52] develops a cheap-talk model of veto threats. There are two players, a Chooser (C), who plays the role of receiver, and a Proposer (P), who plays the role of sender. The players have preferences that are represented by single-peaked utility functions which depend on the real-valued outcome of the game and an ideal point. P's ideal point is common knowledge. C's ideal point is her private information, drawn from a prior distribution that has a smooth positive density on a compact interval, $[\underline{t}, \bar{t}]$. The game form is simple: C learns her type, then sends a cheap-talk signal to P, who responds with a proposal. C then either accepts or rejects the proposal. Accepted proposals become the outcome of the game. If C rejects the proposal, then the outcome is the status quo point.

As usual in cheap-talk games, this game has a babbling outcome in which C's message contains no information and P makes a single, take-it-or-leave-it offer that is accepted with probability strictly between 0 and 1. Matthews shows there may be equilibria in which two outcomes are induced with positive probability (size-two equilibria), but size $n > 2$ (perfect Bayesian) equilibria never exist. In a size-two equilibrium, P offers his ideal outcome to those types of C whose message indicates that their ideal point is low; this offer is always accepted in equilibrium. If C indicated that his ideal point is high, P makes a compromise offer that is sometimes accepted and sometimes rejected.

Future Directions

The most exciting developments in signaling games in the future are likely to come from interaction between economics and other disciplines.

Over the last ten years the influence of behavioral economists have led the profession to rethink many of its fundamental models. An explosion of experimental studies have already influenced the interpretation of signaling models and have led to a re-examination of basic assumptions. There is evidence that economic actors lack the strategic sophistication assumed in equilibrium models. Further, economic agents may be motivated by more than their material well being. Existing experimental evidence provides broad support for many of the qualitative predictions of the theory (Banks, Camerer, and Porter [5] and Brandts and Holt [17]), but also suggests ways in which the theory may be inadequate.

The driving assumption of signaling models is that when informational asymmetries exist, senders will attempt to lie for strategic advantage and that sophisticated receivers will discount statements. These assumptions may be reconsidered in light of experimental evidence that some agents will behave honestly in spite of strategic incentives to lie. For example, Gneezy [32] and Hurkens and Kartik [40] present experimental evidence that some agents are reluctant to lie even when there is a financial gain from doing so. There is evidence from other disciplines that some agents are unwilling or unable to manipulate information for strategic advantage and that people may be well equipped to detect these manipulations in ways that are not captured in standard models (see, for example, Ekman [27] or Trivers [73]). Experimental evidence and, possibly, results from neuroscience may demonstrate that the standard assumption that some agents cannot manipulate information for their strategic advantage (or that other agents have ability to see through deception) will inform the development of novel models of communication that include behavioral types. Several papers study the implications of including behavioral types into the standard paradigm. The reputation models of Kreps and Wilson [44] and Milgrom and Roberts [54] are two early examples. Papers on communication by Chen [19], Crawford [23], Kartik [41], and Olszewski [61] are more recent examples. New developments in behavioral economics will inform future theoretical studies.

There is substantial interest in signaling in philosophy. Indeed, the philosopher David Lewis [48] (first published in 1969) introduced signaling games prior to the contributions of Spence and Zahavi. Recently linguists have been paying more attention to game-theoretic ideas. Benz, Jäger and Van Rooij [12] collects recent work that attempts to formalize ideas from linguistic philosophy due to Grice [36]. While there have been a small number of contributions by economists in this area (Rubinstein [66] and Sally [67] are examples), there is likely to be more active interaction in the future.

Finally, future work may connect strategic aspects of communication to the actual structure of language. Blume [15], Cucker, Smale, Zhou [25], and Nowak and Krakauer [60] present dramatically different models on how structured communication may result from learning processes. Synthesizing these approaches may lead to fundamental insights on how the ability to send and receive signals develops.

Acknowledgments

I thank the Guggenheim Foundation, NSF, and the Secretaría de Estado de Universidades e Investigación del Ministerio de Educación y Ciencia (Spain) for financial support and Jose Penalva for his comments. I am grateful to the Departament d'Economia i d'Història Econòmica and Institut d'Anàlisi Econòmica of the Universitat Autònoma de Barcelona for hospitality and administrative support.

Bibliography

Primary Literature

1. Ambrus A, Takahashi S (2008) Multi-sender cheap talk with restricted state space. Theor Econ 3(1):1–27
2. Aumann R, Hart S (2003) Long cheap talk. Econometrica 71(6):1619–1660
3. Aumann RJ (1990) Nash equilibria and not self-enforcing. In: Gabszewicz JJ, Richard J-F, Wolsey LA (eds) Economic Decision Making: Games, Econometrics and Optimisation. Elsevier, Amsterdam, pp 201–206
4. Ausubel LM, Deneckere RJ (1989) Reputation in bargaining and durable goods monopoly. Econometrica 57(3):511–531
5. Banks J, Camerer C, Porter D (1994) An experimental analysis of nash refinements in signaling games. Games Econ Behav 6(1):1–31
6. Banks JS (1990) Monopoly agenda control and asymmetric information. Q J Econ 105(2):445–464
7. Banks JS (1991) Signaling Games in Political Science. Routledge, Langhorne
8. Banks JS, Sobel J (1987) Equilibrium selection in signaling games. Econometrica 55(3):647–661
9. Battaglini M (2002) Multiple referrals and multidimensional cheap talk. Econometrica 70(4):1379–1401
10. Ben-Porath E (2003) Cheap talk in games with incomplete information. J Econ Theory 108:45–71
11. Ben-Porath E, Dekel E (1992) Signaling future actions and the potential for sacrifice. J Econ Theory 57:36–51
12. Benz A, Jäger G, Van Rooij R (eds) (2005) Game Theory and Pragmatics. Palgrave MacMillan, Houndmills, Basingstoke
13. Bergstrom CT, Lachmann M (1998) Signalling among relatives. III. talk is cheap. Proc Natl Acad Sci USA 95:5100–5015
14. Bhattacharya S (1979) Imperfect information, dividend policy, and the bird in the hand fallacy. Bell J Econ 10(1):259–270
15. Blume A (2000) Coordination and learning with a partial language. J Econ Theory 95:1–36
16. Blume A, Board O, Kawamura K (2007) Noisy talk. Theor Econ 2(4):396–440
17. Brandts J, Holt CA (1992) An experimental test of equilibrium dominance in signaling games. Am Econ Rev 82(5):1350–1365
18. Chakraborty A, Harbaugh R (2007) Comparative cheap talk. J Econ Theory 132(1):70–94
19. Chen Y (2005) Perturbed communication games with honest senders and naive receivers. Technical report. Arizona State University, Tempe
20. Chen Y, Kartik N, Sobel J (2007) On the robustness of informative cheap talk. Technical report, UCSD. In: Econometrica 76(1):117–136, Blackwell Synergy
21. Cho I-K, Kreps DM (1987) Signaling games and stable equilibria. Q J Econ 102(2):179–221
22. Cho I-K, Sobel J (1990) Strategic stability and uniqueness in signaling games. J Econ Theory 50(2):381–413
23. Crawford VP (2003) Lying for strategic advantage: Rational and boundedly rational misrepresentation of intentions. Am Econ Rev 93(1):133–149
24. Crawford VP, Sobel J (1982) Strategic information transmission. Econometrica 50(6):1431–1451
25. Cucker F, Smale S, Zhou D-X (2004) Modeling language evolution. Found Comput Math 4(3):315–343
26. Edlin AS, Shannon C (1998) Strict single-crossing and the strict spence-mirrlees condition: A comment on monotone comparative statics. Econometrica 66(6):1417–1425
27. Ekman P (2001) Telling Lies. W.W. Norton, New York
28. Farrell J (1993) Meaning and credibility in cheap-talk games. Games Econ Behav 5(4):514–531
29. Forges F (1990) Equilibria with communication in a job market example. Q J Econ 105(2):375–398
30. Fudenberg D, Tirole J (1983) Sequential bargaining with incomplete information. Rev Econ Stud 50(2):221–247
31. Giovannoni F, Seidmann DJ (2007) Secrecy, two-sided bias and the value of evidence. Games Econ Behav 59(2):296–315
32. Gneezy U (2005) Deception: The role of consequences. Am Econ Rev 95(1):384–394
33. Grafen A (1990) Biological signals as handicaps. J Theor Biol 144:517–546
34. Grafen A (1990) Sexual selection unhandicapped by the fisher process. J Theor Biol 144:473–516
35. Green JR, Stokey NL (2007) A two-person game of information transmission. J Econ Theory 135(1):90–104
36. Grice HP (1991) Studies in the Way of Words. Harvard University Press, Cambridge
37. Grossman S (1981) The role of warranties and private disclosure about product quality. J Law Econ 24:461–483
38. Grossman S, Perry M (1987) Perfect sequential equilibria. J Econ Theory 39:97–119
39. Gul F, Sonnenchein H, Wilson R (1986) Foundations of dynamic monopoly and the coase conjecture. J Econ Theory 39:155–190
40. Hurkens S, Kartik N (2006) (When) Would I lie to you? Comment on Deception: The role of consequences. Technical report, UCSD
41. Kartik N (2005) Information transmission with almost-cheap talk. Technical report, UCSD
42. Kohlberg E, Mertens J-F (1986) On the strategic stability of equilibria. Econometrica 54(5):1003–1037
43. Kreps DM, Sobel J (1994) Signalling. In: Aumann RJ, Hart S (eds) Handbook of game theory: with economics applications. Handbooks in Economics, no 11, vol 2. Elsevier, Amsterdam, chap 25, pp 849–868
44. Kreps DM, Wilson R (1982) Reputation and imperfect information. J Econ Theory 27:253–277
45. Krishna V, Morgan J (2001) A model of expertise. Q J Econ 116(2):747–775
46. Lachmann M, Bergstrom CT (1998) Signalling among relatives. II. beyond the tower of babel. Theor Popul Biol 54:146–160
47. Leland HE, Pyle DH (1977) Informational asymmetries, financial structure, and financial intermediation informational asymmetries, financial structure, and financial intermediation. J Finance 32(2):371–387

48. Lewis D (2002) Convention: A Philosophical Study. Blackwell, Oxford
49. Mailath GJ (1987) Incentive compatibility in signaling games with a continuum of types. Econometrica 55(6):1349–1365
50. Mailath GJ, Okuno-Fujiwara M, Postlewaite A (1993) On belief based refinements in signaling games. J Econ Theory 60(2):241–276
51. Mathews SA, Okuno-Fujiwara M, Postlewaite A (1991) Refining cheap-talk equilibria. J Econ Theory 55(2):247–273
52. Matthews SA (1989) Veto threats: Rhetoric in a bargaining game. Q J Econ 104(2):347–369
53. Maynard Smith J (1991) Honest signalling: the Philip Sidney game. Anim Behav 42:1034–1035
54. Milgrom P, Roberts J (1982) Limit pricing and entry under incomplete information: An equilibrium analysis. Econometrica 50(2):443–459
55. Milgrom P, Roberts J (1982) Predation, reputation, and entry deterence. J Econ Theory 27:280–312
56. Milgrom PR (1981) Good news and bad news: Representation theorems. Bell J Econ 21:380–391
57. Milgrom PR, Shannon C (1994) Monotone comparative statics. Econometrica 62(1):157–180
58. Modigliani F, Miller MH (1958) The cost of capital, corporation finance and the theory of investment. Am Econ Rev 48(3):261–297
59. Morris S (2001) Political correctness. J Political Econ 109:231–265
60. Nowak MA, Krakauer DC (1999) The evolution of language. Proc Natl Acad Sci 96(14):8028–8033
61. Olszewski W (2004) Informal communication. J Econ Theory 117:180–200
62. Rabin M (1990) Communication between rational agents. J Econ Theory 51:144–170
63. Rabin M, Farrell J (1996) Cheap talk. J Econ Perspectives 10(3):103–118
64. Riley JG (2001) Silver signals: Twenty-five years of screening and signaling. J Econ Lit 39(2):432–478
65. Rothschild M, Stiglitz J (1976) Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. Q J Econ 90(4):629–649
66. Rubinstein A (2000) Economics and language. Cambridge University Press, New York
67. Sally D (2005) Can I say, bobobo and mean, there's no such thing as cheap talk? J Econ Behav Organ 57(3):245–266
68. Seidmann DJ, Winter E (1997) Strategic information transmission with verifiable messages. Econometrica 65(1):163–169
69. Siller S (1998) A note on errors in grafen's strategic handicap models. J Theor Biol 195:413–417
70. Sobel J (1985) A theory of credibility. Rev Econ Stud 52(4): 557–573
71. Sobel J, Takahashi I (1983) A multistage model of bargaining. Rev Econ Stud 50(3):411–426
72. Spence AM (1974) Market Signaling. Harvard University Press, Cambridge
73. Trivers RL (1971) The evolution of reciprocal altruism. Q Rev Biol 46:35–58
74. Vida P (2006) Long Pre-Play Communication in Games. Ph D thesis, Autonomous University of Barcelona
75. Wilson C (1977) A model of insurance markets with incomplete information. J Econ Theory 16:167–207
76. Zahavi A (1975) Mate selection- a selection for a handicap. J Theor Biol 53:205–214

Books and Reviews

Admati A, Perry M (1987) Strategic delay in bargaining. Rev Econ Stud 54:345–364
Austen-Smith D (1990) Information transmission in debate. Am J Political Sci 34(1):124–152
Battigalli P, Siniscalchi M (2002) Strong belief and forward induction reasoning. J Econ Theory 106(2):356–391
Bernheim BD (1994) A theory of conformity. J Political Econ 102(5):841–877
Blume A, Kim Y-G, Sobel J (1993) Evolutionary stability in games of communication. Games Econ Behav 5:547–575
Fudenberg D, Tirole J (1991) Game Theory. MIT Press, Cambridge
Gibbons R (1992) Game Theory for Applied Economists. Princeton University Press, Princeton
Kartik N, Ottaviani M, Squintani F (2007) Credulity, lies, and costly talk. J Econ Theory 134(1):93–116
Krishna V, Morgan J (2001) A model of expertise. Q J Econ 116(2): 747–775
Lo P-Y (2006) Common knowledge of language and iterative admissibility in a sender-receiver game. Technical report. Brown University
Manelli AM (1996) Cheap talk and sequential equilibria in signaling games. Econometrica 64(4):917–942
Milgrom P, Roberts J (1986) Price and advertising signals of product quality. J Political Econ 94(4):796–821
Noldeke G, Samuelson L (1997) A dynamic model of equilibrium selection in signaling markets. J Econ Theory 73(1):118–156
Noldeke G, Van Damme E (1990) Signalling in a dynamic labour market. Rev Econ Stud 57(1):1–23
Ottaviani M, Sørensen PN (2006) Professional advice. J Econ Theory 126(1):120–142
Rabin M (1994) A model of pre-game communication. J Econ Theory 63(2):370–391
Ramey G (1996) D1 signaling equilibria with multiple signals and a continuum of types. J Econ Theory 69(2):508–531
Rasmusen EB (2006) Games and Information: An Introduction to Game Theory, 4th edn. Blackwell, New York
Riley JG (1979) Informational equilibrium. Econometrica 47:331–359
Sobel J, Stole L, Zapater I (1990) Fixed-equilibrium rationalizability in signaling games. J Econ Theory 52(2):304–331
Spence AM (1973) Job market signaling. Q J Econ 90:225–243
Swinkels JM (1999) Education signallling with preemptive offers. Rev Econ Stud 66(4):949–970

Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface

Albert A. M. Holtslag, Gert-Jan Steeneveld
Department of Meteorology and Air Quality, Wageningen University, Wageningen, The Netherlands

Article Outline

Glossary
Definition of the Subject
Introduction
Background
Atmospheric Boundary-Layer Structure
Modeling Basics
Local Mixing Parameterization
More Advanced Mixing Parameterizations
Intercomparison of Single Column Models for Stable Conditions
Modeling Boundary Layers over Land
Impact of Land Surface Conditions on Model Results
Summary
Acknowledgments
Bibliography

Glossary

Atmospheric boundary layer The Atmospheric Boundary Layer (ABL) is the lower part of the atmosphere which is directly influenced by the presence of the earth's surface. As such its major characteristics are turbulence and the diurnal cycle.

Diurnal cycle The depth of the dry atmospheric boundary layer (ABL) can vary over land between tens of meters during night up to kilometers during daytime (see Fig. 2). Over sea the depth is often typical a few hundred meters and rather constant on the time scale of a day.

Turbulence Turbulence in the atmospheric boundary layer is the three-dimensional, chaotic flow of air with time scales typically between a second and an hour. The corresponding length scales are from a millimeter up to the depth of the boundary layer (or more in the case of clouds). Turbulence in the ABL originates due to friction of the flow and heating (convection) at the surface.

Definition of the Subject

In this article we deal with the single column modeling of the Atmospheric Boundary layer (ABL) and the the complex interactions which may occur with the land surface. As such we review the major characteristics of the ABL over land, and summarize the basic parameterizations for the represesentation of atmospheric turbulence and the surface fluxes. The modeling principles are illustrated with the outcome of single-column models for a variety of conditions using field data and fine-scale model results. Our emphasis is on stable conditions which occur over land at night-time under clear skies. For readers not familiar with atmospheric turbulence and meteorological definitions, some background and basic definitions are also given.

Introduction

The Atmospheric Boundary Layer (ABL) is generally characterized by turbulence. Because of its capability to mix air with different properties efficiently, the representation of turbulence is directly relevant for atmospheric and environmental modeling. For instance, turbulence directly impacts on the transfer of momentum, sensible heat, water vapor, ozone, and methane, among many other quantities, between the earth's surface and the atmosphere. Turbulence also defines the mixing of properties inside the atmospheric boundary layer, the transfer of quantities between the boundary layer and the clear or cloudy atmosphere aloft, and the mixing inside clouds.

Turbulence in the ABL is mainly due to the mechanical turbulence by vertical wind shear and turbulence by convection. Most of the atmosphere above the ABL is not turbulent, although turbulence can occur throughout the whole atmosphere. For instance, cumulus-type clouds, which may grow into thunderstorms, are always turbulent through convection produced by the heat released due to the condensation of water vapor. Turbulence can also occur in clear air above the ABL; most of this is produced in layers of strong vertical wind shear at the boundary between air masses (so-called 'Clear-Air Turbulence').

Because of the mixing capacity of turbulence, modeling atmospheric boundary layers is also relevant for many practical applications. For instance, chimney plumes are diluted and spread over larger volumes than they would be without turbulence. As such, strong local peaks of pollution are prevented and otherwise clean air is polluted. In practice turbulence may also cause engineering problems, because it shakes structures such as bridges, towers, and airplanes, causing failure of such systems in extreme cases. Turbulent fluctuations in the horizontal motions during severe storms can be fatal to tall buildings or bridges, particularly if resonance (e. g., forcing of a system at its natural frequency) occurs.

The correct formulation of the overall effects by turbulence, either inside or outside the atmospheric boundary layer, is an essential part of atmospheric models dealing with the prediction and study of weather, climate and air quality. These models are based on solving the equations dealing with atmosphere behavior. With state-of-the-art computers, the number of grid points in atmospheric models is limited to a number of typically 10^8 or

so. This implies that on the regional and global scale the atmospheric model equations are usually applied too fairly large 'air boxes'. Such boxes are often in the order of ten to hundred kilometers wide and ten to a few hundred meters thick. In these large boxes, smaller scale motions make air parcels interact and mix. For example, if a hot parcel is located next to a cold parcel, turbulent motion at their boundaries will heat the cool and cool the hot parcel. Thus, a closure formulation is needed to reproduce mixing by the turbulent motions into the model-resolved scales using the equations for the larger-scale 'mean' motions. It is important to realize that the closure formulation needs to be expressed in terms of variables available in the modeling context. This is called a 'parameterization'.

In this contribution we provide an overview of the modeling principles, the turbulent closures and parameterizations in use for of the atmospheric boundary layer, where we emphasize the modeling and parameterization of turbulence in the atmospheric boundary layer without clouds. Additionally we discuss the performance of models in current use [7,28], and we study the impact of the surface boundary condition over land [18].

Background

Atmospheric models for the forecasting and study of weather, climate, and air quality are typically based on integration of the basic equations governing atmospheric behavior. These equations are the gas law, the equation of continuity (mass), the first law of thermodynamics (heat), the conservation equations for momentum (the so-called "Navier–Stokes equations"), and usually equations expressing the conservation of moisture, trace gases and air pollutants. At one extreme, atmospheric models may deal with the world's climate and climate change; at the other, they may account for the behavior of local flows at coasts, in mountain-valley areas, or even deal with individual clouds. This all depends on the selected horizontal modeling domain and the available computing resources.

Since there is an enormous range of scales in atmospheric motion and turbulence, there is a need to separate the scales of atmospheric turbulence from larger-scale motions. Let C denote an atmospheric variable, such as specific humidity. Then $\overline{C}$ represents a mean or "smoothed" value of C, typically taken on a horizontal scale of order 10 (or more) km and a corresponding time scale in the order of 10 min to one hour. A local or instantaneous value of C would differ from $\overline{C}$. Thus, we have

$$C = \overline{C} + c\,. \tag{1}$$

Here c represents the smaller-scale fluctuations. Note that we use lower case for the latter (often primes are used as well to indicate fluctuations). In principle, the fluctuations around the mean motion also reflect gravity waves and other smaller scale motions, in addition to turbulence. Gravity waves often co-exist with turbulence or are generated by turbulence. If the wind at the same time is weak, there may be no turbulence at all. Anyhow, if turbulence exists, it is usually more important for most atmospheric applications, because it mixes more efficiently than the other small-scale motions.

To make the mathematical handling of c tractable, it must satisfy the so-called "Reynolds postulates". These require, for example, that $\overline{c} = 0$ and that small- and larger-scale values must not be correlated. After a quantity has been averaged to create a larger-scale quantity, further averaging should produce no further changes, in order for this postulate to apply. The mean of the summation of two variables A and C will produce $\overline{A \pm C} = \overline{A} \pm \overline{C}$. A further condition is that a mean variable $\overline{C}$ must be differentiable, since differentials show up in the atmospheric equations (see below). In practice, not all these conditions are rigorously satisfied. If the Reynolds postulates are fulfilled, then the averaging for the product of two variables provides

$$\overline{AC} = \overline{A}\,\overline{C} + \overline{ac}\,. \tag{2}$$

The second term at the right hand side of Eq. (2) is known as the turbulent covariance. Similarly, the turbulence variance of a quantity is given by $\overline{C^2} - (\overline{C})^2$ (which is the square of the standard deviation).

If in Eq. (2), the variable A represents one of the velocity components (U, V, W in the x, y, z direction, respectively), then $\overline{AC}$ is the total flux of C and the second term at the right hand side of Eq. (2) represents a turbulent flux of C. For instance, $\overline{uc}$ and $\overline{wc}$ are the horizontal and vertical turbulent fluxes of the variable C, respectively. Here u and w are the turbulent fluctuations of the horizontal and vertical velocities. Near the surface, the mean vertical wind $\overline{W}$ is usually small, and thus the total vertical fluxes are normally dominated by the turbulent contributions.

Atmospheric Boundary-Layer Structure

Turbulent fluctuations, variances and fluxes of variables are influenced by the vertical boundary-layer structure. Here the variation of temperature in the atmospheric boundary-layer plays an important role. Since pressure decreases with altitude, air parcels, which are forced to rise (sink), do expand (compress). According to the first law of thermodynamics, a rising (sinking) parcel will cool

(warm) if there is no additional energy source such as condensation of water vapor. Then this is called a dry adiabatic process.

It can be shown that in the atmospheric boundary layer, the temperature (T) variation with height for a dry adiabatic process is $\mathrm{d}T/\mathrm{d}z = -g/C_\mathrm{p}$ (here g is gravity constant and C_p is specific heat at constant pressure). The value for g/C_p is approximately 1 K per 100 m. An atmospheric layer which has such a temperature variation with height, is called neutral for dry air (at least when there is no convection arising from other levels). In that case $\Theta = T + (g/C_\mathrm{p})z$ is constant, where Θ is called the potential temperature (Note that the previous definition for potential temperature is not accurate above the boundary layer). Since air normally contains water vapor and because moist air is lighter than dry air, we have to correct for the influence of this on vertical motions. Consequently, a virtual potential temperature is defined as $\Theta_\mathrm{v} = \Theta(1 + 0.61q)$, where q is the specific humidity (defined as the mass of water vapor per unit mass of moist air).

In a neutral layer with constant Θ_v, vertical motions of moist (not saturated) air can maintain themselves. If the virtual potential temperature of the atmospheric layer increases with height, vertical displacements are suppressed. This is called a stable condition (or 'inversion'). At the other hand, when the virtual potential temperature decreases with height, vertical fluctuations may be accelerated. Consequently this is called an unstable condition. Thus in considerations with turbulent fluctuations and atmospheric stability, we have to deal with the virtual potential temperature and not with the actual temperature. Similarly, the vertical flux of sensible heat is connected to turbulent fluctuations of (virtual) potential temperature; e. g. it reads as $\overline{w\theta}_\mathrm{v}$ (in m K/s). The latter relates directly to the energy per time and unit area H by $H = \rho C_\mathrm{p}\overline{w\theta}_\mathrm{v}$ (in W/m^2), where ρ is density of the air (in kg/m^3).

Figure 1 (after [30]), provides the typical, idealized, mean vertical profiles for temperature T, potential temperature Θ, specific humidity q, in addition to the horizontal wind M (defined by $M^2 = U^2 + V^2$). These profiles apply for an atmospheric boundary layer over land in clear sky conditions in the afternoon and around midnight. Note that in the free atmosphere the horizontal wind is mostly a result of the acting of the larger scale pressure differences and the Coriolis force due to the rotation of the earth (but other effects may play a role as well). The resulting wind is known as the 'geostrophic' wind and indicated with G in Fig. 1 (see dashed line). In the daytime boundary layer the actual wind is smaller due to surface friction, while at clear nights the actual wind away from the surface may be substantially stronger than G due to inertial effects (resulting in the so-called 'low level jet').

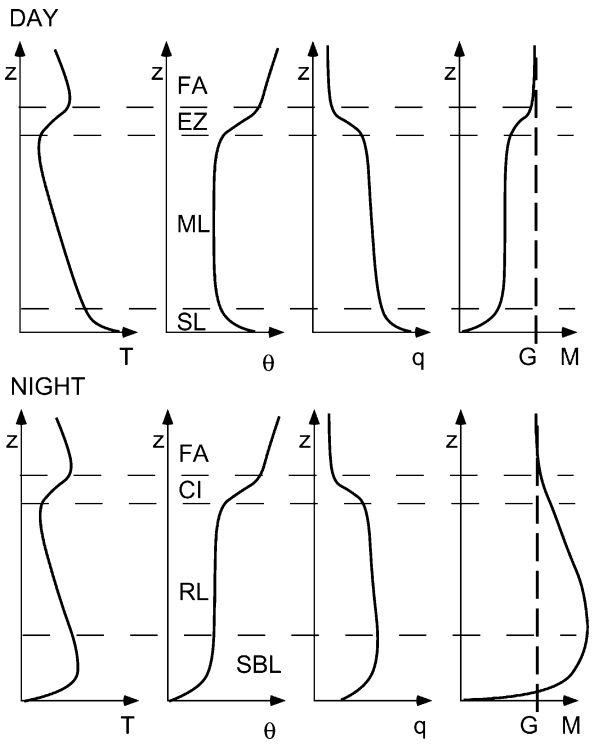

Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface, Figure 1
Idealized vertical profiles of mean variables in the Atmospheric Boundary Layer over land in fair weather (after [30]). See text for additional information

The temporal variation of the mean boundary-layer profiles over land can be quite substantial due to the strong diurnal variation of solar incoming radiation and the nighttime cooling at the land surface. During daytime the turbulent boundary layer may grow to several kilometers into the non-turbulent 'free atmosphere' (indicted as FA in Fig. 1). At night the turbulent part of the stable boundary layer (SBL) may only extend up to a few hundred meters or less (the lowest dashed line in the lower figure). An idealized picture for the temporal variation of the boundary layer over land is given in Fig. 2 (after [30]). Here the arrows with local time indications refer to the day and nighttime figures of Fig. 1.

Figure 1 also indicates that the boundary layer during daytime shows a three-layer structure: an unstable 'surface layer (SL)', a 'well-mixed layer (ML)' with rather uniform (virtual) potential temperature, and a stably-stratified 'entrainment zone (EZ)'. In the latter zone, turbulence acts to exchange heat, momentum, water-vapor and trace gasses

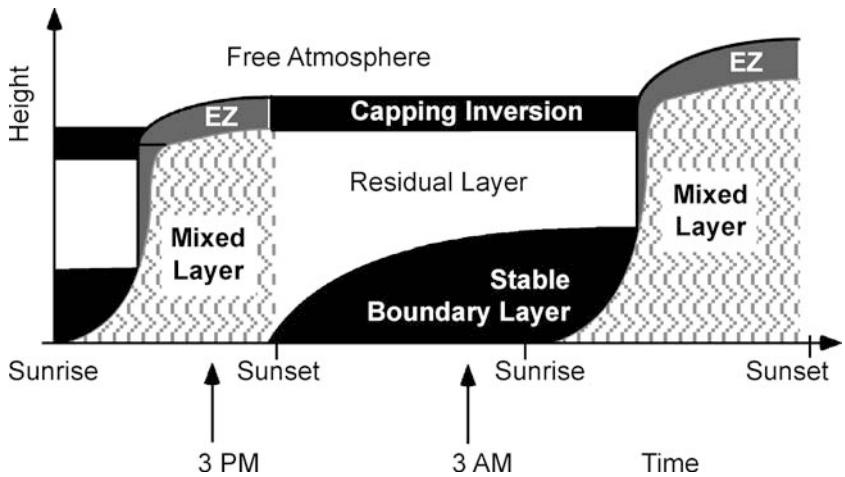

Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface, Figure 2
Idealized diurnal evolution of the Atmospheric Boundary Layer over land in fair weather (after [30])

between the boundary layer and the free atmosphere. During nighttime, often the vertical structure of the previous day persists above the SBL. As such a 'residual layer (RL)' with sporadic turbulence (remaining from the previous day) can be identified as well as a 'capping inversion (CI)'.

Modeling Basics

The challenge of modeling the atmospheric boundary layer is the prediction of the temporal variation of the vertical and horizontal structures in response to the influence of the major processes acting in the atmosphere and at the earth's surface. As such the governing equations have to be integrated. In practice, the variables are split into 'mean' larger-scale motions and smaller-scale fluctuations as in Eq. (1). Inserting this into the basic equations and after averaging this provides a set of equations for the behavior of the larger-scale (mean) variables. The larger-scale variables are then used explicitly in atmospheric models. This can be demonstrated as follows below.

The general character of any of the budget equations dealing with atmospheric motions is

$$\frac{\mathrm{D}C}{\mathrm{D}t} = S_i \,. \tag{3a}$$

Here S_i represents the subsequent sources and sinks for the variable C (such as radiation or chemistry effects). The notation $\mathrm{D}C/\mathrm{D}t$ represents the total rate of change for the variable C by local changes $(\partial/\partial t)$, and changes transported with the fluid motion in the three directions. As such, we have

$$\frac{\partial C}{\partial t} + U\frac{\partial C}{\partial x} + V\frac{\partial C}{\partial y} + W\frac{\partial C}{\partial z} = S_i \,. \tag{3b}$$

Here U, V, W are the wind speeds in the three directions x, y, z, respectively.

If in the atmospheric motion each variable is split into a mean component and a fluctuation then (3b) provides after Reynolds-averaging, some algebraic manipulations and simplifying assumptions, a budget equation for the mean variable $\overline{C}$. This reads as

$$\begin{aligned}\frac{\mathrm{D}\overline{C}}{\mathrm{D}t} &= \frac{\partial \overline{C}}{\partial t} + \overline{U}\frac{\partial \overline{C}}{\partial x} + \overline{V}\frac{\partial \overline{C}}{\partial y} + \overline{W}\frac{\partial \overline{C}}{\partial z} \\ &= \overline{S_i} - \frac{\partial \overline{uc}}{\partial x} - \frac{\partial \overline{vc}}{\partial y} - \frac{\partial \overline{wc}}{\partial z} \,.\end{aligned} \tag{4}$$

We may note that in the derivation of (4), single terms representing fluctuations have disappeared (as above in Eq. (2)). However, terms involving the product of two fluctuations did remain.

Thus because the basic equations are nonlinear, the budget equations for the mean variables contain terms involving smaller-scale motions. The latter terms are of the form of a divergence of fluxes produced by such motions in the three directions and appear as the last three terms in Eq. (4). These motions are said to be sub-grid and consequently, closure formulations or parameterizations are needed to introduce mixing by the smaller-scale, sub-grid, motions into the equations for the larger-scale motions (as resolved by the model). Note that additional terms may also appear in (4) when the source or sink term S_i incorporates nonlinear effects (such as in the case of chemistry).

The atmospheric model equations can also be applied on much smaller spatial and temporal scales then discussed here, for instance by using vertical and horizontal grid elements of 10 to 100 m, and time steps of seconds

only. It is important to realize that in such cases a significant part of the turbulent fluctuations are resolved by the model equations. This type of modeling is nowadays known as 'Large-eddy simulation (LES)'. This has become a powerful and popular tool in the last decade to study turbulence in clear and cloudy boundary layers under well-defined conditions. It is important to realize that in the case of LES the simplifying assumptions leading to Eq. (2) are normally not valid.

A special and simple form of Eq. (4) arises for horizontally homogeneous conditions. In such cases the terms including horizontal derivatives are negligible. If in addition the mean vertical wind is small and if there are no other sources and sinks, then (4) provides

$$\frac{\partial \overline{C}}{\partial t} = -\frac{\partial \overline{wc}}{\partial z} \,. \tag{5}$$

This equation is known as the one-dimensional, vertical diffusion equation. It shows that the local time rate of change for the mean of a variable (such as temperature or wind) at a certain height, is given by the divergence of the turbulent (corresponding heat or momentum) flux in the vertical direction. As such, information on the turbulent flux may produce a local forecast of the variation of a mean variable (but only under the simplifications mentioned).

Equation (5) can be seen as the basis of a single-column model where only local information of the atmosphere is relevant. However, normally the other terms in (4) are also relevant, in particular the terms with mean wind speed (the so-called "advection terms"). This means that in general the budget equations for momentum, heat, and the various scalars are closely coupled in any atmospheric model. Still one can solve for the local time rate of change in a single column model once the advection terms are known form observations or other means. This approach is widely adopted to study atmospheric boundary layers in comparison with observations on a local scale.

Before we proceed with more detailed parameterizations for the fluxes in the boundary layer, let us deal with the derivation of the surface fluxes. These fluxes enter as boundary conditions when solving the budget equations for all the relevant mean variables (in any approach). It is important to realize that near the surface, the average wind must vanish because the mean wind is zero at the earth's surface. At the other hand, we know from observations that the fluxes of heat, momentum and trace gasses are nonzero. Consequently, it is convenient to model an 'effective' surface flux $\overline{wc_0}$ of a conserved variable due to the combined effect of molecular diffusion and turbulence at the surface. This can be achieved by writing

$$\overline{wc_0} = \beta_t w_t (C_0 - C_a) \,. \tag{6}$$

Here C_0, and C_a are the values of the transported variable at the surface and in the air, respectively; β_t is a transfer coefficient, and w_t is an effective transport velocity representing the turbulence. For example, in near-neutral conditions the effective transport velocity is well represented by the well-known surface friction velocity u_{*0}. Then it can be shown that $\beta_t = \kappa/\ln(z/z_0)$, where κ is the 'Von Karman' constant (often specified as $\kappa \cong 0.4$), z is the corresponding height of C_a in the lowest part of the boundary layer and z_0 is the so-called surface roughness length for the variable C. We refer to the literature for a more detailed treatment (e. g., Beljaars and Holtslag [4]).

Local Mixing Parameterization

To solve the budget Eq. (4) for all the mean atmospheric variables involved, the terms involving turbulent fluxes need to be parametrized. As mentioned before, this means that the fluxes need to be expressed in terms of available mean model quantities, both in the atmosphere and at the surface. Once this has been achieved, the atmospheric model equations can be integrated. Thus, starting with proper initial values, new values can be calculated for the following time step and so on.

The most frequently used parameterization for environmental and atmospheric models, is known as first-order closure or often also called K-theory. In this theory it is assumed that the flux $\overline{wc}$ of a variable C in the vertical direction z, is down the vertical gradient of the mean concentration of C per unit mass. Thus

$$\overline{wc} = -K_c \frac{\partial \overline{C}}{\partial z} \,. \tag{7}$$

Here, K_c is known as the 'eddy-diffusivity' or mixing coefficient for the variable C. Similarly, the horizontal fluxes can be represented in terms of horizontal gradients. Note that the corresponding eddy-diffusivities typically are not constant, but that they generally depend on properties of the flow and the variable of interest. This also means that normally no analytic solutions are possible, not even for the simple case in which Eqs. (5) and (8) are combined.

We may note that the dimension of an eddy-diffusivity is a length scale ℓ times a velocity scale. These are proportional to the products of effective eddy sizes and eddy velocities in the corresponding directions. Often a diagnostic expression is used for the eddy-diffusivity, on basis of what is called 'mixing length theory' (in analogy with molecular

diffusion). The result reads as

$$K_c = \ell^2 S f(\mathrm{Ri}) \,. \tag{8}$$

Here S is vertical wind shear (that is the variation of mean horizontal wind with height). Note that the combination ℓS in Eq. (9) has units of velocity. In Eq. (8), $f(\mathrm{Ri})$ denotes a functional dependence on local stability as represented by the gradient Richardson-number Ri defined by

$$\mathrm{Ri} = \frac{g}{\overline{\Theta_\mathrm{v}}} \frac{\partial \overline{\Theta_\mathrm{v}}/\partial z}{(\partial \overline{U}/\partial z)^2 + (\partial \overline{V}/\partial z)^2} \,. \tag{9}$$

Here g is the acceleration due to gravity, and $\overline{\Theta_\mathrm{v}}$ is the mean 'virtual potential temperature'.

The specification of the length scale ℓ is not at all straightforward, except near the surface where so-called 'surface-layer similarity theory' (see cited literature) provides that $\ell \propto z$. A frequently used form for ℓ is:

$$\frac{1}{\ell} = \frac{1}{\kappa z} + \frac{1}{\lambda} \,. \tag{10}$$

Here λ is a turbulent length scale, which should be valid for the turbulence far above the surface. We note that the latter has a rather empirical nature and consequently there is no agreement on the specification of λ in the literature.

Equation (8) is a diagnostic equation, which indicates that the eddy-diffusivity may variy with height, wind speed, stability, et cetera. In combination with the flux parameterization of Eq. (7), it follows that the flux at a certain height depends on the local gradient of the mean variable involved. Consequently the approach is referred to as a 'diagnostic local mixing approach'. Such an approach is mostly suitable for relatively homogeneous conditions with neutral and stable stratification, and is not so suitable for cases with convection (see non-local mixing parameterizations below).

More Advanced Mixing Parameterizations

A physically realistic alternative to the diagnostic approach is to relate the eddy-diffusivity of Eq. (7) to the actual turbulent kinetic energy of the flow, by using the prognostic turbulent kinetic energy equation and an appropriate choice for the turbulent length scale. It is important to realize that the kinetic energy of atmospheric motion per unit of mass E is given by the half of the sum of the velocities squared in the three directions (as in classic mechanics), e. g. $E = (U^2 + V^2 + W^2)/2$. Similar as with respect to Eq. (2), we can separate between the Mean Kinetic Energy $\overline{E}$ of the mean atmospheric motions and the Turbulent Kinetic Energy (TKE or e) of the smaller-scale fluctuating motions by turbulence. Thus e is given by $e = (\overline{u^2} + \overline{v^2} + \overline{w^2})/2$.

The prognostic equation for e reads in its basic form as:

$$\frac{\mathrm{D}e}{\mathrm{D}t} = -\overline{uw}\frac{\partial \overline{U}}{\partial z} - \overline{vw}\frac{\partial \overline{V}}{\partial z} + \frac{g}{\overline{\Theta_\mathrm{v}}}\overline{w\theta_\mathrm{v}} + D - \varepsilon \,. \tag{11}$$

Here $\mathrm{D}e/\mathrm{D}t$ is the total variation of e with time (the sum of local variations and those transported with the mean air motion). The two terms at the immediate right hand side of (11) represent the shear production of turbulence. These depend primarily on vertical variations of wind or, near the ground, on wind speed and surface roughness. The terms are almost always positive. The third term in Eq. (11) represents the rate of production or breakdown of turbulence by buoyancy effects (such as heat convection). It depends directly on density effects, which can be written in terms of the virtual potential temperature $\overline{\Theta_\mathrm{v}}$, and its turbulent flux $\overline{w\theta_\mathrm{v}}$. The term D in Eq. (11) represents divergence and pressure redistribution terms. These have a tendency to cancel near the surface. Finally, the term ε reflects the molecular dissipation of turbulence into heat and this term is always positive. In fact ε is typically proportional to e/τ, where τ is the characteristic time scale for the turbulent mixing process.

Using Eq. (11), turbulent kinetic energy can be calculated for given mean profiles when the corresponding fluxes are calculated using Eq. (7) for all fluxes involved. In this approach the diffusivities are typically calculated with equations of the form

$$K_c = \alpha_\mathrm{c} \ell \sqrt{e} \,. \tag{12}$$

Here α_c is a constant depending on the variable of interest. The length scale is typically calculated with a similar type of diagnostic equation as (10) provides. This approach is known as the 'TKE-length scale approach' and it is an example of so-called 1.5 order closure. Sometimes a prognostic equation is used for the length scale as well, but such an approach is more popular in engineering applications then in the atmospheric sciences.

It can be shown that Eq. (8) is a solution of (11) and (12) in stationary conditions and when other simplifications are made such as the neglect of the influences by advection and turbulence divergence in the TKE equation. A more advanced turbulence scheme is known as 'second-order closure'. In such an approach, prognostic equations are developed for the fluxes and variances themselves. Such equations have a very similar structure as Eq. (12) for kinetic energy. Unfortunately, new unknowns are present in these equations. These must be related to the

other variables in the model equations, always involving assumptions. Thus, second-order closure involves many more than the original equations and is therefore computationally more time consuming ('expensive') than first-order and 1.5 order closure.

One may expect that a model with 1.5 or second-order closure would produce more realistic results then a model with a first order closure. However, in practice this is often not the case, because of complex model interactions and the difficulty of representing all the relevant details with sufficient accuracy (see also below). That is the reason why diagnostic approaches remain popular. Nevertheless, second order equations are useful to gain insight in the governing physics, and after simplification useful extensions of the basic parameterizations may be achieved.

We continue our discussion with mixing parameterizations which have been proposed for boundary-layers with strong atmospheric convection. In such cases, the turbulent flux of a conserved quantity is typically not proportional to the local gradient alone as predicted by Eq. (7). In fact, in a large part of the ABL the mean gradients are small in conditions with dry convection, in particular for potential temperature (see Fig. 1). Then the fluxes depend mostly on the mixing characteristics of the large eddies across the ABL. Theories are available, which have modified K-theory to allow for the influence of convection, for example by including additional terms at the right hand side of Eq. (8) For details we refer to the literature (e. g., Holtslag and Moeng [15]).

In the next sections we apply the modeling concepts above and compare their results with field observations. In addition we present results from model intercomparison studies, and illustrate the role of boundary conditions.

Intercomparison of Single Column Models for Stable Conditions

Atmospheric models for weather and climate need to make an overall representation of the smaller-scale boundary-layer and near surface processes. This appears to be more successful during daytime (e. g. Ek and Holtslag [10]; Holtslag and Ek [17]) then during nighttime stable conditions over land. The modeling of the stable boundary layer over land is rather complex because of the many different physical processes which are "at work" in stable conditions [21]. These small-scale processes are: clear air radiation divergence, drainage flow, generation of gravity waves and shear instabilities, fog and dew formation, the occurrence of a low-level jet and generation of discontinuous or intermittent turbulence [33]. In addition, the phenomenology of stable atmospheric boundary layers is quite divers, e. g. shallow and deep boundary layers with continuous turbulence through most of their depth, and on the other hand boundary layers with intermittent turbulence or even laminar flow.

The small-scale processes influence the vertical and horizontal exchange of quantities between the surface and the atmosphere as well as the mixing in the atmosphere on a variety of scales. In addition, it is known that turbulent mixing in stratified flow has an inherent non-linear character and may, as such, trigger positive feedbacks. These positive feedbacks, in turn, may cause unexpected transitions between totally different SBL regimes (e. g. van de Wiel et al. [34]).

Figure 3 depicts the interactions between relevant processes in the stable boundary layer. The non-linear behavior of the system is seen in e. g. the surface sensible heat-flux (H). A sudden change of the surface temperature can result in 2 different impacts on H. First, in weakly stable conditions (with strong wind and sufficient turbulence), a surface temperature decrease will provide a larger heat flux since H is proportional to the temperature difference between the surface and the atmosphere. The larger heat-flux from the atmosphere to the surface compsensates for the stronger cooling. In contrast for stronger stably stratified conditions, a surface temperature decrease will provide a stronger stratification and inhibits turbulent mixing, and consequently H will decrease. This allows for even stronger surface cooling (positive feedback). Note that a similar diagram for the daytime boundary layer can be found in Ek and Holtslag [10].

Having in mind the above mentioned complexity, one should not be surprised that atmospheric models encounter large forecast errors for stable conditions [20,24]. One strategy to improve model performance is to provide different models the same forecasting task, and analyze which model descriptions are in favor for which atmospheric stability. Recently such an intercomparison of boundary-layer schemes for stable conditions was made within the GEWEX Atmospheric Boundary Layer Study ('GABLS'). This GEWEX project aims to improve the understanding and the representation of the atmospheric boundary layer in regional and large-scale climate models [14]. A rather simple case was selected as a benchmark to review the state of the art and to compare the skills of single column (1D) models [7] and Large–Eddy Simulation models [3]. In this case a stable boundary layer is driven by an imposed, uniform geostrophic wind, with a specified constant surface-cooling rate over (homogeneous) ice. The case is initialized with $\theta = 265\,\mathrm{K}$ for $0 < z < 100$ and a lapse rate of 1 K/100 m aloft.

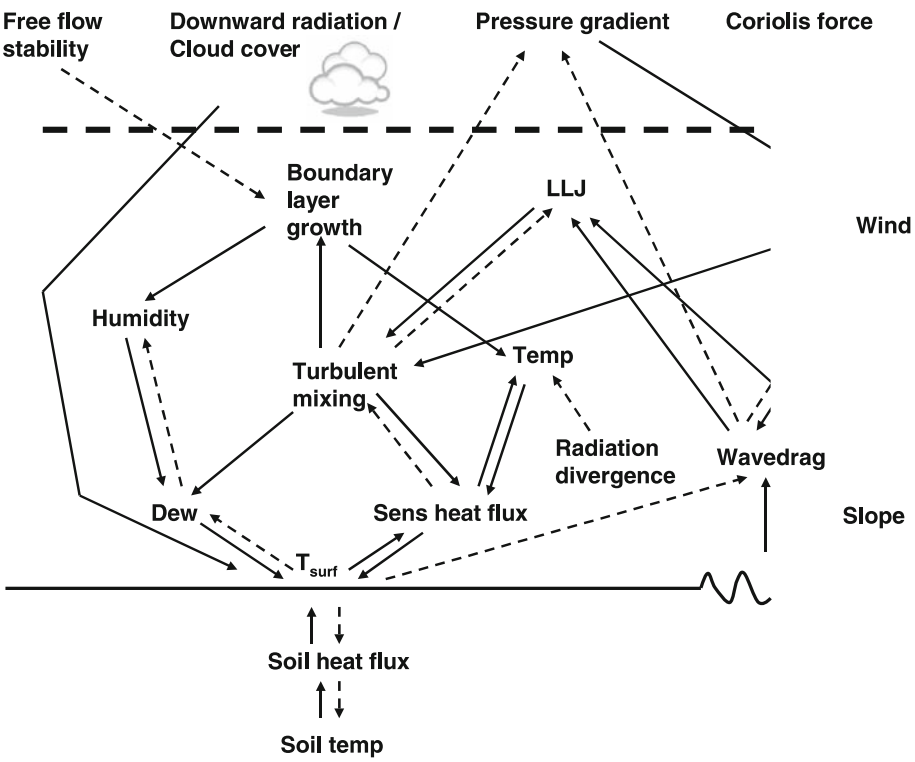

Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface, Figure 3
Interaction diagram for the processes and variables in the stable atmospheric boundary layer over land (after [26])

It turns out that with the same initial conditions and model forcings, the models indicate a large range of results for the mean temperature and wind profiles. Figure 4 shows the mean profiles for several models after nine hours of constant surface cooling (sufficient to achieve a quasi-steady state). The variable results achieved are strongly related to the details of the boundary-layer mixing schemes [7]. An important finding is that the models in use at operational weather forecast and climate centers (as depicted at the left hand side of Fig. 4) typically allow for enhanced mixing resulting in too deep boundary layers, while the typical research models (at the right hand sides) show less mixing in more in agreement with the 'Large Eddy Simulation' results for this case [3].

Because of the enhanced mixing in weather and climate models, these models tend to show a too strong surface drag, too deep boundary layers, and an underestimation of the wind turning in the lower atmosphere [19]. At the other hand, by decreasing the mixing and surface drag, a direct impact on the atmospheric dynamics ('Ekman pumping') is noted (e. g. Beljaars and Viterbo [5]). Consequently, cyclones may become too active, corresponding in too high extremes for wind and precipitation, etc.

Modeling Boundary Layers over Land

To study the interactions of the ABL with the land surface we utilize the model by Duynkerke [9] with the extensions and modifications by Steeneveld et al. [28]. This model has been validated against tower observations at Cabauw, The Netherlands and later with CASES-99 field observations [1,28], and also participated in the GABLS model comparison in the previous section.

Instead of prescribing the surface temperature, and to enable interaction between the ABL and the land surface, the model is extended with a soil and a vegetation layer The soil temperature evolution is calculated by solving the diffusion equation (using a grid spacing of 1 cm) and the heat flux G_h from the soil to vegetation is calculated by:

$$G_h - (1 - f_{veg})K^{\downarrow} = r_g(T_{veg} - T_{s0}) \,. \tag{13}$$

In Eq. (13) $K^{\downarrow}$ is the incoming shortwave (solar) radiation, T_{veg} represents the vegetation surface temperature, and T_{s0} the soil temperature just below the vegetation. We use a vegetation fraction $f_{veg} = 0.9$ and conductance $r_g = 5.9\,\mathrm{Wm^{-2}K^{-1}}$, which are consistent with the obser-

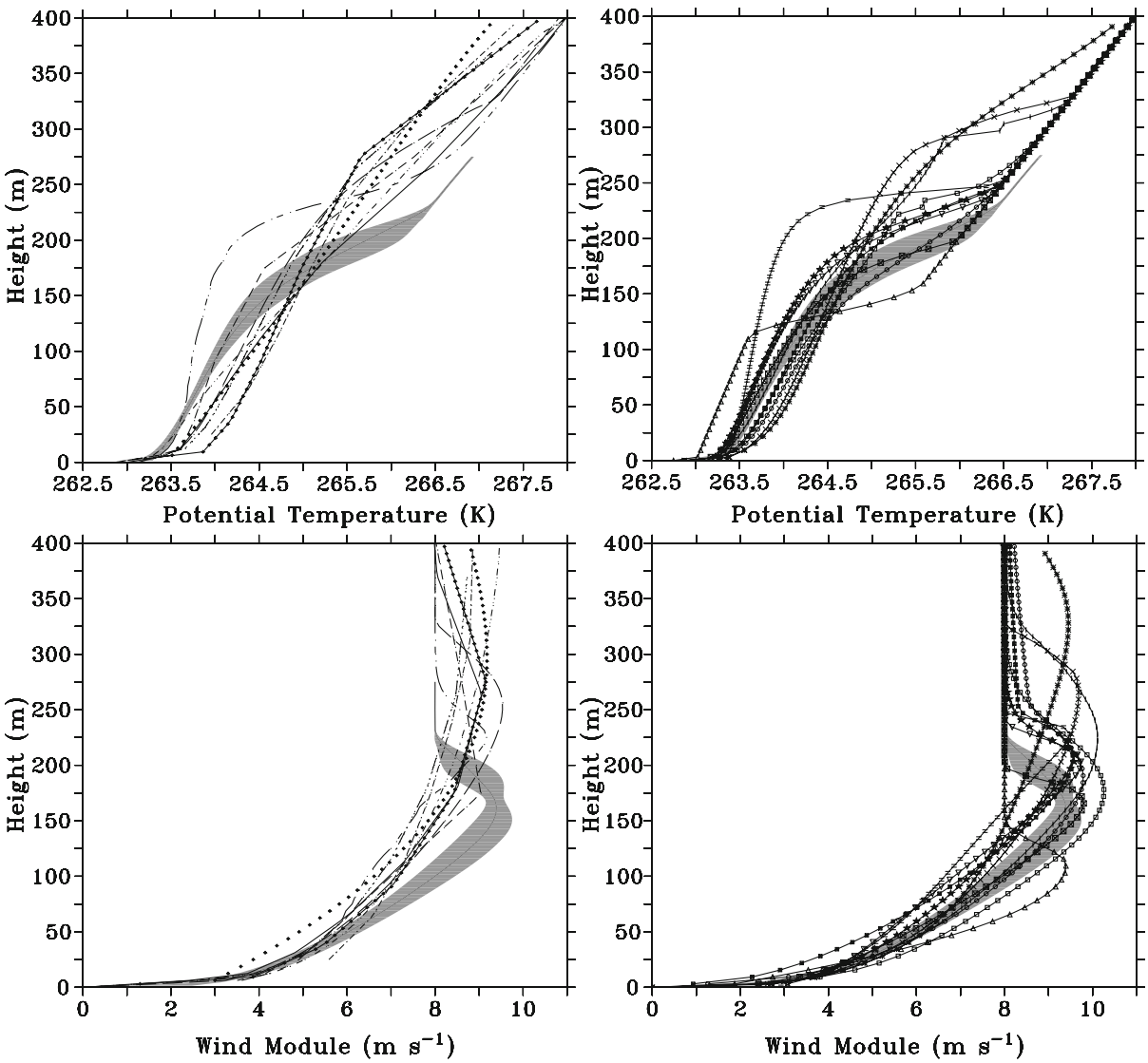

Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface, Figure 4
Modeled potential temperature and wind profiles by an ensemble of column models (after 9 h). *Grey areas* indicate the ensemble of Large Eddy Simulation results [3]. *Left panel* shows the results for first order closure models and the *right panel* for higher order closure models (after [7])

vations of CASES99 [28]. Initial soil and surface temperatures are also taken from the CASES99 observations.

Subsequently, the evolution of T_{veg} is computed by solving the surface energy budget for the vegetation layer:

$$C_{\mathrm{v}} \frac{\partial T_{\mathrm{veg}}}{\partial t} = Q^* - G_{\mathrm{h}} - H - L_{\mathrm{v}} E \,. \tag{14}$$

Here C_{v} is the heat capacity of the vegetation layer per unit of area ($C_{\mathrm{v}} = 2000\,\mathrm{Jm}^{-2}K^{-1}$, van de Wiel [34]), Q^* is the net radiation, H is the sensible heat flux and $L_{\mathrm{v}}E$ the latent heat flux. The turbulent fluxes are calculated basically with the format of Eq. (6) above. Finally, Q^* is calculated by adopting the Garratt and Brost [12] radiation scheme. Note that Eqs. (13) and (14) provide a rather strong coupling of the atmosphere to the vegetated land surface for the current parameter setting which is found to be important [28].

Model forecasts have been compared with CASES-99 observations for contrasting diurnal cycles (i. e. for different wind speeds) for 23–26 Oct. 1999. Note that the model

is only initialized once and that the total run comprises three full days. We distinguish between "radiative nights" with weak winds and small turbulent mixing, when radiative cooling dominates the SBL development. On the contrary, so-called "continuous turbulent" nights are characterized by strong winds. The final archetype is the so called "intermittently turbulent" night where turbulent episodes alternate with calm periods, when the radiative and turbulent forcings are of similar order of magnitude. Here we restrict ourselves to the results for surface fluxes, surface vegetation temperature, and vertical profiles of temperature and wind speed during nighttime. The diurnal cycle of the modeled net radiation Q^* (the balance of all incoming and outgoing short- and longwave radiative fluxes) agrees with the observations (not shown). Net radiation amounts typically $Q^* = 400\,\mathrm{Wm}^{-2}$ during daytime and $Q^* = -70\,\mathrm{Wm}^{-2}$ during nighttime.

The friction velocity (u_*) shows a clear diurnal cycle: u_* is large during the day and small at night, which is in general well captured by the model (Fig. 5a). Looking in more detail we find that during weak winds (1st and 3rd night) the model tends to overestimate u_*. The model lacks a clear turbulence collapse as observed during the first (intermittent) night. In the period 24 Oct., 700 CDT – 25 Oct., 1700 CDT the model performs well, while during the last (radiative) night u_* is slightly too high until midnight but follows the collapse at the end of the night. The overall bias amounts to $0.03\,\mathrm{ms}^{-1}$ for the last night. Sodar observations show much smaller wind speeds at 200 m AGL (which is above the SBL during this night) than the imposed **G**. This may suggest that **G** was overestimated, and this consequently may explain the bias in u_*. In general it is known that models correctly predict u_* for strong winds, but overestimate u_* for weak winds [25,32].

The sensible heat flux differs substantially between day ($\sim 250\,\mathrm{Wm}^{-2}$ here) and night (between 0 and –60 Wm^{-2} depending on the wind speed). In the first (intermittent) night, the modeled $H = -14.1\,\mathrm{Wm}^{-2}$ on average, while –9.1 Wm^{-2} was observed. However, the model does not simulate the observed intermittent character of the surface fluxes (Fig. 5b). Some models with more resolution [23,36] were also able to reproduce intermittent turbulence. On the other hand, the models by Sharan and Gopalakrishnan [25] and Derbyshire [8] did not show any intermittency. This subject needs further investigation.

Just after the day-night transition to the intermittent night, the observed magnitude of H shows a clear maximum (see arrows in Fig. 5b), which is well reproduced by the model. This maximum is caused by a sudden reversal of the stratification near the surface due to longwave radiation emission during the day-night transition,

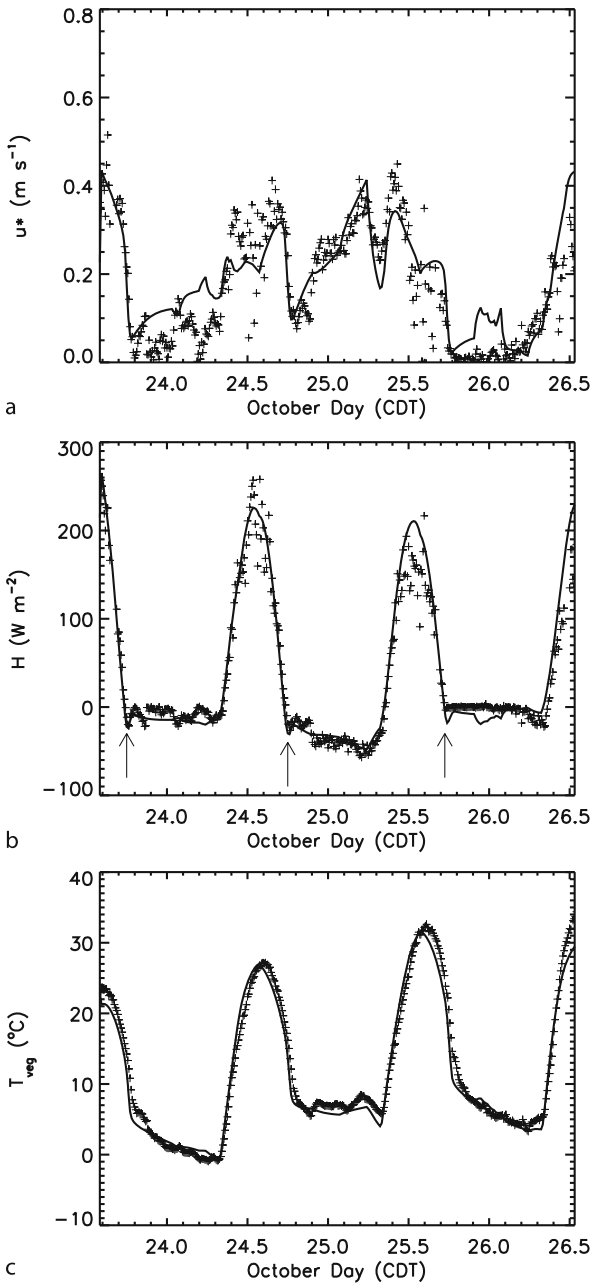

Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface, Figure 5
Modeled and observed (+) friction velocity (a), surface sensible heat flux (b), and surface vegetation temperature (c) for three diurnal cycles in CASES-99 (after [28])

and is maintained by residual turbulence of the convective boundary layer. This is an often observed *realistic feature* (e. g. during 11 of the 30 nights for CASES-99 and in FIFE observations shown by [6]). However, most model-

ing studies rarely show these minima. The reproduction of this detailed feature emphasizes the realism of the model outcome.

During the turbulent night (24–25 Oct.), the predicted H follows the observations. The specific minimum during the day-night transition (24 Oct., 1900 CDT) is present here as well. Radiative flux divergence dominates the last (radiative) night and the observed H is approximately zero. The model slightly overestimates the magnitude of H (–2.9 $\mathrm{Wm^{-2}}$), mainly caused by an overestimation of u_*. This causes a weaker stratification and thus a larger magnitude of H. The second half of this night the model gives good results.

Reliable prediction of T_{veg} is a common problem for large-scale models. Some models show unphysical decoupling of the atmosphere from the surface resulting in so-called "runaway cooling" of T_{veg}. On the other hand, the pragmatic enhanced mixing approach which is commonly used for very stable conditions, leads to overestimation of T_{veg}. For both day- and nighttime T_{veg} *is simulated in very good agreement with the data, despite the fact that we cover a broad range of stability* (Fig. 5c).

We conclude that the present model generates surface fluxes which are in good agreement with observations, because of the detail in the description of the surface scheme, the soil heat flux and radiation physics (with high resolution). In general, the model is also able to estimate temperature and wind profiles (see results and discussion in Steeneveld et al. [28]). To examine the robustness of the results, we performed some sensitivity analysis on the initial conditions and model parameters. Disturbing the initial temperature (by 1 K), wind profiles (by 5%), soil temperature and vegetation temperature (both by 1 K) do not affect the results seriously. Also model re-initialization every 24 h (1400 CDT) with observed radiosonde information showed hardly any impact on the results (not shown).

Impact of Land Surface Conditions on Model Results

Inspired by the result in the previous section that a coupling with the land surface is necessary to obtain satisfactory model results, we now analyze the difference in variability of model results during model intercomparisons, as function of the chosen boundary condition. At first we use the first-order closure model and vary the parameters in the turbulence scheme for stable conditions in a reasonable range to mimic the apparent variability among boundary-layer models. As such first model runs are performed with a prescribed surface temperature as inspired by (but not identical to) the observations in CASES99 [22] and as described in the GABLS2 case description [31]. Subsequently, the model runs are repeated, but then using an interactive prognostic heat budget equation for the surface temperature (Eq. (14)).

To study the impacts of parameter values on the model results, reference runs are made for coupled and uncoupled cases with alternative permutations in some of the parameter settings for stable conditions. The parameter modifications are chosen such that they cover a realistic range in comparison with existing models of the stable boundary layer [7]. The local starting time in the model runs is 14.00 LT on October 22, 1999 (rather than 16.00 LT in the GABLS2 runs). The duration of all runs is 59 hours (so that the axis of all the figures indicates 14.00 until 73.00 h, covering a period of 2.5 diurnal cycles). In all model runs the roughness length for heat z_{oh} and momentum z_{om} (3 mm and 3 cm respectively), and the canopy resistance are constant, and the geostrophic wind is taken at a reference value of 9.5 $\mathrm{ms^{-1}}$ (as in Svensson and Holtslag [31]). The reference model set up has 50 logarithmically distributed layers and the first atmospheric model level is at 2 m.

The model results for all parameter permutations are presented for the sensible heat flux (Fig. 6), In the upper sub-frame of the figure (labeled a) the results achieved with the uncoupled model are given (using prescribed surface temperature). Overall the variety of results in the upper frame is comparable to the variety within the GABLS2 intercomparison study in stable conditions for the uncoupled models (see Svensson and Holtslag [31]). Thus we have a range of –15 to –50 $\mathrm{Wm^{-2}}$ for the sensible heat flux (at the end of the first night e. g. at the time of 30 h). The variability is a result of the range of parameters chosen above and the impact is apparently sufficient to mimic the different parameterizations for stable conditions in the models used within GABLS2.

Next we repeat all model runs and allow for surface feedback using Eqs. (13) and (14). The results for the sensible heat flux with the coupled model are given in the lower frame (Fig. 6b). Now we have a range of –10 to –25 $\mathrm{Wm^{-2}}$ (again the values apply for the end of the first night e. g. at the time of 30 h). Thus it appears that the variety of model results is *smaller* for the sensible heat flux in the coupled case. At the same time it appears that the variability appears to be somewhat larger for friction velocity and boundary-layer depth, which seems to be related to the larger variability in the near surface air temperature and wind speed.

During daytime the sensible heat fluxes are rather similar for all model runs within one category (either coupled or uncoupled), but the maximum values differ. In addition, due to the coupling the sensible heat fluxes show

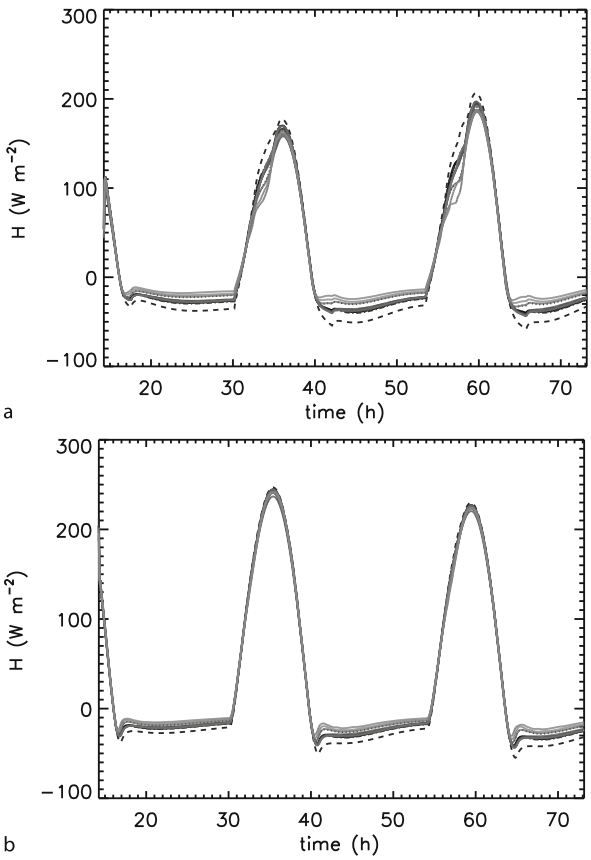

Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface, Figure 6
Time series of an ensemble of model results for the sensible heat flux in a model intercomparison study with a prescribed surface temperatures, and b by solving the surface energy budget (after [18])

a more smooth behavior in the morning hours as compared with the uncoupled results. Thus, surface feedback is influencing the model results and is also able to compensate for some variation in the model parameter values. Note also that the variability in the friction velocities of the first night remains during the morning hours in the uncoupled runs, but not so much in the coupled case.

In Fig. 7 the surface temperatures are given as specified for the uncoupled case (the dashed line), and the temperatures as calculated in the various interactive model runs (various gray lines). It is seen that the latter ones are quite different from each other (in particular at night). It is also important to note that the surface temperature by the ensemble of coupled model runs is clearly different from the specified temperature in the uncoupled case. This impacts also on the absolute values and the range of air temperatures and the wind speeds.

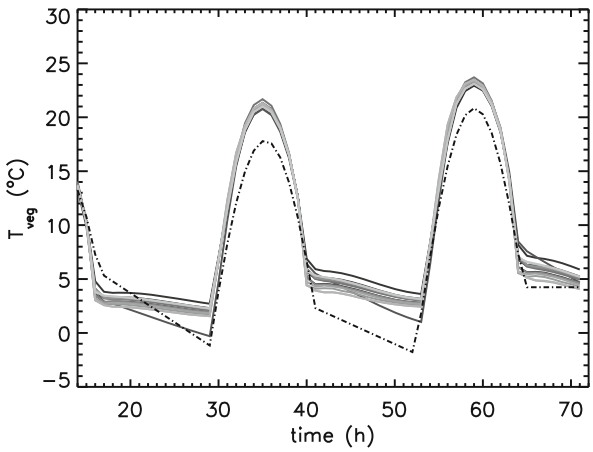

Single Column Modeling of Atmospheric Boundary Layers and the Complex Interactions with the Land Surface, Figure 7
Time series of an ensemble of modeled surface temperature for coupled runs. *Dash-dotted line*: prescribed surface temperature in the uncoupled case (after [18])

Thus it is apparent that the treatment of the surface temperature impacts strongly on the outcome of the boundary-layer model results and their variety (see also Basu et al. [2]). By repeating the uncoupled model runs with a specified surface temperature as given by the ensemble mean value of the interactive runs, we achieve basically the same variety of model outputs for the potential temperature and wind as for the coupled cases. This confirms that in model evaluation studies the surface temperature should be taken consistent with the value of the geostrophic wind (although this may be model dependent).

Summary

In this paper a summary is given of the basic approaches for the modeling and parameterization of turbulence in the atmospheric boundary layer. The treated approaches are in current use in regional and global-scale models for the forecasting and study of weather, climate and air quality. Here we have shown the results of such approaches by using single column models in comparison with field data and fine-scale model results. We have also studied the impact of the surface temperature condition on the variability of results by an atmospheric boundary-layer model. From the coupled model results we achieve that surface feedback can compensate for some of the variety introduced by changing model parameters. Generally much work needs still to be done before we have a full understanding of the complexity of atmospheric turbulence and the interactions with the land surface. A better understanding of atmo-

spheric turbulence hopefully also contributes to our capability in refining and unifying the turbulence parameterizations for modeling of the atmospheric boundary layer in response to the different type of surfaces which are found in reality.

Acknowledgments

This contribution is a compilation of earlier works by the authors, in particular the works by Holtslag [13], Steeneveld et al. [28] and Holtslag et al. [18].

Bibliography

1. Baas P, Steeneveld GJ, van de Wiel BJH, Holtslag AAM (2006) Exploring Self-correlation in flux-gradient relationships for stably stratified conditions. J Atmos Sci 63:3045–3054
2. Basu S, Holtslag AAM, van de Wiel BJH, Moene AF, Steeneveld GJ (2007) An inconvenienth 'truth' about using the sensible heatflux as a surface boundary condition in models under stably stratified regimes. Acta Geophys 56:88–99. doi:10.2478/s11600-007-0038-y
3. Beare R, MacVean M, Holtslag AAM, Cuxart J, Esau I, Golaz J-C, Jimenez M, Khairoutdinov M, Kosovic B, Lewellen D, Lund T, Lundquist J, McCabe A, Moene A, Noh Y, Raasch S, Sullivan P (2006) An intercomparison of Large–Eddy Simulations of the stable boundary layer. Bound-Layer Meteorol 118:247–272
4. Beljaars ACM, Holtslag AAM (1991) Flux parameterization over land surfaces for atmospheric models. J Appl Meteor 30:327–341
5. Beljaars ACM, Viterbo P (1998) Role of the boundary layer in a numerical weather prediction model. In: Holtslag AAM, Duynkerke PG (eds) Clear and Cloudy boundary layers. Royal Netherlands Academy of Arts and Sciences, Amsterdam, 372 pp
6. Chen F, Dudhia J (2001) Coupling an Advanced Land Surface-Hydrology Model with the Penn State-NCAR MM5 Modeling System. Part II: Preliminary Model Validation. Mon Wea Rev 129:587–604
7. Cuxart J, Holtslag AAM, Beare RJ, Bazile E, Beljaars A, Cheng A, Conangla L, Ek MB, Freedman F, Hamdi R, Kerstein A, Kitagawa H, Lenderink G, Lewellen D, Mailhot J, Mauritsen T, Perov V, Schayes G, Steeneveld GJ, Svensson G, Taylor P, Weng W, Wunsch S, Xu K-M (2006) Single-column model intercomparison for a stably stratified atmospheric boundary layer. Bound-Layer Meteorol 118: 273–303
8. Derbyshire SH (1999) Boundary layer decoupling over cold surfaces as a physical boundary instability. Bound-Layer Meteorol 90:297–325
9. Duynkerke PG (1991) Radiation fog: A comparison of model simulation with detailed observations. Mon Wea Rev 119:324–341
10. Ek MB, Holtslag AAM (2004) Influence of Soil Moisture on Boundary Layer Cloud Development. J Hydrometeor 5:86–99
11. Garratt JR (1992) The Atmospheric Boundary Layer. Cambridge University Press, New York, 316 pp
12. Garratt JR, Brost RA (1981) Radiative Cooling effects within and above the nocturnal boundary layer. J Atmos Sci 38:2730–2746
13. Holtslag AAM (2002) Atmospheric Boundary Layers: Modeling and Parameterization. In: Holton JR, Pyle J, Curry JA (eds) Encyclopedia of Atmospheric Sciences, vol 1. Academic Press, pp 253–261
14. Holtslag AAM (2006) GEWEX Atmospheric Boundary Layer Study (GABLS) on Stable Boundary Layers. Bound-Layer Meteorol 118:243–246
15. Holtslag AAM, Moeng C-H (1991) Eddy diffusivity and countergradient transport in the convective boundary layer. J Atmos Sci 48:1690–1698
16. Holtslag AAM, de Bruin HAR (1988) Applied Modeling of the Nighttime Surface Energy Balance over Land. J Clim Appl Meteor 27:689–704
17. Holtslag AAM, Ek MB (2005) Atmospheric Boundary Layer Climates and Interactions with the Land Surface. In: Encyclopedia of Hydrological Sciences. Wiley
18. Holtslag AAM, Steeneveld GJ, van de Wiel BJH (2007) Role of land-surface feedback on model performance for the stable boundary layer. Bound-Layer Meteorol 125:361–376
19. Lenderink G, Van den Hurk B, van Meijgaard E, van Ulden A, Cuijpers H (2003) Simulation of present day climate in RACMO2: First results and model developments. KNMI Technical report TR-252, 24 p
20. Mahrt L (1998) Stratified atmospheric boundary layers and breakdown of models. Theor Comp Fluid Phys 11:263–279
21. Mahrt L (1999) Stratified atmospheric boundary layers, Bound-Layer Meteorol 90:375–396
22. Poulos GS et al (2002) CASES-99: A comprehensive investigation of the stable nocturnal boundary layer. Bull Am Meteor Soc 83:555–581
23. ReVelle DO (1993) Chaos and "bursting" in the planetary boundary layer. J Appl Meteor 32:1169–1180
24. Salmond JA, McKendry IG (2005) A review of turbulence in the very stable boundary layer and its implications for air quality. Prog Phys Geogr 29:171–188
25. Sharan M, Gopalakrishnan SG (1997) Comparative Evaluation of Eddy Exchange coefficients for strong and weak wind stable boundary layer modelling. J Appl Meteor 36:545–559
26. Steeneveld GJ (2007) Understanding and prediction of stable boundary layers over land. Ph D thesis, Wageningen University, 199 pp
27. Steeneveld GJ, van de Wiel BJH, Holtslag AAM (2006) Modeling the arctic stable boundary layer and its coupling to the surface. Bound-Layer Meteorol 118:357–378
28. Steeneveld GJ, van de Wiel BJH, Holtslag AAM (2006) Modeling the Evolution of the Atmospheric Boundary Layer Coupled to the Land Surface for Three Contrasting Nights in CASES-99. J Atmos Sci 63:920–935
29. Steeneveld GJ, Mauritsen T, de Bruijn EIF, Vila-Guerau de Arellano J, Svensson G, Holtslag AAM (2008) Evaluation of limited area models for the representation of the diurnal cycle and contrasting nights in CASES99. J Appl Meteor Clim 47:869–887
30. Stull RB (1988) An introduction to Boundary-Layer Meteorology. Kluwer, Dordrecht, 666 pp. (reprinted 1999)
31. Svensson G, Holtslag AAM (2006) Single column modeling of the diurnal cycle based on CASES99 data -GABLS second intercomparison project. 17th Symposium on Boundary layers and Turbulence, 22–25 May, San Diego. American Meteorol Soc, Boston, Paper 8.1 (available at http://ams.confex.com/ams/pdfpapers)

32. Tjemkes SA, Duynkerke PG (1989) The nocturnal boundary layer: model calculations compared with observations. J Appl Meteor 28:161–175
33. van de Wiel BJH (2002) Intermittency and Oscillations in the Stable Boundary Layer over Land. Ph D thesis, Wageningen University, 129 pp
34. van de Wiel BJH, Moene AF, Hartogensis OK, de Bruin HAR, Holtslag AAM (2003) Intermittent turbulence and oscillations in the stable boundary layer over land, Part III: a classification for observations during CASES99. J Atmos Sci 60:2509–2522
35. van de Wiel BJH, Moene AF, Steeneveld GJ, Hartogensis OK, Holtslag AAM (2007) Predicting the Collapse of Turbulence in Stably Stratified Boundary Layers. Turbul Flow Combust 79:251–274
36. Welch RM, Ravichandran MG, Cox SK (1986) Prediction of Quasi-Periodic Oscillations in Radiation Fogs. Part I: Comparison of Simple Similarity Approaches. J Atmos Sci 43:633–651

Slug Flow: Modeling in a Conduit and Associated Elastic Radiation

Luca D'Auria, Marcello Martini
Osservatorio Vesuviano, Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Napoli, Naples, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Slug Flow and Strombolian Activity
Numerical Modeling
Bubble Ascent
Slug Ascent in a Vertical Pipe
Slug Ascent in a Pipe with a Flare
Seismological Constraints on Numerical Models
Conclusions
Future Directions
Appendix A – Fluid Dynamics of a Two-Phase System
Appendix B – Numerical Implementation
Bibliography

Glossary

Strombolian activity Is a kind of volcanic activity consisting of discrete, intermittent ejections of gas and magma fragments. The height reached by the ejecta seldom exceeds a few hundred meters above the volcanic crater. They consist mostly of molten magma fragments that partially quench as they fly. Strombolian explosions are intermediate between small Hawaiian eruptions and stronger Plinian eruptions. This kind of activity takes its name from the Stromboli volcano (Southern Italy) famous for its perpetual activity consisting of repeated moderate explosions (usually about 10 every hours).

Gas slug A bubble whose diameter is close to the diameter of the conduit where it is flowing. The shape and the motion of the slug is largely controlled by the conduit walls.

Modeling A simplified representation of a system aimed at reproducing some of its features. Mathematical models describe a system through a set of variables related by analytical relations. When these relations are too complex to be solved exactly, they can be solved using approximate numerical methods. Actually these numerical techniques often involve the massive use of computers. An alternative kind of modeling (analogue modeling) involves the use of versions of the system under study rescaled to fit spatial and temporal laboratory scales. Different materials are used to simulate original ones. For instance silicon oil is often used to simulate magma. The scaling relationships are rigorously stated in order to get physically meaningful results.

Computational fluid dynamics A set of mathematical, numerical and computational tools aimed at simulating complex fluid flows on computers. It developed simultaneously with computer science starting from the 1950s mostly with the aim of solving engineering problems. Today CFD spans a wider range of fields, from aeronautics to chemical engineering, to astrophysics to geophysics and much more. The most recent developments of CFD are related with the increase in computer performances and with the wider use of parallel computers.

Diffuse interface theory A molecular theory for describing the variation of the chemical composition across the interface on the basis of a rigorous thermodynamical approach. Starting from the definition of a free-energy function dependent on the chemical composition and its gradient, it is possible to compute all the chemical-physical properties of the interface. Beyond its physical meaning, this theory can be used also as a numerical tool for modeling of multiphase flows.

Very-long-period events (VLP) Seismic events recorded on active volcanoes and geothermal systems having a typical period of 10^2–10^0 s. Their observation and study began during the 1990s, with the spreading of seismic broadband sensors and have shown to be one of the most powerful tools for investigating the geometries of volcanic conduits and the dynamics of volcanic eruptions. Until now they have been observed

in tens of volcanoes with different eruptive styles such as: Aso (Japan), Erebus (Antarctica), Kilauea (Hawaii), Popocatepetl (Mexico), Sakurajima (Japan), Stromboli (Italy).

Moment-tensor Tensor representation of the force systems acting on seismic sources. It can be applied to common earthquake sources (rupturing faults) as well as to volcanic sources (fluid filled conduits). In the former case its trace is null, which from a physical point of view, means that there is no netvolume change during common earthquakes. On the other hand in volcanic sources volumetric variations are very common and furthermore they are accompanied also by a single force component, related to the net acceleration of center of mass of the fluid filling the conduit.

Definition of the Subject

Among the eruptive styles, the Strombolian activity is one of the more easy to study because of its repetitive behavior. For this reason large amount of data can be comfortably collected. Strombolian volcanoes are like natural laboratories repeating the same experiment (individual explosions) many times each day.

The development of quantitative models of eruptive dynamics is driven by the comparison of experimental observations and synthetic data obtained through mathematical, numerical or analogue modeling.

Since Strombolian activity offers a profuse amount of interesting seismic signals, during the last decades there has been growing attention on seismological techniques aimed at retrieving the conduit geometry and the eruption dynamics from the seismological recordings. One of these techniques, the source function inversion, is able to retrieve a summary of the forces acting on the volcanic conduit during the VLP event generation [5]. The comparison of observed source functions with synthetic ones, obtained through numerical modeling, allow us to put constraints on the proposed models.

Quantitative models, able to fit seismological observations, are a powerful tool for interpreting seismic recordings and therefor the seismological monitoring of active volcanoes.

Introduction

In this paper we discuss the mechanism of generation of Very-Long-Period events related to Strombolian explosions. This eruptive style, occurring in many basaltic volcanoes worldwide, is characterized by the ascent and the bursting of large gas slugs. The mechanism of formation, ascent and explosion of bubbles and slugs and their relation to eruptive activity has been studied from a theoretical point of view and by means of analogue simulations. Here we introduce results from numerical simulations, focusing on the pressure variations induced on the conduit walls and responsible for the generation of seismic signals.

We will first illustrate the main features of the fluid dynamics related to Strombolian eruptive activity (Sect. "Slug Flow and Strombolian Activity") and an overview of the numerical modeling (Sect. "Numerical Modeling") Then we will show results obtained using simple conduit model (Sect. "Bubble Ascent", Sect. "Slug Ascent in a Vertical Pipe" and Sect. "Slug Ascent in a Pipe with a Flare") and we will compare the synthetic source functions with actual observations (Sect. "Seismological Constraints on Numerical Models").

Slug Flow and Strombolian Activity

A fundamental distinction can be made between eruptive regimes on the basis of the magma viscosity. In silicic systems, the magma viscosity ($>10^5$ Pa s) is too high to allow an independent motion of gas bubbles [18]. They can only grow by diffusion processes and expand under the effects of pressure variations until fragmentation occurs. In basaltic magmas the viscosity ($<10^3$ Pa s) allows independent motion of the gas bubbles leading to a different behavior with the possibility of bubble coalescence, splitting and turbulent fluid flow [11]. Theoretical laboratory and numerical studies have been published in order to understand the dynamics of Strombolian eruption in terms of gas/magma interaction [12,24].

The distinction between free bubble ascent and slug flow can be made using the ratio between the average bubble diameter d and the conduit diameter D [7]: the parameter λ (Table 1). Values of λ higher than 0.6 are related to slug flow. The range of behavior exhibited by slug flow depends on the physical properties of the fluids (density, viscosity, surface tension) and on the geometry of the conduit (diameter, shape, inclination). It is possible to define adimensional parameters for describing the particular flow regime (see Table 1). The Reynolds number is the ratio between inertial and viscous forces. Low values of Re are typical of laminar flows, while higher values are related to turbulent regimes. The Froude number Fr is the ratio between inertial and gravitational forces. The Eotvos number Eo is the ratio between buoyancy and surface tension effects. The values of Eo determines the bubble shape. High values of Eo are related to distorted bubble shapes, while lower values to sub-spherical shapes. The Morton number Mo has a similar meaning. The dimensionless inverse viscosity N_f assess the relative importance of viscous effects.

Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Table 1
Symbols used in the text

Symbol	Meaning
d	Average bubble diameter
D	Conduit diameter
L	Distance between the top of the bubble and the magma/air interface in the initial conditions.
U	Terminal slug velocity
λ	$\lambda = \frac{d}{D}$
g	Gravity
ρ	Liquid density
$\Delta\rho$	Difference between gas and liquid densities
μ	Dynamic viscosity
ν	Kinematic viscosity $\nu = \frac{\mu}{\rho}$
σ	Surface tension
γ	Isothermal expansion ratio
Re	Reynolds number $Re = \frac{Ud}{\nu}$
Fr	Froude number $Fr = \frac{U}{\sqrt{gD}}\sqrt{\frac{\rho}{\Delta\rho}}$
Mo	Morton number $Mo = \frac{g\mu^4\Delta\rho}{\rho^2\sigma^3}$
Eo	Eotvos number $Eo = \frac{\Delta\rho g D^2}{\sigma}$
N_f	Dimensionless inverse viscosity $N_f = \left(\frac{Eo^3}{Mo}\right)^{\frac{1}{4}}$

Viscous flows are characterized by $N_f < 2$ while inertial flows by $N_f > 200$ [9].

The ascent of gas in a basaltic system has been studied both from a theoretical and an experimental point of view. Observations of basaltic systems have shown that usually the flow conditions are transitional between viscous dominated and inertia dominated systems [17]. This puts constraints on the parameter range to explore both in numerical and analogue modeling. The range of adimensional numbers for basaltic systems, summarized from [17] and [9] is reported in Table 2.

For a basaltic system with given physical properties another variable plays a fundamental role in determining the physics of the flow: the gas/magma volumetric ratio. Flows are characterized by a low ratio consist of isolated bubbles rising with minor interaction between them and the conduit walls (Fig. 1a). In this regime bubbles rise assuming a shape that depends on their size and on the magma viscosity. Surface tension effect depends on the size of the bubbles and the ascent velocity depends on both the bubble size and the magma viscosity. In the range allowed for basaltic systems, larger bubbles have shapes ranging from dimpled ellipsoidal cap to spherical cap, while smaller bubbles have shapes ranging from spherical to ellipsoidal [7]. For higher gas/magma ratios bubbles begin to interact and to coalesce forming gas slugs occupying most of the conduit diameter (Fig. 1b). The ascent of gas slugs is strongly controlled by the conduit walls. During their passage, the magma is completely mixed and the motion of smaller bubbles is dominated by the fluid vorticity released (in a transitional regime) in their wake. For even higher gas/magma ratios (Fig. 1c) the slugs coalesce forming an almost continuous (turbulent) gas flow at the center of the conduit, while walls remains wet with a magma film. The high velocity of the flow produces instabilities on the surface of the magma film pulling out small fragments carried away from the gas.

Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Table 2
Range of adimensional numbers for flows related to basaltic systems (from [17] and [9])

Parameter	Range
Fr	$0.1 \div 0.345$
Re	$5 \div 31^3$
Eo	$51^5 \div 71^7$
Mo	$51^5 \div 101^{10}$
N_f	$16 \div 5000$

Bubbly flows are related to effusive and to moderate Hawaiian explosive activity. Each single bubble explodes at the magma/air interface ejecting small magma fragments resulting from the shattering of the liquid film around the bubble. The bubble is usually overpressurized because of surface tension, viscous and inertial effects [23]. Slug flows are related instead to intermittent Strombolian activity.

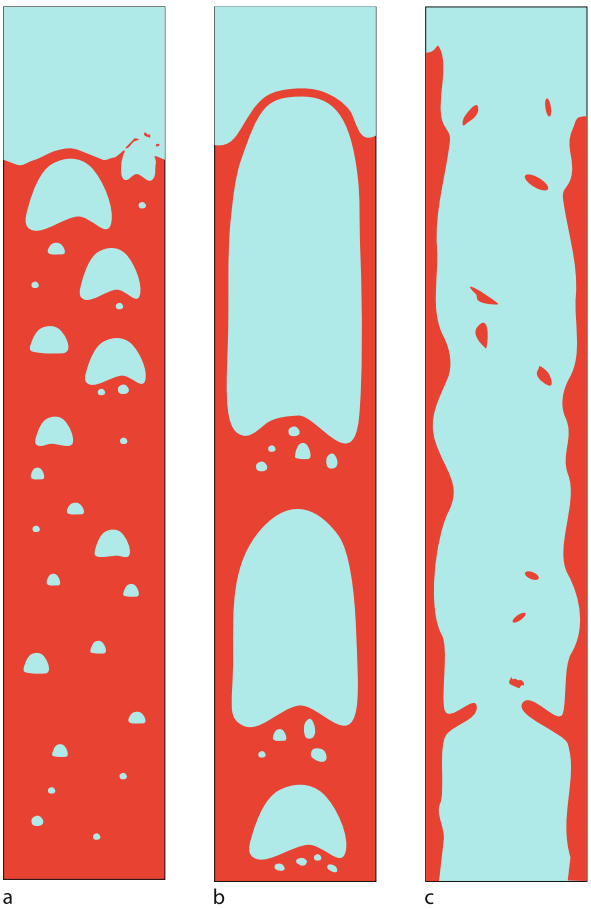

Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 1
Schematic representation of gas/magma flows in basaltic systems. a represents bubbly flow, b slug flow and c annular flow

Each slug exploding at the magma/air interface generates a jet of gas of short duration (from a few seconds to some tens of seconds) [24]. The main Strombolian explosion can be followed by minor bursts caused by the smaller bubbles trapped behind the slug wake. Annular flows are related to continuous lava fountains with the central gas jet carrying molten magma fragments [11].

The nature of the flow can change along the conduit because of the gas expansion due to the magmastatic pressure decrease. This expansion makes the gas/magma volumetric ratio increase along the conduit. At the base of the conduit, where the gas starts to exolve from the magma there is always a bubbly flow. As the pressure decreases the flow can change into slug flow. This can be made easier by an inclined conduit [4] and by constrictions [12] that force the bubbles to coalesce even at lower gas/magma ratios than those required in a straight vertical conduit. The total gas flow can change in the conduit because of variations in the deep feeding system or because of non-linear instabilities due to the complex shape of the conduit system [11,12].

The expansion of bubbles and slugs is a fundamental factor to take into account when studying Strombolian activity [17]. Assuming a simple ideal isothermal state equation for the gas we get the relative volumetric change for a bubble rising in a magma with density ρ:

$$\frac{V_0}{V} = \frac{P_{\text{atm}} + \rho g h}{P_{\text{atm}}} \,. \tag{1}$$

It is easy to show that the expansion ratio in the upper few hundreds of meters is more than one order of magnitude.

The flow regime can change also horizontally, for instance in a subvertical dike the flow rate can be higher on one side, leading to a slug flow, and lower on the other side, leading to a bubbly flow. This explains the coexistence of effusive and Strombolian activity observed during some basaltic eruptions.

Numerical Modeling

The aim of this work is to investigate the pressure variations induced by a gas bubble rising in a magma-filled volcanic conduit. This phenomenon has been also investigated by means of analogue laboratory models [9,10,12,17].

The major drawback of analogue modeling, in this context, is that it provides only a limited number of pressure time series: one for each sensor. Numerical modeling provides a different point of view giving the full set of scalar (density), vector (velocity) and tensor (pressure) quantities over the whole computational domain. This allows quantitative inferences on the flow regimes and on the elastodynamic wavefield generated (see Sect. "Seismological Constraints on Numerical Models").

The modeling of two-phase systems as gas magma is not a simple task in computational fluid dynamics (CFD) [6]. Taking into account surface tension effects at the liquid-gas interface can be done in two different ways. The first consists in explicitly tracking the time evolution of the gas-liquid interface [22]. In situations involving complex flows with extensive occurrences of bubble coalescence and splitting, these methods show a rapid increase in computational effort. Another category of methods model the gas-liquid systems using the diffuse-interface theory. These methods consider two scalar fields, defining the relative local concentrations of the two components and a modified advection-diffusion equation for

Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Table 3
Description and adimensional numbers for the six simulations presented in the text

Sim.	Description	Comp. domain (pts.)	λ	γ	*Re*	log(*Mo*)	*Eo*	*Fr*	N_f
1A	Bubble ascent (low viscosity)	60 × 600	0.5	5	1.93×10^3	0	9×10^4	0.37	5.19×10^3
1B	Bubble ascent (high viscosity)	60 × 600	0.5	5	1.90×10^2	4	9×10^4	0.36	5.19×10^2
2A	Slug ascent in vertical pipe (low viscosity)	30 × 600	0.8	5	1.82×10^3	0	9×10^4	0.35	5.19×10^3
2B	Slug ascent in vertical pipe (high viscosity)	30 × 600	0.8	5	1.54×10^2	4	9×10^4	0.30	5.19×10^2
3A	Slug ascent in vertical pipe with a flare (low viscosity)	60 × 600	0.8	5.3	1.83×10^3	0	9×10^4	0.35	5.19×10^3
3B	Slug ascent in vertical pipe with a flare (high viscosity)	60 × 600	0.8	5.3	1.68×10^2	4	9×10^4	0.32	5.19×10^2

modeling the physical-chemical interaction between them. This is done using a thermodynamically consistent definition of a free-energy function that takes into account the phase equilibria. This approach leads to smooth interfaces where the concentration of one phase gradually decreases while the other increases. The thickness of these interfaces depends on the numerical method and on the surface tension value [25]. From a numerical point of view this problem can be faced using Lattice Boltzmann Methods (LBM) [20] or by classic CFD [6]. Here we use the latter. Some details about the numerical method are reported in the Appendices.

In our models we have not considered the mechanical interaction between the fluid phases and the elastic conduit walls. The exchange of linear momentum between them is a significant factor in the generation of seismic waves [3] in seismo-volcanic sources. However in conduits having ratios between the linear dimension and thickness lower than 10^1–10^2 the effect of the motion of the conduit walls does not affect significantly the fluid flow. Furthermore we show in Sect. "Seismological Constraints on Numerical Models" that we will not compare the result of the simulation directly with seismograms, but with an equivalent system of forces acting on the conduit walls.

In the following we present results of some elementary two-dimensional conduit models focusing on the slug flow regime which occurs during Strombolian activity. The numerical simulations generate snapshots of the physical quantities (composition, density, pressure and velocity) along all the point of the discretized computational domain. The effective computational domains for each simulation are showed in Table 3. In all the simulations we start from a static fluid with a bubble in the lower part of the conduit and a gas/magma interface in the upper part. The boundary conditions keeps a constant pressure value at the bottom and the top of the model. The no-slip boundary conditions are implemented along the conduit walls.

The adimensional numbers for each simulation are reported in Table 3. Velocities for determining *Re* and *Fr* are computed when the bubble/slug is moving in a steady state after a transient due to the initial conditions. These velocity values U are also used for normalizing times:

$$t^* = t\,\frac{U}{L}\,, \tag{2}$$

where L is the distance between the top of the bubble and the magma/air interface. So (2) represents a normalization for the virtual time the bubble needs to reach the surface in an ideal steady motion. We can also normalize distances:

$$x^* = x\,\frac{1}{D}\,, \tag{3}$$

where D is the conduit diameter. We can define then normalized velocities as:

$$v^* = \frac{\partial x^*}{\partial t^*}\,. \tag{4}$$

In our simulations we have also defined an isothermal expansion ratio γ which is the ratio between the bubble volume at the initial conditions and its volume at (simulated) atmospheric pressure.

Bubble Ascent

We first consider the ascent of a single bubble in a vertical conduit. In the first simulation (1A in Table 3) the initial shape of the bubble is an ellipse (see Fig. 2). As the simulation evolves the bubble becomes unstable splitting into three smaller bubbles. The two smaller lateral bubbles are embedded in a symmetric vortex structure. As the flow

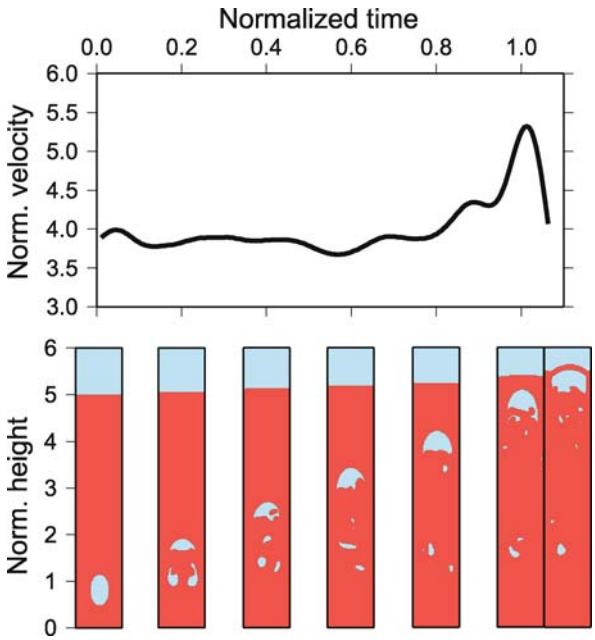

Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 2

Ascent and bursting of a gas bubble (simulation 1A in Table 3). Time, height and velocity are normalized according to (2), (3) and (4)

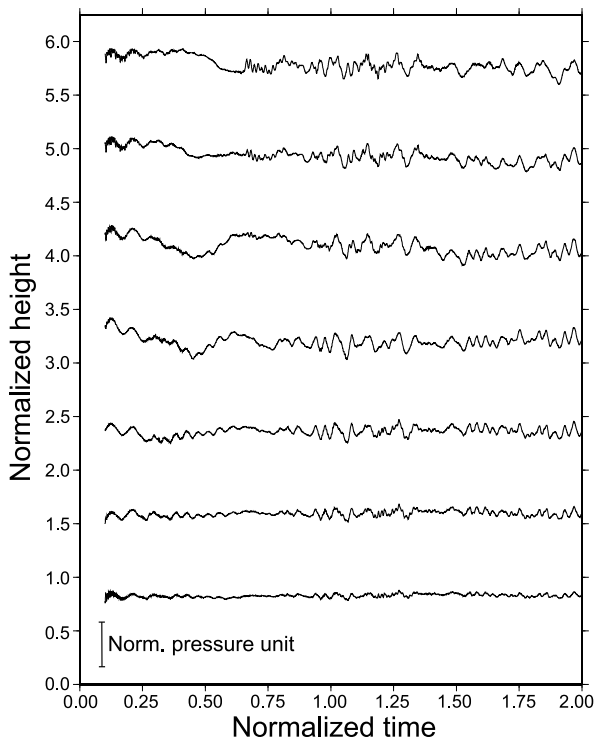

Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 3

Time series of pressure variations for simulation 1A. The represented pressure values are normalized by the external atmospheric pressure P_0. Each time series represents the recording of a virtual sensor located at the center of the conduit at the height of its first point (see *scale on the left*)

evolves the symmetry is broken and a turbulent wake develops behind the largest bubble. This one rises with an almost steady velocity (Fig. 2 top) with a spherical cap shape. In the final part of the simulation, the bubble suffers an expansion that made the bubble accelerate ($t > 0.8$) and causes the flow to become transitional toward the slug flow regime. Then the bubble reaches the surface and a curved liquid film develops above the bubble just before bursting. This phase is accompanied by a sudden deceleration of the bubble ascent ($t > 1.0$). The pressure variations in the conduit are modest, with a lack of low frequency oscillations. Also the pressure transients generated by the bubble burst are limited (Fig. 3).

The adimensional number for this simulation (Table 3) is in a range related to transitional flow regimes typical of basaltic systems [9,17]. An increase of an order of magnitude of the viscosity (1B in Table 3) makes the flow still in the transitional regime (Table 3). The high values of the Eotvos number indicates that the surface tension effect is not relevant in the flow dynamics [7] in both cases.

Slug Ascent in a Vertical Pipe

In these simulations (2A and 2B in Table 3) we model the ascent of a single slug in a vertical pipe. Again we start from an elliptical shape of the bubble. After an initial transient the typical slug shape develops (Fig. 4) and a turbulent wake appears behind the slug. As the slug rises it suffers a volume expansion. In the slug flow regime the liquid is pushed upward causing an overall increase of the hydrostatic pressure in the conduit. After the bursting of the slug the liquid film on the wall falls down to the original hydrostatic level. The slug expansion causes a significant acceleration of the liquid for $t > 1.0$ (Fig. 4). The velocity drops as the slug reaches the bursting point and the liquid film develops.

The major feature in the pressure pattern is the slow ramp-like increase followed by a drop (Fig. 5). The increase reflects the rise of the hydrostatic head above the slug while the gradual decrease is related to the passage of the slug. The bursting of the slug generates a moderate pressure transient in the conduit (marker "T" in Fig. 5) and damped resonant oscillations in the uppermost part of the pipe, filled with gas (above norm. height 11 in Fig. 5).

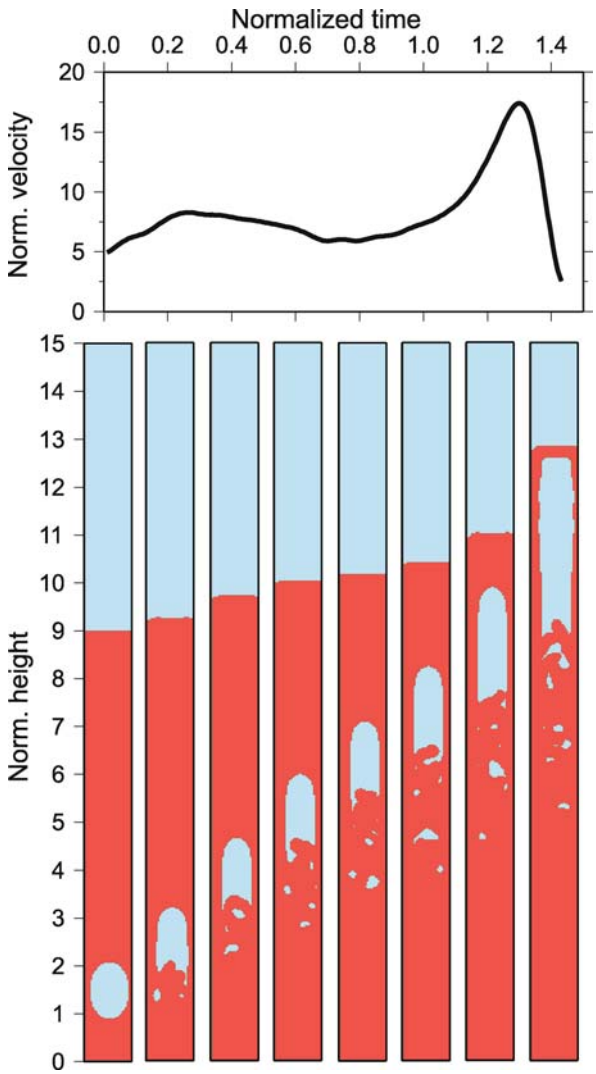

Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 4
Ascent and bursting of a single gas slug in a straight pipe (simulation 2A in Table 3)

The simulated pressure pattern fits very well the observations of analogue modeling [9]. Both the pressure increase linked to the rise of the hydrostatic head and the pressure drop related to the passage of the slug are in good agreement. Together with these long period variations they observe also oscillatory pressure transients just before the slug approaches the surface and after the bursting. The simulations presented here lack this feature because of an intrinsic limit in the numerical method that does not model explicitly the interface (see Appendix).

The overall behavior of simulation 2B, where a tenfold increase of the liquid viscosity (Table 3) is very similar to

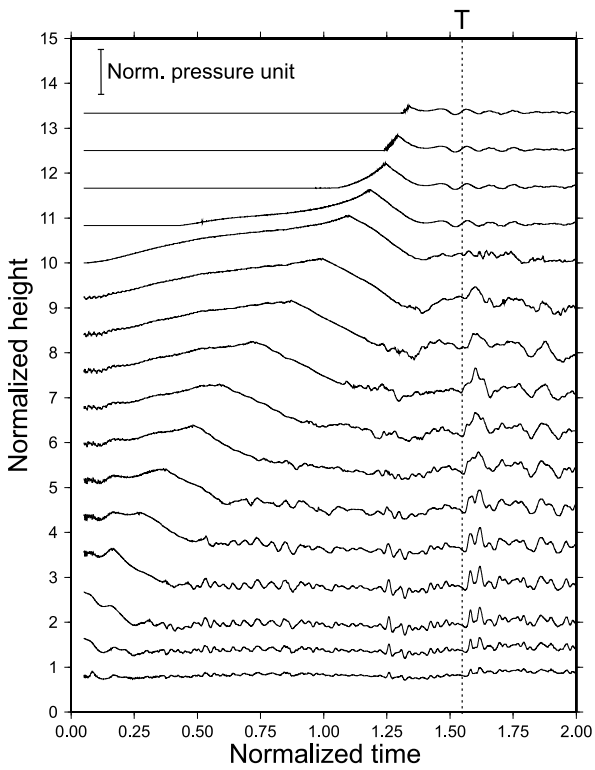

Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 5
Time series of pressure variations for simulation 2A. *T* marks the most significant transient in the pressure patterns

2A. The only remarkable difference is the lack of high frequency oscillations in the pressure (as the transient "T" in Fig. 5) due to the greater viscous damping effect.

Slug Ascent in a Pipe with a Flare

In this set of simulations (3A and 3B in Table 3) we model again the ascent of a single slug in a pipe. However in this case the width of the pipe doubles after a norm. height of 10. In the first part of the simulation, the behavior is similar to the one illustrated in Sect. "Slug Ascent in a Vertical Pipe". As the slug nose enters in the flare (Fig. 6) is starts to expand rapidly making the fluid accelerate upward. The expansion of the slug in the flare is followed by its breakup because of the development of strong turbulence. The velocity of the top of the slug increases until the slug passes through the flare then it drops because of the change in the conduit diameter.

The pressure pattern shows a major difference, compared with simulation 2A (Fig. 7). As the slug passes through the flare there is a sudden pressure increase

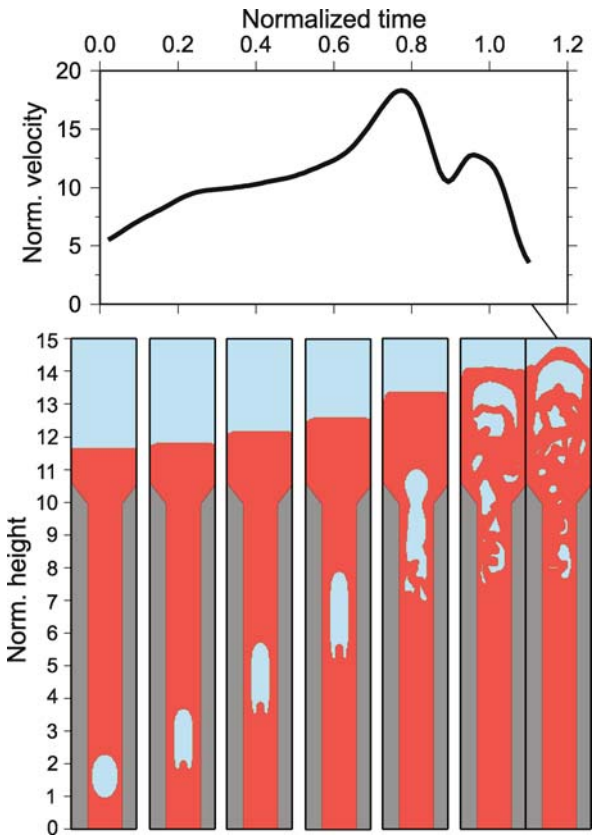

Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 6

Ascent and bursting of a single gas slug in a pipe with a flare (simulation 3A in Table 3)

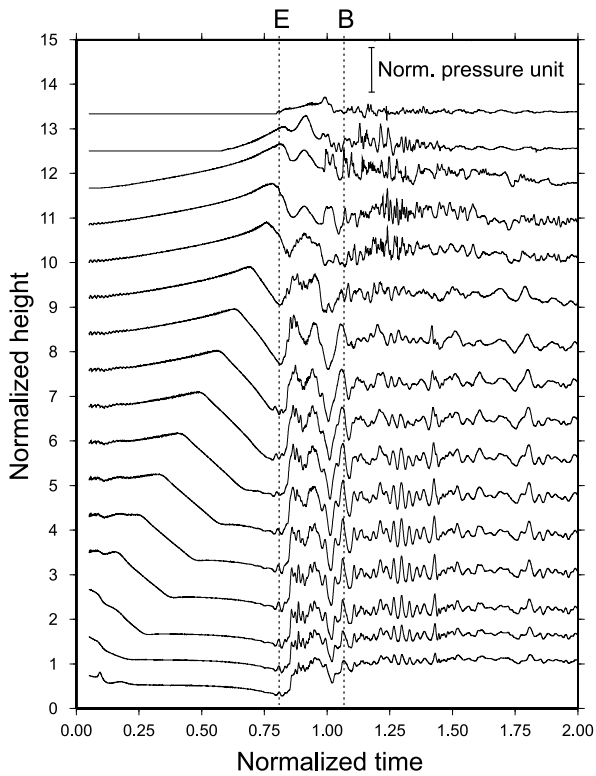

Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 7

Time series of pressure variations for simulation 3A. *E* marks the entering of the slug in the upper conduit while *B* marks the bursting of the main bubble at the magma surface

recorded along the whole conduit. The pressure then drops when the slug expansion has terminated. This strong pressure pulse is clearly related to the varying fluid flow regime as the slug enters the upper conduit and it deserves a closer analysis. In Fig. 8 we have represented the fluid dynamic regime during the slug expansion together with the related pressure variations. We observe that the pressure rises when most of the slug has passed through the flare. The strong acceleration induces turbulence both in the lower and the upper sections of the conduit. The upward acceleration first causes the disruption of the lower part of the slug leaving behind the main bubble a set of smaller ones, whose motion is driven by the fluid turbulence. When the slug has entered in the upper conduit it suffers another splitting due both to the induced vorticity and to the change in the boundary conditions. The behavior of the slug for $t^* > 0.8$ is characterized by a flow regime with a smaller value of λ and so it is similar to the initial part ($t^* < 0.2$) of simulation 1A.

Similar fluid dynamics behavior and pressure patterns have been observed in analogue simulations by [10]. In particular, the breakup of the slug and the generation of a positive pressure pulse as it passes through the flare are observed in analogue models within a wide range of fluid viscosities and conduit widenings.

As in the previous case the simulation with higher liquid viscosity (3B in Table 3) shows a similar behavior. Higher viscosities reduce the fluid vorticity leading to less fragmented slugs.

Seismological Constraints on Numerical Models

Seismological data analysis is a powerful tool for putting constraints on the geometry and the dynamics of volcanic systems. Non-stationary fluid flow in volcanic conduits generates pressure variations on the conduit walls and then seismic waves propagating toward the Earth's surface where they are recorded by seismometers. The frequency band of seismic signals recorded in volcanic ar-

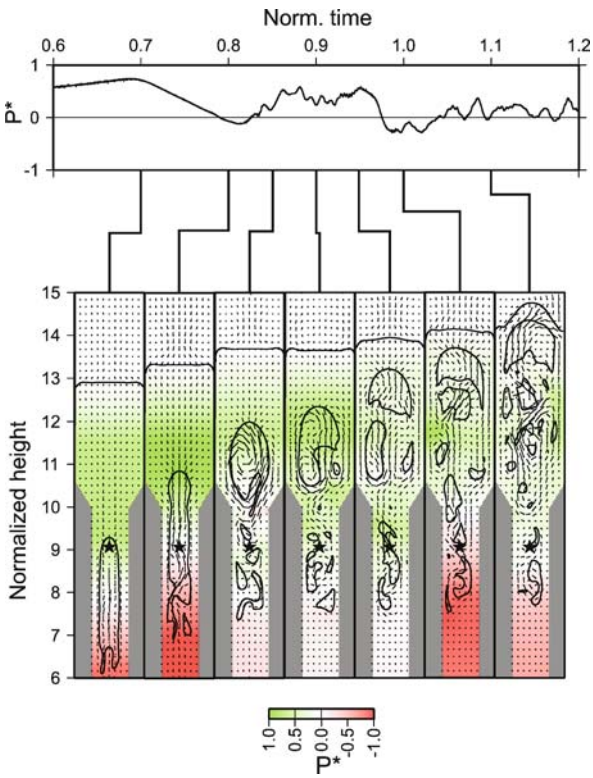

Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 8

Detail of the conduit during the slug expansion phase. Each snapshot represents the normalized pressure variations (see the *color scale* on the bottom) and the local fluid velocity vectors. The *thick contours* are the bubble interfaces. The pressure values represented above are measured at the virtual sensor indicated by a *black star* in the snapshots

eas spans a wide range: from the Ultra-Long-Period (ULP) band ($>10^2$)s, the Very-Long-Period (VLP) band (10^2–10^0 s) and the Long-Period (LP) band (10^0–10^1 s) [5].

From a formal point of view this can be expressed considering the elastodynamic field generated by a general extended source, whose external surface is Σ:

$$u_n(\mathbf{x}, t) = \iint_{\Sigma} [f_q * G_{np} + m_{pq} * G_{np,q}] \mathrm{d}\Sigma \,, \tag{5}$$

where $u_n(\mathbf{x}, t)$ is the nth component of the ground displacement recorded at the position $\mathbf{x}$, f_q is the body-force distribution over Σ, m_{pq} is the moment density tensor and G_{np} are the Green's functions. In (5) we take into account all the details of the source dynamics. However the density of actual seismic networks is not sufficient for retrieving every minor feature of the seismic sources. Volcano monitoring networks usually have a limited extension (10^3–10^4 m) and the wavelengths associated with ULP and VLP signals are higher. Therefore the analysis of such signals the seismic source (i. e. the part of the volcanic conduit responsible of seismic wave generation) can be represented as a point [5]. Under the assumption of a point source we can express (5) as [5]:

$$u_n(\mathbf{x}, t) = F_q * G_{np} + M_{pq} * G_{np,q} \,, \tag{6}$$

where:

$$F_q = \iint_{\Sigma} f_q \mathrm{d}\Sigma \tag{7}$$

and:

$$M_{pq} = \iint_{\Sigma} m_{pq} \mathrm{d}\Sigma \,. \tag{8}$$

In common earthquake sources the single force component F_q is null. In volcanic sources, the acceleration of the center of mass of the fluid makes this component noteworthy.

The inversion of the recorder waveforms $\mathbf{u}(\mathbf{x}, t)$, after the numerical computation of the Green's functions $\mathbf{G}$ allows the retrieval of the single force and moment tensor components of (6) [14]. $\mathbf{F}$ and $\mathbf{M}$ are a synthesis of the force systems acting on volcanic conduits and then can be used for discriminating among numerical models on the basis of their fit with observations.

Numerical simulations provide the pressure tensor field $\mathbf{P}$ over the whole computational domain. The pressure tensor can be used for computing the forces acting on each point of the conduit walls multiplying it for the unit vector normal to the wall $\hat{\mathbf{n}}$:

$$\mathbf{f} = \mathbf{P}\hat{\mathbf{n}} \,. \tag{9}$$

These values can be integrated numerically using an expression similar to (7) to get the single force component $\mathbf{F}$. Then using the definition of moment they can be used for retrieving also the equivalent moment tensor:

$$M_{pq} = \iint_{\Sigma} \left(f_q - \frac{F_q}{\Sigma} \right) r_p \mathrm{d}\Sigma \,, \tag{10}$$

where $\mathbf{r}$ is the arm formed by the force respect to an arbitrary origin $\mathbf{O}$.

In Fig. 9 we have represented the vertical force component and the moment tensor isotropic component (that is the trace $M_{xx} + M_{zz}$) for the three simulations with lower liquid viscosity. The represented values are normalized.

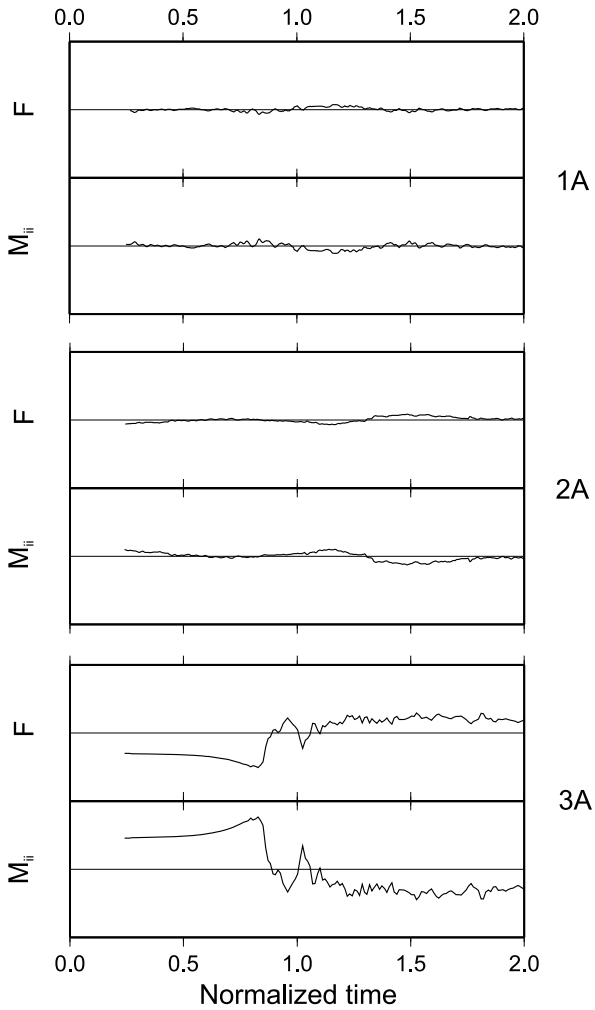

Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 9

Vertical force and isotropic component for simulations 1A, 2A and 3A of the moment tensor for simulations 1A, 2A and 3A (see Table 3)

In the first case (1A) we note the absence of significant signals. There are only minor oscillations related to fluid dynamics instability during the bubble ascent. The rise of a bubble smaller than the conduit dimension seems not to be able to generate VLP signals. Therefore a repetitive bursting of small bubbles at the magma/air interface can be responsible for the generation of continuous infrasonic and seismic tremor recorded in some basaltic volcanoes [16].

In the second case, again, the signal amplitudes are quite low. It is evident that there is an anticorrelation between the vertical force and the isotropic moment. The vertical acceleration of the slug ($1.0 < t^* < 1.3$ in Fig. 4) causes a downward reaction force as well as an increase in the conduit pressure. The bursting ($t^* > 1.3$) causes a downward acceleration of the liquid and a pressure decrease. A similar effect, although with a minor magnitude, is also evident in 1A (Fig. 9).

In the third case the signal amplitude increases dramatically, about five times higher than 2A. The slug entering in the upper, wider portion of the conduit is marked by a downward force and a positive moment. At the end of the expansion phase ($t^* > 0.8$) there is a sudden inversion of the trends in both quantities, followed again by a positive downward peak of the force and positive of the momentum ($1.0 < t^* < 1.1$). This peak occurs when two pieces of the original slug coalesce again (Fig. 8) leading to a sudden, but limited in time, upward acceleration. In real cases the dynamics can be even more complex with multiple slugs or slug fragments entering in the upper conduit and interacting with each other. In this case during the main expansion phase, the lower conduit is filled almost exclusively with liquid (Fig. 6). If we suppose the simultaneous presence of many vertically aligned slugs in the lower conduit, then the pressure variations induced by the expansion of the topmost one would influence the lower ones in a complex non-linear mechanism still to investigate.

The signals presented in Fig. 9 need to be scaled to an actual time scale to be compared with real seismic signals source functions. In basaltic volcanoes the dimensions of the upper conduit is of the order of 10^2 m while the ascent velocities of the slugs are of the order of 10^0–10^1 m/s. This gives a scaling factor of about 10^1–10^2 s. Since the simulated transients have a normalized duration of about 10^{-1} the actual simulated signals should have a characteristic period of 10^0–10^1 s. These values are within the range of VLP and LP signals [5].

The patterns observed in simulation 3A (Fig. 9) closely matches the results obtained by [4] for the source function of VLP events at the Stromboli volcano. Thus this result, together with analogue simulations [10], strongly supports the hypothesis that VLP events at Stromboli are generated by the passage of a gas slug through a conduit widening and its subsequent expansion and bursting. The long ramp-like signals in both force and moment for $t^* < 0.8$ in Fig. 9 has a characteristic period longer than the VLP signals recorded and analyzed by [4]. Recordings at Stromboli, using a seismic sensor with a wider frequency range [13] interestingly suggested in some seismic signals a similar ramp having a length of more than 60 s, compatible with our scaling. This long ramp is related to the slow and gradual increase in the magmastatic head due to the continuous expansion of the ascending slug.

Conclusions

The rapid expansion of the gas slugs in the uppermost part of the conduit plays a fundamental role, both in the eruptive dynamics and the seismic wave generation process. We have focused on the role of conduit geometry in the fluid dynamics of gas slug ascent and its implication in the generation of seismic signals. In one of the simulations (3A) we have shown that the system of forces acting on the conduit is able to generate seismic waves with a higher efficiency compared with other cases. This observation can be generalized to more complex geometry such as an alternating of widening and narrowing [10]. The passage of slugs can occur also in very complex conduits [17]. In this case it is possible that pressure transients are generated in different positions along the conduit.

Future Directions

A deeper understanding of the relationship between slug ascent dynamics and seismic signals generation would require more advanced modeling techniques in various three-dimensional geometries, testing how the effect of changes in slug volumes and magma properties can affect the generation of seismic signals.

These studies are an important step toward more advanced seismic monitoring techniques of active basaltic volcanoes aimed at assigning in real time a volcanological meaning to variations in observed LP and VLP seismic signals.

Appendix A – Fluid Dynamics of a Two-Phase System

Definition of a Two-Phase System

A fluid two-phase system can be described using two scalar fields n_a and n_b, representing the local molar densities of the two components a and b. The actual local density ρ is then:

$$\rho = m_a n_a + m_b n_b \,, \tag{11}$$

where m_a and m_b are the molecular weight of the two phases. As it will be shown in the following, it is convenient to describe the system using an alternative representation based on the variables:

$$n = n_a + n_b \,, \tag{12}$$

and

$$\phi = n_a - n_b \,. \tag{13}$$

ϕ defines the local composition of the fluid. It can span the range $[-n, +n]$ with the value $-n$ representing a pure b composition and $+n$ a pure a composition. The relation between n, ϕ and ρ is:

$$\rho = \tfrac{1}{2}\left[m_a (n+\phi) + m_b (n-\phi)\right] \,. \tag{14}$$

The values of m_a and m_b are set so that in reference conditions the value of n is always equal to 1. So for instance, if we consider pure water at ambient conditions $m_w = 1000$ kg/mol.

State Equation

We assume that pure phases obey to a simple isothermal ideal gas state equation like:

$$P = \rho c^2 \,, \tag{15}$$

where P is the pressure and c is the sound speed. We assume that the pressure is zero in the reference state, so the state equation for the component a is:

$$P = m_a (n-1)\, c_a^2 \,. \tag{16}$$

A similar relation holds for b. Since along interfaces the composition varies continuously, we should define a state equation for a mixture that satisfies two requirements: the state equation for pure phases must match (16) and the isobaric contours in the (ϕ, n) plane must be straight lines. This second requirement follows from the consideration that diffusion processes between the components follows straight paths in this domain. In most of the actual Lattice Boltzmann literature, this question is not issued because the proposed implementations are almost isobaric ($n \simeq 1$) [21,25,26]. A suitable choice for the state equation is then:

$$P(n,\phi) = \frac{1}{4}\left[\Lambda + \sqrt{16 m_a m_b c_a^2 c_b^2 (n-1) + \Lambda^2}\right] \tag{17}$$

with:

$$\Lambda = m_a c_a^2 (n+\phi-2) + m_b c_b^2 (n-\phi-2) \,. \tag{18}$$

In Fig. 11 an example contour plot of isobaric lines is shown. We can define another parameter χ as:

$$\chi(n,\phi) = -1 + 2\sqrt{\frac{(\phi+B)^2 + (n-B)^2}{(A+B)^2 + (A-B)^2}} \,, \tag{19}$$

where:

$$A = \frac{P(n,\phi)}{m_a c_a^2} + 1 \,, \tag{20}$$

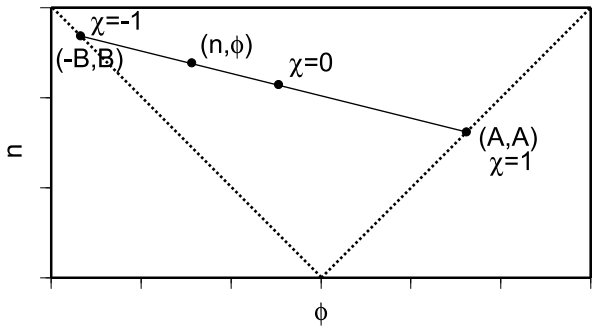

Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 10
Definition of χ. See text for details

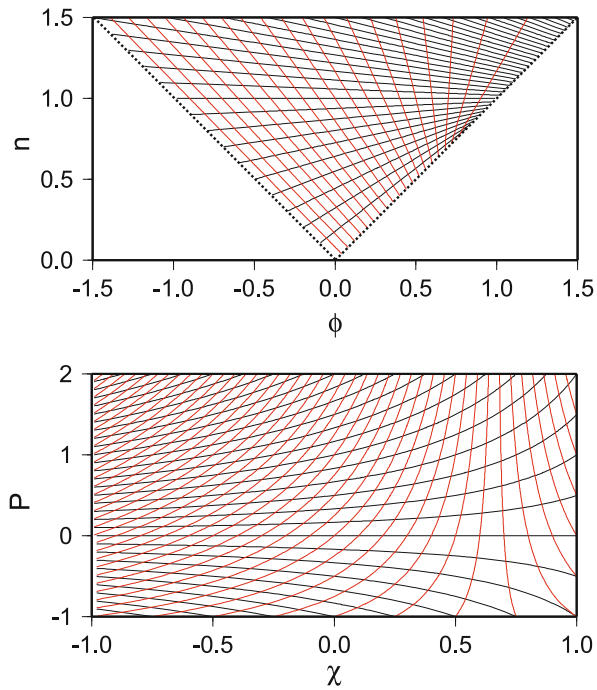

Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 11
In the upper panel, isolines of χ (*red lines*) and P (*black lines*) are represented in the (n, ϕ) domain. In the lower panel isolines of n (*black lines*) and ϕ (*red lines*) are represented in the (χ, P) domain. These plots represent a two-phase system where $m_a c_a^2 = 5$ and $m_b c_b^2 = 1$

and

$$B = \frac{P(n, \phi)}{m_b c_b^2} + 1 \,. \tag{21}$$

The actual meaning of χ is shown in Fig. 10. It is the rescaled ratio of the distance between the points (ϕ, n) and $(-B, B)$ and the distance between the points (A, A) and $(-B, B)$. So $\chi = -1$ along the line $\phi = -n$ and $\chi = 1$ along $\phi = n$. In practice χ is related to the local composition of the fluid. In Fig. 10 the relation between ϕ, n, P and χ is represented. This relation can be inverted and the system (χ, P) can be used as well as the (ϕ, n) coordinate system. In the following we show that the former is the best choice for describing chemical equilibria.

Chemical Equilibria and the Diffuse Interface Theory

The chemical equilibrium between the two phases can be described using a physically consistent thermodynamic approach [2,8,26]. We define a Ginzburg–Landau free energy functional for an heterogeneous mixture as [15]:

$$\mathcal{F} = \int \left[\psi(n, \phi) + \frac{\kappa_n}{2} (\nabla n)^2 + \frac{\kappa_\phi}{2} (\nabla \phi)^2 \right] \mathrm{d}V \,, \tag{22}$$

where κ_n and κ_ϕ are parameters related to the surface tension. The system reaches its final thermodynamical equilibrium making the free energy minimum.

We can rewrite (22) in the (χ, P) representation:

$$\mathcal{F} = \int \left[\psi(\chi, P) + \frac{\kappa_\chi}{2} (\nabla \chi)^2 \right] \mathrm{d}V \,. \tag{23}$$

Note that we have set $\kappa_P = 0$. This implies that surface tension effects are independent of the absolute pressure. In this paper we set:

$$\psi(\chi, P) = \alpha \left(\frac{\chi^4}{4} - \frac{\chi^2}{2} \right) + P \ln \frac{P}{P_0} \,, \tag{24}$$

where P_0 is an arbitrary reference pressure value. Using this definition of ψ we note that this function has two minima for $\chi = \pm 1$, corresponding to the two equilibrium compositions. Starting from an heterogeneous mixture, the system evolves by creating domains having a quite homogeneous composition separated by interfaces where the composition varies smoothly. The equilibrium is reached when the chemical potential:

$$\mu = \frac{\delta \mathcal{F}}{\delta \chi} \,, \tag{25}$$

is everywhere equal to zero [2,26]. So, in our case, using (23) and (24) we obtain the condition of chemical equilibrium for a plane interface orthogonal to the x-direction:

$$\mu = \frac{\partial \psi}{\partial \chi} - \kappa_\chi \frac{\partial^2 \chi}{\partial x^2} \,. \tag{26}$$

Assuming the boundary conditions $\psi(x) = 0$ and $\frac{\partial \chi}{\partial x} = 0$ for $x = \pm\infty$ and $\chi(0) = 0$, the previous ODE can be integrated giving the expression of the spatial variation of the

composition through an equilibrium interface:

$$\chi(x) = \tanh\left(\frac{2x}{\xi}\right), \tag{27}$$

where the interface thickness ξ is the width where the 96% of variation occurs. Its value is:

$$\xi = 2\sqrt{-\frac{2\kappa_\chi}{\alpha}}. \tag{28}$$

On the basis of the definition of surface tension σ [25]:

$$\sigma = \int_{-\infty}^{+\infty} \mathcal{F}(x)\,\mathrm{d}x, \tag{29}$$

we can write:

$$\sigma = \frac{2}{3}\sqrt{-2\alpha\kappa_\chi}. \tag{30}$$

Using (28) and (30) we can also retrieve useful inverse relations:

$$\alpha = 3\frac{\sigma}{\xi}, \tag{31}$$

and

$$\kappa_\chi = \frac{3}{8}\sigma\xi. \tag{32}$$

We should emphasize that in real physical systems expression (27) describes an actual interface having a thickness whose order of magnitude is of molecular scale. In this work we use the diffuse interface theory as a numerical tool for modeling of two-phase systems. In other words, we use unphysical interfaces having a macroscopic thickness (usually $10^{-3} \div 10^{-1} m$).

Cahn–Hilliard and Mass Conservation Equations

The time evolution under non-equilibrium conditions can be expressed by two Cahn–Hilliard equations [2,26]:

$$\frac{\mathrm{D}n}{\mathrm{D}t} = \Gamma\nabla^2\mu_n, \tag{33}$$

$$\frac{\mathrm{D}\phi}{\mathrm{D}t} = \Gamma\nabla^2\mu_\phi. \tag{34}$$

These equations are similar to common advection-diffusion equations with Γ being a diffusion coefficient and the operator $\mathrm{D}/\mathrm{D}t$ the substantial derivative. The explicit expressions for the chemical potentials μ_n and μ_ϕ are:

$$\mu_n = \frac{\delta\mathcal{F}}{\delta n} = \frac{\delta\mathcal{F}}{\delta\chi}\frac{\partial\chi}{\partial n} = \mu_\chi\frac{\partial\chi}{\partial n}, \tag{35}$$

$$\mu_\phi = \frac{\delta\mathcal{F}}{\delta\phi} = \frac{\delta\mathcal{F}}{\delta\chi}\frac{\partial\chi}{\partial\phi} = \mu_\chi\frac{\partial\chi}{\partial\phi}. \tag{36}$$

with:

$$\mu_\chi = \frac{\delta\mathcal{F}}{\delta\chi} = \alpha\chi(\chi^2-1) - \kappa_\chi\nabla^2\chi. \tag{37}$$

Then the Cahn–Hilliard Eqs. (33) and (34) can be rewritten as:

$$\frac{\mathrm{D}}{\mathrm{D}t}\begin{pmatrix} n \\ \phi \end{pmatrix} = \Gamma\nabla^2\mu_\chi\begin{pmatrix} \frac{\partial\chi}{\partial n} \\ \frac{\partial\chi}{\partial\phi} \end{pmatrix} \tag{38}$$

On the basis of the definitions of n (12) and ϕ (13), we can state that the previous equation expresses the mass transfer of components a and b because of chemical disequilibria. Since we are dealing with a compressible flow, we should obviously account for this in the mass balances. Then, following basic fluid dynamics [1], we can rewrite (38) in explicit form as:

$$\frac{\partial n}{\partial t} = -\mathbf{v}\nabla n - n\nabla\cdot\mathbf{v} + \frac{\partial\chi}{\partial n}\Gamma\nabla^2\mu_\chi \tag{39}$$

$$\frac{\partial\phi}{\partial t} = -\mathbf{v}\nabla\phi - \phi\nabla\cdot\mathbf{v} + \frac{\partial\chi}{\partial\phi}\Gamma\nabla^2\mu_\chi. \tag{40}$$

Thermodynamic Pressure Tensor

It can be shown, from statistical mechanics that the pressure tensor can be obtained from (23) [8,15]:

$$P_{ij}^{th} = p_0\delta_{ij} + \kappa_\chi\frac{\partial\chi}{\partial x_i}\frac{\partial\chi}{\partial x_j}, \tag{41}$$

where the isotropic component is:

$$p_0 = P\frac{\delta\mathcal{F}}{\delta P} + \chi\frac{\delta\mathcal{F}}{\delta\chi} - \left(\psi(\chi,P) + \frac{\kappa_\chi}{2}(\nabla\chi)^2\right). \tag{42}$$

so:

$$p_0 = P + \alpha\left(\frac{3}{4}\chi^4 - \frac{1}{2}\chi^2\right) - \kappa\chi(\nabla^2\chi) - \frac{\kappa}{2}(\nabla\chi)^2 \tag{43}$$

The tensor $\mathbf{P}^{th}$ describes the stresses induced by spatial variation in density and composition and has to be added to the viscous stress tensor $\boldsymbol{\tau}$.

Conservation Equations

Let us now apply the results of the previous section for building a system of fluid dynamics equations suitable for a numerical implementation. Together with the mass

conservation Eqs. (39) and (40) we need the conservation equation for the linear momentum [1]:

$$\rho \frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = \nabla \cdot \mathbf{P} + \rho \mathbf{g} \,, \tag{44}$$

where $\mathbf{P} = -\mathbf{P}^{th} + \boldsymbol{\tau}$ is the full pressure tensor of (41), $\mathbf{g}$ is the gravity and $\boldsymbol{\tau}$ is the viscous stress tensor that, for a Newtonian fluid is:

$$\tau_{ij} = \lambda e_{kk} \delta_{ij} + 2\mu e_{ij} \,, \tag{45}$$

with λ and μ the bulk and shear viscosities and e_{ij} is the strain rate tensor expressed by:

$$e_{ij} = \frac{1}{2}\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right) \,. \tag{46}$$

Appendix B – Numerical Implementation

Transformed Equations

A common problem in the numerical solution of the conservation Eqs. (33) and (34) is the output of values having no physical meaning (for instance negative densities for one of the components) with bad effects on the numerical stability of the code. This is due to the fact that both variables are defined over a limited range of values that are $n \in [0, +\infty[$ and $\phi \in [-n, +n]$. These conditions are not explicit in the conservation equations, but this problem can be overcome by making simple changes of variable:

$$n = \mathrm{e}^l \,, \tag{47}$$

$$\phi = \mathrm{e}^l \tanh q \,. \tag{48}$$

Using the new variable l instead of n the conservation Eq. (39) become:

$$\frac{\partial l}{\partial t} = -\mathbf{v} \cdot \nabla l - \nabla \cdot \mathbf{v} + \frac{1}{\mathrm{e}^l}\left(\frac{\partial \chi}{\partial n}\right) \Gamma \nabla^2 \mu \,. \tag{49}$$

Substituting (48) in (40) we first obtain:

$$\frac{\partial q}{\partial t} = -\mathbf{v} \cdot \nabla q - \frac{1}{2}\sinh 2q \left(\frac{\partial l}{\partial t} + \mathbf{v} \cdot \nabla l + \nabla \cdot \mathbf{v}\right) + \frac{1}{\mathrm{e}^l}\cosh^2 q \left(\frac{\partial \chi}{\partial \phi}\right) \Gamma \nabla^2 \mu \,. \tag{50}$$

Using (49) and some algebraic manipulation the previous expression become:

$$\frac{\partial q}{\partial t} = -\mathbf{v} \cdot \nabla q - \frac{1}{\mathrm{e}^l}\cosh q \left[\frac{\partial \chi}{\partial n}\sinh q + \frac{\partial \chi}{\partial \phi}\cosh q\right] \cdot \Gamma \nabla^2 \mu \tag{51}$$

Boundary and Initial Conditions

Equations (49) and (51) have to be solved setting proper boundary and initial conditions.

The no-slip boundary condition, implemented along the conduit walls is:

$$\mathbf{v} = 0 \,. \tag{52}$$

Another boundary condition is set along the walls:

$$\nabla \phi \cdot \mathbf{n} = 0 \,. \tag{53}$$

This is the neutral wetting condition [25] and it is needed in order to avoid the wall to behave sticky respect to one of the phases.

At the conduit top and bottom conditions of constant pressure are implementer. On the top the pressure is kept to a reference atmospheric value, while at the bottom it is kept at the hydrostatic pressure value, computed on the initial conditions.

In the initial conditions we set $\mathbf{v} = 0$ and the computational domain is composed of domains having a homogeneous composition separated by smooth interfaces, close to the equilibrium solution of (27). At the top of the model a gas domain represents the atmosphere, while the remaining part of the conduit is filled with magma. A small elliptical gas bubble in hydrostatic equilibrium is placed in the lower part of the conduit. The details are specified for each set of simulations.

Finite Difference Implementation

Equations (49), (51) and (44) are discretized on a regular grid and the differential equations are transformed in finite-difference equations [19]. Scalar, vector and tensor quantities are discretized on staggered grids [6] (Fig. 12). Scalar quantities (l, q, μ and P_{ii}) are defined at integer grid steps (x_i, z_j). Velocities are staggered half grid step along x and z directions, that is v_x is defined at $(x_{i+1/2}, z_j)$ and v_z at $(x_i, z_{j+1/2})$. The deviatoric part of the pressure tensor P_{xz} is defined at grid points $(x_{i+1/2}, z_{j+1/2})$. This allows a second-order accuracy in the computation of spatial derivative.

We apply also a staggering in time. At the time step k we first solve (49) and (51). Then the current values of l and q are updated:

$$l_{i,j}^{(k)} = l_{i,j}^{(k-1)} + \frac{\mathrm{d}l}{\mathrm{d}t}^{(k)} \Delta t \,, \tag{54}$$

$$q_{i,j}^{(k)} = q_{i,j}^{(k-1)} + \frac{\mathrm{d}q}{\mathrm{d}t}^{(k)} \Delta t \,. \tag{55}$$

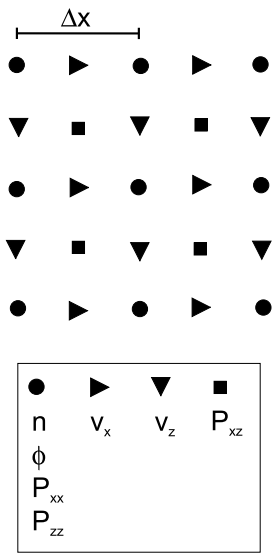

Slug Flow: Modeling in a Conduit and Associated Elastic Radiation, Figure 12
Spatial scheme of the staggered grid used in the computation. *Circle* represents the isotropic grid (density, composition and isotropic pressure), *rightward triangles* are the grids for v_x while *downward triangles* are for v_z. *Squares* are the grids used for discretizing P_{xz}

Using these new values the full pressure tensor $\mathbf{P}^{(k)}$ is computed using the discretized version of (41), (42) and (43). At the time step $k + 1/2$ (44) is solved and the values of the local velocities are updated.

The computation of $\frac{\mathrm{d}l}{\mathrm{d}t}^{(k)}$ and $\frac{\mathrm{d}q}{\mathrm{d}t}^{(k)}$ requires the values of $l^{(k-1)}$, $q^{(k-1)}$ and $v^{(k-1/2)}$. On the other hand the computation of $v^{(k+1/2)}$ requires the values of $l^{(k)}$, $q^{(k)}$ and $v^{(k-1/2)}$.

The time step Δt is chosen in order to satisfy a stability condition. Since we are dealing with isothermal flows with a very low Mach number, the condition can be expressed by:

$$\Delta x < \Delta t \; c_{\max} \; , \tag{56}$$

where $c_{\max}$ is the highest sound speed among the two components. In our simulations we have set:

$$\Delta x = 6 \; \Delta t \; c_{\max} \; , \tag{57}$$

Bibliography

Primary Literature

1. Aris R (1962) Vectors, tensors and the basic equation of fluid mechanics. Dover Publications
2. Cahn JW, Hilliard JE (1958) Free energy of a nonuniform system. i. interfacial free energy. J Chem Phys 28(2):258–267
3. Chouet B (1986) Dynamics of a fluid-driven crack in three dimensions by the finite difference method. J Geophys Res 91:13967–13992
4. Chouet B, Dawson P, Ohminato T, Martini M, Saccorotti G, Giudicepietro F, De Luca G, Milana G, Scarpa R (2003) Source mechanisms of explosions at stromboli volcano, italy, determined from moment-tensor inversions of very-long-period data. J Geophys Res 108(B1)
5. Chouet BA (1996) New methods and future trends in seismological volcano monitoring. In: Scarpa R, Tilling RI (eds) Monitoring and mitigation of volcano hazards. Springer
6. Chung TJ (2002) Computational fluid dynamics. Cambridge University Press, Cambridge
7. Clift R, Grace JR, Weber ME (1978) Bubbles, drops and particles. Dover Publications
8. Evans R (1979) The nature of the liquid-vapour interface and other topics in the statistical mechanics of non-uniform, classical fluids. Adv Phys 28(2):143–200
9. James MR, Lane SJ, Chouet B, Gilbert JS (2004) Pressure changes associated with the ascent and bursting of gas slugs in liquid-filled vertical and inclined conduits. J Volc Geotherm Res 129:61–82
10. James MR, Lane SJ, Chouet BA (2006) Gas slug ascent through changes in conduit diameter: Laboratory insight into a volcano-seismic source process in low-viscosity magmas. J Geophys Res 111
11. Jaupart C (2000) Magma ascent at shallow levels. In: Sigurdsson H (ed) Encyclopedia of Volcanoes. Academic Press
12. Jaupart C, Vergniolle S (1988) Laboratory models of hawaiian and strombolian eruptions. Nature 331:58–60
13. Kirchdorfer M (1999) Analysis and quasistatic fe modeling of long period impulsive events associated with explosions at stromboli volcano (italy). Annali di Geofisica 42(3):379–390
14. Ohminato T, Chouet BA, Dawson P, Kedar S (1998) Waveform inversion of very long period impulsive signals associated with magmatic injection beneath kilauea volcano, hawaii. J Geophys Res 103(B10):23839–23862
15. Pooleay CM, Kuksenok O, Balazs AC (2005) Convection-driven pattern in phase-separating binary fluids. Phys Rev E 71:030501(R)
16. Ripepe M, Poggi P, Braun T, Gordeev E (1996) Infrasonic waves and volcanic tremor at stromboli. Geophys Res Lett 23(2):181–184
17. Seyfried R, Freundt A (2000) Experiments on conduit flow and eruption behaviour of basaltic volcanic eruptions. J Geophys Res B10(B10):23727–23740
18. Sparks RSJ (1978) The dynamics of bubble formation and growth in magmas: a review and analysis. J Volc Geotherm Res 3:137–186
19. Strikwerda JC (2004) Finite difference schemes and partial differential equations. SIAM, Philadelphia
20. Succi S (2001) The Lattice Boltzmann Equation for fluid dynamics and beyond. In: Numerical mathematics and scientific computation. Oxford University Press, Oxford
21. Swift MR, Orlandini E, Osborn WR, Yeomans JM (1996) Lattice Boltzmann simulations of liquid-gas and binary fluid system. Phys Rev E 54(5):5041–5052
22. Tomiyama A, Takagi S, Matsumoto Y (1999) Numerical simulation of bubble flows using interface tracking and bubble tracking method. Trans Model Simul 23

23. Vergniolle S, Brandeis G (1996) Strombolian explosions 1. a large bubble breaking at the surface of a lava column as a source of sound. J Geophys Res 101(B9):20433–20447
24. Vergniolle S, Mangan M (2000) Hawaiian and strombolian eruptions. In: Sigurdsson H (ed) Encyclopedia of Volcanoes. Academic Press
25. Xu A, Gonnella G, Lamura A (2003) Phase-separating binary fluids under oscillatory shear. Phys Rev E 67
26. Yue P, Feng JJ, Liu C, Shen J (2004) A diffuse-interface method for simulating two-phase flows of complex fluids. J Fluid Mech 515:293–317

Books and Reviews

Brennen CE (2005) Fundamentals of multiphase flow. Cambridge University Press, Cambridge
Scarpa R, Tilling RI (eds) (1996) Monitoring and mitigation of volcano hazards. Springer
Sigurdsson H, Bruce Houghton BF, McNutt SR, Rymer H, Stix J (eds) (2000) In: Encyclopedia of Volcanoes. Academic Press
Tannehill JC, Anderson DA, Pletcher RH (1997) Computational fluid mechanics and heat transfer, second edn. Taylor and Francis, London
Zobin V (2003) Introduction to volcanic seismology. Elsevier Science

Smooth Ergodic Theory

Amie Wilkinson
Northwestern University, Evanston, USA

Article Outline

Glossary
Definition of the Subject
Introduction
The Volume Class
The Fundamental Questions
Lebesgue Measure and Local Properties of Volume
Ergodicity of the Basic Examples
Hyperbolic Systems
Beyond Uniform Hyperbolicity
The Presence of Critical Points and Other Singularities
Future Directions
Bibliography

Glossary

Conservative, dissipative Conservative dynamical systems (on a compact phase space) are those that preserve a finite measure equivalent to volume. Hamiltonian dynamical systems are important examples of conservative systems. Systems that are not conservative are called dissipative. Finding physically meaningful invariant measures for dissipative maps is a central object of study in smooth ergodic theory.

Distortion estimate A key technique in smooth ergodic theory, a distortion estimate for a smooth map f gives a bound on the variation of the jacobian of f^n in a given region, for n arbitrarily large. The jacobian of a smooth map at a point x is the absolute value of the determinant of derivative at x, measured in a fixed Riemannian metric. The jacobian measures the distortion of volume under f in that metric.

Hopf argument A technique developed by Eberhard Hopf for proving that a conservative diffeomorphism or flow is ergodic. The argument relies on the Ergodic Theorem for invertible transformations, the density of continuous functions among integrable functions, and the existence of stable and unstable foliations for the system. The argument has been used, with various modifications, to establish ergodicity for hyperbolic, partially hyperbolic and nonuniformly hyperbolic systems.

Hyperbolic A compact invariant set $\Lambda \subset M$ for a diffeomorphism $f: M \to M$ is hyperbolic if, at every point in Λ, the tangent space splits into two subspaces, one that is uniformly contracted by the derivative of f, and another that is uniformly expanded. Expanding maps and Anosov diffeomorphisms are examples of globally hyperbolic maps. Hyperbolic diffeomorphisms and flows are the archetypical smooth systems displaying chaotic behavior, and their dynamical properties are well-understood. Nonuniform hyperbolicity and partial hyperbolicity are two generalizations of hyperbolicity that encompass a broader class of systems and display many of the chaotic features of hyperbolic systems.

Sinai–Ruelle–Bowen (SRB) measure The concept of SRB measure is a rigorous formulation of what it means for an invariant measure to be "physically meaningful". An SRB measure attracts a large set of orbits into its support, and its statistical features are reflected in the behavior of these attracted orbits.

Definition of the Subject

Smooth ergodic theory is the study of the statistical and geometric properties of measures invariant under a smooth transformation or flow. The study of smooth ergodic theory is as old as the study of abstract ergodic theory, having its origins in Bolzmann's Ergodic Hypothesis in the late 19th Century. As a response to Boltzmann's hypothesis, which was formulated in the context of Hamiltonian Mechanics, Birkhoff and von Neumann defined er-

godicity in the 1930s and proved their foundational ergodic theorems. The study of ergodic properties of smooth systems saw an advance in the work of Hadamard and E. Hopf in the 1930s their study of geodesic flows for negatively curved surfaces. Beginning in the 1950s, Kolmogorov, Arnold and Moser developed a perturbative theory producing obstructions to ergodicity in Hamiltonian systems, known as Kolmogorov–Arnold–Moser (KAM) Theory. Beginning in the 1960s with the work of Anosov and Sinai on hyperbolic systems, the study of smooth ergodic theory has seen intense activity. This activity continues today, as the ergodic properties of systems displaying weak forms of hyperbolicity are further understood, and KAM theory is applied in increasingly broader contexts.

Introduction

This entry focuses on the basic arguments and principles in smooth ergodic theory, illustrating with simple and straightforward examples. The classic texts [1,2] are a good supplement.

The discussion here sidesteps the topic of ► Kolmogorov–Arnold–Moser (KAM) Theory, which has played an important role in the development of smooth ergodic theory. For reasons of space, detailed discussion of several active areas in smooth ergodic theory is omitted, including: higher mixing properties (Kolmogorov, Bernoulli, etc.), finer statistical properties (fast decay of correlations, Central Limit Theorem, large deviations), smooth thermodynamic formalism (transfer operators, pressure, dynamical zeta functions, etc.), the smooth ergodic theory of random dynamical systems, as well as any mention of infinite invariant measures. The text [3] covers many of these topics, and the texts [4,5,6] treat random smooth ergodic theory in depth. An excellent discussion of many of the recent developments in the field of smooth ergodic theory is [7].

This entry assumes knowledge of the basic concepts in ergodic theory and of basic differential topology. The texts [8] and [9] contain the necessary background.

The Volume Class

For simplicity, assume that M is a compact, boundaryless C^∞ Riemannian manifold, and that $f\colon M \to M$ is an orientation-preserving, C^1 map satisfying $m(D_x f) > 0$, for all $x \in M$, where

$$m(D_x f) = \inf_{v \in T_x M, \|v\|=1} \|D_x f(v)\| \,.$$

If f is a diffeomorphism, then this assumption is automatically satisfied, since in that case $m(D_x f) = \|D_{f(x)} f^{-1}\|^{-1} > 0$. For non-invertible maps, this assumption is essential in much of the following discussion. The Inverse Function Theorem implies that any map f satisfying these hypotheses is a covering map of positive degree $d \geq 1$.

These assumptions will avoid the issues of infinite measures and the behavior of f near critical points and singularities of the derivative. For most results discussed in this entry, this assumption is not too restrictive. The existence of critical points and other singularities is, however, a complication that cannot be avoided in many important applications. The ergodic-theoretic analysis of such examples can be considerably more involved, but contains many of the elements discussed in this entry. The discussion in Sect. "Beyond Uniform Hyperbolicity" indicates how some of these additional technicalities arise and can be overcome. For simplicity, the discussion here is confined almost exclusively to discrete time evolution. Many, though not all, of the the results mentioned here carry over to flows and semiflows using, for example, a cross-section construction (see Chap. 1 in [2]).

Every smooth map $f\colon M \to M$ satisfying these hypotheses preserves a natural measure *class*, the measure class of a finite, smooth Riemannian volume on M. Fix such a volume ν on M. Then there exists a continuous, positive *jacobian* function $x \mapsto \mathrm{jac}_x f$ on M, with the property that for every sufficiently small ball $B \subset M$, and every measurable set $A \subset B$ one has:

$$\nu(f(A)) = \int_B \mathrm{jac}_x f \, \mathrm{d}\nu(x) \,.$$

The jacobian of f at x is none other than the absolute value of the determinant of the derivative $D_x f$ (measured in the given Riemannian metric). To see that the measure class of ν is preserved by f, observe that the Radon–Nikodym derivative $\frac{df_*\nu}{d\nu}(x)$ at x is equal to $\sum_{y \in f^{-1}(x)} (\mathrm{jac}_y f)^{-1} > 0$. Hence $f_*\nu$ is equivalent to ν, and f preserves the measure class of ν.

In many contexts, the map f has a natural *invariant* measure in the measure class of volume. In this case, f is said to be *conservative*. One setting in which a natural invariant smooth measure appears is Hamiltonian dynamics. Any solution to Hamilton's equations preserves a smooth volume called the *Liouville measure*. Furthermore, along the invariant, constant energy hypermanifolds of a Hamiltonian flow, the Liouville measure decomposes smoothly into invariant measures, each of which is equivalent to the induced Riemannian volume. In this way, many systems of physical or geometric origin, such as billiards, geodesic flows, hard sphere gases, and evolution of the n-body problem give rise to smooth conserva-

tive dynamical systems. See ► Dynamics of Hamiltonian Systems.

Note that even though f preserves a smooth measure class, it might not preserve any measure in that measure class. Consider, for example, a diffeomorphism $f\colon S^1 \to S^1$ of the circle with exactly two fixed points, p and q, $f'(p) > 1 > f'(q) > 0$. Let μ be an f-invariant probability measure. Let I be a neighborhood of p. Then $\bigcap_{n=1}^{\infty} f^{-n}(I) = \{p\}$, but on the other hand, $\mu(f^{-n}(I)) = \mu(I) > 0$, for all n. This implies that $\mu(\{p\}) > 0$, and so μ does not lie in the measure class of volume. This is an example of a dissipative map. A map f is called *dissipative* if every f-invariant measure with full support has a singular part with respect to volume. As was just seen, if a diffeomorphism f has a periodic sink, then f is dissipative; more generally, if a diffeomorphism f has a periodic point p of period k such that $\mathrm{jac}_p f^k \neq 1$, then f is dissipative.

The Fundamental Questions

For a given smooth map $f\colon M \to M$, there are the following fundamental questions.

1. Is f conservative? That is, does there exist an invariant measure in the class of volume? If so, is it unique?
2. When f is conservative, what are its statistical properties? Is it ergodic, mixing, a K-system, Bernoulli, etc.? Does it obey a Central Limit Theorem, fast decay of correlations, large deviations estimates, etc.?
3. If f is dissipative, does there exist an invariant measure, not in the class of volume, but (in some sense) natural with respect to volume? What are the statistical properties of such a measure, if it exists?

There are several plausible ways to "answer" these questions. One might fix a given map f of interest and ask these questions for that specific f. What tends to happen in the analysis of a single map f is that either:

- the question can be answered using "soft" methods, and so the answer applies not only to f but to perturbations of f, or even to *generic* or *typical* f inside a class of maps; or
- the proof requires "hard" analysis or precise asymptotic information and cannot possibly be answered for a specific f, but can be answered for a large set of f_t in a typical (or given) parametrized family $\{f_t\}_{t\in(-1,1)}$ of smooth maps containing $f = f_0$.

Both types of results appear in the discussion that follows.

Lebesgue Measure and Local Properties of Volume

Locally, any measure in the measure class of volume is, after a smooth change of coordinates, equivalent to Lebesgue measure in $\mathbb{R}^n$. In fact, more is true: Moser's Theorem implies that locally any Riemannian volume is, after a smooth change of coordinates, *equal* to Lebesgue measure in $\mathbb{R}^n$. Hence to study many of the local properties of volume, it suffices to study the same properties for Lebesgue measure.

One of the basic properties of Lebesgue measure is that every set of positive Lebesgue measure can be approximated arbitrarily well in measure from the outside by an open set, and from the inside by a compact set. A consequence of this property, of fundamental importance in smooth ergodic theory, is the following statement.

Fundamental Principle #1: Two disjoint, positive Lebesgue measure sets cannot mix together uniformly at all scales.

As an illustration of this principle, consider the following elementary exercise in measure theory. First, some notation. If ν is a measure and A and B are ν-measurable sets with $\nu(B) > 0$, the *density of A in B* is defined by:

$$\nu(A\colon B) = \frac{\nu(A \cap B)}{\nu(B)} .$$

Proposition 1 *Let $\mathcal{P}_1, \mathcal{P}_2, \ldots$ be sequence of (mod 0) finite partitions of the circle S^1 into open intervals, with the properties: a) any element of $\mathcal{P}_n$ is a (mod 0) union of elements of $\mathcal{P}_{n+1}$, and b) the maximum diameter of elements of $\mathcal{P}_n$ tends to 0 as $n \to \infty$.*

Let A be any set of positive Lebesgue measure in S^1. Then there exists a sequence of intervals $I_1, I_2, \ldots$, with $I_n \in \mathcal{P}_n$ such that:

$$\lim_{n\to\infty} \lambda(A\colon I_n) = 1 .$$

Proof Assume that Lebesgue measure has been normalized so that $\lambda(S^1) = 1$. Fix a (mod 0) cover of $S^1 \setminus A$ by pairwise disjoint elements $\{J_i\}$ of the union $\bigcup_{n=1}^{\infty} \mathcal{P}_n$ with the properties:

$$\lambda(J_1) \geq \lambda(J_2) \geq \cdots , \text{ and}$$

$$\lambda\left(\bigcup_{i=1}^{\infty} J_i\right) = \sum_{i=1}^{\infty} \lambda(J_i) < 1 .$$

For $n \in \mathbb{N}$, let U_n be the union of all the intervals J_i that are contained in $\mathcal{P}_n$, and let $V_n = \bigcup_{i=1}^{n} U_n$. This defines an increasing sequence of natural numbers $i_1 = 1 < i_2 < i_3 < \cdots$ such that $U_n = \bigcup_{i=i_n}^{i_{n+1}-1} J_n$ and $V_n = \bigcup_{i=1}^{i_{n+1}-1} J_n$.

For each n, the interval I_n will be chosen to be an element $\mathcal{P}_n$, disjoint from V_n, in which the density of $\bigcup_{i=n+1}^{\infty} U_i$ is very small (approaching 0 as $n \to \infty$). Since $(S^1 \setminus A) \cap I_n$ is contained in $\bigcup_{i=n+1}^{\infty} U_i$, this choice of I_n will ensure that the density of A in I_n is large (approaching 1 as $n \to \infty$).

To make this choice of I_n, note first that the density of $\bigcup_{i=n+1}^{\infty} U_i$ inside of $S^1 \setminus V_n$ is:

$$\frac{\lambda(\bigcup_{i=n+1}^{\infty} U_i)}{\lambda(S^1 \setminus V_n)} = \frac{\sum_{i=i_{n+1}}^{\infty} \lambda(J_i)}{1 - \sum_{i=1}^{i_{n+1}-1} \lambda(J_i)} = a_n \,.$$

Note that, since $\sum_{i=1}^{\infty} \lambda(J_i)1$, one has $a_n \to 0$ as $n \to \infty$. Since the density of $\bigcup_{i=n+1}^{\infty} U_i$ inside of $S^1 \setminus V_n$ is at most a_n, there is an interval I_n in $\mathcal{P}_n$, disjoint from V_n, such that the density of $\bigcup_{i=n+1}^{\infty} U_i$ inside of I_n is at most a_n. Then

$$\lim_{n\to\infty} \lambda(A\colon I_n) \geq \lim_{n\to\infty} 1 - a_n = 1 \,.$$

□

In smooth ergodic theory, it is often useful to use a variation on Proposition 1 (generally, in higher dimensions) in which the partitions $\mathcal{P}_n$ are nested, dynamically-defined partitions. A simple application of this method can be used to prove that the doubling map on the circle is ergodic with respect to Lebesgue measure, which is done in Sect. "Lebesgue Measure and Local Properties of Volume".

Notice that this proposition does not claim that the intervals I_n are nested. If one imposes stronger conditions on the partitions $\mathcal{P}_n$, then one can draw stronger conclusions.

A very useful theorem in this respect is the Lebesgue Density Theorem. A point $x \in M$ is a *Lebesgue density point* of a measurable set $X \subseteq M$ if

$$\lim_{r\to 0} m(X\colon B_r(x)) = 1 \,,$$

where $B_r(x)$ is the Riemannian ball of radius r centered at x. Notice that the notion of Lebesgue density point depends only on the smooth structure of M, because any two Riemannian metrics have the same Lebesgue density points. The Lebesgue Density Theorem states that if A is a measurable set and $\widehat{A}$ is the set of Lebesgue density points of A, then $m(A \,\Delta\, \widehat{A}) = 0$.

Ergodicity of the Basic Examples

This section contains proofs of the ergodicity of two basic examples of conservative smooth maps: irrational rotations on the circle and the doubling map on the circle. See ► Ergodic Theory: Basic Examples and Constructions for a more detailed description of these maps. These proofs serve as an elementary illustration of some of the fundamental techniques and principles in smooth ergodic theory.

Rotations on the circle. Denote by S^1 the circle $\mathbb{R}/\mathbb{Z}$, which is an additive group, and by λ normalized Lebesgue-Haar measure on S^1. Fix a real number $\alpha \in \mathbb{R}$. The rotation $R_\alpha\colon S^1 \to S^1$ is the translation defined by $R_\alpha(x) = x + \alpha$. Since translations preserve Lebesgue-Haar measure, the map R_α is conservative. Note that R_α is a diffeomorphism and an isometry with respect to the canonical flat metric (length) on S^1.

Proposition 2 *If $\alpha \notin \mathbb{Q}$, then the rotation $R_\alpha\colon S^1 \to S^1$ is ergodic with respect to Lebesgue measure.*

Proof Let A be an R_α-invariant set in S^1, and suppose that $0 < \lambda(A) < 1$. Denote by A^c the complement of A in S^1. Fix $\varepsilon > 0$. Proposition 1 implies that there exists an interval $I \subset S^1$ such that the density of A in I is large: $\lambda(A\colon I) > 1 - \varepsilon$. Similarly, one may choose an interval J such that $\lambda(A^c\colon J) > 1 - \varepsilon$. Without loss of generality, one may choose I and J to have the same length. Since α is irrational, R_α has a dense orbit, which meets the interval I. Since R_α is an isometry, this implies that there is an integer n such that $\lambda(R_\alpha^n(I) \,\Delta\, J) < \varepsilon\lambda(I)$. Since $\lambda(I) = \lambda(J)$, this readily implies that $|\lambda(A\colon R_\alpha^n(I)) - \lambda(A\colon J)| < \varepsilon$. Also, since A is invariant, and R_α is invertible and preserves measure, one has:

$$\lambda(A\colon R_\alpha^n(I)) = \lambda(R_\alpha^n(A)\colon R_\alpha^n(I)) = \lambda(A\colon I) > 1 - \varepsilon \,.$$

But for ε sufficiently small, this contradicts the facts that $\lambda(A\colon J) = 1 - \lambda(A^c\colon J) < \varepsilon$ and $|\lambda(A\colon R_\alpha^n(I)) - \lambda(A\colon J)| < \varepsilon$. □

Note that this is not a proof of the strongest possible statement about R_α (namely, minimality and unique ergodicity). The point here is to show how "soft" arguments are often sufficient to establish ergodicity; this proof uses no more about R_α than the fact that it is a transitive isometry. Hence the same argument shows:

Theorem 1 *Let $f\colon M \to M$ be a transitive isometry of a Riemannian manifold M. Then f is ergodic with respect to Riemannian volume.*

One can isolate from this proof a useful principle:

Fundamental Principle #2: Isometries preserve Lebesgue density at all scales, for arbitrarily many iterates.

This principle implies, for example, that a smooth action by a compact Lie group on M is ergodic along typical (nonsingular) orbits. This principle is also useful in studying area-preserving flows on surfaces and, in a refined form, unipotent flows on homogeneous spaces. In

the case of surface flows, ergodicity questions can be reduced to a study of interval exchange transformations. See the entry ► Ergodic Theory: Basic Examples and Constructions for a detailed discussion of interval exchange transformations and flows on surfaces. ► Ergodic Theory, Introduction to contains detailed information on unipotent flows.

Doubling map on the circle. Let $T_2 : S^1 \to S^1$ be the doubling map defined by $T_2(x) = 2x$. Then T_2 is a degree-2 covering map and endomorphism of S^1 with constant jacobian $\mathrm{jac}_x T_2 \equiv 2$. Since $\frac{d(T_2)_*\lambda}{d\lambda} = \frac{1}{2} + \frac{1}{2} = 1$, T_2 preserves Lebesgue-Haar measure. The doubling map is the simplest example of a hyperbolic dynamical system, a topic treated in depth in the next section.

As with the previous example, the focus here is on the property of ergodicity. It is again possible to prove much stronger results about T_2, such as Bernoullicity, by other methods. Instead, here is a soft proof of ergodicity that will generalize readily to other contexts.

Proposition 3 *The doubling map $T_2 : S^1 \to S^1$ is ergodic with respect to Lebesgue measure.*

Proof Let A be a T_2-invariant set in S^1 with $\lambda(A) > 0$. Let $p \in S^1$ be the fixed point of T_2, so that $T_2(p) = p$. For each $n \in \mathbb{N}$, the preimages of p under T_2^{-n} define a (mod 0) partition $\mathcal{P}_n$ into 2^n open intervals of length 2^{-n}; the elements of $\mathcal{P}_n$ are the connected components of $S^1 \setminus T_2^{-n}(\{p\})$. Note that the sequence of partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots$ is nested, in the sense of Proposition 1. Restricted to any interval $J \in \mathcal{P}_n$, the map $T_2{}^n$ is a diffeomorphism onto $S^1 \setminus \{p\}$ with constant jacobian $\mathrm{jac}_x(T_2^n) = (T_2^n)'(x) = 2^n$.

Since A is invariant, it follows that $T_2^{-n}(A) = A$. Fix $\varepsilon > 0$. Proposition 1 implies that there exists an $n \in \mathbb{N}$ and an interval $J \in \mathcal{P}_n$ such that $\lambda(A : J) > 1 - \varepsilon$. Note that $T_2^n(A \cap J) \subset A$. But then

$$\begin{aligned}
\lambda(A) &\geq \lambda(T_2^n(A \cap J)) \\
&= \int_{A\cap J} \mathrm{jac}_x(T_2^n)\, d\lambda(x) \\
&= 2^n \lambda(A \cap J) \\
&= 2^n \lambda(A : J)\lambda(J) \\
&> 2^n(1-\varepsilon)\lambda(J) = 1 - \varepsilon\,.
\end{aligned}$$

Since ε was arbitrary, one obtains that $\lambda(A) = 1$. □

In this proof, the facts that the intervals in $\mathcal{P}_n$ have constant length 2^{-n} and that the jacobian of $T_2{}^n$ restricted to such an interval is constant and equal to 2^n are not essential. The key fact really used in this proof is the assertion that the ratio:

$$\frac{\lambda(T_2^n(A \cap J) : T_2^n(J))}{\lambda(A : J)}$$

is bounded, *independently of n*. In this case, the ratio is 1 for all n because T_2 has constant jacobian.

It is tempting to try to extend this proof to other expanding maps on the circle, for example, a C^1, λ-preserving map $f : S^1 \to S^1$ with $d_{C^1}(f, T_2)$ small. Many of the aspects of this proof carry through *mutatis mutandis* for such an f, save for one. A C^1-small perturbation of T_2 will in general no longer have constant jacobian, and the *variation* of the jacobian of f^n on a small interval can be (and often is) unbounded. The reason for this unboundedness is a lack of control of the modulus of continuity of f'. Hence this argument can fail for C^1 perturbations of T_2. On the other hand, the argument still works for C^2 perturbations of T_2, even when the jacobian is not constant.

The principle behind this fact can be loosely summarized:

Fundamental Principle #3: On controlled scales, iterates of C^2 expanding maps distort Lebesgue density in a controlled way.

This principle requires further explanation and justification, which will come in the following section. The C^2 hypothesis in this principle accounts for the fact that almost all results in smooth ergodic theory assume a C^2 hypothesis (or something slightly weaker).

Hyperbolic Systems

One of the most developed areas of smooth ergodic theory is in the study of hyperbolic maps and attractors. This section defines hyperbolic maps and attractors, provides examples, and investigates their ergodic properties. See [10,11] and ► Hyperbolic Dynamical Systems for a thorough discussion of the topological and smooth properties of hyperbolic systems.

A *hyperbolic structure* on a compact f-invariant set $\Lambda \subset M$ is given by a Df-invariant splitting $T_\Lambda M = E^u \oplus E^s$ of the tangent bundle over Λ and constants $C, \mu > 1$ such that, for every $x \in \Lambda$ and $n \in \mathbb{N}$:

$$\begin{aligned}
v \in E^u(x) &\implies \|D_x f^n(v)\| \geq C^{-1}\mu^n \|v\|\,, \\
\text{and} \quad v \in E^s(x) &\implies \|D_x f^n(v)\| \leq C\mu^{-n}\|v\|\,.
\end{aligned}$$

A hyperbolic attractor for a map $f : M \to M$ is given by an open set $U \subset M$ such that: $f(U) \subset \overline{U}$, and such that the set $\Lambda = \bigcap_{n\geq 0} f^n(U)$ carries a hyperbolic structure. The set Λ is called the *attractor*, and U is an *attracting region*. A map $f : M \to M$ is *hyperbolic* if M decomposes (mod 0)

into a finite union of attracting regions for hyperbolic attractors. Typically one assumes as well that the restriction of f to each attractor Λ_i is topologically transitive.

Every point p in a hyperbolic set Λ has smooth *stable manifold* $\mathcal{W}^s(p)$ and *unstable manifold* $\mathcal{W}^u(p)$, tangent, respectively, to the subspaces $E^s(p)$ and $E^u(p)$. The set $\mathcal{W}^s(p)$ is precisely the set of $q \in M$ such that $d(f^n(p), f^n(q))$ tends to 0 as $n \to \infty$, and it follows that $f(\mathcal{W}s(p)) = \mathcal{W}^s(f(p))$. When f is a diffeomorphism, the unstable manifold $\mathcal{W}^u(p)$ is uniquely defined and is the stable manifold of f^{-1}. When f is not invertible, local unstable manifolds exist, but generally are not unique. If Λ is a transitive hyperbolic attractor, then every unstable manifold of every point $p \in \Lambda$ is dense in Λ.

Examples of Hyperbolic Maps and Attractors

Expanding Maps The previous section mentioned briefly the C^r perturbations of the doubling map T_2. Such perturbations (as well as T_2 itself) are examples of *expanding maps*. A map $f: M \to M$ is *expanding* if there exist constants $\mu > 1$ and $C > 0$ such that, for every $x \in M$, and every nonzero vector $v \in T_x M$:

$$\|D_x f^n(v)\| \geq C\mu^n \|v\| ,$$

with respect to some (any) Riemannian metric on M. An expanding map is clearly hyperbolic, with $U = M$, E^s the trivial bundle, and $E^u = TM$. Any disk in M is a local unstable manifold for f.

Anosov Diffeomorphisms A diffeomorphism $f: M \to M$ is called *Anosov* if the tangent bundle splits as a direct sum $TM = E^u \oplus E^s$ of two Df-invariant subbundles, such that E^u is uniformly expanded and E^s is uniformly contracted by Df. Similarly, a flow $\varphi_t: M \to M$ is called Anosov if the tangent bundle splits as a direct sum $TM = E^u \oplus E^0 \oplus E^s$ of three $D\varphi_t$-invariant subbundles, such that E^0 is generated by $\dot{\varphi}$, E^u is uniformly expanded and E^s is uniformly contracted by $D\varphi_t$. Like expanding maps, an Anosov diffeomorphism is an Anosov attractor with $\Lambda = U = M$.

A simple example of a conservative Anosov diffeomorphism is a hyperbolic linear automorphism of the torus. Any matrix $A \in SL(n, \mathbb{Z})$ induces an automorphism of $\mathbb{R}^n$ preserving the integer lattice $\mathbb{Z}^n$, and so descends to an automorphism $f_A: T^n \to T^n$ of the n-torus $T^n = \mathbb{R}^n/\mathbb{Z}^n$. Since the determinant of A is 1, the diffeomorphism f_A preserves Lebesgue-Haar measure on T^n. In the case where none of the eigenvalues of A have modulus 1, the resulting diffeomorphism f_A is Anosov. The stablebundle E^s at $x \in T^n$ is the parallel translate to x of the sum of the contracted generalized eigenspaces of A, and the unstable bundle E^u at x is the translated sum of expanded eigenspaces.

In general, the invariant subbundles E^u and E^s of an Anosov diffeomorphism are integrable and tangent to a transverse pair of foliations $\mathcal{W}^u$ and $\mathcal{W}^s$, respectively (see, e. g [12]. for a proof of this). The leaves of $\mathcal{W}^s$ are uniformly contracted by f, and the leaves of $\mathcal{W}^u$ are uniformly contracted by f^{-1}. The leaves of these foliations are as smooth as f, but the tangent bundles to the leaves do not vary smoothly in the manifold. The regularity properties of these foliations play an important role in the ergodic properties of Anosov diffeomorphisms.

The first Anosov flows to be studied extensively were the geodesic flows for manifolds of negative sectional curvatures. As these flows are Hamiltonian, they are conservative. Eberhard Hopf showed in the 1930s that such geodesic flows for surfaces are ergodic with respect to Liouville measure [14]; it was not until the 1960s that ergodicity of all such flows was proved by Anosov [15]. The next section describes, in the context of Anosov diffeomorphisms, Hopf's method and important refinements due to Anosov and Sinai.

DA Attractors A simple way to produce a non-Anosov hyperbolic attractor on the torus is to start with an Anosov diffeomorphism, such as a linear hyperbolic automorphism, and deform it in a neighborhood of a fixed point, turning a saddle fixed point into a source, while preserving the stable foliation. If this procedure is carried out carefully enough, the resulting diffeomorphism is a dissipative hyperbolic diffeomorphism, called a *derived from Anosov (DA)* attractor. Other examples of hyperbolic attractors are the Plykin attractor and the solenoid. See [10].

Distortion Estimates

Before describing the ergodic properties of hyperbolic systems, it is useful to pause for a brief discussion of distortion estimates. Distortion estimates are behind almost every result in smooth ergodic theory. In the hyperbolic setting, distortion estimates are applied to the action of f on unstable manifolds to show that the volume distortion of f along unstable manifolds can be controlled for arbitrarily many iterates.

The example mentioned at the end of the previous section illustrates the ideas in a distortion estimate. Suppose that $f: S^1 \to S^1$ is a C^2 expanding map, such as a C^2 small perturbation of T_2. Then there exist constants $\mu > 1$ and $C > 0$ such that $(f^n)'(x) > C\mu^n$ for all x and n.

Let d be the degree of f. If I is a sufficiently small open interval in S^1, then for each n, $f^{-n}(I)$ is a union of d disjoint intervals. Furthermore, each of these intervals has diameter at most $C^{-1}\mu^{-n}$ times the diameter of I. It is now possible to justify the assertion in Fundamental Principle #3 in this context.

Lemma *There exists a constant $K \geq 1$ such that, for all $n \in \mathbb{N}$, and for all $x, y \in f^{-n}(I)$, one has:*

$$K^{-1} \leq \frac{(f^n)'(x)}{(f^n)'(y)} \leq K\,.$$

Proof Since f is C^2 and f' is bounded away from 0, the function $\alpha(x) = \log(f'(x))$ is C^1. In particular, α is Lipschitz continuous: there exists a constant $L > 0$ such that, for all $x, y \in S^1$, $|\alpha(x) - \alpha(y)| < Ld(x, y)$. For $n \geq 0$, let $\alpha_n(x) = \log((f^n)'(x))$. The Chain Rule implies that $\alpha_n(x) = \sum_{i=0}^{n-1} \alpha(f^i(x))$.

The expanding hypothesis on f implies that for all $x, y \in f^n(I)$ and for $i = 0, \ldots, n$, one has $d(f^i(x), f^i(y)) \leq C^{-1}\mu^{i-n} d(f^n(x), f^n(y)) \leq C^{-1}\mu^{i-n}$. Hence

$$\begin{aligned}
|\alpha_n(x) - \alpha_n(y)| &\leq \sum_{i=0}^{n-1} |\alpha(f^i(x)) - \alpha(f^i(y))| \\
&\leq L \sum_{i=0}^{n-1} d(f^i(x), f^i(y)) \\
&\leq L \sum_{i=0}^{n-1} C^{-1}\mu^{i-n} \\
&< LC^{-1}\mu^{-1}(1-\mu^{-1})^{-1}\,.
\end{aligned}$$

Setting $K = \exp(LC^{-1}\mu^{-1}(1-\mu^{-1})^{-1})$, one now sees that $(f^n)'(x)/(f^n)'(y)$ lies in the interval $[K^{-1}.K]$, proving the claim. □

In this distortion estimate, the function $\alpha\colon M \to \mathbb{R}$ is called a *cocycle*. The same argument applies to any Lipschitz continuous (or even Hölder continuous) cocycle.

Ergodicity of Expanding Maps

The ergodic properties of C^2 expanding maps are completely understood. In particular, every conservative expanding map is ergodic, and every expanding map is conservative. The proofs of these facts use Fundamental Principles #1 and 3 in a fairly direct way.

Every C^2 conservative expanding map is ergodic with respect to volume. The proof is a straightforward adaptation of the proof of Proposition 3 (see, e. g [2].). Here is a description of the proof for $M = S^1$. As remarked earlier, the proof of Proposition 3 adapts easily to a general expanding map $f\colon S^1 \to S^1$ once one shows that for every f-invariant set A, and every connected component J of $f^{-n}(S^1 \setminus \{p\})$, the quantity

$$\frac{\lambda(f^n(A \cap J)\colon f^n(J))}{\lambda(A\colon J)}$$

is bounded independently of n. This is a fairly direct consequence of the distortion estimate in Lemma 6.1 and is left as an exercise.

The same distortion estimates show that every C^2 expanding map is conservative, preserving a probability measure ν in the measure class of volume. Here is a sketch of the proof for the case $M = S^1$. To prove this, consider the push-forward $\lambda_n = f_*^n\lambda$. Then λ_n is equivalent to Lebesgue, and its Radon–Nikodym derivative $d\lambda_n \setminus d\lambda$ is the density function

$$\rho_n(x) = \sum_{y \in f^{-n}(x)} \frac{1}{\mathrm{jac}_y f^n}\,.$$

Since $f_*^n\lambda$ is a probability measure, it follows that $\int_{S^1} \rho_n \, \mathrm{d}\lambda = 1$. A simple argument using the distortion estimate above (and summing up over all d^n branches of f^{-n} at x) shows that there exists a constant $c \geq 1$ such that for all $x, y \in S^1$,

$$c^{-1} \leq \frac{\rho_n(x)}{\rho_n(y)} \leq c\,.$$

Since the integral of ρ_n is 1, the functions ρ_n are uniformly bounded away from 0 and ∞. It is easy to see that the measure $\nu_n = \frac{1}{n}\sum_{i=1}^n f_*^i\lambda$ has density $\frac{1}{n}\sum_{i=1}^n \rho_i$. Let ν be any subsequential weak* limit of ν_n; then ν is absolutely continuous, with density ρ bounded away from 0 and ∞. With a little more care, one can show that ρ is actually Lipschitz continuous.

As a passing comment, the ergodicity of ν and positivity of ρ imply that ν is the unique f-invariant measure absolutely continuous with respect to λ. With more work, one can show that ν is exact. See [2] for details.

Ergodicity of Conservative Anosov Diffeomorphisms

Like conservative C^2 expanding maps, conservative C^2 Anosov diffeomorphisms are ergodic. This subsection outlines a proof of this fact. Unlike expanding maps, however, Anosov diffeomorphisms need not be conservative. The subsection following this one describe a type of invariant measure that is "natural" with respect to volume, called a Sinai-Ruelle-Bowen (or SRB) measure. The central result for hyperbolic systems states that every hyperbolic attractor carries an SRB measure.

The Hopf Argument In the 1930s Hopf [14] proved that the geodesic flow for a compact, negatively-curved surface is ergodic. His method was to study the Birkhoff averages of continuous functions along leaves of the stable and unstable foliations of the flow. This type of argument has been used since then in increasingly general contexts, and has come to be known as the Hopf Argument.

The core of the Hopf Argument is very simple. To any $f: M \to M$ one can associate the *stable equivalence relation* $\sim_s$, where $x \sim_s y$ iff $\lim_{n\to\infty} d(f^n(x), f^n(y)) = 0$. Denote by $W^s(x)$ the stable equivalence class containing x. When f is invertible, one defines the *unstable equivalence relation* to be the stable equivalence relation for f^{-1}, and one denotes by $W^u(x)$ the unstable equivalence class containing x.

The first step in the Hopf Argument is to show that Birkhoff averages for continuous functions are constant along stable and unstable equivalence classes. Let $\phi: M \to \mathbb{R}$ be an integrable function, and let

$$\overline{\phi} = \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \phi \circ f^i \,. \tag{1}$$

Observe that if ϕ is continuous, then for every $x \in M$ and $x' \in W^s(x)$, $\lim_{n\to\infty} |\phi(f^n(x)) - \phi(f^i(x'))| = 0$. It follows immediately that $\overline{\phi}_f(x) = \overline{\phi}_f(x')$. In particular, if the limit in (1) exists at x, then it exists and is constant on $W^s(x)$.

Fundamental Principle #4: Birkhoff averages of continuous functions are constant along stable equivalence classes.

The next step of Hopf's argument confines itself to the situation where f is conservative and Anosov. In this case, f is invertible, the stable equivalence classes are precisely the leaves of the stable foliation $\mathcal{W}^s$, and the unstable equivalence classes are the leaves of the unstable foliation $\mathcal{W}^u$. Since f is conservative, the Ergodic Theorem implies that for every L^2 function ϕ, the function $\overline{\phi}_f$ is equal (mod 0) to the projection of ϕ onto the f-invariant functions in L^2. Since this projection is continuous, and the continuous functions are dense in L^2, to prove that f is ergodic, it suffices to show that the projection of any continuous function is trivial. That is, it suffices to show that for every continuous ϕ, the function $\overline{\phi}_f$ is constant (a.e.).

To this end, let $\phi: M \to \mathbb{R}$ be continuous. Since the f-invariant functions coincide with the f^{-1}-invariant functions, one obtains that $\overline{\phi}_f = \overline{\phi}_{f^{-1}}$ a.e. The previous argument shows $\overline{\phi}_f$ is constant along $\mathcal{W}^s$-leaves and $\overline{\phi}_{f^{-1}}$ is constant along $\mathcal{W}^u$-leaves. The desired conclusion is that $\overline{\phi}_f$ is a.e. constant. It suffices to show this in a local chart, since the manifold M is connected. In a local chart, after a smooth change of coordinates, one obtains a pair of transverse foliations $\mathcal{F}_1, \mathcal{F}_2$ of the cube $[-1,1]^n$ by disks, and a measurable function $\psi: [-1,1]^n \to \mathbb{R}$ that is constant along the leaves of $\mathcal{F}_1$ and constant along the leaves of $\mathcal{F}_2$.

When the foliations $\mathcal{F}_1$ and $\mathcal{F}_2$ are smooth (at least C^1), one can perform a further smooth change of coordinates so that $\mathcal{F}_1$ and $\mathcal{F}_2$ are transverse coordinate subspace foliations. In this case, Fubini's theorem implies that any measurable function that is constant along two transverse coordinate foliations is a.e. constant. This completes the proof in the case that the foliations $\mathcal{W}^s$ and $\mathcal{W}^u$ are smooth. In Hopf's original argument, the stable and unstable foliations were assumed to be C^1 foliations (a hypotheses satisfied in the examples he considered, due to low-dimensionality. See also [16], where a pinching condition on the curvature, rather than low dimensionality, implies this C^1 condition on the foliations.)

Absolute Continuity For a general Anosov diffeomorphism or flow, the stable and unstable foliations are not C^1, and so the final step in Hopf's orginal argument does not apply. The fundamental advance of Anosov and Anosov-Sinai was to prove that the stable and unstable foliations of an Anosov diffeomorphism (conservative or not) satisfy a weaker condition than smoothness, called *absolute continuity*. For conservative systems, absolute continuity is enough to finish Hopf's argument, proving that every C^2 conservative Anosov diffeomorphism is ergodic [15,17].

For a definition and careful discussion of absolute continuity of a foliation $\mathcal{F}$, see [13]. Two consequences of the absolute continuity of $\mathcal{F}$ are:

1. (AC1) If $A \subset M$ is any measurable set, then

$$\lambda(A) = 0 \iff \lambda_{\mathcal{F}(x)}(A) = 0 \,, \quad \text{for} \quad \lambda - \text{a.e. } x \in M \,,$$

   where $\lambda_{\mathcal{F}(x)}$ denotes the induced Riemannian volume on the leaf of $\mathcal{F}$ through x.
2. (AC2) If τ is any small, smooth disk transverse to a local leaf of $\mathcal{F}$, and $T \subset \tau$ is a 0-set in τ (with respect to the induced Riemannian volume on τ), then the union of the $\mathcal{F}$ leaves through points in T has Lebesgue measure 0 in M.

The proof that $\mathcal{W}^s$ and $\mathcal{W}^u$ are absolutely continuous has a similar flavor to the proof that an expanding map has a unique absolutely continuous invariant measure (although the cocycles involved are Hölder continuous, rather than Lipschitz), and the facts are intimately related.

With absolute continuity of the stable and unstable foliations in hand, one can now prove:

Theorem 2 (Anosov) *Let f be a C^2, conservative Anosov diffeomorphism. Then f is ergodic.*

Proof By the Hopf Argument, it suffices to show that if ψ^s and ψ^u are L^2 functions with the following properties:

1. ψ^s is constant along leaves of $\mathcal{W}^s$,
2. ψ^u is constant along leaves of $\mathcal{W}^u$, and
3. $\psi^s = \psi^u$ a.e.,

then ψ^s (and so ψ^u as well) is constant a.e.

This is proved using the absolute continuity of $\mathcal{W}^u$ and $\mathcal{W}^s$. Since M is connected, one may argue this locally. Let G be the full measure set of $p \in M$ such that $\psi^s = \psi^u$. Absolute continuity of $\mathcal{W}^s$ (more precisely, consequence (AC1) of absolute continuity described above) implies that for almost every $p \in M$, G has full measure in $\mathcal{W}^s(p)$. Pick such a p. Then for almost every $q \in \mathcal{W}^s(p)$, $\psi^s(q) = \psi^u(p)$; defining G' to be the union over all $q \in \mathcal{W}^s(p) \cap G$ of $\mathcal{W}^u(q)$, one obtains that ψ^s is constant on $G \cap G'$. But now, since $\mathcal{W}^s(p) \cap G$ has full measure in $\mathcal{W}^s(p)$, the absolute continuity of $\mathcal{W}^u$ (consequence (AC2) above) implies that G' has full measure in a neighborhood of p. Hence ψ^s is a.e. constant in a neighborhood of p, completing the proof. □

SRB Measures

In the absence of a smooth invariant measure, it is still possible for a map to have an invariant measure that behaves naturally with respect to volume. In computer simulations one observes such measures when one picks a point x at random and plots many iterates of x; in many systems, the resulting picture is surprisingly insensitive to the initial choice of x. What appears to be happening in these systems is that the trajectory of almost every x in an open set U is converging to the support of a singular invariant probability measure μ. Furthermore, for any open set V, the proportion of forward iterates of x spent in V appears to converge to $\mu(V)$ as the number of iterates tends to ∞.

In the 1960s and 70s, Sinai, Ruelle and Bowen rigorously established the existence of these physically observable measures for hyperbolic attractors [18,19,20]. Such measures are now known as Sinai-Ruelle-Bowen (SRB) measures, and have been shown to exist for non-hyperbolic maps with some hyperbolic features. This subsection describes the construction of SRB measures for hyperbolic attractors.

An f-invariant probability measure μ is called an *SRB (or physical) measure* if there exists an open set $U \subset M$ containing the support of μ such that, for every continuous function $\phi : M \to \mathbb{R}$ and λ-a.e. $x \in U$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \phi(f^i(x)) = \int_M \phi \, \mathrm{d}\mu \, .$$

The maximal open set U with this property is called the *basin* of f. To exclude the possibility that the SRB measure is supported on a periodic sink, one often adds the condition that at least one of the Lyapunov exponents of f with respect μ is positive. Other definitions of SRB measure have been proposed (see [21]). Note that every ergodic absolutely continuous invariant measure with positive density in an open set is an SRB measure. Note also that an SRB measure for f is not in general an SRB measure for f^{-1}, unless f preserves an ergodic absolutely continuous invariant measure.

Every transitive Anosov diffeomorphism carries a unique SRB measure. To prove this, one defines a sequence of probability measures ν_n on M as follows. Fix a point $p \in M$, and define ν_0 to be the normalized restriction of Riemannian volume to a ball B^u in $\mathcal{W}^u(p)$. Set $\nu_n = \frac{1}{n}\sum_{i=1}^{n} f_*^i \nu_0$. Distortion estimates show that the density of ν_n on its support inside $\mathcal{W}^u$ is bounded, above and below, independently of n. Passing to a subsequential weak* limit, one obtains a probability measure ν on M with bounded densities on $\mathcal{W}^u$-leaves.

To show that ν is an SRB measure, choose a point $q \in M$ in the support of ν. Since ν has positive density on unstable manifolds, almost every point in a neighborhood of q in $\mathcal{W}^u(q)$ is a regular point for f (that is, a point where the forward Birkhoff averages of every continuous function exist). A variation on the Hopf Argument, using the absolute continuity of $\mathcal{W}^s$, shows that ν is an ergodic SRB measure.

A similar argument shows that every transitive hyperbolic attractor admits an ergodic SRB measure. In fact this SRB measure has much stronger mixing properties, namely, it is Bernoulli. To prove this, one first constructs a Markov partition [22] conjugating the action of f to a Bernoulli shift. This map sends the SRB measure to a Gibbs state for a mixing Markov shift (see ► Pressure and Equilibrium States in Ergodic Theory). A result that subsumes all of the results in this section is:

Theorem 3 (Sinai, Ruelle, Bowen) *Let $\Lambda \subset M$ be a transitive hyperbolic attractor for a C^2 map $f : M \to M$. Then f has an ergodic SRB measure μ supported on Λ. Moreover: the disintegration of μ along unstable manifolds of Λ is equivalent to the induced Riemannian volume, the Lyapunov exponents of μ are all positive, and μ is Bernoulli.*

Beyond Uniform Hyperbolicity

The methods developed in the smooth ergodic theory of hyperbolic maps have been extended beyond the hyperbolic context. Two natural generalizations of hyperbolicity are:

- partial hyperbolicity, which requires uniform expansion of E^u and uniform contraction of E^s, but allows central directions at each point, in which the expansion and contraction is dominated by the behavior in the hyperbolic directions; and
- nonuniform hyperbolicity, which requires hyperbolicity along almost every orbit, but allows the expansion of E^u and the contraction of E^s to weaken near the exceptional set where there is no hyperbolicity.

This section discusses both generalizations.

Partial Hyperbolicity

Brin and Pesin [23] and independently Pugh and Shub [24] first examined the ergodic properties of partially hyperbolic systems soon after the work of Anosov and Sinai on hyperbolic systems. One says that a diffeomorphism $f\colon M \to M$ of a compact manifold M is *partially hyperbolic* if there is a nontrivial, continuous splitting of the tangent bundle, $TM = E^{\mathrm{s}} \oplus E^{\mathrm{c}} \oplus E^{\mathrm{u}}$, invariant under Df, such that E^{s} is uniformly contracted, E^{u} is uniformly expanded, and E^c is dominated, meaning that for some $n \geq 1$ and for all $x \in M$:

$$\|D_x f^n|_{E^s}\| < m(D_x f^n|_{E^c}) \leq \|D_x f^n|_{E^c}\| < m(D_x f^n|_{E^u}).$$

Partial hyperbolicity is a C^1-open condition: any diffeomorphism sufficiently C^1-close to a partially hyperbolic diffeomorphism is itself partially hyperbolic. For an extensive discussion of examples of partially hyperbolic dynamical systems, see the survey article [25] and the book [26]. Among these examples are: the time-1 map of an Anosov flow, the frame flow for a compact manifold of negative sectional curvature, and many affine transformations of compact homogeneous spaces. All of these examples preserve the volume induced by a Riemannian metric on M.

As in the Anosov case, the stable and unstable bundles E^s and E^u of a partially hyperbolic diffeomorphism are tangent to foliations, again denoted by $\mathcal{W}^s$ and $\mathcal{W}^u$ respectively [23]. Brin-Pesin and Pugh-Shub proved that these foliations are absolutely continuous.

A partially hyperbolic diffeomorphism $f\colon M \to M$ is *accessible* if any point in M can be reached from any other along an *su-path*, which is a concatenation of finitely many subpaths, each of which lies entirely in a single leaf of $\mathcal{W}^s$ or a single leaf of $\mathcal{W}^u$. Accessibility is a global, topological property of the foliations $\mathcal{W}^u$ and $\mathcal{W}^s$ that is the analogue of transversality of $\mathcal{W}^u$ and $\mathcal{W}^s$ for Anosov diffeomorphisms. In fact, the transversality of these foliations in the Anosov case immediately implies that every Anosov diffeomorphism is accessible. Fundamental Principle #4 suggests that accessibility might be related to ergodicity for conservative systems.

Conservative Partially Hyperbolic Diffeomorphisms Motivated by a breakthrough result with Grayson [27], Pugh and Shub conjectured that accessibility implies ergodicity, for a C^2, partially hyperbolic conservative diffeomorphism [28]. This conjecture has been proved under the hypothesis of center bunching [29], which is a mild spectral condition on the restriction of Df to the center bundle E^{c}. Center bunching is satisfied by most examples of interest, including all partially hyperbolic diffeomorphisms with $\dim(E^c) = 1$. The proof in [29] is a modification of the Hopf Argument using Lebesgue density points and a delicate analysis of the geometric and measure-theoretic properties of the stable and unstable foliations.

In the same article, Pugh and Shub also conjectured that accessibility is a widespread phenomenon, holding for an open and dense set (in the C^{r} topology) of partially hyperbolic diffeomorphisms. This conjecture has been proved completely for $r = 1$ [30], and for all r, with the additional assumption that the central bundle E^{c} is one dimensional [31].

Together, these two conjectures imply the third, central conjecture: in [28]:

Conjecture 1 (Pugh–Shub) *For any $r \geq 2$, the C^r, conservative partially hyperbolic diffeomorphisms contain a C^r open and dense set of ergodic diffeomorphisms.*

The validity of this conjecture in the absence of center bunching is currently an open question.

Dissipative Partially Hyperbolic Diffeomorphisms There has been some progress in constructing SRB-type measures for dissipative partially hyperbolic diffeomorphisms, but the theory is less developed than in the conservative case. Using the same construction as for Anosov diffeomorphisms, one can construct invariant probability measures that are smooth along the $\mathcal{W}^u$ foliation [32]. Such measures are referred to as u-Gibbs measures. Since the stable bundle E^{s} is not transverse to the unstable bundle E^{u}, the Anosov argument cannot be carried through to show that u-Gibbs measures are SRB measures.

Nonetheless, there are conditions that imply that a u-Gibbs measure is an SRB measure: for example,

a u-Gibbs measure is SRB if it is the unique u-Gibbs measure [33], if the bundle $E^s \oplus E^c$ is nonuniformly contracted [34], or if the bundle $E^u \oplus E^c$ is nonuniformly expanded [35]. The proofs of the latter two results use Pesin Theory, which is explained in the next subsection.

SRB measures have also been constructed in systems where E^c is nonuniformly hyperbolic [36], and in (noninvertible) partially hyperbolic covering maps where E^c is 1-dimensional [37]. It is not known whether accessibility plays a role in the existence of SRB measures for dissipative, non-Anosov partially hyperbolic diffeomorphisms.

Nonuniform Hyperbolicity

The concept of Lyapunov exponents gives a natural way to extend the notion of hyperbolicity to systems that behave hyperbolically, but in a nonuniform manner. The fundamental principles described above, suitably modified, apply to these nonuniformly hyperbolic systems and allow for the development of a smooth ergodic theory for these systems. This program was initially proposed and carried out by Yakov Pesin in the 1970s [38] and has come to be known as *Pesin theory*.

Oseledec's Theorem implies that if a smooth map f satisfying the condition $m(D_x f) > 0$ preserves a probability measure ν, then for ν-a.e. $x \in M$ and every nonzero vector $v \in T_x M$, the limit

$$\lambda(x, v) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \log \|D_x f^i(v)\|$$

exists. The number $\lambda(x, v)$ is called the *Lyapunov exponent at x in the direction of v*. For each such x, there are finitely many possible values for the exponent $\lambda(x, v)$, and the function $x \mapsto \lambda(x, \cdot)$ is measurable. See the discussion of Oseledec's Theorem in ► Ergodic Theorems.

Let f be a smooth map. An f-invariant probability measure μ is *hyperbolic* if the Lyapunov exponents of μ-a.e. point are all nonzero. Observe that any invariant measure of a hyperbolic map is a hyperbolic measure.

A conservative diffeomorphism $f\colon M \to M$ is *nonuniformly hyperbolic* if the invariant measure equivalent to volume is hyperbolic. The term "nonuniform" is a bit misleading, as uniformly hyperbolic conservative systems are also nonuniformly hyperbolic. Unlike uniform hyperbolicity, however, nonuniform hyperbolicity allows for the *possibility* of different strengths of hyperbolicity along different orbits.

Nonuniformly hyperbolic diffeomorphisms exist on all manifolds [39,40], and there are C^1-open sets of nonuniformly hyperbolic diffeomorphisms that are not Anosov diffeomorphisms [41]. In general, it is a very difficult problem to establish whether a given map carries a hyperbolic measure that is nonsingular with respect to volume.

Hyperbolic Blocks As mentioned above, the derivative of f along almost every orbit of a nonuniformly hyperbolic system looks like the derivative down the orbit of a uniformly hyperbolic system; the nonuniformity can be detected only by examining a positive measure set of orbits. Recall that Lusin's Theorem in measure theory states that every Borel measurable function is continuous on the complement of an arbitrarily small measure set. A sort of analogue of Lusin's theorem holds for nonuniformly hyperbolic maps: every C^2, nonuniformly hyperbolic diffeomorphism is uniformly hyperbolic on a (noninvariant) compact set whose complement has arbitrarily small measure. The precise formulation of this statement is omitted, but here are some of its salient features.

If μ is a hyperbolic measure for a C^2 diffeomorphism, then attached to μ-a.e. point $x \in M$ are transverse, smooth stable and unstable manifolds for f. The collection of all stable manifolds is called the *stable lamination* for f, and the collection of all unstable manifolds is called the *unstable lamination* for f. The stable lamination is invariant under f, meaning that f sends the stable manifold at x into the stable manifold for $f(x)$. The stable manifold through x is contracted uniformly by all positive iterates of f in a neighborhood of x. Analogous statements hold for the unstable manifold of x, with f replaced by f^{-1}.

The following quantites vary measurably in $x \in M$:

- the (inner) radii of the stable and unstable manifolds through x,
- the angle between stable and unstable manifolds at x, and
- the rates of contraction in these manifolds.

The stable and unstable laminations of a nonuniformly hyperbolic system are absolutely continuous. The precise definition of absolute continuity here is slightly different than in the uniformly and partially hyperbolic setting, but the consequences (AC1) and (AC2) of absolute continuity continue to hold.

Ergodic Properties of Nonuniformly Hyperbolic Diffeomorphisms Since the stable and unstable laminations are absolutely continuous, the Hopf Argument can be applied in this setting to show:

Theorem 4 (Pesin) *Let f be C^2, conservative and nonuniformly hyperbolic. Then there exists a (mod 0) partition $\mathcal{P}$ of*

M into countably many f-invariant sets of positive volume such that the restriction of f to each $P \in \mathcal{P}$ is ergodic.

The proof of this theorem is also exposited in [42]. The countable partition can in examples be countably infinite; nonuniform hyperbolicity alone does not imply ergodicity.

The Dissipative Case As mentioned above, establishing the existence of a nonsingular hyperbolic measure is a difficult problem in general. In systems with some global form of hyperbolicity, such as partial hyperbolicity, it is sometimes possible to "borrow" the expansion from the unstable direction and lend it to the central direction, via a small perturbation. Nonuniformly hyperbolic attractors have been constructed in this way [43]. This method is also behind the construction of a C^1 open set of nonuniformly hyperbolic diffeomorphisms in [41].

For a given system of interest, it is sometimes possible to prove that a given invariant measure is hyperbolic by establishing an *approximate* form of hyperbolicity. The idea, due to Wojtkowski and called the *cone method*, is to isolate a measurable bundle of cones in TM defined over the support of the measure, such that the cone at a point x is mapped by $D_x f$ into the cone at $f(x)$. Intersecting the images of these cones under all iterates of Df, one obtains an invariant subbundle of TM over the support of f that is nonuniformly expanded.

Lai-Sang Young has developed a very general method [44] for proving the existence of SRB measures with strong mixing properties in systems that display "some hyperbolicity". The idea is to isolate a region X in the manifold where the first return map is hyperbolic and distortion estimates hold. If this can be done, then the map carries a mixing, hyperbolic SRB measure. The precise rate of mixing is determined by the properties of the return-time function to X; the longer the return times, the slower the rate of mixing.

More results on the existence of hyperbolic measures are discussed in the next section.

An important subject in smooth ergodic theory is the relationship between entropy, Lyapunov exponents, and dimension of invariant measures of a smooth map. Significant results in this area include the Pesin entropy formula [45], the Ruelle entropy inequality [46], the entropy-exponents-dimension formula of Ledrappier–Young [47,48], and the proof by Barreira-Pesin-Schmeling that hyperbolic measures have a well-defined dimension [49]. ► Hyperbolic Dynamical Systems contains a discussion of these results; see this entry there for further information.

The Presence of Critical Points and Other Singularities

Now for a discussion of the aforementioned technical difficulties that arise in the presence of singularities and critical points for the derivative.

Singularities, that is, points where Df (or even f) fails to be defined, arise naturally in the study of billiards and hard sphere gases. The first subsection discusses some progress made in smooth ergodic theory in the presence of singularities.

Critical points, that is, points where Df fails to be invertible, appear inescapably in the study of noninvertible maps. This type of complication already shows up for noninvertible maps in dimension 1, in the study of unimodal maps of the interval. The second subsection discusses the technique of parameter exclusion, developed by Jakobson, which allows for an ergodic analysis of a parametrized family of maps with criticalities.

The technical advances used to overcome these issues in the interval have turned out to have applications to dissipative, nonhyperbolic, diffeomorphisms in higher dimension, where the derivative is "nearly critical" in places. The last subsection describes extensions of the parameter exclusion technique to these near-critical maps.

Hyperbolic Billiards and Hard Sphere Gases

In the 1870s the physicist Ludwig Boltzmann hypothesized that in a mechanical system with many interacting particles, physical measurements (observables), averaged over time, will converge to their expected value as time approaches infinity. The underlying dynamical system in this statement is a Hamiltonian system with many degrees of freedom, and the "expected value" is with respect to Liouville measure. Loosely phrased in modern terms, Boltzmann's hypothesis states that a generic Hamiltonian system of this form will be ergodic on constant energy submanifolds. Reasoning that the time scales involved in measurement of an observable in such a system are much larger than the rate of evolution of the system, Boltzmann's hypothesis allowed him to assume that physical quantities associated to such a system behave like constants.

In 1963, Sinai revived and formalized this ergodic hypothesis, stating it in a concrete formulation known as the Boltzmann-Sinai Ergodic Hypothesis. In Sinai's formulation, the particles were replaced by N hard, elastic spheres, and to compactify the problem, he situated the spheres on a k-torus, $k = 2, 3$. The Boltzmann-Sinai Ergodic Hypothesis is the conjecture that the induced Hamiltonian system

on the $2kN$-dimensional configuration space is ergodic on constant energy manifolds, for any $N \geq 2$.

Sinai verified this conjecture for $N = 2$ by reducing the problem to a billiard map in the plane. As background for Sinai's result, a brief discussion of planar billiard maps follows.

Let $D \subset \mathbb{R}^s$ be a connected region whose boundary ∂D is a collection of closed, piecewise smooth simple curves the plane. The *billiard map* is a map defined (almost everywhere) on $\partial D \times [-\pi, \pi]$. To define this map, one identifies each point $(x, \theta) \in \partial D \times [-\pi, \pi]$ with an inward-pointing tangent vector at x in the plane, so that the normal vector to ∂D at x corresponds to the pair $(x, \pi/2)$. This can be done in a unique way on the smooth components of ∂D. Then $f(x, \theta)$ is obtained by following the ray originating at (x, θ) until it strikes the boundary ∂D for the first time at (x', θ'). Reflecting this vector about the normal at x', define $f(x, \theta) = (x', \pi - \theta')$.

It is not hard to see that the billiard map is conservative. The billiard map is piecewise smooth, but not in general smooth: the degree of smoothness of f is one less than the degree of smoothness of ∂D. In addition to singularities arising from the corners of the table, there are singularities arising in the *second* derivative of f at the tangent vectors to the boundary.

In the billiards studied studied by Sinai, the boundary ∂D consists of a union of concave circular arcs and straight line segments. Similar billiards, but with convex circular arcs, were first studied by Bunimovich [51]. Sinai and Bunimovich proved that these billiards are ergodic and nonuniformly hyperbolic. For the Boltzmann-Sinai problem with $N \geq 3$, the relevant associated dynamical system is a higher dimensional billiard table in euclidean space, with circular arcs replaced by cylindrical boundary components.

In a planar billiard table with circular/flat boundary, the behavior of vectors encountering a flat segment of boundary is easily understood, as is the behavior of vectors meeting a circular segment in a neighborhood of the normal vector. If the billiard map is ergodic, however, every open set of vectors will meet the singularities in the table infinitely many times. To establish the nonuniform hyperbolicity of such billiard tables via conefieds, it is therefore necessary to understand precisely the fraction of time orbits spend near these singularities. Furthermore, to use the Hopf argument to establish ergodicity, one must avoid the singularities in the second derivative, where distortion estimates break down. The techniques for overcoming these obstacles involve imposing restrictions on the geometry of the table (even more so for higher dimensional tables), and are well beyond the scope of this paper.

The study of hyperbolic billiards and hard sphere gases has a long and involved history. See the articles [50] and [52] for a survey of some of the results and techniques in the area. A discussion of methods in singular smooth ergodic theory, with particular applications to the Lorentz attractor, can be found in [53]. Another, more classical, reference is [54], which contains a formulation of properties on a critical set, due to Katok–Strelcyn, that are useful in establishing ergodicity of systems with singularities.

Interval Maps and Parameter Exclusion

The logistic family of maps $f_t\colon x \mapsto tx(1-x)$ defined on the interval $[0, 1]$ is very simple to define but exhibits an astonishing variety of dynamical features as the parameter t varies. For small positive values of t, almost every point in I is attracted under the map f_t to the sink at -1. For values of $t > 4$, the map has a repelling hyperbolic Cantor set. As the value of t increases between 0 and 4, the map f_t undergoes a cascade of period-doubling bifurcations, in which a periodic sink of period 2^n becomes repelling and a new sink of period 2^{n+1} is born. At the accumulation of period doubling at $t \approx 3.57$, a periodic point of period 3 appears, forcing the existence of periodic points of all periods. The dynamics of f_t for t close to 4 has been the subject of intense inquiry in the last 20 years.

The map f_t, for t close to 4, shares some of the features of the doubling map T_2; it is 2-to-1, except at the critical point $\frac{1}{2}$, and it is uniformly expanding in the complement of a neighborhood of this critical point. Because this neighborhood of the critical point is not invariant, however, the only invariant sets on which f_t is uniformly hyperbolic have measure zero. Furthermore, the second derivative of f_t vanishes at the critical point, making it impossible to control distortion for orbits that spend too much time near the critical point.

Despite these serious obstacles, Michael Jakobson [55] found a method for constructing absolutely continuous invariant measures for maps in the logistic family. The method has come to be known as *parameter exclusion* and has seen application far beyond the logistic family. As with billiards, it is possible to formulate geometric conditions on the map f_t that control both expansion (hyperbolicity) and distortion on a positive measure set. As these conditions involve understanding infinitely many iterates of f_t, they are impossible to verify for a given parameter value t.

Using an inductive formulation of this condition, Jakobson showed that the set of parameters t near 4 that fail to satisfy the condition at iterate n have exponentially small measure (in n). He thereby showed that for a posi-

tive Lebesgue measure set of parameter values t, the map f_t has an absolutely continuous invariant measure [55]. This measure is ergodic (mixing) and has a positive Lyapunov exponent. The delicacy of Jakobson's approach is confirmed by the fact that for an open and dense set of parameter values, almost every orbit is attracted to a periodic sink, and so f_t has no absolutely continuous invariant measure [56,57]. Jakobson's method applies not only to the logistic family but to a very general class of C^3 one-parameter families of maps on the interval.

Near-Critical Diffeomorphisms

Jakobson's method in one dimension proved to extend to certain highly dissipative diffeomorphisms. The seminal paper in this extension is due to Benedicks and Carleson; the method has since been extended in a series of papers [59,60,61] and has been formulated in an abstract setting [62].

This extension turns out to be highly nontrivial, but it is possible to describe informally the similarities between the logistic family and higher-dimensional "near critical" diffeomorphisms. The diffeomorphisms to which this method applies are crudely hyperbolic with a one dimensional unstable direction. Roughly this means that in some invariant region of the manifold, the image of a small ball under f will be stretched significantly in one direction and shrunk in all other directions. The directions of stretching and contraction are transverse in a large proportion of the invariant region, but there are isolated "near critical" subregions where expanding and contracting directions are nearly tangent.

The dynamics of such a diffeomorphism are very close to 1-dimensional if the contraction is strong enough, and the diffeomorphism resembles an interval map with isolated critical points, the critical points corresponding to the critical regions where stable and unstable directions are tangent.

An illustration of this type of dynamics is the Hénon family of maps $f_{a,b}\colon (x, y) \mapsto (1 - ax^2 + by, x)$, the original object of study in Benedicks-Carlesson's work. When the parameter b is set to 0, the map $f_{a,b}$ is no longer a diffeomorphism, and indeed is precisely a projection composed with the logistic map. For small values of b and appropriate values of a, the Hénon map is strongly dissipative and displays the near critical behavior described in the previous paragraph. In analogy to Jakobson's result, there is a positive measure set of parameters near $b = 0$ where $f_{a,b}$ has a mixing, hyperbolic SRB measure.

See [63] for a detailed exposition of the parameter exclusion method for Hénon-like maps.

Future Directions

In addition to the open problems discussed in the previous sections, there are several general questions and problems worth mentioning:

- What can be said about systems with everywhere vanishing Lyapunov exponents? Open sets of such systems exist in arbitrary dimension. Pesin theory carries into the nonuniformly hyperbolic setting the basic principles from uniformly hyperbolic theory (in particular, Fundamental Principles #3 and 4 above). To what extent do properties of isometric and unipotent systems (for example, Fundamental Principle #2) extend to conservative systems all of whose Lyaponov exponents vanish?
- Can one establish the existence of and analyze in general the conservative systems on surfaces that have two positive measure regimes: one where Lyapunov exponents vanish, and the other where they are nonzero? Such systems are conjectured exist in the presence of KAM phenomena surrounding elliptic periodic points.
- On a related note, how common are conservative systems whose Lyapunov exponents are nonvanishing on a positive measure set? See [64] for a discussion.
- Find a broad description of those dissipative systems that admit finitely (or countably) many physical measures. Are such systems dense among all dissipative systems, or possibly generic among a restricted class of systems? See [41,65] for several questions and conjectures related to this problem.
- Extend the methods in the study of systems with singularities to other specific systems of interest, including the infinite dimensional systems that arise in the study of partial differential equations.
- Carry the methods of smooth ergodic theory further into the study of smooth actions of discrete groups (other than the integers) on manifolds. When do such actions admit (possibly non-invariant) "physical" measures?

There are other interesting open areas of future inquiry, but this gives a good sample of the range of possibilities.

Bibliography

1. Arnol'd VI, Avez A (1986) Ergodic problems of classical mechanics. Translated from the French by Avez, A. Benjamin WA, Inc., New York Amsterdam
2. Mañé R (1987) Ergodic theory and differentiable dynamics. Translated from the Portuguese by Silvio L. Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)], vol 8. Springer, Berlin

3. Baladi V (2000) Positive transfer operators and decay of correlations. Advanced Series in Nonlinear Dynamics, 16. World Scientific Publishing Co., Inc., River Edge
4. Kifer Y (1986) Ergodic theory of random transformations. Progress in Probability and Statistics, vol 10. Birkhäuser, Boston
5. Kifer Y (1988) Random perturbations of dynamical systems. Progress in Probability and Statistics, 16. Birkhäuser, Boston
6. Liu PD, Qian M (1995) Smooth ergodic theory of random dynamical systems. Lecture Notes in Mathematics, vol 1606. Springer, Berlin
7. Bonatti C, Dìaz LJ, Viana M (2005) Dynamics beyond uniform hyperbolicity. A global geometric and probabilistic perspective. Encyclopaedia of Mathematical Sciences, 102. Mathematical Physics, III. Springer, Berlin
8. Cornfeld IP, Fomin SV, Sinaĭ YG (1982) Ergodic theory. Translated from the Russian by A. B. Sosinskiĭ. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol 245. Springer, New York
9. Hirsch MW (1979) Differential topology. Graduate Texts in Mathematics, No. 33. Springer, New York Heidelberg
10. Robinson C (1995) Dynamical systems. Stability, symbolic dynamics, and chaos. Studies in Advanced Mathematics. CRC Press, Boca Raton
11. Katok A, Hasselblatt B (1995) Introduction to the modern theory of dynamical systems. With a supplementary chapter by Katok and Leonardo Mendoza. Encyclopedia of Mathematics and its Applications, vol 54. Cambridge University Press, Cambridge
12. Hirsch MW, Pugh CC, Shub M (1977) Invariant manifolds. Lecture Notes in Mathematics, vol 583. Springer, Berlin New York
13. Brin M, Stuck G (2002) Introduction to dynamical systems. Cambridge University Press, Cambridge
14. Hopf E (1939) Statistik der geodätischen Linien in Mannigfaltigkeiten negativer Krümmung. Ber Verh Sachs Akad Wiss Leipzig 91:261–304 (German)
15. Anosov DV (1967) Geodesic flows on closed Riemann manifolds with negative curvature. Proceedings of the Steklov Institute of Mathematics, No. 90 1967. Translated from the Russian by S. Feder. American Mathematical Society, Providence
16. Sinaĭ JG (1961) Geodesic flows on compact surfaces of negative curvature. Dokl Akad Nauk SSSR 136:549–552 (Russian); translated as Soviet Math Dokl 2:106–109
17. Anosov DV, Sinaĭ JG (1967) Certain smooth ergodic systems. (Russian) Uspehi Mat Nauk 22 no. 5(137):107–172
18. Sinaĭ JG (1972) Gibbs measures in ergodic theory. (Russian) Uspehi Mat Nauk 27 no. 4(166):21–64
19. Ruelle D (1976) A measure associated with axiom-A attractors. Amer J Math 98(3):619–654
20. Bowen R (1975) Equilibrium states and the ergodic theory of Anosov diffeomorphisms. Lecture Notes in Mathematics, vol 470. Springer, Berlin New York
21. Young LS (2002) What are SRB measures, and which dynamical systems have them? Dedicated to David Ruelle and Yasha Sinai on the occasion of their 65th birthdays. J Statist Phys 108(5–6):733–754
22. Bowen R (1970) Markov partitions for Axiom A diffeomorphisms. Amer J Math 92:725–747
23. Brin MI, Pesin JB (1974) Partially hyperbolic dynamical systems. Izv Akad Nauk SSSR Ser Mat 38:170–212 (Russian)
24. Pugh C, Shub M (1972) Ergodicity of Anosov actions. Invent Math 15:1–23
25. Burns K, Pugh C, Shub M, Wilkinson A (2001) Recent results about stable ergodicity. Smooth ergodic theory and its applications. (Seattle, 1999), 327–366. Proc Sympos Pure Math, 69, Amer Math Soc, Providence
26. Pesin YB (2004) Lectures on partial hyperbolicity and stable ergodicity. Zurich Lectures in Advanced Mathematics. European Mathematical Society (EMS), Zürich
27. Grayson M, Pugh C, Shub M (1994) Stably ergodic diffeomorphisms. Ann of Math 140(2):295–329
28. Pugh C, Shub M (1996) Stable ergodicity and partial hyperbolicity. International Conference on Dynamical Systems (Montevideo, 1995), 182–187, Pitman Res Notes Math Ser, 362, Longman, Harlow
29. Burns K, Wilkinson A, On the ergodicity of partially hyperbolic systems. Ann of Math. To appear
30. Dolgopyat D, Wilkinson A (2003) Stable accessibility is C^1 dense. Geometric methods in dynamics. II. Astérisque No. 287
31. Rodríguez HA, Rodríguez HF, Ures R (2008) Partially hyperbolic systems with 1D-center bundle. Invent Math 172(2)
32. Pesin YB, Sinaĭ YG (1983) Gibbs measures for partially hyperbolic attractors. Ergod Theor Dynam Syst 2(3–4):417–438
33. Dolgopyat D (2004) On differentiability of SRB states for partially hyperbolic systems. Invent Math 155(2):389–449
34. Bonatti C, Viana M (2000) SRB measures for partially hyperbolic systems whose central direction is mostly contracting. Israel J Math 115:157–193
35. Alves JF, Bonatti C, Viana M (2000) SRB measures for partially hyperbolic systems whose central direction is mostly expanding. Invent Math 140(2):351–398
36. Burns K, Dolgopyat D, Pesin Y, Pollicott Mark Stable ergodicity for partially hyperbolic attractors with negative central exponents. Preprint
37. Tsujii M (2005) Physical measures for partially hyperbolic surface endomorphisms. Acta Math 194(1):37–132
38. Pesin JB (1977) Characteristic Ljapunov exponents, and smooth ergodic theory. Uspehi Mat Nauk 32(196):55–112, 287 (Russian)
39. Katok A (1979) Bernoulli diffeomorphisms on surfaces. Ann Math 110(2):529–547
40. Dolgopyat D, Pesin Y (2002) Every compact manifold carries a completely hyperbolic diffeomorphism. Ergod Theor Dynam Syst 22(2):409–435
41. Shub M, Wilkinson A (2000) Pathological foliations and removable zero exponents. Invent Math 139(3):495–508
42. Pugh C, Shub M (1989) Ergodic attractors. Trans Amer Math Soc 312(1):1–54
43. Viana M (1997) Multidimensional nonhyperbolic attractors. Inst Hautes Études Sci Publ Math No 85:63–96
44. Young LS (1998) Statistical properties of dynamical systems with some hyperbolicity. Ann Math 147(2):585–650
45. Pesin JB (1976) Characteristic Ljapunov exponents, and ergodic properties of smooth dynamical systems with invariant measure. Dokl Akad Nauk SSSR 226 (Russian)
46. Ruelle D (1978) An inequality for the entropy of differentiable maps. Bol Soc Brasil Mat 9(1):83–87
47. Ledrappier F, Young LS (1985) The metric entropy of diffeomorphisms. I. Characterization of measures satisfying Pesin's entropy formula. Ann Math 122(2):509–539
48. Ledrappier F, Young LS (1985) The metric entropy of diffeomorphisms. II. Relations between entropy, exponents and dimension. Ann Math 122(2):540–574

49. Barreira L, Pesin Y, Schmeling J (1999) Dimension and product structure of hyperbolic measures. Ann Math 149(2):755–783
50. Szász D (2000) Boltzmann's ergodic hypothesis, a conjecture for centuries? Hard ball systems and the Lorentz gas, Encycl Math Sci, vol 101. Springer, Berlin, pp 421–448
51. Bunimovič LA (1974) The ergodic properties of certain billiards. (Russian) Funkcional. Anal i Priložen 8(3):73–74
52. Chernov N, Markarian R (2003) Introduction to the ergodic theory of chaotic billiards. Second edition. Publicações Matemáticas do IMPA. IMPA Mathematical Publications 24º Colóquio Brasileiro de Matemática. 26th Brazilian Mathematics Colloquium Instituto de Matemática Pura e Aplicada (IMPA), Rio de Janeiro
53. Araújo V, Pacifico MJ (2007) Three Dimensional Flows. Publicações Matemáticas do IMPA. [IMPA Mathematical Publications] 26º Colóquioquio Brasileiro de Matemática. [26th Brazilian Mathematics Colloquium] Instituto de Matemática Pura e Aplicada (IMPA), Rio de Janeiro
54. Katok A, Strelcyn JM, Ledrappier F, Przytycki F (1986) Invariant manifolds, entropy and billiards; smooth maps with singularities. Lecture Notes in Mathematics, 1222. Springer, Berlin
55. Jakobson MV (1981) Absolutely continuous invariant measures for one-parameter families of one-dimensional maps. Comm Math Phys 81(1):39–88
56. Graczyk J, Światek G (1997) Generic hyperbolicity in the logistic family. Ann Math 146(2):1–52
57. Lyubich M (1999) Feigenbaum–Coullet–Tresser universality and Milnor's hairiness conjecture. Ann Math 149(2):319–420
58. Benedicks M, Carleson L (1991) The dynamics of the Hénon map. Ann Math 133(2):73–169
59. Mora L. Viana M (1993) Abundance of strange attractors. Acta Math 171(1):1–71
60. Benedicks M, Young LS (1993) Sinaĭ–Bowen–Ruelle measures for certain Hénon maps. Invent Math 112(3):541–576
61. Benedicks M, Viana M (2001) Solution of the basin problem for Hénon-like attractors. Invent Math 143(2):375–434
62. Wang QD, Young LS (2008) Toward a theory of rank one attractors. Ann Math 167(2)
63. Viana M, Lutstsatto S (2003) Exclusions of parameter values in Hénon-type systems. (Russian) Uspekhi Mat Nauk 58 (2003), no. 6(354), 3–44; translation in Russian Math. Surveys 58(6):1053–1092
64. Bochi J, Viana M (2004) Lyapunov exponents: how frequently are dynamical systems hyperbolic? Modern dynamical systems and applications. Cambridge Univ Press, Cambridge, pp 271–297
65. Palis J (2005) A global perspective for non-conservative dynamics. Ann Inst H Poincaré Anal Non Linéaire 22(4):485–507

Social Cognitive Complexity

Jürgen Klüver
Duisburg-Essen University, Essen, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Social-Cognitive Complexity, Algorithms, and Bottom Up Models
Theoretical Foundations
A Universal Modeling Schema
Examples
Conclusions and Future Directions
Bibliography

Glossary

Attractor A sub space of the state space of a certain system that the system does not leave anymore. If the sub space consists of only one point it is a point attractor.

Dialectical relation A relation between two persons or generally two sides representing a tension and a mutual dependency. In more contemporary terms a dialectical relation is a certain form of feed back relation.

Homo oeconomicus An actor who acts according to the assumptions of rational choice.

Homo sociologicus The assumption that man is basically a being following social rules in contrast to *homo oeconomicus*.

Meta rules Rules that operate on rules of interaction and change these rules.

Rational choice The theoretical assumption that man is a rational and egoistical actor, who tries to maximize his profit and selects in a certain situation the best strategies.

Socialization The biographical process by which an individual learns and internalizes the cultural norms and social rules of his specific society. The result is a social actor.

Social-cognitive systems Complex dynamical systems that consist of social actors and contain at least two different levels, namely a social level and a cognitive one.

Socio-cultural evolution The development of societies in the mutually interdependent dimensions of social structure and culture.

Universal modeling schema A schema that consists of different interdependent levels and each level is modeled as a complex dynamical system.

Definition of the Subject

Social-cognitive systems certainly belong to the most complex systems that we know. The reason for this complexity is due to the fact that social-cognitive systems at least consist of two different levels, namely a social and a cognitive one that permanently interact and mutually influence each other. In terms of the social and cognitive sciences, social-

cognitive systems consist of social actors who interact and whose actions are determined by certain social rules. Yet these actors are also complex dynamical systems, i. e., cognitive ones. The cognitive processes of the individual actors also influence the specific actions. The social rules determine in some part the cognitive processes that in turn generate the according actions. But also in turn the cognitive processes may determine the actions in such a way that the social rules are influenced and even changed. That is for example the case in times of revolutions and/or political reforms. In this sense social-cognitive systems are determined by a permanent interdependent dynamics between these levels and that is why a modeling and precise analysis of these systems is rather difficult. In addition, social-cognitive systems are able to adapt to changing environmental conditions. This means that these systems are able to change their rules and structure.

Introduction

Social-cognitive complexity, although of course not in this term, has been the subject of social thinking since the beginning of reflections on society and social actions. Already one of the earliest works on society in human history, namely Plato's *Politeia* (The Republic), emphasizes the importance of the interdependency of mind and social structure: A certain desired social structure, i. e. the differentiation of society into different social classes, can be generated and reproduced only by influencing the minds of the social actors via education. A suited education cares for the adequate thinking of the individuals, which in turn guarantees the production and reproduction of the class structure by the social actions of the educated actors. The class structure finally determines the minds of the individuals and produces the wished forms of socially determined thinking. Plato's ideal society is certainly not a paradigm for democratic societies but the logic of Plato's arguments already demonstrates the necessity to take account of both social and cognitive levels if one wants to understand society and the socially determined actions of human beings. It is not by chance that all authors who thought about utopian societies also stressed the importance of education, i. e. the influencing of minds to produce and reproduce social structures, for example Huxley and Orwell.

More than two millennia later, Karl Marx in his theory of Historical Materialism [27] dealt in a similar way with the subject of social-cognitive complexity. He postulated a "dialectical" relation between consciousness or mind respectively and social structure – the social conditions (*soziale Verhältnisse*) in his terms. The social conditions determine the minds of social actors and produce that way an ideological consciousness; accordingly the social actors reproduce the social conditions by their ideologically determined actions. That is the essence of Marx's famous base-superstructure theorem: the social conditions determine the consciousness, which in turn produces the ideological parts of society (In the German original: *Das gesellschaftliche oder materielle Sein bestimmt das Bewusstsein*). Yet in times of societal crisis the consciousness of certain social actors, namely the members of revolutionary classes, the consciousness is able to overcome the ideological constraints and produces by the social actions of revolution a new social structure. In such times historical progress is generated by the determination of social structures by the social minds of the revolutionary actors. That is why Marx called the relationship between mind and social structure a dialectical one: both sides of the relationship form a dynamical tension, influence the other and are influenced in turn by the other. If one substitutes the philosophical term of dialectic by the more precise contemporary term of feed back, one may say that Marx postulated social-cognitive systems as systems whose dynamics is characterized by permanent feed back loops between a social and a cognitive level. To be sure, although Marx frequently stressed the importance of mathematically formulated theories he never could think of transforming the venerable concept of dialectic into a mathematical framework.

Marx had in principle captured the whole problem of social-cognitive complexity although only in an informal manner and formulated in classical philosophical terms. Yet his monumental approach was certainly not suited as a common foundation for the development of the social and cognitive sciences. Nearly at the same time the French sociologist Emile Durkheim, one of the founding fathers of modern sociology, postulated that social phenomena must only be explained by other *social* phenomena, which is one of the most important theoretical premises of the contemporary social sciences. As a consequence social scientists only deal with *social* complexity and consciously often neglect the influence of cognitive factors on social processes. Accordingly cognitive scientists only deal with *cognitive* complexity and neglect the impact of social factors on cognitive processes. Paradigmatic in this sense is the great work of Jean Piaget [34,35] who defined cognitive development as a principally autonomous process that can only be slowed down or accelerated by its social environment. With a famous expression of the German social theorist Niklas Luhmann one can say that the different disciplines of social-cognitive complexity reduced the complexity of their subject by concentrating on only one level of the two-level problem [26].

Matters became even more complicated when social theorists discovered that society is not a homogenous system but must itself be differentiated into different levels. The most basic differentiation is of course that of a micro-level and a macro-level [1,22]. In addition it is possible and sometimes necessary to introduce a meso-level, that is a middle range level of social reality. For example, if the micro-level is constituted by individual social actors or their actions respectively, if the macro-level is defined as a whole society or even an aggregation of societies like, e. g. the European Union, then a meso-level may consist of certain institutions, political parties and the like.

An additional complication and hence another increase of complexity must be accepted by taking account of the fundamental difference between social structure and culture. It is not possible for this article to mention all the numerous definitions of culture social theorists have developed in the last century. Following Habermas [15] and Giddens [12] it is sufficient to define the social structure of a society as the set of all social rules that are valid in a certain society and culture as the sum of all accepted knowledge, beliefs, world views, etc. in that society. By considering that fundamental distinction and by taking into account the different levels of society we apparently obtain a multi-dimension system, namely a system consisting of different levels that contain two dimensions – the socio-structural one and the cultural one. It is no wonder that social theorists concentrated on the social level of social-cognitive complexity, because this level is complex enough.

The main problem with such multi-level approaches is of course that the relations between the different levels and hence their interactions could usually be described only in an informal and frequently metaphorical manner. When a social theorist draws a graphical schema with different levels and additionally two dimensions the schema itself explains nothing but just states a problem. Somehow the levels interact and somehow individual actions are determined by the levels above, but in general this *somehow* cannot be precisely explained. To be sure, there are a lot of empirical studies that try, e. g., to analyze the impact of certain social structure on specific actions. But there is no general precise theory of links between the different levels of society [1].

Yet despite these difficulties in several fields of the social and cognitive sciences not only the different levels of society but also the cognitive level of individual actors is taken into account. That is especially the case in socio-psychology and in particular in theories of socialization. The analysis of processes of socialization indeed demands that cognitive levels and social levels alike must be considered, if one understands socialization as the process by which the individual mind is formed according to the social structure and the cultural norms and values of a particular society. Although in the general models of processes of socialization the relation between the social levels and the cognitive one is expressed only one-sided, i. e., only the determination of the personality by social structure is taken into account in contrast to Marx, at least it is suited to speak of attempts to capture the whole social-cognitive complexity of social reality. Such a model of socialization is shown in Fig. 1; it is a shortened version of a more complex model of Geulen and Hurrelmann [11].

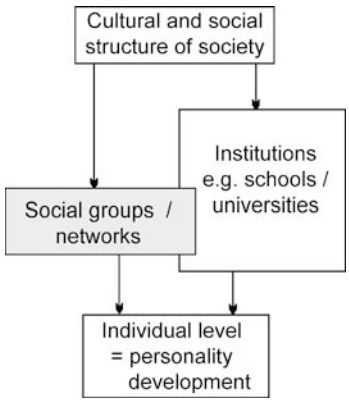

Social Cognitive Complexity, Figure 1
A multi-level model of socialization

The well-known classical study of Berger and Luckmann "The Social Construction of Reality" is a paradigmatic example of such attempts [6] that in contrast to most one-sided models assumes a mutual interdependency of social and cognitive levels: According to them there is, in the tradition of Marx, a dialectical relation between the subjective social reality and the objective one. The subjective social reality consists of the internalized social rules and cultural values that constitute the personal identity, i. e. the individual personality. The objective social reality is generated by the according social actions and interactions of the socialized individual actors. Because the structure of this reality mirrors that of the individual personality the individual actions reproduce the objective reality, which was the determining factor of the constitution of the subjective reality. As in the case of Marx we have again a permanent feedback between the social and the cognitive level. Younger individuals participate in this process by taking over the structure of objective reality from their elders, by reproducing this structure in their own actions, and so forth. Although the theory of Berger and Luckmann does not deal with social progress, i. e., changes of social structure and culture, their model demonstrates how in principle it is possible to describe the interdependency

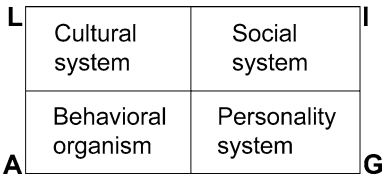

Social Cognitive Complexity, Figure 2
The general action system

of social and cognitive levels, even if only in an informal manner.

The most famous attempt to capture social-cognitive complexity in a whole model is certainly the well-known model of Parsons [8,29,32] of the differentiation of the action system. A simplified version is shown in Fig. 2.

The general assumption of Parsons is that each system is differentiated into four subsystems according to the famous AGIL-schema; the same logic of differentiation is valid for the subsystems and so forth. A means Adaptation, G stands for Goal Orientation, I is Integration, and L means Latent Pattern Maintenance. (The explanation of the AGIL-schema cannot be done in this article [29,33]). The picture shows that the social subsystems (social and cultural systems) constitute together with the cognitive subsystems (personality and behavioral system) the whole action system. The specific form of interaction between the four subsystems is called interpenetration, which roughly means that the four systems influence each other. For example, personalities are formed via socialization by the social and cultural subsystems, social and cultural systems are reproduced by personalities and according behavior and so on. Yet again we only have an informal theory, although the general logic of the AGIL-differentiation is certainly a great achievement.

To put it into a nutshell, the problem of dealing with social-cognitive complexity and not reducing the problem to one level is seen by many theorists. Yet the theoretical models are always just informal descriptions of the problem and programmatic foundations, if one looks at them from a precise, i. e. mathematical point of view.

Social-Cognitive Complexity, Algorithms, and Bottom Up Models

Despite many attempts to formulate the complexity of social-cognitive systems in a mathematical way for a long time it was not possible to transfer the mathematical tools of, e. g., mathematical physics to the problems of social reality. The use of statistical methods in empirical social research and experimental psychology is of course a well tested method since many decades. The problem was and frequently still is the mathematical formulation of *theoretical* models. Apparently the use of in particular differential equations is not suited for the problems one has to deal with in the social and cognitive realms. To be sure, simple cases can be analyzed in a mathematical way e .g. [5], but the complexity of social-cognitive systems in the sense of the great theorists mentioned in the preceding section could not be captured by models founded on the classical methods of the calculus.

The reasons for this are the peculiarities of social-cognitive systems. They are constituted, as was mentioned in the definition, by several levels that permanently interact and thus must be characterized by a multi-dimensional dynamics. In addition, social-cognitive systems are able to change their specific rules and structure by adapting to varying environmental conditions. It is of course possible to undertake a time series analysis in order to find some regularity in the historical development of such systems, but a time series analysis just describes a system's behavior in a phenomenological manner and gives no explanation. To make matters worse, even if one has found some regularity by a careful time series analysis the system may change its behavior via the varying of its rules; accordingly the time series will change and the analysis becomes worthless for the next time span. It is no wonder that more than once social and cognitive scientists alike believed that it would never be possible to develop a mathematical science of social and cognitive systems.

Fortunately things have changed by the development of new mathematical tools for the analysis of complex systems and in particular by the emergence of complex computer programs. These new mathematical modeling techniques, as cellular automata, Boolean networks, evolutionary algorithms, and artificial neural nets are on a first sight rather different methods but on a second sight they have much in common:

a) All modeling techniques are heuristically oriented to some natural processes as the reproductive capacity of living systems in the case of cellular automata and Boolean nets, the logic of biological evolution in the case of evolutionary algorithms, and the cognitive processes of the brain in the case of artificial neural nets. Sometimes these new mathematical tools are put together under the a bit misleading name of Soft Computing [48] but it would be more appropriate to call them nature oriented algorithms.
b) All these algorithms are very well suited for the construction of so-called bottom up models [24]. This term is used in contrast with top down models and means that the model construction is done by starting at the level of the elements of the system that shall be mod-

eled. Accordingly the units of the models represent the elements of the respective systems. The dynamics of the empirical system is represented by certain rules of interaction and the thus caused behavior of the whole model. The model's behavior is generated as an emergent consequence of the local interactions of the units of the model. This procedure allows constructing the respective model by *immediately* transferring the empirical observations of the system's elements and their rules of local interaction into the model. It is not necessary to make an abstract representation of the whole system by an aggregation of the elements.

For example, the modeling of a social group via a cellular automaton [21] can be done the following quite natural way: The cells of the cellular automaton represent the different individual members of the group and their specific rules of interaction determined by the group's social hierarchy are represented by the rules of transitions of the cellular automaton. It is then possible, at least in some cases, to predict the group's behavior by according simulation runs of the cellular automaton, as we did in several successful cases. The work of many researchers demonstrated that in particular cellular automata are very well suited for the analysis of social groups [13,16,25,30,39,41,43].

c) The capability of social-cognitive systems to adapt, i. e. to change their specific rules of interaction in order to fulfill certain environmental conditions can be modeled by the hybridization of the respective models, i. e., by the coupling of different modeling tools and such the constructing of a hybrid model [14]. In order to manage a successful adaptation adaptive systems must have not only rules of interaction but also meta rules, namely additional rules that operate on (*meta*) the rules of local interaction and change them according to the environmental demands. The best-known example for such adaptive capabilities is of course the changing of genomes by the genetic operators of mutation and recombination and by the force of natural selection in biological evolution. In this example the rules are those of individual epigenesis and ontogenesis; the meta rules are the genetic operators, steered by selection, that change the genome by variation.

   A mathematical model of these processes of adaptation can be quite easily done by the construction of a hybrid system, consisting of several Boolean or logical nets respectively, representing the individual genotypes [18], and a genetic algorithm that operates on the rules of the Boolean nets and their specific topology.

   Another example for adaptation is the case of political reforms. In nearly all advanced societies the *official* social rules of interaction are defined by laws and bureaucratic prescriptions. In parliamentary societies the meta rules are defined by the competence of the parliaments to change the laws and prescriptions by discussions and votes of majority. The meta rules then are given by the rules of procedure by which the parliaments have to operate. A variation of the rules of interaction by parliamentary decisions is done when the internal or external conditions of the society demand a changing of social structure. It can be demonstrated that democratic societies are more adaptive than dictatorships because dictatorial societies are both unwilling and unable to change their structure because of the resulting unrest and the danger for the ruling elite to lose its power [19]. An according model for such phases of political changing could, e. g. be constructed by the coupling of a cellular automaton (the model for the population) and an evolutionary strategy (the meta rules).

d) In contrast with a pure time series analysis bottom up models are able to explain, i. e. they can reduce observed phenomena to causes, namely to previous events that are causally linked to the observed event. In cases of complex systems such previous events can be either a previous state of the system. In that case the explanation consists of naming this particular previous state and the rules that generated the present state from the previous one. Or the previous event is an external disturbance of the system. Then the explanation consists of naming the disturbance and the reaction of the system's rules to the disturbance, i. e. to the externally changed state of the system. The third and most difficult case is that of an operation of the system's meta rules, their variation of the rules of interaction, and the generation of the present state via the changed rules of interaction. To be sure, the analysis and explanation of all different cases by the construction of an according valid model needs a lot of information about the empirical system. Yet a thorough explanation of such complex systems as socio-cognitive ones can only be done by the construction of models and simulation runs that use these new mathematical tools. To be sure, explanations of this kind can principally also be done by top down models and have been successfully done in physics and chemistry. Only the third case needs the distinction between rules and meta rules and hence demands models constructed the way sketched above.

Theoretical Foundations

The social sciences are, unfortunately, characterized by a multitude of different theoretical approaches that can-

not be enumerated in this article. Yet for attempts to construct mathematical models of social or even social-cognitive processes there are two main approaches, namely the Rational Choice approach (abbreviated RC) e.g. [17] and a social systems theory approach, founded on social rules and structure. Both are based on certain anthropological foundations about the nature of man or social actors respectively.

The RC-approach is based on the assumption that man is basically an egoistical being who always tries to maximize his own profit without caring for others. This assumption has a venerable tradition and it goes at least back to the Leviathan of Hobbes, namely the question how social order is possible in a world of egoists. The term "rational" means that a social actor tries to find the optimal strategies in each action situation, i. e., he analyzes, which action as means could serve him best for his maximum welfare as goal. Hence each social action situation is considered by an actor as a situation in a strategic game like chess or poker: he rationally analyzes the different possible strategies and chooses the best – therefore rational choice.

The RC-approach has a great advantage for the purpose of mathematical model construction in the social sciences, namely the famous Theory of Games by von Neumann and Morgenstern [45]. Game theory was originally developed for the mathematical analysis of economical decision problems and already Max Weber stated that economical actions are a paradigm of rational actions, oriented to the most favorable proportion between different means and the goal of profit maximizing e.g. [46] (Oscar Morgenstern was an economist). But soon the partisans of RC-approaches generalized this idea and tried to capture the logic of social action in general by game theoretical models.

Usually game theoretical models are based on a so-called pay-off matrix. Because most RC-theorists distinguish only between two forms of behavior, namely cooperation and defection, a pay-off matrix is a two-dimensional one. A characteristic a pay-off matrix that is frequently used for the analysis of social dilemmas, in particular the famous Prisoner's Dilemma (PD), is the following one:

	C	D
C	3	0
D	5	1

C of course stands for cooperation and D for defection. As usual the matrix-values show the result of the combination of the two different action strategies. For example, the combination DC = 5 means that a defective player D gets five points while the cooperative player gets zero points (CD = 0). Accordingly a cooperative player gets three points if the other is also cooperative; the other gets three points too (CC = 3). The absolute values of course are arbitrary; the only condition for strategic games of the PD-type is that DC > CC > DD > CD.

In a famous study about *The Evolution of Cooperation* based on a tournament of different computer programs, i. e. programs that simulated winning strategies, Axelrod [2] showed that the best strategy was *Tit for Tat* (TFT). TFT can roughly be translated as "I do the same to you as you to me". TFT is a very simple strategy because the first player *A* always starts with a cooperative action and waits for the response of the other player *B*. If *B* is also cooperative then *A* keeps his cooperative actions. If *B* is defective then *A* switches his strategy and becomes defective too as long as *B* acts in a defective manner. The first cooperative action of *B* then is answered by *A* with an also cooperative action and so forth. Axelrod used for his own analysis a cellular automaton and confirmed his results later by using a genetic algorithm [3].

TFT is basically a cooperative strategy because each player who acts according to TFT always takes account of the strategies of his opponent and becomes defective only if his opponent does the same. Hence Axelrod concluded that the emergence of cooperation in a world of rational egoists is not only possible but also even probable because it can be mathematically shown that a cooperative strategy like TFT is the most favorable one for an egoistic rational player. In this sense the old question of Hobbes could finally be answered by the methods of mathematical game theory.

Despite the fact that the results of Axelrod are not so generally valid as Axelrod believed [9,19,31] it is no wonder that many partisans of the RC-approach saw these results as confirmation for the RC-approach and for the necessity to study social dilemmas by game theoretical methods. Sometimes it seemed that the social sciences had to be reduced to game theoretical analysis of decision problems. Yet these studies neglect several important difficulties.

Some time ago Lord Dahrendorf [7] formulated the important distinction between *homo oeconomicus* and *homo sociologicus*. *Homo oeconomicus* is the rational and egoistical actor, i. e. the basis of RC. *Homo sociologicus* in contrast is the social actor who orientates his actions to social rules and norms. He may be or not be an egoistical being but the important point is that his actions must be understood as the consequences of certain social rules that are valid in a specific situation. *Homo oeconomicus* understands a social action situation as a decision problem for finding the optimal strategy; his rationality, therefore, is the calculation for the optimal proportion between

means and ends. The rationality of *homo sociologicus* on the other hand is an interpretative one: he must understand the situation in order to know, which rules determine the situation and which according actions he has to perform.

It is important to see that the question of Hobbes, which Axelrod and other partisans of RC tried to answer with a game theoretical analysis of social dilemmas, does not arise with the assumption of *homo sociologicus* as the basic nature of *homo sapiens*. Man is, according to this position, a social being by nature, i. e., his human nature is defined not only by characteristics like language or consciousness but also by his sociality. The basic question is then not how social order is possible at all – the question Hobbes answered by introducing the law enforcing state – but *how the construction* of social order is performed and how and why certain forms of social order were changed and others were not. Dahrendorf of course named this pair of concepts in this manner because he was convinced that *homo sociologicus* is the normal aspect of man as a social being; *homo oeconomicus* is a rather special case in situations where either no valid social rules exist, or where some actors deliberately decide *not to* follow the respective rules, or where social rules leave some space of freedom, which must be filled out by individual decisions.

It is not an arbitrary decision if one chooses the option for one anthropological option or the other. There are many reasons for the validity of the *homo sociologicus* position:

a) Studies in behavioral biology teach us that nearly all biological species consist of social beings, whether predators, prey, or plants; there are exceptions as, e. g., otters that live as loners but these are indeed exceptions. From anthropological studies it is to be learned that all our hominid ancestors lived in social communities and that socio-cultural evolution would not have been possible without human sociality [37]. Therefore, one may safely assume that human sociality is part of our biological heritage and not something humans had to create [47].
b) Social rules, and in particular social institutions, are important tools to reduce the complexity of the world and in particular that of the social world. Without accepted rules each new situation would be a new decision problem that has to be calculated anew. In contrast to that, social institutions offer a relief of the burden of new decisions because they enable actions by just following the rules [6,10,44]. That is particularly favorable if the rules have been proven in past situations. Hence, the evolution of social rules is more probable than Hobbes' famous *bellum omnium contra omnes* ("The war of all against all").
c) By following social rules it is possible to gain a social reputation as being reliable. Such a reputation frequently is favorable if others are undecided if they should trust an actor or not.
d) Finally, countless social studies have demonstrated that social actors mostly are rule-obeying beings. In a strict sense the social sciences could not have been as successful as they are if they would not have made the basic assumption of *homo sociologicus*.

These considerations do not mean that RC is wrong but that it is incomplete. There are certainly cases where social actors indeed act as egoistical rational players in a strategic game as for example in cases of social deviance or political revolutions. In particular, it is obviously a characteristic of modern democratic societies that the respective social rules nearly always leave a space of freedom for individual decisions. The social role of a teacher, for example, defines the necessity to act in a pedagogical way and to evaluate the efforts of the pupils. But it is an individual decision of the teacher if he mainly fosters the pupils or mainly acts as a critical evaluator of the pupils' performances. In this sense RC is an apt theory for some peculiarities of modern societies but it is not a general theory of human social action.

The consequence of these meta theoretical considerations is the following: The construction of mathematical models of social systems and of course of cognitive systems too must be founded on the basic logical unit of *rule*, either social rules or cognitive ones. To be sure, the elements of social systems are individual actors if one follows the bottom up modeling approach. Yet these actors are determined by the respective rules of their social systems; their actions, hence, must be understood as the result of the application of the rules valid in the respective situation. Certainly actors can be in error about the specific rules and/or they can decide to be socially deviant. In these cases it is particularly important to analyze the cognitive processes that lead the actor to social deviance. Yet the center of model construction and model interpretation must be the according rules.

We shall see in one example below that also in cases of social conformity the respective cognitive processes must be taken into account if one wants to understand how certain conformity emerged. Therefore, the interdependency of social and cognitive processes must be always be *principally* taken into account, although frequently it is possible to understand and even predict social behavior without considering the according cognitive processes.

A Universal Modeling Schema

The considerations of the preceding sections can be summarized in form of a universal modeling schema that captures the interdependency of different levels, the bottom up principle, and the logical dominance of the rule concept.

The basic *social* level of the schema consists of individual social actors who interact according to certain social rules. One may visualize this level, e. g., as the grid of a two-dimensional cellular automaton or the structure of a Boolean net. The level constitutes a dynamical complex system whose behavior depends, as usual, on the specific interaction rules, the initial states, and eventually operations of certain meta rules. Figure 3 shows such a level, which we may call level 1:

The concentration on this level and the neglect of a cognitive one apparently follows the mentioned methodical demand of Durkheim – social phenomena must only be explained by other social phenomena. Social actions are exclusively understood by the reconstruction of the respective rules that determined the actions. It is well known that many social processes can be understood and explained by taking into account only this level.

However, particularly in cases of social deviance, but not only there, the knowledge about the respective rules is not sufficient. Because social actors act with respect to certain rules *and* their particular thinking, worldviews etc. in many cases an according cognitive level must considered too. For example, the deeds of religious fanatics can only be understood if one reconstructs their fundamentalist worldviews and the according individual legitimization of criminal acts. In such cases social actors must also be considered as complex dynamical systems, i. e. cognitive ones with according cognitive rules and elements. The schema has to be extended as is shown in Fig. 4; the cognitive level may be called level 0.

Level 0 and level 1 already allow the modeling of many social processes, in particular those where an interplay of social and cognitive dynamics must be taken into account. Yet often even these two levels are not enough, as was mentioned in the introduction, because it is frequently necessary to model aggregations of individual actors. If, for example, one wants to analyze the political situation in a certain country with respect to the different political parties, then the concept of party of course means a certain institution consisting of specific rules and all party members. The interaction between such parties – coalitions, conflicts and so on – should be modeled on another level, namely level 2, the level of aggregations of actors (Fig. 5).

Note that level 2 must also be understood as a complex dynamical system with its own specific rules and elements. The elements there are usually called collective actors. All the three levels, hence, constitute certain complex systems whose dynamics is determined a) by the respective rules, elements, and initial states, b) by the eventual operation of meta rules that change the rules and in consequence the generation of states, and c) the interdependency between the different levels. This interdependency is visualized in

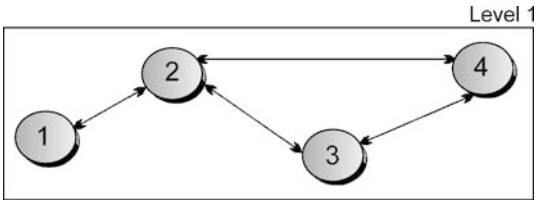

Social Cognitive Complexity, Figure 3
The social level as a complex dynamical system

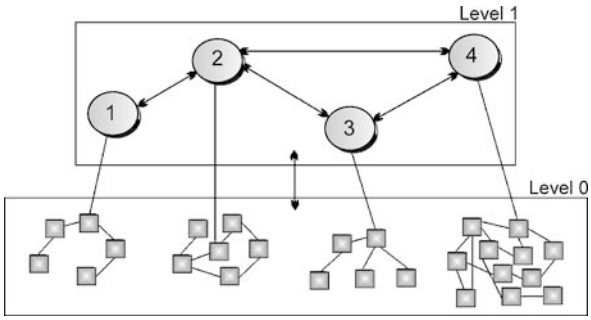

Social Cognitive Complexity, Figure 4
Social actors modeled as cognitive complex systems

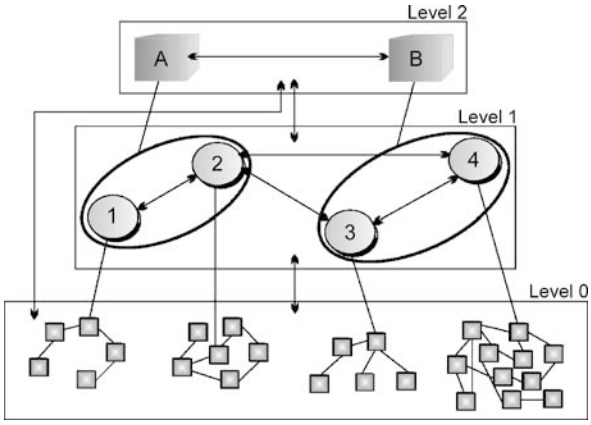

Social Cognitive Complexity, Figure 5
Social actors (level 1) modeled as cognitive complex systems (level 0) and as aggregations (level 2)

the pictures by the arrows that represent the mutual influences.

In contrast to many schemas of multi level models, e. g. the schema shown in the introduction, it must be noted that there is *in principle* a permanent feedback between all different levels. The action of individual actors (level 1) generate and reproduce the collective actors on level 2, yet collective actors in turn influence and determine the individual actors. A member of a political party acts *as* a party member. Accordingly the cognitive processes of individual actors determine their actions but these cognitive processes in turn are influenced by the result of the actions on the one hand and the collective actors the individual actor is a part of on the other. A party member not only acts but also thinks in concepts of the respective party. In addition, individual cognitive processes may directly influence collective actors, particularly in times of social and/or political change. History is full of the impact of individual cognitive processes on whole institutions, if the individual actors operate as reformers or revolutionaries.

It is certainly possible to extend even the schema shown in Fig. 5 by either adding a third social level, consisting of aggregations of collective actors, or by adding a second cognitive level by distinguishing between, e. g. conscious and unconscious thinking. Such extensions would lead to four- or even five-level models. Yet in most cases of social-cognitive modeling it is sufficient to operate with one, two or three levels. To be sure, it is always a question of the specific research interest how many levels and which ones one must take account of. The schema, after all, is just a *schema*, that is a frame work to guide the modeling processes when dealing with socio-cognitive processes – and not only these. The construction of a certain concrete model for a specific problem might be still a difficult task, in particular if one has to evaluate the validity of the specific model.

It is rather evident that this modeling schema can practically be applied to all complex social-cognitive processes. Yet the schema is in an even more ambitious sense a *universal* schema. This is a consequence from the mathematical characteristics of the algorithms mentioned in Sect. "Social-Cognitive Complexity, Algorithms, and Bottom Up Models". Imagine for example that level 1 is modeled by a cellular automaton, level 0 by one or several types of artificial neural networks, and level 3 again by a cellular automaton or a Boolean net. It is known from mathematical investigations that cellular automata are potentially equivalent to Universal Turing machines [36], which roughly means that no formal system is more general than such systems. This is the so-called Church–Turing thesis, which no mathematician seriously doubts. A corollary, the *physical* Church–Turing thesis, states that each physical, i. e. empirical system can be modeled via the use of a suited Universal Turing machine or an equivalent system. Although the physical Church–Turing thesis has not been proven in an exact mathematical manner nobody doubts this thesis either.

Without great difficulties it can be shown that each neural net and each Boolean net is equivalent to an according cellular automaton – the mappings between cellular automaton and the two classes of nets are injective. Hence principally each complex system can be modeled by using this schema *and* by using one or several of the mentioned algorithms. An important consequence from these considerations is the fact that a mathematical science of social-cognitive systems is not only possible but also that there is for each problem a general way how to treat it in a mathematical manner. The schema, therefore, can be understood as a constructive mapping from socio-cognitive phenomena to mathematical structures. The universality of the modeling schema is a constructive proof of the possibility of mathematical social and cognitive sciences.

Examples

These considerations show that and how mathematical models of social-cognitive processes can be constructed. Yet the most convincing proof is of course the demonstration of concrete examples, i. e., the models of specific complex social-cognitive systems. More detailed descriptions of these and other examples can be found in [20] and [21].

The Differentiating of Social Groups into Subgroups

A certain social group that consist of more than, say, three or four members is usually not a homogenous entity but is differentiated into several subgroups. The criteria for the generation of such subgroups are, of course, not always the same. In working teams for example the different qualifications of the members might be a criterion for the forming of specialized sub-teams. Yet in many informal groups a principle is certainly valid that may be called Homans' Principle [17] and that can be stated as follows: "Members of a social group favor interactions with people they like and they avoid interactions, if possible, with people they do not like." Everyday experience shows that this principle is indeed in many social groups an important criterion.

Since the foundation of sociometry by Moreno [28] the usage of a so-called sociomatrix or Moreno matrix respectively for the representation of social relations has become customary. If one wants to represent the mutual feelings

of the members of a certain group via a sociomatrix the matrix may be:

	a	*b*	*c*
a	0	1	−1
b	1	0	1
c	−1	0	0

The matrix is coded with three values. Value 1 means a positive feeling or a liking respectively, value 0 means neutral or indifferent feelings, and −1 means negative feelings or disliking respectively. Hence the matrix can be understood as member *a* likes *b* and dislikes *c*, member *b* likes *a* and *c*, and member *c* dislikes *a* and is indifferent to *b*. The usage of such a sociomatrix as a mirror of the emotional structure of a certain group can be very important for, e. g. teachers, group leaders, group therapists or trainers of sport teams.

According to the Homans principle our research group COBASC constructed a cellular automaton with the goal to simulate and predict the emerging of different subgroups in specific groups, which we called in honor to Moreno the Moreno-CA (COBASC is the abbreviation for *Computer Based Analysis of Social Complexity*). It is a two-dimensional cellular automaton whose cells represent the members of the particular group that shall be modeled. The state of the cells represent the emotional states of the group member that range from "very dissatisfied" to "very satisfied." The cells state is dependent of the eight adjacent cells (a "Moore neighborhood") in the following way: the values of the emotional relations of the center cell to the eight cells of the Moore neighborhood are summed up; the emotional state of the center cell then is the arithmetical mean of these values. (The cells are usually modeled as squares; hence there are eight adjacent cells). The values of the emotional relations are coded as 1, 0, or −1 as in the example above. The task of each cell is to find a Moore neighborhood where its own emotional state is a maximum. Accordingly the rules of the Moreno-CA are:

a) Place as many cells on the grid as there are group members; the rest of the grid's cells are free places.
b) Compute the emotional state for each "member" cell; vacant cells get the value 0.
c) Let each cell look on the grid for a Moore neighborhood where it can obtain a maximum state value.
d) Move each cell into that neighborhood.
e) If an attractor state, preferably a point attractor, has been reached, stop; if no attractor state could be reached after a certain number of runs, stop too. (A point attractor is a state the system will not leave although the rules are still operating; an attractor of longer period consists of several different states that the system will generate in cyclic order.)

The last rule is necessary because it can happen that the group reaches no attractor state at all or only those with rather long periods [21]. Factually the rules are a bit more complicated but this simplified version is the essence.

We validated this model with different social groups and obtained in all cases very satisfactory results. One empirical example shall serve to illustrate the methodical procedure:

With the permission of the teacher we asked a group of eight pupils of a primary school, with which pupils they would like to share a room in a youth hostel, with which pupils they would not share a room under any circumstances, and with which pupils they would share or not, according to the circumstances in the youth hostel. (It is often problematic to ask children (in this case 10 years old) directly if they like or dislike other children. That is why we chose this procedure of asking). These answers were transferred in the respective sociomatrix of the group. Afterwards the pupils were ordered to go into a strange classroom and choose places in order to wait for the teacher. The hypothesis was of course that the pupils would choose places according to their emotional relations to the other pupils. Then the values of the factual sociomatrix were inserted into the Moreno-CA; it reached a point attractor after 8 runs. The prediction of the program was then compared with the factual sitting order of the pupils (Fig. 6).

The program of course did not predict the absolute spatial distribution of the pupils in the classroom, for example if one pupil prefers to sit at a window. The Moreno-CA just predicts the relative positions of the pupils, i. e., which pupils are sitting together or which pupils are outsiders and are sitting alone. A comparison of the factual distribution and the prediction of the cellular automaton shows that the program indeed correctly prognosticated the sitting order for nearly all pupils with the exception of

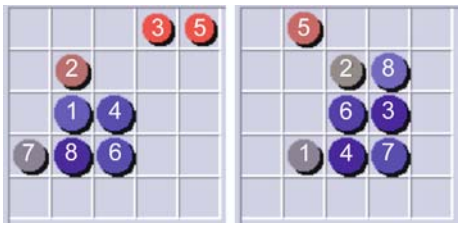

Social Cognitive Complexity, Figure 6
The factual sitting order of the pupils (*left side*) and the prediction of the Moreno-CA

pupil 3. He is factually sitting besides the outsider pupil 5. However, the teacher of the children could give an explanation for this error of the program: Pupil 3 is known to be a practical jester who wanted to win the applause of the other children by teasing the outsider. Such melancholy character traits of course the program could not know.

We inserted the same sociomatrix into another program, namely a so-called self-organizing map or Kohonen feature map respectively (see next example). This is a special type of artificial neural nets. This net obtained even better results than the Moreno-CA because it placed pupil 3 beside pupil 5. An explanation for this result is given in [21].

To be sure, we did not want to predict just sitting orders of children in a classroom. The social experiment and the simulation by the two programs should show which subgroups existed in the whole group of children; this differentiation would become visible in the little experiment. The teacher, by the way, confirmed the results of the experiment and the prediction because according to her thorough knowledge the group of pupils factually consisted of an outsider and a homogenous group of all other pupils.

With respect to the modeling schema of the previous section this model uses only level 1, namely the social one. To be sure, in an implicit way, the emotional state of the children also plays a decisive role: there are no explicit social rules in this situation and the desire of the children to be in the neighborhood of their friends and to avoid other children is, according to Homans' principle, the main determination of the factual behavior. But in the model it was sufficient to represent the children as finite state automata, i. e., units whose state values are a direct consequence of their respective social milieu – in the model the Moore neighborhood. The Moreno-CA and the social experiment with the children demonstrate that indeed in many cases it is sufficient to construct models with only one level, in particular the social one.

Socialization by Learning from a Model

In Sect. "Introduction" it was demonstrated why theories of socialization always must take account of both social level(s) and the cognitive one. The social level represents the social milieu that is decisive for the specific socialization process; if one wants to explain the particular characteristics of this milieu one must consider additional social levels. The cognitive level of course represents the personality or identity of the individual that is formed or influenced by the respective milieu. The term formed though does not necessarily mean that socialization is a passive process. The socialized individual has the task to actively construct a worldview consisting of cultural values, social rules, and a culture specific knowledge.

One of the most known theories of socialization is the theory of learning from a model by Albert Bandura [4]. This theory states that socialization is to a high degree dependent on other people in the social milieu, who can act as models, i. e. as paradigms. The socialized individual perceives these models as positive or negative examples for its form of thinking and behavior. In other words, young people and children do not learn cultural values and specific forms of social behavior by internalizing general rules but by concrete examples they can observe and imitate. (There is a striking similarity between the socialization theory of Bandura and the cognitive theory of prototypes by Eleanor Rosch [23], which can only be mentioned here.)

We applied this theory to a real case, namely a male youth from the *Ruhrgebiet*, an industrial area in the West of Germany. The youth, who shall be named Tom, of course a fictitious name, grew up in a workers family; he told us his story in an interview that was performed by one of our students. The father was an aggressive and violent man who often hit Tom's mother and frequently got involved in brawls and fights in the pubs. The mother was a submissive person who tried to keep up the family by working as a cleaner. Tom, therefore, had to care for his two younger sisters, which he very much disliked although he liked his sisters. When Tom was 10 years old his father left the family and disappeared. At the time of the interview Tom was 17 years old and had never heard any more from his father. Because Tom joined a gang of hooligans and got into trouble with the authorities and because the mother was unable to care for the two girls and Tom alike, Tom was admitted to a hostel of delinquent youths.

Despite the behavior of the father Tom adored him because *the father was a real man*. "Real" men are not afraid of other men, they become violent and hit if they feel insulted and in particular they dominate women, if necessary also by violence. "Real" women, on the other hand, are submissive and obedient to their men; they cook, wash and care for the children while real men work outside the house. As a consequence Tom developed a dichotomous worldview: men are strong and women are weak; women should obey and men command. The worlds of men and women must not be mixed in the sense that women take over tasks of men and men in turn care for the household. Consequently Tom despises the educators in the hostel because the female educators try to give orders to young men like Tom and the male educators perform household tasks like cleaning and cooking.

Tom apparently obtained his worldview by learning from the positive model of his father – "I want to be a man

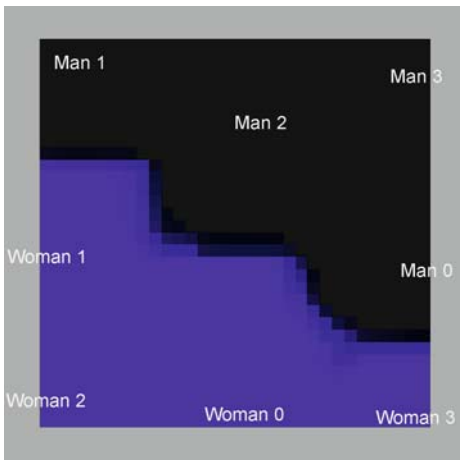

Social Cognitive Complexity, Figure 7
The reconstruction of a dichotomous worldview: female versus male clusters

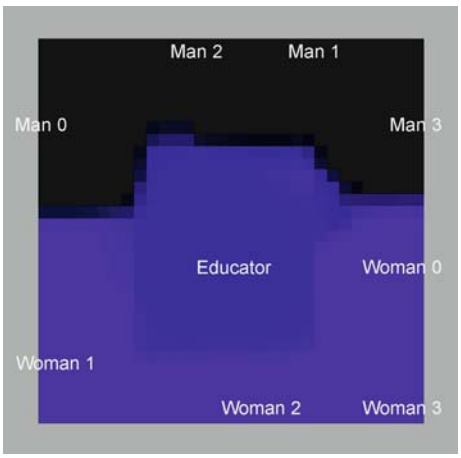

Social Cognitive Complexity, Figure 8
Increasing the female cluster by adding other "real women" and "false" men

like my father" – and the negative model of his mother. He also remarked that most men he met in his childhood were similar to his father and the women were similar to his mother. Hence he was early influenced by a homogenous milieu, where men and women all occupied the same gender specific roles.

For a simulation of Tom's biography with the result of a dichotomous worldview we used a Self-Organizing Map (SOM) or Kohonen Feature Map respectively that belongs to the class of neural networks, which learn in a non-supervised manner. In contrast to the forms of supervised learning non-supervised learning is not determined by externally given targets but the learning process is characterized by an internal logic. To put it into a nutshell, non-supervised learning is the ordering of a set of data according to certain internal criteria. There are different types of SOMs; we used a so-called Ritter–Kohonen model [38]. The technical details do not matter here; they can be looked up in any textbook on neural nets. The main intention of the simulation was to reconstruct the genesis of Tom's worldview and to demonstrate the according behavior in the present. Hence we inserted formal representations of the father and the mother as positive and negative models and inserted additional persons into the SOM. The SOM then constructed clusters of the different persons who are distinguished according to the criterion "real men" versus "real women". A first result is shown in Fig. 7.

The spatial nearness on the screen represents similarities or differences between the clustered persons respectively.

Afterwards we inserted other persons who differed from the first in certain personal respects. In particular we inserted formal representations of male educators who had female characteristics like cooking or cleaning. The SOM placed these men near the female cluster according to Tom's worldview: Tom does not accept these educators as real men but sees them more as women in male disguise. This is shown in Fig. 8.

The SOM apparently is able to simulate the important parts of Tom's socialization processes. As a SOM is a purely deterministic algorithm the successful simulation of Tom's cognitive genesis must be seen as a deterministic process too. Tom, so to speak, had no chance to become another person because the social circumstances determined his biography in a complete way.

The model uses only level 0, i. e., the cognitive one. Yet the social level is implicitly present too because of the inputs we inserted to the program; these inputs represent the social milieus in which Tom grew up and presently lives. In addition, the worldview of Tom has certainly social consequences not only for his biography. His membership in a hooligan gang, where he feels at home, indicates that his biography is representative for many youth members of such gangs. The according consequences for societal problems are obvious. In this respect the example of Tom also contains level 2, e. g. the level of collective actors like the different hooligan gangs and the social institutions that have to deal with them.

Modeling Social-Cultural Evolution

The last example is a rather complex model whose theoretical foundations and historical arguments can be read up in [20]. In contrast to the two preceding examples that be-

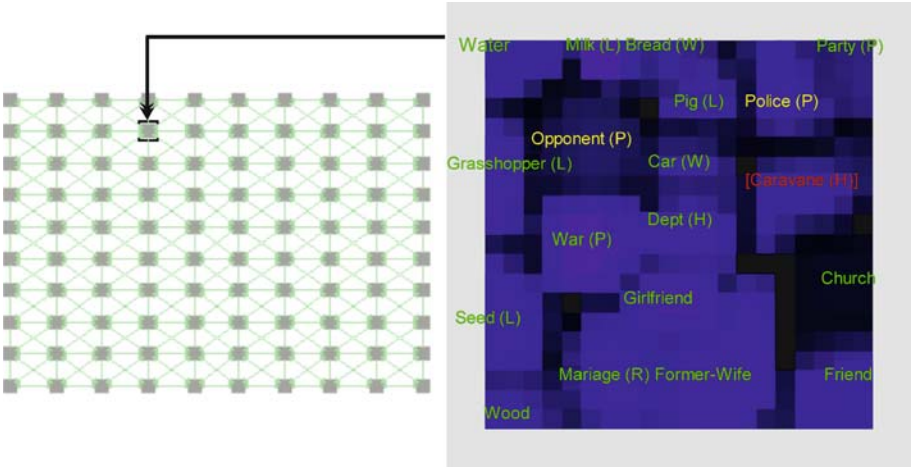

Social Cognitive Complexity, Figure 9
Cognitive state of an individual actor and his place in the artificial society

long to the field of micro-sociology this model deals with the evolution of whole societies. The basic assumption is that socio-cultural evolution is generated by creative inventions of individual actors and by the acceptance of these new ideas by the respective society. The social structure and the cultural norms of a society may promote or hinder the generation and acceptance of new ideas. In particular, the decisive factor is the degree of role autonomy for the occupants of creative roles like engineers, artisans, artists or scientists. If this degree is sufficiently high then the respective society evolves, i. e. increases its cultural capacity and generates social structures that are more efficient to deal with the problems of the society than old ones. If the degree is low, then the society stagnates, i. e. it becomes caught in a cultural and socio-structural attractor.

Comparative studies of the history of many societies always showed the same results, namely an early blossoming and cultural growth, then a period of consolidation, and finally stagnation or even regression, i. e. the return to some previous states [40,42]. According to our theoretical assumptions in these cases the degree of role autonomy was not high enough to allow an unhindered evolution. Only one case is known where apparently cultural evolution did not get caught in an attractor, namely the culture of European Modernity that was the historical basis for the contemporary Western culture. Historical studies confirm our hypothesis: In all cases of socio-cultural evolution the degree of role autonomy was much lower than in the European Modernity and the Western culture in general.

To validate these considerations by the use of a mathematical model we constructed the so-called socio-cultural-cognitive algorithm (SCCA). The model consists of a) a social level that is constructed as a cellular automaton. As usual the cells represent individual actors, i. e. members of a certain society. The rules of the cellular automaton determine the degree of role autonomy, i. e. the proportional ability of the actors to learn and create new ideas on the second level. b) The second level is a cognitive one: Each cell consists of a combination of different neural networks, namely several so-called bi-directional associative nets (BAM) and a SOM from the type used in example "Socialization by Learning from a Model". The task of the BAM is to associate observed characteristics from certain objects with the according semantic concepts – small, furry, and meowing is associated with cat. The task of the SOM is to order the concepts into a semantical network. These cognitive processes can be performed either in the way of learning, i. e. taking over new concepts from other actors, or by creating new concepts. The sum of the results of the individual cognitive processes is defined as the state or level that the respective culture has reached. Figure 9 shows an individual actor and its place on the grid of the cellular automaton.

The degree of role autonomy is inserted into the rules of the cellular automaton. We did many experiments with different values of this degree. The most important result is that in most cases the according culture indeed showed an evolutionary pattern known from human history, in particular described by the great British historian Arnold Toynbee (Fig. 10).

Only in the few cases with a high degree of role autonomy our model showed an evolutionary path known from the history of European Modernity. This is shown in Fig. 11.

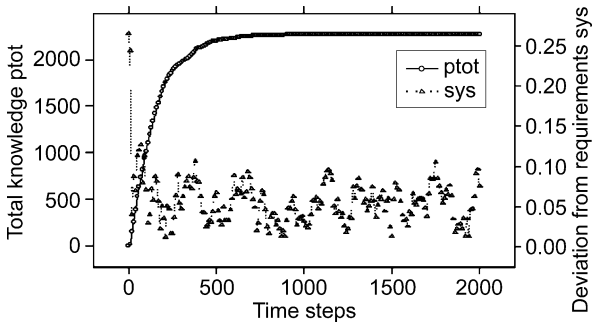

Social Cognitive Complexity, Figure 10
A normal path (Toynbee path) of cultural evolution

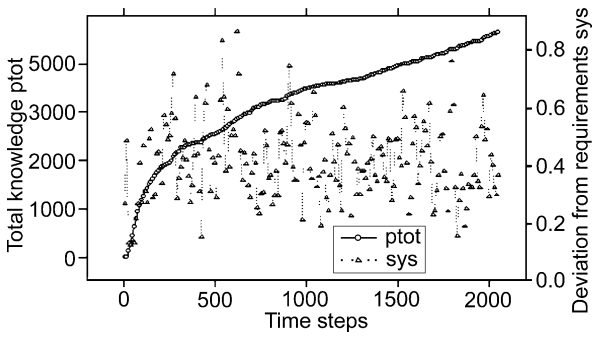

Social Cognitive Complexity, Figure 11
A "Western development"

Societies with a sufficient high degree of role autonomy are not hindered in their cultural development although of course such societies may be stopped in their evolution by external forces. That could have been the reason for the final stagnation of the ancient Greek culture that also was characterized by a high degree of role autonomy for the occupants of the creative roles. Our model confirms that such societies are very seldom, which may be indeed the main reason for the fact, as Toynbee [42] observed, that the overwhelming majority of historically known societies stagnation was their inevitable fate.

The model is characterized by interplay of the two levels with consequences for a third one because the social structure determines the individual creative processes that in turn determine the fate of the whole society. In addition, sufficient developments on the cognitive level may also change the social structure: the better the cognitive improvements are, the higher becomes the degree of role autonomy on the social level. Such a rule is confirmed too by historical observations of the European history: In the Middle Ages the degree of role autonomy was not much higher than in contemporary cultures like feudal China or the Islamic societies. By the cultural progress during European Modernity the degree of role autonomy permanently increased. The model, hence, shows the consequences of such interdependent dynamics of the different levels that may explain important courses of history with respect to the fate of whole cultures.

Conclusions and Future Directions

The general possibility of a mathematically formulated science of social and cognitive processes was demonstrated in the Universal Modeling Schema and the fact that suited formal models like cellular automata or neural networks are equivalent to Universal Turing Machines. How many levels one needs for a concrete model, which ones, and how the relations between the levels are to be formulated is of course a question of the specific research problem. Constructing models of socio-cognitive processes will never be a trivial task. The modeling schema gives general orientations, but the detailed knowledge of the problem and in particular of the empirical data is always a necessary condition for the construction of social-cognitive models. That is of course a truism if it is stated in such a general manner but one that must always be taken into account. In particular, the knowledge about model construction and of suited algorithms never substitutes the profound knowledge about the specific fields of research.

The general relation between a social level and a cognitive one can also be stated in terms of individual actors. If one takes for example a most basic social situation, i. e. the communicative interaction between an actor A and another actor B, and if one assumes that during the interaction neither the social rules of this situation nor the cognitive processes of the actors change, then this interaction can be written the following way:

If f designates the set of social rules decisive for this situation and if g_A and g_B designate the cognitive processes of A and B respectively, then

$$f(A) = g_B(B)\,.$$

In other words, the application of the social rules f on the personality of A causes A to send an appropriate message to B, who in turn is moved to perform the cognitive processes g_B. In this case the social level determines the cognitive one, as is certainly the case in most forms of everyday communication. Accordingly the equations can be written if the cognitive processes determine the social rules, for example if a student becomes more intelligent than his professor by the teaching of the professor. (For general formulations of these equations see also [21]).

The most important problem for a socio-cognitive science that operates with mathematical models and accord-

ing computer simulations is without doubt the proper relation between theory, formal model and simulation program. A complete research program that has to take care of all three components follows the order: theory construction or theory selection respectively – transformation of the theory into a suited model – construction of a simulation program – comparing the results of the simulation runs with known empirical data. If everything went right, the last step will be a confirmation of the whole process. Yet since Murphy's law we know that each step may contain errors and if the data are not compatible with the simulation results then the error(s) may be in the theory, the model, the simulation program, or in several of the steps.

In particular the fundamental role of theory has frequently been neglected in the many attempts to construct a mathematical or computational respectively social science by using computer models of socio-cognitive phenomena. The mentioned dominance of RC-models on the foundations of Game Theory is just one indicator for this problem. Yet the social and cognitive sciences are much more than theories about rational and egoistical players in strategic games. Therefore, the most important problem for the future is the careful analysis of the research program formulated in the three or four steps mentioned above. Only then the use of mathematical models and suited simulation programs will be the methodical basis for a real socio-cognitive science.

Yet we have no alternative to the development of such sciences. If this research program is not successful we will always be the unconscious slaves of the social conditions we produced ourselves, in particular of a future we did not wish for. A science of the complex socio-cognitive reality is certainly not a sufficient condition for the solving of all our social problems. But it is a necessary condition for the task that we must understand our own species, that is to understand our own social and cognitive processes.

Bibliography

Primary Literature

1. Alexander JC, Giesen B, Münch R, Smelser NJ (eds) (1987) The Micro-Macro-Link. University of California Press, Berkeley
2. Axelrod R (1984) The Evolution of Cooperation. Basic Books, New York
3. Axelrod R (1987) The Evolution of Strategies in the Iterated Prisoner's Dilemma. In: Davis L (ed) Genetic Algorithms and Simulated Annealing. Morgan Kauffman, Los Altos
4. Bandura A (1986) Social Foundations of Thought and Action. A Social Cognitive Theory. Prentice Hall, Englewoods Cliff
5. Beltrami E (1993) Mathematical models in the social and biological sciences. Jones and Bartlett, London
6. Berger P, Luckmann T (1966) The Social Construction of Reality. Doubleday, New York
7. Dahrendorf Lord R (1958) Homo Sociologicus. Ein Versuch zur Geschichte. Bedeutung und Kritik der Kategorie der sozialen Rolle. Leske Buderich, Opladen
8. Fararo TJ (2001) Social Action Systems. Foundation and Synthesis in Sociological Theory. Praeger, London
9. Fogel DB (1993) Evolving Behaviors in the Iterated Prizoner's Dilemma. Evol Comput 1:77–97
10. Gehlen A (1956) Urmensch und Spätkultur. Frobenius, Bonn
11. Geulen D, Hurrelmann K (1982) Zur Programmatik einer umfassenden Sozialisationstheorie. In: Hurrelmann K, Ulich K (eds) Handbuch der Sozialisationsforschung. Beltz, Weinheim, pp 5–67
12. Giddens A (1984) The Constitution of Society. Outline of the Theory of Structuration. Polity Press, Cambridge
13. Gilbert N, Troitzsch KG (1999) Simulation for the Social Scientist. Open University Press, Buckingham
14. Goonatilake S, Khebbal S (eds) (1995) Intelligent Hybrid Systems. Wiley, London
15. Habermas J (1981) Theorie des kommunikativen Handelns, vol 2 (Theory of communicative action). Suhrkamp, Frankfurt
16. Hegselmann R, Mueller U, Troitzsch KG (eds) (1996) Modeling and Simulation in the Social Sciences from the Philosophy of Science Point of View. Kluwer, Dordrecht
17. Homans GC (1950) The Human Group. Harcourt, Brace, Jovanovich, New York
18. Kauffman S (1993) The Origins of Order. Oxford University Press, London
19. Klüver J (2000) The Dynamics and Evolution of Social Systems. Kluwer, Dordrecht
20. Klüver J (2002) An Essay Concerning Socio-Cultural Evolution. Theoretical Principles and Mathematical Models. Kluwer, Dordrecht
21. Klüver J, Klüver C (2007) On communication. An Interdisciplinary and Mathematical Approach. Springer, Dordrecht
22. Knorr-Cetina K, Cicourel AV (eds) (1981) Advances in Social Theory and Methodology. Toward an Integration of Micro- and Macro-Sociologies. Routledge and Kegan Paul, Boston
23. Lakoff G (1987) Women, Fire, and Dangerous Things. What Categories reveal about the Mind. The University of Chicago Press, Chicago
24. Langton C (1988) Preface. In: Langton C (ed) Artificial Life. Addison-Wesley, Redwood
25. Levy S (1993) Artificial Life. Penguin Books, London
26. Luhmann N (1984) Soziale Systeme. Suhrkamp, Frankfurt
27. Marx K, Engels F (1969) Deutsche Ideologie (German Ideology). MEW 3
28. Moreno JL (1953) Who shall survive? Foundations of sociometry, group psychotherapy and sociodrama. Beacon House Inc., Beacon
29. Münch R (1988) Theorie des Handelns. Suhrkamp, Frankfurt
30. Nowak A, Levenstein M (1996) Modeling Social Change with Cellular Automata. In: Hegselmann R, Trotzsch K, Muller U (eds) Computer simulations from the philosophy of science point of view. Kluwer, Dordrecht
31. Nowak MA, May RM (1993) The Spatial Dilemma of Evolution. Int J Bifurc Chaos 3:35–78
32. Parsons T (1968) The Structure of Social Action. Academic Press, New York
33. Parsons T, Platt GM (1973) The American University. Harvard University Press, Cambridge

34. Piaget J (1972) The Principles of Genetic Epistemology. Routledge, London
35. Piaget J (1972) The psychology of intelligence. Littlefield Adams, Totowa
36. Rasmussen S, Knudsen C, Feldberg R (1992) Dynamics of Programmable Matter. In: Langton C, Taylor C, Farmer JD, Rasmussen S (eds) Artificial Life II. Addison-Wesley, Redwood
37. Read DW (2005) Change in the Form of Evolution: Transition from Primate to Hominid Form of Social Organization. J Math Sociol, Special Issue Emergence Soc Order 25:3
38. Ritter H, Kohonen T (1989) Self-organizing Semantic Maps. Biol Cybern 61:241–254
39. Schelling TC (1971) Dynamical Models of Segregations. J Math Sociol 1:143–186
40. Spengler O (1926) The Decline of the West. Deutsches Original: (1922) Der Untergang des Abendlandes. Alfred Knopf, New York
41. Suleiman R, Trotzsch KG, Gilbert N (eds) (2000) Tools and Techniques for Social Science Simulation. Physica, Heidelberg
42. Toynbee A (1934–1961) A Study of History (12 vols). Oxford University Press, Oxford
43. Troitzsch KG, Mueller U, Gilbert N, Doran J (eds) (1996) Social Science Microsimulation. Springer, Heidelberg
44. Vanberg VJ (1994) Rules and choice in economics. Routledge, New York
45. von Neumann J, Morgenstern O (1944) Theory of Games and Economic Behavior. Princeton University Press, Princeton
46. Weber M (1981) Die protestantische Ethik. Mohr, Tübingen
47. Weingart P, Mitchell SD, Richerson PJ, Maasen S (eds) (1997) Human by Nature. Between Biology and the Social Sciences. Lawrence Erlbaum, Mahwah
48. Zadeh LA (1994) Fuzzy Logic, Neural Networks and Soft Computing. Comun of the ACM 37:77–84

Books and Reviews

Alexander JC (1983) Theoretical Logic in Sociology, vol 3. The Classical Attempt at Theoretical Synthesis: Max Weber. Melbourne-Henly, London
Brennan G, Buchanan M (1985) The Reason of Rules. Constitutional Political Economy. Cambridge University Press, Cambridge
Coleman JS (1974) Power and the Structure of Society. Norton, New York
Coleman JS (1990) Foundations of Social Theory. Cambridge, London
Diekmann A, Schmidt P (1998) Environmental Sociology, Special Issue of Rationality and Society. Sage, London
Ginsberg A (1990) Connecting Diversification to Performance. A Sociocogn Approach. The Acad Manag Rev 15:3:514–535
Hamilton D, Devine P, Ostrom T (eds) (1994) Social Cognition: Impact on Social Psychology. Academic Press, London
Howard J (1994) A social cognitive conception of social structure. Soc Psychol Q 57:210–27
Howard RW (1995) Learning and Memory. Major Ideas, Principles, Issues and Applications. Praeger, Westport
Parsons T (1951) The Social System. The Free Press of Glencoe, Glencoe
Resnick LB, Levine JM, Reasley SD (ed) (1991) Perspectives on socially shared cognition. American Psychological Association, Washington
Rogoff B (1990) Apprenticeship in thinking: Cognitive development in social context. Oxford University Press, New York
Schech S, Haggis J (2000) Culture and Development. A Critical Introduction. Blackwell Publishers, Malden
Singley MK, Anderson JR (eds) (1989) The transfer of cognitive skill. Harvard University Press, Cambridge
Snijders T, Snijders TAB, Bosker R (2000) Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. Sage Publications, London
Walsh DM (ed) (2001) Naturalism, Evolution and Mind. University Press, Cambridge
Wertsch JV (ed) (1985) Culture, communication and cognition: Vygotskian perspectives. Cambridge University Press, Cambridge
Winch C (1998) The Philosophy of Human Learning. Routledge, London

Social Coordination, from the Perspective of Coordination Dynamics

Olivier Oullier[1,2], James A. S. Kelso[2]
[1] Laboratoire de Neurobiologie Humaine (UMR 6149), Aix-Marseille Université, Marseille, France
[2] Human Brain and Behavior Laboratory, Center for Complex Systems and Brain Sciences, Florida Atlantic University, Boca Raton, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Intentional Interpersonal Coordination
Issues in Quantifying Spontaneous Interpersonal Coordination
Human Spontaneous Synchronization
Shared Behavioral and Neural Social Coordination Dynamics
Social Memory and the Dependence on Initial Conditions
Future Directions
Acknowledgments
Bibliography

Glossary

Self-organization Self-organization lies behind all structure and pattern formation in nature's complex systems, including the human brain. Self-organization is a principle governing a system where no agent-like entity is ordering the elements, telling them where and what to do. In self-organizing systems, low-di-

mensional dynamics are revealed by changing one (or more) control parameter(s) whose role is simply to move the system through a series of state changes without prescribing its behavioral patterns.

Coordination dynamics Coordination dynamics seeks the laws, principles and mechanisms underlying the coordinated behavior of different kinds of components at multiple levels of description (molecules, cells, circuits, etc). It is an overarching conceptual framework that describes, explains and predicts how patterns of coordination form and change at multiple levels of brain and behavior. The brain, mind and behavior are linked by virtue of sharing a common underlying coordination dynamics.

Information exchange A remarkable fact is that in contrast to classical dynamics that deal with fundamental quantities such as mass, length and time and their relations, coordination dynamics is informational in nature, dealing with informational quantities of a relational kind that couple different parts of a system or different systems.

Phase transitions Phase transitions are the true illustration that a system is self-organizing. They are spontaneous qualitative pattern changes occurring as parameters are changed quantitatively. When they occur, abrupt switches from one coordinated pattern to another are observed and the dynamics of the entire self-organizing system is dominated by one or a few collective variables: the order parameters.

Stability Stability is a key concept in coordination dynamics. Here the stability is of coordination or collective variables. The (loss of) stability of a self-organizing system indicates whether a phase transition is to occur. In order to evaluate the stability of a system, one can perturb it and measure the time it takes for the system to return to its initial state, i. e. its *relaxation time*. A number of other converging measures have been used to measure stability in coordination dynamics such as *switching time* (the time it takes for the system to switch from one pattern to another when phase transitions occur) and *critical fluctuations* (the increase of variability of the collective variable in the vicinity of the phase transition).

Definition of the Subject

Social Coordination Dynamics (SCD) explores, at both behavioral and neural levels, the mechanisms mediating the formation and dissolution of bonds between individuals. SCD applies the concepts, methods and tools of informationally coupled self-organizing systems (coordination dynamics) to quantify real time social processes. Just as coordination dynamics deals with how the parts of complex systems work together in a meaningful way to achieve goals, so SCD aims to understand the interplay of forces operating at both individual and collective levels to produce effective social behavior. SCD offers a novel perspective and new metrics to explore systematically a fundamental form of human bonding (or lack thereof), and the self-organizing processes that underlie its persistence and change over space and time. SCD therefore complements recent developments in several fields such as sociology, social cognitive neuroscience, behavioral economics, game theory and neuroeconomics.

Introduction

Coordination can be broadly defined as a functional ordering among interacting components in space and time. Coming in many guises, coordination represents one of the most striking features of living organisms. The science of coordination, *Coordination Dynamics* (abbreviated CD) [36,37,38,41] stems from a complex systems framework based on the theory and methods of informationally coupled self-organizing dynamical systems (see ► Coordination Dynamics). CD explores a number of basic coordination phenomena that cut across a wide range of levels, creatures and functions. Of particular relevance to social coordination are: (i) patterned states of coordination remain stable in time despite perturbations; (ii) component parts and processes (dis)engage in a flexible fashion depending on functional demands and/or changes in environmental conditions; (iii) multiple coordination states exist rendering living things *multi-functional*, effectively satisfying the same (or different) set of circumstances; (iv) switching from partially to fully coordinated states and vice versa is commonplace; (v) selection of coordination patterns is tailored to suit the current needs of the organism; (vi) coordination patterns adapt to changing internal and external contingencies; (vii) depending on a balance between competitive and cooperative processes, learning may take the form of abrupt transitions from one coordinated pattern to another; and (viii) the system may remain in the current pattern of coordination even when conditions change thus exhibiting memory.

The foregoing list contains some of the core aspects of CD reflecting its inherently nonlinear and emergent character. Such phenomena appear so spontaneously and so consistently as to suggest the existence of an underlying lawfulness or regularity that transcends the multitude of differences between different systems and the settings in which they can be observed [41,45].

Coordination achieves its pinnacle in the vast array of cells and connections called the human brain, and in the collection of human beings we call society [41,45]. How social interactions form and change in complex systems and contexts is of great interest to many disciplines, particularly psychology, biology, physics, economics and the social sciences. The primary focus of the present article is to review recent work investigating the coordination dynamics of individuals interacting with each other in real time. At the core of all personal relationships is how the other becomes intertwined with the self. *Social coordination* is the tendency of two or more individuals to coordinate their ongoing actions with each other based on mutual information exchange. *Social Coordination Dynamics* (abbreviated SCD) is a theoretical-empirical framework that investigates the behavioral and neural dynamics of bond formation between individuals, operationalized in terms of how they spontaneously synchronize their behavioral and neural patterns [82,99].

Synchronization is a form of spontaneous pattern formation that operates according to general principles of self-organization described by nonlinear dynamics [25,26,72]. Following on Huygens's analysis of two clocks synchronizing on a wall, many studies have framed the problem of mutual synchronization in terms of a network of oscillators each of whose individual behavior is altered by nearest neighbor interaction [5,7,30,57,106,107]. Under that framework synchronization has been observed among very different entities in a broad range of physical, biological and social systems. Human brains (and behavior) have proven no exception to these principles [19,27,41,50,51,89]. Experiments have revealed that humans exchange information – whether uni- or multimodal in nature – to spontaneously adopt and switch coordination patterns (e. g. [37,53,58]).

The validity of the measures and constructs from coordination dynamics are worth mentioning because they speak to the appropriateness of a dynamical framework for investigating social situations. Whereas it is easy to justify the physical existence of linkages between components in the coordinated behavior of a single entity, no such linkage typically exists between people. *Social coordination occurs via information exchange*, typically through vision, touch and sound. Emotional interactions may also be involved. A natural measure that describes this informational exchange is the relative phase between coordinating behaviors. The relative phase is an informational variable whose dynamics and has been shown to capture quantitatively coordinated patterns of brain and behavior among different kinds of components, events and processes (see ► Coordination Dynamics [26,40,41]). For coupled rhythmic behaviors, the relative phase dynamics is often adequate not only for uncovering basic mechanisms underlying synergy formation and behavioral change but also the strength and directionality of influences during social interaction ([82,99]; see also [62,87]) for recent reviews.

Intentional Interpersonal Coordination

Among the many phenomena of human social coordination, one that most of us have experienced is the synchronized clapping of an audience. Néda and colleagues [70,71] have investigated why applause occurs in unison, with individual "clappers" sometimes acting as a single synchronized ensemble. Although synchronized clapping may vary little from one situation to another, the mechanisms governing the phenomenon are nuanced and context-dependent, even within the same audience. An illustration of this context-dependence comes from the world-famous *New Year's Concert* given every year by the Vienna Philarmonic Orchestra in Austria. Traditionally the concert ends with the Radetzky March by Johann Strauss Jr. This piece of classical music is performed in quite an unusual way. For instance, the conductor leads not only the orchestra but also the audience. Upon a visual cue from the maestro, the audience claps in synchrony with the music. The collective clapping is synchronized both with the music and the visual signals given by the conductor. The reader who is not really into classical music might prefer the example of the song *'We will rock you'* by Queen. Except for a final and unique guitar solo, this song is constituted by a powerful rhythm and a poignant vocal performance by lead singer Freddie Mercury. When this very rhythmic song was performed live, the audience intentionally coordinated its movements with the sound of the drums and the pattern of movements visually provided by the singer. People were therefore intentionally clapping their hands on the first two beats and extending their arms on the third. In the coordination dynamics literature, this is referred to as *intentional sensorimotor coordination* of individuals with external events.

Several studies have employed the sensorimotor coordination tasks to investigate interpersonal coordination dynamics for the case when a person intentionally synchronizes her movements with another by means of visual information exchange see [62] for a review. Following the bimanual and sensorimotor paradigms introduced by Kelso and colleagues [36,37,46,48], Schmidt, Carello and Turvey [88] asked two individuals sitting next to each other to swing their legs in an in-phase or antiphase fashion with respect to the leg movements of the other member of the dyad. As movement frequency was increased

(or decreased) by means of an auditory metronome, they found many of the predicted features of nonequilibrium phase transitions [27,47,89]: (i) *differential stability* between the two coordination patterns; (ii) *phase transitions* from the less stable coordination pattern (antiphase) to the more stable one (in-phase); (iii) *critical fluctuations* (i. e. increase in coordination variability) in the vicinity of the transition region and (iv) *hysteresis* (i. e. a sensitivity to the history of the system). All such hallmarks of coordination dynamics (and others such as *critical slowing down* in the vicinity of transitions) have been repeatedly found in a huge number of studies covering various experimental settings. Those settings include, but are not restricted to inter- (e. g. [36,37]) and intra-limb coordination (e. g. [9]) and coordination beween a limb and its uni- (e. g. [3,9,48]) or multi-modal environment [58] to name only a few. The main contribution of Schmidt and colleagues' research was to demonstrate that coordination phenomena found within a person's brain or body, extend to the interactions between people. It is noteworthy that the observed effects extend outside the typical laboratory setting to include coordination phenomena between an individual and an animal as in Lagarde and colleagues' investigation of the coordination dynamics of the horse–rider system [59]. In this unique experiment horses were ridden while walking, trotting and running on a treadmill. The movement dynamics of the horse, the rider and the horse–rider pair were recorded and analyzed revealing that the human–animal dyad exhibits similar coordination dynamics to human interpersonal coordination [59]. In this respect, it cannot be overemphasized that coordination dynamics deals with emergent cooperative effects across very different coordinating elements from neurons to muscles to limbs to people and across the animal–environment divide ([41,42]; for an excellent discussion, see Turvey [100] and commentaries in Vallacher and Nowak [101]). Both the 'intrinsic dynamics' of the individual elements and the nature of the coupling between different elements must be identified for a full account of the phenomena observed.

Several experiments by Schmidt and co-workers, as well as by other groups, have explored the effects on interpersonal coordination of variables such as the manipulation of objects (e. g. hand-held pendulums) or visual surroundings ([74,86]; see [87] for a recent review). Incorporating both aspects de Rugy and colleagues developed a neuro-mechanical model of visually mediated intentional interpersonal coordination [16]. Their model consists of two cross-coupled neuro-mechanical units, each composed of a neural oscillator driving a wrist-pendulum system moved by a different person. Taken individually, each unit reproduces the natural tendency of the participants to freely oscillate close to resonance frequency. When cross-coupled through the vision of movements of the other individual, each person entrains the other as they adopt a common frequency influenced by their own mechanical properties. Although important, neuromechanical properties are not the only factors that determine the stability of coordination patterns between individuals: attentional load and egocentric constraints also influence interpersonal coordination dynamics [96,97].

A series of experiments has investigated whether the motoric and perceptual constraints that shape the dynamics of inter- and intra-limb coordination play a similar role in the coordination between people (e. g. [12,13,53,58,66,78]). In intrapersonal bimanual coordination the preference for co-activation of homologous muscles appears to be mediated by general principles of symmetry in neural organization such as reciprocal connectivity between homologous brain areas. In a study by Oullier and colleagues [76] investigating the relative role of visual/directional and motor (a) symmetries in interpersonal coordination, two participants made index finger flexions while seated facing each other. One acted as a driver (D) by synchronizing to a metronome that systematically increased in rate. The second participant, or follower (F), was required to coordinate finger movements with D via visual coupling only. F participated in four conditions (Fig. 1) determined by a combination of coordination pattern (in-phase or antiphase) and hand posture (supination or pronation). The relative phase requirement was defined by the spatial configuration (i. e. the position of the endpoint of the finger). In this way, co-activation of homologous muscles (finger flexion by F and D) produced both an in-phase and antiphase relationship between the effector endpoints depending on the experimental condition. If purely directional constraints [92] determine the stability of interpersonal coordination, and D functions only as a generic rhythmic stimulus, perceptual antiphase coordination should display decreased stability regardless of the relative hand position of the participants. Contrary to this hypothesis, a strong role was found for interpersonal homologous muscle co-activation. Coordination between individuals was most stable when they were activating similar muscle groups such that co-flexion was always more stable regardless of the resulting spatial pattern. Directional constraints played only a modulatory role. These initial results are at odds with the concept of social coordination as a form of simple perceptual-motor coupling. Rather, it appears that perception of homologous muscular activation acts as a constraint on coordinative stability, creating a "functional homology" to bimanual coordination. Thus, social coordination may be differentiated from

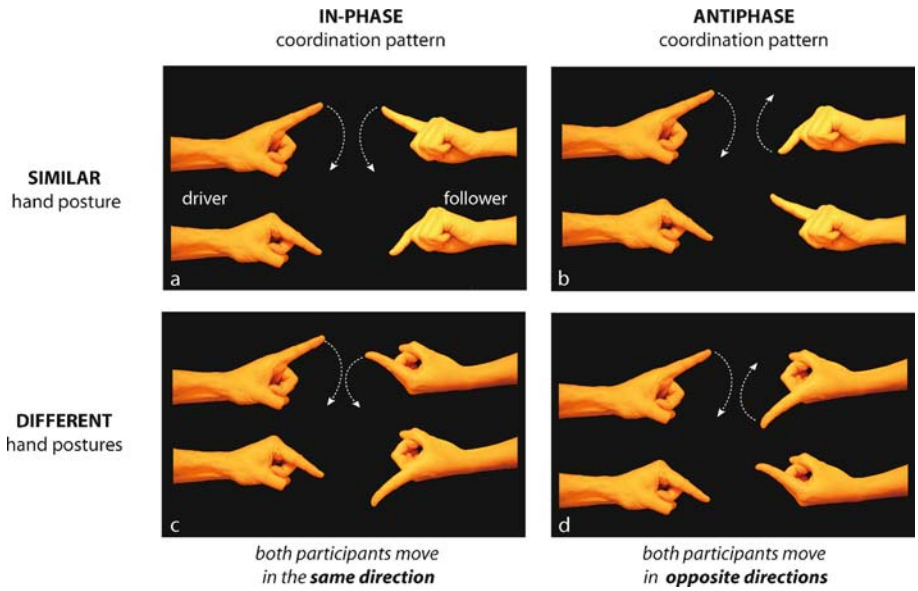

Social Coordination, from the Perspective of Coordination Dynamics, Figure 1

Participants show a preference for homologous muscular activation, irrespective of the visuospatial congruency of their movement. Each column describes a possible configuration of interpersonal coordination. The left participant is paced with a metronome whose frequency increases (driver). Oullier and colleagues [76] studied the frequency at which the right subject (follower) loses stability in each condition. In columns a and b, both subjects are in the same hand position. The pattern a (both flex then both extend: in-phase coordination) is more stable than the pattern b (when one extends the other flexes: anti-phase coordination). In columns c and d, the participants adopt a different hand posture. The pattern d (both flex then both extend: anti-phase coordination) is more stable than the pattern c (where both subjects move in the same direction). These results suggest that coordinative stability is not purely governed by visuospatial congruency (cf. [67]). Rather, the embodiment of the other's movement leads the follower to adopt an anatomically homologous movement. This pattern of behavior is unique to the fact that the follower and the entity with which he/she coordinates are both humans [76]

simple perceptual-motor coupling by virtue of the biological and functional relevance provided when viewing another person.

Although all these studies have employed comparable experimental settings and the common theoretical framework of coordination dynamics with the aim of better understanding intentional interpersonal coordination, it is not yet clear whether spontaneous mutual entrainment actually occurs in a true two-way interaction, or whether one individual simply acts as a pacing stimulus or 'driver' for the other (e. g. [48]). A similar concern can be raised regarding the behavior of the audience during the Radetzky March at the New Year's Concert in Vienna. It seems unlikely that audience members spontaneously synchronized with each other while music was played, since their primary intent was to respond by clapping in rhythm with the music and with the visual cues coming from the stage. This process has been well described in human movement (neuro)science and coordination dynamics and occurs when an individual intentionally coordinates his movements with external physical stimuli [3,41,48]. A sensorimotor interpretation of audience participation is strengthened by results of an experiment in which the auditory metronome used to pace the interpersonal coordination was silenced at times [76]. The study revealed that the presence of an external pacing stimulus (an auditory metronome in that case) actually reduced interpersonal coordinative stability regardless of the adopted directional or muscular pattern adopted. Oullier and colleagues [76] provided evidence for stronger mutual entrainment when no external information could perturb the dyadic interactions, analogous to what the music and the conductor would do during a concert. Hence, from an experimental perspective, this phenomenon is not social interaction per se but rather sensorimotor coordination to an external event. In the case of the "observed" audience clapping in unison *during* the performance, one could argue that the audience is constituted by a collection of individuals coordinating mainly with the music and the conductor with little contribution from neighbor-to-neighbor interactions, (A similar concern can be raised in the study by Schmidt and co-workers [88] as partic-

ipants were instructed to intentionally coordinate with each other *and* with an auditory metronome. Hence, one participant could serve as a visual metronome to the other (and reciprocally) and/or the phase transitions observed could either be interpersonal in nature (from interpersonal antiphase to in-phase) or from syncopation to synchronization as in a single individual coordinating with an auditory metronome (cf. [48]).

A different scenario, however, is characteristic of the end of the performance, when the audience expresses its approval of the orchestra and conductor through applause. At this moment each person applauds according to her preferred/intrinsic pace with no driving stimuli – whether visual or auditory – coming from the stage. In spite of the absence of pacing information, the audience quickly and spontaneously entrains to a common rhythm such that everyone is clapping in unison. Note that at this moment, the only information that can alter an individual's behavior is the sound (and possibly the vision) of the movements made by their neighbors [70,71]. Thus, we have units involved in individual rhythmic behaviors communicating via, at least, one means of information exchange. According to Winfree [108], this is a minimum requirement for self-organized spontaneous synchronization to emerge (see also [41]). In that case, any collective pattern that emerges is more likely to be unintentional compared to situations where the audience follows the conductor and the music.

Issues in Quantifying Spontaneous Interpersonal Coordination

An abundant literature exists addressing unintentional interpersonal coordination in experimental paradigms ranging from people swinging pendulums [86], dancing [33], walking [102] or rocking chairs [84] to performing joint Fitts' tasks [69], talking to each other [83,91] or even boxing [60]. However, many questions remain regarding the nature of the behavioral and neural processes mediating the formation and dissolution of unintended synchronous behavior between individuals and how such processes may be quantified [2,55].

Oullier and colleagues [82] have identified three major problems in investigating spontaneous synchronization in social settings. First, even when the source and nature of the coupling has been identified, it is difficult to manipulate experimentally relevant variables such as the coupling strength (e. g. [71]). Almost by definition, spontaneous behavior is not externally goal directed or explicitly controlled. Most of the results reporting unintentional synchronization in humans are based on observation and categorization methods that rely primarily on the experimenter's appreciation of a given exemplar behavior rather than a quantitative measure of coupling and individual behavior (e. g. [4,14]).

A second problem is the challenge of complexity, both in terms of the large number of units to analyze (e. g. thousands of pairs of clapping hands [71]) and the complexity of the behavior itself (e. g. mother-infant synchronization [14]). Such compositional and behavioral complexity has hindered experimental attempts to record and quantify both the individual and social dynamics. Even the reduction in dimensional complexity afforded in coordinated behavior can only go so far in elucidating the relationship between group behavior and the individual units of which it is composed.

A third problem comes from the possibility that any change in a person's behavior induced by interacting with another may persist even after the encounter is over. So far, there has been very little precise quantification of the mutual influence people have on each other's behavior a posteriori, i. e. how individual behavior is affected after the social encounter when people no longer exchange information (but see Sect. "Social Memory and the Dependence on Initial Conditions").

Human Spontaneous Synchronization

In behavioral experiments that revealed spontaneous interpersonal synchronization, Oullier and colleagues [77,80,82] explored coordinative patterns that emerge only as a function of visual information exchange. The main hypothesis was that even without instructions to do so, spontaneous synchronization between partners would occur as soon as they coupled visually while moving in front of each other. On the other hand, spontaneous interpersonal coordination should disappear whenever exchange of information is no longer possible. In Oullier et al.'s behavioral experiments, pairs of participants executed movements while in full view (or not) of each other's ongoing actions as well as their own [77,80,82]. Each member of the dyad executed movements at their own preferred frequency and amplitude without any external pacing from a metronome or any other sort. Movements were required to be as smooth and continuous as possible throughout an experimental trial. What is important here is that participants were not given any instructions regarding the way to move with respect to each other. The experimental protocol consisted of participants moving with no vision of the other's movements before being allowed to see their own actions at the same time as they saw the other person's. Finally, visual information was removed again. Experi-

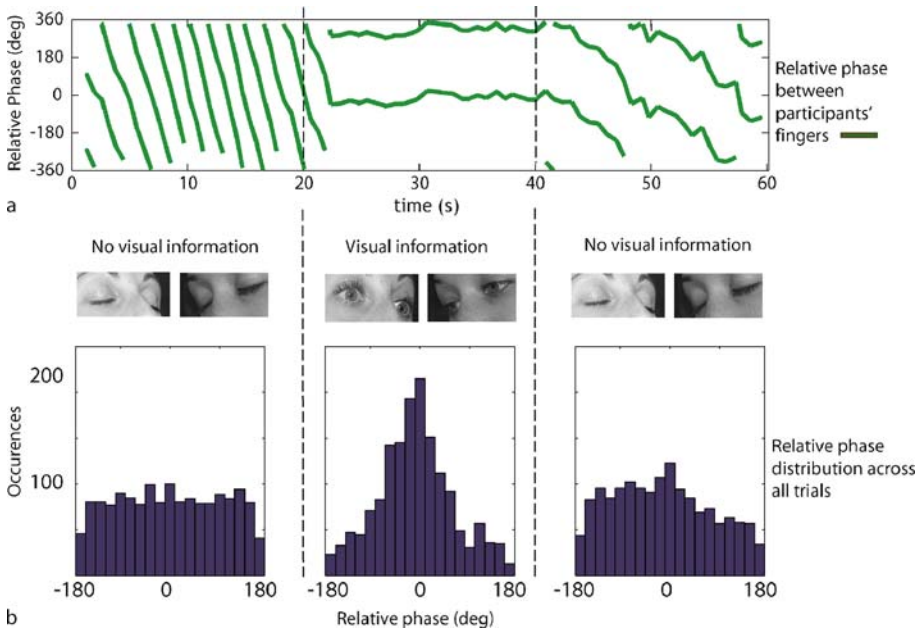

Social Coordination, from the Perspective of Coordination Dynamics, Figure 2

Relative phase between the movements of two individuals. a This panel illustrates the evolution of the relative phase as a function of time in a representative trial of the SCD paradigm. *Left column* No visual information exchange, each individual moves independently, movements are uncoordinated. *Middle column* as soon as people exchange visual information, they spontaneously couple. Their relative phase is therefore close to 0° . *Right column* When visual information is removed, they are no longer synchronized. b This panel represents distributions of relative phase for all the subjects and all the trials (adapted from [82])

mental trials were therefore equally partitioned into three contiguous segments each of equal duration within which both subjects either were allowed to exchange information with each other or not. When visual information was available, participants looked at each other's finger motion and were also able to see their own finger [77,80,82].

In SCD, following theories of cooperative phenomena in open systems [25,26] a central idea is that the behavior of a complex dyadic system may be captured by the value of a low-dimensional collective variable known as the *order parameter*. In the vicinity of critical points, emergent behavior is governed by the dynamics of this collective variable e. g. [25,41]. In experimental cases the order parameters are not known in advance but have to be discovered. For the situation of social coordination as in many other cases treated by CD, an appropriate order parameter describing the system dynamics is the relative phase ϕ between the movements of each member of the pair [80,82]. The relative phase measure allows for a reduction of a potentially very high dimensional system (e. g. where one has to consider, among other components, the neurons, joints and muscles of both individuals) as it captures the macroscopic spatio-temporal behavioral pattern (see Fig. 2). Even at an overt behavioral level, four degrees of freedom (position and velocity of each component) may be compressed onto a single relative phase value that summarizes the organization of the dyadic system. Quantitative evaluation of spontaneous synchrony is also provided by the FFT power spectrum overlap between the movements of each person. The spectrum overlap measures the percentage of movement frequencies common to both partners in a pair [82]. Defined as the area of intersection between each participant's normalized spectral plots, it serves an indicator of the strength of the frequency entrainment between the two participants (see Fig. 4).

When no visual exchange was allowed, each subject produced movements independently at their own frequency. As a result, the relative phase ϕ between the subjects' finger motions exhibited phase wrapping (Fig. 2, left column). However, following a simple auditory cue to open their eyes, subjects spontaneously adopted in-phase motion, ϕ stabilizing around 0° (Fig. 2, middle column). On a signal to close the eyes again, the individual movement frequencies diverged and ϕ fell back into phase wrapping (Fig. 2, right column). These initial results were corroborated by a subsequent more extensive study, in which the order of the vision and no-vision segments was changed. Once again, spontaneous synchro-

nization emerged as soon as vision of the other's movements was allowed [82]. Overall, results reveal that with visual information exchange, participants tend to mutually couple at a common phase and frequency, whereas in the absence of vision, participants' movement trajectories diverge and behave independently. Such emergent mutual coupling is truly a result of spontaneous social interaction and may be distinguished from previous dyadic studies in which one person may simply be intentionally tracking (or driving) the other [16,77,88,97] or maintaining their own rhythm [86].

Why does spontaneous interpersonal coordination occur at all? Compelling examples stretching from human evolution through religious ritual and sports to political, war and economical strategy suggest that *keeping together in time* is one of the most powerful ways to create and sustain communities and communication [65]. Moreover, *not* moving in synchrony may be too costly for the dyad see, (e. g., [56]).

In order to better understand which features of visual information exchange may facilitate spontaneous social coordination one has to bear in mind that human movements can be unintentionally affected by the vision of an object oscillating in the environment. This is illustrated by experiments using the moving-room paradigm in which the walls of the room move but not the floor (e. g. [61,75,78]). Body sway of the observer's couples in time spontaneously with small oscillatory motions of the room. In addition, experimental data show that the mere observation of the movements of another person interferes with one's execution of a similar action [54]. Interestingly, such interference is less noticeable when the movements observed are not generated by humans [15]. In the latter work, one of the members of the dyad was replaced by a computer-generated moving hand, the trajectory of which was driven either by a sinusoidal function or a pre-recorded real finger trajectory. The stimulus movement frequency in the study by de Guzman and coworkers [15] was fixed at either 10% below or 10% above the subject's self-paced rate as determined at the start of the experiment. Results revealed that the human–avatar coordination was strongest when the latter was an image of a hand driven by real movement data. The weakest coupling occurred when the visual stimulus followed a sinusoidal trajectory. Unlike the interpersonal situation [82], spontaneous synchronization was not found for all trials and, when it happened, was supported by a significantly lower frequency overlap [15]. One may invoke a one-way coupling to explain these findings, since the motion of the computer generated hand could not be influenced by the movement of the participant. Taken together, the foregoing results support the hypothesis that biological relevance in general, and biological motion in particular – including its natural variation– play a key role in social coordination.

Shared Behavioral and Neural Social Coordination Dynamics

One explanation for the emergence of spontaneous social coordination may be found at the neurophysiological level. For instance, some areas of the brain are known to be associated with the perception (but not the execution) of biological motion including the posterior superior temporal sulcus (abbreviated STS) [1,23,24,31]. STS is known to be a major source of visual information for the so-called *human mirror system* (abbreviated HMS) [85]. Originally identified in monkeys, *mirror neurons* are (sensori)motor neurons discharging both when one performs a given action and sees the same action performed by someone else. They have been identified primarily in the ventral premotor cortex and the rostral region of the inferior parietal lobule [20]. The HMS constitutes a neural mechanism that is automatically activated by the sight of somebody else's actions, even when the observer does not make overt movements. The main idea is that during observation the HMS provides a simulation of the actions of other people potentially providing a basis for understanding the intentions of others [31].

Since the foregoing behavioral experiments allow participants to both produce and observe movement at the same time it seems possible that the HMS is at least partially involved in the spontaneous coordination observed. In order to investigate this question, Tognoli and colleagues [99] recorded brain activity of each member of the dyad using a specially designed dual-electroencephalography (EEG) system. Each participant wore a 60-electrode EEG-cap that enabled simultaneous recording of their brains to accompany kinematic measurements of their behavior.

To grasp the significance of the work by Tognoli and colleagues, we need to revert to earlier studies conducted within the framework of Coordination Dynamics have employed instabilities in coordination as a means to uncover the link between the dynamics of behavior and the dynamics of the brain [39,42], with the goal of relating levels by virtue of their *shared dynamical properties* (e. g. [19,39,49,50,52]). In such research, the high temporal resolution of electroencephalography (EEG) and magnetoencephalography (MEG) was exploited to quantify the relationship between the large scale neural dynamics emerging from billions of interconnected neurons and the behavioral dynamics revealed in experiments on coordi-

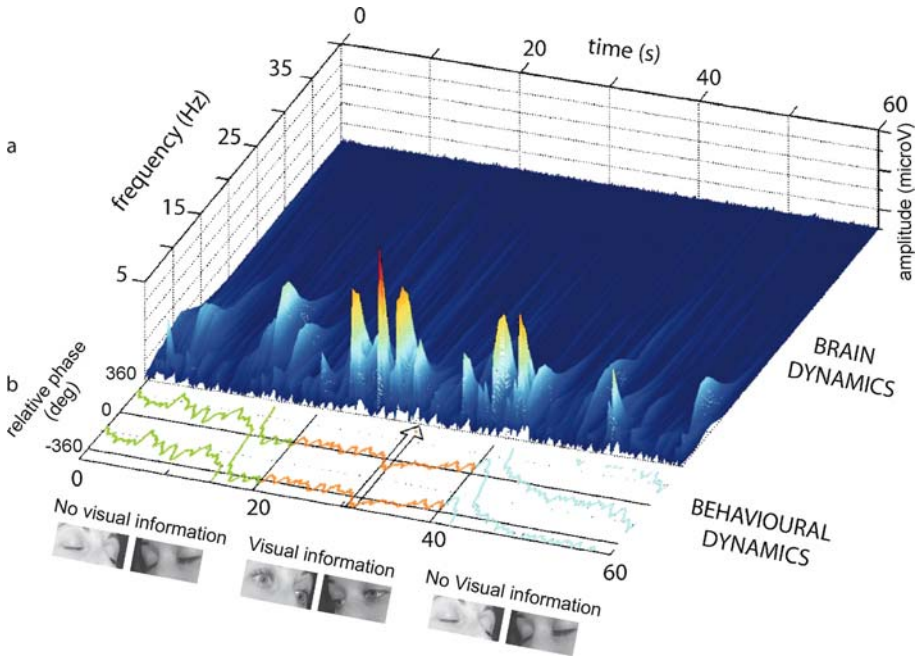

Social Coordination, from the Perspective of Coordination Dynamics, Figure 3

Relation between Phi_2 and social coordination. a Time-frequency spectrum from electrode CP4 (located over parietal brain regions) from a single trial. Phi_2 is low before and after vision but increases during vision. b Corresponding relative phase between finger movements. Synchronized in-phase behavior is observed during visual contact. Notice the temporary disengagement of the rhythm when coordination is lost briefly (adapted from [99])

nation [41]. Observed features of the dynamics expressed at both levels of description such as multistability and phase transitions (i. e. the spontaneous switch from one pattern to another due to loss of stability), were taken as evidence that principles of self-organization govern pattern formation in both brain and behavior [26,41]. Of particular initial interest was the identification of qualitative changes in the pattern of neural activity that occurred simultaneously with transitions between behavioral patterns [19,49,50,63,104]. On the basis of this work, an exciting hypothesis is that the transitions from uncoordinated to spontaneous coordination observed in the SCD paradigm may be accompanied by similar events at the brain level.

In an effort to shed new light on how social processes are integrated in the brain, Tognoli and colleagues [99] identified several neural mechanisms or *neuromarkers* that appear and disappear with the emergence and dissolution of coordinated behavior between two people. Interestingly, these social neuromarkers consist of brain rhythms in the 10 Hz frequency range located over right centro-parietal areas of the cerebral cortex. In particular, a social brain rhythm termed the *Phi Complex* consists of two components: the first, Phi_1, increases during independent behavior i. e. before information exchange between members of the dyad. When subjects saw each other's finger movements and coordinated together, Phi_1 disappeared and Phi_2, a different rhythm within the same frequency band appeared (Fig. 3) [99].

In a subsequent study, Tognoli and colleagues [98] explored the dynamics of the *Phi Complex* by instructing participants to *intentionally* synchronize when visual information exchange was allowed. In this case, participants interact to accomplish a shared goal. Again, Phi_1 appeared during uncoordinated behavior and Phi_2 when social coordination occurred. Analysis of dyads who participated in both experiments [98,99] revealed that the amplitude of Phi_2 was higher during intentional than spontaneous coordination. Thus, Phi_2 appears to be a neural signature of social coordination whether it emerges spontaneously or not.

The cortical location of the *Phi Complex* appears to be consistent with neuro-anatomical sources within the human mirror system. One of the conclusions drawn by Tognoli and colleagues is that Phi_1 might have an inhibiting role on the mirror system. Previous claims by Brass and Heyes [8] have argued that the mirror system is always active by default and thus must be inhibited in non-social contexts. Hence Phi_1 could inhibit Phi_2, the latter being seen as a facilitator of social coordination that participates

in information exchange between the motor cortex and the mirror system [98,99].

In summary, experiments using dual high density electrode arrays to record and measure brain activity from two persons in conjunction with motion capture technology, have allowed an exploration of shared behavioral and neural social coordination dynamics [98,99]. Transitions from uncoordinated dyadic behavior to interpersonal synchronization have been demonstrated to accompany the emergence of a new brain rhythm – the Phi complex – located in the human mirror system. Such work suggests that SCD may serve as a novel framework for identifying behavioral and neural signatures in reciprocal interactions and allows for a more dynamical approach to the study of the mirror neuron system.

Social Memory and the Dependence on Initial Conditions

At first blush, the emergence of spontaneous coordination between individuals [77,80,82,99] might be seen as an instantiation of mutual entrainment that entails nothing more than a couple of oscillators and a medium of information exchange [41,108]. In such generic cases, once the coupling is removed, each oscillator should return to its own intrinsic frequency, that is, any influence of the interaction should disappear. However, the situation between two people is different (see Fig. 4). Theoretically, in a typical coupled clocks scenario, there should be no difference between the movement periods of the 'clocks' before or after coupling-induced synchronization. However, a serendipitous experimental finding [82] was the consistent and persistent influence of the social interaction on subsequent rhythmic behavior despite the absence of information exchange between the pair (Fig. 4). This remnant of a prior social interaction may qualify as a kind of *social memory* [82].

Social memory is thought to play an important role in human actions, and, to a larger extent, on the way we live [32]. In the context of SCD, social memory implies that the intrinsic parameters of the individual components have been altered by virtue of the social interaction. Mathematically [42], one may represent the situation before the interaction as follows:

$$\begin{aligned} \ddot{x}_1 + i_1(x_1, \dot{x}_1, a_1) + \omega_1^2 x_1 &= 0 \\ \ddot{x}_2 + i_2(x_2, \dot{x}_2, a_2) + \omega_2^2 x_2 &= 0\,. \end{aligned} \tag{1}$$

Where x_1 and x_2 represent the coordinating components,

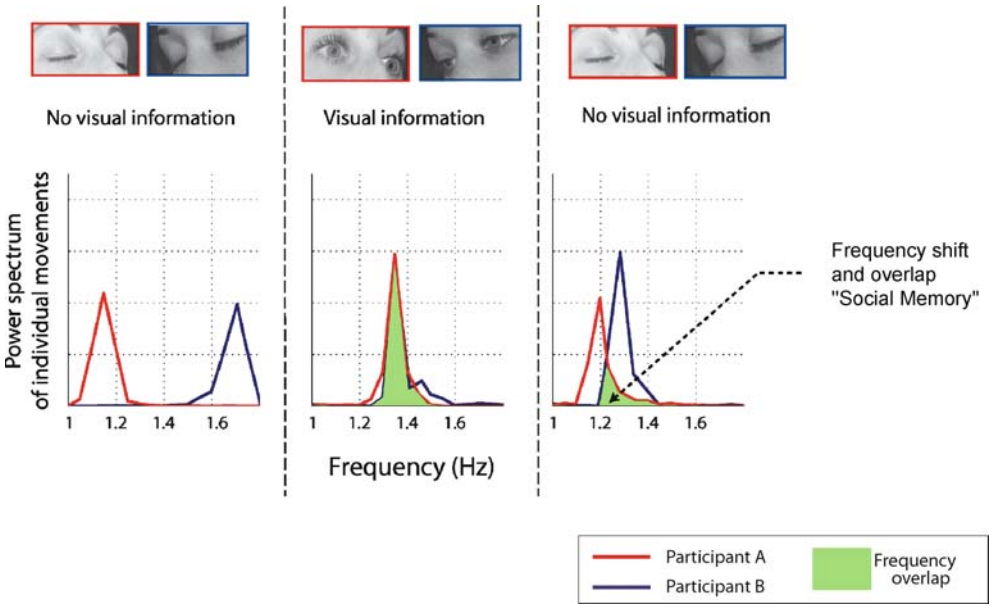

Social Coordination, from the Perspective of Coordination Dynamics, Figure 4

Evidence for social memory. Illustrated is an example of frequency overlap between the movements of both subjects in a representative trial of the SCD paradigm. *Left column* no vision of each other's movements; each individual moves at their own intrinsic frequency so there is no frequency overlap. *Middle column* visual information is exchanged between participants, causing spontaneous synchronization to occur at a common frequency (and phase, see Fig. 2). *Right column* visual information is no longer exchanged but individuals do not revert back to their initial intrinsic frequency. This remnant of frequency overlap as a result of prior social interaction suggests a kind of 'social memory' (adapted from [82])

$i_{1,2}$, $a_{1,2}$ and $\omega_{1,2}$ refer to individually chosen intrinsic parameters such as the chosen frequency and amplitude.

During the interaction, the system is visually coupled, F_1 and F_2 representing a coupling function such as the well-known HKB-coupling [27,44]:

$$\begin{aligned} \ddot{x}_1 + I_1(x_1, \dot{x}_1, A_1) + \omega_1^2 x_1 &= F_1(x_1, \dot{x}_1, x_2, \dot{x}_2) \\ \ddot{x}_2 + I_2(x_2, \dot{x}_2, A_2) + \omega_2^2 x_2 &= F_2(x_2, \dot{x}_2, x_1, \dot{x}_1)\,. \end{aligned} \tag{2}$$

Now notice that the interactive context has formed a coupling (the right hand side of Eq. (2)) but also led of a modification of the individual component parameters, I and A (on the left hand side of Eq. (2)). One may say that the boundary conditions of Eq. (1) have been altered by the social interaction.

After the interaction, the coupling function disappears (F_1 and F_2 terms on the right hand side are zero) and the system is "uncoupled" (cf. Fig. 1):

$$\begin{aligned} \ddot{x}_1 + I_1(x_1, \dot{x}_1, A_1) + \omega_1^2 x_1 &= 0 \\ \ddot{x}_2 + I_2(x_2, \dot{x}_2, A_2) + \omega_2^2 x_2 &= 0\,. \end{aligned} \tag{3}$$

However, notice in Eq. (3) the individual intrinsic parameters of the system which were modified by the interaction are still in place. Though uncoupled, the individual components are still affected by the interaction. How this internalization process occurs remains open to empirical investigation.

A benefit of the SCD paradigm is that one is able to quantify the strength and persistence of prior social influences on an individual's behavior. The finding that the modification of the neural network depends on which modality is engaged during the mutual encounter suggests that additional cortical areas may have been recruited and included into the initial global neural assembly due to social context [32]. However, beyond the Phi complex, and perhaps due to inherent limitations in spatial resolution, examination of the dual EEG data showed no evidence of further cortical engagement. Another possibility (in line with the foregoing mathematical analysis) is that the connectivity and dynamics of the initial network is modified by social interaction, and the new organization retained after the interaction is over. Recent evidence in support of this hypothesis suggests that two people engaging in a common task share a representation of each other's movement dynamics, including trajectory amplitude and frequency [6,17]. Such a (shared) representation may persist when vision is removed, i. e. when information exchange is no longer possible [21]. Moreover, representations at the neural level have been shown to be highly flexible and context-dependent [34,35], influenced both by environmental [105] and task demands [79].

The extent and duration of the carryover or remnant effects observed in behavioral experiments may reflect many factors, including the strength of the bond that is formed between people, place in the social hierarchy, the willingness of each participant to cooperate, gender differences, personality characteristics and the significance each participant attaches to the social encounter [32]. An additional finding from our work favors a motor contribution to social memory as well: the persistence effect was found to be independent of the duration of movement that followed the social encounter [82]. This hypothesis is strengthened by results showing that observation of another person performing movements generates a kinematically specific memory of the observed motions in primary motor cortex [95].

The systematic directionality effect observed in the SCD paradigm is revealing also [82]: the extent to which one member of the dyad is influenced by the other was shown to depend on initial conditions. Obviously for synchronization to occur, the person moving with the lowest/highest intrinsic movement frequency must speed up/slow down during information exchange. A surprising result is that the difference between the initial and the final intrinsic movement frequencies (vision absent) was always greater for the person starting with the higher compared to the lower movement frequency [82]. The extent to which initial, so-called 'intrinsic dynamics' determine behavior after the social encounter is over may be of great interest to understanding social interactions in more complex settings where hierarchical relations are involved.

An important problem in human social behavior concerns understanding the degree to which an individual influences the actions of a group (e. g. peer group, family, class) he/she is in. Due to several factors (personality, situational), a person (the leader) may affect the behavior of others more than the others affect her or him. The concept of leadership is commonly associated with interactions taking place in hierarchical settings such as typical organizations, but is actually broader than that. Strength of behavioral influence is overlooked because behavioral interactions have not been systematically studied. Contemporary complex systems approaches (e. g. [28]) view the formation of leader-follower roles as interactive and emergent but in so doing may have undermined the significance of individual dimensions. The approach of SCD is rather to ask: *What makes two people behave independently and what makes them behave as a unit?* The paradigm of social coordination dynamics exploits inherent asymmetry between two people during behavioral interactions and gauges, e. g. using directional coupling measures, which of the two has a stronger influence.

Future Directions

Human beings are social by nature, and interactions with others represent a substantial portion of their many daily activities. A common and well described consequence of interpersonal activity is that an individual's behavior, whether intentional or not, is modified by interactions with others [32]. Alterations of individual and collective behaviors range from imitation and mimicry to spontaneous synchronization, and have been observed in groups varying in size from dyads to thousands of individuals e. g. [4,68].

Social Coordination Dynamics investigates how the natural (uninstructed) social influence of one person on another evolves in real time and has led to a number of new findings. The first is that humans immediately and spontaneously coordinate their actions with each other when provided vision of the movements of the other together with their own. The second is that a specific brain rhythm underlies social coordination. Transitions from individual to coordinated social behavior are observed at both behavioral and brain levels.

The third finding is that an individual's intrinsic behavior is altered by social interaction: the effect of the previous social encounter persists when vision of the other's movements is no longer available. A fourth and final finding is that social coordination is affected by initial conditions, enabling one to predict which individual is most affected by the social encounter.

Insights into elementary forms of social interaction have been obtained by applying the concepts, methods and tools of coordination dynamics. A notable feature of coordination dynamics is its ability to uncover mechanisms and principles common to different kinds of complex systems at different levels of observation and to relate them by virtue of shared behavioral and neural dynamics [41]. SCD and its dynamical measures have proven to provide adequate quantification of the spontaneous coupling between individuals, the transition to loss of entrainment and the effect of the social encounter at both behavioral and brain levels. The same basic patterns of coordinated behavior and pattern dynamics (multistability, critical fluctuations accompanied by a temporary loss of stability, phase transitions, hysteresis and critical slowing down) have been observed within an individual, between an individual and the environment, and between individuals. In this respect SCD complements recent developments in social cognitive neuroscience, behavioral economics, game theory, socio-economics and neuroeconomics (e. g. [10,11,18,73,81,94,103]).

The field of *social neuroeconomics* serves to illustrate the benefits of considering SCD in contexts other than interpersonal sensorimotor interaction. Social neuroeconomics investigates the neural correlates of economical decision making [18]. One particular feature of this nascent field is that decision making processes are always studied in a body- and movement-independent fashion. Why is that? After all, from the very first months of life, individuals live vicariously through one another adopting, if only temporarily, a similar posture or tempo during interactions with a peer, or yawning [4,64,90]. As Henry Greely [22] recently reminded the readership of Science Magazine *"Human society is the society of human brains. Of course those brains are encased in, affected by, and dependent on the rest of the body, but our most important interactions are with other people's brains, as manifested through their bodies."* Although this statement sounds like common sense, thus far the coordination dynamics between bodies has remained unexplored in the field of social neuroeconomics. Yet how many times have we experienced the feeling that trusting someone will be difficult even before talking to them? Whether it was the way she moved or some other factor, body-related cues play a key role in modulating economic decisions (e. g. [93]). A scientific approach to "body language" might aim to understand how perceived actions of others affect the cognitive and emotional processes involved in economical decision-making. For instance, a finding such as the Phi Complex – especially the modulation of Phi_2 when individuals intentionally coordinate [98,99] – could turn to be crucial to better competition–cooperation mechanisms underlying decision in economic contexts such as public coordination games [29]. In sum, as a conceptual framework that encompasses the dynamics of both neural and behavioral levels, SCD promises to bridge the gaps between levels of analysis [41,81] and clear a path for new multi-level, interdisciplinary investigations of social interactions. Like synchronization itself, the function of SCD is to facilitate communication across heretofore unrelated fields.

Acknowledgments

The authors wish to thank their collaborators at the *Human Brain and Behavior Laboratory* (Center for Complex Systems and Brain Sciences) who co-authored the articles on social coordination dynamics reported in this article: Drs. Gonzalo C. de Guzman (*Florida Atlantic University*), Kelly J. Jantzen (*Washington Western University*), Julien Lagarde (*Université de Montpellier I*), Cyrille Magne (*Middle Tenesse State University*) and Emmanuelle Tognoli (*Florida Atlantic University*). Amélie L. and Au-

drey O. should also be thanked for letting the authors use their beautiful hands and eyes in all figures. We wish to thank the following institutions for supporting our research and the writing of this article. The *US National Institutes of Health* (NIMH Innovation Grant MH 42900 and MH080838), the Pierre de Fermat Chair of the Région Midi-Pyrénées (France) to J.A.S.K, and the *Centre National de la Recherche Scientifique* (Programme CNRS Neuroinformatique to O.O.). Correspondence should be addressed to: Olivier Oullier (olivier@oullier.fr) or J. A. Scott Kelso (kelso@ccs.fau.edu). For additional information please visit the following websites: www.oullier.fr and www.ccs.fau.edu/section_links/HBBLv2/index.html

Bibliography

Primary Literature

1. Allison T, Puce A, McCarthy G (2000) Social perception from visual cues: Role of the STS region. Trends Cogn Sci 4:267–278
2. Balaban E (2004) Neurobiology - Why voles stick together. Nature 429:711–712
3. Bardy BG, Oullier O, Bootsma RJ, Stoffregen TA (2002) Dynamics of human postural transitions. J Exp Psychol Hum Percept Perform 28:499–514
4. Barsalou LW, Niedenthal PM, Barbey AK, Ruppert JA (2003) Social embodiment. Psychol Learn Motiv 43:43–92
5. Bennett M, Schatz MF, Rockwood H, Wiesenfeld K (2002) Huygens's clocks. Proc Roy Soc Lond A Math Physic Engineer Sci 458:563–579
6. Bosbach S, Cole J, Prinz W, Knoblich G (2005) Inferring another's expectation from action: The role of peripheral sensation. Nat Neurosci 8:1295–1297
7. Bottani S (1996) Synchronization of integrate and fire oscillators with global coupling. Phys Rev e 54:2334–2350
8. Brass M, Heyes C (2005) Imitation: is cognitive neuroscience solving the correspondence problem? Trends Cogn Sci 9:489–495
9. Buchanan JJ, Kelso JA (1993) Posturally induced transitions in rhythmic multijoint limb movements. Exp Brain Res 94:131–142
10. Camerer CF (2003) Behavioral game theory: Experiments in strategic interaction. Princeton University Press, Princeton
11. Camerer CF, Loewenstein G, Prelec D (2005) Neuroeconomics: How neuroscience can inform economics. J Econ Lit XLIII:9–64
12. Carson RG (2004) Governing coordination. Why do muscles matter? In: Jirsa VK, Kelso JAS (eds) Coordination dynamics: Issues and trends. Springer, Berlin, pp 141–154
13. Carson RG, Kelso JAS (2004) Governing coordination: behavioural principles and neural correlates. Exp Brain Res 154:267–274
14. Condon WS, Sander LW (1974) Neonate movement is synchronized with adult speech - Interactional participation and language acquisition. Science 183:99–101
15. de Guzman GC, Tognoli E, Lagarde J, Jantzen KJ, Kelso JAS (2005) Effects of biological relevance of the stimulus in mediating spontaneous visual social coordination. Society Neurosci Program 867(21)
16. de Rugy A, Salesse R, Oullier O, Temprado JJ (2006) A neuromechanical model for interpersonal coordination. Biol Cybern 94:427–443
17. Decety J, Sommerville JA (2003) Shared representations between self and other: A social cognitive neuroscience view. Trends Cogn Sci 7:527–533
18. Fehr E, Camerer CF (2007) Social neuroeconomics: the neural circuitry of social preferences. Trends Cogn Sci 11:419–427
19. Fuchs A, Kelso JAS, Haken H (1992) Phase transitions in the human brain: Spatial mode dynamics. Int J Bifurc Chaos 2:917–939
20. Gallese V, Fadiga L, Fogassi L, Rizzolatti G (1996) Action recognition in the premotor cortex. Brain 119:593–609
21. Goldman MS, Levine JH, Major G, Tank DW, Seung HS (2003) Robust persistent neural activity in a model integrator with multiple hysteretic dendrites per neuron. Cereb Cortex 13:1185–1195
22. Greely H (2007) On neuroethics. Science 318:533
23. Grèzes J, Fonlupt P, Bertenthal B, Delon-Martin C, Segebarth C, Decety J (2001) Does perception of biological motion rely on specific brain regions? Neuroimage 13:775–785
24. Grèzes J, Armony JL, Rowe J, Passingham RE (2003) Activations related to mirror and canonical neurons in the human brain: An fMRI study. Neuroimage 18:928–937
25. Haken H (1983) Synergetics: An introduction. Springer, Berlin
26. Haken H (1996) Principles of brain functioning: A synergetic approach to brain activity, behavior and cognition. Springer, Berlin
27. Haken H, Kelso JAS, Bunz H (1985) A theoretical-model of phase-transitions in human hand movements. Biol Cybern 51:347–356
28. Hazy JK, Goldstein JA, Lichtenstein BB (2007) Complex systems leadership theory: New perspectives from complexity science on social and organizational effectiveness. ISCE Publishing Company, Mansfield
29. Hichri W, Kirman AP (2007) The emergence of coordination in public good games. Europ J Phys B 55:149–159
30. Hugenii C (1673) Horologium oscillatorium. Apud F. Muguet, Paris
31. Iacoboni M, Molnar-Szakacs I, Gallese V, Buccino G, Mazziotta JC, Rizzolatti G (2005) Grasping the intentions of others with one's own mirror neuron system. PLoS Biol 3:529–535
32. Insel TR, Fernald RD (2004) How the brain processes social information: Searching for the social brain. Annu Rev Neurosci 27:697–722
33. Issartel J, Marin L, Cadopi M (2007) Unintended interpersonal co-ordination: can we march to the beat of our own drum? Neurosci Lett 411:174–179
34. Jantzen KJ, Steinberg FL, Kelso JAS (2004) Brain networks underlying human timing behavior are influenced by prior context. Proc Nat Acad Sci USA 101:6815–6820
35. Jantzen KJ, Steinberg FL, Kelso JAS (2005) Functional MRI reveals the existence of modality and coordination-dependent timing networks. Neuroimage 25:1031–1042
36. Kelso JAS (1981) Contrasting perspectives on order and regulation in movement. In: Long J, Baddeley A (eds) Attention and performance IX. Erlbaum, Hillsdale, pp 437–457
37. Kelso JAS (1984) Phase-transitions and critical behavior in human bimanual coordination. Am J Physiol 246:1000–1004

38. Kelso JAS (1991) Behavioral and neural pattern generation: The concept of NBDS. In: Koepchen HP, Huopaniemi T (eds) Cardiorespiratory and motor coordination. Springer, Munich, pp 224–238
39. Kelso JAS (1992) Coordination dynamics of human brain and behavior. Springer Proc Phys 69:223–234
40. Kelso JAS (1994) The informational character of self-organized coordination dynamics. Hum Mov Sci 13:393–413
41. Kelso JAS (1995) Dynamic Patterns: The self-organization of brain and behavior. MIT Press, Cambridge
42. Kelso JAS (2000) Principles of dynamic pattern formation and change for a science of human behavior. In: Bergman R, Cairns RB, Nilsson LG, Nystedt L (eds) Developmental science and the holistic approach . Lawrence Erlbaum Associates, Mahaw, pp 63–83
43. Kelso JAS (2005) Context, components and complexity. Invited paper, Plexus Institute Summit, Delray Beach, September 11
44. Kelso JAS (2007) The Haken–Kelso–Bunz Model. Scholarpedia (Computational Neuroscience/Dynamical Systems). http://www.scholarpedia.org/article/Haken-Kelso-Bunz_model
45. Kelso JAS, Engstrøm DA (2006) The complementary nature. MIT Press, Cambridge
46. Kelso JAS, Scholz JP, Schöner G (1986) Dynamics govern switching among patterns of coordination. Phys Lett A 134:8–12
47. Kelso JAS, Schöner G, Scholz JP, Haken H (1987) Phaselocked modes, phase transitions and component oscillators in coordinated biological motion. Physica Scripta 35:79–87
48. Kelso JAS, DelColle J, Schöner G (1990) Action-perception as a pattern formation process. In: Jeannerod M (ed) Attention and Performance XIII. Erlbaum, Hillsdale, pp 139–169
49. Kelso JAS, Bressler SL, de Guzman GC, Ding M, Fuchs A, Holroyd T (1991) Cooperative and critical phenomena in the human brain revealed by multiple SQUIDS. In: Duke D, Pritchards W (eds) Measuring chaos in the human brain. World Scientific, New Jersey, pp 97–112
50. Kelso JAS, Bressler SL, de Guzman GC, Ding M, Fuchs A, Holroyd T (1992) A phase-transition in human brain and behavior. Phys Lett A 169:134–144
51. Kelso JAS, Fuchs A, Lancaster R, Holroyd T, Cheyne D, Weinberg H (1998) Dynamic cortical activity in the human brain reveals motor equivalence. Nature 392:814–818
52. Kelso JAS, Fuchs A, Jirsa VK (1999) Traversing scales of brain and behavioral organization. I.-III. In: Uhl C (ed) Analysis of neurophysiological brain functioning. Springer, Heidelberg, pp 73–125
53. Kelso JAS, Fink PW, DeLaplain CR, Carson RG (2001) Haptic information stabilizes and destabilizes coordination dynamics. Proc R Soc Lond B Biol Sci 268:1207–1213
54. Kilner JM, Paulignan Y, Blakemore SJ (2003) An interference effect of observed biological movement on action. Curr Biol 13:522–525
55. Konner M (2004) The ties that bind - Attachment: the nature of the bonds between humans are becoming accessible to scientific investigation. Nature 429:705
56. Körding KP, Fukunaga I, Howard IS, Ingram JN, Wolpert DM (2004) A neuroeconomics approach to inferring utility functions in sensorimotor control. PLoS Biol 2:1652–1656
57. Kuramoto Y (1984) Chemical oscillations, waves, and turbulences. Springer, Berlin
58. Lagarde J, Kelso JAS (2006) Binding of movement, sound and touch: multimodal coordination dynamics. Exp Brain Res 173:673–688
59. Lagarde J, Kelso JAS, Peham C, Licka T (2005) Coordination dynamics of the horse-rider system. J Mot Behav 37:418–424
60. Lagarde J, de Guzman GC, Oullier O, Kelso JAS (2006) Interpersonal interactions during boxing: Data and model. J Sport Exerc Psychol 28: S108
61. Lee DN, Lishman JR (1975) Visual proprioceptive control of stance. J Hum Mov Stud 1:87–95
62. Marsh KL, Richardson MJ, Baron RM, Schmidt RC (2006) Contrasting approaches to perceiving and acting with others. Ecol Psychol 18:1–37
63. Mayville JM, Bressler SL, Fuchs A, Kelso JAS (1999) Spatiotemporal reorganization of electrical activity in the human brain associated with a timing transition in rhythmic auditory-motor coordination. Exp Brain Res 127:371–381
64. McGarva AR, Warner RM (2003) Attraction and social coordination: Mutual entrainment of vocal activity rhythms. J Psycholing Res 32:335–354
65. McNeill WH (1995) Keeping together in time. Harvard University Press, Cambridge
66. Mechsner F (2004) A perceptual-cognitive approach to bimanual coordination. In: Jirsa VK, Kelso JAS (eds) Coordination dynamics: Issues and Trends. Springer, Berlin, pp 177–195
67. Mechsner F, Kerzel D, Knoblich G, Prinz W (2001) Perceptual basis of bimanual coordination. Nature 414:69–73
68. Motter AE, Nishikawa T, Lai YC (2003) Large-scale structural organization of social networks. Phys Rev E 68:e036105-e036110
69. Mottet D, Guiard Y, Ferrand T, Bootsma RJ (2001) Two-handed performance of a rhythmical fitts task by individuals and dyads. J Exp Psychol Hum Percept Perform 27:1275–1286
70. Néda Z, Ravasz E, Vicsek T, Brechet Y, Barabasi AL (2000) Physics of the rhythmic applause. Phys Rev e 61:6987–6992
71. Néda Z, Ravasz E, Brechet Y, Vicsek T, Barabasi AL (2000) The sound of many hands clapping - Tumultuous applause can transform itself into waves of synchronized clapping. Nature 403:849–850
72. Nicolis G, Prigogine I (1977) Self-organization in non-equilibrium systems. Wiley, New York
73. Oullier O, Kelso JAS (2006) Neuroeconomics and the metastable brain. Trends Cogn Sci 10:353–354
74. Oullier O, Temprado JJ (2005) The structure of visual backgrounds modulates interpersonal coordination dynamics in a virtual environment. J Sport Exerc Psychol 27:S118-S119
75. Oullier O, Bardy BG, Stoffregen TA, Bootsma RJ (2002) Postural coordination in looking and tracking tasks. Hum Mov Sci 21:147–167
76. Oullier O, de Guzman GC, Jantzen KJ, Kelso JAS (2003) On context dependence of behavioral variability in inter-personal coordination. Int J Comput Sci Sport 2:126–128
77. Oullier O, de Guzman GC, Jantzen KJ, Lagarde JF, Kelso JAS (2004) Spontaneous interpersonal synchronization is modulated by the degree of visual coupling. J Sport Exerc Psychol 26:S11-S11
78. Oullier O, Bardy BG, Stoffregen TA, Bootsma RJ (2004) Task-specific stabilization of postural coordination during stance on a beam. Mot Control 8:174–187

79. Oullier O, Jantzen KJ, Steinberg FL, Kelso JAS (2005) Neural substrates of real and imagined sensorimotor coordination. Cereb Cortex 15:975–985
80. Oullier O, de Guzman GC, Jantzen KJ, Lagarde J, Kelso JAS (2005) Spontaneous interpersonal synchronization. In: Peham C, Schöllom WI, Verwey W (eds) European workshop on movement sciences: Mechanics-Physiology-Psychology. Sportverlag, Köln, pp 34–35
81. Oullier O, Kelso JAS, Kirman AP (2008) Neuroeconomics: A dynamical systems perspective. Rev Econ Pol 118:51–62
82. Oullier O, de Guzman GC, Jantzen KJ, Lagarde J, Kelso JAS (2008) Social coordination dynamics: Measuring human bonding. Soc Neurosci 3:178–192
83. Richardson MJ, Marsh KL, Schmidt RC (2005) Effects of visual and verbal interaction on unintentional interpersonal coordination. J Exp Psychol Hum Percept Perform 31:62–79
84. Richardson MJ, Marsh KL, Isenhower RW, Goodman JR, Schmidt RC (2007) Rocking together: dynamics of intentional and unintentional interpersonal coordination. Hum Mov Sci 26:867–891
85. Rizzolatti G, Craighero L (2004) The mirror-neuron system. Annu Rev Neurosci 27:169–192
86. Schmidt RC, O'Brien B (1997) Evaluating the dynamics of unintended interpersonal coordination. Ecol Psychol 9:189–206
87. Schmidt RC, Richardson MJ (2008) Dynamics of interpersonal coordination. In: Fuchs A, Jirsa VK (eds) Springer, Heidelberg, pp 281–308
88. Schmidt RC, Carello C, Turvey MT (1990) Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. J Exp Psychol Hum Percept Perform 16:227–47
89. Schöner G, Kelso JAS (1988) Dynamic pattern generation in behavioral and neural systems. Science 239:1513–1520
90. Schurmann M, Hesse MD, Stephan KE, Saarela M, Zilles K, Hari R, Fink GR (2005) Yearning to yawn: the neural basis of contagious yawning. Neuroimage 24:1260–1264
91. Shockley K, Baker AA, Richardson MJ, Fowler CA (2007) Articulatory constraints on interpersonal postural coordination. J Exp Psychol Hum Percept Perform 33:201–208
92. Simon JR (1969) Reactions toward the source of stimulation. J Exp Psychol 81:174–176
93. Solnick SJ, Schweitzer ME (1999) The Influence of Physical Attractiveness and Gender on Ultimatum Game Decisions. Organ Behav Hum Decis Process 79:199–215
94. Sommerville JA, Decety J (2006) Weaving the fabric of social interaction: Articulating developmental psychology and cognitive neuroscience in the domain of motor cognition. Psychon Bull Rev 13:179–200
95. Stefan K, Cohen LG, Duque J, Mazzocchio R, Celnik P, Sawaki L, Ungerleider L, Classen J (2005) Formation of a motor memory by action observation. J Neurosci 25:9339–9346
96. Temprado JJ, Laurent M (2004) Attentional load associated with performing and stabilizing a between-persons coordination of rhythmic limb movements. Acta Psychol 115:1–16
97. Temprado JJ, Swinnen SP, Carson RG, Tourment A, Laurent M (2003) Interaction of directional, neuromuscular and egocentric constraints on the stability of preferred bimanual coordination patterns. Hum Mov Sci 22:339–363
98. Tognoli E, Magne C, de Guzman GC, Kelso JAS (2007) Brain rhythms underlying intentional social coordination. Society Neurosci Program 304
99. Tognoli E, Lagarde J, de Guzman GC, Kelso JA (2007) The phi complex as a neuromarker of human social coordination. Proc Nat Acad Sci USA 104:8190–8195
100. Turvey MT (2004) Impredicativity, dynamics and the perception-action divide. In: Jirsa VK, Kelso JAS (eds) Coordination Dynamics: Issues and Trends. Springer, Berlin
101. Vallacher RR, Nowak A (1997) The emergence of dynamical social psychology. Psychol Inquiry 8:73–99 (and commentaries therein)
102. van Ulzen NR, Lamoth CJ, Daffertshofer A, Semin GR, Beek PJ (2008) Characteristics of instructed and uninstructed interpersonal coordination while walking side-by-side. Neurosci Lett 432:88–93
103. Vinkovic D, Kirman A (2006) A physical analogue of the Schelling model. Proc Nat Acad Sci USA 103:19261–19265
104. Wallenstein GV, Kelso JAs, Bressler SL (1995) Phase transitions in spatiotemporal patterns of brain activity and behavior. Physica D 84:626–634
105. Wheeler ME, Peterson SE, Buckner RL (2004) Memory's echo: Vivid remembering reactivates sensory-specific cortex (vol 97, pg 11125, 2000). Proc Nat Acad Sci USA 101:5181
106. Winfree AT (1967) Biological rhythms and behavior of populations of coupled oscillators. J Theor Biol 16:15–42
107. Winfree AT (1980) The geometry of biological time. Springer, New York
108. Winfree AT (2002) On emerging coherence. Science 298: 2336–2337

Books and Reviews

Frith C, Wolpert DM (2004) The neuroscience of social interaction: Decoding, imitating and influencing the actions of others. Oxford University Press, Oxford

Fuchs A, Jirsa VK (eds) (2008) Coordination: Neural, behavioral and social dynamics. Springer, Berlin

Jantzen KJ, Kelso JAS (2007) Neural coordination dynamics of human sensorimotor behavior: A Review. In: Jirsa VK, McIntosh AR (eds) Handbook on brain connectivity. Springer, Berlin, pp 421–461

Jeannerod M (2006) Motor cognition: What actions tell to the self. Oxford University Press, Oxford

Jirsa VK, Kelso JAS (2004) Coordination dynamics: Issues and trends. Springer, Berlin

Kelso JAS (1995) Dynamic Patterns: The self-organization of brain and behavior. MIT Press, Cambridge

Kelso JAS, Engstrøm DA (2006) The complementary nature. MIT Press, Cambridge

Oullier O, Jantzen KJ (2008) Neural indices of behavioral instability in coordination dynamics. In: Fuchs A, Jirsa VK (eds) Springer, Berlin, pp 205–227

Pikovsky A, Rosenblum M, Kurths J (2001) Synchronization: A universal concept in nonlinear science. Cambridge University Press, Cambridge

Rizzolatti G, Craighero L (2004) The mirror-neuron system. Ann Rev Neurosci 27:169–192

Sommerville JA, Decety J (2006) Weaving the fabric of social interaction: Articulating developmental psychology and cognitive neuroscience in the domain of motor cognition. Psychon Bull Rev 13:179–200

Strogatz SH (2003) Sync: The emerging science of spontaneous order. Hyperion Press, New York

Social Network Analysis, Estimation and Sampling in

Ove Frank
Statistics Department, Stockholm University, Stockholm, Sweden

Article Outline

Glossary
Definition of the Subject
Introduction
Background
Basic Terminology and Notation
Sampling Designs
Snowball Sampling Designs
Estimation Based on Snowball Samples
Network Walk Sampling Designs
Estimation Based on Network Walk Samples
Methods for Estimating Hidden Populations
Probabilistic Network Models
Bipartite Network Sampling Designs
A Bipartite Network Model for Crime Participation
Future Directions
Bibliography

Glossary

Network model A network model describes how an observed network structure with its data could be generated from probabilistic assumptions.

Network population A network population is a finite or infinite collection of existing or conceivable networks. For instance, the population could consist of network instances of a process changing with time, or the population could consist of network realizations in different sets of units.

Observation scheme An observation scheme describes what variables are known or can be observed for units and relations in the population and in the sample.

Population network A population network is a population of units equipped with relational structure between them. Usually several attributes are attached both to units and to pairs of units, and the network can be described as a valued graph on the population of units with vertex and edge or arc variables defined in the graph.

Sampling design A sampling design describes how units are sampled from a population of units and what the selection probabilities are for all the possible sample sequences or sample subsets. The sampling procedure is usually controlled by the investigator. If the sampling is not controlled by the investigator but voluntary or "controlled by nature", the sample network can be considered as an outcome of a probabilistic network model, and there is no clear separation between design-based and model-based inference.

Sampling frame A sampling frame is a list of identifying labels of all possible sample units.

Social network A social network is a network consisting of units and relations of interest in the social sciences. The term is used both for population networks and for networks belonging to a network population.

Survey sampling Survey sampling of networks refer to the investigation of networks sampled from a network population while survey sampling in networks refer to the investigation of sub-networks sampled from a population network. Survey sampling of networks is usually model-based with an infinite network population. Survey sampling in networks is usually design-based with a finite population network.

Definition of the Subject

Sampling in network analysis can refer either to the sampling of networks from a population of networks or to the sampling of one or more sub-networks from a population network. Sampling designs and model approaches based on probabilistic methods allow statistical tools to be developed for the analysis of data obtained in surveys carried out in a network setting.

From about the 1960s the rapid development of computers and computer power changed the scene for statistics and applied mathematics in general. New possibilities of processing large amounts of data by using fast computer algorithms for calculations, sorting, and searching contributed to an accelerated interest in computer science and discrete mathematics. Simulation techniques and other computer intensive numerical methods revolutionized the arsenal of tools available in statistics and other parts of applied mathematics. Graph theory was established not only as an expanding part of discrete mathematics but also became applied as a tool for investigating relational structures in the social sciences. Friendship, co-operation, dominance, support and other relations between individuals were studied as sociograms or as more general networks in which both the individuals and the relations were attributed with qualitative and quantitative variables like gender and age of the individuals and type and strength of the relations. Other network applications include transmission of infectious diseases between people, work scheduling and other planning, routing systems

for transportation of goods, traffic flows in urban areas, information flows in communication channels, administrative and organizational structure, capital transfer and surveillance systems. Characteristic of such network applications is the inherent complexity of co-variation or correlation that could exist within and between variables defined for different elements of the network.

Introduction

Special sampling designs using network information for successive link tracing or sample adjustment have been described in the literature under various names such as multiplicity sampling, snowball sampling, and Respondent driven sampling. Such designs have been used for investigations of hard-to-reach populations of various kinds, like drug addicts, homeless people and individuals at risk for hiv-infection, tuberculosis, or hepatitis. Such designs have also been proved useful in investigations of rare properties of individuals in populations that could be sampled from available frames but would require very large samples with conventional sampling designs.

This presentation focuses on sampling designs in finite population networks and network models developed for survey sampling of social networks. Estimation methods are described that can be applied under different assumptions about what population data are known and what sample data can be observed.

The next three sections give background material of various kinds. A brief historical background is given that describes the development of a survey sampling theory for networks. Basic terminology and some general notation needed is then specified and serves the purpose of unifying and simplifying some of the concepts and arguments used later. A common and characteristic feature of network surveys is that they allow non-orthodox sampling methods. Hence it is important to specify when such methods are justified. Some fundamentals about sampling designs are given together with an overview of some standard designs.

After this background follows five sections on network sampling designs. Sections "Snowball Sampling Designs"–"Estimation Based on Network Walk Samples" treat snowball sampling designs and the important sub-class of network walk designs. For various special cases and modifications of basic snowball and walk samples, standard estimation methods are discussed. Section "Methods for Estimating Hidden Populations" is devoted entirely to the estimation of sizes and properties of hidden populations and rare parts of large populations.

The discussion of network designs is followed by a presentation of various probabilistic network models used especially in the social sciences. Some of these models are appropriate in combination with designs or as tools for deriving so-called model-assisted estimators. The main interest in these network models stems from their use when an observed network is considered as an instant of a network process or as a sample network from a super-population of networks.

Next there are two sections on sampling designs and models for bipartite networks. Bipartite networks have units of two kinds which generally require very different treatment so that it is not appropriate to consider them as special cases of networks. Section "Bipartite Network Sampling Designs" describes sampling designs and estimators in bipartite networks. A combination of sampling designs and random modeling in bipartite networks is illustrated in Sect. "A Bipartite Network Model for Crime Participation" which treats a specific criminological application.

Finally, Sect. "Future Directions" comments briefly on possible future directions for the development of network surveys in the social sciences.

Background

Some early attempts to incorporate graphs in a statistical investigation are the studies on sociograms [49], on snowball sampling [33], and on sampling in graphs [3]. Of particular interest is a paper by Stephan [60] pointing to the need for network methods in sample survey theory. He suggests the term nexus sampling for surveys that consider data from units as well as from relations between units. Sampling in populations with graph structure is also discussed in the articles [6,14], and a comprehensive monograph on statistical inference in graphs is the thesis [16].

An independent discussion of some first steps in network sampling is given in [34,42]. Various problems in survey sampling and estimation in networks were discussed in a series of papers during the 1970s and 1980s: [7,17,18,19,20,21,22,23,24,25,26,27,48]. These papers use essentially various network sampling designs and no elaborate probabilistic model assumptions.

The theory of random graphs developed mainly from two models introduced by Erdös and Renyi in 1959 and 1960 [12,13], the uniform graph with n vertices and N randomly selected edges among the unordered pairs of vertices, and the Bernoulli graph with n vertices and independent insertions of edges with a common probability p for all the unordered pairs of vertices. These random graphs have challenged many mathematicians, and there is now an extensive literature on combinatorial and asymptotic results for random graphs. A brilliant comprehensive ref-

erence on general graph theory is [11]. Monographs concentrating on random graphs are [4,37,46].

Statistical inference based on probabilistic network models without explicitly addressing the sampling design problem has been treated for some variants of Bernoulli graphs, but generally such models have too little structure to be of interest in a statistical context with network data. A network model with a more elaborate and flexible parametrization that make it applicable in the social sciences is the random digraph model of Holland and Leinhardt introduced in [36]. A comprehensive monograph with many references to the social network literature is [67]. The introduction of Markov graphs in [30] and the log-linear modeling of random networks in [47,57,68] have had a major impact on the development of statistical methods for social network analysis. The edited volumes [5,8] are good sources for the methodological development. Various aspects of network modeling and statistical inference from network samples are the topics of several theses in recent years. Well known to me are the following important contributions to the methodological development: [10,35,38,40,42,55,59,62].

Basic Terminology and Notation

Survey sampling concepts of basic importance are populations of units, sample selection, sampling designs, observation schemes, and variables of interest defined on the units. Two standard reference works for survey sampling in finite populations are [54,63]. In order to make network sampling possible, the population has to be equipped with a relational structure between its units. Usually it is given as a binary variable defined on the ordered pairs of units. A population with a binary relation can be represented as a graph in which the vertices represent the units and the edges (arcs) indicate the unordered (ordered) pairs of units that are related. In a network setting vertices and edges (arcs) are sometimes called nodes and links.

The population from which samples are drawn can be different from the target population of the survey. It can also be changing with time so that particular precautions need to be taken. Disregarding such complications we assume that there is a finite population of N units labeled by integers $1, \ldots, N$. The population is denoted $U = \{1, \ldots, N\}$. Samples are drawn from U either as ordered sequences of units with or without repetitions allowed or as subsets of U. In the first case, a sample sequence of m units drawn is denoted $u = (u_1, \ldots, u_m)$, and in the second case a sample subset is denoted S. Both u and S are random entities when the sample is drawn according to a probability design. The probability design of u is given by a probability distribution over the set U^m of all ordered m-sequences of units from U, and the probability design of S is given by a probability distribution over the class of all subsets of U. If the size n of S is fixed, all subsets of other sizes have zero probability.

The observation scheme specifies which variables are known or observable. A variable defined on the units (a vertex variable) is given by a function x from U to some range space R. Formally,

$$x = \{(i, x_i) \colon i \in U\} ,$$

is a set of pairs (i, x_i) assigning a value x_i in R to each unit i in U. If there is a specified ordering of the units, x could also be represented as a sequence of values ordered in the same way, e.g. an ordered sequence $(x_1, \ldots, x_N)$ corresponding to the order $1, \ldots, N$ of the units. The first representation is more convenient when restrictions to different subsets of U are considered. For any subset A of U we denote by

$$x(A) = \{(i, x_i) \colon i \in A\} ,$$

the restriction of the variable x to units in A. In particular, $x(U) = x$ and $x(\emptyset) = \emptyset$. A binary variable x with $R = \{0, 1\}$ and, more generally, a K-category variable x with $R = \{0, \ldots, K-1\}$ partitions the population U into K disjoint subsets

$$U_j = \{i \in U \colon x_i = j\} \quad \text{for} \quad j \in R .$$

A variable defined on the ordered pairs of units (an arc variable) is given by a function y from U^2 to some range space R. Formally,

$$y = \{(i, j, y_{ij}) \colon i \in U, j \in U\} ,$$

is a set of triplets (i, j, y_{ij}) assigning a value y_{ij} in R to each ordered pair (i, j) in U^2. If there is a specified ordering of the units in U, y could also be represented as a matrix array (y_{ij}) with N rows and N columns ordered according to the unit-ordering. For any two subsets A and B of U we denote by

$$y(A, B) = \{(i, j, y_{ij}) \colon i \in A, j \in B\} ,$$

the restriction of the variable y to ordered pairs of units (i, j) with $i \in A$ and $j \in B$. The set $y(A, B)$ is simply referred to as the set of values of y from A to B. In particular, $y(U, U) = y$ and $y(A, \emptyset) = y(\emptyset, B) = \emptyset$ for any subsets A and B of U. A binary variable y with $R = \{0, 1\}$ and, more generally, a K-category variable y with $R = \{0, \ldots, K-1\}$ partitions U^2 into K disjoint subsets

$$\{(i, j) \in U^2 \colon y_{ij} = r\} \quad \text{for} \quad r \in R .$$

Sampling Designs

In order to gain general information about population units with access only to the information from a sample of units, it is evident that the sample should represent the population in some specified way. Even if the population is partitioned into subpopulations according to several attributes of the units, and the sample is selected so that some units from each subpopulation are included in the sample, one cannot be sure that the sample allows inference to be drawn with confidence about collective properties of all the population units.

Survey sampling theory is based on information from samples that are selected according to probabilistic designs. A sampling design that gives all population units a positive probability of being included in the sample is required to achieve a sample that is representative of the population. The essential property of a sampling design that makes it appropriate for population inference is that its inclusion probabilities should be known or estimable for all units. This allows sample data to be weighted so that they accurately represent population data. A probabilistic sampling design not only makes inference possible; it also makes it possible to specify how inference uncertainty can be judged and quantified. During the early days of survey sampling, the benefits of probabilistic sampling designs were not always properly understood and had to be emphasized. Today this is so well known that it may seem dubious if someone tries to use non-probability samples.

Therefore, it is important to clarify the legitimacy of network sampling designs that involve arbitrarily or conveniently selected samples. The essential principle is still that inclusion probabilities should be known or estimable. Some network sampling designs are based on an initial probability sample followed by sequential samples that may be systematically or probabilistically selected. As a consequence, it might be hard to estimate inclusion probabilities for the final network sample, but there are no doubts about its legitimacy.

When the network sample is generated from an initial sample that is not a probability sample, it is important that the sequential samples generated after the initial sample are selected according to specified probabilistic rules. Under some fairly general assumptions to be specified, the inclusion probabilities for the final network sample can be proved to converge towards limits that are independent of the way the initial sample was selected. Thus, the lack of knowledge about the initial conditions is compensated for by letting the network sample consist of several waves of units sequentially selected according to a known or estimable probability design. This wave design determines the limiting inclusion probabilities of the network sampling design, and the accuracy of these probabilities increases with the number of waves. As we will see later, it is possible to modify the probabilistic wave design so that the limiting inclusion probabilities of the network sample become uniform or equal to any other desired probability distribution over the population of units.

It is convenient to specify a few standard sampling designs and their inclusion probabilities which are fundamental for some of the estimators discussed later. Consider a population U of N units and a sample subset S selected according to a uniform probability design over the n-subsets of U. This design is specified as

$$S \sim \mathrm{Unif}(n, U) .$$

It is often called simple random sampling without replacement. Thus, the probability that $S = s$ is equal to

$$P(s) = n!/N(N-1)\ldots(N-n+1) ,$$

for any n-subset s of U. The probability of inclusion in S of a fixed unit i from U is equal to the sum of the selection probabilities $P(s)$ over all s containing unit i. It is equal to n/N and denoted by

$$\pi_i = n/N ,$$

for any i in U. The probability of inclusion in S of two distinct units i and j from U equals

$$\pi_{ij} = n(n-1)/N(N-1) ,$$

for any distinct i and j in U.

Another subset design called Bernoulli sampling with parameter $p(0 < p < 1)$ is specified as

$$S \sim \mathrm{Bern}(p, U) ,$$

and is defined by selection probabilities

$$P(s) = p^n(1-p)^{N-n} ,$$

for any subset s of U of size n for $n = 0, 1, \ldots, N$. Its inclusion probabilities are

$$\pi_i = p ,$$

for any i in U, and

$$\pi_{ij} = p^2 ,$$

for any distinct i and j in U. An important property of Bernoulli sampling is that its inclusion indicators

$I_i = I(i \in S)$ are independent. The inclusion indicators are 1 or 0 according to whether unit i is included in S or in its complement $U - S$.

A generalization of Bernoulli sampling is obtained if the inclusion indicators are independent but not necessarily identically distributed Bernoulli variables. The sampling design with independent Bernoulli(p_i) distributed inclusion indicators for i in U is sometimes called Poisson sampling (perhaps referring to the fact that the random sample size is approximately Poisson distributed with parameter $\lambda = p_1 + \cdots + p_N$). We specify this design as a Bernoulli sampling with general parameter $p = \{(i, p_i) \colon i \in U\}$ satisfying $0 < p_i < 1$ for all i in U; in short

$$S \sim \mathrm{Bern}(p_i \colon i \in U) .$$

Its selection probabilities are

$$P(s) = \prod_{i \in s} p_i \prod_{j \in U - s} (1 - p_j) ,$$

for any subset s of U, and its inclusion probabilities are

$$\pi_i = p_i ,$$

for any i in U, and

$$\pi_{ij} = p_i p_j ,$$

for any distinct i and j in U.

A probability design for a sample sequence $u = (u_1, \ldots, u_m)$ of m units independently drawn according to a uniform distribution over U has selection probabilities

$$P(i_1, \ldots, i_m) = 1/N^m ,$$

for any sequence $(i_1, \ldots, i_m)$ in U^m. This design specified by

$$u_k \sim \mathrm{Unif}(U) \quad \text{independent for} \quad k = 1, \ldots, m ,$$

is often called simple random sampling with replacement. Its inclusion probabilities are given by

$$\pi_i = 1 - (1 - 1/N)^m ,$$

for any i in U, and

$$\pi_{ij} = 1 - 2(1 - 1/N)^m + (1 - 2/N)^m ,$$

for any distinct i and j in U.

A generalization of this design to m independent identically distributed (IID) draws according to a probability distribution with N arbitrary positive probabilities $p(i)$ summing to 1 for i in U has selection probabilities

$$P(i_1, \ldots, i_m) = p(i_1) \ldots p(i_m) ,$$

for any sequence $(i_1, \ldots, i_m)$ in U^m. We specify this design by

$$u_k \sim \mathrm{IID}(p(i) \colon i \in U) \quad \text{for} \quad k = 1, \ldots, m .$$

Its inclusion probabilities are

$$\pi_i = 1 - [1 - p(i)]^m ,$$

for any i in U, and

$$\pi_{ij} = 1 - [1 - p(i)]^m - [1 - p(j)]^m + [1 - p(i) - p(j)]^m ,$$

for any distinct i and j in U. A further generalization would allow the sequence of independent draws to be made according to different distributions.

For any probability design of a sample sequence $u = (u_1, \ldots, u_m)$ we define the inclusion count or multiplicity m_i of unit i as the number of draws equal to that unit. The multiplicities sum to m, and a positive multiplicity of a unit indicates its inclusion in the sample sequence. The set of distinct units in a sample sequence u is denoted by $s(u)$. This is a random subset of U of size $n(u)$ and expected size

$$\mathrm{E}\ n(u) = \sum_{i \in U} \pi_i ,$$

and variance

$$\mathrm{Var}\ n(u) = \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) .$$

Any sample sequence $u = (u_1, \ldots, u_m)$ with $u_k \sim \mathrm{IID}(p(i) \colon i \in U)$ for $k = 1, \ldots, m$ has multiplicities $(m_1, \ldots, m_N)$ that are multinomially distributed with parameters m and $(p(1), \ldots, p(N))$. The number of distinct units $n(u)$ in u is a sum of independent Bernoulli(π_i)-variables, which is asymptotically Poisson($\sum_i \pi_i$). With dependent draws it is generally difficult to derive the simultaneous probability distribution of the multiplicities or the distribution of the number of distinct units in u.

Snowball Sampling Designs

Let the population $U = \{1, \ldots, N\}$ have a graph structure given by a binary variable y defined on the pairs of population units. Let the sets of units adjacent after and before unit i be

$$A_i = \{j \in U \colon y_{ij} = 1\} \quad \text{and} \quad B_i = \{j \in U \colon y_{ji} = 1\} ,$$

for $i \in U$. For any subset s of U, we denote by $A(s)$ the union of the A_i for $i \in s$. A snowball sampling design starts with an initial subset S_0 of U. This subset can either be

a convenience sample or the outcome of a probability sampling design with a known or unknown probability distribution P_0 over the class of all subsets of U. In the latter case there are selection probabilities $P_0(s) \geq 0$ summing to 1 for all subsets s of U. In the former case S_0 is equal to a fixed set s_0, and it is convenient to put $P_0(s_0) = 1$. Note that in either case we don't even know that all units in the population have positive probabilities of being included in the initial sample.

The snowball sample is obtained by successively extending the initial sample S_0 with units selected from the sets of units that are adjacent after previously joined units. More precisely, a first wave W_1 is selected according to a probability distribution over the class of subsets of

$$(U - S_0) \cap A(S_0) .$$

The union of the initial sample S_0 and the disjoint first wave W_1 is equal to the one-wave snowball S_1. Generally, the kth wave W_k is selected according to a probability distribution over the class of subsets of

$$(U - S_{k-1}) \cap A(W_{k-1}) ,$$

for $k = 1, 2, \ldots$ with $W_0 = S_0$. The k-wave snowball S_k is the union of the $(k-1)$-wave snowball S_{k-1} and the kth wave W_k. The waves are disjoint and

$$W_k = S_k - S_{k-1} \quad \text{for} \quad k = 1, 2, \ldots$$

The k-wave snowball S_k is called saturated if the kth wave W_k is the last wave that is non-empty, so that the sequence of snowballs has reached an absorbing state

$$S_k = S_{k+1} = \cdots$$

According to the selections of the waves, it follows that the k-wave snowball S_k evolves as a stochastic process depending on the $(k-1)$-wave snowball S_{k-1} and the kth wave which depends on S_{k-1} and $W_{k-1} = S_{k-1} - S_{k-2}$. Thus, the snowball process S_k has a two-step memory: The probability distribution of S_k conditional on $S_0, \ldots, S_{k-1}$ depends on S_{k-1} and S_{k-2} only, for $k = 2, 3, \ldots$ Another equivalent way to express this is to say that the stochastic process with states consisting of the current snowball and its last wave (S_k, W_k) is a Markov chain.

The transition probabilities of this Markov chain are

$$\begin{aligned} P(S_{k+1} = s_{k+1}, W_{k+1} = w_{k+1} | S_k = s_k, W_k = w_k) \\ = P(W_{k+1} = w_{k+1} | S_k = s_k, W_k = w_k) \\ = P(w_{k+1} | s_k, w_k) , \end{aligned}$$

for s_{k+1} equal to the union of s_k and w_{k+1}. The Markov chain is assumed to be time-homogeneous so that the transition probabilities do not depend on the stage parameter k. The transition probabilities do depend on graph parameters

$$y(w_k, U - s_k) = \{(i, j, y_{ij}) \colon i \in w_k, j \in U - s_k\} ,$$

governing possible wave selections and possibly also on other parameters governing properties of the wave selections. Assume that saturation occurs at stage k or later. Then the sequence of snowball samples $(S_0, S_1, \ldots, S_k)$ has a probability distribution given by

$$\begin{aligned} P(S_0 = s_0, S_1 = s_1, \ldots, S_k = s_k) &= P(s_0, s_1, \ldots, s_k) \\ &= P_0(s_0) P(w_1 | s_0, w_0) P(w_2 | s_1, w_1) \\ &\quad \cdots P(w_k | s_{k-1}, w_{k-1}) , \end{aligned}$$

for any strictly increasing subset sequence $(s_0, s_1, \ldots, s_k)$ with $s_j - s_{j-1} = w_j$ for $j = 1, 2, \ldots, k$ and $s_0 = w_0$.

The probability distribution of the k-wave snowball S_k is obtained by summing the probabilities $P(s_0, s_1, \ldots, s_k)$ over all strictly increasing subset sequences $(s_0, s_1, \ldots, s_{k-1})$ with s_{k-1} strictly included in s_k. The initial sample and each new wave need to contain at least one unit, so that s_k contains at least $k + 1$ units. If s_k contains more units, the sum is extended over all possible distributions of the non-initial units among the k waves.

In order to be able to perform such calculations, we need to specify the common sampling design of the waves. At stage k, the next wave W_{k+1} has a probability distribution over the class of subsets of that part of U that is not included in the current snowball S_k but is adjacent after its last wave W_k. Now the selection of a subset for W_{k+1} can be considered as obtained by the recruitment of new units made by units in W_k. Each unit i in W_k can recruit only from A_i, the set of units adjacent after unit i. Assume that the recruited set R_i is a Bernoulli sample with inclusion probabilities p_{ij} for j in A_i. The recruited sets R_i are assumed to be selected independently for all i in U. The union of the sets R_i recruited by the units i in W_k is the set $R(W_k)$. This set is a subset of $A(W_k)$, and the part of it that is not included in S_k is the new wave W_{k+1}. Thus

$$W_{k+1} = (U - S_k) \cap R(W_k) ,$$

where $R(W_k)$ as a union of independent Bernoulli samples is itself a Bernoulli sample conditional on W_k. The probability that unit j is included in $R(W_k)$ equals the probability that j is included in at least one of the R_i for i in W_k. This probability is denoted

$$p(W_k, j) = 1 - \prod (1 - p_{ij}) ,$$

where the product is over i in W_k and j is a unit in $A(W_k)$. It follows that conditional on the current state of the Markov chain (S_k, W_k), the next wave W_{k+1} is a Bernoulli sample with inclusion probabilities

$$p(W_k, j) \quad \text{for } j \text{ in } (U - S_k) \cap A(W_k) .$$

The transition probabilities of the Markov chain (S_k, W_k) can now be specified as

$$P(w_{k+1}|s_k, w_k) = \left[\prod p(w_k, j)\right]\left[\prod q(w_k, j)\right] ,$$

where the first product is over j in w_{k+1} and the second product is over j in $U - s_{k+1}$ with probabilities denoted $q(w_k, j) = 1 - p(w_k, j)$. The basic parameters here are the probabilities p_{ij} that unit i would recruit unit j. Only units j in A_i can be recruited, and only units i included in the initial sample or any of the waves will have an opportunity to recruit. We can interpret the parameters p_{ij} as probabilities of independent recruitment arcs in the population graph. When y is considered as specifying a fixed population graph, p_{ij} is zero whenever y_{ij} is zero. If all units in A_i should have the same probability of being recruited, we could put $p_{ij} = \alpha_i y_{ij}$ where α_i is a probability governing the recruitment or co-operation activity of unit i. If recruitment also depends on availability, status or attraction of the recruited unit, we could put $p_{ij} = \alpha_i \beta_j y_{ij}$. The parameters α_i and β_i for units i in U could be related to a categorical variable x defined on U. For a K-category variable x the parameters $\alpha_i = \alpha(x_i)$ and $\beta_i = \beta(x_i)$ have K possible values each, and the parametrization involves only $2K$ parameters.

Recruitment or selection of adjacent members of the population can be considered as a sampling performed by the investigator or by the respondent. The units of the population can be people that could be interviewed by the investigator. The units can also be geographical regions or other entities for which the adjacent units can either be observed directly or determined by interviews with some unit representative. In such cases when the investigator can get access to information about all units that are adjacent after a sampled unit, then the recruitment of further units can be made by the investigator. There are situations when information about adjacent units might be sensitive or embarrassing to the respondent. If the respondent is unwilling to release information about all adjacent units, a possibility might be to ask the respondent just to select some of them at random. Such respondent driven sampling might in some situations give more reliable data and could provide a reasonable way to avoid some problems of integrity. For a more thorough discussion of this, see, for instance, [53]. There are techniques developed with distribution of recruitment cards and various benefits to participants in order to help the investigator to get new waves of the sample. Even if respondent i provides only R_i and not A_i, information could be collected about out-degrees or other statistics needed for the respondent driven probability design of the wave sampling.

The combination of design based and model based probabilities are standard in survey sampling and should be so also when network sampling is applied. Respondent driven sampling could be such a case when a random graph approach would be appropriate for the wave sampling. An illustration of combination of design-based and model-based approaches to network surveys is given in [66].

Estimation Based on Snowball Samples

Estimation of population data from snowball sample data requires inclusion probabilities of all population units. Any fixed unit j is included in a k-wave snowball sample if it is included in the initial sample or in any of its k following waves. Since the initial sample and the waves are disjoint, their inclusion events are mutually exclusive, and the probability that the snowball includes j equals the sum of the probabilities that the initial sample or any of the waves include j:

$$P(j \in S_k) = P(j \in S_0) + P(j \in W_1) + \cdots + P(j \in W_k) .$$

Obviously this inclusion probability increases strictly with the number of waves until the snowball is saturated. If the initial sample is a convenience sample, its inclusion probability is 1 or 0 depending on whether j belongs to it or not. According to the results derived above for the conditional distribution of the next wave W_{k+1} given the current state (S_k, W_k), the conditional probability that a unit j not in S_k should be included in W_{k+1} is given by

$$p(W_k, j) = 1 - \prod(1 - p_{ij}) ,$$

where the product extends over units i in W_k. Therefore

$$P(j \in W_{k+1}) = \mathrm{E}\ [1 - I(j \in S_k)]p(W_k, j) ,$$

where the expectation is over outcomes of (S_k, W_k). The probability of inclusion of unit j in S_{k+1} is obtainable as the sum of the probabilities of inclusion in the waves, but usually it is easier to first consider the conditional probability of inclusion of j in S_{k+1} given the current state (S_k, W_k). This probability is given by $p(S_k, j)$ if we define $p(S_k, j) = 1$ for $j \in S_k$. This is achieved by defining $p_{jj} = 1$ for $j \in U$. Hence, the unconditional inclusion

probability is given by

$$P(j \in S_{k+1}) = \mathrm{E}\ p(S_k, j) = 1 - \mathrm{E} \prod_i [1 - I(j \in S_k) p_{ij}],$$

where the expectation is over outcomes of S_k. Even if the initial sample and all the recruitment samples are independent Bernoulli samples, it doesn't follow that S_k is a Bernoulli sample, and we cannot carry out the expectation factor by factor to obtain a recurrence relation between single unit inclusion probabilities of the current and the next snowball. In fact, it can be shown that for distinct i and j in U

$$\pi_{ij}(k) \geq \pi_i(k)\pi_j(k)\ ,$$

where $\pi_j(k) = P(j \in S_k)$ and $\pi_{ij}(k) = P(i \in S_k\ \&\ j \in S_k)$ for $k = 0, 1, \ldots$ Equality holds true if and only if there are no units of positive probability of inclusion in S_0, which are k steps adjacent before both i and j. An approximation that deserves to be further investigated for snowballs with few waves is

$$\pi_j(k+1) = 1 - \prod_i [1 - \pi_i(k) p_{ij}]\ ,$$

or, equivalently, using logarithms

$$\log[1 - \pi_j(k+1)] = \sum_i \log[1 - \pi_i(k) p_{ij}]\ .$$

When the inclusion probabilities are not too close to 1, we can obtain a linear approximation to the recurrence relation by series expansions of the logarithmic functions:

$$\pi_j(k+1) = \sum_i \pi_i(k) p_{ij}\ .$$

Since the inclusion probabilities are fundamental for many estimators, it would be of interest to know if there are sampling designs for which this recurrence relation is an acceptable approximation.

Consider a one-wave snowball with Bernoulli samples as described above. There are several possibilities for data collection that need to be distinguished. Assume that the value of a vertex variable x and the degree are observed for sample units. If the units are identifiable and individual recruitments are observed, data consist of (i, x_i, y_i) for $i \in S_0$ and (j, x_j, y_j) for $j \in R_i$ and $i \in S_0$. Units have multiplicities 0 or 1 in S_0, but their multiplicities in S_1 might be larger and that applies also to units in S_0 if they are recruited in the first wave. If only collective recruitments are observed, data consist of (i, x_i, y_i) for $i \in S_0$ and (j, x_j, y_j) for $j \in R(S_0)$. In this case, multiplicities in S_1 can be 0, 1, or 2, and 2 applies only to units in S_0. If only collective recruitments not included in S_0 are observed, data consist of (i, x_i, y_i) for $i \in S_0$ and (j, x_j, y_j) for $j \in W_1$. Now multiplicities are never larger than 1. If units are not separated between S_0 and W_1, data consist of (i, x_i, y_i) for $i \in S_1$. Also in this case multiplicities are never larger than 1. If the units are not identifiable and individual recruitments are observed, data consist of (x_i, y_i) for $i \in S_0$ and (x_j, y_j) for $j \in R_i$ and $i \in S_0$ with no possibility to observe multiplicities and no possibility to reduce data to $R(S_0)$, W_1, or S_1. If only collective recruitments are observed, we cannot separate S_0 and W_1 and we don't know S_1 or any multiplicities. If only collective recruitments of units not included in S_0 are observed, we can separate S_0 and W_1 and we know S_1 and we know that no multiplicities are larger than 1. Without identities and without separation of S_0 and W_1, data consist of (x_j, y_j) for $j \in S_1$. Depending on the observation scheme available, different estimators can be used.

To illustrate various estimators, it is convenient to use snowball sample indicators $S_{kj} = I(j \in S_k)$ for $j \in U$ and $k = 0, 1, \ldots$ and recruitment indicators $z_{ij} = I(j \in R_i)$ for $i \in U$ and $j \in U$. According to our assumptions S_{0j} are Bernoulli(p_j) for $j \in U$, and z_{ij} are Bernoulli(p_{ij}) for $i \in U$ and $j \in U$, and all these variables are independent. Here $0 \leq p_j \leq 1$ for $j \in U$, and by allowing the values 0 and 1 initial convenience samples are possible. Recruitments are restricted by the presence of edges in the acquaintance graph, which implies that $0 \leq p_{ij} \leq y_{ij}$ for distinct i and j in U. For convenience we define $p_{jj} = 1$ for $j \in U$. Note that even if the acquaintance graph is undirected with $y_{ij} = y_{ji}$, the recruitment graph given by arc indicators z_{ij} is directed.

Consider the total $T = \sum x_i$ of a vertex variable x and the total $D = \sum y_i$ of degrees in the acquaintance graph. Let us specify $p_j = y_j/D$ and $p_{ij} = y_{ij}/y_i$ for $i \neq j$. Assuming D known and using only data from S_0, we have an unbiased estimator of T given by

$$T' = \sum_i (x_i S_{0i}/p_i) = \sum_i (x_i S_{0i} D/y_i)\ .$$

Similarly the total population size N has an unbiased estimator

$$N' = \sum_i (S_{0i} D/y_i)\ .$$

Without assuming D known, we have an asymptotically unbiased estimator of the population mean of x according to

$$T'/N' = \sum_i (x_i S_{0i}/y_i) \Big/ \sum_i (S_{0i}/y_i)\ .$$

Thus, the average of x-values in S_0 is weighted by reciprocal degrees.

Using individual recruitment data, an unbiased estimator of T is given by

$$T'' = \sum_j \left(x_j \sum_i S_{0i} z_{ij} \Big/ \sum_i p_i p_{ij} \right)$$
$$= \sum_i \sum_j (x_j S_{0i} z_{ij} D / y_j) \,.$$

Also here the degree total D cancels from the asymptotically unbiased estimator of the population mean of x. We get

$$T''/N'' = \sum_i \sum_j (x_j S_{0i} z_{ij} / y_j) \Big/ \sum_i \sum_j (S_{0i} z_{ij} / y_j) \,.$$

Thus, the average of x-values in S_1 is counted with multiplicities and weighted by reciprocal degrees.

Using collective recruitments and only data from S_1, an unbiased estimator of T is given by

$$T''' = \sum_j [x_j S_{1j} / \pi_j(1)] \,,$$

where

$$\pi_j(1) = 1 - \prod_i (1 - p_i p_{ij})$$
$$= 1 - \prod_i (1 - y_{ij}/D) = 1 - (1 - 1/D)^{y_i} \,.$$

This time D does not appear as a factor in T''' or in the corresponding N''', so it does not cancel in the population mean estimator T'''/N'''. Replacing x_j by y_j in the estimators T', T'', and T''', we obtain D', D'', and D''' that all depend on D. The first two wouldn't provide any possibility to iteratively determine a value of a D-estimator. For D''' an iterative method would start with an initial value of D in $\pi_j(1)$, find D''', use this D''' for D in $\pi_j(1)$, find a new D''', etc. Convergence properties of such algorithms and properties of estimators that can be obtained from them haven't been explored. We will comment on similar iterative methods in Sect. "Estimation Based on Network Walk Samples".

When data from snowballs with more than one wave are available, it might be comparatively easy to set up estimators based on a first few waves, but more complicated to find the required inclusion probabilities for snowballs of many waves. An important technique described in [64] is to use Rao-Blackwellization to improve estimators based on data from only a few waves to obtain estimators based on all data. Rao-Blackwellization of an estimator requires extensive numerical calculations of its possible values for all arrangements of available data that are consistent with the reduced data remaining when multiplicities of sample units and order between the sample units are ignored. The expected value of an estimator conditional on a sufficient statistic in the form of the reduced data is the improved estimator. Sufficiency in survey sampling from a finite population was treated in [1]. More recent references are [9,44].

Network Walk Sampling Designs

Assume that the population $U = \{1, \ldots, N\}$ has a graph structure given by a binary variable y defined on the set of ordered pairs of population units. Let A_i and B_i be the sets of units adjacent after and before unit i. The size of A_i is the out-degree of unit i, and the size of B_i is the in-degree of unit i. When the graph is undirected, $A_i = B_i$ and out- and in-degrees are equal and called degrees.

A sampling design based on a random walk in a graph selects the sample as a sequence of units $u_1, u_2, \ldots$ according to a time-homogeneous Markov chain with transition probabilities

$$P(u_{k+1} = j | u_k = i) = P_{ij} \,,$$

that are independent of the stage parameter k for j in A_i and $k = 1, 2, \ldots$ If the graph is directed and strongly connected so that every unit has a directed path to every other unit, and if not all cycle lengths are multiples of a common factor larger than one, then the Markov chain will be irreducible and aperiodic. If the graph is undirected, connected, and contains at least one cycle of length one or three, then the Markov chain will be irreducible and aperiodic. A time-homogeneous, irreducible, and aperiodic Markov chain has a unique limiting distribution

$$\lim P(u_k = j | u_1 = i) = p_j > 0 \,,$$

independent of the initial unit i as k tends to infinity. The limiting distribution is equal to the stationary distribution that can be obtained by solving the equation system

$$p_j = \sum p_i P_{ij} \quad \text{for} \quad j = 1, \ldots, N \,,$$

subject to $\sum p_j = 1$. For instance, let the transition probability be $P_{ij} = y_{ij}/y_i$ for an undirected graph with degree y_i of unit i. This means that the transition from unit i is made with equal probabilities to each one of the y_i units that are adjacent to it. It follows that $p_i = y_i/(y_1 + \cdots + y_N)$. Hence, transition from any unit to one of its adjacent units according to a uniform probability distribution implies that the limiting probabilities of the units are proportional to their degrees.

If the limiting probabilities p_i for the units i in U cannot be determined from known or observable data, ergodic theory proves that they can be estimated by the relative frequencies in a long sample sequence obtained according to any time-homogeneous, irreducible, and aperiodic Markov chain. To be more specific, consider a sequence $u = (u_1, \ldots, u_m)$ of length m. The multiplicities of the units are m_i for i in U. The relative multiplicity sequence $(m_1/m, \ldots, m_N/m)$ converges towards the limiting distribution $(p_1, \ldots, p_N)$ when m tends to infinity.

Let $n(u)$ be the size of the set $s(u)$ of distinct units in u. The probability π_i that unit i is included in $s(u)$ equals the probability that $m_i > 0$. The multiplicity sequence is a reduction of the information in the sample sequence u that ignores the order of selection. The set $s(u)$ is a further reduction that ignores both the order of selection and the multiplicities of the units selected.

The graph employed for random walk sampling determines the zero entries among the transition probabilities $P = (P_{ij})$. These network-induced transition probabilities determine the limiting stationary distribution $p = (p_1, \ldots, p_N)$. If the initial unit u_1 is selected according to the stationary distribution p, this distribution applies to the marginal distribution of any position u_k of the walk. Another initial distribution implies that the stationary distribution is approached asymptotically on the walk for positions u_k as k increases. The limiting stationary distribution is important for appropriate weighting of data collected by network walk samples. Therefore, it is of special interest that it is possible to achieve any pre-assigned limiting distribution by adjusting the transition probabilities P induced by the network.

Assume that it is desirable to obtain a limiting distribution $q = (q_1, \ldots, q_N)$ that is different from the stationary distribution p derived from the transition probabilities P. Modified new transition probabilities $Q = (Q_{ij})$ satisfying the reversibility condition $q_i Q_{ij} = q_j Q_{ji}$ would lead to the stationary distribution q, which is also the limiting distribution when transitions are made according to the new Q. If we define Q by

$$Q_{ij} = P_{ij} \min(1, q_j P_{ji}/q_i P_{ij}) ,$$

for $i \neq j$ and

$$Q_{ii} = 1 - \sum_{i \neq j} Q_{ij} ,$$

it follows that

$$q_i Q_{ij} = \min(q_i P_{ij}, q_j P_{ji}) = q_j Q_{ji} ,$$

i. e. the reversibility condition is satisfied. To implement transitions according to Q, we can tentatively apply P and then decide to accept or reject a suggested transition from i to j with probability Q_{ij}/P_{ij} and $1 - Q_{ij}/P_{ij}$, respectively. More specifically, if the walk at stage k is in position $u_k = i$, and the next transition according to P would lead to unit j, the investigator might have to collect information from j to be able to calculate P_{ji} and compare it with P_{ij}. If $q_j P_{ji} \geq q_i P_{ij}$, then the transition to j has a probability valid also according to Q, and the walk is allowed to jump to $u_{k+1} = j$. Otherwise, the transition to j has not a valid probability according to Q, and, unless a transition is accepted, the walk has to stay at $u_k = i$ for another trial. Equivalently we could put $u_{k+1} = i$ and let the stage parameter count the number of trials instead of the number of jumps to other positions. Note that the probability that the modified walk stays for another trial equals the sum of the probabilities that the walk according to P jumps to a unit to be rejected, which is

$$\sum_{i \neq j} P_{ij}(1 - Q_{ij}/P_{ij}) = 1 - \sum_{i \neq j} Q_{ij} .$$

Consider the following illustrations for a directed network with transition probabilities $P_{ij} = y_{ij}/a_i$ for $j \in A_i$. Here the stationary probabilities can be difficult to find. Assume that we want a stationary distribution with probabilities q proportional to the out-degrees: $q_j = a_j/(a_1 + \cdots + a_N)$. The modified transition probabilities are

$$Q_{ij} = P_{ij} \min(1, q_j P_{ji}/q_i P_{ij}) = \min(y_{ij}, y_{ji})/a_i ,$$

and only jumps to mutual contacts are accepted. We need to collect information about out-degrees and be able to verify mutuality. Assume instead that we want a stationary distribution proportional to in-degrees: $q_j = b_j/(b_1 + \cdots + b_N)$. Then

$$Q_{ij} = (y_{ij} y_{ji}/b_i) \min(b_i/a_i, b_j/a_j) .$$

Only jumps to mutual contacts are possible, and only those with $b_i/a_i \leq b_j/a_j$ have transition probabilities that are valid also according to Q. Thus, unit j can be rejected unless it is a mutual contact with at least the same ratio of out-degree to in-degree as unit i. In this case we need to collect information about out- and in-degrees and be able to verify mutuality. Finally, assume that a uniform limiting distribution $q_j = 1/N$ is wanted. Then

$$Q_{ij} = \min(P_{ij}, P_{ji}) = y_{ij} y_{ji}/\max(a_i, a_j) ,$$

and j can be rejected unless it is a mutual contact that has at most the same out-degree as unit i. This case requires information about out-degrees and mutuality. Further illustrations are provided in [65]. A general presentation

of methods for constructing transition kernels (probabilities or probability densities) with desired properties, including the Metropolis–Hastings algorithm and the Gibbs sampler, can be found in Sect. 3.6 of [39].

The random walk sampling discussed so far has been in undirected graphs that are connected or in directed graphs that are strongly connected. If the connectedness assumption is uncertain for the graph being employed, or if it is known that it is not valid, then the Markov chain is reducible, and the transitions cannot reach beyond the connected component induced by the initial sample. In order not to confine the chain to a part of the population only, transitions to adjacent units could sometimes be replaced by transitions to other units.

A way to control the balance between staying on the current walk and starting on a new one, is to use a mixture distribution for the transitions. Let $P' = (P'_{ij})$ be a matrix of probabilities for transitions to adjacent units, and let $P'' = (P''_{ij})$ be a matrix of probabilities for transitions to arbitrary units in the population. A mixture distribution with transition probabilities given by

$$P = \alpha P' + (1-\alpha)P'' ,$$

where $0 < \alpha < 1$, is said to have a damping factor α measuring the proportion of transitions that are selected according to P'. For instance, $P'_{ij} = y_{ij}/a_i$ for j in A_i and $P''_{ij} = 1/N$ for j in U represent a mixture between uniform transitions from i to A_i and uniform transitions from i to U. The walk stops when it reaches a unit with $a_i = 0$. A slight modification is to define P'_{ij} also when $a_i = 0$, so that the walk is never forced to stop. For α close to 1, the damping effect will be small, and a population with a disconnected graph will be poorly represented by walk sample data. For decreasing α, the damping effect increases, the influence of the graph decreases, and the importance of having an appropriate matrix P'' increases. It might be appropriate to assume, for instance, that global transitions are made with probabilities proportional to out-degrees or in-degrees of the units, which would reflect that activity or attraction of the units is likely to influence the selections. In this way, the damping effect needed to handle a disconnected graph does not necessarily diminish the influence of the graph.

When network walk sampling is applied to webpage transitions in [70], a model is used that in our notation can be given as $P'_{ij} = y_{ij}/a_i$ if $a_i > 0$ and $P'_{ij} = 1/N$ if $a_i = 0$, and $P''_{ij} = p''_j$. Here p''_j reflects the web surfer's preferences for various pages, and the damping factor α reflects tendencies to use page links for surfing rather than own preferences.

Estimation Based on Network Walk Samples

Assume that sample data consist of the values of a variable x defined on the population of units. When units in the sample sequence can be identified, sample data consists of m values of x that can be referred to the units. The values (i, x_i, m_i) are known for all units i in $s(u)$. When units in the sample sequence cannot be identified, sample data consist of m values of x that cannot be referred to the units. Neither the multiplicities nor the set of distinct units can be determined. Sample data are restricted to the frequency distribution of the x-values.

There is an intermediate situation that could appear when sample data consist of the values of a variable x defined on U. If units in the sample sequence cannot be identified, it could still be possible to successively mark them in some way and observe when the same unit is sampled again. Data then consist of a sequence of m values on x that cannot be referred to the units but that can be partitioned into n sub-sequences of constant values belonging to the same unknown unit. Notice that the same value might appear in different sub-sequences. The multiplicities are known but their unit affiliations are not, and the number of distinct units is known but not their identities. The values (x_i, m_i) are known for n unknown distinct units i in U. If x is a K-category variable, the frequency distribution of these values (x_i, m_i) can be displayed in a K by m array representing available sample data when identities of units can not be observed but similarity between units is observed. Let $(n_1, \ldots, n_m)$ be the marginal frequency distribution of the multiplicities. Then

$$n_1 + n_2 + \cdots + n_m = n ,$$

and

$$n_1 + 2n_2 + \cdots + mn_m = m .$$

Knowledge of the frequencies $(n_1, \ldots, n_m)$ represents a reduction of the information in the multiplicity sequence that ignores the identities of the units. The marginal frequency distribution of the K-category variable x is the only information left when neither similarities nor identities between units can be observed.

Consider a network walk sample sequence $u = (u_1, \ldots, u_m)$ obtained by a time-homogeneous, irreducible, and aperiodic Markov chain with transition probabilities $P_{ij} = y_{ij}/a_i$ for $j \in A_i$ and initial probabilities equal to the stationary probabilities $p_j = y_j/(y_1 + \cdots + y_N)$ for $j \in U$. The following reasoning applies asymptotically for large sample sequences even if the initial unit is not selected according to the stationary distribution. Data consist of the values of a vertex variable x and the degrees for

each unit in the sample sequence u, but the units need not be identified. The population mean of x is $\mu = T/N$ where $T = x_1 + \cdots + x_N$. The mean of x in the sample sequence is

$$\sum_{i \in u}(x_i/m) = \sum_{i \in U}(m_i x_i/m) ,$$

where m_i is the unknown number of units equal to i in u. Since

$$\mathrm{E}\ m_i = mp_i = my_i/(y_1 + \cdots + y_N) ,$$

the sample mean is biased as an estimator of the population mean μ unless all degrees are equal. Now

$$T' = \sum_{i \in u}(x_i/mp_i) = \sum_{i \in U}(m_i x_i/mp_i) ,$$

has expected value T, and

$$N' = \sum_{i \in u}(1/mp_i) = \sum_{i \in U}(m_i/mp_i) ,$$

has expected value N. Moreover,

$$\mu' = T'/N' = \sum_{i \in u}(x_i/y_i) \Big/ \sum_{i \in u}(1/y_i) ,$$

has asymptotically expected value μ for large values on m. The quantities T' and N' are not available from data since they contain the unknown total $D = y_1 + \cdots + y_N$ of all degrees. However, this total cancels in the ratio $\mu' = T'/N'$ so μ' is an asymptotically unbiased estimator of μ. We note that the estimator μ' is a weighted mean of the values of x in the sample sequence u with weights that are inversely proportional to the degrees.

The results obtained can be applied with $x_i = y_i$ to get an estimator of the average degree D/N in the population. Its estimator is

$$D'/N' = m \Big/ \sum_{i \in u}(1/y_i) ,$$

which is the harmonic mean of the degrees in the sample sequence. The results can also be applied with $x_i = I(i \in U_1)$, i. e. with x as an indicator variable for a specific subset U_1 of U of unknown size N_1. Then $T = N_1$ and the relative size of U_1 has an asymptotically unbiased estimator

$$N_1'/N' = \sum_{i \in u}(x_i/y_i) \Big/ \sum_{i \in u}(1/y_i) ,$$

which is the ratio between the sum of inverted degrees of units in the sample sequence that belong to U_1 and the sum of inverted degrees of all units in the sample sequence. The average degree in U_1 is given by

$$D_1/N_1 = \left[\sum_{i \in U} x_i y_i\right] \Big/ \left[\sum_{i \in U} x_i\right] ,$$

where $x_i = I(i \in U_1)$, and it can be estimated by

$$D_1'/N_1' = \left[\sum_{i \in u} x_i\right] \Big/ \left[\sum_{i \in u}(x_i/y_i)\right] ,$$

which is the harmonic mean of the degrees of the units in the sample sequence that belong to U_1. We note that the corresponding arithmetic mean is larger than or equal to the harmonic mean and therefore over-estimates the (arithmetic) population mean. This is because we over-sample units with large degrees. The bias is reduced by taking the harmonic sample mean.

If the units in the sample sequence are identifiable, it is possible to observe the multiplicities and the set of distinct sample units $s(u)$. Should the inclusion probabilities $\pi_i = P(i \in s(u))$ be known, this would allow the population total T to be estimated by the unbiased estimator

$$T'' = \sum_{i \in s(u)} (x_i/\pi_i) .$$

Similarly, N would have an unbiased estimator

$$N'' = \sum_{i \in s(u)} (1/\pi_i) ,$$

and μ an asymptotically unbiased estimator

$$\mu'' = \left[\sum_{i \in s(u)} (x_i/\pi_i)\right] \Big/ \left[\sum_{i \in s(u)} (1/\pi_i)\right] .$$

The inclusion probabilities $\pi_i = P(i \in s(u))$ are approximately $1 - \exp(-mp_i)$ and according to our assumptions about the Markov chain, the stationary probabilities are $p_j = y_j/D$ where $D = \sum y_j$ is the degree total. Now the average degree D/N was shown above to be estimated by the harmonic mean of the degrees in the sample sequence. Altogether, this allows us to set up the estimator N'' as a function of N, and by replacing N by N'' we might have a possibility to solve iteratively for N'' in the equations

$$N'' = \sum_{i \in s(u)} [1 - \exp 1/(-my_i/D')] ,$$
$$D' = mN'' \Big/ \sum_{i \in u}(1/y_i) .$$

Possibilities of finding estimators iteratively, when the inclusion probabilities are functions of some of the parameters we want to estimate, deserve further study.

Methods for Estimating Hidden Populations

Hidden populations usually refer to populations of human individuals sharing a stigmatizing, illegal, risky, or embarrassing behavior that make them hard to find. There are no lists or frames to sample them from. Investigation of such populations might be of interest to society and necessary for determining appropriate resources to be allocated for social support and other types of social action. Standard survey sampling methods that require population frames cannot be used to study homeless people, drug users, sexual workers, individuals exposed to criminal or medical risks, etc. Various alternative approaches are discussed in [61].

Convenience samples of such hidden populations might be obtained by social field workers or by the police. The information gained from such sources could indicate the need for further study and data collection to obtain reliable inference about the hidden population. Members of the hidden population that are approached by social field workers might not be unwilling to co-operate and give away information about other members of the hidden population. Several practical methods described in the literature use various kinds of payments and other benefits to individuals in the hidden population that help the investigator to reach other members of the population for interview. When respondents are willing to reveal which other members they know of, the investigator can select the waves of a snowball sample. Respondents might be more willing to co-operate if they are asked to select at random one or two of the other members they know of. Thus, respondent driven random walk or small wave snowball sampling can be applied to the population network.

In cases when members of hidden populations are not willing to reveal themselves, it might still be possible to get information about them by applying network sampling. Assume for instance that the hidden population of interest is part of a larger population that can be sampled conventionally. A respondent sampled from the larger population might without revealing names or identities be able and willing to give some information about how many members of the hidden subpopulation she knows about. Such local network information about the hidden subpopulation can sometimes be used to infer information about the entire hidden subpopulation. As an illustrative example, consider the following setup considered in [41].

The hidden population U_1 consists of an unknown number N_1 of individuals. In order to estimate N_1, a larger population U of N individuals is defined so that it comprises U_1 as a subset; U_1 is embedded in U in such a way that another subset U_2 of U can be assumed to be equally common in U_1 and in U. The subset U_2 should consist of individuals having some easily identified characteristic. Neither the size N of U nor the size N_2 of U_2 need to be known. Select a sample S from U with positive inclusion probabilities π_i for all i in U. Each respondent is asked whether she belongs to U_2 and whether she has some acquaintances in U_2. She is not asked whether she belongs to U_1 but only whether she has some acquaintances in U_1. Let x_1 and x_2 be indicator variables of the sets U_1 and U_2, i. e. $x_{jk} = I(j \in U_k)$ for $j = 1, \ldots, N$ and $k = 1, 2$. Let further $y = (y_{ij})$ be the symmetric adjacency matrix of the acquaintance graph with degree y_i equal to the unknown number of acquaintances of individual i in U. This number is usually quite large and the respondent need not be asked to estimate it but rather to concentrate on giving reliable estimates of her numbers of acquaintances that belong to U_1 and U_2, say

$$a_{ik} = \sum_{j \in U} x_{jk} y_{ij} ,$$

for $i = 1, \ldots, N$ and $k = 1, 2$. If the independence assumptions are met for x_1, x_2, and y, then it could be expected under a randomization model that proportionality is valid according to

$$a_{ik}/y_i = N_k/N \quad \text{for} \quad k = 1, 2 \quad \text{so that} \quad a_{i1}/a_{i2} = N_1/N_2 .$$

The sum of the numbers a_{ik} for $i \in U$ equals the degree sum in U_k, i. e.

$$\sum_{i \in U} a_{ik} = \sum_{j \in U} x_{jk} y_j = D_k .$$

The standard Horvitz–Thompson estimator of D_k is

$$D'_k = \sum_{i \in S} (a_{ik}/\pi_i) ,$$

for $k = 1, 2$. Similarly Horvitz–Thompson estimators of N and N_2 are given by

$$N' = \sum_{i \in S} (1/\pi_i) \quad \text{and} \quad N'_2 = \sum_{i \in S} (x_{i2}/\pi_i) .$$

From the proportionality assumption we get that the average degrees are equal in U, U_1, and U_2:

$$D/N = \sum_{i \in U} y_i/N = D_k/N_k ,$$

for $k = 1, 2$. Hence, N_1 can be estimated by

$$N'_1 = D'_1 N'_2 / D'_2 .$$

By summing $a_{i1}/a_{i2} = N_1/N_2$ over i in U_2 and over i in U we get

$$\sum_{i\in U}(a_{i1}x_{i2}/a_{i2}) = N_1 \quad \text{and} \quad \sum_{i\in U}(a_{i1}/a_{i2}) = NN_1/N_2\,,$$

which leads to two alternative estimators of N_1, namely

$$N_1'' = \sum_{i\in S}(a_{i1}x_{i2}/a_{i2}\pi_i)$$

$$\text{and} \quad N_1''' = (N_2'/N')\sum_{i\in S}(a_{i1}/a_{i2}\pi_i)\,.$$

Here the sums are understood to be over sample units $i \in S$ with acquaintances in U_2. Should $a_{i2} = 0$ we expect $y_i = 0$ and $a_{i1} = 0$. If the three estimators N_1', N_1'', and N_1''' turn out to give very different estimates, the assumptions underlying the calculations are doubtful and another approach is needed.

This example illustrates that properties of hidden groups that are hard to access can still be estimated if the hidden group is embedded in a convenient network. While snowball sampling and random walk sampling use the network to successively extend an initial sample, the hidden group example uses randomization assumptions about stable expected average network degrees as a possible bridge between different subgroup sizes. If the network assumptions needed for the randomization procedure are not valid, then a natural alternative approach would be to combine the probability design of the sampling with a probability model of the network.

An entirely different kind of hidden population are animal populations. In wildlife surveys the technique of catch-recapture sampling is an attempt to handle the lack of population frames. Bird counting, for instance, relies on marking the birds caught in a first sample by identifying rings before they are released again. After some time a new sample is taken, and the number of marked birds in the first sample is divided by the proportion of marked birds in the second sample to get an estimator of the size of the bird population. This idea can be elaborated on by specifying more carefully various assumptions about how the population changes with time and how the risk of being caught depends on different environmental factors that can be controlled for. Catch-recapture techniques have also been discussed as a tool in criminology surveys to estimate the number of offenders of specific types of crimes in different regions and in different periods of time. Such surveys cannot be carried out without a thorough model specification as a complement to the sampling design. In Sect. "Bipartite Network Sampling Designs" an illustration is given of utilizing network models and network sampling methods in a criminological setting.

Area sampling is another technique that can be used for counting animal populations and other populations distributed over geographical areas. A population region is partitioned into a number of area sites. Neighboring sites can be considered as adjacent vertices in a graph. The size of the area population is a vertex variable, and the total population size is a vertex variable total that can be estimated by a snowball sample of area sites with counts of their area populations.

Probabilistic Network Models

Survey sampling in finite populations distinguishes between design-based and model-based inference. Design-based inference uses probabilistic sampling designs but considers population data as fixed, while model-based inference also makes probabilistic assumptions about population data. Essential properties of population data might be summarized in a few parameters in an appropriate model and thereby allow a convenient and simplified description of the population. The graph employed in a network sampling design can be exposed to various unknown influences and uncontrolled effects that make it natural to try to capture its essential features in a random graph model. Random graph models can also be used to describe the outcome when a graph is sampled from a population of graphs, or when an observed graph is generated by some complex process, for instance, being an instant realization of a network changing with time. Random graphs useful for statistical data analysis need to be sufficiently rich in parameters so that meaningful fits can be made. However, too many degrees of freedom in the model make it vulnerable to accidental fits with no explanatory power.

A simple Bernoulli graph on $U = \{1, \ldots, N\}$ with a symmetric adjacency matrix $y = (y_{ij})$ has $N(N-1)/2$ independent edge indicators y_{ij} that are Bernoulli(p)-distributed for $1 \le i < j \le N$. The only parameter p represents edge density. A block model variant of the Bernoulli graph has a K-category vertex variable x and assumes y_{ij} to be independent and Bernoulli distributed with a parameter $p(x_i, x_j) = p(x_j, x_i)$ for $1 \le i < j \le N$. Now there are $(K+1)K/2$ parameters representing edge densities within and between different blocks of vertices. Depending on whether block affiliation is known, can be observed, or is not observable, the possibilities to estimate the parameters are very different. Statistical inference for stochastic block models is considered in many papers. An extensive account with many references can be found in [67].

For a directed graph on U with adjacency matrix y there are $N(N-1)/2$ dyads (y_{ij}, y_{ji}) for $1 \le i < j \le N$. The simple model with IID dyads has four parameters

$p_{00}, p_{01}, p_{10}, p_{11}$ that are probabilities $p_{kl} = P(y_{ij} = k, y_{ji} = l)$, and the probabilities sum to 1 leaving 3 degrees of freedom. More generally, with independent dyads and probabilities $P_{ijkl} = P(y_{ij} = k, y_{ji} = l)$, there are $2N(N-1)$ parameters subject to $N(N-1)/2$ restrictions $\sum_k \sum_l P_{ijkl} = 1$, which leaves $3N(N-1)/2$ degrees of freedom. Assuming a block structure with a K-category vertex variable x and $P_{ijkl} = p_{kl}(x_i, x_j) = p_{lk}(x_j, x_i)$ there are 2 degrees of freedom within each category of size larger than one and 3 degrees of freedom between any pair of distinct categories, in total $2K + 3K(K-1)/2 = K(3K+1)/2$ degrees of freedom. In particular, $x_i = i$ for $i = 1, \ldots, N$ implies that each category consists of only one vertex, and there are 3 degrees of freedom for each pair of vertices, i. e. $3N(N-1)/2$ degrees of freedom in total in accordance with the previous result. Thus, a categorization of 20 vertices based on two binary vertex variables reduces the degrees of freedom from 570 to 26, and a simple dichotomy reduces the degrees of freedom to 7, so block models represent a substantial simplification of dyad independence models. A practical approach might be to start with a rather large number of tentative vertex variables and apply cluster analysis to their dyad distributions in order to find how many of these distributions need to be distinguished. This technique is applied in [31,32].

The block model approach can be considered even if the category affiliation is uncertain and described by independent vertex variables x_i with a probability distribution $P(x_i = k) = p_k$ for $k = 1, \ldots, K$. Let N_k be the number of vertices in category k and N_{kl} the number of pairs of vertices in categories k and l, i. e. $N_{kl} = N_k N_l$ for $k \neq l$ and $N_{kk} = N_k(N_k - 1)/2$. The log-likelihood function when vertex variables and graph adjacencies are observed is given by

$$\log L = \sum_{k=1,\ldots,K} N_k \log p_k + \sum_{1 \le k \le l \le K} \sum_{m=0,1} \sum_{n=0,1} N_{klmn} \log p_{mn}(k, l) ,$$

where

$$N_{klmn} = \sum_{1 \le i < j \le N} \sum I(x_i = k, x_j = l, y_{ij} = m, y_{ji} = n) .$$

This model has $K - 1$ degrees of freedom in addition to the previous $K(3K+1)/2$, so in total $3(K+1)K/2 - 1$ degrees of freedom. This model is treated for instance in [45,58,69].

Consider now the model with independent dyads and probabilities $P_{ijkl} = P(y_{ij} = k, y_{ji} = l)$. A convenient way to re-parametrize this model is to introduce odds and an odds ratio. Let a_{ij} be the odds of y_{ij} when $y_{ji} = 0$, and let b_{ij} be the odds of y_{ij} when $y_{ji} = 1$:

$$a_{ij} = P_{ij10}/P_{ij00}, b_{ij} = P_{ij11}/P_{ij01} .$$

The ratio between these odds is

$$c_{ij} = b_{ij}/a_{ij} = P_{ij11}P_{ij00}/P_{ij01}P_{ij10} .$$

The corresponding odds of y_{ji} are

$$a_{ji} = P_{ji10}/P_{ji00} = P_{ij01}/P_{ij00} ,$$
$$b_{ji} = P_{ji11}/P_{ji01} = P_{ij11}/P_{ij10} ,$$

and the odds ratio is the same

$$c_{ji} = b_{ji}/a_{ji} = c_{ij} .$$

The three parameters a_{ij}, a_{ji}, c_{ij} are arbitrary non-negative numbers, and together with the normalizing constant

$$P_{ij00} = 1/(1 + a_{ij} + a_{ji} + a_{ij}a_{ji}c_{ij}) ,$$

we retain the other probabilities as

$$P_{ij10} = a_{ij}P_{ij00} , \quad P_{ij01} = a_{ji}P_{ij00} ,$$
$$P_{ij11} = a_{ij}a_{ji}c_{ij}P_{ij00} .$$

Finally we put

$$\lambda_{ij} = -\log P_{ij00} , \quad \alpha_{ij} = -\log a_{ij} , \quad \gamma_{ij} = -\log c_{ij} ,$$

where $\lambda_{ij} = \lambda_{ji} > 0$ and $\gamma_{ij} = \gamma_{ji}$. Using the three parameters $\alpha_{ij}, \alpha_{ji}, \gamma_{ij}$ and the normalizing constant λ_{ij} determined by them, it is possible to express the likelihood function L according to

$$-\log L = \sum_{1 \leqslant i < j \leq N} \sum (\lambda_{ij} + \alpha_{ij}y_{ij} + \alpha_{ji}y_{ji} + \gamma_{ij}y_{ij}y_{ji}) .$$

There are $3N(N-1)/2$ degrees of freedom represented by $\alpha_{ij}, \alpha_{ji}, \gamma_{ij}$ for $1 \le i < j \le N$. By introducing an out-degree effect α_i, an in-degree effect β_i, and a mutuality effect γ_i for each vertex, we assume that

$$\alpha_{ij} = \alpha_i + \beta_j , \gamma_{ij} = \gamma_i + \gamma_j .$$

Since $\alpha_{ij} = \alpha + \beta + \alpha_i - \alpha + \beta_j - \beta$ for arbitrary α and β, α_i and β_j are identifiable only up to a translation

$\alpha = -\beta$. If we introduce an arc density effect δ and put $\alpha_{ij} = \delta + \alpha_i + \beta_j$ we can impose the restrictions $\sum \alpha_i = 0$ and $\sum \beta_j = 0$. In this way we have reduced the degrees of freedom from $3N(N-1)/2$ to $3N-1$, and the likelihood L is a log-linear function of the parameters given by

$$-\log L = \sum_{1 \le i < j \le N} \sum \lambda_{ij} + \delta \sum_{i \ne j} \sum y_{ij} + \sum_i \alpha_i \sum_j y_{ij} + \sum_j \beta_j \sum_i y_{ij} + \sum_i \gamma_i \sum_j y_{ij} y_{ji} ,$$

with out-degree, in-degree, and number of mutual arcs at each vertex as the sufficient statistics. A further reduction is obtained by assuming all mutuality effects equal, $\gamma_i = \gamma$, which leads to a model with $2N$ degrees of freedom. This is the original Holland–Leinhardt model for directed graphs introduced and investigated in [36]. Various extensions to valued graphs and block models are possible, and algorithms for model fitting and parameter estimation are described and illustrated in the reference book [67] and in several references given there.

Random graph models with probability distributions of exponential type have been suggested with more elaborate network statistics. Some examples of investigations of such models are [47,52,57,68].

Instead of specifying which statistics are likely to be essential for the application considered and use an exponential type model with these statistics as sufficient statistics, Frank and Strauss in [30] derived the sufficient statistics from basic assumptions about the dependence structure in the adjacency matrix. They specified a Markov dependence structure for the adjacency matrix y and proved that it implied a log-linear likelihood with certain sufficient statistics. The Markov assumption is that y_{ij} and y_{kl} are independent whenever i, j, k, l are all distinct, but dependence is possible between any arc indicators with some common vertex. The sufficient statistics are then given by star and triangle statistics. Adding a homogeneity assumption saying that isomorphic structures have the same probability, the sufficient statistics are given by star counts and triangle counts. Here stars and triangles should be understood in a broad sense so that they can consist of dyads of any type. The fundamental tool for this result in [30] is a theorem of Besag given in [2] together with a dependence graph specifying the dependencies among arc indicators in y. This dependence graph is in fact the so-called line-graph of the graph given by y. The maximal cliques of the dependence graph are the minimal sufficient statistics of y. Further exploration of these ideas can be found in the thesis [50] and the article [51].

Bipartite Network Sampling Designs

A bipartite network consists of vertices of two kinds and vertex and edge variables defined on them. If U is a set of N vertices of the first kind, and V is a set of M vertices of the second kind, different vertex variables could be defined on U and V. Edge variables are defined between U and V. As a simple example, take U as a set of households, V as a set of individuals, and define a binary edge variable y_{ij} that indicates whether individual j belongs to household i for $i \in U$ and $j \in V$. Each individual belongs to exactly one household. Assume that U is a target population, and that we are interested in some variable x defined on U. Only a sample S from V can be obtained. For each individual sampled, the value of x is reported that belongs to the individual's household. Thus

$$\sum_{i \in U} x_i y_{ij} ,$$

is reported by $j \in S$. If the sampling design has inclusion probabilities

$$\pi_j = P(j \in S) ,$$

the household total $T = \sum_{i \in U} x_i$ has a design unbiased estimator

$$T' = \sum_{j \in S} \sum_{i \in U} x_i y_{ij} / \pi_j = \sum_{i \in U} x_i \sum_{j \in S} y_{ij} / \pi_j .$$

If more than one individual from the same household are sampled, the estimator contains this household value x with multiplicity. If household identities are checked and no repeated household values should be used, another design unbiased estimator is available as

$$T'' = \sum_{i \in B(S)} x_i / P(i \in B(S)) ,$$

where $B(S)$ is the set of distinct households reported by individuals in S. The probability $P(i \in B(S))$ is equal to $P(S \cap A_i \ne \emptyset)$, where A_i is the set of individuals belonging to household i. This is the probability that none of those individuals are included in the sample S. If we introduce multiplicities m_i for the number of individuals in S belonging to household i, there is an alternative design unbiased estimator

$$T''' = \sum_{i \in U} x_i m_i / \mathrm{E} m_i ,$$

that requires the expected value of the multiplicity. Comparisons between these estimators and further results on

bipartite network sampling can be found in Sirken's pioneering paper on multiplicity sampling [56].

The possibility to use bipartite networks in survey sampling is not confined to cases with sampling units that are different from the units in the target population. As an illustration of a more complex combination of sampling from both U and V, we consider in the next section a model approach to a simultaneous estimation of crime rates and offender activity that relies on a bipartite network model.

A Bipartite Network Model for Crime Participation

Consider crimes of a particular kind in a specified geographical region during a specified period of time, say assaults in Canada during year 2000. Let M denote the unknown total number of incidents and m the number known to the police. Thus, $M - m$ is an unknown so-called "dark figure" of incidents, and m is the number of identified incidents. The identified incidents involve r incidents reported to the police and $m - r$ incidents identified by other means. Let m_i be the number of identified incidents having i identified offenders, and let r_i be the number of reported incidents having i identified offenders for $i = 0, 1, \ldots$ Incidents with at least one identified offender are said to be cleared by the police. There are unidentified offenders for all the unidentified incidents, and there are unidentified co-offenders for the identified incidents. The unknown numbers of incidents M_{ij} with i actual offenders and j identified offenders for $j = 0, 1, \ldots, i$ and $i = 1, 2, \ldots$ involve many dark numbers in addition to $M - m$. The unknown numbers of incidents

$$M_i = \sum_{j=0,\ldots,i} M_{ij} ,$$

with i actual offenders for $i = 1, 2, \ldots$ provide a partition of M that specifies the distribution of incidents according to size, i. e. their number of co-offenders.

The same offender can be involved in several offences. Let N denote the unknown total number of distinct individuals participating in the M offences. The known number n of distinct offenders identified in the cleared incidents imply that there is a dark number of $N - n$ non-identified offenders. The number of incidents in which an offender participates is called the activity of the offender. We need to distinguish between identified activity and actual activity. Let N_{ij} be the number of offenders with activity i and identified activity j for $j = 0, 1, \ldots, i$ and $i = 1, 2, \ldots$. The partition of N into the numbers

$$N_i = \sum_{j=0,\ldots,i} N_{ij} ,$$

of offenders of activity i for $i = 1, 2, \ldots$ specifies the distribution of offenders according to activity. This distribution involves further dark numbers that cause its discrepancy from the known numbers n_j of identified offenders participating in j identified offences for $j = 1, 2, \ldots$

In order to estimate the distribution of incidents according to size and the distribution of offenders according to activity, a bipartite network model can be defined with the set U of M incidents as vertices of the first kind and the set V of N offenders as vertices of the second kind. Criminal participation can be specified by an edge from $i \in U$ to $j \in V$ if incident i involves offender j. Let A_i be the subset of V specifying the co-offenders of incident i, and let B_j be the subset of U specifying the re-offences of offender j.

Now we can consider the set of identified incidents as a sample S from U, and the set of identified co-offenders of identified incident $i \in S$ as a sample S_i from A_i. These samples have sampling designs depending on factors governing how incidents are reported to the police or identified by other means and how the co-offenders are identified. In a situation like this when sampling is not determined by the investigator it is natural to incorporate the sampling design in a probabilistic network model.

An approach taken in [28] specifies the incidents in U as generated by a homogeneous Poisson process with parameter λ representing the expected number of incidents per time unit. The sample S of identified incidents is a Bernoulli sample with inclusion probability α, which means that identified incidents occur according to a homogeneous Poisson process with parameter $\alpha\lambda$. The parameter α measures the proportion of identified incidents. At least for crimes with a victim, α can be estimated by comparing incidence reporting frequencies and incidence identifications by other means with incidence frequencies estimated from victimization studies.

The set V of offenders is assumed to consist of a fixed number N of individuals that have a positive probability of offending. Independently for each incident $i \in U$, a set A_i of co-offenders in that incident is selected from V according to some common probability sampling design with inclusion probabilities $\pi_j > 0$ for $j \in V$. The sampling design might involve unknown parameters reflecting what is known about structure and behavior in the population V. For each identified incident $i \in S$, a set S_i of identified co-offenders is modeled as an independent Bernoulli sample of A_i with parameter β governing the chance that the police can identify an offender. An optimistic and a pessimistic view of this chance distinguishes between whether or not the identification of one co-offender in an incident should imply the identification of all co-offenders in that incident.

With this model, it turns out to be possible to estimate simultaneously both the distribution of incidents according to size and the distribution of offenders according to activity. The basic results are the following. From the distribution of identified size among identified incidents, it is possible to estimate the parameter β and the distribution of true size among all incidents. Furthermore, if θ is the average identified activity among identified offenders and $\gamma = n/N$ is the unknown proportion of identified offenders, then identified activity has a Poisson($\gamma\theta$) distribution that is truncated at zero, and true activity has a Poisson($\gamma\theta/\alpha\beta$) distribution. This implies that

$$\gamma = 1 - \exp(-\theta\gamma) ,$$

from which γ can be solved, and N estimated by n/γ.

An interesting feature of this bipartite network model is that it makes explicit assumptions about both incident size and offender activity, and these assumptions lead to estimators of all the dark numbers related to unidentified incidents, unidentified offenders, and unidentified co-offenders in identified incidents. In particular, the estimator of the hidden population size N of offenders provides a special insight into how this estimator can be modified if not all parts of the distribution of identified activity are reliable. This also makes interesting connections with the literature that derives or suggests such modifications by using other approaches.

Future Directions

It is likely that the development of more sophisticated network models for specific applications will appear as improvements of Markov models and other network models belonging to the exponential family. Such development is also expected to contribute to more efficient algorithms for computer intensive estimation and goodness-of-fit testing of network models.

The developments of design-based inference methods that are appropriate for sampling of sub-networks from a finite population network are especially challenging for few-waves snowball sampling and for network walk sampling. The iterative estimation methods discussed in Sects. "Estimation Based on Snowball Samples" and "Estimation Based on Network Walk Samples" are examples of problems that deserve to be investigated further. More important is perhaps the need to develop efficient algorithms for Rao-Blackwellization of various initially obtained estimators and to gain more information about minimal sufficient statistics in various snowball sampling designs.

Bibliography

1. Basu D, Ghosh JK (1967) Sufficient statistics in sampling from a finite universe. Proc 36th Session Int Stat Inst, pp 850–859
2. Besag JE (1974) Spatial interaction and the statistical analysis of lattice systems. J Royal Stat Soc B 36:192–236
3. Bloemena AR (1964) Sampling from a graph. Math Centre Tracts, Amsterdam
4. Bollobas B (2001) Random graphs. Cambridge University Press, Cambridge
5. Brandes U, Erlebach T (eds) (2005) Network analysis. Springer, Berlin
6. Capobianco M (1970) Statistical inference in finite populations having structure. Trans New York Acad Sci 32:401–413
7. Capobianco M, Frank O (1982) Comparison of statistical graph-size estimators. J Stat Plan Inference 6:87–97
8. Carrington PJ, Scott J, Wasserman S (eds) (2005) Models and methods in social network analysis. Cambridge University Press, Cambridge
9. Cassel CM, Särndal CE, Wretman J (1993) Foundations of inference in survey sampling. Krieger Publ Comp, Malabar
10. Corander J (2000) On Bayesian graphical model determination. Dissertation, Stockholm University
11. Diestel R (2005) Graph theory. Springer, Berlin
12. Erdös P, Renyi A (1959) On random graphs I. Publ Math Debrecen 6:290–297
13. Erdös P, Renyi A (1960) On the evolution of random graphs. Publ Math Inst Hungar Acad Sci 5:17–61
14. Frank O (1969) Structure inference and stochastic graphs. Swedish Research Institute of National Defence, Stockholm, FOA-Reports 3(6):1–10
15. Frank O (1970) Sampling from overlapping subpopulations. Metrika 16:32–42
16. Frank O (1971) Statistical inference in graphs. Dissertation, Stockholm University
17. Frank O (1977) Estimation of graph totals. Scand J Stat 4:81–89
18. Frank O (1977) A note on Bernoulli sampling in graphs and Horvitz–Thompson estimation. Scand J Stat 4:178–180
19. Frank O (1977) Survey sampling in graphs. J Stat Plan Inference 1:235–264
20. Frank O (1978) Sampling and estimation in large social networks. Soc Netw 1:91–101
21. Frank O (1978) Estimation of the number of connected components in a graph by using a sampled subgraph. Scand J Stat 5:177–188
22. Frank O (1979) Estimating a graph from triad counts. J Stat Comput Simul 9:31–46
23. Frank O (1979) Estimation of population totals by use of snowball samples. In: Holland P, Leinhardt S (eds) Perspectives on Social Network Research. Academic Press, New York, pp 319–347
24. Frank O (1980) Sampling and inference in a population graph. Int Stat Rev 48:33–41
25. Frank O (1980) Estimation of the number of vertices of different degrees in a graph. J Stat Plan Inference 4:45–50
26. Frank O (1981) A survey of statistical methods for graph analysis. In: Leinhardt S (ed) Sociological methodology. Jossey-Bass, San Francisco, pp 110–155
27. Frank O (1987) Random sampling and social networks – a survey of various approaches. Math Inform Sci Hum 26(104):19–33

28. Frank O, Carrington PC (2007) Estimation of offending and co-offending using available data with model support. J Math Sociol 31:1–46
29. Frank O, Snijders T (1994) Estimating the size of hidden populations using snowball sampling. J Off Stat 10:53–67
30. Frank O, Strauss D (1986) Markov graphs. J Am Stat Assoc 81:832–842
31. Frank O, Komanska H, Widaman K (1985) Cluster analysis of dyad distributions in networks. J Classif 2:219–238
32. Frank O, Hallinan M, Nowicki K (1985) Clustering of dyad distributions as a tool in network modelling. J Math Sociol 11:47–64
33. Goodman LA (1961) Snowball sampling. Ann Math Stat 32:148–170
34. Granovetter M (1976) Network sampling: Some first steps. Am J Sociol 81:1287–1303
35. Hagberg J (2003) On degree variance in random graphs. Dissertation, Stockholm University
36. Holland PW, Leinhardt S (1981) An exponential family of probability distributions for directed graphs. J Am Stat Assoc 76:33–65
37. Janson S, Luczak T, Rucinski A (2000) Random graphs. Wiley, New York
38. Jansson I (1997) On statistical modeling of social networks. Dissertation, Stockholm University
39. Kaipio J, Somersalo E (2005) Statistical and computational inverse problems. Springer, New York
40. Karlberg M (1997) Triad count estimation and transitivity testing in graphs and digraphs. Dissertation, Stockholm University
41. Killworth PD, McCarty C, Bernard HR, Johnsen EC, Domini J, Shelley GA (2002) Two interpretations of reports of knowledge of subpopulation sizes. Soc Netw 25:141–160
42. Koskinen J (2004) Essays on Bayesian inference for social networks. Dissertation, Stockholm University
43. Morgan DL, Rytina S (1977) Comment on "Network sampling: Some first steps, by M Granovetter". Am J Sociol 83:722–727
44. Mukhopadhyay P (2001) Topics in survey sampling. Springer, New York
45. Nowicki K, Snijders TAB (2001) Estimation and prediction for stochastic blockstructures. J Am Stat Assoc 96:1077–1087
46. Palmer EM (1985) Graphical evolution. Wiley, New York
47. Pattison P, Wasserman S (1998) Logit models and logistic regressions for social networks. II: Multivariate relations. Br J Math Stat Psychol 52:169–193
48. Proctor CH (1979) Graph sampling compared to conventional sampling. In: Holland PW, Leinhardt S (eds) Perspective on social network research. Academic Press, New York, pp 301–318
49. Proctor CH, Loomis CP (1951) Analysis of sociometric data. In: Jahoda M, Deutsch M, Cook SW (eds) Research methods in social relations. Dreiden Press, New York, pp 561–586
50. Robins GL (1998) Personal attitudes in inter-personal contexts: Statistical models for individual characteristics and social relationships. Dissertation, University of Melbourne
51. Robins GL, Pattison P (2005) Interdependencies and social processes: Dependence graphs and generalized dependence structures. In: Carrington PJ, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, Cambridge, pp 192–214
52. Robins GL, Pattison P, Wasserman S (1999) Logit models and logistic regressions for social networks. III: Valued relations. Psychometrika 64:371–394
53. Salganik MJ, Heckathorn DD (2004) Sampling and estimation in hidden populations using respondent-driven sampling. Sociol Methodol 34:193–239
54. Särndal CE, Swensson B, Wretman J (1992) Model assisted survey sampling. Springer, New York
55. Schweinberger M (2007) Statistical methods for studying the evolution of networks and behaviour. Dissertation, University of Groningen
56. Sirken MG (1970) Household surveys with multiplicity. J Am Stat Assoc 63:257–266
57. Snijders TAB (2002) Markov chain Monte Carlo estimation of exponential random graph models. J Soc Struct 3(2)
58. Snijders TAB, Nowicki K (1997) Estimation and prediction of stochastic blockmodels for graphs with latent block structure. J Classif 14:75–100
59. Spreen M (1999) Sampling personal network structures: Statistical inference in Ego-graphs. Dissertation, University of Groningen
60. Stephan FF (1969) Three extensions of sample survey technique. In: Johnson NL, Smith H Jr (eds) New developments in survey sampling. Wiley, New York
61. Sudman S, Kalton G (1986) New developments in the sampling of special populations. Ann Rev Sociol 12:401–429
62. Tallberg C (2003) Bayesian and other statistical approaches for analyzing network block-structures. Dissertation, Stockholm University
63. Thompson ME (1997) Theory of sample surveys. Chapman & Hall, London
64. Thompson SK (2006) Adaptive web sampling. Biometrics 62(4):1224–1234
65. Thompson SK (2006) Targeted random walk designs. Surv Methodol 32:11–24
66. Thompson S, Frank O (2000) Model-Based Estimation with Link-Tracing Sampling Designs. Surv Methodol 26:87–98
67. Wasserman S, Faust K (1994) Social network analysis. Cambridge University Press, Cambridge
68. Wasserman S, Pattison P (1996) Logit models and logistic regressions for social networks. I: An introduction to Markov random graphs and p*. Psychometrika 60:401–426
69. Wellman B, Frank O, Espinoza V, Lundquist S, Wilson C (1991) Integrating individual, relational and structural analysis. Soc Netw 13:223–249
70. Wills RS (2006) Google's page rank: The math behind the search engine. Math Intell 28(4):6–11

Social Network Analysis, Graph Theoretical Approaches to

Wouter de Nooy
University of Amsterdam, Amsterdam, The Netherlands

Article Outline

Glossary
Definition of the Subject
Introduction

Cohesion
Centrality and Brokerage
Prestige and Ranking
Analyzing Complexity in Social Networks
Future Directions
Bibliography

Glossary

Adjacent Two vertices are adjacent if they are connected by a line.

Arc An arc is a directed line. Formally, an arc is an ordered pair of vertices.

Attribute An attribute is a characteristic of a vertex measured independently of the network.

Bipartite network See: Two-mode network.

Clique A clique is a maximal complete subnetwork containing three vertices or more.

Complete A (sub)network is complete if it has maximum density: All possible lines occur.

Component A (weak) component is a maximal (weakly) connected subnetwork.

Degree The degree of a vertex is the number of lines incident with it.

Density Density of a simple network is the number of lines, expressed as a proportion of the maximum possible number of lines.

Digraph A digraph or directed graph is a graph containing one or more arcs.

Distance The distance from vertex u to vertex v is the length of the geodesic from u to v.

Edge An edge is an undirected line. Formally, an edge is an unordered pair of vertices.

Ego-network The ego-network of a vertex contains this vertex, its neighbors and all lines among the selected vertices.

Geodesic A geodesic is the shortest path between two vertices.

Graph A graph is a set of vertices and a set of lines between pairs of vertices.

Incident A line is defined by its two endpoints, which are the two vertices that are incident with the line.

Indegree The indegree of a vertex is the number of arcs it receives.

Line A line is a tie between two vertices in a network: Either an arc or an edge.

Loop A loop is a line that connects a vertex to itself.

Neighbor A vertex that is adjacent to another vertex is its neighbor.

Network A network consists of a graph and additional information on the vertices or the lines of the graph.

One-mode network In a one-mode network, each vertex can be related to each other vertex.

Outdegree The outdegree of a vertex is the number of arcs it sends.

Path A path is a semipath with the additional condition that none of its lines is an arc of which the end vertex is the arc's tail.

Reachable We say that a vertex is reachable from another vertex if there is a path from the latter to the former.

Semicycle A semicycle is a closed semipath ending at the vertex at which it starts.

Semipath A path is a closed sequence of lines such that the end vertex of one line is the starting vertex of the next line and no vertex in between the first and last vertex of the sequence occurs more than once.

Signed graph A signed graph is a graph in which each line carries either a positive or a negative sign.

Simple graph A simple undirected graph contains neither multiple edges nor loops. A simple directed graph does not contain multiple arcs.

Star-network A star-network is a network in which one vertex is connected to all other vertices but these vertices are not connected among themselves.

Strong component A strong component is a maximal subnetwork in which each pair of vertices is connected by a path.

Strongly connected A (sub)network is strongly connected if each pair of vertices is connected by a path.

Structural property A structural property is a characteristic (value) of a vertex that is a result of network analysis.

Triad A triad is a (sub)network consisting of three vertices.

Two-mode network In a two-mode network, vertices are divided into two sets and vertices can only be related to vertices in the other set.

Undirected graph An undirected graph does not contain arcs: All of its lines are edges.

Vertex (vertices) A vertex (singular of vertices) is the smallest unit in a network.

Weakly connected A (sub)network is weakly connected if each pair of vertices is connected by a semipath.

Definition of the Subject

Social network analysis (SNA) focuses on the structure of ties within a set of social entities or actors, e. g., persons, groups, organizations, and nations, or the products of human activity or cognition such as semantic concepts, web sites, and so on. In a graph theoretical approach, a social network is conceptualized as a graph, that is, a set of ver-

tices (or nodes, units, points) representing social actors and a set of lines representing one or more social relations among them.

A network, however, is more than a graph because it contains additional information on the vertices and lines. Characteristics of the social actors, for instance a person's sex, age, or income, are represented by discrete or continuous attributes of the vertices in the network, and the intensity, frequency, valence, or type of social relation are represented by line weight or value, line sign, or line type. Formally (see pp. 94–95, 127–128 in [1]), a network $\mathbf{N}$ can be defined as $\mathbf{N} = (U, L, F_U, F_L)$ containing a graph $\mathbf{G} = (U, L)$, which is an ordered pair of a unit or vertex set U and a line set L, extended with a function F_U specifying a property of the units ($f\colon U \to X$) and a function F_L specifying a property of the lines ($f\colon L \to Y$). The set of lines L may be regarded as the union of a set of undirected edges E and a set of directed arcs $A(L = E \cup A)$. Each element e of E is an unordered pair of units u and v (vertices) from U, that is, $e(u\colon v)$, and each element a of A is an ordered pair of units u and v (vertices) from U, that is, $a(u\colon v)$.

The application of graph theory to social relations can be traced back to at least the 1940s (see pp. 69–72 in [2]) when the mathematician R. Duncan Luce and the engineer Albert Perry teamed up with the social psychologist Leon Festinger [3] and when the mathematician Frank Harary started his collaboration with Leon Festinger and afterwards with Dorwin Cartwright [4]. They extended pioneering work in SNA that had been done notably in sociometry [5,6] and anthropology [7,8,9]. In the 1960s, advances in graph theoretical approaches to SNA such as the contributions by Øystein Ore [10], Claude Flament [11], Frank Harary [12], and innovative applications such as Everett M. Rogers' work on the diffusion of innovations [13], prepared the ground for the rise of SNA in both the USA [14] and Europe [15] as a new set of methods or a new methodology [16] in the 1970s.

Introduction

The conceptualization of social systems as graphs and networks offered the opportunity for systematic investigation and theorizing of the structure of ties among social actors beyond the pair. Whereas classical sociology tended to make a quantum leap from the individual and the pair to the triple, group, or society [17], graph theory offered the tools to formally describe and visualize social structure consisting of three and more actors. This led to a new awareness of social structure as a system of ties that is both the product of human action and the context and condition for human action. Because this point of view is very relevant to the issue of complexity in social networks, it is briefly presented in the next paragraph.

The prevalent action theory in SNA conceptualizes collective behavior as the socially 'orchestrated' behavior of individuals or other actors. Actors adjust their behavior and attitudes, opinions, and beliefs to the behavior (etc.) of other members of the social system in which they participate. The system itself is not supposed to behave but it constrains actor behavior: It is the social context within which actors operate. As a network of ties, the system defines to whom an actor is exposed. The immediate contacts – the neighbors in graph theory – of an actor are usually most important to its behavior, but indirect contacts such as the neighbors' neighbors may be taken into account as well. In other words, an actor's local context or ego-network is likely to affect its behavior. At the same time, however, by ending ties or creating new ones, the individual changes both local network structure and overall network structure, that is, the system. Thus, individual action changes the local context for its neighbors' (neighbors' etc.) action. Complexity arises in the interplay between individual behavior and the system both as the overall structure of the network of social ties and as the local context for each actor within the system. To the actors, the change of network structure is not necessarily predictable, so the interplay between individual action and network structure may offer surprising results.

Let us turn to an example now, which is one of the earliest applications of graph theory to social networks. This example nicely illustrates both the transition from a focus on the tie within a single pair to the study of group structure in the social sciences and the interplay between local action and overall network structure. In 1946, the psychologist Fritz Heider formulated the theory of psychological balance [18], which stated among other things that a person (P) feels uncomfortable when he or she disagrees with his or her friend (O) on a particular topic (X). Person P is hypothesized to be stressed and to try to change this situation either by adapting its opinion on topic X, so it matches O's opinion, or by changing its opinion on O, regarding him or her no longer as a friend. Figure 1 represents both a situation of imbalance and a situation of balance as a signed digraph. The circles and arrows represent the vertices and arcs of the graph and the valence of the opinions or affections are shown both by the labels and the style of the arrows: Solid arcs show positive opinions or sentiments, dotted arcs show negative opinions or sentiments.

In 1956, the mathematician Frank Harary and the psychologist Dorwin Cartwright realized that psychological

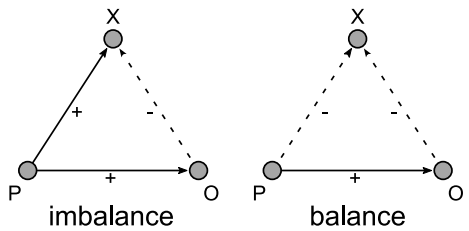

Social Network Analysis, Graph Theoretical Approaches to, Figure 1
The principle of imbalance and balance conceptualized as signed digraphs

balance in this triad (three vertices and the lines among them) may be conceptualized as a specific pattern of arcs in a signed graph, viz., in a balanced triad, the $P - O - X$ semicycle (a closed semipath) always contains zero or an even number of negative arcs, whereas an imbalanced triad is characterized by a semicycle with an uneven number of negative arcs [4]. As a next step, replacing the topic by a third person and perceptions of liking or disliking (by the focal person P) by expressed liking or disliking as ties, they easily generalized this idea to a network of arbitrary size. They proved that a signed network is balanced if and only if all semicycles contain no or an even number of negative arcs.

In addition, they proved that a balanced network either contains one set of vertices with just positive arcs among them, or two sets of vertices with all positive arcs within the sets and all negative arcs between the sets, which is a polarized network. In 1967, this result was generalized to polarization among three or more groups by James A. Davis, who showed that a network can be partitioned into an arbitrary number of subsets of vertices such that all positive ties are within the subsets and all negative arcs are among the subsets if the network does not contain semicycles with exactly one negative line [19]. Figure 2 shows an example from Samuel F. Sampson's [20] study of a network of sentiments among novices in a monastery. It depicts the situation at the fourth measurement wave, which was highly polarized at that time. Vertex color indicates whether the novice had previously attended one particular seminary (black: Yes, white: No).

In this case, a mathematical relation was established between a behavioral hypothesis at the level of the individual, viz., adjusting ones affect relations such that ones situation is balanced, and overall network structure, viz., polarization. Individual behavior proved to have an unexpected outcome at the level of overall network structure and graph theoretical concepts, in this case semicycles, provided the link. In many cases, however, the link between individual behavior and overall network structure has not been established formally and is sometimes even hard to predict intuitively.

This entry of the encyclopedia aims to present an overview of graph theoretical approaches to SNA, highlighting the complex relations between individual action and overall network structure. It intends to explain why current developments focus on local structure rather than overall network structure to unravel the complexity of social networks. For each of the main topics in SNA (cohesion, brokerage, and prestige), behavioral hypotheses are presented stating why actors create, maintain or dissolve ties. The typical local structure of ties produced by this behavior is sketched in combination with graph theoretical indices for measuring it, and finally the expected consequences to the overall structure of the social network are discussed in combination with the graph theoretical measures developed for measuring them. Note that this approach more or less reverses the historical development of SNA, which focused on overall network structure first and gradually became more interested in the behavior of actors that created, maintained, or changed network structure.

Cohesion

One of the first intuitions in SNA concerns the tendency of human beings to form cohesive subgroups. This is a classical topic in the social sciences, see, for instance, George C. Homans' book *The Human Group* [21], and it was central to the sociometry tradition [22]. But where do cohesive subgroups come from and what do they do?

The first and most general behavioral hypothesis merely states that similar people tend to interact more easily and people who interact tend to become or perceive themselves as more similar provided that the interaction is characterized as positive, friendly, and so on. In SNA, this tendency is mainly known as homophily, a concept coined by Paul F. Lazarsfeld and Robert K. Merton [23,24], but it is known under other names in several scientific disciplines, e. g., the phenomenon of attribution [25] and affect control [26,27] in social-psychology, assortative or selective mixing in epidemiology and ecology ([28], p. 2 in [29]), and assortative mating in genetics with efforts at statistical modeling at least as early as 1985 [30].

It is important to note that there are two sides to this behavioral hypothesis, a selection effect, that is, the impact of similarities on the ties that are created, sustained, or broken [31], and an influence effect [32], which hypothesizes that perceived or actual similarities such as the socially constructed identities or opinions [33] result from

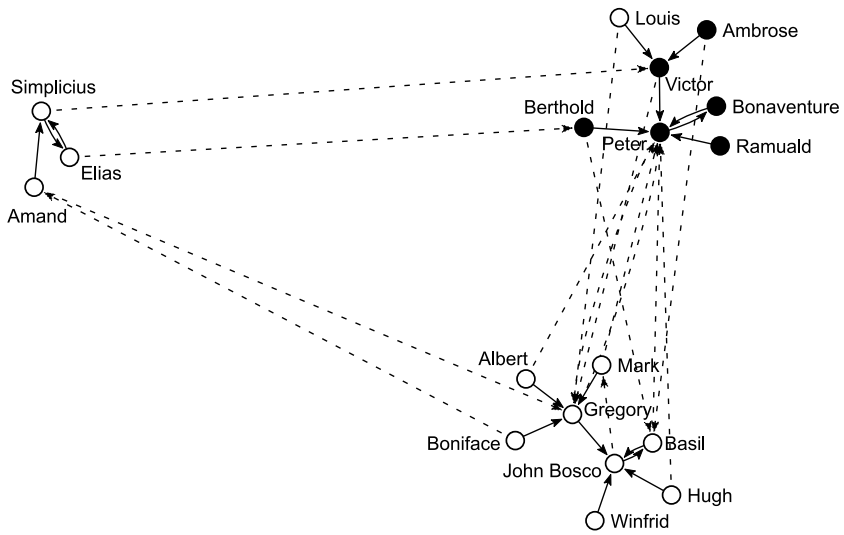

Social Network Analysis, Graph Theoretical Approaches to, Figure 2
Almost perfect polarization in the network of sentiments among novices

the ties among actors. According to these hypotheses, people who are directly linked are or become more similar because of their interaction, so they become more likely to engage in ties and maintain ties among them. Thus, social groups form and persist as tightly linked sets of people that tend to share social and psychological characteristics, producing solidarity.

If we concentrate on the graph theoretical aspects of this behavioral hypothesis, that is, the structure of ties, and ignore the (dis)similarities among the actors for the moment, we find several characteristics of local structures that measure cohesive subgroup formation. At the level of a pair of actors, reciprocity of ties in directed networks signals subgroup formation: Both actors are hypothesized to choose each other when they are similar. At the level of the triple, transitivity results from tendencies toward cohesion. If actor u establishes a tie with actor v because they are similar, and actor v establishes a tie to actor w for the same reason, actors u and w are also similar, so actor u is expected to establish a tie with w as well, creating a so-called transitive triad (Fig. 3). Stated differently, the path from u via v to w is closed by an arc from u to w. In general, the closure of paths or semipaths both in directed and undirected networks signals cohesive subgroup formation at the local level. Closure within an ego-network may be calculated as the percentage of all possible ties among a vertex' neighbors that are present, which is one of the definitions of the clustering coefficient (see pp. 32–33 in [34]) but the concept of closure can be extended beyond a vertex' immediate neighbors, e. g., the number of semi-

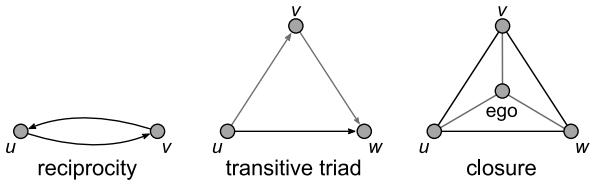

Social Network Analysis, Graph Theoretical Approaches to, Figure 3
Reciprocity, transitivity, and closure in a directed ego-network

cycles of length 4 or larger in which a vertex is involved, e. g., balanced semicycles in signed networks.

If we include vertex attributes in our measures of cohesive subgroup formation, we can calculate homophily quite simply as the probability or ratio of ties between vertices that share a particular characteristic to ties between vertices that do not. Extending this idea to the ego-network, the homogeneity of actors involved in an ego-network may be taken as a measure of tendencies toward homophily. For qualitative attributes of the actors, Blau's index of variability or heterogeneity [35] can be used $(1 - \sum p_i^2)$ where p is the proportion of group members in a particular category and i is the number of different categories), which is conceptually related to the Herfindahl–Hirschman Index in economics measuring the extent of monopoly within an industry. It is interesting to note that Blau's theory hypothesizes that heterogeneity rather than homogeneity of actors within a group enhances the operation and efficiency of the group. If improving group

efficiency is the aim of actors, we would have to use a behavioral hypothesis that is quite the opposite of the homophily hypothesis.

Tendencies toward cohesive subgroup formation at the level of actor behavior are most likely to produce sets of densely connected vertices in the overall network and, if we add attributes of the actors, the vertices within sets tend to have similar characteristics. In the most extreme case, the sets are disconnected, so the network consists of several weak components, that is, maximal weakly connected subnetworks, or they are connected only by ties with a negative social meaning, as in the example of polarization presented in the Introduction.

In the history of SNA, the concept of relatively densely connected subnetworks has yielded a large number of graph theoretical ways for identifying cohesive subgroups at the level of overall network structure. Limiting the discussion to one-mode networks, that is networks in which there can be a tie between any pair of vertices (for two-mode or bipartite networks, ► Social Network Analysis, Two-Mode Concepts in), Wasserman and Faust (see pp. 251–252 in [36]) distinguish 4 approaches to defining cohesive subgroups.

In the first and strictest approach, a cohesive subgroup is defined as a set of vertices in which all vertices are adjacent, that is, directly linked, to one another. In other words, cohesive subgroups are maximal complete subgraphs, which are called cliques [3].

The second approach is based on the notion of reachability and closeness of members within a subgroup. Members of a subgroup must be reachable in the sense that there are paths between them, i. e., a sequence of lines such that the end vertex of one line is the starting vertex of the next line (following the direction of the lines if they are arcs) such that no vertex in between the first and last vertex occurs more than once. In addition, the shorter the geodesics (shortest paths) between them, the closer the vertices are in a graph theoretical sense, so the more cohesive the subgroup is supposed to be.

The criterion of reachability does not necessarily yield very dense subgroups. In sparse networks, any maximal connected subgraph (strong component) may represent a cohesive subgroup: There is a path between each pair of vertices within a component. Increasing the required number of independent paths between any pair of vertices within a cohesive subgroup yields slightly denser subgroups, e. g., requiring at least two independent paths produces bi-components, which may be generalized to k-components for higher minimum numbers of independent (vertex-disjoint) paths between all vertices within a subgroup [37].

Focusing on graph-theoretical distance between vertices usually yields denser subgroups. One may, for instance, set a maximum n to the distance between any two vertices within a subgroup, which is the concept of an n-clique [38,39]. Adding the restriction that the diameter of an n-clique is n or less, one obtains n-clans [38,40]. Alternatively, one may define a cohesive subgroup as a maximal subgraph of diameter n, which is called an n-club [40].

The third approach focuses on the minimum number, strength, or multiplicity of ties among subgroup members. Subgraphs that are maximal with respect to the minimum number of neighbors within the subgraph are called k-cores [41] and a maximal subgraph with respect to the maximum number of vertices in the subgraph that are not adjacent are known as k-plex [42]. In a similar vein, restrictions can be imposed on the minimum strength or multiplicity of ties among members of a cohesive subgroup, generalizing Seidman's concept of a k-core to a valued core, which is called an m-core (see pp. 115–116 in [43]) or m-slice (see pp. 109–110 in[44]): Maximal connected subgraphs considering only lines with minimum value (or multiplicity) m.

In the fourth approach, cohesive subgroups are based on the relative frequency of ties among subgroup members in comparison to non-members: Cohesive subgroups are relatively dense sections within the network, that is, relative to the sections outside (and between) subgroups. An *LS* set [45] is a maximal subgraph such that any of its subsets has more ties to its complement within the *LS* set than to vertices outside the *LS* set. Borgatti, Everett and Shirey [46] generalized this idea to the concept of the lambda set, which requires the number of edge-disjoint paths between any pair of vertices within the lambda set to be larger than between any vertex within and any vertex outside the lambda set. A probabilistic version of plus-clusters in signed networks, discussed in the Introduction, can also be subsumed under this approach as it requires relatively many positive lines within cohesive subgroups and relatively many negative lines among cohesive subgroups [47]. Finally, clustering techniques and some types of blockmodels (see pp. 741–742 in [14]) also detect clusters of vertices that have relatively many ties within clusters and few among clusters. These models offer an alternative way for finding cohesive subgroups (see pp. 133–246 in [1]), ► Positional Analysis and Blockmodeling.

The large number of alternative measures for cohesive subgroups attests to the fact that behavioral tendencies at the actor level do not play out into nicely structured overall networks in a standard way. Especially the density of the social relation under investigation has an impact on

the extent and ways in which cohesive subgroups can be found in the overall structure of the network. As a consequence, measures of network cohesion such as the clustering coefficient (averaged over all vertices in the network) can be quite uninformative about overall network structure: Loosely knit cohesive subgroups which are clearly identified by some of the techniques presented above, yield low clustering coefficients in a sparse network. Cohesion in a network is better summarized by calculating the percentage of vertices that are part of identified subgroups, the number and sizes of cohesive subgroups, and so on [48].

In addition to homophily, there is a second behavioral hypothesis related to cohesion in SNA. This hypothesis is based on the idea that social action is embedded in networks [49,50,51]. Named after the sociologist Georg Simmel, Simmelian ties are ties that are embedded in other ties, e. g., business ties are embedded in family ties, or in complete triads and cliques. They are hypothesized to enforce group norms and enhance trust, hence pressure people into the same behavior because the two actors involved in a tie share common neighbors who supervise their behavior.

Just as with the homophily hypothesis, the embeddedness hypothesis predicts that tightly connected actors will be more similar in their behavior and attitudes. In addition, it predicts that embedded ties are more stable and new ties are more likely to be established when they are embedded in existing cliques or existing ties. Closure again is an important indicator of tendencies to establish and maintain embedded ties but so is the multiplicity of relations: The extent to which a tie on one social relation duplicates a tie on another social relation. At the level of overall network structure, we should expect relatively dense sections, especially cliques, and in a multirelational network, that is, a network containing ties on two or more social relations, we should find that the same subsets of vertices are clustered on each relation. Graph-theoretical measures of the latter are rare. The stochastic blockmodeling technique developed by Krysztof Nowicki and Tom A.B. Snijders [52] is an example. See ► Social Networks, Algebraic Models for.

If data on vertex attributes are available, especially if they concern public behavior, that is, behavior that is easily noticed by third parties such as publicly expressed opinions and statements, Simmelian ties are hypothesized to produce a special effect. Involvement in different groups (cliques) then exposes actors to possibly conflicting sets of norms and loyalties, which may urge them to cut their ties with some or all of these groups [53]. In this case, actors are hypothesized to withdraw from stressful relations, so they discontinue ties that incorporate them into cliques (with a preference for cutting a minimum number of ties) or they discontinue ties such that they are no longer connected to actors voicing different opinions or norms. At the macro level, this would produce disconnected sets of cliques instead of overlapping cliques.

Centrality and Brokerage

The notion of centrality in social networks has a long history in SNA. It is attributed to Alex Bavelas [54]. In discussions of centrality, network ties are usually regarded as channels for the exchange of information, goods, services, and so on. Being central in this exchange system has always been hypothesized to be related to influence and satisfaction. Centralization, as a characteristic of a network, has been linked to the efficiency of a network as an exchange system. More centralized groups, for example, have often been shown to be more efficient.

Linton C. Freeman [55] argued that the approaches to centrality are based on three ideas about what being central means: (1) being active within the network, that is, maintaining many ties, (2) being efficient or independent of go-betweens by having short distances to other vertices in the network, and (3) being an important go-between, that is, being part of many paths between other vertices in the network. Although alternative classifications and approaches exist, for instance, Noah E. Friedkin's alternative classification [56] and the formal graph theoretical approach to centrality by Stephen P. Borgatti and Martin G. Everett [57], Freeman's classification is used here, adding concepts of brokerage that have been developed elsewhere in SNA.

Activity

Being active or prominent in the network means that an actor has many ties, hence access to many sources of information (etc.). As a consequence, this actor is more attractive as a neighbor for other actors, which translates to the behavioral hypothesis that actors have a preference for ties with vertices that already have many ties. The degree of a vertex (the number of lines incident with a vertex), then, is the relevant graph theoretical measure of local structure, which is also known as degree centrality. Note that the concept of centrality expresses a structural property of a vertex.

Centralization is the corresponding structural property of a network and it is defined as the variation in the centrality scores of the vertices in the network because this variation shows the extent to which there is a center (very central vertices) and a periphery (vertices with very low centrality scores). The star and ring networks are defined

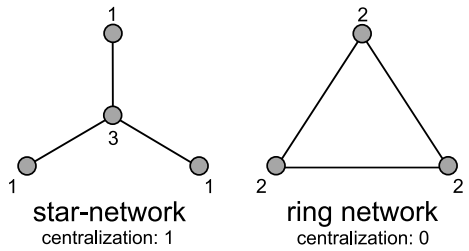

Social Network Analysis, Graph Theoretical Approaches to, Figure 4
Centrality and centralization in a star network and a ring network

as respectively the most and least centralized networks and they are known to exhibit the highest and lowest variation in centrality scores in simple networks, that is, networks without multiple lines (and loops in the case of a directed network). See Fig. 4 for an illustration, showing a star and a ring network labeling the vertices with their degree centrality scores.

From research on the power law in networks [58,59], discussed in another entry of this encyclopedia, it is known that preferential attachment to degree creates networks with a peculiar degree distribution, including many vertices with low or modest degree and few vertices with high degree. According to the definition of centralization in network analysis, this implies large variation in degree centrality scores, hence high degree centralization. In other words, the behavioral hypothesis of preference for high degree actors produces centralized overall network structure.

Note, however, that the way in which the 'power law networks' are assembled – growing from an initial seed without context – is hardly ever applicable to social networks, which usually have no discernable starting point (networks originate from networks) and are always constrained by the historical context. It remains to be seen whether power law distributions in empirical social networks are created by preferential attachment. Some results indicate that even though the degree distributions of cross-sectional snapshots of a large social network follow the power law, there is hardly any continuity in degree centrality of vertices over time, which does not suggest that the actors are driven by preferential attachment [60].

Efficiency and Weak Ties

The second approach to centrality focuses on graph theoretical distances between vertices. The central idea here is that actors try to improve access to information and efficient spreading of information by minimizing the number of go-betweens needed to reach or be reached by all other actors in the network. The behavioral hypothesis states that a vertex prefers to link to actors that give access to parts of the network that are presently remote to this vertex and that can only be reached through some or many go-betweens that may withhold or distort information. A minimum number of go-betweens yields maximum independence and maximum efficiency in the exchange network.

Graph theoretical measures of local structure focus on graph theoretical distance, that is, the minimum length of paths between vertices because path length equals the number of go-betweens in the network plus one. Linton Freeman's closeness centrality [55] is a straightforward implementation of this idea because it merely normalizes the average graph theoretical distance between a vertex and all other reachable vertices in the network. In addition, paths can be weighted by the centrality of the vertices on them, which is done by Phillip Bonacich's eigenvector centrality [61,62]. Finally, the difference between incoming paths and outgoing paths may be added, which is implemented in the version of closeness centrality developed by Thomas W. Valente and Robert K. Foreman [63].

Again, the normalized variation of closeness centrality scores of the vertices in the network yields the appropriate measure of centralization of overall network structure. It is not known yet, however, whether and how tendencies to reduce distances to other vertices at the level of individual actors in the network play out into the closeness centralization of the overall network. On the one hand, if the network contains some vertices with high closeness centrality, they offer rather short paths toward many vertices so they should attract a lot of new ties. This would enhance their centrality and possibly the variation in centrality scores although a general rise in closeness centrality scores of all vertices may also decrease the variation. On the other hand, if the network has low centralization, it is more likely that vertices connect directly to remote parts, reducing path lengths among remote parts, which would not raise the variation in closeness centrality scores and yield or sustain low closeness centralization of overall network structure.

The strength of weak ties argument proposed by Mark Granovetter [64,65] may be regarded as a special application of the notion of efficiency. In his research on finding a job, Granovetter noticed that relatively superficial ties, ties with infrequent contact, give access to new information because they are more likely to link you to someone with whom you are not linked directly or at a short distance. Strong or intense ties tend to be situated within cohesive subgroups, so they are more likely to offer redun-

dant information already received through other ties. Granovetter is only interested in the effects of having weak ties, but if we turn his idea into a behavioral hypothesis, it suggests a preference to relate to distant vertices that are neither connected to yourself nor to your neighbors (or your neighbors' neighbors, and so on).

A tendency to connect to the most remote parts of the network means that actors tend to establish links to vertices at a large or maximum distance in the network. As in the case of maximizing closeness centrality, it is not clear whether this leads to identifiable patterns in overall network structure. It is quite obvious, however, that this tendency acts as a counterforce against tendencies toward cohesion as densely connected subnetworks: Actors are hypothesized to span gaps rather than to close local configurations. Different hypotheses must be developed for strong ties, which are hypothesized to contribute to subgroup formation, and weak ties linking remote parts. The weak ties will probably increase the number of links between dense parts of the network, increasing the k-connectivity (minimum number of node-independent paths between any two vertices) of the network. If so, we should expect high k-connectivity of the network if the weak ties hypothesis is true [37].

Control and Structural Holes

The third approach to centrality focuses on control over flows within the network: The more you are in between other vertices in the network, the more they depend on you to pass on information, the more you are able to control exchange within the network and profit from your control. Using this type of control is called brokerage.

The notion of being in between other vertices has a straightforward translation to graph theory as being part of a path between two other vertices. Limiting paths to the shortest paths between vertices both Linton C. Freeman's [55] betweenness centrality and Jac M. Anthonisse's [66] rush of a vertex are based on the proportion of all geodesics between other vertices that include this vertex. This measure has been extended to handle directed ties [67,68]. Information centrality [69,70] takes into account all paths between vertices, not just the geodesics and flow betweenness [71] or entropy [72] also consider the values of lines. See Stephen P. Borgatti's [73] classification for more details.

Betweenness centralization as the normalized variation of betweenness centrality of the vertices in the network offers a measure of centralization at the network level and so do center-periphery blockmodels (see pp. 741–742 in [14]), e. g., in the world trade system, central countries are able to profit from the lack of trade among countries in the periphery [74,75,76,77]. The link between strategies for maximizing betweenness centrality by actors and the overall structure of the resulting network is unclear. It is not to be expected that control behavior will produce a highly centralized structure because no actor will be satisfied with low betweenness centrality and as a consequence high betweenness centrality for some vertices is unlikely. Therefore, the variation in betweenness centrality scores will not be high.

While betweenness centrality situates an actor with respect to all other actors in the network, Ronald S. Burt [78,79] proposed a local variant, focusing on control within the ego-network of an actor. His behavioral hypothesis rests on the *tertius gaudens* principle: The benefits that accrue to an actor that is in between two actors that are not directly linked because of the opportunity to broker information between them or, in a more malicious variant, to divide and conquer.

This hypothesis translates quite easily into graph theoretical structure. The absence of a tie between two neighbors of an actor is called a structural hole (Fig. 5a). The behavioral hypothesis states that actors try to increase the structural holes that they can exploit. At the same time, however, they try to minimize the structural holes through which they can be exploited. This means, among other things, that an actor will not end a tie to one of its neighbors if the two neighbors are directly linked (see Fig. 5b): That would create an opportunity to broker at the expense of the actor. In this situation, the actor is constrained in its opportunities to change ties.

Structural holes and constraint are the flip sides of the same coin. A tie with low constraint indicates that the tie is involved in (many) incomplete triads (such as Fig. 5a), so there are (many) structural holes offering the actor options for brokering. High constraint on a tie means that it is part of (many) complete triads (as in Fig. 5b), so there are few or no possibilities for brokerage. Because the presence or absence of ties among an actor's neighbors is key

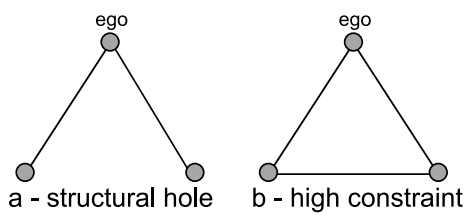

Social Network Analysis, Graph Theoretical Approaches to, Figure 5
A structural hole (a) and a triad with high constraint (b)

to the argument, network analysts have also used the density of the ego-network without the ego as a proxy of constraint: The higher the density, the higher the constraint on the ego. Alternatively, betweenness centrality for ego-networks [80,81] can be used.

If the structural holes hypothesis governs actors' behavior, what overall network structure should we expect to find? A strong tendency toward brokerage at the micro level is not likely to produce a centralized network because each actor would try to maximize the number of structural holes around itself, which would yield a bipartite graph in the extreme case (no ties among any vertex' neighbors), or it would minimize its constraint by ending all ties to neighbors that have contacts outside ego's immediate neighborhood, which would produce a highly clustered network consisting of isolated cliques or isolated vertices in the extreme case.

In summary, the relation between local action and overall network structure is simple and clear only in the case of preferential attachment to high-degree vertices. Tendencies to maximize centrality that looks beyond the immediate neighbors such as closeness and betweenness centrality, do not necessarily yield centralized networks. Even for local structures, alternative hypotheses for actor's behavior are available that are unlikely to produce centralized networks, e. g., a preference to avoid constraint. The interplay between local action and overall structure is quite complex.

Prestige and Ranking

The preceding sections have not distinguished between directed and undirected networks. For prestige and ranking, however, the direction of ties is crucial because asymmetry in networks is assumed to be linked to social prestige [82]. The general idea here is that social inequalities are reflected and possibly created by asymmetric ties, e. g., everyone invites the most popular boy or girl in class but s/he doesn't return each invitation. Of course, the nature of the social relation determines the direction of choices; ties like "reports to" or "pays respect to" point toward higher levels in a hierarchy while "beating up" points in the opposite direction.

A central behavioral hypothesis concerns the popularity or attractiveness of actors. Actors tend to (want to) relate to actors with attractive structural properties or attributes, so attributes related to power or social status increase the probability that an actor will be chosen. From a constructivist point of view, however, being chosen often is also interpreted as a sign of importance and prestige, so receiving many choices (ties) increases the probability of receiving even more. In this way, networks may produce informal status hierarchies. The Matthew Effect, proposed by Robert K. Merton [83], comes to mind here: "For unto every one that hath shall be given, and he shall have abundance: But for him that hath not shall be taken away even that which he hath" (gospel of Matthew XXV, 29).

In graph theoretical terms, the structural attractiveness of an actor refers to the number of incoming arcs on vertices, which is simply the indegree of a vertex. This is called the popularity of a vertex and, of course, we must replace it by the vertex' outdegree if the relation is negative, e. g., submission, beating up, criticizing. If indirect choices must be taken into account as well, attractiveness is measured by proximity prestige [84], which is based on the average distance from all other vertices in the network – a directed variant of closeness centrality. Proximity prestige captures the idea that nominations or choices by actors who are themselves popular, contribute more to ones structural prestige. Bonacich's measure of power [62,85] adds the idea that power may also be derived from being connected to powerless actors rather than to other powerful people.

The popularity hypothesis is another example of preferential attachment to degree with the restriction that we focus either on indegree or outdegree. Therefore, the indegree (or outdegree) distribution of the overall network is expected to follow the power law and network structure will be characterized by high degree centralization.

Adding data on social attributes that make some actors more prestigious such as wealth, social class, beauty, and so on, we should expect preferential attachment to vertices that score high on these attributes. Note that vertex attributes play a slightly different role here than in the case of cohesion. Now we are concerned with attributes of (at least) ordinal level, expressing prestige that an actor possesses to a higher or lower degree. In the case of cohesion, we deal with nominal attributes, which merely express an identity. In contrast to homophily, the attributes of the actor who initiates the directed tie (the tail of the arc) does not matter here because the effect is solely related to structural characteristics or attributes of the actor at the receiving end of the tie (the arc's head).

There is a second, slightly different behavioral hypothesis relating to deference or submission rather than attractiveness. The idea is that actors mainly tend to create positive ties to other actors in their own status group or to actors in a higher status group – the people they are looking up to – to consolidate and improve their social position. Similarly, they tend to direct negative ties to actors in lower status groups. Note the difference with attractiveness: It is hypothesized that actors choose upwardly but

they need not prefer the most attractive (top) actors in the network as they are supposed to do according to the attractiveness hypothesis.

The main difference between attractiveness and deference is that the former only takes into account structural properties or attributes of the tie's receiver, whereas the characteristics of both the sender and receiver of the tie matter to the latter. This distinction has important consequences to the structure of the overall network. Whereas the attractiveness hypothesis yields centralized networks, the deference hypothesis yields layered networks. The layers consist of sets of vertices that are symmetrically linked, e. g., by reciprocal ties, while the ties between layers are asymmetric, all pointing in the same direction. This behavior may be both a consequence of a formal hierarchy, e. g., positions within an organization with formalized relations such as reports to, or actually show an informal social hierarchy, e. g., status differences between men and women in a particular social setting.

In simple directed networks, triads, that is, three vertices and the lines among them, are the key to measuring tendencies toward ranking at the local level. Translating the concepts of balance and clusterability (see Sect. "Introduction") from signed digraphs to unsigned digraphs, James A. Davis and Samuel Leinhardt [86,87] replaced positive ties by symmetric ties within cohesive subgroups and negative ties by no ties among subgroups. Thus, ties within a layer, either within clusters or among clusters, are symmetric. They assumed that asymmetric ties represent the ranking of clusters into a hierarchy, introducing the model of ranked clusters. Moreover, they showed that a network with a perfectly fitting ranked clusters model contains only certain types of triads, whereas other types do not occur (Fig. 6).

The ranked clusters model requires arcs from each vertex to all vertices on higher ranks. This requirement is usually too strict for empirical social networks and it is relaxed in the transitivity model proposed by Paul W. Holland and Samuel Leinhardt [88], which requires that clusters of vertices on different ranks are either completely linked or not linked at all, yielding a partial order, by simply adding one type of triad to the set of allowed triads (Fig. 6). Later, Eugene Johnsen [89] proposed the model of hierarchical clusters to account for asymmetries within clusters. Note the nesting of the models for overall network structure, which is why the sets of permitted triads is extended for more general models. Finally, it was shown that the models developed for unsigned digraphs could also be detected in incomplete signed digraphs using types of semicycles [90].

The triads characterizing ranked structures serve as models for tie creation, maintaining, and breaking behavior of actors under the deference hypothesis. For instance, triad 120U (Fig. 6) predicts that a member of a cluster is likely to establish or maintain a tie to an actor at a higher rank if its neighbors within the cluster have such a tie. In the perfect case, there is a one-to-one relation between sets of occurring types of triads and the overall structure of the network. Therefore, triad census [91,92], which is the frequency distribution of the sixteen types of triads in a directed network, offers an indication of overall network structure. In the imperfect case, the triad census of a network may be compared to the frequency distributions of triad types in randomly generated networks to test the tendency toward ranking.

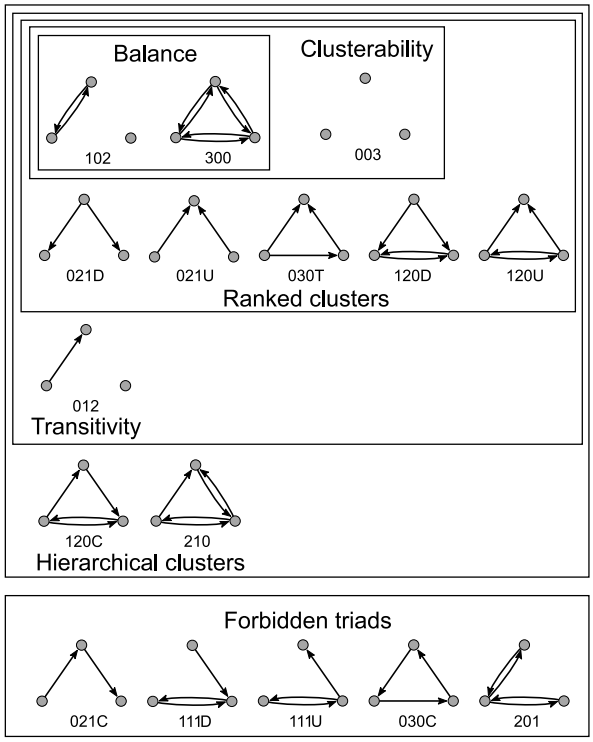

Social Network Analysis, Graph Theoretical Approaches to, Figure 6
Triad types and balance-theoretic models

The triad census does not show the composition of the clusters and ranks; it does not identify the vertices belonging to particular clusters and ranks. This can be done in several ways. Realizing that ranks should be connected asymmetrically in directed networks, strong components cannot include more than one rank because vertices within strong components are by definition mutually reachable. Ties between strong components, then, are asymmetric, so it is easy to establish the ranking among strong components. Strong components, however, do not

require a lot of symmetry in the ties; actually, no tie needs to be reciprocated. The symmetric-acyclic decomposition proposed by Patrick Doreian, Vladimir Batagelj and Anuška Ferligoj [93] does require at least some symmetric ties (mutual choices) within clusters because they define a symmetric cluster as a maximum subset of vertices that are directly or indirectly linked by symmetric ties. Generalized blockmodels [1] that are asymmetric with respect to off-diagonal blocks offer another way to identify hierarchical relations (see ► Positional Analysis and Blockmodeling).

Analyzing Complexity in Social Networks

The preceding sections presented behavioral hypotheses that have similar, different, or even opposite consequences for overall network structure. It is not plausible that one particular type of behavior dominates network formation. Therefore, it is not likely that overall structure of empirical social networks will display one particular form that can be hypothesized in advance or that behavioral tendencies can be adequately tested on particular characteristics of overall network structure. It has been shown, for example, that the degree distribution of a network does not reveal tendencies toward cohesive subgroup formation that are operative during network evolution [94].

Even if overall network structure displays certain characteristics, they may be produced by different types of behavior. Centralization in a social network, for instance, may arise from a tendency of actors to minimize paths to all other actors or from a preference for prestigious actors. Alternatively, it may be a by-product of cohesive subgroup formation: Actors that are marginal to the cohesive subgroups may directly or indirectly connect different subgroups, which gives them a central position with respect to betweenness. Furthermore, pronounced overall network patterns may occur only temporarily in empirical social networks when they create socially unstable situations. The polarization predicted by balance theory in Sampson's network of novices, used as an example in the Introduction, was only temporary. After this polarization and most likely due to it, many novices left the monastery. The network, so to speak, fell apart.

For these reasons, SNA increasingly focuses on local structure using overall network structure merely as a collection of (overlapping) local structures. Behavioral hypotheses translate much more directly to local structure, that is, to the ties of the actor and those of its neighbors (and possibly their neighbors), as we have seen in the preceding sections. Local structure is the part of the network that an actor can easily survey and actually change.

The latest developments in techniques for modeling network structure and evolution apply this actor-oriented approach either in statistical models, see ► Network Analysis, Longitudinal Methods of and ► Social Networks, Exponential Random Graph (p^*) Models for. The techniques test behavioral tendencies by relating the creation, maintenance, and ending of ties by individual actors to the local configuration of ties, to previous ties between the actor and the alter or to present ties on another social relation, and to characteristics of both the actor and the alter.

In principle, the actor-oriented approach is able to test all behavioral hypotheses presented in the preceding sections on characteristics of local structure in which the actors are embedded and properties of the actors themselves. If hypothesized local configurations appear more often than expected by chance, the underlying behavioral hypothesis is assumed to guide individual behavior at least to some extent. If the behavior in a set of actors or individual actor's behavior are in line with several behavioral hypotheses at the same time, the effect of each behavioral tendency can be separated. Thus, it is possible to link complex overall network structure to compound behavior of the actors in the network.

Future Directions

The techniques for analyzing local structure are in development. Models for the co-evolution of relations and quantitative attributes of vertices over time have just been introduced [95]. Not all behavioral hypotheses have been included yet, for instance, because they involve non-standard types of networks such as signed relations, and new ones are bound to be proposed. Incomplete data and external constraints on data collection or conceptual constraints on network structure such as two-mode networks (see ► Social Network Analysis, Two-Mode Concepts in) may limit the applicability of current models and spur the development of new ones.

If the actor-oriented approach is successful, will overall network structure become completely redundant? Will it only serve network exploration – looking for behavioral hypotheses rather than testing them – and for analyzing the consequences of network position on behavior, attitudes, or esteem, for instance, does the subgroup to which an actor belongs or its centrality correlate with its subsequent behavior or attitudes?

Let us return to the balance example, presented in the Introduction. The high degree of polarization among Sampson's novices was followed by the voluntary or forced exit of several novices. This suggests that people are able to survey overall network structure and draw conclusions

from it, rather than just react to their local network environment. A highly polarized social group does not seem to be the kind of situation that we like to be living in. An actor's perception of overall network structure may also affect its behavior.

In this respect, the balance example is interesting because the original theory by Heider referred to perceptions of affect relations and unit relations (having characteristics or possessing items) rather than objectively measured relations. The importance of perception to network analysis is stressed in a more general way by Harrison White, who argues that social ties are stories [96]. According to him, people are linked into social networks by the stories that they tell about their ties. Thus, we should expect people to react to the ties as they remember and tell them rather than to the ties as they are observed by the researcher or registered in, for instance, membership lists.

A similar argument was made by David Krackhardt when he reconsidered balance theory [97]. According to him, we should measure each actor's perception of network structure, for which he proposed the concept of Cognitive Social Structures, compare these perceptions, and use them to explain why actors behave as they do. In his approach, actors are assumed to be able and active in forming impressions of overall network structure.

If humans are capable of imagining overall network structure, then communication of these perceptions may also play a role in network formation, including the accidental and deliberate distortions or simplifications that are likely to happen. This brings us to the link between network structure and mental categories such as social classifications or culture in a more general sense as argued by Ronald L. Breiger [98,99]. If, for example, members of the network perceive and discuss cohesive subgroups, they assign names and meanings to social configurations. Thus, social meanings and identities are created in the process of establishing social ties and interpreting them. The duality of social structure on the one hand and the structure of symbolic categories on the other hand as proposed by John W. Mohr [100], may be the essential condition for cultural meanings that are social in the sense that they affect actors' sense of identity and behavior.

These discrete and qualitative rather than continuous and quantitative classifications are very likely to affect the network behavior of actors: They define the categories that are experienced as being similar or dissimilar in the case of homophily and group formation, or superior versus inferior in the case of prestige and ranking. As a consequence, it is to be expected that the present focus on local structure will be complemented by a focus on overall network structure, especially perceived and communicated network structure.

Bibliography

Primary Literature

1. Doreian P, Batagelj V, Ferligoj A (2005) Generalized blockmodeling. Cambridge University Press, Cambridge
2. Freeman LC (2004) The development of social network analysis. A study in the sociology of science. Empirical, Vancouver
3. Luce RD, Perry A (1949) A method of matrix analysis of group structure. Psychometrika 14:95–116
4. Cartwright D, Harary F (1956) Structural balance: A generalisation of Heider's theory. Psychol Rev 63:277–293
5. Murphy G (1937) Editorial foreword. Sociometry 1:5–7
6. Moreno JL (1937) Sociometry in relation to other social sciences. Sociometry 1:206–219
7. Warner WL, Lunt PS (1941) The social life of a modern community. Yale University Press, New Haven
8. Davis A, Gardner BB, Gardner MR (1946) Deep south: A social anthropological study of caste and class. University of Chicago, Chicago
9. Mitchell JC (1969) The concept and use of social networks. In: Mitchell JC (ed) Social networks in urban settings. Manchester University Press, Manchester
10. Ore Ø (1963) Graphs and their uses. Mathematical Association of America, Washington
11. Flament C (1963) Applications of graph theory to group structure. Prentice-Hall, Englewood Cliffs
12. Harary F, Norman RZ, Cartwright D (1965) Structural models: An introduction to the theory of directed graphs. Wiley, New York
13. Rogers EM (1962) Diffusion of innovations. Free Press, New York
14. White HC, Boorman SA, Breiger RL (1976) Social structure from multiple networks. I. Blockmodels of roles and positions. Am J Sociol 81:730–780
15. Helmers HM, Mokken RJ, Plijter RC, Stokman FN, Anthonisse JM (1975) Graven naar macht: Op zoek naar de kern van de Nederlandse economie. Van Gennep, Amsterdam
16. Wellman B (1988) Structural analysis: From method and metaphor to theory and substance. In: Wellman B, Berkowitz SD (eds) Social structures: A network approach. Cambridge University Press, Cambridge, pp 19–61
17. Simmel G (1902) The number of members as determining the sociological form of the group. I. Am J Sociol 8:1–46
18. Heider F (1946) Attitudes and cognitive organization. J Psychol 21:107–112
19. Davis JA (1967) Clustering and structural balance in graphs. Hum Relat 20:181–187
20. Sampson SF (1968) A novitiate in a period of change. An experimental and case study of social relationships. Cornell University, Ithaca, p 575
21. Homans GC (1950) The human group. Harcourt Brace, New York
22. Moreno JL (1953) Who shall survive?: Foundations of sociometry, group psychotherapy and sociodrama, 1st edn 1934. Beacon House, Beacon
23. Lazarsfeld PF, Merton RK (1954) Friendship as a social process: A substantive and methodological analysis. In: Berger M, Abel

T, Page CH (eds) Freedom and control in modern society. Van Nostrand, Princeton, pp 18–66
24. McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. Ann Rev Sociol 27:415–444
25. Festinger L (1962) A theory of cognitive dissonance, 1st edn 1957. Tavistock, London
26. Heise DR (1979) Understanding events: Affect and the construction of social action. Cambridge University Press, New York
27. Smith-Lovin L, Heise DR (1988) Analyzing social interaction: Advances in affect control theory. Gordon Breach Science, New York, pp 192
28. Morris M (1991) A log-linear modeling framework for selective mixing. Math Biosci 107:349–377
29. Newman MEJ (2003) Mixing patterns in networks. Phys Rev E 67
30. Tallis GM (1985) Transfer systems and covariance under assortative mating. Theor Appl Genet 70:497–504
31. Robins G, Elliott P, Pattison P (2001) Network models for social selection processes. Soc Netw 23:1–30
32. Robins G, Pattison P, Elliott P (2001) Network models for social influence processes. Psychometrika 66:161–189
33. Friedkin NE, Johnsen EC (1990) Social influence and opinions. J Math Sociol 15:193–205
34. Watts DJ (1999) Small worlds. The dynamics of networks between order and randomness. Princeton University Press, Princeton
35. Blau PM (1977) Inequality and heterogeneity: A primitive theory of social structure. Free Press, New York
36. Wasserman S, Faust K (1994) Social network analysis: Methods and applications. Cambridge University Press, Cambridge
37. White DR, Harary F (2001) The cohesiveness of blocks in social networks: Node connectivity and conditional density. Sociol Methodol 31:305–359
38. Alba RD (1973) A graph-theoretical definition of a sociometric clique. J Math Sociol 3:113–126
39. Luce RD (1950) Connectivity and generalized cliques in sociometric group structure. Psychometrika 15:169–190
40. Mokken RJ (1979) Cliques, clubs and clans. Qual Quant 13:161–173
41. Seidman SB (1983) Network structure and minimum degree. Soc Netw 5:269–287
42. Seidman SB, Foster BL (1978) A graph-theoretic generalization of the clique concept. J Math Sociol 6:139–154
43. Scott J (1991) Social network analysis: A handbook. Sage, London
44. de Nooy W, Mrvar A, Batagelj V (2005) Exploratory social network analysis with Pajek. Cambridge University Press, Cambridge
45. Seidman SB (1983) Internal cohesion of ls sets in graphs. Soc Netw 5:97–107
46. Borgatti SP, Everett MG, Shirey PR (1990) LS sets, lambda sets and other cohesive subsets. Soc Netw 12:337–357
47. Doreian P, Mrvar A (1996) A partitioning approach to structural balance. Soc Netw 18:149–168
48. White DR, Owen-Smith J, Moody J, Powell WW (2004) Networks, fields and organizations: Micro-dynamics, scale and cohesive embeddings. Comput Math Organ Theory 10:95–117
49. Friedkin NE (1984) Structural cohesion and equivalance explanations of social homogeneity. Sociol Methods Res 12:235–261
50. Granovetter M (1985) Economic action and social structure: The problem of embeddedness. Am J Sociol 91:481–510
51. Granovetter M (1992) Problems of explanation in economic sociology. In: Nohria N, Eccles RG (eds) Networks and organizations: Structure, form, and action. Harvard Business School Press, Boston, pp 25–56
52. Nowicki K, Snijders TAB (2001) Estimation and prediction for stochastic blockstructures. J Am Statist Assoc 96:1077–1087
53. Krackhardt D (1999) The ties that torture: Simmelian tie analysis in organizations. Res Sociol Organ 16:183–210
54. Bavelas A (1948) A mathematical model for group structures. Hum Organ 7:16–30
55. Freeman LC (1979) Centrality in social networks conceptual clarification. Soc Netw 1:215–239
56. Friedkin NE (1991) Theoretical foundations for centrality measures. Am J Sociol 96:1478–1504
57. Borgatti SP, Everett MG (2006) A graph-theoretic perspective on centrality. Soc Netw 28:466–484
58. Albert R, Barabasi A-L (2000) Topology of evolving networks: Local events and universality. Phys Rev Lett 85:5234–5237
59. Barabási A-L (2002) Linked: The new science of network. Perseus, Cambridge
60. Braha D, Bar-Yam Y (2006) From centrality to temporary fame: Dynamic centrality in complex networks. Complexity 12:59–63
61. Bonacich P (1972) Factoring and weighting approaches to clique identification. J Math Sociol 2:113–120
62. Bonacich P (1987) Power and centrality: A family of measures. Am J Sociol 92:1170–1182
63. Valente TW, Foreman RK (1998) Integration and radiality: Measuring the extent of an individual's connectedness and reachability in a network. Soc Netw 20:89–105
64. Granovetter M (1973) The strength of weak ties. Am J Sociol 78:1360–1380
65. Granovetter M (1995) Getting a job: A study of contacts and careers, 1st edn 1974. The University of Chicago Press, Chicago
66. Anthonisse JM (1971) The rush in a directed graph. Stichting Mathematisch Centrum, Amsterdam, p 10
67. Gould RV (1987) Measures of betweenness in non-symmetric networks. Soc Netw 9:277–282
68. White DR, Borgatti SP (1994) Betweenness centrality measures for directed graphs. Soc Netw 16:335–346
69. Stephenson K, Zelen M (1989) Rethinking centrality: Methods and examples. Soc Netw 11:1–37
70. Newman MEJ (2005) A measure of betweenness centrality based on random walks. Soc Netw 27:39–54
71. Freeman LC, Borgatti SP, White DR (1991) Centrality in valued graphs: A measure of betweenness based on network flow. Soc Netw 13:141–154
72. Tutzauer F (2007) Entropy as a measure of centrality in networks characterized by path-transfer flow. Soc Netw 29:249–265
73. Borgatti SP (2005) Centrality and network flow. Soc Netw 27:55–71
74. Snyder D, Kick E (1979) Structural position in the world system and economic growth 1955–70. A multiple network analysis of transnational interactions. Am J Sociol 84:1096–1126

75. Breiger RL (1981) Structures of economic interdependence among nations. In: Blau PM, Merton RK (eds) Continuities in structural inquiry. Sage, Newbury Park, pp 353–380
76. Bornschier V, Trezzini B (1997) Social stratification and mobility in the world system – Different approaches and recent research. Int Sociol 12:429–455
77. Borgatti SP, Everett MG (2000) Models of core/periphery structures. Soc Netw 21:375–395
78. Burt RS (1992) Structural holes: The social structure of competition. Harvard University Press, Cambridge/London
79. Burt RS (1992) The social structure of competition. In: Nohria N, Eccles RG (eds) Networks and organizations: Structure, form, and action. Harvard Business School Press, Boston, pp 57–91
80. Everett M, Borgatti SP (2005) Ego network betweenness. Soc Netw 27:31–38
81. Marsden PV (2002) Egocentric and sociocentric measures of network centrality. Soc Netw 24:407–422
82. Knoke D, Burt RS (1983) Prominence. In: Burt RS, Minor MJ (eds) Applied network analysis. A methodological introduction. Sage, Beverly Hills, pp 195–222
83. Merton RK (1968) The Matthew effect in science. Science 159:56–63
84. Lin N (1976) Foundations of social research. McGraw-Hill, New York
85. Bonacich P, Lloyd P (2001) Eigenvector-like measures of centrality for asymmetric relations. Soc Netw 23:191–201
86. Davis JA, Leinhardt S (1968) The structure of positive interpersonal relations in small groups. Annual Meeting of the American Sociological Association, Boston, August 1968
87. Davis JA, Leinhardt S (1972) The structure of positive interpersonal relations in small groups. In: Berger J (ed) Sociological theories in progress, vol 2. Houghton Mifflin, Boston, pp 218–251
88. Holland PW, Leinhardt S (1971) Transitivity in structural models of small groups. Comp Group Stud 2:107–124
89. Johnsen EC (1985) Network macrostructure models for the Davis–Leinhardt set of empirical sociomatrices. Soc Netw 7:203–224
90. de Nooy W (1999) The sign of affection: Balance-theoretic models and incomplete signed digraphs. Soc Netw 21:269–286
91. Davis JA (1979) The Davis/Holland/Leinhardt Studies: An overview. In: Holland PW, Leinhardt S (eds) Perspectives on social network research. Academic, New York, pp 51–62
92. Holland PW, Leinhardt S (1978) An omnibus test for social structure using triads. Sociol Methods Res 7:227–256
93. Doreian P, Batagelj V, Ferligoj A (2000) Symmetric-acyclic decompositions of networks. J Classif 17:3–28
94. Snijders TAB (2003) Accounting for degree distributions in empirical analysis of network dynamics. In: Breiger R, Carley K, Pattison P (eds) Dynamic social network modeling and analysis: Workshop summary and papers. National Academic, Washington DC, pp 146–161
95. Snijders TAB, Steglich CEG, Schweinberger M (2007) Modeling the co-evolution of networks and behavior. In: Montfort KV, Oud H, Satorra A (eds) Longitudinal models in the behavioral and related sciences. Erlbaum, Mahwah, pp 41–71
96. White HC (1992) Identity and control; A structural theory of social action. Princeton University Press, Princeton
97. Krackhardt D (1987) Cognitive social structures. Soc Netw 9:109–134
98. Breiger RL (2004) The analysis of social networks. In: Hardy MA, Bryman A (eds) Handbook of data analysis. Sage, London, pp 505–526
99. Bian Yj, Breiger RL, Davis D, Galaskiewicz J (2005) Occupation, class, and social networks in urban China. Social Forces 83:1443–1468
100. Mohr JW (1998) Measuring meaning structures. Ann Rev Sociol 24:345–370

Books and Reviews

Berkowitz SD (1982) An introduction to structural analysis: The network approach to social research. Butterworths, Toronto
Brandes U, Erlebach T (2005) Network analysis: Methodological foundations. Springer, Berlin
Breiger RL (2004) The analysis of social networks. In: Hardy MA, Bryman A (eds) Handbook of data analysis. Sage, London, pp 505–526
Carrington PJ, Scott J, Wasserman S (2005) Models and methods in social network analysis. Cambridge University Press, Cambridge
Degenne A, Forsé M (1999) Introducing social networks. Sage, London
de Nooy W, Mrvar A, Batagelj V (2005) Exploratory social network analysis with Pajek. Cambridge University Press, Cambridge
Freeman LC, White DR, Romney AK (1992) Research methods in social network analysis, 1st edn 1989. Transaction, New Brunswick, pp 530
Knoke D, Kuklinski JH (1982) Network analysis. Sage, Beverly Hills
Marsden PV, Lin N (1982) Social structure and network analysis. Sage, Beverly Hills
Scott J (1991) Social network analysis: A handbook. Sage, London
Wasserman S, Faust K (1994) Social network analysis: Methods and applications. Cambridge University Press, Cambridge
Wellman B, Berkowitz SD (1988) Social structures: A network approach. In: Granovetter M (ed) Structural analysis in the social sciences. Cambridge University Press, Cambridge, p 497

Social Network Analysis, Large-Scale

VLADIMIR BATAGELJ
University of Ljubljana, Ljubljana, Slovenia

Article Outline

Glossary
Definition of the Subject
Introduction
Large Networks and Complexity of Algorithms
Decompositions
Connectivity
Cuts
Dense Groups – Cores and Short Rings
Islands

Pattern Searching
Two Mode Networks
Multiplication of Networks
Statistical Approach
Future Directions
Bibliography

Glossary

For the basic notions on graphs and networks see the Articlea Wouter de Nooy: ► Social Network Analysis, Graph Theoretical Approaches to.

Network consists of vertices linked by lines and additional data about vertices and/or lines.

Network decomposition identification of parts of network and their interconnections. Usually it is described by a partition of set of vertices or set of lines.

Time complexity of algorithm describes how the time needed to run the algorithm depends on the size of the input data.

Reduction of network a network obtained by shrinking each cluster from a given partition into a vertex.

Condensation a reduction for strong connectivity partition.

Cut a subnetwork of vertices/lines with values of selected property above given threshold.

Island a connected subnetwork of selected size of (locally) important, with respect to selected property, vertices/lines.

Pattern searching identification of all appearances of selected small subnetwork (pattern or fragment) in a given network.

Topological sort procedure to determine a compatible ordering in acyclic network.

Definition of the Subject

A *network* is based on two sets: a set of *vertices* (nodes), that represent the selected *units*, and a set of *lines* (links), that represent *ties* between units. Each line has two vertices as its *end-points*; if they are equal it is called a *loop*. Vertices and lines form a *graph*. A line can be *directed* – an *arc*, or *undirected* – an *edge*.

Additional data about vertices or lines are usually known – their *properties* (attributes). For example: name/label, type, value, position, … In general

$$\text{Network} = \text{Graph} + \text{Data}\,.$$

The data can be measured or computed.

Formally, a *network* $\mathcal{N} = (\mathcal{V}, \mathcal{L}, \mathcal{P}, \mathcal{W})$ consists of the following:

- A *graph* $\mathcal{G} = (\mathcal{V}, \mathcal{L})$, where $\mathcal{V}$ is the set of vertices and $\mathcal{L} = \mathcal{E} \cup \mathcal{A}$. $\mathcal{E} \cap \mathcal{A} = \emptyset$ is the set of lines. $\mathcal{A}$ is the set of *arcs* and $\mathcal{E}$ is the set of *edges*.
- $\mathcal{P}$ – set of *vertex value functions* or properties: $p\colon \mathcal{V} \to A$
- $\mathcal{W}$ – set of *line value functions* or weights: $w\colon \mathcal{L} \to B$

The size of a network/graph is expressed by two numbers: number of vertices $n = |\mathcal{V}|$ and number of lines $m = |\mathcal{L}|$. In a *simple undirected* graph (no parallel edges, no loops) $m \le \frac{1}{2}n(n-1)$; and in a *simple directed* graph (no parallel arcs) $m \le n^2$.

For a family of graphs $\mathbb{G}$, we define a *density* of graph $\mathcal{G}$ as $\gamma(\mathcal{G}) = \frac{m(\mathcal{G})}{m_{\max}(\mathbb{G})}$.

Introduction

Small networks (some tens of vertices) can be represented by a picture and analyzed by many algorithms (***UCINET***, ***NetMiner***). Also *middle size* networks (some hundreds of vertices), if they are not dense, can still be represented by a picture, but some analytical procedures can't be used.

Till 1990 most networks were small – they were collected by researchers using surveys, observations, archival records, etc. The advances in IT allowed one to create networks from the data already available in the computer(s) or by browsing on the Internet. *Large* networks became reality. Large networks are too big to be displayed in detail; special algorithms are needed for their analysis (Pajek). The availability of large data sets also provided incentives to the boost of theoretical research in (large) network analysis (not only in the social sciences).

A recent overview of social network analysis software is given in Huisman and Van Duijn [28].

Large Networks and Complexity of Algorithms

Large networks have several thousands or millions of vertices. The upper limit to their size is technologically dependent – they can be stored in computer's memory; otherwise we deal with a *huge* network (see Abello et al. [1]).

Large networks are usually sparse $m \ll n^2$; typically $m = O(n)$ or $m = O(n \log n)$, see Table 1.

A collection of large networks is available from Pajek's datasets.

The *time complexity* of an algorithm describes how the time needed to run the algorithm depends on the size of the input data. In computer science the problems for which only algorithms of exponential (or higher) complexity are known are considered hard or intractable, since the speed-up of a computer only additively increases the size of problems that can be solved in a given period of

Social Network Analysis, Large-Scale, Table 1
Examples of large networks

Network	$n = \|\mathcal{V}\|$	$m = \|\mathcal{L}\|$	Source
ODLIS dictionary	2909	18419	ODLIS online
Citations SOM	4470	12731	Garfield's collection
Molecula 1ATN	5020	5128	Brookhaven PDB
Comput geometry	7343	11898	BiBTEX bibliographies
English words 2-8	52652	89038	Knuth's English words
Internet traceroutes	124651	207214	Internet Mapping Project
Franklin genealogy	203909	195650	Roperld.com gedcoms
World-Wide-Web	325729	1497135	Notre Dame Networks
Internet Movie DB	1324748	3792390	IMDB
Wikipedia	659388	16582425	Wikimedia
US patents	3774768	16522438	Nber
SI internet	5547916	62259968	Najdi Si

time, but the problems for which an algorithm of polynomial complexity exists are considered 'nice'. When dealing with large instances of problems this isn't always true anymore. Let us look to time complexities of some typical algorithms in Table 2.

For the interactive use on large networks already quadratic algorithms, $O(n^2)$, are too slow – we have to restrict our 'toolbox' to a selection of efficient, *subquadratic* algorithms.

How can we deal with large structures? Already Romans knew – ***divide et impera*** (divide and conquer). In case of networks ***divide*** means the use of (recursive) *decomposition* of a large network into several smaller networks (see Fig. 1) that can be visualized and treated further using more sophisticated methods; ***impera*** means that we have to take care about the interlinks among so obtained parts.

Another approach is the use of different *statistical* quantities to describe the properties of a network and using probabilistic models to derive the answers to some questions.

Decompositions

Decompositions of a network are usually described by clusterings of vertices or lines. In the following we shall use mainly the clusterings of vertices.

A nonempty subset $C \subseteq \mathcal{V}$ is called a *cluster* (group). A nonempty set of clusters $\mathbf{C} = \{C_i\}$ forms a *clustering*.

Clustering $\mathbf{C} = \{C_i\}$ is a *partition* iff

$$\cup\mathbf{C} = \bigcup_i C_i = \mathcal{V} \quad \text{and} \quad i \neq j \Rightarrow C_i \cap C_j = \emptyset\,.$$

Clustering $\mathbf{C} = \{C_i\}$ is a *hierarchy* iff $C_i \cap C_j \in \{\emptyset, C_i, C_j\}$. In other words, in a hierarchy two clusters are either disjoint or is one contained in the other.

Hierarchy $\mathbf{C} = \{C_i\}$ is *complete*, iff $\cup\mathbf{C} = \mathcal{V}$; and is *basic* if for all $v \in \cup\mathbf{C}$ also $\{v\} \in \mathbf{C}$.

Contraction of cluster C in a graph $\mathcal{G}$ is called a graph $\mathcal{G}/C$, in which all vertices of the cluster C

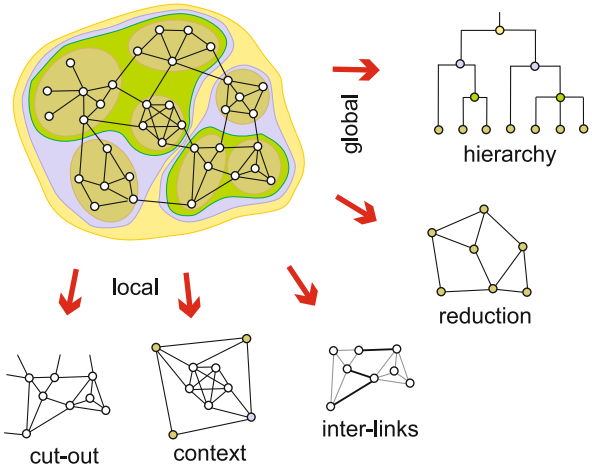

Social Network Analysis, Large-Scale, Figure 1
Decompositions

Social Network Analysis, Large-Scale, Table 2
Complexities of some typical algorithms

Algorithm	T(n)	1000	10 000	100 000	1 000 000	10 000 000
Alg-A	$O(n)$	0.00 s	0.015 s	0.17 s	2.22 s	22.2 s
Alg-B	$O(n \log n)$	0.00 s	0.06 s	0.98 s	14.4 s	**2.8 m**
Alg-C	$O(n\sqrt{n})$	0.01 s	0.32 s	10.0 s	**5.27 m**	**2.78 h**
Alg-D	$O(n^2)$	0.07 s	7.50 s	**12.5 m**	**20.8 h**	**86.8 d**
Alg-E	$O(n^3)$	0.10 s	**1.67 m**	**1.16 d**	**3.17 y**	**3.17 ky**

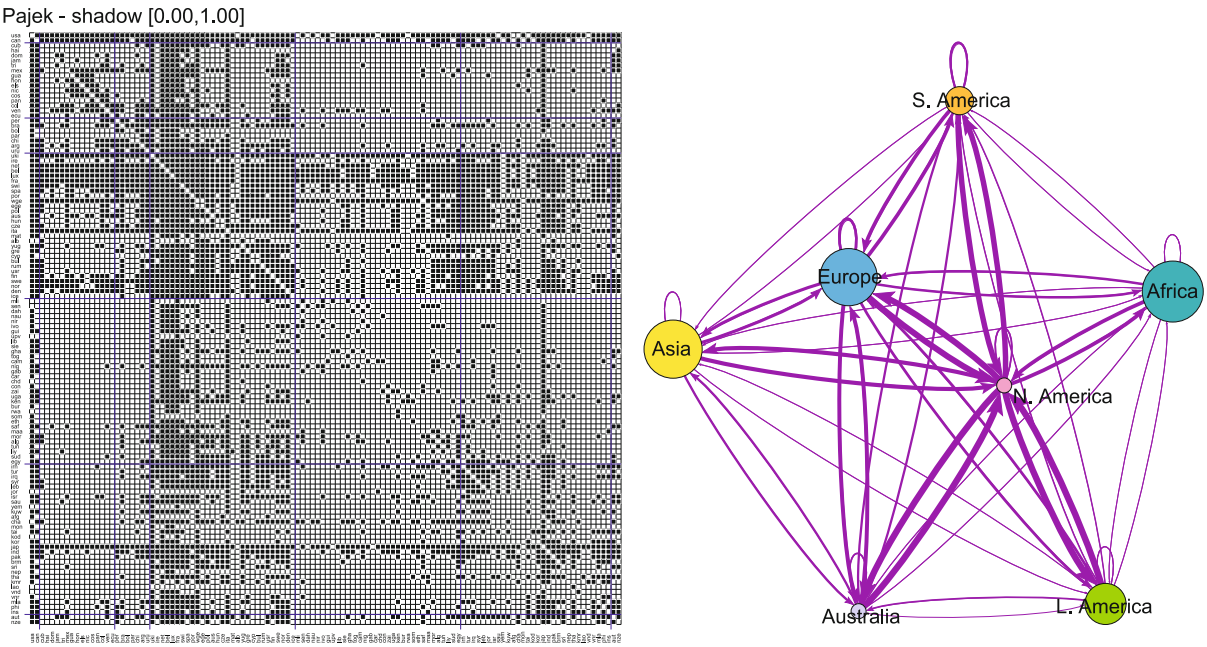

Social Network Analysis, Large-Scale, Figure 2
Snyder and Kick's international trade; matrix display and reduction

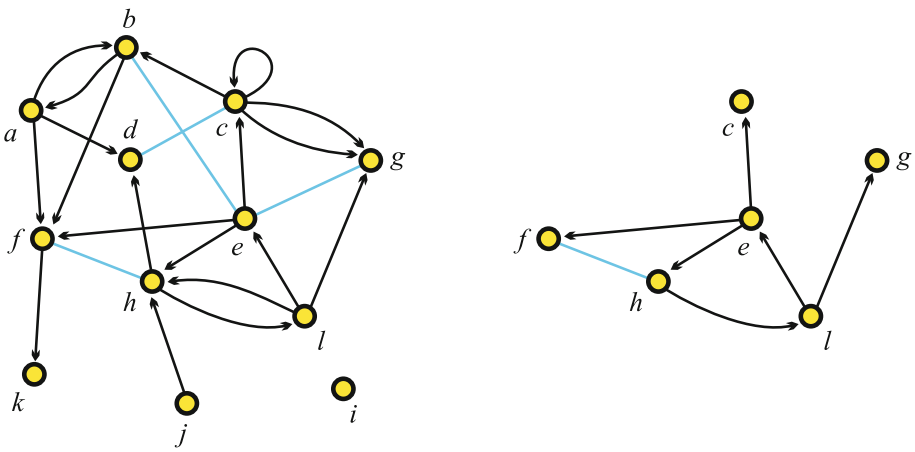

Social Network Analysis, Large-Scale, Figure 3
Graph and its subgraph

are replaced by a single new vertex, say c. More precisely: $\mathcal{G}/C = (\mathcal{V}', \mathcal{L}')$, where $\mathcal{V}' = (\mathcal{V} \setminus C) \cup \{c\}$ and $\mathcal{L}'$ consists of lines from $\mathcal{L}$ that have both end-points in $\mathcal{V} \setminus C$. Beside these it contains also a 'star' with the center c and: arc(v, c), if $\exists p \in \mathcal{L}, u \in C$: $p(v, u)$; or arc(c, v), if $\exists p \in \mathcal{L}, u \in C$: $p(u, v)$. There is a loop (c, c) in c if $\exists p \in \mathcal{L}, u, v \in C$: $p(u, v)$.

In a network over graph $\mathcal{G}$, we have also to specify how the new values/weights are determined in the shrunk part of the network. Usually as the sum or maximum/minimum of the original values.

For a given partition if we contract all clusters except few selected we obtain their *context*; and if we contract all clusters we obtain the *reduction* of a given network.

On the left side of the Fig. 2 the *matrix display* of Snyder and Kick's [40] international trade network is presented. Vertices in the display are reordered according to the partition by (sub)continents. On the right side the

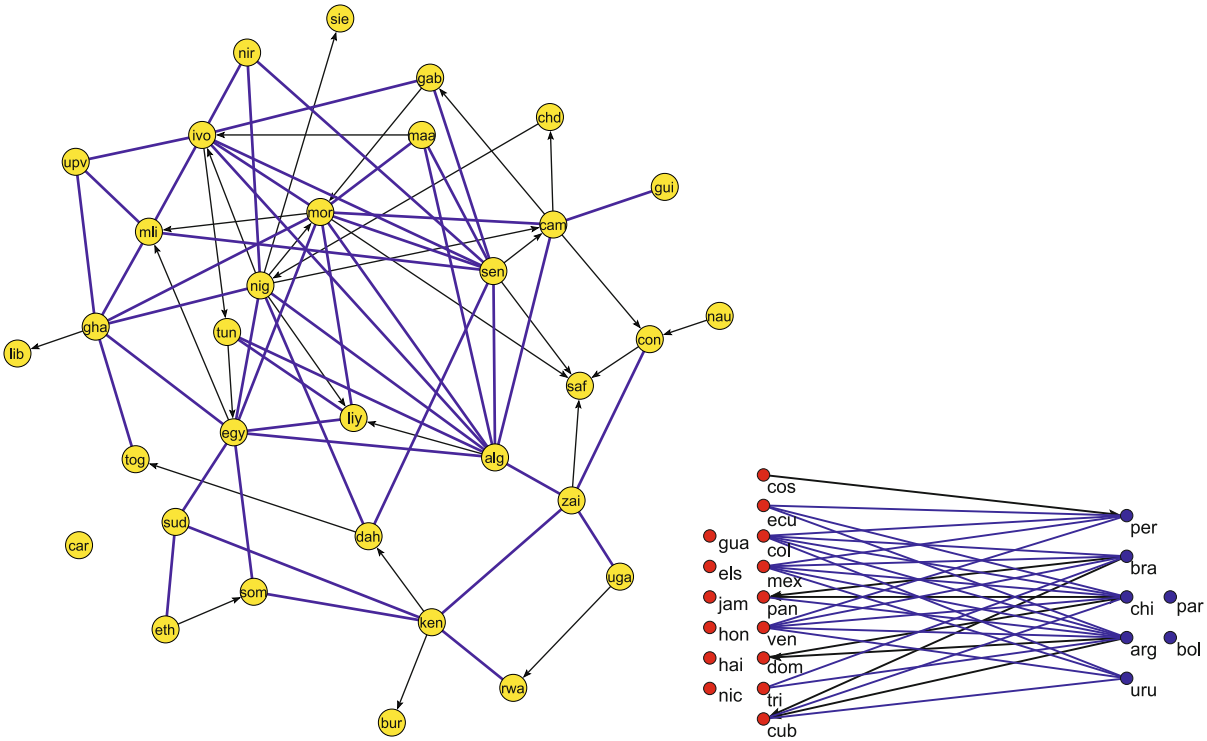

Social Network Analysis, Large-Scale, Figure 4
Africa cut-out and inter-links between South and Latin America

corresponding reduction of the network is presented. The lines in the reduction have the thickness proportional to the weights

$$w(C_i, C_j) = \frac{n(C_i, C_j)}{n(C_i) \cdot n(C_j)}$$

where $n(C_i, C_j)$ is the number of lines from cluster C_i to cluster C_j; and $n(C_i)$ is the number of lines inside the cluster C_i.

A *subgraph* $\mathcal{H} = (\mathcal{V}', \mathcal{L}')$ of a given graph $\mathcal{G} = (\mathcal{V}, \mathcal{L})$ is a graph which set of lines is a subset of set of lines of $\mathcal{G}$, $\mathcal{L}' \subseteq \mathcal{L}$, its vertex set is a subset of set of vertices of $\mathcal{G}$, $\mathcal{V}' \subseteq \mathcal{V}$, and it contains all end-vertices of $\mathcal{L}'$. The graph on the right side of Fig. 3 is a subgraph of the graph on the left side.

A subgraph can be *induced* by a given subset of vertices $\mathcal{V}'$, then $\mathcal{L}' = \mathcal{L}|\mathcal{V}'$ consists of all lines from $\mathcal{L}$ which have both end-points in $\mathcal{V}'$; or lines $\mathcal{L}'$, then $\mathcal{V}' = \mathcal{V}|\mathcal{L}'$ consists of all end-points of lines from $\mathcal{L}'$. It is a *spanning* subgraph iff $\mathcal{V}' = \mathcal{V}$.

On the left side of Fig. 4 the *cut-out* of African countries from the Snyder and Kick's network is presented – the induced subgraph by Africa cluster; and on the right side the *inter-links* between Latin America and South America – the induced subgraph by Latin America and South America clusters with inside cluster lines removed.

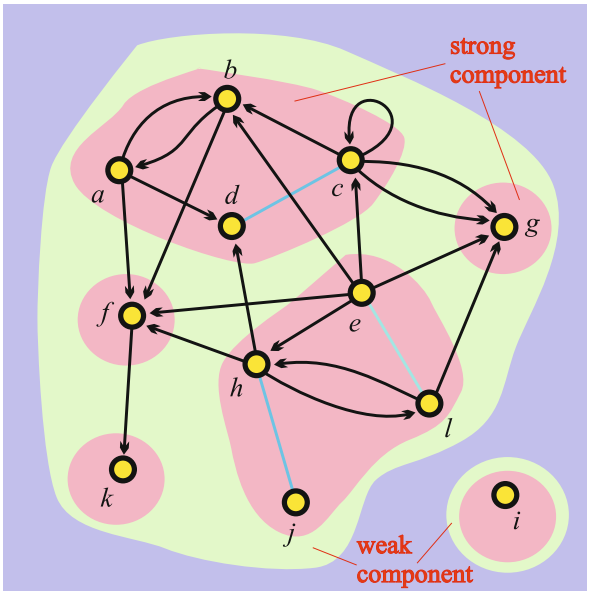

Social Network Analysis, Large-Scale, Figure 5
Weak and strong components

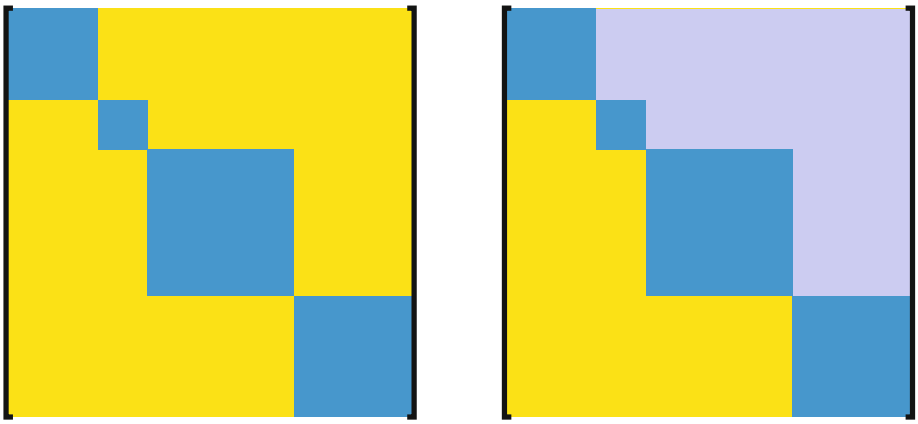

Social Network Analysis, Large-Scale, Figure 6
Weak and strong components in matrix display

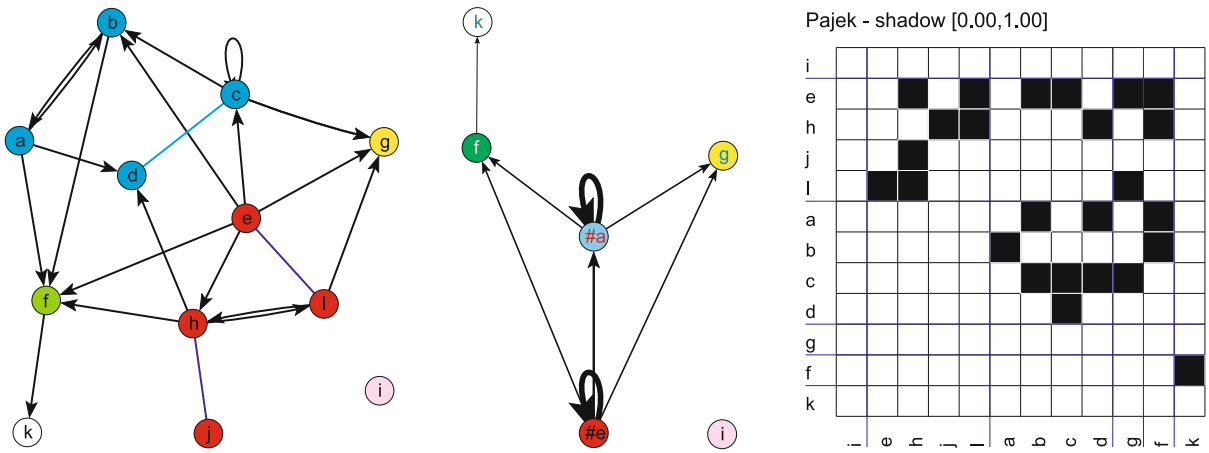

Social Network Analysis, Large-Scale, Figure 7
Condensation

Connectivity

A *walk* from vertex u to vertex v is a sequence of lines $l(v_{i-1}, v_i)$, $i = 1, \ldots, k$ such that $v_0 = u$ and $v_k = v$. k is called the *length* of the walk. If in the definition of a walk we don't care about the direction of its lines we get a *semi-walk*. A walk is *closed* iff $u = v$. A graph is *acyclic* iff it doesn't contain any closed walk. A walk in which all vertices are different is a *path*.

Vertex u is *reachable* from vertex v iff there exists a walk with initial vertex v and terminal vertex u. Vertex v is *weakly connected* with vertex u iff there exists a semi-walk with v and u as its end-vertices. Vertex v is *strongly connected* with vertex u iff they are mutually reachable.

Weak and strong connectivity are equivalence relations. Equivalence classes induce weak/strong *components* (See Fig. 5).

Reordering the vertices of network such that the vertices from the same class of weak partition are put together we get a matrix representation (left side of Fig. 6) consisting of diagonal blocks – weak components. The out-diagonal blocks are zero-blocks. Most problems can be solved separately on each component and afterward these solutions combined into final solution.

If we shrink every strong component of a given graph into a vertex, delete all loops and identify parallel arcs the obtained reduced graph, called also the *condensation* of a given graph, is acyclic [27]. For every acyclic

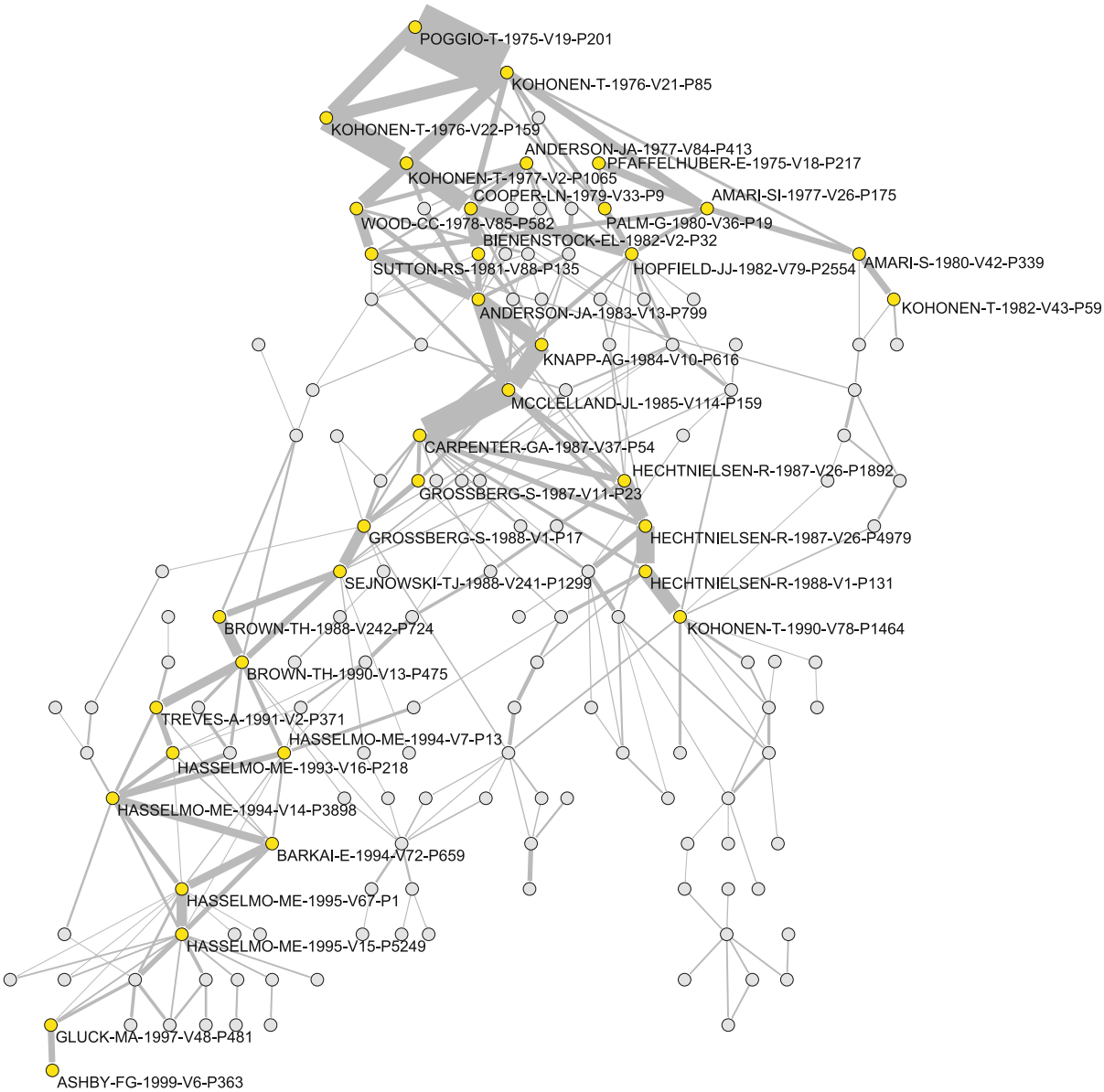

Social Network Analysis, Large-Scale, Figure 8
Main component of arc cut at level 0.007 of the SOM citation network

graph an *ordering/level* function $i\colon \mathcal{V} \to \mathbb{N}$ exists such that $(u, v) \in \mathcal{A} \Rightarrow i(u) < i(v)$. The procedure to determine such ordering is called *topological sort* [19]. Reordering in matrix display the vertices of a network by this ordering we obtain a representation as at the right side of Fig. 6 – the blocks below the diagonal are zero-blocks.

A directed graph, its condensation and its topologically ordered matrix display are presented in Fig. 7.

For several network analysis problems more efficient algorithms exist for acyclic networks.

Cuts

The basic approach to find interesting groups inside a network is to express our intentions (question) with an appropriate property/weight (measured or computed from network structure) and then identify the substructures of elements with the highest (lowest) values of the selected property. This approach is known as a method of *cuts*.

There exist several measures of importance of vertices in a network such as: degree, betweeness, closeness [16,24], hubs and authorities [30], clustering coefficient, …

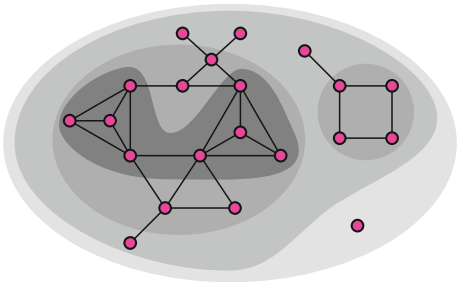

Social Network Analysis, Large-Scale, Figure 9
Cores

The *degree* $\deg(v)$ of vertex v equals to the number of lines having vertex v as their end-point. The *maximum degree* of a graph is denoted by Δ. Similarly the in-degree $\mathrm{indeg}(v)$ of vertex v equals to the number of lines having vertex v as their terminal point, and the out-degree $\mathrm{outdeg}(v)$ ...

The *vertex-cut* of a network $\mathcal{N} = (\mathcal{V}, \mathcal{L}, p)$, for a property $p\colon \mathcal{V} \to \mathbb{R}$, at selected level t is a subnetwork $\mathcal{N}(t) = (\mathcal{V}', \mathcal{L}(\mathcal{V}'), p)$, determined by the set

$$\mathcal{V}' = \{v \in \mathcal{V}\colon p(v) \geq t\}$$

and $\mathcal{L}(\mathcal{V}')$ is the set of lines from $\mathcal{L}$ that have both end-points in $\mathcal{V}'$.

The *line-cut* of a network $\mathcal{N} = (\mathcal{V}, \mathcal{L}, w)$, for a weight $w\colon \mathcal{L} \to \mathbb{R}$, at selected level t is a subnetwork $\mathcal{N}(t) = (\mathcal{V}(\mathcal{L}'), \mathcal{L}', w)$, determined by the set

$$\mathcal{L}' = \{e \in \mathcal{L}\colon w(e) \geq t\}$$

and $\mathcal{V}(\mathcal{L}')$ is the set of all end-points of the lines from $\mathcal{L}'$.

In the analysis of a cut $\mathcal{N}(t)$, we look at its components. Their number and sizes depend on t. Usually there are many small components. Often we consider only components of size at least k and not exceeding K. The components of size smaller than k are discarded as 'less interesting'; and the components of size larger than K are cut again at some higher level.

The values of threshold t and size bounds k and K are determined by inspecting the distribution of vertex/line-values and the distribution of component sizes and considering additional knowledge on the nature of network or goals of analysis.

The p_S-core at level 46 (see Fig. 11) of the collaboration network in the field of computational geometry is an example of vertex cut.

The citation network analysis started in 1964 with the paper of Garfield et al. [25]. In 1989 Hummon and Doreian [29] proposed three indices – weights of arcs that are proportional to the number of different source-sink paths passing through the arc. In Fig. 8 the main component

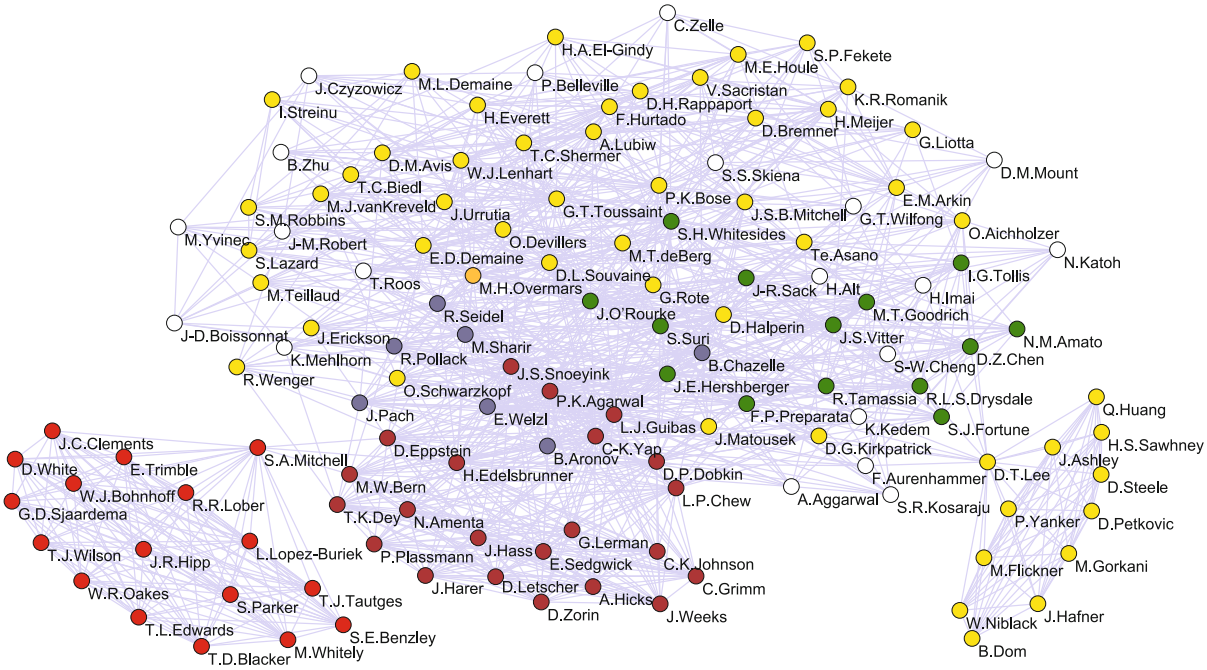

Social Network Analysis, Large-Scale, Figure 10
Cores of orders 10–21 in Computational Geometry collaboration network

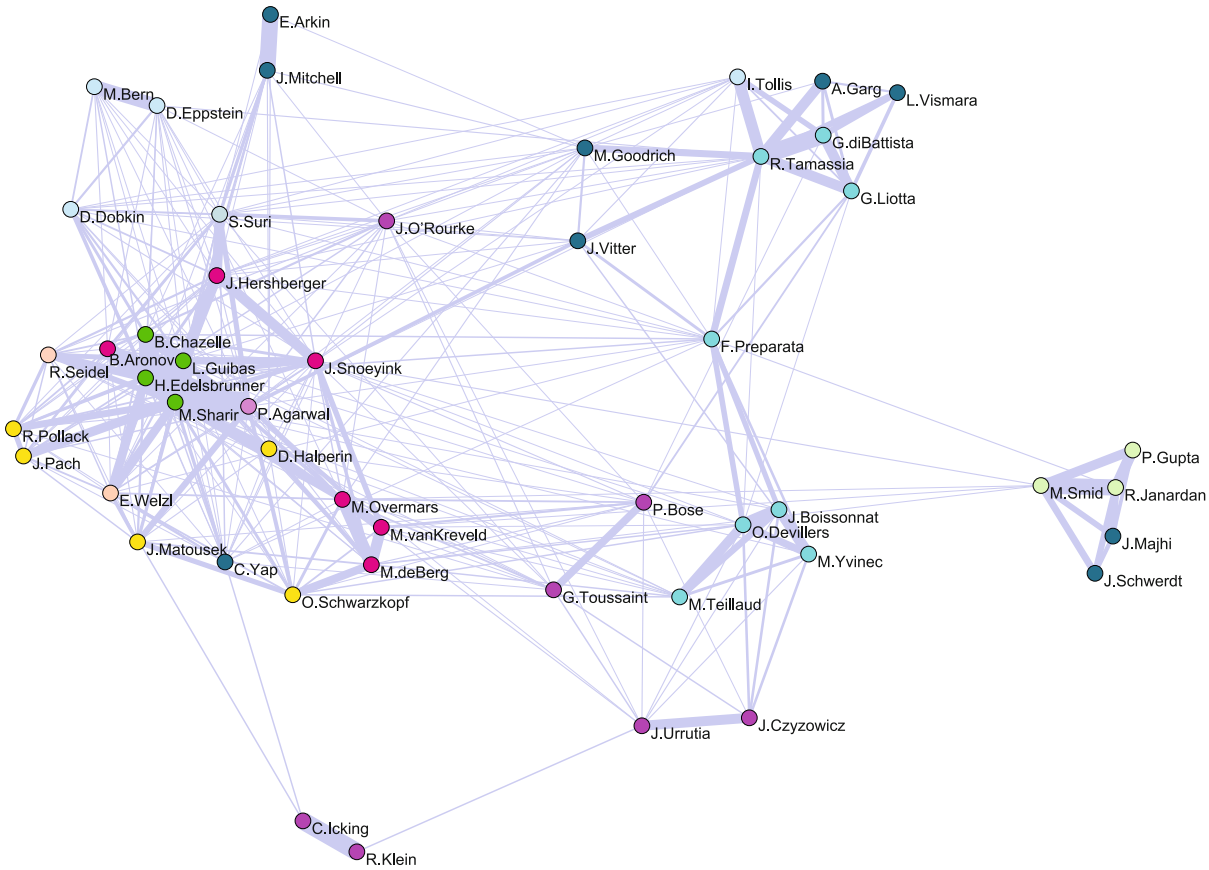

Social Network Analysis, Large-Scale, Figure 11
***p_5*-core at level 46 in Computational Geometry collaboration network**

of the arc cut at level 0.007 for SPC (search path count) weights of the SOM (selforganizing maps) citation network (4470 vertices, 12,731 arcs) is presented.

Dense Groups – Cores and Short Rings

Several notions were proposed in attempts to formally describe dense groups in graphs.

Clique of order k, $k \geq 3$, is a maximal complete subgraph (isomorphic to complete graph K_k – graph with k vertices and all possible edges among them).

Other notions are: s-plexes, s-clans, LS sets, lambda sets, cores, … (Wasserman and Faust [43]). For all of them, except for cores, it turned out that they are difficult (no fast algorithm exists) to determine.

The notion of core was introduced by Seidman in 1983 [38]. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph. A subgraph $\mathcal{H}_k = (\mathcal{W}, \mathcal{E}|\mathcal{W})$ induced by the set $\mathcal{W}$ is a *k-core* or a *core of order k* iff for all $v \in \mathcal{W}$: $\deg_{\mathcal{H}_k}(v) \geq k$, and $\mathcal{H}_k$ is a maximal subgraph with this property. The core of maximum order is also called the *main* core. The *core number* of vertex v is the highest order of a core that contains this vertex. In general graphs instead of the degree $\deg(v)$ we can also use: in-degree, out-degree, in-degree + out-degree, etc., determining different types of cores.

From Fig. 9, representing 0, 1, 2 and 3 core, we can see the following properties of cores:

- The cores are nested: $i < j \Longrightarrow \mathcal{H}_j \subseteq \mathcal{H}_i$. They form a hierarchy.
- Cores are not necessarily connected subgraphs.

An efficient algorithm for determining the cores hierarchy is based on the following property: If from a given graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ we recursively delete all vertices, and edges incident with them, of degree less than k, the remaining graph is the k-core.

The Fig. 10 presents the cores of orders 10 to 21 in the collaboration network ($n = 7343$, $m = 11{,}898$) for the field of computational geometry – two authors are linked

iff they wrote a paper together. The weight of the edge equals to the number of joint papers.

The notion of core can be generalized to networks. Let $\mathcal{N} = (\mathcal{V}, \mathcal{E}, w)$ be a network, where $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a graph and weight $w\colon \mathcal{E} \to \mathbb{R}$ is a function assigning values to edges. A *vertex property function* on $\mathcal{N}$, or a *p-function* for short, is a function $p(v, U), v \in \mathcal{V}, U \subseteq \mathcal{V}$ with real values. Let $N_U(v) = N(v) \cap U$, where $N(v)$ is the set of neighbors of v. Besides degrees, here are some other examples of p-functions [13]:

$$p_S(v, U) = \sum_{u \in N_U(v)} w(v, u), \text{ where } w\colon \mathcal{E} \to \mathbb{R}_0^+$$

$$p_M(v, U) = \max_{u \in N_U(v)} w(v, u), \text{ where } w\colon \mathcal{E} \to \mathbb{R}$$

$$p_k(v, U) = \text{number of cycles of length } k \text{ through vertex } v \text{ in } (U, \mathcal{E}|U)$$

$$p_\gamma(v, U) = \begin{cases} \dfrac{\deg(v, U)}{\max_{u \in N(v)} \deg(u)}, & \text{if } \deg(v) > 0; \\ 0, & \text{otherwise} \end{cases}$$

$$p_\delta(v, U) = \max_{u \in N_U^+(v)} \deg(u) - \min_{u \in N_U^+(v)} \deg(u)$$

$$p_a(v, U) = \begin{cases} \dfrac{1}{|N_U(v)|} \displaystyle\sum_{u \in N_U(v)} w(v, u), & \text{if } N_U(v) \neq \emptyset; \\ 0, & \text{otherwise} \end{cases}$$

The subgraph $\mathcal{H} = (C, \mathcal{E}|C)$ induced by the set $C \subseteq \mathcal{V}$ is a *p-core at level* p-core $t \in \mathbb{R}$ iff for all $v \in C\colon t \leq p(v, C)$ and C is a maximal such set.

The function p is *monotone*, iff it has the property

$$C_1 \subset C_2 \Rightarrow \forall v \in \mathcal{V}\colon (p(v, C_1) \leq p(v, C_2))$$

The degrees and the functions p_S, p_M, p_k, p_γ and p_δ are monotone; and p_a is not. For a monotone function the p-core at level t can be determined, as in the ordinary case, by successively deleting vertices with value of p lower than t; and the cores on different levels are nested

$$t_1 < t_2 \Rightarrow \mathcal{H}_{t_2} \subseteq \mathcal{H}_{t_1}$$

The p-function is *local*, iff $p(v, U) = p(v, N_U(v))$. The degrees, $p_S, p_M, p_\gamma, p_\delta$ and p_a are local; but p_k is **not** local for $k \geq 4$. For a local monotone p-function an $O(m \max(\Delta, \log n))$ algorithm for determining the p-core levels exists, assuming that $p(v, N_C(v))$ can be computed in $O(\deg_C(v))$.

Figure 11 presents the p_S-core at level 46 of the collaboration network in the field of computational geometry. Note, for example, that R. Klein (lower left) has in-core degree only 2, but its in-core sum of weights is at least 46 – he wrote most of his papers with C. Icking.

A k-ring is a simple closed chain of length k. Using k-rings we can define a weight of an edge e as $w_k(e) = \#$ of different k-rings containing the edge $e \in \mathcal{E}$.

Since for each edge e of complete graph K_r, $r \geq k \geq 3$ we have $w_k(e) = (r-2)!/(r-k)!$, the edges belonging to cliques have large weights. Therefore these weights can be used to identify the dense parts of a network. For example: all r-cliques of a network belong to $(r-2)$-edge cut for the weight w_3.

Related to triangular (3-rings) network is the notion of *triangular connectivity* that can be used to operationalize the notion of Granovetter's strong and weak ties [26]. This notion can be generalized to short cycle connectivity. For details see [14]. For efficient algorithms for computing triangles in networks see [10,39], and Latapy.

In Fig. 12 the edge-cut at level 16 of triangular network of Erdős collaboration graph (without Erdős, $n = 6926$, $m = 11{,}343$) is presented [9].

In directed networks there are two types of triangles or 3-rings (cyclic and transitive, see Fig. 14).

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a simple undirected graph. *Clustering* in vertex v is usually measured as a quotient between the number of lines in subgraph $\mathcal{G}^1(v) = \mathcal{G}(N(v))$ induced by the neighbors of vertex v and the number of lines in the complete graph on these vertices:

$$C(v) = \begin{cases} \dfrac{2|\mathcal{L}(\mathcal{G}^1(v))|}{\deg(v)(\deg(v)-1)}, & \deg(v) > 1 \\ 0, & \text{otherwise} \end{cases}.$$

For simple directed graphs, we have to omit the number 2.

So defined clustering coefficient attains largest values mostly on vertices of low degree – it is not useful for data analysis task. A better coefficient is obtained by the following correction

$$C_1(v) = \frac{\deg(v)}{\Delta} C(v)$$

where Δ is the maximum degree in graph $\mathcal{G}$. This measure attains its largest value in vertices that belong to an isolated clique of size Δ.

Islands

Islands are very general and efficient approach to determine the 'important' subnetworks in a given network with respect to a given property of vertices or lines. It is an improvement of the cuts approach. If we represent a given or computed value of vertices/lines as a height of vertices/

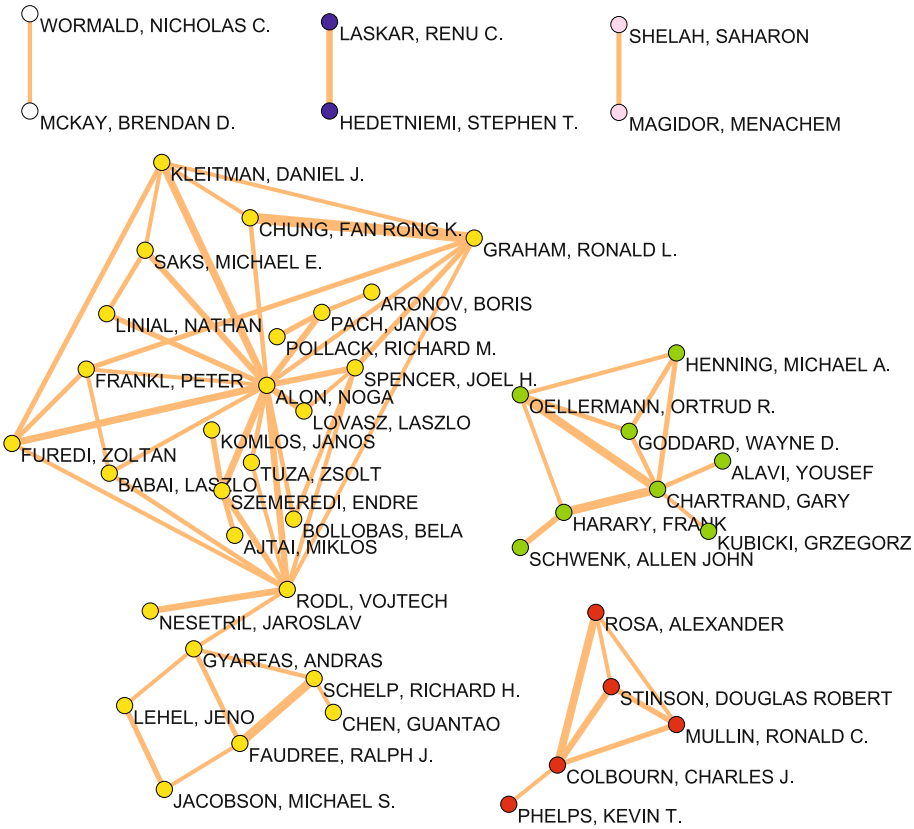

Social Network Analysis, Large-Scale, Figure 12
Edge-cut at level 16 of triangular network of Erdős collaboration graph

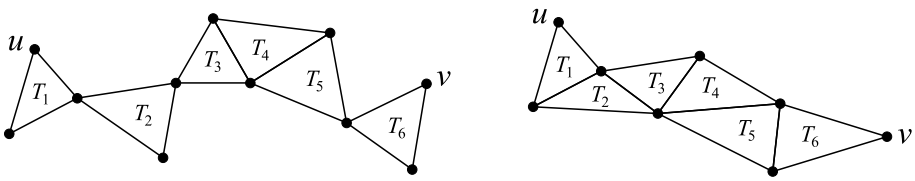

Social Network Analysis, Large-Scale, Figure 13
Vertex and edge triangular connectivity

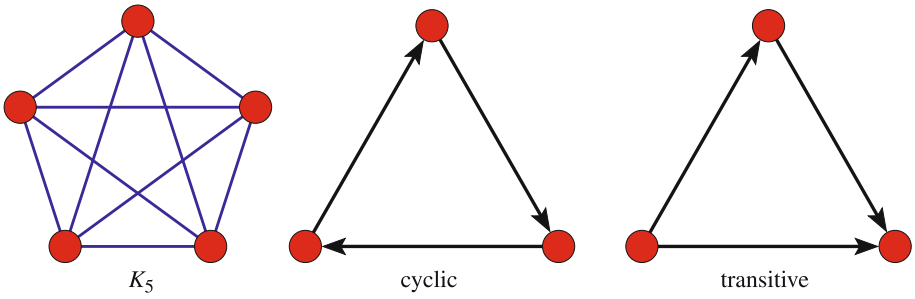

Social Network Analysis, Large-Scale, Figure 14
K_5 and cyclic and transitive 3-ring

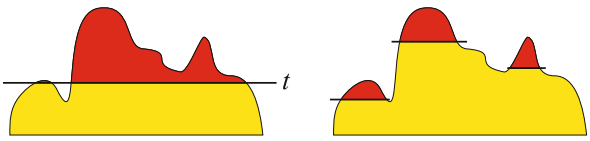

Social Network Analysis, Large-Scale, Figure 15
Cuts and islands

lines and we immerse the network into a water up to selected level we get *islands*. Varying the level we get different islands [45].

In the islands approach we select only maximal islands of sizes inside the given size bounds k to K, but on different levels. In this way we bypass the problems of the cuts approach: determining the 'right' threshold value and too small/large sizes of obtained components. Besides this we can also identify locally important islands with small heights – emerging groups. Very efficient algorithms exist to determine the islands hierarchy and to list all the islands of selected sizes. An island is *simple* iff it has only one peak.

As an example, let us take the Nber network of US Patents. It has 3,774,768 vertices and 16,522,438 arcs. We computed SPC weights in it and determined all (2,90)-islands. The reduced network has 470,137 vertices, 307,472 arcs and for different k: $C_2 = 187{,}610$, $C_5 = 8859$, $C_{30} = 101$, $C_{50} = 30$ islands. The main island turns out

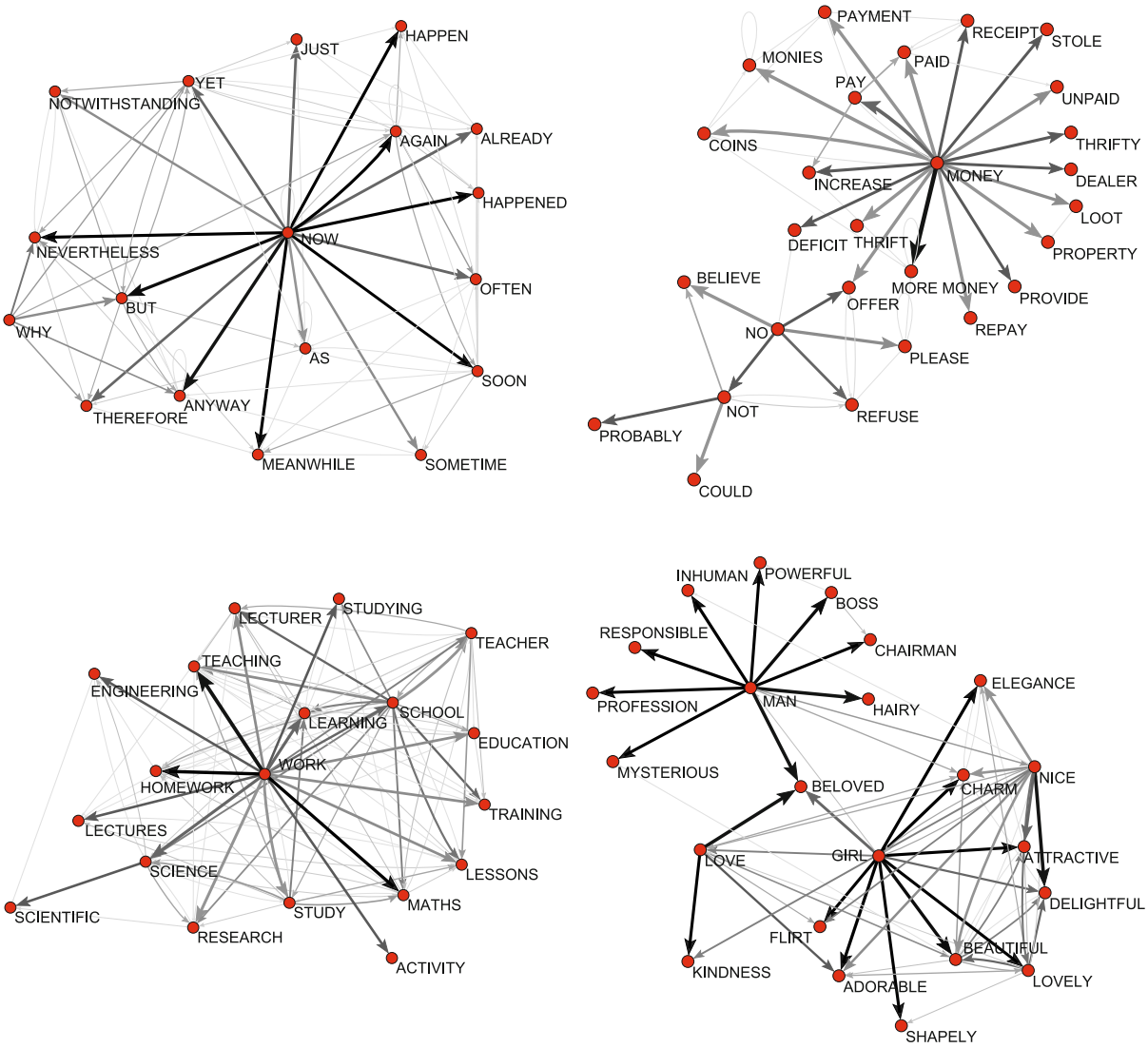

Social Network Analysis, Large-Scale, Figure 16
Selected islands from The Edinburgh Associative Thesaurus

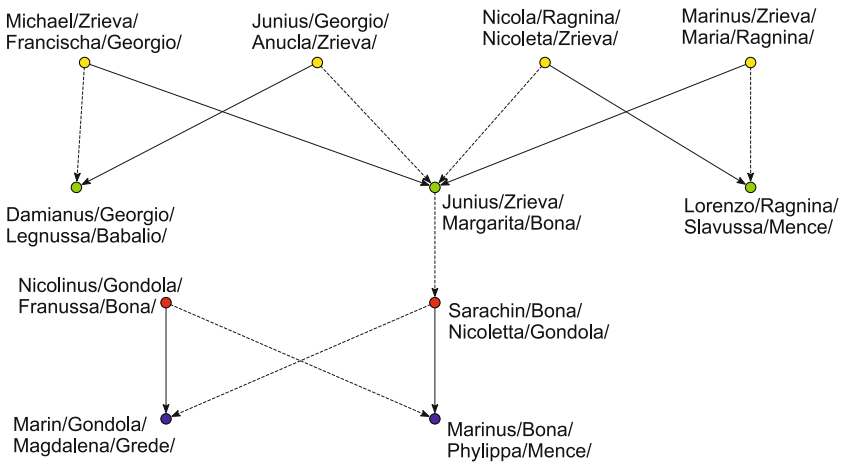

Social Network Analysis, Large-Scale, Figure 17
Marriages among relatives in Ragusa

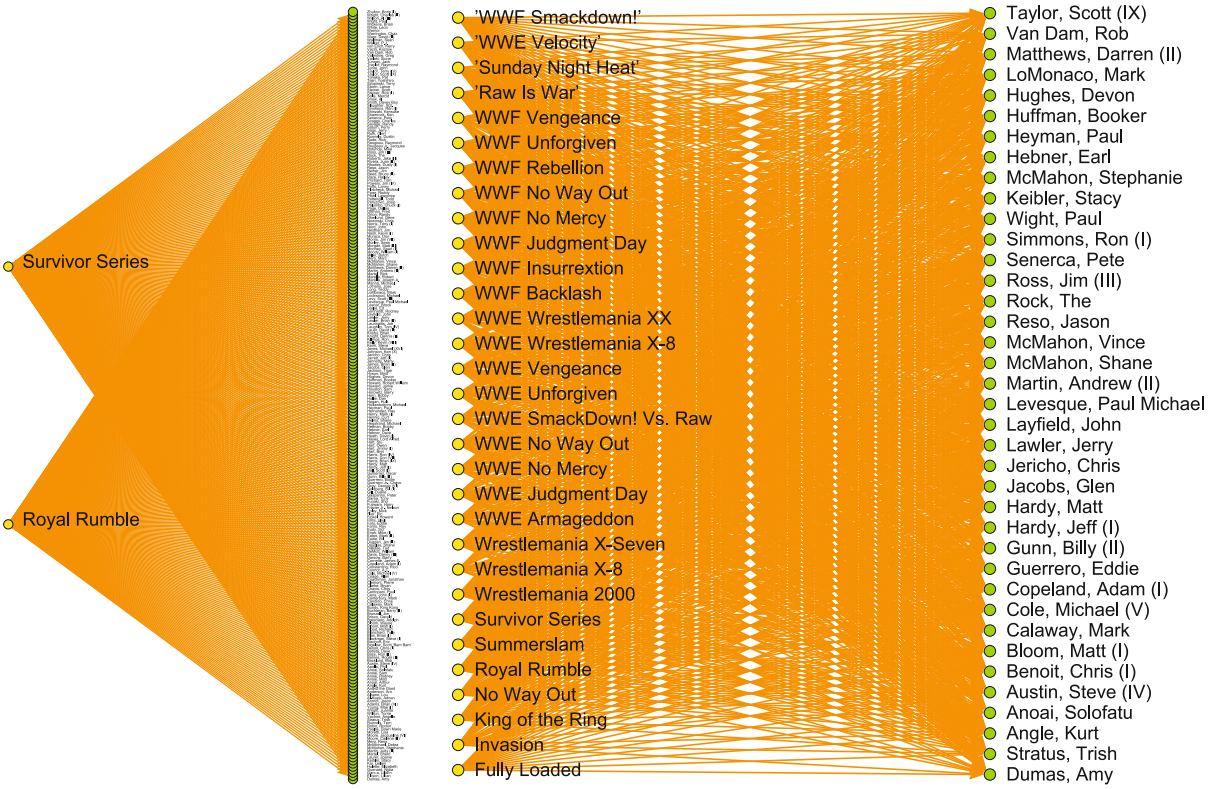

Social Network Analysis, Large-Scale, Figure 18
(247,2)-core and (27,22)-core of IMDB – wrestling

to be the island on the theme *LCD – Liquid crystal display*. In Fig. 16 four islands for transitivity triangular weight from *The Edinburgh Associative Thesaurus* network ($n = 23{,}219$, $m = 325{,}624$) are presented. From the left bottom island of words around the leader 'WORK' we see that the data were collected asking students.

Pattern Searching

If a selected *pattern* determined by a given graph does not occur frequently in a sparse network the straightforward backtracking algorithm applied for pattern searching finds all appearances of the pattern very fast even in

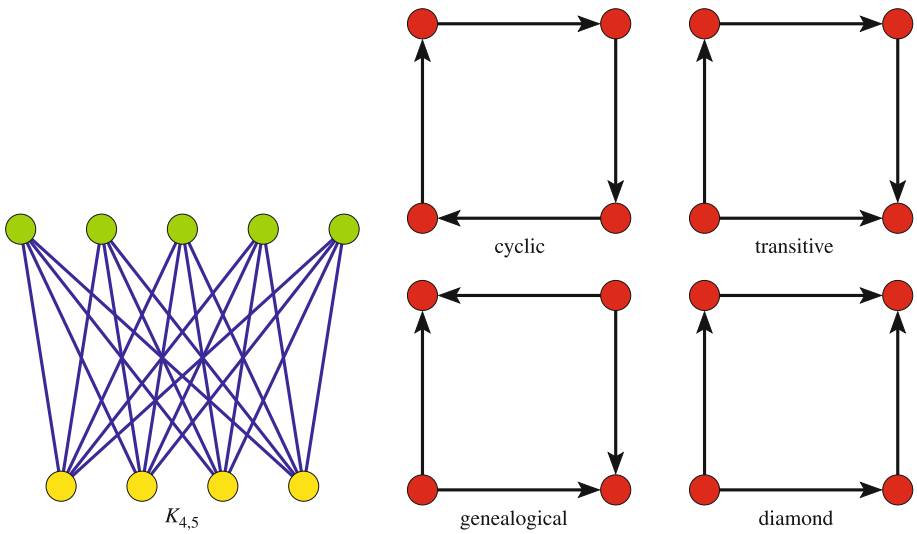

Social Network Analysis, Large-Scale, Figure 19
$K_{4,5}$ and directed 4-rings

the case of very large networks. Pattern searching was successfully applied to searching for patterns of atoms in large organic molecula (carbon rings) and searching for relinking marriages in genealogies (Batagelj and Mrvar [12]; Batagelj [6]).

The Fig. 17 presents three connected relinking marriages in the genealogy (represented as a p-graph) of Ragusan noble families. In a p-graph the vertices represent married couples or nonmarried individuals. A solid arc indicates the __ *is a son of* __ relation, and a dotted arc indicates the __ *is a daughter of* __ relation. In all three patterns a brother and a sister from one family found their partners in the same other family.

Two Mode Networks

A network $\mathcal{N} = (\mathcal{V}, \mathcal{L}, w)$ in which the set of vertices $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2$ is composed of two disjoint sets $\mathcal{V}_1$ and $\mathcal{V}_2$, and $\mathcal{L}$ is a set of *lines* linking $\mathcal{V}_1$ and $\mathcal{V}_2$ is called a *two-mode* or bipartite network.

The two-mode networks often appear in applications, but till recently no directed methods for analysis of larger two-mode networks were available. To identify dense parts of two-mode network, we can use the adapted cores and short rings approaches (Ahmed et al. [3]).

The subset of vertices $C \subseteq \mathcal{V}$ is a (p, q)-core in a two-mode network $\mathcal{N} = (\mathcal{V}_1, \mathcal{V}_2; \mathcal{L})$, $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2$ iff

a. in the induced subnetwork $\mathcal{H} = (C_1, C_2; \mathcal{L}(C))$, $C_1 = C \cap \mathcal{V}_1$, $C_2 = C \cap \mathcal{V}_2$ it holds for all $v \in C_1$: $\deg_{\mathcal{H}}(v) \geq p$ and for all $v \in C_2$: $\deg_{\mathcal{H}}(v) \geq q$;

b. C is the maximal subset of $\mathcal{V}$ satisfying condition **a**.

The two-mode cores have the following properties:

- $C(0, 0) = \mathcal{V}$
- $C(p, q)$ is not always connected
- $(p_1 \leq p_2) \wedge (q_1 \leq q_2) \Rightarrow C(p_2, q_2) \subseteq C(p_1, q_1)$.

To determine a (p, q)-core an algorithm similar to the ordinary core algorithm can be used: recursively remove from the first set all vertices of degree less than p, and from the second set all vertices of degree less than q. It can be implemented to run in $O(m)$ time.

The main question when applying the bipartite cores is what are the right values of p and q? The most interesting are the values on the 'border' that don't produce too large cores.

In Fig. 18 the (247,2)-core and (27,22)-core from the Internet Movie Database (two-mode network actors × movies, $n = 1{,}324{,}748 = 428{,}440 + 896{,}308$ vertices and $m = 3{,}792{,}390$ arcs) are presented. Both deal with wrestling.

In 2-mode network there are no 3-rings. The densest substructures are complete bipartite subgraphs $K_{p,q}$ – see $K_{4,5}$ on the left side of Fig. 19. They contain many 4-rings

$$w_4(e) = (p-1)(q-1)\,, \quad \text{for } e \in K_{p,q}\,.$$

There are four types of directed 4-rings – see the right side of Fig. 19. In the case of transitive rings we can count also on how many transitive rings the arc is a *shortcut*.

In the Internet Movie Database, we obtained for w_4 12,465 simple line islands on 56,086 vertices; 30 among

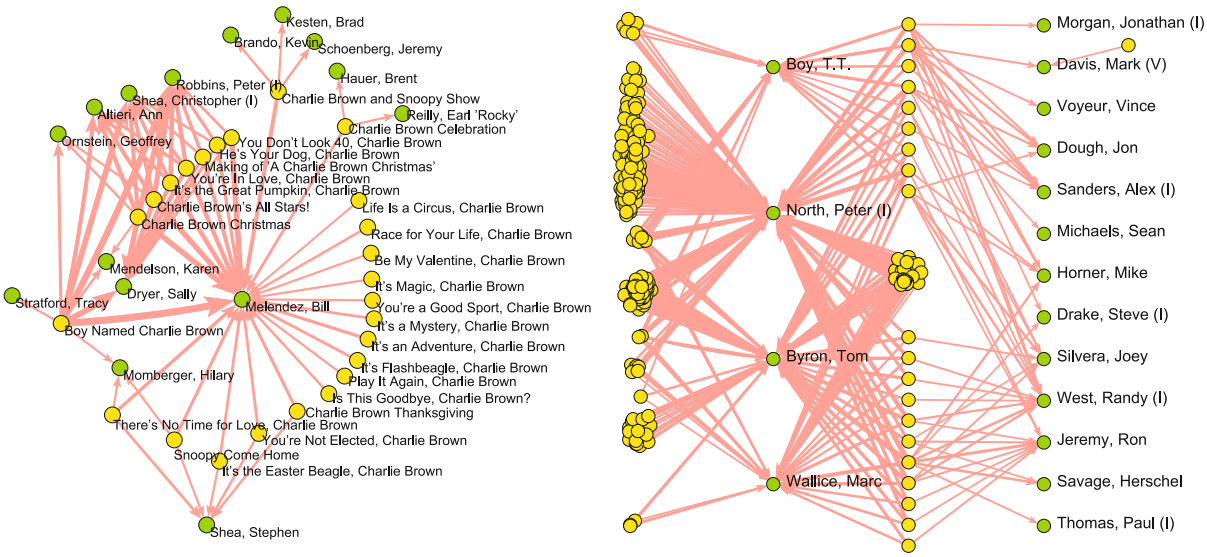

Social Network Analysis, Large-Scale, Figure 20
Islands for w_4 / Charlie Brown and adult

them have size at least 50. Two of them are presented on Fig. 20.

Multiplication of Networks

To a simple two-mode *network* $\mathcal{N} = (\mathcal{I}, \mathcal{J}, \mathcal{E}, w)$; where $\mathcal{I}$ and $\mathcal{J}$ are sets of *vertices*, $\mathcal{E}$ is a set of *edges* linking $\mathcal{I}$ and $\mathcal{J}$, and $w\colon \mathcal{E} \to \mathbb{R}$ is a *weight*; we can assign a *network matrix* $\mathbf{W} = [w_{i,j}]$ with elements: $w_{i,j} = w(i, j)$ for $(i, j) \in \mathcal{E}$ and $w_{i,j} = 0$ otherwise.

Given a pair of compatible networks $\mathcal{N}_A = (\mathcal{I}, \mathcal{K}, \mathcal{E}_A, w_A)$ and $\mathcal{N}_B = (\mathcal{K}, \mathcal{J}, \mathcal{E}_B, w_B)$ with corresponding matrices $\mathbf{A}_{\mathcal{I}\times\mathcal{K}}$ and $\mathbf{B}_{\mathcal{K}\times\mathcal{J}}$ we call a *product of networks* $\mathcal{N}_A$ and $\mathcal{N}_B$ a network $\mathcal{N}_A \star \mathcal{N}_B = \mathcal{N}_C = (\mathcal{I}, \mathcal{J}, \mathcal{E}_C, w_C)$, where $\mathcal{E}_C = \{(i, j)\colon i \in \mathcal{I}, j \in \mathcal{J}, c_{i,j} \neq 0\}$ and $w_C(i, j) = c_{i,j}$ for $(i, j) \in \mathcal{E}_C$. The product matrix $\mathbf{C} = [c_{i,j}]_{\mathcal{I}\times\mathcal{J}} = \mathbf{AB}$ is defined in the standard way

$$c_{i,j} = \sum_{k\in\mathcal{K}} a_{i,k} \cdot b_{k,j} \,.$$

In the case when $\mathcal{I} = \mathcal{K} = \mathcal{J}$ we are dealing with ordinary one-mode networks (with square matrices).

The standard matrix multiplication is too slow to be used for large networks. For sparse large networks, we can multiply much faster considering only nonzero elements. In general the multiplication of large sparse networks is a 'dangerous' operation since the result can 'explode' – it is not sparse. But in many interesting cases, we can assure that also the product is sparse. For example, we can prove:

If at least one of the sparse networks $\mathcal{N}_A$ and $\mathcal{N}_B$ has small maximal degree on K then also the resulting product network $\mathcal{N}_C$ is sparse.

A more detailed analysis gives: Let for $k \in \mathcal{K}$ be $d_{\min}(k) = \min(\deg_A(k), \deg_B(k))$, $\Delta_{\min} = \max_{k\in\mathcal{K}} d_{\min}(k)$, $d_{\max}(k) = \max(\deg_A(k), \deg_B(k))$, $\mathcal{K}(d) = \{k \in \mathcal{K}\colon d_{\max}(k) \geq d\}$, and $d^* = \operatorname{argmin}_d (|\mathcal{K}(d)| \leq d)$. If for the sparse networks $\mathcal{N}_A$ and $\mathcal{N}_B$ the quantities $\Delta_{\min}$ and d^* are small then also the resulting product network $\mathcal{N}_C$ is sparse.

For example, using network multiplication we can in a given genealogy from the basic relations (P – parent-of, L – is a man, J – is a woman) compute all other kinship relations. For details see [12].

An important application of network multiplication is conversion of two-mode network to the corresponding one-mode networks. Often we transform a two-mode network $\mathcal{N}$ into an ordinary (one-mode) network $\mathcal{N}_1 = (\mathcal{I}, \mathcal{E}_1, w_1)$ or/and $\mathcal{N}_2 = (\mathcal{J}, \mathcal{E}_2, w_2)$, where $\mathcal{E}_1$ and w_1 are determined by the matrix $\mathbf{W}^{(1)} = \mathbf{W}\mathbf{W}^{\mathrm{T}}$, $w_{ij}^{(1)} = \sum_{k\in\mathcal{J}} w_{ik} \cdot w_{kj}^{\mathrm{T}}$ and $\mathbf{W}^{\mathrm{T}}$ is the transpose of matrix $\mathbf{W}$. Evidently the matrix $\mathbf{W}^{(1)}$ is symmetric $w_{ij}^{(1)} = w_{ji}^{(1)}$. There is an edge $(i, j) \in \mathcal{E}_1$ in $\mathcal{N}_1$, iff $N(i) \cap N(j) \neq \emptyset$. Its weight is $w_1(i, j) = w_{ij}^{(1)}$. The network $\mathcal{N}_2$ is determined in a similar way by the matrix $\mathbf{W}^{(2)} = \mathbf{W}^{\mathrm{T}}\mathbf{W}$.

The networks $\mathcal{N}_1$ and $\mathcal{N}_2$ are analyzed using standard methods for one-mode networks.

Another very important application of network multiplication is producing different networks from data tables. A *data table* $\mathcal{T}$ is a set of *records* $\mathcal{T} = \{T_k\colon k \in \mathcal{K}\}$,

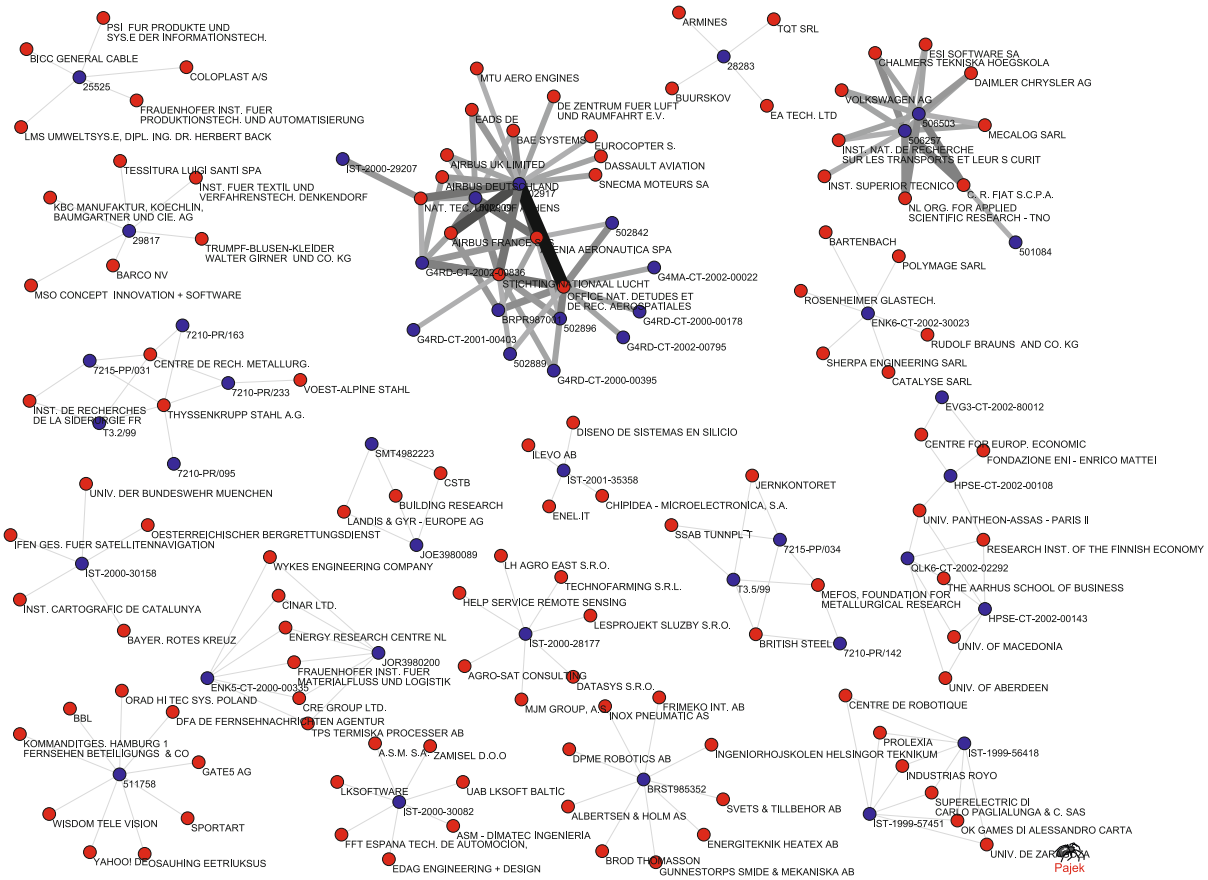

Social Network Analysis, Large-Scale, Figure 21
The main two islands in ProjInst

where $\mathcal{K}$ is the set of *keys*. A record has the form $T_k = (k, q_1(k), q_2(k), \ldots, q_r(k))$ where $q_i(k)$ is the value of the *property* (attribute) $\mathbf{q}_i$ for the key k.

Suppose that the property $\mathbf{q}$ has the range $2^{\mathcal{Q}}$. For example: Authors[WasFau] = {S. Wasserman, K. Faust}, PubYear [WasFau] = {1994}, … If $\mathcal{Q}$ is finite (it can always be transformed in such set by partitioning the set $\mathcal{Q}$ and recoding the values) we can assign to the property $\mathbf{q}$ a two-mode network $\mathcal{K} \times \mathbf{q} = (\mathcal{K}, \mathcal{Q}, \mathcal{E}, w)$ where $(k, v) \in \mathcal{E}$ iff $v \in q(k)$, and $w(k, v) = 1$.

Also, for properties $\mathbf{q}_i$ and $\mathbf{q}_j$ we can define a two-mode network $\mathbf{q}_i \times \mathbf{q}_j = (\mathcal{Q}_i, \mathcal{Q}_j, \mathcal{E}, w)$ where $(u, v) \in \mathcal{E}$ iff $\exists k \in \mathcal{K}\colon (u \in q_i(k) \wedge v \in q_j(k))$, and $w(u, v) = \mathrm{card}\big(\{k \in \mathcal{K}\colon (u \in q_i(k) \wedge v \in q_j(k))\}\big)$.

It holds $[\mathbf{q}_i \times \mathbf{q}_j]^{\mathrm{T}} = \mathbf{q}_j \times \mathbf{q}_i$ and $\mathbf{q}_i \times \mathbf{q}_j = [\mathcal{K} \times \mathbf{q}_i]^{\mathrm{T}} \star [\mathcal{K} \times \mathbf{q}_j] = [\mathbf{q}_i \times \mathcal{K}] \star [\mathcal{K} \times \mathbf{q}_j]$.

We can join a pair of properties $\mathbf{q}_i$ and $\mathbf{q}_j$ also with respect to the third property $\mathbf{q}_s$: we get a two-mode network $[\mathbf{q}_i \times \mathbf{q}_j]/\mathbf{q}_s = [\mathbf{q}_i \times \mathbf{q}_s] \star [\mathbf{q}_s \times \mathbf{q}_j]$.

For the meeting ***The Age of Simulation*** at Ars Electronica in Linz, January 2006, a dataset of EU projects on simulation was collected by FAS research, Vienna and stored in the form of Excel table. The rows are the descriptions of projects participants (idents) and columns correspond to different their properties. From this table three two-mode networks were produced: Project – $\mathbf{P}$ = [idents × projects]; Country – $\mathbf{C}$ = [idents × countries]; and Institution – $\mathbf{U}$ = [idents × institutions]; where $|idents| = 8869$, $|projects| = 933$, $|institutions| = 3438$, and $|countries| = 60$.

Since all three networks have the common set (idents) we can derive from them using network multiplication several interesting networks, such as: ProjInst – $\mathbf{W}$ = [projects × institutions] = $\mathbf{P}^{\mathrm{T}} \star \mathbf{U}$; Countries – $\mathbf{S}$ = [countries × countries] = $\mathbf{C}^{\mathrm{T}} \star \mathbf{C}$; and Institutions – $\mathbf{Q}$ = [institutions × institutions]/projects = $\mathbf{W}^{\mathrm{T}} \star \mathbf{W}$.

For identifying important parts of ProjInst network the 4-rings weights were computed and in the obtained

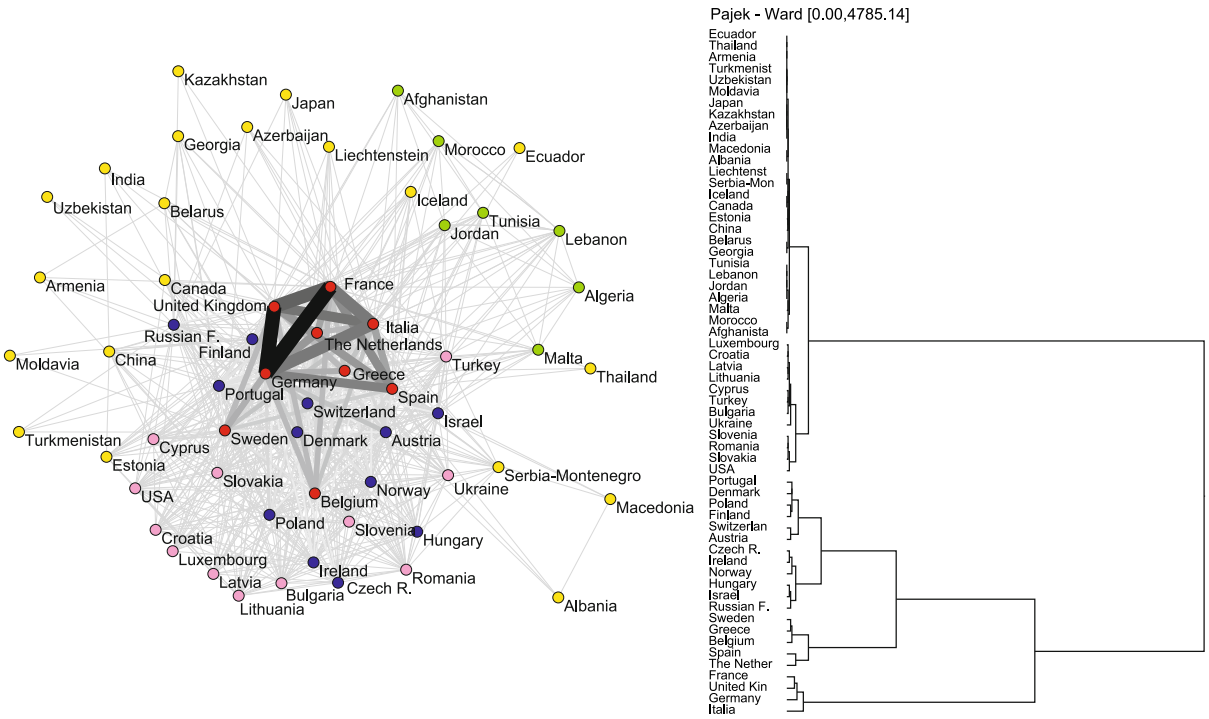

Social Network Analysis, Large-Scale, Figure 22
Collaboration among countries – graph and hierarchical clustering

network the line islands were determined. 101 islands were obtained, 18 of the size at least 5 (see Fig. 21). The two most important islands are: aviation companies and car companies.

In Fig. 22 the collaboration among countries is presented. For dense (sub)-networks we get better visualization by using matrix display. To determine the ordering of vertices we used Ward's clustering procedure with corrected Euclidean distance as dissimilarity measure [21]. The permutation determined by hierarchy can often be improved by changing the positions of clusters in the clustering tree. We get a typical center-periphery structure (see Fig. 23).

Note that in matrix display some details become apparent, such as the collaboration inside the peripherical group Afghanistan, Morocco, Malta, Tunisia, Lebanon, Jordan and Algeria; or a collaboration of Russian Federation with ex-Soviet republics Turkmenistan, Uzbekistan, Moldavia, Kazakhstan, Azerbaijan and Japan.

Statistical Approach

There are many properties *computed* from the network data that give us different information about it. For example:

Global properties Number of vertices, lines (edges/arcs), components; diameter; centralization; maximum core number, …

Local properties Degrees, core numbers, indices (betweeness, hubs, authorities, …). Usually we look at their *distributions* or *inspect* the values of interesting elements.

Another interesting task is searching for associations between computed (structural) data and input (measured) data.

Paul Erdős and Alfréd Rényi introduced in 1959 the notion of random graph in which each pair of vertices is linked with a given probability p. The theory of ER random graphs is well developed (see [15]). Some characteristic results:

- The degree distribution is binomial (in the limit Poisson's) and most of the vertices have degree (very) close to the average degree;
- For $p \geq \frac{1}{n}$ cycles appear in the graph, and soon also the *giant component*;
- For $p \geq \frac{\log_2 n}{n}$ almost all graphs are connected.

Real-life networks are usually not random in the Erdős–Rényi sense. The analysis of their distributions gave new views about their structure.

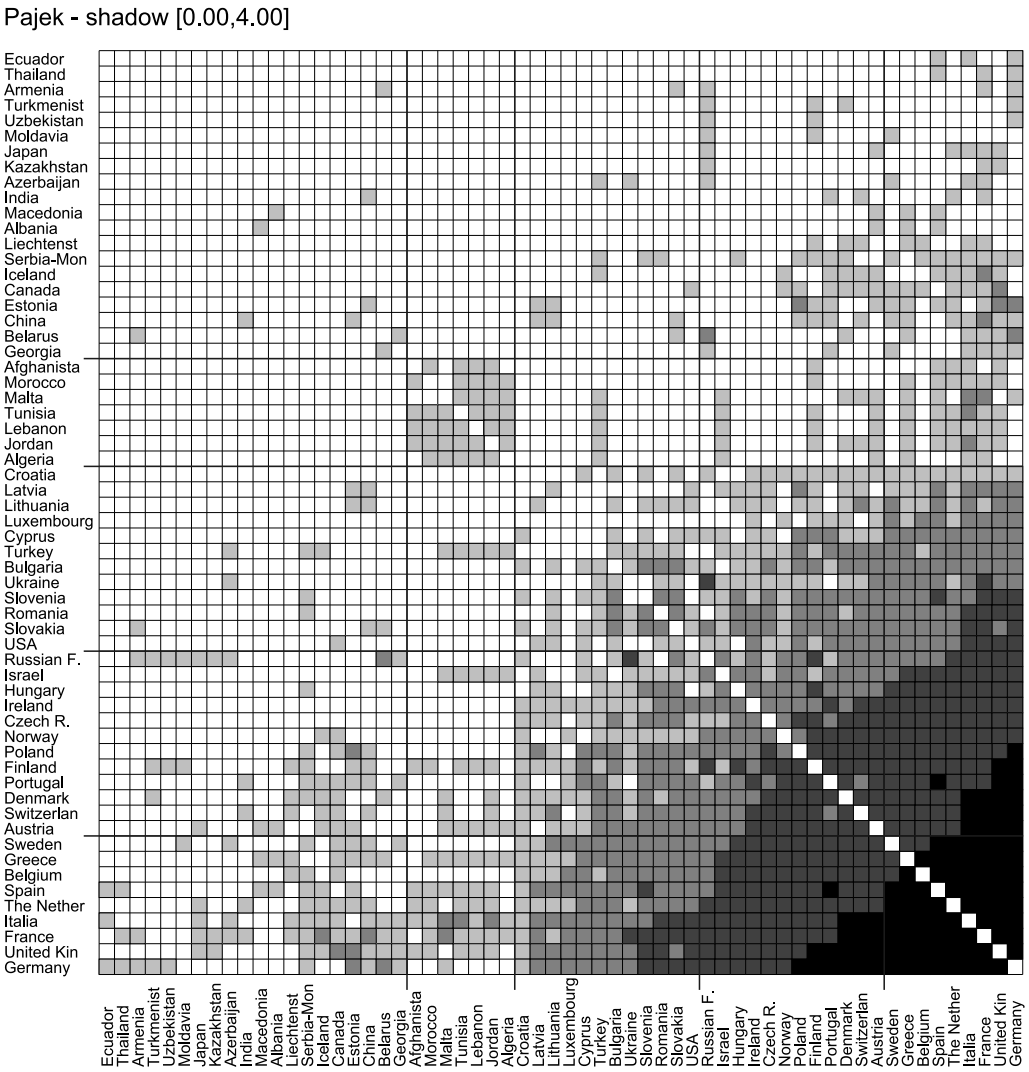

Social Network Analysis, Large-Scale, Figure 23
Matrix display of collaboration among countries

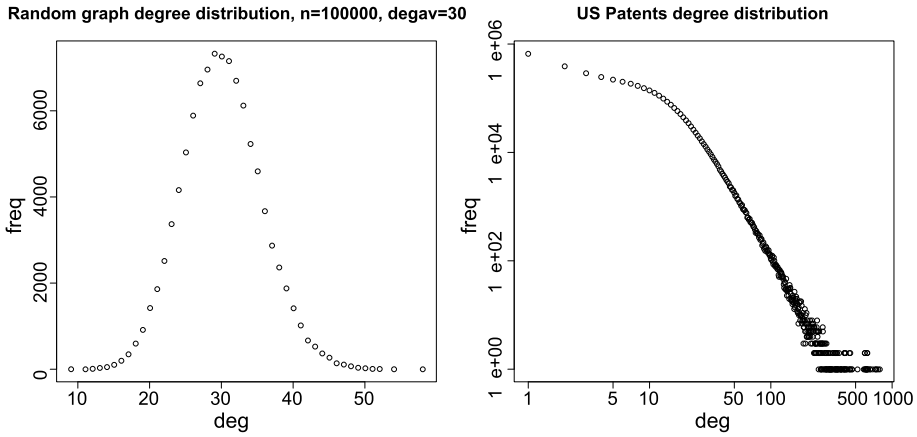

Social Network Analysis, Large-Scale, Figure 24
Distributions: ER-random and US patents

On the left side of Fig. 24 a degree distribution in ER graph on $n = 100{,}000$ vertices with average degree $\overline{\deg} = 30$ is presented. On the right side a degree distribution for US patents citation network is presented (in log–log scale). Evidently this distribution is very far away from Poisson distribution.

In 1967 a psychologist Stanley Milgram made his experiment with letters. The letter should reach a target person. The persons involved in experiment were asked to send the letter with these instructions to his or her acquaintance that is supposed to be closer (in the acquaintances network) to the target person. The letter was sent from Boston to Omaha. The average length of the successful paths was six – *six degrees of separation*. The average path length on the internet is 19 clicks.

The networks in which the average shortest path length is small are called *small worlds*. Duncan Watts and Steven Strogatz developed in late 90-ties a procedure for construction of (random) small worlds by *rewiring* – an edge is randomly selected and one of its endpoints is attached to same other vertex. After each rewiring step the average length of geodesics is usually decreased because the rewiring creates shortcuts.

Albert-László Barabási from University of Notre Dame in 1998 analyzed several networks and noticed:

- The degree distribution follows the *power law* – the probability p_d that a vertex has a degree d equals to $p_d = cd^{-\gamma}$. In a log–log scale diagram it is represented by a line.
- In a network there exist some vertices with *large degree* (very improbable in ER graphs). These vertices link the network into a single component.

It turned out that most of real life networks (persons – e-mail, phone calls, sexual contacts (drug users, AIDS), collaboration; movie actors – playing in the same movie; proteins – interactions; words – semantic relations; ...) have such characteristics. Because for these networks their degree distribution has no natural scale they were named *scale-free* networks. For a discussion about the notion of scale-free network see [33].

The following was the first explanation (Barabási) of scale-free nature of many real-life networks:

- These networks are growing.
- In this process new vertices are added and linked with new edges to already existing vertices. The random selection of vertex to which a new vertex is attached is not uniform but follows the *preferential attachment* rule – the selection probability is proportional to the degree of a vertex.

Based on this model it can be shown the following:

- The degree distribution is the power law.
- The average length of geodesics is $O(\log n)$.
- These networks are resilient against random vertex or edge removals (random attacks), but quickly become disconnected when large degree nodes (Achilles' heel) are removed (targeted attacks).

Mark Granovetter noticed in 1973 that in social networks groups appear linked with *strong ties* [26]. They link in larger networks with *weak ties*. Also in other real-life networks vertices often form groups – the clustering coefficient is larger than in ER networks.

Several improvements and alternative models were proposed that also produce scale-free networks with some additional properties characteristic for real-life networks: copying [31], combining random and preferential attachment [37], R-mat [18], forest fire [32], aging, fitness, nonlinear preferences, ...

There are several applications of the scale-free networks theory. For example searching (Adamic et al. [2]) and spreading of epidemies (Barthélemy, Barrat, Pastor-Sattoras, Vespignani, Complex Networks Collaboratory).

For general overviews see [4,22,35,36].

Future Directions

In 2005 the support for *multi-relational* networks was introduced in Pajek. Combined with *temporal* networks it enables analysis of new kinds of networks – such as KEDS networks (*Kansas Event Data System* or *Tabari*). These networks are usually small in terms of vertices but can be (very) large in terms of lines – different interaction events among actors.

The last developed approach for analysis of large networks is adaptation of hierarchical clustering with relational constraints based on Ferligoj and Batagelj [23] to large networks. The basic idea to get a fast algorithm is to compute the dissimilarities between units (vertices) only for the linked pairs of units [11]. This approach is one of the possible approaches to analysis of spatial networks.

There are still several fields of social network analysis for which efficient approaches to deal with large networks have to be developed such as blockmodeling, probabilistic models, ...

In the near future new versions of network analysis software will appear using very large computer memories enabled by the new 64-bit computer architecture. A special challenge is development of methods and software for analysis of huge networks.

Bibliography

Primary Literature

1. Abello J, Pardalos PM, Resende MG (2002) Handbook of Massive Data Sets. Springer, Heidelberg
2. Adamic LA, Lukose RM, Huberman BA (2002) Local Search in Unstructured Networks. In: Bornholdt S, Schuster HG (eds) Handbook of Graphs and Networks: From the Genome to the Internet. Wiley-VCH, Berlin
3. Ahmed A, Batagelj V, Fu X, Hong SH, Merrick D, Mrvar A (2007) Visualisation and Analysis of the Internet Movie Database. Proceedings of the Asia-Pacific Symposium on Visualisation (APVIS2007), Sydney, Australia, 5–7 Feb. IEEE, New York, pp 17–24
4. Albert R, Barabási AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
5. Alvarez-Hamelin JI, Dall'Asta L, Barrat A, Vespignani A (2005) *k*-core decomposition: a tool for the visualization of large scale networks. cs.NI/0504107 published in Advances in Neural Information Processing Systems 18, Canada, 2006
6. Batagelj V (1989) Similarity measures between structured objects. In: Graovac A (ed) Proceedings of International Course and Conference on the Interfaces between Mathematics, Chemistry and Computer Science, Dubrovnik, 20–25 June 1988. Studies in Physical and Theoretical Chemistry, vol 63. Elsevier/North-Holland, Amsterdam, pp 25–40
7. Batagelj V, Brandes U (2005) Efficient Generation of Large Random Networks. Phys Rev E 71:036113
8. Batagelj V, Ferligoj A (2000) Clustering relational data. In: Gaul W, Opitz O, Schader M (eds) Data Analysis. Springer, Berlin, pp 3–15
9. Batagelj V, Mrvar A (2000) Some Analyses of Erdős Collaboration Graph. Soc Netw 22:173–186
10. Batagelj V, Mrvar A (2001) A Subquadratic Triad Census Algorithm for Large Sparse Networks with Small Maximum Degree. Soc Netw 23:237–43
11. Batagelj V, Mrvar A (2007) Hierarchical clustering with relational constraints of large data sets. 6th Slovenian International Conference on Graph Theory, Bled, 24–30 June
12. Batagelj V, Mrvar A (2008) Analysis of kinship relations with Pajek. Soc Sci Comput Rev 26(2):224–246
13. Batagelj V, Zaveršnik M (2002) Generalized Cores. arxiv cs.DS/0202039
14. Batagelj V, Zaveršnik M (2007) Short cycle connectivity. Discret Math 307(3–5):310–318
15. Bollobás B (2001) Random Graphs. Cambridge University Press, Cambridge
16. Brandes U (2001) A Faster Algorithm for Betweenness Centrality. J Math Soc 25(2):163–177
17. Breiger RL (2004) The analysis of social networks. In: Hardy M, Bryman A (eds) Handbook of data analysis. Sage, London, pp 505–526
18. Chakrabarti D, Zhan Y, Faloutsos C (2004) R-MAT: A Recursive Model for Graph Mining. In: SIAM Data Mining 2004, Orlando, Florida, SIAM
19. Cormen TH, Leiserson CE, Rivest RL, Stein C (2001) Introduction to algorithms. MIT Press, Cambridge
20. Doreian P, Batagelj V, Ferligoj A (2000) Symmetric-Acyclic Decompositions of Networks. J Classif 17(1):3–28
21. Doreian P, Batagelj V, Ferligoj A (2005) Generalized Blockmodeling. Cambridge University Press, Cambridge
22. Dorogovtsev SN, Mendes JFF (2003) Evolution of networks: from biological nets to the internet and www. Oxford University Press, Oxford
23. Ferligoj A, Batagelj V (1983) Some types of clustering with relational constraints. Psychometrika 48(4):541–552
24. Freeman LC (1979) Centrality in Social Networks: A Conceptual Clarification. Soc Netw 1:211–213
25. Garfield E, Sher IH, Torpie RJ (1964) The Use of Citation Data in Writing the History of Science. The Institute for Scientific Information, Philadelphia
26. Granovetter M (1973) The Strength of Weak Ties. Am J Sociol 78:1360–80
27. Harary F, Norman RZ, Cartwright D (1965) Structural Models: An Introduction to the Theory of Directed Graphs. Wiley, New York
28. Huisman M, van Duijn MAJ (2005) Software for social network analysis. In: Carrington PJ, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, Cambridge, pp 270–316
29. Hummon NP, Doreian P (1990) Computational Methods for Social Network Analysis. Soc Netw 12:273–288
30. Kleinberg J (1998) Authoritative sources in a hyperlinked environment. Proc 9th ACM-SIAM Symposium on Discrete Algorithms
31. Kleinberg J, Kumar R, Raghavan P, Rajagopalan S, Tomkins A (1999) The Web as a graph: measurements, models and methods. Proc of the 5th International Computing and combinatorics Conference
32. Leskovec J, Kleinberg J, Faloutsos C (2006) Laws of Graph Evolution: Densification and Shrinking Diameters. ACM Transactions on Knowledge Discovery from Data (TKDD) vol 1, issue 1, article 2
33. Li L, Alderson D, Tanaka R, Doyle JC, Willinger W (2007) Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications. cond-mat/0501169, Internet Math 2(4):431–523
34. Mane KK, Börner K (2004) Mapping topics and topic bursts in PNAS. Proc Natl Acad Sci USA 101:5287–5290
35. Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45:167–256
36. Newman MEJ, Barabási AL, Watts D (2006) The Structure and Dynamics of Networks. Princeton Studies in Complexity. Princeton University Press, Princeton
37. Pennock DM, Flake GW, Lawrence S, Glover EJ, Giles CL (2002) Winners don't take all: Characterizing the competition for links on the web. Proc Natl Acad Sci USA 99(8):5207–5211
38. Seidman SB (1983) Network Structure And Minimum Degree. Soc Netw 5:269–287
39. Schank T, Wagner D (2005) Finding, counting and listing all triangles in large graphs, an experimental study. In: Workshop on Experimental and Efficient Algorithms (WEA). Lecture Notes in Computer Science, vol 3503, Springer, pp 606–609
40. Snyder D, Kick E (1979) The World System and World Trade: An Empirical Exploration of Conceptual Conflicts. Sociol Q 20(1):23–36
41. Snijders TAB (2005) Models for Longitudinal Network Data. In: Carrington P, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, New York

42. Stuckenschmidt H, Klein M (2004) Structure-Based Partitioning of Large Concept Hierarchies. Proc of the 3rd International Semantic Web Conference ISWC 2004, Hiroshima, Japan
43. Wasserman S, Faust K (1994) Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge
44. White DR, Batagelj V, Mrvar A (1999) Analyzing Large Kinship and Marriage Networks with Pgraph and Pajek. Soc Sci Comput Rev 17:245–274
45. Zaveršnik M, Batagelj V (2004) Islands. Slides from Sunbelt XXIV, Portorož, Slovenia, 12–16 May

Books and Reviews

Ahuja RK, Magnanti TL, Orlin JB (1993) Network Flows: Theory, Algorithms, and Applications. Prentice Hall, Englewood Cliffs
Batagelj V, Mrvar A (2003) Pajek – Analysis and Visualization of Large Networks. In: Jünger M, Mutzel P (eds) Graph Drawing Software. Springer, Berlin, pp 77–103
Brandes U, Erlebach T (2005) Network Analysis: Methodological Foundations. Lecture Notes in Computer Science. Springer, Berlin
Carrington PJ, Scott J, Wasserman S (2005) Models and Methods in Social Network Analysis. Cambridge University Press, Cambridge
Degenne A, Forsé M (1999) Introducing Social Networks. SAGE Publications, London
de Nooy W, Mrvar A, Batagelj V (2005) Exploratory Social Network Analysis with Pajek. Cambridge University Press, Cambridge
Knuth DE (1993) The Stanford GraphBase: A Platform for Combinatorial Computing. Addison-Wesley, Reading
Scott JP (2000) Social Network Analysis: A Handbook. SAGE Publications, London

Web Resources

Center for Complex Network Research, Notre Dame: http://www.nd.edu/networks/
Center for Spatially Integrated Social Science: http://www.csiss.org/
Complex Networks Collaboratory: http://cxnets.googlepages.com/
Internet Movie Database http://www.imdb.com/
Matthieu Latapy. Triangle computation web page. http://www-rp.lip6.fr/latapy/Triangles/
Nber: http://www.nber.org/patents/
Netminer: http://www.netminer.com/
Pajek: http://pajek.imfm.si data sets: http://vlado.fmf.uni-lj.si/pub/networks/data/
The Edinburgh Associative Thesaurus: http://www.eat.rl.ac.uk/
The Kansas Event Data System: http://web.ku.edu/keds/
UCINET: http://www.analytictech.com/

Social Network Analysis, Overview of

John Scott
School of Law and Social Science,
University of Plymouth, Plymouth, UK

Article Outline

Glossary
Definition of the Subject
The Development of Social Network Analysis
Graph Theory and Ideas of Balance
Diffusion Processes
Algebraic Models and Blockmodeling
Scaling models and Visualization
Statistical Models for Hypothesis Testing
Agent-Based Computational Models and Temporal Processes
Small World Models and Network Dynamics
Bibliography

Glossary

Algebraic models Approaches to network analysis that use algebraic methods for studying sets of points to produce positional analyses such as blockmodels.

Agent-based computational models Models of dynamic systems that focus on agent-level properties, seeing system level changes as consequences of the interaction of rule-following agents.

Balance theories A number of related theories concerning both the psychological state of consonance or dissonance found in a person's ideas and affects and the equilibrium or disequilibrium of one person's relations with another. The exploration of the mathematical principles of these states.

Blockmodeling An approach to positional analysis that uses algebraic methods to construct image graphs in which blocks represent connections or the absence of a connection between sets of points.

Graph theory A basic method for the analysis of networks in which the relational properties of the members of a set are seen in terms of points and lines. Pairwise connections among points are used to generate and explore system-level phenomena such as density and centralization.

Diffusion processes The processes through which innovations and other changes spread through a network, the flow and pace of change being determined by the structure of the network.

Scaling methods Geometrical techniques for displaying and analyzing a network as a mapping of points located in a multidimensional space.

Small-world models Network models based on graphs of relatively low density but high reachability. Originated in psychological experiments on the communicative effectiveness of interpersonal acquaintance.

Definition of the Subject

The formal idea of using a network as a tool of analysis originated in electrical engineering as a way of envisaging and modeling the flows of electricity through national power grids and in domestic and industrial settings. It was merely a short step to the application of the idea in civil engineering, where water supply, drainage, roads and railways could all be regarded as networks. Much later this idea was central to electronic and computer systems engineering. Parallel methodological innovations had occurred in areas of physical geography, where river drainage systems had been envisaged as networks of flows that could be modeled in mathematical terms, and this led to interest among human geographers about the potential for applying the idea. The idea that social relationships could be seen as forming a network or reticule of relations that are capable of analysis in formal, mathematical terms suggested novel ways of considering social phenomena and promised insights that were unavailable through alternative approaches.

Social network analysis first emerged as an approach distinctive from statistical analysis within anthropology, social psychology, and sociology. It went on to influence theoretical and methodological developments in many other areas of social sciences. In all these areas, researchers took up mathematical models from graph theory, which they used to investigate structures of relations. Researchers who were working in the area of social network analysis soon began to broaden their framework of analysis by exploring the implications of other mathematical models, such as algebraic set theory, scaling methods, and statistical and stochastic methods. In this way, social network analysis became linked to the larger intellectual debates going on in other areas of the natural and social sciences where these same methods were being drawn upon, and a number of social network analysts have contributed to the development of these and other mathematical models. Social network analysts have promoted and undertaken the application of these ideas in many areas of sociology, most notably in economic sociology, political sociology, the study of family and kinship, community studies, the investigation of world systems, and numerous other areas.

Investigation into social networks has now become far more fashionable than ever before. However, it is in the nature of intellectual fashions that much of the new interest comes from those who are least familiar with the area. This is certainly true of social network analysis, where many newcomers were unaware of its long history. They were influenced, in particular, by the exaggerated claims to intellectual novelty that were being made by some physicists who have advocated random-network and small-world models as the basis of what they call a "new science of networks". Indeed, many people have assumed that social network analysis began de novo with these models and they know little about the applications already made using other models. Despite their exaggerated claims, the work of the physicists has provided us with some powerful new models, and these have had a major impact beyond the specifically social sphere – this fact will be apparent from considering chapters in a number of the sections in this *Encyclopedia*. However, the physicists have not provided a complete replacement for social network analysis, as they sometimes imply, but have offered a broadening of its range and concerns.

This overview article will discuss the development of social network analysis and will sketch the various mathematical models that have been used to explore the nature of complex social systems. It begins with an overview of the history of social network analysis and its main applications. This is followed by a consideration of graph theory, the earliest and principal approach to the study of social networks. The following sections look at later and more advanced methods: algebraic methods and blockmodeling, scaling methods and visualization techniques, statistical methods, agent-based computational methods, and finally the small-world methods of the physicists. The latter discussion will place the new social physics in the context of earlier discussions and it will show the variety of mathematical models that have a part to play in contemporary social network analysis. The various methods and models are considered in outline and without technical detail. The various specialist chapters included in this and other sections of the Encyclopedia cover these methods in greater detail.

The Development of Social Network Analysis

Social network analysis has a long history, though it appeared under its specific name only far more recently (see Chap. 2, p. 23 in [51]). Although structural thinking was central to classical sociology, the intellectual roots of a specifically network perspective can be traced back to those sociologists of the late nineteenth and early twentieth centuries who used the idea of a network of relations as a metaphor for understanding social relations as *structures* of relations. Georg Simmel and other German sociologists, in particular, wrote extensively on the "interweaving" of actions to form complex configurations or "webs" of connection. Their metaphorical usage of the idea drew explicitly from the textile industry and its ideas of the weft and warp of woven fabrics. The first proper usage of the

actual word "network" in a sociological context, however, was in the work of the anthropologist Alfred Radcliffe-Brown [45], who used it to suggest a more formal focus for structural analysis. Simultaneously with this, a number of social psychologists working in the area of child study and field theory began to construct diagrams of social relations that they gradually formalized as "sociograms" to show the chains of positive and negative connections that they were finding among the individuals in their experimental studies.

The crucial social psychological work on social networks was that of the psychotherapist Jacob Moreno [41]. Following his arrival in the United States he began to combine psychodrama with studies of the influence of friendship choices on personality development among schoolchildren. His key innovation was to make use of sociograms to depict the friendship patterns that he was able to discover within small groups such as a school class or a neighborhood peer group. Moreno made many intellectual contacts, most importantly with the *Gestalt* psychologist Kurt Lewin, who had himself recently migrated to the United States. Moreno and Lewin exchanged many ideas – and the nature of the exchange became a matter of contention between them in later years. Lewin soon began to move towards a more social form of psychology that stressed the importance of looking at organized patterns of group relations in particular spheres of activity as forming systemic wholes. Such wholes could be seen, by analogy with electromagnetic and gravitational fields, as social fields through which influences are able to flow from one area or part to another. The aim of Lewin's field theory was to map and to explore the "dynamics" of these fields of interpersonal influence.

Both Moreno and Lewin used diagrammatic representations of social relations, but it was Moreno who invented the classic form of the sociogram. A sociogram consists of points, representing individuals, and lines, representing the social relationships among them. Moreno argued that these patterns of points and lines – later described more formally as "vertices" and "edges" – could be described as "social configurations" in which could be charted such things as the direction of friendship choices and their intensity. This work came to be referred to as "sociometry", and a journal with that name was founded in 1937. Through the work published in this journal, the ideas of Moreno and Lewin on configurations of social relations in small groups gradually gave rise to a more mathematically oriented approach that went under the name of "group dynamics". It was in this school of thought that ideas from graph theory – first formulated by René König [33] – began to be employed from the late 1940s, these ideas becoming the basis of a more formal and rigorous study of the properties of the social networks of small groups. By treating the sociogram as a graph, it proved possible to chart dyadic, triadic, and more complex structures and to analyze the states of "balance" and imbalance that existed among positive and negative relations within each network [3,13,28].

It was during the 1930s and 1940s that anthropological and sociological work also began to move in a similar direction. In 1929, the anthropologist W. Lloyd Warner, a former colleague of Radcliffe-Brown, had moved to Harvard University in order to work with the psychologist Elton Mayo on a series of experimental studies of work behavior. Their aim was to use the techniques already employed by anthropologists in studies of small tribes and villages, but to carry out new studies of small groups in the factory and office settings of urban societies. Influenced also by the early system theory constructed by the physiologists Lawrence Henderson and Walter Cannon, they began a series of empirical studies at the Hawthorne electrical works in Chicago. The so-called Hawthorne studies became a milestone in social research – not least because of their significance in developing the idea of the social network.

The Hawthorne studies [47] had begun as management-run investigations into worker efficiency and its relationship to the physical conditions of work. The early experiments examined such things as the consequences of changes in levels of heating and lighting and the provision of work breaks. The studies turned from physical to social factors under Warner and Mayo, and Mayo famously reported that the crucial factor in increasing worker productivity had been their involvement in the experiment itself. The paradoxical results showing increased productivity even when physical conditions worsened were interpreted as showing the high levels of satisfaction that workers felt simply from being part of the management experiment: it was the first time that anyone from management had shown any direct interest in them. Adopting more anthropological methods of observation in the bank wiring room, the researchers documented the gradual building of informal work group relations and forms of social solidarity among those in the experimental group. These informal relations ran in parallel with the formal relations of the managerial organization chart, and the researchers used sociograms to chart the informal social networks among workers.

The sociograms drawn by the Hawthorne researchers allowed them to identify "cliques" and other sub-groups within the larger group, and Warner went on to investigate this kind of group structure further in a series of

investigations into community relations in the New England city of Newburyport (referred to as "Yankee City" in [59,60,61,62]), the Mississippi city of Natchez ("Old City" in [15]), and the Indiana city of Morris ("Jonesville" in [58]). In this work, Warner and his colleagues used network ideas to explore the formation of cliques, "crowds", and "circles" among those they studied, and they examined these in relation to the more formal structures of economic and political life. Their investigations identified these social circles through a rough and ready use of the Venn diagrams of set theory and the methods of matrix display. Using these simple methods, they were able to identify such things as class divisions in clique membership and the effects of ethnicity on the structure of network relations.

It was during the late 1940s and early 1950s that a significant consolidation of the social network approaches that had been developing began to be forged. It was to be another decade, however, before a recognizable tradition of social network analysis was properly established. George Homans, a colleague of Warner and Mayo at Harvard, attempted a synthesis of the anthropological and sociological ideas as the basis of an alternative sociology to the system theory of Talcott Parsons [30]. Homans re-analyzed some of the earlier data using rudimentary techniques of matrix rearrangement, his aim being to bring out the clique structure more clearly. Homans' method of manually "reshuffling" the 0s and 1s in a matrix to bring out blocks of high and low density was a pioneering move towards the more formal methods of blockmodeling that developed during the 1970s. Homans himself interpreted his data on social relations in terms of the frequency and duration of the relationships involved and went on to develop rational choice models of action to explain the findings [31].

The early 1950s also saw the development of more systematic anthropological work on social networks in Britain. The Manchester University anthropologist Max Gluckman established the Rhodes-Livingstone Institute in Northern Rhodesia (now Zambia) to study African social structures in village and town settings. An early associate of the Institute was John Barnes, who undertook independent fieldwork in Bremnes, Norway. It was here that he recognized the importance of the network metaphor – famously while observing the repaired fishing nets of the local fishermen – and he advocated the study of whole communities as networks of points and lines [2]. Elizabeth Bott, a student of Lloyd Warner and then working at the Tavistock Institute (associated with the Michigan Research Center for Group Dynamics), took up this metaphor in her own investigations into the kinship networks of London families [6,7]. Because of her position at the Tavistock, she was encouraged to link this to the work of Moreno and to the early field theory. This work was systematized by Siegfried Nadel and Clyde Mitchell. Nadel [42] aimed to establish a mathematical basis for social network research, but his early death prevented him from carrying this further. Mitchell [39], however, made the network idea the basis for a series of studies at the Rhodes-Livingstone Institute and set out a number of conceptual innovations – density, multiplexity, durability, direction, reachability and so forth – that echoed and took further the ideas of Homans.

By 1969, then, a clear body of network ideas had been established. in anthropology and social psychology, and this was coming to be organized around the mathematical theory of graphs. It was at this point, however, that an alternative framework began to develop in the United States, and this eventually subsumed the earlier approach in a comprehensive framework of social network analysis. John Boyd, François Lorrain, and Harrison White had begun to explore some aspects of Lévi-Strauss's analysis of kinship structures using algebraic set theory [72], while Edward Laumann [34] and Joel Levine [35] began to use multidimensional scaling to investigate social structure. White moved to Harvard University and brought together a strikingly original group of young researchers who were interested in using mathematical methods to explore aspects of social networks and social relations.

One of the earliest products of this group was Mark Granovetter's [26,27] influential argument about "the strength of weak ties" – the first, and most important, of the counter-intuitive results to come out of this more sophisticated form of network analysis. Granovetter's argument held that the chances that a person will receive useful information from those in his or her social network is greater when a person has acquaintances rather than close and intimate friends. One's intimates tend to have access to the same pool of information through their dense pattern of mutual connections. Acquaintances, on the other hand are connected to quite diverse and distant social networks and so can communicate a much greater range of information. Paradoxically, then, it is the weak acquaintanceship ties that are the "strongest" or most important carriers of information. Having a small number of acquaintances can be more useful in a job search than having a large number of intimates.

The ability of White's group to pursue this program of social network analysis was made possible by the advances that were being made in computing technology and that were then becoming available in software programs. Earlier work had used manual techniques and so

had been limited to relatively small groups. Mainframe computers and Fortran programming made possible the analysis of much larger data sets and the use of more powerful mathematics. As the members of the Harvard group spread to Graduate centers across North America, their work became a major influence within sociology and encouraged the global spread of formal social network analysis. Barry Wellman and Steven Berkowitz were particularly active in setting up the *International Network for Social Network Analysis* (http://www.insna.org/) and this became the basis for the founding of the journal *Social Networks* and for their own powerful syntheses in social network analysis [4,71] (see also [32]). Powerful methodological ideas and sociological applications were developed in the United States by Ronald Breiger, Ronald Burt [10,11], Patrick Doreian, and others. Social network analysis was pursued in Canada by Barry Wellman and Peter Carrington, by John Scott in Britain [51], by Rob Mokken and Frans Stokman in the Netherlands (the developers of the *Gradap* program used in [29]), and by Charles Crothers and Malcolm Alexander in Australia. Stanley Wasserman led the way in producing a standard textbook for social network methods [64] and other consolidating texts appeared [24,37]. By the 1990s, social network analysis had become one of the most strongly established research specialisms in Sociology, with a strong intellectual base, advanced methods and texts [12,18], and a comprehensive range of exemplary studies.

During the 1990s, however, a rather paradoxical development occurred. The exhaustion of many of the conventional research programmes in physics had led a number of theoretical physicists to search out new areas in which to apply their theories. In 1998, Duncan Watts and Steven Strogatz published a paper [68] that revisited some of the ideas on random networks that had grown out of Stanley Milgram's work on "small worlds" [38,56]. This caused a great stir among physicists and led Albértó-Laszlo Barabási to propose the building of a "new science of networks" with the potential for applications across the social sciences [1]. In apparent ignorance of the prior existence of social network analysis, he claimed that it was only during the 1990s – and thanks to physicists' models – that investigators had become aware of the fact that social networks have structures and show orderly patterns of development.

This claim startled and dumbfounded most sociologists working on social networks, for whom the existence of social structure was a fundamental axiom. They resented the ignoring of what they had already achieved and the misrepresentation of sociological understandings of the social world that it involved. Despite its shaky historiography, the arguments of Bárabasi and other social physicists had a major influence outside the discipline and among those in sociological specialisms that were, so far, untouched by social network analysis. Works by Duncan Watts were particularly important [66,67]. The new social physics of networks was especially well-received in the popular and scientific press and resulted in a number of popularizing works (see, for example [8]).

It is, nevertheless, true that the work of the physicists has certainly been responsible for a substantial growth of interest in social network analysis during the last decade. The reason for the appeal of this new work lies in its promise to provide a dynamic model of social networks, rather than the more static ones of the past. While the claim that social network analysis lacks a concern for dynamic processes is, perhaps, overstated, it is certainly the case that the new methods do seem to offer the possibility of a nuanced account of structural change in social networks. While much of the work is still presented without any recognition of prior sociological contributions, sociologists, for their part, are beginning to discuss how these ideas might enrich social network analysis. The arguments about network dynamics are, in fact, converging with some of the work on agent-based computational models already being undertaken within Sociology.

A very important issue that has arisen in the use of the mathematics of graphs is that of whether it provides a substantive *theory* or simply a *method* of analysis. Some researchers hold that the mathematical theorems that can be derived from graph theory can be directly translated into substantive theorems concerning particular domains of application. According to this point of view, sociological laws can be derived directly from mathematical laws. Those who see social network analysis as a method – perhaps a majority of those working in the area – hold that mathematical theorems provide simply the formal constraints to which social processes must conform: a knowledge of mathematics does not, in itself, obviate the need to study social phenomena empirically in order to understand the implications of these theorems for particular types of social relations.

The remainder of this overview article will set out the main intellectual strands within contemporary social network analysis. These areas are discussed in greater detail in the specialist articles in this and other sections of the *Encyclopedia*, and their mathematical bases will not be presented in any detail. The overview will concentrate on the general principles involved in each area and on the ways in which they relate to each other.

Graph Theory and Ideas of Balance

Graph theory, as already noted, is that branch of mathematics that had the earliest influence on the development of social network analysis (see ► Social Network Analysis, Graph Theoretical Approaches to). In graph theory, a graph or network is simply a collection of points (vertices) connected by lines (edges). Graph theory comprises a set of mathematical axioms and derivations that describe the actual and potential patterns formed by these points and lines. This basic model of a graph can be applied widely in many substantive domains: electrical wiring networks, river drainage networks, road and transportation networks, and computer networks have all been investigated using graph theory.

A basic graph consists of simple points and lines, but the lines in a graph can be assigned positive or negative attributes, directions, or numerical values, and the attributes of points can be recorded. Complex mathematical theorems can be constructed for each of the different types of graph. Some of the earliest work to use these ideas focused on the question of "balance" in graphs. Balance is a term that describes the pattern of signs and numerical values in a graph in terms of the reciprocity and transitivity of the relations. A graph of friendship relations, for example, might be unbalanced if two people who each like one another have opposite relationships to a third person. If, on the other hand, all three people liked one another with a similar intensity, the graph would be balanced. Balance theorists have analyzed complex graphs in terms of such "triads" of relations. By producing a census of triads and computing the balance or imbalance in each triad, it is possible to arrive at a calculation of the overall balance in the graph as a whole. Such ideas of balance have generally been assimilated to the idea of equilibrium analyzed in classical mechanical models. A graph that is in a state of disequilibrium, therefore, will be marked by tensions and stresses that push and pull it towards a state of equilibrium. Where a social network is mapped as an unbalanced graph, it can be hypothesized that the participants in the network will experience these stresses as push and pull factors on the social relations that they maintain with each other. The overall network can, therefore, be seen as a dynamic field of causal influences.

Much of the work that has been carried out using the idea of balance has considered focal actors to be the points of reference, and much of the early work that applied graph theory pursued such "ego-centric" ideas. From this point of view, the chosen actor is seen as a point that is "adjacent" to certain other points that comprise its immediate "neighborhood". Each point can, therefore, be given an adjacency score. Actors vary in terms of the number of others with whom they are connected – the number of friends, number of enemies, number of political associates, and so on. People or groups may, therefore, be more or less connected into their neighborhood of others and so they may be compared in terms of their adjacency scores. The early work of Moreno had identified school children as sociometric "stars" or "isolates" according to the number of friends they had. The stars were the especially popular class members, while the isolates had few or no friends in the class. The adjacent others may have more or fewer connections among themselves: my friends may or may not be friendly with each other, my business associates may or may not do business with each other, and so on. Thus, a basic ego-centric measure is the "density" of an actor's neighborhood. This measure is the number of links that actually exist among immediate contacts expressed as a proportion of the number of possible contacts that they could, in principle, sustain among themselves. Where a large proportion of connections intersect, the actor's neighborhood has a high density.

Researchers recognized fairly early on the importance of going beyond such ego-centric measures to more global or socio-centric measures of the network as a whole. In considering a village, for example, a researcher may want to know not only the distribution of individuals according to the densities of their friendship relations, but also the density of friendship connections in the village considered as a whole. Such measures of density would indicate the state of social cohesion or social solidarity in the network. The basic ego-centric measure of density can easily be extended to the network as a whole – if appropriate data are available – by computing the proportion of possible links within the village that are actually established. Calculations of density can also be made for large social networks that must be studied through sample investigations. If it can be assumed that the sample is representative, then the distribution of adjacency scores for sample members can be used to compute the overall density of the network.

Density has become one of the most widely used measures in studies that employ graph theory and it has been used to indicate the changing character of social solidarity and the state of "community" that exists in a group. Wellman's work [69,70] has demonstrated that growing amounts of geographical mobility have not led to a commensurate decline in community but have resulted, instead, in the transformation of patterns of solidarity. People in large cities, he shows, are likely to interact less frequently and less intensely with numerous others on a day to day basis as they move to live and work in other parts of the city. Thanks to cars, the telephone, and the internet,

however, they are able to maintain more connections, but on a less frequent and less intense basis.

Particularly important measures of global connectedness within a graph are based on the "distance" from one point to another and the corresponding "centrality" that points have in the overall pattern of connection. The distance between two points in graph theoretical terms is simply the number of lines that must be traversed in moving from one point to another. Thus, two adjacent points stand at distance 1 from each other, while two points that are connected through a common neighbor are at distance 2 from each other. It is generally assumed that influence and communication within a social network attenuates with increasing distance. Nevertheless, the rate of attenuation and the particular level of distance that it is reasonable to regard as constituting a socially significant connection are quite variable from one type of network to another. Distant friends of friends, for example, may be insignificant, while distant financial donors to a political party or candidate may be more significant. The criteria of significance, it will be apparent, rests always on a sociological judgment and never on a mathematical measure alone.

The principal use of distance measures has been to assess the relative centrality or peripherality of points in a graph. A point whose average distance to all other points in the graph is low can be regarded as occupying a central position in the network as a whole, while a point with a very high average distance from all others is peripheral to it. Applications of these ideas in studies of interlocking directorships – the lines created when company directors sit on two or more company boards – have made great use of measures of centrality. They have to documented the existence of "hubs" within large spheres of adjacent companies. Such hubs – typically banks or other financial institutions – are the focal centers for business decision-making and for the allocation of credit at the level of the intercorporate network as a whole [50].

Users of graph theory have often distinguished between measures of centrality based on overall "closeness" and an idea of centrality based on "betweenness" [22]. A betweenness measure of centrality is one that is based on a calculation of the probability that any particular point lies on a path between any two other points. This latter measure can be used to operationalize the idea of the broker or intermediary, a person who may not be central in the network as a whole, but is an important point of contact between parts of the network.

A further extension of the ideas of distance and point centrality has been the attempt to measure the overall centralization of the graph as a whole. A measure of centrality grasps the extent to which a whole graph is organized around a small number of central points that constitute its structural center.

The early work of Warner and his associates was concerned with the formation of cliques within networks. Their use of this idea of the clique was, however, relatively loose and essentially commonsensical. Developments in graph theory have since made possible the construction of a number of structural concepts for sub-graphs that are both precise and distinct from each other. The most straightforward of these sub-graph concepts is that of the "component". A component is a maximally connected sub-graph, the set of all those points that are connected, directly or indirectly, by a continuous sequence of lines. A graph will typically consist of a number of such components, together with a variable number of completely isolated points. Component boundaries may, therefore, indicate the boundaries to the flow of communication, influence, or resources within the corresponding social network. A mapping of the size distribution of components will show the extent to which these flows are fragmented. It is also possible to map the internal structures of the various components by taking account of the sign, value, and direction of the constituent lines in order to uncover any blockages or distortions in the flow of communication, influence, or resources. Using a measure of value, for example, the nesting of intensely connected components within more weakly connected components can be disclosed. The "slicing" techniques available in many software packages allow contour maps of component structure to be built, the resulting map showing the peaks of intensity and the troughs or valleys that separate them.

Component structure can also be studied through an investigation of the "cycles" that it contains. A cycle is a path that returns to its own starting point, and a cyclic component consists of a set of intersecting cycles. A component may consist of a number of such cyclic components that are connected only through non-cyclic "bridges".

Cliques, as defined in graph theory, are sub-sets of points in which every possible pair of points is connected at a specified distance. Thus, a 1-clique is a sub-set of point in which all points are directly connected to each other. A 2-clique, on the other hand, is a sub-set of points that are connected through common neighbors. There is, therefore, a whole family of clique concepts based on varying distance measures. These are generically referred to as n-cliques and any large component in a graph will typically consist of many overlapping *n*-cliques, as well-connected points will tend to be members of large numbers of cliques. An analysis of cliques must always specify the value of n that will be used. Once more, the choice of a value for

identifying cliques is not a matter that can be determined on mathematical grounds alone. The investigator must always decide on the sociological meaning – if any – that can be given to particular values of n.

This basic idea of the clique has been extended into a variety of related ideas, most notably called clans, clubs, plexes, and circles, each of which differs according to the precise criterion that defines sub-set membership. These various sub-graph and sub-set concepts are particularly important in studies of diffusion processes in social networks as they define the "obstacles" and barriers that prevent the smooth diffusion of ideas or resources. This was the insight that lay behind Granovetter's [26] recognition of the strength of weak ties: it is the possession of "weak" connections that bridge a person into sub-graphs that would otherwise be unable to transfer information to the person.

Most uses of graph theory in social network analysis have employed one-mode data. That is to say, they have transformed the initial bipartite or two-mode data into a simpler form. An initial incidence matrix showing, for example, the memberships of people (as columns) in particular organizations (as rows) is transformed into a column-by-column adjacency matrix of people and a row-by-row adjacency matrix of organizations. Thereafter, the analyses of the people and the organizations proceeds separately (see ► Social Network Analysis, Two–Mode Concepts in). Recently, however, attempts have been made to develop graph theory approaches to the analysis of two-mode data, treating the bipartite data as if it were one-mode data. The adjacency matrices are, in effect, seen as sub-matrices of a larger adjacency matrix in which people and organizations, for example, comprise *both* the columns and the rows. Once the larger matrix is generated, many of the conventional graph theory measures and visualization techniques can be applied to it. Not all such measures, however, are sensibly or usefully seen in two-mode terms.

Many of the more complex graph theoretical measures for whole graphs can be difficult, if not impossible, to use in large-scale networks where there is incomplete data. This is typically the case where a sampling methodology has been followed (see ► Social Network Analysis, Estimation and Sampling in).

Some network parameters can be estimated with sample data. If information on the neighborhood or neighborhood density of individuals are collected from a sample, for example, it is straightforward to compare estimates of these measures for the whole population. It would also be possible to estimate the density of the network as a whole from such sample data, as density is an adjacency-based measure. So long as the normal questions of sample size and representativeness are considered, such measures pose few problems. Matters are more complex, however, with many global network measures, where the structure of relations is more likely to be lost in the process of sampling. It is, for example, impossible to measure the number of separate components in a network using sample data. Some ingenious methods of estimation have, however, been suggested as being useful for such measures, so long as appropriate sampling designs are used.

An alternative to sampling is the adoption of explicit methods of data reduction that allow large-scale networks to be analyzed more easily. As Vladimir Batagelj shows in ► Social Network Analysis, Large-Scale, it is especially useful to have methods of visualization for such networks. The *Pajek* program [16] was designed specifically for this purpose. Large networks are those with thousands, or perhaps millions, of points and the computing time required for such networks has been a barrier to studying them. Batagelj uses recursive decomposition methods to divide a large network into several smaller ones. This can be achieved by the clustering of points or lines, compressing them into compound points or lines. Varying forms of reduction can be achieved by adopting different criteria for reduction: for example, collapsing all members of a component onto a point, or collapsing all members of a clique into a single point. Such reduced graphs can then be analyzed with the familiar techniques of graph theory. He also introduces a number of concepts specific to reduced networks. The concept of an island, for example, refers to peaks at specified cut-off levels on a measure of the properties of the compared points or lines.

Diffusion Processes

A key issue in sociological studies of influence has been the ways in which innovations and other social changes are diffused through the social relations that people have established. Similar processes of diffusion were investigated in geography and anthropology, and a number of investigators began to explore mathematical models of diffusion, that had already been begun to be applied in epidemiological studies of the spread of disease through a population [48,49] (see also [14]). Everett Rogers documented the existence of an S-curve describing the pace of innovation: a slow initial uptake was followed by more rapid adoption, and then a tailing-off as diffusion produced a relative saturation of the population and correspondingly fewer people available to take up the changes. The key parameters were the speed of the initial take-off and the speed of subsequent adoption.

Contemporary approaches to diffusion [57] begin from measures of spatial autocorrelation to investigate the speed of the spread of innovations (see ► Social Networks, Diffusion Processes in). Spatial autocorrelation is the correlation of a variable with itself measured spatially. That is, its dependence on neighboring points. Such models measure the extent to which changes spread through contiguous regions of space. Within networks, this idea is generalized to analyze contiguity in social network terms. It is in this way that a "contagion" model of social influence can be operationalized. Network parameters can be weighted in various ways in a model, taking, for example, direct links as more significant than indirect links, and predictive conclusions about the temporal and spatial spread of changes can be drawn. As Tom Valente shows, event history analysis has been used to allow more sophisticated temporal measures of diffusion, and this has become a key area of innovation within the area of diffusion studies. Such work is also now developing closely with the agent-based computational models discussed below.

Algebraic Models and Blockmodeling

The early work on network structures that was carried out by Harrison White and his associates [36,72] made use of algebraic ideas to model graphs of social relations. Their argument was that algebra provided a means for analyzing the properties of sets of points. This work laid the basis for what came to be known as blockmodeling. This is a way of exploring the relations among structurally defined positions in social networks, the positions being understood as "blocks" or sets of points with certain common relational attributes that make them, in crucial respects, structurally equivalent to each other [5,73]. This form of positional analysis has subsequently been elaborated and enlarged in a variety of ways, resulting in the production of a whole family of models for network structure.

An algebraic representation of a social network involves identifying a set of points that is defined by the specific relation that unites its members [43] (see ► Social Networks, Algebraic Models for). Thus, a network is seen as consisting of a set of ordered pairs. This approach can easily be generalized to explore situations in which more than one relation is involved and units can be understood as connected into multiple networks, each of which is defined by a particular type of relationship. From this point of view, a multiple network is a set of sets, and each set is defined by a "primitive relation" or generator that contributes to determining the overall network structure. Therefore, the individuals or groups that comprise the points in a social network are treated as standing in compound relations to each other. Paths in the network may be composed from lines that represent a variety of different relations. The basic step in identifying the structure of such a complex network is the multiplication of the initial binary relations to convert them into binary compound relations. The structure of relationships among all possible paths in a multiple network is termed a partially ordered semigroup and this comprises the underlying relational structure of the network. As with graph theory, this algebraic method allows such global properties of social networks to be linked to the local, egocentric properties of specific units within the network. Thus, the immediate environment of any particular member of the network can be seen as a "local role algebra" that represents the role set of the focal individual or group. Such an algebra defines the vectors that slice through the whole network and so summarize an ordering of the constituent relation vectors in which the individual point is involved.

The basic elements of blockmodeling are based on the ideas first set out by the anthropologist Nadel, who held that social network analysis must take account of social positions and their associated roles rather than simply with individuals [42] (see ► Positional Analysis and Blockmodeling). Graph theory is especially well-suited to the identification of individual-level phenomena and the characteristics of the social groups formed by individuals. It is much less well-suited to the identification of structural positions. For example, a large number of individuals may each occupy the socially determined position of "father", but a graph theoretical analysis of each individual male adult's connections to children will not easily identify this shared social role. This is because graph theory searches for clusters of connections based on distance measures. Blockmodeling, on the other hand, starts out from a recognition that individuals that occupy the same social position are "equivalent" or "substitutable" in some way. Thus, each father relates in similar ways to particular children, but the fathers do not all relate to the *same* children. Positions and their associated roles are enduring structural regularities among sets of points and differ markedly from cliques, components and other such subgraphs.

The typical strategy for constructing a blockmodel is to apply an algorithm to a bipartite, two-mode, matrix of individual-level data in order to reduce it to an "image graph" that comprises blocks of equivalent points. Though typically employing two-mode data, blockmodels can be produced for one-mode data and, with greater difficulty, for three-mode data. Simple binary data have been most widely used in blockmodeling, with valued data being han-

dled by slicing procedures that partition the matrix into 1s and 0s on the basis of a chosen threshold of significance. Attempts are being made, however, to extend blockmodeling directly to valued data. Approaches to blockmodeling vary from each other in terms of the particular criteria that are used in the algorithm. The earliest approach – the CONCOR method of White and his associates – used iterated correlation measures to produce the image graph. In this method, repeated correlations computed on the rows and columns of the initial matrix converge towards a rearranged matrix that shows blocks of 0s and blocks of 1s. The blocks of 0s – "zero-blocks" – are the structural holes in the network that divide the various social positions (the blocks of 1s) from each other. Later work associated with Ronald Burt [9,10], has used simpler clustering approaches that are based on Euclidean distance measures calculated from path distances among pairs of points. Burt's method also allows the construction of an image graph for the identification of blocks of equivalent positions, and Burt terms this a social topology.

Applications of these ideas to intercorporate business connections has shown that blocks in networks of shareholdings and interlocking directorships can be seen as comprising structurally similar agents within the economy. The blocks uncovered in such an analysis are the dominant investors and subordinate enterprises that stand in similar relations to each other but may have few mutual or reciprocal connections among themselves. They are social positions but do not necessarily constitute social groups. The key text on blockmodeling has been produced by Doreian and colleagues [19].

The construction of an algebraic representation of a social network comprises a depiction of the "deep structure" of the network in the sense that was intended by Lévi-Strauss and many structuralist writers. The method allows a formal construction and representation of the deep structures that have, more typically, been identified through more purely interpretative means. As in these structuralist approaches, a key aim of algebraic network analysis is the identification of isomorphic subgroups and local role algebras across different empirical networks that can be compared for commonalities of structure. This comparative approach allows the formulation of statements concerning generic processes in social systems, modeled as lattices. Hierarchical structures, for example, may be common to a number of different empirical domains and may exhibit similar formal properties, regardless of their specific empirical content. Similarly, empirically variant role sets may exhibit a similarity of underlying structure, such as patterns of conflict or cooperation with different categories of others.

This approach does not, however, reduce the variation found in actual social networks to invariant mathematical properties. There can be no derivation of a network *theory* directly from a mathematical theory. As Pattison (see p. 135 in [43]) argues, algebraic approaches have sacrificed mathematical power in order to reflect more adequately the social relationships in the particular field under investigation. Algebraic methods have tended to be used – as shown in the discussion of blockmodeling – as methods of data reduction, as means for decomposing or fragmenting networks into simpler structural components. The construction of empirically relevant theories from these components is far more difficult and has, so far, rarely been pursued.

Scaling models and Visualization

Both graph theory and algebraic approaches offer rudimentary techniques for the visualization of social networks (see ► Social Network Visualization, Methods of). Graph theory, for example, is based around the idea of the sociogram of connected points. For any but the smallest networks, however, sociograms become a confused jumble of cross-cutting lines, and a rearrangement of points aimed at reducing the amount of overlap can actually destroy any visual representation of the structure of the network by distorting the distances between and the relative locations of the points. A blockmodeling approach to large networks reduces the amount of data to be handled and allows simple image graphs of the relations among blocks to be drawn, but this is limited to positional analysis and loses sight of the details that are present in the relations among the individual points. In both forms of analysis, therefore, social network analysts have tended to rely on the purely mathematical manipulation of large networks and many have almost abandoned the attempt to visually display whole complex social networks. The desire to recapture the simple visual impact of the sociogram, however, has motivated a number of researchers to investigate ways of drawing network diagrams that retain the spatial sense inherent in relational data. Unsurprisingly, perhaps, the techniques discovered have followed the methods used in cartography to represent physical landscapes on the page of a map or atlas.

The earliest and most influential visualization techniques drew on multidimensional scaling (MDS), as this was seen as a method for mapping data that retained its fundamental metrics of space and distance. Essentially similar techniques are those of principal component analysis or factor analysis. Instead of calculating distance simply by the number of lines connecting two points – "path

distance" – MDS approximates to the measures of physical distance used in conventional spatial thinking. MDS techniques convert raw data on social networks into Euclidean distances. In this conversion, points are regarded as "close" to each other by virtue of their position in the overall configuration of relations. This is typically measured on the basis of a proximity measure such as the frequency of contact between two points or the intensity of a relationship between them, and these are usually given a strictly metric form by computing correlation coefficients. For example, two points with identical patterns of connection are perfectly correlated with each other, and decreasing similarity is reflected in declining closeness. Using such similarity (or dissimilarity) data, MDS procedures search for the best fit within a Euclidean space of specified dimensions. Once a solution is obtained it can be displayed on the page and rotated into an informative orientation that discloses the meaning of the dimensions that have been identified in the analysis. Non-metric MDS, using the rank order of distances rather than imputed actual distances, is particularly appropriate for much social network data and operates in very much the same way. Useful implementations of this approach are now available in the major software packages, and novel techniques such as multiple correspondence analysis are also beginning to become available.

The use of MDS and similar scaling or dimensional models provides a visual image of a network, but it also provides measures of distance, direction, and location that are not available within graph theory or algebraic approaches. This can allow a more rigorous embedding of social networks in social space and can permit intriguing hypotheses to be formed. For example, cliques and components, centrality and density, can all be mapped onto an MDS solution, and the relations among cliques or central points appear strikingly and can be measured determinately. The procedures made available within the *Pajek* program [18] have been designed specifically to allow such possibilities for large-scale networks. Graph theory and algebraic methods can be used together with the spring embedding of data through a variant of MDS to produce powerful network representations.

Particular problems of visualization occur when research is concerned with two-mode data rather than simple one-mode data. The most easily interpretable results are those that display a bipartite graph by using different symbols or colors for the two sets of data. Such graphs are a logical extension of graph theory approaches to two-mode data. Some interesting work is now being undertaken, however, using Galois lattices. Abandoning graph theoretical and spatial representations, the lattices aim to disclose the pattern of connections and their hierarchical order of "containment" from core to periphery.

One particularly striking advance in visualization techniques has been the use of models that allow the tracking of change over time. These have typically involved the adaptation of animation methods. Further advances in this area might be expected as some of the techniques discussed below begin to offer improved ways of handling temporal data on social networks.

Statistical Models for Hypothesis Testing

Much work on social networks has remained at the level of description and has failed to move towards explanatory concerns. Only rarely have researchers gone beyond a description to formulate hypotheses and test theories aimed at explaining the observed structures. One reason for this absence of theoretical work and hypothesis testing has been the weakness of the available statistical techniques for undertaking this work. While attempts have sometimes been made to use basic statistical measures of probability and significance to test hypotheses about network structure, these involve a number of difficulties. Standard statistical procedures such as significance tests, regression, and the analysis of variance all assume the independence of observations. The relational data studied by social network analysts, however, cannot be seen in this way. Relational data are, by definition, non-independent observations. Thus, novel statistical techniques have been required if methods of statistical inference are to be used. The most important work in recent years has been the generalization of Markov graphs to a larger family of models that has been undertaken by Stanley Wasserman and his colleagues [44,46,65], building on earlier work by Frank and Strauss [21]. Their $\boldsymbol{p}^*$ models – now termed exponential random graph models (ERG models or ERGM) – are designed specifically to allow the easier formulation and testing of theories. An ERGM defines a probability distribution on the set of all networks that can be constructed on a given set of points in terms of a particular parameter vector. ERGMs comprise a family of statistical models that promise a great advance in the understanding of social networks and have the potential to connect with the "small world" methods discussed below. The models are discussed in ► Social Networks, Exponential Random Graph ($\boldsymbol{p}^*$) Models for.

For any set of points it is possible to generate through simulation a set of all the possible graphs connecting them. These randomly generated graphs vary along the full range from completely unconnected to completely connected. A very large number of such graphs can exist along that

range, as the number of possible graphs varies with the size of the network at an exponential rate. Thus, the number of possible configurations for a three-point graph is 64, that of a four-point graph is 1024, and the number for a five point graph is in excess of a million. An actually observed network can be treated as one realization from this set of logically possible graphs. If its probability of occurring by chance is low, then its actual occurrence may be regarded as statistically significant.

To assess this statistical significance, it is necessary to construct a probability distribution of the random graphs. Each random graph has a specific probability of occurring, depending on its particular structural properties. The aim in constructing these distributions is to see whether a particular structural property, such as a particular level of density, centralization, or clustering, is more or less likely to occur by chance than any other. These probabilities are estimated through log linear analysis from the statistical dependencies among the members of the network. The log odds ratios of the conditional probabilities of the relational ties associated with each element constructed from this dependence graph are used to produce an approximate likelihood function that can be used with Monte Carlo estimation techniques to produce maximum likelihood estimates. In this process the successive approximations constitute chains of graphs – a Markov chain – that converge towards a stationary distribution and a stable estimate of network parameters. Thus, ERGM provides a means for testing the probability that the observed network will occur with precisely the structural characteristics that it has on the basis of chance alone. This enables the researcher to highlight the statistically significant – non-chance – results apparent in the data.

Agent-Based Computational Models and Temporal Processes

As well as its descriptive focus, much social network analysis has also concentrated on the static features of social networks. This has also begun to change in recent years as more attention has been given to the dynamic processes involved in changes over time. A key advance in this direction has come from the use of models that depict the ways in which the behavior of individual agents results in global transformations of network structure. In so-called agent-based computational models, agents (whether individuals or groups) are seen as rule-following entities whose decisions to act in one way or another are consequential for the overall network by virtue of their concatenation with the action consequences of others. Therefore, a knowledge of the rules under which agents act can be used to predict broad patterns of change in network structure.

Agent-based models originated in simulation studies, where a simple and determinate set of rules can be applied to generate specific network structures (see also pp. 99ff in [25]; [40]). Applications of this idea to human social networks have tended to rely on assumptions drawn from rational choice theory, according to which individual agents are assumed to act on certain simple principles of utility maximization. However, the more complex models allow for the handling of different categories of agent, the members of each category following a different set of rules, and for a relaxation of the assumption of pure rationality by modeling rules of non-rational or ritualistic behavior. Such models are stochastic – specifying the probability of certain courses of action but not implying a complete determinism.

A particularly powerful example of such an approach is that of Tom Snijders [52,53,54] (see ► Network Analysis, Longitudinal Methods of). Change in network structure is regarded as a probabilistic process, describable by an objective function that measures the attractiveness to the agents of changes in their connectivity. The model sees an incremental adjustment of individual action to the changing network structure, resulting in a continuous – but often non-linear – process of network evolution. It is assumed that agents control the outgoing relations to others within the constraints on their opportunities that are set by the current network structure. However, they act "myopically" in relation to the immediate environment and the immediate consequences of their actions, having no conception of the wider consequences of their choices. At each point in time, the outcome for the overall network is a result of the intersection of the courses of action pursued by these myopic agents and so it will rarely correspond to the intentions of any one actor. At the same time, each actor has only a partial awareness of the changes that have resulted from their actions and so their next acts must be regarded as equally myopic and as likely, once again, to have unintended consequences at the network level. As a result of the continual iteration of actions, there is a co-evolution of network and behavior. Complex patterns of change are likely to occur. Snijders shows that small, incremental changes can accumulate to a point of "crisis" [55], a tipping point at which a radical, non-linear transformation of network structure occurs.

Snijders' longitudinal approach is especially powerful as it explicitly takes account of global network parameters such as density, reciprocity, distance, and balance. It can also be reprogrammed to recognize a wide range of different rule systems and decision procedures. Snijders

also suggests that the approach can handle a greater degree of far-sightedness in agents, though with a corresponding decline in the determinateness of the analyses that can be drawn. Current work is making important connections with the early work of Wasserman [63] and the subsequent development of ERGM procedures. The approach is implemented in the *SIENA* program and has been used in a number of simulation studies.

Small World Models and Network Dynamics

The most recent area of development in social network analysis is the investigation of small world networks inspired by a number of physicists interested in network dynamics (explored in various entries in other sections of this *Encyclopedia*). Although this has, as yet, led to relatively few empirical investigations, there is growing interest in these applications. Many advocates of the approach have failed to discover or to appreciate the amount of work already undertaken in social network analysis and have assumed a very truncated history for the approach. There are, however, some signs of a closer appreciation of social science work, and the approaches are beginning to have an impact in the mainstream of social network analysis.

The "small world" hypothesis suggests that any two people chosen at random are likely to be connected to each other by a path of relatively short length. The hypothesis was formulated by Ithiel de Sola Pool and Manfred Kochen in a working paper of 1958 that was not published for another 20 years [17]. Their ideas achieved a wide currency and in 1967 Stanley Milgram at Harvard undertook some experimental studies to test the hypothesis [38]. He gave experimental subjects the names of people initially unknown to them and asked them to get a message to them using only known persons as intermediaries. That is, the messages were to pass through friend of friends of friends until they reached their destinations. Milgram discovered that subjects and targets were typically separated by paths of distance 6, no matter what the physical distances separating them. This finding embodied the common exclamation "it's a small world" uttered when people discover a mutual connection. Milgram hypothesized that the density of links of acquaintanceship is such that a considerable amount of redundancy is built into social networks: it is level of density in a network that explains its small world characteristics.

In the late 1990s, Watts and Strogatz [68] undertook some empirical studies in the mathematics of random networks. The structural properties of random networks, produced through the random linking of one point to another, had first been studied by Paul Erdös and Alfréd Rényi [20] and had led to a number of interesting conclusions. Watts and Strogatz discovered that a particular subset of random graphs corresponded closely to those actually found in natural and social phenomena, and they demonstrated that these had the small world characteristics identified by Milgram; though they were skeptical about the precise finding of "six degrees of separation". They showed, however, that the high levels of clique and sub-graph formation that were basic to the redundancy found in real networks gave them a number of characteristic features. Crucial to the structure of these networks was the existence of well-connected points – which they termed "hubs" – that make the density and, therefore, the small world properties, possible. It is the existence of hubs that explains why the density does not need to be extremely high for small world conditions to apply. On the contrary, there is a middle range of density in which efficient communication is possible.

The physicist Albértó-Laszlo Barabási, searching for a new topic to investigate, began to look at the findings of Watts and Strogatz. Barabási's central idea was that of the "power law" or "scale free" distribution, according to which small world networks have a frequency distribution of point degrees showing a small number of well-connected points and a large number of less well-connected points. These well-connected points are the hubs that generate the small world properties. He concluded that these findings provided the basis for a completely new approach to networks [1]. As has already been noted, his claims to novelty and to the revolutionary character of the work seriously underestimated the amount of prior work undertaken on social networks and their properties. Nevertheless, he did formulate some intriguing ideas and the work subsequently taken up by Watts does offer the prospect of genuine advances in social network analysis.

Watts holds that a small world graph is one in which there are a large number of "short cuts". These are lines that connect points that would otherwise be quite distant from each other. Neither a completely connected graph nor a sparse graph have short cuts, and so small world properties occur in an intermediate range of graphs. They are locally clustered but globally sparse. It is for these graphs, he holds, that precise mathematical conclusions about structure and structural development can be drawn. He holds, for example, that relatively small changes in network connectivity, such as the addition of a small number of shortcuts, can dramatically change its properties through "phase transitions" at critical threshold points. There are considerable possibilities inherent in using these ideas alongside those from agent-based computational studies already discussed. The common thread

is a concern for the investigation of longitudinal processes in which micro-level incremental changes result in non-linear macro-level transformations of structure.

Bibliography

1. Barabási A-L (2002) Linked: The new science of networks. Perseus, Cambridge
2. Barnes JA (1954) Class and committee in a Norwegian island parish. Hum Relat 7:39–58
3. Bavelas A (1948) A mathematical model for group structures. Appl Anthr 7:16–30
4. Berkowitz SD (1982) An introduction to structural analysis. Butterworths, Toronto
5. Boorman SA, White HC (1976) Social structure from multiple networks, vol II. Am J Sociol 81:1384–1446
6. Bott E (1955) Urban families: Conjugal roles and social networks. Hum Relat 8:253–292
7. Bott E (1956) Urban families: The norms of conjugal roles. Hum Relat 9:325–342
8. Buchanan M (2002) Small world: Uncovering nature's hidden networks. Weidenfeld and Nicolson, London
9. Burt RS (1980) Models of network structure. Ann Rev Sociol 6:79–141
10. Burt RS (1982) Towards a structural theory of action. Academic Press, New York
11. Burt RS (1992) Structural holes. Cambridge University Press, New York
12. Carrington PJ, Scott J, Wasserman S (2005) Models and methods in social network analysis. Cambridge University Press, Cambridge
13. Cartwright D, Zander A (1953) Group dynamics. Tavistock, London
14. Coleman JS, Katz E, Menzel H (1966) Medical innovation: A diffusion study. Bobbs-Merrill, New York
15. Davis AB, Gardner BB, Gardner MR (1941) Deep south. University of Chicago Press, Chicago
16. De Nooy W, Mrvar A, Batagelj V (2005) Exploratory social network analysis with Pajek. Cambridge University Press, New York
17. De Sola Pool I, Kochen M (1978) Contacts and influence. Soc Netw 1(1):5–51
18. DeNooy W, Mrvar A, Batagelj V (2005) Exploratory social network analysis with Pajek. Cambridge University Press, New York
19. Doreian P, Batagelj V, Ferligoj A (2005) Generalized blockmodelling. Cambridge University Press, New York
20. Erdös P, Rényi A (1959) On the evolution of random graphs. Publ Math Inst Hung Acad Sci 5:17–61
21. Frank O, Strauss D (1986) Markov Graphs. J Am Stat Assoc 81:832–842
22. Freeman LC (1979) Centrality in social networks: Conceptual clarification. Soc Netw 1:215–239
23. Freeman LC (2004) The development of social network analysis: A study in the sociology of science. Empirical Press, Vancouver
24. Freeman LC, White DR, Romney AK (1989) Research methods in social network analysis. Transaction Books, New Brunswick
25. Gilbert N, Troizch KG (1999) Simulation for the social scientist. Open University Press, Buckingham
26. Granovetter M (1973) The strength of weak ties. Am J Sociol 78:1360–1380
27. Granovetter M (1974) Getting a job. Harvard University Press, Cambridge
28. Harary F, Norman RZ (1953) Graph theory as a mathematical model in social science. Institute for Social Research, Ann Arbor
29. Helmers HM (1975) Graven naar macht. Van Gennep, Amsterdam
30. Homans G (1950) The human group. Routledge and Kegan Paul, London
31. Homans G (1961) Social behaviour: Its elementary forms. Routledge and Kegan Paul, London
32. Knoke D, Kuklinski JH (1982) Network analysis. Sage Publications, Beverley Hills
33. König D (1936) Theorie der endlichen und unendlichen Graphen. Chelsea, New York
34. Laumann EO (1966) Prestige and association in an urban community. Bobbs-Merrill, Indianapolis
35. Levine JH (1972) The sphere of influence. Am Sociol Rev 37:14–27
36. Lorrain F, White HC (1971) Structural equivalence of individuals in social networks. J Math Sociol 1:49–80
37. Marsden PV, Lin N (1982) Social structure and network analysis. Sage, Beverley Hills
38. Milgram S (1967) The small world problem. Psychol Today 2:60–67
39. Mitchell JC (1969) Social networks in urban situations. Manchester University Press, Manchester
40. Monge PR, Contractor NS (2003) Theories of communication networks. Oxford University Press, Oxford
41. Moreno JL (1934) Who shall survive? Beacon Press, New York
42. Nadel SF (1957) The theory of social structure. Free Press, Glencoe
43. Pattison P (1993) Algebraic models for social networks. Cambridge University Press, Cambridge
44. Pattison P, Wasserman S (1999) Logit models and logistic regressions for social networks: II Multivariate relations. Br J Math Stat Psychol 52:169–193
45. Radcliffe-Brown AR (1940) On social structure. In: Radcliffe-Brown AR (ed) Structure and function in primitive society. Cohen and West, London
46. Robins GL, Pattison P, Wasserman S (1999) Logit models and logistic regressions for social networks: III Valued relations. Psychometrika 64:371–394
47. Roethlisberger FJ, Dickson WJ (1939) Management and the worker. Harvard University Press, Cambridge
48. Rogers E (1962) Diffusion of innovations, 5th edn. Free Press, New York
49. Ryan R, Gross N (1943) The diffusion of hybrid seed corn in two Iowa communities. Rural Sociol 8:15–24
50. Scott J (1991) Networks of corporate power: A comparative assessment. Ann Rev Sociol 17:181–203
51. Scott J (2000) Social network analysis, 2nd edn. Sage, London
52. Snijders TAB (2001) The statistical evaluation of social network dynamics. In: Sobel ME, Becker MP (eds) Sociological methodology. Basil Blackwell, Oxford
53. Snijders TAB (2005) Models for longitudinal network data. In: Carrington PJ, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, Cambridge

54. Snijders TAB, Van Duijn MAJ (1997) Simulation for statistical inference in dynamic network models. In: Conte R, Hegelmann R, Terna P (eds) Simulating social phenomena. Springer, Berlin
55. Thom R (1972) Structural stability and morphogenesis. Benjamin, London
56. Travers J, Milgram S (1969) An experimental study of the small world problem. Sociometry 32:425–443
57. Valente TW (1995) Network models of the diffusion of innovations. Hampton Press, Cresskill
58. Warner WL (1949) Democracy in Jonesville. Harper and Brothers, New York
59. Warner WL, Low JO (1947) The social system of a modern factory. Yale University Press, New Haven
60. Warner WL, Lunt PS (1941) The social life of a modern community. Yale University Press, New Haven
61. Warner WL, Lunt PS (1942) The status system of a modern community. Yale University Press, New Haven
62. Warner WL, Srole L (1945) The social systems of american ethnic groups. Yale University Press, New Haven
63. Wasserman S (1980) Analyzing social networks as stochastic processes. J Am Stat Assoc 75:280–294
64. Wasserman S, Faust K (1994) Social network analysis: Methods and applications. Cambridge University Press, New York
65. Wasserman S, Pattison P (1996) Logit models and logistic regressions for social networks, vol I. An introduction to Markov random graphs and $\boldsymbol{p}^*$. Psychometrika 60:401–426
66. Watts D (1999) Small worlds: The dynamics of networks between order and randomness. Princeton University Press, Princeton
67. Watts D (2003) Six degrees. The science of a connected age. Norton WW, New York
68. Watts DJ, Strogatz SH (1998) Collective dynamics of small-world networks. Nature 393:440–442
69. Wellman B (1979) The community question: The intimate networks of East Yorkers. Am J Sociol 84:1201–1231
70. Wellman B (1988) Networks as personal communities. In: Wellman B, Berkowitz SD (eds) Social structures. Cambridge University Press, Cambridge
71. Wellman B, Berkowitz S (1988) Social structures. Cambridge University Press, New York
72. White HC (1963) An anatomy of kinship. Prentice-Hall, Englewood Cliffs
73. White HC, Boorman SA, Breiger RL (1976) Social structure from multiple networks, I. Am J Sociol 81:1384–1446

Social Network Analysis, Two-Mode Concepts in

Stephen P. Borgatti
Gatton College of Business and Economics, University of Kentucky, Lexington, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Basic Concepts
Two-Mode Data in Social Network Analysis
Unimodal Approaches to Two-Mode Data
Bimodal Approaches to Two-Mode Data
Conclusion
Future Directions
Bibliography

Glossary

2-Mode matrix A (2-dimensional) matrix is said to be 2-mode if the rows and columns index different sets of entities (e. g., the rows might correspond to persons while the columns correspond to organizations). In contrast, a matrix is 1-mode if the rows and columns refer to the same set of entities, such as a city-by-city matrix of distances.

Blockmodel A blockmodel is a partitioning of the cells of a matrix into blocks that is induced by the partitioning the rows and columns into classes and sorting the matrix such that rows (and columns) that belong to the same class are next to each other. More specifically, two matrix cells x_{ij} and x_{mp} are in the same block if $\text{class}(i) = \text{class}(m)$ and $\text{class}(j) = \text{class}(p)$.

Centrality A family of concepts for characterizing the structural importance of a node's position in a network.

Graph cohesion A family of concepts characterizing the extent of connectedness of a graph, such as density (the proportion of pairs of nodes that have ties with each other), or average path distance.

Multidimensional scaling (MDS) A method of locating points in space such that Euclidean distances between the points correspond to a matrix of input similarities/distances among objects. Used to provide visual representations of 1-mode matrices such as correlation matrices or perceptual distances among objects.

Regular equivalence The definition of regular equivalence is recursive. If two nodes are regularly equivalent, then they are connected to regularly equivalent nodes. Regular equivalence is used to identify nodes that are playing the same structural role, even if they are not connected to each other.

Social network (or, in graph theory, a graph) A collection of nodes (also referred to as vertices or actors) together with a set of ties (also known as edges or links) that connect pairs of nodes. Typically used to represent social relations such as who is friends with whom, or who is the supervisor of whom.

Structural equivalence At an intuitive level, a pair of nodes is said to be structurally equivalent to the ex-

Names of Participants of Group I	Code Numbers and Dates of Social Events Reported in *Old City Herald*													
	(1) 6/27	(2) 3/2	(3) 4/12	(4) 9/26	(5) 2/25	(6) 5/19	(7) 3/15	(8) 9/16	(9) 4/8	(10) 6/10	(11) 2/23	(12) 4/7	(13) 11/21	(14) 8/3
1. Mrs. Evelyn Jefferson	×	×	×	×	×	×		×	×					
2. Miss Laura Mandeville	×	×	×		×	×	×	×						
3. Miss Theresa Anderson		×	×	×	×	×	×	×	×					
4. Miss Brenda Rogers	×		×	×	×	×	×	×						
5. Miss Charlotte McDowd			×	×	×		×							
6. Miss Frances Anderson			×		×	×		×						
7. Miss Eleanor Nye					×	×	×	×						
8. Miss Pearl Oglethorpe						×		×	×					
9. Miss Ruth DeSand					×		×	×	×					
10. Miss Verne Sanderson							×	×	×			×		
11. Miss Myra Liddell								×	×	×		×		
12. Miss Katherine Rogers								×	×	×		×	×	×
13. Mrs. Sylvia Avondale							×	×	×	×		×	×	×
14. Mrs. Nora Fayette						×	×		×	×	×	×	×	×
15. Mrs. Helen Lloyd							×	×		×	×	×		
16. Mrs. Dorothy Murchison								×	×					
17. Mrs. Olivia Carleton									×		×			
18. Mrs. Flora Price									×		×			

Social Network Analysis, Two-Mode Concepts in, Figure 1
DGG women-by-events matrix

tent that they occupy identical locations in a network, meaning that they are connected to exactly the same others. Structurally equivalent nodes are identical with respect to all structural properties, such as centrality or subgroup membership.

Definition of the Subject

In social network analysis, 2-mode data refers to data recording ties between two sets of entities. In this context, the term "mode" refers to a class of entities – typically called actors, nodes or vertices – whose members have social ties with other members (in the 1-mode case) or with members of another class (in the 2-mode case). Most social network analysis is concerned with the 1-mode case, as in the analysis of friendship ties among a set of school children or advice-giving relations within an organization. The 2-mode case arises when researchers collect relations between classes of actors, such as persons and organizations, or persons and events. For example, a researcher might collect data on which students in a university belong to which campus organizations, or which employees in an organization participate in which electronic discussion forums. These kinds of data are often referred to as affiliations. Co-memberships in organizations or participation in events are typically thought of as providing opportunities for social relationships among individuals (and also as the consequences of pre-existing relationships). At the same time, ties between organizations through their members are thought to be conduits through which organizations influence each other.

Introduction

Perhaps the best known example of 2-mode network analysis is contained in the study of class and race by Davis, Gardner and Gardner (henceforth DGG) published in the 1941 book *Deep South* [6]. They followed 18 women over a nine-month period, and reported their participation in 14 events, such as a meeting of a social club, a church event, a party, and so on. Their original figure is shown in Fig. 1.

DGG used the data to investigate the extent to which social relations tended to occur within social classes.

Basic Concepts

A typical data matrix has two dimensions or *ways*, corresponding to the rows and columns of the matrix. The number of ways in a matrix X can be thought of as the number of subscripts needed to represent a particular datum, as in x_{ij}. If we stack together a number of similarly sized 2-dimensional matrices, we can think of the result as a 3-dimensional or 3-way matrix.

The *modes* of a matrix correspond to the distinct sets of entities indexed by the ways. In the DGG dataset described above, the rows correspond to women and the columns to a different class of entities, namely events.

Hence, the matrix has two modes in addition to two ways; it is 2-way, 2-mode. In contrast, a persons-by-persons matrix A, in which $a_{ij} = 1$ if person i is friends with person j, is a 2-way, 1-mode matrix, because both ways point to the same set of entities.

In a sense, what constitutes different modes is up to the researcher. If we collect romantic ties among a group of heterosexual people of both genders, we could construct a 2-mode men-by-women matrix X in which $x_{ij} = 1$ if a romantic tie was observed between man i and woman j, and $x_{ij} = 0$ otherwise. Or, one could construct a larger 1-mode person-by-person matrix B also consisting of 1s and 0s in which it just happens that 1s only occur in cells where the row and column correspond to persons of different gender. Use of the men-by-women matrix would imply that same-gender relations were impossible, whereas use of the person-by-person matrix would suggest that same-gender relations were logically possible, even if actually not observed.

Matrices recording relational information such as romantic ties can be represented as mathematical graphs as well. A graph $G(V, E)$ consists of a set of nodes or vertices V together with a set of lines or edges E that connect them. An edge is simply an unordered pair of nodes (u, v). (In directed graphs or *digraphs* we use *ordered* pairs to indicate direction of the tie.) To indicate a tie between two nodes u and v, we simply include the pair (u, v) in the set E. The number of nodes in a graph is denoted by $|V|$ or n.

A bipartite graph is a graph in which we can partition all nodes into two sets, V_1 and V_2, such that all edges include a member of V_1 and a member of V_2. The number of nodes in each vertex set is denoted n_1 and n_2, respectively.

Two-Mode Data in Social Network Analysis

Most social networks are conceived of as relations among a set of nodes, and therefore represented as a 1-mode matrix (typically of 1s and 0s) or a simple graph or digraph. For example, we might collect data on who is friendly with whom within an organization, or who injects drugs with whom in a neighborhood.

However, 2-mode data are common in social network contexts as well. Typical examples include, actor-by-event attendance (as in the DGG data), actor by group membership (such as managers sitting on corporate boards), and actor by trait possession (such as adjective checklist data), and actor by object possession (such as material style of life scales in which inventories are made of household possessions).

In many cases when 2-mode data are collected, the analytical interest is focused on one mode or the other. For example, in the DGG dataset, person-by-event attendances were collected in order to understand social relations among the women, specifically, whether women tended to have social relations primarily within their own social classes. In the interlocking directorate literature, membership of executives on corporate boards is collected mainly in order to understand how corporations are intertwined, and how the structure of this connectivity affects corporate control of society.

However, it can also occur that neither mode dominates our analytical focus and the primary interest is in the correspondence of one mode to the other. For example, a university might ask its faculty which courses they prefer to teach. Here, the objective is typically not to understand how faculty are related to each other through courses, nor how courses are related via faculty, but in the optimal assignments of persons to courses so that courses are staffed and faculty are not complaining.

Unimodal Approaches to Two-Mode Data

One approach to handling 2-mode data in social network analysis is to convert the data to 1-mode data. This is especially appropriate when the analytical interest focuses primarily on just one of the modes. Consider, for example, the case of a person by group membership matrix X in which $x_{ij} = 1$ if person i belongs to group j. Let us assume that the groups are small and everyone in a group knows everyone else. In that case, we could try to infer an acquaintance network by constructing a 1-mode matrix A such that $a_{ij} = 1$ if person i is in at least one group with person j. Better yet, we can construct a valued matrix A such that a_{ij} gives the number of groups that i and j are both members of. In other words,

$$a_{ij} = \sum_k x_{ik} x_{jk} \text{ or } A = \mathrm{XX}' . \tag{1}$$

We might regard a_{ij} as a proxy for the social proximity of i and j, or perhaps as a rough indicator of the potential for information flow between them. In this approach, we analyze each mode of the data separately. Figure 2 shows the values of A for the 2-mode data shown in Fig. 1.

It should be noted that A can be seen as a matrix of profile similarities or correlations among pairs of rows in X. For example, the matrix of Pearson correlations among rows of X is defined as follows:

$$r_{ij} = \frac{\frac{1}{m}\sum_k x_{ik} x_{jk} - u_i u_j}{s_i s_j} , \tag{2}$$

where u_i is the mean of row i and s_i is the standard deviation of row i. It is evident that the correlation r_{ij} is es-

	EVE	LAU	THE	BRE	CHA	FRA	ELE	PEA	RUT	VER	MYR	KAT	SYL	NOR	HEL	DOR	OLI	FLO
EVELYN	8	6	7	6	3	4	3	3	3	2	2	2	2	2	1	2	1	1
LAURA	6	7	6	6	3	4	4	2	3	2	1	1	2	2	2	1	0	0
THERESA	7	6	8	6	4	4	4	3	4	3	2	2	3	3	2	2	1	1
BRENDA	6	6	6	7	4	4	4	2	3	2	1	1	2	2	2	1	0	0
CHARLOTTE	3	3	4	4	4	2	2	0	2	1	0	0	1	1	1	0	0	0
FRANCES	4	4	4	4	2	4	3	2	2	1	1	1	1	1	1	1	0	0
ELEANOR	3	4	4	4	2	3	4	2	3	2	1	1	2	2	2	1	0	0
PEARL	3	2	3	2	0	2	2	3	2	2	2	2	2	2	1	2	1	1
RUTH	3	3	4	3	2	2	3	2	4	3	2	2	3	2	2	2	1	1
VERNE	2	2	3	2	1	1	2	2	3	4	3	3	4	3	3	2	1	1
MYRNA	2	1	2	1	0	1	1	2	2	3	4	4	4	3	3	2	1	1
KATHERINE	2	1	2	1	0	1	1	2	2	3	4	6	6	5	3	2	1	1
SYLVIA	2	2	3	2	1	1	2	2	3	4	4	6	7	6	4	2	1	1
NORA	2	2	3	2	1	1	2	2	2	3	3	5	6	8	4	1	2	2
HELEN	1	2	2	2	1	1	2	1	2	3	3	3	4	4	5	1	1	1
DOROTHY	2	1	2	1	0	1	1	2	2	2	2	2	2	1	1	2	1	1
OLIVIA	1	0	1	0	0	0	0	1	1	1	1	1	1	2	1	1	2	2
FLORA	1	0	1	0	0	0	0	1	1	1	1	1	1	2	1	1	2	2

Social Network Analysis, Two-Mode Concepts in, Figure 2
Women-by-women matrix of overlaps across events

		Row j	
		1	0
Row i	1	a	b
	0	c	d

Social Network Analysis, Two-Mode Concepts in, Figure 3
Contingency table T

sentially a_{ij} corrected for the number of groups that each belongs to. This kind of correction seems eminently desirable, but of course there are many ways of doing this. For example, consider a cross-tabulation T of row i and row j, such that t_{uv} gives the number of columns k of X for which $x_{ik} = u$ and $x_{jk} = v$, as shown in Fig 3.

In the table, the first row marginal $(a + b)$ gives the number of groups that person i belongs to while the first column marginal gives the number of groups that person j belongs to. Note that a_{ij} of Eq. (1) corresponds to a in Fig. 3. An obvious approach is the Jaccard coefficient, which may be defined as

$$c_{ij} = \frac{a}{a + b + c} \,. \tag{3}$$

Thus, c_{ij} is essentially the cardinality of the intersection of the groups belonged to by both persons, divided by the cardinality of their union. The Jaccard coefficient is often recommended when the number of columns in X is large and the number of 1s in each row is highly limited.

An alternative specifically designed for 2-mode affiliation data is given by Bonacich [1]. It normalizes a_{ij} as

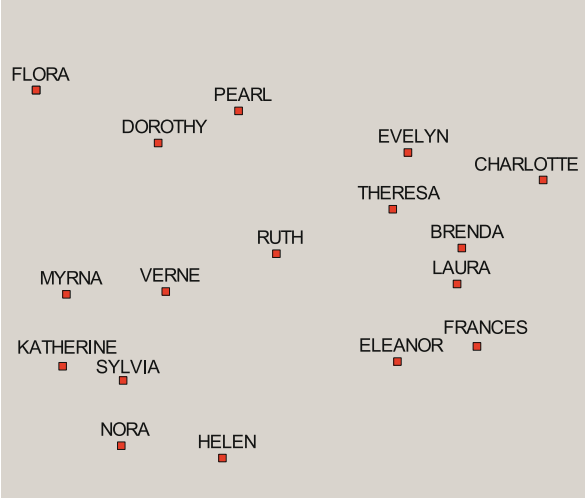

Social Network Analysis, Two-Mode Concepts in, Figure 4
Multidimensional scaling of Jaccard coefficients among rows of DGG matrix

follows:

$$a^*_{ij} = \frac{a_{ij} - \sqrt{adbc}}{ad - bc} \,, \quad \text{for } ad \neq bc \,. \tag{4}$$

Unimodal Visualization of Two-Mode Data

Given that a 2-mode matrix has been transformed into a 1-mode matrix by taking similarities among the rows (or columns), one can visualize the network using all the usual techniques for visualization of valued networks. For exam-

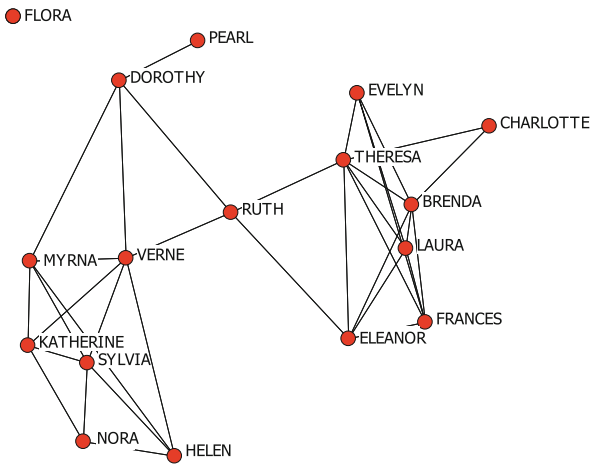

Social Network Analysis, Two-Mode Concepts in, Figure 5
Multidimensional scaling of Jaccard similarity coefficients, with lines indicating similarity scores greater than 0.4. Actor Olivia is obscured by Flora

ple, a standard approach is to use metric multidimensional scaling (MDS) on the matrix A, generating a map in which points corresponding to nodes (e. g., persons) appear close to each other to the extent that they share many groups. Figure 4 shows such an MDS map based on Jaccard similarities.

To highlight structure, it is common to overlay lines between pairs of nodes with a similarity score greater than a certain level. Figure 5 shows the result for similarities greater than 0.4. The diagram effectively shows the bridging role of Ruth, who, based on ethnographic evidence, was seen by DGG as a member of both groups of women.

Alternatively, one can use a standard graph layout algorithm (GLA) to draw the graph induced by dichotomizing the Jaccard similarity matrix. For example, define $(u, v) \in E$ if and only if (iff) $c_{ij} > 0.4$. Compared to multidimensional scaling representations, GLAs have the disadvantage that distances between points cannot strictly be interpreted, but this property also means that nodes need not obscure each other. Figure 6 shows the results of applying a spring-embedding [10] GLA to the dichotomized data.

A similar analysis can be carried out on the events rather than the women. Applying Eq. (3) to the columns of the 2-mode matrix in Fig. 1 yields a matrix of Jaccard coefficients which can be visualized using the same methods used for the women. Figure 7 shows events with Jaccard overlaps greater than 0.35.

Unimodal Analysis of Two-Mode Data

In general, analysis of 2-mode data transformed into valued 1-mode networks proceeds like any other valued network. As with visualization, this often means generating a graph from the valued data via some rule such as $(u, v) \in E$ iff $a_{ij} > q$, where q is chosen by the researcher. Typically, there is no theoretical reason for choosing any particular value of q; hence a series of different values is generally chosen and the analysis repeated for each.

There are, however, a few consequences that stem from the 2-mode origin of the data. By their very nature, many commonly used measures of similarity and dissimilarity satisfy triangle inequality laws. For example, for Euclidean distance, every triple of nodes i, j, k satisfy the following rule:

$$d_{ik} \leq d_{ij} + d_{jk} \,. \tag{5}$$

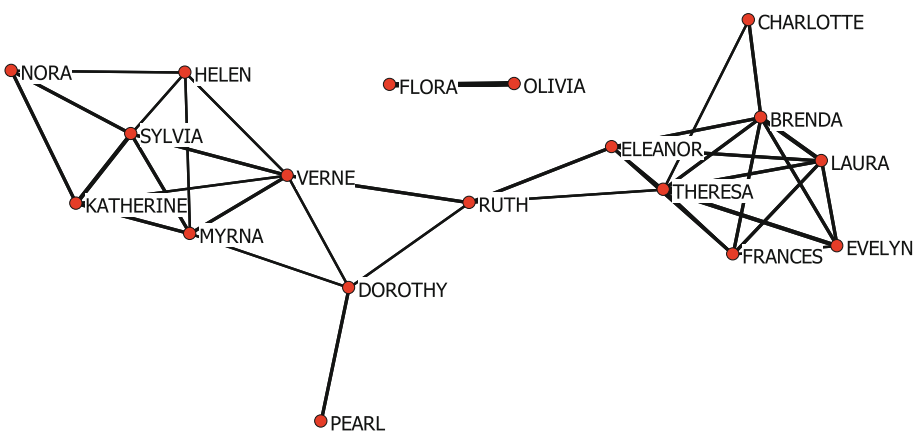

Social Network Analysis, Two-Mode Concepts in, Figure 6
Spring-embedding representation of Jaccard similarities dichotomized at > 0.4

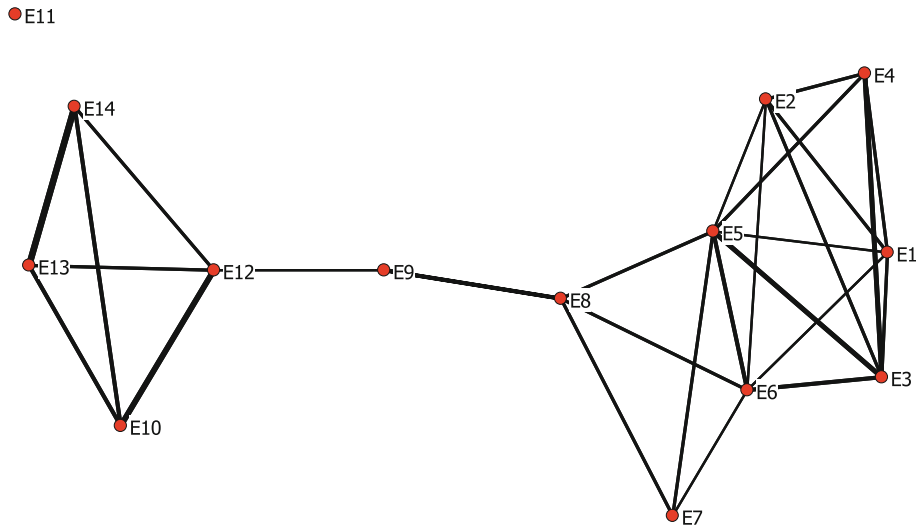

Social Network Analysis, Two-Mode Concepts in, Figure 7
Spring-embedding representation of "ties" among events ($c_{ij} > 0.35$)

As a result, the 1-mode data (especially if not dichotomized) artifactually exhibit a certain level of transitivity that may be higher than baseline models built on simple sociometric choice data would expect. Statistics based on transitivity, such as structural holes and clustering coefficients, must similarly be interpreted with some caution in such data.

Bimodal Approaches to Two-Mode Data

Another approach to working with 2-mode data seeks to analyze both modes simultaneously. The data are seen to represent relations between two sets of nodes, forming a bipartite graph $G_B(V_1 + V_2, E)$ in which, or all u and v, $(u, v) \in E$ if and only if u and v belong to different vertex sets. In other words, all ties are between vertex sets and none are within-group. The matrix representation of such a graph can be a rectangular incidence matrix X (as in Fig. 1) or a square bipartite adjacency matrix B with $n = n_1 + n_2$ rows representing both modes, and an equal number of columns, also representing both modes. In the latter case, the original matrix X forms a submatrix of the larger adjacency matrix B in which both rows and columns index the $V_1 + V_2$ entities. The matrix B is composed of four blocks, two of which are empty, as shown in Fig. 8 Note that the original matrix X forms the top right quadrant of B, and its transpose forms the bottom left quadrant.

Bimodal Visualization of Two-Mode Data

All of the standard ways to visualize networks, such as MDS and GLAs, apply to bipartite graphs. For example, Fig. 9 shows a spring-embedding layout of the bipartite graph represented by the matrix in Fig. 8.

In the representation, two nodes are near each other roughly to the extent that the geodesic distance between them is short. Thus, events are near each other if they are attended by the same women (distance 2), and women are near each other if they attend the same events. In this example, the representation makes clear that there is a set of women on the left (Mryna, Helen, Katherine, Nora, Silvia, etc.) that attend a set of events exclusive to them (events 10 through 13), and another set of women (Evelyn, Theresa, Laura, Brenda, etc.) that have their own events (E1 through E4), and finally a set of events that both "circles" of women attend (events E6 through E9).

For small datasets, this bimodal visualization is often extremely effective for transmitting a holistic understanding of the whole dataset.

It is worth noting that there is a simple mathematical relationship between pairwise overlaps a_{ij} as defined in Eq. (1) and paths in the bipartite graph. Specifically, the number of 2-step paths between any pair of women i and j in the bipartite graph is equal to a_{ij}, the number of events they attended in common. Of course, the number of 2-step paths is simply the matrix product BB, the bipartite adjacency matrix multiplied by itself. As shown in Fig. 10, the top left block and bottom right blocks of BB are the matrices A as calculated by Eq. (1) for rows and columns respectively of X.

Bimodal Analysis of Two-Mode Data

Given a bimodal perspective, one approach to analyzing 2-mode data is to develop entirely new metrics and algo-

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	1	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	1	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	1	1	0	1	0	1	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	1	1	1	1	1	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Social Network Analysis, Two-Mode Concepts in, Figure 8
Bipartite adjacency matrix *B* created from the original DGG 2-mode matrix *X*

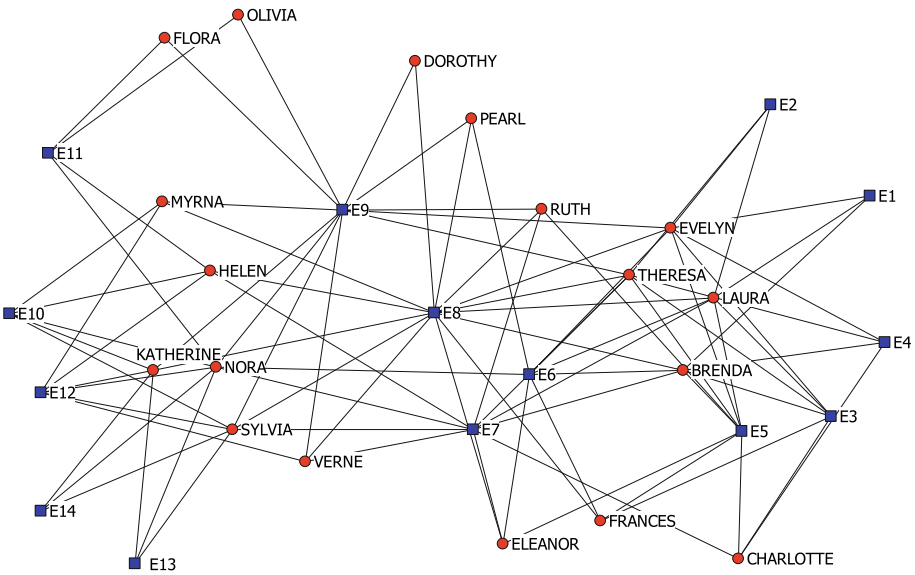

Social Network Analysis, Two-Mode Concepts in, Figure 9
Spring-embedding representation of bipartite graph

rithms designed specifically for 2-mode data. Such techniques take cognizance of the fact that the observed network is not just bipartite by happenstance, but could not have been any other way. In other words, taking account of the fact that the observed lack ties between certain nodes (namely, those belonging to different modes) was by design – similar to the concept of structural zeros in log-linear modeling. To date, few techniques of this kind have

	EVE	LAU	THE	BRE	CHA	FRA	ELE	PEA	RUT	VER	MYR	KAT	SYL	NOR	HEL	DOR	OLI	FLO	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14
EVELYN	8	6	7	6	3	4	3	3	3	2	2	2	2	2	1	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LAURA	6	7	6	6	3	4	4	2	3	2	1	1	2	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
THERESA	7	6	8	6	4	4	4	3	4	3	2	2	3	3	2	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BRENDA	6	6	6	7	4	4	4	2	3	2	1	1	2	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CHARLOTTE	3	3	4	4	4	2	2	0	2	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FRANCES	4	4	4	4	2	4	3	2	2	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ELEANOR	3	4	4	4	2	3	4	2	3	2	1	1	2	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PEARL	3	2	3	2	0	2	2	3	2	2	2	2	2	2	1	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RUTH	3	3	4	3	2	2	3	2	4	3	2	2	3	2	2	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VERNE	2	2	3	2	1	1	2	2	3	4	3	3	4	3	3	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MYRNA	2	1	2	1	0	1	1	2	2	3	4	4	4	3	3	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KATHERINE	2	1	2	1	0	1	1	2	2	3	4	6	6	5	3	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SYLVIA	2	2	3	2	1	1	2	2	3	4	4	6	7	6	4	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NORA	2	2	3	2	1	1	2	2	2	3	3	5	6	8	4	1	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HELEN	1	2	2	2	1	1	2	1	2	3	3	3	4	4	5	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DOROTHY	2	1	2	1	0	1	1	2	2	2	2	2	2	1	1	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OLIVIA	1	0	1	0	0	0	0	1	1	1	1	1	1	2	1	1	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FLORA	1	0	1	0	0	0	0	1	1	1	1	1	1	2	1	1	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	2	3	2	3	3	2	3	1	0	0	0	0	0
E2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3	3	2	3	3	2	3	2	0	0	0	0	0
E3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	6	4	6	5	4	5	2	0	0	0	0	0
E4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	4	4	4	3	3	3	2	0	0	0	0	0
E5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	6	4	8	6	6	7	3	0	0	0	0	0
E6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	5	3	6	8	5	7	4	1	1	1	1	1
E7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	4	3	6	5	10	8	5	3	2	4	2	2
E8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	5	3	7	7	8	14	9	4	1	5	2	2
E9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	2	3	4	5	9	12	4	3	5	3	3
E10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	4	4	5	2	5	3	3
E11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	1	3	2	4	2	1	1
E12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	4	5	5	5	2	6	3	3
E13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	3	3	1	3	3	3
E14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	3	3	1	3	3	3

Social Network Analysis, Two-Mode Concepts in, Figure 10
Matrix *BB* giving the number of paths of length 2 between all pairs of nodes

been developed, the exception being the area of 2-mode centrality measures, which has received significant attention (e. g., [2]).

Another approach is to treat the bipartite graphs as ordinary graphs and apply all the standard algorithms and techniques of social network analysis. Effectively this assumes either that the special nature of the graphs will not affect the techniques, or that we can pretend that ties within modes could have occurred and just didn't. This approach works for a small class of methods, but by no means all. For example, calculating transitivity fails because transitive triples are impossible in bipartite graphs (all ties are between modes, which means that if $a \to b$ and $b \to c$ then a and c must be members of the same class, and therefore cannot be tied, making transitivity impossible). In contrast, if we were to adapt the definition of transitivity to be based on quadruples such that a quad is transitive if $a \to b$, $b \to c$, $c \to d$ and $a \to d$, this would be an example of the first approach.

Finally, a compromise approach is to use the standard metrics and algorithms that apply to general graphs, and then either adjust the outputs via normalization or adjust the baseline expectations for the results [5,9]. In the former case, we develop normalizations that adjust the metrics (typically by dividing by theoretical bipartite maxima), and in the latter case we derive different theoretical distributions for the statistics in question. In general, we choose the former approach when statistics are bounded between 0 and 1 and can be interpreted as proportions of maximum possible values. We choose the latter approach when the statistics are unbounded and have direct interpretations (see example of cohesion below).

Bimodal Approaches to Graph Cohesion The simplest and most common measure of network cohesion is *density* – the number of edges divided by the number of pairs of nodes (using ordered pairs in the case of directed graphs and unordered pairs in the case of undirected graphs). In bipartite graphs, of course, only edges between vertex sets are possible. As a result, the maximum possible undirected ordinary density is

$$\frac{n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 1)} \,. \tag{6}$$

Thus, if density were calculated on a 2-mode network as if it were an ordinary graph, we would probably want to normalize the result by dividing by the sum above, otherwise the calculated density would appear misleadingly low. This is a case where the "compromise approach" discussed above is effective. (Of course, in terms of computation, it would be easier to simply calculate the average of the 2-mode matrix X rather than convert to bipartite form and then perform this adjustment.)

For measures of cohesion that are not expressed as fractions of maximum possible values, such as average geodesic path length, the need to renormalize is not as

great, since the raw values are directly interpretable. In doing so, however, we must be careful not to compare the results with standard rules of thumb or theoretical models based on simple random graphs, as these have not assumed the bipartite restrictions.

Bimodal Approaches to Centrality Four measures of centrality are used with high frequency in social network analysis today: degree, closeness, betweenness and eigenvector. We discuss each in turn.

Degree centrality, d_i, is defined as the number of ties incident upon node i. Degree centrality is typically normalized by dividing by the maximum number of ties possible, which in a graph of n nodes is $n-1$. Hence

$$d_i^* = \frac{d_i}{n-1} \,. \tag{7}$$

However, in any (non-trivial) bipartite graph, no node can be connected to all others, since this would mean within-mode ties. Instead, for a node in V_1, the maximum number of ties it could have is n_2, whereas for a node in V_2, the maximum number of ties is n_1. Hence a natural adjustment is to provide two separate normalization formulas as follows:

$$\begin{aligned} d_i^* &= \frac{d_i}{n_2} \,, \quad \text{for } i \in V_1 \,, \\ d_j^* &= \frac{d_j}{n_1} \,, \quad \text{for } j \in V_2 \,. \end{aligned} \tag{8}$$

Closeness centrality is ordinarily defined as the sum of geodesic distances from a node to all others, as shown in Eq. (8), where d_{ij} is the length of the shortest path from i to j and $n = n_1 + n_2$ is the total number of nodes. The best (smallest) score possible occurs when the node has a tie to every other node, in which case the total distance to all others is $n-1$. Thus, closeness centrality is usually normalized by dividing c_i into $n-1$.

$$\begin{aligned} c_i &= \sum_j^n d_{ij} \\ c_i^* &= \frac{n-1}{c_i} \,. \end{aligned} \tag{9}$$

In the bipartite case, the maximum number of nodes that a node can be distance 1 from is the number of nodes in the other class. For other nodes in its own class, the closest it can be is two links away. Thus, the theoretical bipartite minimum value of c_i (where $i \in V_1$) is $n_2 + 2(n_1 - 1)$ and the minimum for c_j (where $j \in V_2$) is $n_1 + 2(n_2 - 1)$. Therefore, closeness centrality can be normalized in bipartite graphs as follows:

$$\begin{aligned} c_i^* &= \frac{n_2 + 2(n_1 - 1)}{c_i} \,, \quad \text{for } i \in V_1 \,, \\ c_j^* &= \frac{n_1 + 2(n_2 - 1)}{c_j} \,, \quad \text{for } j \in V_2 \,. \end{aligned} \tag{10}$$

Betweenness centrality refers to the sum of shares of shortest paths that pass through a given node. The betweenness of node k in an ordinary graph is defined by Eq. (11), where g_{ij} is the number of geodesic paths from node i to node j, and g_{ikj} is the number of geodesic paths from i to j that pass through k.

$$b_k = \frac{1}{2} \sum_{i \neq k}^{n} \sum_{j \neq k,i}^{n} \frac{g_{ikj}}{g_{ij}} \,. \tag{11}$$

Betweenness is ordinarily normalized by dividing by $(n-1)(n-2) = n^2 - 3n + 2$, which is the maximum betweenness that any node can achieve in a graph with n nodes, which occurs for the node at the center of a star-shaped graph. This maximum is appropriate for bipartite graphs only when one mode has just one node; otherwise we must take account of the sizes of each vertex set. Eq. (12) gives the maximums for nodes in each vertex set as a function of the vertex set sizes. In the equation, x div y refers to integer division of x by y and x mod y refers to the remainder of an integer division of x by y.

$$\begin{aligned} b_{V_1\,\max} &= \tfrac{1}{2}\big[n_2^2(s+1)^2 + n_2(s+1)(2t-s-1) \\ &\qquad - t(2s-t+3)\big] \\ s &= (n_1 - 1) \text{ div } n_2 \,, \quad t = (n_1 - 1) \mod n_2 \\ b_{V_2\,\max} &= \tfrac{1}{2}\big[n_1^2(p+1)^2 + n_1(p+1)(2r-p-1) \\ &\qquad - r(2p-r+3)\big] \\ p &= (n_2 - 1) \text{ div } n_1, \quad r = (n_2 - 1) \mod n_1 \end{aligned} \tag{12}$$

Given these maxima, we can normalize standard betweenness centrality for bipartite graphs by dividing by these maxima, as shown in Eq. (13).

$$\begin{aligned} b_i^* &= \frac{b_i}{b_{V_1\,\max}} \,, \quad \text{for } i \in V_1 \,, \\ b_j^* &= \frac{b_j}{b_{V_2\,\max}} \,, \quad \text{for } j \in V_2 \,. \end{aligned} \tag{13}$$

Cohesive Subgroups Detecting cohesive subgroups is somewhat more difficult in bipartite graphs than in ordinary graphs. For example, one of the earliest formal definitions of a subgroup is the clique [11] which is defined

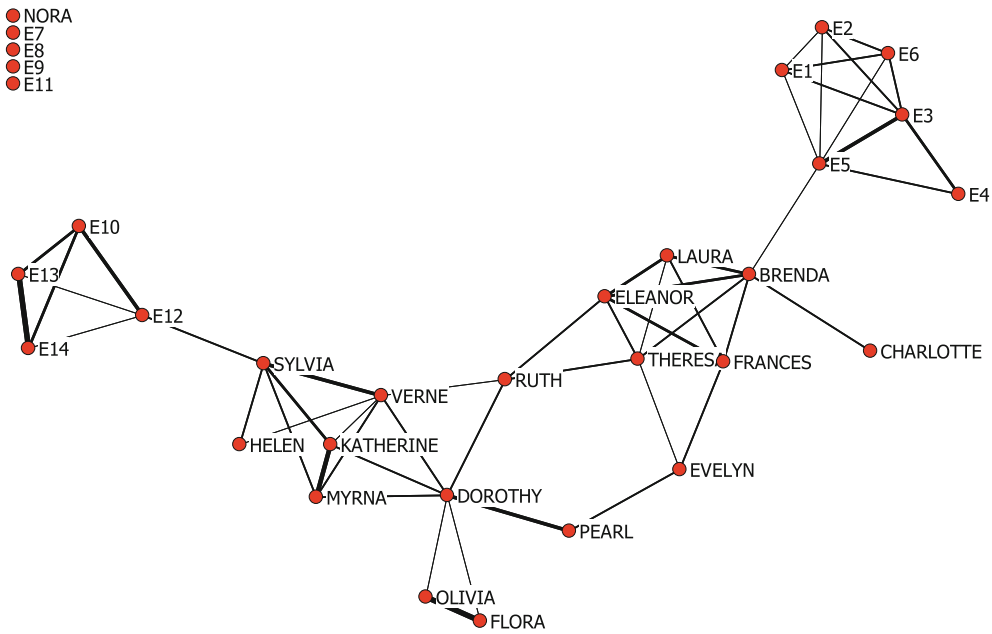

Social Network Analysis, Two-Mode Concepts in, Figure 11
Graph induced by Pearson correlations across biclique membership profiles. Edges indicate a correlation greater than 0.4

as a maximal complete subgraph – i. e., a set of nodes that is as large as possible subject to the condition that every pair is adjacent. However, in bipartite graphs, cliques are impossible, since in any connected triple, two nodes must be non-adjacent. Thus, cliques are not useful in this context. In addition, relaxations of the clique concept based on density or frequency of paths (such as k-plexes, lambda sets and ls-sets) do not work well, due to the sparse nature of bipartite graphs. For example, if we calculate k-plexes for the DGG dataset, we find huge numbers of very small groups. For $k = 2$, there are 394 2-plexes of size 3 or greater. For $k = 3$, there are 5,553 3-plexes, and for $k = 4$, there are 37,633 4-plexes of size 3 or greater.

However, other classical notions of cohesive subgroups make more sense for bipartite graphs. In particular, relaxations of the clique concept based on distance, such as n-cliques, n-clans and n-clubs have good interpretations in bipartite graphs and work well in practice. In fact, 2-cliques defined on the bipartite graph have the property of bipartite completeness, meaning that, within the bipartite subgraph induced by the 2-clique, every node of one mode has a tie with every node of the other mode, which is to say that all possible ties are present. In this sense, for bipartite graphs, 2-cliques capture the underlying idea of a clique better than cliques do. This notion has been formalized in the definition of a *biclique*, which is defined as a maximal complete bipartite subgraph. Mathematically, it is identical to a 2-clique computed on the bipartite representation.

As a practical example, a biclique analysis of the DGG dataset finds 68 bicliques, and a secondary analysis of similarities of nodes across biclique membership profiles does a good job of revealing structure, as shown in Fig. 11. In the figure, an edge is shown between two nodes if the Pearson correlation between their biclique membership profiles is greater than 0.4 (i. e., $(u, v) \in E$ iff $r_{ij} > 0.4$). The two groups are women are clearly shown, as well as the two groups of events that are associated with each of the groups of women. In addition, the separation of Flora and Olivia is clearly shown, as well as the bridging position of Ruth and Pearl.

The success of bicliques as 2-mode analogues of cliques suggests the possibility of creating 2-mode analogues for other cohesive subgroup concepts that we previously dismissed as unusable in the 2-mode context. For example, a *k-plex* is defined in ordinary network analysis as a maximal subgraph $S(V, E)$ of size n such that each member of the k-plex is adjacent to at least $n - k$ others. By analogy, for 2-mode networks we can define a (k_1, k_2)-*biplex* as a maximal bipartite subgraph $G(V_1, V_2, E)$ such that each node in V_1 is adjacent to $|V_2| - k_2$ others and each node in V_2 is adjacent is adjacent to $|V_1| - k_1$ others. Obviously, a (0,0)-biplex is a biclique. A useful feature of this relaxation of a biclique is that we can set different standards of

cohesiveness for each mode of the data network, perhaps reflecting the relative sizes of the modes, or the affiliative capabilities of the nodes themselves.

Bimodal Approaches to Positions and Roles Two concepts of position and role are particularly well known in social network analysis. These are structural equivalence and regular equivalence. Generalizations to 2-mode (and higher) data were developed by Borgatti [3] and Borgatti and Everett [4] and Everett and Borgatti [7]. We consider each of these in turn.

Structural Equivalence In the best definition of structural equivalence [8], two nodes u and v are said to be structurally equivalent if there exists a graph automorphism that would be an identity except that u and v are mapped to each other. In other words, given a diagram of the graph, you could swap nodes u and v (and no other nodes) without changing the structure of the network one iota. A key implication of this definition is that structurally equivalent nodes have identical relational environments – aside from each other, they are connected and not connected to exactly the same third parties. As a result, one approach to identifying structurally equivalent nodes is to compute a similarity measure among rows and columns of the adjacency matrix defining the graph.

This approach requires no modification for the bipartite case. In fact, the unimodal analysis described in Sect. "Unimodal Approaches to Two-Mode Data" (applied to each mode in turn) is precisely the same as computing structural equivalence on the 2-mode incidence matrix (Fig. 1). In addition, for certain measures such as the Jaccard coefficient, computing structural equivalence on the bipartite adjacency matrix B of Fig. 8 gives exactly the same results as computing it on the 2-mode incidence matrix X.

Another approach to structural equivalence is known as blockmodeling. Instead of (or as a result of) measuring the extent of structural equivalence between nodes, we partition the nodes into classes such that nodes in the same class are structurally equivalent (or nearly so). Given such a partition of nodes, we can reorder and partition the corresponding rows and columns of the adjacency matrix. This in turn induces to partition of matrix cells into *matrix blocks*. A characteristic property of perfect structural equivalence partitions is that matrix blocks are necessarily homogeneous with respect to cell values, and each block will consist of all 1s or all 0s (known as 1-blocks and 0-blocks respectively). Figure 12 gives an example with three equivalence classes.

In the 2-mode case, one approach is to blockmodel the bipartite adjacency matrix B. This can be done, but the bipartite structure imposes certain constraints. For example, blocks involving within-mode ties must be 0-blocks. In addition, the best 2-block partition will almost certainly be the mode partition (except in trivial cases), and in general, all other partitions will be refinements of the mode partition (i. e., they will be nested hierarchically within the mode partition).

	A1	A2	A3	B1	B2	B3	B4	C1	C2	C3
A1	0	0	0	1	1	1	1	0	0	0
A2	0	0	0	1	1	1	1	0	0	0
A3	0	0	0	1	1	1	1	0	0	0
B1	1	1	1	0	0	0	0	1	1	1
B2	1	1	1	0	0	0	0	1	1	1
B3	1	1	1	0	0	0	0	1	1	1
B4	1	1	1	0	0	0	0	1	1	1
C1	1	1	1	1	1	1	1	0	0	0
C2	1	1	1	1	1	1	1	0	0	0
C3	1	1	1	1	1	1	1	0	0	0

Social Network Analysis, Two-Mode Concepts in, Figure 12
Perfect structural equivalence blockmodel

A more elegant (and computationally efficient) approach is to work from the 2-mode incidence matrix X. In this case, we redefine a blockmodel to refer to a coordinated pair of partitions, one for the rows and one for the columns. We then redefine structural equivalence as follows. Given a 2-mode matrix X with modes V_1 and V_2, a 2-mode blockmodeling consists of a pair of partitions P_1 and P_2 such that (a) for all u, v in V_1, $p_1(u) = p_1(v)$ iff $x_{uw} = x_{vw}$ for all w in V_2 and (b) for all a, b in V_2, $p_2(a) = p_2(b)$ iff $x_{za} = x_{zb}$ for all z in V_1. (The notation $p(u)$ indicates the equivalence class that node u belongs to in partition P.) Restating this in words, a 2-mode structural equivalence blockmodeling is one in which row nodes in the same class if and only if they have identical rows, and column nodes are in the same class if and only if they have identical columns. An example involving 4 classes of rows and 3 classes of columns is shown in Fig. 13.

In empirical work, of course, we do not expect to obtain perfect 1-blocks and 0-blocks. Instead we seek partitions that minimize the number of errors (where we define an error as either a 1 inside a 0-block or a 0 inside a 1-block).

Regular Equivalence Two nodes are said to be *regularly equivalent* (i. e., play the same structural role) to each other to the extent that they have ties (lack of ties) to corresponding others who are themselves regularly equivalent to each other. That is, u and v are regularly equivalent

	E1	E2	E3	F1	F2	F3	F4	G1	G2	G3
A1	1	1	1	1	1	1	1	0	0	0
A2	1	1	1	1	1	1	1	0	0	0
A3	1	1	1	1	1	1	1	0	0	0
B1	1	1	1	0	0	0	0	0	0	0
B2	1	1	1	0	0	0	0	0	0	0
B3	1	1	1	0	0	0	0	0	0	0
B4	1	1	1	0	0	0	0	0	0	0
C1	0	0	0	1	1	1	1	0	0	0
C2	0	0	0	1	1	1	1	0	0	0
C3	0	0	0	1	1	1	1	0	0	0
D1	0	0	0	1	1	1	1	1	1	1
D2	0	0	0	1	1	1	1	1	1	1

Social Network Analysis, Two-Mode Concepts in, Figure 13
A 2-mode structural equivalence blockmodel in which the row partition has 4 classes and the column partition has 3 classes

if, (a) for all x such that u has a tie to x, there exists a y that is regularly equivalent to x such that v has a tie to y, and (b) for all x such that u has an incoming tie from x, there exists a y that is regularly equivalent to x such that v has an incoming tie from y. Whereas in structural equivalence equivalent nodes are connected to the same others, in regular equivalence equivalent nodes are connected to the same types of others.

Alternatively, we can define regular equivalence in terms of a partition C of nodes into labeled classes (known as colors) such that regularly equivalent nodes are required to have the same colors in their neighborhoods. That is, for all nodes u and v,

$$C(u) = C(v) \text{ implies } C(N(u)) = C(N(v)) \,, \tag{14}$$

where $C(N(u))$ is the set of distinct colors found among the nodes constituting u's immediate neighborhood.[1] Taking a blockmodeling perspective, it can readily be seen that Eq. (14) implies a blockmodel in which each matrix block is either entirely zero (a "zeroblock") or contains at least one 1 in every column and in every row (a "regular oneblock"), as illustrated in Fig. 14.

Equation (14) lends itself easily to the bipartite case. Instead of seeking a single partition of nodes, we seek a different partition for each mode. Labeling these partitions R and C, we modify our definition such that for all $u, v \in V_1$ and $x, y \in V_2$,

$$\begin{aligned} C(u) = C(v) \text{ implies } R(N(u)) = R(N(v)) \,, \quad \text{and} \\ R(x) = R(y) \text{ implies } C(N(x)) = C(N(y)) \,. \end{aligned} \tag{15}$$

[1]For directed graphs, we require $C(N^o(u)) = C(N^o(v))$ and $C(N^i(u)) = C(N^i(v))$, where $N^o(v)$ refers to the set of nodes that v sends a tie to, and $N^i(u)$ refers to the set of nodes that v receives a tie from.

	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10
N1	0	0	0	1	0	0	1	0	0	0
N2	0	0	0	0	1	0	0	0	0	0
N3	0	0	0	1	1	1	1	0	0	0
N4	1	0	0	0	0	0	0	0	1	1
N5	1	0	1	0	0	0	0	1	1	0
N6	0	1	0	0	0	0	0	0	1	1
N7	0	1	1	0	0	0	0	1	0	1
N8	1	0	1	0	1	1	0	0	0	0
N9	0	0	1	0	0	1	0	0	0	0
N10	0	1	0	1	1	0	1	0	0	0

Social Network Analysis, Two-Mode Concepts in, Figure 14
Perfect regular equivalence blockmodel on an ordinary graph

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
C1	1	0	1	0	1	1	0	0	0	0
C2	0	0	1	0	0	1	0	0	0	0
C3	0	1	0	1	1	0	1	0	0	0
C4	1	0	0	0	0	0	0	0	1	1
C5	1	0	1	0	0	0	0	1	1	0
C6	0	1	0	0	0	0	0	0	1	1
C7	0	1	1	0	0	0	0	1	0	1
C8	0	0	0	0	1	1	0	0	0	0
C9	0	0	0	0	0	1	0	0	0	0
C10	0	0	0	1	1	0	1	0	0	0
C11	1	0	1	0	0	0	0	0	1	1
C12	0	1	0	0	0	0	0	1	0	1

Social Network Analysis, Two-Mode Concepts in, Figure 15
A 2-mode regular equivalence blockmodel

Of course, the neighbors N() of any node consist entirely of nodes belonging to the other mode. Thinking in terms of the 2-mode incidence matrix shown in Fig. 1, Eq. (15) implies that we can section the matrix into rectangular blocks such that each block is zeroblock or a regular 1-block. An example of a 2-mode regular blockmodel is shown in Fig. 15.

As with structural equivalence, in empirical work we do not expect to obtain perfect 1-blocks and 0-blocks. Instead we seek partitions induce matrix blocks that are as nearly regular as possible.

Conclusion

In this chapter we have considered methods of visualizing and analyzing 2-mode network data. An examination of the chapter suggests that four strategies are employed for handling 2-mode data. First, there is converting the data into two separate 1-mode datasets, and analyzing each separately. Second, there is using existing 1-mode methods on a bipartite representation of the 2-mode data, and essentially ignoring its special features. Third, there is using 1-mode methods but either adjusting the interpretation or

applying some kind of normalization to adjust the results. Finally, there is developing new methods designed specifically for the 2-mode case. The chapter has provided examples of all four strategies.

It is worth reminding the reader that only 2-mode data viewed in a network context has been considered here. In principle, most data dealt with by social scientists is structured as a 2-mode matrix of cases (rows) and variables (columns) and most statistical techniques assume such data. This chapter obviously does not discuss this vast pantheon of techniques, although it is clear that many would be appropriate, particularly structure-finding techniques such as factor analysis and correspondence analysis.

Future Directions

Although it would be a mistake to think of 2-mode data as an advance over 1-mode data, it is important to note that there are many cases were extending network analysis methodology to more than 2 modes is desirable. For example, we might analyze membership in organizations over time, yielding a 3-way, 3-mode data matrix of relations that is person by organization by period. Similarly, we might be interested in modeling the intellectual landscape of an academic field, representing publications as k-mode bundles of authors, journals, years, institutional affiliations and topics.

Bibliography

1. Bonacich P (1972) Techniques for analyzing overlapping memberships. In: Costner HL (ed) Sociological methodology. Jossey-Bass, San Francisco, pp 176–185
2. Bonacich P (1991) Simultaneous group and individual centralities. Soc Netw 13:155–168
3. Borgatti SP (1989) Regular equivalence in graphs, hypergraphs, and matrices. Doctoral Dissertation, Univ of California, Irvine, Ann Arbor
4. Borgatti SP, Everett M G (1992) Regular blockmodels of multiway, multimode matrices. Soc Netw 14:91–120
5. Borgatti SP, Everett MG (1997) Network analysis of 2-mode data. Soc Netw 19(3):243–269
6. Davis A, Gardner B, Gardner M (1941) Deep South: A Social Anthropological Study of Caste and Class. University of Chicago Press, Chicago
7. Everett MG, Borgatti SP (1993) An extension of regular colouring of graphs to digraphs, networks and hypergraphs. Soc Netw 15:237–254
8. Everett MG, Borgatti SP (1994) Regular equivalence: General theory. J Math Sociol 19(1):29–52
9. Faust K (1997) Centrality in affiliation networks. Soc Netw 19:157–191
10. Kamada T, Kawai S (1989) An algorithm for drawing general undirected graphs. Inf Process Lett 31:7–15
11. Luce RD, Perry AD (1949) A method of matrix analysis of group structure. Psychometrika 20:319–327

Social Networks, Algebraic Models for

Philippa Pattison
Department of Psychology, University of Melbourne, Parkville, Australia

Article Outline

Glossary
Definition of the Subject
Introduction
Algebras from Networks
Algebraic Structure
Algebraic Analysis
Algebraic and Network Mappings
Examples
Future Directions
Bibliography

Glossary

Social network A *social network* comprises a set of relationships among members of a set $N = \{1, 2, \ldots, n\}$ of actors. It can be represented by an $n \times n$ binary array X recording the presence or absence of a social relationship, or *tie*, between each pair (or ordered pair) of members of $N = \{1, 2, \ldots, n\}$. If there is a relationship from actor k to actor l, we write $X(k, l) = 1$; otherwise, $X(k, l) = 0$. If the relationship is a property of a pair of actors, the network is *nondirected*; if it is a property of an ordered pair, the network is termed *directed*. The directed network X may also be regarded as a *binary relation* R_X on the set N with $(k, l) \in R_X$ if and only if $X(k, l) = 1$; equivalently, it may be construed as a *directed graph* with *node set* N and *arc set* R_X, with an *arc* from node k to node l if and only if $(k, l) \in R_X$.

Affiliation network An *affiliation network* is an $n \times g$ binary array X recording the membership of each of a set N of actors in a prescribed set G of *groups*, with $X(k, l) = 1$ if actor k is a member of group l, and $X(k, l) = 0$ otherwise.

Multiple network A *multiple network* is a collection of networks for each of a set of r relations. We let $X_m(k, l) = 1$ if the tie from k to l corresponding to the relation of *type* m is present; and $X_m(k, l) = 0$ if the tie is absent. Nodes k and l are joined by a *labeled walk* with *label* $Y_1\, Y_2 \ldots Y_j$ if there is a sequence of nodes $k = k_0, k_1, \ldots, k_j = l$, for which $Y_h(k_{h-1}, k_h) = 1$ for $h = 1, 2, \ldots, j$.

Local network The *(1-)neighborhood* of a subset P of actors in a network is defined to be the set $P \cup \{l \in N\colon X_m(k,l) = 1$ for some $k \in P$ and some relation $m\}$. The *q-neighborhood* of P is then defined recursively as the 1-neighborhood of the $(q-1)$-neighborhood of P. The *q-local network* of the subset P of N is the network restricted to its q-neighborhood.

Algebra A (*partially ordered*) *algebra* is a triple $[S, F, \leq]$, where S is a nonempty set of elements (usually assumed to be finite), F is a specified set of operations, f_α, each mapping a power $S^{n(\alpha)}$ of S into S, for some non-negative finite integer $n(\alpha)$, and $\leq$ is a partial order on S. Each operation f_α is assumed to be *isotone* in each of its variables: that is, if $x_i \leq y_i\,(x_i, y_i \in S;\ i = 1, 2, \ldots, n(\alpha))$, then $f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)}) \leq f_\alpha(y_1, y_2, \ldots, y_{n(\alpha)})$. A *family of algebras* is a collection of algebras each having the same set F of operations and satisfying a specified set of postulates. Two algebras belonging to the same family are termed *similar*.

Partial algebra A *partial algebra* is a triple $[S, F, \leq]$, where S is a nonempty set of elements (usually assumed to be finite), F is a specified set of *partial* operations, f_α, each mapping some subset $T^{(\alpha)}$ of $S^{n(\alpha)}$ into S, for some non-negative finite integer $n(\alpha)$, and $\leq$ is a partial order on S. Each partial operation f_α is assumed to be *isotone* in each of its variables: that is, if $x_i \leq y_i\,(x_i, y_i \in S;\ i = 1, 2, \ldots, n(\alpha))$, then $f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)}) \leq f_\alpha(y_1, y_2, \ldots, y_{n(\alpha)})$ provided that both $(x_1, x_2, \ldots, x_{n(\alpha)})\,, (y_1, y_2, \ldots, y_{n(\alpha)}) \in T^{(\alpha)}$. A *family of partial algebras* is a collection of algebras each having the same set F of partial operations defined on the same subsets $T^{(\alpha)}$ of the power sets $S^{n(\alpha)}$.

Semigroup A *(partially ordered) semigroup* is an algebra $[S, F, \leq]$ in which F comprises a single binary operation f satisfying the *associativity* condition:

$$f(f(x, y), z) = f(x, f(y, z))\,.$$

Lattice A *lattice* L is an algebra $[S, F, \leq]$ in which F comprises two associative and commutative binary operations, $\wedge$ and $\vee$ (termed *meet* and *join*, respectively) satisfying the identities:

$$x \wedge x = x\,, \quad x \vee x = x$$

and

$$x \wedge (x \vee y) = x \vee (x \wedge y) = x\,.$$

The operations are isotone, so that $x \leq y$ is equivalent to the pair of conditions:

$$x \wedge y = x \quad \text{and} \quad x \vee y = y\,,$$

and the operations of meet and join may be interpreted as the greatest lower bound and least upper bound, respectively. A lattice L is *distributive* if the identity

$$x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$$

holds. A lattice L is *modular* if, whenever $x \leq z$, then

$$x \vee (y \wedge z) = (x \vee y) \wedge z\,.$$

Role algebra A *role algebra* is an algebra $[S, F, \leq]$ in which F comprises a single binary composition operation satisfying the condition: $s \leq t$ in S implies $su \leq tu$ in S, for any $u \in W$.

Homomorphism An (*isotone*) *homomorphism* from an algebra $A = [S, F, \leq]$ onto a similar algebra $B = [T, F, \leq]$ is a mapping $\phi\colon S \to T$ such that,

(i) for all $f_\alpha \in F$ and $x_i \in S$,

$$\phi(f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)})) = f_\alpha(\phi(x_1)\,, \phi(x_2), \ldots, \phi(x_{n(\alpha)}))\,; \quad \text{and}$$

(ii) $x \leq y$ in S implies $\phi(x) \leq \phi(y)$ in T.

The algebra B is termed a (*homomorphic*) *image* of A, and we write $B = \phi(A)$. Each homomorphism ϕ from $A = [S, F, \leq]$ onto $B = [T, F, \leq]$ has a corresponding binary relation π on S (termed here a *π-relation*) in which $(x, y) \in \pi$ if and only if $\phi(y) \leq \phi(x)$. The equivalence relation σ_π defined by $(x, y) \in \sigma_\pi$ if and only if $(x, y) \in \pi$ and $(y, x) \in \pi$ is termed a *congruence relation*.

Homomorphism lattice The *homomorphism lattice* $L(A)$ of the algebra $A = [S, F, \leq]$ is the collection of all homomorphisms of A partially ordered by the relation: $\phi_1 \leq \phi_2$ if, for all $x, y \in S$, $\phi_2(x) \leq \phi_2(y)$ implies $\phi_1(x) \leq \phi_1(y)$. The *lattice $L_\pi(A)$ of π-relations* on A, dual to $L(A)$, has the partial ordering: $\pi_1 \leq \pi_2$ if $(x, y) \in \pi_1$ implies $(x, y) \in \pi_2$, for any $x, y \in S$.

Definition of the Subject

Algebraic models have been proposed to represent structure in social networks. They are usually constructed from a set of operations and relations defined on network constituents such as paths or walks in the network, or vectors of ties directed to or from individual network members. The algebra represents the relationships among these

constituents, for example, relations of ordering among all possible walks in a multiple network, or relations of overlap and ordering among profiles of group membership. To date, they have been used to represent kinship structures [6,9,48], role structures and stability in multiple networks [1,5,8,36], states of diffusion processes in networks [30], informal hierarchy [17], connectivity structures [13] and the structure of membership in affiliation networks [16,38]. In several cases, partial algebraic representations have also been proposed [35,39,40].

Algebraic models have usually been proposed as exact structural representations and, in many cases, therefore, their application to network data has been coupled with some form of prior data aggregation such as blockmodeling [5,50]; see also the entry on ► Positional Analysis and Blockmodeling. More rarely, algebraic construction and data aggregation have been combined into a single step [8,33].

Introduction

Algebraic models for relational structures such as networks have their foundations in anthropology and were developed for general network structures by Lorrain, White and colleagues [5,26,27]. These general network representations based on path structures in networks emerged from more specific models for path structures in kinship systems, including permutation group models for Australian aboriginal kinship systems [48]. Boyd extended this approach in an important way when he proposed that the structural evolution of kinship systems could be analyzed in terms of homomorphic mappings [6]; applications to network semigroups were described later [5,37].

Lattice representations for relational structures have a long history in mathematics [2] and applications to relational data in many domains have been developed extensively by Wille and collaborators in formal concept analysis [19,20,51]. They have also been applied to affiliation network structures [16,45,47].

The form of algebraic structure regarded as appropriate in a particular context is a matter for substantive consideration. The large majority of forms that have been proposed to date have depended upon the importance of paths or walks in the network, the symmetric or asymmetric nature of network ties, and the patterns of overlap among tie partners. A focus on one or more of these features has resulted in algebraic groups, semigroups, partially ordered semigroups, local role algebras and lattices as representations of different aspects of structure in multiple social networks and affiliation networks [5,8,14,16,29,36,48] and it is on these forms that we focus here.

Algebras from Networks

A Language for Discriminating Among Relations

We begin with a finite set $\Sigma = \{x_1, x_2, \ldots, x_r\}$ of r relational terms. We term Σ a relational *alphabet* and call each term x_i in Σ a *letter* of the alphabet. The letters $x_1, x_2, \ldots$ denote relational terms such as "friendship", "co-worker", "confidant", and so on. If $r = 1$, the set Σ simply identifies a single network relation.

In order to construct algebraic representations that are sensitive to social structural form, claims in the social network literature can be used to introduce a *free word algebra* on the alphabet Σ. The algebra is intended to provide a language for describing more complex relational terms that can be constructed from the primitive set Σ and so to provide the means for discriminating among social structural forms. The claims about appropriate ways of constructing new terms from the primitive set may be drawn from a number of sources, but one set of such claims is helpfully laid out by White [49]. White argues that there are three universal ways in which types of tie (i. e., relational terms) are discriminated. The first is in terms of patterns of overlap; the second in terms of asymmetry of relation; and the third in terms of the institutionalization of indirect ties. The overlap of two relational ties refers to their joint occurrence and can arguably be described using an *intersection* operation, $\cap$. The asymmetry of a tie can be described by the relationship between the tie and its *converse*, that is the coding of whether, for two individuals k and l, there is a tie from k to l and from l to k. The tie is *symmetric* if the tie from k to l occurs only in conjunction with the tie from l to k, and the tie is *strictly asymmetric* if a tie from k to l occurs only in the absence of a tie from l to k. We use the expression x' to denote the converse of the relational term x. Indirect ties may be described by a *composition*, or *concatenation* operation, denoted $^\circ$. If x_1 refers to the term "friendship" and x_2 to "coworker", then the concatenation $x_1{}^\circ x_2$ may be used to refer to a relation that holds between a person and his or her friend's co-workers.

We denote by F the set of operations from which our descriptive language is to be built. Unless otherwise specified, we assume here that F comprises the three operations just described; that is $F = \{\cap, ', ^\circ\}$. Each operation f_α in F possesses an *arity*, $n(\alpha)$, that determines the form of new terms created from Σ by f_α: in particular, f_α leads to expressions of the form $f_\alpha(y_1, y_2, \ldots, y_{n(\alpha)})$; $y_m \in \Sigma$, $m = 1, 2, \ldots, n(\alpha)$. The operations $\cap$ and $^\circ$ have arity 2, that is, they are *binary* and lead to expressions that can be written as $y_1 \cap y_2$, and $y_1{}^\circ y_2$, respectively, whereas

the converse operation $'$ is *unary* (that is, it has an arity of 1) and leads to an expression that can be denoted y_1' (where $y_1, y_2 \in \Sigma$).

We describe a recursive construction of a free word algebra $W = [\Sigma, F]$ from the alphabet Σ and the set F of operations [2]. We call each letter in Σ an F-polynomial of *rank* 0. For any positive integer ρ, we define an F-polynomial of rank ρ recursively as an expression, termed a *word*, of the form $f_\alpha(u_1, \ldots, u_{n(\alpha)})$, where each u_j is an F-polynomial of rank $\leq \rho - 1$ (and at least one u_j is an F-polynomial of rank $\rho - 1$). The *(free) word algebra* $W = [\Sigma, F]$ comprises all distinct F-polynomials of finite rank (where $f_\alpha(u_1, \ldots, u_{n(\alpha)})$ and $f_\beta(v_1, \ldots, v_{n(\beta)})$ are distinct unless $\alpha = \beta$ and $u_k = v_k$ for all $k = 1, \ldots, n(\alpha)$).

It is sometimes convenient to restrict attention to the collection of words of rank no greater than some fixed integer ρ. So we also define $W_\rho = [\Sigma, F]_\rho$ as the subset of the free word algebra $W = [\Sigma, F]$ comprising words of rank $\leq \rho$.

For example, suppose that F comprises the operations $\{\cap, ', \circ\}$ and that $\Sigma = \{x, y\}$. Then the elements of rank 0 in the word algebra $W = [\Sigma, F]$ are $\{x, y\}$; the elements of rank 1 are $\{x \cap x, x \cap y, y \cap x, y \cap y, x', y', x^\circ x, x^\circ y, y^\circ x, y^\circ y\}$; elements of rank 3 include $x \cap x \cap x, x \cap (x^\circ x), (x^\circ x) \cap x, x \cap x', x' \cap x$, etc; and so on. In this case, the free word algebra $W = [\Sigma, F]$ comprises all composite relational terms that can be constructed from the primitive terms x and y using the intersection, converse and concatenation operations.

The free word algebra simply provides a language that permits certain forms of discrimination among types of tie to be encoded. It makes no structural assumptions whatever, although we might anticipate from the intended interpretation of the terms in F that certain terms in W will not need to be discriminated. For instance, we might suppose that: the intersection operation can be assumed to be *idempotent* ($x \cap x = x$, for all $x \in W$), *commutative* ($x \cap y = y \cap x$, for all $x, y \in W$) and *associative* ($(x \cap y) \cap z = x \cap (y \cap z)$, for all $x, y, z \in W$); that the converse operation satisfies $x'' = x$, for all $x \in W$; and that the concatenation operation is also associative. Rather than impose such postulates at this abstract level, though, we move directly to the evaluation of the terms in W for an empirical realization of a (multiple) network.

Algebras for Multiple Networks

The language provided by the free algebra $W = [\Sigma, F]$ can be used to determine all possible discriminations among ties that our structural claims suggest as appropriate. To describe the structural forms to which observed network ties give rise, each of the derived relations for an observed multiple network can be computed. There is a straightforward correspondence between the words u in $W = [\Sigma, F]$ expressed in terms of letters in Σ and operations in F, on the one hand, and Boolean matrix manipulations of the binary constituents in the multiple network, on the other. The correspondence is achieved by replacing:

- the relational term x_m by the array X_m;
- the intersection operation $\cap$ by Boolean intersection of arrays (i. e., element-wise Boolean multiplication, denoted $\bullet$); and
- the composition operation $\circ$ by Boolean matrix multiplication.

Boolean intersection, transposition and composition are defined for $n \times n$ binary arrays Y_1 and Y_2 by:

- $Y_1 \bullet Y_2(k, l) = Y_1(k, l) Y_2(k, l)$;
- $Y_1'(k, l) = Y_1(l, k)$; and
- $Y_1 Y_2(k, l) = Y_1(k, 1) Y_2(1, l) + Y_1(k, 2) Y_2(2, l) + \cdots + Y_1(k, n) Y_2(n, l)$.

The term $u \in W$ then corresponds to an $n \times n$ Boolean array, denoted X_u, with $X_u(k, l) = 1$ if and only if there is a relational tie described by the word u from k to l. We also define a partial order on the set of relations $\{X_u : u \in W\}$ by:

$$X_u \leq X_v \quad \text{if and only if } X_u(k, l) \leq X_v(k, l) \quad \text{for all } k, l \in N .$$

Suppose that we evaluate X_u for all $u \in W$. We define a generalized version of the *Axiom of Quality* by setting $u \leq v$ in W whenever $X_u \leq X_v$ [5].

Defining an induced partition θ on W by $(u, v) \in \theta$ if and only if $u \leq v$ and $v \leq u$ and letting $S = W/\theta$ be the set of associated classes with the induced partial order:

$$s \leq t \text{ in } S \quad \text{if } u \in s\,,\ v \in t \text{ and } u \leq v \text{ in } W\,,$$

we can then define a general (partially ordered) algebraic structure on S.

A (*partially ordered*) *algebra* is a triple $[S, F, \leq]$, where S is a nonempty set of elements (usually assumed to be finite), F is a specified set of operations, f_α, each mapping a power $S^{n(\alpha)}$ of S into S, for some non-negative finite integer $n(\alpha)$, and $\leq$ is a partial order on S. Each operation f_α is assumed to be *isotone* in each of its variables: that is, if $x_i \leq y_i$ ($x_i, y_i \in S$; $i = 1, 2, \ldots, n(\alpha)$), then $f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)}) \leq f_\alpha(y_1, y_2, \ldots, y_{n(\alpha)})$. A *family of algebras* is a collection of algebras each having the

same set F of operations and satisfying a specified set of postulates. Two algebras belonging to the same family are termed *similar*.

For $F = \{\circ\}$, the algebra $[S, F, \leq]$ is a *partially ordered semigroup* [36].

The Semigroup of a Multiple Network

For multiple networks, a composite relation constructed as the concatenation of observed relations records the presence of a particular form of *labelled walk*. Specifically, actor k is connected to actor l by a walk with label u if and only if $X_u(k, l) = 1$. If $X_u \leq X_v$, two actors who are joined by a walk with label u are also necessarily joined by a walk with label v.

An algorithm for constructing the semigroup of a multiple network $[S, \{\circ\}, \leq]$ is as follows:

(S1) Let $W_1 = \{X_1, X_2, \ldots, X_r\}$ and set $i = 1$.
(S2) Construct the networks with edge sets WZ for each $W \in W_i$ and each $Z \in W_1$. Place WZ in W_{i+1} if $WZ \neq Y$, for any $Y \in W_j$, $j = 1, 2, \ldots, i+1$; otherwise, $WZ = Y$ for some $Y \in W_j$ $(j = 1, 2, \ldots, i+1)$ is an *equation* in the semigroup S.
(S3) If $W_{i+1} = \emptyset$, stop and compute the partial order among elements in $W_1 \cup W_2 \cup \ldots$; otherwise, set i to i+1 and return to step S2.

The equations $WZ = Y$ may be recorded in a *right multiplication table* in which Y appears as the table entry corresponding to the row labelled W and the column labelled Z. The partial ordering on S is constructed from the collection of networks in $W_1, W_2, \ldots$.

Lattices from Networks

Lattice representations for networks arise in two important ways. In the first, elements of the lattices are elements of $S = W/\theta$ for $F \supseteq \{\cap, \circ\}$. The meet of two elements is their intersection, and their join is the least relation in S which is greater than or equal to both of them. Indeed, the algebra $[S, F, \leq]$ is both a lattice and a partially ordered semigroup in this case.

In the second type of lattice construction, the elements correspond to rows or columns of an affiliation network. Since the affiliation array provides an explicit representation of the co-constitution of groups by actors and actors by groups, the lattice structures to which it gives rise can be used to analyze the duality of actors and groups [10,15,16,31,33,38,47].

A lattice denoted $L(X)$ is generated by the rows of a matrix X under the intersection operation; it is termed the *Galois* or *concept lattice* of X [20]. In $L(X)$, the meet of two elements is equal to their intersection, while their join is the minimal element in the lattice which is greater than or equal to both of them. A lattice can also be constructed from the rows of the matrix X under the union operation. In this lattice, termed the *Zareckii lattice* by Boyd, the join of two vectors is equal to their union, and their meet is the maximal vector which is less than or equal to both of them [8].

To construct the Galois lattice, one simply needs to compute all distinct intersections among rows of X, add the vector $[1\ 1 \ldots 1]$ and compute the partial order among the resulting set of row vectors.

Role Algebras from Networks

Role algebras have been proposed as local network analogues of the semigroup representation of relational structure in an entire social network. The representation has been presented in several slightly different forms but each is derived from an original formulation proposed by Mandel [11,29,53,54].

Denote by $k * X_u$ the vector $[X_u(k, 1), X_u(k, 2), \ldots, X_u(k, n)]$ for $u \in W$. The vector indicates whether there is a relationship of type u from actor k to each other actor in the network. As before, we can envisage evaluating $k * X_u$ for all $u \in W$. We then define a k-centered version of the *Axiom of Quality* by setting

$$u \leq v \text{ in } W \quad \text{whenever} \quad k * X_u \leq k * X_v$$

and an induced partition θ on W by $(u, v) \in \pi$ if and only if $u \leq v$ and $v \leq u$. Also as before, we let $S = W/\theta$ denote the set of associated classes with the induced partial order:

$$s \leq t \text{ in } S \quad \text{if } u \in s\,,\ v \in t \ \text{ and } \ u \leq v \text{ in } W\,.$$

The algebra $[S, F, \leq]$ is termed a *role algebra*.

The local role algebra for actor k may be constructed as follows.

(RA1) Let $W_1 = \{X_1, X_2, \ldots, X_r\}$ and set $i = 1$.
(RA2) Construct the vectors $k * WZ$ for each $W \in W_i$ and each $Z \in W_1$. Place WZ in W_{i+1} if $k * WZ \neq k * Y$, for any $Y \in W_j$, $j = 1, 2, \ldots, i+1$; otherwise, $k * WZ = k * Y$ for some $Y \in W_j$ $(j = 1, 2, \ldots, i+1)$ and $WZ = Y$ is an *equation* in the role algebra.
(RA3) If $W_{i+1} = \emptyset$, stop and compute the partial order among elements in $W_1 \cup W_2 \cup \ldots$; otherwise, set i to i+1 and return to step RA2.

Partial Algebras

To achieve algebraic closure, it may be necessary to compute relations corresponding to words of high rank in W and, in the case of some representations at least, it can be argued that some restriction on these repeated operations should be imposed. Mandel, for instance, made such an argument for the composition operation in social networks; he claimed that short network paths are likely to be more salient to the members of the network than longer ones [29]. These considerations can be formalized with the introduction of partial algebraic structures, whose elements are subject to rank restrictions on some or all of the operations used to construct the algebra.

Let $W_\rho = [\Sigma, F]_\rho$ be the subset of the free word algebra $W = [\Sigma, F]$ comprising words of rank $\leq \rho$. Define a partial order on the elements of W_ρ by the generalized Axiom of Quality:

$$u \leq v \text{ in } W_\rho \quad \text{whenever} \quad X_u \leq X_v \,.$$

The relation $\leq$ is clearly both *reflexive* ($u \leq u$ for all u) and *transitive* ($u \leq v$ and $v \leq w$ implies $u \leq w$ for any u, v, w), that is, a *quasi-order*. Further, each operation f_α is *isotone* in each of its variables whenever the result of the operation is contained in W_ρ; that is, if $x_i \leq y_i$ ($x_i, y_i \in W_\rho$; $i = 1, 2, \ldots, n(\alpha)$), then $f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)}) \leq f_\alpha(y_1, y_2, \ldots, y_{n(\alpha)})$, provided that both $f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)}), f_\alpha(y_1, y_2, \ldots, y_{n(\alpha)}) \in W_\rho$. Thus it follows that the partial ordering $\leq$ on W_ρ gives rise to a *partial* algebra in the following sense.

A *partial algebra* is a triple $[S, F, \leq]$, where S is a non-empty set of elements (usually assumed to be finite), F is a specified set of *partial* operations, f_α, each mapping some subset $T^{(\alpha)}$ of $S^{n(\alpha)}$ into S, for some non-negative finite integer $n(\alpha)$, and $\leq$ is a partial order on S. Each partial operation f_α is assumed to be *isotone* in each of its variables: that is, if $x_i \leq y_i$ ($x_i, y_i \in S$; $i = 1, 2, \ldots, n(\alpha)$), then $f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)}) \leq f_\alpha(y_1, y_2, \ldots, y_{n(\alpha)})$ provided that both $(x_1, x_2, \ldots, x_{n(\alpha)}), (y_1, y_2, \ldots, y_{n(\alpha)}) \in T^{(\alpha)}$. A *family of partial algebras* is a collection of algebras each having the same set F of partial operations defined on the same subsets $T^{(\alpha)}$ of the power sets $S^{n(\alpha)}$.

Partial semigroup algebras may be constructed from labelled walks in a multiple network of some maximum length ρ [40]. For role algebras, the partial role algebra for a node k derived from W_ρ is the role algebra associated with the ρ-local neighborhood of k [39].

Algebraic Structure

In some cases, it is possible to analyze the structural properties of algebraic representations of networks.

The Semigroup of a (Single) Network Under Composition

We first consider the semigroup for a single network array X. We say that there is a *walk* from actor k to actor l if $X(k, l) = 1$. There is a walk of *length* p from k to l if there is a sequence of actors $k = k_0, k_1, \ldots, k_p = l$ such that $X(k_{i-1}, k_i) = 1$, for $i = 1, 2, \ldots, p$. It is readily seen that $X^p(k, l) = 1$ if and only if there is a walk of length p from k to l. The sequence $X, X^2, \ldots$ generated by the composition operation is therefore a sequence recording the presence of walks of different lengths between all pairs of network actors. The semigroup structure of this sequence therefore describes relations of equality and ordering among walks in the network.

The sequence can be described with the help of the following definition. Let p be the least integer such that

$$X^p = X^q \quad \text{for some} \quad q > p \,.$$

Also let $q = p + d$ ($d \geq 1$) be the least integer satisfying this relation. The integers p and d are termed the *index* and *period* of the network, respectively. It can readily be seen that the sequence has the form

$$X, \ldots, X^{p-1}, X^p, \ldots, X^{p+d-1}, X^p, \ldots, X^{p+d-1}, \ldots \,.$$

The set $\{X, X^2, \ldots, X^{p+d-1}\}$ defines a semigroup S of which $G = \{X^p, \ldots, X^{p+d-1}\}$ is a group (defined below).

It is natural to ask what is the relationship between properties of a network and its index and period.

For a strongly connected network X (that is, a network in which there is a walk of some length from each actor to each other actor), the period d is the greatest common denominator of all integers m such that m is the length of a cycle in X (that is, a walk with the same initial and final node). If X has more than one strong component, its period is the least common multiple of the periods of those strong components [24].

A network X with index 1 and period 1 is termed *idempotent*; for a network of index p and period 1, the network X^p is idempotent. It may be shown that every idempotent network is a pseudo-order, defined as follows [43]. Let Z be a quasi-order on N and let $e_Z = Z \cap Z'$. An actor $k \in N$ is termed *Z-strict* if there is no distinct actor $l \neq k$ such that $Z(k, l) = 1$. The subset H of N is *Z-permissible* if each of its actors is Z-strict and there are no pairs $k, l \in N$ such that k covers l or l covers k (recall that k covers l if $Z(k, j) = 1$ and $Z(j, l) = 1$ implies $j = k$ or $j = l$). Define the relation Z_H by

$$Z_H(k, l) = 1 \quad \text{iff} \quad k = l \text{ and } k \in H \,.$$

Then a network X is a *pseudo-order* if

$$X = Z \backslash Z_H$$

for some quasi-order Z and Z-permissible subset H of N.

Structure of Multiple Network Semigroups

In some cases, networks define semigroups with certain structural features.

An element s in a semigroup S possesses a (generalized) *inverse* if there exists some $t \in S$ such that $sts = s$ and $tst = t$. The element t is termed the inverse of s. If every $s \in S$ possesses an inverse, then S is termed an *inverse semigroup*. If, in addition, S possesses an *identity* element, that is, an element e such that for any $s \in S$, there exists some $t \in S$ such that $ts = e$, then S is a *group*. The element t is termed a *(group-)inverse* of s. An element s in a semigroup S is *regular* if there exists some $t \in S$ such that $sts = s$. The semigroup S is *regular* if each of its elements is regular.

It is clear from these definitions that any group is also an inverse semigroup and that any inverse semigroup is also regular. The results below describe some of the conditions under which mutiple networks give rise to regular semigroups, inverse semigroups and groups.

We first describe the conditions under which S is a group [28,42]. Let X be a network relation having a group-inverse Y, that is, a relation Y such that $XY' = I$ where I is the identity relation (for which $I(k, l) = 1$ if $k = l$; $I(k, l) = 0$ otherwise). Then it follows that (i) $XX' = I = X'X$ and $Y = X$ (that is, X' is a two-sided group-inverse for X and is unique); and (ii) X is a permutation relation (that is a relation in which, for each node k, there is exactly one node l for which $X(k, l) = 1$ and exactly one node j for which $X(j, k) = 1$). Hence if each relation X_i of a multiple network is a permutation relation and X_i' is included in S for each i, then the semigroup S is a group.

An important class of social relations which are constructed as permutation relations are "marriage class systems". These systems were proposed by White as models for kinship relations in certain Australian aboriginal tribes [48]. In such systems, persons are uniquely and permanently assigned to a *clan* and clans are related to one another by marriage and descent rules which can be represented as permutation relations. The marriage relation specifies the clan containing the wives of the male members of each clan, while the descent relation identifies the clan containing the children of male members of each clan. It follows that marriage class systems give rise to groups that represent the kinship structures of the societies concerned.

Another structure arising in the description of kinship systems is the inverse semigroup [9]. One can verify whether a network semigroup is an inverse semigroup by determining whether each of its elements has an inverse in S. The following is a characterization of conditions under which a relation possesses an inverse [24].

A network relation X has a *Thierrin–Vagner inverse* if there exists a relation Y satisfying $XYX = X$ and $YXY = Y$. Clearly, X has a Thierrin–Vagner inverse if and only if X possesses an inverse in S. If X is a regular element of S, so that $XYX = X$ for some Y in S, then YXY is a Thierrin–Vagner inverse for X. Thus determining whether X is regular will reveal whether X has an inverse. The regularity of X can be ascertained as follows.

The network relation Y is a *subinverse* of X if $XYX \leq X$. The set of subinverses of X is closed under union (that is, if Y and Z are subinverses of X then so is the relation $Y \cup Z$); hence there is a largest subinverse X^* for any relation X. The largest subinverse X^* of a relation X can be calculated from X according to $X^* = (X^T X^c X^T)^c$ where X^c denotes the complement of X. It therefore follows that a network X is regular if and only if $X \leq X(X^T X^c X^T)^c X$, and that the largest Thierrin–Vagner inverse of a regular relation X is $X^* X X^*$ [24,44].

The Algebraic Structure of Local Role Algebras

Less is known of the general algebraic properties of local role algebras. It is easily seen, however, that any orderings among relations present in the semigroup of a multiple network are also present in the local role algebra of each actor in the network. As a result, a number of the algebraic properties of local role algebras are inherited from the "parent" network semigroup. For example, if $s \leq t$ in the partially ordered semigroup S constructed from the composition operation, then $s \leq t$ in the local role algebra of every node in the network. Indeed, the converse is readily seen to also hold, so that $s \leq t$ in S if and only if $s \leq t$ in the local role algebra of every node in the network.

Algebraic Analysis

In order to describe an algebraic representation, general procedures for analyzing finite algebraic representations into simpler, independent components have been developed [20,36,37,39]. These procedures have been applied to the decomposition of lattice representations [20], partially ordered semigroups and local role algebras [34,35,36]. The procedures extract maximally independent components

of a given algebraic representation and lead to a detailed structural analysis.

The account of these procedures for decomposing a finite algebra into simpler components is based on definitions adapted from [2,18]. Clearly, any decomposition procedure depends on the synthesis rules by which the components are assumed to be combined to produce the algebra. Two important synthesis rules are the direct and subdirect product.

Let $A_1 = [S_1, F, \leq]$, $A_2 = [S_2, F, \leq]$, $\ldots, A_q = [S_q, F, \leq]$ be a collection of similar algebras. The *direct product* $A_1 \times A_2 \times \cdots \times A_q$ of $A_1, A_2, \ldots, A_q$ is the algebra comprising the set $S_1 \times S_2 \times \cdots \times S_q$ and the operations $f_\alpha \in F$ given by

$$f_\alpha([x_1, x_2, \ldots, x_q], \ldots, [z_1, z_2, \ldots, z_q]) = [f_\alpha(x_1, \ldots, z_1), \ldots, f_\alpha(x_q, \ldots, z_q)] .$$

The partial order for $A_1 \times A_2 \times \cdots \times A_q$ is given by

$$[x_1, x_2, \ldots, x_q] \leq [z_1, z_2, \ldots, z_q]$$

if and only if

$$x_1 \leq z_1, x_2 \leq z_2, \ldots, \text{ and } x_q \leq z_q .$$

An algebra $B = [T, F, \leq]$ is a *subalgebra* of $A = [S, F, \leq]$ if T is a subset of S (possibly empty) which is closed under the operations of F (that is, $f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)}) \in T$, for any $x_1, x_2, \ldots, x_{n(\alpha)} \in T$, and any $f_\alpha \in F$); in addition, $x \leq y$ in B if and only if $x \leq y$ in A, for any $x, y \in T$. A subalgebra $C = [S, F, \leq]$ of a direct product $A_1 \times A_2 \times \cdots \times A_q$ of similar algebras $A_i = [S_i, F, \leq]$ $(i = 1, 2, \ldots, q)$ is a *subdirect product* of $A_1, A_2, \ldots, A_q$ if for each $x_i \in S_i$, there exists an element $c \in S$ having x_i as its component in A_i.

In the case of both the direct and subdirect product of algebras, the operations in the composite algebra are performed as the conjunction of their independent operation in the component algebras. The difference between the two constructions lies in the set on which these operations are defined. For the direct product, the appropriate set is the full Cartesian product of elements from each component algebra; in the subdirect product case, the appropriate set is simply a subset of the full Cartesian product with the additional condition that each element in each component algebra appears in *some* element of the subset.

Some well-known theorems in universal algebra establish that the existence and nature of direct and subdirect representations of an algebra are determined by the lattice of homomorphisms of the algebra, or, equivalently, by a lattice of relations on the algebra associated with its homomorphisms.

An (*isotone*) *homomorphism* from an algebra $A = [S, F, \leq]$ onto a similar algebra $B = [T, F, \leq]$ is a mapping $\phi : S \to T$ such that,

(i) for all $f_\alpha \in F$ and $x_i \in S$,

$$\phi(f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)})) = f_\alpha(\phi(x_1), \phi(x_2), \ldots, \phi(x_{n(\alpha)})) ; \quad \text{and}$$

(ii) $x \leq y$ in S implies $\phi(x) \leq \phi(y)$ in T.

The algebra B is termed a (*homomorphic*) *image* of A, and we write $B = \phi(A)$.

Further, each homomorphism ϕ from $A = [S, F, \leq]$ onto $B = [T, F, \leq]$ has a corresponding binary relation π on S (termed here a π*-relation*) in which $(x, y) \in \pi$ if and only if $\phi(y) \leq \phi(x)$; it may readily be established that π is transitive and reflexive (and hence a quasi-order) and that the equivalence relation σ_π defined by $(x, y) \in \sigma_\pi$ if and only if $(x, y) \in \pi$ and $(y, x) \in \pi$ has the substitution property, namely:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_{n(\alpha)}, y_{n(\alpha)}) \in \sigma_\pi$$

implies

$$(f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)}),\ f_\alpha(y_1, y_2, \ldots, y_{n(\alpha)})) \in \sigma_\pi .$$

The relation σ_π is termed a *congruence relation*. The collection of π-relations may be partially ordered by: $\pi_1 \leq \pi_2$ if $(x, y) \in \pi_1$ implies $(x, y) \in \pi_2$, for any $x, y \in S$. Under this partial ordering, the relations π form a lattice $L_\pi(A)$, the *lattice of* π*-relations* on A.

The conditions under which an algebra A may be represented as a direct or subdirect product of similar algebras $A_1, A_2, \ldots, A_r$ (that is, possesses *direct* or *subdirect representations*) can be described as follows.

Let $X = \{\pi_1, \pi_2, \ldots, \pi_q\}$ be a set of π relations in $L_\pi(A)$ for which

$$\pi_1 \wedge \pi_2 \wedge \cdots \wedge \pi_q = \pi_{\min},$$

where $\pi_{\min}$ is the minimal element of $L_\pi(A)$. Then A may be represented as a subdirect product of the algebras $\phi_1(A), \phi_2(A), \ldots, \phi_q(A)$, where $\phi_i(A)$ is the image of A under the homomorphism ϕ_i corresponding to the relation π_i $(i = 1, 2, \ldots, q)$. If, in addition, for all i,

(a) $\pi_1 \wedge \pi_2 \wedge \cdots \wedge \pi_{i-1}$ and π_i are *permutable* (i. e., $(\pi_1 \wedge \pi_2 \wedge \cdots \wedge \pi_{i-1})\pi_i = \pi_i(\pi_1 \wedge \pi_2 \wedge \cdots \wedge \pi_{i-1})$); and

(b) $(\pi_1 \wedge \pi_2 \wedge \cdots \wedge \pi_{i-1}) \vee \pi_i = \pi_{\max}$, where $\pi_{\max}$ is the maximal element of $L_\pi(A)$, then A may be represented as the direct product of the algebras $\phi_1(A), \phi_2(A), \ldots, \phi_q(A)$ [2].

An algebra which possesses no nontrivial representation as the direct product of similar algebras is termed *directly irreducible*, while an algebra which cannot be expressed as a nontrivial subdirect product of similar algebras is *subdirectly irreducible*. Pattison and Bartlett advocated the use of subdirect representations, but sought to restrict attention to a small set of such representations rather than to the (possibly large) class of all subdirect representations of the algebra [37]; see also [7]. Their restriction entailed (a) identifying all *irredundant* subdirect representations of an algebra, and then (b) defining a partial ordering on the irredundant representations and selecting only minimal members of the resulting partially ordered set. The resulting set of minimal, irredundant subdirect representations were termed the factorizations of the algebra.

More formally, an element x of a lattice L is *meet-irreducible* if $x = x_1 \wedge x_2$ implies $x = x_1$ or $x = x_2$. A subset of lattice elements $X = \{x_1, x_2, \ldots, x_r\}$ is *irredundant* if

a) each x_i is meet-irreducible; and
b) $x_1 \wedge \cdots \wedge x_{i-1} \wedge x_{i+1} \wedge \cdots \wedge x_q, \neq x_{\min}$, for each $i = 1, 2, \ldots, q$; where $x_{\min}$ is the minimal element of the lattice L.

The collection of all irredundant subsets of a lattice L may be partially ordered by the relation:

$X \leq Y$ iff, for each $j = 1, 2, \ldots, p$, there exists some i $(i = 1, 2, \ldots, q)$ such that $y_j \leq x_i$ in L, where $X = \{x_1, x_2, \ldots, x_q\}$ and $Y = \{y_1, y_2, \ldots, y_p\}$.

A *factorization* of an algebra A is then the subdirect representation corresponding to any minimal, irredundant subset of elements of $L_\pi(A)$ whose meet is the minimal relation $\pi_{\min}$.

An Algorithm for Factorization

A general algorithm for constructing the set of all factorizations of an algebra has also been developed [37]. The algorithm operates on a subset of relations in the lattice $L_\pi(A)$ and conducts a reasonably efficient search for subsets of π relations which satisfy the conditions of the factorization definition. In particular, the algorithm constructs the set of factorizations of an algebra A from the atoms of its lattice $L_\pi(A)$ of π-relations and their maximal meet-complements.

An *atom* of a lattice L is an element that covers $x_{\min}$. A *meet-complement* of an element $x \in L$ is an element x^* such that $x^* > x_{\min}$ and $x \wedge x^* = x_{\min}$. A meet-complement x^* of x is *maximal* if x has no other meet-complement z such that $z > x^*$.

Let $Z = \{z_1, z_2, \ldots, z_a\}$ be the set of atoms of $L_\pi(A)$. Define the collection of sets of the form $\{z_1^*, z_2^*, \ldots, z_a^*\}$ where z_i^* is a maximal meet-complement of z_i. Then it can be shown that any factorization of A is a subset of such a set [36]. This result is the basis for the algorithm for finding factorizations. Under certain circumstances, we can write down an explicit expression for π-relations in the factorization of A.

Let A be an algebra with π-relation lattice $L_\pi(A)$. Let π_{ab} be the least π-relation in which $a \geq b$ for elements $a, b \in A$; we term π_{ab} the π-relation generated by the ordering $a \geq b$.

Let z be an atom of $L_\pi(A)$; define the relation $\pi(z) = \{(a, b) : \pi_{ab} \wedge z = \pi_{\min}\}$. If z has a unique maximal meet-complement, then $\pi(z)$ is a π-relation and is a unique maximal meet-complement of z. Conversely, if $\pi(z)$ is a π-relation, then it is the unique maximal meet-complement of z. Further, if $Z = \{z_1, z_2, \ldots, z_a\}$ is the set of atoms of $L_\pi(A)$ and if, for each $z_i \in Z$, z_i has a unique maximal meet-complement $\pi(z_i)$, then the factorization of A is unique and corresponds to the π-relations $\{\pi(z_1), \pi(z_2), \ldots, \pi(z_a)\}$ [36]. In other words, if the factorization is unique, it can be identified immediately from the maximal meet-complements of the atoms of $L_\pi(A)$.

Under some circumstances, we can make more general claims about the uniqueness of factorizations. Not surprisingly, the structure of the lattice $L_\pi(A)$ plays a major part. It is known, for instance, that if the lattice $L_\pi(A)$ is distributive, then A possesses a unique irredundant subdirect representation, and hence a unique factorization. In particular, since the lattice of π-relations for a lattice is necessarily distributive, any finite lattice has a unique factorization. If the lattice $L_\pi(A)$ is modular, then any irredundant subdirect representation of A has the same number of components [2].

Factorization for Partial Algebras

Analogous decompositions can be described for partial algebras. An (*isotone*) *homomorphism* from a partial algebra $A = [S, F, \leq]$ onto a similar partial algebra $B = [T, F, \leq]$ is a mapping $\phi\colon S \to T$ such that,

(i) for all $f_\alpha \in F$ and $x_i \in S$,

$$\phi(f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)})) = f_\alpha(\phi(x_1), \phi(x_2), \ldots, \phi(x_{n(\alpha)}))$$

whenever the operations on both sides of the expression are defined; and

(ii) $x \leq y$ in S implies $\phi(x) \leq \phi(y)$ in T.

The algebra B is termed a (*homomorphic*) *image* of A, and we write $B = \phi(A)$.

Further, each homomorphism ϕ from $A = [S, F, \leq]$ onto $B = [T, F, \leq]$ has a corresponding binary relation π on S (termed here a π*-relation*) in which $(x, y) \in \pi$ if and only if $\phi(y) \leq \phi(x)$; it may readily be established that π is transitive and reflexive (so that π is a quasi-order). Further, the equivalence relation σ_π defined by $(x, y) \in \sigma_\pi$ if and only if $(x, y) \in \pi$ and $(y, x) \in \pi$ has the substitution property, namely:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_{n(\alpha)}, y_{n(\alpha)}) \in \sigma_\pi$$

implies

$$(f_\alpha(x_1, x_2, \ldots, x_{n(\alpha)}),\ f_\alpha(y_1, y_2, \ldots, y_{n(\alpha)})) \in \sigma_\pi$$

whenever the latter term is defined. The relation σ_π is often termed a *congruence relation*. The collection of all π-relations form a lattice $L_\pi(A)$, the *lattice of* π*-relations* on A, under the partial ordering: $\pi_1 \leq \pi_2$ if $(x, y) \in \pi_1$ implies $(x, y) \in \pi_2$, for any $x, y \in S$.

In the case of partial algebras, it is convenient to express the algebra as an intersection of the π-relations corresponding to maximally independent homomorphic images of the algebra. Recall that the minimal element $\pi_{\min}$ of the lattice $L_\pi(A)$ corresponds to the algebra A itself. Thus, any expression of $\pi_{\min}$ in the form

$$\pi_1 \wedge \pi_2 \wedge \cdots \wedge \pi_q = \pi_{\min}$$

expresses A as the intersection of π-relations corresponding to homomorphic images of A. Any such expression that involves π-relations $\{\pi_i\,;\, i = 1, 2, \ldots, q\}$ satisfying the conditions of the factorization definition yields an expression in terms of a set of simple and maximally independent images of the algebra. Further, the same algorithm for constructing factorizations of an algebra that was described earlier may be used to find expressions of this type.

Algebraic and Network Mappings

The semigroup of a multiple network records the orderings and equations among different types of labelled paths in the network. A natural question to arise is how properties of a multiple network constrain the relational structure of its semigroup. We therefore review some general conditions under which the semigroup of one network is a homomorphic image of the semigroup of another.

We define two multiple networks to possess the same structure if their semigroups are isomorphic, that is, if the network relations are in one-to-one correspondence and they have identical right multiplication tables and partial orders.

One strategy for investigating whether networks have similar structures is to determine the conditions under which network homomorphisms induce homomorphisms of the network semigroup [4]. A *network homomorphism* from a multiple network $\{X_1, X_2, \ldots, X_r\}$ on actor set N to a multiple network $\{Y_1, Y_2, \ldots, Y_r\}$ on actor set M is a mapping ψ from N onto M such that: (a) $X_m(k, l) = 1$ implies $Y_m(\psi(k), \psi(l)) = 1$, for any k, l, m; and (b) $Y_m(i, j) = 1$ for some i, j implies that $X_m(k, l) = 1$, for some k, l such that $\psi(k) = i$ and $\psi(l) = j$. The network on M is termed the *image* of the network on N under the mapping ψ.

The mapping ψ satisfies the *structural equivalence* condition if for any m, and for any $k, l \in N$, $\psi(k) = \psi(l)$ if and only if:

- $X_m(k, j) = 1$ iff $X_m(l, j) = 1$ for any $j \in N$; and
- $X_m(j, k) = 1$ iff $X_m(j, l) = 1$ for any $j \in N$.

Lorrain and White observed that if two multiple networks are related by a network homomorphism satisfying the structural equivalence condition, then their semigroups are isomorphic and we may argue that they possess the *same* relational structure [27]. More generally, we can ask whether a homomorphism between two networks induces a homomorphism between their semigroups. It is readily seen that a homomorphism is not always guaranteed.

Network homomorphisms induced by certain blockmodels lead to semigroup homomorphisms (see the entry on ► Positional Analysis and Blockmodeling). More general conditions under which a semigroup homomorhism is guaranteed were established by Kim and Roush [25].

A network homomorphism ψ satisfies Kim and Roush's *condition* G_i if the following holds for any pair of equivalence classes ρ_1 and ρ_2 on N induced by the mapping ψ (so that $k, l \in \rho_h$ for some h iff $\psi(k) = \psi(l)$). Let the number of elements in ρ_1 and ρ_2 be n_1 and n_2, respectively. Let D be any subset of ρ_1 of i elements or, if $i > n_1$, let $D = \rho_1$. Then, for any $m = 1, 2, \ldots, r$, the set $\{l : l \in \rho_2 \text{ and } X_m(k, l) = 1 \text{ for some } k \in D\}$ has at least $\min(i, n_2)$ elements. The condition G_1 is also known as the *outdegree condition* and G_n is also termed the *indegree condition*. A network homomorphism that satisfies both G_1 and G_n is termed *regular*.

Kim and Roush demonstrated that if ψ is a network homomorphism from one multiple network onto another that satisfies the condition G_i, then there is a homomor-

phism mapping the semigroup of the first onto the semigroup of the second.

A more general condition combines the condition G_i with what Pattison termed the *central representatives condition* [34]. Let ψ be a network homomorphism. Then ψ satisfies the condition G_{im} if, for each class ρ of elements of N induced by ψ,

- there exists a *central subset* C of ρ such that for any X_m
  - $X_m(k, l) = 1$ for some $k \in \rho$ implies $X_m(k^*, l) = 1$ for some $k^* \in C$, and
  - $X_m(l, k) = 1$ for some $k \in \rho$ implies $X_m(l, k^*) = 1$ for some $k^* \in C$; and
- if C^* denotes the union of central subsets C, then the central subsets C satisfy Kim and Roush's condition G_i on the network defined on C^*.

If each central subset C comprises a single actor, then the condition is equivalent to Pattison's central representatives condition, while if each central subset C comprises the whole of the equivalence class on N induced by ψ, it is equivalent to the condition G_i. Kim and Roush showed that if one network can be mapped onto another by a network homomorphism satisfying the condition G_{im}, then there is a homomorphism from the semigroup of the first to the semigroup of the second [25]. The condition G_{im} is the most general condition known that guarantees the existence of such a homomorphism.

Comparing Network Semigroups

Two network semigroups S_1 and S_2 are strictly comparable only if $S_1 \leq S_2$ or $S_2 \leq S_1$. Loosely speaking, though, they are "similar" if they share many homomorphic images. Boorman and White proposed that the largest shared homomorphic image (the so-called *joint homomorphism*) is a useful construction for comparing network semigroups [5]. Bonacich and McConaghy, on the other hand, argued that the smallest semigroup containing each of S_1 and S_2 (the *common structure semigroup*) was a more appropriate representative of common semigroup structure [3,32]. The resolution of the problem of finding a representative of common structure depends on how a semigroup is viewed [36]. If a semigroup is seen as a list of the homomorphic images that it admits, then the joint homomorphism corresponds to the list of shared features, while if it seen as a collection of semigroup equations and orderings, then the common structure semigroup serves as a representation of shared structure. The common structure semigroup of two network semigroups is the semigroup of the disjoint union of the underlying networks.

Both the joint homomorphic image and common structure semigroup constructions have been shown to be useful in particular applications, the former in identifying common reductions of the networks giving rise to identical semigroups, and the latter in identifying shared cultural forms (such as "the friend of a friend is always an associate").

Linking Algebraic and Network Homomorphisms

The factorization of an algebra identifies relatively independent components of the algebra generated by constituents of the data array under the selected operation. It is natural to ask whether this analysis induces a corresponding decomposition of the network itself. That is, if an algebra A has a homomorphic image B (such as one of the factors of A), is it possible to find one or more homorphisms of the network that is consistent with the congruence relation σ for the homomorphism and that generates the algebra B? If so, we can argue that there is an association between such a reduction of the network and the homomorphic image.

In the case of the lattice of an affiliation network, such a reduction can always be uniquely found [20]. For semigroups and role algebras, there is no guarantee that such a reduction can be found, but it is possible to identify the smallest homomorphic image of the network whose algebra contained B as a homomorphic image [34].

Examples

We illustrate a number of the algebraic constructions just described for a multiple network blockmodel and an approximation to an affiliation network.

Network Semigroup

Table 1 corresponds to a multiple network on 7 nodes with two kinds of edges, F and N. The network is a blockmodel reported by Vickers for relations of friendship and negative ties among members of a high school class in an Australian country town [46]. The questions from which the blockmodel was constructed were (a) "Who are your best friends?" and (b) "Who would you rather not have as a friend?".

The semigroup generated by the multiple network in Table 1 has the right multiplication table and partial order shown in Table 2. The process of constructing the semigroup generates the sets $W_1 = \{F, N\}$, $W_2 = \{F^2, FN, NF, N^2\}$, $W_3 = \{F^3, FNF, FN^2, NF^2, N^2F, N^3\}$, $W_4 =$

Social Networks, Algebraic Models for, Table 1
A multiple network on 7 nodes

		Block									Block						
Relation	Block	1	2	3	4	5	6	7	Relation	Block	1	2	3	4	5	6	7
F	1	1	1	1	0	0	0	0	*N*	1	0	0	0	1	1	1	1
	2	0	1	0	0	0	0	1		2	0	0	0	1	1	1	0
	3	1	1	1	0	0	0	0		3	0	0	0	1	1	1	0
	4	0	0	0	1	0	0	0		4	0	1	1	0	0	1	1
	5	0	0	0	0	1	0	0		5	0	0	0	0	0	0	1
	6	1	0	0	0	0	1	0		6	0	0	0	1	0	0	1
	7	1	1	1	0	0	0	1		7	0	0	0	1	1	1	1

Social Networks, Algebraic Models for, Table 2
Right multiplication table and partial order for the semigroup of the multiple network of Table 1

Right multiplication table				Partial order														
		Generators			Elements													
Element	Word	*F*	*N*		1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	*F*	3	4	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	*N*	5	6	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0
3	*FF*	7	4	3	1	0	1	0	0	0	0	0	0	0	0	0	0	0
4	*FN*	8	9	4	0	1	0	1	0	0	0	0	0	0	0	0	0	0
5	*NF*	10	9	5	0	1	0	0	1	0	0	0	0	0	0	0	0	0
6	*NN*	11	12	6	0	0	0	0	0	1	0	0	0	0	0	0	0	0
7	*FFF*	7	4	7	1	0	1	0	0	0	1	0	0	0	0	0	0	0
8	*FNF*	8	9	8	0	1	0	1	1	0	0	1	0	1	0	0	0	1
9	*FNN*	13	12	9	0	0	0	0	0	1	0	0	1	0	0	0	0	0
10	*NFF*	14	9	10	0	1	0	0	1	0	0	0	0	1	0	0	0	0
11	*NNF*	11	12	11	1	0	1	0	0	1	1	0	0	0	1	0	0	0
12	*NNN*	13	12	12	0	1	0	1	0	1	0	0	1	0	0	1	0	0
13	*FNNF*	13	12	13	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14	*NFFF*	14	9	14	0	1	0	0	1	0	0	0	0	1	0	0	0	1

$\{FN^2F, NF^3\}$ and $W_5 = \emptyset$; equations include $F^2N = FN, NFN = FN^2$, and $F^4 = F^3$.

It is evident from Table 2 that the relations F and N both have index 3 and period 1. The relation F is regular, whereas the relation N is not.

Each of the equations of Table 2 specifies the empirical coincidence of potentially distinct labelled walks. The equation $F^2N = FN$, for example, indicates the those individuals nominated by one's friends as preferred non-friends coincide with individuals so nominated by the friends of one's friends.

Role Algebra

The role algebra for block 4 in the network of Table 1 is presented in Table 3. Equations in the right multiplication table indicate equalities among labelled paths having block 4 as the source. For example, the equation $FF = F$ indicates that friendship paths of length 2 emanating from block 4 reach exactly the same set of blocks as direct friendship ties. Likewise, the equation $FN = N$ allows us to infer that, namely block 4 would prefer not to have as friends precisely those blocks so nominated by their friends. The partial order diagram indicates that block 4's friendship ties are a subset of those to whom block 4 has negative tie paths (block 4) of length 2.

The role algebra of block 4 has a unique factorization. The lattice of π-relations has two atoms and each atom possesses a unique maximal meet-complement; the atoms and their maximal meet-complements are shown in Table 4. These unique maximal meet-complements are associated with the role algebras shown in Table 5. In the first factor, both *FN* and *NF* ties from block 4 coincide with *N* ties, and *NNF* ties include all others. In the second factor, any path from block 4 whose last tie is labelled *N* coincides with an *N* tie.

Social Networks, Algebraic Models for, Table 3
The role algebra for block 4 in the network of Table 1

Right multiplication table				Partial order						
		Generators			Elements					
Element	Word	*F*	*N*		1	2	3	4	5	6
1	*F*	1	2	1	1	0	0	0	0	0
2	*N*	3	4	2	0	1	0	0	0	0
3	*NF*	3	4	3	0	1	1	0	0	0
4	*NN*	5	6	4	1	0	0	1	0	0
5	*NNF*	5	6	5	1	1	1	1	1	1
6	*NNN*	5	6	6	1	1	0	1	0	1

Social Networks, Algebraic Models for, Table 4
Atoms and unique maximal-meet complements for the block 4 role algebra

	Partial order							Partial order						
		Elements							Elements					
	Elements	1	2	3	4	5	6	Elements	1	2	3	4	5	6
Atom	1	1	0	0	0	0	0	1	1	0	0	0	0	0
	2	0	1	0	0	0	0	2	0	1	0	0	0	0
	3	0	1	1	0	0	0	3	0	1	1	0	0	0
	4	1	1	0	1	0	0	4	1	0	0	1	0	0
	5	1	1	1	1	1	1	5	1	1	1	1	1	1
	6	1	1	0	1	0	1	6	1	1	1	1	0	1
Maximal meet-complement	1	1	0	0	0	0	0	1	1	0	0	0	0	0
	2	0	1	1	0	0	0	2	1	1	0	1	0	1
	3	0	1	1	0	0	0	3	1	1	1	1	1	1
	4	1	0	0	1	0	0	4	1	1	0	1	0	1
	5	1	1	1	1	1	1	5	1	1	1	1	1	1
	6	1	1	1	1	1	1	6	1	1	0	1	0	1

Lattice of an Affiliation Network

The affiliation network in Table 6 is an approximation to the so-called "Southern Women" data [12,23]. The approximation was obtained by a method described in [52], and is akin to a form of dual clustering of the rows and columns of the original data array. The Galois lattice of the (approximate) affiliation array is presented in Fig. 1. In Fig. 1, known as the *line diagram* of the lattice, one element s is drawn above and connected to another element t if s covers t, (that is if $s \geq t$ and there is no element u distinct from s and t for which $s \geq u \geq t$) [51]. The meet of two elements is the greatest element lying below both elements and connected to them by a descending path. The join of two elements is the least element to which they are both connected by an ascending path.

Social Networks, Algebraic Models for, Table 5
Factors of the role algebra for block 4

Factor 1	Right multiplication table			Partial order				
		Generators			Elements			
Element	Word	*F*	*N*		1	2	3	4
1	*F*	1	2	1	1	0	0	0
2	*N*	2	3	2	0	1	0	0
3	*NN*	4	4	3	1	0	1	0
4	*NNF*	4	4	4	1	1	1	1

Factor 2	Right multiplication table			Partial order			
		Generators			Elements		
Element	Word	*F*	*N*		1	2	3
1	*F*	1	2	1	1	0	0
2	*N*	3	2	2	0	1	0
3	*NF*	3	2	3	0	1	1

Social Networks, Algebraic Models for, Table 6
An affiliation network: an approximation to the Southern Women data; the approximation differs from the original data in the 20 underlined values

Event														
Woman	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	1	1	1	1	1	$\underline{1}$	1	1	0	0	0	0	0
2	1	1	1	$\underline{1}$	1	1	1	1	$\underline{1}$	0	0	0	0	0
3	$\underline{1}$	1	1	1	1	1	1	1	1	0	0	0	0	0
4	$\underline{0}$	0	1	1	1	1	1	1	0	0	0	0	0	0
5	0	0	1	1	1	$\underline{1}$	1	$\underline{1}$	0	0	0	0	0	0
6	0	0	1	0	1	1	0	1	0	0	0	0	0	0
7	0	0	0	0	$\underline{0}$	$\underline{0}$	1	1	0	0	0	0	0	0
8	0	0	0	0	0	$\underline{0}$	0	1	1	0	0	0	0	0
9	0	0	0	0	$\underline{0}$	0	1	1	1	0	0	0	0	0
10	0	0	0	0	0	0	1	1	1	0	0	$\underline{0}$	0	0
11	0	0	0	0	0	0	0	1	$\underline{0}$	1	0	1	0	0
12	0	0	0	0	0	0	0	1	1	1	0	1	1	1
13	0	0	0	0	0	0	$\underline{0}$	1	1	1	0	1	1	1
14	0	0	0	0	0	$\underline{0}$	1	$\underline{1}$	1	1	1	1	1	1
15	0	0	0	0	0	0	1	1	0	1	1	1	0	0
16	0	0	0	0	0	0	0	1	1	0	0	0	0	0
17	0	0	0	0	0	0	0	$\underline{1}$	1	0	$\underline{0}$	0	0	0
18	0	0	0	0	0	0	0	$\underline{1}$	1	0	$\underline{0}$	0	0	0

The rows and columns of the affiliation array may be mapped to lattice elements in such a way that any row element (in this case, a woman) is at or above any column elements (in this case, events) with which it is affiliated; likewise, any column element (event) is at or below any affiliated row elements (women). It can be seen from Fig. 1, for example, that all women attend event e8, whereas only women W1, W2 and W3 attend events e1 and e2.

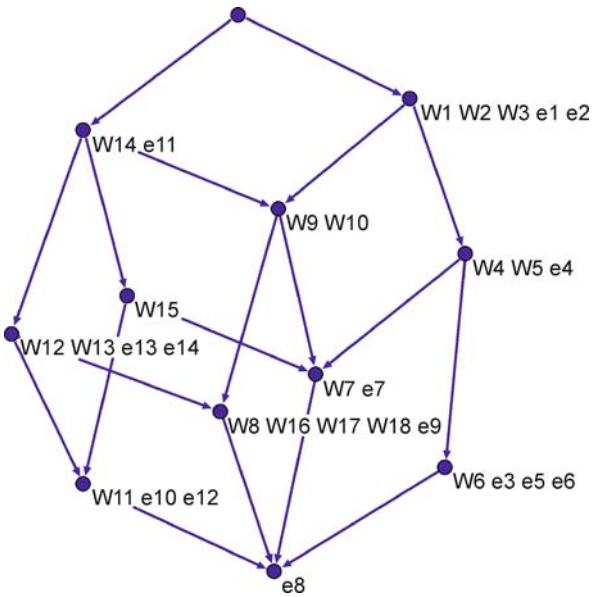

Social Networks, Algebraic Models for, Figure 1
The line diagram of the Galois lattice of Table 4

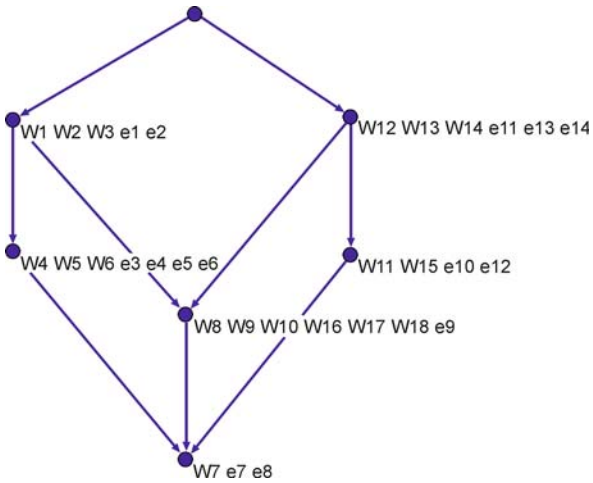

Social Networks, Algebraic Models for, Figure 2
Largest homomorphic image of the lattice of Fig. 1

The lattice displayed in Fig. 2 is a maximal homomorphic image of the lattice of Fig. 1 and provides a simplification of the lattice. In this diagram, a division of the women in the network into three ordered clusters is apparent. One cluster comprises the women W1 to W6, with women W4, W5 and W6 attending only a subset of the events attended by women W1, W2 and W3. A second cluster comprises women W11 to W15, with women W11 and W15 attending only a subset of the events attended by women W12, W13 and W14. The third cluster of women comprises W7, W8, W16, W17 and W18 and is again divided into two sub-clusters, one of whom (W7) attends a subset of the events attended by the other (W8, W16, W17, W18). It is also clear from the lattice of Fig. 2 that the first two clusters of women attend a mixture of distinct events (e1 to e6 in the case of the first cluster, and e10 to e14 in the case of the second) as well as common events (e7 to e9), whereas the third cluster of women attend only those events also attended by women in the first two clusters (e7 to e9). This structural feature of the affiliation network is drawn out very clearly in the unique maximal homomorphic image of the lattice of Fig. 2, shown in Fig. 3.

Future Directions

The algebraic constructions described above provide exact and detailed representations of structural relationships

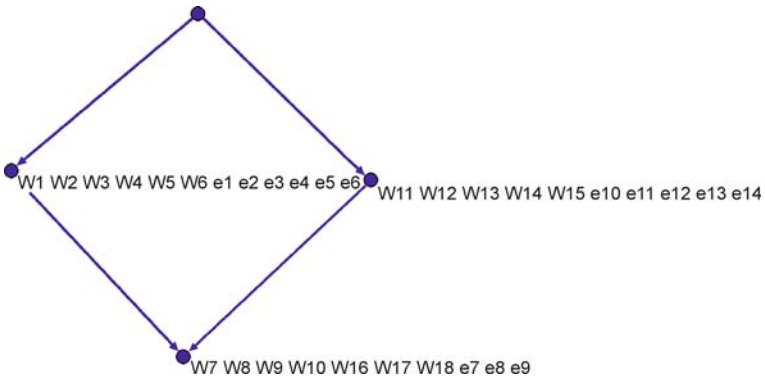

Social Networks, Algebraic Models for, Figure 3
Largest homomorphic image of the lattice of Fig. 2

among network constituents. Of course, this level of detail assumes that the relational observations comprising the network are accurate, as no allowance is made for variability or error. This is often a tenuous assumption, and two important directions for further development utilize the structural sensitivities of algebraic representations in the presence of network tie variability.

The first direction is to use statistical criteria to generate (partial) algebraic structures that summarize key structural features of a network [1,8,39,40]. Instead of the generalized Axiom of Quality introduced earlier, an *Approximate Axiom of Quality* is proposed in which $u \leq v$ in W whenever there is "sufficient evidence" that the relation $X_u \leq X_v$ holds. Such an approach can lead to a theoretically-guided and structurally-focused form of exploratory data analysis for multiple networks. Theoretical guidance comes from choice of operations in the set F, while structural focus resides in the (partial) algebra to which the approach gives rise.

The second direction is to understand how exact algebraic representations can emerge as special, so-called *degenerate*, cases of statistical models such as exponential random graph models for networks (see the entry on ► Social Networks, Exponential Random Graph ($\boldsymbol{p}^*$) Models for *for Networks*). An exponential random graph model defines a probability distribution on the set of all networks on a node set N as a function of some parameter vector. For any parameter vector, a subset of networks have what can be defined as minimum "energy". As the parameter vector is scaled by an increasingly large multiplier, the probability associated with any network whose energy exceeds the minimum tends to zero. The minimum energy networks can therefore be seen as highly constrained or "frozen" structural forms associated with the stochastic model [21,22,41]. The structure of these forms will often warrant algebraic analysis of the type described earlier; in addition, we can understand the non-frozen models as stochastic generalizations of these structural forms.

Bibliography

1. Barnes G, Cerrito P, Levi I (1998) A mathematical model for interpersonal relationships in social networks. Soc Netw 20:179–196
2. Birkhoff G (1963) Lattice theory, 3rd edn. American Mathematical Society, Providence
3. Bonacich P (1980) The common structure semigroup, a replacement for the Boorman and White joint reduction. Am J Soc 86:159–166
4. Bonacich P (1983) Representations for homomorphisms. Soc Netw 5:173–192
5. Boorman SA, White HC (1976) Social structures from multiple networks II. Role structures. Am J Soc 81:1384–1446
6. Boyd JP (1969) The algebra of group kinship. J Math Psychol 6:139–167
7. Boyd JP (1989) Social semigroups and Green relations. In: Freeman LC, White DR, Romney AK (eds) Research methods in social network analysis. George Mason University Press, Fairfax, pp 215–254
8. Boyd JP (1990) Social semigroups: A unified theory of scaling and blockmodelling as applied to social networks. George Mason University Press, Fairfax
9. Boyd JP, Haehl JH, Sailer LD (1989) Kinship systems and inverse semigroups. J Math Soc 2:37–61
10. Breiger RL (1974) The duality of persons and groups. Soc Forces 53:181–190
11. Breiger RL, Pattison PE (1986) Cumulated social roles: the duality of persons and their algebras. Soc Netw 8:215–256
12. Davis A, Gardner B, Gardner M (1941) Deep south. Chicago University Press, Chicago
13. Doreian P (1981) Polyhedral dynamics and conflict mobilization in social networks. Soc Netw 3:107–116
14. Duquenne V (1996) On lattice approximations: syntactic aspects. Soc Netw 18:189–200
15. Freeman LC (1996) Cliques, Galois lattices, and the structure of human social groups. Soc Netw 18:173–87
16. Freeman LC, White DR (1993) Using Galois lattices to represent network data. In: Marsden P (ed) Sociological methodology. American Sociological Association, Washington, pp 127–146
17. Friedell M (1967) Organizations as semilattices. Am Soc Rev 32:46–54
18. Fuchs L (1963) Partially ordered algebraic systems. Pergamon, Oxford
19. Ganter B, Wille R (1989) Conceptual scaling. In: Roberts F (ed) Applications of combinatorics and graph theory in the biological and social sciences. Springer, New York, pp 139–167
20. Ganter B, Wille R (1999) Formal concept analysis: mathematical foundations. Springer, New York
21. Grenander U (1993) General pattern theory. Oxford University Press, Oxford
22. Handcock MS (2003) Statistical models for social networks. In: Breiger RL, Carley KM, Pattison PE (eds) Dynamic social network modeling and analysis. National Academies Press, Washington
23. Homans G (1951) The human group. Routledge & Kegan Paul, London
24. Kim KH (1982) Boolean matrix theory and applications. Dekker, New York
25. Kim KH, Roush FW (1984) Group relationships and homomorphisms of Boolean matrix semigroups. J Math Psychol 28:448–452
26. Lorrain F (1975) Reseaux sociaux et classifications sociales. Hermann, Paris
27. Lorrain F, White HC (1971) Structural equivalence of individuals in social networks. J Math Soc 1:49–80
28. Luce RD (1956) A note on Boolean matrix theory. Proc Am Math Soc 3:382–388
29. Mandel M (1983) Local roles and social networks. Am Soc Rev 48:376–386
30. Martin JL (2002) Some algebraic structures for diffusion in social networks. J Math Soc 26:123–146

31. Martin JL (2006) Jointness and duality in algebraic approaches to dichotomous data. Soc Method Res 35:159–192
32. McConaghy M (1981) The common role structure: improved blockmodelling methods applied to two communities' elites. Soc Method Res 9:267–285
33. Mische A, Pattison PE (2000) Composing a civic arena: Publics, projects and social settings. Poetics 27:163–194
34. Pattison PE (1982) The analysis of semigroups of multirelational systems. J Math Psychol 25:87–118
35. Pattison PE (1989) Mathematical models for local social networks. In: Keats J, Taft R, Heath R, Lovibond S (eds) Mathematical and theoretical systems. North Holland, Amsterdam
36. Pattison PE (1993) Algebraic models for social networks. Cambridge University Press, New York
37. Pattison PE, Bartlett WK (1982) A factorization procedure for finite algebras. J Math Psychol 25:51–81
38. Pattison PE, Breiger RL (2002) Lattices and dimensional representations: matrix decompositions and ordering structures. Soc Netw 24:423–444
39. Pattison PE, Wasserman S (1995) Constructing algebraic models for local social networks using statistical methods. J Math Psychol 39:57–72
40. Pattison PE, Wasserman S, Robins G, Kanfer A (2000) Statistical evaluation of algebraic constraints for social networks. J Math Psychol 44:536–568
41. Robins G, Woolcock J, Pattison P (2005) Small and other worlds: Global network structures from local processes. Am J Sociol 110:894–936
42. Rutherford D (1963) Inverses of Boolean matrices. Proc Glasgow Math Assoc 6:49–53
43. Schein B (1970) A construction for idempotent binary relations. Proc Jpn Acad 46:246–247
44. Schein B (1976) Regular elements of the semigroup of all binary relations. Semigr Forum 13:95–102
45. Schweizer T (1993) The dual ordering of actors and possessions. Curr Anthropol 34:469–483
46. Vickers M (1981) Relational analysis: an applied evaluation. MSc Thesis, University of Melbourne, Melbourne
47. White DR, Duquenne V (1996) Special issue on social networks and discrete structure analysis. Soc Netw 18:169–318
48. White HC (1963) An anatomy of kinship. Prentice-Hall, Englewood Cliffs
49. White HC (1992) Identity and control. University of Chicago Press, Chicago
50. White HC, Boorman SA, Breiger RL (1976) Social structure from multiple networks: I. Blockmodels of roles and positions. Am J Sociol 81:730–780
51. Wille R (1984) Line diagrams of hierarchical concept systems. Intern Classif 11:77–86
52. Wilson S, Bladin P, Saling M, Pattison P (2005) Characterizing psychosocial outcome trajectories following seizure surgery. Epilepsy Behav 6:570–580
53. Winship C, Mandel M (1983) Roles and positions: a critique and extension of the blockmodelling approach. In: Leinhardt S (ed) Sociological methodology. Jossey-Bass, San Francisco, pp 314–344
54. Wu L (1983) Local blockmodel algebras for analyzing social networks. In: Leinhardt S (ed) Sociological methodology. Jossey-Bass, San Francisco, pp 272–313

Social Networks, Diffusion Processes in

Thomas W. Valente
Department of Preventive Medicine, School of Medicine, University of Southern California, Alhambra, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Network Models
Empirical Studies
Agent Based Models
Conclusions
Future Directions
Acknowledgments
Bibliography

Glossary

Adoption A person's change in behavior.

Agent based models Creation of hypothetical (sometimes prototypical) network structures and the simulation of diffusion within those structures.

Homophily Tendency for people to be connected to others like themselves.

Incidence The percent of new adopters at each time period.

Internal versus external influence Internal influence posits that adoption is driven by person-to-person persuasion whereas external influence posits it is driven by sources outside the network such as mass media.

Event history analysis Transformation of data to represent person-time observations.

Network exposure The degree of behavioral adoption in each person's network neighborhood.

Network threshold The number or percent of adopters in a person's neighborhood necessary for a person to adopt the innovation.

Rate The speed of diffusion.

Prevalence The cumulative percent of adopters in the population.

Weight matrix Any N×N matrix representing potential distances or similarities that models potential pathways for adoption influence (e. g., a structural equivalence matrix derived from an adjacency matrix to model influence of structural equivalence relations on adoption).

Definition of the Subject

Diffusion of innovations is the study of how new ideas and behaviors spread within a population, both within and between communities. Diffusion research spans many disciplines since understanding how new ideas and behaviors spread is an important aspect in many different fields of study. For example, marketers are often interested in how and why people change their purchasing habits and sociologists concerned with how people come to accept new norms. The first empirical investigations date back 100 years, while contemporary studies mark the beginning to the 1943 study of diffusion of hybrid seed corn in Iowa. A subset of the hundreds of empirical diffusion studies have focused specifically on the collection and analysis of social network data combined with data on when individuals adopted selected behaviors.

Introduction

Diffusion of innovations theory attempts to explain how new ideas and practices spread within and between communities. The theory has its roots in anthropology, economics, geography, sociology, marketing, mathematics, among other disciplines [1,2,3,4], and has in some ways been adapted from epidemiology [5,6]. The premise, confirmed by empirical research, is that new ideas and practices spread through interpersonal contacts largely consisting of interpersonal communication [1,7,8,9,10,11].

In their 1943 pioneering study, Ryan and Gross [10] laid the groundwork for the diffusion paradigm by showing that, among other things, social factors rather than economic ones were important influences on adoption [11]. Hundreds of diffusion studies were conducted in the 1950s and early 1960s to examine the diffusion process in more detail across a variety of settings [1]. Many studies sought to understand how information created in government or otherwise sponsored programs could be disseminated more effectively. Diffusion research peaked in the early 1960s, but has been reinvigorated recently with the advent of more sophisticated network models and technology making it possible to study the diffusion process more explicitly.

Most diffusion studies focus on trying to understand the factors that lead some members of a population to adopt a new idea, while others do not. Further, studies try to understand why some people adopt the behavior early while others wait a substantial amount of time before accepting the new practice. For example, Ryan and Gross [10] wanted to know why some farmers purchased hybrid seed corn almost immediately upon its availability while others waited until almost all the farmers in the area purchased it before they were willing to do so. Similarly, Coleman and others [12] wanted to know why some physicians began prescribing tetracycline as soon as it was available, while others waited until most physicians prescribed it before they were willing to do so.

This chapter describes a variety of mathematical and network models used to study the diffusion of these and other innovations. The Coleman and others [12] study provided a conceptual leap from other diffusion studies by explicitly measuring who talked to whom within the community about the innovation. Rogers [13] also collected network data to study the diffusion of weed spray in Iowa in his dissertation. Burt [14] unearthed the data and made it available to the network community so that scholars could debate various models used to describe the network diffusion process. Although having data has been useful for clarifying diffusion models, the limitations of these data and this study, make it a poor choice for studying adoption behavior. Rather, scholars should have focused on collecting better data, or reanalyzing diffusion network data in which contagion was more likely.

This chapter chronicles the development of network diffusion models and indicates where such progress is being made. I first present macro models used to estimate the speed of diffusion and, with the Bass [15] model, to estimate rates of innovation and imitation. Next, spatial autocorrelation is presented which is used to estimate the degree contiguous nodes adopt innovations. Spatial autocorrelation led to the network autocorrelation model which is presented statically (cross-sectional data only) and then with one time lag. I then discuss event history analysis applications of network autocorrelation and its extension by including time-based network interaction terms. Throughout the chapter, I attempt to provide a review of recent research conducted in a variety of domains, but mostly drawn from the public health field.

Macro Models

One consistent finding of diffusion research has been that the cumulative pattern of diffusion follows a growth pattern approximated by a simple one-parameter logistic function such as:

$$y_t = b_0 + \frac{1}{1 + e^{-b_1 t}} , \tag{1}$$

where y is the proportion of adopters, b_0 the y intercept, t is time, and b_1 the rate parameter to be estimated. This simple model can be used to compare growth rates for various innovations, but is extremely limited in its applicability. A considerable improvement was advanced by Bass [15]

Social Networks, Diffusion Processes in, Table 1
Diffusion Rate Parameter Estimates and Moran's *I* Estimates for Two Data Sets

	Medical Innovation	Cameroon Tontine 1 Simulation
One Parameter Model		
Coefficient (95% Confidence Intervals)	0.23 (–.053–0.51)	0.06 (.01–0.12)
N	17	50
R^2	0.76	0.71
Two Parameter (Bass) Model		
Innovation Coefficient (95% C. I.)	–0.43(–0.83–0.03)	–0.20 (–0.30–0.09)
Imitation Coefficient (95% C. I.)	4.09 (3.05–5.12)	2.96 (2.58–3.34)
N	16	49
R^2	0.89	0.89
Moran's *I*	–.13	–.08
z-score	–6.73	–7.80

and many others [16,17,18] by creating a two parameter model:

$$y_t = b_0 + (b_1 - b_0)Y_{t-1} - b_1(Y_{t-1})^2 \,, \tag{2}$$

where y is the proportion of adopters, b_0 a rate parameter for innovation, and b_1 a rate parameter for imitation (the degree of adoption due to prior adopters). The Bass model incorporates the percentage adopters at each time point and thus makes a better estimate of the growth attributable to personal network persuasion. The mathematical model in Eq. 2 can be used to: (1) forecast expected levels of diffusion [17], (2) estimate the rate of diffusion attributed to different theoretical aspects of the diffusion processes, b_0, external influence or innovativeness, and b_1, internal influence or interpersonal persuasion [15,16,18]. This model can be used to estimate rate of disease spread from a central source such as contaminated food or from infectious spread through interpersonal contact. In the social realm, one can use the model to estimate rate of adoption from a mass media advertisement or from interpersonal influence. Rate parameter estimates from both models for two diffusion datasets are provided in Table 1. Interpretation of these estimates is highly dependent on the time scale used to measure diffusion.

These rate parameter estimates can be used as outcomes to study factors associated with diffusion at the macro-level by comparing rates between groups and/or populations. For example, parameter estimates for different countries can be compared to study factors associated with the spread of behaviors in different countries. Modeling at this macro-level, however, is imprecise at best because it assumes perfect social mixing, everyone interacting with everyone else [19,20]. These macro models do not measure whether people who are connected to one another engage in the same behaviors. Geographers have devoted considerable attention to trying to determine whether innovations spread between contiguous areas.

Spatial Autocorrelation

Rather than just estimate rate of diffusion, spatial models measure whether artifacts, diseases, farming practices, and other behaviors spread between contiguous areas [2,21,22]. Proximity data are easy to obtain and are relatively unambiguous, thus providing a network of connections based on distance. Moran's I was an early model developed to test for spatial association, geographic clustering of adoption:

$$I = \frac{N \sum_i^N \sum_j^N D_{ij}(y_i - \bar{y})(y_j - \bar{y})}{S \sum_i^N (y_i - \bar{y})^2} \,, \tag{3}$$

where N is the sample size, D a distance matrix (as proximities), y indicates adoption, and S the sum of the distances in the distance matrix. Moran's I measures the degree nodes that are connected to one another deviate from the average behavior in the network similarly or differently. Moran's I is high when connected nodes (positive elements of D) are either positively or negatively different from the average score. The statistical significance of Moran's I can be calculated in two ways: via permutation methods or analytically.

To use a permutation method to calculate the significance of Moran's I assume adoption (y_i) is randomly distributed and calculate I repeatedly to get a sample of estimates based on D and the number of adopters. If Moran's I calculated is significantly different than the random sample generated, Moran's I is considered significant (z-scores can be obtained). The logic then is to calculate the degree neighbors (however defined) have similar adoption behavior compared to that expected if adoption were distributed randomly. Variance estimators for Moran's I can be found in spatial statistics textbooks [21,23] and used to calculate exact significance tests. Moran's I is useful, and has been extended considerably, yet this approach often assumes that geographic proximity equates with communication and influence, which may not be true.

The spatial autocorrelation methodology was seen as a useful approach to measuring network autocorrelation,

the bias inherent in a regression model when "y" appears as both the dependent and independent variable. Erbing and Young [24] wrote an influential paper on measuring network effects and using network autocorrelation methods. Dow [25,26] demonstrated the effects of network autocorrelation on estimate errors, and Doriean and others [27] found considerable bias in the point estimates and their standard errors. Exactly how network autocorrelation applied to diffusion of innovations was not clear since spatial autocorrelation measured diffusion at the macro level, but did not show whether specific individuals were more or less likely to adopt based on their network position. Further, spatial autocorrelation did not show how network structure influenced diffusion. To do so, we turn to network models.

Social Networks, Diffusion Processes in, Table 2
Social network influence weightings

Relational	Positional	Central
1. Direct Ties	1. Percent Positive Matches (Tie Overlap)	1. Degree
2. Indirect Ties	2. Euclidean Distance	2. Closeness
3. Joint Participation in Groups or Events	3. Regular Equivalence	3. Betweenness
		4. Flow
		5. Integration/radiality
		6. Information
		7. Power

Network Models

Figure 1 displays two networks from a study conducted in Cameroon among women in voluntary organizations [28]. Women were asked to name their friends in the organization in an attempt to determine if friendship ties were associated with contraceptive choices (they were). The diffusion network model posits that initial contraceptive choices would be made by some women based on their innovativeness and exposure to outside sources of influence such as their cosmopoliteness, media use, or greater need for the innovation. The new idea, and its practice, then spreads through the network as users persuade non-users to adopt either by exhortation, entreaty, enticement, or example.

These influences are captured by an exposure or contagion model (Fig. 2), each individuals likelihood of adoption increases as the proportion (or number) of users in his/her personal network increases. Personal network exposure is the proportion or number of adopters in each person's network that provide information and influence with regard to some behavior. The equation for non-random mixing, or personal network exposure is:

$$E_i = \frac{\sum w_{ij} \boldsymbol{y}_j}{\sum w_i}, \tag{4}$$

where w is the social network weight matrix, and y is vector of adoptions. For an individual who reported five contacts, network exposure (E_i) is the proportion of those contacts that have adopted (Fig. 2). When network exposure is measured on direct contacts it captures social influence conveyed through overt transmission of information, persuasion or direct pressure. Alternatively, exposure can be calculated by transforming the social network, W, to reflect other social influence processes. For example, W can be transformed to represent the degree of structural equivalence (similarity in network position) among people in the network. Exposure calculated on this network captures social influence conveyed via comparison to equivalent others such as by social comparison or competition [14]. Exposure can also be weighted by network properties such as centrality to reflect social influence by opinion leaders.

These three social influence processes are modeled with three different classes of network weight matrices (relational, positional, and central), constructed from the same social network data (Table 2). All three can be justified theoretically as sources of influence on adoption behavior and all three can be calculated various ways (there are at least 10 centrality measures). It is possible that all three operate for different people or at different times during the diffusion process.

In addition to the social influence process, a second dimension to these influence mechanisms is the weights attached to each based on social distance. For example, in relational influence models, different weights can be assigned to direct ties, ties-of-ties, and even the ties-of-ties-of-ties; in positional equivalence models, different weights can be assigned to those that are more equivalent than others [7]. A potential line of diffusion network research then is to compare different network weighting mechanisms to model and compare different social influence processes.

Diffusion was simulated through the two Cameroon networks in Fig. 1 to illustrate how network exposure and network structure influence diffusion. At each time period, adoption occurred for the non-adopter with the most nominations received, then network exposure was calculated, all nodes with exposure of 50% or higher were categorized as adopters and the process repeated. We compared diffusion in this network to that simulated in a net-

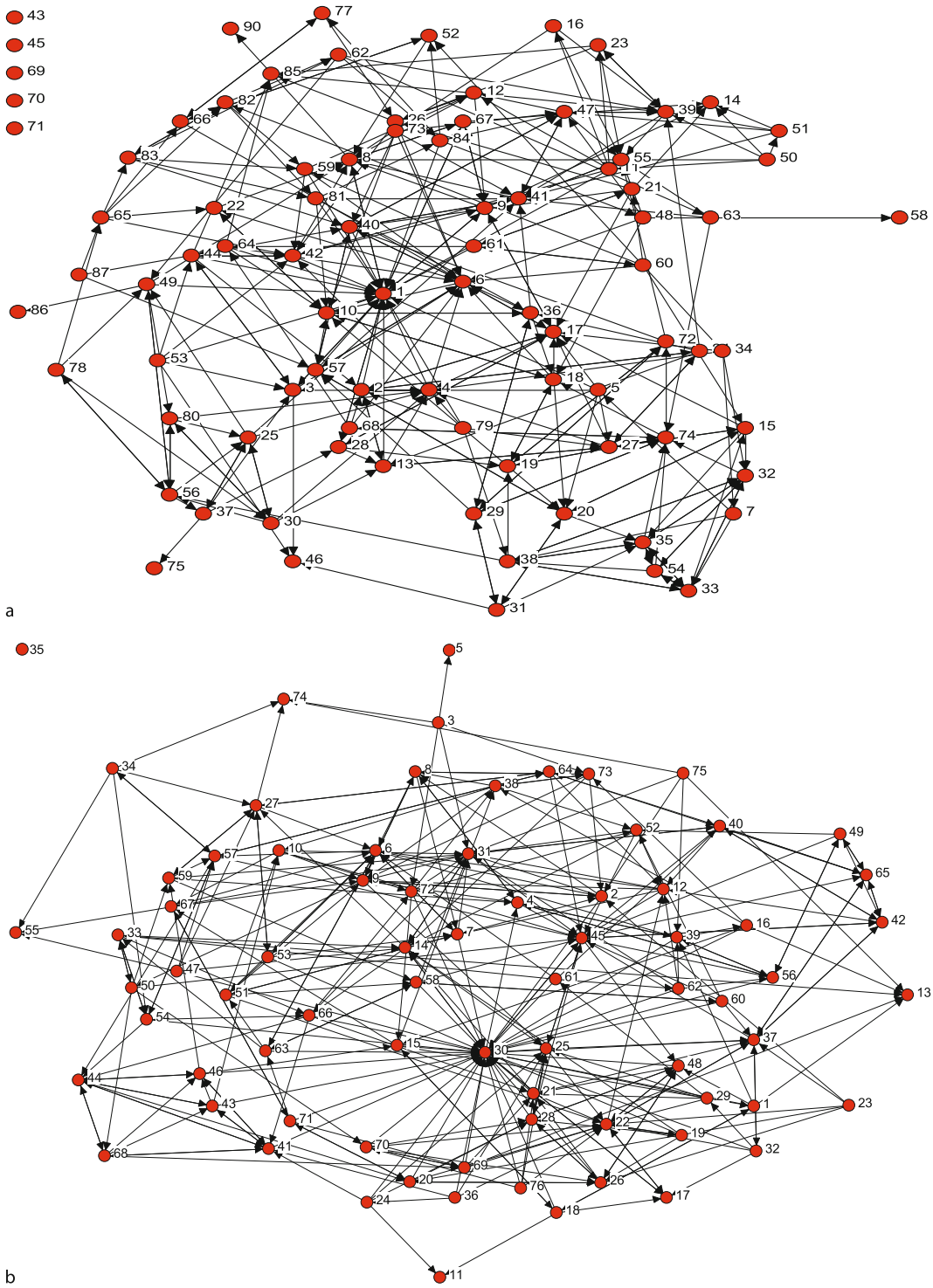

Social Networks, Diffusion Processes in, Figure 1
Networks 1 and 2 from the Cameroon Voluntary Association Study

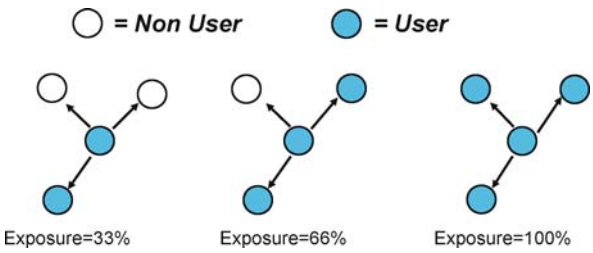

Social Networks, Diffusion Processes in, Figure 2
Personal network exposure from direct contacts

work of the same size and density, but with links allocated randomly. Both conditions were averaged across 1000 runs. Figure 3 shows that, in Network 1, initially the diffusion trajectories are similar, but at about time 10, diffusion in the actual network accelerated.

The network accelerated diffusion because it is somewhat centralized (in-degree 21.7%) and once diffusion reaches the center of the network it can propagate rapidly. Notice that at about time 20 diffusion slowed, accelerated again at about 25, and then slowed from about time 30 to 40. These "fits and starts" are a product of the network structure: diffusion reaches pockets of inter-connectivity and spreads rapidly within these dense pockets, but slows between groups. Network 2 (Fig. 3) had more rapid and sustained diffusion because it was even more centralized (in degree, 47.2%). Note that in the spatial autocorrelation model adoptions were randomized to measure statistical significance and in this simulation, the network structure was randomized to illustrate its influence on the rate of diffusion.

Simulation assumptions regarding influences on adoption could easily be changed to achieve different outcomes. For example, when adoptions were assigned randomly, diffusion was constant in network one (and saturation lower) and similar to the random network in network two. The validity of these diffusion models rests partly on determining whether network exposure influences adoption. To that end, a number of empirical studies have been conducted to measure the degree social network exposure is associated with adoption.

Empirical Studies

Empirical support for an association between individual own behavior and that of his/her peers can be found throughout the behavioral sciences literature. While many scholars assume adoption is associated with network exposure, few studies have traced an innovation through a network of social contacts to empirically validate this proposition. The lack of data on diffusion within an entire network stems largely from the difficulty of trying to collect data over a time period long enough for diffusion to occur. Consequently, most studies have relied on retrospective data which introduces some but not much bias [29,30]. It has also meant that several scholars have re-analyzed two studies that collected network and adoption data, (1) medical innovation study [12], reanalyzed by Burt [14], Marsden and Podolny [31], Strang and Tuma [32], Valente [7,33], and Van den Bulte and Lillien [34]; (2) Korean family planning study [35], reanalyzed by Dozier [36], Valente [7,33], Montgomery and Chung [37], and Kohler [38]. Recent studies in the fields of reproductive health [39] and substance abuse [40] have provided new data, but these classics remain classic.

Because collecting complete network data can be difficult, most empirical research has been ego-centric [41,42], based on respondent reports of their behavior and that of their network peers who are not necessarily connected to one another and not interviewed. Social influence is often based on respondent reports of perceptions of peer behavior or perceptions of peer influence [43,44]. Comparison of exposure scores based on respondent perceptions with alters' reports in one study found that perceptions were more strongly related to behavior than exposure based on alter reports [28,37,45].

Sociometric studies interview members of a bounded community and attempt to gather information from everyone in the community (typically conducted in schools, organizations, and small communities) and record their time of adoption [1,12,46,47,48]. Sociometric studies are useful for understanding how an innovation flows within the community and how certain network structural variables influence the diffusion process. Sociometric data capture network influences by the alters' reports since they were also interviewed. For example, sociometric studies can determine whether structural positions such as centrality are associated with adoption and/or whether centralization is associated with more rapid diffusion [7].

A number of recent diffusion network studies have been cross-sectional and in many cases retrospective involving only one time point. For example, a study in Thailand by Entwisle and others [49] found that contraceptive choices made by early adopters contributed significantly to the contraceptive choices made by later adopters (also see [35]). Valente and others [28] collected sociometric data on contraceptive use among women in voluntary associations in Cameroon and showed that perceptions of these friends' behavior, and in particular perceptions that these friends encouraged contraceptive use, were significantly associated with behavior. In general, statistical ana-

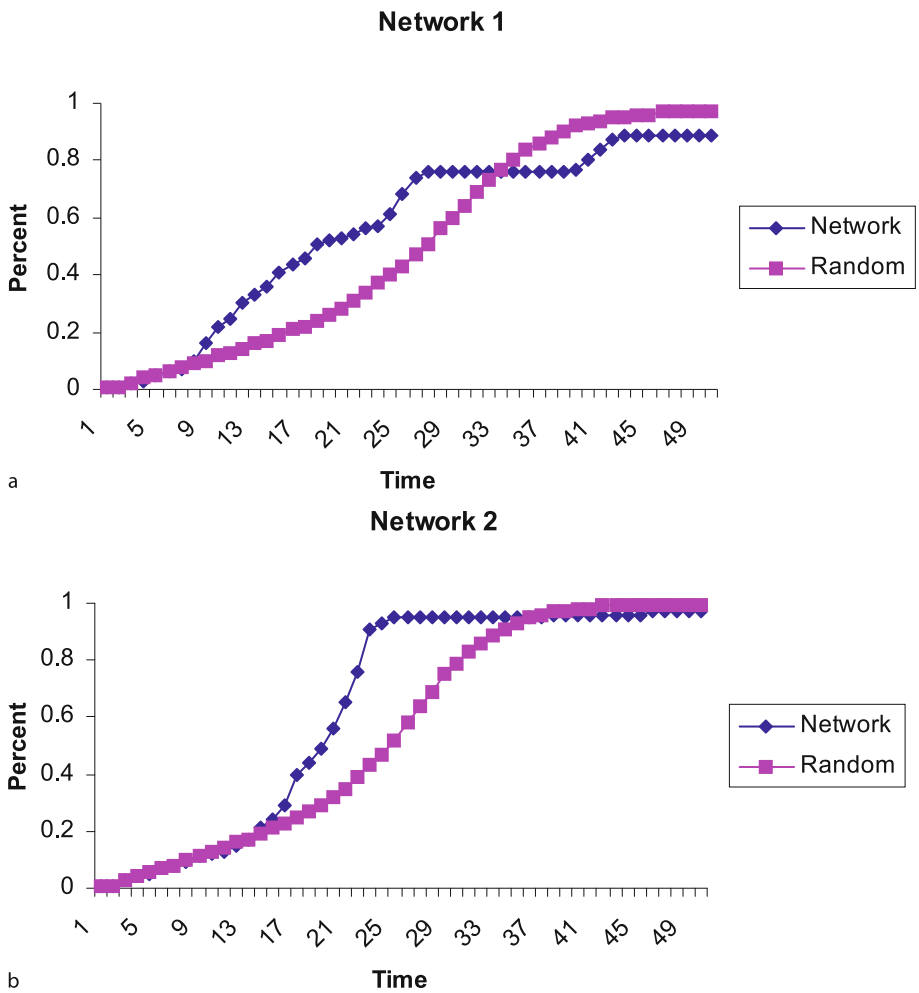

Social Networks, Diffusion Processes in, Figure 3
Simulated diffusion in two networks, each compared to random networks of the same size and density. Network 2 is more centralized than Network 1

lyzes use the following model:

$$\log \frac{\Pr(\boldsymbol{y}_t = 1)}{(1 - \Pr(\boldsymbol{y}_t = 1))} = \alpha + \sum B_k X_k + B_{(k+1)} \omega \boldsymbol{y}_t, \quad (5)$$

where $\boldsymbol{y}$ is a binary vector of adoption behavior, α is the intercept, β_k are parameter estimates for vectors of K socio-demographic characteristics (Xs), and ω represents the social network matrix. The $\omega \boldsymbol{y}_t$ term represents the calculation of contemporaneous network exposure and this vector is usually divided by a count of the number of nominations sent (alternatively the number of nominations can be entered into the regression separately).

Significant estimates for β_{k+1} indicate contagion effects by showing that network exposure is associated with adoption. The variances for these estimates, however, are usually biased since the observations are not independent and hence the errors in prediction are not independent. One partial solution is to obtain robust estimates by controlling for clustering. Clustering is the degree that elements from the same cluster are similar compared to those of different clusters. For example, two individuals chosen at random from the same organization are more likely to be similar than two chosen at random from different organizations. Table 3 reports regression results of the Cameroon data with and without correction for clustering. Without correction, network exposure is strongly and significantly associated with adoption, but with the correction it is only marginally statistically significant (p=.04). Controlling for clustering is particularly important in network exposure models because network choices are often restricted to the cluster.

Social Networks, Diffusion Processes in, Table 3
Logistic regression on the likelihood of contraceptive behavior on controls and network exposure with and without correction for clustering (N=555; Groups=9)

	Contraceptive Method Use			
	Without Correction		With Correction	
	Adjusted		Adjusted	
	Odds Raios	P-value	Odds Raios	P-value
Age	0.97	0.001	0.97	0.010
Education	0.91	0.247	0.91	0.184
Possessions	1.39	0.000	1.39	0.000
Network Exposure	1.14	0.005	1.14	0.047

Even with clustering controlled, social influence as measured through social networks seems to be strongly associated with behavior. For example, a school-based sociometric study was conducted by Alexander and colleagues [50] using the Adolescent health data [51] to show that students with a majority of network ties who were smokers were almost two times as likely to smoke themselves with an additional two times greater likelihood of smoking for those with best friends who smoke. Clustering was controlled and the multi-level model accurately captured micro-level effects within the context of macro-level influences. The study measured the influence of peers on smoking while conditioning on the smoking rate within the school [50].

Estimating the network exposure (autocorrelation) term with a multi-level model can provide contagion estimates across settings and estimate the degree it varies between settings (i. e., communities, schools, organizations, and so on). The models are incomplete however, because there may be factors that influence both adoption and choice of social network contacts. For example, the decision to smoke and to nominate friends who smoke may both be a function of delinquency or rebellion. Hence an association between behavior and peer behavior can be spurious. Testing social influence with network methods then requires longitudinal data involving at least two time points. Boulay and Valente [52] collected data among women in three villages of Nepal and found that having discussion partners who used contraception influenced information seeking behavior and contraceptive choice. Having data from two time points allows testing of a simple dynamic model on adoption:

$$\log \frac{\Pr(\boldsymbol{y}_t = 1)}{(1 - \Pr(\boldsymbol{y}_t = 1))} = \alpha + \sum B_k X_k + B_{(k+1)}\omega_t \boldsymbol{y}_t + B_{(k+2)}\omega_{(t-1)}\boldsymbol{y}_{(t-1)}, \quad (6)$$

where y is a binary indicator of behavior, α is the intercept, β_k are parameter estimates for vectors of K socio-demographic characteristics (Xs), and ω represents the social network matrix. A positive and significant β_{k+2} indicates that respondents with high network exposure at baseline were more likely to adopt at time two. A positive and significant β_{k+1} indicates that change in network exposure is associated with change in behavior. This may indicate contagion but still may be a product of some omitted factor. Panel data collected at two time periods is adequate for most research needs, and can provide evidence of network influences on behavior. However, since there is often a considerable time between the two measures, many factors may account for simultaneous change in behavior and network exposure. To cope with this threat, data can be collected on time of adoption, expanding the micro level dynamic analysis by using event history analysis [53].

Event History Analysis

Event history analysis techniques have been developed to analyze data with a substantive number of time points, estimating coefficients with maximum likelihood estimators [32,53,54,55,56]. There are two types of event history analysis, discrete time, in which the outcome is binary, and continuous time, in which the outcome is time-to-an-event. Since diffusion occurs over time, there is an explicit time dimension in diffusion studies captured by both discrete and continuous time models. The time of adoption variable is the dependent variable and may be influenced by both time-varying and time-constant factors. Some individuals may not have adopted by the time of data collection giving rise to time-censored observations. Right censoring occurs when data are collected before the innovation has finished diffusing or does not diffuse to all members of the community or study. Left censoring occurs when the data are incomplete at the beginning of the process. For example, adoption data for the period 1993 to 2000 may have some people who adopted in 1989–1992 classified as 1992 adopters.

There are a variety of event history techniques including hazard models developed in epidemiology used to understand the hazard or risk to disease or injury over time. Hazard and/or event history analysis generally requires that the data be reshaped from simple observations to a case-time format such that there is a case in the data for each individual at each time period of study up to and including that person's time of adoption. The time-varying and time-constant independent variables are included in each case as well as a binary indicator for whether the individual adopted or not (got sick or not).

Maximum-likelihood estimation can determine whether the independent variables are associated with the dependent variable (adopt/not adopt) [57]. A study of 100 people with an average adoption time of seven translates into 700 person-time cases. Each person-time case has a variable for the network exposure at that time period plus an indicator for whether the person adopted or not (plus additional time-constant and time-varying covariates as desired). The event history model is:

$$\log \frac{\Pr(\boldsymbol{y}_t = 1)}{(1 - \Pr(\boldsymbol{y}_t = 1))} = \alpha + \sum B_j X_j + \sum B_{kt} X_{kt} + \sum B_{(k+1)} \omega \boldsymbol{y}_t \,, \tag{7}$$

where y is a binary indicator of behavior, α is the intercept, β_j are parameter estimates for vectors of J socio-demographic characteristics (X_j), and β_{kt} are parameter estimates for the matrix of time-varying socio-demographic characteristics (X_{kt}) and ω represents the social network weight matrix, and t a time indicator. Note here we have assumed a static (constant) network. Standard statistical packages allow testing of event history or survival data in a relatively straightforward manner, once the data are reformatted. Event history analysis requires the construction of exposure matrices for each time period which can be a formidable task particularly if one uses more than one network weight matrix.

Marsden and Podolny [31] used event history analysis and tested network exposure's association with adoption in the medical innovation data. Results showed that exposure was not associated with adoption in that study. Strang and Tuma [32] revisited the issue with the same data by postulating time variance in network influence, (i. e., how much lag time, if any, is there in the influence). They found evidence of contagion. Van den Bulte and Lillien [34] supplemented the medical innovation data with archival data on media promotion by pharmaceutical firms at the time of the original study and showed that network contagion effects disappear once these data are added. Their analysis demonstrates the importance of omitted variables when studying diffusion through networks. The rapid diffusion measured in the medical innovation study indicates that contagion was probably not the primary factor driving diffusion.

Event history analysis of the three classic diffusion network datasets has been conducted [58]. The analysis controlled for within village and within person covariation, and terms for time and a logistic transformation of time were included to control for macro-level effects. Terms for infection and susceptibility [32,59] were included to measure whether adoption by central individuals (high in-degree) influenced subsequent adoption, infection, and whether centrality (out-degree) influenced a person's likelihood to adopt as diffusion occurred, susceptibility. In-degree and out-degree were also included in the model. Two network exposure terms were computed, direct ties and structural equivalence. Structural equivalence was computed as in Burt's [14] measure, Euclidian distance raised to the 16th power. For this analysis, network exposure was calculated using contemporaneous measures since two of the datasets recorded adoption in one-year intervals. Two control variables representing individual characteristics were included. Analysis was conducted only on those who adopted. The following model was estimated:

$$\log(\Pr(\boldsymbol{y}_t = 1)) = \alpha + \sum \beta_{lm} X_{lm} + \sum \beta_{lmt} V_{lmt} + \sum \beta_{lmt} \omega_s \boldsymbol{y}_t + \lambda_{lm1} C_D(\boldsymbol{y}_+) + \lambda_{lm2} C_D(\boldsymbol{y}_+) \,, \tag{8}$$

where y is a binary indicator of behavior, α is the intercept, Xs are vectors or time-constant socio-demographic and network characteristics, V represents vectors of time-varying terms, in this case time and its transformation, ω_s represents the social network matrices, and λ estimates the effects of centrality degree variables multiplied by the time varying proportion of adopters in the network (infection and susceptibility). Results were mixed, but seem to indicate that both infection and susceptibility effects are present. In all three datasets, infection is positively associated with adoption indicating that as those with high in-degree adopt, the likelihood others in the network will adopt increases. In two studies, Brazilian farmers and Korean women, susceptibility is associated with adoption indicating that those with a high number of nominations sent are more likely to adopt as the innovation diffuses. Ties sent and received are marginally associated with adoption, and only for the Brazilian data is exposure, through structural equivalence, associated with adoption. These results, however, change dramatically when non-adopters are included or when a term for the average exposure at each time period is included such that infection and susceptibility effects disappear.

The event history analysis approached has also been used by Montgomery and others (2001) using ego-centric data to study network exposure's influence on contraceptive use in Ghana. Current analysis of four rounds of data over two years has shown that contraceptive use is strongly associated with use by social network peers. The Ghana field study provides some of the most conclusive evidence of the magnitude of social influence on behavior change

by showing that as the number of social network contacts who use contraceptives increases, the likelihood of contraceptives use by ego also increases. Of all the variables, the network exposure variables were the most significant influences on contraceptive adoption. Another longitudinal field study in Kenya found similar results, again based on ego-centric network data [60].

Montgomery and others [61] also report preliminary analysis in which network influences are weighted by tie characteristics such as the frequency of communication. They found that adding these weights did not change the strength of peer influence. Similar results have been reported in Valente [7] and Valente and Saba, p. 109 in [43]. Consequently, it seems that the influence of social networks on behavior (contraceptive use in these cases) seems broad in nature, and are not conditioned on specific factors such as the frequency of communication between dyads or the their socio-demographic equality. These factors may play a strong, and even pervasive role in determining who is connected to whom [62], but they do not seem to determine the degree of influence social contacts provide.

Network exposure and adoption may not always be strongly correlated for a number of reasons. First, exposure may not be associated with adoption for everyone, but may be most influential during the middle stages of diffusion, when awareness is high, but uncertainty about its relative advantages is also high [63]. Exposure may have less of an effect early in the process when there are few adopters and obvious advantages to waiting; and late in the process when most people have a majority of adopters in their personal network anyway. Second, individuals may have varying thresholds to adoption such that some are innovative and others are not [19]. Valente [7,33] posited a social network threshold model in which contagion (majority rule) is a special case. Most simulation models assume majority influence on adoption decisions as was done in the beginning of this chapter. It is reasonable, however, to expect that individuals vary in the amount of network exposure needed to adopt an innovation. Disproving thresholds may not be possible, but construct validity for the concept has been demonstrated [33]. Valente and Saba [43] replicated the threshold model using egocentric data and showed that people with a minority of network members using contraception had higher campaign recall indicating that the media campaign could substitute for interpersonal sources of influences. If thresholds vary, network exposure is needed for people to reach those thresholds, if they do not and the special case of contagion exists, network exposure will determine when individuals adopt.

In spite of the impressive list of studies showing some support for an association between individual behavior and network exposure, and the theoretical simulations of network structure and thresholds, significant work remains to be done. Most scholars and lay people would agree that social networks influence behavior. The barriers to demonstrating this effect, however, have been challenges of data collection and agreement on appropriate statistical methodology. The most commonly analyzed dataset, Medical Innovation, is 45 years old, consists only of 125 respondents and arguably is not a diffusion study at all. The limitation of available data has forced researchers to rely on simulations and agent based modeling approaches to understand theoretical mechanisms driving diffusion.

Agent Based Models

Agent based models (ABM) are computer simulations created to understand how actors (usually though not necessarily people) behave given a set of preexisting conditions and rules for their behavior [64]. A researcher can hypothetically pose a population of people who have some degree of connectedness and some structure to those connections. Model parameters may be set that vary those structural conditions or they may be taken as fixed and other properties varied. To model the diffusion of innovations one might be interested in how network structure affects the rate of diffusion. So a simulation can be created which varies network structure and then simulates the spread of an innovation within those structures. One could also vary other properties of the diffusion by varying the number of initial adopters or the transmissibility of the innovation.

To illustrate, an empirical dataset was used to simulate theoretical network structures and potential diffusion within those structures. Data were collected among adolescents in ninth grade and they were asked to write the names of their five closest friends. A computer program was written to match the names to the roster in the school so a network could be constructed from the data. The network had 150 nodes with 327 friendship links. The agent based model had 2 parameters: (1) the initial 7% seed adopters would vary under 3 conditions: (a) randomly selected seeds, (b) the most central nodes, and (c) peripheral nodes; (2) the type of network structure would vary by 4 conditions (a) the real network, (b) a random one with the same links, (c) a centralized one, and (d) a clustered one.

A simple diffusion model was then run setting a low threshold for adoption such that at each time interval after the initial seeding those nodes connected to 15% other

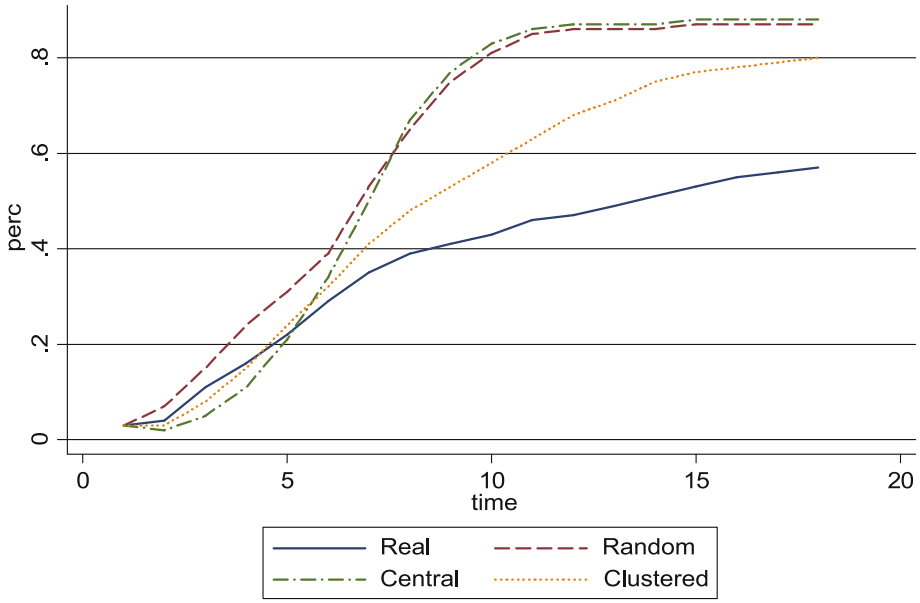

Social Networks, Diffusion Processes in, Figure 4
Comparison of simulated diffusion trajectories for a real network, and 3 simulated ones based on the real network's density. The three simulated networks are random links, centralized links (centralization = 3%), and clustered links (clustering coefficient=12%). Diffusion occurs most rapidly in random and centralized networks

nodes would become adopters. The model was run for 20 time periods which was generally long enough for diffusion to occur in most situations. After a run of 20 time periods was completed, results were tabulated and then the simulation for each condition run 25 times and the results aggregated. The model was run 25 times for each condition to ensure that the results were not the product of some anomaly. It is customary in agent based models to run the simulations for hundreds or thousands to times to guarantee the results are robust.

Figure 4 shows illustrative findings by plotting the diffusion trajectories in the different conditions. The data show that simulated diffusion occurs most rapidly in the random and centralized networks. It is slowest in the real network. The network structure of real networks retards diffusion because of homophily – the tendency for people to be friends with others like themselves. This homophily tendency creates clustered networks so that diffusion gets trapped in these pockets of interconnectivity. Real networks also retard diffusion because they are symmetric (If $A \rightarrow B$ then $B \rightarrow A$) and so adoption by a pair leads to reinforcement, not more diffusion. Bridges are needed to carry the innovation across the gap between clusters and these bridges act as bottlenecks that slow diffusion.

Statistical analysis can be conducted on the data to determine the effect on diffusion. Outcomes can be the final cumulative percent of adopters (the prevalence), or the number of time periods till diffusion has reached 50%, or the rate of adoption estimated using one of the mathematical models described above. In these data we find that all 3 simulated networks have more rapid diffusion than the real network, and we find that seeding the network with central adopters accelerates diffusion while seeding it with peripheral actors slows diffusion.

Conclusions

Much progress has been made since 1943 when Ryan and Gross first laid the foundation for diffusion of innovations theory. Rogers [1] chronicled the many studies conducted since then and helped shaped a general diffusion model with wide applicability now being renewed and reinvigorated with fresh theory and analytic models. Overall, results indicate that social network influences on behavior are important and have consequences for the health and well being of populations and individuals. These new insights have shed light on important aspects of how new ideas and practices spread within and between communities.

Along with new insights have come new questions and new perspectives to be addressed. It is clear that a lack of data on time of adoption coupled with information on net-

work relations has hampered developments. Few diffusion or behavioral studies collect information on networks, conversely few network studies record time of adoption. There are advantages to marrying these two ideas, however, and future research will hopefully try to collect both types of data.

It is also clear that our understanding of how diffusion occurs is still somewhat limited. The Medical Innovation data have often been used to demonstrate the importance of networks in adoption yet analyses by Valente [7] and by Van den Bulte and Lillien [34] have shown that contagion via social influence in this setting was unlikely. Given the number of confounding factors, and some of the data requirements, it may be prohibitively difficult to substantiate the role of social networks in innovation adoption via survey methods alone. Purposively intervening on social networks, however, may prove to be a fruitful avenue of research. If network-based interventions can be used to accelerate innovation diffusion, then a stronger case can be made for the importance of social contagion in the diffusion process.

Nonetheless, it is clear that networks are important influences on behavior since most people acknowledge that they receive information and influence via their social networks and that they model the behavior of others. What is less clear is how to capture that influence in quantitative terms that mimic the theoretical progress made in the network field. Further, verbal accounts on how people make decisions and adopt behaviors usually reveal non-linearities, chance circumstances, and whims that are not independent of networks, but not easily captured in social influence models.

The link between micro and macro levels of analysis represents an opportunity for study of diffusion processes. The opportunity lies in the fact that multi-level modeling techniques enable the separation of micro-level network exposure influences from macro-level contextual factors. Yet both are social network influences and both represent elements of the diffusion paradigm. It is hoped that by controlling for contextual effects we don't "throw the baby out with the bathwater," by eliminating the micro-level influences that provide expressions for those contextual effects.

In spite of controls for macro-level contextual effects, micro-level associations between peer network behavior, and those of respondents are still sometimes strong. Debate remains about the meaning of these associations, is it peer influence, peer selection, or further contextual effects? More rigorous studies may eventually tease this out, in the interim, better study designs and interventions will need to be created. This review has attempted to point out some of the challenges diffusion scholars face and some of the promising new directions it may take.

Scholars have turned to agent based modeling as a technique to investigate diffusion properties. Agent based models (ABM) allow the theoretical exploration of how network structure can be varied and shown to influence diffusion. ABM also enables us to study diffusion parameters such as thresholds, transmissibility, or the importance of initial adopters on diffusion speed and prevalence. Until empirical data can be gathered and analyzed ABM represent a promising approach to understanding how diffusion processes occur on networks.

Future Directions

Future developments in diffusion processes are likely to occur along two interrelated fronts. First, increasingly scholars are using network data to propose and implement social change practices [65]. These activities may involve organizational change in the business context or health promotion and treatment in the medical context. These network based interventions provide the opportunity to change networks and measure the resulting changes in performance or behavior. Thus, scholars can more definitively link changes in networks to changes in outcomes.

Second, considerable progress has been made in advancing statistical models that measure network evolution and behavioral changes that account for multiple structural influences [66]. For example, scholars can test the degree of similarity in behaviors while simultaneously estimating the influence of social network characteristics such as whether two people nominate one another or are part of a linked triad. These models (known as P* or exponential random graph models) provide a statistical estimation of concepts such as network influence, and network selection.

Acknowledgments

Support for this research was provided by NIDA grant #DA10172. Some portions of this chapter appeared in [58].

Bibliography

1. Rogers EM (2003) Diffusion of innovation, 5th edn. The Free Press, New York
2. Hägerstrand T (1967) Innovation diffusion as a spatial process. University of Chicago Press, Chicago
3. Brown L (1981) Innovation diffusion: a new perspective. Methuen, New York
4. Robertson TS (1971) Innovative behavior and communication. Holt, New York

5. Bailey NTJ (1975) The mathematical theory of infectious diseases and its applications. Charles Griffen, London
6. Morris M (1993) Epidemiology and social networks: modeling structured diffusion. Sociol Methods Res 22:99–126
7. Valente T (1995) Network models of the diffusion of innovations. Hampton Press, Cresskill
8. Bael GM, Bohlen JM (1955) How farm people accept new ideas. Cooperative Extension Service Report 15, Ames
9. Katz E, Levine ML, Hamilton H (1963) Traditions of research on the diffusion of innovation. Am Sociol Rev 28:237–253
10. Ryan R, Gross N (1943) The diffusion of hybrid seed corn in two Iowa communities. Rural Sociol 8(1):15–24
11. Valente TW, Rogers EM (1995) The origins and development of the diffusion of innovations paradigm as an example of scientific growth. Sci Commun 16(3):242–73
12. Coleman JS, Katz E, Menzel H (1966) Medical innovation: a diffusion study. Bobbs Merrill, New York
13. Rogers EM (1957) Personality correlates of the adoption of technological practices. Rural Sociol 22:267–268
14. Burt R (1987) Social contagion and innovation: cohesion versus structural equivalence. Am J Sociol 92:1287–1335
15. Bass FM (1969) New product growth for model consumer durables. Manag Sci Ser a-Theory 15(5):215–227
16. Hamblin RL, Jacobsen RB, Miller JLL (1973) A mathematical theory of social change. Wiley, New York
17. Mahajan V, Peterson RA (1985) Models of innovation diffusion. Sage, Newburg Park
18. Valente T (1993) Diffusion of innovations and policy decision-making. J Commun 43(1):30–41
19. Granovetter M (1978) Threshold models of collective behavior. Am J Sociol 83:1420–1443
20. Van den Bulte C, Lillien GL (1997) Bias and systematic change in the parameter estimates of macro-level diffusion models. Marketing Sci 16:338–353
21. Cliff A, Ord JK (1981) Spatial processes: models and applications. Pion, London
22. Griffith DA et al (1999) A casebook for spatial statistical data analysis: a compilation of analyses of different thematic data sets. Oxford University Press, New York
23. Bailey TC, Gatrell AC (1995) Interactive spatial data analysis. Longman, Essex
24. Erbing L, Young A (1979) Individuals and social structure: contextual effects as endogenous feedback. Sociol Methods Res 7:396–430
25. Dow M, Burton ML, White DR (1982) Network autocorrelation: a simulation study of a foundational problem in regression and survey research. Soc Netw 4:169–200
26. Dow M (1986) Model selection procedures for network autocorrelated disturbance models. Sociol Methods Res 14:403–422
27. Doreian P, Teuter K, Wang C (1984) Network autocorrelation models: some Monte Carlo results. Sociol Methods Res 13:155–200
28. Valente TW et al (1997) Social network associations with contraceptive use among Cameroonian women in voluntary associations. Soc Sci Med 45:677–687
29. Coughenour CM (1965) The problem of reliability of adoption data in survey research. Rural Sociol 30:184–203
30. Nischan P et al (1993) Comparison of recalled and validated oral contraceptive histories. 138(9):697–703. Am J Epidemiol 189(9):697–703
31. Marsden PV, Podolny J (1990) Dynamic analysis of network diffusion processes. In: Weesie J, Flap H (eds) Social networks through time. ISOR, Utrecht
32. Strang D, Tuma NB (1993) Spatial and temporal heterogeneity in diffusion. Am J Sociol 99:614–639
33. Valente T (1996) Social network thresholds in the diffusion of innovations. Soc Netw 18:69–89
34. Van den Bulte C, Lillien GL (2001) Medical innovation revisited: social contagion versus marketing effort. Am J Sociol 106:1409–1435
35. Rogers EM, Kincaid DL (1981) Communication Networks: A new paradigm for research. Free Press, New York
36. Dozier DM (1977) Communication networks and the role of thresholds in the adoption of innovations. Ph D Thesis, Stanford University
37. Montgomery MR, Chung W (1999) Social networks and the diffusion of fertility control in the Republic of Korea. In: Leete R (ed) Dynamics of values in fertility change. Oxford University Press, Oxford
38. Kohler HP (1997) Learning in social networks and contraceptive choice. Demography 34:369–383
39. Casterline J (2001) Diffusion processes and fertility transition: selected perspectives. National Academy Press, Washington
40. Neiagus A et al (2001) HIV risk networks and HIV transmission among injecting drug users. Eval Program Plan 24:221–226
41. Marsden PV (1987) Core discussion networks of Americans. Annu Rev Sociol 52:122–131
42. Marsden PV (1990) Network data and measurement. Annu Rev Sociol 16:435–463
43. Valente T, Saba W (1998) Mass media and interpersonal influence in a reproductive health communication campaign in Bolivia. Commun Res 25:96–124
44. Valente T, Vlahov D (2001) Selective risk taking among needle exchange participants in Baltimore: implications for supplemental interventions. Am J Public Health 91:406–411
45. Urberg KA, Degirmencioglu SM, Pilgrim C (1997) Close friend and group influence on adolescent cigarette smoking and alcohol use. Dev Psychol 33:834–844
46. Becker MH (1970) Sociometric location and innovativeness: reformulation and extension of the diffusion model. Am Sociol Rev 35:267–282
47. Scott J (2000) Network analysis: a handbook, 2nd edn. Sage, Newbury Park
48. Wasserman S, Faust K (1994) Social networks analysis: methods and applications. Cambridge University Press, Cambridge
49. Entwisle B et al (1996) Community and contraceptive choice in rural Thailand: a case study of Nang Rong. Demography 33:1–11
50. Alexander C et al (2001) Peers, schools, and adolescent cigarette smoking. J Adolesc Health 29(1):22–30
51. Bearman PS, Jones J, Udry JR (2000) National longitudinal study of adolescent health: research design. http://www.cpc.unc.edu/projects/addhealth/data
52. Boulay M, Valente TW (2005) Dynamic sources of information and dissonance in the discussion networks of women in rural Nepal. J Health Commun 10:519–536
53. Tuma NB, Hannan MT (1984) Social dynamics: models and methods. Academic Press, New York
54. Allison PD (1984) Event history analysis. Sage, Newberry Park
55. Bartholomew DJ (1982) Stochastic models for social processes. Wiley, New York

56. Teachman JD, Hayward MD (1993) Interpreting hazard rate models. Sociol Methods Res 21(3):340–371
57. Eliason SR (1979) Maximum likelihood estimation: logic and practice. Sage, Newburg Park
58. Valente T (2005) Models and methods for innovation diffusion. In: Carrington PJ, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, Cambridge
59. Myers DJ (2000) The diffusion of collective violence: infectiousness, susceptibility, and mass media networks. Am J Sociol 106:173–208
60. Kohler H-P, Behrman JR, Watkins SC (2001) The density of social networks and fertility decisions: evidence from South Nyanza District, Kenya. Demography 38:43–58
61. Montgomery MR et al (1999) Social networks and contraceptive dynamics in Southern Ghana. Paper presented at the annual meeting of the Population Association of America. Washington DC
62. White K, Watkins SC (2000) Accuracy, stability and reciprocity in informal conversational networks in rural Kenya. Soc Netw 22:337–356
63. Carley KM (2001) Learning and using new ideas: a socio-cognitive approach. In: Casterline J (ed) Diffusion processes and fertility transition: selected perspectives. National Academy Press, Washington
64. Epstein JM (2006) Generative social science. Princeton University Press, Princeton
65. Valente TW, Fosados R (2006) Diffusion of innovations and network segmentation: the part played by people in the promotion of health. J Sex Transm Dis 33:S23-S31
66. Robins G, Pattison P, Kalish Y, Lusher D (2007) An introduction to exponential random graph (p*) models for social networks. Soc Netw 29:173–191

Social Networks, Exponential Random Graph (p^*) Models for

Garry Robins
School of Behavioural Science, University of Melbourne, Melbourne, Australia

Article Outline

Glossary
Definition of the Subject
Introduction
Notation and Terminology
Dependence Hypotheses
Bernoulli Random Graph (Erdös–Rényi) Models
Dyadic Independence Models
Markov Random Graphs
Simulation and Model Degeneracy
Social Circuit Dependence:
Partial Conditional Dependence Hypotheses
Social Circuit Specifications
Estimation
Goodness of Fit and Comparisons
with Markov Models
Further Extensions and Future Directions
Bibliography

Exponential random graph models, also known as p^* models, constitute a family of statistical models for social networks. The importance of this modeling framework lies in its capacity to represent social structural effects commonly observed in many human social networks, including general degree-based effects as well as reciprocity and transitivity, and at the node-level, homophily and attribute-based activity and popularity effects. The models can be derived from explicit hypotheses about dependencies among network ties. They are parametrized in terms of the prevalence of small subgraphs (configurations) in the network and can be interpreted as describing the combinations of local social processes from which a given network emerges. The models are estimable from data and readily simulated. Versions of the models have been proposed for univariate and multivariate networks, valued networks, bipartite graphs and for longitudinal network data. Nodal attribute data can be incorporated in social selection models, and through an analogous framework for social influence models.

The modeling approach was first proposed in the statistical literature in the mid-1980s, building on previous work in the spatial statistics and statistical mechanics literature. In the 1990s, the models were picked up and extended by the social networks research community. In this century, with the development of effective estimation and simulation procedures, there has been a growing understanding of certain inadequacies in the original form of the models. Recently developed specifications for these models have shown a substantial improvement in fitting real social network data, to the point where for many network data sets a large number of graph features can be successfully reproduced by the fitted models.

Glossary

Alternating independent-2-paths A parameter (and statistic) in new specification models; a particular combination of *k-independent-2-path* counts into the one statistic.

Alternating k-stars A Markov parameter (and statistic) in the new specification models; a particular combination of Markov *k*-star counts into the one statistic; equivalent to *geometrically weighted degree counts*; useful for modeling the degree distribution.

Alternating k-triangles A parameter (and statistic) in the new specification models; a particular combination of *k-triangle* counts into the one statistic; equivalent to *weighted shared partners*.

Cyclic triad A Markov graph configuration: in a directed network, ties *ij, jk* and *ki* are observed among actors *i*, *j*, and *k*.

Degeneracy (or near-degeneracy) When a model implies that very few distinct graphs are probable, often only empty or complete graphs; degenerate models cannot be good models for social network data.

Dependence assumption Theoretical assumption about dependencies among possible network ties; determines the type of parameters in the model.

Dyad independence Assumes that dyads are independent of one another; the model includes edge and reciprocity parameters, and possibly also node or dyad attributes.

Dyad-wise shared partners A parameter (and statistic) in the higher order models; equivalent to *alternating independent 2-paths*.

Edge-wise shared partner distribution Distribution of the number of dyads who are themselves related and who have a fixed number of shared partners.

Edge-wise shared partners A parameter (and statistic) in the higher order models; equivalent to *alternating k-triangles*.

Geometrically weighted degree counts A statistic (and parameter) in the new specification models: a sum of degree counts with geometrically decreasing weights; equivalent to *alternating k-stars*.

Homogeneity assumption Assumption about which parameters to equate, to make a model identifiable.

***k*-independent-2-paths** Configurations in the higher order models; equivalent to *k*-triangles but without the base.

***k*-in-star** A Markov graph configuration: in a directed graph, *k* arcs are directed to the one actor.

***k*-out-star** A Markov graph configuration: in a directed graph, *k* arcs are expressed by the one actor.

***k*-triangle** A configuration in higher order models; in a non-directed graph, the combination of *k* triangles, each sharing the one edge (the base of the *k*-triangle).

***k*-star** A Markov graph configuration: in a non-directed graph, *k* edges are expressed by the one actor.

Markov dependence assumption Introduced by Frank and Strauss [9], proposes that, conditional on the rest of the graph, two possible ties are independent of each other unless they share an actor.

Mixed-star A Markov graph configuration: a two path in a directed graph.

Monte Carlo Markov chain maximum likelihood estimation (MCMCMLE) Method of estimation based on computer simulation; more principled than pseudolikelihood.

Network configuration A small subgraph that may be observed in the data and that is represented by parameters in the model: e. g. reciprocated ties, triangles.

Parameters Relate to specific network configurations that may be observed in the graph; a large positive parameter is interpreted as the presence of more of the configurations than might be expected from chance (given the other effects in the model); a large negative parameter signifies the relative absence of the configuration.

Partial dependence assumption Assumption for dependencies among possible ties created by the presence of other ties; permits models with higher order configurations than Markov configurations.

Pseudo-likelihood estimation An approximate method of estimation using logistic regression; does not produce reliable standard errors.

p_1 Model An early dyad independence model, including popularity and expansiveness effects.

p_2 Model Elaboration of p_1 model, where popularity and expansiveness effects are random, and independent variables may be used to predict ties.

Simple random graphs, Bernoulli graphs, Erdös–Rényi graphs Assume that edges are independent of one another and are observed with a given probability.

Social circuit dependence Two possible ties are conditionally dependent when, if observed, they would create a 4-cycle.

Transitive triad A Markov graph configuration: in a directed network, ties *ij, jk* and *ik* are observed among actors *i*, *j*, and *k*.

Triangle A Markov graph configuration: in a non-directed network, a clique of three actors, ties *ij, jk* and *ik* are observed among actors *i*, *j*, and *k*.

Definition of the Subject

Exponential random graph models, also known as p^* models, constitute a family of statistical models for social networks. These models take the form of a probability distribution of graphs:

$$\Pr(\mathbf{X} = \mathbf{x}) = (1/\kappa)\exp\{\boldsymbol{\eta}'\mathbf{g}\},$$

for a set of tie indicator variables $\mathbf{X}$ on a network of fixed node size n, where $\mathbf{x}$ is a realization, with a parameter vector $\boldsymbol{\eta}$ and a vector of network statistics $\mathbf{g}$. Each value of the parameter vector corresponds to a probability distribution on the set of all graphs with n nodes.

Introduction

As noted in [34], statistical approaches to social networks have quite a long history, stretching at least back to the work of Moreno in the 1930s [24]. Yet, that history is rather sparsely scattered across various literatures and different eras and, as a result, even today we see that older techniques and approaches are reinvented and repackaged as new. The *statistical modeling* of a social network – as distinct from the use of statistically-based approaches to understanding particular properties of a social network – has a shorter and somewhat more coherent pedigree.

Criteria for a successful statistical model of a social network have been proposed by [34]: it should be possible to estimate the parameters of the model from data in a principled way, and the fitted model should be a good representation of that data; the model should provide theoretically plausible interpretations about the type of effects that might have produced the network; and using the estimated model parameters it should be possible to draw inferences about competing explanations for the data. Most, perhaps all, models for social networks fall short of completely meeting these requirements. Some models are "thought experiments", intended to illuminate but with an uncertain link to data. Some models are estimable from data but still cannot adequately represent important features of the network. Models may imply interpretations that are just not theoretically plausible in terms of social science; and some models cannot include different effects simultaneously in order to test one against the other. Of course, by definition all models are imperfect (else, they are not "models"). So these criteria direct the aim in model development, even under the knowledge that we cannot always hit the target exactly.

A major insight in the statistical modeling of social network structure has been that structural effects can be detected as some form of deviation from what would be expected as "randomness". This was in fact the original proposal of Moreno and colleagues [24] who were the first to compare observed network data to what would be expected from null distributions. More explicitly, Rapoport's biased net theory [32,33] proposed certain structural "biases" away from random tie formation, including the important notions of transitivity (later to be described as clustering), the propensity for human social networks to exhibit a high proportion of triangulated ties.

Of course, in the graph theory literature the best-known model for randomness is the *simple random graph model*, often known as the Erdös–Rényi graph [7] or uniform Bernoulli graph distribution. This model proposes that for a given number of nodes network ties are observed between pairs of nodes independently and with a fixed probability p. Properties of this model have been examined extensively, utilizing its analytic tractability. It has also been used as a null model for various sampling distributions (e. g. see the summary in [8].) For an observed network, p can be readily estimated (the maximum likelihood estimate is simply the density of the network), but unfortunately this model is not a good representation of almost any human social network. There are no structural effects here, only randomness.

For directed networks, [15] extended the Bernoulli model in the early 1980s by parametrizing structural effects for reciprocity and for differential node-level activity (out-degree) and popularity (in-degree). Possible network ties within dyads were dependent on one another, but were independent between dyads. This dyad-independence assumption permitted the model to be estimated as a straight-forward loglinear model. Holland and Leinhardt called this model p_1, the subscripted '1' implying a program of further research, with progressively enlarged dependence assumptions (within arcs, within dyads, within triads, etc.). However, Holland and Leinhardt were uncertain how to progress beyond dyads. The problem was that standard statistical estimation required some level of independence and triads, unlike dyads, overlapped. (A subsequent, more sophisticated extension, the p_2 model, also has dyadic independence at its heart, but conditional on random node-level effects, so that the random effects indirectly introduce dependencies that may extend beyond dyads [53].)

Ove Frank and his colleague, David Strauss, provided the crucial insight: in the world of complex networks, it was dependence that mattered, and assumptions involving traditional statistical independence, although helpful for fitting models, were likely to be inadequate empirically [9]. In contrast, their approach centered on *conditional* independence, whereby contingencies among network ties may transmit across the network in the sense that the presence of one network tie may affect the presence of any other, but most pairs of ties were in some sense "remote", so that the contingencies affecting a given network tie had to be transmitted through "neighboring" ties. So once one had a concept of "neighboring ties", that is, some notion of which possible ties were conditionally independent after taking into account other observed network ties, the form of a model could be specified. Approaches from spatial statistics and statistical mechanics, with notions of dependence within neighborhoods (broadly defined) could be translated in this way into social network models. Frank and Strauss proposed Markov random graph mod-

els, based on an explicit argument about conditional independence among networks ties, that could be regarded as theoretically plausible for the first time. These models incorporated reciprocity, degree-based effects and various forms of triangulation or clustering.

In the 1990s, Markov random graph models became the accepted form of exponential random graph models and they have only recently been superseded. They could be estimated through the rough and ready procedure of pseudo-likelihood [52], a not particularly sophisticated approach but enough to get the field started empirically. Wasserman and Pattison popularized these models among social networks researchers as p^* models [55] and added further generalizations, with extensions to multivariate [29] and valued networks [36]. Nodal attributes were introduced in social selection [37] and social influence models [38].

In the last decade, especially with more principled methods of estimation and simulation, it has been shown that homogeneous Markov random graph models face considerable, often insurmountable, difficulties in dealing with real network data. Data with inhomogeneities – for instance, very high degree nodes, or dense regions of multiple triangulation, both of which are not uncommon in real social networks – typically result in *degeneracy* for Markov random graph models, where no set of parameter estimates can adequately represent the data. Newer specifications have been proposed by Snijders and colleagues [48] that help considerably with these issues of degeneracy, and have the capacity to enhance our ability to model small-scale to medium-sized social networks, in some cases remarkably well. These advances are based on new theorizations of dependence among network ties, echoing the original project of Holland and Leinhardt with a notion of increasingly extended dependence assumptions.

This article begins with a presentation of notation and terminology and then discusses how dependence hypotheses lead to the general form of the model. Specific examples of dependence hypotheses, and the resulting models, are then presented: Bernoulli graph distributions; dyadic independence models; and Markov random graph models. Methods of simulation are briefly introduced before a discussion of degeneracy issues, particularly as they apply to Markov random graph models. Recently proposed dependence hypotheses are introduced, including the so-called "social circuit" dependence, leading to additional specifications for exponential random graph models that substantially improve model performance. Estimation and goodness of fit approaches are discussed before a short empirical example is presented. Extensions are briefly described followed by some concluding remarks about future directions.

Notation and Terminology

A network comprises a set of relational ties between pairs of individual *actors* (be they people or other social entities). A network can be represented as a graph G with node set $N = \{1, 2, \ldots, n\}$ representing the individuals and edge set E representing the relational ties. For the statistical models of this article, relational ties are construed as a set of binary random variables X_{ij} such that $X_{ij} = 1$ if a tie is observed from node i to node j, for $i \neq j$, and $X_{ij} = 0$, otherwise. For a nondirected network X_{ij} and X_{ji} are equivalent, whereas for a directed network the two variables are distinct. X_{ii} is undefined (or may be considered as a structural zero). A nondirected tie is an *edge* and a directed tie an *arc*. We specify x_{ij} as the observed value of the variable X_{ij} and we let $\mathbf{X}$ be the matrix of all variables with $\mathbf{x}$ being a realization (where the diagonals of the matrices are forced to be zero). $\mathbf{X}$ and $\mathbf{x}$ are necessarily symmetric for nondirected networks, but not so for directed networks. More generally, $\mathbf{x}$ may be valued but in this article we restrict attention to binary ties.

There are natural extensions to this basic notation, depending on the data structures. For instance, suppose the data is in the form of a bipartite network, with two distinct sets of nodes and ties between nodes of different types but not between nodes of the same type (e. g. such a data structure can represent people's membership of clubs). Then $\mathbf{X}$ may comprise a set of variables X_{pa} indicating the presence or absence of a tie between a node p of the first type and a node a of the second type. The data may involve *multiple* or *multivariate networks*, with r different types of relations among the one set of nodes. In that case $\mathbf{X}$ may be thought of as a three-way array of variables X_{ijr} such that $X_{ijr} = 1$ indicates that presence of a tie of type r from node i to node j. An analogous three-way array may represent network data collected for the same set of nodes at different time points such that $X_{ijt} = 1$ indicates that presence of a tie at time t from node i to node j.

Nodes themselves can have certain properties (*attributes*) that may be measured as binary, categorical or continuous variables. We denote an attribute variable with the vector $\mathbf{Y}$, where $Y_i = y_i$ indicates that for attribute Y node i has a value y_i.

Dependence Hypotheses

Following work in spatial statistics [1], Frank and Strauss introduced the notion of a *dependence graph* into social network modeling [9] in order to represent possible de-

pendencies among network variables X_{ij}. The nodes of the dependence graph are the network variables X_{ij} and an edge between two nodes indicates dependence between the respective variables, even when the observed values for all remaining variables are known. These edges specify a *neighborhood* relationship between pairs of variables and the cliques of the dependence graph can be thought of as *local social neighborhoods* for the set of tie variables [27]. Conversely, the absence of an edge in the dependence graph indicates that two network variables are conditionally independent, and are not neighbors, so that their interaction need not be taken into account in a model based on local social neighborhoods. (A more technical discussion on dependence graphs for social network models can be found in [35]).

The previous paragraph describes one important result from the Hammersley–Clifford theorem [1], that the form of a probability distribution for a set of interacting variables relates solely to the neighborhood structure (or clique structure) of the dependence graph (with a single network variable also taken as a clique). For a given node set N, once a dependence hypothesis is specified and the dependence graph is defined, it follows *necessarily* from the Hammersley–Clifford theorem that:

$$\Pr(\mathbf{X} = \mathbf{x}) = (1/\kappa(\boldsymbol{\eta})) \exp\{\Sigma_A \eta_A g_A(\mathbf{x})\} , \qquad (1)$$

where:

(i) the summation is over all neighborhoods A;
(ii) η_A is a parameter corresponding to the neighborhood A (and there is no non-zero parameter for any set of ties that do not constitute a neighborhood);
(iii) $g_A(\mathbf{x}) = \prod_{x_{ij} \in A} x_{ij}$ is the *network statistic* corresponding to neighborhood A and indicates whether all the ties in A are observed in the network $\mathbf{x}$;
(iv) κ is a normalizing quantity, that is a function of the parameter vector $\boldsymbol{\eta}$, to ensure that (1) is a proper probability distribution.

All exponential random graph models take this general form, describing a probability distribution of graphs on n nodes. Neighborhoods A are subsets of possible ties, so they represent possible subgraphs that may or may not be observed in the network $\mathbf{x}$. There is one, and no more than one, parameter for each distinct neighborhood (although one neighborhood may be a subgraph of another and so both may have separate parameters).

The form of the model in (1) is too general to be identifiable and some homogeneity constraint needs to be imposed on what is otherwise a large number of neighborhoods A (and hence a large number of parameters). The original proposal [9] was that parameters should be equated across subsets of ties that were isomorphic to each other (that is, subgraphs that are indistinguishable once node labels are removed). With some further constraints (noted below), this assumption can produce identifiable models that can be fitted to data. (Other possible ways to constrain the number of parameters are also described below). Following the terminology first used by Moreno [24], we term these isomorphic neighborhoods as *network configurations* (although some of the network literature of the last decade has adopted the term *motif*). Counts of configurations become the sufficient statistics of the model.

Examination of the parameters for the various configurations of a fitted model can provide insight into the social processes that may underpin the network. The strength and direction of any particular parameter value will affect how frequently the corresponding configuration is observed. If the parameter is large and positive, we expect to observe the corresponding configuration to occur more frequently than if the parameter value were zero [40].

Obviously, good dependence hypotheses are necessary for this approach to work, as they crucially shape the nature of the model by defining the possible configurations.

Bernoulli Random Graph (Erdös–Rényi) Models

Suppose we hypothesize that there are no dependencies within the network, so that the dependence graph has no edges and the only neighborhoods are single tie variables X_{ij}. Then (1) becomes:

$$\Pr(\mathbf{X} = \mathbf{x}) = (1/\kappa) \exp(\Sigma_{i,j} \eta_{ij} x_{ij}) .$$

Constraining parameters to be equal for isomorphic neighborhoods here implies equating all the parameters $\eta_{ij} = \theta$, as there is only one type of neighborhood. Then we have:

$$\Pr(\mathbf{X} = \mathbf{x}) = (1/\kappa) \exp(\Sigma_{i,j} \theta x_{ij}) = (1/\kappa) \exp(\theta L(\mathbf{x})) ,$$

where $L(\mathbf{x})$ is the number of edges (arcs) in the network $\mathbf{x}$ and θ is a *density* or *edge* parameter (sometimes called a *choice* parameter in older literature). This graph distribution is equivalent to that of a distribution of simple random graphs where the probability of a tie between a pair of edges is $\exp\theta/(1 + \exp\theta)$.

The dependence assumption here is unrealistic and this model will fit few, if any, real networks, but it is often a useful null or baseline model.

Dyadic Independence Models

For directed networks, a dyadic independence hypothesis implies edges in the dependence graph between variables

X_{ij} and X_{ji}. Neighborhoods then take the form of single edges and dyadic pairs $\{X_{ij}, X_{ji}\}$. Applying constraints in the form of the one density parameter θ to all single edge neighborhoods and the one *reciprocity parameter* ρ to all dyadic neighborhoods results in the following two parameter model:

$$\Pr(\mathbf{X} = \mathbf{x}) = (1/\kappa)\exp(\theta L(\mathbf{x}) + \rho M(\mathbf{x}))\,,$$

where $L(\mathbf{x})$ is the number of arcs and $M(\mathbf{x})$ the number of mutual dyads (i. e. where $X_{ij} = X_{ji} = 1$) in the network $\mathbf{x}$. This model has the capacity to model reciprocity effects in directed networks. Slightly less sweeping homogeneity constraints result in the p_1 model [15].

The interpretation of a single parameter in an exponential random graph model is relative to other effects in the model, so that in this case, the presence of a large and positive ρ parameter is relative to the density effect. This is an advantage: in this model, for instance, we can infer that there is substantial reciprocity *given the density of the graph*. It is of little use to talk about a large reciprocity effect in absolute terms, because the number of arcs in the graph (i. e. the density) establishes the preconditions for reciprocity. In other words, graphs of very low density have very little opportunity to demonstrate reciprocity anyway, and the absence of mutual dyads in a low density graph may simply be a feature of the low density. Similarly the presence of many mutual dyads in a high density graph may simply be explained by the high density, without the need to infer a reciprocity effect. A large positive ρ, on the other hand, shows that the number of arcs in a graph are arranged in sufficiently many mutual dyads to suggest that a separate and substantial reciprocity process is required to explain the structure of this network, over and above any density effect.

Once again, the dyadic independence assumption does not do a good job of reproducing real networks, because the model has no capacity for triangulation. The model may be useful as a baseline model.

Markov Random Graphs

Frank and Strauss [9] proposed Markov dependence, by postulating that a possible tie from i to j is assumed to be contingent on any other possible tie involving i or j, even if all other ties in the network are fixed. Markov dependence implies that two possible network ties are conditionally independent unless they share a common actor. They showed that this assumption resulted in models with configurations for nondirected graphs of single edges, star-like structures and triangles (or 3-cycles). For directed graph models, configurations include directed counterparts of these as well as reciprocity (see [55] for a fuller description of directed graph parameters). Frank and Strauss termed these *Markov random graph models*.

Once homogeneity is imposed across isomorphic neighborhoods, we obtain configurations (and related parameters) for nondirected networks as depicted in Fig. 1. The resulting model is:

$$\Pr(\mathbf{X} = \mathbf{x}) = (1/\kappa)\exp(\theta L(\mathbf{x}) + \Sigma_{r=2,n-1}\sigma_r S_r(\mathbf{x}) + \tau T(\mathbf{x}))\,, \quad (2)$$

where:

(i) θ is a density parameter, and $L(\mathbf{x})$ the number of edges in $\mathbf{x}$, as before;
(ii) σ_r is a parameter for a *star of size r*, and $S_r(\mathbf{x})$ is the number of stars of size r in $\mathbf{x}$;
(iii) a star of size r is a configuration centered on a single node i such that there are r edges emanating from i (note that stars of size larger than r contain many stars of size r, so that this is not simply a partition into nodes of different degrees, although the full set of counts $S_r(\mathbf{x})$ can be converted into the degree distribution, and vice versa);
(iv) τ is a triangle or clustering parameter and $T(\mathbf{x})$ is a count of the number of triangles in $\mathbf{x}$.

This model still has too many star parameters to be identifiable. Note that the expression $\theta L(\mathbf{x}) + \Sigma_{r=2,n-1}\sigma_r S_r(\mathbf{x})$ completely parametrizes the degree distribution (so the τ parameter can be interpreted as the strength of transitivity conditional on the degree distribution of the graph). An identifiable model can be obtained by restricting the number of star parameters to be substantially less than $n - 2$, for instance, to less than four (an alternative approach is described below). Then the model becomes:

$$\Pr(\mathbf{X} = \mathbf{x}) = (1/\kappa)\exp(\theta L(\mathbf{x}) + \sigma_2 S_2(\mathbf{x}) + \sigma_3 S_3(\mathbf{x}) + \tau T(\mathbf{x}))\,. \quad (3)$$

The parameters can be interpreted in ways that represent plausible social processes [39]:

(i) θ is a baseline propensity for social ties to form;
(ii) σ_2 represents a tendency for individuals to seek multiple partners (if it is positive);
(iii) whereas σ_3 (if negative as it often is when this model is fitted to real data) represents a ceiling effect against having too many partners; so, together the balance between σ_2 and σ_3 represents the benefits of multiple

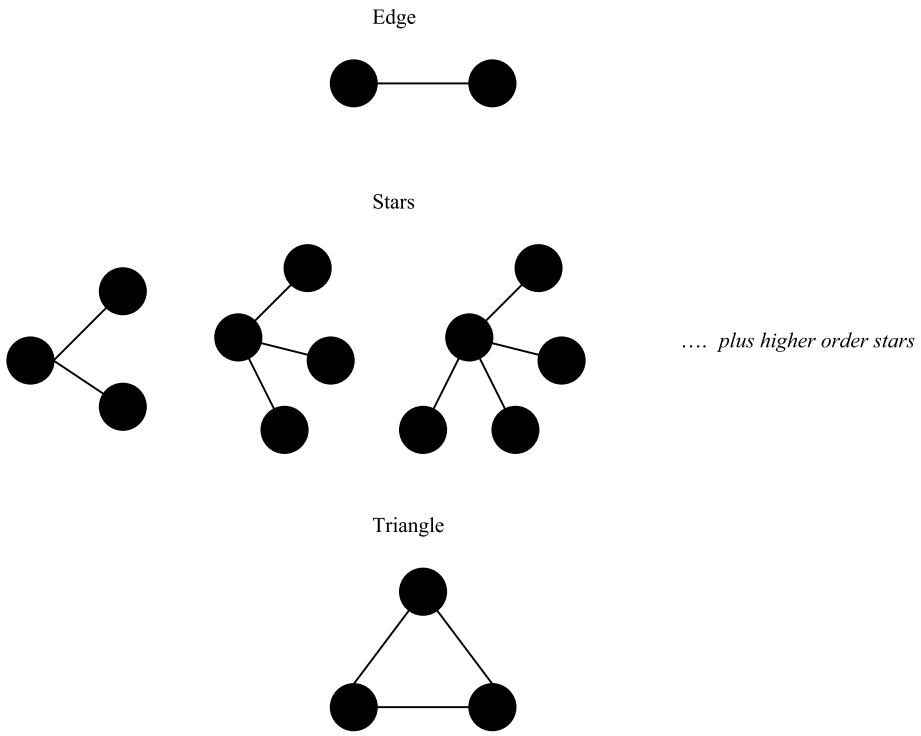

Social Networks, Exponential Random Graph (p^*) Models for, Figure 1
Configurations for a nondirected Markov random graph model

contacts against the costs of maintaining too many contacts;

(iv) while τ represents tendencies towards clustering.

An alternative interpretation is to consider the θ, σ_2 and σ_3 parameters as controlling for the first three moments of the degree distribution, so that while model (2) in effect presents a full parametrization, model (3) provides a more parsimonious control over the degree distribution (saturated models of the degree distribution are discussed by [43] and [47], see also [11]). Of course, for any data set it is an empirical question whether the first three moments are sufficient to capture the degree distribution adequately. Below, we describe ways to investigate that question.

Extensions of this basic Markov random graph model include models for multivariate networks [29], for valued networks [36] and for affiliation networks [42] (see also [28]).

By tuning parameter values even the simple four parameter model (3) can represent many different types of networks, including small world networks, networks with long paths, and highly clustered ("caveman") networks [39]. For some parameter values, however, the models became "frozen" into certain unchanging highly structured patterns. This effect is an illustration of *degeneracy* for these models, which was a research issue given important attention in the early years of the 21st century. The upshot of this body of research was that homogeneous Markov random graph models were recognized as frequently inadequate to deal with real social network data.

Simulation and Model Degeneracy

The problems of degeneracy became apparent through a number of simulation studies. Strauss was the first to describe a rather straightforward application of standard statistical simulation techniques (e. g. Metropolis or Metropolis–Hastings algorithms) to simulate Markov random graph models for a fixed set of nodes and given set of parameter values [51]. More recent treatments and some important results are given in [13,14,17,39,45]. These methods provide a principled statistical means to produce a distribution of graphs with probabilities of each graph being observed in the distribution consistent with any particular model in the form of (1). Typically the simulation is run for a large number of iterations and a sample of graphs is extracted and examined to understand typical properties of the graphs in the distribution.

A graph distribution is termed as *near degenerate* [13,14] if it implies only a very few (possibly only one

or two) distinct graphs with substantial non-zero probabilities. For instance, certain parameter values for Markov random graph models place almost all of the probability mass on either the empty or the full graph. These cannot be good models for human social networks. For certain parameter values, there may be two quite separate regions where graphs are most likely (possibly one region of low density and another of high density graphs), with the possibility of a dramatic phase transition from one region to the other (for instance, as parameter values change very slightly). Such phase transitions and other aspects of degeneracy have been studied for a variety of simple Markov random graph models by a number of authors [2,12,13,14,19,26,39,41,45,48].

Issues of degeneracy call into question whether Markov random graph models can adequately represent most network data. It would not matter if some regions of the parameter space produced degenerate models, as long as those regions applicable to empirical data were non-degenerate. Experience shows the opposite applies: for some data, Markov random graph models turn out to be well-behaved, but for many empirical networks they are degenerate. When applied to social networks with a few very high degree nodes, or with some regions of high triangulation, degeneracy frequently occurs for Markov models. Models such as (3) reflect homogeneity constraints, and they appear to have difficulty when the data presents inhomogeneity in the distribution of degrees or triangles. It is not entirely certain whether the problem with the models arises because of the Markov dependence assumption or the assumption of homogeneity. In any event, to deal more effectively with real networks, a different approach to these models seems necessary. To date, attention has been directed to revisiting the dependence assumption.

Snidjers and colleagues [48] drew two conclusions from the tendency for Markov random graph models to be degenerate when fitted to real networks with a high level of clustering: (1) the Markov dependence assumption may be too restrictive; (2) the representation of the social phenomenon of transitivity by the total number of triangles might be too simplistic. They proposed specifications that drew on both these possibilities. Before we introduce these new specifications, however, we turn to higher order network dependencies.

Social Circuit Dependence: Partial Conditional Dependence Hypotheses

There are several ways in which network dependence assumptions could be extended beyond Markov dependence [27]. One approach is of *partial conditional dependence*, whereby the presence of certain network ties *created* dependence among other possible network ties. This assumption permits the emergence of dependence structures as ties come into and go out of existence. This type of dependence can be incorporated into a version of the Hammersley–Clifford theorem and hence into exponential random graph models [27].

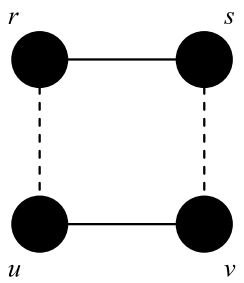

Social Networks, Exponential Random Graph (p^*) Models for, Figure 2
Social circuit dependence (*Broken lines* represent possible edges; *full lines* represent observed edges X_{ru} is conditionally dependent on X_{sv})

A particular hypothesis about this type of emergent dependence was used by [48] and described by [41] as *social circuit dependence*. Social circuit dependence is defined as two possible network ties being conditionally dependent if their observation would lead to a 4-cycle. This type of dependence is depicted in Fig. 2. Here $X_{rs} = X_{uv} = 1$, that is, in the network data there are observed edges between nodes r and s, and between nodes u and v. The presence of these edges leads to conditional dependence between variables X_{ru} and X_{sv} because if edges were also observed between r and u, and between s and v, a 4-cycle would result in the graph.

Interpretation of this hypothesis is discussed by [41]. Generally we would expect two distinct possible edges (r, u) and (s, v) to be conditionally independent. If, however, person r knows person s, and person u knows person v, then the presence of a tie between r and u can make the presence of a tie between s and v more likely, that is, they are conditionally dependent. It is simple to think of real social circumstances where this happens: for instance, in families the presence of a friendship between two children increases the chances of the two mothers coming to know one another; or in a business, cooperation between two bosses may lead to their employees working together; or in research, two postdoctoral researchers might discuss issues with each other because their academic mentors have been collaborators. In all of these cases, what makes the dependence come into effect is the presence of ties that

constitute part of the 4-cycle. Without those ties, the dependence does not arise. So, the mothers are the mothers of the two particular children, not the mothers of entirely different children; the employees are employed by those particular bosses; and the two post-docs are supported by those two mentors.

Social Circuit Specifications

Snijders and colleagues [48] proposed three new statistics for exponential random graph models for nondirected networks: *alternating k-stars*, *alternating k-triangles*, and *alternating independent two-paths*. To derive these statistics they use: (1) social circuit dependence *in addition to* Markov dependence; and (2) a non-linear functional form combining various configurations into the one statistic.

Alternating *k*-Stars

The alternating k-star parameter uses the second idea alone and does not extend dependence beyond the class of Markov random graph models. Rather than adopt the usual practice of limiting the number of higher order stars – as for example in model (3) where the star parameters are deliberately limited to no more than 3-stars – Snijders and colleagues returned to model (2), keeping all σ_r in the model but imposing constraints among these parameter values. Specifically, the *alternating k-star assumption* proposes that for all $k \geq 2, \sigma_{(k+1)} = -\sigma_k/\lambda$ for some λ greater than 1. Then in (1) there is one parameter σ for all star effects, with an associated statistic:

$$u = \sum_{k=2}^{n-1} (-1)^k \frac{S_k}{\lambda^{k-2}} , \tag{4}$$

where the parameter σ is referred to as the *alternating k-star parameter*. As λ is greater than 1, the direct impact of higher order stars (that is, very high degree nodes) is reduced for higher k. Of course, as noted above, high degree nodes still have substantial effect in the model as they produce many 2- and 3-stars, and so on. The alternating sign helps balance these overlapping star counts for high degree nodes. In [48] λ is set at 2 but [17] show how to estimate an optimal value of λ (see also [16]). The alternating k-star parameter is equivalent to a *geometrically weighted degree parameter* that explicitly models the degree distribution but puts more weight on the numbers of nodes with lower degrees, with weights decreasing geometrically as the degrees increase [16].

The interpretation of the parameter is as follows [41, 48]. If the alternating k-star parameter is positive, then highly probable networks are likely to contain some higher degree nodes. A positive alternating k-star parameter (together with a negative density parameter) implies graphs that exhibit preference for connections between a larger number of low degree nodes and a smaller number of higher degree nodes, akin to a core-periphery structure.

Alternating *k*-Triangles

The alternating k-triangle assumption incorporates both social circuit and Markov dependence. Snijders and colleagues show that with both these dependence assumptions operating simultaneously, configurations more complex than simple triangles are possible [48]. In particular, they introduced the notion of a *k-triangle*, a combination of k individual triangles that all share one edge (the common *base* of the k triangles), as represented in Fig. 3.

Let T_k be the count of k-triangles in a graph. Then the *alternating k-triangles assumption* combines these counts into the one statistic in an analogous way as for the alternating k-stars, that is:

$$t = 3T_1 + \sum_{k=1}^{n-3} (-1)^k \frac{T_{k+1}}{\lambda^k} , \tag{5}$$

where the factor of 3 for T_1 is due to symmetry considerations in a non-directed triangle [48].

This is the *alternating k-triangle statistic* with an associated *alternating k-triangle parameter* τ. A positive k-triangle parameter indicates triangulation in the network but also tendencies for triangles themselves group together in larger higher order "clumps". A positive alternating k-triangle effect combined with a negative alternating k-star effect can produce a web of multiple smaller regions of triangulation [41,48].

[16] shows that the alternating k-triangle parameter is equivalent to the *edge-wise shared partner* (or ESP) parameter that models the distribution of shared partners of tied actors, but with weights decreasing geometrically as the number of shared partners increase.

Alternating *k*-Two-Paths

Snijders and colleagues also proposed a parameter that represents a lower order configuration for a k-triangle, namely a *k-two-path* which is a k-triangle without the base [48]. These configurations, represented in Fig. 4, describe the number of distinct two-paths between a pair of nodes. The motivation was to provide a parameter that, in conjunction with k-triangles, would distinguish between tendencies to form edges at the base, or at the sides of a k-triangle.

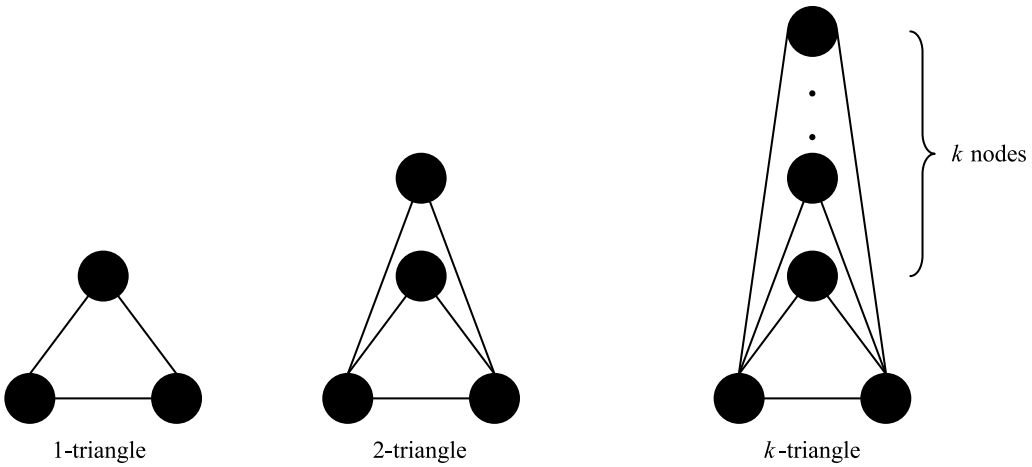

Social Networks, Exponential Random Graph (p^*) Models for, Figure 3
Various *k*-triangles

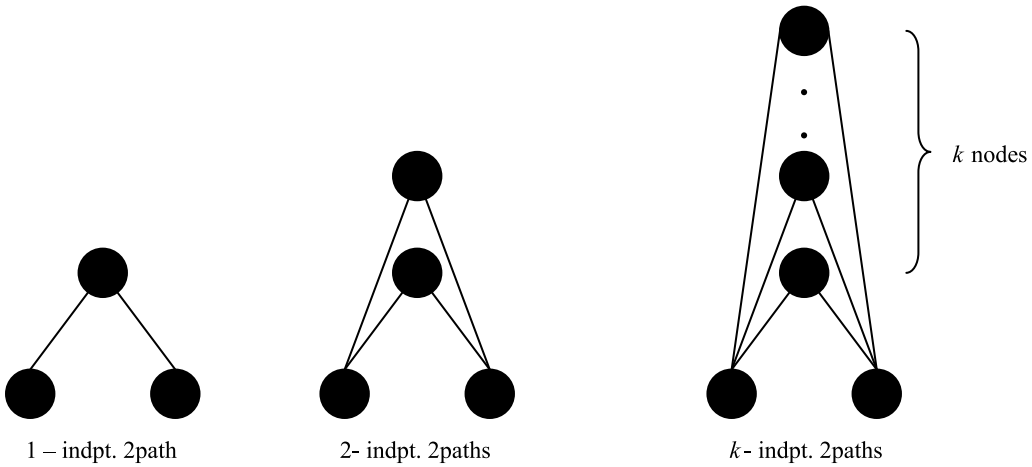

Social Networks, Exponential Random Graph (p^*) Models for, Figure 4
Various alternating *k*-independent 2path configurations

Again the counts of these configurations are combined into the one alternating statistic in an analogous way as in the previous two cases, to produce an alternating 2-path statistic and parameter for the model. [16] shows that this new parameter is equivalent to the *dyad-wise shared partner* (or DSP) parameter that models the distribution of shared partners of actors who may or may not be tied, but with weights decreasing geometrically as the number of shared partners increase.

The parameter can be interpreted as representing localized multiple connectivity between nodes. When this parameter is negative, together with a positive alternating *k*-triangle parameter, there is a tendency against 4-cycles in the network, unless those cycles include triangles (alternatively, the presence of many 2-paths between nodes is related to the formation of triangles).

Specifications for Directed Graphs

Counterpart statistics and parameters for directed graphs have also been proposed. We do not discuss these here but refer readers to [48].

Estimation

Pseudo-likelihood estimation [52], based on logistic regression techniques, was commonly used until more principled methods became available in recent years. Preferred methods of estimation involve simulation proce-

dures (Monte Carlo Markov chain maximum likelihood estimation – MCMCMLE) which for these models has been discussed by a number of authors [4,5,6,17,45,47,56]. Software for Monte Carlo estimation techniques has recently become publicly available ([41] provides a review). The observed graph statistics are compared with those expected under a set of provisional parameter values, using a stochastic simulation based on those values. The goal is to find a set of parameter values whereby the observed statistics are equal to the expected statistics. Parameter values are refined until the observed and expected statistics can be equated, when the model is said to converge. If the parameter estimates never converge, the model is degenerate. Standard errors can also be estimated.

Goodness of Fit and Comparisons with Markov Models

Simulation also provides an innovative means to assess how well the model can represent the data. By simulating the model from the parameter estimates, and extracting a sample of graphs it is possible to examine any graph statistic of interest (whether there is an associated parameter in the model or not). Any graph statistic for the observed graph can be compared to the distribution of such graph statistics from the simulated graphs. If the observed graph statistic is not extreme in the distribution of simulated statistics, then it is plausible that that particular feature of the observed graph could have come from the distribution of graphs implied by the model. In this way, it is possible to check which graph features the model successfully captures. This procedure is described in detail by [18] (see also [41]).

[11] provides a compelling example of how this procedure may be used to make decisions about model selection. On the basis of applying the new parametrization to a large school-based network of over 1000 nodes, [11] concludes that for this data Markov models were degenerate, but the newer parametrization had similar underlying interpretations, avoided degeneracy and was empirically validated. Other comparisons with Markov models demonstrate substantial improvements for the new parametrization in overcoming degeneracy [41].

A Simple Empirical Example

Figure 5 presents an empirical network of social relationships among 16 actors, in this case families from medieval Florence [25].

Markov models do not converge for this network. Table 1 provides parameter estimates and standard errors for models involving the social circuit parameters. The third column in the table provides a convergence statistic which indicates good convergence if it has an absolute value less than 0.1 [41]. In this case, the model exhibits excellent convergence.

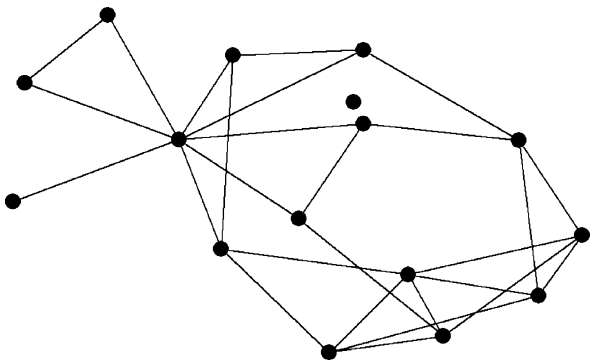

Social Networks, Exponential Random Graph (p^*) Models for, Figure 5
Social relationships among 16 actors

Social Networks, Exponential Random Graph (p^*) Models for, Table 1
Parameter estimates: Florentine families network (NB: $\lambda = 2$)

Parameter	Estimate	Standard error	Convergence
Edge	−0.04	2.09	0.01
Alternating *k*-star	−1.01	0.78	0.02
Alternating *k*-triangle	0.68	0.33	0.02
Alternating *k*-2paths	0.18	0.17	0.03

A parameter may be inferred as important if its estimate in absolute terms is more than twice its standard error. So in this simple model, there is a substantial effect for alternating *k*-triangles. In Fig. 5, regions of triangulation can be seen, consistent with the alternating *k*-triangle effect, perhaps combined with the negative alternating *k*-star effect (although the reliability of that negative effect is not certain, given that the estimate is not more than twice the standard error). In sum, there is good evidence here that this network differs substantially from a simple random network, in particular, that there is substantially more triangulation in this network than would be expected in a simple random graph.

Table 2 shows the goodness of fit analysis for graph features other than those in the model. This analysis was based on extracting 1,000 sampled graphs from a simulation of 1,000,000 (after a burn-in of 100,000), using the parameter estimates in Table 1. The goodness of fit statistic takes the form of a *t*-ratio, calculated as the difference between the graph statistic and the mean from the

Social Networks, Exponential Random Graph (p^*) Models for, Table 2
Goodness of fit analysis: Florentine families network model

Graph feature	Goodness of fit
Number of 2-stars	−0.10
Number of 3-stars	−0.25
Number of triangles	−0.17
Standard deviation, degree distribution	−0.10
Skew, degree distribution	−0.12
Number of isolates	1.06
Clustering coefficient	−0.09

simulated graphs divided by the standard deviation as estimated from the sample. For graph features not in the model, the model is considered to reproduce the data well if the goodness of fit statistic is not extreme (for instance, with t-ratios of less than 2). As can be seen, on the features examined, including counts of various Markov configurations, aspects of the degree distribution, and clustering, the model reproduces the data well. There is one isolate in the data, whereas 71% of simulated graphs had no isolates, hence the rather larger goodness of fit statistic. Nevertheless, with 29% of graphs in the sample having at least one isolate, the data cannot be said to be extreme in this regard.

The clustering coefficient in the Table refers to the proportion of actual to possible triangles and is calculated as thrice the number of triangles as a ratio of the number of two-paths (the factor of three applies because there are three two-paths in any one triangle).

Table 2 does not show features of the geodesic distribution but a separate analysis reveals that the observed geodesic distribution is consistent with those from the simulated graphs: none of the geodesic quartiles of the observed graph is extreme compared to the distribution of quartiles from the simulated graphs. Overall, the model can be said to fit the data well.

Further Extensions and Future Directions

Multiple Networks

It is possible to use the same general framework as (1) when $\mathbf{X}$ is a three-way array representing multiple relations on the one set of nodes. The parametrization, however, becomes more complex. [29] showed how to develop Markov models along these lines, with these models subsequently used in some interesting empirical work, principally in organizations [22,23]. This empirical research, however, utilized pseudo-likelihood estimation and did not take into account possible model degeneracy. See also [20].

Early work on adapting the new specifications proposed by Snijders and colleagues to multiple networks has concentrated on the case of two networks, by using the alternating k-star, -triangle, and -2path parameters for effects within networks and then investigating dyadic associations between networks [57].

Bipartite Networks

Again (1) may be used as a basis for models for bipartite networks. [42] were the first to do this with Markov dependence. [28] proposed additional parameters that derived from partial conditional dependence assumptions. [54] incorporated these and other parameters, combined with an alternating combination of configuration counts analogous to those in the newer specifications, into a model for bipartite networks.

Longitudinal Networks

Perhaps the major approach to modeling network data collected at multiple time points is the actor-oriented models of Snijders and colleagues [44] which utilize a continuous time Markov chain technique. Certain specifications of actor-oriented dynamic models have exponential random graph models as their stationary distribution [44], so the links between the two approaches are quite strong. Accordingly, it is possible to develop longitudinal versions of exponential random graph models to investigate network dynamics. The conceptual difference with the actor-oriented approach is quite subtle: exponential random graph longitudinal models suppose that ties change in response to particular social neighborhoods of other ties, rather than in response to actors seeking to optimize particular structural positions. Early work along these lines, including the evolution of multiple networks, has been reported by [30]. For further discussion on the links between the actor-oriented and tie-based versions of the models, see [46].

Nodal Attributes

One of the important issues that researchers often wish to examine is whether and how nodal attributes relate to network structure. For instance, there are good empirical and theoretical grounds for expecting that people who have social relationships are more likely to be similar to each other (known as *homophily*). Conceptually, two possible processes apply: individuals may develop ties because they are similar (*social selection*) or individuals who are tied may influence each other to be similar (*social influence*).

Of course, both of these processes may occur simultaneously and it is a methodological challenge for a given set of data to disentangle the two to determine the respective strengths of selection and influence effects, respectively. Snijders and colleagues [31,49,50] have recently developed innovative methods for investigating this question for longitudinal data within actor-oriented longitudinal models.

[37] proposed social selection models within the framework of (1), having the form:

$$\Pr(\mathbf{X} = \mathbf{x}|\mathbf{Y} = \mathbf{y}) = (1/\kappa)\exp\{\Sigma_A \eta_A g_A(\mathbf{x}, \mathbf{y})\}\,, \qquad (6)$$

where $\mathbf{Y}$ is a vector of attribute variables (binary, categorical or continuous). Here the configurations A relate to combinations of nodal attributes and network subgraphs, and the network is modeled conditional on a fixed distribution of attributes. The intent of such models may be various: accounting for heterogeneity (that may otherwise cause difficulties in fitting models) as well as assessing selection, homophily, and covariate activity and popularity, while controlling for dependencies naturally occurring in the network setup.

An analogous approach to social influence models and network contagion was proposed by [38] where the patterns of attributes were modeled conditional on a fixed network:

$$\Pr(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}) = (1/\kappa)\exp\{\Sigma_A \eta_A g_A(\mathbf{x}, \mathbf{y})\}\,. \qquad (7)$$

Future Directions

Not only have the new specifications shown a remarkable improvement in the successful modeling of real networks, they have opened up a wide range of network modeling possibilities. The discussion of model extensions mentioned above gives some indication of current work in this area. Of course, the possibility of further improvements in model specification has not been closed. Future directions may include methods to examine very large, community-based networks, not only in terms of data collection (for instance, the development of model-based sampling methods) but also theoretical issues based on possible differences in dependence between very large scale and smaller scale structures. Model-based approaches to missing network data, perhaps within a Bayesian framework, are also in prospect [10,21].

The use of exponential random graph models in empirical research, especially in combination with multiple measures on individuals, opens a wide prospect for network statistical modeling to make a substantial contribution in many applied contexts.

Bibliography

Primary Literature

1. Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. J R Stat Soc Ser B 36:96–127
2. Burda Z, Jurkiewicz J, Kryzwicki A (2004) Network transitivity and matrix models. Phys Rev E69:026106
3. Corander J, Dahmström K, Dahmström P (1998) Maximum likelihood estimation for Markov graphs, Research report. Department of Statistics, Stockholm University, Stockholm
4. Corander J, Dahmström K, Dahmström P (2002) Maximum likelihood estimation for exponential random graph models. In: Hagberg J (ed) Contributions to social network analysis, information theory and other topics in statistics: A festschrift in honour of Ove Frank. Department of Statistics, Stockholm University, Stockholm
5. Crouch B, Wasserman S (1998) Fitting p^*: Monte Carlo maximum likelihood estimation. International Conference on Social Networks, Sitges, Spain, May 28–31
6. Dahmström K, Dahmström P (1993) ML-estimation of the clustering parameter in a Markov graph model, Research report. Department of Statistics, Stockholm University, Stockholm
7. Erdös P, Rényi A (1959) On random graphs. I. Publicationes Mathematicae (Debrecen) 6:290–297
8. Frank O (1981) A Survey of Statistical Methods for Graph Analysis. Sociol Methodol 12:110–155
9. Frank O, Strauss D (1986) Markov graphs. J Am Stat Assoc 81:832–842
10. Gile K, Handcock M (2006) Model-based assessment of the impact of missing data on inference for networks. University of Washington, CSS working paper 66
11. Goodreau S (2007) Advanced in exponential random graph (p^*) models applied to a large social network. Soc Netw 29:231–248
12. Häggström O, Jonasson J (1999) Phase transition in the random triangle model. J Appl Probab 36:1101–1115
13. Handcock M (2002) Statistical models for social networks: Degeneracy and inference. In: Breiger R, Carley K, Pattison P (eds) Dynamic social network modeling and analysis. National Academies Press, Washington, pp 229–240
14. Handcock M (2003) Assessing degeneracy in statistical models of social networks. University of Washington CSS Working Paper no 39
15. Holland P, Leinhardt S (1981) An exponential family of probability distributions for directed graphs (with discussion). J Am Stat Assoc 76:33–65
16. Hunter D (2007) Curved exponential family models for social networks. Soc Netw 29:216–230
17. Hunter D, Handcock M (2006) Inference in curved exponential families for networks. J Comput Graph Stat 15:565–583
18. Hunter D, Goodreau S, Handcock M (2008) Goodness of fit of social network models. J Am Stat Assoc 103:248–258
19. Jonasson J (1999) The random triangle model. J Appl Probab 36:852–867
20. Koehly L, Pattison P (2005) Random graph models for social networks: multiple relations or multiple raters. In: Carrington P, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, New York, pp 162–191

21. Koskinen J (2004) Essays on Bayesian inference for social networks. Department of statistics, Stockholm University, Stockholm
22. Lazega E, Pattison P (1999) Multiplexity, generalized exchange and cooperation in organizations. Soc Netw 21:67–90
23. Lomi A, Pattison P (2006) Manufacturing relations: An empirical study of the organization of production across multiple networks. Organ Sci 17:313–332
24. Moreno J, Jennings H (1938) Statistics of social configurations. Sociometry 1:342–374
25. Padgett JF, Ansell CK (1993) Robust action and the rise of the Medici, 1400-1434. Am J Sociol 98:1259–1319
26. Park J, Newman M (2004) Solution of the 2-star model of a network. Phys Rev E70:066146
27. Pattison P, Robins G (2002) Neighborhood-based models for social networks. Sociol Methodol 32:301–337
28. Pattison P, Robins G (2004) Building models for social space: Neighbourhood based models for social networks and affiliation structures. Mathematiques des science humaines 168:11–29
29. Pattison P, Wasserman S (1999) Logit models and logistic regressions for social networks, II. Multivariate relations. Br J Math Stat Psychol 52:169–194
30. Pattison P, Robins G, Wang P, Snijders TAB, Koskinen J (2006) The co-evolution of multiple networks. Sunbelt XXVI International Social Networks Conference, Vancouver, April 2006
31. Pearson M, Steglich C, Snijders TAB (2006) Homophily and assimilation among sport-active adolescent substance users. Connections 27:51–67
32. Rapoport A (1953) Spread of information through a population with a socio-structural bias: I. assumption of transitivity. Bull Math Biophys 15:523–533
33. Rapoport A (1957) Contributions to the theory of random and biased nets. Bull Math Biophys 19:257–277
34. Robins G, Morris M (2007) Advances in exponential random graph (p^*) models. Soc Netw 29:169–172
35. Robins G, Pattison P (2005) Interdependencies and social processes: Generalized dependence structures. In: Carrington P, Scott J, Wasserman S (eds) Models and Methods in Social Network Analysis. Cambridge University Press, New York, pp 192–214
36. Robins G, Pattison P, Wasserman S (1999) Logit models and logistic regressions for social networks, III. Valued relations. Psychometrika 64:371–394
37. Robins G, Elliott P, Pattison P (2001) Network models for social selection processes. Soc Netw 23:1–30
38. Robins G, Pattison P, Elliott (2001) Network models for social influence processes. Psychometrika 66:161–190
39. Robins G, Pattison P, Woolcock J (2005) Social networks and small worlds. Am J Sociol 110:894–936
40. Robins G, Pattison, Kalish Y, Lusher D (2007) An introduction to exponential random graph (p^*) models for social networks. Soc Netw 29:173–191
41. Robins G, Snijders TAB, Wang P, Handcock M, Pattison P (2007) Recent developments in exponential random graph (p^*) models for social networks. Soc Netw 29:192–215
42. Skvoretz J, Faust K (1999) Logit models for affiliation networks. Sociol Methodol 29:253–280
43. Snijders TAB (1991) Enumeration and simulation methods for 0–1 matrices with given marginals. Psychometrika 56:397–417
44. Snijders TAB (2001) The statistical evaluation of social network dynamics. Sociol Methodol 31:361–395
45. Snijders TAB (2002) Markov chain Monte Carlo estimation of exponential random graph models. J Soc Struct 3:2
46. Snijders TAB (2006) Statistical methods for network dynamics. In: Luchini S et al (eds) Proceedings of the XLIII Scientific Meeting, Italian Statistical Society, pp 281–296
47. Snijders TAB, van Duijn MAJ (2002) Conditional maximum likelihood estimation under various specifications of exponential random graph models. In: Hagberg J (ed) Contributions to social network analysis, information theory, and other topics: A festschrift in honour of Ove Frank. Department of Statistics, University of Stockholm, Stockholm, pp 117–134
48. Snijders TAB, Pattison P, Robins G, Handcock M (2006) New specifications for exponential random graph models. Sociol Methodol 36:99–153
49. Snijders TAB, Steglich C, Schweinberger M (2007) Modeling the co-evolution of networks and behavior. In: van Montfort K, Oud H, Satorra A (eds) Longitudinal models in the behavioral and related sciences. Lawrence Erlbaum, Hillsdale, pp 41–71
50. Steglich C, Snijders TAB, West P (2006) Applying SIENA: An illustrative analysis of the coevolution of adolescents' friendship networks, taste in music, and alcohol consumption. Methodology 2:48–56
51. Strauss D (1986) On a general class of models for interaction. SIAM Rev 28:513–527
52. Strauss D, Ikeda M (1990) Pseudo-likelihood estimation for social networks. J Am Stat Assoc 85:204–212
53. van Duijn M, Snijders TAB, Zijlstra B (2004) p2: a random effects model with covariates for directed graphs. Stat Neerlandica 58:234–254
54. Wang P (2006) Exponential random graph (p^*) models for affiliation networks. Postgraduate diploma thesis, University of Melbourne
55. Wasserman S, Pattison P (1996) Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p^*. Psychometrika 61:401–425
56. Wasserman S, Robins G (2005) An Introduction to Random Graphs, Dependence Graphs, and p^*. In: Carrington P, Scott J, Wasserman S (eds) Models and Methods in Social Network Analysis. Cambridge University Press, New York, pp 148–161
57. Zhao Y, Robins G (2006) Multiple networks: Comparing QAP and exponential random graph (p^*) models. Sunbelt XXVI International Social Networks Conference, Vancouver, April 2006

Books and Reviews

Agneessens F, Roose H, Waege H (2004) Choices of theatre events: p^* models for affiliation networks with attributes. Metodoloski Zvezki 1:419–439

Anderson C, Wasserman S, Crouch B (1999) A p^* primer: Logit models for social networks. Soc Netw 21:37–66

Baddeley A, Möller J (1989) Nearest-neighbour Markov point processes and random sets. Int Stat Rev 57:89–121

Boer P, Huisman M, Snijders TAB, Zeggelink E (2003) StOCNET: an open software system for the advanced analysis of social networks. ProGAMMA/ICS, Groningen

Brieger R (1981) Comment on An exponential family of probability distributions for directed graphs. J Am Stat Assoc 76:51–53

Butts CT (2008) Social network analysis: A methodological introduction. Asian J Soc Psychol 11:13–41

Contractor N, Wasserman S, Faust K (2006) Testing multi-theoretical multilevel hypotheses about organizational networks: An analytic framework and empirical example. Acad Manag J 31:681–703
Frank O (1981) A survey of statistical methods for graph analysis. Sociol Methodol 11:110–155
Frank O (1991) Statistical analysis of change in networks. Stat Neerlandica 95:283–293
Frank O, Nowicki K (1993) Exploratory statistical analysis of networks. Quo Vadis, Graph Theory? Ann Discret Math 55:349–366
Handcock M, Hunter D, Butts C, Goodreau S, Morris M (2006) Statnet: An R package for the statistical analysis and simulation of social networks. University of Washington, Washington
Koehly L, Goodreau S, Morris M (2004) The link between exponential random graph models and loglinear models for networks. Sociol Methodol 34:241–270
Lazega E, van Duijn M (1997) Position in formal structure, personal characteristics and choices of advisors in a law firm: A logistic regression model for dyadic network data. Soc Netw 19:375–397
Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chkovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298:824–827
Monge P, Contractor N (2003) Theories of communication networks. Oxford University Press, New York
Newman M (2003) The structure and function of complex networks. SIAM Rev 45:167–256
Pattison P, Robins GL (2008) Probabilistic network theory. In: Rudas T (ed) Handbook of Probability: Theory with Applications. Sage, Thousand Oaks, pp 291–312
Pattison P, Wasserman S (2002) Multivariate random graph distributions: applications to social network analysis. In: Hagberg J (ed) Contributions to social network analysis, information theory and other topics in statistics: A festschrift in honour of Ove Frank. Department of Statistics, University of Stockholm, Stockholm, pp 74–100
Robins G, Pattison P (2001) Random graph models for temporal processes in social networks. J Math Sociol 25:5–41
Robins G, Pattison P, Woolcock J (2004) Models for social networks with missing data. Soc Netw 26:257–283
Schweinberger M, Snijders TAB (2003) Settings in social networks: A measurement model. Sociol Methodol 33:307–341
Wang P, Robins G, Pattison P (2006) PNet: Program for the estimation and simulation of *p**exponential random graph models, User Manual. School of Behavioural Science, University of Melbourne, Melbourne

Social Networks and Granular Computing

Churn-Jung Liau
Institute of Information Science, Academia Sinica, Taipei, Taiwan

Article Outline

Glossary
Definition of the Subject
Introduction
Mathematical Preliminaries
Rough Set Theory – Functional Granulation
Positional Analysis – Relational Granulation
Modal Logic – The Bridge
From Relational Granulation to Functional Granulation
Conclusion and Future Directions
Bibliography

Glossary

Granulation The process of drawing a set of objects (or points) together by indiscernibility, similarity, proximity, or functionality.

- **Functional granulation** The granulation process is functional if it is based on the attributes of the objects. It is called functional granulation because each attribute is a function from the set of objects to the set of values.
- **Relational granulation** The granulation process is relational if it is based on the relationships between objects.

Rough set A rough set is defined by the lower and upper approximations of a concept. The lower approximation contains all elements that necessarily belong to the concept, while the upper approximation contains those that possibly belong to the concept. In rough set theory, a concept is considered a classical set.

Social network A social network is comprised of a set of actors, called the domain, and a family of relations on the domain. It is usually represented as a graph, where each node represents an actor and an edge between two nodes represents a relational tie between these two actors. An edge can be labeled with the relation it represents.

Positional equivalence Two actors are in equivalent positions if their "pattern" of relationships with other actors is the same.

- **Structural equivalence** Two actors are structurally equivalent if they are related to the same actors.
- **Regular equivalence** Two actors are regularly equivalent if they are equally related to equivalent actors.
- **Exact equivalence** Two actors are exactly equivalent if they are related to the same number of equivalent actors.

Modal logic Modal logic was originally developed as a type of philosophical logic for reasoning about necessity and possibility. However, it has been extended to broadly cover a family of logics for reasoning about modalities including tense, obligation, belief, and knowledge. Semantically, it is also a powerful

mathematical discipline that deals with (restricted) description languages for discussing various kinds of relational structures, where a relational structure comprises a set of elements and a collection of relations on that set.

Hybrid logic Hybrid logic is a branch of modal logic that allows direct reference to the elements in a relational structure. Traditionally, only the properties of the elements could be represented by modal logic formulas.

Definition of the Subject

Granular computing (GrC) is a problem-solving concept deeply rooted in human thinking. Hence, it has played a major role in solving many important problems throughout the history of mathematics. GrC is concerned with the processing of information granules, which are groups of objects drawn together by indiscernibility, similarity, proximity, or functionality [37]. The process of forming information granules is called granulation. If the process is based totally on the attributes of the objects, it is called *functional granulation*, since attributes are mathematical functions from the set of objects to the set of values; if, in addition, the granulation process is also based on the relationship between objects, it is called *relational granulation* [11,20].

Interestingly, social scientists have applied the techniques of relational granulation (albeit by different names) to positional analysis in social networks [7,8,10,19,35]. Social network analysis (SNA) is a methodology used extensively in social and behavioral sciences, as well as in political science, economics, organization theory, and industrial engineering [15,31,34]. Positional analysis of a social network tries to find similarities between actors in the network. While many traditional clustering methods are based on the attributes of the individual actors, SNA is more concerned with the structural similarity between the actors. In SNA, a category, called a *social role* or *social position*, is defined in terms of the similarities of the patterns of relations among the actors, rather than the attributes of the actors.

Analyzing a social position from a GrC perspective has two major advantages. On the one hand, it extends the application scope of the GrC methodology. On the other hand, we can logically transform relational granulation of social networks into functional granulation; as a result, the well-developed techniques of data table analysis can be applied to social position analysis. Thus, this is an important cross-fertilization of GrC and SNA.

Introduction

GrC encompasses a wide range of techniques in computational intelligence, such as interval computing, classification, cluster analysis, and fuzzy and rough set theories. Among them, rough set theory [26,27] is one of the most well-known methods with successful applications in data analysis and uncertainty management. The theory was first proposed by Z. Pawlak as a mathematical tool to deal with vague concepts [26]. The basic assumption of the theory is that the effective use of knowledge is based on the capability to classify objects. Thus, knowledge consists of a family of classifications of a domain of interest. A classification of a domain is represented as a partition of (or an equivalence relation on) the domain. Each equivalence class of a classification is the elementary knowledge about the classification. For example, if the domain is classified according to colors, then the elementary knowledge may include "red", "green", "blue", etc. In this sense, a classification is derivable from a set of features or attributes of objects. An important task of data analysis is to define a concept via a set of features. If the concept can be precisely defined as the objects with some specific attribute values, then it is an exact concept with respect to the given attributes. However, in practice, most concepts are vague and can not be defined precisely by a given set of attributes. In such cases, rough set theory provides an effective way to approximate the vague concepts with some exact concepts in the domain. Because of its applications in data mining, rough set theory has flourished since 1990.

At about the time rough set theory was proposed, the pioneer of fuzzy set theory, L.A. Zadeh, emphasized the importance of information granularity in fuzzy reasoning [36]. Subsequently, T.Y. Lin proposed the term "GrC" to unify different techniques developed for processing information granules. From the GrC perspective, both classification in rough set theory and fuzzification in fuzzy set theory are types of granulation.

In rough set theory, objects are partitioned into equivalence classes based on their attribute values. While such values essentially represent functional information associated with the objects, more complicated information can be utilized in the classification of objects. In particular, relational information between objects can play an important role in the granulation process. The information is defined by general binary relations, which are extensions of the functional attributes of the objects. Geometrically, such granulation is derived from the neighborhood system of topological spaces [32], where each point/object is assigned at most one neighborhood/granule. This kind of granulation is called *relational granulation*, while granu-

lation based on attribute values only is called *functional granulation* [11,20].

Relational granulation is especially useful for the analysis of network data because a network is a collection of relations between nodes. A concrete instance of the application of relational granulation is positional analysis of social networks. Positional analysis attempts to find actors occupying the same position in a social network based on the pattern of their relationships to other actors. Depending on the different patterns of relationships, different notions of positional equivalence have been proposed [19]. These notions can be easily regarded as instances of relational granulation. In this article, we consider three of the most important notions of positional equivalence – *structural equivalence*, *regular equivalence*, and *exact equivalence*.

Recently, it was shown that social positions based on regular equivalence can be syntactically expressed as well-formed formulas (wff) in a kind of modal logic [22]. Thus, actors occupying the same social position based on regular equivalence will satisfy the same set of modal formulas. Traditionally, modal logic has been considered the logic for reasoning about modalities, such as necessity, possibility, time, action, belief, knowledge, and obligation. However, semantically, it is essentially a language for describing relational structures [5]. A relational structure is simply a set combined with a collection of relations on that set. Thus, social networks are mathematically equivalent to relational structures. The logical characterization of social positions implies that relational granulation for positional analysis can be transformed into a functional granulation process. Indeed, since each actor in a social network may satisfy or falsify a modal formula, a modal formula can be seen as a binary attribute associated with each actor. Therefore, based on the results in [22], two actors are regularly equivalent if they have the same values with respect to all attributes defined by the language of the particular modal logic.

By extending previous results, we can find logical characterizations of different positional equivalences and transform them into a functional granulation process. Consequently, the techniques of rough set-based data analysis can also be applied to positional analysis. However, since the set of wffs of a modal logic is usually infinite, the functional granulations corresponding to positional equivalences need to consider infinite sets. To circumvent this problem, we also propose a new definition of positional equivalence, called *observational equivalence*. Let Σ be a set of wffs in a logical language that denotes the set of observable properties of the actors in a network. Two actors a and b are observationally equivalent with respect to Σ, if for each wff $\varphi \in \Sigma$, a satisfies φ iff b satisfies φ.

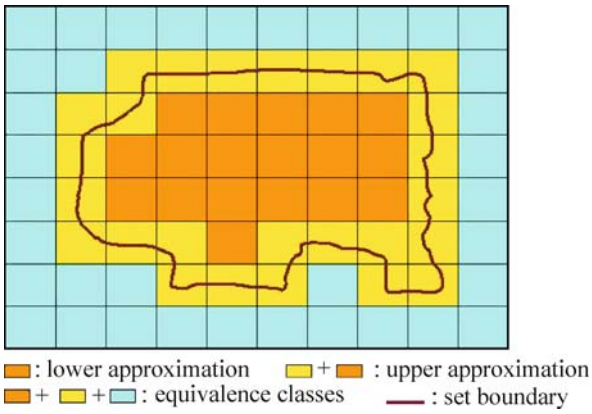

Social Networks and Granular Computing, Figure 1
A rough set: the set delimited by the *curve* (the set boundary) can not be represented exactly as a union of the basic building blocks. The *blocks* totally included in the set form the lower approximation of the set, while the *blocks* that intersect with the set (including the lower approximation) form the upper approximation of the set

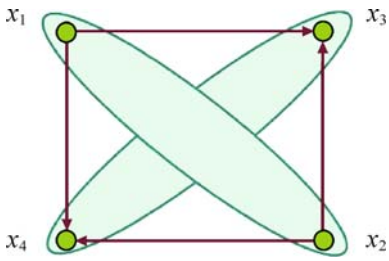

Social Networks and Granular Computing, Figure 2
Structural equivalence: the social network contains 4 actors denoted by the *nodes* and a single binary relation denoted by the *arcs*. Both the out-neighborhoods of x_1 and x_2 are $\{x_3, x_4\}$, and their common in-neighborhood is the empty set. On the other hand, both the in-neighborhoods of x_3 and x_4 are $\{x_1, x_2\}$, and their common out-neighborhood is the empty set. Thus, $x_1 \cong_s x_2$ and $x_3 \cong_s x_4$. The *green shadowed areas* indicate the equivalence classes

The remainder of this article is organized as follows. In Sect.“ Mathematical Preliminaries”, we introduce the background knowledge and mathematical notations of sets and relations. In Sects. “Rough Set Theory – Functional Granulation” to “Modal Logic – The Bridge”, we review rough set theory, social positional analysis, and modal logic respectively. In Sect. “From Relational Granulation to Functional Granulation”, we present the logical characterizations of three main positional equivalences and the definition of observational equivalence. Finally, in Sect. “Conclusion and Future Directions”, we present our conclusions and indicate some future research directions.

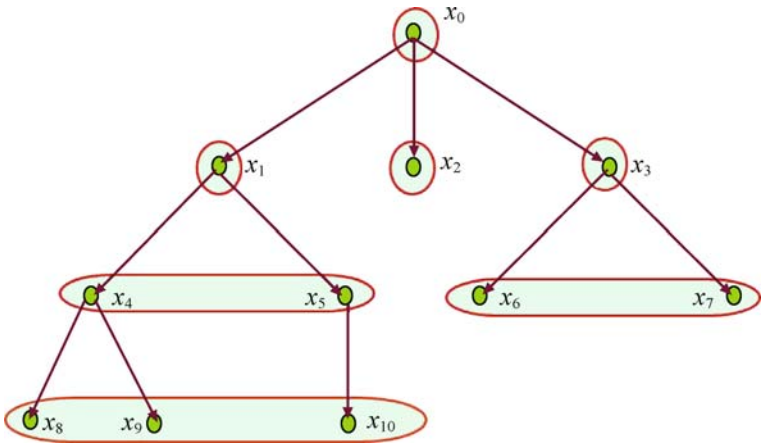

Social Networks and Granular Computing, Figure 3
Regular equivalence in a tree-structured social network: the root x_0 is discernible from other nodes because its in-neighborhood is empty, whereas all the other nodes have nonempty in-neighborhoods. Nodes x_1, x_2, x_3 are discernible from other nodes because they are the only nodes connected to x_0. Moreover, they are mutually discernible, since x_2 has no children, x_1 has children and grand-children, and x_3 has children, but not grand-children. Note that x_4 and x_5 are indiscernible, since they have the same in-neighborhood and equivalent out-neighborhoods; x_6 and x_7 are indiscernible, since they have the same in-neighborhood and both have empty out-neighborhoods; and x_8, x_9, and x_{10} are indiscernible, since they have equivalent in-neighborhoods and all have empty out-neighborhoods

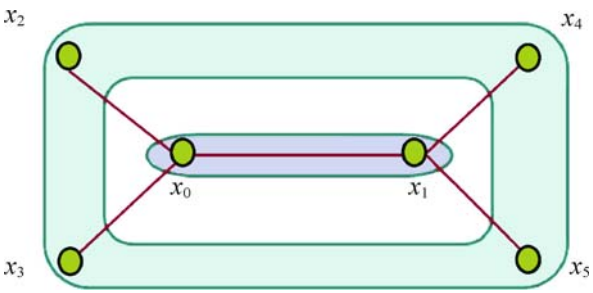

Social Networks and Granular Computing, Figure 4
Exact equivalence: the social network is a undirected graph. In other words, the binary relation is symmetry. The exact equivalence partitions the nodes into two classes, $C_0 = \{x_0, x_1\}$ and $C_1 = \{x_2, x_3, x_4, x_5\}$. To verify that it is indeed an exact equivalence, we can check that both x_0 and x_1 connect to one C_0 node and two C_1 nodes, and all of x_2, x_3, x_4, and x_5 connect to one C_0 node

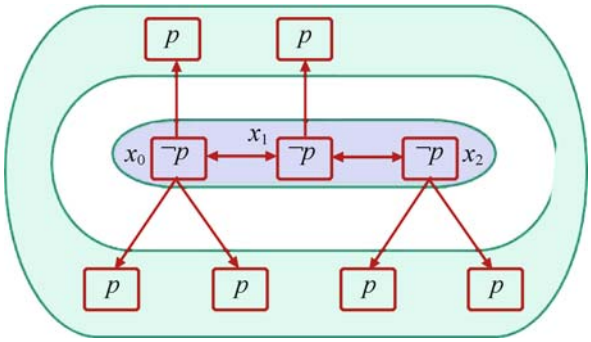

Social Networks and Granular Computing, Figure 5
A social network represented as a Kripke model: we assume the underlying modal language contains one propositional symbol p and one relational symbol r. Since $\{x_0, x_1, x_2\}$ is an equivalence class in the shown regular equivalence, the three nodes all satisfy the same set of PMML wffs. For example, they all satisfy $\neg p \wedge \langle r \rangle p$. However, they are not exactly equivalent, so they are distinguishable by GML wffs. For example, x_0 satisfies both $\langle r \rangle_1 p$ and $\langle r \rangle_2 p$, x_1 does not satisfy $\langle r \rangle_1 p$ or $\langle r \rangle_2 p$, and x_2 satisfies $\langle r \rangle_1 p$ but not $\langle r \rangle_2 p$

Mathematical Preliminaries

In this section, we introduce the basic knowledge about sets and relations used in this article.

Sets

A set is a collection of objects, called the elements of the set. We use capitals A, B, U, X, etc. to denote sets. The elements of a set are denoted by lower-case letters and we write $x \in U$ if x is an element of the set U. We usually use two notations to denote a set. The first one lists its elements. For example, we can write $\{x_1, x_2, \ldots, x_n\}$ for a finite set, where n is a positive integer, and $\{x_1, x_2, \ldots\}$ for an infinite set. The second notation specifies the defining property of the set. For example, we can write $\{x \mid x = 2n, n \in Z\}$ for the set of even numbers. The

number of elements in a set is called its cardinality. The cardinality of U is denoted by $|U|$. A set U is called an empty set if $|U| = 0$, and a singleton if $|U| = 1$. We denote the empty set by $\emptyset$.

The basic operations on sets are defined as follows:

1. Intersection of X and Y: $X \cap Y = \{u \mid u \in X$ and $u \in Y\}$,
2. Union of X and Y: $X \cup Y = \{u \mid u \in X$ or $u \in Y\}$,
3. Cartesian product of X and Y: $X \times Y = \{(x, y) \mid x \in X$ and $y \in Y\}$, and
4. Difference of X by Y: $X - Y = \{u \mid u \in X$ and $u \notin Y\}$.

Since intersection, union, and the Cartesian product are associative (i. e., $(X \cap Y) \cap Z = X \cap (Y \cap Z)$ etc.), we can eliminate the parentheses and apply the operations to more than two sets. For example, the Cartesian product of the sets $U_1, U_2, \ldots, U_k$ is denoted by

$$U_1 \times U_2 \times \cdots \times U_k = \{(x_1, x_2, \ldots, x_k) \mid x_i \in U_i, \forall 1 \le i \le k\} ,$$

where each element is called a k-tuple. A 2-tuple is also called a pair. If $U_1 = U_2 = \cdots = U_k = U$, the Cartesian product is written as U^k. A set X is said to be a subset of another set Y, denoted by $X \subseteq Y$, if for all $x \in X$, it is also the case that $x \in Y$. A function f from the set X to the set Y, denoted by $f: X \to Y$, is a mapping that associates each element in X with an element in Y. The set X is called the domain of the function, and Y is called the range of f. For each $x \in X$, we write $f(x)$ for the element in Y that is associated with x by f and call it the image of x under f. Also, the image of X under f is defined as $f(X) = \{f(x) \mid x \in X\}$.

Multisets

For sets, repeat occurrences of the same elements are only counted once. Thus, the set $\{x, x\}$ is the same as the set $\{x\}$. Sometimes, the number of occurrences of an element is important. In such cases, we consider multisets (or bags). A multiset is formally defined as a pair $M = (U, m)$, where U is some set and $m: U \to N$ is a function from U to the set $N = \{0, 1, 2, 3, \ldots\}$ of natural numbers. U is called the universe and for each element $x \in U$, $m(x)$ is the multiplicity (that is, the number of occurrences) of x in M. A set can be considered as a special multiset in which $m(x) = 0$ or 1 for each x in the universe. By generalizing the notation of set membership, we can write $x \in M$ if $m(x) > 0$. The cardinality of M is defined as $|M| = \sum_{x \in U} m(x)$. Let $M_1 = (U, m_1)$ and $M_2 = (U, m_2)$ be two multisets in the same universe; then, $M_1 \subseteq M_2$ iff $m_1(x) \le m_2(x)$ for all $x \in U$. The intersection, union, and difference between $M_1 = (U, m_1)$ and $M_2 = (U, m_2)$ are defined as follows:

1. $M_1 \cap M_2 = (U, m)$, where $m(x) = \min(m_1(x), m_2(x))$ for all $x \in U$,
2. $M_1 \cup M_2 = (U, m)$, where $m(x) = \max(m_1(x), m_2(x))$ for all $x \in U$,
3. $M_1 - M_2 = (U, m)$, where $m(x) = (m_1(x) - m_2(x)) \cdot \chi(m_1(x) - m_2(x))$ for all $x \in U$, and

$$\chi(k) = \begin{cases} 1, & \text{if} \quad k > 0, \\ 0, & \text{if} \quad k \le 0 . \end{cases}$$

The Cartesian product of $M_1 = (U_1, m_1)$ and $M_2 = (U_2, m_2)$ is defined as $M_1 \times M_2 = (U_1 \times U_2, m)$, where $m(x, y) = m_1(x)m_2(y)$ for all $x \in U_1$ and $y \in U_2$.

Relations

In mathematics, a subset of $U_1 \times U_2 \times \cdots \times U_k$ is called a k-ary relation among the sets $U_1, U_2, \ldots, U_k$, and a subset of U^k is called a k-ary relation on U. We are particularly interested in binary relations. According to the definition, a binary relation on U is a set of pairs of elements in U. A pair (x, y) in a binary relation means that x is related to y by the relation. We use lower-case Greek letters, α, β, ρ, etc. to denote relations. The basic operations on binary relations include the three aforementioned set-theoretic operations – intersection, union, and difference, as well as converse and composition. The converse of a binary relation $\alpha \subseteq U_1 \times U_2$ is $\alpha^- \subseteq U_2 \times U_1$ such that

$$\alpha^- = \{(x, y) \mid (y, x) \in \alpha\} .$$

The composition of two binary relations $\alpha \subseteq U_1 \times U_2$ and $\beta \subseteq U_2 \times U_3$ is $\alpha\beta \subseteq U_1 \times U_3$ such that

$$\alpha\beta = \{(x, y) \mid \exists z \in U, (x, z) \in \alpha \wedge (z, y) \in \beta\} .$$

A binary relation $\rho \subseteq U^2$ is said to be reflexive if $(x, x) \in \rho$ for all $x \in U$; ρ is symmetric if for all $x, y \in U$, $(x, y) \in \rho$ implies $(y, x) \in \rho$; ρ is anti-symmetric if for all $x, y \in U$, $(x, y) \in \rho$ and $(y, x) \in \rho$ implies $x = y$; and ρ is transitive if for all $x, y, z \in U$, $(x, y) \in \rho$ and $(y, z) \in \rho$ implies $(x, z) \in \rho$. A binary relation that is reflexive, anti-symmetric, and transitive is called a partial order. Let $\le$ be a partial order on the universe U and $X \subseteq U$, then, $u \in U$ is called an upper bound (resp. lower bound) of X if for all $x \in X$, $x \le u$ (resp. $u \le x$). The least upper bound (resp. the greatest lower bound) of X is the upper bound (resp. lower bound) u of X such that for any upper bounds (resp. lower bounds) u' of X, $u \le u'$ (resp. $u' \le u$). The least upper bound (resp. the greatest lower bound) of X is called

its supremum (resp. infimum), and is denoted by $\sup(X)$ (resp. $\inf(X)$). For any X (even though X is finite), $\sup(X)$ and $\inf(X)$ do not necessarily exist. If for all $x, y \in U$, $\sup(\{x, y\})$ and $\inf(\{x, y\})$ exist, then $\sup(\{x, y\})$ is called the join of x and y and denoted by $x \sqcup y$, while $\inf(\{x, y\})$ is called the meet of x and y and denoted by $x \sqcap y$. The structure $(U, \sqcap, \sqcup)$ is called a lattice. A subset $X \subseteq U$ is a sublattice of U if for any $x, y \in X$, $x \sqcap y$ and $x \sqcup y \in X$.

A binary relation that is reflexive, symmetric, and transitive is called an equivalence relation, which yields a partition of the universe. A partition of U is a class of subsets of U, $\{X_i \mid i \in I\}$ for some index set I, where $\bigcup_{i \in I} X_i = U$ and $X_i \cap X_j = \emptyset$ for any $i \neq j$. Let ρ be an equivalence relation on U, and let $x \in U$; then, the equivalence class containing x is defined as

$$[x]_\rho = \{y \in U \mid (x, y) \in \rho\} .$$

Obviously, $\{[x]_\rho \mid x \in U\}$ forms a partition of the universe. For any $X \subseteq U$, we denote $[X]_\rho$ by the set $\{[x]_\rho \mid x \in X\}$ and $[\![X]\!]_\rho$ by the multiset $([U]_\rho, m)$ such that $m([x]_\rho) = |[x]_\rho \cap X|$ for all $x \in U$.

Equivalence relations on a fixed domain are partially ordered by set-inclusion. Let ρ_1 and ρ_2 be two equivalence relations and define $\rho_1 \leq \rho_2$ iff $\rho_1 \subseteq \rho_2$; then, $\leq$ is a partial order on the set of all equivalence relations on a fixed domain. We say that ρ_1 is finer than ρ_2 or ρ_2 is coarser than ρ_1 if $\rho_1 \leq \rho_2$. Given a binary relation α, the composition of α with itself k times is denoted by α^k and the transitive closure of α is defined as $\alpha^\infty = \bigcup_{k \geq 1} \alpha^k$. Let ρ_1 and ρ_2 be two equivalence relations; then it can be shown that $\rho_1 \cap \rho_2$ and $(\rho_1 \cup \rho_2)^\infty$ are, respectively, the meet and the join of ρ_1 and ρ_2.

Rough Set Theory – Functional Granulation

The basic construct of rough set theory is the approximation space, which is a pair (U, ρ), where U is a finite set of objects (the universe) and ρ is an equivalence relation on U. From the GrC perspective, the equivalence classes of ρ are information granules of the approximation space.

In rough set theory, a concept is represented as a subset of the universe. For any concept $X \subseteq U$, we can associate the following two subsets with X,

$$\underline{\rho}X = \{x \in U \mid [x]_\rho \subseteq X\}$$
$$\overline{\rho}X = \{x \in U \mid [x]_\rho \cap X \neq \emptyset\} ,$$

where $\underline{\rho}X$ and $\overline{\rho}X$ are called the ρ-lower and ρ-upper approximation of X, respectively. From a practical viewpoint, ρ can be considered as an indiscernibility relation; thus, for a given concept X, we can only know that X contains all the elements in $\underline{\rho}X$ and does not contain any element outside $\overline{\rho}X$. In other words, the concept X cannot be defined exactly in the approximation space. The pair $(\underline{\rho}X, \overline{\rho}X)$ is considered a rough approximation of X, and any such pair is called a rough set (see Fig. 1). A concept X is called an exact concept in the approximation space if $\underline{\rho}X = X = \overline{\rho}X$; otherwise, it is called a rough concept.

Rough set theory is very useful in the analysis of data tables. A data table is a pair $S = (U, F)$, where U is a nonempty, finite set (the universe) and F is a nonempty, finite set of primitive attributes. Every $f \in F$ is a total function $f\colon U \to V_f$, where V_f denotes the possible values of f. If $E = \{f_1, \ldots, f_n\} \subseteq F$, then $V_E = V_{f_1} \times \cdots \times V_{f_n}$. Thus, E is also considered a function from U to V_E. An equivalence relation $\mathrm{IND}(E)$ is associated with every subset of attributes $E \subseteq F$, and defined by

$$(x, y) \in \mathrm{IND}(E) \Leftrightarrow f(x) = f(y) \quad \forall f \in E .$$

$\mathrm{IND}(E)$ is called an indiscernibility relation. We write $\mathrm{IND}(f)$ instead of $\mathrm{IND}(\{f\})$ for all $f \in F$. It is easy to show that $\mathrm{IND}(E) = \bigcap_{f \in E} \mathrm{IND}(f)$. Since $\mathrm{IND}(E)$ is an equivalence relation, $(U, \mathrm{IND}(E))$ is an approximation space and we can define the $\mathrm{IND}(E)$-lower and $\mathrm{IND}(E)$-upper approximation of X for any $X \subseteq U$. We write $\underline{E}X$ and $\overline{E}X$ instead of $\underline{\mathrm{IND}(E)}X$ or $\overline{\mathrm{IND}(E)}X$. Obviously, exact concepts in $(U, \mathrm{IND}(E))$ are those definable using only attributes in E. In effect, the indiscernibility relation $\mathrm{IND}(E)$ granulates (or partitions) the universe according to the functions in E; thus, it results in a functional granulation of the universe.

The definitions are used in the analysis of dependency between attributes in a data table. Let us say that attribute E_2 depends on E_1, denoted by $E_1 \Rightarrow E_2$, iff $\mathrm{IND}(E_1) \subseteq \mathrm{IND}(E_2)$, i. e., any two objects in U with the same attribute values in E_1 will also have the same attribute values in E_2. It is easy to show that $E_1 \Rightarrow E_2$ iff $\underline{E_1}X = X$ for all X that are equivalence classes of $\mathrm{IND}(E_2)$. Given an attribute $f \in F$, if $F - \{f\} \Rightarrow \{f\}$, then f can be eliminated from F without influencing the definability of F. In other words, by using the set of attributes, f is redundant for classification purposes. By repeated elimination of redundant attributes, we can eventually obtain an irreducible set of attributes with the same classification capability as the original set of attributes. Such an irreducible set of attributes is called a reduct in rough set theory. Note that the definition of dependency is relative to a data table, so it does not necessarily mean that an intrinsic connection exists between the inter-dependent attributes.

Positional Analysis – Relational Granulation

Social networks are defined by actors and relations (or nodes and edges in terms of graph theory) [15]. A social network is generally defined as a relational structure $\mathfrak{N} = (A, (\alpha_i)_{i\in I})$, where A is the set of actors in the network, I is an index set, and for each $i \in I$, $\alpha_i \subseteq A^{k_i}$ is a k_i-ary relation on the domain A. If $k_i = 1$, then α_i is also called an attribute. In practice, most SNA literature considers a simplified version of social networks with only binary relations. For ease of presentation, we focus on a social network with only unary and/or binary relations. Thus, the social network considered in this article is a structure $\mathfrak{N} = (A, (P_i)_{i\in I}, (\alpha_j)_{j\in J})$, where A is a *finite set* of actors, $P_i \subseteq A$ for all $i \in I$, and $\alpha_j \subseteq A \times A$ for all $j \in J$. In terms of graph theory, $\mathfrak{N}$ is a labeled graph, where A is a set of nodes labeled with subsets of I, and each α_j denotes a set of (labeled) edges. For each $a \in A$, the out-neighborhood and in-neighborhood of a with respect to a binary relation α, denoted respectively by $N^+_\alpha(a)$ and $N^-_\alpha(a)$, are defined as follows:

$$N^+_\alpha(a) = \{b \in A \mid (a, b) \in \alpha\} ,$$
$$N^-_\alpha(a) = \{b \in A \mid (b, a) \in \alpha\} .$$

If ρ is an equivalence relation on A and a is an actor, the ρ-equivalence class of a is equal to its neighborhood, i. e., $[a]_\rho = N^+_\rho(a) = N^-_\rho(a)$. Note that the latter equality holds because of the symmetry of ρ.

Several equivalence relations have been proposed for exploring the similarity between actors' roles. The simplest definition of positional equivalence is the concept of structural equivalence proposed in [21], which states that two actors are positionally equivalent if they are related to the same individuals (see Fig. 2).

Definition 1 Let $\mathfrak{N} = (A, (P_i)_{i\in I}, (\alpha_j)_{j\in J})$ be a social network and ρ be an equivalence relation on A; then ρ is a (strong) structural equivalence with respect to $\mathfrak{N}$ if $(a, b) \in \rho$ implies

1. $a \in P_i$ iff $b \in P_i$ for all $i \in I$, and
2. $N^+_{\alpha_j}(a) = N^+_{\alpha_j}(b)$ and $N^-_{\alpha_j}(a) = N^-_{\alpha_j}(b)$ for all $j \in J$.

By the definition, there may exist more than one structural equivalence for a given network. However, it has been shown that there always exists a maximum (i. e., coarsest) structural equivalence for a network [19]. In fact, the set of all structural equivalences form a sublattice of all equivalence relations on A. Indeed, if ρ_1 and ρ_2 are structural equivalences with respect to $\mathfrak{N}$, so too are $\rho_1 \sqcap \rho_2$ and $\rho_1 \sqcup \rho_2$. Since the set of all equivalence relations on a finite domain is finite, the join of all structural equivalences is the coarsest structural equivalence with respect to a network. Let $\mathfrak{N} = (A, (P_i)_{i\in I}, (\alpha_j)_{j\in J})$ be a social network and ρ be the coarsest structural equivalence with respect to $\mathfrak{N}$; then, two actors $x, y \in A$ are said to be structurally equivalent, denoted by $x \cong_s y$, if $(x, y) \in \rho$.

Although structural equivalence is conceptually simple, it is sometimes restrictive. Regular equivalence relaxes the requirement that equivalent actors must be connected with identical actors, and suggests that actors occupy the same position if they are connected to positionally equivalent actors. Regular equivalence has been studied extensively [7,8,10,19,35], and there are several alternative definitions of the concept. We present two here. The first is based on the characterization given by Boyd and Everett [8], which states that an equivalence relation ρ is a *regular equivalence* with respect to a binary relation α if it commutes with α, i. e.

$$\alpha\rho = \rho\alpha .$$

By this definition, if ρ is a regular equivalence with respect to α and $(a, b) \in \rho$, then for each $c \in N^+_\alpha(a)(\text{resp.} N^-_\alpha(a))$, there exists $c' \in N^+_\alpha(b)(\text{resp.} N^-_\alpha(b))$ such that $(c, c') \in \rho$. The property naturally leads to an alternative definition of regular equivalence [19], which states that an equivalence relation ρ is a *regular equivalence* with respect to a binary relation α if for $a, b \in A$,

$$(a, b) \in \rho \Rightarrow [N^+_\alpha(a)]_\rho = [N^+_\alpha(b)]_\rho \quad \text{and} \quad [N^-_\alpha(a)]_\rho = [N^-_\alpha(b)]_\rho .$$

According to this definition, if a and b are regularly equivalent, then they are connected to equivalent neighborhoods (see Fig. 3). Obviously, the above definitions are equivalent. Thus, we have the following definition.

Definition 2 Let $\mathfrak{N} = (A, (P_i)_{i\in I}, (\alpha_j)_{j\in J})$ be a social network and ρ be an equivalence relation on A; then ρ is a regular equivalence with respect to $\mathfrak{N}$ if

1. $(a, b) \in \rho$ implies $a \in P_i$ iff $b \in P_i$ for all $i \in I$; and
2. ρ is a regular equivalence with respect to α_j for all $j \in J$.

Analogous to the case of structural equivalences, the join of two regular equivalences is still a regular equivalence, even though their meet is not necessarily a regular equivalence. Consequently, we can find the coarsest regular equivalence of a network. Then, two actors, x and y, are regularly equivalent, denoted by $x \cong_r y$, if (x, y) is in the coarsest regular equivalence of the network.

For regular equivalence, only the occurrence or non-occurrence of a position in the neighborhood of an actor matters. However, the number of occurrences is sometimes an important factor in positional analysis. In such cases, a number restriction can be added to the definition of regular equivalences. An equivalence relation ρ is an *exact equivalence* with respect to a binary relation α if for $a, b \in A$,

$$(a, b) \in \rho \Rightarrow [\![N_\alpha^+(a)]\!]_\rho = [\![N_\alpha^+(b)]\!]_\rho \quad \text{and} \quad [\![N_\alpha^-(a)]\!]_\rho = [\![N_\alpha^-(b)]\!]_\rho \,.$$

By replacing the set equality in the definition with the multiset equality, the number of equivalent neighbors must be the same for two actors to be considered position-equivalent (see Fig. 4).

Definition 3 Let $\mathfrak{N} = (A, (P_i)_{i\in I}, (\alpha_j)_{j\in J})$ be a social network and ρ be an equivalence relation on A; then, ρ is an exact equivalence with respect to $\mathfrak{N}$ if

1. $(a, b) \in \rho$ implies $a \in P_i$ iff $b \in P_i$ for all $i \in I$; and
2. ρ is an exact equivalence with respect to α_j for all $j \in J$.

The set of all exact equivalences also forms a lattice, but it is not a sublattice of the set of all equivalence relations [10]. Consequently, the coarsest exact equivalence for a network exists and we can define two actors x and y as exactly equivalent, denoted by $x \cong_e y$, if they belong to the coarsest exact equivalence. The partition produced by an exact equivalence is called an equitable partition or a divisor of a graph [19].

From the above definition, it is clear that all the attributes in a social network are binary. In other words, for each attribute P_i, an actor either satisfies P_i (has the value 1) or falsifies P_i (has the value 0). When a social network $\mathfrak{N} = (A, (P_i)_{i\in I}, (\alpha_j)_{j\in J})$ has the property $J = \emptyset$, it is in fact a kind of data table. In this case, all three kinds of positional equivalence are reduced to the indiscernibility relation. This means that, if information about the actors' relationship is not available, social positions are fully determined by the attributes of the actors. Thus, without relational information, functional granulation is sufficient for positional analysis. However, if relational information is available, two actors can be further differentiated by their relationship with other actors, even though all their attributes have the same values. In this sense, the definition of positional equivalence is based on the process of relational granulation.

Modal Logic – The Bridge

Modal logics were originally developed as formalizations for reasoning about modalities. The formal study started with alethic modal logics (the logics of necessity and possibility). The modalities $\Box$ (for necessity) and $\Diamond$ (for possibility) have now become standard notations in modal logic literature, and many variants of modal logic have been proposed to deal with different classes of modalities; for example, temporal logic for tense operators, deontic logic for obligation and permission, dynamic logic for actions, and epistemic logic for beliefs and knowledge. The development of these logics was motivated by philosophical enquiry as well as by technical applications in computer science, artificial intelligence, and economic game theory.

While modal logics were initially presented in the form of reasoning systems, the invention of relational semantics has been influential in the continuing development of the field [17,18]. The semantics, proposed independently by J. Hintikka, S. Kanger, and S. Kripke, and now known as Kripke semantics, show that modal logics are in fact logics for reasoning about relational structures. From this semantic perspective, standard modal logics can be regarded as fragments of first-or second-order predicate logics, whereby the necessity and possibility modalities correspond to universal and existential quantifiers, respectively. In spite of this correspondence, quantification in modal logic tends to be bounded in some way to worlds that are "relevant to" or "accessible from" the current one. Consequently, although a number of properties of modal logics follow immediately from those of their classical quantificational counterparts, the modal operators typically have less expressive power than full quantification. This results in many interesting properties not available in classical predicate logic. One of the most striking results is that semantic invariances between models are actually various forms of *bisimulation*, which preserve the local properties of worlds and their transition patterns [5]. Interestingly, it has been shown that bisimulation in Kripke models corresponds exactly to regular equivalence in social networks [22]. The implication of the results is that position-equivalent actors can be characterized by an appropriate set of modal formulas. As each modal formula can be regarded as a binary attribute, from the GrC perspective, such characterization transforms relational granulation into functional granulation.

In this section, we present three modal logics – a multimodal logic, a graded modal logic, and a hybrid logic. The presentation of a modal logic, as with any other formal logic, requires the specification of its syntax and semantics. The syntactic aspect includes the language of the logic – its

alphabet and formation rules for wffs, as well as a deductive system for logical reasoning. For the purpose of this article, we need only present the language part. For the semantic aspect, we have to define the models within which the wffs can be interpreted, as well as the satisfaction of a wff in a model.

Propositional Multi-Modal Logic

We start with propositional modal logic (PML) [9]. The alphabet of PML consists of a set of propositional symbols, PV, and the logical symbols $\neg$(negation), $\wedge$(and), $\vee$(or), $\supset$(material implication), and $\Diamond$(possibility); $\neg$, $\wedge$, $\vee$, and $\supset$ are called Boolean connectives, whereas $\Diamond$ is the only primitive modality of PML. The set of wffs of PML is the smallest set containing PV that satisfies the following conditions:

- if φ is a wff, then $\neg\varphi$ and $\Diamond\varphi$ are wffs,
- if φ and ψ are wffs, then $\varphi \wedge \psi, \varphi \vee \psi$, and $\varphi \supset \psi$ are wffs.

As usual, we abbreviate $(\varphi \supset \psi) \wedge (\psi \supset \varphi)$ as $\varphi \equiv \psi$; $\neg \Diamond \neg\varphi$ as $\Box\varphi$; any tautology $p \vee \neg p$ as $\top$; and any contradiction $p \wedge \neg p$ as $\bot$. A Kripke model for PML is a triple $\mathfrak{M} = (W, \alpha, V)$, where W is a set of possible worlds, α is a binary relation on W, called an accessibility relation, and $V\colon W \times PV \to \{0, 1\}$ is a truth assignment for evaluating the truth value of each propositional symbol in each world. The satisfaction of a wff φ in a world w of the model $\mathfrak{M}$, denoted by $\mathfrak{M}, w \models \varphi$, is defined by the following clauses:

1. $\mathfrak{M}, w \models p$ if $V(w, p) = 1$ for each $p \in PV$;
2. $\mathfrak{M}, w \models \neg\varphi$ iff $\mathfrak{M}, w \not\models \varphi$;
3. $\mathfrak{M}, w \models \varphi \wedge \psi$ iff $\mathfrak{M}, w \models \varphi$ and $\mathfrak{M}, w \models \psi$;
4. $\mathfrak{M}, w \models \varphi \vee \psi$ iff $\mathfrak{M}, w \models \varphi$ or $\mathfrak{M}, w \models \psi$;
5. $\mathfrak{M}, w \models \varphi \supset \psi$ iff $\mathfrak{M}, w \not\models \varphi$ or $\mathfrak{M}, w \models \psi$;
6. $\mathfrak{M}, w \models \Diamond\varphi$ iff there exists $(w, u) \in \alpha$ such that $\mathfrak{M}, u \models \varphi$.

Propositional multi-modal logic (PMML) is an extension of PML that allows more than one modality [16]. PMML is particularly suitable for reasoning about relational structures that contain a number of binary relations, e. g., social networks. The alphabet of PMML consists of a set of propositional symbols PV, a set of relational symbols REL, the Boolean connectives, the relational converse symbol $^-$, and the modality-forming symbol $\langle\rangle$. The set of wffs of PMML is the smallest set containing PV that satisfies the following conditions:

- if φ is a wff, then $\neg\varphi$ is a wff;
- if φ is a wff and r is a relational symbol, then $\langle r\rangle\varphi$ and $\langle r^-\rangle\varphi$ are wffs;
- if φ and ψ are wffs, then $\varphi \wedge \psi, \varphi \vee \psi$, and $\varphi \supset \psi$ are wffs.

We abbreviate $\neg\langle r\rangle\neg\varphi$ and $\neg\langle r^-\rangle\neg\varphi$ as $[r]\varphi$ and $[r^-]\varphi$, respectively. A Kripke model for PMML is $\mathfrak{M} = (W, (\alpha_r)_{r\in REL}, V)$, where W and V are the same as above, and for each $r \in REL$, α_r is a binary relation on W. The satisfaction of PMML wffs is defined in the same way as that of PML wffs, except that clause 6 is replaced by

6. $\mathfrak{M}, w \models \langle r\rangle\varphi$ iff there exists $(w, u) \in \alpha_r$ such that $\mathfrak{M}, u \models \varphi$;
7. $\mathfrak{M}, w \models \langle r^-\rangle\varphi$ iff there exists $(u, w) \in \alpha_r$ such that $\mathfrak{M}, u \models \varphi$.

Note that we need the converse modalities $\langle r^-\rangle$ because all positional equivalences are defined with respect to their in-neighborhoods and out-neighborhoods. The modalities $\langle r\rangle$ allow us to access the out-neighborhood of a world, whereas $\langle r^-\rangle$ is needed to access the in-neighborhood.

Graded Modal Logic

Graded modal logic (GML) is a generalization of PMML that allows us to consider the number of worlds accessible from a given world [33]. The alphabet of GML is the same as that of PMML; however, the second formation rule of wffs is replaced by the following rule:

- if φ is a wff and r is a relational symbol, then $\langle r\rangle_n\varphi$ and $\langle r^-\rangle_n\varphi$ are wffs for each natural number n.

We also write $[r]_n\varphi$ and $[r^-]_\varphi$ as the abbreviations of $\neg\langle r\rangle_n\neg\varphi$ and $\neg\langle r^-\rangle_n\neg\varphi$, respectively.

The Kripke models for GML are the same as those for PMML, but clauses 6 and 7 for the conditions of satisfaction of PMML wffs are modified as follows:

6. $\mathfrak{M}, w \models \langle r\rangle_n\varphi$ iff $|\{u \in W \mid (w, u) \in \alpha_r$ and $\mathfrak{M}, u \models \varphi\}| > n$;
7. $\mathfrak{M}, w \models \langle r^-\rangle_n\varphi$ iff $|\{u \in W \mid (u, w) \in \alpha_r$ and $\mathfrak{M}, u \models \varphi\}| > n$.

According to the semantics, the modalities $\langle r\rangle_0$ and $\langle r^-\rangle_0$ in GML are respectively equivalent to $\langle r\rangle$ and $\langle r^-\rangle$ in PMML. Intuitively, $\mathfrak{M}, w \models \langle r\rangle_n\varphi$ means that φ is true in more than n worlds that are r-accessible from w, whereas $\mathfrak{M}, w \models [r]_n\varphi$ means that φ is false in no more than n worlds that are r-accessible from w. Thus, we abbreviate $\langle r\rangle_{n-1}\varphi \wedge [r]_n\neg\varphi$ (resp. $\langle r^-\rangle_{n-1}\varphi \wedge [r^-]_n\neg\varphi$) as $(r)_n\varphi$ (resp. $(r^-)_n\varphi$) to denote that there are exactly n r-accessible (resp. r^--accessible) worlds that satisfy φ.

Hybrid Logic

Hybrid logics (HL) are extensions of modal logics that include symbols to name possible worlds in models. The development of HL dates back to A. Prior's 1951 work, which had a philosophical foundation [29,30] (see [1] for a brief overview). The main idea is to use a special kind of atomic formula, called a nominal, to refer to possible worlds in models. The idea was adopted by the Sofia school in the late 80s and early 90s [14,24,25]. Subsequently, Blackburn and Seligman, who have been influential in the recent development of HL, adopted it in the late 90s [4,6].

Here, we only need the simplest hybrid logic (SHL). The alphabet of SHL is the same as that of PMML plus a set of nominals NOM that is disjoint with the set PV. The set of wffs of SHL is the smallest set containing $PV \cup NOM$ that satisfies the formation rules of PMML. The key semantic idea of HL is that each nominal must be true in exactly one possible world in any model. In other words, a nominal names a world by being true in that world and nowhere else. The idea can be easily formalized by extending PMML models with a component that assigns a nominal to a possible world. A hybrid model is then $\mathfrak{M} = (W, (\alpha_r)_{r \in REL}, V, f)$, where $(W, (\alpha_r)_{r \in REL}, V)$ is a PMML model and $f\colon NOM \to W$ is the mapping that assigns a nominal to a possible world. The satisfaction of SHL wffs is defined by clauses 1–7 for PMML, with the addition of the following clause for nominals:

8. $\mathfrak{M}, w \models i$ if $f(i) = w$ for each $i \in NOM$.

From Relational Granulation to Functional Granulation

In this section, we explain how to transform positional equivalences from relational granulation to functional granulation. Given a social network $\mathfrak{N} = (A, (P_i)_{i \in I}, (\alpha_j)_{j \in J})$, the steps are as follows:

1. For each kind of equivalence, find a corresponding modal language based on $\mathfrak{N}$. The three kinds of positional equivalence discussed earlier correspond to the following logical languages:
   (a) Structural equivalence: SHL
   (b) Regular equivalence: PMML
   (c) Exact equivalence: GML.
2. Consider $\mathfrak{N}$ as a model in the corresponding logic by a simple transformation. The set of actors is the set of possible worlds in the resultant model.
3. Find a sublanguage $\mathcal{L}$ of the corresponding logic. Prove that for any two actors x and y, $x \cong y$ iff x and y satisfy the same set of wffs in $\mathcal{L}$, where $\cong$ is $\cong_s$, $\cong_r$, or $\cong_e$ (see Fig. 5).

Structural Equivalences and Hybrid Logics

We use SHL to characterize structural equivalences. Given a social network $\mathfrak{N} = (A, (P_i)_{i \in I}, (\alpha_j)_{j \in J})$, we define an SHL language with the following alphabet:

1. $PV = \{p_i \mid i \in I\}$;
2. $REL = J$;
3. $NOM = A$.

The social network $\mathfrak{N}$ is transformed into a hybrid model $\mathfrak{M}_{\mathfrak{N}} = (A, (\alpha_j)_{j \in J}, V, id)$, where V is defined by $V(x, p_i) = 1$ iff $x \in P_i$ for all $x \in A$ and $i \in I$, and id is the identity function on A. Let us now define a sublanguage $\mathcal{H}_{\mathfrak{N}}$ of the SHL as the set of wffs $PV \cup \{\langle j \rangle a, \langle j^- \rangle a \mid j \in REL, a \in NOM\}$. We say that two actors, x and y, are $\mathcal{H}_{\mathfrak{N}}$-equivalent with respect to $\mathfrak{N}$ if for all $\varphi \in \mathcal{H}_{\mathfrak{N}}$, $(\mathfrak{M}_{\mathfrak{N}}, x \models \varphi$ iff $\mathfrak{M}_{\mathfrak{N}}, y \models \varphi)$. Then, we have the following characterization theorem.

Theorem 1 *Let $\mathfrak{N} = (A, (P_i)_{i \in I}, (\alpha_j)_{j \in J})$ be a social network; then, for all $x, y \in A$, $x \cong_s y$ in $\mathfrak{N}$ iff x and y are $\mathcal{H}_{\mathfrak{N}}$-equivalent with respect to $\mathfrak{N}$.*

Proof By the satisfaction condition, $(\mathfrak{M}_{\mathfrak{N}}, x \models \langle j \rangle a$ iff $(x, a) \in \alpha_j$. Thus, x and y are $\mathcal{H}_{\mathfrak{N}}$-equivalent iff the following two conditions are satisfied

1. $x \in P_i$ iff $y \in P_i$ for all $i \in I$ (as $PV \subseteq \mathcal{H}_{\mathfrak{N}}$),
2. for all $j \in J$ and $a \in A$, $(x, a) \in \alpha_j$ iff $(y, a) \in \alpha_j$ and $(a, x) \in \alpha_j$ iff $(a, y) \in \alpha_j$ (since both $\langle j \rangle a, \langle j^- \rangle a$ and $\langle j \rangle a, \langle j^- \rangle a$ are in $\mathcal{H}_{\mathfrak{N}}$).

These two conditions are exactly the same as those in Definition 1. Since $x \cong_s y$ iff (x, y) is in the coarsest structural equivalence, $x \cong_s y$ iff these two conditions are satisfied.□

Regular Equivalences and Multimodal Logics

The first logical characterization of positional equivalences is that of regular equivalences by dynamic logic – a particular type of PMML. This is shown in [22] by the connection between regular equivalence and the well-known notion of bisimulation in modal logics. We present the result here by following the general procedure of the transformation from relational granulation to functional granulation.

Given a social network $\mathfrak{N} = (A, (P_i)_{i \in I}, (\alpha_j)_{j \in J})$, we define a PMML language with the following alphabet:

1. $PV = \{p_i \mid i \in I\}$;
2. $REL = J$.

The social network $\mathfrak{N}$ is transformed into a Kripke model $\mathfrak{M}_{\mathfrak{N}} = (A, (\alpha_j)_{j\in J}, V)$, where V is defined by $V(x, p_i) = 1$ iff $x \in P_i$ for all $x \in A$ and $i \in I$. Let $\mathcal{L}_{\mathfrak{N}}$ be the set of wffs of the PMML and $\mathcal{L}_{\mathfrak{N}}$-equivalence be defined the same as $\mathcal{H}_{\mathfrak{N}}$-equivalence.

Theorem 2 ([22]) *Let $\mathfrak{N} = (A, (P_i)_{i\in I}, (\alpha_j)_{j\in J})$ be a social network; then, for all $x, y \in A$, $x \cong_r y$ in $\mathfrak{N}$ iff x and y are $\mathcal{L}_{\mathfrak{N}}$-equivalent with respect to $\mathfrak{N}$.*

Proof It is shown in [22] that regular equivalences are exactly bisimulations, i. e., $x \cong_r y$ iff x and y are bisimular. By Theorems 2.20 and 2.24 in [5], x and y are bisimular iff they are $\mathcal{L}_{\mathfrak{N}}$-equivalent with respect to $\mathfrak{N}$. □

Exact Equivalences and Graded Modal Logics

Following the procedure used in the previous subsection, we define $\mathcal{G}_{\mathfrak{N}}$ as the set of wffs of the GML based on the given network $\mathfrak{N}$. Then, we have the following theorem.

Theorem 3 *Let $\mathfrak{N} = (A, (P_i)_{i\in I}, (\alpha_j)_{j\in J})$ be a social network; then, for all $x, y \in A$, $x \cong_e y$ in $\mathfrak{N}$ iff x and y are $\mathcal{G}_{\mathfrak{N}}$-equivalent with respect to $\mathfrak{N}$.*

Proof

1. ($\Rightarrow$): Assume that $x \cong_e y$; then, there exists an exact equivalence ρ on $\mathfrak{N}$ such that $(x, y) \in \rho$. We show that $\mathfrak{M}_{\mathfrak{N}}, x \models \varphi$ implies $\mathfrak{M}_{\mathfrak{N}}, y \models \varphi$ by induction on the complexity of φ (the proof of the converse direction is exactly the same). The case where $\varphi \in PV$ follows from condition 1 of Definition 3. The cases for Boolean connectives are straightforward by the induction hypothesis. For formulas of the form $\langle j\rangle_n\psi$ or $\langle j^-\rangle_n\psi$, we prove the case for $\langle j\rangle_n\psi$ by using the condition $[\![N^+_{\alpha_j}(x)]\!]_\rho = [\![N^+_{\alpha_j}(y)]\!]_\rho$. The case for $\langle j^-\rangle_n\psi$ is proved similarly by applying $[\![N^-_{\alpha_j}(x)]\!]_\rho = [\![N^-_{\alpha_j}(y)]\!]_\rho$. Now, if $\mathfrak{M}_{\mathfrak{N}}, x \models \langle j\rangle_n\psi$, then there exist more than n worlds in $N^+_{\alpha_j}(x)$ that satisfy ψ. Assume $u_0, u_1, \ldots, u_n$ are among these worlds. Since $[\![N^+_{\alpha_j}(x)]\!]_\rho = [\![N^+_{\alpha_j}(y)]\!]_\rho$, there exist $u'_0, u'_1, \ldots, u'_n \in N^+_{\alpha_j}(y)$ such that $(u_i, u'_i) \in \rho$ for $0 \le i \le n$. By the induction hypothesis, $\mathfrak{M}_{\mathfrak{N}}, u'_i \models \psi$ for $0 \le i \le n$; thus, $\mathfrak{M}_{\mathfrak{N}}, y \models \langle j\rangle_n\psi$.
2. ($\Leftarrow$): Let us define an equivalence relation ρ on A by

$$(x, y) \in \rho \quad \text{iff} \quad x \text{ and } y \text{ are } \mathcal{G}_{\mathfrak{N}}\text{-equivalent with respect to } \mathfrak{N}.$$

We prove that ρ is an exact equivalence. The first condition of Definition 3 is straightforward. For the second condition, let us assume that $(x, y) \in \rho$, $[\![N^+_{\alpha_j}(x)]\!]_\rho = ([A]_\rho, m_1)$, and $[\![N^+_{\alpha_j}(y)]\!]_\rho = ([A]_\rho, m_2)$ for some $j \in J$. Assume also that $[A]_\rho = \{[z_i]_\rho \mid 1 \le i \le k\}$. First, we consider z_1 as an example. For each $2 \le i \le k$, since $[z_1]_\rho$ is an equivalence class distinct from $[z_i]_\rho$, there exists a $\mathcal{G}_{\mathfrak{N}}$-wff ψ_i such that $\mathfrak{M}_{\mathfrak{N}}, z_i \models \psi_i$, but $\mathfrak{M}_{\mathfrak{N}}, z_1 \models \neg\psi_i$. If $m_1([z_1]_\rho) = n_1 > 0$, then there exist exactly n_1 α_j-successors of x that are $\mathcal{G}_{\mathfrak{N}}$-equivalent to z_1. Thus, $\mathfrak{M}_{\mathfrak{N}}, x \models (j)_{n_1}(\psi \wedge \bigwedge_{i=2}^k \neg\psi_i)$ for all ψ satisfied in z_1. Since $(x, y) \in \rho$, it is that $\mathfrak{M}_{\mathfrak{N}}, y \models (j)_{n_1}(\psi \wedge \bigwedge_{i=2}^k \neg\psi_i)$ for all ψ satisfied in z_1. This means there are exactly $n_1\alpha_j$-successors of y that satisfy $\bigwedge_{i=2}^k \neg\psi_i$ (by letting $\psi = \top$). Thus, these n_1 α_j-successors of y must satisfy all ψ satisfied in z_1 and be $\mathcal{G}_{\mathfrak{N}}$-equivalent to z_1. Since no other α_j-successors of y can be $\mathcal{G}_{\mathfrak{N}}$-equivalent to z_1 (otherwise there would be more than $n_1\alpha_j$-successors of y satisfying $\bigwedge_{i=2}^k \neg\psi_i$), $m_2([z_1]_\rho) = n_1$. In the same way, we can prove that if $m_2([z_1]_\rho) > 0$, then $m_1([z_1]_\rho) > 0$. Thus, $m_1([z_1]_\rho) = m_2([z_1]_\rho)$. Similarly, we can prove $m_1([z_i]_\rho) = m_2([z_i]_\rho)$ for $2 \le i \le k$. Consequently, we have $m_1 = m_2$ and $[\![N^+_{\alpha_j}(x)]\!]_\rho = [\![N^+_{\alpha_j}(y)]\!]_\rho$. The proof of $[\![N^-_{\alpha_j}(x)]\!]_\rho = [|N^-_{\alpha_j}(y)|]_\rho$ is analogous, except that the modality $(j)_{n_1}$ is replaced by $(j^-)_{n_1}$. □

Observational Equivalences

We have presented the logical characterizations of three kinds of positional equivalence. The theoretical implication of these characterizations is that positional equivalences based on relational granulation can be transformed into indiscernibility relations based on a set of binary attributes. Let $\mathfrak{N} = (A, (P_i)_{i\in I}, (\alpha_j)_{j\in J})$ be a social network and let F denote $\mathcal{H}_{\mathfrak{N}}$, $\mathcal{L}_{\mathfrak{N}}$, or $\mathcal{G}_{\mathfrak{N}}$. Then, (A, F) is a (generalized) data table such that IND(F) is the corresponding positional equivalence. However, from a practical viewpoint, since both $\mathcal{L}_{\mathfrak{N}}$ and $\mathcal{G}_{\mathfrak{N}}$ are infinite sets, it is generally difficult (if not impossible) to decide IND(F) based on such a data table representation. This motivates us to define a new kind of positional equivalence based on observations.

Let $\mathfrak{N}$ be a social network and F be a *finite* set of wffs in an appropriate logic language. Then, two actors, x and y, are *F-observationally equivalent*, denoted by $x \cong_F y$, if for any wff $\varphi \in F$, x satisfies φ iff y satisfies φ. The wffs in F denote the observable properties of the actors in the network. Due to the limited capability of the observer, the set of observable properties is usually finite. Depending on the observable properties chosen, we can focus on different parts of the social network. For example, positional analysis for a genealogical study and a political investigation

may have to consider different relationships between actors. As shown by the above results, the choice of an appropriate logic language is also crucial, since it may result in different notions of positional equivalence.

A recent approach to structural equivalence that conforms with our definition of observational equivalence is reported in [28]. In that approach, a set of institutions E is given and each actor may belong to several institutions. Two actors are structurally equivalent if they belong to the same set of institutions. To avoid confusion with the structural equivalence defined in Sect. "Positional Analysis – Relational Granulation", we call the equivalence relation defined in [28] *institutional equivalence*. Note that institutional equivalence is in fact the same as structural equivalence defined in Sect. "Positional Analysis – Relational Granulation", if we consider the universe as the union of A and E and represent the network as a bipartite graph. A place (social position) is then defined as an equivalence class of actors (or equivalently, the set of institutions that those actors belong to). Here, our set of observation sentences corresponds to the set of institutions in [28]. While institutions are primitive entities, our observation sentences are formed from the social relations and attributes of actors. In fact, a wff in F can also be regarded as an intensional definition of an institution. In this sense, observational equivalence is a kind of intensionally institutional equivalence.

Conclusion and Future Directions

This article starts with an introduction to rough set theory – the basic theory of functional granulation. We then review several notions of positional equivalence in social network analysis and consider them as relational granulation from the GrC perspective. By using modal logics as a bridge, we transform the relational granulation-based positional equivalences into functional granulation. However, the transformation is mainly theoretically interesting, since the set of functional attributes used to granulate a social network may be infinite. This practical consideration motivates us to propose the notion of observational equivalence. It conforms with an alternative definition of structural equivalence, which we call institutional equivalence to avoid confusion.

More generally, both observational equivalence and institutional equivalence can be analyzed in the framework of formal concept analysis (FCA) [13]. FCA is a conceptual knowledge representation and data analysis method in which a *formal context* is defined as a triple (A, F, R), where A and F are sets of objects and attributes respectively; and $R \subseteq A \times F$ is a binary relation between A and F, called the incidence of the formal context. For $A' \subseteq A$ and $F' \subseteq F$, let us denote $R(A') = \{f \in F \mid \forall a \in A', (a, f) \in R\}$ and $R^{-}(F') = \{a \in A \mid \forall f \in F', (a, f) \in R\}$. Then, a *formal concept* is defined as a pair (A', F') such that $A' \subseteq A$, $F' \subseteq F$, $R(A') = F'$ and $R^{-}(F') = A'$. Here, we call A' the extent and F' the intent of the formal concept. The application of FCA to sociology has been proposed in [12]. In that application, the formal context (also called the Galois lattice) is used to model two-mode network data, where the set of actors is a mode of the data, the set of events is the other mode of data, and the relation between the two modes is the participation relationship between the actors and events. When we regard A as the set of actors, F as the set of observation sentences (resp. institution), and R as the satisfaction (resp. membership) relation, a formal context can be constructed out of a social network. However, a social position (or place) defined by observational or institutional equivalences may not be the extent of a formal concept. Thus, it is tempting to define a social position alternatively as a formal concept of such formal contexts. Unlike other positional analysis techniques that define a social position as an equivalence class of some positional equivalence, the extents of different formal concepts may overlap. This means that social positions may not be disjoint in such a definition. The theoretical and practical implications of such positional analysis requires further research.

Another future research direction concerns the practical computation of positional equivalences. As mentioned earlier, when a social network is transformed into a data table (A, F) for positional analysis and F is infinite, it is difficult to decide if two actors are IND(F)-equivalent by exhaustively comparing the satisfaction of all wffs in x and y. While observational equivalence provides a practical approximation of the computation, we would like to know if precise computation is possible. To answer this question, the inference of the *most specific concept* (msc) in description logics [2] may be helpful. Description logics are a family of knowledge representation formalisms that have strong connections with modal logics [3,23]. Indeed, the description languages $\mathcal{ALC}$ and $\mathcal{ALCN}$ correspond to the sublanguages of PMML and GML respectively [3,23]. Given a network represented in a description language, the msc of an actor x, denoted by $msc(x)$, corresponds to a wff that satisfies x and implies any wffs that satisfy x. Thus, if $msc(x)$ and $msc(y)$ are logically equivalent, then x and y will satisfy the same set of wffs, i. e., $(x, y) \in \mathrm{IND}(F)$. The problem for researchers is to find logical languages that can characterize the aforementioned positional equivalences adequately and whose msc problems can be solved effectively.

In summary, we have proposed a logical approach to positional analysis in social networks from a GrC perspective. The approach facilitates a connection between more complicated relational granulation and simpler functional granulation. While there remain problems to be solved in the research programme, the approach seems promising for cross-fertilization between the fields of GrC and SNA.

Bibliography

1. Areces C, ten Cate B (2007) Hybrid logics. In: Blackburn P, Van Benthem J, Wolter F (eds) Handbook of Modal Logic. Elsevier, Amsterdam, pp 821–868
2. Baader F, Küsters R (2007) Nonstandard inferences in description logics: the story so far. In: Gabbay DM, Goncharov SS, Zakharyaschev M (eds) Mathematical Problems from Applied Logic I: Logics for XXIst Century. Springer, Berlin
3. Baader F, Nutt W (2002) Basic description logics. In: Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF (eds) Description Logic Handbook. Cambridge University Press, Cambridge, pp 47–100
4. Blackburn P (2000) Representation, reasoning, and relational structures: a hybrid logic manifesto. Logic J IGPL 8(3):339–625
5. Blackburn P, de Rijke M, Venema Y (2001) Modal Logic. Cambridge University Press, Cambridge
6. Blackburn P, Seligman J (1995) Hybrid languages. J Log Lang Inf 4:251–272
7. Borgatti SP, Everett MG (1989) The class of all regular equivalences: algebraic structure and computation. Soc Netw 11(1):65–88
8. Boyd JP, Everett MG (1999) Relations, residuals, regular interiors, and relative regular equivalence. Soc Netw 21(2):147–165
9. Chellas BF (1980) Modal Logic: An Introduction. Cambridge University Press, Cambridge
10. Everett MG, Borgatti SP (1994) Regular equivalences: general theory. J Math Sociol 18(1):29–52
11. Fan TF, Liau CJ, Liu DR, Tzeng GH (2006) Granulation based on hybrid information systems. In: Proc of the 2006 IEEE International Conference on Systems Man and Cybernetics, Taipei, pp 4768–4772
12. Freeman LC, White DR (1993) Using Galois lattices to represent network data. Sociol Methodol 23:127–146
13. Ganter B, Wille R (1999) Formal Concept Analysis: Mathematical Foundations. Springer, Berlin
14. Gargov G, Goranko V (1993) Modal logic with names. J Philos Log 22(6):607–636
15. Hanneman RA, Riddle M (2005) Introduction to Social Network Methods. University of California, Riverside
16. Horrocks I, Hustadt U, Sattler U, Schmidt R (2007) Computational modal logic. In: Blackburn P, Van Benthem J, Wolter F (eds) Handbook of Modal Logic. Elsevier, Amsterdam, pp 181–245
17. Kripke S (1959) A completeness theorem in modal logic. J Symb Log 24(1):1–14
18. Kripke S (1963) Semantic analysis of modal logic I: normal propositional calculi. Z Math Log Grundl Math 9:67–96
19. Lerner J (2005) Role assignments. In: Brandes U, Erlebach T (eds) Network Analysis. LNCS, vol 3418. Springer, Berlin, pp 216–252
20. Liau CJ, Lin TY (2005) Reasoning about relational granulation in modal logics. In: Proc of the First IEEE International Conference on Granular Computing, Beijing, pp 534–558
21. Lorrain F, White HC (1971) Structural equivalence of individuals in social networks. J Math Sociol 1:49–80
22. Marx M, Masuch M (2003) Regular equivalence and dynamic logic. Soc Netw 25(1):51–65
23. Nardi D, Brachman RJ (2002) An introduction to description logics. In: Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF (eds) Description Logic Handbook. Cambridge University Press, Cambridge, pp 5–44
24. Passy S, Tinchev T (1985) Quantifiers in combinatory PDL: completeness, definability, incompleteness. In: Fundamentals of Computation Theory FCT 85. LNCS, vol 199. Springer, Berlin, pp 512–519
25. Passy S, Tinchev T (1991) An essay in combinatory dynamic logic. Inf Comput 93:263–332
26. Pawlak Z (1982) Rough sets. Int J Inf Comput Sci 11(15):341–356
27. Pawlak Z (1991) Rough Sets – Theoretical Aspects of Reasoning about Data. Kluwer, Dordrecht
28. Pizarro N (2007) Structural identity and equivalence of individuals in social networks: Beyond duality. Int Sociol 22(6):767–792
29. Prior A (1967) Past, Present and Future. Oxford University Press, London
30. Prior A (1968) Now. Nous 2:101–119
31. Scott J (2000) Social Network Analysis: A Handbook, 2nd edn. SAGE Publications, London
32. Sierpinski W, Krieger C (1956) General Topology. University of Toronto Press, Toronto
33. van der Hoek W (1992) On the semantics of graded modalities. J Appl Non-Class Log 2(1):81–123
34. Wasserman S, Faust K (1994) Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge
35. White DR, Reitz KP (1983) Graph and semigroup homomorphisms on netwoks and relations. Soc Netw 5(1):143–234
36. Zadeh LA (1979) Fuzzy sets and information granularity. In: Gupta N, Ragade R, Yager R (eds) Advances in Fuzzy Set Theory and Applications. North-Holland, Amsterdam, pp 3–18
37. Zadeh LA (1997) Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets Syst 19:111–127

Social Network Visualization, Methods of

Linton C. Freeman
Department of Sociology and Institute for Mathematical Behavioral Science, School of Social Sciences, University of California, Irvine, USA

Article Outline

Glossary
Definition of the Subject
Introduction

Visualization in Social Network Analysis
Images Based on One Mode Undirected Relations
Images Based on One Mode Directed Relations
Images Based on Two Mode Relations
Images Based on One or Two Mode Data Matrices
Future Directions
Bibliography

Glossary

Adjacent A node is adjacent to another if there is an edge connecting them.

Arrow A line directed from one node to another.

Binary relation A two valued yes/no or on/off relation.

Bipartite graph A graph, $B = \langle N, E \rangle$ where N is a finite set of nodes and E is a collection of pairs of nodes in which N is partitioned into two disjoint subsets, N_1 and N_2, and no edge in E has both end points in the same subset.

Blockmodeling A procedure for clustering actors such that the actors in each cluster share similar patterns of ties both within and between clusters.

Connected Any two nodes in a graph are said to be connected if there is a path from one to the other; a graph is connected if there is a path connecting every pair of nodes.

Cycle Any path that begins and ends at the same node.

Digraph A directed graph.

Directed graph A graph $D = \langle N, A \rangle$ where N is a finite collection of nodes and A is a set of pairs linked by directed lines or arrows.

Directed line A line going from a node to another representing a non-reciprocated link.

Edge A line connecting two nodes representing a reciprocated link.

Edge labeled graph A graph in which at least two kinds of connections between nodes are identified.

Formal concept analysis A method of data analysis based on Galois lattice structure.

Galois lattice A dual structure that displays the dependencies of both objects and their properties.

Geodesic The shortest path between two nodes.

Graph A graph $G = \langle N, E \rangle$ where N is a finite set of nodes and E is a collection of pairs of nodes represented as edges.

Hyperedge An edge in a hypergraph that can enclose more than two nodes.

Hypergraph A hypergraph, $F = \langle N, H \rangle$, consists of a set of nodes N and a collection of hyperedges, H.

Indegree The indegree of a node is the number of directed lines it receives.

Irreflexive A relation in which no edge connects any node with itself.

Multidimensional scaling A search procedure designed to represent an observed set of proximities or distances in a small number of dimensions.

Node A point in a graph.

One mode matrix A data matrix in which the rows and columns both represent the same objects.

Outdegree The outdegree of a node is the number of directed lines it sends out.

Path A path is a sequence of nodes and edges beginning with a node that has an edge connecting it to the next node in the sequence and so on.

Path length The length of a path connecting two nodes is the number of edges it contains.

Permutation A reordering of the rows, columns, or rows and columns of a matrix.

Principle diagonal The set of cells in a square matrix that runs from the upper left to the lower right.

Relation A collection of ordered or unordered pairs of nodes.

Singular value decomposition an algebraic procedure that decomposes a data matrix into its "basic structure".

Sociometry An early version of social network analysis introduced by Jacob Moreno and Helen Jennings.

Spring embedder A kind of multidimensional scaling based on a model in which it is assumed that nodes are connected by springs that pull and push on them.

Symmetric A relation in which if a node a is adjacent to another, b, then b is adjacent to a.

Tree A graph is a tree if it is connected and contains no cycles.

Two mode matrix A data matrix in which the rows and columns represent different objects.

Definition of the Subject

Social network visualization refers to the practice of constructing pictorial images of the connections linking social actors. The use of such images provides two benefits. It allows investigators to gain new insights into the patterning of social connections, and it helps investigators to communicate their results to others.

Introduction

Social network analysis did not emerge as a systematic field of research until early in the twentieth century [1]. But visual images of social networks were produced more than a millennium earlier. The earliest of these images that I have uncovered was produced in Spain in the middle of

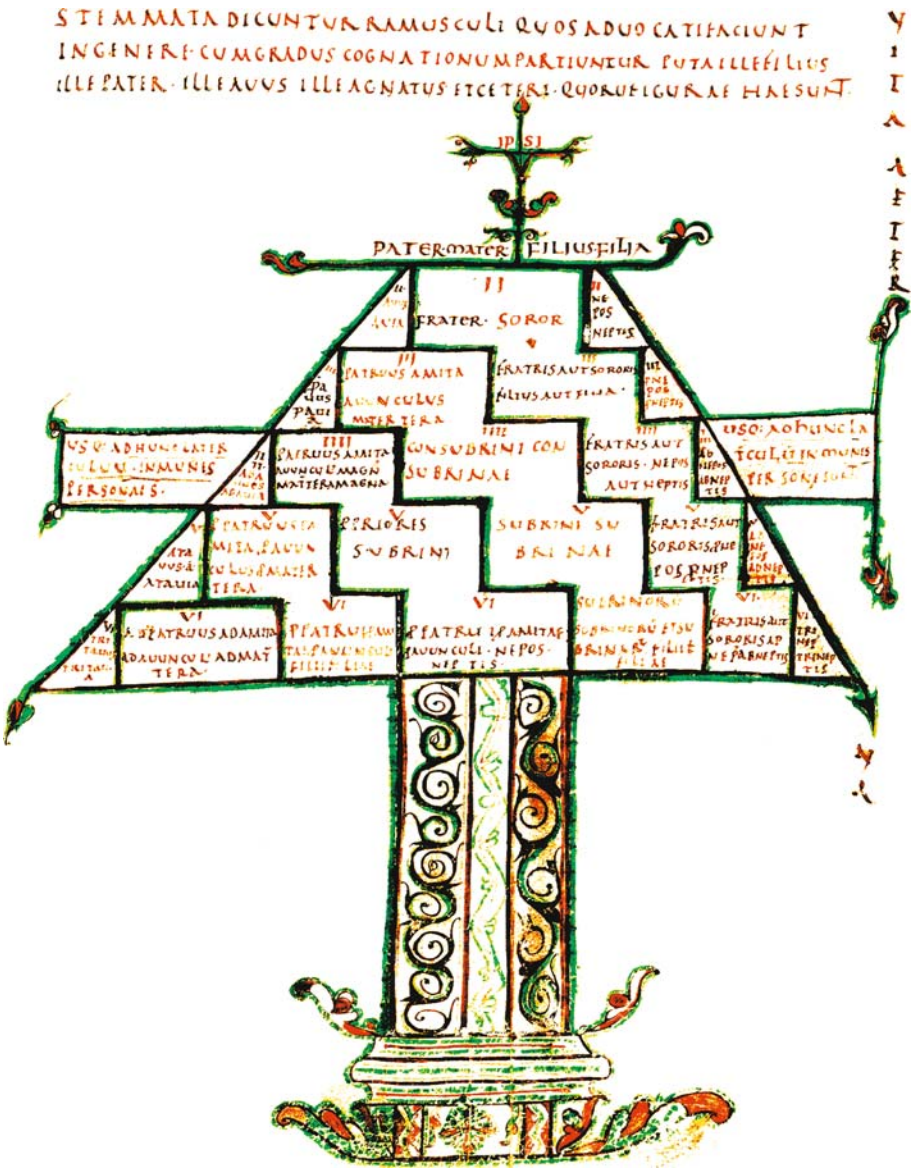

Social Network Visualization, Methods of, Figure 1
Tree of consanguinity with six degrees of relationship

the ninth century. That image is attributed to the prolific writer and Roman Catholic Saint, Isidore de Séville. It is reproduced here as Fig. 1.

The image shown in Fig. 1 displays relationships based on genealogical descent. From the earliest times, people have been interested in kinship ties – in who is related to whom. This interest is evident in the descent lists found in the Christian bible and in the oral genealogies that were required to be memorized by Hawaiian nobles [2].

The fact that Fig. 1 takes the form of a tree shows that as early as the ninth century people saw the analogy between the branching structure of descent and that of trees. This notion was captured in a mathematical formalization in 1857 by Arthur Cayley [3]. Cayley defined a tree in mathematical graph theoretic terms. Biggs, Lloyd and Wilson (p. 38 in [4]) characterized Cayley's definition by saying that his "… use of the word 'tree' in this context is presumably derived from the diagrammatic form of these graphs, and is akin to the traditional use of the word in describing genealogical or 'family' trees."

The use of trees to depict descent was, of course, continued. As time passed, however, their form became sim-

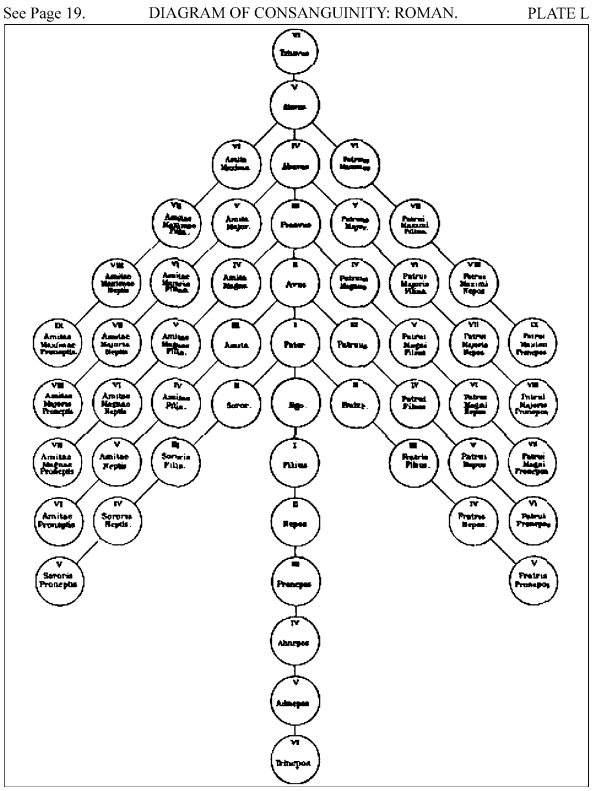

Social Network Visualization, Methods of, Figure 2
Descent in ancient Rome

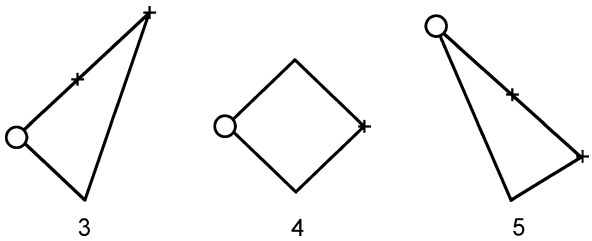

Social Network Visualization, Methods of, Figure 3
Macfarlane's images of two-step marriage prohibitions

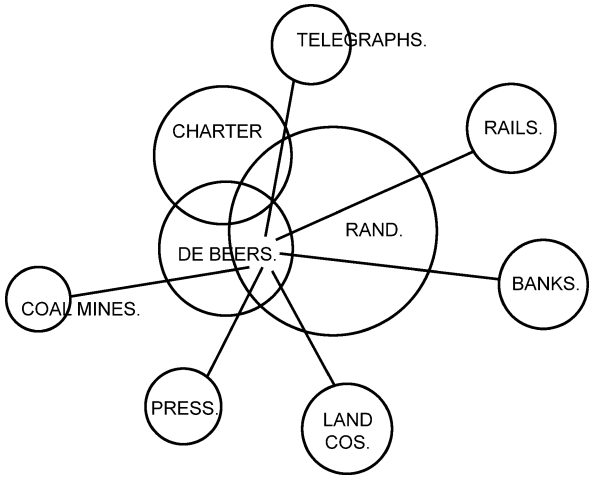

Social Network Visualization, Methods of, Figure 4
Hobson's image of corporate interlocks

plified. Lewis Henry Morgan [5] was an attorney and an anthropologist. He was interested in comparing how different peoples reckoned kinship and in 1871 he published a mammoth work containing a collection of kinship trees. Each tree depicted descent as conceived by a society somewhere in the world. Morgan's trees are quite simple. Figure 2 shows descent as it was reckoned in ancient Rome.

Twelve years later a mathematician-physicist, Alexander Macfarlane [6], produced a different kind of graphic image based still on kinship. Macfarlane set out to examine British marriage prohibitions and he represented them both algebraically and visually. His visual images depict males using plus signs (+) and females with circles (o). Earlier generations he placed higher on the page. Descent is shown by lines connecting points. A short line crossing a descent line indicates another person, of either sex, in an intermediate generation. And the lowest point is always the prohibited offspring.

The illustration shown in Fig. 3 displays all the two-step marriage relations that are prohibited by British law. The left image shows that a male may not marry his granddaughter. The middle image shows that he may not marry his sister. And the right image shows that he may not marry his grandmother. Or, put the other way, a woman may not marry her grandfather, her brother or her grandson.

Macfarlane's paper also included algebraic expressions that captured all of the same marriage prohibitions. But Sir Francis Galton [7], who attended Macfarlane's presentation, declared that his "diagrammatic form" seemed "the most distinctive and self-explanatory" of the two treatments.

Finally, in 1894, John Hobson [8] produced a visual image of a social network that was not based on kinship. He had collected two mode (corporation by director) data on interlocking corporate directorates. He reasoned that, to the degree that corporations shared directors, they could be expected to cooperate and work together.

Hobson's illustration was designed to show the interlock among, as he put it, "the small inner ring of South African finance." Corporations are depicted as circles, and interlock is shown by overlapping or by a line connecting two circles. Hobson's image is reproduced here as Fig. 4.

The important feature of this image is that it displays a connection linking more than two corporations. Hobson's data showed that three corporations, Charter, Rand and De Beers, all shared directors in common. And, at the same time, Rand and De Beers also both shared directors with coalmines, telegraphs, rails, and others. The overlaps in his image allowed him to display which companies shared with which others.

It is clear, then, that a concern with connections among social actors and the use of visual images have a long history of intimate association. It should come as no surprise therefore that images played an important part in the development of social network analysis when it did emerge as an organized field of research.

Visualization in Social Network Analysis

In the book cited above (p. 3 in [1]), I described the modern science of social network analysis as possessing four defining properties. They were:

1. It embodies ideas about the importance of social ties linking social actors.
2. It collects data reflecting those ties.
3. It involves the use of graphic imagery.
4. It employs mathematical and/or computational models.

Pre-network research often included one or two of those properties, but in the late 1920s each of two independent research teams came up with efforts that included all four.

One took place in the early 1930s. It involved a psychiatrist, Jacob L. Moreno, and a psychologist, Helen H. Jennings. Together, they developed an approach they called "sociometry." They reported two huge studies, both focused on examining the structure of social ties. One was conducted among prisoners at Sing Sing Prison in Ossining, New York [9] and the other among young delinquents at the New York State Training School for Girls in Hudson, New York [10].

Both Moreno–Jennings studies involved the extensive use of graphic images. The image shown in Fig. 5 was included in their report on the research at Sing Sing prison (Moreno, 1932). In that figure, individuals or other kinds of social actors are represented as points or *nodes* and links between pairs of social actors are lines or *edges* connecting pairs of nodes. In Fig. 5 Moreno was concerned with the positions of individuals and the patterning of their ties. As he put it, the individuals at the top and the bottom were "dominant" and the image showed that those dominant individuals were linked both "directly" and "indirectly".

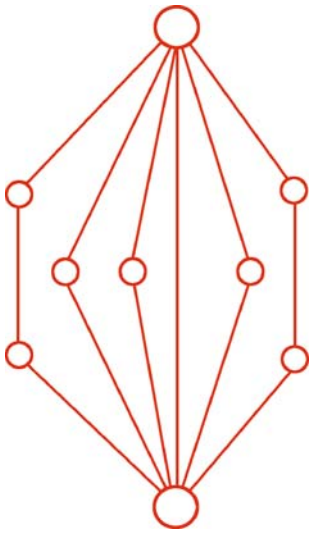

Social Network Visualization, Methods of, Figure 5
Image of a Pattern of Linkages

Most of the data collected by Moreno and Jennings involved asking individuals whom they liked or disliked. In data of that sort, choices are seldom reciprocated. So Moreno and Jennings drew lines with arrowheads to reveal who chose whom. Mutual choices were drawn without arrowheads and they also included a small line bisecting the main line connecting the two nodes.

Moreno and Jennings often required subjects to report both their likes and their dislikes. By using different colors, red for likes and black for dislikes, a single image could display both. The image shown in Fig. 6 was published in the Moreno–Jennings report on the Hudson School (Moreno, 1934). It depicts positive and negative choices among 13 members of an American football team. Moreover, it contains another innovation. The various team members are placed in the drawing in approximately the same relative locations that they occupied on the football field. That arrangement shows the players' positions and permits the viewer to evaluate the impact of physical proximity on the patterning of social linkages.

Figures of the sort used by Moreno and Jennings had a major impact on the style of graphic imagery used subsequently in social network analysis. For the most part, social network analysts have represented social actors as nodes and links between actors as edges or as directed lines with arrowheads.

The second introduction of the social network approach also occurred in the early 1930s. An anthropologist, W. Lloyd Warner, and a collection of his colleagues and students at Harvard, conducted three elab-

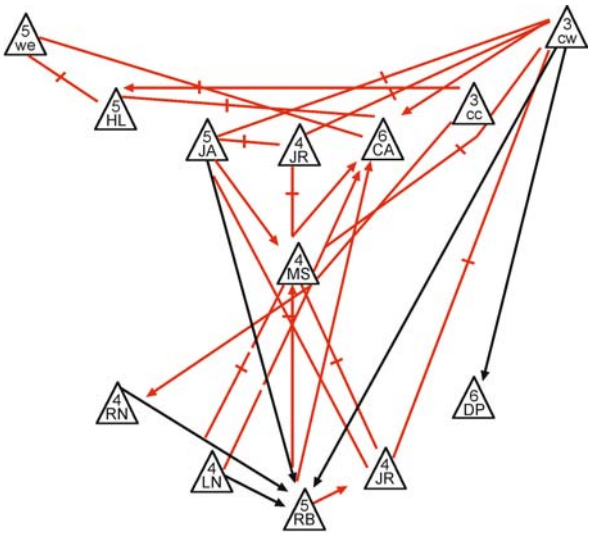

Social Network Visualization, Methods of, Figure 6
Positive and Negative Choices in a Football Team

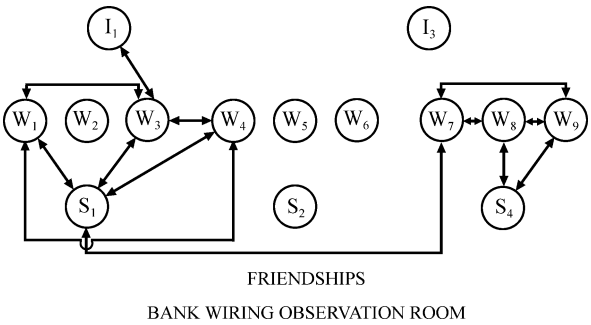

Social Network Visualization, Methods of, Figure 7
Friendships linking factory workers

orate network analytic projects. One was a study of an industrial factory, the Western Electric plant in Cicero, Illinois [11]. The other two were studies of communities: one focused on a New England town, Newburyport, Massachusetts [12], and the other on a southern town, Natchez, Mississippi [13].

The image shown in Fig. 7 was produced as part of the factory study. It displays observed friendship ties among pairs of individuals who worked together in the same workroom. It was drawn using nodes and two-headed lines instead of edges, but it is very similar to the images produced by Moreno and Jennings. In addition, as in Fig. 6 above, the impact of physical space was displayed; workers were placed in the drawing in positions that reflected the locations of their workstations.

In reporting their study of Newburyport, Warner and Lunt used the kind of drawing of overlapping circles that

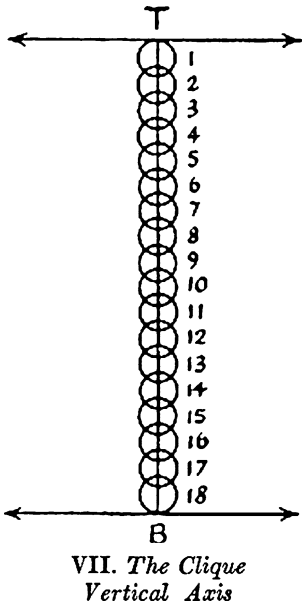

Social Network Visualization, Methods of, Figure 8
An idealized pattern of overlapping 'cliques'

Hobson had used to construct Fig. 4. But here that image was introduced, not to describe data, but to propose an idea they had about social structure. The diagram in Fig. 8 represents the investigators' idealized version of the expected structure of overlaps among subgroups in the presence of social class. The idea is that only subgroups that are close to one another in class ranking are likely to have overlapping memberships.

In their study of Natchez Davis, Gardner and Gardner [13] employed the same diagrammatic form to display two mode data reflecting on the earlier Newburyport hypothesis. Figure 9 shows subgroups of black males and their overlaps. In that image the men are arranged in terms of both social class and age. Both, it turned out, provided important bases for grouping.

Finally, in that same report, Davis, Gardner and Gardner also introduced an entirely different kind of social network image. Like Hobson, they had collected two mode data. Eighteen women were designated in the rows of their data matrix and fourteen social events were depicted in the columns. That matrix is reproduced here as Fig. 10.

The data shown in Fig. 10 were all collected during a single year. But, by examining the column headings, it is clear that Davis and his colleagues did not arrange the social events according to the dates upon which they took place. Instead, they listed both the events and the women who attended them in such a way that the arrangement itself suggests that these women were organized into two

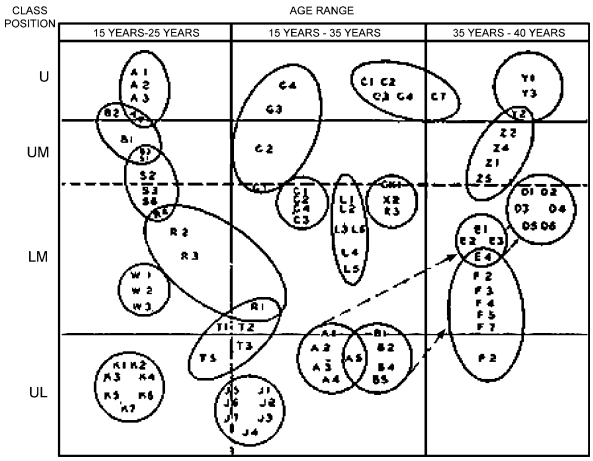

Social stratification of a group of colored cliques: ages 15–60

Social Network Visualization, Methods of, Figure 9
Stratification, age and overlapping groups

groups. The two groups overlap, but for the most part they are distinct. Most of the women in the top half of the matrix attended the leftmost five events. And most of the women in the bottom half attended the rightmost five events. The middle four events apparently brought both groups of women together.

This arrangement of women and events was self-consciously produced by the authors. Davis, Gardner and Gardner were convinced that these women were organized into two groups and they presented their data matrix in a way that would illustrate that conclusion. The interesting thing is that these authors never commented explicitly about how they had rearranged the columns and rows in their matrix. They simply organized their display in a way that would make the point.

From the outset, then, four kinds of images have played important parts in the development of social network analysis. These first network graphics included drawings displaying (1) one mode undirected relations, (2) one mode directed relations, (3) two mode relations and (4) one or two mode data matrices. A few other kinds of network images have been used since then, but the four originals – particularly those based on one mode undirected and one mode directed relations – still dominate the field. In the next four sections we will examine the four original kinds of images and how their use has evolved in the social network context.

Images Based on One Mode Undirected Relations

Mathematically speaking, the node and edge images introduced by Moreno and Jennings in Fig. 5 are *graphs*. A graph $G = \langle N, E \rangle$ where N is a finite set of nodes and E is a collection of pairs of nodes. In graph visualizations, a pair of nodes in E is presented as a line connecting the two nodes in question. Two nodes are called *adjacent* if there is an edge directly connecting them to each other. A graph embodies a *binary* (yes/no or on/off) relation that is *irreflexive* (no node is adjacent to itself) and *symmetric* (if a node a is adjacent to another, b, then b is adjacent to a).

A *path* is a sequence of nodes and edges, beginning with a node that has an edge connecting it to the next node in the sequence. The *length* of a path between two nodes

Names of Participants of Group I	Code Numbers and Dates of Social Events Reported in *Old City Herald*													
	(1) 6/27	(2) 3/2	(3) 4/12	(4) 9/26	(5) 2/25	(6) 5/19	(7) 3/15	(8) 9/16	(9) 4/8	(10) 6/10	(11) 2/23	(12) 4/7	(13) 11/21	(14) 8/3
1. Mrs. Evelyn Jefferson	×	×	×	×	×	×		×	×					
2. Miss Laura Mandeville	×	×	×		×	×	×	×						
3. Miss Theresa Anderson		×	×	×	×	×	×	×	×					
4. Miss Brenda Rogers	×		×	×	×	×	×	×						
5. Miss Charlotte McDowd			×	×	×		×							
6. Miss Frances Anderson			×		×	×		×						
7. Miss Eleanor Nye					×	×	×	×						
8. Miss Pearl Oglethorpe						×		×	×					
9. Miss Ruth DeSand					×		×	×	×					
10. Miss Verne Sanderson							×	×	×			×		
11. Miss Myra Liddell								×	×	×		×		
12. Miss Katherine Rogers								×	×	×		×	×	×
13. Mrs. Sylvia Avondale							×	×	×	×		×	×	×
14. Mrs. Nora Fayette						×	×		×	×	×	×	×	×
15. Mrs. Helen Lloyd							×	×		×	×	×		
16. Mrs. Dorothy Murchison								×	×					
17. Mrs. Olivia Carleton									×		×			
18. Mrs. Flora Price									×		×			

Fig. 3.—Frequency of interparticipation of a group of women in Old City, 1936—Group I.

Social Network Visualization, Methods of, Figure 10
The Davis, Gardner and Gardner data on women's attendance at social events

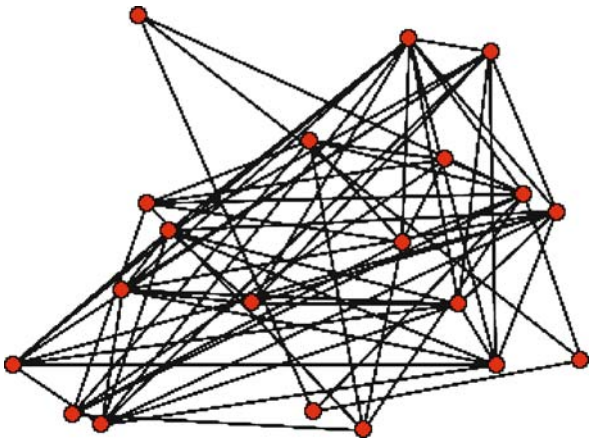

Social Network Visualization, Methods of, Figure 11
Links in the network of a homeless woman I

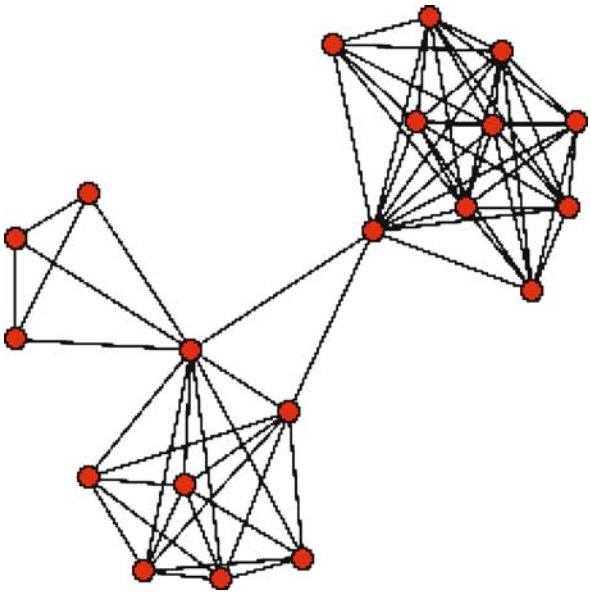

Social Network Visualization, Methods of, Figure 12
Links in the network of a homeless woman II

is the number of edges it contains. And the shortest path connecting two nodes is called the *geodesic*.

Any two nodes in N are *connected* if there is a path from one to the other. And a whole graph G is connected if every pair of nodes in N is connected. If a path begins and ends at the same node, that path is a *cycle*. Finally, a graph is a *tree* if it is both connected and it contains no cycles.

The image in Fig. 11 is a graph. It is based on data recorded by J. Clyde Mitchell [14] on the social ties among the 19 individuals involved in the personal network of a homeless woman in Britain. I used a program called NetDraw to place the nodes representing individuals in Fig. 11 in random positions. That calls attention to the importance of the locations of points in graphic displays. Given the locations of the points in Fig. 11 it is very difficult for the viewer to see anything interesting in the patterning of this woman's network.

Compare the image in Fig. 11 with that in Fig. 12. Figure 12 was also produced using NetDraw, but this time the points were placed using a *spring embedder* [16]. A spring embedder is a computer algorithm that, in effect, places a spring of unit length between every pair of adjacent nodes and a much longer spring between nodes that are not adjacent. It starts with a random placement of nodes, then the whole apparatus is set in motion and the various springs push and pull until they reach an equilibrium.

The advantage of using a spring embedder is that it does not require the investigator to make ad hoc judgments in locating nodes in a graph. It uses a standard computer algorithm to place the nodes automatically. Thus, every user will get the same result. There are several different spring embedding algorithms. And they are all examples of a more general class of computer algorithms that search for optimal locations for nodes in relatively few dimensions. This general class of search algorithms is called *multidimensional scaling* [17].

An alternative method for placing nodes automatically is grounded in algebra. It is called *singular value decomposition* [18]. Singular value decomposition is not search based. Instead, it uses matrix operations to produce a linear transformation of the data, and thus to position the nodes in one, two or three or more dimensions. There is no guarantee that it will always be effective, but often singular value decomposition provides very good placements of the nodes in few enough dimensions that visualization is possible [19]. A NetDraw image based on singular value decomposition of Mitchell's data is shown in Fig. 13.

The images in Figs. 12 and 13 both show that the whole network is organized into three densely connected groups that are only loosely linked to one another. That is interesting, but it does not tell us anything about the bases for the groupings. By adding a little information, and continuing to use NetDraw, we can transform the graph of Fig. 12 into a *node labeled* graph, see Fig. 14.

Given the labels, we can identify the homeless woman, the "respondent." We can also see how her network is split up. One division includes her original family, another her friends along with her social worker, and the third contains her estranged husband and his family, her in-laws.

Mitchell's report, however, included even more details. It included estimates of the strength of the tie linking each

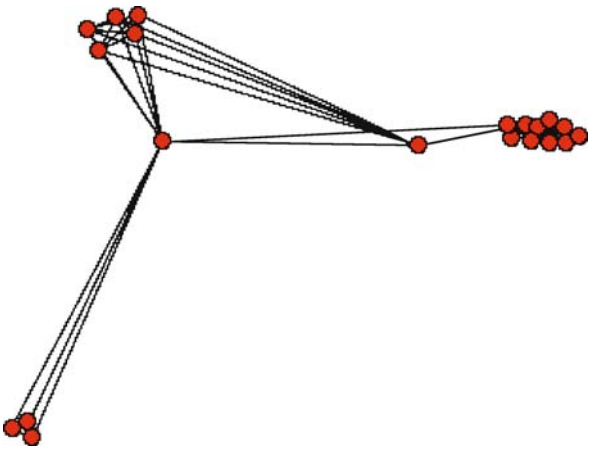

Social Network Visualization, Methods of, Figure 13
Links in the network of a homeless woman III

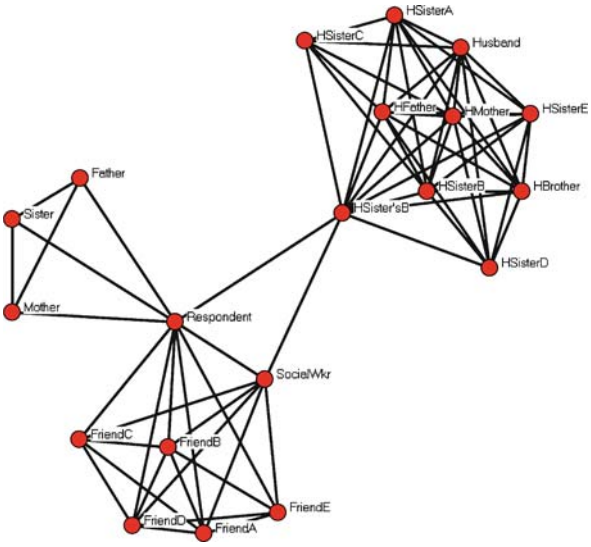

Social Network Visualization, Methods of, Figure 14
Links in the network of a homeless woman IV

of the pairs of individuals. He classified each tie as either strong or weak. We can embody this additional information in our NetDraw image by an adding another component to our graph. Figure 15, then, was produced using the spring embedder, it is node labeled, and, in addition, it is an *edge labeled* graph.

In Fig. 15 strong ties are indicated by wide edges. By examining their patterning, we learn that the individuals within each family are linked together mostly by strong ties, while the homeless woman's friends have fewer strong ties linking them together. This result is not surprising, but it does provide additional insight about the structural position of the woman in question. Clearly, it would be easier for either family to achieve consensus and provide support than it would be for the respondent's loosely connected collection of friends [20].

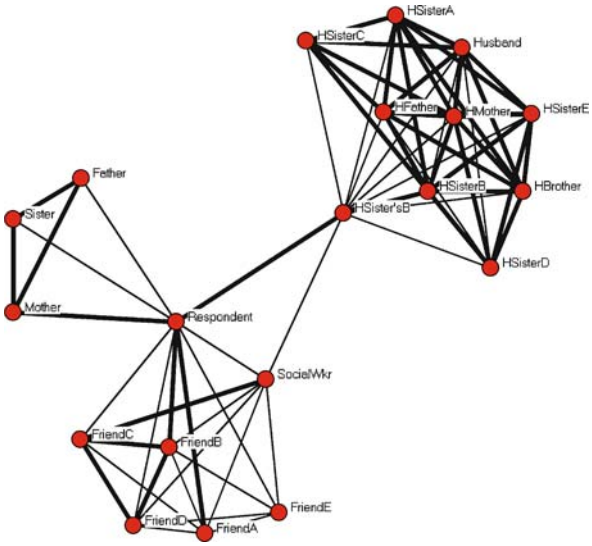

Social Network Visualization, Methods of, Figure 15
Links in the network of a homeless woman V

It should be clear, then, that the placement of nodes and the labeling of both nodes and edges are critical for the ability of a graph to communicate important information. Good images can provide investigators with new insights about the structural properties of the social networks they are studying. And they can, of course, help to communicate the results of social network research to outsiders.

Images Based on One Mode Directed Relations

It was obvious from the outset that these simple graphs would not permit many kinds of displays of interest to social network analysts. Even Moreno and Jennings saw the need to display the direction of choice in their sociograms. The direction of connections can be expressed using *directed graphs* or *digraphs*.

A digraph $D = \langle N, A \rangle$ where N is a finite collection of nodes and A is a set of pairs shown as *directed lines* or *arrows*. When an arrow is directed from node a to node b in a digraph, then a is the *tail* of the arrow and b is the *head*; a is the immediate *predecessor* of b and b is the immediate *successor* of a. The *outdegree* of a node is the number of arrows for which it is the tail and its *indegree* is its number for which it is the head.

In any study that involves social links that are not symmetric, digraphs provide a natural representation. Con-

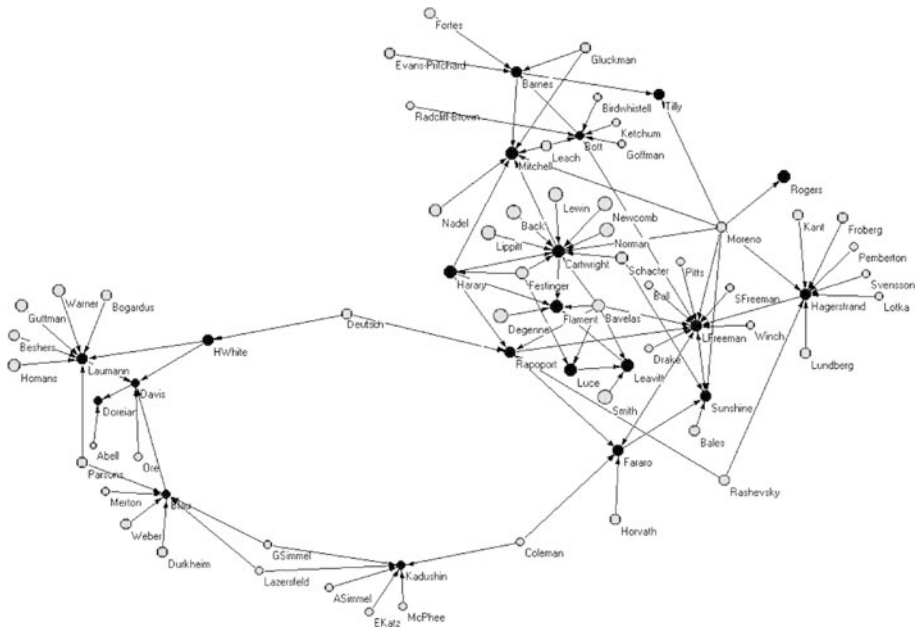

Social Network Visualization, Methods of, Figure 16
Influences on some founders of social network analysis

sider Fig. 16 that was produced by a program called Pajek [21]. In preparing a book on the development of social network analysis, I interviewed a number of the founders. Each was asked to name others who had influenced them to think in network terms. The result is a data set that obviously lacks symmetry.

My interest, however, was with clusters, or blocs, of influentials and nominees. So I placed the nodes using a spring embedder designed by Kamada and Kawai [22]. The resulting figure shows that there seem to be two fairly well defined subgroups, one on the left and one on the right. The two groups are relatively dense but they are only loosely connected together. The people on the left are almost entirely sociologists and those on the right are mostly from other fields. And from the patterning one can suspect that there was some kind of split between these two groups.

For some kinds of data the search for clusters or groups is not appropriate. For example, when we are dealing with data that should embody some sort of ordering, digraph representations are particularly important. To illustrate how digraphs can be used to display ordering, consider the data collected by Forkman and Haskell [23]. They studied several communities, each made up of six domestic hens. In five of these communities the hens formed strict pecking orders in which the top hen pecked all the others; the second pecked all but the top, and so on. Figure 17 shows a visone [24] image of the data from one of those five

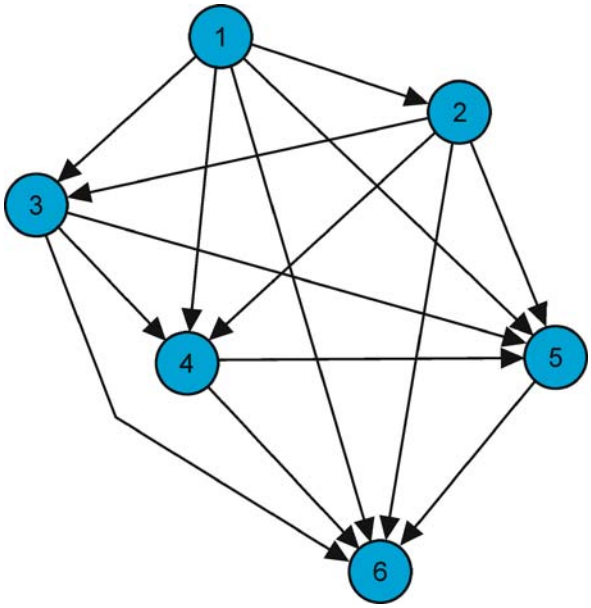

Social Network Visualization, Methods of, Figure 17
Dominance among six hens

communities. There the nodes are arranged, top down, in terms of their outdegrees and the pecking order is obvious.

Often data approach, but do not achieve, a strict order. David Krackhardt [25], for example, collected data on who sought advice from whom among 14 employees in the in-

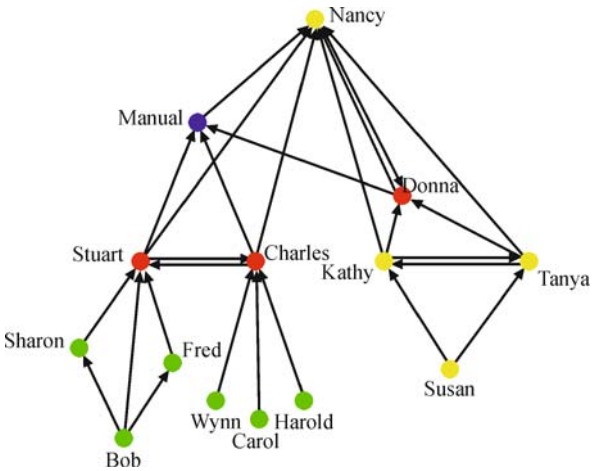

Social Network Visualization, Methods of, Figure 18
Visone image of advice seeking (from [26])

ternal auditing staff of a large company. Krackhardt's data could not be drawn with all the arrows pointing in one direction. So, in Fig. 18 he arranged the individuals in such a way that as many arrows as possible were pointing up. The viewer, then, can immediately see there is an important hierarchical element displayed by these data. From the image, it appears that Nancy is at the top of the advice chain and Bob, Wynn, Carol, Harold and Susan are at the bottom.

There is, however, an important limitation in this figure. Nancy seeks advice from Donna, Donna seeks advice from Manuel and Manuel seeks advice from Nancy. Thus, these three form a directed cycle of advice seeking. Given such a circular arrangement, no possible hierarchy among these three individuals can be established. Any order in which they were arranged would be misleading. In addition, Stuart and Charles cannot be ordered because they chose each other. The same is true for Kathy and Tanya.

The apparent ordering of nodes in Krackhardt's image was imposed by human judgment. There are computer algorithms that can automatically arrange the nodes into a hierarchical form [26]. They are, however, not as well grounded or reliable as multidimensional scaling and singular value decomposition.

Images Based on Two Mode Relations

Any time we deal with a relation that can link more than two social actors, we cannot use graphs or directed graphs. Both graphs and directed graphs can deal only with links between pairs. Two mode data, however, allow for relations that link three or more actors. So, whenever we have two mode data, like that collected by Hobson [8] or Davis, Gardner and Gardner [13] we need another way to construct images.

There are several ways to construct images of two mode data. I will consider three of them in the present section, hypergraphs, bipartite graphs and lattices. Then, in the next section, I will discuss the use of matrix representations for both one mode and two mode data.

Hobson [8] collected two mode data on corporations and their directors. He produced the image shown in Fig. 4 showing corporate interlocks as overlapping areas. Mathematically, images like Hobson's are *hypergraphs*. A hypergraph, $F = \langle N, H \rangle$, consists of a set of nodes N and a collection of *hyperedges*, H. While an edge in an ordinary graph connects two nodes, a hyperedge in a hypergraph may link any arbitrary subset of the nodes in N. Pictorially, hyperedges are represented as boundaries enclosing sets of nodes.

The use of hypergraphs was demonstrated in a recent report by Estrada and Rodríguez-Velázquez [27]. They began with one mode data that showed the patterning of predation among the members of eleven species in a Malaysian rain forest. Their graph, showing who preys on whom, is shown in Fig. 19.

Figure 19 shows which species preys on which other species. But if the investigator is interested, as those who study food webs often are, in defining ecological niches in terms of co-predation, Fig. 19 makes the overall pattern less than obvious. As an alternative we can build a hypergraph.

The matrix shown in Fig. 20 is based on the data in Fig. 19. It was built by considering each of the species in turn as prey. Then all of the species that share each given prey are pooled together. Species 1, 6 and 9 have no prey

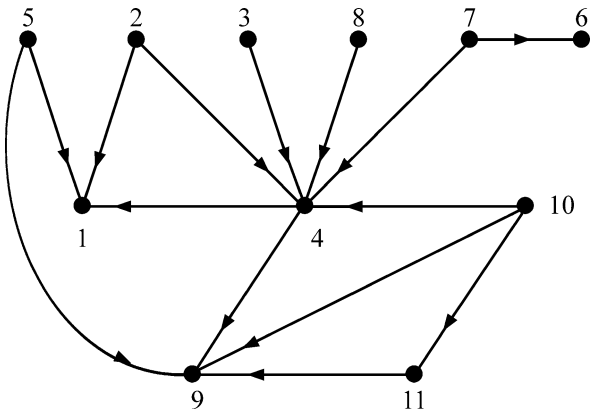

Social Network Visualization, Methods of, Figure 19
Who preys on whom in a Malaysian Rain Forest

Prey on

	4	1	9
2	1	1	0
3	1	0	0
4	0	1	1
5	0	1	1
7	1	0	0
8	1	0	0
10	1	0	1
11	0	0	1

Social Network Visualization, Methods of, Figure 20
Two mode matrix of co-predation

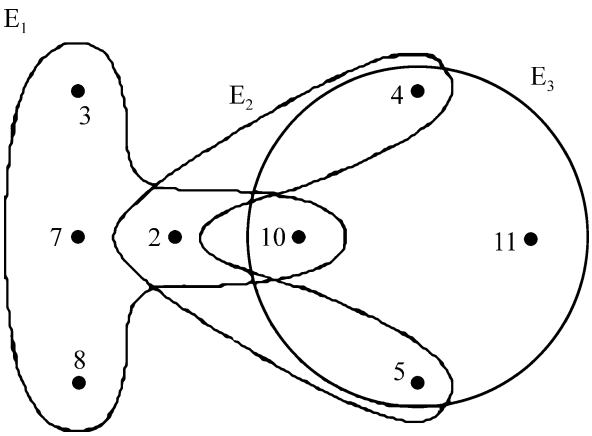

Social Network Visualization, Methods of, Figure 21
Hypergraph of co-predation

in the set. And species 4, 1 and 9 are the targets of co-predation. So the new matrix is two mode. It has the three targets of co-predation as columns and the eight predators as rows.

That matrix is captured visually by the hypergraph in Fig. 19. It immediately reveals that there are three niches. The one labeled E_1 includes all the species who preyed on species 4, E_2 those who preyed on species 1 and E_3 those who preyed on 9. Thus, each edge in Fig. 21 encloses a collection of species that compete directly for at least one prey.

There are, however, other ways to picture two mode data. In a more recent study of corporate interlocks, Joel Levine [29] reported data on the board memberships of seven major American corporations. Those corporations turned out to have ten directors who appeared on two or more of their boards. Levine presented his interlock data using a *bipartite graph*. A bipartite graph, $B = \langle N, E\rangle$ where N is partitioned into two disjoint subsets, N_1 and N_2, and no edge in E has both end points in the same subset. He used singular valued decomposition to place the nodes representing both corporations and board members and produced a bipartite image similar to the one displayed in Fig. 22. I prepared that figure using NetDraw. There, the corporations are shown as red circles and the board members are blue squares. Thus, both the colors and the shapes of the nodes stress the bipartite nature of the graph.

There is still another form of graphic display, one that reveals even more structural information about a two mode data set. It is based on an algebraic procedure called *Galois lattice analysis* or *formal concept analysis* [30,31,32,33]. A Galois or formal concept lattice is defined on an object by property, matrix. Let O be a set of objects and A be a set of attributes. The binary matrix $O \times A$ indicates which objects possess which attributes.

We can define a pair $\langle O_i, A_i\rangle$ such that O_i is a subset of O and A_i is a subset of A and every object in O_i has every attribute in A_i. Moreover, both O and A must be maximal. Thus, for every attribute in A that is not in A_i, there is an object in O_i that does not have that attribute. And for every object in O that is not in O_i, there is an attribute in A_i that the object lacks.

These pairs are dual and they can be partially ordered by inclusion. Given two pairs $\langle O_i, A_i\rangle$ and $\langle O_j, A_j\rangle$ we say that $\langle O_i, A_i\rangle$ is less than $\langle O_j, A_j\rangle$ when O_i is a subset of O_j or, equivalently, when A_j is a subset of A_i. Since all these pairs have unique least upper bounds and greatest lower bounds they form a dual (Galois) lattice.

I will illustrate by considering again the woman by event data collected by Davis, Gardner and Gardner [13]. Let the women (1 through 18) be the objects and the events (A through N) be the attributes. The data, arranged into a Galois lattice by a program called GLAD, are shown in Figure into a lattice in Fig. 23.

The lattice displays the same three classes of events that define the same two groups of women that we saw in Fig. 10. But, in addition to the classes of events and groups of women, we can now see the containment structures of both events and women. To begin with, by following lines up from the bottom we can see which women attended which events. When we get to the top we hit the set of all events and, at the same time, because no woman attended all 14 events, it is also the null set of women.

The uppermost events (E–L) involved the largest sets of women. Other events are contained in the lower inter-

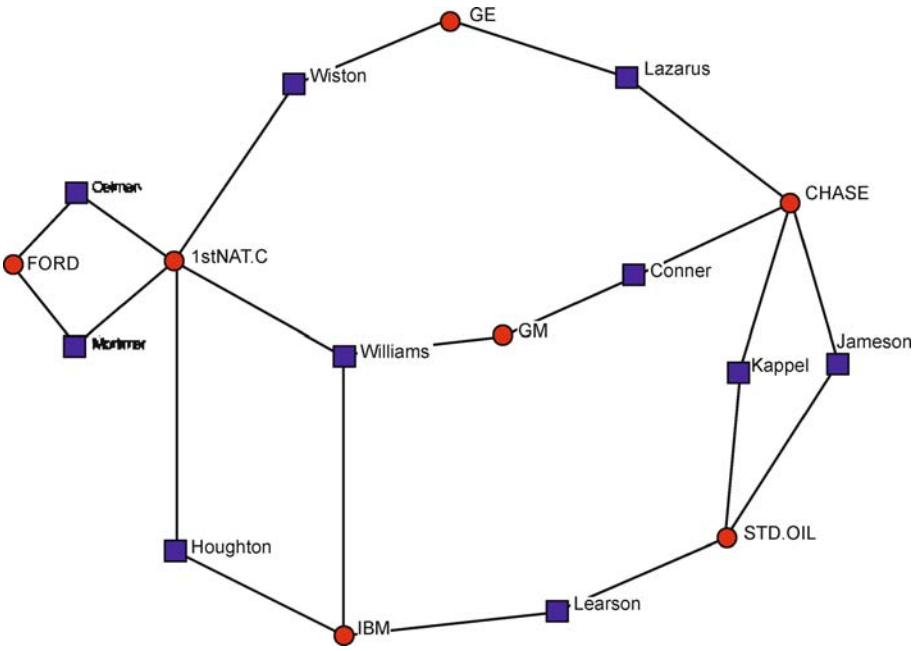

Social Network Visualization, Methods of, Figure 22
A NetDraw Image of Levine's interlock data as a bipartite

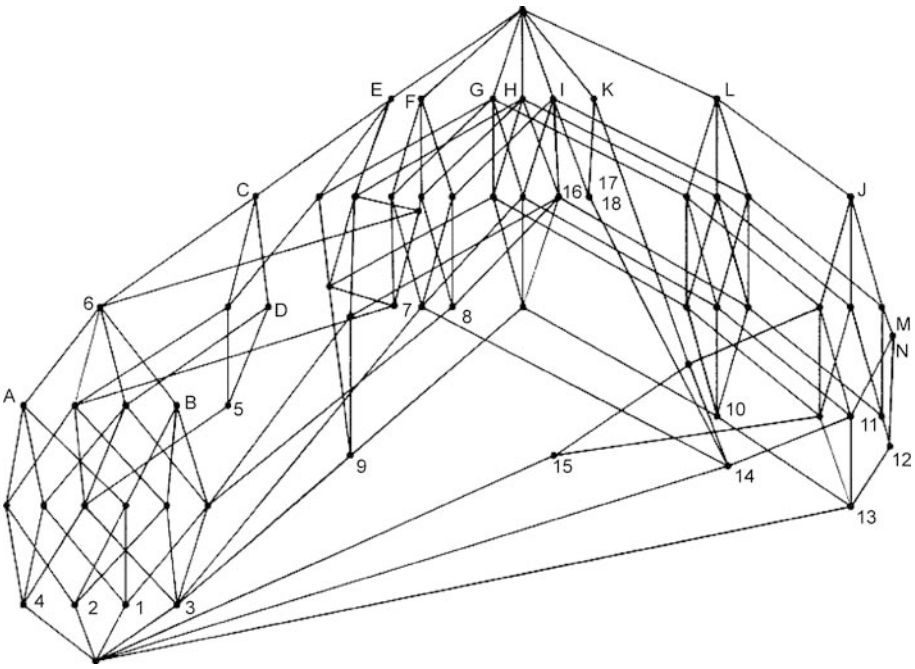

Social Network Visualization, Methods of, Figure 23
The Davis, Gardner and Gardner data as a Galois lattice

sections of these events. Event C, for example, is contained in E; everyone who attended C was present at E. And, at the next lower level, B and D are both contained in C. The events, then, can be seen as varying in their "openness".

At the same time, the figure shows the upward containment structure of the women in terms of their patterns of attendance. Because no event attracted all 18 women, the lowest point represents the set of all women as well as

the null set of events. Then, the lowest set of women (1, 2, 3, 4, 13, 14 and 15) are the "core" attendees, so to speak. The next level contains woman 9 who never attended unless woman 3 was also present, and woman 5 whose attendance depended on that of women 4 and 3. Women 6, 7, 8, 10, 11, 12, 17 and 18 are also at this second level. In some sense, these are all secondary or peripheral participants in these events. And, finally, woman 16 turns out to be a third level participant; she was extremely peripheral. Woman 16 attended events only when secondary attendees 8–12 and core attendees 1, 3 and 13 were all present. All in all, then, the image of the Galois lattice reveals a great deal about the internal structure of attendance.

In this section I have shown three ways of visualizing two mode data. All three of them, however, share one important limitation. That limitation stems from the fact that all three of them can only be used for very small data sets. As the number of cases grows, they all produce images that become increasingly difficult to read.

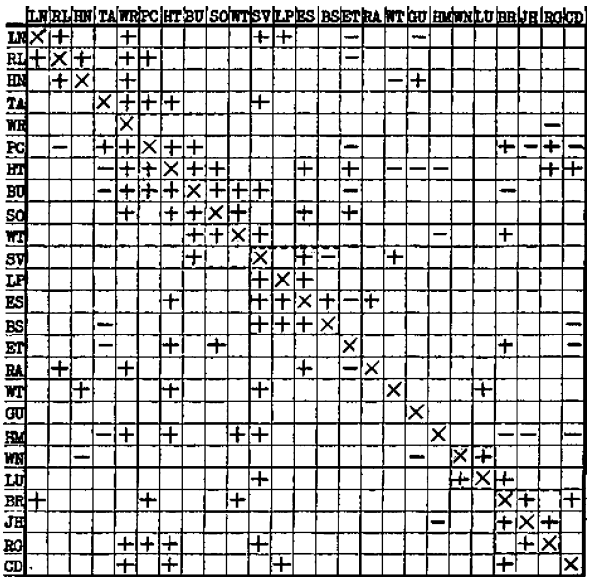

Social Network Visualization, Methods of, Figure 24
The Forsyth and Katz image of sociometric choices

Images Based on One or Two Mode Data Matrices

When Davis, Davis and Gardner [13] first used matrix permutation, they did so without calling attention to the process. But since that first use a number of contributors have suggested procedures explicitly designed to rearrange the rows and columns of matrices. As time has passed, the overall tendency has been to come up with more effective procedures. And, with the introduction of computers, it has become possible to manipulate ever larger matrices. Presently, there is no end in sight.

Matrix permutations, moreover can be used with either one mode or two mode data. Five years after Davis, Gardner and Gardner introduced matrix permutation in their two mode data set, Elaine Forsyth and Leo Katz [34] explicitly proposed permuting matrices as a way to uncover and display social groups in a one mode data set. They illustrated using data from one of Moreno's [10] sociometric studies. The young women in a residence hall had each been asked to name others in their hall for whom they had positive feelings and those for whom their feelings were negative. Positive choices were recorded using plus signs and negative choices were recorded as minus signs.

Forsyth and Katz adopted a brute-force procedure that involved rearranging rows and columns and redrawing the image again and again until as many of the plus signs fell as close to the principle diagonal as possible. At that point, cohesive groups become visible as clusters of plus signs around the diagonal. Their result is shown in Fig. 24.

Obviously, the Forsyth and Katz procedure was extremely cumbersome. But Beum and Brundage [35] soon came up with a systematic iterative procedure for finding groups by rearranging the rows and columns of a one mode matrix. And, by the late 1950s, when computers emerged on the scene, Coleman and MacRae [36] developed a series of Univac programs at the Operations Analysis Laboratory at the University of Chicago that were designed to uncover the groups in large networks.

An entirely different kind matrix permutation procedure was proposed by Harrison White and his students. They introduced the idea of blockmodeling [37]. In so doing, they provided a theoretical basis for reordering network data matrices, and they developed a number of algorithms for doing so.

The aim of this new thrust was to reorder the matrix in such a way that it could be partitioned to reveal two or more collections of social actors who were *not* linked by some social relation of interest. So, instead of arraying actors along the diagonal of a matrix, White et al. sought permutations that would define zero blocks – sets of actors between which there were no social links. They used their approach to examine a great many network data sets. One example is shown in Fig. 25.

The data in Fig. 25 were collected by Sampson [38] in his study of a monastery. Sampson asked each of a collection of 18 novices to report their relationships with each of the others. Figure 25 shows an 18 by 18 matrix of their responses to a question asking the novices about which oth-

	1	2	3	4	5	6	7	8	9	1	11	12	13	14	15	16	17	18
1										2						1		3
2				3									2					
3		1		3		2												
4	2	3												1				
5			1														3	2
6							3										2	1
7			2	2		3		2									1	
8												1	2			3		
9												1					3	2
1																		
11							3										1	2
12																		
13		3				2						1						
14				2									3				1	1
15			1	3									2					
16			1	3							2							
17				2		3		1										
18				3		1					2							

Social Network Visualization, Methods of, Figure 25
Sampson's data on who was a negative influence on whom

	10	5	9	6	4	11	8	12	1	2	14	15	7	16	13	3	17	18
10																		
5																1	3	2
9								1									3	2
6													3				2	1
4									2	3	1							
11													3				1	2
8								1						3	2			
12																		
1	2													1				3
2					3										2			
14					2										3		1	1
15					3										2	1		
7				3	2		2									2	1	
16					3	2										1		
13				2				1		3								
3				2	3					1								
17				3	2		1											
18				1	3	2												

Social Network Visualization, Methods of, Figure 26
White, Boorman and Breiger's partitioning of the negative influence data matrix from Sampson

$$\begin{matrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{matrix}$$

Social Network Visualization, Methods of, Figure 27
The CONCOR reduction of the Sampson matrix

ers had negative influences on them. A response of 3 indicated a first choice. A 2 was a second choice and a 1 was a third choice. White et al. reasoned that only first and second choices represented strong responses, so they ignored the third choices and treated the entries of 1 as if they did not exist.

One of the several procedures procedure White et al. introduced was called CONCOR. CONCOR is a recursive procedure that begins by calculating correlations between the rows (or columns) of a network data matrix. Then correlations are calculated between the rows of the resulting correlation matrix. That procedure continues until it produces a matrix of correlations that uniformly displays values of +1 and −1. Those positive and negative values are used to partition the individuals into two subsets. The CONCOR procedure can be repeated using the data contained within each of the partitions. Thus, the original matrix can thus be refined to any desired degree.

White and his students used CONCOR on the data shown in Fig. 25 in an attempt to uncover blocks that contained only 0s. They could then use these zero blocks to reduce the complexity of the data matrix. That matrix produced three zero blocks. They are shown in Fig. 26.

In the reduced model in Fig. 27 each cell represents one of the blocks in Fig. 26. Thus, the 18 by 18 matrix is reduced to a 3 by 3 array. The reduction is consistent with Sampson's original ethnographic description of subgroups among the novices. Moreover, its pattern of zero blocks in the principle diagonal indicates that no block member saw any fellow block member as having a negative influence. But the members of each block saw at least some of the members of both of the other blocks as negative influences. This makes sense in the light of the ongoing conflict that Sampson described in his report.

Since that time, displays based on matrix permutations have grown in size, complexity and sophistication. One particularly striking example was produced by Richards and Seary [39]. Their data were drawn from a study of participants in a needle exchange program in Baltimore, Maryland [40]. Richards and Seary examined data on 4259 individuals who picked up and returned needles at each of four exchange sites over a 30 month period. Each cell in the matrix is a record of the number of needles picked up by the individual in that row and returned by the individual in that column. About a third of all needles fall in the principle diagonal of the matrix.

Richards and Seary used the data from the largest weak component in the data set. That component involved 100 000 needle exchanges among 36 000 individuals. Richards and Seary used their program MultiNet [41] scaled the data using a form of singular value decomposition called *correspondence analysis* [18]. They used the coordinates provided by the first Eigenvector to reorder the rows and columns of the matrix. They then colored the entries in terms of frequencies. The color scale is logarithmic: gray is 1 needle, blue 2–3, green 4–7, red 8–15, ma-

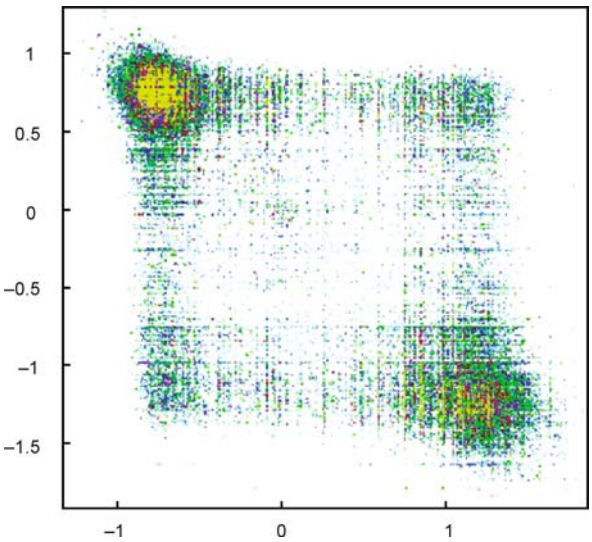

Social Network Visualization, Methods of, Figure 28
The largest component in the Baltimore needle exchange data

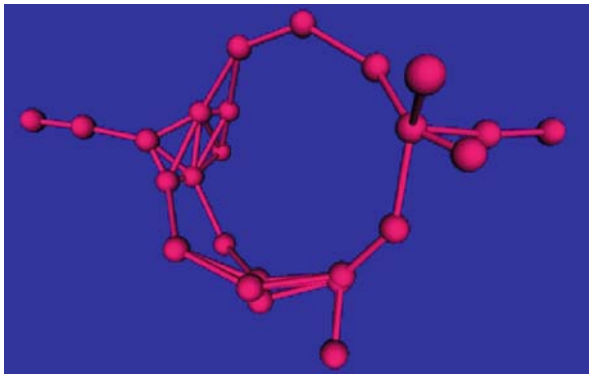

Social Network Visualization, Methods of, Figure 29
VRML image of friendship among teens in a Dublin suburb

genta 16–31, yellow 32 and above. Their image is shown in Fig. 28.

Figure 28 dramatically illustrates the utility of images based on matrix permutation. It shows that there was not a single community of needle users in Baltimore. Instead, there were two distinct communities of individuals who regularly obtain, return and exchange needles with one another. These two relatively large communities were centered around two of the four needle exchange sites.

Future Directions

Overall, the long range the trend in visualizing social networks has been to rely on computers to do more and more of the job. First, computers used a version of singular value decomposition to locate nodes in two dimensional images [42]. Then, soon thereafter, Coleman and MacRae [35] programmed a computer both to permute rows and columns of a matrix and to print out an image of the result. And, in the early 1970s, Alba [43] wrote a program that performed calculations to place nodes and then went on to draw node and edge images of the results [44].

Since the 1970s, then, network analysts have increasingly used computers both for calculations and to draw images. And increasingly, multidimensional scaling and singular value decomposition have been used to determine locations for nodes. Moreover, when two dimensions are not enough to display network structure, three dimensional images are being produced.

When microcomputers became available it quickly became possible to produce images that gave the appearance of being three dimensional. Moreover, with the advent of color screens, color images began to be produced. Figure 29 represents data collected by Kirke [45] on social links among teenagers in a suburb of Dublin. Nodes were first located in three dimensions using multidimensional scaling. And then the virtual reality modeling language (VRML) was used to produce the appearance of three dimensions. Figure 27 was produced as a cover design for a book [46] and the colors were used simply make the image more attractive.

Colors can, however, be used to enhance the ability of an image to communicate important information. In Fig. 30 Höpner and Krempel [47] used a spring embedder and Krempel's own programs to arrange the nodes in two dimensions. The nodes represent the 100 largest German corporations in the year 2000. They used color to label both nodes and directed lines. In their image each company is represented as a node and an arrow pointing from one node to another means that the first node holds shares in the second. The size of a node indicates the number of connections to other nodes it has. Financial companies are shown as yellow nodes and industrial companies are red. Links between financial companies are yellow, those between industrial companies are red and links between financial and industrial companies are orange. Bu using color, then, his directed graph reveals a great deal of information about the organization of German industry and finance.

The image in Fig. 31 was made using a program called MAGE [48]. MAGE was written by Richardson and Richardson [49]. It is designed as a display on computer screens, and it allows the viewer to move into the picture as well as to spin and rotate it. It is useful, then, for exploring the patterning of structural data in three apparent dimensions.

Social Network Visualization, Methods of, Figure 30
Links among German corporations in 2000

In Fig. 31 students in a university professional school program reported who their friends in their class were. Nodes were placed using a multidimensional scaling program and then they were colored according to their program in the school. It turned out that most of their friendship choices linked those that shared a program.

Richards and Seary's [50] program, Multinet, produces a wide range of graphic images. Included are images that actually can be viewed in three dimensions using *anaglyphic* glasses in which one lens is red and one is blue. Obviously, I cannot illustrate their program here, but anyone who wants to see real 3D images should explore Multinet.

The most recent development in visualizing social networks involves the production of animated graphics. As more and more process data are collected and as more process models are constructed, animated images are a natural development. A group at Stanford University has written a Java program, SoNIA [51], that makes it quite simple to produce animated node and edge and node and directed line images [52,53]. These images allow users to explore the changing structural forms generated by process data.

Overall then, in the period between Moreno's hand drawn ad hoc images and the latest animations of dynamic network processes, there has been a dramatic growth in our ability to visualize social network structure. The major contribution has come from computers. Today we can use a wide variety of readily available computer programs to both design images and to produce screen images and/or printed output.

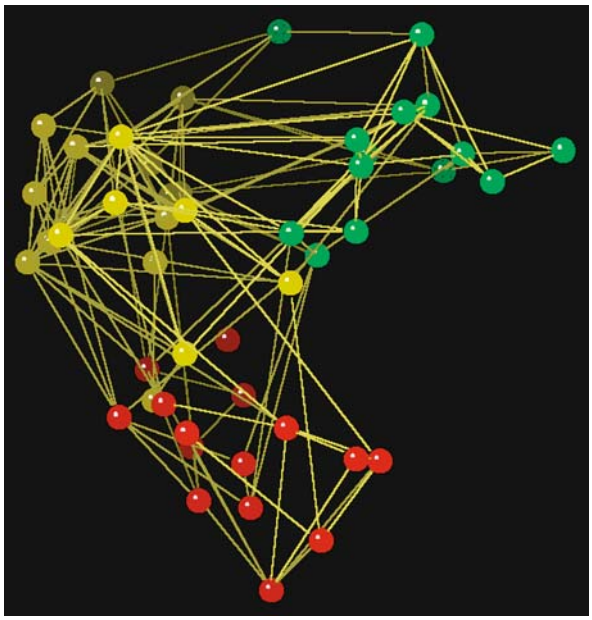

Social Network Visualization, Methods of, Figure 31
MAGE image of friendship among classmates

But, as the job of producing images becomes easier, we must be careful not to lose our sense of why we are pro-

ducing then in the first place. From the very beginning, the important point has always been that the visual images of social networks are not produced simply to be decorative. In every case, the early images were drawn in order to dramatize some feature of social structure. Moreno produced Fig. 5 to illustrate the importance of considering the number of connections in evaluating the structural position of an individual. In Fig. 6 the number of negative ties received by one of the running backs showed, as Moreno (p. 213 in [10]) put it, "It is easy to see that when 5/RB is running with the ball he is not apt to get the maximum of cooperation in interference and blocking."

Figure 7 was a pictorial statement by Warner and Lunt that when cohesive subgroups overlap, they should not be expected to bridge wide differences in social class. Figure 8, from Davis, Gardner and Gardner, demonstrated that the Warner–Lunt hypothesis was supported by data with respect to both social class and age. And, finally, Fig. 9 illustrated the presence of cohesive groups and of the variation of different individuals in their involvement in those groups. In every case each of these early authors had a point to make, and in every case the image helped to make that point. That is the key to the effective use of visual materials in social network analysis.

In future we can expect to see continued development of computer programs designed to aid in visualizing social networks. We can look forward to continued refinement of algorithms for displaying group structure that are based on multidimensional scaling, particularly spring embedding. We can anticipate better algorithms for displaying hierarchies and approximate hierarchies. We can expect to have more powerful programs for animation. And, at the same time, we can expect to be able to produce higher quality and more refined visual displays of all sorts.

Bibliography

Primary Literature

1. Freeman LC (2004) The development of social network analysis: A study in the sociology of science. Empirical Press, Vancouver, BC
2. Schweitzer VS (1998) Words of power. Coffee Times, Fall
3. Cayley A (1857) On the theory of the analytical forms called trees. Philos Mag 13:19–30
4. Biggs NL, Lloyd EK, Wilson RJ (1977) Graph theory 1736–1936. Oxford University Press, Oxford
5. Morgan LH (1871/1997) Systems of Consanguinity and Affinity in the Human Family. University of Nebraska Press, Lincoln
6. Macfarlane A (1883) Analysis of relationships of consanguinity and affinity. J R Anthropol Inst Great Britain Ireland 12:46–63
7. Appendix to Macfarlane A (1883) Analysis of relationships of consanguinity and affinity. J R Anthropol Inst Great Britain Ireland 12:46–63
8. Hobson JA (1884) The evolution of modern capitalism; A study of machine production. Allen & Unwin, Macmillan, London, New York
9. Moreno JL (1932) Application of the group method to classification. National Committee on Prisons and Prison Labor, New York
10. Moreno JL (1934) Who shall survive? Nervous and Mental Disease Publishing Company, Washington
11. Roethlisberger FJ, Dickson WJ (1939) Management and the worker. Harvard University Press, Cambridge
12. Warner WL, Lunt PS (1941) The social life of a modern community. Yale University Press, New Haven
13. Davis A, Gardner B, Gardner MR (1941) Deep south. University of Chicago Press, Chicago
14. Mitchell JC (1994) Situational analysis and network analysis. Connections 17:16–22
15. Available at http://www.analytictech.com/downloadnd.htm
16. Eades P (1984) A heuristic for graph drawing. Congressus Nutnerantiunt 42:149–160
17. Krempel L (1999) Visualizing networks with spring embedder: Two-mode and valued data. In: Proceedings of the section of statistical graphics. American Statistical Association, Alexandria, pp 36–45
18. Weller SC, Romney AK (1990) Metric scaling: Correspondence analysis. Sage Publications, Newbury Park
19. Freeman LC (2005) Graphical techniques for exploring social network data. In: Carrington PJ, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, Cambridge, pp 248–269
20. Bott E (1957) Family and social network. Tavistock, London
21. Available at http://vlado.fmf.uni-lj.si/pub/networks/pajek/
22. Kamada T, Kawai S (1989) A general framework for visualizing abstract objects and relations. ACM Trans Graph 10:1–39
23. Forkman B, Haskell MJ (2004) The maintenance of stable dominance hierarchies and the pattern of aggression: Support for the suppression hypothesis. Ethology 110:737–744
24. Available at http://visone.info/download/
25. Krackhardt D (1996) Social networks and the liability of newness for managers. In: Cooper CL, Rousseau DM (eds) Trends in organizational behavior, vol 3. Wiley, New York, pp 159–173
26. Brandes U, Raab J, Wagner D (2001) Exploratory network visualisation: Simultaneous display of actor status and connections. J Soc Struct 2:4
27. Estrada E, JA Rodríguez-Velázquez. (2005) Complex networks as hypergraphs. Physica A 364:581–594
28. Available on request from its author, Vincent Duquenne (v.duquenne@wanadoo.fr)
29. Levine JH (1979) Joint-space analysis of 'pick-any' data: Analysis of choices from an unconstrained set of alternatives. Psychometrika 44:85–92
30. Wille R (1982) Restructuring lattice theory: An approach based on hierarchies of concepts. In: Rival I (ed) Ordered Sets. Reidel, Dordrecht–Boston, pp 445–470
31. Wille R (1984) Line diagrams of hierarchical concept systems. Int Classif 11:77–86
32. Duquenne V (1987) Contextual implications between attributes and some representation properties for finite lattices. In: Ganter B, Willie R, Wolff K (eds) Beiträge zur Begriffsanalyse. B. I. Wissenschaftsverlag, Berlin, pp 213–240

33. Freeman LC, White DR (1993) Using Galois lattices to represent network data. In: Marsden PV (ed) Sociological methodology. Blackwell, Oxford, pp 127–146
34. Forsyth E, Katz L (1946) A matrix approach to the analysis of sociometric data: Preliminary report. Sociometry 9:340–347
35. Beum CO, Brundage EG (1950) A method for analyzing the sociomatrix. Sociometry 13:141–145
36. Coleman JS, MacRae D (1960) Electronic processing of sociometric data for groups up to 1000 in size. Am Sociol Rev 25:722–727
37. White HC, Boorman SA, Breiger RL (1976) Social structure from multiple networks: I. Blockmodels of roles and positions. Am J Sociol 81:730–779
38. Sampson SF (1969) A noviate in a period of change: An experimental and case study of relationships. Unpublished Ph D dissertation. Department of Sociology, Cornell University
39. Richards WD, Seary AJ (2000) Cover illustration. Connections 23:1
40. Valente TW, Foreman RK, Junge B, Vlahov D (1998) Satellite exchange in the Baltimore needle exchange program. Public Health Reports 113:91–96
41. Availble by contacting Bill Richards (richards@sfu.ca)
42. Bock RD, Husain SZ (1952) Factors of the tele: A preliminary report. Sociometry 15:206–219
43. Alba R (1972) SOCK. Behav Sci 17:326–327
44. Kadushin C (1974) The american intellectual elite. Little, Brown, Boston
45. Kirke DM (1996) Collecting peer data and delineating peer networks in a complete network. Social Netw 18:333–346
46. Kirke DM (2006) Teenagers and substance use: Social networks and peer influence. Palgrave Macmillan, Basingstoke, Hampshire and New York
47. Höpner M, Krempel L (2003) The politics of the German company network. Max Planck Institute for the Study of Societies, Working Paper 03/9
48. MAGE is available at http://kinemage.biochem.duke.edu/software/mage.php
49. Richardson DC, Richardson JS (1992) The kinemage – a tool for scientific communication. Protein Sci 1:3–9
50. Richards WD, Seary AJ (2006) Multinet. http://www.sfu.ca/~richards/Multinet/Pages/multinet.htm
51. SoNIA is available at http://www.stanford.edu/group/sonia/
52. Moody J, McFarland DA, Bender-deMoll S (2005) Visualizing network dynamics. Am J Sociol 110:1206–1241
53. Bender-deMoll S, McFarland DA (2006) The art and science of dynamic network visualization. J Social Struct 7:2

Books and Reviews

Brandes U, Kenis P, Raab J, Schneider V, Wagner D (1999) Explorations into the visualization of policy networks. J Theor Polit 11:75–106

Brandes U, Corman SR (2003) Visual unrolling of network evolution and the analysis of dynamic discourse. Information Visualization 2:40–50,

Brandes U, Kenis P, Wagner D (2003) Communicating centrality in policy network drawings. IEEE Transactions on Visualization and Computer Graphics 9:241–253,

Brandes U, Wagner D (2004) Netzwerkvisualisierung. Inform Technol 46:129–134,

Brandes U, Wagner D (2004) Visone - Analysis and visualization of social networks. In: Jünger M, Mutzel P (eds) Graph drawing software. Springer-Verlag, pp 321–340

Brandes U, Kenis P, Raab J (2006) Explanation through network visualization. Methodology 2:16–23

Frank O (2000) Structural plots of multivariate binary data. J Social Struct 1:4

Freeman LC (2000) Visualizing social groups. In: Proceedings of the section on statistical graphics (1999). American Statistical Association, pp 47–54

Freeman LC (2000) Visualizing social networks. J Social Struct 1:1

Freeman LC, Webster CM, Kirke DM (1998) Exploring social structure using dynamic three-dimensional color images. Social Netw 20:109–118

Johnson JC, Krempel L (2004) Network visualization: The 'Bush Team' in Reuters News Ticker 9/11–11/15/01. J Social Struct 5:1

Klovdahl AS (1981) A note on images of networks. Social Netw 3:197–214

Klovdahl AS (1998) A picture is worth…: Interacting visually with complex network data. In: Liebrand WBG (ed) Computer modeling of dynamic social processes. Sage, London

Krempel L (2005) Visualisierung komplexer Strukturen. Grundlagen der Darstellung mehrdimensionaler Netzwerke. Campus, Frankfurt aM

Krempel L, Schnegg M (2005) About the image: Diffusion dynamics in an historical network. Structure and dynamics: eJ Anthropol Rel Sci 1:1, 10

McGrath C, Blythe J, Krackhardt D (1996) Seeing groups in graph layouts. Connections 19:22–29

McGrath C, Blythe J, Krackhardt D (1997) The effects of spatial arrangements on perceptions of graphs. Social Netw 19:223–242

McGrath C, Krackhardt D, Blythe J (2003) Visualizing complexity in networks: Seeing both the forest and the trees. Connections 25:37–47

McGrath C, Blythe J (2004) Do you see what i want you to see? The effects of motion and spatial layout on viewers' perceptions of graph structure. J Social Struct 5:2

Social Organizations with Complexity Theory: A Dramatically Different Lens for the Knowledge Economy

Russ Marion
Clemson University, Clemson, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Basic Premises
Understanding Organization with Complexity
Complexity Outcomes: Creativity, Adaptability, Learning
Methodology for Studying Complex Organizations

Conclusions
Future Directions
Bibliography

Glossary

Many of these terms are commonly understood by complexity theorists; these definitions, however, describe their meaning within the context of social organizations.

Adaptive behaviors Perturbations (creative ideas, pressures, etc.) in a system that foster some observable level of phase transition. Change leading to systemic reaction.

Adaptive tension Pressures external or internal to the organization that perturb an organization thus pressuring it toward phase transitions and structural or ideational elaboration.

Administrative leadership behaviors related to such things as strategic planning, policy making, and resource acquisition and distribution.

Agent An entity that is the smallest unit of interest in a complex dynamic. An agent could be a person, idea, task, knowledge, etc.

Aggregation Coevolutionary emergence of diverse ideas, agents, etc.

Aggregation mechanisms Macro-level complexity mechanisms. Refers to the dynamic interactions of perturbations (see initiating mechanisms), amplifications (expansion and elaboration of a perturbed state), and phase transition (nonlinear, form-shift in some part of an organization).

Coevolution A process in which "reciprocal selective pressures operate to make the evolution of … (one agent in an interactive process) partially dependent on the evolution of (other agents)" [67]. Coevolution occurs when heterogeneous agents in a niche are interdependent such that they dynamically adjust to each others changes.

Commodity-based economy Refers to a manufacturing economy in which the major assets are physical commodities.

Complexity organizational knowledge Systemic knowledge that imbues organization with capacity for adaptive and creative responses; attributable to the strength and viability of the system's complex structures and to the viability of the relationship between the organization's complex structures and its bureaucratic structures.

Complexity mechanism Social mechanisms that underlie complexity dynamics.

Complexity theory In organizational sciences, the study of emergent dynamics in neural-like networks of adaptive, vision-oriented agents.

Conflicting constraints Conflict that emerges when the preferences of one agent challenges the preferences of another. In complex networks, conflicting constraints generate pressure to elaborate.

Dissipative structures Structures that emerge when far from equilibrium systems release excess energy in a phase transition.

Emergence The appearance of new structures or ideas from the actions of complex interactions.

Enabling behaviors Activities that foster conditions (e. g., interdependency, enabling rules, adaptive tension) in which complex dynamics can emerge.

Enabling rules Rules that govern interactions among adaptive agents in complex systems. Contrast with bureaucratic rules, which delimit the responsibilities of agents in a closed system bureaucracy.

Entanglement A dynamic relationship between the formal administrative forces and informal complexly adaptive, emergent forces of an organizational systems.

Equilibrium A stable, predictable relationship among agents and structures of a relationship. Related to thermodynamic concept of a low energy state.

Extreme event Instances of dramatic change that occur infrequently in social organizations; the US government's experience of the Katrina hurricane in New Orleans illustrates. Extreme events typically require rapid action and can be minimally responsive to top-down control.

Far from equilibrium Typically defined as a high energy state; defined here for organizational studies as an intensely complex, dynamic state driven by excess levels of pressure and perturbations.

Heterogeneity A diversity of skills, worldviews, preferences, beliefs, goals, (etc.) among interactive agents in a complex system.

Initiating mechanisms Micro-level complexity mechanisms. They include coevolutionary interaction and perturbations (an unexpected change in an interaction relationship; attributable to complex, interactive dynamic).

Interdependency Network conditions in which the actions of one agent are influenced by the actions of another.

Knowledge economy Refers to a production economy in which the major assets are the knowledge possessed by individuals and networks of individuals (organizational knowledge).

Meso theory Variously defined as theory that bridges macro and micro level theory; different levels of hierarchy; or different levels of analysis (individual, dyadic, group).

Multi-agent based modeling Agent based modeling procedure that analyzes different types of networks simultaneously (e. g., agent networks, task networks, etc.)

Organizational level From Jaques [40], the level of bureaucracy that lies between the upper echelon levels and the work production level; includes middle management.

Perturbations Events that disturb normal organizational interactions and generate pressures that can lead to phase transitions.

Phase transition Sudden, nonlinear form shifts attributable to the dissipation of accumulated pressure in complex systems. Phase transitions can occur at multiple levels of intensity.

Postmodernism In general, a rejection of scientific modernism. For organizational science, it represents a realization that organizational behaviors cannot be adequately expressed as mathematical relationships among variables. Complexity theory adds that organization is an ultimately unpredictable dynamic whose causal structure is based on interactions among complexity mechanisms.

Production level From Jaques [40], the level of bureaucracy responsible for line production.

Requisite complexity McKelvey and Boisot's [65] modification of Ashby's requisite variety; they maintain that viable organizations are at least as complex as their competition.

Requisite variety Ashby's [5] dictum that viable organizations have at least the same degree of flexibility as their competition.

Spaces between A proposal that creative ideas emerge from interactive dynamics. Creative emergence is generated when agents work to resolve tension; such tension is product of conflicting constraints, heterogeneous preferences, ideas, and knowledge, etc.

Social mechanism A process attributable to dynamic interactions among multiple people, variables, ideas, physical limitations, etc. Mechanisms describe "a set of interacting parts – an assembly of elements producing an effect not inherent in any one of them" p. 336 in [27], p. 74 in [37].

Strategic level From Jaques [40], the upper echelon level of bureaucracy; responsible for organizational strategy, policy making, resource acquisition and allocation, etc.

Swarm behavior Study of the dynamics of swarms in biology and in organizations. Scientists are finding that amazingly complex, "intelligent" behavior can emerge from complex behaviors that are structured by a few, simple enabling rules. Useful in organizations for improving such things as distribution efficiency or flexible response to environmental conditions.

System dynamics Network simulations in which relationships among agents are defined mathematically.

Top-down administration Administrative practices in which decisions are made by superiors to be carried out by subordinates.

Vision A teleological view of the future. Vision can be highly specific, or determinate, as when an organization projects future markets (such visions are more properly, goals; see Sharp Corp.'s strategy statement in this paper for an example). Vision statements appropriate for complex systems are indeterminate in that they do not preclude the future; for example, "This company will strive to enhance its competitive advantage by optimizing its flexibility."

Definition of the Subject

Complexity theory for organizations examines the influence of complex dynamics on (among many things) organizational structure, leadership, power and control, influence, and strategy. It is applicable not only to understanding but of practice in any organizational type whose primary commodity is knowledge and application of knowledge. The core outcomes of complex dynamics are creativity, adaptability, and learning. Complexity theory is particularly germane in what has come to be called, the knowledge economy.

This paper defines basic premises underlying complex dynamics in organizations, and argues in particular that organizational complexity is best understood as the interactions among complexity mechanisms (causal process attributable to dynamic interactions among multiple people, variables, ideas, etc.) rather than as the outcome of defined variables. We explore the relationship between bureaucracy and complexity, and define three levels of behavior in complex organizations: administrative, enabling, and adaptive. We apply these ideas along with complex premises underlying complexity to strategic leadership. The paper describes how complexity produces its core outcomes – creativity, adaptability, and learning – and argues that these outcomes are particularly pertinent to today's economy. It concludes with a discussion of unique research methodologies that can be used to study a process based on mechanisms rather than on variables.

Introduction

The science of complexity offers interesting solutions to some difficult challenges in organizations. These challenges affect nearly all economic sectors of society, from electronics to education, computer programming to automotives, banking, government, and the armed forces. They include hyper-competitive environments [38], globalization, unequal global distribution of wealth and production costs, rapid technological and social change, highly complex operations (e. g., the complexity of sophisticated software programming), and the changing nature of the workforce (less loyal, comfortable with electronic communication, demanding of flexible work arrangements). To address these challenges, practitioners and scholars recognize the need for radically new leadership paradigms [81], revised perspectives of organizational strategy [81], flexible structures [83], rapid adaptability [16], and decentralized problem-solving structures [25]. Complexity science is applicable to these challenges because, among other things, it focuses on flexible, interactive dynamics rather than top-down control, it offers unique descriptions of change in dynamic system, it reveals important clues about creativity, adaptability, and learning in organizations, and it shifts attention from central tendency variables to causality based on dynamic mechanisms [50,81].

Traditional thought in organization science was oriented toward efficient production and top-down coordination. It assumed that organizational behavior was intimately dependent upon structure provided by positional leaders, rational plans and goals, centralized coordination, organizational evolution, and vision [35]. It has explored the nature and impact of power [41], institutional forces [28], and charisma [23]. Traditional thought has, like complexity theory, described change but has assumed that organizational change and innovation must be planned, that effective action must be coordinated, and that organization must be structured in the board room.

By the 1990s, this "top-down" notion began losing credibility among organizational scholars and practitioners (e. g., [35]). The major culprit in this demise was the knowledge economy – the globalization, rapid change, and hyper-competitive environments described above. Top-down perspectives of organization had been oriented toward more stable, commodity based environments [9], and the emerging knowledge economy is oriented toward creativity, adaptability, and learning.

Complexity theorists in the organizational sciences seek relevant strategies for dealing with the knowledge economy. Complexity theory is useful for this because it does not merely improve traditional perspectives of doing organization but rather addresses different sets of behaviors than are typically addressed by organizational theorists. In particular, it focuses on informal interactive processes within an organization, on processes that emerge from informally interacting people and groups, and on productive relationships between informal and formal activities within an organization. It asks such questions as, "How can organizational decision making and interactive patterns be restructured and reorganized to increase flexibility" [49]? "How can leadership be re-conceptualized in a way that is effective for complex organizations" [70]? Or, "How do we re-focus organizational strategy for a knowledge economy" [58]?

In this paper, we look first at basic premises underlying the application of complexity theory to organizations, then look at how complexity can inform understanding of bureaucracy and strategic leadership. We examine the organizational outcomes of complex dynamics (hence the rational for applying this science), and conclude with a discussion of research methodologies that are useful for studying complex organizational behaviors.

Basic Premises

Studies have found that, like biological [30], economic [4], and physical [45] systems, emergent dynamics are readily observed among interacting people, or agents [20,68,69]. Interactive dynamics and emergence are common denominators in complexity studies, and while there are discipline-specific variations (organizational agents, for example, make learned, adaptive choices while weather systems react primarily to the interactions of physical properties), all are shaped by these defining dynamics. Agents interact and the interaction changes them (emergence).

Complexity organization researchers are interested in understanding how these dynamics operate in human systems. Lichtenstein and Plowman [51] have found, for example, that certain interactions among individuals and groups lead to perturbations, or pressure to elaborate. Consistent with Kauffman [43], these tensions pressure the system to adapt and to change. Perturbed, interactive systems also tend toward self-organization [57,62]. This phenomenon was first systematically observed in the 1940s and 1950s as informal groups [39] and was expanded by complexity theorists to include any situation in which ideas or people spontaneously, and without external motivation, act in some degree of synchrony with one another. Chiles et al. [20] identified dissipative structures – fluctuation, positive feedback, stabilization and re-

combination – in the emergence of the music industry at Branson, Missouri.

Researchers are also interested in understanding the conditions that enable emergent, self-organizing behavior in social systems [82]. *Interaction*, of course, is *sine qua non*. Schreiber, Marion, Uhl-Bien, and Carley [74], drawing from Kauffman [43], found that interactive agents must possess a requisite level of *interdependency* to enable learning outcomes (too much or too little will suppress complex dynamics; [43]). This interdependency generates *conflicting constraints*, which are important sources of perturbations in a complex system. McKelvey [64] adds that such pressure is also generated by external sources such as leadership behavior or economic pressures; he refers to this as *adaptive tension*. Uhl-Bien et al. [82] adopted this term to refer to both internal and external pressures.

Bonabeau and Meyer [10] have developed fascinating insights into the nature of swarm behavior in human organizations, and have used it to solve efficiency problems involving such things as package delivery and local surges in phone traffic [10]. Swarms are a function of a set of simple, *enabling rules*, such as "fill orders [in a distribution warehouse] until the next person in line takes over", or "drop 'digital' pheromone trails for subsequent messages to track" (in computerized instructions for phone systems).

Complex behaviors in social systems are enabled by *heterogeneity* of skills, preferences, outlooks, etc. Traditionally, organizational theorists, particularly those who study leadership, have sought to align employees to centralized goals [56]; complexity theory suggests that organizations whose employees pursue a diversity of goals (within the context of an interactive, interdependent network) will have a diversity of ideas to draw from as they seek creativity and adaptable changes.

Finally, Schreiber, Marion, Uhl-Bien, and Carley [74] found in a multi-agent based simulation that moderate levels of *vision* (or general organizational focus) are conducive to complex behaviors. High centralization of vision (in combination with high interdependency among agents) was counter-productive for complexity outcomes in their simulation, as was low levels of vision (low vision was more problematic to outcomes than was high vision).

Emergence and Phase Transitions

Emergence refers to the generation of new structures or ideas due to the actions of complex interactions. This phenomenon is important because it allows the system to rapidly produce adaptive, creative responses to environmental conditions. It occurs in social structures when heterogeneous agents or ideas act in relative synchrony with one another. The emergence of riot conditions out of the behaviors of just a few people illustrates (the emergence of a riot also illustrates that emergence is not necessarily a functional response to environmental contingencies; the notion of counter-productive emergence has been discussed by Uhl-Bien, Marion and McKelvey [82]).

One form of emergence involves the *accumulation* of different ideas or agents. Anderson [1] and Marion [55], for example, described the emergence of technology that led to the invention, in 1975, of the microcomputer. It began with disparate advances – transistors, printed boards, LEDs, etc. Engineers began using these isolated components together, as they did in the transistor radio popular in the 1960s. Accumulation took a big leap with the introduction of the handheld calculator in the late 1960s, which added such things as processors and information storage capacity. The microcomputer was the ultimate "accumulation" or emergence event.

Importantly, the components of such emergence processes accumulate because they "coevolve." Coevolution is defined by Pianka, p. 329 in [67], as a process in which "reciprocal selective pressures operate to make the evolution of either … [agent in the interactive process] partially dependent on the evolution of the other" see also McKelvey [64]. Coevolution occurs when heterogeneous agents in a niche are interdependent such that they dynamically adjust to each other's changes. It raises the important observation that an organizational species' fitness – the capacity to survive and thrive – is influenced by supporting roles from other organizational species [43].

A second form of emergence is called *phase transition*. Phase transitions are sudden, nonlinear form-shifts (changes in structure, knowledge, worldview, etc.). The sudden demise, in 1989, of the USSR illustrates this phenomenon. Phase transitions like that experienced by the USSR result from a build up of tension or pressure [32,64,71]. Haken [32], for example, illustrated phase transition in terms of slowly heated oil; eventually the heat (tension) reaches a point at which the oil precipitously transitions to a new state, observed as a gentle boil.

Phase transitions, occur at multiple levels of intensity, and a number of researchers have observed that, in complex systems, intensity of transitions plotted against their frequency describes a power law relationship ($1/f^x$); [2,7,75]. Power law curves are steeply descending slopes that quickly bottom out into a more gently descending slope; when intensity and frequency are converted to log equivalents, the slope becomes a straight line descending left to right like a negative regression line in statistics. Power law curves describe situations in which high intensity events

occur infrequently while low intensity events occur often.

This has opened a line of exploration into what are called, extreme events, or analyzes of phenomena that are highly intense but of low frequency (the 2008 Organization Science Winter Conference was devoted to this topic). Stephenson and Bonabeau [76] concluded, for example, that bottom-up, or adaptive, leadership strategies may be more effective in extreme events such as Katrina or the Twin Tower attacks of 9/11 than are top-down leadership strategies.

Far from Equilibrium

The accrued tension described by Haken generates what Prigigine [71] calls, a far from equilibrium (FFE) state. At FFE; the system will often act spontaneously to release (or dissipate) built up tension by moving to a new structural state [64].

In organizational studies, the far from equilibrium notion is unique, for organizations have traditionally been described in terms of equilibrium structures ([12]; see Maruyama [60], for a notable exception). The idea that far from equilibrium states exist suggests an equilibrium continuum for organizations that stretches from an equilibrium state to far from equilibrium conditions. Hazy and Marion [34] argue, however, that while this may be true for some physical systems, it is not true of human systems. Human systems always function in a state of change, albeit with varying degrees of complexity, and it makes no sense to juxtapose such systems against an equilibrium end point. They conclude, then, that far from equilibrium in a social system occurs when the dynamical conditions and resulting complexity of a social system are increased beyond a certain point, not when it moves away from equilibrium.

Even so, the assumption that organization exists in a state of equilibrium is common in the traditional literature on organizations. Although it has been clarified by earlier organizational theorists to be a moving equilibrium (called homeostasis) that adjusts to environmental conditions (the influential classics of this argument are [12]; and [84]), the basic premises in organizational studies have implicitly or explicitly equated organization with adapted variations of the equilibrium view from thermodynamics. Complexity theory contradicts this notion, arguing instead that adaptive systems can only survive and thrive in a state variously called far from equilibrium [64], complexity, edge of chaos [42], or Type IV order [47]. If Hazy and Marion [34] are correct, perhaps the better terms for this are those that do not imply an equilibrium state at all.

Complexity Mechanisms

Complexity is unique in organizational studies because it is oriented around social processes rather than around variables. A variable, or measure of central tendency, is static and it reveals little about the evolutionary social processes in which the variable is embedded. Complexity seeks to understand and to describe the behaviors of those social processes, or what Davis and Marquis [27] and Hedström and Swedberg [36] and others have labeled, social mechanisms.

Mechanisms are defined as commonly observed patterns of behavior [29]; they describe "a set of interacting parts – an assembly of elements producing an effect not inherent in any one of them" p. 336 in [27]; see also p. 74 in [37]. Mechanisms are processes, or dynamic interactions among agents, ideas, and structures. Elster's [29] definition proposes that certain mechanisms can be "commonly" observed across multiple phenomena (that is, they are not necessarily unique to given conditions). They illustrate by arguing that the logistic curve (emergence begins slowly, picks up speed, then levels off) is a mechanism and that it can be observed in numerous situations (Mozart's lifetime productivity, emergence of product popularity, etc.)

Uhl-Bien and Marion [81] identify two broad, interdependent categories of universal mechanisms that they call, complexity mechanisms: These are initiating mechanisms, or interactive processes that generate perturbations, or disturbances, in a system, and aggregation mechanisms, which include accumulation mechanisms (accruing of ideas or structures) and phase transitions (described earlier as nonlinear changes that generate new order). Initiating mechanisms are more micro level while aggregation mechanisms are more macro level.

Practitioners who grasp this notion of complexity mechanisms do not perceive organization in terms of average motivation or efficiency, but rather perceive organization as sets of interacting, evolving processes. They do not ask, for example, whether satisfaction will increase if leaders engage in certain specific acts, or whether high leadership scores on a transformational scale will enhance productivity. Rather they want to know how interactive, interdependent mechanisms influence the capacity of social networks within organizations to generate unpredictable but creative outcomes. These contrasting sets of illustrative questions differ in that the former are based on scientific assumptions about linear relationships among central tendencies and the latter are based on postmodern assumptions regarding complex interactions that are indeterminate in the detail but determinate in the general

nature of their outcome. For example, we can predict dynamic creativity capacity but we cannot predict what that creativity will look like.

Understanding Organization with Complexity

We turn now to an examination of some specific applications of complexity to organizational theory and behavior. This discussion is framed around the relationship of complexity to bureaucracy and around strategic leadership.

Bureaucracy and Complexity

The discussion to this point has implicitly portrayed organization as an adaptive dynamic, or as a system driven exclusively by agents acting informally within complex networks to produce emergent outcomes. In reality, organizations are bureaucratic structures, and despite predictions by some of a post-bureaucratic economy [31,35,54], it is unrealistic to ignore this pervasive element of organization (as Weber [85], said, bureaucracy will be with us until the last shovel of coal is dug from the earth).

For this reason, complexity theorists Uhl-Bien et al. [82] have chosen to depict informal, complex dynamics as embedded within each of Jaques' (1989) three levels of bureaucracy: production, organizational, and strategic. They observe that, while bureaucratic functions differ at each level, the nature of complex dynamics differ only in the tasks they address. That is, complex behavior is similar across the system, whether observed in the board room or in the production divisions. Dynamic, complex behaviors are pervasive, and are universally driven by the enabling conditions described in the last section. Specific complex processes are entangled with one another and even across bureaucratic level. Further, the inter-influences across these dynamics is more horizontal than hierarchical: Dynamics on the shop floor – Jaques' [40] production level – can directly and quickly influence dynamics in the executive suite – Jaques' strategic level. Using complexity terminology, complex dynamics in organizations are, therefore, fractal in nature. From an organizational perspective, they represent a meso view of organization; that is, they describe the linkages that unite micro, or individual, level structures with macro, or organizational level, analysis [26]. Uhl-Bien and Marion [81] also argue that they are meso in that they link different hierarchical levels.

From this, Uhl-Bien, Marion, and McKelvey [82] propose three key dynamics within organizations (which they describe relative to leadership): administrative, adaptive, and enabling. *Administrative leadership* is related to traditional perspectives of leaders as those who generate organizational strategy, gather and distribute resources, and create policy. They differ from traditional perspectives in that their actions must serve to nurture the somewhat fragile, bottom-up (or adaptive) complex leadership processes (fragile in that they are somewhat sensitive to political power behaviors).

Adaptive leadership is defined as any interactive behavior that creates a perturbation which leads to a phase transition. The key elements are interaction, perturbation, and phase transition. Uhl-Bien et al. argue that change originates in the "spaces between" interacting agents (see also [52]); that is, it is the product of the push and shove that can occur between or among agents with heterogeneous skills or worldviews. Perturbations are resulting disturbances in a network (change, new ideas or perspectives, inventions, etc.) that reshape the interaction patterns in that network [81]. Phase transitions are nonlinear shifts within a network (these are described more fully below).

Enabling leadership, the third of Uhl-Bien et al. [82] types, refers to individuals who foster the enabling conditions described above (interaction, interdependency, heterogeneity, enabling rules, etc.) and who mediate the relationship between adaptive and administrative functions.

Emergence, Niche, and Strategic Leadership

The earlier discussion of emergence and phase transitions help us re-visualize the role of another strand in organization sciences: strategic leadership. Strategic leadership theorists study upper echelon behaviors of leader/agents responsible for positioning an organization in its environment. Traditional literature has perceived strategic leadership as a top-down, goal oriented process [13] in which the leader manages the organization's environment. For example, Sir Howard Stringer articulated the strategic goal of Sony Corporation at the 2006 International Consumer Electronics Show in Las Vegas, as:

> Sharpening our focus, shattering internal silos, streamlining our product offerings and growing ever more consumer-centric. We have analyzed our product lines and reorganized our corporate structure to become more nimble and better able to deliver champion products and a focused product line-up. [78].

The website for Sharp Corporation articulates both its philosophy and its business strategy in terms of expanding its market (sharp-world.com/corporate/info/philosophy/index.html). Strategic goals from other companies could be cited, but most will, like these, focus on the company's competitive position in its environment.

Given such assumptions, traditional strategic leadership literature has examined such things as the personality and decisions of top echelon leaders, relationships between the CEO and his or her board of directors, transformational leadership, and visionary leadership (see [13], for a summary).

More recent work by strategic leadership scholars is still concerned about position in the environment, but has focused not on upper echelon behavior but on the internal dynamics of organizations and how organizational flexibility contributes to organizational strategy [8]. Interestingly, Sharp has heeded this call: Stringer's articulation of their goal (above) includes a call for a "more nimble" firm. However, a review of the complexity literature has revealed little that has addressed this issue, despite the fact that it seems a natural subject for complexity analysis. One exception is an article by Marion and Uhl-Bien [58], which argues that flexible, complex structuring is key to organizational strategy. They propose that firms work to increase interaction, interdependency, and adaptive tension.

Marion and Uhl-Bien [58] also suggest that that traditional external-oriented strategic leadership has missed an important point. Much of this literature is, at least implicitly, about dominating the environment or beating the competition. Complexity theory argues that different organizations coevolve [43], and in doing so, they grow to depend upon one another. That is, coevolving systems aggregate and form a niche, or interdependent infrastructure upon which each part of the niche depends to some extent. Automobile makers depend upon fuel outlets and automotive repair shops, for instance – and vice versa. Automotive makers even depend on one another, for their competitive relationship generate pressures that foster structural and ideational elaboration and generate a heterogeneous set of creative solutions that foster even greater creativity. External strategy, then, may be about more than competition; it may also be about fostering effective infrastructure: Marion and Uhl-Bien called this, survival of the cooperative.

Complexity Outcomes: Creativity, Adaptability, Learning

From all this, we can generalize the outcomes of complex dynamics for organizations as creativity, adaptability, and learning. *Creativity* is defined as an ideational phase transition in which an outcome reasonably could not have been predicted based on preexisting conditions or paradigmatic assumptions. Described more basely, it is something that "no one saw coming." Since it is a phase transition, it can be a minor bit of creativity or a major insight.

The major creative insights, of course, capture our imagination and dramatically change our way of seeing things (e. g., Einstein's relativity or Schrödinger's cat). However, small creative events are more common and, arguably, just as important. Creative ideas exist within complex networks that enable them to coevolve and aggregate. That is, they become part of the emergence dynamic within an organization. Creative ideas and associated structures change, aggregate with other creative ideas or structures, and generate adaptive pressures that maintain or increase the complexity of a system. Without the pressures that creative ideas generate, an organization would move towards increased isomorphism and decreased complexity – they move to a state that some would mistakenly call, homeostasis.

Adaptability is enhanced when an organization increases its level of complexity. This translates into (among other things) increased heterogeneity of skills and preferences. Traditional organizational theory has assumed that everyone should be on the "same page" (e. g., "Leaders … are responsible for the dissemination of strategic organizational goals, as well as for convincing their constituents to effectively implement those goals" p. 626 in [6]). Complexity theorists argue that, although organizational agent should be united by a vision, dynamic heterogeneity enables organizations and agents to rapidly and effectively deal with environmental exigencies. Ashby [5] refers to this as "requisite variety;" he defines it by stating that it takes variety to defeat variety. McKelvey and Boisot [65] call it "requisite complexity," thus underscoring the dynamic, co-evolving nature of complex systems and their adaptive response capabilities.

Learning is defined for organizations in several ways. Many of us would define it as something that an individual does. Some organizational theorists [3,22,66], including some complexity scholars (e. g., [9]), argue that learning is best understood as an organizational process. The functioning of the brain exemplifies this logic (see [44]). Individual neurons "learn" only small pieces of an environmental event; it is only when those pieces are put together within a network of pieces that we began to make sense of that event. This, crudely, is what is meant by organizational learning. Organizational knowledge can be defined as a network of different but synchronized bits of individual knowledge. Organizational learning is related to the complex nature of those networks – that is, to its dynamic, changing capabilities. If, then, a system increases its complexity, it will increase its capacity to learn.

Still others propose that organizational learning is a function of the capacity of the network to disseminate information [15]. Like the complexity perspective of organizational knowledge in the previous paragraph, this perspective is intimately linked to the degree of complexity in the organizational network.

Methodology for Studying Complex Organizations

Before developing the methodological strategies that are useful for studying complex organizational system, it will be helpful to summarize a few of the assumptions that underlie complexity, all of which have been stated or implied in this manuscript. These assumptions guide our selection of methodological strategies in organizational studies.

1. Complexity science examines the dynamics of interactions among large numbers of adaptive agents [24].
2. Complex systems evolve dynamically and change nonlinearly (unpredictably; [21]).
3. Human systems do not exist in a state of equilibrium.
4. Complex systems are best described in terms of interactions among complexity mechanisms rather than as interaction among variables.
5. Complex dynamics are enabled rather than planned or controlled.

Traditional Methodologies in the Social Science

Perhaps the most common methodology used to study organizations involves field research strategies and regression statistical procedures. Field strategies permit researchers to draw conclusions in non-laboratory conditions. Regression techniques are based on assumptions of regular, predictable relationships among variables; if one finds such a relationship between variable A and variable B, the relationship was true yesterday and it will be true tomorrow [61]. Such relationships are predictable in that they assume that if you increase variable A, then a resulting change in variable B can be calculated.

While statistical methodology is certainly useful, it cannot deal with many of the conditions posed by complexity science. It is not suited, for example, for exploring dynamic aggregation in organizations, nor is it helpful for explaining nonlinear changes that occur at far from equilibrium states. In particular, its usefulness lapses when one seeks to understand interactions among, and outputs of, complexity mechanisms. Complexity mechanisms are a function of dynamic interactions among numerous agents, ideas, and even variables. Traditional statistical methodology is just not suited for this type of analysis. Statistical assumptions are better suited to equilibrium systems in which desired outcomes involve planning and control. Complexity science is not particularly interested in such outcomes.

A growing trend in organizational studies involves the use of qualitative strategies. Qualitative research requires careful observation of a social dynamic followed by thorough analysis of what was observed. Qualitative methodology is useful for observing dynamic activities and social mechanisms, and is rather widely used for that purpose [20,49,68,70]. Its drawbacks are that it is time and resource intensive and it does not allow one to freely experiment with, or freely change, a social mechanisms – one cannot readily play what – if games in qualitative analysis.

Methodologies for Complexity

The most widely used methodologies in complexity study in organizations are agent-based modeling (ABM), systems dynamics, and qualitative research. ABM replicates or mimics interactive dynamics among organizational agents. ABM is a simulation procedure in which individually programmed agents interact in a computer according to certain rules. Commonly used examples are NetLogo by Uri Wilinsky, Dynamic Network Analysis by Kathleen Carley, and Repast Agent Simulation Toolkit by Michael J. North. These simulations allow researchers to simulate interactions under a variety of conditions and to test the effects of various changes in those conditions. ABM has been used to explore such things as the dynamics that underlie personnel turnover [14], organizational design and performance under stress [19], and leadership as enabler of complex organizational functioning [73].

System dynamic analysis is similar to ABM in that it examines evolving relationships among agents. It differs in that those relationships are not governed by social rules as they are with ABM's, rather they're governed by mathematical defined relationships. Like ABM, it permits the researcher to experiment with various conditions. Sterman [77], for example, has used it to examine interactions among individual decisions and the effects of feedback on the decision process.

Both ABM and system dynamic computer programs permit researchers to base their simulations on data collected in real-world situations. Once that data is entered, however, the quality of the subsequent simulation is only as good as the quality of the rules or the equations that govern it. However the intent is not to predict specific outcomes; rather, the intent is to understand how dynamic relationships can affect outcomes. Carley's Dynamic Network Analysis, for example, permits a user to evaluate the effects of an evolving network on the capacity of the or-

ganization to diffuse knowledge. One can play with such things as the degree of interdependency among agents or the degree of centrality around vision (as Schreiber et al. did in their [74] study) to see how they influence the learning dynamic.

Qualitative research performs generally the same things that ABM and system dynamic analysis do, but it allows the researcher to continuously examine real life events rather than merely simulating those dynamics. The benefit is a closer description of reality; the drawbacks are that it is highly resource and time intensive (especially if one wants to obtain results as in-depth as are results available through simulations), and it does not generally permit much control over conditions that might influence outcomes. Interesting qualitive studies of complex dynamics have been conducted by such organizational theorists as Plowman, Baker et al. [68], who examined emergent conditions in an inner-city church, and Chiles, Meyer, and Hench [20], who studied the complex evolution of Branson, Missouri.

Conclusions

I believe that the best definition of complexity behavior is provided by Paul Cilliers [21]. To frame that definition, he argues that we often confuse complicated with complexity. A complicated system is one with many components interacting to perform complicated tasks, but all of those components maintain their original integrity. A jet plane, he said, is complicated: the components of the plane are unchanged by their relationship. Complex systems are composed of many interacting units, all of which are changed by that interaction such that they quickly loose their original form and function. Organizational theorists who study complex dynamics are seeking to observe and explain that dynamic [48,63]; to understand how organizations might be restructured to improve their capacity for complex outcomes [11]; to improve general leadership in organizations [51,57,80]; to restructure our understanding and implementation of strategic leadership [58], and to restructure how we think about reality, understanding, predictability, and scientific methodology [16,17,18,53,72].

Complexity is genuinely an interdisciplinary science, applicable equally to biology, physics, and the social sciences. Organizational sciences have experienced a number of cumbersome imports in the past: systems theory borrowed from biology [12], as did population ecology theory [33], and scientific management theory borrowed from mechanics [79]. Complexity theory, however, is not about biology or mechanics, it is about interaction dynamics – and this is common across multiple phenomena. It is also about change, not gradual change but change that surprises. It helps us understand how systems learn and adapt to their environment. It comes at a good time for organizational studies, and will help us adapt organizations to the dramatic, knowledge-oriented changes being experienced in world economies.

Future Directions

One is tempted to argue that organizational complexity theory represents a paradigm shift in the way we think about reality (in the Kuhnian sense; [46]), and there is considerable reason to do so. Complexity theorists, for example, acknowledges that the problems facing organizations are far too complex to be solved by rational thinking and planning, and that organizations must engage the interactive, complex capacity of all its employees to solve those problems [59]. Because of this, organizational complexity theory is probably one of the first truly postmodern theories of organization. Organizational complexity theory is premised on complexity mechanisms rather than on variables, thus the questions that emerge from this perspective are dramatically different from any that have been asked in the past (e. g., "how does level of interdependency interact with level of vision to influence temporal emergence of, and phase transitions in, organizational learning"). Complexity theory clearly separates itself from the notions of equilibrium that have so influenced 20th century organizational theory. It moves us away from the centrality of administration and focuses on the centrality of bottom-up, adaptive leadership. Its notions of leadership do not revolve around interpersonal relationships, as does most of leadership theory in the 20th century; rather, it revolves around interactive dynamics. Complexity theory proposes that organizational strategy should seek effective relationships, rather than competition, with other members of its niche. Simply put, complexity theory is a dramatically different way to understand organization.

Complexity theory has captured the attention of academics and practitioners. Important indications of this include the large number of articles on the subject that are appearing in scholarly journals (e. g., the August, 2007, edition of *The Leadership Quarterly* was a special issue on complexity leadership) and the fact that complexity-related articles are being published in top-tier journals (e. g., [11,68]); the increasing number of complexity-related grants in organizations (NSF began a complexity category for organizational scholars in 2007); and a significant surge in the number of practitioners who are seeking consultant relationships with complexity theorists.

A number of interesting research questions are emerging from the complexity paradigm. Andriani and McKelvey's work on extreme events could have significant implications for the way we approach Katrina-like tragedies in the future, for example. The work to develop methodologies such as agent based modeling can have tremendous impact on how we analyze organizational dynamics. The idea that organizational behavior is about the actions of mechanisms is beginning to sink in among scholars, and could dramatically realign the way we think about organization. Complexity leadership theory poses unique questions for the role of strategic leadership and challenges the very way we perceive leadership.

Arguably, there has not been a true paradigm shift in organization studies since Weber introduced his concept of bureaucracy [59]. At the risk of being overly optimistic, organizational complexity theory may be changing this.

Bibliography

1. Anderson P (1995) Microcomputer manufacturers. In: Carroll GR, Hannan MT (eds) Organizations in Industry. Oxford University Press, New York, pp 37–58
2. Andriani P, McKelvey B (2006) Extremes and scale-free dynamics in organization science: Some theory, research, statistics, and power law implications. Manuscript submitted for publication
3. Argyris C, Schön DA (1978) Organizational learning: A theory of action perspective. Addison-Wesley, Reading
4. Arthur WB (1989) The economy and complexity. In: Stein DL (ed) Lectures in the sciences of complexity, vol: 1. Addison-Wesley, Redwood City, pp 713–740
5. Ashby WR (1960) Design for a Brain, 2nd edn. Wiley, New York
6. Berson Y, Avolio BJ (2004) Transformational leadership and the dissemination of organizational goals: A case study of a telecommunication firm. Leadersh Q 15(5):625–646
7. Blau JR (1995) Art museums. In: Carroll GR, Hannan MT (eds) Organization in Industry. Oxford University Press, New York, pp 87–114
8. Boal K (2004) Strategic Leadership, Organizational Learning and Network Ties. Paper presented at the Conference on Strategic Leadership on Both Sides of the Atlantic, International Institute for Management Development, Lausanne, 2004
9. Boisot M (1998) Knowledge assets: Securing competitive advantage in the information economy. Oxford University Press, Oxford
10. Bonabeau E, Meyer C (2001) Swarm intelligence: A whole new way to think about business. Harvard Bus Rev 79(5):107–114
11. Brown SL, Eisenhardt KM (1997) The art of continuous change: Linking complexity theory and time-paced evolution in relentlessly shifting organizations. Adm Sci Q 42(1):1–34
12. Buckley W (1967) Sociology and modern systems theory. Prentice Hall, Englewood Cliffs
13. Canella AA, Monroe MJ (1997) Contrasting perspectives on strategic leaders: Toward a more realistic view of top managers. J Manag 23(3):213–230
14. Carley K (1992) Organizational learning and personnel turnover. Organ Sci 3:20–46
15. Carley K (1996) Organizational learning and personnel turnover. In: Cohen MD, Sproull LS (eds) Organizational Learning. Sage, Thousand Oaks
16. Carley K (1997) Organizational adaptation. An Oper Res 75: 25–47
17. Carley K (1999) On the evolution of social and organizational networks. In: Bacharach S, Knoke D, Andrews SB (eds) Research in the Sociology of Organizations: Networks In and Around Organizations, vol: 16. Press JAI, Greenwich, pp 3–30
18. Carley K (2002) Intra-organizational complexity and computation. In: Baum JAC (ed) The Blackwell Companion to Organizations. Blackwell, Oxford, pp 208–232
19. Carley K, Lin Z (1995) Organizational designs suited to high performance under stress. Trans IEEE Syst Man Cybern 25:221–230
20. Chiles T, Meyer A, Hench T (2004) Organizational emergence: The origin and transformation of Branson, Missouri's musical theaters. Organ Sci 15(5):499–519
21. Cilliers P (1998) Complexity and postmodernism: Understanding complex systems. Routledge, London
22. Cohen MD, Sproull LS (1996) Organizational Learning. Sage, Thousand Oaks
23. Conger JA (1999) Charismatic and transformational leadership in organizations: An insider's perspective on these developing streams of research. Leadersh Q 10(2):145–179
24. Coveney P (2003) Self-organization and complexity: A new age for theory, computation and experiment. Paper presented at the Nobel symposium on self-organization, Karolinska Institutet, Stockholm, 2003
25. Cusumano M (2001) Focusing creativity: Microsoft's "Synch and Stabilize" approach to software product development. In: Nonaka I, Nishiguchi T (eds) Knowledge emergence: Social, technical, and evolutionary dimensions of knowledge creation. Oxford University Press, Oxford, pp 111–123
26. Dansereau F, Yammarino FJ (eds) (1998) Leadership: Multiple-level approaches: Contemporary and alternative, vol: 2. Press JAI, Stamford
27. Davis GF, Marquis C (2005) Prospects for organization theory in the early twenty-first century: institutional fields and mechanisms. Organ Sci 16(4):332–343
28. DiMaggio PJ, Powell WW (1991) The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. In: Powell WW, DiMaggio PJ (eds) The new institutionalism in organizational analysis. University of Chicago Press, Chicago, pp 63–82
29. Elster J (1998) A plea for mechanisms. In: Hedström P, Swedberg R (eds) Social Mechanisms. Cambridge University Press, New York, pp 45–73
30. Goodwin B (1994) How the leopard changed its spots: The evolution of complexity. Charles Scribner's Sons, New York
31. Grey C, Garsten C (2001) Trust, Control and the Post-Bureaucracy. Organization Studies, vol: 22. Walter de Gruyter, New York, pp 229
32. Haken H (1983) Synergetics, an introduction, vol: 3. Springer-Verlag, Berlin
33. Hannan MT, Freeman J (1977) The population ecology of organizations. Am J Sociol 82:929–964
34. Hazy J, Marion R (2008) When Organizing in Extreme Situations, Is Far-From-Equilibrium (FFE) a Meaningful Concept?

Poster presented at the Organization Science Winter Conference, Squaw Creek, 2008
35. Heckscher C, Donnellon A (1994) Defining the post-bureaucratic type. New perspectives on organizational change Thousand Oaks: Sage, In: Heckscher C, Donnellon A (eds) The post-bureaucratic organization
36. Hedström P, Swedberg R (1998) Social Mechanisms: An analytical approach to social theory. Cambridge University Press, Cambridge
37. Hernes G (1998) Real virtuality. In. Hedström P, Swedberg R (eds) Social mechanisms: An analytical approach to social theory. Cambridge University Press, Cambridge, pp 74–101
38. Hitt MA (1998) Presidential Address: Twenty-first century organizations: Business firms, business schools, and the academy. Acad Manag Rev 23:218–224
39. Homans GC (1950) The human group. Harcourt, New York
40. Jaques E (1989) Requisite organization. Cason Hall, Arlington
41. Jermier JM (1998) Introduction: Critical perspectives on organizational control. Adm Sci Q 43(2):235–256
42. Kauffman SA (1993) The origins of order. Oxford University Press, New York
43. Kauffman SA (1995) At home in the universe: The search for the laws of self-organization and complexity. Oxford University Press, New York
44. Kelso JA (1995) Dynamic patterns: The self-organization of brain and behavior. Press MIT, Cambridge
45. Koschmieder EL, Davis S, Tvergaard V, Batchelor GK, Leibovich S (eds) (1993) Bernard cells and Taylor vortices–Cambridge monographs on mechanics and applied mathematics. Cambridge University Press, New York
46. Kuhn TS (1970) The structure of scientific revolutions, 2nd edn. The University of Chicago Press, Chicago
47. Langston CG (1986) Studying artificial life with cellular automata. Phys 22D:120–149
48. Lichtenstein B (2000) Dynamics of rapid growth and change: A complexity theory of entrepreneurial transitions. In: Liebcap G (ed) Advances in the study of entrepreneurship, innovation, and economic growth, vol: 6. Press JAI, Westport CT, pp 161–192
49. Lichtenstein B, Carter N, Dooley K, Gartner W (2007) Complexity dynamics of nascent entrepreneurship. J Bus Ventur 22:236–261
50. Lichtenstein B, Plowman D (2007a) Emergent leadership: A meso-model of leadership inspired by complexity. Manuscript submitted for publication
51. Lichtenstein B, Plowman D (2007b) The leadership of emergence: Meso-level leadership. Paper presented at the Festschrift for Jerry Hunt, Texas Tech, 2007
52. Lichtenstein B, Uhl-Bien M, Marion R, Seers A, Orton D, Schreiber C (2006) Leadership in Emergent Events: Exploring the Interactive Process of Leading in Complex Situations. Emerg Complex Organ 8(4):2–12
53. Lin F, Pai Y (2000) Using Multi-Agent Simulation and Learning to Design New Business Processes. Trans IEEE Syst Man Cybern Part A Syst Hum 30(3):380–384
54. Maravelias C (2003) Post-bureaucracy–control through professional freedom. J Organ Chang Manag 16(5):547
55. Marion R (1999) The edge of organization: Chaos and complexity theories of formal social organizations. Sage, Newbury Park
56. Marion R (2006) Complexity in organizations: A paradigm shift. In: Sengupta A (ed) Chaos, Nonlinearity, Complexity: The Dynamical Paradigm of Nature, vol: 206. Springer-Verlag, Berlin, pp 248–270
57. Marion R, Uhl-Bien M (2001) Leadership in complex organizations. Leadersh Q 12:389–418
58. Marion R, Uhl-Bien M (2007a) Complexity and strategic leadership. In: Hooijberg R, Hunt J, Antonakis J, Boal K, Lane N (eds) Being There Even When You Are Not: Leading Through Structures, Systems, and Processes, vol: 4. Elsevier, Amsterdam, pp 273–287
59. Marion R, Uhl-Bien M (2007b) Paradigmatic Influence and Leadership: The Perspectives of Complexity Theory and Bureaucracy Theory. In: Hazy JK, Goldstein J, Lichtenstein B (eds) Complex Systems Leadership Theory. Publishing ISCE, New York, pp 143–159
60. Maruyama M (1963) The second cybernetics: Deviation amplifying mutual causal processes. Am Sci 51:164–179
61. Maxwell JA (2004) Causal explanation, qualitative research, and scientific inquiry in education. Educ Res 33(2):3–11
62. McKelvey B (1999) Complexity theory in organization science: Seizing the promise or becoming a fad? Emerg 1(1):5–32
63. McKelvey B (2001) Energizing order-creating networks of distributed intelligence. Int J Innov Manag 5:181–212
64. McKelvey B (2008) Emergent Strategy via Complexity Leadership: Using Complexity Science and Adaptive Tension to Build Distributed Intelligence. In: Uhl M-Bien, Marion R (eds) Complexity leadership. Conceptual foundations, vol: 1. Information Age Publishing, Charlotte, pp. 225–268
65. McKelvey B, Boisot MH (2003) Transcendental organizational foresight in nonlinear contexts. Paper presented at the Conference INSEAD on Expanding Perspectives on Strategy Processes, Fontainebleau, France, 2003
66. Nonaka I, Takeuchi H (eds) (1995) The knowledge-creating company. Oxford University Press, Oxford
67. Pianka ER (1994) Evolutionary Ecology, 5th edn. HarperCollins, New York
68. Plowman D, Baker LT, Beck T, Kulkarni M, Solansky S, Travis D (2007) Radical change accidentally: The emergence and amplification of small change. Acad Manag J 50(3):515–543
69. Plowman D, Duchon D (2008) Dispelling the myths about leadership: From cybernetics to emergence. In: Uhl M-Bien, Marion R (eds) Complexity leadership. Conceptual foundations, vol: 1. Information Age Publishing, Charlotte, pp. 129–154
70. Plowman D, Solansky S, Beck T, Baker L, Kulkarni M, Travis D (2007) The role of leadership in emergent, self-organization. Leadersh Q 18:341–356
71. Prigogine I (1997) The end of certainty. The Free Press, New York
72. Schreiber C, Marion R, Uhl-Bien M, Carley K (2006) Multi-Agent Based Simulation of a Model of Complexity Leadership. Paper presented at the International Conference on Complex Systems, Boston, 2006
73. Schreiber C, Carley K (2006) Leadership style as an enabler of organizational complex functioning. Emergence Complex Organ 8(4):61–76
74. Schreiber C, Marion R, Uhl-Bien M, Carley K (2007) Multi-Agent Based Modeling of Complexity Leadership Theory. Paper presented at the North American Association for Computational Social and Organizational Sciences, Atlanta, 2007
75. Schroeder M (1991) Fractals, chaos, power laws. Freeman, New York

76. Stephenson WD, Bonabeau E (2007) Expecting the Unexpected: The Need for a Networked Terrorism and Disaster Response Strategy (Electronic Version). Homeland Security Affairs, 3 http://www.hsaj.org/?article=3.1.3
77. Sterman JD (1989) Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. Manag Sci 35(3):321–339
78. Stringer SH (2006) Entertaining the Future. International Consumer Electronics Show Las Vegas January – 5:2006 Retrieved November, 2007, http://www.sony.com/SCA/speeches/060105_stringer.shtml
79. Taylor F (1911) The principles of scientific management. Harper, New York
80. Uhl-Bien M, Marion R (2005) Complexity leadership theory. Paper presented at the Workshop on Leadership and Complexity, George Washington University
81. Uhl-Bien M, Marion R (2007) The mechanisms of emergence in complexity leadership theory: A meso-model of adaptive dynamics in organizations. Paper presented at the Festschrift honoring Jerry Hunt, Lubbock, 2007
82. Uhl-Bien M, Marion R, McKelvey B (2007) Complexity Leadership Theory: Shifting Leadership from the Industrial Age to the Knowledge Era. Leadersh Q 18:298–318
83. Volberda HW (1996) Toward the flexible form: How to remain vital in hypercompetitive environments. Organ Sci 7(4):359–374
84. von Bertalanffy L (1956) General systems theory. In: L v. Bertalanffy, Rapoport A (eds) General systems: Yearbook of the society for the advancement of general systems theory, vol: 1. George Braziller Publishers, New York, pp 1–10
85. Weber M (1947) The theory of social and economic organization (trans: Henderson AH, Parsons T). Free Press, Glencoe

Social Phenomena Simulation

Paul Davidsson[1], Harko Verhagen[2]
[1] Blekinge Institute of Technology, Ronneby, Sweden
[2] Stockholm University and Royal Institute of Technology, Stockholm, Sweden

Article Outline

Glossary
Definition of the Subject
Introduction
Why Simulate Social Phenomena?
Simulating Social Phenomena
Future Directions
Further Reading
Bibliography

Glossary

Agent (or software agent) A self-contained entity that has a state and that is situated (able to perceive and act) in an environment. In addition, agents are often assumed to be rational and autonomous.

Cellular automaton A mathematical structure modeling a set of cells that interact with their neighbors. Each cell has a set of neighbors and a state. All the cells update their values simultaneously at discrete time steps. The new state of a cell is determined by the current state of its neighbors according to a local function or rule.

Microlevel simulation A type of simulation in which the specific behaviors of specific individuals are explicitly modeled.

Definition of the Subject

Social phenomena simulation in the area of agent-based modeling and simulation concerns the emulation of the individual behavior of a group of social entities, typically including their cognition, actions, and interaction. Agent-based social simulation constitutes the intersection of three scientific fields, namely, agent-based computing, the social sciences, and computer simulation [6]. Agent-based computing is a research area mainly within computer science and includes, e. g., agent-based modeling, design, and programming. By the social sciences we here refer to a large set of different sciences that study the interaction among social entities, e. g., social psychology, management science, policy, and some areas of biology. Computer simulation concerns the study of different techniques for simulating phenomena on a computer, e. g., discrete-event, object-oriented, and equation-based simulation.

Introduction

Computer simulation consists of three main steps: (i) designing a model of an actual or theoretical system, (ii) executing the model on a computer, and (iii) analyzing the execution output. Already in the early days of computer development, simulation was used in different research areas to predict the behavior of complex systems. Such simulations were typically based on differential equations and focused on results at the aggregate level. These models of, for instance, predator-prey populations could result in fairly accurate models but were limited in the sense that the models excluded individual behavior and decision making, as well as interaction between individuals, and were based on homogeneous agents. The development of agent-based modeling offers a possible solution to this problem with its (seemingly) natural mapping onto interacting individuals with incomplete information and capabilities, no global control, decentralized data, asynchronous computing, and inclusion of hetero-

geneous agents. Agent-based simulation models also offer the possibility of studying the dynamics of the interaction processes instead of focusing on the (static) results of these processes [16,26].

Agent-based modeling can be traced back to von Neumann, who in the 1950s invented what was later termed *cellular automata*. These were used by Conway in the 1970s when he constructed the well-known *Game of Life*. It is based on very simple rules determining the life and death of the cells in a virtual world in the form of a 2-D grid. Inspired by this work, researchers developed more-refined models, often modeling the social behavior of groups of animals or artificial creatures. One example is the Boid model by Reynolds [24], which simulates coordinated animal motion such as bird flocks and fish schools. With respect to human societies, Epstein and Axtell [8] developed in the 1990s one of the first agent-based models, called Sugarscape, to explore the role of social phenomena such as seasonal migrations, pollution, sexual reproduction, combat, and transmission of disease. This work is in spirit closely related to one of the best-known and earliest examples of the use of simulation in social science, namely, the Schelling model [27], in which cellular automata were used to simulate the emergence of segregation patterns in neighborhoods based on a few simple rules expressing the preferences of the agents. Another pioneer from the 1950s worth mentioning is Barricelli [2], who to some extent used agent-based modeling for simulating biological systems.

The cellular automata models closely resemble the models used in statistical physics, which has inspired physicists to include the simulation of social phenomena in large-scale social systems in their research agenda. In this area, sometimes referred to as sociophysics, phenomena such as opinion spreading in a society and competition between languages have been studied. These models originally described the behavior of atoms and molecules, which are quite simple objects, and the macrolevel phenomena caused by their interaction (rather than by complex behavior of the individual as in the case of humans). Thus, in these models little attention is paid to individual variation and the individual decision making is rather primitively modeled. A prominent example of sociophysics is the work of Galam [10].

To sum up, we can identify two main approaches to social simulation:

- Macrolevel (or equation-based) simulation, which is typically based on mathematical models. It views the set of individuals (the population) as a structure that can be characterized by a number of variables.
- Microlevel (or agent-based) simulation, in which the specific behaviors of specific individuals are explicitly modeled. In contrast to macrolevel simulation, it views the structure as emerging from the interactions between individuals and thus exploring the standpoint that complex effects need not have complex causes.

As argued by Van Parunak et al. [21], agent-based modeling is most appropriate for domains characterized by a high degree of localization and distribution and dominated by discrete decision. Equation-based modeling, on the other hand, is most naturally applied to systems that can be modeled centrally and in which the dynamics are dominated by physical laws rather than information processing. We will here focus on agent-based models, particularly those that have a richer representation of the individual than the cellular automata and statistical physics models.

Why Simulate Social Phenomena?

Simulation of social phenomena can be done for different purposes, e. g.,

- Supporting social-theory building;
- Supporting the engineering of systems, e. g., validation, testing, etc.;
- Supporting planning, policy making, and other decision making;
- Training, in order to improve a person's skills in a certain domain.

It is possible to distinguish between four types of end users: *scientists*, who use social phenomena simulation in the research process to gain new knowledge, *policymakers*, who use it for making strategic decisions, *managers* (of systems), who use it to make operational decisions, and *other professionals*, such as architects, who use it in their daily work. We will now describe how these types of end users may use simulation of social phenomena for different purposes.

Supporting Social-Theory Building

In the context of social-theory building, agent-based simulation can be seen as an experimental method or as theories in themselves [26]. In the former case, simulations are run to test the predictions of theories, whereas in the latter case simulations in themselves are formal models of theories. Formalizing the ambiguous, natural-language-based theories of the social sciences helps to find inconsistencies and other problems, and thus contributes to theory building.

Using agent-based simulation studies as an experimental tool offers great possibilities. Many experiments with human societies are either unethical or even impossible to conduct. Experiments in silico, on the other hand, are fully possible. These can also breathe new life into the ever-present debate in sociology on the micro–macro link [1]. Agent-based models mostly focus on the emergence of macrolevel properties from the local interaction of adaptive agents that influence one another [17,26]. However, simulations in computational organization theory [4,22], for example, often try to analyze the influence of macrolevel phenomena on individuals. Using agent-based models to simulate the bidirectional relation between micro- and macrolevel concepts would provide tools to analyze the theoretical consequences of the work done by theorists such as Habermas, Giddens, and Bourdieu, to name a few [26].

Supporting the Engineering of Systems

Many new technical systems are distributed and involve complex interactions between humans and machines. The properties of agent-based simulation make it especially suitable for simulating these kinds of systems. The idea is to model the behavior of human users in terms of software agents. In particular, this seems useful in situations where it is too expensive, difficult, inconvenient, tiresome, or even impossible for real human users to test out a new technical system. Of course, also the technical system, or parts thereof, may be simulated. For instance, if the technical system includes hardware that is expensive and/or special purpose, it is natural to simulate also this part of the system when testing out the control software. An example of such a case is the testing of control systems for "intelligent buildings," where agents simulate the behavior of the people in the building [5].

Supporting Planning, Policy Making, and Other Decision Making

Here the focus is on exploring different possible future scenarios in order to choose between alternative actions. Besides this type of prediction, simulation of social phenomena may be used for analysis, i. e., to gain deeper knowledge and understanding of a certain phenomenon.

An area in which several studies of this kind have been carried out is disaster management, such as experiments concerning different roles and the efficiency of reactions to emergencies [18]. Based on individuals' observations, personal characteristics and skills, past experience and role characteristics, and social network, the agents create a plan to execute. Each agent represents a human being (acting in a particular role). The effect of adding a role (floor warden) in a fire alarm scenario upon the evacuation efficiency in an abstract environment is analyzed. In another approach, the agents are placed in an environment based on GIS (geographical information system) data, thereby tying the simulation closer to the physical reality [29]. In yet another study, real-world data were used for both the environment and the agents' internal decision-making model to analyze the effect of different insurance policies on the willingness of agents to pay for a disaster insurance policy [3].

Another application area for this type of simulation study is disease spreading. Typically, agents are used to represent human beings and the simulation model is linked to real-world geographical data. One study [32] also included agents that represent towns acting as the epicenter of disease outbreak. The town agent's behavior repertoire consisted of different containment strategies. The simulation model can be quickly adapted to local circumstances via the geographical data (given that there is data on the population as well) and is used to determine the effects of different containment strategies.

A third area where agent-based social simulation has been used to support planning and policy making is traffic and transport. An example of this is the simulation of all car travel in Switzerland during morning peak traffic [23].

Training

The main advantage of using simulation for training purposes is to be part of a real-world-like situation without real-world consequences. Especially in the military the use of simulation for training purposes is widespread. Also in medicine, where mistakes can be very expensive in terms of money and lives, the use of simulation in education is on the rise.

An early product in this area was a tool to help train police officers to manage large public gatherings such as crowds, demonstrations, and marches [31]. Another early example of agent-based simulation for training purposes is Steve [19,25]. Steve was an agent integrated with voice synthesis software and virtual reality software providing a very realistic training environment for controlling the engine room of a virtual US Navy surface ship.

An example of a more recent project is the PSI agent [15]. Whereas in most cases the simulator training is aimed at training practical skills or decision making, this work focuses on acquiring theoretical insights in the realm of psychological theory. The simulation enables students to explore psychological processes without ethical problems.

Simulating Social Phenomena

One of the first, and most simple, way of performing microlevel simulation is often called *dynamic microsimulation* [11,12]. It is used to simulate the effect of the passing of time on individuals. Data from a (preferably large) random sample from the population to be simulated is used to initially characterize the simulated individuals. Some examples of sampled features are: age, sex, employment status, income, and health status. A set of transition probabilities are used to describe how these features will change over a given time period, e. g., there is a probability that an employed person will become unemployed over the course of a year. The transition probabilities are applied to the population for each individual in turn and then repeatedly reapplied for a number of simulated time periods. Sometimes it is necessary to also model changes in the population, e. g., birth, death, and marriage. This type of simulation can be used to, e. g., predict the outcome of different social policies. However, the quality of such simulations depends on the quality of:

- the random sample, which must be representative, and
- the transition probabilities, which must be valid and complete.

In traditional microsimulation, the behavior of each individual is regarded as a "black box." The behavior is modeled in terms of probabilities and no attempt is made to justify these in terms of individual preferences, decisions, plans, etc. Also, each simulated individual is considered in isolation without regard to their interaction with others. Thus, better results may be gained if cognitive processes and communication between individuals are also simulated.

Opening the black box of individual decision making can be done in several ways. The first layer to add is often individual psychology; for instance, the so-called beliefs, desires, and intentions (BDI) model is often used. Models of individual cognition used in agent-based social simulation include the use of Soar (a computer implementation of Allen Newell's unified theory of cognition [20]), which was used in Steve (discussed in Sect. "Why Simulate Social Phenomena?").

For the simulation of social behavior the agents need to be equipped with mechanisms for reasoning at the social level (unless the social level is regarded as emerging from individual behavior and decision making). Several models have been based on theories from economics, social psychology, sociology, etc. An example of this is provided by Guye-Vuillème [13], who has developed an agent-based model for simulating human interaction in a virtual-reality environment. The model is based on sociological concepts such as roles, values, and norms and motivational theories from social psychology to simulate persons with social identities and relationships. Another example is the Consumat model [14], a metamodel combining several psychological theories on decision making in a consumer situation, used, for instance, to investigate different flood-management policies [3]. Also, nonsymbolic approaches such as neural networks have been used to model agents' decision making [18].

Future Directions

In a recent study of applications of agent-based simulation [7], it was concluded that even if agent-based simulation seems a promising approach to many problems involving the simulation of complex systems of interacting entities such as social phenomena, it seems that the full potential of the agent concept often is not utilized. For instance, most models have very primitive agent cognition, in particular if the number of agents involved is large.

Regarding future applications, Fiedrich and Burghardt [9] argue that agent-based simulation is a very promising approach to disaster management practice. In particular, agent-based social simulation in combination with sophisticated visualization techniques, such as virtual reality, in the form of "serious games," has the potential to provide very powerful training environments. In the context of military training, Stone [28] provides some interesting applications.

Further Reading

Classic books in the area of simulation of social behavior include "Growing Artificial Societies: Social Science from the Bottom Up" by Epstein and Axtell [8] and "Simulation for the Social Scientist" by Gilbert and Troitzsch [12].

More recent findings can be found in, e. g., the Journal of Artificial Societies and Social Simulation (http://jasss.soc.surrey.ac.uk) and the proceedings of, e. g., the International Workshop series on Multi-Agent-Based Simulation (MABS) (http://www.pcs.usp.br/~mabs/), the World Congress in Social Simulation (WCSS) [30], the conference of the European Social Simulation Association (ESSA) (http://www.essa.eu.org/), and the series of Agent Workshops in Chicago (http://www.agent2005.anl.gov/).

Bibliography

1. Alexander JC, Giesen B, Münch R, Smelser NJ (eds) (1987) The Micro-Macro Link. University of California Press, Berkeley
2. Barricelli NA (1957) Symbiogenetic evolution processes realized by artificial methods. Methodos 9(35–36):143–182

3. Brouwers L, Verhagen H (2004) Applying the Consumat model to flood management policies. In: Müller J, Seidel M (eds) 4th Workshop on Agent-Based Simulation. SCS, Montpellier, pp 29–34
4. Carley KM, Prietula M (eds) (1994) Computational Organization Theory. Erlbaum, Hillsdale
5. Davidsson P (2000) Multi-agent-based simulation: beyond social simulation. In: Moss S, Davidsson P (eds) Multi-Agent-Based Simulation. Lecture Notes in Computer Science, vol 1979. Springer, Berlin, pp 98–107
6. Davidsson P (2002) Agent-based social simulation: a computer science view. J Artif Soc Soc Simul 5(1)
7. Davidsson P, Holmgren J, Kyhlbäck H, Mengistu D, Persson M (2007) Applications of multi-agent-based simulation. In: Antunes L, Takadama K (eds) Multi-Agent-Based Simulation VII. Lecture Notes in Computer Science, vol 4442. Springer, Berlin
8. Epstein JM, Axtell RL (1996) Growing Artificial Societies: Social Science from the Bottom Up. MIT Press, Cambridge
9. Fiedrich F, Burghardt P (2007) Emergency response information systems: emerging trends and technologies: agent-based systems for disaster management. Commun ACM 50(3):41–42
10. Galam S (2008) Sociophysics: a review of Galam models. arXiv:0803.1800 http://arxiv.org/abs/0803.1800. Accessed 06 Aug 2008
11. Gilbert N (1994) Computer Simulation of Social Processes. Social Research Update, Issue 6, Department of Sociology, University of Surrey, UK. http://www.soc.surrey.ac.uk/sru/. Accessed 06 Aug 2008
12. Gilbert N, Troitzsch KG (2005) Simulation for the Social Scientist, 2nd edn. Open University Press, Maidenhead
13. Guye-Vuillème A (2004) Simulation of nonverbal social interaction and small groups dynamics in virtual environments. Ph D thesis, Ècole Polytechnique Fédérale de Lausanne, No 2933
14. Janssen MA, Jager W (1999) An Integrated Approach to Simulating Behavioural Processes: A Case Study of the Lock-in of Consumption Patterns. J Artif Soc Soc Simul 2(2)
15. Künzel J, Hämmer V (2006) Simulation in university education: the artificial agent PSI as a teaching tool. Simulation 82(11):761–768
16. Lansing JS (2002) "Artificial societies" and the social sciences. Artif Life 8:279–292
17. Macy MW, Willer R (2002) From factors to actors: computational sociology and agent-based modeling. Annu Rev Sociol 28:143–166
18. Massaguer D, Balasubramanian V, Mehrotra S, Venkatasubramanian N (2006) Multi-agent simulation of disaster response. In: Proceedings of the 1st International Workshop on Agent Technology for Disaster Management 2006. Hakodate
19. Méndez G, Rickel J, de Antonio A (2003) Steve meets Jack: the integration of an intelligent tutor and a virtual environment with planning capabilities. In: Intelligent Virtual Agents. Lecture Notes on Artificial Intelligence, vol 2792. Springer, Berlin, pp 325–332
20. Newell A (1994) Unified Theories of Cognition. Harvard University Press, Cambridge
21. Parunak HVD, Savit R, Riolo RL (1998) Agent-based modeling vs. equation-based modeling: a case study and users' guide. In: Sichman JS, Conte R, Gilbert N (eds) Multi-Agent Systems and Agent-Based Simulation. Lecture Notes in Computer Science, vol 1534. Springer, Berlin, pp 10–26
22. Prietula MJ, Carley KM, Gasser L (eds) (1998) Simulating Organizations: Computational Models of Institutions and Groups. MIT Press, Cambridge
23. Raney B, Cetin N, Völlmy A, Vrtic M, Axhausen K, Nagel K (2003) An Agent-Based Microsimulation Model of Swiss Travel: First Results. J Netw Spatial Econ 3(1):23–41
24. Reynolds CW (1987) Flocks, herds, and schools: a distributed behavioral model. Comput Graph 21(4):25–34 (SIGGRAPH '87 Conference Proceedings)
25. Rickel J, Lewis Johnson W (1999) Animated agents for procedural training in virtual reality: perception, cognition, and motor control. Appl Artif Intell 13:343–382
26. Sawyer RK (2003) Artificial societies – multi-agent systems and the micro-macro link in sociological theory. Sociol Meth Res 31(3):325–363
27. Schelling TC (1971) Dynamic models of segregation. J Math Sociol 1(1):143–186
28. Stone B (2005) Serious gaming. Def Manag J (31):142–144
29. Takahashi T (2007) Agent based disaster simulation evaluation and its probability model interpretation. In: Proceedings of the 4th International Conference on Intelligent Human-Computer Systems for Crisis Response and Management, 2007, Deltft, pp 369–376
30. Takahashi S, Sallach D, Rouchier J (2007) Advancing Social Simulation – The First World Congress. Springer, Berlin
31. Williams R (1993) An agent based simulation environment for public order management training. In: Western Simulation Multiconference, Object-Oriented Simulation Conference (OOS '93), San Diego 151–156
32. Yergens D, Hiner J, Denzinger J, Noseworthy T (2006) IDESS – A multi-agent-based simulation system for rapid development of infectious disease models. Int Trans Syst Sci Appl 1(1):51–58

Social Processes, Physical Models of

František Slanina[1,2]

[1] Institute of Physics, Academy of Sciences of the Czech Republic, Prague, Czech Republic

[2] Center for Theoretical Study, Prague, Czech Republic

Article Outline

Glossary
Definition of the Subject
Historical Introduction
Individualism Versus Cooperation
Opinion Dynamics
Bounded Confidence
Emergence of Social Classes
Future Directions
Acknowledgments
Bibliography

Glossary

Absorbing state In a stochastic process, a state from which there is no way out. Once the system reaches the absorbing state, it stays there forever.

Complete graph A graph where every pair of vertices is connected by an edge.

Configuration space The ensemble of all allowed states (configurations) a model system can reach. Any point in configuration space represents one such state.

Graph an assembly of points (vertices), some of which are connected by lines (edges).

Hypercubic lattice A generalization of a rectangular plane mesh; a graph embedded in a space of arbitrary dimensionality d. Assuming rectangular coordinates $x_1, x_2, \ldots, x_d$, the coordinates of the vertices in a hypercubic lattice are all whole numbers.

Markov process A special type of stochastic process. In a Markov process the system changes its state randomly and, most importantly, the changes of state are independent of history. A Markov process is a memoryless system.

Random walk A stochastic process describing hops of a particle to randomly chosen neighbor sites on a lattice.

Stochastic process A sequence of random events. At each time t, which may be either discrete ($t = 1, 2$, etc.) or continuous, a new random variable is introduced representing the outcome of the process in that time. Also called a random process.

Definition of the Subject

Modeling social phenomena as if they were manifestations of mutual interactions of physical objects is the ultimate goal of the reductionist approach to reality. Both the inanimate and animate worlds, including all the behavior of humans, would be traced back to the properties of atoms and molecules. This program is absolutely unrealizable, though. On the other hand, the discipline of **sociophysics** tries to bypass the brute-force approach by developing schematically effective models which aim at describing reality at a "macroscopic", rather than microscopic, level.

For example, when one wants to model the behavior of a large assembly of humans facing the necessity of choosing between two options, it is customary to neglect all details of the behavior of the people involved and describe their states by two-value quantities, such as $s = +1$ or $s = -1$; physicists call them spins. The interactions are often expressed using a cost function, which physicists call energy. The state with lowest energy is favored, but there are also external perturbations, or noise, which prevent the system from settling in that state. Physicists call the measure of the noise the temperature. The system of interacting spins at a certain temperature then serves as a model of the particular situation in human society.

It is a non-trivial question if such approach could work and if so, why. The essential ingredient for its functioning is **universality**. This notion, borrowed from statistical physics, means, that the behavior of the system does not ultimately depend on the details of that system, but only on its generic features. In plain words, for the description of the human mind, the behavior of a driver on a highway, or of a stockbroker, it is irrelevant whether the matter is composed of quarks, strings or whatever elementary particles may be discovered. The description of the world at the "macroscopic" level is detached from "microscopic" details.

It is the task of **sociophysics** to develop suitable macroscopic models that conform to observed facts about society and, at the same time, yield themselves to the methods of theoretical physics. There are non-negligible successes in this effort, yet still many areas remain unexplored. Consequently, the subject is alive and in a state of rapid progress.

Historical Introduction

It has been nearly two hundred years since the notion of social physics was introduced by the French philosopher Auguste Comte [1] in his attempt to revolutionize all thinking under the banner of positive philosophy. He himself abandoned that term later [2], but the idea of connecting social and physical phenomena in a unified frame has been around since. Later on, it somehow degenerated into a purely descriptive science, borrowing from physics nothing more than the tendency to exactitude in measurements; a prominent name to cite in this context is the Belgian statistician Adolphe Quételet. Little use stems from twopenny analogies, advocated by some later followers of Hegelian philosophy, between phase transitions, such as boiling water, and abrupt changes, mostly violent ones, in human society. Nonetheless, you may still find these considerations, based on the "quantity turns into quality" principle, in popular-philosophy brochures.

Skipping ahead a little more than a century and half since Comte, the first modern attempt to bridge the gap between social sciences and physics was marked by a memorable conference scheduled to take place in Moscow, 1 to 5 July 1974. Scientists both from the West and from the USSR were to discuss implications of physics for other fields, including social sciences and humanities. The organizing committee included people such as Kenneth Arrow, Nobel laureate for Economy and Hans

Bethe, Nobel laureate for Physics. Among the speakers was Andrei Sakharov, the well-known dissident and Nobel laureate for Peace. However, the Communist authorities stepped in, the conference was banned, most of the Russian participants were arrested, and a majority of them were exiled. However, many of the manuscripts were smuggled from the USSR to the West, and eventually the contributions were published in a proceedings volume [3]. A small sample of it appeared in [4].

In the 1980s, a very fruitful and general concept of synergetics was developed by H. Haken and his followers. The main idea consists in spontaneous emergence of coherent structures. In addition to being illustrated by examples from several branches of physical sciences (e. g. lasers) the idea was also elaborated in the context of social phenomena [5].

About the same time, the science of complexity started to gain wider popularity (and continues to do so to this day). The Santa Fe Institute in Santa Fe, USA, pioneered this research; it was here that complex system theory began to be systematically applied to economics and, more generally, to all human social behavior. Another Physics Nobel laureate, P. W. Anderson [6], has played a prominent role in promoting this path.

To finish this historical overview, the journey returns to France, where since the early 1980s the group of Serge Galam has been busy developing genuine physical models of social phenomena, also using he term sociophysics consistently [7,8,9]. Let us stop here and turn to several selected topical problems, where physically based models may help answer social scientists' questions.

Individualism Versus Cooperation

People never earn their living alone. From the darkest prehistorical past they have joined together to form gangs, stalking and hunting their prey. Without communication within the band, our ancestors could hardly have caught a mammoth, language would never have evolved and human brain capacity would probably never have exceeded that of a lemur. In short, cooperation between humans has been decisive in shaping the world around us. A tendency to cooperation is an inherent feature of human nature.

It was not until the middle of the eighteenth century or so that a different view started to spread. In this view, people are presented as selfish profit-seekers and if any cooperation is observed, it occurs *despite* the natural tendency to maximize personal gain, or as a by-product of that tendency. Under vague names such as neo-liberalism and social Darwinism, these and related ideas pervade the current thinking on human society.

On the other hand, empirically, cooperation is much more evolved now and assumes more sophisticated and complex forms than in the old times when it was considered self-evident. To detect this more complex cooperation, however strong it may be, requires external explanation. The starting point is a zero hypothesis of non-cooperation, from which cooperation emerges by some non-trivial mechanism to be discovered.

Prisoner's Dilemma Game

Games in General *Game theory* [10,11], introduced in 1940s by John von Neumann and Oskar Morgenstern provides very fertile ground on which to model the starting assumption of selfish individuals. In this model, agents are not only utterly selfish, but also absolutely rational. In a typical setup, agents meet in pairs and each of them chooses one of S possible strategies. If the first agent adopts strategy i and the second agent strategy j, then the first one gains an amount denoted A_{ij}, while the other's gain is A_{ji}. The matrix A is called a *payoff matrix*, and various types of games are distinguished according to its properties. For example, if the gain of one agent equals the loss of the other agent, so $A_{ij} + A_{ji} = 0$ for all i and j, it is a zero-sum game. If the gain of the winning party is always smaller than the loss of the adversary, the game is a negative-sum one, so it holds that $A_{ij} + A_{ji} < 0$ for all i and j.

It is assumed that the rules of the game, quantified in the matrix A, are known to both players. Therefore, they can build their strategies on rational analysis of the payoff matrix. It may happen that one strategy gives highest gain, irrespective of the action taken by the opponent. Formally, the first player's strategy, k, is such that $A_{ij} \leq A_{kj}$ for any strategy i of the first player and any strategy j of the second player. The same may also hold for the second player. Suppose that l is her best strategy, irrespective of the action of the first player. Obviously then, the first person always plays k while the second always plays l. If either player changes his or her strategy unilaterally, he or she is instantly worse off. Such a situation, if it happens, is called Nash equilibrium, after John F. Nash, a mathematician who devoted much of his career to applications of game theory in economics [12,13].

It is vital to understand that the notion of Nash equilibrium differs fundamentally from the usual equilibrium studied in various branches of physics. Whether the problem is to find the equilibrium of solid bodies on a lever or thermodynamic equilibrium in a system of steam, water and ice, the situation is formalized by finding a unique function to be minimized (energy, for example). On the contrary, Nash equilibrium means that every player maxi-

mizes his or her own function, with state of all other players fixed [14].

Introduction to Prisoner's Dilemma Collaboration has been studied using a very simple two-player game with two strategies, called the prisoner's dilemma game. Imagine that the police have arrested two accomplices for a burglary, but do not have enough evidence to prove that they actually committed the crime. The prisoners are kept well separated and the investigator offers a deal to each of them independently. Each criminal is given a promise that if he parts company with the other and confesses to the robbery they did together, he will be rewarded. If both of them confess, they will be jailed only for a short period; if one confesses and the other does not, then the first one is released and rewarded and the other gets a severe punishment; if neither of them confesses, both are released. Each must choose without knowing the other's choice.

The strategy of collaboration (C) dictates not to confess. In fact, it would be most beneficial to both suspects, if both collaborate. However, individually it is more tempting to defect (D) and confess to the police, as it prevents the situation in which the other confesses and the individual is punished alone. If both prisoners reckon in this way, the result is that both defect and must suffer some time in prison. Hence the dilemma.

The payoff matrix of this prisoner's dilemma game is characterized by four numbers, the gain if both defect, $A_{DD} = P$, or both cooperate, $A_{CC} = R$, and the gain of the defector $A_{DC} = T$ and the collaborator $A_{CD} = S$, if one of them defects and the other does not. So,

$$A_{PD} = \begin{array}{cc} & \begin{array}{cc} C & D \end{array} \\ & \begin{pmatrix} R & S \\ T & P \end{pmatrix} \begin{array}{c} C \\ D \end{array} \end{array} . \tag{1}$$

In order for the prisoner's dilemma to work as described above, the values must satisfy the inequalities $T > R > P > S$ and $2R > S + T$. The most studied values of the parameters are

$$T = 5\,, \quad R = 3\,, \quad P = 1\,, \quad S = 0\,. \tag{2}$$

It is easy to see that there is a single Nash equilibrium: the case in which both players defect. We arrived at this conclusion intuitively, above, and an exact check is even quicker. If the second player collaborates, the rewards of the first are R and T if he collaborates or defects, respectively; because $T > R$, defection is better. Similarly, if the second player defects, the gains for the first are S and P, and as $P > S$, it is again preferable to defect. So, defection is the optimal strategy.

Evolution of Cooperation

Iterated Prisoner's Dilemma The prisoner's dilemma game is rather trivial: if the players meet only once, defecting is certainly the best strategy. The Nash equilibrium excludes cooperation a priori, coinciding with the neo-liberalist view of society. What, then, is the mechanism behind the cooperation seen in reality? If there is a way for emergence of cooperation in the prisoner's dilemma game, perhaps we can see why reality sometimes follows that path.

Things become much more complex if the players face each other in an unlimited series of encounters. Robert Axelrod showed [15,16,17], in a series of computer tournaments, that complex strategies, taking into account past actions of the adversary, perform much better than simple defection. The generic and repeatedly confirmed outcome was that the strategy called Tit-For-Tat (TFT) outperformed all the rest of the strategies tried. It consists in playing cooperation at first and then repeating the last action of the other player.

Even more complicated behavior emerges if there is some level of noise. The players can make mistakes, or sometimes play at random, not obeying their usual strategies. New strategies may emerge due to mutations. Also relevant is the length of the players' memories, i. e. how many steps in the past they remember their opponents' actions. When all agents are put into a common room, where all of them they can play with all others, mixtures of strategies emerge [18,19]. The composition of the mixture depends on the length of the memory of the agents; if the memory length is unlimited, the evolution of the system proceeds indefinitely, introducing more and more complex strategies all the time.

Spatial Prisoner's Dilemma The trivial prisoner's dilemma game becomes complex when the players play again and again with the same opponent. But, extending the model to cover a wider field, how it could be that among billions of people a player finds the same opponent he or she has already faced? If the two were molecules in a well-stirred container, their chance to meet repeatedly would be virtually nil. Some *compartmentalization* of the agents must be imposed, in order to apply the iterative prisoner's dilemma to a large number of players. But the influence of compartmentalization can also be studied independently, preferably in its extreme version, where immobile agents are fixed at the vertices of a network, for example a two-dimensional square lattice: that is the structure used throughout this subsection.

Agents play the usual prisoner's dilemma game with their eight neighbors. The complexity stems from their

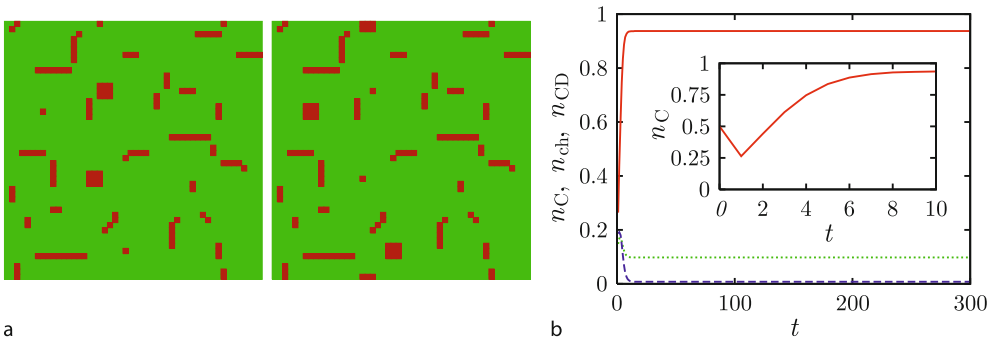

Social Processes, Physical Models of, Figure 1
Spatial prisoner's dilemma. In the *left two panels*, the configuration is drawn in times t_1, t_2, $t_1 < t_2$, on the lattice with $L = 50$. Cooperating agents are shown in *green*, the defectors in *red*. In the *right panel*, time evolution of the density of cooperators n_C (*solid line*), density of agents changing their state from one step to the next n_{ch} (*dashed line*), and density of interfaces n_{CD} (*dotted line*) are shown. System size is $L = 100$, data are averaged over 100 realizations. The incentive to defect is, $b = 1.05$. The initial condition is random spatial distribution of cooperators with density $n_C = 0.5$. In the *inset*, the detail for short times is shown

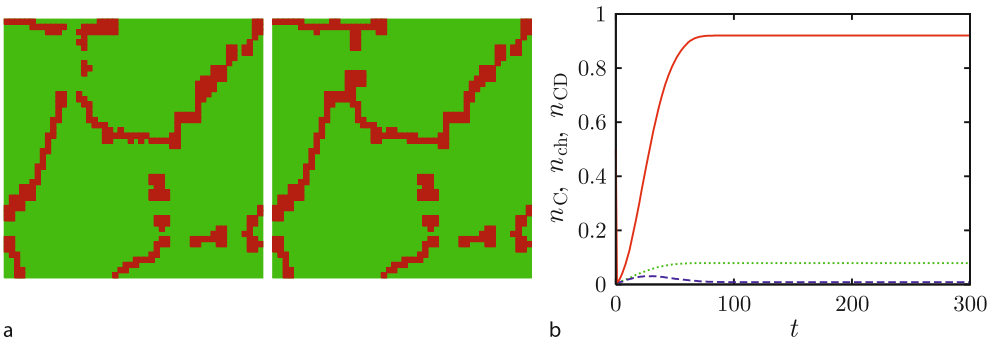

Social Processes, Physical Models of, Figure 2
Same as Fig. 1 with $b = 1.5$

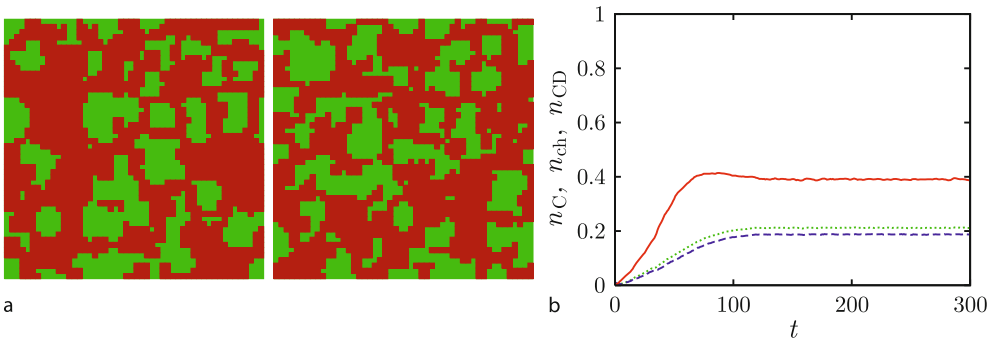

Social Processes, Physical Models of, Figure 3
Same as Fig. 1 with $b = 1.6$

ability to change their strategies from one step to the next [20]. For simplicity, let us consider only the 0-memory strategies: collaboration and defection. At each round, the agent adds her gain from the eight plays with her neighbors, who do the same in their turn. Each agent then decides upon her strategy for the next step. She looks at the gains of her neighbors and compares these numbers with her own. If any of the neighbors earns more than she, the agent adopts the current strategy of the most successful neighbor for the next step. Otherwise, the agent repeats her previous strategy. The evolution of the strategies proceeds by imitation.

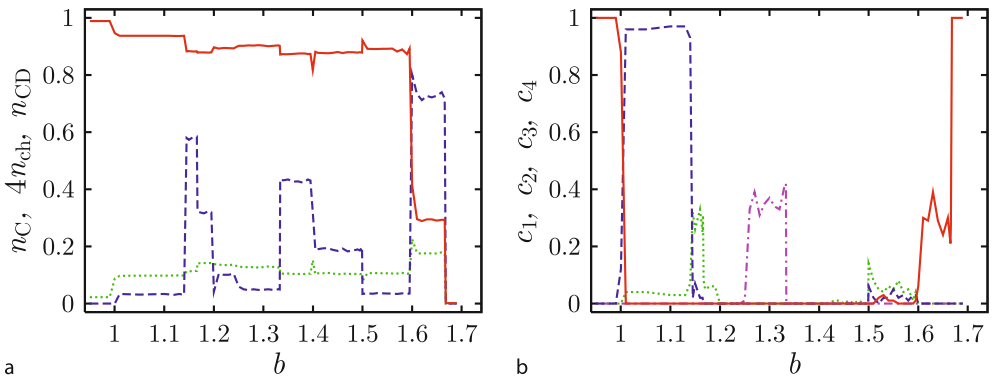

Social Processes, Physical Models of, Figure 4

Spatial prisoner's dilemma. Various parameters of the stationary state depending on the incentive to defect *b*. All data are calculated for a lattice with $L = 100$; initial configuration is a random spatial distribution of cooperators with density $n_C = 0.5$; data are averaged over 100 realizations. The *left panel* shows the density of cooperators (*solid line*), density of sites changing their state from one step to the next (*dashed line*), and density of bonds connecting cooperators with defectors (*dotted line*). The *right panel* shows the fraction of realizations reaching static configuration (*solid line*), and fractions reaching cyclic attractors with period 2 (*dashed line*), 3 (*dot-dashed line*), and 4 (*dotted line*)

The payoff matrix can be simplified so that there is only one control parameter, the temptation to defect b. It is assumed that $T = b > 1$, $R = 1$, and $P = S = 0$. The spatial prisoner's dilemma game is then simulated, starting from a random configuration, where every agent chooses cooperation or defection with equal probability. Further evolution proceeds by deterministic parallel dynamics, as described above. In fact, the system is a cellular automaton with specific, relatively complicated, update rules. The resulting configurations are illustrated in Figs. 1 to 3. The cooperators survive in significant proportion even if the temptation for defection b is large. However, their spatial arrangement strongly depends on the value of b. Figure 1 shows the situation for the value $b = 1.05$, only slightly above the lowest limit compatible with prisoner's dilemma inequalities. Small groups of defectors survive within a large sea of cooperators. Most of the defector groups are stable or change cyclically with a short period. When the temptation is increased to $b = 1.5$, isolated islands of defectors grow into strings joined together at some places, as seen in Fig. 2. Cooperation still prevails, but only within patches encircled by defectors. The spatial structure exhibits only small variations in time. This changes when the temptation grows further. Figure 3, where $b = 1.6$, already shows more defectors than collaborators and the arrangement changes chaotically.

Systematic study of the dependence on b is summarized in Fig. 4 showing data from simulations, averaged over many realizations of the initial conditions. First, the fraction of cooperators, n_C, remains quite high, above 80%, up to about $b = 1.6$, where it drops suddenly to about 40%; only above $b = 1.7$ do the cooperators vanish. Clearly, repeated games induced by fixed spatial arrangement of the players strongly encourages cooperation. Note that the players do not follow any strategy based on observation of the past behavior of the agents. The cooperation emerges spontaneously.

Further inspection of the dependence of n_C on b reveals sudden jumps at specific values of the temptation parameter. The jumps are even more pronounced in some other parameters characterizing the stationary state, namely the concentration n_{ch} of sites changing their state from one step to the next, and the concentration of bonds (neighbor pairs) connecting a defector and a cooperator, n_{CD}, which is the measure of the density of interfaces between cooperating and defecting domains. Furthermore, some realizations end in static configurations, while other reach a cyclic attractor with short periods 2, 3, or 4. The fraction of realizations corresponding to these four types are denoted c_1, c_2, c_3, and c_4, respectively. Of course, some initial conditions may also lead to longer stationary periods or to a quasi-chaotic state with a barely identifiable periodicity. Interestingly, the quantities c_i depend on b in a very irregular fashion. For example, the period-2 states dominate in an interval from $b = 1$ to about $b = 1.14$, but are rare elsewhere. In contrast, the period-3 states occur practically only in the interval $5/4 < b < 4/3$.

To understand these features it is necessary to analyze various spatial structures produced in the dynamics [21]. For example, an isolated defector in the sea of cooperators, as shown in Fig. 5a, has 8 cooperating neighbors, so its gain is $g_D = 8p$. The neighbors themselves have 7 cooperating neighbors, resulting in gain $g_C = 7$. This means that for $p < 7/8$ the defector becomes cooperator in the next step,

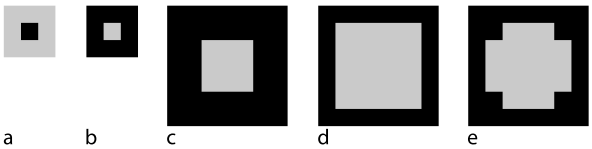

Social Processes, Physical Models of, Figure 5
Several configurations of the defectors (*black*) and cooperators (*gray*) in the spatial prisoner's dilemma game on a square lattice. Their stability is discussed in the text

but for $p \geq 7/8$ it survives. Similar analysis shows that an isolated cooperator surrounded by defectors, Fig. 5b, is never stable. Indeed, the cooperator gains 0, but each of the the neighboring defectors gains p.

It is possible to proceed further to more and more complicated geometries. For example, a 2×2 square of cooperators, Fig. 5c, is much more stable than an isolated cooperator. Each of the four gains $g_C = 3$ from its three cooperating neighbors. The defecting neighbors are of two types. At the corners, they have one cooperator to exploit, so their gain is p. The defectors adjacent to the edges of the square have two cooperating neighbors, giving higher gain $g_C = 2p$. Therefore, the square persists for $p \leq 3/2$. But that is not all; the defecting neighbors see that the cooperators gain more, so they become cooperators themselves in the next step. For $p < 3/2$ the initial 2×2 square grows into a 4×4 square of cooperators, Fig. 5d, which in turn expands into 6×6 square, and so on. Cooperation spreads despite the relatively large value of the temptation to defect!

Some configurations may exhibit periodic changes. A multitude of possible generalizations can be found in the literature, either concerning update rules [22,23,24,25] or geometry of the links connecting interacting neighbors [26,27,28,29,30]. However, the general features of the spatial prisoner's dilemma game remain in force, namely the fact that repeated plays against the same agents, which are dictated by the geometry of the social network, leads naturally to coexistence of large patches of collaborators alternated with various arrangements of defectors.

To sum up, spontaneous emergence of cooperation seems to be reproduced in model situations, on condition that the game is played repeatedly. The overall picture of cooperation is typically rather complex, precluding any simplistic ideology-based conclusions.

Opinion Dynamics

Voter Model

People can make up their minds by simply looking around and picking the opinions of randomly chosen neighbors.

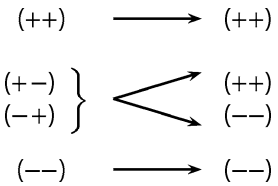

Social Processes, Physical Models of, Figure 6
Illustration of the dynamics of opinions in the voter model. The scheme shows how pairs of neighboring sites are updated. If the two sites have identical states, they remain unchanged. If the single-site states differ, they can become either both + or both − with equal probability. Conservation of average magnetization follows directly

That is the idea behind the stochastic process introduced in 1970s [31,32] and called the *voter model*. The voter model plays a special role among other models of opinion spreading and consensus formation, because it is exactly soluble in any spatial dimension, while showing highly non-trivial dynamics [33,34,35,36,37,38,39,40,41]. Physicists are interested in the voter model, as it can shed light on spinodal decomposition [35] and chemists use it to model catalytic reactions [36,37,42].

Definition In fact, there is a whole family of diverse voter models [43]; the exactly soluble class comprises the so-called linear voter models.

The most studied geometry is the d-dimensional hypercubic lattice Λ of linear dimension L, with periodic boundary conditions. The coordinates of the point $x \in \Lambda$ will be denoted x_α, $\alpha = 1, 2, \ldots, d$.

On each lattice site there is an agent whose state can be either $+1$ or -1. These two choices can represent a person's political preferences in a two-party system, hence the name voter model. The configuration of the entire system is described by a point in the configuration space $\sigma \in S = \{-1, +1\}^\Lambda$. The state of the site $x \in \Lambda$ is denoted $\sigma(x)$.

The dynamics of the model are very simple. In each step, one site x and its neighbor y are chosen randomly. Then x adopts the state of y, so $\sigma(x)$ is replaced by $\sigma(y)$, as illustrated in Fig. 6. This scheme also demonstrates one important property of the voter model. The average opinion, which a physicist would call magnetization, defined as $m = (1/|\Lambda|)\sum_{x\in\Lambda}\sigma(x)$ is conserved after averaging all possible realizations of the process, even though in individual realizations it may fluctuate. It is also evident that the uniform states where all sites are either $+1$ or -1 never change. These two configurations are *absorbing states* of the voter model.

In one dimension the dynamics can be easily understood. The configuration is determined by the sequence

of "domain walls", separating regions uniformly populated by $+1$ or -1. The configuration can change only by flipping the state of the sites beside the domain wall. Either the agent to the left of the wall adopts the state of the agent on the right side, or vice versa. In the former case the domain wall moves one step leftwards, in the latter case it jumps rightwards. Both possibilities have the same probability, so the domain wall performs a random walk. Moreover, when two domain walls meet, the region bordered by them disappears and the domain walls themselves annihilate. So, the one-dimensional voter model is exactly mapped onto the system of annihilating random walkers. This model is quite well understood [44]. Unfortunately, in any dimension larger than one such equivalence is no longer valid.

A slightly more formal description goes as follows: The voter model is a continuous-time Markov process σ_t which takes values in the configuration space $S = \{-1, +1\}^{\Lambda}$. For any $\sigma \in S$ denote as σ^x the state which is obtained from σ by flipping the state of site $x \in \Lambda$, so $\sigma^x(y) = (1 - 2\delta_{xy})\sigma(y)$. It is necessary to know the set of nearest neighbors of x on the lattice Λ. In the case of a d-dimensional hypercubic lattice, there are $2d$ neighbors obtained by shifting the point x by either $+1$ or -1 along the d Cartesian axes. Denote μth neighbor of x by x^{μ}. The transition rates describing a single flip are

$$w(\sigma, \sigma^x) = \frac{1}{2}\left[1 - \sigma(x)\frac{1}{2d}\sum_{\mu=1}^{2d}\sigma(x^{\mu})\right] \tag{3}$$

while all other transition rates are zero

$$w(\sigma, \sigma') = 0\,, \quad \text{if} \quad |\{x \in \Lambda : \sigma(x) \neq \sigma'(x)\}| \neq 1\,. \tag{4}$$

The dynamics proceeds according to the master equation

$$\frac{\mathrm{d}}{\mathrm{d}t}p_t(\sigma) = \sum_{x\in\Lambda}\left[w(\sigma^x, \sigma)p_t(\sigma^x) - w(\sigma, \sigma^x)p_t(\sigma)\right]. \tag{5}$$

Solution It is possible to directly write the equation for the average state on a single site $S(x, t) \equiv \langle\sigma_t(x)\rangle = \sum_{\sigma\in S} p_t(\sigma)\sigma(x)$, starting with the master equation (5). The result is

$$\frac{\mathrm{d}}{\mathrm{d}t}S(x, t) = \Delta_x S(x, t)\,, \tag{6}$$

where the notation $\Delta_x f(x) = -f(x) + \frac{1}{2d}\sum_{\mu=1}^{2d} f(x^{\mu})$ is used for the discrete Laplace operator on the d-dimensional hypercubic lattice. Equation (6) is in fact the discrete diffusion equation, describing the movement of a random walker over the lattice.

For two-site correlations $R(x - y, t) \equiv \langle\sigma_t(x)\sigma_t(y)\rangle$, the equation is

$$\frac{\mathrm{d}}{\mathrm{d}t}R(x, t) = 2\Delta_x R(x, t) \tag{7}$$

for $x \neq 0$. If the two points coincide, it holds trivially $R(0, t) = 1$ for all times $t \geq 0$. Technically, this is the boundary condition for the solution of the discrete diffusion Eq. (7). In this way it is possible to get closed equations for correlations of all orders.

Equation (6) can be solved by standard mathematical techniques, including Fourier and Laplace transforms. The initial condition is as follows: at the origin the state is $+1$, while all other sites are $+1$ or $+1$ with equal probability, so $S(x, 0) = \delta_{x0}$. The solution is

$$S(x, t) = \mathrm{e}^{-t}\prod_{\alpha=1}^{d} I_{x_\alpha}\left(\frac{t}{d}\right)\,, \tag{8}$$

where $I_\nu(z)$ is the modified Bessel function [45].

Asymptotic behavior of the Bessel function gives that, for large times, the average state of any site decays to zero as $S(x, t) \sim t^{-1/2}$, $t \to \infty$. Recall that the initial condition was $S(x, 0) = \delta_{x0}$. So, the average state of the site at the origin decays to zero monotonically, while the other sites, $x \neq 0$, exhibit first an increase in $S(x, t)$ and then decay at later times. This can be understood as propagation of a diffusive wave of $+1$'s from the origin to the rest of the lattice, eventually vanishing at large times.

More information on the dynamics is contained in the two-site correlation function $R(x, t)$. Apart from the factor 2, it obeys the same equation as $S(x, t)$. However, there is an important difference. Equation (7) holds only for $x \neq 0$ and, besides the initial condition $R(x, 0) = \delta_{0x}$, the solution must also satisfy the boundary condition $R(0, t) = 1$. Nevertheless, it is possible, again, to proceed by Fourier- and Laplace-transforming the Eq. (7). The boundary condition enters through the yet unknown function

$$n_{+-}(t) = \frac{1}{2}\left(1 - \frac{1}{2d}\sum_{\mu=1}^{2d} R(x - x^{\mu}, t)\right)\,, \tag{9}$$

which is simply the concentration of interfaces, i. e., the fraction of the bonds connecting sites with unequal state, also called *active bonds*. This quantity is an important measure of the level of activity in the system, as changes of the configuration can occur only at sites adjacent to active bonds.

Let us now separately discuss the results in dimensions $d = 1$, $d = 2$, and $d \geq 3$. In one dimension, the density of

interfaces can be expressed in closed form

$$n_{+-}(t) = \frac{1}{2}\mathrm{e}^{-2t}[I_0(2t) + I_1(2t)] \,. \tag{10}$$

For large times, the behavior is $n_{+-}(t) \simeq (4\pi t)^{-1/2}$ and $\lim_{t\to\infty} R(x,t) = 1$. This means that the activity measured by $n_{+-}(t)$ slowly decays to zero and eventually all the agents become fully correlated. Both features are signatures of complete ordering, which is the fate of the voter model in one dimension.

For $d = 2$ the situation is similar. Again, the density of interfaces decays to zero and the correlation function approaches 1 for large times. However, the evolution is much slower. The solution can be expressed in closed form using the so-called elliptic integrals, but let us mention here only the asymptotic behavior

$$n_{+-}(t) \simeq \frac{\pi}{2}\frac{1}{\ln(16t)} \,, \quad t \to \infty \,. \tag{11}$$

The behavior changes dramatically for $d \geq 3$, because $n_{+-}(t)$ has a positive limit for $t \to \infty$. In short, the activity never ceases and the voter model never reaches a totally ordered state. Explicit calculations show that the stationary density of interfaces becomes

$$n_{+-}(\infty) = \left(2d\int_0^\infty \left[\mathrm{e}^{-t}I_0(t)\right]^d \mathrm{d}t\right)^{-1} \,. \tag{12}$$

The asymptotic behavior of the Bessel function guarantees that the integral converges for $d > 2$ and $n_{+-}(\infty) > 0$. Numerical values for some values of d are listed in Table 1 below. One can also find how the density of interfaces approaches its asymptotic value, $n_{+-}(t) - n_{+-}(\infty) \sim t^{1-(d/2)}$.

All the above results for the voter model tacitly assume an infinitely large lattice. In practice, e. g., in numerical simulations, the system always consists of a finite number, $N = L^d$, of sites. This implies that eventually the dynamics lead to one of the two absorbing states, even for $d \geq 3$, where the *infinite* system never orders. The time needed to reach any of the absorbing states in a particular realization of the process is called *stopping time* τ_{st}. The typical scale for the stopping time, called *consensus time* $\tau(N)$, diverges in the thermodynamics limit. It is possible to show rigorously [46] that the consensus time grows like $\tau(N) \sim N^2$ in $d = 1$, $\sim N \ln N$ in $d = 2$ and $\sim N$ for $d \geq 3$. These formulae can be understood on an intuitive level, at least in one dimension, where the dynamics is equivalent to diffusion of domain walls, annihilating upon encounter. The size of homogeneous areas grows, therefore, as the mean displacement of a random walk, $\sim t^{-1/2}$. The time needed to cover the whole system of size N is therefore $\sim N^2$. A similar consideration is also feasible in higher dimensions [36].

The asymptotics of the voter model on d-dimensional hypercubic lattice are summarized in Table 1. Where the behavior is indicated by "$\sim$", the meaning is $t \to \infty$ or $x \to \infty$ or $N \to \infty$, according to the context.

Local Majority Models

Galam Model In democracy, consensus on an issue is rarely achieved by simply waiting until one of the options pervades the whole system through pairwise contact of individuals. Instead, there are various hierarchical levels of decision making, each of which comes to a conclusion based on the principle of majority. This is supposed to lead to a state in which most people are satisfied, when the opinion of the majority of those who participate in the decision process is declared a law. If there were only a single hierarchical level, namely a public referendum (Switzerland may serve as a model example), one could be quite sure that the outcome really represents the majority opinion in the society. However, as soon as there are more levels, and decisions on lower level are passed on to levels above, there is no obvious guarantee that the opinions are not distorted or even reversed. Serge Galam devised a simple model demonstrating that the distortions may not be an exception, but rather a rule [47,48,49,50,51].

The reason for breaking the society into several hierarchical levels is that the cost of communication in a too large collective is prohibitive. It is interesting that this is also the main technical obstacle hindering the introduction of secure electronic elections. One can imagine voting for a presidential candidate from one's home computer over the Internet, but to make the protocol reliable enough to ensure proper secrecy *and* resistance to any attempt at fraud on the scale of an entire nation seems to be beyond currently available technical capabilities [52].

So, the model takes it for granted that the society is organized hierarchically and, on each level, the decision is made within a small group, say, of 3 people. Within each group, majority rule tells what opinion shall be held by the representative of the group when sent to make decision on the upper level. Now, suppose there are only two possible choices, A and B and the fraction of people with opinion A on the level l is $n_A(l)$. On the basic level, $n_A(0)$ represents the concentration of A in the whole population.

Supposing that the people are not correlated in any way, the concentrations $n_A(l)$ provide full information on the system. Essentially, this is a kind of mean-field approx-

Social Processes, Physical Models of, Table 1
Table of some properties of the voter model on a hypercubic lattice in *d* spatial dimensions

d	$n_{+-}(\infty)$	$n_{+-}(t) - n_{+-}(\infty)$	$R(x,\infty)$	$\tau(N)$		d	$n_{+-}(\infty)$
1	0	$\sim t^{-1/2}$	1	$\sim N^2$		3	0.329731 . . .
2	0	$\sim (\ln t)^{-1}$	1	$\sim N \ln N$		4	0.403399 . . .
≥ 3	> 0	$\sim t^{1-d/2}$	$\sim \lvert x\rvert^{2-d}$	$\sim N$		5	0.432410 . . .
						∞	1/2

imation, neglecting any social structure or network within one hierarchical level. The model of majority decisions on a complete graph, also studied in the literature [53,54], is to large extent equivalent to the Galam model.

The dynamics of the model consists in determining the fraction of A at level $n+1$ on the basis of the concentration on level n. For groups of size 3 it leads to the recurrence relation

$$n_A(l+1) = n_A^3(l) + 3\,(1 - n_A(l))\, n_A^2(l)\,. \tag{13}$$

The evolution according to this rule is depicted in Fig. 7a. There are three fixed points at values $n_{A\,\mathrm{fix}} = 0, 1$, and $1/2$, which can be easily checked by insertion in (13). The two extreme values are stable, while the middle point is an unstable fixed point. This reminds us of physical systems with a phase transition, treated by renormalization group techniques [55]. In that formalism, the unstable fixed point corresponds to the critical value of a control parameter and the stable fixed points represent the various possible phases. Here there are two phases, populated uniformly by either all A or all B opinions, constituting total consensus on the issue.

The transition occurs at the symmetric point $n_{A\,\mathrm{fix}} = 1/2$, where exactly half of the people have opinion A. In this sense, the decision making is "fair" because it leads to consensus according to the initial majority. However, a seemingly minor modification makes a big difference. Imagine that the groups formed at each level are composed of an even number of persons, e. g., four. The decision must be taken as to what to do if exactly half of the group pushes for decision A but the rest for B. No majority emerges and the tie must be resolved. If there is an arbitrarily small bias towards one of the choices, say, B, the recurrence relation for the concentrations of opinion A is

$$n_A(l+1) = n_A^4(l) + 4\,(1 - n_A(l))\, n_A^3(l)\,. \tag{14}$$

Figure 7b shows what happens. The unstable fixed point is shifted significantly to higher values of n_A, namely to

$$n_{A\,\mathrm{fix}} = \frac{1+\sqrt{13}}{6} = 0.76759\ldots \tag{15}$$

which means that opinion B may win even if it was initially in a minority!

The situation can be even more complicated if some of the people systematically vote against the majority, instead of following the crowd. Such agents are called *contrarians* [56,57]. For simplicity, take (again) groups of size 3. First, the majority option within the group in question is

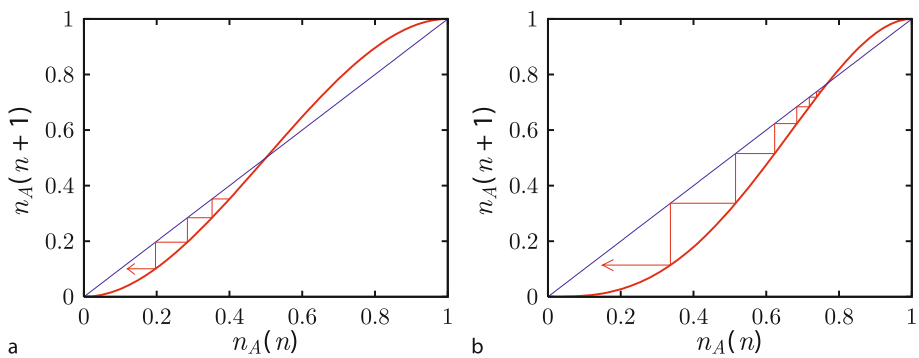

Social Processes, Physical Models of, Figure 7
Galam model. In a, graph of the recurrence relation for groups of size 3. The *stairs* indicate how the iterations of the relation drive the system away from the unstable fixed point $n_{A\,\mathrm{fix}} = 1/2$. In b, an analogous scheme for groups of size 4 with infinitesimal bias in favor of option *B*. Note the shift of the unstable fixed point to value $n_{A\,\mathrm{fix}} = (1+\sqrt{13})/6$

found, then each member of the group flips her opinion with probability p_c, which is to be interpreted as the concentration of contrarians in the population. The recursion relation is modified to

$$n_A(l+1) = (1-p_c)\left[n_A^3(l) + 3\,(1-n_A(l))\,n_A^2(l)\right] + p_c\left[(1-n_A(l))^3 + 3(1-n_A(l))^2 n_A(l)\right] \tag{16}$$

and results in a shift of the stable fixed points away from the endpoints of the interval $[0, 1]$. This means that total consensus is never reached; the contrarians always introduce a certain level of dissidence, but generally the final decision respects the initial majority option in the society. For $p_c < 1/6$ this picture remains valid, as there are three fixed points, with the unstable one keeping its position at $n_{A\,\mathrm{fix}} = 1/2$. However, if the concentration of contrarians rises and $p_c > 1/6$, the three fixed points coalesce to a single stable fixed point at $n_{A\,\mathrm{fix}} = 1/2$. No consensus is ever reached and the distribution of opinions tend to a precisely equilibrated state of equally represented opinions A and B. Several recent cases of popular referenda or presidential elections ending in extremely narrow victories come to mind, showing that the presence of contrarians may have palpable consequences for our lives.

A natural question is, how fast is consensus approached as one climbs higher and higher up ladder of hierarchies? It is evident that as we start closer to the unstable fixed point, it takes us more time to approach one of the stable points. To get a quantitative estimate, let us turn to the simplest case described by Eq. (13) and replace the discrete level index l by a continuous variable l which can be interpreted as time elapsed during the formation of the consensus. Thus, the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}l} n_A(l) = -n_A(l)(n_A(l)-1)(2n_A(l)-1) \tag{17}$$

is obtained, which can be solved relatively easily. The solution for several initial conditions is shown in Fig. 8a. Of course, starting from any point $n_A(0)$ inside the interval $(0, 1)$, the time to reach either of the stable fixed points 0 or 1 is infinite. However, in reality it is not necessary to reach the fixed point exactly, because the population consists of a finite number N of people, so full consensus is reached if the procedure is stopped at a distance $1/N$ from the fixed point. The time to achieve that diverges for increasing N as $\sim \ln N$, as will be seen from the explicit calculation.

To be more precise, the stopping time t_{st} as a function of the initial concentration $n_0 \equiv n_A(0) < 1/2$ can be defined by the formulae $n_A(0) = n_0$ and $n_A(t_{\mathrm{st}}(n_0)) = 1/N$, expressing the initial and final conditions, respectively. The case $n_0 > 1/2$ differs only in the final condition, which is $n_A(t_{\mathrm{st}}(n_0)) = 1 - 1/N$. Knowing the general solution of (17) the formula is

$$t_{\mathrm{st}}(n_0) = \ln\frac{n_0(1-n_0)}{(2n_0-1)^2} + \ln\frac{(N-2)^2}{N-1}\,. \tag{18}$$

Hence the already-mentioned logarithmic divergence of the stopping time for $N \to \infty$. The N-independent part of the stopping time can be seen in Fig. 8b. Note the divergence for $n_0 \to 1/2$, which can be regarded as a sign of a certain kind of dynamic phase transition.

The Galam model has potential for extensions in various directions and, indeed, it was further generalized [58,59,60,61,62,63,64,65,66,67,68]. For us, the most physically relevant question is how the dynamics change

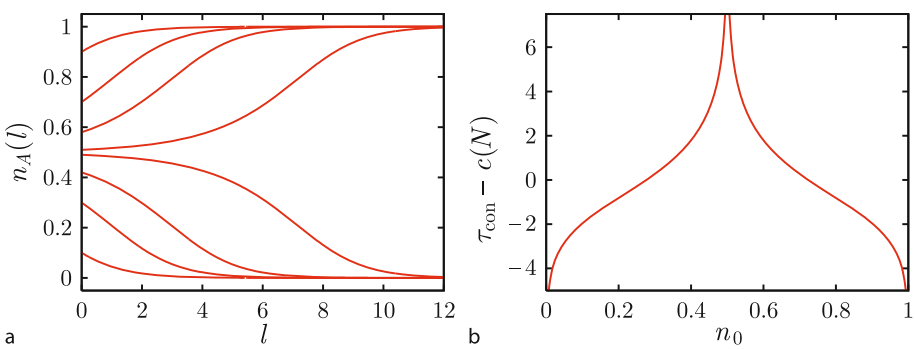

Social Processes, Physical Models of, Figure 8

Galam model with continuous levels, for group size 3. The evolution according to Eq. (17) is shown in a, with initial concentration of opinion *A* equal n_0 = 0.9, 0.7, 0.58, 0.51, 0.49, 0.42, 0.3, and 0.1 (from *top* to *bottom*). In b, the time to reach consensus in a society of *N* people is shown. In this graph, $c(N) = \ln\frac{(N-2)^2}{N-1}$

if the people are distributed on a fixed lattice or network [69,70,71]. After all, the Galam model has all the attributes of a mean-field version of some more complicated model, even though it may not be obvious which one [72]. The next section will make the point more clear.

Sznajd Model Clearly, a group of people puts stronger pressure on an individual than each member of the group taken separately. Nature designed humans so that they follow the crowd. If you see two or more people sharing an opinion on certain issue, you are tempted to join. This is the basic idea behind the model invented by Katarzyna Sznajd-Weron and Józef Sznajd (the daughter and the father) [73]. In its first version, the model was defined on a one-dimensional lattice of length N. Each site is inhabited by an agent which can be in two states, denoted $+1$ and -1, as in the voter or majority-rule models. It may correspond to people choosing between two dominant brands of a certain product in the market or voting in a two-party political system. In each step of the dynamics, a pair of neighbors is chosen randomly. If they are in the same state, say, $+1$, then the two sites adjacent to the pair adopt the same opinion $+1$, propagating the consensus outwards. Conversely, if they differ in opinion, they propagate the dissensus. The rule is shown schematically in Fig. 9.

Formally, suppose the chosen pair is in state $(\sigma(x), \sigma(x+1))$, at time t. If $\sigma(x) = \sigma(x+1)$, the neighbors of the pair are updated as $\sigma_{t+1/N}(x-1) = \sigma_{t+1/N}(x+2) = \sigma(x)$, while for $\sigma(x) \neq \sigma(x+1)$ the update rule is $\sigma_{t+1/N}(x-1) = \sigma(x+1), \sigma_{t+1/N}(x+2) = \sigma(x)$. Unlike the voter or majority rule model, there are three absorbing states. Besides the obvious "ferromagnetic" states of all $+1$ or all -1, there is the "antiferromagnetic" state where the sites in states $+1$ and -1 alternate regularly. Strictly speaking, there are two such states, one characterized by $+1$ at sites with odd coordinates and the other by $+1$ at even sites.

Looking at the rules more thoroughly, it is evident that the linear chain can be divided into two sublattices, one of them containing all odd coordinates x, the other all even sites. The state of the agents in one sublattice is never influenced by the agents in the other one. Moreover, the dynamics within one sublattice are a trivial modification of the voter model. The only difference is that in the voter model, one site induces change of state of one of its neighbors, while here the agent induces *both* of its neighbors adopt its state. Indeed, the update rule can be written as $\sigma_{t+1/N}(x-1) = \sigma(x+1)$, $\sigma_{t+1/N}(x+2) = \sigma(x)$, irrespective of the relation between $\sigma(x)$ and $\sigma(x+1)$ [74]. This observation simplifies the model to a large extent. Indeed, it is equivalent to two copies of the voter model evolving in parallel, where the only coupling between them comes from the fact that the sites to be updated are neighbors on the original lattice. The two voter models evolve so that the update occurs at the same place in both copies. However, the initial state of both sublattices is chosen randomly and independently and many properties, for example the concentrations of $+1$ in either of the sublattices, remain uncorrelated forever.

Restricting the study to one sublattice only, the solution follows the same way as was used successfully for the voter model. The flipping rates are analogous to (3), differing only in factor 2, originating from the fact that a site induces a change in state of its two neighbors. This factor means only a rescaling of time, so it is quite safe to neglect it. The result for the voter model can be translated directly to our case.

As a first application, let us look at the asymptotic states. The probability of reaching the state of all $+1$ is simply the concentration of $+1$ opinions in the initial state. This holds for both sublattices independently. From here, it is straightforward to deduce the probabilities P_+, P_- and P_{AF} that the system ends in absorbing states with all $+1$, all -1, and the antiferromagnetic state, respectively, as a function of the initial concentration of $+1$ opinions. Indeed, a state with all $+1$ means that both sublattices reached the uniform $+1$ state (and similarly for -1) while the antiferromagnetic state is obtained if one of the sublattices ended in a $+1$ state and the other in a -1 state. Hence

$$\begin{aligned} P_+(n_+) &= n_+^2 \\ P_-(n_+) &= (1-n_+)^2 \\ P_{AF}(n_+) &= 2(1-n_+)n_+ \,. \end{aligned} \tag{19}$$

The history of a single agent is also interesting. During its evolution, it may change its state several (including zero) times. The time elapsed between two subsequent changes of state is called *decision time* t_{dec}. It is natural to ask, what is the probability distribution of decision times of all agents? The answer relies on mapping the evolution of the voter model onto diffusive motion of domain walls. As explained in Sect. "Voter Model", the domain walls separating the regions of $+1$ and -1 opinions, evolve in time like annihilating random walkers. Selecting an agent at site x, its state flips if and only if the domain wall crosses the point x. The decision time is nothing other than the time between two successive visits of the random walker to the same position x, or, using language familiar to experts in random walks, the decision time is the time of first return of the random walk to the origin. Calculation of this quantity can be found in probability theory textbooks.

$$\text{Model of Ref. [73]} \quad \left\{ \begin{array}{lcl} (\cdot\,+\,+\,\cdot) & \longrightarrow & (+\,+\,+\,+) \\ (\cdot\,-\,-\,\cdot) & \longrightarrow & (-\,-\,-\,-) \\ (\cdot\,-\,+\,\cdot) & \longrightarrow & (-\,+\,-\,+) \\ (\cdot\,+\,-\,\cdot) & \longrightarrow & (-\,+\,-\,+) \end{array} \right\} \text{Sznajd model}$$

Social Processes, Physical Models of, Figure 9
Illustration of the dynamics of opinions in the the model of Sznajd–Weron and Sznajd, as it appears in the first paper [73] and as it was modified subsequently and named the "Sznajd model". The dots replace any state of the site

The decision time distribution behaves asymptotically as a power law

$$P^{>}_{\text{dec}}(t) \equiv \text{Prob}\{t_{\text{dec}} > t\} \sim t^{-1/2}\,, \quad t \to \infty \tag{20}$$

and this is exactly what was found numerically in the founding work of the Sznajds.

The solution obtained by mapping to the voter model is only the beginning of the story. Things start to be exciting again as soon as the "simplification" of the dynamic rules introduced above is made, allowing only neighbor pairs with equal states to influence their neighborhood. If the two agents in the pair do not agree in their opinions, nothing happens. In the scheme shown in Fig. 9 this corresponds to taking only first two rows as allowed updates. It is this modification that has been widely studied, and common consensus has assigned it the name Sznajd model (although calling it "Sznajds' " would perhaps do more justice to the authors).

As a first step in the analysis of the Sznajd model, it is important to note that there are again three absorbing states, all $+1$, all -1 and the antiferromagnetic one. But now the antiferromagnetic state is unstable, because randomly flipping a single site results in a nucleus of three sites in the same state, which irresistibly invades the whole system. But although unstable, the existence of such an absorbing state leaves important traces in the dynamics, as will be more clear later.

The transition rates for the underlying Markov process are

$$\begin{aligned} w(\sigma, \sigma^x) = \frac{1}{8}\Big[&\sigma(x-2)\sigma(x-1) + \sigma(x+1)\sigma(x+2) \\ &- \sigma(x)\big(\sigma(x-2) + \sigma(x-1) + \sigma(x+1) \\ &+ \sigma(x+2)\big) + 2\Big]\,. \end{aligned} \tag{21}$$

It is instructive to compare this with Eq. (3) describing voter model dynamics. The Sznajd model is a member of the family of non-linear voter models. This means that the transition rates depend non-linearly (quadratically, in fact) on the states of sites other than x.

There is a rather standard technique of approximate solution of similar systems, called the Kirkwood approximation. In our case, its use yields the probability for ending in the all $+1$ state in the form

$$P_+(n_+) = \frac{n_+^2}{(1-n_+)^2 + n_+^2}\,. \tag{22}$$

Figure 10 compares this result with numerical simulations, showing very good agreement.

Two more questions have been asked regarding the one-dimensional Sznajd model. The first is the decision time, introduced a short while ago. It was found numerically [74] that it follows the same power law (20) as in the previous case, where it resulted from underlying linear voter dynamics. For the Sznajd model, no proof is available showing that this is the exact result, but it is possible to understand this result on an intuitive level. Although the transition rates (21) correspond to a non-linear, rather than linear, voter model, the evolution of domain walls, responsible for the behavior of the decision time, remains very similar. If the domain walls are far from each other, they again perform a random walk. Non-linearity comes into play only when the domain walls come

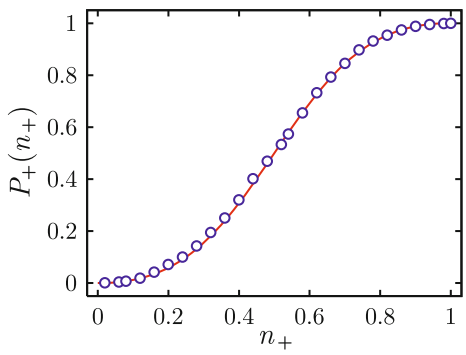

Social Processes, Physical Models of, Figure 10
One-dimensional Sznajd model. Probability of reaching the final configuration of all sites in state $+1$, depending on the initial concentration n_+ of $+1$ sites. The *points* are numerical simulation data extracted from [75], the *line* is the analytical result (22) obtained using the Kirkwood approximation

close together, more precisely when they are one lattice spacing apart. Then, they can no longer move freely, they can make only one step towards one another, thus annihilating each other. The dynamics is again those of a diffusion-annihilation process, but the domain walls are not independent; in addition to annihilation they interact at a short distance.

The interaction is in fact a trace of the unstable antiferromagnetic absorbing state. Two walls at distance one form a small nucleus of antiferromagnetic state, thus slowing down the dynamic with respect to free annihilating random walkers. If by chance several domain walls come so close that the distance between neighboring walls is 1, they form an antiferromagnetically ordered cluster, where the walls squeezed between another two walls from both the left and right sides cannot move at all and the dynamics is hindered even more.

However, for large times, the density of domain walls is low and the influence of such antiferromagnetic islands can be neglected. Therefore, it is no big surprise that the power-law dependence of the decision time remains unaffected, at least for large times.

The second interesting question is how many agents never changed their opinion up to certain time t, or, what is the *persistence* in the dynamics of the Sznajd model? In the voter model in one dimension, there is an exact analytical solution [44] showing that the fraction of sites which retain their initial opinion at least to time t decays as $\sim t^{-3/8}$. It may seem quite surprising that the same behavior was found numerically in the one-dimensional Sznajd model [76,77]. However, applying the same arguments as above, showing that the long-time dynamics is essentially dominated by annihilating random walks of the domain walls, it follows that it is quite plausible that the long-time behavior of the persistence is governed by the same exponent 3/8 both in the voter and the Sznajd models in one dimension.

Agents organized on a line are now understood, and it is natural to ask what changes if they are arranged in a two-dimensional mesh, like the people scattered on the surface of the Earth. For example, on a square lattice, one-dimensional dynamics can be generalized in such a way that a pair of neighboring agents is chosen, and if the two have the same state, they make all their 6 nearest neighbors share their state [78,79]. Slightly more formally, let x be a random coordinate on a square lattice with $N = L^2$ sites and $y = x^\nu$ one of the 4 neighbors of x. If $\sigma_t(x) = \sigma_t(y)$, then the configuration is updated as $\sigma_{t+1/N}(x^\mu) = \sigma_{t+1/N}(y^\mu) = \sigma_t(x)$ for all $\mu \in \{1, 2, 3, 4\}$. Numerical simulations show an interesting difference from the one-dimensional case. The probability $P_+(n_+)$ has discontinuity at $n_+ = 1/2$, indicating a dynamical phase transition, while in one dimension the dependence is smooth. More strongly, the data suggest that $P_+(n_+) = 0$ for $n_+ < 1/2$ and $P_+(n_+) > 0$ in the opposite case $n_+ > 1/2$. This result is equally valid in any dimension larger than 1. In the limit of infinite dimension, the behavior should coincide with the mean-field approximation, which implies putting the Sznajd model on a complete graph [54]. Such a modification is analytically solvable, as we now show.

On a complete graph with N vertices, the configuration of the system is fully described by the number N_+ of agents in state $+1$ or by the magnetization $m = 2N_+/N - 1$. The evolution of the latter is governed by transition rates which in the limit $N \to \infty$ yield the Fokker–Planck equation

$$\frac{\partial}{\partial t} P_m(m, t) = -\frac{\partial}{\partial m}\left[\frac{1}{2}(1 - m^2) m P_m(m, t)\right] \tag{23}$$

for the probability density $P_m(m, t)$ of the magnetization at time t. Before proceeding further, note that the average magnetization $\langle m\rangle(t) = \int m P_m(m, t)\mathrm{d}m$ satisfies the equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle m\rangle(t) = \frac{1}{2}\left(\langle m\rangle(t) - \langle m^3\rangle(t)\right) , \tag{24}$$

which has a structure similar to Eq. (17) describing the dynamics of the Galam model. Indeed, equating $m = 2n_A - 1$, identifying the level index with time, $l = t$, and making the approximation $\langle m^3\rangle = \langle m\rangle^3$, thus neglecting fluctuations, the Eqs. (17) and (24) coincide. So, the Galam and Sznajd models are solved in one shot.

It can be easily verified that the general solution of Eq. (23) has the form

$$P_m(m, t) = \frac{1}{(1 - m^2)m} f\left(\mathrm{e}^{-t/2}\frac{m}{\sqrt{1 - m^2}}\right) \tag{25}$$

for arbitrary function $f(y)$. The latter has to be determined from initial conditions, and assuming that at the beginning the magnetization was m_0 for all realizations of the process, it occurs that $f(m/\sqrt{1 - m^2}) = (1-m^2)m\delta(m-m_0)$. Clearly, in this case the probability density consists of a single δ-function which moves, as time passes, towards one of the ends of the allowed interval for the magnetization, $m \in [-1, 1]$. The dynamics is those of a pure deterministic drift and no diffusion ever smears out the evolving probability packet, as can be seen immediately from the absence of a second-derivative term in Eq. (23). Clearly, this is due to the limit of infinite system size and for finite N there is an additional diffusive term in the Fokker–Planck equation, proportional to N^{-1}.

The deterministic nature of the evolution of opinions has profound consequences. First, the initial sharp δ-function distribution of magnetization remains sharp until the end, thus justifying the neglect of fluctuations and the replacement $\langle m^3 \rangle = \langle m \rangle^3$. The fluctuations are set on only at times which diverge when $N \to \infty$ and the equivalence to the Galam model is exact in the thermodynamic limit. In a finite system of size N, the typical time $t_{\rm st}$ to reach the absorbing state, starting from magnetization m_0, can be estimated by requiring that drift brings the magnetization to the distance of order $\sim 1/N$ from either of the extremal points $m = \pm 1$. Expressed in terms of the initial concentration of the $+1$ opinion, $n_+ = (m_0 + 1)/2$, it is

$$t_{\rm st} \simeq \ln \left(\frac{(1 - n_+) n_+ N}{(2n_+ - 1)^2} \right) . \tag{26}$$

So, having clarified the close relation to the Galam model, it comes as no surprise that the Eqs. (26) and (18) coincide for large N.

For any positive initial magnetization the final state is always the uniform configuration of all agents in state $+1$, and vice versa. The probability of ending in the all $+1$ state is the step function $P_+ = \theta(n_+ - 1/2)$. This is consistent with simulations of the Sznajd model in dimensions larger than 1 and confirms the existence of a dynamical phase transition at $n_+ = 1/2$. It also shows that in a society where everybody interacts with everybody, but no person changes her opinion unless she meets at least two other people who compel her to do so, the initial majority, however narrow it is, always takes all. No chance is left to minorities. Fortunately enough, reality is more complex. The next section describes model implementations of one common "complication", that the number of choices is not limited to two but can be large, or even very large.

Bounded Confidence

Axelrod Model

Definition One of the features characteristic of the way culture is shared and propagated around the globe is that similar cultures are much more prone to mutual convergence, while incompatible lifestyles often coexist side by side without visibly influencing each other. Robert Axelrod introduced a model [80] nicely describing such a situation.

In the Axelrod model, contrary to the voter or Sznajd model, the character of each of the agents is given by more than one feature. One can think of tastes regarding food, sports, music, etc. These categories represent the features. For each feature the taste can assume various values, e. g., somebody likes eating raw vegetables, spending whole days in the fitness center and listening to Mozart in the evenings, while somebody else feeds on French fries, watches football on TV and adores the pop-star of the season. If two neighbors do not agree on any feature, they are so different that they do not influence each other. Conversely, if they find at least one feature where they share the same preference, one of them may find a second feature in which they do differ and change the preference on that second feature so that it agrees with the preference of the neighboring agent. The fact that the agents do not always interact, but do so only if they have something in common, is called *bounded confidence*.

More precisely, there are $N = L^d$ agents placed again on the d-dimensional hypercubic lattice $\Lambda = \{0, 1, \ldots, L-1\}^d$ with periodic boundary conditions and endowed with $F > 1$ integers. The numbers represent the values of its F features, so the state of the agent at site $x \in \Lambda$ is described by the vector with coordinates $\sigma(x, i) \in \mathbb{Z}$, $i = 1, 2, \ldots, F$. The configuration space of the model is $\mathbb{Z}^{\Lambda \times \{1,\ldots,F\}}$ and the evolution of the configuration σ_t is a Markov process determined by transition rates

$$w(\sigma, \sigma^{x,f,a}) = \frac{1}{2d} \sum_{\mu=1}^{2d} \theta \left(A(x, x^\mu) \right) \times \left[F - A(x, x^\mu) \right]^{-1} \delta_{a,\sigma(x^\mu,f)} , \tag{27}$$

where the summation goes over the set of $2d$ neighbors x^μ of the site x. We denoted here $\sigma^{x,f,a}$ the configuration which differs from σ only in feature f, so $\sigma^{x,f,a}(x, f) = a$ and $\sigma(x, f) \neq a$. The function $A(x, y) = \sum_{g=1}^{F} \delta_{\sigma(x,g),\sigma(y,g)}$ counts the number of features on which the agents at positions x and y agree, and $\theta(x)$ is the Heaviside function, $\theta(x) = 1$ for $x > 0$ and zero otherwise. The first factor after the sum in (27) accounts for the condition that the neighbors must agree in at least one feature and the second factor is here due to the fact that in each update the agent chooses randomly among $F - A(\sigma(x), \sigma(x^\mu))$ features in which she disagrees with her neighbor.

Simulations If the two factors after the sum in (27) were absent, the transition rates would depend linearly on the states of the neighbors and a generalization of the linear voter model would emerge, with all its beautiful solubility. However, the non-linearity is there and makes the model non-trivial. Lacking suitable analytical tools, it is necessary to rely on numerical simulations.

To be precise, the rules of the dynamics of the Axelrod model implemented in simulations are as follows: On

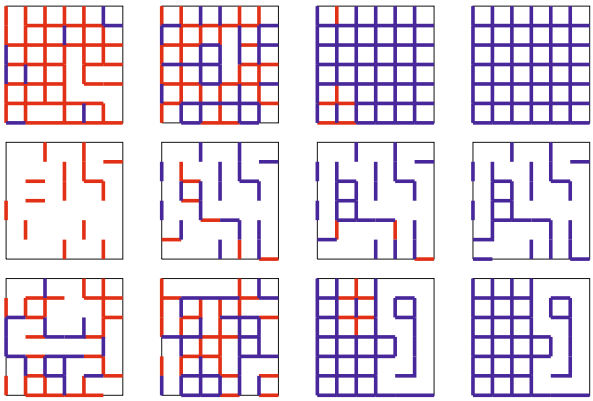

Social Processes, Physical Models of, Figure 11
Typical evolution of the Axelrod model on a square lattice with periodic boundary conditions. In each row, snapshots of the configuration in four times $t_1 < t_2 < t_3 < t_4$, from *left* to *right*, where the last configuration is the absorbing state. The active bonds are drawn in *red*, the bonds with full consensus in *blue*, the bonds with absolute disagreement are left *white*. The initial condition is drawn from $\mathbb{Z}_q$ with uniform probability. The parameters are, from *top* to *bottom*, $F = 3$ and $q = 2$; $F = 3$ and $q = 14$; $F = 2$ and $q = 2$. The frame in *light gray* around each configuration is there only for visual convenience

a lattice of size $N = L^d$, time t is discrete and proceeds in chunks of $1/N$, so from time t to $t + 1$ there are as many elementary updates as there are sites. One update step consists of:

1. Choose a site x at random; randomly choose one if its neighbors x^μ. Count the number A of features in which the agent at x agrees with the chosen neighbor.
2. If $A = 0$, do nothing. If $0 < A < F$, randomly choose a feature f among those in which the agent and the neighbor differ. Then set the value of the feature f equal, $\sigma_{t+1/N}(x, f) = \sigma_t(x^\mu, f)$.

One of the principal quantities of interest will be the density of active bonds, which is a generalization of the density of interfaces investigated in the voter model. A bond connecting agents at sites x and y is called active if the agents differ in at least one feature and also agree in at least one feature, i. e., $0 < A(\sigma(x), \sigma(y)) < F$. Their fraction relative to the total number of bonds in the lattice will be denoted $n_A(t)$.

When the number of active bonds drops to zero, after the stopping time τ_{st}, evolution freezes and the system reaches one if its absorbing states. It means that all neighbors are either identical or absolutely incompatible. Clusters of agents which share the same values of all features can be identified, while the borders of such clusters are marked by bonds with no shared value. Even for the simplest case of two features $F = 2$ and the set of allowed values for these features constrained to only two elements, there are infinitely many possible absorbing states characterized by various cluster configurations. (For a finite lattice this number is of course finite, but grows very fast with the system size.) This leads to an important note: Speaking of culture, there might be a good many stable configurations and it is impossible (and improper) to discriminate as to which is the best one.

A full characterization of the set of absorbing states would be too difficult. A rough measure is the number of agents in the largest cluster $s_{\max}$, compared to the total number of agents N. When this figure is close to the total system size, it means that the absorbing state is uniform, or very close to it. If the size of the largest cluster is small, or even remains finite when $N \to \infty$, the absorbing state is "multicultural" or fragmented into many isolated islands without mutual interaction, much like the hundreds of villages in Papua-New Guinea, each of them having its own particular language.

The evolution and shape of the eventual absorbing state depends critically on the initial condition. First, note that if, at the beginning, all the values of the features belonged to certain subset $V \subset \mathbb{Z}$ of integers, all features will remain within the set V forever. The most important thing is how large the set V is. But even when V is infinite, the values can be mostly concentrated on a finite subset of V, very rarely going beyond it. A simple but reliable measure of the effective size of the set V, taking into account how often each member is actually present, is the *inverse participation ratio* (IPR). Suppose that, in the initial state, the probability of some feature of a randomly picked agent having the ith value from the set V is p_i. Then, the IPR is

$$q^{-1} = \sum_{i=1}^{|V|} p_i^2 \,. \tag{28}$$

Such a definition is motivated by a simple consideration: if set V consists of the q lowest non-negative integers, i. e., $V = \mathbb{Z}_q \equiv \{0, 1, \ldots, q-1\}$, and all members have the same probability $p_i = 1/q$, the inverse participation ratio is exactly q^{-1}. The integer q is simply the number of elements which are participating, hence the name.

Phase Transition Let us look at how the Axelrod model behaves on a two-dimensional square lattice. Recall that the linear size of the lattice is L, so there are $N = L^2$ agents. The initial condition will consist in choosing the values of all features for all agents independently from the same uniform distribution on $\mathbb{Z}_q$. The typical evolution is shown in

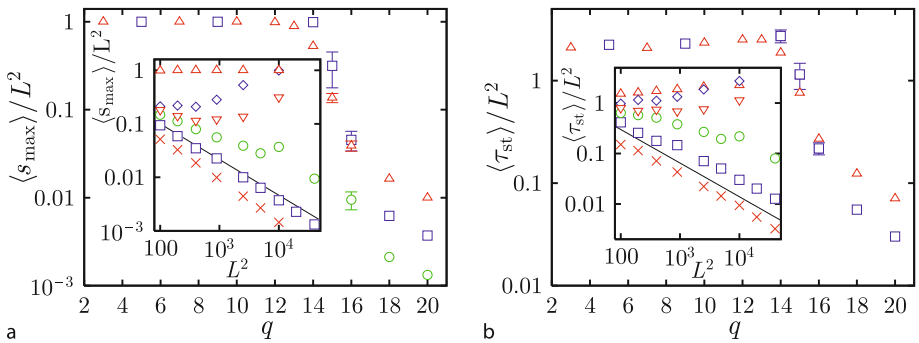

Social Processes, Physical Models of, Figure 12

Phase transition in Axelrod model on a square lattice of size $L \times L$, for $F = 3$, when changing the localization parameter q of the initial state. In a, relative size of the largest cluster, in b, average stopping time relative to the system size are shown. Different symbols correspond to sizes $L = 50$ (△), $L = 100$ (□), and $L = 200$ (○). In the *insets*, dependence of the corresponding quantities on the system size L^2 are shown. Symbols correspond to parameters $q = 5$ (△), 14 (◇), 15 (▽), 16 (○), 20 (□), and 30 (×). The *straight line* is the power $\sim L^{-4/3}$

Fig. 11. For larger F and q the absorbing state is very fragmented, but decreasing q it becomes totally uniform and a single cluster covers the entire lattice. On the other hand, for small F the absorbing state contains a few moderately-sized clusters.

These vague observations are put on a quantitative basis in Figs. 12 and 13. Let us look first at the behavior of the average size of the largest cluster, $\langle s_{max}\rangle$, when the parameter q is changed. The most important finding is that, at a certain value $q = q_c$, there is a phase transition separating the regime $q < q_c$, in which the maximum cluster makes up a finite fraction of the whole lattice, from the phase with $q > q_c$, where the fraction of sites within the largest cluster goes to zero for $N \to \infty$. The quantity $\langle s_{max}\rangle/N$ can serve as an order parameter.

Another quantity which illustrates phase transition is the average stopping time, i. e., the time to reach the absorbing state. Generically, it grows with the system size as $\langle \tau_{st}\rangle \sim N^{\eta}(q)$, but below the transition it is proportional to the system size, so $\eta(q) = 1$, while for $q > q_c$ it grows more slowly, with the exponent $\eta(q) \simeq 1/3$. Interestingly, comparison of the stopping time with the size of the largest cluster leads to the conclusion that, for $F = 3$, the two quantities are roughly proportional, $\langle \tau_{st}\rangle \propto \langle s_{max}\rangle$, both of them exhibiting a jump at the transition, suggesting that the clusters grow linearly in time until they reach the absorbing state. On the other hand, for $F = 2$, such proportionality is not observed. Instead, the average maximum cluster decreases continuously to zero when q_c is approached from below and at the same time the stopping time seems to diverge.

With discrete q, it is impossible to determine the position of the phase transition with high precision, but in practice it is possible to make some reasonable estimates, based on the dependence of $\langle s_{max}\rangle/N$ on N, as shown in the insets. For $F = 3$ the critical value of the parameter q lies somewhere close to $q_c \simeq 15$, while for $F = 2$ the estimated value is $q_c \simeq 5$.

A more important difference between cases $F = 2$ and $F = 3$ than the numerical value of the transition point q_c is the very nature of the phase transition. It was already shown that the stopping time normalized to the system size blows up at the transition only for $F = 2$, remaining finite for $F = 3$. With fair trustworthiness it is possible to say that for $F = 2$ the transition is continuous, i. e., second-order, while for $F = 3$ the transition is clearly a first-order one. Although it may be dangerous to rely only on numerical data for the order parameter, the conclusion on the nature of the phase transition was also supported by results for the distribution of cluster sizes close to the transition. In Fig. 14a it can be seen that the probability of finding in the absorbing state a cluster of size larger than or equal to s decays as a power, i. e., $P^{>}_{clust}(s) \sim s^{1-\alpha}$, with an obvious cutoff at large s due to the finiteness of the lattice. The value of the exponent α plays critical role. For $F = 2$ the estimate is $\alpha_{F=2} \simeq 1.6$ and for $F = 3$ it is $\alpha_{F=3} \simeq 2.65$. It has been shown that the exponent does not depend on the number of features as long as $F \geq 3$ [81]. How does the distribution of cluster sizes relate to the type of the phase transition? Denoting $N_{clust}(N)$ the total number of clusters and $P_{clust}(s, N) = P^{>}_{clust}(s) - P^{>}_{clust}(s+1)$ the fraction of clusters of size s on the lattice with N sites, the result is

$$N = N_{clust}(N) \sum_{s=1}^{N} s P_{clust}(s, N) \,. \tag{29}$$

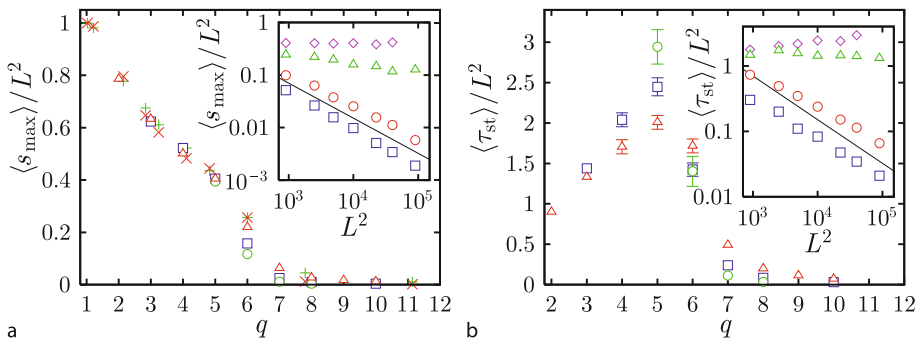

Social Processes, Physical Models of, Figure 13

Phase transition in the Axelrod model on a square lattice of size $L \times L$, for $F = 2$, when changing the localization parameter q of the initial state. In a, relative size of the largest cluster, in b, average stopping time relative to the system size. In both panels, results for the uniform initial distribution is plotted for sizes $L = 50$ (△), 100 (□), 200 (○). In the *insets*, dependence of the corresponding quantities on the system size, for $q = 5$ (◇), 6 (△), 7 (○), and 8 (□). The *straight lines* are power laws $\sim L^{-4/3}$ In a, there are also data extracted from [81], where the initial condition was assembled from Poisson-distributed integers; sizes were $L = 50$ (+) and $L = 150$ (×)

If $P_{\text{clust}}(s, N) \sim s^{-\alpha}$ and $\alpha < 2$, the sum on the right hand side diverges when $N \to \infty$. But the sum is simply the average cluster size, the quantity which plays the role of correlation length. The behavior resembles the percolation transition and the diverging correlation length is an unmistakable characteristic of a second-order phase transition.

On the other hand, if $P_{\text{clust}}(s, N) \sim s^{-\alpha}$ with $\alpha > 2$, the sum converges. To keep the equality in (29) for $N \to \infty$, either the number of clusters must be proportional to the system size, or, in addition to the power-law complement, there must be an additional term in the distribution of cluster sizes, accounting for the largest cluster of size $s_{\max} \simeq N$, i. e.,

$$P_{\text{clust}}(s, N) \simeq a s^{-\alpha} + b \ \delta_{s, s_{\max}} \ . \tag{30}$$

The former possibility occurs for $q > q_c$ and the latter for $q < q_c$. Indeed, the simulation data in Fig. 14a show the presence of a δ-function part in (30), with positive weight $b > 0$. Below the transition, the largest cluster spans essentially whole system, and at q_c its size drops discontinuously to a value negligible with respect to N. This is typical of first-order transitions.

These considerations are also compatible with the behavior of the stopping time as a function of q. For $F = 2$, data indicate that $\langle \tau_{\text{st}} \rangle / N$ diverges as $q \to q_c$, while for $F = 3$ it remains constant. Interpreting the stopping time as the correlation time of the dynamics, the model exhibits exactly the same behaviors commonly observed in the second- and first-order phase transitions, respectively.

To conclude, the phase transition in the Axelrod model is driven by the parameter q, measuring the localization of values in the initial condition, and belongs to the class of first-order transitions, if the number of features is at least 3, while for only 2 features the transition is continuous. Interestingly, the same dependence in terms of number of components is known in the Potts model. However, one should bear in mind that the transition in the Axelrod model has a purely dynamical origin. There is no equilibrium besides the absorbing states, so the analogy with phase transitions in equilibrium statistical physics must be taken cautiously and with certain reserve.

Although the most interesting question touches the properties of the absorbing states, it is instructive, also, to see how the system approaches them. In Fig. 14b the evolution of the average number of active bonds is shown. The averages are taken only over realizations which have not yet reached the absorbing state. Around the time $t \simeq 1$, i. e., after as many updates on the computer as there are lattice sites, the density of interfaces decreases drastically. When $q > q_c$ it brings the system quickly to the absorbing state, while in the opposite case, $q < q_c$, the activity rises again and n_A increases to a value close to $n_A \simeq 0.5$. Then, a very slow evolution follows. Most often, the evolution ends due to an occasional hit into an absorbing state, but this is, rather, a finite size effect, as demonstrated in the inset in Fig. 14b. When system size increases, slow evolution is increasingly prolonged. To see the behavior at these very long times is very difficult on a computer, but it seems that the density of active bonds decays very slowly to zero, resembling the logarithmic decay of the density of interfaces in the two-dimensional voter model.

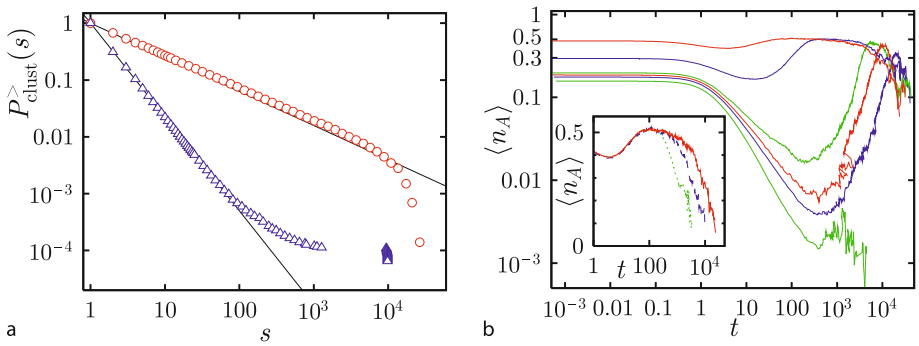

Social Processes, Physical Models of, Figure 14

Axelrod model on a square lattice. In a, distribution of cluster sizes close to the transition $q \simeq q_c$, averaged over 100 realizations. The parameters are $F = 2$, $q = 5$, $L = 200$ (◯), and $F = 3$, $q = 15$, $L = 100$ (△). The *straight lines* are power laws $\sim s^{-0.6}$ and $\sim s^{-1.65}$. In the data for $F = 3$, note the the isolated points around $s \simeq 10^4 = L^2$ indicating that the distribution is composed of a power-law part plus a δ-function located close to L^2. In b, time dependence of the number of active bonds, averaged over realizations of the process. The lattice size is $L = 100$, number of features $F = 3$ and different lines correspond to $q = 5, 9, 14, 15, 16$, and 18, from *top* to *bottom*. In the *inset*, the dynamics is shown for $F = 3$, $q = 5$, and three lattice sizes, $L = 30$ (*dotted line*), 50 (*dashed line*), and 100 (*solid line*)

Bounded Confidence with Continuous Opinions

There may be a different view on consensus formation, assuming that opinions on a certain issue can vary continuously, but a discussion on that subject and possible convergence of the opinions cannot take place unless the actual opinions are close enough. If the people differ too much, they may even avoid mutual contact completely or decide to use "different means", an often-used euphemism for killing the opponent. If, on the contrary, the difference in their opinions, measured by a continuous variable, does not exceed a certain threshold, then the difference may be further diminished by discussions and peaceful persuasion. This idea is called *bounded confidence* and a kind of it is also a key ingredient of the Axelrod model.

Linear Dynamics Modeling consensus formation using continuous variables has been around for quite a long time. Let us denote F_i as the variable describing the opinion of the individual i. The simplest way to implement the convergence of opinions, introduced by DeGroot [82], is to assume that in an upcoming period, opinions are linear combinations of the current opinions of all individuals

$$F_i(t+1) = \sum A_{ij} F_j(t) \,, \tag{31}$$

where A is a stochastic matrix, i. e., all its elements are non-negative, $A_{ij} \geq 0$, and the sum of all its columns is one, $\sum_i A_{ij} = 1$. This is called a *confidence matrix*. The diagonal elements are assumed to be strictly positive, $A_{ii} > 0$, which means that the individuals have at least some non-vanishing belief in their own opinions. The new values of F_i are weighted averages of previous opinions, the weights being stored in the matrix A. As such, the dynamics is very simple and is equivalent to those of a Markov chain with transition probabilities A_{ij}.

In DeGroot's model, the question of reaching a consensus or not is reduced to the study of connected components of the graph of direct interactions. When there are more of them around, each of them reaches its own consensus, independent of the others. This is quite simple. Introducing bounded confidence adds complications.

Hegselmann and Krause The new opinion of individual i takes into account others' opinions with weights A_{ij}. What if these weights are not constant, but depend on the current opinions themselves? Although some attempts in this direction have been made earlier [83], the principle of bounded confidence was first applied in the models developed by the groups of Deffuant et al. [84] and Krause and Hegselman [85,86]. Let us discuss the latter model first.

The Hegselmann–Krause (HK) model [86] considers opinion formation as a fully deterministic process, as in DeGroot's model investigated in the previous section. The new opinion of an individual is the average of current opinions of those (including the individual of concern) whose opinions lie within a fixed confidence bound $\epsilon > 0$ from the individual's own opinion.

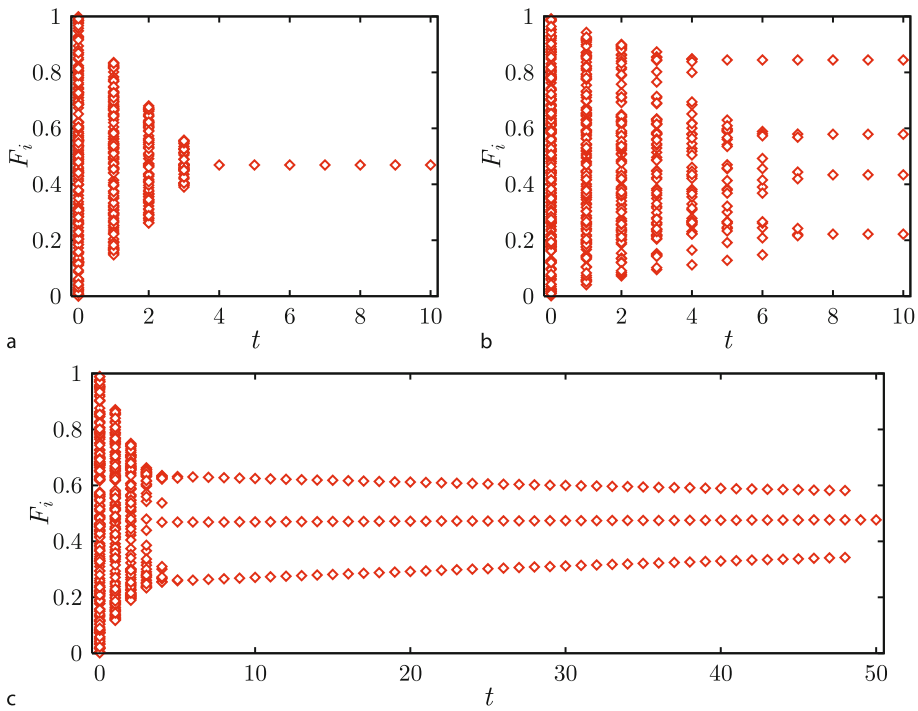

Social Processes, Physical Models of, Figure 15

Evolution of the Hegselmann–Krause model. Each point represents one or more agents with specified value F_i. Total number of agents is $N = 200$. In a, the confidence bound is $\epsilon = 0.3$, while in b it is $\epsilon = 0.1$. c shows how the approach to consensus is slowed down when the confidence bound is close to its critical value. Here $\epsilon = 0.24$

For those pairs which do not differ more than ε in their opinions, the weights A_{ij} are uniform, so

$$A_{ij} = \begin{cases} 0 & \text{for} \quad |F_i - F_j| > \epsilon \\ \frac{1}{N_{i\epsilon}} & \text{for} \quad |F_i - F_j| \le \epsilon \,. \end{cases} \tag{32}$$

Obviously, the normalization constant is the number of individuals within the confidence bound, $N_{i\epsilon} = |\{j\colon |F_i - F_j| \le \epsilon\}|$. Note that the dependence of the confidence matrix on current opinions is very strongly non-linear. From the formal point of view, the most important thing is that the zero-pattern of the confidence matrix can also change in time. It is possible to interpret it as changing the structure of the graph of directly communicating individuals. To see if consensus is going to be reached, it is important to observe whether the graph remains connected during evolution. Clearly, once it splits into disconnected parts, it will never join together again and consensus will not be achieved.

To get an impression of how the evolution of the HK model proceeds, look at Fig. 15. For a large enough confidence bound ε, the system approaches full consensus, while lower values of ε induce several stable communicating subgroups which do not interact with each other, and the system splits into several clusters with different opinions. The number of such clusters grows as $1/\epsilon$ when the confidence bound shrinks. Numerical simulations show that the critical value ϵ_c, at which the full consensus breaks down, approaches the value about $\epsilon_c \simeq 0.2$ when the number of individuals increases. Interestingly, when the confidence bound comes close to the critical value ϵ_c, the number of steps needed to reach consensus increases, suggesting that this is indeed a kind of a dynamical phase transition. This behavior is rather robust, as it persists in various modifications of the HK model, most notably if the individuals can communicate only through a regular lattice or a random network [87,88,89,90].

Deffuant et al. Apart from the initial condition, the model of Hegselmann and Krause is fully deterministic and opinions are updated in parallel. A similar model, but with random sequential update, was introduced by G. Deffuant et al. [84].

There are, again, N individuals with opinions F_i. In each update step, two of them, say, i and j, are chosen randomly. Then, we check to see whether their opinions

differ less than (or equally to) the confidence bound ε. In the positive case, their opinions are slightly shifted towards each other

$$\left.\begin{aligned} &F_i(t+1/N) \\ &\quad = (1-\mu)F_i(t) + \mu F_j(t) \\ &F_j(t+1/N) \\ &\quad = \mu F_i(t) + (1-\mu)F_j(t) \end{aligned}\right\} \text{for} \quad |F_i(t) - F_j(t)| \le \epsilon \tag{33}$$

where μ is a parameter fixing the rate of convergence.

For very large numbers of individuals, $N \to \infty$, the dynamics can be expressed in terms of the continuous distribution of opinions $P(f,t)$, which can be written formally as $\int_0^f P(f',t)\mathrm{d}f' = \lim_{N\to\infty} \frac{1}{N}\sum_j \theta(f - F_j(t))$. The following innocent-looking rate equation is obtained [91]

$$\frac{\partial}{\partial t}P(f,t) = \int\limits_{|f_1-f_2|\le\epsilon} P(f_1,t)P(f_2,t) \times \left[\delta((1-\mu)f_1 + \mu f_2 - f) - \delta(f_1 - f)\right]\mathrm{d}f_1\mathrm{d}f_2\,, \tag{34}$$

which exhibits fairly complex behavior. Starting from the uniform initial condition $P(f,0) = 1$ for $f \in [0,1]$, the configuration evolves into a stationary state composed of one or several δ-functions, each of them corresponding to one cluster of individuals sharing the same opinion.

To see this in a simple example, consider the case $\epsilon > 1$. All individuals interact with each other, so their opinions converge to a common limit $f = 1/2$. It is possible to see this when we use Eq. (34) to investigate the time evolution of the moments of the distribution. Obviously $\frac{\mathrm{d}}{\mathrm{d}t}\langle f\rangle = 0$, so the average opinion is independent of time, $\langle f\rangle = 1/2$. The dispersion from the mean $\langle f^2\rangle_\mathrm{c} = \langle f^2\rangle - \langle f\rangle^2$ obeys

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle f^2\rangle_\mathrm{c} = -2\mu(1-\mu)\langle f^2\rangle_\mathrm{c}\,, \tag{35}$$

which means that the dispersion decays exponentially to zero with rate $\mu(1-\mu)$. This confirms our intuition, noted earlier, that μ determines the speed of the evolution towards the stationary state.

The behavior in the complementary regime of very small ε can be guessed, assuming that $P(f,t)$ does not vary too wildly at scales comparable to ε or shorter. This also means that f is assumed to be farther than ε from the extremal values 0 and 1. In this case, $P(f_{1,2},t)$ is expanded in a Taylor series and the integral on the right hand side of (34) can be performed explicitly. Finally, the following partial differential equation is obtained

$$\frac{\partial}{\partial t}P(f,t) = -\frac{\epsilon^3}{3}\mu(1-\mu)\frac{\partial^2}{\partial x^2}P^2(f,t)\,. \tag{36}$$

There is a trivial homogeneous solution $P(f,t) = C$, independent of f and t, which is, however, unstable at all length scales. To see it, it is enough to linearize (36) close to the uniform solution, $P(f,t) = C + D(f,t)$, where $|D(f,t)| \ll C$, and express the result in terms of the Fourier transform $\widetilde{D}(k,t)$ of the small perturbation. The result is $\frac{\partial}{\partial t}\widetilde{D}(k,t) = \frac{2}{3}k^2\epsilon^3 C\mu(1-\mu)\widetilde{D}(k,t)$, indicating that perturbations increase as their wavelength decreases. On the other hand, if a stationary state is reached, the structures created cannot be closer to each other than ε. This feature is lost in the derivation of Eq. (36), but nevertheless the conclusion is that, for small ε, the stationary state will be composed of regularly spaced δ-functions with period proportional to ε.

To see explicitly how the stationary state is approached for general values of the confidence bound ε, the solution of Eq. (34) should be found numerically. Examples of such evolution are shown in Fig. 16. For large enough ε, all opinions converge to the common value 1/2, as in the case $\epsilon > 1$. However, low ε produces two or more distinct peaks. Detailed analysis [91,92] reveals that a series of bifurcations occurs in the system when the confidence bound is decreased. The consensus breaks into separated groups suddenly at a critical value ϵ_{c1}, which break themselves into more groups at $\epsilon_{c2} < \epsilon_{c1}$ and so forth. The numerically found value of the first critical point is $\epsilon_{c1} = 0.269\ldots$ But how do the peaks emerge from the originally homogeneous distribution? It can be seen in Fig. 16 that they are triggered by the original inhomogeneity at the edges of the interval $[0,1]$. A structure of gradually sharpening peaks propagates towards the middle of the interval, where non-linear waves from opposite endpoints meet and form the resulting pattern.

This observation can be translated into common language by saying that the ultimate fate of the opinions in the society is largely in hands of the extremists. Indeed, individuals close to the edges of the scale of possible opinions are, from the very beginning, drawn towards the middle of the interval $[0,1]$, i. e., close to the centrist position, while those who are farther from the edges than ε keep their opinions unchanged for some finite time. The crucial question is, how far from the extreme will the former extremists move before they settle? It may happen that they will drift so close to the midpoint that the centrists will attract them all; full consensus follows. But the extremists themselves can attract enough centrists and, instead of creating one peak at the center, two (or even more) peaks oc-

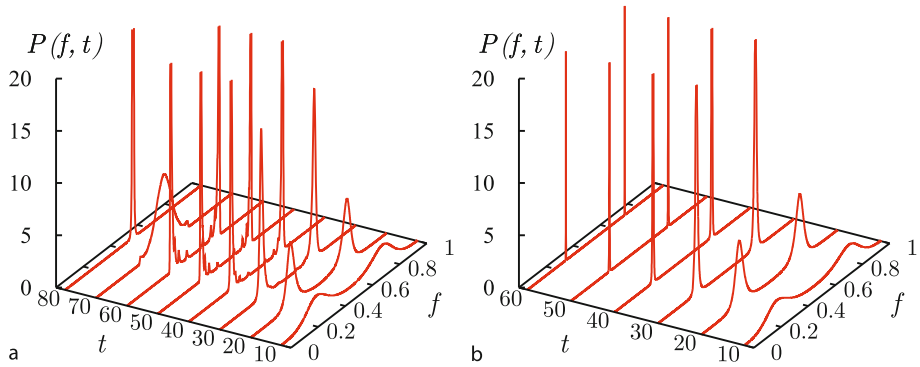

Social Processes, Physical Models of, Figure 16
Evolution of the Deffuant model with $\mu = 0.5$. In a, approach to consensus for confidence bound $\epsilon = 0.27$, slightly larger than the critical value ϵ_{c1}. In a, two peaks are formed in the distribution of opinions, for $\epsilon = 0.25$

cur. The resulting society lives with polarized (or diversified) opinions. The parameter which decides the outcome is the confidence bound ε. Higher confidence leads to an overwhelming consensus among people, while lower confidence causes the society split into several non-communicating communities.

Emergence of Social Classes

People are not equal. Not only do they differ in the color of their eyes, in their ability to play chess or to run a marathon race, but even individuals with very similar talents may find themselves very different in social status. All attempts to bring more justice into such evident disequilibrium has ended in desperate or even catastrophic failure. Perhaps the best one can do is to make the membranes separating social levels as permeable as possible, so that no one is a priori disqualified. The ubiquity of social stratification in animal as well as human collectives is certainly a phenomenon which calls for an explanation, and the fact that rabbits, dogs, apes and *Homo sapiens* exhibit similar behavior suggests some common mechanisms which may not be too complicated after all, although they produce highly complex outcomes.

Bonabeau Model

A newcomer in an animal group always has to undergo some fighting before its placement in the social ladder is commonly accepted. If, on the other hand, an individual leaves the group for some prolonged period and returns back, it has to fight again, as the previously established level has faded away. These two observations motivated Bonabeau et al. [93] to introduce a model of self-organized hierarchies [93,94,95,96,97,98].

To be clear, the hierarchies are not understood in the sense of trees with a king or a marshal on the top and lesser ranks below. Instead, it means an ordering, each individual bearing a single number called strength, indicating its position among others. When two agents meet, the stronger one has higher probability to be strengthened, while the weaker is most likely pushed down even more. The strengths of the agents who do not meet at this time relax towards zero by a fixed fraction. More formally, the configuration of the system at time t is described by the collection of strengths $F_i(t)$ of the agents $i = 1, 2, \ldots, N$. In each step, a pair of agents is chosen to fight, i and j, say, resulting in a change of the strengths by ± 1 for the winner and loser, respectively. Moreover, all strengths relax to the reference (zero) level deterministically. Thus

$$\begin{aligned} F_i\left(t+\frac{1}{N}\right) &= \left(1-\frac{\mu}{N}\right)F_i(t) + \Delta_{ij} \\ F_j\left(t+\frac{1}{N}\right) &= \left(1-\frac{\mu}{N}\right)F_j(t) - \Delta_{ij} \\ F_k\left(t+\frac{1}{N}\right) &= \left(1-\frac{\mu}{N}\right)F_k(t)\,, \quad k \neq i\,, \quad k \neq j\,, \end{aligned} \tag{37}$$

with

$$\text{Prob}\left\{\Delta_{ij} = \pm 1\right\} = \frac{1}{1 + \mathrm{e}^{\mp\eta(F_i(t)-F_j(t))}}\,. \tag{38}$$

The parameter η tunes the level of randomness in the dynamics, where $\eta \to \infty$ corresponds to purely deterministic outcomes, the stronger agent always beating the weaker, while $\eta = 0$ means that the strengths are increased and decreased by mere chance.

So far, the question of which pairs of agents interact and when, has not been dealt with. The most natural choice is to place the individuals on a network, or simply

on a square lattice, leaving some sites empty, and to allow the agents to diffuse along the edges of the network. When two agents happen to meet at one site, they fight. More than two agents on a site is not allowed; they behave like hard hemispheres, each site being able to accommodate one full sphere. Numerical simulations of such systems [93,94,95,96,97] show that the change in the density of agents, or, equivalently, the relaxation rate μ, induces a phase transition from a uniform state with all agents' strength close to 0 to a hierarchical state where the strengths are highly non-homogeneous. The order parameter is the dispersion of the number of fights won by the agents, i. e.,

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(w_i - \frac{1}{2}\right)^2} ,$$

where $w_i = n_i^+/(n_i^+ + n_i^-)$ and n_i^+, n_i^- are the number of encounters won and lost by the agent i, respectively. In the homogeneous phase $\sigma = 0$ while in the hierarchical phase it has a finite value.

Mean-Field Solution

Analytical study is possible in the mean-field approximation. Indeed, if diffusion is fast enough to ensure many encounters with various agents during the typical time given by the speed of relaxation of the strengths towards zero, the spatial structure of the lattice on which the diffusion takes place becomes irrelevant. In other words, at each time step, two agents are chosen at random and allowed to fight. In this case the dynamics is much simpler, and for a large system the following deterministic evolution equations for the strengths is obtained:

$$\frac{\mathrm{d}F_i}{\mathrm{d}t} = -\mu F_i + \frac{1}{N}\sum_{j=1}^{N}\frac{\sinh\eta(F_i - F_j)}{1+\cosh\eta(F_i - F_j)} . \tag{39}$$

The stochastic term decreases as $1/N$ for large N, so it is neglected here.

It can be easily seen that the average strength $\overline{F} = \frac{1}{N}\sum_i F_i$ relaxes exponentially to 0 according to $\frac{\mathrm{d}\overline{F}}{\mathrm{d}t} = -\mu\overline{F}$. So, it is sufficient to consider only stationary states with zero mean, $\overline{F} = 0$. The simplest of them, satisfying (39), is the trivial uniform state $F_i = 0$. This may be unstable, though, and its linear stability must be investigated.

Using the notation $\frac{\mathrm{d}F_i}{\mathrm{d}t} = R_i(F_1, \ldots, F_N)$ for the expression occurring in (39), the eigenvalues of the matrix $H_{ij} = \frac{\partial}{\partial F_j}R_i(0, \ldots, 0)$ should be inspected. Thus

$$H_{ij} = \left(-\mu + \frac{\eta}{2}\right)\delta_{ij} - \frac{\eta}{2N} . \tag{40}$$

One of the eigenvectors is uniform, $x_i = 1$, corresponding to eigenvalue $-\mu$, which is always negative. The remaining $N-1$ eigenvectors have the form $x_i = 1 - N\delta_{ik}$ for some k and they all belong to the same eigenvalue $\frac{\eta}{2} - \mu$. If the latter is positive, the uniform solution of the Eq. (39) is unstable. This happens for

$$\mu \le \mu_c \equiv \frac{\eta}{2} \tag{41}$$

and the question naturally occurs, what are the stationary configurations beyond the critical value μ_c? Not very far beyond the critical point, linear stability analysis can offer a useful hint, but it needs to be complemented by the influence of the lowest non-linear terms. Therefore, expanding the right-hand side of (39) up to the third order in the differences $F_i - F_j$, one gets for the stationary state

$$(2\mu - \eta)F_i = \frac{\eta^3}{12N}\sum_j (F_i - F_j)^3 . \tag{42}$$

The solution is assumed in the form

$$F_i = a\delta_{ik} - b \tag{43}$$

suggested by the eigenvectors of H_{ij} corresponding to unstable modes. Inserting the trial solution (43) into (42) leads to a set of equations for the parameters a and b which can be solved easily

$$a = \begin{cases} 0 & \text{for} \quad \mu > \frac{\eta}{2} \\ \pm\sqrt{\frac{12N(\eta-2\mu)}{(N-2)\eta^3}} & \text{for} \quad \mu < \frac{\eta}{2} \end{cases} \tag{44}$$

$$(2\mu - \eta)b = -\frac{1}{12N}\eta^3 a^3 . \tag{45}$$

Hence, the order parameter σ is deduced. Note that in the stationary state, the fraction w_i of fights won by the agent i should be balanced by a relaxation of the strength. Indeed, $2w_i - 1$ is the average increase of the strength of agent i in one step, which should be equal to μF_i. So, $\sigma = \mu\sqrt{\sum_i F_i^2/N}/2$ and inserting the result (45) yields

$$\sigma = \begin{cases} 0 & \text{for} \quad \mu > \frac{\eta}{2} \\ \frac{\mu}{\eta}\sqrt{\frac{3}{N}\left(1 - \frac{2\mu}{\eta}\right)} & \text{for} \quad \mu < \frac{\eta}{2} , \\ & \mu \to \frac{\eta}{2} \quad \text{and} \quad N \gg 1 . \end{cases} \tag{46}$$

How can this solution be interpreted? One may compare it to a situation in a society with one master and many servants, a single emperor while all the rest are servants, equal to each other in their subordination. This looks nice, but it follows from the formulae (45) that this "realistic" solution

always comes together with a mirror image of itself, a society of a single servant subject to many equal masters. This is an absurd situation, revealing something artificial in the Bonabeau model itself. Indeed, the dynamics is invariant with respect to inversion of all strengths, $F_i \to -F_i$, because no a priori advantage of being stronger was explicitly introduced. The only thing which was assumed was that the strong grow stronger and the weak grow weaker.

Now let us turn back to mathematical aspects of the model. It can be shown that a solution of type (43) exists for any value of μ and η, with only the complication that the equations for a and b become transcendental, thus analytically insoluble. This would be too simple, though. In fact, it is easy to find a stable solution in the limit $\eta \to \infty$, with μ kept finite, and see that it differs completely from (43). The expression within the sum in (39) becomes the sign function of the difference $F_i - F_j$ and the set of equations for the stationary state becomes

$$\mu F_i = \frac{1}{N} \operatorname{sign}(F_i - F_j) \tag{47}$$

and reordering the agents so that their strengths make an increasing sequence $F_1 < F_2 < \ldots < F_N$ one gets

$$F_i = \frac{1}{\mu N}(2i - N - 1)\,, \tag{48}$$

a society organized as a regular ladder of ranks. The order parameter corresponding to such a state,

$$\sigma = \frac{1}{3}\left(1 - \frac{1}{N^2}\right) \tag{49}$$

approaches a non-zero limit for $N \to \infty$, contrary to the configuration of (43) which is appropriate only in close proximity to the critical point, and whose order parameter decreases as $N^{-1/2}$, as can be seen from (44).

Despite the obvious criticism that the essential mirror symmetry $F_i \to -F_i$ is unrealistic and leads to unacceptable solutions, the Bonabeau model manifests quite well the basic idea of how the various social classes emerge. Within this framework, hierarchies are due to dynamical instabilities of the uniform state. The same general mechanism is also (at least partially) responsible for the stratification of human society. Any departure from the mean, be it positive or negative, is amplified. Any accidental misfortune sends one almost invariably even deeper. That is the whole mystery.

Future Directions

The prospects of physical modeling of social behavior are larger than what is presented here. The study of the emergence of cooperation within the framework of spatial and repeated prisoner's dilemma game is one of the classics. It may be, and has been, generalized in various ways. For example, one may wonder what happens if the agents could move through the lattice, or along the edges of the general graph representing the structure of the society. Indeed, people do not sit on the same lattice point forever. It turns out that, depending on the average velocity of the agents' movement, there is a phase transition. Less mobile agents form a phase with a macroscopic fraction of collaborators, while beyond a certain level of mobility the collaborators vanish.

Bounded confidence models pose some subtle and still unsolved problems. One of them concerns how the final state depends on the number of agents, or, stated differently, on the granularity of the distribution of opinions on the opinion axis. If the number of agents were arbitrarily large, the density could be arbitrarily small and it could happen that between any two agents there might be other agents which may mediate ultimate consensus. If that were true, the splitting of society into disparate groups would be the effect of limited size. The conclusion sounds somewhat paradoxical: small societies are more likely to be divided into non-communicating sub-cultures, while large societies turn to be more homogeneous. Another question naturally follows: how small must a social group be to allow disparate opinions to coexist? Implications for multicultural studies are evident.

The dynamics of social strata are relatively less explored and the state of the problem presented here must be considered as preliminary. The societies sketched in these models do not possess any internal structure, which would either stabilize or slow-down the emergence of stratification. Surely there is much space for future work here.

Some important areas were not covered at all, for example the whole family of physically based models of the birth and competition of languages [41,99]. Some social phenomena also closely touch economics, but these are covered by other articles in this Encyclopedia.

Acknowledgments

The original results in this article were obtained within the projects AVOZ10100520 and MSM0021620845.

Bibliography

Primary Literature

1. Comte A (1822) Plan des travaux scientifiques nécessaires pour réorganiser la société
2. Comte A (1839) Cours de philosophie positive, tome IV, 46e leçon. Bachelier Paris

3. Chigier NA, Stern EA (eds) (1975) Collective phenomena and the applicatios of physics to other fields of science. Brain Research Publications, Fayetteville
4. Callen E, Shapero D (1974) A theory of social imitation. Phys Today 27(7):23–28
5. Weidlich W (1991) Physics and social science – The approach of synergetics. Phys Rep 204:1–163
6. Anderson PW, Arrow KJ, Pines D (1988) The economy as an evolving complex system. Addison Wesley, Reading
7. Galam S (2004) Sociophysics: A personal testimony. Physica A 336:49–55
8. Galam S, Gefen Y, Shapir Y (1982) Sociophysics: A new approach of sociological collective behaviours. I. Mean-behaviour description of a strike. J Math Sociol 9:1–13
9. Galam S, Moscovici S (1991) Towards a theory of collective phenomena: Consensus and attitude changes in groups. Eur J Soc Psychol 21:49–74
10. von Neumann J, Morgenstern O (1944) Theory of games and economic behavior. Princeton University Press, Princeton
11. Vega-Redondo F (1996) Evolution, games, and economic behaviour. Oxford University Press, Oxford
12. Nash JF (1950) Equilibrium points in n-person games. Proc Natl Acad Sci USA 36:48–49
13. Nash JF (1950) The bargaining problem. Econometrica 18:155–162
14. Marsili M, Zhang YC (1997) Fluctuations around Nash equilibria in game theory. Physica A 245:181–188
15. Axelrod R (1980) Effective choice in the prisoner's dilemma. J Confl Resolut 24:3–25
16. Axelrod R (1980) More effective choice in the prisoner's dilemma. J Confl Resolut 24:379–403
17. Axelrod R, Hamilton WD (1981) The evolution of cooperation. Science 211:1390–1396
18. Lindgren K (1991) Evolutionary phenomena in simple dynamics. In: Langton CG, Taylor C, Farmer JD, Rasmussen S (eds) Artificial Life II. Addison-Wesley, Readwood City, pp 295–312
19. Lindgren K (1997) Evolutionary dynamics in game-theoretic models. In: Arthur WB, Durlauf SN, Lane DA (eds) The economy as an evolving complex system II. Perseus, Reading, pp 337–367
20. Nowak MA, May M (1992) Evolutionary games and spatial chaos. Nature 359:826–829
21. Schweitzer F, Behera L, Muhlenbein H (2002) Evolution of cooperation in a spatial prisoner's dilemma. Adv Compl Syst 5:269–299
22. Szabó G, Töke C (1998) Evolutionary prisoner's dilemma game on a square lattice. Phys Rev E 58:69–73
23. Chiappin JRN, de Oliveira MJ (1999) Emergence of cooperation among interacting individuals. Phys Rev E 59:6419–6421
24. Lim YF, Chen K, Jayaprakash C (2002) Scale-invariant behavior in a spatial game of prisoner's dilemma. Phys Rev E 65:026134
25. Szabó G, Vukov J, Szolnoki A (2005) Phase diagrams for an evolutionary prisoner's dilemma game on two-dimensional lattices. Phys Rev E 72:047107
26. Abramson G, Kuperman M (2001) Social games in a social network. Phys Rev E 63:030901(R)
27. Ebel H, Bornholdt S (2002) Evolutionary games and the emergence of complex networks. arXiv:cond-mat/0211666 (Preprint)
28. Ebel H, Bornholdt S (2002) Coevolutionary games on networks. Phys Rev E 66:056118
29. Zimmermann MG, Eguíluz VM (2005) Cooperation, social networks, and the emergence of leadership in a prisoner's dilemma with adaptive local interactions. Phys Rev E 72:056118
30. Vukov J, Szabó G, Szolnoki A (2006) Cooperation in noisy case: Prisoner's dilemma game on two types of regular random graphs. cond-mat/0603419
31. Clifford P, Sudbury A (1973) A model for spatial conflict. Biometrika 60:581–588
32. Holley RA, Liggett TM (1975) Ergodic theorems for weakly interacting infinite systems and the voter model. Ann Prob 3:643–663
33. Liggett TM (1985) Interacting particle systems. Springer, Berlin
34. Redner S (2001) A guide to first-passage processes. Cambridge University Press, Cambridge
35. Scheucher M, Spohn H (1988) A soluble kinetic model for spinodal decomposition. J Stat Phys 53:279–294
36. Krapivsky PL (1992) Kinetics of monomer-monomer surface catalytic reactions. Phys Rev A 45:1067–1072
37. Frachebourg L, Krapivsky PL (1996) Exact results for kinetics of catalytic reactions. Phys Rev E 53:R3009–R3012
38. Ben-Naim E, Frachebourg L, Krapivsky PL (1996) Coarsening and persistence in the voter model. Phys Rev E 53:3078–3087
39. Dornic I, Chaté H, Chave J, Hinrichsen H (2001) Critical coarsening without surface tension: The universality class of the voter model. Phys Rev Lett 87:045701
40. Al Hammal O, Chaté H, Dornic I, Muñoz MA (2005) Langevin description of critical phenomena with two symmetric absorbing states. Phys Rev Lett 94:230601
41. Castellano C, Fortunato S, Loreto V (2007) Statistical physics of social dynamics. arXiv:0710 3256
42. ben-Avraham D, Considine D, Meakin P, Redner S, Takayasu H (1990) Saturation transition in a monomer-monomer model of heterogeneous catalysis. J Phys A: Math Gen 23:4297–4312
43. Liggett TM (1999) Stochastic interacting systems: Contact, voter, and exclusion processes. Springer, Berlin
44. Derrida B, Hakim V, Pasquier V (1996) Exact exponent for the number of persistent spins in the zero-temperature dynamics of the one-dimensional Potts model. J Stat Phys 85:763–797
45. Gradshteyn IS, Ryzhik IM (1994) Table of integrals, series, and products, 5th edn. Academic Press, San Diego
46. Cox JT (1989) Coalescing random walks and voter model consensus times on the torus in $\mathbb{Z}^d$. Ann Prob 17:1333–1366
47. Galam S (1986) Majority rule, hierarchical structures, and democratic totalitarianism: A statistical approach. J Math Psychol 30:426–434
48. Galam S (1990) Social paradoxes of majority rule voting and renormalization group. J Stat Phys 61:943–951
49. Galam S (1999) Application of statistical physics to politics. Physica A 274:132–139
50. Galam S (2000) Real space renormalization group and totalitarian paradox of majority rule voting. Physica A 285:66–76
51. Galam S, Wonczak S (2000) Dictatorship from majority rule voting. Eur Phys J B 18:183–186
52. Schneier B (1996) Applied cryptography, 2nd edn. Wiley, New York

53. Krapivsky PL, Redner S (2003) Dynamics of majority rule in two-state interacting spin systems. Phys Rev Lett 90:238701
54. Slanina F, Lavička H (2003) Analytical results for the Sznajd model of opinion formation. Eur Phys J B 35:279–288
55. Plischke M, Bergersen B (1994) Equilibrium statistical physics. World Scientific, Singapore
56. Galam S (2004) Contrarian deterministic effects on opinion dynamics: The hung elections scenario. Physica A 333:453–460
57. Stauffer D, Sá Martins JS (2004) Simulation of Galam's contrarian opinions on percolative lattices. Physica A 334:558–565
58. Florian R, Galam S (2000) Optimizing conflicts in the formation of strategic alliances. Eur Phys J B 16:189–194
59. Galam S (2002) Minority opinion spreading in random geometry. Eur Phys J B 25:403–406
60. Galam S (2002) The September 11 attack: A percolation of individual passive support. Eur Phys J B 26:269–272
61. Galam S (2003) Modelling rumors: The no plane Pentagon french hoax case. Physica A 320:571–580
62. Galam S (2003) Global physics: From percolation to terrorism, guerilla warfare and clandestine activities. Physica A 330:139–149
63. Galam S, Mauger A (2003) On reducing terrorism power: A hint from physics. Physica A 323:695–704
64. Galam S, Vignes A (2005) Fashion, novelty and optimality: An application from Physics. Physica A 351:605–619
65. Galam S (2004) The dynamics of minority opinions in democratic debate. Physica A 336:56–62
66. Galam S (2004) Unifying local dynamics in two-state spin systems. cond-mat/0409484
67. Galam S (2005) Local dynamics vs. social mechanisms: A unifying frame. Europhys Lett 70:705–711
68. Gekle S, Peliti L, Galam S (2005) Opinion dynamics in a three-choice system. Eur Phys J B 45:569–575
69. Galam S, Chopard B, Masselot A, Droz M (1998) Competing species dynamics: Qualitative advantage versus geography. Eur Phys J B 4:529–531
70. Tessone CJ, Toral R, Amengual P, Wio HS, San Miguel M (2004) Neighborhood models of minority opinion spreading. Eur Phys J B 39:535–544
71. Galam S (2005) Heterogeneous beliefs, segregation, and extremism in the making of public opinions. Phys Rev E 71:046123
72. Sousa AO, Malarz K, Galam S (2005) Reshuffling spins with short range interactions: When sociophysics produces physical results. Int J Mod Phys C 16:1507–1517
73. Sznajd-Weron K, Sznajd J (2000) Opinion evolution in closed community. Int J Mod Phys C 11:1157–1165
74. Behera L, Schweitzer F (2003) On spatial consensus formation: Is the Sznajd model different from a voter model? cond-mat/0306576
75. Krupa S, Sznajd-Weron K (2005) Relaxation under outflow dynamics with random sequential updating. Int J Mod Phys C 16:177–1783
76. Stauffer D, de Oliveira PMC (2002) Simulation of never changed opinions in Sznajd consensus model using multi-spin coding. cond-mat/0208296
77. Stauffer D, de Oliveira PMC (2002) Persistence of opinion in the Sznajd consensus model: Computer simulation. Eur Phys J B 30:587–592
78. Stauffer D, Sousa AO, Moss de Oliveira S (2000) Generalization to square lattice of Sznajd sociophysics model. Int J Mod Phys C 11:1239–1245
79. Bernardes AT, Costa UMS, Araujo AD, Stauffer D (2001) Damage spreading, coarsening dynamics and distribution of political votes in Sznajd model on square lattice. Int J Mod Phys C 12:159–167
80. Axelrod R (1997) The dissemination of culture: A model with local convergence and global polarization. J Confl Resolut 41:203–226
81. Castellano C, Marsili M, Vespignani A (2000) Nonequilibrium phase transition in a model for social influence. Phys Rev Lett 85:3536–3539
82. DeGroot MH (1974) Reaching a consensus. J Am Stat Assoc 69:118–121
83. Chatterjee S, Seneta E (1977) Toward consensus: Some convergence theorems on repeated averaging. J Appl Prob 14:89–97
84. Deffuant G, Neau D, Amblard F, Weisbuch G (2000) Mixing beliefs among interacting agents. Adv Compl Syst 3:87–98
85. Krause U (2000) A discrete nonlinear and non-autonomous model of consensus formation. In: Elaydi S, Ladas G, Popenda J, Rakowski J (eds) Communications in difference equations. Gordon and Breach, Amsterdam, pp 227–236
86. Hegselmann R, Krause U (2002) Opinion dynamics and bounded confidence models, analysis and simulation. J Artif Soc Soc Simul 5: http://jasss.soc.surrey.ac.uk/5/3/2.html
87. Fortunato S (2004) Damage spreading and opinion dynamics on scale free networks. cond-mat/0405083
88. Fortunato S (2004) The Krause–Hegselmann consensus model with discrete opinions. Int J Mod Phys C 15:1021–1029
89. Fortunato S (2005) On the consensus threshold for the opinion dynamics of Krause–Hegselmann. Int J Mod Phys C 16:259–270
90. Pluchino A, Latora V, Rapisarda A (2005) Compromise and synchronization in opinion dynamics. physics/0510141
91. Ben-Naim E, Krapivsky PL, Redner S (2003) Bifurcations and patterns in compromise processes. Physica D 183:190
92. Weisbuch G, Deffuant G, Amblard F, Nadal JP (2001) Interacting agents and continuous opinions dynamics. cond-mat/0111494
93. Bonabeau E, Theraulaz G, Deneubourg JL (1995) Phase diagram of a model of self-organizing hierarchies. Physica A 217:373–392
94. Sousa AO, Stauffer D (2000) Reivestigation of self-organizing social hierarchies. Int J Mod Phys C 11:1063–1066
95. Stauffer D, Sá Martins JS (2003) Asymmetry in hierarchy model of Bonabeau et al. cond-mat/0308437
96. Schulze C, Stauffer D (2004) Phase diagram in Bonabeau social hierarchy model with individually different abilities. cond-mat/0405697
97. Malarz K, Stauffer D, Kułakowski K (2005) Bonabeau model on a fully connected graph. physics/0502118
98. Lacasa L, Luque B (2005) Bonabeau hierarchy models revisited. physics/0511105
99. Schulze C, Stauffer D, Wichmann S (2008) Birth, survival and death of languages by Monte Carlo simulation. Commun Comput Phys 3:271–294

Books and Reviews

Castellano C, Fortunato S, Loreto V (2007) Statistical physics of social dynamics. arXiv:0710 3256

Schweitzer F (ed) (2002) Modeling complexity in economic and social systems. World Scientific, Singapore

Weidlich W (1991) Physics and social science – The approach of synergetics. Phys Rep 204:1–163

Social Processes, Simulation Models of

Klaus G. Troitzsch
Universität Koblenz-Landau, Koblenz, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Social Processes and Social Systems
System Dynamics
Microanalytical Simulation
Multilevel Simulation
Multi-Agent Simulation
Tools
Future Directions
Bibliography

Glossary

Social process A social process in the current context can be defined as any series of events occurring where several human beings enter into interaction. The definition may be extended to processes in which social animals are involved.

Model A model in the current context is defined as the representation of some part of the real world, which is called a target system, where this representation is done in terms of some language, be it natural language or some formal graphic and/or computer-executable language. It represents real-world entities with nouns in natural language or objects in typical programming languages, properties of these entities with adjectives in natural language or instance variables in typical programming language, actions of these entities with verbs in natural language or methods and procedures in programming languages, and relations between such entities with relations between nouns or objects in the respective languages.

Simulation Simulation is the execution of a formal model of some process described in a programming language on a computer.

Agent An agent in the current context is usually defined as a piece of software that has some autonomy, i. e. operates without other parts of a larger computer program having direct control of its internal state and actions, that can communicate with other agents in some kind of language, i. e. without direct access to other agents' inner state, that can perceive its environment and can react on it, and that can also take the initiative, i. e. can act without a stimulus from outside and display goal-directed behavior.

System A system in the current context is a mathematical structure consisting of a set of components (composition) belonging to the system, another set of things of the same kind not belonging to the system (forming the environment of the system) and a set of relations defined on the set of things forming both composition and environment and containing at least one relation that changes the history of the involved things [6]. In other contexts, "system" may have other connotations.

Definition of the Subject

The simulation of social processes has been a topic of both the social sciences and computer science for more than five decades and goes back to the first applications of computers. This is, among others, due to the fact that the first developers of computers were not only mathematicians and people from electric engineering, but also economists and even political scientists. As one can read in biographies of Herbert A. Simon, a political scientist, [61] and John von Neumann, a mathematician and game theorist, [49] persons like these were fascinated by the idea that beyond mathematical analysis computer applications could be made fruitful for the understanding of social processes.

Whereas cross-sectional analysis in the social sciences had been supported by statistical methods even earlier, any longitudinal analysis of societies needed more advanced methods of analysis than statistics can provide.

Simulation of social processes comes in different forms, as this entry will deal with in much more detail. One usually distinguishes between pure macro-level simulations where one is only interested in what happens on one level, usually a nation, a society, or an organization and where only one entity – the whole system – is represented, from different kinds of multi-level simulations where at least a system and its components or elements are represented together with the interactions between levels and interactions among the components.

Simulation of social processes has several distinct aims. On one hand, and this is usually the primary, but most primitive aim, one is interested in the prediction of processes (this is often done before the process is properly understood and often leads to predictions that can only be invalidated by empirical research performed after the prediction). On the other hand, one is first interested to understand what is going on in a social process (i. e. to communicate to oneself what one believes are the ingredients of a social process), to make this communicable to others in order to be able to discuss one's understanding with colleagues, and to experiment with the model in order to find out whether the executable simulation model behaves like the target system it represents. Simulation thus plays a similar role as earlier kinds of models, e. g. a mathematical model in physics, which were qualified by Max Planck [48] as "a system of terms and sentences, the so-called physicist's world view, which he endows to the best of his knowledge in order that it – put in place of the real world – should send him – if ever possible – the same message as the real world would send". And a computer program written for simulation is just "a system of terms and sentences" that – even literally – is able to send messages to its user.

Introduction

This entry is organized as follows. The next section tries to give an overview of social processes that have so far been modelled for the purpose of simulation. As one will see, there is a very wide range of social processes that have been simulated with quite different simulation tools and approaches. And this section will, of course, give a more detailed definition what a social process is, and it will shortly deal with some system theories (unfortunately there are very different ones under the same name) and analyze how their adherents have used simulation to analyze what they described as systems.

Section "Microanalytical Simulation" and Sect. "Multi-level Simulation" will deal with two classical approaches to social simulation. The two classical approaches are those of system dynamics and of microanalytical simulation. Both of them have a history of more than 50 years (their learned societies celebrated their 50th anniversaries in August 2007) and have found a wide audience, with system dynamics still being an approach used from ecology to management sciences, modelling ecosystems as well as the planet as a whole, but also firms and markets, and with microanalytical simulation still being an approach used to predict consequences of tax and transfer reforms and the age structure of future generations. While system dynamics is a purely one-level, one-of-a-kind approach as it always represents exactly one undivided real-world system in terms of its properties, thus hiding its components, microanalytical simulation represents a large number of non-interacting men and women and/or households that react on the stimuli given by the experimenter who is only interested in the aggregate statistics which in turn never respond to the behavior of the micro-entities. Thus, in both of these classical approaches an important feature of social processes is in a way missing, as both of them do not actually represent interactions between human beings and, moreover, do not represent interactions between individual and group or individual and society.

Section "Multi-Agent Simulation" will deal with several contemporary approaches that have not yet made their way into the ministries or the boards of enterprises (with a few important exceptions that will be discussed). Common to all of these disparate approaches is the attempt at representing processes in which interactions among individuals and between levels play a major role. Interactions can occur in topological neighborhoods (as for instance in cellular automata) or in networks (as in many agent-based or multi-agent simulations), sometimes neither neighborhood nor network topology is represented, and models restrict themselves to the micro-macro interaction. Approaches of this type most often try to represent cognitive processes beside social processes, as at least in social systems of human beings much of the interaction is controlled by cognitive processes in the minds of participants. And it is still an open question whether a valid representation of human social system can do without taking account of cognitive processes, as it is an open question whether taking only cognitive processes into account leads to an explanation of social processes ("mind is not enough" [8], but "society is not enough" seems also to be true).

Social Processes and Social Systems

Talking about social processes makes it necessary first to define what a process is. Following Bunge [5] a process is a sequence of states a thing had or has or will have over time. It can be written as a time series whose elements denote the state of a thing at a certain point of time, the sequence of states being ordered with respect to time. Thus a process can be represented with a sequence of vectors, each vector representing the state of the thing at a certain time, and each element of each vector representing one of the properties of the thing that describe its state at this time.

When talking about social processes, what was called "thing" in the preceding paragraph (and thus following Bunge [5]) will always be a system composed of human beings (and perhaps containing several levels of subsystems) where the state of the system can and should be described both on the system level and on the component level (and perhaps on the levels defined by different subsystems). Although system dynamics describes the state of the system only on this upper level and neglects any subsystems and components, and although microanalytic simulation describes the state of the system only on the level of either households or individual (but usually not on both) and not on the aggregate level (this is done only for calculating results), both of these approaches describe social processes but in an incomplete manner.

But it will not be sufficient just to describe a social process (as if videotaping it and looking at the video frame after frame), instead it is necessary for understanding and explaining why the process developed the way it developed to construct a model as a means that can create the next frame from the former frames, thus (re-) constructing the laws the process obeys. For simple things such as those elementary physics deals with this law was often represented with a differential equation, one of the classical means of modelling processes. For social systems, this representation is only rarely if ever appropriate, as the calculus does not easily lend itself to describe a population and its members at the same time. This is why system dynamics sets the members of a population (members of an organization, citizens of a country) aside and concentrates on the relatively few properties of the system as a whole, which is also done in some of the classic system theories – for Rapoport (pp. 8–9 in [52]), for instance, "the simplest mathematical description of a system – a model of a spring balance" is the "equation $L = L_0 + kW$ where L is the length of the stretched spring, W the suspended weight, L_0 the 'natural' length of the spring (when $W = 0$), and k a constant that represents the 'stretchability' of the spring." This is exactly the same view on a system as the one taken by system dynamics, where only the variable and constants attributes of the system seem to be interesting, but not the interactions between the parts of a system.

At least from the time of Coleman, but certainly even earlier, e. g. in Durckheim [14], social processes have been superficially represented with the so-called "Coleman boat" or "Coleman bath-tub", a figure [9] that mirrors the notion that processes on the level of, say, society, can be observed on this level, but it is clear that it is not society that "acts", but whatever happens on the level of society is performed by the actions of individual human beings that are in a way under the control of the society and at the same time making up society. In Coleman's example, there is an observable process in which protestant religious doctrine seems to have led to capitalism. This "single proposition breaks into three: one with an independent variable characterizing the society and a dependent variable characterizing the individual; a second with both independent and dependent variable characterizing the individual; and a third with the independent variable characterizing the individual and the dependent variable characterizing the society. Thus the proposition begins and ends in the macro level, but in between it dips to the level of the individual" [8]. The original macro proposition is usually represented with the deck of a boat, while the first of the three new proposition forms its stern, the second is the keel and the third is the bow of the boat. Thus, protestant religious doctrine introduced in a population (wherever this doctrine might originate) influences the value system of individuals, and the changed value system modifies their economic behavior; finally this economic behavior of the individuals modifies the characteristics of the population as a whole (and this latter process is often called "emergence").

The first of these inter-level processes is meanwhile known as "immergence", a term coined by Castelfranchi [7] for the process by which – in Colemans's example – an individual becomes aware of a doctrine ("cognitive emergence") and decides to act according to this doctrine; this process was already described by Durckheim [14] when he talks about "sociological phenomena [that] penetrate into us by force or at the very least by bearing down more or less heavily upon us". The other process, by which individuals change the state of the system (population, organization, ...) qualitatively is often called "emergence".

System Dynamics

As superficially mentioned in earlier sections, system dynamics is a tool for modelling processes in social systems on a macro level. The approach originated in the 1950s, when Jay W. Forrester made his first steps in representing social processes in terms of differential equations (as others had done before him, e. g. Herbert A. Simon [59,60]) and found it necessary to include non-differentiable functions into his models. This was how the programming language DYNAMO was first invented, meanwhile superseded by more modern attempts at designing modelling languages with graphical user interfaces, a functional, i. e. declarative programming language, in which the programmer does not (have to) describe how the computer should perform its calculations, but only

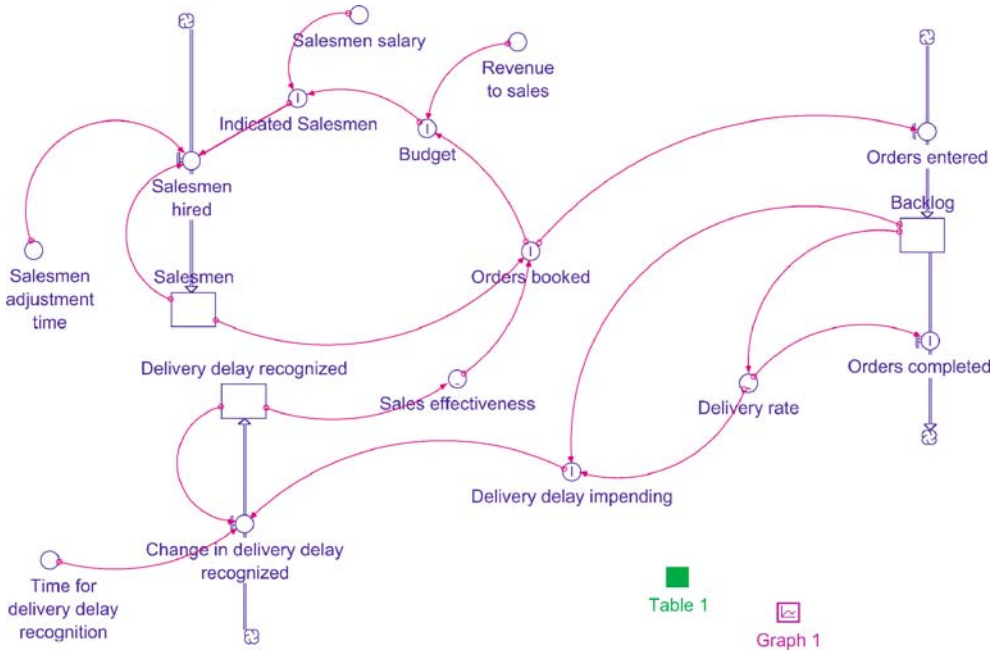

Social Processes, Simulation Models of, Figure 1
Graphical representation of Forrester's sales growth model, using the STELLA simulation tool

which invariants should hold over time. An invariant in this sense is an equation that holds true even when the states of a thing or system change (for instance, even if the size of population is changing over time, its growth rate could be constant for a long period, such that for all times the equation $\text{size}(t+1) = \text{size}(t) + \text{growthrate} * \text{size}(t)$ holds). System dynamics is of course much more sophisticated, but the equation in the preceding sentence is a typical representative of equations in system dynamics models. The first step in modelling is usually a graphical model showing the dependencies and feedbacks between the attributes of the modelled macro system. These attributes come in several different kinds, the two main kinds of variables being levels and rates. Levels are state variables, usually describing something like material stocks (inventory of a warehouse, size of a population etc.), whereas rates are state change variables, usually describing the material flow into a stock and from a stock per time unit. Beside these, auxiliary variables serve as abbreviations, and constants can be defined.

Among the target systems modelled with the help of system dynamics and the respective tools, there are agglomerations [18], industrial enterprises [16], ecosystems [2] as well as the world as a whole [19,42,43].

A relatively simple example of system dynamics modelling can be found in [17] where a firm with its "sales growth and saturation" is modelled and simulated. The only properties of this firm considered are the current salesperson staff, the backlog of unprocessed orders and the delivery delay recognized. These "level variables" are connected to each other via several "rate variables" – salespersons hired, change in delivery delay recognized, orders entered and orders completed – and several "auxiliary variables" – sales effectiveness, orders booked, budget, indicated salespersons, delivery rate, delivery delay impending – and a few constants – salespersons adjustment time, revenue to sales and time for delivery delay recognition. A STELLA representation of the model makes the idea behind the model sufficiently clear (see Fig. 1).

The simulation of this model shows the time series in several variables of this fictitious firm: The model starts with 10 salespersons and a high budget (stemming from a large number of orders), thus more salespersons are hired who bring more and more orders, more than can be processed such that the backlog quickly increases and the sales effectiveness decreases after some time. This is the signal for hiring less new salespersons, but the delivery delay cannot be decreased in the long run, as a still rising number of salespersons continue to bring more orders.

The graph in Fig. 2 shows some of the drawbacks of the approach: at any time (except at the start) the number of salespersons is a non-integer number, and at any time the number of newly hired persons is also a non-integer

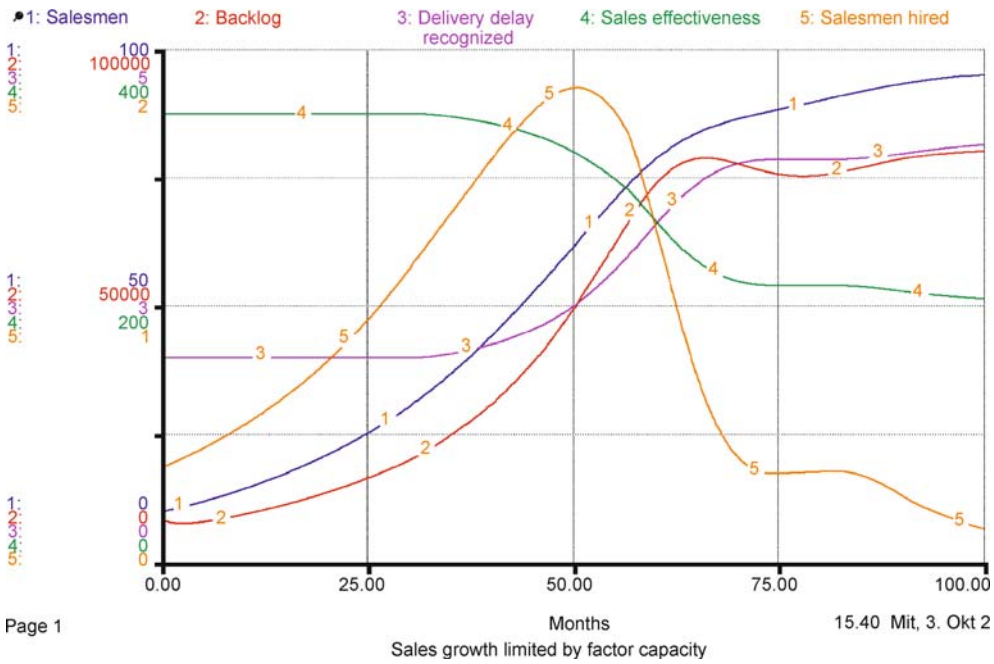

Social Processes, Simulation Models of, Figure 2
Graphical output of the sales growth STELLA model

number ranging from 0.13 to 1.85 whereas, of course, in a real firm, persons would be hired on a full-time or half-time base, but numbers such as 0.13 or 1.85 would be quite unlikely.

At the other end of the range of applications, the world models of the early 1970 can be placed, which became very famous as the Club of Rome funded and published them. Their objective was to predict the consequences of the world population growth for the use of resources and the pollution. Forrester's version [19] is replicated in a NetLogo 4.0 reimplementation shown in Fig. 3. This model has only five level variables – population, resources, pollution, capital investment (in industry) and capital investment in agriculture – which are linked together by nine rate variables and a large number of auxiliary variables and constants. The growth of the world population, for instance, is determined by two rate variables that contain the numbers of births and deaths, respectively, per time unit, and these in turn are determined by natural birth and death rates (biological constants, as it were) and several multipliers, auxiliary variables that transport the influences of causes such as pollution, food availability etc.; resources is another level variable, and this one is only determined by one rate variable, namely the decrease in resources per time unit, as Forrester (and also his successors in Meadows's research group) did not foresee that new resources might be discovered that were not known at the beginning of the simulation run (1900) or at the time when the simulation model was designed (1970).

The output of the model (see Fig. 4) shows what was predicted in the 1970s: a continued increase in the world population up to 8.6 billion in 2040, a decrease in available resources from 900 billion units (whatever these mean) in 1900 to 445 billion in 2040, and an increase in pollution from 200 million to 20 billion units (of another kind) in about 2050 – a result that was discussed worldwide in the 1970 (and perhaps this discussion was the reason why the predictions did not come out true 40 years later – although Meadows in [43] seems to have believed that his predictions were validated by the two decades after [42] appeared.

It is one of the big advantages of system dynamics models that they are easy to understand, easy to develop and to maintain and that they deliver plausible predictions that can easily be validated (or invalidated) against empirical data, though only after several decades. One of their important drawbacks is that the estimation of their input parameters is difficult if not impossible. Take as an example the estimation of a function between the pollution ratio and the pollution dependent death-rate multiplier (the pollution ratio is the actual pollution divided by the "pollution standard", which is defined as 3.6 billion pollution units – approximately the pollution in 1975). In Forrester's world model this function is programmed as

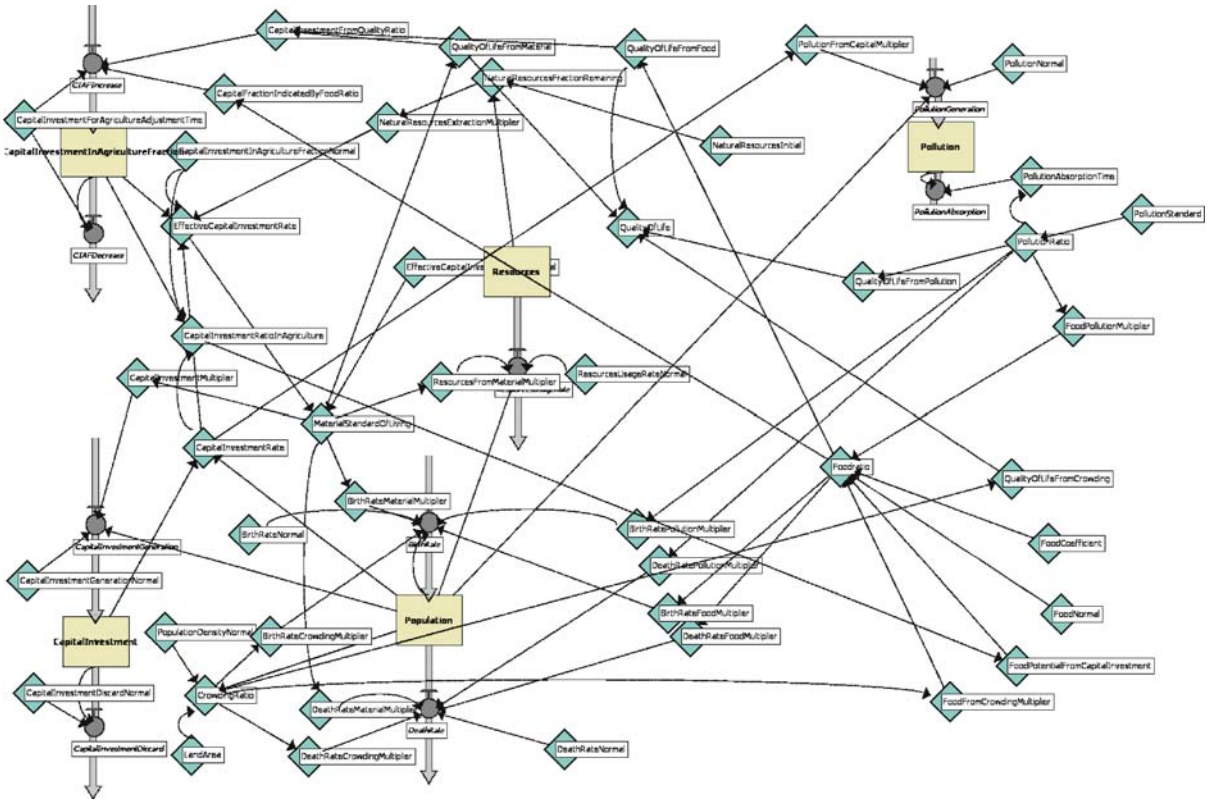

Social Processes, Simulation Models of, Figure 3
Forrester's WORLD2 model in a NetLogo 4.0 implementation

a table function (see Fig. 5), but in many cases it is quite questionable how the form of such a function was estimated.

In [42], the estimation of parameters for many functions between auxiliary variables is explained. The strategy of estimating the parameters of these functions is in most cases the same: it is usually the result of a nonlinear regression between – to take the preceding example – the pollution rate and the death rate pollution multiplier estimated from as many countries of the world as data could be found from, thus believing that synchronously found data from various places in the world were representative of data for the world as a whole collected at different points of time.

Microanalytical Simulation

Microanalytical simulation [22,47] has its origin in the necessity of predicting tax income and tax loads as well as the cost of social transfer. Every modern state needs information on how much taxes must be levied on its citizens in order to perform all the tasks of the state and how much social security contributions must be levied in order to pay retirement income, and it needs to know how many children will enter and leave the school system, how many students will enter and leave universities, how many people will enter the labor force at what time of their life and leave it at the time of their retirement, and, lastly, how many people will have near relatives who might be ready and willing to nurse them when they are in need. Whereas the age structure can be estimated even with difference equations, thus with system dynamics techniques, at least the last of the mentioned questions (which takes kinship networks into account) can only be answered on the micro level.

Microanalytical simulation always starts with a large representative sample of the target population. The sampling units are usually households (of which all members are interviewed), and the attributes sampled are at least age, gender, marital, educational and socio-economic status as well as the links to other members of the same household. This sample is then updated for every time step (usually one year) to answer the questions mentioned above for some time in the future.

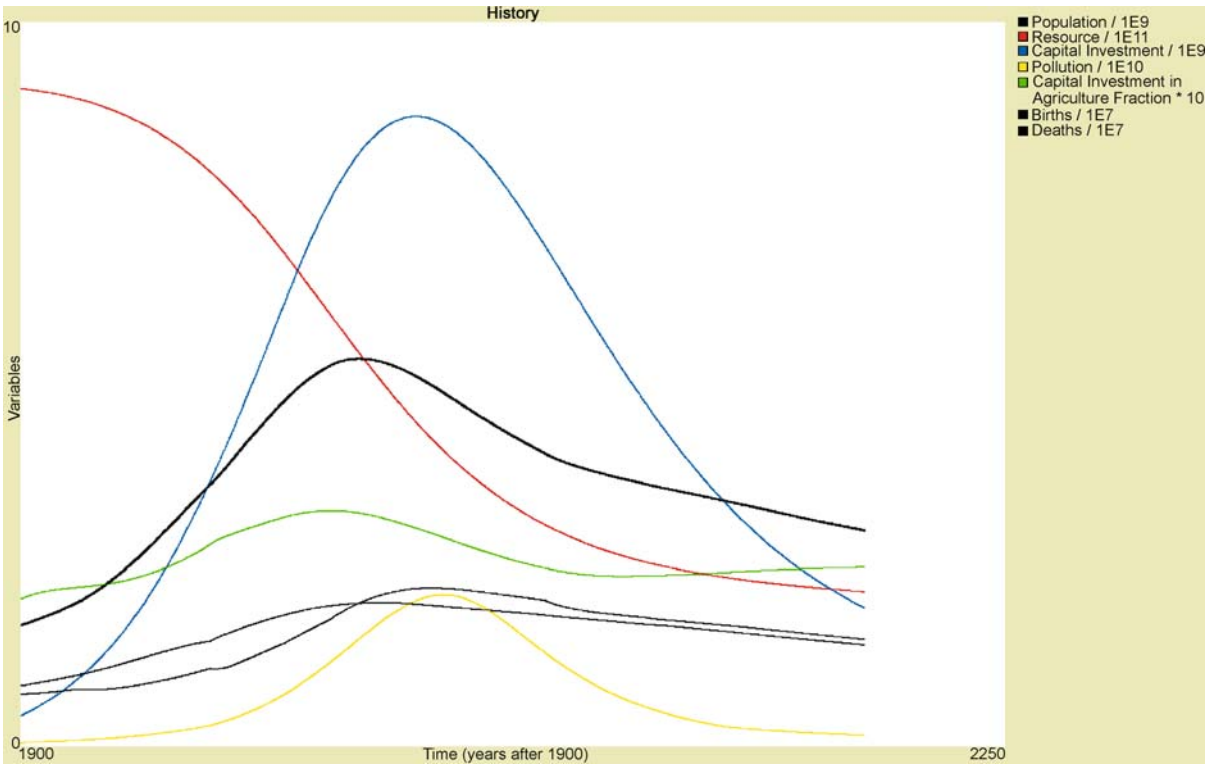

Social Processes, Simulation Models of, Figure 4
Some time series of Forrester's WORLD2 model, from a NetLogo 4.0 implementation

Microanalytical simulation has two easily distinguishable variants. They are usually labelled "static microsimulation" and "dynamic microsimulation". The former is static in so far as the attributes of the sample individuals are not changed over time, only the weights of the individual are updated in order to be representative for some macro variables where the information about the composition are taken from other (demographic or econometric) models. This static simulation is used only for short-time predictions, for instance for estimating tax loads due to revised tax regulation. Dynamic microsimulation changes all individual attributes according to known or estimated probabilities of state changes (for instance age-dependent birth, death, marriage and divorce rates or frequencies of transitions within the educational system and from the educational system into the labor force). For every time step, every micro entity has to be updated, and its link to other micro entities has to be revised, too. As compared to static microsimulation, this variant has the advantage that all state changes are endogenous and, more important, that state transition probabilities can be disaggregated to small sectors of the target and sample population (for instance, assuming different probabilities of giving birth to a first, second, third etc. child for women of different age groups, of different educational and socio-economic state), given that sufficient evidence for estimating such probabilities is available.

Early microanalytic simulation models were always programmed in general purpose programming languages from scratch, without using any particular tools. Their results were only partially published as research was mostly done for state agencies who were not always interested in having all the details of models published. Only recently, toolboxes such as UMDBS [56] and CoMicS [31] became available which lend themselves to be used in academic research and teaching, the more so as in many countries data sets of large representative samples and even panels have become available to the academic public for more than a decade now.

Multilevel Simulation

Multilevel and multi-agent simulations, different as they may be, always had their focus on interactions either among agents of the same kind or between agents of different levels or both. With levels, these approaches under-

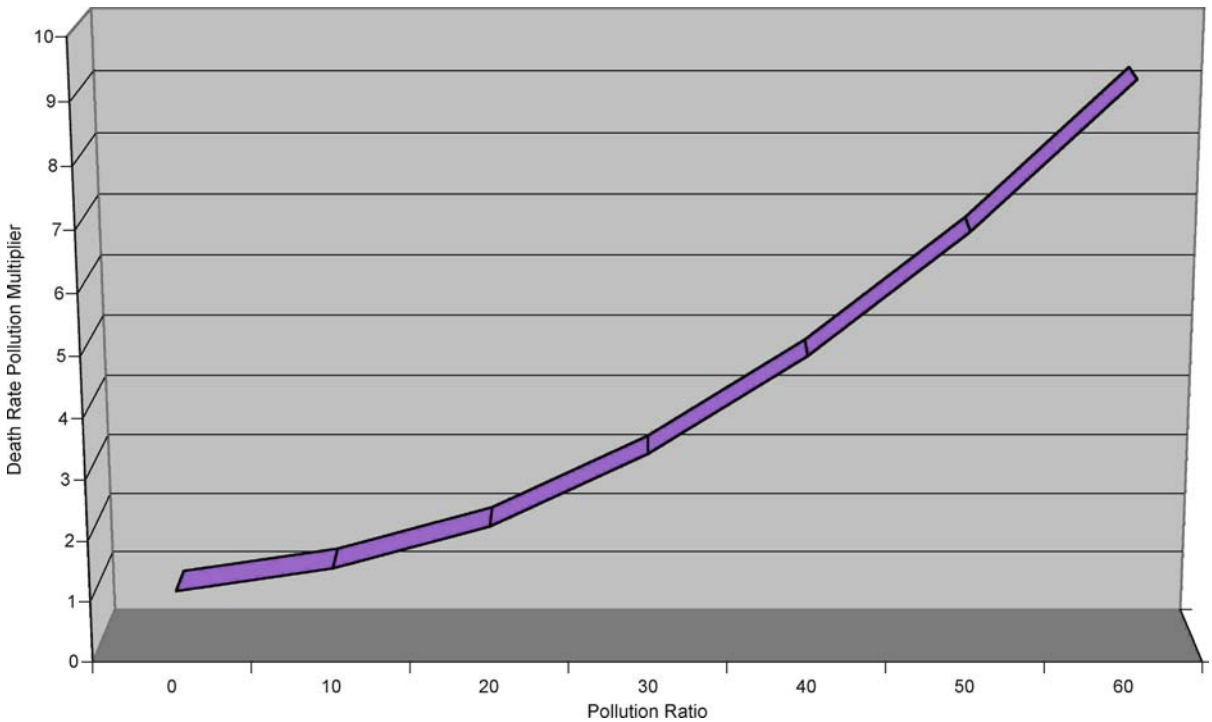

Social Processes, Simulation Models of, Figure 5
How system dynamics defines functions between its variables: the dependence of the death rate multiplier on the pollution ratio

stand different kinds of systems, where the elements of the systems of one kind are systems of the other kind [6]. To give an example: From the biologist's point of view, human beings are systems of one kind (composed of different kinds of tissues, or of cells, etc.), and on the other hand they can be seen as elements of systems called human groups, or populations, or societies etc. And what multilevel simulations are interested in is the interaction between, for instance, the group and an individual, between the macro and the micro level, thus focusing on the stern and bow of the "Coleman boat" [9]. Multi-agent simulation models are usually less focused on the inter-level interaction and more on the inter-agent interaction, but they, too, are usually interested in detecting properties emerging on the macro level.

More often than not, the label "multi1-agent simulation" is used for models in which the micro entities display only very simple behavior, have only very few states and do not communicate with other micro entities (in any reasonable sense of the word "communicate"), have no memory and make no decision. To distinguish these from multi-agent models proper (see the next section), they will here be labelled as "multilevel" models.

A first, quite well-known example may illustrate this. Thomas Schelling [57,58] designed a model in which (rather primitive) agents have to decide whether they move from one neighborhood into another, in which they can expect to be happier. Their happiness depends on the composition of their neighborhood, given that agents differ only with respect to one dichotomous property (say, their native language) – they are the happier, the more people in their neighborhood speak the same language, and there is a threshold where happiness turns into unhappiness. The world in Schelling's model is represented with a cellular automaton (see the chapter on cellular automata in this Encyclopaedia), and the neighborhood consists of the eight cells around an agent, which may be empty or inhabited by another agent. The central agent is happy if the language distribution in its neighborhood is such that a certain percentage of its neighbors speak the same language. In each time step, every agent scans its neighborhood and decides whether to move or stay, if it decides to move it tries to find another cell where the neighborhood is more attractive than in its original cell. One realization [67] of this model is shown in the two parts of Fig. 6.

It shows that the random distribution of red and green agents soon changes into a distribution with clearly distinguishable red and green clusters, part of which are even separated by uninhabited zones. The threshold between

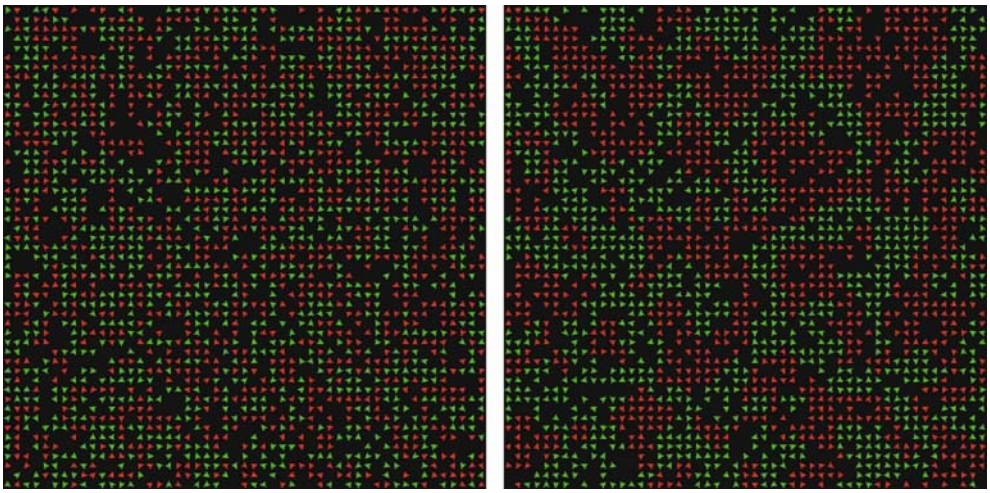

Social Processes, Simulation Models of, Figure 6
Schelling's segregation model (*left:* initialization, *right:* state of the macro level when all individuals are happy with their neighborhoods)

happiness and unhappiness in this example is 30 per cent, i. e. an agent moves only if more than 70 per cent in its neighborhood are different from itself. Although these agents are quite tolerant with respect to strangers, clusters occur; and these clusters are larger and better separated when the tolerance threshold is higher. One should note here that the Schelling model takes only local interactions between agents into account; there is no interaction between levels if only in so far as the structure of the macro level changes due to the individual interactions, but it does not feedback into the micro level as the agents are not capable of recognizing what happens on the macro level.

Another early approach to multilevel modelling of social processes is mainly interested in the interactions between the levels. It was originally developed by Hermann Haken [21], a German physicist who tried to apply the methods with which he had done research in laser physics to other disciplines, including sociology, and extended by his colleagues Wolfgang Weidlich and Günter Haag [66]. This approach, called synergetics by Haken, is currently better known as sociophysics or econophysics. Although the micro level entities are often also called agents by several authors, they are only particles as they only react on something like social forces or fields that are modelled in a manner very similar to the description of gravitational, electrostatic and electromagnetic forces and fields.

The original work by Weidlich and Haag [66] describes populations consisting of a large number of individuals where the individuals usually have only one discrete state variable which is either binary or can take only very few values, and the state change of the individuals depends on the macro state which is in turn determined by the individual state changes. The simplest example may make this clear: Imagine a population whose members have to decide between two options (yes and no) and whose decision depends on the perceived current majority in such a way that the probability of switching from no to yes is the higher, the larger the current yes majority is and the other way round. Then under certain assumptions of modelling the exact function between the current yes majority and the individual transition probabilities, the population as a whole shows some interesting behavior. In the example the two individual transition probabilities have the following form:

$$\mu_{\text{yes}\leftarrow\text{no}} = \nu \exp(\pi + \kappa x)$$

$$\mu_{\text{no}\leftarrow\text{yes}} = \nu \exp[-(\pi + \kappa x)] \,,$$

where ν is a flexibility parameter, π is a preference parameter, and κ is a coupling parameter, whereas x describes the current state of the population ($x = -1$ means "all no", $x = +1$ means "all yes"). The flexibility parameter only determines how likely any change between options is; the preference parameter determines whether changes from yes to no are generally more likely than those from no to yes (if it is negative) or the other way round (if it is positive); and the coupling parameter determines the strength of the influence the population exerts on the individual. It goes without saying that any individual change changes the state of the population. Figure 7 shows a typical outcome of a simulation of this very abstract social process. Each of the curves represents the history of the state of one out of twenty populations (all of them with a vanishing preference parameter and a coupling parameter of 1.4);

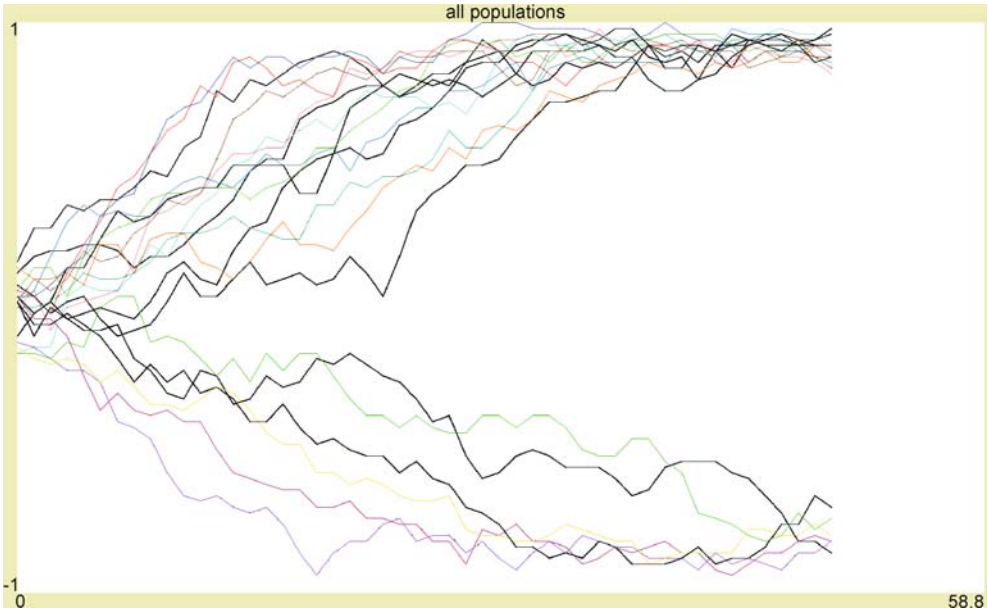

Social Processes, Simulation Models of, Figure 7
History of 20 populations whose members change their state with a probability that depends on the state of the population

some of the curves show that their populations soon develop into populations with a very high yes majority while others become populations with a high no probability, but none of them is ever entirely homogeneous. A more concrete social process that is likely to behave similarly is the distribution of two alternative goods that are not easily compatible, e. g. computer operating systems or videotape machines; here one could expect a high coupling as being with the majority makes the exchange of programs or videos easier. A low coupling constant results in an outcome where all populations are more or less evenly split in yes and no members – which in reality could be the case when the two alternative goods are easily compatible or substitutable, e. g. cigarette brands.

Extensions of this model were proposed by Lumsden and Wilson [41] who replaced the exponential function in the formula for the individual transition probabilities with other types of functions, to explain the co-evolution between "genes, mind, and culture". Other extensions, already given in [66], deal with several populations interacting with each other, the individual transition probabilities not only depending on the state of the members' own population, but also on the state of other populations, for instance modelling migration behavior of two populations between two regions.

Helbing [26] extended the models of his teachers, Wolfgang Weidlich and Günter Haag, in a way that he described as "quantitative sociodynamics". Much like his teachers he used closed solutions wherever possible to analyze the abstract social processes that he formalized, at the same time extending the focus not only to interactions between levels but also to interactions between individuals [24,25] which he also used to model the behavior of traffic participants, mainly pedestrians [24,27]. Here again, individual human beings are modelled as particles in a field and not endowed with the familiar human capabilities of decision making. In Helbing's pedestrian models (e. g. [27]) the pedestrians are just particles moving in a "social field", a term originally coined by Lewin [38], according to whose theory "behavioral changes are guided by so-called social fields or social forces, which has later on been put into mathematical terms" (p. 625 in [27]). These particle-pedestrians are accelerated at any time by a force which is determined by their desired directions and speeds and a repulsive force which keeps them apart from each other and from obstacles. Although the comparison between simulation results and video recordings of real pedestrian behavior shows nearly no qualitative difference, the particle-pedestrian are far from the definition of agents that is used by other authors who describe their simulations as multi-agent models.

Another kind of "social field" models was used to analyze attitude formation processes in an artificial society where the attitudes were not binary or categorical, but continuous. Models like these were inspired by Anthony Downs's early theory on party affiliation [13]. Downs as-

sumed that voters moved in a continuous attitude space (one-dimensional, from left to right) and were attracted or repelled by parties, and he believed that under certain circumstances a polarization of the electorate would emerge with high frequencies on the left and right wings and low frequencies in the middle, but he never formalized this theory. Several different formalizations were published in the sequel which tried to explain polarization without the interference of parties. One of these formalizations was inspired by empirical data which – for some German elections in the past – showed such a polarization in terms of a bi- or multimodal frequency distribution of voters over their attitude space which emerged during election campaigns and vanished shortly after the election [63]. The simulation runs started with a sample of a large numbers of model entities which obeyed an approximate bivariate normal distribution at the beginning (samples are never perfectly normal) entities then tried to move according to a function which was composed of the gradient of the frequency density distribution and some random effect. These movements changed the frequency density distribution, which in turn changed the gradient. In most simulation runs the frequency density functions became bimodal or multimodal after some time. Another representative of models inspired by Downs was presented by Guillaume Deffuant and is colleagues [12]. Here the simulated entities have a one-dimensional continuous attitude (much like in [63]), but additionally something like persuasive power, and as in [63], typically multimodal distributions emerge with modes both in the center but also at the ends of the attitude range, representing extremists in real electorates.

The approach in [12] is in some way similar to social impact theory developed by Bibb Latané [34] who also combined attitude and persuasive power, but let simulated entities interact on a grid, thus introducing topography and restricting interaction to near-neighbor interaction. The simulation models following his theory [35,36] show, similarly to the synergetic models, a survival of minorities, but also the clusters observed in Schelling's migration models (although the simulated entities in the social impact theory do not move in topological space).

Multi-Agent Simulation

An example that can easily be compared to the pedestrian models is the riot behavior model of Wander Jager and colleagues [29]. Here, agents belong to two different subpopulations (party 1, party 2) which can be imagined as the fans of two competing football clubs. Besides moving in a two-dimensional space, they can decide to fight or withdraw. In a version with police as a third subpopulation, the agents of this kind have additional behavioral options. Which action an agent takes does not depend on anything like a social field (at least not in the sense of Helbing, but perhaps in the sense of Lewin), but on their perceptions of the behavior of other agents in their neighborhood. These agents follow more or less the definition of agents that was given by Wooldridge and Jenkins [69] and which is currently the most often cited definition of agents:

An agent is a piece of software that

- has some autonomy, i. e. operates without other parts of a larger computer program having direct control of its internal state and actions,
- can communicate with other agents in some kind of language, i. e. without direct access to other agents' inner state,
- can perceive its environment and can react on it, and
- can also take the initiative, i. e. can act without a stimulus from outside and display goal-directed behavior.

This results in several requirements: In order to be able to take the initiative (to be pro-active), agents must have goals (descriptions of desired and possible states of their neighborhood or surroundings, including themselves) and must be able to compare their perceived state of their environment to their goals. According to the outcome of such a comparison, they must decide between several possible actions which they have to evaluate, and moreover they must be able to design a plan (i. e. a course of possible actions) if there is no action that leads them directly to their goal. Another requirement can be that agents communicate with other agents, for instance asking them for their help when only co-operation enables them to achieve their goals.

Among the early forerunners of multi-agent systems in the social sciences at least two can be named where processes of voter attitude changes are modelled and simulated. Although the poor computer languages of the late 1950s and early 1960s did not allow for agents in the sense of our days, any re-implementation would nowadays be a multi-agent system with several classes of agents (representing voters, candidates, media channels, as in [1] or in the Simulmatics project supporting John F. Kennedy's election campaign [11]) as they dealt with the communications among citizens, between citizens and candidates as well as between citizens and media channels and modelled their behavior and actions in a rule-based manner.

A classical example that fulfills most of the requirements raised in [69] is Epstein's and Axtell's attempt at building "Social Science from the Bottom Up" [15], an artificial world in which agents make their living on two dif-

ferent crops which they harvest, store, consume and barter according to their needs and the available resources. An extension of this artificial world can be found in [32]. In addition to all the features of the artificial society in [15], these agents can subordinate to others if the latter offer themselves as co-ordinators and promise their subordinates to inform them about the state of a wider environment than they can perceive individually, while the subordinates in turn provide their chieftains with their local information and pay a certain tribute from what they harvest. To be able to do all this, all agents have a long-term memory in which they store information about their world and which can guide them to places in their world that promise a rich harvest of the kind of crop they actually need. This extended model shows that a world with co-ordinators and subordinates is more sustainable than a world in which all agents try to harvest on their own; it also shows that lonely agents are best off when times are getting better, whereas subordinating pays best when times are getting worse and being a co-ordinator pays best when times are worst. Figure 8 is a part of two plots written by the simulation program: the green curve in the upper graph shows the development of food available over time, the blue one shows the number of surviving agents over time, while the three curves in the lower graph show the average "income" of the three types of agents (blue. the lonely agents, magenta: the chieftains, yellow: the subordinates) – "bad times" is when the overall available food is at its minimum, "good times" is when it is at a maximum. Among others, Fig. 8 shows that the macro behavior of the process is the same as the behavior of the well-known Lotka–Volterra model [39,65], but the classical coupled differential equations of this model or the stochastic process described in [66] would not have been able to display the additional information about the fate of the three kinds of agents.

Validating a model [64] such as the one described above is nearly impossible, as the model describes an abstract social process which is very unlikely to have ever occurred in reality. Nevertheless, this model makes a prediction that is quite similar to the prediction of another agent-based simulation [52,54]: "In a society of herdsmen and farmers in Western Africa, decisions which rest on

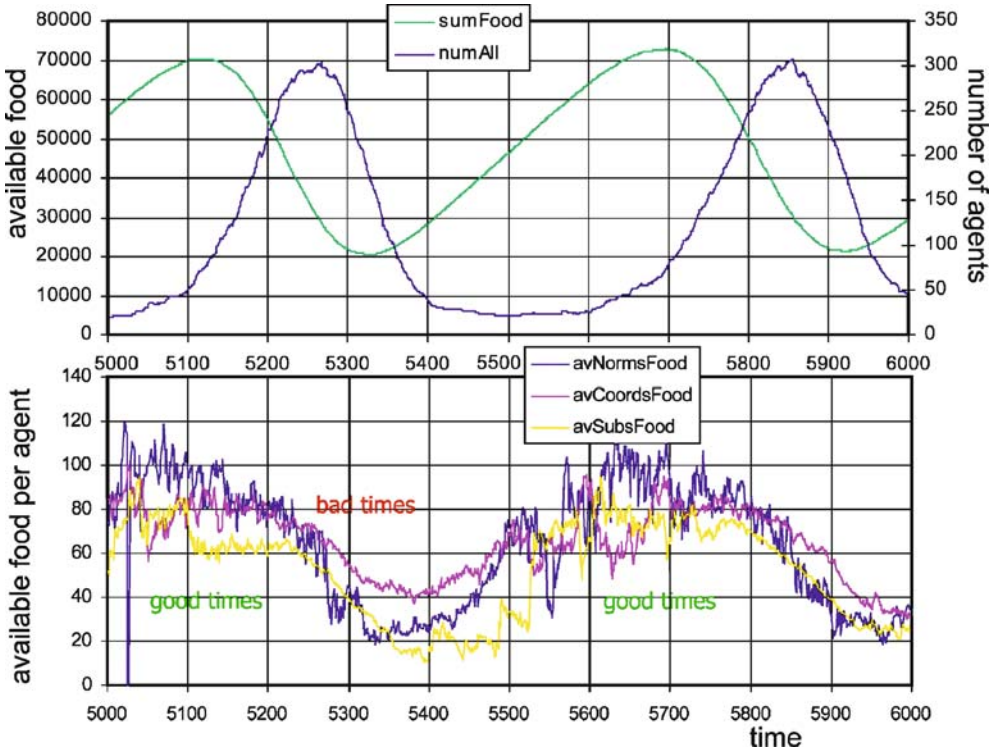

Social Processes, Simulation Models of, Figure 8
Development of an artificial society whose members can declare themselves co-ordinators and subordinates, respectively (graphs produced by the simulation model described in [32])

friendship networks ('friend-priority' decisions) proved to be much more effective then decisions which were made on pure cost deliberations ('cost priority' decisions)." See p. 189 in [52]. The papers quoted here are at the same time representatives of a style of modelling and simulating social processes which is known as "participative simulation" and usually used to discuss models with stakeholders who work with the simulation models and give hints at improving and amending models to adapt them better to their personal experience. This style of modelling and simulation is being used more and more often in consultancy to give stakeholders an opportunity to work with models on their own, instead having to believe what the consultant presents in the end.

Tools

Computer simulation always needs some tools to be run. In the early times of computer simulation, simulation programs were written in general purpose programming languages (like FORTRAN), but over the decades special simulation toolboxes were designed and implemented, first by researchers in fields such as organization science, production management (particularly for discrete event and queuing systems) or physics and engineering (particularly for numerical solutions of differential equations). Both of these kinds of tools were also used by social scientists, though only rarely. DYNAMO was the first simulation tool designed for and used by economists and social scientists [50], but only for system dynamics models. It was later on superseded by STELLA [53] with the same purpose, but a graphical user interface (see Figs. 1 and 2). There are several simulation tools for cellular automata (see the entry on cellular automata for more details). Multi-level simulation models are often written in general purpose programming languages (but see [44] describing a language – MIMOSE – explicitly designed for this kind of models). Multi-agent simulation proper has been developed with the help of toolboxes such as SWARM [30], RePAST [45,46], MASON [40], CORMAS [37] and NetLogo [68] (the latter was also used to produce the majority of graphics in this entry). All of these are freeware and capable of supporting most social simulation techniques (with the exception of microanalytic simulation, as this always needs a large database connected to the simulation tool). With the exception of NetLogo, which has a simulation language of its own, the actual description of agent behavior has to be done in object-oriented languages such as Smalltalk, Objective C or Java (for more detailed reviews of these toolboxes see [51,62]).

Future Directions

In a way, "agents cover all the world" [4] in that multi-agent systems can be used for all simulation purposes, as agents can always be programmed in a way that they behave as continuous or discrete models, can be activated according to event scheduling or synchronously or in a round-robin manner, can use rule bases as well as stochastic state transitions, and all these kinds of agents can even be nested into each other, thus supporting a wider range of applications than any of the classical simulation approaches. This leads to a third aspect of complexity (after the complexity of domains and the complexity of time): agent-based models can encompass several different approaches, both from a technical and implementation point of view, but also from the disciplines making use of simulation (for instance, disciplines such as neurophysiology, cognitive psychology, social psychology and sociology can combine their contributions into a deeply structured simulation model).

It goes too far to say that all this would not be possible in a non-agents world (as everything is programmable in Assembler), but the examples above will have made clear that models combining aspects from even neighboring disciplines as the ones enumerated in this entry are only understandable and communicable when they come in a modular form that is typical for agent-based models – and the same applies to ecological models where disciplines from physics and biochemistry to population biology and economics would play their co-operative roles.

Although agent communication languages (see the special issue of Autonomous Agents and Multi-Agent Systems, vol. 14 no. 2) such as KQML [33] have been developed for a long time (back to 1993), it is still an open question how agents in a simulation model could develop a communication means on their own and/or extend their communication tool to be able to refer to a changing environment (see the special issue of Autonomous Agents and Multi-Agent Systems, vol. 14 no. 1). Hutchins and Hazlehurst [28] made a first step into the field of the emergence of a lexicon, but their agents were only able to agree on names of things (patterns) they saw. The NEWTIES project [20], ambitious as it is, aims at creating an artificial society that develops its own culture and will also need to define agent capabilities that allow them to develop something like a language although it is still questionable whether the experimenters will be able to understand what their artificial agents talk about. The same objective is aimed at in the current EMIL project (Emergence in the Loop [3,10]) which attempts at creating an agent society in which norms emerge as agents observe each

other and draw conclusions about which behavioral feature is desirable and which is misdemeanor in the eyes of other agents – which, as in the case of language emergence, makes it necessary that agents can make abstractions and generalizations from what they observe in order that ambiguities are resolved.

This means that multi-agent models in the proper sense – where software agents model cognitive capacities of human beings – will be the aim of any future development in the simulation of social processes, although even these models will still contain features of the classical approaches, as even human beings display behavior that can be categorized as reflex acts and instinctive and which can be modelled with models borrowed from particle kinematics in which human beings are represented by stochastic automata. On the other hand it should be clear that only part of human behavior in very large groups can be explained this way, as an important part of this behavior consists in deliberate and conscious actions that are determined by cognition.

Bibliography

Primary Literature

1. Abelson RP, Bernstein A (1963) A Computer Simulation of Community Referendum Controversies. Pub Opin Q 27(1):93–122
2. Anderson JM (1973) The Eutrophication of Lakes. In: Meadows D, Meadows D (eds) Toward Global Equilibrium. Wright Allen, Cambridge, pp 171–140
3. Andrighetto G, Campenni M, Conte R, Paolucci M (2007) On the Immergence of Norms: a Normative Agent Architecture. In: Proceedings of AAAI Symposium, Social and Organizational Aspects of Intelligence, Washington
4. Brassel KH, Möhring M, Schumacher E, Troitzsch KG (1997) Can Agents Cover All the World? In: Conte R, Hegselmann R, Terna P (eds) Simulating Social Phenomena. Springer, Berlin, pp 55–72
5. Bunge M (1977) Ontology I: The Furniture of the World. Treatise of Basic Philisophy, vol 3. Reidel, Dordrecht
6. Bunge M (1979) Ontology II: A World of Systems. Treatise of Basic Philosophy, vol 4. Reidel, Dordrecht
7. Castelfranchi C (1998) Simulating with Cognitive Agents: The Importance of Cognitive Emergence. In: Sichman JS, Conte R, Gilbert N (eds) Multi-Agent Systems and Agent-Based Simulation. Springer, Berlin, pp 26–44
8. Castelfranchi C, Conte R (1994) Mind is not Enough. The Precognitive Bases of Social Interaction. In: Doran JE, Gilbert N (eds) Simulating Societies: The Computer Simulation of Social Phenomena. UCL Press, London, pp 267–286
9. Coleman JS (1990) Foundations of Social Theory. The Belknap Press of Harvard University Press, Cambridge
10. Conte R, Andrighetto G, Campenni M, Paolucci M (2007) Emergent and Immergent Effects in Complex Social Systems. In: Proceedings of AAAI Symposium, Social and Organizational Aspects of Intelligence, Washington
11. de Sola Pool I, Abelson RP (1962) The Simulmatics Project. Pub Opin Q 25:167–183
12. Deffuant G, Amblard F, Weisbuch G, Faure T (2002) How can Extremism Prevail? A study based on the relative agreement interaction model. J Artif Soc Soc Simul 5(4) http://www.soc.surrey.ac.uk/JASSS/5/4/1.html
13. Downs A (1957) An Economic Theory of Democracy. Harper, New York
14. Durckheim E (1982) The rules of the sociological method (trans: Halls WD). The Free Press, New York
15. Epstein JM, Axtell R (1996) Growing Artificial Societies – Social Science from the Bottom Up. MIT Press, Cambridge
16. Forrester JW (1961) Industrial Dynamics. MIT Press, Cambridge
17. Forrester JW (1968) Principles of Systems. MIT Press, Cambridge
18. Forrester JW (1969) Urban Dynamics. MIT Press, Cambridge
19. Forrester JW (1971) World Dynamics. MIT Press, Cambridge
20. Gilbert N, den Besten M, Botovics A, Craenen BGW, Divina F, Eiben AE, Griffioen R, Hévízi G, Lõrincz A, Paechter B, Schuster S, Schut M,C Tzolov C, Vogt P, Yang Lu (2006) Emerging Artificial Societies Through Learning. J Artif Soc Soc Simul 9(2) http://www.soc.surrey.ac.uk/JASSS/9/2/9.html
21. Haken H (1978) Synergetics. An Introduction. In: Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry and Biology, 2nd edn. Springer, Berlin
22. Harding A (ed) (1996) Microsimulation and Public Policy. Contributions to Economic Analysis, vol 232. Elsevier, Amsterdam
23. Helbing D (1991) A Mathematical Model for the Behavior of Pedestrians. Behav Sci 36:298–310
24. Helbing D (1991/1992) A Mathematical Model for Behavioral Changes by Pair Interactions and Its Relation to Game Theory. Angew Soz 17:179–194
25. Helbing D (1992) A Mathematical Model for Attitude Formation by Pair Interactions. Behav Sci 37:190–214
26. Helbing D (1994) Quantitative Sociodynamics. Stochastic Methods and Models of Social Interaction Processes. Kluwer, Dordrecht
27. Helbing D, Johansson A (2007) Quantitative Agent-Based Modeling of Human Interactions in Space and Time. In: Amblard F (ed) Proceedings of the 4th Conference of the European Social Simulation Association (ESSA'07), Toulouse, 10–14 Sept, 2007, pp 623–637
28. Hutchins E, Hazlehurst B (1995) How to invent a lexicon: the development of shared symbols in interaction. In: Gilbert N, Conte R (eds) Artificial Societies. The Computer Simulation of Social Life. UCL Press, London, pp 157–189
29. Jager W, Popping R, Sande HVD (2001) Clustering and fighting in two-party crowds: Simulating the approach-avoidance conflict. J Artif Soc Soc Simul 4(3) http://www.soc.surrey.ac.uk/JASSS/4/3/7.html
30. Johnson P, Lancaster A (2001) Swarm User Guide. http://www.swarm.org/swarmdocs-2.1.1/userbook/userbook.html
31. Klein C, Fuchs D, Berger P, Hassenpflug P (2006) MicS. Agentenbasierte Mikrosimulation, Project Thesis (Studienarbeit), Universität Koblenz-Landau, Koblenz
32. König A, Möhring M, Troitzsch KG (2003) Agents, Hierarchies and Sustainability. In: Billari F, Fürnkranz-Prskawetz A (eds) Agent-Based Computational Demography. Springer, Berlin, pp 197–210
33. Labrou Y, Finin T (1997) A Proposal for a new KQML Specification, Computer Science and Electrical Engineering Depart-

ment (CSEE). University of Maryland Baltimore County (UMBC), Baltimore
34. Latané B (1981) The psychology of social impact. Amer Psychol 36:343–356
35. Latané B (1996) Dynamic Social Impact. Robust Predictions from Simple Theory. In: Hegselmann R, Mueller U, Troitzsch KG (eds) Modelling and Simulation in the Social Sciences from a Philosophy of Science Point of View. Kluwer, Dordrecht, pp 287–310
36. Latané B, Nowak A (1994) Attitudes as Catastrophes: From Dimensions to Categories with Increasing Involvement. In: Vallacher R, Nowak A (eds) Dynamical Systems in Social Psychology. Academic Press, San Diego, pp 219–250
37. LePage C, Bousquet F, Bakam I, Bah A, Baron C (2000) CORMAS: A multiagent simulation toolkit to model natural and and social dynamics at multiple scales. In: Proceedings of the Workshop "The ecology of scales", Wageningen
38. Lewin K (1951) Field Theory in Social Science. Harper, New York
39. Lotka AJ (1925) Elements of Physical Biology. Williams & Wilkins, Baltimore
40. Luke S, Cioffi-Revilla C, Panait L, Sullivan K (2004) MASON: A New Multi-Agent Simulation Toolkit. In: Proceedings of the 2004 SwarmFest Workshop, Santa Fe
41. Lumsden CJ, Wilson EO (1981) Genes, Mind, and Culture. The Coevolutionary Process. Harvard University Press, Cambridge
42. Meadows D et al (1974) Dynamics of Growth in a Finite World. MIT Press, Cambridge
43. Meadows D et al (1992) Beyond the Limits. MIT Press, Cambridge
44. Möhring M (1996) Social Science Multilevel Simulation with MIMOSE. In: Troitzsch KG, Mueller U, Gilbert N, Doran JE (eds) Social Science Microsimulation. Springer, Berlin, pp 123–137
45. North MJ, Howe TR, Collier NT, Vos RJ (2005) The Repast Simphony Runtime System. In: Agent 2005 Conference on Generative Social Processes, Models and Mechanisms. Argonne National Laboratory, Argonne
46. North MJ, Collier NT, Vos JR (2006) Experiences Creating Three Implementations of the Repast Agent Modeling Toolkit. ACM Trans Model Comp Simul 16(1):1–25
47. Orcutt GH, Merz J, Quinke H (1986) Microanalytic simulation models to support social and financial policy. Information Research and Resource Reports, vol 7. North-Holland, Amsterdam
48. Planck M (1930) Positivismus und reale Außenwelt. In: Planck M (1949) Vorträge und Erinnerungen. Hirzel, Stuttgart, pp 228–245
49. Poundstone W (1992) Prisoners' Dilemma. John von Neumann, Game Theory, and the Puzzle of the Bomb. Oxford UP, Oxford
50. Pugh III AL (1976) DYNAMO User's Manual. MIT Press, Cambridge
51. Railsback SF, Lytinen SL, Jackson SK (2006) Agent-based Simulation Platforms: Review and Development Recommendations. Simulation 82(9):609–623
52. Rapoport A (1986) General Systems Theory. Essential Concepts & Applications. Abacus, Tunbridge Wells and Cambridge
53. Richmond B (2001) An introduction to systems thinking: STELLA software. ISEE Systems, Lebanon
54. Rouchier J, Bousqet F, Barreteau O, LePage C, Bonnefoy JL (2000) Multi-Agent Modelling and Renewable Resources Issues: The Relevance of Shared Representations for Interacting Agents. In: Moss S, Davidsson P (eds) Multi-Agent-Based Simulation. Springer, Berlin, pp 181–197
55. Rouchier J, Bousquet F, Requier-Desjardins M, Antona M (2001) A multi-agent model for describing transhumance in North Cameroon: comparison of different rationality to develop a routine. J Econo Dyn Control 25:527–559
56. Sauerbier T (2002) UMDBS – A New Tool for Dynamic Microsimulation. J Artif Soc Soc Simul 5:2–5 http://jasss.soc.surrey.ac.uk/5/2/5.html
57. Schelling TC (1971) Dynamic Models of Segragation. J Math Soc 1:143–186
58. Schelling TC (1978) Micromotives and Macrobehavior. Norton, New York
59. Simon HA (1957) A Formal Theory of Interactions in a Social Group. Amer Soc Rev 17:202–211
60. Simon HA (1957) Models of Man, Social and Rational. Math Essays Ration Hum Behav Soc Setting. Wiley, New York
61. Simon HA (1996) Models of My Life. MIT Press, Cambridge
62. Tobias R, Hofmann C (2004) Evaluation of free Java-libraries for social-scientific agent based simulation. J Artif Soc Soc Simul 7 http://jasss.soc.surrey.ac.uk/7/1/6.html
63. Troitzsch KG (1998) Multilevel Process Modeling in the Social Sciences: Mathematical Analysis and Computer Simulation. In: Liebrand WBG et al (eds) Computer Modeling of Social Processes. Sage, London, pp 20–36
64. Troitzsch KG (2004) Validating Simulation Models. In: Horton G (ed) Networked Simulations and Simulated Networks. 18th European Simulation Multiconference. SCS, Erlangen, San Diego, pp 265–270
65. Volterra V (1926) Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. Allt Accad Naz Lincei 6(2):31–113
66. Weidlich W, Haag G (1983) Concepts and Models of a Quantitative Sociology. The Dynamics of Interacting Populations. Springer, Berlin
67. Wilensky U (1998) NetLogo Segregation model. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston. http://ccl.northwestern.edu/netlogo/models/Segregation
68. Wilensky U (2007) NetLogo 4.0 User Manual. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston. http://ccl.northwestern.edu/netlogo/docs/NetLogo%20User%20Manual.pdf
69. Wooldridge M, Jennings NR (1995) Intelligent agents: theory and practice. Knowl Eng Rev, 10:115–152

Books and Reviews

Ahrweiler P, Gilbert N (1998) Computer Simulations in Science and Technology Studies. Springer, Berlin

Amblard F (2007) Interdisciplinary Approaches to the Simulation of Social Phenomena. In: Proceedings of the 4th Conference of the European Social Simulation Association. IRIT, Toulouse

Casti JL (1997) Would-Be Worlds. How Simulation Is Changing the Frontier of Science. Wiley, New York

Coelho H, Espinasse B (2004) 5th Workshop on Agent-Based Simulation. SCS, Erlangen

Conte R, Hegselmann R, Terna P (eds) (1997) Simulating Social Phenomena. Springer, Berlin

Doran JE, Gilbert N (1994) Simulating Societies: The Computer Simulation of Social Phenomena. UCL Press, London

Edmunds B, Hernández C, Troitzsch KG (2008) Social Simulation. Technologies, Advances, and New Discoveries. Information Science Reference. Hershey, New York
Gilbert N, Conte R (eds) (1995) Artificial Societies. The Computer Simulation of Social Life. UCL Press, London
Gilbert N, Troitzsch KG (2005) Simulation for the Social Scientist. 2nd edn. Open University Press, Maidenhead
Hegselmann R, Mueller U, Troitzsch KG (eds) (1996) Modelling and Simulation in the Social Sciences from a Philosophy of Science Point of View. Kluwer, Dordrecht
Liebrand WBG et al (1998) Computer Modeling of Social Processes. Sage, London
Moss S, Davidsson P (2000) Multi-Agent-Based Simulation. Lecture Notes in Artificial Intelligence, vol 1979. Springer, Berlin
Müller JP, Seidel MM (2003) 4th Workshop on Agent-Based Simulation. SCS, Erlangen
Sichman JS, Conte R, Gilbert N (1998) Multi-Agent Systems and Agent-Based Simulation. Springer, Berlin
Simon HA (1977) Models of Discovery and Other Topics in the Methods of Science. Reidel, Dordrecht
Suleiman R, Troitzsch KG, Gilbert N, Mueller U (1998) Social Science Microsimulation. Tools for Modeling, Parameter Optimization and Sensitivity Analysis. Springer, Heidelberg
Sun R (2006) Cognition and Multi-Agent Interaction. From Cognitive Modeling to Social Simulation. Cambridge UP, New York
Troitzsch KG (2005) Representing Social Reality. In: Pre-Proceedings of the Third Conference of the European Social Simulation Association. Fölbach, Koblenz
Troitzsch KG, Mueller U, Gilbert N, Doran JE (eds) (1996) Social Science Microsimulation. Springer, Berlin
Vallacher R, Nowak A (eds) (1994) Dynamical Systems in Social Psychology. Academic Press, San Diego

Social Psychology, Applications of Complexity to

Robin R. Vallacher
Florida Atlantic University, Boca Raton, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Intrinsic Dynamics
Attractors in Psychological Systems
Dynamical Minimalism
The Dynamics of Social Influence
Dynamics of Interpersonal Coordination
Future Directions
Bibliography

Glossary

Dynamical system A dynamical system is a set of interconnected elements that change due to their mutual influences. A change in each element depends on the nature of the influences from other elements. Due to these mutual influences, the system as a whole evolves in time. Because the state of every element is the effect of the states of all the relevant elements in the previous moment, and because the state of every element is one of the causes determining the state of the other elements in the next moment, a dynamical system is characterized by bi-directional causality.

Intrinsic dynamics In a dynamical system, the current state of the system is directly caused by the preceding state of the system. More precisely, the state of each element at a given moment is caused by the states of other elements in the previous moment and by its own previous state. This chain of cause and effect iterated over time gives rise to internally-generated or "intrinsic" dynamics.

Self-organization In a dynamical system consisting of inter-connected elements, the state of each element adjusts to the current state of other elements to which it is connected. Because of this mutual influence, higher order units develop that represent the organization of the basic elements. No higher-order agent is required for such order to emerge. Hence, the process is referred to as self-organization.

Emergence Emergence occurs when the individual elements of a dynamical system achieve organization by means of their mutual influence. The development of the higher level state is said to be emergent because this state was not inherent in the properties of the lower-level elements and because the higher-level state was not imposed on the system from forces outside the system. The higher-order properties that result from the mutual adjustment among lower-level elements provide coordination for the lower-level elements. Emergence thus provides for substantial growth in the complexity of a system's processes and properties. Because of emergence, very complex systems can often be described by very simple models.

Attractor An attractor represents a subset of a system's phase space to which the system evolves over time and which resists forces that would perturb this temporal trajectory. This subset can consist of a region of nearby states or it can represent two (or more) regions among which the system oscillates over time in a periodic, quasi-periodic, or chaotic manner. In a system governed by attractor dynamics, a relatively wide range of starting points (initial states) will eventually converge on a much smaller set of states or on a pattern of change between states defining the attractor. External influences may push the system out of the attrac-

tor, but over time the system will return to this equilibrium.

Cellular automata Cellular automata are programmable dynamical systems. In this approach, a finite set of elements is specified, each of which can adopt a finite number of discrete states. The elements are arranged in a specific spatial configuration that usually takes the form of a two-dimensional lattice or grid. The location of each element on this grid specifies the element's neighborhood. The elements evolve in discrete units of time, such that the state of an element at $t + 1$ depends on the states of the neighboring elements at time t. The dynamics of cellular automata depend on the nature of the updating rule and on the format of the grid dictating the neighborhood structure.

Definition of the Subject

The seminal insights that launched the field of social psychology over a century ago have stood the test of time. Such early scholars as James [41], Cooley [21], Mead [65], Lewin [59], and Asch [4] emphasized the multiplicity of interacting forces operating in individual minds and in social groups and the potential for sustained patterns of change resulting from such complexity. The inherent complexity and dynamism of human experience, however, proved problematic for mainstream social psychology during much of the 20th century. The canonical paradigm in this period reduced dynamics to a one-step process involving a purported cause, operationalized as an independent variable at an arbitrary time 1, and its effect, operationalized as a dependent variable that was assessed at an arbitrary time 2. The complexity of experience, with multiple variables interacting over time to generate a stream of thought or action, was difficult to investigate with available statistical techniques (e. g., correlation, regression, analysis of variance). Social psychology thus fostered an image of human experience that emphasized simplicity and stability at the expense of complexity and dynamism.

Contemporary social psychology shows signs of returning to the deep intuitions concerning human experience articulated by the field's founding fathers. The renewed appreciation for complexity and dynamism was made possible by developments in the understanding of nonlinear dynamical systems in the 1970s and 1980s, and the application of these developments to social processes within the last two decades. Computer simulations capture the complexity of social processes and document the emergence of higher-order properties from the interaction of basic elements in a mental or social system. Innovative means of collecting and analyzing time-series data provide rigorous insight into the intrinsic dynamics of mental, affective, behavioral, and interpersonal processes. And formal models enable researchers to identify the parameters that are critical for understanding the dynamism and complexity of social psychological phenomena. In short, the contemporary emphasis on complexity and dynamics provides a new paradigm for the field, one that holds promise for advancing the social psychology as a precise science while preserving the basic insights that launched the field over a century ago.

Introduction

Despite its ubiquity, dynamism is a poor candidate for theory construction. Theories are expressed in terms of invariant properties representing stable "signals" that are obscured by the "noise" associated with personal, interpersonal, and societal processes. The stream of consciousness may be an accurate depiction of subjective experience, for example, but its turbulent nature seems inconsistent with fundamental properties that transcend particular individuals and their moment-to-moment mental states. To identify regularities and invariant properties, social psychologists commonly ignore the ever-changing undercurrent of mind and action, focusing instead on those elements of thought and behavior that admit to stability and structure. Accordingly, they typically emphasize higher order units of mental and behavioral phenomena (e. g., traits, schemata, global evaluations, norms) that presumably lend stability and coherence to subjective experience. The turbulent flow of lower level elements (thoughts, feelings, movements) is effectively rendered irrelevant to "true" understanding and meaningful prediction.

The gulf between the reality of experiential turbulence and the focus on stability in theory construction is unnecessary. Dynamism and structure represent complementary aspects of experience that together provide the basis for cognitive, emotional, and behavioral accommodation. The basic notion is that the flow of lower-level elements gives rise to higher order structures, which in turn constrain the dynamics of lower-level elements. This reciprocal causal relationship between lower-level and higher-level units of experience is central to *dynamical social psychology* [74,106] a recently developed paradigm that represents an adaptation of the concepts, principles, and tools developed within dynamical systems and complexity science.

It is not feasible to describe every version of the dynamical perspective, nor every application of this general approach. This may have been possible in the

1990s [26,31,35,45,102] when the dynamical perspective was a promissory note rather than an established paradigm. Since that time, there has been a proliferation of many innovative research strategies—some providing formal models implemented in computer simulations, others offering empirical means for investigating the dynamics of personal, interpersonal, and societal processes—that have been used to investigate a wide variety of phenomena. The aim herein is to highlight the crucial elements of the dynamical perspective that find expression in otherwise distinct theories, research strategies, and topical agendas.

The first two sections (Sects. "Intrinsic Dynamics" and "Attractors in Psychological Systems") describe basic concepts from the study of nonlinear dynamical systems that are directly relevant to social psychology. Section "Dynamical Minimalism" describes dynamical minimalism [72], an approach that provides a workable entrée into the nature and expression of dynamic processes at different levels of social reality. This approach is illustrated in the next two sections with respect to two research agendas—one emphasizing the emergence of group-level properties from the self-organization of individual agents (Sect. "The Dynamics of Social Influence") the other exploring the tendency of individuals to coordinate their behavioral and mental dynamics in service of forming dyads and social groups (Sect. "Dynamics of Interpersonal Coordination"). The final section ("Future Directions") reflects on the trajectory of dynamical social psychology thus far and offers caveats concerning the relevance of this approach to the unique features of human experience.

Intrinsic Dynamics

People's mental and emotional states, overt behavior, and social relations evolve and change in the absence of external influence. The ubiquity of intrinsic dynamics is apparent at different levels of social reality, from basic intrapersonal processes to macro-level societal phenomena. At the level of the mind, the temporal pattern of cognitive and affective elements in the stream of thought [41] often provides a more accurate depiction of a person's mental make-up than do the summary aspects of the person's mental process (e. g., overall attitude, final decision) that are more often the focus of investigation [103]. For example, research has shown that simply thinking about an attitude object (e. g., another person) in the absence of external influence or new information tends to promote more extreme (polarized) evaluations of the object over time [97].

Research on social judgment has shown that internally generated thoughts and feelings about a target person often reflect elaborate but identifiable patterns of change that convey important information. A judgment that is neutral when average over time, for instance, can have very different meanings and implications, depending on the intrinsic dynamics of the judgment process [109]. When neutrality represents little variation in evaluation occurring on a relatively slow time-scale, the summary judgment might well reflect a truly neutral sentiment. If neutrality reflects oscillation between highly positive and highly negative judgments on a rapid timescale, however, the summary judgment signifies heightened involvement and ambivalence rather than neutrality per se.

Intrinsic dynamics also characterize personal action. Actions have a hierarchical structure, in that the performance of an action entails the coordinated interplay of more basic actions or sub-acts. "Going to work," for example, may involve getting dressed, leaving the house, driving a car, parking the car, and entering a building. Each of the lower-level acts, in turn, can be decomposed into yet more basic lower-level elements. "Driving," for instance, consists of starting the car, turning the steering wheel, making turns, and braking. Each level in an action hierarchy is associated with a different time scale, with increasingly lower-level acts taking place in correspondingly shorter intervals of time [69]. "Going to work" unfolds on a longer time scale than does "driving," for example, and the time scale for "driving" is longer than that for each instance of "turning the steering wheel." There is evidence that the embedded time scales in an action hierarchy often have a fractal structure [69].

The intrinsic dynamics of action, in turn, span the levels of action in an overall action hierarchy. What may appear to be a continual succession of momentary movements when defined in low-level, mechanistic terms can take on the appearance of switching between qualitatively different actions, each occurring on a longer time scale, when defined in higher-level terms. Research on action identification [107] has identified several factors that determine the level at which an action is regulated. This line of work has also demonstrated that people reliably differ in their default level of action identification across many action domains [108]. Individuals who tend to think about their actions in lower-level terms are predisposed to the emergence of higher-level action understanding. Emergence occurs when such individuals are exposed to cues (e. g., feedback from other people) that suggest higher-level meaning for the actions [114] or when they privately reflect on their actions, allowing the lower-level action elements to self-organize into a higher-level act identity that provides subjective integration for these elements [104].

Social interaction has also been investigated with respect to intrinsic dynamics. For example, research has focused on the interpersonal coordination of relatively low-level actions, such as speaking (e. g. [20,23]) and limb movement (e. g. [6,48,69,101]). In one approach, two individuals are asked to swing their legs while sitting down across from one another [6]. One person swings his or her legs in time to a metronome and the other person tries to match those movements. This research has revealed two forms of coordination: *in-phase*, with the individuals swinging their legs in unison, and *anti-phase*, with the individuals swinging their legs with the same frequency but in the opposite direction. Individuals can maintain anti-phase coordination only up to a certain frequency of movement, at which point they switch to in-phase coordination. When the frequency is subsequently decreased, at some value they are able to coordinate anti-phase again, but this tempo is significantly lower than the point at which they originally started to coordinate in-phase. Such *hysteresis* indicates that movement coordination can be analyzed as a nonlinear dynamical system [38,48]. This line of research has identified modes of coordination more complex than in-phase and anti-phase [5,87,101].

Interpersonal dynamics are not confined to the coordination of speech and motor movements, but also include the temporal coordination of higher-level actions (e. g., plans, goals) and internal states (e. g., moods, judgments). Although this topic has not been heavily researched, there is evidence that the quality of a social relationship is reflected in partners' ability to coordinate in-phase with respect to their respective higher-level actions, opinions, and feelings (e. g. [5,36,64,82,83,100]). The ebb and flow of feelings, information exchange, and action conveys deeper insight into the nature of a relationship than do global indices such as the average sentiment, the amount of information exchanged, or the summary action tendencies. Colloquially, people who feel positively about one another are said to "be in synch" or "on the same wavelength" with respect to their internal states.

At the societal level, tracking the temporal trajectory associated with the emergence of norms and public opinion provides greater insight into the society's future make-up and likely response to external threat than simply knowing what the societal norms and opinions are at a single point in time [76]. When norms and opinions develop gradually over a long period of time, for example, the society displays resistance to external threats or even to new information that might promote better economic conditions. But societal change in political and economic ideology can occur in a rapid, nonlinear manner (e. g. [75,84]), with a temporal trajectory that resembles phase transitions in physical systems [58]. Societies that undergo such nonlinear transitions are vulnerable to later rebounds of the earlier ideologies and are highly responsive to threats and new information, and they can experience a period of sustained oscillation between conflicting worldviews [75].

Attractors in Psychological Systems

Psychological systems are characterized by intrinsic dynamics, but they also demonstrate stability and resistance to change. Each day, people encounter vast amounts of information relevant to social judgment and interpersonal relations, much of which is mutually contradictory. People nonetheless manage to form and maintain coherent patterns of thought and behavior in their social lives. The partners to a romantic relationship, for example, experience a wide variety of thoughts and feelings about one another, but over time each partner's mental state tends to converge on positive sentiment toward the other. Thus, despite the ever-changing nature of intrapersonal and interpersonal experience, people's mental, emotional, and behavioral states tend to converge on relatively narrow sets of specific states or on patterns of change between specific states. These states or patterns of change are referred to as *attractors*.

When a system is at its attractor, it tends to maintain that state despite potentially destabilizing forces and influences. An external influence may move the system to another state, but the system will return fairly quickly to one of its attractors. Several social psychological phenomena imply the existence of an attractor. For example, self-regulation is defined in terms of resistance to temptations and distractions, impulse control, and the maintenance of states representing salient personal standards and values [14,15,105]. In the same fashion, self-esteem maintenance [98], self-verification [95], and psychological reactance [8] all reflect the tendency of mental systems to converge on a particular state (e. g., a level of self-esteem) and to resist influences that threaten to dislodge the person's judgments and beliefs from that state.

Three basic types of attractors are commonly distinguished: fixed-point attractors, periodic (including multiperiodic) attractors, and deterministic chaos (cf. [25,73,89]). Each type is likely to have relevance for understanding different intrapersonal and interpersonal processes, although work designed to establish such relevance is in its nascent stage. A system can be characterized by another set of points that have the opposite effect of attractors. Termed *repellors*, they represent unstable equilibria of the system, so that the smallest departure from their exact values will result in the system rapidly escaping

from the region surrounding that value. Whereas attractors may be described as states that the system "seeks" over long periods of time, repellors reflect states that the system selectively "avoids." The repellor concept has been used to describe how neural networks avoid particular memories [105].

Fixed-Point Attractors

A *fixed-point attractor* is similar to the notion of equilibrium or homeostasis [12,66]. It represents the case in which the state of the system converges to a stable value. A desired end-state or goal [14,15,104,106] for example, can be described as a fixed-point attractor. This is apparent when a person maintains a belief, an interpersonal evaluation, or an action tendency despite forces or sources of information that challenge these tendencies. Attractors, however, are not limited to goals, intentions, or other desired states. Thus, a person might display a pattern of antagonistic behavior in his or her social relations, despite efforts to avoid behaving in this manner. Similarly, a person with low self-esteem may initially embrace flattering feedback from someone, but over time the person may discount or reinterpret this feedback, displaying instead a pattern of self-evaluative thought that converges on a negative state [96]. In an inter-group context, meanwhile, warring factions may display conciliatory gestures when prompted to do so, but revert to a pattern of antagonistic thought and behavior when the outside interventions are relaxed [19]. In short, when a system's dynamics are governed by a fixed-point attractor, the system will consistently evolve to a particular state, even if this state is not hedonically pleasant, and will return to this state despite being perturbed by forces that might promote a more pleasant state.

A psychological system may have more than one attractor, with each corresponding to a distinct equilibrium. Which attractor governs a system's dynamics in a particular instance depends on the starting values of the system's evolution. The set of initial states that converge on each attractor represents the *basin of attraction* for that attractor. For a person or a group characterized by multiple fixed-point attractors, then, the process in question can display different equilibrium tendencies, each associated with a distinct basin of attraction. Within each basin, different initial states will eventually converge on the same stable value. However, even a slight change in the system's initial state will promote a large change in the system's trajectory if this change represents a state that falls just outside the original basin of attraction and within a basin for a different attractor. For example, in conflict situations there are

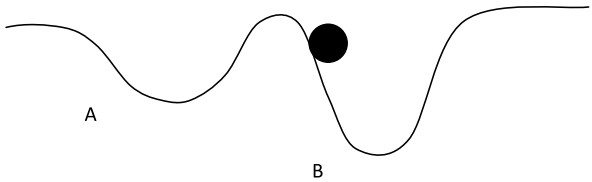

A dynamical system with two fixed-point attractors (A and B)

Social Psychology, Applications of Complexity to, Figure 1
A dynamical system with two fixed-point attractors [106]

typically two dominant responses, one corresponding to aggression and one corresponding to conciliation. Slight differences in the circumstances surrounding the conflict can thus promote dramatically different behaviors, with no option for a response that integrates the two tendencies (e. g. [19]).

The attractor concept can be captured in a simple metaphor. Figure 1 shows a ball on a hilly landscape. The ball represents the current state of the system and the valleys (A and B) represent different fixed-point attractors. The evolution of the system toward an attractor is characterized by the ball rolling down a hill and coming to rest in the bottom of a valley. Each attractor in Fig. 1 has a basin of attraction, represented by the width of the valley. The basin of attraction for Attractor A in Fig. 1 is wider than the basin for Attractor B, which means that a wider range of states will evolve toward Attractor A. Attractors can also vary in their respective strength, which is represented by the corresponding depth of the two valleys. Attractor B, then, is stronger than Attractor A. Hence, it is more difficult for external influence to change the current state of the system when the system is within the basin of attraction for Attractor B as opposed to Attractor A.

The existence of multiple fixed-point attractors in a system captures the intuition that people may have different (even mutually contradictory) goals, self-concepts, recurring emotional states, and patterns of social behavior. Someone may have two or more standards for self-regulation, for example, with each providing for action guidance under different sets of conditions. The person's behavior may reflect an affiliation standard under one set of conditions, but reflect an achievement standard under a different set of conditions. In similar fashion, a person may have multiple self-views (e. g. [37,60,61,93]), each representing an integrated and stable way of thinking about him or herself. One self-view is likely to become salient when a specific set of self-relevant information is made salient by virtue of context or role expectations. Characteristic patterns of emotional response, meanwhile, can be viewed as distinct attractors, with each attractor reflecting an "emotional Gestalt" that provides coherence for a person's cog-

nitive-affective dynamics under a unique set of conditions and constraints [57,99]. More generally, apparent inconsistency in personality can be viewed as the existence of multiple attractors, each associated with a different basin of attraction for thought, feeling, and action [82,86,92]. One set of conditions might promote a trajectory that evolves toward dominance and competition, for instance, whereas another might promote empathy and cooperation.

The strength of an attractor and the size of its basin of attraction are potentially independent, and different combinations of these basic properties may have unique implications for psychological processes [82,83]. Note again the two attractors in Fig. 1. Attractor A is relatively weak but has a wide basin of attraction. Thus, a relatively small force may change the state of the system (i. e., move the ball up the gradual slope), but the system is likely to return to the attractor (i. e., it will roll back into the valley) even if these changes are relatively large. In contrast, Attractor B has a narrower basin of attraction but is relatively strong. A great deal of force is required to promote even a slight impact on the system (i. e., move the ball up the steep slope), but if this effect is achieved, the system will not return to the attractor (i. e., it will escape the valley).

As an example, consider a romantic couple that has two attractors: a strong attractor associated with positive feelings and a weak attractor associated with negative feelings. Assume the couple has a wider range of attraction for positive feelings than for negative feelings. The partners are likely to evolve toward positive feelings about one another if they begin an interaction within a broad range of affective states (e. g., neutral to very positive), but they may end up feeling negative if they begin an interaction within a different (more restricted) range of affective states (e. g., mildly to highly negative). A broader range of initial states are likely to promote a communication trajectory that results in an exchange of warm sentiments as opposed to critical comments. If, however, the couple typically starts out with negative feelings, the negative attractor, despite having a narrow basin, may dictate the trajectory for feelings expressed in the couple's interactions. It is conceivable, of course, that the couple has a wider basin of attraction for negative feelings, so that anything short of a highly positive initial state will dissolve into a negatively toned exchange (see [33,34]).

Latent Attractors

A system may have multiple attractors, but when the system is at one of them, the others may not be visible to observers, or perhaps not even to the actors themselves. The existence of these potential states might not even be suspected. Such *latent attractors* may be highly important in the long run, though, because they define which states are possible for the system when conditions change. By identifying possibilities for a system that have yet to be experienced, the concept of latent attractor goes beyond the traditional notion of equilibrium (e. g. [2]). Critical changes in a system might not be reflected in the manifest state of the system, but rather in the creation or destruction of a latent attractor representing a potential state that is currently invisible.

The implications of latent attractors have been explored in the context of inter-group characterized by seemingly intractable conflict [18,19,85]. Factors such as objectification, dehumanization, and stereotyping of outgroup members are preconditions for the development of intractable conflict [17,22] but their immediate impact may not be apparent. Rather, these factors may create a latent attractor to which the system can abruptly switch in response to a provocation that is relatively minor, even trivial. By the same token, efforts at conflict resolution that seem fruitless in the short run may have the effect of creating a latent positive attractor for inter-group relations, thus establishing a potential relationship to which the groups can switch if other conditions permit. In this case, the existence of a latent positive attractor might promote a rapid de-escalation of conflict, even between groups with a long history of and seemingly intractable conflict.

Periodic Attractors

Rather than converging on a stable value, systems can display sustained rhythmic or oscillatory behavior. Such a temporal pattern is referred to as a periodic or limit-cycle attractor. Periodicity is obviously associated with many biological phenomena, such as circadian rhythms and menstrual cycles [30], but this dynamic tendency may also underlie certain psychological phenomena [32]. Moods have been shown to have a periodic structure, for example, often corresponding to a weekly cycle (e. g. [9,50,51]). Research on the intrinsic dynamics of both social judgment [109] and self-evaluation [110], meanwhile, has demonstrated that the stream of thought often oscillates between positive and negative assessments, sometimes in accordance with remarkably fast time scales. Periodic structure also characterizes human action [69] and is a feature of social interaction as well (e. g. [6,32,70]).

Distinguishing a periodic attractor from the existence of multiple fixed-point attractors can prove difficult, since the system in both cases displays movement between different states over time. The difference centers on the reg-

ularity of the movement between states and the role of external factors in producing such movement. A periodic attractor reflects a repetitive temporal pattern, such that the values of the dynamical variable repeat after a time $T, x_i(t) = (t + T)$, where T is the period of motion. The state of the system undergoes constant change even in the absence of noise or external influence. To qualify as a periodic attractor, then, a pattern of change must represent a pattern on which the system converges, and to which it returns after small perturbations. In a daily activity cycle, for example, a sleepless night might temporarily disrupt the pattern (e. g., oversleeping the next few days), but eventually the pattern will be restored.

A system with fixed-point attractors, in contrast, tends to stabilize on a particular state or set of states. Such attractors capture all trajectories within their respective basins, so a disturbance, noise, or an external influence is necessary to move the system from one stable state to another. A person with self-regulatory standards for both compromise and confrontation, for example, will display one of these tendencies as long as the action context is within the basin of attraction for that tendency. If the attractors differ in the size of their respective basins, and if contexts are avoided that attract the person's mental, emotional, or behavioral state toward the smaller basin, the person may behave for long periods of time in line with the stronger attractor. In similar fashion, a romantic couple may have fixed-point attractors for both positive and negative affective states, but whether they display periodic movement between them will depend on the starting conditions associated with their interactions. Even if the couple oscillates between positive and negative states, each state provides at least temporary stability. In periodic evolution, stability is not provided by any particular state, but rather by the pattern of changes between states.

The distinction between fixed-point and periodic attractors was observed in a study investigating the temporal trajectories of affective states on the part of bipolar depressive individuals [42]. Time-series analysis of mood and other symptoms revealed that many of these patients oscillated between a normal and a depressed state. However, patients whose temporal dynamics did not reflect fixed-point attractor tendencies were at highest risk for suicide and were hospitalized more often for their depression. These risks were low for individuals whose moods oscillated around a single attractor, even one corresponding to a depressed state, and for those whose moods switched between two distinct attractors reflecting a normal state and a depressed state. These results imply an interesting connection between attractor dynamics and self-regulatory tendencies. Self-regulation implies stabilization with respect to some states and de-stabilization of other states. The stable states reflect fixed-point attractors for a person's mental and emotional dynamics. The lack of fixed-point attractors for one's internal state signals a breakdown in the capacity for self-regulation.

Deterministic Chaos

The best known phenomena concerning nonlinear dynamical systems centers on *deterministic chaos* (cf. [89]). Many researchers and scholars, especially those from fields other than mathematics and physics, commonly discuss the primary insights from the work on nonlinear dynamics as chaos theory. When investigating a chaotic system, anything less than infinite precision in the knowledge of a system at one point in time can undermine prediction of the system's future states. This decoupling between determinism and practical predictability happens because all initial inaccuracies are amplified by the system's intrinsic dynamics, so that the inaccuracies grow exponentially over time. After some time, exponential growth assures that the size of the error will exceed the possible range of states of the system's behavior.

Chaos clearly represents a possibility in nonlinear dynamical systems [31,73], and has been demonstrated in many biological and physical phenomena. In principle, then, human thought and behavior may sometimes follow a chaotic trajectory. Despite this potential, though, unequivocal evidence for deterministic chaos in human thought and behavior remains to be documented. Human dynamics always contain some degree of randomness and human behavior is often unpredictable. It can be quite difficult, however, to determine the degree to which such unpredictability reflects deterministic chaos, the stochastic nature of the laws governing human nature, or the multitude of influences unaccounted for by measurement that can be treated as noise.

Dynamical Minimalism

In canonical social psychology, the complexity of human thought and behavior is typically assumed to reflect complex interactions among a large number of variables in the traditional approach to theory construction. The approach of *dynamical minimalism* [72], in contrast, assumes that complex properties can result from simple rules specifying the interactions among very simple elements. The emergence of complexity occurs when the elements are nonlinear and interact over time [39]. For example, complex cognitive phenomena (e. g., pattern recognition, error correction), have been observed in a simple network of binary elements, where each element reacts to the input it

receives from other elements [40]. In this spirit, dynamical minimalism attempts to identify the simplest set of assumptions capable of producing a phenomenon of interest. In formal models, this is equivalent to identifying the simplest mathematical rules to express what is known about a phenomenon. The goal of dynamical minimalism, then, is to achieve parsimony in theory construction without stripping the phenomenon of its subtlety and nuance.

This approach provides a new perspective on the relation between micro and macro levels of description. Reductionism is commonly assumed in mainstream models, such that the rules observed at one level of description correspond to the rules operating at another level. In effect, the properties at a macro level of description are reduced to the properties of elements at a micro level. One might, for example, explain the relation between poverty and crime in a social system by reducing this relation to the relation between frustration and crime at the level of individuals in the society. Dynamical models, however, do not assume isomorphism among levels of description. To the contrary, the rules specifying the interaction among a system's elements are likely to generate very different rules at higher levels of system behavior (see [24] for an early appreciation of this idea).

The emergence of new properties at a macro level is illustrated in the *society of self* model of self-structure [81]. This model assumes very simple rules by which self-relevant information is integrated in forming a self-concept. Each element of information (e. g., an episodic memory, a physical feature, a self-perceived trait) adopts the prevailing valence (positive vs. negative) of related elements. This simple rule, when iterated over time, generated interesting but largely unexpected consequences at the global level of self-understanding. The self-structure became differentiated into locally coherent regions (e. g., social roles, areas of competence), each of which displayed resistance to discrepant information (e. g., negative social feedback). Global self-esteem and high self-concept certainty also emerged at the macro level from the simple rule of influence among elements at the micro level.

Emergence seems to represent a paradox for theory construction. How can knowledge of the lower-level elements provide an explanation of the higher-level properties if properties at a macro level cannot be derived from properties of the system's lower level elements? The role of computer simulations in dynamical minimalism helps to resolve this paradox. With computer models, one can specify the properties of system elements and the rules of interaction among these elements. As the elements interact in accordance with these rules over time, dynamics appear at the system level that were not assumed or programmed for the elements themselves. With computer simulation, then, a theory constructed at a basic level of psychological reality (e. g., moment-to-moment thought process, dyadic social interaction) can be tested at a higher level of psychological reality (e. g., social judgment, group norms).

Computer simulations play another important role in dynamical minimalism. The basic elements comprising a system are often uninteresting and trivial, and the interactions among them may have only minor impact on the systems' global properties. But some properties of system elements may have an important impact on the system's higher-order properties as the elements interact over time. It may not be obvious which properties are trivial and which are essential for the emergence process. Dynamical minimalism makes this distinction and thus constructs a model that incorporates only the properties that are critical for the emergence of macro level phenomena. Computer simulations enable one to systematically vary the assumptions regarding the properties of elements and their interactions, and then observe which assumptions promote meaningful changes at the macro level. Those properties that have trivial consequences at the macro level are eliminated from the model. In short, computer simulations enable one to distill the minimal set of components necessary to capture the essence of a phenomenon.

Computer simulations cannot substitute for empirical verification of a theoretical model. They are critical in identifying the properties that are central to the model and investigating the consequences of these properties for the functioning of the system in question. Knowledge of these consequences then provides the basis for framing hypotheses to be investigated in empirical research. The relationship between computer simulations and empirical research work in the other direction as well. Empirical studies can refine a model, which is then implemented in computer simulations. The results of the simulations, in turn, may generate new hypotheses to be tested in subsequent empirical research. The reciprocal feedback loops among theory, computer simulation, and empirical research is central to the approach of dynamical minimalism.

The Dynamics of Social Influence

Social influence refers to any change in an individual's thoughts, feelings, or behavior that occurs as the result of the real or imagined presence of others [3]. It is widely considered to be the core process of social experience [111] and is manifest in a wide array of phenomena, including obedience to authority, conformity, imitation and modeling, bystander intervention in emergencies, social loaf-

ing, stage fright, persuasion, and groupthink. There is evidence, however, that the essence of social influence reflects the interplay of three basic factors: the number, strength, and immediacy of the sources of influence [52]. Empirical research has shown that the magnitude and nature of social influence represents a multiplicative function of these sources.

Nowak et al. [76] modeled the dynamics of social influence using cellular automata. In this approach, each individual is represented with three properties: an opinion on a topic, a degree of persuasive strength, and a position in a social space. Individuals are commonly assumed to have one of two opinions on an issue (e. g., pro vs. con). The group consists of n individuals located on a two-dimensional grid, with each cell corresponding to an individual (see Fig. 2). The color of each cell specifies that individual's current opinion (light gray denotes pro, dark gray denotes con), whereas the height of the cell represents the individual's strength (e. g., expertise, confidence, charisma). Each individual discusses the issue with other group members ascertains the degree of support for each position. As a result of these assessments, each individual adopts the opinion that is most prevalent. The following formula expresses the strength of influence of each opinion:

$$I_i = \left(\sum_{1}^{N} \left(\frac{s_j}{d_{ij}^2} \right)^2 \right)^{1/2} ,$$

where I_i denotes total influence, s_j represents the strength of each individual, and d_{ij} represents the distance between individuals i and j. The opinions of those who are closest to the individual and have the greatest strength are weighted most heavily by the individual. An individual's own position is also considered and is weighted most heavily because of its immediacy (0 distance). Influence grows with the square root of the number of people exerting influence.

In each round of the simulations, one individual is chosen at random and influence is computed for each opinion in the group. A simple updating rule is employed: the individual changes his or her opinion to match the prevailing opinion if the resultant strength of this opinion position is greater than the strength of the individual's current position. This basic process is performed for each individual in the group. This procedure is repeated until there are no further changes in opinion. This typically involves several rounds of simulation, because an individual who had previously changed his or her position to match that of his or her neighbors may revert to the original position if the neighbors change their opinions.

Figure 2 depicts representative results of the computer simulations. In the initial configuration (Fig. 2a), there is a majority of 60% (light gray) and a minority of 40% (dark gray), with the majority and minority opinions randomly distributed in social space. The majority and minority groups have the same relative proportions of strong and weak members (tall vs. short cells). Figure 2b shows that an equilibrium is reached after six rounds of simulated discussion. Although the majority has grown (to 90%) at the expense of the minority (now 10%), the minority opinion has survived by forming clusters of like-minded people. Note that these clusters are primarily formed around strong individuals.

These two group-level outcomes—polarization and clustering—are routinely observed in computer simulations of this process [55]. Each is reminiscent of well-documented social processes. Research on group dynamics (e. g. [68]), for example, has shown that the average attitude in a group becomes polarized in the direction of the prevailing attitude as a result of group discussion. In the model, polarization reflects the greater influence of the majority opinion. In the initial (random) configuration (Fig. 2a) the average proportion of neighbors holding a particular opinion (pro or con) reflects the proportion of this opinion in the total group. The average group member, in other words, is surrounded by more majority than minority members, so more minority members are converted to the majority position than vice versa. However, some majority members are converted to the minority position because they happen to be located close to an especially strong minority member or because more minority members happen to be at this location.

Clustering is also a pervasive feature of social life, having been documented for such diverse facets of social life as attitudes, political beliefs, religions, clothing fashions, and farming techniques. It has been show, for example, that attitudes tend to cluster in residential neighborhoods [28]. When opinions are distributed randomly, the sampling of opinions through social interaction provides a realistic portrayal of the distribution of opinions in the larger society. However, when opinions are clustered, the same sampling process produces a very biased result because the opinions of one's nearby neighbors are weighted the most heavily. The prevalence of one's own opinion is therefore likely to be over-estimated. Opinions that are in the minority in global terms, then, can form a local majority. This process enables individuals with a minority opinion to maintain this opinion in the belief that it actually represents a majority position.

In the spirit of dynamical minimalism, this model focuses on the minimal set of processes responsible for the emergence of group-level properties that are invariant across diverse areas of topical interest. In so doing, it pro-

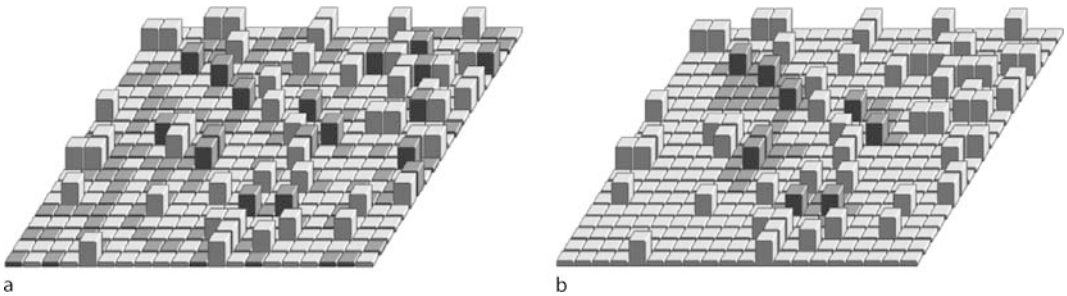

Social Psychology, Applications of Complexity to, Figure 2
a Initial distribution of opinions in the simulated group. b Final equilibrium of opinions in the simulated group [106]

vides a platform for investigating how group-level properties emerge for different domains of social functioning. This potential has been developed in recent years. Kenrick and his colleagues, for example, have simulated the emergence of cultural norms from the mutual influence among individuals with conflicting strategies (decision-rules) pertaining to fundamental human goals (e. g. [49]). These goals, reflecting domains of adaptive functioning confronted by human groups throughout history [10,11,29], include self-protection, coalition formation, status-seeking, mate choice, relationship maintenance, and offspring care.

In their computer simulations, individuals with different decision-rules regarding a particular domain interacted with one another over time, and the group-level consequences of these interactions were observed. One series of simulations investigated how individual differences in decision-rules for cooperation versus competition affected community level propensities for such behavior. Another series explored the mutual impact of male and female mating strategies on the emergence of societal-level norms regarding mating. Clustering was observed in both cases, with an initial random configuration of decision rules giving way over time to local communities characterized by coherent norms regarding the behavior in question. In the mating simulations, for example, some communities were characterized by relatively unrestricted mating strategies (reflecting the decision-rule preferred by males), whereas other communities developed norms sanctioning restricted mating strategies (reflecting the decision-rule preferred by females).

Dynamics of Interpersonal Coordination

The dynamical account of social influence provides a concise description of how the state (e. g., attitude) of a single individual depends on the state of other individuals. Because many psychological processes reflect intrinsic dynamics, however, individuals can be conceptualized as displaying patterns of change rather than as a set of states. In this view, social influence (and social interaction generally) can be investigated as the coordination over time of individual dynamics.

The most basic form of coordination is positive correlation or in-phase relation. In social interaction, this occurs when the overt behaviors, attitudes, or emotions of one person induce similar behaviors or states in the other person at the same time (e. g. [16]). This basic form of coordination is reflected in such familiar phenomena as imitation, mimicry, and empathy [62]. Other forms of coordination can be identified, however, with counterparts in different contexts for social interaction [62,69]. Turn taking in conversation represents negative correlation or anti-phase relation between individuals in their respective talking and listening (when one person speaks, the other is silent). Negative correlation also characterizes antagonistic relationships, in that the sadness or despair of one person induces satisfaction or happiness in the other person and vice versa. More complex forms of synchronization can also be identified that reflect nonlinear relationships and higher-order interactions between the partners' respective behaviors and internal states (cf. [74,82]).

A Model of Synchronization Dynamics

This perspective provides the foundation for a recently developed model of *synchronization* [82,83], a phenomenon that characterizes coupled dynamical systems [44,91]. Because positive correlation represents the most fundamental and common form of coordination, it has provided the primary focus to date in this class of models. The Nowak et al. [82,83] model assumes that each individual attempts to achieve synchronization by adjusting his or her internal state or overt behavior in response to the state or behavior of the individual with whom he or she is interacting. Individuals in social interaction, in other words, modify their respective thoughts, feelings, or action tendencies to promote positive correlation over time in these features

of experience. The synchronization of individuals' dynamics results in a higher order system with its own dynamic properties.

Coupled logistic are used to model interpersonal synchronization [74,81,82,83]. The dynamics of each individual are represented with a logistic equation, which is the simplest dynamical system capable of displaying complex (e. g., chaotic) behavior [27,89]. However, the behavior of each person not only depends on his or her preceding state but also to a certain extent on the preceding state of the other person. The coupling of individuals' dynamics is specified in the following equation:

$$x_1(t+1) = \frac{r_1 x_1(t)(1 - x_1(t)) + \alpha r_2 x_2(t)(1 - x_2(t))}{1 + \alpha}$$
$$x_2(t+1) = \frac{r_2 x_2(t)(1 - x_2(t)) + \alpha r_1 x_1(t)(1 - x_1(t))}{1 + \alpha} .$$

The dynamical variable (x) represents the intensity of behavior, and the control parameter, r, corresponds to internal states (e. g., personality traits, moods, values) that shape the person's pattern of behavior (i. e., changes in x over time). To the value of the dynamical variable representing one individual's behavior x_1, one adds a fraction α of the value of the dynamical variable representing the other individual's behavior x_2. The magnitude of α represents the strength of coupling and can be viewed as the degree of mutual influence characterizing the interaction. Depending on the social context, it might reflect the intensity of communication or the degree of mutual imitation. When α is 0, there is no coupling (e. g., no influence or communication) on the behavior level, whereas when α is 1, each individual's behavior is determined equally by his or her preceding behavior and the influence of the other individual. Intermediate values of α represent moderate values of coupling in the relationship.

Modeling Behavioral Synchronization

The respective control parameters of two individuals are unlikely to be identical when they first interact with one another. And the degree of influence between individuals differs across interactions and relationships. To represent this variability in social reality, Nowak, Vallacher [74] systematically varied the similarity of partners' control parameters (r), representing their internal states, and their degree of coupling (α), representing their mutual influence (e. g., communication, imitation). Each simulation began with a random value of x for each individual, drawn from a uniform distribution that varied from 0 to 1. They let the simulations run for 300 steps, allowing each system to converge on its pattern of intrinsic dynamics and both systems to synchronize. For the next 500 simulation steps, they recorded the values of x for each system and the degree of synchronization of the two systems.

The results demonstrated that behavioral synchronization increased both with increases in α and in the similarity in r. At each degree of coupling, synchronization increased with greater similarity in partner's internal states. Likewise, at each level of similarity in internal states, synchronization increased with increased strength of coupling. These results have straightforward implications. If two people have similar control parameters (internal states), relatively little coupling (e. g., influence, communication, mutual reinforcement, monitoring) is necessary for them to achieve a high degree of synchronization in their behavior. If the partners have different internal states, on the other hand, high mutual influence is required to maintain the same level of synchronization. In a close relationship, this suggests that constant and intense communication may be a sign that the partners are not well coordinated with respect to relevant internal states (e. g., temperament, desires values). But when the partners are similar with respect to such internal states, they can devote their energy to common pursuits rather than to constant clarification, monitoring, and other forms of influence.

Modeling Internal Synchronization

Some internal parameters, such as moods and emotional states, vary considerably across time and settings. If a particular context induces a common mood or emotion (e. g., joy or sadness) in a dyad or group, interpersonal synchronization is easy to achieve. Other internal states, however, demonstrate greater stability and less likely to vary across different contexts. Attitudes, values, and personality traits, in particular, are commonly considered to be enduring (cross-situational) features of a person's psychological make-up. Yet, even these internal parameters admit to variability and even modification. The Nowak et al. [82,83] model has been used to understand and investigate the nature of such change. The core assumption is that individuals are motivated to achieve coordination in their internal parameters [13,62]. To satisfy this motive, individuals vary their internal parameters in a direction that leads to increasing synchronization. When synchronization is achieved, the value of the resultant control parameter is engraved as an attractor for that internal state. In essence, people are assumed to develop stable internal states through social synchronization (e. g. [112]).

To model this process, Nowak et al. simply assumed that the value of each individual's control parameter drifts somewhat in the direction of the value of the other indi-

vidual's control parameter on each simulation step. How quickly the respective control parameters begin to match depends on the size of the initial discrepancy between them and on the rate at which they drift. This process occurs with each individual knowing the exact value of the other individual's control parameter. This is an important feature of the model, since experimental research has shown that people's internal states are often difficult to infer [43,47,71,113]. The model assumes only that each individual remembers the other individual's most recent behaviors (i. e., the most recent values of x) as well as his or her own most recent behaviors. Each individual compares his or her own behavior with that of the other individual, and then adjusts his or her internal parameter in order to promote increased similarity with the other person's behavior pattern, until a match is achieved [115]. Thus, if the other individual's behavior is more complex than the individual's own pattern of behavior, the individual slightly increases the value of his or her own control parameter. Conversely, the individual slightly decreases the value of his or her own control parameter if the other individual's behavior is less complex than his or her own. In short, interacting individuals estimate one another's internal state by monitoring the evolution of one another's behavior.

Nowak et al. [82,83] ran simulations to investigate the convergence of behavior and internal states under both relatively weak and relatively strong coupling ($\alpha = .25$ and .7, respectively). The y-axis in Fig. 3 corresponds to the magnitude of difference between the two systems in their behavior or internal states, and the x-axis corre-

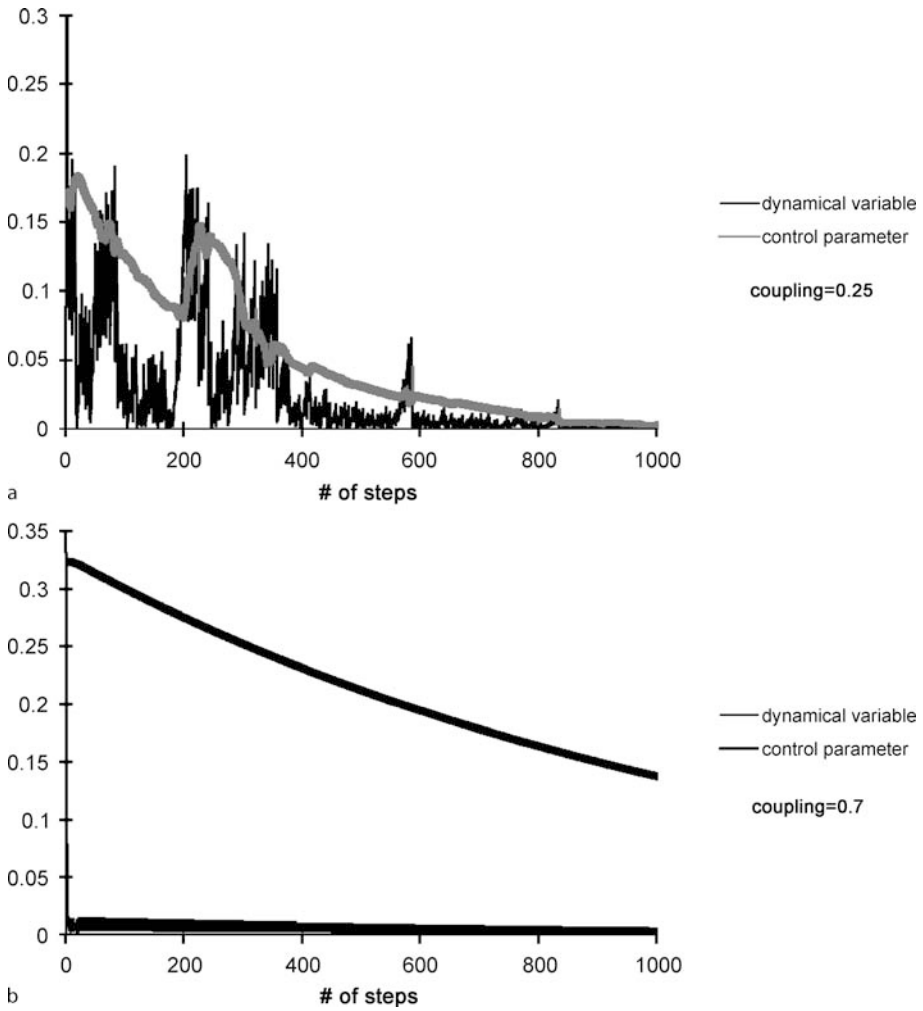

Social Psychology, Applications of Complexity to, Figure 3
a Convergence of behavior and internal states under weak coupling b Convergence of behavior and internal states under strong coupling [82]

sponds to time, as reflected in the number of iterations. Consider first the results observed when the coupling was weak (Fig. 3a). The behavior of the two systems converged in a relatively slow and non-linear manner. The control parameters also showed a clear tendency towards convergence. When a match in internal states was achieved, moreover, full synchronization of behavior was also obtained. Compare these results with those observed under strong coupling (Fig. 3b). Note that although there was immediate synchronization of behavior, the control parameters failed to synchronize, even after 1000 iterations.

The differential results obtained for weak and strong coupling may have noteworthy implications for interpersonal relations. First, even for people with very different internal parameters, strong coupling tends to promote full synchronization of behavior. Once synchronization is achieved, the two people may be totally unaware that their internal states are different. This suggests that if the coupling were to be reduced in magnitude (or removed altogether), the dynamics of the two individuals would immediately diverge. Hence, people who employ very strong influence (e. g., reinforcement, monitoring) to obtain coordination of behavior may effectively hinder synchronization of their respective internal parameters. In more general terms, there may be an optimal level of influence and control over one another's behavior in interpersonal relations [111]. When influence is too weak, synchronization between individuals may fail to develop. But when influence is very strong, it can prevent the development of a relationship based on mutual understanding and empathy. Intermediate levels of mutual influence, then, may be most effective for the development of synchronization on a deep level. Stated differently, the most advantageous degree of coupling (e. g., influence) is the minimal amount necessary to achieve synchronization.

These implications are consistent with extensive research in social psychology suggesting that behavior attributed to external causes is less likely to promote psychological change than is behavior attributed to internal causes. For example, people are resistant to changing their preferences and attitudes if they believe that their behavior is in response to direct orders, rewards, threats, and other external influences [7,56]. Salient external influences, in fact, may activate mechanisms to counter the influences, creating an internal state that is opposite of the intended effect of the influence [8,111].

Future Directions

The application of complexity science to social psychology makes for an ironic discipline. The tools of complexity and dynamical systems are ideally suited to capture the subtlety and uniqueness of human experience, yet they are grounded in concepts and methods that provide meaningful integration with the natural sciences. Disciplines such as mathematics, physics, and chemistry have affirmed the significance of intrinsic dynamics, nonlinear phenomena, self-organization, and complexity. In recent years, psychologists have become increasingly cognizant that these features of systems in nature have counterparts in mental, interpersonal, and collective experience. It would be odd for a contemporary social psychologist to discount the potential for emergence or to ignore the role of computer simulations and time series in illuminating how minds, groups, and societies work.

A word of caution is in order, however. Social reality is not the same as physical reality. Unlike atoms, individuals are not interchangeable, and groups are more than self-organized ensembles of simple particles. People have values and beliefs, universal concerns and idiosyncratic tendencies, and moments of self-reflection and sudden impulse. One of the basic rules of human psychology, moreover, is the capacity for reflecting on one's operating rules and attempting to override them. People exist in a symbolically constructed world and do not respond in a reflexive way to objective reality. These unique features of human experience cannot be chalked up to the recognition that people are dynamic and complex. The task of dynamical social psychology is thus more daunting than discovering the differential equations that govern interpersonal and collective dynamics. Models derived from complexity and dynamical systems provide a foundational science for the discipline and much remains to be discovered within this new paradigm. Ultimately, though, the properties that separate us from other systems in nature must be incorporated into theoretical models. It is an ironic testament to the complexity perspective that a coherent theory of social psychology should be assembled from elements that are both universal within nature and unique to human experience.

Bibliography

Primary Literature

1. Abelson RP (1979) Social clusters and opinion clusters. In: Holland PW, Leinhardt S (eds) Perspectives in social network research. Academic Press, New York, pp 239–256
2. Abraham RH, Shaw CD (1992) Dynamics, the geometry of behavior, 2nd edn. Addison-Wesley, Reading
3. Allport GW (1968) The historical background of modern social psychology. In: Lindzey GA, Aronson E (eds) The Handbook of Social Psychology, vol 1. Addison-Wesley, Reading

4. Asch SE (1946) Forming impressions of personalities. J Abnorm Soc Psychol 41:258–290
5. Baron RM, Amazeen PM, Beek PJ (1994) Local and global dynamics of social relations. In: Vallacher RR, Nowak A (eds) Dynamical systems in social psychology. Academic Press, San Diego, pp 111–138
6. Beek PJ, Hopkins B (1992) Four requirements for a dynamical systems approach to the development of social coordination. Huma Mov Sci 11:425–442
7. Bem DJ (1967) Self-perception: An alternative interpretation of cognitive dissonance phenomena. Psychol Rev 74:183–200
8. Brehm SS, Brehm JW (1981) Psychological reactance: A theory of freedom and control. Academic Press, New York
9. Brown KW, Moskowitz DS (1998) Dynamic stability of behavior: The rhythms of our interpersonal lives. J Personal 66:105–134
10. Bugental DB (2000) Acquisition of the algorithms of social life: A domain-base approach. Psychol Bull 126:187–219
11. Buss DM (1999) Evolutionary psychology: The new science of the mind. Allyn Bacon, Boston
12. Cannon WB (1932) The wisdom of the body. Norton, New York
13. Caporeal LR, Baron RM (1997) Groups as the mind's natural environment. In: Simpson JA (ed) Evolutionary social psychology. Lawrence Erlbaum Associates, Hillsdale, pp 317–344
14. Carver CS, Scheier MF (1999) Themes and issues in the self-regulation of behavior. In: Wyer RS Jr (ed) Advances in social cognition, vol 12. Erlbaum, Mahwah, pp 1–105
15. Carver CS, Scheier MF (2002) Control processes and self-organization as complementary principles underlying behavior. Personal Soc Psychol Rev 6:304–315
16. Chartrand TL, Bargh JA (1999) The chameleon effect: The perception-action link and social interaction. J Pers Soc Psychol 76:893–910
17. Coleman PT (2003) Characteristics of protracted, intractable conflict: Towards the development of a meta-framework - I. First paper in a three-paper series. Peace Confl: J Peace Psychol 9:1–37
18. Coleman PT, Bui-Wrzosinska L, Vallacher RR, Nowak A (2006) Protracted conflicts as dynamical systems. In: Schneider AK, Honeyman C (eds) The negotiator's fieldbook: The desk reference for the experience negotiator. American Bar Association Books, Chicago, pp 61–74
19. Coleman PT, Vallacher RR, Nowak A, Bui-Wrzosinska L (2007) Intractable conflict as an attractor: A dynamical systems approach to conflict escalation and intractability. Am Behav Sci 50:1454–1475
20. Condon WS, Ogston WD (1967) A segmentation of behavior. J Psychiatr Res 5:221–235
21. Cooley CH (1902) Human nature and the social order. Scribner, New York
22. Deutsch M (1973) The resolution of conflict: Constructive and destructive processes. Yale University Press, New Haven
23. Dittman AT, Llewellyn LG (1969) Body movement and speech rhythm in social conversation. J Personal Soc Psychol 11:98–106
24. Durkheim E (1938) The rules of sociological method. University of Chicago Press, Chicago
25. Eckmann JP, Ruelle D (1985) Ergodic theory of chaos and strange attractors. Rev Mod Phys 57:617–656
26. Eiser JR (1994) Attitudes, chaos, and the connectionist mind. Blackwell, Oxford
27. Feigenbaum MJ (1978) Quantitative universality for a class of nonlinear transformations. J Stat Phys 19:25–52
28. Festinger L, Schachter S, Back K (1950) Social pressures in informal groups. Stanford University Press, Stanford
29. Fiske AP (1992) The four elementary forms of sociality: Framework for a unified theory of social relations. Psychol Rev 99:689–723
30. Glass L, Mackey MC (1988) From clocks to chaos: The rhythms of life. Princeton University Press, Princeton
31. Goldstein J (1996) Causality and emergence in chaos and complexity theories. In: Sulis W, Combs A (eds) Nonlinear dynamics and human behavior. World Scientific, Singapore, pp 161–190
32. Gottman JM (1979) Detecting cyclicity in social interaction. Psychol Bull 86:338–348
33. Gottman J, Swanson C, Swanson K (2002) A general systems theory of marriage: Nonlinear difference equation modeling of marital interaction. Personal Soc Psychol Rev 4:326–340
34. Gottman JM, Murray JD, Swanson CC, Tyson R, Swanson KR (2002) The Mathematics of Marriage. MIT Press, Cambridge
35. Guastello SJ (1995) Chaos, catastrophe, and human affairs: Applications of nonlinear dynamics to work, organizations, and social evolution. Erlbaum, Mahwah
36. Guastello SJ, Pincus D, Gunderson PR (2006) Electrodermal arousal between participants in a conversation: Nonlinear dynamics and linkage effects. Nonlinear Dyn Psychol Life Sci 10:365–399
37. Higgins ET (1987) Self-discrepancy: A theory relating self and affect. Psychol Rev 94:319–340
38. Hock HS, Kelso JAS, Schoner G (1993) Bistability and hysteresis in the organization of apparent motion pattern. J Exp Psychol Hum Percept Perform 19:63–80
39. Holland JH (1995) Emergence: From chaos to order. Addison-Wesley, Reading
40. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences USA 79:2554-2558
41. James W (1890) Principles of psychology. Holt, New York
42. Johnson SL, Nowak A (2002) Dynamical patterns in bipolar depression. Personal Soc Psychol Rev 6:380–387
43. Jones EE, Davis KE (1965) From acts to dispositions: The attribution process in person perception. In: Berkowitz L (ed) Advances in experimental social psychology, vol 2. Academic Press, New York, pp 220–266
44. Kaneko K (ed) (1993) Theory and applications of coupled map lattices. World Scientific, Singapore
45. Kaplowitz SA, Fink EL (1992) Dynamics of attitude change. In: Levine RL, Fitzgerald HE (eds) Analysis of dynamic psychological systems, vol 2. Plenum, New York, pp 341–369
46. Kelley HH (1967) Attribution in social psychology. Neb Symp Motiv 15:419–422
47. Kelley HH (1971) Attribution in social interaction. Morristown
48. Kelso JAS (1995) Dynamic patterns: The self-organization of brain and behavior. The MIT Press, Cambridge
49. Kenrick DT, Maner JK, Butner J, Li NP, Becker DV, Schaller M (2002) Dynamical evolutionary psychology: Mapping the domains of the new interactionist paradigm. Personal Soc Psychol Rev 6:347–356

50. Larsen RJ (1987) The stability of mood variability: A spectral analytic approach to daily mood assessments. J Personal Soc Psychol 52:1195–1204
51. Larsen RJ, Kasimatis M (1990) Individual differences in entrainment of mood to the weekly calendar. J Personal Soc Psychol 58:164–171
52. Latané B (1981) The psychology of social impact. Am Psychol 36:343–356
53. Latané B, Nowak A (1994) Attitudes as catastrophes: From dimensions to categories with increasing involvement. In: Vallacher RR, Nowak A (eds) Dynamical systems in social psychology. Academic Press, San Diego, pp 219–249
54. Latané B, Nowak A (1997) The causes of polarization and clustering in social groups. Prog Commun Sci 13:43–75
55. Latané B, Nowak A, Liu J (1994) Measuring emergent social phenomena: dynamism, polarization and clustering as order parameters of social systems. Behav Sci 39:1–24
56. Lepper MR, Greene D (eds) (1978) The hidden costs of reward. Erlbaum, Hillsdale
57. Lewis MD (2005) Bridging emotion theory and neurobiology through dynamic systems modeling. Behav Brain Sci 28:169–194
58. Lewenstein M, Nowak A, Latané B (1993) Statistical mechanics of social impact. Phys Rev 45:703–716
59. Lewin K (1936) Principles of topological psychology. McGraw-Hill, New York
60. Linville PW (1985) Self-complexity and affective extremity: Don't put all your eggs in one cognitive basket. Soc Cogn 3:94–120
61. Markus H, Nurius P (1986) Possible selves. Am Psychol 41:954–969
62. Marsh KL, Richardson MJ, Baron RM (2006) Contrasting approaches to perceiving and acting with others. Ecol Psychol 18:1–37
63. Mazanov J, Byrne DG (2006) A cusp catastrophe model analysis of changes in adolescent substance use: Assessment of behavioural intention as a bifurcation variable. Nonlinear Dyn Psychol Life Sci 10:445–470
64. McGrath JE, Kelley JR (1986) Time and human interaction: Toward a psychology of time. Guilford Publications, New York
65. Mead GH (1934) Mind, self, and society. University of Chicago Press, Chicago
66. Miller NE (1944) Experimental studies of conflict. In: Hunt JM (ed) Personality and the behavior disorders. Ronald, New York
67. Moscovici S, Lage E, Naffrechoux M (1969) Influence of a consistent minority on responses of a majority in a color perception task. Sociometry 32:365–379
68. Myers DG, Lamm H (1976) The group polarization phenomenon. Psychol Bull 83:602–627
69. Newtson D (1994) The perception and coupling of behavior waves. In: Vallacher RR, Nowak A (eds) Dynamical systems in social psychology. Academic Press, San Diego, pp 139–167
70. Nezlek JB (1993) The stability of social interaction. J Personal Soc Psychol 65:930–941
71. Nisbett R, Ross L (1980) Human inference: Stategies and shortcomings of social judgment. Prentice-Hall, Englewood Cliffs
72. Nowak A (2004) Dynamical minimalism: Why less is more in psychology. Personality and Social Psychology Review 8:183-192
73. Nowak A, Lewenstein M (1994) Dynamical systems: A tool for social psychology? In: Vallacher RR, Nowak A (eds) Dynamical systems in social psychology. Academic Press, San Diego, pp 17–53
74. Nowak A, Vallacher RR (1998) Dynamical social psychology. Guilford, New York
75. Nowak A, Vallacher RR (2001) Societal transition: Toward a dynamical model of social change. In: Wosinska W, Cialdini RB, Barrett DW, Reykowski J (eds) The practice of social influence in multiple cultures. Lawrence Erlbaum, Mahwah, pp 151–171
76. Nowak A, Szamrej J, Latané B (1990) From private attitude to public opinion: A dynamic theory of social impact. Psychol Rev 97:362–376
77. Nowak A, Lewenstein M, Szamrej J (1993) Social transitions occur through bubbles. Sci Am Pol vers 12:16–25
78. Nowak A, Urbaniak J, Zienkowski L (1994) Clustering processes in economic transition. RECESS Res Bull 3:43–61
79. Nowak A, Lewenstein M, Frejlak P (1996) Dynamics of public opinion and social change. In: Hegselman R, Pietgen HO (eds) Modeling social dynamics: Order, chaos, and complexity. Helbin, Vienna, pp 54–78
80. Nowak A, Vallacher RR, Borkowski W (2000) Modeling the temporal coordination of behavior and internal states. In: Ballot G, Weisbuch G (eds) Applications of simulation to the social sciences. Hermes Science Publications, Oxford, pp 67–86
81. Nowak A, Vallacher RR, Tesser A, Borkowski W (2000) Society of self: The emergence of collective properties in self-structure. Psychol Rev 107:39–61
82. Nowak A, Vallacher RR, Zochowski M (2002) The emergence of personality: Personal stability through interpersonal synchronization. In: Cervone D, Mischel W (eds) Advances in personality science. Guilford Press, New York, pp 292–331
83. Nowak A, Vallacher RR, Zochowski M (2005) The emergence of personality: Dynamic foundations of individual variation. Dev Rev 25:351–385
84. Nowak A, Vallacher RR, Kus M, Urbaniak J (2005) The dynamics of societal transition: Modeling nonlinear change in the Polish economic system. Int J Sociol 35:65–88
85. Nowak A, Vallacher RR, Bui-Wrzosinska L, Coleman PT (2007) Attracted to conflict: A dynamical perspective on malignant social relations. In: Golec A, Skarzynska K (eds) Understanding social change: Political psychology in Poland. Nova Science Publishers, Haauppague
86. Read SJ, Miller LC (2002) Virtual personalities: A neural network model of personality. Personal Soc Psychol Rev 6:357–369
87. Rosenblum LD, Turvey MT (1988) Maintenance tendency in coordinated rhythmic movements: Relative fluctuations and phase. Neuroscience 27:289–300
88. Schachter S (1951) Deviation, rejection and communication. J Abnorm Soc Psychol 46:199–207
89. Schuster HG (1984) Deterministic chaos. Physik, Vienna
90. Sherif M (1936) The psychology of social norms. Harper, New York
91. Shinbrot T (1994) Synchronization of coupled maps and stable windows. Phys Rev E 50:3230–3233
92. Shoda Y, LeeTiernan S, Mischel W (2002) Personality as a dynamical system: Emergence of stability and distinctiveness

from intra- and interpersonal interactions. Personal Soc Psychol Revi 6:316–325
93. Showers CJ (1995) The evaluative organization of self-knowledge. In: Kernis MH (ed) Efficacy, agency, and self-esteem. Plenum, New York, pp 101-120
94. Simon D, Holyoak KJ (2002) Structural dynamics of cognition: From consistency theories to constraint satisfaction. Personal Soc Psychol Rev 6:283–294
95. Swann WB Jr (1990) To be adored or to be known? The interplay of self-enhancement and self-verification. In: Higgins ET, Sorrentino RM (eds) Handbook of motivation and cognition: Foundations of social behavior, vol 2. Guilford, New York, pp 408–448
96. Swann WB, Hixon JG, Stein-Seroussi A, Gilbert D (1990) The fleeting gleam of praise: Cognitive processes underlying behavioral reactions to self-relevant feedback. J Personal Soc Psychol 59:17–26
97. Tesser A (1978) Self-generated attitude change. In: Berkowitz L (ed) Advances in experimental social psychology, vol 11. Academic Press, New York, pp 85–117
98. Tesser A, Martin L, Cornell D (1996) On the substitutability of self-protective mechanisms. In: Gollwitze PM, Bargh JA (eds) The psychology of action. Guilford Publications, New York, pp 48–68
99. Thagard P, Nerb J (2002) Emotional Gestalts: Appraisal, change, and the dynamics of affect. Personal Soc Psychol Rev 6:274–282
100. Tickle-Degnen L, Rosenthal R (1987) Group rapport and nonverbal behavior. Rev Personal Soc Psychol 9:113–136
101. Turvey MT (1990) Coordination. Am Psychol 4:938–953
102. Vallacher RR, Nowak A (eds) (1994) Dynamical systems in social psychology. Academic Press, San Diego
103. Vallacher RR, Nowak A (1994) The stream of social judgment. In: Vallacher RR, Nowak A (eds) Dynamical systems in social psychology. Academic Press, San Diego, pp 251–277
104. Vallacher RR, Nowak A (1997) The emergence of dynamical social psychology. Psychol Inq 8:73–99
105. Vallacher RR, Nowak A (1999) The dynamics of self-regulation. In: Wyer RS Jr (ed) Advances in self-regulation, vol 12. Lawrence Erlbaum, Mahwah, pp 241–259
106. Vallacher RR, Nowak A (2007) Dynamical social psychology: Finding order in the flow of human experience. In: Kruglanski AW, Higgins ET (eds) Social psychology: Handbook of basic principles, 2nd edn. Guilford Publications, New York, pp 734–758
107. Vallacher RR, Wegner DM (1987) What do people think they're doing? Action identification and human behavior. Psychol Rev 94:1–15
108. Vallacher RR, Wegner DM (1989) Levels of personal agency: Individual variation in action identification. J Personal Soc Psychol 57:660–671
109. Vallacher RR, Nowak A, Kaufman J (1994) Intrinsic dynamics of social judgment. J Personal Soc Psychol 66:20–34
110. Vallacher RR, Nowak A, Froehlich M, Rockloff M (2002) The dynamics of self-evaluation. Personal Soc Psychol Rev 6:370–379
111. Vallacher RR, Nowak A, Miller ME (2003) Social influence and group dynamics. In: Weiner I, Millon T, Lerner MJ (eds) Handbook of psychology, vol 5. Personality and social psychology. Wiley, New York, pp 383–417
112. Vallacher RR, Nowak A, Zochowski M (2007) Dynamics of social coordination: The synchronization of internal states in close relationships. In: Hauf P, Forsterling F (eds) Making minds: The shaping of human minds through social context. John Benjamins Publishing Company, Amsterdam, pp 31–46
113. Wegner DM, Vallacher RR (1977) Implicit psychology: An introduction to social cognition. Oxford University Press, New York
114. Wegner DM, Vallacher RR, Kiersted G, Dizadji D (1986) Action identification in the emergence of social behavior. Soc Cogn 4:18–38
115. Zochowski M, Liebovitch LS (1999) Self-organizing dynamics of coupled map systems. Phys Rev E 59:2830

Books and Reviews

Guastello S, Koopmans M, Pincus D (eds) (2008) Chaos and Complexity in Psychology: The Theory of Nonlinear Dynamical Systems. Cambridge University Press, New York
Hegselman R, Pietgen HO (eds) (1996) Modeling social dynamics: Order, chaos, and complexity. Helbin, Vienna
Lewin K (1936) Principles of topological psychology. McGraw-Hill, New York
Liebrand W, Nowak A, Hegselman R (eds) (1998) Computer modeling and the study of dynamic social processes. Sage, New York
Read SJ, Miller LC (eds) (1998) Connectionist models of social reasoning and social behavior. Erlbaum, Mahwah
Smith ER (1996) What do connectionism and social psychology offer each other? J Personal Soc Psychol 70:893–912
Vallacher RR, Read SJ, Nowak A (eds) (2002) The dynamical perspective in social psychology. Personal Soc Psychol Rev, Special Issue 6:264–387

Soft Computing, Introduction to

Janusz Kacprzyk
Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

As science progresses, scholars and researchers continue to develop models of various phenomena and problems. These models have become more and more formal, notably mathematical, and hence have relied more and more on computing.

The initial period has been characterized by computing meant as well-defined algorithmic procedures implemented by using computing architectures, tools and techniques based on traditional binary (yes – no or 0 – 1) logic. This logic has also been a basis for all kinds of approaches within the field of artificial intelligence, a field of science and technology that has been vividly developed since the mid-1900s and aimed at devising machines capable of performing human-specific tasks that require

intelligence. Symbolic computation has become a main paradigm within artificial intelligence, and the role of numerical computing has been neglected to a large extent.

As promises of artificial research gurus that machines will be able to perform virtually all sophisticated human-specific tasks and practically replace humans in a short time have failed, it has become more and more evident that sticking to symbolic computations only is a mistake. Clearly, there may be tasks for which symbolic computations are adequate, but there may be tasks, presumably even more, when numerical computations are indispensable.

Even if this necessity of a synergistic use of symbolic and numerical computations in the construction of intelligent systems has become obvious, people have been, for a long time, not fully convinced that, since the human being is a crucial element in all kinds of research, analysis and construction of broadly perceived intelligent systems, computational techniques and processes involved therein should reflect human-specific features.

In our context, the main feature of the human element in computation is that natural language is the only fully natural means of articulation and communication of a human being, and natural language is plagued by inherent vagueness, ambiguity, imprecision, etc. These kinds of information imperfections are out of scope of conventional mathematical tools based on – for instance – probabilistic or statistical tools.

The necessity of developing a new framework which would differ from the traditional "hard" computing based on traditional mathematics, i. e., basically on binary logic, and the dichotomy of true and false, has been recognized quite early. However, it was Lotfi A. Zadeh from the University of California at Berkeley who first pronounced that necessity – in the context of the analysis of complex animate systems, notably human-centered – in a more comprehensive way in the early 1960s. In 1965 he introduced fuzzy sets theory, presumably the first comprehensible, yet simple and efficient calculus of imprecision. A fuzzy set is a class to which elements can belong to a degree, from full membership, through all intermediate values, to full non-membership. Clearly, this has provided a constructive means for the representation of imprecisely specified concepts, relations, reasoning schemes, etc. Zadeh then proceeded to the formulation of possibility theory, and finally based his ideas on a more direct link to natural language which has culminated in his computing with words and perceptions.

The introduction of a fuzzy set, followed by numerous applications of fuzzy logic-based tools and techniques, notably for control, ranging from the control of domestic appliances, automatic transmissions in cars, to the control of industrial installations, has certainly been very important for the proliferation of fuzzy logic, maybe even more generally for non-conventional computational tools and techniques.

As a next step, Zadeh has advocated a wider view of those new computational tools and techniques and coined the term of *soft computing*. Basically, soft computing may be viewed as a departure from conventional (hard and "precise") computing in that it is tolerant of imprecision, uncertainty, partial truth, and approximate relationships, approximately satisfied properties, etc.

The main principle underlying soft computing is to be able to exploit a tolerance for imprecision, uncertainty, partial truth, and approximation to attain tractable, robust and computationally inexpensive modeling and solution tools and techniques. A natural connection to neural computing and evolutionary computing has more recently been recognized.

It is usually assumed that the principal components of soft computing are fuzzy logic (complemented with rough sets theory which may be viewed to address related problems of imprecise information but understood and dealt with in a different way), neural computation, and evolutionary computation. These areas are sometimes complemented with machine learning and probabilistic reasoning, with the latter subsuming belief networks, chaos theory and parts of learning theory. Moreover, for a long time, rough sets have been considered to be a relevant part of soft computing.

However, it should be noticed that soft computing is not a confluence of more or less independent fields, but is meant as a synergistic combination of various methodologies and methods in which each one contributes in a synergistic and complementary way to the analysis and solution of a problem.

In this section we will present the main components of soft computing. The concepts of synergy and complementarity mentioned above will provide a pivotal perspective.

We start with a brief introduction to fuzzy sets and systems (see J. Kacprzyk, ► Fuzzy Sets Theory, Foundations of). First, we define the concept of a fuzzy set followed by the definitions of its main properties. The idea of a fuzzy relation is then introduced, and shown to be a convenient apparatus for devising approximate reasoning schemes. The concept of a linguistic variable is shown to be a basis for a natural-language-based approach to modeling and reasoning. Fuzzy arithmetic is presented making it possible to use imprecisely specified (fuzzy) numbers in models and algorithms. A universal problem of decision making

is dealt with under fuzzy goals, constraints, etc., in a static and dynamic context.

Though fuzzy sets in the traditional Zadeh sense do provide a simple yet efficient means for the representation and handling of imprecision, some natural extensions of the basic concept of a fuzzy set have appeared. One of the most relevant and popular is the concept of a type 2 fuzzy set in which the degree of membership itself, which is a real number from the unit interval in the traditional fuzzy set, is now a fuzzy set defined in the unit interval. This natural extension has implied relevant extensions of fuzzy modeling and reasoning tools and techniques which are presented in detail, and potentials for applications are underlined (see R.I. John and J.M. Mendel, ► Fuzzy Logic, Type-2 and Uncertainty).

As in any formal model in which, as is true in virtually all cases in reality, multiple aspects, points of view and entities have to be accounted for, a proper aggregation is of a paramount importance. A comprehensive account of various aggregation operators is provided, both from a strictly mathematical point of view and from a more constructive, application oriented one (see V. Torra, ► Aggregation Operators and Soft Computing).

The discussion of the foundations of fuzzy logic is completed with a brief account of possibility theory initiated by Zadeh in the late 1970s as an extension of fuzzy sets theory and fuzzy logic, a mathematical theory for dealing with certain types of imperfect (uncertain) information in a way that is an alternative to probability theory. Main concepts related to possibility theory are presented, and a historical perspective is provided (see D. Dubois and H. Prade, ► Possibility Theory).

The comprehensive presentation of foundational concepts and properties related to fuzzy logic is then followed by sections presenting some more relevant areas which can be of use in the analysis of many systems and in the solution of many problems in diverse areas of science and technology.

First, fuzzy optimization and fuzzy mathematical programming are outlined starting with basic concepts of optimality, feasibility, etc., under fuzzy information. Then, various classes of fuzzy optimization and fuzzy mathematical programming problem classes are briefly outlined and basic solution concepts and algorithms are presented. Some examples of applications are mentioned (see W. Lodwick and E. Untiedt, ► Fuzzy Optimization).

The topic of statistics with imprecise data is considered next, a very interesting topic from a conceptual point of view, and practically relevant. First, a sound motivation is presented by pointing out that, in reality, much available data is imprecise, and so they cannot be used by powerful and well founded, yet too "hard" traditional statistical tools and techniques. New concepts are formulated and extensions of well known statistical means under an imprecise information are presented. Some applications are outlined (see M.A. Gil and O. Hryniewicz, ► Statistics with Imprecise Data).

The next papers are concerned with the two other key elements of soft computing: neural computation and evolutionary computation. To emphasize their synergy and complementarity, these tools have been presented in a hybrid context, that is as a presentation of new fields in which the confluence of fuzzy, neural and evolutionary tools yields a new quality, and results in the emergence of a new class of tools and techniques which have already changed the landscape of soft computing.

We start with the neuro-fuzzy systems, which refer to a combination of (artificial) neural networks and fuzzy logic, and may be viewed as a main, fundamental step into the development of hybrid intelligent systems (see L. Rutkowski, K. Cpałka, R. Nowicki, A. Pokropiñska, R. Scherer, ► Neuro-fuzzy Systems). Basically, neuro-fuzzy systems try to combine transparent, human-like reasoning types of fuzzy rule-based systems with a parallel, connectionist type of neural networks. Therefore, they are universal approximators with an ability to yield interpretable IF-THEN rules, bridging the gap between accuracy and interpretability. A broad coverage of main architectures and implementations is presented.

Though neuro-fuzzy systems have been a considerable step forward towards in developing modeling tools of a better expressive power, effectiveness and efficiency, some natural extensions have been proposed that take advantage of the increasing popularity of evolutionary computation. It has been shown theoretically and experimentally that the incorporation in broadly perceived fuzzy systems, including neuron-fuzzy systems, of a learning mechanism which can make it possible to adjust the systems' structure and parameters may greatly enhance performance. The use of evolutionary computation has led in this context to the concept of an evolving fuzzy system that has been considered a promising alternative for some time (see P. Angelov, ► Evolving Fuzzy Systems).

The above mentioned hybridization techniques, which basically boil down to architectures and algorithms being a synergistic combination of fuzzy systems, notably fuzzy rule based systems, neural networks and evolutionary computation, in particular evolutionary learning, can be further extended when applied to real world problems which always provide much inspiration. Applications for systems modeling and control are of a universal importance for many fields of science and technology and hence

are dealt with in detail (see O. Castillo and P. Melin, ► Hybrid Soft Computing Models for Systems Modeling and Control).

The chapters briefly described above have been related to fuzzy sets and some of their hybridizations by including elements of neural networks and evolutionary computation to attain some "superadditivity" of the strengths and power of the particular single tools and techniques.

Rough sets theory, introduced by the late Zdzislaw Pawlak in the early 1980s, has been for some time considered a very promising tool for dealing with imperfect information, notably in data analysis and decision making.

Basically, a rough set is a formal approximation of a conventional crisp set in that it is represented as a pair of two traditional crisp sets which constitute the lower and the upper approximation of the original. These can be viewed as sets of elements that possibly and surely belong to the original set. Rough sets theory is presented in much detail, first in a more traditional, pure setting in Pawlak's spirit. Various concepts are presented, relations between them are formulated and proved, and their possible applications are outlined, notably in the relevant areas of data mining and knowledge discovery. Some extensions of the basic concept of a rough set are also presented (see J. Peters, A. Skowron and J. Stepaniuk, ► Rough Sets: Foundations and Perspectives).

Next, a "meta-problem" in science, that is decision making, is dealt with in the context of rough sets. In addition to the basic definition of a rough set, which provides much insight into the formalization and solution of various classes of decision making problems, a novel idea of a dominance-based rough set is presented. Its use to derive new, more human consistent solution concepts in decision making and data analysis is proposed (see R. Słowiński, S. Greco and B. Matarazzo, ► Rough Sets in Decision Making).

To summarize, we have tried to present the vast and diversified area of soft computing in a comprehensive, illustrative and constructive way. The basic perspective assumed may be viewed as presenting soft computing as a significant paradigm shift in computing in that it tries to exploit the fact that the human mind, unlike the computer we have now, has a remarkable ability to store and process information which is predominantly imprecise, uncertain and granular.

We have started with some basic fundamental concepts and properties, and then have devoted much space to the presentation of what is characteristic for soft computing: a synergistic, complementary use of various tools and techniques to solve problems in an effective and efficient way. This has led to various types of hybridization in which tools from fuzzy systems, rough sets, neural networks and evolutionary computation have been employed. It seems that soft computing should play a more and more important role in the years to come, in which more and more emphasis should be put on human-consistent and human-centric tools and techniques, and on hybrid systems.

Software Architectures for Autonomy

Robert A. Touchton[1], Carl D. Crane III[2]
[1] Honeywell International, Phoenix, USA
[2] University of Florida, Gainesville, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Joint Architecture for Unmanned Systems (JAUS)
NIST 4D/RCS
Service Oriented Architecture/Component Oriented Architecture
Distributed Architecture for Mobile Navigation
Situation Assessment
Planning and Decision Making
Knowledge Representation
Adaptive Planning Framework
Future Directions
Bibliography

Glossary

Reference architecture (RA) A reference architecture is a framework containing valuable implementation guidance for meeting the requirements of an enterprise.

Automated guided vehicle (AGV) A vehicle that can be programmed to automatically drive to designated points and perform preprogrammed functions.

JAUS The Joint Architecture for Unmanned Systems (JAUS) Reference Architecture defines a set of reusable components and their interfaces. The architecture emphasizes vehicle platform independence, mission isolation, computer hardware independence, and technology independence.

NIST 4D/RCS The acronym NIST 4D/RCS refers to the National Institute of Standards and Technology Real-time Control System. 4D/RCS describes in detail the functions and associated interfaces necessary to provide sensory processing, world modeling, knowledge

management, cost/benefit analysis, and behavior generation.

Adaptive planning framework The Adaptive Planning Framework addresses dynamic situation assessment, behavior management, and decision-making. It incorporates a three-stage process of 1) understanding the current situation, 2) understanding the suitability and viability of the available behaviors in light of that situation, and 3) providing the capability to autonomously make and execute behavior-related decisions, all in real-time.

Service oriented architecture (SOA) A Service Oriented Architecture maintains a strictly enforced standardized interface among entities. A standardized messaging construct enables one entity to request a service from another entity and for that service provider to send its response.

Definition of the Subject

Developing an autonomous system, such as an autonomous ground vehicle, that can operate and maneuver in an unstructured environment is a complicated task. One of the most daunting issues facing autonomous vehicle researchers is how to best exploit sensor and other information discovered during the execution of a plan. If the system takes too long to deliberate on the possible meanings and implications of this newfound data and knowledge, the vehicle may well have progressed beyond the point where it can benefit from it. Indeed, it may now be sitting atop the unforeseen obstacle that spawned the influx of new information that was being processed.

The execution of specific autonomous behaviors is becoming reasonably well understood, such as "waypoint-following with obstacle detection", though improvements and breakthroughs in these areas continue. However, the autonomous selection of which behavior(s) should be invoked, and in what sequence and by what method, is in need of movement in the state of the practice. Advanced ways of thinking about, organizing, and applying situational knowledge to macro-level planning and decision-making are needed by the autonomous robotics community in order to achieve the full potential of the field. This article discusses various approaches and various reference architectures that have been developed to address these problems.

Introduction

Three predominant standards are discussed in the following sections, i. e. the Joint Architecture for Unmanned Systems (JAUS) Reference Architecture, the National Institute of Standards and Technology (NIST) 4D/RCS (Real-time Control System), and the Service Oriented Architecture. A fourth architecture referred to as the Distributed Architecture for Mobile Navigation (DAMN) is also discussed although this architecture is not widely adopted. Lastly, the Adaptive Planning Framework approach that was developed at the University of Florida to address the problem of decision making is introduced.

Joint Architecture for Unmanned Systems (JAUS)

The Joint Architecture for Unmanned Systems (JAUS) Reference Architecture, Version 3.2 [18] defines a set of reusable components and their interfaces. In order to ensure that the architecture will be applicable to the entire domain of unmanned mobile systems, the following four characteristics have been considered by the JAUS Working Group in the creation of the Reference Architecture:

1. Vehicle platform independence. In order for JAUS components to be interoperable, no assumptions about the underlying vehicle or its means of propulsion are made.
2. Mission isolation. The JAUS components can typically be assembled such that a variety of missions can be supported.
3. Computer hardware independence. No assumption of or requirement of particular computer hardware is made. This allows for future adaptability and enhancement as new computer hardware becomes available.
4. Technology independence. This is similar to the computer hardware independence, but focuses more on the technical approach rather than the computer hardware. For example, many approaches could be used to determine vehicle position and orientation. No one approach, such as GPS, inertial dead reckoning, or landmark-based navigation for example, is specified.

As currently defined, JAUS Reference Architekture (RA) establishes a pre-defined set of standard, yet flexible, components that provide a menu of capabilities that can be drawn from to design an unmanned system. Components are divided into five categories:

- Command and control components
- Communications components
- Platform components
- Manipulator components
- Environmental sensor components

The RA also defines a standardized messaging construct (header and content) that enables JAUS components to exchange information in an efficient and robust

fashion. The messaging approach first defines the content and usage of a standardized JAUS Header. It then prescribes the legal JAUS data types that can be incorporated into a message. Then it defines each JAUS message.

The Adaptive Planning Framework Reference Implementation was developed within such components and using such messages as defined by JAUS. Specifically, the Reference Implementation for the Adaptive Planning Framework was cast in an operational JAUS-compliant vehicle. However, the concepts and ideas that make up the Adaptive Planning Framework are not tied to nor specifically depend on JAUS. This enables other organizations to implement the framework in an alternative architecture or in future evolutions of the JAUS Reference Architecture itself.

NIST 4D/RCS

The National Institute of Standards and Technology (NIST) has been working for over two decades on establishing a standardized approach to the intelligent control of unmanned vehicle systems. The most comprehensive summary of their approach is given in NISTIR 6910, 4D/RCS: A Reference Model Architecture [27]. The 4D/RCS architecture is itself derived from NIST RCS, a domain-independent architecture developed by NIST a decade plus earlier (see [26] for a good overview of the generic RCS methodology). 4D/RCS goes on to specialize RCS to the domain of intelligent vehicle systems for military use.

4D/RCS focuses on ways to ensure that military missions involving unmanned vehicles can be analyzed, decomposed, distributed, planned, and executed in an intelligent, effective, efficient, and coordinated fashion. 4D/RCS describes in detail the functions and associated interfaces necessary to provide sensory processing, world modeling, knowledge management, cost/benefit analysis, and behavior generation. Of particular interest is its hierarchical treatment of time, providing a temporally layered set of eight planning/execution regimes (see Fig. 1). For example, it suggests that a vehicle Subsystem Planner (Level 3) ought to execute at ∼1–5 Hz with a 5 second, 50 meter planning horizon at a 40 cm grid map resolution, while a Section Planner (Level 5) might need to re-plan every 50 seconds, with a 10 minute, 5 km planning horizon at a 40 m grid map resolution.

A challenge posed by 4D/RCS is that their hierarchy of nodes calls for each node to possess a complete set of functional capabilities (i. e., World Model, Value Judgment, Behavior Generation, etc.), scaled and scoped to its level of operation in the hierarchy. The partitioning, decomposition, and distribution of the Adaptive Planning Framework Specialists and Decision Brokers across a 4D/RCS hierarchy will be a completely new research area. Of greater concern is that 4D/RCS puts a great deal of power and functionality into their World Model, including prediction and simulation.

Service Oriented Architecture/ Component Oriented Architecture

Several facets of the Information Technology (IT) sector have been working to establish standards that support software interoperability across diverse organizations under the moniker of Service Oriented Architecture or SOA. SOA enables loose coupling among diverse software entities across a common network. This is accomplished by maintaining a strictly enforced standardized interface among the entities and a standardized messaging construct that enables one entity to request a service from another entity and for that service provider to send its response. This rapidly emerging standard is of interest here because the JAUS Working Group has begun a transition to a SOA-style architecture.

The most mature of these efforts is sponsored by the World Wide Web Consortium (W3C), which relies heavily upon SOA as the foundation of its Web Services initiative and, therefore, is leading the way in its maturation and adoption as a standard. Web Services extend the SOA concept to address anonymous entities that can discover one another and engage one another's services autonomously over the World Wide Web. They have published a treatise on the Web Services Architecture that includes an excellent overview of SOA in Sect. 3.1 of [45]. They go on in that section to outline some of the pitfalls of a SOA, such as network reliability and latency, lack of shared memory between service provider and consumer (i. e., everything that must be conveyed from one entity to another must be done explicitly via message content, and side effects of receipt of a message must be well understood and agreed upon), concurrency mismatches, and so on.

Industry has also taken a strong role in promoting SOA as a de facto standard. IBM (http://www-128.ibm.com/developerworks/webservices/standards/), Sun Microsystems (http://java.sun.com/developer/technicalArticles/WebServices/soa2/SOATerms.html#soaterms), and Microsoft (http://msdn.microsoft.com/architecture/soa/default.aspx?pull=/library/en-us/dnmaj/html/aj1soa.asp), to name three, have all embraced the notion.

Academia has also been active in this arena. IEEE Computer Society has formed a Technical Committee on Services Computing (http://tab.computer.org/tcsc/

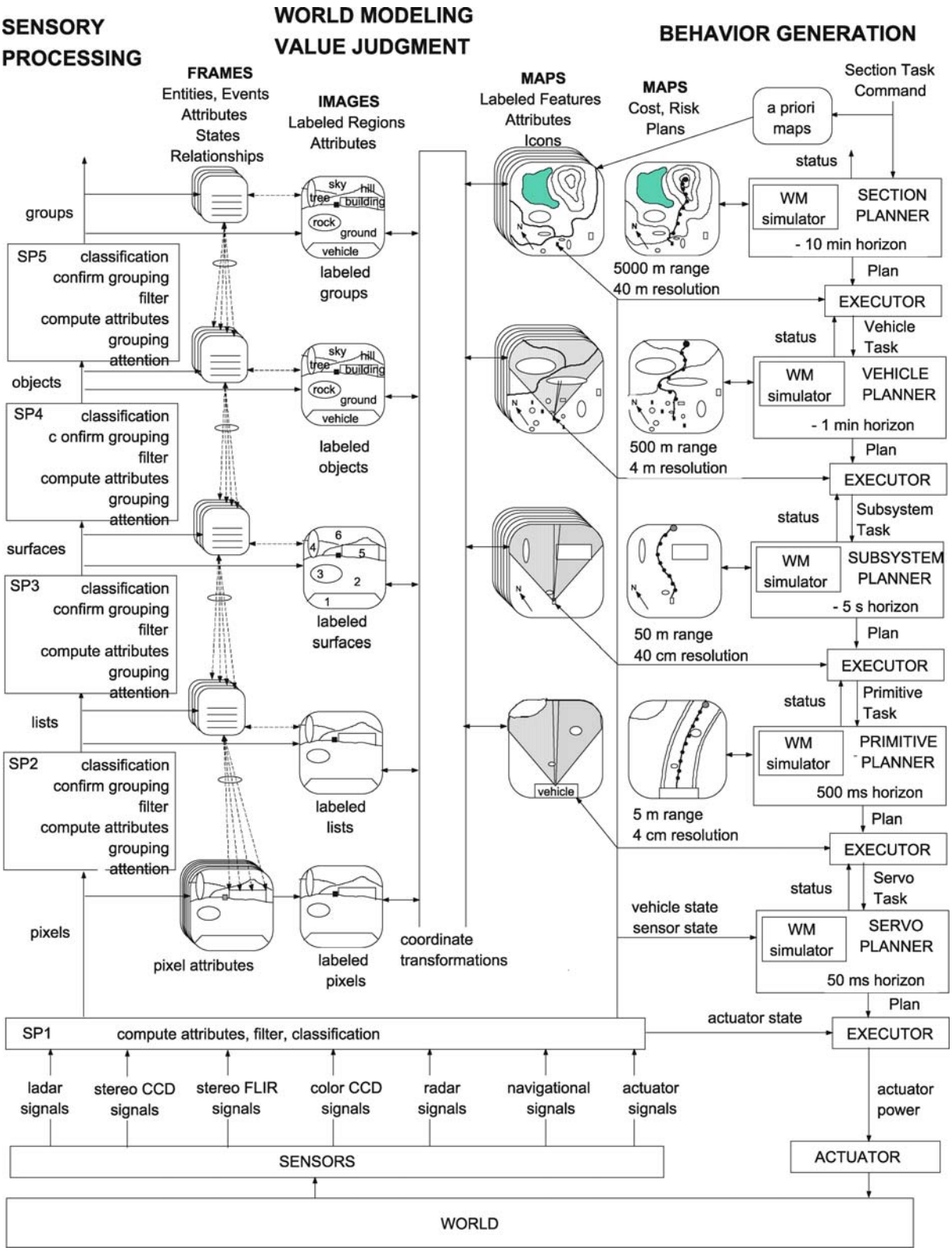

Software Architectures for Autonomy, Figure 1

Excerpt from NIST PowerPoint Presentation (Source: http://www.isd.mel.nist.gov/projects/rcs/presentationhui/sld019.htm)

link.htm), and ACM has been actively including SOA topics in many of their conferences and symposiums.

A closely related predecessor to SOA is component-based architecture (COA), which differs primarily in its stronger predisposition of what services a software entity ("component") will provide and less standardization of how components communicate with each other. In other words, COA does not worry so much about a component performing a single task (as in SOA) as long as the multiple services provided by a given component, and the interface for executing those services, are well documented. The emphasis is on providing a good platform for problem decomposition and loose coupling among components, with less emphasis on component interoperability. Aksit [1] provides an excellent compilation of articles on the topic of COA, especially Chap. 3, "Component–Based Architecting for Distributed Real-Time Systems", which, in turn, includes a detailed example of using a COA to devise a Car Navigation System (page 85). All in all, SOA can be considered a maturation, and perhaps specialization, of COA.

Distributed Architecture for Mobile Navigation

The Distributed Architecture for Mobile Navigation (DAMN) was originally published as a Ph.D. Dissertation [35] and, while not as widely adopted as the architectures discussed above, it has provided many useful insights. Even though the scope of DAMN is limited to navigation and obstacle avoidance, its distributed approach, its support of hybrid planning and implementation styles, its blend of centralized and decentralized processing, and its thoughtful treatment of salient challenges to real-time decision-making all make it worthy of elaboration here.

The basic premise behind DAMN is that centralized arbitration of distributed decision-making processes provides a reasonable and useful balance between the demands for real-time responsiveness and the challenges brought about by the asynchronous, latency-filled, heterogeneous, uncertain environment encountered by an autonomous ground vehicle. As in the other architectures discussed, DAMN provides a modular, extensible, and interoperable framework for supporting the generation and arbitration of sensor data, behaviors, and commands to the mobility platform, controllers, and actuators. This notion is shown schematically in Fig. 2, where sensor data and high-level commands have been bundled with the assortment of behaviors depicted.

The treatise goes on to present and analyze alternative structuring of the placement of arbitration (e. g., sensor vs. command vs. effect) and to explore various action selection schemes. A detailed presentation of the DAMN implementation on a CMU Navlab AGV and the experimental results achieved provides further insights into the merits and shortcomings of the architecture. Another major contribution of that research was the application of utility theory to the behavior arbitration process, as further discussed later in this article.

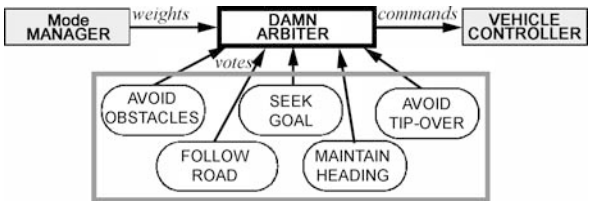

Software Architectures for Autonomy, Figure 2
DAMN Arbiter and Behaviors (Source: [35], page 9, Figure 1–2)

Situation Assessment

The situation assessment domain of interest to the topic of software architectures for autonomy is that of an unmanned system understanding its surroundings and status at a higher, more abstract level than that provided directly by its perception systems. In reviewing the literature, one must filter the use of the term when used in the context of the design of manned combat systems; such references often address such topics as own and enemy radars, missile tracking, and weapon lethality. Most such references are in the context of providing situational assessment for a human [16,48], such as pilot support on board a combat aircraft. Of interest here, however, is the applicability to unmanned systems, wherein the raw sensor and signal data is processed into more general situational conclusions, usually as a result of some form of inference or deduction. For clarity, the term "situation assessment" when used in this document will refer to this latter connotation. This domain is sometimes mentioned in the literature as "situational awareness" and could be referring to either of the connotations discussed above.

Work at USC [49] described the use of templates and patterns to provide situation assessment in virtual humans. They demonstrate a way to use situation assessment to improve decision-making by allowing the software system to better focus its attention (i. e., computing resources) with the goal of improved utilization of onboard resources.

Of particular interest is the work underway at NIST. They are working in several areas that address situation assessment. One has to do with incorporating situation assessment feedback to human operators of robotic devices [40]. While their emphasis is on the human-machine

interface, there are insights to be gained from the situation identification and classification schemes that they developed. An even more relevant front is their work on using 4D/RCS to control on-road robotic vehicles. There are both formal papers [38] and materials and presentations available on the NIST web site (see Fig. 3) that demonstrate ways to incorporate situation assessment notions into the 4D/RCS architecture.

Weiss, Philipps and To et al. [46] present a capability that could be adapted into a Traffic Specialist. It provides situation classification and prediction for an assortment of expressway-related conditions (such as same/different lane assignment for other vehicles that are detected), and states (Such as approach rate {approach | approach with distance warning | approach with collision warning}). Similarly, [13] introduce material that could form a Collision Avoidance Specialist that could manage interactions with moving obstacles using such notions as "time to collision", "time to brake", and "time to disappear". Another area of interest is vehicle self-awareness and work such as [33] sheds light on how a Vehicle Health Specialist might be devised.

Finally, it should be noted that much of the discussion of situation assessment in the literature was secondary to a broader discussion and is, thus, of most use in providing insights into possible nomenclature and classification. References such as these are discussed in the Knowledge Representation section rather than here.

Planning and Decision Making

Since the scope of this topic is so broad, its treatment here will be, first of all, limited to the domain of real-time planning and decision-making on an AGV and then further organized as an assortment of "views". The notion of planning refers to the orchestration of executable behaviors to achieve a goal (e. g., find a series of waypoints that will take the vehicle to a desired goal, then drive the vehicle to those waypoints while avoiding obstacles, obeying driving rules and maintaining stability), as well as the low-level planning conducted by a given behavior (e. g., finding an obstacle-free path towards the next waypoint within the perception horizon of the vehicle). The following list of behavior primitives are but a sampling of those gleaned from the literature:

- Seek Goal
- Avoid Obstacles
- Follow Road
- Respond to Blocked Path
- Explore
- Wander
- Maintain Stability
- Seek Target/Intruder
- Intercept Target
- Mark Location

Viewed as a Sense-Plan-Act Problem

This is perhaps the most fundamental view of autonomous control of a mobile robot and one into which many autonomous robotic implementations can be cast. The notion is to neutralize uncertainties in the robot's perception of its world, its understanding of its own state, and the effects of its own actions by indirectly "closing the loop" through the continuous gathering of feedback from its environment while executing its plan [25]. Since it is anticipated that the plan itself will be divided into a sequence of steps, the idea is that the results of executing the initial steps can be observed and compared with expected results. If expectations are not being adequately met (in essence, forming an "error" signal), then the subsequent steps can be adjusted accordingly, or an entirely new plan can be published. It is presumed that the robot will have an ability to store its perception and state knowledge in some form of a world model, which can, in turn, be used by the planner.

This design style best describes the autonomous control used on the NAVIGATOR [8]. The four environmental sensors publish their findings, in the form of a traversability grid, to a sensor arbiter. Two additional pseudo-sensors each publish a traversability grid to the sensor arbiter denoting the a priori route boundary and a priori path plan. The sensor arbiter then fuses these inputs and publishes to the receding horizon planner a comprehensive traversability grid, which represents a localized view of a world model. The receding horizon planner uses an A* search algorithm and a simple vehicle model to iteratively produce viable plans that achieve a desired goal state and to choose the one that minimizes traversability cost. The goal itself is based on the a priori path plan and is replaced with a new goal once the vehicle nears it. The planning search that occurs has as its only objective the publishing of an instantaneous wrench command (steering, throttle, brake) to the vehicle's primitive driver, whose job is to execute that wrench as actuator positions. Thus, every cycle of the planner produces a new wrench command. Since every component in the chain executes at a nominal rate of 20 Hz, a new "plan" (as manifested in the instantaneous wrench command) is always being issued, thus providing a responsive behavior, with some deliberation on how that behavior is generated. Figure 4 shows a snapshot of an arbitrated traversability grid and the instantaneous plan. Red, orange, and yellow indicate lessen-

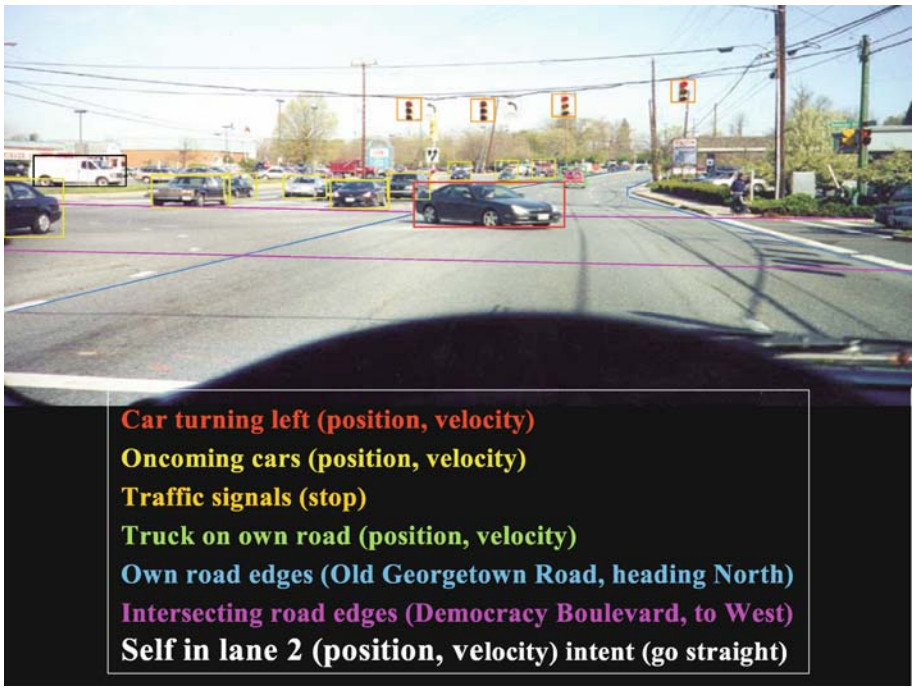

Software Architectures for Autonomy, Figure 3
Excerpt from NIST PowerPoint Presentation (Source: http://www.isd.mel.nist.gov/projects/rcs/presentationhui/sld061.htm)

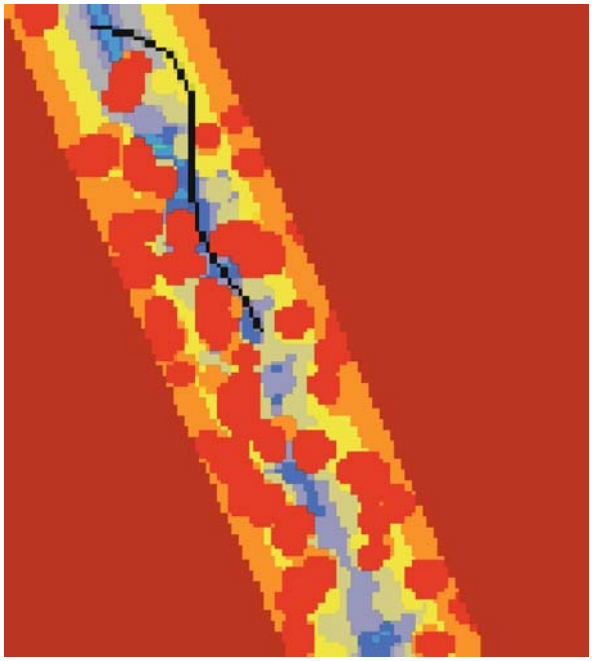

Software Architectures for Autonomy, Figure 4
Example Traversability Grid taken from the NAVIGATOR while in a Cluttered Roadway

ing severity of obstacles, gray and blue indicate improving degrees of smoothness of terrain. The instantaneous plan is indicated in black.

One difference in the NAVIGATOR's implementation of the Sense-Plan-Act paradigm is that, by encapsulating the a priori plan into a pseudo-sensor whose findings compete with those of the other sensors, the conventional aspects of planning provide only "suggestions" for a preferred action, rather than forcing the vehicle onto a defined course. Although implemented quite differently, this notion is in concert with the findings of Payton, Rosenblatt and Keirsey [29], who go on to note that "In general, internalized plans should be conceived as representations that allow the raw results of search in an abstract state space to be made available as advice to continuous real-time decision-making processes".

There are many good examples of robotic systems that have implemented some fashion of the Sense-Plan-Act paradigm. Most have to do with navigation and obstacle avoidance, such as Batavia and Nourbakhsh [6]. Examples of this paradigm applied to other aspects of robotic planning and decision-making are much harder to find.

Viewed as a Subsumption Problem

The notion of empowering a mobile robot to operate without any centralized control was first introduced by Rod-

ney Brooks as he devised a self-managing, layered control scheme dubbed the "subsumption architecture" [7]. By decomposing a robot's control system into layers of task-achieving behaviors, control is governed by the dominant layer in play at an instance in time, which, in turn, "subsumes" the behaviors of the lower level layers. For example, let "Wander" be considered a level 1 behavior and "Explore" a level 2 behavior. Since Explore is the higher level, it will self-determine whether exploring is an appropriate behavior under current circumstances. If so, then it will alter the Wander behavior to be not random, but to fulfill its wishes to visit new areas. If not, then it will allow the Wander behavior to proceed without any alteration. This notion is extrapolated across all possible behaviors. This style of planning and decision-making is often referred to as "reactive". The resultant behavior of the robot is referred to as "emergent" since it is likely that the observed behavior is some extemporaneous blend of the possible behaviors that the robot could execute.

This approach to planning and decision-making has a dedicated following and is especially appealing for multi-agent and swarm applications. For example, the subsumption architecture and reactive behavior play a major (though not exclusive) role in the design of robots at the Idaho National Lab (see www.inl.gov/adaptiverobotics).

The differences between these first two views can be captured by the relative importance placed on each of the three components of the Sense-Plan-Act view. For example, a purely reactive system does almost no local planning since every stimulus anticipated during the sensing stage has a prescribed behavioral action, thus relegating the planning stage to simply resolving action conflicts when more than one stimulus is perceived or queuing and dispensing actions when one stimulus invokes multiple actions (i. e., managing the subsumption process). Conversely, a deliberative system will have a large emphasis on the planning stage, attempting to formulate a new plan that incorporates newly sensed information along with any changes in state of the vehicle or its mission while simulating the effect of alternative actions on the quality and viability of the plan. The juxtaposition of the Sense-Plan-Act view's emphasis on deep planning through possibly time-consuming deliberation and the Subsumption view's potentially unpredictable, but fast, reaction to stimuli, explains why researchers are still seeking other, hybrid or blended, planning and decision-making styles.

Viewed as a Decision Theory Problem

Another rich area of exploration is how classical decision theory might be applied to the AGV domain. For example, Karacapilidis and Papadias [20] describe how argumentation can be automated and used to support collaborative decision-making. Their ideas for automating argumentation constructs include evaluations such as "Scintilla of Evidence", "Beyond Reasonable Doubt", and "Preponderance of Evidence".

Rauenbusch and Grosz [32] and others speak of devising explicit "Plan Trees" whose nodes encapsulate the desired action/behavior, associated constraints, and contextual applicability and whose structure models the desired decision-making outcomes. The search through the tree is conducted using some measure of cost or value such that the correct path through the tree delivers the correct series of actions/behaviors.

Hoffman and Yates [15] present a synopsis of what has become known as the "three-step decision-making process". In this paper, they report that most, if not all decisions can be modeled as a cascading set of three-step activities. One of the models specifically referenced for use in process control is 1) Situation Assessment, 2) Planning, and 3) Commitment to a course of action. Each of these steps may be expanded into another three-step decision-making process, such as deciding which situational conditions are present or relevant, or whether to keep or abandon a committed action.

A final realm under decision theory is known as Hierarchical Task Network (HTN) planning [9] provide an overview of this concept and cite the seminal works that have contributed to it on the way to introducing a formalism of HTN planning semantics. The basic premise of HTN planning is to iteratively decompose tasks until primitive tasks are reached (defined as tasks that cannot be further decomposed and that are actionable). These primitive tasks are assembled into a network of increasingly abstract tasks allowing a planning algorithm to select a high-level task, recursively expand its children until its primitive tasks are reached. Some expansions may be constrained based on the current situation, thus pruning the search when compared with an unconstrained expansion of the network. Each reachable path from the high-level tasks to the primitive tasks becomes a candidate plan. While this exploration of the HTN is taking place, the candidates are being evaluated by so-called "critics" so that any arising conflicts can be identified and the winning candidate declared. Because of its deep reasoning, HTN-based planning is not typically used for real-time applications.

Viewed as a Behavior Arbitration Problem

The concept of Behavior Arbitration was introduced as part of the Distributed Architecture for Mobile Navigation

(DAMN) [35] as a key ingredient for achieving its goal of balancing centralized and decentralized design styles. All (decentralized) behavior generators submit their control output (referred to as a "vote") and the (centralized) DAMN Arbiter fuses their votes into a single command set to the vehicle.

"Utility fusion", which uses traditional utility theory to provide an alternative to command fusion, is another concept that evolved from DAMN [36]. This notion requires each behavior generator to submit a probabilistic utility estimate along with its vote, thus enabling a "utility arbiter" to compute the Maximum Expected Utility and use it to select the optimal behavior.

Viewed as an Action Selection Problem

Action Selection is another way to view planning and decision-making on an AGV. In this view, the mobile robot is tasked with selecting the most appropriate action based on the current situation, in spite of inaccurate, incomplete, and possibly unforeseen information. For this to happen, there must exist some catalog or repository of possible actions from which to select and the criteria upon which to base a selection decision. Pirjanian [30] provides an excellent overview of ten varying approaches to the action selection problem. In this treatise, he summarizes each (including DAMN), then compares and contrasts them in terms of eight criteria, including planning vs. reactivity, synchronous vs. asynchronous, hierarchy vs. no hierarchy, and knowledge representation which all have a direct bearing on the current research.

NIST has also developed an approach to action selection via its hierarchical planning and control scheme [21, 22]. The scheme enables the system to plan at different rates at each level, with the scope of planning fixed for each level. For example, high-level goal planning might take place at a lower resolution and update rate, but would cover a larger expanse than say planning for obstacle avoidance. The plans themselves are broken into a tree or graph of subgoals and subtasks (task decomposition itself is discussed under Knowledge Representation) and the actions are selected, executed, and monitored in accordance with the defined planning levels. The planning levels are chosen to be consistent with the time, duty cycle, and range horizon parameters established in the 4D/RCS architecture. For example, AGV mobility planning is broken into four levels: Servo, Prim(itive), Autonomous Mobility, and Vehicle System. Balakirsky and Lacaze [2] elaborate how planning, in the form of Value Judgment and Behavior Generation, takes place for the Vehicle System planning level.

Viewed as an Adaptive Planning Problem

Note that in some literature, "adaptive" is used to mean that the system "learns" from its experience, thus improving its performance over time, whereas the connotation used here is that the system alters its plan based on new, situational information that has been provided by upstream knowledge and data processing. Thus, while the possibility of actually changing the a priori behaviors from which to choose through learning should not be ruled out for future generations of the Adaptive Planning Framework, it is certainly not the emphasis or the motivation for using the term "adaptive" in its moniker. The genesis of adaptive planning as used here was a search to improve the performance of (manual) military mission planning through the use of expert systems, such as the Adaptive Mission Planning System in [41]. The quest continues as military planners seek to reduce 24-month planning cycles down to a year or less for complex deployments and even less for Crisis Action Planning [14]. In fact, their definition, "Adaptive Planning is the systematic, on-demand creation and revision of executable plans, with up-to-date options, as circumstances require", could suffice for the work conducted here as long as its transition to an autonomous, real-time setting is understood.

The need to alter a plan already in progress can have a number of causes, including insufficient time for completion, ineffective results, changes in the situation, and receipt of a new objective to name a few. The Artificial Intelligence community has driven related work in this area, but application to mobile autonomous robotics has not been at the forefront. For example, [12] presents an excellent treatise of an adaptive planning architecture based on the premise that an "agent dynamically constructs explicit control plans to guide its choices among situation-triggered behaviors". To accomplish this, she identified and explored five areas where an intelligent system might require adaptive behavior, depending on the situation encountered:

1. Perception Strategy – Adapt to information requirements and resource limitations
2. Control Mode – Adapt to goal-based constraints and environmental uncertainty
3. Reasoning Tasks – Adapt to perceived and inferred conditions
4. Reasoning Methods – Adapt to available information and current performance criteria
5. Meta-Control Strategy – Adapt to dynamic configurations of demands and opportunities

As an example of more recent work that does focus on mobile robotics in a real-time setting, Hassan, Simo and Crespo [11] offer a behavior-based architecture that will adapt to temporal constraints by allowing itself to utilize more deliberative techniques when time is available, but moving towards more reactive behaviors when time is at a premium. They also introduce the notion of adjusting the quality of service that a given element might deliver based on the situation encountered. For example, this approach might allow the system to attempt to achieve its goal with a "rough" plan if a "complete" plan could not be delivered in a timely enough manner. Musliner [23] and his Adaptive Mission Planner, provides another view on how to empower an autonomous system to alter its plans based on temporal constraints and in light of changing environments, objectives, and system capabilities. That work built upon his earlier efforts to devise the Cooperative Intelligent Real-time Control Architecture (CIRCA) [24], which provides formalisms on how to represent tasks and decisions in a LISP setting. While CIRCA has not been applied in the mobile robotic domain (making its suitability to support an AGV unknown), there are insights to be gained from this work. Finally, NIST has incorporated an element of adaptive planning in their recent work on autonomous on-road driving as part of 4D/RCS. For example, Balakirsky and Scrapper [3] discuss an expert system and knowledge representation scheme that support adaptive planning for autonomous lane and speed management.

Knowledge Representation

In this section, related work on Knowledge Representation relevant to the domain of AGVs is explored. Knowledge Representation refers to the schemas and constructs used to document, standardize, normalize, and utilize the entities within the domain of interest. It must capture the semantics and meanings of the relationships among the entities, as well as their names, descriptions, attributes, and the method or reasoning mechanism for determining their current state or value.

Sources of such domain knowledge include technical documents, specifications, training manuals, etc. (many of which can be accessed via the web). Example knowledge sources include a table of Autonomous Mobility Situation Coverage Requirements from Demo III requirements analysis [34], a Functional Taxonomy chart for an AGV from a TACOM (the U. S. Army's Tank-Automotive COMmand) PowerPoint presentation [31], and by drawing analogies from human military operations as found in the Army Universal Task List [44]. Remaining knowledge gaps must be filled in by interviews of subject matter experts or perhaps empirically through experimentation.

By far, the most work in knowledge representation for intelligent vehicles has been done by NIST. Thus, this section will conclude with an extended example, demonstrating their approach to representing knowledge about situational conditions, states and events, planning, and behaviors within the 4D/RCS context.

Lexicons, Taxonomies, and Ontologies

One technique for knowledge representation is to progress from a lexicon (a domain-specific dictionary of terms), to a taxonomy (a logical ordering and categorization of those terms), to an ontology (an explicit specification of those terms along with the semantics and relationships among them). One on-line dictionary defines ontology as follows:

> An explicit formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships that hold among them … Definitions associate the names of entities in the universe of discourse (e. g. classes, relations, functions or other objects) with human-readable text describing what the names mean and formal axioms that constrain the interpretation and well-formed use of these terms … The hierarchical structuring of knowledge about things by subcategorizing them according to their essential (or at least relevant and/or cognitive) qualities [17].

Much work is in progress attempting to build general purpose, or even "common sense" ontologies that would be useful to all domains. The most famous of these is the OpenCyc project (http://www.opencyc.org/), which, with 47,000 concepts and 306,000 assertions about them to date, is well on its way to achieving its vision to become "the world's largest and most complete general knowledge base and commonsense reasoning engine". Another initiative of interest is the DARPA Agent Markup Language (DAML) project (http://www.daml.org/), which was sponsored by DARPA to create an xml extension that provides (among other things) a rich and suitable language for the creation of general-purpose ontologies (282 distinct ontologies had been created using DAML by the time the program funding was terminated in 2006 and the work absorbed by W3C).

For the AGV domain, the scope of the knowledge that must be represented via the techniques discussed in this subsection is still quite broad. Situational knowledge spans from urban to highway to off-road environments,

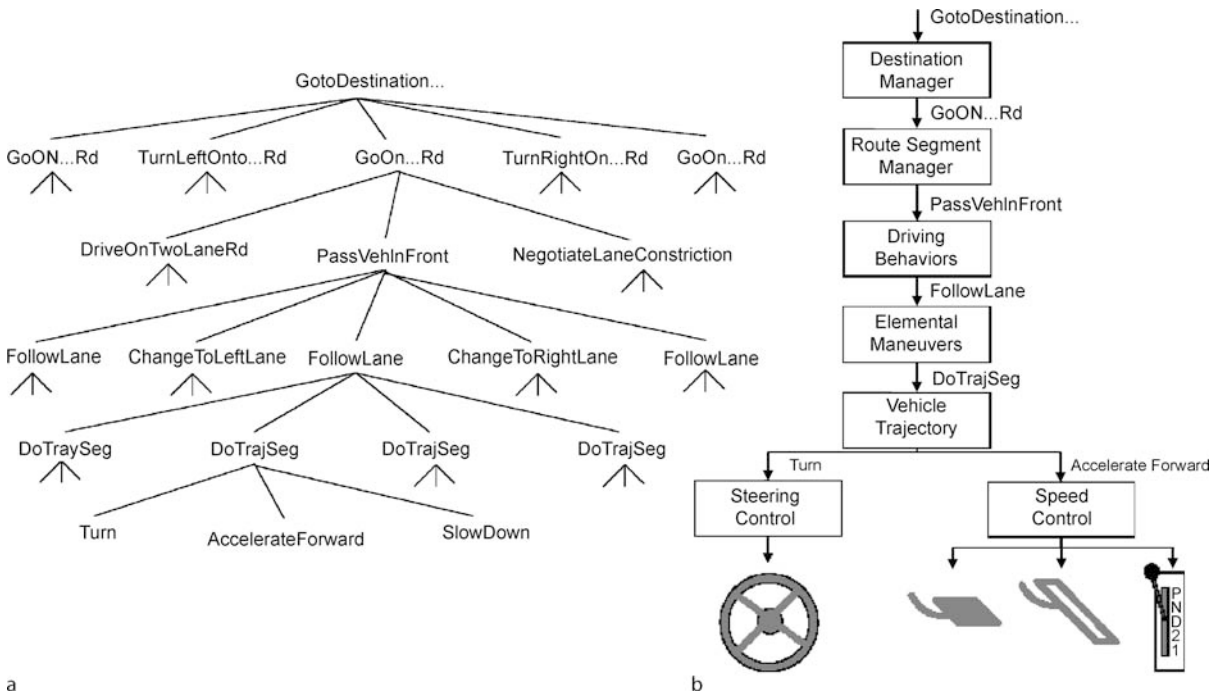

Software Architectures for Autonomy, Figure 5
NIST Knowledge representation schemes for on-road driving. (Source: [4,5], Figures 3 and 4)

potential obstacles and hazards that might be present, traffic rules and driving best practices, and so on. Planning and behavior knowledge encompasses a wide variety of missions and tasks, whether they are high-level (conduct search and rescue operation), tactical (pass the vehicle), elementary (change lane), behaviors (avoid obstacles, maintain stability), planning rules and processes, and so on. Even knowledge about "self" or "ego" must be represented, such as capabilities, limitations and constraints, or current status. There is a major initiative under way at NIST, sponsored by TACOM, to develop an Intelligent Systems Ontology. Although still a work in progress, this intelligent-vehicle-specific ontology is expected to provide a standard set of domain concepts, their attributes and their interrelationships, delivered in a fashion that facilitates knowledge capture and reuse [39]. This ontology is beginning to gain traction as it makes its way into the AGV navigation planning community outside of NIST [37].

World Model Knowledge Store (WMKS)

Another dimension of knowledge representation is how data, information, and knowledge are stored. Whether it is provided a priori, or it is perceived, inferred, or received by the AGV, there must be a place and a format for storing, accessing, and analyzing it. Such data, information, and knowledge are often referred to as the "world model" and the place where they are stored as the "knowledge store". The breadth and sophistication of the world model knowledge store for a given AGV design will vary widely, depending on its degree of autonomy, the scope of its behavior, the complexity of its design, etc.

Situation Assessment findings, which must also be managed, fit into what some communities refer to as "meta-knowledge", i. e., knowledge about the knowledge. For example, while pumping out its perception data, a sensor could independently assess and report on its own confidence in its findings and its own health, and perhaps even declare that its own results should not be used right now (say, due to a camera white-out).

Although not always so, the knowledge store is usually persistent, using either a relational database or an object-oriented knowledge based system. Since much of the information stored is of a geo-spatial nature, the knowledge store often includes geo-spatial extensions for explicitly representing GIS and topographical data, polygonal objects, etc. Another consideration is whether the WMKS contents are stored in a central location, accessible by all AGV modules (sometimes referred to as a "blackboard architecture") or each module maintains a subset of the WMKS containing just the content it needs, with data, information, and knowledge marshaled among the AGV

modules on an as-needed/as-requested basis (sometimes referred to as a "publish/subscribe architecture").

Knowledge Representation at NIST

NIST advocates task decomposition as a key knowledge representation technique to support the hierarchical control strategy emphasized in its 4D/RCS architecture and has published widely on various ways to accomplish it [4,5]. This technique for representing the actionable elements that could be assembled to create a plan strives to break high-level tasks (e. g., a mission objective) into distinct hierarchical levels and also to identify multiple subtasks at a given level. Figure 5a–5b shows an example of how the "GoToDestination" task is decomposed into a "planning graph" that ultimately leads to a specific wrench command to the vehicle. The system must know (or be able to infer) the state of each node in the tree along with the cost of each arc in order for the associated control module to formulate the appropriate plan. Extending the example in Fig. 2–5b, a Destination Manager has determined that staying on the current road is appropriate and a Route Segment Manager has decided that passing the vehicle in front of it is the most desirable way to reach the destination. A Driving Behaviors module knows that its own vehicle has already changed into the passing lane and has further determined that the best thing to do right now is to stay in that lane, while a low-level Elemental Maneuvers module has found a wrench that ought to produce the requested outcome. Each Manager or module manages its own situational understanding either from direct sensory input or from its own local subset of the World Model Knowledge Store. Naturally, there are other tree elements and control modules that address following distance, speed, and so on, in addition to non-mobility-related tasks, such as payload management, communications, etc.

Once a plan is devised and approved, its elements must be executed by invoking one or more actions or behaviors, or perhaps by unleashing an entire subsystem to take over low-level control of the vehicle. NIST advocates the use of State Tables to represent the action decision-making knowledge [4]. A State Table is crafted for each node in the Task Decomposition Tree containing the rules that the control module is to use for mapping node inputs (states or situations) to allowable output actions.

To trigger the appropriate and desired state response, the matching situation must be known. The NIST approach to this is to determine and store the cascading precursor situational knowledge as a collection of "world states", but, in conformance to the 4D/RCS architecture, only that subset relevant to a given module. The lane-changing example concludes with a glimpse of the dozens of situational findings that lead up to the finding of interest ("ConditionsGoodToPass").

Adaptive Planning Framework

The Adaptive Planning Framework was developed by the authors at the University of Florida in order to address the requirements of the DARPA Urban Challenge that was held in November 2007. Figure 6 shows the University of Florida Navigator vehicle on which the Adaptive Planning Framework was implemented for the DARPA Urban Challenge. In the Adaptive Planning Framework, the system is assumed to be able to operate in a finite number of behavior modes. These behavior modes govern how the vehicle operates under various driving conditions. The framework is predominantly used to make intelligent decisions pertaining to these behaviors. The framework is scalable to systems of varying complexity and size and is compatible with existing architectures such as JAUS RA-3.2, NIST 4D/RCS, and others. The Adaptive Planning Framework is composed of three principle elements tasked with assessing the situation, determining the suitability and viability of all possible solutions, and executing the most suitable of all recommended solutions.

Situation Assessment Specialist

Dynamic environment information, originating from any array of sensors is monitored and managed by the Situation Assessment Specialists. Each specialist design is tailored to the sensor or collection of sensors whose data it analyzes. These specialists can, but are not required to, "live" on the same computing node that directly receives the sensor input. While the inputs to the specialist can come from any data source, the output or "finding" must adhere to specific guidelines outlined by the framework. Findings can be in the form of conditions, state, or events. A condition may have a value of present or absent only. All conditions are by default absent and must be proven present at each iteration. A finding classified as a state can only exhibit one of many a priori states. The event category is reserved for findings whose occurrence at some point in time is of significance even after the initial finding has passed. Once the findings have been generated the information is disseminated to all other components that might need it.

An example of a situation assessment specialist would be a software component whose sole function was to determine if it is safe to move to the adjacent lane. This component would monitor sensor data and reach a Boolean

Software Architectures for Autonomy, Figure 6
Team Gator Nation NaviGATOR

conclusion that would be stored as metadata for use by other processes. A second example would be a software component whose sole function was to determine if it is 'legal' to move to an adjacent lane. Here 'legal' is defined as not crossing a yellow line or not changing lanes when approaching an intersection.

Behavior Specialist

The findings rendered by the Situation Assessment Specialists are consumed by the behavior specialists. There is a one-to-one mapping of each behavior with a behavior specialist. The role of the specialist is to monitor the findings and evaluate the suitability of its behavior under the current perceived operating conditions. As with the specialist findings, the default recommendation is unsuitable and must be proven appropriate at every iteration of the program to ensure truth of the results and operating safety. A specialist does not possess the ability to activate or deactivate its associated behavior; such authority is only given to the Decision Broker.

For the DARPA Urban Challenge problem, the vehicle was programmed with six behavior modes. The corresponding behavior specialist constantly evaluates the appropriateness of its behavior. The six behavior modes were:

1. Roadway Navigation. The Roadway Navigation behavior is the primary driving behavior deriving commands to be sent to the vehicle actuators while the objective is lane following. This behavior allows the vehicle to navigate the roadway within the lines of its desired lane and maintain a safe following distance behind any vehicles ahead.
2. Change Lane Maneuver. The change lane maneuver is used in passing situations or in cases where the vehicle must change lanes in a multi-lane road in order to pass through a mission goal point. The behavior constrains the vehicle to remain within the lane boundaries of the new lane.
3. Reverse Direction. This behavior is called whenever it is determined that the current lane is blocked and there is no alternate clear lane available for passing. It is also applicable in cases where the vehicle has entered a 'dead end' road that it must "escape" to reach a mission goal point.
4. Intersection Traversal. The intersection traversal behavior is applicable when the vehicle enters the vicinity of an intersection. This is one of the most complicated behavior modes in that the system must rely on a series of situation assessment specialists to safely navigate the intersection. This behavior mode must handle queuing, stopping at the stop line, determining right of way, and ultimately traveling through the intersection while avoiding other vehicles.
5. Open Area Navigation. Open area navigation is a behavior that is only needed in special circumstances. This behavior allows the vehicle to move towards a goal location without striking any object, while avoiding any

rough terrain. This is in effect the only behavior mode that was required in the 2005 DARPA Urban Challenge. It is useful in the Urban Challenge when the vehicle is in an open area such as a parking lot prior to performing an actual parking maneuver.

6. Parking Lot. This behavior must deal with the problems that arise in the parking lot scenario where precise motion is necessary. When the vehicle approaches the vicinity of an assigned parking space, precise path planning will be initiated to align the vehicle as required. Situation assessment specialists will be monitoring the near surroundings of the vehicle to center the vehicle in its parking space while avoiding any static or dynamic objects.

Decision Broker

At the highest level of the framework lies the Decision Broker. Its role is to monitor all Behavior Specialist recommendations. It assumes ultimate authority over how the vehicle will operate while in autonomous mode. Like the other entities within the framework, the Decision Broker can base its conclusions on not only the recommendations and findings of other specialists, but it may also look at data from any other pertinent source. The author's implementation of the Adaptive Planning Framework centralized all the Decision Broker functionality within the JAUS Subsystem Commander component. The framework architecture employs an asynchronous, iterative, forward chaining reasoning approach to decision making.

Future Directions

JAUS is migrating to the Society of Automotive Engineering (SAE) Standard AS-4, Unmanned Systems. The transition to the SAE offered opportunity to further advance the standard through the introduction of more formalized specifications and methodologies.

The AS-4A Architecture Framework subcommittee has prepared document AIR 5665, Architecture Framework for Unmanned Systems. This document describes the concepts, capabilities and interoperability concerns of unmanned systems, and lays the foundation for all subsequent Aerospace Standards to be released by AS-4.

The AS-4B Network Environment subcommittee is preparing the document AIR5645, "JAUS Transport Considerations". This document examines aspects of JAUS, aspects of communications media, and aspects of the domain of unmanned systems, explaining how these affect the design of transport mechanisms for JAUS messaging. The committee has also prepared an initial draft version of AS5669, "JAUS Transport Specification".

The AS-4C Information Modeling and Definition subcommittee has prepared a working draft of the document AS5684, JAUS Service Interface Definition Language. This specification allows for the unambiguous definition of JAUS services. AS5684 is specified using Relax NG Compact notation, a machine-readable language, allowing for the creation of various tools to aid in the unmanned system design process. The AS-4C Subcommittee has also begun work on AS5710, JAUS Service Set. AS5710 is a collection of standard unmanned system services. Each service is defined using the JAUS Service Interface Definition Language. Current work on this document includes defining a discovery service and transport service, and transitioning the JAUS Reference Architecture 3.2 components to services.

Bibliography

1. Aksit M (2002) Software architectures and component technology. Kluwer, Boston
2. Balakirsky S, Lacaze A (2000) World modeling and behavior generation for autonomous ground vehicle. In: Proceedings of the 2000 IEEE International Conference on Robotics & Automation. San Francisco, pp 1201–1206
3. Balakirsky S, Scrapper C (2004) Knowledge representation and planning for on-road driving. Robotics Auton Syst 49(2004):57–66
4. Barbera A, Albus J, Messina E, Schlenoff C, Horst J (2004) How task analysis can be used to derive and organize the knowledge for the control of autonomous vehicles. In: Proceedings of the 2004 AAAI Spring Symposium Series on Knowledge Representation and Ontology for Autonomous Systems. Palo Alto, pp 67–78
5. Barbera A, Messina E, Huang H-M, Schlenoff C, Balakirsky S (2004) Software engineering for intelligent control systems. Künstliche Intell 3(4):22–26
6. Batavia PH, Nourbakhsh I (2000) Path planning for the cye personal robot. In: Proceedings of International Conference on Intelligent Robots and Systems, 2000 (IROS 2000). Takamatsu, Japan, 15–20
7. Brooks RA (1986) A robust layered control system for a mobile robot. IEEE J Robotics Autom 2(1):14–23
8. Crane CD, Armstrong DG, Touchton R, Galluzzo T, Solanki S, Lee J, Kent D, Ahmed M, Montane R, Ridgeway S, Velat S, Garcia G, Griffis M, Gray S, Washburn J, Routson G (2006) Team CIMAR's NaviGATOR: An unmanned ground vehicle for application to the 2005 DARPA grand challenge. J Field Robotics 23(8):599–623
9. Erol K, Hendler J, Nau D (1994) Semantics for hierarchical task-network planning. CS-TR-3239, UMIACS-TR-94-31, ISR-TR-95-9, University of Maryland, Institute for Advanced Computer Studies College Park, MD
10. Franklin S, Graesser A (1996) Is it an agent, or just a program?: A taxonomy for autonomous agents. In: Third International

Workshop on Agent Theories, Architectures, and Languages. Budapest, pp 21–35
11. Hassan H, Simo J, Crespo A (2001) Flexible real-time mobile robotic architecture based on behavioural models. Eng Appl Artif Intell 14(5):685–702
12. Hayes-Roth B (1995) An architecture for adaptive intelligent systems. Artif Intell 72:329–365
13. Hillenbrand J, Kroschel K, Schmid V (2005) Situation assessment algorithm for a collision prevention assistant. In: Intelligent Vehicle Symposium, 2005. Las Vegas, pp 459–465
14. Hoffman H (2004) Adaptive planning overview. Presented at Military Operations Research Society, Capabilities-Based Planning: The Road Ahead. The Military Operations Research Society (http://www.mors.org/), Alexandria
15. Hoffman RR, Yates JF (2005) Decision(?)Making(?). IEEE Intell Syst 20(4):76–83
16. Howard C, Stumptner M (2005) Probabilistic reasoning techniques for situation assessments. In: Third International Conference on Information Technology and Applications, 2005 (ICITA '05). Sydney, pp 383–386
17. Howe D (2005) The Free Online Dictionary of Computing. http://foldoc.org/foldoc.cgi?query=ontology. Accessed Oct 2006
18. JAUS (2005) JAUS Reference Architecture, version 3.2. JAUS Working Group, (http://www.jauswg.org/)
19. JAUS-OPC (2005) OPC 2.75 Interface Control Document, version 1.0, Payload Interface section. JAUS Working Group – OCU and Payloads Committee
20. Karacapilidis N, Papadias D (2001) Computer supported argumentation and collaborative decision making: the HERMES system. Inf Syst 26(4):259–277
21. Lacaze A (2002) Hierarchical planning algorithms. Presented at SPIE 16th Annual International Symposium on Aerospace/Defense Sensing, Simulation and Controls, Orlando
22. Murphy K, Abrams M, Balakirsky S, Coombs D, Hong T, Legowik S, Chang T, Lacaze A (2000) Intelligent control for unmanned vehicles. Presented at World Automation Conference (WAC 2000), Maui
23. Musliner DJ (2001) Imposing real-time constraints on self-adaptive controller synthesis. Lect Notes Comput Sci 1936:143–160
24. Musliner DJ, Durfee EH, Shin KG (1995) World modeling for the dynamic construction of real-time control plans. Artif Intell 74(1):83–127
25. Nilsson NJ (1998) Artificial intelligence: a new synthesis. Morgan Kaufmann, San Francisco
26. NIST (National Institute of Standards and Technology) (1992) A real-time control system methodology for developing intelligent control systems. NISTIR 4936. National Institute of Standards and Technology, Gaithersburg
27. NIST (National Institute of Standards and Technology) (2002) 4D/RCS: A Reference Model Architecture for Unmanned Vehicle System, Version 2.0. NISTIR 6910. National Institute of Standards and Technology, Gaithersburg
28. Panzarasa P, Jennings NR, Norman TJ (2002) Formalising collaborative decision-making and practical reasoning in multi-agent systems. J Logic Comput 12(1):55–117
29. Payton DW, Rosenblatt JK, Keirsey DM (1990) Plan guided reaction. IEEE Trans Syst, Man, Cybernetics 20(6):1370–1382
30. Pirjanian P (1997) An Overview of System Architecture for Action Selection in Mobile Robotics. Laboratory of Image Analysis, Aalborg University, Aalborg
31. Pritchett W (2002) A domain framework for intelligent ground combat vehicle systems. Presented at Software Technology Conference (STC 2002). TACOM (Tank-automotive & Armaments COMmand), Salt Lake City
32. Rauenbusch TW, Grosz BJ (2003) A decision making procedure for collaborative planning. Presented at AAMAS'03 – Second International Joint Conference on Autonomous Agents and Multi-Agent Systems, Melbourne
33. Reichard KM, Crow EC (2005) Intelligent self-situational awareness for unmanned and robotic platforms. AUVSI Unmanned Systems Conference, Baltimore
34. Robotic Systems Technology (1998) Demo III experiemental unmanned vehicle (XUV) program autonomous mobility requirements analysis. prepared for TACOM
35. Rosenblatt JK (1997) DAMN: A distributed architecture for mobile navigation. Dissertation, Carnegie Mellon
36. Rosenblatt JK (2000) Optimal selection of uncertain actions by maximizing expected utility. Auton Robots 9(1):17–25
37. Schlenoff C, Balakirsky S, Uschold M, Provine R, Smith S (2003) Using ontologies to aid navigation planning in autonomous vehicles. Knowl Eng Rev 18(3):243–255
38. Schlenoff C, Madhavan R, Barbera T (2004) A hierarchical, multi-resolutional moving object prediction approach for autonomous on-road driving. 2004 ICRA Conference, New Orleans, pp 1956–1961
39. Schlenoff C, Washington R, Barbera T, Manteuffel C (2005) A standard intelligent system ontology. Proceedings of SPIE – Unmanned Ground Vehicle Technology VII, Orlando, pp 46–56
40. Scholtz J, Antonishek B, Young J (2004) Evaluation of a human-robot interface: Development of a situational awareness methodology. National Institute of Standards and Technology, last accessed Oct 2006
41. Seares CDF (1987) Adaptive mission planning: Squeezing out greater combat capability. Air University Review, http://www.airpower.maxwell.af.mil/airchronicles/aureview/1987/seares2.html. Accessed Oct 2006
42. Touchton RA (1988) Reactor emergency action level monitor. In: Majumdar MC, Majumdar D, Sackett JI (eds) Artificial Intelligence and Other Innovative Computer Applications in the Nuclear Industry. Plenum, New York, pp 189–197
43. Touchton RA, Gunter AD, Wilson KM, Eldredge II TD, Weaver ME (1988) Reactor emergency action level monitor vol 1: REALM Technical Report. NP-5719. Electric Power Research Institute, Palo Alto
44. US Army (2003) Army Universal Task List. FM 7-15, Washington
45. W3C (2004) Web Services Architecture. World Wide Web Consortium, http://www.w3.org/TR/2004/NOTE-ws-arch-20040211/. Accessed Oct 2006
46. Weiss K, Philipps H, To TB, Kirchner A (2005) Environmental perception and situation assessment for an advanced highway assistant. In: 2005 IEEE Intelligent Vehicles Symposium Proceedings. Las Vegas, pp 472–477
47. Winston P, Horn B (1989) LISP: Third Edition. Addison-Wesley, Reading
48. Yanco HA, Drury J (2004) Where am I? Acquiring situation awareness using a remote robot platform. In: 2004 IEEE International Conference on Systems, Man and Cybernetics, The Hague, pp 2835–2840

49. Zhang W, Hill RW (2000) A template-based and pattern-driven approach to situation awareness and assessment in virtual humans. In: Fourth International Conference on Autonomous Agents, Barcelona, pp 116–123

Soliton Perturbation

Ji-Huan He
Modern Textile Institute, Donghua University, Shanghai, China

Article Outline

Glossary
Definition of the Subject
Introduction
Methods for Soliton Solutions
Future Directions
Bibliography

Glossary

Soliton A soliton is a nonlinear pulse-like wave that can exist in some nonlinear systems. The isolated wave can propagate without dispersing its energy over a large region of space; collision of two solitons leads to unchanged forms, solitons also exhibit particlelike properties.

Soliton perturbation theory The soliton perturbation theory is used to study the solitons that are governed by the various nonlinear equations in presence of the perturbation terms.

Homotopy perturbation method The homotopy perturbation method is a useful tool to the search for solitons without the requirement of presence of small perturbations. In this method, a homotopy is constructed with a homotopy parameter, p. When $p = 0$, it becomes a nonlinear wave equation such as a KdV equation with a known soliton solution; when $p = 1$, it turns out to be the original nonlinear equation. To change p from zero to unity, one must only change from a trial soliton to the solved soliton.

Variational iteration method The variational iteration method is a new method for obtaining soliton-type solutions of various nonlinear wave equations. The method begins with a soliton-type solution with some unknown parameters which can be determined after few iterations. The iteration formulation is constructed by a general Lagrange multiplier which can be identified optimally via variational theory.

Exp-function method The exp-function method is a new method for searching for both soliton-type solutions and periodic solutions of nonlinear systems. The method assumes that the solutions can be expressed in arbitrary forms of the exp-function.

Definition of the Subject

The soliton is a kind of nonlinear wave. There are many equations of mathematical physics which have solutions of the soliton type. The first observation of this kind of wave was made in 1834 by John Scott Russell [1]. In 1895, the famous KdV equation, which possesses soliton solutions, was obtained by D. J. Korteweg and H. de Vries [2], who established a mathematical basis for the study of various solitary phenomena.

From a modern perspective, the soliton is used as a constructive element to formulate the complex dynamical behavior of wave systems throughout science: from hydrodynamics to nonlinear optics, from plasmas to shock waves, from tornados to the Great Red Spot of Jupiter, from traffic flow to the Internet, from Tsunamis to turbulence [3]. More recently, solitary waves are of key importance in the quantum fields: on extremely small scales and at very high observational resolution equivalent to a very high energy, space–time resembles a stormy ocean and particles and their interactions have soliton-type solutions [4].

Introduction

The soliton was first discovered in 1834 by John Scott Russell, who observed that a canal boat stopping suddenly gave rise to a solitary wave which traveled down the canal for several miles, without breaking up or losing strength. Russell named this phenomenon the 'soliton'.

In a highly informative as well as entertaining article [1] J.S. Russell gave an engaging historical account of the important scientific observation:

> *I was observing the motion of a boat which was rapidly drawn along a narrow channel by a pair of horses, when the boat suddenly stopped – not so the mass of water in the channel which it had put in motion; it accumulated round the prow of the vessel in a state of violent agitation, then suddenly leaving it behind, rolled forward with great velocity, assuming the form of a large solitary elevation, a rounded, smooth and well-defined heap of water, which continued its course along the channel apparently without change of form or diminution of speed. I followed it on horseback, and overtook it still rolling on at*

a rate of some eight or nine miles an hour, preserving its original figure some thirty feet long and a foot to a foot and a half in height. Its height gradually diminished, and after a chase of one or two miles I lost it in the windings of the channel. Such, in the month of August 1834, was my first chance interview with that singular and beautiful phenomenon which I have called the Wave of Translation.

His ideas did not earn attention until 1965 when N.J. Zabusky and M.D. Kruskal began to use a finite difference approach to the study of KdV equation [5], and various analytical methods also led to a complete understanding of Solitons, especially the inverse scattering transform proposed by Gardner, Greene, Kruskal, and Miura [6] in 1967. The significance of Russell's discovery was then fully appreciated. It was discovered that many phenomena in physics, electronics and biology can be described by the mathematical and physical theory of the 'Soliton'.

The particle-like properties of solitons [7] also caught much attention, and were proposed as models for elementary particles [8]. More recently it has been realized that some of the quantum fields which are used to describe particles and their interactions also have solutions of the soliton type [9].

Methods for Soliton Solutions

The investigation of soliton solutions of nonlinear evolution equations plays an important role in the study of nonlinear physical phenomena. There are many analytical approaches to the search for soliton solutions, such as soliton perturbation, tanh-function method, projective approach, F-expansion method, and others [10,11,12,13,14].

Soliton Perturbation

We consider the following perturbed nonlinear evolution equation [15,16]

$$u_T + N(u) = \varepsilon R(u), \quad 0 < \varepsilon \ll 1 . \tag{1}$$

When $\varepsilon = 0$, we have the un-perturbed equation

$$u_T + N(u) = 0 , \tag{2}$$

which is assumed to have a soliton solution.

When $\varepsilon \neq 0$, but $0 < \varepsilon \ll 1$, we can use perturbation theory [15,16], and look for approximate solutions of Eq. (1), which are close to the soliton solutions of Eq. (2). Using multiple time scales (a slow time τ and a fast time t, such that $\partial_T = \partial_t + \varepsilon\partial_\tau$), we assume that the soliton solution can be expressed in the form

$$u(x, T) = u_0(\xi, \tau) + \varepsilon u_1(\xi, \tau, t) + \varepsilon^2 u_2(\xi, \tau, t) + \cdots \tag{3}$$

where $\xi = x - ct$, and τ is a slow time and t is a fast time.

Substituting Eq. (3) into Eq. (1) and then equating like-powers of ε, we can obtain a series of linear equations for u_i $(i = 0, 1, 2, 3, \ldots)$.

In most cases the nonlinear term $R(u)$ in Eq. (1) plays an import role in understanding various solitary phenomena, and the coefficient ε is not limited to a "small parameter".

Variational Approach

Recently, variational theory and homotopy technology have been successfully applied to the search for soliton solutions [17,18] without requiring the small parameter assumption. Both variational and homotopy technologies can lead to an extremely simple and elementary, but rigorous, derivation of soliton solutions.

Considering the KdV equation

$$\frac{\partial u}{\partial t} - 6u\frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3} = 0 , \tag{4}$$

we seek its traveling wave solutions in the following frame

$$u(x, t) = U(\xi) \quad v(x, t) = V(\xi), \quad \xi = x - ct , \tag{5}$$

where c is angular frequency. Substituting Eq. (5) into Eq. (4) yields

$$-cu' - 6uu' + u''' = 0 , \tag{6}$$

where a prime denotes the differential with respect to ξ.

Integrating Eq. (6) yields the result

$$-cu - 3u^2 + u'' = 0 . \tag{7}$$

By the semi-inverse method [19], the following variational formulation is established

$$J = \int_0^\infty \left(\frac{1}{2}cu^2 + u^3 + \frac{1}{2}\left(\frac{\mathrm{d}u}{\mathrm{d}\xi}\right)^2 \right) \mathrm{d}\xi . \tag{8}$$

The semi-inverse method is a powerful mathematical tool to the search for variational formulae for real-life physical problems.

By the Ritz method, we search for a solitary wave solution in the form

$$u = p\,\mathrm{sech}^2(q\xi) , \tag{9}$$

where p and q are constants to be further determined.

Substituting Eq. (9) into Eq. (8) results in

$$J = \int_0^\infty \left[\frac{1}{2}cp^2\,\mathrm{sech}^4(q\xi) + p^3\,\mathrm{sech}^6(q\xi) + \frac{1}{2}(4p^2q^2\,\mathrm{sech}^4(q\xi)\tanh^2(q\xi)\right]\mathrm{d}\xi$$
$$= \frac{cp^2}{2q}\int_0^\infty \mathrm{sech}^4(z)\mathrm{d}z + \frac{p^3}{q}\int_0^\infty \mathrm{sech}^6(z)\mathrm{d}z + 2p^2q\int_0^\infty \left\{\mathrm{sech}^4(z)\tanh^4(z)\right\}\mathrm{d}z$$
$$= \frac{cp^2}{3q} + \frac{8p^3}{15q} + \frac{4p^2q}{15}\,. \tag{10}$$

Making J stationary with respect to p and q results in

$$\frac{\partial J}{\partial p} = \frac{2cp}{3q} + \frac{24p^2}{15q} + \frac{8pq}{15} = 0\,, \tag{11}$$
$$\frac{\partial J}{\partial q} = -\frac{cp^2}{3q^2} - \frac{8p^3}{15q^2} + \frac{4p^2}{15} = 0\,, \tag{12}$$

or simplifying

$$5c + 12p + 4q^2 = 0\,, \tag{13}$$
$$-5c - 8p + 4q^2 = 0\,. \tag{14}$$

From Eqs. (13) and (14), we can easily obtain the following relations:

$$p = -\frac{1}{2}c, \quad q = \sqrt{\frac{c}{4}}\,. \tag{15}$$

So the solitary wave solution can be approximated as

$$u = -\frac{c}{2}\,\mathrm{sech}^2\sqrt{\frac{c}{4}}\,(x - ct - \xi_0)\,, \tag{16}$$

which is the exact solitary wave solution of KdV equation (4).

The preceding analysis has the virtue of utter simplicity. The suggested variational approach can be readily applied to the search for solitary wave solutions of other nonlinear problems, and the present example can be used as paradigms for many other applications in searching for solitary wave solutions of real-life physics problems.

Variational Iteration Method

The variational iteration method [20] is an alternative approach to soliton solutions without the requirement of establishing a variational formulation for the discussed problems [17,21,22,23,24]. As an illustrating example, we consider the $K(3,1)$ equation in the form [17]:

$$u_t + u^2u_x + u_{xxx} = 0\,. \tag{17}$$

According to the variational iteration method, its iteration formulation can be constructed as follows

$$u_{n+1}(x,t) = u_n(x,t) - \int_0^t \left\{(u_n)_t + u_n^2(u_n)_x + (u_n)_{xxx}\right\}\mathrm{d}t\,. \tag{18}$$

To search for its compacton-like solution, we assume the solution has the form

$$u_0(x,t) = \frac{a\sin^2(kx + wt)}{b + c\sin^2(kx + wt)}\,, \tag{19}$$

where a, b, k, and w are unknown constants further to be determined after few iterations [17].

Homotopy Perturbation Method

The homotopy perturbation method [25] provides a simple mathematical tool for searching for soliton solutions without any small perturbation [18,26]. Considering the following nonlinear equation

$$\frac{\partial u}{\partial t} + au\frac{\partial u}{\partial x} + b\frac{\partial^3 u}{\partial x^3} + N(u) = 0, \quad a > 0,\ b > 0\,, \tag{20}$$

we can construct a homotopy in the form

$$(1-p)\left\{\frac{\partial u}{\partial t} + 6u\frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3}\right\} + p\left\{\frac{\partial u}{\partial t} + au\frac{\partial u}{\partial x} + b\frac{\partial^3 u}{\partial x^3} + N(u)\right\} = 0\,. \tag{21}$$

When $p = 0$, we have

$$\frac{\partial u}{\partial t} + 6u\frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3} = 0\,, \tag{22}$$

a well-known KdV equation whose soliton solution is known. When $p = 1$, Eq. (21) turns out to be the original equation. According to the homotopy perturbation method, we assume

$$u = u_0 + pu_1 + p^2u_2 + \cdots \tag{23}$$

Substituting Eq. (23) into Eq. (21), and proceeding with the same process as the traditional perturbation method does, we can easily solve u_0, u_1 and other components. The solution can be expressed finally in the form

$$u = u_0 + u_1 + u_2 + \cdots \tag{24}$$

The homotopy perturbation method always stops before the second iteration, so the solution can be expressed as

$$u = u_0 + u_1 \tag{25}$$

for most cases.

Parameter-Expansion Method

The parameter-expansion method [27,28,29,30] does not require one to construct a homotopy. To illustrate its solution procedure, we re-write Eq. (20) in the form

$$\frac{\partial u}{\partial t} + au\frac{\partial u}{\partial x} + b\frac{\partial^3 u}{\partial x^3} + 1 \cdot N(u) = 0 \,. \tag{26}$$

Supposing that the parameters a, b, and 1 can be expressed in the forms

$$a = a_0 + pa_1 + p^2 a_2 + \cdots \tag{27}$$

$$b = b_0 + pb_1 + p^2 b_2 + \cdots \tag{28}$$

$$1 = pc_1 + p^2 c_2 + \cdots \tag{29}$$

where p is a bookkeeping parameter, $p = 1$. Substituting Eqs. (23), (27), (28) and (29) into Eq. (26) and proceeding the same way as the perturbation method, we can easily obtain the needed solution.

Exp-function Method

The exp-function method [31,32,33] provides us with a straightforward and concise approach to obtaining generalized solitonary solutions and periodic solutions and the solution procedure, with the help of Matlab or Mathematica, is utterly simple. Consider a general nonlinear partial differential equation of the form

$$F(u, u_x, u_y, u_z, u_t, u_{xx}, u_{yy}, u_{zz}, u_{tt}, u_{xy}, u_{xt}, u_{yt}, \ldots) = 0 \,. \tag{30}$$

Using a transformation

$$\eta = ax + by + cz + dt \,, \tag{31}$$

we can re-write Eq. (30) in the form of the following nonlinear ordinary differential equation:

$$G(u, u', u'', u''', \ldots) = 0 \,, \tag{32}$$

where a prime denotes a derivation with respect to η.

According to the exp-function method, the traveling wave solutions can be expressed in the form

$$u(\eta) = \frac{\sum_{n=-k}^{l} a_n \exp(n\eta)}{\sum_{m=-i}^{j} b_m \exp(m\eta)} \,, \tag{33}$$

where i, j, k, and l are positive integer which could be freely chosen, a_n and b_m are unknown constants to be determined. The solution procedure is illustrated in [32].

Future Directions

It is interesting to point out the connection of catastrophe theory to loop soliton chaos, and finally to chaotic Cantorian spacetime [34,35].

El Naschie [34,35] studied the Eguchi–Hanson gravitational instanton solution and its interpretation by 't Hooft in the context of a quantum gravitational Hilbert space, as an event and a possible solitonic "extended" particle. Transferring a certain solitonic solution of Einstein's field equations in Euclidean "real" space–time to the mathematical infinitely-dimensional Hilbert space, it is possible to observe a new non-standard process by which a definite mass can be assigned to massless particles. Thus by invoking Einstein's gravity in "solitonic" gauge theory and vice versa, an alternative explanation for how massless particles acquire mass is found, which is also in harmony with the basic structure of our standard model as it stands at present [34,35].

Bibliography

Primary Literature

1. Russell JS (1844) Report on Waves. Fourteenth Meeting of the British Association for the Advancement of Science, John Murray, London, pp 311–390
2. Korteweg DJ, De Vires G (1895) On the change of form of long waves advancing in a rectangular channel, and a new type of long stationary wave. Phil Mag Ser 539:422–443
3. Eilbeck C (2007) John Scott Russell and the solitary wave. Heriot-Watt University, Edinburgh http://www.ma.hw.ac.uk/~chris/scott_russell.html
4. El Naschie MS (2004) A review of E infinity theory and the mass spectrum of high energy particle physics. Chaos Solit Fract 19(1):209–236
5. Zabusky NJ, Kruskal MD (1965) Interaction of 'Solitons' in a collisionless plasma and the recurrence of initial states. Phys Rev Lett 15:240–243
6. Gardner CS, Greene JM, Kruskal MD, Miura RM (1967) Method for solving the KdV equation. Phys Rev Lett 19:1095–1097
7. Bode M, Liehr AW, Schenk CP et al (2002) Interaction of dissipative solitons: particle-like behavior of localized structures in a three-component reaction-diffusion system. Physica D (1–2):45–66
8. Braun HB, Kulda J, Roessli B et al (2005) Emergence of soliton chirality in a quantum antiferromagnet. Nat Phys 1(3):159–163
9. Ahufinger V, Mebrahtu A, Corbalan R et al (2007) Quantum switches and quantum memories for matter-wave lattice solitons. New J Phys 9:4
10. Ye JF, Zheng CL, Xie LS (2006) Exact solutions and localized excitations of general Nizhnik–Novikov–Veselov system in (2+1)-dimensions via a projective approach. Int J Nonlinear Sci Num Simul 7(2):203–208
11. Bogning JR, Tchakoutio-Nguetcho AS, Kofane TC (2005) Gap solitons coexisting with bright soliton in nonlinear fiber arrays.

Int J Nonlinear Sci Num Simul 6(4):371–385; Abdusalam HA (2005) On an improved complex tanh-function method. Int J Nonlinear Sci Num Simul 6(2):99–106
12. El-Sabbagh MF, Ali AT (2005) New exact solutions for (3+1)-dimensional Kadomtsev–Petviashvili equation and generalized (2+1)-dimensional Boussinesq equation. Int J Nonlinear Sci Num Simul 6(2):151–162
13. Shen JW, Xu W (2004) Bifurcations of smooth and non-smooth travelling wave solutions of the Degasperis-Procesi equation. Int J Nonlinear Sci Num Simul 5(4):397–402
14. Sheng Z (2007) Further improved F-expansion method and new exact solutions of Kadomstev–Petviashvili equation. Chaos Solit. Fract 32(4):1375–1383
15. Yu H, Yan J (2006) Direct approach of perturbation theory for kink solitons. Phys Lett A 351(1–2):97–100
16. Herman RL (2005) Exploring the connection between quasistationary and squared eigenfunction expansion techniques in soliton perturbation theory. Nonlinear Anal 63(5–7): e2473–e2482
17. He JH, Wu XH (2006) Construction of solitary solution and compacton-like solution by variational iteration method. Chaos Solit Fract 29(1):108–113
18. He JH (2005) Application of homotopy perturbation method to nonlinear wave equations. Chaos Solit Fract 26(3): 695–700
19. He JH (2004) Variational principles for some nonlinear partial differential equations with variable coefficients. Chaos Solit Fract 19(4):847–851
20. He JH (1999) Variational iteration method – a kind of non-linear analytical technique: Some examples. Int J Non-Linear Mech 34(4):699–708
21. Abulwafa EM, Abdou MA, Mahmoud AA (2007) Nonlinear fluid flows in pipe-like domain problem using variational-iteration method. Chaos Solit Fract 32(4):1384–1397
22. Inc M (2007) Exact and numerical solitons with compact support for nonlinear dispersive K(*m*,*p*) equations by the variational iteration method. Phys A 375(2):447–456
23. Soliman AA (2006) A numerical simulation and explicit solutions of KdV-Burgers' and Lax's seventh-order KdV equations. Chaos Solit Fract 29(2):294–302
24. Abdou MA, Soliman AA (2005) Variational iteration method for solving Burger's and coupled Burger's equations. J Comput Appl Math 181(2):245–251
25. He JH (2000) A coupling method of a homotopy technique and a perturbation technique for non-linear problems. Int J Non-Linear Mech 35(1):37–43
26. Ganji DD, Rafei M (2006) Solitary wave solutions for a generalized Hirota-Satsuma coupled KdV equation by homotopy perturbation method. Phys Lett A 356(2):131–137
27. Shou DH, He JH (2007) Application of Parameter-expanding Method to Strongly Nonlinear Oscillators. Int J Nonlinear Sci Numer Simul 8:113–116
28. He JH (2001) Bookkeeping parameter in perturbation methods. Int J Nonlinear Sci Numer Simul 2:257–264
29. He JH (2002) Modified Lindstedt-Poincare methods for some strongly non-linear oscillations. Part I: expansion of a constant. Int J Non-Linear Mech 37:309–314
30. Xu L (2007) He's parameter-expanding methods for strongly nonlinear oscillators. J Comput Appl Math 207(1):148–157
31. He J-H, Wu X-H (2006) Exp-function method for nonlinear wave equations. Chaos Solit Fract 30(3):700–708
32. He J-H, Abdou MA (2007) New periodic solutions for nonlinear evolution equations using Exp-function method. Chaos Solit Fract 34(5):1421–1429
33. Wu X-H, He J-H (2008) EXP-function method and its application to nonlinear equations. Chaos Solit Fract 38(3):903–910
34. El Naschie MS (2004) Gravitational instanton in Hilbert space and the mass of high energy elementary particles. Chaos Solit Fract 20(5):917–923
35. El Naschie MS (2004) How gravitational instanton could solve the mass problem of the standard model of high energy particle physics. Chaos Solit Fract 21(1):249–260

Books and Reviews

He JH (2006) Some Asymptotic Methods for Strongly Nonlinear Equations. Int J Mod Phys B 20(10):1141–1199; 20(18): 2561–2568

He JH (2006) Non-perturbative methods for strongly nonlinear problems. dissertation.de-Verlag im Internet, Berlin

Drazin PG, Johnson RS (1989) Solitons: An Introduction. Cambridge University Press, Cambridge

Solitons and Compactons

Ji-Huan He[1], Shun-dong Zhu[2]

[1] Modern Textile Institute, Donghua University, Shanghai, China

[2] Department of Science, Zhejiang Lishui University, Lishui, China

Article Outline

Glossary
Definition of the Subject
Introduction
Solitons
Compactons
Generalized Solitons and Compacton-like Solutions
Future Directions
Cross References
Bibliography

Glossary

Soliton A soliton is a stable pulse-like wave that can exist in some nonlinear systems. The soliton, after a collision with another soliton, eventually emerges unscathed.

Compacton A compacton is a special solitary traveling wave that, unlike a soliton, does not have exponential tails.

Generalized soliton A generalized soliton is a soliton with some free parameters. Generally a generalized soliton can be expressed by exponential functions.

Compacton-like solution A compcton-like solution is a special wave solution which can be expressed by the squares of sinusoidal or cosinoidal functions.

Definition of the Subject

Soliton and compacton are two kinds of nonlinear waves. They play an indispensable and vital role in all ramifications of science and technology, and are used as constructive elements to formulate the complex dynamical behavior of wave systems throughout science: from hydrodynamics to nonlinear optics, from plasmas to shock waves, from tornados to the Great Red Spot of Jupiter, from traffic flow to Internet, from Tsunamis to turbulence. More recently, soliton and compacton are of key importance in the quantum fields and nanotechnology especially in nanohydrodynamics.

Introduction

Solitary waves were first observed by John Scott Russell in 1895, and were studied by D. J. Korteweg and H. de Vries in 1895 . Compactons are special solitons with finite wavelength. It was Philip Rosenau and his colleagues who first found compactons in 1993. Please refer to "▶ Soliton Perturbation" for detailed information.

Solitons

A soliton is a special solitary traveling wave that after a collision with another soliton eventually emerges unscathed. Solitons are solutions of partial differential equations that model phenomena like water waves or waves along a weakly anharmonic mass-spring chain.

The Korteweg–de Vries (KdV) equation is the generic model for the study of nonlinear waves in fluid dynamics, plasma and elastic media. KdV equation is one of the most fundamental equations in nature and plays a pivotal role in nonlinear phenomena. We consider the KdV equation in the form

$$\frac{\partial u}{\partial t} + 6u\frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3} = 0\,. \tag{1}$$

Its solitary traveling wave solution can be solved as

$$u(x,t) = \tfrac{1}{2}c\,\mathrm{sech}^2\left\{\tfrac{1}{2}c^{1/2}(x-ct)\right\}\,. \tag{2}$$

The bell-like solution as illustrated in Fig. 1 is called a soliton.

We re-write Eq. (2) in an equivalently form:

$$u(\xi) = p\,\mathrm{sec}\,h^2(q\xi) = \frac{4p}{\mathrm{e}^{2q\xi} + \mathrm{e}^{-2q\xi} + 2}\,. \tag{3}$$

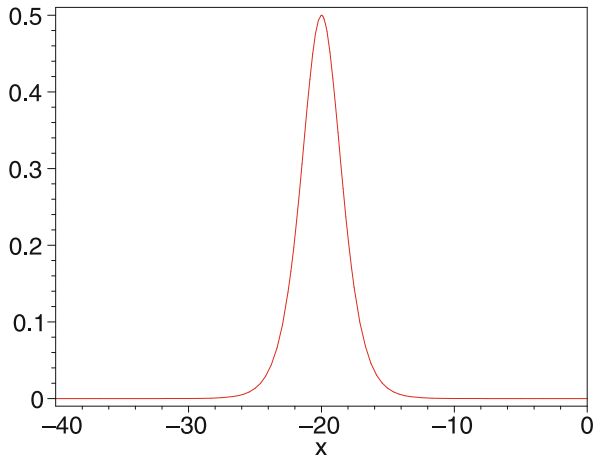

Solitons and Compactons, Figure 1
Bell-like solitary wave

where $u(x,t) = u(\xi)$, $\xi = x - ct$, c is the wave velocity.

It is obvious that

$$\lim_{\xi\to\infty} u(\xi) = 0 \quad \text{and} \quad \lim_{\xi\to-\infty} u(\xi) = 0\,. \tag{4}$$

The soliton has exponential tails, which are the basic character of solitary waves. The soliton obeys a superposition-like principle: solitons passing through one another emerge unmodified, see Fig. 2.

Compactons

Now consider a modified version of KdV equation in the form

$$u_t + (u^2)_x + (u^2)_{xxx} = 0\,. \tag{5}$$

Introducing a complex variable ξ defined as $\xi = x - ct$, where c is the velocity of traveling wave, integrating once, we have

$$-cu + u^2 + (u^2)_{\xi\xi} = D\,, \tag{6}$$

where D is an integral constant.

To avoid singular solutions, we set $D = 0$. We re-write Eq. (6) in the form

$$v_{\xi\xi} + v - cv^{1/2} = 0\,, \tag{7}$$

where $u^2 = v$.

In case $c = 0$, we have periodic solution: $v(\xi) = A\cos\xi + B\sin\xi$. Periodic solution of nonlinear oscillators can be approximated by sinusoidal function. It helps understanding if an equation can be classified as oscillatory by direct inspection of its terms.

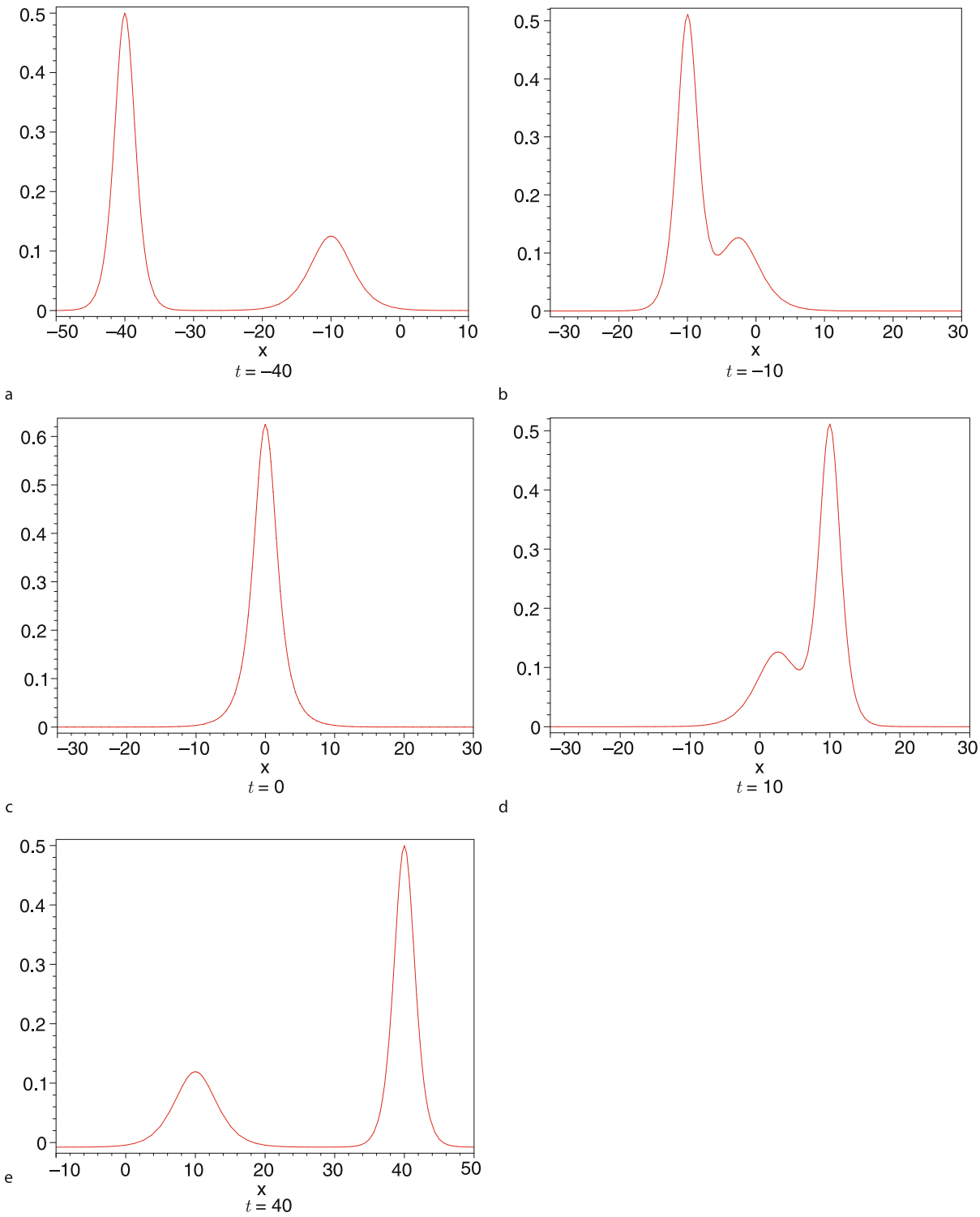

Solitons and Compactons, Figure 2
Collision of two solitary waves

We consider two common order differential equations whose exact solutions are important for physical understanding:

$$u'' - k^2 u = 0\,, \tag{8}$$

and

$$u'' + \omega^2 u = 0\,. \tag{9}$$

Both equations have linear terms with constant coefficients.

The crucial difference between these two very simple equations is the sign of the coefficient of u in the second term. This determines whether the solutions are exponential or oscillatory. The general solution of Eq. (8) is

$$u = Ae^{kt} + Be^{-kt}\,. \tag{10}$$

The second Eq. (9) has a positive coefficient of u, and in this case the general solution reads

$$u = A\cos\omega t + B\sin\omega t\,. \tag{11}$$

This solution describes an oscillation at the angular velocity ω.

Equation (7) behaves sometimes like an oscillator when $1 - cv^{-1/2} > 0$, i. e., $u = v^{1/2}$ has a periodic solution, we assume v can be expressed in the form

$$v = u^2 = A^2\cos^4\omega\xi\,. \tag{12}$$

Substituting Eq. (12) into Eq. (7) results in

$$12A^2\omega^2\cos^2\omega\xi - 16A^2\omega^2\cos^4\omega\xi + A^2\cos^4\omega\xi - cA\cos^2\omega\xi = 0\,. \tag{13}$$

We, therefore, have

$$\begin{aligned} 12A^2\omega^2 - cA &= 0 \\ -16A^2\omega^2 + A^2 &= 0\,. \end{aligned} \tag{14}$$

Solving the above system, Eq. (14), yields

$$\omega = \frac{1}{4}\,, \quad A = \frac{4}{3}c\,. \tag{15}$$

We obtain the solution in the form

$$u = v^{1/2} = \frac{4c}{3}\cos^2\left[\frac{1}{4}(x - ct)\right]\,. \tag{16}$$

By a careful inspection, v can tend to a very small value or even zero, as a result, $1 - cv^{-1/2}$ tends to negative infinite, and Eq. (7) behaves like Eq. (8) with $k \to \infty$, the exponential tails vanish completely at the edge of the bell-shape (see Fig. 3):

$$u = \begin{cases} \frac{4c}{3}\cos^2\left[\frac{1}{4}(x - ct)\right], & |x - ct| \le 2\pi \\ 0\,, & \text{otherwise.} \end{cases} \tag{17}$$

This is a compact wave. Unlike solitons (Fig. 4), compacton does not have exponential tails (Fig. 3).

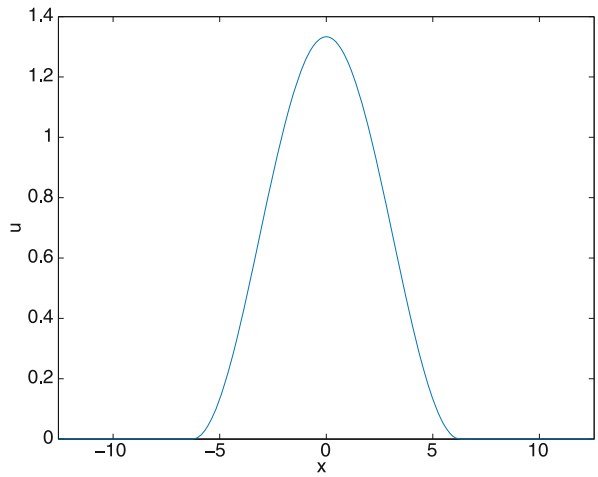

Solitons and Compactons, Figure 3
Compaton wave without tails

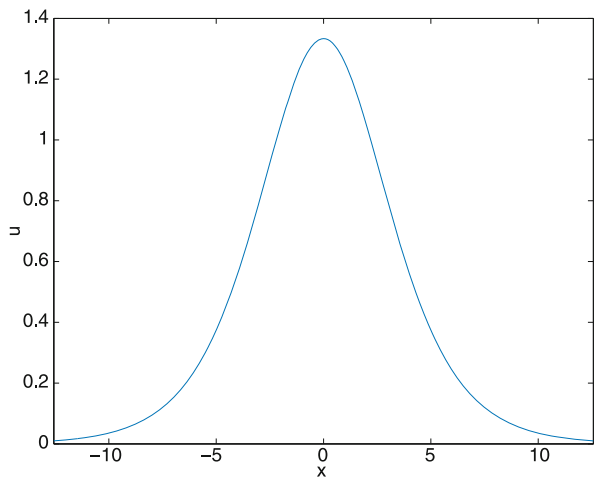

Solitons and Compactons, Figure 4
Solitary wave with two tails

Generalized Solitons and Compacton-like Solutions

Solitary solutions have tails, which can be best expressed by exponential functions. We can assume that a solitary

solution can be expressed in the following general form

$$u(\eta) = \frac{\sum_{n=-c}^{d} a_n \exp(n\eta)}{\sum_{m=-p}^{q} b_m \exp(m\eta)} , \tag{18}$$

where c, d, p, and q are positive integers which are unknown to be further determined, a_n and b_m are unknown constants. The unknown constants can be easily determined using Matlab, the method is called the Exp-function method.

We consider the modified KdV equation in the form:

$$u_t + u^2 u_x + u_{xxx} = 0 . \tag{19}$$

Using a transformation: $u(x,t) = u(\xi), \xi = kx + \omega t$, we have

$$\omega u' + k u^2 u' + k^3 u''' = 0 , \tag{20}$$

where prime denotes the differential with respect to ξ.

We suppose that the solution of Eq. (20) can be expressed as

$$u(\xi) = \frac{a_c \exp(c\xi) + \cdots + a_{-d} \exp(-d\xi)}{b_p \exp(p\xi) + \cdots + b_{-q} \exp(-q\xi)} . \tag{21}$$

To determine values of c, d, p and q, we balance the linear term of highest order in Eq. (20) with the highest order nonlinear term. According to the homogeneous balance principle, we obtain the result $c = p$ and $d = q$. For simplicity, we set $c = p = 1$ and $d = q = 1$, so Eq. (21) reduces to

$$u(\xi) = \frac{a_1 \exp(\xi) + a_0 + a_{-1} \exp(-\xi)}{\exp(p\xi) + b_0 + b_{-1} \exp(-\xi)} . \tag{22}$$

Substituting Eq. (22) into Eq. (20), and by the help of Matlab, clearing the denominator and setting the coefficients of power terms like $\exp(j\xi)$, $j = 1, 2, \cdots$ to zero yield a system of algebraic equations, solving the obtained system, we obtain the following exact solutions:

$$\begin{cases} a_0 = a_1 b_0 + \dfrac{3k^2 b_0}{a_1} , & a_{-1} = \dfrac{b_0^2 \left(3k^2 + 2a_1^2\right)}{8a_1} , \\ b_{-1} = \dfrac{b_0^2 \left(3k^2 + 2a_1^2\right)}{8a_1^2} , & \omega = -ka_1^2 - k^3 , \end{cases} \tag{23}$$

where a_1 and b_0 are free parameters, which depends upon the initial conditions and/or boundary conditions. The property that stability may depend on initial/boundary conditions is characteristic only for nonlinear systems.

The relationship between wave speed and frequency is

$$\omega = -ka_1^2 - k^3 . \tag{24}$$

Note that the value of a_1 is determined from the initial/boundary conditions, so frequency or wave speed may not independent of initial/boundary conditions.

Then, the closed form solution of Eq. (19) reads

$$\begin{aligned} & u(x,t) \\ & = \frac{\left(a_1 \exp[kx - (ka_1^2 + k^3)t] + a_1 b_0 + \frac{3k^2 b_0}{a_1} + \frac{b_0^2(3k^2+2a_1^2)}{8a_1} \exp[-kx + (ka_1^2 + k^3)t]\right)}{\left(\exp[kx - (ka_1^2 + k^3)t] + b_0 + \frac{b_0^2(3k^2+2a_1^2)}{8a_1^2} \exp[-kx + (ka_1^2 + k^3)t]\right)} \\ & = a_1 + \frac{\frac{3k^2 b_0}{8}}{\left(\exp[kx - (ka_1^2 + k^3)t] + b_0 + \frac{b_0^2(3k^2+2a_1^2)}{8a_1^2} \exp[-kx + (ka_1^2 + k^3)t]\right)} . \end{aligned} \tag{25}$$

Generally a_1, b_0 and k are real numbers, and the obtained solution, Eq. (25), is a generalized soliton solution.

If we choose $k = 1, a_1 = 1, b_0 = \sqrt{8/5}$, Eq. (25) becomes

$$u(x,t) = 1 + \frac{3\sqrt{1/40}}{\exp[x - 2t] + \sqrt{8/5} + \exp[-x + 2t]} . \tag{26}$$

The bell-like solution is illustrated in Fig. 5.

In case k is an imaginary number, the obtained solitary solution can be converted into periodic solution or compact-like solution. We write $k = iK$, Eq. (25) becomes

$$\begin{aligned} u(x,t) &= a_1 \\ &+ \frac{-\frac{3K^2 b_0}{8}}{\left((1+p)\exp[Kx - (Ka_1^2 - K^3)t] + b_0 + i(1-p)\frac{b_0^2(3k^2+2a_1^2)}{8a_1^2} \exp[-Kx + (Ka_1^2 - K^3)t]\right)} , \end{aligned} \tag{27}$$

where $p = \frac{b_0^2(-3K^2+2a_1^2)}{8a_1^2}$.

If we search for a periodic solution or compact-like solution, the imaginary part in the denominator of Eq. (27) must be zero, that requires that

$$1 - p = 1 - \frac{b_0^2(-3K^2 + 2a_1^2)}{8a_1^2} = 0 . \tag{28}$$

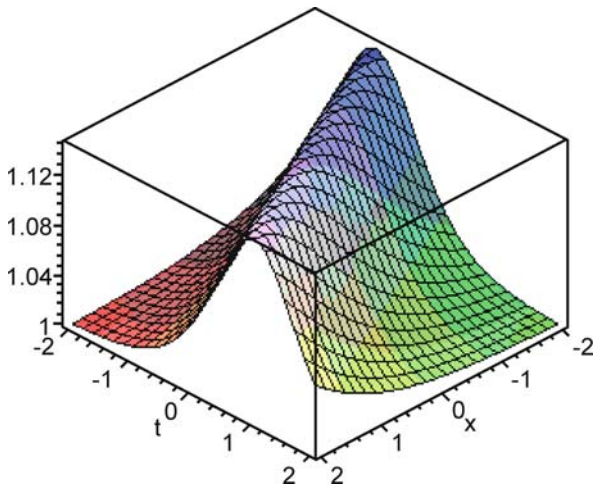

Solitons and Compactons, Figure 5
Propagation of a solution with respect to time

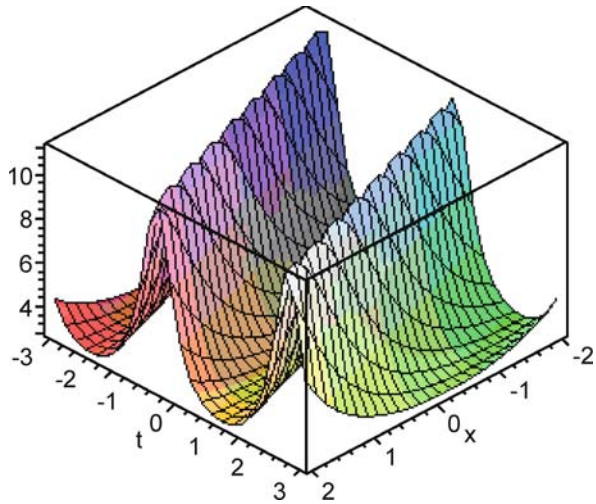

Solitons and Compactons, Figure 6
Periodic solution

Solving b_0 from Eq. (28), we obtain

$$b_0 = \pm\sqrt{\frac{8}{-3K^2 + 2a_1^2}}\,. \tag{29}$$

Substituting Eq. (29) into Eq. (27) results in a periodic solution, which reads

$$u(x,t) = a_1 + \frac{\pm 3K^2\sqrt{\frac{2}{-3K^2+2a_1^2}}}{\cos[Kx - (Ka_1^2 - K^3)t] \pm \sqrt{\frac{2}{-3K^2+2a_1^2}}} \tag{30}$$

or a generalized compact-like solution:

$$u(x,t) = \begin{cases} a_1 + \dfrac{\pm 3K^2\sqrt{\frac{2}{-3K^2+2a_1^2}}}{\left(\cos\left[Kx-\left(Ka_1^2-K^3\right)t\right]\pm\sqrt{\frac{2}{-3K^2+2a_1^2}}\right)}\,, \\ a_1 + 3K^2\,, \quad \text{otherwise} \end{cases}$$
$$\left|Kx - \left(Ka_1^2 - K^3\right)t\right| \le \frac{\pi}{2} \tag{31}$$

where a_1 and K are free parameters, and it requires that $2a_1^2 > 3K^2$. If we choose $k = 1, a_1 = 1, b_0 = \sqrt{8/5}$, Eq. (30) becomes

$$u(x,t) = 1 + \frac{3\sqrt{2}}{\cos[x - 3t] + \sqrt{2}}\,. \tag{32}$$

The periodic solution is illustrated in Fig. 6.

Now we give an heuristical explanation of why Eq. (19) behaves sometimes periodically and sometimes compaton-like.

We re-written Eq. (20) in form

$$u'' + \frac{\omega}{k^3}u + \frac{1}{3k^2}u^3 = 0. \tag{33}$$

It is a well-known Duffing equation with a periodic solution for all $\omega > 0$ and $k > 0$.

Actually in our study ω can be negative, we re-write Eq. (33) in the form

$$u'' - \frac{\omega}{k^3}u + \frac{1}{3k^2}u^3 = 0\,, \quad \omega > 0\,. \tag{34}$$

This equation, however, has not always a periodic solution. We use the parameter-expansion method to find its period and the condition to be an oscillator. In order to carry out a straightforward expansion like that in the classical perturbation method, we need to introduce a parameter, λ, because none appear explicitly in this equation. To this end, we seek an expansion in the form

$$u = u_0 + \lambda u_1 + \lambda^2 u_2 + \lambda^3 u_3 + \cdots\,. \tag{35}$$

The parameter λ is used as a bookkeeping device and is set equal to unity.

The coefficients of the linear term and nonlinear term can be, respectively, expanded in a similar way:

$$-\frac{\omega}{k^3} = \Omega^2 + m_1\lambda + m_2\lambda^2 + \ldots \tag{36}$$

$$\frac{1}{3k^2} = n_1\lambda + n_2\lambda^2 + \ldots\,, \tag{37}$$

where m_i and n_i are unknown constants to be further determined.

Interpretation of why such expansions work well is given by [1].

Substituting Eqs.(35)–(37) to (34), we have

$$\begin{aligned}&\left(u_0+\lambda u_1+\lambda^2 u_2+\ldots\right)''\\&+\left(\Omega^2+m_1\lambda+m_2\lambda^2+\ldots\right)\left(u_0+\lambda u_1+\lambda^2 u_2+\ldots\right)\\&+\left(n_1\lambda+n_2\lambda^2+\ldots\right)\cdot\left(u_0+\lambda u_1+\lambda^2 u_2+\ldots\right)^3=0\end{aligned} \tag{38}$$

and equating coefficients of like powers of λ, we obtain

Coefficient of λ^0

$$u_0''+\Omega^2 u_0=0\,. \tag{39}$$

Coefficient of λ^1

$$u_1''+\Omega^2 u_1+m_1 u_0+n_0 u_0^3=0\,. \tag{40}$$

The solution of Eq. (39) is

$$u_0=A\cos\Omega t\,. \tag{41}$$

Substituting u_0 into (40) gives

$$\begin{aligned}u_1''+\Omega^2 u_1+A(m_1+\tfrac{3}{4}n_0A^2)\cos\Omega t\\+\tfrac{1}{4}n_0A^3\cos 3\Omega t=0\,.\end{aligned} \tag{42}$$

No secular term in u_1 requires that

$$m_1+\frac{3}{4}n_0A^2=0 \quad \text{or} \quad A=0\,. \tag{43}$$

If the first-order approximate solution is searched for, then we have

$$-\frac{\omega}{k^3}=\Omega^2+m_1 \tag{44}$$

$$\frac{1}{3k^2}=n_1\,. \tag{45}$$

We finally obtain the following relationship

$$\Omega^2=-\frac{\omega}{k^3}+\frac{1}{4k^2}A^2\,. \tag{46}$$

To behave like an oscillator requires that

$$-\frac{\omega}{k^3}+\frac{1}{4k^2}A^2>0 \tag{47}$$

or

$$\frac{\omega}{k}<\frac{1}{4}A^2\,. \tag{48}$$

The amplitude A may strongly depend upon initial/boundary conditions which may determine the wave type of a nonlinear equation.

Now we approximate Eq. (34) in the form

$$u''+\frac{1}{k^2}\left(-\frac{\omega}{k}+\frac{A^2\cos^2\Omega t}{3}\right)u=0\,. \tag{49}$$

In case $|\Omega t|\to\pi/2$, the above equation behaves exponentially, resulting in a compact-like wave as discussed above.

Future Directions

It is interesting to identify the conditions for a nonlinear equation to have solitary, or periodic, or compacton-like solutions. In most open literature, many papers on soliton and compacton are focused themselves on a special solution with either a soliton or a compacton without considering the initial/boundary conditions, which might be vital important for its actual wave type.

Solitons and compactons for difference-differential equations (e. g. Lotka–Volterra-like problems) have been caught much attention due to the fact that discrete space-time may be the most radical and logical viewpoint of reality (refer to E-infinity theory detailed concept). For small scales, e. g., nano scales, the continuum assumption becomes invalid, and difference equations have to be used for space variables.

Fractional differential model is another compromise between the discrete and the continuum, and can best describe solitons and compactons.

Many interesting phenomena arise in nanohydrodynamics recently, such as remarkably excellent thermal and electric conductivity, and extremely extraordinary fast flow in nanotubes. Consider a single compacton wave along a nanotube, and its wavelength is as same as the diameter of the nanotubes, under such a case, almost no energy is lost during the transportation, resulting in extremely extraordinary fast flow in the nanotubes.

The physical understanding of the transformation $k=iK$ is also worth further studying.

Cross References

► Soliton Perturbation

Bibliography

Primary Literature

1. He JH (2006) New interpretation of homotopy perturbation method. Int J Mod Phys 20(18):2561–2568

Some Famous Papers on Solitons and Compactons

2. Burger S, Bongs K, Dettmer S et al (1999) Dark solitons in Bose–Einstein condensates. Phys Rev Lett 83(25):5198–5201
3. Denschlag J, Simsarian JE, Feder DL et al (2000) Generating solitons by phase engineering of a Bose–Einstein condensate. Science 287(5450):97–101
4. Diakonov D, Petrov V, Polyakov M (1997) Exotic anti-decuplet of baryons: prediction from chiral solitons. Z Phys A 359(3):305–314
5. Duff MJ, Khuri RR, Lu JX (1995) Sting Solitons. Phys Rep 259(4–5):213–326
6. Eisenberg HS, Silberberg Y, Morandotti R et al (1998) Discrete spatial optical solitons in waveguide arrays. Phys Rev Lett 81(16):3383–3386
7. Fleischer JW, Segev M, Efremidis NK, et al (2003) Observation of two-dimensional discrete solitons in optically induced nonlinear photonic lattices. Nature 422(6928):147–150
8. He JH (2005) Application of homotopy perturbation method to nonlinear wave equations. Chaos Soliton Fract 26(3):695–700
9. He JH, Wu XH (2006) Construction of solitary solution and compacton-like solution by variational iteration method. Chaos Soliton Fract 29(1):108–113
10. Khaykovich L, Schreck F, Ferrari G et al (2002) Formation of a matter-wave bright soliton. Science 296(5571):1290–1293
11. Rosenau P (1997) On nonanalytic solitary waves formed by a nonlinear dispersion. Phys Lett A 230(5–6):305–318
12. Rosenau P (2000) Compact and noncompact dispersive patterns. Phys Lett A 275(3):193–203
13. Strecker KE, Partridge GB, Truscott AG et al (2002) Formation and propagation of matter-wave soliton trains. Nature 417(6885):150–153
14. Torruellas WE, Wang Z, Hagan DJ et al (1995) Observation of 2-dimensional spatial solitary waves in a quadratic medium. Phys Rev Lett 74(25):5036–5039

Review Article

15. He JH (2006) Some asymptotic methods for strongly nonlinear equations. Int J Mod Phys B 20(10):1141–1199 and 20(18):2561–2568

Exp-function Method

16. He JH, Wu XH (2006) Construction of solitary solution and compacton-like solution by variational iteration method. Chaos Soliton Fract 29(1):108–113
17. He JH, Wu XH (2006) Exp-function method for nonlinear wave equations. Chaos Soliton Fract 30(3):700–708
18. Zhu SD (2007) Exp-function method for the Hybrid–Lattice system. Int J Nonlinear Sci 8(3):461–464
19. Zhu SD (2007) Exp-function method for the discrete mKdV lattice. Int J Nonlinear Sci 8(3):465–468

Parameter-Expansion Method

20. He JH (2001) Bookkeeping parameter in perturbation methods. Int J Nonlinear Sci Numer Simul 2(3):257–264
21. He JH (2002) Modified Lindstedt–Poincare methods for some strongly non-linear oscillations Part I: Expansion of a constant. Int J Nonlinear Mech 37(2):309–314
22. He JH (2006) Non-perturbative methods for strongly nonlinear problems. dissertation.de-Verlag im Internet GmbH, Berlin
23. Shou DH, He JH (2007) Application of parameter-expanding method to strongly nonlinear oscillators. Int J Nonlinear Sci Numer Simul 8(1):121–124
24. Xu L (2007) Application of He's parameter-expansion method to an oscillation of a mass attached to a stretched elastic wire. Phys Lett A 368(3–4):259–262
25. Xu L (2007) Determination of limit cycle by He's parameter-expanding method for strongly nonlinear oscillators. J Sound Vib 302(1–2):178–184

Nanohydrodynamics and Nano-effect

26. He JH, Wan Y-Q, Xu L (2007) Nano-effects, quantum-like properties in electrospun nanofibers. Chaos Solitons Fract 33(1), 26–37
27. Majumder M, Chopra N, Andrews R et al (2005) Nanoscale hydrodynamics – Enhanced flow in carbon nanotubes. Nature 438(7064):44–44

E-Infinity Theory

28. El Naschie MS (2007) A review of applications and results of E-infinity theory. Int J Nonlinear Sci Numer Simul 8(1):11–20
29. El Naschie MS (2007) Deterministic quantum mechanics versus classical mechanical indeterminism. Int J Nonlinear Sci Numer Simul 8(1):5–10

Fractional-Order Differential Equations

30. Draganescu GE (2006) Application of a variational iteration method to linear and nonlinear viscoelastic models with fractional derivatives. J Math Phys 47(8):082902
31. He JH (1998) Approximate analytical solution for seepage flow with fractional derivatives in porous media. Comput Methods Appl Mech Eng 167(1–2):57–68
32. Momani S, Odibat Z (2007) Homotopy perturbation method for nonlinear partial differential equations of fractional order. Phys Lett A 365(5–6):345–350
33. Odibat ZM, Momani S (2006) Application of variational iteration method to Nonlinear differential equations of fractional order. Int J Nonlinear Sci Numer Simul 7(1):27–34
34. Wang Q (2007) Homotopy perturbation method for fractional KdV equation. Appl Math Comput 190(2):1795–1802

Solitons: Historical and Physical Introduction

François Marin
Laboratoire Ondes et Milieux Complexes, Fre CNRS 3102, Le Havre Cedex, France

Article Outline

Glossary
Definition of the Subject

Introduction
Historical Discovery of Solitons
Physical Properties of Solitons and Associated Applications
Mathematical Methods Suitable for the Study of Solitons
Future Directions
Bibliography

Glossary

Breaking waves As waves increase in height through the shoaling process, the crest of the wave tends to speed up relative to the rest of the wave. Waves break when the speed of the crest exceeds the speed of the advance of the wave as a whole.

Crystal lattice A geometric arrangement of the points in space at which the atoms, molecules, or ions of a crystal occur.

Deep water Water sufficiently deep that surface waves are little affected by the ocean bottom. Water deeper than one-half the surface wave length is considered deepwater.

Fluxon Quantum of magnetic flux.

Freak waves Single waves which result from a local focusing of wave energy. They are of considerable danger to mariners because of their unexpected nature.

Geostrophic adjustment The process by which an unbalanced atmospheric flow field is modified to geostrophic equilibrium, generally by a mutual adjustment of the atmospheric wind and pressure fields depending on the initial horizontal scale of the disturbance.

Geostrophic equilibrium A state of motion of an inviscid fluid in which the horizontal Coriolis force exactly balances the horizontal pressure force at all points of the field.

Hydraulic jump A sudden turbulent rise in water level, such as often occurs at the foot of a spillway when the velocity of rapidly flowing water is instantaneously slowed.

Katabatic wind Most widely used in mountain meteorology to denote a downslope flow driven by cooling at the slope surface during periods of light larger-scale winds.

Lightning Lightning is a transient, high-current electric discharge.

Plasma Hot, ionized gas.

Shallow water Water depths less than or equal to one half of the wavelength of a wave.

Solitary wave Localized wave that propagates along one space direction only, with undeformed shape.

Soliton Large-amplitude pulse of permanent form whose shape and speed are not altered by collision with other solitary waves, the exact solution of a nonlinear equation.

Spillway A feature in a dam allowing excess water to pass without overtopping the dam.

Thermocline A layer in which the temperature decreases significantly (relative to the layers above and below) with depth.

Synoptic scale Used with respect to weather systems ranging in size from several hundred kilometers to several thousand kilometers.

Thunder The sound emitted by rapidly expanding gases along the channel of a lightning discharge.

Thunderstorm In general, a local storm, invariably produced by a cumulonimbus cloud and always accompanied by lightning and thunder, usually with strong gusts of wind, heavy rain, and sometimes with hail.

Tidal bore Tidal wave that propagates up a relatively shallow and sloping estuary or river, in a solitary wave form. The leading edge presents an abrupt rise in level, sometimes with continuous breaking and often immediately followed by several large undulations. The tidal bore is usually associated with high tidal range and a sharp narrowing and shoaling at the entrance. Also called pororoca (Brazilian) and mascaret (French).

Troposphere The portion of the atmosphere from the earth's surface to the tropopause, that is the lowest 10–20 km of the atmosphere.

Tsunami Long period ocean wave generated by an earthquake or a volcanic explosion.

Definition of the Subject

The interest in nonlinear physics has grown significantly over the last fifty years. Although numerous nonlinear processes had been previously identified the mathematic tools of nonlinear physics had not yet been developed. The available tools were linear, and nonlinearities were avoided or treated as perturbations of linear theories. The solitary water wave, experimentally discovered in 1834 by John Scott Russell, led to numerous discussions. This hump-shape localized wave that propagates along one space-direction with undeformed shape has spectacular stability properties. John Scott Russell carried out many experiments to obtain the properties of this wave. The theories which were based on linear approaches concluded that this kind of wave could not exist. The controversy was resolved by J. Boussinesq [5] and by Lord Rayleigh [64] who showed that if dissipation is neglected, the increase in local wave velocity associated with finite amplitude is balanced by the decrease associated with dispersion, leading

to a wave of permanent form. A model equation describing the unidirectional propagation of long waves in water of relatively shallow depth with a localized solution representing a single hump as discovered by Russell was obtained by Korteweg and de Vries [42]. This equation has become very famous and is now known as the Korteweg–de Vries equation or KdV equation. These results were not considered very important when they were obtained. A remarkable numerical discovery was made by E. Fermi, J. Pasta and S. Ulam [18] as they studied the flow of incoherent energy in a solid modeled by a one-dimensional lattice of equal masses connected by nonlinear springs. The initial injected energy was not shared among all the degrees of freedom of the lattice, but returned almost entirely to the original excited mode. The explanation was given only ten years later by Zabusky and Kruskal [82] from numerical solutions of the KdV equation used to model in the continuum approximation a nonlinear atomic lattice with periodic boundary conditions. This led to the concept of solitons and ultimately to the development of integrable systems. The term soliton was chosen as it is a localized wave which propagates preserving its shape and velocity as a particle would propagate. A soliton is then a large-amplitude pulse of permanent form whose shape and speed are not altered by collision with other solitary waves, and it is the exact solution of a nonlinear equation. The mathematical features of solitons are often well developed in soliton literature since they are at the origin of very nice theoretical developments such as the powerful inverse scattering transform which solves a complex nonlinear equation through a series of linear steps. Nevertheless, the physics of solitons is very rich and it is a very actual research topic in numerous fields, for example in hydrodynamics, in optics, in electricity, in solid physics, in chemistry and in biology. Nonlinearity and dispersion are very common in macroscopic physics as well as in microscopic physics. Nonlinearity tends to localize the signals while dispersion spreads them. These two opposite effects sometimes compensate and the regime of solitons takes place. Real physical systems are only approximately described by the equations of the theory of solitons, but a remarkable feature of solitons is their very high stability relative to perturbations. The soliton concept is now firmly established. They are widely accepted as a structural basis for viewing and understanding the dynamic behavior of complex nonlinear systems.

Introduction

The first experimental observation of a solitary wave was made in August 1834 by a Scottish engineer named John Scott Russell (1808–1882). He was mounted on a horse along the Union Canal linking Edinburgh with Glasgow when he saw a rounded smooth well-defined heap detach itself from the prow of a boat brought to rest and continue its course without change of shape or diminution of speed. Scott–Russell reported his observation in the British Association Report [71] as follows:

"I was observing the motion of a boat which was rapidly drawn along a narrow channel by a pair of horses, when the boat suddenly stopped—not so the mass of water in the channel which it had put in motion; it accumulated round the prow of the vessel in a state of violent agitation, then suddenly leaving it behind, rolled forward with great velocity, assuming the form of a large solitary elevation, a rounded, smooth and well-defined heap of water, which continued its course along the channel apparently without change of form or diminution of speed. I followed it on horseback, and overtook it still rolling on at a rate of some eight or nine miles an hour, preserving its original figure some thirty feet long and a foot to a foot and a half in height. Its height gradually diminished, and after a chase of one or two miles I lost it in the windings of the channel. Such, in the month of August 1834, was my first chance interview with that singular and beautiful phenomenon which I have called the Wave of Translation".

Russell wrote that this August day in 1834 was the happiest in his life. In July 1995, an international gathering of scientists witnessed a re-creation of the solitary wave observed by Russell on an aqueduct of the Union Canal. This aqueduct was named after this occasion *Scott Russell aqueduct*; the scientists were attending a conference on nonlinear waves in physics and biology at Heriot-Watt University (Edinburgh), near the canal.

Throughout his life Russell remained convinced that "his" wave that he initially called the "Wave of Translation" and later the "Great Solitary Wave" was of fundamental importance. He therefore carried out extensive experiments in wave-tanks and established the velocity formula for solitary waves:

$$V_S = \sqrt{g(d + A_S)} \tag{1}$$

where g is the acceleration due to gravity, d the undisturbed water depth and A_S the amplitude of the solitary wave. The "quasi-cycloidal" form of the wave was found to be independent of the way it was generated. Russell [69] described the induced movement of fluid particles: "By the transit of the wave the particles of the fluid are raised from their places, transferred forwards in the direction of the motion of the wave, and permanently deposited at rest in a new place at a considerable distance from their original position". He also deduced from his numerous experiments that two solitary waves can cross

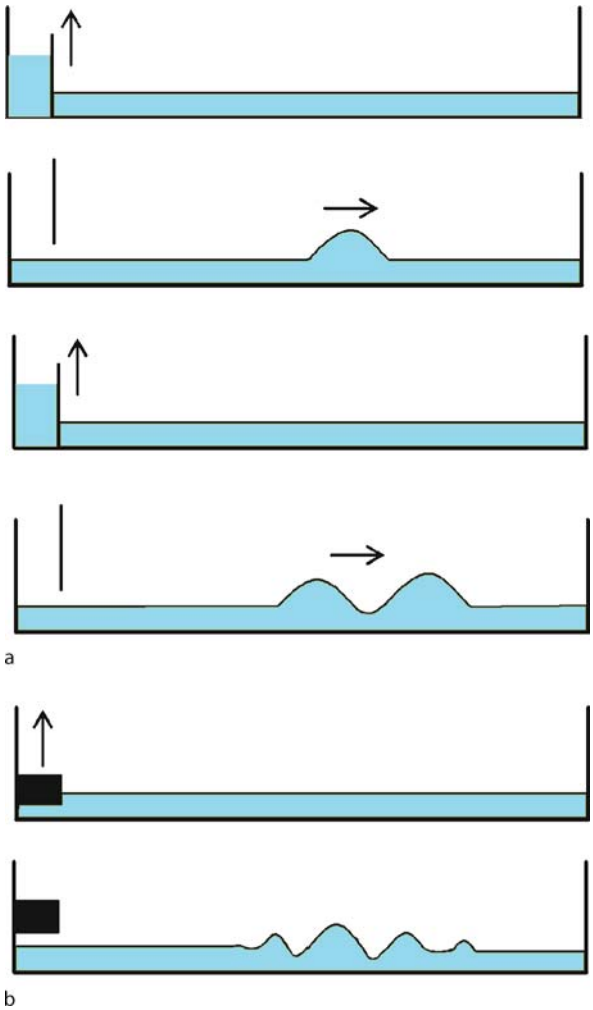

Solitons: Historical and Physical Introduction, Figure 1
Schematic view of Scott Russell's experiments

each other "without change of any kind". Figure 1 shows a schematic view of his experiments in tanks, adapted from Remoissenet [65]. An initial elevation of water may induce one or two solitary waves depending on the relation between its height and length (Fig. 1a); the case where an initial depression does not evolve into solitary waves but leads to an oscillatory wave train of gradually increasing length and decreasing amplitude is shown in Fig. 1b.

In Sect. "Historical Discovery of Solitons", the historical discovery of solitons is presented. The physical concept of solitons and the associated applications are described in Sect. "Physical Properties of Solitons and Associated Applications". In Sect. "Mathematical Methods Suitable for the Study of Solitons", the mathematical methods suitable for the study of solitons are considered. Finally, Sect. "Future Directions" is devoted to future directions.

Historical Discovery of Solitons

After Russell had carried out his observation of the solitary wave in August 1834, he was put in charge by the Union Canal Company of determining the efficiency of canals for the transport of steam-driven barges. He worked on the resistance of boats towed through water of finite depth. A resistance proportional to the square of the velocity V_b of the boat was predicted by the theory at that time. Convinced of the imperfection of this theory, Russell performed numerous experiments in the summers of 1834 and 1835 [12]. Several boats were towed in canals of various depths at velocities ranging from 3 to 15 miles per hour. The resistance was measured by a dynamometer. Resistance curves such as the one depicted in Fig. 2 were obtained, with a resistance local maximum R_m for a boat velocity V_{cr} and a growing deviation from the theoretical curve for increasing values of boat velocity. The critical velocity V_{cr} was found to depend on the water depth. Russell noticed that this critical velocity was the same as the velocity of the solitary wave for the given water depth d, and gave a qualitative explanation for the diminution of flow resistance for increasing values of the boat velocity just above the critical velocity V_{cr}. For this description, he considered the variation of the shape of the water surface around the moving boat with the velocity. For velocities smaller than the critical value, the water level is raised around the prow where a wave of displacement is generated, leading to an inclination of the boat. The resulting effect is an augmentation of the section of immersion and of the corresponding flow resistance. When the velocity of the boat reaches the critical value, the wave of displacement has the velocity of a solitary wave and the vessel travels on this solitary wave. For velocities greater than the critical value, the boat stays on the wave summit leading to a smaller flow resistance. Russell [70] tried a quantitative explanation of this process, but this paper contained several errors. However, Russell was a very fine observer; he was far ahead of his time in considering the fundamental importance of solitary waves. He mentioned [70] that "a large or high wave had a greater velocity than a small one. When a small wave preceded a large one, the latter invariably overtook the other, and when the large wave was before the less, their mutual distance invariably became greater". He showed that the shape of solitary waves depends on their height, the width of the wave decreasing for increasing values of height. It is also fascinating that Russell had observed the collision properties of solitary waves [8]. Nevertheless, the observations of Russell induced numerous critical comments. Airy [2] wrote in his treatise "Tides and Waves": "We are not disposed to recognize this wave (discovered by Scott

Russell) as deserving the epithets "great" or "primary", and we conceive that ever since it was known that the theory of shallow waves of great length was contained in the equation $\partial^2 X/\partial t^2 = gd\partial^2 X/\partial x^2$ the theory of the solitary wave has been perfectly well known". In the equation quoted by Airy, t is the time and x is the horizontal coordinate of a fluid particle. The wave velocity in shallow water at the first order of approximation, when the ratio H/d between the wave height H and the water depth may be neglected is $V_0 = \sqrt{gd}$, as found by Lagrange in 1786; this velocity differs from the formulation proposed by Russell for solitary waves (Eq. (1)). Airy mentioned further: "Some experiments were made by Mr. Russell on what he calls a negative wave. But (we know not why) he appears not to have been satisfied with these experiments and had omitted them in his abstract. All of the theorems of our IVth section, without exception, apply to these as well as to positive waves, the sign of the coefficient only being changed". Also, according to Airy [2] and Lamb [46], long waves could not propagate in a canal of rectangular section without changing their shape. The controversy arose because the dispersion is neglected by the nonlinear shallow water theory. Stokes [73] considered Russell results and tried to obtain them analytically; however, his linear or weakly nonlinear approach did not permit him to retrieve the results of Russell, which made him doubtful about them. In order to distinguish his "great wave" from other waves, Russell [71] introduced four orders of waves:

1. Waves of translation. These waves involve mass transfer. The solitary waves observed by Russell are included in this order.
2. Oscillatory waves. These are the waves which can be the most often observed; they do not involve mass transfer.
3. Capillary waves. The surface tension effects are important for those waves.
4. Corpuscular waves. Those waves are rapid successions of solitary waves.

Russell was most concerned with the first order, but he also carried out experiments with waves of the second and third order. The fourth order (corpuscular waves) suggested by Russell had been seriously criticized, and the manuscripts he submitted on this order were never published due to ignorance of mechanics principles.

The great difficulties encountered by Russell to prove the importance of his findings strongly disappointed him. Tired of discussions, he stopped his work on solitary waves and started to build large steam ships with great success. A detailed biography of Russell may be found in [16].

A French mathematician, Joseph Boussinesq, knew of the existence of Russell's observations, and also of detailed experiments performed by Bazin [3] in the long branch of the canal de Bourgogne close to Dijon (France), which confirmed Russell's observations. Boussinesq tried to obtain a solution of Euler's equations compatible with Russell's and Bazin's results. He developed the horizontal and vertical velocity components of fluid velocity in a rectangular channel in power of the distance from the bed. Including nonlinear terms which had been neglected by Lagrange, Boussinesq [5] obtained a solution with the properties of the solitary wave observed by Russell. In particular, the shape of the wave was found to be given by:

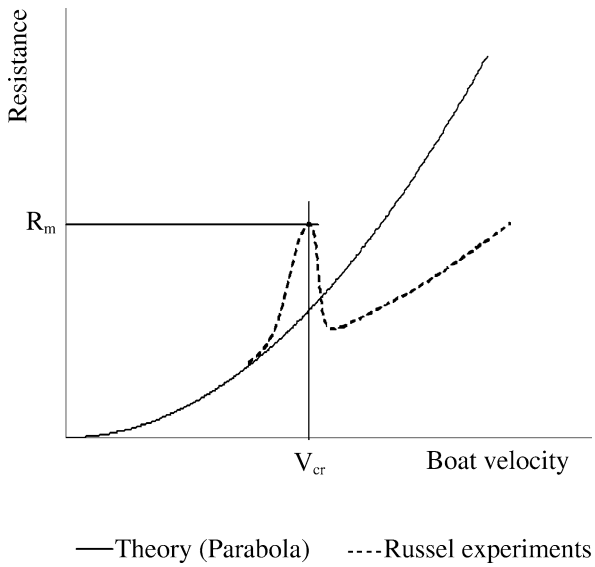

Solitons: Historical and Physical Introduction, Figure 2
Resistance as a function of towing velocity according to Russell

$$\eta = A_S \operatorname{sech}^2\left[\sqrt{\frac{3A_S}{4d^3}}(x - V_S t)\right] \tag{2}$$

where η is the free surface displacement and $\operatorname{sech} x = 1/\cosh x$; the velocity of the wave V_S is given by Eq. (1). The article of Boussinesq [5] had been almost ignored in England, and five years after its publication Lord Rayleigh [64] independently obtained the solitary wave profile. Following Rayleigh's work and including the effects of surface tension, the Amsterdam professor of mathematics Diederik Johannes Korteweg and his doctoral student Gustav de Vries derived a model equation [42] which describes the unidirectional propagation of long waves in water of relatively shallow depth. This famous equation is now known as the Korteweg–de Vries equation or KdV equation, and it has the following form:

$$\frac{\partial \eta}{\partial t} + V_0 \frac{\partial \eta}{\partial x} + \frac{3}{2}\frac{V_0}{d}\eta\frac{\partial \eta}{\partial x} + \frac{1}{6}V_0 d^2 \frac{\partial^3 \eta}{\partial x^3} = 0\,. \tag{3}$$

Korteweg and de Vries showed that this nonlinear wave equation has a localized solution which represents a single hump of positive elevation corresponding to the solitary wave observed by Scott Russell. The KdV equation had been in fact formulated in 1877 by Boussinesq in an impressive paper [6], but this author did not use directly this formulation. Russell's experiments had found their theory and one could think that many scientists would rapidly extend the results of this theory. However, the KdV equation had a quiet life for many decades.

In the early 1950s, the physicists Enrico Fermi, John Pasta and Stan Ulam could use one of the first computers, the "Maniac" to work on systems without closed analytic solutions. In particular, they studied the way a crystal evolves towards the thermal equilibrium through the numerical simulation of the dynamics of a one-dimensional lattice consisting of N equal masses (atoms) connected by nonlinear springs. The problem is described by the Hamiltonian

$$H = \sum_{i=0}^{N-1} \frac{1}{2} p_i^2 + \sum_{i=0}^{N-1} \frac{1}{2} K (x_{i+1} - x_i)^2 + \frac{K\alpha}{3} \sum_{i=0}^{N-1} (x_{i+1} - x_i)^3 \quad (4)$$

where x_i is the atom i displacement, p_i the corresponding momentum, K the constant of the quadratic potential, and $\alpha (\ll 1)$ a coefficient of nonlinear interaction. The boundary condition $x_0 = x_N = 0$ is assumed. Considering a normal mode decomposition

$$A_k = \sqrt{\frac{2}{N}} \sum_{i=0}^{N-1} x_i \sin(ik\pi/N) \,, \quad (5)$$

the Hamiltonian (Eq. (4)) may be written in the following way

$$H = \frac{1}{2} \sum_k \left(\dot{A}_k^2 + \omega_k^2 A_k^2 \right) + \alpha \sum_{k,l,m} c_{klm} A_k A_l A_m \quad (6)$$

where the pulsations ω_k are given by $\omega_k^2 = 4K \sin^2(k\pi/2N)$, and c_{klm} are constants. When the masses have very small displacements from the equilibrium positions, the dynamics of the lattice may be described by a superposition of normal modes. It was supposed that if only one mode (the fundamental mode $k = 1$ in the circumstances) is excited in a crystal lattice, nonlinear interactions would lead to energy equipartition or thermal equilibrium. This was suggested by the results at the beginning of the calculations since the modes $2, 3, \ldots$ were successively excited. However, after 157 fundamental periods, almost all the energy returned to the fundamental mode. Furthermore, the energy returned almost periodically to the original excited mode leading to a near recurrence process, now known as the FPU paradox. Fermi et al. [18] did not get the expected result and the modeled crystal lattice did not approach energy equipartition. This remarkable discovery is at the origin of the comprehension of the concept of the soliton and of the developments of "numerical experiments". Fermi died just after this work; this had a rather inhibitory effect on the significance of this result and no paper was published after the preprint "Studies of nonlinear problems" [18]. The results were communicated to the scientific community by Ulam through several conferences. Complementary studies [20,39] confirmed that the introduction of nonlinearity in a system does not guarantee the equipartition of energy. In 1960, the KdV equation was rederived by Gardner and Morikawa who worked on collisionless hydromagnetic waves. These authors attracted the attention of Zabrusky and Kruskal on the KdV equation, and the FPU paradox was solved a few years later in terms of solitons [82]. Zabrusky and Kruskal considered the following motion equations corresponding to the Hamiltonian (Eq. (4))

$$\ddot{x}_i = K(x_{i+1} + x_{i-1} - 2x_i) + K\alpha\left[(x_{i+1} - x_i)^2 - (x_i - x_{i-1})^2\right] . \quad (7)$$

Taking into account that the recurrence occurs before the excitation of high-order modes, these authors carried out an asymptotic analysis in the continuum approximation and obtained an equation including dispersive and nonlinear terms which correspond to the KdV equation. This equation, obtained in a totally different physical context as the one to explain the observations of Scott Russell, is at the origin of the explanation of the FPU paradox. As depicted in Fig. 3, the numerical experiments of Zabrusky and Kruskal showed that the initial sinusoidal condition evolves into steep fronts and then into a finite number of short pulses which are solitons. These solitons travel around the lattice with velocities depending on their amplitude; they collide preserving their individual shapes and velocities with only a small change in their phases. As time increases, there is an instant at which the solitons collide at the same point, and the initial state comes close to recurrence. The recurrence period calculated by Zabrusky and Kruskal closely approximates the actual FPU recurrence period, showing that the KdV equation is a suitable approximation of the FPU system. The system had to be considered as totally nonlinear to be solved, and the linear normal modes of the system could not be considered. The pulse-like waves were called solitons by Zabrusky and

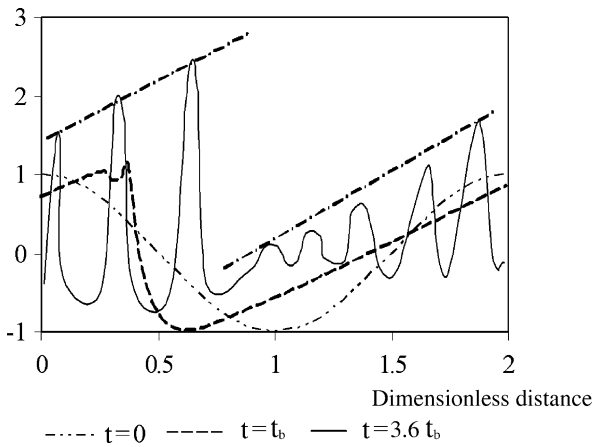

Solitons: Historical and Physical Introduction, Figure 3
Evolution of an initially periodic profile from Zabusky and Kruskal numerical analysis of the KdV equation. The breaking time for the wave profile is t_b

Kruskal to indicate the remarkable quasi-particle properties of these very stable solitary waves. They were at first called solitrons to introduce them into the family of particle names such as electron and neutron, but the name solitron was a trade mark, and was therefore not suitable. The concept of solitons has spread over numerous branches of physics, such as optical physics, biological physics, astronomy, particle physics, condensed matter, fluid dynamics, ferromagnetism. The numerical experiments of Zabrusky and Kruskal led to the development of integrable systems. Taniuti and Wei [76] and Su and Gardner [74] have shown that a large class of nonlinear evolution equations may be reduced to consideration of the KdV equation which may be regarded as a very important canonical form. A famous method, the so-called Inverse Scattering Transform (IST) was proposed by Gardner et al. [24] to integrate the KdV equation. This procedure may be considered as the nonlinear analogue of the Fourier transform method suitable to linear dispersive systems.

Physical Properties of Solitons and Associated Applications

Properties of Solitons

Let us consider the KdV equation (Eq. (3)). Introducing the variable $X = x - V_0 t$, this equation may be written in the following way

$$\frac{1}{V_0}\frac{\partial \eta}{\partial t} + \frac{3}{2d}\eta\frac{\partial \eta}{\partial X} + \frac{d^2}{6}\frac{\partial^3 \eta}{\partial X^3} = 0\,. \tag{8}$$

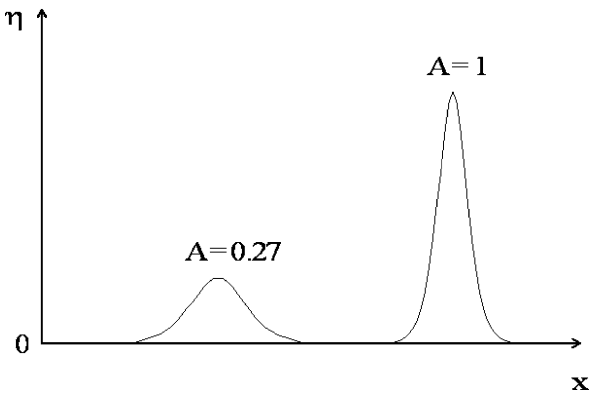

Solitons: Historical and Physical Introduction, Figure 4
Comparison between two solutions of the KdV equation with different amplitude *A*

Using the dimensionless variables $\phi = \eta/d$, $\xi = X/X_0$ and $\tau = t/t_0$ where X_0 and t_0 are respectively a typical length and time, the KdV equation may be formulated in its standard form:

$$\frac{\partial \phi}{\partial \tau} + 6\phi\frac{\partial \phi}{\partial \xi} + \frac{\partial^3 \phi}{\partial \xi^3} = 0\,. \tag{9}$$

This equation has many solutions, as all nonlinear equations. Among these solutions, there are the spatially localized solutions:

$$\phi = A\,\mathrm{sech}^2\left[\sqrt{\frac{A}{2}}(\xi - 2A\tau)\right], \tag{10}$$

where A is a positive constant. These solutions quantitatively correspond to the observations of Scott Russell. In particular, the width of the soliton decreases for increasing values of its amplitude, as depicted in Fig. 4. In dimensional variables and in a fixed referential, the corresponding solution is given by Eq. (2).

The KdV equation appears widely in physics, each time that waves propagate in a weakly nonlinear and weakly dispersive medium. The nonlinear term $\phi(\partial\phi/\partial\xi)$ in this equation causes the steepness of the wave form. In a dispersive medium, the Fourier components of a pulse propagate at different velocities, inducing a spreading of the pulse. The dispersive term $\partial^3\phi/\partial\xi^3$in the KdV equation makes the wave form spread. The soliton solution of the KdV equation results from the balance of nonlinearity and dispersion. This balance is generally very stable, which explains the numerous applications of the theory of solitons, even if the real physical situations (where friction and defects take place) are only approximately described by the soliton equations, and in particular by the KdV

equation. It would be strictly speaking more appropriate to use the terminology quasi-soliton in physics; however, bearing in mind that the word soliton has a quite ideal connotation in the context of systems with exact solutions, we will use the word soliton. The conservation of the shape and of the velocity of the soliton after a collision is a manifestation of stability. The high stability of solitons relative to perturbations, combined with the fact that they can spontaneously emerge in a physical system in which energy is supplied, lead to numerous applications of solitons in macroscopic and microscopic physics. Nevertheless, a soliton may sometimes vanish. For example, this is the case when such a wave propagates in a media where the water depth decreases during the wave propagation. The dispersive effect progressively decreases when the nonlinear effect increases, leading to the wave breaking as waves on a beach. The vanishing of a soliton may also result from energy dissipation. Marin et al. [52] have shown that sand ripples may form when solitons propagate in shallow water over a sandy bed, and that a strong interaction between the free surface and the bed occurs. This interaction leads to a decrease of the soliton amplitude, which may result under certain circumstances in the disappearance of the soliton. Let us now consider some of the most typical soliton applications.

Solitons in Fluid Mechanics

In the hydrodynamic field, a very nice case where solitons take place is the tidal bore, also called hydraulic jump in translation, and known as the mascaret in France. A tidal bore is a positive wave of translation which can occur in a river with a gently slopping bottom and a broad funnel-shaped estuary, where a difference of more than six meters between high and low water takes place [50]. The tidal bore appears as a single wave or a series of waves which propagate upstream. At a given instant, it is localized at the upstream extremity of the flood tide propagation in the inter-tidal zone. The flow direction is towards the sea before the tidal bore arrival, and may be in the opposite direction after its occurrence. Such bores are distributed widely throughout the world. The most powerful can be seen in the Amazon river in South America, where it is called pororoca, which means gigantic noise. An approximately 5 m-high wave propagates with a velocity of about 30 m/s, and surfers take advantage of this exceptional wave to organize competitions. In England, bores may be contemplated on several rivers, the most famous occurring in the Severn river, near Bristol. In France, bores take place mainly in the Mont Saint-Michel bay, in the Gironde, Dordogne and Garonne rivers. The mascaret of the Seine river

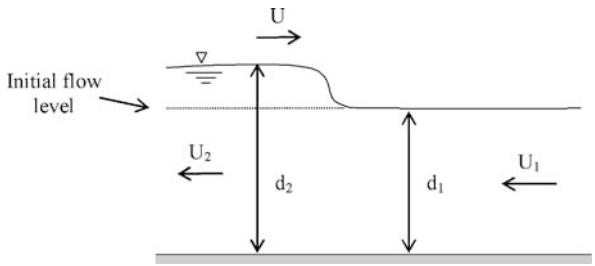

Solitons: Historical and Physical Introduction, Figure 5
Sketch of the propagation of a positive wave of translation

almost disappeared between 1960 and 1970 because of the construction of dikes and the dredging works which had been carried out in order for large boats to reach the city of Rouen. Increasing the water depth induces a decrease of the nonlinear effects which become insufficient to balance the dispersion. In Canada, several bores are observed in Fundy bay. In Asia, an important bore occurs in the Qiantang river, while the one in the river Pungue in Africa can propagate upstream over more than 80 km. Two types of bores may be observed: the breaking bore and the undulated bore, depending on the value of the Froude number F_r defined by $F_r = (U_1 + U)/\sqrt{gd_1}$, where U_1 is the flow velocity at the depth d_1 before the passage of the bore and U the bore velocity (Fig. 5). In Fig. 5, d_2 depicts the water depth after the passage of the bore and U_2 the flow velocity at the depth d_2. For a value of F_r lower than about 1.5, the bore is undulated, and for a value of F_r greater than 1.5, the breaking bore occurs. Most of the bores are undulated bores. The tidal bore is a very important turbulent phenomenon for the inter-tidal zone of a river, inducing a significant mixing of waters and complex motions of sediment on the river bed [14]. River bores have been known for many centuries but only qualitative observations were carried out. Quantitative measurements came much later, mainly because solitons are inherently nonlinear when only linear processes were considered.

Long waves such as tsunamis often behave like solitary waves [48]. In particular, the run-up and shoreward inundation are often simulated using solitary waves [49]. The tsunamis may be described by the KdV equation. The theory of very long shallow water waves is valid since the tsunamis have a wavelength which is generally greater than 100 km and they propagate in oceans whose mean depth is about 4000 m. Their propagation velocity is greater than $V_0 = \sqrt{gd}$, that is greater than approximately 700 km/h. When they approach a coast, their velocity decreases and their height increases to satisfy the equation of energy conservation, and they finally break inducing often

very significant damages on the shore. Most tsunamis have a seismic origin, such as the dramatic one which occurred in the Indian Ocean on 26 December 2004 and which led to the death of more than 220 000 people. These solitons may also result from a quick arrival in the sea of volcanic products, through a similar process to that in Scott Russell's experiments depicted in Fig. 1. This was the case for the tsunami generated in 1883 by the eruption of Krakatau in Indonesia, when approximately 36 000 people died. The inundation phase for a tsunami highly depends on the bottom bathymetry. The breaking can be progressive and structural damages are mainly caused by inundation [13]. The breaking can also be explosive and induce the formation of a plunging jet. When the breaking takes place in very shallow water, the tsunami amplitude becomes so high that an undulation appears on the long wave, which develops into a progressive bore [10]. This turbulent front is of the same type as the wave which occurs when a dam breaks, and the water can rise very rapidly, typically from 0 to 3 meters in 1.5 minutes. The event of December 2004 has shown that the available models are not able to predict accurately the wave run-up heights under severe conditions.

Helal and El-Eissa [37] presented the connection between shallow water waves and the KdV equation. More recently, Mei and Li [55] considered the propagation of long waves over a randomly rough seabed. These authors have shown analytically and numerically that, assuming a randomness height to mean depth ratio comparable to the one of mean depth to the characteristic wavelength, disorder causes diffusion, leading to spatial attenuation of the wave amplitude. Mei and Li [55] proposed a modified KdV equation which includes terms representing the effects of disorder on amplitude attenuation. After a propagation over a region of finite length, the transmitted wave is a pulse which disintegrates into several small solitons of vanishing energy.

Hydrodynamic solitons may also be observed in deep water. Stokes showed in 1847 through the use of a small amplitude expansion of a sinusoidal wave, that periodic waves of finite amplitude are possible in deep water. The Stokes waves are unstable to infinitesimal modulation perturbations [4]. Zakharov and Shabat [83] were the first to show that an initial wave packet may evolve into a number of envelope solitons and a dispersive tail, the envelope solitons consisting of a carrier wave modulated by an envelope signal. This was experimentally verified by Yuen and Lake [81], by carrying out detailed experiments in deep water in a wave tank. The stability properties of envelope solitons were also considered. The freak waves, also called rogue waves, extreme or giant waves, may also have soliton-like shape. These waves appear surprisingly in the sea ("wave from nowhere") in deep or shallow water, and can cause severe damages to ships and fixed ocean structures. They are characterized by their brief occurrence in space and time, resulting from a local focusing of wave energy [19]. The main features of their physical processes may be obtained performing numerical simulations in the framework of classical nonlinear evolution equations, such as the KdV equation [41].

Oceanographers have also observed internal solitary waves in many regions around the world's oceans [61]. Most of the observed solitons propagate at the interface between the thermocline and the deep sea [56]. They are mainly excited by tidal flows over bottom topography [25], and they contribute to significant vertical mixing. Brandt et al. [7] developed a rotationless Boussinesq-like model for the generation and propagation of internal waves; the results of the simulations are in good agreement with large solitary waves recorded with airborne Radar images. Michallet and Barthélemy [56] carried out experiments in a 3-m-long flume to study interfacial long-waves in a two-immiscible-fluid system involving water and petrol. The comparison between the experiments and nonlinear theories shows that the KdV solitary waves match the experiments for small amplitude waves for all layer thickness ratios. When the crest is close to the critical level, that is approximately located at mid-depth, the large amplitude waves are well predicted by a modified KdV equation including both quadratic and cubic nonlinear terms (KdV-mKdV equation). However, only very few laboratory or field measurements of internal solitary waves are available, due to the prohibitive cost of obtaining precise measurements in the field and to the difficulty of generating and measuring these waves in the laboratory.

Blood pressure pulses may be considered as solitons [63]. There is a balance between the nonlinearity due to the blood flow and the dispersion connected to the elasticity of the artery. Previous studies [32,33,80] have shown that the dynamics of weakly nonlinear pressure waves in a thin nonlinear elastic tube filled with an incompressible fluid may be governed by the KdV equation. A Boussinesq-like equation was obtained by Paquerot and Remoissenet [62] to describe the blood pressure propagation in large arteries, in the limit of an ideal fluid and for slowly varying arterial parameters.

Several methods have been tested to generate hydrodynamic solitons in a flume; the method used by Scott Russell has been described in Sect. "Introduction" (Fig. 1). Hammack and Segur [29] showed experimentally and theoretically that from any net positive volume of water above the still water level, at least one solitary wave will emerge fol-

lowed by a train of dispersive waves. Maxworthy [54] and Weidman and Maxworthy [79] pushed the liquid in the horizontal direction by a single displacement of a piston. The generation of solitary waves using a piston-type wave maker has been studied in detail by Goring [26]. Guizien and Barthélemy [28] have proposed an experimental procedure to generate solitary waves in a flume using a piston type wave maker from Rayleigh's (1876) solitary wave solution. The advantage of this method is that solitary waves are rapidly established with very little loss of amplitude in the initial stage of the propagation of the solitary waves. The generation of solitons using a wave flume in resonant mode, without an absorbing beach, has been considered by Marin et al. [52]. Surface waves are produced in shallow water by an oscillating paddle at one end of the flume; a near-perfect reflection takes place at the other end. The frequency of the oscillating paddle was chosen close to the resonant frequency of the mode whose wavelength is equal to the flume length. For small values of the amplitude of displacement of the oscillating paddle, only standing harmonic waves are generated in the flume. For values of this amplitude greater than a critical value, pulses propagating from one end of the channel to the other end are excited on the background of the standing harmonic wave. From one to four pulses propagating in each direction of the flume could be generated over the time period of the flow, depending on the frequency and the amplitude of horizontal displacement of the oscillating paddle. These pulses were identified as solitons (one pulse) and bound states of solitons (several pulses). The generation of solitons in the flume results from the excitation of high harmonics. The spatiotemporal properties of solitons generated in this way were studied in detail by Ezersky et al. [17]. Space-time diagrams have been constructed to highlight the spatiotemporal dynamics of nonlinear fields for two solitons colliding in the resonator. Period doublings and the multistability of the nonlinear waves, i. e. the occurrence of different regimes for the same values of the control parameters but under different initial conditions, have been shown. It is important to keep in mind that the solitary waves correspond to the limit of cnoidal waves of period tending to infinity. The velocity of cnoidal waves V_{cn} is known to depend on the so-called elliptic parameter m and is given by the formula [11]

$$V_{\mathrm{cn}} = \left[1 - \frac{A_{\mathrm{S}}}{2d} + \frac{A_{\mathrm{S}}}{md}\left(1 - \frac{3E(m)}{2K(m)}\right)\right]\sqrt{gd} \qquad (11)$$

where $K(m)$ and $E(m)$ are complete elliptic functions of the first and second kind. The parameter m, which depends on the wave amplitude to width ratio, is responsible for the shape of the wave with $m = 0$ corresponding to the harmonic wave, and $m = 1$ to the soliton. Strictly speaking, the waves generated in a flume used in resonant mode are not exactly solitons, but cnoidal waves with an elliptic parameter very close to 1; the value of this parameter was found to be 0.9996 for the tests carried out by Ezersky et al. [17]. The shapes of pulses are almost indistinguishable for $m = 1$ and $m = 0.9996$, but the pulse repetition rates differ appreciably, the period tending to infinity as $m \to 1$.

Atmospheric solitons also exist, such as the Morning Glory Cloud of the Gulf of Carpentaria in northern Australia, which can be seen as a spectacular propagating roll cloud. Atmospheric solitons are horizontally propagating nonlinear internal gravity waves that can travel over large distances with minimal change in form [68]. The solitary waves that have been observed in the atmosphere can be divided into two classes: those that propagate in a fairly shallow stratified layer near the ground and those that occupy the entire troposphere. The generation mechanisms appear to be quite different for these two classes of waves. The waves that occupy the lower part of the troposphere mainly involve a gravity current such as a thunderstorm outflow, a katabatic wind, a sea breeze front, or a downslope windstorm, interacting with a low-level stable layer. The motion of the gravity current produces perturbations that are trapped in the low-level stable layer and eventually evolve over time into a series of solitary waves. For the tropospheric scale waves, they are very probably generated by synoptic scale features such as large-scale convective systems and geostrophic adjustments. For the two types of atmospheric solitons, there must exist some feature in the atmosphere that serves to prevent the wave energy from propagating away in the vertical direction. These trapping mechanisms are either deep layers of the atmosphere with very low values of the buoyancy frequency, or critical layers. The KdV equation, eventually enhanced with higher-order nonlinearity, gives a reasonable agreement with the observations [68].

Solitons in Nonlinear Transmission Lines

In another area of physics, transmission lines with nonlinear elements are found to propagate in a soliton mode. Nonlinear electrical transmission lines are simple experimental devices to observe and consider quantitatively the propagation and properties of solitons. The propagation of these waves is used for picosecond impulse generation and broadband millimeter-wave frequency multiplication [66]. Let us consider the elementary LC network depicted in Fig. 6, with linear inductors L and nonlinear capacitors C. The differential capacitance $C(V_n)$ is supposed to depend nonlinearly on the voltage across the nth

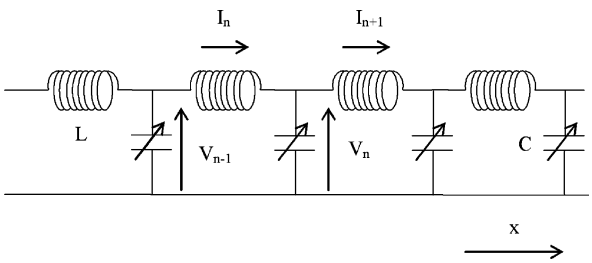

Solitons: Historical and Physical Introduction, Figure 6
Elementary circuit of an electrical network with linear inductance's *L* and nonlinear capacitance's *C*

capacitor V_n:

$$C(V_n) = \frac{\mathrm{d}Q_n(V_n)}{\mathrm{d}V_n} \tag{12}$$

where $Q_n(V_n)$ is the charge stored in the nth capacitor. Following Kirchhoff's law, we have:

$$V_{n-1} - V_n = \frac{\mathrm{d}\phi_n}{\mathrm{d}t}\,, \quad I_n - I_{n+1} = \frac{\mathrm{d}Q_n}{\mathrm{d}t} \tag{13}$$

where the magnetic flux ϕ_n is related to the current I_n by the relation: $\phi_n = LI_n$. Using Eqs. (12) and (13), it is easy to obtain:

$$\frac{\mathrm{d}^2 Q_n}{\mathrm{d}t^2} = \frac{1}{L}(V_{n+1} + V_{n-1} - 2V_n)\,, \quad n = 1, 2, \ldots \tag{14}$$

Assuming the following capacitance-voltage relation,

$$C(V_n) = C_0(1 - 2aV_n) \tag{15}$$

where a is a small nonlinear coefficient, we find:

$$LC_0 \frac{\mathrm{d}^2 V_n}{\mathrm{d}t^2} - LC_0 a \frac{\mathrm{d}^2 V_n^2}{\mathrm{d}t^2} = (V_{n+1} + V_{n-1} - 2V_n)\,, \quad n = 1, 2, \ldots \tag{16}$$

The system (Eq. 16) of nonlinear equations cannot be solved analytically. The continuum limit is used to get approximate solutions; setting the position $x = n\delta$ where δ is a hypothetical cellule length, we get

$$\frac{\partial^2 V}{\partial t^2} - \frac{\delta^2}{LC_0}\frac{\partial^2 V}{\partial x^2} = \frac{\delta^4}{12LC_0}\frac{\partial V}{\partial x^4} + a\frac{\partial^2 V^2}{\partial t^2}\,. \tag{17}$$

This weakly dispersive and nonlinear wave equation describes waves that can travel both to the left and to the right; the dispersive nature is due to the discreteness of the electrical network. A localized wave solution of the Eq. (17) that does not change its shape as it propagates with constant velocity v may be found [65]:

$$V = \frac{3(v^2 - v_0^2)}{2av^2}\,\mathrm{sech}^2\left(\frac{\sqrt{3(v^2 - v_0^2)}}{v_0}\left(n - \frac{v}{\delta}t\right)\right) \tag{18}$$

where $v_0 = \delta/\sqrt{LC}$. This solitary wave solution represents a pulse with amplitude

$$V_m = \frac{3}{2a}\frac{v^2 - v_0^2}{v^2} \tag{19}$$

which depends on the velocity v. The width L_w at half height of this pulse is given by

$$L_w = 1.76\frac{v_0}{\sqrt{3(v^2 - v_0^2)}}\,. \tag{20}$$

The waveforms of this solitary wave and of the KdV soliton are similar, and the width L_w depends on the amplitude V_m:

$$V_m L_w^2 = C^t \tag{21}$$

where C^t is a constant.

Solitons in Plasmas

Solitons may propagate in plasmas; this has received much attention because of a possible relevance to final state configurations in fusion devices. The combination of dispersion and nonlinearity in plasmas may be described by the KdV equation. A direct analogy between the equations for shallow water and those for plasmas has been mentioned by Gardner and Morikawa [23]. However, the KdV equation is not always adapted to the excitations which can be generated in plasmas by electromagnetic waves, since wave packets are often produced instead of pulses characterizing the solutions of the KdV equation. Another type of soliton equation is then more suitable, the nonlinear Schrödinger equation (NLS); this second class of soliton equation will be considered in Subsect. "Solitons in Optical Fibers".

Solitons in a Chain of Pendulums

Hydrodynamic solitons, solitons in nonlinear transmission lines are nontopological solitons, because the system returns to its initial state after the passage of the wave. There is also another type of soliton, the kink solitons which are topological solitons, since the structure of

the system is sometimes modified after the passage of the wave. A mechanical system consisting of a chain of pendulums, each pendulum being elastically connected to its neighbors by springs, is a typical example for which topological solitons can merge. It is easy to show that this mechanical system may be described by the following equation:

$$\frac{\partial^2\theta}{\partial t^2} - c_0^2\frac{\partial^2\theta}{\partial x^2} + \omega_0^2 \sin\theta = 0 \tag{22}$$

where θ is the angle of rotation of the pendulums, x is the axis along which the pendulums are distributed, $c_0^2 = l_1^2\beta/I_i$, l_1 being the distance between two pendulums, β the torque constant of a section of spring between two pendulums, I_i the moment of inertia of a single pendulum of mass m and length l_2, and $\omega_0^2 = mgl_2/I_i$. The first term in Eq. (22) represents the inertial effects of the pendulums, the second term corresponds to the restoring torque due to the coupling between pendulums, and the third term represents the gravitational torque. The Eq. (22) is called the Sine–Gordon (SG) equation [47]; it can be totally integrated and admits exact solitons solutions. It has become a very famous equation containing dispersion and nonlinearity which is used to model various phenomena in physics. It constitutes the third class of nonlinear equation leading to solitons.

Fluxons in a Josephson Tunnel Junction

Solitons arise also in a Josephson tunnel junction, which is a junction between two superconductors. These junctions are very attractive for information processing and storage [78]. The physical quantity of interest in the long Josephson junction is a quantum of magnetic flux, or a fluxon, which has a soliton behavior. The relevant nonlinear equation which describes the physical processes is the Sine–Gordon equation. This equation is particularly suitable in solid physics.

Solitons in Optical Fibers

One of the fundamental applications of solitons concerns optical fibers. In optical fiber communication using linear pulses, dispersion and losses in the fiber limit the information capacity which can be transported and the distance of transmission. Hasegawa and Tappert [31] first proposed to balance the dispersive effect by nonlinearity using soliton-based communications. The Kerr effect, in other words the nonlinear change of the refractive index of the fiber material, is used and an initial optical pulse may form a nonlinear stable pulse, an optical soliton, which is in fact an envelope soliton. Hasegawa and Tappert proposed the previously mentioned nonlinear Schrödinger equation for the description of solitons propagating in optical fibers. The first observations of these solitons were made in 1980 by Mollenauer, Stolen, and Gordon [58]. In real media, the light intensity of the soliton decreases due to losses in the fiber, and suitable reshaping of the pulse is required. Numerous theoretical and experimental studies on nonlinear guided waves have been carried out, because of their wide range of applications [30,43]. Inserting an optical fiber in a laser cavity, a "soliton laser" may be obtained, a femtosecond laser which is now in use. A real revolution in international communications is taking place at present, based on optical soliton technology. It is anecdotal that a fiber-optic cable linking Edinburgh and Glasgow now runs beneath the path from which John Scott Russell made his initial observations, and along the aqueduct which now bears his name.

The NLS equation may be obtained from the Sine–Gordon equation. Previously considered systems as water waves on the free surface or chains of pendulums have also weak amplitude solutions, plane waves which have the form:

$$\theta = \psi e^{i(qx-\omega t)} + \text{c.c.} \tag{23}$$

where ψ is the wave amplitude, q is the wave number, ω is the pulsation and c.c. denotes the complex conjugate of a complex number. When the wave amplitude is sufficiently increased for the nonlinearity to take place, modulation can spontaneously arise because harmonics are generated by the nonlinearity. The initial wave may split in waves packets whose properties correspond to the ones of solitons. These solitons consisting of a carrier wave which is modulated by an envelope signal are called envelope solitons. It can be shown [63] that the evolution of the envelope ψ is described by the NLS equation:

$$i\frac{\partial\psi}{\partial t} + P\frac{\partial^2\psi}{\partial x^2} + Q\,|\psi|^2\,\psi = 0 \tag{24}$$

where P and Q are coefficients which depend on the physical problem, as the significance of the variables t and x. The NLS equation is formally similar to the Schrödinger equation of quantum mechanics:

$$i\hbar\frac{\partial\psi}{\partial t} + \frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} - U\psi = 0\,. \tag{25}$$

The potential U in Eq. (25) is proportional to the absolute square of the wave envelope ψ, m is the mass of the quasiparticle and $\hbar = h/2\pi$ where h is the Planck constant. The potential represents the self-trapping of the wave energy.

Solitons in Solid Physics

Solitons are involved in the atomic structure of matter. At the beginning of the twentieth century, a central question in solid state physics was how a plastic deformation of a metal takes place in the corresponding crystal lattice. This problem induced numerous discussions, and it was suggested that if one atom in the lattice is pulled away, a neighbor atom follows the one which had been moved, jumping into the new hole, that is the new free lattice space. Then, a dislocation runs through the crystal. This proposition was improved in 1939 by Frenkel and Kontorova [22] who developed a model where not only each atom moves alone over one lattice distance, but also many atoms form a dislocation line generating a wave of translation. The process of traveling of the dislocation line through the crystal was assumed to occur without losses of energy. This is equivalent to Russell's solitary wave of translation which travels over long distance without change of form and velocity. This led Frenkel and Kontorova to make an analogy between the dislocation line and the soliton, and they obtained the one-soliton solution of the Sine–Gordon equation.

The solitons permit us to interpret properties of dielectric materials. In other respects, magnetic materials are interesting examples to experimentally verify in a very accurate way the theory of solitons at the atomic scale [57]. The concept of soliton in polymer physics constitutes a very nice case of an interdisciplinary approach. Their occurrence was suggested in 1988 by theoretician physicists [34]. Many studies have since then been carried out, in chemistry and experimental physics. Alan J. Heeger, Alan G. MacDiarmid and Hideki Shirakawa received the Nobel prize in chemistry in October 2000 for their works on the electric conduction of conductive plastics which is performed by solitons.

Solitons in Biology

Interest in the physical modeling of biological processes has grown significantly since the beginning of the 1990s. Nonlinear localization phenomena have been recently shown in systems very close to biological systems [15]. The notion of the soliton is now used to explain the dynamics of biological macromolecules such as proteins and the molecule ADN; an approach to the dynamics of the molecule ADN can be obtained using the envelope solitons solution of the NLS equation [63]. Intense researches are presently being carried out on this subject. The great size of biological molecules allows collective behaviors of atom groups, which are with nonlinearity important components for the existence of solitons.

Mathematical Methods Suitable for the Study of Solitons

Solitons may appear in many fields, such as fluid mechanics, solid state physics, plasma physics, optical fibers and biology, as shown in Sect. "Physical Properties of Solitons and Associated Applications". The description of these physical systems often leads to dispersive and nonlinear equations which are not presently solvable. It is then important to refer to the great classes of soliton equations, which are idealized models for numerous systems. These great classes of soliton equations correspond to the Korteweg–de Vries equation, known as the KdV equation (Eq. (9)), the Sine–Gordon equation (SG; Eq. (22)), and the nonlinear Schrödinger equation (NLS, Eq. (24)). These three famous equations can be totally integrated. They are not completely dissociated. For instance, it is possible to obtain the NLS equation from the SG equation and from the KdV equation. The soliton solutions represent only a first approximation of the physical properties of real systems. Dissipation or a weak spatial or temporal variation of the physical parameters in real physical systems are examples of phenomena which are very common and which are not taken into account in the soliton equations. The modeling has to be adapted to the physical system and to the excitation conditions of this system. In the case of plasmas, the KdV equation is more suitable for a description of the system when intense pulses excite it, and the NLS equation is more adapted for sinusoidal perturbations.

There are two broad theoretical approaches available to obtain the solutions of the great classes of soliton equations, the analytic and the algebraic methods. Analytic approaches include the powerful inverse scattering transform (IST) [24] and the remarkable Hirota method [38]. One-soliton solutions or multi-soliton solutions can be obtained. Moreover, the IST method can predict the number of solitons that can emerge from an initial disturbance applied to a physical system [59]. This method shows how the solitons have a role in nonlinear normal modes as the Fourier modes for a linear equation [44]. There are also the Bäcklund transformation method [67], the direct linearizing transform method [1], the Painlevé analysis [45], and the method of position-like solutions [40]. The algebraic methods involve the Lie group theoretic method [60], the direct algebraic method [53], and the tangent hyperbolic method [51].

The theoretical methods cannot always be used since they need some conditions to make them applicable. Numerical methods can be applied [36]. Zabusky and Kruskal [82] were the pioneers in studying in 1965 the KdV equation numerically, by using the leapfrog method

as an explicit finite difference scheme. Many methods have since then been used: a split-step Fourier method [77], the hopscotch method [27], a pseudo-spectral method [21], spectral methods (the Galerkin and Chebyshev methods) [35,72], the finite element method [75], and Fourier collocation calculations [9].

Future Directions

Solitons may occur in the macroscopic and microscopic scales of physics. This results from their very general existence conditions, the coexistence of dispersion and nonlinearity. The concept of solitons is not restricted to a specific field of physics, and the investigation of solitons remains a very active research area. We have previously mentioned (Sect. "Physical Properties of Solitons and Associated Applications") that one important application of solitons concerns optic fibers. The propagation of optical solitons through nonlinear fibers has been extensively studied; however, in real media, the dynamics of these solitons and the conditions for their generation are significantly affected by various inhomogeneities in the media. The problem of nonlinear wave propagation in the form of solitons in inhomogeneous optical fiber media is actually not well understood, despite the wide range of applications [30]. The investigation of nonlinear processes in coupled optical waveguides is another research direction for the future design of optical computers and sensor elements. Recent advances have shown that solitons have a great potential for the improvement of optical systems which demand fast and reliable data transfer. In the hydrodynamic field, the tsunamis often behave like solitary waves. The dramatic tsunami which occurred in the Indian Ocean on 26 December 2004 has clearly shown that the actual numerical models do not accurately predict the water propagation on the coastal plains. An important direction for future research is the (3D) three-dimensional simulation of tsunami breaking along a coast. The 3D simulation of the interaction of a tsunami with different beach profiles, with and without obstacles has also to be considered. This will permit the design of protecting devices and the setting up of security zones. In biology, nonlinear localization phenomena have been proved in model systems close to biological systems, but the challenge is open for biological molecules. Other important application fields for solitons concern magneto-electronics and secure communications. The formation of spatio-temporal patterns on perturbations of soliton systems are important problems to be studied in the years to come. The existence of instabilities in the solitary wave solutions of the great classes of soliton equations have to be tackled. Nonlinear excitations in systems whose spatial dimension is greater than one raise many questions. Among these questions, there are the problems of the possible stable structures, their collision properties and the effect of external forces. New applications will certainly emerge from these studies.

Bibliography

Primary Literature

1. Ablowitz MJ, Clarkson PA (1991) Solitons, nonlinear evolutions equations and inverse scattering. Cambridge University Press, Cambridge
2. Airy GB (1845) Tides and waves. Encycl Metropolitana 5:291–396
3. Bazin H (1865) Recherches expérimentales relatives aux remous et à la propagation des ondes. In: Darcy H and Bazin H (eds) Recherches hydrauliques, imprimerie impériale, Paris
4. Benjamin TB, Feir JE (1967) The disintegration of wavetrains on deep water. Part 1 Theory. J Fluid Mech 27:417–430
5. Boussinesq J (1871) Théorie de l'intumescence liquide appelée onde solitaire ou de translation, se propageant dans un canal rectangulaire. CR Acad Sci 72:755–759
6. Boussinesq J (1877) Essai sur la théorie des eaux courantes. MSE 23:1–680
7. Brandt P, Rubino A, Alpers W, Backhaus JO (1997) Internal waves in the Strait of Messina studied by a numerical model and synthetic aperture radar images from the ERS 1/2 satellites. J Phys Oceanogr 27:648–663
8. Bullough RK (1988) The Wave Par Excellence, the solitary progressive great wave of equilibrium of the fluid: an early history of the solitary wave. In: Lakshmanan M (ed) Solitons: Introduction and applications. Springer Ser Nonlinear Dyn. Springer, New York, pp 150–281
9. Canuto C, Hussaini MY, Quarteroni A, Zang TA (1988) Spectral methods in fluid dynamics. Springer, Berlin
10. Chanson H (2005) Le tsunami du 26 décembre 2004: un phénomène hydraulique d'ampleur internationale. Premiers constats. Houille Blanche 2:25–32
11. Cooker MJ, Weidman PD, Bale DS (1997) Reflection of high amplitude wave at a vertical wall. J Fluid Mech 342:141–158
12. Darrigol O (2003) The spirited horse, the engineer, and the mathematician: water waves in nineteenth-century hydrodynamics. Arch Hist Exact Sci 58:21–95
13. Dias F, Dutykh D (2007) Dynamics of tsunami waves. Extreme Man-Made and Natural Hazards in Dynamics of Structures. In: Ibrahimbegovic A, Kozar I (eds) Proc NATO Adv Res Workshop on Extreme Man-Made and Natural Hazards in Dynamics of Structures. Springer, Opatija, Croatia
14. Donnelly C, Chanson H (2002) Environmental impact of a tidal bore on tropical rivers. In: Proc 5th Int River Management Symp. Brisbane, Australia
15. Edler J, Hamm P (2002) Self-trapping of the amide I band in a peptide model crystal. J Chem Phys 117:2415–2424
16. Emerson GS (1977) John Scott Russell: a great Victorian engineer and naval architect. John Murray, London
17. Ezersky AB, Polukhina OE, Brossard J, Marin F, Mutabazi I (2006) Spatiotemporal properties of solitons excited on the surface of shallow water in a hydrodynamic resonator. Phys Fluids 18:067104

18. Fermi E, Pasta J, Ulam S (1955) Studies of nonlinear problems. Los Alamos report, LA-1940. published later In: Segré E (ed)(1965) Collected Papers of Enrico Fermi. University of Chicago Press
19. Fochesato C, Grilli S, Dias F (2007) Numerical modelling of extreme rogue waves generated by directional energy focusing. Wave Motion 44:395–416
20. Ford JJ (1961) Equipartition of energy for nonlinear systems. J Math Phys 2:387–393
21. Fornberg B, Whitham GB (1978) A numerical and theoretical study of certain nonlinear wave phenomena. Philos Trans R Soc London 289:373–404
22. Frenkel J, Kontorova T (1939) On the theory of plastic deformation and twinning. J Phys 1:137–149
23. Gardner CS, Morikawa GK (1960) Similarity in the asymptotic behaviour of collision-free hydromagnetic waves and water waves. Technical Report NYO-9082, Courant Institute of Mathematical Sciences. New York University, New York
24. Gardner CS, Green JM, Kruskal MD, Miura RM (1967) Method for solving the Korteweg-de Vries equation. Phys Rev Lett 19:1095–1097
25. Gerkema T, Zimmerman JTF (1994) Generation of nonlinear internal tides and solitary waves. J Phys Oceanogr 25:1081–1094
26. Goring DG (1978) Tsunamis – The propagation of long waves onto a shelf. Ph D thesis. California Inst Techn, Pasadena, California
27. Greig IS, Morris JL (1976) A hopscotch method for the KdV equation. J Comput Phys 20:64–80
28. Guizien K, Barthélemy E (2002) Accuracy of solitary wave generation by a piston wave maker. J Hydraulic Res 40(3):321–331
29. Hammack JL, Segur H (1974) The Korteweg-de Vries equation and water waves. Part 2. Comparisons with experiments. J Fluid Mech 65:289–314
30. Hao R, Li L, Li ZH, Xue W, Zhou GS (2004) A new approach to exact soliton solutions and soliton interaction for the nonlinear Schrödinger equation with variable coefficients. Opt Commun 236:79–86
31. Hasegawa A, Tappert F (1973) Transmission of stationary nonlinear optical pulses in dispersive dielectric fiber: II. Normal dispersion. Appl Phys Lett 23:171–172
32. Hashizume Y (1985) Nonlinear pressure waves in a fluid-filled elastic tube. J Phys Soc Japan 54:3305–3312
33. Hashizume Y (1988) Nonlinear pressure wave propagation in arteries. J Phys Soc Japan 57:4160–4168
34. Heeger AJ, Kivelson S, Schrieffer JR, Su WP (1988) Solitons in conducting polymers. Rev Modern Phys 60:781–850
35. Helal MA (2001) Chebyshev spectral method for solving KdV equation with hydrodynamical application. Chaos Solit Fractals 12:943–950
36. Helal MA (2002) Review: Soliton solution of some nonlinear partial differential equations and its applications in fluid mechanics. Chaos, Solitons and Fractals 13:1917–1929
37. Helal MA, El-Eissa HN (1996) Shallow water waves and KdV equation (oceanographic application). PUMA 7(3–4):263–282
38. Hirota R (1971) Exact solution of the Korteweg–de Vries equation for multiple collisions of solitons. Phys Rev Lett 27:1192–1194
39. Jackson EA (1963) Nonlinear coupled oscillators. I. Perturbation theory: ergodic problems. J Math Phys 4:551–558
40. Jaworski M, Zagrodzinski J (1995) Position and position-like solution of KdV and Sine–Gordon equations. Chaos Solit Fractals 5(12):2229–2233
41. Kharif C, Pelinovsky E (2003) Physical mechanisms of the rogue wave phenomenon. Eur J Mech B/Fluids 22:603–634
42. Korteweg DJ, De Vries G (1895) On the change of form of long waves advancing in a rectangular channel, and on a new type of long stationary waves. Phil Mag 39(5):442–443
43. Kruglov VI, Peacok AC, Harvey JD (2003) Exact Self–Similar Solutions of the Generalized Nonlinear Schrödinger Equation with Distributed Coefficients. Phys Rev Lett 90(11):113902 http://prola.aps.org/abstract/PRL/v90/i11/e113902
44. Lakshmanan M (1997) Nonlinear physics: integrability, chaos and beyond. J Franklin Inst 334B(5/6):909–969
45. Lakshmanan M, Sahadevan R (1993) Painlevé analysis, Lie symmetries and integrability of coupled nonlinear oscillators of polynomial type. Phys Rep 224:1–93
46. Lamb H (1879) Treatise on the motion of fluids. Hydrodynamics, 6th edn 1952. Cambridge University Press, Cambridge
47. Lamb H (1971) Analytical descriptions of ultrashort optical pulse propagation in a resonant medium. Rev Mod Phys 43:99–124
48. Liu PL-F, Synolakis CE, Yeh HH (1991) Report on the International Workshop on long-wave run-up. J Fluid Mech 229:675–688
49. Lo EYM, Shao S (2002) Simulation of near-shore solitary wave mechanisms by an incompressible SPH method. Appl Ocean Res 24:275–286
50. Lynch DK (1982) Tidal bores. Sci Am 247:134–143
51. Malfliet W (1992) Solitary wave solutions of nonlinear wave equations. Am J Phys 60(7):650–654
52. Marin F, Abcha N, Brossard J, Ezersky AB (2005) Laboratory study of sand bedforms induced by solitary waves in shallow water. J Geophys Res 110(F4):F04S17
53. Martnez Alonso L, Olmedilla Morino E (1995) Algebraic geometry and soliton dynamics. Chaos Solit Fractals 5(12):2213–2227
54. Maxworthy T (1976) Experiments on collision between solitary waves. J Fluid Mech 76:177–185
55. Mei CC, Li Y (2004) Evolution of solitons over a randomly rough seabed. Phys Rev E 70:016302
56. Michallet H, Barthélemy E (1998) Experimental study of interfacial solitary waves. J Fluid Mech 366:159–177
57. Mireska HJ, Steiner M (1991) Solitary excitations in one-dimensional magnets. Adv Phys 40:196–356
58. Mollenauer LF, Stolen RH, Gordon JP (1980) Experimental observation of picosecond pulse narrowing and solitons in optical fibers. Phys Rev Lett 45:1095–1098
59. Newell AC (1985) Solitons in mathematics and physics. SIAM, Philadelphia
60. Olver PJ (1986) Applications of Lie groups to differential equations. Graduate Texts in Mathematics, vol 107. Springer, Berlin
61. Ostrovsky LA, Stepanyants YA (1989) Do internal solitons exist in the ocean? Rev Geophys 27:293–310
62. Paquerot JF, Remoissenet M (1994) Dynamics of nonlinear blood pressure waves in large arteries. Phys Lett A 194:77–82
63. Peyrard M, Dauxois T (2004) Physique des solitons. EDP Sciences. CNRS Editions, Paris
64. Rayleigh L (1876) On waves. Phil Mag 5(1):257–279
65. Remoissenet M (1999) Waves called solitons – Concepts and experiments. Springer, Berlin

66. Rodwell MJ, Allen ST, Yu RY, Case MG, Bhattacharya U, Reddy M, Carman E, Kamegawa M, Konishi Y, Pusl J, Pullela R (1994) Active and nonlinear wave propagation devices in ultrafast electronics and optoelectronics. Proc IEEE 82:1037–1058
67. Rogers C, Shadwick WF (1982) Bäcklund transformations and applications. Academic, New York
68. Rottman JW, Grimshaw R (2001) Atmospheric internal solitary waves. In: Environmental stratified flows. Kluwer, Boston, pp 61–88
69. Russell JS (1837) Report on the committee on waves. In: Murray J (ed) Bristol, Brit Ass Rep, London, pp 417–496
70. Russell JS (1839) Experimental researches into the laws of certain hydrodynamical phenomena that accompany the motion of floating bodies, and have not previously been reduced into conformity with the known laws of the resistance of fluids. Trans Royal Soc Edinb 14:47–109
71. Russell JS (1844) Report on waves. In: Murray J (ed) Brit Ass Rep Adv Sci 14, London, pp 311–390
72. Sanz–Serna JM, Christie I (1981) Petrov-Galerkin method for nonlinear dispersive waves. J Comput Phys 39:94–102
73. Stokes GS (1847) On the theory of oscillatory waves. Trans Cambridge Phil Soc 8:441–473
74. Su CH, Gardner CS (1969) Korteweg-de Vries equation and generalizations. III. Derivation of the Korteweg-de Vries equation and Burgers equation. J Math Phys 10:536–539
75. Taha TR, Ablowitz MJ (1984) Analytical and numerical aspects of certain nonlinear evolution equations, (III) numerical, KdV equation. J Comput Phys 55:231–253
76. Taniuti T, Wei CC (1968) J Phys Soc Jpn 24:941–946
77. Tappert F (1974) Numerical solution of the KdV equation and its generalisation by split-step Fourier method. Lec Appl Math Am Math Soc 15:215–216
78. Ustinov AV (1998) Solitons in Josephson junctions. Phys D 123:315–329
79. Weidman PD, Maxworthy T (1978) Experiments on strong interaction between solitary waves. J Fluid Mech 85:417–431
80. Yomosa S (1987) Solitary waves in large blood vessels. J Phys Soc Japan 56:506–520
81. Yuen HC, Lake BM (1975) Nonlinear deep water waves: theory and experiments. Phys Fluids 18:956–960
82. Zabusky NJ, Kruskal MD (1965) Interaction of solitons in a collisionless plasma and the recurrence of initial states. Phys Rev Lett 15:240–243
83. Zakharov VE, Shabat AB (1972) Exact theory of two-dimensional self-focusing and onedimensional self-modulation of waves in nonlinear media. Sov Phys JETP 34:62–69

Books and Reviews

Agrawal GP (2001) Nonlinear Fiber Optics. Academic Press, Elsevier
Akmediev NN, Ankiewicz A (1997) Solitons, Nonlinear Pulses and Beams. Chapman and Hall, London
Braun OM, Kivshar YS (2004) The Frenkel–Kontorova Model. Concepts, Methods and Applications. Springer, Berlin
Bullough RK, Caudrey P (1980) Solitons. Springer, Heidelberg
Davydov AS (1985) Solitons in Molecular Systems. Reidel, Dordrecht
Dodd RK, Eilbeck JC, Gibbon JD, Morris HC (1982) Solitons and Nonlinear Wave Equations. Academic Press, London
Drazin PG, Johnson RS (1993) Solitons: an Introduction. Cambridge University Press, Cambridge
Eilenberger G (1981) Solitons: Mathematical Methods for Physicists. Springer, Berlin
Hasegawa A (1989) Optical Solitons in Fibers. Springer, Heidelberg
Infeld E, Rowlands G (2000) Nonlinear waves, Solitons and Chaos. Cambridge University Press, Cambridge
Lamb GL (1980) Elements of Soliton Theory. Wiley, New York
Toda M (1978) Theory of Nonlinear Lattices. Springer, Berlin

Solitons Interactions

Tarmo Soomere
Center for Nonlinear Studies, Institute of Cybernetics, Tallinn University of Technology, Tallinn, Estonia

Article Outline

Glossary
Definition of the Subject
Introduction: Key Equations, Milestones, and Methods
Extended Definitions
Elastic Interactions of One-Dimensional and Line Solitons
Geometry of Oblique Interactions of KP Line Solitons
Soliton Interactions in Laboratory and Nature
Effects in Higher Dimensions
Applications of Line Soliton Interactions
Future Directions
Bibliography

Glossary

Solitary wave A localized structure which travels with constant speed and shape in otherwise homogeneous medium.

Soliton In the classical nomenclature, a soliton is a nonlinear spatially or temporally localized entity (solitary wave, impulse, wave packet, kink, etc.) of permanent form that retains its identity and shape in interactions with other similar entities. Also frequently used term for solutions of the relevant nonlinear partial differential equations approximately describing such physical entities.

Soliton (colloquially as well as in certain domains) Used for denoting any long-lived localized structures (e. g. solitary waves, self-trapped optical beams, localized vortices) which exhibit low energy radiation and approximately keep their shape.

Soliton solution A solution, usually in closed form, of the nonlinear partial differential equations that exhibit properties of a single soliton. A multi-soliton solution exhibits properties and collective behavior of groups of solitons.

Elastic interaction A generalization of the concept of elastic collisions of ideal mechanical bodies: interaction of localized entities in which the dynamically significant properties of the counterparts such as the total energy, linear and angular momentum, and topological charge or spin are conserved. For low-dimensional and scalar solitons denotes the case in which all the properties of interacting solitons are completely restored after interaction.

Inelastic interaction Interaction of soliton-like structures that modifies one or more dynamically significant properties of the counterparts.

Phase shift A change of the position of a soliton owing to interactions with other solitons compared to the location in which it would be in the absence of other solitons.

Resonant interaction A specific form of soliton interactions leading to the fusion of two or more solitons into a new soliton. Generally associated with an infinite phase shift of the counterparts.

Definition of the Subject

The concept of solitons reflects one of the most important developments in science of the 20th century: the nonlinear description of the world. It is customary to start an introduction to solitons by recalling that it is not easy to give a comprehensive and precise definition of a soliton. Frequently, a soliton is explained as a spatially localized (solitary) wave with spectacular stability properties. Although the combination of simultaneously being a localized structure propagating while mostly keeping its shape as waves do, and surviving for a long time in realistic (that is, nonlinear) conditions, already is a fascinating property in itself, yet another key quality defines whether a particular entity is a soliton. The distinction is made based on the way in which two or more of these objects interact with each other.

The classical nomenclature associates the term soliton with (i) nonlinear and (ii) localized entities, which (iii) have a permanent form and (iv) retain their identity in interactions with other entities from the same class (e. g. Drazin and Johnson [37]; also the entry ► Solitons and Compactons). The fundamental property of a soliton is thus to retain its identity in nonlinear interactions. In low-dimensional systems (understood here as environments admitting solitons in one spatial dimension and line solitons in two dimensions), a soliton's amplitude, shape, and velocity are restored after each interaction, whereas in more complex systems this request embraces all dynamically significant qualities. In other words, a structure is a soliton if and only if its *interactions* are *fully elastic*. Thus, the interaction of solitons is always an intrinsic topic whenever solitons are considered.

In the nonlinear world, *linear superposition* does not hold and *nonlinear interaction* takes place. Mathematically, classes of entities are called linear, whenever any linear combination of entities belongs to the same class. Among structures corresponding to the solutions of certain equations, exclusively those that satisfy linear equations are denoted as linear ones. Any linear combination of solutions of a linear equation also solves this equation. The principle of linear superposition is that the resultant structure is simply the sum of the parties and it implies that the individual linear wave shapes are perfectly restored after meeting with each other.

The property of elasticity of soliton interactions has a fundamentally different nature, because, as a rule, a linear superposition of soliton solutions is not a solution of the governing nonlinear equation. This property distinguishes nonlinear *solitary* traveling waves (that is, spatially or temporally localized pulses, whereas the relevant single-wave solution of the underlying equation may be completely stable) from *solitons*, which may be observed and analyzed as separate entities but which always express property (iv) should similar entities appear in the system.

The propagation and interaction of both solitons and solitary waves in realistic conditions (that is, in the presence of dissipation, external fields, and inhomogeneities of the medium) frequently results in a certain loss of energy (usually in the form of radiation of relatively short waves); in certain cases collapse or fission of the structure, or a net exchange of energy between the waves occurs. Even if a solitary wave travels without radiation losses in an ideal environment, it may only be called a soliton if neither radiation loss nor net energy exchange occurs when it meets a similar structure; otherwise a perfect re-emerging of each soliton after the interaction would be impossible. More generally, no net changes of any dynamically or topologically significant quantities are permitted in elastic collisions. Only changes of certain dynamically less significant properties such as the exact location (phase) of the solitons are common in elastic interactions. Solitary waves or structures possessing a selection of properties of solitons are sometimes called quasi-solitons.

Although the first evidence of solitons was extracted from observations in nature, the theory of solitons and their interactions has been developed very much in terms of the mathematical theory of the relevant nonlinear partial differential equations (PDEs) of the motion. The definitions of solitons and their interactions are commonly formulated in terms of solutions to such equations. It is

even customary to speak of solitary waves and solitons, and of their interactions, having in mind the relevant exact solutions to these equations. Therefore, in mathematical terms, a soliton – either a solitary wave or a more complex entity – is a solution of a nonlinear PDE that elastically interacts with other similar solutions. In the contemporary nomenclature it is also customary to associate solitons and their interactions with numerical (multi-soliton) solutions to these equations.

Interactions between solitons are the most fascinating features of soliton phenomena. The most instructive quality of soliton interactions is the universality of many of their features that exists in spite of the extremely diverse physical origins of the counterparts. The presentation of the basic conceptual ideas involving soliton interactions, a discussion of several particular cases of such interactions and an overview of observations of the relevant phenomena in natural conditions are mostly limited to effects occurring in the elastic interactions of one-dimensional and line solitons, in which the solitons behave very much like ideal mechanical bodies and which serve as the kernel of the classical concept of solitons and their interactions. Some spectacular features of nearly elastic interactions in higher dimensions are interpreted in terms of generalizations of the classical elastic interactions. Finally, a selection of practical applications of specific features of soliton interactions is again discussed from the viewpoint of line solitons.

Introduction: Key Equations, Milestones, and Methods

Integrable Equations

The quality of some nonlinear PDEs to admit soliton solutions is associated with the property of integrability. Integrable equations usually admit an infinite number of integrals (constants) of motion. Many nonintegrable equations possess localized shape-preserving traveling wave solutions that resemble solitons. However, only integrable equations have the universal property of possessing exact multi-soliton solutions that reflect perfectly elastic interactions between individual solitons. Thus, the integrability of the underlying equation is a primary constituent of the classical soliton interactions. Arnold [10] defines, for example, a soliton as a solitary wave (solution) of an integrable PDE having additional stability and robustness features which are inherited directly from the integrability of this PDE. Nonlinear interactions even between the greatly different (but physically relevant) solutions of integrable equations are also elastic.

Among the variety of integrable PDEs that admit multi-soliton solutions, examples related to the Korteweg–de Vries (KdV) equation, the (focusing) nonlinear Schrödinger (NLS) equation, the integrable Boussinesq equation, the sine-Gordon (SG), and the Kadomtsev–Petviashvili (KP) equation will be used below. Below we shall refer to the listed equations as the classical soliton equations. Since the KdV and the KP equations represent a large number of different physical systems and their solutions can be easily visualized in terms of the easily accessible environment of shallow-water waves, solutions to and results from studies of these equations will be largely used for illustrations of the explanations below.

The rapid development of the theory of solitons has led to the discovery of many integrable equations which show multi-soliton solutions and describe soliton interactions (see, e. g. Arnold [10] and the entry ► Partial Differential Equations that Lead to Solitons). A number of analytical methods have been developed to obtain the (multi-) soliton solutions of integrable equations, starting from the late 1960s. A major tool is the *Inverse Scattering Transform* (IST), first developed by Gardner, Green, Kruskal and Miura [47,50] for the KdV equation. The method consists of associating to the evolution equation a Sturm–Liouville problem. In the case of the KdV equation the Sturm–Liouville equation is just the time independent Schrödinger equation of quantum mechanics, where the potential is the wave function of the KdV equation. The single- and multi-soliton solutions are associated with the discrete spectrum of eigenvalues. A generalization of the Inverse Scattering Transform, known as the *AKNS method* (named after Ablowitz, Kaup, Newell and Segur [2] who developed the method), applies to a large class of integral equations. See the entry ► Inverse Scattering Transform and the Theory of Solitons for detailed information.

At the same time, Hirota [62] found that soliton solutions can be obtained by reducing, through a suitable transformation, the evolution equation to a *bilinear form*. Using some properties of the bilinear operator, it is possible to find the multi-soliton solution of an integrable equation. The proof of integrability of a particular PDE is usually nontrivial. The experience with the Hirota method, applied to a variety of nonlinear wave equations, is that exact multi-soliton solutions are found in many cases when it is not known if the equation is integrable. It is natural to expect that the existence of such solutions reflects the integrability in some sense; however, no proof of this conjecture is known at this time.

Mathematically speaking, it is always possible to build an integrable equation representing certain features of an

observed phenomenon. Nevertheless, only a few of these equations can be properly derived, normally using asymptotic expansions, from the corresponding primitive equations describing physical phenomena. Even the classical soliton equations only approximately describe the natural phenomena. A more exact description of the natural processes requires the introduction of additional contributions to these equations. Such contributions usually destroy the property of integrability and lead to nonelasticity of interactions of soliton-like entities. Consequently, most of the times solitons are just a good approximation of solitary waves in nature. This limitation is to some extent balanced by the fact that in many cases integrable soliton-admitting equations adequately represent the basic features of physical phenomena even far beyond their formal scope of validity. In spite of such intrinsic limitations, the tools developed in soliton theory have allowed researchers to reach a very deep understanding of some physical phenomena which would have hardly been explained by other means.

Milestones

Several milestones of the solitons' history (► Solitons: Historical and Physical Introduction) are particularly significant to soliton interactions. Already the famous first observations in 1834 of shallow-water solitons [132] implicitly involved certain effects created through soliton interactions. The observed wave ahead of a ship probably had a straight crest, as the one reproduced in 1995 in the Union Canal near Edinburgh (see, e. g. p. 11 in [34]). It is a remarkable feature of ship-induced solitons in relatively narrow and shallow channels that a straight-crested upstream soliton exists ahead of a two-dimensional (2D) disturbance, whereas the solitons generated ahead of a ship sailing in wider shallow water areas have curved crests. The straight crest of the Russell's soliton and of her sister structures are thus not intrinsically one-dimensional (1D) structured solitons. They are formed owing to a specific mechanism of soliton interactions as described below. The KP equation, allowing a simple but adequate description of the crest-straightening phenomenon, was derived even later than the word soliton was coined.

A substantial step towards the concept of solitons was made in the mid-1950s by Fermi, Pasta and Ulam [40] who numerically analyzed the evolution of phonons in an anharmonic lattice. From the mathematical point of view this problem can be described by a discretization of the KdV equation. Surprisingly at the time, the process did not lead to an energy equipartition among the modes, although nonlinearity generally tends to create such an equipartition. Instead, a sort of recurrence, an early evidence of the elastic soliton interactions, was observed.

The basic features of soliton interactions were first demonstrated for media described by the KdV equation [64,174]. The evolution of a sinusoidal initial wave with periodic boundary conditions gives rise to an ensemble of solitary waves that move with different speeds and gradually overtake each other. Originally eight of them were mentioned; later revisitations increased the count to nine [106]. The difference in the number of solitons is not principal since small-amplitude solitons only become evident through additional phase shifts in extremely long-term simulatons [39,133]. One surprise of the system was that after a very long time the whole profile reappears. The other, conceptually much more striking effect, was the persistence of the waves: after a certain phase of interaction with each other, they continued thereafter as if there had been no interaction at all. This persistence, which reflects the essence of the soliton interactions, led Zabusky and Kruskal to invent the name "soliton".

Solitons and new features of their interactions were later discovered in a large number of different physical systems. The universal principle generalizing the physical forces affecting the birth of all solitons is that the propagating disturbance (a pulse in a fiber, a wave-packet on a water surface, a self-focusing optical beam, a localized vortex, or a vortex dipole, etc.) is captured in a potential well jointly induced by its motion and by virtue of the nonlinearity of the underlying system. The solitons can thus be interpreted as the bound states of the relevant potential wells, and soliton interactions – as interactions between the bound states of a jointly induced potential well, or between bound states of different wells located at close proximity. This concept of potential well becomes vividly evident in the inverse scattering theory due to the conservation of the eigenvalues of the underlying spectral problem. This universality explains why, in spite of the immense diversity of the solitons and the underlying physical systems, the interactions (collisions) between solitons in all of these systems follow the same principles.

Classical Soliton-Admitting Equations and Appearance of Solitons

The particular appearance of soliton interactions may vary a lot, depending on the physical system and the nature of the solitons. The Russell soliton is an example of nontopological solitons that often become evident in the form of traveling waves and that cannot exist in rest. On the contrary, single topological solitons (that interpolate between

two different states of the system which have the same energy) may exist in rest.

A localized entity consists of an infinite number of Fourier harmonics and generally experiences (linear) spreading or wave radiation. A localized entity of permanent form therefore may only appear if the spreading effects are balanced by a certain restoring force that evidently can only be of a nonlinear nature. In other words, all solitons correspond to a certain balance between spreading and nonlinear effects, the latter becoming evident, for example, as nonlinear steepening of the wave shape in the KdV system, or focusing of diffractive structure in nonlinear optics.

The physical meaning of spreading depends on the particular system. Dispersion is responsible for the spreading of pulses in all media in which the group velocity depends on the wave properties. The classical examples of dispersive solitons are surface [132], internal [33] and Rossby solitons [128] in geophysical systems. Similar structures occur in a large variety of environments such as drift solitons or ion-acoustic solitons in plasma [169], or solitons in optical fibers [60].

Many examples of dispersive solitons and their interactions occur in the framework of the KdV equation and its generalizations. It is a characteristic equation governing weakly nonlinear, spatially one-dimensional (1D) waves whose phase speed attains a simple maximum for infinitely long waves

$$\tilde{\eta}_t + \frac{3c_0}{2h}\tilde{\eta}\tilde{\eta}_x + \frac{c_0 h^2}{6}\tilde{\eta}_{xxx} = 0\,,$$
$$\text{in nondimensional variables}\quad \eta_t + 6\eta\eta_x + \eta_{xxx} = 0\,, \tag{1}$$

where η is the relative water surface elevation in the shallow water environment (Fig. 1), h is the still water depth, and c_0 is the maximum phase speed (celerity) of linear waves at this depth. Its generalization to the case of solitons propagating in slightly different directions

$$(\eta_t + 6\eta\eta_x + \eta_{xxx})_x + 3\eta_{yy} = 0\,, \tag{2}$$

was derived by Kadomtsev and Petviashvili [66]. Named the KP equation after them, it describes features of a so-called weakly 2D environment (also called 1.5-dimensional systems). The nondimensional (x, y, t, η) and physical variables $(\tilde{x}, \tilde{y}, \tilde{t}, \tilde{\eta})$ in Eq. (2) are related as follows: $x = \sqrt{\varepsilon}(\tilde{x} - \tilde{t}\sqrt{gh})/h$, $y = \varepsilon\tilde{y}/h$, $t = \sqrt{\varepsilon^3 gh}\tilde{t}/h$, $\eta = 3\tilde{\eta}/(2\varepsilon h) + O(\varepsilon)$ whereas $\varepsilon = |\tilde{\eta}_{\max}|/h \ll 1$.

Since the KdV equation only admits nontopological elevation solitons, the KdV/KP framework misses several basic features of soliton interactions, such as collisions of solitons with different topological charge and interactions of solitons of different type. These classes of dispersive solitons are represented by the sine-Gordon (SG) equation

$$\theta_{tt} - c_0^2\theta_{xx} + \omega_0^2 \sin\theta = 0\,, \tag{3}$$

where θ in many applications has the meaning of a certain 2π-periodic angle and ω_0 is the minimum angular frequency of linear waves. This equation, with its name stemming from a play of words regarding its form which is similar to the Klein–Gordon equation, arose already in the 19th century in studies of certain surfaces. It grew greatly in importance when it was realized that it led to kink and antikink solutions with the collisional properties of solitons [116]. Its major field of application is solid state physics, yet it also appears in a number of other physical environments such as the propagation of fluxons in Josephson junctions (a junction between two superconductors), the motion of rigid pendulums attached to a stretched wire, and dislocations in crystals. Its kink solutions (Fig. 1)

$$\theta_S = 4\arctan\exp\left(\pm\frac{\omega_0}{c_0}\frac{x - vt - x_0}{\sqrt{1 - v^2/c_0^2}}\right)\,, \tag{4}$$

where x_0 is the initial position of the single soliton at the x-axis and $v < c_0$ is the speed of translation of the soliton, represent so-called topological solitons that interpolate between two different states of the system, which have the same energy and may exist in rest (e. g. a frozen (anti)kink, $v = 0$). The SG equation is a simple model admitting breather solitons and soliton pairs with different topological charges. The bulk topological charge is an invariant of the system. In a more complex manner the topological charge becomes evident in the theory of optical spatial solitons, where it can be interpreted in terms of spin of elementary particles. The kink and antikink solutions also represent the possibility of the existence of solitons and antisolitons. Both attractive and repulsive interactions can be vividly demonstrated in this framework and, after all, interactions of physically meaningful solitons of different types are possible.

Another key equation for dispersive waves is the nonlinear Schrödinger (NLS) equation that has a nondimensional form

$$i\psi_t + p\psi_{xx} + q\,|\psi|\,\psi^2 = 0 \tag{5}$$

and only has soliton solutions in the focusing case $pq > 0$. This is a universal equation describing the evolution of complex wave envelopes ψ in a dispersive weakly nonlinear medium. It applies, among other phenomena, to

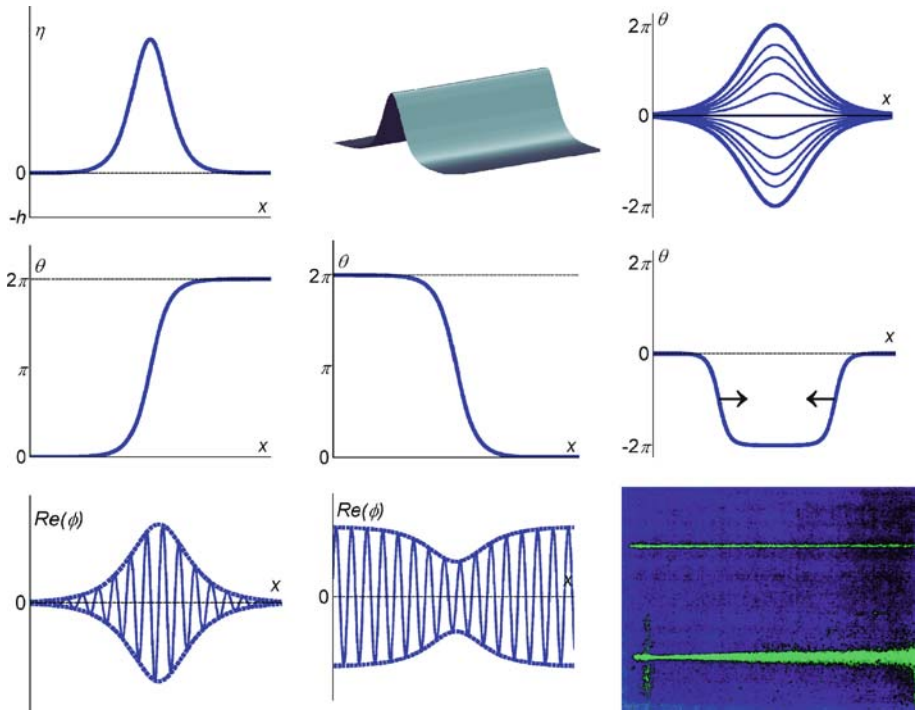

Solitons Interactions, Figure 1

A selection of solitons. From *left* to *right*: *top* – KdV traveling wave soliton, KP plane (*line*) soliton, SG breather; *middle* – kink, antikink, and a kink-antikink pair; *bottom* – bright and gray envelope soliton, and a top view photograph of a 10-μm-wide optical spatial soliton propagating in a strontium barium niobate photorefractive crystal and the same beam diffracting naturally. The *bottom right* image is from [152]. Optical spatial solitons and their interactions: Universality and diversity. Reprinted with permission from AAAS

deep water waves. The studies of its soliton solutions and their interactions also date back to the 1960s [175,176]. Its major application in the context of soliton interactions is nonlinear optics, where its different versions describe the propagation and interactions of both dispersive solitons (e. g. envelope solitons in optical fibers) and diffractive optical spatial solitons.

The basic reason for the existence of optical solitons is that the optical properties (refractive index or absorption) of some materials are modified by the presence of light. This feature introduces nonlinearity into the system and alters the propagation of optical pulses either in space or in time. A dispersive optical soliton is formed when the linear dispersion effects are balanced by a sort of lensing (of short light pulses of permanent form that are called temporal solitons in the optical nomenclature) in an appropriate material (for example, an optical fiber). They were predicted by Hasegawa and Tappert [60] and first observed by Mollenauer et al. [97]. The basic properties of their interactions [59] are analogous to those occurring for KdV or SG solitons. There is continuous interest for their studies because of their applications, e. g. in long-distance optical communication systems.

The generic example of diffractive solitons, evolution and interactions of a part of which are described by the NLS equation in two space dimensions, are optical spatial solitons. The natural tendency to broaden in space owing to the diffraction of optical beams (Fig. 1) can be balanced, for example, due to the optical Kerr effect that consists of a local refractive index change induced by light. As a first approximation, this change is assumed to take place instantaneously and to be proportional to the local intensity of light. In the focusing case (when an increase of the intensity of light causes an increase in the refractive index), the light-induced lensing can make an optical beam stable with respect to perturbations in both its width and intensity. The resulting beam in a 2D or 3D medium is an example of a diffractive soliton. Interaction of such beams occurs owing to overlapping of the modified regions of the medium. Usually such an interaction is local and has a long range only under specific conditions [129].

The first studies suggesting the existence of such solitons in nonlinear optical Kerr media date back to the 1960s [11,27]. In some cases (such as the 1D Kerr solitons in a planar medium) they are classical solitons and

described by a fully integrable equation (the NLS equation in two space dimensions [176]). Self-trapped light beams and more complex structures in a 2D and 3D medium attracted the attention of researchers starting from the 1980s, when the appropriate materials were found [13].

In multi-dimensional media, both diffraction and dispersion affect the propagation of solitary waves. A classical example of a diffractive-dispersive system is the motion of short-crested waves on the water surface. Most of the analysis of soliton interactions in this system has been made for effectively 1D solitons with infinitely long crests (so-called *plane* or *line solitons*); the effects of diffraction being implicitly neglected in their propagation and interactions. A practically realized example of diffractive-dispersive solitons are short pulses of incoherent (or white) light that is self-trapped both in the direction of its propagation and in the transversal direction(s), and that propagate changing neither their shape nor their length [94].

Extended Definitions

The classical nomenclature of low-dimensional solitons and their interactions was formulated several decades ago, and it is not surprising that many later discoveries have led to attempts to extend its content. A part of the extensions add several consistent features (such as the conservation of linear and angular momentum, and topological charge or spin during the soliton interactions); however, in several schools substantial generalizations of the term soliton and of the interpretation of the soliton interactions have been introduced. The only component of the classical nomenclature kept in all the interpretations is the essential role of nonlinearity. Since it is virtually impossible to reflect all the extensions, mostly the material and results following the classical definition are presented below.

The solitons are commonly interpreted as a subset of traveling solitary waves. The classical nomenclature favors structures that are stationary in a proper coordinate system; yet it does not exclude the resting topological solitons (e. g. single-kink solutions to the sine-Gordon equation).

A principal extension of the classical definition of solitons that caused quite a serious discussion in the 1970s [101] consists of accounting for the *resonant interactions*. They may result in the fusion of several solitons into a new soliton or, equivalently, in the decay of solitons into counterparts [177]. Although the number of solitons is not conserved, all the parties of the resonant system are solitons at the limit of exact resonance and no energy radiation occurs. It is commonly accepted now as a variation (or a limiting case) of elastic interactions.

The classical soliton interactions preserve the shape of the solitons. The shape is understood in a wide sense; e. g. the shape of the envelope of the NLS solitons is preserved as well as the time-dependent shape of breathers that have an internal oscillation. This quality, which is universal for all scalar solitons, has a more general interpretation in the case of *vector* (composite) *solitons*. For example, the otherwise elastic interactions of Manakov vector solitons exhibit a shape-changing nature [125] in an integrable system consisting of two coupled NLS equations [82]. This property has a paramount practical importance and bright perspectives e. g. in optical soliton-based computations ([155], see also entry ► Optical Computing), being one of the key options of practical applications of specific features of soliton interactions.

An obvious reason for a wider view on soliton interactions is that the relatively simple integrable equations admitting soliton solutions stem from certain asymptotic expansions and only approximately describe the processes in nature. A more exact representation of the physical processes through inclusion of higher order terms of those expansions generally destroys the integrability of these equations. In many cases, though, the perturbing terms are small. The influence of a small dissipative perturbation is usually obvious: it brakes the moving solitons and/or damps the oscillating solitons. Such structures are sometimes called dissipative solitons although such a name is somewhat self-contradicting. Effects due to conservative (Hamiltonian) perturbations are much richer in content. Usually they do not destroy or brake solitons, but they may, e. g., render otherwise elastic soliton interactions to inelastic ones owing to extra wave radiation. A well-known example of the effect of perturbations is dissipation-induced annihilation of a kink-antikink pair in a perturbed sine-Gordon equation that otherwise would lead to a breather [114]. A monumental survey of the relevant problems is presented in [77]; see also entry ► Soliton Perturbation.

Since none of the physical and numerical environments perfectly matches the governing equations, solitary waves both in realistic conditions and in numerical simulations always are approximations to solitons. In numerical experiments practically ideal solitons can be represented and errors basically occur due to purely computational inaccuracies. Laboratory experiments with solitons encounter problems with perfect excitation of even single solitons and moreover with controlling their interaction properties; however the relevant results are of fundamental importance in establishing the adequacy, the limits and the practical applicability of the properties of solitons and their interactions.

In the theory of elementary particles frequently a soliton is defined merely as a localized solution (resp. entity) of permanent form and the constraint of its survival in collisions is released (e. g. Rebbi [127]). A popular and deep example is the quantum field theory, where e. g. the Yang–Mills field equations admit solutions localized in space (which represent very heavy elementary particles), and also solutions, localized in time as well as in space (instantons). This position has certain support by the conjecture that the classical soliton equations can be interpreted as reductions of self-dual Yang–Mills equations and thus the classical solitons form a subset of the above solutions.

A similar position is also widely adopted in studies of optical beams and vortices. A large part of such beams are described by nonintegrable equations [135] and thus lie beyond the classical soliton nomenclature. Many features of their behavior and interactions, however, demonstrate a striking similarity with that of classical solitons, and exhibit universal properties [152], resembling the similar universality of the classical solitons interactions. Soliton-like beams organized by different nonlinear effects were mostly called solitary waves until the mid-1990s. The modern nonlinear optics nomenclature treats all self-trapped optical beams as solitons [136] although their interactions are not always elastic. A selection of results of studies of their interactions that give a flavor of the richness of soliton interaction phenomena in 3D media is presented below.

The constraint of energy conservation both in the motion and the interactions of the solitons is frequently disregarded in studies of long-lived structures which exhibit reasonable radiation [42] but that still survive in certain collisions. On the other hand, other highly interesting solitary structures such as (Rossby) dipole modons [83,84] do not radiate energy but only survive under certain conditions. A certain amount of results and concepts in this domain is presented to highlight the overlapping of the properties of soliton interactions and interactions of other long-living structures.

Elastic Interactions of One-Dimensional and Line Solitons

The term soliton was originally introduced for nonlinear disturbances, the interaction of which resembles the collision of particles and is fully elastic (Table 1), so that the number of solitons is always conserved and their amplitudes are fully restored afterwards. In low-dimensional cases such as those described by the KdV equation, the amplitudes, directions and propagation velocities of each soliton always recover their initial values (Fig. 2). The shape of each soliton is prescribed by the structure of (single-) soliton solutions to this equation and, as the simplest evidence of the recurrence of the system, it is not surprising that the shape is perfectly restored as well.

Three phases of the interactions of solitons can be distinguished either in time or in space (Fig. 2). First, well-defined, clearly separated solitons approach each other. In the second (interaction) phase, they usually lose their identity and merge into a composite structure. After a while, the solitons emerge again. The appearance of the composite structure depends on the particular physical system; as a rule it is neither a linear superposition of (properly shifted) counterparts nor a new soliton. Therefore solitons survive the collision event, even though they completely merge for a while. The variety of the composites range from the complete vanishing of the counterparts (e. g. kink-antikink collision in the SG equation) over certain interactions of KdV solitons, in which the individual identities are almost conserved and two humps are always visible, to a fourfold elevation of the water surface in a resonant interaction of the KP solitons.

In some environments, one cannot say whether the solitons propagate through each other as waves do or collide as particles do. Both interpretations have a solid ground. The interaction of unidirectional KdV solitons resembles the collision of two particles whereas, e. g. in head-on interactions of Rossby dipole solitons[84] the identity of both counterparts can be continuously tracked and the interaction process resembles a sort of complex overtaking.

Attraction and Repulsion

Another feature of soliton interactions resembling collisions of real massive and/or electrically charged particles that exert certain forces on each other is that soliton interactions may show an *attractive* or *repulsive* nature. This feature can be demonstrated in the framework of the kink-antikink solution of the sine-Gordon equation

$$\theta_{SA} = 4\arctan\frac{c_0 \sinh \nu t\xi}{\nu \cosh x\xi}, \quad \text{where } \xi = \frac{\omega_0}{c_0}\frac{1}{\sqrt{1-\nu^2/c_0^2}} \tag{6}$$

and that only exists if $\nu > 0$. The motion of the counterparts speeds up when they move towards each other, and slows down when they move apart after passing each other. If one examines the positions of the centers of the counterparts of this solution, the sequences of their positions in the vicinity of their meeting point are not the same

Solitons Interactions, Table 1

Properties of solitons in elastic interactions of 1D and line solitons. Analogous rules hold for angular momentum and spin in elastic collisions in higher dimensions

Property	1D interactions and nonresonant interactions of line solitons	Exceptions in resonant interactions
Number of solitons	Conserved except in a short phase of interaction	Changed
Energy	Restored after interactions for each soliton	Merging possible; total energy conserved.
Amplitude Shape (steepness)	Substantial local changes may occur in the interaction region; restored after interactions	Durable changes may occur
Phase or location	Finite phase shifts commonly occur; yet no phase shifts in certain environments. The total phase shift is exactly equal to the sum of partial shifts that would result from separate collisions with each of the solitons	Infinite phase shifts
Geometry	Substantial changes in the interaction region, restored after interaction	Durable changes may occur
Propagation direction, linear momentum	Conserved for each soliton	Conservation of bulk linear momentum
Topological charge	Conserved for each soliton	

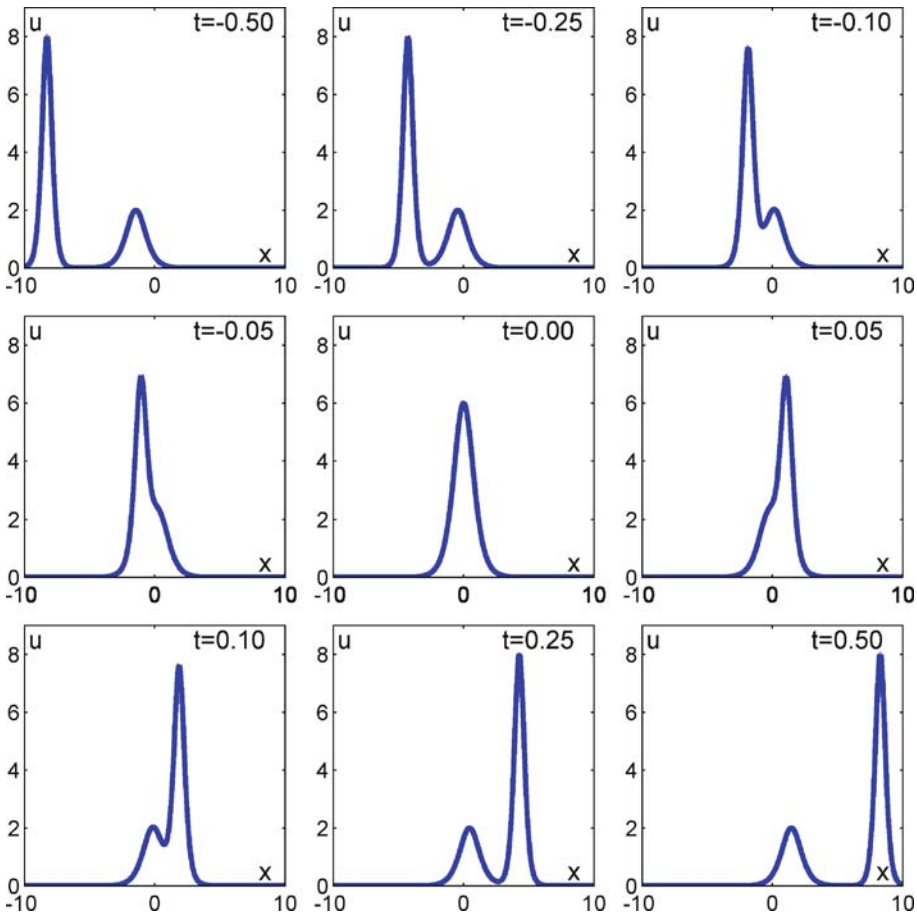

Solitons Interactions, Figure 2

Temporal evolution of KdV solitons described by the two-soliton solution $u(x, t) = 12 \times (3 + 4\cosh(2x - 8t) + \cosh(4x - 64t))/([3\cosh(x - 28t) + \cosh(3x - 36t)]^2)$ [37]. The amplitudes of the counterparts are $u_{1\,\max} = 8$ and $u_{2\,\max} = 2$. Notice the gradual decrease of the taller soliton when it catches the shorter, and its gradual increase when it moves further ahead of the shorter one, and the relatively small amplitude $\tilde{u} = 6$ of the composite structure

as they would be if they had been alone in the system, or far from the other counterpart.

If the solitons were massive or charged particles, one would say that an attractive force exists between them which decreases as their separation increases. On the contrary, the speed of two kinks as well as two antikinks approaching each other slows down as their separation decreases, which is characteristic for repulsive interaction. Another classical reflection of "forces" between solitons is the effect of attraction or repulsion of (envelope) solitons in optical fibers [53].

In two dimensions, the attractive interactions of two obliquely interacting line solitons become evident as an attraction of the relevant wave crests. In the case of optical spatial solitons the attraction affects their centroids. For repulsive interactions the opposite holds. The relevant effects are vividly demonstrated in the framework of the KP equation for the description of drift vortex solitons in environment, otherwise described by the Hasegawa–Mima equation [158]. This environment is equivalent to the one described by the Charney–Obukhov equation for large-scale motions in geophysical hydrodynamics admitting Rossby waves and solitons.

Transient Amplitude Changes, Durable Phase Shifts and Recurrence Patterns

Already the first study of solitons [174] revealed some other universal aspects of soliton interactions. First, certain durable phase shifts may occur, the magnitude of which depends on the interaction details. For example, during the collision presented in Fig. 2, the taller soliton is shifted forward by an amount $\Delta x_1 = \frac{1}{2}\log 3$ and the shorter one backwards by $\Delta x_2 = \log 3$ compared to the case when the solitons are alone in the system. The above-discussed "forces", if present in the system, may be interpreted as the cause of the phase shifts.

The total phase shift of a soliton induced by elastic collisions with any number of solitons is exactly equal to the sum of (partial) shifts that would result from separate collisions with each of the solitons involved. This property is commonly referred to as the *absence of many-particle effects*.

In the case of line solitons the phase shifts become evident in the form of durable shifts of the counterparts during the interaction (Fig. 3). In the negative phase shift case the solitons are delayed compared to their position in the absence of interactions, whereas in the positive phase shift case they seem to be accelerated.

Second, the amplitudes of the interacting solitons may gradually change as they approach to each other (Fig. 2).

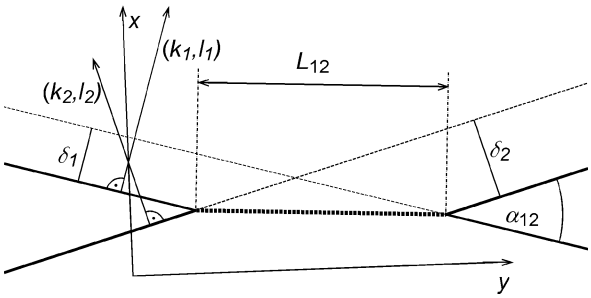

Solitons Interactions, Figure 3
Idealized patterns of crests of interacting KP solitons (*bold lines*), their position in the absence of interaction (*dashed lines*) and the crest of the residue s_{12} (*bold dashed line*, see Eq. (7)) corresponding to the negative phase shift case (adapted from [120])

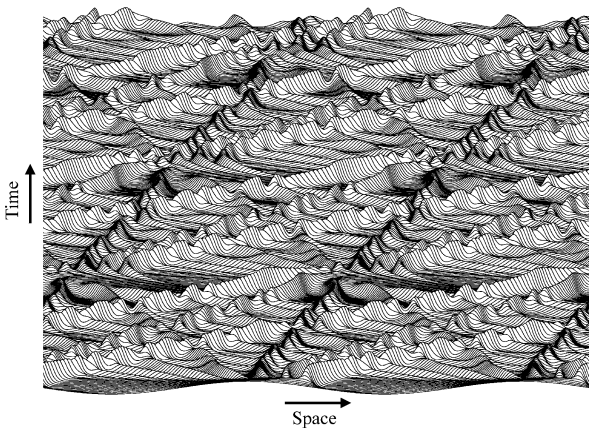

Solitons Interactions, Figure 4
Time-slice plot over two 2π periods of the evolution of the ensemble of the KdV solitons generated from a sinusoidal initial condition in space for $0 \le t \le 97$ at $\log d = -2.3$. Reprinted from [134]

The amplitude of the composite structure is generally not equal to the sum of amplitudes of the interacting solitons. In interactions of a small number of KdV solitons running along an infinite medium, the behavior of each counterpart and its influence on the partners can always be identified from the shape of the elevation since after some time the solitons become separated enough to detect all of them and their perfectly restored properties in the limit of infinite separation.

More insight into the properties of the soliton interaction is provided by long-term simulations of ensembles of solitons excited from a periodical signal (Fig. 4). The total number of interacting parties is infinite then, the solitons are tightly packed and such a separation not necessarily occurs. In some cases the presence of several shorter solitons can only be identified through their contributions

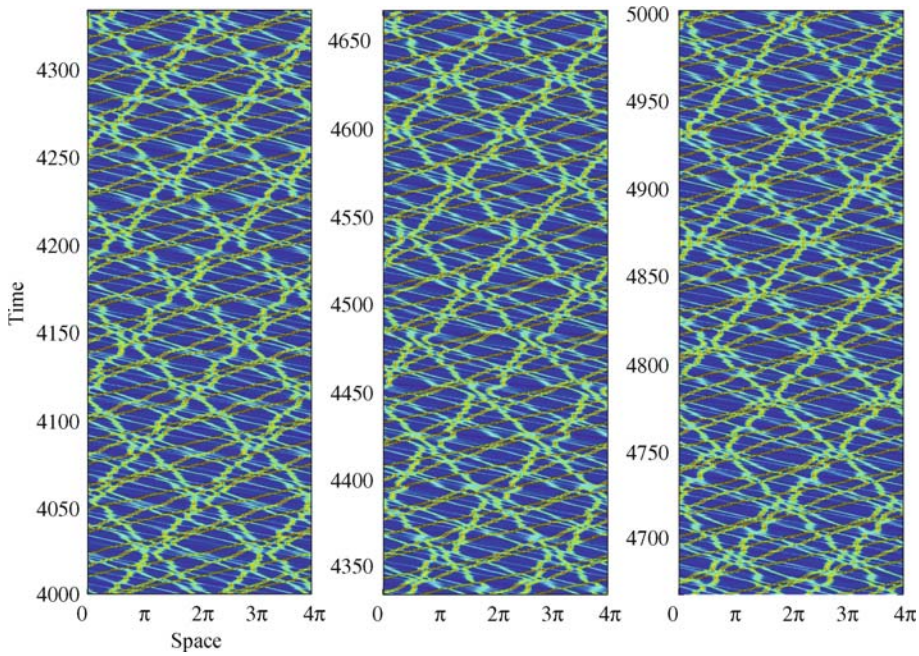

Solitons Interactions, Figure 5
Patterns of locations of crests of the KdV solitons generated from sinusoidal initial conditions at $\log d = -2.3209$. Reused with permission from [39]

to the phase shifts of the partners since they are seldom visually identifiable; such solitons are called hidden solitons [39].

The trajectories of the (crests of the) counterparts may show quite a complex pattern and may exhibit both simple recurrence (at which the initial system is approximately restored) as well as super-recurrence (understood as a nearly perfect restoration of the initial system over longer times).

Practically periodical in time rhombus-like patterns (suggesting that interactions of quite a small number of solitons govern the properties of the system) optionally occur in cases when a super-recurrence is possible (Fig. 5), whereas in other cases these patterns show gradual variation also over extremely long times. There seems to exist a critical value d^* in the range $-1.8 < \log d^* < -1.9$ of the dispersion parameter d in the KdV equation (presented in the form $\eta_t + \eta\eta_x + d\eta_{xxx} = 0$), which distinguishes if a super-recurrence exist. If $d < d^*$, no super-recurrence can be detected even within extremely long calculation times [39,134].

Durable Local Amplitude Changes in Oblique Interactions of Line Solitons

The interaction of unidirectional KdV solitons does not create any large changes in the wave amplitudes ([37], see also ► Korteweg–de Vries Equation (KdV), History, Exact *N*-Soliton Solutions and Further Properties of the). However, amplitude amplification may occur under certain conditions when line solitons propagating in slightly different directions meet each other. Extensive evidence comes from coastal engineering, where it was known already a long time ago that ocean waves approaching breakwaters and seawalls under a certain angle caused unexpectedly large overtopping. This phenomenon was systematically studied already in the 1950s [117], well before the word soliton was coined.

When a shallow-water wave is obliquely launched along an ideally reflecting wall, the reflection does not necessarily follow the rules of geometrical optics. For a certain range of incidence angles, the crests of the approaching and the reflected wave fuse together near the wall and the process resembles the *Mach reflection*. The common crest, an analogue of the *Mach stem*, is generally unsteady and gradually lengthens for sine waves [15] and Stokes waves [173]. This type of reflection also occurs during perfect reflection of random [89] and solitary waves of different nature [45,92,93,115,159].

For a certain set of parameters of the approaching KdV soliton, the resulting structure is equivalent to half of the pattern created by two interacting KP solitons of equal amplitude (Fig. 3). The prominent feature of both the (Mach)

reflection and oblique interactions of shallow-water solitons is that the height of the common crest may considerably exceed the sum of the heights of the interacting solitons. For interacting waves with equal amplitudes the hump can be up to four times as high as the incoming waves. Originally established in the context of the reflection of Boussinesq solitons ([92,93]; named Mach reflection by Melville 1980 [91]), the amplitude amplification in a certain region occurs in all oblique attractive interactions of line KP solitons (that is, in interactions with a negative phase shift). Since the interaction pattern is stationary in an appropriately moving coordinate system for a range of the parameters of the interacting solitons, it may lead to durable local changes of the amplitude of the resulting pattern of elevation.

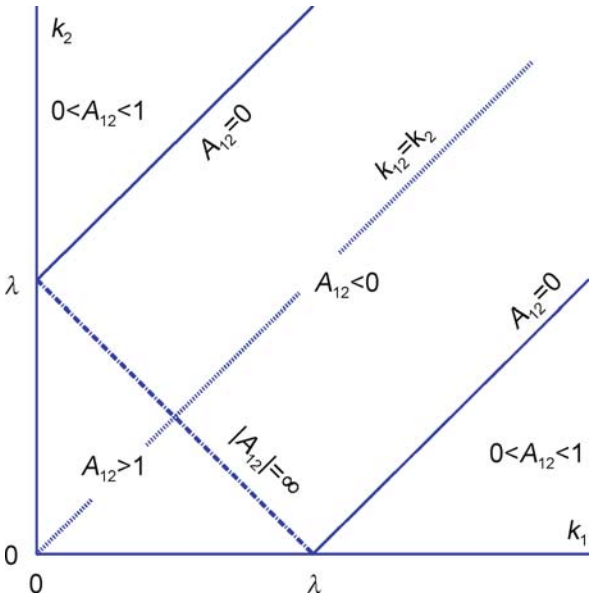

Solitons Interactions, Figure 6
Negative and positive phase shift areas for the two-soliton solution of the KP equation for fixed $l_{1,2}$. The phase shift $\Delta_{12} = -\ln A_{12}$ has a discontinuity along the line $k_1 + k_2 = \lambda$ at which $|A_{12}| = \infty$ and the resonant interactions occur. No two-soliton solution exists in the area where $A_{12} < 0$. An analogue of linear superposition occurs when $A_{12} = 0$. Adapted from [120]

Resonance

While the theory of integrable systems requires that the 1D solitons retain their identity in interactions, collisions of solitons in multiple dimensions may lead to the formation of new solitons. The phenomenon called *resonant interaction* leads to the emergence of new structures that combine the energy of the counterparts and to modifications of the number of solitons and of some of their geometrical features.

This possibility was first recognized by Newell and Redekopp [101] in the context of the KP and NLS equations. Miles [93] gives an early answer to their question about the practical consequences of this phenomenon in terms of an essential increase of the wave amplitude occurring during the resonant Mach reflection. Originally named "breakdown of the Zakharov–Shabat theory of integrable systems with more than one spatial dimension", it highlights the more complex nature of soliton interactions in more than one dimension, where large phase shifts are possible.

Remarkably, the resonance conditions $\boldsymbol{\kappa}_\infty = \boldsymbol{\kappa}_1 + \boldsymbol{\kappa}_2$, $\omega_\infty = \omega_1 + \omega_2$, where $\boldsymbol{\kappa}_i = (k_i, l_i)$, $i = 1, 2, \infty$ are the wave vectors and ω_i are the frequencies of the solutions of the corresponding linearized KP equation, are precisely the same as those for the resonant interaction of three weakly nonlinear waves. Conceptually, however, such interactions more resemble so-called double resonance, in which not only the above resonance conditions are met but also the group velocities of two or more counterparts are equal and otherwise sporadic energy exchange within the resonant triplets is replaced by much more intense interactions (e. g. [145]).

The patterns of line soliton crests (Fig. 3) suggest that only attractive interactions may result in resonance. The resonant interactions between line solitons correspond to infinitely large phase shifts and infinitely strong attraction, and lead to durable changes of the geometry of the system. In the situation depicted in Fig. 3, the resonance would lead to an infinitely long common crest and thus render the "four-arm" pattern to a "three-arm" one.

In the KP framework, the resonance phenomenon is connected with the fact that its multi-soliton solutions representing the interaction of line solutions only exist for a subset of the space of parameters of the interacting solitons (Fig. 6). Resonance occurs when the parameters of the interacting solitons lie at a certain part of the border of this subset.

A new soliton born at the exact resonance may resonantly interact with yet another soliton. This mechanism has been employed in [17] to construct a family of exact solutions to the KP equation corresponding to resonance of several solitons. Such solutions consist of unequal numbers of incoming and outgoing line solitons, the parameters of which can be obtained from asymptotic analysis [16]. This class contains a variety of multisoliton solutions, many of which exhibit nontrivial spatial interaction patterns that are generally characteristic to multisoliton solutions to the KP equation [118].

Geometry of Oblique Interactions of KP Line Solitons

Patterns

A relatively simple but instructive, easily visualizable and rich-in-content model demonstrating the discussed features of interactions of line solitons is the KP equation. It admits simple explicit formulae for multi-soliton solutions and extensive analytical treatment of their geometrical properties. The two-soliton solution to the KP equation can be written as a sum $\eta = s_1 + s_2 + s_{12}$ of terms, reflecting to some extent the two interacting solitons s_1, s_2 and residue s_{12}

$$\begin{aligned} s_{1,2} &= \sqrt{A_{12}}k_{1,2}^2\Theta^{-2}\cosh\left(\varphi_{2,1}x + \ln\sqrt{A_{12}}\right) , \\ s_{12} &= 2\Theta^{-2}\left[(k_1-k_2)^2 + A_{12}(k_1+k_2)^2\right] , \\ \Theta &= \cosh\frac{\varphi_1-\varphi_2}{2} + \cosh\frac{\varphi_1+\varphi_2+\ln A_{12}}{2} . \end{aligned} \tag{7}$$

Here $\varphi_i = k_i x + l_i y + \omega_i t$, $a_i = \frac{1}{2}k_i^2$, $i = 1, 2$, are the phases and amplitudes of the interacting solitons, the 'frequencies' ω_i satisfy the dispersion relation $k_i\omega_i + k_i^4 + 3l_i^2 = 0$ of the linearized KP equation, $A_{12} = [\lambda^2 - (k_1-k_2)^2]/[\lambda^2 - (k_1+k_2)^2]$ is the phase shift parameter and $\lambda = l_1k_1^{-1} - l_2k_2^{-1}$.

The two-soliton solution only exists in a part of the parameter space $R^{(4)}(k_1, k_2, l_1, l_2)$. An interaction may result in either the positive or the negative phase shifts $\delta_{1,2} = -\ln A_{12}/|\kappa_{1,2}|$ of the counterparts (Fig. 6). The negative phase shift (cf. Fig. 3) can be attributed to an attractive interaction and the positive phase shift to a repulsive interaction.

The residue s_{12} (with an amplitude a_{12}) is at times considered as a virtual 2D "interaction soliton" [118], which grows into the resonant soliton at the limit of exact resonance $A_{12} \to \infty$ (cf. Fig. 7). This virtual structure can be heuristically interpreted as supported by both dispersive and diffractive effects, the latter being balanced by the oblique motion of the interacting solitons.

The patterns of both idealized (Fig. 3) and factual (Fig. 7) wave crests are symmetric with respect to a particular point called interaction center, and are stationary in a properly moving coordinate frame. If the amplitudes of the solitons are equal, the pattern is equivalent to the reflection of the incoming soliton from the wall following the motion of the interaction center. For unequal amplitude solitons the equivalence is not complete because mass and energy flux occur through such a wall.

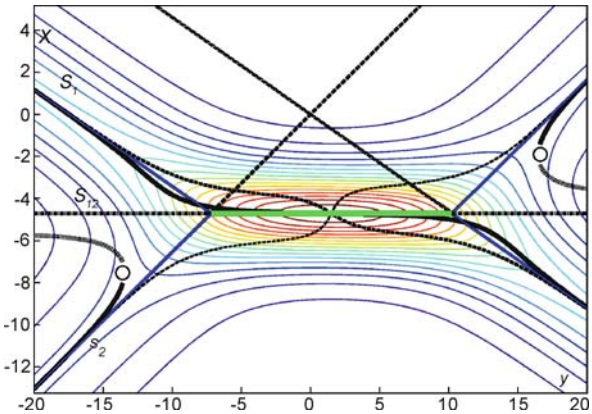

Solitons Interactions, Figure 7
Patterns of idealized (*solid/dashed straight lines*) and factual (*dashed curves*) crests of the incoming solitons s_1, s_2, the crest of the residue s_{12} (*green line*), the visible wave crests and troughs (*bold* and *bold dashed curves*, respectively), and isolines of the two-soliton solution in the case $l = l_1 = -l_2 = 0.3$, $k_1 = 0.6507$, $k_2 = 0.4599$ in normalized coordinates (x, y). The interaction center is the crossing point of all factual crests. Contour interval is $\frac{1}{4}\max(a_1, a_2) = 0.0689$. *Circles* show the common points of the troughs and crests of the two-soliton solution. Reprinted from [146]

Amplitudes

The phase shifts $\delta_{1,2}$ of the counterparts only depend on the amplitudes of the incoming solitons and the angle $\alpha_{12} = 2\arctan(\frac{1}{2}\lambda)$ between their crests. For the negative phase shift case $A_{12} > 1$ (that is typical in interactions of solitons with comparable amplitudes, Fig. 5), an interaction pattern emerges, the height of which exceeds the sum of the heights of two incoming solitons. The maximum surface elevation for equal amplitude solitons is $a_{\max} = 4a_{1,2}/(1 + A_{12}^{-1/2})$; thus, the nonlinear superposition of solitons may lead to a fourfold amplification of the surface elevation. In a highly idealized case of interactions of five solitons the maximum surface elevation may exceed the amplitude of the incoming solitons by more than an order. The largest elevation occurs if the interacting solitons are in exact resonance $A_{12} = \infty$, when solitons intersect under a physical angle $\tilde{\alpha}_{12} = 2\arctan\sqrt{3\tilde{\eta}/h}$. For unequal amplitude solitons the maximum elevation $a_{\max}$ for finite A_{12} and the amplitude of the resonant soliton a_∞ at $A_{12} = \infty$ are [46,146]

$$a_{\max} = a_{12} + 2A_{12}^{1/2}\frac{a_1 + a_2}{\left(A_{12}^{1/2} + 1\right)^2}, \quad a_\infty = \frac{(k_1+k_2)^2}{2}. \tag{8}$$

The near-resonant high hump is particularly narrow and its front is very steep. The maximum slope of the front of the two-soliton solution may be eight times as large as

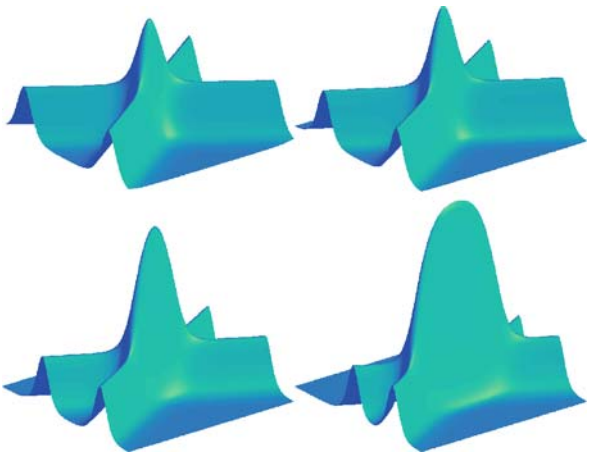

Solitons Interactions, Figure 8
Surface elevation in the vicinity of the interaction area, corresponding to the incoming solitons with equal amplitudes intersecting under different angles [120]

the slope of the incoming solitons [150]. For unequal amplitude solitons, the amplification of slope of the front of the interaction pattern is roughly proportional to the amplitude amplification. The extraordinary steepness of the front of the near-resonant hump, although intriguing, is not unexpected, because the resonant soliton is higher and therefore narrower than the incoming solitons.

The length L_{12} of the idealized common crest (Fig. 3) is $L_{12} \sim \ln A_{12}$ and therefore is modest unless the interacting solitons are near-resonant. The length of the area, where the total elevation exceeds the sum of the amplitudes of the counterparts, may be considerably longer; however it is also roughly proportional to $\ln A_{12}$ [150]. For largely different amplitudes of the interacting solitons, the amplitude amplification remains modest but the spatial extent of substantial influence of nonlinear interaction is roughly as large as if the amplitudes were equal [151]. Their near-resonant interaction becomes evident as a wave with its crest oriented and propagating differently compared with the counterparts. The described features are universal for that part of the wave vector space, in which the two-soliton solution with a negative phase shift exists. In the case of a positive phase shift (that is typical for interactions of solitons with largely different amplitudes) the highest elevation does not exceed the amplitude of the larger soliton.

Numerical simulations of the process of formation of the composite structure show that a high wave hump is formed quite fast in the interaction region and its height soon considerably exceeds the sum of the heights of the counterparts [58,124]. A high and gradually lengthening wave hump is also formed in cases when no exact two-soliton solution of the KP equations exists. The interaction in such cases is inelastic: neither the orientation of the crests nor the height of the solitons before and after interaction match each other, and a certain amount of wave radiation takes place. The described features become evident in a number of simulations of soliton interactions in different environments [105,168].

Soliton Interactions in Laboratory and Nature

Ion-Acoustic Solitons

The most productive studies of soliton interactions in the 1970s and the 1980s were performed in the framework of ion-acoustic waves. Their behavior is approximately described by the KP equation. An analytical solution describing interactions and resonance of two plane ion-acoustic solitons of small amplitude in the 3D collisionless plasma was given in [171]. This solution is wider than the two-soliton solution of the KP equation in the sense that the nearly unidirectional approximation is not used.

The first successful experiment proving that the 2D interaction of solitons has a wider spectrum of features compared with the 1D collisions is reported in [178]. A collision of two equal amplitude spherical solitons led to significant changes and to the emergence of a new 2D object. This interaction was not perfectly elastic: the solitons were not only shifted in phase but also reduced in amplitude. Soon afterwards, a near-resonant interaction of two KP solitons, with a large-amplitude wave hump of the counterparts oriented across the principal propagation direction, was observed near the interaction center, whereas the counterparts experienced clear phase shifts [43]. The amplitude amplification was found to be close to twofold the sum of the amplitudes of the incoming solitons [103].

The first adequate evidence of the nonlinear increase in amplitude during collisions of unequal amplitude KP solitons was also obtained in this framework [46]. For largely different amplitudes the measured maximum amplitudes were less than the predicted ones. A probable reason for the discrepancy is that the near-resonant wave hump occurs and covers a substantial area (equivalently, can be reliably estimated with the use of relatively large probes) only if the incoming solitons are very close to resonance. The common part of the crests of interacting solitons in the reported experiments was pretty short suggesting that the solitons were not exactly in resonance. Moreover, the complete identification of all the crests in the interactions of solitons with considerably different amplitudes is not possible because the crests are partially invisible (Fig. 7, [146]). The generic reason for experimental difficulties in this and similar experiments is the above-

discussed feature that large-amplitude structures tend to mask the presence of shorter ones.

Shallow-Water and Internal Solitons

Phase shifts, formation of a common crest of the interacting soliton-like waves, and accompanying amplitude and orientation changes are commonly observable in very shallow areas (Fig. 9). A famous photo by Terry Toedtemeier (1978), in which two solitonic shallow-water waves have significant phase shifts and quite a long common crest on a beach in the US state of Oregon, has been reproduced in a number of sources (e. g. p. 24 in [34]).

Although [93] predicted up to fourfold amplification of the surface elevation, much weaker amplitude amplification was found in early experiments [91]. An amplification close to the theoretical one was detected as a by-product of studies of shallow-water channel superconductivity [25].

Russell's "great wave of translation" [132] probably had a practically straight crest, reflecting a well-known feature of the generation of solitons in relatively narrow channels: a nearly perfectly 1D upstream soliton is created also by a 2D disturbance. The KdV model, which is frequently used to describe this phenomenon, is intrinsically 1D and certainly results in straight wave crests. An analysis of the simplified problem of ship motion in the KP equation and the Boussinesq equation shows, in accordance with common experience, that the wave crests are curved [71]. In fact, the essentially 2D waves with paraboloidal crests generally emerge in front of the disturbance. Their curvature monotonously decreases as the waves propagate [85,87].

Although the sidewall of the channel is not essential for the radiation of the upstream solitons, still it plays a crucial role in the transformation of already radiated waves. The straight-crested solitons are formed in the process of the (Mach) reflection of solitons with curved crests from the sidewalls [113]. This feature can be distinguished in many numerical studies, where an initially curved soliton starts to straighten when it comes in touch with the sidewall [149]. Since the reflection of solitons, which in this case is equivalent to soliton interactions, plays the key important role in creating straight-crested waves in channels, with a certain exaggeration it can be said that Russell's soliton exists due to soliton interactions.

Probably the most common solitonic phenomenon in nature next to the shallow water waves is represented by internal solitary waves in the oceans and in the atmosphere ([56,107,111,130], see also the entry ► Non-linear Internal Waves). In many cases, the KdV equation is an appropriate model for internal waves and the effects occurring

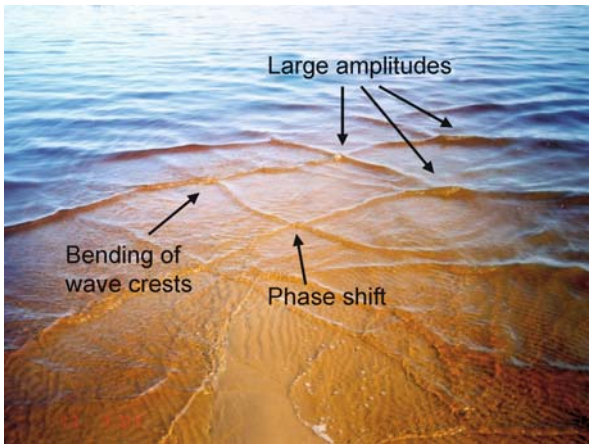

Solitons Interactions, Figure 9
Interaction patterns of solitonic surface waves in very shallow water near Kauksi resort on Lake Peipsi, Estonia. Photo and copyright by Lauri Ilison, July 2003. First published in [151]

during their interactions are equivalent to those described above.

Large-scale internal solitary waves and their groups exhibit many features unique to soliton interactions and are often called simply internal solitons. Intersecting solitary wave packets on Synthetic Aperture Radar (SAR) images frequently exhibit phase shifts while crossing each other [51], similarly to phase shifts occurring in the KP model, where the agreement between theory and experiments is quite good. Apel et al. [9] summarize advances in the descriptions and observations of internal solitons and their interactions.

The other environment for internal solitons, which provides some instructive aspects and is eligible, e. g., for the description of small-scale internal solitons in a two-layer medium, is the Benjamin–Ono (BO) equation [1]

$$\eta_t + c\eta_x + \alpha\eta\eta_x + \frac{\beta}{\pi}\frac{\partial^2}{\partial x^2} \quad P.V. \left[\int_{-\infty}^{\infty} \frac{\eta(x',t)}{x-x'}\mathrm{d}x'\right] = 0. \tag{9}$$

First, the BO solitons do not acquire a phase shift after a collision. Second, this environment demonstrates that generalization of integrable 1D soliton-admitting equations to even a weakly 2D case may lead to the loss of integrability, thus the preservation of this feature in the KP equation is not universal. The weakly 2D analogues of the BO equation [1] apparently are not integrable [9] and do not admit simple analytic multi-soliton solutions. Numerically simulated oblique interactions of weakly 2D BO solitons show that a phenomenon resembling the resonant

interaction of the KP solitons and an accompanied amplitude amplification do occur but the newly generated wave in the zone of nonlinear interaction is far from the BO soliton [105,167].

Solitons at Planetary Scale

Large-scale internal waves and solitons in nature are affected by the joint influence of the Earth's rotation and sphericity, and perfect internal line solitons generally cannot exist in the ocean. The physical reason of this "antisoliton theorem" is that due to rotational dispersion, there is always a phase synchronism between a source moving at an arbitrary speed and linear perturbations in such environments. This leads to unavoidable wave radiation from a large-scale solitary internal wave [49]; however this effect is negligible for relatively small-scale solitons, the typical time scales of which in the oceans are a few tens of minutes.

The existence of nonstationary solitary waves is still possible [52]. They exhibit a phenomenon similar to the recurrence of solitons – the repeating decay and reemergence process, and formation of a nearly localized wave packet consisting of a long-wave envelope and shorter, faster solitary-like waves that propagate through the envelope [61]. The robust, long-lived structure may contain as much as 50% of the energy in the initial solitary wave. Interacting packets may either pass through one another, or merge to form a longer packet. Their further studies may reveal interesting features concerning the functioning of oceans and other large stratified water bodies. Yet in the majority of practically interesting cases such long-term behavior is masked by topographic effects that modify the internal waves much more dramatically.

Another medium of particular practical importance as well as an instructive one from the viewpoint of soliton interactions is the large-scale geophysical flow in the oceans and the atmosphere. It is frequently treated as anisotropic quasi-2D turbulence, which has the property of energy concentration in large-scale structures. Eddies in such flows are stabilized by the background rotation and also affected by the joint effect of the Earth's rotation and sphericity. The latter becomes evident through the North-South variation of the Coriolis force, the so-called (planetary) beta-effect. In this environment, large-scale vortices often maintain their coherence for surprisingly long times and show an amazing variety of interactions; perhaps the most well-known example being the Jovian Great Red Spot.

Form-preserving, uniformly translating, horizontally localized solutions (of the equations of planetary dynamics) are called modons. In nondissipative quasi-geostrophic dynamics they exhibit a variety of properties of solitons. They carry a certain amount of fluid with them, which distinguishes them from (solitary) waves that cause, if at all, a finite displacement of the medium. The word modon was coined a decade after soliton was first used [157].

The standard quasi-geostrophic models predict that the accompanying Rossby-wave radiation should cause a decay of isolated eddies. Several models explain the persistence of monopolar vortices using external features of the flow, or interactions with a background shear that suppresses the Rossby-wave radiation [19,126,144]. A number of generalizations of the quasi-geostrophic model predict that, within a certain range of parameters, a nonlinear anticyclonic vortex of a special form can exist for a long time [121,153,170] due to the mutual compensation of weak (scalar) nonlinearity and weak dispersion. Some authors call such structures (Petviashvili-type) Rossby solitons, although their collisions are not elastic.

Long-living coherent vortices and their interactions have been studied in many experiments with rapidly rotating parabolic vessels [102]. The increase of the effective gravity due to the centrifugal acceleration towards the side walls modifies the governing equations, so that the (Petviashvili-type) Rossby soliton does not exist in the paraboloidal geometry [104]; yet anticyclonic vortices have a very long lifetime [154]. Another option to study the evolution of such vortices is a tank with a sloping bottom to imitate the planetary beta-effect (e. g. [41,90], among others; for an overview of earlier studies see [63]).

The majority of interactions of solitary vortices in such an environment are inelastic. The 2D localized vortices of comparable size and intensity, and like sign attract each other and generally coalesce soon. Only practically equivalent vortices may form a quasi-stable pair rotating around each other. The vortices of different signs either repulse from each other or form a relatively stable dipole [54,55]. The simultaneous evolution of a large number of soliton-like vortices has mostly been studied theoretically or numerically (e. g. in terms of a cluster of point vortices or hetons, [57]).

The simplest stable barotropic modon in flat-bottom beta-plane dynamics is the Larichev–Reznik dipole [83]. It is a traveling dipolar vortex with a characteristic north/south antisymmetry. It propagates eastward at any speed, or westward at speeds greater than the longest-wave speed. The speed of translation is therefore out of the range of the phase speed of Rossby waves, which is the reason why this modon does not radiate waves (unlike large-scale internal solitons). The significance of such structures in everyday

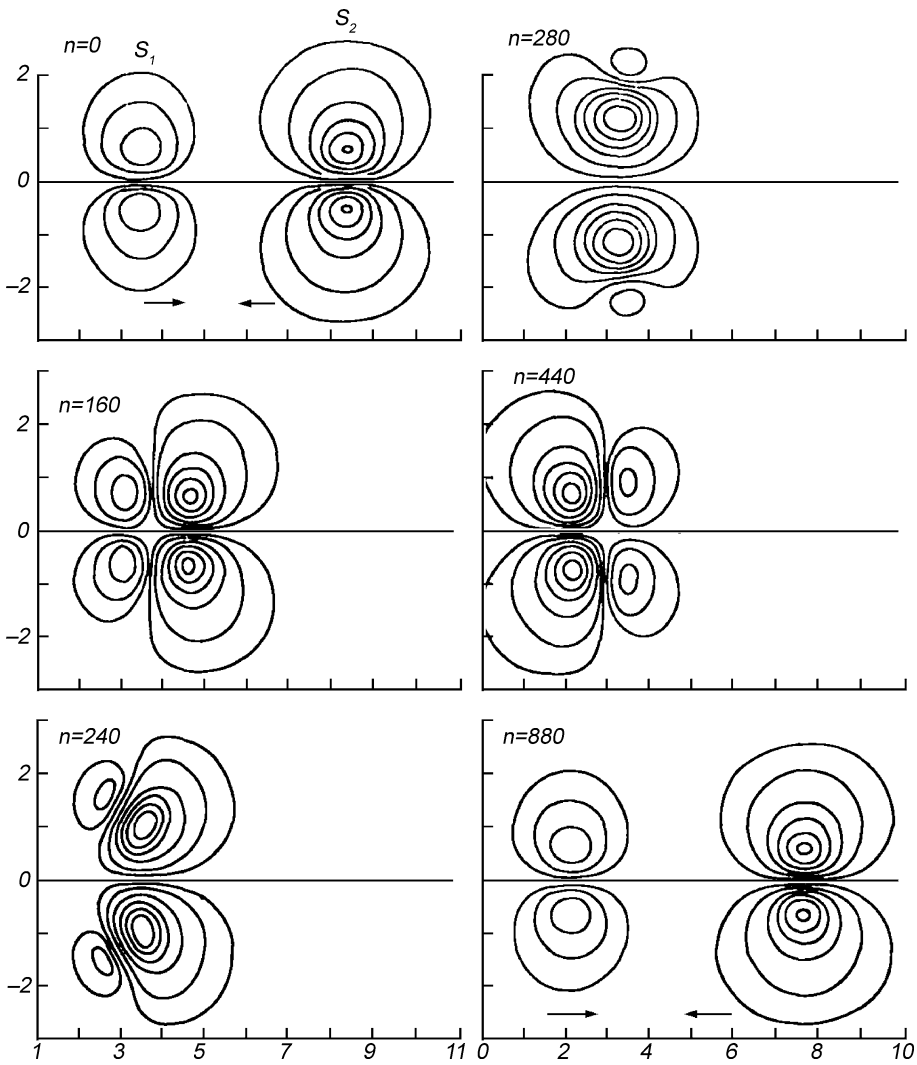

Solitons Interactions, Figure 10
Head-on collision of Larichev–Reznik dipoles [67]

life and climate dynamics (where they may be responsible, e. g., for certain unusual weather events since the vortex pair may behave completely differently compared with a single cyclone or anticyclone) is far from being well understood.

A part of the numerically simulated direct head-on and overtaking collisions of Larichev–Reznik modons are elastic [84]. An instructive feature of the head-on collisions is that the identity of all four vortices can be sometimes tracked during all the interaction phase (Fig. 10). The process resembles a collision of soft particle pairs connected by an elastic chain. Their oblique collisions generally are not elastic and may lead to the formation of new dipoles, tripoles, or to the destruction of the structures.

Effects in Higher Dimensions

Optical Spatial Solitons

Starting from the mid-1980s, various optical solitary entities have become a key arena for studies of the properties and interactions of solitary structures in multiple dimensions. The interest in these studies is continuously supported by a variety of existing and emerging applications of such solitons and their interactions in communication and computation technology (see also the entry ► Optical Computing). Another reason for the rapid progress in these studies is the relative ease of their generation in suitable materials, combined with the precision of contemporary optical experiments and the possibilities of precise

control over almost every parameter [136,152]. They offered a lot of the further conceptual progress of soliton interactions through making possible 3D propagation of solitary entities in laboratory conditions, albeit in many cases these entities did not match the classical nomenclature of solitons. They were called solitary waves in the field for several decades, because the governing equations are not necessarily integrable and the interactions not necessarily elastic. Still their interaction at times shows amazingly elastic nature and they are commonly called optical spatial solitons (OSS) in the modern nonlinear optics nomenclature [136].

A large number of various kinds of OSSs have been identified [76,135]. Observations of spatial Kerr solitons became possible from the mid-1980s, when so-called slab waveguides were built [3,13]. Their behavior substantially depends on the dimensions of the soliton and the waveguide. For example, the bright Kerr solitons are stable only in planar systems whereas the 2D solitons were found to collapse. Quasi-stable 2D solitons in Kerr media described by the cubic NLS in two spatial dimensions exist in the form of so-called necklace-ring beams [143].

Optical spatial solitons occurring due to the saturation of the nonlinear change in the refractive index were first demonstrated in the 1970s [18] and realized in a solid medium in the 1990s [74]. The net effect of the multiple physical effects involved in photorefractive materials is that the underlying nonlinearities are also saturable. Such solitons were also discovered in the early 1990s [137] and were found to be stable in both slab waveguides [75] and in bulk media [38].

So-called quadratic solitons (that consist of multifrequency waves coupled by means of a second-order nonlinearity) were predicted in the mid-1970s [70] and realized experimentally in the 1990s [164]. They can be thought of as vector solitons because they involve mutual self-trapping of two or more components [152]. Their interactions are similar to those occurring in other saturable nonlinear media.

Coherent and Incoherent Collisions

Coherent interactions of OSSs occur when the nonlinear medium responds quickly to the (interference) effects in the overlapping area of the beams. The increase of the intensity of light in the overlapping region of two parallel launched equivalent in-phase solitons in the focusing medium leads to an increase in the refractive index in that region. This in turn forces the centroid of each soliton toward it. Hence the solitons attract each other. The spatial distribution of the intensity of light resembles the temporal distribution of the amplitude of attracting interacting line solitons. In the case of antiphase beams, an opposite process occurs and such solitons exert repulsion. Coherent collisions of Kerr solitons have been demonstrated and the attraction for in-phase and the repulsion for antiphase solitons were clearly observed at the beginning of the 1990s [4,5,139]. The collision of nonequivalent coherent OSSs results in an energy exchange between the beams and in a repulsive force that makes the beams diverge.

Some reactions (e. g. photorefractive and thermal) of the optical medium are relatively slow. Solitons within which the phase varies randomly across the beam were first demonstrated in [95] and later extended to the beams of incoherent white light [94]. So-called incoherent soliton interactions occur when the response of the medium only follows the average intensity of light. Interactions of incoherent bright solitons are always attractive (Andersen and Lisak 1995).

The long-distance behavior of the counterparts in attractive interactions depends on the properties of the medium and the orientation of the beams. Pure Kerr solitons launched along nearly parallel directions result in periodic paths of the solitons' centroids. As in the case of the recurrence of the KdV solitons, Kerr solitons exactly return to their initial conditions after each cycle. Solitons launched under large enough converging angles exhibit a slight lateral deflection (which is equivalent to the phase shift of the KdV solitons). The solitons launched under large enough diverging angles never collide.

If the energy transfer were neglected, the force between the equivalent 1D Kerr solitons would vary from the maximum attractive between in-phase solitons to the maximum repulsive for antiphase solitons. This process leads to the (generally nonperiodic) energy transfer, where one soliton may gather net energy from the other. The details of the trajectories and other properties of the interacting solitons can be quite complex [152].

Interactions of 2D OSSs (e. g. in media with saturating nonlinearities) have a much larger variety of scenarios [152]; for example, interactions of incoherent photorefractive solitons may be both attractive and repulsive [156]. A number of resonance-like inelastic phenomena have been observed in which the number of solitons is not conserved. The *fusion* of two or more solitons into one structure resembles the resonance phenomenon. Here it happens owing to the gradual change of the orientation of the waveguides of attracting solitons launched under small relative angles into a new waveguide [48]. Differently from the resonance of line solitons, this happens for a certain range of initial parameters. The threshold is the maximum total internal reflection angle in the induced

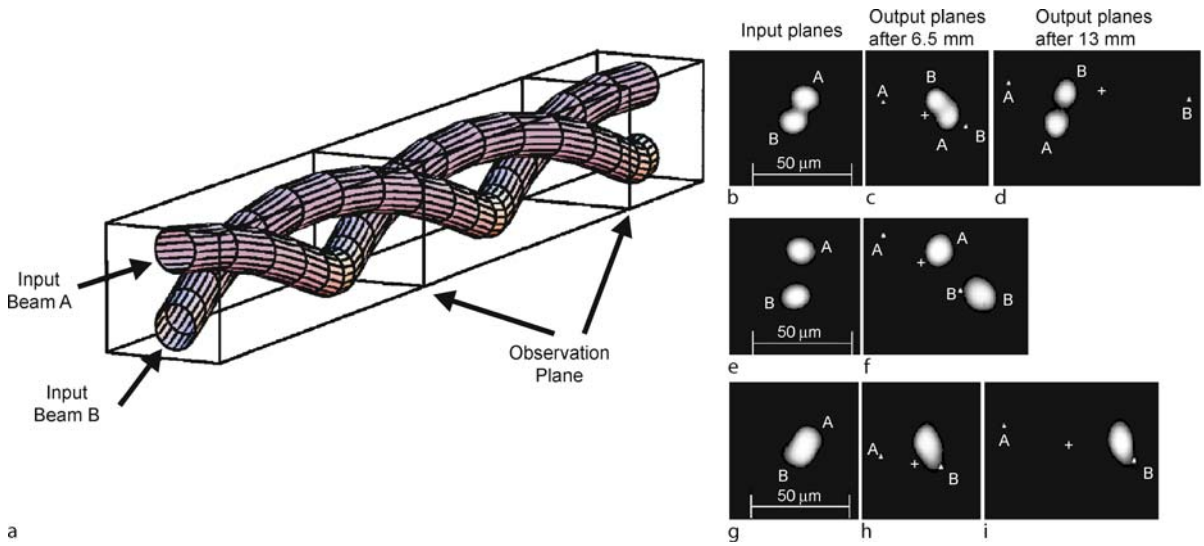

Solitons Interactions, Figure 11

a The experimentally observed soliton spiraling process. The arrows indicate the initial trajectories. b, e, and g show different input conditions and c, f, and h are the outputs after 6.3 mm and d and i after 13 mm. The triangles indicate the centers of the corresponding diffracting beams. From [152]

waveguide [140,142]. For larger collision angles, the solitons interact as described above. The solitons can fuse together already on the first merging or after a certain number of oscillations of their trajectories with decreasing amplitudes and periods [12,78,160,161]. A sort of inverse of this process is the breakup (*fission*, [142]) of optical solitons into new solitons upon their interaction [163]. *Annihilation* of solitons upon collision also may occur, when three solitons collided and only two emerged from the collision process [79].

Three-Dimensional Effects

In 3D space the trajectories of interacting beam solitons do not necessarily lie on a single plane and the system possesses nonzero initial angular momentum. The trajectories of solitons that are launched along skew lines passing close to each other may bend in three dimensions. The solitons *spiraling* around each other form a double helix orbit [123], much like two celestial objects or two moving charged particles moving along nearly parallel trajectories do. The interaction generally leads to spiraling-fusion or spiraling-repulsion of the trajectories [14,160,161]. Since the mutual rotation encounters a centrifugal force (which is always repulsive), a perfect double-helix DNA-like system, an interesting class of elastic interactions, requires soliton attraction. It may be formed from identical in-phase coherent solitons; yet such a double helix is unstable with respect to perturbations of both the relative phase and amplitude of one of the solitons. This effect resembles processes occurring during the interactions of vortices of different and like sign in quasi-2D rotating fluids.

Attractive interactions of incoherent solitons have made possible the observation of spiraling solitons in saturable media [141]. The two solitons of equal power orbit periodically about each other. Such a pair conserves angular momentum and is more or less stable for a certain range of the initial orientation and distance between the beams. Although the underlying saturable nonlinearities are described by nonintegrable equations, the spiraling process does not lead to measurable energy radiation [20]. When the initial distance between the solitons is increased, they do not form a long-living interacting system. The double spiral for beams located initially very close to each other converges fast and the counterparts eventually fuse. The pair apparently has a limited lifetime. A likely reason for its fusion or off-spiraling is the potential coherence of the light in the beams.

Interactions of Vector Solitons

Vector (composite) solitons consist of two or more components (also called modes) that mutually self-trap in a nonlinear medium. The simplest vector solitons, first suggested by Manakov [88], consist of two orthogonally

polarized components in a nonlinear Kerr medium in which self-phase modulation is identical to cross-phase modulation. They are described by an integrable system and interact elastically. An appropriate nonlinear material and experimental conditions for their demonstration were developed a long time after their discovery [68]. Similar solitons may exist if each field component is at a different frequency, and the frequency difference between components is much larger than the inverse of the nonlinearity relaxation time [138,166], or if the field components are incoherent with one another [22,30]. The energy exchange between the components of the interacting vector solitons takes place without radiative losses [7].

Vector solitons can also consist of different modes of their jointly induced waveguide [29,165]. These multimode solitons may possess multiple humps, may contain both bright and dark components [28,166], may have a certain internal dynamics, and were found to be stable (or weakly unstable) in large regions of their parameter space [96,109,110]. Interactions between vector solitons have, additionally to the above-discussed generic properties of solitons interactions (e. g. [86]), also some unique features. The shape transformations of colliding multimode solitons can lead to two different multimode solitons emerging from the collision process [6,69,80]. This "polarization switching" was predicted initially for Manakov solitons [125].

Composite multimode multihump solitons may carry topological charge in one of the vector components [21, 98]. This charge carried by a soliton of finite energy can be interpreted as spin of real particles. Elastic collisions of such 3D solitons should, ideally, conserve not only energy, and linear and angular momentum [122], but also the equivalent of spin. Two colliding composite solitons carrying opposite spins may form a metastable state that later decays into two or three new solitons. If the solitons interact under a certain critical angle, angular momentum is transferred from spin to orbital angular momentum. Finally, the shape transformation of the vortex component occurs at large collision angles, for which scalar solitons of all types simply go through one another unaffected or only have phase shifts [99,100]. The recent progress in optical vortices, vortex solitons, and their interactions is reviewed in [35].

Applications of Line Soliton Interactions

A generic use of soliton interactions in every case where solitary waves or topological solitons may exist, follows from the definition of a soliton. It consists of establishing if the structure is a solitonic one or not from the properties of their interactions.

The practical use of specific features of soliton interactions was started largely in parallel with the first technologies based on the use of solitons (e. g. optical soliton-based communications) more than a decade ago. The most promising is the technology of ► Optical Computing (see the relevant entry), which may be partially based on the interactions of vector solitons [155] and which may open completely new horizons in the architecture of computers.

Soliton Interactions on a Water Surface

Several applications based on properties of elastic soliton interactions (or changes of certain fields during such interactions) have been proposed in the framework of water waves. From the purely geometrical properties of oblique interactions of plane solitons one may extract, e. g., the water wave height. The relations for the phase shifts and for the intersection angle can be reduced to the following transcendental equation:

$$\delta_1 \sqrt{2a_1\left(1+\lambda^2/4\right)} = \pm \ln \frac{\delta_2^2\lambda^2 - 2\left(\delta_2-\delta_1\right)^2 a_1^2}{\delta_2^2\lambda^2 - 2\left(\delta_2+\delta_1\right)^2 a_1^2}. \quad (10)$$

If the intersection angle α_{12}, the phase shifts $\delta_{1,2}$ and their signs (or the wave propagation directions) are known, equation (10) uniquely defines the heights of the interacting solitons [118,119].

The phenomenon of drastic increase of surface elevation and slope of wave fronts described in Sect. "Geometry of Oblique Interactions of KP Line Solitons" can be attributed to formation of "freak" or "giant" waves in the ocean, the height and steepness of which are considerably larger than expected based on the classical wave statistics [73]. Their sudden appearance is a feature of paramount importance to and a generic source of danger for navigation and in coastal and offshore engineering. Interactions of envelope (Schrödinger) solitons and breather solutions of the NLS [108] in deep water are their potential source, although the model based on envelope soliton interactions underestimates the probability of large wave formation compared with the fully nonlinear model [31]. Interacting solitary wave groups that emerge from a long wave packet can produce freak wave events and may lead to a threefold increase of the wave slope [32].

The problem of extremely large-amplitude internal waves is by no means less important, as they can pose acute (currently neither well understood nor accounted for) danger for submarines, oil and gas platforms, oil risers and pipelines, and to other engineering constructions

in relatively deep water. The problem of "freak" internal waves and their generation due to internal soliton interactions has only recently attracted the attention of researchers [81].

There is a clear potential in the use of the advances concerning the geometry of the line soliton interactions in studies of both internal and shallow water solitons. While phase shifts usually have no dynamical significance, unexpectedly large elevations, extreme slopes, or changes in orientation of wave crests owing to the (Mach) reflection or oblique interactions of solitonic waves frequently cause an acute danger. The particularly high (stem) waves may lead to hits by high waves arriving from an unexpected direction. They may break during propagation into shallow water and attack armor blocks from an unprotected side [89] or overtop the breakwaters in unexpected locations [172]. These effects may be particularly pronounced in the case of ship wakes that often approach seawalls or breakwaters from other directions than wind waves do [148]. The possibility of drastic steepening of the front of the near-resonant structure is immaterial in many physical systems but is a crucial component of danger from freak waves [73], since specifically steep and high waves present an acute danger (e. g. [162]). Even if the steepness of the wave front is not large in absolute terms, relatively steep and long, tsunami-like waves tend to become asymmetric and exhibit unexpectedly large runup heights [36].

The amplitude amplification, however, evidently becomes effective relatively seldom, because two or more systems of long-crested solitonic waves must approach a certain area from different directions, and extreme elevations and slopes only occur if the amplitudes of the interacting solitons, the angle between their crests and the water depth are specifically balanced. The long life-time of the resulting wave hump in favorable conditions [73] may drastically increase the probability of the occurrence of abnormally high waves.

Ship-Induced Solitons and Wave Resistance

An increasing source of solitonic waves is the fast ship traffic in relatively shallow areas. The generation of shallow-water solitons is most effective for large ships sailing at speeds roughly equal to the maximum phase speed $\sqrt{gh}$ of surface gravity waves [87]. The low decay rates and exceptional compactness of the solitonic ship wakes has led to a significant impact on the safety of people, property and craft, unusually high hydrodynamic loads in a part of the nearshore, and to a considerable remote impact of the ship traffic in shallow areas [112,147]. Such a wake has probably caused a fatal accident as far as about 10 km from the sailing line already in 1912 [149]. The interactions of ship-induced solitons may lead to dangerous waves in the vicinity of coastal fairways and harbor entrances in otherwise sheltered areas [120].

A intriguing use of soliton interactions, combining effects occurring during an elastic oblique interaction of shallow-water solitons, followed by an annihilation of solitons and forced antisolitons, has been made for reducing the wave resistance in channels [25] and for catamaran design [26]. In a channel, the bow wave reflected from a side wall is used to cancel the stern wave, whereas the bow wave of one hull of the catamaran is used to cancel the stern wave of the other hull. Its practical use is possible for ships sailing at near-critical speeds, the wake of which may consist of a single soliton. The adequate calculation of the phase shift occurring during its Mach reflection (equivalently, during the interaction of two bow solitons) is a key component in achieving a perfect annihilation of properly timed bow solitons and the sterns' waves of depression.

The use of this effect by adjusting the channel width, the water depth and the ship's speed probably is the first intentional use of the features occurring during soliton interactions in the design of a certain technology [23,24]. At the exact design conditions the reflected bow wave completely cancels the stern wave and leads to a sort of channel superconductivity [25]. The same effect may occur between two ships moving in parallel [65].

Finally, it is interesting to note that already J.S. Russell may have been aware of this possibility. He describes efforts of a spirited horse which, pulling a boat in a canal, had drawn the boat up into its own wave leading to a significant reduction in resistance. The boat owner had noted this, and the event led to 'high-speed' service on some canals in the 1820s and 1830s [131]. We shall probably never know if the decrease of wave resistance occurred due to simple crossing of the critical speed or owing to the channel superconductivity; however, even the remote possibility that the practical use of soliton interactions may have been reported even before the first report about solitons, is remarkable.

Future Directions

Although the theory of soliton interactions has been extensively developed during four decades and its basic features for low-dimensional environments have been well understood, it is still in the stage of rapid development. Its first practical applications appeared only a decade ago and there is evidently room for relevant developments in (geo)physical applications. Analysis of long-term evolution of ensembles of solitons is still a challenge and reveals

ever new interesting features even in the simplest soliton-admitting systems such as the KdV equation. Although several formal procedures of building explicit multi-soliton solutions to many equations have been known since the 1970s, studies of properties of (optionally resonant) interactions of several line solitons only started at the turn of the millennium. Many fascinating features of elastic interactions of solitons of different kinds in more complex systems will obviously become evident in the nearest future. The results in this direction may largely widen understanding of the integrability of the underlying equations.

Extremely rapid progress may be expected in studies into properties of long-lived soliton-like structures, in particular, of optical spatial solitons. The classification of their nomenclature and interactions is far from being completed. This environment is the key framework for identifying the new features of soliton interactions in higher dimensions, even though interactions of such structures are not necessarily elastic.

Bibliography

1. Ablowitz MJ, Segur H (1981) Solitons and the inverse scattering transform. SIAM, Philadelphia
2. Ablowitz MJ, Kaup DJ, Newell AC, Segur H (1974) The inverse scattering transform – Fourier analysis for nonlinear problems. Stud Appl Math 53:249–315
3. Aitchison JS, Weiner AM, Silberberg Y, Oliver MK, Jackel JL, Leaird DE, Vogel EM, Smith PWE (1990) Observation of spatial optical solitons in a nonlinear glass wave-guide. Opt Lett 15(9):471–473
4. Aitchison JS, Silberberg Y, Weiner AM, Leaird DE, Oliver MK, Jackel JL, Vogel EM, Smith PWE (1991) Spatial optical solitons in planar glass wave-guides. J Opt Soc Am B-Optical Phys 8(6):1290–1297
5. Aitchison JS, Weiner AM, Silberberg Y, Leaird DE, Oliver MK, Jackel JL, Smith PWE (1991) Experimental-observation of spatial soliton-interactions. Opt Lett 16(1):15–17
6. Akhmediev N, Krolikowski W, Snyder AW (1998) Partially coherent solitons of variable shape. Phys Rev Lett 81(21):4632–4635
7. Anastassiou C, Segev M, Steiglitz K, Giordmaine JA, Mitchell M, Shih MF, Lan S, Martin J (1999) Energy-exchange interactions between colliding vector solitons. Phys Rev Lett 83(12):2332–2335
8. Anderson D, Lisak M (1985) Bandwidth limits due to incoherent soliton interaction in optical-fiber communication-systems. Phys Rev A 32(4):2270–2274
9. Apel JR, Ostrovsky LA, Stepanyants YA, Lynch JF (2007) Internal solitons in the ocean. J Acoust Soc Am 121(2):695–722
10. Arnold JM (1998) Varieties of solitons and solitary waves. Opt Quantum Electron 30:631–647
11. Askar'yan GA (1962) Effects of the gradient of strong electromagnetic beam on electrons and atoms. Sov Phys JETP 15(6):1088–1090
12. Baek Y, Schiek R, Stegeman GI, Baumann I, Sohler W (1997) Interactions between one-dimensional quadratic solitons. Opt Lett 22(20):1550–1552
13. Barthelemy A, Maneuf S, Froehly C (1985) Soliton propagation and self-confinement of laser-beams by Kerr optical nonlinearity. Opt Commun 55(3):201–206
14. Belic MR, Stepken A, Kaiser F (1999) Spiraling behavior of photorefractive screening solitons. Phys Rev Lett 82(3):544–547
15. Berger V, Kohlhase S (1976) Mach-reflection as a diffraction problem. In: Proc 25th Int Conf on Coastal Eng. ASCE, New York, pp 796–814
16. Biondini G, Chakravarty S (2006) Soliton solutions of the Kadomtsev–Petviashvili II equation. J Math Phys 47(3):033514
17. Biondini G, Kodama Y (2003) On a family of solutions of the Kadomtsev–Petviashvili equation which also satisfy the Toda lattice hierarchy. J Phys A-Math General 36(42):10519–10536
18. Bjorkholm JE, Ashkin A (1974) cw self-focusing and self-trapping of light in sodium vapor. Phys Rev Lett 32(4):129–132
19. Busse FH (1994) Convection driven zonal flows and vortices in the major planets. Chaos 4:123–134
20. Buryak AV, Kivshar YS, Shih MF, Segev M (1999) Induced coherence and stable soliton spiraling. Phys Rev Lett 82(1):81–84
21. Carmon T, Anastassiou C, Lan S, Kip D, Musslimani ZH, Segev M, Christodoulides D (2000) Observation of two-dimensional multimode solitons. Opt Lett 25(15):1113–1115
22. Chen ZG, Segev M, Coskun TH, Christodoulides DN (1996) Observation of incoherently coupled photorefractive spatial soliton pairs. Opt Lett 21(18):1436–1438
23. Chen X-N, Sharma SD (1994) Nonlinear theory of asymmetric motion of a slender ship in a shallow channel. In: Rood EP (ed) 20th Symposium on Naval Hydrodynamics. US Office on Naval Research, Santa Barbara, pp 386–407
24. Chen X-N, Sharma SD (1997) Zero wave resistance for ships moving in shallow channels at supercritical speeds. J Fluid Mech 335:305–321
25. Chen X-N, Sharma SD, Stuntz N (2003) Zero wave resistance for ships moving in shallow channels at supercritical speeds. Part 2. Improved theory and model experiment. J Fluid Mech 478:111–124
26. Chen X-N, Sharma SD, Stuntz N (2003) Wave reduction by S-catamaran at supercritical speeds. J Ship Res 47:145–154
27. Chiao RY, Garmire E, Townes CH (1964) Self-trapping of optical beams. Phys Rev Lett 13(15):479–482
28. Christodoulides DN (1988) Black and white vector solitons in weakly birefringent optical fibers. Phys Lett A 132(8–9):451–452
29. Christodoulides DN, Joseph RI (1988) Vector solitons in birefringent nonlinear dispersive media. Opt Lett 13(1):53–55
30. Christodoulides DN, Singh SR, Carvalho MI, Segev M (1996) Incoherently coupled soliton pairs in biased photorefractive crystals. Appl Phys Lett 68(13):1763–1765
31. Clamond D, Grue J (2002) Interaction between envelope solitons as a model for freak wave formations. Part I: Long time interaction. Comptes Rendus Mecanique 330(8):575–580
32. Clamond D, Francius M, Grue J, Kharif C (2006) Long time interaction of envelope solitons and freak wave formations. Eur J Mech B-Fluids 25(5):536–553
33. Davis RE, Acrivos A (1967) Solitary internal waves in deep water. J Fluid Mech 29(3):593–608

34. Dauxois T, Peyrard M (2006) Physics of solitons. Cambridge University Press, Cambridge
35. Desyatnikov AS, Kivshar YS, Torner L (2005) Optical vortices and vortex solitons. Prog Opt 47:291–391
36. Didenkulova II, Zahibo N, Kurkin AA, Levin BV, Pelinovsky EN, Soomere T (2006) Runup of nonlinearly deformed waves on a coast. Doklady Earth Sci 411(8):1241–1243
37. Drazin PG, Johnson RS (1989) Solitons: An introduction. In: Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge
38. Duree GC, Shultz JL, Salamo GJ, Segev M, Yariv A, Crosignani B, Diporto P, Sharp EJ, Neurgaonkar RR (1993) Observation of self-trapping of an optical beam due to the photorefractive effect. Phys Rev Lett 71(4):533–536
39. Engelbrecht J, Salupere A (2005) On the problem of periodicity and hidden solitons for the KdV model. Chaos 15:015114
40. Fermi E, Pasta J, Ulam S (1955) Studies of nonlinear problems. I Los Alamos report LA-1940; (1965) In: Segré E (ed) Collected papers of Enrico Fermi. University of Chicago Press, Chicago
41. Firing E, Beardsley RC (1976) The behaviour of a barotropic eddy on a beta-plane. J Phys Oceanogr 6:57–65
42. Flierl GR, Larichev VD, Mcwilliams JC, Reznik GM (1980) The dynamics of baroclinic and barotropic solitary eddies. Dyn Atmos Oceans 5(1):1–41
43. Folkes PA, Ikezi H, Davis R (1980) Two-dimensional interaction of ion-acoustic solitons. Phys Rev Lett 45(11):902–904
44. Fuerst RA, Canva MTG, Baboiu D, Stegeman GI (1997) Properties of type II quadratic solitons excited by imbalanced fundamental waves. Opt Lett 22(23):1748–1750
45. Funakoshi M (1980) Reflection of obliquely incident large-amplitude solitary wave. J Phys Soc Japan 49:2371–2379
46. Gabl EF, Lonngren KE (1984) On the oblique collision of unequal amplitude ion-acoustic solitons in a field-free plasma. Phys Lett A 100:153–155
47. Gardner CS, Greene JM, Kruskal MD, Miura RM (1974) Korteweg–de Vries equations and generalizations: methods for exact solutions. Commun Pure Appl Math 27:97–133
48. Gatz S, Herrmann J (1992) Soliton collision and soliton fusion in dispersive materials with a linear and quadratic intensity depending refraction index change. IEEE J Quantum Electron 28(7):1732–1738
49. Galkin VM, Stepanyants YA (1991) On the existence of stationary solitary waves in a rotating fluid. PMM J Appl Math Mech 55(6):939–943
50. Gardner CS, Green JM, Kruskal MD, Miura RM (1967) Method for solving the Kortweg–de Vries equation. Phys Rev Lett 19:1095–1097
51. Gasparovic RF, Apel JR, Kasischke ES (1988) An overview of the SAR internal wave signature experiment. J Geophys Res-Oceans 93(C10):12304–12316
52. Gilman OA, Grimshaw R, Stepanyants YA (1996) Dynamics of internal solitary waves in a rotating fluid. Dyn Atmos Oceans 23(1–4):403–411
53. Gordon JP (1983) Interaction forces among solitons in optical fibers. Opt Lett 8(11):596–598
54. Griffits RW, Hopfinger EJ (1986) Experiments with baroclinic vortex pairs in a rotating fluid. J Fluid Mech 173:501–518
55. Griffiths RW, Hopfinger EJ (1987) Coalescing of geostrophic vortices. J Fluid Mech 178:73–97
56. Grimshaw R (2001) Internal solitary waves. In: Grimshaw R (ed) Environmental Stratified Flows. Kluwer, Dordrecht, pp 1–30
57. Gryanik VM, Doronina TN, Olbers DJ, Warncke TH (2000) The theory of three-dimensional hetons and vortex-dominated spreading in localized turbulent convection in a fast rotating stratified fluid. J Fluid Mech 423:71–125
58. Haragus-Courcelle M, Pego RL (2000) Spatial wave dynamics of steady oblique wave interactions. Physica D 145:207–232
59. Hasegawa A (1989) Optical Solitons in Fibers. Springer, Berlin
60. Hasegawa A, Tappert F (1973) Transmission of stationary nonlinear optical pulses in dispersive dielectric fiber: II Normal dispersion. Appl Phys Lett 23:171–172
61. Helfrich KR (2007) Decay and return of internal solitary waves with rotation. Phys Fluid 19(2):026601
62. Hirota R (1971) Exact solution of the Korteweg–de Vries equation for multiple collisions of solitons. Phys Rev Lett 27:1192–1194
63. Hopfinger EJ, van Heijst GJF (1993) Vortices in rotating fluids. Annu Rev Fluid Mech 25:241–289
64. Ikezi H, Taylor RJ, Baker DR (1970) Formation and interaction of ion-acoustic solitons. Phys Rev Lett 25(1):11–14
65. Jiankang W, Lee TS, Chu C (2001) Numerical study of wave interaction generated by two ships moving parallely in shallow water. Comput Meth Appl Mech Engrg 190:2099–2110
66. Kadomtsev BB, Petviashvili VI (1970) The stability of solitary waves in weakly dispersive media. Dokl Akad Nauk SSSR 192:532–541
67. Kamenkovich et al. (1987) Synoptic eddies in the ocean. Russian version. Gidrometeoizdat, Leningrad, pp 124
68. Kang JU, Stegeman GI, Aitchison JS, Akhmediev N (1996) Observation of Manako vspatial solitons in AlGaAs planar waveguides. Phys Rev Lett 76(20):3699–3702
69. Kanna T, Lakshmanan M (2003) Exact soliton solutions of coupled nonlinear Schrödinger equations: Shape-changing collisions, logic gates, and partially coherent solitons. Phys Rev E 67(4):046617 Part 2
70. Karamzin YN, Sukhorukov AP (1976) Mutual focusing of high-power light. Sov Phys JETP 41:414–420
71. Katsis C, Akylas TR (1987) On the excitation of long nonlinear water waves by a moving pressure distribution. Part 2. Three-dimensional effects. J Fluid Mech 177:49–65
72. Kaup DJ, Newell AC (1978) Solitons as particles, oscillators, and in slowly changing media: A singular pertubation theory. Proc R Soc London A 361(4):413–446
73. Kharif C, Pelinovsky N (2003) Physical mechanisms of the rogue wave phenomenon. Eur J Mech B Fluids 22:603–634
74. Khitrova G, Gibbs HM, Kawamura Y, Iwamura H, Ikegami T, Sipe JE, Ming L (1993) Spatial solitons in a self-focusing semiconductor gain medium. Phys Rev Lett 70(7):920–923
75. Kip D, Wesner M, Shandarov V, Moretti P (1998) Observation of bright spatial photorefractive solitons in a planar strontium barium niobate waveguide. Opt Lett 23(12):921–923
76. Kivshar YS, Luther-Davies B (1998) Dark optical solitons: physics and applications. Phys Rep – Rev Sect Phys Lett 298(2–3):81–197
77. Kivshar YS, Malomed BA (1989) Dynamics of solitons in nearly integrable systems. Rev Mod Phys 61(4):768–915
78. Krolikowski W, Holmstrom SA (1997) Fusion and birth of spatial solitons upon collision. Opt Lett 22(6):369–377

79. Krolikowski W, Luther-Davies B, Denz C, Tschudi T (1998) Annihilation of photorefractive solitons. Opt Lett 23(2):97–99
80. Krolikowski W, Akhmediev N, Luther-Davies B (1999) Collision-induced shape transformations of partially coherent solitons. Phys Rev E 59(4):4654–4658
81. Kurkin AA, Pelinovsky EN (2004) Freak waves: Facts, theory and modelling. Nizhny Novgorod State Technical University, Nizhny Novgorod (in Russian)
82. Lakhsmanan M, Rajasekhar S (2003) Nonlinear dynamics: integrability, chaos and patterns. Springer, Berlin
83. Larichev VD, Reznik GM (1976) Strongly non-linear, two-dimensional isolated Rossby waves. Oceanology 16:961–967
84. Larichev VD, Reznik GM (1983) Collision of two-dimensional solitary Rossby waves. Oceanology 23(5):725–734
85. Lee SJ, Grimshaw RHJ (1990) Upstream-advancing waves generated by three-dimensional moving disturbances. Phys Fluids A 2:194–201
86. Leo G, Assanto G (1997) Collisional interactions of vectorial spatial solitary waves in type II frequency-doubling media. J Opt Soc Am B-Opt Phys 14(11):3151–3161
87. Li Y, Sclavounos PD (2002) Three-dimensional nonlinear solitary waves in shallow water generated by an advancing disturbance. J Fluid Mech 470:383–410
88. Manakov SV (1974) On the theory of two-dimensional stationary self-focusing of electromagnetic waves. Sov Phys JETP 38:248–253
89. Mase H, Memita T, Yuhi M, Kitano T (2002) Stem waves along vertical wall due to random wave incidence. Coast Eng 44:339–350
90. Masuda A, Marubayashi K, Ishibashi M (1990) A laboratory experiment and numerical simulation of an isolated barotropic eddy in a basin with topographic β. J Fluid Mech 213:641–659
91. Melville WK (1980) On the Mach reflection of solitary waves. J Fluid Mech 98:285–297
92. Miles JW (1977) Obliquely interacting solitary waves. J Fluid Mech 79:157–169
93. Miles JW (1977) Resonantly interacting solitary waves. J Fluid Mech 79:171–179
94. Mitchell M, Segev M (1997) Self-trapping of incoherent white light. Nature 387(6636):880–883
95. Mitchell M, Chen ZG, Shih MF, Segev M (1996) Self-trapping of partially spatially incoherent light. Phys Rev Lett 77(3):490–493
96. Mitchell M, Segev M, Christodoulides DN (1998) Observation of multihump multimode solitons. Phys Rev Lett 80(21):4657–4660
97. Mollenauer LF, Stolen RH, Gordon JP (1980) Experimental-observation of picosecond pulse narrowing and solitons in optical fibers. Phys Rev Lett 45(13):1095–1098
98. Musslimani ZH, Segev M, Christodoulides DN, Soljacic M (2000) Composite multihump vector solitons carrying topological charge. Phys Rev Lett 84(6):1164–1167
99. Musslimani ZH, Soljacic M, Segev M, Christodoulides DN (2001) Interactions between two-dimensional composite vector solitons carrying topological charges. Phys Rev E 63(6):066608
100. Musslimani ZH, Soljacic M, Segev M, Christodoulides DN (2001) Delayed-action interaction and spin-orbit coupling between solitons. Phys Rev Lett 86(5):799–802
101. Newell AC, Redekopp LG (1977) Breakdown of Zakharov–Shabat theory and soliton creation. Phys Rev Lett 38(8):377–380
102. Nezlin VM, Sneshkin EN (1993) Rossby vortices, spiral structures, solitons: Astrophysics and plasma physics in shallow water experiments. Springer, Berlin
103. Nishida Y, Nagasawa T (1980) Oblique collision of plane ion-acoustic solitons. Phys Rev Lett 45(20):1626–1629
104. Nycander J (1993) The difference between monopole vortices in planetary flows and laboratory experiments. J Fluid Mech 254:561–577
105. Oikawa M, Tsuji H (2006) Oblique interactions of weakly nonlinear long waves in dispersive systems. Fluid Dyn Res 38:868–898
106. Osborne AR, Bergamasco L (1986) The solitons of Zabusky and Kruskal revisited: perspective in terms of the periodic spectral transform. Physica D 18:26–46
107. Osborne AR, Burch TL (1980) Internal solitons in the Andaman Sea. Science 208(4443):451–460
108. Osborne AR, Onorato M, Serio M (2000) The nonlinear dynamics of rogue waves and holes in deep-water gravity wave trains. Phys Lett A 275(5–6):386–393
109. Ostrovskaya EA, Kivshar YS, Chen ZG, Segev M (1999) Interaction between vector solitons and solitonic gluons. Opt Lett 24(5):327–329
110. Ostrovskaya EA, Kivshar YS, Skryabin DV, Firth WJ (1999) Stability of multihump optical solitons. Phys Rev Lett 83(2):296–299
111. Ostrovsky LA, Stepanyants YA (1989) Do internal solitons exist in the ocean? Rev Geophys 27:293–310
112. Parnell KE, Kofoed-Hansen H (2001) Wakes from large high-speed ferries in confined coastal waters: Management approaches with examples from New Zealand and Denmark. Coastal Manage 29:217–237
113. Pedersen G (1988) Three-dimensional wave patterns generated by moving disturbances at transcritical speeds. J Fluid Mech 196:39–63
114. Pedersen NF, Samuelsen MR, Welner D (1984) Soliton annihilation in the perturbed sine-Gordon system. Phys Rev B 30(7):4057–4059
115. Peregrine DH (1983) Wave jumps and caustics in the propagation of finite-amplitude water waves. J Fluid Mech 136:435–452
116. Perring JK, Skyrme THR (1962) A Model Unified Field Equation. Nucl Phys 31:550–555
117. Perroud PH (1957) The Solitary Wave Reflection Along a Straight Vertical Wall at Oblique Incidence. Univ. of Calif. Berkeley, IER Rept 99-3, pp 93
118. Peterson P, van Groesen E (2000) A direct and inverse problem for wave crests modelled by interactions of two solitons. Physica D 141:316–332
119. Peterson P, van Groesen E (2001) Sensitivity of the inverse wave crest problem. Wave Motion 34:391–399
120. Peterson P, Soomere T, Engelbrecht J, van Groesen E (2003) Soliton interaction as a possible model for extreme waves in shallow water. Nonlinear Process Geophys 10:503–510
121. Petviashvili VI (1980) Red Spot of Jupiter and the drift soliton in plasma. JETP Lett 32:619–622
122. Pigier C, Uzdin R, Carmon T, Segev M, Nepomnyaschchy A, Musslimani ZH (2001) Collisions between $(2+1)$D rotating propeller solitons. Opt Lett 26(20):1577–1579

123. Poladian L, Snyder AW, Mitchell DJ (1991) Spiraling spatial solitons. Opt Commun 85(1):59–62
124. Porubov AV, Tsuji H, Lavrenov IV, Oikawa M (2005) Formation of the rogue wave due to non-linear two-dimensional waves interaction. Wave Motion 42:202–210
125. Radhakrishnan R, Lakshmanan M, Hietarinta J (1997) Inelastic collision and switching of coupled bright solitons in optical fibers. Phys Rev E 56(2):2213–2216
126. Read PL, Hide R (1983) Long-lived eddies in the laboratory and in the atmosphere of Jupiter and Saturn. Nature 302(10):126–129
127. Rebbi C (1979) Solitons. Sci Am 240(2):76–91
128. Redekopp LG, Weidman PD (1978) Solitary Rossby waves in zonal shear flows and their interactions. J Atm Sci 35:790–804
129. Rotschild C, Alfassi B, Cohen O, Segev M (2006) Long-range interactions between optical solitons. Nat Phys 2(11):769–774
130. Rottman JW, Grimshaw R (2001) Atmospheric internal solitary waves. In: Grimshaw R (ed), Environmental Stratified Flows. Kluwer, Dordrecht, pp 91–129
131. Russell JS (1837) Applications and illustrations of the law of wave in the practical navigation of canals. Trans R Soc Edin 14:33–34
132. Russell JS (1844) Report on waves. In: Report of the 14th Meeting of the British Association for the Advancement of Science. Murray, York, pp 311–390
133. Salupere A, Peterson P, Engelbrecht J (2002) Long-time behaviour of soliton ensembles. Part I – Emergence of ensembles. Chaos Solitons Fractals 14(9):1413–1424
134. Salupere A, Peterson P, Engelbrecht J (2003) Long-time behaviour of soliton ensembles. Part II – Periodical patterns of trajectories. Chaos Solitons Fractals 15:29–40
135. Segev M (1998) Optical spatial solitons. Opt Quantum Electron 30(7–10):03–533
136. Segev M, Stegeman G (1998) Self-trapping of optical beams: Spatial solitons. Phys Today 51(8):42–48
137. Segev M, Crosignani B, Yariv A, Fischer B (1992) Spatial solitons in photorefractive media. Phys Rev Lett 68(7):923–926
138. Shalaby M, Barthelemy AJ (1992) Observation of the self-guided propagation of a dark and bright spatial soliton pair in a focusing nonlinear medium. IEEE J Quantum Electron 28(12):2736–2741
139. Shalaby M, Reynaud F, Barthelemy A (1992) Experimental-observation of spatial soliton-interactions with a pi-2 relative phase difference. Opt Lett 17(11):778–780
140. Shih MF, Segev M (1996) Incoherent collisions between two-dimensional bright steady-state photorefractive spatial screening solitons. Opt Lett 21(19):1538–1540
141. Shih MF, Segev M, Salamo G (1997) Three-dimensional spiraling of interacting spatial solitons. Phys Rev Lett 78(13):2551–2554
142. Snyder AW, Sheppard AP (1993) Collisions, steering, and guidance with spatial solitons. Opt Lett 18(7):482–484
143. Soljacic M, Sears S, Segev M (1998) Self-trapping of "necklace" beams in self-focusing Kerr media. Phys Rev Lett 81(22):4851–4854
144. Sommeria J, Nore C, Dumont T, Robert R (1991) Statistical theory of the Great Red Spot of Jupiter. C R Acad Sci II 312:999–1005
145. Soomere T (1992) Geometry of the double resonance of Rossby waves. Ann Geophys 10:741–748
146. Soomere T (2004) Interaction of Kadomtsev–Petviashvili solitons with unequal amplitudes. Phys Lett A 332:74–81
147. Soomere T (2005) Fast ferry traffic as a qualitatively new forcing factor of environmental processes in non-tidal sea areas: a case study in Tallinn Bay, Baltic Sea. Environ Fluid Mech 5:4 293–323
148. Soomere T (2006) Nonlinear ship wake waves as a model of rogue waves and a source of danger to the coastal environment: a review. Oceanologia 48(S):185–202
149. Soomere T (2007) Nonlinear components of ship wake waves. Appl Mech Rev 60(3):120–138
150. Soomere T, Engelbrecht J (2005) Extreme elevations and slopes of interacting solitons in shallow water. Wave Motion 41:179–192
151. Soomere T, Engelbrecht J (2006) Weakly two-dimensional interaction of solitons in shallow water. Eur J Mech B Fluids 25(5):636–648
152. Stegeman GI, Segev M (1999) Optical spatial solitons and their interactions: Universality and diversity. Science 286(5444):1518–1523. doi:10.1126/science.286.5444.1518
153. Stegner A, Zeitlin V (1996) Asymptotic expansions and monopolar solitary Rossby vortices in barotropic and two-layer model. Geophys Astrophys Fluid Dyn 83:159–195
154. Stegner A, Zeitlin V (1998) From quasi-geostrophic to strongly non-linear monopolar vortices in a paraboloidal shallow-water experiment. J Fluid Mech 356:1–24
155. Steiglitz K (2001) Time-gated Manakov spatial solitons are computationally. Phys Rev E 63(1):016608
156. Stepken A, Kaiser F, Belic MR, Krolikowski W (1998) Interaction of incoherent two-dimensional photorefractive solitons. Phys Rev E 58(4):R4112–R4115
157. Stern ME (1975) Minimal properties of planetary eddies. J Mar Res 33:1–13
158. Taijiri M, Maesono H (1997) Resonant interactions of drift vortex solitons in a convective motion of a plasma. Phys Rev E 55(3):3351–3357
159. Tanaka M (1993) Mach reflection of a large-amplitude solitary wave. J Fluid Mech 248:637–661
160. Tikhonenko V, Christou J, Luther-Davies B (1995) Spiraling bright spatial solitons formed by the breakup of an optical vortex in a saturable self-focusing medium. J Opt Soc Am B-Opt Phys 12(11):2046–2052
161. Tikhonenko V, Christou J, Luther-Davies B (1996) Three dimensional bright spatial soliton collision and fusion in a saturable nonlinear medium. Phys Rev Lett 76(15):2698–2701
162. Toffoli A, Lefevre JM, Bitner-Gregersen E, Monbaliu J (2005) Towards the identification of warning criteria: Analysis of a ship accident database. Appl Ocean Res 27(6):281–291
163. Torner L, Torres JP, Menyuk CR (1996) Fission and self-deflection of spatial solitons by cascading. Opt Lett 21(7):462–464
164. Torruellas WE, Wang Z, Hagan DJ, Vanstryland EW, Stegeman GI, Torner L, Menyuk CR (1995) Observation of 2-dimensional spatial solitary waves in a quadratic medium. Phys Rev Lett 74(25):5036–5039
165. Tratnik MV, Sipe JE (1988) Bound solitary waves in a birefringent optical fiber. Phys Rev A 38(4):2011–2017
166. Trillo S, Wabnitz S, Wright EM, Stegeman GI (1988) Optical solitary waves induced by cross-phase modulation. Opt Lett 13(10):871–873

167. Tsuji H, Oikawa M (2001) Oblique interaction of internal solitary waves in a two-layer fluid of infinite depth. Fluid Dyn Res 29:251–267
168. Tsuji H, Oikawa M (2004) Two-dimensional interaction of solitary waves in a modified Kadomtsev–Petviashvili equation. J Phys Soc Japan 73:3034–3043
169. Washimi H, Taniuti T (1966) Propagation of ion-acoustic solitary waves of small amplitude. Phys Rev Lett 17:996–998
170. Williams GP, Yamagata T (1984) Geostrophic regimes, intermediate solitary vortices and Jovian eddies. J Atm Sci 41:453–478
171. Yajima N, Oikawa M, Satsuma J (1978) Interaction of ion-acoustic solitons in three-dimensional space. J Phys Soc Japan 44:1711–1714
172. Yoon SB, Liu PLF (1989) Stem waves along breakwater. J Waterw Port Coast Ocean Eng – ASCE 115:635–648
173. Yue DK, Mei CC (1980) Forward diffrection of Stokes waves by a thin wedge. J Fluid Mech 99:33–52
174. Zabusky NJ, Kruskal MD (1965) Interaction on "solitons" in a collisionless plasma and the recurrence of initial states. Phys Rev Lett 15(6):240–243
175. Zakharov VE (1968) Stability of periodic waves of finite amplitude on the surface of a deep fluid. Zh Prikl Mekh Tekh Fiz 9:86–94; J Appl Mech Tech Phys 9:190–194
176. Zakharov VE, Shabat AB (1972) Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media. Sov Phys JETP 34:62–69
177. Zakharov VE, Manakov SV, Novikov SP, Pitaevsky LP (1980) Theory of solitons. Nauka, Moscow (in Russian) (English translation, 1984, Consultants Bureau, New York)
178. Ze F, Hershkowitz N, Chan C, Lonngren KE (1979) Inelastic collision of spherical ion-acoustic solitons. Phys Rev Lett 42(26):1747–1750

Solitons, Introduction to

Mohamed A. Helal
Department of Mathematics, Faculty of Science,
Cairo University, Giza, Egypt

The concept of solitons reflects one of the most important developments in science at the second half of the 20th century: the nonlinear description of the world. It is not easy to give a comprehensive and precise definition of a soliton. Frequently, a *soliton* is explained as a spatially localized wave in a medium that can interact strongly with other solitons but will afterwards regain its original form.

It is a nonlinear pulse-like wave that can exist in some nonlinear systems. The soliton wave can propagate without dispersing its energy over a large region of space; collision of two solitons leads to unchanged forms, solitons also exhibit particle-like properties.

The most remarkable property of solitons is that they do not disperse and thus conserve their form during propagation and collision. ► Solitons Interactions.

The nonlinear science has been growing for approximately fifty years. However, numerous nonlinear processes had been previously identified, but the nonlinear mathematical tools were not developed. The available tools were linear, and nonlinearities were avoided or treated as perturbations of linear theories.

The first experimental observation of a solitary wave was made in August 1834 by a Scottish engineer named John Scott Russell (1808–1882). Scott Russell reported his observation; to the Fourteenth Meeting of the British Association for Advancement of Science, held in 1844; in a long report entitled: "Report on waves" as follows:

> "*I was observing the motion of a boat which was rapidly drawn along a narrow channel by a pair of horses, when the boat suddenly stopped – not so the mass of water in the channel which it had put in motion; it accumulated round the prow of the vessel in a state of violent agitation, then suddenly leaving it behind, rolled forward with great velocity, assuming the form of a large solitary elevation, a rounded, smooth and well-defined heap of water, which continued its course along the channel apparently without change of form or diminution of speed. I followed it on horseback, and overtook it still rolling on at a rate of some eight or nine miles an hour, preserving its original figure some thirty feet long and a foot to a foot and a half in height. Its height gradually diminished, and after a chase of one or two miles I lost it in the windings of the channel. Such, in the month of August 1834, was my first chance interview with that singular and beautiful phenomenon which I have called the Wave of Translation*".

This hump-shape localized wave that propagates along one space-direction with undeformed shape has spectacular stability properties. John Scott Russell carried out many experiments to get the properties of this wave. ► Solitons: Historical and Physical Introduction.

A model equation describing the unidirectional propagation of long waves in water of relatively shallow depth with a localized solution (representing a single hump as discovered by Russell) was obtained by Korteweg and de Vries. This equation has become very famous and is now known as the Korteweg–de Vries equation or KdV equation. ► Water Waves and the Korteweg–de Vries Equation.

There exists a certain class of nonlinear partial differential equations that leads to solitons. The Korteweg–de Vries equation, Kadomtsev–Petviashvili (KP) equation, Klein–Gordon (KG) equation, Sine–Gordon (SG) equa-

tion, nonlinear Schrodinger (NLS) equation, Korteweg–de Vries Burger's (KdVB) equation, etc… are some known equations that lie in this specific class of this NLPDE. This class of nonlinear equations has a special type of the traveling wave solutions which are either solitary waves or solitons. ► Partial Differential Equations that Lead to Solitons.

The Korteweg–de Vries equation is a canonical equation in nonlinear wave physics demonstrating the existence of solitons and their elastic interactions. This KdV equation attracted many authors to publish an enormous number of publications. Different analytical methods for solving this well-known equation are presented in ► Korteweg–de Vries Equation (KdV), Different Analytical Methods for Solving the.

One of the most famous analytical methods for solving the KdV equation is the *Inverse Scattering Transformation*. This method was first introduced by the well-known scientists Gardener, Green, Kruskal, and Miura. In 1967, Gardner et al. published a magnificent and original paper on the analytical solution to the initial value problem due to a disturbance of finite amplitude in an infinite domain. This systematic method is not easy, and needs several long steps to get the complete solution. ► Inverse Scattering Transform and the Theory of Solitons.

There are various algebraic and geometric characteristics that the KdV equation possesses. More significantly, a lot of physically important solutions to the KdV equation can be presented explicitly through a simple, specific form, called *the Hirota bilinear form*. More complicated types of solutions leading to N-solitons are presented in ► Korteweg–de Vries Equation (KdV), History, Exact *N*-Soliton Solutions and Further Properties of the.

It was suitable to search for an easier method to solve this famous and well-known nonlinear partial differential equation (KdV). Many semi-analytical methods have been developed and used to solve the KdV equation. The most popular and famous semi-analytical method that was introduced to solve this nonlinear partial differential equation (KdV), was the *Adomian Decomposition Method*. ► Adomian Decomposition Method Applied to Non-linear Evolution Equations in Soliton Theory.

Other semi-analytical methods like the variational iteration method (VIM), homotopy analysis method (HAM), and homotopy perturbation method (HPM) are usable to solve these types of NLPDE such as the Korteweg–de Vries equation and even the modified Korteweg–de Vries equation (mKdV). ► Korteweg–de Vries Equation (KdV) and Modified Korteweg–de Vries Equations (mKdV), Semi–analytical Methods for Solving the.

It is indispensable to introduce other methods to illustrate the solution of this NLPDE. The numerical technique is a powerful tool which is always needed to present and compare the results in the easiest manner. There exist different numerical tools for solving the KdV and mKdV equations, and even for any evolution equations ► Korteweg–de Vries Equation (KdV), Some Numerical Methods for Solving the.

In hydrodynamical problems, the motion of incompressible inviscid fluid subject to a constant vertical gravitational force (where the fluid is bounded below by an impermeable bottom and above by a free surface) with some physical assumptions lead to the shallow water model. The shallow water wave equations are usually used in oceanography and atmospheric science. The shallow water approximation theory leads to the KdV equation. This means that the soliton solution as well as the solitary waves are the straight forward solutions for many physical problems. ► Shallow Water Waves and Solitary Waves.

The study of the water waves in the Pacific Ocean leads directly to a special type of long gravity waves named *tsunamis*. This word is a Japanese one, which means harbor's wave.

A tsunami is essentially a long wavelength water wave train or a series of waves generated in a body of water (mostly in oceans) that vertically displaces the water column. Earthquakes, landslides, volcanic eruptions, nuclear explosions and the impact of cosmic bodies can generate tsunamis. Tsunamis, as they approach coastlines, can rise enormously and savagely attack and inundate to cause huge damage to properties and cost thousands of lives.

The theoretical study of waves in oceanography is very complicated. The simplest mathematical model includes the KdV equation. This KdV represents the governing equation that lead to tsunamis. ► Solitons, Tsunamis and Oceanographical Applications of.

The rapid change in water density with depth is called *Pycnocline*. This happens in open or closed seas, as well as in oceans. In the frame of shallow water approximation theory that governs the oceans, and due to the pycnocline the nonlinear internal gravity waves are presented. They arise from perturbations to hydrostatic equilibrium, where balance is maintained between the force of gravity and the buoyant restoring force. An illustrative study of nonlinear waves (that was generated inside a stratified fluid occupying a semi infinite channel of finite and constant depth by a wave maker situated in motion at the finite extremity of the channel) is presented in this section. ► Non-linear Internal Waves.

The soliton perturbation theory is used to study solitons that are governed by the various nonlinear equations in the presence of the perturbation terms. ► Soliton Perturbation.

A *compacton* is a special solitary traveling wave that, unlike a soliton, does not have exponential tails.

Another expression and terminology called *compacton-like soliton* is a special wave solution which can be expressed by the squares of sinusoidal or cosinusoidal functions.

Soliton and compacton are two kinds of nonlinear waves. They play an indispensable and vital role in all branches of science and technology, and are used as constructive elements to formulate the complex dynamical behavior of wave systems throughout science: from hydrodynamics to nonlinear optics, from plasmas to shock waves, from tornados to the Great Red Spot of Jupiter, and from tsunamis to turbulence. More recently, soliton and compacton have been of key importance in the quantum fields and nanotechnology especially in nano-hydrodynamics. ► Solitons and Compactons.

Solitons, Tsunamis and Oceanographical Applications of

M. Lakshmanan
Center for Nonlinear Dynamics, Bharathidasan University, Tiruchirapalli, India

Article Outline

Glossary
Definition of the Subject
Introduction
Shallow Water Waves and KdV Type Equations
Deep Water Waves and NLS Type Equations
Tsunamis as Solitons
Internal Solitons
Rossby Solitons
Bore Solitons
Future Directions
Bibliography

Glossary

Soliton A class of nonlinear dispersive wave equations in (1+1) dimensions having a delicate balance between dispersion and nonlinearity admit localized solitary waves which under interaction retain their shapes and speeds asymptotically. Such waves are called solitons because of their particle like elastic collision property. The systems include Korteweg–de Vries, nonlinear Schrödinger, sine-Gordon and other nonlinear evolution equations. Certain (2+1) dimensional generalizations of these systems also admit soliton solutions of different types (plane solitons, algebraically decaying lump solitons and exponentially decaying dromions).

Shallow and deep water waves Considering surface gravity waves in an ocean of depth h, they are called shallow-water waves if $h \ll \lambda$, where λ is the wavelength (or from a practical point of view if $h < 0.07\lambda$). In the linearized case, for shallow water waves the phase speed $c = \sqrt{gh}$, where g is the acceleration due to gravity. Water waves are classified as deep (practically) if $h > 0.28\lambda$ and the corresponding wave speed is given by $c = \sqrt{g/k}$, $k = \frac{2\pi}{\lambda}$.

Tsunami Tsunami is essentially a long wavelength water wave train, or a series of waves, generated in a body of water (mostly in oceans) that vertically displaces the water column. Earthquakes, landslides, volcanic eruptions, nuclear explosions and impact of cosmic bodies can generate tsunamis. Propagation of tsunamis is in many cases in the form of shallow water waves and sometimes can be of the form of solitary waves/solitons. Tsunamis as they approach coastlines can rise enormously and savagely attack and inundate to cause devastating damage to life and property.

Internal solitons Gravity waves can exist not only as surface waves but also as waves at the interface between two fluids of different density. While solitons were first recognized on the surface of water, the commonest ones in oceans actually happen underneath, as internal oceanic waves propagating on the pycnocline (the interface between density layers). Such waves occur in many seas around the globe, prominent among them being the Andaman and Sulu seas.

Rossby solitons Rossby waves are typical examples of quasigeostrophic dynamical response of rotating fluid systems, where long waves between layers of the atmosphere as in the case of the Great Red Spot of Jupiter or in the barotropic atmosphere are formed and may be associated with solitonic structures.

Bore solitons The classic bore (also called mascaret, poroca and aeger) arises generally in funnel shaped estuaries that amplify incoming tides, tsunamis or storm surges, the rapid rise propagating upstream against the flow of the river feeding the estuary. The profile depends on the Froude number, a dimensionless ratio of intertial and gravitational effects. Slower bores can take on oscillatory profile with a leading dispersive shock-wave followed by a train of solitons.

Definition of the Subject

Surface and internal gravity waves arising in various oceanographic conditions are natural sources where one

can identify/observe the generation, formation and propagation of solitary waves and solitons. Unlike the standard progressive waves of linear dispersive type, solitary waves are localized structures with long wavelengths and finite energies and propagate without change of speed or form and are patently nonlinear entities. The earliest scientifically recorded observation of a solitary wave was made by John Scott Russel in August 1834 in the Union Canal connecting the Scottish cities of Glasgow and Edinburgh. The theoretical formulation of the underlying phenomenon was provided by Korteweg and de Vries in 1895 who deduced the now famous Korteweg–de Vries (KdV) equation admitting solitary wave solutions. With the insightful numerical and analytical investigations of Martin Kruskal and coworkers in the 1960s the KdV solitary waves have been shown to possess the remarkable property that under collision they pass through each other without change of shape or speed except for a phase shift and so they are solitons. Since then a large class of soliton possessing nonlinear dispersive wave equations such as the sine-Gordon, modified KdV (MKdV) and nonlinear Schrödinger (NLS) equations occurring in a wide range of physical phenomena have been identified.

Several important oceanographic phenomena which correspond to nonlinear shallow water wave or deep water wave propagation have been identified/interpreted in terms of soliton propagation. These include tsunamis, especially earthquake induced ones like the 1960 Chilean or 2004 Indian Ocean earthquakes, internal solitary waves arising in stratified stable fluids such as the ones observed in Andaman or Sulu seas, Rossby waves including the Giant Red Spot of Jupiter and tidal bores occurring in estuaries of rivers. Detailed observations/laboratory experiments and theoretical formulations based on water wave equations resulting in the nonlinear evolution equations including KdV, Benjamin–Ono, Intermediate Long Wave (ILW), Kadomtsev–Petviashivili (KP), NLS, Davey–Stewartson (DS) and other equations clearly establish the relevance of soliton description in such oceanographic events.

Introduction

Historically, the remarkable observation of John Scott Russel [1,2] of the solitary wave in the Union Canal connecting the cities of Edinburgh and Glasgow in the month of August 1834 may be considered as the precursor to the realization of solitons in many oceanographic phenomena. While riding on a horse back and observing the motion of boat drawn by a pair of horses which suddenly stopped but not so the mass of water it had set in motion, the wave (which he called the 'Great Wave of Translation') in the form of large solitary heap of water surged forward and travelled a long distance without *change of form or diminution of speed*. The wave observed by Scott Russel is nothing but a solitary wave having a remarkable staying power and a patently nonlinear entity. Korteweg and de Vries in 1895, starting from the basic equations of hydrodynamics and considering unidirectional shallow water wave propagation in rectangular channels, deduced [3] the now ubiquitous KdV equation as the underlying nonlinear evolution equation. It is a third order nonlinear partial differential equation in (1+1) dimensions with a delicate balance between dispersion and nonlinearity. It admits elliptic function cnoidal wave solutions and in a limiting form exact solitary wave solution of the type observed by John Scott Russel thereby vindicating his observations and putting to rest all the controversies surrounding them.

It was the many faceted numerical and analytical study of Martin Kruskal and coworkers [4,5] which firmly established by 1967 that the KdV solitary waves have the further remarkable feature that they are solitons having elastic collision property (Fig. 1).

It was proved decisively that the KdV solitary waves on collision pass through each other except for a finite phase shift, thereby retaining their forms and speeds asymptotically as in the case of particle like elastic collisions. The inverse scattering transform (IST) formalism developed for this purpose clearly shows that the KdV equation is a completely integrable infinite dimensional nonlinear Hamiltonian system and that it admits multisoliton solutions [6,7,8,9]. Since then a large class of nonlinear dispersive wave equations such as the sine-Gordon (s-G), modified Korteweg–de Vries (MKdV), NLS, etc. equations in (1+1) dimensions modeling varied physical phenomena have also been shown to be completely integrable soliton systems [6,7,8,9]. Interesting (2+1) dimensional versions of these systems such as Kadomtsev–Petviashvile (KP), Davey–Stewartson (DS) and Nizhnik–Novikov–Veselov (NNV) equations have also been shown to be integrable systems admitting basic nonlinear excitations such as line (plane) solitons, algebraically decaying lump solitons and exponentially localized dromion solutions [7,8].

It should be noted that not all solitary waves are solitons while the converse is always true. An example of a solitary wave which is not a soliton is the one which occurs in double-well $\lambda\phi^4$ wave equation which radiates energy on collision with another such wave. However even such solitary waves having finite energies are sometimes referred to as solitons in the condensed matter, particle physics and fluid dynamics literature because of their localized structure.

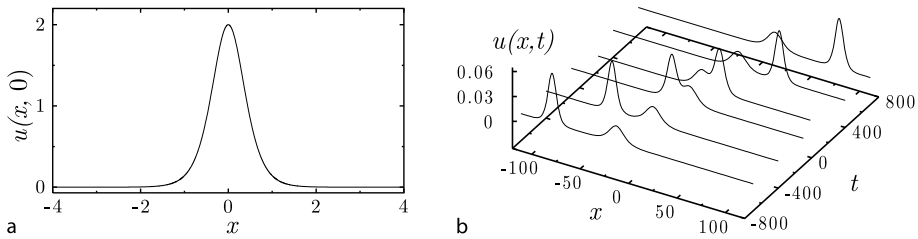

Solitons, Tsunamis and Oceanographical Applications of, Figure 1
KdV equation: a One soliton solution (solitary wave) at a fixed time, say $t = 0$, b Two soliton solution (depicting elastic collision)

As noted above solitary waves and solitons are abundant in oceanographic phenomena, especially involving shallow water wave and deep water wave propagations. However, these events are large scale phenomena very often difficult to measure experimentally, rely mostly on satellite and other indirect observations and so controversies and differing interpretations do exist in ascribing exact solitonic or solitary wave properties to these phenomena. Yet it is generally realized that many of these events are closely identifiable with solitonic structures. Some of these observations deserve special attention.

Tsunamis

When large scale earthquakes, especially of magnitude 8.0 in Richter scale and above, occur in seabeds at appropriate geological faults tsunamis are generated and can propagate over large distances as small amplitude and long wavelength structures when shallowness condition is satisfied. Though the tsunamis are hardly felt in the mid-sea, they take monstrous structures when they approach land masses due to conservation of energy. The powerful Chilean earthquake of 1960 [10] led to tsunamis which propagated for almost fifteen hours before striking Hawaii islands and a further seven hours later they struck the Japanese islands of Honshu and Hokkaido. More recently the devastating Sumatra–Andaman earthquake of 2004 in the Indian Ocean generated tsunamis which not only struck the coastlines of Asian countries including Indonesia, Thailand, India and Srilanka but propagated as far as Somalia and Kenya in Africa, killing more than a quarter million people. There is very good sense in ascribing soliton description to such tsunamis.

Internal Solitons

Peculiar striations, visible on satellite photographs of the surface of the Andaman and Sulu seas in the far east (and in many other oceans around the globe), have been interpreted as secondary phenomena accompanying the passage of "internal solitons". These are solitary wavelike distortions of the boundary layer between the warm upper layer of sea water and cold lower depths. These internal solitons are travelling ridges of warm water, extending hundreds of meters down below the thermal boundary, and carry enormous energy. Osborne and Burch [11] investigated the underwater currents which were experienced by an oil rig in the Andaman sea, which was drilling at a depth of 3600 ft. One drilling rig was apparently spun through ninety degrees and moved one hundred feet by the passage of a soliton below.

Rossby Waves and Solitons

Rossby waves [12] are long waves between layers of the atmosphere, created by the rotation of the planet. In particular, in the atmosphere of a rotating planet, a fluid particle is endowed with a certain rotation rate, determined by its latitude. Consequently its motion in the north-south direction is constrained by the conservation of angular momentum as in the case of internal waves where gravity inhibits the vertical motion of a density stratified fluid. There is an analogy between internal waves and Rossby waves under suitable conditions. The KdV equation has been proposed as a model for the evolution of Rossby solitons [13] and NLS equation for the evolution of Rossby wave packets. The Great Red Spot of the planet of Jupiter is often associated with a Rossby soliton.

Bore Solitons

Rivers which flow to the open oceans are very often affected by tidal effects, tsunamis or storm surges. Tidal motions generate intensive water flows which can propagate upstream on tens of kilometers in the form of step-wise perturbation (hydraulic jumps) analogous to shock waves in acoustics. This phenomenon is known as a bore or mascaret (in French). Examples of such bores include the tidal bore of Seine river in France, Hooghly bore of the Ganges in India, the Amazon river bore in Brazil and Hangzhou

bore in China. Bore disintegration into solitons is a possible phenomenon in such bores [14].

Besides these there are other possible oceanographic phenomena such as capillary wave solitons [15], resonant three-wave or four wave interaction solitons [16], etc., where also soliton picture is useful.

In this article we will first point out how in shallow channels the long wavelength wave propagation is described by KdV equation and its generalizations (Sect. "Shallow Water Waves and KdV Type Equations"). Then we will briefly point out how NLS family of equations arise naturally in the description of deep water waves (Sect. "Deep Water Waves and NLS Type Equations"). Based on these details, we will point how the soliton picture plays a very important role in the understanding of tsunami propagation (Sect. "Tsunamis as Solitons"), generation of internal solitons (Sect. "Internal Solitons"), formation of Rossby solitons (Sect. "Rossby Solitons") and disintegration of bores into solitons (Sect. "Bore Solitons").

Shallow Water Waves and KdV Type Equations

Kortweg and de Vries [3] considered the wave phenomenon underlying the observations of Scott Russel from first principles of fluid dynamics and deduced the KdV equation to describe the unidirectional shallow water wave propagation in one dimension.

Consider the one-dimensional (x-direction) wave motion of an incompressible and inviscid fluid (water) in a shallow channel of height h, and of sufficient width with uniform cross-section leading to the formation of a solitary wave propagating under gravity. The effect of surface tension is assumed to be negligible. Let the length of the wave be l and the maximum value of its amplitude, $\eta(x, t)$, above the horizontal surface be a (see Fig. 2).

Then assuming $a \ll h$ (shallow water) and $h \ll l$ (long waves), one can introduce two natural small parameters into the problem $\epsilon = a/h$ and $\delta = h/l$. Then the analysis proceeds as follows [3,6].

Equation of Motion: KdV Equation

The fluid motion can be described by the velocity vector $\boldsymbol{V}(x, y, t) = u(x, y, t)\boldsymbol{i} + v(x, y, t)\boldsymbol{j}$, where $\boldsymbol{i}$ and $\boldsymbol{j}$ are the unit vectors along the horizontal and vertical directions, respectively. As the motion is irrotational, we have $\nabla \times \boldsymbol{V} = 0$. Consequently, we can introduce the velocity potential $\phi(x, y, t)$ by the relation $\boldsymbol{V} = \nabla\phi$.

Conservation of Density The system obviously admits the conservation law for the mass density $\rho(x, y, t)$ of the fluid, $\mathrm{d}\rho/\mathrm{d}t = \rho_t + \nabla \cdot (\rho \boldsymbol{V}) = 0$. As ρ is a constant, we have $\nabla \cdot \boldsymbol{V} = 0$. Consequently ϕ obeys the Laplace equation

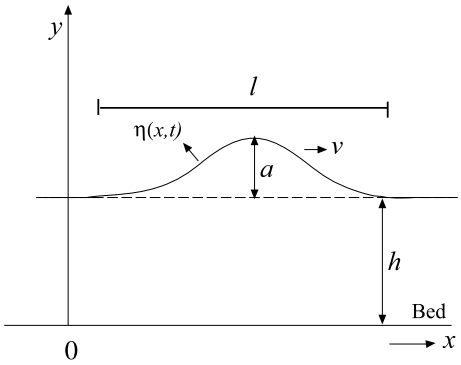

Solitons, Tsunamis and Oceanographical Applications of, Figure 2
One-dimensional wave motion in a shallow channel

$$\nabla^2\phi(x, y, t) = 0 \,. \tag{1}$$

Euler's Equation As the density of the fluid $\rho = \rho_0 =$ constant, using Newton's law for the rate of change of momentum, we can write $\mathrm{d}\boldsymbol{V}/\mathrm{d}t = \partial\boldsymbol{V}/\partial t + (\boldsymbol{V}\cdot\nabla)\boldsymbol{V} = -\frac{1}{\rho_0}\nabla p - g\boldsymbol{j}$, where $p = p(x, y, t)$ is the pressure at the point (x, y) and g is the acceleration due to gravity, which is acting vertically downwards (here $\boldsymbol{j}$ is the unit vector along the vertical direction). Since $\boldsymbol{V} = \nabla\phi$ we obtain (after one integration)

$$\phi_t + \frac{1}{2}(\nabla\phi)^2 + \frac{p}{\rho_0} + gy = 0 \,. \tag{2}$$

Boundary Conditions The above two Eqs. (1) and (2) for the velocity potential $\phi(x, y, t)$ of the fluid have to be supplemented by appropriate boundary conditions, by taking into account the fact (see Fig. 2) that (a) the horizontal bed at $y = 0$ is hard and (b) the upper boundary $y = y(x, t)$ is a free surface. As a result
(a) the vertical velocity at $y = 0$ vanishes, $v(x, 0, t) = 0$, which implies

$$\phi_y(x, 0, t) = 0 \,. \tag{3}$$

(b) As the upper boundary is free, let us specify it by $y = h+\eta(x, t)$ (see Fig. 2). Then at the point $x = x_1$, $y = y_1 \equiv y(x, t)$, we can write $\frac{\mathrm{d}y_1}{\mathrm{d}t} = \frac{\partial\eta}{\partial t} + \frac{\partial\eta}{\partial x}\cdot\frac{\mathrm{d}x_1}{\mathrm{d}t} = \eta_t + \eta_x u_1 = v_1$. Since $v_1 = \phi_{1y}$, $u_1 = \phi_{1x}$, we obtain

$$\phi_{1y} = \eta_t + \eta_x\phi_{1x} \,. \tag{4}$$

(c) Similarly at $y = y_1$, the pressure $p_1 = 0$. Then from (2), it follows that

$$u_{1t} + u_1 u_{1x} + v_1 v_{1x} + g\eta_x = 0 \,. \tag{5}$$

Thus the motion of the surface of water wave is essentially specified by the Laplace Eq. (1) and Euler Eq. (2) along with one fixed boundary condition (3) and two *variable nonlinear* boundary conditions (4) and (5). One has to then solve the Laplace equation subject to these boundary conditions.

Taylor Expansion of $\phi(x, y, t)$ in y Making use of the fact $\delta = h/l \ll 1$, $h \ll l$, we assume $y(= h + \eta(x, t))$ to be small to introduce the Taylor expansion

$$\phi(x, y, t) = \sum_{n=0}^{\infty} y^n \phi_n(x, t) \,. \tag{6}$$

Substituting the above series for ϕ into the Laplace Eq. (1), solving recursively for $\phi_n(x, t)$'s and making use of the boundary condition (4), $\phi_y(x, 0, t) = 0$, one can show that

$$u_1 = \phi_{1x} = f - \frac{1}{2} y_1^2 f_{xx} + \text{ higher order in } y_1 \,, \tag{7}$$

$$v_1 = \phi_{1y} = -y_1 f_x + \frac{1}{6} y_1^3 f_{xxx} + \text{ higher order in } y_1 \,, \tag{8}$$

where $f = \partial\phi_0/\partial x$. We can then substitute these expressions into the nonlinear boundary conditions (4) and (5) to obtain equations for f and η.

Introduction of Small Parameters ϵ and δ So far the analysis has not taken into account fully the shallow nature of the channel ($a/h = \epsilon \ll 1$) and the solitary nature of the wave ($a/l = a/h \cdot h/l = \epsilon\delta \ll 1$, $\epsilon \ll 1$, $\delta \ll 1$), which are essential to realize the Scott Russel phenomenon. For this purpose one can stretch the independent and dependent variables in the above equations through appropriate scale changes, but retaining the overall form of the equations. To realize this one can introduce the natural scale changes

$$x = lx' \,, \quad \eta = a\eta' \,, \quad t = \frac{l}{c_0} t' \,, \tag{9}$$

where c_0 is a parameter to be determined. Then in order to retain the form of (7) and (8) we require

$$\begin{aligned} &u_1 = \epsilon c_0 u_1' \,, \quad v_1 = \epsilon\delta c_0 v_1' \,, \quad f = \epsilon c_0 f' \,, \\ &y_1 = h + \eta(x, t) = h\left(1 + \epsilon\eta'\left(x', t'\right)\right) \,. \end{aligned} \tag{10}$$

Then

$$u_1' = f' - \frac{1}{2}\delta^2 \left(1 + \epsilon\eta'\right)^2 f'_{x'x'} = f' - \frac{1}{2}\delta^2 f'_{x'x'} \,, \tag{11}$$

where we have omitted terms proportional to $\delta^2\epsilon$ as small compared to terms of the order δ^2. Similarly from (8), we obtain

$$v_1' = -\left(1 + \epsilon\eta'\right) f'_{x'} + \frac{1}{6}\delta^2 f'_{x'x'x'} \,. \tag{12}$$

Now considering the nonlinear boundary condition (4) in the form $v_1 = \eta_t + \eta_x u_1$, it can be rewritten as

$$\eta'_{t'} + f'_{x'} + \epsilon\eta' f'_{x'} + \epsilon f' \eta'_{x'} - \frac{1}{6}\delta^2 f'_{x'x'x'} = 0 \,. \tag{13}$$

Similarly considering the other boundary condition (5) and making use of the above transformations, it can be rewritten, after neglecting terms of the order $\epsilon^2\delta^2$, as

$$f'_{t'} + \epsilon f' f'_{x'} + \frac{ga}{\epsilon c_0^2}\eta'_{x'} - \frac{1}{2}\delta^2 f'_{x'x't'} = 0 \,. \tag{14}$$

Now choosing the arbitrary parameter c_0 as $c_0^2 = gh$ so that $\eta'_{x'}$ term is of order unity, (14) becomes

$$f'_{t'} + \eta'_{x'} + \epsilon f' f'_{x'} - \frac{1}{2}\delta^2 f'_{x'x't'} = 0 \,. \tag{15}$$

(Note that $c_0 = \sqrt{gh}$ is nothing but the speed of the water wave in the linearized limit). Omitting the primes for convenience, the evolution equation for the amplitude of the wave and the function related to the velocity potential reads

$$\eta_t + f_x + \epsilon\eta f_x + \epsilon f \eta_x - \frac{1}{6}\delta^2 f_{xxx} = 0 \,, \tag{16}$$

$$f_t + \eta_x + \epsilon f f_x - \frac{1}{2}\delta^2 f_{xxt} = 0 \,. \tag{17}$$

Note that the small parameters ϵ and δ^2 have occurred in a natural way in (16), (17).

Perturbation Analysis Since the parameters ϵ and δ^2 are small in (16), (17), we can make a perturbation expansion of f in these parameters:

$$f = f^{(0)} + \epsilon f^{(1)} + \delta^2 f^{(2)} + \text{ higher order terms} \,, \tag{18}$$

where $f^{(i)}$, $i = 0, 1, 2, \ldots$ are functions of η and its spatial derivatives. Note that the above perturbation expansion is an asymptotic expansion. Substituting this into Eqs. (16) and (17) and regrouping and comparing different powers proportional to (ϵ, δ^2) and solving them successively one

can obtain (see for example [6] for further details) in a self consistent way,

$$f^{(0)} = \eta, \quad f^{(1)} = -\frac{1}{4}\eta^2 , \quad f^{(2)} = \frac{1}{3}\eta_{xx} . \tag{19}$$

Using these expressions into (18) and substituting it in (16) and (17), we ultimately obtain the KdV equation in the form

$$\eta_t + \eta_x + \frac{3}{2}\epsilon\eta\eta_x + \frac{\delta^2}{6}\eta_{xxx} = 0 , \tag{20}$$

describing the unidirectional propagation of long wavelength shallow water waves.

The Standard (Contemporary) Form of KdV Equation Finally, changing to a moving frame of reference, $\xi = x - t$, $\tau = t$, and introducing the new variables $u = (3\epsilon/2\delta^2)\eta$, $\tau' = (6/\delta^2)\tau$ and redefining the variables τ' as t and ξ as x, we finally arrive at the ubiquitous form of the KdV equation as

$$u_t + 6uu_x + u_{xxx} = 0 . \tag{21}$$

The Korteweg–de Vries Eq. (21) admits cnoidal wave solution and in the limiting case solitary wave solution as well. This form can be easily obtained [6,17] by looking for a wave solution of the form $u = 2f(\xi)$, $\xi = (x - ct)$, and reducing the KdV equation into a third order nonlinear ordinary differential equation of the form

$$-c\frac{\partial f}{\partial \xi} + 12f\frac{\partial f}{\partial \xi} + \frac{\partial^3 f}{\partial \xi^3} = 0 . \tag{22}$$

Integrating Eq. (22) twice and rearranging, we can obtain

$$\left(\frac{\partial f}{\partial \xi}\right)^2 = -4f^3 + cf^2 - 2df - 2b \equiv P(f) , \tag{23}$$

where b and d are integration constants. Calling the three real roots of the cubic equation $P(f) = 0$ as α_1, α_2 and α_3 such that

$$\left(\frac{\partial f}{\partial \xi}\right)^2 = -4(f - \alpha_1)(f - \alpha_2)(f - \alpha_3), \tag{24}$$

the solution can be expressed in terms of the Jacobian elliptic function as

$$f(\xi) = f(x - ct) = \alpha_3 - (\alpha_3 - \alpha_2)\mathrm{sn}^2 \cdot \left[\sqrt{\alpha_3 - \alpha_1}\,(x - ct)\,, m\right] , \tag{25a}$$

where

$$(\alpha_1 + \alpha_2 + \alpha_3) = \frac{c}{4}, \quad m^2 = \frac{\alpha_3 - \alpha_2}{\alpha_3 - \alpha_1}, \ \delta = \text{constant}. \tag{25b}$$

Here m is the modulus parameter of the Jacobian elliptic function. Eq. (25a) represents in fact the so called cnoidal wave (because of its elliptic function form). In the limiting case $m = 1$, the form (25a) reduces to

$$f = \alpha_2 + (\alpha_3 - \alpha_2)\ \mathrm{sech}^2\left[\sqrt{\alpha_3 - \alpha_1}\,(x - ct)\right] . \tag{26}$$

Choosing now $\alpha_1 = 0$, $\alpha_2 = 0$, and using (25b) we have

$$f = \frac{c}{4}\,\mathrm{sech}^2\left[\frac{\sqrt{c}}{2}\,(x - ct)\right] . \tag{27}$$

Then the solitary wave solution to the KdV Eq. (21) can be written in the form

$$u(x, t) = \frac{c}{2}\mathrm{sech}^2\frac{\sqrt{c}}{2}(x - ct + \delta), \quad \delta : \text{constant} \tag{28}$$

Note that the velocity of the solitary wave is directly proportional to the amplitude: larger the wave the higher is the speed. More importantly, the KdV solitary wave is a soliton: it retains its shape and speed upon collision with another solitary wave of different amplitude, except for a phase shift, see Fig. 1 [6,7,8]. In fact for an arbitrary initial condition, the solution of the Cauchy initial value problem consists of N-number of solitons of different amplitudes in the background of small amplitude dispersive waves. All these results ultimately lead to the result that the KdV equation is a completely integrable, infinite dimensional, nonlinear Hamiltonian system. It possesses [6,7,8]

(i) a Lax pair of linear differential operators and is solvable through the so called inverse scattering transform (IST) method,
(ii) infinite number of conservation laws and associated infinite number of involutive integrals of motion,
(iii) N-soliton solution,
(iv) Hirota bilinear form,
(v) Hamiltonian structure

and a host of other interesting properties (see for example [6,7,8,9]).

KdV Related Integrable and Nonintegrable NLEEs

Depending on the actual physical situation, the derivation of the shallow water wave equation can be suitably modified to obtain other forms of nonlinear dispersive wave

equations in (1+1) dimensions as well as in (2+1) dimensions relevant for the present context. Without going into the actual derivations, some of the important equations possessing solitary waves are listed below [6,7,8,9].

1. Boussinesq equation [7]

$$u_t + uu_x + g\eta_x - \frac{1}{3}h^2 u_{txx} = 0, \tag{29a}$$

$$\eta_t + [u(h+\eta)]_x = 0 \tag{29b}$$

2. Benjamin–Bona–Mahoney (BBM) equation [18]

$$u_t + u_x + uu_x - u_{xxt} = 0 \tag{30}$$

3. Camassa–Holm equation [19]

$$u_t + 2\kappa u_x + 3uu_x - u_{xxt} = 3u_x u_{xx} + uu_{xxx} \tag{31}$$

4. Kadomtsev–Petviashville (KP) equation [7]

$$(u_t + 6uu_x + u_{xxx})_x + 3\sigma^2 u_{yy} = 0 \tag{32}$$

($\sigma^2 = -1$: KP-I, $\sigma^2 = +1$: KP-II).

In the derivation of the above equations, generally the bottom of water column or fluid bed is assumed to be flat. However in realistic situations the water depth varies as a function of the horizontal coordinates. In this situation, one often encounters inhomogeneous forms of the above wave equations. Typical example is the variable coefficient KdV equation [14]:

$$u_t + f(x,t)uu_x + g(x,t)u_{xxx} = 0, \tag{33}$$

where f and g are functions of x, t. More general forms can also be deduced depending upon the actual situations, see for example [14].

Deep Water Waves and NLS Type Equations

Deep water waves are strongly dispersive in contrast to the weakly dispersive nature of the shallow water waves (in the linear limit). Various authors (see for details [8]) have shown that nonlinear Schrödinger family of equations models the evolution of a packet of surface waves in this case. There are several oceanographic situations where such waves can arise [8]:

(i) A localized storm at sea can generate a wide spectrum of waves, which then propagates away from the source region in all horizontal directions. If the propagating waves have small amplitudes and encounter no wind away from the source region, these waves can eventually sort themselves into nearly one-dimensional packets of nearly monochromatic waves. For appropriately chosen scales, the underlying evolution of each of these packets can be shown to satisfy the nonlinear Schrödinger equation and its generalizations in (2+1) dimensions.

(ii) Nearly monochromatic, nearly one-dimensional waves can cover a broad range of surface waves in the sea that results due to a steady wind of long duration and fetch. Then the generalization of the NLS equation in (2+1) dimensions can describe the waves that result in when the wind stops.

In all the above situations one looks for the solution of the equations of motion (1)-(5) but generalized in three dimensions which consists mainly in the form of a small amplitude, nearly monochromatic, nearly one-dimensional wave train. Assuming that this wave train travels in the x-direction with a mean wave number $\boldsymbol{\kappa} = (k, l)$ with a characteristic amplitude 'a' of the disturbance and a characteristic variation δk in k, one can deduce the NLS equation in (2+1) dimensions under the following conditions:

(i) small amplitudes such that $\hat{\epsilon} = \epsilon\delta \equiv \kappa a \ll 1$
(ii) slowly varying modulations, $\frac{\delta k}{\kappa} \ll 1$
(iii) nearly one dimensional waves, $\frac{|l|}{k} \ll 1$
(iv) balance of all three effects, $\frac{\delta k}{\kappa} = \frac{|l|}{k} \approx O(\hat{\epsilon})$
(v) for finite and deep water waves, $(kh)^2 \gg \hat{\epsilon}$

In the lowest order approximation (linear approximation) of water waves, the prediction is that a localized initial state will generally evolve into wave packets with a dominant wave number $\boldsymbol{\kappa}$ and corresponding frequence ω, given by the dispersion relation $\omega = (g\kappa + \sigma\kappa^3)\tanh\kappa h, \kappa = |\boldsymbol{\kappa}| = \sqrt{k^2 + l^2}$ within which each wave propagates with the phase speed $c = \omega/k$, while the envelope propagates with the group velocity $c_g = \mathrm{d}\omega/\mathrm{d}\kappa$. After a sufficiently long time the wave packet tends to disperse around the dominant wave number.

This tendency for dispersion can be offset by cumulative nonlinear effects. In the absence of surface tension, the outcome for unidirectional waves can be shown to be describable by the NLS equation. If the surface wave in the lowest order is $\phi \approx A\exp i(kx - \omega t)$+c.c, where ϕ is the velocity potential, then to leading order the wave amplitude evolves as

$$i(A_t + c_g A_x) + \frac{1}{2}\lambda A_{xx} + \mu|A|^2 A = 0. \tag{34}$$

The coefficients here are given by

$$\lambda = \frac{\partial^2 \omega}{\partial k^2}\,,$$
$$\mu = -\frac{\omega k^2}{16S^4}(8C^2S^2 + g - 2T^2) + \frac{\omega}{8C^2S^2}\frac{(2\omega C^2 + kc_g)^2}{(gh - c_g^2)}, \tag{35}$$

where $C = \cosh(kh)$, $S = \sinh(kh)$, $T = \frac{S}{C}$. Equation (34) has been obtained originally by Zakharov in 1968 for deep water waves [20] and by Hasimoto and Ono for waves of finite depth in 1972 [21].

The NLS equation is also a soliton possessing integrable system and is solvable by the IST method [6,7,8,9]. For $\lambda > 0$ (focusing case), the envelope solitary (soliton) wave solution (also called bright solitons in the optical physics context) is given by

$$A(x,t) = a\ \mathrm{sech}\gamma(x - c_g t)\exp(-i\Omega t), \tag{36}$$

where $\mu a^2 = \lambda\gamma^2$, $\Omega = +\frac{1}{2}\mu a^2$.

When the effects of modulation in the transverse y-direction are taken into account, so that the wave amplitude is now given by $A(x, y, t)$, the NLS equation is replaced by the Benney–Roskes system [22] also popularly known as the Davey–Stewartson equations [23],

$$i(A_t + c_g A_x) + \frac{1}{2}\lambda A_{xx} + \frac{1}{2}\delta A_{yy} + \mu|A|^2 A + UA = 0\,, \tag{37a}$$

$$\alpha U_{xx} + U_{yy} + \beta(|A|^2)_{yy} = 0\,, \tag{37b}$$

where $\delta = \frac{c_g}{k}$, $\alpha = 1 - \left(\frac{c_g^2}{gh}\right)$, $gh\beta = \frac{\omega}{8C^2S^2}(2\omega C^2 + kc_g)^2$. Here $U(x, y, t)$ is the wave induced mean flow. In the deep water wave limit, $kh \to \infty$, and $U \to 0$, $\beta \to 0$ and one has the nonintegrable (2+1) dimensional NLS equation. On the other hand in the shallow water limit, one has the integrable Davey–Stewartson (DS) equations. For details see [7,8]. The DS-I equation admits algebraically decaying lump solitons and exponentially decaying dromions [24] besides the standard line solitons for appropriate choices of parameters. A typical dromion solution of DS-I equation is shown in Fig. 3.

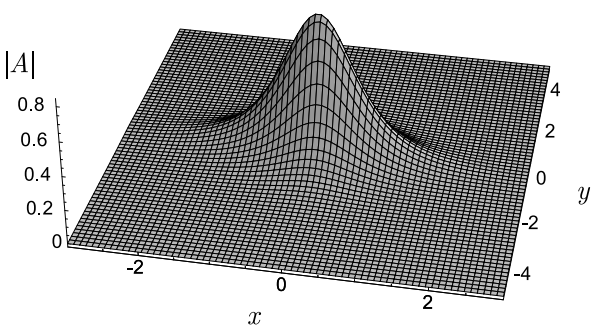

Solitons, Tsunamis and Oceanographical Applications of, Figure 3
Exponentially localized dromion solution of the Davey–Stewartson equation at a fixed time ($t = 0$) for suitable choice of parameters

Tsunamis as Solitons

The term 'tsunami' (tsu:harbour, nami:wave in Japanese) which was perhaps an unknown word even for scientists in countries such as India, Srilanka, Thailand, etc. till recently has become a house-hold word since that fateful morning of December 26, 2004. When a powerful earthquake of magnitude 9.1–9.3 on the Richter scale, epicentered off the coast of Sumatra, Indonesia, struck at 07:58:53, local time, described as the 2004 Indian Ocean earthquake or Sumatra–Andaman earthquake (Fig. 4), it triggered a series of devastating tsunamis as high as 30 meters that spread throughout the Indian Ocean killing about 275,000 people and inundating coastal communities across South and Southeast Asia, including parts of Indonesia, Srilanka, India and Thailand and even reaching as far as the east coast of Africa [25]. The catastrophe is considered to be one of the deadliest disasters in modern history.

Since this earthquake and consequent tsunamis, several other earthquakes of smaller and larger magnitudes keep occurring off the coast of Indonesia. Even as late as July 17, 2006 an earthquake of magnitude 7.7 on the Richter scale struck off the town of Pandering at 15.19 lo-

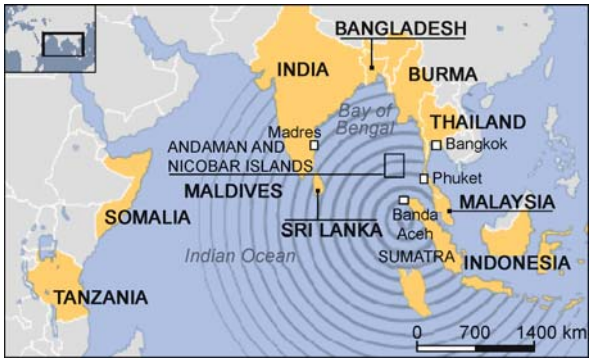

Solitons, Tsunamis and Oceanographical Applications of, Figure 4
26 December 2004 Indian Ocean tsunami (adapted from the website www.blogaid.org.uk with the courtesy of Andy Budd)

cal time and set off a tsunami of 2m high which had killed more than 300 people.

These tsunamis, which can become monstrous tidal waves when they approach coastline, are essentially triggered due to the sudden vertical rise of the seabed by several meters (when earthquake occurs) which displaces massive volume of water. The tsunamis behave very differently in deep water than in shallow water as pointed out below. By no means the tsunami of 2004 and later ones are exceptional; More than two hundred tsunamis have been recorded in scientific literature since ancient times. The most notable earlier one is the tsunami triggered by the powerful earthquake (9.6 magnitude) off southern Chile on May 22, 1960 [10] which traveled almost 22 hours before striking Japanese islands.

It is clear from the above events that the tsunami waves are fairly permanent and powerful ones, having the capacity to travel extraordinary distances without practically diminishing in size or speed. In this sense they seem to have considerable resemblance to shallow water nonlinear dispersive waves of KdV type, particularly solitary waves and solitons. It is then conceivable that tsunami dynamics has close connection with soliton dynamics.

Basics of Tsunami Waves

As noted above tsunami waves of the type described above are essentially triggered by massive earthquakes which lead to vertical displacement of a large volume of water. Other possible reasons also exist for the formation and propagation of tsunami waves: underwater nuclear explosion, larger meteorites falling into the sea, volcano explosions, rockslides, etc. But the most predominant cause of tsunamis appear to be large earthquakes as in the case of the Sumatra–Andaman earthquake of 2004. Then there are three major aspects associated with the tsunami dynamics [26]:

1. Generation of tsunamis
2. Propagation of tsunamis
3. Tsunami run up and inundation

There exist rather successful models to approach the generation aspects of tsunamis when they occur due to the earthquakes [27]. Using the available seismic data it is possible to reconstruct the permanent deformation of the sea bottom due to earthquakes and simple models have been developed (see for example, [28]). Similarly the tsunami run up and inundation problems [29] are extremely complex and they require detailed critical study from a practical point of view in order to save structures and lives when a tsunami strikes.

However, here we will be more concerned with the propagation of tsunami waves and their possible relation to wave propagation associated with nonlinear dispersive waves in shallow waters. In order to appreciate such a possible connection, we first look at the typical characteristic properties of tsunami waves as in the case of 2004 Indian Ocean tsunami waves or 1960 Chilean tsunamis.

The Indian Ocean Tsunami of 2004

Considering the Indian Ocean 2004 tsunami, satellite observations after a couple of hours after the earthquake establish an amplitude of approximately 60 cms in the open ocean for the waves. The estimated typical wavelength is about 200 kms [30]. The maximum water depth h is between 1 and 4 kms. Consequently, one can identify in an average sense the following small parameters (ϵ and δ^2) of roughly equal magnitude:

$$\epsilon = \frac{a}{h} \approx 10^{-4} \ll 1, \quad \delta^2 = \frac{h^2}{l^2} \approx 10^{-4} \ll 1 \tag{38}$$

As a consequence, it is possible that a nonlinear shallow water wave theory where dispersion (KdV equation) also plays an important role (as discussed in Sect. "Shallow Water Waves and KdV Type Equations") has considerable relevance [26]. However, we also wish to point out here that there are other points of view: Constantin and Johnson [31] estimate $\epsilon \approx 0.002$ and $\delta \approx 0.04$ and conclude that for both nonlinearity and dispersion to become significant the quantity $\delta\epsilon^{-3/2}\times$wavelength estimated as 90,000 kms is too large and shallow water equations with variable depth (without dispersion) should be used. However, it appears that these estimates can vary over a rather wide range and with suitable estimates it is possible that the range of 10,000–20,000 kms could be also possible ranges and hence taking into account the fact that both the Indian Ocean 2004 and Chilean 1960 tsunamis have traveled over 10 hours or more (in certain directions) before encountering land mass appears to allow for the possibility of nonlinear dispersive waves as relevant features for the phenomena. Segur [32] has argued that in the 2004 tsunamis, the propagation distances from the epicenter of the earthquake to India, Srilanka, or Thailand were too short for KdV dynamics to develop. In the same way one can argue that the waves that hit Somalia and Kenya in the east coast of Africa (or as in the case of Chilean earthquake see also [32]) have traveled sufficiently long distance for KdV dynamics to become important [33]. In any case one can conclude that at least for long distance tsunami propagation solitary wave and soliton picture of KdV like equations become very relevant.

Internal Solitons

For a long time seafarers passing through the Strait of Malacca on their journeys between India and the Far East have noticed that in the Andaman sea, between Nicobar islands and the north east coast of Sumatra, often bands of strongly increased surface roughness (ripplings or bands of choppy water) occur [11,34]. Similar observations have been reported in other seas around the globe from time to time. In recent times there has been considerable progress in understanding these kind of observations in terms of internal solitons in the oceans [7]. These studies have been greatly facilitated by photographs taken from satellites and space-crafts orbiting the earth, for example by synthetic aperture radar (SAR) images of ERS-1/2 satellites [34,35].

Peculiar striations of 100 km long, separated by 6 to 15 km and grouped in packets of 4 to 8, visible on satellite photographs (see Fig. 5) of the surface of the Andaman and Sulu seas in the Far East, have been interpreted as secondary phenomena accompanying the passage of 'internal solitons', which are solitary wavelike distortions of the boundary layer between warm upper layer of sea water and cold lower depths. These internal solitons are traveling edges of warm water, extending hundreds of meters down below the thermal boundary [7]. They carry enormous energy with them which is perhaps the reason for unusually strong underwater currents experienced by deep-sea drilling rigs. Thus these internal solitons are potentially hazardous to sub-sea oil and gas explorations. The ability to predict them can improve substantially the cost effectiveness and safety of offshore drilling.

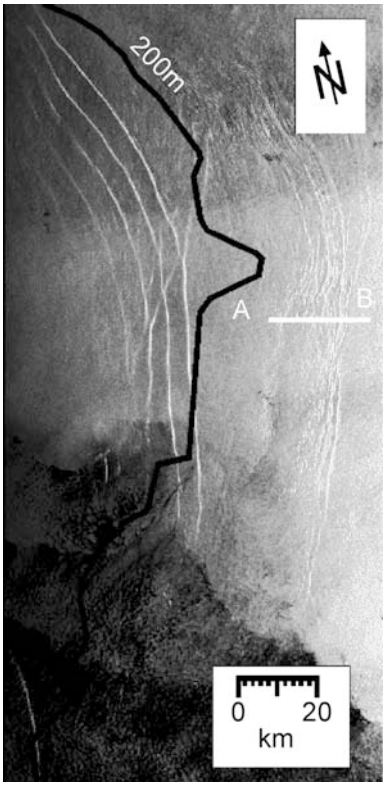

Solitons, Tsunamis and Oceanographical Applications of, Figure 5

SAR image of a 200 km × 200 km large section of the Andaman Sea acquired by the ERS-2 satellite on April 15, 1996 showing sea surface manifestations of two internal solitary wave packets, see [34]. Figure reproduced from ESA website www.earth.esa.int/workshops/ers97/papers/alpers3 with the courtesy of European Space Agency and W. Alpers

A systematic study of the underwater currents experienced by an oil rig in the Andaman sea which was drilling at a depth of 3600 ft was carried out by Osborne and Burch in 1980 [11]. They spent four days measuring underwater currents and temperatures. The striations seen on satellite photographs turned out to be kilometer-wide bands of extremely choppy water, stretching from horizon to horizon, followed by about two kilometers of water "as smooth as a millpond". These bands of agitated water are called "tide rips", they arose in packets of 4 to 8, spaced about 5 to 10 km apart (they reached the research vessel at approximately hourly intervals) and this pattern was repeated with the regularity of tidal phenomenon.

As described in [7], Osborne and Burch found that the amplitude of each succeeding soliton was less than the previous one, is precisely what is expected for solitons (note that the velocity of a solitary wave solution of KdV equation increases with amplitude, vide Eq. (22)). Thus if a number of solitons are generated together, then we expect them eventually to be arranged in an ordered sequence of decreasing amplitude. From the spacing between successive waves in a packet and the rate of separation calculated from the KdV equation, Osborne and Burch were able to estimate the distance the packet had traveled from its source and thus identify possible source regions [7]. They concluded that the solitons are generated by tidal currents off northern Sumatra or between the islands of the Nicobar chain that extends beyond it and that their observations have good general agreement with the predictions for internal solitons as given by the KdV equation. Numerous recent observations and predictions of solitons in the Andaman sea have clearly established that it is a site where extraordinarily large internal solitons are encountered [34,35].

Further, Apel and Holbrook [36] undertook a detailed study of internal waves in the Sulu sea. Satellite photographs had suggested that the source of these waves was near the southern end of the Sulu sea and their research

ship followed one wave packet for more than 250 miles over a period of two days – an extraordinary coherent phenomenon [7]. These internal solitons travel at speeds of about 8 kilometers per hour (5 miles per hour), with amplitude of about 100 meters and wavelength of about 1700 meters.

Similar observations elsewhere have confirmed the presence of internal solitons in oceans including the strait of Messina, the strait of Gibraltar, off the western side of Baja California, the Gulf of California, the Archipelago of La Maddalena and the Georgia strait [7]. There has also been a number of experimental studies of internal solitons in laboratory tanks in the last few decades [37]. These experiments provide detailed quantitative information usually unavailable in the field conditions, and are also an efficient tool for verifying various theoretical models.

As a theoretical formulation of internal solitons [7], consider two incompressible, immiscible fluids, with densities ρ_1 and ρ_2 and depths h_1 and h_2 respectively such that the total depth $h = h_1 + h_2$. Let the lighter fluid of height h_1 be lying over a heavier fluid of height h_2, in a constant gravitational field (Fig. 6). The lower fluid is assumed to rest on a horizontal impermeable bed, and the upper fluid is bounded by a free surface.

Then as in Sect. "Shallow Water Waves and KdV Type Equations", we denote the characteristic amplitude of wave by 'a' and the characteristic wavelength $l = k^{-1}$. Then the various nonlinear wave equations to describe the formation of internal solitons follow by suitable modification of the formulation in Sect. "Shallow Water Waves and KdV Type Equations", assuming viscous effects to be negligible. Each of these equations is completely integrable and admits soliton solutions [7].

(a) KdV equation (Eq. (21)) follows when

(i) the waves are of long wavelength $\delta = \frac{h}{l} \ll 1$,
(ii) the amplitude of the waves are small, $\epsilon = \frac{a}{h} \ll 1$, and
(iii) the two effects are comparable $\delta^2 = O(\epsilon)$

(b) Intermediate-Long-Wave (ILW) equation [38]

$$u_t + u_x + 2uu_x + Tu_{xx} = 0, \tag{39}$$

where Tu is the singular integral operator

$$(Tf)(x) = \frac{1}{2l} \fint_{-\infty}^{\infty} \coth\left\{\frac{\pi}{2l}(y-x)\right\} f(y)\mathrm{d}y \tag{40}$$

with $\fint_{-\infty}^{\infty}$ the Cauchy principal value integral is obtained under the assumption that

(a) there is a thin (upper) layer, $\epsilon = \frac{h_1}{h_2} \ll 1$,
(b) the amplitude of the waves is small, $a \ll h_1$,
(c) the above two effects balance, $\frac{a}{h_1} = O(\epsilon)$,
(d) the characteristic wavelength is comparable to the total depth of the fluid, $l = kh = O(1)$ and
(e) the waves are long waves in comparison with the thin layer, $kh_1 \ll 1$.

(c) Benjamin–Ono equation [39]

$$u_t + 2uu_x + Hu_{xx} = 0, \tag{41}$$

where Hu is the Hilbert transform

$$(Hf)(x) = \frac{1}{\pi} \fint_{-\infty}^{\infty} \frac{f(y)}{y-x}\mathrm{d}y, \tag{42}$$

is obtained under the assumption

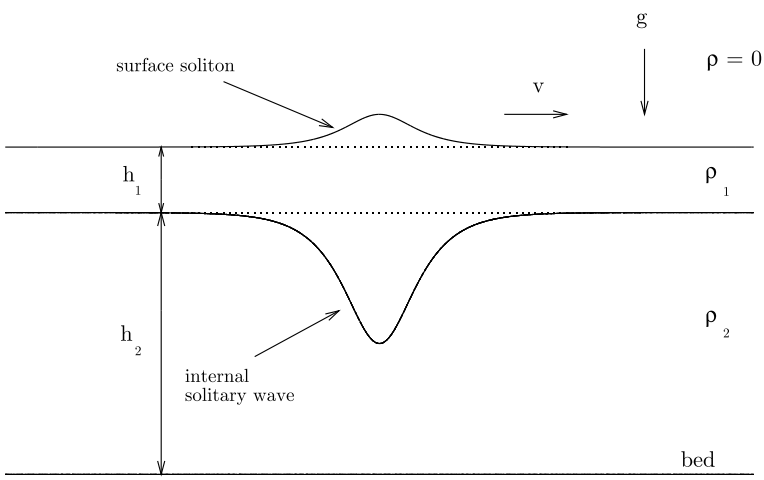

Solitons, Tsunamis and Oceanographical Applications of, Figure 6
Formation of internal soliton (note that under suitable conditions small amplitude surface soliton can also be formed)

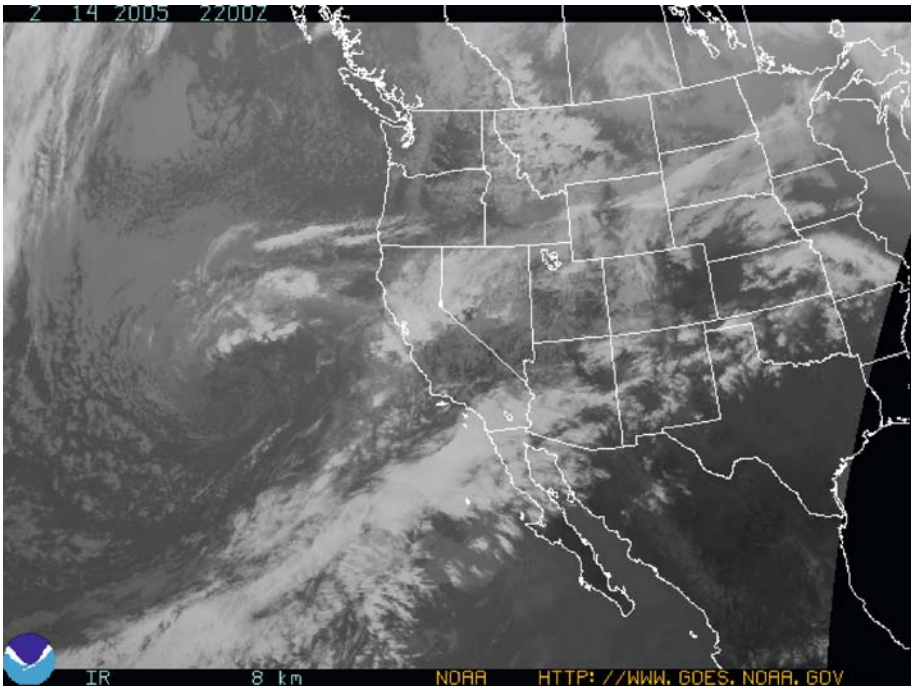

Solitons, Tsunamis and Oceanographical Applications of, Figure 7
A coupled high/low pressure systems in the form of a Rossby soliton formed off the coast of California/Washington on Valentine's Day 2005. The low pressure front hovers over Los Angeles dumping 30 inches of rain on the city in two weeks. The high pressure system lingers off the coast of Washington state providing unseasonally warm and sunny weather. This particular Rossby soliton proved exceptionally stable because of its remarkable symmetry. In an accompanying animated gif in this website one can watch the very interesting phenomenon as the jet stream splits around the soliton suctioning warm wet air from the off the coast of Mexico to Arizona, leaving behind a welcome drenching of rain. The figure and caption have been adapted from the website http://mathpost.la.asu.edu/~rubio/rossby_soliton/rs.html with the courtesy of the National Oceanic and Atmospheric Administration (NOAA) and Antonio Rubio

(a) there is a thin (upper) layer $h_1 \ll h_2$,
(b) the waves are long waves in comparison with the thin layer, $kh_1 \ll 1$,
(c) the waves are short in comparison with the total depth of the fluid, $kh \gg 1$ and
(d) the amplitude of the waves is small, $a \ll h_1$.

It may be noted that in the shallow water limit, as $\delta \to 0$, the ILW equation reduces to the KdV equation, while the Benjamin–Ono equation reduces to it in the deep water wave limit as $\delta \to \infty$. Each of these equations have their own ranges of validity and admit solitary wave and soliton solutions to represent internal solitons of the oceans.

Rossby Solitons

The atmospheric dynamics is an important subject of human concern as we live within the atmosphere and are continuously affected by the weather and its rather complex behavior. The motion of the atmosphere is intimately connected with that of the ocean with which it exchanges fluxes of momentum, heat and moisture [40]. Then the atmospheric dynamics is dominated by the rotation of the earth and the vertical density stratification of the surrounding medium, leading to newer effects. Similar effects are also present in other planetary dynamics as well.

In the atmosphere of a rotating planet, a fluid particle is endowed with a certain rotation rate (Coriolis frequency), determined by its latitude. Consequently its motion in the north-south direction is inhibited by conservation of angular momentum. The large scale atmospheric waves caused by the variation of the Coriolis frequency with latitude are called *Rossby waves*. In Sect. "Internal Solitons" we saw that KdV equation and its modifications model internal waves. Since there is a resemblance between internal waves and Rossby waves, it is expected that KdV like equations can model Rossby waves as well [8]. Under the simplest assumptions like long waves, incompressible fluid, β-plane approximations, etc. Benney [41] had derived the KdV equation as a model for Rossby

waves in the presence of east-west zonal flow. Boyd [42] had shown that long, weakly nonlinear, equatorial Rossby waves are governed either by KdV or MKdV equation. Recently it has been shown by using the eight years of Topex/Poseidon altimeter observations [43] that a detailed characterization of major components of Pacific dynamics confirms the presence of equatorial Rossby solitons.

Also observational and numerical studies of propagation of nonlinear Rossby wave packets in the barotropic atmosphere by Lee and Held [44,45] have established their presence, notably in the Northern Hemisphere as storm tracks, but more clearly in the Southern Hemisphere (see Fig. 7). They also found that the wavepackets both in the real atmosphere and in the numerical models behave like the envelope solitons of the nonlinear Schrödinger equation (Fig. 7).

An interesting application of KdV equation to describe Rossby waves is the conjecture of Maxworthy and Redekopp [46] that the planet Jupiter's Great Red Spot might be a solitary Rossby wave. Photographs taken by the spacecraft Voyager of the cloud pattern show that the atmospheric motion on Jupiter is dominated by a number of east-west zonal currents, corresponding to the jet streams of the earth's atmosphere [8]. Several oval-shaped spots are also seen, including the prominent Great Red Spot in the southern hemisphere. The latter one has been seen approximately this latitude for hundreds of years and maintains its form despite interactions with other atmospheric objects. One possible explanation is that the Red Spot is a solitary wave/soliton of the KdV equation, deduced from the quasigeostrophic form of potential vorticity equation for an incompressible fluid. A second test of the model is the combined effect of interaction of the Red Spot and Hollow of the Jupiter atmosphere on the South Tropical Disturbance which occurred early in the 20th century and lasted several decades. Maxworthy and Redekopp [46] have interpreted this interaction as that of two soliton collision of KdV equation, with the required phase shift [6,7,8].

The above connection between the KdV solitary wave and the rotating planet may be seen more concretely by the following argument as detailed in [8]. One can start with the quasigeostrophic form of the potential vorticity equation for an incompressible fluid in the form

$$\left\{\left(\frac{\partial}{\partial t}+v\frac{\partial}{\partial x}\right)+\epsilon\left(\frac{\partial\psi}{\partial y}\right)\left(\frac{\partial}{\partial x}-\frac{\partial\psi}{\partial x}\frac{\partial}{\partial y}\right)\right\}$$
$$\left\{\mu^2\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial}{\partial z}\left(K^2\frac{\partial}{\partial z}\right)\right\}\psi+(\beta-U'')\frac{\partial\psi}{\partial x}=0, \tag{43}$$

where x, y and z represent the east, north and vertical directions, while the function U is related to the horizontal stream function Φ as $\Phi(x,y,z,t) = \int_{y'}^{y} U(\eta)d\eta + \epsilon\psi(x,y,z,t)$ in terms of a zonal shear flow and a perturbation. In (43), in the β-plane approximation, the Coriolis parameter is given as $f = 2\Omega\sin\theta_0 + \beta y$ and the function $K(z)$ is given by $K(z) = 2\Omega\sin\theta_0 l_2/N(z)d$, where $N(z)$ is the Brunt–Väisälä frequency and l_2 is the characteristic length scales in the north-south direction while d is the length scale in the vertical direction. Note that K compares the effects of rotation and density of variation. Then $\mu = l_2/l_1$ represents the ratio of the length scales in the north-south and east-west directions.

In the linear ($\epsilon \to 0$), long wave ($\mu \to 0$) limit, the potential function ψ has the form $\psi = \sum_n A_n(x - c_n t)\phi_n(y)p_n(z)$, where c_n is deduced by solving two related eigenvalue problems [8], $(K^2 p_n')' + k_n^2 p_n = 0$, $p_n(0) = p_n(1) = 0$ and $\phi_n'' - k_n^2\phi_n + [(\beta - U'')/(U - c_n)]\phi_n = 0$, $\phi_n(y_s) = \phi_n(y_N) = 0$. If the various modes are separable on a short time scale, one can then deduce an evolution equation for the individual modes, by eliminating secular terms at higher order in the expansion. Depending on the nature and existence of a stable density of stratification, which is characterized by the function $N(z)$ mentioned above, either KdV or mKdV equation can be deduced for a given mode and thereby establishing the soliton connection in the present problem.

In spite of the complexity of the phenomenon underlying Rossby solitons there is clear evidence of the significance of solitonic picture.

Bore Solitons

Rivers which flow into the open oceans are usually affected by tidal flows, tsunami or storm surge. For a typical estuary as one moves towards the mouth of the river, the depth increases and width decreases. When a tidal wave, tsunami or storm surge hits such an estuary, it can be seen as a hydraulic jump (step-wise perturbations like a shock-wave) in the water height and speed which will propagate upstream [47]. Far less dangerous but very similar is the bore (mascaret in French), a tidal wave which can propagate in a river for considerable distances.

Typical examples of bores occur in Seine river in France and the Hooghli river in West Bengal in India. A bore existed in the Seine river upto 1960 and disappeared when the river was dredged. The tide amplitude here is one of the largest in the world. The Hughli river is a branch of the Ganges that flows through Kolkata where a bore of 1m is present, essentially due to the shallowness of the river. Another interesting example is that at the time

Solitons, Tsunamis and Oceanographical Applications of, Figure 8
Tidal bore at the mouth of the Araguari River in Brazil. The bore is undular with over 20 waves visible. (Adapted from the book "Gravity Currents in the Environment and the Laboratory" by John E. Simpson, Cambridge University Press with the courtesy of D.K. Lynch, John E. Simpson and Cambridge University Press)

of the 1983 Japan sea tsunami, waves in the form of a bore ascended many rivers [48]. In some cases, bores had the form of one initial wave with a train of smaller waves and in other cases only a step with flat water surface behind was observed (Fig. 8). The Hangzhou bore in China is a tourist attraction. Other well known bores occur in the Amazon in Brazil and in Australia.

Another interesting situation where bore solitons were observed was in the International H_2O Project (IHOP), as a density current (such as cold air from thunderstorm) intrudes into a fluid of lesser density that occurs beneath a low level inversion in the atmosphere. A spectacular bore and its evolution into a beautiful amplitude-ordered train of solitary waves were observed and sampled during the early morning of 20 June 2002 by the Leandre-II abroad the P-3 aircraft. The origin of this bore was traceable to a propagating cold outflow boundary from a mesoscale convective system in extreme western Kansas [49].

In the process of propagation the bore undergoes dissipation, dispersive disintegration, enhancement due to decrease of the river width and depth, influence of nonlinear effects and so on. The profile depends on the Froude number which is a dimensionless ratio of inertial and gravitational effects. Theoretical models have been developed to study these effects based on KdV and its generalizations [14]. For example, bore disintegration into solitons in channel of constant parameters can be studied in signal coordinates in terms of the KdV like equation for the perturbation of water surface,

$$\eta_x + \frac{1}{c_0}(1-\alpha\eta)\eta_t - \beta\eta_{ttt} = 0\,, \tag{44}$$

where $c_0 = \sqrt{gh}$, $\alpha = 3/2h$, $\beta = h^2/6c_0^3$, h being the depth of the river, with the bore represented by a Heaviside step function as the boundary condition at $x = 0$. Other effects then can be incorporated into a variable KdV equation of the form (26).

Future Directions

We have indicated a few of the most important oceanographical applications of solitons including tsunamis, internal solitons, Rossby solitons and bore solitons. There are other interesting phenomena like capillary wave solitons [15], resonant three and four wave interaction solitons [16], etc. which are also of considerable interest depending on whether the wave propagation corresponds to shallow, intermediate or deep waters. Whatever be the situation, it is clear that experimental observations as well as their theoretical formulation and understandings are highly challenging complex nonlinear evolutionary problems. The main reason is that the phenomena are essentially large scale events and detailed experimental observations require considerable planning, funding, technology and manpower. Often satellite remote sensing measurements need to be carried out as in the case of internal solitons and Rossby solitons. Events like tsunami propagation are rare and time available for making careful measurements are limited and heavily dependent on satellite imaging, warning systems and after event measurements. Consequently developing and testing theoretical models are extremely hazardous and difficult. Yet the basic modeling in terms of solitary wave/soliton possessing nonlinear dispersive wave equations such as the KdV and NLS family of equations and their generalizations present fascinating possibilities to understand these large scale complex phenomena and opens up possibilities of prediction. Further understanding of such nonlinear evolution equations, both integrable and nonintegrable systems particularly in higher dimensions, can help to understand the various phenomena clearly and provide means of predicting events like tsunamis and damages which occur due to internal solitons and bores. Detailed experimental observations can also help in this regard. It is clear that what has been understood so far is only the qualitative aspects of these phenomena and much more intensive work is needed to understand the quantitative aspects to predict them.

Bibliography

Primary Literature

1. Russel JS (1844) Reports on Waves. 14th meeting of the British Association for Advancement of Science. John Murray, London, pp 311–390

2. Bullough RK (1988) The Wave Par Excellence. The solitary progressive great wave of equilibrium of fluid: An early history of the solitary wave. In: Lakshmanan M (ed) Solitons: Introduction and Applications. Springer, Berlin
3. Korteweg DJ, de Vries G (1895) On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves. Philos Mag 39:422–443
4. Zabusky NJ, Kruskal MD (1965) Interactions of solitons in a collisionless plasma and recurrence of initial states. Phys Rev Lett 15:240–243
5. Gardner CS, Greene JM, Kruskal MD, Miura RM (1967) Method for solving the Korteweg–de Vries equation. Phys Rev Lett 19:1095–97
6. Lakshmanan M, Rajasekar S (2003) Nonlinear Dynamics: Integrability and Chaos. Springer, Berlin
7. Ablowitz MJ, Clarkson PA (1991) Solitons, Nonlinear Evolution Equations and Inverse Scattering. Cambridge University Press, Cambridge
8. Ablowitz MJ, Segur H (1981) Solitons and Inverse Scattering Transform. Society for Industrial and Applied Mathematics, Philadelphia
9. Scott AC (1999) Nonlinear Science: Emergence and Dynamics of Coherent Structures. Oxford University Press, New York
10. Dudley WC, Miu L (1988) Tsunami! University of Hawaii Press, Honololu
11. Osborne AR, Burch TL (1980) Internal solitons in the Andaman Sea. Science 258:451–460
12. Rossby GG (1939) Relation between variations in the intensity of the zonal circulation of the atmosphere. J Mar Res 2:38–55
13. Redekopp L (1977) On the theory of solitary Rossby waves. J Fluid Mech 82:725–745
14. Caputo JG, Stepanyants YA (2003) Bore formation and disintegration into solitons in shallow inhomogeneous channels. Nonlinear Process Geophys 10:407–424
15. Longuet–Higgins MS (1993) Capillary gravity waves of solitary type and envelope solitons in deep water. J Fluid Mech 252:703–711
16. Philips OM (1974) Nonlinear dispersive waves. Ann Rev Fluid Mech 6:93–110
17. Helal MA, Molines JM (1981) Nonlinear internal waves in shallow water. A theoretical and experimental study. Tellus 33:488–504
18. Benjamin TB, Bona JL, Mahoney JJ (1972) Model equations for long waves in nonlinear dispersive systems. Philos Trans A Royal Soc 272:47–78
19. Camassa R, Holm D (1992) An integrable shallow water equation with peaked solitons. Phys Rev Lett 71:1661–64
20. Zakharov VE (1968) Stability of periodic waves of finite amplitude on the surface of a deep fluid. J Appl Mech Tech Phys 2:190–194
21. Hasimoto H, Ono H (1972) Nonlinear modulation of gravity waves. J Phys Soc Jpn 33:805–811
22. Benney DJ, Roskes G (1969) Wave instabilities. Stud Appl Math 48:377–385
23. Davey A, Stewartson K (1974) On three dimensional packets of surface waves, Proc Royal Soc Lond A 338:101–110
24. Fokas AS, Santini PM (1990) Dromions and a boundary value problem for the Davey–Stewartson I equation. Physica D 44:99–130
25. Kundu A (ed) (2007) Tsunami and Nonlinear Waves. Springer, Berlin
26. Dias F, Dutykh D (2007) Dynamics of tsunami waves. In: Ibrahimbegovic A, Kozar I (eds) Extreme man-made and natural hazards in dynamics of structures. NATO security through Science Series. Springer, Berlin, pp 201–224
27. Dutykh D, Dias F (2007) Water waves generated by a moving bottom. In: Kundu A (ed) Tsunami and Nonlinear Waves. Springer, Berlin, pp 65–94
28. Okada Y (1992) Internal deformation due to shear and tensile faults in a half space. Bull Seism Soc Am 82:1018–1040
29. Carrier GF, Wu TT, Yeh H (2003) Tsunami runup and drawdown on a plane beach. J Fluid Mech 475:79–99
30. Banerjee P, Politz FF, Burgman R (2005) The size and duration of the Sumatra–Andaman earthquake from far-field static offsets. Science 308:1769–1772
31. Constantin A, Johnson RS (2006) Modelling tsunamis. J Phys A39:L215-L217
32. Segur H (2007) Waves in shallow water, with emphasis on the tsunami of (2004). In: Kundu A (ed) Tsunami and Nonlinear Waves. Springer, Berlin, pp 3–29
33. Lakshmanan M (2007) Integrable nonlinear wave equations and possible connections to tsunami dynamics. In: Kundu A (ed) Tsunami and Nonlinear Waves. Springer, Berlin, pp 31–49
34. Alpers W, Wang-Chen H, Cook L (1997) Observation of internal waves in the Andaman Sea by ERS SAR. IEEE Int 4:1518–1520
35. Hyder P, Jeans DRG, Cauqull E, Nerzic R (2005) Observations and predictability of internal solitons in the northern Andaman Sea. Appl Ocean Res 27:1–11
36. Apel JR, Holbroook JR (1980) The Sulu sea internal soliton experiment, 1. Background and overview. EOS Trans AGU 61:1009
37. Ostrovsky LA, Stepanyants YA (2005) Internal solitons in laboratory experiments: Comparison with theoretical models. Chaos 15(1–28):037111
38. Joseph RI (1977) Solitary waves in a finite depth fluid. J Phys A10:L225-L227
39. Benjamin TB (1967) Internal waves of permanent form in fluids of great depth. J Fluid Mech 29:559–592
40. Kundu PK, Cohen IM (2002) Fluid Mechanics, Second Edition. Academic Press, San Diego
41. Benney DJ (1966) Long nonlinear waves in fluid flows. Stud Appl Math 45:52–63
42. Boyd JP (1980) Equatorial solitary waves. Part I: Rossby solitons. J Phys Oceanogr 10:1699–1717
43. Susanto RD, Zheng Q, Xiao-Hai Y (1998) Complex singular value decomposition analysis of equatorial waves in the Pacific observed by TOPEX/Poseidon altimeter. J Atmospheric Ocean Technol 15:764–774
44. Lee S, Held I (1993) Baroclinic wave packets in models and observations. J Atmospheric Sci 50:1413–1428
45. Tan B (1996) Collision interactions of envelope Rossby solitons in a baratropic atmosphere. J Atmospheric Sci 53:1604–1616
46. Maxworthy T, Redekopp LG (1976) Theory of the Great Red Spot and other observed features of the Jovian atmosphere. Icarus 29:261–271
47. Caputo JG, Stepanyants YA (2007) Tsunami surge in a river: a hydraulic jump in an inhomogeneous channel. In: Kundu A (ed) Tsunami and Nonlinear Waves. Springer, Berlin, pp 97–112
48. Tsuji T, Yanuma T, Murata I, Fujiwara C (1991) Tsunami ascending in rivers as an undular bore. Nat Hazard 4:257–266
49. Koch SE, Pagowski M, Wilson JW, Fabry F, Flamant C, Feltz W, Schwemmer G, Geerts B (2005) The structure and dynamics of

atmospheric bores and solitons as determined from remote sensing and modelling experiments during IHOP, AMS 32nd Conference on Radar Meteorology, Report JP6J.4

Books and Reviews

Lamb H (1932) Hydrodynamics. Dover, New York
Miles JW (1980) Solitary waves. Ann Rev Fluid Mech 12:11–43
Stoker JJ (1957) Water Waves. Interscience, New York
Dauxois T, Peyrard M (2006) Physics of Solitons. Cambridge University Press, Cambridge
Johnson RS (1997) An Introduction to the Mathematical Theory of Water Waves. Cambridge University Press, Cambridge
Hammack JL (1973) A note on tsunamis: their generation and propagation in an ocean of uniform depth. J Fluid Mech 60:769–800
Drazin PG, Johnson RS (1989) Solitons: An Introduction. Cambridge University Press, Cambridge
Mei CC (1983) The Applied Dynamics of Ocean Surface Waves. Wiley, New York
Scott AC et al (1973) The soliton: a new concept in applied science. Proc IEEE 61:1443–83
Scott AC (ed) (2005) Encyclopedia of Nonlinear Science. Routledge, New York
Sulem C, Sulem P (1999) The Nonlinear Schrödinger Equation. Springer, Berlin
Helfrich KR, Melville WK (2006) Long nonlinear internal waves. Ann Rev Fluid Mech 38:395–425
Bourgault D, Richards C (2007) A laboratory experiment on internal solitary waves. Am J Phys 75:666–670

Space Plasmas, Dynamical Complexity in

Tom Chang
Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Plasma Resonances and Coherent Structures
Coarse-Grained Dissipation and Magnetic Reconfiguration
Non-Gaussian Probability Distribution Functions
Wavelet Transforms and Intermittency
Multifractals
Invariant Scaling
Dynamical Modeling – The Lu–Klimas Magnetic Field Reversal Model
Future Directions
Acknowledgments
Bibliography

Glossary

Alfvén wave Transverse magnetohydrodynamic wave that propagates along the magnetic field line direction in an electrically conducting fluid.

Anomalous resistivity Nonclassical resistivity in plasma dynamics.

Astronomical unit (AU) Unit of length nearly equal to the semi-major axis of Earth's orbit around the Sun.

Auroral electrojet index (AE) Empirical index that measures magnetic perturbations in the auroral zone. It provides a gauge for the horizontal current strength in the auroral ionosphere.

Auroral zone Region where auroral activities are observed.

Castaing distribution A convolution of Gaussians with variances distributed according to a log-normal distribution.

Chaos Nonsteady, nonperiodic, complex dynamical motion.

Coarse-grained dissipation Apparent dissipation due to interactions of coherent structures.

Coarse-graining Replacement of a detailed description with a lower-resolution coarse-grained model.

Coherent structure Large scale structure of nonlinearly interacting dynamical systems.

Correlation function Ensemble average of the correlation of fluctuations.

Criticality Behavior of extended system at which scale-invariance prevails.

Crossover Transition from one critical state to another or multitudes of critical states; or, transition from one fractal state to another or multitudes of fractal states.

Current sheet Sheet-like region with strong current density usually generated by strong local magnetic shear.

Cusp Funnel shaped region dividing the sunward and tailward sides of the magnetic field lines of the Earth.

Drift Alfvén vortex Field-aligned coherent Alfvénic vortex structure that drifts across the magnetic field.

Dynamical complexity Complex stochastic behavior of numerous nonlinearly interacting coherent structures of various sizes.

Dynamical renormalization group Group generated by coarse-graining scale transformations of a nonlinear dynamical system.

Extended similarity scaling (ESS) Scale invariant property among the structure or partition functions of a fluctuating event.

Exponent relation Algebraic relation among scaling exponents.

Extreme value theory Statistical theory dealing with extreme deviations from the mean.

Flatness (kurtosis) A normalized scale dependent measure related to the fourth order moment of the scale dependent coarse-grained probability distribution. It may be evaluated by taking the ensemble average of the square of the normalized power using the wavelet transform.

Forced and/or self-organized criticality (FSOC) Organization of a dynamical system into a scale invariant critical state with or without some tuning or forcing.

Fractal Fluctuations that exhibit generalized dimension of an irrational number.

Gaussian distribution Normal distribution.

Generalized dimension Dimension of the measure of fractal characteristics of a fluctuating event based on the limiting characteristic of the logarithm of the partition function at small scales.

Helicity Extent to which the magnetic field wraps around itself.

Heliosphere Magnetic bubble containing our solar system, including the solar wind.

Hysteresis History dependence of physical system that dissipates energy.

Intermittency Random fluctuations that exhibit non-self-similar characteristics.

Irrelevant parameter Parameter that becomes irrelevant under repeated coarse-graining transformations.

Ionosphere Upper most portion of atmosphere that is ionized and connected to the magnetosphere.

Local intermittency measure, Lim Normalized power in wavelet transform that is intensity, scale and location dependent.

Low-dimensional chaos Chaotic motion characterized by a small number of parameters.

Lu/Klimas model One-dimensional dynamical model with anomalous diffusion involving hysteresis, which exhibits scale invariant stochastic processes.

Magnetic reconfiguration (and/or reconnection) Reconfiguration of magnetic topology, which may or may not involve the phenomenon of magnetic reconnection of field lines.

Magnetohydrodynamics Mathematical description of electrically conducting fluids moving within electromagnetic environments.

Magnetosphere Region of space around the Earth (or any astrophysical object), that is controlled by its magnetic field.

Magnetotail Tail region (away from the Sun) of the magnetosphere of the Earth or any object in the heliosphere.

Multifractals Fluctuations characterized by multitudes of fractal dimensions that are not linearly related.

P-model Two-scale cascade model with "p" as the parameter that characterizes the fragmentation probability of the cascading process of intermittency.

Partition function of order q qth moment-order of the segmental coarse-grained probabilities.

PDF Probability distribution function.

Plasma resonance Site at which a particular plasma propagation mode vanishes.

Rank-ordered multifractal analysis, ROMA Analysis of multifractal characteristics based on the ordering of fractal dimensions in terms of scaled sizes of the intermittent fluctuations.

Relevant parameter Parameter that becomes more and more relevant under repeated coarse-graining transformations.

Response function Ensemble average of the response of the system due to fluctuations.

Scale invariants Quantities that do not vary under coarse-graining scale transformations.

Singularity spectrum $f(\alpha)$ Continuum distribution of differential measures with singularity index α.

Self-organized criticality (SOC) Auto-organization of a dynamical system into a scale-invariant critical state without significant tuning.

Solar wind Stream of charged particles in interplanetary space ejected from the upper atmosphere of the Sun.

Space plasma Ionized medium in the space environment.

Structure function of order q qth order moment of the probability distribution function of scale-dependent, coarse-grained fluctuations. It may be evaluated in terms of the ensemble average of the qth power of the scale-dependent fluctuations.

Taylor hypothesis Hypothesis that interprets temporal fluctuations as spatial fluctuations when the transit time of the fluctuations is much less than the characteristic evolution time. Also called Taylor's "frozen-in" hypothesis.

Turbulence Chaotic stochastic flow of a continuum or discrete medium with infinite number of degrees of freedom.

Wavelet transform Transform generated by a basis set of wave-packet-like functions, which provides information of intensity, location and scale of a fluctuating quantity.

Whistler wave Right-hand polarized electromagnetic wave in a plasma with frequency well below the electron cyclotron frequency and above the ion cyclotron frequency. The high frequency whistler waves travel

faster than the low frequency ones along the magnetic field lines in the plasma medium.

Definition of the Subject

Commonly called the fourth state of matter, plasmas make up 99 per cent of the Universe. They are the major constituents of the solar-terrestrial, interstellar, and intergalactic media – generally in the form of ionized gases. According to Tonks [118] who co-authored the celebrated Tonks and Langmuir [119] paper on oscillations in ionized gases, Langmuir [70] coined the terminology "plasma" to describe, in today's language, a quasineutral mixture – probably because of the likeness between the electrons and ions in an ionized medium and the red and white corpuscles in a blood plasma [99]. Recent observations, particularly in-situ measurements in the heliosphere and solar-terrestrial regions, have indicated that space plasmas commonly exhibit random, intermittent, and anisotropic fluctuating characteristics at all spatiotemporal scales while interacting with the electromagnetic environment. Such observations have led theoreticians and experimentalists alike to regard the stochastic behaviors of space plasmas as prime candidates of naturally occurring examples of dynamical complexity.

By definition, "dynamical complexity" is a phenomenon exhibited by a nonlinearly interacting dynamical system within which multitudes of different sizes of large scale "coherent structures" are formed, resulting in a complicated global nonlinear stochastic behavior for the dynamical system – vastly different from that could be surmised from the original dynamical equations.

The main purpose of this article is to demonstrate the prevalence of dynamical complexity in space plasmas and to indicate the various statistical methods of analyzing this kind of dynamical processes as well as to discuss theoretical and numerical methods that can provide the basic understanding of such processes.

Introduction

Traditionally, analyzes of space plasma processes are based on fluid or kinetic formulations. Among the fluid descriptions, the simplest is that based on the so-called one-component magnetohydrodynamic (MHD) equations, which includes the equations of motion, continuity, and energy conservation, the Ohm's law and the Maxwell equations for a single conducting fluid medium (see, e. g. [7]). Associated with such a formulation are the concepts of magnetic field lines, streamlines and other continuum variables such as velocity, electric and magnetic fields, as well as plasma, charge and current densities. The dynamical state of the system is understood in terms of a topology generally characterized by the relatively smooth or piecewise continuous variations of such entities with space and time.

Most of the observed space plasma processes, however, generally exhibit discernible turbulent fluctuations of such quantities. The standard approach to the theoretical analyzes of dynamically fluctuating states is based on the concepts of linear instabilities, nonlinear growths and interactions of the wave modes. Although the basic governing equations generally contain strong nonlinearities, one is led to believe that turbulent motions may be understood by expressing the fluctuations in Fourier modes (plane waves) and then considering the interactions among these nonlocal modes by requiring them to satisfy the "basic" equations. This procedure produced such intractable complications that have led to decades of futile search for a workable theory of "turbulence".

In reality, turbulent fluctuations in space plasmas are generally composed of the simultaneous coexistence of propagating modes and intermittent nonlinearly interacting spatiotemporal structures [32,33]. The "physics" of the bimodal state of such type of admixture of turbulent fluctuations may be understood from the point of view of the development and interactions of coherent structures arising from plasma resonance sites and the ensuring dynamical complexity resulting from such developments and interactions. In this article, we shall consider the dynamical complexity in space plasmas from such a concept. Sample results of direct numerical simulations and dynamical modeling including the calculated fluctuation probability distribution functions and local intermittency measures based on the wavelet transforms are provided to characterize the sporadic, localized, and scale-dependent nature of the intermittent turbulence. The concepts of multifractals, scale invariance and symmetry-breaking will be introduced. Observational examples demonstrating the existence of such phenomena and the associated invariance properties of intermittent space plasma turbulence will be provided.

Applications of the ideas of dynamical complexity to space plasmas were relatively recent. The first contact of these ideas with space plasmas was probably contained in a series of papers addressing the apparent low-dimensional chaotic behavior related to the dynamics of the Earth's magnetosphere [4,65,101,103,104,105]. These ideas were followed by a paper by Chang [22] that suggested the possibility of interpreting such phenomenon from the point of view of forced and/or self-organized criticality (FSOC) surmised from the method of the dynamic renormalization group (DRG). This suggestion was moti-

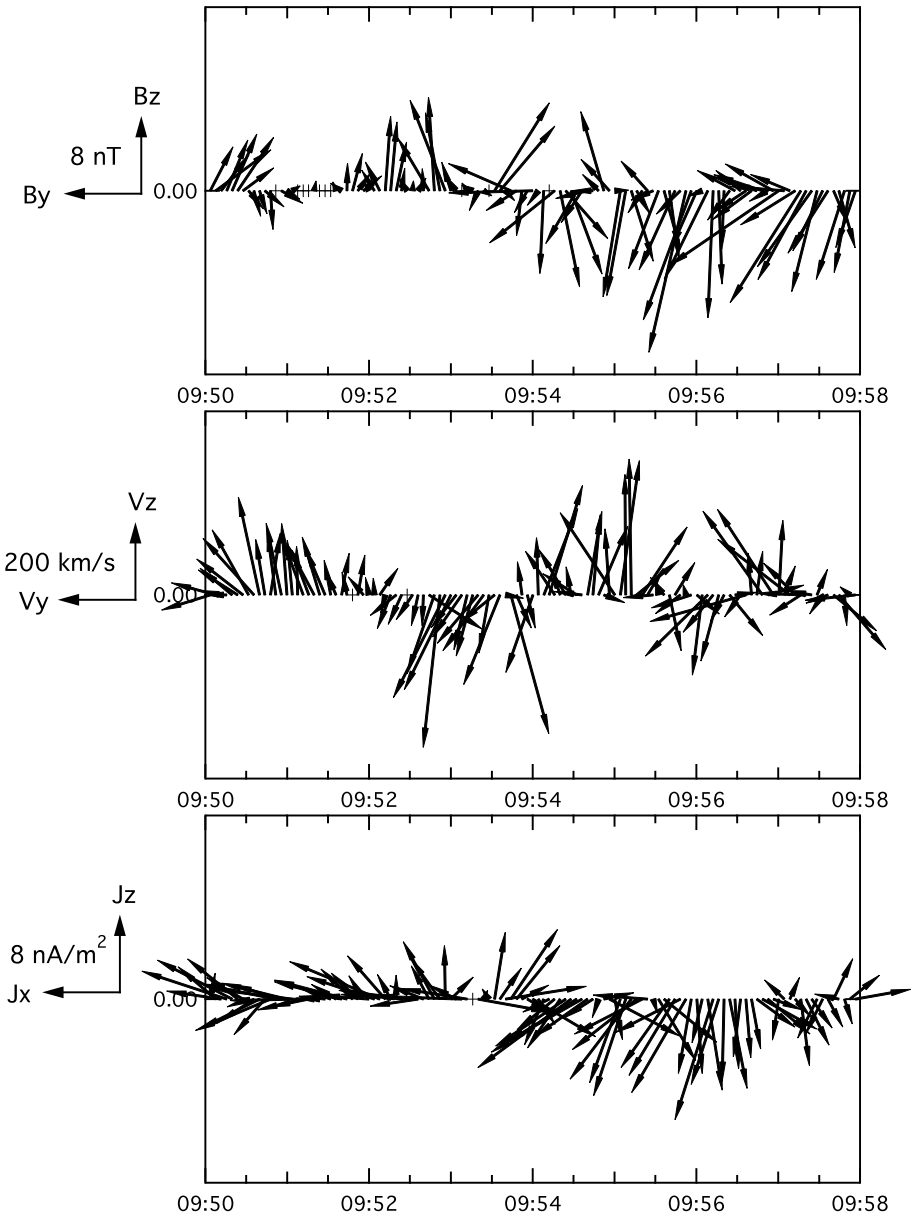

Space Plasmas, Dynamical Complexity in, Figure 1

Vector representation of the components of the magnetic field (B_y, B_z), ion plasma flow (V_y, V_z), and current density (J_x, J_z) during a magnetic substorm (i. e., when there was intense auroral activity) on August 22, 2001 observed by the Cluster spacecraft at a downstream distance of about 19 Re (Earth radii) in the magnetotail showing the large variability of these plasma parameters. Reprinted from [82] with thanks to the European Geophysical Union

vated by the low-dimensional chaos papers as well as the interpretation of the observed large variability of plasma fluctuations in the Earth's magnetotail as the result of sporadic and localized current disruptions [78], Figs. 1 and 2. At approximately the same time, Lu and Hamilton [74] used the idea of self-organized criticality (SOC) of Bak et al. [3] to explain the curious scale-independent behavior of the solar flare occurrence rate on the flare sizes. Such ideas suggest that complex dynamical systems generally organize themselves into states with statistical properties describable by power laws. (A succinct tutorial review of the SOC concept may be found in Jensen [63]). An alternative viewpoint of dynamical complexity is based on the concept of multifractals [54,98] which is a generaliza-

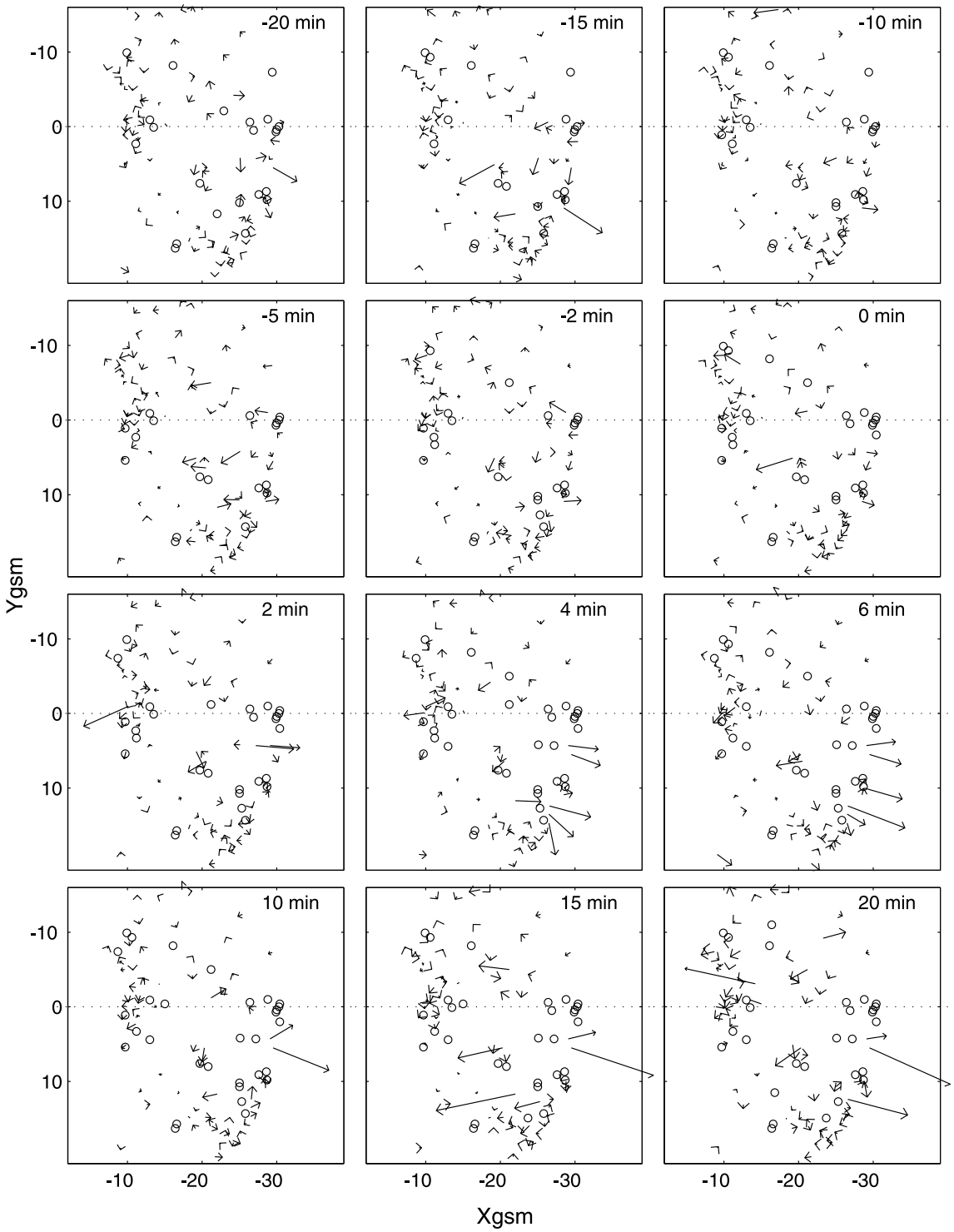

Space Plasmas, Dynamical Complexity in, Figure 2

A temporal sequence of synoptic patterns of total plasma flow showing large variabilities constructed from superposed epoch analysis of 102 magnetic substorm events from Geotail spacecraft observations. The scale is 100 km/s per Re (Earth radius). The acronym GSM stands for Geocentric Solar Magnetospheric coordinate system. The *X*-axis points to the Sun. The *XZ*-plane contains the Earth's dipole axis, and the *Y*-axis completes the right-handed coordinate system in units of Re. Reprinted from [80] with thanks to the American Geophysical Union

tion of the original ideas of fractals introduced by Mandelbrot [83]. Burlaga [12,13,14] was the first to incorporate such ideas to interpret the intermittent turbulent behavior of the solar wind. Other solar wind studies based on such a point of view followed, e. g. [16,17,18,19,20,58,59,60,84,102]. These ideas were followed by Consolini et al. [41] and Consolini [37] in their interpretations of the observed fractal time series of the Auroral Electrojet (AE) indices.

Since then, a flurry of activities has blossomed into the current investigations of the phenomenon of complexity in space plasmas. Perhaps the most noteworthy are the connection of such phenomena with the traditional ideas of intermittent turbulence in plasmas and the development of the concept of crossovers (transitions) from one criticality to another and onto multifractals [27,28]. This article will provide simple descriptions of some of these modern ideas in conjunction with actual observational results and numerical simulations.

Due to the limitation of space, this entire article will center on the discussion of the complexity of space plasmas from the fluid description. It should be understood that situations also exist in space plasmas where the physical phenomena are related to the complexity of kinetic plasma effects, see, e. g., [33,111,115]. Since the purpose of this article is to introduce to the readers the basic concepts of – and recent developments in – the physics of dynamical complexity in space plasmas, only those primary references that touch upon these introductory ideas are included. No attempt has been made to include the hundreds of outstanding contributions that are available on the subject.

The contents of this article are structured as follows. In the next section, the concepts of plasma resonances and coherent structures are introduced. This is followed by a discourse on coarse-grained dissipation and magnetic reconfiguration. The ideas of non-Gaussian probability distributions of turbulent fluctuations, wavelet transforms and intermittency are then discussed in the two subsequent sections. Next, we introduce the basic ideas of multifractals in terms of structure functions, partition functions, singularity measures and rank-ordered multifractal spectra in Sect. "Multifractals". The concepts of scale invariants, forced and/or self-organized criticality are described in Sect. "Invariant Scaling" along with a discourse relating these ideas to multifractals. In Sect. "Dynamical Modeling – The Lu–Klimas Magnetic Field Reversal Model", we discuss the utility of dynamical models. A simple one-dimensional model, the Lu–Klimas model, is provided to illustrate the usefulness of such methods. This is followed by a brief discussion of what future entails in the study and research of dynamical phenomena of complexity related to space plasmas. In addition to the primary list of references that are cited in the article, we also provide a secondary list of review articles and books in the bibliography for further in-depth study.

Plasma Resonances and Coherent Structures

Plasmas are known for their ability to form numerous types of coherent structures of varied sizes with scales that are generally much larger than those of the constituent particles (ions, electrons, neutrals) of the plasma medium. The reason behind this is closely related to the nonlinear and long range forces of the dynamical couplings among the constituent particles and species interacting within the electromagnetic environment.

Examples of coherent structures abound in the literature on theories and observations of nonlinear space plasma processes. They may appear as convective forms, nonlinear solitary structures, pseudo-equilibrium configurations, or other types of spatiotemporal entities. They may be locally generated or convected from elsewhere. Some of these structures may be more stable and long-lived; others may be less stable and relatively short-lived and sometimes partially formed. Generally, such structures are not purely laminar entities; they are likely composed of bundled fluctuations of all frequencies and scales. Due to the nature of the physics of complexity, it would be futile to attempt to evaluate and/or study the details and stabilities of each of these countless interacting structures or partially formed structures, though some basic understanding of the different types of structures would be helpful in the comprehension of the full stochastic complexity that results from their underlying nonlinear dynamics.

As an example of such coherent structures, let us consider the behavior of Alfvénic flux tubes in magnetized plasmas. We shall base our theoretical discussion by considering the plasma medium as a single, charge-neutral, conducting fluid moving in and interacting self-consistently with its electromagnetic environment. Because the plasmas in the space environment are generally highly rarefied, we can typically neglect the dissipative effects from particle collisions and assume the plasma is perfectly conducting and inviscid. The most elemental mathematical formulation for magnetized plasmas under such simplifying assumptions is the so-called ideal incompressible magnetohydrodynamics (MHD). The basic equations are the equations of continuity and motion, and the relevant Maxwell's equations where the displacement current is ne-

glected for nonrelativistic motion.

$$\nabla \cdot \mathbf{v} = 0 \quad \text{(continuity equation for incompressible medium)} \tag{1}$$

$$\rho \mathrm{d}\mathbf{v}/\mathrm{d}t = \mathbf{j} \times \mathbf{B} - \nabla p \quad \text{(equation of motion for an inviscid medium)} \tag{2}$$

$$\nabla \cdot \mathbf{B} = 0 \text{ (Gauss' law)} \tag{3}$$

$$\partial \mathbf{B}/\partial t + \nabla \times \mathbf{E} = 0 \quad \text{(Faraday's law)} \tag{4}$$

$$\mathbf{j} = \nabla \times \mathbf{B} \quad \text{(Ampère's law neglecting the displacement current)} \tag{5}$$

where $\mathrm{d}/\mathrm{d}t = \partial/\partial t + \mathbf{v} \cdot \nabla$ is the convective derivative, ρ is the density of the medium, $\mathbf{v}$ is the velocity, $\mathbf{j}$ is the current density, $\mathbf{B}$ is the magnetic field, $\mathbf{j} \times \mathbf{B}$ is the Lorentz force, p is the fluid pressure, and $\mathbf{E}$ is the electric field. These equations are written in SI units with the vacuum permeability set equal to 1.

To complete the set of equations of such a formulation, we complement the above equations with the Ohm's law for a perfectly conducting medium:

$$\mathbf{E} + \mathbf{v} \times \mathbf{B} = 0 \tag{6}$$

These equations may be combined to form the following set of equations of motion and induction:

$$\rho \mathrm{d}\mathbf{v}/\mathrm{d}t = (\mathbf{B} \cdot \nabla)\mathbf{B} + \cdots \tag{7}$$

$$\mathrm{d}\mathbf{B}/\mathrm{d}t = (\mathbf{B} \cdot \nabla)\mathbf{v} \tag{8}$$

where the eclipses represent the gradient of the total pressure ($p + B^2/2$). Standard arguments lead Eqs. (7) and (8) to a linearized wave equation characterizing the fundamental propagation of small fluctuations of $\mathbf{v}$ and $\mathbf{B}$. This wave equation admits the well-known Alfvén waves which are transverse waves to the magnetic field $\mathbf{B}$ and propagate along the magnetic field direction. The phase velocity v_A of the Alfvén wave is $B/\sqrt{\rho}$, i. e., proportional to the strength of the magnetic field and inversely proportional to the square root of the density ρ of the plasma. For such waves to propagate, the operators on the right hand sides of (7) and (8) must not vanish, i. e., $\mathbf{B} \cdot \nabla \to i\mathbf{k} \cdot \mathbf{B} = k_\parallel \neq 0$. Therefore, the propagation vector $\mathbf{k}$ must contain a field-aligned component $k_\parallel$. When the parallel component $k_\parallel$ of the propagation vector vanishes (i. e., at the resonance sites), the fluctuations are localized. Around these resonance sites (usually in the form of three-dimensional space curves), it may be shown that the fluctuations – which will try to propagate away as Alfvén waves – are held back by the background magnetic field and the plasmas medium, thereby forming the so-called Alfvénic coherent structures (which are actually domains of bundled stochastic fluctuations).

Coarse-Grained Helicity

We now consider such magnetized domains near the Alfvenic resonance sites. For an ideal MHD system, any physically acceptable magnetic field must satisfy the solenoidal condition (Gauss' law): $\nabla \cdot \mathbf{B} = 0$. Also, any variation of the field away from the initial value must satisfy the constraints (4) and (6), i. e., Faraday's law and Ohm's law for infinite conductivity.

It may be easily demonstrated that these constraints are equivalent to an infinite set of integral constraints involving the helicity K, such that

$$K = \int_V \mathbf{A} \cdot \mathbf{B} \mathrm{d}V \tag{9}$$

is an invariant for any single connected volume V enclosed by a flux surface, where $\mathbf{A} = \nabla \times \mathbf{B}$ is the vector potential.

Taylor's Conjecture

We are interested in the situation where there are domains with stochastic turbulent fluctuations within which field lines merge and mix indistinguishably. Thus, it will be difficult – in fact, unpractical – to discuss the topology of individual field lines. Nonetheless, it was suggested by Taylor [114] that when the volume integral of (9) is taken over the stochastic region, the coarse-grain averaged helicity in a relaxing state would be essentially conserved. As the domain of the stochastic region near the Alfvenic resonance site relaxes to a statistically stationary minimum energy state under the constraint of this conjecture of conservation of coarse-grained helicity, it may be shown using the variational principle that such a domain (coherent structure) will be essentially force-free in the sense that $\mathbf{j} \times \mathbf{B} = 0$ where $\mathbf{j}$ and $\mathbf{B}$ are the mean current density and magnetic field, respectively.

We are, of course, interested in dynamical states that are far from equilibrium. Thus, in visualizing the relaxed states from the point of view of this *Taylor's conjecture*, we shall consider timescales such that "nearly coherent" structures are formed. These structures actually move, mix

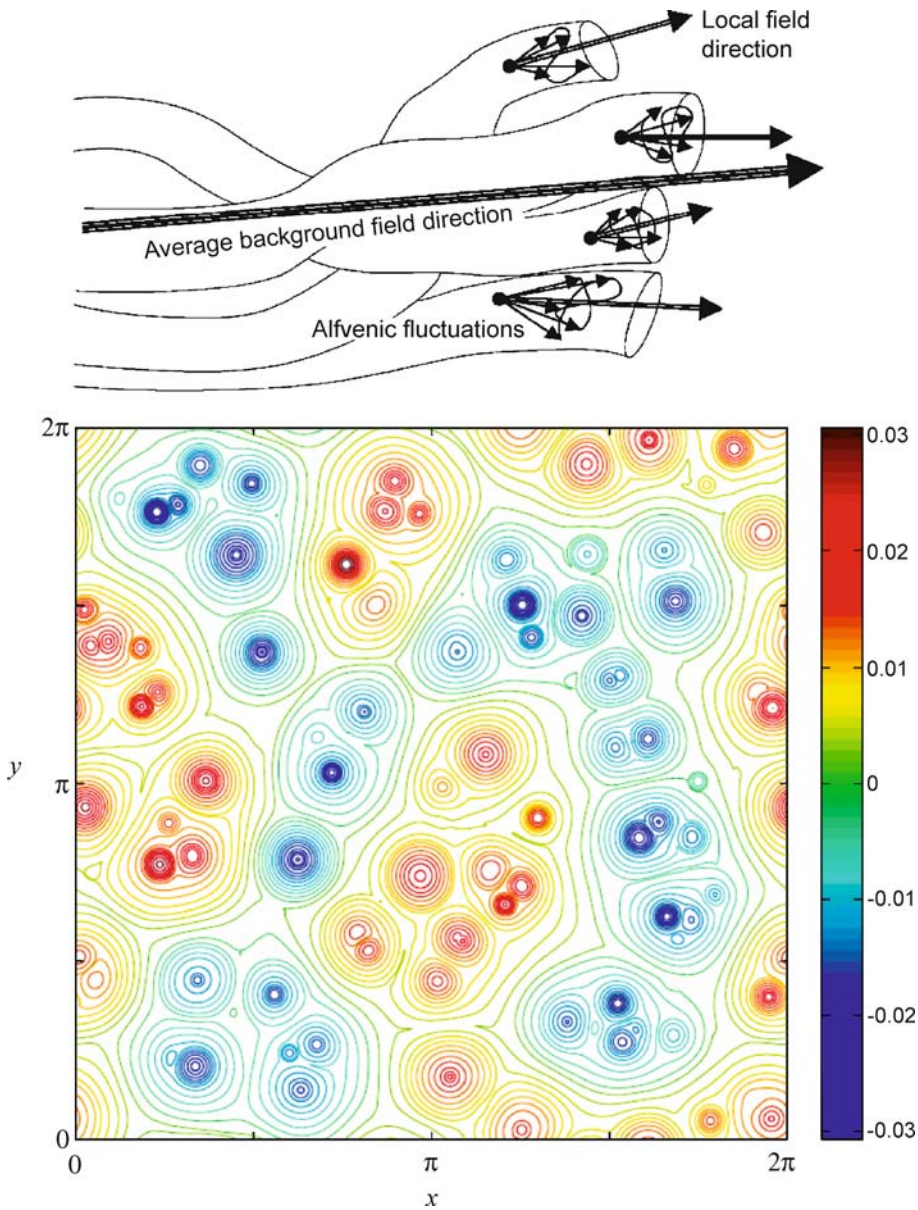

Space Plasmas, Dynamical Complexity in, Figure 3
***Top*: Schematics of tangled flux tubes in the solar wind. Each flux tube is characterized by a local magnetic field direction aligned approximately with the background field and the presence of Alfvénic fluctuations makes the magnetic field vector wander randomly about this direction. Moving across the tubes, strong intermittent interactions of predominantly nonpropagating (i. e., resonant) fluctuations are expected. Reprinted from [10] with permission from Elsevier. *Bottom*: 2D MHD simulation of cross-sectional view of interacting flux tubes for homogeneous intermittent turbulence. *Colors* represent current intensities and directions. Adapted from [33] with thanks to the American Institute of Physics**

and sometimes merge together while immersed in an otherwise turbulently diffusing plasma medium.

To obtain some physical insight of the geometries of these magnetic coherent structures, let us consider the special situation in the solar wind and make the reasonable assumption that the perturbed magnetic field fluctuations are much smaller than – and essentially transverse to – the mean magnetic field B_0 which will be temporarily assumed to be uniform for the current discussion [89]. Thus, let us write $\mathbf{B} = (\delta B_x, \delta B_y, B_0)$, where z is in the direction of

the mean magnetic field, and (x, y) are orthogonal coordinates normal to z. The force-free condition for constant B_0 and $\nabla \cdot \mathbf{j} = 0$ then leads approximately to the scalar condition $\mathbf{B} \cdot \nabla j_z = 0$, obtained by taking the z-component of the curl of $\mathbf{j} \times \mathbf{B} = 0$. We have, then,

$$B_0 \partial j_z / \partial z = -(\delta B_x \partial/\partial x + \delta B_y \partial/\partial y) j_z \,. \tag{10}$$

For convenience, let us introduce the flux function ψ by writing $(\partial\psi/\partial y, -\partial\psi/\partial x) = (\delta B_x, \delta B_y)$ for the perturbed transverse components of the magnetic field in the (x, y) directions such that the Gauss law of magnetism, $\nabla \cdot \mathbf{B} = 0$ is satisfied. Then, j_z and ψ are governed by Eq. (10) and the Ampere's law, $\nabla \times \mathbf{B} = \mathbf{j}$.

A simple example of the flux function and axial current density satisfying the above conditions would be the class of circularly cylindrical solutions of $\psi(r)$ and $j_z(r)$, [25,26]. Generally, the solutions would be more involved because of the variabilities of the local conditions of the plasma and the three-dimensional geometry. Moreover, the dynamic coherent structures with the inclusion of plasma pressure and other modifying effects (including electron-inertia terms) would be even more complicated. However, we expect these structures to be in the form of nearly field-aligned flux tubes, Fig. 3 [10,25,26,136,137,138]. Existence of Alfvénic flux tubes in the magnetopause and magnetotail have also been suggested by Tetreault [116,117] and Chang [23,24], respectively.

Generally, there exist various types of propagation modes (whistler modes, electromagnetic ion cyclotron waves, etc.) in magnetized plasmas. Thus, we envision a corresponding number of different types of plasma resonances and associated coherent structures that typically characterize the dynamics of the plasma medium under the influence of an electromagnetic background.

Coarse-Grained Dissipation and Magnetic Reconfiguration

When coherent Alfvénic flux tubes with the same polarity migrate toward each other, strong local magnetic shears are created, Fig. 4. Wu and Chang [136,137,138] have demonstrated that the existing sporadic nonpropagating fluctuations in the strong local shear region, particularly those close to the neutral sheet (i. e., at the location where the local magnetic field vanishes), will stay in the region and continue to interact nonlinearly. On the other hand, fluctuations away from the neutral sheet region, are nonresonant and will therefore propagate away along magnetic field lines as Alfvén waves. Combined with the magnetic shear geometry, the resonant fluctuations will induce a nonlinear instability near the neutral sheet region, which will produce more fluctuations – nonresonant and resonant. The nonresonant fluctuations will again propagate away as Alfvén waves while the resonant ones will join the other resonant fluctuations and interact nonlinearly, thereby broadening the resonance region, Fig. 5.

This combined phenomenon of "coarse-grained dissipation" depletes the energy originally contained in the coarse-grained magnetic fields near the shear region, initiating a reconfiguration of the coherent structures. In coarse-grained sense, it breaks some of the closed field lines of each of the coherent structures and then reconnects them into single closed field lines. And this process continues with the system adjusting intermittently with the surrounding environment until all free energies are exhausted, eventually leading to the formation of one single combined coherent structure with one set of coarse-grained concentric closed field lines, Fig. 6.

The final state of the resulting coherent structure will have less average energy due to the combined dissipation of Alfvénic propagation of nonresonant fluctuations and nonlinear interactions of the resonant fluctuations (i. e., resonance broadening). Such is the manifestation of continuous magnetic topological reconfiguration due to the dynamic "fluctuation-induced nonlinear instability". And this merging process may repeat over and over again among the coherent structures of the same polarity, from the largest scales to the smallest scales where kinetic effects may have to be included.

On the other hand, when coherent structures of opposite polarities approach each other due to the forcing of the surrounding plasma, they might repel each other, scatter, or induce magnetically quiescent localized regions.

Under any of the conditions of the above interaction scenarios, new fluctuations will be generated. And these new fluctuations can provide new resonance sites, thereby nucleating new coherent structures of varied sizes. This kind of tangled geometry of interacting flux tubes leading to dynamical complexity has been described by Bruno et al. [10] who first deduced their existence in the solar wind, as "cooked spaghetti", Fig. 3 (top panel), and demonstrated by Wu and Chang [136,137,138], Fig. 3 (bottom panel), Matthaeus et al. [90], Fig. 7, and others with MHD simulations. And they are the source of the observed sporadic and localized current disruptions in the Earth's magnetotail and elsewhere [75,78,82].

Topological reconfigurations of such nature occur quite frequently for the dynamical interactions of coherent structures in space plasmas and are not limited just to flux tubes; for example, Sundkvist et al. [109] and Alexandrova et al. [1] have observed intermittent interactions of drift Alfvén vortices in the cusp and magnetosheath, respec-

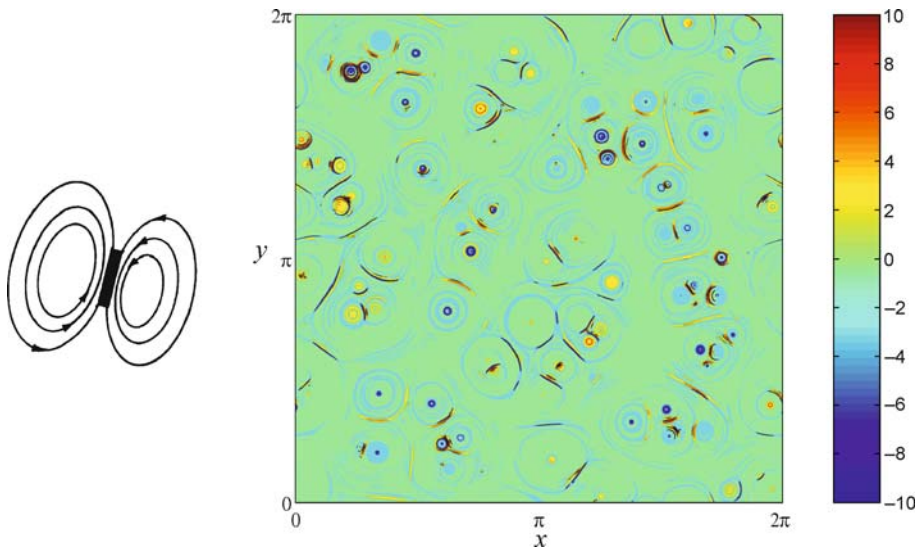

Space Plasmas, Dynamical Complexity in, Figure 4
Cross-sectional views of interacting Alfvénic flux tubes. *Left*: Schematic of merging. Arrows indicate directions of magnetic field and blackened area indicates location of strong shear. *Right*: 2D MHD simulation of the current intensities for homogeneous intermittent turbulence. *Colors* represent the intensities and directions. Adapted from [137] with permission from Elsevier

tively. Similarly, such enhanced intermittency at the intersection regions of whistler coherent structures has also been surmised by Consolini and Lui [40] and Consolini et al. [42] in the plasma sheet. All such intermittent interactions are quite akin to the avalanche phenomenon prevalent in sandpile models [3,35,36,62,133]. They are the origin of the various observed magnetic reconnection signatures in space plasmas.

This stochastic behavior of the interactions of the plasma coherent structures is a phenomenon of "dynamical complexity" as defined in Sect. "Definition of the Subject", which, for emphasis, is reproduced below:

"Dynamical complexity" is a phenomenon exhibited by a nonlinearly interacting dynamical system within which multitudes of different sizes of large scale "coherent structures" are formed, resulting in a complicated global nonlinear stochastic behavior for the dynamical system – vastly different from that could be surmised from the original dynamical equations.

Non-Gaussian Probability Distribution Functions

The fluctuations that are induced by the interactions and mergings of coherent structures are sporadic and localized. Since the coherent structures are numerous and outsized, we expect the fluctuations within the interaction regions of these structures (resonance overlap regions) to be large and occur relatively more frequently than those that would have been expected from a medium of the original minute plasma particles (electrons, ions and neutrals). A technique useful in gauging the degree of such effects is by studying the shapes of the probability distribution functions (PDFs) of intermittent fluctuations at varying scales.

To demonstrate this, let us refer to one of the 2D MHD simulations described previously [32,33,136,137,138]. (In the following discussions, the measured fluctuations may be those of any physical property of the MHD medium. We shall choose the physical property as the strength of the magnetic field $B(x_i)$ to render the discussions more specific). Consider, for example, the spatial series of the strength of the magnetic field $B(x_i)$ for a given time t and at fixed y, with $x_i = i\delta$ where $i = 0, 1, 2, \ldots, N$, and δ is the grid size of the simulation. We can then form the coarse-grained differences

$$\delta B_i^2(\Delta) = B^2(x_i + \Delta) - B^2(x_i) \tag{11}$$

within the interval with $\Delta = k\delta$ (k: an integer) and generate the probability distribution function $P(\delta B^2, \Delta)$. Figure 8 displays the calculated results of $P(\delta B^2, \Delta)$ for a 2D MHD simulation for the homogeneous case for several coarse-grained scales, Δ. From this figure, we note that the distribution for each scale falls nearly onto a smooth curve with the exception at the tails where scatterings from the mean are more visible. And the shapes of the PDF curves deviate more and more from that of a Gaussian at smaller and smaller scales.

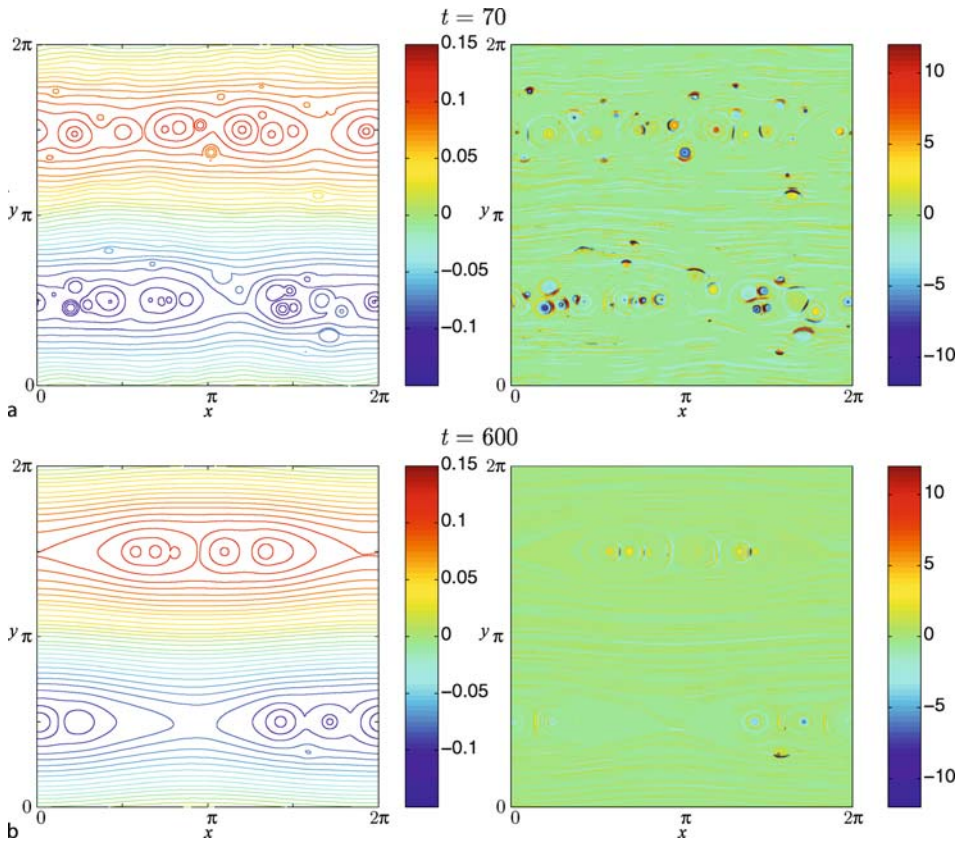

Space Plasmas, Dynamical Complexity in, Figure 5

2D MHD simulation of coarse-grained dissipation in sheared magnetic field. *Left*: Contours of magnetic flux. *Right*: Corresponding current density distributions. Initially, new fluctuations are excited and the resonant fluctuations begin to interact nonlinearly near the neutral sheet region (near $y = \pi/2, 3\pi/2$). Eventually, the field lines are reconnected in the coarse-grained sense. Adapted from [137] with permission from Elsevier

Such PDFs sometimes satisfy the one-parameter scaling form:

$$P(\delta B^2, \Delta)\Delta^s = P_s(\delta B^2/\Delta^s) \tag{12}$$

where s is the parameter (or scaling exponent) such that all PDFs essentially collapse onto one master scaling function P_s. Such scaling behavior seems to be approximately satisfied for the above simulated result with $s \approx 0.335$ (Fig. 9), although closer examination reveals some discrepancies – particularly in the tail regions [33]. The more subtle nature of the scaling properties of such PDFs will be considered further in Sect. "Multifractals".

We shall now proceed to make contact with in-situ spacecraft observations. Until recently, most of the turbulence data in space came from measurements obtained by one single spacecraft. The data are then analyzed based on the Taylor hypothesis [113] which assumes that the transit time of eddies (bundled fluctuations) is much less than the characteristic evolution time. For fully developed MHD turbulence, the characteristic eddy turnover (evolution) time may be estimated:

$$t(\text{evolution}) \sim (\lambda/2\pi)/[(\delta B/B_0)v_A] \tag{13}$$

where λ is the typical size of the eddy, v_A is the Alfvén speed based on the mean magnetic field B_0 and δB is the average fluctuating magnetic field of the eddy in question. The transit time of the eddy is:

$$t(\text{transit}) \sim \lambda/v_s \tag{14}$$

where v_s is the relative streaming speed between the plasma medium and the spacecraft.

Thus, as it has been demonstrated by Matthaeus and Goldstein [88] that the requirement of $t(\text{evolution}) \gg t(\text{transit})$ becomes:

$$v_s/v_A \gg 2\pi(\delta B/B_0) \tag{15}$$

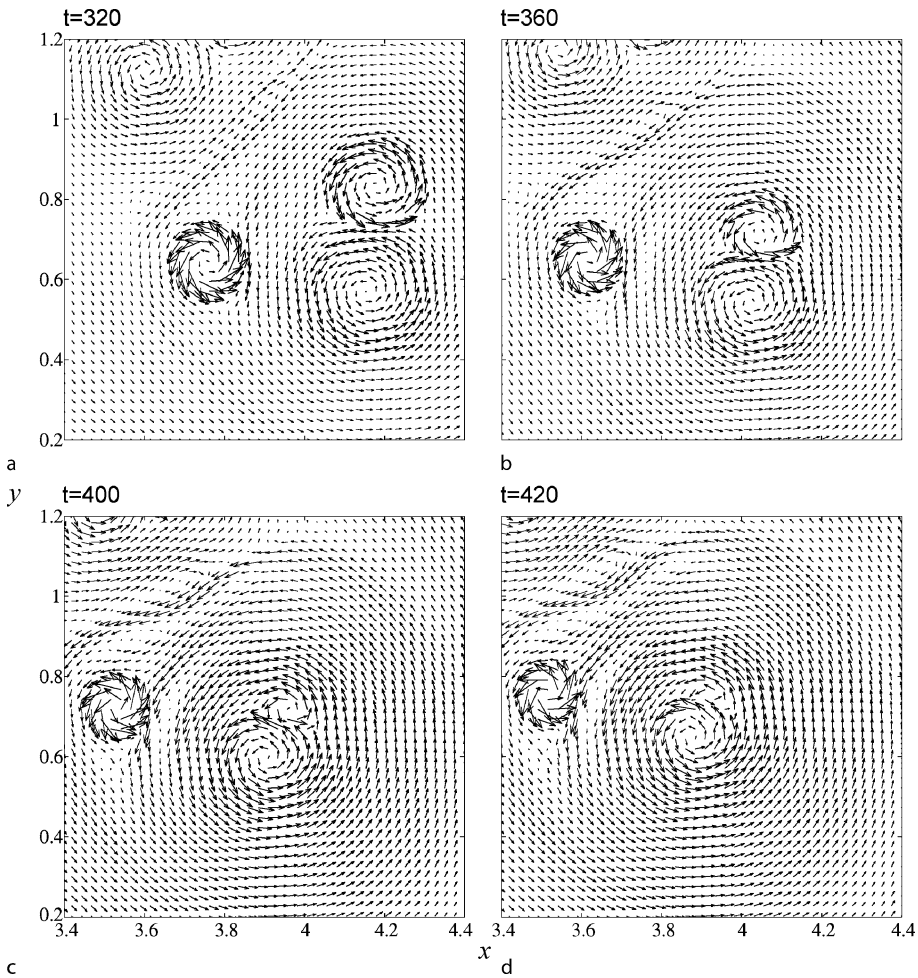

Space Plasmas, Dynamical Complexity in, Figure 6

2D MHD simulation of merging of flux tubes. *Arrows* indicate magnitudes and directions of magnetic fields. Reprinted from [137] with permission from Elsevier

As pointed out by Tu and Marsch [120], at 1 AU for solar wind observations the wind speed (which is essentially the relative speed between the plasma medium and the spacecraft since it is much larger than the spacecraft speed) is of the order of 10 times that of the ambient Alfvén speed and $\delta B/B_0 \simeq 0.5$. Therefore, condition (15) is satisfied for solar wind turbulence in that region. Because the spacecraft speed is much smaller than the solar wind speed, the data gathered by instruments aboard the spacecraft essentially sample the spatial fluctuations in the direction of the solar wind, particularly at times when the solar wind turbulence is homogeneous at large scales and fully developed. Thus, for example, the time series of such measurements of the strength of the magnetic field $B(t_i)$, where $t_i = i\delta$ with $i = 0, 1, 2, \ldots, N$ and δ is the sampling interval, may be interpreted based on the hypothesis as the spatial series of the ambient turbulence fluctuations with $x = v_s t$ (x: direction of the solar wind velocity). We can then construct the probability distribution function $P(\delta B, \tau)$ or $P(\delta B^2, \tau)$ with the coarse-grained scale $\tau = k\delta$ (k: integer) in analogy to the calculations related to the simulation results discussed above.

Non-Gaussian probability distribution functions from such statistics have been commonly observed in the solar wind for fluctuations related to the magnetic, velocity, density and other field variables, e. g. [15,47,48,56,71,72,85,107,108]. Figure 9 depicts the typical PDFs of magnetic field fluctuations obtained from such an analysis for both the slow ($\sim (400 \pm 200)$ km/s) and fast ($\sim (700 \pm 100)$ km/s) solar wind [108]. The similarities between the simulated results and results from such observations as exhibited in Fig. 9 are quite striking.

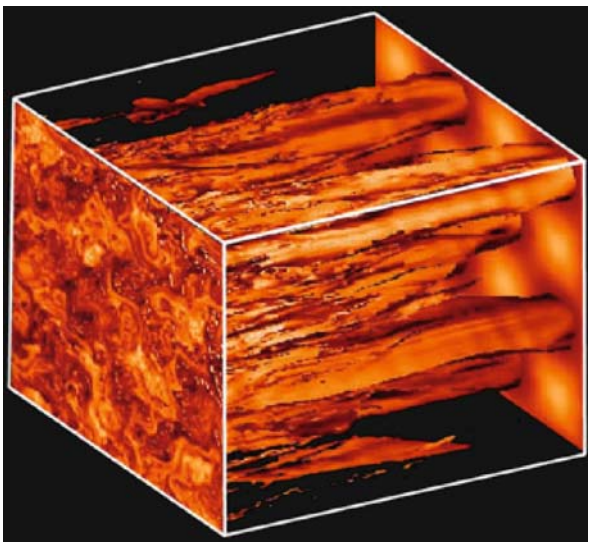

Space Plasmas, Dynamical Complexity in, Figure 7
MHD simulation of a two-component model, with 80% of its energy in two-dimensional modes, exhibiting tangled flux tubes with considerable transverse complexity. Reprinted from [90] with permission from the American Astronomical Society

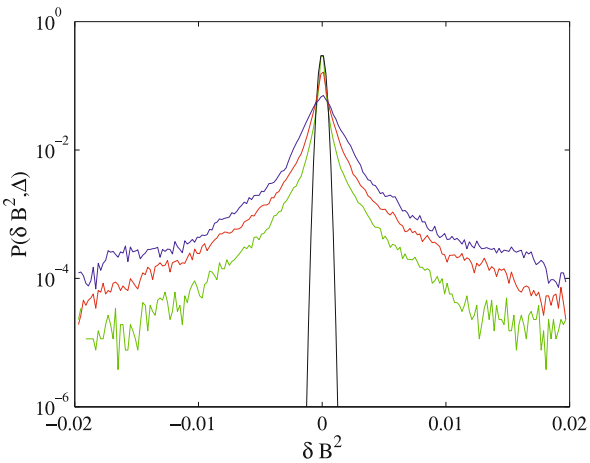

Space Plasmas, Dynamical Complexity in, Figure 8
Probability distribution function $P(\delta B^2, \Delta)$ from a 2D MHD simulation for homogeneous turbulence for $\Delta = 2$ (*green*), 8 (*red*), 32 (*blue*) units of grid spacing. The *black* curve is the PDF for Gaussian fluctuations. Reprinted from [34] with permission from Springer Science and Business Media

One popular empirical technique in expressing the shape of the PDFs of the observed intermittent turbulent fluctuations is the Castaing distribution. In 1990, Castaing et al. [21] suggested that the intermittent fluctuations might perhaps be viewed as an ensemble of subsets of fluctuations ξ, each subset having a normal distribution:

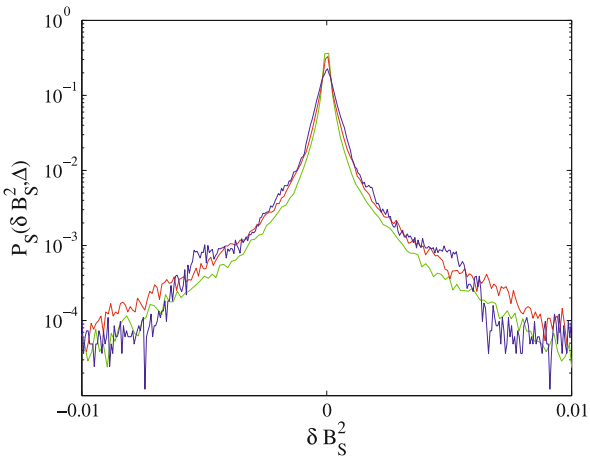

Space Plasmas, Dynamical Complexity in, Figure 9
Scaled probability distribution function $P_s(\delta B_s^2, \Delta)$ with $\delta B_s^2 = \delta B^2/\Delta^s$, $s = 0.335$, and $\Delta = 2$ (*green*), 8 (*red*), 32 (*blue*) units of grid spacing. Adapted from [33] with thanks to the American Institute of Physics

$$P_\sigma(\xi) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\xi^2/2\sigma^2) \tag{16}$$

where σ is the variance. The intermittency is then assumed to be due to the fluctuations of the variances that satisfy a log-normal distribution:

$$Q_\lambda(\sigma)\mathrm{d}\sigma = \frac{1}{\lambda\sqrt{2\pi}} \exp\left(-\frac{\ell n^2(\sigma/\sigma_0)}{2\lambda^2}\right) \mathrm{d}\ell n\sigma \tag{17}$$

where σ_0 is the most probable variance of ξ, and λ is the variance of $\ell n\sigma$. Combining (16) and (17) gives the Castaing distribution

$$\Pi_\lambda(\xi) = \frac{1}{2\pi\lambda} \int_0^{+\infty} \exp\left(-\frac{\xi^2}{2\sigma^2}\right) \exp\left(-\frac{\ell n^2(\sigma/\sigma_0)}{2\lambda^2}\right) \frac{\mathrm{d}\sigma}{\sigma^2} . \tag{18}$$

Such fits of data to (18) are shown as solid curves in Fig. 10. How well does such an ansatz characterize the data depends very much on the type and scale of the fluctuations [47]. Other empirical fits such as the kappa-distribution and gamma-distribution have also been suggested, but we do not have space here to enumerate and discuss these and other phenomenological models.

Similar observational results have been detected for the magnetic field and velocity fluctuations in the plasma sheet [2,134], in the cusp [44,139], for the electric field

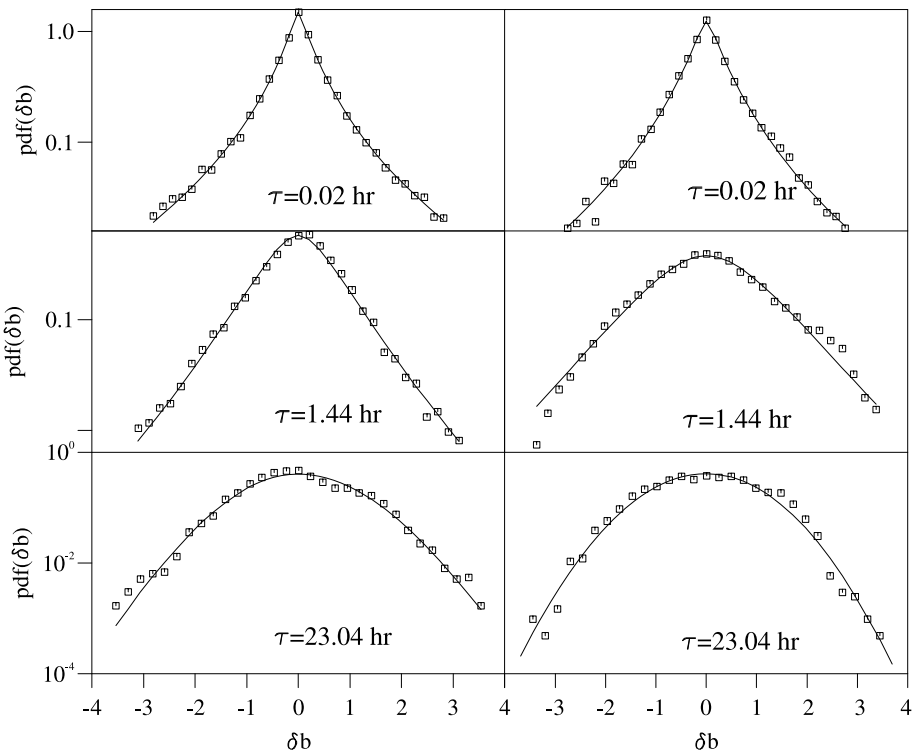

Space Plasmas, Dynamical Complexity in, Figure 10
PDF for δB calculated from the observational data in the (*left*) fast solar wind (∼ 700 km/s) and (*right*) slow solar wind (∼ 400 km/s) streams. *Solid lines* represent fits obtained with the model suggested by Castaing et al. [21]. Reprinted from [108] with permission from Elsevier

fluctuations in the auroral zone [111,112], as well as for fluctuations of other field variables, though caution must be exercised in interpreting some of these results as the Taylor hypothesis may or may not be satisfied in the observational regions.

Wavelet Transforms and Intermittency

Traditional procedure for analyzing fluctuations is to perform Fourier transforms of the observational, numerical or experimental data. The basis functions of the Fourier transform are sines and cosines. Each such basis function extends in space (or time) all the way from minus infinity to plus infinity. Thus, Fourier transform essentially wipes out the localized information of the intermittent fluctuations resulting from the confined interactions of the coherent structures. Therefore, for intermittent turbulence, it will be more appropriate to consider transforms whose basis functions are localized and scale dependent instead of sines and cosines. Such a transform has the generic name of "wavelet transform" [43,46]. Below, we restrict our discussions to one-dimensional wavelet transforms.

To be more specific, a wavelet transform replaces (transforms) the fluctuating spatial (or temporal) series into the coefficients of a set of basis functions that are localized in space (or time) with different scales in space (or time). The basis functions are usually chosen such that they are square-integrable and generated by a single mother wavelet, a rescalable and translatable function $\psi[(x-b)/a]$ where a is the scale and b defines the translation along x. Two of the more popular mother wavelets are the complex-valued Morlet wavelet [94] defined by

$$\psi_0(y) = \mathrm{e}^{ik\cdot y}\mathrm{e}^{-|y|^2/2} \tag{19}$$

and the Haar wavelet [52] given by

$$\psi_0(y) = \begin{cases} 1 & \text{if } y \in [0, 1/2); \\ -1 & \text{if } y \in [1/2, 1), \text{ and} \\ 0 & \text{if } y \notin [0, 1)\,. \end{cases} \tag{20}$$

To generate the full set of functions, we simply replace the argument in (19) or (20) by $(x-b)/a$ and allow the parameters (a, b) to vary within the domain of interest.

The Morlet wavelet is appealing to the physicists because it fits in with the usual understanding of wave packets. And due to its continuous nature, the transform is amenable to analytical manipulations. One disadvantage of the Morlet wavelets is that they are not entirely orthogonal to each other. The Haar wavelet, on the other hand, is truly local and the basis functions are real and orthogonal to each other. The simple and discrete nature of the Haar wavelets renders it ideal for numerical applications. One disadvantage of the Haar wavelets is that they are not symmetrical with respect to the midpoint, thus producing slanted spectral plots. This asymmetry, however, becomes less and less noticeable for smaller and smaller scales.

To obtain the coefficients of a wavelet transform for a particular fluctuating spatial (or temporal) series $S(x)$, we evaluate the convolution integral as follows:

$$C(a,b) = (K/a)\int S(x')\psi[(x'-b)/a]\mathrm{d}x' \tag{21}$$

where K is a normalization constant.

From $C(a,b)$, we can calculate, for example, the normalized power $\mathrm{NP}(a,b)$:

$$\mathrm{NP}(a,b) = |C(a,b)|^2 / \left\langle \left|C(a,b)^2\right| \right\rangle_x \tag{22}$$

where $\langle \ldots \rangle_x$ denotes the spatial (or temporal) average; and, here, the average is essentially the Fourier transformed power for the entirety of the fluctuations. In usual applications, only a truncated finite set of coefficients are used to perform practical calculations. The normalized power is a power spectrum of both scale and location; thus, it is sometimes called the Local Intermittency Measure, LIM. By definition, LIM is equal to 1 for the Fourier transform. Figure 11(top) is a color representation of the Local Intermittency Measure or LIM of the magnetic field intensity using the Haar transform for the 2D homogeneous MHD simulation discussed previously at a given time t and for fixed y. The scale-dependent, sporadic and localized powers of the fluctuations are vividly displayed. We note that the intermittency increases as the scale is reduced. And, the intensity can be rather strong at small scales – even when the fluctuations are strongly intermittent and localized.

Normalized power of wavelet transforms have been used to evaluate the Local Intermittency Measure LIM in the solar wind (e. g. [9,10]), for the AE index [38,39], in the magnetotail (e. g. [42,77,131]), in the cusp (e. g. [44, 139]), and in the auroral zone [112]. Figure 11 (bottom) is a color representation of the Local Intermittency Measure LIM using the Haar transform for the electrical field fluctuations in the transverse direction to the magnetic field in the auroral zone [112]. The characteristics of strong intermittency at small scales are clearly visible.

As it has been pointed out in Sect. "Non-Gaussian Probability Distribution Functions", intermittent fluctuations have non-Gaussian probability distributions especially at small scales. We may look for a quantity which provides a measure of the non-Gaussian nature of the fluctuations as a function of scale. A normalized scale dependent measure related to the fourth order moment of the coarse-grained probability distribution is called the Flatness. It indicates whether the data are more peaked or flatter relative to the Gaussian distribution. It is also called the Kurtosis. It has been suggested by Meneveau [91] that the average over space (or time) of the square of the normalized power is essentially the scale dependent Flatness of the probability distribution.

$$\text{Flatness} = \left\langle [\mathrm{NP}(a.b)]^2 \right\rangle_x \tag{23}$$

For Gaussian probability distributions, the value of the Flatness may be shown to be equal to 3. For an intermittent event, the Flatness increases without bound at small scales. The rate at which the Flatness increases with the decrease of scale is a measure of intermittency. Figure 12(top) plots the values of the Flatness calculated as a function of scale corresponding to the fluctuations discussed above for the magnetic field fluctuations for the 2D MHD simulation for given y and fixed t. We note that the value of the Flatness becomes larger than 3 and continues to increase as the scale becomes smaller and smaller.

Flatness has been evaluated for the solar wind turbulence by Bruno et al. [10,11], Fig. 12(bottom). It is one of the tools used by Bruno et al. [10] to arrive at the conclusion that solar wind turbulence is the manifestation of interacting tangled flux tubes or "cooked spaghetti" as depicted in Fig. 3(top). Wavelet transforms have been used by various authors to evaluate Flatnesses of intermittent turbulence in other regions of space plasmas, e. g. [44,77, 112,131,139]. Other more generalized techniques of gauging the intermittency in space plasma turbulence have also been employed by Vörös [129] and Vörös et al. [130,132], which we shall not discuss here.

The normalized power and Flatness are only two tools that provide gauges of intermittency of a fluctuating series. One might ask: Are there other measures such that in their totality fully characterize the intermittency nature of the fluctuating series? This will be addressed briefly in the next section.

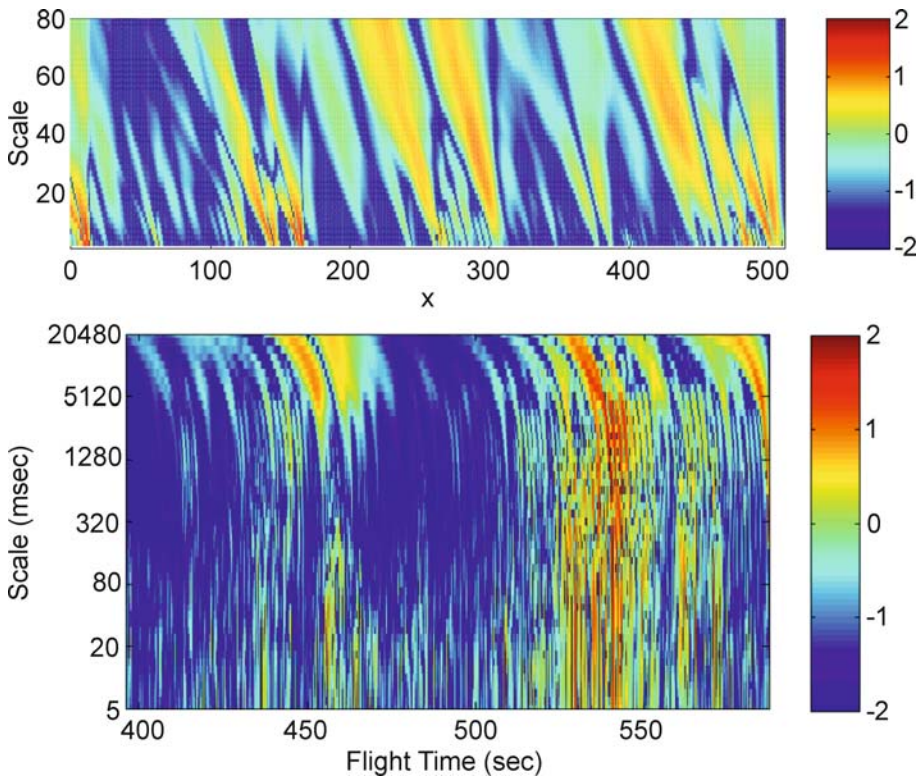

Space Plasmas, Dynamical Complexity in, Figure 11
***Top*: Normalized power or LIM for the intermittent magnetic field fluctuations *B*(*x*) using the Haar transform for a 2D homogeneous MHD simulation at given time *t* and for fixed *y*, in units of grid spacing. *Color* indicates intensity. *Bottom*: Normalized power or LIM for the intermittent transverse electric field fluctuations in the auroral zone. *Color* indicates intensity. Adapted from [112] with thanks to the American Geophysical Union**

Multifractals

Another popular modus operandi designed to study the phenomenon of intermittency is based on the concept of multifractals. As we have seen above, turbulence in space plasmas generally encompass fluctuations of all varieties and sizes, which interact and propagate throughout the plasma medium. For illustrative purposes, let us visualize some particular fluctuations that have conventional geometrical properties in a three-dimensional Euclidean space. Because of their sporadic and localized nature, it is easy to imagine that they generally cannot fill the full three-dimensional space that they occupy at a given time. Or said in another way, the space these fluctuations occupy is only a fraction of the full three-dimensional space. Such geometrical property was popularized by Mandelbrot [83] when he coined the word "fractals" or fractal geometry.

Actual fluctuations in plasma turbulence generally do not have the conventional geometrical properties. We must then devise some abstract "measure" to characterize the properties of the fluctuations and evaluate its fractal characteristics which may be interpreted with geometrical analogs. Consider, for example, the spatial series of the simulated fluctuations of the strength of the magnetic field, $B(x_i)$, along some constant value of y at time t for the two-dimensional homogeneous MHD turbulence discussed in the previous sections, where $x_i = i\delta$ with $i = 0, 1, 2, \ldots, N$ and δ is the length between grid points. From this series, we can construct a spatial series by considering, for example, the absolute value of the fluctuations due to the coarse-grained difference of the strength of the magnetic field between two spatial values $x_i + \Delta$ and x_i with $\Delta = k\delta$, some multiple of δ:

$$\delta B_i = |B(x_i + \Delta) - B(x_i)| \tag{24}$$

within the spatial interval $X = N\delta$. The simulation is statistically homogeneous over the interval X. Thus, we may calculate the ensemble average of δB_i over the interval X,

$$S_1(\delta B, \Delta) = \langle |B(x_i + \Delta) - B(x_i)| \rangle \tag{25}$$

and use it as a "measure" for the coarse-grained fluctuations of the simulated spatial series. The choice of taking

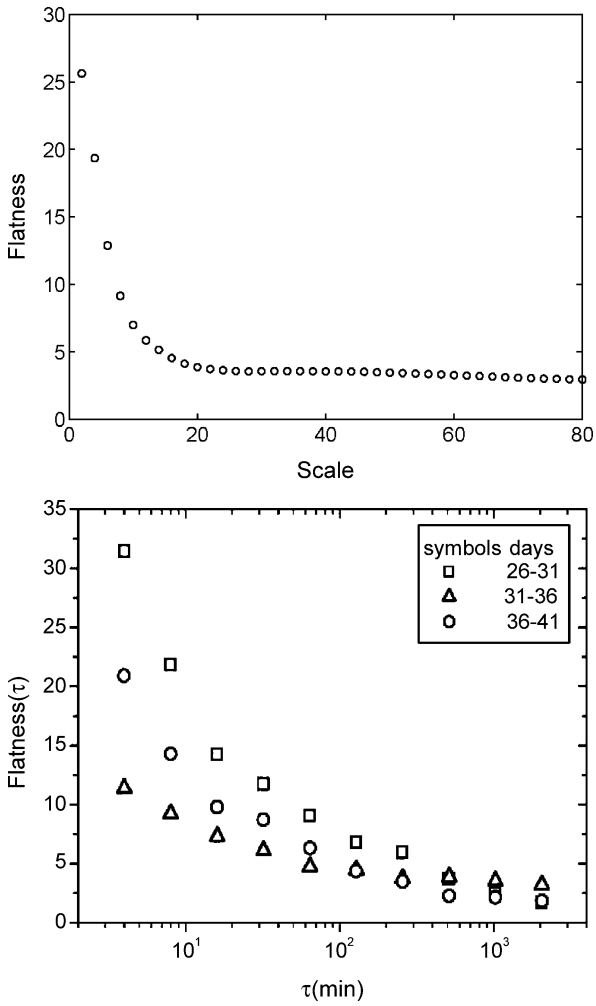

Space Plasmas, Dynamical Complexity in, Figure 12
***Top*: Flatness as a function of scale in units of grid spacing for the intermittent magnetic field fluctuations *B*(*x*) using the Haar transform for a 2D homogeneous MHD simulation at given time *t* and for fixed *y*. *Bottom*: Flatness versus time scale for three different time intervals for the solar wind during solar minimum (i. e., when the solar activity is very weak) as measured by the WIND spacecraft. Reprinted from [11] with thanks to the American Institute of Physics**

the ensemble average of the absolute values of the coarse-grained differences instead of the values of the raw differences is for the purpose of better statistical convergence [8, 127]. We can then plot $\log S_1(\delta B, \Delta)$ against $\log \Delta$ for different choices of Δ (or k). If the result approximates a straight line for some range of Δ for small Δ, we can then assign a fractal number $\zeta_1(\delta B)$ to the fluctuations of the strength of the magnetic field as:

$$\zeta_1(\delta B) = \mathrm{d}(\log S_1(\delta B, \Delta))/\,\mathrm{d}(\log \Delta) \tag{26}$$

for this range of small Δ. In other words, within this range of small Δ we may represent $S_1(\delta B, \Delta)$ as power of Δ with exponent $\zeta_1(\delta B)$. This exponent may be considered as an analog to the classical concept of dimension. It is generally an irrational number and usually cannot be surmised from simple dimensional analysis arguments; thus it is sort of a "fractal dimension" for the particular choice of "measure" described above.

Structure Functions

It is obvious that if another choice of measure is made, the corresponding fractal number for the same spatial series will generally be different. For example, one might wish to look at higher order moments of the PDFs $P(\delta B, \Delta)$ of δB, the so-called structure functions:

$$\begin{aligned} S_q(\delta B, \Delta) &= \int (\delta B)^q P(\delta B, \Delta)\,\mathrm{d}(\delta B) \\ &= \left\langle |B(x_i + \Delta) - B(x_i)|^q \right\rangle \end{aligned} \tag{27}$$

The motivation here is that different moments emphasize different peaks in the fluctuating series. Generally, corresponding to each ζ_q there will be a fractal exponent ζ_q for small Δ. If $\zeta_q = \zeta_1 q$, then the fractal property of the fluctuating series is fully characterized by the value of ζ_1. Such fluctuations are then said to be "monofractal" or "self-similar", i. e., the fractal characteristics for all the moment orders are similar to each other. For general intermittent turbulence, on the other hand, ζ_q would be a nonlinear function of q, or "multifractal".

In practical calculations, it has been suggested by Benzi et al. [6] for hydrodynamic turbulence studies that the correlations among the structure functions seemed to be much more apparent than those exhibited by the correlations of individual structure functions directly with coarse-grained scaling. This is particularly true for structure functions based on the absolute values of the differences of fluctuations as it is defined here [51]. This also seemed to be true for plasma turbulence studies [19,20,95, 100]. Thus, it has become very popular to first determine the relative values of the structure function exponents with respect to the structure function exponent of a particular order, and then determine the exponent for that particular order more accurately by some phenomenological, dimensional, or empirical reasoning. This procedure is called "extended similarity scaling" (ESS). Figure 13 is the result of such a statistical calculation for the fluctuations of the magnetic field strength $B(x_i)$ of the 2D homogeneous MHD simulation for given y and t. We note that the structure function exponents exhibit nonlinear, i. e., multifractal, behavior as expected for intermittent turbulence.

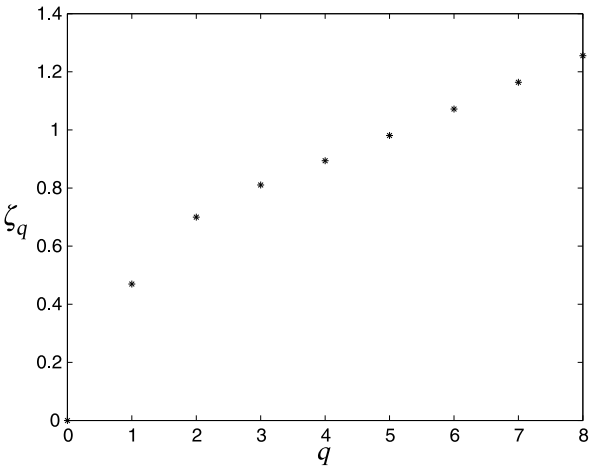

Space Plasmas, Dynamical Complexity in, Figure 13
Structure function exponents for the magnetic field intensity at given time *t* and for fixed *y* using ESS for a 2D MHD simulation for homogeneous intermittent turbulence. The values of the exponents at high orders are less accurate than those for the low orders due to the size limitation of the simulation box. Reprinted from [27] with thanks to the American Institute of Physics

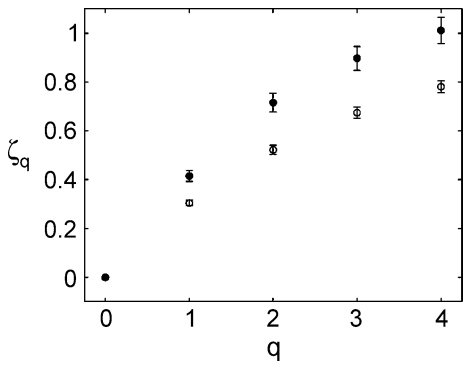

Space Plasmas, Dynamical Complexity in, Figure 14
Structure function exponents using ESS for the time series of the strength of the magnetic field collected for extended periods during solar minimum (low solar activity) by the WIND spacecraft (*circles*) and solar maximum (high solar activity) by the ACE spacecraft (*filled circles*). Adapted from [11] with thanks to the American Institute of Physics

Calculations for structure function exponents have been carried out for magnetic and velocity fluctuations in the solar wind, e. g. [11,12,13,14,17,58,59,60,86,97,121] and elsewhere in the space environment, e. g. [57], usually by assuming the applicability of the Taylor hypothesis. We caution here again that except for most of the solar wind studies, the assumption of the Taylor hypothesis must be exercised with a critical eye. In Fig. 14, the results for such calculations for the time series of the strength of the magnetic field collected for extended periods during solar minimum (low activity) and solar maximum (high activity) by two separate spacecraft (Wind and ACE) are presented. The multifractal nature of the intermittent turbulence is apparent in both calculated results [11].

Structure function calculations may be performed conveniently for a fluctuating series for positive values of q; but will generally diverge for $q < 0$. An alternative procedure to evaluate the fractal characteristics of different moment orders is the so-called "singularity analysis" described below.

Partition Functions, Generalized Dimensions and Singularity Spectra

As we have seen in the previous discussions, turbulence in space plasmas is generally intermittent and therefore probably composed of an admixture of fluctuations of different fractal characteristics. We may attempt to extract this "multifractal" nature of the observed fluctuating time series by searching for the "dominant singular behavior" for different moment orders at small scales following the technique suggested by Parisi and Frisch [98] and Halsey et al. [54] and first applied to solar wind turbulence by Burlaga [12,13,14,15], and later for the auroral electrojet (AE) index by Consolini [41], for the high latitude geomagnetic fluctuations by Vörös [129,130], for the plasma sheet by Lui [76] and Weygand et al. [134], to the magnetospheric cusp by Yordanova et al. [139], and to the auroral electric field fluctuations by Chang et al. [34].

The basic idea here is to define an appropriate measure that is scale dependent. (Instead of referring to a spatial series such as that considered above for the numerical simulation, for variance, we consider below a generic time series based on temporal observations). To proceed, we first define an "incremental measure":

$$\delta\mu_j = \left|B(t_j + \delta) - B(t_j)\right| \tag{28}$$

where δ is the sampling interval. We now subdivide the total time interval T into $M = T/\tau$ segments with $\tau = k\delta$ and calculate the normalized scale-dependent "segmental measure"

$$\mu_i(\tau) = \sum_{j=(i-1)k+1}^{ik} \delta\mu_j/\mu \quad \text{with} \quad \mu = \sum_{j=1}^{N} \delta\mu_j \tag{29}$$

For convenience, we will generally choose M to be an integer.

The normalization of $\mu_i(\tau)$ is an attempt to represent the segmental measure as a measure of probability. This

is the scale dependent measure that we seek to define. We shall assume that each such measure varies with the scale τ in a singular manner as a power law. We now form the qth moment order of the coarse-grained probabilities $\mu_i(\tau)$, traditionally called the "partition function" [54]:

$$\Gamma(q,\tau) = \sum_{i=1}^{M} \mu_i^q(\tau) \tag{30}$$

where q can be any real number. We shall now search for the dominant singular behavior of $\Gamma(q,\tau)$ as characterized by a power law in τ with exponent $\gamma(q)$ for small τ similar to that for the structure function analysis. In general, for each moment order, $\gamma(q)$ is a different number characterizing the particular fractal behavior of the subset of fluctuations, which dominates the (singular) scaling behavior of that particular moment order.

As mentioned above, Consolini et al. [41] have carried out such a singularity measure multifractal analysis for a typical time series of the auroral electrojet (AE) index, Fig. 15 (top left panel). In their calculations, the incremental measure was chosen as:

$$\delta\mu_j = \left|\phi_{\mathrm{AE}}(t_j+\delta) - \phi_{\mathrm{AE}}(t_j)\right|^2 \tag{31}$$

where $\phi_{\mathrm{AE}}(t_j)$ is the measured AE index at time t_j and δ is the sampling interval. We note from the figure that $\gamma(q)$ is indeed a nonlinear function of q, and thus, a "multifractal". There are two distinct asymptotic regimes of q ($q < -2$, and $q > +2$) for which $\gamma(q)$ varies with q nearly linearly and a nonlinear crossover region between the two asymptotes.

Henschel and Procaccia [55] suggested the concept of the so-called "generalized dimensions" for fractals. Their definition of the generalized dimension D_q is related to the partition function $\Gamma(q)$ as follows:

$$D_q = (q-1)^{-1} \lim_{\tau\to 0} \log\Gamma(q)/\log\tau = \gamma(q)/(q-1). \tag{32}$$

It is called the generalized dimension because in the limit of $q \to 0$, D_0 is essentially the fractal (or similarity) dimension defined by Mandelbrot [83]. Also, it may be shown that, in the limit of $q \to 1$, D_1 corresponds to the so-called information dimension [5]. In addition, D_q's for $q \geq 2$ are the corresponding qth correlation dimensions [50]. The calculated generalized dimension D_q as a function of moment order q for the AE index is shown in Fig. 15 (top right panel). As shown by Consolini et al. [41], such a nonlinear variation may be modeled rather accurately by the so-called p-model which is a "two-scale cascade model" as follows [16,92]:

$$D_q = \log_2[p^q + (1-p)^q]^{1/(1-q)} \quad \text{with} \quad p = 0.746 \pm 0.002. \tag{33}$$

where p is the single fitting parameter associated with the fragmentation probability of the cascade process. This one-parameter representation of the generalized dimension for the multifractal phenomenon of the AE index reinforces the idea stated earlier that the multifractal phenomenon can be thought of as the crossover phenomenon between two invariant regions.

Another way to gauge the multifractal characteristics in terms of the partition function formalism is the method of the singularity spectrum $f(\alpha)$ [54,98]. If we associate with each segmental measure μ_i a singularity index α_i such that $\mu_i = \tau^{\alpha_i}$ where α_i is within some range between α and $\alpha + \Delta\alpha$ and pass on to the continuum limit by writing the number of differential measures whose singularity index is between α and $\alpha + \mathrm{d}\alpha$ as $\rho(\alpha)\tau^{-f(\alpha)}\mathrm{d}\alpha$, then the partition function as defined in (30) in the continuum limit is:

$$\Gamma(q,\tau) = \int \mathrm{d}\alpha' \rho(\alpha')\tau^{-f(\alpha')+q\alpha'}. \tag{34}$$

Since we are interested in the phenomenon for very small τ, the integral is dominated by the value of α' which makes the expression $-f(\alpha') + q\alpha'$ the smallest. Assuming $\rho(\alpha') \neq 0$ and $f(\alpha')$ smooth, the minimum is located at $\alpha' = \alpha(q)$ with

$$f'(\alpha(q)) = q \quad \text{and} \quad f''(\alpha(q)) < 0 \tag{35}$$

Therefore the partition function for very small τ is approximately

$$\Gamma(q,\tau) \sim \exp\left\{[q\alpha(q) - f(\alpha(q))]\ln\tau\right\} \tag{36}$$

This leads immediately to the identification of

$$\gamma(q) = D_q(1-q) = [q\alpha(q) - f(\alpha(q))] \tag{37}$$

Differentiating (37) with respect to q and using the first expression of (35), we obtain

$$\alpha(q) = \gamma'(q) \equiv \mathrm{d}\gamma/\mathrm{d}q \tag{38}$$

Thus, given $\gamma(q)$, we may determine $\alpha(q)$ and then, $f(\alpha)$. Such a singularity spectrum for the AE index considered by Consolini et al. [41] is given in Fig. 15 (bottom panel). Because $\alpha(q)$ is intimately related to $\gamma(q)$ (and therefore

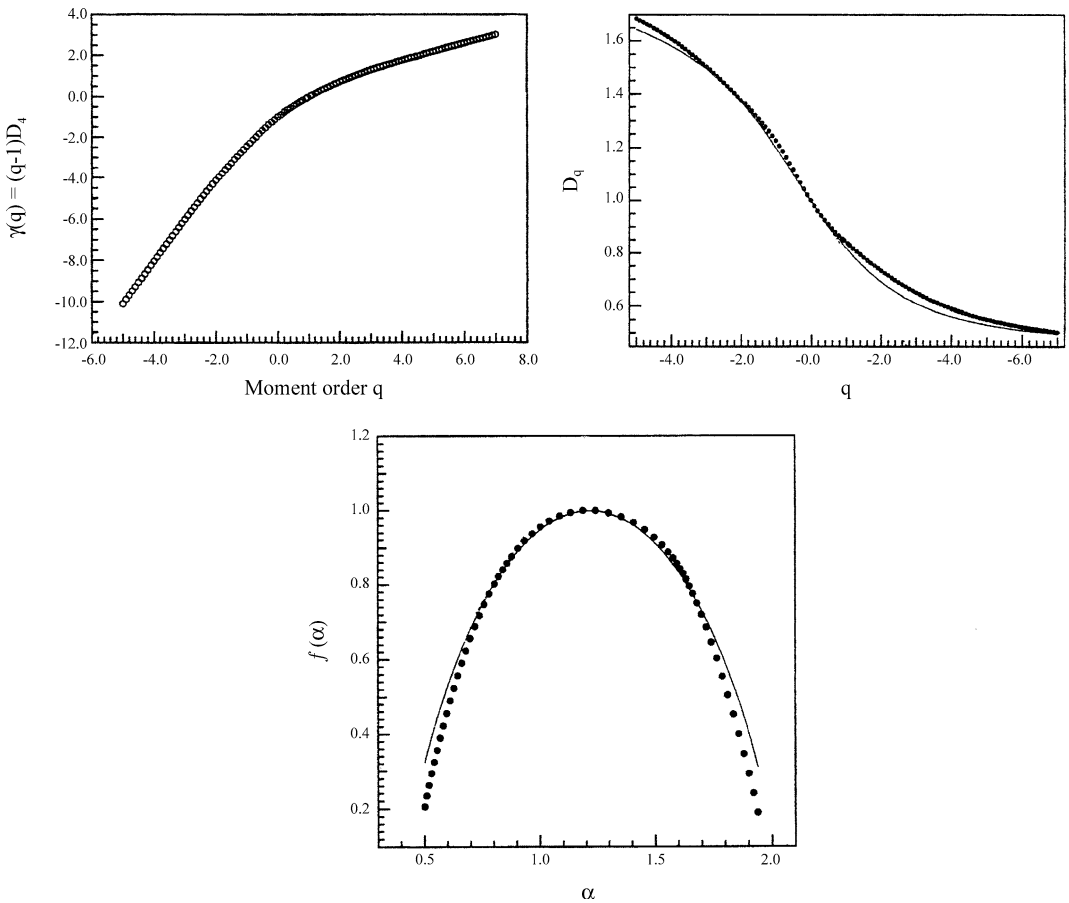

Space Plasmas, Dynamical Complexity in, Figure 15
***Top left*: Partition function exponent $\gamma(q)$ for a time series of the AE (auroral electrojet) index. *Top right*: Generalized dimension D_q for the same AE time series. Solid line is the best fit using the *P*-model. *Bottom*: Singularity spectrum $f(\alpha)$ for the same AE time series obtained from the Legendre transform. Adapted from [41] with thanks to the American Physical Society**

also to the partition function $\Gamma(q)$), $f(\alpha)$ sorts out the spectrum of singularities of the intermittent fluctuations, that may be mapped to the multifractal dimensions D_q as defined by (32).

In the above, we have introduced a sequence of algebraic manipulations mainly for the purpose of connecting three different popular approaches of identifying the singularity nature of the multifractal characteristics of a time (or spatial) series based on the partition function formulation. We must emphasize that these are merely different approaches of viewing fracture characteristics of intermittent fluctuations with no additional physics involved, except perhaps with the following useful thermodynamic analogy. That is, the relationship between $(\gamma(q), q)$ – or equivalently (D_q, q) – and $(f(\alpha), \alpha)$ is formally a Legendre transform. Thus, one may easily associate the relationships among such measures with the standard notion of Legendre transforms of classical thermodynamics [96]. Such a connection may become useful in identifying analogies between singularity measures of multifractals with multicriticality measures in thermodynamics.

It is interesting to note that in multifractal analyzes of space plasma turbulence in regions where intermittency are expected, similar functional dependence of $\gamma(q)$ on q as above is also generally obtained [12,13,14,15,18,34,87,134,139]. Although, we must recall that for turbulence in the plasma sheet or the auroral zone and even sometimes in the solar wind, the assumption of the validity of the Taylor hypothesis must be applied with caution.

In the previous section, it was demonstrated that intermittency characteristics of a fluctuating series could be brought out using the technique of wavelet transforms. Certain quantities, Local Intermittency Measure LIM or Normalized Power and Flatness, were introduced using

the wavelet transforms to gauge the intermittency effects. Since the complete set of coefficients of the wavelet transforms essentially represents the full fluctuating event, it is reasonable to inquire if it can be used to advantage to describe the full characteristics of intermittency of a fluctuating series. The answer is a definite *yes*. For example, wavelet transforms have been used to study multifractal characteristics of time series related to intermittent turbulence in space plasmas by Yordonova et al. [139] for the magnetospheric cusp and Vörös et al. [131] for the plasma sheet.

Rank-Ordered Multifractal Analysis (ROMA)

In the above, intermittent fluctuations were analyzed using the structure function and/or partition function methods based on the statistics of the full set of fluctuations. Since most of the observed or simulated intermittent fluctuations are dominated by fluctuations with small amplitudes, the subdominant fractal characteristics of the minority fluctuations – generally of larger amplitudes – are easily masked by those characterized by the dominant population. It therefore appears useful to search for a procedure that explores the fractal nature of the subdominant fluctuations by first appropriately isolating out the minority populations and then perform statistical investigations for each of the isolated populations.

Phrasing it differently, for intermittent fluctuations exhibiting multifractal characteristics, we visualize that these fluctuations are composed of many types, each type being characterized by a particular fractal dimension. The questions to ask then are (i) what are the different fractal dimensions? and (ii) how are the various types or classes of fluctuations distributed within the turbulent medium? We answer these questions quantitatively by constructing a "rank-ordered multifractal spectrum".

Using the example of a 2D MHD simulation that was discussed in Sect. "Non-Gaussian Probability Distribution Functions", we demonstrate below how this idea may be achieved with a rank-order method that subdivides the fluctuations into groupings based on the range of the scaled-sizes of the fluctuations [28]. In the simulation, ideal compressible MHD equations expressed in conservative forms are solved numerically with 1024×1024 grid points in a doubly periodic (x, z) domain of length 2π in both directions using the WENO code [64] so that the total mass, energy, magnetic fluxes and momenta are conserved. The initial condition consists of random magnetic field and velocity with a constant total pressure for a high beta plasma. As discussed previously, after sufficient elapsed time, the system evolves into a set of randomly interacting multiscale coherent structures exhibiting classical aspects of intermittent fluctuations.

We recall that the PDFs $P(\delta B^2, \Delta)$ of the square of the magnitude of the magnetic field δB^2 for such intermittent fluctuations were non-Gaussian and became more and more heavy-tailed at smaller and smaller scales (Fig. 8). An attempt to collapse the unscaled PDFs according to the one-parameter scaling formula (12) re-displayed below as (39):

$$P(\delta B^2, \delta)\delta^s = P_s(\delta B^2/\delta^s) \tag{39}$$

indicated approximate scaling with an estimated scaling exponent of $s = 0.335$ (Fig. 9). The scaling formula (39) may be interpreted from the concept of fractals. We begin the interpretation by assuming that δB^2 and P vary with δ as power laws: $\delta B^2/\delta^a = I$ and $P/\delta^b = J$, where (a, b) are the power-law (fractal) exponents and (I, J) are constants – i. e. invariants with respect to the scale δ. If the form of $P(\delta B^2, \delta)$ is invariant as the scale changes, then it has been shown that a functional relation exists between the two invariants, (I, J) [29]. Imposing the normalization condition for the PDFs, we obtain the one-parameter scaling form (39) with $s = a = -b$.

Such a phenomenon is self-similar and the PDFs collapse onto one scaled PDF $P_s(Y)$ where $Y = \delta B^2/\delta^s$ is a scale invariant. In such a case, the structure function exponents for the PDFs will satisfy the monofractal property of $\zeta_q = sq$. The scaling exponent s may be interpreted as a single fractal measure that characterizes the fluctuations of all scales through the scaling relation (39). If the PDFs are self-similar Gaussian distributions for all scales, representing random diffusion, the scaling exponent s is equal to 0.5. For other monofractal distributions, the scaling exponent may take on any positive real value.

For the current example, structure function calculations based on the full set of simulated fluctuations showed a nonlinear relation between the exponents and the moment order, indicating the phenomenon is multifractal in nature [27]. However, the physical interpretation of the multifractal nature is not easily deciphered by merely examining the curvature of the deviation from linearity.

We shall therefore attempt to perform statistical analyzes individually for subsets of the fluctuations that characterize the various fractal behaviors within the full multifractal set. Such grouping of fluctuations must depend somehow on the sizes of the fluctuations. However, the groupings cannot depend merely on the raw values of the sizes of the fluctuations because the ranges will be different for different scales. Thus, we proceed to rank-order the sizes of the fluctuations based on the ranges of the scaled variable $Y = \left|\delta B^2\right| / \Delta^s$ where s is the scaling exponent for

each grouping defined in (39). (Absolute values of the fluctuations are used to take advantage of the symmetry property of the PDFs and for the purpose of better statistical convergence.) For each chosen range we shall assume that the fluctuations of all scales will exhibit monofractal behavior satisfying the form of (39) and be characterized by an scaling exponent s. The question is then how can this procedure be accomplished.

We begin by considering a differential range of scales dY in the vicinity of some scaled size $Y = \left|\delta B^2\right| / \Delta^s$. Fluctuations whose sizes fall within such a differential range will probably exhibit monofractal behavior characterized by some scaling exponent s such that the differential structure function $\mathrm{d}S_q$ will vary with the scale as Δ^{sq}, i. e.,

$$\mathrm{d}S_q \triangleq (|\delta B^2|)^q P(|\delta B^2|, \Delta)\,\mathrm{d}(|\delta B^2|) = \Delta^{sq} Y^q P_s(Y)\mathrm{d}Y \tag{40}$$

Given a set of PDFs $P(|\delta B^2|, \Delta)$, the corresponding multifractal spectrum $s(Y)$ may be obtained approximately (if the ansatz is valid) by integrating the functional differential expression (40) over small contiguous ranges of ΔY with the assumption that within each incremental range the scaling exponent s is essentially a constant [28]. Thus, for a generic range of ΔY within (Y_1, Y_2), we form such a range-limited structure function as follows:

$$\begin{aligned}\Delta S_q(\left|\delta B^2\right|, \Delta) &= \int_{a_1}^{a_2} \left|\delta B^2\right|^q P(\left|\delta B^2\right|, \Delta)\,\mathrm{d}\left|\delta B^2\right| \\ &\simeq \Delta^{sq} \int_{Y_1}^{Y_2} Y^q P_s(Y)\mathrm{d}Y \end{aligned} \tag{41}$$

where $a_1 = Y_1 \Delta^s$ and $a_2 = Y_2 \Delta^s$. We then search the value of s such that the scaling property of the range-limited structure function that varies with s is $\Delta S_q(s) \sim \Delta^{sq}$. If such a value of s exists, then we have found one region of the multifractal spectrum of the fluctuations such that the PDFs in the range of ΔY collapses onto one scaled PDF. Performing this procedure for all contiguous ranges of ΔY produces the rank-ordered multifractal spectrum $s(Y)$ that we are looking for. The determined value of s for each grouping should be un-affected by the statistics of other subsets of fluctuations that are not within the chosen range ΔY and therefore should be quantitatively quite accurate. The physical meaning of this spectrum is that the PDFs for all time lags collapse onto one master multifractal scaled PDF. The spectrum is implicit since Y is a defined as a function of s.

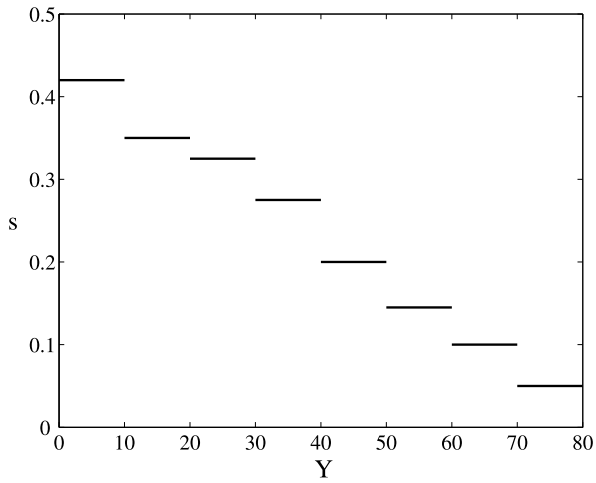

Space Plasmas, Dynamical Complexity in, Figure 16
Rank-ordered multifractal spectrum, $s(Y)$. $Y = |\delta B^2|/\Delta^s$; $|\delta B^2|$ in units of bin size and Δ in grid spacing. The spectrum is calculated for 8 contiguous ranges of ΔY. Reproduced from [28] with thanks to the American Physical Society

Interestingly, for this particular example, there exists one and only one value of s in each range of ΔY, that satisfies the above constraint. Unlike the structure functions defined for the full range of fluctuations, the range-limited structure functions based on (41) exists also for negative real values of q (except for the range including $Y = 0$). Figure 16 displays the calculated rank-ordered spectrum $s(Y)$ based on eight contiguous ranges of ΔY [28]. It is noted that the spectrum has values of s ranging between 0.5 and 0.0. The spectrum can be refined by choosing more range intervals with smaller range sizes of ΔY, although in practice this procedure is limited by the availability of simulated data points. At $Y = 0$, the scaling exponent s appears to approach the self-similar Gaussian value of 0.5 representing random diffusion. As the value of the scaled fluctuation size Y increases the scaling exponent decreases accordingly indicating the fluctuations are becoming less and less space filling. At very large value of Y, the scaling exponent s seems to asymptotically approach the value of zero indicating the fluctuations have become extremely sparse.

Such an implicit multifractal spectrum has several advantages over the results obtainable using the conventional structure function calculations. Firstly, the utility of the spectrum is to fully collapse the unscaled PDFs. Secondly, the physical interpretation is clear. It indicates how space-filling (in terms of the value of s) are the scaled fluctuations once the value of Y is given. Thirdly, the determination of the values of the fractal nature of the grouped fluctuations

is not affected by the statistics of other fluctuations that do not exhibit the same fractal characteristics. Fourthly, it provides a natural connection between the one-parameter scaling idea (39) and the multifractal behavior of intermittency.

Invariant Scaling

Forced and/or Self-Organized Criticality (FSOC)

In the above sections, we provided some convincing arguments as well as numerical and observational evidences indicating that space plasma turbulence is generally in a state of dynamic "topological complexity". By "complex" topological states we mean magnetic topologies that are not immediately deducible from the basic (e. g., MHD) equations. We also discovered that the structure/partition functions for turbulence generally scale as power laws at least at small scales. Expressed another way, we found that there were power-law scale invariants with respect to coarse-graining at small scales where intermittency prevails. Below, we shall briefly address the salient features of such "topological phase transitions" as well as the associated scale invariant properties in turbulent plasmas. (We note here that the concept of phase transitions has also been employed to address the phenomenon of complexity in space plasmas by Sitnov et al., [106] using the singular spectrum method.)

For nonlinear stochastic systems exhibiting complexity, the correlations among the fluctuations of the random dynamical fields (electric, magnetic and velocity fields, etc.) are generally very long-ranged and there exist many correlation scales (as exhibited by the various outsized plasma coherent structures). The dynamics of such systems are notoriously difficult to handle either analytically or numerically. On the other hand, since the correlations are extremely long-ranged, it is reasonable to expect that the system will exhibit some sort of invariance property under coarse-graining scale transformations. A powerful analytical technique that exploits this invariance property is the method of the dynamic renormalization group (DRG) which we shall briefly describe below. The technique is a generalization of the static renormalization group introduced by Wilson in 1972 [135] for equilibrium phase transitions. Dynamic renormalization group was originally developed as a perturbation theory by Halperin et al. [53] and first applied to hydrodynamic problems by Forster et al. [49]. A closed form exact theory of DRG was developed by Chang et al. [30]. And a review of the basic elements of this theory may be found in Sect. "Non-Gaussian Probability Distribution Functions" of a Physics Report by Chang et al. [31].

No matter how complex the stochastic behavior is for the turbulent system, we may conjecture that the system is somehow characterized by a set of dynamically evolving parameters $\{P_n\}$. We now transform the stochastic system by coarse-graining and ask how these parameters transform accordingly (subject to the rules stipulated by the underlying dynamical equations, e. g., those for the interacting flux tubes). Symbolically, we write

$$\partial\{P_n\}\partial\ell = R\{P_n\} \tag{42}$$

where R represents the coarse-graining (renormalization-group) transformation operator and ℓ *is* the coarse-graining parameter. The operator is generally very complicated for even relatively simple stochastic systems. Instead of actually delving into the mathematical formalism of DRG, we shall merely explore some of the basic concepts related to such ideas.

Forced and/or Self-organized Criticality Generally, if one starts at some initial state $\{P_n(0)\}$ and proceeds with the coarse-graining procedure following the prescription dictated by (42), the result would be a trajectory in the phase space characterized by the parameters. In analogy to nonlinear dynamics, we expect that there may be fixed points (singular points) in the phase space flow field, at which $d\{P_n\}/d\ell = 0$. At such a point, e. g., $\{P_n^*\}$, the correlation length should not be changing. However, the coarse-graining transformation requires that all length scales to change with ℓ. Therefore, to satisfy both requirements, the correlation length must be either infinite or zero at the fixed point. When it is at infinity, the state of the dynamical system would then be at a state of dynamic criticality, analogous to the state of criticality in equilibrium phase transitions.

Generally, a dynamical system will not be near a parametric state such as $\{P_n^*\}$. But, if the coarse-graining procedure (i. e., the renormalization group trajectory) leads it eventually to a point in the parametric space close to a fixed point such as $\{P_n^*\}$, then in a coarse-grained sense the stochastic system would be close to a state of dynamic criticality. We may then approximate the coarse-graining transformation (42) by linearizing the operator R. The result is a set of coupled linear differential equations characterizing the variations of the deviations of the parameters $\{P_n\}$ from $\{P_n^*\}$.

$$\mathrm{d}\{Q_n\}/\mathrm{d}\ell = R_{\mathrm{L}}\{Q_n\} \tag{43}$$

where $Q_n = P_n - P_n^*$ and R_{L} is the linearized renormalization-group matrix operator.

We assume that expression (43) may be expressed in terms of a set of eigenvalue equations:

$$\mathrm{d}\mu_i/\mathrm{d}l = \lambda_i \mu_i \tag{44}$$

where the eigenvectors μ_i's are linear combinations of the parameters Q_n and the λ_i's are the corresponding eigenvalues. We shall assume that the eigenvectors are complete and the eigenvalues are all real. (Nondiagonalizable operator R_{L} may render the eigenvectors incomplete and lead to logarithmic corrections to the discussion below, which will not be considered here).

The solutions of (44) are simply $\mu_i = \mu_i^0 \exp \lambda_i \ell$ where $\{\mu_i^0\}$ represents some arbitrary state that is close to $\{P_n^*\}$. From this, it can be verified immediately that

$$\mu_i/\mu_j^{\lambda_i/\lambda_j} = \text{constants} \tag{45}$$

i. e., the μ_i's form power law invariants which are independent of the coarse-graining scale transformation ℓ.

Expressing this result in words: The mathematical consequence of the linear approximation is that, close to dynamic criticality, certain linear combinations of the perturbations of the parameters that characterize the stochastic phenomenon will correlate with each other in the form of power laws. These include the (k, ω), i. e. mode number and frequency, spectra of the correlations of the various fluctuations of the dynamic field variables in space and time, etc. The fact that the dynamical (stochastic) system must initiate from some state so that the coarse-graining trajectory will carry it to the close proximity of a fixed point such as $\{P_n*\}$ means that there may need some initial tuning of the state for the system to have such behavior. For this reason, we shall refer to such dynamical systems as systems near Forced and/or Self-organized Criticality (FSOC) [22]. The phenomenon of FSOC is spatiotemporal and the invariant scaling idea permeates throughout the time-space domain. In addition, the scaling between time and space generally is not isotropic.

Such power law behavior has been detected in probability distributions related to the solar flares [74,126], Fig. 17, in the AE burst occurrences as a function of the AE burst strength [37], Fig. 18, in the global auroral UVI imagery of the statistics of size and energy dissipated by the magnetospheric system [81], in the probability distributions of spatiotemporal magnetospheric disturbances as seen in the UVI images of the nighttime ionosphere [124], in the probability distributions of durations of Bursty Bulk Flows [2], and in the magnetic field fluctuations of the Earth's magnetotail [61,93]. Similar behavior has also been obtained in numerical simulations for the distributions for solar flares [45], for dynamical models for the current sheet [66,110,122] and other applications. This invariant behavior had been called self-organized criticality (SOC) [3], i. e., tuning is generally not necessary.

An eigenvector (eigen-parameter) μ_i is "irrelevant" if the corresponding eigenvalue λ_i is less than zero because in the limit of large ℓ the value of μ_i will become negligible. The reverse is true for positive λ_i's. We shall not consider the special case of $\lambda_i = 0$ which introduces singularities of fractional powers of logarithms. If there are only a small number of relevant eigenvalues with $\lambda_i > 0$, the corresponding FSOC system has only a small number of "relevant parameters" that are needed to characterize the stochastic state, i. e., the system may appear to have low dimensionality in parameter space near criticality [22].

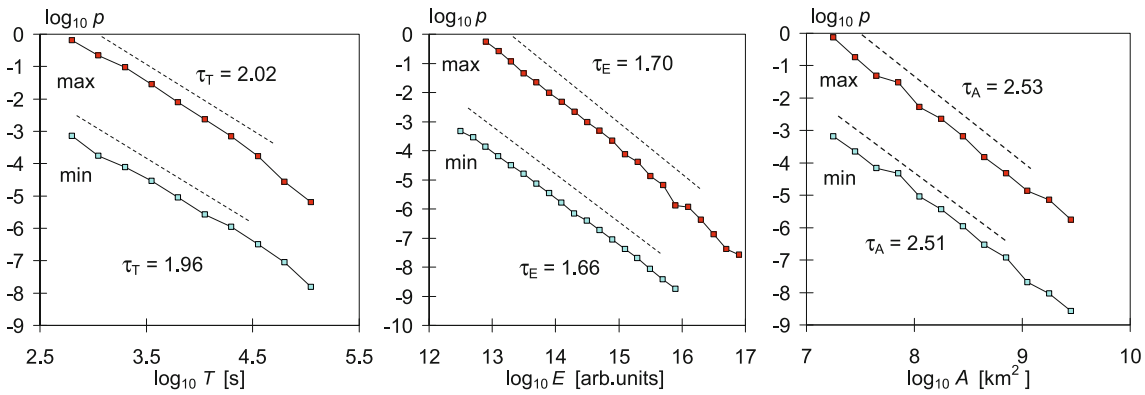

Space Plasmas, Dynamical Complexity in, Figure 17

Flare probability distributions of avalanche lifetime (*left*), emission flux (*middle*) and peak area (*right*) for solar max (high activity) and min (low activity) obtained for emissions detected by the EIT (extreme ultraviolet imaging telescope) aboard the SOHO spacecraft in the 195 Å wavelength above a threshold of 40% of the average emission. The min distributions are shifted downward for comparison. Adapted from [126] with thanks to the American Physical Society

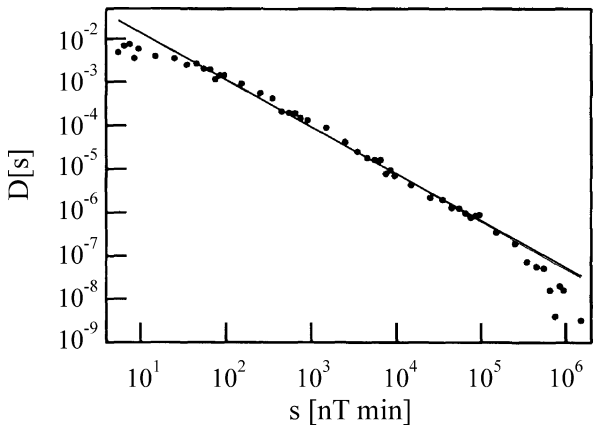

Space Plasmas, Dynamical Complexity in, Figure 18
Distribution $D(s)$ of the burst size $s = \int_{\Omega} [\mathrm{AE}(t) - L]dt$ of an AE index time series, where the integration is performed over the time interval Ω when the quiet level of $L = (45 \pm 15)nT$ is exceeded. The solid line is a power-law best fit. Reprinted from [37] with permission from the Italian Society of Physics

From Criticality to Intermittency

We now ask a deeper question: How does the invariant behavior of the structure or partition functions, which changes with moment order for intermittent turbulence amalgamate (reconcile) with the concept of criticality (SOC or FSOC) outlined above?

For systems near criticality the correlation functions are generally related to the correlation time and space as power laws. Since the structure functions are related to the spatial correlation functions [7], it is therefore not surprising that for systems near SOC/FSOC, the structure function exponents for intermittent turbulence are power law scale-invariants in coarse-graining scale transformations.

For a stochastic process that is monofractal, the structure function exponents vary with the moment order linearly. Thus a single invariant, such as $I_1 = S_1/\Delta^{\zeta_1}$ would suffice to characterize the behavior of fully developed turbulence that is self-similar. For example, it was demonstrated in Subsect. "Rank-Ordered Multifractal Analysis (ROMA)", that if the probability distributions functions (PDFs) satisfy the one-parameter scaling law of (39), the structure function exponents $\zeta_q = sq$ where s is the scaling exponent, and the PDFs collapse onto one master scaling function $P_s(\delta B^2/\Delta^s)$. This is approximately satisfied by the PDFs for the numerical simulation results discussed in Sect. "Non-Gaussian Probability Distribution Functions" [33] (Fig. 9 and re-displayed here as Fig. 19, middle panel). Similar results have been found by Hnat et al. [56] for their study of the PDFs of solar wind turbulence as seen by the WIND spacecraft and Tam et al. [112] for the electric field intensity fluctuations in the auroral zone as detected by the SIERRA rocket experiment (Fig. 19, top and bottom panels). However, it is to be noted that in all three scaling plots the deviations from the mean become larger for larger fluctuations, which are indications that multifractal characteristics will still need to be addressed. This was demonstrated in terms of the rank-ordered multifractal spectrum as discussed in Subsect. "Rank-Ordered Multifractal Analysis (ROMA)".

We note that generally – e. g., Fig. 8 – the portions of the PDFs for small fluctuations for most of the observed and simulated intermittent turbulence have shapes very close to that of a Gaussian. This is the reason why the calculated exponent s for small scaled fluctuations is close to the value of 0.5 (characteristic of self-similar random diffusion). In the numerical example considered in Subsec. "Plasma Resonances and Coherent Structures", $s \approx 0.42$ for the range of ΔY between $(0, 10)$. And the behavior of the fluctuations within this range was shown to be essentially monofractal. This would have given the value of the first structure function exponent $\zeta_1 = 0.42$ and the related power spectrum index would therefore be approximately $1 + \zeta_2 = 1 + 2\zeta_1 \approx 1.84$ (Frisch, 1995). If no attempt were made in rank-ordering and the value of s were determined for the full set of fluctuations by looking for power-law behavior of $P(0, \Delta) \sim \Delta^s$, the best value was found to be approximately 0.335. Assuming approximate monofractal scaling, this number would give a power spectrum index of approximately 1.67. Both numbers are not very different from that obtainable using classical arguments for hydrodynamic or MHD turbulence.

Most observational or numerical "critical exponents" of SOC are obtained either using first order box counting or from calculations of density, optical, and other field intensities such as electric, magnetic, velocity, and energy fields, which generally involve statistical averages over the zeroth, first, or second (i. e. low) order moments of their related PDFs and therefore each represents an approximate monofractal behavior of some sort of weighted average of the fractal values for small scaled fluctuations.

In fact, if the experimental or numerical evaluations are based on rough estimates, the values of the exponents will generally take on essentially the Gaussian – i. e., mean-field – values. And such observed "critical exponents" would satisfy the standard scaling relations. We suggest that this is the reason why recent observations [126] and numerical simulations [69] seem to indicate the coexistence of the phenomenon of SOC and multifractal signatures for intermittent turbulence. See, e. g. Figs. 17 and 20.

Thus, based on the example of rank-order procedure of the multifractal behavior of intermittent fluctuations

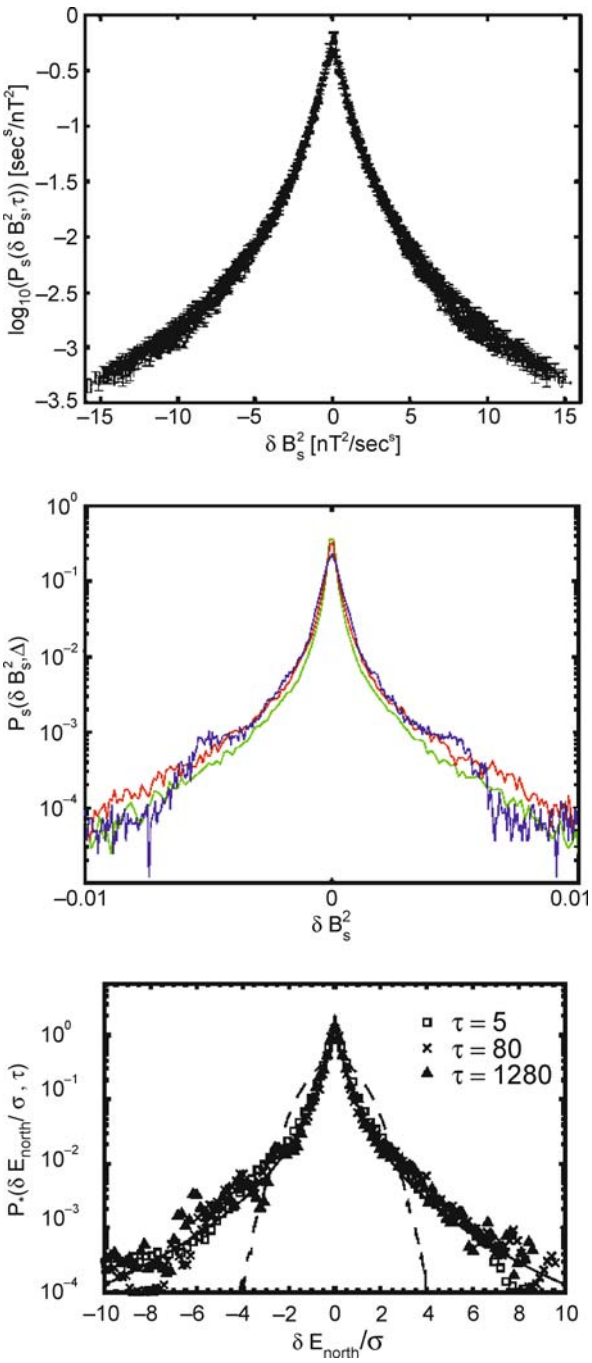

◄ **Space Plasmas, Dynamical Complexity in, Figure 19**
***Top*: Scaled PDF based on the one-parameter scaling form of (39) with $\zeta_1 = s \approx 0.42$ for the fluctuations of the magnetic field energy density as seen by the WIND spacecraft. The data shown correspond to τ between 46 seconds and 26 hours. Adapted from [56] with thanks to the American Geophysical Union. *Middle*: Scaled PDF for magnetic field energy density fluctuations based on the one-parameter scaling form of (39) with $\zeta_1 = s \approx 0.335$ from a 2D MHD simulation for homogeneous turbulence for $\Delta = 2$ (*green*), 8 (*red*), 32 (*blue*) units of grid spacing. Adapted from [33] with thanks to the American Institute of Physics. *Bottom panel*: Normalized PDF for fluctuations of transverse electric field intensity E_{north}/σ at $\tau = 5, 80, 1280$ ms, where $\sigma = \tau^{\zeta_1}(\zeta_1 \approx 0.8)$ is the root mean square of the fluctuations. The solid line corresponds to the normalized Castaing distribution with $\lambda = 1.02$. The results of the analyzes for Gaussian fluctuations (*dashed*) are shown for comparison. Reprinted from [112] with thanks to the American Geophysical Union**

described in Subsect. "Rank-Ordered Multifractal Analysis (ROMA)", we suggest that power-law relations of SOC can be made more precise if the exponents are determined more meticulously using rank-ordered fluctuations. For each rank-order, particularly for the rank-order of the smallest scaled fluctuations, one can determine a corresponding set of more accurately determined "critical exponents" that satisfy the classical exponent relations.

In summary, the conventional concepts of SOC or FSOC as tested by the lower order spatiotemporal statistical averages would generally be satisfied for intermittent turbulence. However, the ideas of invariant scaling and scaling relations will need to be generalized to include the "crossover effects" for large scaled fluctuations.

Dynamical Modeling – The Lu–Klimas Magnetic Field Reversal Model

By utilizing results obtained from direct numerical simulations, investigators were able to confirm a number of the theoretical predictions of intermittent turbulence and complexity phenomena in space plasmas that could arise from the process of merging and interactions of plasma coherent structures. The system sizes of direct numerical simulations, however, are generally limited by the capabilities of the present day computers. To circumvent this limitation, an alternative approach to study the intermittency in space plasma turbulence numerically would be to mimic the coarse-grained dissipation and the development of complexity by means of dynamical modeling which, in turn, may lead to a better theoretical understanding of the fundamental underlying stochastic processes.

While there exists a number of dynamical models that mimic avalanche processes exhibiting scaling and some also intermittency, only a few have attempted to incorporate the basic physics of plasmas in their formulations. From these, we select one simple model, the Lu–Klimas fast-field annihilation model of magnetic field reversal, for further discussion and comments. The model has the basic

ingredient of a diffusive equation [66]:

$$\frac{\partial B_x}{\partial t} = \frac{\partial}{\partial z}\left[D(z,t)\frac{\partial}{\partial z}\right]B_x + S(z,t)\,. \tag{46}$$

It may be obtained through the drastic reduction of the resistive MHD system in an idealized one-dimensional limit. In (46), $B_x(z,t)$ is the magnetic flux and $S(z,t)$ is a source term which could be interpreted as that characterizes the convection of the magnetic flux, although in discussions below it will be arbitrarily prescribed and assumed to be independent of time. The diffusion coefficient $D(z,t)$ is linearly proportional to an anomalous resistivity with a hysteresis characteristic as stipulated below:

$$\frac{\partial D}{\partial t} + \frac{D}{\tau} = \frac{Q(|\partial B_x/\partial z|)}{\tau} \tag{47}$$

where $Q = D_{\min}$ (low state) for $|\partial B_x/\partial z| < k$, and $Q = D_{\max} \gg D_{\min}$ (high state) for $|\partial B_x/\partial z| > \beta k$ with $\beta < 1$. Thus, Q is double-valued and dependent upon the history for $\beta k < |\partial B_x/\partial z| < k$. The value k may be thought of as a stability threshold which Klimas et al. [66] attributed to some sort of current driven instability (e. g., Lui et al. [79]). The value of Q remains in the low state ($= D_{\min}$) until the threshold is reached and switches to the high state ($= D_{\max}$). It stays at the high state until $|\partial B_x/\partial z| < \beta k$. This anomalous diffusive model is essentially that originally suggested by Lu [73], which was motivated by his studies related to the phenomena of self-organized criticality [3]; in particular, those related to the probability distributions of solar flares [74]. Unlike the Klimas' model, the source term in Lu's original formulation was assumed to be random and the boundary conditions were chosen to allow transport through the boundaries.

Klimas et al. [66] and Uritsky et al. [123] have studied their model numerically on a spatial interval $-L \le z \le L$ subject to the boundary conditions $\partial B_x/\partial z = 0$ at $z = \pm L$, i. e., no transport of B_x through the boundaries. They considered antisymmetric solutions with a field reversal at $z = 0$ and a time-independent source term $S(z,t) = S_0 \sin(\pi z/2L)$ which may be considered as a steady convection of B_x into the field reversal region. Setting $\tau = 1$ as the fundamental time scale and $L = 20$, they found that there exists a range of choices of parameters for which the dynamical model evolves in a sequence of quasi-periodic loading ($\mathrm{d}E/\mathrm{d}t > 0$)-unloading ($\mathrm{d}E/\mathrm{d}t < 0$) cycles where the total field energy $E(t)$ is defined by:

$$E(t) \equiv \int_{-L}^{L} \mathrm{d}z B_x^2(z,t)\,. \tag{48}$$

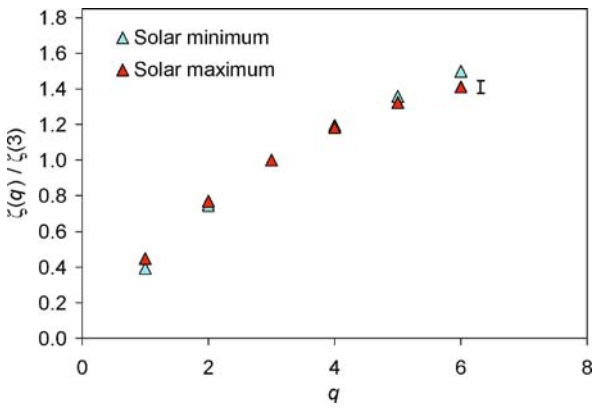

Space Plasmas, Dynamical Complexity in, Figure 20
Relative structure function exponents with respect to the third order exponent using ESS for flare statistics at solar maximum (high activity) and solar minimum (low activity) based on emissions detected by the EIT (extreme ultraviolet imaging telescope) aboard the SOHO spacecraft in the 195 Å wavelength. Adapted from [126] with thanks to the American Physical Society

During the loading intervals, when the system is quiet, the slow increase of $E(t)$ is the sole evolutionary feature of the system. As loading continues, the field gradient $|\mathrm{d}B_x/\mathrm{d}z|$, i. e. the current density, reaches the critical level for Q somewhere in the field distribution (usually at a pair of points symmetrically placed near $z = 0$). Subsequently a complicated process of unloading ensues, which is best understood in terms of the spatiotemporal evolution of the diffusion coefficient $D(z,t)$ as depicted in Fig. 21 during the initial stage of development of the unloading process, where the color scale of the figure indicates the intensity of the diffusion coefficient. In the figure, a white dot is placed at each point where the critical level of $|\mathrm{d}B_x/\mathrm{d}z|$ – or current density – for Q is reached and $Q(z,t) = D_{\max}$. As the instabilities spread and propagate toward both the center and the edges of the field distribution, more current sheets (i. e., sheet regions with strong current density) are excited and their topology becomes increasingly complex. The current sheets generate diffusion and the diffusion coefficient $D(z,t)$ is at a local maximum just after the passage of a current sheet at any given position. This strong localized diffusion enhances the transport of the magnetic flux toward $z = 0$ where fluxes of opposite signs meet and annihilate. This process of generation of complicated current sheet topology, transport of magnetic flux and annihilation adjusts self-consistently with the imposed source function such that there is a dynamically pseudo-stationary state. And, the loading-unloading process repeat quasi-periodically after an initial transient.

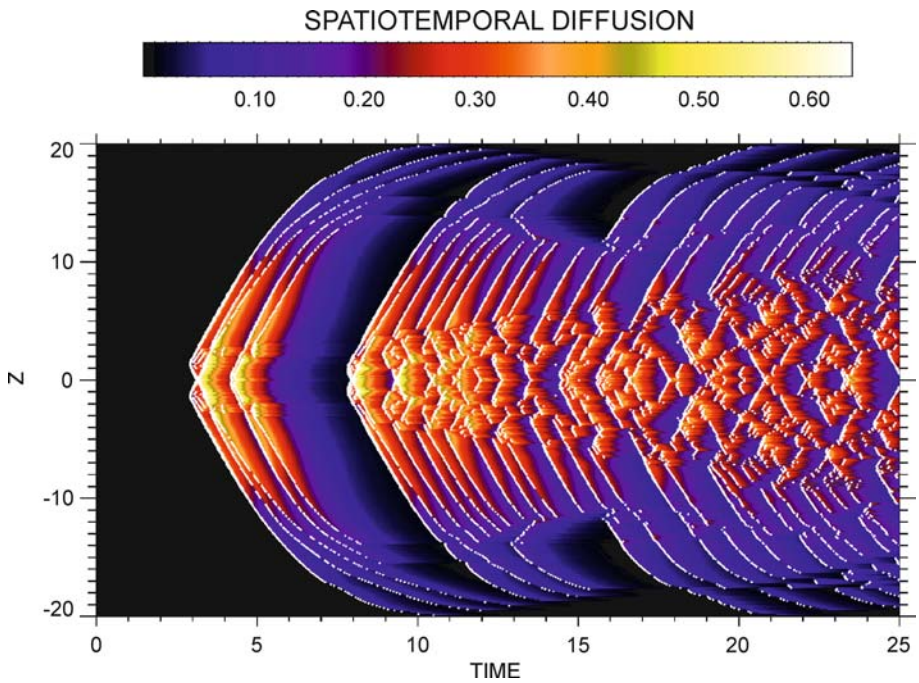

Space Plasmas, Dynamical Complexity in, Figure 21
Spatiotemporal evolution of the diffusion coefficient $D(z, t)$ for the Lu–Klimas model during the initial stage of development of an unloading process. The *color scale* of the figure indicates the intensity of the diffusion coefficient. A *white dot* is placed at each point where the critical level for Q is reached and $Q(z, t) = D_{max}$. Reprinted from [66] with thanks to the American Geophysical Union

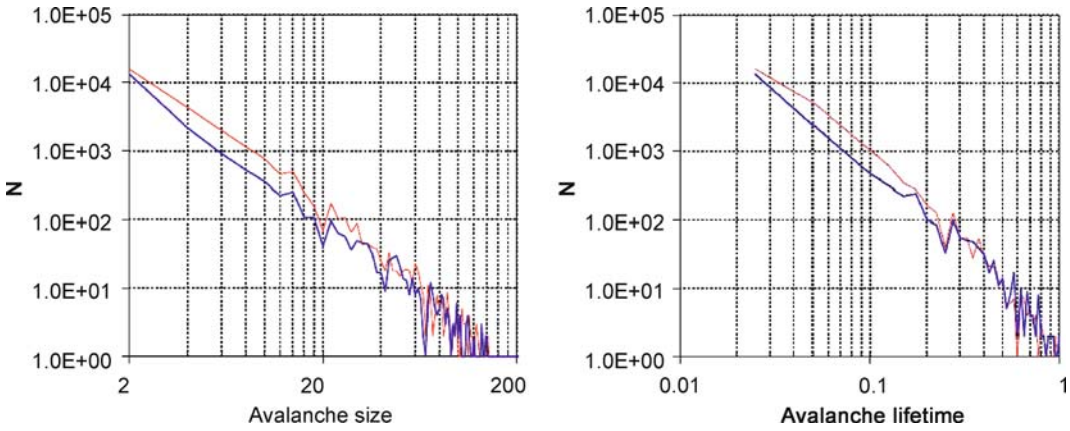

Space Plasmas, Dynamical Complexity in, Figure 22
Avalanche size (*left*) and lifetime (*right*) distributions for the Lu–Klimas model for early (*blue*) and late (*red*) intervals during a single unloading event. Reprinted from [123] with permission from Elsevier

The above phenomenon of continued sporadic generation of unstable current sheets and the accompanying broadening of the diffusive layer during the unloading process is reminiscent of the resonance broadening phenomenon of the "apparent reconnection" process in the 2D MHD simulation result for the sheared magnetic field geometry discussed earlier. Klimas et al. [67,68,69], Uritsky et al. [125] and Uritsky and Klimas [122] have recently extended their dynamical modeling effort based on the anomalous resistivity idea with hysteresis to the 2D geometry and the similarity between their calculated results of sporadic localized reconnections and the direct 2D MHD simulation results of resonance-broadening and coarse-grained dissipation – as well as the ensuing mag-

netic reconfiguration process – due to the sporadic, localized merging of coherent structures is even more noticeable, even though the boundary and initial conditions as well as the presumed instabilities of both sets of calculations were quite different. The main point to note is that the phenomenon of coarse-grained dissipation in the results of direct numerical simulation is replaced in the dynamical model by the assumed anomalous resistivity with hysteresis. Thus, other than the arbitrariness of the dissipation mechanism, such type of dynamical modeling has almost achieved its purported aim; i. e., using a much simplified dissipation model, large scale simulations of complexity of space plasmas may perhaps be achieved with relatively small scale computers.

However, the existing dynamical models may be too crude to lead to quantitatively comparable results of actual observations. For example, in the Lu–Klimas model, the critical threshold was arbitrary and kept at a constant value. In addition the phenomenon of hysteresis was again quite arbitrary and the double-valuedness of Q was also triggered at arbitrary levels. In the coarse-grained dissipation scenario acquired from, for example, direct numerical simulations, there was actually no specific threshold for triggering the "fluctuation induced nonlinear instability" and the dissipation during "apparent reconnection" was due to the combination of wave (e. g., Alfvén wave) propagation and resonance broadening. The hope is that perhaps even with some of such arbitrary ad hoc assumptions in the present and/or future modeling of local dissipation, the hypothesized model(s) will still lead to some of the global features such as the scale-invariant properties that will fall in the same universality class as the actual physical phenomena.

We now proceed to discuss the invariant properties of the simulated result of the one-dimensional magnetic field reversal and annihilation model. There are at least two sets of statistics that may shed light on the scale free invariant properties of SOC for the loading-unloading events that emerge from the numerical calculations of the Lu–Klimas model. Firstly, after an initial transient, the sequence of burst-like unloading events seem to occur by clustering around an average field strength, i. e., on average, the increasing strength of the field reversal is more-or-less balanced by the annihilation of the magnetic field in the current sheet. Thus, we expect that the unloading events exhibit power law distributions in, for example, energy dissipation and burst duration. These results were essentially uncovered by Klimas et al. [66] for the special situation of very low driving rates and $D_{\min} = 0$.

We may also ask the question about the scale free invariant behavior for the complicated current sheet distributions triggered by the instabilities associated with the anomalous resistivity for a single unloading event. Such small scale fluctuations again seem to self-consistently balance between the anomalous transport and field annihilation. Thus, we may again expect certain universal SOC behavior for these small scale fluctuations. To search for such behavior, Uritsky et al. [123] defined avalanches in terms of the size and duration of contiguous unstable grid points in the spatiotemporal domain. Power law statistics of the probability distributions of avalanche size and life time emerged from the statistics of their numerical calculations, Fig. 22. And, these exponents have been demonstrated to satisfy the standard exponent relations known for SOC processes [128]. There are other statistical verifications of SOC behavior from both the 1D and 2D dynamical modeling results, which we shall not discuss here.

More interestingly, Klimas et al. [69] recently demonstrated that the velocity fluctuations of their 2D simulation of the dynamical model exhibited both SOC and multifractal characteristics very similar to those expected for intermittent MHD turbulence; thus, corroborating with some of the basic concepts of interwoven connection and crossover phenomenon between SOC/FSOC and intermittent turbulence as espoused in the previous section.

Future Directions

We have now come to perhaps the most important part of this article. As discussed in the introduction, the development of the modern concepts of dynamical complexity in space plasmas is very recent. Despite a flurry of activities, current investigations – both from the theoretical and observational points of view – have barely scratched the surface of what may be achieved and understood. For example, we need to know more about the complexity behavior of the plasmas from the kinetic point of view. We also need to know how the plasma particle distributions respond to the fluctuations of its turbulent environment. In addition, the global space environment is generally inhomogeneous with complex initial and boundary conditions. Thus we need to address processes that are statistically nonstationary and spatially inhomogeneous as well as those that evolve temporally and nonuniformly. We also need to understand how the complexity and multifractal phenomena of space plasmas influence the global dynamics of the space plasma environment.

Throughout the discourse, we have placed our emphasis of dynamical complexity in space plasmas on the basic underlying physics. We have come to realize that such fundamental ideas must come from the understanding of the complicated stochastic behavior arisen from the inter-

actions of the multitudes of coherent structures that are prevalent in the natural development of the dynamics of space plasmas. We learned that there were various types of invariant concepts related to coarse-graining, which characterize the primary phenomena that we are attempting to study, analyze and understand. We studied a suite of statistical methodology to analyze complexity, some of which are new and some are carryovers from other fields of complexity, including hydrodynamic turbulence.

We realized that most of the so-called theories that exist today in dynamical complexity of space plasmas are mainly phenomenological concepts and generally mimic those that were prevalent before the modern developments of complexity. We know that these theories cannot carry us too far as we know that they had not advanced the true understanding of complexity phenomena in the past. For this reason, we have broken the tradition in this article and tried not to continually compare our statistically acquired results and theoretically developed understandings with the prevalent classical ideas and numerology. Such comparisons abound in the literature and generally murk the basic concepts that we are trying to convey.

We have learned that most of the phenomena of complexity in space plasmas are scale invariant concepts and yet very little theory has been developed to try to explore these invariant characteristics. We are still stuck with the dogma that we must develop our theories from the "basic elemental equations", while fully aware that the phenomena that we are trying to address generally have nothing to do with these equations directly. After all, nearly every invariant result gained from the coarse-graining concept breaks most of the symmetries and invariance properties of the basic elemental equations.

We have learned that the elemental dissipation such as classical resistivity and viscosity generally has nothing to do with the apparent dissipation that are everywhere in dynamical complexity. And yet every existing theoretical idea insists on including these almost irrelevant dissipative terms in its basic development.

We learned that intermittency generally involve large deviations from the mean and yet we insist on fitting every statistical result into neatly packed curves as if all physics and theories must result in smooth curves in parametric studies. It is time to bring in the ideas of Lebesgue measure, extreme value statistics and thermodynamics of rare events, as well as rank-ordered groupings and associated analyzes to our future developments of new theories and statistical methods

With enthusiasm, we look forward to the dawning of a new era from the future horizon when these new concepts and theoretical understandings finally establish their footings in the modern development of dynamical complexity in space plasmas.

Acknowledgments

The author is extremely grateful to Cheng–chin Wu for his unremitting collaboration and for many of the numerical results that are displayed in this article. He is indebted to Vassilis Angelopoulos, Roberto Bruno, Sandra Chapman, Abe Chian, Giuseppe Consolini, Marius Echim, Bogdan Hnat, Jack Jasperse, Charlie Kennel, Margy Kivelson, Alex Klimas, Hervé Lamy, Tony Lui, Eckart Marsch, Bill Matthaeus, John Podesta, Surja Sharma, Sunny Tam, Vadim Uritsky, Dimitri Vassiliadis, Zoltan Vörös, Nick Watkins, Jim Weygand, and Gary Zank for cooperative research and/or useful discussions. He is very appreciative to those who kindly allowed him to quote and reprint their published results and to Rich Chang for a critical reading of the manuscript and for his helpful suggestions. The author wishes to thank the Section Editor of Chaos and Complexity in Astrophysics for this encyclopedia volume, for inviting him to contribute this article and for his careful editing and useful suggestions. He is most grateful to members of his family for their constant encouragement and particularly to his wife, Monica, for her understanding and persistent spurring. Permissions from Elsevier. Springer Science and Business Media, the American Astrophysical Society, and the Italian Society of Physics and the liberal copyright policies of the American Physical Society, the American Institute of Physics, the American Geophysical Union, and the European Geophysical Union, which allowed the author to reprint figures and illustrations from their publications, are gratefully acknowledged. This project was partially supported by the AFOSR, NASA and NSF of the US government.

Bibliography

Primary Literature

1. Alexandrova O, Mangeney A, Maksimovic M, Cornilleau-Wehrlin N, Bosqued JM, André M (2006) Alfvén vortex filaments observed in magnetosheath downstream of a quasi-perpendicular bow shock. J Geophys R 111:A12208; doi:10.1029/2006JA011934
2. Angelopoulos V, Mukai T, Kokubun S (1999) Evidence for intermittency in Earth's plasma sheet and implications for self-organized criticality. Phys Plasmas 6:4161–4168
3. Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality: an explanation of 1/f noise. Phys Rev Lett 59:381–384
4. Baker DN, Klimas AJ, McPherson RL, Büchner J (1990) The evolution from weak to strong geomagnetic activity: An interpretation in terms of deterministic chaos. Geophys Res Lett 17:41–44
5. Balatoni J, Renyi A (1956) Pub Math Inst Hung Acad Sci 1:9

6. Benzi R, Ciliberto S, Tripiccione R, Baudet C, Massaioli F, Succi S (1993) Extended self-similarity in turbulent flows. Phys Rev E 48:R29–32
7. Biskamp D (1993) Nonlinear magnetohydrodynamics. Cambridge University Press, Cambridge
8. Biskamp D (2003) Magnetohydrodynamic turbulence. Cambridge University Press, Cambridge
9. Bruno R, Bavassano B, Pietropaolo E, Carbone V, Veltri P (1999) Effects of intermittency on interplanetary velocity and magnetic field fluctuations anisotropy. Geophys Res Lett 26:3185–3188
10. Bruno R, Carboni V, Veltri P, Pietropaolo E, Bavassano B (2001) Identifying intermittency events in the solar wind. Planet Space Sci 49:1201–1210
11. Bruno R, Carbone V, Chapman S, Hnat B, Noullez A, Sorriso-Valvo L (2007) Intermittent character of interplanetary magnetic field fluctuations. Phys Plasmas 14:032901
12. Burlaga LF (1991) Intermittent turbulence in the solar wind. J Geophys Res 96:5847–5851
13. Burlaga LF (1991) Multifractal structure of the interplanetary magnetic field: Voyager 2 observations near 25 AU, 1987–1988. Geophys Res Lett 18:69–72
14. Burlaga LF (1991) Multifractal structure of speed fluctuations in recurrent streams at 1AU and near 6 AU. Geophys Res Lett 18:1651–1654
15. Burlaga LF (1993) Intermittent turbulence in large-scale velocity fluctuations at 1 AU near solar maximum. J Geophys R 98:17467–17473
16. Carbone V (1993) Cascade model for intermittency in fully developed magneto-hydrodynamic turbulence. Phys Rev Lett 71:1546–1548
17. Carbone V (1994) Scaling exponents of the velocity structure functions in the interplanetary medium. Ann Geophys 12:585–590
18. Carbone V, Bruno R (1996) Cancellation exponents and multifractal scaling laws in the solar wind magnetohydrodynamic turbulence. Ann Geophys 14:777–785
19. Carbone V, Veltri P, Bruno R (1995) Experimental evidence for differences in the extended self-similarity scaling laws between fluid and magnetohydrodynamic turbulent flows. Phys Rev Lett 75:3110–3113
20. Carbone V, Bruno R, Veltri P (1996) Evidence for extended self-similarity in hydromagnetic turbulence. Geophys Res Lett 23:121–124
21. Castaing B, Gagne Y, Hopfinger EJ (1990) Velocity probability density functions of high Reynolds number turbulence. Physica D 46:177–200
22. Chang T (1992) Low-dimensional behavior and symmetry breaking of stochastic systems near criticality – Can these effects be observed in space and in the laboratory? IEEE Trans Plasma Sci 20:691–694
23. Chang T (1998) Sporadic localized reconnections and multiscale intermittent turbulence in the magnetotail. In: Horwitz JL, Gallagher DI, Peterson WK (eds) Geospace mass and energy flow. Geophysical Monograph, vol 104. AGU, Washington, pp 193–199
24. Chang T (1999) Self-organized criticality, multifractal spectra, sporadic localized reconnections and intermittent turbulence in the magnetotail. Phys Plasmas 6:4137–4145
25. Chang T (2001) An example of resonances, coherent structures and topological phase transitions – the origin of the low frequency broadband spectrum in the auroral zone. Nonlinear Process Geophys 8:175–179
26. Chang T (2001) Colloid-like behavior and topological phase transitions in space plasmas: intermittent low frequency turbulence in the auroral zone. Phys Scr T89:80–83
27. Chang T, Wu CC (2007) Dynamical complexity, intermittent turbulence, coarse-grained dissipation, criticality and multifractal processes. In: Shaikh D (ed) Turbulence and nonlinear processes in astrophysical plasmas. AIP Proc 932:161–166
28. Chang T, Wu CC (2008) Rank-ordered multifractal spectrum for intermittent fluctuations. Phys Rev E 77:045401(R)
29. Chang T, Hankey A, Stanley HE (1973) Generalized scaling hypothesis in multicomponent systems, vol 1. Classication of critical points by order and scaling at tricriical points. Phys Rev B 8:346–364
30. Chang T, Nicoll JF, Young JF (1978) A closed-form differential renormalization-group generator for critical dynamics. Phys Lett A 67:287–290
31. Chang T, Vvedensky DD, Nicoll JF (1992) Differential renormalization-group generators for static and dynamic critical phenomena. Phys Reports 217:279–362
32. Chang T, Tam SWY, Wu CC, Consolini G (2003) Complexity, forced and/or self-organized criticality, and topological phase transitions in space plasmas. Space Sci Rev 107:425–445
33. Chang T, Tam SWY, Wu CC (2004) Complexity induced anisotropic bimodal intermittent turbulence in space plasmas. Phys Plasmas 11:1287–1299
34. Chang T, Tam SWY, Wu CC (2006) Complexity in space plasmas – a brief review. Space Sci Rev 122:281–291
35. Chapman SC, Watkins NW, Dendy RO, Helander P, Rowlands G (1998) A simple avalanche model as an analogue for the magnetospheric activity. Geophys Res Lett 25:2397–2400
36. Chapman SC, Dendy RO, Rowlands G (1999) A sandpile model with dual scaling regimes for laboratory, space, and astrophysical plasmas. Phys Plasmas 6:4169–4177
37. Consolini G (1997) Sandpile cellular automata and the magnetospheric dynamics. In: Aiello S, Iucci N, Sironi G, Treves A, Villante U (ed) Cosmic Physics in the Year 2000. Proc VIII GIFCO Conference, Società Italiana di Fisica, Bologna, pp 123–126
38. Consolini G, Chang T (2001) Magnetic field topology and criticality in geotail dynamics: Relevance to substorm phenomena. Space Sci Rev 95:309
39. Consolini G, Chang T (2002) Complexity, magnetic field topology, criticality, and metastability in magnetotail dynamics. J Atmos Solar Terr Phys 64:541
40. Consolini G, Lui ATY (2000) Symmetry Breaking and Nonlinear Wave-Wave Interaction in Current disruption: Possible evidence for a phase Transition. In: Ohtani S et al (eds) Magnetospheric Current Systems. Geophysical Monograph, vol 118. AGU, Washington, p 395
41. Consolini G, Marcucci MF, Candidi M (1996) Multifractal structure of auroral electrojet index data. Phys Rev Lett 76:4082–4085
42. Consolini G, Chang T, Lui ATY (2005) Complexity and topological disorder in the Earth's magnetotail dynamics. In: Sharma AS, Kaw PK (eds) Nonequilibrium phenomena in plasmas. Springer, Dordrecht, pp 51–69
43. Daubechies I (1992) Ten lectures on wavelets. Society for Industrial and Applied Mathematics, Philadelphia, p 357

44. Echim MM, Lamy H, Chang T (2007) Multipoint observations of intermittency in the cusp regions. Nonlinear Processes in Geophysics 14:525–534
45. Einaudi G, Velli M (1999) The distribution of flares, statistics of magnetohydrodynamic turbulence and coronal heating. Phys Plasmas 6:4146–4153
46. Farge M (1992) Wavelet transforms and their applications to turbulence. Ann Rev Fluid Mech 24:395–457
47. Feynman J, Ruzmaikin A (1994) Distributions of the interplanetary field revisited. J Geophys Res 99:17645–17651
48. Forman M, Burlaga LF (2003) Exploring the Castaing distribution function to study intermittence in the solar wind at L1 in June 2000. In: Velli M, Bruno R, Malara F (eds) Solar wind ten. AIP Conf Proc 679:554–557
49. Forster D, Nelson DR, Stephen MJ (1977) Large distance and long time properties of a randomly stirred fluid. Phys Rev A 16:732–749
50. Grassberger P, Procaccia I (1983) Characterization of strange attractors. Phys Rev Lett 50:346–349
51. Grossmann S, Lohse D, Reeh A (1997) Application of extended self-similarity in turbulence. Phys Rev E 56:5473–5478
52. Haar A (1910) Zur Theorie der Orthogonalen Functionensysteme. Math Ann 69:331–371
53. Halperin BI, Hohenberg PC, Ma SK (1972) Calculation of dynamical critical properties using Wilson's expansion methods. Phys Rev Lett 29:1548–1551
54. Halsey T, Jensen MH, Kadanoff LP, Procaccia I, Schraiman BI (1986) Fractal measures and their singularities: The characterization of strange sets. Phys Rev A 33:1141–1151
55. Hentschel HGE, Procaccia I (1983) The infinite number of generalized dimensions of fractals and strange attractors. Physica 8D:435–444
56. Hnat B, Chapman SC, Rowlands G, Watkins NW, Farrell WM (2002) Finite size scaling in the solar wind magnetic field energy density as seen by WIND. Geophys Res Lett 29:1446; doi:10.1029/2001GL014587
57. Hnat B, Chapman SC, Rowlands G, Watkins NW, Freeman MP (2003) Scaling in long term data sets of geomagnetic indices and solar wind ε as seen by WIND spacecraft. Geophys Res Lett 30:2174; doi:10.1029/2003GL018209
58. Horbury TA, Balogh A, Forsyth RJ, Smith EJ (1995) Ulysses magnetic field observations of fluctuations within polar coronal flows. Ann Geophys 13:105–107
59. Horbury TA, Balogh A, Forsyth RJ, Smith EJ (1995) Observations of evolving turbulence in the polar solar wind. Geophys Res Lett 22:3401–3404
60. Horbury TA, Balogh A, Forsyth RJ, Smith EJ (1997) Ulysses observations of intermittent heliospheric turbulence. Adv Space Phys 19:847–850
61. Hosino M, Nishida A, Yamamoto T, Kokubun S (1994) Turbulent magnetic field in the distant magnetotail: Bottom-up process of plasmoid formation? Geophys Res Lett 21:2935–2938
62. Hwa T, Kardar M (1992) Avalanches, hydrodynamics, and discharge events in models of sandpiles. Phys Rev A 45:7002–7023
63. Jensen HJ (1998) Self-organized criticality – Emergent complex behavior in physical and biological systems. Cambridge University Press, Cambridge
64. Jiang GS, Wu CC (1999) A high-order WENO finite difference scheme for the equations of ideal magnetohydrodynamics. J Comp Phys 150:561–594
65. Klimas AJ, Baker DN, Roberts DA, Fairfield DH, Büchner J (1992) A nonlinear dynamical analogue model of geomagnetic activity. J Geophys Res 97:12253–12266
66. Klimas AJ, Valdivia JA, Vassiliadis D, Baker DN, Hesse M, Takalo J (2000) Self-organized criticality in the substorm phenomenon and its relation to localized reconnection in the magnetospheric plasma sheet. J Geophys Res 105:18765–18780
67. Klimas AJ, Uritsky VM, Vassiliadis D, Baker DN (2004) Reconnection and scale-free avalanching in a driven current-sheet model. J Geophys Res 109:A02218; doi:10.1029/2003JA010036
68. Klimas AJ, Uritsky VM, Vassiliadis D, Baker DN (2005) A mechanism for the loading-unloading substorm cycle missing in MHD global magnetospheric simulation models. Geophys Res Lett 32:L14108; doi:10.1029/2005GL022916
69. Klimas AJ, Uritsky VM, Paczuski M (2007) Self-organized criticality and intermittent turbulence in an MHD current sheet with a threshold Instability. arXiv:astro-ph/0701486v2
70. Langmuir I (1928) Oscillations in ionized gases. Proc Natl Acad Sci 14:628
71. Leubner LP (2004) Fundamental issues on kappa-distributions in space plasmas and interplanetary proton distributions. Phys Plasmas 11:1308–1316
72. Leubner MP, Vörös Z (2005) A nonextensive entropy approach to solar wind intermittency. Astrophys J 618:547–555
73. Lu ET (1995) Avalanches in continuum driven dissipative systems. Phys Rev Lett 74:2511–2514
74. Lu ET, Hamilton RJ (1991) Avalanches and the distribution of solar flares. Astrophys J 380:L89–L92
75. Lui ATY (1996) Current disruption in the Earth's magnetosphere: observations and models. J Geophys Res 101:13067–13088
76. Lui ATY (2001) Multifractal and intermittent nature of substorm-associated magnetic turbulence in the magnetotail. J Atmos Solar Terr Phys 63:1379–1385
77. Lui ATY (2002) Multiscale phenomena in the near-Earth magnetosphere. J Atmos Solar Terr Phys 64:125
78. Lui ATY, Lopez RE, Krimigis SM, McEntire RW, Zanetti LJ, Potemra TA (1988) A case study of magnetotail current sheet disruption and diversion. Geophys Res Lett 15:721–724
79. Lui ATY, Chang CL, Mankofsky A, Wong HK, Winske D (1991) A cross-field current instability for substorm expansions. J Geophys Res 96:11389–11401
80. Lui ATY, Liou K, Newell PT, Meng CI, Ohtani SI, Kokubun S, Ogino T, Brittnacher M, Parks G (1998) Plasma and magnetic flux transport associated with auroral breakups. Geophys Res Lett 25:4059–4062
81. Lui ATY, Chapman SC, Liou K, Newell PT, Meng CI, Brittnacher M, Parks GK (2000) Is the dynamic magnetosphere an avalanching system? Geophys Res Lett 27:911–914
82. Lui ATY, Zheng Y, Zhang Y, Rème H, Dunlop MW, Gustafsson G, Mende SB, Mouikis C, Kistler LM (2006) Cluster observation of plasma flow reversal in the magnetotail during a substorm. Ann Geophys 24:2005–2013
83. Mandelbrot B (1977) Fractals: Form, chance and dimension. Freeman, San Francisco

84. Marsch E, Liu S (1993) Structure functions and intermittency of velocity fluctuations in the inner solar wind. Ann Geophys 11:227–238
85. Marsch E, Tu CY (1994) Non-Gaussian probability distributions of solar wind fluctuations. Ann Geophys 12:1127–1138
86. Marsch E, Tu CY (1997) Intermittency, non-Gaussian statistics and fractal scaling of MHD fluctuations in the solar wind. Nonlinear Process Geophys 4:101–124
87. Marsch E, Tu CY, Rosenbauer H (1996) Multifractal scaling of the kinetic energy flux in solar wind turbulence. Ann Geophys 14:259–269
88. Matthaeus WH, Goldstein ML (1982) Measurements of the rugged invariants of magnetohydrodynamic turbolence in the solar wind. J Geophys Res 87:6011–6028
89. Matthaeus WH, Goldstein ML, Roberts DA (1990) Evidence for the presence of quasi-two-dimensional nearly incompressible fluctuations in the solar wind. J Geophys Res 95:20673–20683
90. Matthaeus WH, Qin G, Bieber JW, Zank GP (2003) Nonlinear collisionless perpendicular diffusion of charged particles. Astrophys J Lett 590:L53–L56
91. Meneveau C (1991) Analysis of turbulence in the orthonormal wavelet representation. J Fluid Mech 232:469–521
92. Meneveau C, Sreenivasan KR (1987) Simple multifractal cascade model for fully developed turbulence. Phys Rev Lett 59:1424–1427
93. Milovanov AV, Zelenyi LM, Zimbardo G (1996) Fractal structures and power law spectra in the distant Earth's magnetotail. J Geophys Res 101:19903–19910
94. Morlet J (1983) Sampling theory and wave propagation. In: Chen CH (ed) Issues on acoustic signal/image processing and recognition, NATO ASI, vol F1. Springer, Berlin
95. Müller WC, Biskamp D (2002) Scaling properties of three-dimensional magnetohydrodynamic turbulence. Phys Rev Lett 84:475–478
96. Ott E (1993) Chaos in dynamical systems. Cambridge University Press, New York
97. Pagel C, Balogh A (2001) A study of magnetic fluctuations and their anomalous scaling in the solar wind: the Ulysses fast-latitude scan. Nonlinear Process Geophys 8:313–330
98. Parisi G, Frisch U (1985) On the singularity structure of fully developed turbulence. In: Ghil M, Benzi R, Parisi G (eds) Turbulence and predictability in geophysical fluid dynamics. Proc Intern School of Physics, 'Erico Fermi' 1983, Varenna, Italy. North Holland, Amsterdam, pp 84–87
99. Pinheiro MJ (2007) Plasmas: the genesis of the word. arXiv:physics/0703260v1
100. Politano H, Pouquet A, Carbone V (1998) Determination of anomalous exponents of structure functions in two-dimensional magnetohydrodynamic turbulence. Europhys Lett 43:516–521
101. Roberts DA, Baker DN, Klimas AJ, Bargatze LF (1991) Indications of low dimensionality in magnetospheric dynamics. Geophys Res Lett 18:151–154
102. Rusmaikin A, Lyannaya IR, Styashkin VA, Yeroshenko E (1993) The spectrum of the interplanetary magnetic field near 1.3 AU. J Geophys Res 98:13303–13306
103. Shan LH, Hansen P, Goertz CK, Smith RA (1991) Chaotic appearance of the AE index. Geophys Res Lett 18:147–150
104. Shan LH, Goertz CK, Smith RA (1991) On the embedding-dimension analysis of AE and AL time series. Geophys Res Lett 18:1647–1650
105. Sharma AS, Vassiliadis DV, Papadopoulos K (1993) Reconstruction of low-dimensional magnetospheric dynamics by singular spectrum analysis. Geophys Res Lett 20:335–338
106. Sitnov GL, Sharma AS, Papadopoulos K, Vassiliadis D, Valdivia JA, Klimas AJ, Baker DN (2000) Phase transition-like behavior of the magnetosphere during substorms. J Geophys Res 105:12955–12974
107. Sorriso-Valvo L, Carbone V, Veltri P, Consolini G, Bruno R (1999) Intermittency in the solar wind turbulence through probability distribution functions of fluctuations. Geophys Res Lett 26:1801–1804
108. Sorriso-Valvo L, Carbone V, Giuliani P, Veltri P, Bruno R, Antoni V, Martines E (2001) Intermittency in plasma turbulence. Planet Space Sci 49:1193–1200
109. Sundkvist D, Krasnoselskikh V, Shukla PK, Vaivads A, André M, Buchert S, Rème H (2005) In situ multi-satellite detection of coherent vortices as a manifestation of Alfvénic turbulence. Nature 436:825–828
110. Takalo J, Timonen J, Klimas AJ, Valdivia JA, Vassiliadis D (2001) A coupled map as a model of the dynamics of the magnetotail current sheet. J Atmos Solar Terr Phys 63:1407–1414
111. Tam SWY, Chang T (2005) Energization of ions by bimodal intermittent fluctuations. In: Lui ATY, Kamide Y, Consolini G (eds) Multiscale coupling of sun-earth processes. Elsevier, Netherlands, pp 375
112. Tam SWY, Chang T, Kintner PM, Klatt E (2005) Intermittency analyses on the SIERRA measurements of the electric field fluctuations in the auroral zone. Geophys Res Lett 32:L05109; doi:10.1029/2004GL021445
113. Taylor GI (1938) The spectrum of turbulence. Proc R Soc Lond A164:476–490
114. Taylor JB (1974) Relaxation of toroidal plasma and generation of reverse magnetic fields. Phys Rev Lett 33:1139–1141
115. Tetreault D (1991) Theory of electric fields in the auroral acceleration region. J Geophys Res 96:3549–3563
116. Tetreault D (1992) Turbulent relaxation of magnetic fields, 1. Coarse-grained dissipation and reconnection. J Geophys Res 97:8531–8540
117. Tetreault D (1992) Turbulent relaxation of magnetic fields, 2. Self-organization and intermittency. J Geophys Res 97:8541–8547
118. Tonks L (1967) Am J Phys 35:857–858
119. Tonks L, Langmuir I (1929) Oscillations in ionized gases. Phys Rev 33:195–210
120. Tu CY, Marsch E (1995) MHD structures, waves and turbulence in the solar wind – observations and theories. Kluwer, Dordrecht
121. Tu CY, Marsch E, Rosenbauer H (1996) An extended structure function model and its application to the analysis of solar wind intermittency properties. Ann Geophys 14:270–285
122. Uritsky VM, Klimas AJ (2005) Hysteresis-controlled instability waves in a scale-free driven current sheet model. Nonlinear Process Geophys 12:827–833
123. Uritsky VM, Klimas AJ, Valdivia JA, Vassiliadis D, Baker DN (2001) Stable critical behavior and fast field annihilation in a magnetic field reversal model. J Atmos Solar Terr Phys 63:1425–1433

124. Uritsky VM, Klimas AJ, Vassiliadis D, Chua D, Parks G (2002) Scale-free statistics of spatiotemporal auroral emissions as depicted by POLAR UVI images: Dynamic magnetosphere is an avalanching system. J Geophys Res 107:1426; doi:10.1029/2001JA000281
125. Uritsky VM, Klimas AJ, Vassiliadis D (2002) Multiscale dynamics and robust critical scaling in a continuum current sheet model. Phys Rev E 65:046113
126. Uritsky VM, Paczuski M, Davila JM, Jones SI (2007) Coexistence of self-organized criticality and intermittent turbulence in the solar corona. Phys Rev Lett 99:025001
127. Vahnstein S, Sreenivasan K, Pierrehumbert R, Kashyap V, Juneja A (1994) Scaling exponents for turbulence and other random processes and their relationships with multifractal structure. Phys Rev E 50:1823–1835
128. Vespignani A, Zapperi S (1998) How self-organized criticality works: A unified mean-field picture. Phys Rev E 57:6345–6362
129. Vörös Z (2000) On multifractality of high-latitude geomagnetic fluctuations. Ann Geophys 18:1273–1283
130. Vörös Z, Baumjohann W, Nakamura R, Runov A, Zhang TL, Volwerk M, Eichelberger HU, Balogh A, Horbury TS, Glassmeier KH, Klecker B, Rème H (2003) Multi-scale magnetic field intermittence in the plasma sheet. Ann Geophys 21:1955–1964
131. Vörös Z, Baumjohann W, Nakamura R, Runov A, Volwerk M, Zhang TL, Balogh A (2004) Wavelet analysis of magnetic turbulence in the Earth's plasma sheet. Phys Plasmas 11:1333–1338
132. Vörös Z, Baumjohann W, Nakamura R, Runov A, Volwerk M, Schwarzl H, Balogh A, Rème H (2005) Dissipation scales in the Earth's plasma sheet estimated from cluster measurements. Nonlinear Process Geophys 12:725–732
133. Watkins NW, Chapman SC, Dendy RO, Rowlands G (1999) Robustness of collective behaviour in a strongly driven avalanche model: magnetospheric implications. Geophys Res Lett 26:2617–2620
134. Weygand JM, Kivelson MG, Khurana KK, Schwarzl HK, Thompson SM, McPherron RL, Balogh A, Kistler LM, Goldstein ML, Borovsky J, Roberts DA (2005) Plasma sheet turbulence observed by Cluster, vol II. J Geophys Res 110:A01205; doi:10.1029/2004JA010581
135. Wilson KG, Kogut J (1994) The renormalization group and the ε expansion. Phys Reports 12C:75–200
136. Wu CC, Chang T (2000) 2D MHD simulation of the emergence and merging of coherent structures. Geophys Res Lett 27:863–866
137. Wu CC, Chang T (2001) Further study of the dynamics of two-dimensional MHD coherent structures – a large scale simulation. J Atmos Solar Terr Phys 63:1447–1453
138. Wu CC, Chang T (2005) Intermittent turbulence in 2D MHD simulation. In: Lui ATY, Kamide Y, Consolini G (eds) Multiscale coupling of sun-earth processes. Elsevier, Netherlands, p 321
139. Yordanova E, Grezesiak M, Wernik AW, Popielawska B, Stasiewicz K (2005) Multifractal structure of turbulence in the magnetospheric cusp. Ann Geophys 22:2431–2440

Books and Reviews

Bruno R, Carbone V (2005) The solar wind as a turbulence laboratory. Living Rev Solar Phys 2:4; http://www.livingreviews.org/lrsp-2005-4. Accessed 15 June 2007
Chang T, Chapman S, Klimas A (eds) (2001) Forced and/or self-organized criticality (FSOC) in space plasmas. J Atmos Solar Terr 63:1359–1453
Forster D (1975) Hydrodynamic fluctuations, broken symmetry, and correlation functions. Benjamin, Reading
Frisch U (1995) Turbulence. Cambridge University Press, Cambridge
Lavenda BH (1995) Theromodynamics of Extremes. Albion Publishers, Chichester
Leubner MP, Baumjohann W, Chian ACL (eds) (2006) Advances in space environment research, vol 2. Space Sci Rev 122:1–337
Lui ATY, Kamide Y, Consolini G (eds) (2005) Multiscale coupling of sun-earth processes. Elsevier, the Netherlands
Ma SK (1976) Modern theory of critical phenomena. Benjamin, Reading
Sharma AS, Kaw PK (eds) (2005) Nonequilibrium phenomena in plasmas. Astrophysics and Space Science Library, vol 321. Springer, Dordrecht, pp 1–340
Sornette D (2000) Critical phenomena in natural sciences; Chaos, fractals, selforganization and disorder: concepts and tools. Springer, Berlin

Spectral Theory of Dynamical Systems

Mariusz Lemańczyk
Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Toruń, Poland

Article Outline

Glossary
Definition of the Subject
Introduction
Maximal Spectral Type of a Koopman Representation, Alexeyev's Theorem
Spectral Theory of Weighted Operators
The Multiplicity Function
Rokhlin Cocycles
Rank-1 and Related Systems
Spectral Theory of Dynamical Systems of Probabilistic Origin
Inducing and Spectral Theory
Special Flows and Flows on Surfaces, Interval Exchange Transformations
Future Directions
Acknowledgments
Bibliography

Glossary

Spectral decomposition of a unitary representation If $\mathcal{U} = (U_a)_{a\in\mathbb{A}}$ is a continuous unitary representation of a locally compact second countable (l.c.s.c.)

Abelian group $\mathbb{A}$ in a separable Hilbert space H then a decomposition $H = \bigoplus_{i=1}^{\infty} \mathbb{A}(x_i)$ is called *spectral* if $\sigma_{x_1} \gg \sigma_{x_2} \gg \ldots$ (such a sequence of measures is also called *spectral*); here $\mathbb{A}(x) := \overline{span}\{U_a x \colon a \in \mathbb{A}\}$ is called the *cyclic space* generated by $x \in H$ and σ_x stands for the spectral measure of x.

Maximal spectral type and the multiplicity function of $\mathcal{U}$ The *maximal spectral type* $\sigma_{\mathcal{U}}$ of $\mathcal{U}$ is the type of σ_{x_1} in any spectral decomposition of H; the *multiplicity function* $M_{\mathcal{U}} \colon \widehat{\mathbb{A}} \to \{1, 2, \ldots\} \cup \{+\infty\}$ is defined $\sigma_{\mathcal{U}}$-a.e. and $M_{\mathcal{U}}(\chi) = \sum_{i=1}^{\infty} 1_{Y_i}(\chi)$, where $Y_1 = \widehat{\mathbb{A}}$ and $Y_i = \operatorname{supp} \mathrm{d}\sigma_{x_i}/\mathrm{d}\sigma_{x_1}$ for $i \geq 2$.

A representation $\mathcal{U}$ is said to have *simple spectrum* if H is reduced to a single cyclic space. The multiplicity is *uniform* if there is only one essential value of $M_{\mathcal{U}}$. The essential supremum of $M_{\mathcal{U}}$ is called the *maximal spectral multiplicity*. $\mathcal{U}$ is said to have *discrete spectrum* if H has an orthonormal base consisting of eigenvectors of $\mathcal{U}$; $\mathcal{U}$ has *singular* (*Haar*, *absolutely continuous*) spectrum if the maximal spectral type of $\mathcal{U}$ is singular with respect to (equivalent to, absolutely continuous with respect to) a Haar measure of $\widehat{\mathbb{A}}$.

Koopman representation of a dynamical system $\mathcal{T}$ Let $\mathbb{A}$ be a l.c.s.c. (and not compact) Abelian group and $\mathcal{T} \colon a \mapsto T_a$ a representation of $\mathbb{A}$ in the group $Aut(X, \mathcal{B}, \mu)$ of (measure-preserving) automorphisms of a standard probability Borel space $(X, \mathcal{B}, \mu)$. The *Koopman representation* $\mathcal{U} = \mathcal{U}_{\mathcal{T}}$ of $\mathcal{T}$ in $L^2(X, \mathcal{B}, \mu)$ is defined as the unitary representation $a \mapsto U_{T_a} \in U(L^2(X, \mathcal{B}, \mu))$, where $U_{T_a}(f) = f \circ T_a$.

Ergodicity, weak mixing, mild mixing, mixing and rigidity of $\mathcal{T}$ A measure-preserving $\mathbb{A}$-action $\mathcal{T} = (T_a)_{a \in \mathbb{A}}$ is called *ergodic* if $\chi_0 \equiv 1 \in \widehat{\mathbb{A}}$ is a simple eigenvalue of $\mathcal{U}_{\mathcal{T}}$. It is *weakly mixing* if $\mathcal{U}_{\mathcal{T}}$ has a continuous spectrum on the subspace $L_0^2(X, \mathcal{B}, \mu)$ of zero mean functions. $\mathcal{T}$ is said to be *rigid* if there is a sequence (a_n) going to infinity in $\mathbb{A}$ such that the sequence $(U_{T_{a_n}})$ goes to the identity in the strong (or weak) operator topology; $\mathcal{T}$ is said to be *mildly mixing* if it has no non-trivial rigid factors. We say that $\mathcal{T}$ is *mixing* if the operator equal to zero is the only limit point of $\{U_{T_a}|_{L_0^2(X,\mathcal{B},\mu)} \colon a \in \mathbb{A}\}$ in the weak operator topology.

Spectral disjointness Two $\mathbb{A}$-actions S and $\mathcal{T}$ are called *spectrally disjoint* if the maximal spectral types of their Koopman representations $\mathcal{U}_{\mathcal{T}}$ and $\mathcal{U}_S$ on the corresponding L_0^2-spaces are mutually singular.

SCS property We say that a Borel measure σ on $\widehat{\mathbb{A}}$ satisfies the *strong convolution singularity* property (SCS property) if, for each $n \geq 1$, in the disintegration (given by the map $(\chi_1, \ldots, \chi_n) \mapsto \chi_1 \cdot \ldots \cdot \chi_n$) $\sigma^{\otimes n} = \int_{\widehat{\mathbb{A}}} \nu_\chi \, d\sigma^{(n)}(\chi)$ the conditional measures ν_χ are atomic with exactly $n!$ atoms ($\sigma^{(n)}$ stands for the nth convolution $\sigma * \ldots * \sigma$). An $\mathbb{A}$-action $\mathcal{T}$ satisfies the SCS property if the maximal spectral type of $\mathcal{U}_{\mathcal{T}}$ on L_0^2 is a type of an SCS measure.

Kolmogorov group property An $\mathbb{A}$-action $\mathcal{T}$ satisfies the *Kolmogorov group property* if $\sigma_{\mathcal{U}_{\mathcal{T}}} * \sigma_{\mathcal{U}_{\mathcal{T}}} \ll \sigma_{\mathcal{U}_{\mathcal{T}}}$.

Weighted operator Let T be an ergodic automorphism of $(X, \mathcal{B}, \mu)$ and $\xi \colon X \to \mathbb{T}$ be a measurable function. The (unitary) operator $V = V_{\xi,T}$ acting on $L^2(X, \mathcal{B}, \mu)$ by the formula $V_{\xi,T}(f)(x) = \xi(x) f(Tx)$ is called a *weighted operator*.

Induced automorphism Assume that T is an automorphism of a standard probability Borel space $(X, \mathcal{B}, \mu)$. Let $A \in \mathcal{B}$, $\mu(A) > 0$. The *induced automorphism* T_A is defined on the conditional space $(A, \mathcal{B}_A, \mu_A)$, where $\mathcal{B}_A$ is the trace of $\mathcal{B}$ on A, $\mu_A(B) = \mu(B)/\mu(A)$ for $B \in \mathcal{B}_A$ and $T_A(x) = T^{k_A(x)}x$, where $k_A(x)$ is the smallest $k \geq 1$ for which $T^k x \in A$.

AT property of an automorphism An automorphism T of a standard probability Borel space $(X, \mathcal{B}, \mu)$ is called *approximatively transitive* (AT for short) if for every $\varepsilon > 0$ and every finite set $f_1, \ldots, f_n$ of non-negative L^1-functions on $(X, \mathcal{B}, \mu)$ we can find $f \in L^1(X, \mathcal{B}, \mu)$ also non-negative such that $\|f_i - \sum_j \alpha_{ij} f \circ T^{n_j}\|_{L_1} < \varepsilon$ for all $i = 1, \ldots, n$ (for some $\alpha_{ij} \geq 0$, $n_j \in \mathbb{N}$).

Cocycles and group extensions If T is an ergodic automorphism, G is a l.c.s.c. Abelian group and $\varphi \colon X \to G$ is measurable then the pair (T, φ) generates a *cocycle* $\varphi^{(\cdot)}(\cdot) \colon \mathbb{Z} \times X \to G$, where

$$\varphi^{(n)}(x) = \begin{cases} \varphi(x) + \ldots + \varphi(T^{n-1}x) & \text{for} \quad n > 0\,, \\ 0 & \text{for} \quad n = 0\,, \\ -(\varphi(T^n x) + \ldots + \varphi(T^{-1}x)) & \text{for} \quad n < 0\,. \end{cases}$$

(That is $(\varphi^{(n)})$ is a standard 1-cocycle in the algebraic sense for the $\mathbb{Z}$-action $n(f) = f \circ T^n$ on the group of measurable functions on X with values in G; hence the function $\varphi \colon X \to G$ itself is often called a *cocycle*.)

Assume additionally that G is compact. Using the cocycle φ we define a *group extension* T_φ on $(X \times G, \mathcal{B} \otimes \mathcal{B}(G), \mu \otimes \lambda_G)$ (λ_G stands for Haar measure of G), where $T_\varphi(x, g) = (Tx, \varphi(x) + g)$.

Special flow Given an ergodic automorphism T on a standard probability Borel space $(X, \mathcal{B}, \mu)$ and a positive integrable function $f \colon X \to \mathbb{R}^+$ we put

$$X^f = \{(x, t) \in X \times \mathbb{R} \colon 0 \leq t < f(x)\}\,,$$
$$\mathcal{B}^f = \mathcal{B} \otimes \mathcal{B}(\mathbb{R})|_{X^f}\,,$$

and define μ^f as normalized $\mu \otimes \lambda_{\mathbb{R}}|_{X^f}$. By a *special flow* we mean the $\mathbb{R}$-action $T^f = (T_t^f)_{t\in\mathbb{R}}$ under which a point $(x, s) \in X^f$ moves vertically with the unit speed, and once it reaches the graph of f, it is identified with $(Tx, 0)$.

Markov operator A linear operator $J\colon L^2(X, \mathcal{B}, \mu) \to L^2(Y, C, \nu)$ is called *Markov* if it sends non-negative functions to non-negative functions and $J1 = J^*1 = 1$.

Unitary actions on Fock spaces If H is a separable Hilbert space then by $H^{\odot n}$ we denote the subspace of n-tensors of $H^{\otimes n}$ symmetric under all permutations of coordinates, $n \geq 1$; then the Hilbert space $F(H) := \bigoplus_{n=0}^{\infty} H^{\odot n}$ is called a *symmetric Fock space*. If $V \in U(H)$ then $F(V) := \bigoplus_{n=0}^{\infty} V^{\odot n} \in U(F(H))$ where $V^{\odot n} = V^{\otimes n}|H^{\odot n}$.

Definition of the Subject

Spectral theory of dynamical systems is a study of special unitary representations, called Koopman representations (see the glossary). Invariants of such representations are called spectral invariants of measure-preserving systems. Together with the entropy, they constitute the most important invariants used in the study of measure-theoretic intrinsic properties and classification problems of dynamical systems as well as in applications. Spectral theory was originated by von Neumann, Halmos and Koopman in the 1930s. In this article we will focus on recent progresses in the spectral theory of finite measure-preserving dynamical systems.

Introduction

Throughout $\mathbb{A}$ denotes a non-compact l.c.s.c. Abelian group ($\mathbb{A}$ will be most often $\mathbb{Z}$ or $\mathbb{R}$). The assumption of second countability implies that $\mathbb{A}$ is metrizable, σ-compact and the space $C_0(\mathbb{A})$ is separable. Moreover the dual group $\widehat{\mathbb{A}}$ is also l.c.s.c. Abelian.

General Unitary Representations

We are interested in *unitary*, that is with values in the unitary group $U(H)$ of a Hilbert space H, (weakly) continuous representations $V\colon \mathbb{A} \ni a \mapsto V_a \in U(H)$ of such groups (the scalar valued maps $a \mapsto \langle V_a x, y\rangle$ are continuous for each $x, y \in H$).

Let $H = L^2(\widehat{\mathbb{A}}, \mathcal{B}(\widehat{\mathbb{A}}), \mu)$, where $\mathcal{B}(\widehat{\mathbb{A}})$ stands for the σ-algebra of Borel sets of $\widehat{\mathbb{A}}$ and $\mu \in M^+(\widehat{\mathbb{A}})$ (whenever X is a l.c.s.c. space, by $M(X)$ we denote the set of complex Borel measures on X, while $M^+(X)$ stands for the subset of positive (finite) measures). Given $a \in \mathbb{A}$, for $f \in L^2(\widehat{\mathbb{A}}, \mathcal{B}(\widehat{\mathbb{A}}), \mu)$ put

$$V_a^{\mu}(f)(\chi) = i(a)(\chi) \cdot f(\chi) = \chi(a) \cdot f(\chi) \quad (\chi \in \widehat{\mathbb{A}}),$$

where $i\colon \mathbb{A} \to \widehat{\widehat{\mathbb{A}}}$ is the canonical Pontriagin isomorphism of $\mathbb{A}$ with its second dual. Then $V^{\mu} = (V_a^{\mu})_{a\in\mathbb{A}}$ is a unitary representation of $\mathbb{A}$. Since $C_0(\widehat{\mathbb{A}})$ is dense in $L^2(\widehat{\mathbb{A}}, \mu)$, the latter space is separable. Therefore also direct sums $\bigoplus_{i=1}^{\infty} V^{\mu_i}$ of such type representations will be unitary representations of $\mathbb{A}$ in separable Hilbert spaces.

Lemma 1 (Wiener Lemma) *If $F \subset L^2(\widehat{\mathbb{A}}, \mu)$ is a closed V_a^{μ}-invariant subspace for all $a \in \mathbb{A}$ then $F = 1_Y L^2(\widehat{\mathbb{A}}, \mathcal{B}(\widehat{\mathbb{A}}), \mu)$ for some Borel subset $Y \subset \widehat{\mathbb{A}}$.*

Notice however that since $\mathbb{A}$ is not compact (equivalently, $\widehat{\mathbb{A}}$ is not discrete), we can find μ continuous and therefore V^{μ} has no irreducible (non-zero) subrepresentation. From now on only unitary representations of $\mathbb{A}$ in separable Hilbert spaces will be considered and we will show how to classify them.

A function $r\colon \mathbb{A} \to \mathbb{C}$ is called *positive definite* if

$$\sum_{n,m=0}^{N} r(a_n - a_m) z_n \overline{z_m} \geq 0 \tag{1}$$

for each $N > 0$, $(a_n) \subset \mathbb{A}$ and $(z_n) \subset \mathbb{C}$. The central result about positive definite functions is the following theorem (see e. g. [173]).

Theorem 1 (Bochner–Herglotz) *Let $r\colon \mathbb{A} \to \mathbb{C}$ be continuous. Then r is positive definite if and only if there exists (a unique) $\sigma \in M^+(\widehat{\mathbb{A}})$ such that*

$$r(a) = \int_{\widehat{\mathbb{A}}} \chi(a)\, \mathrm{d}\sigma(\chi) \quad \textit{for each} \quad a \in \mathbb{A}.$$

If now $\mathcal{U} = (U_a)_{a\in\mathbb{A}}$ is a representation of $\mathbb{A}$ in H then for each $x \in H$ the function $r(a) := \langle U_a x, x\rangle$ is continuous and satisfies (1), so it is positive definite. By the Bochner–Herglotz Theorem there exists a unique measure $\sigma_{\mathcal{U},x} = \sigma_x \in M^+(\widehat{\mathbb{A}})$ (called the *Spectral measure* of x) such that

$$\widehat{\sigma}_x(a) := \int_{\widehat{\mathbb{A}}} i(a)(\chi)\, \mathrm{d}\sigma_x(\chi) = \langle U_a x, x\rangle$$

for each $a \in \mathbb{A}$. Since the partial map $U_a x \mapsto i(a) \in L^2(\widehat{\mathbb{A}}, \sigma_x)$ is isometric and equivariant, there exists a unique extension of it to a unitary operator

$W: \mathbb{A}(x) \to L^2(\widehat{\mathbb{A}}, \sigma_x)$ giving rise to an isomorphism of $\mathcal{U}|_{\mathbb{A}(x)}$ and V^{σ_x}. Then the existence of a spectral decomposition is proved by making use of separability and a choice of maximal cyclic spaces at every step of an induction procedure. Moreover, a spectral decomposition is unique in the following sense.

Theorem 2 (Spectral Theorem) *If* $H = \bigoplus_{i=1}^{\infty} \mathbb{A}(x_i) = \bigoplus_{i=1}^{\infty} \mathbb{A}(x_i')$ *are two spectral decompositions of* H *then* $\sigma_{x_i} \equiv \sigma_{x_i'}$ *for each* $i \geq 1$.

It follows that the representation $\mathcal{U}$ is entirely determined by types (the sets of equivalent measures to a given one) of a decreasing sequence of measures or, equivalently, $\mathcal{U}$ is determined by its maximal spectral type $\sigma_{\mathcal{U}}$ and its multiplicity function $M_{\mathcal{U}}$.

Notice that eigenvalues of $\mathcal{U}$ correspond to Dirac measures: $\chi \in \widehat{\mathbb{A}}$ *is an eigenvalue* (i. e. *for some* $\|x\| = 1$, $U_a(x) = \chi(a)x$ *for each* $a \in \mathbb{A}$) *if and only if* $\sigma_{\mathcal{U},x} = \delta_\chi$. Therefore $\mathcal{U}$ has a discrete spectrum if and only if the maximal spectral type of $\mathcal{U}$ is a discrete measure.

We refer the reader to [64,105,127,147,156] for presentations of spectral theory needed in the theory of dynamical systems – such presentations are usually given for $\mathbb{A} = \mathbb{Z}$ but once we have the Bochner–Herglotz Theorem and the Wiener Lemma, their extensions to the general case are straightforward.

Koopman Representations

We will consider *measure-preserving* representations of $\mathbb{A}$. It means that we fix a probability standard Borel space $(X, \mathcal{B}, \mu)$ and by $Aut(X, \mathcal{B}, \mu)$ we denote the group of automorphisms of this space, that is $T \in Aut(X, \mathcal{B}, \mu)$ if $T: X \to X$ is a bimeasurable (a.e.) bijection satisfying $\mu(A) = \mu(TA) = \mu(T^{-1}A)$ for each $A \in \mathcal{B}$. Consider then a representation of $\mathbb{A}$ in $Aut(X, \mathcal{B}, \mu)$ that is a group homomorphism $a \mapsto T_a \in Aut(X, \mathcal{B}, \mu)$; we write $\mathcal{T} = (T_a)_{a \in \mathbb{A}}$. Moreover, we require that the associated Koopman representation $\mathcal{U}_{\mathcal{T}}$ is continuous. Unless explicitly stated, $\mathbb{A}$-actions are assumed to be *free*, that is we assume that for μ-a.e. $x \in X$ the map $a \mapsto T_a x$ is $1-1$. In fact, since constant functions are obviously invariant for U_{T_a}, that is the trivial character 1 is always an eigenvalue of $\mathcal{U}_{\mathcal{T}}$, the Koopman representation is considered only on the subspace $L_0^2(X, \mathcal{B}, \mu)$ of zero mean functions. We will restrict our attention only to *ergodic* dynamical systems (see the glossary). It is easy to see that $\mathcal{T}$ is ergodic if and only if whenever $A \in \mathcal{B}$ and $A = T_a(A)$ (μ-a.e.) for all $a \in A$ then $\mu(A)$ equals 0 or 1. In case of ergodic Koopman representations, all eigenvalues are simple. In particular, (ergodic) Koopman representations with discrete spectra have simple spectra. The reader is referred to monographs mentioned above as well as to [26,158,177,196,204] for basic facts on the theory of dynamical systems.

The passage $\mathcal{T} \mapsto \mathcal{U}_{\mathcal{T}}$ can be seen as functorial (contravariant). In particular a measure-theoretic isomorphism of $\mathbb{A}$-systems $\mathcal{T}$ and $\mathcal{T}'$ implies spectral isomorphism of the corresponding Koopman representations; hence spectral properties are measure-theoretic invariants. Since unitary representations are completely classified, one of the main questions in the spectral theory of dynamical systems is to decide which pairs $([\sigma], M)$ can be realized by Koopman representations. The most spectacular, still unsolved, is the Banach problem concerning $([\lambda_{\mathbb{T}}], M \equiv 1)$. Another important problem is to give complete spectral classification in some classes of dynamical systems (classically, it was done in the theory of Kolmogorov and Gaussian dynamical systems). We will also see how spectral properties of dynamical systems can determine their statistical (ergodic) properties; a culmination given by the fact that a spectral isomorphism may imply measure-theoretic similitude (discrete spectrum case, Gaussian–Kronecker case). We conjecture that a dynamical system $\mathcal{T}$ either is spectrally determined or there are uncountably many pairwise non-isomorphic systems spectrally isomorphic to $\mathcal{T}$.

We could also consider Koopman representations in L^p for $1 \leq p \neq 2$. However whenever $W: L^p(X, \mathcal{B}, \mu) \to L^p(Y, C, \nu)$ is a surjective isometry and $W \circ U_{T_a} = U_{S_a} \circ W$ for each $a \in \mathbb{A}$ then by the Lamperti Theorem (e. g. [172]) the isometry W has to come from a nonsingular pointwise map $R: Y \to X$ and, by ergodicity, R "preserves" the measure ν and hence establishes a measure-theoretic isomorphism [94] (see also [127]). Thus spectral classification of such Koopman representations goes back to the measure-theoretic classification of dynamical systems, so it looks hardly interesting. This does not mean that there are no interesting dynamical questions for $p \neq 2$. Let us mention still open Thouvenot's question (formulated in the 1980s) for $\mathbb{Z}$-actions: *For every ergodic* T *acting on* $(X, \mathcal{B}, \preceq)$, *does there exist* $f \in L^1(X, \mathcal{B}, \mu)$ *such that the closed linear span of* $f \circ T^n$, $n \in \mathbb{Z}$ *equals* $L^1(X, \mathcal{B}, \mu)$?

Iwanik [79,80] proved that if T is a system with positive entropy then its L^p-multiplicity is ∞ for all $p > 1$. Moreover, Iwanik and de Sam Lazaro [85] proved that for Gaussian systems (they will be considered in Sect. "Spectral Theory of Dynamical Systems of Probabilistic Origin") the L^p-multiplicities are the same for all $p > 1$ (see also [137]).

Markov Operators, Joinings and Koopman Representations, Disjointness and Spectral Disjointness, Entropy

We would like to emphasize that spectral theory is closely related to the theory of joinings (see ▶ Joinings in Ergodic Theory for needed definitions). The elements ρ of the set $J(S,\mathcal{T})$ of joinings of two $\mathbb{A}$-actions S and $\mathcal{T}$ are in a 1-1 correspondence with Markov operators $J = J_\rho$ between the L^2-spaces equivariant with the corresponding Koopman representations (see the glossary and ▶ Joinings in Ergodic Theory). The set of ergodic self-joinings of an ergodic $\mathbb{A}$-action $\mathcal{T}$ is denoted by $J_2^e(\mathcal{T})$.

Each Koopman representation $\mathcal{U}_\mathcal{T}$ consists of Markov operators (indeed, U_{T_a} is clearly a Markov operator). In fact, if $U \in U(L^2(X,\mathcal{B},\mu))$ is Markov then it is of the form U_R, where $R \in Aut(X,\mathcal{B},\mu)$ [133]. This allows us to see Koopman representations as unitary Markov representations, but since a spectral isomorphism does not "preserve" the set of Markov operators, spectrally isomorphic systems can have drastically different sets of self-joinings.

We will touch here only some aspects of interactions (clearly, far from completeness) between the spectral theory and the theory of joinings.

In order to see however an example of such interactions let us recall that the simplicity of eigenvalues for ergodic systems yields a short "joining" proof of the classical isomorphism theorem of Halmos-von Neumann in the discrete spectrum case: *Assume that* $S = (S_\dashv)_{\dashv\in\mathbb{A}}$ *and* $\mathcal{T} = (T_a)_{a\in\mathbb{A}}$ *are ergodic* $\mathbb{A}$*-actions on* $(X,\mathcal{B},\mu)$ *and* (Y,C,ν) *respectively. Assume that both Koopman representations have purely discrete spectrum and that their group of eigenvalues are the same. Then* S *and* $\mathcal{T}$ *are measure-theoretically isomorphic*. Indeed, each ergodic joining of $\mathcal{T}$ and S is the graph of an isomorphism of these two systems (see [127]; see also Goodson's proof in [66]). Another example of such interactions appear in the study Rokhlin's multiple mixing problem and its relation with the *pairwise independence property* (PID) for joinings of higher order. We will not deal with this subject here, referring the reader to ▶ Joinings in Ergodic Theory (see however Sect. "Lifting Mixing Properties").

Following [60], two $\mathbb{A}$-actions S and $\mathcal{T}$ are called *disjoint* provided the product measure is the only element in $J(S,\mathcal{T})$. It was already noticed in [72] that spectrally disjoint systems are disjoint in the Furstenberg sense; indeed, $Im(J_\rho|_{L_0^2}) = \{0\}$ since $\sigma_{\mathcal{T},J_\rho f} \ll \sigma_{S,f}$.

Notice that whenever we take $\rho \in J_2^e(\mathcal{T})$ we obtain a new ergodic $\mathbb{A}$-action $(T_a \times T_a)_{a\in\mathbb{A}}$ defined on the probability space $(X\times X,\rho)$. One can now ask how close spectrally to $\mathcal{T}$ is this new action? It turns out that except of the obvious fact that the marginal σ-algebras are factors, $(\mathcal{T}\times\mathcal{T},\rho)$ can have other factors spectrally disjoint with $\mathcal{T}$: the most striking phenomenon here is a result of Smorodinsky and Thouvenot [198] (see also [29]) saying that each zero entropy system is a factor of an ergodic self-joining system of a fixed Bernoulli system (Bernoulli systems themselves have countable Haar spectrum). The situation changes if $\rho = \mu\otimes\mu$. In this case for $f,g \in L^2(X,\mathcal{B},\mu)$ the spectral measure of $f\otimes g$ is equal to $\sigma_{\mathcal{T},f} * \sigma_{\mathcal{T},g}$. A consequence of this observation is that an ergodic action $\mathcal{T} = (T_a)_{a\in\mathbb{A}}$ is weakly mixing (see the glossary) if and only if the product measure $\mu\otimes\mu$ is an ergodic self-joining of $\mathcal{T}$.

The entropy which is a basic measure-theoretic invariant does not appear when we deal with spectral properties. We will not give here any formal definition of entropy for amenable group actions referring the reader to [153]. Assume that $\mathbb{A}$ is countable and discrete. We always assume that $\mathbb{A}$ is Abelian, hence it is amenable. For each dynamical system $\mathcal{T} = (T_a)_{a\in\mathbb{A}}$ acting on $(X,\mathcal{B},\mu)$, we can find a largest invariant sub-σ field $\mathcal{P}\subset\mathcal{B}$, called the *Pinsker* σ-algebra, such that the entropy of the corresponding quotient system is zero. Generalizing the classical Rokhlin-Sinai Theorem (see also [97] for $\mathbb{Z}^d$-actions), Thouvenot (unpublished) and independently Dooley and Golodets [31] proved this theorem for groups even more general than those considered here: *The spectrum of* $\mathcal{U}_\mathcal{T}$ *on* $L^2(X,\mathcal{B},\mu)\ominus L^2(\mathcal{P})$ *is Haar with uniform infinite multiplicity*. This general result is quite intricate and based on methods introduced to entropy theory by Rudolph and Weiss [179] with a very surprising use of Dye's Theorem on orbital equivalence of all ergodic systems. For $\mathbb{A}$ which is not countable the same result was recently proved in [17] in case of unimodular amenable groups which are not increasing union of compact subgroups. It follows that spectral theory of dynamical systems essentially reduces to the zero entropy case.

Maximal Spectral Type of a Koopman Representation, Alexeyev's Theorem

Only few general properties of maximal spectral types of Koopman representations are known. The fact that a Koopman representation preserves the space of real functions implies that its maximal spectral type is the type of a symmetric (invariant under the map $\chi\mapsto\overline{\chi}$) measure.

Recall that the *Gelfand spectrum* $\sigma(\mathcal{U})$ of a representation $\mathcal{U} = (U_a)_{a\in\mathbb{A}}$ is defined as the of *approximative eigenvalues* of $\mathcal{U}$, that is $\sigma(\mathcal{U}) \ni \chi \in \widehat{\mathbb{A}}$ if for a sequence (x_n) bounded and bounded away from zero, $\|U_a x_n - \chi(a)x_n\| \to 0$ for each $a\in\mathbb{A}$. The spectrum

is a closed subset in the topology of pointwise convergence, hence in the compact-open topology of $\widehat{\mathbb{A}}$. In case of $\mathbb{A} = \mathbb{Z}$, the above set $\sigma(U)$ is equal to $\{z \in \mathbb{C} : U - z \cdot Id \text{ is not invertible}\}$.

Assume now that $\mathbb{A}$ is countable and discrete (and Abelian). Then there exists a Fölner sequence $(B_n)_{n\geq 1}$ whose elements tile $\mathbb{A}$ [153]. Take a free and ergodic action $\mathcal{T} = (T_a)_{a\in\mathbb{A}}$ on $(X, \mathcal{B}, \mu)$. By [153] for each $\varepsilon > 0$ we can find a set $Y_n \in \mathcal{B}$ such that the sets $T_b Y_n$ are pairwise disjoint for $b \in B_n$ and $\mu(\bigcup_{b\in B_n} T_b Y_n) > 1 - \varepsilon$. For each $\chi \in \widehat{\mathbb{A}}$, by considering functions of the form $f_n = \sum_{b\in B_n} \chi(b) 1_{T_b Y_n}$ we obtain that $\chi \in \sigma(\mathcal{U}_\mathcal{T})$. It follows that the topological support of the maximal spectral type of the Koopman representation of a free and ergodic action is full [105,127,147]. The theory of Gaussian systems shows in particular that there are symmetric measures on the circle whose topological support is the whole circle but which cannot be maximal spectral types of Koopman representations.

An open well-known question remains whether an absolutely continuous measure ρ is the maximal spectral type of a Koopman representation if and only if ρ is equivalent to a Haar measure of $\widehat{\mathbb{A}}$ (this is unknown for $\mathbb{A} = \mathbb{Z}$).

Another interesting question was recently raised by A. Katok (private communication): *Is it true that the topological supports of all measures in a spectral sequence of a Koopman representation are full*? If the answer to this question is positive then for example the essential supremum of $M_{\mathcal{U}_\mathcal{T}}$ is the same on all balls of $\widehat{\mathbb{A}}$.

Notice that the fact that all spectral measures in a spectral sequence are symmetric means that $\mathcal{U}_\mathcal{T}$ is isomorphic to $\mathcal{U}_{\mathcal{T}^{-1}}$. A. del Junco [89] showed that generically for $\mathbb{Z}$-actions, T and its inverse are not measure-theoretically isomorphic (in fact he proved disjointness).

Let $\mathcal{T}$ be an $\mathbb{A}$-action on $(X, \mathcal{B}, \mu)$. One can ask wether a "good" function can realize the maximal spectral type of $\mathcal{U}_\mathcal{T}$. In particular can we find a function $f \in L^\infty(X, \mathcal{B}, \mu)$ that realizes the maximal spectral type? The answer is given in the following general theorem (see [139]).

Theorem 3 (Alexeyev's Theorem) *Assume that $\mathcal{U} = (U_a)_{a\in\mathbb{A}}$ is a unitary representation of $\mathbb{A}$ in a separable Hilbert space H. Assume that $F \subset H$ is a dense linear subspace. Assume moreover that with some F-norm $|\!|\cdot|\!|$ – stronger than the norm $\|\cdot\|$ given by the scalar product – F becomes a Fréchet space. Then, for each spectral measure σ ($\ll \sigma_\mathcal{U}$) there exists $y \in F$ such that $\sigma_y \gg \sigma$. In particular, there exists $y \in F$ that realizes the maximal spectral type.*

By taking $H = L^2(X, \mathcal{B}, \mu)$ and $F = L^\infty(X, \mathcal{B}, \mu)$ we obtain the positive answer to the original question. Alexeyev [14] proved the above theorem for $\mathbb{A} = \mathbb{Z}$ using analytic functions. Refining Alexeyev's original proof, Frączek [52] showed the existence of a sufficiently regular function realizing the maximal spectral type depending only on the "regularity" of the underlying probability space, e. g. when X is a compact metric space (compact manifold) then one can find a continuous (smooth) function realizing the maximal spectral type.

By the theory of systems of probabilistic origin (see Sect. "Spectral Theory of Dynamical Systems of Probabilistic Origin"), in case of simplicity of the spectrum, one can easily point out spectral measures whose types are not realized by (essentially) bounded functions. However, it is still an open question whether for each Koopman representation $\mathcal{U}_\mathcal{T}$ there exists a sequence $(f_i)_{i\geq 1} \subset L^\infty(X, \mathcal{B}, \mu)$ such that the sequence $(\sigma_{f_i})_{i\geq 1}$ is a spectral sequence for $\mathcal{U}_\mathcal{T}$. For $\mathbb{A} = \mathbb{Z}$ it is unknown whether the maximal spectral type of a Koopman representation can be realized by a characteristic function.

Spectral Theory of Weighted Operators

We now pass to the problem of possible essential values for the multiplicity function of a Koopman representation. However, one of known techniques is a use of cocycles, so before we tackle the multiplicity problem, we will go through recent results concerning spectral theory of compact group extensions automorphisms which in turn entail a study of weighted operators (see the glossary).

Assume that T is an ergodic automorphism of a standard Borel probability space $(X, \mathcal{B}, \mu)$. Let $\xi : X \to \mathbb{T}$ be a measurable function and let $V = V_{\xi,T}$ be the corresponding weighted operator. To see a connection of weighted operators with Koopman representations of compact group extensions consider a compact (metric) Abelian group G and a cocycle $\varphi : X \to G$. Then U_{T_φ} (see the glossary) acts on $L^2(X \times G, \mu \otimes \lambda_G)$. But

$$L^2(X\times G, \mu\otimes\lambda_G) = \bigoplus_{\chi\in\widehat{G}} L_\chi, \quad \text{where } L_\chi = L^2(X, \mu)\otimes\chi,$$

where L_χ is a U_{T_φ}-invariant (clearly, closed) subspace. Moreover, the map $f \otimes \chi \mapsto f$ settles a unitary isomorphism of $U_{T_\varphi}|_{L_\chi}$ with the operator $V_{\chi\circ\varphi,T}$. Therefore, spectral analysis of such Koopman representations reduces to the spectral analysis of weighted operators $V_{\chi\circ\varphi,T}$ for all $\chi \in \widehat{G}$.

Maximal Spectral Type of Weighted Operators over Rotations

The spectral analysis of weighted operators $V_{\xi,T}$ is especially well developed in case of rotations, i. e. when

we additionally assume that T is an ergodic rotation on a compact monothetic group X: $Tx = x + x_0$, where x_0 is a topologically cyclic element of X (and μ will stand for Haar measure λ_x of X). In this case Helson's analysis [74] applies (see also [68,82,127,160]) leading to the following conclusions:

- The maximal spectral type $\sigma_{V_{\xi,T}}$ is either discrete or continuous.
- When $\sigma_{V_{\xi,T}}$ is continuous it is either singular or Lebesgue.
- The spectral multiplicity of $V_{\xi,T}$ is uniform.

We now pass to a description of some results in case when $Tx = x + \alpha$ is an irrational rotation on the additive circle $X = [0, 1)$. It was already noticed in the original paper by Anzai [16] that when $\xi\colon X \to \mathbb{T}$ is an affine cocycle ($\xi(x) = \exp(nx + c)$, $0 \neq n \in \mathbb{Z}$) then $V_{\xi,T}$ has a Lebesgue spectrum. It was then considered by several authors (originated by [123], see also [24,26]) to which extent this property is stable when we perturb our cocycle. Since the topological degree of affine cocyles is different from zero, when perturbing them we consider smooth perturbations by cocycles of degree zero.

Theorem 4 ([82]) *Assume that $Tx = x + \alpha$ is an irrational rotation. If $\xi\colon [0, 1) \to \mathbb{T}$ is of non-zero degree, absolutely continuous, with the derivative of bounded variation then $V_{\xi,T}$ has a Lebesgue spectrum.*

In the same paper, it is noticed that if we drop the assumption on the derivative then the maximal spectral type of $V_{\xi,T}$ is a Rajchman measure (i. e. its Fourier transform vanishes at infinity). It is still an open question, whether one can find ξ absolutely continuous with non-zero degree and such that $V_{\xi,T}$ has singular spectrum. "Below" absolute continuity, topological properties of the cocycle seem to stop playing any role – in [82] a continuous, degree 1 cocycle ξ of bounded variation is constructed such that $\xi(x) = \eta(x)/\eta(Tx)$ for a measurable $\eta\colon [0, 1) \to \mathbb{T}$ (that is ξ is a *coboundary*) and therefore $V_{\xi,T}$ has purely discrete spectrum (it is isomorphic to U_T). For other results about Lebesgue spectrum for Anzai skew products see also [24,53,81] (in [53] $\mathbb{Z}^d$-actions of rotations and so called winding numbers instead of topological degree are considered).

When the cocycle is still smooth but its degree is zero the situation drastically changes. Given an absolutely continuous function $f\colon [0, 1) \to \mathbb{R}$ M. Herman [76], using the Denjoy–Koksma inequality (see e. g. [122]), showed that $f_0^{(q_n)} \to 0$ uniformly (here $f_0 = f - \int_0^1 f \, d\lambda_{[0,1)}$ and (q_n) stands for the sequence of denominators of α). It follows that $T_{e^{2\pi i f}}$ is rigid and hence has a singular spectrum. B. Fayad [37] shows that this result is no longer true if one dimensional rotation is replaced by a multi-dimensional rotation (his counterexample is in the analytic class). See also [130] for the singularity of spectrum for functions f whose Fourier transform satisfies $o(1/|n|)$ condition or to [84], where it is shown that sufficiently small variation implies singularity of the spectrum.

A natural class of weighted operators arises when we consider Koopman operators of rotations on nil-manifolds. We only look at the particular example of such a rotation on a quotient of the Heisenberg group $(\mathbb{R}^3, *)$ (a general spectral theory of nil-actions was mainly developed by W. Parry [157]) – these actions have countable Lebesgue spectrum in the orthocomplement of the subspace of eigenfunctions) that is take the nil-manifold $\mathbb{R}^3/_*\mathbb{Z}^3$ on which we define the nil-rotation $S((x, y, z) * \mathbb{Z}^3) = (\alpha, \beta, 0)*(x, y, z)*\mathbb{Z}^3 = \big(x + \alpha, y + \beta, z + \alpha y\big)*\mathbb{Z}^3$, where α, β and 1 are rationally independent. It can be shown that S is isomorphic to the skew product defined on $[0, 1)^2 \times \mathbb{T}$ by

$$T_\varphi\colon (x, y, z) \mapsto \Big(x + \alpha, y + \beta, z \cdot e^{2\pi i \varphi(x,y)}\Big) \ ,$$

where $\varphi(x, y) = \alpha y - \psi(x + \alpha, y + \beta) + \psi(x, y)$ with $\psi(x, y) = x[y]$. Since nil-cocycles can be considered as a certain analog of affine cocycles for one-dimensional rotations, it would be nice to explain to what kind of perturbations the Lebesgue spectrum property is stable.

Yet another interesting problem which is related to the spectral theory of extensions given by so called *Rokhlin cocycles* (see Sect. "Rokhlin Cocycles") arises, when given $f\colon [0, 1) \to \mathbb{R}$, we want to describe spectrally the one-parameter set of weighted operators $W_c := V_{e^{2\pi i c f},T}$; here T is a fixed irrational rotation by α. As proved by quite sophisticated arguments in [84], if we take $f(x) = x$ then for all non-integer $c \in \mathbb{R}$ the spectrum of W_c is continuous and singular (see also [68] and [145] where some special α's are considered). It has been open for some time if at all one can find $f\colon [0, 1) \to \mathbb{R}$ such that for each $c \neq 0$, the operator W_c has a Lebesgue spectrum. The positive answer is given in [205]: for example if $f(x) = x^{-(2+\varepsilon)}$ ($\varepsilon > 0$) and α has bounded partial quotients then W_c has a Lebesgue spectrum for all $c \neq 0$. All functions with such a property considered in [205] are non-integrable. It would be interesting to find an integrable f with the above property.

We refer to [66] and the references therein for further results especially for transformations of the form $(x, y) \mapsto (x+\alpha, 1_{[0,\beta)}(x)+y)$ on $[0, 1)\times\mathbb{Z}/2\mathbb{Z}$. Recall however that earlier Katok and Stepin [104] used this kind of transfor-

mations to give a first counterexample to the Kolmogorov group property (see the glossary) for the spectrum.

The Multiplicity Problem for Weighted Operators over Rotations

In case of perturbations of affine cocycles, this problem was already raised by Kushnirenko [123]. Some significant results were obtained by M. Guenais. Before we state her results let us recall a useful criterion to find an upper bound for the multiplicity: *If there exist $c > 0$ and a sequence $(F_n)_{n\geq 1}$ of cyclic subspaces of H such that for each $y \in H$, $\|y\| = 1$ we have $\liminf_{n\to\infty} \|proj_{F_n} y\|^2 \geq c$, then $esssup(M_U) \leq 1/c$* which follows from a well-known lemma of Chacon [23,26,111,127]. Using this and a technique which is close to the idea of local rank one (see [44,111]) M. Guenais [69] proved a series of results on multiplicity which we now list.

Theorem 5 *Assume that $Tx = x + \alpha$ and let $\xi\colon [0,1) \to \mathbb{T}$ be a cocycle.*

(i) If $\xi(x) = e^{2\pi i c x}$ then $M_{V_{\xi,T}}$ is bounded by $|c| + 1$.
(ii) If ξ is absolutely continuous and ξ is of topological degree zero, then $V_{\xi,T}$ has a simple spectrum.
(iii) if ξ is of bounded variation, then $M_{V_{\xi,T}} \leq \max(2, 2\pi\, Var(\xi)/3)$.

Remarks on the Banach Problem

We already mentioned in Introduction the Banach problem in ergodic theory, which is simply the question whether there exists a Koopman representation for $\mathbb{A} = \mathbb{Z}$ with simple Lebesgue spectrum. In fact no example of a Koopman representation with Lebesgue spectrum of finite multiplicity is known. Helson and Parry [75] asked for the validity of a still weaker version: *Can one construct T such that U_T has a Lebesgue component in its spectrum whose multiplicity is finite*? Quite surprisingly in [75] they give a general construction yielding for each ergodic T a cocycle $\varphi\colon X \to \mathbb{Z}/2\mathbb{Z}$ such that the unitary operator U_{T_φ} has a Lebesgue spectrum in the orthocomplement of functions depending only on the X-coordinate. Then Mathew and Nadkarni [144] gave examples of cocycles over so called dyadic adding machine for which the multiplicity of the Lebesgue component was equal to 2. In [126] this was generalized to so called *Toeplitz $\mathbb{Z}/2\mathbb{Z}$-extensions* of adding machines: for each even number k we can find a two-point extension of an adding machine so that the multiplicity of the Lebesgue component is k. Moreover, it was shown that Mathew and Nadkarni's constructions from [144] in fact are close to systems arising from number theory (like the famous Rudin–Shapiro sequence, e. g. [160]), relating the result about multiplicity of the Lebesgue component to results by Kamae [96] and Queffelec [160]. Independently of [126], Ageev [8] showed that one can construct 2-point extensions of the Chacon transformation realizing (by taking powers of the extension) each even number as the multiplicity of the Lebesgue component. Contrary to all previous examples, Ageev's constructions are weakly mixing.

Still an open question remains whether for $\mathbb{A} = \mathbb{Z}$ one can find a Koopman representation with the Lebesgue component of multiplicity 1 (or even odd).

In [70], M. Guenais studies the problem of Lebesgue spectrum in the classical case of Morse sequences (see [107] as well as [124], where the problem of spectral classification in this class is studied). All dynamical sytems arising from Morse sequences have simple spectra [124]. It follows that if a Lebesgue component appears in a Morse dynamical system, it has multiplicity one. Guenais [70] using a Riesz product technique relates the Lebesgue spectrum problem with the still open problem of whether a construction of "flat" trigonometric polynomials with coefficients ± 1 is possible. However, it is proved in [70] that such a construction can be carried out on some compact Abelian groups and it leads, for an Abelian countable torsion group $\mathbb{A}$, to a construction of an ergodic action of $\mathbb{A}$ with simple spectrum and a Haar component in its spectrum.

Lifting Mixing Properties

We now give one more example of interactions between spectral theory and joinings (see Introduction) that gives rise to a quick proof of the fact that r-fold mixing property of T ($r \geq 2$) lifts to a weakly mixing compact group extension T_φ (the original proof of this fact is due to D. Rudolph [175]). Indeed, to prove r-fold mixing for a mixing($=$2-mixing) transformation S (acting on (Y, C, ν)) one has to prove that each weak limit of off-diagonal self-joinings (given by powers of S, see ► Joinings in Ergodic Theory) of order r is simply the product measure $\nu^{\otimes r}$. We need also a Furstenberg's lemma [62] about relative unique ergodicity (RUE) of compact group extensions T_φ: *If $\mu \otimes \lambda_G$ is an ergodic measure for T_φ then it is the only (ergodic) invariant measure for T_φ whose projection on the first coordinate is μ.* Now the result about lifting r-fold mixing to compact group extensions follows directly from the fact that whenever T_φ is weakly mixing, $(\mu \otimes \lambda_G)^{\otimes r}$ is an ergodic measure (this approach was shown to me by A. del Junco). In particular if T is mixing and T_φ is weakly

mixing then for each $\chi \in \widehat{G} \setminus \{1\}$, the maximal spectral type of $V_{\chi\circ\varphi,T}$ is Rajchman.

See Sect. "Rokhlin Cocycles" for a generalization of the lifting result to Rokhlin cocycle extensions.

The Multiplicity Function

In this chapter only $\mathbb{A} = \mathbb{Z}$ is considered (for other groups, even for $\mathbb{R}$, much less is known; see however the case of so called *product* $\mathbb{Z}^d$*-actions* [50]). Contrary to the case of maximal spectral type, it is rather commonly believed that there are no restrictions for the set of essential values of Koopman representations.

Cocycle Approach

We will only concentrate on some results of the last twenty years. In 1983, E.A. Robinson [164] proved that for each $n \geq 1$ there exists an ergodic transformation whose maximal spectral multiplicity is n. Another important result was proved in [165] (see also [98]), where it is shown that given a finite set $M \subset \mathbb{N}$ containing 1 and closed under the least common multiple one can find (even a weakly mixing) T so that the set of essential values of the multiplicity function equals M. This result was then extended in [67] to infinite sets and finally in [125] (see also [11]) to all subsets $M \subset \mathbb{N}$ containing 1. In fact, as we have already noticed in the previous section the spectral theory for compact Abelian group extensions is reduced to a study of weighted operators and then to comparing maximal spectral types for such operators. This leads to sets of the form

$$M(G, v, H) = \Big\{\sharp(\{\chi \circ v^i : i \in \mathbb{Z}\} \cap anih(H)) : \chi \in anih(H)\Big\}$$

($H \subset G$ is a closed subgroup and v is a continuous group automorphism of G). Then an algebraic lemma has been proved in [125] saying that each set M containing 1 is of the form $M(G, v, H)$ and the techniques to construct "good" cocycles and a passage to "natural factors" yielded the following: *For each* $M \subset \{1, 2, \ldots\} \subset \{\infty\}$ *containing 1 there exists an ergodic automorphim such that the set of essential values for its Koopman representation equals* M. See also [166] for the case of non-Abelian group extensions.

A similar in spirit approach (that means, a passage to a family of factors) is present in a recent paper of Ageev [13] in which he first applies Katok's analysis (see [98,102]) for spectral multiplicities of the Koopman representation associated with Cartesian products $T^{\times k}$ for a generic transformation T. In a natural way this approach leads to study multiplicities of tensor products of unitary operators. Roughly, the multiplicity is computed as the number of atoms (counted modulo obvious symmetries) for conditional measures (see [98]) of a product measure over its convolution. Ageev [13] proved that for a typical automorphism T the set of the values of the multiplicity function for $U_{T^{\times k}}$ equals $\{k, k(k-1), \ldots, k!\}$ and then he just passes to "natural" factors for the Cartesian products by taking sets invariant under a fixed subgroup of permutations of coordinates. In particular, he obtains all sets of the form $\{2, 3, \ldots, n\}$ on L^2_0. He also shows that such sets of multiplicities are realizable in the category of mixing transformations.

Rokhlin's Uniform Multiplicity Problem

The Rokhlin multiplicity problem (see the recent book by Anosov [15]) was, given $n \geq 2$, to construct an ergodic transformation with uniform multiplicity n on L^2_0. A solution for $n = 2$ was independently given by Ageev [9] and Ryzhikov [188] (see also [15] and [66]) and in fact it consists in showing that for some T (actually, any T with simple spectrum for which $1/2(Id + U_T)$ is in the weak operator closure of the powers of U_T will do) the multiplicity of $T \times T$ is uniformly equal to 2 (see also Sect. "Future Directions").

In [12], Ageev proposed a new approach which consists in considering actions of "slightly non-Abelian" groups; and showing that for a "typical" action of such a group a fixed "direction" automorphism has a uniform multiplicity. Shortly after publication of [12], Danilenko [27], following Ageev's approach, considerably simplified the original proof. We will present Danilenko's arguments.

Fix $n \geq 1$. Denote $\overline{e}_i = (0, \ldots, 1, \ldots, 0) \in \mathbb{Z}^n$, $i = 1, \ldots, n$. We define an automorphism L of $\mathbb{Z}^n$ setting $L(\overline{e}_i) = \overline{e}_{i+1}$, $i = 1, \ldots, n-1$ and $L(\overline{e}_n) = \overline{e}_1$. Using L we define a semi-direct product $G := \mathbb{Z}^n \rtimes \mathbb{Z}$ defining multiplication as $(u, k) \cdot (w, l) = (u + L^k w, k + l)$. Put $e_0 = (0, 1)$, $e_i = (\overline{e}_i, 0)$, $i = 1, \ldots, n$ (and $Le_i = (L\overline{e}_i, 0)$). Moreover, denote $e_{n+1} = e_0^n = (0, n)$. Notice that $e_0 \cdot e_i \cdot e_0^{-1} = Le_i$ for $i = 1, \ldots, n$ ($L(e_{n+1}) = e_{n+1}$).

Theorem 6 (Ageev, Danilenko) *For every unitary representation* $\mathcal{U}$ *of* G *in a separable Hilbert space* H*, for which* $U_{e_1 - L^r e_1}$ *has no non-trivial fixed points for* $1 \leq r < n$*, the essential values of the multiplicity function for* $U_{e_{n+1}}$ *are contained in the set of multiples of* n*. If, in addition,* U_{e_0} *has a simple spectrum, then* $U_{e_{n+1}}$ *has uniform multiplicity* n*.*

It is then a certain work to show that the assumption of the second part of the theorem is satisfied for a typical action of the group G. Using a special (C, F)-construction with all the cut-and-stack parameters explicit Danilenko [27] was also able to show that each set of the form $k \cdot M$, where $k \geq 1$ and M is an arbitrary subset of natural numbers containing 1, is realizable as the set of essential values of a Koopman representation.

Some other constructions based on the solution of the Rokhlin problem for $n = 2$ and the method of [125] are presented in [103] leading to sets different than those pointed above; these sets contain 2 as their minimum.

Rokhlin Cocycles

We consider now a certain class of extensions which should be viewed as a generalization of the concept of compact group extensions. We will focus on $\mathbb{Z}$-actions only.

Assume that T is an ergodic automorphism of $(X, \mathcal{B}, \mu)$. Let G be a l.c.s.c. Abelian group. Assume that this group acts on (Y, C, ν), that is we have a G-action $S = (S_g)_{g \in G}$ on (Y, C, ν). Let $\varphi\colon X \to G$ be a cocycle. We then define an automorphism $T_{\varphi,S}$ of the space $(X \times Y, \mathcal{B} \otimes C, \mu \otimes \nu)$ by

$$T_{\varphi,S}(x, y) = (Tx, S_{\varphi(x)}(y)).$$

Such an extension is called a *Rokhlin cocycle extension* (the map $x \mapsto S_{\varphi(x)}$ is called a *Rokhlin cocycle*). Such an operation generalizes the case of compact group extensions; indeed, when G is compact the action of G on itself by rotations preserves Haar measure. (It is quite surprising, that when only we admit G non-Abelian, then, as shown in [28], **each** ergodic extension of T has a form of a Rokhlin cocycle extension.) Ergodic and spectral properties of such extensions are examined in several papers: [63,65,129,131,132,133,167,176]. Since in these papers rather joining aspects are studied (among other things in [129] Furstenberg's RUE lemma is generalized to this new context), we will mention here only few results, mainly spectral, following [129] and [133]. We will constantly assume that G is non-compact. As $\varphi\colon X \to G$ is then a cocycle with values in a non-compact group, the theory of such cocycles is much more complicated (see e. g. [193]), and in fact the theory of Rokhlin cocycle extensions leads to interesting interactions between classical ergodic theory, the theory of cocycles and the theory of non-singular actions arising from cocycles taking values in non-compact groups – especially, the Mackey action associated to φ plays a crucial role here (see the problem of invariant measures for $T_{\varphi,S}$ in [132] and [28]); see also monographs [1,98,101,193]. Especially, two Borel subgroups of $\widehat{G}$ are important here:

$$\Sigma_\varphi = \{\chi \in \widehat{G}\colon \chi \circ \varphi = c \cdot \xi / \xi \circ T \text{ for a measurable } \xi\colon X \to \mathbb{T} \text{ and } c \in \mathbb{T}\}.$$

and its subgroup Λ_φ given by $c = 1$. Λ_φ turns out to be the group of L^∞-eigenvalues of the Mackey action (of G) associated to the cocycle φ. This action is the quotient action of the natural action of G (by translations on the second coordinate) on the space of ergodic components of the skew product T_φ – the Mackey action is (in general) not measure-preserving, it is however non-singular. We refer the reader to [2,78,147] for other properties of those subgroups.

Theorem 7 ([132,133])

(i) $\sigma_{T_{\varphi,S}}|_{L^2(X\times Y,\mu\otimes\nu)\ominus L^2(X,\mu)} = \int_{\hat{G}} \sigma_{V_{\chi\circ\varphi,T}}\, \mathrm{d}\sigma_S$.

(ii) *$T_{\varphi,S}$ is ergodic if and only if T is ergodic and $\sigma_S(\Lambda_\varphi) = 0$.*

(iii) *$T_{\varphi,S}$ is weakly mixing if and only if T is weakly mixing and S has no eigenvalues in Σ_φ.*

(iv) *if T is mixing, S is mildly mixing, φ is recurrent and not cohomologous to a cocycle with values in a compact subgroup of G then $T_{\varphi,S}$ remains mixing.*

(v) *If T is r-fold mixing, φ is recurrent and $T_{\varphi,S}$ is mildly mixing then $T_{\varphi,S}$ is also r-fold mixing.*

(vi) *If T and R are disjoint, the cocycle φ is ergodic and S is mildly mixing then $T_{\varphi,S}$ remains disjoint with R.*

Let us recall [61,195] that an $\mathbb{A}$-action $S = (S_a)_{a\in\mathbb{A}}$ is mildly mixing (see the glossary) if and only if the $\mathbb{A}$-action $(S_a \times \tau_a)_{a\in\mathbb{A}}$ remains ergodic for every properly ergodic non-singular $\mathbb{A}$-action $\tau = (\tau_a)_{a\in\mathbb{A}}$.

Coming back to Smorodinsky–Thouvenot's result about factors of ergodic self-joinings of a Bernoulli automorphism we would like to emphasize here that the disjointness result (*vi*) above was used in [132] to give an example of an automorphism which is disjoint from all weakly mixing transformations but which has an ergodic self-joining whose associated automorphism has a non-trivial weakly mixing factor. In a sense this is opposed to Smorodinsky–Thouvenot's result as here from self-joinings we produced a "more complicated" system (namely the weakly mixing factor) than the original system.

It would be interesting to develop the theory of spectral multiplicity for Rokhlin cocycle extensions as it was done in the case of compact group extensions. However a difficulty is that in the compact group extension case we deal with a countable direct sum of representations of the form

$V_{\chi\circ\varphi,T}$ while in the non-compact case we have to consider an integral of such representations.

Rank-1 and Related Systems

Although properties like mixing, weak (and mild) mixing as well as ergodicity, are clearly spectral properties, they have "good" measure-theoretic formulations (expressed by a certain behavior on sets). Simple spectrum property is another example of a spectral property, and it was a popular question in the 1980s whether simple spectrum property of a Koopman representation can be expressed in a more "measure-theoretic" way. We now recall rank-1 concept which can be seen as a notion close to Katok's and Stepin's theory of cyclic approximation [104] (see also [26]).

Assume that T is an automorphism of a standard probability Borel space $(X, \mathcal{B}, \mu)$. T is said to have *rank one* property if there exists an increasing sequence of Rokhlin towers tending to the partition into points (a *Rokhlin tower* is a family $\{F, TF, \ldots, T^{n-1}F\}$ of pairwise disjoint sets, while "tending to the partition into points" means that we can approximate every set in $\mathcal{B}$ by unions of levels of towers in the sequence). Baxter [20] showed that the maximal spectral type of such a T is realized by a characteristic function. Since the cyclic space generated by the characteristic function of the base contains characteristic functions of all levels of the tower, by the definition of rank one, the increasing sequence of cyclic spaces tends to the whole L^2-space, therefore rank one property implies simplicity of the spectrum for the Koopman representation. It was a question for some time whether rank-1 is just a characterization of simplicity of the spectrum, disproved by del Junco [88]. We refer the reader to [46] as a good source for basic properties of rank-1 transformations.

Similarly to the rank one property, one can define *finite rank* automorphisms (simply by requiring that an approximation is given by a sequence of a fixed number of towers) – see e. g. [152], or even, a more general property, namely the *local rank one* property can be defined, just by requiring that the approximating sequence of single towers fills up a fixed fraction of the space (see [44,111]). Local rank one (so the more finite rank) property implies finite multiplicity [111]. Mentzen [146] showed that for each $n \geq 1$ one can construct an automorphism with simple spectrum and having rank n; in [138] there is an example of a simple spectrum automorphism which is not of local rank one. Ferenczi [45] introduced the notion of funny rank one (approximating towers are Rokhlin towers with "holes"). Funny rank one also implies simplicity of the spectrum. An example is given in [45] which is even not loosely Bernoulli (see Sect. "Inducing and Spectral Theory", we recall that local rank one property implies loose Bernoullicity [44]).

The notion of AT (see the glossary) has been introduced by Connes and Woods [25]. They proved that AT property implies zero entropy. They also proved that funny rank one automorphisms are AT. In [32] it is proved that the system induced by the classical Morse-Thue system is AT (it is an open question by S. Ferenczi whether this system has funny rank one property). A question by Dooley and Quas is whether AT implies funny rank one property. AT property implies "simplicity of the spectrum in L^1" which we already considered in Introduction (a "generic" proof of this fact is due to J.-P. Thouvenot).

A persistent question was formulated in the 1980s whether rank one itself is a spectral property. In [49] the authors maintained that this is not the case, based on an unpublished preprint of the first named author of [49] in which there was a construction of a Gaussian–Kronecker automorphism (see Sect. "Spectral Theory of Dynamical Systems of Probabilistic Origin") having rank-1 property. This latter construction turned out to be false. In fact de la Rue [181] proved that no Gaussian automorphism can be of local rank one. Therefore the question whether: *Rank one is a spectral property* remains one of the interesting open questions in that theory. Downarowicz and Kwiatkowski [33] proved that rank-1 is a spectral property in the class of systems generated by generalized Morse sequences.

One of the most beautiful theorems about rank-1 automorphisms is the following result of J. King [110] (for a different proof see [186]).

Theorem 8 (WCT) *If T is of rank one then for each element S of the centralizer $C(T)$ of T there exists a sequence (n_k) such that $U_T^{n_k} \to U_S$ strongly.*

A conjecture of J. King is that in fact for rank-1 automorphisms each indecomposable Markov operator $J = J_\rho$ ($\rho \in J_2^e(T)$) is a weak limit of powers of U_T (see [112], also [186]). To which extent the WCT remains true for actions of other groups is not clear. In [214] the WCT is proved in case of rank one flows, however the main argument seems to be based on the fact that a rank one flow has a non-zero time automorphism T_{t_0} which is of rank one, which is not true. After the proof of the WCT by Ryzhikov in [186] there is a remark that the rank one flow version of the theorem can be proved by a word for word repetition of the arguments. He also proves that if the flow $(T_t)_{t\in\mathbb{R}}$ is mixing, then T_1 does not have finite rank. On the other hand, for $\mathbb{A} = \mathbb{Z}^2$, Downarowicz and Kwiatkowski [34] gave recently a counterexample to the WCT.

Even though it looks as if rank one construction is not complicated, mixing in this class is possible; historically the first mixing constructions were given by D. Ornstein [151] in 1970, using probability type arguments for a choice of spacers. Once mixing was shown, the question arose whether absolutely continuous spectrum is also possible, as this would give automatically the positive answer to the Banach problem. However Bourgain [21], relating spectral measures of rank one automorphisms with some classical constructions of Riesz product measures, proved that a certain subclass of Ornstein's class consists of automorphisms with singular spectrum (see also [5] and [6]). Since in Ornstein's class spacers are chosen in a certain "non-constructive" way, quite a lot of attention was devoted to the rank one automorphism defined by cutting a tower at the nth step into $r_n = n$ subcolumns of equal "width" and placing i spacers over the ith subcolumn. The mixing property conjectured by M. Smorodinsky, was proved by Adams [7] (in fact Adams proved a general result on mixing of a class of staircase transformations). Spectral properties of rank-1 transformations are also studied in [114], where the authors proved that whenever $\sum_{n=1}^{\infty} r_n^{-2} = +\infty$ (r_n stands for the number of subcolumns at the nth step of the construction of a rank-1 automorphism) then the spectrum is automatically singular. H. Abdalaoui [5] gives a criterion for singularity of the spectrum of a rank one transformation; his proof uses a central limit theorem. It seems that still the question whether rank one implies singularity of the spectrum remains the most important question of this theory.

We have already seen in Sect. "Spectral Theory of Weighted Operators" that for a special class of rank one systems, namely those with discrete spectra ([87]), we have a nice theory for weighted operators. It would be extremely interesting to find a rank one automorphism with continuous spectrum for which a substitute of Helson's analysis exists.

B. Fayad [39] constructs a rank one differentiable flow, as a special flow over a two-dimensional rotation. In [40] he gives new constructions of smooth flows with singular spectra which are mixing (with a new criterion for a Rajchman measure to be singular). In [35] a certain smooth change of time for an irrational flows on the 3-torus is given, so that the corresponding flow is partially mixing and has the local rank one property.

Spectral Theory of Dynamical Systems of Probabilistic Origin

Let us just recall that when $(Y_n)_{n=-\infty}^{\infty}$ is a stationary process then its distribution μ on $\mathbb{R}^{\mathbb{Z}}$ is invariant under the shift S on $\mathbb{R}^{\mathbb{Z}}$: $S((x_n)_{n\in\mathbb{Z}}) = (y_n)_{n\in\mathbb{Z}}$, where $y_n = x_{n+1}$, $n \in \mathbb{Z}$. In this way we obtain an automorphism S defined on $(\mathbb{R}^{\mathbb{Z}}, \mathcal{B}(\mathbb{R}^{\mathbb{Z}}), \mu)$. For each automorphism T we can find $f\colon X \to \mathbb{R}$ measurable such that the smallest σ-algebra making the stationary process $(f \circ T^n)_{n\in\mathbb{Z}}$ measurable is equal to $\mathcal{B}$, therefore, for the purpose of this article, by a system of probabilistic origin we will mean (S, μ) obtained from a stationary infinitely divisible process (see e. g. [142,192]). In particular, the theory of Gaussian dynamical systems is indeed a classical part of ergodic theory (e. g. [149,150,211,212]). If $(X_n)_{n\in\mathbb{Z}}$ is a stationary real centered Gaussian process and σ denotes the *spectral measure of the process*, i. e. $\widehat{\sigma}(n) = E(X_n \cdot X_0)$, $n \in \mathbb{Z}$, then by $S = S_\sigma$ we denote the corresponding Gaussian system on the shift space (recall also that for each symmetric measure σ on $\mathbb{T}$ there is exactly one stationary real centered Gaussian process whose spectral measure is σ). Notice that if σ has an atom, then in the cyclic space generated by X_0 there exists an eigenfunction Y for S_σ – if now S_σ were ergodic, $|Y|$ would be a constant function which is not possible by the nature of elements in $\mathbb{Z}(X_0)$. In what follows we assume that σ is continuous.

It follows that U_{S_σ} restricted to $\mathbb{Z}(X_0)$ is spectrally the same as $V = V^\sigma$ acting on $L^2(\mathbb{T}, \sigma)$, and we obtain that $(U_{S_\sigma}, L^2(\mathbb{R}^{\mathbb{Z}}, \mu_\sigma))$ can be represented as the symmetric Fock space built over $H = L^2(\mathbb{T}, \sigma)$ and $U_{S_\sigma} = F(V)$ – see the glossary ($H^{\odot n}$ is called the *n-th chaos*). In other words the spectral theory of Gaussian dynamical systems is reduced to the spectral theory of special tensor products unitary operators. Classical results (see [26]) which can be obtained from this point of view are for example the following:

(A) ergodicity implies weak mixing,
(B) the multiplicity function is either 1 or is unbounded,
(C) the maximal spectral type of U_{S_σ} is equal to $\exp(\sigma)$, hence Gaussian systems enjoy the Kolmogorov group property.

However we can also look at a Gaussian system in a different way, simply by noticing that the variables $e^{2\pi i f}$ (f is a real variable), where $f \in \mathbb{Z}(X_0)$ generate $L^2(\mathbb{R}^{\mathbb{Z}}, \mu_\sigma)$. Now calculating the spectral measure of $e^{2\pi i f}$ is not difficult and we obtain easily (C). Moreover, integrals of type $\int e^{2\pi i f_0} e^{2\pi i f_1 \circ T^n} e^{2\pi i f_2 \circ T^{n+m}} \, d\mu_\sigma$ can also be calculated, whence in particular we easily obtain Leonov's theorem on the multiple mixing property of Gaussian systems [141].

One of the most beautiful parts of the theory of Gaussian systems concerns ergodic properties of S_σ when σ is concentrated on a thin Borel set. Recall that a closed sub-

set $K \subset \mathbb{T}$ is said to be a *Kronecker set* if each $f \in C(K)$ is a uniform limit of characters (restricted to K). Each Kronecker set has no rational relations. Gaussian–Kronecker automorphisms are, by definition, those Gaussian systems for which the measure σ (always assumed to be continuous) is concentrated on $K \cup \overline{K}$, K a Kronecker set. The following theorem has been proved in [51] (see also [26]).

Theorem 9 (Foiaş–Stratila Theorem) *If T is an ergodic automorphism and f is a real-valued element of L_0^2 such that the spectral measure σ_f is concentrated on $K \cup \overline{K}$, where K is a Kronecker set, then the process $(f \circ T^n)_{n\in\mathbb{Z}}$ is Gaussian.*

This theorem is indeed striking as it gives examples of weakly mixing automorphisms which are spectrally determined (like rotations). A relative version of the Foiaş–Stratila Theorem has been proved in [129].

The Foiaş–Stratila Theorem implies that whenever a spectral measure σ is Kronecker, it has no realization of the form σ_f with f bounded. We will see however in Sect. "Future Directions" that for some automorphisms T (having the SCS property) the maximal spectral type σ_T has the property that S_{σ_T} has a simple spectrum.

Gaussian–Kronecker automorphisms are examples of automorphisms with simple spectra. In fact, whenever σ is concentrated on a set without rational relations, then S_σ has a simple spectrum (see [26]). Examples of mixing automorphisms with simple spectra are known [149], however it is still unknown (Thouvenot's question) whether the Foiaş–Stratila property may hold in the mixing class. F. Parreau [154] using independent Helson sets gave an example of mildly mixing Gaussian system with the Foiaş–Stratila property.

In [165] there is a remark that the set of finite essential values of the multiplicity function of U_{S_σ} forms a (multiplicative) subsemigroup of $\mathbb{N}$. However, it seems that there is no "written" proof of this fact.

Joining theory of a class of Gaussian system, called GAG, is developed in [136]. A Gaussian automorphism S_σ with the Gaussian space $H \subset L_0^2(\mathbb{R}^{\mathbb{Z}}, \mu_\sigma)$ is called a GAG if for each ergodic self-joining $\rho \in J_2^e(S_\sigma)$ and arbitrary $f, g \in H$ the variable

$$(\mathbb{R}^{\mathbb{Z}} \times \mathbb{R}^{\mathbb{Z}}, \rho) \ni (x, y) \mapsto f(x) + g(y)$$

is Gaussian. For GAG systems one can describe the centralizer and factors, they turn out to be objects close to the probability structure of the system. One of the crucial observations in [136] was that all Gaussian systems with simple spectrum are GAG.

It is conjectured (J.P. Thouvenot) that in the class of zero entropy Gaussian systems the PID property holds true.

For the spectral theory of classical factors of a Gaussian system see [137]; also spectrally they share basic spectral properties of Gaussian systems. Recall that historically one of the classical factors namely the σ-algebra of sets invariant for the map

$$(\ldots, x_{-1}, x_0, x_1, \ldots) \mapsto (\ldots, -x_{-1}, -x_0, -x_1, \ldots)$$

was the first example with zero entropy and countable Lebesgue spectrum (indeed, we need a singular measure σ such that $\sigma * \sigma$ is equivalent to Lebesgue measure [150]). For factors obtained as functions of a stationary process see [83].

T. de la Rue [181] proved that Gaussian systems are never of local rank-1, however his argument does not apply to classical factors. We conjecture that Gaussian systems are disjoint from rank-1 automorphisms (or even from local rank-1 systems).

We now turn the attention to Poissonian systems (see [26] for more details). Assume that $(X, \mathcal{B}, \mu)$ is a standard Borel space, where μ is infinite, σ-finite. The new configuration space $\widetilde{X}$ is taken as the set of all countable subsets $\{x_i : i \geq 1\}$ of X. Once a set $A \in \mathcal{B}$, of finite measure is given one can define a map $N_A : \widetilde{X} \to \mathbb{N}(\cup\{\infty\})$ just counting the number of elements belonging to A. The measure-theoretic structure $(\widetilde{X}, \widetilde{\mathcal{B}}, \widetilde{\mu})$ is given so that the maps N_A become random variables with Poisson distribution of parameter $\mu(A)$ and such that whenever $A_1, \ldots, A_k \subset X$ are of finite measure and are pairwise disjoint then the variables $N_{A_1}, \ldots, N_{A_k}$ are independent.

Assume now that T is an automorphism of $(X, \mathcal{B}, \mu)$. It induces a natural automorphism on the space $(\widetilde{X}, \widetilde{\mathcal{B}}, \widetilde{\mu})$ defined by $\widetilde{T}(\{x_i : i \geq 1\} = \{Tx_i : i \geq 1\}$. The automorphism $\widetilde{T}$ is called the *Poisson suspension* of T (see [26]). Such a suspension is ergodic if and only if no set of positive and finite μ-measure is T-invariant. Moreover ergodicity of $\widetilde{T}$ implies weak mixing. In fact the spectral structure of $U_{\widetilde{T}}$ is very similar to the Gaussian one: namely the first chaos equals $L^2(X, \mathcal{B}, \mu)$ (we emphasize that this is about the whole L^2 and not only L_0^2) on which $U_{\widetilde{T}}$ acts as U_T and the $L^2(\widetilde{X}, \widetilde{\mu})$ together with the action of $U_{\widetilde{T}}$ has the structure of the symmetric Fock space $F(L^2(X, \mathcal{B}, \mu))$ (see the glossary).

We refer to [22,86,168,169] for ergodic properties of systems given by symmetric α-stable stationary processes, or more generally infinitely divisible processes. Again, they share spectral properties similar to the Gaussian case: er-

godicity implies weak mixing, while mixing implies mixing of all orders.

In [171], E. Roy clarifies the dynamical "status" of such systems. He uses Poisson suspension automorphisms and the Maruyama representation of an infinitely divisible process mixed with basic properties of automorphisms preserving infinite measure (see [1]) to prove that as a dynamical system, a stationary infinitely divisible process (without the Gaussian part), is a factor of the Poisson suspension over the Lévy measure of this process. In [170] a theory of ID-joinings is developed (which should be viewed as an analog of the GAG theory in the Gaussian class). Parreau and Roy [155] give an example of a Poisson suspension with a minimal possible set of ergodic self-joinings.

Many natural problems still remain open here, for example (assuming always zero entropy of the dynamical system under consideration): Are Poisson suspensions disjoint from Gaussian systems? What is the spectral structure for dynamical systems generated by symmetric α-stable process? Are such systems disjoint whenever $\alpha_1 \neq \alpha_2$? Are Poissonian systems disjoint from local rank one automorphisms (cf. [181])? In [91] it is proved that Gaussian systems are disjoint from so called simple systems (see ► Joinings in Ergodic Theory and [93,208]); we will come back to an extension of this result in Sect. "Future Directions". It seems that flows of probabilistic origin satisfy the Kolmogorov group property for the spectrum. One can therefore ask how different are systems satisfying the Kolmogorov group property from systems for which the convolutions of the maximal spectral type are pairwise disjoint (see also Sect. "Future Directions" and the SCS property).

We also mention here another problem which will be taken up in Sect. "Special Flows and Flows on Surfaces, Interval Exchange Transformations" – *Is it true that flows of probabilistic origin are disjoint from smooth flows on surfaces?* Recently A. Katok and A. Windsor announced that it is possible to construct a Kronecker measure so that the corresponding Gaussian system ($\mathbb{Z}$-action (!)) has a smooth representation on the torus.

Yet one more (joining) property seems to be characteristic in the class of systems of probabilistic origin, namely they satisfy so called ELF property (see [30] and ► Joinings in Ergodic Theory). Vershik asked whether the ELF property is spectral – however the example of a cocycle from [205] together with Theorem 7 (i) yields a certain Rokhlin extension of a rotation which is ELF and has countable Lebesgue spectrum in the orthocomplement of the eigenfunctions (see [206]); on the other hand any affine extension of that rotation is spectrally the same, while it cannot have the ELF property.

Prikhodko and Thouvenot (private communication) have constructed weakly mixing and non-mixing rank one automorphisms which enjoy the ELF property.

Inducing and Spectral Theory

Assume that T is an ergodic automorphism of a standard probability Borel space $(X, \mathcal{B}, \mu)$. Can "all" dynamics be obtained by inducing (see the glossary) from one fixed automorphism was a natural question from the very beginning of ergodic theory. Because of Abramov's formula for entropy $h(T_A) = h(T)/\mu(A)$ it is clear that positive entropy transformations cannot be obtained from inducing on a zero entropy automorphism. However here we are interested in spectral questions and thus we ask how many spectral types we obtain when we induce. It is proved in [59] that the family of $A \in \mathcal{B}$ for which T_A is mixing is dense for the (pseudo) metric $d(A_1, A_2) = \mu(A_1 \triangle A_2)$. De la Rue [182] proves the following result: *For each ergodic transformation T of a standard probability space $(X, \mathcal{B}, \mu)$ the set of $A \in \mathcal{B}$ for which the maximal spectral type of U_{T_A} is Lebesgue is dense in $\mathcal{B}$.* The multiplicity function is not determined in that paper. Recall (without giving a formal definition, see [152]) that a zero entropy automorphism is *loosely Bernoulli* (LB for short) if and only if it can be induced from an irrational rotation (see also [43,99]). The LB theory shows that not all dynamical systems can be obtained by inducing from an ergodic rotation. However an open question remained whether LB systems exhaust spectrally all Koopman representations. In a deep paper [180], de la Rue studies LB property in the class of Gaussian–Kronecker automorphisms, in particular he constructs S which is not LB. Suppose now that T is LB and for some $A \in \mathcal{B}$, U_{T_A} is isomorphic to U_S. Then by the Foiaş–Stratila Theorem, T_A is isomorphic to S, and hence T_A is not LB. However an induced automorphism from an LB automorphism is LB, a contradiction.

Special Flows and Flows on Surfaces, Interval Exchange Transformations

We now turn our attention to flows. The cases of the geodesic flow, horocycle flows on homogenous spaces of $SL(2, \mathbb{R})$ and nilflows are classical (we refer the reader to [105] with a nice description of the first two cases, while for nilflows we refer to [157]: these classes of flows on homogenous spaces have countable Lebesgue spectrum, in the third case – in the orthocomplement of the eigenspace). On the other hand the classical cyclic approximation theory of Katok and Stepin [104] (see [26]) leads to examples of smooth flows on the torus with simple continuous singular spectra.

Given an ergodic automorphism T on $(X, \mathcal{B}, \mu)$ and a positive integrable function $f\colon X \to \mathbb{R}^+$ consider the corresponding special flow T^f (see the glossary). Obviously, such a flow is ergodic. Special flows were introduced to ergodic theory by von Neumann in his fundamental work [148] in 1932. Also in that work he explains how to compute eigenvalues for special flows, namely: *$r \in \mathbb{R}$ is an eigenvalue of T^f if and only if the following functional equation*

$$e^{2\pi i r f(x)} = \frac{\xi(x)}{\xi(Tx)}$$

has a measurable solution $\xi\colon X \to \mathbb{T}$. We recall also that the classical Ambrose-Kakutani theorem asserts that practically each ergodic flow has a special representation ([26], see also Rudolph's theorem on special representation therein).

A classical situation when we obtain "natural" special representations is while considering smooth flows on surfaces (we refer the reader to Hasselblatt's and Katok's monograph [73]). They have transversals on which the Poincaré map is piecewise isometric, and this entails a study of interval exchange transformations (IET), see [26,108,163]. Formally, to define IET of m intervals we need a permutation π of $\{1, \ldots, m\}$ and a probability vector $\lambda = (\lambda_1, \ldots, \lambda_m)$ (with positive entries). Then we define $T = T_{\lambda,\pi}$ of $[0, 1)$ by putting

$$T_{\lambda,\pi}(x) = x + \beta_i^\pi - \beta_i \quad \text{for} \quad x \in [\beta_i, \beta_{i+1}) ,$$

where $\beta_i = \sum_{j<i} \lambda_j$, $\beta_i^\pi = \sum_{\pi j<\pi i} \beta_j$. Obviously, each IET preserves Lebesgue measure. One of possible approaches to study ergodic properties of IET is an "a.e" approach "seen" in the definition of $T_{\lambda,\pi}$. It is based on the fundamental fact that the induced transformation on a subinterval of $[0, 1)$ is also IET (see [26]). This leads to a very delicate and deep mathematics based on Rauzy induction, which is a way of inducing on special intervals, considering only irreducible permutations whose set is partitioned into orbits of some maps (any such an orbit is called a *Rauzy class*). If now $\mathcal{R}$ is a Rauzy class of permutations and λ lies in the standard simplex Δ_{m-1} then the Rauzy induction together with a natural renormalization leads to a map $\mathcal{P}\colon \mathcal{R} \times \Delta_{m-1} \to \mathcal{R} \times \Delta_{m-1}$. For a better understanding of the dynamics of the Rauzy map Veech [209] introduced the space of *zippered rectangles*. A zippered rectangle associated to the Rauzy class $\mathcal{R}$ is a quadruple (λ, h, a, π), where $\lambda \in \mathbb{R}_+^m$, $h \in \mathbb{R}_+^m$, $a \in \mathbb{R}_+^m$, $\pi \in \mathcal{R}$ and the vectors h and a satisfy some equations and inequalities. Every zippered rectangle (λ, h, a, π) determines a Riemann structure on a compact connected surface. Denote by $\Omega(\mathcal{R})$ the space of all zippered rectangles, corresponding to a given Rauzy class $\mathcal{R}$ and satisfying the condition $\langle \lambda, h\rangle = 1$. In [209], Veech defined a flow $(P^t)_{t\in\mathbb{R}}$ on the space $\Omega(\mathcal{R})$ putting

$$P^t(\lambda, h, a, \pi) = (e^t\lambda, e^{-t}h, e^{-t}a, \pi)$$

and extended the Rauzy map. On so called *Veech moduli space* of zippered rectangles, the flow (P^t) becomes the *Teichmüller flow* and it preserves a natural Lebesgue measure class; by passing to a transversal and projecting the measure on the space of IETs $\mathcal{R} \times \Delta_{m-1}$ Veech has proved the following fundamental theorem ([209], see also [143]) which is a generalization of the fact that Gauss measure $1/(\ln 2)1/(1 + x)dx$ is invariant for the Gauss map which sends $t \in (0, 1)$ into the fractional part of its inverse.

Theorem 10 (Veech, Masur, 1982) *Assume that $\mathcal{R}$ is a Rauzy class. There exists a σ-finite measure $\mu_{\mathcal{R}}$ on $\mathcal{R} \times \Delta_{m-1}$ which is $\mathcal{P}$-invariant, equivalent to "Lebesgue" measure, conservative and ergodic.*

In [209] it is proved that a.e. (in the above sense) IET is then of rank one (and hence is ergodic and has a simple spectrum). He also formulated as an open problem whether we can achieve the weak mixing property a.e. This has been recently answered in positive by A. Avila and G. Forni [19] (for π which is not a rotation).

Katok [100] proved that IET have no mixing factors (in fact his proof shows more: the IET class is disjoint with the class of mixing transformations). By their nature, IET transformations are of finite rank (see [26]) so they are of finite multiplicity. They need not be of simple spectrum (see remarks in [105] pp. 88–90). It remains an open question whether an IET can have a non-singular spectrum. Joining properties in the class of exchange of 3 and more intervals are studied in [47,48]. An important question of Veech [208] whether a.e. IET is simple is still open.

When we consider a smooth flow on a surface preserving a smooth measure, whose only singularity (we assume that we have only finitely many singularities) are simple (non-degenerated) saddles then such a flow has a special representation over an interval exchange automorphism under a smooth function which has finitely many logarithmic singularities (see [73]). In the article by Arnold [18] the quasi-periodic Hamiltonian case is considered: $H\colon \mathbb{R}^2 \to \mathbb{R}$ satisfies $H(x+m, y+n) = H(x, y)+ n\alpha_1+m\alpha_2$, $\alpha_1/\alpha_2 \notin \mathbb{Q}$, and we then consider the following system of differential equations on $\mathbb{T}^2$

$$\frac{dx}{dt} = \frac{\partial H}{\partial y}, \quad \frac{dy}{dt} = \frac{-\partial H}{\partial x} . \tag{2}$$

As Arnold shows the dynamical system arising from the system (2) has one ergodic component which has a special representation over the irrational rotation by $\alpha := \alpha_1/\alpha_2$ under a smooth function with finitely many logarithmic singularities (all other ergodic components are periodic orbits and separatrices). By changing a speed as it is done in [57] so that critical points of the vector field in (2) become singular points, Arnold's special representation is transformed to a special flow over the same irrational rotation however under a piecewise smooth function. If the sum of jumps is not zero then in fact we come back to von Neumann's class of special flows considered in [148]. Similar classes of special flows (when the roof function is of bounded variation) are obtained from ergodic components of flows associated to billiards in convex polygones with rational angles [106]. Kochergin [115] showed that special flows over irrational rotations and under bounded variation functions are never mixing. This has been recently strengthened in [54] to the following: *If T is an irrational rotation and f is of bounded variation then the special flow T^f is spectrally disjoint from all mixing flows*. In particular all such flows have singular spectra. Moreover, in [54] it is proved that whenever the Fourier transform of the roof function f is of order $O(1/n)$ then T^f is disjoint from all mixing flows (see also [55]). In fact in the papers [54,55,56,57] the authors discuss the problem of disjointness of those special flows with all ELF-flows conjecturing that no flow of probabilistic origin has a smooth realization on a surface. In [140] the analytic case is considered leading to a "generic" result on disjointness with the ELF class generalizing the classical Shklover's result on the weak mixing property [197].

Kochergin [117] proved the absence of mixing for flows where the roof function has finitely many singularities, whenever the sum of "left logarithmic speeds" and the sum of "right logarithmic speeds" are equal – this is called a *symmetric logarithmic case*, however some Diophantine restriction is put on α.

In [128], where also the absence of mixing is considered for the symmetric logarithmic case, it was conjectured (and proved for arbitrary rotation) that a necessary condition for mixing of a special flow T^f (with arbitrary T and f) is the condition that the sequence of distributions $((f_0^{(n)})_*)_n$ tends to δ_∞ in the space of probability measures on $\overline{\mathbb{R}}$. K. Schmidt [194] proved it using the theory of cocycles and extending a result from [3] on tightness of cocycles.

A. Katok [100] proved the absence of mixing for special flows over IET when the roof function is of bounded variation (see also [187]). Katok's theorem was strengthened in [56] to the disjointness theorem with the class of mixing flows.

On the other hand there is a lot of (difficult) results pointing out classes of special flows over irrational rotations which are mixing, especially (but not only) in the class of non-symmetric logarithmic singularities: [36,38] (B. Fayad was able to give a speed of convergence to zero for Fourier coefficients), [109,119,120]. Recently mixing property has been proved in a non-symmetric case in [203] when the base transformation is a special class of IETs.

The eigenvalue problem (mainly how many frequencies can have the group of eigenvalues) for special flows over irrational rotations is studied in [41,42,71].

A. Avila and G. Forni [19] proved that a.e. translation flow on a surface (of genus at least two) is weakly mixing (which is a drastic difference with the linear flow case of the torus, where the spectrum is always discrete).

The problem of whether mixing flows indicated in this chapter are mixing of all orders is open (it is also unknown whether they have singular spectra). One of several possible approaches (proposed by B. Fayad and J.-P. Thouvenot) toward positive solution of this problem would be to show that such flows enjoy so called Ratner's property (R-property). This property may be viewed as a particular way of divergence of orbits of close points; it was shown to hold for horocycle flows by M. Ratner [162]. We refer the reader to [162] and the survey article [201] for the formal definitions and basic consequences of R-property. In particular, R-property implies "rigidity" of joinings and it also implies the PID property; hence mixing and R-property imply mixing of all orders. In [57,58] a version of R-property is shown for the class of von Neumann special flows (however α is assumed to have bounded partial quotients). This allowed one to prove there that such flows are even mildly mixing (mixing is excluded by a Kochergin's result). We conjecture that an R-property may also hold for special flows over multidimensional rotations with roof functions given by nil-cocycles which we mentioned in Sect. "Spectral Theory of Weighted Operators".

If indeed the R-property is ubiquitous in the class of smooth flows on surfaces it may also be useful to show that smooth flows on surfaces are disjoint with flows of probabilistic origin – see [91,92,135,190,202].

B. Fayad [40] gives a criterion that implies singularity of the maximal spectral type for a dynamical system on a Riemannian manifold. As an application he gives a class of smooth mixing flows (with singular spectra) on $\mathbb{T}^3$ obtained from linear flows by a time change (again this is a drastic difference with dimension two, where a smooth time change of a linear flow leads to non-mixing flows [26]).

The spectral multiplicity problem for special flows (with sufficiently regular roof functions) over irrational rotations seems to be completely untouched (except for the case of a sufficiently smooth f – the spectrum of T^f is then simple [26]). It would be nice to have examples of such flows with finite bigger than one multiplicity. In particular, is it true that the von Neumann class of special flows have finite multiplicity? This was partially solved by A. Katok (private communication) on certain subspaces in L^2, but not on the whole L^2-space.

Problem. *Given $Tx = x + \alpha$ (with α irrational) can we find $f\colon [0,1) \to \mathbb{R}^+$ sufficiently regular (e.g. with finitely many "controllable" singularities) such that T^f has a Lebesgue spectrum?*

Of course the above is related to the question whether at all one can find a smooth flow on a surface with a Lebesgue spectrum (for $\mathbb{Z}$-actions we can even see positive entropy diffeomorphisms on the torus).

We mention at the end that if we drop here (and in other problems) the assumption of regularity of f then the answers will be always positive because of the LB theory; in particular there is a section of any horocycle flow (it has the LB property [161]) such that in the corresponding special representation T^f the map T is an irrational rotation. Using a Kochergin's result [118] on cohomology (see also [98,176]) the L^1-function f is cohomologous to a positive function g which is even continuous, thus T^f is isomorphic to T^g.

Future Directions

We have already seen several cases where spectral properties interact with measure-theoretic properties of a system. Let us mention a few more cases which require further research and deeper understanding.

We recall that the weak mixing property can be understood as a property complementary to discrete spectrum (more precisely to the distality [62]), or similarly mild mixing property is complementary to rigidity. This can be phrased quite precisely by saying that T is not weakly (mildly) mixing if and only if it has a non-trivial factor with discrete spectrum (it has a non-trivial rigid factor). It has been a question for quite a long time if in a sense mixing can be "built" on the same principle. In other words we seek a certain "highly" non-mixing factor. It was quite surprising when in 2005 F. Parrreau (private communication) gave the positive answer to this problem.

Theorem 11 (Parreau) *Assume that T is an ergodic automorphism of a standard probability space $(X, \mathcal{B}, \mu)$. Assume moreover that T is not mixing. Then there exists a non-trivial factor (see below) of T which is disjoint with all mixing automorphisms.*

In fact, Parreau proved that each factor of T given by $\mathcal{B}_\infty(\rho)$ (this σ-algebra is described in [136]), where $U_T^{n_k} \to J_\rho$, is disjoint from all mixing transformations. This proof leads to some other results of the same type, for example: *Assume that T is an ergodic automorphism of a standard probability space. Assume that there exists a non-trivial automorphism S with a singular spectrum which is not disjoint with T. Then T has a non-trivial factor which is disjoint with any automorphism with a Lebesgue spectrum.*

The problem of spectral multiplicity of Cartesian products for "typical" transformation studied by Katok [98] and then its solution in [13] which we already considered in Sect. "The Multiplicity Function" lead to a study of those T for which

$$(CS) \quad \sigma^{(m)} \perp \sigma^{(n)} \text{ whenever } m \neq n\,,$$

where $\sigma = \sigma_T$ just stands for the reduced maximal spectral type of U_T (which is constantly assumed to be a continuous measure), see also Stepin's article [199].

The usefulness of the above property (CS) in ergodic theory was already shown in [90], where a spectral counterexample machinery was presented using the following observation: *If $\mathcal{A}$ is a $T^{\times\infty}$-invariant sub-σ-algebra such that the maximal spectral type on $L^2(\mathcal{A})$ is absolutely continuous with respect to σ_T then $\mathcal{A}$ is contained in one of the coordinate sub-σ-algebras $\mathcal{B}$.* Based on that in [90] it is shown how to construct two weakly isomorphic action which are not isomorphic or how to construct two non-disjoint automorphisms which have no common non-trivial factors (such constructions were previously known for so called minimal self-joining automorphisms [174]). See also [200] for extensions of those results to $\mathbb{Z}^d$-actions.

Prikhodko and Ryzhikov [159] proved that the classical Chacon transformation enjoys the (CS) property. The SCS property defined in the glossary is stronger than the (CS) condition above; the SCS property implies that the corresponding Gaussian system S_{σ_T} has a simple spectrum. Ageev [10] shows that Chacon's transformation satisfies the SCS property; moreover in [13] he shows that the SCS property is satisfied generically and he gives a construction of a rank one mixing SCS-system (see also [191]). In [134] it is proved that some special flows considered in Sect. "Special Flows and Flows on Surfaces, Interval Exchange Transformations" (including the von Neumann class, however with α having unbounded partial quotients) have the SCS property. Since the corresponding Gaussian systems have simple spectra, it would be interesting

to decide whether σ_T (for an SCS-automorphism) can be concentrated on a set without rational relations. It is quite plausible that the SCS property is commonly seen for smooth flows on surfaces.

Katok and Thouvenot (private communication) considered systems called *infinitely divisible*. These are systems T on $(X, \mathcal{B}, \mu)$ which have a family of factors $\mathcal{B}_\omega$ indexed by $\omega \in \bigcup_{n=0}^{\infty}\{0,1\}^n$ ($\mathcal{B}_\varepsilon = \mathcal{B}$) such that $\mathcal{B}_{\omega 0} \perp \mathcal{B}_{\omega 1}$, $\mathcal{B}_{\omega 0} \vee \mathcal{B}_{\omega 1} = \mathcal{B}_\omega$ and for each $\eta \in \{0,1\}^{\mathbb{N}}$, $\cap_{n\in\mathbb{N}} \mathcal{B}_{\eta[0,n]} = \{\emptyset, X\}$. They showed (unpublished) that there are discrete spectrum transformations which are ID, and that there are rank one transformations with continuous spectra which are also ID (clearly Gaussian systems are ID). It was until recently that a relationship between ID automorphisms and systems coming from stationary ID processes was unclear. In [135] it is proved that dynamical systems coming from stationary ID processes are factors of ID automorphisms; moreover, ID automorphisms are disjoint with all systems having the SCS property. It would be nice to decide whether Koopman representations associated to ID automorphisms satisfy the Kolmogorov group property.

Acknowledgments

Research supported by the EU Program Transfer of Knowledge "Operator Theory Methods for Differential Equations" TODEQ and the Polish Ministry of Science and High Education.

Bibliography

1. Aaronson J (1997) An introduction to infinite ergodic theory, mathematical surveys and monographs. Am Math Soc 50:284
2. Aaronson J, Nadkarni MG (1987) L^∞ eigenvalues and L^2 spectra of nonsingular transformations. Proc London Math Soc 55(3):538–570
3. Aaronson J, Weiss B (2000) Remarks on the tightness of cocycles. Coll Math 84/85:363–376
4. Abdalaoui H (2000) On the spectrum of the powers of Ornstein transformations Ergodic theory and harmonic analysis (Mumbai, 1999). Sankhya Ser. A 62:291–306
5. Abdalaoui H (2007) A new class of rank 1 transformations with singular spectrum. Ergodic Theor Dynam Syst 27:1541–1555
6. Abdalaoui H, Parreau F, Prikhodko A (2006) A new class of Ornstein transformations with singular spectrum. Ann Inst H Poincaré Probab Statist 42:671–681
7. Adams T (1998) Smorodinsky's conjecture on rank one systems. Proc Am Math Soc 126:739–744
8. Ageev ON (1988) Dynamical systems with a Lebesgue component of even multiplicity in the spectrum. Mat Sb (NS) 136(178):307–319, (Russian) 430 (1989) Translation in Math. USSR-Sb. 64:305–317
9. Ageev ON (1999) On ergodic transformations with homogeneous spectrum. J Dynam Control Syst 5:149–152
10. Ageev ON (2000) On the spectrum of cartesian powers of classical automorphisms. Mat Zametki 68:643–647, (2000) Translation Math. Notes 68:547–551, (Russian)
11. Ageev ON (2001) On the multiplicity function of generic group extensions with continuous spectrum. Ergodic Theory Dynam Systems 21:321–338
12. Ageev ON (2005) The homogeneous spectrum problem in ergodic theory. Invent Math 160:417–446
13. Ageev ON (2007) Mixing with staircase multiplicity function. preprint
14. Alexeyev VM (1958) Existence of bounded function of the maximal spectral type. Vestnik Mosc Univ 5:13-15 and (1982) Ergodic Theory Dynam Syst 2:259–261
15. Anosov DV (2003) On Spectral multiplicities in ergodic theory. Institut im. V.A. Steklova, Moscow
16. Anzai H (1951) Ergodic skew product transformations on the torus. Osaka J Math 3:83–99
17. Arni N (2005) Spectral and mixing properties of actions of amenable groups. Electron Res Announc AMS 11:57–63
18. Arnold VI (1991) Topological and ergodic properties of closed 1-forms with incommensurable periods. Funktsional Anal Prilozhen 25:1–12, (Russian)
19. Avila A, Forni G (2007) Weak mixing for interval exchange transformations and translation flows. Ann Math 165:637–664
20. Baxter JR (1971) A class of ergodic transformations having simple spectrum. Proc Am Math Soc 27:275–279
21. Bourgain J (1993) On the spectral type of Ornstein's class of transformations. Isr J Math 84:53–63
22. Cambanis S, Podgórski K, Weron A (1995) Chaotic behaviour of infinitely divisible processes. Studia Math 115:109–127
23. Chacon RV (1970) Approximation and spectral multiplicity. In: Dold A, Eckmann B (eds) Contributions to ergodic theory and probability. Springer, Berlin, pp 18–27
24. Choe GH Products of operators with singular continuous spectra. In: Operator theory: operator algebras and applications, Part 2. (Durham, 1988), pp 65–68, Proc Sympos Pure Math, 51, Part 2, Amer Math Soc, Providence, 1990
25. Connes A, Woods E (1985) Approximately transitive flows and ITPFI factors. Ergod Theory Dynam Syst 5:203–236
26. Cornfeld IP, Fomin SV, Sinai YG (1982) Ergodic theory. Springer, New York
27. Danilenko AL (2006) Explicit solution of Rokhlin's problem on homogeneous spectrum and applications. Ergod Theory Dyn Syst 26:1467–1490
28. Danilenko AL, Lemańczyk M (2005) A class of multipliers for $\mathcal{W}^\perp$. Isr J Math 148:137–168
29. Danilenko AI, Park KK (2002) Generators and Bernouillian factors for amenable actions and cocycles on their orbits. Ergod Theory Dyn Syst 22:1715–1745
30. Derriennic Y, Frączek K, Lemańczyk M, Parreau F (2007) Ergodic automorphisms whose weak closure of off-diagonal measures consists of ergodic self-joinings. Coll Math 110:81–115
31. Dooley A, Golodets VY (2002) The spectrum of completely positive entropy actions of countable amenable groups. J Funct Anal 196:1–18
32. Dooley A, Quas A (2005) Approximate transitivity for zero-entropy systems. Ergod Theory Dyn Syst 25:443–453

33. Downarowicz T, Kwiatkowski J (2000) Spectral isomorphism of Morse flows. Fundam Math 163:193–213
34. Downarowicz T, Kwiatkowski J (2002) Weak Cosure Theorem fails for $\mathbb{Z}^d$-actions. Stud Math 153:115–125
35. Fayad B (2001) Partially mixing and locally rank 1 smooth transformations and flows on the torus $\mathbb{T}^d$, $d \geq 3$. J London Math Soc 64(2):637–654
36. Fayad B (2001) Polynomial decay of correlations for a class of smooth flows on the two torus. Bull Soc Math France 129:487–503
37. Fayad B (2002) Skew products over translations on T^d, $d \geq 2$. Proc Amer Math Soc 130:103–109
38. Fayad B (2002) Analytic mixing reparametrizations of irrational flows. Ergod Theory Dyn Syst 22:437–468
39. Fayad B (2005) Rank one and mixing differentiable flows. Invent Math 160:305–340
40. Fayad B (2006) Smooth mixing flows with purely singular spectra. Duke Math J 132:371–391
41. Fayad B, Katok AB, Windsor A (2001) Mixed spectrum reparameterizations of linear flows on $\mathbb{T}^2$, dedicated to the memory of I. G. Petrovskii on the occasion of his 100th anniversary. Mosc Math J 1:521–537
42. Fayad B, Windsor A (2007) A dichotomy between discrete and continuous spectrum for a class of special flows over rotations. J Mod Dyn 1:107–122
43. Feldman J (1976) New K-automorphisms and a problem of Kakutani. Isr J Math 24:16–38
44. Ferenczi S (1984) Systèmes localement de rang un. Ann Inst H Poincaré Probab Statist 20:35–51
45. Ferenczi S (1985) Systèmes de rang un gauche. Ann Inst H Poincaré Probab Statist 21:177–186
46. Ferenczi S (1997) Systems of finite rank. Colloq Math 73:35–65
47. Ferenczi S, Holton C, Zamboni LQ (2005) Joinings of three-interval exchange transformations. Ergod Theory Dyn Syst 25:483–502
48. Ferenczi S, Holton C, Zamboni LQ (2004) Structure of three-interval exchange transformations III: ergodic and spectral properties. J Anal Math 93:103–138
49. Ferenczi S, Lemańczyk M (1991) Rank is not a spectral invariant. Stud Math 98:227–230
50. Filipowicz I (1997) Product $\mathbb{Z}^d$-actions on a Lebesgue space and their applications. Stud Math 122:289–298
51. Foiaş C, Stratila S (1968) Ensembles de Kronecker dans la théorie ergodique. CR Acad Sci Paris, Ser A-B 267:A166–A168
52. Frączek K (1997) On a function that realizes the maximal spectral type. Stud Math 124:1–7
53. Frączek K (2000) Circle extensions of Z^d-rotations on the d-dimensional torus. J London Math Soc 61(2):139–162
54. Frączek K, Lemańczyk M (2004) A class of special flows over irrational rotations which is disjoint from mixing flows. Ergod Theory Dyn Syst 24:1083–1095
55. Frączek K, Lemańczyk M (2003) On symmetric logarithm and some old examples in smooth ergodic theory. Fund Math 180:241–255
56. Frączek K, Lemańczyk M (2005) On disjointness properties of some smooth flows. Fund Math 185:117–142
57. Frączek K, Lemańczyk M (2006) On mild mixing of special flows over irrational rotations under piecewise smooth functions. Ergod Theory Dyn Syst 26:719–738
58. Frączek K, Lemańczyk M, Lesigne E (2007) Mild mixing property for special flows under piecewise constant functions. Disc Contin Dyn Syst 19:691–710
59. Friedman NA, Ornstein DS (1973) Ergodic transformations induce mixing transformations. Adv Math 10:147–163
60. Furstenberg H (1967) Disjointness in ergodic theory, minimal sets, and a problem of Diophantine approximation. Math Syst Theory 1:1–49
61. Furstenberg H, Weiss B (1978) The finite multipliers of infinite ergodic transformations. Lect Notes Math 668:127–132
62. Furstenberg H (1981) Recurrence in ergodic theory and combinatorial number theory. Princeton University Press, Princeton
63. Glasner E (1994) On the class of multipliers for $\mathcal{W}^{\perp}$. Ergod Theory Dyn Syst 14:129–140
64. Glasner E (2003) Ergodic theory via joinings. Math Surv Monogr, AMS, Providence, RI 101:384
65. Glasner E, Weiss B (1989) Processes disjoint from weak mixing. Trans Amer Math Soc 316:689–703
66. Goodson GR (1999) A survey of recent results in the spectral theory of ergodic dynamical systems. J Dyn Control Syst 5:173–226
67. Goodson GR, Kwiatkowski J, Lemańczyk M, Liardet P (1992) On the multiplicity function of ergodic group extensions of rotations. Stud Math 102:157–174
68. Gromov AL (1991) Spectral classification of some types of unitary weighted shift operators. Algebra Anal 3:62–87, (Russian) (1992) Translation in St. Petersburg Math J 3:997–1021
69. Guenais M (1998) Une majoration de la multiplicité spectrale d'opérateurs associés à des cocycles réguliers. Isr J Math 105:263–284
70. Guenais M (1999) Morse cocycles and simple Lebesgue spectrum. Ergod Theory Dyn Syst 19:437–446
71. Guenais M, Parreau F (2005) Valeurs propres de transformations liées aux rotations irrationnelles et aux fonctions en escalier. (in press)
72. Hahn F, Parry W (1968) Some characteristic properties of dynamical systems with quasi-discrete spectra. Math Syst Theory 2:179–190
73. Hasselblatt B, Katok AB (2002) Principal structures. In: Handbook of dynamical systems, vol 1A. North-Holland, Amsterdam, pp 1–203
74. Helson H (1986) Cocycles on the circle. J Oper Theory 16:189–199
75. Helson H, Parry W (1978) Cocycles and spectra. Arkiv Math 16:195–206
76. Herman M (1979) Sur la conjugaison différentiable des difféomorphismes du cercle à des rotations. Publ Math IHES 49:5–234
77. Host B (1991) Mixing of all orders and pairwise independent joinings of systems with singular spectrum. Isr J Math 76:289–298
78. Host B, Méla F, Parreau F (1991) Non-singular transformations and spectral anaysis of measures. Bull Soc Math France 119:33–90
79. Iwanik A (1991) The problem of L^p-simple spectrum for ergodic group automorphisms. Bull Soc Math France 119:91–96
80. Iwanik A (1992) Positive entropy implies infinite L^p-multiplicity for $p > 1$. Ergod Theory Rel Top (Güstrow 1990) III:124–127. Lecture Notes in Math, vol 1514. Springer, Berlin

81. Iwanik A (1997) Anzai skew products with Lebesgue component of infinite multiplicity. Bull London Math Soc 29:195–199
82. Iwanik A, Lemańczyk M, Rudolph D (1993) Absolutely continuous cocycles over irrational rotations. Isr J Math 83:73–95
83. Iwanik A, Lemańczyk M, de Sam Lazaro J, de la Rue T (1997) Quelques remarques sur les facteurs des systèmes dynamiques gaussiens. Stud Math 125:247–254
84. Iwanik A, Lemańczyk M, Mauduit C (1999) Piecewise absolutely continuous cocycles over irrational rotations. J London Math Soc 59(2):171–187
85. Iwanik A, de Sam Lazaro J (1991) Sur la multiplicité L^p d'un automorphisme gaussien. CR Acad Sci Paris Ser I 312:875–876
86. Janicki A, Weron A (1994) Simulation and chaotic behavior of α-stable stochastic processes. Monographs and Textbooks in Pure and Applied Mathematics, vol 178. Marcel Dekker, New York, pp 355
87. del Junco A (1976) Transformations with discrete spectrum are stacking transformations. Canad J Math 28:836–839
88. del Junco A (1977) A transformation with simple spectrum which is not of rank one. Canad J Math 29:655–663
89. del Junco A (1981) Disjointness of measure-preserving transformations, minimal self-joinings and category. Prog Math 10:81–89
90. del Junco A, Lemańczyk M (1992) Generic spectral properties of measure-preserving maps and applications. Proc Amer Math Soc 115:725–736
91. del Junco A, Lemańczyk M (1999) Simple systems are disjoint with Gaussian systems. Stud Math 133:249–256
92. del Junco A, Lemańczyk M (2005) Joinings of distally simple systems. (in press)
93. del Junco A, Rudolph D (1987) On ergodic actions whose self-joinings are graphs. Ergod Theory Dyn Syst 7:531–557
94. Kachurovskii AG (1990) A property of operators generated by ergodic automorphisms. Optimizatsiya 47(64):122–125, (Russian)
95. Kalikow S (1984) Two fold mixing implies three fold mixing for rank one tranformations. Ergod Theory Dyn Syst 4:237–259
96. Kamae T Spectral properties of automata generating sequences. (in press)
97. Kamiński B (1981) The theory of invariant partitions for $\mathbb{Z}^d$-actions. Bull Acad Polon Sci Ser Sci Math 29:349–362
98. Katok AB Constructions in ergodic theory. (in press)
99. Katok AB (1977) Monotone equivalence in ergodic theory. Izv Akad Nauk SSSR Ser Mat 41:104–157, (Russian)
100. Katok AB (1980) Interval exchange transformations and some special flows are not mixing. Isr J Math 35:301–310
101. Katok AB (2001) Cocycles, cohomology and combinatorial constructions in ergodic theory. In: Robinson Jr. EA (ed) Proc Sympos Pure Math 69, Smooth ergodic theory and its applications, Seattle, 1999. Amer Math Soc, Providence, pp 107–173
102. Katok AB (2003) Combinatorial constructions in ergodic theory and dynamics. In: University Lecture Series 30. Amer Math Soc, Providence
103. Katok AB, Lemańczyk M (2007) Some new cases of realization of spectral multiplicity function for ergodic transformations. (in press)
104. Katok AB, Stepin AM (1967) Approximations in ergodic theory. Uspekhi Mat Nauk 22(137):81–106, (Russian)
105. Katok AB, Thouvenot J-P (2006) Spectral Properties and Combinatorial Constructions in Ergodic Theory. In: Handbook of dynamical systems, vol 1B. Elsevier, Amsterdam, pp 649–743
106. Katok AB, Zemlyakov AN (1975) Topological transitivity of billiards in polygons. Mat Zametki 18:291–300, (Russian)
107. Keane M (1968) Generalized Morse sequences. Z Wahr Verw Geb 10:335–353
108. Keane M (1975) Interval exchange transformations. Math Z 141:25–31
109. Khanin KM, Sinai YG (1992) Mixing of some classes of special flows over rotations of the circle. Funct Anal Appl 26:155–169
110. King JL (1986) The commutant is the weak closure of the powers, for rank-1 transformations. Ergod Theory Dyn Syst 6:363–384
111. King JL (1988) Joining-rank and the structure of finite rank rank mixing transformations. J Anal Math 51:182–227
112. King JL (2001) Flat stacks, joining-closure and genericity. (in press)
113. Klemes I (1996) The spectral type of the staircase transformation. Tohoku Math J 48:247–248
114. Klemes I, Reinhold K (1997) Rank one transformations with singular spectral type. Isr J Math 98:1–14
115. Kocergin AV (1972) On the absence of mixing in special flows over the rotation of a circle and in flows on a two-dimensional torus. Dokl Akad Nauk SSSR 205:949–952, (Russian)
116. Kochergin AV (1975) Mixing in special flows over rearrangment of segments and in smooth flows on surfaces. Math USSR Acad Sci 25:441–469
117. Kochergin AV (1976) Non-degenerate saddles and absence of mixing. Mat Zametky 19:453–468, (Russian)
118. Kochergin AV (1976) On the homology of function over dynamical systems. Dokl Akad Nauk SSSR 231:795–798
119. Kochergin AV (2002) A mixing special flow over a rotation of the circle with an almost Lipschitz function. Sb Math 193:359–385
120. Kochergin AV (2004) Nondegenerate fixed points and mixing in flows on a two-dimensional torus II. Math Sb 195:15–46, (Russian)
121. Koopman BO (1931) Hamiltonian systems and transformations in Hilbert space. Proc Nat Acad Sci USA 17:315–318
122. Kuipers L, Niederreiter H (1974) Uniform distribution of sequences. Wiley, London, pp 407
123. Kushnirenko AG (1974) Spectral properties of some dynamical systems with polynomial divergence of orbits. Vestnik Moskovskogo Univ 1–3:101–108
124. Kwiatkowski J (1981) Spectral isomorphism of Morse dynamical systems Bull Acad Polon Sci Ser Sci Math 29:105–114
125. Kwiatkowski Jr. J, Lemańczyk M (1995) On the multiplicity function of ergodic group extensions II. Stud Math 116:207–215
126. Lemańczyk M (1988) Toeplitz Z_2–extensions. Ann Inst H Poincaré 24:1–43
127. Lemańczyk M (1996) Introduction to ergodic theory from the point of view of the spectral theory. In: Lecture Notes of the tenth KAIST mathematics workshop, Taejon, 1996, pp 1–153
128. Lemańczyk M (2000) Sur l'absence de mélange pour des flots spéciaux au dessus d'une rotation irrationnelle. Coll Math 84/85:29–41
129. Lemańczyk M, Lesigne E (2001) Ergodicity of Rokhlin cocycles. J Anal Math 85:43–86

130. Lemańczyk M, Mauduit C (1994) Ergodicity of a class of cocycles over irrational rotations. J London Math Soc 49:124–132
131. Lemańczyk M, Mentzen MK, Nakada H (2003) Semisimple extensions of irrational rotations. Stud Math 156:31–57
132. Lemańczyk M, Parreau F (2003) Rokhlin extensions and lifting disjointness. Ergod Theory Dyn Syst 23:1525–1550
133. Lemańczyk M, Parreau F (2004) Lifting mixing properties by Rokhlin cocycles. (in press)
134. Lemańczyk M, Parreau F (2007) Special flows over irrational rotation with simple convolution property. (in press)
135. Lemańczyk M, Parreau F, Roy E (2007) Systems with simple convolutions, distal simplicity and disjointness with infinitely divisible systems. (in press)
136. Lemańczyk M, Parreau F, Thouvenot J-P (2000) Gaussian automorphisms whose ergodic self–joinings are Gaussian. Fundam Math 164:253–293
137. Lemańczyk M, de Sam Lazaro J (1997) Spectral analysis of certain compact factors for Gaussian dynamical systems. Isr J Math 98:307–328
138. Lemańczyk M, Sikorski A (1987) A class of not local rank one automorphisms arising from continuous substitutions. Prob Theory Rel Fields 76:421–428
139. Lemańczyk M, Wasieczko M (2006) A new proof of Alexeyev's theorem. (in press)
140. Lemańczyk M, Wysokińska M (2007) On analytic flows on the torus which are disjoint from systems of probabilistic origin. Fundam Math 195:97–124
141. Leonov VP (1960) The use of the characteristic functional and semi–invariants in the ergodic theory of stationary processes. Dokl Akad Nauk SSSR 133:523–526, (1960) Sov Math 1:878–881
142. Maruyama G (1970) Infinitely divisible processes. Theory Prob Appl 15(1):1–22
143. Masur H (1982) Interval exchange transformations and measured foliations. Ann Math 115:169–200
144. Mathew J, Nadkarni MG (1984) Measure-preserving transformation whose spectrum has Lebesgue component of multiplicity two. Bull London Math Soc 16:402–406
145. Medina H (1994) Spectral types of unitary operators arising from irrational rotations on the circle group. Michigan Math J 41:39–49
146. Mentzen MK (1988) Some examples of automorphisms with rank *r* and simple spectrum. Bull Pol Acad Sci 7–8:417–424
147. Nadkarni MG (1998) Spectral theory of dynamical systems. Hindustan Book Agency, New Delhi
148. von Neumann J (1932) Zur Operatorenmethode in der Klassichen Mechanik. Ann Math 33:587–642
149. Newton D (1966) On Gaussian processes with simple spectrum. Z Wahrscheinlichkeitstheorie Verw Geb 5:207–209
150. Newton D, Parry W (1966) On a factor automorphism of a normal dynamical system. Ann Math Statist 37:1528–1533
151. Ornstein D (1970) On the root problem in ergodic theory. In: Proc. 6th Berkeley Symp Math Stats Prob, University California Press, Berkeley, 1970, pp 348–356
152. Ornstein D, Rudolph D, Weiss B (1982) Equivalence of measure preserving transformations. Mem Amer Math Soc 37(262):116
153. Ornstein D, Weiss B (1987) Entropy and isomorphism theorems for actions of amenable groups. J Anal Math 48:1–141
154. Parreau F (2000) On the Foiaş and Stratila theorem. In: Proceedings of the conference on Ergodic Theory, Toruń, 2000
155. Parreau F, Roy E (2007) Poissson suspensions with a minimal set od self-joinings. (in press)
156. Parry W (1981) Topics in ergodic theory. Cambridge Tracts in Mathematics, 75. Cambridge University Press, Cambridge-New York
157. Parry W (1970) Spectral analysis of *G*-extensions of dynamical systems. Topology 9:217–224
158. Petersen K (1983) Ergodic theory. Cambridge University Press, Cambridge
159. Prikhodko AA, Ryzhikov VV (2000) Disjointness of the convolutions for Chacon's automorphism. Dedicated to the memory of Anzelm Iwanik. Colloq Math 84/85:67–74
160. Queffelec M (1988) Substitution dynamical systems – spectral analysis. In: Lecture Notes in Math 1294. Springer, Berlin, pp 240
161. Ratner M (1978) Horocycle flows are loosely Bernoulli. Isr J Math. 31:122–132
162. Ratner M (1983) Horocycle flows, joinings and rigidity of products. Ann Math 118:277–313
163. Rauzy G (1979) Echanges d'intervalles et transformations induites. Acta Arith 34:315–328
164. Robinson EA Jr (1983) Ergodic measure preserving transformations with arbitrary finite spectral multiplicities. Invent Math 72:299–314
165. Robinson EA Jr (1986) Transformations with highly nonhomogeneous spectrum of finite multiplicity. Isr J Math 56:75–88
166. Robinson EA Jr (1988) Nonabelian extensions have nonsimple spectrum. Compos Math 65:155–170
167. Robinson EA Jr (1992) A general condition for lifting theorems. Trans Amer Math Soc 330:725–755
168. Rosiński J, Zak T (1996) Simple condition for mixing of infinitely divisible processes. Stoch Process Appl 61:277–288
169. Rosiński J, Zak T (1997) The equivalence of ergodicity and weak mixing for infinitely divisible processes. J Theor Probab 10:73–86
170. Roy E (2005) Mesures de Poisson, infinie divisibilité et propriétés ergodiques. In: Thèse de doctorat de l'Université Paris 6
171. Roy E (2007) Ergodic properties of Poissonian ID processes. Ann Probab 35:551–576
172. Royden HL (1968) Real analysis. McMillan, New York
173. Rudin W (1962) Fourier analysis on groups. In: Interscience Tracts in Pure and Applied Mathematics, No. 12 Interscience Publishers. Wiley, New York, London, pp 285
174. Rudolph DJ (1979) An example of a measure-preserving map with minimal self-joinings and applications. J Anal Math 35:97–122
175. Rudolph D (1985) *k*-fold mixing lifts to weakly mixing isometric extensions. Ergod Theor Dyn Syst 5:445–447
176. Rudolph D (1986) $\mathbb{Z}^n$ and $\mathbb{R}^n$ cocycle extension and complementary algebras. Ergod Theor Dyn Syst 6:583–599
177. Rudolph D (1990) Fundamentals of measurable dynamics. Oxford Sci Publ, pp 169
178. Rudolph D (2004) Pointwise and L^1 mixing relative to a subsigma algebra. Illinois J Math 48:505–517
179. Rudolph D, Weiss B (2000) Entropy and mixing for amenable group actions. Ann Math 151(2):1119–1150
180. de la Rue T (1996) Systèmes dynamiques gaussiens d'entropie nulle, lâchement et non lâchement Bernoulli. Ergod Theor Dyn Syst 16:379–404

181. de la Rue T (1998) Rang des systèmes dynamiques gaussiens. Isr J Math 104:261–283
182. de la Rue T (1998) L'ergodicité induit un type spectral maximal équivalent à la mesure de Lebesgue. Ann Inst H Poincaré Probab Statist 34:249–263
183. de la Rue T (1998) L'induction ne donne pas toutes les mesures spectrales. Ergod Theor Dyn Syst 18:1447–1466
184. de la Rue T (2004) An extension which is relatively twofold mixing but not threefold mixing. Colloq Math 101:271–277
185. Ryzhikov VV (1991) Joinings of dynamical systems. Approximations and mixing. Uspekhi Mat Nauk 46 5(281):177–178 (in Russian); Math Surv 46(5)199–200
186. Ryzhikov VV (1992) Mixing, rank and minimal self-joining of actions with invariant measure. Math Sb 183:133–160, (Russian)
187. Ryzhikov VV (1994) The absence of mixing in special flows over rearrangements of segments. Math Zametki 55:146–149, (Russian)
188. Ryzhikov VV (1999) Transformations having homogeneous spectra. J Dyn Control Syst 5:145–148
189. Ryzhikov VV (2000) The Rokhlin problem on multiple mixing in the class of actions of positive local rank. Funkt. Anal Prilozhen 34:90–93, (Russian)
190. Ryzhikov VV, Thouvenot J-P (2006) Disjointness, divisibility, and quasi-simplicity of measure-preserving actions. Funkt Anal Prilozhen 40:85–89, (Russian)
191. Ryzhikov VV (2007) Weak limits of powers, simple spectrum symmetric products and mixing rank one constructions. Math Sb 198:137–159
192. Sato K-I (1999) Lévy Processes and infinitely divisible distributions. In: Cambridge Studies in Advanced Mathematics, vol 68. Cambridge University Press, Cambridge, pp 486
193. Schmidt K (1977) Cocycles of Ergodic Transformation Groups. In: Lecture Notes in Math. 1. Mac Millan, India
194. Schmidt K (2002) Dispersing cocycles and mixing flows under functions. Fund Math 173:191–199
195. Schmidt K, Walters P (1982) Mildly mixing actions of locally compact groups. Proc London Math Soc 45(3)506–518
196. Sinai YG (1994) Topics in ergodic theory. Princeton University Press, Princeton
197. Shklover M (1967) Classical dynamical systems on the torus with continuous spectrum. Izv Vys Ucebn Zaved Matematika 10:(65)113–124, (Russian)
198. Smorodinsky M, Thouvenot J-P (1979) Bernoulli factors that span a transformation. Isr J Math 32:39–43
199. Stepin AM (1986) Spectral properties of generic dynamical systems. Izv Akad Nauk SSSR Ser Mat 50:801–834, (Russian)
200. Tikhonov SV (2002) On a relation between the metric and spectral properties of $\mathbb{Z}^d$-actions. Fundam Prikl Mat 8:1179–1192
201. Thouvenot J-P (1995) Some properties and applications of joinings in ergodic theory. London Math Soc Lect Note Ser 205:207–235
202. Thouvenot J-P (2000) Les systèmes simples sont disjoints de ceux qui sont infiniment divisibles et plongeables dans un flot. Coll Math 84/85:481–483
203. Ulcigrai C (2007) Mixing of asymmetric logarithmic suspension flows over interval exchange transformations. Ergod Theor Dyn Syst 27:991–1035
204. Walters P (1982) An Introduction to Ergodic Theory. In: Graduate Texts in Mathematics, vol 79. Springer, Berlin, pp 250
205. Wysokińska M (2004) A class of real cocycles over an irrational rotation for which Rokhlin cocycle extensions have Lebesgue component in the spectrum. Topol Meth Nonlinear Anal 24:387–407
206. Wysokińska M (2007) Ergodic properties of skew products and analytic special flows over rotations. In: Ph D thesis. Toruń, 2007.
207. Veech WA (1978) Interval exchange transformations. J Anal Math 33:222–272
208. Veech WA (1982) A criterion for a process to be prime. Monatshefte Math 94:335–341
209. Veech WA (1982) Gauss measures for transformations on the space of interval exchange maps. Ann Math 115(2):201–242
210. Veech WA (1984) The metric theory of interval exchange transformations, I Generic spectral properties. Amer J Math 106:1331–1359
211. Vershik AM (1962) On (1962) the theory of normal dynamic systems. Math Sov Dokl 144:625–628
212. Vershik AM (1962) Spectral and metric isomorphism of some normal dynamical systems. Math Sov Dokl 144:693–696
213. Yassawi R (2003) Multiple mixing and local rank group actions. Ergod Theor Dyn Syst 23:1275–1304
214. Zeitz P (1993) The centralizer of a rank one flow. Isr J Math 84:129–145

Spin Dependent Exchange and Correlation in Two-Dimensional Electron Layers

M. W. Chandre Dharma-wardana
Institute of Microstructural Sciences,
National Research Council of Canada, Ottawa, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
Ideally Thin 2D Electron Layers
2D Layers with Finite Thickness
Multi-Component 2D Layers
Graphene: 2D Two-Valley System on a Honey-Comb Lattice
Exchange and Correlation in a Magnetic Field
Exchange and Correlation at Finite Temperatures
Conclusion
Bibliography

Glossary

Atomic units, a. u. The electron-charge $|e|$, and the mass m_e are taken as unity. The unit of time is fixed by setting the Plank constant $\hbar$ to unity. The Bohr radius $a_0 = \hbar^2/(m_e e^2)$ is one a.u. of length in the

centimeter-gram-second(CGS) system which uses the "esu"(electrostatic unit of charge). The SI system uses the meter, kilogram, second (with the Ampere as the unit of current), $a_0 = (4\pi\varepsilon_0\hbar^2/(m_e e^2)$ where ε is the electric permittivity of the vacuum. The value of a_0 is 5.29177×10^{-9} cm. The unit of energy, the Hartree, is e^2/a_0 in the CGS system, being 27.2116 eV. In semiconductor physics, *effective atomic units* are used, with e^2/ε, replacing e^2, where ε is the dielectric constant. The band mass m_b is used instead of m_e, so that the effective Bohr radius $a_0^* = a_0(\varepsilon/m^*)$, where m^* is the effective mass. Then the effective Hartree is of the order of millivolts.

Confining potential This potential keeps the electron in a given spatial region. It is due to the physical structure of the device, the applied gate voltages etc.

Correlation energy The contribution to the total energy beyond the Hartree–Fock approximation, denoted by E_c.

Correlation hole is the depletion of electron density near an electron due to Coulomb repulsion effects.

Coupling constant is the ratio of the potential energy(PE) to the kinetic energy (KE). In a classical electron fluid, the KE (thermal energy) is T, and the PE is $1/r_s$ (atomic units), and r_s is the Wigner–Seitz radius. The coupling constant $=$ PE/KE $= \Gamma = 1/(r_s T)$. In quantum systems, the Fermi energy E_F is used instead of T for the KE, and $\Gamma = r_s$.

Classical-map hyper-netted-chain (CHNC) A method for using the classical hyper-netted-chain equation to calculate the correlation functions of quantum systems.

Effective mass is denoted by m^*, and is the band mass m_b in units of the electron mass m_e.

Exchange energy The part of the Hartree–Fock energy due to electron exchange, i. e., the "Fock" part, denoted by E_x. It is first order in the Coulomb interaction.

Fermi hole This denotes the reduction in the probability of finding a like-spin electron near another, due to Fermi statistics.

Hartree–Fock Hartree's self-consistent one-body approximation for interacting electrons is based on a product wavefunction. Fock included exchange using an antisymmetrized product. "Hartree–Fock" is the label for calculations of the energy, wavefunctions etc., where the electron moves in this mean potential generated by the electrostatics and the exchange effects. The Hartree term is zero in uniform systems.

HIGFET Heterojunction-Insulated-Gate Field-effect Transistor.

Jellium A model "metal" where the positive ionic charges are replaced by a uniform static charge which neutralizes the free-electron charge.

MOSFET Metal-oxide semiconductor field-effect transistor.

Heterojunction A semiconductor-interface involving two dissimilar materials.

Hyper-netted-chain (HNC) A classical integral equation due to van Leeuwen, Groenveld and de Boer (1959) which non-perturbatively sums "hyper-netted-chain" diagrams, going beyond mean-field theory.

Electron gas parameter $\mathbf{r_s}$ See Wigner–Seitz radius.

Pair-correlation function Denoted by $h(\vec{r}) = g(\vec{r}) - 1$ where $g(r)$ is the pair-distribution function(PDF).

Pair-distribution function The pair-distribution function(PDF), $g(\vec{r})$, is the probability of finding a particle at the location $\vec{r}$, given a reference particle at the origin.

Plasma analogy A class of methods for approximately replacing a charged quantum fluid by an equivalent classical fluid at a finite temperature.

Pseudo-spin Discrete degrees of freedom beside the electron spin. The electrons in Si/SiO_2 interfaces occupy two valleys. The valley index is a pseudospin.

Quantum-Monte Carlo (QMC) In molecular dynamics (MD), Newton's equations of motion are integrated using a stochastic scheme based on the Metropolis algorithm. In QMC a trial wavefunction provides a probability measure for the Metropolis algorithm. The wavefunction is optimized in various ways, leading to "variational QMC", where the nodes of the trial wavefunction are held fixed. In "Diffusion QMC", the nodes are also relaxed.

Random-phase approximation (RPA) A time-dependent self-consistent field method where an electron with momentum $\vec{k}$ moves in an effective potential which contains the external potential and a $\vec{k}, \omega$ dependent screened potential. It reduces in the static $k \to 0$ limit to Thomas-Fermi screening, (or Debye–Hŭkel screening in classical systems). It is also called the "ring sum" or "bubble sum", and contains no exchange effects.

Subbands The electrons with the z-motion confined to a quantum well have discrete energy levels (index n). Each level carries with it a band of energies for the x, y motion. These are energy "subbands".

Singwi–Tosi–Land–Sjölander (STLS) A method due to Singwi et al. for determining the density-density correlation function of electrons (and other quantum systems) by truncating the equation of motion via an intuitive decoupling scheme involving the PDFs. STLS has been extended by Vashista, Ichimaru and others.

Wigner–Seitz radius In 2D this is the radius, denoted by r_s, of the circle containing, on the average, just one electron.

Definition of the Subject

Since the advent of density-functional theory (DFT), the exchange-correlation energy E_{xc} of an interacting system has become a basic quantity in many-particle theory. Here we study the E_{xc} of two-dimensional (2D) electron layers. Such layers contain electrons which move in the x and y directions, while confined in z. 2D layers are formed at insulator-semiconductor interfaces in heterojunctions, and more particularly at metal-oxide-semiconductor(MOS) interfaces. These include two types of semiconductors (e. g., GaAs and the alloyed form $Al_xGa_{1-x}As$, containing a small fraction x of Al, and written as AlGaAs for brevity). The interface region defines a "confining potential" where an electron layer may from [2]. SiO_2 is an insulator with a large bandgap, while Si can be doped in a controlled manner to behave as a conductor. The Si/SiO_2 interface supports the formation of an electron layer at the interface. The electron density n in such layers can be controlled using external potentials, and it is this which is the key to the importance of these materials systems. These structures are the basis of metal-oxide field-effect transistors (MOSFETS), ubiquitous in modern electronic devices (Fig. 1). The Si-MOSFET was developed in the 1960s, while it had been fore-shadowed as early as the 1930s in the work of Shockley and others.

Electron layers confined at the air-liquid interface of a fluid (e. g., Helium) were studied in the 1970s, but did not provide easy tunability of the density. High-mobility 2D systems in GaAs/GaAlAs heterostructures and in Si-MOSFETS ushered the more recent phase, unearthing new physics and new technologies, with the quantum Hall effect(QHE) [4], discovered in 1980, providing a classic example of fundamental physics directly leading to practical applications in metrology. Modern nanotechnology, spintronics, plasmonics etc., greatly dependent on the theory of electron layers.

Graphite is made of 2D sheets of carbon atoms, i. e., "graphene" sheets, held together by weak inter-layer bonds The 2D electrons in graphene behave as massless Fermions with $\epsilon(\vec{k}) = \hbar v_F \vec{k}$, where v_F is the Fermi velocity. The hexagonal unit cell of graphene contains two inequivalent Carbon atoms. Thus there are two degenerate conduction bands and two valance bands, with a zero band gap located at two inequivalent points, labeled **K** and **K′**, in the hexagonal 2D-Brillouin zone of graphene. This system needs a spin index, a pseudospin index and also a band index. Graphene is a rich 2D system with novel physics (e.g, QHE) and the promise of new technological applications [9].

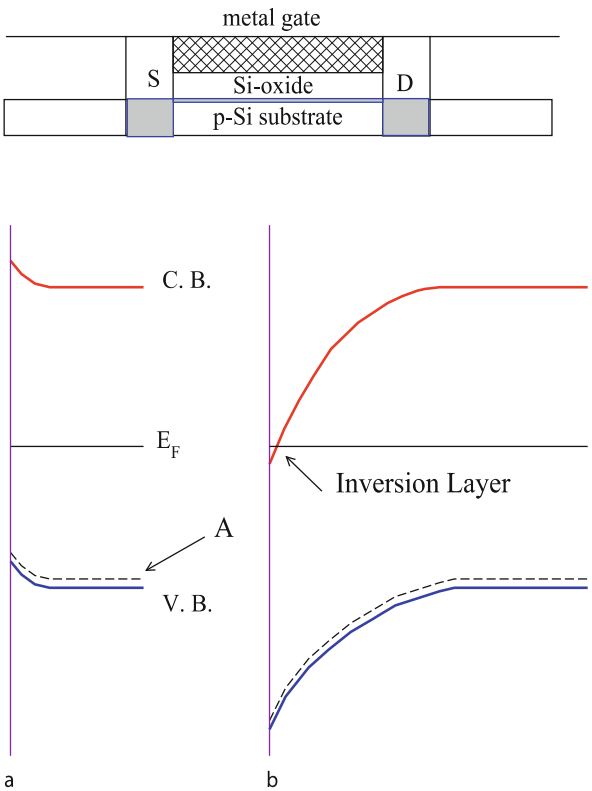

Spin Dependent Exchange and Correlation in Two-Dimensional Electron Layers, Figure 1
***Top*: A cross section of a *p*-type MOSFET. The *n*-type contacts labeled S (source) and D (drain) are made by ion-implantation into the *p*-type substrate. A voltage on the metal gate controls the current between S and D flowing in the Si/SiO_2 interface (channel region). Panel a shows the conduction band (c.b.), the acceptor levels (A), and the valance band (v.b.) modified by the application of negative bias to the metal gate. A sufficiently strong negative bias can produce an accumulation layer of holes. In b a sufficiently positive gate voltage is applied, bending the conduction band *below* the Fermi energy E_F, and creating a potential well which supports a 2D layer of electrons, known as an 'inversion layer'**

The Coulomb interaction conspires to destroy the "single-particle" (i. e., "free" particle) picture of electron systems. Only the kinetic energy and the confining potential (called the "external potential") are needed in the simplest models. Mean-field (e.g, Hartree–Fock) models provide an "effective" single-particle picture. The "many-body effects", which transcend the simple picture, lead to new effects like plasmons, charge density waves, spin-polarized states, Wigner crystallization, fractional-QHE, and super-

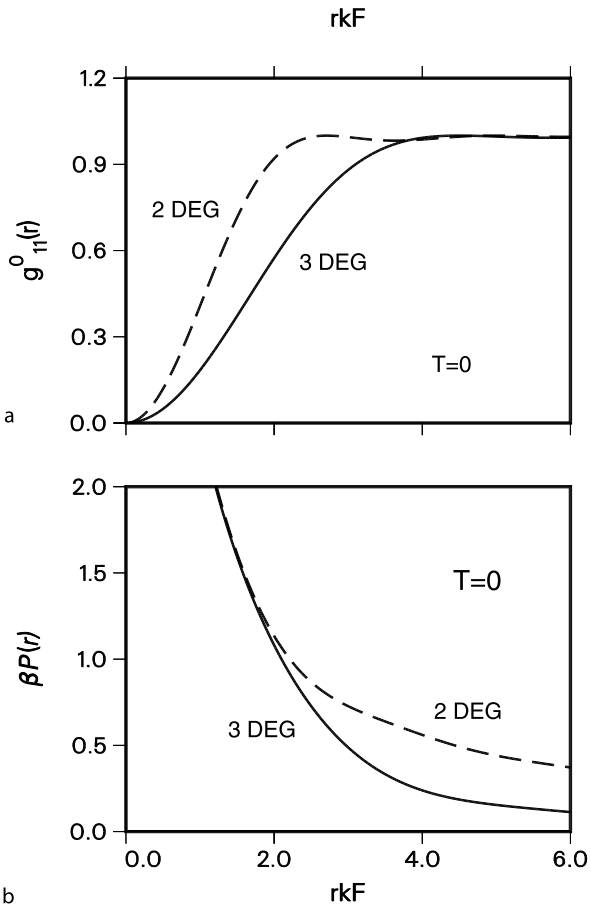

Spin Dependent Exchange and Correlation in Two-Dimensional Electron Layers, Figure 2
Upper panel **shows the pair-distribution function (PDF) for non-interacting parallel-spin electrons in 2D and 3D electron gases. The depletion near the origin is the "Fermi hole". The** ***lower panel*** **shows the equivalent** ***classical*** **potential (Pauli exclusion potential, in $\beta = 1/T$ units) which would create such a depletion in the PDF at the temperature T (at $T = 0$, only the product $\beta P(r)$ is meaningful). The x-axis is the electron-pair separation in reciprocal k_F units**

conductivity [18]. Many-body effects are more enhanced in low dimensional (e. g., 1D and 2D systems in comparison to 3D) systems. The Coulomb interactions force two electrons to repel each other and form a "Coulomb hole" or correlation hole, i. e., a region of low probability of approach. The energy consequence of this is the "correlation energy". Similarly, since the Pauli principle forbids the presence of two non-interacting electrons of the same spin in a common spatial eigenfunction, a "Fermi hole " is formed and leads to an exchange energy (see Fig. 2).

There is no Fermi hole for antiparallel spins, and their non-interacting PDF, i. e., $g^0_{12}(r)$ is unity for all r. When interactions are included, a Coulomb hole is formed. Hence the evaluation of E_{xc} is directly related to obtaining good PDFs.

Thus the theory of E_{xc} is of great importance in modern density-functional many-body approaches. From a technological point of view, the subject is of critical importance in the proper design of modern nano-structure devices and quantum-well lasers [16].

Introduction

We use "effective" atomic units, with the effective Bohr radius a_0^* as the unit of length. The dielectric constant ε is of the order of 10–12 for common semiconductors, while m^* may be 0.06 in GaAs and 0.19 in Si. Hence the effective Hartree is measured in meV, unlike the atomic Hartree which is 27.12 eV. Thus the properties of the material enter into the theory only via ε, m^*, and the effective Landé factor g^*, and these are usually absorbed into the effective atomic units. Since the electrons reside in the interface between two materials (say A, B),the dielectric constant ε relevant to the electron layer has to be evaluated from the individual dielectric constants of A and B. This is trivial in the case of GaAs/AlGaAs systems, since the dielectric constants are very close. This is no longer the case with Si/SiO_2 systems where the Si and the Si-oxide have nominal dielectric constants of 11.5 and 3.9 respectively. Then an average dielectric constant has to be used, as discussed in the appendix of [2], and in [8].

If the z-motion of an electron is restricted to a finite length a_z, as in an inversion layer, the allowed quantum states form *discrete subbands*, with energies $E(n, k_x, k_y)$:

$$E_n(k_x, k_y) = \epsilon_n + \epsilon(k_x, k_y) \,.$$

The energy of the in-plane motion, $\epsilon(k_x, k_y)$, depends on the in-plane momenta $\hbar k_x$ and $\hbar k_y$ with $\vec{k} = \vec{k}_x + \vec{k}_y$. For materials like GaAs/AlGaAs, or Si/SiO_2 interfaces, the in-plane energy dispersion is essentially parabolic, with

$$\epsilon(k_x, k_y) = \frac{(\hbar k_x)^2 + (\hbar k_y)^2}{2m^*} \,.$$

Here m^* is an effective mass, assumed to be the same in the x and y directions Only the lowest subband, say $n = 1$, is occupied in a 2D layer. Let ϵ_1 be the energy zero. The highest occupied state is given by the in-plane momentum $\hbar k_F$, known as the *Fermi momentum*. The Fermi energy (ϵ_{k_F}) has to be far below the bottom of the second subband to ensure a 2D layer. The electron density, i. e., the number of electrons n per unit area has n_1 "up-spin", and n_2 "down-spin" electrons. If $n_1 \neq n_2$ we have a *spin-polarized* electron system. The *degree of spin polarization*

ζ is:

$$\zeta = \frac{n_1 - n_2}{n_1 + n_2} .$$

Thus the state of a clean 2D-electron layer may be specified by n, ζ, and the temperature T. Additional pseudo-spin or valley indices are sometimes necessary.

The Fermi Energy and the Role of r_s as a Coupling Constant

When the electron-number density per a.u. of area is n, the Wigner–Seitz radius r_s is given by $r_s = 1/\sqrt{(\pi n)}$. We also have $n_1 = nx_1$, $x_1 = (1+\zeta)/2$, $n_2 = nx_2$, $x_2 = (1-\zeta)/2$. The Fermi momentum $k_{F\sigma}$ of a given spin species σ is evaluated by requiring that the sum of all occupied states of the species σ adds upto the density n_σ.

$$\int_0^{k_F} 2\pi \frac{k\mathrm{d}k}{(2\pi)^2} = n_\sigma .$$

This gives $k_\sigma = (4\pi n_\sigma)^{1/2}$. Since $n = n_1 + n_2$, the overall $k_F = (2\pi n)^{1/2}$, i. e., $k_F = \sqrt{2}/r_s$ and $E_F = k_F^2/2 = 1/r_s^2$.

Hence, the "coupling strength", i. e., ratio of the Coulomb energy $(1/r_s)$ to the kinetic energy $(\sim E_F)$ at $T = 0$ is clearly r_s. High density systems $(r_s < 1)$ are weakly coupled, while low density systems (high r_s) are strongly correlated. When $r_s \sim 26 - 27$, the 2D electron system acquires a ferromagnetic ground state. At stronger coupling, $r_s \simeq 35$, a 2D Wigner crystal is formed. The strength of the 2D massless electron interactions in graphene cannot be specified by an r_s since a Bohr radius cannot be specified. Instead, the coupling constant g is taken as the ratio of a typical Coulomb energy $(1/a_0)$ to the hopping energy (t) on the Hexagonal 2D unit cell with a lattice constant of a_0.

$$g = \frac{e^2/a_0\varepsilon_0}{\hbar E_F} = \left\{e^2/a_0\varepsilon_0\right\}\left\{t\sqrt{3}/2\right\} .$$

If $\varepsilon_0 = 1$, then $g \simeq 2.7$. Thus, when the graphene sheet is placed on most substrates, $\varepsilon > 1$, and the electron interactions become weak, i. e., $g \leq 1$.

Fourier Transforms

The Fourier transform of 2D-functions relating their r-space forms with q-space forms is very useful. If the 2D electron system is uniform in all directions of the $x - y$ plane, i. e., isotropic, then we only need *radial* Fourier transforms. These are obtained using the relations:

$$F(\vec{r}) = \int_0^\infty F(\vec{q}) J_0(qr) q \mathrm{d}q/(2\pi)$$

$$F(\vec{q}) = \int_0^\infty F(\vec{r}) J_0(qr) 2\pi r \mathrm{d}r .$$

Here $J_0(qr)$ is a Bessel function [1] such that:

$$J_0(qr) = 2\int_0^\pi cos(qr\cos\theta) d\theta .$$

Then it is easy to show that the Coulomb interaction $1/r$ has the Fourier form $2\pi/q$. Here we assume ideally thin layers (unlike in Sect. "Graphene: 2D Two-Valley System on a Honey-Comb Lattice"). A screened Coulomb interaction $V_s(r)$, with an exponential damping is often encountered. This potential has an analytic Fourier transform and is:

$$V_s(r) = \exp(-k_s r)/r \qquad V_s(q) = 2\pi/\left\{\left(q^2 + k_s^2\right)^{1/2}\right\} .$$

The potential contains a "screening wavevector" k_s, and appears in what is known as "Thomas-Fermi" screening. This type of potential is also called a "Yukawa potential". The 'range' of the potential is of the order of $1/k_s$, and is known as the screening length.

The Hamiltonian of the System

The total Hamiltonian contains a contribution from a *uniform* neutralizing background of positive charge n_b which is static. In a real system, e. g., a metal, the neutralizing background is provided by the ion subsystem. Each ion has a short-ranged core-region and a Coulomb-like long-range potential which overlaps the long-range potential of the neighbours. In many metals (e. g., Sodium) the resulting overlapping positive potential is, to a very good approximation, similar to a smudged-out, structureless "jellium". In a plasma, the positive ions are dynamical but extremely heavy, and hence a jellium approximation, known as the one-component plasma (OCP) is often used. The jellium model is used in theories of electron layers as the physics then focuses entirely on electron-electron effects.

We write the electron density operator and the Coulomb interaction as:

$$n(\vec{r}) = \sum_i \delta(\vec{r} - \vec{r}_i) , \quad V_{ee} = \frac{1}{2}\sum_{ij} \frac{n(\vec{r}_i) n(\vec{r}_j)}{|\vec{r}_i - \vec{r}_j|} .$$

The latter has a divergence when $i \to j$, with the electron interacting with "itself". We also have the terms $V_{\rm bb}$ and $V_{\rm be}$, involving the background. For example,

$$V_{\rm be} = -\sum_{ij} n_{\rm b} \frac{n(\vec{r}_j)}{|\vec{r}_i - \vec{r}_j|}$$

where $n_{\rm b} = n$ is the uniform-background density. The background term $V_{\rm bb}$ has a divergent self-interaction term similar to that in $V_{\rm ee}$, with (positive) sign. However, the interaction of the electron charge with the background, i. e., $V_{\rm be}$, has a negative divergent term, and cancels both divergent (positive) self-interaction terms. Thus the total Hamiltonian, inclusive of the background is free of divergences. In Fourier space, the coulomb interaction is $2\pi/q$, and hence the divergence manifests in the limit $q \to 0$. This is exactly canceled by the corresponding $q \to 0$ terms in the background. Also, since the background is uniform, there are no other Fourier components. The above discussion involves the "subtraction" of divergent quantities, and can be made more rigorous using the screened Coulomb potential $V_{\rm s}(r)$, and taking the limit $k_{\rm s} \to 0$ at the end.

The electron number-density operator $n(r)$ in real space has a Fourier transform $n(q)$, defining the occupation number in the momentum state q. The expectation value of $n(r)$ is of course the constant n, which may be written as $\overline{n}$ where needed.

The momentum eigenstates of the uniform noninteracting system are the complete set of plane waves.

$$\phi_q = \phi(\vec{q}) = {\rm e}^{\vec{q}\cdot\vec{r}}/\Omega^{1/2}$$

Here we use a normalization volume Ω, which is sometimes set to unity, or the infinite-volume limit is taken, when the normalization becomes $1/(2\pi)^{1/2}$. The occupation number in each momentum state, for non-interacting Fermions at a temperature T, and chemical potential μ is given by:

$$n_k = \frac{1}{1 + {\rm e}^{\beta(\epsilon_k - \mu)}} \,.$$

Here $\beta = 1/T$ where T is expressed in energy units. At $T = 0$, the chemical potential μ becomes the Fermi energy $E_{\rm F}$, and the occupation numbers reduce to unity for energies $\epsilon_k < E_F$, and zero for higher energies. The occupation number n_k given above is really the expectation value of the operator n_k. When we need to emphasize this distinction, we write the operator as $\hat{n}_k$, and the meanvalue as $\overline{n}_k$. In a uniform system, most properties are dependent only on the modulus $|k|$ of the momentum $\vec{k}$, and we sometimes suppress the vector notation for brevity. We also introduce creation and annihilation operators a_k^+ and a_k which add an electron to a momentum state $\vec{k}$, or remove an electron from the momentum state $\vec{k}$. Thus the number operator acting on the eigenstate $|\phi_k\rangle$ is given as:

$$\hat{n}_k|\phi_k >= a_k^+ a_k|\phi_k\rangle = \overline{n}_k|\phi_k\rangle \,.$$

The total Hamiltonian of the system is:

$$H = (H^0 - \mu N) + V'_{\rm ee} \,.$$

Here we are using a thermal ensemble which allows for the definition of a chemical potential μ and a temperature T. Also, $N = \sum_{\vec{k}} \hat{n}_k$ is an operator fixing the mean

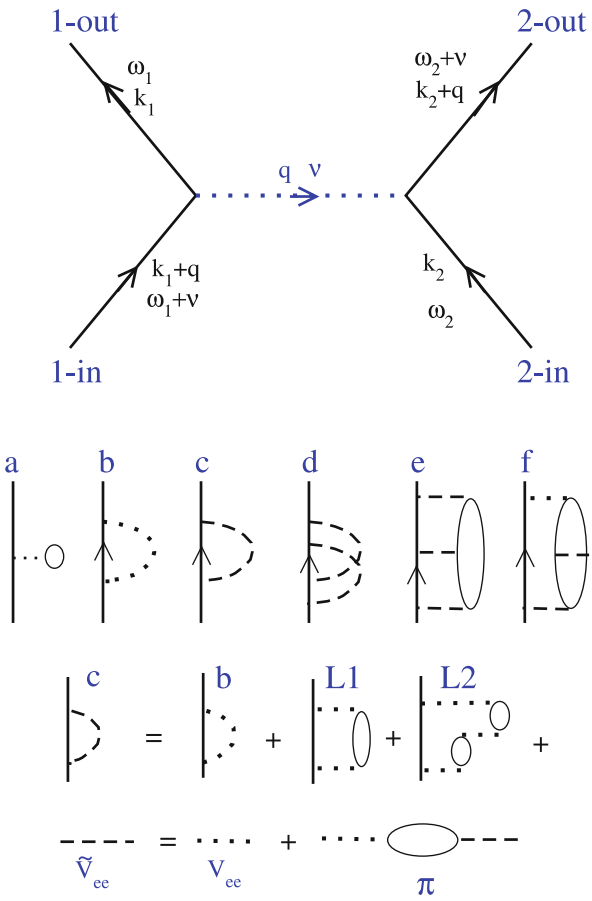

Spin Dependent Exchange and Correlation in Two-Dimensional Electron Layers, Figure 3

***Top Panel*: Feynman diagram for the bare Coulomb interaction $V_{\rm ee}(q)$ of two electrons. *Bottom panel*: a shows the exchange interaction which is first order in the bare interaction $V_{\rm ee}(q)$. b is the bare exchange, and c has the RPA "bubble sum", where $\tilde{V}_{\rm ee}(q, \nu)$ is given by an integral equation (a geometric series in this case) involving the polarization "bubble" $\pi(q, \nu)$. Contributions from diagrams like c–d are not included in the RPA propagators**

number of particles $\langle N \rangle$ in the volume Ω, given μ and T. The Coulomb interaction V'_{ee} carries a prime to indicate that a jellium background has been included to remove the divergencies at zero momentum transfer. Then we have, in momentum space:

$$H = \sum_{\vec{k}} (\epsilon_k - \mu) a_k^+ a_k + \frac{1}{2\Omega} \sum_{\vec{k}_1 \vec{k}_2 \vec{q}}' V_q a_{k_1}^+ a_{k_2+q}^+ a_{k_2} a_{k_1+q}.$$

The prime on the summation indicates that the case $q = 0$ is excluded, as it is exactly cancelled by the jellium background. The suffixes on the cluster of four operators are such that the total momentum of the two electrons is conserved during the interaction. That is, an electron in the initial state $\vec{k}_1 + \vec{q}$ is in the final state $\vec{k}_1$, while the second electron, in the initial state $\vec{k}_2$ is transfered to the state $\vec{k}_2 + \vec{q}$. The momentum transfer $\vec{q}$ is effected through the Coulomb interaction $V_q = 2\pi/q$. The momentum is conserved in the Coulomb collision since we are dealing with a *uniform system* (see top panel in Fig. 3).

Ideally Thin 2D Electron Layers

This system is most closely approximated in Si/SiO_2 systems, or in specially fabricated GaAs/GaAlAs quantum wells (HIGFETs tend to yield thick 2D electron layers and are discussed in Sect. "2D Layers with Finite Thickness"). Microscopic many-body theories have used this system as the basic "work horse". Early work used diagrammatic and other perturbation methods, or truncated equation of motion methods, like STLS, developed by Singwi and coworkers. The objective of the latter is to develop PDFs non-perturbatively, and calculate the exchange and correlation energies from an integration over the coupling constant (described below). However, the perturbation methods (restricted to the high-density regime in validity) provide basic reference results.

A Brief Outline of the Diagrammatic Method

Much of the initial work was based on a perturbation expansion of the energy or the electron propagator in powers of the Coulomb interaction $V_{ee}(q)$, as in Fig. 3. In the graphs labeled a–f we look at the self-energy (i.e, the effective potential) of an electron propagating through the interacting fluid. In (a) an electron propagates forward, while the interaction with the second electron occurs with zero-momentum transfer ($q = 0$), i.e, one may imagine that the "2-out" line "falls into" the "2-in" line to from a bubble, when $\vec{k}_2 + \vec{q} = \vec{k}_2$ by momentum conservation. Such $q = 0$ terms are Hartree terms, and have already been eliminated via the static background. The graph (b) is the first-order exchange diagram (the "Fock" term). This may be pictured as the "1-out" line "falling" into the "2-in" line, exchanging electron identities. This is possible only if both electrons have the same spin. The graph (c) is the *screened exchange* diagram where the Coulomb interaction is *dynamically screened* by summing the simple polarization loops L1, L2 etc. This is known as the RPA-screened exchange, where RPA (random-phase approximation) is a name originating from methods of derivation using self-consistent equations. However, each term in such an expansion is infinite; e. g., the second-order perturbation term L1, and those beyond it (L2 etc.) are all divergent, as they involve integrals of the form $\int (2\pi/q)^n q \mathrm{d}q \ldots$, where $n \geq 2$ (there are standard "Feynman rules" for converting the graphs shown in Fig. 3 into algebraic expressions [18]; they are not needed for our purpose). The divergence arises from perturbation terms containing simple polarization loops (denoted by $\pi(q, \omega)$ in Fig. 3) connected by two or more direct interaction lines. However, although each term is infinite, the *sum* L1 + L2 + ... is finite. The bare interaction $V_{ee}(q, \omega)$ gets "screened" by the polarization processes denoted by the bubble (particle- hole pairs), and gives the well-behaved (non-divergent) $\tilde{V}_{ee}(q)$ which is used in the graph (c). The contributions from scattering processes shown in Fig. 3d–f, etc., are *not* included in the RPA sum.

Thus the contribution to the effective potential seen by an electron at the Fermi level can be written as

$$V_{\mathrm{eff}} = V_{\mathrm{xc}} = V_{\mathrm{x}} + V_{\mathrm{c}}$$

where V_{x} is the exchange potential, arising from the first-order diagram Fig. 3b, while V_{c} is the contribution from *all other* higher processes. Thus V_{c} in the RPA is simply the sum of contributions from L1, L2 etc. A better approximation would be a V_{c} which involves Fig. 3e–f, and all other terms. But the numerous higher terms are virtually impossible to evaluate, and their partial inclusion is problematic as various sum rules, Ward identities etc., have to be satisfied [18]. In analogy with the effective potential, we can also divide the total energy of the system (beyond the kinetic energy) as:

$$E_{\mathrm{xc}} = E_{\mathrm{x}} + E_{\mathrm{c}}; \quad \epsilon_{\mathrm{x}} = E_{\mathrm{x}}/N; \quad \epsilon_{\mathrm{c}} = E_{\mathrm{c}}/N$$

where E_{x} is the exchange energy, and E_{c} is the correlation energy. The latter by definition contains all contributions other than exchange. A large part of the spin dependence of E_{xc} is contained in E_{x}. The RPA is an evaluation of E_{c} via the series L1, L2 etc.

The Adiabatic Connection Formula

Another approach to E_{xc} is to begin with the non-interacting system and "build-up" the interacting system by increasing the value of the "coupling constant" λ from $0 \to 1$ in the scaled Coulomb potential

$$V_{ee}(\lambda, r) = \lambda / r \,.$$

The usual coupling parameter r_s is here assumed to be merely a density parameter ($r_s = 1/\sqrt{\pi n}$) held constant during this "charging process". As λ increases to unity, the PDFs $g_{ij}(\lambda, r)$, where i, j indicate spin species, change from the non-interacting $g^0_{ij}(r)$, Fig. 2, to the interacting from $g_{ij}(r)$. This is accompanied by the "build up" of the energies arising from the interactions. Thus the exchange-correlation energy E_{xc} appears as an integration over the coupling constant (if we are dealing with a system at a finite temperature T, then we have Helmholtz free energies F_{xc} instead of E_{xc}). Thus, if x_1, x_2 are the fractional compositions n_i/n, then the exchange-correlation energy per particle E_{xc}/N is

$$E_{xc}/N = \int_0^1 \frac{d\lambda}{\lambda} \frac{n}{2} \int 2\pi r dr \frac{\lambda}{r} \sum_{ij} x_i x_j (g_{ij}(\lambda, r) - 1) \,. \quad (1)$$

This is known as the adiabatic connection formula [11], and is due to Pauli. Thus if the fully interacting PDF is known, the exchange-correlation energy can be obtained without the limitations of perturbation theory. The full PDF can be evaluated by (i) Quantum Monte Carlo methods (QMC), (ii) integral-equation methods developed from correlated wavefunctions using Feenberg-type analyses, (iii) classical-map hyper-netted chain (CHNC) methods, (iv) methods for specific applications as in Giorgi et al. [10] (v) SLTS methods for moderately coupled fluids.

The Exchange Energy

Note that the exchange energy alone is given by

$$E_x/N = \frac{n}{2} \int 2\pi r dr v(r) \sum_{ij} x_i x_j \left(g^0_{ij}(r) - 1 \right) \,. \quad (2)$$

That is, the Fermi hole of the non-interacting PDF determines the exchange energy. This was *defined* as a contribution taken to *first order*, Fig. 3b. Since one factor of $v(r)$ is already included in Eq. (2), we can only use $g^0_{ij}(r)$ as a better form would have go beyond first order in V_{ee}. The noninteracting PDF $g^0_{ij}(r)$ can be evaluated [18] directly from the noninteracting one-body wavefunctions (plane waves):

$$\begin{aligned} h_{ij} &= g_{ij} - 1 \\ h^0_{ij}(r) &= -\delta_{ij} F_i(r) F_j(r) \\ F_i(r) &= \frac{1}{n_i} \int \frac{d\vec{k}}{(2\pi)^2} f_i(k) e^{i\vec{k}\cdot\vec{r}} \,. \end{aligned}$$

Here $f_i(k)$ is the Fermi occupation number for the momentum state k. At $T = 0$, $F_i(r) = 2J_1(k_i r)/(k_i r)$; $k_i = \sqrt{4\pi n_i}$ where $J_1(x)$ is a Bessel function [1], and k_i is the Fermi momentum of the spin species i. At finite temperatures, the integrations have to be done numerically.

Formulae for the Kinetic and Exchange Energies.

At zero temperature, the Hartree–Fock energy can be written in terms of the spin polarization ζ and the density parameter r_s. The Hartree energy E_H is zero.

$$E_{HF} = E_0 + E_H + E_X \,; \quad E_H = 0 E_0/N = \frac{1+\zeta}{2r_s^2} \,;$$

$$E_x/N = -\frac{2\sqrt{2}}{3\pi r_s} \left[(1+\zeta)^{3/2} + (1-\zeta)^{3/2} \right]$$

The spin-dependent energies, e. g. $E_x(\zeta)$, may be written as:

$$\begin{aligned} E_x(\zeta) &= E_x(0) + (E_x(1) - E_x(0)) P(\zeta) \\ P(\zeta) &= \frac{\zeta_+^\alpha + \zeta_-^\alpha - 2}{2^\alpha - 2} \,. \end{aligned}$$

Here $\zeta_\pm = (1 \pm \zeta)$, and $P(\zeta)$ is called the "polarization function"; $\alpha = 3/2$ in Hartree–Fock theory.

The Correlation Energy

The calculation of the correlation energy E_c is the heart of the many-body problem, and is the main challenge. If E_c were known, the total ground state energy is $E_T = E_{HF} + E_c$. A widely used QMC evaluation was given in 1989 by Tanatar and Ceperley [23] using a variational Monte Carlo method (VMC) as well as a Greens Function Monte Carlo method. They gave parametrized forms for the unpolarized ($\zeta = 0$) and fully polarized ($\zeta = 1$) energies per electron, i. e., $\epsilon_c(r_s)$, in the form

$$\epsilon_c(r_s) = a_0 \frac{1 + a_1 x}{1 + a_1 x + a_2 x^2 + a_3 x^3}$$

where $x = \sqrt{r_s}$. The parameters have the values, $a_0 \ldots a_3 = -0.1784,\ 1.1300,\ 0.9052,\ 0.4165$ for $\zeta = 0$ and $-0.02575,\ 340.5813,\ 75.2293,\ 37.0170$ for $\zeta = 1$. No simulations were done for intermediate ζ, and hence it

was often *assumed* that the exchange-like polarization function $P(\zeta)$ could be used for the correlation energy as well.

$$\epsilon_c(\zeta) = \epsilon_c(0) + (\epsilon_c(1) - \epsilon_c(0))P(\zeta)$$

A more detailed study, given by Attaccalite et al. in 2002 is currently the best available QMC form of E_c. These authors [3] did simulations for finite-ζ as well and proposed a parameterized form.

$$\begin{aligned}
\epsilon_c(r_s,\zeta) &= (e^{-\beta r_s} - 1)\epsilon_x^6(r_s,\zeta) \\
&= +\epsilon_c(r_s,0) + \alpha_1(r_s)\zeta^2 + \alpha_2\zeta^4 \\
\alpha_1(r_s) &= (1/2\left[\partial^2\epsilon_c(r_s,\zeta)/\partial^2\zeta^2\right]_{\zeta=0} \\
\alpha_2(r_s) &= (1/24)[\partial^4\zeta^4]_{\zeta=0} \\
\epsilon_x^6(r_s,\zeta) &= \epsilon_x(r_s,\zeta) - \{1 + (3\zeta^2)(1+\zeta^2/16)\}
\end{aligned}$$

Here parameters α_1, α_2 are spin-stiffness constants. A short computer program for the calculation of ϵ_c is available [24].

CHNC Approaches

QMC procedures are computationally very demanding, and hence the development of other, simpler methods is of great interest. The accuracy of such methods can be ascertained by comparison with QMC results, and then such methods can be used in areas where QMC is prohibitive (e. g., at finite temperatures, many valley systems). In this respect, the classical-map HNC approach (CHNC) is worthy of note. In CHNC [21], the quantum fluid at $T = 0$ is replaced by a classical Coulomb fluid at a finite temperature T_q. The latter is chosen requiring that (i) the non-interacting classical fluid has the same $g_{ii}^0(r)$ as the quantum fluid. (ii) the correlation energy of the interacting classical fluid has the same correlation energy as the quantum fluid. The first requirement is easily met, and involves the construction of a classical potential which would give the same Fermi hole as the quantum fluid (see Fig. 2). The classical potential, $\beta V_{pau}(r)$ is called the "Pauli potential" and is a universal function of rk_F. This potential is zero for anti-parallel spins. This procedure fixes a 'Pauli potential' but not the temperature T_q, since only the product $\beta \times V_{pau}(r)$ is obtained. The second condition is used to determines T_q by evaluating $\epsilon_c(r_s)$ using the adiabatic connection formula, for a trial value of T_q., and a chosen ζ, say $\zeta = 1$. Then T_q is adjusted till the QMC value of $\epsilon_c(r_s, \zeta = 1)$ is obtained. Thus a table of T_q for each r_s, i. e., $T_q(r_s)$, is obtained. The classical PDFs needed here are easily calculated using a well established hyper-netted-chain technique. Only the ϵ_c value at $\zeta = 1$ is used as the input to the fit. The output is the full tabulation of the interacting $g_{ij}(r)$. It is found to be in *remarkably good agreement with the QMC generated* $g_{ij}(r)$. This is perhaps a consequence of DFT where it is asserted that the true density distribution is obtained if the true ground state energy (i.e, $E_0 + E_X + E_c$) is captured. It is also found that this CHNC procedure yields accurate $\epsilon_c(r_s)$ values at spin-polarizations different to the input value. The $T_q(r_s)$ function is found to be transferable to many Coulomb-fluid problems including hydrogen plasmas [6].

Fermi-Liquid Parameters

A knowledge of the correlation energy as a function of the r_s, ζ, and T enables an easy evaluation of Landau Fermi liquid parameters (LFLP). These are usually evaluated from complicated perturbation theory calculations whose domain of validity is, strictly speaking, restricted to small r_s. Three quantities are of great interest for the LFLP. These are, the inverse compressibility κ which is the density derivative of the chemical potential μ, the enhanced spin-susceptibility χ compared to the Pauli susceptibility χ_P, (incorporated in g^*, the effective Landé g factor), and the effective mass m^*. The latter involves the interacting and ideal specific heats C_v, C_v^0. AT finite T, the correlation energy is replaced by the correlation Free energy, $F_c(r_s, \zeta, T)$. This is the correction to the Helmholtz free energy beyond the Hartree–Fock approximation. If $F_c(r_s, \zeta, T)$ is known, a non-perturbative result becomes available for κ, m^* and g^* (for detailed, see [8]). Thus:

$$\begin{aligned}
m^* &= C_v/C_v^0 = 1 + \frac{\left[\partial^2 F_{xc}(t)/\partial t^2\right]}{\left[\partial^2 F_0(t)/\partial t^2\right]} \\
\chi_P/\chi_s &= (m^* g^*)^{-1} = 1 + \frac{\left[\partial^2 F_{xc}(\zeta)/\partial\zeta^2\right]}{\left[\partial^2 F_0(\zeta)/\partial\zeta^2\right]} \\
1/\kappa &= n^2\partial\mu/\partial n = n^2\partial^2[F_0 + F_x + F_c]/\partial n^2 .
\end{aligned}$$

Calculations of the ground-state energy as a function of r_s, ζ, as well as the spin-susceptibility enhancement using the QMC and the CHNC methods show that the 2D fluid (which is paramagnetic $\zeta = 0$ at high densities), becomes a ferromagnetic liquid (see [3,7]) for $r_s \simeq 26$. Thus the gain in kinetic energy due to spin-polarization is compensated by the exchange-correlation energy at sufficiently low densities. If simple perturbation methods are used for evaluating $E_{xc}(r_s)$, incorrect predictions (e. g., a magnetic transition at $r_s \sim 2$ in RPA) are obtained. The spin-susceptibility enhancement "blows up" towards infinity, as the transition is approached. These theoretical predictions have been confirmed experimentally (see [22]).

2D Layers with Finite Thickness

The assumption that the 2D layer is infinitely thin (i. e., "ideal") is often incorrect in practice. The lowest subband of a quantum well supporting a 2D layer may have an z-extension w of 10–20 nm, with a density distribution $n(z)$. If the next subband is inaccessible to the electrons, then they are 2D systems with a finite thickness. The HIGFETS used in many experiments contain a nearly triangular potential well (see Fig. 1) and the lowest subband wavefunction is approximated by the Fang-Howard(FH) function [2] $\phi_{\mathrm{fh}}(z)$. The difficulty of modeling exchange and correlation in such systems using either the ideal-2D parametrized E_{xc} forms, or the 3D forms (to allow for the z-extension) was noted by Martin et al. [14]. In fact, new exchange-correlational functionals have to be constructed for the effective-2D potential found in thick-2D layers [5]. We denote the Coulomb potential in an infinitely thin layer by $V(r) = 1/r$, while $W(r)$ is used for the effective 2-D potential of a thick layer. The potential $W(r)$ between two electrons having coordinates $(\vec{r}_1, z_1)$ and $(\vec{r}_2, z_2)$, with $\vec{r} = \vec{r}_1 - \vec{r}_2$ is given by,

$$W(r) = \int_0^{z_m}\int_0^{z_m} \frac{\mathrm{d}z_1 \mathrm{d}z_2 n(z_1) n(z_2)}{\left[r^2 + (z_1 - z_2)^2\right]^{1/2}} \,. \tag{3}$$

Here z_{m} is ∞ for FH, while $z_{\mathrm{m}} = w$ for a quantum well. The potential $W(r) = (1/r)F(r)$ and the form factor $F(r)$ reflects the effect of the z-extension of the density. It has been shown that any arbitrary density distribution $n(z)$ can be replaced by a electrostatically equivalent rectangular slab of width w by choosing w such that: in the 2D plane as $n(z)$.

$$n_{\mathrm{cd}} = 1/w = \int n(z)^2 \mathrm{d}z \,. \tag{4}$$

The quasi-2D potential for a constant-density slab of width w is given by

$$W(r) = V(r)F(s), \quad s = r/w, \quad t = \sqrt{(1+s^2)} \tag{5}$$

$$F(s) = 2s\left[\log\frac{1-t}{s} + 1 - t\right] \,. \tag{6}$$

This potential tends to $1/r$ for large r, and behaves as

$$\frac{2}{w}\left(\ln\frac{2w}{r} + \frac{r}{w} - 1\right)$$

for $r < w$. Thus the short-range behaviour is logarithmic and weaker than the Coulomb potential. The k-space form of the CDM potential is:

$$W(k, w) = V(k)F(p), \quad p = kw \tag{7}$$

$$F(p) = (2/p)\{(\mathrm{e}^{-p} - 1)/p + 1\} \,. \tag{8}$$

The form factors $F(s)$ and $F(p)$ tend to unity as $w \to 0$. The effective w in HIGFETS depends on r_{s}, i. e., in atomic units.

$$w = 16/(3b)b^3 = 33/\left(2r_{\mathrm{s}}^2\right) \,.$$

Using these potentials, layer-thickness dependent exchange-correlations functionals $E_{\mathrm{x}}c(r_{\mathrm{s}}, \zeta, w)$ can be evaluated using perturbation methods, or using the CHNC. The resulting xc-functionals have been given in [5], and displayed in Fig. 4. It is found that the exchange energy is reduced due to layer thickness which enters as w/r_{s}, thus

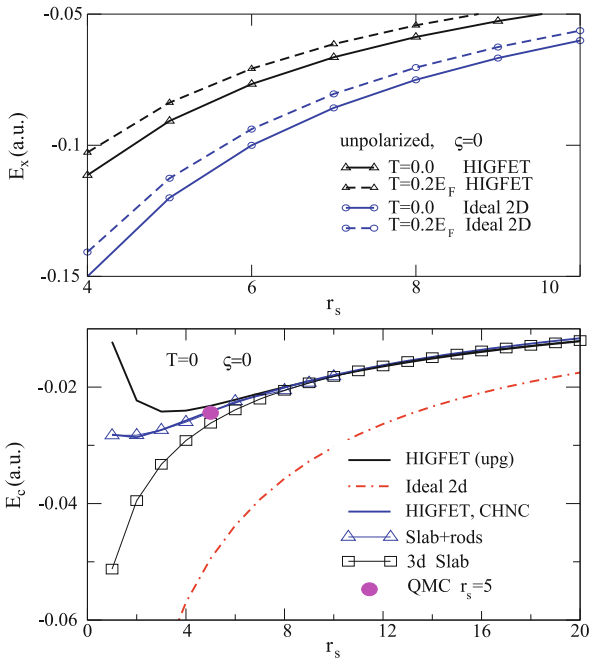

Spin Dependent Exchange and Correlation in Two-Dimensional Electron Layers, Figure 4

a The exchange energy E_x (Hartree a.u.) of a 2DES in a HIGFET compared to that of an ideal 2DES, at $T/E_{F=0}$ and 0.2, with $\zeta = 0$. *Solid line with circles*, ideal 2D, $T = 0$, *Solid line with triangles*, HIGFET at $T = 0$. Corresponding *broken lines* are for $T = 0.2E_F$. b The correlation energy E_c at $T = 0$ and $\zeta = 0$, for the HIGFET layer. HIGFET(upg), *black solid line*, is the "unperturbed-*g*" approximation. The *deep grey dashed line*, HIGFET(CHNC), is the full calculation. This is compared with the correlation energy of a 3D slab model (*line with squares*), and the "slab+rod"model (*triangles*). For details, see [5]. The QMC datum (*blacked hatched circle*) for a HIGFET, $r_s = 5$, is from [22]

affecting high densities and thick layers. A most interesting observation is that the transition to a ferromagnetic state which occurs at $r_s \simeq 26 - 27$ in ideal 2D, is pushed to higher r_s as the thickness w is increased.

Multi-Component 2D Layers

The two most important multi-component systems are 2D layers found in Si/SiO_2 interfaces, and in graphene monolayers.

Two-Valley System in Silicon-Silica Interfaces

Bulk Si has six equivalent valleys. However, when SiO_2 layers are grown on the (001) surface, with z the growth direction, the 2D layer occupies two degenerate layers in the (001) plane. This is a 4-component fluid, with two spin states and two valley states. Thus using an index $\nu = 1, 2, 3, 4$ where the first two are for spin, and the last two are for the valleys, there are 10 PDFS in the upper triangle of the matrix $g_{\nu\eta}(r)$. Only five of these are independent even for $\zeta \neq 0$. Four of these 10 are diagonal and contribute to both exchange and correlation, where as the off diagonals contribute *only* to the correlation energy E_c. This contrasts with the usual single valley system where there are two diagonals, and one off-diagonal PDF in the matrix of PDFs. Thus correlation effects dominate over exchange effects in multi-valley systems, and they show less tendency to spin-polarization effects which are driven by exchange effects. Thus there is no spontaneous spin-polarized state in the 2D-two valley system. However, the large increase in the correlation energy has been argued to be responsible for a rapid increase in the effective mass of electrons in Si MOSFETS as the density n decreases towards $\simeq 8 \times 10^{10}$ electrons/cm^2 [8,15]. However, this is found to be sensitive to slight changes in the CHNC parametrization.

Quantum Monte Carlo results of $E_{xc}(r_s)$ at $\zeta = 0, 1$ and CHNC results (at many values of ζ) are available, and are in excellent agreement with each other. If the total electron number density per unit area in the 2D layer is n in atomic units, then the overall $r_s = 1/\sqrt{\pi n}$, while the valley density, $n/2$ yields an $r_{sv} = \sqrt{2} r_s$. This r_{sv} and the corresponding k_{Fv} are used in constructing the non-interacting PDFs $g^0_{\nu\eta}$ from which the exchange energy can be calculate. Clearly, this is $2\epsilon_x(r_{sv})$, where ϵ_x is the usual one-valley formula for E_x/N .

The correlation energy cannot be completely mapped into results based on the 1-valley system. Thus E_c is determined via a coupling constant integration over the 10 PDFs, obtained from QMC or CHNC, using the adiabatic connection formula. Detailed results and numerical fits are available in [8].

Graphene: 2D Two-Valley System on a Honey-Comb Lattice

The electrons in the two-valley 2D system of Si/SiO_2 interfaces carry a valley index, but the single-particle wavefunctions are simple plane waves. In graphene, the two-component wavefunctions of massless fermions have a "chiral" character, since they carry phase angles ϕ_k, where $\vec{k}$ is the momentum vector in the 2D plane. The valence and conduction bands touch at the $\mathbf{K}, \mathbf{K}'$ points, and we need a band index $b = 1, -1$, to indicate the conduction and valence bands. Thus, we may have simultaneous electron and hole occupations, leading to an eight-component problem. Although the ϵ_x can be calculated explicitly, the calculation of the ϵ_c becomes very demanding. However, as the coupling constant g is independent of of the electron density, being $\simeq 2.7/\varepsilon$, the interactions are weak. Also, for pure electron doping or pure hole doping, the ϵ_c of the 4-component system in Si/SiO_2 at the same coupling strength as g (i.e, $r_s = g$) may be used as an estimate. Hence we focus on E_x and introduce a unit of energy $E_u = v_F k_c$ Here v_F is the Fermi velocity of the massless fermions, and k_c is a cutoff wavenumber chosen so that the number of states in the Brillouin zone is conserved, that is, if A_0 is the area per carbon atom, $k_c^2 = 4\pi(1/A_0)$. Then the exchange energy per Carbon atom can be written as:

$$E_x/E_u = -\frac{A_0 g^0/k_c}{(2\pi)^2}\frac{1}{4}\sum_{b_1,b_2,\sigma}\int_0^{2\pi} d\theta dk dp \times kp\frac{1 + b_1 b_2 \cos(\theta)}{|\mathbf{k} - \mathbf{p}|} n_{b_1,\sigma}(k) n_{b_2,\sigma}(p) \,. \quad (9)$$

In the above $g^0 = 2.7/\varepsilon$. A special feature, not found in the usual (non-chiral) 2D system is the occurrence of a term which depends on θ, the scattering angle. Because of this modulation of the interaction potential, the exchange energy in graphene is weaker than in a normal 2D layer with the same coupling strength. The above expressions for the exchange energy may be expressed [19] in terms of integrals over PDFs involving Bessel and Struve functions, or as elliptic elliptic integrals which depend on ζ and the fractional doping. Calculations of the total energy, (i.e, kinetic+E_{xc}) for graphene layers doped with electrons, holes or both, *do not* support a transition to a spin-polarized ground state. This is not surprising since the strength of the coulomb interaction is $\sim 2.7/\varepsilon$, i. e., weak, given that the spin transition in usual 2D systems occur for $r_s \sim 26$.

Exchange and Correlation in a Magnetic Field

Electrons respond to an external magnetic field by forming Landau Levels. If B is the magnetic field in the z-direction, then the energy spectrum is that of a harmonic oscillator, where the cyclotron frequency ω_c and the magnetic length ℓ are given by:

$$\omega_c = |e|B/(m_{bc}) , \quad \ell = \{\hbar c/(|e|B)\}^{1/2} . \tag{10}$$

Here m_b is the band mass and c is the velocity of light. The Harmonic oscillator energy levels with index η are:

$$\epsilon_\eta = (\eta + 1/2)\hbar\omega_c + \epsilon_{sb} \tag{11}$$

contains the lowest subband energy ϵ_{sb} of z-motion, and this may be taken as the zero of energy. The momentum states of the 2D system collapse into degenerate states, with each Landau level having $N_L = eBA/hc$ states, where A is the area of the system. If there are N electrons in A, the *filling factor* $\nu = N/N_L = n/(2\pi\ell^2)$. The integer part η of ν gives the number of *filled* Landau levels, while $\nu - \eta$ is the part of the highest landau level which is partially field. The filled Landau levels are like "inert shells" of atoms, while the partially filed part is polarizable and accounts for the more interesting aspects of exchange and correlation among electrons in magnetic fields. The total filling factor $\nu = \nu_1 + \nu_2$ where ν_i is for the ith spin species.

Since we are interested in ϵ_{xc}, we need the PDFs for use in Eq. (1). Thus it is convenient to work in the symmetric gauge where the eigenstates are functions of the radial distance r and the angle ϕ in the 2D plane, having radial and angular momentum quantum numbers n_r and m. Then, with $\hbar = 1$,

$$\epsilon_{n_r,m} = \{n_r + (|m| - m)/2 + 1/2\}\omega_c . \tag{12}$$

When the field is weak, there are a large number of Landau levels, and the usual 2D formulae for exchange and correlation can be used. This can be improved by using the summations over the density of states instead of integrations over k_x, k_y. A discussion may be found in, e. g., [4]. The high magnetic field limit, where only a few occupied Landau levels is of great importance. The integer and fractional QHE are found in this 'high-field' regime. If the number of full Landau levels of spin σ is ν_σ, the non-interacting PDF $g^0(r)$ is:

$$g^0(r) = 1 - \frac{e^{-x/2}}{\nu^2} \sum_\sigma \left[L^1_{\nu_\sigma - 1}(x^2/2)\right]^2 \tag{13}$$

where $x = r/\ell$, and $L_n{}^1$ is an associated Laguerre polynomial. The spin summation is unimportant in high fields, when the LLL is fully polarized by the field. ϵ_x can be evaluated from Eq. (2). The 2D Coulomb interaction may contain finite-layer thickness corrections. The effect of the magnetic field is totally contained in $e_{mag} = e^2/(\varepsilon\ell)$ and forms a natural energy unit for this system. The exchange energy ϵ_ν (in e_{mag} units), is $-\sqrt{\pi/8}$ at $\nu = 1$, and increases almost linearly to $2.35\epsilon_1$ at $\nu = 6$.

The correlation energy $\epsilon_c(r_s, \nu)$ of filled Landau levels can be evaluated using the adiabatic connection formula and the interacting PDFs. However, only an RPA evaluation is available, and $\epsilon_c(r_s, \nu)$ is very close to the zero. Hence the $B = 0$ value of ϵ_c is a good approximation.

The highest occupied Landau level is partially filled if $\nu_L = \nu - \eta$ is nonzero. The correlations in partially field states are clearly manifest in the high field limit when the lowest Landau level is fractionally occupied. Here ϵ_c has to be calculated from Quantum Monte Carlo simulations, or from methods based on the plasma analogy, via an ansatz for the many particle wavefunction. The case $\nu_L = 1/3$ and other "odd-integer" fractional-QHE states were elegantly explained by Laughlin [4], leading to a Nobel prize. This can be extended to other fractions within the plasma analogy to construct PDFs and obtain ϵ_c via the adiabatic connection formula [17]. Quantum simulations as well as composite-Fermion models have now largely superseded such methods. The literature on the quantum Hall effect [4] should be consulted for more details.

Exchange and Correlation at Finite Temperatures

The first-order (exchange) free energy F_x consists of F_x^i, where i denotes the two spin species. At $T = 0$ these reduce to the exchange energies:

$$E_i^x/n = \frac{8}{3\sqrt{\pi}} n_i^{1/2} . \tag{14}$$

Here $n_1 = n(1 + \zeta)/2$, and $n_2 = n(1 - \zeta)/2$. Then the exchange energy per particle at $T = 0$, i. e., the total E_x/n becomes

$$E_x/n = \left(E_1^x + E_2^x\right)/n = -\frac{8}{3\pi r_s}\left[x_1^{3/2} + x_2^{3/2}\right] \tag{15}$$

where x_1 and x_2 are the fractional compositions $(1 \pm \zeta)/2$ of the two spin species.

We define a reduced temperature $t = T/E_F$, $E_F = \pi n$, and the species-dependent reduced chemical potentials μ_i^0/T by η_i, reduced temperatures $t_1 = t/(1 + \zeta)$ and $t_2 = t/(1 - \zeta)$, based on the two Fermi energies E_{F1} and E_{F2} which are $E_F(1 \pm \zeta)$. Then we have:

$$F_i^x/E_i^x = \frac{3}{16} t_i^{3/2} \int_{-\infty}^{\eta_i} \frac{I^2_{-1/2}(u)\mathrm{d}u}{(\eta_i - u)^{1/2}} . \tag{16}$$

The $I_{-1/2}$ is the Fermi integral defined as usual:

$$I_\nu(z) = \int_0^\infty \frac{\mathrm{d}x x^\nu}{1 + \mathrm{e}^{x-z}} . \quad (17)$$

The η_i are given by

$$\eta_i = \log(\mathrm{e}^{1/t_i} - 1) . \quad (18)$$

In the paramagnetic case Eq. (16) reduces to the result given by Isihara et al. [13]. For small values of t, the exchange energy is of the form,

$$E_\mathrm{x}(r_\mathrm{s}, t) = E_\mathrm{x}(r_\mathrm{s}, 0)[1 + (\pi^2/16)t^2 \log(t) - 0.56736t^2 + \ldots] . \quad (19)$$

The total exchange free energy is $F_\mathrm{x} = \Sigma F_i^x$.

A real-space formulation of $F_\mathrm{x} = F_1^x + F_2^x$ using the zeroth-order PDFs fits naturally with the approach of our study. Thus

$$F_\mathrm{x}/n = n \int \frac{2\pi r \mathrm{d}r}{r} \sum_{i<j} h_{ij}^0(r) . \quad (20)$$

Here $h_{ij}^0(r) = g_{ij}^0(r) - 1$. In the non-interacting system at temperature T, the antiparallel h_{12}^0, viz., $g_{12}^0(r, T) - 1$, is zero while

$$h_{11}^0(\mathbf{r}) = -\frac{1}{n_i^2} \Sigma_{\mathbf{k}_1,\mathbf{k}_2} n(k_1) n(k_2) \mathrm{e}^{i(\mathbf{k}_1 - \mathbf{k}_2)\cdot\mathbf{r}} = -[f(r)]^2 .$$

Here k, r are 2D vectors and $n(k)$ is the Fermi occupation number at the temperature T. At $T = 0$ $f(r) = 2J_1(k_i r)/k_r$ where $J_1(x)$ is a Bessel function.

The exchange free energy is a universal function $F_\mathrm{x}(t)/E_\mathrm{x}$, for arbitrary ζ. That is, the same function applies to any component, on using the reduced Fermi temperature of the spin species. The total F_x is the sum of both spin terms. A parametrized fit is:

$$F_i^x(t, \zeta)/E_i^x(\zeta) = \frac{1 + C_1 t_i + C_2 t_i^2}{1 + C_3 t_i + C_4 t_i^2} \tanh(1/\sqrt{t_i}) . \quad (21)$$

The fit coefficients C_i are 3.27603, 4.81484, 3.33100, 6.51436. The temperature t_i is $t/(1 \pm \zeta)$, appropriate to the spin polarization. The exchange effects in the 2DES decay more slowly with temperature than in the 3D case where a $\tanh(1/t)$ factor appears in Eq. 3.2 of [20]. The above form does not explicitly contain the low-temperature logarithmic term [12], but it reproduces the value of 0.99382 at $t = 0.05$, while the numerical integration gives 0.9939497. Similarly, at $t = 1$, 10 and 30 the fit (integral) returns 0.63839 (0.63839), 0.22999 (0.22990), and 0.13421 (0.13410) respectively.

If the 2D layer has a finite thickness, E_x is reduced. For numerical procedures and parametrized forms, the reader is referred to [5]. The correlation energy $\epsilon_\mathrm{c}(r_\mathrm{s}, \zeta, T)$ can be easily calculated using the PDFs from CHNC [7], but they have not been presented in a parametrized form.

Conclusion

An understanding of electron-exchange and correlation, i. e., many-body effects, is crucial to the design of quantum-well lasers, field-effect transistors, spintronics and other areas of nanostructure technology. We currently have excellent results for uniform electron systems at zero magnetic fields. The case of moderate magnetic fields is poorly understood. The very high-field case has progressed with our understanding of the fractional and integer quantum Hall effects.

Bibliography

1. Abramovitz M, Stegun IA (1965) Handbook of mathematical functions. Dover, New York
2. Ando T, Fowler B, Stern F (1982) Rev Mod Phys 54:437
3. Attaccalite C, Moroni S, Gori-Giorgi P, Bachelet GB (2002) Phys Rev Lett 88:256601; (2003) Erratum 91:109902
4. Chakraborty T (1996) The fractional quantum hall effect. Springer, New York
5. Dharma-wardana MWC (2005) Phys Rev B 72:125339
6. Dharma-wardana MWC, Perrot F (2002) Phys Rev B 66:14110
7. Dharma-wardana MWC, Perrot F (2003) Phys Rev Lett 90:136601
8. Dharma-wardana MWC, Perrot F (2004) Phys Rev B 70:035308
9. Geim AK, Novoselov KS (2007) Nat Mater 6:183–191
10. Gori-Giorgi P, Perdew JP (2001) Phys Rev B 64:155102
11. Harris J, Jones RO (1974) J Phys F 4:1170
12. Hong S, Mahan GD (1995) Phys Rev B 52:7860
13. Isihara A, Toyoda T (1980) Phys Rev B 21:p3358
14. Kim Y-H, Lee I-H, Nagaraja S, Leburton J-P, Hood RQ, Martin RM (2000) Phys Rev B 61:5202
15. Kravchenko SV, Sarachik MP (2004) Rep Prog Phys 67:1
16. Liu HC, Song CY, Wasilewski ZR, Springthorpe AJ, Cao JC, Dharma-wardana C, Aers GC, Lockwood DJ, Gupta JA (2003) Phys Rev Lett 90:077402
17. MacDonald AH, Aers GC, Dharma-wardana MWC (1985) Phys Rev B 31:5529
18. Mahan GD (1981) Many-particle physics. Plenum, New York
19. Peres NMR, Guinea F, Castro Neto AH (2007) Phys Rev 72:174406; Dharma-wardana MWC (2007) Ibid 75:075427
20. Perrot F, Dharma-wardana MWC (1984) Phys Rev A 30:2619
21. Perrot F, Dharma-wardana MWC (2001) Phys Rev Lett 87:206404
22. Tan YW, Zhu J, Stormer HL, Pfeiffer LN, Baldwin KN, West KW (2005) cond-matt/0412260, Phys Rev Lett 94:16405; De Palo S,

Botti M, Moroni S, Senatore G (2005) Phys Rev Lett 94:226405; Vakili K, Shkolnikov YP, Tutuc E, DePoortere EP, Shayegan M (2004) Phys Rev Lett 92:226401
23. Tanatar B, Ceperley DM (1989) Phys Rev B 39:5005
24. chandre.dharma-wardana@nrc-cnrc.gc.ca, http://babylon.phy.nrc.ca/ims/qp/chandre/

Spin Dynamics in Disordered Solids

Fridrikh S. Dzheparov
Institute for Theoretical and Experimental Physics, Moscow, Russia

Article Outline

Glossary
Definition of the Subject
Introduction
Delocalization of Nuclear Polarization in a Disordered Spin System
Nuclear Relaxation via Paramagnetic Impurities
Resonance Line Form Function for Magnetically Diluted Solids
Saturation on the Wing of a Dipole-Broadened EPR Line and Cluster Expansions
Future Developments
Bibliography

Glossary

Averaging Two kinds of averaging are necessary for any observable $\hat{Q}$: a standard quantum mechanical one $Q = \langle \hat{Q} \rangle = \mathrm{Tr}(\hat{Q}\rho)$, where ρ is the density matrix, and consequent averaging over all possible positions of particles in disordered media $\langle Q \rangle_c = \langle \langle \hat{Q} \rangle \rangle_c$.

Continuum media approximationj Positions of particles forming the disordered solids can be considered as a subset (impurity sites) of crystal lattice sites, randomly distributed on the lattice with a small probability $c \ll 1$ to find a given site occupied. Many important results can be received in continuum media approximation (CMA) when prime cell volume $\Omega_c \to 0$ together with impurity concentration $c \to 0$ at a fixed value of impurity density $n = c/\Omega_c$.

Disordered solids Statically disordered media are considered, this means that constituent particles (atoms, ions and so on) do not participate in significant translational motions during the relaxation time under discussion, and therefore their positions are fixed (frozen) in the main approximation.

High-temperature approximation High-temperature approximation (HTA) in spin dynamics consists of using the simplest density matrix $\rho = 1/\mathrm{Tr}\,1$, corresponding to the limit of infinite temperature T of canonical Gibbs distribution $\rho_G = \exp(-H/T)/\mathrm{Tr}\exp(-H/T)$; here H is the Hamiltonian. Spin dynamics remains nontrivial and rich in this limit. Sometimes HTA means application of $\rho = (1 - H/T)/\mathrm{Tr}\,1$.

Local field See "Secular part of dipole–dipole interactions".

Local frequency The frequency $\omega_{\mathrm{loc}} = \gamma H_{\mathrm{loc}}$ of rotation of the spin in local field H_{loc}.

Secular part of dipole–dipole interactions As a rule any spin is considered as subjected to an external static magnetic field $\mathbf{H}_0 = H_0\mathbf{n}_z$ (directed along the z-axis) and local field, produced by dipole magnetic moments of surrounding spins. The Hamiltonian of the dipole–dipole interaction of spins $\mathbf{I}_i$ and $\mathbf{I}_j$ is of the form $H_d^{(0)}(i,j) = r_{ij}^{-3}((\mathbf{m}_i\mathbf{m}_j) - 3(\mathbf{n}_{ij}\mathbf{m}_i)(3\mathbf{n}_{ij}\mathbf{m}_j))$, where $\mathbf{r}_{ij} = \mathbf{n}_{ij}r_{ij} = \mathbf{r}_i - \mathbf{r}_j$, and $\mathbf{r}_i$ is the position of spin "i" having magnetic moment $\mathbf{m}_i = \hbar\gamma_i I_i$. If $H_0 \gg H_{\mathrm{loc}}$, where H_{loc} is the mean square field produced at any spin by surrounding spins (local field), then the Hamiltonian $H_d^{(0)}(i,j)$ can be substituted by the so-called secular dipole–dipole Hamiltonian $H_d(i,j) = \frac{1}{2r_{ij}^3}(1-3(\mathbf{n}_{ij}\mathbf{n}_z)^2)(\mathbf{m}_i\mathbf{m}_j - 3(\mathbf{m}_i\mathbf{n}_z)(\mathbf{m}_j\mathbf{n}_z))$. The accuracy of the substitution is not less than $\sim H_{\mathrm{loc}}/H_0$.

Spin dynamics Spin dynamics is considered as a time evolution of correlation functions directly connected with measurable quantities of paramagnetic samples.

Units As a rule $\hbar = 1$ is supposed. Part of the equations contains $\hbar$ written explicitly.

Definition of the Subject

The main specifics of the theory of spin dynamics in disordered solids result from the fact that calculation of observable values must start from the solution of the equation of motion, and then they should be averaged over random distribution of spins in the sample. Nominally any problem in statistical physics looks analogous, but the content of existing text books contains much more simple bypass methods to achieve the results. An important bypass class, determined in the preceding century, consists of problems that are equivalent to the motion of weak (or seldom) interacting (quasi)particles in translational invariant media. The basis of this solution is formed by the Boltzmann equation (devised in the 19th century) and by methods of deriving hydrodynamic equations based on this foundation [1]. Other important advancements in physical kinet-

ics are connected with the invention of a projection technique by Nakajima–Zwanzig for deriving various master equations and with the development of effective approximations for corresponding memory functions [1,2]. Both these bypass methods produce satisfactory predictions if the system has such small parameters as, for example, the ratio of collision duration to time between collisions, or ratio of memory time to time of variation of substantial observables (the observables whereon the system evolution is projected). Similar small parameters do not exist in disordered solids after a beginning stage, or they are totally absent. Therefore, other more refined methods are necessary to predict experimental results. They are partially presented below.

Introduction

The first correct solutions of the problems of disordered media kinetics, which are directly connected with the spin dynamics, were obtained by Forster [3] and Anderson [4].

Forster's problem (in application to spin kinetics) describes the relaxation of impurity nuclei via paramagnetic impurities in the absence of nuclear spin diffusion. The evolution of the polarization $p_j(t) = \langle I_j^z(t)\rangle$ of the jth nucleus, placed at site $\mathbf{r}_j$, and having the spin I, is described by the kinetic equation

$$\dot{p}_j = -\sum\nolimits_{a=1}^{N} \nu_{aj} p_j = -\sum\nolimits_{\mathbf{r}} n_{\mathbf{r}} \nu_{\mathbf{r}j} p_j, \quad p_j(t=0) = 1. \tag{1}$$

Here ν_{aj} is the depolarization rate under influence of the ath paramagnetic center (acceptor), N is the total number of acceptors in the sample, $\nu_{\mathbf{r}j} = \nu_{aj}(\mathbf{r}_a = \mathbf{r})$, and $n_{\mathbf{r}}$ is the occupation number of the site $\mathbf{r}$ by an acceptor ($n_{\mathbf{r}} = 1\,(0)$ if the site $\mathbf{r}$ is (not) occupied by an acceptor). Occupations of different sites will be assumed to be independent and having no dependence on $\mathbf{r}$ (with a small probability of occupation $c \ll 1$ as a rule):

$$\langle n_{\mathbf{r}}\rangle_c = c\,, \quad \langle n_{\mathbf{r}} n_{\mathbf{x}}\rangle_c = c\delta_{\mathbf{rx}} + c^2(1-\delta_{\mathbf{rx}})\,,$$
$$\left\langle \prod\nolimits'^{m}_{j=1} n_{\mathbf{r}_j} \right\rangle_c = \prod\nolimits'^{m}_{j=1} \langle n_{\mathbf{r}_j}\rangle_c = c^m\,. \tag{2}$$

All $\mathbf{r}_j$ are different in the last relation. Coincidence in indexes can be treated using the identity $n_{\mathbf{r}}^2 = n_{\mathbf{r}}$. The problem consists of calculation of the observable polarization, averaged over all possible positions of acceptors in the sample. Ensemble averaging can be used for macroscopic samples: $p(t) = \langle p_j(t)\rangle_c$. Occupation number representation gives the simplest and general solution of the last problem [5]. Indeed

$$\begin{aligned} p(t) = \langle p_j(t)\rangle_c &= \left\langle \exp\left(-\sum\nolimits_{\mathbf{r}} n_{\mathbf{r}} \nu_{\mathbf{r}j} t\right)\right\rangle_c \\ &= \prod\nolimits_{\mathbf{r}} \langle \exp\left(-n_{\mathbf{r}} \nu_{\mathbf{r}j} t\right)\rangle_c \\ &= \prod\nolimits_{\mathbf{r}} \langle 1 + n_{\mathbf{r}}\left[\exp\left(-\nu_{\mathbf{r}j} t\right) - 1\right]\rangle_c \\ &= \prod\nolimits_{\mathbf{r}} \langle 1 + c\left[\exp\left(-\nu_{\mathbf{r}j} t\right) - 1\right]\rangle_c \\ &= \exp\left\{\sum\nolimits_{\mathbf{r}} \ln\left[1 + c\left(\exp\left(-\nu_{\mathbf{r}j} t\right) - 1\right)\right]\right\}\,. \end{aligned} \tag{3}$$

The identity $f(n_{\mathbf{r}}) = f(0) + n_{\mathbf{r}}(f(1) - f(0))$ is applied here. It is valid for any realistic function $f(x)$. The relation (3) is exact for any c.

The typical depolarization rate is of the form $\nu_{aj} = \nu_0 r_0^6 \chi(\mathbf{n}_{aj})/r_{aj}^6$, where ν_0 scales the transport at minimal distance r_0, $\mathbf{n}_{aj} = \mathbf{r}_{aj}/r_{aj}$, $\mathbf{r}_{aj} = \mathbf{r}_a - \mathbf{r}_j$, $\chi(\mathbf{n}) = |Y_{21}(\mathbf{n})|^2$, and $Y_{lm}(\mathbf{n})$ is the spherical harmonics. Additional simplification is possible for the continuum media approximation (CMA), when $c \to 0$, but $p(0) - p(t) \neq 0$:

$$\begin{aligned} p(t, c \to 0) &= \exp\left(n \int \mathrm{d}^d r \left(\mathrm{e}^{-\nu_{\mathbf{r}j} t} - 1\right)\right) \\ &= \exp\left(-(\beta_F t)^{d/6}\right)\,, \quad \beta_F \propto n^{6/d} \nu_0 r_0^6\,. \end{aligned} \tag{4}$$

Here arbitrary spatial dimension d is considered, and impurity density $n = c/\Omega_{\mathrm{c}}$ is introduced together with the prime cell volume Ω_c.

The most important properties of Forster's result (4) consist of replacement of the simple exponential kinetics by a slower law, and in proportionality of β_F to transfer rate at average distance $\bar{r} = n^{-1/d}$: $\beta_F = \mathrm{const} \cdot \nu_{\mathbf{r}j}(|\mathbf{r} - \mathbf{r}_j| = \bar{r})$, where const doesn't depend on concentration.

Anderson introduced the model to calculate the EPR line form function $g(\omega) = \int_{-\infty}^{\infty} \frac{\mathrm{d}t}{2\pi} \mathrm{e}^{-i\omega t} F(t)$ and free induction decay $F(t) = \langle\langle S_+(t) S_-\rangle\rangle_c / \langle\langle S_+ S_-\rangle\rangle_c$ for the system of equivalent spins $S_j = 1/2$, randomly distributed inside a sample and having dipole–dipole interaction. The model consists of substitution of exact secular interaction

$$\begin{aligned} H &= \frac{1}{2} \sum_{ij} b_{ij} \left(S_i^z S_j^z - \frac{1}{3} \mathbf{S}_i \mathbf{S}_j\right) \\ &= \frac{1}{2} \sum_{\mathbf{r},\mathbf{q}} n_{\mathbf{q}} n_{\mathbf{r}} b_{\mathbf{rq}} \left(S_{\mathbf{r}}^z S_{\mathbf{q}}^z - \frac{1}{3} \mathbf{S}_{\mathbf{r}} \mathbf{S}_{\mathbf{q}}\right) \end{aligned} \tag{5}$$

by a simpler one, depending on spin z-components only:

$$H_A = \frac{1}{2} \sum_{\mathbf{r},\mathbf{q}} n_{\mathbf{q}} n_{\mathbf{r}} b_{\mathbf{rq}} S_{\mathbf{r}}^z S_{\mathbf{q}}^z\,, \tag{6}$$

where $S_{\mathbf{r}}^{\alpha}$ is the α component of spin $\mathbf{S}_{\mathbf{r}}$ placed at site $\mathbf{r}$ and $b_{\mathbf{rq}} = b_0 r_0^3 (1 - 3\cos^2 \vartheta_{\mathbf{rq}})/|\mathbf{r} - \mathbf{q}|^3$ is the standard

dipole–dipole coefficient [2] (here and below $b_{\mathbf{rr}} = 0$). It is of great importance, that free induction signals $F_2(t) = \langle S_+(t)S_-\rangle_0/\langle S_+S_-\rangle_0$ are the same for both Hamiltonians for two spin problem, if $S_j = 1/2$. Anderson's model has an exact solution for arbitrary concentration [6]. Indeed, the equation of motion for orthogonal spin components has the solution

$$n_{\mathbf{r}}S_{\mathbf{r}}^+(t) = n_r \exp(iH_A t)S_{\mathbf{r}}^+ \exp(-iH_A t) = n_{\mathbf{r}}S_{\mathbf{r}}^+ \exp\left(i\sum\nolimits_{\mathbf{q}} n_{\mathbf{q}} b_{\mathbf{rq}} S_{\mathbf{q}}^z t\right) .$$

Introducing total moment of the sample $\mathbf{S} = \sum_{\mathbf{r}} n_{\mathbf{r}}\mathbf{S}_{\mathbf{r}}$, and using standard high-temperature approximation for free induction decay (FID) we have

$$\begin{aligned} F(t) &= \frac{\langle\langle S_+(t)S_-\rangle\rangle_c}{\langle\langle S_+S_-\rangle\rangle_c} = \left\langle\left\langle \exp\left(i\sum\nolimits_{\mathbf{q}} n_{\mathbf{q}} b_{\mathbf{rq}} S_{\mathbf{q}}^z t\right)\right\rangle\right\rangle_c \\ &= \prod\nolimits_{\mathbf{q}} \left\langle\cos(n_{\mathbf{q}} b_{\mathbf{qr}} t/2)\right\rangle_c \\ &= \prod\nolimits_{\mathbf{q}} \left(1 + c\left(\cos\left(b_{\mathbf{qr}} t/2\right) - 1\right)\right) \\ &= \exp\left(\sum\nolimits_{\mathbf{q}} \ln\left(1 + c\left(\cos(b_{\mathbf{qr}} t/2) - 1\right)\right)\right) . \end{aligned} \tag{7}$$

Applying CMA to (7) we arrive at Anderson's result:

$$\begin{aligned} F(t, c \to 0) &= \exp\left(-n\int \mathrm{d}^d q\left(1 - \cos\left(b_{\mathbf{qr}} t/2\right)\right)\right) \\ &= \exp\left(-(D_A t)^{d/3}\right) . \end{aligned} \tag{8}$$

For three-dimensional systems the model evidently has simple Lorentz form function:

$$g(\omega) = g_A(\omega) = \frac{D_A}{\pi\left(\omega^2 + D_A^2\right)} .$$

Anderson's parameter D_A, as well as β_F, is proportional to the rate of the process taken at the average distance: $D_A \propto b_{\mathbf{qr}}(|\mathbf{q} - \mathbf{r}| = \bar{r})$. Other derivations of Eq. (8) (closer to the original Anderson's treatment fulfilled for $d = 3$) can be found in [7].

Delocalization of Nuclear Polarization in a Disordered Spin System

One of the most important generalizations of the Forster process is presented by delocalization of nuclear polarization in the system of impurity nuclei, randomly distributed in a diamagnetic matrix, when host nuclei have faster phase relaxation and flip-flop transitions then impurities. Examples of such systems: nuclei ^{6}Li in the single crystal ^{7}Li^{19}F or spins ^{107}Ag in the single crystal ^{109}Ag^{19}F. The former system is of special interest, because it is accessible for direct experimental study due to unique coincidence of g-factors of stable nuclei ^{6}Li and β-active nuclei ^{8}Li (β-nuclei) [8]. The system has simple and instructive ergodic properties as well [9]. Therefore, we will consider below a specific system (consisting of ^{6}Li nuclei in a LiF single crystal with addition of one β-nucleus ^{8}Li) where nuclear polarization transfers from initially polarized ^{8}Li nucleus to the nearest nonpolarized ^{6}Li nuclei and then migrates over other ^{6}Li nuclei and might return back to the ^{8}Li.

It was shown in [10] that modern master equation treatment leads to a description of the system by following kinetic equations, which have been formulated in [5]:

$$\frac{\partial p_{i0}}{\partial t} = -\sum\nolimits_j \left(\nu_{ji}p_{i0} - \nu_{ij}p_{j0}\right) , \quad p_{i0}(t = 0) = \delta_{i0} . \tag{9}$$

Here $p_{i0} = \langle I_i^z\rangle$ is the quantum statistical average value of the z-component (polarization) of the ith nucleus of the system ^{8}Li-^{6}Li, placed at $\mathbf{r}_i$ ($i = 0$ corresponds to ^{8}Li with a spin $I_0 = I = 2$, and $i \neq 0$ to ^{6}Li with a spin $I_i = S = 1$). The rates of polarization transfer are of the form:

$$\begin{aligned} \nu_{ji} &= \xi_j \nu_{ji}^0 r_0^6 \cdot \left(\frac{1 - 3\cos^2\theta_{ji}}{r_{ji}^3}\right)^2 , \\ \nu_{ji}^0 &= \frac{\pi}{6} S(S+1)\left(\frac{g_i g_j \beta_n^2}{\hbar r_0^3}\right)^2 g_{ij}(\omega_{ij}) . \end{aligned} \tag{10}$$

Here $\xi_j = I_j(I_j + 1)/[S(S + 1)]$, $g_0 = g_I = 0.8267$ and $g_{i\neq 0} = g_S = 0.8220$ are g-factors of ^{8}Li and ^{6}Li correspondingly, β_n is nuclear magneton, θ_{ji} is the angle between external static field H_0 and $\mathbf{r}_{ji} = \mathbf{r}_j - \mathbf{r}_i$, $r_0 = 2.01 \cdot \sqrt{2}$ Å is the minimal distance between Li nuclei, and ω_{ij} is the difference of the Larmor frequencies. The cross-relaxation form function $g_{ij}(\omega)$ was studied in [11], and it can be taken as Gaussian $g_{ij}(\omega) \approx \exp(-\omega^2/(2M))/(2\pi M)^{1/2}$ in the main approximation. Here $M = 2M_2$, where M_2 is the second moment of the ^{8}Li NMR line. As a result ν_{ij}^0 have two values only, $\nu_{ij}^0 = \nu_0$ for transfer between ^{6}Li spins (at that $\omega_{ij} = 0$), and $\nu_{i0}^0 = \nu_1$ for transfer between ^{8}Li and ^{6}Li (with $\omega_{ij} = \Delta$).

The occupation number representation of the Eq. (9) is of the form [5]

$$\frac{\partial \tilde{P}_{\mathbf{x0}}}{\partial t} = -\sum\nolimits_{\mathbf{z}} \left(n_{\mathbf{z}} \nu_{\mathbf{zx}} \tilde{P}_{\mathbf{x0}} - n_{\mathbf{x}} \nu_{\mathbf{xz}} \tilde{P}_{\mathbf{z0}}\right) , \quad \tilde{P}_{\mathbf{x0}}(t=0) = n_x \delta_{\mathbf{x0}}/c , \quad (11)$$

where propagator $\tilde{P}_{\mathbf{x0}}$ gives polarization of the lattice site $\mathbf{x}$ when initially the site $\mathbf{0}$ was polarized, and $\nu_{\mathbf{zx}} = \nu_{ij}(\mathbf{r}_i = \mathbf{z}, \mathbf{r}_j = \mathbf{x})$. Equation (9) is a direct consequence of Eq. (11), that is evident, if we omit all empty sites for which $n_{\mathbf{x}} = 0$ and, consequently, $\tilde{P}_{\mathbf{x0}} = 0$.

The experimentally observable value is polarization of the β-nucleus, averaged over a random distribution of ^{8}Li-^{6}Li nuclei $P_{\mathbf{00}}(t) = \langle p_{00}\rangle_c = \langle \tilde{P}_{\mathbf{00}}\rangle_c$. A calculation of such a value belongs to the problems of random walks in disordered media (RWDM), which is one of the most complex modern fields of statistical physics, which defines the actuality of this kind of study. To clarify the status of the problem we note that its solution in CMA can be connected [12] with the calculation of the path integral of the form:

$$\mathcal{P}_{\mathbf{xy}}(t) = \int_{\mathbf{q}(0)=\mathbf{x}}^{\mathbf{q}(t)=\mathbf{y}} D\mathbf{p}(\tau) D\mathbf{q}(\tau) \exp\left[i \int_{\mathbf{x}}^{\mathbf{y}} \mathbf{p}\mathrm{d}\mathbf{q} + L\left[\mathbf{q}, \mathbf{p}, t\right]\right] \quad (12)$$

$$L\left[\mathbf{q}, \mathbf{p}, t\right] = n \int \mathrm{d}^3 z \left(\mathrm{e}^{-\int_0^t \mathrm{d}\tau A^z(\mathbf{q}(\tau), \mathbf{p}(\tau))} - 1\right) , \quad (13)$$

where $A^z(\mathbf{q}, \mathbf{p}) = \nu_{\mathbf{zq}} \left(1 - \mathrm{e}^{-i\mathbf{p}(\mathbf{z}-\mathbf{q})}\right)$. The representation (12) is similar to, but more complex than path integrals in the famous polaron problems (PP) [12]. The main difference consists of multi-time action L in (13) instead of two-time action in PP, in strong singularity of $\nu_{\mathbf{zq}}$ instead of the less singular kernel $1/|\mathbf{z}-\mathbf{q}|$ in PP, in additional path integral over all $\mathbf{p}(\tau)$ and in strong dependence of the infinite-fold integral (12) on the exact form of approximations by the integrals with finite multiplicity.

Field and superfield path integral representations for $P_{\mathbf{xy}}(t) = \langle \tilde{P}_{\mathbf{xy}}\rangle_c$ exist as well [13]. They demonstrate the relation of the RWDM with general problems of the modern field theory. It should be noted, that Eq. (9) and corresponding propagator $P_{\mathbf{xy}}(t)$ have wide applications in many fields of physics. For example they describe (after minimal corrections, but retaining the dipole long-range action) incoherent spatial transport of the localized electronic excitations (which is of importance in optics and biophysics) [14], and with $\ln(\nu_{ij}) \propto r_{ij}$ the same equations are used in the theory of the hopping conductivity [15], where their applications are combined with percolation theory.

The prognosis of the measurements with dipole transport has been based on the relation [10]

$$P_{\mathbf{00}}(t) = F(t) = \exp\left(-\sqrt{\beta_1 t}\right) + \xi \frac{1 - \exp\left(-\sqrt{\beta_1 t}\right)}{(\mu\beta(t+\tau))^{3/2}} \cdot \left(1 + \frac{\varphi}{\sqrt{\mu\beta(t+\tau)}}\right) , \quad (14)$$

where $\xi = \xi_0 = I(I+1)/[S(S+1)] = 3$, Forster parameters $\beta = \frac{256}{243}\frac{r_0^6}{\Omega_c^2}\pi^3 c^2 \nu_0$ and $\beta_1 = \beta \cdot \nu_1/\nu_0$, and the limit of small c is assumed (CMA). The values $\varphi = 2.09$ and $\mu\beta\tau = 5.11$ were chosen here to construct Eq. (14) as an interpolating formula between exact results [16] of the expansion of $P_{\mathbf{00}}$ in terms of c^m (which is an expansion in powers of $(\beta t)^{m/2}$ in reality) and expansion in terms of $1/(\beta t)^{m/2}$, produced by the approximative or numerical treatment of the long time asymptotics (where dipole long ranging induces exact dependence between the first and second terms [10]). The value of μ is connected with the main values D_α of the diffusion tensor: $\mu\beta = 4\pi(c/\Omega_c)^{2/3}(\prod_{\alpha=1}^3 D_\alpha)^{1/3}$. It should be stressed that an analytical solution for the nature of the long time asymptotics is absent up to now, but studies of realistic models [17] and modern numerical-analytical studies [13,18,19] indicate that it is of diffusion type. The diffusion tensor is calculated now with 1% uncertainty [18,19] giving $\mu = 0.71$. As a result the relation (14) holds to within $(\beta t)^{1/2}$ at small βt, and it holds to within $(\beta t)^{-2}$ at large βt.

The applicability of the Eq. (14) (having no fitting parameters) was checked at $\beta_1 t \leq 10$ in [8] and at $\beta_1 t \leq 15$ in [20]. The last experimental results [21,22] of the ITEP group indicate that at $\beta_1 t \sim 25$ some correction is necessary. It can be introduced in simplest form as

$$P_{\mathbf{00}}(t) = F(t)G(t) \qquad G(t) = G_{\exp} = \left(1 - \frac{\left(\frac{1}{8} + \alpha\right)\beta_1 t - u(\beta_1 t)^2}{(1 + \nu\beta t)^3}\right) . \quad (15)$$

Here $F(t)$ is defined in (14), $\alpha = \alpha(\Delta)$ is tabulated in [10] (at that $\alpha(\Delta \to 0) = 0.013$), and fitting parameters u and ν should be determined by experimental data. The relation (15) holds to within βt at small βt, and it holds to within $(\beta t)^{-2}$ at large βt. The fitting produces $u \approx 0.06$ and $\nu \approx 0.12$ [21]. Direct numerical simulation of the $P_{\mathbf{00}}(t)$ fulfilled in [19] for a small external magnetic field corresponding to $\beta \approx \beta_1$ justified this correction and gave more precise form and values for $F(t)$ indicating that

$0.9 \le F(t) \le 2$. New experimental results [23] have been obtained for larger magnetic fields, when $\beta - \beta_1 \propto \beta_1$. More detailed numerical simulation is expected for description of the process in this region of fields. This should take into account correlations of local fields on impurity spins [11].

The numerical simulation [13,18,19] is based on the substitution of infinite disordered media by a crystal with a large disordered elementary cell, containing $N_d \sim 1000$ impurity spins, and on checking of stability of the results with regard to variation of N_d. It starts from Eq. (9) written as

$$\frac{\partial p_{i0}}{\partial t} = -\sum_j B_{ij} p_{j0} \,, \quad B_{ij} = \delta_{ij} \sum_k \nu_{ki} - \nu_{ij} \,,$$
$$p_{i0}(t=0) = \delta_{i0} \,, \tag{16}$$

and N_d impurities randomly placed in sites of a supercell, having $N = N_d/c = N_g^3$ lattice sites, and edges $\mathbf{R}_\alpha = N_g \mathbf{b}_\alpha$, where $\mathbf{b}_\alpha$ form the basis of the matrix crystal (LiF for the system ^{8}Li-^{6}Li). Then the supercell is continued periodically to cover all space. If $N_d \to \infty$, then we go back to random media. It is very important that the initial condition remains of the correct form and that it is not continued periodically. The eigenvalue problem

$$\sum_j B_{ij} \phi_j(m) = \varepsilon_m \phi_i(m) \tag{17}$$

has Bloch's solution

$$\phi_j(m) = \exp(i\mathbf{k}\mathbf{r}_j)\chi_j(\mathbf{k}, \mu) \,, \quad m = \{\mathbf{k}, \mu\} \,, \tag{18}$$

where $\chi_j(\mathbf{k}, \mu)$ has periodical dependence on impurity position $\mathbf{r}_j$ and $\mathbf{k}$ belongs to Brillouin zone V_B formed by all $\mathbf{k}$ satisfying the condition

$$|\mathbf{k}\mathbf{R}_\alpha| \le \pi \,.$$

Therefore, we have a finite eigenproblem for $\chi_j(\mathbf{k}, \mu)$:

$$\sum_{j=0}^{N_d-1} \bar{B}_{ij}(\mathbf{k})\chi_j(\mathbf{k}, \mu) = \varepsilon_{\mathbf{k},\mu}\chi_j(\mathbf{k}, \mu) \,, \tag{19}$$

$$\bar{B}_{ij}(\mathbf{k}) = \sum_{n_\alpha} \mathrm{e}^{-i\mathbf{k}(\mathbf{r}_i - \mathbf{r}_j - \mathbf{R}(\mathbf{n}))} B_{ij}\left(\mathbf{r}_j + \mathbf{R}(\mathbf{n})\right) \,. \tag{20}$$

Here $\mathbf{R}(\mathbf{n}) = \sum_{\alpha=1}^3 n_\alpha \mathbf{R}_\alpha$, and $B_{ij}\left(\mathbf{r}_j + \mathbf{R}(\mathbf{n})\right)$ is B_{ij} with substitution $\mathbf{r}_j \to \mathbf{r}_j + \mathbf{R}(\mathbf{n})$. As a result

$$p_{j0}(t) = \frac{1}{V_B} \int_{V_B} \mathrm{d}^3 k \exp\left(i\mathbf{k}\mathbf{r}_j\right) \left(\exp\left(-\bar{B}(\mathbf{k})t\right)\right)_{j0} \,, \tag{21}$$

$$P_{\mathbf{00}}(t) = \frac{1}{V_B} \int_{V_B} \mathrm{d}^3 k \left\langle \left(\exp\left(-\bar{B}(\mathbf{k})t\right)\right)_{00} \right\rangle_c \,, \tag{22}$$

$$P_{\mathbf{x0}}(t) = \frac{c}{V_B} \int_{V_B} \mathrm{d}^3 k \exp\left(i\mathbf{k}\mathbf{x}\right) \left\langle \left(\exp\left(-\bar{B}(\mathbf{k})t\right)\right)_{10} \right\rangle_c^{(\mathbf{r}_1 = \mathbf{x})} \,. \tag{23}$$

Here $\langle(\exp(-\bar{B}(\mathbf{k})t))_{10}\rangle_c^{(\mathbf{r}_1=\mathbf{x})}$ means that averaging is fulfilled under condition that site $\mathbf{x}$ is occupied by an impurity spin having number 1 (the site $\mathbf{0}$ is always occupied by spin "0").

For identical impurities

$$P_{\mathbf{00}}(t) = \frac{1}{V_B N_d} \int_{V_B} \mathrm{d}^3 k \left\langle \mathrm{Tr}\left\{\exp\left(-\bar{B}(\mathbf{k})t\right)\right\} \right\rangle_c \,. \tag{24}$$

The Eqs. (22–24) were applied for numerical simulation basing on modern programs of matrix diagonalization. The results for the ^{8}Li-^{6}Li system and for electrodipole transport of localized electronic excitations among identical impurities can be found in [19].

Nuclear Relaxation via Paramagnetic Impurities

Another important generalization of Forster's process is presented by nuclear relaxation via paramagnetic impurities, which is the main relaxation channel in isolators if nuclear spin $I = 1/2$. The fundamental studies [2,24,25,26] were based on calculation of the linear (in impurity concentration c) term of the time dependence of the sample magnetization $P(t) = 1 - cQ(t)$ (for 3d-systems) with following substitution $P(t) = \exp(-cQ(t))$. Analysis of two- and one-dimensional problems was absent whereas experiments are already aimed at fractal objects [27]. Therefore, new theory was constructed [28]. It produced the relation $P(t) = \exp(-cQ(t))$ as the main approximation, and the function $Q(t)$ is calculated for arbitrary $d \le 3$.

The process used to be described by the kinetic equation

$$\frac{\partial p(\mathbf{x}, t)}{\partial t} = D\Delta p(\mathbf{x}, t) - \sum\nolimits_{\mathbf{z}} n_{\mathbf{z}} \nu_{\mathbf{zx}} p(\mathbf{x}, t) \,,$$
$$\nu_{\mathbf{zx}} = \frac{\nu_0 r_0^6}{|\mathbf{x} - \mathbf{z}|^6} = \frac{C}{|\mathbf{x} - \mathbf{z}|^6} \,, \tag{25}$$

with the initial condition $p(\mathbf{x}, t) = p_0$. Here $p(\mathbf{x}, t)$ is polarization of the nucleus, placed at the crystal site $\mathbf{x}$, D is spin-diffusion coefficient, and the angular dependence of $\nu_{\mathbf{zx}}$ is neglected together with the difference in eigenvalues of the diffusion tensor. The observable nuclear polariza-

tion (normalized to $\bar{p}(t=0)=1$) is

$$\bar{p}(t) = \frac{1}{\Omega}\int \mathrm{d}^d x p(\mathbf{x},t) = \langle 0\,|G(t)|\,0\rangle = \langle 0\,|\langle G(t)\rangle_c|\,0\rangle\,. \tag{26}$$

Here d is the space dimensionality, Ω is the crystal volume, the symbol $|0\rangle$ presents a vector having components $\langle \mathbf{x}|0\rangle = 1/\sqrt{\Omega}$, and the propagator $G_{\mathbf{xy}}(t) = \langle \mathbf{x}|G(t)|\mathbf{y}\rangle$ obey Eq. (25), but for initial condition $G_{\mathbf{xy}}(t=0) = \delta(\mathbf{x}-\mathbf{y})$.

Expansion of the observable $\bar{p}(t)$ in concentration powers [16] gives in first terms

$$\begin{aligned}\bar{p}(t) &= \langle 0\,|G(t)|\,0\rangle = \left\langle 0\left|G^{(0)}(t)\right|0\right\rangle \\ &\quad + n\int \mathrm{d}^d r\left\langle 0\left|\left[G^{(1)}(t,\mathbf{r}) - G^{(0)}(t)\right]\right|0\right\rangle + O(n^2) \\ &= \exp\left(-M_0(t)(1+O(n^2))\right),\end{aligned} \tag{27}$$

$$M_0(t) = n\int \mathrm{d}^d r\left\langle 0\left|\left[G^{(0)}(t) - G^{(1)}(t,\mathbf{r})\right]\right|0\right\rangle\,. \tag{28}$$

The propagator $G^{(0)}(t)$ corresponds to evolution in the absence of acceptors, and $G^{(1)}(t,\mathbf{r})$ is the propagator of the system, having only one acceptor, placed at $\mathbf{r}$. Then, using operator representation of Eq. (25), translational invariance, resolvent identity and spectral expansion, we have

$$\begin{aligned}M_0(\lambda) &= \int_0^\infty \mathrm{d}t e^{-\lambda t} M_0(t) \\ &= \frac{n\Omega}{\lambda}\left\langle 0\left|U_0\frac{1}{\lambda + A + U_0}\right|0\right\rangle \\ &= \frac{n\Omega}{\lambda}\left\langle 0\left|U_0\frac{1}{(A+U_0)(\lambda+A+U_0)}U_0\right|0\right\rangle \\ &= \frac{n\Omega}{\lambda}\sum_n \frac{|\langle n\,|U_0|\,0\rangle|^2}{(\lambda+E_n)\,E_n}\,.\end{aligned} \tag{29}$$

Here $A = -D\Delta$, $\langle \mathbf{x}|U_0|\mathbf{z}\rangle = \delta_{\mathbf{xz}}v_{\mathbf{x0}}$, and $(A+U_0)|n\rangle = E_n|n\rangle$. This representation is too complex for direct calculations, but it is useful for different asymptotical treatments (see, for example, [8]). The most important are short time and long time studies, because a satisfactory precision can be achieved with the aid of the representation

$$M_0(t) = M_F(t) + M_1(t)\,. \tag{30}$$

Here the first term

$$M_F(t) = n\int \mathrm{d}^d x\,(1-\exp(-v_{\mathbf{x0}}t)) = (\beta_F t)^{d/s}$$

describes the initial (Forster's) part of the relaxation, while $M_1(t)$ is the long-time asymptotic expression, calculated on the basis of the representation (30) and formally continued to arbitrary positive values of t. Here we use $v_{\mathbf{x0}} = v_0 r_0^s/x^s$ to clarify some parametrical dependencies. The results, obtained for $M_1(t)$ at arbitrary $d \le 3$, are presented in detail in Ref [28]. In particular, at long time

$$\begin{aligned}M_1(t,d=1) &= 4n\sqrt{\frac{Dt}{\pi}}\,, \\ M_1(t,d=2) &= \frac{4\pi Dnt}{\ln\left(Dt/b^2\right)}\,, \\ M_1(t,d=3) &= 4\pi Dbnt\,,\end{aligned} \tag{31}$$

where $b \propto (C/D)^{1/(s-2)}$ is a "scattering length", which incorporates all dependence on the "potential" $v_{\mathbf{x0}}$. It is evident, that at $d=1$ the dependence on b is absent here, and for $d=2$ it is rather weak.

In order to find an argument for regrouping (27) of the concentration expansion, we represent the propagator $G(t)$ in the form

$$\begin{aligned}G(t) &= \langle \exp(-(A+U)t)\rangle_c = \exp(-B(t))\,, \\ B(t) &= At + M(t)\,,\end{aligned} \tag{32}$$

$$U = \sum\nolimits_{\mathbf{z}} n_{\mathbf{z}} U^{\mathbf{z}}\,, \quad U^{\mathbf{z}}_{\mathbf{xq}} = \delta_{\mathbf{xq}} v_{\mathbf{xz}}\,. \tag{33}$$

The operator $M(t)$, which is as-yet undefined, can be written in the form $M(t) = \sum_{\mathbf{z}} M_{\mathbf{z}}(t)$ similar to the form of U in (33). It can be said that the operators $M_{\mathbf{z}}(t)$ must adequately describe the effect of acceptors in the so-called effective medium that appears upon averaging over the configurations of acceptors. It is therefore natural to assume that, on average, the propagator $G(t)$ undergoes no changes if one of the sites of the effective medium is replaced by an actual one and if the result is thereupon averaged over the distribution of acceptors; that is,

$$G(t) = \left\langle \exp\left(-At - M(t) + M^z(t) - n_z U^z t\right)\right\rangle_c\,. \tag{34}$$

Relations (32)–(34) form a closed set of nonlinear operator equations. The solution, according to [28], practically coincides with $P(t) = \exp(-M_0(t))$, if $cQ(t) = M_0(t) \sim 1$ and $c \ll 1$, and corrections are important for longer time.

It should be stressed, that (1) to clarify the influence of corrections to this solution in more detail we can calculate the next ($\propto c^2$) term of the concentration expansion, and (2) there exist physical [29] and mathematical [30,31] studies, giving the law $\log\left(1/\bar{p}(t\to\infty)\right) \propto t^{d/(d+2)}$, which is expected [32] to be valid at $\bar{p}(t) \lesssim 10^{-12}$. The asymptotics can be received by methods of field theory, but up to a preexponential multiplier only [33], therefore these

methods did not produce useful results for the problem of random walks in disordered media, discussed in Sect. "Delocalization of Nuclear Polarization in a Disordered Spin System", where the exponential part of the asymptotics is absent.

Resonance Line Form Function for Magnetically Diluted Solids

The line shape and the Fourier-transform-related free induction decay (FID) belong to the most important observable values in the physics of magnetic resonance. In the study of nuclear spin systems forming a crystal lattice the first ($\propto t^2$ and $\propto t^4$) terms of the expansion of FID in powers of time carry information of high importance [2]. In the theory of the line shape of disordered (magnetically diluted) electron spin systems, the first ($\propto c$ and $\propto c^2$) terms in the expansion in powers of the concentration c of paramagnetic centers play the same role [6]. The third ($\propto c^2$) term was calculated in a recent study [34] for the first time. This consideration is particularly topical in connection with the new experiments on measuring the EPR spectra of paramagnetic impurities distributed at the solid surface [35,36]. It can be expected, that modern pulse methods will produce new possibilities for measurement of FID in such magnetically disordered nuclear systems as ^{29}Si in silicon crystals for example. Some efforts in this direction can be found in the recent article [37] and in references therein.

Let the paramagnetic centers (PCs) be randomly distributed in a d-dimensional crystal lattice with the primecell volume Ω_c. The free induction decay in the high-temperature approximation is given by

$$G(t) = \left\langle \left\langle S^+(t)S^- \right\rangle_0 \right\rangle_c / \left\langle \left\langle S^+S^- \right\rangle_0 \right\rangle_c \ , \tag{35}$$

where $S^{\pm} = \sum_{\mathbf{r}} n_{\mathbf{r}} S^{\pm}_{\mathbf{r}}$, $S^{\pm}_{\mathbf{r}} = S^x_{\mathbf{r}} \pm i S^y_{\mathbf{r}}$, $S^+(t) = \mathrm{e}^{iHt} S^+ \mathrm{e}^{-iHt}$, $n_{\mathbf{r}}$ is the occupation number, $\langle \cdots \rangle_0 = \mathrm{Tr}\,(\cdots)\,/\mathrm{Tr}1$, $\langle \cdots \rangle_c$ stands for the averaging over the spatial spin distributions (over occupation numbers), and H is the secular part of dipole–dipole interactions:

$$H = \frac{3}{4} \sum_{rq} n_{\mathbf{r}} n_{\mathbf{q}} A(\mathbf{r}, \mathbf{q}) \left(S^z_{\mathbf{r}} S^z_{\mathbf{q}} - \frac{a}{3} \mathbf{S}_{\mathbf{r}} \mathbf{S}_{\mathbf{q}} \right) \ . \tag{36}$$

Here $A(\mathbf{r}, \mathbf{q}) = \hbar \gamma^2 \left(1 - 3\cos^2 \vartheta_{rq}\right) / \left|\mathbf{r} - \mathbf{q}\right|^3$, γ is the gyromagnetic ratio, and ϑ_{rq} is the angle between $\mathbf{r} - \mathbf{q}$ and external static field $\mathbf{H}_0$. Parameter $a = 0$ in the Anderson model and $a = 1$ for pure dipole interaction. For other a values, Hamiltonian (36) corresponds to a system with the anisotropic axisymmetric g-factor. In what follows, $S = 1/2$ for all PCs. Let us expand FID in power of $n_{\mathbf{x}}$ and perform configurational averaging [6]. To terms $\propto O(c^3)$ one has

$$G(t) = 1 + c \sum_{\mathbf{r}_1} (2K_{01}(t) - 1) + \frac{c^2}{2} \sum_{\mathbf{r}_1 \neq \mathbf{r}_2} (2K_{012}(t) - 2K_{01}(t) - 2K_{02}(t) + 1) \ , \tag{37}$$

and, applying CMA we have up to terms $\propto O(n^3)$

$$G(t) = 1 + n \int \mathrm{d}^d r_1 \, (2K_{01}(t) - 1) + \frac{n^2}{2} \int \mathrm{d}^d r_1 d^d r_2 \, (2K_{012}(t) - 2K_{01}(t) - 2K_{02}(t) + 1) \ , \tag{38}$$

where $n = c/\Omega_c$ is the d-dimensional PC density, and

$$K_{01}(t) = \left\langle \mathrm{e}^{iH_{01}t} S^+_0 \mathrm{e}^{-iH_{01}t} (S^-_0 + S^-_1) \right\rangle_0 \ ,$$
$$K_{012}(t) = \left\langle \mathrm{e}^{iH_{012}t} S^+_0 \mathrm{e}^{-iH_{012}t} (S^-_0 + S^-_1 + S^-_2) \right\rangle_0 \ .$$

Here, according to Eq. (36),

$$H_{ij} = \tfrac{1}{2} A_{ij} (3 S^z_i S^z_j - a \mathbf{S}_i \mathbf{S}_j), \quad H_{012} = H_{01} + H_{02} + H_{12}, \tag{39}$$

with $A_{ij} = A(\mathbf{r}_i, \mathbf{r}_j)$. The interaction $A_{ij} \propto \left|\mathbf{r}_i - \mathbf{r}_j\right|^{-3}$. Therefore, substitution of integration variables $\mathbf{r}_i \to t^{1/3} \mathbf{r}_i$ excludes time from the integrands and reveals that the mth term in (38) is $\propto (n t^{d/3})^m$. That is Eq. (38) is expanded [16] in terms of dimensionless parameter

$$(D_d t)^{d/3} = n \int \mathrm{d}^d r_1 \, (1 - 2K_{01}(t)) \ , \tag{40}$$

and, hence, can be represented as

$$G(t) = 1 - (D_d t)^{\frac{d}{3}} + \frac{1}{2} \xi_d(a) (D_d t)^{\frac{2d}{3}} + O\left((D_d t)^d\right) . \tag{41}$$

The functions $\xi_d(a)$ were calculated numerically basing on the Eq. (38). It was supposed, that at $d \leq 2$ all field directions are equally probable. For this case, after averaging of (41) we have

$$\bar{G}(t) = 1 - (\bar{D}_d t)^{\frac{d}{3}} + \frac{\bar{\xi}_d(a)}{2} (\bar{D}_d t)^{\frac{2d}{3}} + O\left((\bar{D}_d t)^d\right) , \tag{42}$$

and

$$\bar{D}_d = \beta_d n^{d/3} \gamma^2 \hbar \ , \quad \beta_3 = \frac{2\pi^2}{3\sqrt{3}} \ ,$$
$$\beta_2 = 4.647 \ , \qquad \beta_1 = 6.348 \ . \tag{43}$$

It should be noticed, that $\bar{D}_3 = D_3$, and D_3 coincides with Anderson's result [4].

The results are presented in the following table:

a	0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6
ξ_3	1	1.01	1.03	1.05	1.08	1.11	1.13	1.15	1.18
$\bar{\xi}_2$	1.027	1.07	1.11	1.15	1.19	1.22	1.25	1.27	1.29
$\bar{\xi}_1$	1.062	1.16	1.20	1.25	1.28	1.32	1.35	1.37	1.40

(44)

These relations can be used directly for analysis of the wings of experimental EPR lines [34]. But treatment of the full line or FID requires regrouping of the expansion (42) in such a way to receive a physically adopted result for all t. One of the simplest approaches [38,39,40] was generalized for this aim in [34]. It introduces the most essential properties of the disordered systems into the Anderson–Weiss–Kubo (AWK) theory [41,42], which was originally developed for the description of line narrowing by motion. The result is of the form

$$G(t) = \exp\left(-\left(2B_d^2 \int_0^t \mathrm{d}\tau (t-\tau) F(B_d \tau)\right)^{d/6}\right) ,$$
$$F(x) = \exp\left(-(q_d x)^{d/3}\right) . \tag{45}$$

Being approximative Eq. (45) nevertheless reproduces the structure of the expansion (41), non-negativity of the resonance line shape and parameters B_d and q_d can be defined using D_d and ξ_d, that gives

$$B_3 = D_3 , \qquad B_2 = \bar{D}_2 ,$$
$$q_3 = 3(\xi_3 - 1) , \quad q_2 = \left((10/3)(\bar{\xi}_2 - 1)\right)^{3/2} . \tag{46}$$

The dimensionality $d = 1$ requires more refined treatment, because system orientations near the "magic" direction $\theta = \arccos(1/\sqrt{3})$ produce a singularity $g_1(\omega \to 0) \propto \ln(1/\omega)$.

In application to pure dipole interaction with $a = 1$ the Eqs. (46) and (44) give $q_3 = 0.33 \approx 1/3$ and $q_2 = 0.63$. Fitting of the experimental data of [35] in the region, where the sample was considered as two-dimensional gives $q_2 = q_2^{\mathrm{exp}} \approx 0.05$ with significant distinctions from both $q_2 = 0$, and $q_2 = 0.63$. Additional studies are necessary to clarify the nature of this deviation from pure dipole evolution.

Application of the approach to magnetically diluted nuclear spin systems (like ^{29}Si in silicon) requires one important modification connected with the fact that typical nuclear concentration $c = 0.01 \div 0.1$ is not sufficiently small to replace lattice sums by integrals in Eq. (38). Therefore, FID for such concentrations can be represented by the Eq. (45) with substitution

$$(D_d t)^{d/3} \to \varkappa(t) = c \sum_{\mathbf{r}_1} (1 - 2K_{01}(t)) . \tag{47}$$

For example, if $d = 3$ and $a = 1$, then Eqs. (45), (46) produce

$$G(t) = \exp\left(-\left(6\left(D_3 t + 3\exp(-D_3 t/3) - 3\right)\right)^{1/2}\right) . \tag{48}$$

Applying the recipe (47) we receive the representation

$$G(t) = \exp\left(-\left(6\left(\varkappa(t) + 3\exp\left(-\varkappa(t)/3\right) - 3\right)\right)^{1/2}\right) , \tag{49}$$

which is exact at small $\varkappa(t)$ up to $O(\varkappa^2(t))$ and coincides with (48) for $\varkappa(t) \gtrsim 1$.

Saturation on the Wing of a Dipole-Broadened EPR Line and Cluster Expansions

A bright demonstration of peculiarities of spin dynamics in disordered media is realized in the shape of the hole burned on the wing of the EPR line. The EPR spectrum in dipole-broadened $3d$ solids has the Lorentz wing, but experimentation [43] revealed that the wing of the hole falls down exponentially contrary to expectations, that a dipole broadened line should be homogeneous. To solve this contradiction a new method of cluster expansions [39,40] was invented, which gives a constructive alternative to both concentration expansion, discussed in Sect. "Resonance Line Form Function for Magnetically Diluted Solids", and spin pockets conception [44], used in [43] for interpretation of the data.

If interaction inside a pair of spins exceeds interactions with any other spin, then the spectrum of the pair is similar to a discrete one, broadened by small interactions with surrounding spins. The pair forms a two-spin cluster (2-cluster). Analogically we can define 3- and other many-spin clusters. It is sufficient for many tasks to divide all spins into three groups: 2- and 3-clusters and all other "mass" spins. Numerical analysis [39] shows, that 2- and 3-clusters contain 51(1)% and 11(1)% of all spins correspondingly, but 2-clusters define all Lorentz wings of the EPR line. Mass spins (similar to nuclear spins in a crystal) have finite heat capacity, which is defined by interaction at average distance contrary to infinite heat capacity of a full dipole–dipole reservoir, calculated in CMA. Mass spins produce fluctuating fields, which gave the main broadening of the clusters spectra. Transitions between states of the cluster take place due to interac-

tions with other clusters and with mass spins, and they are slow. As a result saturation at the EPR line wing induce transitions between states of the 2-cluster, and the line shape of the hole is defined by interaction of the cluster with mass spins. If this interaction is estimated within the AWK model, then any mass configuration produces exponential wings of the hole $g(\Delta) \propto \exp(-\Delta/\mu)$, where Δ is detuning of the saturating field from the cluster transition frequency, and μ is defined by magnitude of the fields produced by other spins on the cluster, and by the rate of their fluctuations. After configurational averaging the wings became $\langle g(\Delta)\rangle_c \propto \Delta^{-4}$, but averaged hole shape described the transitions for small magnitude ω_1 of the saturating field only. In a general case the observable area of the hole σ should be averaged. For large ω_1 we have $\sigma \propto \ln\left(t_p\omega_1^2/\mu\right)$, where t_p is the duration of the saturating pulse. Logarithm is a slow function, therefore $\langle\sigma\rangle_c \propto \ln\left(t_p\omega_1^2/\langle\mu\rangle_c\right)$, which was observed in [43] and was interpreted as the exponential wing of Portis's pockets. The described theory produces a microscopic picture of the phenomenon and indicates that Portis's pockets have limited relation to the problem, because they are homogeneous by definition, i. e. $g(\Delta) = \langle g(\Delta)\rangle_c$, while in the cluster theory this relation isn't fulfilled.

Future Developments

Methods described in Sect. "Delocalization of Nuclear Polarization in a Disordered Spin System" are expected to be useful for understanding of studies of delocalization of relaxed excitons in different problems which can be studied by optical methods (fluorescence and multi-wave mixing) first of all. We expect that results obtained in this section will also produce a reliable basis for development of general methods of statistical physics and path integration theory in order to reproduce and clarify these results and many other problems of nonlinear field theory.

The results of Sect. "Nuclear Relaxation via Paramagnetic Impurities" create a new regular method for treatment of a wide class of problems of disordered media. The method is similar to, but, nevertheless, different from standard coherent (effective) potential approximation. We expect that the same ideas applied to the second term of the concentration expansion can produce better understanding for many related problems.

The results of Sect. "Resonance Line Form Function for Magnetically Diluted Solids" require new experimental studies for all spatial dimensions. Checking on nuclear spin systems like ^{29}Si in silicon crystals is most desirable, because the details of interaction of nuclei are known much better then for paramagnetic centers.

Cluster expansion, described in Sect. "Saturation on the Wing of a Dipole-Broadened EPR Line and Cluster Expansions" should be useful for different problems of spectral transport in disordered solids.

The work was supported by the Russian Foundation for Basic Research, project #06-02-17088.

Bibliography

1. Balesku R (1975) Equlibrium and Non-Equilibrium Statistical Mechanics. Wiley, New York
2. Abragam A, Goldman M (1982) Nuclear Magnetism: Order and Disorder. Clarendon Press, Oxford
3. Forster T (1949) Z Naturforsch (A) 4:321
4. Anderson PW (1951) Phys Rev 82:342
5. Dzheparov FS, Lundin AA (1978) Sov Phys JETP 48:514
6. Dzheparov FS, Lundin AA, Khazanovich TN (1987) Sov Phys JETP 65:314
7. Feldman EB, Lacelle S (1996) J Chem Phys 104:2000
8. Abov YG, Bulgakov MI, Borovlev SP, Gul'ko AD, Garochkin VM, Dzheparov FS, Stepanov SV, Trostin SS, Shestopal VE (1991) Sov Phys JETP 72:534
9. Dzheparov FS (1999) JETP 89:753
10. Dzheparov FS (1991) Sov Phys JETP 72:546
11. Abov YG, Gulko AD, Dzheparov FS, Stepanov SV, Trostin SS (1995) Phys Part Nucl 26:692
12. Feynman RP (1972) Statistical mechanics. W.A. Benjamin, Massachusetts
13. Dzheparov FS, L'vov DV, Shestopal VE (1998) JETP 87:1179
14. Rieger PT, Palese SP, Miller RJD (1997) Chem Phys 221:85
15. Ziman JM (1979) Models of Disorder: the Theoretical Physics of Homogeneously Disordered Systems. Cambridge University Press, Cambridge
16. Dzheparov FS, Smelov VS, Shestopal VE (1980) JETP Lett 32:47
17. Dzheparov FS, Shestopal VE (1993) Theor Mat Phys 94:496
18. Dzheparov FS, L'vov DV, Shestopal VE (2007) J Supercond Nov Magn 20(2):175
19. Dzheparov FS (2005) JETP Lett 82(8):521
20. Dzheparov FS, Gul'ko A, Heitjans P, L'vov D, Schirmer A, Shestopal V, Stepanov S, Trostin S (2001) Physica B297:288
21. Abov YG, Gul'ko AD, Dzheparov FS (2006) Phys At Nuclei 69:1701
22. Gul'ko AD, Ermakov ON, Stepanov SV, Trostin SS (2007) J Supercond Nov Magn 20(2):169
23. Dzheparov FS, Gul'ko AD, Ermakov ON et al (2008) In: Salikhov KM (ed) Modern development of magnetic resonance, Abstracts of the International conference Zavoisky-100. Kazan, pp 47–48, Appl Magn Reson 35(3)
24. Khutsishvili G (1965) Sov Phys Usp 8:743
25. Alexandrov IV (1975) Theory of Magnetic Relaxation. Nauka, Moscow (in Russian)
26. Atsarkin VA (1980) Dynamical Nuclear Polarization. Nauka, Moscow (in Russian)
27. Tabti T, Goldman M, Jacquinot JF, Fermon C, LeGoff G (1997) J Chem Phys 107:9239
28. Dzheparov FS, Jacquinot JF, Stepanov SV (2002) Phys At Nuclei 65:2052
29. Balagurov VY, Vaks VT (1973) Sov Phys JETP 38:968
30. Pastur LA (1977) Teor Mat Fiz 32:88

31. Donsker M, Varadhan S (1975) Commun Pure Appl Math 28:525
32. Havlin S, Dishon M, Kiefer JE, Weiss GH (1984) Phys Rev Lett 53:407
33. Renn SR (1986) Nucl Phys B275(FS17):273
34. Dzheparov FS, Kaganov IV (2002) JETP Lett 75:259
35. Atsarkin VA, Vasneva GA, Demidov VV, Dzheparov FS, Odintsov BM, Clarkson RB (2000) JETP Lett 72:369
36. Atsarkin VA, Demidov VV, Vasneva GA, Dzheparov FS, Ceroke PJ, Odintsov BM, Clarkson RB (2001) J Magn Res 149:85
37. Li D, Dementyev AE, Dong Y, Ramos R, Barrett SE (2007) Phys Rev Lett 98:190401
38. Grinberg ES, Kochelaev BI, Khaliullin GG (1981) Sov Phys Solid State 23:224
39. Dzheparov FS, Henner EK (1993) JETP 77:753
40. Dzheparov FS, Kaganov IV, Khenner EK (1997) JETP 85:325
41. Anderson PW, Weiss PR (1953) Rev Mod Phys 25:269
42. Kubo R (1962) J Phys Soc Jpn 17:1100
43. Atsarkin VA, Vasneva GA, Demidov VV (1986) Sov Phys JETP 64:898
44. Portis AM (1953) Phys Rev 91:1071

Spin-Polarized Quantum Transport in Mesoscopic Conductors: Computational Concepts and Physical Phenomena

Michael Wimmer, Matthias Scheid, Klaus Richter
Institut für Theoretische Physik, Universität Regensburg, Regensburg, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Numerical Quantum Transport
Matrix Reordering Strategies for Quantum Transport
Spin Filtering in Nanostructures
Pure Spin Current Generation
Future Directions
Acknowledgments
Bibliography

Glossary

Aharonov Bohm effect The magnetic flux enclosed in between propagating quantum mechanical waves shifts their relative phases as a result of the underlying electromagnetic vector potential. This gives rise to distinct oscillations in the magnetoconductance of a ring conductor.

Landauer–Büttiker formalism For phase-coherent quantum transport, the Landauer–Büttiker formalism relates the conductance of a device to the transmission probability of charge carriers.

Rashba- and Dresselhaus spin-orbit coupling Coupling of the spin degree of freedom to the orbital motion of charge carriers due to structural or bulk inversion asymmetry in semiconductors.

Ratchets Devices that convert unbiased fluctuations or perturbations into directed motion.

Spintronics Extension of charge-based electronics in metals or semiconductors by utilizing the spin degree of freedom of the charge carriers.

Definition of the Subject

Mesoscopic conductors are electronic systems of sizes in between nano- and micrometers, and often of reduced dimensionality. In the phase-coherent regime at low temperatures, the conductance of these devices is governed by quantum interference effects, such as the Aharonov–Bohm effect and conductance fluctuations as prominent examples. While first measurements of quantum charge transport date back to the 1980s, spin phenomena in mesoscopic transport have moved only recently into the focus of attention, as one branch of the field of spintronics. The interplay between quantum coherence with confinement-, disorder- or interaction-effects gives rise to a variety of unexpected spin phenomena in mesoscopic conductors and allows moreover to control and engineer the spin of the charge carriers: spin interference is often the basis for spin-valves, -filters, -switches or -pumps. Their underlying mechanisms may gain relevance on the way to possible future semiconductor-based spin devices.

A quantitative theoretical understanding of spin-dependent mesoscopic transport calls for developing efficient and flexible numerical algorithms, including matrix-reordering techniques within Green function approaches, which we will explain, review and employ.

Introduction

Charge and spin transport through phase-coherent conductors of mesoscopic scales carry imprints of wave interference as predominant and characteristic features: in the simplest case of a point contact, the conductance increases stepwise with Fermi energy, reflecting the discrete number of quantized open transverse channels contributing to transport; for more complex mesoscopic systems, such as ballistic quantum dots or diffusive conductors, the conductance typically wildly fluctuates, upon varying the Fermi energy or other parameters, around its classical

mean value. Among the different effects on charge transport, the Aharonov–Bohm (AB) effect represents possibly the most genuine interference phenomenon at the heart of mesoscopic physics: The magnetoconductance of a ring conductor coupled to two leads exhibits distinct sinusoidal oscillations when monitored as a function of a perpendicular magnetic field threading the ring, with a period given by the magnetic flux quantum. As the AB signal stems from interfering waves traveling through the two different arms of the ring, it requires phase coherent wave functions extending over the ring typically on micron scales [1]. Hence the AB effect is frequently being used as a tool to investigate phase coherence and dephasing mechanisms of the orbital part of the wave functions, while the spin degree of freedom was usually neglected.

With rising interest in spin-dependent transport, the interplay of the electron spin and charge degree of freedom has been exploited in a variety of spin interference devices, to be discussed below. Different types of couplings to the electron spin have been considered for spin engineering in non-magnetic conductors. On the one hand this is possible through the Zeeman coupling to an externally applied magnetic field. Non-uniform B-fields with spatially varying direction are being employed to achieve a tailored spin dynamics, including the possibility for guided spin evolution or triggering spin flips. On the other hand, intrinsic spin-orbit (SO) interaction proves to be relevant in spin-dependent transport. It exists in systems with bulk inversion asymmetry and/or structure inversion asymmetry, e. g. due to the vertical confinement in semiconductor heterostructures (Rashba SO coupling [2,3]).

Among the mesoscopic spin interference systems considered in the literature, ring geometries have again played an important role, opening up the field of spin-based AB physics (see [4,5] for recent accounts including overviews over the literature). This includes topics such as Berry phase-signatures in transport, both theoretically [6,7,8,9,10,11,12] and experimentally [13,14,15,16], spin-related conductance modulation [17,18,19], persistent currents [6,20], spin Hall effect [21], spin filters [22,23] and detectors [24], and spin switching mechanisms [4,10].

To illustrate how the spin polarization can be tuned by exploiting (orbital and spin) interference in such an AB setup, we consider as an introductory example spin switching in a two-dimensional (2D) ballistic phase-coherent ring symmetrically coupled to two single-channel leads (Fig. 1, [17]). We assume Rashba SO coupling which is relevant in conductors laterally defined on GaAs- or InAs-based two-dimensional electron gases (2DEGs). Rashba SO coupling will be defined and discussed in Sect. "Spin Filtering in Nanostructures". It can be viewed

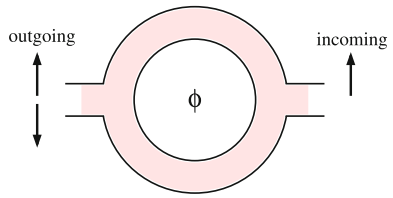

Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 1

Aharonov–Bohm physics with spin: Two-dimensional Aharonov–Bohm ring of mean radius r_0 used for numerical calculations of the conductance presented in Fig. 2. An additional perpendicular magnetic field B generates a flux $\phi = \pi r_0^2 B$. The *gray zone* corresponds to the region subject to a finite Rashba coupling switched adiabatically on and off in the leads (from [17])

as the coupling of the spin to a fictitious in-plane magnetic field directed perpendicular to the electron momentum. Hence in a ring it points mainly in radial direction. The strength of the SO field can be tuned by an external gate voltage [25] allowing to control experimentally the spin evolution.

We assume *spin-polarized* spin-up electrons entering the ring from the right (see Fig. 1). Figure 2 displays numerically computed (see Sect. "Numerical Quantum Transport") conductance traces as a function of the external magnetic flux ϕ for weak and moderate Rashba strengths. The overall conductance is presented as a solid line, and its spin-resolved components, $G^{\uparrow\uparrow}$ and $G^{\downarrow\uparrow}$, corresponding to outgoing spin-up and -down channels, are shown as dashed and dotted lines, respectively. In the weak SO coupling limit, Fig. 2a, the overall conductance (solid line) shows the usual AB oscillations of period ϕ_0 and is dominated by $G^{\uparrow\uparrow}$ (dashed line). As expected for weak spin-coupling, the spin polarization is almost conserved during transport. Interesting features appear for the case of moderate SO coupling depicted in panel (b). There, both components, $G^{\uparrow\uparrow}$ (dashed line) and $G^{\downarrow\uparrow}$ (dotted line), contribute similarly to the overall conductance (solid line). However, the spin polarization of the transmitted electrons varies as a function of the magnetic flux ϕ: $G^{\downarrow\uparrow} = 0$ at $\phi = 0$, while $G^{\uparrow\uparrow} = 0$ at $\phi = \phi_0/2$. Hence, for zero flux all transmitted carriers conserve their original (incoming) spin-orientation, while for $\phi = \phi_0/2$ the transmitted particles reverse their spin polarization. In other words: by tuning the magnetic flux from 0 to $\phi_0/2$ we can reverse the spin-polarization of transmitted particles in a controlled way. Hence the AB ring with SO coupling acts as a tunable spin-switch. This switching can be traced back to constructive or destructive AB interference [4].

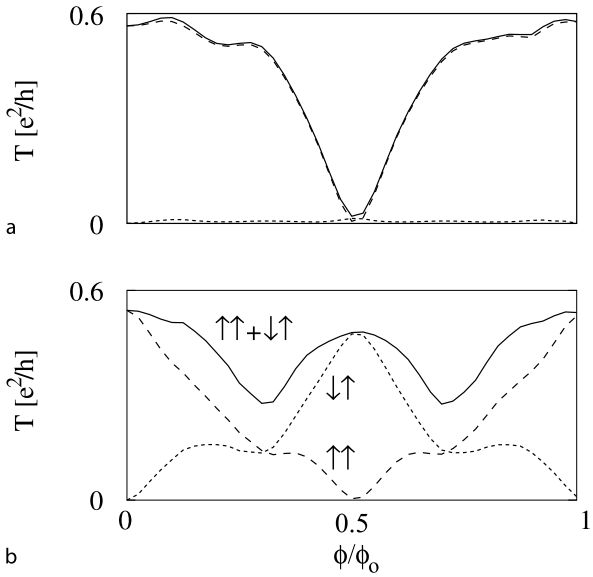

Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 2

Mesoscopic Aharonov–Bohm spin interference: The conductance of spin-up polarized carriers entering a single-channel two-dimensional Aharonov–Bohm ring (see Fig. 1) is shown as a function of flux $\phi = \pi r_0^2 B$ (in units of the flux quantum $\phi_0 = hc/e$) through the ring in the presence of Rashba spin-orbit coupling. For panel a and b the scaled spin-orbit coupling strength takes the values 0.2 and 1.0, expressed as the product $\omega_R T_0/2\pi$ of the precession frequency ω_R of the spin around the effective spin-orbit magnetic field and the time T_0 for traveling of the electrons around the ring. The overall conductance (*solid line*) is split into its components $G^{\uparrow\uparrow}$ (*dashed line*) and $G^{\downarrow\uparrow}$ (*dotted line*). Note in panel b the continuous change of the spin polarization, related to $G^{\uparrow\uparrow} - G^{\downarrow\uparrow}$, with ϕ and the spin switching at $\phi = \phi_0/2$ (adapted from [17])

Switching a given spin polarization requires the generation of spin-polarized particles in non-magnetic mesoscopic conductors in the first place. Since spin injection from ferromagnets into a semiconductor remains problematic [26], alternative proposals have been made to achieve spin-polarized currents or spin accumulation without magnets, which we will briefly review in Sect. "Spin Filtering in Nanostructures". Among those are the spin Hall effect [27] and, in the context of coherent mesoscopic transport, concepts for Zeeman- and SO-mediated adiabatic spin pumping and spin ratchets.

For a recent account on spin phenomena in systems of reduced dimensions see [28]; for a review on the related field of magnetization dynamics and pumping in layered magnetic heterostructures see [29].

A complete and quantitative understanding of spin phenomena in the mesoscopic realm requires computational approaches to quantum transport which also serve as reference calculations for analytical predictions usually based on model assumptions. However, also numerical approaches cannot cope with the full many-body transport problem without relying on approximations. Here we focus on mesoscopic conductors, i. e. systems with a considerable number of electrons, with strong coupling to external leads. Then, a mean-field treatment is usually justified which allows one to reduce the Hamiltonian to a single-particle problem with an effective confinement potential resulting from a combination of external and mean-field potentials.

We further consider coherent transport close to equilibrium at relatively low bias, excluding inelastic effects, such that the Landauer approach to transport is justified. However, even in this case brute-force computational approaches quickly reach their limits: Conductors at mesoscopic scales are typically characterized by extensions which are (at least in one direction) much larger than the Fermi wavelength of the charge carriers, the shortest quantum scale involved. This implies rapidly oscillating, complex and often irregular spinor wave functions extending throughout the systems which require for the quantum mechanical numerical solution either huge sets of basis functions or, in tight-binding approaches, the use of rather fine, preferentially adapted grids. The strength of the widespread tight-binding approaches for transport discussed and reviewed here lies in their flexibility and general applicability. Moreover, tight-binding transport codes can be combined with density-functional (DFT) calculations for structure and electronic properties of the nano-conductors by using the parameters computed within DFT as input. This approach is frequently applied to transport in nano- or molecular electronics.

The paper is organized as follows: In the methodological Sect. "Numerical Quantum Transport" we will first briefly summarize and provide the key relations for spin quantum transport within the Landauer framework. In Sect. "Matrix Reordering Strategies for Quantum Transport" we focus on and explain in some detail advanced computational concepts, making use of graph-theory, to implement powerful and flexible algorithms for tight-binding transport codes. The numerical strength of the codes is demonstrated for ring-type geometries showing that by efficient tight-binding implementation one can gain orders of magnitude in performance. In Sect. "Spin Filtering in Nanostructures" and "Pure Spin Current Generation" we employ these numerical schemes to address two important aspects of spin-dependent transport, namely spin filtering and generating pure spin currents. To this end we focus on laterally-confined 2D ballistic nanostructures with Rashba SO interaction. We conclude

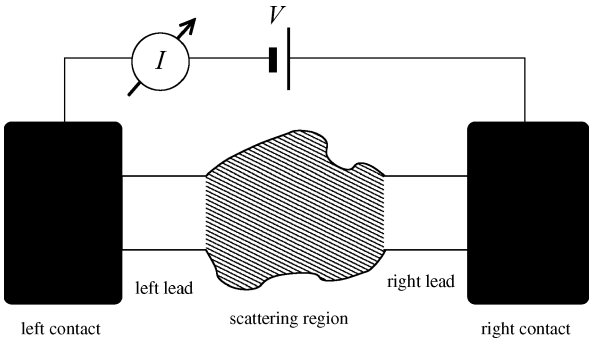

Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 3
Schematic view of a two-point measurement setup

with an outline of future research directions in mesoscopic spin transport which we consider important.

Numerical Quantum Transport

Landauer–Büttiker Transport Theory

If the dimensions of a device get smaller than the phase coherence length l_ϕ of charge carriers, classical transport theories are not valid any more. Instead, carrier dynamics is now governed by quantum mechanics and the wave-like nature of particles becomes important. In general, the conductance/resistance of such a device does not follow Ohm's law.

Consider a two-point measurement setup as shown in Fig. 3: A scattering region is connected to large (phase-breaking) reservoirs by leads. The leads are assumed to be perfect and infinitely long to define asymptotic eigenstates $\phi_{n,\sigma}(y)\mathrm{e}^{\pm \mathrm{i}kx}$ at energy E, where n is the quantum number of transverse confinement —also called the *channel* number —and σ is the spin index. The total scattering eigenstate $\psi_{n,\sigma}$ originating from channel n with spin σ in the left lead is, within the lead region, given by

$$\psi_{n,\sigma}(\mathbf{x}) = \begin{cases} \phi_{n,\sigma}(y)\mathrm{e}^{\mathrm{i}kx} + \sum_m r_{m,\sigma';n,\sigma}\ \phi_{m,\sigma'}(y)\mathrm{e}^{-\mathrm{i}kx} & \text{for } \mathbf{x} \text{ in left lead} \\ \sum_m t_{m,\sigma';n,\sigma}\ \phi_{m,\sigma'}(y)\mathrm{e}^{\mathrm{i}kx} & \text{for } \mathbf{x} \text{ in right lead} \end{cases}, \tag{1}$$

and obeys the stationary Schrödinger equation $H\psi_{n,\sigma} = E\psi_{n,\sigma}$. The conductance G in linear response can then be calculated within the *Landauer–Büttiker* formalism [30,31,32] (for tutorials see [1,33]):

$$G = \frac{\mathrm{e}^2}{h}\sum_{n,m}\sum_{\sigma,\sigma'} T_{m,\sigma';n,\sigma} = \frac{\mathrm{e}^2}{h}\ T_\mathrm{C}\,, \tag{2}$$

where $T_{m,\sigma';n,\sigma} = |t_{m,\sigma';n,\sigma}|^2$ is given by the squared transmission amplitudes of the scattering states $\psi_{n,\sigma}$. The fraction e^2/h is called the conductance quantum. The *scattering matrix* $S_{m,\sigma';n,\sigma}$ is a useful definition that combines reflection and transmission amplitudes for both leads into a single matrix. In this notation the index n,σ then also contains information about the respective lead.

The problem of calculating the conductance G is thus reduced to calculating the scattering eigenstates $\psi_{n,\sigma}$. Alternatively, the scattering amplitudes $r_{m,\sigma';n,\sigma}$ and $t_{m,\sigma';n,\sigma}$ can also be derived from the retarded Greens function $G^\mathrm{R}(\mathbf{x},\mathbf{x}')$ of the system. The retarded Greens function for a given energy E obeys the equation

$$(E - H + \mathrm{i}\eta)\ G^\mathrm{R}(\mathbf{x},\mathbf{x}') = \delta(\mathbf{x}-\mathbf{x}')\,, \tag{3}$$

where H is the Hamiltonian of the system and η an infinitesimally small number. Formally this equation can be solved as

$$G^\mathrm{R} = (E - H + \mathrm{i}\eta)^{-1}\,. \tag{4}$$

The transmission and reflection amplitudes are then given by the *Fisher–Lee* relation [34]:

$$t_{m,\sigma';n,\sigma} = -\mathrm{i}\hbar\sqrt{v_m v_n}\int_{C_\mathrm{R}} \mathrm{d}y \times \int_{C_\mathrm{L}} \mathrm{d}y'\phi_{m,\sigma'}(y)\,G^\mathrm{R}(\mathbf{x},\mathbf{x}')\,\phi_{n,\sigma}(y')\,, \tag{5}$$

$$r_{m,\sigma';n,\sigma} = \delta_{mn}\delta_{\sigma'\sigma} - \mathrm{i}\hbar\sqrt{v_m v_n}\int_{C_\mathrm{L}} \mathrm{d}y \times \int_{C_\mathrm{L}} \mathrm{d}y'\phi_{m,\sigma'}(y)\,G^\mathrm{R}(\mathbf{x},\mathbf{x}')\,\phi_{n,\sigma}(y')\,, \tag{6}$$

where v_n denotes the velocity of channel n and the integration runs over the cross-section $C_\mathrm{L}(C_\mathrm{R})$ of the left (right) lead. Equations (5) and (6) are valid only for leads without magnetic fields and no spin-orbit interaction. Baranger and Stone [35] have extended the formalism to also account for arbitrary magnetic fields in the leads, and their description can also be applied to finite spin-orbit interaction.

Tight-binding Representation of the Hamiltonian

Except for very simple geometries, the scattering problem cannot be solved analytically. Therefore, the use of computers for a numerical solution of the scattering problem is very often the method of choice. However, the related stationary Schrödinger equation $H\psi = E\psi$ is a differential equation with continuous degrees of freedom that are difficult to treat on a computer. In general, a numerical solution is thus only attempted within a *discrete* basis set which converts the differential equation into a matrix equation.

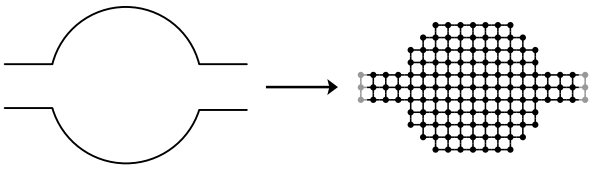

Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 4
Discretizing a continuous region on a square grid

The method of finite differences is a very simple and yet powerful way to introduce such a discrete basis set and has been applied to the Schrödinger equation already as early as 1934 [36,37] (for an introduction see [1,33]). Here we illustrate, as an example, the application of the method for a simple one-dimensional effective mass Hamiltonian including a potential

$$H = -\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + V(x)\,. \tag{7}$$

In the method of finite differences, space is approximated by a grid of discrete lattice points spaced equidistantly with lattice constant a. For the 2D case, usually a square grid is used as depicted in Fig. 4. Using the Taylor expansion of the wave function ψ we can write

$$\begin{aligned}\psi(x+a) = \psi(x) + \psi'(x)a + \frac{1}{2}\psi''(x)a^2 \\ + \frac{1}{6}\psi^{(3)}(x)a^3 + \frac{1}{24}\psi^{(4)}(x)a^4 + \ldots\,,\end{aligned} \tag{8}$$

$$\begin{aligned}\psi(x-a) = \psi(x) - \psi'(x)a + \frac{1}{2}\psi''(x)a^2 \\ - \frac{1}{6}\psi^{(3)}(x)a^3 + \frac{1}{24}\psi^{(4)}(x)a^4 + \ldots\,.\end{aligned} \tag{9}$$

Adding Eq. (8) and (9), we arrive at an expression for the second derivative of the wave function in terms of values of the wave function on the grid,

$$\begin{aligned}&\frac{\mathrm{d}^2}{\mathrm{d}x^2}\psi(x) \\ &\quad = \frac{1}{a^2}\left(\psi(x+a) + \psi(x-a) - 2\psi(x)\right) + \mathcal{O}(a^2)\,,\end{aligned} \tag{10}$$

valid up to second order in the lattice spacing a. The differential equation

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\psi(x) + V(x)\psi(x) = E\psi(x) \tag{11}$$

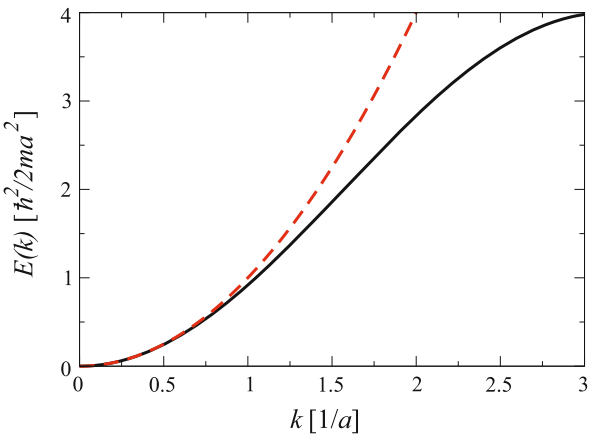

Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 5
Energy spectrum *E*(*k*) for the continuous Schrödinger equation (*red dashed line*) and the tight-binding approximation (*black solid line*)

is thus replaced by a set of difference equations

$$\begin{aligned}-\frac{\hbar^2}{2ma^2}\left(\psi(x_{i+1}) + \psi(x_{i-1}) - 2\psi(x_i)\right) + V(x)\psi(x_i) \\ = E\psi(x_i)\,,\end{aligned} \tag{12}$$

where $x_{i\pm1} = x_i \pm a$, yielding the tight-binding representation of the Hamiltonian:

$$\begin{aligned}H = \sum_{x_i} -\frac{\hbar^2}{2ma^2}\left(|x_i\rangle\langle x_{i+1}| + \text{h.c.}\right) \\ + \left(2\frac{\hbar^2}{2ma^2} + V(x_i)\right)|x_i\rangle\langle x_i|\,.\end{aligned} \tag{13}$$

Here, $|x_i\rangle$ denotes a state localized at grid point x_i.

In principle, the quality of the finite differences approximation can be improved up to a desired precision by reducing the lattice spacing a. However, since this leads to a larger problem size, the minimum lattice spacing achievable is set by the available computing time and memory. Thus, one must keep an eye on the validity of the finite differences approximation. In Fig. 5 we show the energy spectrum $E(k)$ for the continuous one-dimensional Schrödinger equation and the tight-binding (finite differences) approximation. The tight-binding approximation only holds for $ka \ll 1$ and $E \ll \hbar^2/(2ma^2)$, and it does not make sense to consider the whole energy spectrum given by the tight-binding band width. The method of finite differences presented here can straight-forwardly be applied to more complex Hamiltonians, including for example spin-orbit interactions [17], and will be later

used to calculate transport phenomena including spin in Sect. "Spin Filtering in Nanostructures" and "Pure Spin Current Generation".

A tight-binding representation of the Hamiltonian can also be obtained by applying the finite element method [38]. Furthermore tight-binding Hamiltonians are also used in treatments beyond the effective mass approximation, such as from atomic orbitals in empirical tight-binding models [39,40,41], or from orbitals of the Kohn–Sham equations within DFT [42,43,44].

Numerical Algorithms

Within the tight-binding approximation the Hamiltonian H can be represented by a matrix, even though this matrix is still infinite as the leads are infinitely long. However, the infinite matrix problem can be reduced to a finite problem by partitioning the system into three isolated parts: left lead, scattering region, and right lead. The Hamiltonian then reads

$$H = \begin{pmatrix} H_{\mathrm{L}} & V_{\mathrm{LS}} & 0 \\ V_{\mathrm{SL}} & H_{\mathrm{S}} & V_{\mathrm{SR}} \\ 0 & V_{\mathrm{RS}} & H_{\mathrm{R}} \end{pmatrix} , \tag{14}$$

where $H_{\mathrm{L(R)}}$ is the (infinite) Hamiltonian of the left (right) lead, H_S is the Hamiltonian of the scattering region and of finite size. $V_{\mathrm{SL}} = V_{\mathrm{LS}}^{+}$ and $V_{\mathrm{SR}} = V_{\mathrm{RS}}^{+}$ represent the coupling of the left and right lead to the scattering region. Since the leads are always chosen such that asymptotic eigenstates can be defined, the Hamiltonian of the isolated leads must contain some periodicity that facilitates calculating their Greens functions $g^{\mathrm{R}}_{\mathrm{L,R}}$. This can be done analytically for simple systems [1,33], for more complex situations the Greens function can be calculated numerically either by iteration [45,46] or semianalytical formulas [40,44,47]. Introducing the retarded self-energy $\Sigma^{\mathrm{R}} = \sum_{i=L,R} V_{\mathrm{S}i}\, g^{\mathrm{R}}_{i}\, V_{i\mathrm{S}}$, [1,33] the Greens function G_{S} of the scattering region is given by

$$G^{\mathrm{R}}_{\mathrm{S}} = \left(E - H - \Sigma^{\mathrm{R}}\right)^{-1} , \tag{15}$$

reminiscent of Eq. (4) but with an effective Hamiltonian $H + \Sigma^{\mathrm{R}}$.

The original infinite-dimensional problem has thus been reduced to a finite size matrix problem that can, in principle, be solved straight-forwardly on a computer. However, for any but rather small problems, the computational task of the direct inversion in Eq. (15) is prohibitive. Therefore, many algorithms make use of the *sparsity* of the Hamiltonian matrix in tight-binding representation – in particular they employ the property that this matrix can be written in block-tridiagonal form

$$H = \begin{pmatrix}
\ddots & & & & & & & & \\
H_{\mathrm{L}} & V_{\mathrm{L}} & & & & & & & \\
V_{\mathrm{L}}^{\dagger} & H_{\mathrm{L}} & H_{01} & & & \ddots & & & \\
 & H_{10} & H_{11} & H_{12} & & & 0 & & \\
 & & H_{21} & H_{22} & H_{23} & & & \ddots & \\
 & & & H_{32} & \ddots & & & & \\
 & \ddots & & & \ddots & H_{N-1N} & & & \\
 & & 0 & & H_{NN-1} & H_{NN} & H_{NN+1} & & \\
 & & & \ddots & & H_{N+1N} & H_{\mathrm{R}} & V_{\mathrm{R}} & \\
 & & & & & & V_{\mathrm{R}}^{\dagger} & H_{\mathrm{R}} & \\
 & & & & & & & & \ddots
\end{pmatrix} . \tag{16}$$

Here the index L(R) denotes the blocks in the left (right) lead, $1 \ldots N$ the blocks within the scattering region, and 0, $(N+1)$ the first block in the left (right) lead. Such a form arises naturally in the method of finite differences, when grid points are grouped into vertical slices according to their x-coordinates, as shown in Fig. 6, but also applies to any other sparse tight-binding Hamiltonians.

The block-tridiagonal form of the Hamiltonian is the foundation of several quantum transport algorithms, together with the fact that, according to Eq. (5) and (6), only the blocks $G^{\mathrm{R}}_{N+1\,0}$ and G^{R}_{00} of the Greens function G^{R} are needed for the calculation of transmission and reflection probabilities. The transfer matrix approach applies naturally to block-tridiagonal Hamiltonians, but becomes unstable for larger systems. However, a stabilized version has been developed by Usuki et al. [48,49]. In the dec-

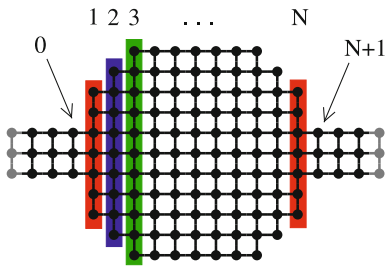

Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 6

Block-tridiagonal matrix form arising in the method of finite differences. Grid points with the same *x*-coordinate are placed into the same block

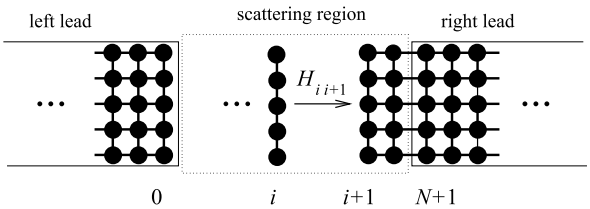

Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 7
Schematic view of the recursive Greens function algorithm: the system is built up adding block by block

imation technique [50,51], the Hamiltonian of the scattering region is replaced by an effective Hamiltonian between the two leads by eliminating internal degrees of freedom. The contact block reduction method [52] calculates the full Greens function of the system using a limited set of eigenstates. The recursive Greens function (RGF) technique [53,54,55] uses Dyson's equation to build up the system's Greens function block by block. It has also been adapted to Hall geometries with four terminals [56] and to calculate non-equilibrium densities [57,58]. Furthermore, the RGF algorithm has been formulated to be suitable for parallel computing [59], and the modular recursive Greens function (MRGF) method is an extension to take advantage of special geometries, such as circles or rectangles [60,61].

Here we give a simple form of the RGF algorithm as described in [33,55]. The RGF method makes use of Dyson's equation

$$G = G_0 + G_0 V G \tag{17}$$

to add successively blocks to the system, as depicted in Fig. 7. Let $G^{\mathrm{R},(i)}$ denote the Greens function for the system containing all blocks $\geq i$. Then, at energy E, the Greens function $G^{\mathrm{R},(i)}$ is related to $G^{\mathrm{R},(i+1)}$ by

$$G^{\mathrm{R},(i)}_{i\,i} = \left(E - H_{i\,i} - H_{i\,i+1}\, G^{\mathrm{R},(i+1)}_{i+1\,i+1}\, H_{i+1\,i}\right)^{-1} \tag{18}$$

and

$$G^{\mathrm{R},(i)}_{N+1\,i} = G^{\mathrm{R},(i+1)}_{N+1\,i+1}\, H_{i+1\,i}\, G^{\mathrm{R},(i)}_{i\,i}\,. \tag{19}$$

Starting from $G^{\mathrm{R},(N+1)}_{N+1\,N+1} = g^{\mathrm{R}}_{\mathrm{R}}$, the surface Greens function of the right lead, N slices are added recursively, until $G^{\mathrm{R},(1)}$ has been calculated. The blocks of the Greens function of the full system necessary for transport are then given by

$$G^{\mathrm{R}}_{00} = \left(\left(g^{\mathrm{R}}_{\mathrm{L}}\right)^{-1} - H_{01}\, G^{\mathrm{R},(1)}_{11}\, H_{10}\right)^{-1} \tag{20}$$

and

$$G^{\mathrm{R}}_{N+1\,0} = G^{\mathrm{R},(1)}_{N+1\,1}\, H_{10}\, G^{\mathrm{R}}_{00}\,, \tag{21}$$

where $g^{\mathrm{R}}_{\mathrm{L}}$ is the surface Greens function of the left lead.

Each step of the algorithm performs inversions and matrix multiplications with matrices of size M_i. Since the computational complexity of matrix inversion and multiplications scales as ${M_i}^3$, the complexity of the RGF algorithm is $\propto \sum_{i=0}^{N+1} M_i^3$. Thus, it scales linearly with the "length" N, and cubically with the "width" M_i of the system.

While for certain geometries the RGF algorithm cannot compete with more specialized algorithms such as MRGF, it is very versatile and easily adapted to many situations, and is thus our method of choice. In the next section we will discuss matrix reordering techniques that improve the runtime of the RGF algorithm considerably and allow the treatment of arbitrary systems.

Matrix Reordering Strategies for Quantum Transport

Graph-theoretical Approaches to Matrix Reordering

As shown in the previous section, the structure of a Hamiltonian matrix H does influence the runtime of the RGF algorithm. Thus, the runtime of the algorithm can potentially be improved by reordering the matrix with a permutation P,

$$H' = P\, H\, P^{-1}\,. \tag{22}$$

At first glance, such an effort may seem pointless: For example, the block-tridiagonal structure naturally associated with a finite difference grid (as discussed in Sect. "Numerical Quantum Transport") leads to a sparse matrix with a small bandwidth, as shown in Fig. 8. However, as we show below, the choice of a suitable permutation P can still lead to a significant speed-up of the RGF algorithm.

For this we define a weight $w(H)$ associated with a given matrix H as

$$w(H) = \sum_{i=0}^{N+1} M_i^3 \quad \text{where } M_i \text{ is the size of block } H_{ii}\,. \tag{23}$$

Optimizing the matrix for the RGF algorithm is then equivalent to minimizing the weight $w(H)$. Since $\sum_{i=0}^{N+1} M_i = N_{\mathrm{grid}}$, where N_{grid} is the total number of grid points, $w(H)$ is minimal, if all M_i are equal, and $M_i = N_{\mathrm{grid}}/(N+2)$. Therefore, a matrix tends to have

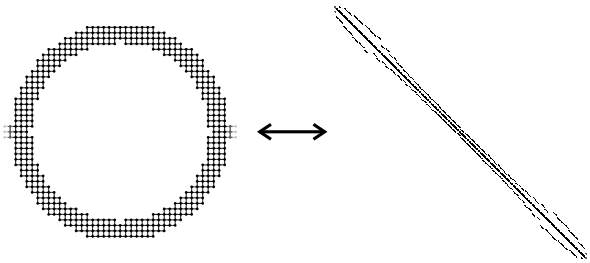

Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 8

Example of a sparsity structure of a matrix associated with a finite difference grid. *Black dots* mark non-zero entries. The picture of the finite difference grid can also be interpreted as a graphical representation of a graph

small weight, if the number N of blocks is large, and all blocks are equally sized. The reordering problem of the matrix H is thus summarized as follows:

Matrix reordering problem: Find a reordered matrix H' such, that

1. H'_{00} and H'_{N+1N+1} are blocks given by the left and right leads (as required by the RGF algorithm),
2. H' is block-tridiagonal ($H'_{ij} \neq 0$, iff $j = i+1, i, i-1$),
3. the number N of blocks is as large as possible, and all blocks are equally sized.

These requirements define the optimization problem of reordering the matrix H. Usually, in such optimization problems finding the best solution deterministically is prohibitively expensive, and one has to resort to heuristic strategies.

In order to do that, we reformulate our matrix problem using concepts from graph theory. A *graph* G is a pair $G = (V, E)$, where V is a set of *vertices* i, and E a set of pairs of vertices $(i, j) \in V \times V$. Such a pair is called an *edge*. A graph is called *undirected*, if for every edge $(i, j) \in E$ also $(j, i) \in E$. Two vertices i and j are called *adjacent*, if $(i, j) \in E$. A graph can be visualized by drawing dots for each vertex i and lines connecting these dots for every edge (i, j). It should be noted, that all the pictures of finite difference grids shown so far can directly be interpreted as graphs . There is a natural one-to-one correspondence between graphs and the structure of sparse matrices. For a given $n \times n$ matrix H, we define a graph $G = (V, E)$ with $V = \{1, \ldots, n\}$ and $(i, j) \in E$ iff $H_{ij} \neq 0$. Thus, the symmetric zero–nonzero structure of Hermitian matrices, as considered in quantum transport, leads to associated undirected graphs. An example of such a correspondence between a graph and a matrix has already been shown in Fig. 8. With respect to matrices, graphs are also very convenient for storing and handling sparse matrix data structures on a computer.

In terms of graph theory, matrix reordering corresponds to renumbering the vertices of a graph. Since we are only interested in reordering the matrix in terms of matrix blocks (the order within a block should not matter too much), we define a *partitioning* of G as a set $\{V_i\}$ of disjoint subsets $V_i \subset V$ such that $\bigcup_i V_i = V$ and $V_i \cap V_j = \emptyset$ for $i \neq j$. Using these concepts, we can now reformulate the original matrix reordering problem into a graph partitioning problem:

Graph partitioning problem: Find a partitioning $\{V_0, \ldots, V_{N+1}\}$ of G such that:

1. V_0 and V_{N+1} contain the vertices belonging to left and right leads,
2. there are edges between V_i and V_j iff $j = i+1, i, i-1$ (block-tridiagonality),
3. the number of sets V_i is as large as possible and all sets V_i have the same cardinality. Such a partitioning with all V_i equally sized is called *balanced*.

A partitioning that meets requirement 2 is called a *level set* with levels V_i and appears commonly as an intermediate step in algorithms for bandwidth reduction of a matrix [62,63,64,65]. These algorithms seek to find a level set of minimal width, i. e. $\max_{i=0\ldots N+1} |V_i|$ as small as possible which is equivalent to requirement 3. The major difference between our graph partitioning problem and the bandwidth reduction algorithms is requirement 1: In our case the start and end blocks are given by the geometry of the system, whereas in the bandwidth reduction methods these can be chosen freely. The implications of this difference will be discussed below.

The bandwidth reduction algorithms start with the observation that a *breadth first search* (BFS) starting from any vertex in the graph creates a level set: In our situation, the BFS starts from level V_0 of the left lead. Then, successively for $i = 0, 1, 2, \ldots$, all vertices adjacent to V_i that have not been assigned to a level yet are placed in V_{i+1}. This construction ensures that each level V_i only has edges connecting to vertices in $V_{i+1,i,i-1}$ and thus ensures block-tridiagonality. The search stops, once a vertex adjacent to the right lead is encountered, and all unassigned vertices are placed into the last level V_N. The number of BFS steps determines therefore the maximum number N of blocks and is related to the minimum distance between left and right lead in the graph. However, this construction of the level set also suffers from a serious problem: Depending on the distance between the leads, the last level V_N can potentially contain a large number of vertices leading to a very unbalanced partitioning. In the bandwidth reduc-

tion methods, the first and the last vertex are chosen to have (to a good approximation) maximum distance, and thus this problem does not occur there. Hence, conventional bandwidth reduction algorithms can only be applied to quantum transport problems, if the leads are —in terms of the underlying graph —furthest apart.

In this study, we will consider two reordering algorithms: First, the Gibbs–Poole–Stockmeyer (GPS) algorithm [65], a widely used bandwidth reduction algorithm. The GPS algorithm combines the level sets originating from a BFS from both left and right lead to give an optimized level set. Due to the limitations discussed above, it can only be used efficiently for systems with leads far apart. To overcome this difficulty partly, we also propose a second algorithm, later referred to as *BFS partitioning*: The system is bisected recursively by means of a simultaneous BFS from left and right leads. In a bisection process, vertices that are closer (as given by the number of BFS steps) to the left (right) lead are placed into the left (right) level. The resulting two levels are then further bisected recursively until the final level set has been constructed. This algorithm tries to avoid an unbalanced partitioning, as every step tries to create a balanced bisection. We have found this global approach —as opposed to the local approach in the BFS —to yield balanced partitionings for systems where there are only few local minima in the weight $w(H)$. For general systems, a more sophisticated method should be used [66].

Obviously, reordering the matrix will only improve the runtime of the RGF algorithm, if the overhead of the reordering process is small compared to the actual transport calculation. Because of this reason, applying general optimization algorithms to the matrix reordering problem is not an option. Instead, heuristics designed for graph problems give much better performance. The GPS algorithm scales linearly with the number of edges $|E|$, and since in a tight-binding representation $|E| \propto N_{\text{grid}}$, its computational complexity is $\mathcal{O}(N_{\text{grid}})$, whereas the BFS partitioning algorithm scales as $\mathcal{O}(N_{\text{grid}} \log N_{\text{grid}})$. In any case, the scaling is much more favorable than that of the RGF algorithm, $\propto \sum_{i=0}^{N+1} M_i^3$, so that for systems of typical size the overhead of the reordering process becomes negligible, as we will demonstrate in the next section.

Example: Ring Geometry

In order to demonstrate the performance of the algorithms discussed above, we consider their application to a ring geometry in finite difference approximation. The performance of the RGF algorithm after matrix reordering is compared with the performance using the ordering that arises naturally in finite difference grids as shown in Fig. 6 (in the remainder referred to as *natural partitioning*).

In Fig. 9a–d we show four different approaches for calculations in a ring geometry. A ring can be treated as a circular cavity (see Fig. 9a), with a large potential on the lattice points inside the inner ring diameter. This approach is easier to implement than a real ring but less efficient, as a large additional number of lattice points enters the calculation. However, this approach has been used frequently in the past, and therefore we also consider its performance. Transport calculations in a real ring require somewhat more bookkeeping because of the non-trivial geometry, but can be easily done describing the grid as a graph. For the circular cavity, we only consider natural partitioning, as shown in Fig. 9a, for the ring we consider natural partitioning (Fig. 9b), GPS partitioning (Fig. 9c) and BFS partitioning (Fig. 9d).

The partitionings in Fig. 9c and d are dramatically different from the natural partitioning. The levels align mainly in vertical or diagonal directions as these are the preferred directions in the square lattice. The number of levels is increased with respect to the natural partitioning, as the distance between both leads is much larger than the number of vertical slices, leading to a larger number of blocks in the block-tridiagonal matrix, as can be seen from Fig. 9e. Both GPS and BFS partitioning lead to a drastically reduced block size with respect to the natural partitioning and the result is rather balanced. Though the actual partitionings in Fig. 9c and d look quite different, with respect to minimizing the weight $w(H)$ they are equally good. The BFS partitioner conforms to the geometric structure, as it puts vertices in levels according to their distance from the leads. Except for a small number of blocks in the beginning and the end of the block-tridiagonal matrix (see Fig. 9e), the GPS partitioner leads to an equally balanced structure, although the partitioning looks quite different. The GPS partitioner works well in this case, as the leads are almost at maximum distance in this ring structure.

We now apply the recursive Greens function algorithm from Sect. "Numerical Quantum Transport" to the different partitionings. In Fig. 10a we show the runtime of the algorithm as a function of the number of lattice points in the leads, which is also the number of lattice points across the arms of the ring. Note that in all cases the runtime includes both the time spent in calculating the matrix reordering and the time spent in the actual transport calculation.

The runtime scales similarly in all cases, as this is the scaling of the RGF algorithm. Nevertheless, the runtimes of the different approaches can differ by a factor that is significant. As expected, the circular cavity is slowest, due

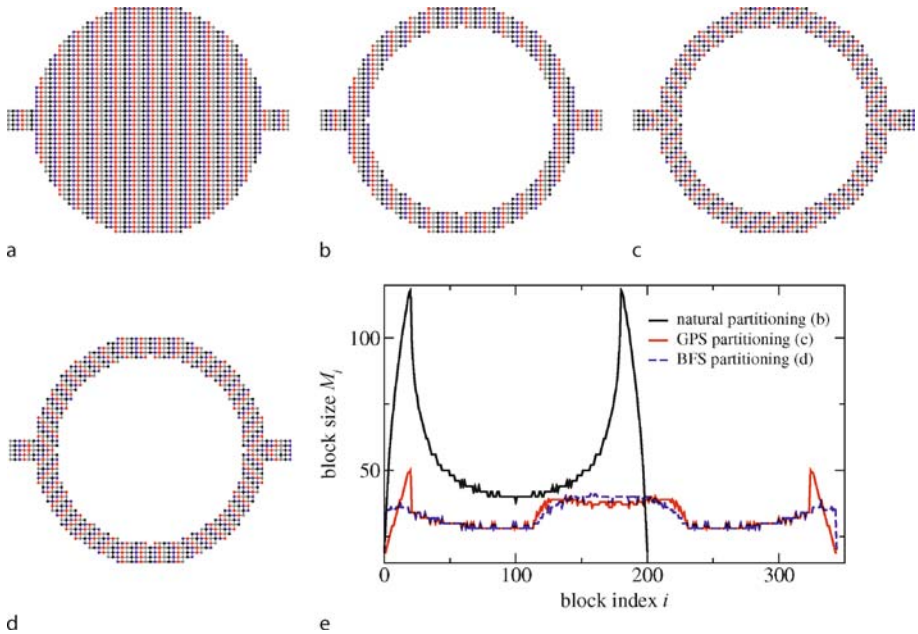

Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 9

Partitioning of the different systems considered in the performance study. a Circular cavity in natural partitioning, where the ring structure is enforced by a potential term in the middle of the cavity, b ring in natural partitioning, c ring in GPS partitioning and d ring in BFS partitioning. In a–d, vertices belonging to the same level in the partitioning are marked with the *same color*, different levels are marked in *alternating colors*. Note that, in order to reveal the partitioning structure more clearly, the grids shown here are much smaller than in a typical calculation. e Size M_i of the blocks H_{ii} in the block-tridiagonal matrix for a ring with 20 lattice points in the leads for natural, GPS and BFS partitioning (the circular cavity is not shown here)

to the extra number of lattice points. GPS and BFS partitionings lead to a rather similar performance that is *significantly* better than the ring in natural partitioning. It outperforms the circular cavity even by a factor of up to 100. In the remainder, we examine the performances of the ring for different partitionings in more detail and leave out the circular cavity. In Fig. 10b we show the relative performance gain of GPS and BFS partitionings over the natural partitioning. Except for the smallest of systems that are too small to be useful in practice, the performance of GPS and BFS partitionings is better than the natural partitionings. Even for moderately sized systems the performance gain through the matrix reordering is approximately 3, with the BFS partitioner being slightly better than the GPS partitioner. In Fig. 10b we also show estimates of the performance gain calculated from the weights $w(H)$ of the different partitionings. These estimates predict a performance gain of approximately 4. For small system sizes, we see deviations from these estimates because of the overhead of the partitioning process, for larger system sizes we almost reach the estimated value. On modern computer architectures, runtime does not only depend on the number of arithmetic operations [67], and thus we do not achieve the full theoretical potential of the reordering, yet still significant improvements. In Fig. 10c we show the fraction of time spent in calculating the matrix reordering with respect to the total computation time, and as expected the overhead becomes negligible already for moderately sized systems. It should be noted that in actual calculations the partitioning is commonly only done once, and transport calculations can be done repeatedly with the same partitioning: Usually one is interested in transport properties depending on some parameters, and these generally do not change the *structure* of the Hamiltonian matrix but only the values of the respective entries. In this case, the partitioning overhead becomes even more irrelevant.

It should be emphasized, that for all the situations, the *same* transport code was used. In addition to the significant speedup through the graph techniques considered here, the abstraction of the system through graph structures allows for very generic transport codes. This is an additional strength of this approach, as the well-established RGF algorithm is thus readily applied to arbitrary systems that would require special treatment otherwise, such as a scattering region with perpendicular leads, as depicted in Fig. 11.

Of course, for any system, an algorithm designed for that special system will probably outperform generic ap-

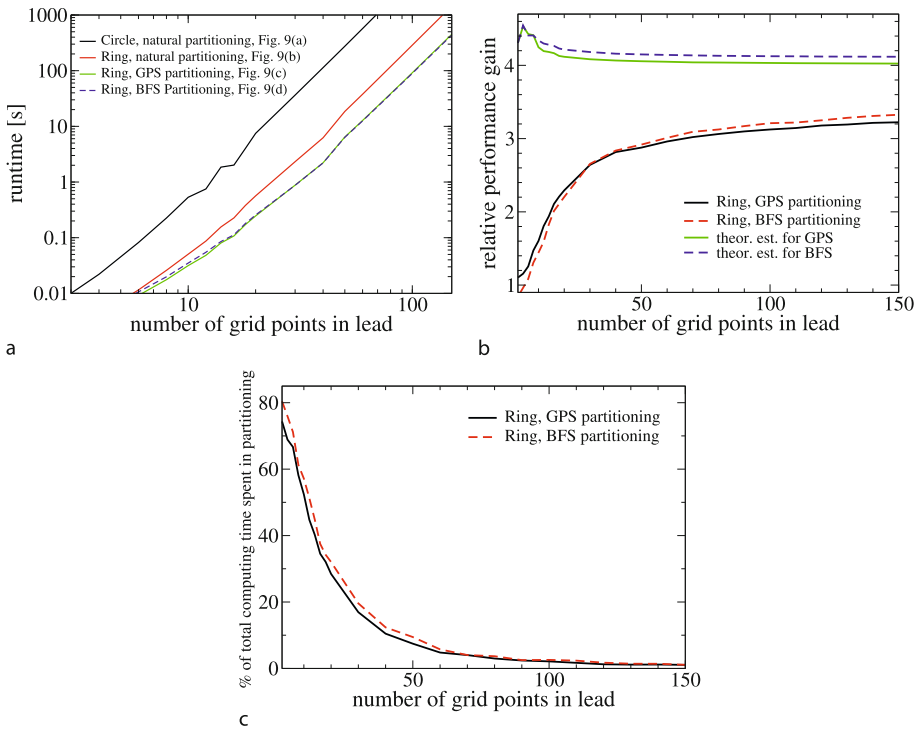

Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 10

Performance of the partitionings (for rings, Fig. 9): a runtime of the different systems as a function of the number of lattice points in the leads, i. e. as a function of system size. The calculations were performed on a Core2Duo T5500 processor and 1GB of memory, and the runtime includes both partitioning overhead and transport calculation. b Performance gain of GPS and BFS partitionings with respect to natural partitioning and c overhead of the matrix reordering as a function of system size

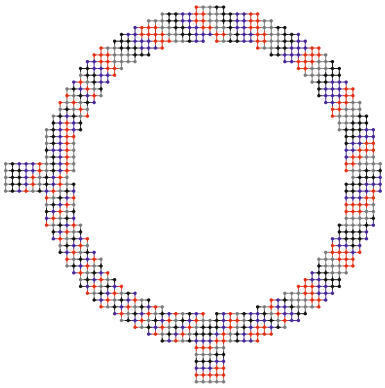

Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 11

BFS partitioning of a ring with perpendicular leads. Note that in this case the two leads are closer together and thus the number of blocks is reduced. Therefore the block sizes tend to be larger than in the previous examples

proaches. However, the combination of matrix reordering and RGF algorithm can be applied to arbitrary systems and is thus probably the most versatile transport approach. In addition, all the algorithms relying on the block-tridiagonal structure of the Hamiltonian mentioned in Sect. "Numerical Quantum Transport" can benefit from these matrix reordering strategies.

Modern transport calculations tend to be more and more complex and time-consuming. For example, an electronic calculation including the electron spin typically takes $2^3 = 8$ times longer than a calculation on spinless electrons. Furthermore, calculations including disorder involve averages over many disorder configurations. Any increase in computation speed is therefore beneficial, and the added versatility through matrix reordering methods makes these techniques even more useful.

Spin Filtering in Nanostructures

In the introductory example in Sect. "Introduction" it was shown how to realize systems that work as spin switches making use of the interference of wavefunctions propagating clockwise and counterclockwise in Aharonov–Bohm rings with SO-interaction. However, several other device proposals have been put forward utilizing different concepts in order to achieve spin filtering in mesoscopic systems.

A very prominent category is transverse focusing of ballistic electrons/holes in two dimensional electron/hole gases (2DEGs/2DHGs) [68]. In materials that exhibit SO interaction the cyclotron radius of ballistic electrons/holes due to a magnetic field perpendicular to the 2DEG/2DHG depends on the spin state of the charge carriers. Therefore, it is possible to filter out either spin-up or spin-down electrons by an appropriate arrangement of quantum point contacts and a proper choice of the perpendicular magnetic field. Apart from several theoretical treatments of this topic [69,70,71], spin filtering by transverse focusing has already been experimentally verified in a GaAs-based 2DHG [72]. In related experiments, spin-polarized currents in 2DEGs/2DHGs were detected by a setup consisting of point contacts and making use of transverse focusing of the charge carriers [73]. With such a detector it was possible to confirm the presence of spin-polarized currents emitted from mesoscopic quantum dots utilizing quantum interference at high in-plane magnetic fields [74] and from quantum point contacts, which were either made spin sensitive with high in-plane magnetic fields [73] or showed a pronounced "0.7-anomaly" [75].

A further appealing approach to filter spins are three terminal structures, that act as mesoscopic Stern–Gerlach type spin filters [76]. In these devices one of the leads injects spin-unpolarized current and, after passing a region where the spin-degeneracy is lifted, oppositely-polarized output currents exit through the other two leads. This separation of up and down spins can be accomplished, e. g., by utilizing Rashba SO interaction [77,78,79].

However, three terminals are not required to create spin polarized currents. Many devices, as e. g. the AB-ring presented in Sect. "Introduction", typically rely on two terminals only, where transport through tailored geometries with SO interaction [80,81,82], magnetic fields [83,84,85,86] or a combination of both [87] can result in a significant spin filter effect.

As a representative example for the methods mentioned above, in the present section we present spin filtering due to Rashba SO interaction in quantum wires connected to two terminals. We consider a quantum wire in y-direction realized in a 2DEG in the (x, y) plane connected to two nonmagnetic leads. The Hamiltonian of the system, with spatially dependent Rashba SO interaction is given by

$$H_0 = \frac{\hat{p}^2}{2m^*} + \frac{\alpha(x)}{2\hbar}(\hat{\sigma}_x \hat{p}_y - \hat{\sigma}_y \hat{p}_x) + (\hat{\sigma}_x \hat{p}_y - \hat{\sigma}_y \hat{p}_x)\frac{\alpha(x)}{2\hbar} + V(x, y) + U_\mathrm{B}(x, y) \,. \tag{24}$$

Here $V(x, y)$ is the lateral transverse confinement potential forming the quantum wire, while $U_\mathrm{B}(x, y)$ is an additional electrostatic potential in the system, e. g., realized by gate voltages. Furthermore, $\hat{\sigma}_i$ denote the Pauli spin matrices, and m^* is the effective electron mass of the semiconducting material. We consider a constant Rashba SO interaction strength ($= \alpha_\mathrm{C}$) in the central region of the system, which is connected to two semi-infinite leads on opposite sides, where $\alpha(x)$ is chosen to be zero avoiding ambiguities in the definition of spin current that arise for leads with SO interaction [88]. In order not to introduce additional effects due to an abrupt jump in the SO coupling strength, the parameter $\alpha(x)$ is changed sufficiently smooth from zero to α_C between the leads and the central region. For the numerical calculations presented in the next two sections the Hamiltonian (24) is discretized as shown in Sect. "Numerical Quantum Transport" yielding a tight-binding Hamiltonian on a square grid.

In the rest of this section we investigate the transport properties in the linear response regime due to an infinitesimal bias voltage δU applied between the left and right contact. The charge (C) and spin (S) current in the Landauer–Büttiker formalism are then given by

$$I_\mathrm{C/S} = G_\mathrm{C/S}\, T_\mathrm{C/S}\, \delta U \,,$$

where $G_\mathrm{C} = \mathrm{e}^2/h$ and $G_\mathrm{S} = \mathrm{e}/4\pi$ are the conductance quanta of charge and spin respectively. Since, opposite to charge current, the spin current can be different in the right and left lead [85], here we choose to evaluate the respective currents in the right lead. Then the transmission probabilities $T_\mathrm{C/S}$ at the Fermi energy are given by

$$\begin{aligned} T_\mathrm{C} &= T_{+,+} + T_{+,-} + T_{-,+} + T_{-,-} \,, \\ T_\mathrm{S} &= T_{+,+} + T_{+,-} - T_{-,+} - T_{-,-} \,, \end{aligned} \tag{25}$$

where $T_{\sigma,\sigma'} = \sum_{n\in\mathrm{R},n'\in\mathrm{L}} |S_{n,\sigma;n',\sigma'}|^2$ is the probability for an electron, injected into the left (L) lead with spin state σ' to be transmitted to the right (R) lead and end up there in spin state σ. In the present and the following section we fix the spin quantization axis to the y-axis. The scattering matrix elements $S_{n,\sigma;n',\sigma'}$ and therefore also the spin resolved transmission probabilities $T_{\sigma,\sigma'}$ are evaluated using the recursive Greens function algorithm presented in Sect. "Numerical Quantum Transport" and "Matrix Reordering Strategies for Quantum Transport".

One general feature of Landauer transport in a quantum wire with SO interaction and non-magnetic leads is the absence of spin-polarized currents in a lead that supports only a single transversal mode. This property can be derived from the invariance of the system under the time-

reversal operator $\hat{\mathcal{T}} = -\mathrm{i}\hat{C}\sigma_y$ [89], where $\hat{C}$ is the operator of complex conjugation. For a perfect quantum wire which is translationally invariant in the direction of transport all occupied transversal subbands transmit without reflection and spin polarization is not possible due to SO interaction. However, if backreflection, caused by deviations from a perfect quantum wire, is present, it is possible to observe spin-polarized currents in leads with at least two transversal channels. There the typical mechanism responsible for spin polarization is the mixing of spins from different transversal subbands due to the SO interaction.

In Eq. (24) this translational invariance in y-direction is already broken by the spatially varying SO interaction $\alpha(x)$ even if the quantum wire was perfect in y-direction otherwise, i. e. $V(x, y) = V(y)$ and $U_\mathrm{B}(x, y) = U_\mathrm{B}(y)$. However, if the region where $\alpha(x)$ is turned on/off is sufficiently long, reflection due to the change of $\alpha(x)$ is negligible.

There exist several device proposals relying on this mixing of spins from different transversal subbands due to x-dependent lateral confinement potentials $V(x, y)$ or other superimposed electrostatic potentials $U_\mathrm{B}(x, y)$. These device designs include, e. g., constrictions [80,90], lateral side pockets [81], or electrostatic barriers [82], to name only a few.

In most of those proposals, systems symmetric with respect to inversion of the x-coordinate were considered, i. e. $V(x, y) = V(-x, y)$, $U_\mathrm{B}(x, y) = U_\mathrm{B}(-x, y)$ and $\alpha(x) = \alpha(-x)$. Then the Hamiltonian (24) is left invariant upon application of the symmetry operation

$$\hat{\mathcal{P}} = -\mathrm{i}\,\hat{C}\hat{R}_x\hat{\sigma}_z\,, \tag{26}$$

where the operator $\hat{R}_x$ inverses the x-coordinate. The scattering wavefunctions inside the leads are changed by the operator $\hat{\mathcal{P}}$ in the following way: $\hat{R}_x$ exchanges the leads, i. e., a transversal mode index n is replaced by the corresponding mode index $\bar{n}$ in the other lead. The operator of complex conjugation transforms incoming (outgoing) states into outgoing (incoming) states with complex conjugated amplitude. Moreover, the combined effect of $\hat{C}\sigma_z$ is to rotate a spinor with the coordinates (θ, ϕ) on the Bloch sphere to the coordinates $(\theta, -\phi + \pi)$. Exploiting this invariance of the Hamiltonian one can derive the relation $S_{n,\sigma;n',\sigma'}(E) = S_{\bar{n}',\sigma';\bar{n},\sigma}(E)$ between the elements of the scattering matrix (see [85,89] for related expressions). This results in the equality of the spin flip transmissions $T_{+,-} = T_{-,+}$. Therefore, for those devices to be able to work as a spin filter, in view of Eq. (25) the spin conserving transmissions $T_{+,+}$ and $T_{-,-}$ need to be different.

As an example, in the following we consider Landauer transport through a Rashba SO quantum wire with a constriction. Similar calculations were carried out in [71,80] where it was shown, that this setup is able to produce a spin polarized current of sizeable quantity. It was conjectured, that the mechanism responsible for the spin polarization was the depletion of higher transversal modes of the wire and a subsequent spin dependent repopulation of those modes when traversing the constriction [80]. To experimentally observe the predicted spin polarization the use of a transverse electron focusing technique was suggested [71].

A typical grid (with lattice spacing a) used in the calculation for the symmetric point contact in a wire of width $W = 15a$ is shown in Fig. 12a, where the constriction of length $L_\mathrm{PC} = 10a$ is formed by hard-wall potentials:

$$V(x, y) = \begin{cases} 0 & \text{for } C(x) < y < 15a - C(x) \\ \infty & \text{else} \end{cases}$$

with

$$C(x) = \begin{cases} 2.05a\left(1 - \cos\left(\frac{2\pi\left(x + \frac{L_\mathrm{PC}}{2}\right)}{L_\mathrm{PC}}\right)\right) & \text{for } |x| < \frac{L_\mathrm{PC}}{2} \\ 0 & \text{otherwise}\,. \end{cases} \tag{27}$$

Additionally $U_\mathrm{B}(x, y)$ is set to zero. In Fig. 12c we present the relevant transmission probabilities with respect to the Fermi energy E for this system. There we observe that the total conductance T_C is reduced in comparison with that of a perfect quantum wire. In the latter case T_C exhibits sharp steps due to conductance quantization [91,92] which are washed out here due to tunneling processes through the constriction. Furthermore, for energies, where only a single transversal mode is supported in the quantum wire, the spin transmission vanishes as expected, i. e. $T_{+,+} = T_{-,-}$ for energies $\bar{E} \lesssim 0.175$. Also the relation $T_{+,-} = T_{-,+}$ is fulfilled as required by the symmetry of the setup. Finally, at energies where a new transversal mode opens up ($\bar{E}_2 \approx 0.18$, $\bar{E}_3 \approx 0.39$) dips in the transmission probabilities become apparent. Those dips can be explained by interference of localized states in the central region where $\alpha(x) = \alpha_\mathrm{C}$ and the scattering states in the quantum wire where $\alpha(x) = 0$ [93].

In order to study the influence of the adiabaticity of the constriction on the degree of spin polarization that can be extracted, in Fig. 13 we show T_S as a function of the length of the constriction L_PC for several values of α_C. For all of the curves we clearly observe an increase in spin transmission with increasing L_PC, in accordance with the mechanism suggested in [80]. There it was argued in the context of Landau–Zener transitions that the repopulation of higher transversal subbands will be more efficient for more

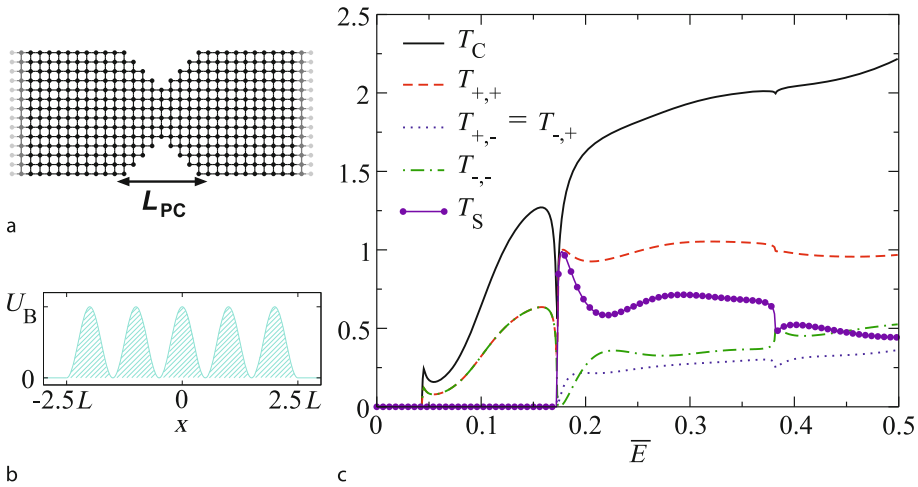

Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 12

Panel a The square lattice discretization of a quantum wire (width $W = 15a$) with a single constriction of length $L_{PC} = 10a$, see also Eq. (27). Panel b Periodic array of $N = 5$ electrostatic barriers with period length L and barrier height U_B. Panel c Charge T_C, spin T_S and spin resolved transmission probabilities $T_{\sigma,\sigma'}$ for the system depicted in panel a) and specified in the text at fixed SO interaction strength $\bar{\alpha} = [(m^* a)/\hbar^2]\,\alpha_C = 0.1$ with respect to the Fermi energy $\bar{E} = [(2m^* a^2)/\hbar^2]\,E$. The nth transversal mode in the wire opens at $\bar{E}_n = (\pi^2 n^2)/(W/a)^2$

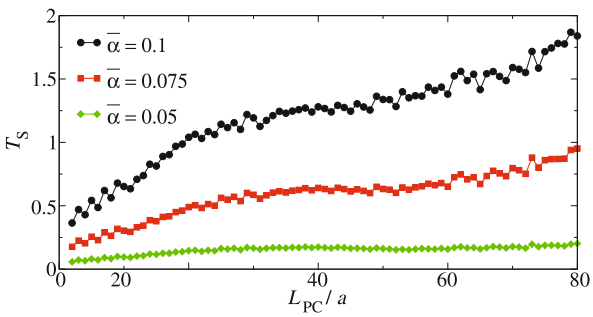

Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 13

Spin transmission probability T_S at fixed injection energy $\bar{E} = 0.25$ within the second transversal mode for three different SO interaction strengths $\bar{\alpha} = 0.1$ (*black dots*), $\bar{\alpha} = 0.075$ (*red squares*) and $\bar{\alpha} = 0.05$ (*green diamonds*) with respect to the length of the constriction L_{PC}

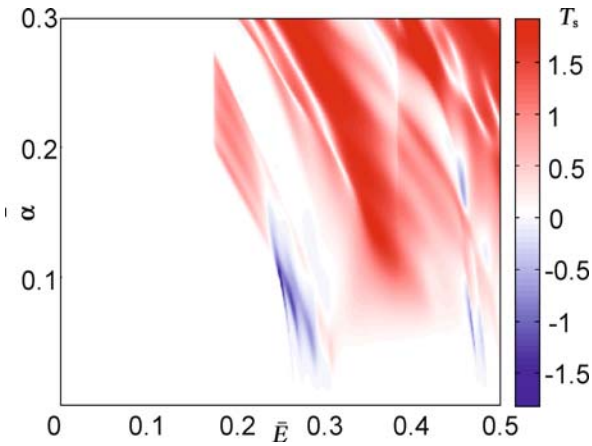

Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 14

Spin transmission probability T_S plotted as a function of the Fermi energy $\bar{E}$ and SO interaction strength $\bar{\alpha}$ for a quantum wire with $N = 5$ electrostatic barriers of length $L = 10a$ and height $\bar{U}_{Barr} = [(2m^* a^2)/\hbar^2]\,U_{Barr} = 0.2$

adiabatic constrictions or barriers, resulting in a higher degree of spin polarization. For $\bar{\alpha} = [(m^* a)/\hbar^2]\,\alpha_C = 0.1$ the spin transmission even approaches the highest possible value $T_S = 2$. One drawback of the presented system is its restriction to unidirectional spin polarization. In agreement with the model of [80], Fig. 12 and 13 give evidence that $T_S \geq 0$ for the parameter range considered. This limitation to output current with fixed spin polarization direction restricts the usability of the spin filter to special purposes. A possible way to circumvent this constraint is the use of a periodic array of electrostatic barriers [82], which we now briefly investigate. Figure 14 shows the spin transmission of a straight hard-wall quantum wire ($C(x) = 0$ in Eq. (27)) subject to $N = 5$ electrostatic barriers,

$$U_B(x, y) = U_B(x) = \begin{cases} \frac{U_{Barr}}{2}\left(1 - \cos\left(\frac{2\pi\left(x+\frac{L}{2}\right)}{L}\right)\right) & \text{for } -N\frac{L}{2} < x < N\frac{L}{2} \\ 0 & \text{otherwise}, \end{cases}$$

(as shown in Fig. 12b). The spin transmission is plotted as a function of the Fermi energy and the SO interaction strength. Again, one observes $T_S = 0$ for energies below $\bar{E}_2$. Furthermore, different regions in parameter space exhibit opposite sign of T_S, enabling to change the sign of the output polarization, e. g., by tuning α_c via gate voltages [25]. This additional functionality is due to resonant tunneling, which is absent for a quantum wire with only a single constriction or barrier.

Pure Spin Current Generation

In the preceding sections we focused on mesoscopic geometries that exhibit functionalities such as spin filtering or spin switching when applying an external bias between the contacts in the system. In the case of a static dc bias in a two terminal geometry those spin currents arise due to different currents of spin-up and spin-down electrons flowing into the direction of the contact with the lower chemical potential. However, over the last few years a growing number of device proposals has been put forward, that exhibit the interesting feature of pure spin current generation, i. e. spin currents in the absence of net charge transport. This intriguing case appears, when the direction of motion of spin-up electrons is opposite to the direction of motion of spin-down electrons and both currents equal in absolute magnitude.

Among the devices sharing the prospect of creating pure mesoscopic spin currents are systems with more than two terminals realized in 2DEGs exhibiting Rashba [2,3] and/or the Dresselhaus [94] SO coupling. Here a charge current is induced in this multiterminal structure by the application of bias voltages between the different contacts of the system. However, if the voltage of one of the leads is adjusted to make it work as a voltage probe, no charge current passes this lead but a pure spin current can appear owing to the SO coupling present in the system. The basic working principle behind these devices is the so-called mesoscopic spin Hall effect [95], a version of the intrinsic spin Hall effect [27], where the typical system size does not exceed the phase coherence length of the electrons. The SO interaction leads to different transport dynamics for different spin species, which can be used to extract the desired pure spin currents by a clever design of the multiterminal geometry [21,96,97].

Complementary to the generation of pure spin current in multiterminal geometries, there are other types of devices not relying on the application of a net dc-bias. In spin pumping, the cyclic variation of two or more system parameters, such as e. g. gate voltages, induces spin-polarized currents at zero bias, where the induced charge current can be tuned to disappear, leaving pure spin currents. Several realizations of spin pumps in mesoscopic systems have been proposed, relying on SO interaction [98,99] or the Zeeman coupling of electrons to external magnetic fields [100]. The latter proposal has been experimentally confirmed [101] by detecting spin-polarized currents making use of a transverse electron focusing technique [73] mentioned in the previous section.

Complementary to pumps, ratchets only require a single driving parameter to achieve directed transport, and the current direction can be switched upon tuning external parameters such as temperature. In addition to the requirement of a broken spatial symmetry the ratchet has to be operated out of thermal equilibrium. The concept of particle ratchets, which has been addressed in numerous works [102], has recently been extended to systems called spin ratchets. To be specific, the mesoscopic spin ratchets proposed so far [82,85,103], are based on a quantum wire realized in a 2DEG. Between the two contacts attached to the quantum wire an ac bias voltage $U_R(t)$ is applied (rocking ratchet) with zero net (time-averaged) bias. Furthermore, in the central region of the quantum wire the spin degeneracy is lifted due to either SO interaction [82] or the Zeeman coupling to an external magnetic field [85,103]. Upon appropriate choice of the system geometry and tuning of the external driving the charge transported in the forward ($U_R > 0$) and backward bias ($U_R < 0$) situation can be made equal allowing for spin currents in the absence of net charge transport.

In the following we outline the model for the spin ratchets introduced in [82,85,103]. There, driving with a period t_0 is considered. It is implied, that this period is much larger than characteristic time scales related to the electron transport through the quantum wire. Therefore, the system is assumed to be in a steady state at every instance of time, and the Landauer–Büttiker approach to transport is used for the computation of the ratchet currents. To be specific, we consider an unbiased square wave driving $U_R(t) = U_0 \,\mathrm{sign}[\sin(2\pi t/t_0)]$, where $U_R(t)$ is restricted to the values $\pm U_0$. The net current is then given by the average of the steady-state currents in the two rocking situations (labeled $+U_0$ and $-U_0$ respectively) for both charge and spin,

$$\langle I_{C/S}(U_0)\rangle = \frac{I_{C/S}(+U_0) + I_{C/S}(-U_0)}{2} \,. \tag{28}$$

Since the spin ratchet effect requires nonlinear transport [103], i. e. finite bias voltages, the Hamiltonian (24) introduced in the previous section has to be extended to additionally include the effective electrostatic potential in the conductor due to the applied bias. Therefore, we add

the term $H_R = eU_R g(x, y; U_R)$ to the Hamiltonian (24) yielding the full Hamiltonian at finite bias:

$$H = H_0 + H_R \,. \tag{29}$$

Furthermore at finite bias a generalized version of the expressions for charge and spin current valid at $U_R \neq 0$ has to be used. For coherent Landauer transport those currents can be obtained from an integration of the transmission probabilities over the Fermi window [1]. Finally, the averaged charge $\langle I_C \rangle$ and spin $\langle I_S \rangle$ currents can be written as [85]

$$\langle I_{C/S}(U_0) \rangle = \frac{G_{C/S}}{2e} \int_{E_C}^{\infty} \mathrm{d}E \, \Delta f(E, U_0) \Delta T_{C/S}(E, U_0) \,.$$

Here E_C is the energy of the conduction band edge and $\Delta f(E, U_0) = [f(E, E_F + eU_0/2) - f(E, E_F - eU_0/2)]$ is the difference between the Fermi functions of the leads at bias voltage U_0, defining the Fermi window. The averaged charge/spin transmission is just the difference between the steady state transmissions in the two rocking situations:

$$\Delta T_{C/S}(E, U_0) = T_{C/S}(E, +U_0) - T_{C/S}(E, -U_0) \,. \tag{30}$$

Considering the Hamiltonian (29) we now show under what conditions the net charge transported after one full rocking period is zero, i. e., $\langle I_C(U_0) \rangle = 0$. If the electrostatic potentials $V(x, y)$, $U_B(x, y)$ and the Rashba SO strength $\alpha(x)$ are invariant under inversion of the x-coordinate,

$$\begin{aligned} V(x, y) &= V(-x, y) \,, \\ U_B(x, y) &= U_B(-x, y) \,, \\ \alpha(x) &= \alpha(-x) \,, \end{aligned} \tag{31}$$

it is appropriate to assume that the electrostatic potential distribution due to the finite applied voltage $g(x, y; U_R)$ also possesses this symmetry. Then the total Hamiltonian (29) is invariant under the action of the symmetry operation $\hat{\mathcal{P}} = -\mathrm{i}\hat{C}\hat{R}_U \hat{R}_x \sigma_z$ where R_U switches the sign of the bias voltage ($\pm U_0 \leftrightarrow \mp U_0$), yielding the relation

$$T_{\sigma,\sigma'}(E, \pm U_0) = T_{\sigma',\sigma}(E, \mp U_0)$$

between the spin-resolved transmission probabilities in the two rocking situations [85]. Inserting this relation into Eq. (30), we observe that the expression for the net charge transmission ΔT_C is zero, resulting in vanishing net charge current. Furthermore it can be used to simplify the expression for the net spin current:

$$\begin{aligned} \langle I_S(U_0) \rangle = &\frac{G_S}{e} \int_{E_C}^{\infty} \mathrm{d}E \; \Delta f(E; U_0) \\ &\times [T_{+,-}(E, +U_0) - T_{-,+}(E, +U_0)] \,. \end{aligned}$$

At $U_0 = 0$ the Hamiltonian (29) reduces to Eq. (24), which is invariant under the operation of Eq. (26), yielding $T_{+,-}(E, 0) = T_{-,+}(E, 0)$. This absence of net spin current $\langle I_S(U_0 = 0) \rangle = 0$ in the linear response regime is in agreement with the theoretical prediction [85]. However, for finite rocking voltages $U_0 \neq 0$ the additional potential introduced via H_R breaks this symmetry and therefore enables different spin flip transmissions and thus a resulting net spin current.

We now turn our attention to the quantum wire shown in Fig. 12a. For this system, which exhibits the symmetries of Eq. (31), we perform numerical calculations at finite bias U_0, in order to confirm its operability as a spin ratchet. In general, the function $g(x, y; U_R)$ has to be obtained by self-consistently solving the Schrödinger and Poisson equation of the system. However, in the present treatment we make use of a simple model for $g(x, y; U_R)$ assuming a linear voltage drop in the region where the point contact is formed in the quantum wire:

$$g(x, y; U_R) = g(x) = \begin{cases} \frac{1}{2} & \text{for } x < -\frac{L_{PC}}{2} \,, \\ -\frac{x}{L_{PC}} & \text{for } -\frac{L_{PC}}{2} < x < \frac{L_{PC}}{2} \,, \\ -\frac{1}{2} & \text{for } x > \frac{L_{PC}}{2} \,. \end{cases}$$

It is well known that a constriction in a quantum wire acts as an effective potential barrier constituting a region where the voltage applied across the wire is likely to drop. Since the bias voltages we consider are small compared to the energy shift introduced by the constriction, this assumption of a linear voltage drop should be an appropriate approximation for the actual distribution of the electrostatic potential in this wire [104]. Self-consistent calculations have confirmed that this approximation is valid [105].

In Fig. 15 we plot the net spin transmission ΔT_S as a function of $\bar{\alpha}$ and the injection energy in the range, where both leads support two transversal modes. We observe that $|\Delta T_S|$ reaches values of up to 0.9 in the parameter range shown. Furthermore, for a given value of injection energy, the sign of ΔT_S can be switched by changing α_C. Since the strength of the Rashba SO interaction α_C can be tuned via gate voltages [25] the presented system offers also the possibility for experimentally steering and switching the spin current direction.

Future Directions

In the present work we outlined general theoretical and computational concepts of coherent spin-dependent transport at low temperatures and focussed, with regard to numerical examples and possible experimental realiza-

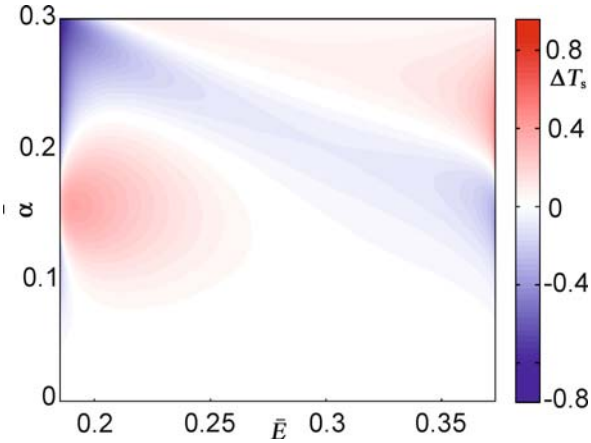

Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 15

Net ratchet spin transmission probability $\Delta T_S(U_0) = 2\left(T_{+,-}(+U_0) - T_{-,+}(+U_0)\right)$ for a stripe with a junction (see text) presented as a function of injection energy $\bar{E}$ and SO interaction strength $\bar{\alpha}$ at finite applied voltage $eU_0 = 0.02\,\hbar^2/(2m^* a^2)$. Note the sign change in the spin transmission upon tuning the SO coupling

tions, onto ballistic two-dimensional nanostructures based on non-magnetic high-mobility semiconductors.

In order to experimentally achieve high spin polarizations and reasonable spin currents, if possible at room temperature, broad efforts are made to investigate and design novel materials for spintronics. Here, prominent and promising examples, both with respect to fundamental physics and possible applications, are magnetic semiconductors such as GaMnAs [106] or semimagnetic materials with huge g-factors, for instance HgTe [107]. Charge transport in these materials is based on holes. However, relatively few theoretical papers deal with phase coherence effects for hole (spin) transport, though the rich band structure and the interplay between heavy and light hole (or electron- and hole-like) degree of freedoms promise interesting additional features.

The theoretical methods for quantum transport, presented here in the context of mesoscopic systems, are also applied and extended to treat transport in a further prospective field, namely through single-molecule junctions [108], for instance (break) junctions with a molecule bridging the gap between two leads or scanning tunneling microscope measurements of tunnel current through molecules at surfaces. In *Molecular Spintronics* [44,109] spin effects in transport through molecules are addressed. This subfield of spin electronics is still in its infancy. On the computational side these systems pose considerable problems since an adequate approach requires an appropriate description of the electronic and possibly vibrational properties of the molecule including the coupling to and effects of the leads. Whether (spin) DFT calculations for such an embedded molecule, combined with Landauer-type transport calculations, are appropriate, remains to be an issue, in particular if charging or non-equilibrium effects are involved.

As a further future direction we expect that spin transport in graphene, monolayers of graphite, may evolve as another future research line. After its experimental discovery in 2004 [110], graphene has gained much experimental and huge theoretical attention owing to its many exotic properties such as the massless charge carriers, internal spin-like degree of freedoms and unconventional transport characteristics. Also first experiments on graphene-based nanoconductors, e. g., measurements of the Aharonov–Bohm effect in graphene rings [111], are on their way. Graphene is also viewed as a prospective candidate for spin-electronics, since the spin decoherence and spin relaxation times in graphene are expected to be long [112,113]. Recent promising experiments already succeeded in injecting spins from ferromagnetic metallic contacts into graphene, although the conductance mismatch between graphene and the ferromagnetic leads is expected to suppress the efficiency. Recent theoretical proposals predict efficient spin injection into bulk graphene from graphene ribbons employing the occurrence of current-carrying spin-polarized edge states in the ribbons [114].

Acknowledgments

Parts of the results presented are based on previous work together with D. Bercioux, D. Frustaglia, and A. Pfund. We further thank I. Adagideli, M. Grifoni and D. Weiss for useful discussions. Two of us, KR and MW, acknowledge financial support by the Deutsche Forschungsgemeinschaft (SFB 689) and MS through the Studienstiftung des Deutschen Volkes.

Bibliography

Primary Literature

1. Datta S (2002) Electronic Transport in Mesoscopic Systems. Cambridge University Press, Cambridge
2. Rashba EI (1960) Properties of semiconductors with an extremum loop .1. cyclotron and combinational resonance in a magnetic field perpendicular to the plane of the loop. Sov Phys Solid State 2:1109–1122
3. Bychkov YA, Rashba EI (1984) Oscillatory effects and the magnetic susceptibility of carriers in inversion layers. J Phys C: Solid State Phys 17:6039–6045

4. Hentschel M, Schomerus H, Frustaglia D, Richter K (2004) Aharonov–Bohm physics with spin. I. Geometric phases in one-dimensional ballistic rings. Phys Rev B 69:155326
5. Cohen G, Hod O, Rabani E (2007) Constructing spin interference devices from nanometric rings. Phys Rev B 76:235120
6. Loss D, Goldbart P, Balatsky AV (1990) Berry's phase and persistent charge and spin currents in textured mesoscopic rings. Phys Rev Lett 65:1655–1658
7. Stern A (1992) Berry's phase, motive forces, and mesoscopic conductivity. Phys Rev Lett 68:1022–1025
8. Aronov AG, Lyanda-Geller YB (1993) Spin-orbit Berry phase in conducting rings. Phys Rev Lett 70:343–346
9. Qian TZ, Su ZB (1994) Spin-orbit interaction and Aharonov–Anandan phase in mesoscopic rings. Phys Rev Lett 72: 2311–2315
10. Frustaglia D, Hentschel M, Richter K (2001) Quantum transport in nonuniform magnetic fields: Aharonov–Bohm ring as a spin switch. Phys Rev Lett 87:256602
11. Frustaglia D, Hentschel M, Richter K (2004) Aharonov–Bohm physics with spin. II. Spin-flip effects in two-dimensional ballistic systems. Phys Rev B 69:155327
12. Nitta J, Bergsten T (2007) Time reversal Aharonov–Casher effect using Rashba spin-orbit interaction. New J Phys 9:341
13. Morpurgo AF, Heida JP, Klapwijk TM, van Wees BJ, Borghs G (1998) Ensemble-average spectrum of Aharonov–Bohm conductance oscillations: Evidence for spin-orbit-induced Berry's phase. Phys Rev Lett 80:1050–1053
14. Yau JB, De Poortere EP, Shayegan M (2002) Aharonov–Bohm oscillations with spin: Evidence for Berry's phase. Phys Rev Lett 88:146801
15. König M, Tschetschetkin A, Hankiewicz EM, Sinova J, Hock V, Daumer V, Schäfer M, Becker CR, Buhmann H, Molenkamp LW (2006) Direct observation of the Aharonov–Casher phase. Phys Rev Lett 96:076804
16. Grbić B, Leturcq R, Ihn T, Ensslin K, Reuter D, Wieck AD (2007) Aharonov–Bohm oscillations in the presence of strong spin-orbit interactions. Phys Rev Lett 99:176803
17. Frustaglia D, Richter K (2004) Spin interference effects in ring conductors subject to Rashba coupling. Phys Rev B 69:235310
18. Nitta J, Meijer FE, Takayanagi H (1999) Spin-interference device. Appl Phys Lett 75:695–697
19. Mal'shukov AG, Shlyapin VV, Chao KA (1999) Effect of the spin-orbit geometric phase on the spectrum of Aharonov–Bohm oscillations in a semiconductor mesoscopic ring. Phys Rev B 60:R2161–R2164
20. Splettstoesser J, Governale M, Zülicke U (2003) Persistent current in ballistic mesoscopic rings with Rashba spin-orbit coupling. Phys Rev B 68:165341
21. Souma S, Nikolić BK (2005) Spin Hall current driven by quantum interferences in mesoscopic Rashba rings. Phys Rev Lett 94:106602
22. Popp M, Frustaglia D, Richter K (2003) Spin filter effects in mesoscopic ring structures. Nanotechnology 14:347–351
23. Bellucci S, Onorato P (2007) Filtering of spin currents based on a ballistic ring. J Phys: Condens Matter 19:395020
24. Ionicioiu R, D'Amico I (2003) Mesoscopic Stern–Gerlach device to polarize spin currents. Phys Rev B 67:041307
25. Nitta J, Akazaki T, Takayanagi H, Enoki T (1997) Gate control of spin-orbit interaction in an inverted $In_{0.53}$ $Ga_{0.47}$ As/$In_{0.52}$ $Al_{0.48}$ As heterostructure. Phys Rev Lett 78:1335–1338
26. Schmidt G, Ferrand D, Molenkamp LW, Filip AT, van Wees BJ (2000) Fundamental obstacle for electrical spin injection from a ferromagnetic metal into a diffusive semiconductor. Phys Rev B 62:R4790–R4793
27. Sinova J, Culcer D, Niu Q, Sinitsyn NA, Jungwirth T, MacDonald AH (2004) Universal intrinsic spin Hall effects. Phys Rev Lett 92:126603
28. New J Phys (2007) Focus on Spintronics in Reduced Dimensions. New J Phys 9
29. Corresponding settings, where a precessing magnetization in a ferromagnet emits spin currents, were reviewed by Tserkovnyak Y, Brataas A, Bauer GEW, Halperin BI (2005) Nonlocal magnetization dynamics in ferromagnetic heterostructures. Rev Mod Phys 77:1375
30. Landauer R (1957) Spatial variation of currents and fields due to localized scatterers in metallic conduction. IBM J Res Dev 1:223–231
31. Büttiker M, Imry Y, Landauer R, Pinhas S (1985) Generalized many-channel conductance formula with application to small rings. Phys Rev B 31:6207–6215
32. Stone AD, Szafer A (1988) What is measured when you measure a resistance? – The Landauer forumula revisited. IBM J Res Dev 32:384–413
33. Ferry DK, Goodnick SM (2001) Transport in Nanostructures. Cambridge University Press, Cambridge
34. Fisher DS, Lee PA (1981) Relation between conductivity and transmission matrix. Phys Rev B 23:6851–6854
35. Baranger HU, Stone AD (1989) Electrical linear-response theory in an arbitrary magnetic field: A new fermi-surface formation. Phys Rev B 40:8169–8193
36. Kimball GE, Shortley GH (1934) The numerical solution of Schrödinger's equation. Phys Rev 45:815–820
37. Pauling L, Wilson EB (1935) Introduction to Quantum Mechanics. Dover, New York
38. Havu P, Havu V, Puska MJ, Nieminen RM (2004) Nonequilibrium electron transport in two-dimensional nanostructures modeled using Green's functions and the finite-element method. Phys Rev B 69:115325
39. Bowen RC, Klimeck G, Lake RK, Frensley WR, Moise T (1997) Quantitative simulation of a resonant tunneling diode. J Appl Phys 81:3207–3213
40. Sanvito S, Lambert CJ, Jefferson JH, Bratkovsky AM (1999) General Green's-function formalism for transport calculations with *spd* Hamiltonians and giant magnetoresistance in Co- and Ni-based magnetic multilayers. Phys Rev B 59:11936–11948
41. Luisier M, Schenk A, Fichtner W, Klimeck G (2006) Atomistic simulation of nanowires in the $sp^3d^5s^*$ tight-binding formalism: From boundary conditions to strain calculations. Phys Rev B 74:205323
42. Brandbyge M, Mozos JL, Ordejón P, Taylor J, Stokbro K (2002) Density-functional method for nonequilibrium electron transport. Phys Rev B 65:165401
43. Di Carlo A, Pecchia A, Latessa L, Frauenheim T, Seifert G (2005) Tight-binding DFT for molecular electronics (gDFTB). In: Cuniberti G, Fagas G, Richter K (eds) Introducing Molecular Electronics. Springer, Berlin, pp 153–184
44. Rocha AR, García-Suárez VM, Bailey S, Lambert C, Ferrer J, Sanvito S (2006) Spin and molecular electronics in atomically generated orbital landscapes. Phys Rev B 73:085414

45. Lopez Sancho MP, Lopez Sancho JM, Rubio J (1984) Quick iterative scheme for the calculation of transfer matrices: application to Mo (100). J Phys F: Met Phys 14:1205–s1215
46. Lopez Sancho MP, Lopez Sancho JM, Rubio J (1985) Highly convergent schemes for the calculation of bulk and surface Green functions. J Phys F: Met Phys 15:851–858
47. Krstić PS, Zhang XG, Butler WH (2002) Generalized conductance formula for the multiband tight-binding model. Phys Rev B 66:205319
48. Usuki T, Takatsu M, Kiehl RA, Yokoyama N (1994) Numerical analysis of electron-wave detection by a wedge-shaped point contact. Phys Rev B 50:7615–7625
49. Usuki T, Saito M, Takatsu M, Kiehl RA, Yokoyama N (1995) Numerical analysis of ballistic-electron transport in magnetic fields by using a quantum point contact and a quantum wire. Phys Rev B 52:8244–8255
50. Lambert CJ, Weaire D (1980) Decimation and Anderson localization. Phys Status Solid B 101:591–595
51. Leadbeater M, Lambert CJ (1998) A decimation method for studying transport properties of disordered systems. Ann Phys 7:498–502
52. Mamaluy D, Vasileska D, Sabathil M, Zibold T, Vogl P (2005) Contact block reduction method for ballistic transport and carrier densities of open nanostructures. Phys Rev B 71:245321
53. Thouless DJ, Kirkpatrick S (1981) Conductivity of the disordered linear chain. J Phys C: Solid State Phys 14:235–245
54. Lee PA, Fisher DS (1981) Anderson localization in two dimensions. Phys Rev Lett 47:882–885
55. MacKinnon A (1985) The calculation of transport properties and density of states of disordered solids. Z Phys B 59:385–390
56. Baranger HU, DiVincenzo DP, Jalabert RA, Stone AD (1991) Classical and quantum ballistic-transport anomalies in microjunctions. Phys Rev B 44:10637–10675
57. Lake R, Klimeck G, Bowen RC, Jovanovic D (1997) Single and multiband modeling of quantum electron transport through layered semiconductor devices. J Appl Phys 81:7845–7869
58. Lassl A, Schlagheck P, Richter K (2007) Effects of short-range interactions on transport through quantum point contacts: A numerical approach. Phys Rev B 75:045346
59. Drouvelis P, Schmelcher P, Bastian P (2006) Parallel implementation of the recursive Green's function method. J Comp Phys 215:741–756
60. Rotter S, Tang JZ, Wirtz L, Trost J, Burgdörfer J (2000) Modular recursive Green's function method for ballistic quantum transport. Phys Rev B 62:1950–1960
61. Rotter S, Weingartner B, Rohringer N, Burgdörfer J (2003) Ballistic quantum transport at high energies and high magnetic fields. Phys Rev B 68:165302
62. Cuthill E, McKee J (1969) Reducing the bandwidth of sparse symmetric matrices. In: Proc 24th Nat Conf, ACM, New York, pp 157–172
63. George A (1971) Computer implementation of the finite element method. Tech Rep STAN-CS-71-208, Computer Sci Dept, Stanford Univ, Stanford
64. Liu WH, Sherman AH (1976) Comparative analysis of the Cuthill–McKee and the reverse Cuthill–McKee ordering algorithms for sparse matrices. SIAM J Num Anal 13:198–213
65. Gibbs NE, William G, Poole J, Stockmeyer PK (1976) An algorithm for reducing the bandwidth and profile of a sparse matrix. SIAM J Num Anal 13:236–250
66. Wimmer M, Richter K (2008) Optimal block-tridiagonalization of matrices for coherent charge transport. arXiv:0806.2739v1
67. Whaley RC, Petitet A, Dongarra JJ (2001) Automated empirical optimization of software and the ATLAS project. Parallel Comput 27:3–35
68. van Houten H, Beenakker CWJ, Williamson JG, Broekaart MEI, van Loosdrecht PHM, van Wees BJ, Mooij JE, Foxon CT, Harris JJ (1989) Coherent electron focusing with quantum point contacts in a two-dimensional electron gas. Phys Rev B 39:8556–8575
69. Usaj G, Balseiro CA (2004) Transverse electron focusing in systems with spin-orbit coupling. Phys Rev B 70:041301
70. Govorov AO, Kalameitsev AV, Dulka JP (2004) Spin-dependent transport of electrons in the presence of a smooth lateral potential and spin-orbit interaction. Phys Rev B 70:245310
71. Reynoso A, Usaj G, Balseiro CA (2007) Detection of spin polarized currents in quantum point contacts via transverse electron focusing. Phys Rev B 75:085321
72. Rokhinson LP, Larkina V, Lyanda-Geller YB, Pfeiffer LN, West KW (2004) Spin separation in cyclotron motion. Phys Rev Lett 93:146601
73. Potok RM, Folk JA, Marcus CM, Umansky V (2002) Detecting spin-polarized currents in ballistic nanostructures. Phys Rev Lett 89:266602
74. Folk JA, Potok RM, Marcus CM, Umansky V (2003) A Gate-Controlled Bidirectional Spin Filter Using Quantum Coherence. Science 299:679–682
75. Rokhinson LP, Pfeiffer LN, West KW (2006) Spontaneous spin polarization in quantum point contacts. Phys Rev Lett 96:156602
76. Fabian J, Das Sarma S (2002) Spin transport in inhomogeneous magnetic fields: A proposal for Stern–Gerlach-like experiments with conduction electrons. Phys Rev B 66:024436
77. Kiselev AA, Kim KW (2001) T-shaped ballistic spin filter. Appl Phys Lett 78:775–777
78. Ohe JI, Yamamoto M, Ohtsuki T, Nitta J (2005) Mesoscopic Stern–Gerlach spin filter by nonuniform spin-orbit interaction. Phys Rev B 72:041308
79. Cummings AW, Akis R, Ferry DK (2006) Electron spin filter based on rashba spin-orbit coupling. Appl Phys Lett 89:172115
80. Eto M, Hayashi T, Kurotani Y (2005) Spin polarization at semiconductor point contacts in absence of magnetic field. J Phys Soc Jpn 74:1934
81. Zhai F, Xu HQ (2007) Spin filtering and spin accumulation in an electron stub waveguide with spin-orbit interaction. Phys Rev B 76:035306
82. Scheid M, Pfund A, Bercioux D, Richter K (2007) Coherent spin ratchets: A spin-orbit based quantum ratchet mechanism for spin-polarized currents in ballistic conductors. Phys Rev B 76:195303
83. Song JF, Ochiai Y, Bird JP (2003) Fano resonances in open quantum dots and their application as spin filters. Appl Phys Lett 82:4561–4563
84. Zhai F, Xu HQ (2006) Spin filtering in single magnetic barrier structures revisited. Appl Phys Lett 88:032502

85. Scheid M, Bercioux D, Richter K (2007) Zeeman ratchets: pure spin current generation in mesoscopic conductors with non-uniform magnetic fields. New J Phys 9:401
86. Shi QW, Zhou J, Wu MW (2004) Spin filtering through a double-bend structure. Appl Phys Lett 85:2547–2549
87. Zhai F, Xu HQ (2005) Generation of spin polarization in two-terminal electron waveguides by spin-orbit interaction and magnetic field modulations. Phys Rev B 72:085314
88. Shi J, Zhang P, Xiao D, Niu Q (2006) Proper definition of spin current in spin-orbit coupled systems. Phys Rev Lett 96:076604
89. Zhai F, Xu HQ (2005) Symmetry of spin transport in two-terminal waveguides with a spin-orbital interaction and magnetic field modulations. Phys Rev Lett 94:246601
90. Silvestrov PG, Mishchenko EG (2006) Polarized electric current in semiclassical transport with spin-orbit interaction. Phys Rev B 74:165301
91. van Wees BJ, van Houten H, Beenakker CWJ, Williamson JG, Kouwenhoven LP, van der Marel D, Foxon CT (1988) Quantized conductance of point contacts in a two-dimensional electron gas. Phys Rev Lett 60:848–850
92. Wharam DA, Thornton TJ, Newbury R, Pepper M, Ahmed H, Frost JEF, Hasko DG, Peacock DC, Ritchie DA, Jones GAC (1988) One-dimensional transport and the quantisation of the ballistic resistance. J Phys C: Solid State Phys 21:L209–L214
93. Sánchez D, Serra L (2006) Fano–Rashba effect in a quantum wire. Phys Rev B 74:153313
94. Dresselhaus G (1955) Spin-orbit coupling effects in zinc blende structures. Phys Rev 100:580–586
95. Bardarson JH, Adagideli I, Jacquod P (2007) Mesoscopic spin Hall effect. Phys Rev Lett 98:196601
96. Hankiewicz EM, Molenkamp LW, Jungwirth T, Sinova J (2004) Manifestation of the spin Hall effect through charge-transport in the mesoscopic regime. Phys Rev B 70:241301
97. Sheng L, Sheng DN, Ting CS (2005) Spin-Hall effect in two-dimensional electron systems with Rashba spin-orbit coupling and disorder. Phys Rev Lett 94:016602
98. Governale M, Taddei F, Fazio R (2003) Pumping spin with electrical fields. Phys Rev B 68:155324
99. Sharma P, Brouwer PW (2003) Mesoscopic effects in adiabatic spin pumping. Phys Rev Lett 91:166801
100. Mucciolo ER, Chamon C, Marcus CM (2002) Adiabatic quantum pump of spin-polarized current. Phys Rev Lett 89:146802
101. Watson SK, Potok RM, Marcus CM, Umansky V (2003) Experimental realization of a quantum spin pump. Phys Rev Lett 91:258301
102. Reimann P (2002) Brownian motors: noisy transport far from equilibrium. Phys Rep 361:57–265
103. Scheid M, Wimmer M, Bercioux D, Richter K (2006) Zeeman ratchets for ballistic spin currents. Phys Status Solidi (c) 3:4235
104. McLennan MJ, Lee Y, Datta S (1991) Voltage drop in mesoscopic systems: A numerical study using a quantum kinetic equation. Phys Rev B 43:13846–13884
105. Lassl A, Scheid M, Richter K (2008) Unpublished
106. Ohno H, Shen A, Matsukura F, Oiwa A, Endo A, Katsumoto S, Iye Y (1996) (Ga,Mn)As: A new diluted magnetic semiconductor based on GaAs. Appl Phys Lett 69:363–365
107. König M, Wiedmann S, Brüne C, Roth A, Buhmann H, Molenkamp LW, Qi XL, Zhang SC (2007) Quantum spin Hall insulator state in HgTe quantum wells. Science 318:766–770
108. Cuniberti G, Fagas G, Richter K (eds) (2005) Introducing Molecular Electronics. Springer, Berlin
109. Emberly E, Kirczenow G (2002) Molecular spintronics: spin-dependent electron transport in molecular wires. Chemical Physics 281:311–324
110. Novoselov KS, Geim AK, Morozov SV, Jiang D, Zhang Y, Dubonos SV, Grigorieva IV, Firsov AA (2004) Electric field effect in atomically thin carbon films. Science 306:666–669
111. Russo S, Oostinga JB, Wehenkel D, Heersche HB, Sobhani SS, Vandersypen LMK, Morpurgo AF (2007) Aharonov–Bohm effect in graphene. arXiv:0711.1508v1
112. Huertas-Hernando D, Guinea F, Brataas A (2006) Spin-orbit coupling in curved graphene, fullerenes, nanotubes, and nanotube caps. Phys Rev B 74:155426
113. Min H, Hill JE, Sinitsyn NA, Sahu BR, Kleinman L, MacDonald AH (2006) Intrinsic and Rashba spin-orbit interactions in graphene sheets. Phys Rev B 74:165310
114. Wimmer M, Adagideli I, Berber S, Tománek D, Richter K (2008) Spin transport in rough graphene nanoribbons. Phys Rev Lett 100:177207

Books and Reviews

Bruus H, Flensberg K (2004) Many-body Quantum Theory in Condensed Matter Physics: An Introduction. Oxford University Press, Oxford

Datta S (2002) Electronic Transport in Mesoscopic Systems. Cambridge University Press, Cambridge

Fabian J, Matos-Abiague A, Ertler C, Stano P, Žutić I (2007) Semiconductor Spintronics. Acta Physica Slovaca 57:565–907

Ferry DK, Goodnick SM (2001) Transport in Nanostructures. Cambridge University Press, Cambridge

Stability and Feedback Stabilization

Eduardo D. Sontag
Department of Mathematics, Rutgers University,
New Brunswick, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Linear Systems
Nonlinear Systems: Continuous Feedback
Discontinuous Feedback
Sensitivity to Small Measurement Errors
Future Directions
Bibliography

Glossary

Stability A globally asymptotically stable equilibrium is a state with the property that all solutions converge to this state, with no large excursions.

Stabilization A system is stabilizable (with respect to a given state) if it is possible to find a feedback law that renders that state a globally asymptotically stable equilibrium.

Lyapunov and control-Lyapunov functions A control-Lyapunov functions is a scalar function which decreases along trajectories, if appropriate control actions are taken. For systems with no controls, one has a Lyapunov function.

Definition of the Subject

The problem of stabilization of equilibria is one of the central issues in control. In addition to its intrinsic interest, it represents a first step towards the solution of more complicated problems, such as the stabilization of periodic orbits or general invariant sets, or the attainment of other control objectives, such as tracking, disturbance rejection, or output feedback, all of which may be interpreted as requiring the stabilization of some quantity (typically, some sort of "error" signal). A very special case, when there are no inputs, is that of stability.

Introduction

This article discusses the problem of stabilization of an equilibrium, which we take without loss of generality to be the origin, for a finite-dimensional system $\dot{x} = f(x, u)$. The objective is to find a *feedback law* $u = k(x)$ which renders the origin of the "closed-loop" system $\dot{x} = f(x, k(x))$ globally asymptotically stable. The problem of stabilization of equilibria is one of the central issues in control. In addition to its intrinsic interest, it represents a first step towards the solution of more complicated problems, such as the stabilization of periodic orbits or general invariant sets, or the attainment of other control objectives, such as tracking, disturbance rejection, or output feedback, all of which may be interpreted as requiring the stabilization of some quantity (typically, some sort of "error" signal). A very special case (when there are no inputs u) is that of stability.

After setting up the basic definitions, we consider *linear* systems. Linear systems are widely used as models for physical processes, and they also play a major role in the general theory of local stabilization. We briefly review pole assignment and linear-quadratic optimization as approaches to obtaining feedback stabilizers.

In general, there is a close connection between the existence of continuous stabilizing feedbacks and smooth *control-Lyapunov functions*, (cfl's), which constitute an extension of the classical concept of Lyapunov functions from dynamical system theory. We discuss the role of clf's in design methods and "universal" formulas for feedback controls.

For nonlinear systems, it has been known since the late 1970s that, in general, there are topological obstructions to the existence of even continuous stabilizers. We review these obstructions, using tools from degree theory.

Finally, we turn to discontinuous stabilization and the associated issue of defining precisely a "solution" for a differential equation with discontinuous right-hand side. We introduce techniques from nonsmooth analysis and differential games, in order to deal with discontinuous controllers. In particular, we discuss the effect of measurement errors on the performance of such controllers.

Preliminaries

In this article, we restrict attention to continuous-time deterministic systems whose states evolve in finite-dimensional Euclidean spaces $\mathbb{R}^n$. (This excludes many other equally important objects of study in control theory: systems which evolve on infinite dimensional spaces and are described by PDE's, systems evolving on manifolds which serve to model state constraints, discrete-time systems described by difference equations, and stochastic systems, among others.) In order to streamline the presentation, we suppose throughout that controls take values in $\mathcal{U} = \mathbb{R}^m$ (constraints in controls would lead to proper subsets $\mathcal{U}$). A *control* (other names: *input*, *forcing function*) is any measurable locally essentially bounded map $u(\cdot)\colon [0, \infty) \to \mathcal{U} = \mathbb{R}^m$. In general, we use the notation $|x|$ for Euclidean norms, and use $\|u\|$ to indicate the essential supremum of a function $u(\cdot)$. For basic terminology and facts about control systems, see [25].

Given a map $f\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ which is locally Lipschitz and satisfies $f(0, 0) = 0$, we consider the associated forced system of ordinary differential equations

$$\dot{x}(t) = f(x(t), u(t)) \,. \tag{1}$$

The maximal solution $x(\cdot)$ of (1) which corresponds to a given initial state $x(0) = x^0$ and to a given control u is defined on some maximal interval $[0, t_{\max}(x^0, u))$, and is denoted by $x(t, x^0, u)$. In the special case when f does not depend on u, we have an unforced system, or system with no inputs

$$\dot{x}(t) = f(x(t)) \,. \tag{2}$$

Unforced systems are associated to a controlled system (1) in two different ways. The first is when one substitutes a feedback law $u = k(x)$ in (1) to obtain a "closed-loop" system $\dot{x} = f(x, k(x))$. The second is when one considers

instead the autonomous system $\dot{x} = f(x, 0)$ which models the behavior of (1) in the absence of any controls. All definitions stated for unforced systems are implicitly applied also to systems with inputs (1) by setting $u \equiv 0$; for instance, we define the global asymptotic stability (GAS) property for (2), but we say that (1) is GAS if $\dot{x} = f(x, 0)$ is. For systems with no inputs (2) we write just $x(t, x^0)$ instead of $x(t, x^0, u)$.

Stability and Asymptotic Controllability Stability is one of the most important objectives in control theory, because a great variety of problems can be recast in stability terms. This includes questions of driving a system to a desired configuration (e. g., an inverted pendulum on a cart, to its upwards position), or the problem of tracking a reference signal (such as a pilot's command to an aircraft). We focus in this talk on global asymptotic stabilization.

Recall that the class of $\mathcal{K}_\infty$ functions consists of all $\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ which are continuous, strictly increasing, unbounded, and satisfy $\alpha(0) = 0$. The class of $\mathcal{KL}$ functions consists of those $\beta : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with the properties that

(1) $\beta(\cdot, t) \in \mathcal{K}_\infty$ for all t, and
(2) $\beta(r, t)$ decreases to zero as $t \to \infty$.

We will also use $\mathcal{N}$ to denote the set of all nondecreasing functions $\sigma : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$. Expressed in terms of such comparison functions, the property of *global asymptotic stability (GAS)* of the origin for a system with no inputs (2) becomes:

$$(\exists \beta \in \mathcal{KL}) \quad |x(t, x^0)| \leq \beta\left(|x^0|, t\right) \quad \forall x^0 , \forall t \geq 0 .$$

This definition is equivalent to a more classical "ε-δ" definition usually provided in textbooks, which defines GAS as the combination of stability and global attractivity. For one implication, simply observe that

$$|x(t, x^0)| \leq \beta\left(|x^0|, 0\right)$$

provides the stability (or "small overshoot") property, while

$$|x(t, x^0)| \leq \beta\left(|x^0|, t\right) \underset{t\to\infty}{\longrightarrow} 0$$

gives attractivity. The converse implication is an easy exercise.

More generally, we define what it means for the system with inputs (1) to be (open loop, globally) *asymptotically controllable (AC)* (to the origin). The definition amounts to requiring that for each initial state x^0 there exists some control $u = u_{x^0}(\cdot)$ defined on $[0, \infty)$, such that the corresponding solution $x(t, x^0, u)$ is defined for all $t \geq 0$, and converges to zero as $t \to \infty$, with "small" overshoot. Moreover, we wish to rule out the possibility that $u(t)$ becomes unbounded for x near zero. The precise formulation is as follows.

$$(\exists \beta \in \mathcal{KL})(\exists \sigma \in \mathcal{N}) \quad \forall x^0 \in \mathbb{R}^n \exists u(\cdot),\ \|u\| \leq \sigma(|x^0|),$$

$$|x(t, x^0, u)| \leq \beta\left(|x^0|, t\right) \quad \forall\, t \geq 0 .$$

In particular, (global) asymptotic stability amounts to the existence of $\beta \in \mathcal{KL}$ such that $|x(t, x^0, u)| \leq \beta\left(|x^0|, t\right)$ holds for all $t \geq 0$. A very special case is that of *exponential* stability, in which an estimate of the type $|x(t, x^0, u)| \leq Me^{-\lambda t}|x^0|$ holds. For linear systems (see below), asymptotic stability and exponential stability coincide. It is a puzzling fact that for general systems, one can find continuous coordinate changes that make asymptotically stable systems exponentially stable [8,13] (a fact of little practical utility, since finding such coordinate changes is as hard as establishing stability to being with).

Feedback Stabilization A map $k : \mathbb{R}^n \to \mathcal{U}$ is a *feedback stabilizer* for the system with inputs (1) if k is locally bounded (that is, k is bounded on each bounded subset of $\mathbb{R}$), $k(0) = 0$, and the closed-loop system

$$\dot{x} = f(x, k(x)) \tag{3}$$

is GAS, i. e. there is some $\beta \in \mathcal{KL}$ so that $|x(t)| \leq \beta\left(|x(0)|, t\right)$ for all solutions and all $t \geq 0$. (A technical difficulty with this definition lies the possible lack of solutions of (1) when k is not regular enough. We ignore this for now, but will most definitely return to this issue later.)

For example, if (1) is a model of an undamped spring/mass system, where u represents the net effect of external forces, one obvious way to asymptotically stabilize the system is to introduce damping. In control-theoretic terms, this means that we choose $u(t) = k(x(t))$ to be a negative linear function of the velocity. Physically, one may implement a feedback controller by means of an analog device. In the example of the spring/mass system, one could achieve this by adding friction or connecting a dashpot. Alternatively, in modern control technology, one uses a digital computer to measure the state x and compute the appropriate control action to be applied. (There are many implementation issues which arise in digital control and are ignored in our theoretical formulation "$u = k(x)$", among them the effect of delays in the actual computation of the control $u(t)$ to be applied at time t, and the effect of quantization due to the finite precision of measuring devices and the digital nature of the computer. These issues

are addressed in the literature, although a comprehensive theoretical framework is still lacking.)

Observe that, obviously, if there exists a feedback stabilizer for (1), then (1) is also AC (we just use $u(t) := k(x(t, x^0))$ as u_{x^0}). Thus, it is very natural to ask whether the converse holds: *is every asymptotically controllable system also feedback stabilizable?*

Linear Systems

A *linear system* is a system (1) for which the map f is linear. In other words, there are two linear transformations $A\colon \mathbb{R}^n \to \mathbb{R}^n$ and $B\colon \mathbb{R}^m \to \mathbb{R}^n$ so that the equations take the form

$$\dot{x} = Ax + Bu\,. \tag{4}$$

Such a system is completely specified once that we are given A and B, which we identify by abuse of notation with their respective matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ with respect to the canonical bases in $\mathbb{R}^n$ and $\mathbb{R}^m$. We also say "the system (A, B)" when referring to (4).

It is natural to look specifically for *linear* feedbacks $k\colon \mathbb{R}^n \to \mathbb{R}^m$ which stabilize a linear system (just as a linear term, inversely proportional to velocity, stabilizes an undamped harmonic oscillator). (In fact, this is no loss of generality, since it can be easily proved for linear systems [25] that if a feedback stabilizer $u = k(x)$ exists, then there also exists a linear feedback stabilizer.) We write $u = Fx$, when expressing $k(x) = Fx$ in matrix terms with respect to the canonical bases. Substituting this control law into (4) results in the equation $\dot{x} = (A + BF)x$. Thus, the mathematical problem reduces to:

given $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$, find $F \in \mathbb{R}^{m\times n}$ such that $A + BF$ is Hurwitz.

(Recall that a Hurwitz matrix is one all whose eigenvalues have negative real parts, and that the origin of the system $\dot{x} = Hx$ is globally asymptotically stable if and only if H is a Hurwitz matrix.) The fundamental stabilization result for linear systems is as follows [25]:

Theorem 1 *A linear system is asymptotically controllable if and only if it admits a linear feedback stabilizer.*

A Remark on Linearization

If the dynamics map f in (1) is continuously differentiable, we may expand to first order $f(x, u) = Ax + Bu + o(x, u)$. Let us suppose that the linearized system (A, B) is AC, and pick a linear feedback stabilizer $u = Fx$, whose existence is guaranteed by Theorem 1. Then, the same feedback law $k(x) = Fx$, when fed back into the original system (1), results in $\dot{x} = f(x, Fx) = (A + BF)x + o(x)$. Thus, k locally stabilizes the origin for the nonlinear system. Of course, the assumption that the linearization is AC is not always satisfied. Systems in which inputs enter multiplicatively, such as those controlling reaction rates in chemical problems, lead to degenerate linearizations. In addition, even if the linearized system (A, B) is AC, in general a linear stabilizer $u = Fx$ will not work as a global stabilizer. For example, the system $\dot{x} = x + x^2 + u$ can be locally stabilized with $u := -2x$, but any linear feedback $u = -fx$ $(f > 1)$ results in an additional equilibrium away from the origin (at $x = f - 1$). Nonlinear feedback must be used (obviously, in this, example, $u = -2x - x^2$ works for global stabilization).

Returning to linear systems, let us note that Theorem 1 is of great interest because (1) there is a simple algebraic test to check the AC property, and (2) there are several practically useful algorithms for obtaining a stabilizing F, including pole placement and optimization, which we sketch next.

Pole Placement

The first technique for stabilization is purely algebraic. In order to simplify this exposition, we will suppose that the system (4) is not just AC but is in fact *controllable*, meaning that every state can be steered, in finite time, to every other state. (Any AC system (4) can be decomposed into two components, of which one is already GAS and the other one is controllable, cf. [25], so this represents no loss of generality.) Controllability is characterized by the property – generically satisfied for pairs (A, B) – that

$$\operatorname{rank}\left[B \middle| AB \middle| A^2B \middle| \ldots \middle| A^{n-1}B\right] = n$$

(note that the composite matrix shown has n rows and nm columns).

Two pairs (A, B) and $(\widetilde{A}, \widetilde{B})$ are said to be *feedback equivalent* if there exist $T \in GL(n, \mathbb{R})$, $F_0 \in \mathbb{R}^{m\times n}$, and $V \in GL(m, \mathbb{R})$ so that

$$(A + BF_0)T = T\widetilde{A} \quad \text{and} \quad BV = T\widetilde{B}\,. \tag{5}$$

Feedback equivalence corresponds to changes of basis in the state and control-value spaces (invertible matrices T and V, respectively) and feedback transformations $u = F_0x + u'$, where u' is a new control. (An equivalent way to describe feedback equivalence is by the requirement that two pairs should be in the same orbit un-

der the action of a "feedback group" which is obtained as a suitable semidirect product of $GL(n,\mathbb{R})$, $(\mathbb{R}^{m\times n},+)$, and $GL(m,\mathbb{R})$.) Controllability is preserved under feedback equivalence. Moreover, if (5) holds and if one finds a matrix $\widetilde{F}$ so that $\widetilde{A}+\widetilde{B}\widetilde{F}$ is Hurwitz, then

$$A+BF = T(\widetilde{A}+\widetilde{B}\widetilde{F})T^{-1}$$

is also Hurwitz, where $F := F_0 + V\widetilde{F}T^{-1}$. Thus, the task of finding a stabilizing feedback F can be reduced to the same problem for any pair $(\widetilde{A},\widetilde{B})$ which is feedback equivalent to the given pair (A,B).

One then proceeds to show that there always exists an equivalent pair $(\widetilde{A},\widetilde{B})$ which has a form simple enough that the existence of $\widetilde{F}$ is trivial to establish. In order to find such a pair, it is useful to study the classification of controllable pairs under feedback equivalence. This classification is closely related to Kronecker's theory of "matrix pencils" applied to polynomial matrices $[\lambda I - A, B] = \lambda[I,0] + [-A,B]$ modulo matrix equivalence, cf. [25]. The orbits under feedback equivalence are in one-to-one correspondence with the possible partitions of $n = \kappa_1 + \ldots + \kappa_r$ into the sum of r positive integers, $r \le m$, and in each orbit one can find a pair $(\widetilde{A},\widetilde{B})$ which is in "controller canonical form", for which F can be trivially found. For simplicity, let is just discuss here the very special case of single-input systems ($m=1$). For this case, the action of the feedback group is transitive, and each controllable system is feedback equivalent to the following special system:

$$A := \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix} \quad B := \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} .$$

For this system, a stabilizing feedback is trivial to obtain. Indeed, take the polynomial $p(\lambda) = (\lambda+1)^n = \lambda^n - \alpha_n\lambda^{n-1} - \ldots - \alpha_2\lambda - \alpha_1$. With $F = (\alpha_1,\alpha_2,\alpha_3,\ldots,\alpha_n)$,

$$A+BF := \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \\ \alpha_1 & \alpha_2 & \alpha_3 & \ldots & \alpha_n \end{pmatrix}$$

has characteristic polynomial $(\lambda+1)^n$, and hence is a Hurwitz matrix, as required for stabilization.

Observe that, instead of the particular $p(\lambda)$ which we used, we could have picked any polynomial all whose roots have negative real parts, and the same argument applies. The conclusion is that, not only can we make $A+BF$ Hurwitz, but we can assign to it any desired set of n eigenvalues (as long as they form a set closed under conjugation). This is the reason that the technique is called *eigenvalue placement* (or "pole placement" because the eigenvalues of A are the poles of the resolvent $(\lambda I - A)^{-1}$). See Chap. 5 in [25] for a detailed treatment of the pole placement problem.

Variational Approach

A second technique for stabilization is based on optimal control techniques. We first pick any two symmetric positive definite matrices $R \in \mathbb{R}^{m\times m}$ and $Q \in \mathbb{R}^{n\times n}$ (for instance the identity matrices of the respective sizes). Next, we consider the problem of minimizing, for each initial state x^0 at time $t=0$, the infinite-horizon cost

$$\mathcal{J}_{x^0}(u) := \int_0^\infty \left\{u(t)'Ru(t) + x(t)'Qx(t)\right\} \mathrm{d}t$$

over all controls $u\colon [0,\infty) \to \mathbb{R}^m$ which make $\mathcal{J}_{x^0}(u) < \infty$, where $x(t) = x(t,x^0,u)$ and prime indicates transpose. The main result from linear-quadratic optimal control (cf. Sect. 8.4 in [25]) implies that, for AC systems, there is a unique solution u to this problem, which is given in the following form: there is a matrix $F \in \mathbb{R}^{m\times n}$ such that solving $\dot{x} = (A+BF)x$ with $x(0) = x^0$ gives that $u(t) := Fx(t)$ minimizes $\mathcal{J}_{x^0}(\cdot)$. Moreover, this F stabilizes the system (which is intuitively to be expected, since $\mathcal{J}_{x^0}(u) < \infty$ implies that solutions $x(t)$ must be in L^2), and F can be computed by the formula

$$F := -R^{-1}B'P\,, \tag{6}$$

where P is a symmetric and positive definite solution of the *Matrix Algebraic Riccati Equation*

$$PBR^{-1}B'P - PA - A'P - Q = 0\,. \tag{7}$$

A Sufficient Nonlinear Condition

Although of limited applicability, it is worth remarking that there is a partial extension to nonlinear systems of the stabilization method which was just described. For simplicity, we specialize our discussion to control-affine systems, i. e., those for which the input appears only in an affine form. This class is sufficient for the study of most forced mechanical systems. The Eq. (1) becomes:

$$\dot{x} = g_0(x) + \sum_{i=1}^m u_i g_i(x) = g_0(x) + G(x)u\,.$$

(See Sect. 8.5 in [25] for general f, and for proofs). We now pick two continuous functions $Q\colon \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ and $R\colon \mathbb{R}^n \to \mathbb{R}^{n\times n}$, so that $R(x)$ is a symmetric positive definite matrix for each x.

In general, we say that a continuous function

$$V\colon \mathbb{R}^n \to \mathbb{R}_{\geq 0}$$

is *positive definite* if $V(x) = 0$ only if $x = 0$, and it is *proper* (or "weakly coercive") if for each $a \geq 0$ the set $\{x|V(x) \leq a\}$ is compact, or, equivalently, $V(x) \to \infty$ as $|x| \to \infty$ (radial unboundedness). Given any such V which is also differentiable, we denote the vector function whose components are the directional derivatives of V in the directions of the various control vector fields g_i, $i \geq 1$ by:

$$L_G V(x) := \nabla V(x) G(x) = \left(L_{g_1} V(x), \ldots, L_{g_m} V(x)\right) ,$$

and also write $L_{g_0} V(x) := \nabla V(x) g_0(x)$.

We consider the following PDE on such functions V:

$$\forall x \quad Q(x) + L_{g_0} V(x) - \frac{1}{4} L_G V(x) R(x)^{-1} \left(L_G V(x)\right)' = 0. \tag{8}$$

This reduces to the Algebraic Riccati Eq. (7) in the special case of linear systems, quadratic $V(x) = x'Px$ and $Q(x) = x'Qx$, and constant matrices $R(x) \equiv R$. We also take the following generalization of the feedback law (6):

$$k(x) := -\frac{1}{2} R(x)^{-1} \left(L_G V(x)\right)' . \tag{9}$$

Finally, we assume that Q is a positive definite function. One then has [25]:

Theorem 2 *Suppose that V is a twice continuously differentiable, positive definite, and proper solution of the PDE (8). Then, k defined by (9) stabilizes the system.*

This theorem arises from the following optimization problem: for each state $x^0 \in \mathbb{R}^n$, minimize the cost

$$\mathcal{J}_{x^0}(u) := \int_0^\infty u(t)' R(x(t)) u(t) + Q(x(t)) \mathrm{d}t ,$$

where $x(t) = x(t, x^0, u)$, over all those controls $u\colon [0, \infty) \to \mathcal{U}$ for which the solution $x(t, x^0, u)$ of (1) is defined for all $t \geq 0$ and satisfies $\lim_{t\to\infty} x(t) = 0$. Under the above assumptions, and as for linear systems, one also concludes that for each state x^0 the solution of $\dot{x} = f(x, k(x))$ with initial state $x(0) = x^0$ exists for all $t \geq 0$, the control $u(t) = k(x(t))$ is optimal, and $V(x^0)$ is the optimal cost from initial state x^0. Moreover, the formula for k arises from the Hamilton–Jacobi–Bellman equation of optimal control theory, because

$$k(x) = \operatorname*{argmin}_u \left\{\nabla V(x) \cdot f(x, u) + u' R(x) u + Q(x)\right\}$$

when $f(x, u) = g_0(x) + G(x)u$.

There are applications where this method has proven useful. Unfortunately, however, and in contrast to the linear case, in general there exists no positive definite, proper, and C^2 solution V of the above PDE. On the other hand, the formula (9) does appear, with variations, in other contexts, including generalizations of the idea of adding damping to systems, cf. Sect. 5.9 in [25], and, more generally, the use of auxiliary positive definite and proper functions V, in similar roles, will be central to the control-Lyapunov ideas discussed later.

Nonlinear Systems: Continuous Feedback

One of the central topics which we will address here concerns possibly discontinuous feedback laws k. Before turning to that subject, however, we study continuous feedback. When dealing with linear systems, linear feedback is natural, and indeed sufficient from a theoretical standpoint, as shown by the results just reviewed. However, for our general study, major technical questions arise in even deciding on just what degree of regularity should be imposed on the feedback maps k.

It turns out that the precise requirements away from 0, say asking whether k is merely continuous or smooth, are not very critical; it is often the case that one can "smooth out" a continuous feedback (or, even, make it real-analytic, via Grauert's Theorem) away from the origin. So, in order to avoid unnecessary complications in exposition due to nonuniqueness, let us call a feedback k *regular* if it is locally Lipschitz on $\mathbb{R}^n \setminus \{0\}$. For such k, solutions of initial value problems $\dot{x} = f(x, k(x))$, $x(0) = x^0$, are well defined (at least for small time intervals $[0, \varepsilon)$) and, provided k is a stabilizing feedback, are unique (cf. Exercise 5.9.9 in [25]).

On the other hand, behavior at the origin cannot be "smoothed out" and, at zero, the precise degree of smoothness plays a central role in the theory [12]. For instance, consider the system

$$\dot{x} = x + u^3 .$$

The continuous (and, in fact, smooth away from zero) feedback $u = k(x) := -\sqrt[3]{2x}$ globally stabilizes the system (the closed-loop system becomes $\dot{x} = -x$). However, there is no possible stabilizing feedback which is differentiable at the origin, since $u = k(x) = O(x)$ implies that

$$\dot{x} = x + O\left(x^3\right)$$

about $x = 0$, which means that the solution starting at any positive and small point moves to the right, instead of towards the origin. (A general result, assuming that A has no purely imaginary eigenvalues, cf. [25], Section 5.8, is that if – and only if – $\dot{x} = Ax + Bu + o(x, u)$ can be locally asymptotically stabilized using a feedback which is differentiable at the origin, the linearization $\dot{x} = Ax + Bu$ must be AC itself. In the example that we gave, this linearization is just $\dot{x} = x$, which is not AC.)

We now turn to the question of existence of regular feedback stabilizers. We first study a comparatively trivial case, namely systems with one state variable and one input. After that, we turn to multidimensional systems.

The Special Case $n = m = 1$

There are algebraic obstructions to the stabilization of $\dot{x} = f(x, u)$ if the input u appears nonlinearly in f. Ignoring the requirement that there be a $\sigma \in \mathcal{N}$ so that controls can be picked with $\|u\| \leq \sigma(|x^0|)$, asymptotic controllability is, for $n = m = 1$, equivalent to:

$$(\forall x \neq 0)(\exists u)\ xf(x, u) < 0 \tag{10}$$

(this is proved in [28]; it is fairly obvious, but some care must be taken to deal with the fact that one is allowing arbitrary measurable controls; the argument proceeds by first approximating such controls by piecewise constant ones). Let us introduce the following set:

$$\mathcal{O} := \{(x, u) | xf(x, u) < 0\} \ ,$$

and let $\pi : (x, u) \mapsto x$ be the projection into the first coordinate in $\mathbb{R}$. Then, (10) is equivalent to:

$$\pi\mathcal{O} = \mathbb{R} \setminus \{0\} \ .$$

(One can easily include the requirement "$\|u\| \leq \sigma(|x^0|)$" by asking that for each interval $[-K, K] \subset \mathbb{R}$ there must be some compact set $C_K \subset \mathbb{R}^2$ so that $[-K, K] \subseteq \pi(C_K)$. For simplicity, we ignore this technicality.) In these terms, a stabilizing feedback is nothing else than a locally bounded map $k\colon \mathbb{R} \to \mathbb{R}$ such that $k(0) = 0$ and so that k is a section of π on $\mathbb{R} \setminus \{0\}$:

$$(x, k(x)) \in \mathcal{O} \ \ \forall x \neq 0 \ .$$

For a regular feedback, we ask that k be locally Lipschitz on $\mathbb{R} \setminus \{0\}$.

Clearly, there is no reason for Lipschitz, or for that matter, just continuous, sections of π to exist. As an illustration, take the system

$$\dot{x} = x\left[(u-1)^2 - (x-1)\right]\left[(u+1)^2 + (x-2)\right] \ .$$

Let

$$\mathcal{O}_1 = \left\{(u+1)^2 < (2-x)\right\} \text{ and } \mathcal{O}_2 = \{(u-1)^2 < (x-1)\}$$

(these are the interiors of two disjoint parabolas). Here, $\mathcal{O}$ has three connected components, namely $\mathcal{O}_2$, $\mathcal{O}_1$ intersected with $x < 0$, and $\mathcal{O}_1$ intersected with $x < 0$. It is clear that, even though $\pi\mathcal{O} = \mathbb{R}$, there is no continuous curve (graph of $u = k(x)$) which is always in $\mathcal{O}$ and projects onto $\mathbb{R} \setminus \{0\}$. On the other hand, there exist many possible feedback stabilizers provided that we allow one discontinuity. It is also possible to provide examples, even with $f(x, u)$ smooth, for which an infinite number of switches are needed in any possible stabilizing feedback law. Finally, it may even be possible to stabilize semiglobally with a regular feedback, meaning that for each compact subset K of the state-space there is a regular, even smooth, feedback law $u = k_K(x)$ such that all states in K get driven asymptotically to the origin, but yet it may be impossible to find a single $u = k(x)$ which works globally. See [26] for details.

When feedback laws are required to be continuous at the origin, new obstructions arise. The case of systems with $n = m = 1$ is also a good way to introduce this subject. The first observation is that stabilization about the origin (even if just local) means that we must have, near zero:

$$f(x, k(x)) \begin{cases} > 0 & \text{if} \quad x < 0 \\ < 0 & \text{if} \quad x > 0 \\ = 0 & \text{if} \quad x = 0 \end{cases} .$$

In fact, all that we need is that $f(x_1, k(x_1)) < 0$ for some $x_1 > 0$ and $f(x_2, k(x_2)) > 0$ for some $x_2 < 0$. This guarantees, via the intermediate-value theorem that, if k is continuous, the projection

$$(-\varepsilon, \varepsilon) \to \mathbb{R} \ , \quad x \mapsto f(x, k(x))$$

is onto a neighborhood of zero, for each $\varepsilon > 0$. It follows, in particular, that the image of

$$(-\varepsilon, \varepsilon) \times (-\varepsilon, \varepsilon) \to \mathbb{R} \ , \quad (x, u) \mapsto f(x, u)$$

also contains a neighborhood of zero, for any $\varepsilon > 0$ (that is, the map $(x, u) \mapsto f(x, u)$ is open at zero). This last property is intrinsic, being stated in terms of the original data $f(x, u)$ and not depending upon the feedback k. Brockett's condition, to be described next, is a far-reaching generalization of this argument; in its proof, degree theory replaces the use of the intermediate value theorem.

Obstructions and Necessary Degree Conditions

If there are "obstacles" in the state-space, or more precisely if the state-space is a proper subset of $\mathbb{R}^n$, discontinuities in feedback laws cannot in general be avoided,

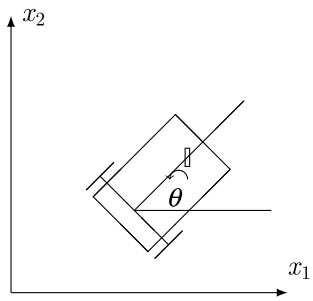

Stability and Feedback Stabilization, Figure 1
Shopping cart

since the domain of attraction of an asymptotically stable vector field must be diffeomorphic to Euclidean space. But even if states evolve in Euclidean spaces, similar obstructions may arise. These are due not to the topology of the state space, but to "virtual obstacles" implicit in the form of the system equations. These obstacles occur when it is impossible to move instantaneously in certain directions, even if it is possible to move eventually in every direction, the phenomenon of "nonholonomy". As an illustration, let us consider a model for the "shopping cart" shown in Fig. 1 ("knife-edge" or "unicycle" are other names for this example). The state is given by the orientation θ, together with the coordinates x_1, x_2 of the midpoint between the back wheels. The front wheel is a castor, free to rotate. There is a non-slipping constraint on movement: the velocity $(\dot{x}_1, \dot{x}_2)'$ must be parallel to the vector $(\cos\theta, \sin\theta)'$. This leads to the following equations:

$$\begin{aligned}\dot{x}_1 &= u_1 \cos\theta\\ \dot{x}_2 &= u_1 \sin\theta\\ \dot{\theta} &= u_2\end{aligned}$$

where we may view u_1 as a "drive" command and u_2 as a steering control. (In practice, one would implement these controls by means of differential forces on the two back corners of the cart.) The feedback transformation $z_1 := \theta$, $z_2 := x_1 \cos\theta + x_2 \sin\theta$, $z_3 := x_1 \sin\theta - x_2 \cos\theta$, $v_1 := u_2$, and $v_2 := u_1 - u_2 z_3$ brings the system into the system with equations $\dot{z}_1 = v_1, \dot{z}_2 = v_2, \dot{z}_3 = z_2 v_1$ known as "Brockett's example" or "nonholonomic integrator" (yet another change can bring the third equation into the form $\dot{z}_3 = z_1 v_2 - z_2 v_1$). We view the system as having state space $\mathbb{R}^3$. Although a physically more accurate state space would be the manifold $\mathbb{R}^2 \times \mathbb{S}^1$, the necessary condition to be given is of a local nature, so the global structure is unimportant.

This system is (obviously) completely controllable (formally, controllability can be checked using the Lie algebra rank condition, as in e. g. [25], Exercise 4.3.16), and in particular is AC. However, we may expect that discontinuities are unavoidable due to the non-slip constraint, which does not allow moving from, for example the position $x_1 = 0$, $\theta = 0$, $x_2 = 1$ in a straight line towards the origin. Indeed, one then has [3]:

Theorem 3 *If there is a stabilizing feedback which is regular and continuous at zero, then the map $(x, u) \mapsto f(x, u)$ is open at zero.*

The test fails here, since no points of the form $(0, \varepsilon, *)$ belong to the image of the map

$$\begin{aligned}\mathbb{R}^5 \to \mathbb{R}^3 \colon (x_1, x_2, \theta, u_1, u_2)' \mapsto f(x, u)\\ = (u_1 \cos\theta, u_1 \sin\theta, u_2)'\end{aligned}$$

for $\theta \in (-\pi/2, \pi/2)$, unless $\varepsilon = 0$.

More generally, it is impossible to continuously stabilize any system without drift

$$\dot{x} = u_1 g_1(x) + \ldots + u_m g_m(x) = G(x)u$$

if $m < n$ and $\operatorname{rank}[g_1(0), \ldots, g_m(0)] = m$ (this includes all totally nonholonomic mechanical systems). Indeed, under these conditions, the map $(x, u) \mapsto G(x)u$ cannot contain a neighborhood of zero in its image, when restricted to a small enough neighborhood of zero. Indeed, let us first rearrange the rows of G:

$$G(x) \rightsquigarrow \begin{pmatrix} G_1(x) \\ G_2(x) \end{pmatrix}$$

so that $G_1(x)$ is of size $m \times m$ and is nonsingular for all states x that belong to some neighborhood N of the origin. Then,

$$\begin{pmatrix} 0 \\ a \end{pmatrix} \in \operatorname{Im}\left[N \times \mathbb{R}^m \to \mathbb{R}^n \colon (x, u) \mapsto G(x)u\right] \Rightarrow a = 0$$

(since $G_1(x)u = 0 \Rightarrow u = 0 \Rightarrow G_2(x)u = 0$ too).

If the condition $\operatorname{rank}[g_1(0), \ldots, g_m(0)] = m$ is violated, we cannot conclude a negative result. For instance, the system $\dot{x}_1 = x_1 u$, $\dot{x}_2 = x_2 u$ has $m = 1 < 2 = n$ but it can be stabilized by means of the feedback law $u = -(x_1^2 + x_2^2)$.

Observe that for linear systems, Brockett's condition says that

$$\operatorname{rank}[A, B] = n$$

which is the Hautus controllability condition (see e. g. [25], Lemma 3.3.7) at the zero mode.

Idea of the Proof One may prove Brockett's condition in several ways. A proof based on degree theory is probably easiest, and proceeds as follows (for details see for instance [25], Sect 5.9). The basic fact, due to Krasnosel'ski, is that if the system $\dot{x} = F(x) = f(x, k(x))$ has the origin as an asymptotically stable point and F is regular (since k is), then the degree (index) of F with respect to zero is $(-1)^n$, where n is the system dimension. In particular, the degree is also nonzero with respect to points p in a neighborhood of 0, which means that the equation $F(x) = p$ can be solved for small p, and hence $f(x, u) = p$ can be solved as well. The proof that the degree is $(-1)^n$ follows by exhibiting a homotopy, namely

$$F_t(x^0) = \frac{1}{t}\left[x\left(\frac{t}{1-t}, x^0\right) - x^0\right],$$

between $F_0 = F$ and $F_1(x) = -x$, and noting that the degree of the latter is obviously $(-1)^n$. An alternative proof uses Lyapunov functions. Asymptotic stability implies the existence of a smooth Lyapunov function V for $\dot{x} = F(x) = f(x, k(x))$, so, on the boundary ∂B of a sublevel set $B = \{x | V(x) \le c\}$ we have that F points towards the interior of B. Thus, for p small, $F(x) - p$ still points to the interior, which means that B is invariant with respect to the perturbed vector field $\dot{x} = F(x) - p$. Provided that a fixed-point theorem applies to continuous maps $B \to B$, this implies that $F(x) - p$ must vanish somewhere in B, that is, the equation $F(x) = p$ can be solved. (Because, for each small $h > 0$, the time-h flow ϕ of $F - p$ has a fixed point $x_h \in B$, i. e. $\phi(h, x_h) = x_h$, so picking a convergent subsequence $x_h \to \bar{x}$ gives that $0 = (\phi(h, x_h) - x_h)/(h) \to F(\bar{x}) - p$.) A fixed point theorem can indeed be applied, because B is a retract of $\mathbb{R}^n$ (use the flow itself); note that this argument gives a weaker conclusion than the degree condition.

Control-Lyapunov Functions

The method of control-Lyapunov functions ("clf's") provides a powerful tool for studying stabilization problems, both as a basis of theoretical developments and as a method for actual feedback design.

Before discussing clf's, let us quickly review the classical concept of Lyapunov functions, through a simple example. Consider first a damped spring-mass system $\ddot{y} + \dot{y} + y = 0$, or, in state-space form with $x_1 = y$ and $x_2 = \dot{y}$, $\dot{x}_1 = x_2$, $\dot{x}_2 = -x_1 - x_2$. One way to verify global asymptotic stability of the equilibrium $x = 0$ is to pick the (Lyapunov) function $V(x_1, x_2) := \frac{3}{2}x_1^2 + x_1x_2 + x_2^2$, and observe that $\nabla V(x).f(x) = -|x|^2 < 0$ if $x \neq 0$, which means that

$$\frac{\mathrm{d}V(x(t))}{\mathrm{d}t} = -\left|x(t)^2\right| < 0$$

along all nonzero solutions, and thus the energy-like function V decreases along all trajectories, which, since V is a nondegenerate quadratic form, implies that $x(t)$ decreases, and in fact $x(t) \to 0$. Of course, in this case one could compute solutions explicitly, or simply note that the characteristic equation has all roots with negative real part, but Lyapunov functions are a general technique. (In fact, the classical converse theorems of Massera and Kurzweil [17,19] show that, whenever a system is GAS, there always exists a smooth Lyapunov function V.)

Now let us modify this example to deal with a control system, and consider a forced (but undamped) harmonic oscillator $\ddot{x} + x = u$, i. e. $\dot{x}_1 = x_2$, $\dot{x}_2 = -x_1 + u$. The damping feedback $u = -x_2$ stabilizes the system, but let us pretend that we do not know that. If we take the same V as before, now the derivatives along trajectories are, using "$\dot{V}(x, u)$" to denote $\nabla V(x).f(x, u)$ and omitting arguments t in $x(t)$ and $u(t)$:

$$\dot{V}(x, u) = -x_1^2 + x_1x_2 + x_2^2 - (x_1 + 2x_2)u\,.$$

This expression is affine in u. Thus, if x is a state such that $x_1 + 2x_2 \neq 0$, then we may pick a control value u (which depends on this current state x) such that $\dot{V} < 0$. On the other hand, if $x_1 + 2x_2 = 0$, then the expression reduces to $\dot{V} = -5x_2^2$ (for any u), which is negative unless x_2 (and hence also $x_1 = -2x_2$) vanishes.

In conclusion, for each $x \neq 0$ there is some u so that $\dot{V}(x, u) < 0$. This is, except for some technicalities to be discussed, the characterizing property of control-Lyapunov functions. For any given compact subset B in $\mathbb{R}^n$, we now pick some compact subset $\mathcal{U}_0 \subset \mathcal{U}$ so that

$$\forall x \in B, x \neq 0\,, \quad \exists u \in \mathcal{U}_0 \quad \text{such that} \quad \dot{V}(x, u) < 0\,. \tag{11}$$

In principle, then, we could then stabilize the system, for states in B, by using the steepest descent feedback law:

$$k(x) := \operatorname*{argmin}_{u \in \mathcal{U}_0} \nabla V(x) \cdot f(x, u) \tag{12}$$

("argmin" means "pick any u at which the min is attained"; we restricted $\mathcal{U}$ to be assured that $\dot{V}(x, u)$ attains a minimum). Note that the stabilization problem becomes, in these terms, a set of static nonlinear programming problems: minimize a function of u, for each x. Global stabilization is also possible, by appropriately picking $\mathcal{U}_0$ as a func-

tion of the norm of x; later we discuss a precise formulation.

Control–Lyapunov functions, if understood non-technically as the basic paradigm "look for a function $V(x)$ with the properties that $V(x) \approx 0$ if and only if $x \approx 0$, and so that for each $x \neq 0$ it is possible to decrease $V(x)$ by some control action," constitute a very general approach to control (sometimes expressed in a dual fashion, as maximization of some measure of success). They appear in such disparate areas as A.I. game-playing programs (position evaluations), energy arguments for dissipative systems, program termination (Floyd/Dijkstra "variant"), and learning control ("critics" implemented by neural-networks). More relevantly to this paper, the idea underlies much of modern feedback control design, as illustrated for instance by the books [7,11,15,16,25].

Differentiable clf's: Precise Definition A *differentiable control-Lyapunov function* (clf) is a differentiable function $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ which is proper, positive definite, and *infinitesimally decreasing*, meaning that there exists a positive definite continuous function $W : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$, and there is some $\sigma \in \mathcal{N}$, so that

$$\sup_{x \in \mathbb{R}^n} \min_{|u| \leq \sigma(|x|)} \nabla V(x) \cdot f(x,u) + W(x) \leq 0 \,. \tag{13}$$

This is basically the same as condition (11), with $\mathcal{U}_0 =$ the ball of radius $\sigma(|x|)$ picked as a function of x. The main difference is that, instead of saying "$\nabla V(x) \cdot f(x,u) < 0$ for $x \neq 0$" we write $\nabla V(x) \cdot f(x,u) \leq -W(x)$, where W is negative when $x \neq 0$. The two definitions are equivalent, but the "Hamiltonian" version used here is the correct one for the generalizations to be given, to nonsmooth V.

The basic result is due to Artstein [2]:

Theorem 4 *A control-affine system $\dot{x} = g_0(x) + \sum u_i g_i(x)$ admits a differentiable clf if and only if it admits a regular stabilizing feedback.*

The proof of sufficiency is easy: if there is such a k, then the converse Lyapunov theorem, applied to the closed-loop system $F(x) = f(x, k(x))$, provides a smooth V such that

$$L_F V(x) = \nabla V(x) F(x) < 0 \quad \forall x \neq 0 \,.$$

This gives that for all nonzero x there is some u (bounded on bounded sets, because k is locally bounded by definition of feedback) so that $\dot{V}(x,u) < 0$; and one can put this in the form (13).

The necessity is more interesting. The original proof in [2] proceeds by a nonconstructive argument involving partitions of unity, but it is also possible [24,25] to exhibit explicitly a feedback, written as a function:

$$k\left(\nabla V(x) \cdot g_0(x), \ldots, \nabla V(x) \cdot g_m(x)\right)$$

of the directional derivatives of V along the vector fields defining the system (*universal formulas* for stabilization). Taking for simplicity $m = 1$, one such formula is:

$$k(x) := -\frac{a(x) + \sqrt{a(x)^2 + b(x)^4}}{b(x)} \quad (0 \text{ if } b = 0)$$

where $a(x) := \nabla V(x) \cdot g_0(x)$ and $b(x) := \nabla V(x) \cdot g_1(x)$. The expression for k is analytic in a,b when $x \neq 0$, because the clf property means that $a(x) < 0$ whenever $b(x) = 0$ [24,25].

Thus, the question of existence of regular feedback, for control-affine systems, reduces to the search for differentiable clf's, and this gives rise to a vast literature dealing with the construction of such V's, see [7,15,16,25] and references therein. Many other theoretical issues are also answered by Artstein's theorem. For example, via Kurzweil's converse theorem one has that the existence of k merely continuous on $\mathbb{R}^n \setminus \{0\}$ suffices for the existence of smooth (infinitely differentiable) V, and from here one may in turn find a k which is smooth on $\mathbb{R}^n \setminus \{0\}$. In addition, one may easily characterize the existence of k continuous at zero as well as regular: this is equivalent to the *small control property*: for each $\varepsilon > 0$ there is some $\delta > 0$ so that $0 < |x| < \delta$ implies that $\min_{|u| \leq \varepsilon} \nabla V(x) \cdot f(x,u) < 0$ (if this property holds, the universal formula automatically provides such a k). We should note that Artstein provided a result valid for general, not necessarily control-affine systems $\dot{x} = f(x,u)$; however, the obtained "feedback" has values in sets of relaxed controls, and is not a feedback law in the classical sense. Later, we discuss a different generalization.

Differentiable clf's will in general not exist, because of obstructions to regular feedback stabilization. This leads us naturally into the twin subjects of discontinuous feedbacks and non-differentiable clf's.

Discontinuous Feedback

The previous results and examples show that, in order to develop a satisfactory general theory of stabilization, one in which one proves the implication "asymptotic controllability implies feedback stabilizability," we must allow discontinuous feedback laws $u = k(x)$. But then, a major technical difficulty arises: solutions of the initial-value problem $\dot{x} = f(x, k(x))$, $x(0) = x^0$, interpreted in the classical sense of differentiable functions or even as (absolutely) continuous solutions of the integral equa-

tion $x(t) = x^0 + \int_0^t f(x(s), k(x(s)))\mathrm{d}s$, do not exist in general. The only general theorems apply to systems $\dot{x} = F(x)$ with continuous F. For example, there is no solution to $\dot{x} = -\operatorname{sign} x$, $x(0) = 0$, where $\operatorname{sign} x = -1$ for $x < 0$ and $\operatorname{sign} x = 1$ for $x \geq 0$. So one cannot even pose the stabilization problem in a mathematically consistent sense.

There is, of course, an extensive literature addressing the question of discontinuous feedback laws for control systems and, more generally, differential equations with discontinuous right-hand sides. One of the best-known candidates for the concept of solution of (3) is that of a *Filippov solution* [6,9], which is defined as the solution of a certain differential inclusion with a multivalued right-hand side which is built from $f(x, k(x))$. Unfortunately, there is no hope of obtaining the implication "asymptotic controllability implies feedback stabilizability" if one interprets solutions of (3) as Filippov solutions. This is a consequence of results in [5,22], which established that the existence of a discontinuous stabilizing feedback in the Filippov sense implies the Brockett necessary conditions, and, moreover, for systems affine in controls it also implies the existence of regular feedback (which we know is in general impossible).

A different concept of solution originates with the theory of discontinuous positional control developed by Krasovskii and Subbotin in the context of differential games in [14], and it is the basis of the new approach to discontinuous stabilization proposed in [4], to which we now turn.

Limits of High-Frequency Sampling

By a *sampling schedule* or *partition* $\pi = \{t_i\}_{i\geq 0}$ of $0, +\infty$ we mean an infinite sequence

$$0 = t_0 < t_1 < t_2 < \ldots$$

with $\lim_{i\to\infty} t_i = \infty$. We call

$$\mathbf{d}(\pi) := \sup_{i\geq 0}(t_{i+1} - t_i)$$

the *diameter* of π. Suppose that k is a given feedback law for system (1). For each π, the *π-trajectory starting from* x^0 of system (3) is defined recursively on the intervals $[t_i, t_{i+1})$, $i = 0, 1, \ldots$, as follows. On each interval $t_i, t_{i+1})$, the initial state is measured, the control value $u_i = k(x(t_i))$ is computed, and the constant control $u \equiv u_i$ is applied until time t_{i+1}; the process is then iterated. That is, we start with $x(t_0) = x^0$ and solve recursively

$$\dot{x}(t) = f(x(t), k(x(t_i)))\,, t \in t_i, t_{i+1})\,, \quad i = 0, 1, 2, \ldots$$

using as initial value $x(t_i)$ the endpoint of the solution on the preceding interval. The ensuing π-trajectory, which we denote as $x_\pi(\cdot, x^0)$, is defined on some maximal nontrivial interval; it may fail to exist on the entire interval $[0, +\infty)$ due to a blow-up on one of the subintervals $t_i, t_{i+1})$. We say that it is *well defined* if $x_\pi(t, x^0)$ is defined on all of $[0, +\infty)$.

Definition The feedback $k\colon \mathbb{R}^n \to \mathcal{U}$ *stabilizes* the system (1) if there exists a function $\beta \in \mathcal{KL}$ so that the following property holds: For each

$$0 < \varepsilon < K$$

there exists a $\delta = \delta(\varepsilon, K) > 0$ such that, for every sampling schedule π with $\mathbf{d}(\pi) < \delta$, and for each initial state x^0 with $|x^0| \leq K$, the corresponding π-trajectory of (3) is well-defined and satisfies

$$\left|x_\pi(t, x^0)\right| \leq \max\{\beta\,(K, t)\,, \varepsilon\} \quad \forall t \geq 0\,. \tag{14}$$

In particular, we have

$$\left|x_\pi(t, x^0)\right| \leq \max\left\{\beta\left(\left|x^0\right|, t\right), \varepsilon\right\} \quad \forall t \geq 0 \tag{15}$$

whenever $0 < \varepsilon < |x^0|$ and $\mathbf{d}(\pi) < \delta(\varepsilon, |x^0|)$ (just take $K := |x^0|$).

Observe that the role of δ is to specify a lower bound on intersampling times. Roughly, one is requiring that

$$t_{i+1} \leq t_i + \theta\,(|x(t_i)|)$$

for each i, where θ is an appropriate positive function.

Our definition of stabilization is physically meaningful, and is very natural in the context of sampled-data (computer control) systems. It says in essence that a feedback k stabilizes the system if it drives all states asymptotically to the origin and with small overshoot when using *any fast enough sampling schedule*. A high enough sampling frequency is generally required when close to the origin, in order to guarantee small displacements, and also at infinity, so as to preclude large excursions or even blow-ups in finite time. This is the reason for making δ depend on ε and K.

This concept of stabilization can be reinterpreted in various ways. One is as follows. Pick any initial state x^0, and consider any sequence of sampling schedules π_ℓ whose diameters $\mathbf{d}(\pi_\ell)$ converge to zero as $\ell \to \infty$ (for instance, constant sampling rates with $t_i = i/\ell$, $i = 0, 1, 2, \ldots$). Note that the functions $x_\ell := x_{\pi_\ell}(\cdot, x^0)$ remain in a bounded set, namely the ball of radius $\beta(|x^0|, 0)$ (at least for ℓ large enough, for instance, any ℓ so that $\mathbf{d}(\pi_\ell)\delta(|x^0|/2, |x^0|)$). Because $f(x, k(x))$ is bounded

on this ball, these functions are equicontinuous, and (Arzela–Ascoli's Theorem) we may take a subsequence, which we denote again as $\{x_\ell\}$, so that $x_\ell \to x$ as $\ell \to \infty$ (uniformly on compact time intervals) for some absolutely continuous (even Lipschitz) function $x\colon [0,\infty) \to \mathbb{R}^n$. *We may think of any limit function $x(\cdot)$ that arises in this fashion as a generalized solution of the closed-loop Eq.* (3). That is, generalized solutions are the limits of trajectories arising from arbitrarily high-frequency sampling when using the feedback law $u = k(x)$. Generalized solutions, for a given initial state x^0, may not be unique – just as may happen with continuous but non-Lipschitz feedback – but there is always existence, and, moreover, for any generalized solution, $|x(t)| \le \beta(|x^0|, t)$ for all $t \ge 0$. This is precisely the defining estimate for the GAS property. Moreover, if k happens to be regular, then the unique solution of $\dot{x} = f(x, k(x))$ in the classical sense is also the unique generalized solution, so we have a reasonable extension of the concept of solution. (This type of interpretation is somewhat analogous, at least in spirit, to the way in which "relaxed" controls are interpreted in optimal trajectory calculations, namely through high-frequency switching of approximating regular controls.) The definition of stabilization was given in [4] in a slightly different form; see [26] for a discussion of the equivalence.

Stabilizing Feedbacks Exist

The main result is [4]:

Theorem 5 *The system (1) admits a stabilizing feedback if and only if it is asymptotically controllable.*

Necessity is clear. The sufficiency statement is proved by construction of k, and is based on the following ingredients:

- Existence of a nonsmooth control-Lyapunov function V.
- Regularization on shells of V.
- Pointwise minimization of a Hamiltonian for the regularized V.

In order to sketch this construction, we start by quickly reviewing a basic concept from nonsmooth analysis.

Proximal Subgradients Let V be any continuous function $\mathbb{R}^n \to \mathbb{R}$ (or even, just lower semicontinuous and with extended real values). A *proximal subgradient* of V at the point $x \in \mathbb{R}^n$ is any vector $\zeta \in \mathbb{R}^n$ such that, for some $\sigma > 0$ and some neighborhood $\mathcal{O}$ of x,

$$V(y) \ge V(x) + \zeta \cdot (y - x) - \sigma^2 \left| y - x^2 \right| \quad \forall y \in \mathcal{O} .$$

In other words, proximal subgradients are the possible gradients of supporting quadratics at the point x. The set of all proximal subgradients at x is denoted $\partial_p V(x)$.

Nonsmooth Control–Lyapunov Functions A continuous (but not necessarily differentiable) $V\colon \mathbb{R}^n \to \mathbb{R}_{\ge 0}$ is a *control-Lyapunov function* (clf) if it is proper, positive definite, and infinitesimally decreasing in the following generalized sense: there exist a positive definite continuous $W\colon \mathbb{R}^n \to \mathbb{R}_{\ge 0}$ and a $\sigma \in \mathcal{N}$ so that

$$\sup_{x \in \mathbb{R}^n} \max_{\zeta \in \partial_p V(x)} \min_{|u| \le \sigma(|x|)} \zeta \cdot f(x,u) + W(x) \le 0 . \quad (16)$$

This is the obvious generalization of the differentiable case in (13); we are still asking that one should be able to make $\nabla V(x) \cdot f(x,u) < 0$ by an appropriate choice of $u = u_x$, for each $x \ne 0$, except that now we replace $\nabla V(x)$ by the proximal subgradient set $\partial_p V(x)$. An equivalent property is to ask that V be a viscosity supersolution of the corresponding Hamilton–Jacobi–Bellman equation.

For nonsmooth clf's, the main basic result is [23,26]:

Theorem 6 *The system (1) is asymptotically controllable if and only if it admits a continuous clf.*

The proof is based on first constructing an appropriate W, and then letting V be the optimal cost (Bellman function) for the problem $\min \int_0^\infty W(x(s))\mathrm{d}s$. However, some care has to be taken to insure that V is continuous, and the cost has to be adjusted in order to deal with possibly unbounded minimizers. An important and very useful refinement of this result is the fact that a locally Lipschitz clf can also be shown to exist [21].

Regularization Once V is known to exist, the next step in the construction of a stabilizing feedback is to obtain Lipschitz approximations of V. For this purpose, one considers the Iosida–Moreau inf-convolution of V with a quadratic function:

$$V_\alpha(x) := \inf_{y \in \mathbb{R}^n} \left[V(y) + \frac{1}{2\alpha^2} |y - x|^2 \right]$$

where the number $\alpha > 0$ is picked constant on appropriate regions. One has that $V_\alpha(x) \nearrow V(x)$, uniformly on compacts. Since V_α is locally Lipschitz, Rademacher's Theorem insures that V_α is differentiable almost everywhere. The feedback k is then made equal to a pointwise minimizer k_α of the Hamiltonian, at the points of differentiability (compare with (12) for the case of differentiable V):

$$k_\alpha(x) := \operatorname*{argmin}_{u \in \mathcal{U}_0} \nabla V_\alpha(x) \cdot f(x,u) ,$$

where α and the compact $\mathcal{U}_0 = \mathcal{U}_0(\alpha)$ are chosen constant on certain compacts and this choice is made in between level curves. The critical fact is that V_α is itself a clf for the original system, at least when restricted to the region where it is needed. More precisely, on each shell of the form

$$C = \{x \in \mathbb{R}^n | r \leq |x| \leq R\} ,$$

there are positive numbers m and α_0 and a compact subset $\mathcal{U}_0$ such that, for each $0 < \alpha \leq \alpha_0$, each $x \in C$, and every $\zeta \in \partial_{\mathrm{p}} V_\alpha(x)$,

$$\min_{u \in \mathcal{U}_0} \zeta \cdot f(x, u) + m \leq 0 .$$

See [26]. (Actually, this description is oversimplified, and the proof is a bit more delicate. One must define, on appropriate compact sets

$$k(x) := \operatorname*{argmin}_{u \in \mathcal{U}_0} \zeta_\alpha(x) \cdot f(x, u) ,$$

where $\zeta_\alpha(x)$ is carefully chosen. At points x of nondifferentiability, $\zeta_\alpha(x)$ is not a proximal subgradient of V_α, since $\partial_{\mathrm{p}} V_\alpha(x)$ may well be empty. One uses, instead, the fact that $\zeta_\alpha(x)$ happens to be in $\partial_{\mathrm{p}} V(x')$ for some $x' \approx x$. See [4] for details.)

Sensitivity to Small Measurement Errors

We have seen that every asymptotically controllable system admits a feedback stabilizer k, generally discontinuous, which renders the closed-loop system $\dot{x} = f(x, k(x))$ GAS. On the other hand, one of the main motivations for the use of feedback is in order to deal with uncertainty, and one possible source of uncertainty are measurement errors in state estimation. The use of discontinuous feedback means that undesirable behavior – chattering – may arise. In fact, one of the main reasons for the focus on continuous feedback is precisely in order to avoid such behaviors. Thus, we turn now to an analysis of the effect of measurement errors. Suppose first that k is a continuous function of x. Then, if the error e is small, using the control $u' = k(x + e)$ instead of $u = k(x)$ results in behavior which remains close to the intended one, since $k(x + e) \approx k(x)$; moreover, if $e \ll x$ then stability is preserved. This property of robustness to small errors when k is continuous can be rigorously established by means of a Lyapunov proof, based on the observation that, if V is a Lyapunov function for the closed-loop system, then continuity of $f(x, k(x + e))$ on e means that

$$\nabla V(x) \cdot f(x, k(x + e)) \approx \nabla V(x) \cdot f(x, k(x)) < 0 .$$

Unfortunately, when k is not continuous, this argument breaks down. However, it can be modified so as to avoid invoking continuity of k. Assuming that V is continuously differentiable, one can argue that

$$\nabla V(x) \cdot f(x, k(x + e)) \approx \nabla V(x + e) \cdot f(x, k(x + e)) < 0$$

(using the Lyapunov property at the point $x + e$ instead of at x). This observation leads to a theorem, formulated below, which says that a discontinuous feedback stabilizer, robust with respect to small observation errors, can be found provided that there is a C^1 clf. In general, as there are no C^1, but only continuous, clf's, one may not be able to find any feedback law that is robust in this sense.

There are many well-known techniques for avoiding chattering, and a very common one is the introduction of deadzones where no action is taken. The feedback constructed in [4], with no modifications needed, can always be used in a manner robust with respect to small observation errors, using such an approach. Roughly speaking, the general idea is as follows. Suppose that the true current state, let us say at time $t = t_i$, is x, but that the controller uses $u = k(\tilde{x})$, where $\tilde{x} = x + e$, and e is small. Call x' the state that results at the next sampling time, $t = t_{i+1}$. By continuity of solutions on initial conditions, $|x' - \tilde{x}'|$ is also small, where $\tilde{x}'$ is the state that would have resulted from applying the control u if the true state had been $\tilde{x}$. By continuity, it follows that $V_\alpha(x) \approx V_\alpha(\tilde{x})$ and also $V_\alpha(x') \approx V_\alpha(\tilde{x}')$. On the other hand, the construction in [4] provides that $V_\alpha(\tilde{x}') < V_\alpha(\tilde{x}) - \mathrm{d}(t_{i+1} - t_i)$, where d is some positive constant (this is valid while we are far from the origin). Hence, if e is sufficiently small compared to the intersample time $t_{i+1} - t_i$, it will necessarily be the case that $V_\alpha(x')$ must also be smaller than $V_\alpha(x)$. This discussion may be formalized in several ways; see [26] for a precise statement.

If we insist upon fast sampling, a necessary condition arises, as was proved in the recent paper [18] (which, in turn, represented an extension of the work by Hermes [10] for classical solutions under observation error). We next discuss the main result from that paper. We consider systems

$$\dot{x}(t) = f(x(t), k(x(t) + \mathrm{e}(t)) + \mathrm{d}(t)) \tag{17}$$

in which there are observation errors as well as, now, possible actuator errors $d(\cdot)$. Actuator errors $d(\cdot)\colon [0, \infty) \to \mathcal{U}$ are Lebesgue measurable and locally essentially bounded, and observation errors $\mathrm{e}(\cdot)\colon [0, \infty) \to \mathbb{R}^n$ are locally bounded. We define solutions of (17), for each sampling schedule π, in the usual manner, i. e., solving recursively on the intervals $t_i, t_{i+1})$, $i = 0, 1, \ldots$, the differen-

tial equation

$$\dot{x}(t) = f\big(x(t), k(x(t_i) + \mathrm{e}(t_i)) + \mathrm{d}(t)\big) \tag{18}$$

with $x(0) = x^0$. We write $x(t) = x_\pi(t, x^0, \mathrm{d}, \mathrm{e})$ for the solution, and say it is *well-defined* if it is defined for all $t \geq 0$.

Definition The feedback $k\colon \mathbb{R}^n \to \mathcal{U}$ *stabilizes* the system (17) if there exists a function $\beta \in \mathcal{KL}$ so that the following property holds: For each

$$0 < \varepsilon < K$$

there exist $\delta = \delta(\varepsilon, K) > 0$ and $\eta = \eta(\varepsilon, K)$ such that, for every sampling schedule π with $\mathbf{d}(\pi) < \delta$, each initial state x^0 with $|x^0| \leq K$, and each e, d such that $|\mathrm{e}(t)| \leq \eta$ for all $t \geq 0$ and $|\mathrm{d}(t)| \leq \eta$ for almost all $t \geq 0$, the corresponding π-trajectory of (17) is well-defined and satisfies

$$\left|x_\pi(t, x^0, \mathrm{d}, \mathrm{e})\right| \leq \max\{\beta(K, t), \varepsilon\} \quad \forall t \geq 0\,. \tag{19}$$

In particular, taking $K := |x^0|$, one has that

$$\left|x_\pi(t, x^0, d, e)\right| \leq \max\left\{\beta\left(\left|x^0\right|, t\right), \varepsilon\right\} \quad \forall t \geq 0$$

whenever $0 < \varepsilon < |x^0|$, $\mathbf{d}(\pi) < \delta(\varepsilon, |x^0|)$, and for all t, $|\mathrm{e}(t)| \leq \eta(\varepsilon, |x^0|)$, and $|\mathrm{d}(t)| \leq \eta(\varepsilon, |x^0|)$.

The main result in [18] is as follows.

Theorem 7 *There is a feedback which stabilizes the system (17) if and only if there is a C^1 clf for the unperturbed system (1).*

It is interesting to note that, as a corollary of Artstein's Theorem, for control-affine systems $\dot{x} = g_0(x) + \sum u_i g_i(x)$ we may conclude that if there is a discontinuous feedback stabilizer that is robust with respect to small noise, then there is also a regular one, and even one that is smooth on $\mathbb{R}^n \setminus \{0\}$. For non control-affine systems, however, there may exist a discontinuous feedback stabilizer that is robust with respect to small noise, yet there is no regular feedback.

Briefly, the sufficiency part of Theorem 7 proceeds by taking a pointwise minimization of the Hamiltonian, for a given C^1 clf, i. e. $k(x)$ is defined as any u with $|u| \leq \sigma(|x|)$ which minimizes $\nabla V(x) \cdot f(x, u)$. The necessity part is based on the following technical fact: if the perturbed system can be stabilized, then the differential inclusion

$$\dot{x} \in F(x) := \bigcap_{\varepsilon > 0} \overline{\mathrm{co}} f(x, k(x + \varepsilon B))$$

(where B denotes the unit ball in $\mathbb{R}^n$) is strongly asymptotically stable. One may then apply converse Lyapunov theorems for upper semicontinuous compact convex differential inclusions to deduce the existence of V.

We now summarize exactly which implications hold, writing "robust" to mean stabilization of the system subject to observation and actuator noise:

$$\begin{array}{ccccc} C^1 V & \Longleftrightarrow & \exists \text{ robust } k & & \\ \Downarrow & & \Downarrow & & \\ C^0 V & \Longleftrightarrow & \exists k & \Longleftrightarrow & \mathrm{AC}\,. \end{array}$$

Future Directions

There are several alternative approaches to feedback stabilization, notably the very appealing approach to discontinuous stabilization throgh *patchy feedbacks* [1], as well as other related "hybrid" approaches [20]. It is also extremely important to understand the effect of "large" disturbances on the behavior of feedback systems. This study leads one to the very active area of input to state stability (ISS) and related notions (output to state stability as a model of detectability, input to output stability for the study of regulation problems, and so forth), see [27].

Bibliography

Primary Literature

1. Ancona F, Bressan A (2003) Flow stability of patchy vector fields and robust feedback stabilization. SIAM J Control Optim 41:1455–1476
2. Artstein Z (1983) Stabilization with relaxed controls. Nonlinear Anal Theory Methods Appl 7:1163–1173
3. Brockett R (1983) Asymptotic stability and feedback stabilization. In: Differential geometric control theory. Birkhauser, Boston, pp 181–191
4. Clarke FH, Ledyaev Y, Sontag E, Subbotin A (1997) Asymptotic controllability implies feedback stabilization. IEEE Trans Autom Control 42(10):1394–1407
5. Coron J-M, Rosier L (1994) A relation between continuous time-varying and discontinuous feedback stabilization. J Math Syst Estim Control 4:67–84
6. Filippov A (1985) Differential equations with discontinuous right-hand side. Nauka, Moscow. English edition: Kluwer, Dordrecht, (1988)
7. Freeman R, Kokotović P (1996) Robust nonlinear control design. Birkhäuser, Boston
8. Grüne L, Sontag E, Wirth F (1999) Asymptotic stability equals exponential stability, and ISS equals finite energy gain—if you twist your eyes. Syst Control Lett 38(2):127–134
9. Hájek O (1979) Discontinuous differential equations, i & ii. J Diff Equ 32(2):149–170, 171–185
10. Hermes H (1967) Discontinuous vector fields and feedback control. In: Differential equations and dynamical systems. Proc Internat Sympos, Mayaguez 1965. Academic Press, New York, pp 155–165

11. Isidori A (1995) Nonlinear control systems: An introduction. Springer, Berlin
12. Kawski M (1989) Stabilization of nonlinear systems in the plane. Syst Control Lett 12:169–175
13. Kawski M (1995) Geometric homogeneity and applications to stabilization. In: Krener A, Mayne D (eds) Nonlinear control system design symposium (NOLCOS), Lake Tahoe, June 1995. Elsevier
14. Krasovskii N, Subbotin A (1988) Game-theoretical control problems. Springer, New York
15. Krstić M, Deng H (1998) Stabilization of uncertain nonlinear systems. Springer, London
16. Krstić M, Kanellakopoulos I, Kokotović PV (1995) Nonlinear and adaptive control design. Wiley, New York
17. Kurzweil J (1956) On the inversion of Ljapunov's second theorem on stability of motion. Am Math Soc Transl 2:19–77
18. Ledyaev Y, Sontag E (1999) A Lyapunov characterization of robust stabilization. Nonlinear Anal 37(7):813–840
19. Massera JL (1956) Contributions to stability theory. Ann Math 64:182–206
20. Prieur C (2006) Robust stabilization of nonlinear control systems by means of hybrid feedbacks. Rend Sem Mat Univ Pol Torino 64:25–38
21. Rifford L (2002) Semiconcave control-Lyapunov functions and stabilizing feedbacks. SIAM J Control Optim 41:659–681
22. Ryan E (1994) On Brockett's condition for smooth stabilizability and its necessity in a context of nonsmooth feedback. SIAM J Control Optim 32:1597–1604
23. Sontag E (1983) A Lyapunov-like characterization of asymptotic controllability. SIAM J Control Optim 21(3):462–471
24. Sontag E (1989) A "universal" construction of Artstein's theorem on nonlinear stabilization. Syst Control Lett 13(2):117–123
25. Sontag E (1998) Mathematical control theory. In: Deterministic finite-dimensional systems. Texts in Applied Mathematics, vol 6. Springer, New York
26. Sontag E (1999) Stability and stabilization: Discontinuities and the effect of disturbances. In : Nonlinear analysis, differential equations and control (Montreal, 1998). NATO Sci Ser C Math Phys Sci, vol 528. Kluwer, Dordrecht, pp 551–598
27. Sontag E (2006) Input to state stability: Basic concepts and results. In: Nistri P, Stefani G (eds) Nonlinear and optimal control theory. Springer, Berlin, pp 163–220
28. Sontag E, Sussmann H (1980) Remarks on continuous feedback. In: Proc IEEE Conf Decision and Control, Albuquerque, Dec 1980. pp 916–921

Books and Reviews

Clarke FH, Ledyaev YS, Stern RJ, Wolenski P (1998) Nonsmooth analysis and control theory. Springer, New York

Freeman R, Kokotović PV (1996) Robust nonlinear control design. Birkhäuser, Boston

Isidori A (1995) Nonlinear control systems, 3rd edn. Springer, London

Isidori A (1999) Nonlinear control systems II. Springer, London

Khalil HK (1996) Nonlinear systems, 2nd edn. Prentice-Hall, Upper Saddle River

Krstić M, Deng H (1998) Stabilization of uncertain nonlinear systems. Springer, London

Sepulchre R, Jankovic M, Kokotović PV (1997) Constructive nonlinear control. Springer, New York

Stability Theory of Ordinary Differential Equations

Carmen Chicone
Department of Mathematics, University of Missouri-Columbia, Columbia, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Mathematical Formulation of the Stability Concept and Basic Results
Stability in Conservative Systems and the KAM Theorem
Averaging and the Stability of Perturbed Periodic Orbits
Structural Stability
Attractors
Generalizations and Future Directions
Bibliography

Glossary

Ordinary differential equation An equation for an unknown vector of functions of a single variable that involves derivatives of the unknown functions. The order of a differential equation is the highest order of the derivatives that appear. The most important class of differential equations are first-order systems of ordinary differential equations that can be written in the form $\dot{u} = f(u, t)$, where f is a given smooth function $f\colon U \times J \to \mathbb{R}^n$, U is an open subset of $\mathbb{R}^n$, and J is an open subset of $\mathbb{R}$. The unknown functions are the components of u, and the vector of their first-order derivatives with respect to the independent variable t is denoted by $\dot{u}$. A solution of this differential equation is a function $u\colon K \to \mathbb{R}^n$, where K is an open subset of J such that $\frac{\mathrm{d}u}{\mathrm{d}t}(t) = f(u(t), t)$ for all $t \in K$.

Dynamical system A set and a law of evolution for its elements. The first-order differential equation $\dot{u} = f(u, t)$, where $f\colon U \times J \to \mathbb{R}^n$, is the law of evolution for the set $U \times J$; it defines a continuous (time) dynamical system: Given $(v, s) \in U \times J$, the solution $t \mapsto \psi(t, s, v)$ such that $\phi(s, s, v) = v$ determines the evolution of the state v: the state v at time s evolves to the state $\psi(t, s, v)$ at time t. Similarly, a continuous function $f\colon X \to X$ on a metric space X defines a discrete dynamical system. The state $x \in X$ evolves to $f^k(x)$ (which denotes the value of f composed with itself k times and evaluated at x) after k time-steps. The images of $t \mapsto \phi(t, s, v)$ and $k \mapsto f^k(x)$ are called the orbits of the corresponding states v and x.

Stability theory The mathematical analysis of the behavior of the distances between an orbit (or set of orbits) of a dynamical system and all other nearby orbits.

Definition of the Subject

An orbit or set of orbits of a dynamical system is stable if all solutions starting nearby remain nearby for all future times. This concept is of fundamental importance in applied mathematics: the stable solutions of mathematical models of physical processes correspond to motions that are observed in nature.

Introduction

Stability theory began with a basic question about the natural world: Is the solar system stable? Will the present configuration of the planets and the sun remain forever; or, might some planets collide, radically change their orbits, or escape from the solar system?

With the advent of Isaac Newton's second law of motion and the law of universal gravitation, the motions of the planets in the solar system were understood to correspond to the solutions of the Newtonian system of ordinary differential equations that modeled the positions and velocities of the planets and the sun according to their mutual gravitational attractions. Short-term approximate predictions (up to a few years in the future) verified this theory; but, due to the complexity of the differential equations of motion, the problem of long-term prediction was not solved; it still occupies a central place in mathematical research.

Does the Newtonian model predict the stability of the solar system?

During the 18th Century, Pierre Simon Laplace [52,53] asserted a proof of the stability of the solar system. He considered the changes in the semi-major axes and eccentricities of the elliptical motions of the planets around the sun. Using (reasonable) approximations of the Newtonian equations of motion, he showed that for his approximate model these orbital elements do not change over long periods of time due to the disturbances caused by the gravitational attractions of the other bodies in the solar system. If true for the full Newtonian equations of motion, these assertions would imply the stability of the Newtonian solar system. In fact, no proof of the stability of the solar system is known (see [69]). Laplace's results merely provide evidence in favor of stability of the solar system; on the other hand, this work was a primary stimulus for the later development of a general theory of stability.

One of the first rigorous results in stability theory was stated by Joseph-Louis Lagrange [49] and proved by Lejeune Dirichlet [23,24]; it states that an isolated minimum of the potential energy of a conservative mechanical system is the position of a stable equilibrium point.

Joseph Liouville [56,57,58,59] discussed the problem of the stability of rotating fluid bodies. The further development of this theory was suggested to Aleksandr Mikhailovich Lyapunov as a thesis topic by his advisor Pafnuty Chebyshev. This led to the fundamental and foundational work of Lyapunov on stability theory [63].

Henri Poincaré's introduction of the qualitative theory of differential equations [71] influenced Lyapunov's treatment of stability theory and laid much of the foundation for the modern theory of nonlinear dynamical systems. In addition, Poincaré's work on celestial mechanics [72,73,74,75,76] discusses stability theory.

Mathematical Formulation of the Stability Concept and Basic Results

Consider the first-order ordinary differential equation

$$\dot{u} = f(u, t)\,, \tag{1}$$

where the dependent variable u is an n-dimensional real vector, $\dot{u}$ denotes the derivative of u with respect to the independent variable t, and f (a mapping, from the cross product of an open subset U of n-dimensional space $\mathbb{R}^n$ and an open subset J of the real line $\mathbb{R}$, into $\mathbb{R}^n$) is at least class C^1. The existence and uniqueness theorem for differential equations (see, for example, [19]) states that if $(v, s) \in U \times J$, then there is a unique solution $t \mapsto \psi(t, s, v)$ of the differential equation such that $\psi(s, s, v) = v$. Moreover, the function ψ is as smooth as the function f.

Definition 1 For $v \in \mathbb{R}^n$, the notation $|v|$ denotes the length of v.

A solution $t \mapsto \psi(t, s, v)$ of the differential Eq. (1) is called stable if it is defined for all time $t \geq s$ and for every positive number ϵ there is a positive number δ such that $|\phi(t, s, w) - \phi(t, s, v)| < \epsilon$ whenever $t \geq s$ and $|w - v| < \delta$ (see Fig. 1).

A stable solution $t \mapsto \psi(t, s, v)$ is called orbitally asymptotically stable if there is a choice of δ such that the distance between the point $\psi(t, s, w)$ and the set $\{\psi(t, s, v)\colon t \geq s\}$—called the forward orbit of the solution—converges to zero as $t \to \infty$.

A stable solution $t \mapsto \psi(t, s, v)$ is called asymptotically stable if there is a choice of δ such that the distance $|\phi(t, s, w) - \phi(t, s, v)|$ converges to zero as $t \to \infty$ whenever $|w - v| < \delta$ (see Fig. 2).

A solution is called unstable if it is not stable.

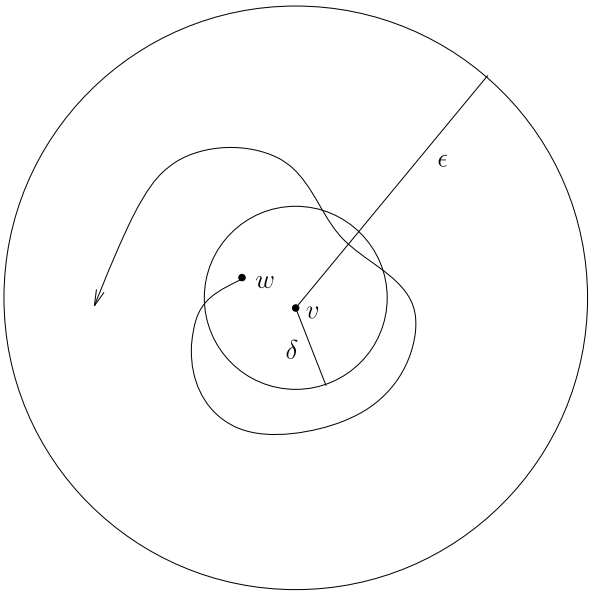

Stability Theory of Ordinary Differential Equations, Figure 1
The figure depicts a stable rest point *v*. The orbit starting at *w*, an arbitrary point whose distance from *v* is less than δ, remains within distance ϵ for all positive time

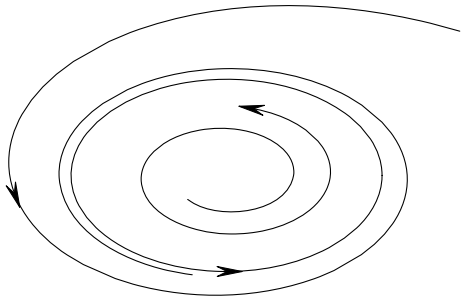

Stability Theory of Ordinary Differential Equations, Figure 2
The figure depicts an orbitally asymptotically stable elliptical periodic orbit together with two additional orbits that approach the periodic orbit, one from the outside and one from the inside of the region bounded by the periodic orbit

While the definitions just given are widely accepted, the words "asymptotic stability" are often used to mean orbital asymptotic stability. For most situations orbital asymptotic stability is the desired concept. It is clear from the definitions that the concept of asymptotic stability is much stronger than the concept of orbital asymptotic stability for stable solutions whose orbits consist of more than one point. The position of the evolving state along an asymptotically stable orbit is approached by an open set of evolving states. On the other hand, for a solution to be orbitally asymptotically stable only the distances between the elements of this open set of evolving states and the point set consisting of the forward orbit of the stable solution is required to approach zero. There is no requirement that the open set of evolving states all converge to the evolving state on the stable orbit in unison as time increases without bound. An important concept introduced by Poincaré, the return map, is useful in the study of orbital asymptotic stability; it will be discussed below.

To illustrate the concept of stability, let us consider the (dimensionless) harmonic oscillator

$$\ddot{x} + x = 0\,. \tag{2}$$

This second-order differential equation is recast as the first-order system

$$\dot{x} = y\,, \quad \dot{y} = -x \tag{3}$$

by introducing the velocity $y := \dot{x}$ as a new variable. It is a classical mechanical system with kinetic energy $\frac{1}{2}\dot{x}^2$ and potential energy $\frac{1}{2}x^2$. According to the principle of Lagrange, the equilibrium solution $(x, y) = (0, 0)$ (which corresponds to an isolated minimum of the potential energy) is stable. This result is easily verified by inspection of the general solution of system (3) that is given by

$$\phi(t, \xi, \eta) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix}\,, \tag{4}$$

where (ξ, η) is an arbitrary point in $\mathbb{R}^2$. All non-equilibrium solutions are periodic. In fact, the orbit of the solution starting at (ξ, η) is a circle with radius $\sqrt{\xi^2 + \eta^2}$. Given an open subset containing the origin, the disk bounded by one of these circles defines an open subset. Every orbit with initial value inside this disk remains inside the disk (and hence the given open set) for all $t \geq 0$. On the other hand, the origin is not orbitally asymptotically stable; indeed, the periodic solutions do not approach the origin in positive time.

The origin is also an equilibrium of the damped harmonic oscillator

$$\ddot{x} + \epsilon\dot{x} + x = 0\,, \tag{5}$$

where $\epsilon > 0$. Again, inspection of the general solution shows that the origin is asymptotically stable and therefore orbitally asymptotically stable. In effect, all terms in the formula for the general solution contain the factor $e^{-\epsilon t/2}$, which converges to zero as $t \to \infty$.

For a nonautonomous example (that is, a differential Eq. (1) such that the partial derivative of f with respect to t is not zero), consider the first-order scalar differential equation

$$\dot{x} = -x + \sin t \tag{6}$$

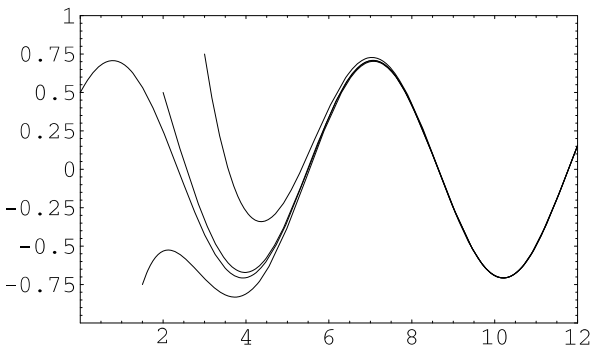

Stability Theory of Ordinary Differential Equations, Figure 3
A graph of the asymptotically stable solution (7) of the scalar first-order differential Eq. (6) is depicted together with the graphs of three additional solutions given by $t \mapsto \phi(t, s, v)$ for (s, v) equal to (1.5, −0.75), (3.0, 0.75), and (2.0, 0.5)

whose solutions are given by

$$\phi(t, s, v) = \left(v - \frac{1}{2}(\sin s + \cos s)\right) e^{s-t} + \frac{1}{2}(\sin t + \cos t) .$$

The solution starting at the state 1/2 at time zero,

$$\phi\left(t, 0, \frac{1}{2}\right) = \frac{1}{2}(\sin t + \cos t) , \tag{7}$$

is asymptotically stable. This result is true due to the presence of the exponential factor in the general solution; it decreases to zero for each fixed s as $t \to \infty$ (cf. Fig. 3).

The origin is an unstable equilibrium for the scalar first-order differential equation $\dot{x} = x^2$, whose solutions are given by

$$\phi(t, s, v) = \frac{v}{1 - (t - s)v} .$$

Indeed, for every $\delta > 0$ (no matter how small) there is a some $v \in \mathbb{R}$ such that $|v| < \delta$ and $v > 0$. The corresponding solution converges to infinity as $t \to (1 + sv)/v$.

Because explicit solutions of differential equations are rare, the main subject of stability theory is the determination of criteria for stability that do not require knowledge of the general solution of the differential equation.

While the theory of stability is mature with many branches of development, the main results (originally obtained by Poincaré and Lyapunov) follow from two fundamental ideas: linearization and Lyapunov functions.

The Principle of Linearized Stability

A linear first-order differential equation is a differential equation of the form

$$\dot{u} = A(t)u , \tag{8}$$

where $A(t)$ denotes an $n \times n$ matrix-valued function of the independent variable t.

Most of the analysis of autonomous linear differential equations can be reduced to linear algebra. Indeed, a function of the form $e^{\lambda t}v$, where λ is a complex number and $v \neq 0$ is a complex vector with n components, is a complex solution of the differential equation

$$\dot{u} = Au , \tag{9}$$

if and only if $Av = \lambda v$; that is, λ is an eigenvalue of the matrix A and v is a corresponding eigenvector. The real and imaginary parts of a complex solution are automatically real solutions of the real differential Eq. (9). Also, linear combinations of solutions of linear equations are again solutions. Thus, in case there is a basis of eigenvectors, all solutions of the linear system can be expressed as linear combinations (superpositions) of the special exponential solutions. The general solution, for an arbitrary system matrix A, can always be obtained from linear combinations of solutions of the form $p(t)e^{\lambda t}v$, where $p(t)$ is a polynomial of degree at most $n - 1$ and $Av = \lambda v$. In fact, there is always a linear change of variables $w = Bu$, given by some invertible matrix B, such that the new system

$$\dot{w} = Jw ,$$

where the transformed system matrix $J := BAB^{-1}$ is in Jordan canonical form (see, for example, [78]). This block diagonal system can be solved by linear combinations of solutions of the form $p(t)e^{\lambda t}v$, and the solution of the original system (9) is obtained by the inverse change of variables. Alternatively, it is not difficult to prove that the general solution of the system (9) is given by

$$\phi(t, s, v) = e^{(t-s)A}v , \tag{10}$$

where

$$e^{tA} := I + \sum_{k=1}^{\infty} (tA)^k \tag{11}$$

and v is an arbitrary point in $\mathbb{R}^n$.

Solutions of the form $t \mapsto p(t)e^{\lambda t}v$ converge to zero as $t \to \infty$ whenever the real part of the eigenvalue λ is less than zero. This observation leads to a basic result:

Theorem 2 *If all eigenvalues of the matrix A have negative real parts, then the zero solution of the autonomous system (9) is asymptotically stable.*

To examine the stability of a solution ψ of the (perhaps nonlinear) differential Eq. (1), consider a second solution γ, define the deviation $\delta = \gamma - \psi$, and note that

$$\dot{\delta} = f(\gamma, t) - f(\psi, t) = f(\delta + \psi, t) - f(\psi, t)$$

Definition 3 The linearized equation along the solution ψ of the differential Eq. (1) is

$$\dot{w} = f_u(\psi(t), t)w \,, \tag{12}$$

where the subscript denotes the partial derivative with respect to u.

The linearized equation may be viewed as an approximation of the differential equation for the deviation δ because the linearized equation is also obtained by Taylor expanding the function

$$\delta \mapsto f(\delta + \psi, t) - f(\psi, t)$$

to first-order at $\delta = 0$ and ignoring the remainder term. The principle of linearized stability states that the original solution ψ is stable whenever the zero solution of the linearized equation is stable.

The principle of linearized stability is not a theorem; rather, it serves as the underlying idea for the theory of stability by linearization. A basic result of this theory is the following theorem (see [63]).

Theorem 4 *Let*

$$\dot{u} = Au + B(t)u + g(u, t) \,, \quad u(t_0) = u_0 \,, \quad u \in \mathbb{R}^n$$

be a smooth initial value problem with solution $t \mapsto \psi(t)$. If

(1) A is a constant matrix and all its eigenvalues have negative real parts,
(2) B is an $n \times n$ matrix valued function, continuously dependent on t, such that the matrix norm $\|B(t)\|$ converges to zero as $t \to \infty$,
(3) g is smooth and there are constants $a > 0$ and $k > 0$ such that $|g(v, t)| \le k|v|^2$ for all $t \ge 0$ and $|v| > a$,

then there are constants $C > 1$, $\beta > 0$, and $\lambda > 0$ such that

$$|\psi(t)| \le C|u_0|e^{-\lambda(t-t_0)} \,, \quad t \ge t_0$$

whenever $|u_0| \le \beta/C$. In particular, the zero solution is asymptotically stable.

Theorem 4 is used for the stability analysis of rest points of autonomous first-order differential equations. If $f\colon \mathbb{R}^n \to \mathbb{R}^n$ and $f(\xi) = 0$, for some $\xi \in \mathbb{R}^n$, then the constant function $\phi(t, \xi) = \xi$ is a solution of the autonomous differential equation

$$\dot{u} = f(u) \,. \tag{13}$$

In this case, ξ is called a rest point for the differential equation. By Taylor expansion at the rest point—this time keeping the remainder term, the differential equation is recast in the form

$$\dot{\gamma} = Df(\xi)\gamma + R(\xi, \gamma) \,,$$

where Df denotes the derivative of f and $\gamma := \phi - \xi$ for an arbitrary solution ϕ. If f is class C^2, we have that

$$|R(\xi, v)| \le k|v|^2 \,.$$

Thus, under the hypothesis that all eigenvalues of the constant system matrix $Df(\xi)$ have negative real parts, Theorem 4 implies that the rest point ξ is asymptotically stable. The hypothesis that f is class C^2 can be relaxed (see, for example, [19]):

Theorem 5 *If $f\colon \mathbb{R}^n \to \mathbb{R}^n$ is class C^1, $\xi \in \mathbb{R}^n$, $f(\xi) = 0$, and all eigenvalues of the matrix $Df(\xi)$ have negative real parts, then the rest point ξ is asymptotically stable. If the matrix $Df(\xi)$ has an eigenvalue with positive real part, then the rest point ξ is unstable.*

The damped harmonic oscillator (5) is equivalent to the first-order system of (linear) differential equations

$$\dot{u} = Au \,, \tag{14}$$

where the system matrix is

$$A := \begin{pmatrix} 0 & 1 \\ -1 & -\epsilon \end{pmatrix} . \tag{15}$$

If $\epsilon > 0$, the system matrix has eigenvalues $\frac{1}{2}(-\epsilon \pm \sqrt{\epsilon^2 - 4})$ whose real parts are always negative. Thus, the rest point at the origin is asymptotically stable in agreement with physical intuition.

The power of the linearization method is demonstrated by its applications to systems whose general solutions are not known explicitly; for example, the nonlinear oscillator

$$\ddot{x} + \epsilon x + x + g(x) = 0 \,, \tag{16}$$

where g is an arbitrary (smooth) function such that $g(0) = Dg(0) = 0$. In this case, the origin is a rest point of the equivalent first-order system and the system matrix of the linearized system at the origin is again the matrix (15); therefore, by Theorem 5, the origin is asymptotically stable. This result applies to the rest point $(\theta, \dot{\theta}) = (0, 0)$ of the damped pendulum

$$\ddot{\theta} + \epsilon\dot{\theta} + \sin\theta = 0 \,. \tag{17}$$

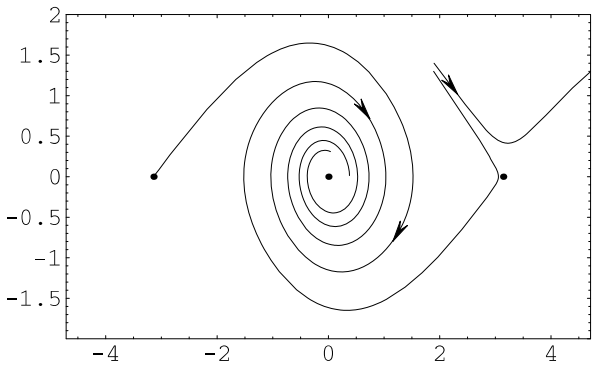

Stability Theory of Ordinary Differential Equations, Figure 4
The figure depicts portions of six orbits in the phase portrait of the damped pendulum (17) with $\epsilon = 0.2$: the rest points at $(\theta, \dot{\theta})$ equal to $(-\pi, 0)$, $(0, 0)$, and $(\pi, 0)$; and, the orbits starting at $(-\pi, 0.001)$, $(\pi - 1.25, 1.3)$, and $(\pi - 1.243, 1.4)$. The last two orbits pass close to the (saddle type) unstable rest point at $(\pi, 0)$

The differential Eq. (17) also has a rest point at $(\theta, \dot{\theta}) = (\pi, 0)$ where the system matrix of the linearized differential equation is

$$\begin{pmatrix} 0 & 1 \\ 1 & -\epsilon \end{pmatrix} . \tag{18}$$

If $\epsilon > 0$, this matrix has an eigenvalue with positive real part; therefore, this rest point—which corresponds to the upward vertical equilibrium of the physical pendulum—is unstable (cf. Fig. 4).

Definition 6 An $n \times n$ matrix is called infinitesimally hyperbolic if all of its eigenvalues have non-zero real parts. A rest point of an autonomous system is called hyperbolic if the system matrix of the linearized equation at this rest point is infinitesimally hyperbolic. The rest point is called nondegenerate if the system matrix is invertible.

An important result on the principle of linearized stability for rest points is the Grobman–Hartman theorem (see [30,31,34,35] and [19,22]):

Theorem 7 *If v is a hyperbolic rest point for the autonomous differential equation $\dot{u} = f(u)$, then there is an open set V containing v and a homeomorphism H with domain V such that the orbits of this differential equation in V are mapped by H to orbits of the linearized system $\dot{w} = Df(v)w$.*

In other words, the qualitative behavior of an autonomous differential equation in a sufficiently small neighborhood of a hyperbolic rest point is the same as the behavior of its linearization at this rest point.

The related problem of the existence of a change of variables (a diffeomorphism) taking a given system to its linearization in a neighborhood of a rest point is more difficult to resolve. Fundamental work in this area is due to Poincaré [74], Carl Siegel [88], Henri Dulac [25], Shlomo Sternberg [90,91] and A.D. Brjuno [16] (see also [9,89]). The non-existence of certain relationships (resonances) among the eigenvalues of the system matrix of the linearization plays an important role in the theory.

Definition 8 The vector of eigenvalues $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ of the $n \times n$-matrix A is called resonant if there is an n-vector of nonnegative integers $M = (m_1, m_2, \ldots, m_n)$ such that $\sum_{k=1}^{n} |m_k| \geq 2$ and at least one of the eigenvalues λ_k is given by

$$\lambda_k = \langle M, \lambda \rangle ,$$

where the angle brackets denote the usual inner product. The eigenvalues are nonresonant if no such relationship occurs.

The first theorem of the subject was proved by Poincaré:

Theorem 9 *If the eigenvalues of the linearization of a vector field given by a formal power series are nonresonant, then there is a change of variables in the form of a formal power series that transforms the vector field to its linearization.*

In case the linearization is resonant, there is a formal change of variables that removes all of the terms in the formal series that defines the vector field except those that are resonant (that is, the corresponding monomials are given by products of variables whose powers form vectors with the properties of M in the definition of resonance).

Definition 10 The vector of eigenvalues $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ of the $n \times n$-matrix A belongs to the Poincaré domain if the convex hull of its components in the complex plane does not contain the origin; otherwise, the vector belongs to the Siegel domain. The vector λ is called type (C, ν) for $C > 0$ and $\nu > 0$, if the inequality

$$|\lambda_k - \langle M, \lambda \rangle| \geq \frac{C}{(\sum_k |m_k|)^\nu}$$

is satisfied for each component λ_k of λ and all vectors M as in Definition 8.

Theorem 11 *If the eigenvalues of the linearization of an analytic vector field are nonresonant and in the Poincaré domain, then there is an analytic change of coordinates that transforms the vector field to its linearization. If the eigenvalues of the linearization of an analytic vector field are of*

type (C, ν), then there is an analytic change of coordinates that transforms the vector field to its linearization.

For modern proofs of Poincaré's and Siegel's theorems see [71] or [87]. Sternberg proved Siegel's result for vector fields in the Siegel domain for sufficiently smooth vector fields (see [90,91]).

To successfully apply linearized stability for rest points of autonomous systems, the fundamental problem is to determine the real parts of the eigenvalues of a square matrix; especially, it is important to determine conditions that ensure all eigenvalues have negative real parts. The most important result in this direction is the Routh-Hurwitz criterion (see [40,84]):

Theorem 12 *Suppose that the characteristic polynomial of the real matrix A is written in the form*

$$\lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n \,,$$

let $a_m = 0$ for $m > n$, and define the determinants Δ_k for $k = 1, 2, \ldots, n$ by

$$\Delta_k := \det \begin{pmatrix} a_1 & 1 & 0 & 0 & 0 & 0 & \cdots 0 \\ a_3 & a_2 & a_1 & 1 & 0 & 0 & \cdots 0 \\ a_5 & a_4 & a_3 & a_2 & a_1 & 1 & \cdots 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \; \vdots \\ a_{2k-1} & a_{2k-2} & a_{2k-3} & a_{2k-4} & a_{2k-5} & a_{2k-6} & \cdots a_k \end{pmatrix}.$$

If $\Delta_k > 0$ for $k = 1, 2, \ldots n$, then all roots of the characteristic polynomial have negative real parts.

The principle of linearized stability is not valid in general. In particular, the stability of the linearized equation at a rest point of an autonomous differential equation does not always determine the stability of the rest point. For example, consider the scalar differential equation $\dot{u} = -u^m$, where $m > 1$ is a positive integer. The origin $u = 0$ is a rest point for this differential equation and the corresponding linearized system is $\dot{w} = 0$. The origin (and every other point on the real line) is a stable rest point of this linear differential equation. On the other hand, by solving the original equation, it is easy to see that the origin is asymptotically stable in case m is odd and unstable in case m is even. Thus, the stability of the linearization does not determine the stability of the rest point.

There is no theorem known that can be used to determine the stability of a general nonautonomous first-order linear differential Eq. (8). In particular, there is no obvious relation between the (time-dependent) eigenvalues of a time-dependent matrix and the stability of the zero solution of the corresponding first-order linear differential equation. Consider, as an example, the π-periodic system $\dot{u} = A(t)u$, where

$$A(t) = \begin{pmatrix} -1 + \frac{3}{2}\cos^2 t & 1 - \frac{3}{2}\sin t \cos t \\ -1 - \frac{3}{2}\sin t \cos t & -1 + \frac{3}{2}\sin^2 t \end{pmatrix}, \quad (19)$$

which was constructed by Lawrence Marcus and Hidehiko Yamabe [66]. The real parts of the eigenvalues of $A(t)$ (given by $\frac{1}{4}(-1 \pm \sqrt{7}\, i)$) are negative; but,

$$u(t) = \mathrm{e}^{t/2} \begin{pmatrix} -\cos t \\ \sin t \end{pmatrix}$$

is a solution that grows without bound.

While a general theory of linearized stability for non-constant solutions of differential equations remains an area of research, much is known in case the non-constant solution is periodic.

Suppose that $t \mapsto \psi(t)$ is a periodic solution with period T of the differential Eq. (1) defined on $\mathbb{R}^n$. Linearization leads to the differential equation

$$\dot{\gamma} = f_u(\psi(t), t)\gamma \,, \quad (20)$$

where the matrix (valued function) $A(t) := f_u(\psi(t), t)$ is periodic with period T. Thus, the theory of linearized stability for periodic solutions starts with the stability theory for the periodic linear differential Eq. (8). The most basic results in this setting are due to Gaston Floquet [27]. To state them, let us first recall that a linear differential equation admits matrix solutions; that is, matrices whose columns are vector solutions. A fundamental matrix solution is an $n \times n$-matrix solution with linearly independent columns; or, equivalently, a matrix solution whose values are all invertible as $n \times n$-matrices.

Floquet's main results are stated in the following theorem (see [19] and recall the definition of the matrix exponential in display (11)).

Theorem 13 *If $\Phi(t)$ is a fundamental matrix solution of the T-periodic system (8), then*

$$\Phi(t + T) = \Phi(t)\Phi^{-1}(0)\Phi(T)$$

for all $t \in \mathbb{R}$. In addition, there is a matrix B (which may be complex) such that

$$e^{TB} = \Phi^{-1}(0)\Phi(T)$$

and a T-periodic matrix function $t \mapsto P(t)$ (which may be complex valued) such that $\Phi(t) = P(t)e^{tB}$ for all $t \in \mathbb{R}$. Also, there is a real matrix R and a real $2T$-periodic matrix function $t \to Q(t)$ such that $\Phi(t) = Q(t)e^{tR}$ for all $t \in \mathbb{R}$.

The stability of the zero solution of the periodic linear differential equation is intimately connected with the eigenvalues of the matrix R in Floquet's theorem. Indeed, the time-dependent (real) change of variables $u = Q(t)v$ transforms the T-periodic differential Eq. (8) to the autonomous linear system

$$\dot{v} = Rv\,.$$

Definition 14 The representation $\Phi(t) = P(t)e^{tB}$ of the fundamental matrix solution Φ in Floquet's theorem is called a Floquet normal form. The eigenvalues of the matrix e^{tB} are called the characteristic (or Floquet) multipliers of the corresponding linear system. A complex number μ is called a characteristic (or Floquet) exponent if there is a characteristic multiplier ρ such that $e^{\mu T} = \rho$.

Theorem 15

(1) The characteristic multipliers and the characteristic exponents do not depend on the choice of the fundamental matrix solution of the T-periodic differential Eq. (8).
(2) If the characteristic multipliers of the periodic system (8) all have modulus less than one; equivalently, if all characteristic exponents have negative real parts, then the zero solution is asymptotically stable.
(3) If the characteristic multipliers of the periodic system (8) all have modulus less than or equal to one; equivalently, if all characteristic exponents have nonpositive real parts, and if the algebraic multiplicity equals the geometric multiplicity of each characteristic multiplier with modulus one; equivalently, if the algebraic multiplicity equals the geometric multiplicity of each characteristic exponent with real part zero, then the zero solution is stable.
(4) If at least one characteristic multiplier of the periodic system (8) has modulus greater than one; equivalently, if a characteristic exponent has positive real part, then the zero solution is unstable.

While Theorem 15 gives a complete description of the stability of the zero solution, no general method is known to determine the characteristic multipliers (cf. [98]). The power of the theorem is an immediate corollary: A finite set of numbers (namely, the characteristic multipliers) determine the stability of the zero solution, at least in the case where none of them have modulus one. The characteristic multipliers can be approximated by numerical integration. And, in special cases, their moduli can be determined by mathematical analysis.

One of the most important applications for Floquet theory and the method of linearization, is the analysis of Hill's equation

$$\ddot{x} + a(t)x = 0\,, \tag{21}$$

or, equivalently, the first-order linear system

$$\dot{x} = y\,, \quad \dot{y} = -a(t)x\,, \tag{22}$$

where x is a scalar and a is a T-periodic function (see [19, 64]). This equation arose from George Hill's study of lunar motion. It is ubiquitous in stability theory.

The stability of the zero solution of Hill's system (22) can be reduced to the identification of a single number; namely, the magnitude of the trace of the principal fundamental matrix solution $\Phi(t)$ at $t = T$ (that is, the fundamental matrix solution $\Phi(t)$ such that $\Phi(0) = I$, where I is the $n \times n$ identity matrix).

Theorem 16 *Suppose that $\Phi(t)$ is the principal fundamental matrix solution of system (22). If $|\mathrm{trace}\,\Phi(T)| < 2$, then the zero solution is stable. If $|\mathrm{trace}\,\Phi(T)| > 2$, then the zero solution is unstable.*

This result, of course, requires knowledge of the solutions of the system. On the other hand, a beautiful theorem of Lyapunov [55] gives a stability criterion using only properties of the function a:

Theorem 17 *If $a\colon \mathbb{R} \to \mathbb{R}$ is a positive T-periodic function such that*

$$T\int_0^T a(t)\,dt \le 4\,,$$

then all solutions of the Hill's equation $\ddot{x} + a(t)x = 0$ are bounded. In particular, the trivial solution is stable.

The principle of linearized stability for periodic solutions of nonlinear systems motivates a theory similar to the stability theory for rest points. There is, however, one essential difference that will be explained for a T-periodic solution γ of the autonomous first-order differential equation

$$\dot{u} = f(u). \tag{23}$$

The linearized equation at γ is $\dot{w} = Df(\gamma(t))w$. Let $\phi(t) = f(\gamma(t))$ and note that

$$\dot{\phi}(t) = Df(\gamma(t))f(\gamma(t)) = Df(\gamma(t))\phi(t)\,;$$

that is, $t \mapsto f(\gamma(t))$ is a solution of the linearized equation. This solution is given by

$$f(\gamma(t)) = \Phi(t)f(\gamma(0))\,,$$

where $\Phi(t)$ is the principal fundamental matrix solution of the linearized equation. If follows that $\Phi(T)f(\gamma(0)) =$

$f(\gamma(0))$; therefore, the number 1 is a characteristic multiplier. It corresponds to motion in the direction of the periodic solution and does not contribute to the stability or instability of the periodic solution. This characteristic multiplier must be treated separately in the stability theory.

Theorem 18 *If the number 1 occurs with algebraic multiplicity one in the set of characteristic multipliers of the linearized differential equation at a periodic solution and all other characteristic multipliers have modulus less than one, then the periodic solution is orbitally asymptotically stable. If at least one characteristic multiplier has modulus larger than one, then the periodic solution is unstable.*

Poincaré introduced a geometric approach to stability for periodic solutions. His idea is simple and useful: Consider a periodic solution γ and some point, say $\gamma(0)$, on the orbit of this solution in $\mathbb{R}^n$. Construct a hypersurface S in $\mathbb{R}^n$ that contains $\gamma(0)$ and is transverse to the orbit of γ at this point. By the implicit function theorem, there is an open neighborhood Σ of $\gamma(0)$ in S such that the solution of the differential equation starting at every point in Σ returns to S. This defines a local diffeomorphism $\mathcal{P}\colon \Sigma \to S$, called the Poincaré (or return) map, which assigns p to its point of first return on S (cf. Fig. 5).

The eigenvalues of the derivative of $\mathcal{P}$ at $\gamma(0)$ are exactly the characteristic multipliers associated with γ. Thus, if all eigenvalues of this derivative have modulus less than one, then γ is orbitally asymptotically stable. The derivative of the Poincaré map can be computed using the linearized equations of the corresponding differential equation along its periodic orbit. While explicit formulas for the desired derivative are only available for special cases, the Poincaré map provides an indispensable tool for organizing the study of stability for periodic orbits in general.

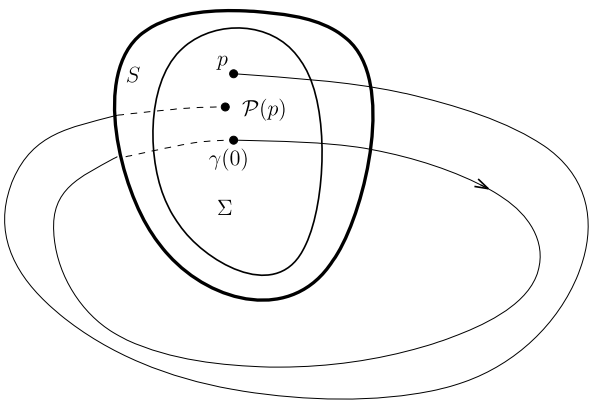

Stability Theory of Ordinary Differential Equations, Figure 5
A schematic diagram of a Poincaré map $\mathcal{P}$ defined on a Poincaré section Σ near a periodic orbit containing the point $\gamma(0)$

It should be clear that the discrete dynamical system defined by $\mathcal{P}$ codes all the information about the qualitative behavior of the solutions of the differential equation that start near the orbit of γ. The existence of the Poincaré map is thus the main connection between discrete and continuous dynamical systems.

Often, but not always, the dynamical properties of diffeomorphisms (or maps) are simpler to analyze than the corresponding properties of differential equations. Thus, theorems are often proved first for maps and later for differential equations. For instance, the Grobman–Hartman theorem was first proved for diffeomorphisms. This result can be applied to the Poincaré map of a periodic orbit to prove the existence of a homeomorphism that maps nearby orbits to the orbits of a corresponding linearization.

For the special case of periodic solutions of autonomous first-order differential equations on the plane, there is exactly one important Floquet multiplier; it is identified in the following fundamental result.

Theorem 19 *If γ is a periodic solution with period T of the differential Eq. (23) defined on $\mathbb{R}^2$, then the number*

$$\mu = \int_0^T \operatorname{div} f(\gamma(t))\, dt$$

(where div *denotes the divergence) is (up to a positive scalar multiple) a characteristic exponent of γ. If μ is negative, then γ is orbitally asymptotically stable. If μ is positive, then γ is unstable.*

The system of differential equations

$$\dot{x} = x - y - (x^2 + y^2)x\,, \quad \dot{y} = x + y - (x^2 + y^2)y$$

has the periodic solution $t \to (\cos t, \sin t)$ (depicted with a non-unit aspect ratio in Fig. 2). Also, the divergence of the vector field is $2 - 4(x^2 + y^2)$; it has value -2 along the corresponding periodic orbit $\Gamma = \{(x, y)\colon x^2 + y^2 = 1\}$. By Theorem 19, Γ is orbitally asymptotically stable. This is an example of a limit cycle; that is, an isolated periodic orbit. By changing to polar coordinates (r, θ), the transformed system

$$\dot{r} = r(1 - r^2)\,, \quad \dot{\theta} = 1$$

decouples, and its flow is given by

$$\phi_t(r, \theta) = \left(\left(\frac{r^2 e^{2t}}{1 - r^2 + r^2 e^{2t}} \right)^{\frac{1}{2}}, \theta + t \right). \tag{24}$$

The positive x-axis is a Poincaré section with corresponding Poincaré map P given by

$$P(x) = \left(\frac{x^2 e^{4\pi}}{1 - x^2 + x^2 e^{4\pi}} \right)^{\frac{1}{2}} .$$

The periodic orbit Γ intersects the Poincaré section at $x = 1$ where we have $P(1) = 1$ and $P'(1) = e^{-4\pi} < 1$.

A periodic solution of an autonomous system cannot be asymptotically stable. This fact depends on a special property of the solutions of autonomous systems: if the differential Eq. (1) is autonomous, then the function ψ giving the solutions, $t \mapsto \psi(t, s, v)$, is a function of $t - s$. Hence, by replacing $t - s$ by t, each solution can be expressed in the form $t \mapsto \phi(t, v)$ where $\phi(0, v) = v$ and $\phi(t, \phi(s, v)) = \phi(t + s, v)$ whenever both sides are defined. The family of functions $v \mapsto \phi(t, v)$, parametrized by t is called the flow of the differential Eq. (1).

Suppose that $t \mapsto \phi(t, v)$ is an asymptotically stable periodic solution with period T. Let $\delta > 0$ be as in Definition 1. If $|w - v| < \delta$, then $\lim_{t\to\infty} |\phi(t, w) - \phi(t, v)| = 0$. For all sufficiently small s, we have that $|\phi(s, w) - v| < \delta$. Hence, $\lim_{t\to\infty} |\phi(t, \phi(s, w)) - \phi(t, v)| = 0$. By the triangle law,

$$\begin{aligned} &|\phi(t + s, v) - \phi(t, v)| \\ &\le |\phi(t + s, v) - \phi(t + s, w)| + |\phi(t + s, w) - \phi(t, v)| \\ &\le |\phi(s, \phi(t, v)) - \phi(s, \phi(t, w)| + |\phi(t + s, w) - \phi(t, v)|. \end{aligned}$$

Since $\xi \mapsto \phi(s, \xi)$ is smooth and a periodic orbit is compact, the law of the mean implies that there is a constant $C > 0$ such that

$$\begin{aligned} &|\phi(t + s, v) - \phi(t, v)| \\ &\quad \le C|\phi(t, v) - \phi(t, w)| + |\phi(t + s, w) - \phi(t, v)| . \end{aligned}$$

Both summands on the right-hand side of the last inequality converge to zero as $t \to \infty$. Hence, we have that

$$\lim_{k\to\infty} |\phi(kT + s, v) - \phi(kT, v)| = 0 ,$$

where k denotes an integer variable. It follows that $\phi(s, v) = v$ for all s in some open set. The point v must be a rest point on a periodic orbit, in contradiction to the uniqueness of solutions.

While a periodic solution of an autonomous system cannot be asymptotically stable, we can ask that an orbitally asymptotically stable periodic orbit satisfy a stronger type of stability.

Definition 20 Let ϕ denote the flow of the autonomous differential Eq. (23) and suppose that Γ is a periodic orbit corresponding to the solution $t \mapsto \phi(t, p)$. We say that $q \in \mathbb{R}^n$ has asymptotic phase p with respect to Γ if the solution $t \mapsto \phi(t, q)$ is such that

$$\lim_{t\to\infty} |\phi(t, q) - \phi(t, p)| = 0 .$$

We say that Γ is isochronous if there exists an open neighborhood of its orbit such that every point in this neighborhood has asymptotic phase with respect to the period solution.

Theorem 21 *If all eigenvalues of the derivative of a Poincaré map at the periodic solution $t \mapsto \phi(t, p)$ of (23) lie strictly inside the unit circle in the complex plane (equivalently, the corresponding periodic orbit is hyperbolic), then every point q in some open neighborhood of this periodic orbit has asymptotic phase.*

A theory of asymptotic phase that includes orbitally stable periodic orbits such that some eigenvalues of the derivative of an associated Poincaré map have modulus one was recently completed (see [21,26]).

Lyapunov Functions

In case the linearized differential equation along a solution is stable but not orbitally asymptotically stable, the principle of linearized stability usually cannot be justified. One of the great contributions of Lyapunov is the introduction of a method that can be used to determine the stability of such solutions.

The fundamental ideas of Lyapunov are most clear for the stability analysis of rest points of autonomous systems (see [63]); but, the theory goes far beyond this case (see [51]).

Definition 22 Let u_0 be a rest point of the autonomous differential Eq. (23). A continuous function $L\colon U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^n$ is an open set with $u_0 \in U$, is called a Lyapunov function at u_0 if

(1) $L(u_0) = 0$,
(2) $L(x) > 0$ for $x \in U \setminus \{u_0\}$,
(3) the function L is continuously differentiable on the set $U \setminus \{u_0\}$, and, on this set, $\dot{L}(x) := \operatorname{grad} L(x) \cdot f(x) \le 0$.

The function L is called a strict Lyapunov function if, in addition,

(4) $\dot{L}(x) < 0$ for $x \in U \setminus \{u_0\}$.

Property (3) states that L does not increase along solutions.

Theorem 23 *If there is a Lyapunov function defined on an open neighborhood of a rest point of an autonomous first-order differential equation, then the rest point is stable. If, in*

addition, the Lyapunov function is a strict Lyapunov function, then the rest point is asymptotically stable.

Lyapunov's theorem was motivated by Lagrange's principle: a rest point of a mechanical system corresponding to an isolated minimum of the potential energy is stable. To obtain this result, recall that the equation of motion of a (conservative) mechanical system for the motion of a particle is

$$m\ddot{x} = -\operatorname{grad} G(x)$$

where m is the mass of the particle and $x \in \mathbb{R}^n$ is its position. The kinetic energy of the particle is defined to be $K = \frac{m}{2}\langle \dot{x}, \dot{x}\rangle$ (where the angle brackets denote the usual inner product) and the potential energy is G, a quantity defined up to an additive constant.

Consider the corresponding equivalent first-order system

$$\dot{x} = y\,, \quad \dot{y} = -\frac{1}{m}\operatorname{grad} G(x)\,,$$

and suppose that x_0 be an isolated minimum of G. The state $(x, y) = (x_0, 0)$ is a rest point of the mechanical system. By inspection, the total energy

$$L(x, y) := \frac{m}{2}\langle y, y\rangle + G(x) - G(x_0)$$

(with an appropriate translation of the potential energy) is a Lyapunov function at the rest point. Thus, the rest point is stable.

The proof of the first part of Lyapunov's theorem is simple: Choose an open ball B with center at the rest point and radius sufficiently small so that its closure is in the domain of the Lyapunov function L. The continuous positive function L on the closed and bounded spherical boundary of B has a non-zero minimum value b. Because the continuous function L vanishes at the rest point, there is a second ball A that is concentric with B, has strictly smaller radius, and is such that the maximum value of L on all of A is smaller than b. A solution of the differential equation starting in A cannot reach the boundary of B because L does not increase along solutions. Thus, the rest point is stable.

Lyapunov's method extends to nonautonomous systems with an appropriate modification of the notion of a Lyapunov function. Suppose that the first-order differential Eq. (1) has a rest point in the sense that for some $T \in \mathbb{R}$, $f(u_0, t) = 0$ for all $t \geq T$. A Lyapunov function is a class C^1 function defined in a neighborhood of u_0 for all $t \geq T$ such that $L(u_0, t) = 0$ for $t \geq T$, there is a continuous nonnegative function M defined on the same neighborhood of u_0 such that $L(u, t) \geq M(u)$ for all $t \geq T$ and $\dot{L} = L_t + L_u \cdot \dot{u} \leq 0$ for all $t \geq T$. If such a Lyapunov function exists, then the rest point is stable. If, in addition, there is a positive function N with the same domain as M such that $\dot{L}(u, t) \leq -N(u)$ for all $t \geq T$, then the rest point is asymptotically stable.

Lyapunov's direct method can be used to prove the principle of linearized stability for rest points of autonomous systems. To see how this might be done, consider the differential equation

$$\dot{u} = Au + g(u)\,, \quad u \in \mathbb{R}^n\,,$$

where A is a real $n \times n$ matrix and $g\colon \mathbb{R}^n \to \mathbb{R}^n$ is a smooth function. Suppose that every eigenvalue of A has negative real part, and that for some $a > 0$, there is a constant $k > 0$ such that, using the usual norm on $\mathbb{R}^n$,

$$|g(x)| \leq k|x|^2$$

whenever $|x| < a$. Let $\langle \cdot, \cdot \rangle$ denote the usual inner product on $\mathbb{R}^n$, and let A^* denote the transpose of the real matrix A. Suppose that there is a real symmetric positive definite $n \times n$ matrix that also satisfies Lyapunov's equation

$$A^*B + BA = -I$$

and define $L\colon \mathbb{R}^n \to \mathbb{R}$ by

$$L(x) = \langle x, Bx\rangle\,.$$

Using Schwarz's inequality, it can be proved that the restriction of L to a sufficiently small neighborhood of the origin is a strict Lyapunov function. The proof is completed by showing that there is a symmetric positive-definite solution of Lyapunov's equation. In fact,

$$B := \int_0^\infty \mathrm{e}^{tA^*}\mathrm{e}^{tA}\,\mathrm{d}t$$

is a symmetric positive definite $n \times n$ matrix that satisfies Lyapunov's equation. Alternatively, it is possible to prove the existence of a solution to Lyapunov's equation using purely algebraic methods (see, for example, [97]).

The instability of solutions can also be detected by Lyapunov's methods. A simple result in this direction is the content of the following theorem.

Theorem 24 *Suppose that L is a smooth function defined on an open neighborhood U of the rest point u_0 of the autonomous differential Eq. (23) such that $L(u_0) = 0$ and $\dot{V}(u) > 0$ on $U \setminus \{u_0\}$. If L has a positive value somewhere in each open set containing u_0, then u_0 is unstable.*

One indication of the subtlety of the determination of stability is the insolvability of the center-focus problem [41].

Consider a planar system of first-order differential equations in the form

$$\dot{x} = y + P(x, y)\,, \quad \dot{y} = -x + Q(x, y)\,,$$

where P and Q are analytic at the origin with leading-order terms at least quadratic in x and y. Is the origin a focus (that is, asymptotically stable or asymptotically unstable) or a center (that is, all orbits in some neighborhood of the origin are periodic)? There is no algorithm that can be used to solve this problem for all such systems in a finite number of steps. On the other hand, the center-focus problem is solved, by using Lyapunov's stability theorem, if a certain sequence of numbers called Lyapunov quantities (which can be computed iteratively and algebraically) has a non-zero element. One problem is that the algorithm for computing the Lyapunov quantities may not terminate in a finite number of steps (see, for example, [19,89]).

The center-focus problem is solved for some special cases. The most important theorem in this area of research is due to Nikolai Bautin (see [14] and [99,100] for a modern proof); he found (among other results) a complete solution of the center-focus problem for quadratic systems (that is, where P and Q are homogeneous quadratic polynomials). While the center-focus problem for quadratics had been solved by Dulac [25] and others [28,43,44,45] before his paper appeared; Bautin introduced a method, which involves the use of polynomial ideals, that has had far reaching consequences.

Stability in Conservative Systems and the KAM Theorem

Stability theory for conservative systems leads to many delicate problems, some of which—for example Lagrange's problem—can be resolved by the method of Lyapunov functions. While the total energy, the total angular momentum, or other first integrals all have the property that their derivatives vanish along trajectories, these integrals do not always serve as Lyapunov functions because they often fail to be positive definite in punctured neighborhoods of rest points or periodic orbits. Thus, other methods are required.

As in the quest to determine the stability of the solar system, a basic problem in mechanics is to determine the stability of periodic motions (or rest points) of conservative differential equations with respect to small perturbations. While no general solution is known, great progress has been made culminating in the Kolmogorov–Arnold–Moser (KAM) theorem (see [5,46,47,68,87,89,92]).

A mechanical system with N degrees-of-freedom (that is, the positions are determined by N-coordinates) is called completely integrable if it has N independent first integrals whose Poisson brackets are in involution (see p. 271 in [8]). In this case, the system can be transformed to a first-order system of differential equations in the simple form

$$\dot{I} = 0, \quad \dot{\theta} = \Omega(I)\,, \tag{25}$$

where I is an N-dimensional variable (of actions), θ is an N-dimensional variable (of angles), and Ω is a smooth function. The new coordinates are called action-angle variables (see, for generalizations, [39,42]). This system is in Hamiltonian form ($\dot{I} = \partial H/\partial\theta$, $\dot{\theta} = -\partial H/\partial I$) with Hamiltonian $H(I, \theta) := -\omega(I)$, where ω is an anti-derivative of Ω; and, of course, the Hamiltonian is constant along solutions of the system.

The corresponding perturbed system is taken to have the form

$$\dot{I} = \epsilon\frac{\partial F}{\partial\theta}(I, \theta), \quad \dot{\theta} = \Omega(I) - \epsilon\frac{\partial F}{\partial I}(I, \theta) \tag{26}$$

(or, more generally, a similar form with additional terms that are higher-order in ϵ), where F is a smooth function that is 2π-periodic in θ. This system is in Hamiltonian form with Hamiltonian

$$\mathcal{H}(I, \theta, \epsilon) := -\omega(I) + \epsilon F(I, \theta)\,.$$

The unperturbed integrable system can be solved explicitly as

$$I(t) = I_0\,, \quad \theta(t) = \Omega(I_0)t + \theta_0;$$

hence, by inspection, all orbits are periodic or quasi-periodic (when the angular variables are defined modulo 2π) according to whether or not the vector of frequencies $\Omega(I_0)$ satisfies a resonance relation, that is, $\langle M, \Omega(I_0)\rangle = 0$ for some N-vector M with integer components.

Every orbit of the unperturbed system with one degree-of-freedom is periodic; but, for example for two degrees-of-freedom, orbits are periodic with period T only if there are integers m_1 and m_2 such that

$$T\Omega_1(I_0) = 2\pi m_1, \quad T\Omega_2(I_0) = 2\pi m_2$$

or $m_1\Omega_2(I_0) - m_2\Omega_1(I_0) = 0$.

Geometrically, a choice of actions fixes an energy level and the angles correspond to an N-dimensional invariant torus. The flow on this torus is either periodic or quasi-periodic according to the existence of a corresponding resonance relation. For this reason, the tori with periodic flows

are called resonant; the others are called nonresonant. The KAM theorem states that most of the nonresonant tori in the unperturbed system (25) survive after small perturbations as invariant tori of the perturbed system (26).

Theorem 25 *If (in the Hamiltonian system (26)) $\mathcal{H}$ is sufficiently smooth, $D\Omega(I)$ is invertible, and ϵ is sufficiently small, then almost all (in the sense of Lebesgue measure) nonresonant unperturbed invariant tori persist.*

There are similar theorems for time-dependent perturbations of integrable systems and for area-preserving maps (see [8]).

For a mechanical system with N degrees-of-freedom, the perturbed invariant tori are N-dimensional manifolds embedded in the $2N$-dimensional phase space. After restriction to a regular energy surface (that is, a level set of $\mathcal{H}$ corresponding to one of its regular values), these N-dimensional tori are embedded in a $(2N-1)$-dimensional manifold. For $N \leq 2$, the invariant tori that exist by the KAM theorem separate the energy hypersurface into two components, one inside and one outside of each invariant torus. A motion starting in a bounded region whose boundary is one of these invariant tori cannot escape to the corresponding unbounded component. Since the invariant tori are dense, such motions are stable (cf. Fig. 6). More generally, each perturbed invariant torus is stable as an invariant set. This result can be used, for example, to prove the stability of certain periodic orbits in the restricted three-body problem (see [54]). On the other hand, for $N > 2$, the invariant tori no longer separate energy surfaces into bounded and unbounded components. Thus, it is possible for orbits to migrate outside nearby tori in a conjectured process called Arnold diffusion (see [6]). The validity of Arnold diffusion for the general Hamiltonian system (26) is an area of current research (see, for a review, [60]).

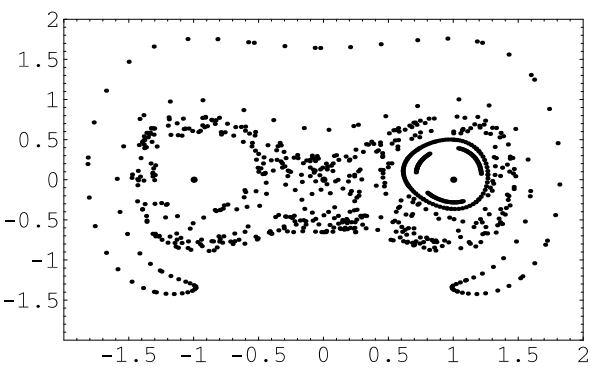

Stability Theory of Ordinary Differential Equations, Figure 6
The figure depicts portions of seven orbits of the (stroboscopic) Poincaré map $\xi \in \mathbb{R}^2 \mapsto \phi(2\pi, \xi)$ for the forced oscillator $\dot{x} = y$, $\dot{y} = x - x^3 + 0.05 \sin t$: three fixed points at (x, y) equal to $(-1, 0)$, $(0, 0)$, and $(1, 0)$; an orbit starting at $(x, y) = (1.0, 0.4)$ at time $t = 0$, which is close to a (3 : 1) resonance; an orbit starting at $(1.0, 0.5)$, which is on an invariant torus that surrounds the resonant orbit; an orbit starting at $(1.0, 0.6)$, which appears to be "chaotic"; and an orbit starting at $(1, 1)$, which is on a large invariant torus. The (3 : 1) resonant orbit is stable

Averaging and the Stability of Perturbed Periodic Orbits

As we have seen, systems of the form

$$\dot{I} = \epsilon P(I, \theta)\,, \quad \dot{\theta} = \Omega(I) + \epsilon Q(I, \theta)\,, \tag{27}$$

where P and Q are 2π-periodic, often arise in mechanics. Here, the properties of system (27) are considered for general perturbations, which are not necessarily conservative.

Note that, in case ϵ is small, the vector of actions (components of I) in the differential Eq. (27) evolves slowly relative to the evolution of the vector of angles θ. The averaging principle states that the evolution of the actions for such a system is well-approximated by the corresponding averaged equation given by

$$\dot{\mathcal{I}} = \epsilon \bar{P}(\mathcal{I})\,, \tag{28}$$

where

$$\bar{P}(\mathcal{I}) := \frac{1}{(2\pi)^N} \int_{\mathbb{T}^N} P(\mathcal{I}, \theta)\, \mathrm{d}\theta$$

and $\mathbb{T}^N$ denotes the N-dimensional torus of angles.

Exactly what is meant by "well-approximated" is clarified by the averaging theorem, which requires a severe restriction: the vector of angles is one-dimensional (cf. [61,86]).

Theorem 26 *If θ is a scalar variable and Ω is bounded away from zero, then there is a near-identity change of variables of the form $\mathcal{I} = I + \epsilon k(I, \theta)$ that is 2π-periodic in θ which transforms system (27) into the form*

$$\dot{\mathcal{I}} = \epsilon \bar{P}(\mathcal{I}) + \epsilon^2 P_1(\mathcal{I}, \theta, \epsilon)\,, \quad \dot{\theta} = \Omega(\mathcal{I}) + \epsilon Q_1(\mathcal{I}, \theta, \epsilon)\,.$$

where P_1 and P_2 are 2π-periodic in θ. Moreover, if the evolution of I starting at I_0 is given by $I(t)$ and the evolution (according to the differential Eq. (28)) of the averaged action $\mathcal{I}$ starting at I_0 is given by $\mathcal{I}(t)$, then there are constants $C > 0$ and $T(I_0) > 0$ such that

$$|\mathcal{I}(t) - I(t)| \leq C\epsilon$$

on the time interval $0 \leq \epsilon t < T(I_0)$.

The existence and stability of perturbed period solutions is the content of the following theorem.

Theorem 27 *Consider the system*

$$\dot{I} = \epsilon F(I,\theta) + \epsilon^2 F_2(I,\theta,\epsilon)\,, \quad \dot{\theta} = \Omega(I) + \epsilon G(I,\theta,\epsilon)\,, \tag{29}$$

where $I \in \mathbb{R}^M$, θ is a scalar, F, F_2, and G are 2π-periodic functions of θ, and there is some number c such that $\Omega(I) \geq c > 0$. If the averaged system has a nondegenerate rest point (see Definition 6) and ϵ is sufficiently small, then system (29) has a periodic orbit. If in addition $\epsilon > 0$ and the rest point is hyperbolic, then the periodic orbit has the same stability type as the hyperbolic rest point; that is, the dimensions of the corresponding stable and unstable manifolds are the same (see Definition 29).

For a typical application of Theorem 27, consider the system of differential equations

$$\begin{aligned} \dot{I} &= \epsilon(a + b\sin k\theta)\sin\psi + \epsilon^2 P(I,\psi,\theta), \\ \dot{\psi} &= \epsilon c I + \epsilon^2 Q(I,\psi,\theta), \\ \dot{\theta} &= 1, \end{aligned} \tag{30}$$

where a, b, and c are non-zero constants, k is a positive integer, and both P and Q are 2π-periodic in θ. The averaged system of differential equations

$$\dot{I} = \epsilon a \sin\psi\,, \quad \dot{\psi} = \epsilon c I$$

has hyperbolic rest points at $(I,\psi) = (0,0)$ and $(I,\psi) = (0,\pi)$. According to Theorem 27, if ϵ is sufficiently small, the system of differential Eq. (30) has corresponding hyperbolic periodic solutions. In case $ac > 0$, the solution of the perturbed system corresponding to $\psi = 0$ is stable, and the solution corresponding to $\psi = \pi$ is unstable (in fact, it is a hyperbolic saddle). If $ac < 0$, the stability types are switched.

Structural Stability

An important aspect of the global theory of dynamical systems is the stability of the orbit structure as a whole. The motivation for the corresponding theory comes from applied mathematics. Mathematical models always contain simplifying assumptions. Dominant features are modeled; supposed small disturbing forces are ignored. Thus, it is natural to ask if the qualitative structure of the set of solutions—the phase portrait—of a model would remain the same if small perturbations were included in the model. The corresponding mathematical theory is called structural stability.

The correct mathematical formulation of the definition of structural stability requires the introduction of a topology on the space of dynamical systems under consideration. The natural setting for the theory is the space of class C^1 vector fields on a smooth compact manifold.

There is a one-to-one correspondence between vector fields and differential equations: the right-hand side of a first-order autonomous differential equation defines a vector field on $\mathbb{R}^n$. The local representatives, with respect to coordinate charts, of a vector field on an n-dimensional manifold M are vector fields on $\mathbb{R}^n$. The set of all smooth vector fields $\mathfrak{X}(M)$ on M has the structure of a Banach space after the introduction of a Riemannian metric. Thus, two vector fields are close if the distance between them in this Banach space is small.

Definition 28 A vector field X on a smooth manifold M is called structurally stable, if there is some neighborhood U of X in $\mathfrak{X}(M)$ such that for each $Y \in U$ there is a homeomorphism of M that maps orbits of Y to orbits of X and preserves the orientation of orbits according to the direction of time. Such a homeomorphism is called a topological equivalence.

In other words, all perturbations of a structurally stable vector field have the same qualitative orbit structures.

At first impression, it might seem that the homeomorphism in the definition of structural stability should be a diffeomorphism. But, this requirement is too strong. To see why, consider a hyperbolic rest point u_0 of the first-order autonomous differential Eq. (23) in $\mathbb{R}^n$. Let $\dot{u} = g(u)$ be a differential equation on $\mathbb{R}^n$, and consider the perturbation of the differential Eq. (23) given by $\dot{u} = f(u) + \epsilon g(u)$. We have that $f(u_0) = 0$. Thus, $F(u,\epsilon) := f(u) + \epsilon g(u)$ is such that $F(u_0,0) = 0$ and $DF(u_0,0) = Df(u_0)$. Because of the hyperbolicity, the linear transformation $Df(u_0)$ is invertible. By an application of the implicit function theorem, there is a smooth function $\beta\colon J \to \mathbb{R}^n$, where J is an open interval in $\mathbb{R}$ containing the origin, such that $F(\beta(\epsilon),\epsilon)) = 0$ for all ϵ in J. In other words, every small perturbation of (the vector field) f has a rest point near u_0. Moreover, for sufficiently small ϵ (and in view of the continuity of eigenvalues with respect to the components of the corresponding matrix), the perturbed rest point $\beta(\epsilon)$ is hyperbolic with the same number of eigenvalues (counting multiplicities) with positive and negative real parts as the eigenvalues of $Df(u_0)$. But, of course, the perturbed eigenvalues are in general not the same as the eigenvalues of $Df(u_0)$. The hyperbolic rest point u_0 is structurally stable in the sense that the local phase portraits at the corresponding rest points of sufficiently small perturbations of f are qual-

itatively the same as the phase portrait of the differential Eq. (23). This fact can be proved using the Grobman–Hartman theorem together with an argument to show that the phase portraits of hyperbolic linear first-order systems are topologically equivalent if their system matrices have the same numbers of eigenvalues (counting multiplicities) with positive and negative real parts. Suppose there were a diffeomorphism h taking orbits to orbits. It would correspond to a change of variables $v = h(u)$ (a smooth conjugacy) taking the differential Eq. (23) to the differential equation $\dot{v} = F(v)$, where the parameter ϵ is suppressed for simplicity. By differentiation of the equation $v = h(u)$ with respect to the independent variable t, it follows that $F(h(u)) = Dh(u)f(u)$. By linearization at the rest point u_0, we have

$$DF(h(u_0))Dh(u_0) = Dh(u_0)Df(u_0)\,.$$

In particular the system matrix $Df(h(u_0))$ for the linearized system at the perturbed rest point is similar, via the invertible matrix $Dh(u_0)$, to the system matrix $Df(u_0)$ for the linearized system at the unperturbed rest point. Thus, the two rest points would have to have exactly the same eigenvalues, contrary to the general situation for which the eigenvalues of the perturbed linearization are different from the eigenvalues of the unperturbed linearization.

The notion of hyperbolicity is an essential hypothesis in structural stability theory; indeed, hyperbolic rest points are structurally stable. Likewise hyperbolic periodic orbits, defined to be those periodic orbits such that there is an associated hyperbolic Poincaré map are structurally stable. This notion can be extended to general invariant sets. The essential requirement is that all of the solutions of the linearized differential equations along the solutions in the invariant set, which are not initially in the direction of the invariant set, either grow or decay exponentially with uniform exponential growth rates over the entire invariant set.

The global theory of structural stability requires a global hypothesis to ensure that the rest points and period orbits are connected with orbits in general position.

Definition 29 Let A be an invariant set (a union of orbits) of an autonomous first-order differential equation, vector field, or discrete dynamical system. The stable manifold of A, denoted $W^s(A)$ is the set of all solutions that converge to A as $t \to \infty$. The unstable manifold $W^u(A)$ is the set of all solutions that converge to A as $t \to -\infty$.

The stable manifold theorem asserts that the stable and unstable manifolds for rest points and periodic orbits of smooth vector fields are indeed (immersed) smooth manifolds (see, for example, [37,38]).

Two immersed manifolds are called transversal if they do not intersect; or, if the sum of their tangent spaces at each point of their intersection is the tangent space of the ambient manifold.

While periodicity is the most familiar form of recurrence for dynamical systems, a more subtle notion of recurrence is required for structural stability theorems.

Definition 30 The point u is called a nonwandering point for an autonomous differential equation if, for each neighborhood U of u and each time $t_0 > 0$, there is some time $t > t_0$ such that at least one solution starting in U at time zero returns to U at time t. Likewise for a discrete dynamical system defined by a diffeomorphism h, the point u is nonwandering if for each neighborhood U of u there is some integer $k > 0$ such that $h^k(U) \cap U$ is not empty. The set of all nonwandering points is called the nonwandering set.

For vector fields on compact orientable two-dimensional manifolds, the classical structural stability theorem is due to M. Peixoto [70] and is a generalization of earlier work for vector fields in the plane by L. Pointryagin and A. Andronov [2,3].

Theorem 31 *A class C^1 vector field on a compact orientable two-dimensional manifold M is structurally stable if and only if*

(1) all rest points are hyperbolic,
(2) all periodic solutions are hyperbolic,
(3) stable and unstable manifolds of all pairs of saddle points (rest points whose linearizations have one positive and one negative eigenvalue) are disjoint,
(4) the nonwandering set consists of only rest points and periodic solutions.

Moreover, the structurally stable vector fields form an open and dense set in $\mathfrak{X}(M)$.

The notion of transversality in Peixoto's theorem is embodied in the property (3). The generalization of this property is an essential ingredient of the theory.

Definition 32 A dynamical system satisfies the strong transversality property if for every pair of points $\{u, v\}$ in its nonwandering set, the corresponding invariant manifolds $W^s(u)$ and $W^u(y)$ intersect transversally (at all their points of intersection).

Definition 33 A dynamical system satisfies Axiom A if its nonwandering set is hyperbolic and the periodic orbits are dense in this set.

The most important result of the theory is the structural stability theorem:

Theorem 34 *A class C^1 diffeomorphism is structurally stable if and only if it satisfies Axiom A and the strong transversality property.*

Joel Robbin [77] proved that a C^2 diffeomorphism is C^1 structurally stable if it satisfies Axiom A and the strong transversality property. This result was improved by Clark Robinson [81] by relaxing the C^2 requirement to C^1; he also proved the analogous result for differential equations [79,80]. Ricardo Mañé [65] proved the converse.

A simple example of a structurally stable diffeomorphism is the north-pole south-pole map of the Riemann sphere; it is given by $z \mapsto 2z$ for the complex variable z. There are two hyperbolic rest points, an unstable fixed point at $z = 0$ and a stable fixed point at $z = \infty$. All other points are attracted to ∞ under forward iteration and to zero under backward (inverse) iteration.

A more exotic example is provided by the hyperbolic toral automorphisms. To define these maps, note that a 2×2-integer matrix with determinant one determines a diffeomorphism of the plane that preserves the integer lattice. Thus, it projects to a map of the torus formed by the quotient space $\mathbb{R}^2/\mathbb{Z}^2$. In case the matrix is hyperbolic, the corresponding diffeomorphism on the torus has a dense set of periodic orbits and thus its nonwandering set is the entire torus. It can be proved that the linear hyperbolic structure of the transformation of the plane (that is, one stable and one unstable eigenspace) projects to a uniform hyperbolic structure on the entire torus; and, moreover, the strong transversality condition is satisfied. Thus, small perturbations of such a map have the same properties. This example was generalized by Dmitri Anosov and others to include a class of diffeomorphisms and vector fields in the continuous case now called Anosov dynamical systems. The most important example of a continuous flow with this property is the geodesic flow in the unit tangent bundle of a compact Riemannian manifold with negative sectional curvatures (see [4]).

Attractors

A natural outgrowth of the theory of stability for rest points and periodic solutions is its generalization to invariant sets. The fundamental objects of study are the attractors. Unfortunately, the definition of this concept is not universally accepted; different authors use different definitions. The following definition is perhaps the most popular:

Definition 35 A closed invariant set A for a dynamical system is called an attracting set if it is contained in an open neighborhood U of the ambient space such that the intersection of the sequence of all forward motions of U under the action of the dynamical system is equal to A. An attracting set is called an attractor if, in addition, there is no closed invariant attracting subset of A.

The existence of attractors is part of the stability theory of dynamical systems; the structure of attractors is the subject of another rich theory. For most of the history of differential equations the only known attractors (the classical attractors) were rest points, periodic orbits, and tori, which are all manifolds. The existence of attractors with fractal dimensions, called strange attractors, was noticed and proved only recently [62,82,85,95].

For differential equations in $\mathbb{R}^n$, the existence of an attractor for autonomous systems is usually proved by demonstrating the existence of an $(n-1)$-dimensional sphere (or other separating hypersurface) such that the vector field corresponding to the differential equation points into the bounded region of space bounded by the sphere (more precisely, the inner product of the vector field and the inner normal is positive at all points on the sphere).

As an example, consider the Lorenz equations on $\mathbb{R}^3$

$$\begin{aligned} \dot{x} &= \sigma(y - x), \\ \dot{y} &= rx - y - xz, \\ \dot{z} &= xy - bz, \end{aligned}$$

where σ, r, and b are positive constants. The origin is a rest point for the Lorenz system and, if $r < 1$, then

$$L(x, y, z) := \frac{1}{\sigma}x^2 + y^2 + z^2$$

is a Lyapunov function. In this case, the origin is globally asymptotically stable. More generally (that is, with no restriction on the size of r), there is a constant $c > 0$ such that the vector field points into the bounded region of space bounded by the ellipsoid $rx^2 + \sigma y^2 + \sigma(z - 2r)^2 = c$. Thus, the Lorenz equations have an attractor. For some parameter values, for example $\sigma = 10$, $b = 8/3$, and $r = 28$, the attractor is a fractal, called the Lorenz attractor (see [29,62,93,95]).

For the special case of two-dimensional first-order autonomous systems, the existence and nature of attractors can be more precisely specified. The main result is the Poincaré-Bendixson theorem (see [15] and, for a proof, [1]):

Theorem 36 *An attractor of the class C^1 differential Eq. (23) on the plane that contains no rest points is a periodic orbit.*

Corollary 37 *If the differential Eq. (23) on the plane is analytic and has a positively invariant annulus containing no rest points, then the annulus contains an orbitally asymptotically stable periodic solution (a stable limit cycle).*

As an example, consider the model of a damped pendulum with torque given by the first-order planar system

$$\dot{\theta} = v\,, \quad \dot{v} = -\sin\theta + \mu - \lambda v\,, \tag{31}$$

where λ and μ are positive constants. By viewing the variable θ as an angular variable modulo 2π, the phase space is the cylinder rather than the plane. Nonetheless, the Poincaré-Bendixson theory is valid on the cylinder because a bounded region of the cylinder can be flattened by a change of variables to a region of the plane. If $|\mu| > 1$, then the system of differential Eq. (31) has a globally attracting periodic orbit. Three main observations can be used to construct a proof: there are no rest points, the quantity $-\sin\theta + \mu - \lambda v$ is negative for sufficiently large values of $v > 0$, and it is positive for negative values of v with sufficiently large absolute values. These facts imply the existence of an invariant annulus containing no rest points. The remainder of the proof uses Theorem 19 and the analyticity of solutions (see p. 98 in [19]).

Generalizations and Future Directions

The concepts of stability theory have been successfully generalized to infinite-dimensional dynamical systems (which arise from the analysis of partial differential equations) defined on Banach spaces and to abstract dynamical systems on metric spaces [13,32,33,36,83,94]. Research in this direction has produced some results on stability and the existence of attractors.

No general theorem akin to Lyapunov's Theorem 4 is known for infinite-dimensional dynamical systems. A main difficulty is that, for an unbounded operator A, the non-zero elements in the spectrum of e^A may not all be given by the exponentials of elements in the spectrum of A (see, for a discussion, Chap. 2 in [20]). Thus, to determine the stability of a rest point by linearization, this spectral mapping property must be checked separately for most cases of interest.

Part of the motivation for the determination of attractors in infinite-dimensional dynamical systems is the possibility of reducing the long-term dynamical properties to the analysis of a finite-dimensional dynamical system on the attractor. The fundamental organizing concept in this direction is the "inertial manifold".

Definition 38 An inertial manifold is a finite-dimensional Lipschitz manifold that is invariant under forward motions of the dynamical system and attracts all solutions of the dynamical system at an exponential rate.

Inertial manifolds have been proved to exist for many dynamical systems defined by evolution-type partial differential equations, for example, reaction-diffusion equations with appropriate boundary conditions. Much current research is devoted to determining conditions that imply the existence of inertial manifolds (or their generalizations) for the equations of fluid dynamics.

The conceptual framework and basic abstract theory of stability is mature and well-understood for finite-dimensional dynamical systems. This theory is currently being extended to an abstract theory for infinite-dimensional dynamical systems. The most important direction for further development is the application of this theory to specific differential equations (including partial differential equations) that arise as mathematical models in applied mathematics. Even the problem that originally motivated stability theory (that is, the stability of periodic and quasi-periodic motions of the N-body problem of classical mechanics) remains open to future research.

Bibliography

Primary Literature

1. Alligood KT, Sauer TD, Yorke JA (1997) Chaos: An introduction to dynamical systems. Springer, New York
2. Andronov AA, Vitt EA, Khaiken SE (1966) Theory of Oscillations. Pergamon Press, Oxford
3. Andronov AA, Leontovich EA, Gordon II, Maier AG (1973) Qualitative Theory of Second-Order Dynamic Systems. Wiley, New York
4. Anosov DV (1967) Geodesic Flows on Closed Riemannian Manifolds with Negative Curvature. Proc Steklov Inst Math 90:1–209
5. Arnold VI (1963) Proof of an Kolmogorov's theorem on the preservation of quasi-periodic motions under small perturbations of the Hamiltonian. Usp Mat Nauk SSSR 113(5):13–40
6. Arnold VI (1964) Instability of dynamical systems with many degrees of freedom. Soviet Math 5(3):581–585
7. Arnold VI (1973) Ordinary Differential Equations. MIT Press, Cambridge
8. Arnold VI (1978) Mathematical Methods of Celestial Mechanics. Springer, New York
9. Arnold VI (1982) Geometric Methods In: The Theory of Ordinary Differential Equations. Springer, New York
10. Arnold VI (ed) (1988) Dynamical Systems I. Springer, New York
11. Arnold VI (ed) (1988) Dynamical Systems III. Springer, New York
12. Arscott FM (1964) Periodic Differential Equations. MacMillan, New York
13. Babin AV, Vishik MI (1992) Attractors of Evolution Equations. North-Holland, Amsterdam

14. Bautin NN (1954) On the number of limit cycles which appear with the variation of coefficients from an equilibrium position of focus or center type. Am Math Soc Transl 100:1–19
15. Bendixson I (1901) Sur les coubes définiés par des équations différentielles. Acta Math 24:1–88
16. Brjuno AD (1971) Analytic form of differential equations. Trudy MMO 25:119–262
17. Chang K-C (2005) Methods in Nonlinear Analysis. Springer, New York
18. Chetayev NG (1961) The Stability of Motion. Pergamon Press, New York
19. Chicone C (2006) Ordinary Differential Equations with Applications, 2nd edn. Springer, New York
20. Chicone C, Latushkin Y (1999) Evolution Semigroups. In: Dynamical Systems and Differential Equations. Am Math Soc, Providence
21. Chicone C, Liu W (2004) Asymptotic phase revisited. J Diff Eq 204(1):227–246
22. Chicone C, Swanson R (2000) Linearization via the Lie derivative. Electron J Diff Eq Monograph 02:64
23. Dirichlet GL (1846) Über die Stabilität des Gleichgewichts. J Reine Angewandte Math 32:85–88
24. Dirichlet GL (1847) Note sur la stabilité de l'équilibre. J Math Pures Appl 12:474–478
25. Dulac H (1908) Déterminiation et intégratin d'une certaine classe d'equatins différentielles ayant pour point singulier un centre. Bull Soc Math Fr 32(2):230–252
26. Dumortier F (2006) Asymptotic phase and invariant foliations near periodic orbits. Proc Am Math Soc 134(10):2989–2996
27. Floquet G (1883) Sur les équations différentielles linéaires à coefficients péiodiques. Ann Sci École Norm Sup 12(2):47–88
28. Frommer M (1934) Über das Auftreten von Wirbeln und Strudeln (geschlossener und spiraliger Integralkurven) in der Umgebung rationaler Unbestimmtheitsstellen. Math Ann 109:395–424
29. Guckenheimer J, Holmes P (1986) Nonlinear Oscillations, Dynamical Systems, and Bifurcation of Vector Fields, 2nd edn. Springer, New York
30. Grobman D (1959) Homeomorphisms of systems of differential equations. Dokl Akad Nauk 128:880–881
31. Grobman D (1962) Topological classification of the neighborhood of a singular point in *n*-dimensional space. Mat Sb 98:77–94
32. Hale JK (1988) Asymptotic Behavior of Dissipative Systems. Am Math Soc, Providence
33. Hale JK, Magalhães LT, Oliva WM (2002) Dynamics in Infinite Dimensions, 2nd edn. Springer, New York
34. Hartman P (1960) A lemma in the theory of structural stability of differential equations. Proc Am Math Soc 11:610–620
35. Hartman P (1960) On local homeomorphisms of Euclidean space. Bol Soc Math Mex 5(2):220–241
36. Henry D (1981) Geometric Theory of Semilinear Paabolic Equations. Lecture Notes in Mathematics, vol 840. Springer, New York
37. Hirsch M, Pugh C (1970) Stable manifolds and hyperbolic sets In: Global Analysis XIV. Am Math Soc, Providence, pp 133–164
38. Hirsch MC, Pugh C, Shub M (1977) Invariant Manifolds. Lecture Notes in Mathematics, vol 583. Springer, New York
39. Hofer H, Zehnder E (1994) Symplectic invariants and Hamiltonian dynamics. Birkhäuser, Basel
40. Hurwitz A (1895) Über die Bedingungen, unter welchen eine Gleichung nur Wurzeln mit negativen reellen Theilen besitzt. [On the conditions under which an equation has only roots with negative real parts.] Math Ann 46:273–284; (1964) In: Ballman RT et al (ed) Selected Papers on Math Trends in Control Theory. Dover, New York (reprinted)
41. Ilyashenko JS (1972) Algebraic unsolvability and almost algebraic solvability of the problem for the center-focus. Funkcional Anal I Priložen 6(3):30–37
42. Jost R (1968) Winkel- und Wirkungsvariable für allgemeine mechanische Systeme. Helvetica Phys Acta 41:965–968
43. Kapteyn W (1911) On the centra of the integral curves which satisfy differential equations of the first order and the first degree. Proc Kon Akak Wet Amsterdam 13:1241–1252
44. Kapteyn W (1912) New researches upon the centra of the integrals which satisfy differential equations of the first order and the first degree. Proc Kon Akak Wet Amsterdam 14:1185
45. Kapteyn W (1912) New researches upon the centra of the integrals which satisfy differential equations of the first order and the first degree. Proc Kon Akak Wet Amsterdam 15:46–52
46. Kolmogorov AN (1954) La théorie générale des systèmes dynamiques et la méchanique. Amsterdam Congress I, pp 315–333
47. Kolmogorov AN (1954) On the conservation of conditionally periodic motions under small perturbations of the Hamiltonian. Dokl Akad Nauk SSSR 98:527–530
48. Lagrange JL (1776) Nouv Mem Acad Roy Sci. Belles-Lettres de Berlin, p 199
49. Lagrange JL (1788) Mechanique Analitique. Desaint, Paris, pp 66–73, pp 389–427; (1997) Boston Studies in the Philosophy of Science 191. Kluwer, Dordrecht (reprinted)
50. Lagrange JL (1869) Oeuvres de Lagrange. Gauthier, Paris 4:255
51. LaSalle JP (1976) The Stability of Dynamical Systems. SIAM, Philadelphia
52. Laplace PS (1784) Mem Acad Roy Sci. Paris, p 1
53. Laplace PS (1895) Oeuvres Complètes de Laplace. Gauthier, Paris 11:47
54. Leontovich AM (1962) On the stability of the Lagrange periodic solutions for the reduced problem of three bodies. Sov Math Dokl 3(2):425–430
55. Liapounoff AM (1947) Problème général de la stabilité du movement. Princeton University Press, Princeton
56. Liouville J (1842) Extrait d'une lettre à M Arago sur le Mémoire de M Maurice relatif à l'invariabilité des grands axes des orbites planétaires. Compt Rend Séances l'Acad Sci 15:425–426
57. Liouville J (1842) Extrait d'un Mémoire sur un cas particulier du problème des trois corps. J Math Sér I 7:110–113
58. Liouville J (1843) Sur la loi de la pesanteur à la surface ellipsoïdale d'équilibre d'une masse liquide homogène douée d'un mouvement de rotation. J Math Sér I 8:360
59. Liouville J (1855) Formules générales relatives à la question de la stabilité de l'équilibre d'une masse liquide homogène douée d'un mouvement de rotation autour d'un axe. Sér I, J Math Pures Appl 20:164–184
60. Lochak P (1999) Arnold diffusion; a compendium of remarks and questions. Hamiltonian systems with three or more degrees of freedom (S'Agaró 1995). In: NATO Adv Sci Inst Ser C Math Phys Sci, 533. Kluwer, Dordrecht, pp 168–183

61. Lochak P, Meunier C (1988) Multiphase averaging for classical systems. In: Trans H.S. Dumas. Springer, New York
62. Lorenz EN (1963) Deterministic nonperiodic flow. J Atm Sci 20:130–141
63. Lyapunov AM (1892) The General Problem of the Stability of Motion. Russ Math Soc, Kharkov (russian); (1907) Ann Facult Sci l'Univ Toulouse 9:203–474 (french trans: Davaux É); (1949) Princeton University Press, Princeton (reprinted); (1992) Taylor & Francis, London (english trans: Fuller AT)
64. Magnus W, Winkler S (1979) Hill's Equation. Dover, New York
65. Mañé R (1988) A proof of the C^1 stability conjecture. Inst Hautes Études Sci Publ Math 66:161–210
66. Marcus L, Yamabe H (1960) Global stability criteria for differential equations. Osaka Math J 12:305–317
67. Melnikov VK (1963) On the stability of the center for time periodic perturbations. Trans Mosc Math Soc 12:1–57
68. Moser J (1962) On invariant curves of area-preserving mappings of an annulus. Nachr Akad Wiss Göttingen, Math Phys K1:1–10
69. Moser J (1978/79) Is the solar system stable? Math Intell 1: 65–71
70. Peixoto MM (1962) Structural stability on two-dimensional manifolds. Topology 1:101–120
71. Poincaré H (1881) Mémoire sur les courbes définies par une équation différentielle. J Math Pures Appl 7(3):375–422; (1882) 8:251–296; (1885) 1(4):167–244; (1886) 2:151–217; all reprinted (1928) Tome I. Gauthier–Villar, Oeuvre, Paris
72. Poincaré H (1890) Sur le problème des trois corps et les équations del la dynamique. Acta Math 13:83–97
73. Poincaré H (1892) Les méthodes nouvelles de la mécanique céleste. Tome III. Invariants intégraux Solutions périodiques du deuxième genre, Solutions doublement asymptotiques. Dover, New York (1957)
74. Poincaré H (1951) Oeuvres, vol 1. Gauthier-Villans, Paris
75. Poincaré H (1957) Les méthodes nouvelles de la mécanique céleste. Tome I. Solutions périodiques. Non-existence des intègrales uniformes, Solutions asymptotiques. Dover, New York
76. Poincaré H (1957) Les méthodes nouvelles de la mécanique céleste. Tome II. Méthodes de MM. Newcomb, Gyldén, Lindstedt et Bohlin. Dover, New York
77. Robbin JW (1971) A structural stability theorem. Ann Math 94(2):447–493
78. Robbin JW (1995) Matrix algebra. A.K. Peters, Wellesley
79. Robinson C (1974) Structural stability of vector fields. Ann Math 99(2):154–175
80. Robinson C (1975) Errata to: "Structural stability of vector fields". (1974) Ann Math 99(2):154–175, Ann Math 101(2):368
81. Robinson C (1976) Structural Stability for C^1 diffeomorphisms. J Diff Eq 22(1):28–73
82. Robinson C (1989) Homoclinic bifurcation to a transitive attractor of Lorenz type. Nonlinearity 2(4):495–518
83. Robinson J (2001) Infinite Dimensional Dynamical Systems. Cambridge Univ Press, Cambridge
84. Routh EJ (1877) A Treatise on the Stability of a Given State of Motion. Macmillan, London; Fuller AT (ed) (1975) Stability of Motion. Taylor, London (reprinted)
85. Rychlik MR (1990) Lorenz attractors through Šil'nikov-type bifurcation I. Ergodic Theory Dynam Syst 10(4):793–821
86. Sanders JA, Verhulst F, Murdock J (2007) Averaging Methods. In: Nonlinear Dynamical Systems, 2nd edn. Appl Math Sci 59. Springer, New York
87. Siegel CL (1942) Iterations of anayltic functions. Ann Math 43:607–612
88. Siegel CL (1952) Über die Normalform analytischer Differentialgleichungen in der Nähe einer Gleichgewichtslösung. Nachr Akak Wiss Gottingen, Math Phys K1:21–30
89. Siegel CL, Moser J (1971) Lectures on Celestial Mechanics. Springer, New York
90. Sternberg S (1958) On the structure of loacal homeomorphisms of Euclidean *n*-space. Am J Math 80(3):623–631
91. Sternberg S (1959) On the structure of loacal homeomorphisms of Euclidean *n*-space. Am J Math 81(3):578–604
92. Sternberg S (1969) Celestial Mechanics. Benjamin, New York
93. Strogatz SH (1994) Nonlinear Dynamics and Chaos. Perseus Books, Cambridge
94. Temam R (1997) Infinite-Dimensional Dynamical Systems. In: Mechanics and Physics. Springer, New York
95. Viana M (2000) What's new on Lorenz strange attractors? Math Intell 22(3):6–19
96. Wiggins S (2003) Introduction to Applied Nonlinear Dynamical Systems and Chaos, 2nd edn. Springer, New York
97. Willems JL (1970) Stability Theory of Dynamical Systems. Wiley, New York
98. Yakubovich VA, Starzhinskii VM, Louvish D (trans) (1975) Linear Differential Equations with Periodic Coefficients. Wiley, New York
99. Żołądek H (1994) Quadratic systems with center and their perturbations. J Diff Eq 109(2):223–273
100. Żołądek H (1997) The problem of the center for resonant singular points of polynomial vector fields. J Diff Eq 135:94–118

Books and Reviews

Arnold VI, Avez A (1968) Ergodic Problems of Classical Mechanics. Benjamin, New York
Coddington EA, Levinson N (1955) Theory of ordinary differential equations. McGraw-Hill, New York
Farkas M (1994) Periodic Motions. Springer, New York
Glendinning PA (1994) Stability, Instability and Chaos. Cambridge University Press, Cambridge
Grimshaw R (1990) Nonlinear Ordinary Differential Equations. Blackwell, Oxford
Hale JK (1980) Ordinary Differential Equations, 2nd edn. RE Krieger, Malabar
Hahn W, Baartz AP (trans) (1967) Stability of Motion. Springer, New York
Hartman P (1964) Ordinary Differential Equations. Wiley, New York
Hirsch M, Smale S (1974) Differential Equations, Dynamical Systems, and Linear Algebra. Acad Press, New York
Katok AB, Hasselblatt B (1995) Introduction to the Modern Theory of Dynamical Systems. Cambridge Univ Pres, Cambridge
Krasovskiĭ NN (1963) Stability of Motion. Stanford Univ Press, Stanford
LaSalle J, Lefschetz S (1961) Stability by Lyapunov's Direct Method with Applications. Acad Press, New York
Lefschetz S (1977) Differential Equations: Geometric Theory, 2nd edn. Dover, New York
Miller RK, Michel AN (1982) Ordinary Differential Equations. Acad Press, New York

Moser J (1973) Stable and Random Motions. In: Dynamical Systems Ann Math Studies 77. Princeton Univ Press, Princeton
Nemytskiĭ VV, Stepanov VV (1960) Qualitative Theory of Differential Equations. Princeton Univ Press, Princeton
Perko L (1996) Differential Equations and Dynamical Systems, 2nd edn. Springer, New York
Robbin JW (1970) On structural stability. Bull Am Math Soc 76:723–726
Robbin JW (1972) Topological conjugacy and structural stability for discrete dynamical systems. Bull Am Math Soc 78(6):923–952
Robinson C (1995) Dynamical Systems: Stability, Symbolic Dynamics, and Chaos. CRC Press, Boca Raton
Smale S (1967) Differentiable dynamical systems. Bull Am Math Soc 73:747–817
Smale S (1980) The Mathematics of Time: Essays on Dynamical Systems, Economic Processes and Related Topics. Springer, New York
Sotomayor J (1979) Lições de equações diferenciais ordinárias. IMPA, Rio de Janerio
Verhulst F (1989) Nonlinear Differential Equations and Dynamical Systems. Springer, New York

Static Games

Oscar Volij
Ben-Gurion University, Beer-Sheva, Israel

Article Outline

Glossary
Definition of the Subject
Introduction
Nash Equilibrium
Existence
Mixed Strategies
Equilibrium in Beliefs
Correlated Equilibrium
Rationality, Correlated Equilibrium and Equilibrium in Beliefs
Bayesian Games
Evolutionary Stable Strategies
Future Directions
Bibliography

Glossary

Player A participant in a game.
Action set The set of actions that a player may choose.
Action profile A list of actions, one for each player.
Payoff The utility a player obtains from a given action profile.

Definition of the Subject

Game theory concerns the interaction of decision makers. This interaction is modeled by means of games. There are various approaches to constructing games. One approach is to focus on the possible outcomes of the decision-makers' interaction by abstracting from the actions or decisions that may lead to these outcomes. The main tool used to implement this approach is the *cooperative game*. Another approach is to focus on the actions that the decision-makers can take, the main tool being the *non-cooperative game*. Within this approach, strategic interactions are modeled in two ways. One is by means of *dynamic*, or extensive form games, and the other is by means of *static*, or strategic games. Dynamic games stress the sequentiality of the various decisions that agents can make. An essential component of a dynamic game is the description of who moves first, who moves second, etc. Static games, on the other hand, abstract from the sequentiality of the possible moves, and model interactions as simultaneous decisions, where the decisions may well be complicated plans of actions that dictate different moves for different situations that may arise. All extensive form games can be modeled as static games, and all strategic form games can be modeled as extensive form games. But some situations may be more conveniently modeled as one or the other kind of game.

This chapter reviews the main ideas and results related to static games, as well as some interesting relationships that connect equilibrium concepts with the idea of rationality. The objective is to introduce the reader to the area of static games and to stimulate his interest for further knowledge of game theory in general. For a comprehensive exposition of some results not covered in this chapter, the reader is referred to the many excellent textbooks available on game theory. Binmore [6], Fudenberg and Tirole [9], Osborne [20], Osborne and Rubinstein [21] constitute only a partial list.

Although the definition of a static game is a very simple one, static games are a very flexible model which allows us to analyze many different situations. In particular, one can use them to analyze strategic interactions that involve either common interests or diametrically opposed interests. Similarly, one can also use static games to model situations where players have either symmetric or asymmetric information. The range of applications of static games is very wide and covers many disciplines, such as economics, political science, biology, philosophy, and computer science among others.

Introduction

In this section we introduce some examples that will be used later to motivate different concepts. We also introduce the definition of a static game.

The prisoner's dilemma involves a donor who is interested in donating some amount of money to two universities. The donor decides that the amount each university will receive depends on the content of the messages the presidents of the respective universities will send to him. Each university will send simultaneously one of two messages. One possible message is "Give him 2" and the other is "Give me 1". The donor will do exactly as told. For instance, if University I sends the message "Give me 1" and University II sends "Give him 2", the donor will donate \$3 to University I and \$0 to University II. This game can be described by means of the following matrix, where the entries represent the payoffs for University I and University II, respectively, that result from the corresponding action choices.

		University II	
		Give him 2	*Give me 1*
University I	*Give him 2*	2, 2	0, 3
	Give me 1	3, 0	1, 1

The battle of the sexes consists of two friends, She and He, who want to go out together, but have no means of communication. They have to decide, each one separately but both simultaneously, whether to go to a boxing match or to a ballet show. For both of them, the worst possible outcome would be to choose different events and not meet. But if they meet, he would rather meet her at the boxing match, while she would rather meet him at the ballet. The battle of the sexes can be described by the following matrix.

		She	
		Box	Ballet
He	Box	2, 1	0, 0
	Ballet	0, 0	1, 2

Again, the entries of this matrix represent the payoffs that he and she get, as a result of their corresponding choices.

Chicken models two drivers who approach each other on a narrow street. If none of them slows down they'll have an accident and their corresponding payoffs will be 0. But if at least one of them slows down, the accident is prevented. The problem is that both of them would like the other to slow down. If only one driver slows down, this driver gets a payoff of 2 and the other driver gets a payoff of 7. If both drivers slow down, then both drivers get a payoff of 6. This situation can be described by the following matrix.

		Driver 2	
		Slow Down	Speed up
Driver 1	Slow Down	6, 6	2, 7
	Speed up	7, 2	0, 0

Matching Pennies involves two friends, each of whom places a coin on a table. If both coins are placed heads up or tails up, then friend 1 gets one dollar from friend 2. If one coin is placed heads up and the other tails up, then friend 1 pays one dollar to friend 2. Matching pennies can be described by the following matrix, where the entries are the amounts of money that the friends get from each other.

		Friend 2	
		Heads	Tails
Friend 1	Heads	1, −1	−1, 1
	Tails	−1, 1	1, −1

The above examples of strategic interactions can be modeled as static games. A static game is a formalization of a strategic situation according to which players choose their actions separately and simultaneously, and as a result obtain certain payoffs. The interaction that a static game models need not require that players take their actions simultaneously. But the interaction is modeled by defining actions in such a way that lets us think of the players as acting simultaneously.

All of the above examples involve a set of players, and for each player there is a set of available actions and a function that associates a payoff level to each of the profiles of actions that may result from the players' choices. These are the three essential components of a static game, as formalized in the following definition.

Definition 1 A static game is a triple $\langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$ where N is a finite set of players, and for each player $i \in N$, A_i is i's set of actions, and $u_i : \times_{k\in N} A_k \to \mathbb{R}$ is player i's utility function.

In the prisoner's dilemma the set of players is $N = \{\text{University I, University II}\}$; the sets of actions are $A_{\mathrm{I}} = A_{\mathrm{II}} = \{\textit{Give me 1}, \textit{Give him 2}\}$; the utility function of University I is $u_{\mathrm{I}}(\textit{Give me 1}, \textit{Give me 1}) = 1$, $u_{\mathrm{I}}(\textit{Give me 1}, \textit{Give him 2}) = 3$, $u_{\mathrm{I}}(\textit{Give him 2}, \textit{Give me 1}) = 0$, $u_{\mathrm{I}}(\textit{Give him 2}, \textit{Give him 2}) = 2$; and the utility function of University II is $u_{\mathrm{II}}(\textit{Give me 1}, \textit{Give me 1}) = 1$, $u_{\mathrm{II}}(\textit{Give me 1}, \textit{Give him 2}) = 0$, $u_{\mathrm{II}}(\textit{Give him 2}, \textit{Give me 1}) = 3$, $u_{\mathrm{I}}(\textit{Give him 1}, \textit{Give him 1}) = 1$.

In this chapter we sometimes refer to static games simply as games. For any game $\langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$, the set

of action profiles $\times_{k\in N} A_k$ is denoted by A, and a typical action profile is denoted by $a = (a_i)_{i\in N} \in A$. If A is a finite set, then we say that the game is finite. Player i's utility function represents his preferences over the set of action profiles. For instance, for any two action profiles a and a' in A, $u_i(a) \geq u_i\left(a'\right)$ means that player i prefers action profile a to action profile a'. Clearly, although player i has preferences over action profiles, he can only affect his own component, a_i, of the profile.

Nash Equilibrium

One objective of game theory is to select, for each game, a set of action profiles that are interesting in some way. These action profiles may be interpreted as predictions of the theory, or prescriptions for the players to follow, or simply as equilibrium outcomes in the sense that if they occur, the players do not wish that they had acted differently. These action profiles are formally given by solution concepts, which are functions that associate each strategic game with the selected set of action profiles. The central solution concept in game theory is known as *Nash equilibrium*. The hypothesis behind this solution concept is that each player chooses his actions so as to maximize his utility, given the profile of actions chosen by the other players. To give a formal definition of the Nash equilibrium concept, we first introduce some useful notation. For each player $i \in N$, let $A_{-i} = \times_{k\in N\setminus\{i\}} A_k$ be the set of the other players' profiles of actions. Then we can write $A = A_i \times A_{-i}$, and each action profile can be written as $a = (a_i, a_{-i}) \in A_i \times A_{-i}$, thereby distinguishing player i's action from the other players' profile of actions.

Definition 2 The action profile $a^* = (a_i^*)_{i\in N} \in A$ in a game $\langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$ is a *Nash equilibrium* if for each player, $i \in N$, and every action $a_i \in A_i$ of player i, a^* is at least as good for player i as the action profile (a_i, a_{-i}^*). That is, if

$$u_i(a^*) \geq u_i(a_i, a_{-i}^*) \quad \text{for all } a_i \in A_i \quad \text{and for all } i \in N\,.$$

It is a *strict Nash equilibrium* if the above inequality is strict for all alternative actions $a_i \in A_i \setminus \{a_i^*\}$.

Analysis of Some Finite Games

Prisoner's Dilemma Recall that the prisoner's dilemma can be described by the following matrix.

		University II	
		Give him 2	*Give me 1*
University I	*Give him2*	2, 2	0, 3
	Give me1	3, 0	1, 1

The action profile (*Give me 1, Give me 1*) is a Nash equilibrium. Indeed,

$$u_{\mathrm{I}}(\textit{Give me 1, Give me 1}) = 1 \geq u_{\mathrm{I}}(\textit{Give him } 2,\ \textit{Give me } 1) = 0$$

and

$$u_{\mathrm{II}}(\textit{Give me } 1,\ \textit{Give me } 1) = 1 \geq u_{\mathrm{II}}(\textit{Give me } 1,\ \textit{Give him } 2) = 0\,.$$

On the other hand, the action profile (*Give him 2, Give him 2*) is not a Nash equilibrium, since University I prefers action "*Give me 1*" if University II chooses action "*Give him 2*":

$$2 = u_{\mathrm{I}}(\textit{Give him } 2,\ \textit{Give him } 2) < u_{\mathrm{I}}(\textit{Give me } 1,\ \textit{Give him } 2) = 3\,.$$

Battle of the Sexes Recall that the battle of the sexes can be described by the following matrix.

		She	
		Box	Ballet
He	Box	2, 1	0, 0
	Ballet	0, 0	1, 2

One can check that (*Box, Box*) is a Nash equilibrium and (*Ballet, Ballet*) is a Nash equilibrium as well. It can also be checked that these are the only two action profiles that constitute a Nash equilibrium.

Matching Pennies The reader can check that Matching Pennies has no Nash equilibrium.

Before we analyze the next example, we introduce a technical tool that allows us to reformulate the definition of Nash equilibrium more conveniently. More importantly, this alternative definition is the key to the standard proof of the existence of Nash equilibrium.

Definition 3 Let $G = \langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$ be a strategic game and let $i \in N$ be a player. Consider a list of actions $a_{-i} = (a, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n) \in \times_{k\in N\setminus\{i\}} A_k$ of all the players other than i. The set of player i's best responses to a_{-i} is

$$\mathcal{B}_i(a_{-i}) = \{a_i \in A_i \colon u_i(a_i, a_{-i}) \geq u_i(b_i, a_{-i}) \text{ for all } b_i \in A_i\}\,.$$

The correspondence $\mathcal{B}_i \colon \times_{j\neq i} A_j \twoheadrightarrow A_i$ that assigns to each $(n-1)$-tuple of actions in A_{-i} the set of best responses to it is called the *best response correspondence* of player i.

The definition of a Nash equilibrium may be stated in terms of the players' best response correspondences, as stated in the following proposition.

Proposition 1 *The action profile $a^* \in A$ is a Nash equilibrium if and only if every player's action is a best response to the other players' actions. That is, if*

$$a_i^* \in \mathcal{B}_i(a_{-i}^*) \quad \textit{for all } i \in N\,.$$

Until now, all the examples involved games where the action sets contained two actions. The next example is a game where the players' action sets are infinite. We will use the player's best response correspondences to find all its Nash equilibria.

The War of Attrition Two animals, 1 and 2, are fighting over a prey. Each animal chooses a time at which it intends to give up. Once one animal has given up, the other obtains the prey; if both animals give up at the same time then they split the prey equally. For each $i = 1, 2$, animal i's willingness to fight for the prey is given by $v_i > 0$. The value v_i is the maximum amount of time that animal i is willing to spend to obtain the prey. Since fighting is costly, each animal prefers as short a fight as possible. If animal i obtains the prey after a fight of length t, his utility will be $v_i - t$. We can model the situation as the game $G = \langle \{1, 2\}, (A_1, A_2), (u_1, u_2)\rangle$ where

- $A_1 = [0, \infty] = A_2$ (an element $t \in A_i$ represents a time at which player i plans to give up)
- $$u_1(t_1, t_2) = \begin{cases} -t_1 & \text{if } t_1 < t_2 \\ \frac{1}{2}v_1 - t_2 & \text{if } t_1 = t_2 \\ v_1 - t_2 & \text{if } t_1 > t_2 \end{cases}$$
- $$u_2(t_1, t_2) = \begin{cases} -t_2 & \text{if } t_2 < t_1 \\ \frac{1}{2}v_2 - t_1 & \text{if } t_1 = t_2 \\ v_2 - t_1 & \text{if } t_2 > t_1\,. \end{cases}$$

We are interested in the best response correspondences. First, we calculate player 1's best response correspondence, $\mathcal{B}_1(t_2)$. There are three cases to consider.

Case 1: $t_2 < v_1$ In this case, $v_1 - t_2 > \frac{1}{2}v_1 - t_2$ and $v_1 - t_2 > -t_1$. Consequently, given that player 2's action is t_2, player 1's utility function has a maximum value of $v_1 - t_2$, which is attained at any $t_1 > t_2$. Therefore, $\mathcal{B}_1(t_2) = (t_2, \infty)$.

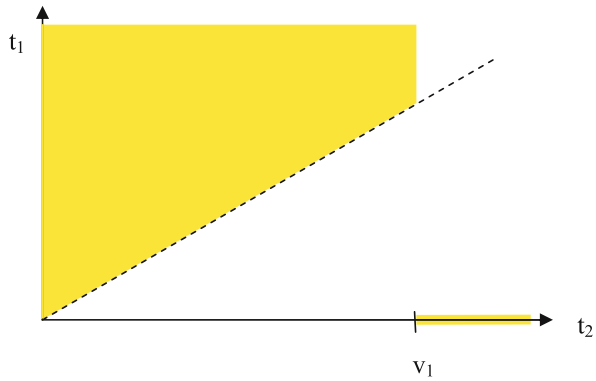

Static Games, Figure 1
Player 1's best response correspondence

Case 2: $t_2 = v_1$ In this case, $0 = v_1 - t_2 > \frac{1}{2}v_1 - t_2$. Therefore, player's 1 utility function $u_1(\cdot, t_2)$ has a maximum value of 0, which is attained at $t_1 = 0$ and at $t_1 > t_2$. Therefore, $\mathcal{B}_1(t_2) = \{0\} \cup (t_2, \infty)$.

Case 3: $t_2 > v_1$ In this case $\frac{1}{2}v_1 - t_2 < v_1 - t_2 < 0$. As a result, player 1's utility function $u_1(\cdot, t_2)$ has a maximum value of 0, which is attained at $t_1 = 0$. Therefore, $\mathcal{B}_1(t_2) = \{0\}$.

Summarizing, player 1's best response correspondence is:

$$\mathcal{B}_1(t_2) = \begin{cases} (t_2, \infty) & \text{if } t_2 < v_1 \\ \{0\} \cup (t_2, \infty) & \text{if } t_2 = v_1 \\ \{0\} & \text{if } t_2 > v_1 \end{cases}$$

which is depicted in Fig. 1.
Similarly, player 2's best response correspondence is:

$$\mathcal{B}_2(t_1) = \begin{cases} (t_1, \infty) & \text{if } t_1 < v_2 \\ \{0\} \cup (t_1, \infty) & \text{if } t_1 = v_2 \\ \{0\} & \text{if } t_1 > v_2\,. \end{cases}$$

Combining the two best response correspondences we get that (t_1^*, t_2^*) is a Nash equilibrium if and only if either $t_1^* = 0$ and $t_2^* \geq v_1$ or $t_2^* = 0$ and $t_1^* \geq v_2$. Figure 2 depicts the set of all the Nash equilibria as the intersection of the two best response correspondences. Two things are worth noting. First, it is not necessarily the case that the player who values the prey most wins the war. That is, there are Nash equilibria of the war of attrition where the player with the highest willingness to fight for the prey gives in first, and as a result the object goes to the other player. Second, in none of the Nash equilibria is there a physical fight. All Nash

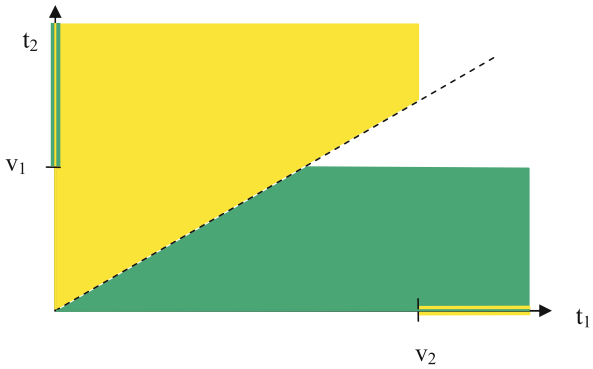

Static Games, Figure 2
The equilibria

equilibria involve one player giving in immediately to the other. This second feature seems rather unrealistic, since fights in "war of attrition"-like situations are commonly observed. If one wants to obtain a fight of positive length in the war of attrition one needs to either drop the Nash equilibrium concept and adopt an alternative one, or model the war of attrition differently. We will adopt this second course of action later.

Existence

As the matching pennies example shows, not all games have a Nash equilibrium. The following theorem, which dates back to Nash [18] and Glicksberg [11], states sufficient conditions on a game for it to have a Nash equilibrium. An earlier version of this theorem for the smaller but prominent class of zero-sum games can be found in von Neumann [23] (translated in von Neumann [24]). The standard proofs use Kakutani's fixed point theorem. We present here an alternative proof, due to Geanakoplos [10], which uses Brouwer's fixed point theorem instead.

Theorem 1 *The game $\langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$ has a Nash equilibrium if for all $i \in N$*

- *the set A_i of actions of player i is a nonempty compact convex subset of an Euclidean space,*
- *the utility function u_i is continuous,*
- *the utility function u_i is concave in A_i.*

Proof (Geanakoplos) Define the correspondence $\varphi_i : A \twoheadrightarrow A_i$ by

$$\varphi_i(\bar{a}) = \arg\max_{a_i \in A_i} \{U_i(a_i, \bar{a}_{-i}) - \|a_i - \bar{a}_i\|^2\},$$

where, $\|\cdot\|$ denotes a norm in the relevant Euclidean space. Note first that φ_i is a nonempty valued correspondence because the maximand is a continuous function and A_i is compact. Second, note that the function $\|a_i - \bar{a}_i\|$ is convex:

$$\begin{aligned}
&\|(\lambda a_i + (1-\lambda)b_i) - \bar{a}_i\| \\
&\quad = \|(\lambda a_i - \lambda \bar{a}_i) + ((1-\lambda)b_i - (1-\lambda)\bar{a}_i)\| \\
&\quad \le \|(\lambda a_i - \lambda \bar{a}_i)\| + \|((1-\lambda)b_i - (1-\lambda)\bar{a}_i)\| \\
&\quad \le |\lambda|\|a_i - \bar{a}_i\| + |1-\lambda|\|b_i - \bar{a}_i\| .
\end{aligned}$$

Since the quadratic function is strictly convex, then the maximand is a strictly concave function. Therefore, the correspondence φ_i is in fact a function. Furthermore, since the maximand is continuous in the parameter $\bar{a}$, φ_i is also continuous. To see this, let $\bar{a}_n \to \bar{a}$ be a convergent sequence of action profiles and let $a_{i\,n} = \varphi_i(\bar{a}_n)$. This means that $U(a_{i\,n}, (\bar{a}_n)_{-i}) \ge U(b_i, (\bar{a}_n)_{-i})$ for all $b_i \in A_i$. Since A_i is a compact set, $a_{i\,n}$ has a convergent subsequence. Denoting by a_i the limit of this subsequence and applying limits to the above inequality, we obtain that

$$U(a_i, \bar{a}_{-i}) \ge U(b_i, \bar{a}_{-a}) \quad \text{for all } b_i \in A_i,$$

namely $a_i = \varphi_i(\bar{a})$. Since this is true for every convergent subsequence of $a_{i\,n}$, we have that $\varphi_i(\bar{a}_n) = a_{i\,n} \to a_i = \varphi_i(\bar{a})$, which means that φ is continuous.

Now define $\varphi : A \to A$ by $\varphi = (\varphi_1, \ldots, \varphi_N)$. Clearly, φ is a continuous function mapping a compact set to itself. Therefore, by Brouwer's fixed point theorem, it has a fixed point: $\varphi(\bar{a}) = \bar{a}$. We now show that $\bar{a}$ is a Nash equilibrium of the game. Assume not. Then, there is some $i \in N$ with $a_i \in A_i$ such that $U_i(a_i, \bar{a}_{-i}) - U_i(\bar{a}) = E > 0$. Then, by concavity of U_i, for all $0 < \epsilon < 1$,

$$\begin{aligned}
&U_i(\epsilon a_i + (1-\epsilon)\bar{a}_i, \bar{a}_{-i}) - U_i(\bar{a}) \\
&\quad \ge \epsilon U_i(a_i, \bar{a}_{-i}) + (1-\epsilon)U_i(\bar{a}) - U_i(\bar{a}) \\
&\quad \ge \epsilon E > 0,
\end{aligned}$$

while $\|\epsilon a_i + (1-\epsilon)\bar{a}_i - \bar{a}_i\|^2 = \epsilon^2\|a_i - \bar{a}_i\|^2 < \epsilon E$, for small enough ϵ. Therefore, for such small ϵ, the action $\epsilon a_i + (1-\epsilon)\bar{a}_i$ satisfies

$$U_i(\epsilon a_i + (1-\epsilon)\bar{a}_i, \bar{a}_{-i}) - \|\epsilon a_i + (1-\epsilon)\bar{a}_i - \bar{a}_i\|^2 > U_i(\bar{a})$$

which contradicts the fact that $\varphi_i(\bar{a}) = \bar{a}_i$. □

Mixed Strategies

So far, we have formally defined a game, and have introduced the solution concept of Nash equilibrium which is arguably the central solution concept of game theory. However, there seem to be two problems with this concept. One is that although Nash equilibria exist in a wide class

of games, there are many simple games that do not have a Nash equilibrium. The most troubling example is Matching Pennies. If game theory cannot provide a prediction for this simple game then one must wonder if there is any value to the theory. The second problem is that the concept of Nash equilibrium predicts a very unrealistic outcome in the war of attrition. One would expect that game theory would not only provide nonempty predictions, but also ones that look reasonable and help explain what we see around us.

One way to approach these problems is not to abandon the theory or the concept of Nash equilibrium altogether, but to modify the way we model the problematic situations. The idea behind mixed strategies is to first modify the game by extending the set of actions available to the players, and then to apply the concept of Nash equilibrium to this extended game. In this way one may obtain additional Nash equilibria, some of which may provide reasonable predictions to the game.

Let $G = \langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$ be a finite game. For any A_i, a probability distribution on A_i is a function

$$x_i : A_i \to \mathbb{R}_+$$

such that

$$\sum_{a_i \in A_i} x_i(a_i) = 1 .$$

The set of all probability distributions on A_i is denoted by $\Delta(A_i)$. A *mixed strategy on* A_i is a random choice over elements of A_i, namely an element of $\Delta(A_i)$. If x_i is a mixed strategy on A_i, $x_i(a_i)$ denotes the probability that action $a_i \in A_i$ is selected when x_i is adopted. Since elements of $\Delta(A_i)$ can have an alternative interpretation, such as beliefs about the choice of player i, we denote the set of mixed strategies by X_i to distinguish it from the more abstract set of probability distributions on A_i. Also, we denote the set of mixed strategy profiles as $X = \times_{i\in N} X_i$. Denoting for each player $i \in N$, $X_{-i} = \times_{k\in N\setminus\{i\}} X_k$, a typical mixed strategy profile can be written as $(x_k)_{k\in N} = (x_i, x_{-i}) \in X_i \times X_{-i}$. The *mixed extension* of the strategic game G is the strategic game $\langle N, (X_i)_{i\in N}, (U_i)_{i\in N}\rangle$ where the set of actions of player i is the set of mixed strategies, X_i, and the payoff function $U_i : \times_{i\in N} X_i \to \mathbb{R}$ of player i is defined by

$$U_i((x_k)_{k\in N}) = \sum_{a=(a_k)_{k\in N}\in A} u_i(a) \Pi_{k\in N} x_k(a_k) .$$

Remark 1 Since each mixed strategy of player i, x_i, can be identified with a vector $x_i = (x_i(a_i))_{a_i\in A_i} \in \mathbb{R}^{|A_i|}$, the function U_i is multinomial in the coordinates of its variables, and, as a result, it is continuous as a function of the players' mixed strategies.

Definition 4 An *equilibrium in mixed strategies* of the game $\langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$ is a Nash equilibrium of the mixed extension of the game. In other words, it is a list of mixed strategies $\left(x_k^*\right)_{k\in N} \in X$ such that for all players $i \in N$ and for all his mixed strategies x_i,

$$U_i\left(\left(x_k^*\right)_{k\in N}\right) \geq U_i\left(\left(x_i, x_{-i}^*\right)\right) .$$

Alternatively, $\left(x_k^*\right)_{k\in N} \in X$ is a mixed strategy equilibrium if

$$x_i^* \in \mathcal{B}_i\left(x_{-i}^*\right) \quad \text{for all } i \in N .$$

Note that for every finite game $G = \langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$, its mixed extension is a strategic game that satisfies the conditions of Theorem 1. As a result, every finite game has a mixed strategy equilibrium.

Example 1 Consider again Matching Pennies. Its mixed extension is the game $\langle N, (X_i)_{i\in N}, (U_i)_{i\in N}\rangle$ where the set of players is $N = \{1, 2\}$, the sets of mixed strategies are $X_1 = \{(p_H, p_T) \geq (0, 0) : p_H + p_T = 1\}$, and $X_2 = \{(q_H, q_T) \geq (0, 0) : q_H + q_T = 1\}$, and the utility functions are given by $U_1((p_H, p_T), (q_H, q_T)) = p_H q_H + p_T q_T - p_H q_T - p_T q_H$ and $U_2((p_H, p_T), (q_H, q_T)) = p_H q_T + p_T q_H - p_H q_H - p_T q_T$. It can be checked that the only Nash equilibrium of this mixed extension is $((1/2, 1/2), (1/2, 1/2))$. Indeed, since $U_1((p_H, p_T), (1/2, 1/2))$ is identically 0, it attains its maximum at, among other strategies, $(1/2, 1/2)$. The same is true for $U_2((1/2, 1/2), (q_H, q_T))$. To see that there is no other equilibrium, note that for (q_H, q_T) with $q_H > q_T$, player 1's best response is $(1, 0)$. But player 2's best response to $(1, 0)$, is $(0, 1)$. Since $0 \leq 1$, (q_H, q_T) with $q_H > q_T$ cannot be part of an equilibrium. Similarly, for any (q_H, q_T) with $q_H < q_T$, player 1's best response is $(0, 1)$. But player 2's best response to $(0, 1)$ is $(1, 0)$. Since $1 \geq 0$, (q_H, q_T) with $q_H < q_T$ cannot be part of an equilibrium.

We next present a characterization of the mixed strategy equilibria of a game that will sometimes allow us to compute them in an easy way. Further, this characterization serves as the basis of an interesting interpretation of the mixed strategy equilibrium concept that we will discuss later. For this purpose, we identify the action $a_i \in A_i$ of player i with the mixed strategy of player i that assigns probability 1 to action a_i, and 0 to all other actions. There-

fore, given a player i, one of his actions $a_i \in A_i$, and a profile $x = (x_k)_{k\in N}$ of the players' mixed strategies, (a_i, x_{-i}) denotes the mixed strategy profile obtained from x by replacing i's mixed strategy x_i by the mixed strategy of player i that assigns probability 1 to action a_i. With this notation we can state the following identity:

$$U_i((x_k)_{k\in N}) = \sum_{a_i\in A_i} x_i(a_i)U_i((a_i, x_{-i}))\,. \tag{1}$$

Indeed,

$$\begin{aligned} U_i((x_k)_{k\in N}) &= \sum_{a=(a_k)_{k\in N}\in A} u_i(a)\Pi_{k\in N}x_k(a_k) \\ &= \sum_{a_i\in A_i}\sum_{a_{-i}\in A_{-i}} u_i(a)\Pi_{k\in N}x_k(a_k) \\ &= \sum_{a_i\in A_i} x_i(a_i)\sum_{a_{-i}\in A_{-i}} u_i(a)\Pi_{k\in N\setminus\{i\}}x_k(a_k) \\ &= \sum_{a_i\in A_i} x_i(a_i)U_i((a_i, x_{-i}))\,. \end{aligned}$$

Identity (1) is useful to prove the following characterization of the mixed strategy Nash equilibria.

Lemma 1 *The strategy profile $x^* = \left(x_k^*\right)_{k\in N}$ is an equilibrium of the mixed extension of $\langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$ if and only if for all players $i \in N$ and for all $a_i \in A_i$,*

$$\text{If } x_i^*(a_i) > 0 \quad \text{then } U_i(\left(a_i, x_{-i}^*\right)) = U_i(x^*) \tag{2}$$

$$\text{If } x_i^*(a_i) = 0 \quad \text{then } U_i(\left(a_i, x_{-i}^*\right)) \le U_i(x^*)\,. \tag{3}$$

Proof Assume that $x^* = \left(x_k^*\right)_{k\in N}$ satisfies conditions (2) and (3). Let $i \in N$, and let x_i be a mixed strategy of player i. Then, by (1)

$$\begin{aligned} U_i(x_i, x_{-i}^*) &= \sum_{a_i\in A_i} x_i(a_i)U_i(\left(a_i, x_{-i}^*\right)) \\ &\le \sum_{a_i\in A_i} x_i(a_i)U_i(x^*) = U_i(x^*) \end{aligned}$$

and therefore x^* is an equilibrium.

Assume now that $x^* = \left(x_k^*\right)_{k\in N}$ is an equilibrium. Let $i \in N$. Then

$$U_i(x^*) \ge U_i(\left(a_i, x_{-i}^*\right)) \quad \forall a_i \in A_i \tag{4}$$

and, in particular, condition (3) holds for all $a_i \in A_i$ such that $x_i(a_i) = 0$. Also, using (1) we can write

$$\sum_{a_i\in A_i} x_i^*(a_i)U_i(x^*) = \sum_{a_i\in A_i} x_i^*(a_i)U_i\left(a_i, x_{-i}^*\right)\,. \tag{5}$$

If there is $a_i \in A_i$ such that $x_i^*(a_i) > 0$ and $U_i(x^*) > U_i\left(a_i, x_{-i}^*\right)$ then, using (4),

$$\sum_{a_i\in A_i} x_i^*(a_i)U_i(x^*) > \sum_{a_i\in A_i} x_i^*(a_i)U_i(\left(a_i, x_{-i}^*\right))$$

in contradiction to (5). □

Corollary 1 *The strategy profile $x^* = \left(x_k^*\right)_{k\in N}$ is an equilibrium of the mixed extension of $\langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$ if and only if for all players $i \in N$ and for all $a_i \in A_i$,*

$$x_i^*(a_i) > 0 \quad \textit{implies } a_i \in \mathcal{B}_i(x_{-i}^*)\,.$$

According to the standard interpretation, a player's mixed strategy in a game G is an action, but in a different game, namely in the mixed extension of G. According to this interpretation, a mixed strategy is a deliberate choice of a player to use a random device. A mixed strategy equilibrium then is a profile of *independent* random devices, each of which is a best response to the others. Corollary 1 provides an alternative interpretation of a mixed strategy equilibrium. According to this interpretation, a player's mixed strategy represents the uncertainty in the minds of the other players concerning the player's action. In other words, a player's mixed strategy is interpreted not as a deliberate choice of the player but the belief, shared by all the other players, about the player's choice. That is, if $(x_k)_{k\in N}$ is a profile of mixed strategies, then x_i is the conjecture, shared by all the players other than i, about i 's ultimate choice of action. Consequently, x_{-i} are the conjectures entertained by player i about his opponents' actions. According to this interpretation, Corollary 1 says that a mixed strategy equilibrium $\left(x_k^*\right)_{k\in N}$ is a profile of beliefs about each player's actions (entertained by the other players) according to which each player chooses an action that is a best response to his own beliefs.

The War of Attrition (cont.)

We have seen in Sect. "Analysis of Some Finite Games" that all the Nash equilibria of the war of attrition predict no real fight for the prey. We will now see that there is a mixed strategy equilibrium of the war of attrition that predicts a positive-length fight with probability one.

The players' action sets in the war of attrition are intervals of real numbers. A mixed strategy for player i in that game can be represented by a cumulative distribution function $F_i\colon [0,\infty] \to [0,1]$. For each $t \in (0,\infty]$, $F_i(t)$ is the probability that player i gives up at or before t. We will look for a Nash equilibrium (F_1, F_2) that consists of two strictly increasing, differentiable cumulative distribution

functions. The density of F_i is denoted by f_i. We will try to find an equilibrium at which each player is indifferent between all pure actions.

Consider player i. Given that his opponent is using mixed strategy F_j, $j \neq i$, if he chooses to give in at time t, then he will face a lottery according to which,

- With probability $1 - F_j(t)$, player i does not obtain the prey and gets a payoff of $-t$,
- With probability $F_j(t)$, player i obtains the prey at time t_j, where t_j is a random variable whose cumulative distribution function is $F_j(t_j)/F_j(t)$ (the distribution player j's surrender time, conditional on his having given in before t).

Therefore, the corresponding expected utility of choosing time t is

$$\begin{aligned} U_i(t, F_j) &= (1 - F_j(t))(-t) + F_j(t) \int_0^t (v_i - t_j) \mathrm{d} \frac{F_j(t_j)}{F_j(t)} \\ &= (1 - F_j(t))(-t) + \int_0^t (v_i - t_j) \mathrm{d}F_j(t_j) \,. \end{aligned}$$

Since in the equilibrium we are looking for, player i is indifferent among all his actions, the above expression is independent of t. Namely, $U_i(t, F_j) \equiv c$. As a result, the derivative of the above utility with respect to t equals 0. Formally,

$$\begin{aligned} \frac{\partial U_i(t, F_j)}{\partial t} &= t f_j(t) - (1 - F_j(t)) + (v_i - t) f_j(t) \\ &= (1 - F_j(t)) + v_i f_j(t) = 0. \end{aligned}$$

This is a differential equation whose general solution is

$$F_j(t) = 1 - K \mathrm{e}^{-\frac{t}{v_i}} \,.$$

If we want it to satisfy $F_j(0) = 0$, we obtain that $K = 1$. As a result, the distribution function is given by

$$F_j(t) = 1 - \mathrm{e}^{-\frac{t}{v_i}} \,.$$

Consequently, the equilibrium we are looking for is

$$(F_1(t), F_2(t)) = \left(1 - \mathrm{e}^{-\frac{t}{v_2}}, 1 - \mathrm{e}^{-\frac{t}{v_1}}\right) \,.$$

According to this equilibrium, for any t, the probability that there is a fight that lasts at least t is $(1 - F_1(t))(1 - F_2(t)) > 0$. Consequently, there is a fight with probability one. The introduction of mixed strategies allowed the concept of Nash equilibrium to be consistent with fights that last a positive length of time. However, the mixed strategy equilibrium has the following unfortunate property. If $v_1 < v_2$, then for all $t > 0$, $F_1(t) < F_2(t)$. In other words, it is more likely that the player with the highest willingness to fight for the prey gives up earlier than any given t, than that the player with the lowest willingness to fight gives in earlier than the same t. Therefore, in equilibrium it is more likely that the player with the lower willingness to fight wins the war than the other way around. In particular, the probability that player 1 gets the object is given by

$$\int_0^\infty F_2(t) \mathrm{d}F_1(t)$$

which can be checked to be equal to $\frac{v_2}{v_1 + v_2} > 1/2$. In order to obtain the more intuitive result that the higher the willingness to fight for the prey, the higher is the probability to obtain it, we will need to model the war of attrition in yet a different way. We'll return to this when we introduce asymmetric information to the games.

Equilibrium in Beliefs

The mixed extension of the game $\langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$ is constructed in two steps. First, we enlarge the set of actions available to each player by allowing him to choose any mixed strategy on his original action set. Second, since the action choices are now probability distributions over actions, we extend the players' original preferences to preferences over profiles of mixed strategies. We do so by evaluating each mixed strategy profile according to the expected value of the original utilities with respect to the probability distribution over action profiles induced by the mixed strategy.

The first step seems uncontroversial since it is certainly possible for players to use random devices. But the second step is somewhat problematic because, by evaluating mixed strategies according to the expected utility of the resulting lotteries, one is implicitly imposing on the players a certain kind of risk preferences. One may wonder what the implications would be if instead of extending the preferences by assuming that players are expected utility maximizers, we assume that players have more general preferences over profiles of mixed strategies. In particular, we would like to know if there is a suitable generalization of Corollary 1.

Let $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$ be a finite game. We define the mixed extension of G as the strategic game $\langle N, (X_i)_{i \in N}, (U_i)_{i \in N} \rangle$ where, as in Sect. "Mixed Strategies", X_i is the set of probability distributions over the actions in A_i, for $i \in N$, but unlike there, the utility function $U_i : X \to \mathbb{R}^N$ is not necessarily a multilinear function of the probabilities, but a general continuous function of the

mixed strategies. The only requirement on U_i is that for all profiles of degenerate mixed strategies $(a_k)_{k\in N}$, we have $U_i\left((a_k)_{k\in N}\right) = u_i\left((a_k)_{k\in N}\right)$. As before, a mixed strategy Nash equilibrium of $\langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$ is a Nash equilibrium of its mixed extension $\langle N, (X_i)_{i\in N}, (U_i)_{i\in N}\rangle$. In other words, it is a list of mixed strategies $\left(x_k^*\right)_{k\in N}$ such that for all players $i \in N$ and for all of his mixed strategies x_i,

$$U_i(\left(x_k^*\right)_{k\in N}) \geq U_i(\left(x_i, x_{-i}^*\right)).$$

Alternatively, $\left(x_k^*\right)_{k\in N}$ is a mixed strategy equilibrium if

$$x_i^* \in \mathcal{B}_i(x_{-i}^*) \quad \text{for all } i \in N.$$

Observation 1 It is important to note that two different actions of a player may be best responses to a given mixed strategy profile of the other players, and yet no probability mixture of the two actions will be a best response to the given mixed strategy profile. This will typically be the case when the function U_i is strictly convex in X_i, since strictly convex functions attain their maximum at boundary points.

Theorem 1 shows that Nash equilibria exist when the extended utility function U_i is concave in X_i. However, Observation 1 indicates that a Nash equilibrium may fail to exist when U_i is strictly convex in X_i. Indeed, take a game $G = \langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$ with no pure strategy Nash equilibrium, like Matching Pennies, and consider its mixed extension $\Gamma = \langle N, (X_i)_{i\in N}, (U_i)_{i\in N}\rangle$ where for all players, their extended utility function is strictly convex. Then, for any player $i \in N$ and for any profile of mixed strategies x_{-i} of the other players, the set of i's best responses $\mathcal{B}_i(x_{-i})$ consists of only degenerate mixed strategies. Since G has no pure strategy Nash equilibrium, we conclude that Γ does not have a Nash equilibrium.

Observation 2 It is also important to note that, unlike in the standard expected utility case, a player's mixed strategy x_i^* may very well be a best response to some profile x_{-i}^* of the other players' mixed strategies and at the same time may assign positive probability to an action that (when regarded as a degenerate mixed strategy) is not a best response to x_{-i}^*. Formally, it may very well be the case that

$$U_i(\left(x_k^*\right)_{k\in N}) \geq U_i(\left(x_i, x_{-i}^*\right)) \quad \text{for all } x_i \in X_i$$

and yet

$$U_i(\left(a_i, x_{-i}^*\right)) < U_i(\left(x_k^*\right)_{k\in N}) \quad \text{for some } a_i \text{ such} \quad x_i^*(a_i) > 0.$$

This will typically occur when the function U_i is strictly concave in X_i.

The definition of mixed strategy equilibrium requires from each strategy in the equilibrium profile that it be a best response to the other strategies. Corollary 1 stated that when preferences have the expected utility form, each mixed strategy in a mixed strategy equilibrium is also a probability mixture over best responses to the other strategies in the profile. This result allowed us to interpret a mixed strategy Nash equilibrium as a profile of beliefs, rather than as a profile of probability mixtures. As explained in Observation 2, however, when preferences over mixed strategies are not expected utility preferences, a mixture over best responses is not necessarily a best response. Therefore, Corollary 1 does not extend to the mixed extension where preferences are not of the expected utility form.

In this setup, however, one can still interpret a player's mixed strategy as a belief entertained by the other players about the actions chosen by that player. And a profile of such beliefs will be in equilibrium if the probability distribution over the player's actions that represents i's beliefs is obtained as a mixture of best responses of this player to his beliefs about the other players' actions. With this idea in mind, Crawford [8] defined the notion of an equilibrium in beliefs. Before we formally present his definition we need to introduce some notation.

Since when the extended utility functions U_i are concave in i's own strategy a best response to a given profile of the other players' strategies may be a non-degenerate mixed strategy, a mixture of best responses will typically be a mixture over non-degenerate mixed strategies. This mixture induces a probability distribution over actions in a natural way by reducing the compound mixture to a simple mixture. This induced probability distribution can be interpreted as a belief over the actions ultimately chosen. For example, in Matching Pennies, if player 1 believes that there is a probability of 1/2 that player 2 will choose the mixed strategy $(1/3, 2/3)$ and a probability of 1/2 that player 2 will choose the mixed strategy $(2/3, 1/3)$, then player 1 believes that player 2 will choose each one of his two actions with equal probability. More generally, if player i assigns probability p_k to the event that player j will choose mixed strategy $x^k \in X_j$, for $k = 1, \ldots, K$, then player i's beliefs about player j's actions are given by $\sum_{k=1}^{K} p_k x^k \in X_j$. That is, for each action $a_j \in A_j$ of player j, player i believes that player j will choose a_j with probability $\sum_{k=1}^{K} p_k x^k(a_j)$. For each set $T \subset X_i$ of mixed strategies, let $D[T] \subset X_i$ denote the set of probability distributions over i's actions that are in-

duced by mixtures over elements of T. With this notation in hand, we can define the concept of equilibrium in beliefs.

Definition 5 Let $G = \langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$ be a game. For each $i \in N$, let $\mathcal{B}_i : X \to X_i$ be the best response correspondence in the mixed extension of G. The profile of beliefs $(x_k^*)_{k\in N} \in \times_{k\in N}\Delta(A_k)$ is an *equilibrium in beliefs* if

$$x_i^* \in D\left[\mathcal{B}_i(x_{-i}^*)\right] \quad \text{for all } i \in N\,.$$

An equilibrium in beliefs is a profile of beliefs $\left(x_k^*\right)_{k\in N}$. For each $i \in N$, x_i^* is the common belief of the players other than i about player i's choice of actions. In order for this profile of beliefs to be in equilibrium, we require that for each player $i \in N$ all the other players believe that i chooses a mixed strategy that is a best response to his beliefs, which are given by $\left(x_k^*\right)_{k\in N\setminus\{i\}}$, about the other players' choices of actions. In other words, x_i^* must be a convex combination of best responses of i to $\left(x_k^*\right)_{k\in N\setminus\{i\}}$.

Example 2 Consider again the mixed extension of Matching Pennies $\langle N, (X_i)_{i\in N}, (U_i)_{i\in N}\rangle$ where the set of players is $N = \{1, 2\}$, the sets of mixed strategies are $X_1 = \{(p_H, p_T) \geq (0, 0) : p_H + p_T = 1\}$ and $X_2 = \{(q_H, q_T) \geq (0, 0) : q_H + q_T = 1\}$, and the utility functions are now given by $U_1((p_H, p_T), (q_H, q_T)) = (p_H q_H)^2 + (p_T q_T)^2 - p_H q_T - p_T q_H$ and $U_2((p_H, p_T), (q_H, q_T)) = (p_H q_T)^2 + (p_T q_H)^2 - p_H q_H - p_T q_T$. Since the utility functions are strictly convex in the players's own mixed strategies, the best response to any strategy of the opponent is a pure strategy. In particular, one can verify that

$$\mathcal{B}_1(q_H, q_T) = \begin{cases} (1, 0) & \text{if } q_H > q_T \\ \{(1, 0), (0, 1)\} & \text{if } q_H = q_T \\ (0, 1) & \text{if } q_H < q_T \end{cases}$$

and

$$\mathcal{B}_2(p_H, p_T) = \begin{cases} (0, 1) & \text{if } p_H > p_T \\ \{(1, 0), (0, 1)\} & \text{if } p_H = p_T \\ (1, 0) & \text{if } p_H < p_T. \end{cases}$$

It can also be verified that $((p_H^*, p_T^*), (q_H^*, q_T^*)) = ((1/2, 1/2), (1/2, 1/2))$ is an equilibrium in beliefs. Indeed, for both $i = 1, 2$, $(1/2, 1/2) \in X_i$ is a convex combination of $(1, 0)$ and $(0, 1)$, which are both in $\mathcal{B}_j(1/2, 1/2)$, $j \neq i$. In this equilibrium,

1. Player 1 believes that player 2 will choose $(1, 0)$ with probability 1/2, and $(0, 1)$ with probability 1/2.
2. Therefore player 1 believes that player 2 will ultimately choose H and T, each with probability 1/2.
3. Given these beliefs, player 1's only best replies are $(1, 0)$ and $(0, 1)$, and
4. Player 2 believes that player 1 will choose each one with probability 1/2. As a result,
5. Player 2 believes that player 1 will ultimately choose H and T each with probability 1/2.
6. Given these beliefs, player 2's only best replies are $(1, 0)$ and $(0, 1)$, and
1. Player 1 believes that player 2 will choose $(1, 0)$ with probability 1/2, and $(0, 1)$ with probability 1/2.

The following result is a direct implication of the definition of an equilibrium in beliefs.

Proposition 2 (Crawford [8]) *Let $G = \langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$ be a strategic game, and $\Gamma = \langle N, (X_i)_{i\in N}, (U_i)_{i\in N}\rangle$ be the mixed extension of G, where U_i is continuous but not necessarily multilinear.*

1. *Every mixed strategy Nash equilibrium of G is an equilibrium in beliefs.*
2. *If for all $i \in N$, U_i is quasiconcave in X_i, then every equilibrium in beliefs is a mixed strategy Nash equilibrium of G.*

Proof

1. Since $\mathcal{B}_i(x_{-i}^*) \subset D\left[\mathcal{B}_i(x_{-i}^*)\right]$ for all $i \in N$, every Nash equilibrium is an equilibrium in beliefs.
2. When the utility function U_i is quasiconcave in i's mixed strategy, the set of best responses $\mathcal{B}_i(x_{-i}^*)$ is a convex set. Therefore, $D\left[\mathcal{B}_i(x_{-i}^*)\right] = \mathcal{B}_i(x_{-i}^*)$, and any equilibrium in beliefs is a Nash equilibrium. □

Crawford [8] shows that although some games have no Nash equilibrium, every game has an equilibrium in beliefs.

Correlated Equilibrium

In the mixed extension of a game, players do not choose their actions directly, but rather choose probability distributions over their action sets according to which the actions are ultimately selected. The important feature about these probability distributions is that they represent independent random variables. The realization of one player's random variable does not give any information about the realization of the other players' random variables. There is nothing in the bare notion of equilibrium, however, that requires players' behavior to be independent. The basic

feature of an equilibrium is that each player is best responding to the behavior of others, and that each player is free to choose any action in his action set. But one thing is that players can, if they so wish, change their behavior without the consent of others, and another different thing is to expect players' choices to be independent. Therefore, one could ask what would happen if the random devices players use to ultimately choose their actions were correlated. In that case, knowledge of the realization of one's random device would provide some partial information about the realization of the other players' random devices, and therefore of their choices. In equilibrium, a player should take this information into account. To illustrate this point, consider the game of Chicken.

		Driver 2	
		Slow Down	Speed up
Driver 1	Slow Down	6, 6	2, 7
	Speed up	7, 2	0, 0

This game has two pure-action Nash equilibria, and one equilibrium in mixed strategies. According to the mixed strategy Nash equilibrium, each player chooses Slow Down with probability 2/3 and Speed Up with probability 1/3. This mixed strategy equilibrium can be implemented by the following random device. Consider two random variables S_1 and S_2, whose joint distribution is given by the following table:

Driver 1 chooses his action as a function of the realization of S_1 and Driver 2 chooses his action as a function of the realization of S_2. (Neither player is informed of the realization of the other player's random variable.) In particular, Driver 1 chooses Slow Down if $S_1 = 1$ and Speed Up otherwise. Similarly, Driver 2 chooses Slow Down if $S_2 = 1$, and Speed Up otherwise. Note that according to this pattern of behavior, each player chooses to slow down with probability 2/3. But more importantly, since S_1 and S_2 are independent random variables, knowledge of the realization of one random variable does not give any information about the realization of the other one. Therefore, after Driver 1 learns the realization of S_1, he still believes that Driver 2 will choose Slow Down with probability 2/3 and consequently any choice is optimal, in particular the one described above. Similarly, after Driver 2 learns the realization of S_2, he still believes that Driver 1 will choose to slow down with probability 2/3, and his planned behavior continues to be optimal.

Static Games, Table 1
A random device

		S_2	
		1	2
S_1	1	4/9	2/9
	2	2/9	1/9

But what would happen if the joint distribution of S_1 and S_2, was not as presented in Table 1, but rather as follows?

		S_2	
		1	2
S_1	1	1/3	1/3
	2	1/3	0

To answer this question, assume that both players still choose their actions according to the previous pattern of behavior: Driver 1 chooses Slow Down if $S_1 = 1$, and Speed Up otherwise. The same holds for Driver 2. As a result, it is still true that each player chooses Slow Down with probability 2/3 and Speed Up with probability 1/3. However, since this time the conditioning random variables S_1 and S_2 are not independent, knowledge of the realization of S_1 affects the beliefs of Driver 1 about the probability with which Driver 2 chooses his actions. In particular, if $S_1 = 1$, Driver 1 updates his beliefs and assigns probability 1/2 to Driver 2 choosing either action, and consequently, Driver 1's only optimal action is Slow Down, which is precisely the choice dictated by the above pattern of behavior. Similarly, if $S_1 = 2$, Driver 1 should update his beliefs and assign probability one that Driver 2 will choose Slow Down. Consequently, Driver 1's best reply is to follow the above pattern of behavior and choose Speed Up. One can see that, given that the players know that the random variables S_1 and S_2 are correlated and they use this information accordingly, there is no incentive for either of them to deviate from the proposed pattern of behavior. Therefore, we can say that this pattern of behavior is an equilibrium. This notion of a correlated equilibrium was introduced in Aumann [2]. Before we give a formal definition we introduce the concept of a correlated strategy profile, which will play a central role not only in this section, but in the next one as well.

Definition 6 Let $G = \langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$ be a game. A *correlated strategy profile in* G consists of

- A finite probability space (Ω, π)
- For each player $i \in N$, a partition $\mathcal{P}_i$ of Ω into events of positive probability
- For each player $i \in N$, a function $\sigma_i : \Omega \to A_i$ which is measurable with respect to $\mathcal{P}_i$.

A correlated strategy profile is a description of what players do and know while playing the game G. The collec-

tion $\langle(\Omega,\pi),(\mathcal{P}_i)_{i\in N}\rangle$ represents the random devices used by the players to ultimately choose their actions. The underlying probability space that governs the players' random devices is (Ω,π). Ω is the set of states, and for each state ω, $\pi(\omega)$ is the probability that ω occurs. For each $i\in N$, the partition $\mathcal{P}_i$ represents player i's information. Each element of the partition represents a different realization of the random device used by i to choose his action. Two states that belong to the same element of the partition $\mathcal{P}_i$ cannot be distinguished by i, while two states that belong to different partition cells can be distinguished by him. For each player i, $\sigma_i\colon\Omega\to A_i$ is the random variable that describes players i's choice of action, $\sigma_i(\omega)$ being the action chosen by him at state ω. The measurability of σ_i with respect to $\mathcal{P}_i$ formalizes the requirement that the actions chosen by player i depend only on his information about the state of the world. Therefore, for any two states that belong to the same element of his partition, the actions chosen by i at those states must be the same. That is, for any $\omega,\omega'\in P\in\mathcal{P}_i$ we have $\sigma_i(\omega)=\sigma_i(\omega')$.

For example, the correlated strategy profile described earlier for the game of chicken can be formalized as $\langle(\Omega,\pi),(\mathcal{P}_i)_{i\in N},(\sigma_i)_{i\in N}\rangle$, where $N=\{\mathrm{I},\mathrm{II}\}$, and

- $\Omega=\{(1,1),(1,2),(2,1))\}$
- $\pi(\omega)=1/3$ for all $\omega\in\Omega$
- $\mathcal{P}_{\mathrm{I}}=\{\{(1,1),(1,2)\},\{(2,1)\}\}$ and $\mathcal{P}_{\mathrm{II}}=\{\{(1,1),(2,1)\}\},\{(1,2)\}$
- $\sigma_{\mathrm{I}}(\omega)=\begin{cases}\text{Slow Down} & \text{if } \omega\in\{(1,1),(1,2)\}\\ \text{Speed up} & \text{if } \omega\in\{(2,1)\}\end{cases}$
- $\sigma_{\mathrm{II}}(\omega)=\begin{cases}\text{Slow Down} & \text{if } \omega\in\{(1,1),(2,1)\}\\ \text{Speed up} & \text{if } \omega\in\{(1,2)\}\end{cases}.$

According to this correlated strategy profile, there are three equally likely states, and the players can distinguish only one component of the state, namely the realization of their random variable. The players' actions are described by the functions σ_I and σ_{II} which depend only on the respective player's information.

In what follows we denote by $\sigma\colon\Omega\to A$ the function that associates with each $\omega\in\Omega$ the action profile induced by the strategies σ_k, for $k\in N$. That is, $\sigma=(\sigma_k)_{k\in N}$. Also, for any $i\in N$, $\sigma_{-i}=(\sigma_k)_{k\in N\setminus\{i\}}$ so that $\sigma=(\sigma_{-i},\sigma_i)$. We are interested in correlated strategy profiles in which no player benefits by altering his behavior. These special profiles are introduced in the following definition.

Definition 7 Let $G=\langle N,(A_i)_{i\in N},(u_i)_{i\in N}\rangle$ be a strategic game. A *correlated equilibrium of* G is a correlated strategy $\langle(\Omega,\pi),(\mathcal{P}_i)_{i\in N},(\sigma_i)_{i\in N}\rangle$ such that for every $i\in N$ and every function $\tau_i\colon\Omega\to A_i$ that is measurable with respect to $\mathcal{P}_i$,

$$\sum_{\omega\in\Omega}\pi(\omega)u_i(\sigma_{-i}(\omega),\sigma_i(\omega))\geq\sum_{\omega\in\Omega}\pi(\omega)u_i(\sigma_{-i}(\omega),\tau_i(\omega))\,. \quad (6)$$

The value $v_i=\sum_{\omega\in\Omega}\pi(\omega)u_i(\sigma_{-i}(\omega),\sigma_i(\omega))$ is player i's correlated equilibrium payoff.

In a correlated strategy profile each player plans to condition his choice of action on the realization of a random variable, and the players' random variables may be correlated. A correlated strategy profile is a correlated equilibrium if no player can find an alternative way to condition his choice on the same random device, so that his expected utility is increased. Note that the player presumably chooses his strategy (his way to condition his actions on the outcomes of the random device) before he learns the realization of the device. Nonetheless, he evaluates the outcomes generated by the players' strategies by taking into account the precise correlation of the random devices on which outcomes players are conditioning their behavior.

Although strictly speaking mixed strategy Nash equilibria are not correlated equilibria, they do induce a correlated equilibrium distribution over action profiles. In order to state this claim, we need the following definition.

Definition 8 Let $\langle(\Omega,\pi),(\mathcal{P}_i,\sigma_i)_{i\in N}\rangle$ be a correlated strategy profile for G. Its induced probability distribution over action profiles is given by the function $p\colon A\to[0,1]$ defined by

$$\begin{aligned}p(a)&=\pi(\{\omega\in\Omega\colon\sigma(\omega)=a\})\\&=\sum_{\{\omega\in\Omega\colon\sigma(\omega)=a\}}\pi(\omega)\quad\text{for all } a\in A\,.\end{aligned}$$

Proposition 3 *Let $G=\langle N,(A_i)_{i\in N},(u_i)_{i\in N}\rangle$ be a strategic game, and let $x=(x_1,\cdots,x_n)$ be a mixed strategy Nash equilibrium of G. Then, there is a correlated equilibrium $\langle(\Omega,\pi),(\mathcal{P}_i)_{i\in N},(\sigma_i)_{i\in N}\rangle$ whose induced probability distribution over action profiles is the same as x's distribution.*

Proof Let $\langle(\Omega,\pi),(\mathcal{P}_i)_{i\in N},(\sigma_i)_{i\in N}\rangle$ be defined as follows:

- $\Omega=A$
- $\pi(a)=\prod_{i\in N}x_i(a_i)$
- $\mathcal{P}_i(a)=\{b\in A\colon b_i=a_i\}$
- $\sigma_i(a)=a_i$.

We claim that $\langle(\Omega,\pi),(\mathcal{P}_i)_{i\in N},(\sigma_i)_{i\in N}\rangle$ is a correlated equilibrium whose probability distribution is the same as x's distribution. Let $i \in N$. Since x is a mixed strategy Nash equilibrium, we know by Lemma 1 that for all $a_i \in A_i$

$$\begin{aligned}&\text{if } x_i(a_i) > 0 \quad \text{then } U_i(x) = U_i(a_i, x_{-i})\\ &\text{if } x_i(a_i) = 0 \quad \text{then } U_i(x) \geq U_i(a_i, x_{-i})\,.\end{aligned}$$

Consequently, for all $a_i \in A_i$

$$x_i(a_i)U_i(a_i, x_{-i}) \geq x_i(a_i)U_i(b_i, x_{-i}) \quad \text{for all } b_i \in A_i\,. \tag{7}$$

Now let $\tau_i : A \to A_i$ be a function that is measurable with respect to $\mathcal{P}_i$. Let $a_{-i} \in A_{-i}$ be a fixed profile of actions for players other than i. Letting $b_i = \tau_i(a_i, a_{-i})$, Eq. (7) implies that

$$x_i(a_i)U_i(a_i, x_{-i}) \geq x_i(a_i)U_i(\tau_i(a_i, a_{-i}), x_{-i}) \quad \text{for all } a_i \in A_i\,.$$

Adding over all $a_i \in A_i$,

$$\sum_{a_i\in A_i} x_i(a_i)U_i(a_i, x_{-i}) \geq \sum_{a_i\in A_i} x_i(a_i)U_i(\tau_i(a_i, a_{-i}), x_{-i})\,.$$

Taking into account the definition of $U_i(a_i, x_{-i})$ and $U_i(\tau_i(a), x_{-i})$, and using the measurability of τ_i with respect to $\mathcal{P}_i$, we get

$$\begin{aligned}&\sum_{a_i\in A_i} x_i(a_i) \sum_{a_{-i}\in A_{-i}} \left(\prod_{j\in N\setminus\{i\}} x_j(a_j)\right) u_i(a_i, a_{-i})\\ &\quad\geq \sum_{a_i\in A_i} x_i(a_i) \sum_{a_{-i}\in A_{-i}} \left(\prod_{j\in N\setminus\{i\}} x_j(a_j)\right) u_i(\tau_i(a), a_{-i})\\ &\sum_{a\in A} \left(\prod_{j\in N} x_j(a_j)\right) u_i(a_i, a_{-i})\\ &\quad\geq \sum_{a\in A} \left(\prod_{j\in N} x_j(a_j)\right) u_i(\tau_i(a), a_{-i})\\ &\sum_{a\in A} \pi(a)u_i(a_i, a_{-i}) \geq \sum_{a\in A} \pi(a)u_i(\tau_i(a), a_{-i})\\ &\sum_{a\in A} \pi(a)u_i(\sigma_i(a), \sigma_{-i}(a)) \geq \sum_{a\in A} \pi(a)u_i(\tau_{-i}(a), \sigma_{-i}(a)).\end{aligned}$$

This shows that $\langle(\Omega,\pi),(\mathcal{P}_i)_{i\in N},(\sigma_i)_{i\in N}\rangle$ is a correlated equilibrium of G. Its induced probability distribution over action profiles is

$$\begin{aligned}p(a) &= \pi(\{b \in A\colon \sigma(b) = a\})\\ &= \pi(\{b \in A\colon b = a\})\\ &= \pi(a)\\ &= \prod_{i\in N} x_i(a_i)\,.\end{aligned}$$

□

Although a correlated strategy profile consists of a randomizing device used by the players, it turns out that the only feature of the device that determines whether or not the correlated strategy profile constitutes a correlated equilibrium is its induced probability distribution over the action profiles. This is shown by the next proposition.

Proposition 4 *Let $G = \langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$ be a finite strategic game. Every correlated equilibrium probability distribution over action profiles can be obtained in a correlated equilibrium of G in which*

- $\Omega = A$
- $\mathcal{P}_i(a) = \{b \in A\colon b_i = a_i\}$.

Proof Let $\left\langle(\Omega',\pi'),(\mathcal{P}'_i,\sigma'_i)_{i\in N}\right\rangle$ be a correlated equilibrium of G. Consider the correlated strategy profile $\langle(\Omega,\pi),(\mathcal{P}_i,\sigma_i)_{i\in N}\rangle$ defined by

- $\Omega = A$
- $\pi(a) = \pi'(\{\omega \in \Omega\colon \sigma'(\omega) = a\})$ for each $a \in A$
- $\mathcal{P}_i(a) = \{b \in A\colon b_i = a_i\}$ for each $i \in N$ and for each $a \in A$
- $\sigma_i(a) = a_i$ for each $i \in N$.

It is clear that this correlated strategy profile induces the required distribution over action profiles. Indeed,

$$\begin{aligned}p(a) &= \pi(\{\omega \in \Omega\colon \sigma(\omega) = a\})\\ &= \pi(\{a' \in A\colon a' = a\})\\ &= \pi(a)\\ &= \pi'(\{\omega \in \Omega'\colon \sigma'(\omega) = a\})\,.\end{aligned}$$

It remains to show that this profile is a correlated equilibrium. Take a function $\tau_i\colon A \to A_i$ that is measurable with respect to $\mathcal{P}_i$. Define $\tau'_i\colon \Omega' \to A_i$ by $\tau'_i(\omega) = \tau_i(\sigma'(\omega)) = \tau_i(\sigma'_{-i}(\omega), \sigma'_i(\omega))$. The function τ'_i is measurable with respect to $\mathcal{P}'_i$. Indeed, if $\omega' \in \mathcal{P}'_i(\omega)$, then $\sigma'_i(\omega') = \sigma'_i(\omega)$ by measurability of σ'_i with respect to $\mathcal{P}'_i$. Therefore, by definition of $\mathcal{P}_i$, $\mathcal{P}_i(\sigma'(\omega')) = \mathcal{P}_i(\sigma'(\omega))$, and both $\sigma'_i(\omega')$ and $\sigma'_i(\omega)$ belong to the same element of

$\mathcal{P}_i$. Since τ_i is measurable with respect to $\mathcal{P}_i$, we conclude that $\tau'_i(\omega') = \tau_i(\sigma'_i(\omega')) = \tau_i(\sigma'_i(\omega)) = \tau'_i(\omega)$.

Also,

$$\begin{aligned}
&\sum_{\omega\in\Omega} \pi(\omega)u_i(\sigma_{-i}(\omega), \tau_i(\omega)) \\
&\quad = \sum_{a\in A} \pi(a)u_i(a_{-i}, \tau_i(a)) \\
&\quad = \sum_{a\in A} \sum_{\{\omega\in\Omega' : \sigma'(\omega)=a\}} \pi'(\omega)u_i(\sigma'_{-i}(\omega), \tau_i(\sigma'(\omega))) \\
&\quad = \sum_{a\in A} \sum_{\{\omega\in\Omega' : \sigma'(\omega)=a\}} \pi'(\omega)u_i(\sigma'_{-i}(\omega), \tau'_i(\omega)) \\
&\quad = \sum_{\omega\in\Omega'} \pi'(\omega)u_i(\sigma'_{-i}(\omega), \tau'_i(\omega)) .
\end{aligned}$$

In particular, for $\tau_i = \sigma_i$,

$$\sum_{\omega\in\Omega} \pi(\omega)u_i(\sigma_{-i}(\omega), \sigma_i(\omega)) = \sum_{\omega\in\Omega'} \pi'(\omega)u_i(\sigma'_{-i}(\omega), \sigma'_i(\omega)) .$$

Since $\langle(\Omega', \pi'), (\mathcal{P}'_i, \sigma'_i)_{i\in N}\rangle$ is a correlated equilibrium,

$$\sum_{\omega\in\Omega'} \pi'(\omega)u_i(\sigma'_{-i}(\omega), \sigma'_i(\omega)) \geq \sum_{\omega\in\Omega'} \pi'(\omega)u_i(\sigma'_{-i}(\omega), \tau'_i(\omega))$$

and therefore

$$\sum_{\omega\in\Omega} \pi(\omega)u_i(\sigma_{-i}(\omega), \sigma_i(\omega)) \geq \sum_{\omega\in\Omega} \pi(\omega)u_i(\sigma_{-i}(\omega), \tau_i(\omega)) .$$

□

Rationality, Correlated Equilibrium and Equilibrium in Beliefs

As mentioned earlier, Nash equilibrium and correlated equilibrium are two examples of what is known as solution concepts. Solution concepts assign to each game a pattern of behavior for the players in the game. The interpretation of these patterns of behavior is not always explicit, but it is fair to say that they are usually interpreted either as descriptions of what rational people do, or as prescriptions of what rational people should do. There is a growing literature that tries to connect various game theoretic solution concepts to the idea of rationality. Rationality is generally understood as the characteristic of a player who chooses an action that maximizes his preferences, given his information about the environment in which he acts. Part of the information a player has is represented by his beliefs about the behavior of other players, their beliefs about the behavior of other players, and so on. So when one speaks of the rationality of players, one needs to take into account their *epistemic state*. There is a formal framework which is appropriate for discussing the actions, knowledge, beliefs and rationality of players. Namely, the framework of a correlated strategy profile. As defined in Sect. "Correlated Equilibrium", a correlated strategy profile in a game G consists of

- A finite probability space (Ω, π)
- For each player $i \in N$ a partition $\mathcal{P}_i$ of Ω into events of positive probability
- For each player $i \in N$ a function $\sigma_i : \Omega \to A_i$ which is measurable with respect to $\mathcal{P}_i$.

For the present discussion we interpret a correlated strategy profile $\langle(\Omega, \pi), (\mathcal{P}_i)_{i\in N}, (\sigma_i)_{i\in N}\rangle$ as a description of the players' behavior and beliefs, as observed by an outside observer. The set Ω is the set of possible states of the world and π is the prior probability on Ω shared by all the players. For each player $i \in N$, $\mathcal{P}_i$ is a partition of Ω that represents i's information. At state $\omega \in \Omega$, player i is informed not of the state that actually occurred, but of the element $\mathcal{P}_i(\omega)$ of his partition that contains ω. Player i then uses this information and his prior π to update his beliefs about the true state of the world. Finally, the function σ_i represents the actions taken by player i at each state. In particular, $\sigma_i(\omega)$ is the action chosen by i at state ω. Although a correlated equilibrium can be interpreted as a correlated strategy profile *prescribed* by a given solution concept (that of a correlated equilibrium), here we want to interpret a correlated strategy profile as a description of what players actually do and believe. Although players cannot freely choose their beliefs (in the same way as they cannot choose their preferences), they can choose their actions. Furthermore, they have no obligation to behave according to the specified correlated strategy profile. However, ultimately players do behave in a certain way and that behavior is what is represented by the given correlated strategy profile.

Once we fix a correlated strategy profile we can address the rationality of the players. Formally,

Definition 9 Player $i \in N$ is *Bayes rational* at $\omega \in \Omega$ if his expected payoff at ω, $E(u_i(\sigma)|\mathcal{P}_i)(\omega)$, is at least as large

as the amount $E(u_i(\sigma_{-i}, a_i)|\mathcal{P}_i)(\omega)$ that he would have got had he chosen action $a_i \in A_i$ instead of $\sigma_i(\omega)$.

In other words, player i is rational at a given state of the world if the action $\sigma_i(\omega)$ he chooses at that state maximizes his expected utility given his information, $\mathcal{P}_i(\omega)$, and, in particular, given his beliefs about the actions of the other players.

As before, for any finite set T, let $\Delta(T)$ be the set of all probability distributions on T. The beliefs of player i about the actions of the other players are represented by his conjectures. A conjecture of i is a probability distribution $\psi_i \in \Delta(A_{-i})$ over the elements of A_{-i}. For any $j \neq i$, the marginal of ψ_i on A_j is the conjecture of i about j induced by ψ_i. Given a correlated strategy profile $\langle(\Omega, \pi), (\mathcal{P}_i)_{i\in N}, (\sigma_i)_{i\in N}\rangle$, one can determine the conjectures that each player is entertaining at each state of the world about the actions of the other players. These conjectures are given by the following definition.

Definition 10 Given a correlated strategy profile $\langle(\Omega, \pi), (\mathcal{P}_i, \sigma_i)_{i\in N}\rangle$, the conjectures of $i \in N$ about the other players' actions are given by the function $\phi_i : \Omega \to \Delta(A_{-i})$ defined by

$$\phi_i(\omega)(a_{-i}) = \frac{\pi\left[\{\omega' \in \mathcal{P}_i(\omega) : \sigma_{-i}(\omega') = a_{-i}\}\right]}{\pi\left[\mathcal{P}_i(\omega)\right]} .$$

For each ω, $\phi_i(\omega) \in \Delta(A_{-i})$ is the conjecture of i at ω. For $j \neq i$, the marginal of $\phi_i(\omega)$ on A_j is the conjecture of i at ω about j's actions.

Given a correlated strategy profile, we can speak about what each player knows. The object of knowledge are called *events*, which are the subsets of the set of states of the world Ω. We say that player i *knows* event $E \subset \Omega$ at state ω, if $P_i(\omega) \subset E$. That is, i knows E at ω if whatever state he deems possible at ω is in E.

The next result, proved by Aumann and Brandenburger [5], shows a remarkable relationship between the rationality of players and the concept of equilibrium in beliefs.

Theorem 2 *Fix a two-person game, $G = \langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$, and let $\langle(\Omega, \pi), (\mathcal{P}_i)_{i\in N}, (\sigma_i)_{i\in N}\rangle$ be a correlated strategy profile for G. Let $\psi_1 \in \Delta(A_1)$ and $\psi_2 \in \Delta(A_2)$ be two conjectures, one about player 1's actions and the other about player 2's actions. Assume that at some state $\omega \in \Omega$ each player knows that the other is rational and that their conjectures at ω are $(\phi_1(\omega), \phi_2(\omega)) = (\psi_2, \psi_1)$. Then, (ψ_1, ψ_2) is an equilibrium in beliefs.*

Proof The fact that player i knows at ω that j's conjecture is ψ_i means that

$$\mathcal{P}_i(\omega) \subset \{\omega' \in \Omega : \phi_j(\omega')(a_i) = \psi_i(a_i) \quad \text{for all } a_i \in A_i\} .$$

Therefore

$$\phi_j(\omega)(a_i) = \psi_i(a_i) \quad \text{for all } a_i \in A_i . \tag{8}$$

Given Proposition 2 and Corollary 1, we need to show that if $\psi_i(a_i^*) > 0$, a_i^* is a best response to ψ_j, for $i, j = 1, 2$, $i \neq j$. For this purpose, assume that $\psi_i(a_i^*) > 0$ for some $a_i^* \in A_i$. Then, by definition of ϕ_j and (8), $\phi_j(\omega)(a_i^*) = \pi\left[\{\omega' \in \mathcal{P}_j(\omega) : \sigma_i(\omega') = a_i^*\}\right] > 0$. Consequently, there is $\omega' \in \mathcal{P}_j(\omega)$ such that $\sigma_i(\omega') = a_i^*$. Since player j knows at ω that player i is rational,

$$\omega' \in \mathcal{P}_j(\omega) \subset \{\omega'' \in \Omega : E\left[u_i(\sigma)|\mathcal{P}_i\right](\omega'') \geq E\left[u_i(\sigma_{-i}, a_i)|\mathcal{P}_i\right](\omega'') \quad \text{for all } a_i \in A_i\} .$$

Therefore,

$$E\left[u_i(\sigma)|\mathcal{P}_i\right](\omega') \geq E\left[u_i(\sigma_{-i}, a_i)|\mathcal{P}_i\right](\omega') \quad \text{for all } a_i \in A_i$$

and since $\sigma_i : \Omega \to A_i$ is measurable with respect to $\mathcal{P}_i$, $\sigma_i(\omega') = a_i^*$ is the action that player i chooses at all states in $\mathcal{P}_i(\omega')$. Then we can write

$$E\left[u_i(\sigma_{-i}, a_i^*)|\mathcal{P}_i\right](\omega') \geq E\left[u_i(\sigma_{-i}, a_i)|\mathcal{P}_i\right](\omega') \quad \text{for all } a_i \in A_i .$$

That is, for all $a_i \in A_i$

$$\begin{aligned}
&\sum_{\omega''\in\mathcal{P}_i(\omega')} \frac{\pi(\omega'')}{\pi(\mathcal{P}_i(\omega'))} u_i(\sigma_{-i}(\omega''), a_i^*) \\
&\quad \geq \sum_{\omega''\in\mathcal{P}_i(\omega')} \frac{\pi(\omega'')}{\pi(\mathcal{P}_i(\omega'))} u_i(\sigma_{-i}(\omega''), a_i) \\
&\sum_{a_j\in A_j} \sum_{\substack{\omega''\in\mathcal{P}_i(\omega') \\ \sigma_j(\omega'')=a_j}} \frac{\pi(\omega'')}{\pi(\mathcal{P}_i(\omega'))} u_i(a_j, a_i^*) \\
&\quad \geq \sum_{a_j\in A_j} \sum_{\substack{\omega''\in\mathcal{P}_i(\omega') \\ \sigma_j(\omega'')=a_j}} \frac{\pi(\omega'')}{\pi(\mathcal{P}_i(\omega'))} u_i(a_j, a_i)
\end{aligned}$$

$$\sum_{a_j \in A_j} \frac{\pi\left[\left\{\omega'' \in \mathcal{P}_i(\omega') : \sigma_j(\omega'') = a_j\right\}\right]}{\pi(\mathcal{P}_i(\omega'))} u_i(a_j, a_i^*)$$
$$\geq \sum_{a_j \in A_j} \frac{\pi\left[\left\{\omega'' \in \mathcal{P}_i(\omega') : \sigma_j(\omega'') = a_j\right\}\right]}{\pi(\mathcal{P}_i(\omega'))} u_i(a_j, a_i)$$
$$\sum_{a_j \in A_j} \phi_i(\omega')(a_j) u_i(a_j, a_i^*)$$
$$\geq \sum_{a_j \in A_j} \phi_i(\omega')(a_j) u_i(a_j, a_i) \,. \tag{9}$$

Since $\omega' \in \mathcal{P}_j(\omega)$ and player j knows at ω that i's conjecture is ψ_j, then

$$\omega' \in \mathcal{P}_j(\omega) \subset \left\{\omega'' \in \Omega : \phi_i(\omega'')(a_j) = \psi_j(a_j) \text{ for all } a_j \in A_j\right\} .$$

Therefore $\phi_i(\omega')(a_j) = \psi_j(a_j)$ for all $a_j \in A_j$, or $\phi_i(\omega') = \psi_j$. That is, i's conjecture at ω' about j's actions is ψ_j. Consequently, substituting in (9),

$$\sum_{a_j \in A_j} \psi_j(a_j) u_i(a_j, a_i^*) \geq \sum_{a_j \in A_j} \psi_j(a_j) u_i(a_j, a_i) \quad \forall a_i \in A_i \,.$$

That is, a_i^* is a best response to player i's beliefs about j's actions. □

The only assumptions required by Theorem 2 is that players know they are rational, and that they know each other's conjectures. In a correlated strategy profile for a two-player game, there is only one player entertaining a conjecture about the actions of player 1, namely, player 2. Similarly, player 1 is the only one who entertains a conjecture about the actions of player 2. In an n-person game, with $n > 2$, for each player, there is more than one player entertaining a conjecture about his actions. Therefore, since an equilibrium in beliefs consists of a profile of beliefs, each of which is shared by $n-1$ players, a generalization of Theorem 2 would require the players' beliefs about player i's actions, for $i \in N$, to be identical. In order to obtain these common beliefs it is not sufficient to assume that players know each other's conjectures. One need to strengthen this assumption. Also, in an equilibrium in beliefs, the common belief about player i's actions assigns positive probability only to best responses to i's conjectures about the choices of the other players. Furthermore, i's conjectures about the other players' choices is the product of his beliefs about each of the other players. In other words, an equilibrium in beliefs implicitly assumes that players believe that the other players' choices are independent. Aumann and Brandenburger [5] show that one way to obtain common conjectures and, simultaneously, that players believe that the other players act independently, is to assume that players' conjectures are commonly known. This surprising and deep result is stated in the next theorem.

Theorem 3 *Let $G = \langle N, (A_i)_{i\in N}, (u_i)_{i\in N}\rangle$ be a strategic game, and let $\langle(\Omega, \pi), (\mathcal{P}_i)_{i\in N}, (\sigma_i)_{i\in N}\rangle$ be a correlated strategy profile for G. Also let $(\psi_i)_{i\in N} \in \times_{i\in N}\Delta(A_{-i})$ be a profile of conjectures, one for each player. Assume that at some state $\omega \in \Omega$ each player knows that the others are rational. Further, assume that at ω their conjectures are commonly known to be $(\psi_i)_{i\in N}$. Then, for each j, all the conjectures ψ_i of players i other than j, induce the same belief $\varphi_j \in \Delta(A_j)$ about j's actions, and the resulting profile of beliefs, $(\varphi_i)_{i\in N}$, is an equilibrium in beliefs.*

Rationality and Correlated Equilibrium

The previous result shows a surprising relationship between the players' rationality and the concept of equilibrium in beliefs. If at some state of the world players know that everybody is rational, and if their conjectures are commonly known at that state, then their beliefs about each player's actions are in equilibrium. It is not that their actions constitute an equilibrium, but that their beliefs do. The question that naturally arises is: are there any epistemic conditions on the players that would induce them to play according to equilibrium? To answer this we turn to Aumann [3], where it is stated that if players are rational at every state, then their behavior constitutes a correlated equilibrium. Therefore, in order to obtain an equilibrium behavior, a sufficient condition is not that players be rational, or that they know that they are rational at some particular state, but that their rationality be common knowledge. And if it is common knowledge that all players are rational, then their behavior is not necessarily a Nash equilibrium, but a correlated equilibrium.

Theorem 4 *Let G be a strategic game, and let $\langle(\Omega, \pi), (\mathcal{P}_i)_{i\in N}, (\sigma_i)_{i\in N}\rangle$ be a correlated strategy profile for G. If each player is rational at each state of the world, then $\langle(\Omega, \pi), (\mathcal{P}_i)_{i\in N}, (\sigma_i)_{i\in N}\rangle$ is a correlated equilibrium.*

Proof Let $\tau_i : \Omega \to A_i$ be a function that is measurable with respect to $\mathcal{P}_i$. Since i is Bayes rational at ω

$$E(u_i(\sigma)|\mathcal{P}_i)(\omega) \geq E(u_i(\sigma_{-i}, a_i)|\mathcal{P}_i)(\omega) \quad \forall a_i \in A_i \,.$$

That is,

$$\sum_{\omega' \in \mathcal{P}_i(\omega)} \frac{\pi(\omega')}{\pi(\mathcal{P}_i(\omega))} u_i(\sigma_{-i}(\omega'), \sigma_i(\omega'))$$
$$\geq \sum_{\omega' \in \mathcal{P}_i(\omega)} \frac{\pi(\omega')}{\pi(\mathcal{P}_i(\omega))} u_i(\sigma_{-i}(\omega'), a_i) \quad \forall a_i \in A_i .$$

In particular, for $a_i = \tau(\omega) = \tau(\omega')$ for all $\omega' \in \mathcal{P}_i(\omega)$,

$$\sum_{\omega' \in \mathcal{P}_i(\omega)} \frac{\pi(\omega')}{\pi(\mathcal{P}_i(\omega))} u_i(\sigma_{-i}(\omega'), \sigma_i(\omega'))$$
$$\geq \sum_{\omega' \in \mathcal{P}_i(\omega)} \frac{\pi(\omega')}{\pi(\mathcal{P}_i(\omega))} u_i(\sigma_{-i}(\omega'), \tau(\omega')) .$$

Multiplying both sides by $\pi(\omega)$ and adding over all the elements of $\mathcal{P}_i$ we get

$$\sum_{\omega \in \Omega} \pi(\omega) u_i(\sigma_{-i}(\omega), \sigma_i(\omega))$$
$$\geq \sum_{\omega \in \Omega} \pi(\omega) u_i(\sigma_{-i}(\omega), \tau(\omega)) .$$

□

Bayesian Games

Thus far, we have considered static games, which are objects of the form $\langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$. Although these games have many applications, they are not readily suitable for the analysis of situations involving asymmetric information. Indeed, an implicit assumption behind the definition of a static game is that all players have the same information about the relevant aspects of the situation. In particular, all players have the same information about the sets of actions and preferences of all players. A static game seems suitable to model strategic interactions like the prisoner's dilemma, rock scissors paper, and even chess. At the time they choose their actions, all the players have exactly the same information. There might be what is called strategic uncertainty, namely, uncertainty about what the players will do, but there is no uncertainty about the rules of the game and about the preferences of the players. But how would one translate a game of cards like bridge or poker into a static game? In a game of cards, at the time of choosing his actions, each player knows the cards he holds in his hand, but does not know the cards of his opponents. He only has a belief about the cards held by his opponents. In order to make a sound choice, a player will try to predict the actions of his opponents, but for this it is crucial to use his beliefs about the cards they hold. For the same reason, his opponents should use their beliefs about their own opponents' cards in order to make a sound choice. Thus, the beliefs about the cards held by each player should be part of a description of a game with asymmetric information. Further, in order to predict his opponents' actions, a player also needs to assess his opponents' beliefs about his own cards. This seems to induce an intractable infinite regress of beliefs, and beliefs about beliefs. Harsanyi [12] provided the basic structure to describe and analyze strategic situations where players are asymmetrically informed. This structure is called a Bayesian game.

Definition 11 A *Bayesian Game* is a system $\langle N, (\Omega, \mu), (A_i, \mathcal{P}_i, u_i)_{i \in N} \rangle$ where

- N is the set of players
- Ω is the set of states of nature
- μ is the players' common prior belief (a probability measure over the set of states)
- A_i is player i's set of actions
- $\mathcal{P}_i$ is player i's information partition (a partition of Ω into sets of positive measure). Each element of the partition is referred to as a player's type.
- $u_i : \times_{i \in N} A_i \times \Omega$ is player i's Bernoulli utility function (a function over pairs (a, ω) where $a \in A$ and $\omega \in \Omega$, the expected value of which represents the player's preferences among lotteries over the set of such pairs).

The interpretation of a Bayesian game is as follows. The basic uncertainty is represented by the probability space (Ω, μ) of all states of nature and the prior probability over them. Each state represents a realization of all the parametric uncertainty of the model. For instance, in a game of cards, each state represents each of the possible card deals. The information of player $i \in N$ is represented by his information partition $\mathcal{P}_i$. While states in the same element of the partition cannot be distinguished by the player, he can distinguish between states that belong to different partition cells. In a game of cards, for instance, each partition cell represents a particular set of cards dealt to the player. The probability measure μ represents the players' prior belief about the state of nature. This prior belief will be used along with the information obtained by each player to form beliefs about the other players' information. The set of actions of player i is A_i. Note that there is no loss of generality in assuming that this set does not depend on the state of nature. One can always add unavailable actions and assign them intolerable disutility. Finally, u_i is the payoff function that associates to each state of nature and action profile a utility level. Note that since the state of the world is unknown to the player at the time of making his choice, a player faces a lottery for any given action profile. The assumption is that the player evaluates this lottery

according to the expected value of u_i with respect to that lottery.

Let $\langle N, (\Omega, \mu), (A_i, \mathcal{P}_i, u_i)_{i \in N} \rangle$ be a Bayesian game. A strategy for player $i \in N$ is a function $\sigma_i : \Omega \to A_i$ that is measurable with respect to $\mathcal{P}_i$. We denote the set of strategies for player i by $\mathcal{B}_i$. That is, $\mathcal{B}_i = \{\sigma_i : \Omega \to A_i : \sigma_i \text{ is measurable w.r.t. } \mathcal{P}_i\}$. The interpretation of a strategy in a Bayesian game is the usual one. For each state of nature $\omega \in \Omega$, $\sigma_i(\omega)$ is the action chosen by player i at ω. The measurability requirement imposes that player i's actions depend only on his information. If player i cannot distinguish between two states of nature, then he must choose the same action at both states. Player i evaluates a profile $\sigma : \Omega \to A$ of strategies according to the expected value of u_i with respect to μ.

In order to define an equilibrium notion for Bayesian games we follow the same idea used for the definition of a mixed strategy equilibrium. Namely, we translate the Bayesian game into a standard game, and then define an equilibrium of the Bayesian game as the Nash equilibirum of the induced game.

Definition 12 A *Bayesian equilibrium* of a Bayesian game $\langle N, (\Omega, \mu), (A_i, \mathcal{P}_i, u_i)_{i \in N} \rangle$ is a Nash equilibrium of the strategic game: $\langle N, (\mathcal{B}_i)_{i \in N}, (U_i)_{i \in N} \rangle$ where for each profile $\sigma : \Omega \to A$ of strategies, $U_i(\sigma) = E_\mu[u_i(\sigma(\omega), \omega)]$ is i's expected utility with respect to μ.

A Bayesian equilibrium of a Bayesian game is a Nash equilibrium of a properly defined static game. As such, conditions for its existence can be derived from Theorem 1. However, in many situations one is interested in particular kinds of equilibria. Specifically, in the analysis of auctions or of the war of attrition, one is often interested in efficient outcomes. In a single object auction, efficient outcomes are characterized by the fact that in equilibrium the object is allocated to the buyer who values it most. According to many standard auction rules, the object goes to the highest bidder. Therefore, in such auctions, to guarantee an efficient outcome, one would need a monotone equilibrium, namely, one in which bidders bids are higher the higher their valuations for the object are. Athey [1] shows conditions under which a Bayesian equilibrium exists where strategies are non-decreasing. The crucial conditions are that the players' types can be represented by a one-dimensional variable, and that, fixing a nondecreasing strategy for each of a player's opponents, this player's expected payoffs satisfies a single-crossing property. This single-crossing property roughly says that if a high action is preferred to a low action for a given type t, then the same must be true for all types higher than t. McAdams [16] extended Athey's result to the case where types and actions are multidimensional and partially ordered.

The Asymmetric Information Version of the War of Attrition

We have seen that, when applied to the war of attrition, as modeled by a standard strategic game or by its mixed extension, the notion of Nash equilibrium does not yield a satisfactory prediction.[1] In the former case all the equilibria involve no fight, and in the latter case the equilibrium dictates a more aggressive behavior to the player who values the contested object less. In what follows, we analyze the war of attrition as a Bayesian game. That is, we assume that the players are ex-ante symmetric but they have private information about their value for the contested object.

A Bayesian game that represents the war of attrition is given by $\langle N, \Omega, (A_i, \mu_i, \mathcal{P}_i, u_i)_{i \in N} \rangle$ where

- $N = \{1, 2\}$
- $\Omega = [0, \infty)^2 = \{(v_1, v_2) : 0 \le v_i < \infty, i = 1, 2\}$
- $A_i = [0, \infty)$ for $i = 1, 2$
- $\mathcal{P}_i(\hat{v}_1, \hat{v}_2) = \{(v_1, v_2) \in \Omega : v_i = \hat{v}_i\}$ for $i = 1, 2$
- $\mu((v_1, v_2) \le (\hat{v}_1, \hat{v}_2)) = F(\hat{v}_1) \times F(\hat{v}_2)$
- $u_i((a_1, a_2), (v_1, v_2)) = \begin{cases} -a_i & \text{if } a_i \le a_j \\ v_i - a_j & \text{if } a_i > a_j \end{cases}.$

Here the set of types of player i, for $i = 1, 2$, is represented by the player's willingness to fight, v_i. The players' willingness to fight are drawn independently from the same distribution F. A state of the world is, therefore, a realization (v_1, v_2) of the players' types, and at that state, each player is informed only of his type. Finally, the utility of a player is his valuation for the prey, if he obtains it, net of the time spent fighting for it. We are interested in a symmetric equilibrium in which both players use a symmetric, strictly increasing strategy $\beta : [0, \infty) \to [0, \infty)$, where $\beta(v)$ is the time at which a player with willingness to fight v is dictated by the equilibrium to give up. Such an equilibrium would imply that types who value the prey more, are willing to fight more. Further, the probability of observing a fight in equilibrium would not be 0 (in fact, it would be 1.)

It turns out that a symmetric equilibrium strategy is given by

$$\beta(v) = \int_0^v \frac{x f(x)}{1 - F(x)} \mathrm{d}x,$$

[1] The war of attrition was analyzed in Maynard Smith [14]. For an analysis of the asymmetric information version of the war of attrition, see Krishna and Morgan [13].

where f denotes the derivative of F. To see this, assume that player j behaves according to β and that player i chooses to give up at t. Letting z be the type such $\beta(z) = t$, the expected utility of player i from choosing t is

$$U(v_i, z) = \int_0^z (v_i - \beta(y))f(y)\mathrm{d}y - \beta(z)(1 - F(z)) .$$

Taking derivatives with respect to z, and using the fact that $\beta'(z) = zf(z)/[1 - F(z)]$ we obtain

$$\begin{aligned}\frac{\partial U}{\partial z}(v_i, z) &= v_i f(z) - \beta'(z)(1 - F(z)) \\ &= (v_i - z)f(z),\end{aligned}$$

which is positive for $z < v_i$, and negative for $z > v_i$. As a result, the expected utility of player i with willingness to pay v_i is maximized at $z = v_i$, which implies that the optimal choice is $\beta(v_i)$.

Thus, modeling the war of attrition as an asymmetric game has allowed us to find an equilibrium in which players with higher willingness to fight fight more, and there is a non-negligible probability of observing a fight.

Evolutionary Stable Strategies

The notion of the Nash equilibrium concept involves players choosing actions that maximize their payoffs given the choices of the other players. The usual interpretation of a Nash equilibrium is as a pattern of behavior that rational players *should* adopt. However, Nash equilibria are sometimes interpreted more descriptively as patterns of behavior that rational players *do* adopt. Certainly, rationality of players is neither a necessary condition nor a sufficient one for players to play a Nash equilibrium. The relationship between rationality and the various solution concepts is not apparent and has been the focus of an extensive literature (see, for example, [3,4,5,7]). Nonetheless, the notion of a Nash equilibrium evokes the idea of players consciously making choices with the deliberate objective of maximizing their payoffs. It is therefore quite remarkable that a concept almost identical to that of Nash equilibrium has emerged from the biology literature. This concept describes a population equilibrium where unconscious organisms are programmed to choose actions with no deliberate aim. In this equilibrium, members of the population meet at random over and over again to interact. At each interaction, these players act in a pre-programmed way and the result of their actions is a gain in biological fitness. Fitness is a concept related to the reproductive value or survival capacity of an organism. In a temporary equilibrium, the fitness gains are such that the proportions of individuals that choose each one of the possible actions remain constant. However, this temporary equilibrium may be disturbed by the appearance of a mutation, which is a new kind of behavior. This mutation may upset the temporary equilibrium if its fitness gains are such that the new behavior spreads over the population. Alternatively, if the fitness gains of the original population outweigh those of the mutation, then the new behavior will fail to propagate and will eventually disappear. In a population equilibrium, the interaction of any mutant with the whole population awards the mutant insufficient fitness gains, and as a result the mutants disappear. The notion of a population equilibrium is formalized by means of the concept of an evolutionary stable strategy, introduced by Maynard Smith and Price [15].

In what follows we restrict our attention to symmetric two-player games. So let $G = \langle\{1, 2\}, \{A_1, A_2\}, \{u_1, u_2\}\rangle$ be a game such that $A_1 = A_2 = \overline{A}$, and such that for all $a, b \in \overline{A}$, $u_1(a, b) = u_2(b, a)$. An evolutionary stable strategy is an action in $\overline{A}$ such that if all members of the population were to choose that action, no sufficiently small proportion of mutants choosing an alternative action would succeed in invading the population. Alternatively, an evolutionary stable strategy is an action in $\overline{A}$ such that if all the members of the population were to choose that action, the population would reject all sufficiently small mutations involving a different action.

More specifically, suppose that all members of the population are programmed to choose $a \in \overline{A}$, and then a proportion ε of the population mutates and adopts action $b \in \overline{A}$. In that case, the probability that a given member of the population meets a mutant is ε, while the probability of meeting a member that plays a is $1 - \varepsilon$. Therefore, the mutation will not propagate and will vanish if the expected payoff of a mutant is less than the expected payoff of a member of the majority. Otherwise it will propagate. This leads to the following definition.

Definition 13 An action $a \in \overline{A}$ is an evolutionary stable strategy of G if there is an $\bar{\varepsilon} \in (0, 1)$ such that for all $\varepsilon \in (0, \bar{\varepsilon})$, and for all $b \in \overline{A}$

$$(1-\varepsilon)u_1(a, a) + \varepsilon u_1(a, b) > (1-\varepsilon)u_1(b, a) + \varepsilon u_1(b, b) . \tag{10}$$

The following result shows that the concept of an evolutionary stable strategy is very close to the notion of a Nash equilibrium.

Proposition 5 *If $a \in \overline{A}$ is an evolutionary stable strategy of G, then (a, a) is a Nash equilibrium. And if (a, a) is*

a strict Nash equilibrium then a is an evolutionary stable strategy.

Proof If $u_1(a,a) > u_1(b,a)$ for all $b \in \overline{A} \setminus \{b\}$, then inequality (10) holds for all sufficiently small $\varepsilon > 0$. If $u_1(b,a) > u_1(a,a)$ for some $b \in \overline{A}$, the reverse inequality holds for all sufficiently small ε. □

Future Directions

Static games have been shown to be a useful framework for analyzing and understanding many situations that involve strategic interaction. At present, a large body of literature is available that develops various solution concepts, some of which are refinements of Nash equilibrium and some of which are coarsenings of it. Nonetheless, several areas for future research remain. One is the application of the theory to particular games to better understand the situations they model, for example auctions. In many markets trade is conducted by auctions of one kind or another, including markets for small domestic products as well as some centralized electricity markets where generators and distributors buy and sell electric power on a daily basis. Also, auctions are used to allocate large amounts of valuable spectrum among telecommunication companies. It would be interesting to calculate the equilibria of many real life auctions. Simultaneously, future research should also focus on the design of auctions whose equilibria have certain desirable properties.

Another future direction would be to empirically and experimentally test the theory. The various equilibrium concepts predict certain kinds of behavior in certain games. Our confidence in the predictive and explanatory power of the theory depends on its performance in the field and in the laboratory. Moreover, the experimental and empirical results should provide valuable feedback for further development of the theory. Although some valuable experimental and empirical tests have already been performed (see [17,19,22,25] to name a few), the empirical aspect of game theory in general, and of static games in particular, remains underdeveloped.

Bibliography

1. Athey S (2001) Single crossing properties and the existence of pure strategy equilibria in games of incomplete information. Econometrica 69:861–889
2. Aumann RJ (1974) Subjectivity and correlation in randomized strategies. J Math Econ 1:67–96
3. Aumann RJ (1987) Correlated equilibrium as an expression of bayesian rationality. Econometrica 55:1–18
4. Aumann RJ (1995) Backward induction and common knowledge of rationality. Games Econ Behav 8:6–19
5. Aumann RJ, Brandenburger A (1995) Epistemic conditions for Nash equilibrium. Econometrica 63:1161–1180
6. Binmore K (2007) Playing for Real. Oxford University Press, New York
7. Brandenburger A, Dekel E (1987) Rationalizability and correlated equilibria. Econometrica 55:1391–1402
8. Crawford V (1990) Equilibrium without independence. J Econ Theory 50:127–154
9. Fudenberg D, Tirole J (1991) Game Theory. MIT Press, Cambridge
10. Geanakoplos J (2003) Nash and Walras equilibrium via Brouwer. Econ Theory 21:585–603
11. Glicksberg IL (1952) A further generalization of the Kakutani fixed point theorem, with application to Nash equilibrium points. Proc Am Math Soc 3:170–174
12. Harsanyi J (1967) Games with incomplete information played by 'Bayesian' players. Manag Sci Part i 14:159–82; Part ii 14:320–34; Part iii 14:486–502
13. Krishna V, Morgan J (1997) An analysis of the war of attrition and the all-pay auction. J Econ Theory 72:343–362
14. Smith MJ (1974) The theory of games and the evolution of animal conflicts. J Theor Biol 47:209–221
15. Smith MJ, Price GR (1973) The logic of animal conflict. Nature 246:15–18
16. McAdams D (2003) Isotone equilibrium in games of incomplete information. Econometrica 71:1191–1214
17. McKelvey RD, Palfrey TR (1992) An experimental study of the centipede game. Econometrica 60:803–836
18. Nash JF (1950) Equilibrium points in n-person games. Proc Natl Acad Sci USA 36:48–49
19. O'Neill B (1987) Nonmetric test of the minimax theory of two-person zero-sum games. Proc Natl Acad Sci 84:2106–2109
20. Osborne MJ (2004) An Introduction to Game Theory. Oxford University Press, New York
21. Osborne MJ, Rubinstein A (1994) A Course in Game Theory. MIT Press, Cambridge
22. Palacios-Huerta I (2003) Professionals play minimax. Rev Econ Stud 70:395–415
23. von Neumann J (1928) Zur Theorie der Gesellschaftsspiele. Math Ann 100:295–320
24. von Neumann J (1959) On the theory of games of strategy. In: Tucker AW, Luce RD (eds) Contributions to the Theory of Games, vol IV. Princeton University Press, Princeton
25. Walker M, Wooders J (2001) Minimax play at Wimbledon. Am Econ Rev 91:1521–1538

Statistical Applications of Wavelets

Sofia Olhede
Department of Statistical Science,
University College London, London, UK

Article Outline

Glossary
Definition of the Subject
Introduction

Decomposition View of Random Processes
Wavelets
Signal Estimation or Denoising
2-D Extensions of Wavelets
Covariance Estimation Using Wavelets
Future Directions
Bibliography

Glossary

Locally stationary process A stochastic process that when sampled in time, its sampled points have joint characteristics in a short time window that only depend on the difference between the sampled time points, and not the global time point of the samples within the time window.

Multiresolution analysis The analysis of a phenomenon of interest over different temporal or spatial scales and locations.

Shrinkage A method of estimation where the initial estimator is decreased by multiplication with a factor of magnitude less than or equal to one. If the object being estimated is small, then a reduced variance will arise, that sets off the increased bias in estimation.

Sparsity Sparsity in a vector corresponds to the support of the vector being limited. The sparsity can either be strict, i. e. the vector is perfectly supported in a small subset of all its entries, or the magnitudes of the entries decay sufficiently rapidly, if ordered in magnitude.

Stationary process A stochastic process that when sampled in time, its sampled points have joint characteristics that only depend on the difference between the sampled time points, and not the global time point of the samples.

Thresholding A method of estimation where the initial estimator is set to zero if it does not exceed a given threshold. Thresholding is a special case of shrinkage.

Wavelet and scaling functions The functions that the wavelet and scaling filters converge to in increasing scale, when using discrete wavelet filters. These are not for any finite scale equivalent to the wavelet and scaling filters.

Wavelet and scaling filters The high and low pass digital filters that are used to calculate the discrete wavelet transform of a given sequence or vector. These are a pair of quadrature mirror filters that satisfy the perfect reconstruction property.

Definition of the Subject

A wavelet is a function essentially supported over a small time-window, that is oscillatory, with a local period that ranges over a limited sets of values. A wavelet function can be considered jointly local in time and frequency. Wavelets have become an integral part of statistical analysis of structured phenomena. Their utility for inference is based on several of their important properties, namely compression, separation of multiscale features, and differencing of a given set of moments. The statistical analysis of processes using wavelets presents a more faceted view of the phenomenon of interest than traditional methods. The utility of this in many areas is unquestionable, and statistical applications of wavelets include areas as disparate as climatology, econometrics, finance, geophysics, mathematics, as well as signal and image processing.

Introduction

The field of wavelets in statistics can roughly be divided into two areas, namely the usage of continuous wavelets, and the usage of discrete wavelets. Continuous wavelets are defined at a continuum of locations, whilst discrete wavelet filters are sequences defined only at a countable number of locations and scales. These sequences are not strictly speaking wavelet functions, but converge in increasing scale to such functions. Unfortunately in the literature both functions and filters are referred to as 'wavelets'.

Continuous wavelets are applied to the characterization of random processes, but are not usually used for actual signal estimation, as the transform does not possess an exact discretely implementable inverse. Continuous wavelets do not necessarily satisfy the constraint of compactness in time, and this permits more freedom in their design. A third class of wavelets that often receive separate treatment are complex-valued discrete wavelet filters. These can be thought of as bridging the gap between continuous and discrete representations. Complex wavelets are often chosen to be redundant, and have interpretable magnitude and phase, see for example the review of these in [48].

The statistical representation of structured phenomena using decompositions was first introduced by [47] who analyzed time series using Fourier decompositions. This analysis method corresponds to decomposing a sample into cosines and sines to determine the power associated with each frequency or mode. The interpretation of the observed weight attached to each frequency depends on the model posited for the observed data. If the process is assumed to be *stationary*, then the observed weight can be shown to converge in expectation to the true weight, namely the Fourier transform of the autocovariance of the

time series. By calculating the full set of weights across frequencies, we may obtain a picture of the frequency content of the process. Even for processes that are not *stationary*, decomposing the set of observations into sinusoids allows us to visualize the frequency content of the signal. As long as the process is *harmonizable*, see [36], such decompositions are suitable.

Unfortunately whilst the decomposition of a time-varying process into sines and cosines is interpretable in a global sense, and can be shown to possess some suitable properties, it is not as easily interpreted as might be the case. Many processes are generated by mechanisms that fundamentally alter over time. It is therefore suitable to use decomposition functions that unlike sines are limited in time, but still correspond to oscillatory functions. A popular choice of such functions is wavelets.

A wavelet is a time-limited oscillation, or a "little wave". The first family was constructed by Alfred Haar in 1910, see [26]. The subject lay dormant for many years, until the mid 1980s when their usage in geophysics revitalized interest in the set of decompositions, by Morlet, and one of the groundbreaking new articles corresponds to [24]. A great theoretical leap forward was the construction of discrete wavelet filters, see [10]. Most of the statistical usage of wavelets has focused around the usage of discrete wavelets.

The development of statistical methods can really be grouped into three main strands of development: i) the usage of compression for non-parametric signal estimation, following work by [15,16,17], ii) the characterization of stochastic difference stationary processes, see [38,51], and iii) the inference of locally stationary wavelet processes, see [40]. Theoretical results have also been achieved in terms of the characterization of stochastic processes, see [3,4,8,34], but these results as of yet have had limited practical implications.

After initial usage of the wavelet transform for signal estimation it became obvious that despite the wavelet transform's orthogonalizing effects, relationships between coefficients must be used, to derive better representations, see [58] The development of data driven decompositions, and usage of the 'lifting algorithm', [12], followed in the second strand of developments. Further steps forward included the construction of two dimensional decompositions for image estimation, where notable contributions include the ridgelet, curvelet and bandelet transforms, see [9,33,52]. Outstanding problems remain in the statistical usage of these transforms, and investigations have mainly been limited to modeling the observed phenomena as deterministic and immersed in some form of noise.

Decomposition View of Random Processes

Wavelets are used to analyze observations whose generating mechanism can be characterized in terms of the location the observation is made at. This could be time, space or some other suitable indexing. We shall denote the location of any observation by $t \in \mathbb{R}^d$, where d denotes the dimension of the indexing, this could be $d = 1$, like a time series, or $d = 2$, like an image. If $\{x(t)\}$ is a random process then its structure can be determined from its moments. The first order structure of $x(t)$ is given by its mean, defined by

$$\mu_x(t) = \mathrm{E}\{x(t)\}, \quad t \in \mathbb{R}^d, \tag{1}$$

and the second order structure of $x(t)$ is given by its covariance, namely

$$\gamma_x(t_1, t_2) = \mathrm{cov}\{x(t_1), x(t_2)\}, \quad t_1, \ t_2 \in \mathbb{R}^d. \tag{2}$$

$\mathrm{E}\{\cdot\}$ denotes the expectation operator, and $\mathrm{cov}\{\cdot,\cdot\}$ the covariance operator. If the process is assumed to be a Gaussian process, i. e. if any sample of n observations from $\{x(t)\}$ is multivariate Gaussian, then $\{\mu_x(t)\}$ in combination with $\{\gamma_x(t_1, t_2)\}$ fully characterize the generation of the process $\{x(t)\}$.

Often we only take n observations from the process $\{x(t)\}$ and still want to fully characterize the generating mechanism, so that we can test hypotheses of interest regarding $\{x(t)\}$ or just estimate some parameters of interest. To be able to do this we must make some further assumptions about $x(t)$, and these could correspond to fully parametric or non-parametric assumptions on $\{\mu_x(t)\}$ or $\{\gamma_x(t_1, t_2)\}$. Typical examples might be $\mu_x(t) = a_x + b_x t$, which is a parametric specification of the mean, or $\mu_x(t)$ possessing a certain degree of smoothness. If the latter approach is taken then it is convenient to represent $\mu_x(t)$ in terms of some functions where the assumed smoothness can be easily incorporated into the modeling. Furthermore the covariance structure can be assumed to simplify if we choose a suitable representation. If $x(t)$ is represented in terms of linear combinations of known basis functions, such as wavelets, then Gaussian processes will be transformed into Gaussian processes with a different mean and covariance function. For certain classes of auto-covariance functions the analysis is much simplified by considering the transformed rather than the original Gaussian process.

Wavelets

A wavelet function, is a complex-valued square integrable function satisfying the admissibility condition, see [27]. We denote a generic wavelet by $\psi(t)$. The Continuous

Wavelet Transform (CWT) of a signal $x(t)$ with respect to a wavelet $\psi(t)$ is defined as a filtration of $x(t)$:

$$w^{(x)}(t,s) = \int_{-\infty}^{\infty} x(t')\psi_{t,s}(t')\,\mathrm{d}t' ,$$
$$\psi_{t,s}(t') = s^{\frac{-1}{2}}\psi\left(\frac{t'-t}{s}\right) . \quad (3)$$

The choice of normalization, here $s^{-1/2}$ varies across application fields. Sometimes 1/2 is replaced by 1 if more convenient for analysis, see [54]. The former is referred to as the *energy normalization*, and the latter as the *compression normalization*. The former normalization conserves the variance of the transformed process.

If $\{x(t)\}$ is a stochastic process satisfying some regularity conditions, see [34], then $\left\{w^{(x)}(t,s)\right\}$ is a doubly indexed stochastic process with mean

$$\mathrm{E}\left\{w^{(x)}(t,s)\right\} = \int_{-\infty}^{\infty} \mu_x(t')\psi_{t,s}(t')\,\mathrm{d}t' ,$$
$$\psi_{t,s}(t') = s^{\frac{-1}{2}}\psi\left(\frac{t'-t}{s}\right) \quad (4)$$

$$\equiv w^{(\mu_x)}(t,s) . \quad (5)$$

The form of this expression simplifies if $x(t)$ takes the simpler form of a time-varying oscillation, see [13], or a singularity [28,55], or is a sufficiently regular function, see [56]. Under such circumstances very few of $\left\{w^{(\mu_x)}(t,s)\right\}$ are non-zero and inferences can be made about the model from the few non-zero coefficients. In the special case of time-varying oscillations the theory of *wavelet ridge analysis* has been developed to characterize the local oscillations. Applications of such methods include mechanical vibratory signals, see [31,59], seismic signals [44,45], as well as ocean eddy time series, see [6,35].

The usage of continuous wavelets does not extend greatly beyond the aforementioned methods of signal inference. The characterization of various stochastic processes have been investigated using continuous wavelets, i. e. [3,8,25,32,34], but this has been put to little practical usage. Some interesting developments for the prediction of continuous signals using wavelets have been developed in [1,2].

The main focus of wavelet usage in statistics has instead evolved around the usage of the Discrete Wavelet Transform (DWT). The DWT is defined as a reversed convolution between a discretely sampled signal and the wavelet filter. We denote the jth level wavelet filter $\left\{h_{j,l}\right\}$, and the scaling filter $\left\{g_{j,l}\right\}$. For a complete review of the properties of these objects see [11,37,46]. These are constructed from the $j = 1$ level filters $\{h_l\}$ and $\{g_l\}$, that satisfy the quadrature mirror relationship of $g_l = (-1)^{l+1}h_{L-1-l}$ where L is the length of the two filters. Given a sampled process $x_n = x(n\Delta t)$ the wavelet coefficients are defined by:

$$w_{j,l}^{(x)} = \sum_n h_{j,l}x_n , \quad h_{j,l} = \sum_{k=0}^{L-1} h_k g_{j-1,l-2^{j-1}k} , \quad (6)$$

where the wavelet filters given $\left\{g_{j-1,l}\right\}$ can now be determined iteratively. To complement the wavelet coefficients of x_n, additionally the scaling coefficients are computed, of

$$v_{j,l}^{(x)} = \sum_n g_{j,l}x_n , \quad g_{j,l} = \sum_{k=0}^{L-1} g_k g_{j-1,l-2^{j-1}k} , \quad (7)$$

this completing the iterative specification of the wavelet and scaling filters. If we collect the wavelet and scaling coefficients in a vector $\mathbf{W}^{(x)} = [w_{1,0}^{(x)}, \ldots, w_{1,N/2-1}^{(x)}, w_{2,0}^{(x)}, \ldots, v_{J/2,0}^{(x)}]$, then the DWT can be represented as:

$$\mathbf{W}^{(x)} = \mathcal{W}\mathbf{X} , \quad (8)$$

where $\mathcal{W}$ is composed of $\left\{h_{j,l}\right\}$ and $\left\{g_{J,l}\right\}$, see [46]. Equation (8) is not the method to compute the DWT, popularly the transform is implemented using Mallat's pyramid algorithm, see [37], but it enables the easy determination of the stochastic properties of the wavelet transform. The CWT and the DWT are not two disparate objects: if we possess a sequence of coefficients $\left\{v_{0,l}^{(x)}\right\}$ such that

$$x(t) = \sum_{n=-\infty}^{\infty} v_{0,l}^{(x)}\phi_{0,k}(t) \quad (9)$$

then $w_{j,l}^{(x)}$ can be calculated from $v_{0,l}^{(x)}$, as can $v_{j,l}^{(x)}$, and correspond to the CWT as well as scaling coefficients of the continuous rather than the sampled process. In practise x_n is equated to $v_{0,n}^{(x)}$, even if this is formally inappropriate.

Signal Estimation or Denoising

The most popular statistical application of wavelets is denoising, or function estimation. This proceeds from a model of the observed signal of:

$$x_n = \mu(n\Delta t) + \epsilon_n, \quad n = 0, \ldots, N-1 , \quad (10)$$

where Δt is referred to as the sampling period of the signal, and the sample size is N. To fully specify the generation of the observations $\{\epsilon_n\}$ must be modeled, and a popular model is to assume this sequence as zero-mean, uncorrelated (white) and Gaussian, with some constant vari-

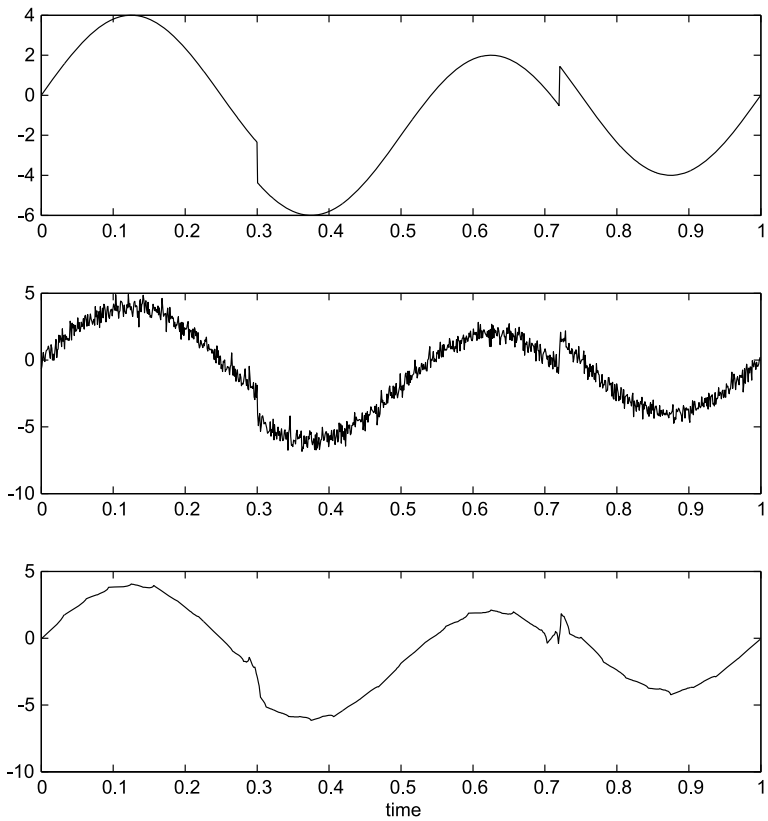

Statistical Applications of Wavelets, Figure 1

The clean Heavisine signal (*top*), the noisy Heavisine signal (*middle*) and the denoised Heavisine signal (*bottom*). The plots show the typical features of a wavelet estimator, namely the adaptation to local smoothness, where the sinusoidal parts of the signal are smoothed, and the discontinuities, are kept without over smoothing

ance σ^2, that needs to be determined to implement estimation. To see a typical realization of such a signal, see Fig. 1. This signal, 'Heavisine' is famous as one of the four typical signals that were introduced by [16]. Heavisine has been immersed in noise, see Fig. 1. This signal is not trivial to smooth as it possesses varying degrees of smoothness across the realization, compare for example the discontinuities with the smooth portions of the signal.

One important inference problem in this setting is to estimate $\{\mu(n\Delta t)\}$ with low expected square deviation, i. e. to chose an estimator $\widehat{\mu}(n\Delta t)$ such that

$$\mathrm{MSE}(\widehat{\mu},\mu) = \sum_{n=0}^{N-1} \mathrm{E}\left\{\left(\mu(n\Delta t) - \widehat{\mu}(n\Delta t)\right)^2\right\}, \qquad (11)$$

is as small as possible. Common methods of estimating $\mu(n\Delta t)$ include using many different adaptive linear estimation methods, such as kernel smoothers, smoothing splines, truncated Fourier methods etc. Smoothing using the DWT was introduced by [15,16], and has enjoyed an incredible success. The simplest form of wavelet estimation is based on the compression of the DWT of $x(t)$ alone, combined with the uniformity of the white process across time.

The wavelet coefficients of $\mu(\cdot)$ are given as $\left\{w_{j,l}^{(\mu)}\right\}$, and together with the scaling coefficients, $\left\{v_{j,l}^{(\mu)}\right\}$, these are sufficient to reconstruct $\mu(n\Delta t)$ via calculating the Inverse DFT (IDFT) or:

$$\begin{pmatrix} \mu(0\Delta t) \\ \dots \\ \mu((N-1)\Delta t) \end{pmatrix} = \mathcal{W}^{-1} \begin{pmatrix} w_{1,0}^{(\mu)} \\ \dots \\ w_{1,\frac{N}{2}-1}^{(\mu)} \\ w_{2,0}^{(\mu)} \\ \dots \\ w_{2,\frac{N}{4}-1}^{(\mu)} \\ \dots \\ v_{J,0} \end{pmatrix}. \qquad (12)$$

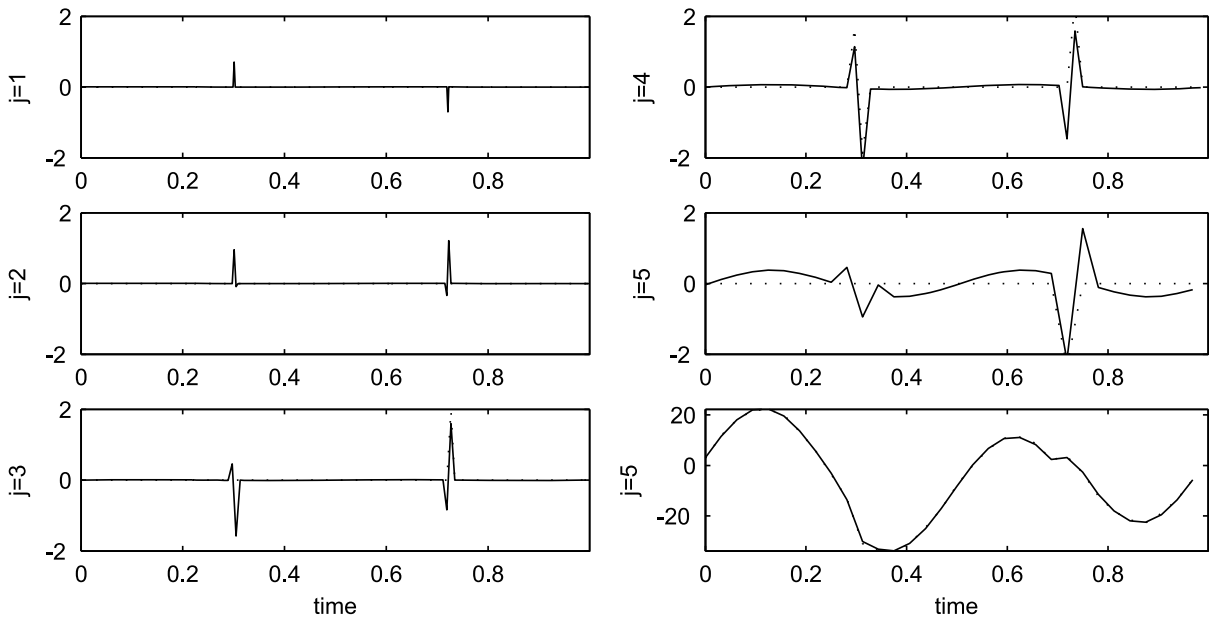

Statistical Applications of Wavelets, Figure 2
The wavelet coefficients of the Heavisine signal, used in the original paper by Donoho and Johnstone [15]. We show the first five level wavelet coefficients, and the scaling coefficients of the signal in *solid line*. Added to the picture are the thresholded wavelet and scaling coefficients of a noisy realization in *dotted line*

Thus a good estimator of $\left\{w_{j,l}^{(\mu)}\right\}$ can easily be converted into a good estimator of $\mu(n\Delta t)$, using Eq. (12). To see the motivation behind the most commonly adopted choice of estimation for the wavelet coefficients, observe the wavelet coefficients of the Heavisine signal, that we plotted in Fig. 1, plotted in Fig. 3. Most of the wavelet coefficients are quite small. If a given signal $\mu(\cdot)$ is sufficiently regular then it can be shown that $\left\{w_{j,l}^{(\mu)}\right\}$ must decrease in magnitude. For this reason one may argue that most of this set are zero, and the transform exhibits compression or sparsity, see [56]. We are therefore trying to estimate a sequence of variables of which most are zero, and only a few are large. It is *not* optimal to equate the estimator to the sample DWT, but instead shrinkage estimation is suitable. Consider the generic setting of an estimator for w denoted by $\widehat{w}$ which is unbiased, but potentially has large variance. A shrinkage estimator of this quantity is given by $\widehat{w}^{(s)} = c\widehat{w}$ for some $0 \le c \le 1$. Clearly this estimator has smaller variance as compared to $\widehat{w}$ (or possibly equal variance), but introduces some bias into the estimation. The mean square error is in fact given by:

$$\mathrm{MSE}(\widehat{w}, w) = c^2\sigma^2 + (1-c)^2 w^2 \,. \tag{13}$$

Thus if the quantity we are trying to estimate is small compared to the variance of the empirical estimator $\widehat{w}$ then by choosing $|c|$ smaller than one, we reduce the mean square error in estimation. Thus for $|w|$ taking a range of values it would be better to use the shrinkage estimator. This may seem non-intuitive, but corresponds to a whole set of theory derived due to [53]. If there was an oracle that could for a given shrinkage rule inform the estimator how coefficients should be shrunk, then good estimators would be obtained. Fortunately it is not necessary to have an oracle, for large sample sizes, to expect to get an estimator with equivalent performance to using an oracle. Looking at Fig. 2 and Fig. 3 we certainly see that replacing most of the empirical wavelet coefficients by zero, should greatly improve the estimation of these coefficients.

Two special shrinkage functions are commonly used, namely the *soft* and *hard* threshold functions. The soft threshold estimator modifies the empirical estimator $\widehat{w}$ by

$$\widehat{w}_\lambda^{(st)} = \begin{cases} \widehat{w} - \lambda & \text{if} \quad \widehat{w} \ge \lambda \\ 0 & \text{if} \quad |\widehat{w}| < \lambda \\ \widehat{w} + \lambda & \text{if} \quad \widehat{w} \le -\lambda \end{cases} \tag{14}$$

i. e. if the observed empirical estimate is sufficiently large a set amount is removed from its magnitude, whilst if it is not sufficiently large, the estimator is set to zero. Also

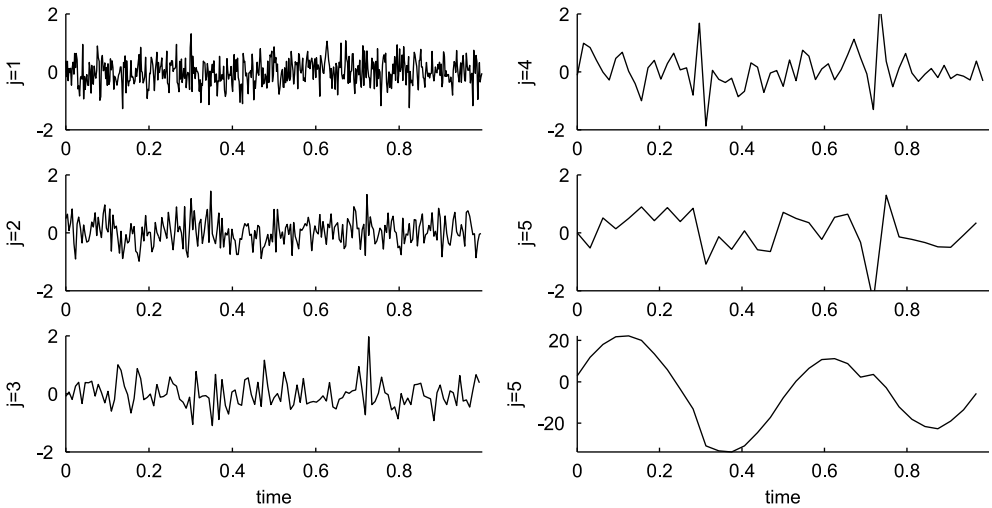

Statistical Applications of Wavelets, Figure 3
The noisy wavelet coefficients of the Heavisine signal, used in the original paper by Donoho and Johnstone. The decomposition has been stopped at level five, and then the scaling coefficients have been calculated to complete the representation

popular is the hard threshold estimator, defined by:

$$\widehat{w}_{\lambda}^{(ht)} = \begin{cases} \widehat{w} & \text{if} \quad \widehat{w} \geq \lambda \\ 0 & \text{if} \quad |\widehat{w}| < \lambda \\ \widehat{w} & \text{if} \quad \widehat{w} \leq -\lambda \end{cases} . \tag{15}$$

Most wavelet coefficients of sufficiently regular functions (i. e. functions in a suitable Besov space) will be very small, or even zero. One would greatly reduce the mean square error of the estimated mean function if one estimated those by zero, rather than their observed value. Thus we define the "oracle risk" in estimation as the mean square error which arises from using the best keep or kill procedure on the empirical wavelet coefficients, denoted by R_N:

$$R_N = \sum_{n=0}^{N-1} \min\left(w_n^2, \sigma^2\right) \tag{16}$$

If instead the estimator $\widehat{w}_{\sigma\lambda_N}^{(ht)}$ for $\lambda_N = \sigma\sqrt{2\log(N)}$ (the *universal threshold*) is used, then Donoho and Johnstone showed that the mean square error of this procedure is close to the oracle risk, see [56]. Alternative methods chose a value of λ that is permitted to vary with the value of j. σ is estimated using a robust scale estimator on the first level wavelet coefficients, such as the median absolute deviation estimator.

If there is only noise present, then using the universal threshold with a hard thresholding procedure will ensure that noise only signals are estimated as zero, see [56]. [18] give a more flexible definition of a universal thresholding procedure in terms of the probability of keeping the largest noise coefficient.

The simplest wavelet estimation method, suitable for regularly spaced, uncorrelated and Gaussian data has subsequently been extended to deal with irregularly spaced data, that is contaminated with correlated noise, and is immersed in non-Gaussian perturbations. A whole class of methods arise from transforming non-Gaussian observations, using say the Anscombe transform. Of particular note is the Haar–Fisz transform, introduced by [21]. The wavelet Haar–Fisz transform is defined by:

$$f_j^k = \begin{cases} 0 & \text{if} \quad v_{jk} = 0 \\ \frac{w_{jk}}{\sqrt{v_{jk}}} & \text{if} \quad v_{jk} \neq 0 \end{cases} , \tag{17}$$

and the distribution of this variable is substantially more Gaussian than the original wavelet coefficients. Fryzlewicz and Nason have established the properties of using this transform in signal estimation in a series of papers, see [21,22,23]. The denoising of correlated data was treated by [30].

Many different extensions also exist to the basic procedure for white Gaussian random variables: usually the DWT coefficients exhibit some form of structured behavior, and either clustering across time (l) or persistence across scale (j) is observed in the magnitude of the coefficients. Various methods have been proposed to use these observed characteristics, such as treating whole blocks of coefficients together, see [7], using Bayesian methods where the dependence of the coefficients is modeled in the

prior and other methods, such as wavelet footprints due to [19], and 'Analytic' denoising proposed by Olhede and Walden and later extended to two dimensions, see [43] and [42]. To observe the clustered structure of wavelet coefficients, see Fig. 2. The discontinuity in the considered signal causes wavelet coefficients to be non-zero near the discontinuities in time across all scales, and across scales at the given time location of the discontinuity. To represent the discontinuities perfectly, and avoid wavelet Gibb's effects, all of these coefficients should be kept, thus estimating a single discontinuity without artifacts. Similar statements can be made for other features than perfect discontinuities, and the above mentioned methods all address this problem in a chosen manner.

A key characteristic of observed coefficients is the sparsity level, i. e. how many of the coefficients that are approximately zero in expectation. Choosing the threshold λ should be done in light of the sparsity of the data. Optimally such sparsity is learned from the data. [29] proposed the EBAYES procedure whereby at each scale the sparsity of the data is estimated and a threshold chosen using marginal likelihood.

2-D Extensions of Wavelets

The development of statistical estimation methods for 1-D signals observed in noise is a mature field. Greater challenges lie in higher dimensional objects such as spatial, or spatiotemporal signals. In the early days when methods for extended into 2-D, the two spatial directions were treated as if the observed phenomena were two different phenomena. This caused a number of unpalatable effects in the reconstruction of the estimated images, such as edge effects and blocking, as well as using too many coefficients to reproduce simple phenomena. To see an example of this, observe Fig. 4. The woman is mainly a smooth picture, and should only correspond to a limited set of coefficients in a compressed representation. Using separable wavelets to represent the picture, because the curve of the woman's nose is misaligned with the decomposition, causes too many coefficients to be used. This lack of compression, leads to poor performance in the estimation.

The solution to these problems came by the development of non-separable decompositions, this leading to fewer non-zero coefficients in the decompositions, or by statistical processing of the separable decompositions, using the combined magnitude of the many coefficients to improve estimation. The main methods of note are ridgelets and curvelets proposed by [9,52], the dual-tree complex wavelet transform, proposed by Kingsbury (see [48]), and bandelets (see [33]). The implementation of several of these transforms is slightly more complicated than the 1-D wavelet transform, but all correspond to linear operations on the observed signal. This means that the statistical properties of the transform coefficients can be determined by direct (if sometimes numerical) calculations. Again, thresholding and inverse transformation forms the basis for most estimation procedures, much in the same manner as the 1-D estimation procedures.

Open areas of research in this general setting, is the development of higher dimensional processing, and suitable transformations for discrete analysis on the sphere.

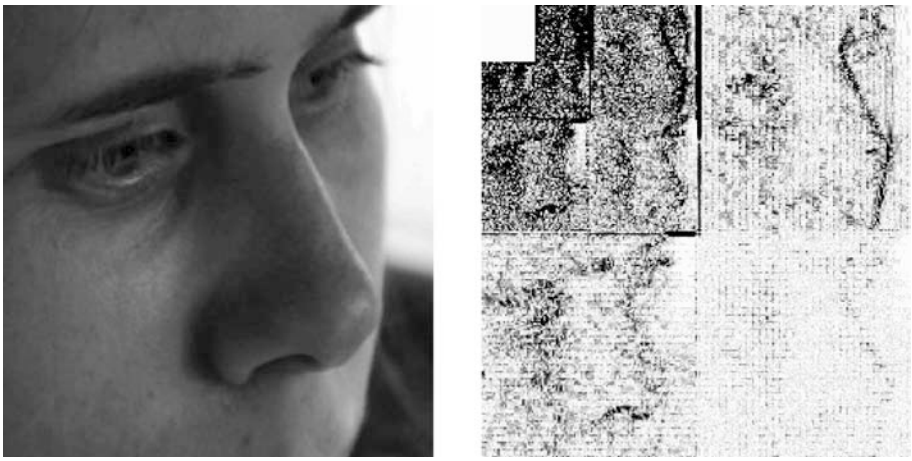

Statistical Applications of Wavelets, Figure 4
A picture of a woman (*left*), and the absolute value of the separable 2-D wavelet coefficients of the picture (*right*). Observe that the nose of the woman takes too many coefficients to record, despite the curve of the nose corresponding to a smooth curve. This illustrates the lack of compression of the 2-D separable wavelet transform

Covariance Estimation Using Wavelets

Wavelets have been used to estimate properties of zero-mean stochastic processes, characterized by their dependence structure alone. One of the fundamental aspects is noting classes of stochastic processes for which the transform provides a simplification of the description of the signal. One set of developments has defined a set of locally stationary processes using wavelet filters as the building block, so called Locally Stationary Wavelet Processes, see [41]. The estimation of the fundamental representation of this process, which is very much like a time varying spectrum, was addressed by Nason and his coauthors, and has subsequently been improved upon by Fryzlewicz and Nason, see [23]. The model has been used to classify sleep states of babies, and has been used in modeling financial times series, see [5,20]. This corresponds to the most natural modification of existing theory for locally stationary time series from local Fourier to wavelet bases.

The usage of wavelets for estimating time series that when differenced are stationary has also seen substantial investigation. By using wavelets the number of differencing steps needs often not be determined, which is a clear advantage. Likelihood theory has been extended from the Fourier domain to the wavelet domain for such time series, see [39], and various procedures for estimating the decay of long memory processes have been adapted for wavelet coefficients, see [38]. Characterization of sets of time series that are jointly stationary when differenced, has also been developed, see [49,50,51,57]. A fundamental method of analysis is to characterize the covariances of the process associated with different time-scales, this yielding a multiresolution analysis of the stochastic process, both considering the auto-covariance of a single process, and cross-covariances between several processes.

Future Directions

The usage of the wavelet transform in 1-D is a mature field. Current developments are mainly in the application and adaption of existing methods to novel applications. The development of higher dimensional transforms has also slowed down considerably. The modern thrust and drive is instead found in using several concepts which the development of analysis methods from the wavelet transform were founded upon. The key two notions for these developments are sparsity and incoherence. For many signals of interest, their wavelet coefficient representation was compressed. This means the sparsity could be used to substantially improve the estimation procedure. The area of 'compressed sensing' is based on using the compression of a signal in a known decomposition, to improve the estimation of the signal, see [14]. Surprisingly the limitations of Nyquist sampling has been circumvented, and near miraculous reconstruction performance obtained. As the data deluge that faces modern signal processing will increase, this stands as a very interesting area of future development.

Bibliography

1. Antoniadis A, Paparoditis E, Sapatinas T (2006) A functional wavelet-kernel approach for time series prediction. J Roy Stat Soc B 68:837–857
2. Antoniadis A, Sapatinas T (2003) Wavelet methods for continuous-time prediction using Hilbert-valued autoregressive processes. J Multivar Anal 87:133–158
3. Averkamp R, Houdre C (1998) Some distributional properties of the continuous wavelet transform of random processes. IEEE Trans Info Theory 44:1111–1124
4. Averkamp R, Houdre C (2000) A note on the discrete wavelet transform of second-order processes. IEEE Trans Info Theory 46:1673–1676
5. Bellegem SV, Fryzlewicz P, von Sachs R (2003) A wavelet-based model for forecasting non-stationary processes. Inst Phys Conf Ser 173:955–958
6. Buresti G, Lombardi G, Bellazzini J (2004) On the analysis of fluctuating velocity signals through methods based on the wavelet and hilbert transforms. Chaos Solitons Fractals 20:149–158
7. Cai T, Silverman BW (2001) Incorporating information on neighbouring coefficients into wavelet estimation. Sankhyā Ser B 63:127–148
8. Cambanis S, Houdre C (1995) On the continuous wavelet transform of second-order random processes. IEEE Trans Info Theory 41:628–642
9. Candès E, Donoho DL (1999) Ridgelets: a key to higher-dimensional intermittency? Philos Trans: Math Phys Eng Sci 357:2495–2509
10. Daubechies I (1988) Orthonormal bases of compactly supported wavelets. Comm Pure Appl Math 41:909–996
11. Daubechies I (1992) Ten lectures on wavelets. SIAM, Philadelphia
12. Daubechies I, Sweldens W (1998) Factoring wavelet transforms into lifting steps. J Fourier Anal Appl 4:247–269
13. Delprat N, Escudié B, Guillemain P, Kronland-Martinet R, Tchamitchian P, Torresani B (1992) Asymptotic wavelet and gabor analysis: extraction of instantaneous frequencies. IEEE Trans Inform Theory 38:644–64
14. Donoho DL (2006) Compressed sensing. IEEE Trans Inf Theory 52:1289–1306
15. Donoho DL, Johnstone IM (1994) Ideal spatial adaptation by wavelet shrinkage. Biometrika 81:425–455
16. Donoho DL, Johnstone IM (1995) Adapting to unknown smoothness via wavelet shrinkage. J Am Stat Assoc 90:1200–1224
17. Donoho DL, Johnstone IM, Kerkyacharian G, Picard D (1995) Wavelet shrinkage – asymptotia. J Roy Stat Soc B 57:301–337
18. Downie TR, Silverman BW (1998) The discrete multiple wavelet transform and threshold methods. IEEE Trans Signal Process 46:2558–2561

19. Dragotti PL, Vetterli M (2003) Wavelet footprints: theory, algorithms, and applications. IEEE Trans Signal Process 51:1306–1323
20. Fryzlewicz P, Bellegem SV, von Sachs R (2003) Forecasting non-stationary time series by wavelet process modelling. Ann Inst Stat Math 55:737–764
21. Fryzlewicz P, Nason GP (2004a) A Haar–Fisz algorithm for poisson intensity estimation. J Comp Graph Stat 13:621–638
22. Fryzlewicz P, Nason GP (2004b) A Haar–Fisz algorithm for poisson intensity estimation. J Comp Graph Stat 13:621–638
23. Fryzlewicz P, Nason GP (2006) Haar–Fisz estimation of evolutionary wavelet spectra. J Roy Stats Soc B 68:611–634
24. Grossman A, Morlet J (1984) Decomposition of Hardy functions in square integrable wavelets of constant shape. Siam J Math Anal 15:723–736
25. Guérin C-A (2000) Wavelet analysis and covariance structure of some classes of non-stationary processes. J Fourier Anal Appl 6:403–425
26. Haar A (1910) On the theory of orthogonal function systems. Math Ann 69:331–371
27. Holschneider M (1995) Wavelets: an analysis tool. Oxford University Press, Oxford
28. Jaffard S, Melot S (2005) Wavelet analysis of fractal boundaries. Commun Math Phys 258:513–159
29. Johnstone I, Silverman BW (2005) Empirical Bayes selection of wavelet thresholds. Ann Stat 33:1700–1752
30. Johnstone IM, Silverman BW (1997) Wavelet threshold estimators for data with correlated noise. J Roy Stat Soc Ser B 59:319–351
31. Kim IK, Kim YY (2005) Damage size estimation by the continuous wavelet ridge analysis of dispersive bending waves in a beam. J Sound Vib 287:707–722
32. Krim H, Pesquet J-C (1995) Multiresolution analysis of a class of nonstationary processes. IEEE Trans Inf Theory 41:1010–1020
33. le Pennec E, Mallat S (2005) Sparse geometric image representations. IEEE Trans Image Process 14:423–438
34. Li T-H, Oh H-S (2002) Wavelet spectrum and its characterization property for random processes. IEEE Trans Inf Theory 48:2922–2937
35. Lilly JM, Gascard J-C (2006) Wavelet ridge diagnosis of time-varying elliptical signals with application to an oceanic eddy. Nonlinear Process Geophys 13:467–483
36. Loève M (1945) Sur le fonctions aléatoire de second ordre. Rev Sci 83:297–303
37. Mallat S (1999) A Wavelet Tour of Signal Processing, 2nd edn. Academic Press, New York
38. McCoy EJ, Walden AT (1996) Wavelet analysis and synthesis of stationary long-memory processes. J Comp Graph Stat 5:26–56
39. Moulines E, Roueff F, Taqqu MS (2007) On the spectral density of the wavelet coefficients of long-memory time series with application to the log-regression estimation of the memory parameter. J Time Ser Anal 28:155–187
40. Nason GP, von Sachs R (1999) Wavelets in time-series analysis. Philos Trans Math Phys Eng Sci 357:2511–2526
41. Nason GP, von Sachs R, Kroisandt G (2000) Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. J Roy Stat Soc B 62:271–292
42. Olhede S (2007) Hyperanalytic denoising. IEEE Trans Image Process 16:1522–1537
43. Olhede S, Walden AT (2004a) 'Analytic' denoising. Biometrika 91:955–973
44. Olhede SC, Walden AT (2004b) The Hilbert spectrum via wavelet projections. Proc Roy Soc Lond A 460:955–975
45. Olhede SC, Walden AT (2005) Wavelet denoising for signals in quadrature. Integr Computer-Aided Eng 12:109–117
46. Percival DB, Walden AT (2000) Wavelet methods for time series analysis. Cambridge University Press, Cambridge
47. Schuster A (1898) On the investigation of hidden periodicities with application to a supposed 26-day period of meteorological phenomena. Terr Mag Atmos Elect 3:13–41
48. Selesnick IW, Baraniuk RG, Kingsbury NG (2005) The dual-tree complex wavelet transform. IEEE Signal Process Mag 22(6):123–151
49. Serroukh A, Walden AT (2000a) Wavelet scale analysis of bivariate time series i: motivation and estimation. J Nonparametr Stat 13:1–36
50. Serroukh A, Walden AT (2000b) Wavelet scale analysis of bivariate time series ii: statistical properties for linear processes. J Nonparametr Stat 13:37–56
51. Serroukh A, Walden AT, Percival DB (2000) Statistical properties and uses of the wavelet variance estimator for the scale analysis of time series. J Am Stat Assoc 95:184–196
52. Starck JL, Candès EJ, Donoho DL (2002) The curvelet transform for image denoising. IEEE Trans Image Process 11:670–84
53. Stein C (1981) Estimation of the mean of a multivariate normal distribution. Ann Stat 9:1135–1151
54. Torrence C, Compo GP (1998) A practical guide to wavelet analysis. Bull Am Meteorol Soc 79:61–78
55. Tu C-L, Hwang W-L, Ho J (2005) Analysis of singularities from modulus maxima of complex wavelets. IEEE Trans Inf Theory 51:1049–1062
56. Wasserman L (2006) All of nonparametric statistics. Springer, New York
57. Whitcher B, Guttorp P, Percival DB (2000) Wavelet analysis of covariance with application to atmospheric time series. J Geophys Res Atmos 105(14):14941–14962
58. Wolfe PJ, Godsill SJ, Ng WJ (2004) Bayesian variable selection and regularisation for time-frequency surface estimation. J Roy Stat Soc Ser B 66:575–589
59. Zhang Z, Ren Z, Huang W (2003) A novel detection method of motor broken rotor bars based on wavelet ridge. IEEE Trans Energy Convers 18:417–423

Statistical and Non-linear Physics, Introduction to

M. Cristina Marchetti
Physics Department, Syracuse University, Syracuse, USA

The field of Statistical and Nonlinear Physics combines the venerable subject of statistical mechanics with the newer area of nonlinear science to create a highly interdisciplinary and exciting area of research. It has its roots in equilibrium statistical physics, but it has evolved to encompass and emphasize nonequilibrium and dynamical phenomena. It is a field in rapid evolution, with a constantly changing focus. It overlaps naturally with many

other disciplines, including fluid dynamics, computational physics, biological physics, condensed matter physics and polymer physics.

Statistical physics aims to describe the large-scale collective behavior of systems composed of a large number of interacting units or degrees of freedom. Starting with a microscopic model, one performs various types of "coarse-graining" to describe phenomena that occur on length and time scales large compared to microscopic ones, such as the size of the particles or the characteristic interaction times. At this large scale the system exhibits collective or emergent behavior that is far richer than that of the individual units. One of the hallmarks of collective behavior is phase transitions. An everyday example is the change from liquid to solid that occurs upon tuning a parameter such as temperature or pressure. Another familiar example arises in the study of the properties of magnetic materials and is highlighted in the article ► Complex Systems and Emergent Phenomena, which is a good starting point for the reader of this section.

When going from the deterministic Hamiltonian description of a system of many interacting units to the large scale coarse-grained description of the same system in terms of a continuum or hydrodynamic theory, an element of randomness is naturally introduced into the problem as one looses the invariance under time reversal that was present in the Hamiltonian description. At the same time randomness on the other hand is an intrinsic property of the dynamics of individual nonlinear systems that are known to often exhibit sensitive dependence to initial conditions and chaotic behavior. There is in fact a deep connection between the randomness of chaotic systems and the irreversible transport properties of extended systems (see ► Chaotic Dynamics in Nonequilibrium Statistical Mechanics). This connection highlights the unity of the two areas of statistical and nonlinear science.

Rather than presenting an exhaustive review of the field of statistical and nonlinear physics, the articles in this section focus on nonequilibrium phenomena that are the subject of current research. Even then, only a fraction of the problems and physical systems that are studied using the tools and ideas of statistical and nonlinear physics are described here. Many of the others can be found in articles throughout the rest of the Encyclopedia.

The topics highlighted in this section of the Encyclopedia may at a first glance seem disparate, but are unified by the ideas that are used to study them. Principal among those are the notions of scaling and universality. The concept of universality, which has been around for some time, has its roots in the study of phase transitions and critical phenomena. Near a phase transition, the system is universal in that its behavior at large scales does not depend on the microscopic physics. The systems and physical problems described in this section of the Encyclopedia may be different in the details, but generally involve the onset of collective or emergent behavior with "universal" properties.

The ideas of scaling universality have had a crucial impact in the study of long-standing problems in nonlinear physics, such as pattern formation and turbulence (see ► Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence). The same tools and ideas have more recently been applied to the study of the complex dynamics of neuronal systems (see ► Neuronal Dynamics), lasers (see ► Noise and Stability in Modelocked Soliton Lasers), and even the physics of economics (see ► Econophysics, Statistical Mechanics Approach to). Rich chaotic behavior has also been discovered in quantum systems (see ► Quantum Chaos).

Recently, a new field has emerged that uses many of the ideas of statistical mechanics to explore problems in "soft" condensed matter physics. Soft materials are literally materials that yield easily to mechanical deformations, They encompass an extremely wide range of systems, from complex fluids such as colloidal suspensions, liquid crystals and polymers (see ► Polymer Physics) to soft elastic gels (see ► Anisotropic Networks, Elastomers and Gels), and granular matter in both its fluidized (see ► Granular Flows) and its jammed (see ► Jamming of Granular Matter) states. The jamming or structural arrest of granular matter when the density reaches a critical value bears striking similarities with the glass transition. Jammed granular matter and glassy systems share many of the intriguing dynamical properties discussed in ► Glasses and Aging, A Statistical Mechanics Perspective on. Jamming also provides an example of the nonequilibrium phase transitions that are ubiquitous in the systems studied by modern statistical physics. Nonequilibrium phase transitions do occur in "pure" systems, i. e., ones without any extrinsic disorder, as a result of the dynamical constraints arising from interactions, but are even more common in systems with quenched disorder, induced for instance by impurities and materials defects (see ► Disordered Elastic Media for a general perspective on such systems). Notable examples include the depinning transition that occurs in driven extended systems such as vortex lattices in type-II superconductors under the action of a uniform external driving force (see ► Collective Transport and Depinning) and the slip-stick motion that characterizes the physics of friction and tectonic motion (see ► Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of). Living matter, such as the cytoskeleton of eukaryotic cell or

even the whole cell, has also been described as a novel form of nonequilibrium soft matter (see ► Cytoskeleton and Cell Motility).

The complexity of the systems that are studied by statistical physicists has naturally led to the development of sophisticated numerical methods. Principal among those is the technique of Monte Carlo simulations (see ► Monte Carlo Simulations in Statistical Physics). More recently physicists have also begun to adapt ideas and algorithms from computer science (see ► Optimization Problems and Algorithms from Computer Science).

Novel experimental techniques have recently allowed us to begin to study the properties of soft and living matter on length and time scales that require a qualitative new way of thinking and a modification of the familiar tools from statistical physics (see ► Fluctuation Theorems, Brownian Motors and Thermodynamics of Small Systems). These range from optical techniques for study of the properties of individual biomolecules (see ► Protein Mechanics at the Single-Molecule Level) to microfluidics devices for the study of the rheological and mechanical properties of complex materials and cells on novel time and length scales (see ► Microfluidics).

Yet another class of problems that has caught the attention of statistical physicists lately is the study of networks (see ► Networks: Structure and Dynamics). Examples of networks are as disparate as the world-wide web, human societies, languages and their structure, and power grids. An especially important class of networks occurs in biology. It consists of the cell signaling networks that control many of the functions of living cells (see ► Cell Biology: Networks, Regulation and Pathways).

Finally, complex emergent behavior is also ubiquitous in quantum matter, where the interactions and fluctuations are controlled by quantum mechanics, as described in the article ► Ultracold Atomic Gases: Novel States of Matter.

Statistics with Imprecise Data

María Ángeles Gil[1], Olgierd Hryniewicz[2]
[1] Faculty of Sciences, University of Oviedo, Oviedo, Spain
[2] Systems Research Institute, Warsaw, Poland

Article Outline

Glossary
Definition of the Subject
Introduction
Mathematical Modeling of Imprecise Data
Fuzzy Random Variables
Statistical Analysis of Fuzzy Data Corresponding to Fuzzy Perceptions of Existing Real–Valued Data
Statistical Analysis of Existing Fuzzy-Valued Data
Future Directions
Bibliography

Glossary

Fuzzy estimators Estimators of "parameters" of probability distributions, or other characteristics, of random variables/fuzzy random variables (such as e. g. the expected value/the fuzzy expected value) when statistical data are imprecise and are described by means of fuzzy sets.

Fuzzy random variable Random element whose observed values are described by fuzzy sets.

Fuzzy set Generalization of a classical notion of a set. In contrast to the case of a classical set, each element x of a fuzzy set may belong to it to a degree described by the so-called membership function $\mu(x)$. Thus, the fuzzy set may be defined as a set of ordered pairs $(x, \mu(x))$, where x belongs to a set $\mathbb{X}$ called the universe of discourse or referential. Alternatively, the fuzzy set can be identified with its membership function (in the same way that a classical set can be identified with its indicator function).

Fuzzy statistical tests Statistical tests used for the verification of hypotheses about the values of "parameters" of probability distributions, or other characteristics, of random variables/fuzzy random variables when statistical data are imprecise and are described by fuzzy sets.

Fuzzy statistics Generalization of traditional statistics that allows to handle imprecise data described in terms of fuzzy sets.

Imprecise data Data that cannot be described by either real numbers or vectors with real-valued components.

Random sets Random elements whose observed values are sets (e. g. intervals, subsets of a plane).

Definition of the Subject

Traditional statistics deals usually with precisely defined data. The source of these data may be of different nature. Usually the data are collected from observations of random experiments, i. e. experiments whose performance leads to an outcome which cannot be predicted in advance with one hundred percent sureness. When the knowledge about the nature of these random events is available we can use the methods of mathematical statistics and draw conclusions about the source of available data. Uncertainty being intrinsic to random outcomes/events is properly de-

scribed by using the formalism of the mathematical theory of probability, and generally it is attributed to *future* observations of these outcomes/events.

Traditionally, statistical experimental data are described by real numbers or by vectors whose components are real numbers. These numbers are either observed directly as results of measurements (e. g. height and weight of a person) or correspond to observed counts of certain categories representing labeled events (e. g. a gender of that person). However, in real life applications the results of measurements are never precise. The precision of every measurement is limited by the precision of a measuring device. In the majority of practical applications this lack of precision may be neglected, and for statistical analysis of available data we can use traditional methods of statistics. If the measurement error cannot be neglected statistical analysis of interval data is recommended by specialists in metrology. However, when statistical data are presented by human beings we need more sophisticated methods for the description of their lack of precision. Therefore, there is a practical need to generalize traditional statistical methods in order to make them applicable for handling imprecise data, e. g. the data described by statements expressed by using a plain language (the so-called "linguistic data").

Introduction

There is a common agreement that uncertainty characterized by randomness shall be described by considering the theory of probability. Thus, mathematical statistics is a proper tool for dealing with data generated by random experiments and described by precisely defined numbers. However, there also exist other types of uncertainty which are related to vagueness, imprecision, existence of only partial information about experimental outcomes/events of interest, etc. In contrast to randomness, uncertainties of such types are attributed rather to *current* perceptions/observations. It has to be noted that the mathematical modeling of all these types of uncertainty which are different from simple randomness is still the subject of controversies. A good overview of the related problems can be found in the paper by P. Walley [36]. In this article we confine ourselves to the case when randomness is observed together with vagueness understood as the lack of precision.

Specialists in measurement theory recognize different types of uncertainty. For instance, in the ISO/IEC Guide [18] it is recommended to distinguish between uncertainty related to pure randomness and uncertainty of other nature, such as a lack of precision. However, the lack of precision of statistical data is usually omitted in the statistical analysis of measurements. Consider, for example, the analysis of results of measurements coming from a digital meter. The results of measurements coming from such a meter are always rounded in order to have their representation by a certain number of digits. When we observe a displayed result of a measurement we never know what is the actual value of the measured quantity. What is more important, and often overlooked by statisticians, we may not increase our knowledge about that value by averaging the results of repeated measurements, as it is frequently recommended by statisticians. Consider, for example, a series of measurements showing exactly the same result. Having such results we are in principle not able to distinguish between two fundamentally different cases when the measured quantity is constant and its value belongs to the range of rounding or when it is varying from measurement to measurement with values belonging to that range.

The situation described in the previous paragraph is relatively simple as the range of possible actual values corresponding to the observed result of a measurement is usually precisely defined. There exist, however, situations when either this range is not precisely known or values in the range are not equally compatible with the available data/information/events. We face such cases when results of measurements or classifications are evaluated by humans (e. g., by evaluating indications of an analog meter or classifying individuals in accordance with their height, and so on). The result of a measurement is even more imprecise when it has been obtained without the usage of any meter (e. g., when we visually evaluate the distance between two points); in such situations statistical data may consist of imprecise statements like "around 5 meters", "more or less between 5 and 10 seconds", etc. The classification of a person in accordance with his/her height leads often to imprecise assessments, like "very short", "rather tall", etc. A similar situation occurs when we deal with retrospective data recalled by human beings; for example, in reliability analysis of field lifetime data we may face situations when failure times are reported imprecisely using statements like 'the failure occurred about one month ago', etc. In all these cases statistical data consist of *imprecise perceptions* of actual real values.

In the previous paragraphs we considered situations when actual values of measured quantities exist, but they are imprecisely perceived. There exist, however, situations when we have to analyze statistical data that represent imprecisely defined concepts. Take for example the color of human hair described using categories such as "blond" or "dark blond"; it is obvious that the border between these two categories is vague; one can try to establish a precise

border between these two categories in terms of the results of precise measurements of the spectrum of reflected light, but such attempts seem to be practically senseless. We face a similar situation in classifying a client of a bank office in accordance with his/her degree of aversion to investment which leads to imprecise assessments, like "very low", "moderate", etc. In both these cases we could try to collect precise statistical data, e. g., by either asking a respondent to a questionnaire to indicate exactly one choice or coding it numerically. However, it seems to be more prudent, natural and informative to expect imprecise answers to questions pertained to vague notions. If we do so, we may face imprecise statistical data for further analysis. In all these cases statistical data consist of *imprecise actual values* themselves.

There also exists another source of imprecision while dealing with statistical data. For example, in reliability lifetime tests we perform tests in more severe "over-stress" conditions, and then we try to recalculate test results to conditions which are considered "normal". For this recalculation we can utilize some partial knowledge about possible values of stress-dependent recalculation coefficients. In such a case we have originally precise lifetime data, but after recalculation these data become imprecise.

In all considered cases the lack of precision can be appropriately described by using fuzzy sets introduced by Lotfi A Zadeh. In the second section of the article we recall some basic definitions related to fuzzy sets. We will present the fuzzy sets methodology as a useful tool for the description of imprecise data. When fuzzy lack of precision is mixed with randomness, either in the sense that available fuzzy data are supposed to come from the perception of real- or vectorial-valued data generated by a random mechanism, or in the sense of they being directly generated by a random mechanism, a convenient tool to use is that of the notion of a fuzzy random variable. This notion is introduced in the third section of this article. The interpretation of the fuzzy random variable depends upon the type of observed fuzzy data and events. In the paper we distinguish between the two types of data which have been described in this section. We start with the description of statistical methods which are useful for the analysis of fuzzy perceptions of existing precise values. Then, we present statistical methods which are useful in the analysis of intrinsically fuzzy-valued data.

Mathematical Modeling of Imprecise Data

There exist competitive methods for the description of vagueness. For example, some statisticians claim that the theory of subjective probability is sufficient for the description of all types of uncertainty. However, many other researchers have shown examples of situations when the application of the classical theory of probability is not sufficient for modeling these situations. Therefore, other formalisms have been proposed for the description of vagueness/imprecision. One of those formalisms, namely the theory of *fuzzy sets* proposed by Lotfi A. Zadeh [37], has been slowly but widely accepted as a good methodology for the description of imprecise data, both from practical and theoretical (see, e. g., the paper by Terán [33]) points of view.

The basic concept of the theory of fuzzy sets is the universe of discourse or referential $\mathbb{X}$ which may be understood as the set of all possible (feasible) elements that are relevant for the description of a certain concept (quantity). Mathematically speaking a ***fuzzy subset*** A of a set $\mathbb{X} \neq \emptyset$ (or a fuzzy set, for short) is a map $A : \mathbb{X} \to [0, 1]$, where $A(x)$ can be interpreted as the degree of compatibility of x with the ill-defined property characterizing A, or degree of truth of the assertion "x is A", or degree of membership of x to A. Equivalent, but more intuitive for some purposes, a fuzzy set A of $\mathbb{X}$ can be defined as a set of ordered pairs $\{(x, \mu_A)\}$, where $x \in \mathbb{X}$ and $\mu_A : \mathbb{X} \to [0, 1]$ is the so-called *membership function* of A. In other words, a fuzzy set can be identified with its membership function, in the same way that a classical set can be identified with its indicator function. In what follows, we will consider indistinctly A or μ_A to denote and refer to a fuzzy subset.

Unfortunately, there is no one generally accepted methodology for the construction of membership functions. The majority of researchers assume that the membership function $\mu_A(x)$ is a purely subjective function provided by a person who describes his/her perception of a certain phenomenon or quantity. Some authors provide practical methods for the construction of the membership function when it is interpreted in terms of the theory of possibility as the possibility distribution (see [8] for more information). Some other authors, e.g Bandemer and Näther [1] or Viertl [35], present methods which may be used for the construction of membership functions in a more objective way from measurements of physical quantities. Anyway, our purpose is not entering a discussion here about this point.

In the analysis of imprecise data we are usually interested in the description of interesting phenomena by numbers. For this purpose we can use the concept of a fuzzy number defined as follows (see [5]):

Definition 1 The fuzzy subset A of the space of real numbers $\mathbb{R}$, with the membership function $\mu_A : \mathbb{R} \to [0, 1]$, is a **fuzzy number** if and only if

(a) A is normal, i. e. there exists at least an element $x_0 \in \mathbb{R}$ such that $\mu_A(x_0) = 1$;
(b) A is fuzzy convex, i. e., $\mu_A(\lambda x_1 + (1-\lambda)x_2) \geq \mu_A(x_1) \wedge \mu_A(x_2)$, for all $x_1, x_2 \in \mathbb{R}$, and $\lambda \in [0,1]$;
(c) μ_A is upper semicontinuous;
(d) the support set, $\operatorname{supp} A = \{x \in \mathbb{R} : \mu_A(x) > 0\}$, is bounded (that is, its closure $\operatorname{cl}(\operatorname{supp} A)$ is compact).

It is easily seen that if A is a fuzzy number then its membership function can be expressed as follows:

$$\mu_A(x) = \begin{cases} 0 & \text{for } x < a_1 \\ r_{l_A}(x) & \text{for } a_1 \leq x < a_2 \\ 1 & \text{for } a_2 \leq x \leq a_3 \\ r_{u_A}(x) & \text{for } a_3 < x \leq a_4 \\ 0 & \text{for } x > a_4\,, \end{cases} \tag{1}$$

where $a_1, a_2, a_3, a_4 \in \mathbb{R}$, $a_1 \leq a_2 \leq a_3 \leq a_4$, $r_{l_A} : [a_1, a_2] \to [0,1]$ is a nondecreasing upper semicontinuous function, and $r_{u_A} : [a_3, a_4] \to [0,1]$ is a nonincreasing upper semicontinuous function. Functions r_{l_A} and r_{u_A} are called sometimes the left and the right "arms" (or "sides") of the fuzzy number, respectively.

A useful notion for dealing with a fuzzy number is the so-called α-level set, also known as the *α-cut*. For $\alpha \in (0,1]$ the *α-level set* of the fuzzy number A is the ordinary (non-fuzzy) set defined as

$$A_\alpha = \{x \in \mathbb{R} : \mu_A(x) \geq \alpha\} \tag{2}$$

and the 0-level set is usually intended to be given by $A_0 = \operatorname{cl}(\operatorname{supp} A)$. The family $\{A_\alpha : \alpha \in [0,1]\}$ is a set representation of the fuzzy number A. According to the resolution identity proposed by Zadeh, we can represent the membership function as:

$$\mu_A(x) = \sup\{\alpha \cdot \mathbf{1}_{A_\alpha}(x) : \alpha \in (0,1]\}\,, \tag{3}$$

where $\mathbf{1}_{A_\alpha}(x)$ denotes the indicator function of A_α.

On the basis of the notion of α-level, a fuzzy number A can be viewed as a fuzzy subset of $\mathbb{R}$ such that its α-level sets are nonempty compact and convex sets of $\mathbb{R}$, that is, nonempty compact intervals. Hence, for each $\alpha \in [0,1]$ we have that $A_\alpha = [A_L(\alpha), A_U(\alpha)]$, where

$$\begin{aligned} A_L(\alpha) &= \inf\{x \in \mathbb{R} : \mu_A(x) \geq \alpha\}\,, \\ A_U(\alpha) &= \sup\{x \in \mathbb{R} : \mu_A(x) \geq \alpha\}\,. \end{aligned} \tag{4}$$

If the sides of the fuzzy number A are strictly monotonic functions, then from Eq. (1) one can see easily that $A_L(\alpha)$ and $A_U(\alpha)$ are inverse functions of r_{l_A} and r_{u_A}, respectively.

In statistical analysis of random data we use functions (and, in particular, operations) of the observed random samples. These functions, called statistics, can be also defined for fuzzy random data. Their membership functions can be derived by using Zadeh's *extension principle*. This principle has the following formulation [38]:

Let $\mathbb{X}$ be a Cartesian product of universes $\mathbb{X} = \mathbb{X}_1 \times \ldots \times \mathbb{X}_r$, and $A_1, \ldots, A_r$ be r fuzzy subsets of $\mathbb{X}_1, \ldots, \mathbb{X}_r$, respectively. Let f be a mapping from $\mathbb{X} = \mathbb{X}_1 \times \ldots \times \mathbb{X}_r$ to a universe $\mathbb{Y}$ such that $y = f(x_1, \ldots, x_r)$. The extension principle allows us to induce from the r fuzzy sets A_i a fuzzy set B on $\mathbb{Y}$ through f, $B = f(A_1, \ldots, A_r)$ such that

$$\mu_B(y) = \begin{cases} \sup_{x_1,\ldots,x_r | y = f(x_1,\ldots,x_r)} \\ \quad \min\{\mu_{A_1}(x_1), \ldots, \mu_{A_r}(x_r)\} & \text{if } f^{-1}(y) \neq \emptyset \\ 0 & \text{if } f^{-1}(y) = \emptyset \end{cases} \tag{5}$$

In case of A_i, $i = 1, \ldots, n$ being fuzzy numbers, and f being either an injective or a continuous function, then for each $\alpha \in [0,1]$ the α-level of $B = f(A_1, \ldots, A_r)$ can be shown to be equal to $B_\alpha = \big(f(A_1, \ldots, A_r)\big)_\alpha$ with

$$\begin{aligned} &\big(f(A_1, \ldots, A_r)\big)_L(\alpha) \\ &\quad = \min_{(x_1,\ldots,x_r) \in (A_1)_\alpha \times \ldots \times (A_r)_\alpha} f(x_1, \ldots, x_r)\,, \\ &\big(f(A_1, \ldots, A_r)\big)_U(\alpha) \\ &\quad = \max_{(x_1,\ldots,x_r) \in (A_1)_\alpha \times \ldots \times (A_r)_\alpha} f(x_1, \ldots, x_r)\,. \end{aligned} \tag{6}$$

Thus, the application of the extension principle for the calculation of the membership function of $y = f(x_1, \ldots, x_r)$ is equivalent to the application of the interval arithmetics on α-level sets of the arguments of this function (see [29] as a basis for the proof). For instance, if A and B are fuzzy numbers, then,

$$\begin{aligned} &(A+B)_L(\alpha) = A_L(\alpha) + B_L(\alpha), \\ &(A+B)_U(\alpha) = A_U(\alpha) + B_U(\alpha)\,, \\ &(\lambda \cdot A)_L(\alpha) = \begin{cases} \lambda \cdot A_L(\alpha) & \text{if } \lambda \geq 0 \\ \lambda \cdot A_U(\alpha) & \text{if } \lambda < 0 \end{cases} \\ &(\lambda \cdot A)_U(\alpha) = \begin{cases} \lambda \cdot A_U(\alpha) & \text{if } \lambda \geq 0 \\ \lambda \cdot A_L(\alpha) & \text{if } \lambda < 0 \end{cases} \end{aligned}$$

for any $\lambda \in \mathbb{R}$, and $\alpha \in [0,1]$.

Fuzzy Random Variables

Uncertainty, understood as randomness, is well described in Probability Theory. The concept of a random variable,

which is basic in this theory, is well-known and its definition does not need to be recalled in this article.

However, when we observe random experimental data which are imprecise, a useful tool to model either the imprecise perception of values coming from real-valued random variables or the random mechanisms generating directly these imprecisa data is the one associated with the so-called concept of fuzzy random variables. Actually, we can consider two different approaches to the concept of fuzzy random variable; the motivation for these approaches and the situations they apply to are different, but the formalization of the second notion and the associated statistical methodology can be applied to the first one.

Historically, the first widely accepted definition of the fuzzy random variable was proposed by Kwakernaak [21,22]. Kruse [19] proposed an interpretation of this notion, and according to this interpretation a fuzzy random variable $\mathcal{Z}$ may be considered as a fuzzy perception of an unknown true real-valued random variable Z_0 associated with a random experiment, and referred to as 'the original' of $\mathcal{Z}$. Below, we recall the version of this definition elaborated by Kruse and Meyer [20].

Definition 2 (Kruse and Meyer [20]) Let $(\Omega, \mathcal{A}, P)$ be a probability space, where Ω is the set of all possible outcomes of a random experiment, $\mathcal{A}$ is a σ-field of subsets of Ω (the set of all possible events of interest), and P is a probability measure associated with $(\Omega, \mathcal{A})$.

A mapping $\mathcal{X} : \Omega \to \mathcal{F}_c(\mathbb{R})$, where $\mathcal{F}_c(\mathbb{R})$ is the space of all fuzzy numbers, is called a **fuzzy random variable** if it satisfies the following properties:

i) $\{\mathcal{X}_\alpha(\omega) : \alpha \in [0,1]\}$, where $\mathcal{X}_\alpha(\omega) = (\mathcal{X}(\omega))_\alpha$ is a set representation of $\mathcal{X}(\omega)$ for all $\omega \in \Omega$;
ii) for each $\alpha \in [0,1]$ both $\mathcal{X}_\alpha^L : \Omega \to \mathbb{R}$ and $\mathcal{X}_\alpha^U : \Omega \to \mathbb{R}$, with $\mathcal{X}_\alpha^L(\omega) = \inf \mathcal{X}_\alpha(\omega)$ and $\mathcal{X}_\alpha^U(\omega) = \sup \mathcal{X}_\alpha(\omega)$, are usual real-valued random variables associated with $(\Omega, \mathcal{A}, P)$.

Values of a fuzzy random variable in Definition 2 have been conceived to model fuzzy perceptions of existing real-valued values (the values of the original, see Fig. 1). For instance, when we qualify the price of a given item in a specific store, we can perceive/label it as being 'cheap', but there is an existing price (although assumed to be unknown for the person receiving and processing data information).

Puri and Ralescu [31] introduced the concept of the also called fuzzy random variable as a generalization of the concept of random set or set-valued random element (and hence, as a generalization also of the concept of random variable). According to this definition a fuzzy random variable $\mathcal{Z}$ may be considered as a random element associating with each experimental outcome a value which is intrinsically fuzzy. Below, we recall Puri and Ralescu's definition

Definition 3 (Puri and Ralescu [31]) Given a probability space $(\Omega, \mathcal{A}, P)$, a mapping $\mathcal{X} : \Omega \to \mathcal{F}_c(\mathbb{R})$ is said to be a *fuzzy random variable* (also referred to as *random fuzzy set*) if for each $\alpha \in [0,1]$ the set-valued mapping $X_\alpha : \Omega \to \mathcal{K}_c(\mathbb{R})$, where $\mathcal{K}_c(\mathbb{R})$ is the class of the nonempty compact intervals and $X_\alpha(\omega) = (X(\omega))_\alpha$ for all $\omega \in \Omega$, is a compact convex random sets (that is, a Borel-measurable mapping with respect to the Borel σ-field generated by the topology associated with the Haussdorf metric on $\mathcal{K}_c(\mathbb{R})$).

Remark 1 The notion of fuzzy random variable has been in fact introduced in a more general way by considering as the codomain the space of fuzzy sets of the p-dimensional Euclidean space, or even more general Banach spaces, whose α-levels are nonempty compact subsets of this space. In case one constrains to $p = 1$ and fuzzy sets being convex, then one gets the last definition.

Remark 2 Although motivation to introduce fuzzy random variables was different in the approaches by Kwakernaak/Kruse and Meyer and by Puri and Ralescu, one can prove that the notion in Definition 3 implies the one in Definition 2. As a consequence, probabilistic ideas and results for the notion in Definition 3 (or for the more general one in Remark 1) apply to the notion in Definition 2, and the same happens for statistical developments. However, many probabilistic conclusions, and most of the statistical procedures for Definition 2 are based on the assumption

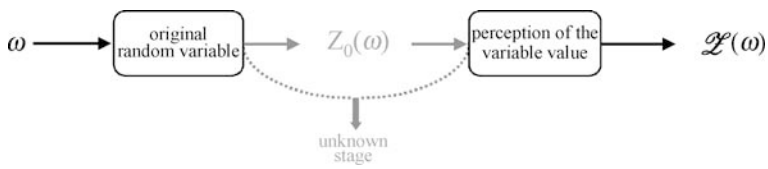

Statistics with Imprecise Data, Figure 1
Fuzzy random variables in Kwakernaak/Kruse and Meyer's sense

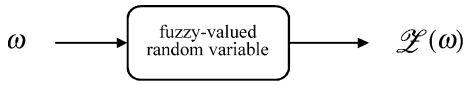

Statistics with Imprecise Data, Figure 2
Fuzzy random variables in Puri and Ralescu's sense

of having an unknown but existing original, and considering Zadeh's extension principle, so that these conclusions and procedures are not usually applicable to deal with data coming from fuzzy random variables in Definition 3.

Remark 3 The concept of fuzzy random variable in Definition 3 can be alternatively formalized as a Borel-measurable mapping with respect to the Borel σ-field generated by the topology associated with some metrics on the space $\mathcal{F}_c(\mathbb{R})$, among them an operational one we will later refer to. Borel-measurability allows us to guarantee that notions like those of the induced distribution by a fuzzy random variable, independence of fuzzy random variables, identically distributed fuzzy random variables, and so on, can be immediately formalized in the probabilistic setting.

Values of a fuzzy random variable in Definition 3 have been conceived to model existing fuzzy values (see Fig. 2). For instance, when we classify a client of a bank in accordance with the degree of aversion to investment as having a 'rather high' degree, there is no underlying real-valued degree, but the classification itself is essentially imprecise.

Statistical Analysis of Fuzzy Data Corresponding to Fuzzy Perceptions of Existing Real–Valued Data

Fuzzy Estimation

When imprecise statistical data correspond to fuzzy perceptions of unobserved/unknown precise (i. e. crisp) statistical data we can treat them as observed values of fuzzy random variables in the sense of Kwakernaak/Kruse and Meyer. In such a case we can analyze imprecise data in terms of probability distributions of their unobserved originals in a similar way as precise statistical data are analyzed using methods of traditional mathematical statistics. The only difference stems from the fact that having imprecise input information in the form of fuzzy data we cannot precisely evaluate the characteristics of the underlying probability distribution. Therefore, instead of finding precise values of the estimators of the 'parameters' describing the underlying probability distribution (the one of the original), it seems more coherent finding their imprecise fuzzy perceptions.

Assume that we observe a fuzzy random sample $\mathcal{X}_1, \ldots, \mathcal{X}_n$ which is viewed as a fuzzy perception of an unobserved random sample $X_1, \ldots, X_n$. Let $F(x;\theta)$ be the cumulative probability function of the original random variable X characterized by a crisp parameter $\theta \in \Theta$. Suppose now that an estimator of θ which is given by a statistic $\hat{\theta} = \phi(X_1, \ldots, X_n)$ is considered. By using Zadeh's extension principle we can consider as a *fuzzy estimator* of θ based on $\mathcal{X}_1, \ldots, \mathcal{X}_n$ the one associating with each fuzzy sample information $(\tilde{x}_1, \ldots, \tilde{x}_n)$ the *fuzzy estimate* $\phi(\tilde{x}_1, \ldots, \tilde{x}_n)$ given by

$$\mu_{\phi(\tilde{x}_1,\ldots,\tilde{x}_n)}(t) = \begin{cases} \sup_{(x_1,\ldots,x_n)\,|\,t=\phi(x_1,\ldots,x_n)} \min\{\mu_{\tilde{x}_1}(x_1), \ldots, \mu_{\tilde{x}_n}(x_n)\} \\ \quad \text{if } t \in \mathrm{Im}(\phi(X_1, \ldots, X_n) \\ 0 \quad \text{otherwise} \end{cases} \tag{7}$$

Alternatively, $\phi(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ can be viewed as a fuzzy estimate of the induced fuzzy parameter $\vartheta = \theta(\mathcal{X})$ with $\theta(\mathcal{X})(t) = \sup_{X \in \mathcal{E}(\Omega,\mathcal{A},P)\,|\,\theta(X)=t} \inf_{\omega\in\Omega} \mu_{\mathcal{X}(\omega)} X(\omega)$ and $\mathcal{E}(\Omega, \mathcal{A}, P)$ being the class of all possible originals of $\mathcal{X}$. In many practical cases, when $\phi(x_1, \ldots, x_n)$ has a simple form, the calculation of (7) is straightforward. For example, when $\phi(x_1, \ldots, x_n) = \text{sample mean} = (x_1 + \ldots + x_n)/n$ the α-levels of the estimate of its expected value θ are expressed as

$$(\phi(\tilde{x}_1, \ldots, \tilde{x}_n))_\alpha = \left[\frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_i)_L(\alpha),\ \frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_i)_U(\alpha)\right] \tag{8}$$

Analogously, whenever the estimator of θ is given by a continuous function or an injective function, the α-levels becomes also rather simple. In other cases, the limits of the α-levels should often be found by solving nonlinear mathematical programming problems defined by (7).

A similar approach may be applied when we try to construct fuzzy versions of the confidence intervals of the unknown parameter θ. Let us assume that we are able to find confidence intervals of θ using precise (crisp) statistical data. For example, let $[\underline{\pi}_l, +\infty)$, where $\underline{\pi}_l = \underline{\pi}_l(X_1, \ldots, X_n; \delta)$, be the one-sided confidence interval for the parameter θ on the confidence level $1-\delta$. Kruse and Meyer [20] have shown that when we replace $\underline{\pi}_l$ with the lower limits of the α-cuts of its fuzzy version we obtain a proper confidence interval for the fuzzy perception of θ. These lower limits can be found for different values of $\alpha \in (0, 1]$ from the formula [15]

$$\begin{aligned} \underline{\Pi}^L_\alpha &= \underline{\Pi}^L_\alpha(\mathcal{X}_1, \ldots, \mathcal{X}_n; \delta) \\ &= \inf\{t \in \mathbb{R} : \forall i \in \{1, \ldots n\}\, \exists x_i \in X_{i,\alpha} \\ &\qquad \text{such that } \underline{\pi}_l(x_1, \ldots, x_n; \delta) \le t\} \end{aligned} \tag{9}$$

where $X_{i,\alpha}$, $i = 1, \ldots, n$ are respective α-cuts of the observed fuzzy data.

In a similar way we can define a fuzzy equivalent of the one-sided confidence interval $(-\infty, \overline{\pi}_u]$:

$$\begin{aligned}\overline{\Pi}_\alpha^U &= \overline{\Pi}_\alpha^U(\mathcal{X}_1, \ldots, \mathcal{X}_n; \delta) \\ &= \sup\big\{t \in \mathbb{R} : \forall i \in \{1, \ldots n\}\, \exists z_i \in X_{i,\alpha} \\ &\quad \text{such that } \overline{\pi}_u(z_1, \ldots, z_n; \delta) \geq t\big\}\end{aligned} \tag{10}$$

where $\overline{\pi}_u(z_1, \ldots, z_n; \delta) = \underline{\pi}_l(z_1, \ldots, z_n; 1-\delta)$. Moreover, exactly the same approach can be applied when we look for two-sided confidence intervals.

Summing up this section we can say that when imprecise observations may be treated as fuzzy perceptions of precise but unobserved realizations of ordinary random variables the problem of point and interval estimation of the unknown parameters of the underlying probability distribution can be reduced to finding fuzzy versions of the formulae known from traditional mathematical statistics.

Fuzzy Statistical Tests

Testing statistical hypotheses is the second main branch of mathematical statistics. Tests of statistical hypotheses have to be applied if we want to make decisions based on the analysis of random data. When our decisions depend on the values of the parameters of probability distributions that describe observed statistical data we use parametric statistical methods. In such a case we test statistical hypotheses about the values of the parameters of probability distributions utilizing a well known equivalence between the set of values of the considered probability distribution parameter for which the null hypothesis is accepted and a certain confidence interval for this parameter. Kruse and Meyer [20] have shown that the same equivalence exists in the case of statistical tests with fuzzy data.

Let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ denote a fuzzy sample, i. e. a fuzzy perception of the usual random sample $X_1, \ldots, X_n$, from the population with the distribution P_θ. Let δ be a given number from the interval $(0, 1)$. Grzegorzewski ([15]) proposed the following definition of the *fuzzy test* for vague data:

Definition 4 A function $\varphi : \big(\mathcal{F}_c(\mathbb{R})\big)^n \to \mathcal{F}(\{0, 1\})$ is called a fuzzy test for the hypothesis H, at the significance level δ, if

$$\sup_{\alpha \in [0,1]} P\{\omega \in \Omega : \varphi_\alpha(\mathcal{X}_1(\omega), \ldots, \mathcal{X}_n(\omega)) \subseteq \{1\} \,|H\} \leq \delta, \tag{11}$$

where φ_α is the α-level set (α-cut) of $\varphi(\mathcal{X}_1, \ldots, \mathcal{X}_n)$.

The fuzzy test defined above can be regarded as a family of classical tests $\{\varphi_\alpha : \alpha \in (0, 1]\}$ for which the significance level is given as the upper bound of type I error for the whole family $\{\varphi_\alpha : \alpha \in (0, 1]\}$.

In order to give an example of a fuzzy statistical test let us consider a following simple null hypothesis: $H : \theta = \theta_0$, against the composite two-sided alternative: $K : \theta \neq \theta_0$. Suppose we know a two-sided symmetrical confidence interval $[\pi_1, \pi_2]$ for θ, on a confidence level $1 - \delta$, where $\pi_1 = \pi_1(X_1, \ldots, X_n; \delta/2)$ and $\pi_2 = \pi_2(X_1, \ldots, X_n; \delta/2)$ are the limits of the ordinary two-sided confidence interval. The fuzzy equivalent of this confidence interval can be calculated using the α-cuts $\Pi_\alpha = [\Pi_\alpha^L, \Pi_\alpha^U]$ for all $\alpha \in (0, 1]$, where the limits of these α-cuts can be computed from Eq. (9)-(10) by replacing δ with $\delta/2$. The fuzzy two-sided statistical test for $H : \theta = \theta_0$, against $K : \theta \neq \theta_0$, on the significance level δ, has been defined by Grzegorzewski [15] as a function $\varphi : (\mathcal{F}_c(\mathbb{R}))^n \to \mathcal{F}(\{0, 1\})$ with following α-cuts

$$\varphi_\alpha(X_1, \ldots, X_n) = \begin{cases} \{0\} & \text{if } \theta_0 \in (\Pi_\alpha \setminus (\neg\Pi)_\alpha)\,, \\ \{1\} & \text{if } \theta_0 \in ((\neg\Pi)_\alpha \setminus \Pi_\alpha)\,, \\ \{0, 1\} & \text{if } \theta_0 \in (\Pi_\alpha \cap (\neg\Pi)_\alpha)\,, \\ \emptyset & \text{if } \theta_0 \notin (\Pi_\alpha \cup (\neg\Pi)_\alpha)\,, \end{cases} \tag{12}$$

Similarly we may obtain fuzzy tests for one-sided hypotheses using the one-to-one correspondence between the acceptance regions of the tests designated for testing these hypotheses on the significance level δ and one-sided confidence intervals for the parameter θ on the confidence level $1 - \delta$.

Grzegorzewski [15] has shown that the membership function of the fuzzy test for the hypothesis H against K is given by

$$\begin{aligned}\mu_\varphi(t) &= \mu_\Pi(\theta_0) I_{\{0\}}(t) + \mu_{\neg\Pi}(\theta_0) I_{\{1\}}(t) \\ &= \mu_\Pi(\theta_0) I_{\{0\}}(t) + (1 - \mu_\Pi(\theta_0)) I_{\{1\}}(t)\,, \quad t \in \{0, 1\},\end{aligned} \tag{13}$$

where Π is a fuzzy acceptance region depending on the considered hypotheses. Thus, the fuzzy fuzzy test defined by Eq. (12), contrary to the classical crisp test, does not lead to the binary decision – to accept or to reject the null hypothesis – but to a fuzzy decision. One may get $\mu_\varphi(0) = 1, \mu_\varphi(1) = 0$ which indicates that we should accept H, or $\mu_\varphi(0) = 0, \mu_\varphi(1) = 1$ which means the rejection of H. However, one may also get $\mu_\varphi(0) = \mu_0, \mu_\varphi(1) = 1 - \mu_0$, where $\mu_0 \in (0, 1)$, which can be interpreted as a degree of conviction that we should accept (μ_0) or reject $(1 - \mu_0)$ the hypothesis H. Thus, in

situation when μ_0 is neither 0 nor 1, a user must decide using other criteria whether to reject or to accept the considered hypothesis. There exist several approaches that are suitable for solving this problem. One of these approaches which is formulated in the language of the possibility theory has been proposed by Hryniewicz [16] who used the results of Dubois et al. [9] who proposed to use statistical confidence intervals of parameters of probability distributions for the construction of possibility distributions of these parameters. According to their approach, the family of two-sided confidence intervals

$$[\pi_L(x_1,\ldots,x_n;1-\delta/2), \quad \pi_U(x_1,\ldots,x_n;1-\delta/2)], \quad \delta\in(0,1) \tag{14}$$

forms the *possibility distribution* $\tilde{\vartheta}$ of the estimated value of the unknown parameter ϑ. In a similar way it is possible to construct one-sided possibility distributions based on one-sided nested confidence intervals. Hryniewicz [16] proposed to compare this possibility distribution with a hypothetical value of the tested parameter. For this purpose he proposed to use the necessity of strict dominance measure introduced by Dubois and Prade [7] for measuring the necessity of the strict dominance relation $\tilde{A}\succ\tilde{B}$, where $\tilde{A}$ and $\tilde{B}$ are fuzzy sets. This measure, called the *Necessity of Strict Dominance* index (*NSD*), is defined as

$$\begin{aligned} NSD &= \mathrm{Ness}\left(\tilde{A}\succ\tilde{B}\right) \\ &= 1-\sup_{x,y;x\le y}\min\left\{\mu_A(x),\mu_B(y)\right\}. \end{aligned} \tag{15}$$

Hryniewicz [16] has shown that in the classical case of precise statistical data and precisely defined statistical hypotheses the value of the *NSD* index is equal to the p-value of the test.

In case of fuzzy data the confidence intervals used for the construction of the possibility distribution of the estimated parameter θ can be replaced by their fuzzy equivalents presented in the previous sections of this article. In his paper Hryniewicz [16] assumes that the value of the significance level of the corresponding statistical test δ is equal to the possibility degree α that defines the respective α-cut of the possibility distribution of $\tilde{\theta}$. He also assumes that in the possibilistic analysis of statistical tests on the significance level δ we should take into account only those possible values of the fuzzy sample whose possibility is not smaller than δ. Thus, the α-cuts of the membership function $\mu_F(\theta)$ denoted by $\left[\mu_{F,L}^{(\alpha)},\mu_{F,U}^{(\alpha)}\right]$ are equivalent to the α-cuts of the respective fuzzy confidence intervals on a confidence level $1-\alpha$. Having the possibility distribution of the test statistic we can use Eq. (15) for the calculation of the degree on necessity that the considered statistical hypothesis has to be accepted. When we set a critical value for this characteristic we arrive at unequivocal (crisp) decisions. It is worthy to note that this approach has been generalized in [16] to the case of testing imprecisely defined hypothesis using fuzzy statistical data.

In the previous paragraphs we have presented fuzzy statistical tests when the class the underlying probability distribution belongs to is known. Verification of this assumption when the available statistical data are imprecise may be very difficult indeed. Therefore, it would be advisable to use fuzzy equivalents of non-parametric (distribution-free) statistical methods. Unfortunately, there exist only few papers devoted to such fuzzy tests. The most interesting result has been obtained by Denœux et al. [4] who proposed a general methodology for the construction of fuzzy tests based on rank statistics.

Statistical analysis of fuzzy random data can be also done in the Bayesian framework. First results presenting the Bayesian decision analysis for imprecise data were given in papers by Casals et al. [3] and Gil [11]. Other approaches have been proposed by such authors as Viertl [34], Frühwirth-Schnatter [10], and Taheri and Behboodian [32]. Comprehensive Bayesian model comprising fuzzy data, fuzzy hypotheses, and fuzzy utility function has been proposed in the paper by Hryniewicz [17].

Statistical Analysis of Existing Fuzzy-Valued Data

When imprecise statistical data correspond to intrinsic fuzzy-valued data we can treat them as observed values of fuzzy random variables in the sense of Puri and Ralescu. In the literature these data are often treated as categorical/ordinal/interval-valued ones. It should be emphasized that the model given by fuzzy random variables allows us to describe and handle these data in a more expressive scale and way (in contrast to just ranking or stating simply the interval support of the values). Thus, many statistical developments for real-valued data are based on distances/deviations between values rather than on the diversity of these values. The use of the fuzzy scale allows to consider metrics with a meaning similar to that for the real-valued case (i. e., distinguishing not only the ranks of variable values w.r.t. a certain criterion, but a physical distance between them).

The distance we will consider here is the one stated by Bertoluzza et al. [2], so that if $A,B\in\mathcal{F}_c(\mathbb{R})$

$$D_W^{\varphi}(A,B)=\sqrt{\int_{[0,1]}\left[\int_{[0,1]}\left[\mathrm{f}_A(\alpha,\lambda)-\mathrm{f}_B(\alpha,\lambda)\right]^2\,\mathrm{d}W(\lambda)\right]\mathrm{d}\varphi(\alpha)} \tag{16}$$

with $f_A(\alpha, \lambda) = \lambda \cdot A_U(\alpha) + (1-\lambda) \cdot A_L(\alpha)$, where

- W and φ are normalized weighted measures on $[0,1]$ formalized as probability measures on $([0,1], \mathcal{B}_{[0,1]})$,
- W is associated with a non-degenerate distribution,
- φ is associated with a strictly increasing distribution function on $[0,1]$.

Remark It should be remarked on D_W^φ that

- W and φ have no stochastic meaning.
- To consider W is equivalents to consider a measure weighting points 0, 1 and a certain $t_0(W) \in (0,1)$; in case W is symmetric, then $t_0(W) = .5$.
- For each α, the choice of W allows us to weight
  - the effect of the distance between the "widths" of the α- levels (i. e., effect of the "shape" difference),
  - in comparison with the effect of the distance between their $t_0(W)$-points (i. e., effect of the "location" difference).
- The choice of φ allows us to weight the influence of each level (i. e., degree of "imprecision", "consensus", "subjectivity",...).
- D_W^φ is a versatile and operational in statistical developments with fuzzy numbers, and it behaves especially well when we consider least-squares approaches.

Since the concept of fuzzy random variable in Puri and Ralescu's sense has been properly stated in a probabilistic context as a random element (i. e., as a Borel-measurable function), concepts like independent and identically distributed fuzzy random variables make immediate sense. Furthermore, all the main ideas, aims and concepts, and several developments can be immediately considered to deal with fuzzy data when coming from fuzzy-valued Borel-measurable mappings. In this respect, notions like either unbiasedness or consistency of a "point" (fuzzy- or real- valued) estimator of a (fuzzy- or real- valued) 'parameter' associated with the distribution of the fuzzy random variable, or the p-value and power of a test concerning such a 'parameter', make the same sense as in the classical case.

Several developments have been made in connection with both estimation and testing of fuzzy- and real-valued parameters associated with the distribution of a fuzzy random variable in Puri and Ralescu's sense. We are now just recalling a few results concerning the "point" fuzzy estimation and testing of the population mean of a fuzzy random variable, which is formalized as follows:

Definition 5 (Puri & Ralescu, [31]) Given a probability space $(\Omega, \mathcal{A}, P)$ and an associated fuzzy random variable $\mathcal{X} : \Omega \to \mathcal{F}_c(\mathbb{R})$ such that $\max\{|\inf \mathcal{X}_0|, |\sup \mathcal{X}_0|\}$ is integrable, then, the **fuzzy expected value** (or **fuzzy mean**) of $\mathcal{X}$ is the fuzzy number $\tilde{E}(\mathcal{X}|P) \in \mathcal{F}_c(\mathbb{R})$ such that for all $\alpha \in [0,1]$

$$\begin{aligned}
&\left(\tilde{E}(\mathcal{X}|P)\right)_\alpha \\
&= \text{Aumann integral of } \mathcal{X}_\alpha \\
&= \Big\{E(X|P) \Big| X : \Omega \to \mathbb{R}, \\
&\qquad\qquad X \in L^1(\Omega, \mathcal{A}, P),\ X \in \mathcal{X}_\alpha \text{ a.s. } [P]\Big\} \\
&= \left[E(\inf \mathcal{X}_\alpha|P), E(\sup \mathcal{X}_\alpha|P)\right].
\end{aligned}$$

Remark 5 The fuzzy mean satisfies that

- If $\mathcal{X}(\Omega) = \{\tilde{x}_1, \ldots, \tilde{x}_m, \ldots\} \subset \mathcal{F}_c(\mathbb{R})$, then, it is coherent with fuzzy arithmetic, that is,

$$\begin{aligned}
\tilde{E}(\mathcal{X}|P) = {} & P(\{\omega \in \Omega \mid \mathcal{X}(\omega) = \tilde{x}_1\}) \cdot \tilde{x}_1 + \ldots \\
& + P(\{\omega \in \Omega \mid \mathcal{X}(\omega) = \tilde{x}_m\}) \cdot \tilde{x}_m + \ldots
\end{aligned}$$

- Strong Laws of Large Numbers are satisfied for different metrics (like D_W^φ, and stronger ones), which also corroborates the suitability of the defined fuzzy mean as the stochastic limit of the sample ones.
- $\tilde{E}(\mathcal{X}|P)$ is the "Fréchet expectation" of $\mathcal{X}$ w.r.t. D_W^φ, i. e., for all $A \in \mathcal{F}_c(\mathbb{R})$:

$$E\left(\left[D_W^\varphi(\mathcal{X}, \tilde{E}(\mathcal{X}|P))\right]^2 |P\right) \le E\left(\left[D_W^\varphi(\mathcal{X}, A)\right]^2 |P\right).$$

The interpretation of the fuzzy mean becomes more clear if we notice that for interval-valued random variables (i. e., random intervals) the expected value is given in a form of an interval with the lower limit equal to the expected value of the lower limits of observed random variables, and the upper limit equal to the expected value of the upper limits of observed random variables. Therefore, the expected value of the fuzzy random variable can be given as a nested set of such intervals represented by respective α-cuts, as it is formally described in the definition.

The definition of other characteristics describing fuzzy random variables may be much more complicated. For example, the definition of the variance requires the introduction of the "Fréchet expectation" (defined above), as it was proposed by Körner and Näther [24].

Fuzzy Estimation

Assume that we observe a fuzzy simple random sample $\mathcal{X}_1, \ldots, \mathcal{X}_n$ which is viewed now as an n-tuple of n independent fuzzy random variables which are identically distributed as $\mathcal{X}$. The associated ***fuzzy sample mean*** is the

statistic given by

$$\overline{\mathcal{X}}_n = \frac{1}{n} \cdot [\mathcal{X}_1 + \ldots + \mathcal{X}n] \ .$$

Then, one can prove that (see Lubiano ([25])

Theorem 1 *The fuzzy sample mean satisfies that*

i) $\overline{\mathcal{X}}_n[\cdot]$ *in an "unbiased fuzzy-valued estimator" of* $\tilde{E}(\mathcal{X}|P)$ *(in the sense of the fuzzy expected value defined by Puri & Ralescu). For most of the metrics we can consider, it is also a 'strongly consistent' fuzzy-valued estimator of* $\tilde{E}(\mathcal{X}|P)$.

ii) One can quantify the mean squared-type error in the fuzzy estimation by considering the real-valued expected value $E\left(\left[D_W^{\varphi}(\overline{\mathcal{X}}_n, \tilde{E}(\mathcal{X}|P))\right]^2\right)$.

Actually, Lubiano et al. [26,27] proposed several developments in connection with the estimation and testing about the D_W^{φ}-mean squared error associated with the estimation of the population fuzzy means by means of the sample one.

Statistical Tests

One of the problems which has received a deep attention in connection with testing from fuzzy random variables in Puri and Ralescu's sense is that concerning two-sided tests about the mean of a fuzzy random variable (one-sample case). In order to define this problem one can decide which metrics could be used for measuring the distance between the hypothetical and observed fuzzy values of considered characteristics of the fuzzy random variables. In the case of the metrics defined by Eq. (16), the problem can be formalized as follows:

Given n independent observations from $\mathcal{X}$, $\mathcal{X}_1, \ldots, \mathcal{X}_n$, we wish to test the null hypothesis $H_0\colon \tilde{E}(\mathcal{X}|P) = A \in \mathcal{F}_c(\mathbb{R})$, which can be equivalently expressed as $H_0\colon D_W^{\varphi}\left(\tilde{E}(\mathcal{X}|P), A\right) = 0$.

Despite simple formulation of the testing problem the construction of statistical tests for the verification of hypotheses related to fuzzy mean values is not that simple. For example, an exact test has been developed in Montenegro et al. [28] for so called "normal" fuzzy random variables (in Puri and Ralescu's sense, [30]). Although the method is exact and easy-to-apply, the assumption of $\mathcal{X}$ being fuzzy "normal" ($\mathcal{X} = V + \mathcal{N}(0,1)$, with $V \in \mathcal{F}_c(\mathbb{R})$) is quite restrictive and often unrealistic, as it means that *all* observed fuzzy values are described by the same membership function shape with maybe different location.

In a more general case asymptotic tests based on Large Sample Theory have been developed for the same purpose. However, they are usually hardly applicable (except for some simple special cases) in practice. For example, the asymptotic distribution of the statistic proposed by Körner [23] involves unknown parameters such as correlations between random variables describing membership functions of observed fuzzy variables. In the paper by Montenegro et al. [28] devoted to the problem of testing the equality of two fuzzy mean values, it is assumed that $\mathcal{X}$ takes on a finite number of values, and large sample sizes would be required anyway in order to estimate unknown parameters of the model.

Simulation studies have been considered to analyze the extent and applicability of the asymptotic test by Montenegro et al. [28]. These studies have confirmed that in estimating the unknown parameters (the eigenvalues of a certain correlation matrix) entails a substantial loss of precision w.r.t. the nominal significance level. Based on this empirical conclusion, the use of D_W^{φ} and the Generalized Bootstrapped CLT (Giné and Zinn, [13]) allow us to consider bootstrap techniques in this context.

The bootstrap technique consists in taking random samples (i. e. re-sampling) from the original one and calculating the value of the considered statistic. By repeating this procedure many times we obtain the sample distribution of the test statistic which may be thus used for testing purposes. In case of fuzzy random data we get the following method proposed by González-Rodríguez et al. [14]:

Theorem 2 *Given a fuzzy random variable* $\mathcal{X}\colon \Omega \to \mathcal{F}_c(\mathbb{R})$ *associated with the probability space* $(\Omega, \mathcal{A}, P)$ *and such that*

- $\max\left\{(\inf \mathcal{X}_0)^2, (\sup \mathcal{X}_0)^2\right\}$ *is integrable* ,
- $\mathcal{X}_1, \ldots, \mathcal{X}_n$ *are i.i.d. as* $\mathcal{X}$,
- $\mathcal{X}_1^*, \ldots, \mathcal{X}_n^*$ *is a bootstrap sample from* $\mathcal{X}_1, \ldots, \mathcal{X}_n$,

then, to test H_0 *at the nominal significance level* $\alpha \in [0,1]$, H_0 *should be rejected whenever*

$$\frac{\left[D_W^{\varphi}\left(\overline{\mathcal{X}_n}, U\right)\right]^2}{\widehat{S}_n^2} > z_\alpha \ ,$$

where z_α *is the* $100(1-\alpha)$ *fractile of the bootstrap distribution of*

$$T_n = \left[D_W^{\varphi}\left(\overline{\mathcal{X}_n^*}, \overline{\mathcal{X}_n}\right)\right]^2 \Big/ \widehat{S}_n^{*\,2}$$

with

$$\overline{\mathcal{X}_n^*} = \sum_{i=1}^{n} \mathcal{X}_i^*/n, \quad \widehat{S}_n^{*\,2} = \sum_{i=1}^{n} \left[D_W^{\varphi}\left(\mathcal{X}_i^*, \overline{\mathcal{X}_n^*}\right)\right]^2 /(n-1) \, .$$

Other bootstrap tests about means of fuzzy random variables which have been recently developed are the following ones:

- One-sided hypotheses tests in the one-sample case.
- Tests for the equality of means of two FRVs
  - for two independent samples
  - for two linked samples.
- Tests for the equality of means of J FRVs (ANOVA):
  - for J independent samples.

From comparative extensive simulation studies that have been performed recently at the University of Oviedo and the European Center for Soft Computing we can draw the following practical conclusions:

- For small/medium samples, the bootstrap method performs and behaves much better than the asymptotic one.
- For large sample sizes (over 300), the improvement is not that remarkable, but the bootstrap approach still provides the best approximation to the nominal significance level.

Taking into account that fuzzy statistical test for testing hypotheses about other characteristics of fuzzy random variables are even more complicated than the procedures designed for testing the hypotheses about mean values we can claim that the bootstrap technique is probably the most promising for dealing with random fuzzy observations.

Future Directions

Statistical analysis of imprecise data is still a developing area of science. Future directions of its development are tightly connected with the development of methods that may be used for the description of uncertainties of different types. It has been pointed out by P. Walley (see, e. g. [36] for a good overview) that traditional probability is not sufficient for good description of different types of uncertainty. Different mathematical models, such as e. g. Dempster–Schafer belief functions, possibility distributions, lower and upper probabilities, lower and upper previsions, and many others, have been proposed for this purpose. However, for the most general models describing uncertainty appropriate statistical methods have not been proposed yet. Therefore, statistical methods for handling very general imprecise data have to be developed in the future.

Another, but much more specific, future direction for the development of statistical analysis of imprecise data is related to the analysis of intrinsically fuzzy data. In contrast to the situation when fuzzy observations may be considered as fuzzy perceptions of real-valued observations, many notions known from traditional statistics are still waiting for their widely accepted definitions, and statistical methods of analysis.

The most challenging future direction is related to Zadeh's paradigm of "computing with words". First of all, we need operational methods for the representation of linguistic concepts which could be useful is statistical analysis of imprecisely reported (with words!) statistical data. Moreover, we also need methods for convincing presentation of the results of computations to users who have only limited knowledge of mathematics and statistics.

Bibliography

Primary Literature

1. Bandemer H, Näther W (1992) Fuzzy Data Analysis. Kluwer, Dordrecht
2. Bertoluzza C, Corral N, Salas A (1995) On a new class of distances between fuzzy numbers, Mathware and Soft Computing, vol 2. Departament of Computer Science and Artificial Intellingence of the University of Granada and Secció de Matemàtiques i Informàtica of the Universitat Politècnica de Catalunya, Granada, Barcelona, pp 71–84. http://docto-si.ugr.es/Mathware/ENG/mathware.html
3. Casals R, Gil MA, Gil P (1986) The fuzzy decision problem. In: An approach to the problem of testing statistical hypotheses with fuzzy information. Eu J Oper Res 27:371–382
4. Denœux T, Masson M-H, Hébert PA (2005) Nonparametric rank-based statistics and significance tests for fuzzy data. Fuzzy Sets Syst 153:1–28
5. Dubois D, Prade H (1978) Operations on Fuzzy Numbers. Int J Syst Sci 9:613–626
6. Dubois D, Prade H (1980) Fuzzy Sets and Systems. In: Theory and Applications. Academic Press, New York
7. Dubois D, Prade H (1983) Ranking fuzzy numbers in the setting of possibility theory. Inf Sci 30:184–244
8. Dubois D, Prade H (1988) Possibility Theory. Plenum Press, New York
9. Dubois D, Foulloy L, Mauris G, Prade H (2002) Probability-possibility transformations, triangular fuzzy-sets and probabilistic inequalities, Proc of the Ninth International Conference IPMU, Annecy, pp 1077–1083
10. Frühwirth-Schatter S (1993) Fuzzy Bayesian inference. Fuzzy Sets Syst 60:41–58
11. Gil MÁ, Kacprzyk J, Fedrizzi M (eds) (1988) Probabilistic-Possibilistic approach to some Statistical Problems with Fuzzy Experimental Observations. In: Combining Fuzzy Imprecision with Probabilistic Uncertainty in Decision Making. Springer, Berlin, pp 286–306
12. Gil MÁ, López-Díaz M, Ralescu DA (2006) Overview on the development of fuzzy random variables. Fuzzy Sets Syst 157:2546–2557
13. Giné E, Zinn J (1990) Bootstrapping general empirical measures. Ann Prob 18:851–869

14. González-Rodríguez G, Montenegro M, Colubi A, Gil MA (2006) Bootstrap techniques and fuzzy random variables. Synergy in hypothesis testing with fuzzy data. Fuzzy Sets Syst 157: 2608–2613
15. Grzegorzewski P (2000) Testing statistical hypotheses with vague data. Fuzzy Sets Syst 112:501–510
16. Hryniewicz O (2006) Possibilistic decisions and fuzzy statistical tests. Fuzzy Sets Syst 157:2665–2673
17. Hryniewicz O, Grzegorzewski P, Gil MÁ (eds) (2002) Possibilistic Approach to the Bayes Statistical Decisions. In: Soft Methods in Probability, Statistics and Data Analysis. Physica, Heidelberg, pp 207–218
18. ISO/IEC (1995) Guide to the expression of uncertainty in measurement (GUM). ISO/IEC, Geneva
19. Kruse R (1982) The strong law of large numbers for fuzzy random variables. Inf Sci 28:233–241
20. Kruse R, Meyer KD (1987) Statistics with Vague Data. D. Riedel Publishing Company, Dordrecht
21. Kwakernaak H (1978) Fuzzy Random Variables. Definitions and Theorems. Inf Sci 15:1–15
22. Kwakernaak H (1979) Fuzzy Random Variables. Algorithms and Examples for the Discrete Case. Inf Sci 17:253–278
23. Körner R (2000) An asymptotic α-test for the expectation of random fuzzy variables. J Stat Plann Inference 83:331–346
24. Körner R, Näther W (2002) On the variance of random fuzzy variables. In: Bertoluzza C, Gil MÁ, Ralescu DA (eds) Statistical Modeling, Analysis and Management of Fuzzy Data. Physica, Heidelberg, pp 25–42
25. Lubiano MA (1999) Medidas de variación de elementos aleatorios. Ph D Thesis, University of Oviedo
26. Lubiano MA, Gil MA (1999) Estimating the expected value of fuzzy random variables in random samplings from finite populations. Stat Pap 40:277–295
27. Lubiano MA, Gil MA, López-Díaz M, López-García MT (2000) The $\vec{\lambda}$-mean squared dispersion associated with a fuzzy random variable. Fuzzy Sets Syst 111:307–317
28. Montenegro M, Colubi A, Casals MR, Gil MA (2004) Asymptotic and Bootstrap techniques for testing the expected value of a fuzzy random variable. Metrika 59:31–49
29. Nguyen HT (1978) A note on the extension principle for fuzzy sets. J Math Anal Appl 64:369–380
30. Puri ML, Ralescu DA (1985) The concept of normality for fuzzy random variables. Ann Prob 13:1373–1379
31. Puri ML, Ralescu DA (1986) Fuzzy Random Variables. J Math Anal Appl 114:409–422
32. Taheri SM, Behboodian J (2001) A Bayesian approach to fuzzy hypotheses testing. Fuzzy Sets Syst 123:39–48
33. Terán P (2007) Probabilistic foundations for measurement modelling with fuzzy random variables. Fuzzy Sets Syst 158:973–986
34. Viertl R (ed) (1987) Is it necessary to develop a fuzzy Bayesian inference. In: Probability and Bayesian Statistics. Plenum, New York, pp 471–475
35. Viertl R (1996) Statistical Methods for Non-Precise Data. CRC Press, Boca-Raton
36. Walley P (1996) Measures of uncertainty in expert systems. Artif Intell 114:1–58
37. Zadeh LA (1956) Fuzzy sets. Inf Control 8:338–353
38. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning. Inf Sci 8(1):199–249; 8(2):301–353; 9(3):43–80

Books and Reviews

Bertoluzza C, Gil MÁ, Ralescu DA (eds) (2002) Statistical Modeling, Analysis and Management of Fuzzy Data. Physica, Heidelberg

Dubois D, Lubiano MA, Prade H, Gil MÁ, Grzegorzewski P, Hryniewicz O (eds) (2008) Soft methods for handling variability and imprecision. Springer, Berlin

Gil MÁ, López-Díaz M, Ralescu DA (2006) Overview on the development of fuzzy random variables. Fuzzy Sets Syst 157: 2546–2557

Grzegorzewski P, Hryniewicz O, Gil MÁ (eds) (2002) Soft methods in probability. In: Statistics and data analysis. Physica, Heidelberg

Kruse R, Gebhardt J, Gil MÁ (1999) Fuzzy Statistics. In: Webster JG (ed) Encyclopedia of Electrical and Electronics Engineering. Wiley, New York

Lawry J, Miranda E, Bugarin A, Li S, Gil MÁ, Grzegorzewski P, Hryniewicz O (eds) (2006) Soft Methods for Integrated Uncertainty Modeling. Springer, Berlin

López-Díaz M, Gil MÁ, Grzegorzewski P, Hryniewicz O, Lawry J (eds) (2004) Soft Methodology and Random Information Systems. Springer, Berlin

Taheri SM (2003) Trends in Fuzzy Statistics. Austrian J Stat 32:239–257

Stellar Dynamics, *N*-body Methods for

Junichiro Makino
Center for Computational Astrophysics, National Astronomical Observatory of Japan, Tokyo, Japan

Article Outline

Glossary
Definition of the Subject
Introduction
Statistical Description
Numerical Methods
Future Directions
Bibliography

Glossary

Binary Two stars which orbit around each other.

Distribution function A density function in the six-dimensional phase space which gives the distribution of stars in a stellar system.

Fokker–Planck equation Partial differential equation for the thermal evolution of the distribution function of a stellar system, expressed in the form of the advection-diffusion equation in the phase space.

Monte-Carlo method The method to numerically follow the thermal evolution of stellar systems using Monte-Carlo integration of Fokker–Planck equation for the distribution function.

Relaxation The process which leads the system to thermal equilibrium. In the case of a stellar system, the main mechanism of relaxation is the gravitational encounter between stars.

Stellar systems A system composed of a large number of stars, i. e., particles which interact through mutual gravity.

Thermal stability Stability of the system against the redistribution of the thermal energy.

Definition of the Subject

Most of astronomical objects are in the first approximation the collection of point mass particles (stars or planets). We call such systems stellar systems, and stellar dynamics is the theoretical framework for the study of such systems. Since the equation of motion for stellar systems with N stars is analytically solvable only for the case of $N = 2$, numerical integration of the orbit of stars has been very important tool for the study of stellar systems.

Introduction

Stellar Dynamics deals with the evolution of systems which consist of a large number of stars interacting with other stars through mutual gravitational force. Examples of such systems include star clusters, galaxies, clusters of galaxies. Star clusters consist of up to around 10 million stars. They are classified into two categories: open clusters and globular clusters. Open clusters are less massive and younger than globular clusters. Globular clusters in our galaxy are all very old (more than 10 Gyr) and fairly massive (more than 10^5 solar mass). Both types of clusters are compact, with the radius of the order of 10 parsec or less.

Galaxies are larger and more massive than star clusters. The total mass of our Galaxy is around 10^{12} solar mass. Around 80% of the total mass is in dark matter, the existence of which we only indirectly know through the measurement of the motions of visible matter such as gas or stars.

In the first approximation, stellar systems such as star clusters and galaxies are just a collection of a large number of point masses. Therefore, one might expect that a statistical description is applicable. In other words, one might expect that the behavior of the system is described by the statistical mechanics. If that were the case, however, the universe would look very different from its present form. All galaxies would look similar, and rather boring, without any structures such as spirals, bars, shells etc., which make the diversity of galaxies. In this article, we first review how the statistical description can be or cannot be applied to stellar systems, and then present an overview of our current understanding for the evolution of stellar systems. Then we discuss the numerical methods for N-body simulations. For a more complete coverage of the subject see textbooks [5,25].

Statistical Description

Thermal Stability and Gravothermal Catastrophe

There are two basic reasons why stellar systems cannot be described by statistical mechanics. The first one is that the thermal equilibrium state does not exist for stellar systems. An equilibrium state for a system of classical particles is given by the Maxwell-Boltzmann statistics. This means that the local velocity distribution is Maxwellian with a single temperature everywhere in the system. In other words, the velocity distribution function is given by

$$f_0(v) = \frac{n_f}{(2\pi\sigma^2)^{3/2}} \exp\left(\frac{-v^2/2}{\sigma^2}\right) , \tag{1}$$

where n_f is the volume number density of particles and σ is the velocity dispersion.

However, with a stellar system, this Maxwellian distribution cannot be realized, simply because the stars with sufficiently high velocities will escape from the system. The depth of the gravitational potential of a stellar system is finite. Thus, if a star gained kinetic energy higher than that of the depth of the gravitational potential at its position, it would escape from the system.

As a thought experiment, we can think of a stellar system with a spherical adiabatic wall around it. The thermal equilibrium state of the stellar system can be described by that of an ideal gas, since the distribution function is Maxwellian for both cases. Therefore, even though we cannot apply the terms like pressure and temperature to stellar systems in general, we can apply these concepts to stellar systems in the thermal equilibrium, since the equilibrium distribution function is the same as that of an ideal gas with self gravity.

In this case, since the stars do not escape, the equilibrium state does exist. In the limit of very high temperature (large kinetic energy), the distribution of stars within the wall is uniform, with the number density of stars same everywhere. However, as we lower the temperature, the central density becomes higher due to the effect of self gravity. There must be a pressure gradient to support the gravity, and that means a density gradient should exist. However, this increase of density gradient means an increase of the

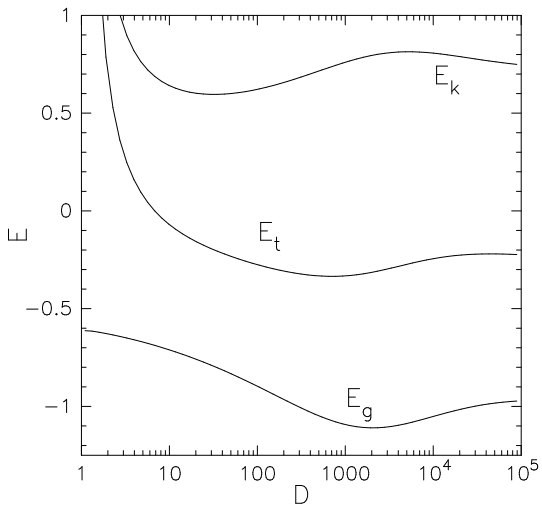

Stellar Dynamics, *N*-body Methods for, Figure 1
Energies as function of the ratio between the density at the center and that at the place of the wall, *D*. Three curves indicate the kinetic energy (*top*), total energy (*middle*) and gravitational energy (*bottom*), respectively

gravitational force, because more mass is concentrated to the center. Thus, if the increase of the gravitational force is larger than the increase of pressure, than the pressure gradient cannot support the system and there is no equilibrium state.

Figure 1 shows this phenomenon more clearly. Here, we plot the thermal energy (E_k), gravitational energy(E_g) and total energy ($E_t = E_k + E_g$) as a function of the ratio of the density at the center and that of just inside the wall. Here, for simplicity, we use the system of units where the radius of the wall R and mass of the gas M are both unity and the gravitational constant G is also unity. The limit as $D \to 1$ corresponds to infinite temperature. In this limit, the gravitational energy converges to -0.6, and the kinetic energy and total energy diverge. As we increase D, the kinetic energy decreases, gravitational energy increases, and the total energy decreases. However, at $D = 709$, the total energy E_t takes a minimum.

The fact that E_t takes a minimum means that this system has neutral stability against the perturbation which conserves the total energy, since an infinitesimal change of D in either direction does not change the total energy. In other words, for this system we can redistribute the thermal energy in such a way that after that perturbation is applied the system is still isothermal. For $D < 709$, when we change D the total energy also changes. In other words, if we redistribute the energy in any way, as far as we keep the total energy unchanged, the system cannot remain isothermal.

This change of behavior corresponds to a sign of the (second-order) variation of total entropy $\delta^2 S$. For $D < 709$, for any thermal perturbation, $\delta^2 S < 0$, and the second law of thermodynamics guarantees that the isothermal state is stable. However, for $D = 709$, there is one form of perturbation for which $\delta^2 S = 0$. For $D > 709$, there are one or more perturbations for which $\delta^2 S > 0$. In other words, the isothermal state is unstable for $D > 709$.

The eigenfunction which corresponds to $\delta^2 S = 0$ has a single node, which means it either transfers the thermal energy from central area to the outer area, or vise versa. The system shows neutral stability against this heat transfer. Thus, even though the thermal energy is moved, the system remains isothermal.

In the case of $D < 709$, any perturbation in the distribution of the thermal energy results in the heat flux which cancels out the perturbation. In the case of $D = 709$, there is one mode of perturbation for which there will be no heat flux. Finally, in the case of $D > 709$, there is at least one mode of perturbation, for which the resulting heat flows in the direction that enhances perturbation. Thus, if we remove some energy from the outer region and give it to the central region, the central region becomes cooler than the outer region, and therefore heat starts to flow inwards. This inward heat flow results in more cooling.

The reason why the central region cools when it gets more thermal energy is that it expands. As a result of the expansion, the gravitational attraction becomes weaker, and the central region can expand more. Through this additional expansion the central region lowers its temperature. In this case, however, the final state is another isothermal state with the same E_t but a smaller value of D. Since the minimum value of E_t corresponds to $D = 709$, for any other value of D, there is one value of D with same E_t which is thermally stable.

If the initial perturbation removes the thermal energy from the central region and deposits it to the outer region, the central region becomes hotter, and the heat starts to flow outwards. In this case, there is no stable final state, and the central region continues to contract and becomes hotter indefinitely. Exactly what happens depends on the efficiency of the heat transport.

Roughly speaking, if the thermal timescale is shorter in the center than in the outer region, as the central region contracts, the timescale becomes even shorter. In this case, the central density can reach infinity in a finite time. Of course, in this case, the mass of the region with infinite density should become zero. On the other hand, if the timescale is longer in the center, the timescale becomes even longer as the contraction proceeds, and the overall system contracts slowly. In the case of an ideal gas,

the timescale is determined by the mechanism of the heat transfer.

In the case of stellar systems, heat transfer proceeds through close encounters between two stars. Since the rate of close encounters is directly proportional to the number density of stars, the thermal relaxation timescale is shorter for the central region, and becomes shorter as the central density goes up. This phenomenon is called "gravothermal catastrophe".

This gravothermal catastrophe occurs in real stellar systems without the adiabatic wall, since real stellar systems always have the temperature decreasing outward, which naturally drives the outward heat flux. Of course, the distribution function is not Maxwellian and one cannot strictly define the temperature. Even so, comparisons between numerical simulations with different approaches have shown that this picture gives a very good description.

Hachisu et al. [9] performed a simulation of the evolution of the self-gravitating gas system with several different forms of heat conductivity. They found that if the heat conductivity is such that the central thermal timescale is shorter than that of the outside region, the evolution of the system would become self-similar. In other words, the central region continues to shrink, leaving a power-law halo behind it. The mass of the core decreases in time and the central density reaches infinity within a finite time.

In the following, we give a simple explanation of this power-law behavior, following the description by Lynden-Bell and Eggleton [14]. Formally, a self-similar solution for a physical quantity y as a function of radius r and time t is expressed as

$$y(r,t) = y_0(t) y_*[r/r_0(t)] \,. \tag{2}$$

We can set $y_*(0) = 1$ without loss of generality. We can assume that in the limit of $r \to \infty$ there is no evolution, since the thermal timescale is longer at the outskirts. We can further assume that functions r_0 and y_0 are powers of the time t, since otherwise the self-similar solution cannot be constructed. Thus, if we express

$$r_0 = (t_0 - t)^{\beta} \,, \tag{3}$$

and

$$y_0 = (t_0 - t)^{\gamma} \,, \tag{4}$$

we have

$$y_0 = r_0^{\gamma/\beta} \,. \tag{5}$$

The ratio between the gravitational binding energy of the core and the thermal energy of the core should be constant. Therefore, we have

$$\sigma^2 \propto \frac{GM_c}{r_c} \sim \rho_0 r_0^2 \,. \tag{6}$$

If we express ρ_0 as

$$\rho_0 = r_0^{\alpha} \,, \tag{7}$$

we have

$$r_0 = (t_0 - t)^{2/(6+\alpha)} \,. \tag{8}$$

Lynden-Bell and Eggleton [14] numerically obtained the self-similar solution for the gaseous model with heat conductivity which they believed would mimic the radial energy transfer in an N-body system, and found that the self-similar solution has the characteristic power law index of

$$\rho = r^{-2.21} \,. \tag{9}$$

Henon [11] demonstrated that the N-body system would exhibit core collapse using Monte-Carlo calculation with 1000 shells. At that time, N-body simulation did show some collapse-like behavior, but it was difficult to see whether the collapse is really self-similar or not because of the limitation in the number of particles. The most beautiful demonstration of the self-similar nature of the collapse is by Cohn [6], who is the first to use the direct integration of the Fokker–Planck equation in the study of the thermal evolution of the globular clusters. He found the power index to be −2.23, which is strikingly close to the value obtained by the gas model calculation of Lynden-Bell and Eggleton [14].

In the study of the gravothermal catastrophe and self-similar solution, N-body simulation did not play a major role. This is partly because more approximate methods, such as the gaseous models and Fokker–Planck calculations, gave reasonable results, and partly because computer power available was quite limited in the 1980s.

Binary Formation and Gravothermal Oscillation

In the continuous limit, the central density diverges in a finite time, and at that time the core mass becomes zero. Since real stellar systems consist of stars with finite mass and finite size, the central density cannot reach infinity. One possibility is that stars start to physically collide, but there is another possibility.

When the central density becomes very high, the cross-section of the three-body close encounter, which causes the formation of a bound binary, becomes non-negligible.

The binding energy of the binary is transferred to the kinetic energy of the third particle and the center-of-mass motion of the binary. This is essentially the same as the nuclear fusion reaction. The difference is that in this case the interaction which makes the two stars bound is the same gravitational force which makes the entire system bound, and that means there is no ground state for a binary star. Thus, in theory, the binding energy of a binary star can become arbitrarily large. In practice, in real star clusters, there is a practical limit for the binding energy of a binary star. When a binary star interacts with another star, its center-of-mass motion acquires some energy, which is typically a fraction of its binding energy. When the kinetic energy of the center-of-mass motion becomes large enough, the binary would be ejected out the cluster. In fact, this ejection of the binary itself and the other star is the main channel which heats the central region of a star cluster.

This energy production by binaries halts the collapse. The collapse of the core is driven by the outward heat flux. If a sufficient amount of energy is generated in the core, the collapse is halted. In the case of a star, gravitational contraction of the star is halted by the energy production from the nuclear fusion reaction, and it becomes a main-sequence star.

In the case of a star cluster, it is possible to construct a steady-state solution quite similar to the main-sequence stage of a star [8,10]. However, it turned out that this steady-state solution is again thermally unstable [26]. This instability drives what is now called "gravothermal oscillation". This oscillation was first found in a gas sphere model, but then confirmed with Fokker–Planck (FP) calculations [7], and later with direct N-body simulations [16].

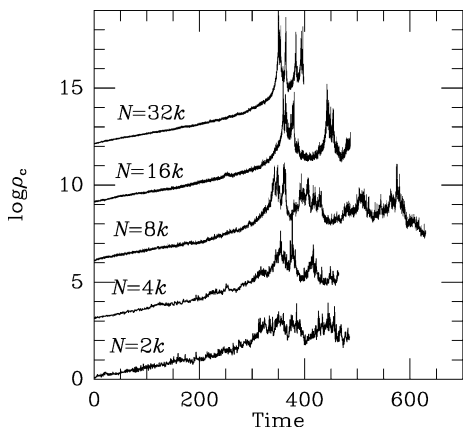

Stellar Dynamics, *N*-body Methods for, Figure 2
The time variation of the central density of simulated star clusters. (Reproduced from [16])

Figure 2 shows the time variation of the central density for simulations with 2–32k particles. The time is scaled so that the initial thermal relaxation time is the same for all runs. The curves are shifted vertically.

From Fig. 2 we can see that the oscillation is not strictly periodic. One reason is simply that the binary formation and its interaction with other stars is a stochastic process. The other reason is that, even in the continuous limit with a smooth and deterministic heating term, this oscillation is chaotic. Note that here what exhibits the chaotic behavior is the central density of the whole N-body system, which is a macroscopic variable of a system with large degrees of freedom.

Numerical Methods

In this section, we discuss the numerical methods used to obtain the solution of gravitational N-body problem. The problem itself is rather simple. We integrate the equation of motion for stars, using some numerical methods for the initial-value problem of ordinary differential equations. The calculation cost of the mutual gravitation interaction is $O(N^2)$ and dominate the calculation cost.

In practice, however, what is done is far more complex for a variety of reasons. Here, let us overview the approaches used.

First of all, for some problems, it is not necessary to calculate the interaction of all $N^2/2$ pairs at each timestep. There are several algorithms which reduce the calculation cost from $O(N^2)$ to $O(N \log N)$ or even $O(N)$. These algorithms include particle-mesh schemes, Barnes-Hut tree algorithm, and the Fast Multipole Method.

Secondly, it is not practical to apply the same timestep to all stars. While two stars undergo close encounters, their distance can become arbitrary small, depending on the impact parameter and the relative velocity. They have to be integrated with the adaptive timestep which can resolve their relative orbit. Moreover, as we discussed in Sect. "Statistical Description", many stellar systems tend to develop high-density cores, in which the stars orbit on a timescale much shorter than that of average stars. Also, in many stellar systems, interactions between binary stars (two stars orbiting around each other) and other stars play an important role, and the evolution of the binary orbit due to the interaction with other stars must be accurately integrated. The orbital timescale of binaries can become as small as a fraction of a second, while stars in globular clusters or galaxies have the orbital timescale much longer than one million years. Thus, it is clearly necessary to apply some

algorithm with which we can change the timestep of individual stars individually.

Finally, an important characteristic of a stellar system is, from both the theoretical and practical point of views, that it is a Hamiltonian system with various symmetry and conservation laws. Numerical schemes which make use of these characteristics play important roles.

In the following, we briefly overview these schemes.

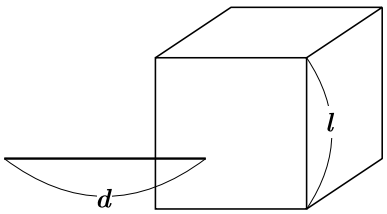

Stellar Dynamics, *N*-body Methods for, Figure 4
Opening criterion for tree traversal

Fast Interaction Calculations

The basic idea of the tree code [4] is to replace the force from a group of distant particles with the force from their center of mass or by a multipole expansion. To ensure accuracy, we make groups for distant particles large and groups for nearby particles small.

We use a tree structure to construct the appropriate grouping for each particle. Before calculating the forces on particles, we first organize particles into a tree structure. Barnes and Hut [4] used an oct-tree based on the recursive subdivision of a cube into eight subcubes. We stop the recursive subdivision if the cube has only one particle (or no particles). See [15] for details concerning an efficient tree construction algorithm. Figure 3 shows the Barnes–Hut tree in two dimensional space.

After the tree is constructed, for each node of the tree, which corresponds to a cube of a certain size, we calculate the coefficient of the multipole expansion of the gravitational force exerted by particles in that cube. The fast algorithm was described in [12].

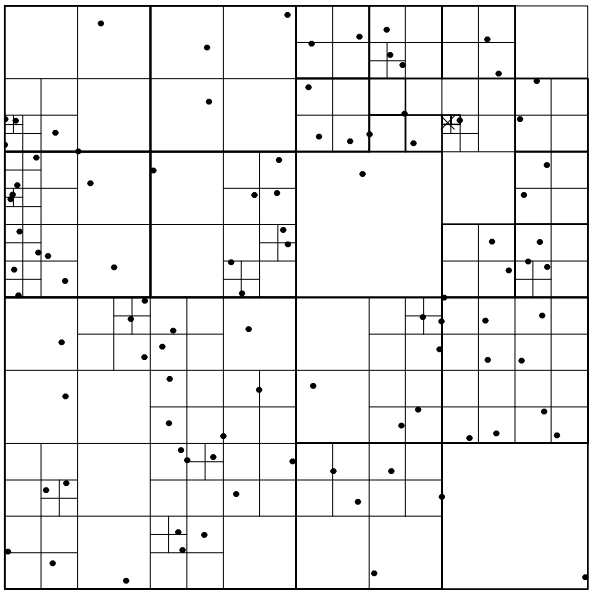

Stellar Dynamics, *N*-body Methods for, Figure 3
Barnes–Hut tree in two dimensions

The force calculation is expressed as a recursive procedure. To calculate the force on a particle, we start from the root node, which corresponds to the total system. We calculate the distance between the node and the particle (d) and compare it with the size of the node (l). If they satisfy the convergence criterion

$$\frac{l}{d} < \theta \,, \tag{10}$$

where θ is the accuracy parameter, we calculate the force from that node to the particle using the coefficients of the multipole expansion. If criterion (10) is not satisfied, the force is calculated as a summation of the forces from eight sub-nodes.

Usually, we use the distance between the particle and the center of mass of the node to determine whether the force is accurate enough. When θ is very large, this criterion can cause unacceptably large error [21]. For most calculations, however, such a pathological situation is not realized.

The fast multipole algorithm (FMM) is in some sense a natural extension of the tree algorithm. In the tree algorithm, we use the multipole expansion to express the gravitational force for a group of stars with other stars sufficiently far away (Fig. 5a). In FMM, the forces on another group of stars are expressed in terms of one spherical harmonics expansion, and the forces on these stars are obtained by evaluating the expansion at the locations of stars. The tree-based recursive approach is now applied to both the path to calculate multipole expansions and the path to calculate the spherical harmonics expansions (usually called local expansions). This treatment removes the log N term in the calculation cost of the tree algorithm, since the cells of a given size do not directly interact with stars. They interact only with the cells of the same or similar sizes.

After the invention of the tree algorithm and FMM in the late 1980s, a large number of works followed on various topics such as the theoretical analysis of the algo-

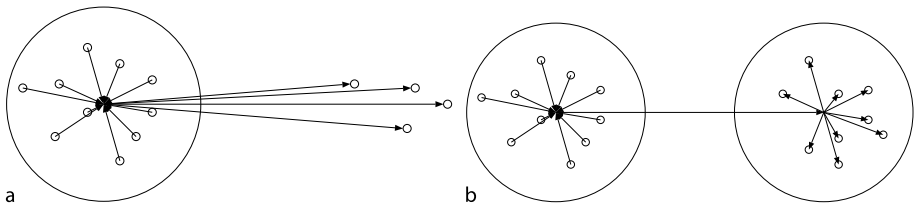

Stellar Dynamics, *N*-body Methods for, Figure 5
Tree algorithm (a) and FMM (b)

rithms, efficient implementation, a fast method to translate multipole and local expansions, parallel implementation, etc. Some of these are covered in [19].

In many other fields of computational science, the Poisson equation associated with the distribution of particles is numerically solved using FFT. If the number of grid points one needs to resolve the potential field is of the same order or less than the number of particles, it is highly advantageous to use FFT, since the calculation cost of FFT is $O(M \log M)$, where M is the number of grid points, with relatively small coefficient. However, FFT requires that the grid is regular with equal spacing. Thus, FFT alone is not very suitable for astronomical simulations, where small-scale structures develop through gravitational instability. Recently, the combination of the tree algorithm and FFT potential solver has become the standard scheme for simulations which requires periodic boundary conditions (see, e. g., Yoshikawa and Fukushige [27]).

Individual Timestep Algorithms

The individual timestep scheme [1,2] has been the only algorithm that can be used for simulations of gravitational many-body systems, such as open clusters, globular clusters and a system of planetesimals. In a simulation of these systems, we are interested in the change of orbit of each particle due to gravitational encounters with other particles, and the evolution of the total system driven by such changes. In that sense, simulation of these systems is similar to the molecular dynamics simulation, in which we are interested in the thermodynamical process.

In simulations of these collisional systems, we need to follow the changes of the orbit of a star due to individual encounters with a reasonable accuracy, since the encounters drive the evolution of the system. Therefore, we cannot use the softening parameter to simulate the evolution of these systems. In the study of the galaxies, for example, we can modify the $1/r$ potential to $1/\sqrt{r^2+\epsilon^2}$, with small constant ϵ. This softened potential makes it possible to use a constant and global timestep. When we use a strict $1/r$ potential, the force between two particles changes very rapidly when two particles undergo a close encounter. Therefore, during a close encounter, the timestep should be sufficiently small to resolve this rapid change. Roughly speaking, the timestep is determined by the distance to the nearest neighbor. Therefore, if we integrate the system in lockstep, the timestep is determined by the pair of particles with minimum separation. Even in a nearly homogeneous system, the minimum separation is proportional to $N^{-2/3}$, and the dependence on N is stronger when the system is highly inhomogeneous [17].

To reduce the total calculation cost, Aarseth [1] developed an algorithm which assigns each particle its own timestep. In this individual timestep algorithm, each particle adjusts its timestep so that it satisfies the required accuracy. Thus, when two particles undergo a close encounter, although the timesteps of these particles shrink as required, the timesteps of other particles remain long. Makino and Hut [17] showed that the individual timestep scheme is faster than a shared timestep scheme by a factor in the range of $O(N^{1/3})$ and $O(N)$, depending on the distribution of particles.

In the individual timestep algorithm, each particle has its own timestep, and therefore its own time. To calculate the force on a particle due to other particles, we must know their positions at the time of the particle for which we calculate the force. To calculate the position of a particle at a time different from the time of the particle, Aarseth used a third order polynomial extrapolation. This polynomial is evaluated each time a pairwise force is calculated, therefore the calculation cost of the polynomial evaluation is of the same order as that of the force calculation itself. The basic algorithm looks like the following in the case of Aarseth's program, which uses a predictor-corrector scheme:

(a) Select particle i with a minimum $t_i + \Delta t_i$. Set the global time (t) to be this minimum, $t_i + \Delta t_i$.
(b) Predict the positions of all the particles at time t using the extrapolation polynomial.
(c) Calculate the acceleration (a_i) for particle i at time t, using the predicted positions.
(d) Apply the corrector for the position and velocity of particle i.
(e) Go back to step (a).

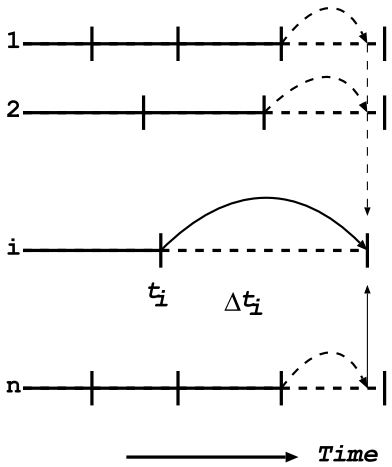

Stellar Dynamics, *N*-body Methods for, Figure 6
Schematic description of the individual timestep algorithm

Figure 6 illustrates how the individual timestep scheme works. When particle i is integrated from t_i to $t_i + \Delta t_i$, the positions of all other particles are predicted at that time, and forces from those particles are calculated using these predicted positions.

Not all time integration schemes can be used with this individual timestep, since the predicted values of the positions of all other particles need to be calculated at the time of the particle to be integrated. Most integration schemes do not provide the solutions at arbitrary points in time. For example, there is no simple way to obtain the approximate solution at the intermediate time with the usual Runge–Kutta schemes or extrapolation schemes.

For the treatment of binaries and small-N systems which can formed in larger stellar systems, see Aarseth [3].

Special Numerical Methods

A stellar system is a Hamiltonian system, and therefore the symplectic algorithms [22] can in principle be applied. These schemes show behavior much better than that of traditional schemes, and can dramatically reduce the integration error for long calculations. However, in practice there are a number of difficulties. One is that with symplectic schemes we cannot change timesteps [24]. There has been a number of works to make the symplectic scheme effectively work as variable-timestep scheme. Most of them rely on the partition of the potential energy term of the Hamiltonian into multiple parts with different timescales, and apply different timesteps to different parts of the potential term [20,23]. A very different approach is to use time-symmetric scheme instead of symplectic schemes. Symmetric schemes offer most of the practical advantages of symplectic schemes, and remain symmetric if we change the timestep. Here, the timestep must be calculated in such a way that does not destroy the time symmetry [13]. Recently, a way to combine this symmetric algorithm and the individual timestep algorithm was proposed [18].

Future Directions

In Sect. "Statistical Description" we over viewed the evolution of a system of point-mass particles. When we try to understand the evolution of real stellar systems such as globular clusters, open clusters and galactic nuclei, we need to take into account various effects which were neglected in Sect. "Statistical Description". For example, real stars are not point masses but have finite radii, and can physically collide with each other. In the case of normal stars like our sun in a galaxy, the chance that it will collide with its neighbors before the end of its lifetime is negligible. However, the number density of stars in the central regions of globular clusters or other massive star clusters can reach more than $10^6/\mathrm{pc}^3$, or about a million times that of the solar neighborhood. In such circumstances the collision is not rare. Also, individual stars have different masses. Even in the case of thermal equilibrium, stars with different masses have different spacial distribution, since the kinetic energy per unit mass is different. Thus, massive stars tend to lose their kinetic energy and segregate to the central region of the cluster, resulting in strong enhancement of collision rates between most massive stars in star clusters. To make things even more complex, each star evolves from the Main Sequence to giant, and depending on their masses, goes through Type II supernova and finally becomes a white dwarf, a neutron star, or a black hole. These stellar evolutions and stellar collisions, coupled with the stellar dynamics, makes stellar systems very interesting and an important research subject.

The ultimate research direction is to model all of these processes, stellar evolution, stellar collisions, stellar dynamics, and at some point the initial gas dynamics of star cluster formation, in one single unified simulation. From the viewpoint of the computational cost, such a simulation is not impossible, since the calculation cost of stellar dynamics is high and others are relatively minor. However, from the viewpoint of software engineering and physical modeling, we do not really know how we can actually develop and maintain a simulation program which can handle gravitational interaction between stars, evolution of individual stars, physical collisions and tidal interaction between stars, and other more complex things like the evo-

lution of binary stars. MODEST (Modeling Dense Stellar systems)[1] is one of such efforts to develop a simulation code, or a loosely coupled collection of codes, to handle complex systems.

Bibliography

1. Aarseth SJ (1963) MN 126:223
2. Aarseth SJ (1985) In: Blackbill JU, Cohen BI (eds) Multiple Time Scales. Academic Press, New York, pp 377–418
3. Aarseth SJ (2003) Gravitational N-Body Simulations. Cambridge University Press, Cambridge, pp 430
4. Barnes J, Hut P (1986) Nature 324:446
5. Binney J, Tremaine S (2008) Galactic Dynamics, 2nd edn. Princeton University Press, Princeton
6. Cohn H (1980) ApJ 242:765
7. Cohn H, Hut P, Wise M (1989) ApJ 342:814
8. Goodman J (1984) ApJ 280:298
9. Hachisu I, Nakada Y, Nomoto K, Sugimoto D (1978) Prog Theor Phys 60:393
10. Heggie DC (1984) MN 206:179
11. Henon M (1971) Astrophys Space Sci 13:284
12. Hernquist L (1990) J Comp Phys 87:137
13. Hut P, Makino J, McMillan S (1995) ApJL 443:L93
14. Lynden-Bell D, Eggleton PP (1980) MN 191:483
15. Makino J (1990) J Comput Phys 87:148
16. Makino J (1996) Dynamical Evolution of Star Clusters. In: Hut P, Makino J (eds) Proc IAU Symp 174. Kluwer, Amsterdam, pp 151–160
17. Makino J, Hut P (1988) ApJS 68:833
18. Makino J, Hut P, Kaplan M, Saygın H (2006) New Astron 12:124
19. Pfalzner S, Gibbon P (1996) Many-Body Tree Methods in Physics. Cambridge University Press, Cambridge
20. Preto M, Tremaine S (1999) AJ7 118:2532
21. Salmon JK, Warren MS (1994) J Comp Phys 111:136
22. Sanz-Serna JM, Calvo MP (1994) Numerical Hamiltonian Problems. Chapman and Hall, London
23. Skeel RD, Biesiadecki JJ (1994) Ann Numer Math 1:191
24. Skeel RD, Gear CW (1992) Physica D 60:311
25. Spitzer Jr L (1987) Dynamical Evolution of Globular Clusters. Princeton University Press, Princeton
26. Sugimoto D, Bettwieser E (1983) MN 204:19P
27. Yoshikawa K, Fukushige T (2005) PASJ 57:849

Stochastic Games

Eilon Solan
The School of Mathematical Sciences,
Tel Aviv University, Tel Aviv, Israel

Article Outline

Glossary
Definition of the Subject
Strategies, Evaluations and Equilibria
Zero-Sum Games
Multi-Player Games
Correlated Equilibrium
Imperfect Monitoring
Algorithms
Additional and Future Directions
Acknowledgments
Bibliography

[1] http://www.manybody.org/modest/

Glossary

A stochastic game A repeated interaction between several participants in which the underlying state of the environment changes stochastically, and it depends on the decisions of the participants.

A strategy A rule that dictates how a participant in an interaction makes his decisions as a function of the observed behavior of the other participants and of the evolution of the environment.

Evaluation of stage payoffs The way that a participant in a repeated interaction evaluates the stream of stage payoffs that he receives (or stage costs that he pays) along the interaction.

An equilibrium A collection of strategies, one for each player, such that each player maximizes (or minimizes, in case of stage costs) his evaluation of stage payoffs given the strategies of the other players.

A correlated equilibrium An equilibrium in an extended game in which at the outset of the game each player receives a private signal, and the vector of private signals is chosen according to a known joint probability distribution. In the extended game, a strategy of a player depends, in addition to past play, on the signal he received.

Definition of the Subject

Stochastic games, first introduced by Shapley [60], model dynamic interactions in which the environment changes in response to the behavior of the players. Formally, a stochastic game is a tuple $G = \langle N, S, (\mathcal{A}_i, A_i, u_i)_{i\in N}, q\rangle$ where

- N is a set of players.
- S is a state space. If S is uncountable, it is supplemented with a σ-algebra of measurable sets.
- For every player $i \in N$, $\mathcal{A}_i$ is a set of actions for that player, and $A_i\colon S \to \mathcal{A}_i$ is a set-valued (measurable) function that assigns to each state $s \in S$ the set of actions $A_i(s)$ that are available to player i in state s. If $\mathcal{A}_i$ is uncountable, it is supplemented with a σ-algebra of measurable sets. Denote $SA = \{(s, a)\colon s \in S, a =$

$(a_i)_{i\in N}, a_i \in A_i(s) \ \ \forall i \in N\}$. This is the set of all possible *action profiles*.

- For every player $i \in N$, $u_i : SA \to \mathbf{R}$ is a (measurable) *stage payoff function* for player i.
- $q: SA \to \Delta(S)$ is a (measurable) *transition function*, where $\Delta(S)$ is the space of probability distributions over S.

The game starts at an initial state s^1, and is played as follows. At each stage $t \in \mathbf{N}$, each player $i \in N$ chooses an action $a_i^t \in A_i(s^t)$, receives the stage payoff $u_i(s^t, a^t)$, where $a^t = (a_i^t)_{i\in N}$, and the game moves to a new state s^{t+1} that is chosen according to the probability distribution $q(\cdot \mid s^t, a^t)$.

A few comments are in order.

1. A stochastic game lasts infinitely many stages. However, the model also captures finite interactions (of length t), by assuming the play moves, at stage t, to an absorbing state with payoff 0 to all players.
2. In particular, by setting $t = 1$, we see that stochastic games are a generalization of matrix games (games in normal formgames in normal form), which are played only once.
3. Stochastic games are also a generalization of repeated games, in which the players play the same matrix game over and over again. Indeed, a repeated game is equivalent to a stochastic game with a single state.
4. Stopping games are also a special case of stochastic games. In these games every player has two actions in all states, *continue* and *quit*. as long as all players choose *continue* the stage payoff is 0; once at least one player chooses *quit* the game moves to an absorbing state.
5. Markov decision problems (see, e. g., [49]) are stochastic games with a single player.
6. The transition function q governs the evolution of the game. It depends on the actions of all players and on the current state, so that all the players influence the evolution of the game.
7. The payoff function u_i of player i depends on the current state as well as on the actions chosen by all players. Thus, a player's payoff depends not only on that player's choice, but also on the behavior of the other players.
8. Though we refer to the functions $(u_i)_{i\in N}$ as "stage payoffs", with the implicit assumption that each player tries to maximize his payoff, in some applications these functions describe a stage *cost*, and then the implicit assumption is that each player tries to minimize his cost.
9. The action of a player at a given stage affects both his stage payoff and the evolution of the state variable, thereby affecting his future payoffs. These two, sometimes contradicting effects make the optimization problem of the players quite intricate, and the analysis of the game challenging.
10. The players receive a stage payoff at each stage. So far we did not mention how the players evaluate the infinite stream of stage payoffs that they receive, nor did we say what is their information at each stage: Do they observe the current state? Do they observe the actions of the other players? These issues will be discussed later.
11. The actions that are available to the players at each stage, the payoff functions, and the transition function, all depend on the current state, and not on past play (that is, past states that the game visited, and past actions that the players chose). This assumption is without loss of generality. Indeed, suppose that the actions available to the players at each stage, the payoff functions, and the transition function, all depend on past play, as well as on the current state. For every $t \in \mathbf{N}$ let H_t be the set of all possible *histories* of length t, that is, all sequences of the form $(s^1, a^1, s^2, a^2, \ldots, s^t)$, where $s^k \in S$ for every $k = 2, 3, \ldots, t$, $a^k = (a_i^k)_{i\in N}$ and a_i^k is an available action to player i at stage k, for every $k = 1, 2, \ldots, t-1$. Then the game is equivalent to a game with state space $H := \bigcup_{t\in\mathbf{N}} H_t$, in which the state variable captures past play, and the state at stage t lies in H_t. In the new game, the sets of available actions, the payoff function, and the transition function, depend on the current state rather than on all past play.

The interested reader is referred to [20,42,72] for further reading on stochastic games. We now provide a few applications.

Example 1 (Capital Accumulation ([7,18,19,34,45])) Two (or more) agents jointly own a natural resource or a productive asset; at every period they have to decide the amount of the resource to consume. The amount that is not consumed grows by a known (or an unknown) fraction. Such a situation occurs, e. g., in fishery: Fishermen from various countries fish in the same area, and each country sets a quota for its fishermen. Here the state variable is the current amount of resource, the action set is the amount of resource to be exploited in the current period, and the transition is influenced by the decisions of all the players, as well as possibly by the random growth of the resource.

Example 2 (Taxation ([14,48])) A government sets a tax rate at every period. Each citizen decides at every period how much to work, and, from the total amount of money he or she has, how much to consume; the rest is saved for the next period, and grows by a known interest rate. Here the state is the amount of savings each citizen has, the stage payoff of a citizen depends on the amount of money that he consumed, on the amount of free time he has, and on the total amount of tax that the government collected. The stage payoff of the government may be the average stage payoff of the citizens, the amount of tax collected, or a mixture of the two.

Example 3 (Communication Network [58]) A single-cell system with one receiver and multiple uplink transmitters share a single, slotted, synchronous classical collision channel. Assume that all transmitted packets have the same length, and require one time unit, which is equal to one time slot, for transmission. Whenever a collision occurs, the users attempt to retransmit their packets in subsequent slots to resolve collision for reliable communication.

Here a state lists all relevant data for a given stage: e. g., the number of packets waiting at each transmitter, or the length of time each has been waiting to be transmitted. The players are the transmitters, and the action of each transmitter is which packet to transmit, if any. The stage cost may depend on the number of time slots that the transmitted packet waited, on the number of packets that have not been transmitted at that period, and possibly on additional variables. The transition depends on the actions chosen by the players, but it has a stochastic component, which captures the number of new packets that arrive at the various transmitters during every time slot.

Example 4 (Queues [1]) Individuals that require service have to choose whether to be served by a private slow service provider, or by a powerful public service provider. This situation arises, e. g., when jobs can be executed on either a slow personal computer or a fast mainframe. Here a state lists the current load of the public and private service providers, and the cost is the time to be served.

The importance of stochastic games stems from the wide range of applications they encompass. Many repeated interactions can be recast as stochastic games; the wide range of theoretical results that have been obtained provide insights that can help in analyzing specific situations and suggesting proper behavior to the participants. In certain classes of games algorithms that have been developed may be used to calculate such behavior.

Strategies, Evaluations and Equilibria

So far we have not described the information that the players have at each stage. In most of the chapter we assume that the players have complete information of past play; that is, at each stage t, they know the sequence $s^1, a^1, s^2, a^2, \ldots, s^t$ of states that were visited in the past (including the current state) and the actions that were chosen by all players. This assumption is too strong for most applications, and in the sequel we will mention the consequences of its relaxation.

Since the players observe past play, a *pure strategy* for player i is a (measurable) function σ_i that assigns to every finite history $(s^1, a^1, s^2, a^2, \ldots, s^t)$ an action $\sigma_i(s^1, a^1, s^2, a^2, \ldots, s^t) \in A_i(s^t)$, with the interpretation that, at stage t, if the finite history $(s^1, a^1, s^2, a^2, \ldots, s^t)$ occurred, player i plays the action $\sigma_i(s^1, a^1, s^2, a^2, \ldots, s^t)$. If the player does not know the complete history, then a strategy for player i is a function that assigns to every possible information set, an action that is available to the player when the player has this information. A *mixed strategy* for player i is a probability distribution over the set of his pure strategies. The space of mixed strategies of player i is denoted by σ_i.

A simple class of strategies is the class of *stationary* strategies; a strategy σ_i for player i is *stationary* if $\sigma_i(s^1, a^1, s^2, a^2, \ldots, s^t)$ depends only on the current state s^t, and not on past play $s^1, a^1, s^2, a^2, \ldots, a^{t-1}$. A stationary strategy of player i can be identified with an element $x = (x_s)_{s\in S} \in \times_{s\in S}\Delta(A_i(s))$, with the interpretation that player i plays the mixed action x_s whenever the current state is s. Denote by $X_i = \times_{s\in S}\Delta(A_i(s))$ the space of stationary strategies of player i.

There are three common ways to evaluate the infinite stream of payoffs that the players receive in a stochastic game: The *finite-horizon evaluation*, in which a player considers the average payoff during the first T stages, the *discounted evaluation*, in which a player considers the discounted sum of his stage payoffs, and the *limsup evaluation*, in which a player considers the limsup of his long-run average payoffs. We now formally define these evaluations.

Every profile $\sigma = (\sigma_i)_{i\in N}$ of mixed strategies, together with the initial state, induces a probability distribution $\mathbf{P}_{s_1,\sigma}$ over the space of infinite plays $H_\infty := SA^{\mathbf{N}}$. We denote the corresponding expectation operator by $\mathbf{E}_{s_1,\sigma}$.

Definition 5 Let σ be a profile of mixed strategies. For every finite horizon $T \in \mathbf{N}$, the *T-stage payoff* under σ for player i is

$$\gamma_i^T(s_1, \sigma) := \mathbf{E}_{s_1,\sigma}\left[\frac{1}{T}\sum_{t=1}^{T} u_i(s^t, a^t)\right].$$

For every discount factor $\lambda \in (0, 1]$, the *λ-discounted payoff* under σ for player i is

$$\gamma_i^{\lambda}(s_1, \sigma) := \mathbf{E}_{s_1,\sigma}\left[\lambda \sum_{t=1}^{\infty}(1-\lambda)^{t-1} u_i(s^t, a^t)\right] .$$

The *limsup payoff* under σ for player i is

$$\gamma_i^{\infty}(s_1, \sigma) := \mathbf{E}_{s_1,\sigma}\left[\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} u_i(s^t, a^t)\right] .$$

The T-stage payoff captures the situation in which the interaction lasts exactly T stages. The λ-discounted evaluation captures the situation in which the game lasts "many" stages, and the player discounts stage payoffs – it is better to receive \$1 today than tomorrow. The limsup payoff also captures the situation in which the game lasts "many" stages, but here the player does not discount his payoffs, and the payoff at each given stage is insignificant as compared to the payoff in all other stages. Equivalently, one could consider the liminf payoff in which the player considers the liminf of the long-run average payoffs.

As usual, an equilibrium is a vector of strategies such that no player can profit by a unilateral deviation. For every player i and every strategy profile $\sigma = (\sigma_i)_{i\in N}$ we denote the strategy profile of all other players, except player i, by $\sigma_{-i} = (\sigma_j)_{j\neq i}$.

Definition 6 Let $\varepsilon \geq 0$. A profile of strategies σ is a *T-stage ε-equilibrium* if

$$\gamma_i^T(s_1, \sigma) \geq \gamma_i^T(s_1, \sigma_i', \sigma_{-i}) - \varepsilon , \quad \forall s_1 \in S, \forall i \in N, \forall \sigma_i' \in \Sigma_i .$$

It is a *λ-discounted ε-equilibrium* if

$$\gamma_i^{\lambda}(s_1, \sigma) \geq \gamma_i^{\lambda}(s_1, \sigma_i', \sigma_{-i}) - \varepsilon , \quad \forall s_1 \in S, \forall i \in N, \forall \sigma_i' \in \Sigma_i .$$

It is a *limsup ε-equilibrium* if

$$\gamma_i^{\infty}(s_1, \sigma) \geq \gamma_i^{\infty}(s_1, \sigma_i', \sigma_{-i}) - \varepsilon , \quad \forall s_1 \in S, \forall i \in N, \forall \sigma_i' \in \Sigma_i .$$

The payoff that corresponds to an ε-equilibrium, that is, either one of the quantities $\gamma^T(s_1, \sigma)$, $\gamma^{\lambda}(s_1, \sigma)$ and $\gamma^{\infty}(s_1, \sigma)$, is called an *$\varepsilon$-equilibrium payoff* at the initial state s_1.

As we will see below, when both state and action spaces are finite, a T-stage and a λ-discounted 0-equilibrium exist. However, when the state or action spaces are infinite such a 0-equilibrium may fail to exist, yet ε-equilibria may exist for every $\varepsilon > 0$.

As the length of the game T varies, or as the discount factor λ varies, the equilibrium strategy profile varies as well. A strategy profile that is an ε-equilibrium for every T sufficiently large and for every λ sufficiently small is called a *uniform ε-equilibrium*.

Definition 7 Let $\varepsilon > 0$. A strategy profile σ is a *uniform ε-equilibrium* if there are $T_0 \in \mathbf{N}$ and $\lambda_0 \in (0, 1)$ such that for every $T \geq T_0$ the strategy profile σ is a T-stage ε-equilibrium, and for every $\lambda \in (0, \lambda_0)$ it is a λ-discounted ε-equilibrium.

If for every $\varepsilon > 0$ the game has a (T-stage, λ-discounted, limsup or uniform) ε-equilibrium with corresponding payoff g_ε, then any accumulation point of $(g_\varepsilon)_{\varepsilon>0}$ as ε goes to 0 is a (T-stage, λ-discounted, limsup or uniform) equilibrium payoff.

Zero-Sum Games

A two-player stochastic game is *zero-sum* if $u_1(s, a) + u_2(s, a) = 0$ for every $(s, a) \in SA$. As in matrix games, every two-player zero-sum stochastic game admits at most one equilibrium payoff at every initial state s_1, which is termed the *value* of the game at s_1. Each player's strategy which is part of an ε-equilibrium is termed *ε-optimal*. The definition of ε-equilibrium implies that an ε-optimal strategy guarantees the value up to ε; for example, in the T-stage evaluation, if σ_1 is an ε-optimal strategy of player 1, then for every strategy of player 2 we have

$$\gamma_1^T(s_1, \sigma_1, \sigma_2) \geq v^T(s_1) - \varepsilon ,$$

where $v^T(s_1)$ is the T-stage value at s_1.

In his seminal work, Shapley [60] presented the model of two-player zero-sum stochastic games with finite state and actions spaces, and proved the following.

Theorem 8 [60] *For every two-player zero-sum stochastic game, the λ-discounted value at every initial state exists. Moreover, both players have λ-discounted 0-optimal stationary strategies.*

Proof Let $\mathcal{V}$ be the space of all functions $v\colon S \to \mathbf{R}$. For every $v \in \mathcal{V}$ define a zero-sum matrix game $G_s^{\lambda}(v)$ as follows:

- The action spaces of the two players are $A_1(s)$ and $A_2(s)$ respectively.
- The payoff function (that player 2 pays player 1) is

$$\lambda u_1(s, a) + (1-\lambda)\sum_{s'\in S} q(s' \mid s, a) v(s') .$$

The game $G_s^\lambda(v)$ captures the situation in which, after the first stage, the game terminates with a terminal payoff $v(s')$, where s' is the state reached after stage 1. Define an operator $\varphi\colon \mathcal{V} \to \mathcal{V}$ as follows:

$$\varphi_s(v) = \mathrm{val}(G_s^\lambda(v)) ,$$

where $\mathrm{val}(G_s^\lambda(v))$ is the value of the matrix game $G_s^\lambda(v)$. Since the value operator is non-expansive, it follows that the operator φ is contracting: $\|\varphi(v) - \varphi(w)\|_\infty \le (1-\lambda)\|v - w\|_\infty$, so that this operator has a unique fixed point $\widehat{v}^\lambda$. One can show that the fixed point is the value of the stochastic game, and every strategy σ_i of player i in which he plays, after each finite history $(s^1, a^1, s^2, a^2, \ldots, s^t)$, an optimal mixed action in the matrix game $G_{s^t}^\lambda(\widehat{v}^\lambda)$, is a λ-discounted 0-optimal strategy in the stochastic game. □

Example 9 Consider the following two-player zero-sum game with three states, s_0, s_1 and s_2; each entry of the matrix indicates the payoff that player 2 (the column player) pays player 1 (the row player, the payoff is in the middle), and the transitions (which are deterministic, and are denoted at the top-right corner).

	L	R
T	$0\ s^2$	$1\ s^1$
B	$1\ s^1$	$0\ s^0$

State s_2

	L
T	$1\ s^1$

State s_1

	L
T	$0\ s^0$

State s_0 .

The states s_0 and s_1 are absorbing: Once the play reaches one of these states it never leaves it. State s_2 is non-absorbing. Stochastic games with a single non-absorbing state are called *absorbing games*. For every $v = (v_0, v_1, v_2) \in \mathcal{V} = \mathbf{R}^3$ the game $G_{s_2}^\lambda(v)$ is the following matrix game:

	L	R
T	$(1-\lambda)v_2$	$\lambda + (1-\lambda)v_1$
B	$\lambda + (1-\lambda)v_1$	$(1-\lambda)v_0$

The game $G_{s_2}^\lambda$

	L
T	$\lambda + (1-\lambda)v_1$

The game $G_{s_1}^\lambda$

	L
T	$(1-\lambda)v_0$

The game $G_{s_0}^\lambda$.

The unique fixed point of the operator $\mathrm{val}(G^\lambda)$ must satisfy

- $\widehat{v}_0 = \mathrm{val}(G_{s_0}^\lambda(\widehat{v}))$, so that $\widehat{v}_{s_0}^\lambda = \widehat{v}_0 = 0$;
- $\widehat{v}_1 = \mathrm{val}(G_{s_1}^\lambda(\widehat{v}))$, so that $\widehat{v}_{s_1}^\lambda = \widehat{v}_1 = 1$;
- $\widehat{v}_2 = \mathrm{val}(G_{s_1}^\lambda(\widehat{v}))$. By Theorem 8 both players have a stationary λ-discounted 0-optimal strategy. Denote by x (resp. y) a mixed action for player 1 (resp. player 2) that is part of a λ-discounted 0-optimal strategy at the state s_2. Since we know that in the fixed point $\widehat{v}_0 = 0$ and $\widehat{v}_1 = 1$, $\widehat{v}_2$ must be the unique solution of

  $$v_2 = y(1-\lambda)v_2 + (1-y) = y ,$$

  so that $\widehat{v}_{s_2}^\lambda = \widehat{v}_2 = (1-\sqrt{\lambda})/(1-\lambda)$. The 0-optimal strategy of player 2 at state s_2 is $y = \widehat{v}_2 = (1-\sqrt{\lambda})/(1-\lambda)$, and the 0-optimal strategy of player 1, $x = \widehat{v}_2 = (1-\sqrt{\lambda})/(1-\lambda)$, can be found by finding his 0-optimal strategy in $G_{s_2}^\lambda(\widehat{v})$.

Bewley and Kohlberg [11] proved that when the state and action spaces are finite, the function $\lambda \mapsto v_s^\lambda$, that assigns to every state s and every discount factor λ the λ-discounted value at the initial state s, is a *Puiseux function*, that is, it has a representation $v_s^\lambda = \sum_{k=K}^\infty a_k \lambda^{k/M}$ that is valid for every $\lambda \in (0, \lambda_0)$ for some $\lambda_0 > 0$, where M is a natural number, K is a non-negative integer, and $(a_k)_{k=K}^\infty$ are real numbers. In particular, the function $\lambda \mapsto v_s^\lambda$ is monotone in a neighborhood of 0, and its limit as λ goes to 0 exists. This result turned out to be crucial in subsequent study on games with finitely many states and actions.

Shapley's work has been extended to general state and action spaces; for a recent survey see [46]. The tools developed in [46], together with a dynamic programming argument, prove that under proper conditions on the payoff function and on the transitions the two-player zero-sum stochastic game has a T-stage value.

Maitra and Sudderth [35] proved that the limsup value exists in a very general setup. Their proof follows closely that of Martin [36] for the determinacy of Blackwell games.

The study of the uniform value emanated from an example, called the "Big Match", due to Gillette [28], that was solved by Blackwell and Ferguson [13].

Example 10 Consider the following stochastic game with two absorbing states and one non-absorbing state.

	L	R
T	$0\ s^2$	$1\ s^2$
B	$1\ s^1$	$0\ s^0$

State s_2

	L
T	$1\ s^1$

State s_1

	L
T	$0\ s^0$

State s_0

Suppose the initial state is s_2. As long as player 1 plays T the play remains at s_2; once he plays B the play moves to either s_0 or s_1, and is effectively terminated. By finding the fixed point of the operator φ one can show

that the discounted value at the initial state s_2 is $\frac{1}{2}$, and a λ-discounted stationary 0-optimal strategy for player 2 is[1] $[\frac{1}{2}(L), \frac{1}{2}(R)]$. Indeed, if player 1 plays T then the expected stage payoff is $\frac{1}{2}$ and play remains at s_2, while if player 1 plays B then the game moves to an absorbing state, and the expected stage payoff from that stage onwards is $\frac{1}{2}$. In particular, this strategy guarantees $\frac{1}{2}$ for player 2 both in the limsup evaluation and uniformly. A λ-discounted 0-optimal strategy for player 1 is $[\frac{1}{1+\lambda}(T), \frac{\lambda}{1+\lambda}(B)]$.

What can player 1 guarantee in the limsup evaluation and uniformly? If player 1 plays the stationary strategy $[x(T), (1-x)(B)]$ that plays at each stage the action T with probability x and the action B with probability $1-x$, then player 2 has a reply that ensures that the limsup payoff is 0: If $x = 1$ and player 2 always plays L, the payoff is 0 at each stage; if $x < 1$ and player 2 always plays R, the payoff is 1 until the play moves to s_0, and then it is 0 forever. Since player 1 plays the action B with probability $1 - x > 0$ at each stage, the distribution of the stage in which play moves to s_0 is geometric. Therefore, the limsup payoff is 0, and if λ is sufficiently small, the discounted payoff is close to 0.

One can verify that if player 1 uses a bounded-recall strategy, that is, a strategy that uses only the last k actions that were played, player 2 has a reply that guarantees that the limsup payoff is 0, and the discounted payoff is close to 0, provided λ is close to 0. Thus, in the limsup payoff and uniformly finite memory cannot guarantee more than 0 in this game (see also [27]).

Intuitively, player 1 would like to condition the probability of playing T on the past behavior of player 2: If in the past player 2 played the action L more often than the action R, he would have liked to play T with higher probability; if in the past player 2 played the action R more often than the action L, he would have liked to play B with higher probability. Blackwell and Ferguson [13] constructed a family of good strategies $\{\sigma_1^M, M \in \mathbf{N}\}$ for player 1. The parameter M determines the amount that the strategy guarantees: The strategy σ_1^M guarantees a limsup payoff and a discounted payoff of $\frac{M}{2M+1}$, provided the discount factor is sufficiently low. In other words, player 1 cannot guarantee $\frac{1}{2}$, but he may guarantee an amount as close to $\frac{1}{2}$ as he wishes by choosing M to be sufficiently large. The strategy σ_1^M is defined as follows: At stage t, play B with probability $\frac{1}{(M+l_t-r_t)^2}$, where l_t is the number of stages up to stage t in which player 2 played L, and r_t is the number of stages up to stage t in which player 2 played R.

[1] That is, at each stage player 2 plays L with probability $\frac{1}{2}$ and R with probability $\frac{1}{2}$.

Since $r_t + l_t = t - 1$ one has $r_t - l_t = 2r_t - (t-1)$. The quantity r_t is the total payoff that player 1 received in the first $t-1$ stages if player 1 played T in those stages (and the game was not absorbed). Thus, this total payoff is a linear function of the difference $r_t - l_t$. When presented this way, the strategy σ_1^M depends on that total payoff. Observe that as r_t increases, $r_t - l_t$ increases as well, and the probability to play B decreases.

Mertens and Neyman [38] generalized the idea presented at the end of Example 10 to stochastic games with finite state and action spaces.[2]

Theorem 11 *If the state and action spaces of a two-player zero-sum stochastic game are finite, the game has a uniform value v_s^0 at every initial state $s \in S$. Moreover, $v_s^0 = \lim_{\lambda\to 0} v_s^\lambda = \lim_{T\to\infty} v_s^T$.*

In their proof, Mertens and Neyman describe a uniform ε-optimal strategy. In this strategy the player keeps a parameter, λ_t, which is a fictitious discount factor to use at stage t. This parameter changes at each stage as a function of the stage payoff; if the stage payoff at stage t is high then $\lambda_{t+1} < \lambda_t$, whereas if the stage payoff at stage t is low then $\lambda_{t+1} > \lambda_t$. The intuition is as follows. As mentioned before, in stochastic games there are two forces that influence the player's behavior: He tries to get high stage payoffs, while keeping future prospects high (by playing in such a way that the next stage that is reached is favorable). When considering the λ-discounted payoff there is a clear comparison between the importance of the two forces: The weight of the stage payoff is λ and the weight of future prospects is $1-\lambda$; the lower the discount factor, the more weight is given to the future. When considering the uniform value (or the uniform equilibrium) the weight of the stage payoff is 0. However, if the player never attempts to receive a high stage payoff, the overall payoff in the game will not be high. Therefore, the player has a fictitious discount factor; if past payoffs are low and they do not meet the expectation, player 1 increases the weight of the stage payoff by increasing the fictitious discount factor; if past payoffs are high, player 1 increases the weight of the future by lowering this fictitious discount factor.

Multi-Player Games

Takahashi [75] and Fink [21] extended Shapley's [60] result to discounted equilibria in non-zero-sum games.

[2] Mertens and Neyman's [38] result actually holds in every stochastic game that satisfies a proper condition, which is always satisfied when the state and action spaces are finite.

Theorem 12 *Every stochastic game with finite state and action spaces has a λ-discounted equilibrium in stationary strategies.*

Proof The proof utilizes Kakutani's fixed point theorem [31]. Let $M = \max_{i,s,a} |u_i(s,a)|$ be a bound on the absolute values of the payoffs. Set $X = \times_{i\in N, s\in S} (\Delta(A_i(s)) \times [-M, M])$. A point $x = (x^A_{i,s}, x^V_{i,s})_{i\in N, s\in S} \in X$ is a collection of one mixed action and one payoff to each player at every state. For every $v = (v_i)_{i\in N} \in [-M, M]^{N\times S}$ and every $s \in S$ define a matrix game $G^\lambda_s(v)$ as follows:

- The action spaces of each player i is $A_i(s)$;
- The payoff to player i is

$$\lambda u_i(s,a) + (1-\lambda) \sum_{s'\in S} q(s' \mid s, a) v_i(s') .$$

We define a set-valued function $\varphi\colon X \to X$ as follows.

- For every $i \in N$ and every $s \in S$, $\varphi^A_{i,s}$ is the set of all best responses of player i to the strategy vector $x_{-i,s} := (x_{j,s})_{j\neq i}$ in the game $G^\lambda_s(v)$. That is,

$$\varphi^A_{i,s}(x,v) := \Big\{ \operatorname{argmax}_{y_{i,s}\in\Delta(A_i(s))} \lambda r_i(s, y_{i,s}, x_{-i,s}) + (1-\lambda) \sum_{s'\in S} q(s' \mid s, y_{i,s}, x_{-i,s}) v_{i,s'} \Big\} .$$

- For every $i \in N$ and every $s \in S$, $\varphi^V_{i,s}(x,v)$ is the maximal payoff for player i in the game $G^\lambda_s(v)$, when the other players play x_{-i}:

$$\varphi^V_s(x,v) := \max_{y_{i,s}\in\Delta(A_i(s))} \left(\lambda r(s, y_{i,s}, x_{-i,s}) + (1-\lambda) \times \sum_{s'\in S} q(s' \mid s, y_{i,s}, x_{-i,s}) v_{i,s'} \right) .$$

The set-valued function φ has convex and non-empty values and its graph is closed, so that by Kakutani's fixed point theorem it has a fixed point. It turns out that every fixed point of φ defines a λ-discounted equilibrium in stationary strategies. □

This result has been extended to general state and action spaces by various authors. These results assume a strong continuity on either the payoff function or the transition function, see, e. g., [39,44,62].

As in the case of zero-sum games, a dynamic programming argument shows that under a strong continuity assumption on the payoff function or on the transitions a T-stage equilibrium exists.

Regarding the existence of the limsup equilibrium and uniform equilibrium little is known. The most significant result in this direction is Vieille [77,78], who proved that every two-player stochastic game with finite state and action spaces has a uniform ε-equilibrium for every $\varepsilon > 0$. This result has been proved for other classes of stochastic games, see, e. g., [6,24,25,61,63,65,76]. Several influential works in this area are [22,32,71,79]. Most of the papers mentioned above rely on the vanishing discount factor approach, which constructs a uniform ε-equilibrium by studying a sequence of λ-discounted equilibria as the discount factor goes to 0.

For games with general state and action spaces, a limsup equilibrium exists under an ergodicity assumption on the transitions, see e. g. Nowak [44], Remark 4 and Jaskiewicz and Nowak [30].

A game has perfect information if there are no simultaneous moves, and both players observe past play. Existence of equilibrium in this case was proven by Mertens [37] in a very general setup.

Correlated Equilibrium

The notion of correlated equilibrium was introduced by Aumann [8,9], see also Forges ▶ Correlated Equilibria and Communication in Games. A *correlated equilibrium* is an equilibrium of an extended game, in which each player receives at the outset of the game a private signal such that the vector of signals is chosen according to a known joint probability distribution. In repeated interactions, such as in stochastic games, there are two natural notions of correlated equilibria: (a) each player receives one signal at the outset of the game (*normal-form correlated equilibrium*); (b) each player receives a signal at each stage (*extensive-form correlated equilibrium*). It follows from Forges [26] that when the state and action sets are finite, the set of all correlated T-stage equilibrium payoffs (either normal-form or extensive-form) is a polytope.

Nowak and Raghavan [47] proved the existence of an extensive-form correlated discounted equilibrium under weak conditions on the state and action spaces. In their construction, the strategies of the players are stationary, and so is the distribution according to which the signals are chosen after each history; both depend only on the current state, rather than on the whole past play. Roughly, their approach is to apply Kakutani's fixed point theorem to the set-valued function that assigns to each game $G^\lambda_s(v)$ the set of all correlated equilibrium payoffs in this game, which is convex and compact.

Solan and Vielle [66] proved the existence of an extensive-form correlated uniform equilibrium payoff when the

state and action spaces are finite. Their approach is to let each player play his uniform optimal strategy in a zero-sum game in which all other players try to minimize his payoff. Existence of a normal-form correlated equilibrium was proved for the class of absorbing games [68].

Solan [64] characterized the set of extensive-form correlated equilibrium payoffs for general state and action spaces and a general evaluation on the stage payoffs, and provided a sufficient condition that ensures that the set of normal-form correlated equilibrium payoffs coincides with the set of extensive-form correlated equilibrium payoffs.

Imperfect Monitoring

So far it has been assumed that at each stage the players know the past play. There are cases in which this assumption is too strong; in some cases players do not know the complete description of the current state (Examples 3 and 4), and in others players do not fully observe the actions of all other players (Examples 2, 3 and 4). For a most general description of stochastic games see Mertens, see Chapter IV in Sorin and Zamir [40] and Coulomb [17].

In the study of the discounted equilibrium, the T-stage equilibrium or the limsup equilibrium, one may consider the game as a one-shot game: The players simultaneously choose strategies, and the payoff is either the discounted payoff, the T-stage payoff or the limsup payoff. If the strategy spaces of the players are compact (e. g., if the state and action spaces are finite), and if the payoff is upper-semi-continuous in each player's strategy, keeping the strategies of the other players fixed, then an equilibrium exists. This approach can be used successfully for the discounted equilibrium or the T-stage equilibrium under weak conditions (see, e. g., [4]), and may be used for the limsup equilibrium under a proper ergodicity condition.

Whenever there exists an equilibrium in stationary strategies (e. g., a discounted equilibrium in games with finitely many states and actions) the only information that players need in order to follow the equilibrium strategies is the current state. In particular, they need not observe past actions of the other players. As we now show, in the "Big Match" (Example 10) the limsup value and the uniform value may fail to exist when each player does not observe the past actions of the other player.

Example 13 (Example 10: Continued.) Assume that no player observes the actions of the other player, and assume that the initial state is s_2. Player 2 can still guarantee $\frac{1}{2}$ in the limsup evaluation by playing the stationary strategy $[\frac{1}{2}(L), \frac{1}{2}(R)]$. One can show that for every strategy of player 2, player 1 has a reply such that the limsup payoff is at least $\frac{1}{2}$. In other words, $\inf_{\sigma_2} \sup_{\sigma_1} \gamma^{\infty}(s_2, \sigma_1, \sigma_2) = \frac{1}{2}$. We now argue that $\sup_{\sigma_1} \inf_{\sigma_2} \gamma^{\infty}(s_2, \sigma_1, \sigma_2) = 0$. Indeed, fix a strategy σ_1 for player 1, and $\varepsilon > 0$. Let θ be sufficiently large such that the probability that under σ_1 player 1 plays B for the first time after stage t is at most ε. Observe that as t increases, the probability that player 1 plays B for the first time after stage t decreases to 0, so that such a θ exists. Consider the following strategy σ_2 of player 2: Play R up to stage θ, and play L from stage $t+1$ and on. By the definition of θ, either player 1 plays B before or at stage θ, and then the game moves to s_0, and the payoff is 0 at each stage thereafter, or player 1 plays T at each stage, and then the stage payoff after stage θ is 0, or, with probability less than ε, player 1 plays B for the first time after stage θ, the play moves to s_1, and the payoff is 1 thereafter. Thus, the limsup payoff is at most ε. A similar analysis shows that

$$\inf_{\sigma_2} \sup_{\sigma_1} \gamma^{\lambda}(s_2, \sigma_1, \sigma_2) = \tfrac{1}{2} ,$$

$$\sup_{\sigma_1} \inf_{\sigma_2} \lim_{\lambda \to 0} \gamma^{\lambda}(s_2, \sigma_1, \sigma_2) = 0 ,$$

so that the uniform value does not exist as well.

This example shows that in general the limsup value and the uniform value need not exist when the players do not observe past play. Though in general the value (and therefore also an equilibrium) need not exist, in many classes of stochastic games the value and an equilibrium do exist, even in the presence of imperfect monitoring.

Rosenberg et al. [55] and Renault [54] showed that the uniform value exists in the one player setup (Markov Decision Problem), in which the player receives partial information regarding the current state. Thus, a single decision maker who faces a dynamic situation and does not fully observe the state of the environment can play in such a way that guarantees high payoff, provided the interaction is sufficiently long or the discount factor is sufficiently low.

Altman et al. [5,6] and Flesch et al. [24] studied stochastic games in which each player has a "private" state, which only he can observe, and the state of the world is composed of the vector of private states. Altman et al. [5,6] studied the situation in which players do not observe the actions of the other players, and Flesch et al. [24] studied the situation in which players do observe each others payoffs. Such games arise naturally in wireless communication (see [5]); take for example several mobiles who periodically send information to a base station. The private state of a mobile may depend, e. g., on its exact physical environment, and it determines the power attenuation between the mobile and the base station. The throughput (the amount of bits per second) that a mobile can send to the base station depends on the power attenuations of

all the mobiles. Finally, the stage payoff is the stage power consumption.

Rosenberg et al. [57] studied the extreme case of two player zero-sum games in which the players observe neither the current state nor the action of the other player, and proved that the uniform value does exist in two classes of games, which capture the behavior of certain communication protocols. Classes of games in which the actions are observed but the state is not observed were studied, e. g., by Sorin [69,70], Sorin and Zamir [74], Krausz and Rieder [33], Flesch et al. [23], Rosenberg et al. [56], Renault [52,53]. For additional results, see [16,73].

Algorithms

There are two kinds of algorithms: Those that terminate in a finite number of steps, and those that iterate and approximate solutions. Both kinds of algorithms were devised to calculate the value and optimal strategies (or equilibria) in stochastic games.

It is well known that the value of a two-player zero-sum matrix game and optimal strategies for the two players can be calculated efficiently using a linear program. Equilibria in two-player non-zero-sum games can be calculated by the Lemke–Howson algorithm, which is usually efficient, however, its worst running time is exponential in the number of pure strategies of the players [59]. Unfortunately, to date there are no efficient algorithms to calculate either the value in zero-sum stochastic games, or equilibria in non-zero-sum games. Moreover, in Example 9 the discounted value may be irrational for rational discount factors, even though the data of the game (payoffs and transitions) are rational, so it is not clear whether linear programming methods can be used to calculate the value of a stochastic game. Nevertheless, linear programming methods were used to calculate the discounted and uniform value of several classes of stochastic games, see [20,50]. Other methods that were used to calculate the value or equilibria in discounted stochastic games include fictitious play [80], value iterates, policy improvement, and general methods to find the maximum of a function (see [20,51]), a homotopy method [29], and algorithms to solve sentences in formal logic [15,67].

Additional and Future Directions

The research on stochastic games extends to additional directions than those mentioned in earlier sections. We mention a few here. Approximation of games with infinite state and action spaces by finite games was discussed by Whitt [81], and further developed by Nowak [43]. Stochastic games in continuous time have also been studied, as well as hybrid models that include both discrete and continuous aspects, see, e. g., [2,10].

Among the many directions of future research in this area, we will mention here but a few. One challenging question is the existence of a uniform equilibrium and a limsup equilibrium in multi-player stochastic games with finite state and action spaces. Another is the development of efficient algorithms that calculate the value of two-player zero-sum games. A third direction concerns the identification of applications that can be recast in the framework of stochastic games, and that can be successfully analyzed using the theoretical tools that the literature developed. Another problem that is of interest is the characterization of approachable and excludable sets in stochastic games with vector payoffs (see [12] for the presentation of matrix games with vector payoffs, and [41] for partial results regarding this problem).

Acknowledgments

I thank Eitan Altman, János Flesch, Yuval Heller, Jean-Jacques Herings, Ayala Mashiach-Yakovi, Andrzej Nowak, Ronald Peeters, T.E.S. Raghavan, Jérôme Renault, Nahum Shimkin, Robert Simon, Sylvain Sorin, William Sudderth, and Frank Thuijmsman, for their comments on an earlier version of the entry.

Bibliography

Primary Literature

1. Altman E (2005) Applications of dynamic games in queues. Adv Dyn Games 7:309–342
2. Altman E, Gaitsgory VA (1995) A hybrid (differential-stochastic) zero-sum game with fast stochastic part. Ann Int Soc Dyn Games 3:47–59
4. Altman E, Solan E (2007) Games with constraints with networking applications. Preprint
5. Altman E, Avrachenkov K, Marquez R, Miller G (2005) Zero-sum constrained stochastic games with independent state processes. Math Methods Oper Res 62:375–386
6. Altman E, Avrachenkov K, Bonneau N, Debbah M, El-Azouzi R, Sadoc Menasche D (2008) Constrained cost-coupled stochastic games with independent state processes. Oper Res Lett 36:160–164
7. Amir R (1996) Continuous stochastic games of capital accumulation with convex transitions. Games Econ Behav 15:111–131
8. Aumann RJ (1974) Subjectivity and correlation in randomized strategies. J Math Econ 1:67–96
9. Aumann RJ (1987) Correlated equilibrium as an expression of bayesian rationality. Econometrica 55:1–18
10. Başar T, Olsder GJ (1995) Dynamic noncooperative game theory. Academic Press, New York
11. Bewley T, Kohlberg E (1976) The asymptotic theory of stochastic games. Math Oper Res 1:197–208

12. Blackwell D (1956) An analog of the minimax theorem for vector payoffs. Pac J Math 6:1–8
13. Blackwell D, Ferguson TS (1968) The big match. Ann Math Stat 39:159–163
14. Chari V, Kehoe P (1990) Sustainable plans. J Political Econ 98:783–802
15. Chatterjee K, Majumdar R, Henzinger TA (2008) Stochastic limit-average games are in EXPTIME. Int J Game Theory 37:219–234
16. Coulomb JM (2003) Absorbing games with a signalling structure. In: Neyman A, Sorin S (eds) Stochastic games and applications. NATO Science Series. Kluwer, Dordrecht, pp 335–355
17. Coulomb JM (2003) Games with a recursive structure. In: Neyman A, Sorin S (eds) Stochastic games and applications. NATO Science Series. Kluwer, Dordrecht, pp 427–442
18. Dutta P, Sundaram RK (1992) Markovian equilibrium in a class of stochastic games: Existence theorems for discounted and undiscounted models. Econ Theory 2:197–214
19. Dutta P, Sundaram RK (1993) The tragedy of the commons? Econ Theory 3:413–426
20. Filar JA, Vrieze K (1996) Competitive Markov decision processes. Springer
21. Fink AM (1964) Equilibrium in a stochastic *n*-person game. J Sci Hiroshima Univ 28:89–93
22. Flesch J, Thuijsman F, Vrieze K (1997) Cyclic Markov equilibria in stochastic games. Int J Game Th 26:303–314
23. Flesch J, Thuijsman F, Vrieze OJ (2003) Stochastic games with non-observable actions. Math Meth Oper Res 58:459–475
24. Flesch J, Schoenmakers G, Vrieze K (2008) Stochastic games on a product state space. Math Oper Res 33:403–420
25. Flesch J, Thuijsman F, Vrieze OJ (2007) Stochastic games with additive transitions. Europ J Oper Res 179:483–497
26. Forges F (1990) Universal mechanisms. Econometrica 58:1341–1364
27. Fortnow L, Kimmel P (1998) Beating a finite automaton in the big match. In: Proceedings of the 7th conference on theoretical aspects of rationality and knowledge. Morgan Kaufmann, San Francisco, pp 225–234
28. Gillette D (1957) Stochastic games with zero stop probabilities, contributions to the theory of games, vol 3. Princeton University Press, Princeton
29. Herings JJP, Peeters RJAP (2004) Stationary equilibria in stochastic games: Structure, selection, and computation. J Econ Theory 118:32–60
30. Jaskiewicz A, Nowak AS (2006) Zero-sum ergodic stochastic games with Feller transition probabilities. SIAM J Control Optim 45:773–789
31. Kakutani S (1941) A generalization of Brouwer's fixed point theorem. Duke Math J 8:457–459
32. Kohlberg E (1974) Repeated games with absorbing states. Ann Stat 2:724–738
33. Krausz A, Rieder U (1997) Markov games with incomplete information. Math Meth Oper Res 46:263–279
34. Levhari D, Mirman L (1980) The great fish war: An example using a dynamic Cournot–Nash solution. Bell J Econ 11(1):322–334
35. Maitra A, Sudderth W (1998) Finitely additive stochastic games with Borel measurable payoffs. Int J Game Theory 27:257–267
36. Martin DA (1998) The determinacy of Blackwell games. J Symb Logic 63:1565–1581
37. Mertens JF (1987) Repeated games. In: Proceedings of the international congress of mathematicians, American Mathematical Society, Berkeley, California, pp 1528–1577
38. Mertens JF, Neyman A (1981) Stochastic games. Int J Game Th 10:53–66
39. Mertens JF, Parthasarathy T (1987) Equilibria for discounted stochastic games, CORE Discussion Paper No. 8750. (Also published in Stochastic Games and Applications, Neyman A, Sorin S (eds), NATO Science Series, Kluwer, 131–172)
40. Mertens JF, Sorin S, Zamir S (1994) Repeated games, CORE Discussion Paper 9420-9422
41. Milman E (2006) Approachable sets of vector payoffs in stochastic games. Games Econ Behav 56:135–147
42. Neyman A, Sorin S (2003) Stochastic games and applications. NATO Science Series. Kluwer
43. Nowak AS (1985) Existence of equilibrium stationary strategies in discounted noncooperative stochastic games with uncountable state space. J Optim Theory Appl 45:591–620
44. Nowak AS (2003) *N*-person stochastic games: Extensions of the finite state space case and correlation. In: Neyman A, Sorin S (eds) Stochastic games and applications. NATO Science Series. Kluwer, Dordrecht, pp 93–106
45. Nowak AS (2003) On a new class of nonzero-sum discounted stochastic games having stationary Nash equilibrium points. Int J Game Theory 32:121–132
46. Nowak AS (2003) Zero-sum stochastic games with Borel state spaces. In: Neyman A, Sorin S (eds) Stochastic games and applications. NATO Science Series. Kluwer, Dordrecht, pp 77–91
47. Nowak AS, Raghavan TES (1991) Existence of stationary correlated equilibria with symmetric information for discounted stochastic games. Math Oper Res 17:519–526
48. Phelan C, Stacchetti E (2001) Sequential equilibria in a Ramsey tax model. Econometrica 69:1491–1518
49. Puterman ML (1994) Markov decision processes: Discrete stochastic dynamic programming. Wiley, Hoboken
50. Raghavan TES, Syed Z (2002) Computing stationary Nash equilibria of undiscounted single-controller stochastic games. Math Oper Res 27:384–400
51. Raghavan TES, Syed Z (2003) A policy improvement type algorithm for solving zero-sum two-person stochastic games of perfect information. Math Program Ser A 95:513–532
52. Renault J (2006) The value of Markov chain games with lack of information on one side. Math Oper Res 31:490–512
53. Renault J (2007) The value of repeated games with an informed controller. Preprint
54. Renault J (2007) Uniform value in dynamic programming. Preprint
55. Rosenberg D, Solan E, Vieille N (2002) Blackwell optimality in Markov decision processes with partial observation. Ann Statists 30:1178–1193
56. Rosenberg D, Solan E, Vieille N (2004) Stochastic games with a single controller and incomplete information. SIAM J Control Optim 43:86–110
57. Rosenberg D, Solan E, Vieille N (2006) Protocol with no acknowledgement. Oper Res, forthcoming
58. Sagduyu YE, Ephremides A (2003) Power control and rate adaptation as stochastic games for random access. Proc 42nd IEEE Conf Decis Control 4:4202–4207
59. Savani R, von Stengel B (2004) Exponentially many steps for finding a Nash equilibrium in a bimatrix game. Proc 45th Ann IEEE Symp Found Comput Sci 2004:258–267

60. Shapley LS (1953) Stochastic games. Proc Nat Acad Sci USA 39:1095–1100
61. Simon RS (2003) The structure of non-zero-sum stochastic games. Adv Appl Math 38:1–26
62. Solan E (1998) Discounted stochastic games. Math Oper Res 23:1010–1021
63. Solan E (1999) Three-person absorbing games. Math Oper Res 24:669–698
64. Solan E (2001) Characterization of correlated equilibria in stochastic games. Int J Game Theory 30:259–277
65. Solan E, Vieille N (2001) Quitting games. Math Oper Res 26:265–285
66. Solan E, Vieille N (2002) Correlated equilibrium in stochastic games. Games Econ Behav 38:362–399
67. Solan E, Vieille N (2007) Calculating uniform optimal strategies and equilibria in two-player stochastic games. Preprint
68. Solan E, Vohra R (2002) Correlated equilibrium payoffs and public signalling in absorbing games. Int J Game Theory 31:91–122
69. Sorin S (1984) Big match with lack of information on one side (part 1). Int J Game Theory 13:201–255
70. Sorin S (1985) Big match with lack of information on one side (part 2). Int J Game Theory 14:173–204
71. Sorin S (1986) Asymptotic properties of a non-zerosum stochastic games. Int J Game Theory 15:101–107
72. Sorin S (2002) A first course on zero-sum repeated games. Mathématiques et Applications, vol 37. Springer
73. Sorin S (2003) Stochastic gameswith incomplete information. In: Neyman A, Sorin S (eds) Stochastic Games and Applications. NATO Science Series. Kluwer, Berlin, pp 375–395
74. Sorin S, Zamir S (1991) Big match with lack of information on one side (part 3). In: Raghavan TES et al (eds) Stochastic games and related topics. Kluwer, pp 101–112
75. Takahashi M (1962) Stochastic games with infinitely many strategies. J Sci Hiroshima Univ Ser A-I 26:123–134
76. Thuijsman F, Raghavan TES (1997) Perfect information stochastic games and related classes. Int J Game Theory 26:403–408
77. Vieille N (2000) Equilibrium in 2-person stochastic games I: A Reduction. Israel J Math 119:55–91
78. Vieille N (2000) Equilibrium in 2-person stochastic games II: The case of recursive games. Israel J Math 119:93–126
79. Vrieze OJ, Thuijsman F (1989) On equilibria in repeated games with absorbing states. Int J Game Theory 18:293–310
80. Vrieze OJ, Tijs SH (1982) Fictitious play applied to sequences of games and discounted stochastic games. Int J Game Theory 12:71–85
81. Whitt W (1980) Representation and approximation of noncooperative sequential games. SIAM J Control Optim 18:33–48

Books and Reviews

Başar T, Olsder GJ (1995) Dynamic noncooperative game theory. Academic
Filar JA, Vrieze K (1996) Competitive Markov decision processes. Springer
Maitra AP, Sudderth WD (1996) Discrete gambling and stochastic games. Springer
Mertens JF (2002) Stochastic games. In: Aumann RJ, Hart S (eds) Handbook of game theory with economic applications, vol 3. Elsevier, pp 1809–1832
Mertens JF, Sorin S, Zamir S (1994) Repeated games. CORE Discussion Paper 9420-9422
Raghavan TES, Shapley LS (1991) Stochastic games and related topics: In honor of Professor L.S. Shapley. Springer
Vieille N (2002) Stochastic games: Recent results. In: Aumann RJ, Hart S (eds) Handbook of game theory with economic applications, vol 3. Elsevier, pp 1833–1850

Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems

Hans C. Fogedby[1,2]
[1] Department of Physics and Astronomy, University of Aarhus, Aarhus, Denmark
[2] Niels Bohr Institute, Copenhagen, Denmark

Article Outline

Glossary
Definition of the Subject
Introduction
Scaling
Conformal Invariance
Loewner Evolution
Stochastic Loewner Evolution
Results and Discussion
Future Directions
Acknowledgments
Bibliography

Glossary

Bessel process The Bessel process operates in d dimensions and describes the radial distance r from the origin of a particle performing Brownian motion. The random motion is governed by the Langevin equation $\mathrm{d}r/\mathrm{d}t = \kappa(d-1)/2r + \xi$, where $\langle \xi\xi \rangle(t) = \kappa\delta(t)$. For $d \leq 2$ the motion is recurrent, i. e., returns to the origin; for $d > 2$ the motion goes off to infinity. In SLE the Bessel process describes the transition between simple curves and self-intersecting curves.

Brownian motion Brownian motion, B_t, is the scaling limit of random walk. Brownian motion is plane-filling, has the fractal dimension $D = 2$, and is described by the Langevin equation $\mathrm{d}B_t/\mathrm{d}t = \eta_t, \langle \eta_t \eta_s \rangle = \delta(t-s)$. B_t is characterized by i) the stationarity property, $B_{t+t'} - B_t$ and $B_{t'}$ identical in distribution and ii) the independence property, $B_{\Delta t'}$ and $B_{\Delta t}$ independent for $\Delta t \neq \Delta t'$. The correlations are given by

$\langle |B_t - B_s|^2 \rangle = |t - s|$ and B_t is distributed according to the Gaussian (normal) distribution $P(B, t) = (2\pi t)^{-1/2} \exp(-B^2/2t)$.

Chordal SLE In order to map the geometry of the growing hull to a real function by means of SLE one chooses a reference domain. Chordal SLE refers to the case where the SLE trace is grown between two boundary points, usually the origin and the point at infinity in the upper half complex plane.

Conformal invariance Conformal invariance or local scale invariance is a larger symmetry than scale invariance. Conformal invariance implies invariance under both a local rotation, translation, and dilatation. Conformal invariance is particularly powerful in 2D where a conformal transformation is implemented by an analytic function.

Conformal transformation A conformal transformation is a transformation which preserves angles but allows for rotation, dilatation, and translation. In an elastic medium picture a conformal transformation corresponds to translation, rotation, and compression (dilatation), but not shear. In 2D an analytic complex function $w = f(z)$ from the complex z plane to the complex w plane generates a conformal transformation.

Continuum limit The limit where a lattice model approaches a continuum model; also called the scaling limit. The resulting continuum field theories define the universality classes of the lattice models.

Correlation length The correlation length measures the size of correlations. At the critical point the correlation length diverges signaling that the system becomes scale invariant.

Critical curves Critical curves are domain walls or cluster boundaries at the critical point. Critical curves are scale invariant and characterized by a fractal dimension.

Exploration Domain walls at the critical point can mathematically be generated by an exploration process where the domain wall is initiated at a boundary point and constructed in steps across the domain. In the percolation case the local step is generated by 'flipping a coin', in the Ising case by evaluating the magnetization at the tip of the 'growing' domain wall. The construction by an exploration process is essential in generating a critical curve by stochastic Loewner evolution.

Fortuin–Kasteleyn representation (FK) Based on a high temperature expansion one can represent the configurations in the Ising, $O(n)$, and Potts models by means of the Fortuin–Kasteleyn representation in terms of random clusters. The FK transformation is essential in identifying random curves for the lattice models which then can be accessed by SLE in the scaling limit.

Fractal dimension Irregular objects can be characterized by a fractal dimension. The fractal dimension D is derived by covering the object with $N(\ell)$ intervals, disks, or spheres of linear dimension ℓ. By letting $\ell \to 0$ this definition resolves the fine scale fractal structure and the scaling relation $N(\ell) \propto (\ell)^{-D}$, or $D = -\lim_{\ell \to 0} \ln(N(\ell)/\ln(\ell)$, yields the fractal dimension. Examples of deterministic fractals are for example the Cantor set, the Koch curve, the Sierpinski gasket, and plane-filling Hilbert and Peano curves. Random walk and diffusion limited aggregation (DLA) are examples of random physical fractals.

Hulls For a self-intersecting SLE trace the trace and enclosed regions form a so-called hull. Points in the hull cannot be connected to infinity without crossing the trace. The growing hull of a self-intersecting SLE trace eventually exhausts the half plane.

Ising model The Ising model originates from the theory of magnetism and plays an important role in the theory of phase transitions and in general in statistical mechanics. The model is defined on a lattice where each lattice sites is endowed with a local spin variable assuming two distinct values. The spins interact by a short range exchange interaction. Above 1D the Ising model has a second order phase transition.

Locality In the percolation case the domain wall is constructed by an exploration process and a local rule for assigning the next step. This locality property is specific to percolation which has a geometric phase transition. In the SLE context the locality property implies $\kappa = 6$, i. e., the percolation case.

Loewner equation The Loewner equation is the first order nonlinear differential equation for the uniformizing map $g_t(z)$ which maps the complement of a growing hull or curve in the upper half plane back to the upper half plane. The geometrical properties of the hull is encoded in the real function a_t. The Loewner equation has the form $\mathrm{d}g_t/\mathrm{d}t = 2/(g_t - a_t)$.

Loewner evolution Loewner evolution refers to the parametrized map $g_t(z)$ which uniformizes the complement of a curve or hull, say in the upper half plane.

Loop erased random walk (LERW) Loop erased random walk is a variant of random walk where loops formed are removed as the walk progresses. LERW has a conformally invariant scaling limit and can be generated by SLE for $\kappa = 2$.

Markov property The Markov property refers to a stochastic process without memory. A typical ex-

ample is random walk or in the scaling limit Brownian motion. The Markov property is essential for the application of SLE to domain walls in the scaling limit.

Measures For lattice systems one can define probability distributions according to the rules of statistical mechanics. In the scaling limit the distributions typically diverge and have to be replaced by the more abstract mathematical concept of probability measures.

$O(n)$ models The $O(n)$ model is a generalization of the Ising model. At each lattice site is associated an n-component spin variable. For $n = 1$ we recover the Ising model, $n = 2$ corresponds to the XY model, and $n = 3$ to the Heisenberg model.

Peano curve A Peano curve is a non-crossing curve which is dense in the plane, i. e., it gets arbitrarily close to every point. The Peano curve has the fractal dimension $D = 2$. In a SLE context a random Peano curve winds around a UST and corresponds to $\kappa = 8$.

Percolation Percolation on a lattice is constructed by filling lattice sites at random with a common probability p. At a critical concentration a spanning cluster extends across the system.

Potts model The Potts model is a generalization of the Ising model where the local lattice variable can assume q different values and where sites only interact when they are in the same Potts state. The Potts model has a FK cluster representation. The Ising model corresponds to $q = 2$; $q = 1$ corresponds to percolation and $q = 0$ to the UST.

Random walk Random walk is an ubiquitous phenomenon in nature. In random walk on a square lattice a particle jump from site to site with a given fixed probability. Each step is independent of the past history, there is no memory, this is the so-called Markov property. The random walk path or history is plane-filling and has the fractal dimension $D = 2$. The mean square displacement of random walk scales with the number of steps.

Restriction The restriction property in a SLE context implies that the measure on a curve in a domain D conditioned not to hit a bulge L is the same as the measure in the domain $D \setminus L$. The restriction property holds in the case of a uniform measure and applies to self-avoiding random walk.

Riemann's mapping theorem Riemann's mapping theorem states that an arbitrary simply connected domain D, i. e., without holes, can be mapped to another simply connected domain D' by means of a suitable complex function $g(z)$, i. e., $g(D) = D'$. Often the disk or half plane are used as reference domains. The mapping theorem does not make assumptions about the domain boundaries which can be fractal.

Scale invariance The scaling property refers to the case where a phenomenon is devoid of a characteristic scale or unit. Scale invariance is typically characterized by a power law behavior and critical exponents.

Scaling limit The limit where the lattice parameter in a lattice model approaches zero. The scaling limit is equivalent to the continuum limit.

Self-avoiding random walk (SAW) A self-avoiding random walk is a walk conditioned not to cross itself. SAW has been used to model polymers. SAW has a uniform probability measure and conforms in a SLE context to the restriction condition.

Stochastic Loewner evolution (SLE) Stochastic Loewner evolution is Loewner evolution driven by a real stochastic function a_t with a distribution given by 1D Brownian motion, i. e., $a_t = \sqrt{\kappa} B_t$. SLE is governed by the stochastic equation of motion $\mathrm{d}g_t/\mathrm{d}t = 2/(g_t - a_t)$.

Schramm's theorem Schramm's theorem refers to Schramm's derivation of SLE for LERW. Schramm showed that the scaling limit of LERW is described by SLE for $\kappa = 2$. Schramm also conjectured that the random Peano curve winding around an UST is described by SLE for $\kappa = 8$ and that percolation is described by SLE for $\kappa = 6$.

Uniformizing maps Conformal transformations which map a domain D to a standard reference domain, e. g., the half plane or the disk, are called uniformizing maps.

Uniform spanning tree (UST) A spanning tree is a collection of vertices and links forming a tree (no loops or cycles). A USF is a random tree picked among all spanning trees with equal probability.

Definition of the Subject

Stochastic Loewner evolution also called Schramm Loewner evolution (abbreviated, SLE) is a rigorous tool in mathematics and statistical physics for generating and studying scale invariant or fractal random curves in two dimensions (2D). The method is based on the older deterministic Loewner evolution introduced by Karl Löwner [64], who demonstrated that an arbitrary curve not crossing itself can be generated by a real function by means of a conformal transformation. A real function defined in one spatial dimension (1D) thus encodes a curve in 2D, in itself an intriguing result. In 2000 Oded Schramm [77] extended this method and demonstrated that driving the Loewner evolution by a 1D Brownian mo-

tion, the curves in the complex plane become scale invariant; the fractal dimension turns out to be determined by the strength of the Brownian motion.

The one parameter family of scale invariant curves generated by SLE is conjectured and has in some cases been proven to represent the continuum or scaling limit of a variety of interfaces and cluster boundaries in lattice models in statistical physics, ranging from self-avoiding random walks to percolation cluster boundaries, and Ising domain walls.

SLE operates in the 2D continuum where it generates extended scale invariant objects. SLE delimits scaling universality classes by a single parameter κ, the strength of the 1D Brownian drive, yielding the fractal dimension D of the scale invariant shapes according to the relation $D = 1 + \kappa/8$. Moreover, SLE provides the geometrical aspects of conformal field theory (CFT). The central charge c, delimiting scaling universality classes in CFT, is thus related to κ by means of the expression $c = (3\kappa - 8)(6 - \kappa)/2\kappa$.

Stochastic Loewner evolution derives its importance from the fact that it addresses the issue of extended random fractal shapes in 2D by direct analysis in the continuum. It thus supplements and extends earlier lattice results and also allows for the determination of new scaling exponents. From the point of view of conformal field theory based on the concept of a local field, operator expansions, and correlations, the geometrical approach afforded by SLE, directly addressing conformally invariant random shapes in the continuum, represents a novel point of view of maybe far reaching consequences; so far only explored in two dimensions.

Introduction

The field of SLE has mainly been driven by mathematicians presenting their results in long and difficult papers. There are, however, presently several excellent reviews of SLE both addressing the theoretical physics community [11,13,25,27,39,43] and the mathematical community [56,85], see also a complete biography up to 2003 [39]. The purpose of this article is to present a heuristic and simple discussion of some of the key aspects of SLE, for details and topics left out we refer the reader to the reviews mentioned above. However, in order to provide the necessary background and set the stage for SLE we begin with some general remarks on scaling in statistical physics.

General Remarks

In statistical physics we study macroscopic systems composed of many interacting components. In the limit of many degrees of freedom the macroscopic behavior roughly falls in two categories. In the most common case the macroscopic behavior is deterministic and governed by phenomenological theories like for example thermodynamics and hydrodynamics operating entirely on a macroscopic level. This behavior is basically a result of the law of large numbers, permitting an effective coarse graining and yielding for example a macroscopic density or velocity field [29]. In the other case, the macroscopic behavior is dominated by fluctuations and shows a random behavior [29,73]. Typical cases are random walk and equilibrium systems tuned to the critical point of a second order phase transition. Other random cases are for example self-organized critical systems purported to model earth quake dynamics, flicker noise and turbulence in fluids [5,40].

The distinction between the deterministic and random cases of macroscopic behavior is illustrated by the simple example of a biased random walk described by the Langevin equation $\mathrm{d}x(t)/\mathrm{d}t = v + \xi(t)$, $\langle \xi(t)\xi(0)\rangle \propto \delta(t)$. Here $x(t)$ is a macroscopic variable sampling the statistically independent microscopic steps $\xi(t)$; the velocity v is an imposed drift or bias. Averaging over the steps we obtain for the deviation of x, $R = [\langle x^2\rangle]^{1/2} = [(vt)^2 + t]^{1/2}$. For large t the deviation $R \sim \langle x\rangle = vt$ and the mean value or deterministic part dominates the behavior, the fluctuational or random part $R \sim t^{1/2}$ being subdominant. Fine tuning the random walk to vanishing bias $v = 0$ we have $\langle x\rangle = 0$ and $R \sim t^{1/2}$, i. e., the random fluctuations control the phenomenon.

The study of complexity encompasses a broader field than statistical physics and is concerned with the emergence of universal properties on a mesoscopic or macroscopic scale in large interacting systems. For example particle systems, networks in biology and sociology, cellular automata, etc. The class of complex systems generally falls in the category of random systems. The emergent properties are a result of many interacting agents or degrees of freedom and can in general not be directly inferred from the microscopic details. A major issue is thus the understanding of generic emerging properties of complex systems [45,70,83]. Here, however, the methods of statistical physics is an indispensable tool in the study of complexity.

The evolution of statistical physics, a branch of theoretical physics, has occurred in steps and is driven both by the introduction of new concepts and the concurrent development of new mathematical methods and techniques. A well-known case is the long standing problem of second order phase transitions or critical phenomena which gave way to a deeper understanding in the sixties and seventies and spun off the renormalization group techniques

for the determination of critical exponents and universality classes [20,24,65,81,86].

Scaling Ideas

This brings us to the fundamental scaling ideas and techniques developed particularly in the context of critical phenomena in equilibrium systems and which now pervade a good part of theoretical physics and, moreover, play an important role in the analysis of complex systems in physical sciences [20,24,29]. Scaling is synonymous with no scale in the sense that a system exhibiting scale invariance is characterized by the absence of any particular scale or unit. Scaling occurs both in the space and/or time behavior and is typically characterized by power law dependencies controlled by scaling exponents.

A classical case is random walk discussed above, characterized by the Langevin equation $\mathrm{d}x/\mathrm{d}t = \xi(t)$, $\langle\xi\xi\rangle \sim \delta(t)$ [73]. Here the mean square displacement scales like $\langle x^2\rangle(t) \sim t^{2H}$, where H is the Hurst scaling exponent; for random walk $H = 1/2$ [32,68]. Correspondingly, the power spectrum $P(\omega) = |x_\omega|^2$, $x_\omega = \int \mathrm{d}t x(t)\exp(i\omega t)$, scales like $P(\omega) \sim \omega^{-1-2H}$, i. e., $P(\omega) \sim \omega^{-2}$ for random walk; we note that the underlying reason for the universal scaling behavior of random walk is the central limit theorem [37,73].

Scaling in Equilibrium

Scaling ideas and associated techniques first came to the forefront in statistical physics in the context of second order phase transitions or critical phenomena [20,24,29,65]. More specifically, consider the usual Ising model defined on a lattice with a local spin degrees of freedom, $\sigma_i = \pm 1$ at site i, subject to a short range interaction J favoring spin alignment. The Ising Hamiltonian has the form $H = -J\sum_{\langle ij\rangle}\sigma_i\sigma_j$, where $\langle ij\rangle$ indicates nearest neighbor sites. The thermodynamic phases are characterized by the order parameter $m = \langle\sigma_i\rangle$. Above one dimension the Ising model exhibits a second order phase transition at a finite critical temperature T_c. Above T_c the model is in a disordered paramagnetic state with $m = 0$ and only microscopic domain of ordered spins. Below T_c the model favors a ferromagnetic state with long range order and macroscopic domains of ordered spins, corresponding to $m \neq 0$; at $T = 0$ the model assumes the ferromagnetic ground state configuration with totally aligned spins. Regarding the spatial organization, the size of the domains of ordered spins is characterized by the correlation length $\xi(T)$. As we approach the critical point at T_c the order parameter vanishes $m \propto |T - T_c|^\beta$ with critical exponent β, but more significantly, the correlation length $\xi(T)$ diverges like $\xi(T) \sim |T - T_c|^{-\nu}$ with scaling exponent ν. This indicates that the system is scale invariant at T_c. Regarding the domains of ordered spins, the system is spatially self-similar on scales from the microscopic lattice distance to the system size; the system size diverging in the thermodynamic limit. The scaling behavior at the critical point T_c is an emergent property in the sense that the scaling exponents β and ν do not depend on microscopic details like the type of lattice, strength of interaction, etc., but only on the dimension of the system and the symmetry of the order parameter [72].

The diverging correlation length at the critical point is the central observation which in the 60-ties and 70-ties gave rise to a detailed understanding of critical phenomena, beginning with the coarse graining block scheme proposed by Kadanoff [41] and culminating with the development and application of field theoretical renormalization group techniques by Wilson and others [20,24,29,65,72,86].

For a diverging correlation length much larger than the lattice distance the local spin σ_i can be replaced by a coarse-grained local field $\phi(r)$ and the Ising Hamiltonian H by the Ginzburg–Landau functional $F = \int \mathrm{d}^d r[(\nabla\phi)^2 + R\phi^2 + U\phi^4]$, where the 'mass' term $R \sim |T - T_c|$. Consequently, the universality class of Ising-type models is described by a scalar field theory. The renormalization group techniques basically quantify the Kadanoff block construction in momentum space and extract scaling properties in an expansion about the upper critical dimension $d = 4$. To leading order in $4 - d$ one obtains $\beta = 1/2 - (4-d)/6$ and $\nu = 1/2 + (4-d)/12$. Alternatively, keeping the correlation length ξ fixed and letting the lattice distance approach zero, we obtain at T_c the so-called scaling limit or continuum limit of the Ising model. The Ising spin σ_i becomes a local field $\phi(r)$ and the weight of a configuration is determined by $\exp(-F)$. Note that in order to implement the scaling limit we must be at T_c. The two scenarios of a growing correlation length for fixed lattice distance implementing the Kadanoff construction and a fixed correlation length for a vanishing lattice distance yielding a continuum field theory are related by an overall scale transformation [86].

It is generally assumed that the global or nonlocal scale invariance at the critical point in the continuum limit can be extended to a local scale invariance including translation and rotation, that is an angle-preserving conformal transformation. This follows heuristically from an a local implementation of the Kadanoff coarse-graining block construction and applies to lattice models with short range interactions and discrete translational and rotational invariance [22,24]. The resulting continuum the-

ories then fall in the category of conformal field theories (CFT).

In 2D the group of conformal transformations is particularly rich since it corresponds to the class of analytical functions $w = f(z)$, mapping the complex plane z to the complex plane w. The infinite group structure imposes sufficient constraints on the structure of conformal field theories in 2D that the scaling form of correlations, e. g., $\langle\phi\phi\rangle(r)$, and in particular the critical exponents can be determined explicitly [23,24]. On finds $\beta = 1/8$ and $\nu = 1$ for the order parameter and correlation length exponents, respectively, in accordance with lattice theory results [14]. Here we also mention the Coulomb gas method for the determination of critical exponents [71].

It is a common feature of both renormalization group calculations based on an expansion about a critical dimension and conformal field theory in 2D that the local field $\phi(r)$ and its correlations are the basic building blocks and that the critical properties are encoded in their scaling form, yielding critical exponents, scaling laws, scaling functions, etc. Notwithstanding the fact that the seminal Kadanoff construction [41] was based on a geometrical picture corresponding to coarse-graining the degrees of freedom over larger and larger scales, keeping track of ordered domains on all scales, the actual geometry of critical phenomena such as the scaling properties of critical clusters was not well-understood and seemed inaccessible in the continuum limit within the context of local field theories. Whereas it is not difficult to generate critical domain walls, interfaces, and clusters for lattice models with appropriate boundary conditions by means of standard Monte Carlo simulation techniques, the continuum or scaling limit of critical shapes appeared until recently, with a few exceptions, beyond present techniques.

Stochastic Loewner Evolution

Here stochastic Loewner evolution (SLE) represents a new insight in 2D critical phenomena with respect to a deeper understanding of scale invariant curves, clusters, and shapes. Also, there appears to be deep connections between SLE and conformal field theory.

SLE is an ingenious way of generating critical curves and shapes in the 2D continuum using conformal transformations. Let us mention a characteristic example. Consider an Ising model on a lattice in a chosen domain. Imposing spin up on a continuous part of the domain boundary and spin down on the remaining part of the boundary, it follows that a specific domain wall or interface will connect the two points on the boundary where a bond is broken. At low temperature the bond energies dominate and the free energy is lowest for a straight domain wall with few kinks. As we approach the critical point entropy or fluctuations come strongly into play and the domain wall meanders balancing energy and entropy. At the critical point the system becomes scale invariant with a diverging correlation length and likewise the domain wall becomes scale invariant, i. e., it has kinks on all scales larger that the lattice distance. In the continuum limit the Ising domain wall becomes a random fractal curve with a particular fractal dimension. Here SLE provides a direct analytical method in the continuum to generate such a random curve and, moreover, provides the fractal dimension in terms of the strength of the 1D random walk driving the SLE evolution.

Outline

The outline of the present article is as follows. In Sect. "Scaling" on scaling we introduce some of the basic models and concepts necessary for a discussion of SLE: A) Random walk, B) Percolation, C) Ising model, D) Critical curves and exploration, and E) Distributions, Markov properties, and measures. In Sect. "Conformal Invariance" we turn to the essential ingredient in SLE, namely, conformal invariance: A) Conformal maps and B) Measures and conformal invariance. Section "Loewner Evolution" is devoted to deterministic or classical Loewner evolution: A) Growing stick, B) Loewner equation and C) Exact solutions. In Sect. "Stochastic Loewner Evolution", constituting the core part of this article, we discuss stochastic Loewner evolution: A) Schramm's theorem, B) SLE properties, C) Curves, hulls, and the Bessel process and D) Fractal dimension. Section "Results and Discussion" is devoted to results and discussions: A) Phase transitions, locality, restriction, and duality, B) Loop erased random walk, C) Self-avoiding random walk, D) Percolation, E) Ising model and $O(n)$ models, F) SLE and conformal field theory, G) Application to 2D turbulence, H) Application to 2D spin glass and I) Further remarks. Finally, in Sect. "Future Directions" we discuss future directions of the field. The bibliography includes books, general reviews, and more technical papers.

Scaling

Stochastic Loewner evolution, has been applied to a host of lattice models proved or conjectured to possess scaling limits. However, for the present purpose we will focus on three lattice models: Random walk, Percolation, and the Ising model.

Random Walk

Random walk is a simple and much studied random process [4,32,73]. Consider an unbiased random walk in the plane composed of N steps, where the ith step, $\vec{\eta}_i$, is random, isotropic and uncorrelated, i. e., $\langle\vec{\eta}_i\rangle = 0$ and $\langle\eta_i^\alpha \eta_j^\beta\rangle \propto \delta_{\alpha\beta}\delta_{ij}$. For the end-to-end distance we have $\vec{x} = \sum_{i=1}^N \vec{\eta}_i$ and for the size $R = [\langle\vec{x}^2\rangle]^{1/2} \sim N^{1/2}$. Assuming one step pr unit time, $N \propto t$, we obtain $R \sim t^{1/2}$, characteristic of diffusive motion. Introducing the fractal dimension D by the usual box counting procedure [32,68]

$$N(R) \sim R^D \,, \tag{1}$$

where N is the number of boxes and R the size of the object, and covering the random walk we readily infer $D = 2$, i. e., the self-crossing random walk is plane-filling modulo the lattice distance. Introducing the scaling exponent ν according to $R \sim N^\nu$ we have for random walk $\nu = 1/2$; note that $D = 1/\nu$.

The scaling limit of unbiased random walk is Brownian motion (BM) [4] and is obtained by scaling the step size η down and the number of steps N up in such a manner that the size $R \sim N^{1/2}\eta$ stays constant. The resulting BM path is a continuous non-differentiable random curve with fractal dimension $D = 2$. The BM path is plane-filling and recurrent in 2D, i. e., it returns to a given point with probability one. Focusing on one of the independent cartesian components 1D BM, B_t, is also described by the Langevin equation

$$\frac{\mathrm{d}B_t}{\mathrm{d}t} = \eta_t \,, \quad \langle\eta_t\eta_s\rangle = \delta(t-s) \,, \tag{2}$$

where η_t is uncorrelated Gaussian white noise with a flat power spectrum. Integrating Eq. (2) B_t samples the step η_t and we find, assuming $B_0 = 0$, $B_t = \int_0^t \eta_{t'}\mathrm{d}t'$ from which we directly infer the fundamental properties of BM, namely, independence and stationarity,

$$B_{T+\Delta T} - B_T \approx B_{\Delta T} \,, \qquad \text{(stationarity)} \tag{3}$$

$$B_{\Delta T}, B_{\Delta T'} \text{ indep. for } \Delta T \neq \Delta T', \quad \text{(independence)} \tag{4}$$

where $\approx$ indicates identical distributions. Moreover, the correlations are given by

$$\langle|B_t - B_s|^2\rangle = |t-s| \tag{5}$$

and B_t is distributed according to the normal (Gaussian) distribution $P(B,t) = (2\pi t)^{-1/2}\exp(-B^2/2t)$.

Whereas 1D BM drives SLE, 2D BM is not itself generated by SLE since the path is self-crossing on all scales; by construction SLE is limited to the generation of non-crossing random curves. However, variations of BM have played an important role in the development of SLE. We mention here the scaling limit of loop erased random walk (LERW) and self-avoiding random walk (SAW), to be discussed later.

Percolation

The phenomenon of percolation is relevant in the context of clustering, diffusion, fractals, phase transitions and disordered systems. As a result, percolation theory provides a theoretical and statistical background to many physical and natural sciences [82].

Percolation is the simplest lattice model exhibiting a geometrical phase transition. The site percolation model is constructed by occupying sites on a lattice with a given common probability p. Let an occupied site be denoted 'plus' and an 'empty' site denoted 'minus'. For p close to zero the sites are mainly unoccupied and the lattice is 'minus'. For p close to one the sites are predominantly occupied and the lattice is 'plus'. At a critical concentration p_c, the percolation threshold, an infinite cluster of 'plus' sites embedded in the 'minus' background extends across the lattice. In the scaling limit of vanishing lattice distance the critical cluster has a fractal boundary which can be accessed by SLE. Whereas the scaling properties of critical percolation clusters define a universality class and is independent of the lattice structure, the critical concentration p_c in general depends on the lattice. For site percolation on a triangular lattice the percolation threshold is known to be $p_c = 1/2$.

In Fig. 1 we depict a realization of site percolation on a triangular lattice in the upper half plane at the percolation threshold $p_c = 1/2$. The occupied sites are denoted 'plus', the empty sites 'minus'. By imposing appropriate boundary conditions we induce a meandering domain wall across the system from A to B. In the scaling limit the domain wall becomes a fractal non-crossing critical curve.

Ising Model

The Ising model is probably the simplest interacting many particle system in statistical physics [20,29,81]. The model has its origin in magnetism but has become of paradigmatic importance in the context of phase transitions. The model is defined on a lattice where each lattice site i is occupied by a single degree of freedom, a spin variable σ_i, assuming two values, $\sigma_i = \pm 1$, i. e., spin up or spin down.

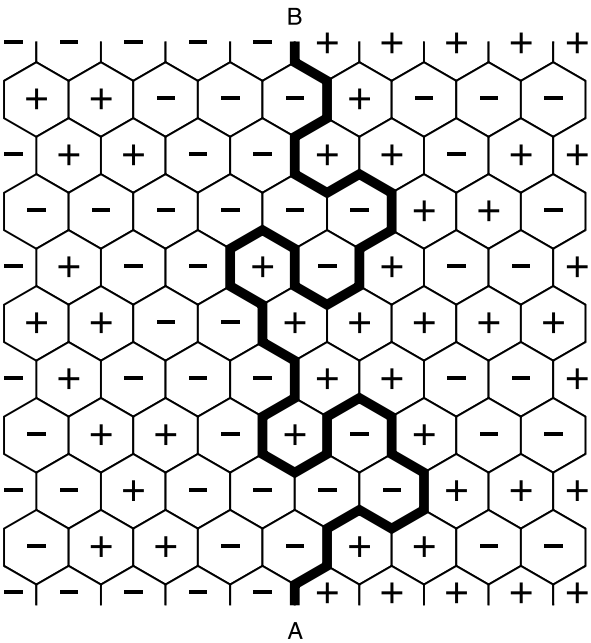

Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 1
We depict site percolation on a triangular lattice in the *upped half plane* at the percolation threshold. The critical concentration is $p_c = 1/2$. The occupied sites are denoted 'plus', the empty sites 'minus'. The boundary conditions enforce a meandering domain wall from *A* to *B*

In the ferromagnetic case considered here the spins interact via a short range exchange interaction J favoring parallel spin alignment. The model is described by the Ising Hamiltonian

$$H = -J \sum_{\langle ij \rangle} \sigma_i \sigma_j \,, \tag{6}$$

where $\langle ij \rangle$ indicates nearest neighbor spin sites i and j. The statistical weight or probability of a specific spin configuration $\{\sigma_i\}$ is determined by the Boltzmann factor

$$P(\{\sigma_i\}) = \exp[-H/kT]/Z \,, \tag{7}$$

where T is the temperature and k Boltzmann's constant. The partition function Z has the form

$$Z = \sum_{\{\sigma_i\}} \exp(-H/kT) \,, \tag{8}$$

yielding the thermodynamic free energy F according to $F = -kT \log Z$. The entropy is given by $S = -\mathrm{d}F/\mathrm{d}T$ and the energy follows from $F = E - TS$. The magnetization or order parameter and correlations are given by

$$m = \sum_{\{\sigma_i\}} \sigma_i P(\{\sigma_i\}) \quad \text{and} \quad \langle \sigma_i \sigma_j \rangle = \sum_{\{\sigma_i\}} \sigma_i \sigma_j P(\{\sigma_i\}) \,, \tag{9}$$

respectively.

The Ising model possesses a phase transition at a critical temperature T_c from a disordered paramagnetic phase above T_c with vanishing order parameter $m = 0$ to a ferromagnetic ordered phase below T_c with non-vanishing order parameter $m \neq 0$. At the critical temperature T_c the order parameter vanishes like $m \sim |T - T_c|^{\beta}$ with critical exponent β. The correlation function $\langle \sigma_i \sigma_j \rangle$ monitoring the spatial organization of ordered domains behaves like

$$\langle \sigma_i \sigma_j \rangle \sim \frac{\exp[-|i-j|/\xi]}{|i-j|^{\eta}} \,, \tag{10}$$

where $|i - j|$ denotes the distance between position i and position j. The algebraic behavior is characterized by the critical exponent η and the correlation length ξ scales like $\xi \sim |T - T_c|^{-\nu}$ with critical exponent ν. At the critical point the correlation length ξ diverges and the Ising model becomes scale invariant with an algebraically decaying correlation function $\langle \sigma_i \sigma_j \rangle \sim |i - j|^{-\eta}$. Assigning 'plus' to spin up and 'minus' to spin down, Fig. 1 also illustrates a typical configuration of the 2D Ising model on a triangular lattice at the critical temperature, including an Ising domain wall from A to B.

Critical Curves – Exploration

In the percolation case at the percolation threshold a critical curve is induced by fixing the boundary conditions. Occupying sites from A to B along the right hand side of the boundary and assigning empty sites along the left hand side of the boundary from A to B a critical curve will meander across the system from A to B. Imagining painting the two sides of the curve, one side is painted with occupied sites, the other side with empty sites. Typically the curve meanders on all scales but by construction does not cross itself. For later purposes the configuration is depicted in Fig. 1.

In order to make contact with SLE we observe that a critical interface in the percolation case also can be constructed by an exploration process. We initiate the curve at the boundary point A and toss a coin. In the case of 'head' the site or hexagon in front is chosen to be occupied and the path bends left; in the event of 'tail' the hexagon in front is left unoccupied and the path bends right. In this manner a critical meandering non-crossing curve is generated terminating eventually at B. The percolation growth

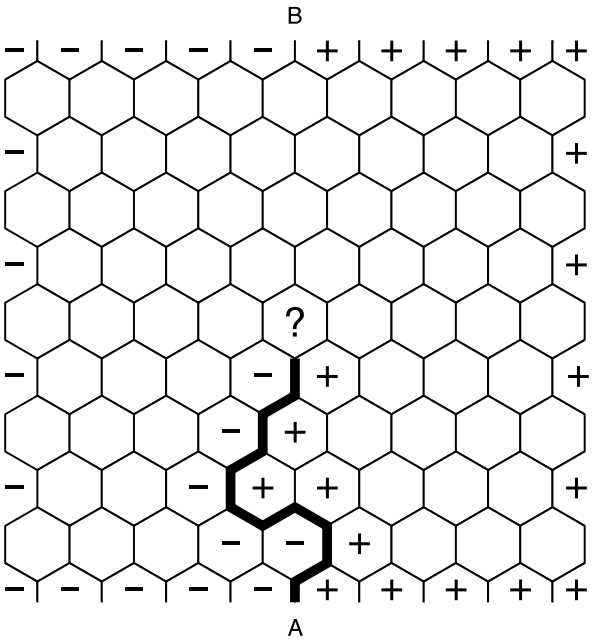

Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 2
We depict the growth process in the percolation case. The percolation threshold is at $p_c = 1/2$. The interface imposed by the boundary conditions originates at the boundary point *A* and progresses towards the boundary point *B*

process is depicted in Fig. 2, where we for clarity only have indicated the sites involved in the growth process. We note that since there is no interaction between the sites the path depends entirely on the local properties.

In the case of the Ising model a critical interface or domain walls at the critical point is again fixed by assigning appropriate boundary conditions with spin up along the boundary from A to B and spin down from B to A. The critical curve can be constructed in two ways: Globally or by an exploration process. In the global case we generate a spin configuration by means of a Monte Carlo simulation, i. e., perform a biased importance sampling implementing the probability distribution in Eq. (7), and identify a critical interface. Alternatively, we can generate an interface by an exploration process like in the percolation case, occupying a site i according to the weight $(1/2)(1 + \langle\sigma_i\rangle)$, where $\langle\sigma_i\rangle$ is evaluated in the domain with the spins along the interface fixed, see again Fig. 2.

Distributions – Markov Properties – Measures

In the case of random walk the number of walks of length L grows like μ^L, where μ is a lattice dependent number, i. e., at each step there are μ lattice-dependent choices for choosing a direction of the next step. Consequently, the weight or probability of a particular walk of length L is proportional to μ^{-L},

$$P(L) \sim \mu^{-L} \,, \tag{11}$$

and all walks of a given length have the same weight. We note here the important Markov property, characteristic of random walk, which can be formulated in the following manner. Assume that the first part γ' of the walk has taken place and thus conditions the remaining part γ of the walk. In a suggestive notation the conditional distribution is given by $P(\gamma|\gamma') = P(\gamma\gamma')/P(\gamma')$. The Markov property implies that the conditional probability is equal to the probability of the walk γ in a new domain where the first part of the walk γ' has been removed, i. e., the identity

$$P(\gamma|\gamma') = P'(\gamma) \,, \quad \text{(Markov property)} \tag{12}$$

where the prime refers to the new domain. The Markov property is self-evident for random walk and follows directly from Eq. (11), i. e., $P(\gamma|\gamma') = P(\gamma\gamma')/P(\gamma') = \mu^{-(L+L')}/\mu^{-L'} = \mu^{-L} = P'(\gamma)$, where L and L' are the lengths of segments γ' and γ, respectively.

In the scaling limit the lattice distance goes to zero whereas the length of the walk diverges. Consequently, the distribution diverges and must be replaced by an appropriate probability measure [4,13,43]. However, in order to define the probability distribution or measure in the scaling limit we shall assume that the Markov property continues to hold and interpret the P in Eq. (12) as probability measures. The Markov property is essential in carrying over the lattice probability distributions in the scaling limit.

For the critical interfaces defined by an exploration processes in the case of site percolation, the Markov property follows by inspection since the propagation of the interface is entirely determined by the local process of occupying the next site with probability 1/2. This locality property is specific for percolation which has a geometrical phase transition and does in the SLE context, to be discussed later, determine the scaling universality class.

In the case of an interface in the Ising model the Markov property also holds, but since the spins interact a little calculation is required [13]. Consider an interface γ defined by an exploration process. According to the rules of statistical mechanics the probability distribution for the interface γ is given by

$$P(\gamma) = \frac{Z(\gamma)}{Z} \,. \tag{13}$$

Here Z is the full partition function defined in Eq. (8) with appropriate boundary conditions imposed. $Z(\gamma)$ is the partial partition function with the spins associated with the interface γ fixed,

$$Z(\gamma) = \sum_{\{\sigma_i\},\gamma} \exp[-H(\gamma)/kT] \,. \tag{14}$$

The Hamiltonian $H(\gamma)$ inferred from Eq. (6) is the energy of a spin configuration with the spins determining γ fixed; $\{\sigma_i\}, \gamma$ indicate the configurations to be summed over. Evidently, we have the identity $Z = \sum_\gamma Z(\gamma)$, i. e., $\sum_\gamma P(\gamma) = 1$.

Whereas in the random walk case we only considered an individual path and in the percolation case the interface only feels the nearby sites, an Ising interface is imbedded in the interacting spin systems and we have to define the Markov property more precisely with respect to a domain D. Consider an interface across the domain D from A to B and assume that the last part γ is conditioned on the determination of the first part γ', i. e., given by the distribution $P_D(\gamma|\gamma')$. Next imagine that we cut the domain D along the interface γ', i. e., break the interaction bonds between the spins determining γ'. The right and left face of γ' can then be considered part of the domain boundary and the Markov property states that the distribution of γ in the cut domain $D \setminus \gamma'$ (i. e., D minus γ') equals $P_D(\gamma|\gamma')$,

$$P_D(\gamma|\gamma') = P_{D\setminus\gamma'}(\gamma) \,. \quad \text{(Markov property)} \tag{15}$$

In order to demonstrate Eq. (15) we use the definition in Eq. (13). The conditional probability $P_D(\gamma|\gamma') = P_D(\gamma\gamma')/P_D(\gamma') = (Z_D(\gamma\gamma')/Z_D)/(Z_D(\gamma')/Z_D) = Z_D(\gamma\gamma')/Z_D(\gamma')$. Correspondingly, the distribution in the cut domain $D \setminus \gamma'$ is $P_{D\setminus\gamma'}(\gamma) = Z_{D\setminus\gamma'}(\gamma)/Z_{D\setminus\gamma'}$. However, it follows from the structure of the partition function in Eqs. (6–8) that $Z_{D\setminus\gamma'} = \exp[E(\gamma')/kT]Z_D(\gamma')$ and $Z_{D\setminus\gamma'}(\gamma) = \exp[E(\gamma')/kT]Z_D(\gamma\gamma')$, where $E(\gamma')$ is the energy of the broken bonds. By insertion the interface Boltzmann factors cancel out and we obtain Eq. (15) expressing the Markov property.

Conformal Invariance

In the complex plane analysis implies geometry. The representation of a complex number $z = x + iy$ directly associates complex function theory with 2D geometrical shapes. This connection is of importance in mathematical physics in for example 2D electrostatics and 2D hydrodynamics. In the context of SLE Riemann's mapping theorem plays an essential role [1].

Conformal Maps

Briefly, Riemann's mapping theorem [2] states that any simply connected domain, i. e., topologically deformable to a disk, in the complex plane z can be uniquely mapped to a unit disk $|w| < 1$ in the complex w plane by mean of a complex function $w = g(z)$. By combining complex functions we can map any simply connected domain to any other simply connected domain. For example, if $g_1(z)$ and $g_2(z)$ map domains D_1 and D_2 to the unit disk, respectively, then $g_2^{-1}(g_1(z))$ maps D_1 to D_2; here g_2^{-1} is the inverse function of g_2, i. e., $g_2^{-1}(g_2(z)) = z$. As an example, the transformation $g(z) = i(1+z)/(1-z)$ maps the the unit disk centered at the origin to the upper half plane. Likewise, the Möbius transformation $g(z) = (az+b)/(cz+d)$ determined by four real parameters, $ad - bc > 0$, maps the upper half plane onto itself. Conformal transformations are angle-preserving and basically correspond to a combination of a local rotation, local translation, and local dilatation. In terms of an elastic medium picture conformal transformations are equivalent to deformations without shear [54]. Expressing $g(z)$ in terms of its real and imaginary parts, $g(z) = u(x,y) + iv(x,y)$, the Cauchy–Riemann equations $\partial u/\partial x = \partial v/\partial y$ and $\partial u/\partial y = -\partial v/\partial x$ hold implying that u and v are harmonic functions satisfying Laplace's equations $\nabla^2 u = 0$ and $\nabla^2 v = 0$, $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$. In Fig. 3 we have depicted a conformal angle-preserving transformation effectuated by the complex function $g(z)$ from the complex z plane to the complex w plane.

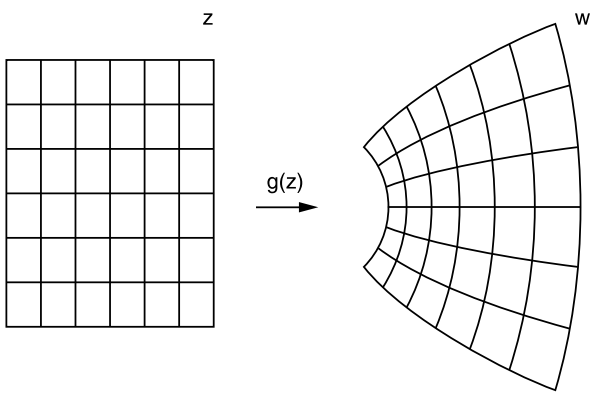

Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 3
We depict a conformal transformation from the complex z plane to the complex w plane. We note the angle-preserving property, i. e., a shear-free transformation. The map in the figure is given by $w = z^2$

Measures – Conformal Invariance

Whereas the Markov property discussed above holds for lattice curves even away from criticality, we here want to assume another property which only holds in the scaling limit at the critical point, namely conformal invariance. In the scaling limit we anticipate that the probability measure $P(\gamma)$ for an interface γ is invariant under a conformal transformation. More precisely, consider a lattice model, say the Ising or percolation model, and specify two domains D and D' on the lattice. Next, consider an interface, cluster boundary or domain wall γ from the boundary points A and B across the domain D. In terms of the partition functions the probability distribution for γ is given by Eq. (13). We now perform the scaling or continuum limit of the lattice model keeping the domains D and D' fixed. The continuous random interface approaches its scaling form and is characterized by the measure $P_D(\gamma)$. At the critical point we assume that the interface is scale invariant under the larger symmetry of conformal transformations. The next step is to consider a specific conformal transformation $g(z)$ which according to Riemann's theorem precisely maps domain D to domain D', i. e., $D' = g(D)$ and the interface to $g(\gamma)$. The assumption of conformal invariance then states that the probability measure P is invariant under this transformation expressing the scale invariance, i. e.,

$$P_D(\gamma) = P_{g(D)}(g(\gamma))\,. \quad \text{(Conformal property)} \tag{16}$$

Both the Markov property and the conformal property are sufficient in combination with Loewner evolution to determine the measures in the scaling limit.

Loewner Evolution

The original motivation of Loewner's work was to examine the so called Bieberbach conjecture which states that $|a_n| \le n$ for the coefficients in the Taylor expansion $f(z) = \sum_{n=0} a_n z^z$, where $f(z)$ maps the unit disk to the complex plane. The conjecture was proposed in 1916 and finally proven in 1984 by de Branges [2,38,39]. For that purpose Loewner [64] considered growing parametrized conformal maps to a standard domain. In the present context Loewner's method allow us to access growing shapes in 2D in an indirect manner by means of a 'time dependent' conformal transformation $g_t(z)$.

Growing Stick

Before we address the derivation of the Loewner equation let us consider the specific conformal transformation

$$g_t(z) = \sqrt{z^2 + 4t}\,. \tag{17}$$

For $t = 0$ we have $g_0(z) = z$, i. e., the identity map. Likewise, for $z \to \infty$ we obtain

$$g_t(z) \sim z + \frac{2t}{z}\,, \tag{18}$$

showing that far away in the complex plane we again have the identity map. The coefficient in the next leading term, C_t/z, is called the capacity C_t; here parametrized by the 'time variable', $t = C_t/2$. The map (17) has a branch point at $z = 2it^{1/2}$ and it follows by inspection that g_t maps the upper half plane minus a stick from the origin $\mathbf{0}$ to $2it^{1/2}$ back to the upper half plane. From the inverse map $f_t(w) = g_t^{-1}(w)$,

$$f_t(w) = \sqrt{w^2 - 4t}\,, \tag{19}$$

we infer that the right face of the stick is mapped to the real axis from 0 to $2t^{1/2}$, the tip to the origin, and the left face to the interval $-2t^{1/2}$ to 0. Under the map the growing stick thus becomes part of the boundary in the w plane. The growing stick is depicted in Fig. 4.

The stick shows up as an imaginary contribution along the interval $-2t^{1/2}$ to $2t^{-1/2}$. More precisely, since $f_t(w)$ is analytic in the upper half plane, implementing the asymptotic behavior $f_t(w) \sim w$ for $w \to \infty$, and using Cauchy's

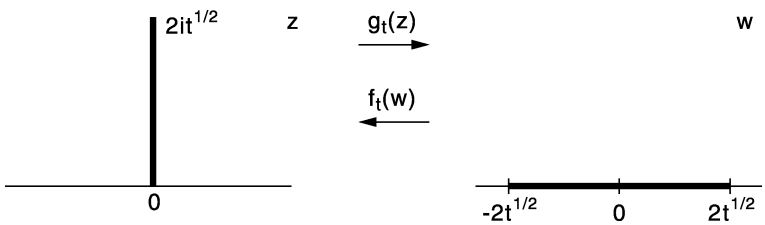

Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 4
We depict the growing stick corresponding to the conformal transformation $w = \sqrt{z^2 + 4t}$. The vertical cut in the z plane extends from the origin to the point $(0, 2it^{1/2})$. The right and left faces of the cut are mapped to the real axis from $-2t^{1/2}$ to $+2t^{1/2}$, the endpoint to the origin, in the complex w plane

theorem, we obtain the dispersion relation or spectral representation

$$f_t(w) = w - \int \frac{\mathrm{d}\omega}{\pi} \frac{A_t(\omega)}{w-\omega}\,, \tag{20}$$

with spectral weight $A_t(\omega)$. In the case of the growing stick we find $A_t(\omega) = (4t-\omega^2)^{1/2}$ for $\omega^2 < 4t$ and otherwise $A_t(\omega) = 0$. Using $1/(\omega + i\epsilon) = \mathrm{P}\ 1/\omega - i\pi\delta(\omega)$ (P denotes principal value) we also have $\mathrm{Im} f_t(\omega) = A_t(\omega)$ and $\mathrm{Re} f_t(\omega) = \omega - \mathrm{P}\int(\mathrm{d}\omega'/\pi)A_t(\omega')/(\omega-\omega')$. The time dependent spectral weight $A_t(\omega)$ thus characterizes the growing stick. With the chosen parametrization we also have the sum rule

$$\int \frac{\mathrm{d}\omega}{\pi} A_t(\omega) = C_t = 2t\,. \tag{21}$$

Finally, we note that the map g_t satisfies the equation of motion

$$\frac{\mathrm{d}g_t(z)}{\mathrm{d}t} = \frac{2}{g_t(z)}\,, \tag{22}$$

i. e., solving Eq. (22) with the initial condition $g_0(z) = z$ and the boundary condition $g_t(z) \sim z$ for $z \to \infty$ we arrive at the map in Eq. (17).

Loewner Equation

The growing stick nicely illustrates the idea of accessing a growing shape indirectly by the application of Riemann's theorem mapping the domain adjacent to the shape to a standard reference domain, here the upper half plane. This so-called uniformizing map effectively absorbs the shape and encodes the information about the shape into the spectral weight $A_t(\omega)$ along the real axis.

Let us consider a general shape or hull $\mathbf{K_t}$ in the upper half plane $\mathbf{H}$. Together with the real axis the shape form part of the boundary of the domain $\mathbf{D}$. In other words, the domain in question is the upper half plane $\mathbf{H}$ with the shape $\mathbf{K_t}$ subtracted, $\mathbf{D} = \mathbf{H} \setminus \mathbf{K_t}$. Applying Riemann's theorem we map the simply connected domain $\mathbf{D}$ back to the upper half plane $\mathbf{H}$ by means of the conformal transformation $g_t(z)$, i. e., g_t absorbs the shape $\mathbf{K_t}$. Imagine that the shape grows a little bit further from $\mathbf{K_t}$ to $\mathbf{K}_{t+\delta t} = \mathbf{K_t} + \delta\mathbf{K_t}$, where $\mathbf{K_t}$ is now part of $\mathbf{K}_{t+\delta t}$; $\delta\mathbf{K_t}$ is the shape increment. Correspondingly, the map $g_{t+\delta t}$ is designed to absorb $\mathbf{K}_{t+\delta t}$, i. e., $\mathbf{H} \setminus \mathbf{K}_{t+\delta t} \to \mathbf{H}$ by means of the map $g_{t+\delta t}$. We now carry out the elimination in two ways. Either we absorb $\mathbf{K_t}$ by means of the map g_t and subsequently $\delta\mathbf{K_t}$ by means of the map δg_t or we absorb $\mathbf{K}_{t+\delta t}$ directly in one step by means of the map $g_{t+\delta t}$. Consequently, combining maps we have $g_{t+\delta t}(z) = \delta g_t(g_t(z))$ or $g_t(z) = \delta g_t^{-1}(g_{t+\delta t}(z))$. Since $\delta g_t^{-1}(w)$ is analytic in $\mathbf{H}$ we obtain the spectral representation

$$\delta g_t^{-1}(w) = w - \int \frac{\mathrm{d}\omega}{\pi} \frac{\delta A_t(\omega)}{w-\omega}\,, \tag{23}$$

with infinitesimal spectral weight δA_t, or inserting $w = g_{t+\delta}(z)$

$$g_t(z) = g_{t+\delta t}(z) - \int \frac{\mathrm{d}\omega}{\pi} \frac{\delta A_t(\omega)}{g_{t+\delta t} - \omega}\,. \tag{24}$$

The last step is to set $\delta A_t(\omega) = \rho_t(\omega)\delta t$, yielding a differential equation for the evolution of the map g_t eliminating the shape $\mathbf{K_t}$,

$$\frac{\mathrm{d}g_t(z)}{\mathrm{d}t} = \int \frac{\mathrm{d}\omega}{\pi} \frac{\rho_t(\omega)}{g_t(z) - \omega}\,. \tag{25}$$

Specifying the weight or measure $\rho_t(\omega)$ along the real ω-axis this equation determines, through the uniformizing map g_t, how the shape $\mathbf{K_t}$ grows. The spectral weight encodes the 2D shape into the real function $\rho_t(\omega)$. Note that since $\rho_t(\omega)$ is not specified and can depend nonlinearly on the map, Eq. (25) still represent a highly nonlinear problem. Invoking the asymptotic condition $g_t(z) \sim z + C_t/z$, where C_t is the time dependent capacity we infer $\mathrm{d}C_t/\mathrm{d}t = \int(\mathrm{d}\omega/\pi)\rho_t(\omega)$ or since $\rho_t(\omega) = \mathrm{d}A_t(\omega)/\mathrm{d}t$ the sum rule in Eq. (21). The conformal mapping procedure is depicted in Fig. 5.

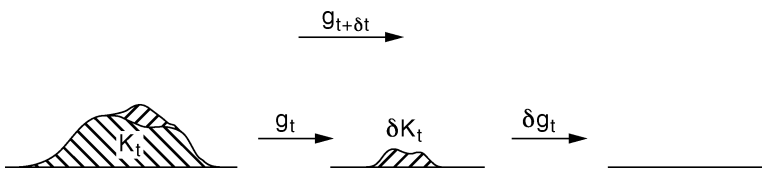

Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 5
The combination of maps involved in the derivation of the Loewner equation. First the map g_t eliminates the hull K_t. Subject to the growth in the time interval δt the incremental hull δK_t is subsequently absorbed by the infinitesimal map δg_t. Correspondingly, the hull $K_{t+\delta t}$ is absorbed by the map $g_{t+\delta t}$ in one step

In the special case where the growth takes place at a point the equation simplifies considerably. Assuming that the spectral weight is concentrated at the point $\omega = a_t$, where a_t is a real function of t and setting $\rho_t(\omega) = 2\pi\delta(\omega - a_t)$ we arrive at the Loewner equation

$$\frac{\mathrm{d}g_t(z)}{\mathrm{d}t} = \frac{2}{g_t - a_t} \,. \tag{26}$$

The Loewner equation describes the growth of a curve or trace γ_t with endpoint z_t, $0 < t < \infty$, in the upper half complex z-plane. The time-dependent conformal transformation g_t maps the simply connected domain $\mathbf{H} \setminus \gamma_t$, i. e., the half plane excluding the curve γ_t back to the w half plane. At a given time instant t the tip of the curve z_t is determined by $g_t(z_t) = a_t$, i. e., the point where Eq. (26) develops a singularity. The topological properties and shape of the curve are encoded in the real function a_t which lives on the real axis in the w-plane. As a_t develops in time the tip of the curve z_t determined by $z_t = g_t^{-1}(a_t)$ traces out a curve. Since the domain $\mathbf{H} \setminus \gamma_\mathbf{t}$ must be simply connected for the Riemann theorem to apply the curve or trace cannot cross itself or cross the real axis. Whenever the curve touches or intersect itself or the real axis the enclosed part will be excluded from the domain. In other, words, during the time progression the curve effectively absorbs part of the upper half plane. It is a deep property of Loewner evolution that the topological properties of a 2D non-crossing curve are entirely encoded by the real function a_t. The encoding works both ways: A given 2D non-crossing curve γ_t corresponds to a specific real function a_t, a given real function a_t yields a specific 2D non-crossing curve γ_t. A continuous a_t will yield a continuous curve γ_t. A discontinuous a_t in general gives rise to branching. Whether or not the curve intersects or touches itself is determined by the singularity structure of the drive a_t. In the case where the Hölder condition $\lim_{\tau\to 0} |(a_{t+\tau} - a_\tau)/\tau^{1/2}|$ is greater that 4 we have self-intersection. Note again that since the curve is defined indirectly by the singularity structure in Eq. (26) we cannot easily identify a curve parametrization and for example determine a tangent vector, etc. The mechanism underlying the Loewner equation is shown in Fig. 6.

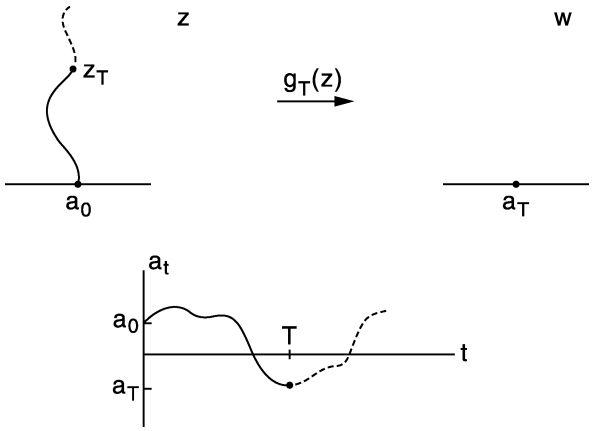

Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 6
The mechanism in the Loewner equation. The curve in the upper half complex z plane generated by the Loewner equation is mapped onto a finite but growing segment of the real axis of the complex w plane. The endpoint z_T is mapped to the real number a_T. As a_t develops in time and makes excursions along the real axis the endpoint z_t of the curve grows into the upper half plane

Exact Solutions

In a series of simple cases one can solve the Loewner equation analytically [42,44]. For vanishing drive $a_t = 0$ we obtain the growing vertical stick discussed above. Correspondingly, a constant drive $a_t = a$ yields a vertical stick growing up from the point a on the real axis.

In the case of a linear drive $a_t = t$ the tip of the curve z_t is given by $z_t = 2 - 2\phi_t \cos\phi_t + 2i\phi_t$, where the phase ϕ_t is determined from the equations: $2\ln r_t - r_t \cos\phi_t = 2\ln 2 + t - 2$ and $r_t = 2\phi_t/\sin\phi_t$. By inspection $\phi_0 = 0$ and $\phi_\infty = \pi$. The curve thus approaches the asymptote $2\pi i$ for $t \to \infty$. For small t analysis yields $z_t \sim (2/3)t + 2i\sqrt{t}$, i. e., the trace approaches the origin with infinite slope. The square root drives $a_t = 2\sqrt{\kappa t}$ and $a_t = 2\sqrt{\kappa(1-t)}$, $0 < t < 1$ with a finite-time singularity can also be treated. In the first case, $a_t = 2\sqrt{\kappa t}$, the trace is a straight line $z_t = B\exp(i\phi)\sqrt{t}$ forming the angle ϕ with respect to the real axis. The amplitude B and phase ϕ depend on the parameter κ. The angle $\phi = (\pi/2)(1 - \kappa^{1/2}/(\kappa+4)^{1/2})$. For $\kappa = 0$, $\phi = \pi/2$ and we recover the perpendicular stick; for $\kappa \to \infty$, $\phi \to 0$ and the angle of intersection decreases to zero. In the second case, $a_t = 2\sqrt{\kappa(1-t)}$, the behavior of the trace is more complex. For $0 < \kappa < 4$ the trace forms a finite spiral in the upper half plane; for $\kappa = 4$ the trace has a glancing intersection with the real axis. For $4 < \kappa < \infty$ the trace hits the real axis in accordance with the Hölder condition discussed above.

Stochastic Loewner Evolution

After these preliminaries we are in position to address stochastic Loewner evolution (SLE). The essential observation made by Oded Schramm [77] within the context of

loop erased random walk was that the Markov and conformal properties of the measures or probability distributions for random curves generated by Loewner evolution imply that the random drive a_t must be proportional to an unbiased 1D Brownian motion.

Schramm's Theorem

The Loewner Equation (26) generates a non-crossing curve in the upper half plane $\mathbf{H}$ originating at the origin $\mathbf{O}$, given a continuous function a_t with initial value $a_0 = 0$. As a_t develops in time the tip of the curve z_t determined by the condition $g_t(z_t) = a_t$ traces out a curve or trace. In the case where a_t is a continuous random function the Loewner Equation (26) likewise becomes a stochastic equation of motion yielding a stochastic map $g_t(z)$. As a result the trace determined by $g_t(z_t) = a_t$ or $z_t = g_t^{-1}(a_t)$ is a random curve. The issue is to establish a contact between the exploration processes defining interfaces in the lattice models, the scaling limit of these curves, and the curves generated by SLE. In the scaling limit we thus invoke the two properties discussed above: i) the Markov property in Eq. (12) and ii) the conformal property in Eq. (16).

In order to demonstrate the surprising property that the Markov and conformal properties in combination imply that a_t must be a 1D Brownian motion we focus on chordal SLE which applies to a random curve or trace connecting two boundary points. Since the probability distribution or measure $P(\gamma)$ on the random curve γ using property ii) is assumed to be conformally invariant and since we by Riemann's theorem can map any simply connected domain to the upper half plane by mean of a conformal transformation, we are free to consider curves in the upper half plane from the origin $\mathbf{O}$ to ∞ parametrized with a time coordinate $0 < t < \infty$.

Imagine that we grow the curve from time $t = 0$ to time T driven by the function a_t, $0 < t < T$. The curve or trace is generated by the Loewner Equation (26) with boundary condition $a_0 = 0$ and the trace z_t by $g_t(z_t) = a_t$ or $z_t = g_t^{-1}(a_t)$. With the chosen time parametrization we have $g_t(z) \sim z + 2t/z$ for $z \to \infty$ in the upper half plane. The map g_t thus uniformizes the trace, i. e., the tip z_t is mapped to a_t on the real axis in the w-plane. In order to invoke the Markov property we let the curve grow the time increment ΔT corresponding to the curve segment $\Delta\gamma$. The Markov property then implies that the distribution on $\Delta\gamma$ conditioned on the distribution on γ is the same as the distribution on $\Delta\gamma$ in the cut domain $\mathbf{H} \setminus \gamma$, i. e., the domain with the curve γ deleted; this stage is illustrated in Fig. 7.

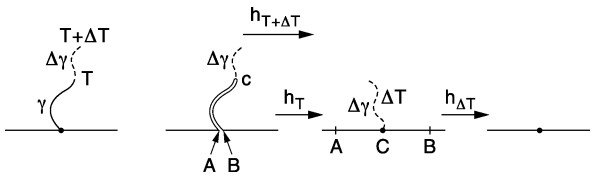

Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 7
The figure depicts the construction in the derivation of SLE. The first step implements the Markov property by turning the curve γ into a cut. Subsequently, the conformal transformation h_T maps γ back to the origin. Finally, the map $h_{\Delta T}$ maps the segment $\Delta\gamma$ to the origin. The complete process is also implemented by $h_{T+\Delta T}$. The combination of the Markov property and conformal invariance implies that a_t performs a Brownian motion

Next, in order to implement the conformal property we shift the image by a_T in such a way that the curve segment $\Delta\gamma$ again starts at the origin $\mathbf{O}$. This is achieved by using the map $h_T = g_T - a_T$ which since $g_T(z_T) = a_T$ maps the tip z_T back to the origin; this construction is also depicted in Fig. 7. Moreover, the asymptotic behavior of h_T for large z is $h_T(z) \sim z - a_T - 2T/z$. Since the measure by assumption is unchanged under the conformal transformation h_T we infer that $\Delta\gamma$ growing the time ΔT from the origin has the same distribution as the segment $\Delta\gamma$ grown from time T to time $T + \Delta T$ conditioned on the segment γ grown up to time T. Moreover, since the segment γ from $\mathbf{O}(AB)$ to C subject to the Markov property has become part of the boundary, as shown in Fig. 7, we also infer that the measure on $\Delta\gamma$ is independent of the measure on γ. Finally, applying $h_{\Delta T}$ we map the segment $\Delta\gamma$ to the origin as a common reference point as indicated in Fig. 7.

Since the random curves are determined by the random maps g_t and h_t driven by the random function a_t the issue is how to transfer the properties of the measure on the curve determined by the Markov and conformal properties to the measure on the random driving function a_t.

In order to combine the Markov properties arising from the analysis of the lattice models and the conformal invariance pertaining to critical random curves, we carry out the following steps. First we grow the curve γ from the origin $\mathbf{O}$ to the tip z_T. Implying conformal invariance the curve is then uniformized back to the origin by means of h_T. The next step is to grow the curve segment $\Delta\gamma$ in time ΔT. This segment is subsequently absorbed by means of the map $h_{\Delta T}$. According to the Markov property the distribution of $\Delta\gamma$ from $\mathbf{O}$ to $z_{\Delta T}$ is the same as the distribution of $\Delta\gamma$ grown from T to $T + \Delta T$ conditioned on γ grown from 0 to T. Since the curve γ is deter-

mined by the map h_t the distribution is reflected in h_t. In particular, the stochastic properties of the curve is transferred to the random function a_t generating the curve by the Loewner evolution. The last step is now to observe that absorbing the segment $\Delta\gamma$ from $\mathbf{O}$ to $z_{\Delta T}$ by means of $h_{\Delta T}$ is the same transformation as first applying the inverse map h_T^{-1} followed by the map $h_{T+\Delta T}$; in both cases the end result is the absorption of the initial curve $\gamma + \Delta\gamma$, see Fig. 7. As regards the measure or distribution we have the equivalence $h_{\Delta T}(z) \approx h_{T+\Delta T}(h_T^{-1}(z))$. Using the asymptotic form $h_t(z) \sim z - a_t - 2t/z$ we obtain $a_{T+\Delta T} - a_T \approx a_{\Delta T}$; note that $\approx$ indicates identical distributions or measures.

In conclusion, the Markov property in combination with conformal invariance implies that $a_{T+\Delta T} - a_T$ is distributed like $a_{\Delta T}$ (stationarity) and that $a_{\Delta T}$ and $a_{\Delta T'}$ are independently distributed for non-overlapping time intervals ΔT and $\Delta T'$ (Markov property). Referring to Sect. "Scaling", A on Brownian motion as expressed in Eqs. (3) and (4), i. e., stationarity and independence, we infer that a_t is proportional to a Brownian motion of arbitrary strength κ, i. e., $a_t = \sqrt{\kappa} B_t$. Note that the reflection symmetry $x \to -x$ holding in the present context rules out a bias or drift in the 1D Brownian motion [13,27].

This is the basic conclusion reached by Schramm in the context of loop erased random walk. Driving Loewner evolution by means of 1D Brownian motion with different diffusion coefficient or strength κ we generate a one-parameter family of conformally invariant or scale invariant non-crossing random curves in the plane.

SLE Properties

Stochastic Loewner evolution is determined by the nonlinear stochastic equation of motion

$$\frac{\mathrm{d}g_t}{\mathrm{d}t} = \frac{2}{g_t - a_t}\,, \quad a_t = \sqrt{\kappa} B_t\,. \tag{27}$$

In the course of time a_t performs a 1D Brownian motion on the real axis starting at the origin $a_0 = 0$. a_t is a random continuous function of t and distributed according to $\sqrt{\kappa} B_t$. More precisely, a_t is given by the Gaussian distribution

$$P(a,t) = (2\pi t)^{-1/2} \exp[-a^2/2\kappa t]\,, \tag{28}$$

with correlations

$$\left\langle (a_t - a_s)^2 \right\rangle = \kappa |t - s|\,. \tag{29}$$

First we notice that a constant shift of the drive a_t, $a_t \to a_t + b$, is readily absorbed by a corresponding shift of the map, $g_t \to g_t + b$. Moreover, using the scaling property of Brownian motion, $B_{\lambda^2 t} = \lambda B_t$, following from e. g. Eq. (5), we have $a_{\lambda^2 t} = \lambda a_t$ and we conclude from Eq. (27) that $g_t(z)$ has the same distribution as $(1/\lambda) g_{\lambda^2 t}(\lambda z)$, i. e., $g_{\gamma^2 t}(\gamma z) \approx \gamma g_t(z)$. Note that this dilation invariance is consistent since the origin $z = 0$ and $z = \infty$, the endpoints of curves, are preserved. Note also that the strength of the drive κ is an essential parameter which cannot be scaled away.

Curves – Hulls – Bessel Process

For vanishing drive, $\kappa = 0$, the SLE yields a non random vertical line from $z = 0$, i. e., the growing stick discussed in Subsect. "Growing Stick". As we increase κ the curve becomes random with excursions to the right and to the left in the upper half plane. Up to a critical value of κ the random curve is simple; i. e., non-touching or non self-intersecting. At a critical value of κ the Brownian drive is so strong that the curve begins to intersect itself and the real axis. These intersection take place on all scales since the curve is self-similar or scale invariant. Denoting the curve by γ_t we observe that since Riemann's theorem uniformizing $\mathbf{H} \setminus \gamma_t$ to $\mathbf{H}$ only applies to a simply connected domain, the regions enclosed by the self-intersections do not become uniformized but are effectively removed from $\mathbf{H}$. The curve γ_t together with the enclosed parts is called the hull $\mathbf{K_t}$ and the mapping theorem applies to $\mathbf{H} \setminus \mathbf{K_t}$.

In order to analyze the critical value of κ we consider the stochastic equation for $h_t(z) = g_t(z) - a_t$. From Eq. (27) it follows that

$$\frac{\mathrm{d}h_t}{\mathrm{d}t} = \frac{2}{h_t} + \xi_t\,, \tag{30}$$

where $\xi_t = -\mathrm{d}a_t/\mathrm{d}t$ is white noise with correlations $\langle \xi_t \xi_s \rangle = \kappa \delta(t - s)$.

The nonlinear complex Langevin Equation (30) maps the tip of the curve z_t back to the origin. Likewise $\mathbf{H} \setminus \gamma_t$ is mapped onto $\mathbf{H}$. A point x on the real axis is mapped to $x_t = h_t(x)$ where x_t satisfies

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = \frac{2}{x_t} + \xi_t\,. \tag{31}$$

The Langevin Equation (31) is known as the Bessel equation and governs the radial distance R from the origin of a Brownian particle in d dimensions.

Introducing $R = (\sum_{i=1}^{d} B_i^2)^{1/2}$ where B_i, $i = 1, \ldots d$, is a 1D Brownian motion with distribution $P(B,t) = (2\pi\kappa t)^{-1/2} \exp[-B^2/2\kappa t]$, we find $P(R,t) \propto (2\pi\kappa t)^{-d}$

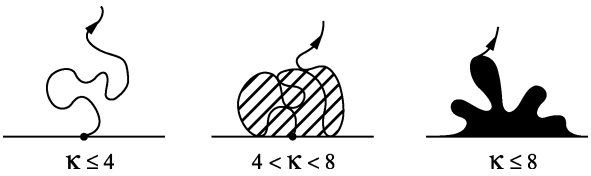

Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 8
The figures depict the phases of SLE. For $\kappa \leq 4$ the SLE trace is a simple non-intersecting scale invariant random curve from the origin to infinity with a fractal dimension between 1 and 3/2. For $4 < \kappa \leq 8$ the SLE curve is self-intersecting on all scales and also intersects the real axis on all scales. The curve together with the enclosed regions, the hull, eventually exhausts the upper half plane. The scale invariant hull has a fractal dimension ranging between 3/2 and 2. For $\kappa \geq 8$ the fractal dimension of the hull locks onto 2 and the scale invariant hull is dense and plane-filling

$R^{d-1}\exp[-R^2/2\kappa t]$ satisfying the Fokker–Planck equation $\partial P/\partial t = (\kappa/2)\partial^2 P/\partial R^2 - (\kappa(d-1)/2)\partial(P/R)/\partial R$, corresponding to the Langevin equation

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \kappa\frac{d-1}{2R} + \xi_t \,. \tag{32}$$

For $d \leq 2$ Brownian motion is recurrent, i. e., the particle returns to the origin $R = 0$, for $d > 2$ the particle goes off to infinity. Referring to Eq. (31) and setting $\kappa(d-1)/2 = 2$ we obtain $\kappa = 4/(d-1)$, i. e., $R \to \infty$ for $\kappa < 4$ and $R \to 0$ for $\kappa > 4$.

Since the tip of the curve z_t is mapped to a_t, i. e., $h(z_t) = 0$, the case $x_t \to \infty$ for $\kappa < 4$ corresponds to a curve never intersecting the real axis, i. e., the curve is simple. For $\kappa > 4$ we have $x_t \to 0$ corresponding to the case where the tip z_t intersects the real axis forming a hull. Since the curve is self-similar the intersections takes place on all scales and eventually the whole upper half plane is engulfed by the hull.

The marginal value $\kappa = 4$ can also be inferred from a simple heuristic argument [27]. For small κ we can ignore the noise in Eq. (31)and the particle is repelled according to the solution $x_t^2 \sim 4t$. For large noise we ignore the nonlinear term and the noise can drive x_t to zero; we have $x_t^2 \sim \kappa t$. The balance is obtained for $\kappa = 4$.

In conclusion, for $\kappa \leq 4$ the curve is simple, for $\kappa > 4$ the curve intersects itself and the real axis infinitely many times on all scales, eventually the hull swallows the whole plane. For large κ the trace turns out to be plane-filling. The two cases $\kappa \leq 4$ and $\kappa > 4$ are depicted in Fig. 8.

Fractal Dimension

The SLE random curves are fractal. An important issue is thus the determination of the fractal dimension D in terms of the Brownian strength or SLE parameter κ. It has been shown that [15,16,74]

$$D = 1 + \frac{\kappa}{8} \quad \text{for} \quad 0 \leq \kappa \leq 8 \,, \tag{33}$$

for $\kappa \geq 8$ the fractal dimension locks onto 2 and the SLE curve is plane-filling.

In order to illustrate a typical SLE calculation we follow Cardy [27] in a heuristic derivation of Eq. (33). In order to evaluate D and in accordance with its definition [32,68] the standard procedure is to cover the object with disks of size ε and follow how the number of disks $N(\epsilon)$ of size ε scales with ε for small ε, i. e., $N(\epsilon) \propto \epsilon^{-D}$. However, since the SLE curve is random the argument has to be rephrased. Alternatively, we consider a disc of size ε located at a fixed position z and ask for the probability $P(z,\epsilon)$ that the SLE curve crosses the disk. The number of disks covering an area A is $N_A = A/\epsilon^2$, i. e., $P \propto \epsilon^{-D}/N_A$, and we infer $P(z,\epsilon) \propto \epsilon^{2-D}$, where D is the fractal dimension. Incorporating the Markov property we subject the curve to an infinitesimal conformal transformation, $h_{\delta t} = g_{\delta t} - a_{\delta t}$, transforming the point z to $w = g_{\delta t}(z) - a_{\delta t}$; moreover, all lengths are scaled by $|h'_{\delta t}(z)|$ [1]. Setting $z = x + iy$ and $w = x' + iy'$ we obtain expanding Eq. (30) $x' + iy' = 2\delta t/(x+iy) - \sqrt{\kappa}\delta B_t$ or $x' = x + 2x\delta t/(x^2+y^2) - \sqrt{\kappa}\delta B_t$, $y' = y - 2y\delta t/(x^2+y^2)$, $\epsilon' = (1-|h_{\delta t}(z)|)\epsilon$, and $|h_{\delta t}| = 2(x^2-y^2)/(x^2+y^2)^2$. By conformal invariance the probability measure $P(x,y,\epsilon)$ is unchanged and we infer

$$P(x,y,\epsilon) = \langle P(x',y',\epsilon')\rangle_{\delta B} \,, \tag{34}$$

where we have averaged over the Brownian motion referring to the initial part of the curve which has been eliminated by the map $h_{\delta t}$. Expanding Eq. (34) to first order in δt and noting that $\langle(\delta B_t)^2\rangle = \delta t$ we arrive at a partial differential equation for $P(x,y,\epsilon)$

$$\left(\frac{2x}{x^2+y^2}\frac{\partial}{\partial x} - \frac{2y}{x^2+y^2}\frac{\partial}{\partial y} + \frac{\kappa}{2}\frac{\partial^2}{\partial x^2} - \frac{2(x^2-y^2)}{(x^2+y^2)^2}\epsilon\frac{\partial}{\partial\epsilon}\right)P = 0 \,. \tag{35}$$

Since $P(x,y,\epsilon) \propto \epsilon^{2-D}$ we have $\epsilon\partial P/\partial\epsilon = (2-D)P$ and the determination of D is reduced to an eigenvalue problem. By inspection one finds

$$P \propto \epsilon^{1-\kappa/8} y^{(\kappa-8)^2/8\kappa}(x^2+y^2)^{(\kappa-8)/2\kappa} \,, \tag{36}$$

and we identify the fractal dimension $D = 1 + \kappa/8$ for $\kappa < 8$; for $\kappa > 8$ another solution yields $D = 2$.

Results and Discussion

Stochastic Loewner evolution based on Eq. (27) generates conformally invariant non-crossing random curves in the upper half plane starting at the origin and going off to infinity. This is the case of chordal SLE, where the random curve connects two boundary points (the origin **O** and infinity ∞). Another case is radial SLE for random curves connecting a boundary point and an interior point in a simply connected domain [13,27,43]. Radial SLE is governed by another stochastic equation and will not be discussed here.

Phase Transitions – Locality – Restriction – Duality

Phase Transitions SLE exhibits two phase transitions; for $\kappa = 4$ and $\kappa = 8$. For $0 < \kappa \leq 4$ the random curve is non-intersecting, i. e., a simple random continuous curve from **O** to ∞. For $4 < \kappa \leq 8$ the curve is self-intersecting on all scales. The curve together with the excluded regions form a hull which in the course of time absorbs the upper half plane. For κ just above 4 the half plane is eventually absorbed but the trace does not visit all regions, i. e., the hull is not dense. As we approach $\kappa = 8$ the trace becomes more dense and the hull becomes plane-filling. This is also reflected in the fractal dimension $D = 1 + \kappa/8$. For $\kappa > 8$ the hull is plane-filling, i. e., $D = 2$. As we increase the strength of the Brownian drive further the excursions of the trace to the right and left in the upper half plane become more pronounced and the hull becomes vertically compressed. These results have been obtained by Rohde and Schramm [74] and Lawler et al. [57]. The various phases of SLE are depicted in Fig. 8. In Fig. 9 we have depicted numerical renderings of SLE traces for various values of κ (with permission from V. Beffara, http://www.umpa.ens-lyon.fr/~vbeffara/simu.php).

Locality – Restriction In addition to the phase transitions at $\kappa = 4$ and $\kappa = 8$, there are special values of κ where SLE shows a behavior characteristic of the scaling limit of specific lattice models: The locality property for $\kappa = 6$ and the restriction property for $\kappa = 8/3$. The issue here is the influence of the boundary on the SLE trace.

To illustrate the locality property, consider for example the SLE trace originating at the origin and purporting to describe the scaling limit of a domain wall in the lattice model. Due to the long range correlations at the critical point it is intuitively clear that a deformation of the boundary, e. g., a bulge **L** on the real axis to the right of the origin, will influence the trace and push it to the left. A detailed analysis show that only for $\kappa = 6$ is the trace independent of a change of the boundary, i. e., the trace does not feel the boundary until it encounters a boundary point [59,63]. Returning to the lattice models the locality property for $\kappa = 6$ applies specifically to the percolation case where the interface generated by the exploration process is governed by a local rule and the model has a geometric phase transition.

The restriction property is less obvious to visualize but basically states that the distribution of traces conditioned not to hit a bulge **L** on the real axis away from the origin is the same as the distribution of traces in the domain where **L** is part of the boundary, i. e., in the domain $\mathbf{H} \setminus \mathbf{L}$. Analysis shows that the restriction property only applies in the case for $\kappa = 8/3$. Among the lattice models only the scaling limit of self-avoiding random walk (SAW), where the measure is uniform, conforms to the restriction property and thus corresponds to $\kappa = 8/3$ [63].

Duality For $\kappa > 4$ the SLE generates a hull of fractal dimension $D > 3/2$. The boundary, external perimeter, or frontier of the hull is again a simple conformally invariant random curve characterized by the fractal dimension $\bar{D}$. Using methods from 2D quantum gravity Duplantier [31] has proposed the relationship,

$$(D-1)(\bar{D}-1) = \tfrac{1}{4}\,, \tag{37}$$

between the fractal dimension of the hull and its frontier. This result has been proved by Beffara for $\kappa = 6$, i. e., the percolation case [16]. Inserting in Eq. (33) we obtain for the corresponding SLE parameter the duality relation

$$\kappa\bar{\kappa} = 16\,. \tag{38}$$

Loop Erased Random Walk (LERW)

Whereas the scaling limit of random walk, i. e., Brownian motion, does not fall in the SLE category because of self-crossings rendering Riemann's mapping theorem inapplicable, variations of Brownian motion are described by SLE.

Loop erased random walk (LERW) where loops are removed along the way is by construction self-avoiding and was introduced as a simple model of a self-avoiding random walk. LERW was studied by Schramm in his pioneering work [77]. LERW has the Markov property and has been proved to be conformally invariant in the scaling limit and described by SLE for $\kappa = 2$ [57]. According to Eq. (33) LERW has the fractal dimension $D = 5/4$ [67]. Also, since $\kappa < 4$ LERW is non-intersecting. A simulation of LERW based on SLE is shown in Fig. 9a.

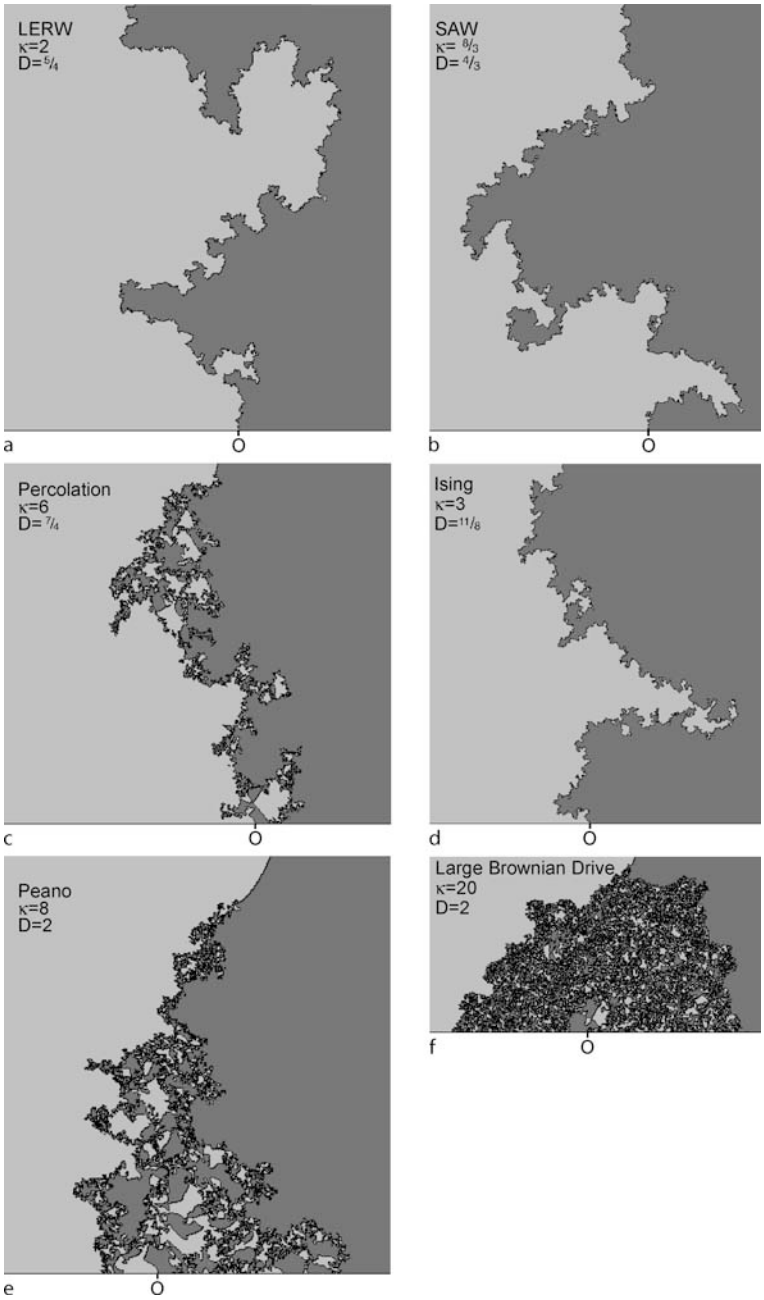

Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 9
We depict a numerical rendering of SLE for a variety of κ values. In **a** we show loop erased random walk (LERW) for $\kappa = 2$ with fractal dimension $D = 5/4$. In **b** we illustrate the case of self-avoiding random walk (SAW) for $\kappa = 8/3$ and fractal dimension $D = 4/3$; both LERW and SAW have $\kappa > 4$ and are simple scale invariant random curves. In **c** we depict site percolation for $\kappa = 6$ with fractal dimension $D = 7/4$. Since $\kappa > 4$ the percolation case is self-intersecting and duality implies that the boundary or frontier of the hull is described by a SLE curve for $\kappa = 16/6 = 8/3$, i. e., the case of SAW. In **d** we show the Ising case for $\kappa = 3$ and fractal dimension $D = 11/8$. In **e** we depict the limiting case $\kappa = 8$ and fractal dimension $D = 2$. The hull is dense and plane-filling. The frontier of the hull corresponds to the SLE case $\kappa = 16/8 = 2$, i. e., the case of LERW. The so-called uniform spanning tree (UST) has the same properties as LERW and the SLE case for $\kappa = 8$ can thus be thought of as a random plane filling Peano curve wrapping around the UST. Finally, in **f** we show the SLE trace and hull for $\kappa = 20$ and $D = 2$. Because of the large Brownian excursions the plane-filling hull is vertically compressed (with permission from V. Beffara: http://www.umpa.ens-lyon.fr/~vbeffara/simu.php)

Self-Avoiding Random Walk (SAW)

Self-avoiding random walk (SAW) is a random walk conditioned not to cross itself. SAW has been used to model polymers in a dilute solution and has a uniform probability measure. Since SAW satisfies the restriction property it is conjectured in the scaling limit to fall in the SLE class with $\kappa = 8/3$ [46,47,48,61], yielding the fractal dimension $D = 4/3$. We note that Flory's mean field theory [30] for the size R of a polymer composed of N links (monomers) scales like $R \sim N^{\nu}$, where $\nu = 3/(2+d)$ for $d \leq 4$. By a box covering we infer $N \sim R^D$ where D is the fractal dimension, i. e., $D = (2+d)/3$. In $d = 2$ we obtain $D = 4/3$ in accordance with the SLE result. SLE induced SAW in the scaling limit is shown in Fig. 9b.

Percolation

The scaling limit of site percolation was conjectured by Schramm [77] to fall in the SLE class for $\kappa = 6$. Subsequently, the scaling limit of site percolation on a triangular lattice has been proved by Smirnov [78,80]. Percolation exhibits a geometrical phase transition. In the exploration process defining a critical interface the rule for propagation is entirely local. The lack of stiffness as for example in the Ising case to be discussed below results in a strongly meandering path winding back and in the scaling limit intersecting earlier part of the path. Since $\kappa > 4$ the path together with the enclosed part, i. e., the hull, eliminates the whole plane in the course of time. As discussed above the locality property is specific to percolation and yields $\kappa = 6$. The fractal dimension of the percolation interface is according to Eq. (33) $D = 7/4$. We note that D is close to 2, i. e., the percolation interface nearly covers the plane densely. A series of new results and a proof of Cardy's conjectured formula for the crossing probability have appeared; we refer to [13,27,43] for details. Using the duality relation (38) the frontier of the percolation hull is a simple SLE curve for $\kappa = 8/3$, corresponding to SAW. In Fig. 9c we have depicted a SLE generated percolation interface.

Ising Model – $O(n)$ Models

The Ising model in Eq. (6) is a special case of the $O(n)$ model defined by the Hamiltonian

$$H = -J \sum_{\langle ij \rangle} \vec{\sigma}_i \vec{\sigma}_j \,, \tag{39}$$

where $\vec{\sigma}_i = (\sigma_1, \ldots \sigma_n)$ is an n-component unit vector associated with the site i. For $n = 1$ we recover the Ising model, $n = 2$ is the XY-model [29], and $n = 3$ the Heisenberg model.

By means of the Fortuin–Kasteleyn (FK) transformation based on a high temperature expansion the configurations of the $O(n)$ model can be described by clusters or graphs on a dual lattice [35]. The crossing domain wall in Fig. 1 is thus a special case of a FK graph if we interpret the representation as a triangular Ising model. It has been conjectured that n is related to the SLE parameter κ by

$$n = -2\cos(4\pi/\kappa) \quad \text{for} \;\; 8/3 \leq \kappa \leq 4 \,. \tag{40}$$

In the Ising case $n = 1$ and we have $\kappa = 3$ yielding the fractal dimension $D = 11/8$ for the Ising domain wall [84]. Since $\kappa < 4$ the Ising domain wall is non-intersecting. Unlike the percolation case, the Ising interface is stiffer due to the interaction. We also note that the scaling limit of spin cluster boundaries in the Ising model recently has been proved to correspond to SLE for $\kappa = 3$ [79]. The interface is shown in Fig. 9d.

SLE – Conformal Field Theory

Whereas conformal field theory (CFT) is based on the concept of a local field $\phi(r)$ and its correlations and therefore only accesses the underlying geometry indirectly through field correlations, SLE directly produces conformally invariant geometrical objects. A major issue is therefore the connection between CFT and SLE [6,7,26]. In CFT the central charge c plays an important role in delimiting the universality classes of the variety of lattice models yielding conformal field theories in the scaling limit. Percolation thus corresponds to the central charge $c = 0$, whereas the Ising model is associated with the central charge $c = 1/2$. It has been conjectured that the connection between the SLE parameter κ and the central charge c is given by

$$c = \frac{(6-\kappa)(3\kappa-8)}{2\kappa} = 1 - 6\frac{(\kappa-4)^2}{4\kappa} \,. \tag{41}$$

We note that $c < 1$ and, moreover, invariant under the duality transformation $\kappa \to 16/\kappa$.

SLE – 2D Turbulence

There is an interesting application of SLE ideas in the context of 2D turbulence. The issue here is to analyze conformal invariance by comparing the statistical properties of geometrical shapes like domain walls with SLE traces with the view of determining the SLE parameter κ and the corresponding universality class.

In 3D turbulence is governed by the incompressible Navies–Stokes equation for the velocity field. Since the viscosity is only effective at small length scales 3D turbulence is characterized by a cascade of kinetic energy $(1/2)v^2$ from large scales (driving scale) to small scales (dissipation scale). In the inertial regime the energy spectrum $E(k)$ (k is the wavenumber) is characterized by the celebrated Kolmogorov 5/3 law [46], $E(k) \propto k^{-5/3}$, indicating an underlying scale invariance in turbulence.

In 2D the cascade picture is different. Since both kinetic energy and squared vorticity (enstrophy) are conserved in the absence of dissipation and forcing, two cascades coexist [52,53]. A direct cascade to small scales for the squared vorticity $\omega^2 = (\nabla \times v)^2$ with scaling exponent -3 and an inverse cascade to larger scales for the kinetic energy $(1/2)v^2$ with Kolmogoroff scaling exponent $-5/3$. The system is thus characterized by a fine scale vorticity structure together with a large scale velocity structure. Moreover, we can assume that the vorticity structure is equipartitioned, i.e, in equilibrium.

In order to investigate whether the scale invariance of the small scale vorticity structure can be extended to conformal invariance Bernard et al. [17] have considered the statistics of the boundaries of vorticity clusters. By comparing the zero-vorticity isolines with SLE traces they find that cluster boundaries fall in the universality class corresponding to $\kappa = 6$, i. e., the case of percolation. Since 2D turbulence is a driven nonequilibrium system, this observation is very intriguing in particular since the correlations between vortices are long-ranged. A similar analysis [19] of the isolines in the inverse cascade in surface quasi-geostrophic turbulence corresponds to $\kappa = 4$, i. e., from Eq. (40) the domain walls in the equilibrium XY model for $n = 2$. For comments on the application of SLE in turbulence we refer to Cardy [28].

SLE – 2D Spin Glass

It is a standing issue whether conformal field theory can be applied to disordered systems, in particular systems with quenched disorder. In recent work Amoruso et al. [3] and Bernard et al. [18] have considered zero temperature domain walls in the Ising spin glass [34]; see also [33]. The Ising spin glass is an equilibrium system with quenched disorder. The system is described by the Hamiltonian $H = -\sum_{\langle ij \rangle} J_{ij}\sigma_i\sigma_j$, where the random exchange constants J_{ij} are picked from a Gaussian distribution with zero mean. The glass transition is at $T = 0$ and the system has a two-fold degenerate ground state. Inducing a scale invariant domain wall between the two ground states and comparing with an SLE trace, it is found that both the Markov and conformal properties are obeyed and that the universality class corresponds to $\kappa \approx 2.3$.

Further Remarks

In this discussion we have left out several topics which have played an important role in the development and applications of SLE. We mention some of them below.

There is an interesting connection between LERW and the so-called uniform spanning tree (UST) [13,43,57,77]. A spanning tree is a collection of vertices and edges which form a tree, i. e., without loops or cycles. A uniform spanning tree is a random spanning tree picked among all possible spanning trees with equal probability. Consider the unique path between two vertices on a UST. Since the path lives on a tree it is by construction non-crossing and it turns out that it has the same distribution as LERW. The winding random curve enclosing the UST can be visualized as a random plane-filling Peano curve. In the scaling limit the Peano curve is described by SLE for $\kappa = 8$ with fractal dimension $D = 2$.

The q-state Potts model [87] constitutes a generalization of the Ising model; here the lattice variable takes q values. The model is defined by the Hamiltonian $H = -J\sum_{\langle ij \rangle} \delta_{\sigma_i\sigma_j}$, where $\sigma_i = 1, \ldots q$; the Ising model obtains for $q = 2$. Applying the high temperature FK representation the configurations can be represented by loops and domain walls. From considerations involving the fractal dimension [76] it has been conjectured that domain walls in the scaling limit of the Potts model fall in the SLE category for $q = 2 + 2\cos(8\pi/\kappa)$, where $4 \le \kappa \le 8$. For $q = 2$ we recover the Ising case for $\kappa = 3$. In the limit $q \to 0$, the graph representation is equivalent to the uniform spanning tree described by SLE for $\kappa = 8$. For a numerical study of the three-state Potts model and its relation to SLE consult [36].

Standard SLE is driven by 1D Brownian motion producing a fractal curve. Ruskin et al. [75] have considered the case of adding a stable Lévy process with shape parameter α to the Brownian motion. Backing their analysis with numerics they find that the SLE trace branches and exhibit a 'phase transitions' related to self-intersections.

2D Brownian motion, the scaling limit of 2D random walk, is an incredibly complex fractal coil owing to the self-crossings on all scales. Although 2D Brownian motion because of self-crossing itself falls outside the SLE scheme, the outer frontier or perimeter of 2D random walk is a non-crossing and non-intersecting fractal curve which can be accessed by SLE. Verifying an earlier conjecture by Mandelbrot [68] it has been proven using SLE techniques [58] that the fractal dimension of the Brownian

perimeter is $D = 4/3$, i.e, the same as the fractal dimension of self-avoiding random walk and the external perimeter of the percolation hull. Other characteristics of Brownian motion such as intersection exponents have also been obtained [59,60,62]; see also [66].

In recent work Zoia et al. [88] have considered the distribution of first passage times and distances along critical curves generated by SLE for different values of κ.

Future Directions

Stochastic Loewner evolution represents a major step in our understanding of fractal shapes in the 2D continuum limit. By combining the Markov property (stationarity) with conformal invariance SLE provides a minimal scheme for the generation of a one-parameter family of fractal curves. The SLE scheme also provides calculational tools which have led to a host of new results. SLE is a developing field and we can on the mathematical front anticipate progress and proofs of some yet unproved scaling limits, e. g., the scaling limit of the FK representation of the Potts model and the scaling limit of SAW.

On the more physical front many issues also remain open. First there is the fundamental issue of the connection between the hugely successful but non-rigorous CFT and SLE. Here progress is already under way. In a series of papers Bauer and Bernard [6,7,8,9,10,12] have shown how SLE results can be derived using CFT methods. Cardy [26] have considered a multiple SLE process and the connection to Dyson's Brownian process and random matrix theory. The analysis of the CFT–SLE connection still remains to be investigated further.

An obvious limitation of SLE is that it only addresses critical domain walls and not the full configuration of clusters and loops in for example the FK representation of the Potts model. In the case of critical percolation this problem has been addressed by Camia and Newman [21]. Another issue is how to provide SLE insight into spin correlations in the Potts or $O(n)$ models.

In the original formulation of SLE the Markov and conformal properties essentially requires a Brownian drive. It is clearly of interest to investigate the properties of random curves generated by other random drives. Such a program has been initiated by Ruskin et al. [75] who considered adding a Lévy drive to the Brownian drive; see also work by Kennedy [49,50].

Since the SLE trace lives in the infinite upper half plane the whole issue of finite size effects remain open. In ordinary critical phenomena the concept of a Kadanoff block construction and the diverging correlation length near the transition lead to a theory of finite size scaling and corrections to scaling which can be accessed numerically. It is an open problem how to develop a similar scheme for SLE.

In statistical physics it is customary and natural to associate a free energy to a domain wall and an interaction energy associated with several domain walls. These free energy considerations are entirely absent in the SLE framework which is based on conformal transformations. A major issue is thus: Where is the free energy in all this and how do we reintroduce and make use of ordinary physical considerations and estimates [69].

Acknowledgments

The present work has been supported by the Danish Natural Science Research Council.

Bibliography

1. Ahlfors LV (1966) Complex analysis: An introduction to the theory of analytical functions of one complex variable. McGraw-Hill, New York
2. Ahlfors LV (1973) Conformal invariance: Topics in geometric function theory. McGraw-Hill, New York
3. Amoruso C, Hartmann AK, Hastings MB, Moore MA (2006) Conformal invariance and stochastic Loewner evolution processes in two-dimensional Ising spin glasses. Phys Rev Lett 97(4):267202. arXiv:cond-mat/0601711
4. Ash RB, Doléans CA (2000) Probability and measure theory. Academic, San Diego
5. Bak P (1999) How nature works: The science of self-organized criticality. Springer, New York
6. Bauer M, Bernard D (2002) SLE$_\kappa$ growth processes and conformal field theory. Phys Lett B 543:135–138. arXiv:math.PR/0206028
7. Bauer M, Bernard D (2003) Conformal field theories of stochastic Loewner evolutions. Comm Math Phys 239:493–521. arXiv:hep-th/0210015
8. Bauer M, Bernard D (2003) SLE martingales and the Viasoro algebra. Phys Lett B 557:309–316. arXiv:hep-th/0301064
9. Bauer M, Bernard D (2004) CFTs of SLEs: The radial case. Phys Lett B 583:324–330. arXiv:math-ph/0310032
10. Bauer M, Bernard D (2004) Conformal transformations and the SLE partition function martingale. Ann Henri Poincare 5:289–326. arXiv:math-ph/0305061
11. Bauer M, Bernard D (2004) Loewner chains. arXiv:cond-mat/0412372
12. Bauer M, Bernard D (2004) SLE, CFT and zig-zag probabilities. Proceedings of the conference 'Conformal Invariance and Random Spatial Processes', Edinburgh, July 2003. arXiv:math-ph/0401019
13. Bauer M, Bernard D (2006) 2D growth processes: SLE and Loewner chains. Phys Rep 432:115–221
14. Baxter RJ (1982) Exactly solved models in statistical mechanics. Academic, London
15. Beffara V (2002) The dimension of SLE curves. arXiv:math.PR/0211322

16. Beffara V (2003) Hausdorff dimensions for SLE_6. Ann Probab 32:2606–2629. arXiv:math.PR/0204208
17. Bernard D, Boffetta G, Celani A, Falkovich G (2006) Conformal invariance in two-dimensional turbulence. Nat Phys 2:124–128
18. Bernard D, Le Doussal P, Middleton AA (2006) Are domain walls in 2D spin glasses described by stochastic Loewner evolutions. arXiv:cond-mat/0611433
19. Bernard D, Boffetta G, Celani A, Falkovich G (2007) Inverse turbulent cascades and conformally invariant curves. Phys Rev Lett 98:024501(4). arXiv:nlin.CD/0602017
20. Binney JJ, Dowrick NJ, Fisher AJ, Newman MEJ (1992) The theory of critical phenomena. Clarendon, Oxford
21. Camia F, Newman CM (2003) Continuum nonsimple loops and 2D critical percolation. arXiv:math.PR/0308122
22. Cardy J (1987) Conformal invariance. In: Domb C, Lebowitz JL (eds) Phase transitions and critical phenomena, vol 11. Academic, London
23. Cardy J (1993) Conformal field theory comes of age. Physics World, June, 6:29–33
24. Cardy J (1996) Scaling an renormalization in statistical physics. Cambridge University Press, Cambridge
25. Cardy J (2002) Conformal invariance in percolation, self-avoiding walks and related problems. Plenary talk given at TH-2002, Paris. arXiv:cond-mat/0209638
26. Cardy J (2003) Stochastic Loewner evolution and Dyson's circular ensembles. J Phys A 36:L379–L408. arXiv:math-ph/0301039
27. Cardy J (2005) SLE for theoretical physicists. Ann Phys 318:81–118. arXiv:cond-mat/0503313
28. Cardy J (2006) The power of two dimensions. Nat Phys 2:67–68
29. Chaikin PM, Lubensky TC (1995) Principles of condensed matter physics. Cambridge University Press, Cambridge
30. de Gennes PG (1985) Scaling concepts in polymer physics. Cornell University Press, Ithaca
31. Duplantier B (2000) Conformally invariant fractals and potential theory. Phys Rev Lett 84:1363–1367. arXiv:cond-mat/9908314
32. Feder J (1988) Fractals (physics of solids and liquids). Springer, New York
33. Fisch R (2007) Comment on conformal invariance and stochastic Loewner evolution processes in two-dimensional Ising spin glasses. arXiv:0705.0046
34. Fischer KH, Hertz JA (1991) Spin glasses. Cambridge University Press, Cambridge
35. Fortuin CM, Kasteleyn PW (1972) On the random cluster model. Physica 57:536–564
36. Gamsa A, Cardy J (2007) SLE in the three-state Potts model – a numerical study. arXiv:0705.1510
37. Gardiner CW (1997) Handbook of stochastic methods. Springer, New York
38. Gong S (1999) The Bieberbach conjecture. RI American 19. Mathematical Society. International, Providence
39. Gruzberg IA, Kadanoff LP (2004) The Loewner equation: Maps and shapes. J Stat Phys 114:1183–1198. arXiv:cond-mat/0309292
40. Jensen HJ (2000) Self-organized criticality: Emergent complex behavior in physical and biological systems. Cambridge University Press, Cambridge
41. Kadanoff LP (1966) Scaling laws for Ising models near T_c. Physics 2:263–271
42. Kadanoff LP, Berkenbusch MK (2004) Trace for the Loewner equation with singular forcing. Nonlinearity 17:R41–R54. arXiv:cond-mat/0402142
43. Kager W, Nienhuis B (2004) A guide to stochastic Loewner evolution and its application. J Stat Phys 115:1149–1229
44. Kager W, Nienhuis B, Kadanoff LP (2004) Exact solutions for loewner evolutions. J Stat Phys 115:805–822
45. Kauffman SA (1996) At home in the universe: The search for the laws of self-organization and complexity. Oxford University Press, Oxford
46. Kennedy T (2002) Monte Carlo tests of stochastic Loewner evolution predictions for the 2D self-avoiding walk. Phys Rev Lett 88(4):130601. arXiv:math.PR/0112246
47. Kennedy T (2004) Conformal invariance and stochastic Loewner evolution predictions for the 2D self-avoiding walk – Monte Carlo tests. J Stat Phys 114:51–78. arXiv:math.PR/0207231
48. Kennedy T (2005) Monte Carlo comparisons of the self-avoiding walk and SLE as parameterized curves. arXiv:math.PR/0510604v1
49. Kennedy T (2006) The length of an SLE – Monte Carlo studies. arXiv:math.PR/0612609v1
50. Kennedy T (2007) Computing the Loewner driving process of random curves in the half plane. arXiv:math.PR/0702071v1
51. Kolmogorov AN (1941) Dissipation of energy in the locally isotropic turbulence. Dokl Akad Nauk SSSR 30:9–13. (Reprinted in Proc Royal Soc Lond A 434:9–13 (1991))
52. Kraichnan RH (1967) Inertial ranges in two-dimensional turbulence. Phys Fluids 10:1417–1423
53. Kraichnan RH, Montgomery D (1980) Two-dimensional turbulence. Rep Prog Phys 43:567–619
54. Landau LD, Lifshitz EM (1959) Theory of elasticity. Pergamon, Oxford
55. Lawler GF (2004) ICTP Lecture Notes Series. 17:307–348
56. Lawler GF (2005) Conformally invariant processes in the plane. In: Mathematical Surveys and Monographs, vol 114. AMS, Providence
57. Lawler GF, Schramm O, Werner W (2001) Conformal invariance of planar loop-erased random walks and uniform spanning trees. Ann Prob 32:939–995. arXiv:math.PR/0112234
58. Lawler GF, Schramm O, Werner W (2001) The dimension of the planar Brownian frontier is 4/3. Math Res Lett 8:401–411. arXiv:math.PR/00010165
59. Lawler GF, Schramm O, Werner W (2001) Values of Brownian intersections exponents I: Half plane exponents. Acta Math 187:237–273. arXiv:math.PR/9911084
60. Lawler GF, Schramm O, Werner W (2001) Values of Brownian intersections exponents II: Plane exponents. Acta Math 187:275–308. arXiv:math.PR/0003156
61. Lawler GF, Schramm O, Werner W (2002) On the scaling limit of planar self-avoiding walk. Fractal geometry and application, a jubilee of Benoit Mandelbrot, Part 2, 339–364, Proc. Sympos. Pure Math., 72, Part 2, Amer. Math. Soc., Providence, RI, 2004. arXiv:math.PR/0204277
62. Lawler GF, Schramm O, Werner W (2002) Values of Brownian intersections exponents III: Two-sided exponents. Ann Inst Henri Poincare 38:109–123. arXiv:math.PR/0005294
63. Lawler GF, Schramm O, Werner W (2003) Conformal restriction: The chordal case. J Amer Math Soc 16:917–955. arXiv:math.PS/0209343

64. Löwner K (Loewner C) (1923) Untersuchungen über schlichte konforme Abbildungen des Einheitskreises. I Math Ann 89:103–121
65. Ma S-K (1976) Modern theory of critical phenomena. Frontiers in Physics, vol 46. Benjamin, Reading
66. Mackenzie D (2000) Taking the measure of the wildest dance on earth. Science 290:1883–1884
67. Majumbar SN (1992) Exact fractal dimension of the loop-erased self-avoiding walk in two dimensions. Phys Rev Lett 68:2329–2331
68. Mandelbrot B (1987) The fractal geometry of nature. Freeman, New York
69. Moore M (2007) private communication
70. Nicolis G (1989) Exploring complexity: An introduction. Freeman, New York
71. Nienhuis B (1987) Coulomb gas formulation of two-dimensional phase transitions. In: Domb C, Lebowitz JL (eds) Phase transitions and critical phenomena, vol 11. Academic, London
72. Pfeuty P, Toulouse G (1977) Introduction to the renormalization group and to critical phenomena. Wiley, New York
73. Reichl LE (1998) A modern course in statistical physics. Wiley, New York
74. Rohde S, Schramm O (2001) Basic properties of SLE. Ann Math 161:879–920. arXiv:mathPR/0106036
75. Rushkin I, Oikonomou P, Kadanoff LP, Gruzberg IA (2006) Stochastic Loewner evolution driven by Levy processes. J Stat Mech (2006) 01:P01001. arXiv:cond-mat/0509187
76. Saleur H, Duplantier B (1987) Exact determination of the percolation hull exponent in two dimensions. Phys Rev Lett 58:2325–2328
77. Schramm O (2000) Scaling limit of loop-erased random walks and uniform spanning trees. Israel J Math 118:221–288. arXiv:math.PR/9904022
78. Smirnov S (2001) Critical percolation in the plane: Conformal invariance, Cardy's formula, scaling limits. C R Acad Sci Paris Ser I Math 333(3):239–244
79. Smirnov S (2006) Towards conformal invariance of 2D lattice models. Proceedings of the international congress of mathematicians (Madrid, August 22–30, 2006). Eur Math Soc 2:1421–1451
80. Smirnov S, Werner W (2001) Critical exponents for two-dimensional percolation. Math Res Lett 8:729–744
81. Stanley HE (1987) Introduction to phase transitions and critical phenomena. Oxford University Press, Oxford
82. Stauffer D, Aharony A (1994) Introduction to percolation theory. CRC, Boca Raton
83. Strogatz S (2003) Sync: The emerging science of spontaneous order. Hyperion, New York
84. Vanderzande C, Stella AL (1989) Bulk, surface and hull fractal dimension of critical Ising clusters in $d = 2$. J Phys A: Math Gen 22:L445–L451
85. Werner W (2004) Random planar curves and Schramm–Loewner evolutions. Springer Lecture Notes in Mathematics 1840:107–195. arXiv: math.PR/0303354
86. Wilson KG, Kogut J (1974) The renormalization group and the ε expansion. Phys Rep 12:75–199
87. Wu FY (1982) The Potts model. Rev Mod Phys 54:235–268
88. Zoia A, Kantor Y, Kardar M (2007) Distribution of first passage times and distances along critical curves. arXiv:0705.1474v1 [cond-mat.stat-mech]

Stochastic Models of Biological Processes

Steven S. Andrews[1], Tuan Dinh[1], Adam P. Arkin[1,2]

[1] Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, USA

[2] Department of Bioengineering and Howard Hughes Medical Institute, University of California, Berkeley, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Non-Spatial Stochastic Modeling
Spatial Stochastic Modeling
Future Directions
Box 1: The *E. Coli* Min System
Acknowledgments
Bibliography

Glossary

Brownian dynamics A level of detail in which each molecule is represented by a point-like particle and molecules move in response to diffusion and collisions

Chemical Fokker–Planck equation (CFPE) Master equation for well-mixed systems that corresponds to the chemical Langevin equation

Chemical master equation (CME) Master equation for the probability that the system has specific integer copy numbers for each type of chemical species; it is exact for a well-mixed system

Chemical Langevin equation (CLE) Approximate stochastic differential equation for well-mixed systems which is based on continuous Gaussian statistics

Direct method An implementation of the Gillespie algorithm

Extrinsic noise In genetic noise studies, expression fluctuations of a gene that arise from upstream genes or global fluctuations

First-reaction method An implementation of the Gillespie algorithm

Gillespie algorithm Exact algorithm for simulating individual trajectories of the CME

Hybrid algorithms Algorithms that are designed to efficiently simulate systems that have multiple timescales

Individual-based spatial models Models that track individual molecules as they diffuse or react

Intrinsic noise Expression fluctuations of a gene that arise from that particular gene

Jump process A process in which the system abruptly changes from one state to another

Optimized direct method A computationally efficient implementation of the Gillespie algorithm

Population-based spatial models Models that track how many molecules of each chemical species are in various spatial compartments

Reaction channel A possible reaction between specified reactant and product chemical species (the terminology distinguishes this meaning from an individual reaction event between single molecules)

Reaction-diffusion equation Deterministic partial differential equation that combines mass action reaction kinetics and normal chemical diffusion

Reaction-diffusion master equation (RDME) Chemical master equation that accounts for diffusion as well as reactions

Reaction rate equation (RRE) Deterministic ordinary differential equation for the net production rate of each chemical species from chemical reactions

Spatial chemical Langevin equation Chemical Langevin equation that accounts for diffusion as well as reactions

Stochastic simulation algorithm (SSA) Alternative term for the Gillespie algorithm

Stoichiometric matrix (ν) Matrix that gives the net production of each chemical species, for each chemical reaction

Tau-leaping method Approximate simulation method for well-mixed systems in which molecule numbers are updated using discrete Poisson statistics

Well-mixed hypothesis Assumption that mixing processes occur faster than the relevant reaction processes

Definition of the Subject

Many processes in cell biology, such as those that carry out metabolism, the cell cycle, and various types of signaling, are comprised of biochemical reaction networks. It has proven useful to study these networks using computer simulations because they allow us to quantitatively investigate hypotheses about the networks. Deterministic simulations are sufficient to predict average behaviors at the population level, but they cannot address questions about noise, random switching between stable states of the system, or the behaviors of systems with very few molecules of key species. These topics are investigated with stochastic simulations. In this article, we review the dominant types of stochastic simulation methods that are used to investigate biochemical reaction networks, as well as some of the results that have been found with them. As new biological experiments continue to reveal more detail about biological systems, and as computers continue to become more powerful, researchers will increasingly turn to simulation methods that can address stochastic and spatial details.

Introduction

Random events are ubiquitous throughout biology. Diffusion, chemical reactions, gene expression, homologous recombination, and most other fundamental biological processes are governed to a large extent by the inherently discrete and stochastic interactions of molecules [1]. In many cases, the random events that occur on very small length and time scales become averaged out when one focuses on larger length or time scales. However, there also exist many examples in which stochastic fluctuations at small scales propagate up to and then influence the system behavior at large scales. Examples range from the swimming trajectories of individual bacteria all the way up to the genetic diversity on which evolution depends.

Research on stochasticity in biochemical systems has received a great deal of attention lately, leading to many recent reviews [2,3,4,5,6,7,8,9]. One reason for its popularity is that the basic designs of many biochemical systems, such as metabolism, cell division, and chemotaxis, are becoming reasonably well understood. Starting from this understanding, researchers are delving deeper to examine the quantitative behaviors of these systems, including the roles of stochastic influences. Also, there is an increasing awareness of the importance of stochasticity in biological systems. For example, it has become clear that noisy gene expression is the rule rather than the exception [9]; this leads to important questions about non-genetic individuality and about biological robustness to gene expression noise. Thirdly, stochasticity is often investigated using computationally demanding simulations. Cheaper and faster computers, as well as improved simulation algorithms, are making it feasible for researchers to investigate more complex biochemical systems, at increasingly realistic levels of detail. Finally, and perhaps most importantly, the last ten years have witnessed incredible progress in experimental biochemical methods, some of which allow direct measurements of stochasticity on microscopic size scales. These methods include gene expression measurements [10,11], flow cytometry [12,13,14,15,16], single molecule detection methods [17,18,19,20], and applications of synthetic biology [21].

At the most fundamental level, the quantum mechanics that describe the dynamics of all physical systems are well-understood and completely deterministic [22]. It is

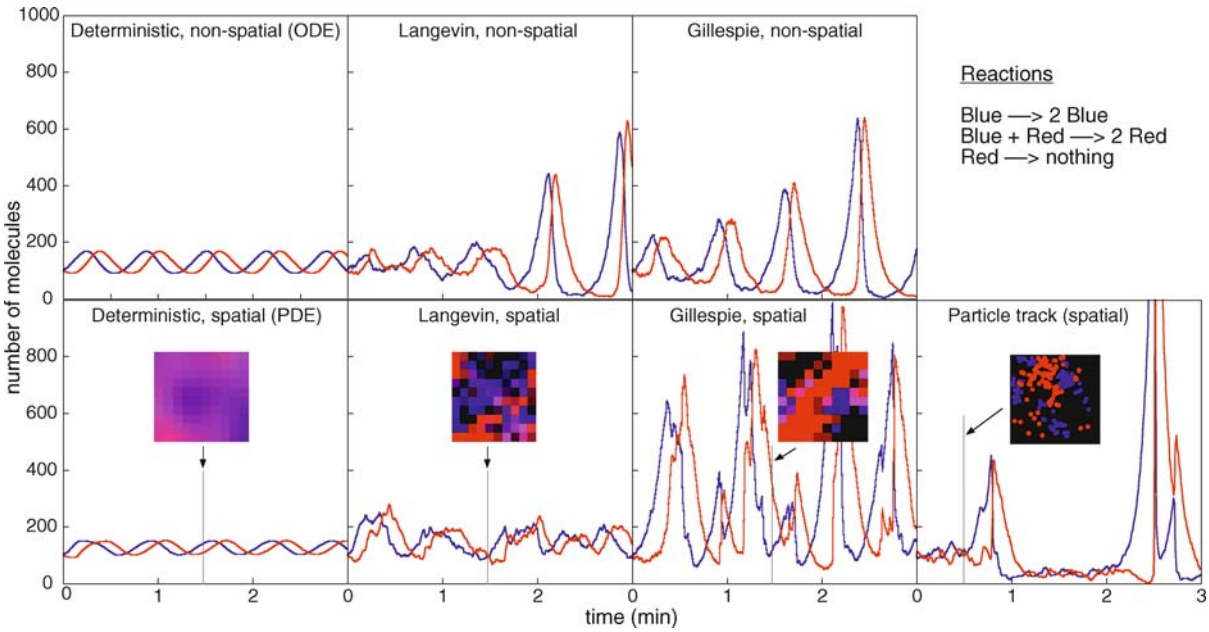

Stochastic Models of Biological Processes, Figure 1

Simulation results for a simple chemical oscillator using different simulation methods. The Lotka–Volterra system is shown, which shares key features with cellular oscillators such as circadian rhythms. *Insets* show the spatial distributions of molecules at the indicated times. In the *top panels*, note that stochasticity allows the system to drift to large amplitude oscillations and that the Langevin and Gillespie methods yield similar results. In the *bottom panels*, all of which were started with nearly homogeneous initial states, differences arise from the approximations: the PDE simulation has predictable oscillations due to the minimal stochasticity (which is only in the initial state); the Gillespie simulation has larger peaks than the Langevin one because it only allows integer numbers of molecules in each bin; and the particle tracking simulation shows larger and fewer bursts than does the Gillespie simulation because it accurately treats diffusion at all length scales (this difference was reduced with a spatial Gillespie simulation that used smaller subvolumes). Parameters: rate constants are 10 min^{-1}, 8000 nm^3 $molec^{-1}$ min^{-1}, and 10 min^{-1}, for the respective reactions shown in the *top-right corner*, systems start with 100 of each *blue* and *red* molecules, their diffusion coefficients are 100 nm^2 min^{-1}, the volume is 100 nm high and wide by 10 nm deep, and the first three spatial simulations divide this volume into cubic subvolumes that are 10 nm on a side. This figure is reproduced from [8]

only when systems are observed that there arises unavoidable randomness, although these aspects of quantum mechanics remain murky and as close to philosophy as science. More importantly, essentially any system that is governed by nonlinear dynamics, including nearly all physical systems, rapidly becomes chaotic as the system size is increased beyond a few molecules, and thus becomes effectively unpredictable [23]. For all intents and purposes, the diffusive trajectories of individual molecules, and the probabilities of chemical reactions occurring between neighboring reactants at specific times, are fundamentally stochastic.

In the laboratory, partly by design, this stochasticity usually averages out. For systems comprised of many particles, diffusion is observed to be described well by Fick's law of diffusion and chemical reaction kinetics are described well by the deterministic reaction rate equations [24]. The situation is often different within biological cells for several reasons: (i) the program for cellular behavior, the genome, is present at low copy number and yet each gene governs the expression of possibly thousands of proteins, (ii) the low copy numbers of many proteins and mRNA transcripts within cells make random variation of their numbers a relatively large fraction of the total, (iii) stochasticity is often amplified during sequential biochemical steps, of which an especially important case is the sequence of DNA transcription followed by mRNA translation, and (iv) because of spatial organization within cells, it often takes very few proteins in specific locations to achieve large effects on the entire cell dynamics.

This review provides a broad account of stochastic modeling of biological processes. The emphasis is placed on stochastic processes at the cellular level although much of the work presented here also applies to other scales and systems. Our goal is to briefly familiarize the reader with the mathematical forms of the most important equations,

the tools for analyzing and simulating them, and some applications for which they have been particularly successful. As shown below, the mathematics, the software implementation, and even the applications of modeling methods are often closely linked.

There are several ways to categorize stochastic biochemical modeling methods. A key designation is whether a method is spatial or non-spatial: spatial models treat spatial organization of proteins and membranes explicitly, whereas non-spatial models include an implicit assumption that mixing processes occur faster than the relevant reaction processes, which is called the *well-mixed hypothesis*. No single modeling method can efficiently capture stochastic dynamics over wide ranges of time scales, so separate methods have been developed that operate at levels of temporal detail that range from nanoseconds to hours. Hybrid simulators combine methods that operate at different timescales to allow the efficient simulation of systems that include both fast and slow processes. Whereas many modeling methods are designed solely to address the reactions in a biochemical reaction network, others also consider system boundaries, mechanics, or the multiple states that proteins can be in. Here, we present non-spatial modeling methods first, followed by spatial methods, with diversions along the way to touch on as many of the other topics as possible. Simulation results from many of the methods that are discussed are compared in Fig. 1, where it is seen that the differences can be quite significant.

Non-Spatial Stochastic Modeling

Deterministic Modeling and Notation

Before focusing on stochastic modeling, it is helpful to introduce the notation and some terminology by summarizing a few aspects of deterministic modeling. As applied to chemical reaction networks, deterministic modeling is based upon ordinary differential equations (ODEs) for the individual chemical reactions. Consider the generic elementary reversible reaction

$$A + B \underset{k_r}{\overset{k_f}{\rightleftharpoons}} C \,, \tag{1}$$

where k_f and k_r are the forward and reverse reaction rates, respectively. We assume that the system is kept well-mixed so that diffusion effects can be ignored. The reaction rate equations for components A, B, and C are the ODEs

$$\frac{d[A]}{dt} = \frac{d[B]}{dt} = -k_f[A][B] + k_r[C] \,, \tag{2a}$$

$$\frac{d[C]}{dt} = k_f[A][B] - k_r[C] \,. \tag{2b}$$

More complex reaction networks are expressed analogously, with one equation for each chemical species and with terms in the equations that represent chemical reactions. From these equations, the reactions can be simulated to show how the concentrations change over time. Or, after setting the left sides of the equations to zero, they can be solved to yield the steady-state chemical concentrations. One can also investigate the dynamic or steady-state behaviors as the reaction rate parameters (k_f and k_r), or initial concentrations, are varied [25]; this can yield phase diagrams for the reactions and additional insight.

It is helpful to generalize the rate equations given above to make them more convenient for computational or analytical work and to show their forms more clearly. First, each chemical concentration is replaced by the variable $Z_i(t)$, where i is an index for A, B, or C and the time-dependence is written out explicitly. Next, the product terms in the equations are replaced by the functions $\tilde{a}_f(\mathbf{Z}(t))$ and $\tilde{a}_r(\mathbf{Z}(t))$ for the forward and reverse reactions, respectively; the tildes indicate that molecule quantities are given as concentrations rather than as molecule numbers. These functions are called the *reaction propensities*. Finally, the '+' or '–' signs show the reaction stoichiometry. They are replaced by ν_{fi} and ν_{ri} for the forward and reverse reactions, respectively, which are elements of the so-called *stoichiometric matrix* (throughout this review, we follow Gillespie's notation [6]). In this example, one unit of each A and B are lost in a unit amount of forward reaction ($\nu_{fA} = \nu_{fB} = -1$) as one unit of C is formed ($\nu_{fC} = 1$); the ν_{ri} values have the opposite signs. With these substitutions, the rate equations become

$$\frac{dZ_i(t)}{dt} = \nu_{fi}\tilde{a}_f(\mathbf{Z}(t)) + \nu_{ri}\tilde{a}_r(\mathbf{Z}(t)) \,, \tag{3a}$$

$$\tilde{a}_f(\mathbf{Z}(t)) = k_f Z_A Z_B \,, \tag{3b}$$

$$\tilde{a}_r(\mathbf{Z}(t)) = k_r Z_C \,. \tag{3c}$$

These equations are trivially extended to arbitrarily large reaction networks. Consider a system with N chemical species that can react via M different *reaction channels* (a "reaction channel" is simply unambiguous terminology for a reaction between specific reactant and product chemical species). The dynamics of this system are given with the *reaction rate equation* (RRE):

$$\frac{dZ_i(t)}{dt} = \sum_{j=1}^{M} \nu_{ji}\tilde{a}_j(\mathbf{Z}(t)) \,. \tag{4}$$

The reaction propensity equations are typically the products of reaction rate constants and the appropriate chemical concentrations, as shown above (Eq. (3b) and (3c)),

but they may also describe non-elementary processes such as Michaelis–Menten kinetics. It is worth noting that the state of the system, at any point in time, is fully expressed with the vector $\mathbf{Z}(t)$. This means that the entire trajectory of the system can be deterministically calculated from the RRE and any single $\mathbf{Z}(t)$ snapshot.

The RRE is at the heart of many branches of quantitative biochemistry and systems biology. A great deal of metabolic theory is based on either the steady-state solutions of the RRE, or the set of steady-state solutions that are possible, given only knowledge of the stoichiometric matrix [26,27,28,29]. Studies of biochemical oscillations [25], including the cell cycle [30,31], circadian rhythms [32,33], and certain spatial patterns [34,35] are usually based on the dynamics of the deterministic RRE. Research on biochemical switches, as are found in prion diseases [36], developmental processes [37], and some protein kinase cascades [38], often focuses on the multiple steady-state solutions of the RRE [39]. Deterministic modeling has been, and still is, the conventional modeling method for most biological systems.

The Chemical Master Equation

Although the RRE is tremendously useful, it cannot address the stochastic processes that are inherent to biochemical systems. This is because the RRE arises from a series of approximations to a more physically rigorous stochastic model of chemical reactions [40,41].

As above, we assume that diffusive processes are much faster than reactive ones [8], which allows us to ignore spatial organization (this assumption is usually valid for genetic and metabolic networks, but often invalid for signaling networks). Nevertheless, the stochasticity of diffusion plays an essential role because it makes the precise timing of individual reactions effectively random. These reactions occur in abrupt transitions, in which reactants are converted effectively instantaneously into products, making this a type of *jump process*. Also, random diffusion causes the system to rapidly lose any memory of its prior states, and thus of the sequence of reactions that led up to the current state. This independence of the system dynamics on its history, called the Markov property, implies that the probability that a specific reaction occurs depends only on the state of the system at that time [41].

Because of the random reaction timing, reactant concentrations do not follow the deterministic trajectory that is predicted by the RRE. Instead, many concentration trajectories are possible, of which a single effectively random one actually occurs. There are two primary ways to investigate the possible trajectories with computational methods. One can simultaneously track the probability of every possible outcome or one can simulate many independent stochastic trajectories and then analyze them as one would with several repetitions of an experiment. These methods are described in this and subsequent sections, respectively.

To mathematically track the probability that the system is in each possible state, it is helpful to first replace the vector of chemical concentrations that was introduced earlier, $\mathbf{Z}(t)$, with a vector of integer-valued molecule numbers, $\mathbf{X}(t)$. These are related to each other, within round-off error, through the volume of the system, which is given as Ω,

$$\mathbf{X}(t) \simeq \Omega \mathbf{Z}(t) \,. \tag{5}$$

The state of the stochastic system is fully captured by $\mathbf{X}(t)$. The probability that the system is in state $\mathbf{x}$ at time t, given that it started in state $\mathbf{x}_0$ at time t_0, is written as $P(\mathbf{x}, t \mid \mathbf{x}_0, t_0)$. This probability changes over time because chemical reactions can transfer the system either into this state from other ones, or out of this state and into others. These possible transitions are combined to yield the *chemical master equation* (CME) [6,42]:

$$\frac{\partial P(\mathbf{x}, t \mid \mathbf{x}_0, t_0)}{\partial t} = \sum_{j=1}^{M} \Big[a_j(\mathbf{x} - \underline{\boldsymbol{\nu}}_j)\, P(\mathbf{x} - \boldsymbol{\nu}_j, t \mid \mathbf{x}_0, t_0) - a_j(\mathbf{x})\, P(\mathbf{x}, t \mid \mathbf{x}_0, t_0) \Big] \,. \tag{6}$$

The sum is over the reaction channels that can occur in the system. The two terms within brackets give the rate at which the probability of being in state $\mathbf{x}$ increases or decreases over time because of reactions into or out of state $\mathbf{x}$, respectively. These are proportional to the reaction propensities for the respective reactions. They are also proportional to the probability that the system was in the starting state, because the system can only leave a state if it was there in the first place.

The reaction propensity $a_j(\mathbf{x})$, given here without a tilde because it is for molecule numbers rather than concentrations, is a probability density: $a_j(\mathbf{x})\mathrm{d}t$ is the probability that exactly one reaction of type j will occur in a system in state $\mathbf{x}$ within the next $\mathrm{d}t$ amount of time. This microscopic propensity function is as central to stochastic chemical kinetics as its macroscopic analog is to the reaction rate equation. However, the microscopic propensity rests on a solid microphysical basis, and has in fact been shown to have an exact solution for a well-stirred thermally-equilibrated gas-phase system [43]. For such a system, the propensity function is

$$a_j = h_j c_j \,, \tag{7}$$

where h_j is the number of distinct combinations of individual reactants for reaction j and c_j is the probability density for one of those reactions to occur. That is, $c_j(t)\mathrm{d}t$ is the probability that a randomly selected set of reactants for reaction j will collide and react in the next infinitesimal time interval $\mathrm{d}t$. For a variety of reaction mechanisms, the c_j values can be calculated quite accurately from only the physical properties of the system [43].

The primary approximation made for the CME is that the reactive system is well-mixed. This implies both that there is no spatial organization and that there are no significant correlations between successive reactions (the Markov property). Examples of correlated reactions include metabolite channeling, which is the transfer of a metabolite from one enzyme to the next before it has a chance to equilibrate into the cytoplasm [44], and geminate recombinations, which are multiple bindings between molecules that bind reversibly [45]. The well-mixed statement also encompasses the assumption that the system is isothermal, which is typically the case in biological systems.

Because the CME becomes computationally intractable with any but the simplest systems, some recent work has focused on efficient solution methods. An algorithm called the finite state projection method accomplishes this by projecting a matrix form of the CME onto a smaller space [46,47]. By choosing the size of the projection space, the accuracy can be adapted to any level of precision. Less formal methods for state-space reduction of the CME have been proposed as well [48]. Another method uses a sparse grid, which can work efficiently for up to 10 proteins [49]. Work has also gone into separations of the CME into fast and slow components, as described in the section on hybrid methods [42,50,51].

Applications of the Chemical Master Equation

The solution of the CME suffers from the so-called "curse of dimensionality" as the size of the state space, and hence the number of equations, increases exponentially with the number of chemical species involved. Except for very small and simple systems, it is extremely difficult to obtain solutions of the CME, either analytically or numerically. However, a few papers do report quite interesting results from direct simulations of the CME.

The master equation was used to investigate the dynamics of transiently denatured segments of double stranded DNA [52]. The authors derived the dynamics, formation rates, and lifetimes of these "bubbles", which can be compared to fluorescence correlation microscopy experiments of fluorescently tagged base pairs [53]. In another use of the CME, studies on molecular motors [54,55] demonstrate how the load-velocity curve, including rectified motion, arises from nucleotide triphosphate binding free energies. These works more fully investigate ideas on thermal ratchets that were presented previously [56]. A particularly intriguing study on the copy number control system for bacterial plasmids [57] showed that stochasticity in a regulatory portion of a system can actually decrease the stochasticity elsewhere in the system. This "stochastic-focusing" changes the behaviors of gradual-response systems towards those of threshold systems [58,59] in a manner that is analogous to the oscillation enhancement that stochastic resonance can create in oscillating systems [60]. These results contradict the widely held belief that an increase in stochasticity in one portion of a system will necessarily increase the stochasticity everywhere downstream of it. Studies of simple signal transduction motifs have shown how the predictions of the RRE can be qualitatively wrong compared to the CME treatment in that the stochastic systems might be bistable or oscillate when the deterministic system has one stable state [40,61]. Many of these studies that used the CME also used other theoretical techniques as well, which allow fruitful comparisons between the methods.

The Gillespie Algorithm

Because of the challenges in working with the CME, it is most often investigated using a Monte Carlo approach in which individual sample trajectories are simulated. These simulations can be exact or approximate. In this context, "exact" means that if the simulation were run many times, the distribution of simulated trajectories would agree exactly with that which would be predicted by an analytical solution, were it obtainable, of the chemical master equation. Exactness implies nothing about the validity of the CME or about the limitations of the computational accuracy (such as round-off errors and imperfect pseudo-random number generators), but only that no further approximations are made beyond those that are assumed by the CME.

In 1976, Gillespie introduced an exact algorithm for simulating the CME [6,62,63] which is called the *stochastic simulation algorithm* (SSA) in his papers, but is better known as the *Gillespie algorithm*. This algorithm cycles through three portions: (i) generate the time step until the next reaction, (ii) determine which reaction that will be, and (iii) execute the reaction by advancing the time and molecule counts to reflect it. The Gillespie algorithm was introduced with two varieties, called the *direct method* and the *first-reaction method*. In the former, the time step to

the next reaction, τ, and the reaction number, j, are chosen from the following probability distributions:

$$P(\tau) = a\mathrm{e}^{-a\tau} , \tag{8a}$$

$$P(j) = \frac{a_j(\mathbf{X})}{a} . \tag{8b}$$

The variable a represents the summed reaction propensity,

$$a \equiv \sum_j a_j(\mathbf{X}) . \tag{8c}$$

In the first-reaction method, a time step, τ_j, is generated for each possible reaction channel. Again, these are exponentially distributed random numbers,

$$\tau_j = a_j(\mathbf{X})\,\mathrm{e}^{-a_j(\mathbf{X})\tau} . \tag{9}$$

The smallest of these time steps is chosen as the next simulation time step, while its subscript dictates the reaction channel that is executed at that time. The direct method is usually preferred because it is a little easier to program and runs slightly faster with a simple implementation. However, the latter has been favored as a basis for improvements on computational efficiency.

The exactness of the Gillespie algorithm comes at the cost of its being computationally demanding. Even if one simulation can be performed reasonably quickly, such as in a few minutes, this can still be too slow to investigate the behaviors of hundreds of mutant cells or to explore different regions of parameter space. Thus, much effort has been devoted to improving the efficiency of the Gillespie algorithm, while still maintaining exactness. These methods are all based upon either the direct method or the first-reaction method, but are carefully designed, usually with priority queues or other indexing methods, so that internal variables are recalculated as infrequently as possible [6,64,65,66,67]. Of these, it appears that the *optimized direct method* is probably best for most practical problems [66]. Because the faster methods are significantly more difficult to program than the original ones, both sets of methods are still commonly used.

The computational intensity of the Gillespie algorithm, even with more efficient implementations, makes it difficult to perform sensitivity analyzes. In these analyzes, one investigates the extent to which the results depend upon input parameters, which can helpful for determining which parameters need additional experimental investigation or which are particularly important for system control. An algorithm for stochastic sensitivity analysis was recently developed and applied to biochemical reaction networks [68]. It involves the addition of just two steps to the basic loop of the Gillespie algorithm.

Another difficulty of the Gillespie algorithm, which also applies to simulations of the RRE and other algorithms presented below, is called *combinatorial explosion*. Suppose a scaffold protein has several sites with which it can bind other proteins, and suppose that each of those proteins can bind to none, one, or two phosphate groups (this is the situation for the Ste5 protein in the yeast pheromone response pathway [69]). There are clearly a tremendous number of possible binding states that the scaffold protein can be in, each of which has to be treated as a separate chemical species. Just listing all of these states is tedious, and simulating their reaction dynamics with the Gillespie algorithm is very slow. One solution is to not list every possibility when the simulation starts, but to create states when they are needed and to destroy them when they are no longer required [65]. Another option is to use an algorithm that is implemented in a program called StochSim [70,71]. Unlike the Gillespie algorithm, this one does not stochastically choose reactions to execute, but it instead chooses reactant pairs from the pool of existing molecules. A probabilistic scheme is used to determine if these reactants should be made to react with each other.

What if the system volume changes as a function of time? This might seem like an unusual concern, but it occurs during cell growth (and cell division) and it affects the reaction propensities. The necessary modifications to the Gillespie algorithm were recently derived, which are likely to be particularly useful for relatively slow processes, such as protein production from infrequently expressed genes [72].

Applications of the Gillespie Algorithm

Models based on the Gillespie algorithm have provided critical insights into the stochastic nature of gene expression [3,73,74]. In particular, fluctuations in the rates of gene transcription are amplified at the translation stage to yield highly erratic time patterns of protein production [75]. When multiple regulatory proteins act together, or compete with each other, this randomness is amplified further because of the random sequence of protein bursts [75]. These effects were shown to stochastically switch a model of phage-λ between the bistable lysis and lysogeny states [76], with results that are consistent with experimental ones. Stochastic gene expression is also used by many pathogenic organisms to randomly switch their surface features so they can evade host responses [77], may be used by the HIV virus to stochastically delay viral expression long enough for transformation of its activated T-cell host to a memory cell and thereby trap HIV

as a latent phage [16], can establish asymmetries that determine cell differentiation [78], and can cause circadian clocks to lose synchrony [21,79]. In fluctuating environments, stochastic gene expression can permit an isogenic bacterial population to grow faster than it would if all individuals were phenotypically homogeneous [77,80,81].

From combined modeling and experimental approaches, the dominant noise sources in the stochastic expression of a specific gene are: (i) the expression fluctuations of that particular gene, which is called the *intrinsic noise*, (ii) noise that is transmitted to it from upstream genes, and (iii) global noise that affects all genes. The latter two sources are often combined and called *extrinsic noise*, which is the total noise source that is extrinsic to that specific gene [12,74,82,83,84]. Noise arises at both the transcription and translation stages, for which the relative importance depends on the strength of the promoter and on whether prokaryotic or eukaryotic transcriptional machinery is used [14,15,75,85,86]. Direct measurements of gene expression have generally confirmed the predictions made by stochastic simulations [11,12,13,15,17,18,21,87].

Gene expression appears to be unavoidably stochastic, and this randomness is usually amplified at each stage, so how does biology function reliably amidst all the noise? This is a central topic of many papers on biological robustness [5,9,88,89,90,91], several of which use the Gillespie algorithm or other types of stochastic modeling. One answer is that many reaction network structures are inherently less susceptible to noise than others. These include ones in which the reaction rates do not depend on the number of mRNA transcripts [92], certain scale-free reaction networks [93], and networks that are designed to function near saturation [94] (analogous to binary logic). Secondly, there are several mechanisms for biological robustness to noise, including negative feedback, integral feedback [95], checkpoints, and redundancy [9]. Because gene expression noise is usually detrimental to biological function, it has been suggested that there is active selection for robustness mechanisms [96,97].

Approximate Stochastic Methods

Because every reaction is simulated individually in the Gillespie algorithm, it is unavoidably computationally demanding, even with the algorithmic methods that have been developed to speed it up. To address this, several approximate methods have been developed.

The most accurate of these approximate methods is called the *tau-leaping* method [6,98,99,100]. In contrast to the Gillespie algorithm, the τ-leaping method uses a simulation time step which is long enough that many individual reactions are likely to occur during the time interval. The reaction propensities ought to change slightly as each reaction occurs to reflect the new chemical populations, although this algorithm uses the assumption that changes within a single time step are negligible. This is the sole approximation made for the τ-leaping method. Each reaction is considered to be an independent event (a consequence of both the well-mixed hypothesis and the constant reaction propensities), so the number of reactions that occur during time step τ for reaction channel j is a Poisson-distributed random variable; it is denoted k_j and has a mean value equal to the reaction propensity $a_j(\mathbf{X}(t))$. The formula used by τ-leaping that updates the system state over one time step is

$$\mathbf{X}(t+\tau) = \mathbf{X}(t) + \sum_{j=1}^{M} k_j \boldsymbol{\nu}_j \,. \tag{10}$$

The algorithm alternates steps in which the state of the system is updated and those in which new reaction propensities are calculated.

As the time step is reduced to zero, the τ-leaping simulation method approaches that of the Gillespie algorithm, although with more computational overhead. In the other direction, increasing the time step makes the simulation becomes more and more approximate. It has been suggested that τ should be chosen by first predicting the change in molecular populations over time using deterministic methods; then, the time step is chosen so that no molecular species is likely to change its population during this time step by more than some pre-determined fraction of its total population [98]. A difficulty that can occur with τ-leaping (which is also an issue with ODE integration), is that it is possible for a molecular species to be assigned a negative population. Several methods have been proposed to avoid this problem, some of which are also able to improve the performance of the algorithm in other ways as well [101,102,103]. Despite several papers on the development of τ-leaping, this method has yet to be applied to novel biological problems.

Two additional approximations allow the application of many more theoretical methods. First, the vector that defines the state of the system, $\mathbf{X}(t)$, is allowed to take on real values as well as integer values. As one would expect, this is usually a reasonable assumption for large chemical populations and a poor assumption for low copy numbers. In particular, it can a very poor approximation in cases where there might become no copies of a chemical species at all; an approximate value of, say, 0.01 protein copies can lead a system to entirely different outcomes than one would find with exactly 0 copies. The second

approximation is to replace the Poisson-distributed random variables that were used in the τ-leaping algorithm with Gaussian-distributed random variables [104]. The effect of this change again decreases as the copy numbers of chemical species are increased. It also decreases as longer time steps are used because that leads to more individual chemical reactions per time step. These approximations allow the updating equation of the τ-leaping algorithm to be replaced by a stochastic differential equation called the *chemical Langevin equation* [105,106,107] (CLE),

$$\frac{\mathrm{d}X_i(t)}{\mathrm{d}t} = \sum_{j=1}^{M} \nu_{ji}\, a_j(\mathbf{X}(t)) + \sum_{j=1}^{M} \Gamma_j(t)\, \nu_{ji} \sqrt{a_j(\mathbf{X}(t))}. \quad (11)$$

The first term is simply the reaction rate equation that is given above (Eq. (4)), but for molecule counts rather than concentrations. The second term adds Gaussian noise to the deterministic result, where $\Gamma_j(t)$ represents a temporally uncorrelated, statistically independent Gaussian white noise with mean 0 and variance 1. In other words, the integral of $\Gamma_j(t)$ is a one-dimensional continuous random walk. The chemical Langevin equation describes a continuous Markov process which is an approximation of the jump Markov process that underlies the chemical master equation. Simulations with the CLE, which are based on an equation that is quite similar to Eq. (11) [106], yield single stochastic trajectories of the system state, much like simulations with the Gillespie algorithm or the τ-leaping method.

Alternatively, instead of following a single system as it moves along one of its many possible stochastic trajectories, it is possible to focus on a single portion of the state space to see how likely it is that the system will be in this region of state space as a function of time. This latter picture is described by *chemical Fokker–Planck equation* [41,105,107] (CFPE),

$$\begin{aligned}\frac{\partial P(\mathbf{x}, t \mid \mathbf{x}_0, t_0)}{\partial t} = & \\ & -\sum_{i=1}^{N} \frac{\partial}{\partial x_i}\left[\left(\sum_{j=1}^{M} \nu_{ji} a_j(\mathbf{x})\right) P(\mathbf{x}, t \mid \mathbf{x}_0, t_0)\right] \\ & + \frac{1}{2}\sum_{i=1}^{N} \frac{\partial^2}{\partial x_i^2}\left[\left(\sum_{j=1}^{M} \nu_{ji}^2 a_j(\mathbf{x})\right) P(\mathbf{x}, t \mid \mathbf{x}_0, t_0)\right] \\ & + \sum_{\substack{i,i'=1 \\ i<i'}}^{N} \frac{\partial^2}{\partial x_i \partial x_{i'}}\left[\left(\sum_{j=1}^{M} \nu_{ji}\nu_{ji'} a_j(\mathbf{x})\right) P(\mathbf{x}, t \mid \mathbf{x}_0, t_0)\right].\end{aligned} \quad (12)$$

The first term, often called the drift term, arises from the deterministic behavior of the system. The latter two terms, collectively called the diffusion term, represent the stochastic deviations away from deterministic behavior. Because the CFPE represents continuous processes, it is significantly more analytically tractable than the chemical master equation.

To compute the probabilities of possible system behaviors using the CFPE, state space is usually discretized into a grid and then the CFPE is integrated using standard numerical methods [108,109]. This analysis method is similar to that employed for the CME, but is usually less computationally demanding because the discretized state space is typically significantly coarser. Nevertheless, the dimensionality of state space still increases exponentially with additional chemical species, so the CFPE still suffers from the curse of dimensionality.

Applications of Approximate Stochastic Methods

Perhaps because stochastic simulations are still a relatively new field of study, many studies with the CLE and CFPE focus more on the mathematical techniques than on the biological applications [109,110,111,112,113]. The approximate CLE and CFPE have been shown to yield results that are in good agreement with exact simulations for a reversible isomerization reaction, even with very few molecules [107].

The CLE was used to analytically investigate the role of noise-induced phenomena in enzymatic futile cycles, which is a motif that is common to many biochemical networks [61]. The analysis indicated that the presence of external noise is sufficient to induce switching bistability in the system, a phenomenon that is often attributed to feedback loops [25]. In combination with experimental data, the CLE was also used to show that translational efficiency is the predominant source of intracellular noise for a single-gene system [15]. The Fokker–Planck equation has been used to model cell growth [111,112] and cell migration [113,114]. Of particular interest, the Fokker–Planck equation has provided a convenient framework to describe the behaviors of molecular motors [109]. A motor protein is approximated as a diffusion particle in a periodic asymmetric free-energy surface. Under the input of chemical energy, the motor switches stochastically between different potentials that describe distinct biochemical states of the motor. The model has been used to explain key experimental observations for molecular motors, most notably for the F_1F_0-ATPase system [115] and a bacterial flagellar motor [109,116].

Hybrid Algorithms

Systems that involve multiple time scales provide major simulation challenges. If the fast time scale is simulated with high precision, then the simulation takes too long for the dynamics of the slow one to be observed with any reasonable efficiency. On the other hand, if the simulation time steps are optimized for the slow timescale, then they are too long for the fast reactions and numerical errors become problematic. In the language of differential equations, these are stiff systems which require special solution techniques. For stochastic simulations with multiple timescales, several methods have been developed recently.

One class of hybrid methods focuses on new mathematics to allow approximations of the Gillespie algorithm, or related algorithms, to function with reasonable accuracy over a wide range of timescales [42,50,51,117,118,119,120,121]. The other class generally involves the coupling of multiple simulators, usually including ODE, Langevin, and/or Gillespie; the high-population molecular species are simulated with less stochastic detail and the low-population species are simulated with more stochastic detail [122,123].

Spatial Stochastic Modeling

Most biological systems are highly organized. For example, *Escherichia coli* bacteria have helical cytoskeletons, polar-localized proteins, centrally positioned chromosomes, and elaborate flagellar motor complexes. Eukaryotes are even more organized, with elaborate organelles, microtubules and other complex cytoskeletal elements, motor proteins that shuttle back and forth, and carefully controlled cell shapes. Even phages display remarkable order in the way the DNA is packed into the outer shell. Where does this order come from? And how does this order influence the biochemical reaction network? These questions are being investigated with new imaging experiments [124,212] and with new computer simulation methods that can account for spatial heterogeneity. These spatial simulation methods are the focus of this section.

Spatial simulations have been used to investigate a wide variety of topics. These include: morphogen gradients across *Drosophila* and *Xenopus* oocytes [125,126,127], the *Escherichia coli* cell division plane localization system [128,129,130,131,132,133,134] (see Sect. "Box 1: The *E. Coli* Min System"), intracellular signaling [135,136,137,138], and rebinding of ligands to receptor complexes [139,140,141].

As described above, many successful biochemical models do not account for spatial heterogeneity; in fact, non-spatial models are in the vast majority. Typically, non-spatial models get away with ignoring space because they model dynamics that occur more slowly than the time it takes for a molecule to diffuse across a cell, because they investigate processes that are not intrinsically spatial, and because they do not demand high quantitative accuracy. As the tools are becoming available, including both fast computers and new software algorithms [45,142,143,144], the interest in including spatial detail is increasing. These spatial models can be either deterministic or stochastic, of which our primary focus is on the latter ones.

As with the non-spatial methods that are described above, stochastic effects in spatial models arise from the discreteness of molecules. This leads to fluctuations in the numbers of molecules, which are typically on the order of the square root of the number of molecules in the appropriate characteristic volume (near steady-state and equilibrium points, but frequently greater near critical points). In spatial models, the characteristic length scale is no longer the size of the entire system but is dictated by the length scale of the spatial heterogeneity. With the shorter length scale, the characteristic volume size is reduced, fewer molecules are in these volumes, and stochastic effects increase. Thus, stochastic simulations can be required for spatial models, even if they were not needed for the corresponding non-spatial model. There are also other good reasons to model stochastic effects in spatial simulations. Many spatial phenomena, such as noise that arises from ligand rebinding [141], cannot be adequately treated without considering the detailed molecular interactions. Finally, a model is only as good as its weakest aspect. If one increases the accuracy of a model in one way, such as by accurately treating either space or stochastics, then the benefits may not be realized until the other aspect is addressed as well.

Schemes for investigating a chemical system with spatial and stochastic detail can be classified by whether they consider molecules within populations or as individuals. In the former case, space is divided it into small subvolumes, whereas in the latter, space is continuous. These classes are described in detail below. Another approach is lattice-based methods [145,146,147,148,149]. However, we do not discuss them here because they are rarely used for quantitative modeling. Furthermore, the underlying lattice geometry usually affects the results, thus making them less realistic.

Population-Based Spatial Models

In a top-down approach towards spatial modeling, one starts with a simple, deterministic, macroscopic description and then adds successive layers of detail. In this case,

the natural starting point is with the standard textbook descriptions of chemical reactions and diffusion [24]. Reactions are described with mass action reaction kinetics expressed with the reaction rate equation that was discussed above (Eq. (4)). Diffusion is described with the diffusion equation [24], also called Fick's second law of diffusion, which is

$$\frac{\partial Z_i(\mathbf{r},t)}{\partial t} = D_i \nabla^2 Z_i(\mathbf{r},t)\,. \tag{13}$$

In an extension of the definition given before, $Z_i(\mathbf{r},t)$ is the concentration of component i at the 3-dimensional position $\mathbf{r}$ and time t. D_i is the diffusion coefficient for component i.

Because reactions and diffusion occur simultaneously, the respective equations are combined to express the simultaneous effects of both processes to yield the *reaction-diffusion equation*,

$$\frac{\partial Z_i(\mathbf{r},t)}{\partial t} = \sum_{j=1}^{M} \nu_{ji}\tilde{a}_j(\mathbf{Z}(\mathbf{r},t)) + D_i \nabla^2 Z_i(\mathbf{r},t)\,. \tag{14}$$

This partial differential equation (PDE) underlies a great deal of theory on chemical and biological pattern formation [126,136,150,151]. The Virtual Cell computer program [152] is general-purpose software that simulates the reaction-diffusion equation. It has been used primarily to explore spatial effects in intracellular signaling [153,154,155].

The reaction-diffusion equation is deterministic, so it captures neither the discreteness of reaction events nor the Brownian motion processes that underlie diffusion. It is possible to add this stochasticity directly into the deterministic theory but that would create a set of coupled stochastic scalar field equations, which would be extraordinarily complicated. Neither the deterministic nor the stochastic PDEs are tractable to work with analytically for any but the very simplest systems except, perhaps, in steady-state. Thus, most analysis is either computational or approximate.

In most such analyses, the equations are first simplified by dividing the system volume into an array of small cubic subvolumes, each with width l. This spatial discretization changes the diffusion portion of the reaction-diffusion equation into a discrete form:

$$\frac{\mathrm{d}Z_{i,k}(t)}{\mathrm{d}t} = \sum_{j=1}^{M} \nu_{ji}\,\tilde{a}_j(\mathbf{Z}_k(t)) + \frac{D_i}{l^2}\sum_{k'}\left[Z_{i,k'}(t) - Z_{i,k}(t)\right]\,. \tag{15}$$

The index k denotes the subvolume number, much as $\mathbf{r}$ represented the spatial location. The latter summation in this discrete reaction-diffusion equation extends over all nearest neighbors of subvolume k, denoted by k'. Because the description of space was changed from continuous states to discrete states, much like the discrete kinds of molecules that are labeled by the index i, diffusion is now formally identical to reactions. The "reaction rate constant" for diffusion [156] between one subvolume and its neighbor is D_i/l^2. Because of this mathematical equivalence, much of the following discussion on the stochastic simulation of the reaction-diffusion equation parallels the discussion presented earlier on non-spatial stochastic simulations.

The first spatial stochastic equation that we present is the one that accounts for the least detail. It is the *spatial chemical Langevin equation*, which results from adding white Gaussian noise to the discrete reaction-diffusion equation. It is

$$\begin{aligned}\frac{\mathrm{d}Z_{i,k}(t)}{\mathrm{d}t} = &\sum_{j=1}^{M} \nu_{ji}\left[\tilde{a}_j(\mathbf{Z}_k(t)) + \Gamma_j(t)\sqrt{\tilde{a}_j(\mathbf{Z}_k(t))}\right] \\ &+ \sum_{k'}\left\{\frac{D_i}{l^2}\left[Z_{i,k'}(t) - Z_{i,k}(t)\right]\right. \\ &\left. + \Gamma_{k'}(t)\sqrt{\frac{D_i}{l^2}}\left[\sqrt{Z_{i,k'}(t)} - \sqrt{Z_{i,k}(t)}\right]\right\}\,.\end{aligned} \tag{16}$$

In an extension to what was presented before, $\Gamma_j(t)$ and $\Gamma_{k'}(t)$ represent temporally uncorrelated, statistically independent Gaussian white noises [106]. This is a specific example of the more general multivariate Langevin equation; it, and the multivariate Fokker–Planck equation, have been explored in depth [41,105]. However, the more specific spatial chemical Langevin equation has essentially never been used, investigated mathematically, or simulated. The sole exception that we are aware of was its simulation for a figure for a tutorial article [8] (those results are reproduced in Fig. 1).

The spatial chemical Langevin equation captures stochasticity reasonably accurately for systems in which there are many molecules per subvolume but not for those with few molecules per subvolume. Errors arise both because Gaussian white noise is the incorrect fluctuation distribution [100,104] and because it treats molecule amounts as continuously variable quantities. These are addressed by moving to the next level of detail in which the continuous molecular concentrations are replaced by discrete numbers of molecules. This changes the temporally

continuous reaction and diffusion processes to stochastic jump processes. The *reaction-diffusion master equation* [156,157,158,159] (RDME) describes the time dependence of the system at this level of description. It is

$$\frac{\mathrm{d}P(\mathbf{x},t)}{\mathrm{d}t} = \sum_{j=1}^{M}\left[a_j(\mathbf{x}-\underline{\boldsymbol{\nu}}_j)P(\mathbf{x}-\boldsymbol{\nu}_j,t) - a_j(\mathbf{x})P(\mathbf{x},t)\right] + \sum_{i=1}^{N}\frac{D_i}{l^2}\sum_{k,k'}\Big[(X_{i,k}+1)\,P(\ldots,X_{i,k}+1, X_{i,k'}-1,\ldots,t) - X_{i,k}P(\mathbf{x},t)\Big]. \tag{17}$$

$P(\mathbf{x},t)$ is the probability that the system is in state $\mathbf{X}=\mathbf{x}$ at time t, $X_{i,k}$ is the number of molecules of type i in subvolume k, $\mathbf{X}$ is the vector of all $X_{i,k}$ values, and a_j is the propensity of reaction j.

The RDME expresses as much detail as is possible through these successive improvements of the reaction-diffusion equation. It is tempting to think of it as the fundamental equation for reactions and diffusion, and thus the basis for a statistical theory of chemistry. In fact, it is sufficiently accurate for most systems, but it nevertheless involves approximations that can be important in some situations. Firstly, neither of the starting equations, which are mass action kinetics and Fickian diffusion, are completely accurate even for very large systems. Mass action kinetics does not address the increased reaction rates that occur on extremely short time scales, which arise from reduced spatial correlations [160,161]. Nor does it address the geminate recombinations that can occur between the products of a dissociation reaction [162,163]. The diffusion equation is usually quite accurate for dilute solutions but fails for highly crowded ones [164,165], including most biological systems [166]. Secondly, the discretization of space into small subvolumes can also lead to inaccuracies, or exacerbate the inaccuracies just mentioned. The subvolume sizes must not be so small that they impinge on the microscopic details of the reaction or diffusion processes. This means that they need to be significantly larger than single molecules and larger than the mean free path lengths of diffusion [156,167]. Conversely, the subvolumes must not be so large that there would be appreciable concentration gradients across them. This means that the subvolume width needs to be less than the reactant correlation length. The correlation length is hard to predict but is at least as large as the average distance that a reactant travels before it reacts, called the reactive mean free path [156,167].

The RDME is even more intractable than the nonspatial chemical master equation because of the addition of spatial states and the many transitions that can occur between the spatial states. These additional states and transitions also make stochastic simulations of the RDME with Gillespie's direct algorithm extremely slow [157]. Several faster algorithms have been developed to address this problem. The "next reaction method" of Gibson and Bruck [64] was adapted to spatial simulations [168], and then further improved, to yield the "next subvolume method" [132,142]. Also, a fast version of the direct method [169] has been developed for spatial simulations [167]. All of these methods yield exactly the same results as Gillespie's original methods [62,63] but use carefully optimized data structures to minimize the number of computations.

A separate challenge with simulating the RDME concerns the cubical subvolumes into which space was discretized. Biological systems rarely have square corners, so the basic theory requires adaptation to account for realistic boundaries. In one approach, the mathematics was developed for dividing boundary subvolumes into two separate portions [170]. Using another approach, the theory was developed for curved surfaces, which was implemented in the MesoRD program [132,171]. Although it has not been developed yet, it has been proposed that automatic mesh refinement could simultaneously account for complex boundaries and lead to significant computational efficiencies [167].

Along with simulations of the *E. coli* Min system, presented in Sect. "Box 1: The E. Coli Min System", population-based spatial stochastic models have been used for a variety of test systems. In the first implementation of a spatial Gillespie algorithm, Stundzia and Lumsden used a one-dimensional simulation to demonstrate stochastic calcium wave propagation [157]. Elf and Ehrenberg showed that spatial and stochastic effects can cause an intrinsically bistable system to lose its global hysteresis through the formation of spatial domains [142]. In a third study, Isaacson and Peskin demonstrated their method for simulating porous boundaries with a model that includes transcription, translation, and nuclear membrane transport [170].

Individual-Based Spatial Models

In a bottom-up approach to spatial modeling, one starts with a very detailed consideration and then makes successive approximations. A convenient place to start is by considering every individual molecule in the system, along with some of the molecular structures. The motions of

these molecules are governed by physical forces including steric repulsion, bond mechanics, and electrostatics. The simulation of the motions that result from these forces is called molecular dynamics [172]. Molecular dynamics can yield very accurate results but is so computationally intensive that it is rarely used for more than hundreds of cubic nanometers of volume or more than tens of nanoseconds of time. These size and time scales are too confining for studying biochemical reaction networks, so approximations are made.

At the Smoluchowski level of detail, all solvent molecules are ignored, solute molecules are treated as spheres, diffusion proceeds stochastically, and molecular rotation, molecular momentum, and long-range intermolecular forces are all ignored. This is a vast simplification, but is often valid. It is usually reasonably accurate for size scales that are larger than a few nanometers and for timescales that are longer than a few nanoseconds, constraints that are acceptable for an enormous range of chemical and biological phenomena.

For diffusion at the Smoluchowski level of detail, the effects of solvent-solute interactions on the solute motion are approximated by assuming that solute molecules diffuse with mathematically ideal Brownian motion [173,174]. This is a key approximation that replaces the deterministic molecular motions that result from solvent collisions with stochastic trajectories. It is often the only source of stochasticity in the theory, or in simulations that derive from this individual-based approach. More precisely, the position of molecule i at time t is given with the probability density $p_i(\mathbf{r}, t)$, which evolves over time according to the master equation

$$\frac{\partial p_i(\mathbf{r}, t)}{\partial t} = D_i \nabla^2 p_i(\mathbf{r}, t) \,. \tag{18}$$

This equation is nearly identical to the diffusion equation, given above (Eq. (13)), differing only in the definitions of the variables and the interpretation. Now, it is not a population of molecules that diffuse, but the positional probability density for a single molecule.

Because it is so simple, the diffusion master equation is analytically tractable, in contrast to the other master equations that were discussed. One result is an entire body of analytical theory on diffusion-influenced reactions [160,175]. Nevertheless, it too becomes unmanageable for systems that have several interacting molecules, so it is simulated with a technique called *Brownian dynamics* [176,177,178,179,180]. In this method, molecules have well-defined point-like positions which are updated at each simulation time step using random displacements. The displacements are chosen by solving the diffusion master equation for molecule i, which is taken to be at the well-defined position $\mathbf{r}_0$ at time t_0. One simulation time step later, at time $t_0 + \Delta t$, the probability density for the molecule's position is found to be a 3-dimensional Gaussian density that is centered at $\mathbf{r}_0$,

$$p_i(r, \Delta t) = \frac{1}{s_i^3 (2\pi)^{3/2}} \exp\left[-\frac{(\mathbf{r} - \mathbf{r}_0)^2}{2 s_i^2} \right] . \tag{19a}$$

The standard deviation of this Gaussian, called the root mean square step length, is

$$s_i = \sqrt{2 D_i \Delta t} \,. \tag{19b}$$

Brownian dynamics simulations provide accuracy that is below that of molecular dynamics, but still captures single molecule behavior.

Brownian dynamics has been used extensively for examining the rates of diffusion-influenced chemical reactions in solution [139,177,179,180,181,182] and for the rates of binding between ligands and receptor arrays [139,141,178,183,184]. In these studies, simulated molecules diffuse in solution; at the moment that a reactant pair, or a ligand and its cognate receptor, come into contact, they undergo a chemical reaction. While diffusing, intermolecular forces are often ignored, although some studies account for these interactions as well [185,186].

To achieve the necessary level of detail, Brownian dynamics simulations usually use very short simulation time steps, often on the order of picoseconds [139]. Adaptive time steps, such that time steps are long when reactants are widely separated and short when they are close, can speed simulations up by several orders of magnitude, but are easy to implement only if there is just one diffusing particle present in the simulation volume [141]. A more sophisticated method that has the same general goal of computational efficiency is called Green's function reaction dynamics [143,187,188] (GFRD). In GFRD, which works with any number of molecules, the system is inspected to see how soon the next molecular collision or reaction could occur. The system is then advanced to that time using a single simulation time step, the event is executed if appropriate, and the cycle repeats. Yet another method, used in a program written by one of us (SSA) called Smoldyn, achieves computational efficiency by modifying the effective radii of simulated molecules so that the same reaction rate is achieved with long simulation time steps as with short ones [45,189,190]. This method does not achieve the same spatial or temporal precision as classical Brownian dynamics or GFRD, but the level of detail is still more than adequate for most biological applications and has been

shown to be indistinguishable from more accurate simulations in many cases [191].

Technically, all of these algorithms execute Brownian dynamics. However, the term "Brownian dynamics" is typically used to describe highly detailed studies in which reaction rates, rebinding dynamics, or similar phenomena are found from fundamental molecular properties such as molecular radii and intermolecular forces. In contrast, GFRD and the methods used in Smoldyn are more often used to determine system-level behaviors from known or estimated reaction rates. These are more often called particle-based stochastic simulation methods [192].

MCell is another program that performs particle-based stochastic simulations [144]. Unlike the others, it cannot simulate reactions that occur in free solution, but instead only treats reactions at surfaces. Despite the decrease of versatility, it is still useful for studying a wide variety of biological phenomena [193,194,195]; in particular, it was developed to investigate the neuromuscular junction [196,197,198]. In MCell, surface-bound receptors are not modeled as single molecules as they would be in Smoldyn or GFRD methods, but as a uniform binding probability that applies to an entire surface tile. This decreases the spatial resolution some, but increases the computational efficiency.

Future Directions

The classic advice of using the right tool for the job is as true in biochemical modeling as it is elsewhere. Several modeling tools have been presented here. Deterministic ordinary differential equation models are simple, easy to use, and can be analyzed with many powerful theoretical and analytical methods. They are the right tool for systems that can be treated as being well-mixed and that are both large enough and sufficiently far from critical points that stochastic effects are unimportant. In contrast, systems that include low copy numbers of important components, and/or that can be triggered by random events, require stochastic modeling methods for their investigation. These include integration of the chemical master equation and random sampling of the stochastic trajectories using the Gillespie algorithm, both of which are exact methods. Approximate methods include τ-leaping stochastic simulations, integration of the chemical Fokker–Planck equation, and sampling with the chemical Langevin equation. Of these, the Gillespie algorithm has proven to be the most popular. Finally, if the system cannot be considered to be well-mixed, then yet different tools are needed. These include spatial variants of the same list of simulation methods, including partial differential equations for deterministic simulations and a spatial Gillespie algorithm for stochastic simulations. Particle-tracking simulation methods allow an even greater degree of detail.

In general, more detailed simulation methods yield more accurate results and are based more closely on underlying processes and less on phenomenological descriptions. However, they are also more computationally intensive and require more model parameters. This parametrization poses a significant problem for current models because the necessary quantitative experimental data are typically only marginally adequate or are completely non-existent. For an ODE model, it is sometimes possible to address this problem by exploring model behaviors over wide ranges of parameter space, from which one can draw phase diagrams that graphically depict how the model behaves for different parameter choices. From this, one can sometimes constrain parameters or gain additional insight into the model; for example, Tyson showed how two enzyme concentrations can be used to regulate the cell cycle, bringing an oocyte from metaphase arrest to autonomous oscillations, and on to growth-controlled cell division [30]. Because of the computational demands of spatial and stochastic models, as well as the richer behavior possibilities, it is much more difficult to explore parameter space with these more complex models. Thus, much work is needed on this topic.

More generally, the mathematical infrastructure for designing and interpreting stochastic models lags far behind that for non-spatial deterministic models. This poses some challenges for theorists. For example, what new theories and graphical tools will help scientists gain intuition into the dynamics of stochastic systems? and what are the controlling elements of stochastic systems? The theory is even more unexplored when spatial organization is considered as well. Nevertheless, spatial considerations are essential because no biological life has been found that is well-mixed; instead, a tremendous amount of biochemical activity involves membranes, polymers, protein scaffolds, large multimeric complexes, and other spatial structures. Theories that address these topics will not be as elegant as those that focus on the chemical master equation, but biology is not always elegant either.

Although research on stochastic modeling of biochemistry grew slowly from the 1950s to the 1990s, the pace has accelerated dramatically during the last 10 to 15 years. This acceleration will likely continue for many more years, in response to the faster computers that become available every year and to the ever-increasing complexity of biochemical data. With this growth, stochastic modeling may open up entire new ways to understand cell biology.

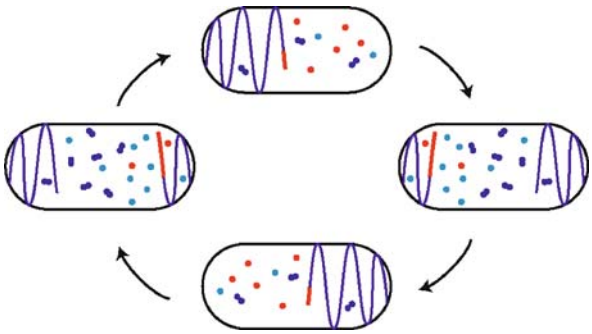

Stochastic Models of Biological Processes, Figure 2
Diagram of the *E. coli* Min system, which is used to position the cell division plane at the cell center. *Dots* represent cytoplasmic proteins, while *curved lines* represent helical membrane-bound protein polymers. *Colors* identify the proteins: *light blue* for MinD bound to ADP, *dark blue* for MinD protein bound to ATP, and *red* for MinE. The system dynamics are summarized in the text of Sect. "Box 1: The E. Coli Min System"

Box 1: The *E. Coli* Min System

The *E. coli* Min system has served as a proving ground for spatial stochastic simulation methods. The Min system is used by *E. coli*, in conjunction with other systems, to position the cell division plane accurately at the cell center [131]. The system is comprised of the proteins MinC, MinD and MinE, which oscillate back and forth across the cell, from one pole to the other, with a period of about 40s (Fig. 2). Of these, only MinD and MinE are required for the oscillation, making this a relatively simple system that exhibits remarkably interesting dynamics. Cytoplasmic MinD proteins bind ATP, dimerize, and polymerize on the inside of the cell membrane to form long helical structures that extend outwards from one of the two cell poles. When MinE binds to the cell-center end of a MinD polymer, it activates ATP hydrolysis which depolymerizes the terminal subunit. As MinE progressively disassembles a MinD polymer at one end of the cell, it reassembles again from the opposite pole to start the next oscillation cycle. The oscillating Min proteins continually inhibit cell division plane formation near the poles using MinC, which colocalizes with MinD, thus only permitting cell division at the cell center.

This system was explored for several years with deterministic reaction-diffusion models [128,199,200]. One of these models, by Howard, Rutenberg, and de Vet [199], was also explored by the same group using a one-dimensional population-based stochastic method [129] (it uses discrete particle numbers and fixed time steps, thus conceptually placing it between the spatial Langevin and spatial Gillespie methods). The authors found that stochastic effects were essential for generating oscillations in some parameter regimes, in a spatial version of stochastic resonance [60,201]. These models helped direct new experiments [202,203,204,205] that clarified the processes of the system.

Building on the prior models and the new experimental data, Huang, Meir, and Wingreen [130] developed a new reaction-diffusion model that was more closely connected with the biology than were previous models and that accounted for several mutant phenotypes. This model became the basis of several stochastic simulations. The spatial Gillespie method was employed by Fange and Elf [132] using their MesoRD program. They showed that a stochastic model can account for a "spotty" phenotype and for oscillations in spherical mutant cells, neither of which can be explained by the deterministic model. The MCell particle-tracking program was used by Kerr and coworkers [133] to show that the Min system alone is insufficient to center the cell division plane with high accuracy.

Yet unexplained with these simulations were convincing experimental results that MinD forms polymers on the cell membrane [205,206,207]. These were explored with another particle-tracking model [208], using a method based on Smoluchowski dynamics [45]. Although this group did simulate spontaneous polymer formation, they observed many randomly oriented short filaments, in contrast to the few helical polymers that are observed experimentally. This inherent difficulty with the reaction-diffusion model [209], whether deterministic or stochastic, has led to several studies that have focused specifically on the polymer dynamics and shapes [134,210,211].

The *E. coli* Min system is already well on its way to becoming the prototypical system for studying spatial biochemical dynamics, much as *E. coli* chemotaxis has become the prototypical system for investigating bacterial signaling.

Acknowledgments

This work was funded by the US Department of Energy.

Bibliography

1. McQuarrie DA (1967) Stochastic approach to chemical kinetics. J Appl Probab 4:413–478
2. Turner TE, Schnell S, Burrage K (2004) Stochastic approaches for modelling in vivo reactions. Comp Biol Chem 28:165–178
3. Raser JM, O'Shea EK (2005) Noise in gene expression: Origins, consequences, and control. Science 309:2010–2013
4. Samoilov MS, Price G, Arkin AP (2006) From fluctuations to phenotypes: The physiology of noise. Sci STKE 2006:re17

5. Rao CV, Wolf DM, Arkin AP (2002) Control, exploitation, and tolerance of intracellular noise. Nature 420:231–237
6. Gillespie DT (2007) Stochastic simulation of chemical kinetics. Ann Rev Phys Chem 58:35–55
7. Wolf DM, Arkin AP (2003) Motifs, modules and games in bacteria. Curr Opin Microbiol 6:125–134
8. Andrews SS, Arkin AP (2006) Simulating cell biology. Curr Biol 16:R523–R527
9. McAdams HH, Arkin A (1999) It's a noisy business! Genetic regulation at the nanomolar scale. Trends Genet 15:65–69
10. Singer RH, Lawrence DS, Ovryn B, Condeelis J (2005) Imaging of gene expression in living cells and tissues. Biomed J Optics 10:051406
11. Levsky JM, Shenoy SM, Pezo RC, Singer RH (2002) Single-cell gene expression profiling. Science 297:836–840
12. Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. Science 297:1183–1186
13. Raser JM, O'Shea EK (2004) Control of stochasticity in eukaryotic gene expression. Science 304:1811–1814
14. Blake WJ, Kaern M, Cantor CR, Collins JJ (2003) Noise in eukaryotic gene expression. Nature 422:633–637
15. Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A (2002) Regulation of noise in the expression of a single gene. Nature Genet 31:69–73
16. Weinberger LS, Burnett JC, Toettcher JE, Arkin AP, Schaffer DV (2005) Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. Cell 122:169–182
17. Cai L, Friedman N, Xie XS (2006) Stochastic protein expression in individual cells at the single molecule level. Nature 440:358–362
18. Yu J, Xiao J, Ren X, Lao K, Xie XS (2006) Probing gene expression in live cells, one protein molecule at a time. Science 311:1600–1603
19. Golding I, Cox EC (2006) Protein synthesis molecule by molecule. Genome Biol 7:212
20. Fusco D, Accornero N, Lavoie B, Shenoy SM, Blanchard J-M, Singer RH, Bertrand E (2003) Single mRNA molecules demonstrate probabilistic movement in living mammalian cells. Curr Biol 13:161–167
21. Elowitz MB, Leibler S (2000) A synthetic oscillatory network of transcriptional regulators. Nature 403:335–338
22. Sakurai JJ (1994) Modern Quantum Mechanics. Addison-Wesley, Boston
23. Strogatz SH (1994) Nonlinear Dynamics and Chaos. Westview Press, Cambridge
24. Atkins PW (1986) Physical Chemistry. Freeman, New York
25. Tyson JJ, Chen KC, Novak B (2003) Sniffers, buzzers, toggles, and blinkers: dynamics of regulatory and signaling pathways in the cell. Curr Opin Cell Biol 15:221–231
26. Covert MW, Schilling CH, Famili I, Edwards JS, Goryanin II, Selkov E, Palsson BO (2001) Metabolic modeling of microbial strains in silico. Trends Biochem Sci 26:179–186
27. Varma A, Palsson BO (1994) Metabolic flux balancing: Basic concepts, scientific and practical use. Nature Biotech 12:994–998
28. Kauffman KJ, Prakash P, Edwards JS (2003) Advances in flux balance analysis. Curr Opin Biotech 14:491–496
29. Fell D (1997) Understanding the Control of Metabolism. Portland Press, London
30. Tyson JJ (1991) Modeling the cell division cycle: cdc2 and cyclin interactions. Proc Natl Acad Sci USA 88:7328–7332
31. Chen KC, Calzone L, Csikasz-Nagy A, Cross FR, Novak B, Tyson JJ (2004) Integrative analysis of cell cycle control in budding yeast. Mol Biol Cell 15:3841–3862
32. Goldbeter A (2002) Computational approaches to cellular rhythms. Nature 420:238–245
33. van Zon JS, Lubensky DK, Altena PRH, ten Wolde PR (2007) An allosteric model of circadian KaiC phosphorylation. Proc Natl Acad Sci USA 104:7420–7425
34. Reinitz J, Mjolsness E, Sharp DH (1995) Model for cooperative control of positional information in Drosophila by bicoid and maternal hunchback. Exp J Zool 271:47–56
35. von Dassow G, Meir E, Munro EM, Odell GM (2000) The segment polarity network is a robust developmental module. Nature 406:188–192
36. Kellershohn N, Laurent M (2001) Prion diseases: dynamics of the infection and properties of the bistable transition. Biophys J 81:2517–2529
37. Ferrell JEJ, Machleder EM (1998) The biochemical basis of an all-or-none cell fate switch in Xenopus oocytes. Science 280:895–898
38. Huang C-YF, Ferrell JEJ (1996) Ultrasensitivity in the mitogen-activated protein kinase cascade. Proc Natl Acad Sci USA 93:10078–10083
39. Laurent M, Kellershohn N (1999) Multistability: a major means of differentiation and evolution in biochemical systems. Trends Biochem Sci 24:418–422
40. Samoilov MS, Arkin AP (2006) Deviant effects in molecular reaction pathways. Nature Biotech 24:1235–1240
41. van Kampen NG (1992) Stochastic Processes in Physics and Chemistry. Elsevier, Amsterdam
42. Haseltine EL, Rawlings JB (2005) On the origins of approximations for stochastic chemical kinetics. Chem J Phys 123:164115
43. Gillespie DT (1992) A rigorous derivation of the chemical master equation. Physica A 188:404–425
44. Rohwer JM, Postma PW, Kholodenko BN, Westerhoff HV (1998) Implications of macromolecular crowding for signal transduction and metabolite channeling. Proc Natl Acad Sci USA 95:10547–10552
45. Andrews SS, Bray D (2004) Stochastic simulation of chemical reactions with spatial resolution and single molecule detail. Phys Biol 1:137–151
46. Munsky B, Khammash M (2006) The finite state projection algorithm for the solution of the chemical master equation. Chem J Phys 124:044104
47. Peles S, Munsky B, Khammash M (2006) Reduction and solution of the chemical master equation using time scale separation and finite state projection. Chem J Phys 125:204104
48. Kuwahara H, Myers CJ, Samoilov MS, Barker NA, Arkin AP (2006) Automated abstraction methodology for genetic regulatory networks. Trans Comput Syst Biol 6:150–175
49. Hegland M, Burden C, Santoso L, MacNamara S, Booth H (2007) A solver for the stochastic master equation applied to gene regulatory networks. Comp J Appl Math 205:708–724
50. Nedea SV, Jansen APJ, Lukkien JJ, Hilbers PAJ (2003) Infinitely fast diffusion in single-file systems. Phys Rev E 67:046707
51. Chatterjee A, Vlachos DG (2006) Multiscale spatial Monte Carlo simulations: Multigriding, computational singular per-

turbation, and hierarchical stochastic closures. Chem J Phys 124:064110
52. Ambjörnsson T, Banik SK, Lomholt MA, Metzler R (2007) Master equation approach to DNA breathing in heteropolymer DNA. Phys Rev E 75:021908
53. Altan-Bonnet G, Libchaber A, Krichevsky O (2003) Bubble dynamics in double-stranded DNA. Phys Rev Lett 90:138101
54. Lattanzi G, Maritan A (2001) Master equation approach to molecular motors. Phys Rev E 64:061905
55. Wang H-Y, Elston T, Mogilner A, Oster G (1998) Force generation in RNA polymerase. Biophys J 74:1186–1202
56. Peskin CS, Odell GM, Oster GF (1993) Cellular motions and thermal fluctuations: the Brownian ratchet. Biophys J 65:316–324
57. Paulsson J, Ehrenberg M (2000) Random signal fluctuations can reduce random fluctuations in regulated component of chemical regulatory networks. Phys Rev Lett 84:5447–5450
58. Paulsson J, Berg OG, Ehrenberg M (2000) Stochastic focusing: fluctuation-enhanced sensitivity of intracellular regulation. Proc Natl Acad Sci USA 97:7148–7153
59. Berg OG, Paulsson J, Ehrenberg M (2000) Fluctuations in repressor control: thermodynamic constraints on stochastic focusing. Biophys J 79:2944–2953
60. Li H, Hou Z, Xin H (2005) Internal noise stochastic resonance for intracellular calcium oscillations in a cell system. Phys Rev E 71:061916
61. Samoilov M, Plyasunov S, Arkin AP (2005) Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations. Proc Natl Acad Sci USA 102:2310–2315
62. Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Comp J Phys 22:435–450
63. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. Phys J Chem 81:2340–2361
64. Gibson MA, Bruck J (2000) Efficient exact stochastic simulation of chemical systems with many species and many channels. Phys J Chem A 104:1876–1889
65. Lok L, Brent R (2005) Automatic generation of cellular reaction networks with Molecularizer 1.0. Nature Biotech 23:131–136
66. Cao Y, Li H, Petzold L (2004) Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. Chem J Phys 121:4059–4067
67. McCollum JM, Peterson GD, Cox CD, Simpson ML, Samatova NF (2006) The sorting direct method for stochastic simulation of biochemical systems with varying reaction execution behavior. Comp Biol Chem 30:39–49
68. Plyasunov S, Arkin AP (2007) Efficient stochastic sensitivity analysis of discrete event systems. Comput J Phys 221:724–738
69. Bardwell L (2004) A walk-through of the yeast mating pheromone response pathway. Peptides 25:1465–1476
70. Morton-Firth CJ, Bray D (1998) Predicting temporal fluctuations in an intracellular signalling pathway. Theor J Biol 192:117–128
71. LeNovère N, Shimizu TS (2001) StochSim: modelling of stochastic biomolecular processes. Bioinformatics 17:575–576
72. Lu T, Volfson D, Tsimring L, Hasty J (2004) Cellular growth and division in the Gillespie algorithm. Syst Biol 1:121–128
73. McAdams H, Arkin A (1998) Simulation of prokaryotic genetic circuits. Annu Rev Biophys Biomol Struct 27:199–224
74. Paulsson J (2004) Summing up the noise in gene networks. Nature 427:415–418
75. McAdams HH, Arkin A (1997) Stochastic mechanisms in gene expression. Proc Natl Acad Sci USA 94:814–819
76. Arkin A, Ross J, McAdams HH (1998) Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. Genetics 149:1633–1648
77. Wolf DM, Vazirani VV, Arkin AP (2005) Diversity in times of adversity: probabilistic strategies in microbial survival games. Theor J Biol 234:227–253
78. Fiering S, Whitelaw E, Martin DIK (2000) To be or not to be active: the stochastic nature of enhancer action. BioEssays 22:381–387
79. Barkai N, Leibler S (2000) Biological rhythms: Circadian clocks limited by noise. Nature 403:267–268
80. Thattai M, van Oudenaarden A (2004) Stochastic gene expression in fluctuating environments. Genetics 167:523–530
81. Kussell E, Leibler S (2005) Phenotypic diversity, population growth, and information in fluctuating environments. Science 309:2075–2078
82. Pedraza JM, van Oudenaarden A (2005) Noise propagation in gene networks. Science 307:1965–1969
83. Swain PS, Elowittz MB, Siggia ED (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc Natl Acad Sci USA 99:12795–12800
84. Mettetal JT, Muzzey D, Pedraza JM, Ozbudak EM, van Oudenaarden A (2006) Predicting stochastic gene expression dynamics in single cells. Proc Natl Acad Sci USA 103:7304–7309
85. Kierzek AM, Zaim J, Zielenkiewicz P (2001) The effect of transcription and translation initiation frequencies on the stochastic fluctuations in prokaryotic gene expression. Biol J Chem 276:8165–8172
86. Peccoud J, Ycart B (1995) Markovian modeling of gene-product synthesis. Theor Popul Biol 48:222–234
87. Rosenfeld N, Young JW, Alon U, Swain PS, Elowittz MB (2005) Gene regulation at the single-cell level. Science 307:1962–1965
88. Kitano H (2004) Biological robustness. Nature Rev Genet 5:826–837
89. Alon U, Surette MG, Barkai N, Leibler S (1999) Robustness in bacterial chemotaxis. Nature 397:168–171
90. Barkai N, Leibler S (1997) Robustness in simple biochemical networks. Nature 387:913–917
91. Stelling J, Sauer U, Szallasi Z, Doyle FJI, Doyle J (2004) Robustness of cellular functions. Cell 118:675–685
92. Vilar JMG, Kueh HY, Barkai N, Leibler S (2002) Mechanisms of noise-resistance in genetic oscillators. Proc Natl Acad Sci USA 99:5988–5992
93. Aldana M, Cluzel P (2003) A natural class of robust networks. Proc Natl Acad Sci USA 100:8710–8714
94. Thattai M, van Oudenaarden A (2002) Attenuation of noise in ultrasensitive signaling cascades. Biophys J 82:2943–2950
95. Yi T-M, Huang Y, Simon MI, Doyle J (2000) Robust perfect adaptation in bacterial chemotaxis through integral feedback control. Proc Natl Acad Sci USA 97:4649–4653
96. Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB (2004) Noise minimization in eukaryotic gene expression. PLoBiol S 2:1–5
97. Voigt CA, Wolf DM, Arkin AP (2005) The Bacillus subtilis sin operon: an evolvable network motif. Genetics 169:1187–1202

98. Cao Y, Gillespie DT, Petzold LR (2006) Efficient step size selection for the tau-leaping simulation method. Chem J Phys 124:044109
99. Gillespie DT, Petzold LR (2003) Improved leap-size selection for accelerated stochastic simulation. Chem J Phys 119:8229–8234
100. Gillespie DT (2001) Approximate accelerated stochastic simulation of chemically reacting systems. Chem J Phys 115:1716–1733
101. Cao Y, Gillespie DT, Petzold LR (2005) Avoiding negative populations in explicit Poisson tau-leaping. Chem J Phys 123:054104
102. Chatterjee A, Mayawala K, Edwards JS, Vlachos DG (2005) Time accelerated Monte Carlo simulations of biological networks using the binomial tau-leap method. Bioinformatics 21:2136–2137
103. Pettigrew MF, Resat H (2007) Multinomial tau-leaping method for stochastic kinetic simulations. Chem J Phys 126:084101
104. Zwanzig R (2001) A chemical Langevin equation with non-Gaussian noise. Phys J Chem B 105:6472–6473
105. Gillespie DT (1996) The multivariate Langevin and Fokker–Planck equations. Am Phys J 64:1246–1257
106. Gillespie DT (2000) The chemical Langevin equation. Chem J Phys 113:297–306
107. Gillespie DT (2002) The chemical Langevin and Fokker–Planck equations for the reversible isomerization reaction. Phys J Chem A 106:5063–5071
108. Wang H, Peskin CS, Elston TC (2003) A robust numerical algorithm for studying biomolecular transport processes. Theor J Biol 221:491–511
109. Xing J, Wang H, Oster G (2005) From continuum Fokker–Planck models to discrete kinetic models. Biophys J 89:1551–1563
110. Tao Y (2004) Intrinsic noise, gene regulation and steady-state statistics in a two-gene network. Theor J Biol 231:563–568
111. van der Mee CVM, Zweifel PF (1987) A Fokker–Planck equation for growing cell populations. Math J Biol 25:61–72
112. Sato K, Kaneko K (2006) On the distribution of state values of reproducing cells. Phys Biol 3:74–82
113. Hill NA, Häder D-P (1997) A biased random walk model for the trajectories of swimming micro-organisms. Theor J Biol 186:503–526
114. Schienbein M, Gruler H (1993) Langevin equation, Fokker–Planck equation and cell migration. Bull Math Biol 55:585–608
115. Xing J, Liao J-C, Oster G (2005) Making ATP. Proc Natl Acad Sci USA 102:16539–16546
116. Elston TC, Oster G (1997) Protein turbines I: the bacterial flagellar motor. Biophys J 73:703–721
117. Allen RJ, Frenkel D, ten Wolde PR (2006) Simulating rare events in equilibrium or nonequilibrium stochastic systems. Chem J Phys 124:024102
118. Rathinam M, Petzold LR, Cao Y, Gillespie DT (2003) Stiffness in stochastic chemically reacting systems: The implicit tau-leaping method. Chem J Phys 119:12784
119. Cao Y, Gillespie DT, Petzold LR (2005) The slow-scale stochastic simulation algorithm. Chem J Phys 122:014116
120. Cao Y, Gillespie DT, Petzold LR (2005) Accelerated stochastic simulation of the stiff enzyme-substrate reaction. Chem J Phys 123:144917
121. Rao CV, Arkin AP (2003) Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm. Chem J Phys 118:4999–5010
122. Adalsteinsson D, McMillen D, Elston TC (2004) Biochemical network stochastic simulator (BioNetS): software for stochastic modeling of biochemical networks. Bioinformatics BMC 5:24
123. Vasudeva K, Bhalla US (2004) Adaptive stochastic-deterministic chemical kinetic simulations. Bioinformatics 20:78–84
124. Baumeister W (2002) Electron tomography: towards visualizing the molecular organization of the cytoplasm. Curr Opin Struct Biol 12:679–684
125. Gierer A, Meinhardt H (1972) A theory of biological pattern formation. Biol Cyber 12:30–39
126. Maini PK, Painter KJ, Chau HNP (1997) Spatial pattern formation in chemical and biological systems. Chem J Soc Faraday Trans 93:3601–3610
127. Gurdon JB, Bourillot P-Y (2001) Morphogen gradient interpretation. Nature 413:797–803
128. Meinhardt H, de Boer PAJ (2001) Pattern formation in *Escherichia coli*: A model for the pole-to-pole oscillations of Min proteins and the localization of the division site. Proc Natl Acad Sci USA 98:14202–14207
129. Howard M, Rutenberg AD (2003) Pattern formation inside bacteria: fluctuations due to the low copy number of proteins. Phys Rev Lett 90:128102
130. Huang KC, Meir Y, Wingreen NS (2003) Dynamic structures in *Escherichia coli*: spontaneous formation of MinE rings and MinD polar zones. Proc Natl Acad Sci USA 100:12724–12728
131. Lutkenhaus J (2007) Assembly and dynamics of the bacterial MinCDE system and spatial regulation of the Z ring. Ann Rev Biochem 76:14.11–14.24
132. Fange D, Elf J (2006) Noise-induced Min phenotypes in *E coli*. PLoComp S Biol 2:637–648
133. Kerr RA, Levine H, Sejnowski TJ, Rappel W-J (2006) Division accuracy in a stochastic model of Min oscillations in *Escherichia coli*. Proc Natl Acad Sci USA 103:347–352
134. Cytrynbaum E, Marshall BDL (2007) A multi-stranded polymer model explains MinDE dynamics in *E coli* cell division. Biophys J 93:1134–1150
135. Bray D (1998) Signaling complexes: biophysical constraints on intracellular communication. Annu Rev Biophys Biomol Struct 27:59–75
136. Slepchenko BM, Schaff JC, Carson JH, Loew LM (2002) Computational cell biology: Spatiotemporal simulation of cellular events. Annu Rev Biophys Biomol Struct 31:423–441
137. Meyers J, Craig J, Odde DJ (2006) Potential for control of signaling pathways via cell size and shape. Curr Biol 16:1685–1693
138. Rao CV, Kirby JR, Arkin AP (2005) Phosphatase localization in bacterial chemotaxis: divergent mechanisms, convergent principles. Phys Biol 2:148–158
139. Agmon N, Edelstein AL (1997) Collective binding properties of receptor arrays. Biophys J 72:1582–1594
140. Lagerholm BC, Thompson NL (1998) Theory for ligand rebinding at cell membrane surfaces. Biophys J 74:1215–1228
141. Andrews SS (2005) Serial rebinding of ligands to clustered receptors as exemplified by bacterial chemotaxis. Phys Biol 2:111–122

142. Elf J, Ehrenberg M (2004) Spontaneous separation of bi-stable biochemical systems into spatial domains of opposite phases. Syst Biol 1:230–236
143. van Zon JS, ten Wolde PR (2005) Green's function reaction dynamics: A particle-based approach for simulating biochemical networks in time and space. Chem J Phys 123:234910
144. Stiles JR, Bartol TM (2001) Monte Carlo methods for simulating realistic synaptic microphysiology using MCell. In: De Schutter E (ed) Computational Neuroscience: Realistic Modeling for Experimentalists. Press CRC, Boca Raton
145. Dab D, Boon J-P, Li Y-X (1991) Lattice-gas automata for coupled reaction-diffusion equation. Phys Rev Lett 66:2535–2539
146. Ermentrout GB, Edelstein-Keshet L (1993) Cellular automata approaches to biological modeling. Theor J Biol 160:97–133
147. Duke TAJ, LeNovère N, Bray D (2001) Conformational spread in a ring of proteins: a stochastic approach to allostery. Mol J Biol 308:541–553
148. Goldman J, Andrews SS, Bray D (2004) Size and composition of membrane protein clusters predicted by Monte Carlo analysis. Eur Biophys J 33:506–512
149. Grima R, Schnell S (2006) A systematic investigation of the rate laws valid in intracellular environments. Biophys Chem 124:1–10
150. Turing AM (1990) The chemical basis of morphogenesis. Bull Math Biol 52:153–197
151. Cross MC, Hohenberg PC (1993) Pattern formation outside of equilibrium. Rev Mod Phys 65:851–1123
152. Slepchenko B, Schaff J, Macara I, Loew LM (2003) Quantitative cell biology with the Virtual Cell. Cell TRENDS Biol 13:570–576
153. Fink CC, Slepchenko B, Moraru II, Watras J, Schaff JC, Loew LM (2000) An image-based model of calcium waves in differentiated neuroblastoma cells. Biophys J 79:163–183
154. Fink CC, Slepchenko B, Moraru II, Schaff J, Watras J, Loew LM (1999) Morphological control of inositol-1,4,5-triphosphate-dependent signals. Cell J Biol 147:929–935
155. Hernjak N, Slepchenko B, Fernald K, Fink CC, Fortin D, Moraru II, Watras J, Loew LM (2005) Modeling and analysis of calcium signaling events leading to long-term depression in cerebellar Purkinje cells. Biophys J 89:3790–3806
156. Baras F, Malek-Mansour M (1996) Reaction-diffusion master equation: A comparison with microscopic simulations. Phys Rev E 54:6139–6148
157. Stundzia AB, Lumsden CJ (1996) Stochastic simulation of coupled reaction-diffusion processes. Comput J Phys 127:196–207
158. Nicolis G, Malek-Mansour M (1980) Systematic analysis of the multivariate master equation for a reaction-diffusion system. Stat J Phys 22:495–512
159. Kruse K, Elf J (2006) Kinetics in spatially extended systems. In: Szallasi Z, Stelling J, Periwal V (eds) System Modeling in Cell Biology From Concepts to Nuts and Bolts. Press MIT, Cambridge, pp 177–198
160. Hynes JT (1985) The theory of reactions in solution. In: Baer M (ed) Theory of Chemical Reaction Dynamics. Press CRC, Boca Raton, pp 171–234
161. Cohen B, Huppert D, Agmon N (2000) Non-exponential Smoluchowski dynamics in fast acid-base reaction. Am J Chem Soc 122:9838–9839
162. Noyes RM (1955) Kinetics of competitive processes when reactive fragments are produced in pairs. Am J Chem Soc 77:2042–2045
163. Pines E, Huppert D (1988) Geminate recombination in excited-state proton transfer reactions: Numerical solution of the Debye–Smoluchowski equation with backreaction and comparison with experimental results. Chem J Phys 88:5620–5630
164. Verkman AS (2002) Solute and macromolecule diffusion in cellular aqueous compartments. Trends Biochem Sci 27:27–33
165. Schnell S, Turner TE (2004) Reaction kinetics in intracellular environments with macromolecular crowding: simulations and rate laws. Prog Biophys Mol Biol 85:235–260
166. Fulton AB (1982) How crowded is the cytoplasm? Cell 30:345–347
167. Bernstein D (2005) Simulating mesoscopic reaction-diffusion systems using the Gillespie algorithm. Phys Rev E 71:041103
168. Ander M, Beltrao P, Di Ventura B, Ferkinghoff-Borg J, Foglierini M, Kaplan A, Lemerle C, Tomás-Oliveira I, Serrano L (2004) SmartCell, a framework to simulate cellular processes that combines stochastic approximation with diffusion and localisation: analysis of simple networks. Syst Biol 1: 129–138
169. Fricke T, Schnakenberg J (1990) Monte-Carlo simulation of an inhomogeneous reaction-diffusion system in the biophysics of receptor cells. Z Phys B 83:277–284
170. Isaacson SA, Peskin CS (2006) Incorporating diffusion in complex geometries into stochastic chemical kinetics simulations. Sci SIAMJ Comput 28:47–74
171. Hattne J, Fange D, Elf J (2005) Stochastic reaction-diffusion simulation with MesoRD. Bioinformatics 21:2923–2924
172. Frenkel D, Smit B (2002) Understanding molecular simulation: from algorithms to applications. Academic, San Diego
173. Berg HC (1993) Random Walks in Biology. Princeton Univ Press, Princeton
174. Gillespie DT (1996) The mathematics of Brownian motion and Johnson noise. Am Phys J 64:225–240
175. Rice SA (1985) Diffusion Limited Reactions. Elsevier, Amsterdam
176. Ermak DL, McCammon JA (1978) Brownian dynamics with hydrodynamic interactions. Chem J Phys 69:1352–1360
177. Northrup SH, Allison SA, McCammon JA (1984) Brownian dynamics simulation of diffusion-influenced bimolecular reactions. Chem J Phys 80:1517–1524
178. Northrup SH (1988) Diffusion-controlled ligand binding to multiple competing cell-bound receptors. Phys J Chem 92:5847–5850
179. Northrup SH, Erickson HP (1992) Kinetics of protein-protein association explained by Brownian dynamics computer simulation. Proc Natl Acad Sci USA 89:3338–3342
180. Edelstein AL, Agmon N (1993) Brownian dynamics simulations of reversible reactions in one dimension. Chem J Phys 99:5396–5404
181. Oh C, Kim H, Shin KJ (2002) Excited-state diffusion-influenced reversible association-dissociation reaction: Brownian dynamics simulation in three dimensions. Chem J Phys 117:3269–3277
182. Kim H, Yang M, Shin KJ (1999) Dynamic correlation effect in reversible diffusion-influenced reactions: Brownian dynamics simulation in three dimensions. Chem J Phys 111:1068–1075
183. Agmon N, Edelstein A (1995) Geometric and many-particle aspects of transmitter binding. Biophys J 68:815–825

184. Edelstein AL, Agmon N (1997) Brownian simulation of many-particle binding to a reversible receptor array. Comput J Phys 132:260–275
185. Agmon N, Szabo A (1990) Theory of reversible diffusion-influenced reactions. Chem J Phys 92:5270–5284
186. Kim H, Shin KJ (1999) Exact solution of the reversible diffusion-influenced reaction for an isolated pair in three dimensions. Phys Rev Lett 82:1578–1581
187. van Zon JS, ten Wolde PR (2005) Simulating biochemical networks at the particle level in time and space: Green's function reaction dynamics. Phys Rev Lett 94:128103
188. van Zon JS, Morelli MJ, Tanase-Nicola S, ten Wolde PR (2006) Diffusion of transcription factors can drastically enhance the noise in gene expression. Biophys J 91:4350–4367
189. Lipkow K (2006) Changing cellular location of CheZ predicted by molecular simulations. Comp PLOS Biol 2:301–310
190. Lipkow K, Andrews SS, Bray D (2004) Simulated diffusion of CheYp through the cytoplasm of *E coli*. J Bact 187:45–53
191. Tournier AL, Fitzjohn PW, Bates PA (2006) Probability-based model of protein-protein interactions on biological timescales. Algorithms Molec Biol 1:25
192. Tolle DP, Le Novère N (2006) Particle-based stochastic simulation in systems biology. Curr Bioinformatics 1:1–6
193. Franks KM, Bartol TM, Sejnowski TJ (2002) A Monte Carlo model reveals independent signaling at central glutamatergic synapses. Biophys J 83:2333–2348
194. Coggan JS, Bartol TM, Esquenazi E, Stiles JR, Lamont S, Martone ME, Berg DK, Ellisman MH, Sejnowski TJ (2005) Evidence for ectopic neurotransmission at a neuronal synapse. Science 309:446–451
195. Koh X, Srinivasan B, Ching HS, Levchenko A (2006) A 3D Monte Carlo analysis of the role of dyadic space geometry in spark generation. Biophys J 90:1999–2014
196. Stiles JR, van Helden D, Thomas J, Bartol M, Salpeter EE, Salpeter MM (1996) Miniature endplate current rise times < 100 microseconds from improved dual recordings can be modeled with passive acetylcholine diffusion from a synaptic vesicle. Proc Natl Acad Sci USA 93:5747–5752
197. Stiles JR, Kovyazina IV, Salpeter EE, Salpeter MM (1999) The temperature sensitivity of miniature endplate currents is mostly governed by channel gating: Evidence from optimized recordings and Monte Carlo simulations. Biophys J 77:1177–1187
198. Bartol TMJ, Land BR, Salpeter EE, Salpeter MM (1991) Monte Carlo simulation of miniature endplate current generation in the vertebrate neuromuscular junction. Biophys J 59:1290–1307
199. Howard M, Rutenberg AD, de Vet S (2001) Dynamic compartmentalization of bacteria: accurate division in *E coli*. Phys Rev Lett 87:278102
200. Kruse K (2002) A dynamic model for determining the middle of *Escherichia coli*. Biophys J 82:618–627
201. Wio HS (1996) Stochastic resonance in a spatially extended system. Phys Rev E 54:R3075–R3078
202. Hu Z, Gogol EP, Lutkenhaus J (2002) Dynamic assembly of MinD on phospholipid vesicles regulated by ATP and MinE. Proc Natl Acad Sci USA 99:6761–6766
203. Hu Z, Saez C, Lutkenhaus J (2003) Recruitment of MinC, an inhibitor of Z-ring formation, to the membrane in *Escherichia coli*: role of MinD and MinE. J Bact 185:196–203
204. Shih Y-L, Fu X, King GF, Le T, Rothfield L (2002) Division site placement in *E coli*: mutations that prevent formation of the MinE ring lead to loss of the normal midcell arrest of growth of polar MinD membrane domains. EMBOJ 21:3347–3357
205. Shih Y-L, Le T, Rothfield L (2003) Division site selection in *Escherichia coli* involves dynamic redistribution of Min proteins within coiled structures that extend between the two cell poles. Proc Natl Acad Sci USA 100:7865–7870
206. Shih Y-L, Kawagishi I, Rothfield L (2005) The MreB and Min cytoskeletal-like systems play independent roles in prokaryotic polar differentiation. Mol Microbiol 58:917–928
207. Suefuji K, Valluzzi R, RayChaudhuri D (2002) Dynamic assembly of MinD into filament bundles modulated by ATP, phospholipids, and MinE. Proc Natl Acad Sci USA 99:16776–16781
208. Pavin N, Paljetak C, Krstic V (2006) Min-protein oscillations in *Escherichia coli* with spontaneous formation of two-stranded filaments in a three-dimensional stochastic reaction-diffusion model. Phys Rev E 73:021904
209. Adelman JL, Andrews SS (2004) Intracellular pattern formation: A spatial stochastic model of bacterial division site selection proteins MinProc CDE. Complex Systems Summer School Final Project Papers, Santa Fe Institute, Santa Fe
210. Drew DA, Osborn MJ, Rothfield LI (2005) A polymerization-depolymerization model that accurately generates the self-sustained oscillatory system involved in bacterial division site placement. Proc Natl Acad Sci USA 102:6114–6118
211. Andrews SS, Arkin AP (2007) A mechanical explanation for cytoskeletal rings and helices in bacteria. Biophys J 93:1872–1884
212. Lippincott-Schwartz J, Snapp E, Kenworthy A (2001) Studying protein dynamics in living cells. Nat Rev Mol Cell Biol 2:444–456

Stochastic Noises, Observation, Identification and Realization with

Giorgio Picci
Department of Information Engineering,
University of Padua, Padua, Italy

Article Outline

Glossary
Introduction
Stochastic Realization
Wide-Sense Stochastic Realization
Geometric Stochastic Realization
Dynamical System Identification
Future Directions
Bibliography

Glossary

Wiener process A *Wiener process* is a continuous time stochastic process $w = \{w(t)\,;\, t \in \mathbb{R}\}$ with continuous sample paths and stationary independent incre-

ments of zero mean and finite variance. The variance of the increments $w(t) - w(s)$ must then have the form $\sigma^2(t-s)$ and $w(t) - w(s)$ must have a Gaussian distribution with mean zero (Levy's theorem [30]). The Wiener process is *normalized* (or *standard*) if $\sigma^2 = 1$. A *p-dimensional* Wiener process is a vector stochastic process having p components, w_k; $k = 1, \ldots, p$, which are independent Wiener processes. A *wide-sense* p-dimensional normalized Wiener process is a continuous time stochastic process $w = \{w(t)\,;\, t \in \mathbb{R}\}$ which is continuous in mean square and has stationary orthogonal increments; i. e.,

$$\mathbb{E}\{[w(t_1) - w(s_1)][w(t_2) - w(s_2)]'\} = I_p|[s_1,\ t_1] \cap [s_2,\ t_2]\,|$$

where $\mathbb{E}$ denotes expectation, I_p is the $p \times p$ identity matrix, the prime denotes transpose and $|\cdot|$ denotes length (Lebesgue measure). This is a weaker concept than that of a Wiener process.

A Wiener process with $w(0) = 0$ is sometimes called a *Brownian motion process*. Usually a Brownian motion is only defined on the half line $\mathbb{R}_+$.

Wiener processes and stochastic integrals with respect to a Wiener process, introduced by Wiener and Itô, are the basic building blocks of *stochastic calculus* [14,25]. Since stochastic integrals only depend on the increments of w, it is immaterial whether an arbitrary random vector is added to w. The equivalence class of (wide sense) Wiener processes differing from each other by the addition of an arbitrary (constant) random vector of finite variance (e. g. $w(0)$), will still be called a Wiener process and denoted by the symbol $\mathrm{d}w$.

White noise Continuous time *white noise* is the formal time derivative of a Wiener process. This derivative must be understood in a suitable distributional sense since it fails to exist as a limit in any of the standard senses of probability theory.

Much more standard is the notion of *discrete-time* (strict or wide-sense) normalized *white noise process*. It is a sequence of independent equally distributed (respectively orthogonal) random vectors $w = \{w(t)\,;\, t \in \mathbb{Z}\}$ with unit variance; i. e.,

$$\mathbb{E}\{w(t)w(s)'\} = I\delta_{ts} := \begin{cases} I & \text{if } s = t \\ 0 & \text{if } s \neq t\,. \end{cases}$$

This last condition alone characterizes wide-sense white noise. We shall always require that w should have zero mean.

Stochastic differential and difference equations Stochastic differential equations are equations that define pathwise a continuous time stochastic process x by "local" evolution laws of the type

$$\mathrm{d}x(t) = f(x)\mathrm{d}t + G(x)\mathrm{d}w(t)\,,$$

where $\mathrm{d}w$ is a vector Wiener process. The equation should in reality be interpreted as an integral equation, the last term being a stochastic integral in the sense of Itô, see [14,25]. Under certain growth conditions on the coefficients the equation can be shown to have a unique solution which, in case f and G depend pointwise on $x(t)$, is a continuous Markov process, in fact a *diffusion process*.

In discrete time ($t \in \mathbb{Z}$) a *stochastic difference equation* is an object of the type

$$x(t+1) = f(x(t))\mathrm{d}t + G(x(t))w(t)$$

where now w is white noise. The solution of an equation of this type is a (discrete-time) Markov process.

Stochastic dynamical systems In continuous time, a (finite-dimensional) *stochastic system* is a pair (x, y) of vector, say n and m dimensional, stochastic processes satisfying equations of the type

$$\mathrm{d}x(t) = f(x(t))\mathrm{d}t + G(x(t))\mathrm{d}w(t)$$
$$\mathrm{d}y(t) = h(x(t))\mathrm{d}t + J(x(t))\mathrm{d}w(t)\,.$$

The process x, which is Markov, is called the *state* of the system while y is the *output* process. The differential notation in the second equation is merely to allow a possible additive white noise component in the output process so y is actually the integral of the variable which would physically be called the output of the system. If there is no additive white noise component ($J \equiv 0$) this trick is unnecessary and the output variable is simply expressed as a memoryless function of the state, $\mathrm{d}y/\mathrm{d}t = h(x(t))$. A similar definition, with the obvious modifications, serves to introduce the concept of discrete-time stochastic systems.

The probabilistic essence of the concept of stochastic system is more generally captured in terms of *conditional independence* of the sigma-algebras induced by the underlying processes. From this it can be clearly seen that the state process at time t plays the role of *dynamic memory* of the system: the future and the past of the output and of the state sigma-algebras are conditionally independent given the present state $x(t)$. For a formal definition the reader is for example referred to p. 219 in [51]. We shall introduce these concepts in a wide-sense context later on.

Introduction

In this article we discuss some general ideas which motivate the use of stochastic dynamical models in applied sciences. We discuss the inherent mathematical problem of stochastic model building, called *stochastic realization* and the statistical problem of *(dynamic) system identification*, i. e. procedures for estimating dynamic stochastic models starting from observed data. We shall in particular discuss linear stochastic state-space models with random inputs, their construction and the relevant identification techniques. We shall assume that the reader has some general background on stationary stochastic processes, statistics [12] and linear system theory as for example exposed in the textbook [24].

There are no physical laws telling us how to obtain stochastic descriptions of nature. The laws of mechanics, electromagnetism, fluid dynamics etc. are, by nature, essentially deterministic. A fundamental empirical observation is that stochastic models generally come about as *aggregate* descriptions of complicated large deterministic systems. Think for example of describing mathematically the outcome of a dice-throwing experiment. Doing this by the laws of physics involves an enormous number of complicated factors such as the mechanics of a cubic-shaped rigid body flying in the air after an initial impulse applied on a specific region of one of its six faces. To describe the trajectory (i. e. the motion of the barycenter, the body orientation and the angular momenta) one should take into account, besides gravity, lift phenomena due to portance, drag, added-mass effects etc., describing the interaction with the surrounding fluid. Then one should describe the discontinuous mechanics of the dice landing on an elastic surface etc. Modeling all of this by the known laws of physics is perhaps possible and we may in principle be able to set up a set of algebraic-differential equations allowing us to predict exactly (or "almost exactly") which upper face of the dice will eventually show. This however, provided we knew an enormous number of geometrical and physical parameters of the system such as, for example, the dimensions of the dice, its initial position and orientation, its mass and moments of inertia, the location, direction and strength of the initial impulse, the density of the air surrounding the trajectory (depending on the temperature, humidity etc.), the possible air convection phenomena, the mechanical structure and geometry of the landing medium, etc. Indeed, this predictive model based on the (deterministic) laws of physics would require a very large number of equations and involve an unimaginable detailed prior knowledge of the parameters of the experiment. Trivially, the rudimentary stochastic model consisting of a probability distribution on the six possible outcomes $\{1, 2, \ldots, 6\}$, although not allowing an "exact", i. e. "deterministic" prediction of the outcome, is incomparably simpler. One could venture to say that most stochastic models used in science and engineering come about, for the same reason, as simple aggregate descriptions of complicated deterministic systems. Thermodynamics, to name just one of the most conspicuous instances of this phenomenon, can be seen as stochastic aggregation of large ensembles of microscopic particles evolving in time according to the deterministic laws of mechanics [28,43].

Simplification is payed in terms of uncertainty and the predictions based on stochastic aggregate models are by nature uncertain. The modeling process hence requires us to quantify the uncertainty of aggregate modes. In a more precise sense, we should investigate their mathematical structure and describe how to construct them. This will be one of the main concerns of this article.

A Brownian Particle in a Heat Bath

The process of constructing aggregate models can be elucidated by a famous example, the *Ford–Kac–Mazur model* [20] which is probably the simplest explicitly solvable example of a problem of this nature.

Consider the problem of describing the dynamics of a particle, called the *Brownian particle*, coupled to a "heat bath" consisting of a very large number of identical particles obeying the laws of classical Hamiltonian mechanics and moving under the influence of a quadratic potential. Mathematically, the ensemble can be described as a large number of coupled harmonic oscillators. The Brownian particle is the only particle of the ensemble which is assumed to be accessible to external observation. One assumes that the macroscopic observables of the system are the relative position $q(t)$ and/or the momentum $p(t)$ of the Brownian particle, at each instant of time.

It is shown in [20] that in the limit when the number of particles in the heat bath tends to infinity, the motion of the Brownian particle is governed (exactly!) by the stochastic differential equations,

$$\dot{q}(t) = p(t) \tag{1}$$

$$dp(t) = -ap(t)\mathrm{d}t + k\mathrm{d}w(t)\,, \quad a > 0 \tag{2}$$

the second of which is the ubiquitous *Langevin equation* of statistical physics. This equation contains a dynamical *friction term*, $(-ap(t))$, and a forcing function term $w(t)$ which is described mathematically as a Wiener process. Physically, these terms can be interpreted as the influence of the surrounding medium on the observed particle, the

friction accounting for the energy transferred from the particle to the heat bath and the forcing term $w(t)$ representing the sum of infinitesimal "random shocks" inflicted to the particle by the surrounding medium.

The stochastic process describing the temporal evolution of the observables $q(t)$ and $p(t)$ is a Markov diffusion process (in fact, both $p(t)$ and the two-dimensional vector process $[q(t)\ p(t)]$ are Markov). The evolution of the macroscopic observables is hence *dissipative*, because of the friction term, and *irreversible*, since diffusion processes evolve in time by the action of a semigroup which cannot be inverted to become a group. This is in accordance with the basic postulates of (macroscopic) thermodynamics.

The Ford–Kac–Mazur example can be generalized [48,49,50,52]. Assume a microscopic system described by a Hamiltonian H on the phase space Ω, a smooth manifold of very large dimension $2N$. The solution of the Hamilton equations determines the phase of the system, say configurations and momenta $z(t) := [q(t)\ \ p(t)]'$ at each time t, uniquely in terms of the initial value $z(0) \in \Omega$. The correspondence defines a flow $z(t) = \Phi(t)z(0)$ on the phase space which leaves invariant the total energy, $H(z(t))$, of the system. The microscopic evolution is in this sense, reversible and conservative. However, microscopic phenomena are *complex* in the sense that they involve an enormous number of interacting components and it is impossible to keep track of the dynamics of such a complex system. One must then resort to a statistical description.

A probability distribution on Ω which is *invariant for the Hamiltonian flow* $U(t)$ defines "thermal equilibrium". It is known that in a finite-dimensional space any absolutely continuous $U(t)$-invariant probability measure admits a one parameter family of densities $\rho(z)$ of the Maxwell–Boltzmann type, equal to a normalization constant times $\exp[-\frac{1}{2\beta}H(z)]$, $\beta > 0$, being interpreted as the "absolute temperature". For a quadratic Hamiltonian in a linear phase space these distributions are Gaussian. The choice of an initial phase $z(0)$ of a system in thermal equilibrium can then be interpreted as a random choice of an elementary event in the elementary outcome space $\{\Omega,\ \rho\}$.

It follows that in thermal equilibrium any observable; i.e. any measurable function $h\colon \Omega \to \mathbb{R}$ on the phase space can be regarded as a random variable. In fact, once combined with the (measure-preserving) Hamiltonian flow $z(t) = U(t)z(0)$, any measurable observable defines a *stationary stochastic process*. In general, the time evolution of any finite, say m-dimensional, family of observables $\{h_1, \ldots, h_m\}$ can be identified with a vector-valued m-dimensional stationary process $\{y(t)\}$ on Ω, with components

$$y_k(t, z(0)) := h_k(U(t)z(0)) \tag{3}$$

where $z(0) \in \Omega$ is the elementary event and $U(t)$ the measure preserving group of transformations on the underlying probability space.

The main point of this story is that the "randomization" of the phase space may, under certain conditions, lead to a description of certain observable processes $\{y(t)\}$ of (3) by a stochastic dynamical model of a *much simpler structure than the microscopic deterministic Hamiltonian description*. These could for example be representations of $\{y(t)\}$ of the type $y(t) = h(x(t))$ where $\{x(t)\}$ a finite dimensional, say $\mathbb{R}^n$-valued Markov process and h is a suitable function $h\colon \mathbb{R}^n \to \mathbb{R}^m$. This description, in case the Markovian representation has a smaller complexity (dimension) than that of the microscopic description, would precisely meet both the general physical plausibility for a thermodynamic description and the mathematical requirement of reduction of complexity.

Remark 1 This view generalizes the way of conceiving stochastic aggregation in the physical literature, e.g [31,32], where it is generally required that the observables themselves should be Markov and satisfy a Langevin equation.

In general the requirement that the observables should themselves be Markov processes is seldom a reasonable one. For example, in the Brownian particle example the macroscopic observable $q(t)$ by itself need *not* satisfy any stochastic differential equation. Formally, $q(t)$ is just a memoryless function of a two-dimensional Markov process $x(t) := [q(t)\ p(t)]'$ and it is instead the "thermodynamic state" $x(t)$ which satisfies a stochastic differential equation.

Stochastic Realization

Stochastic realization is the abstract version of the aggregation problem mentioned above. One studies the problem of representing an m-dimensional stationary process $\{y(t)\}$ as a *memoryless function of a Markov process*. In this setting the physical phase space and the Hamiltonian group $\{U_t\}$ are generalized to an abstract probability space and to a measure preserving one-parameter group of transformations mapping Ω onto itself (the shift group of the process). In this article we shall assume that $\{y(t)\}$ is smooth with continuous values but the problem area also concerns stationary finite state-processes, see [5,19,46,66], which have important applications in communications and biology. In the present context, the question is when

does a given stationary, m-dimensional, stochastic process $\{y(t)\}$ admit representations of the type $y(t) = h(x(t))$ where $\{x(t)\}$ is a Markov diffusion process taking values on a finite-dimensional state space X and h is say a continuous function from X to $\mathbb{R}^m$.

Since a smooth Markov diffusion process $\{x(t)\}$ is a solution of a stochastic differential equation (Chap. VI, Par. 3 in [14]), any such process must admit representations as the output of a stochastic dynamical system of the type

$$\mathrm{d}x(t) = f(x(t))\mathrm{d}t + G(x(t))\mathrm{d}w(t) \tag{4}$$

$$y(t) = h(x(t)) \tag{5}$$

where $\{w(t)\}$ is a vector m Wiener process. This stochastic dynamical system generalizes the Langevin equation representation. For a Gaussian mean square continuous stationary process, the representation simplifies to a linear model of the form

$$\mathrm{d}x(t) = Ax(t)\mathrm{d}t + B\mathrm{d}w(t) \tag{6}$$

$$y(t) = Cx(t) \tag{7}$$

where A, B, C are constant matrices of appropriate dimensions. There may be a need for introducing an additive noise term also in the second (output) equation when the process y has stationary increments, a situation more general than stationarity, see [37].

Representations as the output of a stochastic dynamical system are called *stochastic realizations* or *state-space realizations* of the process y and the Markov process $x(t)$ is accordingly called the *state process* of y, or of the realization.

Signal descriptions by a finite dimensional stochastic realization can be transformed into other equivalent forms such as ARMA models etc., which are also widely used in the engineering and econometric literature. It should be stressed that, irrespective of which particular equivalent form of the stochastic model, virtually all sequential signal processing, prediction and control algorithms (Kalman filtering to name just the most popular) require availability of a finite-dimensional realization of one form or another. Hence signal representation by a stochastic system is indeed a crucial prerequisite for system analysis and design in engineering and applied sciences.

A substantial amount of literature on the stochastic realization problem has appeared in the last four decades. The stationary (and stationary increments) Gaussian case is now covered by a rather complete linear theory, and is surveyed e. g. in [35,37]. For a more up to date account, see the forthcoming book [41]. The nonlinear realization problem is still underdeveloped, see [45,51,62].

Wide-Sense Stochastic Realization

We shall discuss only the wide-sense version of the realization problem, where random variables and processes are described in terms of first- and second-order moments. Since first- and second-order moments individuate a Gaussian distribution uniquely, the theory will in particular cover the Gaussian case. Hereafter, wide-sense stationarity will be simply referred to as stationarity. Also, a "wide-sense Markov process" will simply be called a "Markov process".

The wide-sense realization problem may look like a very particular modeling problem to deal with, but it is completely solvable and is general enough to reveal quite explicitly some of the important features of the solution set of stochastic realizations, some of which are quite unexpected. It is also motivated by its wide applicability since most times in practice the only reasonable way to describe a priori random phenomena is by second-order moments, say by covariances or spectra.

We advise the reader that in many applied fields one may want to construct dynamical system models involving also exogenous input (or decision) variables. The stochastic realization problem of constructing linear state-space representations of a stationary process with inputs has been studied in [11,53,54,55]. In the last two references one may find also material related to the germane problem of *identification with inputs*.

For reasons of space limitations we shall not touch upon this subject. A pointer to the relevant identification literature will be given in the final section of the article.

We shall discuss in some detail only stochastic realization of *discrete-time* random processes. An analogous (and in fact somewhat simpler) treatment can be given for continuous time processes [35] but we choose discrete-time since this theory makes contact with system identification which we shall discuss later in the section entitled "Dynamical System Identification".

Given a vector (say m-dimensional) wide-sense stationary purely nondeterministic (purely nondeterministic will be abbreviated to p.n.d. in the following; this property is called *linear regularity* in the Russian literature, see [57]) process $y = \{y(t)\}$ with $t \in \mathbb{Z}$ (the integers), one wants to find representations of y in terms of a finite dimensional Markov process. In other words one wants to find linear representations of the stationary process y, of the form

$$x(t+1) = Ax(t) + Bw(t) \tag{8a}$$

$$y(t) = Cx(t) + Dw(t) \tag{8b}$$

and procedures for constructing them. Here $\{w(t)\}$ is a vector, p-dimensional normalized white noise process, i. e. $\mathbb{E}\{w(t)w(s)'\} = I\delta(t-s)$, $\mathbb{E}\{w(t)\} = 0$, δ being the Kronecker delta function. Note that a white noise w could be seen as a (degenerate) kind of Markov process. The reason for having the term $Dw(t)$ in (8b) is to avoid the presence of such degenerate components in the dynamic equation for the state process x which would artificially increase its dimension. With this convention, in the extreme circumstance where $y = w$, the state dimension (of a minimal realization) is zero.

The linear representation (8a) involves auxiliary variables, i. e. random quantities which are not given as a part of the original data, especially the n-dimensional state process x (a stationary Markov process) and the generating white noise w. The peculiar properties of these processes actually embody the desired system structure (8a). Constructing these auxiliary processes is an essential part of the realization problem.

We shall restrict ourselves to representations (8a) for which

- (A,B,C) is a minimal triplet; i. e. (A, B) is a reachable pair and (A, C) is an observable pair, see e. g. [24] for these notions;
- $\begin{bmatrix} B \\ D \end{bmatrix}$ has independent columns.

These are classical conditions of nonredundancy of the model [24] and entail no loss of generality. From standard spectral representation theory, see e. g., Chap. 1 in [57], it follows that Eq. (8a) admits stationary solutions if and only if the rows of the $n \times p$ matrix $(e^{i\theta}I - A)^{-1}B$ are square integrable functions of $\theta \in [-\pi, \pi]$. This is equivalent to the absence of poles of the function $z \to (zI - A)^{-1}B$ on the unit circle. Equivalently, the eigenvalues of A of modulus one, if any, must be "unreachable" for the pair (A, B).

When eigenvalues on the unit circle are present, to guarantee stationarity they must be simple roots of the minimal polynomial of A. These eigenvalues, necessarily in even number say $2k$, give then rise to a sum of k uncorrelated sinusoidal oscillations with random amplitude, the so-called *purely deterministic* component of the process. This purely deterministic component of the stationary Markov process x obeys a fixed undriven (i. e. deterministic) linear difference equation of the type $x(t+1) = A_o x(t)$, which is the restriction of (8a) to the modulus one eigenspace of A. The initial conditions are random variables independent of the driving noise w. Since there is no stochastic forcing term in this restricted state equation, it follows that the presence of a purely deterministic component in the Markov process necessarily implies that the pair (A, B) must be nonreachable. The first of the two conditions of nonredundancy listed above hence excludes the presence of a purely deterministic component in the state, and therefore also in the output process y.

From this discussion it is immediately seen that there are stationary solutions x of the difference Eq. (8a) which are p.n.d. Markov processes if and only if A does not have eigenvalues of modulus one; i. e.

$$|\lambda(A)| \neq 1 \,. \tag{9}$$

When $|\lambda(A)| < 1$, the representation (8a) is called *forward* or *causal*. When $|\lambda(A)| > 1$ the representation is called *backward* or *anticausal*. Traditionally the causality of a representation, i. e. the condition $|\lambda(A)| < 1$ has been regarded as being equivalent to stationarity. In fact, stationarity permits a whole family of representations (8a) of the (same) process x, where the matrices A can have very different spectra. Formulas for computing a backward representation starting from a forward one or vice versa are given in [34,36]. The two representations (in particular the relative A matrices) are related by a particular family of *linear state feedback* transformations [47].

The *covariance function* of a zero-mean wide sense stationary process y is the $m \times m$ matrix function $k \mapsto \Lambda_k$, defined by

$$\Lambda_k := \mathbb{E}\{y(t+k)y(t)'\} = \mathbb{E}\{y(k)y(0)'\} \qquad k \in \mathbb{Z} \,.$$

This matrix function will be the initial data of our problem. Since y is p.n.d. the function Λ admits a Fourier transform called the *spectral density* matrix of y, given by

$$\Phi(z) = \sum_{t=-\infty}^{\infty} \Lambda_t z^{-t}$$

where $z = e^{i\theta}$; $\theta \in [-\pi, \pi]$. It is easy to see that the spectral density matrix of a process admitting a state-space realization (8a) must be a rational function of z. This fact follows easily by eliminating the variable x in (8a) and thereby expressing y as the output of a linear filter as in Fig. 1 below whose transfer function $W(z) = C(zI - A)^{-1}B + D$ is a rational function of z, and then using the classical Kintchine and Wiener formula for the spectrum of a filtered

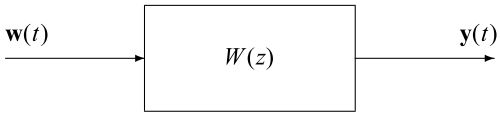

Stochastic Noises, Observation, Identification and Realization with, Figure 1
Shaping filter representation of a stochastic process

stationary process [28,69] so that,

$$\Phi(z) = W(z)W(1/z)' \,. \tag{10}$$

Hence we have

Proposition 1 *The transfer function $W(z) = C(zI - A)^{-1}B + D$ of any state space representation (8a) of the stationary process y, is a* spectral factor *of Φ, in the sense that it satisfies the spectral factorization equation (10).*

It is easy and instructive to check this directly. Supposing that one has a *causal* realization (8a), one can compute for $k \geq 0$,

$$\begin{aligned}\Lambda_k &= \mathbb{E}[\, CA^k x(t) \\ &\quad + \sum_{s=0}^{k-1} CA^{k-1-s} Bw(t+s) + Dw(t+k)\,] \\ &\qquad\qquad [Cx(t) + Dw(t)]' \qquad (11)\\ &= CA^{k-1}\bar{C}' + \frac{1}{2}\Lambda_0\delta(t)\,, \qquad (12)\end{aligned}$$

where $\bar{C}' = \mathbb{E}x(t+1)y(t)'$ and $\Lambda_0 = \mathbb{E}y(t)y(t)'$. By identifying terms, these matrices can be expressed directly in terms of the realization parameters A, B, C, D, by

$$P = APA' + BB' \tag{13a}$$

$$\bar{C}' = APC' + BD' \tag{13b}$$

$$\Lambda_0 = CPC' + DD' \tag{13c}$$

the matrix $P = \mathbb{E}x(t)x(t)'$ being the covariance matrix of the state process. Note that P obeys Eq. (13a) since by causality ($|\lambda(A)| < 1$) $x(t)$ is a function of the infinite past $\{w(s);\ s < t\}$ and hence is uncorrelated with $w(t)$ so that the two terms in the right hand side of (8a) are also uncorrelated and their variances add. In fact, by reachability of the model, P is actually (symmetric and) positive definite.

Now, introduce the decomposition

$$\Phi(z) = \Phi_+(z) + \Phi_+(1/z)' \tag{14}$$

where $\Phi_+(z) = \frac{1}{2}\Lambda_0 + \Lambda_1 z^{-1} + \Lambda_2 z^{-2} + \cdots$ is analytic outside of the unit circle. Clearly $\Phi_+(z)$ must be the Fourier transform of the "causal" tract of the covariance function Λ, computed in (12). Hence

$$\Phi_+(z) = C(zI - A)^{-1}\bar{C}' + \frac{1}{2}\Lambda_0\,. \tag{15}$$

This shows that the spectrum of a process y described by a state-space model (8a), besides being a rational function of z, is directly expressible in parametric form in terms of the parameters of a causal realization. This explicit computation of the spectrum seems to be due to R.E. Kalman and B.D.O. Anderson [4,21].

"Classical" stochastic realization theory, developed in the late 1960s [3,4,17], deals with the inverse problem of computing the parameters A, B, C, D of a state space realization starting from the spectrum (or, equivalently, of the covariance function) of a stationary process y. Of course one assumes here that y is p.n.d. and has a spectral density Φ which is a rational function of $z = e^{j\theta}$.

This inverse problem is just a parametric version of the spectral factorization problem: Given a rational spectrum $\Phi(z)$ one looks for rational spectral factors parametrized as $W(z) = C(zI - A)^{-1}B + D$, solutions of (10). Since rational factors of a given spectral density matrix are in general infinitely many, one must preliminarily single out the interesting solutions by imposing certain restrictions. One very reasonable restriction which is imposed on the solutions of (10) is that the spectral factors should be of *minimal degree*. The degree of a proper rational matrix function is by definition the dimension n of the state of a minimal realization. Minimal degree (for short, *minimal*) spectral factors must have a degree which is exactly one half of the degree of the spectral density matrix $\Phi(z)$.

The solution of the inverse problem is described in the following proposition.

Proposition 2 *Assume the spectrum is given in the parametric form (15) where $|\lambda(A)| < 1$ and $(A, C, \bar{C}')$ is a minimal triplet of dimension n. Then the minimal degree spectral factors of $\Phi(z)$,* analytic in$\{|z| > 1\}$, *are in one-to-one correspondence with the symmetric $n \times n$ matrices P solving the* Linear Matrix Inequality *(LMI)*

$$M(P) := \begin{bmatrix} P - APA' & \bar{C}' - APC' \\ \bar{C} - CPA' & \Lambda_0 - CPC' \end{bmatrix} \geq 0 \tag{16}$$

where the symbol $\geq$ means positive semidefinite. The correspondence is to be understood in the following sense:

Corresponding to each solution $P = P'$ of (16), consider the full column rank matrix factor $\begin{bmatrix} B \\ D \end{bmatrix}$ of $M(P)$,

$$M(P) = \begin{bmatrix} B \\ D \end{bmatrix} [B'D'] \tag{17}$$

(this factor is unique modulo right multiplication by orthogonal matrices) and form the rational matrix

$$W(z) := C(zI - A)^{-1}B + D\,. \tag{18}$$

Then (18) is a minimal realization of a minimal analytic spectral factor of $\Phi(z)$ and all minimal analytic factors can be obtained in this way.

Note that any matrix P solving the LMI satisfies the system (12) and hence can be interpreted as the state covariance of a causal realization of y. It follows that any P solving the LMI is necessarily positive definite.

In fact it has been shown [17] that the set of symmetric solutions to the LMI (16)

$$\mathcal{P} := \{P = P' \,;\, M(P) \geq 0\}$$

is a closed, bounded and convex set with maximal and minimal elements $P_-,\ P_+ \in \mathcal{P}$ which satisfy

$$P_- \leq P \leq P_+ \quad \text{for all} \quad P \in \mathcal{P}$$

where $P_1 \leq P_2$ means that $P_2 - P_1 \geq 0$ (is positive semidefinite).

The existence of solutions to the LMI is equivalent to the property of $\Phi_+(z)$ defined in (14) of being a *positive-real matrix function*, namely that Φ_+ should be analytic in $\{|z| > 1\}$ and that $\Re e\, \Phi_+(e^{j\theta}) \geq 0$. The second condition is automatically satisfied when Φ admits spectral factors since

$$2\Re e\, \Phi_+(e^{j\theta}) = W(e^{j\theta})W(e^{j\theta})^* \geq 0$$

(the star meaning complex conjugate transpose). This property can be characterized in general, for not necessarily stable representations, by the so-called *positive-real lemma* discovered in the 1960s by Kalman, Yakubovich, and Popov [22,56,71] in the context of dissipative systems and stability theory. The first application of the Kalman–Yakubovich–Popov theory to factorization of spectral density functions and to stochastic realization (called *stationary covariance generation*) is due to B.D.O. Anderson [3,4]. Far reaching generalizations to not necessarily stable systems have been pursued by J.C. Willems in his widely quoted paper [70]. The terminology "positive real" derives from network synthesis.

Under certain conditions on the zeros of the spectrum, discussed e. g. in [18], the matrix $\Delta(P) := \Lambda_0 - CPC'$ is nonsingular (and hence positive definite) for all $P \in \mathcal{P}$. In this case (16) is equivalent to the nonnegativity of the Schur complement of $\Delta(P)$ in the matrix $M(P)$, which can be written

$$P - APA' - (\bar{C}' - APC')\Delta(P)^{-1}(\bar{C}' - APC')' \leq 0\,. \quad (19)$$

This is the *Algebraic Riccati Inequality* (ARI) of spectral factorization, first analyzed in the work of [3,17]. The solutions of the ARI with equality sign; i. e. the solutions of the Algebraic Riccati Equation (ARE)

$$P - APA' - (\bar{C}' - APC')\Delta(P)^{-1}(\bar{C}' - APC')' = 0 \quad (20)$$

make $M(P)$ of minimal rank $m = \dim y$ and hence correspond to *square spectral factors* (18). These spectral factors all have the same denominator matrix A and hence only differ by the location of their zeros. In particular one may show that P_-, P_+ solve the ARE and the two spectral factors W_- and W_+ corresponding to P_-, P_+ (are square and) have all of their zeros inside and respectively outside of the unit circle. In other words, W_- is the *minimum phase* and W_+ the *maximum phase* spectral factor. These spectral factors are essentially unique. All other square spectral factors are obtained (in rough terms) by "flipping" some of the zeros of W_- to their reciprocals outside of the unit circle. The maximum phase factor, W_+, is obtained by flipping *all* of the zeros of W_- to their reciprocals. More information on the zero structure of the minimal spectral factors can be found in [39].

In the next section we shall analyze in more detail the correspondence between spectral factors and stochastic realizations. As we shall see, for square spectral factors this correspondence is one-to-one and hence this particular class of stochastic realizations, for which $\dim w = \dim y = m$ can be classified according to their zero location.

The LMI plays an important role in many areas of system theory such as stability theory, dissipative systems and is central in H^∞ control and estimation theory. It seems to be much less appreciated in the scientific community that it plays a very basic role in modeling of stationary random signals as well. As we shall see in the next section certain solutions of the LMI (or of the ARI) have special probabilistic properties and are related to Kalman-filter or "innovations-type" realizations.

Geometric Stochastic Realization

The classical stationary covariance realization theory of the previous section is purely distributional as it says nothing about representation of random quantities in a truly probabilistic sense (i. e. how to generate the random variables or the sample paths of a given process, not just its covariance function). This was implicitly pointed out by Kalman already in [23]. In the last decades a *geometric coordinate-free* approach to stochastic modeling has been put forward in a series of papers by Lindquist, Picci, Ruckebusch et al. [33,34,35,58,59,60] which aims at the representation of random processes in this more specific sense. This motivation is also present in the early papers by Akaike [1,2].

A main point of the geometric approach is to provide procedures for the construction of the random quantities defining a stochastic state-space realization. In particular

the *state space* of a realization is defined in terms of the conditional independence relation between past and future of the signals involved. This relation is intrinsically *coordinate-free* and in the present setting involves only linear subspaces of a given ambient Hilbert space of random variables, typically made of linear functionals of the variables of the process y to be modeled (but in some situations other random data may be used to construct the model).

Constructing the State Space of a Stationary Process

The theme of this section will be a review of geometric realization theory for the stochastic process y. The geometric theory centers on the idea of *Markovian Splitting Subspace* for the process y. This concept is the probabilistic analog of the deterministic notion of state space of a dynamical system and captures at an abstract level the property of "dynamic memory" that the state variables have in deterministic dynamical system theory. Once a stochastic state space is given the construction of the auxiliary random quantities which enter in a state space model and in particular the state process is fairly obvious. The state vector $x(t)$ of a particular realization can be regarded just as a particular basis for the state space, hence once a state-space is constructed, finding state equations is just a matter of choosing a basis and computing coordinates.

Notation 1 *In what follows the symbols $\vee$, $+$ and $\oplus$ will denote vector sum,* direct *vector sum and* orthogonal *vector sum of subspaces, the symbol $\mathbf{x}^{\perp}$ will denote the orthogonal complement of a (closed) subspace $\boldsymbol{X}$ of a Hilbert space with respect to some predefined ambient space. Given a collection $\{X_\alpha \mid \alpha \in A\}$ of subsets of a Hilbert space $\boldsymbol{H}$ we shall denote by $\overline{\text{span}}\{X_\alpha \mid \alpha \in A\}$ the closure in $\boldsymbol{H}$ of the linear (real) vector space generated by the collection. The orthogonal projection onto the subspace $\boldsymbol{X}$ will be denoted by the symbol $\mathbb{E}(\cdot \mid \mathbf{X})$ or by the shorthand $\mathbb{E}^{\mathbf{X}}$. The notation $\mathbb{E}(z \mid \mathbf{X})$ will be used also when z is vector-valued. The symbol will then denote the vector with components $E(z_k \mid \mathbf{X})$, $k = 1, \ldots$. For vector quantities, $|v|$ will denote Euclidean length (or absolute value in the scalar case).*

Let y be a zero mean stationary vector process with finite second-order moments defined on some underlying probability space $\{\Omega, \mathcal{A}, \mu\}$ and let $L^2\{\Omega, \mathcal{A}, \mu\}$ denote the Hilbert space of second-order random variables defined on Ω, endowed with the inner product $< \xi, \eta >= \mathbb{E}\{\xi\eta\}$. Let $\mathbf{H}(y)$ be the (closed) linear subspace of $L^2\{\Omega, \mathcal{A}, \mu\}$, linearly generated by the variables of the process; i. e.

$$\mathbf{H}(y) := \overline{\text{span}}\{y_k(t); k = 1, 2, \ldots, m;\ t \in \mathbb{Z}\}.$$

If y is generated by a linear model of the type (8a) this Hilbert space will in general be a proper subspace of the space $\mathbf{H}(w)$, generated by the input white noise of the model (8a). All random variables of the stochastic system (8a) belong to $\mathbf{H}(w)$ and for this reason $\mathbf{H}(w)$ is called the *ambient space* of the model. By stationarity of w, $\mathbf{H}(w)$ comes naturally equipped with a unitary operator U, called the *shift* of w, such that $Uw_i(t) = w_i(t+1)$ for $i = 1, 2, \ldots, p$ and all $t \in \mathbb{Z}$ (stationarity). Note that the family $\{U^t; t \in \mathbb{Z}\}$ forms a one parameter group of unitary operators on the ambient space; in particular, $U^{-t} = (U^*)^t$. The pair $(\mathbf{H}(w), U)$ is also called a *stationary Hilbert space*. Clearly the processes x and y will also be stationary with respect to U.

The *past* and *future* subspaces (at time zero) of a stationary process are

$$\mathbf{H}^-(y) := \overline{\text{span}}\{y_k(t); k = 1, 2, \ldots, m;\ t < 0\}$$
$$\mathbf{H}^+(y) := \overline{\text{span}}\{y_k(t); k = 1, 2, \ldots, m;\ t \geq 0\}.$$

Because of stationarity everything propagates in time by the action of the unitary group, e. g. the past and future at time t are obtained as $\mathbf{H}_t^-(y) = U^t\mathbf{H}^-(y)$ and $\mathbf{H}_t^+(y) = U^t\mathbf{H}^+(y)$. For this reason the geometric definitions and constructions to follow, although valid for an arbitrary time instant will usually be referred to the time instant $t = 0$.

Let $\mathbf{X}$ be a subspace of some large stationary Hilbert space $\mathbf{H}$ of wide-sense random variables containing $\mathbf{H}(y)$. Define

$$\mathbf{X}_t := U^t\mathbf{X}, \quad \mathbf{X}_t^- := \vee_{s \leq t}\mathbf{X}_s, \quad \mathbf{X}_t^+ := \vee_{s \geq t}\mathbf{X}_s.$$

Definition 1 A *Markovian Splitting Subspace* $\mathbf{X}$ for the process y is a subspace of $\mathbf{H}$ making the vector sums $\mathbf{H}(y)^- \vee \mathbf{X}^-$ and $\mathbf{H}(y)^+ \vee \mathbf{X}^+$ conditionally orthogonal (i. e. conditionally uncorrelated) given $\mathbf{X}$, denoted,

$$\mathbf{H}(y)^- \vee \mathbf{X}^- \perp \mathbf{H}(y)^+ \vee \mathbf{X}^+ \mid \mathbf{X}. \tag{21}$$

The conditional orthogonality condition (21) can be equivalently written as

$$\mathbb{E}[\mathbf{H}(y)^+ \vee \mathbf{X}^+ \mid \mathbf{H}(y)^- \vee \mathbf{X}^-] = \mathbb{E}[\mathbf{H}(y)^+ \vee \mathbf{X}^+ \mid \mathbf{X}] \tag{22}$$

$$\mathbb{E}[\mathbf{H}(y)^- \vee \mathbf{X}^- \mid \mathbf{H}(y)^+ \vee \mathbf{X}^+] = \mathbb{E}[\mathbf{H}(y)^- \vee \mathbf{X}^- \mid \mathbf{X}] \tag{23}$$

which formalize the intuitive meaning of the splitting subspace $\mathbf{X}$ as a dynamic memory of the past (future) for the purpose of predicting the joint future (past). It follows in particular that $\mathbb{X}$ is Markovian (i. e. $\mathbf{X}^- \perp \mathbf{X}^+ \mid \mathbf{X}$) and

any basis vector $x := [x_1, x_2, \ldots, x_n]'$ in a (finite-dimensional) Markovian splitting subspace $\mathbf{X}$ generates a stationary Markov process $x(t) := U^t x$, $t \in \mathbb{Z}$ which, as we shall see in a moment, serves as a *state* for the process y.

The subspace $\mathbf{X}$ is called *proper*, or *p.n.d.* if

$$\cap_t \mathbf{Y}_t^- \vee \mathbf{X}_t^- = \{0\}, \text{ and } \cap_t \mathbf{H}(y)_t^+ \vee \mathbf{X}_t^+ = \{0\}.$$

Obviously for the existence of proper splitting subspaces y must also be purely nondeterministic [57]. Properness is, by the Wold decomposition theorem, equivalent to the existence of two vector white noise processes w and $\bar{w}$ such that,

$$\mathbf{H}(y)^- \vee \mathbf{X}^- = \mathbf{H}^-(w), \quad \mathbf{H}(y)^+ \vee \mathbf{X}^+ = \mathbf{H}^+(\bar{w}).$$

If $\mathbf{X}$ is proper, every Markov process $x(t) := U^t x$ is purely nondeterministic.

The subspaces

$$\mathbf{S} := \mathbf{H}(y)^- \vee \mathbf{X}^- \text{ and } \bar{\mathbf{S}} := \mathbf{Y}^+ \vee \mathbf{X}^+ \tag{24}$$

associated to a Markovian Splitting subspace $\mathbf{X}$, play an important role in the geometric theory of stochastic systems. They are called the *scattering pair* of $\mathbf{X}$ as they can be seen to form an incoming–outgoing pair in the sense of Lax–Phillips Scattering Theory [29].

Definition 2 Given a stationary Hilbert space $(\mathbf{H}, U)$ containing $\mathbf{H}(y)$, a *scattering pair* for the process y is a pair of subspaces $(\mathbf{S}, \bar{\mathbf{S}})$ satisfying the following conditions for a stationary process,

1. $U^*\mathbf{S} \subset \mathbf{S}$ and $U\bar{\mathbf{S}} \subset \bar{\mathbf{S}}$, i. e. $\mathbf{S}$ and $\bar{\mathbf{S}}$ are invariant for the left and right shift semigroups (this means that $\mathbf{S}_t$ is increasing and $\bar{\mathbf{S}}_t$ is decreasing with time).
2. $\mathbf{S} \vee \bar{\mathbf{S}} = \mathbf{H}$
3. $\mathbf{S} \supset \mathbf{H}(y)^-$ and $\bar{\mathbf{S}} \supset \mathbf{H}(y)^+$
4. $\mathbf{S}^\perp \subset \bar{\mathbf{S}}$ or, equivalently, $\bar{\mathbf{S}}^\perp \subset \mathbf{S}$

The following representation Theorem provides the link between Markovian splitting subspaces and scattering pairs.

Theorem 1 ([35]) *The intersection*

$$\mathbf{X} = \mathbf{S} \cap \bar{\mathbf{S}} \tag{25}$$

of any scattering pair of subspaces of $\mathbf{H}$ *is a Markovian splitting subspace. Conversely every Markovian splitting subspace can be represented as the intersection of a scattering pair. The correspondence* $\mathbf{X} \leftrightarrow (\mathbf{S}, \bar{\mathbf{S}})$ *is one-to-one, the scattering pair corresponding to* $\mathbf{X}$ *being given by*

$$\mathbf{S} = \mathbf{H}(y)^- \vee \mathbf{X}^- \qquad \bar{\mathbf{S}} = \mathbf{H}(y)^+ \vee \mathbf{X}^+. \tag{26}$$

The process of forming scattering pairs associated to $\mathbf{X}$ should be thought of as an "extension" of the past and future spaces of y. The rationale for this extension is that scattering pairs have an extremely simple splitting geometry due to the fact that

$$\mathbf{S} \perp \bar{\mathbf{S}} \mid \mathbf{S} \cap \bar{\mathbf{S}} \tag{27}$$

equivalent to

$$\mathbf{S} \vee \bar{\mathbf{S}} = \bar{\mathbf{S}}^\perp \oplus (\mathbf{S} \cap \bar{\mathbf{S}}) \oplus \mathbf{S}^\perp \tag{28}$$

which is called *perpendicular intersection*. It is easy to show that Property (4) in the definition of a scattering pair is actually equivalent to perpendicular intersection. This property of conditional orthogonality given the intersection can also be seen as a natural generalization of the Markov property. Indeed for a Markovian family of subspaces $\mathbf{X} = \mathbf{X}^- \cap \mathbf{X}^+$ and $\mathbf{S} = \mathbf{X}^-$, $\bar{\mathbf{S}} = \mathbf{X}^+$ intersect perpendicularly with intersection $\mathbf{X}$.

Note that $\mathbf{A} \perp \mathbf{B} \mid \mathbf{X} \Rightarrow \mathbf{A} \cap \mathbf{B} \subset \mathbf{X}$ but the inclusion of the intersection in the splitting subspace $\mathbf{X}$ is only *proper* in general. For perpendicularly intersecting subspaces, the intersection is actually the *unique minimal subspace* making them conditionally orthogonal.

Theorem 1 is the fundamental device for the construction and classification of Markovian splitting subspaces.

Denote by $\mathbf{W}_t$, $\bar{\mathbf{W}}_t$ the (wandering) subspaces spanned by the components, at time t, of the generating noises $w(t)$ and $\bar{w}(t)$, of the scattering pair of $\mathbf{X}$. Since

$$\mathbf{S}_{t+1} = \mathbf{S}_t \oplus \mathbf{W}_t, \tag{29}$$

we can write,

$$\mathbf{X}_{t+1} = \mathbf{S}_{t+1} \cap \bar{\mathbf{S}}_{t+1} \subset (\mathbf{S}_t \cap \bar{\mathbf{S}}_{t+1}) \oplus (\mathbf{W}_t \cap \bar{\mathbf{S}}_{t+1}) \tag{30}$$

Since $\bar{\mathbf{S}}_t$ is decreasing in time, we have $\mathbf{S}_t \cap \bar{\mathbf{S}}_{t+1} \subset \mathbf{X}_t$ and by projecting the shifted basis $Ux(t) := x(t+1)$, onto the last orthogonal direct sum above, the time evolution of any basis vector $x(t) := [x_1(t), x_2(t), \ldots, x_n(t)]'$ in $\mathbf{X}_t$, is described by a linear equation of the type

$$x(t+1) = Ax(t) + Bw(t).$$

It is also easy to see that, by the p.n.d. property, A must have all its eigenvalues strictly inside of the unit circle. Naturally, by decomposing instead $\bar{\mathbf{S}}_{t-1} = \bar{\mathbf{S}}_t \oplus \bar{\mathbf{W}}_t$ one could have obtained a *backward difference equation* model for the Markov process x, driven by the backward generating noise $\bar{w}$.

To complete the representation, note that by definition of the past space, $y(t) \in (\mathbf{S}_{t+1} \cap \bar{\mathbf{S}}_t)$. Inserting the decomposition (29) and projecting $y(t)$, leads to a state-output

equation of the form

$$y(t) = Cx(t) + Dw(t)\,.$$

Here one could also obtain a state-output equation driven by the backward noise $\bar{w}$, the same noise driving the backward state model obtained before.

As we have just seen, any basis in a Markovian splitting subspaces produces a stochastic realization of y. It is easy to reverse the implication. In fact the following fundamental characterization holds.

Theorem 2 ([33,37]) *The state space* $\mathbf{X} = \operatorname{span}\{x_1(0), x_2(0), \ldots, x_n(0)\}$ *of any stochastic realization (8a) is a Markovian splitting subspace for the process y.*

Conversely, given a finite-dimensional Markovian splitting subspace $\mathbf{X}$*, to any choice of basis* $x(0) = [\,x_1(0), x_2(0), \ldots, x_n(0)\,]'$ *in* $\mathbf{X}$ *there corresponds a stochastic realization of y of the type (8a).*

Once a basis in $\mathbf{X}$ is available, there are obvious formulas to compute the parameters $(A, C, \bar{C})$ of the corresponding stochastic realization, namely

$$A = \mathbb{E}x(t+1)x(t)'\,P^{-1} \tag{31}$$

$$C = \mathbb{E}y(t)x(t)'\,P^{-1} \tag{32}$$

$$\bar{C} = \mathbb{E}y(t-1)x(t)' \tag{33}$$

where $P = \mathbb{E}x(t)x(t)'$ is the state covariance matrix (i. e. the Gramian matrix of the basis). The matrices B and D, however, are related to the (unobservable) generating white noise w and require the solution of the LMI. This abstract procedure which permits us to compute the parameters of a stochastic realization once the state has been constructed, can be rendered quite concrete and forms the conceptual basis of subspace identification.

Stochastic realizations are called *internal* when $\mathbf{H} = \mathbf{H}(y)$, i. e. the state space is built from the Hilbert space made just of the linear statistics of the process y. For identification the only realizations of interest are the internal ones.

A central problem of geometric realization theory is to construct and to classify all minimal state spaces, i. e. the minimal Markovian splitting subspaces for the process y.

The obvious ordering of subspaces of $\mathbf{H}$ by inclusion, induces an ordering on the family of Markovian splitting subspaces. The notion of minimality is most naturally defined with respect to this ordering. Note that this definition is independent of assumptions of finite-dimensionality and applies also to infinite-dimensional Markovian splitting subspaces, i. e. to situations where comparing dimension would not make much sense.

Definition 3 A Markovian splitting subspace is *minimal* if it does not contain (properly) other Markovian splitting subspaces.

The study of minimality forms an elegant chapter of stochastic system theory. There are several known geometric and algebraic characterizations of minimality of splitting subspaces and of the corresponding stochastic state-space realizations. Since, however, the discussion of this topic would take us too far from the main theme of the paper we shall refer the reader to the literature [35,37].

Contrary to the deterministic situation minimal Markovian splitting subspaces are *nonunique*. Two very important examples are the *forward and backward predictor spaces* (at time zero):

$$\mathbf{X}_- := \mathrm{E}^{\mathbf{H}^-}\mathbf{H}^+ \quad \mathbf{X}_+ := \mathrm{E}^{\mathbf{H}^+}\mathbf{H}^- \tag{34}$$

for which we have the following characterization [35].

Proposition 3 *The subspaces* $\mathbf{X}_-$ *and* $\mathbf{X}_+$ *are minimal Markovian splitting subspaces contained in the past* $\mathbf{H}^-$*, and, respectively, in the future* $\mathbf{H}^+$*, of the process y.*

A basis in the forward predictor space $\mathbf{X}_-$ originates a stationary state-space model in which the state variables are linear functionals of the past history of the process y, i. e. $x(t) \in \mathbf{H}_t^-(y)$. In other words the state coincides with its best estimate (the orthogonal projection), $E[x(t) \mid \mathbf{H}_t^-(y)]$ given the past of y. It follows that the dynamical equations (8a) describe in this case a steady-state Kalman predictor and the input white noise $w = w_-$ is the steady state *innovation process* of y.

Before closing this section we should remark that a scattering picture emerges also in some of the early papers on the "unitary dilation" approach to statistical mechanics, e. g. [31,32]. It is remarkable that the abstract scattering picture described in this section makes contact with this literature. The dilation approach, however, deals (in our language) with the reverse problem of describing the stationary shift group starting from (Markov) processes which are described by a Langevin-type equation. This is a curious and (in our view) somewhat unnatural point of view which apparently does not lead to capture the phenomenon of *nonuniqueness of the macroscopic description* of the observables. Whether this has any physical relevance is however not clear to us.

Dynamical System Identification

System identification is the problem of describing an observed time series (e. g. a sequence of real numbers representing stock prices, temperatures in a room, sampled

audio or video signals, etc.) by a linear dynamic model, in particular by a state-space model of the type (8a). It is a widely accepted viewpoint that system identification should be regarded as a *statistical* problem [40,42]. This is so because the experimental data which one wants to model very often come from complicated, often unknown, nonlinear physical phenomena subject to unknown interactions with the environment, and it would in general be impossible to attempt a "physical" modelization from first principles. Consequently the models describing the data should be stochastic. The general goal is to get accurate models which are as simple as possible. The problem can then be approached from (at least) two conceptually different viewpoints.

Identification by Parametric Optimization

This is the mainstream "optimization" approach, based on the principle of minimizing a suitable measure of discrepancy between the observed data and the data predicted by a probability law underlying a certain chosen model class. Well-known examples of distance functions are the *likelihood function*, or the average squared *prediction-error* of the observed data corresponding to a particular model. Except for rather trivial model classes, these distance functions depend nonlinearly on the model parameters and the minimization can only be done numerically. Hence the optimization approach leads to iterative algorithms in the parameter space, say in the space of minimal (A, B, C, D) matrix quadruples which parametrize a chosen model class. In spite of the fact that this has been almost the only accepted paradigm in system identification for many decades, [42,61], this approach has several drawbacks, including the need of unique parametrization of multivariable models, the fact that the cost function generally has complicated local minima which, for moderate or large dimension of the model are very difficult to detect, and the inherent insensitivity of the cost to variations of some parameters and the corresponding ill-posedness of the estimation problem.

It seems that these limitations are a consequence of the intrinsically "blind" philosophy which underlies setting the problem as a parameter optimization problem. Almost all problems of controller and/or estimator design could in principle be formulated just as parametric optimization after choosing a certain family of parametric structures for the controller or the estimator. Pushing this philosophy to the extreme, in principle one would not need the maximum principle, Kalman filtering, H^∞ theory, etc. (in fact one would not need system and control theory altogether), since everything could be reduced to a nonlinear programming problem in a suitable space of controller or estimator parameters. It is dubious however, whether any real progress in the field of control and estimation could have occurred by following this type of paradigm.

Identification by Stochastic Realization

This is commonly referred to as "subspace methods" identification or "subspace identification" *tout-court*. Subspace identification is essentially a sample version of stochastic realization. Under a reasonable assumption of second-order ergodicity of the process y which has produced the observed data, one can set up a "sample" counterpart of the abstract Hilbert space operations of stochastic realization theory. In this setting the geometric operations of stochastic realization can be translated into statistical operations on the observed data. One obtains a statistical theory of model building which is in a sense perfectly isomorphic to the abstract probabilistic realization theory. The main point is the idea of constructing first a (sample) *state space* for the process y, starting from certain vector spaces, called the *future* and *past* spaces associated with the observations. According to the recipe of (34) the state space is constructed by orthogonal projection of the future onto the past, as seen in the previous section. Successively, a well conditioned basis is chosen in the state space e. g. by principal components (canonical correlation) analysis. Once a "robust" basis; i. e. a state vector, is chosen, the parameters A, C, $\bar{C}$ of the model are computed by formulas analogous to (31),(32),(33). The final step of finding the B and D matrices requires solving a Riccati equation.

By similarity with realization theory, one recognizes that the inherent nonlinearity of model identification has to do with the quadratic nature of the spectral factorization problem. As we have seen, spectral factorization for state-space models involves the solution of a Riccati equation (or more generally of a linear matrix inequality), a problem which has been the object of intensive theoretical and numerical studies in the past three decades. The nonlinearity of the stochastic system identification problem is hence of a well-known and well-understood kind and is much better dealt with by the explicit methods of Riccati solution developed in system theory rather than by nonspecific optimization algorithms. This has allowed a systematic introduction of very reliable and efficient numerical linear algebra tools to solve the problem. The "subspace" approach has been suggested more or less implicitly by several authors in the past [6,13,15,44] but is first clearly presented in [63]. Various extensions to cover identification of systems with inputs have appeared since this paper and there

is now a large literature on the subject. For references we refer the reader to the recently published book [27].

Since the optimization approach is well covered in the literature, we shall just analyze subspace identification below.

The Hilbert Space of a Time Series

The crucial conceptual step in subspace identification is a proper identification of the Hilbert space of random data in which modeling takes place. We shall briefly illustrate this point, following [38].

We shall initially consider an idealized situation in which the observed data (time series)

$$\{y_0, y_1, \ldots, y_t, \ldots\} \qquad y_t \in \mathbb{R}^m \tag{35}$$

is infinitely long.

We shall assume that the limit for $N \to \infty$ and for any $\tau \geq 0$, of the time averages

$$\frac{1}{N+1}\sum_{t=0}^{N} y_{t+\tau} y_t' \tag{36}$$

exists. It can be shown that this limit is a function $\tau \to \Lambda_\tau$ of *positive type*. In the continuous-time setting, functions admitting an "ergodic" limit of the sample correlation function (36), have been studied in depth by Wiener in his famous work on Generalized Harmonic Analysis. Although a systematic translation of the continuous-time results of Wiener into discrete-time seems not to be available in the literature, it is quite evident that a totally analogous set of results holds also for discrete-time signals. In particular it is rather easy to show, by adapting Wiener's proof for continuous time, that the limits of the time averages (36) form a matrix function Λ of positive type, in other words a *bona-fide stationary covariance matrix*, see [67,68] which can then be identified with the "true" covariance function of an underlying stochastic process.

The assumption actually implies that

$$\frac{1}{N+1}\sum_{t=t_0}^{N+t_0} y_{t+\tau} y_t' \to \Lambda_\tau \tag{37}$$

for arbitrary t_0, which can be read as a kind of "statistical regularity" of the (future) data. The ergodicity (36) is of course unverifiable in practice as it says something about data which have not been observed yet. Some assumption of this sort about the mechanism generating future data seems however to be necessary to even formulate the identification problem. We shall call Λ the *true covariance* of the time series $\{y_t\}$.

Now, for each $t \in \mathbb{Z}_+$ define the $m \times \infty$ matrices

$$\mathbb{y}(t) := [y_t, y_{t+1}, y_{t+2}, \ldots] \tag{38}$$

and consider the sequence $\mathbb{y} := \{\mathbb{y}(t) \mid t \in \mathbb{Z}_+\}$. We shall make this sequence of semi-infinite matrices into an object isomorphic to the stationary processes y.

Define the vector space $\mathbb{Y}$ of scalar semi-infinite real sequences obtained as finite linear combinations of the rows of y,

$$\mathbb{Y} := \left\{\sum a_k' \mathbb{y}(t_k); \quad a_k \in \mathbb{R}^m,\ t_k \in \mathbb{Z}_+\right\} \tag{39}$$

This space can be naturally made into an inner product space in the following way.

First, define the bilinear form $\langle \cdot, \cdot \rangle$ on the generators by letting

$$\begin{aligned}\langle a'\mathbb{y}(k),\ b'\mathbb{y}(j)\rangle &:= \lim_{N\to\infty} \frac{1}{N+1}\sum_{t=0}^{N} a' y_{t+k} y_{t+j}' b \\ &= a' \Lambda_{k-j} b, \end{aligned} \tag{40}$$

for $a,\ b \in \mathbb{R}^m$, and then extend it by linearity to all elements of $\mathbb{Y}$.

Let $\mathbf{a} := \{a_k,\ k \in \mathbb{Z}_+\}$ be a sequence of vectors $a_k \in \mathbb{R}^m$, with compact support in $\mathbb{Z}_+$, and let $\mathbb{a}' := \{a_k'\}$. A generic element ξ of the vector space $\mathbb{Y}$ can be represented as

$$\xi = \sum_k a_k' \mathbb{y}(k) = \mathbf{a}'\mathbb{y}$$

Let us assume that the infinite block-symmetric Toeplitz matrix

$$T = \begin{bmatrix} \Lambda_0 & \Lambda_1 & \ldots & \Lambda_k & \ldots \\ \Lambda_1' & \Lambda_0 & \Lambda_1 & \ldots & \ldots \\ \vdots & & \ddots & & \vdots \\ \Lambda_k' & & & \Lambda_0 & \\ \ldots & & & & \end{bmatrix} \tag{41}$$

constructed from the "true" covariance sequence $\{\Lambda_0, \Lambda_1, \ldots, \Lambda_k, \ldots\}$ of the data, is positive definite. Since the bilinear form (40) on $\mathbb{Y}$ can be represented by the quadratic form

$$\langle \xi, \eta \rangle = \langle \mathbf{a}'\mathbb{y}, \mathbf{b}'\mathbb{y}\rangle = \sum_{kj} a_k' \Lambda_{k-j} b_j = \mathbf{a}' T \mathbf{b}$$

it can be seen that the bilinear form is nondegenerate (unless $\Lambda = 0$ identically) and defines a bona-fide inner prod-

uct. As usual one identifies elements whose difference has norm zero (this means $\langle \xi, \xi \rangle = 0 \Leftrightarrow \xi = 0$).

By closing the vector space $\mathbb{Y}$ with respect to the norm induced by the inner product (40), one obtains a real Hilbert space of semi-infinite sequences which hereafter we shall still denote $\mathbb{Y}$. This is the basic data space on which our models will be defined.

The *shift operator* U operates on the family of semi-infinite matrices (38), by the rule

$$Ua'\mathbb{y}(t) = a'\mathbb{y}(t+1) \qquad t \in \mathbb{Z}\,, \quad a \in \mathbb{R}^m$$

It is easy to see that U is a linear map which is isometric with respect to the inner product (40) and can be extended by linearity to all of $\mathbb{Y}$. This Hilbert space framework for time series was introduced in [38]. It is shown in this reference that the "stationary Hilbert space" $(\mathbb{Y}, U)$ is isomorphic to the standard stochastic Hilbert space setup in the L^2 theory of second-order stationary random processes. By virtue of this isomorphism one can formally think of the observed time series (35) as an ergodic sample path of some Gaussian stationary stochastic process $\mathbf{y}$ defined on a true probability space, having covariance matrices equal to the limit Λ of the sum (36) as $N \to \infty$.

Linear functions and operators on the "tail sequences" $\mathbb{y}(t)$ correspond to the same linear functions and operators on the random variables of the process y. In particular the second-order moments of y can be computed in terms of the tail sequences $\mathbb{y}$, by substituting expectations with ergodic limits of the type (40). Since second-order properties are all what matters in our setting, one may even regard the tail sequence $\mathbb{y}$ of (38) as being the *same object* as the underlying stochastic process y. The usual probabilistic language can be adopted in the present setting provided one identifies real random variables as semi-infinite strings of numbers having the "ergodic property" described at the beginning of this section. The inner product of two semi-infinite strings ξ and η in $\mathbb{Y}$ corresponds in particular to the expectation $\mathbb{E}\{\xi\eta\}$; i. e.

$$\langle \xi, \eta \rangle = \mathbb{E}\{\xi\eta\}. \tag{42}$$

This unification of language permits us to carry over in its entirety the geometric theory of stochastic realization derived in the abstract L^2 setting to the time series framework. In particular, the *past* and *future* subspaces of the "processes" $\mathbb{y}$ at time t are defined as the closure of the linear vector spaces spanned by the relative past or future "random variables" $\mathbb{y}(t)$, in the metric of the Hilbert space $\mathbb{Y}$. The only difference to keep in mind here is the different interpretation that representation formulas like (8a) have in this context. The equalities involved in the representation

$$\begin{cases} x(t+1) = Ax(t) + Bw(t) \\ y(t) = Cx(t) + Dw(t) \end{cases} \tag{43}$$

are now to be understood in the sense of equalities of elements of $\mathbb{Y}$, i. e. as asymptotic equality of sequences in the sense of Cesàro limits. In particular the equality signs in the model (43) do not necessarily imply that the same relations hold for the sample values y_t, x_t, w_t at a particular instant of time t. This is in a certain sense similar to the "with probability one" interpretation of the equality sign to be given to the model (43) in case the variables are bona-fide random variables in a probability space.

Consider the orthogonal projection $\mathbb{E}[\,\xi \mid \mathbf{X}]$ of a (row) random variable ξ onto a subspace $\mathbf{X}$ of the space $\mathbb{Y}$. In the probabilistic L^2 setting this has the well-known interpretation of wide-sense conditional expectation given the random variables in $\mathbf{X}$ (a true conditional expectation, in the case of Gaussian distributions). In this setting the projection operator has an immediate and useful *statistical* meaning.

Assume for simplicity that $\mathbf{X}$ is given as the rowspace of some matrix of generators X, then the projection $\mathbb{E}[\,\xi \mid \mathbf{X}]$ has exactly the familiar aspect of the least squares formula expressing the best approximation of the vector ξ as a linear combination of the rows of X. For, writing $\mathbb{E}[\,\xi \mid X\,]$ to denote the projection expressed (perhaps nonuniquely) in terms of the rows of X, the classical linear "conditional expectation" formula leads to

$$\mathbb{E}[\,\xi \mid X\,] = \xi X'[XX']^{\sharp} X\,, \tag{44}$$

which is the universally known "least squares" formula of statistics. The pseudoinverse $\sharp$ can be substituted by a true inverse in case the rows of X are linearly independent.

The Question of Statistical Efficiency

There are questions about the statistical significance (what are the uncertainty bounds on the parameters and on the estimated transfer functions etc.) which are easily and naturally addressed in the optimization framework but harder to be addressed in the subspace/realization approach to identification. The main difference with the mainstream statistical approach is that the estimation of $(A, C, \bar{C})$ is not done *directly* by optimizing a likelihood or other distance functions but by just matching second-order moments, i. e. by solving the Eqs. (46).

This way of proceeding can be seen as an instance of *estimation by the method of moments* described in the statistical textbooks e. g. p. 497 in [12], a very old idea used extensively by K. Pearson in the beginning of the last century. The underlying estimation principle is that the parameter estimates should match exactly the sample second-order moments and is close in spirit to the wide-sense setting that we are working in. It does not involve optimality or minimal distance criteria between the "true" and the model distributions. For this reason it is generally claimed in the literature that one should expect better results (in the sense of smaller asymptotic variance of the estimates) by optimization methods. However, a proof that subspace methods can be efficient (under certain conditions which will be too long to report here) has appeared recently [8].

Usable expressions for the asymptotic covariance of the parameter estimates are given in [9]. The arguments in the derivations are, however, far from trivial.

Subspace Algorithms

Most standard algorithms for subspace identification can be shown to be essentially equivalent to the following three-step procedure (see e. g. [6,38]),

1. The first step is the estimation of a *finite* sequence of sample covariance matrices

$$\{\hat{\Lambda}_0, \hat{\Lambda}_1, \ldots, \hat{\Lambda}_\nu\} \tag{45}$$

from the observed data. Since the data are necessarily finite, ν must be also finite (and in general small compared to the data length).

2. The second step is identification of a rational model for the covariance sequence. This is a *minimal partial realization* (also called "rational extension") problem. Given a finite set of experimental covariance data one is asked to find a minimal value of n and a *minimal* triplet of matrices $(A, C, \bar{C})$, of dimensions $n \times n$, $m \times n$ and $m \times n$ respectively, such that

$$\hat{\Lambda}_k = CA^{k-1}\bar{C}' \qquad k = 1, \ldots, \nu. \tag{46}$$

Recall that $(A, C, \bar{C}')$ is a *minimal triplet*, in the sense of deterministic linear system theory, if the pairs of matrices (A, C) and $(A, \bar{C}')$ are an observable and, respectively, a reachable pair, see e. g. [24]. There are well-known algorithms for computing minimal partial realizations (see e. g. [72] for an efficient algorithm) thereby producing "estimates" of the parameters $(A, C, \bar{C})$ of a minimal realization of a rational spectral density matrix of the process.

3. The third step is to compute, starting from the realization of the rational spectrum estimated in step two, a stationary state-space model. Typically one is interested in the innovation model. This is accomplished by computing the minimal solution P_- of the Linear Matrix Inequality (16), or, equivalently, the minimal solution of the associated algebraic Riccati equation.

A warning is in order concerning the practical use of these algorithms in that some nontrivial mathematical questions related to positivity of the estimated spectrum are often overlooked in the implementation. This issue is thoroughly discussed in [38] and here we shall just give a short summary.

In determining a minimal triplet $(A, C, \bar{C})$ interpolating the partial sequence (45) so that $CA^{k-1}\bar{C}' = \hat{\Lambda}_k \quad k = 1, 2, \ldots, \nu$, we also completely determine the infinite sequence

$$\{\hat{\Lambda}_0, \hat{\Lambda}_1, \hat{\Lambda}_2, \hat{\Lambda}_3, \ldots\} \tag{47}$$

by setting $\hat{\Lambda}_k = CA^{k-1}\bar{C}'$ for $k = \nu + 1, \nu + 2, \ldots$. This sequence is called a *minimal rational extension* of the finite sequence (45). The attribute "rational" is due to the fact that

$$\begin{aligned}\hat{\Phi}_+(z) &:= \frac{1}{2}\hat{\Lambda}_0 + \hat{\Lambda}_1 z^{-1} + \hat{\Lambda}_2 z^{-2} + \ldots \\ &= \frac{1}{2}\hat{\Lambda}_0 + C(zI - A)^{-1}\bar{C}'\end{aligned} \tag{48}$$

is a rational function. In order for (47) to be a bona fide covariance sequence, however, it is necessary that the *infinite* block-Toeplitz matrix obtained by extending the finite matrix

$$T_\nu = \begin{bmatrix} \hat{\Lambda}_0 & \hat{\Lambda}_1 & \hat{\Lambda}_2 & \cdots & \hat{\Lambda}_\nu \\ \hat{\Lambda}_1' & \hat{\Lambda}_0 & \hat{\Lambda}_1 & \cdots & \hat{\Lambda}_{\nu_1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{\Lambda}_\nu' & \hat{\Lambda}_{\nu-1}' & \hat{\Lambda}_{\nu-2} & \ldots & \hat{\Lambda}_0 \end{bmatrix} \tag{49}$$

using the rational extension sequence (47), should be nonnegative definite. Equivalently, the spectral density function corresponding to (47)

$$\hat{\Phi}(z) = \hat{\Lambda}_0 + \sum_{k=1}^{\infty} \hat{\Lambda}_k (z^k + z^{-k}) = \hat{\Phi}_+(z) + \hat{\Phi}_+(z^{-1})' \tag{50}$$

should be nonnegative definite on the unit circle. This is equivalent to the function $\hat{\Phi}_+(z)$ being *positive real*. Consequently, the partial realization needs to be done subject to the extra constraint of positivity.

The constraint of positivity is a rather tricky one and in some of the methods described in the literature it is disregarded. For this reason some subspace algorithms may fail to provide a positive extension and hence may lead to data $(A, C, \bar{C})$ for which there are no solutions of the LMI and hence to totally inconsistent results.

It is important to appreciate the fact that the problem of positivity of the extension has little to do with the noise or sample variability of the covariance data and is present equally well for any *finite* covariance sequence extracted from a true (infinitely long) rational covariance sequence. For there is no guarantee that, even in this idealized situation, the order of a minimal rational extension (47) of the finite covariance subsequence would be sufficiently large to equal the order of the infinite sequence and hence to generate a positive extension. A minimal partial realization may well fail to be positive because its order is too low to guarantee positivity. It is shown in [38] that there are relatively "large" sets of data (45) for which this happens and the matrix A may even fail to be stable [7].

Future Directions

Because of the pervasive presence of fast digital computers in modern society, we are witnessing a sharp change of paradigms in the analysis, design, prediction and control of both man made and natural (say biological, economical etc) systems. Because of the tremendous computing capabilities at our disposal nowadays, stochastic modeling, realization and identification have become an essential ingredient of modern applied sciences. In particular modeling and simulation of complex technological, biological, economic, and environmental systems is becoming more and more essential for understanding the behavior and for prediction and control of these systems. New problems continually arise and much work remains to be done.

We just mention here problems where there is a need of modeling the effect of *input* or *decision variables*, possibly in the presence of feedback. Such *Input–Output* models have important applications in many applied areas such as control of industrial processes, econometrics, etc. There is a large body of results concerning stochastic realization and subspace identification of systems with inputs which we could not discuss in this article, see [10,11,26,27,53,54,55,64,65]. The area is important for applications and still needs research. Modeling and realization of finite state processes and of certain classes of highly non-Gaussian signals is needed in diverse applications such as source coding, image processing, and computer vision.

Bibliography

Primary Literature

1. Akaike H (1975) Markovian representation of stochastic processes by canonical variables. SIAM J Control 13:162–173
2. Akaike H (1974) Stochastic Theory of Minimal Realization. IEEE Trans Autom Control AC-19(6):667–674
3. Anderson BDO (1969) The inverse problem of Stationary Covariance Generation. J Stat Phys 1(1):133–147
4. Anderson BDO (1967) An algebraic solution to the spectral factorization problem. IEEE Trans Autom Control AC-12:410–414
5. Anderson BDO (1999) The realization problem for Hidden Markov Models. Math Control Signals Syst 12:80–120
6. Aoki M (1990) State Space Modeling of Time Series, 2nd edn. Springer, New York
7. Byrnes CI, Lindquist A (1982) The stability and instability of partial realizations. Syst Control Lett 2:2301–2312
8. Bauer D (2005) Comparing the CCA Subspace Method to Pseudo Maximum Likelihood Methods in the case of No Exogenous Inputs. J Time Ser Anal 26:631–668
9. Chiuso A, Picci G (2004) The Asymptotic Variance of Subspace Estimates. J Econom 118(1–2):257–291
10. Chiuso A, Picci G (2004) On the Ill-conditioning of subspace identification with inputs. Automatica 40(4):575–589
11. Chiuso A, Picci G (2005) Consistency Analysis of some Closed-loop Subspace Identification Methods. Automatica 41:377–391, special issue on System Identification
12. Cramèr H (1949) Mathematical Methods of Statistics. Princeton University Press, Princeton
13. Desai UB, Pal D, Kirkpatrick RD (1985) A Realization Approach to Stochastic Model Reduction. Int J Control 42:821–838
14. Doob JL (1953) Stochastic Processes. Wiley, New York
15. Faurre P (1969) Identification par minimisation d'une representation Markovienne de processus aleatoires. Symposium on Optimization, Nice 1969. Lecture Notes in Mathematics, vol 132. Springer, New York
16. Faurre P, Chataigner P (1971) Identification en temp reel et en temp differee par factorisation de matrices de Hankel. French–Swedish colloquium on process control, IRIA Roquencourt
17. Faurre P (1973) Realisations Markovienne de processus stationnaires. Report de Recherche no 13, INRIA, Roquencourt
18. Ferrante A, Picci G, Pinzoni S (2002) Silverman algorithm and the structure of discrete-time stochastic systems. Linear Algebra Appl (special issue on systems and control) 351–352:219–242
19. Finesso L, Spreij PJC (2002) Approximate realization of finite Hidden Markov Chains. Proceedings of the 2002 IEEE Information Theory Workshop, Bangalore, pp 90–93
20. Ford GW, Kac M, Mazur P (1965) Statistical mechanics of assemblies of coupled oscillators. J Math Phys 6:504–515
21. Kalman RE (1963) On a new Criterion of Linear Passive Systems. In: Proc. of the First Allerton Conference. University of Illinois, Urbana Ill, Nov 1963, pp 456–470
22. Kalman RE (1963) Lyapunov Functions for the problem of Lur'e in Automatic Control. Proc Natl Acad Sci 49:201–205
23. Kalman RE (1965) Linear stochastic filtering: reappraisal and outlook. In: Proc. Symposium on Syatem Theory, Politechnic Institute of Brooklin, pp 197–205
24. Kalman RE, Falb PL, Arbib MA (1969) Topics in Mathematical Systems Theory. McGraw–Hill, New York

25. Karatzas I, Shreve S (1987) Brownian motion and stochastic calculus. Springer, New York
26. Katayama T, Picci G (1999) Realization of Stochastic Systems with Exogenous Inputs and Subspace Identification Methods. Automatica 35(10):1635–1652
27. Katayama T (2005) Subspace methods for System Identification. Springer, New York
28. Kintchine A (1934) Korrelationstheorie Stationäre Stochastischen Prozessen. Math Annalen 109:604–615
29. Lax PD, Phillips RS (1967) Scattering Theory. Academic Press, New York
30. Levy P (1948) Processus stochastiques et Mouvement Brownien. Gauthier–Villars, Paris
31. Lewis JT, Thomas LC (1975) How to make a heat bath. In: Arthurs AM (ed) Functional Integration and its Applications. Clarendon, Oxford
32. Lewis JT, Maassen H (1984) Hamiltonian models of classical and quantum stochastic processes. In: Accardi L, Frigerio A, Gorini V (eds) Quantum Probability and Applications to the Quantum Theory of Irreversible processes. Springer Lecture Notes in Mathematics, vol 1055. Springer, New York, pp 245–276
33. Lindquist A, Picci G, Ruckebusch G (1979) On minimal splitting subspaces and Markovian representation. Math Syst Theory 12:271–279
34. Lindquist A, Picci G (1979) On the stochastic realization problem. SIAM J Control Optim 17:365–389
35. Lindquist A, Picci G (1985) Realization theory for multivariate stationary Gaussian processes. SIAM J Control Optim 23:809–857
36. Lindquist A,Pavon M (1984) On the structure of state space models of discrete-time vector processes. IEEE Trans Autom Control AC-29:418–432
37. Lindquist A, Picci G (1991) A geometric approach to modelling and estimation of linear stochastic systems. J Math Syst Estim Control 1:241–333
38. Lindquist A, Picci G (1996) Canonical correlation analysis approximate covariance extension and identification of stationary time series. Automatica 32:709–733
39. Lindquist A, Michaletzky G, Picci G (1995) Zeros of Spectral Factors, the Geometry of Splitting Subspaces, and the Algebraic Riccati Inequality. SIAM J Control Optim 33:365–401
40. Lindquist A, Picci G (1996) Geometric Methods for State Space Identification. In: Identification, Adaptation, Learning. (Lectures given at the NATO-ASI School, From Identification to Learning, Como, Italy, Aug 1994. Springer
41. Lindquist A, Picci G, Linear Stochastic Systems: A Geometric Approach. (in preparation)
42. Ljung L (1987) System Identification – Theory for the User. Prentice–Hall, New York
43. Martin-Löf A (1979) Statistical Mechanics and the foundations of Thermodynamics. Lecture Notes in Physics, vol 101. Springer, New York
44. Moonen M, De Moor B, Vanderberghe L, Vandewalle J (1989) On- and Off-Line Identification of Linear State–Space Models. Int J Control 49:219–232
45. Taylor TJ, Pavon M (1988) A solution of the nonlinear stochastic realization problem. Syst Control Lett 11:117–121
46. Picci G (1978) On the internal structure of finite-state stochastic processes. In: Mohler R, Ruberti A (eds) Recent developments in Variable Structure Systems. Springer Lecture Notes in Economics and Mathematical Systems. vol 162. Springer, New York, pp 288–304
47. Picci G, Pinzoni S (1994) Acausal Models and Balanced realizations of stationary processes. Linear Algebra Appl (special issue on Systems Theory) 205–206:957–1003
48. Picci G (1986) Application of stochastic realization theory to a fundamental problem of statistical physics. In: Byrnes CI, Lindquist A (eds) Modeling, Identification and Robust Control. North Holland, Amsterdam
49. Picci G, Taylor TJ (1990) Stochastic aggregation of linear Hamiltonian systems with microcanonical distribution. In: Kaashoek MA, van Schuppen JH, Ran ACM (eds) Realization and modeling in System Theory. Birkhäuser, Boston, pp 513–520
50. Picci G (1992) Markovian representation of linear Hamiltonian systems. In: Guerra F, Loffredo MI, Marchioro C (eds) Probabilistic Methods in Mathematical Physics. World Sciedntific, Singapore, pp 358–373
51. Picci G (1991) Stochastic realization theory. In: Antoulas A (ed) Mathematical System Theory. Springer, New York
52. Picci G, Taylor TSJ (1992) Generation of Gaussian Processes and Linear Chaos. In: Proc 31st IEEE Conf on Decision and Control. Tucson, pp 2125–2131
53. Picci G, Katayama T (1996) Stochastic realization with exogenous inputs and "Subspace Methods" Identification. Signal Process 52(2):145–160
54. Picci G (1996) Geometric methods in Stochastic Realization and System Identification. CWI Q (invited paper) 9:205–240
55. Picci G (1997) Oblique Splitting Susbspaces and Stochastic Realization with Inputs. In: Helmke U, Prätzel–Wolters D, Zerz E (eds) Operators, Systems and Linear Algebra. Teubner, Stuttgart, pp 157–174
56. Popov VM (1964) Hyperstability and Optimality of Automatic Systems with several Control functions. Rev Rumaine Sci Tech Ser Electrotech Energ 9:629–690
57. Rozanov NI (1963) Stationary Random Processes. Holden-Day, San Francisco
58. Ruckebusch G (1976) Répresentations markoviennes de processus gaussiens stationnaires. Comptes Rendues Acad Sci Paris Series A 282:649–651
59. Ruckebusch G (1978) A state space approach to the stochastic realization problem. Proc 1978 IEEE Intern Symp Circuits and Systems, pp 972–977
60. Ruckebusch G (1978) Factorisations minimales de densités spectrales et répresentations markoviennes. In: Proc 1re Colloque AFCET–SMF, Palaiseau
61. Söderström T, Stoica P (1989) System Identification. Prentice–Hall, New York
62. van Schuppen JH (1989) Stochastic realization problems. In: Nijmeijer H, Schumacher JM (eds) Three decades of mathematical system theory. Lecture Notes in Control and Information Sciences, vol 135. Springer, New York, pp 480–532
63. Van Overschee P, De Moor B (1993) Subspace algorithms for the stochastic identification problem. Automatica 29:649–660
64. Van Overschee P, De Moor B (1994) N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems. Automatica 30:75–93
65. Verhaegen M (1994) Identification of the deterministic part of MIMO State Space Models given in Innovations form from Input–Output data. Automatica 30:61–74
66. Vidyasagar M, (Hidden) Markov Models: Theory and Applications to Biology. Princeton University Press, Princeton

67. Wiener N (1930) Generalized Harmonic Analysis. Acta Math 55:117–258
68. Wiener N (1933) Generalized Harmonic Analysis. In: The Fourier Integral and Certain of its Applications. Cambridge University Press, Cambridge
69. Wiener N (1949) Extrapolation, Interpolation and Smoothing of Stationary Time Series. The M.I.T. Press, Cambridge
70. Willems JC (1971) Least Squares Stationary Optimal Control and the Algebraic Riccati Equation. IEEE Trans Autom Control AC-16:621–634
71. Yakubovich VA (1963) Absolute stability of Nonlinear Control Systems in critical case: part I and II. Avtom Telemeh 24:293–303, 717–731
72. Zeiger HP, McEwen AJ (1974) Approximate linear realization of given dimension via Ho's algorithm. IEEE Trans Autom Control AC-19:153

Books and Reviews

Doob JL (1942) The Brownian movement and stochastic equations. Ann Math 43:351–369
Faurre P, Clerget M, Germain F (1979) Opérateurs Rationnels Positifs. Dunod, Paris
Kalman RE (1981) Realization of covariance sequences. Proc Toeplitz Memorial Conference, Tel Aviv, Birkhäuser
Nelson E (1967) Dynamical theories of Brownian motion. Princeton University press, Princeton
Picci G (1976) Stochastic realization of Gaussian processes. Proc IEEE 64:112–122
Picci G (1992) Stochastic model reduction by aggregation. In: Isidori A, Tarn TJ (eds) Systems Models and Feedback: Theory and Applicatons. Birkhäuser Verlag, Basel, pp 169–177

Stochastic Processes

Alan J. McKane
Theory Group, School of Physics and Astronomy, University of Manchester, Manchester, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Markov Chains
The Master Equation
The Fokker–Planck Equation
Stochastic Differential Equations
Path Integrals
System Size Expansion
Future Directions
Bibliography

Glossary

Fokker–Planck equation A partial differential equation of the second order for the time evolution of the probability density function of a stochastic process. It resembles a diffusion equation, but has an extra term which represents the deterministic aspects of the process.

Langevin equation A stochastic differential equation of the simplest kind: linear and with an additive Gaussian white noise. Introduced by Langevin [30] in 1908 to describe Brownian motion; many stochastic differential equations in physics go by this name.

Markov process A stochastic process in which the current state of the system is only determined from its state in the immediate past, and not by its entire history.

Markov chain A Markov process where both the states and the time are discrete and where the process is stationary.

Master equation The equation describing a continuous-time Markov chain.

Stochastic process A sequence of stochastic variables. This sequence is usually a time-sequence, but could also be spatial.

Stochastic variable A random variable. This is a function which maps outcomes to numbers (real or integer).

Definition of the Subject

The most common type of stochastic process comprises of a set of random variables $\{x(t)\}$, where t represents the time which may be real or integer valued. Other types of stochastic process are possible, for instance when the stochastic variable depends on the spatial position $\mathbf{r}$, as well as, or instead of, t. Since in the study of complex systems we will predominantly be interested in applications relating to stochastic dynamics, we will suppose that it depends only on t. One of the earliest investigations of a stochastic process was carried out by Bachelier [2], who used the idea of a random walk to analyze stock market fluctuations. The problem of a random walk was more generally discussed by Pearson [42], and applied to the investigation of Brownian motion by Einstein [8,9], Smoluchowski [54] and Langevin [30]. The example of a random walk illustrates the fact that in addition to time being discrete or continuous, the stochastic variable itself can be discrete (for instance, the walker moves with fixed step size in one dimension) or continuous (for instance, the velocity of a Brownian particle). The modeling of the process may lead to an equation for the stochastic variable, such as a stochastic differential equation, or for an equation which predicts how the probability density function (pdf) for the stochastic variable changes in time. Stochastic processes are ubiquitous in the physical, biological and social sciences; they may come about through the perception of

very complicated processes being essentially random (the toss of a coin, roll of a die, birth or death of individuals in populations), the inclusion of diverse and poorly characterized effects external to the system under consideration, or thermal fluctuations, among others.

Introduction

Deterministic dynamical processes are typically formulated as a set of rules which allow for the state of the system at time $t+1$ (or $t+\delta t$) to be found from the state of the system at time t. By contrast, for stochastic systems, we can only specify the probability of finding the system in a given state. If this only depends on the state of the system at the previous time step, but not those before this, the stochastic process is said to be Markov. Fortunately many stochastic processes are Markovian to a very good approximation, since the theory of non-Markov processes is considerably more complicated than Markov processes and much less well developed. In this article we will deal almost exclusively with Markov processes.

The mathematical definition of a Markov process follows from the definition of the hierarchy of pdfs for a given process. This involves the joint pdfs $P(x_1, t_1; x_2, t_2; \ldots; x_n, t_n)$, which are the probability that the system is in state x_1 at time t_1, state x_2 at time t_2,..., and state x_n at time t_n, and also the conditional pdfs $P(x_1, t_1; \ldots; x_m, t_m | x_{m+1}, t_{m+1}; \ldots; x_n, t_n)$, which are the probability that the system is in state x_1 at time $t_1, \ldots, x_m$ at time t_m, *given* that it was in state x_{m+1} at time $t_{m+1}, \ldots, x_n$ at time t_n. These pdfs are all non-negative and normalizable, and relations exist between them due to symmetry and reduction (integration over some of the state variables). Nevertheless, for a general non-Markov process, a whole family of these pdfs will be required to specify the process. On the other hand, for a Markov process the history of the system, apart from the immediate past, is forgotten, and so $P(x_1, t_1; \ldots; x_m, t_m | x_{m+1}, t_{m+1}; \ldots; x_n, t_n) = P(x_1, t_1; \ldots, x_m; t_m | x_{m+1}, t_{m+1})$. A direct consequence of this is that the whole hierarchy of pdfs can be determined from only two of them: $P(x, t)$ and $P(x, t|x', t')$. The hierarchy of defining equations then collapses to only two:

$$P(x_2, t_2) = \int \mathrm{d}x_1 P(x_2, t_2 | x_1, t_1) P(x_1, t_1) \tag{1}$$

and

$$P(x_3, t_3 | x_1, t_1) = \int \mathrm{d}x_2 P(x_3, t_3 | x_2, t_2) P(x_2, t_2 | x_1, t_1), \quad t_1 < t_2 < t_3 . \tag{2}$$

The pdf $P(x, t|x', t')$ is referred to as the transition probability and Eq. (2) as the Chapman–Kolmogorov equation. While the pdfs for a Markov process must obey Eqs. (1) and (2), the converse also holds: any two non-negative functions $P(x, t)$ and $P(x, t|x', t')$ which satisfy Eqs. (1) and (2), uniquely define a Markov process.

We will begin our discussion in Sect. "Markov Chains" with what is probably the simplest class of Markov processes: the case when both the state space and time are discrete. These are called Markov chains and were first investigated, for a finite number of states, by Markov [32] in 1906. The extension to an infinite number of states was carried out by Kolmogorov [27] in 1936. If time is continuous, an analogous formalism may be developed, which will be discussed in Sect. "The Master Equation". In physics the equation describing the time evolution of the pdf in this case is called the master equation and was introduced by Pauli [41] in 1928, in connection with the approach to equilibrium for quantum systems, and also by Nordsieck, Lamb and Uhlenbeck [39] in 1940, in connection with fluctuations in cosmic ray physics. The term "master equation" refers to that fact that many of the quantities of interest can be derived from this equation. The connection with previous work on Markov processes was clarified by Siegert [49] in 1949.

In many instances when the master equation cannot be solved exactly, it is useful to approximate it by a rather coarser description of the system, known as the Fokker–Planck equation. This approach will be discussed in Sect. "The Fokker–Planck Equation". This equation was used in its linear form by Rayleigh [44], Einstein [8,9], Smoluchowski [54,55], and Fokker [14], but it was Planck [43] who derived the general nonlinear form from a master equation in 1917, and Kolmogorov who made the procedure rigorous in 1931. All the descriptions which we have mentioned so far have been based on the time evolution of the pdfs. An alternative specification is to give the time evolution of the stochastic variables themselves. This will necessarily involve random variables appearing in the equations describing this evolution, and they will therefore be *stochastic differential equations*. The classic example is the Langevin equation [30] used to describe Brownian motion. This equation is linear and can therefore be solved exactly. The Langevin approach, and its relation to the Fokker–Planck equation is described in Sect. "Stochastic Differential Equations".

A good summary of the understanding of stochastic processes that had been gained by the mid-1950s is given in the book edited by Wax. This covers the basics of the subject, and what is discussed in the first six sections of this article. The article by Chandrasekhar [4], first

published in 1943, and reprinted in Wax, gives an extensive bibliography of stochastic problems in physics before 1943. Since then the applications of the subject have grown enormously, and the equations modeling these systems have correspondingly become more complex. We illustrate some of the procedures which have been developed to deal with these equations in the next two sections. In Sect. "Path Integrals" we discuss how the path-integral formalism may be applied to stochastic processes and in Sect. "System Size Expansion" we describe how master equations can be analyzed when the size of the system is large. We end with a look forward to the future in Sect. "Future Directions".

Markov Chains

The simplest version of the Markov process is when both the states and the time are discrete, and when the stochastic process is *stationary*. When the states are discrete we will denote them by n or m, rather than x, which we reserve for continuous state variables. In this notation the two Eqs. (1) and (2) governing Markov processes read

$$P(n_2, t_2) = \sum_{n_1} P(n_2, t_2|n_1, t_1)P(n_1, t_1) \tag{3}$$

$$P(n_3, t_3|n_1, t_1) = \sum_{n_2} P(n_3, t_3|n_2, t_2)P(n_2, t_2|n_1, t_1)\,, \qquad t_1 < t_2 < t_3\,. \tag{4}$$

A stationary process is one in which the conditional pdf $P(n, t|n', t')$ only depends on the time difference $(t - t')$. For such processes, when time is discrete so that $t = t' + 1, t' + 2, \ldots$, we may write $P(n, t' + k|n', t')$ as $p^{(k)}_{n\,n'}$. The most elementary form of the Chapman–Kolmogorov equation (4) may then be expressed as

$$p^{(2)}_{n\,m} = \sum_{n'} p^{(1)}_{n\,n'} p^{(1)}_{n'\,m}\,. \tag{5}$$

This corresponds to the matrix multiplication of $p^{(1)}$ with itself, and therefore $p^{(2)}$ is simply $(p^{(1)})^2$. In the same way $p^{(k)} = (p^{(1)})^k$, and from now on we drop the superscript on $p^{(1)}$ and denote the matrix by $\mathcal{P}$. The entries of $\mathcal{P}$ are non-negative, with the sum of entries in each column being equal to unity, since

$$\sum_n p_{n\,n'} = \sum_n P(n, t + 1|n', t) = 1\,. \tag{6}$$

Such matrices are called *stochastic matrices*. From Eq. (5) it is clear that $\mathcal{P}^2$ is also a stochastic matrix, and by induction it follows that $\mathcal{P}^k$ is a stochastic matrix if $\mathcal{P}$ is.

The other defining relation for a Markov process, Eq. (3), now becomes

$$P(n, t + 1) = \sum_{n'} p_{n\,n'} P(n', t)\,. \tag{7}$$

This relation defines a *Markov chain*. It has two ingredients: the probability that the system is in state n at time t, $P(n, t)$ – which is usually what we are trying to determine, and the stochastic matrix with entries $p_{n\,n'}$ which gives the probabilities of transitions from the state n' to the state n. The transition probabilities are typically given; they define the model. Note that in many texts the probability of making a transition from n' to n is written as $p_{n'\,n}$, not $p_{n\,n'}$. If we write $P(n, t)$ as a vector $\mathbf{P}(t)$, then we may write Eq. (7) as $\mathbf{P}(t + 1) = \mathcal{P}\mathbf{P}(t)$. Therefore,

$$\mathbf{P}(t) = \mathcal{P}\mathbf{P}(t - 1) = \mathcal{P}\mathcal{P}\mathbf{P}(t - 2) = \cdots = \mathcal{P}^t\mathbf{P}(0)\,, \tag{8}$$

and so if the initial state of the system $\mathbf{P}(0)$ is given, then we can find the state of the system at time t ($\mathbf{P}(t)$) by matrix multiplication by the tth power of the transition matrix.

Examples of Markov Chains

1. *A one-dimensional random walk*. The most widely known example of a Markov chain is a random walk on the real axis, where the walker takes single steps between integers on the line. The simplest version is where the walker has to move during every time interval:

$$p_{n\,n'} = \begin{cases} p\,, & \text{if} \quad n = n' + 1 \\ q\,, & \text{if} \quad n = n' - 1 \\ 0\,, & \text{otherwise}\,, \end{cases} \tag{9}$$

where $p + q = 1$. There are many variants. For instance, the walker could have a non-zero probability of staying put, in which case $p_{n\,n} = r$, with $p + q + r = 1$. The walk could be heterogeneous, in which case p and q (and r), could depend on n'. If there are boundaries, the boundary conditions have to be given. The most common two are *absorbing boundaries* defined by

$$\begin{aligned} p_{n+1\,n} &= p\,, \quad p_{n-1\,n} = q\,, \quad (n = 2, \ldots, N - 1) \\ p_{11} &= 1\,, \quad p_{NN} = 1\,, \\ p_{n\,n'} &= 0\,, \quad \text{otherwise}\,, \end{aligned} \tag{10}$$

and *reflecting boundaries* defined by

$$\begin{aligned} p_{n+1\,n} &= p\,, \quad p_{n-1\,n} = q\,, \quad (n = 2, \ldots, N-1) \\ p_{21} &= p\,, \quad p_{N-1\,N} = q\,, \\ p_{11} &= q\,, \quad p_{NN} = p\,, \\ p_{n\,n'} &= 0\,, \quad \text{otherwise}\,. \end{aligned} \tag{11}$$

With absorbing boundaries (10), if we reach the state 1 or N, we can never leave it, since we stay there with probability 1. When the boundary is reflecting, we can never move beyond it; the only options are to move back towards the other boundary, or stay put. Well known examples of absorbing boundaries include the gambler's ruin problem, where a gambler bets a given amount against the house at each time step and can win with a probability p. Here the absorbing boundary is situated at $n = 0$. Eventually he will arrive at the state $n = 0$, where is has no money left, and so cannot re-enter the game. Birth/death processes will also have absorbing states at $n = 0$: if n is the number of individuals at a given time, and p is the probability of a birth and q of a death, then if there are no individuals left ($n = 0$), none can be born. This condition will be automatically applied if the transition probabilities are proportional to the number of individuals in the population.

2. *The Ehrenfest urn* This Markov chain was introduced by the Ehrenfests [7] in 1907, to illustrate the approach to equilibrium in a gas. Two containers, A and B, contain molecules of the same gas, the sum of the number of molecules in A and B being fixed to be N. At time t a molecule is removed at random from the containers and put into the other. If n' is the number of molecules in container A at a certain time, then at the next time step the transition probabilities will be:

$$p_{n\,n'} = \begin{cases} \frac{n'}{N}\,, & \text{if} \quad n' = n+1 \\ (1 - \frac{n'}{N})\,, & \text{if} \quad n' = n-1 \\ 0\,, & \text{otherwise}\,. \end{cases} \tag{12}$$

This is clearly a heterogeneous random walk of the type (9), and another interpretation of this model is as a random walk, but with a central force.

The most frequently asked question concerning Markov chains is: what is their eventual fate; how does the system behave at large time? Clearly if it tends towards a non-trivial stationary state, $P_{\rm st}(n)$, then from Eq. (7):

$$P_{\rm st}(n) = \sum_{n'} p_{n\,n'} P_{\rm st}(n')\,, \tag{13}$$

and so $P_{\rm st}(n)$ is a right eigenvector of $\mathcal{P}$ with unit eigenvalue. It follows from the properties of a general stochastic matrix that the eigenvalues of a stochastic matrix are such that $|\lambda| \le 1$ [15]. Furthermore every stochastic matrix has an eigenvalue equal to 1, however it may not be simple – there may be a multiplicity of unit eigenvalues. The classification of Markov chains can be used to decide which of these possibilities is the case. For example, Markov chains may be reducible or irreducible, and states recurrent or transient. We shall not discuss this in detail; Feller [10] gives a clear account of this classification and Cox and Miller [5] explore the consequences for the nature of the eigenvalues. Instead we will examine a specific example, that of the Ehrenfest urn introduced above, and focus on the explicit calculation of the eigenvalues and eigenvectors in that case.

Suppose that $\Psi^{(k)}$ and $\Theta^{(k)}$ are the right- and left-eigenvectors of $\mathcal{P}$, respectively, corresponding to the eigenvalue $\lambda^{(k)}$, so that $\Theta^{(k)}.\,\Psi^{(\ell)} = \delta_{k\,\ell}$. Then, in general, and for the Ehrenfest urn in particular,

$$\mathcal{P}^t_{n\,n'} = \sum_{k=0}^{N} \psi_n^{(k)} \left(\lambda^{(k)}\right)^t \theta_{n'}^{(k)}\,, \tag{14}$$

where $\psi_n^{(k)}$ is the nth component of the vector $\Psi^{(k)}$ and similarly for the left-eigenvector. The eigenvalues and eigenvectors for the Ehrenfest urn can be found exactly (Kac [23]; see also Krafft and Schaefer [28]). The eigenvalues are $\lambda^{(k)} = 1 - (2k/N)$, $k = 0, \ldots, N$. Thus in this case there is a single eigenvalue $\lambda = 1$. The corresponding right-eigenvector, which gives the stationary state, is a binomial distribution:

$$P_{\rm st}(n) = \psi_n^{(0)} = \frac{N!}{n!(N-n)!}\,\frac{1}{2^N}\,. \tag{15}$$

The left-eigenvector is $\theta_n^{(0)} = 1$ for all n. The other eigenvectors have the form $\theta_n^{(k)} = a_{kn}$ and $\psi_n^{(k)} = a_{kn} P_{\rm st}(n)$, where the a_{kn} are the Krawtchouk polynomials [1]. Clearly, $a_{0n} = 1$, and the first non-trivial polynomial is $a_{1n} = \sqrt{N}[1 - (2n/N)]$. Therefore, for a suitable choice of initial conditions and using Eq. (8), the large t behavior of the Ehrenfest urn can be found from

$$\mathcal{P}^t_{n\,n'} \approx P_{\rm st}(n) \left\{ 1 + c_{n\,n'} \left(1 - \frac{2}{N}\right)^t \right\}\,, \tag{16}$$

where $c_{n\,n'} = N[1 - (2n/N)][1 - (2n'/N)]$.

The Master Equation

The master equation is a Markov chain in the limit where time is continuous. To derive it we will assume that the states are discrete (the derivation is essentially identical

if they are continuous) and write down the Chapman–Kolmogorov equation (4) in the form:

$$P(n, t+\Delta t|n_0, t_0) = \sum_{n'} P(n, t+\Delta t|n', t)P(n', t|n_0, t_0). \tag{17}$$

We consider only stationary processes, so that we may take $t_0 = 0$ without loss of generality and $P(n, t + \Delta t|n', t)$ is independent of t. We now assume that

$$P(n, t+\Delta t|n', t) = \begin{cases} 1 - \kappa(n)\Delta t + o(\Delta t)\,, & \text{if } n = n' \\ T(n|n')\Delta t + o(\Delta t)\,, & \text{if } n \neq n'\,, \end{cases} \tag{18}$$

where $o(\Delta t)$ means that $o(\Delta t)/\Delta t$ tends to zero as $\Delta t \to 0$. This reasonable: after a very short times the transition probability to stay put is unity minus a term of order Δt and the transition probabilities to move to any other state is of order Δt, but this is still an additional assumption on the process. The quantity $T(n|n')$ is the transition rate, and is only defined for $n \neq n'$. Since $\sum_n P(n, t + \Delta t|n', t) = 1$ for all n', we have that

$$\kappa(n') = \sum_{n\neq n'} T(n|n')\,. \tag{19}$$

Substituting Eq. (18) into Eq. (17), and making use of Eq. (19) we find that

$$\begin{aligned} &\frac{P(n, t + \Delta t|n_0, 0) - P(n, t|n_0, 0)}{\Delta t} \\ &= \sum_{n'\neq n} \left[T(n|n')P(n', t|n_0, 0)\right] \\ &\qquad - P(n, t|n_0, 0) \sum_{n'\neq n} \left[T(n'|n)\right] + \frac{o(\Delta t)}{\Delta t}\,. \end{aligned} \tag{20}$$

Taking the limit $\Delta t \to 0$ gives *the master equation* for how the probability of finding the system in state n at time t changes with time:

$$\frac{\mathrm{d}P(n, t)}{\mathrm{d}t} = \sum_{n'\neq n} T(n|n')\,P(n', t) - \sum_{n'\neq n} T(n'|n)\,P(n, t)\,. \tag{21}$$

We have dropped the initial conditions, assuming that they are understood. It should be noticed that an analogous analysis starting from Eq. (3), rather than Eq. (4), may be carried out, leading to identical equations for $P(n, t|n_0, 0)$ and $P(n, t)$. If the state space is continuous the master equation reads

$$\frac{\partial P(x, t)}{\partial t} = \int \mathrm{d}x' \left[T(x|x')\,P(x', t) - T(x'|x)\,P(x, t)\right]\,. \tag{22}$$

In most applications transitions only take place between states whose label differs by one. That is, $T(n|n')$ and $T(n'|n)$ are zero unless $n' = n + 1$ and $n' = n - 1$. These are called *one-step processes*. For such processes the master equation takes the simpler form

$$\begin{aligned} \frac{\mathrm{d}P(n, t)}{\mathrm{d}t} &= T(n|n + 1)P(n + 1, t) \\ &\quad + T(n|n - 1)P(n - 1, t) \\ &\quad - \left[T(n - 1|n) + T(n + 1|n)\right] P(n, t)\,. \end{aligned} \tag{23}$$

For simplicity let us write

$$g_n = T(n + 1|n) \quad \text{and} \quad r_n = T(n - 1|n)\,, \tag{24}$$

then the master equation may be written as

$$\begin{aligned} \frac{\mathrm{d}P(n, t)}{\mathrm{d}t} &= r_{n+1}P(n + 1, t) + g_{n-1}P(n - 1, t) \\ &\quad - \left[r_n + g_n\right] P(n, t)\,. \end{aligned} \tag{25}$$

Examples of Master Equations

1. *The simple birth–death process*. For a population of simple organisms, for example a colony of bacteria, it might be reasonable to assume that the rate of birth of new bacteria is proportional to the number present at that time, and similarly for the rate of death. This is clearly a Markov process with $g_n = bn$ and $r_n = \mathrm{d}n$, where b and d are rate constants. A variant is to include "immigrants" coming into the population from the outside at a constant rate c, so that $g_n = bn + c$.
   In this example g_n and r_n are linear in n. Such *linear one-step processes* can be solved by the introduction of the generating function $F(z, t) = \sum_n P(n, t)z^n$. This converts the master equation (which is a differential-difference equation) into a partial differential equation for $F(z, t)$ which can be solved if the process is linear. The simplest case of a pure death process ($b = c = 0$ in the above) can illustrate the general procedure. By rescaling the time ($t = \tau/d$), we may write the master equation in the very simple form

$$\frac{\mathrm{d}P(n, \tau)}{\mathrm{d}\tau} = (n + 1)\,P(n + 1, \tau) - n\,P(n, \tau)\,.$$

Multiplying this equation by z^n and summing over all $n \geq 0$ gives

$$\begin{aligned}
&\frac{\partial}{\partial \tau}\left\{\sum_{n=0}^{\infty} z^n P(n,\tau)\right\} \\
&= \sum_{n=0}^{\infty}(n+1)z^n P(n+1,\tau) - \sum_{n=1}^{\infty} n z^n P(n,\tau) \\
&= \sum_{m=1}^{\infty} m z^{m-1} P(m,\tau) - z\sum_{n=1}^{\infty} n z^{n-1} P(n,\tau)\,,
\end{aligned}$$

that is,

$$\frac{\partial F}{\partial \tau} = (1-z)\,\frac{\partial F}{\partial z}\,.$$

A change of variable to $\xi = (1-z)\mathrm{e}^{-\tau}$ and $\eta = \tau$ shows F to be a function of ξ only: $F(z,\tau) = \phi([1-z]\mathrm{e}^{-\tau})$. The function ϕ may be determined from the initial condition. For instance, if $P(n,0) = \delta_{n\,N}$, then $F(z,0) = z^N$ and so $\phi(\xi) = (1-\xi)^N$. This gives the solution for F to be

$$F(z,\tau) = \left[\left(1-\mathrm{e}^{-\tau}\right) + z\mathrm{e}^{-\tau}\right]^N\,.$$

In this case F can easily be expanded as a power series in z, and the $P(n,\tau)$ read off, but even if this is difficult, the moments of the distribution can be readily found by differentiation with respect to z and then setting $z = 1$. It should now be clear why F is called a generating function. In the general case the partial differential equation for F may be solved by standard methods [50]. The solution for a birth–death process without immigration is given in the book by Reichl [45]. The solution with immigration was first given by Kendall [25], who also introduced the technique of the generating function as a method of solution of the master equation.

2. *The Moran model of genetic drift*. Stochastic processes occur extensively in population genetics. The simplest, and most widely known, is a model of genetic drift introduced by Fisher [13] and Wright [58], in which a population of individuals in generation t mate randomly to produce the new generation $t+1$. We assume, for simplicity, that each individual has only one gene of a particular type, and that this may exist in one of two forms (alleles) denoted by A and B. The Wright–Fisher model is based on the sampling of the gene pool at generation t, which consists of n genes of type A and $(N-n)$ genes of type B, to produce the next generation of N genes. Although this may be formulated as a Markov chain, neither Fisher nor Wright did so; this was first carried out by Malécot [31] in 1944. Here we will describe a variant of the model introduced by Moran [36,37], which is a one-step process and can be formulated as a master equation.
The Moran model does not have non-overlapping generations, as in the Wright–Fisher model, and is more akin to a birth–death process where birth and death are coupled. At a given time, two individuals are sampled with replacement: one is designated the parent which is copied to create an offspring and the other is sacrificed to make way for the new offspring. Clearly if a B (chosen with probability $(N-n)/N$) is sacrificed and an A (chosen with probability n/N) is copied, this gives a contribution to $T(n+1|n)$. If the choice is that with A and B interchanged, this gives a contribution to $T(n-1|n)$. The transition rates for the Moran model are thus

$$\begin{aligned}
T(n+1|n) &= \beta\left(1-\frac{n}{N}\right)\left(\frac{n}{N}\right)\,, \\
T(n-1|n) &= \beta\left(\frac{n}{N}\right)\left(1-\frac{n}{N}\right)\,,
\end{aligned} \tag{26}$$

where β is a rate constant, which may be absorbed into the time t.
This may be extended in various ways. For example, mutations may be included: $A \xrightarrow{u} B$ and $B \xrightarrow{v} A$. With probability $(1-u-v)$ the offspring is taken to be a copy of the parent without mutation, as previously described. For the rest of the time (that is, with probability $u+v$), a mutation occurs. If the parent is an A, the offspring becomes a B with probability $u/(u+v)$, and if the parent is a B, the offspring becomes an A with probability $v/(u+v)$. This leads to the transition rates

$$\begin{aligned}
T(n+1|n) &= (1-u-v)\left(1-\frac{n}{N}\right)\left(\frac{n}{N}\right) \\
&\qquad + v\left(1-\frac{n}{N}\right)\,, \\
T(n-1|n) &= (1-u-v)\left(1-\frac{n}{N}\right)\left(\frac{n}{N}\right) + u\left(\frac{n}{N}\right)\,.
\end{aligned} \tag{27}$$

The master equation for the Moran model will be discussed again in the next section.

3. *Competition in a single species model*. The birth–death process described in Example 1 can be generalized to more complex ecological situations. As it stands it consists of the two processes $A \xrightarrow{d} E$ and $A \xrightarrow{b} A + A$ representing death and birth respectively. Here A represents an individual and E is a null state. To model the finite resources available in a given patch, we

put a limit on the number of allowed individuals: $n = 0, 1, \ldots, N$. We can also only allow a birth if enough space and/or other resources are available: $A + E \xrightarrow{b} A + A$ and include competition for these resources: $A + A \xrightarrow{c} A + E$. Since the probability of obtaining an A when sampling the patch is n/N and of obtaining an E is $(N - n)/N$, the birth term is now proportional to $n(N - n)/[N(N - 1)]$ and the competition term proportional to $n(n - 1)/[N(N - 1)]$. This gives the transition rates to be

$$\begin{aligned} T(n+1|n) &= \frac{2bn(N-n)}{N(N-1)}\,, \\ T(n-1|n) &= \frac{cn(n-1)}{N(N-1)} + \frac{\mathrm{d}n}{N}\,. \end{aligned} \tag{28}$$

This approach can be extended to more than one species, for instance competition between and within two species [34] or predator-prey interactions [35]. In these cases the state space is multi-dimensional: $\mathbf{n} = (n_1, n_2, \ldots)$. The master equation still has the form (21), but with n replaced everywhere by the vector $\mathbf{n}$.

Whether or not a stationary state of the master equation exists depends on the nature of the boundary conditions. There are many types of boundary conditions, but two are particularly important. If the boundaries are reflecting, then the probability current vanishes there. If they are absorbing, then the probability of being at the boundary is zero. In the former case probability is conserved, in the latter case it is not, and leaks out of the system.

So to find a non-zero pdf as $t \to \infty$ (a stationary distribution) we therefore assume that the system lies within two reflecting boundaries. For a one-step process, the net flow of probability from the state n to the state $n + 1$, is $J(n, t) = g_n P(n, t) - r_{n+1} P(n + 1, t)$, where $J(n, t)$ is the probability current. The master Eq. (25) may be written as

$$\frac{\mathrm{d}P(n,t)}{\mathrm{d}t} = J(n-1,t) - J(n,t)\,.$$

For a stationary state, the left-hand side of this equation is zero, and the currents will be time-independent. Therefore $J(n - 1) = J(n)$ for all n, that is, all the currents are equal. Since the current vanishes at the boundaries, this constant must be zero. Therefore, for reflecting boundary conditions, $r_{n+1} P_{\mathrm{st}}(n + 1) = g_n P_{\mathrm{st}}(n)$ for all n. If we suppose that one boundary is at $n = 0$ and the other at $n = N$, then we have that $P_{\mathrm{st}}(1) = (g_0/r_1) P_{\mathrm{st}}(0)$, $P_{\mathrm{st}}(2) = (g_1/r_2) P_{\mathrm{st}}(1), \ldots$, which implies $P_{\mathrm{st}}(2) = (g_1 g_0)/(r_2 r_1) P_{\mathrm{st}}(0), \ldots$ Iterating, the stationary state can be expressed as a simple product:

$$P_{\mathrm{st}}(n) = \frac{g_{n-1} g_{n-2} \cdots g_0}{r_n r_{n-1} \ldots r_1} P_{\mathrm{st}}(0)\,, \quad n = 1, \ldots, N\,. \tag{29}$$

The constant $P_{\mathrm{st}}(0)$ is determined by normalization:

$$\begin{aligned} &\sum_{n=0}^{N} P_{\mathrm{st}}(n) = P_{\mathrm{st}}(0) + \sum_{n>0} P_{\mathrm{st}}(n) = 1 \\ &\qquad \Rightarrow (P_{\mathrm{st}}(0))^{-1} = 1 + \sum_{n=1}^{N} \frac{g_{n-1} g_{n-2} \cdots g_0}{r_n r_{n-1} \ldots r_1}\,. \end{aligned} \tag{30}$$

As an example, we return to the Ehrenfest urn (12), which in the language of the master equation is defined by $g_n = (N - n)/N$ and $r_n = n/N$ (any overall rate may be absorbed into the time, and this is irrelevant as far as the stationary state is concerned). Here $n = 0, 1, \ldots, N$ and the molecules never go outside this range, so the boundaries are reflecting. Applying Eqs. (29) and (30) shows that the stationary state is the binomial distribution given by Eq. (15).

The Fokker–Planck Equation

The Fokker–Planck equation describes stochastic processes at a more coarse grained level than those that we have discussed so far. It only involves continuous stochastic variables; these could be for instance the fraction of individuals or genes of a certain kind in a population, whereas the master equation recognized the individuals or genes as discrete entities. To obtain the Fokker–Planck equation we first derive the Kramers–Moyal expansion [29,38].

We begin by defining the *jump moments* for the system:

$$M_\ell(x, t, \Delta t) = \int \mathrm{d}\xi\ (\xi - x)^\ell\, P(\xi, t + \Delta t | x, t)\,. \tag{31}$$

We will assume that these are known, that is, they can be obtained by some other means. They will, however, only be required in the limit of small Δt.

The starting point for the derivation is the Chapman–Kolmogorov equation (4), with the choice of variables analogous to that used in Eq. (17) for the discrete case:

$$P(x, t + \Delta t) = \int \mathrm{d}x'\, P(x, t + \Delta t | x', t)\, P(x', t)\,, \tag{32}$$

again dropping the dependence on the initial conditions. The integrand may be written as

$$\begin{aligned} &P(x, t+\Delta t|x', t)\, P(x', t) \\ &= P([x-\Delta x]+\Delta x, t+\Delta t|[x-\Delta x], t) P([x-\Delta x], t) \\ &= \sum_{\ell=0}^{\infty} \frac{(-1)^\ell}{\ell!} (\Delta x)^\ell \\ &\qquad \cdot \frac{\partial^\ell}{\partial x^\ell} \{P(x+\Delta x, t+\Delta t|x, t) P(x, t)\} \ , \end{aligned} \tag{33}$$

where $\Delta x = x - x'$. Integrating over x' gives for Eq. (32):

$$P(x, t+\Delta t) = \sum_{\ell=0}^{\infty} \frac{(-1)^\ell}{\ell!} \frac{\partial^\ell}{\partial x^\ell} \{M_\ell(x, t, \Delta t) P(x, t)\} \ . \tag{34}$$

Since $P(\xi, t|x, t) = \delta(\xi - x)$, it follows from Eq. (31), that $\lim_{\Delta t \to 0} M_\ell(x, t, \Delta t) = 0$ for $\ell \geq 1$. Also $M_0(x, t, \Delta t) = 1$. Bearing these results in mind, we will now assume that the jump moments for $\ell \geq 1$ take the form

$$M_\ell(x, t, \Delta t) = D^{(\ell)}(x, t)\Delta t + o(\Delta t) \ . \tag{35}$$

Substituting this into Eq. (34), dividing by Δt and taking the limit $\Delta t \to 0$ gives

$$\frac{\partial P}{\partial t} = \sum_{\ell=1}^{\infty} \frac{(-1)^\ell}{\ell!} \frac{\partial^\ell}{\partial x^\ell} \left\{D^{(\ell)}(x, t) P(x, t)\right\} \ . \tag{36}$$

Equation (36) is the Kramers–Moyal expansion. So far nothing has been assumed other than the Markov property and the existence of Taylor series expansions. However, in many situations, examination of the jump moments reveal that in a suitable approximation they may be neglected for $\ell > 2$. In this case, we may truncate the Kramers–Moyal expansion (36) at second order and obtain the *Fokker–Planck equation*:

$$\frac{\partial P}{\partial t} = -\frac{\partial}{\partial x}\left[A(x, t)P(x, t)\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[B(x, t)P(x, t)\right] \ , \tag{37}$$

where $A = D^{(1)}$ and $B = D^{(2)}$ are independent of t if the process is stationary.

To calculate the jump moments (31), it is convenient to write them in terms of the underlying stochastic process $x(t)$. We use the notation

$$\langle x(t)\rangle_{x(t_0)=x_0} = \int \mathrm{d}x\, x P(x, t|x_0, t_0) \ , \tag{38}$$

for the mean of the stochastic variable at time t, conditional on the value of $x(t)$ being given to be x_0 at time t_0. With this notation $x(t)$ denotes the process and the angle brackets are averages over realizations of this process. More generally, we may define $\langle f(x(t))\rangle$ in a similar way, and in particular the jump moments are given by

$$M_\ell(x, t, \Delta t) = \langle (x(t+\Delta t) - x)^\ell \rangle_{x(t)=x} \ . \tag{39}$$

Examples of Fokker–Planck Equations

1. *Simple diffusion*. For the simple symmetric random walk, $g_n = 1$ and $r_n = 1$ when expressed in the language of the master equation (after a rescaling of the time so that the rates may taken to be equal to unity). From Eqs. (18) and (19) we find that for a one-step stationary process,

$$\begin{aligned} &\langle (n(t+\Delta t) - n)^\ell \rangle_{n(t)=n} \\ &= \begin{cases} (g_n - r_n)\,\Delta t + o(\Delta t) \ , & \text{if } \ell \text{ is odd} \\ (g_n + r_n)\,\Delta t + o(\Delta t) \ , & \text{if } \ell \text{ is even} \ , \end{cases} \end{aligned} \tag{40}$$

and so for the symmetric random walk the odd moments all vanish, and the even moments are equal to $2\Delta t + o(\Delta t)$. We now make the approximation which will yield the Fokker–Planck equation: we let $x = nL$, where L is the step size, and let $L \to 0$. Since, for ℓ even, $\langle (x(t+\Delta t) - x)^\ell \rangle = (2L^\ell)\Delta t$, if we rescale the time by introducing $\tau = L^2 t$, then all jump moments higher than the second disappear in the limit $L \to 0$, and Eq. (36) becomes

$$\frac{\partial P}{\partial t} = \frac{\partial^2 P}{\partial x^2} \ . \tag{41}$$

This is the familiar diffusion equation obtained from a continuum approximation to the discrete random walk.

2. *The diffusion limit of the Moran model*. For the Moran model with no mutation, we have from Eqs. (26) and (40) that the odd moments again vanish. If we describe the process by $x(t) = n(t)/N$, the fraction of the genes that are of type A at time t, then the even jump moments are given by

$$\begin{aligned} &\left\langle (x(t+\Delta t) - x)^\ell \right\rangle_{x(t)=x} \\ &\qquad = \frac{1}{N^\ell}\, 2x\,(1-x)\,\Delta t + o(\Delta t) \ , \end{aligned}$$

and so introducing a rescaled time $\tau = 2t/N^2$, and letting $N \to \infty$ we obtain the Fokker–Planck equation

$$\frac{\partial P}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[x(1-x)P\right] \ . \tag{42}$$

The factor of 2 in the rescaling of the time is included so that the diffusive form of the Moran model agrees with that found in the Wright–Fisher model [6].

Now suppose mutations are included. The transition rates are given by Eq. (27) and lead to jump moments $\langle (x(t+\Delta t)-x)^{\ell}\rangle \sim N^{-\ell}$. So the first and second jump moments are not of the same order, and the introduction of a rescaled time $\tau = t/N$, and subsequently letting $N \to \infty$, gives a Fokker–Planck equation of the form (37), but with $B = 0$ and $A(x) = v - (u+v)x$. This corresponds to a *deterministic* process with $\mathrm{d}x/\mathrm{d}t = v - (u+v)x$ [22]. The fact that the system tends to a macroscopic equation when $N \to \infty$ has to be taken into account when determining the nature of the fluctuations for large N. We will discuss this further in Sect. "System Size Expansion".

On the other hand, suppose that the mutation rates scale with N according to $u = 2\tilde{u}/N$ and $v = 2\tilde{v}/N$, where $\tilde{u}$ and $\tilde{v}$ have a finite limit as $N \to \infty$, and where the 2 has again been chosen to agree with the Wright–Fisher model. Now both the first and second jump moments are of order N^{-2}, with the higher moments falling off faster with N. Therefore once again introducing the rescaled time $\tau = 2t/N^2$ and letting $N \to \infty$, we obtain the Fokker–Planck equation

$$\frac{\partial P}{\partial t} = -\frac{\partial}{\partial x}\left[\{\tilde{v} - (\tilde{u}+\tilde{v})\,x\}\,P\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[x(1-x)P\right] . \tag{43}$$

So depending on the precise scaling with N, the Moran model gives different limits as $N \to \infty$ [24]. In the first case, the mutational effects are strong enough that a macroscopic description exists, with fluctuations about the macroscopic state, as discussed in Sect. "System Size Expansion". In the second case, the mutational effects are weaker, and there is no macroscopic equation describing the system; the large N limit is a nonlinear Fokker–Planck equation of the diffusive type.

The Fokker–Planck equation (37) for stationary processes, where A and B are functions of x only, can be solved by separation of variables, with solutions of the form $p(x)\mathrm{e}^{-\lambda t}$, where λ is a constant. The equation for $p(x)$ is then a second order differential equation which, when boundary conditions are given, is of the Sturm–Liouville type. To specify the boundary conditions we once again introduce the probability current, this time through the continuity equation,

$$\frac{\partial P(x,t)}{\partial t} + \frac{\partial J(x,t)}{\partial x} = 0 , \tag{44}$$

where the probability current $J(x,t)$ is given by

$$J(x,t) = A(x,t)P(x,t) - \frac{1}{2}\frac{\partial}{\partial x}\left[B(x,t)P(x,t)\right] . \tag{45}$$

Let us suppose that the system is defined on the interval $[a,b]$. Then if the boundaries are reflecting, there is no net flow of probability across $x = a$ and $x = b$. This implies that $J(a,t) = 0$ and $J(b,t) = 0$. If we integrate the equation of continuity (44) from $x = a$ to $x = b$, and apply these boundary conditions, we see that $\int_a^b P(x,t)\mathrm{d}x$ is independent of time. Therefore, if the pdf is initially normalized, it remains normalized. This is in contrast with the case of absorbing boundary conditions, defined by $P(a,t) = 0$ and $P(b,t) = 0$. If the boundary conditions are at infinity we require that $\lim_{x\to\pm\infty} P(x,t) = 0$, so that if P is well-behaved it is normalizable, and also that $\partial P/\partial x$ is well-behaved in this limit: $\lim_{x\to\pm\infty} \partial P/\partial x = 0$. If A or B do not diverge as $x \to \pm\infty$ this implies that $\lim_{x\to\pm\infty} J(x,t) = 0$. Other types of boundary conditions are possible, and we do not attempt a complete classification here [16,46].

If A and B are independent of time, then from Eq. (44), the stationary state of the system must be given by $\mathrm{d}J(x)/\mathrm{d}x = 0$, that is, J is a constant. For reflecting boundary conditions this constant is zero, and so from Eq. (45) the stationary pdf, $P_{\mathrm{st}}(x)$ must satisfy

$$0 = A(x)P_{\mathrm{st}}(x) - \frac{1}{2}\frac{\partial}{\partial x}\left[B(x)P_{\mathrm{st}}(x)\right] . \tag{46}$$

This may be integrated to give

$$P_{\mathrm{st}}(x) = \frac{C}{B(x)}\exp\left\{2\int^x \mathrm{d}x' \;\frac{A(x')}{B(x')}\right\} , \tag{47}$$

where C is a constant which has to be chosen so that $P_{\mathrm{st}}(x)$ is normalized.

The Fokker–Planck equation, with A independent of t and B constant, can be transformed into the Schrödinger-like problem

$$-B\frac{\partial \psi}{\partial t} = -\frac{B^2}{2}\frac{\partial^2 \psi}{\partial x^2} + U(x)\psi , \tag{48}$$

by the transformation

$$P(x,t) = \left[P_{\mathrm{st}}(x)\right]^{1/2}\psi(x,t) , \tag{49}$$

where

$$U(x) = \frac{1}{2}\left[A(x)\right]^2 + \frac{B}{2}\frac{\mathrm{d}A}{\mathrm{d}x} . \tag{50}$$

So a one-dimensional stationary stochastic process, under certain conditions (such as constant second jump moment) is equivalent to quantum mechanics in imaginary time, with B taking over the role of Planck's constant.

As an example we consider the *Ornstein–Uhlenbeck process* defined by

$$\frac{\partial P}{\partial t} = \frac{\partial}{\partial x}[axP] + D\frac{\partial^2 P}{\partial x^2}, \quad x \in (-\infty, \infty), \quad a > 0. \quad (51)$$

In this case, the potential (50) is given by $U(x) = [a^2x^2 - 2aD]/2$, and so the problem is equivalent to the one-dimensional simple harmonic oscillator in quantum mechanics, but with an energy shift. As in that problem [47], the eigenfunctions are Hermite polynomials. Specifically, the right-eigenfunctions are

$$p_m(x) = P_{\mathrm{st}}(x)\frac{1}{[2^m m!]^{1/2}}H_m(\alpha x)\,, \quad (52)$$

where H_m are the Hermite polynomials [1] and $\alpha = (a/2D)^{1/2}$. The eigenvalue corresponding to the eigenfunction (52) is $\lambda_m = am$, $m = 0, 1, \ldots$ and the left-eigenfunctions are $q_m(x) = [P_{\mathrm{st}}(x)]^{-1}\, p_m(x)$. From these explicit solutions we may calculate other quantities of interest, such as correlation functions. We note that the eigenvalues are all non-negative and that the stationary state corresponds to $\lambda = 0$. The left-eigenfunction for the stationary state is equal to 1. These latter results hold true for a wide-class of such problems.

Stochastic Differential Equations

So far we have described stochastic processes in terms of equations which give the time evolution of pdfs. In this section, we will describe equations for the stochastic variables themselves. The most well known instance of such an equation is the Langevin equation for the velocity of a Brownian particle, and so we begin with this particular example.

Suppose that a small macroscopic particle of mass m (such as a pollen grain) is immersed in a liquid at a temperature T. In addition to any macroscopic motion that the particle may have, its velocity fluctuates due to the random collisions of the particle with the molecules of the liquid. For simplicity, we confine ourselves to one-dimensional motion – along the x-axis. Then the equation of motion of the particle may be written in the form

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = -\alpha\frac{\mathrm{d}x}{\mathrm{d}t} - \frac{\mathrm{d}V}{\mathrm{d}x} + \mathcal{F}(t)\,. \quad (53)$$

The first term on the right-hand side is due to the viscosity of the fluid and α is the friction constant. The second term, where $V(x)$ is a potential, represents the interaction of the particle with any external forces, such as gravity. The final term is the random force due to collisions with the molecules of the liquid. Clearly to complete the specification of the dynamics of the particle we need to give (i) the initial position and velocity of the particle, and (ii) the statistics of the random force $\mathcal{F}(t)$.

To make progress with these points, we imagine a large number of realizations of the dynamics, in which the particle starts with the same initial position, x_0, and velocity, v_0, but where the initial positions and velocities of the molecules in the liquid will be different. Taking the average over a large number of such realizations will give the average position $\langle x(t)\rangle$ and velocity $\langle v(t)\rangle$ at time t, conditional on $x(0) = x_0$ and $v(0) = v_0$. The statistics of the fluctuating force $\mathcal{F}(t)$ are assumed to be such that

(a) $\langle \mathcal{F}(t)\rangle = 0$, since we do not expect one direction to be favored over the other.
(b) $\langle \mathcal{F}(t)\mathcal{F}(t')\rangle = 2D\delta(t - t')$, since we expect that after a few molecular collisions the value that $\mathcal{F}$ takes on will be independent of its former value. That is, the force $\mathcal{F}$ becomes uncorrelated over times of the order of a few collision times between molecules. This is tiny on observational time scales, and so taking the correlation function to be a delta-function is an excellent approximation. The weight of the delta-function is denoted by $2D$, where at this stage D is undetermined.
(c) $\mathcal{F}(t)$ is taken to be Gaussianly distributed on grounds of simplicity, but also because by the central limit theorem it is assumed that the net effect of the large number of molecules which collide with the pollen grain will lead to a distribution which is Gaussian.

Since a Gaussian distribution is specified by its first two moments, conditions (a), (b) and (c) completely define the statistics of $\mathcal{F}(t)$.

Finally, Eq. (53) as it stands does not define a Markov process. This is most easily seen if we write down a discrete time version of the equation. The second derivative means that $x(t + \delta t)$ not only depends on $x(t)$, but also on $x(t - \delta t)$. Therefore only first order derivatives should be included in such equations if the process is to be Markov. This is easily achieved by promoting $v(t)$ to be a second stochastic variable in addition to $x(t)$. Then Eq. (53) may be equivalent written as

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= v\,, \\ m\frac{\mathrm{d}v}{\mathrm{d}t} &= -\alpha v - \frac{\mathrm{d}V}{\mathrm{d}x} + \mathcal{F}(t)\,, \end{aligned} \quad (54)$$

which does define a Markov process. Although, as we remarked in the Introduction, we deal almost exclusively with Markov processes in this article, the situation we have just discussed is a good illustration of one way of

dealing with processes which are presented as being non-Markovian. The method simply consists of adding a sufficient number of supplementary variables to the definition of the state variables of the process until it becomes Markovian. There is no guarantee that this will be possible or require only a small number of additional variables to be promoted in this way, but it is the most straightforward and direct way of rendering non-Markovian processes tractable.

To begin the analysis of Eq. (54) we assume that there are no external forces and so the term $\mathrm{d}V/\mathrm{d}x$ is equal to zero. We may then write Eq. (54) as the Langevin equation

$$\frac{\mathrm{d}v}{\mathrm{d}t} = -\gamma v + F(t)\,; \quad v(0) = v_0\,, \tag{55}$$

where $\gamma = \alpha/m$ and $F(t) = \mathcal{F}(t)/m$. This implies that

$$\langle F(t)\rangle = 0 \;\text{ and }\; \langle F(t)F(t')\rangle = \frac{2D}{m^2}\delta(t-t')\,. \tag{56}$$

Note that since $F(t)$ is a random variable, solving the Langevin equation will give $v(t)$ as a random variable (having a known distribution). It is therefore a *stochastic differential equation*. The function F is frequently called "the noise term" or simply "the noise". It is *white noise* since the Fourier transform of a delta-function is a constant – all frequencies are present in equal amounts.

Multiplying the Langevin equation (55) by the integrating factor $\mathrm{e}^{\gamma t}$ gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left[v(t)\mathrm{e}^{\gamma t}\right] = F(t)\mathrm{e}^{\gamma t} &\Rightarrow v(t)\\ &= v_0\mathrm{e}^{-\gamma t} + \mathrm{e}^{-\gamma t}\int_0^t \mathrm{d}t'\,F(t')\mathrm{e}^{\gamma t'}\,.\end{aligned} \tag{57}$$

By taking the average of the expression for $v(t)$ we find $\langle v(t)\rangle = v_0\mathrm{e}^{-\gamma t}$. More interestingly, if we square the expression for $v(t)$ and take the average, then we find

$$\langle v^2(t)\rangle = v_0^2\mathrm{e}^{-2\gamma t} + \frac{D}{\alpha m}\left[1-\mathrm{e}^{-2\gamma t}\right]\,, \tag{58}$$

which implies that

$$\lim_{t\to\infty}\langle v^2(t)\rangle = \frac{D}{\alpha m}\,. \tag{59}$$

On the other hand, as $t\to\infty$, the Brownian particle will be in thermal equilibrium:

$$\lim_{t\to\infty}\langle v^2(t)\rangle = v_{\mathrm{eq}}^2 \quad\text{and}\quad \frac{1}{2}mv_{\mathrm{eq}}^2 = \frac{1}{2}kT\,,$$

where T is the temperature of the liquid and k is Boltzmann's constant. This implies that

$$\frac{1}{2}m\left(\frac{D}{\alpha m}\right) = \frac{1}{2}kT \;\Rightarrow\; D = \alpha kT\,. \tag{60}$$

The molecules of the liquid are acting as a heat bath for the system – which in this case is a single Brownian particle. The equation $D = \alpha kT$ is a simple example of a fluctuation-dissipation theorem, and determines D in terms of the friction constant, α, and of the temperature of the liquid, T.

Although we have presented a somewhat heuristic rationale for Eq. (54), it may be derived in a more controlled way. A particularly clear derivation has been given by Zwanzig [60], where the starting point is a Hamiltonian which contains three terms: for the system, the heat bath and the interaction between the system and the heat bath. Taking the heat bath to be made up of coupled harmonic oscillators and the interaction term between the system and heat bath to be linear, it is possible to integrate out the bath degrees of freedom exactly, and be left only with the equations of motion of the system degrees of freedom plus the initial conditions of the bath degrees of freedom. Assuming that the bath is initially in thermal equilibrium, so that these initial values are distributed according to a Boltzmann distribution, adds extra "noise" terms to the equations of motion which, with a few more plausible assumptions, make them of the Langevin type.

Examples of Langevin–like Equations

1. *Overdamped Brownian motion*. Frequently the viscous damping force $-\alpha v$ is much larger than the inertial term $m\mathrm{d}^2x/\mathrm{d}t^2$ in Eq. (53), and so to a good approximation the left-hand side of Eq. (53) can be neglected. Scaling time by α, we arrive at the Langevin equation for the motion of an overdamped Brownian particle:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -V'(x) + \mathcal{F}(t)\,; \quad x(0) = x_0\,, \tag{61}$$

where $'$ denotes differentiation with respect to x and where, due to the rescaling of time by α,

$$\langle\mathcal{F}(t)\rangle = 0\,; \quad \langle\mathcal{F}(t)\mathcal{F}(t')\rangle = 2\tilde{D}\delta(t-t')\,; \quad \tilde{D} = \frac{D}{\alpha}\,. \tag{62}$$

A particularly well-known case is when the Brownian particle is moving in the harmonic potential $V(x) = ax^2/2$. Then

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -ax + \mathcal{F}(t)\,; \quad x(0) = x_0\,. \tag{63}$$

Since Eq. (63) relating $x(t)$ to $\mathcal{F}(t)$ is linear, and since the distribution of $\mathcal{F}(t)$ is Gaussian, then $x(t)$ is also distributed according to a Gaussian distribution. Comparing with Eqs. (55) and (56), which also define a linear system, we find that $\langle x(t)\rangle = x_0\mathrm{e}^{-at}$ and, from

Eq. (58), $\langle x^2(t)\rangle = x_0^2 e^{-2at} + (\tilde{D}/a)\left[1 - e^{-2at}\right]$. This gives

$$P(x,t|x_0,0) = \sqrt{\frac{a}{2\pi\tilde{D}\left[1-e^{-2at}\right]}}\,\exp\left\{-\frac{a\left(x-x_0e^{-at}\right)^2}{2\tilde{D}\left[1-e^{-2at}\right]}\right\}. \tag{64}$$

It is straightforward to check that this conditional pdf satisfies the Fokker–Planck equation (51) for the Ornstein–Uhlenbeck process. Below we will show this more directly, by starting from the Langevin equation (61) and deriving Eq. (51) if $V(x)$ is quadratic.

Another case of interest is when $V(x)$ is a double-well potential, $V(x) = -ax^2/2 + bx^4/4$. If the particle is initially located near the bottom of one of the potential wells, it will take on average a time of the order of $e^{\Delta V/D}$ to hop over the barrier and into the well on the other side. Here ΔV is the height of the barrier that it has to hop over [29].

2. *Environmental noise in population biology*. One of the simplest models of two species with population sizes N_1 and N_2 which are competing for a common resource, is the two coupled deterministic ordinary differential equations $\dot{N}_i = r_i N_i$, $i = 1, 2$. The growth rates, r_i, depend on the population sizes in such a way that as the population sizes increase, the r_i decrease to reflect the increased competition for resources. This could be modeled, for instance, by taking $r_i = a_i - b_{ii}N_i - b_{ij}N_j$ with $i, j = 1, 2$ and $i \neq j$. In reality, external factors such as climate, terrain, the presence of other species, and indeed any factor which has an uncertain influence on these two species, will also affect the growth rate. This can be modeled by adding an external random term to the r_i which represents this *environmental stochasticity* [33]. Then the equations become

$$\begin{aligned}\frac{dN_1}{dt} &= a_1N_1 - b_{11}N_1^2 - b_{12}N_1N_2 + N_1\zeta_1(t)\\ \frac{dN_2}{dt} &= a_2N_2 - b_{22}N_2^2 - b_{21}N_2N_1 + N_2\zeta_2(t)\,.\end{aligned} \tag{65}$$

Since the noise terms, $\zeta_i(t)$ are designed to reflect the large number of coupled variables omitted from the description of the model, it is natural, by virtue of the central limit theorem, to assume that they are Gaussianly distributed. It also seems reasonable to assume that any temporal correlation between these external influences is on scales very much shorter than those of interest to us here, and that the noises have zero mean. We therefore assume that

$$\langle\zeta_i(t)\rangle = 0\,; \quad \langle\zeta_i(t)\zeta_j(t')\rangle = 2D_i\delta_{ij}\delta(t-t')\,, \tag{66}$$

where the D_i describe the strength of the stochastic effects. The deterministic equations (that is, Eq. (65) without the noise terms) have a fixed point at the origin, one on each of the N_1 and N_2 axes, and may have another at non-zero N_1 and N_2. For some values of the parameters this latter fixed point may be a saddle, with those on the axes being stable and the origin unstable. In this situation the eventual fate of the species depends significantly on the noise: if the combination of the nonlinear dynamics and the noise drives the system to the vicinity of the fixed point on the N_1 axis, then species 2 will become extinct, and vice-versa.

Langevin equations with Gaussian white noise are equivalent to Fokker–Planck equations. This can be most easily seen by calculating the jump moments (39) from the Langevin equation. For instance, if we begin from the Langevin equation for an overdamped Brownian particle (61),

$$\begin{aligned}\Delta x(t) \equiv x(t+\Delta t) - x(t) &= \int_t^{t+\Delta t} dt'\,\dot{x}(t')\\ &= -\int_t^{t+\Delta t} dt'\,V'(x(t')) + \eta(t)\,,\end{aligned} \tag{67}$$

where $\eta(t) = \int_t^{t+\Delta t} dt'\,\mathcal{F}(t')$. From Eq. (62) it is straightforward to calculate the moments of $\eta(t)$: $\langle\eta(t)\rangle = 0$,

$$\langle\eta^2(t)\rangle = \int_t^{t+\Delta t} dt' \int_t^{t+\Delta t} dt''\langle\mathcal{F}(t')\mathcal{F}(t'')\rangle = 2\tilde{D}\Delta t \tag{68}$$

and, since $\eta(t)$ is Gaussian, $\langle\eta^n(t)\rangle$ is zero if n is odd, and at least of order $(\Delta t)^2$ for $n \geq 4$. This implies that

$$\begin{aligned}M_1(x,\Delta t) &= -V'(x)\Delta t + \mathcal{O}\left(\Delta t\right)^2\,,\\ M_2(x,\Delta t) &= 2\tilde{D}\Delta t + \mathcal{O}\left(\Delta t\right)^2\,,\end{aligned} \tag{69}$$

with all moments of order $(\Delta t)^2$ or higher for $\ell > 2$. The notation $\mathcal{O}\left(\Delta t\right)^2$ means that the magnitude of this quantity is less than a constant times $(\Delta t)^2$, for sufficiently small nonzero $(\Delta t)^2$. This is a weaker, but more specific, statement than saying it is $o\left(\Delta t\right)$. Using Eqs. (35) and (36), the Fokker–Planck equation which is equivalent to the

Langevin equation (61) is found to be

$$\frac{\partial P}{\partial t} = \frac{\partial}{\partial x}\left[V'(x)P\right] + \tilde{D}\frac{\partial^2 P}{\partial x^2} . \tag{70}$$

From Eq. (47), the stationary pdf is $P_{\rm st}(x) = C\exp\{-V(x)/\tilde{D}\} = C\exp\{-V(x)/kT\}$, as expected.

Although in this article we have largely restricted our attention to stochastic processes involving one variable, the construction of a Fokker–Planck equation from the Langevin equation goes through in a similar way for an n-dimensional process $\mathbf{x} = (x_1, \ldots, x_n)$. In this case the jump moments are

$$\left\langle \Delta x_{i_1}(t)\Delta x_{i_2}(t)\ldots\Delta x_{i_\ell}(t)\right\rangle_{\mathbf{x}(t)=\mathbf{x}} = D_{i_1\ldots i_\ell}(\mathbf{x},t)\Delta t + o(\Delta t) , \tag{71}$$

where $\Delta x_{i_\alpha} = x_{i_\alpha}(t+\Delta t) - x_{i_\alpha}$. The Kramers–Moyal expansion is then

$$\frac{\partial P}{\partial t} = \sum_{\ell=1}^{\infty}\frac{(-1)^\ell}{\ell!}\frac{\partial^\ell}{\partial x_{i_1}\ldots\partial x_{i_\ell}}\left\{D_{i_1\ldots i_\ell}(\mathbf{x},t)P\right\} . \tag{72}$$

The Langevin equation for Brownian motion (54), without going to the overdamped limit, serves as a simple illustration of this generalization. Here $\Delta x(t) = v\Delta t$ and $\Delta v(t) = -\gamma v\,\Delta t - m^{-1}V'(x)\,\Delta t + m^{-1}\eta(t)$. This results in the Fokker–Planck equation

$$\frac{\partial P}{\partial t} = -\frac{\partial}{\partial x}[vP] + \frac{\partial}{\partial v}\left[\left\{\gamma v + m^{-1}V'(x)\right\}P\right] + \frac{\gamma kT}{m}\frac{\partial^2 P}{\partial v^2} . \tag{73}$$

This is Kramer's equation. It has a stationary pdf $P_{\rm st}(x,v) = C\exp\{-E/kT\}$, where $E = mv^2/2 + V(x)$.

We end this section by finding the Fokker–Planck equation which is equivalent to the general set of Langevin equations of the form

$$\dot{x}_i = A_i(\mathbf{x},t) + \sum_{\alpha=1}^{m} g_{i\alpha}(\mathbf{x},t)\,\zeta_\alpha(t); \quad i = 1,\ldots,n, \tag{74}$$

where $\zeta_\alpha(t), \alpha = 1,\ldots,m$, is a Gaussian white noise with zero mean and with

$$\langle\zeta_\alpha(t)\zeta_\beta(t')\rangle = \delta_{\alpha\beta}\delta(t-t') . \tag{75}$$

Proceeding as in Eq. (67), but noting the dependence of the function $g_{i\alpha}$ on the stochastic variable, yields

$$\begin{aligned} M_i(\mathbf{x},t,\Delta t) &= \left[A_i(\mathbf{x},t) + \theta(0)\sum_{j=1}^{n}\sum_{\alpha=1}^{m} g_{j\alpha}(\mathbf{x},t)\frac{\partial}{\partial x_j}g_{i\alpha}(\mathbf{x},t)\right]\Delta t \\ &\quad + \mathcal{O}(\Delta t)^2 , \\ M_{ij}(\mathbf{x},t,\Delta t) &= \sum_{\alpha=1}^{m}\left[g_{i\alpha}(\mathbf{x},t)g_{j\alpha}(\mathbf{x},t)\right]\Delta t + \mathcal{O}(\Delta t)^2 , \end{aligned} \tag{76}$$

with all jump moments higher than the second being of order $(\Delta t)^2$ or higher. The quantity $\theta(0)$ is the value of the Heaviside theta function, $\theta(x)$, at $x = 0$ and is indeterminate. This indicates that the Langevin description does not correspond to a unique Fokker–Planck equation. This situation occurs whenever the white noise in a Langevin equation is multiplied by a function which depends on the state variable, as in Eq. (74). For systems such as this acted upon by *multiplicative noise* the Langevin description has to be supplemented by a rule which says whether the state variable in the multiplying function ($g_{i\alpha}$ in Eq. (74)) is that before or after the noise pulse acts [52]. If it is taken to be the value immediately after the noise pulse acts then $\theta(0) = 0$ (Itô rule), whereas if it is taken to be the average of the values before and after, then $\theta(0) = 1/2$ (Stratonovich rule). The Fokker–Planck equation is now found from Eq. (72) to be

$$\begin{aligned} \frac{\partial P}{\partial t} &= -\sum_{i=1}^{n}\frac{\partial}{\partial x_i}\left[A_i(\mathbf{x},t)P(\mathbf{x},t)\right] \\ &\quad + \frac{1}{2}\sum_{i,j=1}^{n}\sum_{\alpha=1}^{m}\frac{\partial^2}{\partial x_i\partial x_j}\left[g_{i\alpha}(\mathbf{x},t)g_{j\alpha}(\mathbf{x},t)P(\mathbf{x},t)\right] , \end{aligned} \tag{77}$$

in the Itô case and

$$\begin{aligned} \frac{\partial P}{\partial t} &= -\sum_{i=1}^{n}\frac{\partial}{\partial x_i}\left[A_i(\mathbf{x},t)P(\mathbf{x},t)\right] \\ &\quad + \frac{1}{2}\sum_{i,j=1}^{n}\sum_{\alpha=1}^{m}\frac{\partial}{\partial x_i}\left[g_{i\alpha}(\mathbf{x},t)\frac{\partial}{\partial x_j}\left\{g_{j\alpha}(\mathbf{x},t)P(\mathbf{x},t)\right\}\right] , \end{aligned} \tag{78}$$

in the Stratonovich case.

Path Integrals

While most early work on stochastic processes was concerned with linear systems, naturally attention soon

moved on to the many interesting systems which could be modeled as nonlinear stochastic processes. These systems are much more difficult to analyze. For example, a nonlinear Langevin equation cannot be solved directly, and so the averaging procedure cannot be carried out in the same explicit way as described in Sect. "Stochastic Differential Equations". There is however one method which is applicable to many nonlinear stochastic differential equations of interest: the solution of these equations can be formally written down as a path-integral, and from this correlation functions and other quantities of physical interest can be obtained. This also has the advantage that all the formalism and approximation schemes developed to study functional integrals over the years can be called into play.

Path-integrals are intimately related to Brownian motion and the earliest work on the subject by Wiener [56, 57], emphasized this. If the problem of interest is formulated as a set of Langevin equations, the derivation of the path-integral representation is particularly straightforward, if rather heuristic. For clarity we begin with the simplest case: an overdamped system with a single degree of freedom, x, acted upon by white noise. The Langevin equation is given by Eq. (61) and the noise is defined by Eq. (62). Since the noise is assumed to be Gaussian, Eq. (62) is a complete specification. An equivalent way of giving it is through the pdf [12]:

$$P[\mathcal{F}]\,\mathcal{D}\mathcal{F} \propto \exp\left(-\frac{1}{4\tilde{D}}\int \mathrm{d}t\,\mathcal{F}^2(t)\right)\mathcal{D}\mathcal{F}\,, \tag{79}$$

where $\mathcal{D}\mathcal{F}$ is the functional measure. The idea is now to regard the Langevin equation (61) as defining a mapping $\mathcal{F} \mapsto x$. The pdf for the x variable is then given by

$$\begin{aligned} P[x] &= P[\mathcal{F}]|_{\mathcal{F}=\dot{x}+V'(x)}\,J[x] \\ &\propto \exp\left(-\frac{1}{4\tilde{D}}\int \mathrm{d}t\,[\dot{x}+V'(x)]^2\right)J[x]\,, \end{aligned} \tag{80}$$

where

$$J[x] = \det\left[\frac{\delta\mathcal{F}}{\delta x}\right]\,, \tag{81}$$

is the Jacobian of the transformation. An explicit expression for the Jacobian may be obtained either by direct calculation of a discretized form of the Langevin equation [21] or through use of the identity relating the determinant of a matrix to the exponential of the trace of the logarithm of that matrix [59]. One finds that $J[x] \propto \exp\{\theta(0)\int \mathrm{d}t\,V''(x)\}$. The quantity $\theta(0)$ is once again the indeterminate value of the Heaviside theta function $\theta(x)$ at $x = 0$. Its appearance is a reflection of the fact that, due to the Brownian-like nature of the paths in the functional integral, the nature of the discretization appears explicitly through this factor [48]. If we consistently use the mid-point rule throughout, then we may take $\theta(0) = 1/2$, which gives

$$\begin{aligned} P[x] &\propto \exp\left(-\frac{1}{4\tilde{D}}\int \mathrm{d}t\,[\dot{x}+V'(x)]^2 + \frac{1}{2}\int \mathrm{d}t\,V''(x)\right) \\ &= \exp\left(-S[x]/\tilde{D}\right)\,. \end{aligned} \tag{82}$$

All quantities of interest can now be found from expression (82). For example, the conditional probability distribution, $P(x, t|x_0, t_0)$ is given by

$$\langle\delta(x-x(t))\rangle_{x(t_0)=x_0} = \int_{x(t_0)=x_0}\mathcal{D}x\,\delta(x-x(t))\,P[x]. \tag{83}$$

The expression (82) has much in common with Feynman's formulation of quantum mechanics as a path-integral [11]. In fact another way to obtain the result is to exploit the transformation (49) to write the Fokker–Planck equation (70) as a Schrödinger equation in imaginary time $\tau = it$, with a potential $U(x) = (1/2)[V'(x)]^2 - \tilde{D}V''(x)$, following Eq. (50). The action in the quantum-mechanical path-integral is

$$\begin{aligned} &\frac{i}{\hbar}\int \mathrm{d}t\left[\frac{1}{2}\dot{x}^2 - U(x)\right] \longrightarrow \\ &\quad \frac{1}{2\tilde{D}}\int \mathrm{d}\tau\left[-\frac{1}{2}\dot{x}^2 - \frac{1}{2}\left[V'(x)\right]^2 + \tilde{D}V''(x)\right]\,, \end{aligned} \tag{84}$$

which is Eq. (82) since $\int_{t_0}^{t}\mathrm{d}t\,\dot{x}V'(x) = \int_{x_0}^{x}\mathrm{d}x\,V'(x) = V(x) - V(x_0)$ does not depend on the path, only on the end-points. The functional $S[x]$ is analogous to the action in classical mechanics, and is frequently referred to as such. It is also sometimes referred to as the generalized Onsager–Machlup functional, in recognition of the original work carried out by Onsager and Machlup [40], in the case of a linear Langevin equation, in 1953.

The above discussion can be generalized in many ways. For example, if the Langevin equation for an n-dimensional process takes the form

$$\dot{x}_i = A_i(\mathbf{x}) + \zeta_i(t)\,, \quad \langle\zeta_i(t)\zeta_j(t')\rangle = 2D_{ij}\delta(t-t')\,, \tag{85}$$

where $\zeta_i(t)$ is a Gaussian noise with zero mean and D_{ij} is independent of $\mathbf{x}$, then the general Onsager–Machlup

functional is [21]

$$S[x] = \int \mathrm{d}t \left[\frac{1}{4} \sum_{i,j} \{\dot{x}_i - A_i(\mathbf{x})\} D_{ij}^{-1} \{\dot{x}_j - A_j(\mathbf{x})\} + \frac{1}{2} \sum_i \frac{\partial A_i}{\partial x_i} \right] , \quad (86)$$

if the matrix D_{ij} is non-singular. The generalization to the situation where the noise is multiplicative is more complicated, and is analogous to the path-integral formulation of quantum mechanics in curved space [20].

System Size Expansion

In Example 2 of Sect. "The Fokker–Planck Equation" we explicitly showed how the master equation may have different limits when the size of the system, N, becomes large. In one case both the first and second jump moments were of the same order (and much larger than the higher jump moments) and so a nonlinear Fokker–Planck equation of the diffusion type was obtained in the limit $N \to \infty$. In another case, the first jump moment scaled in a different way to the second moment, and so the $N \to \infty$ limit gave a deterministic macroscopic equation of the form $\dot{x} = f(x)$, with finite N effects presumably consisting of $1/\sqrt{N}$ fluctuations about the macroscopic state, $x(t)$. It is the second scenario that we will explore in this section. It can be formalized by writing

$$\frac{n}{N} = x(t) + \frac{\xi}{\sqrt{N}} , \quad (87)$$

and substituting this into the master equation, then equating terms of the same order in $1/\sqrt{N}$. The leading order equation obtained in this way will be the macroscopic equation, and the function $f(x)$ will emerge from the analysis. The next-to-leading order equation turns out to be a *linear* Fokker–Planck equation in the variable ξ. Higher order terms may also be included. This formalism was first developed by van Kampen [51] and is usually referred to as van Kampen's system-size expansion. We will describe it in the specific case of a one-step process for a single stochastic variable in order to bring out the essential features of the method.

When using this formalism it is useful to rewrite the master equation (23) using step operators which act on an arbitrary function of n according to $\mathcal{E} f(n) = f(n+1)$ and $\mathcal{E}^{-1} f(n) = f(n-1)$. This gives

$$\frac{\mathrm{d}P(n,t)}{\mathrm{d}t} = (\mathcal{E} - 1)\left[T(n-1|n)P(n,t)\right] + \left(\mathcal{E}^{-1} - 1\right)\left[T(n+1|n)P(n,t)\right] . \quad (88)$$

We begin by using Eq. (87) to write the pdf which appears in the master Equation (88) as

$$P(Nx(t) + \sqrt{N}\xi, t) = \Pi(\xi, t) \quad \Rightarrow \quad \dot{P} = \frac{\partial \Pi}{\partial t} - N^{1/2} \frac{\mathrm{d}x}{\mathrm{d}t} \frac{\partial \Pi}{\partial \xi} . \quad (89)$$

This gives an expression for the left-hand side of the master equation, $\dot{P}$. To get an expression for the right-hand side,

(a) the step operators are expanded in powers of $1/\sqrt{N}$ [53]:

$$\mathcal{E}^{\pm 1} = 1 \pm \frac{1}{\sqrt{N}} \frac{\partial}{\partial \xi} + \frac{1}{2!} \frac{1}{N} \frac{\partial^2}{\partial \xi^2} + \mathcal{O}\left(\frac{1}{N^{3/2}}\right) , \quad (90)$$

(b) $T(n \pm 1|n)$ is expressed in terms of ξ and N,
(c) $P(n,t)$ is replaced by $\Pi(\xi, t)$.

Steps (a), (b) and (c) gives the right-hand side of the master equation as a power-series in $1/\sqrt{N}$. Equating the left-hand and right-hand sides order by order in $1/\sqrt{N}$ (this may require a rescaling of the time, t, by a power of $\sqrt{N}$), gives to leading order (the $\partial \Pi / \partial \xi$ cancels) an equation of the form $\mathrm{d}x/\mathrm{d}t = f(x)$. This may be solved subject to the condition $x(0) = x_0 = n_0/N$, if we take the initial condition on the master equation to be $P(n,0) = \delta_{n,n_0}$. We denote the solution of this macroscopic equation by $x_{\mathrm{M}}(t)$.

To next order in $1/\sqrt{N}$, the Fokker–Planck equation

$$\frac{\partial \Pi}{\partial t} = -f'(x) \frac{\partial}{\partial \xi} [\xi \Pi] + \frac{1}{2} g(x) \frac{\partial^2 \Pi}{\partial \xi^2} , \quad (91)$$

describing a linear stochastic process is found. The functions $f'(x)$ and $g(x)$ are to be evaluated when $x = x_{\mathrm{M}}(t)$, and so are simply functions of time. If the macroscopic system tends to a fixed point: $x_{\mathrm{M}}(t) \to x^*$, as $t \to \infty$, then $f'(x)$ and $g(x)$ may be replaced by constants in order to study the fluctuations about this stationary state.

To illustrate the method we use Example 3, Sect. "The Master Equation". Equating both sides of the master equation in this case one finds

$$-N^{1/2} \frac{\mathrm{d}x}{\mathrm{d}t} \frac{\partial \Pi}{\partial \xi} + \frac{\partial \Pi}{\partial t} = \frac{1}{\sqrt{N}} \left[f_-(x) - f_+(x)\right] \frac{\partial \Pi}{\partial \xi} + \frac{1}{2} \frac{1}{N} \left[f_-(x) + f_+(x)\right] \frac{\partial^2 \Pi}{\partial \xi^2} + \frac{1}{N} \left[f'_-(x) - f'_+(x)\right] \frac{\partial}{\partial \xi} [\xi P] + \ldots , \quad (92)$$

where the functions $f_-(x)$ and $f_+(x)$ are given by:

$$f_-(x) = cx^2 + \mathrm{d}x , \quad f_+(x) = 2bx(1-x) . \quad (93)$$

For the left- and right-hand sides to balance in Eq. (92), a rescaled time $\tau = t/N$ needs to be introduced. Then to leading order one finds $\mathrm{d}x/\mathrm{d}\tau = f(x)$, where $f(x) = f_+(x) - f_-(x)$. At next to leading order Eq. (91) is found with $g(x) = f_+(x) + f_-(x)$. The explicit form of the macroscopic equation is

$$\frac{\mathrm{d}x}{\mathrm{d}\tau} = x\,(r - ax)\ , \tag{94}$$

where $r = 2b - d$ and $a = 2b + c$. Equation (94) is the logistic equation, which is the usual phenomenological way to model intraspecies competition.

Since the Fokker–Planck equation (91) describes a linear process, its solution is a Gaussian. This means that the probability distribution $\Pi(\xi, t)$ is completely specified by the first two moments $\langle\xi(t)\rangle$ and $\langle\xi^2(t)\rangle$. Multiplying Eq. (91) by ξ and ξ^2 and integrating over all ξ one finds

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\langle\xi(t)\rangle &= f'(x_\mathrm{M}(t))\langle\xi(t)\rangle\ ,\\ \frac{\mathrm{d}}{\mathrm{d}t}\langle\xi^2(t)\rangle &= 2f'(x_\mathrm{M}(t))\langle\xi^2(t)\rangle + g(x_\mathrm{M}(t))\ .\end{aligned} \tag{95}$$

We have chosen the initial condition to be $x_0 = n_0/N$, which implies that $\xi(0) = 0$. The first equation in (95) then implies that $\langle\xi(t)\rangle = 0$ for all t. Multiplying the second equation by $f^{-2}(x_\mathrm{M}(t))$ one finds that

$$\langle\xi^2(t)\rangle = f^2(x_\mathrm{M}(t))\int_0^t \mathrm{d}t'\,\frac{g(x_\mathrm{M}(t'))}{f^2(x_\mathrm{M}(t'))}\ , \tag{96}$$

and so the determination of $\langle\xi^2(t)\rangle$ is reduced to quadrature. The method can be applied to systems with more than one stochastic variable and those which are not one-step processes. Details are given in van Kampen's book [53].

Future Directions

This article has focused largely on classical topics in the theory of stochastic processes, since these form the foundations on which the subject is built. Much of the current work, and one would expect future work, will be numerical in character. Some of this will begin from a basic Markovian description in the form of reactions among chemical species – even if the system is not chemical in nature (Example 3 of Sect. "The Master Equation" is an example). A straightforward algorithm developed by Gillespie [17,18], and since then extended and improved [3,19], provides an efficient way of simulating such systems. It thus provides a valuable method of investigating systems which may be formulated as complicated multivariable master equations, which complements the methods we have discussed here. However, many current studies do not begin from a system which can be described in this way, and there is every indication that this will be more true in the future. For instance, in agent based models the stochastic element may be due to mutations in characteristics, traits or behavior, which may be difficult or impossible to formulate mathematically. Such agent based models are certainly individually based, but each individual may have different attributes and generally behave in such a complex way that only numerical simulations can be used to explore the behavior of the system as a whole. Although these complex systems may be used to model more realistic situations, the well-known problems associated with the large number of parameters typically required to describe such systems, will mean that simplified versions will need to be analyzed in order to understand them at a deeper level. These simpler models are likely to include those where the agents of a particular species are essentially identical. In this article we have discussed how the classical equations of the theory of stochastic processes, such as the Fokker–Planck equation, can be obtained from such models. They will therefore form a bridge between the agent-based approaches which are expected to become more prevalent in the future, and the analytic approaches which lie at the heart of the theory of stochastic processes.

Bibliography

Primary Literature

1. Abramowitz M, Stegun I (Eds) (1965) Handbook of mathematical functions. Dover, New York
2. Bachelier L (1900) Théorie de la spéculation. Annales Scientifiques de L'Ecole Normale Supérieure III(17):21–86
3. Cao Y, Li H, Petzold L (2004) Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. J Chem Phys 121:4059–4067
4. Chandrasekhar S (1943) Stochastic problems in physics and astronomy. Rev Mod Phys 15:1–89. Reprinted in Wax (1954)
5. Cox DR, Miller HD (1968) The theory of stochastic processes, Chap 3. Chapman and Hall, London
6. Crow JF, Kimura M (1970) An introduction to population genetics theory. Harper and Row, New York
7. Ehrenfest P, Ehrenfest T (1907) Über zwei bekannte Einwände gegen das Boltzmannsche H-Theorem. Phys Z 8:311–314
8. Einstein A (1905) Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. Ann Physik 17:549–560. For a translation see: A. Einstein, "Investigations on the Theory of the Brownian Movement" Fürth R (ed), Cowper AD (tr) (Dover, New York, 1956). Chapter I

9. Einstein A (1906) Zur Theorie der Brownschen Bewegung. Ann Physik 19:371–381. For a translation see: A. Einstein, "Investigations on the Theory of the Brownian Movement" Fürth R (ed), Cowper AD (tr) (Dover, New York, 1956). Chapter II
10. Feller W (1968) An introduction to probability theory and its applications, Chap XV, 3rd edn. Wiley, New York
11. Feynman RP (1948) Space-time approach to non-relativistic quantum mechanics. Rev Mod Phys 20:367–387
12. Feynman RP, Hibbs AR (1965) Quantum mechanics and path integrals, Chap 12. McGraw-Hill, New York
13. Fisher RA (1930) The genetical theory of natural selection. Clarendon Press, Oxford
14. Fokker AD (1914) Die mittlere Energie rotierende elektrischer Dipole im Strahlungsfeld. Ann Physik 43:810–820
15. Gantmacher FR (1959) The theory of matrices, Chap 13, Sect 6, vol 2. Chelsea Publishing Co., New York
16. Gardiner CW (2004) Handbook of stochastic methods, 3rd edn. Springer, Berlin
17. Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J Comput Phys 22:403–434
18. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. J Phys Chem 81:2340–2361
19. Gillespie DT (2001) Approximate accelerated stochastic simulation of chemically reacting systems. J Chem Phys 115:1716–1733
20. Graham R (1977) Path integral formulation of general diffusion processes. Z Physik B26:281–290
21. Graham R (1975) Macroscopic theory of fluctuations and instabilities. In: Riste T (ed) Fluctuations, Instabilities, and Phase Transitions. Plenum, New York, pp 215–293
22. Haken H (1983) Synergetics. Springer, Berlin. Sect 6.3
23. Kac M (1947) Random walk and the theory of Brownian motion. Amer Math Mon 54:369–391
24. Karlin S, McGregor J (1964) On some stochastic models in genetics. In: Gurland J (ed) Stochastic problems in medicine and biology. University of Wisconsin Press, Madison, pp 245–279
25. Kendall DG (1948) On some modes of population growth leading to R. A. Fisher's logarithmic series distribution. Biometrika 35:6–15
26. Kolmogorov AN (1931) Über die analytischen Methoden in der Wahrscheinlichkeitsrechung. Math Ann 104:415–458
27. Kolmogorov AN (1936) Anfangsgründe der Theorie der Markoffschen Ketten mit unendlich vielen möglichen Zuständen. Mat Sbornik (N.S.) 1:607–610
28. Krafft O, Schaefer M (1993) Mean passage times for tridiagonal transition matrices and a two-parameter Ehrenfest urn model. J Appl Prob 30:964–970
29. Kramers HA (1940) Brownian motion in a field of force and the diffusion model of chemical reactions. Physica 7:284–304
30. Langevin P (1908) Sur la théorie du mouvement brownien. C R Acad Sci Paris 146:530–533. For a translation see: D. S. Lemons and A. Gythiel, Am. J. Phys. 65: 1079–1081 (1997)
31. Malécot G (1944) Sur un problème de probabilités en chaine que pose la génétique. C R Acad Sci Paris 219:379–381
32. Markov AA (1906) Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga. Izv Fiz-Matem Obsch Kazan Univ (Series 2) 15:135–156. See also: "Extension of the limit theorems of probability theory to a sum of variables connected in a chain", in Appendix B of R. Howard "Dynamic Probabilistic Systems, vol 1: Markov Chains" (John Wiley and Sons, 1971)
33. May RM (1973) Model ecosystems, Chap 5. Princeton University Press, Princeton
34. McKane AJ, Newman TJ (2004) Stochastic models in population biology and their deterministic analogs. Phys Rev E 70:041902
35. McKane AJ, Newman TJ (2005) Predator-prey cycles from resonant amplification of demographic stochasticity. Phys Rev Lett 94:218102
36. Moran PAP (1958) Random processes in genetics. Proc Cambridge Philos Soc 54:60–72
37. Moran PAP (1962) The statistical processes of evolutionary theory, Chap 4. Clarendon Press, Oxford
38. Moyal JE (1949) Stochastic processes and statistical physics. J Roy Stat Soc (London) B 11:150–210
39. Nordsieck A, Lamb WE Jr, Uhlenbeck GE (1940) On the theory of cosmic-ray showers. I. The Furry model and the fluctuation problem. Physica 7:344–360
40. Onsager L, Machlup S (1953) Fluctuations and irreversible processes. Phys Rev 91:1505–1512
41. Pauli W (1928) Probleme der modernen Physik. In: Debye P (ed) Festschrift zum 60. Geburtstag A. Sommerfeld. Hirzel, Leipzig, p 30
42. Pearson K (1905) The problem of the random walk. Nature 72:294,342
43. Planck M (1917) Über einen Satz der statistischen Dynamik und seine Erweiterung in der Quantentheorie. Abh Preuss Akad Wiss Berl 24:324–341
44. Rayleigh L (1891) Dynamical problems in illustration of the theory of gases. Phil Mag 32:424–445
45. Reichl LE (1998) A modern course in statistical physics, Chap 5, 2nd edn. Wiley, New York
46. Risken H (1989) The Fokker–Planck equation, 2nd edn. Springer, Berlin
47. Schiff LI (1968) Quantum mechanics, Chap 4, 3rd edn. McGraw-Hill, Tokyo
48. Schulman LS (1981) Techniques and applications of path–integration, Chap 5. Wiley, New York
49. Siegert AJF (1949) On the approach to statistical equilibrium. Phys Rev 76:1708–1714
50. Sneddon IN (1957) Elements of partial differential equations, Chap 2. McGraw-Hill, New York
51. van Kampen NG (1961) A power series expansion of the master equation. Can J Phys 39:551–567
52. van Kampen NG (1981) Itô versus Stratonovich. J Stat Phys 24:175–187
53. van Kampen NG (1992) Stochastic processes in physics and chemistry, 2nd edn. North-Holland, Amsterdam
54. von Smoluchowski M (1906) Zur kinetishen Theorie der Brownschen Molekularbewegung und der Suspensionen. Ann Physik 21:756–780
55. von Smoluchowski M (1916) Drei Vortage über Diffusion, Brownsche Bewegung, und Koagulation von Kolloidteilchen. Phys Z 17:571–599
56. Wiener N (1921) The average of an analytic functional. Proc Natl Acad Sci USA 7:253–260
57. Wiener N (1921) The average of an analytic functional and the Brownian movement. Proc Natl Acad Sci USA 7:294–298

58. Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159
59. Zinn-Justin J (2002) Quantum field theory and critical phenomena, 4th edn. Clarendon Press, Oxford. Sect 4.8.2
60. Zwanzig R (1973) Nonlinear generalized Langevin equations. J Stat Phys 9:215–220

Books and Reviews

Wax N (1954) Selected papers on noise and stochastic processes. Dover, New York

Stochastic Volatility

Torben G. Andersen[1,2,3] , Luca Benzoni[4]
[1] Kellogg School of Management, Northwestern University, Evanston, USA
[2] NBER, Cambridge, USA
[3] CREATES, Aarhus, Denmark
[4] Federal Reserve Bank of Chicago, Chicago, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Model Specification
Realized Volatility
Applications
Estimation Methods
Future Directions
Acknowledgments
Bibliography

Glossary

Implied volatility The value of asset return volatility which equates a model-implied derivative price to the observed market price. Most notably, the term is used to identify the volatility implied by the Black and Scholes [63] option pricing formula.

Quadratic return variation The ex-post sample-path return variation over a fixed time interval.

Realized volatility The sum of finely sampled squared asset return realizations over a fixed time interval. It is an estimate of the quadratic return variation over such time interval.

Stochastic volatility A process in which the return variation dynamics include an unobservable shock which cannot be predicted using current available information.

Definition of the Subject

Given the importance of return volatility on a number of practical financial management decisions, the efforts to provide good real-time estimates and forecasts of current and future volatility have been extensive. The main framework used in this context involves stochastic volatility models. In a broad sense, this model class includes GARCH, but we focus on a narrower set of specifications in which volatility follows its own random process, as is common in models originating within financial economics. The distinguishing feature of these specifications is that volatility, being inherently unobservable and subject to independent random shocks, is not measurable with respect to observable information. In what follows, we refer to these models as *genuine* stochastic volatility models.

Much modern asset pricing theory is built on continuous-time models. The natural concept of volatility within this setting is that of genuine stochastic volatility. For example, stochastic volatility (jump-)diffusions have provided a useful tool for a wide range of applications, including the pricing of options and other derivatives, the modeling of the term structure of risk-free interest rates, and the pricing of foreign currencies and defaultable bonds. The increased use of intraday transaction data for construction of so-called realized volatility measures provides additional impetus for considering genuine stochastic volatility models. As we demonstrate below, the realized volatility approach is closely associated with the continuous-time stochastic volatility framework of financial economics.

There are some unique challenges in dealing with genuine stochastic volatility models. For example, volatility is truly latent and this feature complicates estimation and inference. Further, the presence of an additional state variable – volatility – renders the model less tractable from an analytic perspective. We review how such challenges have been addressed through development of new estimation methods and imposition of model restrictions allowing for closed-form solutions while remaining consistent with the dominant empirical features of the data.

Introduction

The label Stochastic Volatility is applied in two distinct ways in the literature. For one, it is used to signify that the (absolute) size of the innovations of a time series displays random fluctuations over time. Descriptive models of financial time series almost invariably embed this feature nowadays as asset return series tend to display alternating quiet and turbulent periods of varying length and intensity. To distinguish this feature from models that operate with an a priori known or deterministic path

for the volatility process, the random evolution of the conditional return variance is termed stochastic volatility. The simplest case of deterministic volatility is the constant variance assumption invoked in, e. g., the Black and Scholes [63] framework. Another example is modeling the variance purely as a given function of calendar time, allowing only for effects such as time-of-year (seasonals), day-of-week (institutional and announcement driven) or time-of-day (diurnal effects due to, e. g., market microstructure features). Any model not falling within this class is then a stochastic volatility model. For example, in the one-factor continuous-time Cox, Ingersoll, and Ross [113] (CIR) model the (stochastic) level of the short term interest rate governs the dynamics of the (instantaneous) drift and diffusion term of all zero-coupon yields. Likewise, in GARCH models the past return innovations govern the one-period ahead conditional mean and variance. In both models, the volatility is known, or deterministic, at a given point in time, but the random evolution of the processes renders volatility stochastic for any horizon beyond the present period.

The second notion of stochastic volatility, which we adopt henceforth, refers to models in which the return variation dynamics is subject to an unobserved random shock so that the volatility is inherently latent. That is, the current volatility state is not known for sure, conditional on the true data generating process and the past history of all available discretely sampled data. Since the CIR and GARCH models described above render the current (conditional) volatility known, they are not stochastic volatility models in this sense. In order to make the distinction clear cut, we follow Andersen [10] and label this second, more restrictive, set *genuine* stochastic volatility (SV) models.

There are two main advantages to focusing on SV models. First, much asset pricing theory is built on continuous-time models. Within this class, SV models tend to fit more naturally with a wide array of applications, including the pricing of currencies, options, and other derivatives, as well as the modeling of the term structure of interest rates. Second, the increasing use of high-frequency intraday data for construction of so-called realized volatility measures is also starting to push the GARCH models out of the limelight as the realized volatility approach is naturally linked to the continuous-time SV framework of financial economics.

One drawback is that volatility is not measurable with respect to observable (past) information in the SV setting. As such, an estimate of the current volatility state must be filtered out from a noisy environment and the estimate will change as future observations become available. Hence, in-sample estimation typically involves smoothing techniques, not just filtering. In contrast, the conditional variance in GARCH is observable given past information, which renders (quasi-)maximum likelihood techniques for inference quite straightforward while smoothing techniques have no role. As such, GARCH models are easier to estimate and practitioners often rely on them for time-series forecasts of volatility. However, the development of powerful method of simulated moments, Markov Chain Monte Carlo (MCMC) and other simulation based procedures for estimation and forecasting of SV models may well render them competitive with ARCH over time on that dimension.

Direct indications of the relations between SV and GARCH models are evident in the sequence of papers by Dan Nelson and Dean Foster exploring the SV diffusion limits of ARCH models as the case of continuous sampling is approached, see, e. g., Nelson and Foster [219]. Moreover, as explained in further detail in the estimation section below, it can be useful to summarize the dynamic features of asset returns by tractable pseudo-likelihood scores obtained from GARCH-style models when performing simulation based inference for SV models. As such, the SV and GARCH frameworks are closely related and should be viewed as complements. Despite these connections we focus, for the sake of brevity, almost exclusively on SV models and refer the interested reader to the GARCH chapter for further information.

The literature on SV models is vast and rapidly growing, and excellent surveys are available, e. g., Ghysels et al. [158] and Shephard [239,240]. Consequently, we focus on providing an overview of the main approaches with illustrations of the scope for applications of these models to practical finance problems.

Model Specification

The original econometric studies of SV models were invariably cast in discrete time and they were quite similar in structure to ARCH models, although endowed with a more explicit structural interpretation. Recent work in the area has been mostly directly towards a continuous time setting and motivated by the typical specifications in financial economics. This section briefly reviews the two alternative approaches to specification of SV models.

Discrete-Time SV Models and the Mixture-of-Distributions Hypothesis

Asset pricing theory contends that financial asset prices reflect the discounted value of future expected cash flows, implying that all news relevant for either discount rates or cash flows should induce a shift in market prices. Since

economic news items appear almost continuously in real time, this perspective rationalizes the ever-changing nature of prices observed in financial markets. The process linking news arrivals to price changes may be complex, but if it is stationary in the statistical sense it will nonetheless produce a robust theoretical association between news arrivals, market activity and return volatility. In fact, if the number of news arrival is very large, standard central limit theory will tend to imply that asset returns are approximately normally distributed *conditional* on the news count. More generally, variables such as the trading volume, the number of transactions or the number of price quotes are also naturally related to the intensity of the information flow. This line of reasoning has motivated specifications such as

$$y_t|s_t \rightsquigarrow N(\mu_y s_t\ ,\ \sigma_y^2 s_t)\ , \tag{1}$$

where y_t is an "activity" variable related to the information flow, s_t is a positive intensity process reflecting the rate of news arrivals, μ_y represents the mean response to an information event, and σ_y is a pure scaling parameter.

This is a normal mixture model, where the s_t process governs or "mixes" the scale of the distribution across the periods. If s_t is constant, this is simply an i.i.d. Gaussian process for returns and possible other related variables. However, this is clearly at odds with the empirical evidence for, e. g., return volatility and trading volume. Therefore, s_t is typically stipulated to follow a separate stochastic process with random innovations. Hence, each period the return series is subject to two separate shocks, namely the usual idiosyncratic error term associated with the (normal) return distribution, but also a shock to the variance or volatility process, s_t. This endows the return process with genuine stochastic volatility, reflecting the random intensity of news arrivals. Moreover, it is typically assumed that only returns, transactions and quotes are observable, but not the actual value of s_t itself, implying that σ_y cannot be separately identified. Hence, we simply fix this parameter at unity.

The time variation in the information flow series induces a fat-tailed unconditional distribution, consistent with stylized facts for financial return and, e. g., trading volume series. Intuitively, days with a lot of news display more rapid price fluctuations and trading activity than days with a low news count. In addition, if the s_t process is positively correlated, then shocks to the conditional mean and variance processes for y_t will be persistent. This is consistent with the observed clustering in financial markets, where return volatility and trading activity are contemporaneously correlated and each display pronounced positive serial dependence.

The inherent randomness and unobserved nature of the news arrival process, even during period t, renders the true mean and variance series latent. This property is the major difference with the GARCH model class, in which the one-step-ahead conditional mean and variance are a known function of observed variables at time $t-1$. As such, for genuine SV models, we must distinguish the full, but infeasible, information set ($s_t \in \mathcal{F}_t$) and the observable information set ($s_t \notin \mathcal{I}_t$). This basic latency of the mixing variable (state vector) of the SV model complicates inference and forecasting procedures as discussed below.

For short horizon returns, μ_y is nearly negligible and can reasonably be ignored or simply fixed at a small constant value, and the series can then be demeaned. This simplification produces the following return (innovation) model,

$$r_t = \sqrt{s_t}\, z_t\ , \tag{2}$$

where z_t is an i.i.d. standard normal variable, implying a simple normal-mixture representation,

$$r_t|s_t \rightsquigarrow N(0, s_t)\ . \tag{3}$$

Univariate return models of the form (3) as well as multivariate systems including a return variable along with other related market activity variables, such as the transactions count, the quote intensity or the aggregate trading volume, stem from the Mixture-of-Distributions Hypothesis (MDH).

Actual implementation of the MDH hinges on a particular representation of the information-arrival process s_t. Clark [102] uses trading volume as a proxy for the activity variable, a choice motivated by the high contemporaneous correlation between return volatility and volume. Tauchen and Pitts [247] follow a structural approach to characterize the joint distribution of the daily return and volume relation governed by the underlying latent information flow s_t. However, both these models assume temporal independence of the information flow, thus failing to capture the clustering in these series. Partly in response, Gallant et al. [153] examine the joint conditional return-volume distribution without imposing any structural MDH restrictions. Nonetheless, many of the original discrete-time SV specifications are compatible with the MDH framework, including Taylor [249][1], who proposes an autoregressive parametrization of the latent log-volatility (or information flow) variable

$$\log(s_{t+1}) = \eta_0 + \eta_1 \log(s_t) + u_t,\ u_t \rightsquigarrow \text{i.i.d}(0, \sigma_u^2)\ , \tag{4}$$

[1] Discrete-time SV models go father back in time, at least to the easly paper by Rosenberg [232] recently reprinted in Shephard [240].

where the error term, u_t, may be correlated with the disturbance term, z_t, in the return Eq. (2) so that $\rho = \mathrm{corr}(u_t, z_t) \neq 0$. If $\rho < 0$, downward movements in asset prices result in higher future volatility as also predicted by the so-called 'leverage effect' in the exponential GARCH, or EGARCH, form of Nelson [218] and the asymmetric GARCH model of Glosten et al. [160].

Early tests of the MDH include Lamoureux and Lastrapes [194] and Richardson and Smith [231]. Subsequently, Andersen [11] studies a modified version of the MDH that provides a much improved fit to the data. Further refinements of the MDH specification have been pursued by, e. g., Liesenfeld [198,199] and Bollerslev and Jubinsky [67]. Among the first empirical studies of the related approach of stochastic time changes are Ané and Geman [29], who focus on stock returns, and Conley et al. [109], who focus on the short-term risk-free interest rate.

Continuous-Time Stochastic Volatility Models

Asset returns typically contain a predictable component, which compensates the investor for the risk of holding the security, and an unobservable shock term, which cannot be predicted using current available information. The conditional asset return variance pertains to the variability of the unobservable shock term. As such, over a non-infinitesimal horizon it is necessary to first specify the conditional mean return (e. g., through an asset pricing model) in order to identify the conditional return variation. In contrast, over an infinitesimal time interval this is not necessary because the requirement that market prices do not admit arbitrage opportunities implies that return innovations are an order of magnitude larger than the mean return. This result has important implications for the approach we use to model and measure volatility in continuous time.

Consider an asset with log-price process $\{p(t)\ ,\ t \in [0, T]\}$ defined on a probability space $(\Omega, \mathcal{F}, P)$. Following Andersen et al. [19] we define the continuously compounded asset return over a time interval from $t - h$ to t, $0 \leq h \leq t \leq T$, to be

$$r(t, h) = p(t) - p(t - h)\,. \tag{5}$$

A special case of (5) is the cumulative return up to time t, which we denote $r(t) \equiv r(t, t) = p(t) - p(0)$, $0 \leq t \leq T$. Assume the asset trades in a frictionless market void of arbitrage opportunities and the number of potential discontinuities (jumps) in the price process per unit time is finite. Then the log-price process p is a semi-martingale (e. g., Back [33]) and therefore the cumulative return $r(t)$ admits the decomposition (e. g., Protter [229])

$$r(t) = \mu(t) + M^{\mathrm{C}}(t) + M^{\mathrm{J}}(t)\,, \tag{6}$$

where $\mu(t)$ is a predictable and finite variation process, $M^{\mathrm{C}}(t)$ a continuous-path infinite-variation martingale, and $M^{\mathrm{J}}(t)$ is a compensated finite activity jump martingale. Over a discrete time interval the decomposition (6) becomes

$$r(t, h) = \mu(t, h) + M^{\mathrm{C}}(t, h) + M^{\mathrm{J}}(t, h)\,, \tag{7}$$

where $\mu(t, h) = \mu(t) - \mu(t - h)$, $M^{\mathrm{C}}(t, h) = M^{\mathrm{C}}(t) - M^{\mathrm{C}}(t - h)$, and $M^{\mathrm{J}}(t, h) = M^{\mathrm{J}}(t) - M^{\mathrm{J}}(t - h)$.

Denote now with $[r, r]$ the quadratic variation of the semi-martingale process r, where (Protter [229])

$$[r, r]_t = r(t)^2 - 2\int r(s-)\mathrm{d}r(s)\,, \tag{8}$$

and $r(t-) = \lim_{s\uparrow t} r(s)$. If the finite variation process μ is continuous, then its quadratic variation is identically zero and the predictable component μ in decomposition (7) does not affect the quadratic variation of the return r. Thus, we obtain an expression for the quadratic return variation over the time interval from $t - h$ to t, $0 \leq h \leq t \leq T$ (e. g., Andersen et al. [21] and Barndorff-Nielsen and Shephard [51,52]):

$$\begin{aligned}
\mathrm{QV}(t, h) &= [r, r]_t - [r, r]_{t-h} \\
&= [M^{\mathrm{C}}, M^{\mathrm{C}}]_t - [M^{\mathrm{C}}, M^{\mathrm{C}}]_{t-h} \\
&\quad + \sum_{t-h<s\leq t} \Delta M^2(s) \\
&= [M^{\mathrm{C}}, M^{\mathrm{C}}]_t - [M^{\mathrm{C}}, M^{\mathrm{C}}]_{t-h} \\
&\quad + \sum_{t-h<s\leq t} \Delta r^2(s)\,.
\end{aligned} \tag{9}$$

Most continuous-time models for asset returns can be cast within the general setting of Eq. (7), and Eq. (9) provides a framework to study the model-implied return variance. For instance, the Black and Scholes [63] model is a special case of the setting described by Eq. (7) in which the conditional mean process μ is constant, the continuous martingale M^C is a standard Brownian motion process, and the jump martingale M^J is identically zero:

$$\mathrm{d}p(t) = \mu\mathrm{d}t + \sigma\mathrm{d}W(t)\,. \tag{10}$$

In this case, the quadratic return variation over the time interval from $t - h$ to t, $0 \leq h \leq t \leq T$, simplifies to

$$\mathrm{QV}(t, h) = \int_{t-h}^{t} \sigma^2\mathrm{d}s = \sigma^2 h\,, \tag{11}$$

that is, return volatility is constant over any time interval of length h.

A second notable example is the jump-diffusion model of Merton [214],

$$\mathrm{d}p(t) = (\mu - \lambda\bar{\xi})\mathrm{d}t + \sigma\mathrm{d}W(t) + \xi(t)\mathrm{d}q_t\,, \tag{12}$$

where q is a Poisson process uncorrelated with W and governed by the constant jump intensity λ, i. e., $\mathrm{Prob}(\mathrm{d}q_t = 1) = \lambda\mathrm{d}t$. The scaling factor $\xi(t)$ denotes the magnitude of the jump in the return process if a jump occurs at time t. It is assumed to be normally distributed,

$$\xi(t) \rightsquigarrow N(\bar{\xi}, \sigma_\xi^2)\,. \tag{13}$$

In this case, the quadratic return variation process over the time interval from $t-h$ to t, $0 \le h \le t \le T$ becomes

$$\begin{aligned}\mathrm{QV}(t,h) &= \int_{t-h}^{t} \sigma^2\mathrm{d}s + \sum_{t-h\le s\le t} J(s)^2 \\ &= \sigma^2 h + \sum_{t-h\le s\le t} J(s)^2\,,\end{aligned} \tag{14}$$

where $J(t) \equiv \xi(t)\mathrm{d}q(t)$ is non-zero only if a jump actually occurs.

Finally, a broad class of stochastic volatility models is defined by

$$\mathrm{d}p(t) = \mu(t)\mathrm{d}t + \sigma(t)\mathrm{d}W(t) + \xi(t)\mathrm{d}q_t\,, \tag{15}$$

where q is a constant-intensity Poisson process with log-normal jump amplitude (13). Equation (15) is also a special case of (7) and the associated quadratic return variation over the time interval from $t-h$ to t, $0 \le h \le t \le T$, is

$$\begin{aligned}\mathrm{QV}(t,h) &= \int_{t-h}^{t} \sigma(s)^2\mathrm{d}s + \sum_{t-h\le s\le t} J(s)^2 \\ &\equiv IV(t,h) + \sum_{t-h\le s\le t} J(s)^2\,.\end{aligned} \tag{16}$$

As in the general case of Eq. (9), Eq. (16) identifies the contribution of diffusive volatility, termed 'integrated variance' (IV), and cumulative squared jumps to the total quadratic variation.

Early applications typically ignored jumps and focused exclusively on the integrated variance component. For instance, IV plays a key role in Hull and White's [174] SV option pricing model, which we discuss in Sect. "Options" below along with other option pricing applications. For illustration, we focus here on the SV model specification by Wiggins [256]:

$$\mathrm{d}p(t) = \mu\mathrm{d}t + \sigma(t)\mathrm{d}W_p(t) \tag{17}$$

$$\mathrm{d}\sigma(t) = f(\sigma(t))\mathrm{d}t + \eta\sigma(t)\mathrm{d}W_\sigma(t)\,, \tag{18}$$

where the innovations to the return $\mathrm{d}p$ and volatility σ, W_p and W_σ, are standard Brownian motions. If we define $y = \log(\sigma)$ and apply Itô's formula we obtain

$$\begin{aligned}\mathrm{d}y(t) &= \mathrm{d}\log(\sigma(t)) \\ &= \left[-\frac{1}{2}\eta^2 + \frac{f(\sigma(t))}{\sigma(t)}\right]\mathrm{d}t + \eta\mathrm{d}W_\sigma(t)\,.\end{aligned} \tag{19}$$

Wiggins approximates the drift term $f(\sigma(t)) \approx \{\alpha + \kappa[\log(\overline{\sigma}) - \log(\sigma(t))]\}\sigma(t)$. Substitution in Eq. (19) yields

$$\mathrm{d}\log(\sigma(t)) = [\overline{\alpha} - \kappa\log(\sigma(t))]\mathrm{d}t + \eta\mathrm{d}W_\sigma(t)\,, \tag{20}$$

where $\overline{\alpha} = \alpha + \kappa\log(\overline{\sigma}) - \frac{1}{2}\eta^2$. As such, the logarithmic standard deviation process in Wiggins has diffusion dynamics similar in spirit to Taylor's discrete time AR(1) model for the logarithmic information process, Eq. (4). As in Taylor's model, negative correlation between return and volatility innovations, $\rho = \mathrm{corr}(W_p, W_\sigma) < 0$, generates an asymmetric response of volatility to return shocks similar to the leverage effect in discrete-time EGARCH models.

More recently, several authors have imposed restrictions on the continuous-time SV jump-diffusion (15) that render the model more tractable while remaining consistent with the empirical features of the data. We return to these models in Sect. "Options" below.

Realized Volatility

Model-free measures of return variation constructed only from concurrent return realizations have been considered at least since Merton [215]. French et al. [148] construct monthly historical volatility estimates from daily return observations. More recently, the increased availability of transaction data has made it possible to refine early measures of historical volatility into the notion of 'realized volatility', which is endowed with a formal theoretical justification as an estimator of the quadratic return variation as first noted in Andersen and Bollerslev [18]. The realized volatility of an asset return r over the time interval from $t-h$ to t is

$$RV(t,h;n) = \sum_{i=1}^{n} r\left(t - h + \frac{ih}{n}, \frac{h}{n}\right)^2\,. \tag{21}$$

Semi-martingale theory ensures that the realized volatility measure RV converges to the return quadratic variation QV, previously defined in Eq. (9), when the sampling frequency n increases. We point the interested reader to, e. g., Andersen et al. [19] to find formal arguments in support of

this claim. Here we convey intuition for this result by considering the special case in which the asset return follows a continuous-time diffusion without jumps,

$$\mathrm{d}p(t) = \mu(t)\mathrm{d}t + \sigma(t)\mathrm{d}W(t)\,. \tag{22}$$

As in Eq. (21), consider a partition of the $[t-h, t]$ interval with mesh h/n. A discretization of the diffusion (22) over a sub-interval from $(t-h+(i-1)h/n)$ to $(t-h+ih/n)$, $i = 1, \ldots, n$, yields

$$r\left(t-h+\frac{ih}{n}, \frac{h}{n}\right) \approx \mu\left(t-h+\frac{(i-1)h}{n}\right)\frac{h}{n} + \sigma\left(t-h+\frac{(i-1)h}{n}\right)\Delta W\left(t-h+\frac{ih}{n}\right)\,, \tag{23}$$

where $\Delta W(t-h+ih/n) = W(t-h+ih/n) - W(t-h+(i-1)h/n)$.

Suppressing time indices, the squared return r^2 over the time interval of length h/n is therefore:

$$r^2 = \mu^2\left(\frac{h}{n}\right)^2 + 2\mu\sigma\Delta W\left(\frac{h}{n}\right) + \sigma^2(\Delta W)^2\,. \tag{24}$$

As $n \to \infty$ the first two terms vanish at a rate higher than the last one. In particular, to a first order approximation the squared return equals the squared return innovation and therefore the squared return conditional mean and variance are

$$\mathrm{E}\left[r^2|\mathcal{F}_t\right] \approx \sigma^2\frac{h}{n} \tag{25}$$

$$\mathrm{Var}\left[r^2|\mathcal{F}_t\right] \approx 2\sigma^4\left(\frac{h}{n}\right)^2\,. \tag{26}$$

The no-arbitrage condition implies that return innovations are serially uncorrelated. Thus, summing over $i = 1, \ldots, n$ we obtain

$$\begin{aligned}\mathrm{E}\left[RV(t,h,n)|\mathcal{F}_t\right] &= \sum_{i=1}^{n}\mathrm{E}\left[r\left(t-h+\frac{ih}{n}, \frac{h}{n}\right)^2|\mathcal{F}_t\right] \\ &\approx \sum_{i=1}^{n}\sigma\left(t-h+\frac{(i-1)h}{n}\right)^2\frac{h}{n} \\ &\approx \int_{t-h}^{t}\sigma(s)^2\mathrm{d}s\end{aligned} \tag{27}$$

$$\begin{aligned}\mathrm{Var}\left[RV(t,h,n)|\mathcal{F}_t\right] &= \sum_{i=1}^{n}\mathrm{Var}\left[r\left(t-h+\frac{ih}{n}, \frac{h}{n}\right)^2|\mathcal{F}_t\right] \\ &\approx \sum_{i=1}^{n}2\sigma\left(t-h+\frac{(i-1)h}{n}\right)^4\left(\frac{h}{n}\right)^2 \\ &\approx 2\left(\frac{h}{n}\right)\int_{t-h}^{t}\sigma(s)^4\mathrm{d}s\,.\end{aligned} \tag{28}$$

Equation (27) illustrates that realized volatility is an unbiased estimator of the return quadratic variation, while Eq. (28) shows that the estimator is consistent as its variance shrinks to zero when we increase the sampling frequency n and keep the time interval h fixed. Taken together, these results suggest that RV is a powerful and model-free measure of the return quadratic variation. Effectively, RV gives practical empirical content to the latent volatility state variable underlying the models previously discussed in Sect. "Continuous-Time Stochastic Volatility Models".

Two issues complicate the practical application of the convergence results illustrated in Eqs. (27) and (28). First, a continuum of instantaneous return observations must be used for the conditional variance in Eq. (28) to vanish. In practice, only a discrete price record is observed, and thus an inevitable discretization error is present. Barndorff-Nielsen and Shephard [52] develop an asymptotic theory to assess the effect of this error on the RV estimate (see also [209]). Second, market microstructure effects (e. g., price discreteness, bid-ask spread positioning due to dealer inventory control, and bid-ask bounce) contaminate the return observations, especially at the ultra-high frequency. These effects tend to generate spurious correlations in the return series which can be partially eliminated by filtering the data prior to forming the RV estimates. However, this strategy is not a panacea and much current work studies the optimal sampling scheme and the construction of improved realized volatility in the presence of microstructure noise. This growing literature is surveyed by Hansen and Lunde [165], Bandi and Russell [46], McAleer and Medeiros [205], and Andersen and Benzoni [14]. Recent notable contributions to this literature include Bandi and Russell [45], Barndorff-Nielsen et al. [49], Diebold and Strasser [121], and Zhang, Mykland, and Aï t-Sahalia [262]. Related, there is the issue of how to construct RV measures when the market is rather illiquid. One approach is to use a lower sampling frequency and focus on longer-horizon RV measure. Alternatively the literature has explored volatility measures that are more robust to situations in which the noise-to-

signal ratio is high, e. g., Alizadeh et al. [8], Brandt and Diebold [72], Brandt and Jones [73], Gallant et al. [151], Garman and Klass [157], Parkinson [221], Schwert [237], and Yang and Zhang [259] consider the high-low price range measure. Dobrev [122] generalizes the range estimator to high-frequency data and shows its link with RV measures.

Equations (27) and (28) also underscore an important difference between RV and other volatility measures. RV is an ex-post model-free estimate of the quadratic variation process. This is in contrast to ex-ante measures which attempt to forecast future quadratic variation using information up to current time. The latter class includes parametric GARCH-type volatility forecasts as well as forecasts built from stochastic volatility models through, e. g., the Kalman filter (e. g., [167,168]), the particle filter (e. g., [186,187]) or the reprojection method (e. g., [152,155]).

More recently, other studies have pursued more direct time-series modeling of volatility to obtain alternative ex-ante forecasts. For instance, Andersen et al. [21] follow an ARMA-style approach, extended to allow for long memory features, to model the logarithmic foreign exchange rate realized volatility. They find the fit to dominate that of traditional GARCH-type models estimated from daily data. In a related development, Andersen, Bollerslev, and Meddahi [24,25] exploit the general class of Eigenfunction Stochastic Volatility (ESV) models introduced by Meddahi [208] to provide optimal analytic forecast formulas for realized volatility as a function of past realized volatility. Other scholars have pursued more general model specifications to improve forecasting performance. Ghysels et al. [159] consider Mixed Data Sampling (MIDAS) regressions that use a combination of volatility measures estimates at different frequencies and horizons. Related, Engle and Gallo [137] exploit the information in different volatility measures, captured by a multivariate extension of the multiplicative error model suggested by Engle [136], to predict multi-step volatility. Finally, Andersen et al. [20] build on the Heterogeneous AutoRegressive (HAR) model by Barndorff-Nielsen and Shephard [50] and Corsi [110] and propose a HAR-RV component-based regression to forecast the h-steps ahead quadratic variation:

$$\begin{aligned}\mathrm{RV}(t+h,h) = \beta_0 + \beta_D \mathrm{RV}(t,1) + \beta_W \mathrm{RV}(t,5) \\ + \beta_M \mathrm{RV}(t,21) + \varepsilon(t+h)\,. \end{aligned} \tag{29}$$

Here the lagged volatility components $\mathrm{RV}(t,1)$, $\mathrm{RV}(t,5)$, and $\mathrm{RV}(t,21)$ combine to provide a parsimonious approximation to the long-memory type behavior of the realized volatility series, which has been documented in several studies (e. g., Andersen et al. [19]). Simple OLS estimation yields consistent estimates for the coefficients in the regression (29), which can be used to forecast volatility out of sample.

As mentioned previously, the convergence results illustrated in Eqs. (27) and (28) stem from the theory of semi-martingales under conditions more general than those underlying the continuous-time diffusion in Eq. (22). For instance, these results are robust to the presence of discontinuities in the return path as in the jump-diffusion SV model (15). In this case the realized volatility measure (21) still converges to the return quadratic variation, which is now the sum of the diffusive integrated volatility IV and the cumulative squared jump component:

$$\mathrm{QV}(t,h) = \mathrm{IV}(t,h) + \sum_{t-h\le s\le t} J(s)^2\,. \tag{30}$$

The decomposition in Eq. (30) motivates the quest for separate estimates of the two quadratic variation components, IV and squared jumps. This is a fruitful exercise in forecasting applications, since separate estimation of the two components increases predictive accuracy (e. g., [20]). Further, this decomposition is relevant for derivatives pricing, e. g., options are highly sensitive to jumps as well as large moves in volatility (e. g., [141,220]).

A consistent estimate of integrated volatility is the k-skip bipower variation, BV (e. g., Barndorff-Nielsen and Shephard [53]),

$$\begin{aligned}\mathrm{BV}(t,h;k,n) = \frac{\pi}{2}\sum_{i=k+1}^{n}\left|r\left(t-h+\frac{ih}{n},\frac{h}{n}\right)\right| \\ \times\left|r\left(t-h+\frac{(i-k)h}{n},\frac{h}{n}\right)\right|\,. \end{aligned} \tag{31}$$

Liu and Maheu [202] and Forsberg and Ghysels [147] show that realized power variation, which is robust to the presence of jumps, can improve volatility forecasts. A well-known special case of (31) is the 'realized bipower variation', which has $k=1$ and is denoted $\mathrm{BV}(t,h;n) \equiv \mathrm{BV}(t,h;1,n)$. We can combine bipower variation with the realized volatility RV to obtain a consistent estimate of the squared jump component, i. e.,

$$\begin{aligned}\mathrm{RV}(t,h;n) - \mathrm{BV}(t,h;n) \underset{n\to\infty}{\longrightarrow} \mathrm{QV}(t,h) - \mathrm{IV}(t,h) \\ = \sum_{t-h\le s\le t} J(s)^2\,. \end{aligned} \tag{32}$$

The result in Eq. (32) are useful to design tests for the presence of jumps in volatility, e. g., Andersen et al. [20], Barndorff-Nielsen and Shephard [53,54], Huang and

Tauchen [172], and Mizrach [217]. More recently, alternative approaches to test for jumps have been developed by Aït-Sahalia and Jacod [6], Andersen et al. [23], Lee and Mykland [195], and Zhang [261].

Applications

The power of the continuous-time paradigm has been evident ever since the work by Merton [212] on intertemporal portfolio choice, Black and Scholes [63] on option pricing, and Vasicek [255] on bond valuation. However, the idea of casting these problems in a continuous-time diffusion context goes back all the way to the work in 1900 by Bachelier [32].

Merton [213] develops a continuous-time general-equilibrium intertemporal asset pricing model which is later extended by Cox et al. [112] to a production economy. Because of its flexibility and analytical tractability, the Cox et al. [112] framework has become a key tool used in several financial applications, including the valuation of options and other derivative securities, the modeling of the term structure of risk-free interest rates, the pricing of foreign currencies and defaultable bonds.

Volatility has played a central role in these applications. For instance, an option's payoff is non-linear in the price of the underlying asset and this feature renders the option value highly sensitive to the volatility of underlying returns. Further, derivatives markets have grown rapidly in size and complexity and financial institutions have been facing the challenge to manage intricate portfolios exposed to multiple risk sources. Risk management of these sophisticated positions hinges on volatility modeling. More recently, the markets have responded to the increasing hedging demands of investors by offering a menu of new products including, e. g., volatility swaps and derivatives on implied volatility indices like the VIX. These innovations have spurred an even more pressing need to accurately measure and forecast volatility in financial markets.

Research has responded to these market developments. We next provide a brief illustrative overview of the recent literature dealing with option pricing and term structure modeling, with an emphasis on the role that volatility modeling has played in these two key applications.

Options

Rubinstein [233] and Bates [55], among others, note that prior to the 1987 market crash the Black and Scholes [63] (BS) formula priced option contracts quite accurately whereas after the crash it has been systematically underpricing out-of-the-money equity-index put contracts. This feature is evident from Fig. 1, which is constructed from options on the S&P 500 futures. It shows the implied volatility function for near-maturity contracts traded both before and after October 19, 1987 ('Black Monday'). The mild u-shaped pattern prevailing in the pre-crash implied volatilities is labeled a 'volatility smile,' in contrast to the asymmetric post-1987 'volatility smirk'. Importantly, while the steepness and level of the implied volatility curve fluctuate day to day depending on market conditions, the curve has been asymmetric and upward sloping ever since 1987, so the smirk remains in place to the current date, e. g., Benzoni et al. [60]. In contrast, before the crash the implied volatility curve was invariably flat or mildly u-shaped as documented in, e. g., [57]. Finally, we note that the post-1987 asymmetric smirk for index options contrasts sharply with the pattern for individual equity options, which possess flat or mildly u-shaped implied volatility curves (e. g., [37,65]).

Given the failures of the BS formula, much research has gone into relaxing the underlying assumptions. A natural starting point is to allow volatility to evolve randomly, inspiring numerous studies that examine the option pricing implications of SV models. The list of early contributions includes [174,188,211,238,244,245,256]. Here we focus in particular on the Hull and White [174] model,

$$\mathrm{d}p(t) = \mu_p \mathrm{d}t + \sqrt{V(t)}\mathrm{d}W_p(t) \tag{33}$$

$$\frac{\mathrm{d}V(t)}{V(t)} = \mu_V \mathrm{d}t + \sigma_V \mathrm{d}W_V(t)\,, \tag{34}$$

where W_p and W_V are standard Brownian motions. In general, shocks to returns and volatility may be (negatively) correlated, however for tractability Hull and White assume $\rho = \mathrm{corr}(\mathrm{d}W_p, \mathrm{d}W_V) = 0$. Under this assumption they show that, in a risk-neutral world, the premium C^{HW} on a European call option is the Black and Scholes price C^{BS} evaluated at the average integrated variance $\overline{V}$,

$$\overline{V} = \frac{1}{T-t}\int_t^T V(s)\mathrm{d}s\,, \tag{35}$$

integrated over the distribution $h(\overline{V}|V(t))$ of $\overline{V}$:

$$C^{\mathrm{HW}}(p(t), V(t)) = \int C^{\mathrm{BS}}(\overline{V})h(\overline{V}|V(t))\mathrm{d}\overline{V}\,. \tag{36}$$

The early efforts to identify a more realistic probabilistic model for the underlying return were slowed by the analytical and computational complexity of the option pricing problem. Unlike the BS setting, the early SV specifications do not admit closed-form solutions. Thus, the evaluation of the option price requires time-consuming computations through, e. g., simulation methods or nu-

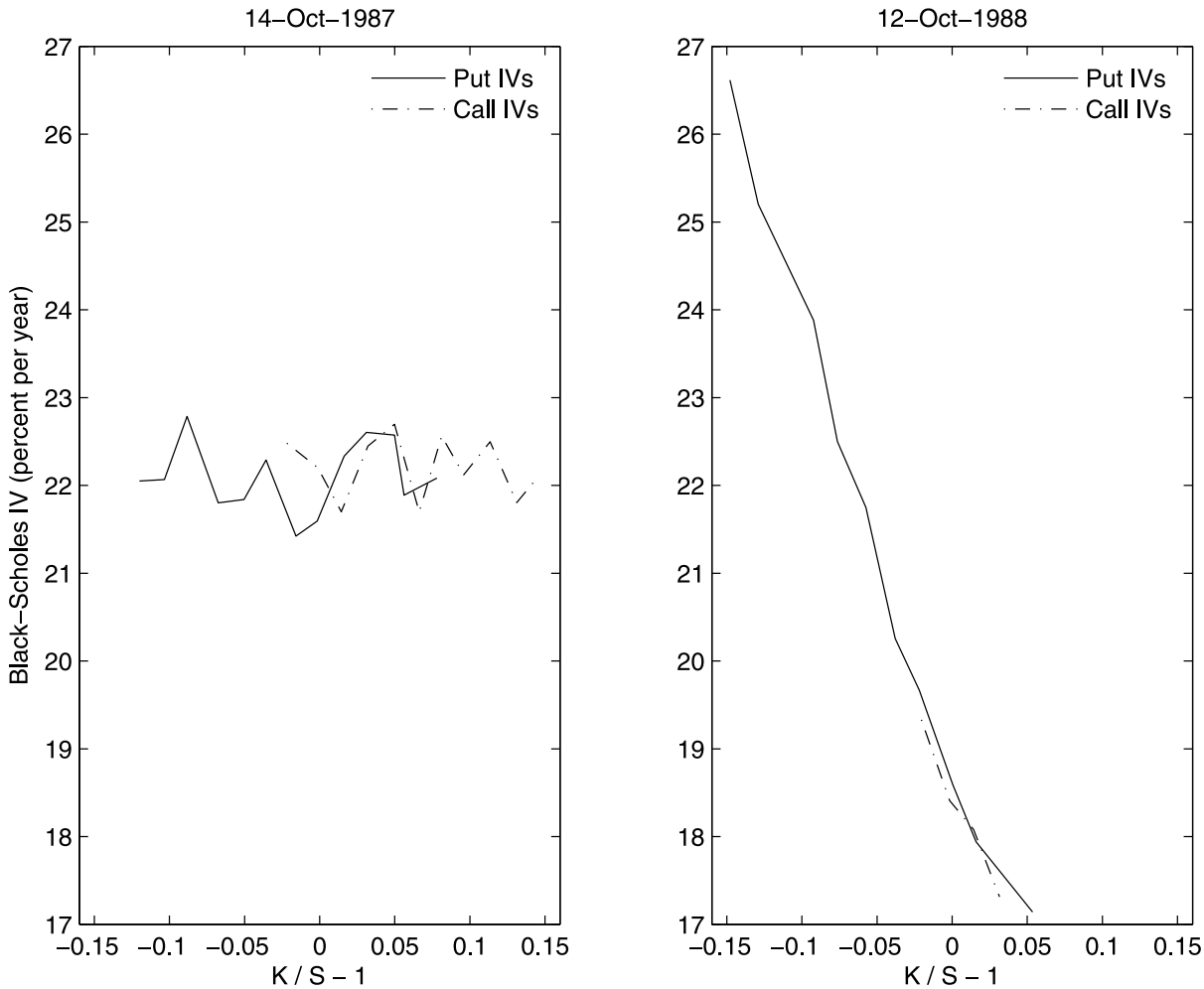

Stochastic Volatility, Figure 1

Pre- and post-1987 crash implied volatilities. The plots depict Black-Scholes implied volatilities computed from near-maturity options on the S&P 500 futures on October 14, 1987 (the week before the 1987 market crash) and a year later

merical solution of the pricing partial differential equation by finite difference methods. Further, the presence of a latent factor, volatility, and the lack of closed-form expressions for the likelihood function complicate the estimation problem.

Consequently, much effort has gone into developing restrictions for the distribution of the underlying return process that allow for (semi) closed-form solutions and are consistent with the empirical properties of the data. The 'affine' class of continuous-time models has proven particularly useful in providing a flexible, yet analytically tractable, setting. Roughly speaking, the defining feature of affine jump-diffusions is that the drift term, the conditional covariance term, and the jump intensity are all a linear-plus-constant (affine) function of the state vector. The Vasicek [255] bond valuation model and the Cox et al. [112] intertemporal asset pricing model provide powerful examples of the advantages of the affine paradigm.

To illustrate the progress in option pricing applications built on affine models, consider the return dynamics

$$\mathrm{d}p(t) = \mu \mathrm{d}t + \sqrt{V(t)}\mathrm{d}W_p(t) + \xi_p(t)\mathrm{d}q(t) \tag{37}$$

$$\mathrm{d}V(t) = \kappa(\overline{V} - V(t))\mathrm{d}t + \sigma_V \sqrt{V(t)}\mathrm{d}W_V(t) + \xi_V(t)\mathrm{d}q(t)\,, \tag{38}$$

where W_p and W_V are standard Brownian motions with non-zero correlation $\rho = \mathrm{corr}(\mathrm{d}W_p, \mathrm{d}W_V)$, q is a Poisson process, uncorrelated with W_p and W_V, with jump inten-

sity

$$\lambda(t) = \lambda_0 + \lambda_1 V(t)\,, \tag{39}$$

that is, $\mathrm{Prob}(\mathrm{d}q_t = 1) = \lambda(t)\mathrm{d}t$. The jump amplitudes variables ξ_p and ξ_V have distributions

$$\xi_V(t) \rightsquigarrow \exp(\overline{\xi}_V) \tag{40}$$

$$\xi_p(t)|\xi_V(t) \rightsquigarrow N\left(\overline{\xi}_p + \rho_\xi \xi_V(t), \sigma_p^2\right)\,. \tag{41}$$

Here volatility is not only stochastic but also subject to jumps which occur simultaneously with jumps in the underlying return process. The Black and Scholes model is a special case of (37)–(41) for constant volatility, $V(t) = \sigma^2, 0 \le t \le T$, and no jumps, $\lambda(t) = 0, 0 \le t \le T$. The Merton [214] model arises from (37)–(41) if volatility is constant but we allow for jumps in returns.

More recently, Heston [170] has considered a special case of (37)–(41) with stochastic volatility but without jumps. Using transform methods he derives a European option pricing formula which may be evaluated readily through simple numerical integration. His SV model has GARCH-type features, in that the variance is persistent and mean reverts at a rate κ to the long-run mean $\overline{V}$. Compared to Hull and White's [174] setting, Heston's model allows for shocks to returns and volatility to be negatively correlated, i. e., $\rho < 0$, which creates a leverage-type effect and skews the return distribution. This feature is consistent with the properties of equity index returns. Further, a fatter left tail in the return distribution results in a higher cost for crash insurance and therefore makes out-of-the-money put options more expensive. This is qualitatively consistent with the patterns in implied volatilities observed after the 1987 market crash and discussed above.

Bates [56] has subsequently extended Heston's approach to allow for jumps in returns and using similar transform methods he has obtained a semi-closed form solution for the option price. The addition of jumps provides a more realistic description of equity returns and has important option pricing implications. With diffusive shocks (e. g., stochastic volatility) alone a large drop in the value of the underlying asset over a short time span is very unlikely whereas a market crash is always possible as long as large negative jumps can occur. This feature increases the value of a short-dated put option, which offers downside protection to a long position in the underlying asset.

Finally, Duffie et al. [130] have introduced a general model with jumps to volatility which embeds the dynamics (37)–(41). In model (37)–(41), the likelihood of a jump to occur increases when volatility is high ($\lambda_1 > 0$) and a jump in returns is accompanied by an outburst of volatility. This is consistent with what is typically observed during times of market stress. As in the Heston case, variance is persistent with a mean reversion coefficient κ towards its *diffusive* long-run mean $\overline{V}$, while the total long-run variance mean is the sum of the diffusive and jump components. In the special case of constant jump intensity, i. e., $\lambda_1 = 0$, the total long-run mean is $\overline{V} + \overline{\xi}_V \lambda_0/\kappa$. The jump term $(\xi_V(t)\mathrm{d}q(t))$ fattens the right tail of the variance distribution, which induces leptokurtosis in the return distribution. Two effects generate asymmetrically distributed returns. The first channel is the diffusive leverage effect, i. e., $\rho < 0$, the second is the correlation between the volatility and the jump amplitude of returns generated through the coefficient ρ_ξ. Taken together, these effects increase model-implied option prices and help produce a realistic volatility smirk.

Several empirical studies rely on models of the form (37)–(41) in option-pricing applications. For instance, Bates [56] uses Deutsche Mark options to estimate a model with stochastic volatility and constant-intensity jumps to returns, while Bates [57] fits a jump-diffusion model with two SV factors to options on S&P 500 futures. In the latter case, the two SV factors combine to help capture features of the long-run memory in volatility while retaining the analytical tractability of the affine setting (see, e. g., [101] for another model with similar features). Alternative approaches to model long memory in continuous-time SV models rely on the fractional Brownian motion process, e. g., Comte and Renault [108] and Comte et al. [107], while Breidt et al. [76], Harvey [166] and Deo et al. [118] consider discrete-time SV models (see [175] for a review). Bakshi et al. [34,37] estimate a model similar to the one introduced by Bates [56] using S&P 500 options.

Other scholars rely on underlying asset return data alone for estimation. For instance, Andersen et al. [15] and Chernov et al. [95] use equity-index returns to estimate jump-diffusion SV models within and outside the affine (37)–(41) class. Eraker et al. [142] extend this analysis and fit a model that includes constant-intensity jumps to returns and volatility.

Finally, another stream of work examines the empirical implications of SV jump-diffusions using a joint sample of S&P 500 options and index returns. For example, Benzoni [59], Chernov and Ghysels [93], and Jones [190] estimate different flavors of the SV model without jumps. Pan [220] fits a model that has jumps in returns with time-varying intensity, while Eraker [141] extends Pan's work by adding jumps in volatility.

Overall, this literature has established that the SV jump-diffusion model dramatically improves the fit of underlying index returns and options prices compared to the

Black and Scholes model. Stochastic volatility alone has a first-order effect and jumps further enhance model performance by generating fatter tails in the return distribution and reducing the pricing error for short-dated options. The benefits of the SV setting are also significant in hedging applications.

Another aspect related to the specification of SV models concerns the pricing of volatility and jump risks. Stochastic volatility and jumps are sources of uncertainty. It is an empirical issue to determine whether investors demand to be compensated for bearing such risks and, if so, what the magnitude of the risk premium is. To examine this issue it is useful to write model (37)–(41) in so-called risk-neutral form. It is common to assume that the volatility risk premium is proportional to the instantaneous variance, $\eta(t) = \eta_V V(t)$. Further, the adjustment for jump risk is accomplished by assuming that the amplitude $\tilde{\xi}_p(t)$ of jumps to returns has mean $\overline{\tilde{\xi}}_p = \overline{\xi}_p + \eta_p$. These specifications are consistent with an arbitrage-free economy. More general specifications can also be supported in a general equilibrium setting, e. g., a risk adjustment may apply to the jump intensity $\lambda(t)$. However, the coefficients associated to these risk adjustments are difficult to estimate and to facilitate identification they typically are fixed at zero. Incorporating such risk premia in model (37)–(41) yields the following risk-neutral return dynamics (e. g., Pan [220] and Eraker [141]):

$$\mathrm{d}p(t) = (r - \mu^*)\mathrm{d}t + \sqrt{V(t)}\mathrm{d}\widetilde{W}_p(t) + \tilde{\xi}_p(t)\mathrm{d}q(t) \quad (42)$$

$$\mathrm{d}V(t) = [\kappa(\overline{V} - V(t)) + \eta_V V(t)]\mathrm{d}t + \sigma_V \sqrt{V(t)}\mathrm{d}\widetilde{W}_V(t) + \xi_V(t)\mathrm{d}q(t)\,, \quad (43)$$

where r is the risk-free rate, μ^* a jump compensator term, $\widetilde{W}_p$ and $\widetilde{W}_V$ are standard Brownian motions under this so-called $\mathcal{Q}$ measure, and the risk-adjusted jump amplitude variable $\tilde{\xi}_p$ is assumed to follow the distribution,

$$\tilde{\xi}_p(t)|\xi_V(t) \rightsquigarrow N\left(\overline{\tilde{\xi}}_p + \rho_\xi \xi_V(t), \sigma_p^2\right)\,. \quad (44)$$

Several studies estimate the risk-adjustment coefficients η_V and η_p for different specifications of model (37)–(44); see, e. g., Benzoni [59], Broadie et al. [78], Chernov and Ghysels [93], Eraker [141], Jones [190], and Pan [220]. It is found that investors demand compensation for volatility and jump risks and these risk premia are important for the pricing of index options. This evidence is reinforced by other studies examining the pricing of volatility risk using less structured but equally compelling procedures. For instance, Coval and Shumway [111] find that the returns on zero-beta index option straddles (i. e., combinations of calls and puts that have offsetting covariances with the index) are significantly lower than the risk-free return. This evidence suggests that in addition to market risk at least a second factor (likely, volatility) is priced in the index option market. Similar conclusions are obtained by Bakshi and Kapadia [36], Buraschi and Jackwerth [79], and Broadie et al. [78].

Risk-Free Bonds and Their Derivatives

The market for (essentially) risk-free Treasury bonds is liquid across a wide maturity spectrum. No-arbitrage restrictions constrain the allowable dynamics in the cross-section of bond yields. Much work has gone into the development of tractable dynamic term structure models capable of capturing the salient time-series properties of interest rates while respecting such cross-sectional no-arbitrage conditions. The class of so-called 'affine' dynamic term structure models provides a flexible and arbitrage-free, yet analytically tractable, setting for capturing the dynamics of the term structure of interest rates. Following Duffie and Kan [129], Dai and Singleton [114,115], and Piazzesi [226], the short term interest rate, $y_0(t)$, is an affine (i. e., linear-plus-constant) function of a vector of state variables, $X(t) = \{x_i(t),\ i = 1, \ldots, N\}$:

$$y_0(t) = \delta_0 + \sum_{i=1}^{N} \delta_i x_i(t) = \delta_0 + \delta_X' X(t)\,, \quad (45)$$

where the state-vector X has risk-neutral dynamics

$$\mathrm{d}X(t) = \tilde{\mathcal{K}}(\tilde{\Theta} - X(t))\mathrm{d}t + \Sigma\sqrt{S(t)}\mathrm{d}\widetilde{W}(t)\,. \quad (46)$$

In Eq. (46), $\widetilde{W}$ is an N-dimensional Brownian motion under the so-called $\mathcal{Q}$-measure, $\tilde{\mathcal{K}}$ and $\tilde{\Theta}$ are $N \times N$ matrices, and $S(t)$ is a diagonal matrix with the ith diagonal element given by $[S(t)]_{ii} = \alpha_i + \beta_i' X(t)$. Within this setting, the time-t price of a zero-coupon bond with time-to-maturity τ is given by

$$P(t, \tau) = \mathrm{e}^{A(\tau) - B(\tau)' X(t)}\,, \quad (47)$$

where the functions $A(\tau)$ and $B(\tau)$ solve a system of ordinary differential equations (ODEs); see, e. g., Duffie and Kan [129]. Semi-closed form solutions are also available for bond derivatives, e. g., bond options as well as caps and floors (e. g., Duffie et al. [130]).

In empirical applications it is important to also establish the evolution of the state vector X under the physical probability measure $\mathcal{P}$, which is linked to the $\mathcal{Q}$-dynamics (46) through a market price of risk, $\Lambda(t)$. Following Dai

and Singleton [114] the market price of risk is often given by

$$\Lambda(t) = \sqrt{S(t)}\lambda \,, \tag{48}$$

where λ is an $N \times 1$ vector of constants. More recently, Duffee [127] proposed a broader 'essentially affine' class, which retains the tractability of standard models but, in contrast to the specification in Eq. (48), allows compensation for interest rate risk to vary independently of interest rate volatility. This additional flexibility proves useful in forecasting future yields. Subsequent generalization are in Duarte [124] and Cheridito et al. [92].

Litterman and Scheinkman [201] demonstrate that virtually all variation in US Treasury rates is captured by three factors, interpreted as changes in 'level', 'steepness', and 'curvature'. Consistent with this evidence, much of the term-structure literature has focused on three-factor models. One problem with these models, however, is that the factors are latent variables void of immediate economic interpretation. As such, it is challenging to impose appropriate identifying conditions for the model coefficients and in particular to find the ideal representation for the 'most flexible' model, i. e., the model with the highest number of identifiable coefficients. Dai and Singleton [114] conduct an extensive specification analysis of multi-factor affine term structure models. They classify these models into subfamilies according to the number of (independent linear combination of) state variables that determine the conditional variance matrix of the state vector. Within each subfamily, they proceed to identify the models that lead to well-defined bond prices (a condition they label 'admissibility') and among the admissible specifications they identify a 'maximal' model that nests econometrically all others in the subfamily. Joslin [191] builds on Dai and Singleton's [114] work by pursuing identification through a normalization of the drift term in the state vector dynamics (instead of the diffusion term, as in Dai and Singleton [114]). Duffie and Kan [129] follow an alternative approach to obtain an identifiable model by rotating from a set of latent state variables to a set of observable zero-coupon yields. Collin-Dufresne et al. [104] build on the insights of both Dai and Singleton [114] and Duffie and Kan [129]. They perform a rotation of the state vector into a vector that contains the first few components in the Taylor series expansion of the yield curve around a maturity of zero and their quadratic variation. One advantage is that the elements of the rotated state vector have an intuitive and unique economic interpretation (such as level, slope, and curvature of the yield curve) and therefore the model coefficients in this representation are identifiable. Further, it is easy to construct a model-independent proxy for the rotated state vector, which facilitates model estimation as well as interpretation of the estimated coefficients across models and sample periods.

This discussion underscores an important feature of affine term structure models. The dependence of the conditional factor variance $S(t)$ on one or more of the elements in X introduces stochastic volatility in the yields. However, when a square-root factor is present parametric restrictions (admissibility conditions) need to be imposed so that the conditional variance $S(t)$ is positive over the range of X. These restrictions affect the correlations among the factors which, in turn, tend to worsen the cross-sectional fit of the model. Specifically, CIR models in which $S(t)$ depends on all the elements of X require the conditional correlation among the factors to be zero, while the admissibility conditions imposed on the matrix $\mathcal{K}$ renders the unconditional correlations non-negative. These restrictions are not supported by the data. In contrast, constant-volatility Gaussian models with no square-root factors do not restrict the signs and magnitude of the conditional and unconditional correlations among the factors but they do, of course, not accommodate the pronounced and persistent volatility fluctuations observed in bond yields. The class of models introduced by Dai and Singleton [114] falls between these two extremes. By including both Gaussian and square-root factors they allow for time-varying conditional volatilities of the state variables and yet they do not constrain the signs of some of their correlations. This flexibility helps to address the trade off between generating realistic correlations among the factors while capturing the time-series properties of the yields' volatility.

A related aspect of (unconstrained) affine models concerns the dual role that square-root factors play in driving the time-series properties of yields' volatility and the term structure of yields. Specifically, the time-t yield $y_\tau(t)$ on a zero-coupon bond with time-to-maturity τ is given by

$$P(t,\tau) = \mathrm{e}^{-\tau y_\tau(t)} \,. \tag{49}$$

Thus, we have

$$y_\tau(t) = -\frac{A(\tau)}{\tau} + \frac{B(\tau)'}{\tau} X(t) \,. \tag{50}$$

It is typically assumed that the B matrix has full rank and therefore Eq. (50) provides a direct link between the state-vector $X(t)$ and the term-structure of bond yields. Further, Itô's Lemma implies that the yield y_τ also follows a diffusion process:

$$\mathrm{d}y_\tau(t) = \mu_{y_\tau}(X(t),t)\mathrm{d}t + \frac{B(\tau)'}{\tau}\Sigma\sqrt{S(t)}\mathrm{d}\widetilde{W}(t) \,. \tag{51}$$

Consequently, the (instantaneous) quadratic variation of the yield given as the squared yield volatility coefficient for y_τ is

$$V_{y_\tau}(t) = \frac{B(\tau)'}{\tau} \Sigma S(t) \Sigma' \frac{B(\tau)}{\tau} . \tag{52}$$

The elements of the $S(t)$ matrix are affine in the state vector $X(t)$, i. e., $[S(t)]_{ii} = \alpha_i + \beta_i' X(t)$. Further, invoking the full rank condition on $B(\tau)$, Eq. (50) implies that each state variable in the vector $X(t)$ is an affine function of the bond yields $Y(t) = \{y_{\tau_j}(t),\ j = 1, \ldots, J\}$. Thus, for any τ there is a set of constants $a_{\tau,j},\ j = 0, \ldots, J$, so that

$$V_{y_\tau}(t) = a_{\tau,0} + \sum_{j=1}^{J} a_{\tau,j} y_{\tau_j}(t) . \tag{53}$$

Hence, the current quadratic yield variation for bonds at any maturity is a linear combination of the term structure of yields. As such, the market is complete, i. e., volatility is perfectly spanned by a portfolio of bonds.

Collin-Dufresne and Goldstein [103] note that this spanning condition is unnecessarily restrictive and propose conditions which ensures that volatility no longer directly enters the main bond pricing Eq. (47). This restriction, which they term 'unspanned stochastic volatility' (USV), effectively breaks the link between the yields' quadratic variation and the level of the term structure by imposing a reduced rank condition on the $B(\tau)$ matrix. Further, since their model is a special (nested) case of the affine class it retains the analytical tractability of the affine model class. Recently Joslin [191] has derived more general conditions for affine term structure models to exhibit USV. His restrictions also produce a market incompleteness (i. e., volatility cannot be hedged using a portfolio of bonds) but do not constrain the degree of mean reversion of the other state variables so that his specification allows for more flexibility in capturing the persistence in interest rate series. (See also the USV conditions in the work by Trolle and Schwartz [253]).

There is conflicting evidence on the volatility spanning condition in fixed income markets. Collin-Dufresne and Goldstein [103] find that swap rates have limited explanatory power for returns on at-the-money 'straddles', i. e., portfolios mainly exposed to volatility risk. Similar findings are in Heidari and Wu [169], who show that the common factors in LIBOR and swap rates explain only a limited part of the variation in the swaption implied volatilities. Moreover, Li and Zhao [197] conclude that some of the most sophisticated multi-factor dynamic term structure models have serious difficulties in hedging caps and cap straddles, even though they capture bond yields well. In contrast, Fan et al. [143] argue that swaptions and even swaption straddles can be well hedged with LIBOR bonds alone, supporting the notion that bond markets are complete.

More recently other studies have examined several versions of the USV restriction, again coming to different conclusions. A direct comparison of these results, however, is complicated by differences in the model specification, the estimation method, and the data and sample period used in the estimation. Collin-Dufresne et al. [105] consider swap rates data and fit the model using a Bayesian Markov Chain Monte Carlo method. They find that a standard three-factor model generates a time series for the variance state variable that is essentially unrelated to GARCH estimates of the quadratic variation of the spot rate process or to implied variances from options, while a four-factor USV model generates both realistic volatility estimates and a good cross-sectional fit. In contrast, Jacobs and Karoui [178] consider a longer data set of US Treasury yields and pursue quasi-maximum likelihood estimation. They find the correlation between model-implied and GARCH volatility estimates to be high. However, when estimating the model with a shorter sample of swap rates, they find such correlations to be small or negative. Thompson [250] explicitly tests the Collin-Dufresne and Goldstein [103] USV restriction and rejects it using swap rates data. Bikbov and Chernov [62], Han [164], Jarrow et al. [183], Joslin [192], and Trolle and Schwartz [254] rely on data sets of derivatives prices and underlying interest rates to better identify the volatility dynamics.

Andersen and Benzoni [12] directly relate model-free realized volatility measures (constructed from high-frequency US Treasury data) to the cross-section of contemporaneous bond yields. They find that the explanatory power of such regressions is very limited, which indicates that volatility is not spanned by a portfolio of bonds. The evidence in Andersen and Benzoni [12] is consistent with the USV models of Collin-Dufresne et al. [105] and Joslin [191], as well as with a model that embeds weak dependence between the yields and volatility as in Joslin [192]. Moreover, Duarte [125] argues that the effects of mortgage-backed security hedging activity affects both the interest rate volatility implied by options and the actual interest rate volatility. This evidence suggests that variables that are not in the span of the term structure of yields and forward rates contribute to explain volatility in fixed income markets. Also related, Wright and Zhou [258] find that adding a measure of market jump volatility risk to a regression of excess bond returns on the term structure of forward rates nearly doubles the R^2 of the regression.

Taken together, these findings suggest more generally that genuine SV models are critical for appropriately capturing the dynamic evolution of the term structure.

Estimation Methods

There are a very large number of alternative approaches to estimation and inference for parametric SV models and we abstain from a thorough review. Instead, we point to the basic challenges that exist for different types of specifications, how some of these were addressed in the early literature and finally provide examples of methods that have been used extensively in recent years. Our exposition continues to focus on applications to equity returns, interest rates, and associated derivatives.

Many of the original SV models were cast in discrete time, inspired by the popular GARCH paradigm. In that case, the distinct challenge for SV models is the presence of a strongly persistent latent state variable. However, more theoretically oriented models, focusing on derivatives applications, were often formulated in continuous time. Hence, it is natural that the econometrically-oriented literature has moved in this direction in recent years as well. This development provides an added complication as the continuous-time parameters must be estimated from discrete return data and without direct observations on volatility. For illustration, consider a fully parametric continuous-time SV model for the asset return r with conditional variance V and coefficient vector Ψ. Most methods to estimate Ψ rely on the conditional density f for the data generating process,

$$f(r(t), V(t)|\mathcal{I}(t-1), \Psi) = f_{r|V}(r(t)|V(t), \mathcal{I}(t-1), \Psi) \times f_V(V(t)|\mathcal{I}(t-1), \Psi) , \quad (54)$$

where $\mathcal{I}(t-1)$ is the available information set at time $t-1$. The main complications are readily identified. First, analytic expressions for the discrete-time transition (conditional) density, f, or the discrete-time moments implied by the data generating process operating in continuous time, are often unavailable. Second, volatility is latent in SV models, so that even if a closed-form expression for f is known, direct evaluation of the above expression is infeasible due to the absence of explicit volatility measures. The marginal likelihood with respect to the observable return process alone is obtained by integrating over all possible paths for the volatility process, but this integral has a dimension corresponding to sample size, rendering the approach infeasible in general.

Similar issues are present when estimating continuous-time dynamic term structure models. Following Piazzesi [227], a change of variable gives the conditional density for a zero-coupon yield y on a bond with time to maturity τ:

$$f(y_\tau(t)|\mathcal{I}(t-1), \Psi) = f_X(g(y_\tau(t), \Psi)|\mathcal{I}(t-1), \Psi) \times |\nabla_y g(y_\tau(t), \Psi)| . \quad (55)$$

Here the latent state vector X has conditional density f_X, the function $g(\cdot, \Psi)$ maps the observable yield y into X, $X(t) = g(y_\tau(t), \Psi)$, and $\nabla_y g(y_\tau(t), \Psi)$ is the Jacobian determinant of the transformation. Unfortunately, analytic expressions for the conditional density f_X are known only in some special cases. Further, the mapping $X(t) = g(y_\tau(t), \Psi)$ holds only if the model provides an exact fit to the yields, while in practice different sources of error (e. g., model mis-specification, microstructure effects, measurement errors) inject a considerable degree of noise into this otherwise deterministic linkage (for correct model specification) between the state vector and the yields. As such, a good measure of X might not be available to evaluate the conditional density (55).

Estimation via Discrete-Time Model Specification or Approximation

The first empirical studies have estimated discrete-time SV models via a (Generalized) Method of Moments procedure by matching a number of theoretical and sample moments, e. g., Chan et al. [89], Ho et al. [171], Longstaff and Schwartz [204], and Melino and Turnbull [211]. These models were either explicitly cast in discrete time or were seen as approximate versions of the continuous-time process of interest. Similarly, several authors estimate diffusive affine dynamic term structure models by approximating the continuous-time dynamics with a discrete-time process. If the error terms are stipulated to be normally distributed, the transition density of the discretized process is multivariate normal and computation of unconditional moments then only requires knowledge of the first two moments of the state vector. This result facilitates quasi-maximum likelihood estimation. In evaluating the likelihood function, some studies suggest using closed-form expressions for the first two moments of the continuous-time process instead of the moments of the discretized process (e. g., Fisher and Gilles [145] and Duffee [127]), thus avoiding the associated discretization bias. This approach typically requires some knowledge of the state of the system which may be obtained, imperfectly, through matching the system, given the estimated parameter vector, to a set of observed zero-coupon yields to infer the state vector X. A modern alternative is to use the so-

called particle filter as an efficient filtering procedure for the unobserved state variables given the estimated parameter vector. We provide more detailed accounts of both of these procedures later in this section.

Finally, a number of authors develop a simulated maximum likelihood method that exploit the specific structure of the discrete-time SV model. Early examples are Danielsson and Richard [117] and Danielsson [116] who exploit the Accelerated Gaussian Importance Sampler for efficient Monte Carlo evaluation of the likelihood. Subsequent improvements were provided by Fridman and Harris [149] and Liesenfeld and Richard [200], with the latter relying on Efficient Importance Sampling (EIS). In a second step, EIS can also be used for filtering the latent volatility state vector. In general, these inference techniques provide quite impressive efficiency but the methodology is not always easy to generalize beyond the structure of the basic discrete-time SV asset return model. We discuss the general inference problem for continuous-time SV models for which the lack of a closed-form expression for the transition density is an additional complicating factor in a later section.

Filtering the Latent State Variable Directly During Estimation

Some early studies focused on direct ways to extract estimates of the latent volatility state variable in discrete-time SV asset return models. The initial approach was based on quasi-maximum likelihood (QML) methods exploiting the Kalman filter. This method requires a transformation of the SV model to a linear state-space form. For instance, Harvey and Shephard [168] consider a version of the Taylor's [249] discrete-time SV model,

$$p(t) = p(t-1) + \beta + \sqrt{V(t)}\varepsilon(t) \tag{56}$$

$$\log(V(t)) = \alpha + \phi \log(V(t-1)) + \eta(t)\,, \tag{57}$$

where p is the logarithmic price, ε is a zero-mean error term with unit variance, and η is an independently-distributed error term with zero mean and variance σ_η^2.

Define $y(t) = p(t) - p(t-1) - \beta$, square the observations in Eq. (56), and take logarithms to obtain the *measurement equation*,

$$\ell(t) = \omega + h(t) + \xi(t)\,, \tag{58}$$

where $\ell(t) \equiv \log y(t)^2, h(t) \equiv \log(V(t))$. Further, ξ is a zero-mean disturbance term given by $\xi(t) = \log(\varepsilon(t)^2) - \mathrm{E}[\log(\varepsilon(t)^2)]$, $\omega = \log(\sigma^2) + \mathrm{E}[\log(\varepsilon(t)^2)]$, and σ is a scale constant which subsumes the effect of the drift term α in Eq. (57). The autoregression (57) yields the *transition equation*,

$$h(t) = \phi h(t-1) + \eta(t)\,, \tag{59}$$

Taken together, Eqs. (58) and (59) are the linear state-space transformation of the SV model (56)–(57). If the joint distribution of ε and η is symmetric, i. e., $f(\varepsilon,\eta) = f(-\varepsilon,-\eta)$, then the disturbance terms in the state-space form are uncorrelated even if η and ε are not. A possible dependence between ε and η allows the model to pick up some of the asymmetric behavior often observed in stock returns. Projection of $[h(t) - \mathrm{E}_{t-1}\, h(t)]$ over $[\,\ell(t) - \mathrm{E}_{t-1}\,\ell(t)\,]$ yields the Kalman filter estimate of the latent (logarithmic) variance process:

$$\begin{aligned}\mathrm{E}_t\, h(t) &= \mathrm{E}_{t-1}\, h(t) \\ &\quad + \frac{\mathrm{E}\{[h(t) - \mathrm{E}_{t-1}\, h(t)] \times [\ell(t) - \mathrm{E}_{t-1}\,\ell(t)]\}}{\mathrm{E}\{[\ell(t) - \mathrm{E}_{t-1}\,\ell(t)]^2\}} \\ &\quad \times [\ell(t) - \mathrm{E}_{t-1}\,\ell(t)]\,,\end{aligned} \tag{60}$$

where the conditional expectations $\mathrm{E}_{t-1}\,\ell(t)$ and $\mathrm{E}_{t-1}\, h(t)$ are given by:

$$\mathrm{E}_{t-1}\,\ell(t) = \omega + \mathrm{E}_{t-1}\, h(t) \tag{61}$$

$$\mathrm{E}_{t-1}\, h(t) = \phi\, \mathrm{E}_{t-1}\, h(t-1)\,. \tag{62}$$

To start the recursion (60)–(62), the initial value $\mathrm{E}_0\, h(0)$ is fixed at the long-run mean $\log(\overline{V})$.

Harvey and Shephard [168] estimate the model coefficients via quasi-maximum likelihood, i. e. by treating the errors ξ and η as though they were normal and maximizing the prediction-error decomposition form of the likelihood function obtained via the Kalman filter. Inference is valid as long as the standard errors are appropriately adjusted. In their application they rely on daily returns on the value-weighted US market index over 1967–1987 and daily returns for 30 individual stocks over 1974–1983. Harvey et al. [167] pursue a similar approach to fit a multivariate SV model to a sample of four exchange rate series from 1981 to 1985. One major drawback of the Kalman filter approach is that the finite sample properties can be quite poor because the error term, ξ, is highly non-Gaussian, see, e. g., Andersen, Chung, and Sørensen [27]. The method may be extended to accommodate various generalizations including long memory persistence in volatility as detailed in Ghysels, Harvey, and Renault [158].

A related literature, often exploited in multivariate settings, specifies latent GARCH-style dynamics for a state vector which governs the systematic evolution of a higher dimensional set of asset returns. An early representative of these specifications is in Diebold and Nerlove [120], who exploit the Kalman filter for estimation, while Fiorentina et al. [144] provide a likelihood-based estimation procedure using MCMC techniques. We later review the MCMC approach and the associated filtering application, e. g, the 'particle filter', in some detail.

The state-space form is also useful to characterize the dynamics of interest rates. Following, e. g., Piazzesi [226], for a discrete-time dynamic term structure model the measurement and transition equations are

$$y_\tau(t) = -\frac{A(\tau)}{\tau} + \frac{B(\tau)'}{\tau} X(t) + \xi_\tau(t) \tag{63}$$

$$X(t) = \mu + \Phi X(t-1) + \Sigma \sqrt{S(t)}\, \varepsilon(t)\,, \tag{64}$$

where $S(t)$ is a matrix whose elements are affine functions of the state vector X, and A and B solve a system of difference equations. When all the yields are observed with error (i. e., $\xi_\tau \neq 0 \forall \tau,\ 0 \leq \tau \leq T$), QML estimation of the system (63)–(64) via the extended Kalman filter method yields an estimate of the coefficient vector. Applications of this approach for the US term structure data include Campbell and Viceira [81], Gong and Remolona [161], and Pennacchi [225]. The extended Kalman filter involves a linear approximation of the relation between the observed data and the state variables, and the associated approximation error will produce biased estimates. Christoffersen et al. [99] raise this concern and recommend the use of the so-called unscented Kalman filter for estimation of systems in which the relation between data and state variables is highly non-linear, e. g., options data.

Methods Accommodating the Lack of a Closed-Form Transition Density

We have so far mostly discussed estimation techniques for models with either a known transition density or one that is approximated by a discrete-time system. However, the majority of empirically-relevant continuous-time models do not possess explicit transition densities and alternative approaches are necessary. This problem leads us naturally towards the large statistics and econometric literature on estimation of diffusions from discretely-observed data. The vast majority of these studies assume that all relevant variables are observed so the latent volatility or yield curve state variables, integral to SV models, are not accounted for. Nonetheless, it may be feasible to extract the requisite estimates of the state variable by alternate means, thus restoring the feasibility, albeit not efficiency, of the basic approach. Since the literature is large and not directly geared towards genuine SV models, we focus on methods that have seen use in applications involving latent state variables.

A popular approach is to invert the map between the state vector and a subset of the observables assuming that the model prices specific securities exactly. In applications to equity markets this is done, e. g., by assuming that one option contract is priced without error, which implies a specific value (estimate) of the variance process given the model parameters Ψ. For instance, Pan [220] follows this approach in her study of S&P 500 options and returns, which we review in more detail in Sect. "Estimation from Option Data". In applications to fixed income markets it is likewise stipulated that certain bonds are priced without error, i. e., in Eq. (63) the error term $\xi_{\tau_i}(t)$ is fixed at zero for a set of maturities $\tau_1, \ldots, \tau_N$, where N matches the dimension of the state vector X. This approach yields an estimate for the latent variables through the inverse-map $X(t) = g(y_\tau(t), \Psi)$.

One criticism of the state vector inversion procedure is that it requires ad hoc assumptions regarding the choice of the securities that are error-free (those used to compute model-implied measures of the state vector) vis-a-vis those observed with error (used either for estimation or to assess model performance in an 'out-of-sample' cross-sectional check). In fact, the extracted state vector can be quite sensitive to the choice of derivatives (or yields) used. Nevertheless, this approach has intuitive appeal. Model-implied measures of the state vector, in combination with a closed-form expression for the conditional density (55), allow for efficient estimation of the coefficient vector Ψ via maximum likelihood. Analytic expressions for f_X in Eq. (55) exist in a limited number of cases. For instance, if X is Gaussian then f_X is multivariate normal, while if X follows a square-root process then f_X can be expressed in terms of the modified Bessel function (e. g., [113]). Different flavors of these continuous-time models are estimated in, e. g., [91,106,132,182,223]. In more general cases, including affine processes that combine Gaussian and square-root state variables, closed-form expressions for f_X are no longer available. In the rest of this section we briefly review different methods to overcome this problem. The interested reader may consult, e. g., [226] for more details.

Lo [203] warns that the common approach of estimating parameters of an Itô process by applying maximum likelihood to a discretization of the stochastic differen-

tial equation yields inconsistent estimators. In contrast, he characterizes the likelihood function as a solution to a partial differential equation. The method is very general, e. g., it applies not only to continuous-time diffusions but also to jump processes. In practice, however, analytic solutions to the partial differential equations (via, e. g., Fourier transforms) are available only for a small class of models so computationally-intensive methods (e. g., finite differencing or simulations) are generally required to solve the problem. This is a severe limitation in the case of multivariate systems like SV models.

For general Markov processes, where the above solution is infeasible, a variety of procedures have been advocated in recent years. Three excellent surveys provide different perspectives on the issue. Aït-Sahalia, Hansen, and Scheinkman [5] discuss operator methods and mention the potential of applying a time deformation technique to account for genuine SV features of the process, as in Conley, Hansen, Luttmer, and Scheinkman [109]. In addition, the Aït-Sahalia [3,4] closed-form polynomial expansions for discretely-sampled diffusions are reviewed along with the Schaumburg [235] extension to a general class of Markov processes with Lévy-type generators. Meanwhile, Bibby, Jacobsen, and Sørensen [61] survey the extensive statistics literature on estimating functions for diffusion-type models and Bandi and Phillips [42] explicitly consider dealing with nonstationary processes (see also the work of Bandi [39], Bandi and Nguyen [41], and Bandi and Phillips [43,44]).

The characteristic function based inference technique has been particularly widely adopted due to the natural fit with the exponentially affine model class which provides essentially closed-form solutions for many pricing applications. Consequently, we dedicate a separate section to this approach.

Characteristic Functions Singleton [242] proposes to exploit the information contained in the conditional characteristic function of the state vector X,

$$\phi(iu, X(t), \Psi) = \mathrm{E}\left[\mathrm{e}^{iu'X(t+1)}\middle|X(t)\right], \tag{65}$$

to pursue maximum likelihood estimation of affine term structure models. In Equation (65) we highlight the dependence of the characteristic function on the unknown parameter vector Ψ. When X is an affine (jump-)diffusion process, ϕ has the exponential affine form,

$$\phi(iu, X(t), \Psi) = \mathrm{e}^{\alpha_t(u)+\beta_t(u)'X(t)}, \tag{66}$$

where the functions α and β solve a system of ODEs. As such, the transition density f_X is known explicitly up to an inverse-Fourier transformation of the characteristic function (65),

$$f_X(X(t+1)\big|X(t);\Psi) = \frac{1}{\pi^N}\int_{\mathbb{R}_+^N} \mathrm{Re}\left[\mathrm{e}^{-iu'X(t+1)}\phi(iu, X(t), \Psi)\right]\mathrm{d}u\,. \tag{67}$$

Singleton shows that Gauss–Legendre quadrature with a relatively small number of quadrature points allows to accurately evaluate the integral in Eq. (67) when X is univariate. As such, the method readily delivers efficient estimates of the parameter vector, Ψ, subject to an auxiliary assumption, namely that the state vector may be extracted by assuming that a pre-specified set of security prices is observed without error while the remainder have non-trivial error terms.

When X is multivariate the Fourier inversion in Eq. (67) is computationally more demanding. Thus, when estimating multi-dimensional systems Singleton suggests focusing on the conditional density function of the individual elements of X, but conditioned on the full state vector,

$$f_{X_j}(X_j(t+1)|X(t);\Psi) = \frac{1}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{-i\omega\mathbf{I}_j'X(t+1)}\phi(i\omega\mathbf{I}_j, X(t), \Psi)\mathrm{d}\omega\,, \tag{68}$$

where the vector $\mathbf{I}_j$ has 1 in the jth element and zero elsewhere so that the jth element of X is $X_j(t+1) = \mathbf{I}_j'X(t+1)$. Maximization of the likelihood function obtained from f_{X_j}, for a fixed j, will often suffice to obtain a consistent estimate of Ψ. Exploiting more than one of the conditional densities (68) will result in more efficient Ψ estimate. For instance, the scores of multiple univariate log-likelihood functions, stacked in a vector, yield moment conditions that allow for generalized method of moment (GMM) estimation of the system. Alternatively, Joslin [192] proposes a change-of-measure transformation which reduces the oscillatory behavior of the integrand in Eq. (67). When using this transformation, Gauss-Hermite quadrature more readily provides a solution to the integral in (67) even if the state vector X is multi-dimensional, thus facilitating full ML estimation of the system.

Related, several studies have pursued GMM estimation of affine processes using characteristic functions. Definition (65) yields the moment condition

$$\mathrm{E}\left[(\phi(iu, X(t), \Psi) - \mathrm{e}^{iu'X(t+1)})z(u, X(t))\right] = 0\,, \tag{69}$$

where X is an N-dimensional (jump-)diffusion, $u \in \mathbb{R}^N$, and z is an instrument function. When X is affine, the characteristic function takes the exponential form (66). Different choices of u and z yield a set of moment conditions that can be used for GMM estimation and inference. Singleton [242] derives the optimal instrument in terms of delivering efficient estimates. Carrasco et al. [86] approximate the optimal instrument with a set of basis functions that do not require the knowledge of the conditional likelihood function, thus avoiding one of the assumptions invoked by Singleton. Further, they build on Carrasco and Florens [87] to implement estimation using a continuum of moment conditions, which yields maximum-likelihood efficiency. Other applications of GMM-characteristic function methods to affine (jump-) diffusions for equity index returns are in Chacko and Viceira [88] and Jiang and Knight [184].

In some cases the lack of closed-form expressions for the moment condition in Eq. (69) can hinder GMM estimation. In these cases the expectation in Eq. (69) can be evaluated by Monte Carlo integration. This is accomplished by simulating a long sample from the discretized process for a given value of the coefficient vector Ψ. The parameter Ψ is then estimated via the simulated method of moments (SMM) of McFadden [206] and Duffie and Singleton [131]. Singleton [242] proposes SMM characteristic function estimators that exploit the special structure of affine term structure models.

Efficient Estimation of General Continuous-Time Processes

A number of recent approaches offer excellent flexibility in terms of avoiding approximations to the continuous-time model-implied transition density while still facilitating efficient estimation of the evolution of the latent state vector for the system.

Maximum Likelihood with Characteristic Functions

Bates [58] develops a filtration-based maximum likelihood estimation method for affine processes. His approach relies on Bayes' rule to recursively update the joint characteristic function of latent variables and data conditional on past data. He then obtains the transition density by Fourier inversion of the updated characteristic function.

Denote with $y(t)$ and $X(t)$ the time-t values of the observable variable and the state vector, respectively, and let $Y(t) \equiv \{y(1), \ldots, y(t)\}$ be the data observed up to time t. Consider the case in which the characteristic function of $z(t+1) \equiv (y(t+1), X(t+1))$ conditional on $z(t) \equiv (y(t), X(t))$, is an exponential affine function of $X(t)$:

$$\begin{aligned}\phi(is, iu, z(t), \Psi) &= \mathrm{E}\left[\mathrm{e}^{is'y(t+1)+iu'X(t+1)}\middle| z(t)\right] \\ &= \mathrm{e}^{\alpha(is,iu,y(t))+\beta(is,iu,y(t))'X(t)} \,. \end{aligned} \quad (70)$$

Next, determine the value of the characteristic function conditional on the observed data $Y(t)$:

$$\begin{aligned}&\phi(is, iu, Y(t), \Psi) \\ &\quad = \mathrm{E}\left[\mathrm{E}\left[\mathrm{e}^{is'y(t+1)+iu'X(t+1)}\middle| z(t)\right]\middle| Y(t)\right] \\ &\quad = \mathrm{E}\left[\mathrm{e}^{\alpha(is,iu,y(t))+\beta(is,iu,y(t))'X(t)}\middle| Y(t)\right] \\ &\quad = \mathrm{e}^{\alpha(is,iu,y(t))}\psi(\beta(is, iu, y(t)), Y(t), \Psi) \,, \end{aligned} \quad (71)$$

where $\psi(iu, Y(t), \Psi) \equiv \mathrm{E}\left[\mathrm{e}^{iu'X(t)}\middle| Y(t)\right]$ denotes the (marginal) characteristic function for the state vector conditional on the observed data. Fourier inversion then yields the conditional density for the observation $y(t+1)$ conditional on $Y(t)$:

$$\begin{aligned}&f_y(y(t+1)|Y(t); \Psi) \\ &\quad = \frac{1}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{-is'y(t+1)}\phi(is, 0, Y(t), \Psi)\mathrm{d}s \,. \end{aligned} \quad (72)$$

The next step updates the characteristic function ψ (Bartlett [48]):

$$\begin{aligned}\psi(iu, Y(t+1), \Psi) &= \frac{1}{2\pi f_y(y(t+1)|Y(t); \Psi)} \\ &\quad \cdot \int_{\mathbb{R}} \mathrm{e}^{-is'y(t+1)}\phi(is, iu, Y(t), \Psi)\mathrm{d}s \,. \end{aligned} \quad (73)$$

To start the recursion, Bates initializes ψ at the unconditional characteristic function of the latent variable X. The log-likelihood function is then given by

$$\begin{aligned}\log \mathcal{L}(Y(T); \Psi) &= \log(f_y(y(1); \Psi) \\ &\quad + \sum_{t=2}^{T} \log(f_y(y(t)|Y(t-1); \Psi)) \,. \end{aligned} \quad (74)$$

A nice feature is that the method provides a natural solution to the filtering problem. The filtered estimate of the latent state X and its variance are computed from the first and second derivatives of the moment generating function $\psi(u, Y(t); \Psi)$ in Eq. (73), evaluated at $u = 0$:

$$\begin{aligned}\mathrm{E}[X(t+1)|Y(t+1); \Psi] &= \frac{1}{2\pi f_y(y(t+1)|Y(t); \Psi)} \\ &\quad \times \int_{\mathbb{R}} \mathrm{e}^{-is'y(t+1)}\phi_u(is, 0, Y(t); \Psi)\mathrm{d}s \end{aligned} \quad (75)$$

$$\mathrm{Var}[X(t+1)|Y(t+1);\Psi] = \frac{1}{2\pi f_y(y(t+1)|Y(t);\Psi)} \times \int_{\mathbb{R}} \mathrm{e}^{-is'y(t+1)}\phi_{uu}(is,0,Y(t);\Psi)\mathrm{d}s - \{\mathrm{E}[X(t+1)|Y(t+1)]\}^2 \,. \quad (76)$$

A drawback is that at each step t of the iteration the method requires storage of the entire characteristic function $\psi(iu, Y(t);\Psi)$. To deal with this issue Bates recommends to approximate the true ψ with the characteristic function of a variable with a two-parameter distribution. The choice of the distribution depends on the X-dynamics while the two parameters of the distribution are determined by the conditional mean $\mathrm{E}[X(t+1)|Y(t+1);\Psi]$ and variance $\mathrm{Var}[X(t+1)|Y(t+1);\Psi]$ given in Equations (75)–(76).

In his application Bates finds that the method is successful in estimating different flavors of the SV jump-diffusion for a univariate series of daily 1953–1996 S&P 500 returns. In particular, he shows that the method obtains estimates that are equally, if not more, efficient compared to the efficient method of moments and Markov Chain Monte Carlo methods described below. Extensions of the method to multivariate processes are theoretically possible, but they require numerical integration of multi-dimensional functions, which is computationally demanding.

Simulated Maximum Likelihood In Sect. "Filtering the Latent State Variable Directly During Estimation" we discussed methods for simulated ML estimation and inference in discrete-time SV models. Pedersen [224] and Santa-Clara [234] independently develop a simulated maximum likelihood (SML) method to estimate continuous-time diffusion models. They divide each interval in between two consecutive data points X_{t+1} and X_t into M sub-intervals of length $\Delta = 1/M$ and they discretize the X process using the Euler scheme,

$$X_{t+(i+1)\Delta} = X_{t+i\Delta} + \mu(X_{t+i\Delta})\Delta + \Sigma(X_{t+i\Delta})\sqrt{\Delta}\varepsilon_{t+(i+1)\Delta}\,, \quad i = 0,\ldots,M-1\,, \quad (77)$$

where μ and Σ are the drift and diffusion terms of the X process and ε is multivariate normal with mean zero and identity variance matrix. The transition density of the discretized process is multivariate normal with mean μ and variance matrix $\Sigma\Sigma'$. As Δ goes to zero, this density converges to that of the continuous-time process X. As such, the transition density from X_t to X_{t+1} is given by

$$f_X(X_{t+1}|X_t;\Psi) = \int f_X(X_{t+1}|X_{t+1-\Delta};\Psi) \times f_X(X_{t+1-\Delta}|X_t;\Psi)\mathrm{d}X_{t+1-\Delta}\,. \quad (78)$$

For sufficiently small values of Δ the first term in the integrand, $f_X(X_{t+1}|X_{t+1-\Delta};\Psi)$, is approximated by the transition density of the discretized process, while the second term, $f_X(X_{t+1-\Delta}|X_t;\Psi)$, is a multi-step-ahead transition density that can be computed from the recursion from X_t to $X_{t+1-\Delta}$. Writing the right-hand side of Eq. (78) as a conditional expectation yields

$$f_X(X_{t+1}|X_t;\Psi) = E_{X_{t+1-\Delta}|X_t}\left[f_X(X_{t+1}|X_{t+1-\Delta};\Psi)\right]. \quad (79)$$

The expectation in Eq. (79) can be computed by Monte Carlo integration over a large number of paths for the process X, simulated via the Euler scheme (77). As Δ vanishes, the Euler scheme is consistent. Thus, when the size of the simulated sample increases the sample average of the function f_X, evaluated at the random draws of $X_{t+1-\Delta}$, converges to the true transition density. Application of the principles in Bladt and Sørensen [64] may well be useful in enhancing the efficiency of the simulation scheme and hence the actual efficiency of the inference procedure in practice.

Brandt and Santa-Clara [75] apply the SML method to estimate a continuous-time model of the joint dynamics of interest rates in two countries and the exchange rate between the two currencies. Piazzesi [227] extends the SML approach for jump-diffusion processes with time-varying jump intensity. She considers a high-frequency policy rule based on yield curve information and an arbitrage-free bond market and estimates the model using 1994–1998 data on the Federal Reserve target rate, the six-month LIBOR rate, and swap yields.

An important issue is how to initialize any unobserved component of the state vector, $X(t)$, such as the volatility state at each observation to provide a starting point for the next Monte Carlo integration step. This may be remedied through application of the particle filter, as mentioned earlier and discussed below in connection with MCMC estimation. Another possibility is, as also indicated previously, to extract the state variable through inversion from derivatives prices or yields assumed observed without pricing errors.

Indirect Inference There are also other method-of-moments strategies to estimate finitely-sampled continuous-

time processes of a general type. One prominent approach approximates the unknown transition density for the continuous-time process with the density of a semi-nonparametric (SNP) auxiliary model. Then one can use the score function of the auxiliary model to form moment conditions for the parameter vector Ψ of the continuous-time model. This approach yields the efficient method of moments estimator (EMM) of Gallant and Tauchen [154], Gallant et al. [150], and Gallant and Long [152], and the indirect inference estimator of Gouriéroux et al. [162] and Smith [243].

To fix ideas, suppose that the conditional density for a continuous-time return process r (the 'structural' model) is unknown. We intend to approximate the unknown density with a discrete-time model (the 'auxiliary' model) that is tractable and yet sufficiently flexible to accommodate the systematic features of the actual data sample well. A parsimonious auxiliary density for r embeds ARMA and EGARCH leading terms to capture the conditional mean and variance dynamics. There may be residual excess skewness and kurtosis that elude the ARMA and EGARCH forms. As such, the auxiliary density is rescaled using a nonparametric polynomial expansion of order K, which yields

$$g_K(r(t)|x(t);\xi) = \left(\nu + (1-\nu) \times \frac{[P_K(z(t),x(t))]^2}{\int_{\mathbb{R}}[P_K(z(t),x(t))]^2\phi(u)\mathrm{d}u}\right)\frac{\phi(z(t))}{\sqrt{h(t)}}\,, \tag{80}$$

where ν is a small constant, $\phi(.)$ is the standard normal density, $x(t)$ contains lagged return observations, and

$$z(t) = \frac{r(t)-\mu(t)}{\sqrt{h(t)}}\,, \tag{81}$$

$$\mu(t) = \phi_0 + ch(t) + \sum_{i=1}^{s}\phi_i r(t-1) + \sum_{i=1}^{u}\delta_i\varepsilon(t-1)\,, \tag{82}$$

$$\log h(t) = \omega + \sum_{i=1}^{p}\beta_i \log h(t-1) + (1+\alpha_1 L + \cdots + \alpha_q L^q) \times \left[\theta_1 z(t-1) + \theta_2(b(z(t-1)) - \sqrt{2/\pi})\right], \tag{83}$$

$$P_K(z,x) = \sum_{i=0}^{K_z} a_i(x)z^i = \sum_{i=0}^{K_z}\left(\sum_{|j|=0}^{K_x} a_{ij}x^j\right)z^i\,, \quad a_{00} = 1\,. \tag{84}$$

Here j is a multi-index vector, $x^j \equiv (x_1^{j_1},\ldots,x_M^{j_M})$, and $|j| \equiv \sum_{m=1}^{M} j_m$. The term $b(z)$ is a smooth (twice-differentiable) function that closely approximates the absolute value operator in the EGARCH variance equation.

In practice, the representation of P_K is given by Hermite orthogonal polynomials. When the order K of the expansion increases, the auxiliary density will approximate the data arbitrarily well. If the structural model is indeed the true data generating process, then the auxiliary density will converge to that of the structural model. For a given K, the QML estimator $\hat{\xi}$ for the auxiliary model coefficient satisfies the score condition

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\partial \log g_K(r(t)|x(t);\hat{\xi})}{\partial\xi} = 0\,. \tag{85}$$

Suppose now that the structural model is correct and Ψ_0 is the true value of its coefficient vector. Consider a series $\{r(t;\Psi), x(t;\Psi)\}$, $t = 1,\ldots,\mathcal{T}(T)$, simulated from the structural model. Then we expect that the score condition (85) holds when evaluated by averaging over the simulated returns rather than over the actual data:

$$m_{\mathcal{T}(T)}(\Psi_0,\hat{\xi}) = \frac{1}{\mathcal{T}(T)}\sum_{t=1}^{\mathcal{T}(T)}\frac{\partial \log g_K(r(t,\Psi_0)|x(t,\Psi_0);\hat{\xi})}{\partial\xi} \approx 0\,. \tag{86}$$

When T and $\mathcal{T}(T)$ tend to infinity, condition (86) holds exactly.

Gallant and Tauchen [154] propose the EMM estimator $\hat{\Psi}$ defined via

$$\hat{\Psi} = \arg\min_{\Psi} m_{\mathcal{T}(T)}(\Psi,\hat{\xi})'\,\hat{W}_T\, m_{\mathcal{T}(T)}(\Psi,\hat{\xi})\,, \tag{87}$$

where the weighting matrix $\hat{W}_T$ is a consistent estimate of the inverse asymptotic covariance matrix of the auxiliary score function, e. g., the inverse outer product of the SNP gradient:

$$\hat{W}_T^{-1} = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\partial \log g_K(r(t)|x(t);\hat{\xi})}{\partial\xi}\right] \times \left[\frac{\partial \log g_K(r(t)|x(t);\hat{\xi})}{\partial\xi}\right]'\,. \tag{88}$$

An important advantages of the technique is that EMM estimates achieve the same degree of efficiency as the ML procedure, when the score of the auxiliary model asymptotically spans the score of the true model. It also delivers powerful specification diagnostics that provide guidance in the model selection. Gallant and Tauchen [154] show that the EMM estimator is asymptotically normal. Further, under the assumption that the structural model is correctly specified, they derive a χ^2 statistic for the test of over-identifying restrictions. Gallant et al. [150] normalize the vector $m_{\mathcal{T}(T)}(\hat{\Psi}, \hat{\xi})$ by its standard error to obtain a vector of score t-ratios. The significance of the individual score elements is often informative of the source of model mis-specification, with the usual caveat that failure to capture one characteristic of the data may result in the significance of a moment condition that pertains to a coefficient not directly related to that characteristic (due to correlation in the moment conditions). Finally, EMM provides a straightforward solution to the problem of filtering and forecasting the latent return variance process V, i. e., determining the conditional densities $f(V(t)|x(t), \Psi)$ and $f(V(t+j)|x(t), \Psi)$, $j \geq 0$. This is accomplished through the *reprojection* method discussed in, e. g., Gallant and Long [152] and Gallant and Tauchen [155]. In applications to dynamic term structure models, the same method yields filtered and forecasted values for the latent state variables.

The reprojection method assumes that the coefficient vector Ψ is known. In practice, Ψ is fixed at the EMM estimate $\hat{\Psi}$. Then one simulates a sample of returns and latent variables from the structural model and fits the auxiliary model on the simulated data. This is equivalent to the first step of the EMM procedure except that, in the reprojection step, we fit the auxiliary model assuming the structural model is correct, rather than using actual data. The conditional density of the auxiliary model, estimated under the null, approximates the unknown density of the structural model:

$$g_K(r(t+j)|x(t); \tilde{\xi}) \approx f(r(t+j)|x(t); \hat{\Psi}), \quad j \geq 0\,, \tag{89}$$

where $\tilde{\xi}$ is the QML estimate of the auxiliary model coefficients obtained by fitting the model on simulated data. This approach yields filtered estimates and forecasts for the conditional mean and variance of the return via

$$\mathrm{E}\left[r(t+j)|x(t); \hat{\Psi}\right] = \int y g_K(y|x(t); \tilde{\xi}) \mathrm{d}y\,, \tag{90}$$

$$\mathrm{Var}\left[r(t+j)|x(t); \hat{\Psi}\right] = \int \left(y - \mathrm{E}\left[r(t+j)|x(t); \hat{\Psi}\right]\right)^2 \times g_K(y|x(t); \tilde{\xi}) \mathrm{d}y\,. \tag{91}$$

An alternative approach consists in fitting an auxiliary model for the latent variable (e. g., the return conditional variance) as a function of current and lagged returns. It is straightforward to estimate such model using data on the latent variable and the associated returns simulated from the structural model with the EMM coefficient $\hat{\Psi}$. Also in this case the auxiliary model density approximates the true one, i. e.,

$$g_K^V(V(t+j)|x(t); \tilde{\xi}) \approx f^V(V(t+j)|x(t); \hat{\psi})\,, \quad j \geq 0. \tag{92}$$

This approach yields a forecast for the conditional variance process,

$$\mathrm{E}\left[V(t+j)|x(t); \hat{\Psi}\right] = \int v g_K^V(v|x(t); \tilde{\xi}) \mathrm{d}v\,. \tag{93}$$

In sum, reprojection is a simulation approach to implement a non-linear Kalman-filter-type technique, which yields effective forecasts for the unobservable state vector.

The indirect inference estimator by Gouriéroux et al. [162] and Smith [243] is closely related to the EMM estimator. Indirect inference exploits that the following two quantities should be close when the structural model is correct and the data are simulated at the true parameter Ψ_0: (i) the QML estimator $\hat{\xi}$ for the auxiliary model computed from actual data; (ii) the QML estimator $\hat{\xi}(\Psi)$ for the auxiliary model fitted on simulations from the structural model. Minimizing the distance between $\hat{\xi}$ and $\hat{\xi}(\Psi)$ in an appropriate metric yields the indirect inference estimator for Ψ. Similar to EMM, asymptotic normality holds and a χ^2 test for over-identifying restrictions is available. However, the indirect inference approach is computationally more demanding, because finding the value of Ψ that minimizes the distance function requires re-estimating the auxiliary model on a different simulated sample for each iteration of the optimization routine. EMM does not have this drawback, since the EMM objective function is evaluated at the same fitted score at each iteration. Nonetheless, there may well be circumstances where particular auxiliary models are of primary economic interest and estimation based on the corresponding moment conditions may serve as a useful diagnostic tool for model performance in such directions.

Several studies have used EMM to fit continuous-time SV jump-diffusion models for equity index returns, e. g., Andersen et al. [15], Benzoni [59], Chernov and Ghysels [93], and Chernov et al. [94,95]. Andersen and Lund [28] and Andersen et al. [16] use EMM to estimate SV jump-diffusion models for the short-term inter-

est rate. Ahn et al. [1,2], Brandt and Chapman [71], and Dai and Singleton [114] fit different flavors of multi-factor dynamic term structure models. Andersen et al. [27] document the small-sample properties of the efficient method of moments estimator for stationary processes, while Duffee and Stanton [128] study its properties for near unit-root processes. A. Ronald Gallant and George E. Tauchen at Duke University have prepared well-documented general-purpose EMM and SNP packages, available for download at the web address ftp.econ.duke.edu in the directories pub/get/emm and pub/get/snp. In applications it is often useful to customize the SNP density to allow for a more parsimonious fit of the data under investigation. For instance, Andersen et al. [15,16], Andersen and Lund [28], and Benzoni [59] rely on the SNP density (80)–(84).

Markov Chain Monte Carlo The MCMC method provides a Bayesian solution to the inference problem for a dynamic asset pricing model. The approach treats the model coefficient Ψ as well as the vector of latent state variables X as random variables and computes the posterior distribution $f(\Psi, X|Y)$, conditional on certain observable variables Y, predicted by the model. The setting is sufficiently general to deal with a wide range of situations. For instance, X and Y can be the (latent) volatility and (observable) return processes as is the case of an SV model for asset returns. Or X and Y can be the latent state vector and observable yields in a dynamic term structure model.

The posterior distribution $f(\Psi, X|Y)$ is the main tool to draw inference not only on the coefficient Ψ but also on the latent vector X. Since $f(\Psi, X|Y)$ is unknown in closed-form in relevant applications, MCMC relies on a simulation (a Markov Chain) from the conditional density $f(\Psi, X|Y)$ to compute mode, mean, and standard deviations for the model coefficients and state variables via the Monte Carlo method.

The posterior $f(\Psi, X|Y)$ is analytically untractable and extremely high-dimensional, so that simulation directly from $f(\Psi, X|Y)$ is typically infeasible. The MCMC approach hinges on the Clifford–Hammersley theorem, which determines conditions under which the posterior $f(\Psi, X|Y)$ is uniquely determined by the marginal posterior distributions $f(\Psi|X, Y)$ and $f(X|\Psi, Y)$. In turn, the posteriors $f(\Psi|X, Y)$ and $f(X|\Psi, Y)$ are determined by a set or univariate posterior distributions. Specifically, denote with $\Psi(i)$ the ith element of the coefficient Ψ, $i = 1, \ldots, K$, and with $\Psi(-i)$ the vector consisting of all elements in Ψ except for the ith one. Similarly denote with $X(t)$ the tth row of the state vector, $t = 1, \ldots, T$, and with $X(-t)$ the rest of the vector. Then the Clifford–Hammersley theorem allows to characterize the posterior $f(\Psi, X|Y)$ via $K + T$ univariate posteriors,

$$f(\Psi(i)|\Psi(-i), X, Y)\ , \quad i = 1, \ldots, K \tag{94}$$

$$f(X(t)|X(-t), \Psi, Y)\ , \quad t = 1, \ldots, T\ . \tag{95}$$

The construction of the Markov Chain relies on the so-called Gibbs sampler. The first step of the algorithm consists in choosing initial values for the coefficient and the state, Ψ_0 and X_0. When (one of or both) the multi-dimensional posteriors are tractable, the Gibbs sampler generates values Ψ_1 and X_1 directly from $f(\Psi|X, Y)$ and $f(X|\Psi, Y)$. Alternatively, each element of Ψ_1 and X_1 is drawn from the univariate posteriors (94)–(95). Some of these posteriors may also be analytically intractable or efficient algorithms to draw from these posteriors may not exist. In such cases the Metropolis-Hastings algorithm ensures that the simulated sample is consistent with the posterior target distribution. Metropolis-Hastings sampling consists of an accept-reject procedure of the draws from a 'proposal' or 'candidate' tractable density, which is used to approximate the unknown posterior (see, e. g., Johannes and Polson [187]).

Subsequent iterations of Gibbs sampling, possibly in combination with the Metropolis-Hastings sampling, yield a series of 'sweeps' $\{\Psi_s, X_s\}$, $s = 1, \ldots, S$, with limiting distribution $f(\Psi, X|Y)$. A long number of sweeps may be necessary to 'span' the whole posterior distribution and obtain convergence due to the serial dependence of subsequent draws of coefficients and state variables. When the algorithm has converged, additional simulations provide a sample from the joint posterior distribution.

The MCMC approach has several advantages. First, the inference automatically accounts for parameter uncertainty. Further, the Markov Chain provides a direct and elegant solution to the *smoothing* problem, i. e., the problem of determining the posterior distribution for the state vector X conditional on the entire data sample, $f(X(t)|Y(1), \ldots, Y(T), \Psi)$, $t = 1, \ldots, T$. The limitation on the approach is largely that efficient sampling schemes for the posterior distribution must be constructed for each specific problem at hand which by nature is case specific and potentially cumbersome or inefficient. Nonetheless, following the development of more general simulation algorithms, the method has proven flexible for efficient estimation of a broad class of important models.

One drawback is that MCMC does not deliver an immediate solution to the *filtering* problem, i. e., determining $f(X(t)|Y(1), \ldots, Y(t), \Psi)$, and the *forecasting* problem, i. e., determining $f(X(t+j)|Y(1), \ldots, Y(t), \Psi)$, $j > 0$. However, recent research is overcoming this limitation

through the use of the 'particle filter'. Bayes rule implies

$$f(X(t+1)|Y(1),\ldots,Y(t+1),\Psi) \propto f(Y(t+1)| X(t+1),\Psi) f(X(t+1)|Y(1),\ldots,Y(t),\Psi)\,, \quad (96)$$

where the symbol $\propto$ denotes 'proportional to'. The first density on the right-hand side of Eq. (96) is determined by the SV model and it is often known in closed form. In contrast, the second density at the far-right end of the equation is given by an integral that involves the unknown filtering density at the prior period, $f(X(t)|Y(1),\ldots,Y(t),\Psi)$:

$$f(X(t+1)|Y(1),\ldots,Y(t),\Psi) = \int f(X(t+1)|X(t),\Psi) \times f(X(t)|Y(1),\ldots,Y(t),\Psi)\mathrm{d}X(t)\,. \quad (97)$$

The particle method relies on simulations to construct a finite set of weights $w^i(t)$ and particles $X^i(t)$, $i=1,\ldots,N$, that approximate the unknown density with a finite sum,

$$f(X(t)|Y(1),\ldots,Y(t),\Psi) \approx \sum_{i=1}^{N} w^i(t)\delta_{X^i(t)}\,, \quad (98)$$

where the Dirac function $\delta_{X^i(t)}$ assigns mass one to the particle $X^i(t)$. Once the set of weights and particles are determined, it is possible to re-sample from the discretized distribution. This step yields a simulated sample $\{X^s(t)\}_{s=1}^{S}$ which can be used to evaluate the density in Eq. (97) via Monte Carlo integration:

$$f(X(t+1)|Y(1),\ldots,Y(t),\Psi) \approx \frac{1}{S}\sum_{s=1}^{S} f(X(t+1)|X^s(t),\Psi)\,. \quad (99)$$

Equation (99) solves the forecasting problem while combining formulas (96) and (99) solves the filtering problem. The challenge in practical application of the particle filter is to identify an accurate and efficient algorithm to construct the set of particles and weights. We point the interested reader to Kim et al. [193], Pitt and Shephard [228] and Johannes and Polson [187] for a discussion on how to approach this problem.

The usefulness of the MCMC method to solve the inference problem for SV models has been evident since the early work by Jacquier et al. [180], who develop an MCMC algorithm for the logarithmic SV model. Jacquier et al. [181] provide extensions to correlated and non-normal error distributions. Kim et al. [193], Pitt and Shephard [228] and Chib et al. [96] develop simulation-based methods to solve the filtering problem, while Chib et al. [97] use the MCMC approach to estimate a multivariate SV model. Elerian et al. [135] and Eraker [140] discuss how to extend the MCMC inference method to a continuous-time setting. Eraker [140] uses the MCMC approach to estimate an SV diffusion process for interest rates, while Jones [189] estimates a continuous-time model for the spot rate with non-linear drift function. Eraker et al. [142] estimate an SV jump-diffusion process using data on S&P 500 return while Eraker [141] estimates a similar model using joint data on options and underlying S&P 500 returns. Li et al. [196] allow for Lévy-type jumps in their model. Collin-Dufresne et al. [104] use the MCMC approach to estimate multi-factor affine dynamic term structure model using swap rates data. Johannes and Polson [186] give a comprehensive survey of the still ongoing research on the use of the MCMC approach in the general nonlinear jump-diffusion SV setting.

Estimation from Option Data

Options' payoffs are non-linear functions of the underlying security price. This feature renders options highly sensitive to jumps in the underlying price and to return volatility, which makes option data particularly useful to identify return dynamics. As such, several studies have taken advantage of the information contained in option prices, possibly in combination with underlying return data, to estimate SV models with or without discontinuities in returns and volatility.

Applications to derivatives data typically require a model for the pricing errors. A common approach is to posit that the market price of an option, O^*, normalized by the underlying observed security price S^*, is the sum of the normalized model-implied option price, O/S^*, and a disturbance term ε (e. g., Renault [230]):

$$\frac{O^*}{S^*} = \frac{O(S^*,V,K,\tau,\Psi)}{S^*} + \varepsilon\,, \quad (100)$$

where V is the latent volatility state, K is the option strike price, τ is time to maturity, and Ψ is the vector with the model coefficients. A pricing error ε could arise for several reasons, including measurement error (e. g., price discreteness), asynchroneity between the derivatives and underlying price observations, microstructure effects, and perhaps most importantly specification error. The structure imposed on ε depends on the choice of a specific 'loss function' used for estimation (e. g., Christoffersen and Jacobs [98]). Several studies have estimated the coefficient vector Ψ by minimizing the sum of the squared option pricing errors normalized by the underlying price S^*, as

in Eq. (100). Others have focused on either squared dollar pricing errors, or squared errors normalized by the options market price (instead of S^*). The latter approach has the advantage that a \$1 error on an expensive in-the-money option carries less weight than the same error on a cheaper out-of-the-money contract. The drawback is that giving a lot of weight to the pricing errors on short-maturity deep-out-of-the-money options could bias the estimation results. Finally, the common practice of expressing option prices in terms of their Black-Scholes implied volatilities has inspired other scholars to minimize the deviations between Black-Scholes implied volatilities inferred from model and market prices (e. g., Mizrach [216]). An alternative course is to form a moment-based loss function and follow a GMM- or SMM-type approach to estimate Ψ. To this end moment conditions stem from distributional assumptions on the pricing error ε (e. g., $\mathrm{E}[\varepsilon] = 0$) or from the scores of a reduced-form model that approximates the data.

In estimating the model, some researchers have opted to use a panel of options consisting of contracts with multiple strikes and maturities across dates in the sample period. This choice brings a wealth of information on the cross-sectional and term-structure properties of the implied volatility smirk into the analysis. Others rely on only one option price observation per time period, which shifts the focus to the time-series dimension of the data. Some studies re-estimate the model on a daily basis rather than seeking a single point estimate for the coefficient Ψ across the entire sample period. This ad hoc approach produces smaller in-sample pricing errors, which can be useful to practitioners, but at the cost of concealing specification flaws by over-fitting the model, which tends to hurt out-of-sample performance. The different approaches are in part dictated by the intended use of the estimated system as practitioners often are concerned with market making and short-term hedging while academics tend to value stable relations that may form the basis for consistent modeling of the dominant features of the system over time.

Early contributions focus on loss functions based on the sum of squared option pricing errors and rely entirely on option data for estimation. This approach typically yields an estimate of the model coefficient Ψ that embeds an adjustment for risk, i. e., return and volatility dynamics are identified under the risk-neutral rather than the physical probability measure. For instance, Bates [56] considers an SV jump-diffusion model for Deutsche Mark foreign currency options and estimates its coefficient vector Ψ via nonlinear generalized least squares of the normalized pricing errors with daily option data from January 1984 to June 1991. A similar approach is followed by Bates [57] who fits an SV model with two latent volatility factors and jumps using daily data on options on the S&P 500 futures from January 1988 to December 1993. Bakshi et al. [34] focus on the pricing and hedging of daily S&P 500 index options from June 1988 to May 1991. In their application they re-calibrate the model on a daily basis by minimizing the sum of the squared dollar pricing errors across options with different maturities and strikes. Huang and Wu [173] explore the pricing implications of the time-changed Lévy process by Carr and Wu [84] for daily S&P 500 index options from April 1999 to May 2000. Their Lévy return process allows for discontinuities that exhibit higher jump frequencies compared to the finite-intensity Poisson jump processes in Equations (37)–(41). Further, their model allows for a random time change, i. e., a monotonic transformation of the time variable which generates SV in the diffusion and jump components of returns. In contrast, Bakshi et al. [35] fit an SV jump-diffusion model by SMM using daily data on long-maturity S&P 500 options (LEAPS).

More recent studies have relied on joint data on S&P 500 option prices and underlying index returns, spanning different periods, to estimate the model. This approach forces the same model to price securities in two different markets and relies on information from the derivatives and underlying securities to better pin down model coefficients and risk premia. For instance, Eraker [141] and Jones [190] fit different flavors of the SV model (with and without jumps, respectively) by MCMC. Pan [220] follows a GMM approach to estimate an SV jump-diffusion model using weekly data. She relies on a single at-the-money option price observation each week, which identifies the level of the latent volatility state variable (i. e., at each date she fixes the error term ε at zero and solves Eq. (100) for V). Aït-Sahalia and Kimmel [7] apply Aït-Sahalia's [4] method to approximate the likelihood function for a joint sample of options and underlying prices. Chernov and Ghysels [93] and Benzoni [59] obtain moment conditions from the scores of a SNP auxiliary model. Similarly, other recent studies have found it useful to use joint derivatives and interest rate data to fit dynamic term structure models, e. g., Almeida et al. [9], and Bikbov and Chernov [62].

Finally, a different literature has studied the option pricing implications of a model in which asset return volatility is a deterministic function of the asset price and time, e. g., Derman and Kani [119], Dupire [134], Rubinstein [233], and Jackwerth and Rubinstein [177]. Since volatility is not stochastic in this setting, we do not review these models here and point the interested reader to, e. g., [133] for an empirical analysis of their performance.

Future Directions

In spite of much progress in our understanding of volatility new challenges lie ahead. In recent years a wide array of volatility-sensitive products has been introduced. The market for these derivatives has rapidly grown in size and complexity. Research faces the challenge to price and hedge these new products. Moreover, the recent developments in model-free volatility modeling have effectively given empirical content to the latent volatility variable, which opens the way for a new class of estimation methods and specification tests for SV systems. Related, improved volatility measures enable us to shed new light on the properties and implications of the volatility risk premium. Finally, more work is needed to better understand the linkage between fluctuations in economic fundamentals and low- and high-frequency volatility movements. We conclude this chapter by briefly reviewing some open issues in these four areas of research.

Volatility and Financial Markets Innovation

Volatility is a fundamental input to any financial and real investment decision. Markets have responded to investors' needs by offering an array of volatility-linked instruments. In 1993 the Chicago Board Option Exchange (CBOE) has introduced the VIX index, which measures the market expectations of near-term volatility conveyed by equity-index options. The index was originally computed using the Black-Scholes implied volatilities of eight different S&P 100 option (OEX) series so that, at any given time, it represented the implied volatility of a hypothetical at-the-money OEX option with exactly 30 days to expiration (see [257]). On September 22, 2003, the CBOE began disseminating price level information using a revised 'model-free' method for the VIX index. The new VIX is given by the price of a portfolio of S&P 500 index options and incorporates information from the volatility smirk by using a wider range of strike prices rather than just at-the-money series (see [77]). On March 26, 2004, trading in futures on the VIX Index started on the CBOE Futures Exchange (CFE) while on February 24, 2006, options on the VIX began trading on the Chicago Board Options Exchange. These developments have opened the way for investors to trade on option-implied measures of market volatility. The popularity of the VIX prompted the CBOE to introduce similar indices for other markets, e. g., the VXN NASDAQ 100 Volatility Index.

Along the way, a new over-the-counter market for volatility derivatives has also rapidly grown in size and liquidity. Volatility derivatives are contracts whose payments are expressed as functions of realized variance. Popular examples are variance swaps, which at maturity pay the difference between realized variance and a fixed strike price. According to estimates by BNP Paribas reported by the Risk [176] magazine, the daily trading volume for variance swaps on indices reached $4–5 million in vega notional (measured in dollars per volatility point) in 2006, which corresponds to payments in excess of $1 billion per percentage point of volatility on an annual basis (Carr and Lee [82]). Using variance swaps hedge fund managers and proprietary traders can easily place huge bets on market volatility.

Finally, in recent years credit derivatives markets have evolved in complexity and grown in size. Among the most popular credit derivatives are the credit default swaps (CDS), which provide insurance against the risk of default by a particular company. The buyer of a single-name CDS acquires the right to sell bonds issued by the company at face value when a credit event occurs. Multiple-name contracts can be purchased simultaneously through credit indices. For instance, the CDX indices track the credit spreads for different portfolios of North American companies while the iTraxx Europe indices track the spreads for portfolios of European companies. At the end of 2006 the notional amount of outstanding over-the-counter single- and multi-name CDS contracts stood at $19 and $10 trillion, respectively, according to the September 2007 Bank for International Settlements Quarterly Review.

These market developments have raised new interesting issues for research to tackle. The VIX computations based on the new model-free definition of implied volatility used by the CBOE requires the use of options with strike prices that cover the entire support of the return distribution. In practice, liquid options satisfying this requirement often do not exist and the CBOE implementation introduces random noise and systematic error into the index (Jiang and Tian [185]). Related, the VIX implementation entails a truncation, i. e., the CBOE discards illiquid option prices with strikes lying in the tails of the return distribution. As such, the notion of the VIX is more directly linked to that of corridor volatility [26]. In sum, robust implementation of a model free measure of implied volatility is still an open area of research. Future developments in this direction will also have important repercussions on the hedging practices for implied-volatility derivatives.

Pricing and hedging of variance derivatives is another active area of research. Variance swaps admit a simple replication strategy via static positions in call and put options on the underlying asset, similar to model-free implied volatility measures (e. g., [77,83]). In contrast, it is still an open area of research to determine the replication

strategy for derivatives whose payoffs are non-linear function of realized variance, e. g., volatility swaps, which pay the square-root of realized variance, or call and put options on realized variance. [82] is an interesting paper in this direction.

Limited liability gives shareholders the option to default on the firm's debt obligation. As such, a debt claim has features similar to a short position in a put option. The pricing of corporate debt is therefore sensitive to the volatility of the firms' assets: higher volatility increases the probability of default and therefore reduces the price of debt and increases credit spreads. The insights and techniques developed in the SV literature could prove useful in credit risk modeling and applications (e. g., [179,248,260]).

The Use of Realized Volatility for Estimation of SV Models

Another promising line of research aims at extracting the information in RV measures for the estimation of dynamic asset pricing models. Early work along these lines includes Barndorff-Nielson and Shephard [51], who decompose RV into actual volatility and realized volatility error. They consider a state-space representation for the decomposition and apply the Kalmann filter to estimate different flavors of the SV model. Moreover, Bollerslev and Zhou [68] and Garcia et al. [156], build on the insights of Meddahi [210] to estimate SV diffusion models using conditional moments of integrated volatility. More recently, Todorov [252] generalizes the analysis for the presence of jumps.

Related, recent studies have started to use RV measures to test the implications of models previously estimated with lower-frequency data. Since RV gives empirical content to the latent quadratic variation process, this approach allows for a direct test of the model-implied restrictions on the latent volatility factor. Recent work along these lines includes Andersen and Benzoni [12], who use model-free RV measures to show that the volatility spanning condition embedded in some affine term structure models is violated in the US Treasury market. Christoffersen et al. [100] note that the Heston square-root SV model implies that the dynamics for the standard deviation process are conditionally Gaussian. They reject this condition by examining the distribution of the changes in the square-root RV measure for S&P 500 returns.

Volatility Risk Premium

More work is needed to better understand the link between asset return volatility and model risk premia. Also in this case, RV measures are a fruitful source of information to shed new light on the issue. Among the recent studies that pursue this venue is Bollerslev et al. [66], who exploit the moments of RV and option-implied volatility to gauge a measure of the volatility risk premium. Todorov [251] explores the variance risk premium dynamics using high-frequency S&P 500 index futures data and data on the VIX index. He finds the variance risk premium to vary significantly over time and to increase during periods of high volatility and immediately after big jumps in underlying returns. Carr and Wu [85] provide a broader analysis of the variance risk premium for five equity indices and 35 individual stocks. They find the premium to be large and negative for the indices while it is much smaller for the individual stocks. Further, they also find the premium to increase (in absolute value) with the level of volatility. Additional work on the volatility risk premium embedded in individual stock options is in Bakshi and Kapadia [36], Driessen et al. [123], and Duarte and Jones [126]. Other studies have examined the linkage between volatility risk premia and equity returns (e. g., [69]) and hedge-fund performance (e. g., [70]). New research is also examining the pricing of aggregate volatility risk in the cross-section of stock returns. For instance, Ang et al. [30] find that average returns are lower on stocks that have high sensitivities to innovations in aggregate volatility and high idiosyncratic volatility (see also the related work by Chen [90] Ang et al. [32], Bandi et al, Guo et al. [42]). This evidence is consistent with the findings of the empirical option pricing literature, which suggests that there is a negative risk premium for volatility risk. Intuitively, periods of high market volatility are associated to worsened investment opportunities and tend to coincide with negative stock market returns (the so-called leverage effect). As such, investors are willing to pay higher prices (i. e., accept lower expected returns) to hold stocks that do well in high-volatility conditions.

Determinants of Volatility

Finally, an important area of future research concerns the linkage between asset return volatility and economic uncertainty. Recent studies have proposed general equilibrium models that produce low-frequency fluctuations in conditional volatility, e. g., Campbell and Cochrane [80], Bansal and Yaron [47], McQueen and Vorkink [207], and Tauchen [246]. Related, Engle and Rangel [139] and Engle et al. [138] link macroeconomic variables and long-run volatility movements. It is still an open issue, however, to determine the process through which news about economic fundamentals are embedded into prices to generate high-frequency volatility fluctuations. Early research

by Schwert [236] and Shiller [241] has concluded that the amplitude of the fluctuations in aggregate stock volatility is difficult to explain using simple models of stock valuation. Further, Schwert [236] notes that while aggregate leverage is significantly correlated with volatility, it explains a relatively small part of the movements in stock volatility. Moreover, he finds little evidence that macroeconomic volatility (measured by inflation and industrial production volatility) helps predict future asset return volatility. Model-free realized volatility measures are a useful tool to further investigate this issue. Recent work in this direction includes Andersen et al. [22] and Andersen and Bollerslev [17], who explore the linkage between news arrivals and exchange rates volatility, and Andersen and Benzoni [13], who investigate the determinants of bond yields volatility in the US Treasury market. Related, Balduzzi et al. [38] and Fleming and Remolona [146] study the reaction of trading volume, bid-ask spread, and price volatility to macroeconomic surprises in the US Treasury market, while Brandt and Kavajecz [74] and Pasquariello and Vega [222] focus instead on the price discovery process and explore the implications of order flow imbalances (excess buying or selling pressure) on day-to-day variation in yields.

Acknowledgments

We are grateful to Olena Chyruk, Bruce Mizrach (the Section Editor) and Neil Shephard for helpful comments and suggestions. Of course, all errors remain our sole responsibility. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Chicago or the Federal Reserve System. The work of Andersen is supported by a grant from the NSF to the NBER and support from CREATES funded by the Danish National Research Foundation.

Bibliography

Primary Literature

1. Ahn DH, Dittmar RF, Gallant AR (2002) Quadratic Term Structure Models: Theory and Evidence. Rev Finance Stud 15:243–288
2. Ahn DH, Dittmar RF, Gallant AR, Gao B (2003) Purebred or hybrid?: Reproducing the volatility in term structure dynamics. J Econometrics 116:147–180
3. Aït-Sahalia Y (2002) Maximum-Likelihood Estimation of Discretely-Sampled Diffusions: A Closed-Form Approximation Approach. Econometrica 70:223–262
4. Aït-Sahalia Y (2007) Closed-Form Likelihood Expansions for Multivariate Diffusions. Annals of Statistics, forthcoming
5. Aït-Sahalia Y, Hansen LP, Scheinkman J (2004) Operator Methods for Continuous-Time Markov Processes. In: Hansen LP, Aït-Sahalia Y (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming
6. Aït-Sahalia Y, Jacod J (2007) Testing for Jumps in a Discretely Observed Process. Annals of Statistics, forthcoming
7. Aït-Sahalia Y, Kimmel R (2007) Maximum Likelihood Estimation of Stochastic Volatility Models. J Finance Econ 83:413–452
8. Alizadeh S, Brandt MW, Diebold FX (2002) Range-Based Estimation of Stochastic Volatility Models. J Finance 57:1047–1091
9. Almeida CIR, Graveline JJ, Joslin S (2006) Do Options Contain Information About Excess Bond Returns? Working Paper, UMN, Fundação Getulio Vargas, MIT, Cambridge
10. Andersen TG (1994) Stochastic Autoregressive Volatility: A Framework for Volatility Modeling. Math Finance 4:75–102
11. Andersen TG (1996) Return Volatility and Trading Volume: An Information Flow Interpretation of Stochastic Volatility. J Finance 51:169–204
12. Andersen TG, Benzoni L (2006) Do Bonds Span Volatility Risk in the US Treasury Market? A Specification Test for Affine Term Structure Models. Working Paper, KSM and Chicago FED, Chicago
13. Andersen TG, Benzoni L (2007) The Determinants of Volatility in the US Treasury market. Working Paper, KSM and Chicago FED, Chicago
14. Andersen TG, Benzoni L (2007) Realized volatility. In: Andersen TG, Davis RA, Kreiss JP, Mikosch T (eds) Handbook of Financial Time Series. Springer, Berlin (forthcoming)
15. Andersen TG, Benzoni L, Lund J (2002) An Empirical Investigation of Continuous-Time Equity Return Models. J Finance 57:1239–1284
16. Andersen TG, Benzoni L, Lund J (2004) Stochastic Volatility, Mean Drift and Jumps in the Short Term Interest Rate. Working Paper, Northwestern University, University of Minnesota, and Nykredit Bank, Copenhagen
17. Andersen TG, Bollerslev T (1998) Deutsche Mark-Dollar Volatility: Intraday Activity Patterns, Macroeconomic Announcements, and Longer Run Dependencies. J Finance 53:219–265
18. Andersen TG, Bollerslev T (1998) Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts. Int Econ Rev 39:885–905
19. Andersen TG, Bollerslev T, Diebold FX (2004) Parametric and nonparametric volatility measurement. In: Hansen LP, Aït-Sahalia Y (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming
20. Andersen TG, Bollerslev T, Diebold FX (2007) Roughing It Up: Including Jump Components in Measuring, Modeling and Forecasting Asset Return Volatility. Rev Econ Statist 89:701–720
21. Andersen TG, Bollerslev T, Diebold FX, Labys P (2003) Modeling and Forecasting Realized Volatility. Econometrica 71:579–625
22. Andersen TG, Bollerslev T, Diebold FX, Vega C (2003) Micro Effects of Macro Announcements: Real-Time Price Discovery in Foreign Exchange. Ammer Econom Rev 93:38–62
23. Andersen TG, Bollerslev T, Dobrev D (2007) No-arbitrage semi-martingale restrictions for continuous-time volatility models subject to leverage effects, jumps and i.i.d. noise: Theory and testable distributional implications. J Econome 138:125–180

24. Andersen TG, Bollerslev T, Meddahi N (2004) Analytic Evaluation of Volatility Forecasts. Int Econ Rev 45:1079–1110
25. Andersen TG, Bollerslev T, Meddahi N (2005) Correcting the Errors: Volatility Forecast Evaluation Using High-Frequency Data and Realized Volatilities. Econometrica 73:279–296
26. Andersen TG, Bondarenko O (2007) Construction and Interpretation of Model-Free Implied Volatility. Working Paper, KSM and UIC, Chicago
27. Andersen TG, Chung HJ, Sørensen BE (1999) Efficient Method of Moments Estimation of a Stochastic Volatility Model: A Monte Carlo Study. J Econom 91:61–87
28. Andersen TG, Lund J (1997) Estimating continuous-time stochastic volatility models of the short term interest rate diffusion. J Econom 77:343–377
29. Ané T, Geman H (2000) Order Flow, Transaction Clock, and Normality of Asset Returns. J Finance 55:2259–2284
30. Ang A, Hodrick RJ, Xing Y, Zhang X (2006) The Cross-Section of Volatility and Expected Returns. J Finance 51:259–299
31. Ang A, Hodrick RJ, Xing Y, Zhang X (2008) High idiosyncratic volatility and low returns: international and further U.S. evidence. Finance Econ (forthcoming)
32. Bachelier L (1900) Théorie de la Spéculation. Annales de École Normale Supérieure 3, Gauthier-Villars, Paris. English translation: Cootner PH (ed) (1964) The Random Character of Stock Market Prices. MIT Press, Cambridge
33. Back K (1991) Asset Prices for General Processes. J Math Econ 20:371–395
34. Bakshi G, Cao C, Chen Z (1997) Empirical Performance of Alternative Option Pricing Models. J Finance 52:2003–2049
35. Bakshi G, Cao C, Chen Z (2002) Pricing and hedging long-term options. J Econom 94:277–318
36. Bakshi G, Kapadia N (2003) Delta-Hedged Gains and the Negative Market Volatility Risk Premium. Rev Finance Stud 16:527–566
37. Bakshi G, Kapadia N, Madan D (2003) Stock Return Characteristics, Skew Laws, and the Differential Pricing of Individual Equity Options. Rev Finance Stud 16:101–143
38. Balduzzi P, Elton EJ, Green TC (2001) Economic News and Bond Prices: Evidence from the US Treasury Market. J Finance Quant Anal 36:523–543
39. Bandi FM (2002) Short-term interest rate dynamics: a spatial approach. J Financ Econ 65:73–110
40. Bandi FM, Moise CE, Russel JR (2008) Market volatility, market frictions, and the cross section of stock returns. Working Paper, University of Chicago, and Case Western Reverse University, Cleveland
41. Bandi FM, Nguyen T (2003) On the functional estimation of jump-diffusion models. J Econom 116:293–328
42. Bandi FM, Phillips PCB (2002) Nonstationary Continuous-Time Processes. In: Hansen LP, Aït-Sahalia Y (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming
43. Bandi FM, Phillips PCB (2003) Fully nonparametric estimation of scalar diffusion models. Econometrica 71:241–283
44. Bandi FM, Phillips PCB (2007) A simple approach to the parametric estimation of potentially nonstationary diffusions. J Econom 137:354–395
45. Bandi FM, Russell J (2006) Separating Microstructure Noise from Volatility. J Finance Econ 79:655–692
46. Bandi FM, Russell J (2007) Volatility. In: Birge J, Linetsky V (eds) Handbook of Financial Engineering. Elsevier, Amsterdam
47. Bansal R, Yaron A (2004) Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles. J Finance 59:1481–1509
48. Bartlett MS (1938) The Characteristic Function of a Conditional Statistic. J Lond Math Soc 13:63–67
49. Barndorff-Nielsen OE, Hansen P, Lunde A, Shephard N (2007) Designing Realized Kernels to Measure the Ex-Post Variation of Equity Prices in the Presence of Noise. Working Paper, University of Aarhus, Aarhus. Stanford University, Nuffield College, Oxford
50. Barndorff-Nielsen OE, Shephard N (2001) Non-Gaussian Ornstein-Uhlenbeckbased models and some of their uses in financial economics. J R Stat Soc B 63:167–241
51. Barndorff-Nielsen OE, Shephard N (2002) Econometric Analysis of Realised Volatility and its Use in Estimating Stochastic Volatility Models. J R Stat Soc B 64:253–280
52. Barndorff-Nielsen OE, Shephard N (2002b) Estimating quadratic variation using realized variance. J Appl Econom 17:457–477
53. Barndorff-Nielsen OE, Shephard N (2004) Power and bipower variation with stochastic volatility and jumps. J Finance Econom 2:1–37
54. Barndorff-Nielsen OE, Shephard N (2006) Econometrics of testing for jumps in financial economics using bipower variation. J Finance Econom 4:1–30
55. Bates DS (1991) The Crash of '87: Was It Expected? The Evidence from Options Markets. J Finance 46:1009–1044
56. Bates DS (1996) Jumps and stochastic volatility: exchange rate processes implicit in deutsche mark options. Rev Finance Stud 9:69–107
57. Bates DS (2000) Post-'87 crash fears in the S&P 500 futures option market. J Econom 94:181–238
58. Bates DS (2006) Maximum Likelihood Estimation of Latent Affine Processes. Rev Finance Stud 19:909–965
59. Benzoni L (2002) Pricing Options under Stochastic Volatility: An Empirical Investigation. Working Paper, Chicago FED
60. Benzoni L, Collin-Dufresne P, Goldstein RS (2007) Explaining Pre- and Post-1987 Crash Prices of Equity and Options within a Unified General Equilibrium Framework. Working Paper, Chicago FED, UCB, and UMN, Minneapolis
61. Bibby BM, Jacobsen M, Sorensen M (2004) Estimating Functions for Discretely Sampled Diffusion-Type Models. In: Hansen LP, Aït-Sahalia Y (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming
62. Bikbov R, Chernov M (2005) Term Structure and Volatility: Lessons from the Eurodollar Markets. Working Paper, LBS, Deutsche Bank, New York
63. Black F, Scholes M (1973) The Pricing of Options and Corporate Liabilities. J Political Econ 81:637–654
64. Bladt M, Sørensen M (2007) Simple Simulation of Diffusion Bridges with Application to Likelihood Inference for Diffusions. Working Paper, University of Copenhagen, Copenhagen
65. Bollen NPB, Whaley RE (2004) Does Net Buying Pressure Affect the Shape of Implied Volatility Functions? J Finance 59:711–753
66. Bollerslev T, Gibson M, Zhou H (2004) Dynamic Estimation of Volatility Risk Premia and Investor Risk Aversion from Option-Implied and Realized Volatilities. Working Paper, Duke University, Federal Reserve Board, Washington D.C.

67. Bollerslev T, Jubinsky PD (1999) Equity Trading vol and Volatility: Latent Information Arrivals and Common Long-Run Dependencies. J Bus Econ Stat 17:9–21
68. Bollerslev T, Zhou H (2002) Estimating stochastic volatility diffusion using conditional moments of integrated volatility. J Econom 109:33–65
69. Bollerslev T, Zhou H (2007) Expected Stock Returns and Variance Risk Premia. Working Paper, Duke University and Federal Reserve Board, Washington D.C.
70. Bondarenko O (2004) Market price of variance risk and performance of hedge funds. Working Paper, UIC, Chicago
71. Brandt MW, Chapman DA (2003) Comparing Multifactor Models of the Term Structure. Working Paper, Duke University and Boston College, Chestnut Hill
72. Brandt MW, Diebold FX (2006) A No-Arbitrage Approach to Range-Based Estimation of Return Covariances and Correlations. J Bus 79:61–73
73. Brandt MW, Jones CS (2006) Volatility Forecasting with Range-Based EGARCH Models. J Bus Econ Stat 24:470–486
74. Brandt MW, Kavajecz KA (2004) Price Discovery in the US Treasury Market: The Impact of Orderflow and Liquidity on the Yield Curve. J Finance 59:2623–2654
75. Brandt MW, Santa-Clara P (2002) Simulated likelihood estimation of diffusions with an application to exchange rate dynamics in incomplete markets. J Finance Econ 63:161–210
76. Breidt FJ, Crato N, de Lima P (1998) The detection and estimation of long memory in stochastic volatility. J Econom 83:325–348
77. Britten-Jones M, Neuberger A (2000) Option Prices, Implied Price Processes, and Stochastic Volatility. J Finance 55:839–866
78. Broadie M, Chernov M, Johannes MJ (2007) Model Specification and Risk Premia: Evidence from Futures Options. J Finance 62:1453–1490
79. Buraschi A, Jackwerth J (2001) The price of a smile: hedging and spanning in option markets. Rev Finance Stud 14:495–527
80. Campbell JY, Cochrane JH (1999) By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior. J Polit Economy 107:205–251
81. Campbell JY, Viceira LM (2001) Who Should Buy Long-Term Bonds? Ammer Econ Rev 91:99–127
82. Carr P, Lee R (2007) Robust Replication of Volatility Derivatives. Working Paper, NYU and Uinversity of Chicago, Chicago
83. Carr P, Madan D (1998) Towards a theory of volatility trading. In: Jarrow R (ed) Volatility. Risk Publications, London
84. Carr P, Wu L (2004) Time-changed Lévy processes and option pricing. J Finance Econ 71:113–141
85. Carr P, Wu L (2007) Variance Risk Premia. Rev Finance Stud, forthcoming
86. Carrasco M, Chernov M, Florens JP, Ghysels E (2007) Efficient estimation of general dynamic models with a continuum of moment conditions. J Econom 140:529–573
87. Carrasco M, Florens JP (2000) Generalization of GMM to a continuum of moment conditions. Econom Theory 16:797–834
88. Chacko G, Viceira LM (2003) Spectral GMM estimation of continuous-time processes. J Econom 116:259–292
89. Chan KC, Karolyi GA, Longstaff FA, Sanders AB (1992) An Empirical Comparison of Alternative Models of the Short-Term Interest Rate. J Finance 47:1209–1227
90. Chen J (2003) Intertemporal CAPM and the Cross-Section of Stock Return. Working Paper, USC, Los Angeles
91. Chen RR, Scott L (1993) Maximum likelihood estimation for a multifactor equilibrium model of the term structure of interest rates. J Fixed Income 3:14–31
92. Cheridito P, Filipović D, Kimmel RL (2007) Market Price of Risk Specifications for Affine Models: Theory and Evidence. J Finance Econ, 83(1):123–170
93. Chernov M, Ghysels E (2002) A study towards a unified approach to the joint estimation of objective and risk neutral measures for the purpose of options valuation. J Finance Econ 56:407–458
94. Chernov M, Gallant AR, Ghysels E, Tauchen G (1999) A New Class of Stochastic Volatility Models with Jumps: Theory and Estimation. Working Paper, London Business School, Duke University, University of Northern Carolina
95. Chernov M, Gallant AR, Ghysels E, Tauchen G (2003) Alternative models for stock price dynamics. J Econom 116:225–257
96. Chib S, Nardari F, Shephard N (2002) Markov chain Monte Carlo methods for stochastic volatility models. J Econom 108:281–316
97. Chib S, Nardari F, Shephard N (2006) Analysis of high dimensional multivariate stochastic volatility models. J Econom 134:341–371
98. Christoffersen PF, Jacobs K (2004) The importance of the loss function in option valuation. J Finance Econ 72:291–318
99. Christoffersen PF, Jacobs K, Karoui L, Mimouni K (2007) Estimating Term Structure Models Using Swap Rates. Working Paper, McGill University, Montreal
100. Christoffersen PF, Jacobs K, Mimouni K (2006a) Models for S&P 500 Dynamics: Evidence from Realized Volatility, Daily Returns, and Option Prices. Working Paper, McGill University, Montreal
101. Christoffersen PF, Jacobs K, Wang Y (2006b) Option Valuation with Long-run and Short-run Volatility Components. Working Paper, McGill University, Montreal
102. Clark PK (1973) A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices. Econometrica 41:135–156
103. Collin-Dufresne P, Goldstein RS (2002) Do Bonds Span the Fixed Income Markets? Theory and Evidence for Unspanned Stochastic Volatility. J Finance 57:1685–1730
104. Collin-Dufresne P, Goldstein R, Jones CS (2008) Identification of Maximal Affine Term Structure Models. J Finance 63(2):743–795
105. Collin-Dufresne P, Goldstein R, Jones CS (2007b) Can Interest Rate Volatility be Extracted from the Cross Section of Bond Yields? An Investigation of Unspanned Stochastic Volatility. Working Paper, UCB, USC, UMN, Minneapolis
106. Collin-Dufresne P, Solnik B (2001) On the Term Structure of Default Premia in the Swap and LIBOR Markets. J Finance 56:1095–1115
107. Comte F, Coutin L, Renault E (2003) Affine Fractional Stochastic Volatility Models with Application to Option Pricing. Working Paper, CIRANO, Montreal
108. Comte F, Renault E (1998) Long memory in continuous-time stochastic volatility models. Math Finance 8:291–323
109. Conley TG, Hansen LP, Luttmer EGJ, Scheinkman JA (1997) Short-term interest rates as subordinated diffusions. Rev Finance Stud 10:525–577

110. Corsi F (2003) A Simple Long Memory Model of realized Volatility. Working paper, University of Southern Switzerland, Lugano
111. Coval JD, Shumway T (2001) Expected Option Returns. J Finance 56:983–1009
112. Cox JC, Ingersoll JE, Ross SA (1985) An Intertemporal General Equilibrium Model of Asset Prices. Econometrica 53:363–384
113. Cox JC, Ingersoll JE, Ross SA (1985) A Theory of the Term Structure of Interest Rates. Econometrica 53:385–407
114. Dai Q, Singleton KJ (2000) Specification Analysis of Affine Term Structure Models. J Finance 55:1943–1978
115. Dai Q, Singleton KJ (2003) Term Structure Dynamics in Theory and Reality. Rev Finance Stud 16:631–678
116. Danielsson J (1994) Stochastic Volatility in Asset Prices: Estimation by Simulated Maximum Likelihood. J Econom 64:375–400
117. Danielsson J, Richard JF (1993) Accelerated Gaussian Importance Sampler with Application to Dynamic Latent Variable Models. J Appl Econom 8:S153–S173
118. Deo R, Hurvich C (2001) On the Log Periodogram Regression Estimator of the Memory Parameter in Long Memory Stochastic Volatility Models. Econom Theory 17:686–710
119. Derman E, Kani I (1994) The volatility smile and its implied tree. Quantitative Strategies Research Notes, Goldman Sachs, New York
120. Diebold FX, Nerlove M (1989) The Dynamics of Exchange Rate Volatility: A Multivariate Latent Factor ARCH Model. J Appl Econom 4:1–21
121. Diebold FX, Strasser G (2007) On the Correlation Structure of Microstructure Noise in Theory and Practice. Working Paper, University of Pennsylvania, Philadelphia
122. Dobrev D (2007) Capturing Volatility from Large Price Moves: Generalized Range Theory and Applications. Working Paper, Federal Reserve Board, Washington D.C.
123. Driessen J, Maenhout P, Vilkov G (2006) Option-Implied Correlations and the Price of Correlation Risk. Working Paper, University of Amsterdam and INSEAD, Amsterdam
124. Duarte J (2004) Evaluating An Alternative Risk Preference in Affine Term Structure Models. Rev Finance Stud 17:370–404
125. Duarte J (2007) The Causal Effect of Mortgage Refinancing on Interest-Rate Volatility: Empirical Evidence and Theoretical Implications. Rev Finance Stud, forthcoming
126. Duarte J, Jones CS (2007) The Price of Market Volatility Risk. Working Paper, University of Washington and USC, Washington
127. Duffee GR (2002) Term Premia and Interest Rate Forecasts in Affine Models. J Finance 57:405–443
128. Duffee G, Stanton R (2008) Evidence on simulation inference for near unit-root processes with implications for term structure estimation. J Finance Econom 6:108–142
129. Duffie D, Kan R (1996) A yield-factor model of interest rates. Math Finance 6:379–406
130. Duffie D, Pan J, Singleton KJ (2000) Transform Analysis and Asset Pricing for Affine Jump-Diffusions. Econometrica 68:1343–1376
131. Duffie D, Singleton KJ (1993) Simulated Moments Estimation of Markov Models of Asset Prices. Econometrica 61:929–952
132. Duffie D, Singleton KJ (1997) An Econometric Model of the Term Structure of Interest-Rate Swap Yields. J Finance 52:1287–1321
133. Dumas B, Fleming J, Whaley RE (1996) Implied Volatility Functions: Empirical Tests. J Finance 53:2059–2106
134. Dupire B (1994) Pricing with a smile. Risk 7:18–20
135. Elerian O, Chib S, Shephard N (2001) Likelihood Inference for Discretely Observed Nonlinear Diffusions. Econometrica 69:959–994
136. Engle RF (2002) New frontiers for ARCH models. J Appl Econom 17:425–446
137. Engle RF, Gallo GM (2006) A multiple indicators model for volatility using intra-daily data. J Econom 131:3–27
138. Engle RF, Ghysels E, Sohn B (2006) On the Economic Sources of Stock Market Volatility. Working Paper, NYU and UNC, Chapel Hill
139. Engle RF, Rangel JG (2006) The Spline-GARCH Model for Low Frequency Volatility and Its Global Macroeconomic Causes. Working Paper, NYU, New York
140. Eraker B (2001) MCMC Analysis of Diffusions with Applications to Finance. J Bus Econ Stat 19:177–191
141. Eraker B (2004) Do Stock Prices and Volatility Jump? Reconciling Evidence from Spot and Option Prices. J Finance 59:1367–1404
142. Eraker B, Johannes MS, Polson N (2003) The Impact of Jumps in Volatility and Returns. J Finance 58:1269–1300
143. Fan R, Gupta A, Ritchken P (2003) Hedging in the Possible Presence of Unspanned Stochastic Volatility: Evidence from Swaption Markets. J Finance 58:2219–2248
144. Fiorentina G, Sentana E, Shephard N (2004) Likelihood-Based Estimation of Latent Generalized ARCH Structures. Econometrica 72:1481–1517
145. Fisher M, Gilles C (1996) Estimating exponential-affine models of the term structure. Working Paper, Atlanta FED, Atlanta
146. Fleming MJ, Remolona EM (1999) Price Formation and Liquidity in the US Treasury Market: The Response to Public Information. J Finance 54:1901–1915
147. Forsberg L, Ghysels E (2007) Why Do Absolute Returns Predict Volatility So Well? J Finance Econom 5:31–67
148. French KR, Schwert GW, Stambaugh RF (1987) Expected stock returns and volatility. J Finance Econ 19:3–29
149. Fridman M, Harris L (1998) A Maximum Likelihood Approach for Non-Gaussian Stochastic Volatility Models. J Bus Econ Stat 16:284–291
150. Gallant AR, Hsieh DA, Tauchen GE (1997) Estimation of Stochastic Volatility Models with Diagnostics. J Econom 81:159–192
151. Gallant AR, Hsu C, Tauchen GE (1999) Using Daily Range Data to Calibrate Volatility Diffusions and Extract the Forward Integrated Variance. Rev Econ Stat 81:617–631
152. Gallant AR, Long JR (1997) Estimating stochastic differential equations efficiently by minimum chi-squared. Biometrika 84:125–141
153. Gallant AR, Rossi PE, Tauchen GE (1992) Stock Prices and Volume. Rev Finance Stud 5:199–242
154. Gallant AR, Tauchen GE (1996) Which Moments to Match. Econ Theory 12:657–681
155. Gallant AR, Tauchen G (1998) Reprojecting Partially Observed Systems With Application to Interest Rate Diffusions. J Ammer Stat Assoc 93:10–24
156. Garcia R, Lewis MA, Pastorello S, Renault E (2001) Estimation of Objective and Risk-neutral Distributions based on Moments of Integrated Volatility, Working Paper. Univer-

sité de Montréal, Banque Nationale du Canada, Università di Bologna, UNC
157. Garman MB, Klass MJ (1980) On the Estimation of Price Volatility From Historical Data. J Bus 53:67–78
158. Ghysels E, Harvey AC, Renault E (1996) Stochastic Volatility. In: Maddala GS, Rao CR (eds) Handbook of Statistics, vol 14. North Holland, Amsterdam
159. Ghysels E, Santa-Clara P, Valkanov R (2006) Predicting Volatility: How to Get the Most Out of Returns Data Sampled at Different Frequencies. J Econom 131:59–95
160. Glosten LR, Jagannathan R, Runkle D (1993) On the Relation Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. J Finance 48:1779–1801
161. Gong FF, Remolona EM (1996) A three-factor econometric model of the US term structure. Working paper, Federal Reserve Bank of New York, New York
162. Gouriéroux C, Monfort A, Renault E (1993) Indirect Inference. J Appl Econom 8:S85–S118
163. Guo H, Neely CJ, Higbee J (2007) Foreign Exchange Volatility Is Priced in Equities. Finance Manag, forthcoming
164. Han B (2007) Stochastic Volatilities and Correlations of Bond Yields. J Finance 62:1491–1524
165. Hansen PR, Lunde A (2006) Realized Variance and Market Microstructure Noise. J Bus Econ Stat 24:127–161
166. Harvey AC (1998) Long memory in stochastic volatility. In: Knight J, Satchell S (eds) Forecasting Volatility in Financial Markets. Butterworth-Heinemann, London
167. Harvey AC, Ruiz E, Shephard N (1994) Multivariate Stochastic Variance Models. Rev Econ Stud 61:247–264
168. Harvey AC, Shephard N (1996) Estimation of an Asymmetric Stochastic Volatility Model for Asset Returns. J Bus Econ Stat 14:429–434
169. Heidari M, Wu L (2003) Are Interest Rate Derivatives Spanned by the Term Structure of Interest Rates? J Fixed Income 13:75–86
170. Heston SL (1993) A closed-form solution for options with stochastic volatility with applications to bond and currency options. Rev Finance Stud 6:327–343
171. Ho M, Perraudin W, Sørensen BE (1996) A Continuous Time Arbitrage Pricing Model with Stochastic Volatility and Jumps. J Bus Econ Stat 14:31–43
172. Huang X, Tauchen G (2005) The relative contribution of jumps to total price variation. J Finance Econom 3:456–499
173. Huang J, Wu L (2004) Specification Analysis of Option Pricing Models Based on Time-Changed Levy Processes. J Finance 59:1405–1440
174. Hull J, White A (1987) The Pricing of Options on Assets with Stochastic Volatilities. J Finance 42:281–300
175. Hurvich CM, Soulier P (2007) Stochastic Volatility Models with Long Memory. In: Andersen TG, Davis RA, Kreiss JP, Mikosch T (eds) Handbook of Financial Time Series. Springer,Berlin
176. Jung J (2006) Vexed by variance. Risk August
177. Jackwerth JC, Rubinstein M (1996) Recovering Probability Distributions from Option Prices. J Finance 51:1611–1631
178. Jacobs K, Karoui L (2007) Conditional Volatility in Affine Term Structure Models: Evidence from Treasury and Swap Markets. Working Paper, McGill University, Montreal
179. Jacobs K, Li X (2008) Modeling the Dynamics of Credit Spreads with Stochastic Volatility. Manag Sci 54:1176–1188
180. Jacquier E, Polson NG, Rossi PE (1994) Bayesian Analysis of Stochastic Volatility Models. J Bus Econ Stat 12:371–389
181. Jacquier E, Polson NG, Rossi PE (2004) Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. J Econom 122:185–212
182. Jagannathan R, Kaplin A, Sun S (2003) An evaluation of multifactor CIR models using LIBOR, swap rates, and cap and swaption prices. J Econom 116:113–146
183. Jarrow R, Li H, Zhao F (2007) Interest Rate Caps "Smile" Too! But Can the LIBOR Market Models Capture the Smile? J Finance 62:345–382
184. Jiang GJ, Knight JL (2002) Efficient Estimation of the Continuous Time Stochastic Volatility Model via the Empirical Characteristic Function. J Bus Econ Stat 20:198–212
185. Jiang GJ, Tian YS (2005) The Model-Free Implied Volatility and Its Information Content. Rev Finance Stud 18:1305–1342
186. Johannes MS, Polson N (2003) MCMC Methods for Continuous-Time Financial Econometrics. In: Hansen LP, Aït-Sahalia I (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming
187. Johannes MS, Polson N (2006) Particle Filtering. In: Andersen TG, Davis RA, Kreiss JP, Mikosch T (eds) Handbook of Financial Time Series. Springer, Berlin
188. Johnson H, Shanno D (1987) Option Pricing when the Variance Is Changing. J Finance Quant Anal 22:143–152
189. Jones CS (2003) Nonlinear Mean Reversion in the Short-Term Interest Rate. Rev Finance Stud 16:793–843
190. Jones CS (2003) The dynamics of stochastic volatility: evidence from underlying and options markets. J Econom 116:181–224
191. Joslin S (2006) Can Unspanned Stochastic Volatility Models Explain the Cross Section of Bond Volatilities? Working Paper, MIT, Cambridge
192. Joslin S (2007) Pricing and Hedging Volatility Risk in Fixed Income Markets. Working Paper, MIT, Cambridge
193. Kim SN, Shephard N, Chib S (1998) Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models. Rev Econ Stud 65:361–393
194. Lamoureux CG, Lastrapes WD (1994) Endogenous Trading vol and Momentum in Stock-Return Volatility. J Bus Econ Stat 14:253–260
195. Lee SS, Mykland PA (2006) Jumps in Financial Markets: A New Nonparametric Test and Jump Dynamics. Working Paper, Georgia Institute of Technology and University of Chicago, Chicago
196. Li H, Wells MT, Yu CL (2006) A Bayesian Analysis of Return Dynamics with Lévy Jumps. Rev Finance Stud, forthcoming
197. Li H, Zhao F (2006) Unspanned Stochastic Volatility: Evidence from Hedging Interest Rate Derivatives. J Finance 61:341–378
198. Liesenfeld R (1998) Dynamic Bivariate Mixture Models: Modeling the Behavior of Prices and Trading Volume. J Bus Econ Stat 16:101–109
199. Liesenfeld R (2001) A Generalized Bivariate Mixture Model for Stock Price Volatility and Trading Volume. J Econom 104:141–178
200. Liesenfeld R, Richard J-F (2003) Univariate and Multivariate Stochastic Volatility Models: Estimation and Diagnostics. J Empir Finance 10:505–531
201. Litterman R, Scheinkman JA (1991) Common Factors Affecting Bond Returns. J Fixed Income 1:54–61
202. Liu C, Maheu JM (2007) Forecasting Realized Volatility: A Bayesian Model Averaging Approach. Working Paper, University of Toronto, Toronto

203. Lo AW (1988) Maximum likelihood estimation of generalized Itô processes with discretely-sampled data. Econom Theory 4:231–247
204. Longstaff FA, Schwartz ES (1992) Interest Rate Volatility and the Term Structure: A Two-Factor General Equilibrium Model. J Finance 47:1259–1282
205. McAleer M, Medeiros MC (2007) Realized Volatility: A Review. Econom Rev, forthcoming
206. McFadden D (1989) A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration. Econometrica 57:995–1026
207. McQueen G, Vorkink K (2004) Whence GARCH? A Preference-Based Explanation for Conditional Volatility. Rev Finance Stud 17:915–949
208. Meddahi N (2001) An Eigenfunction Approach for Volatility Modeling. Working Paper, Imperial College, London
209. Meddahi N (2002) A Theoretical Comparison Between Integrated and Realized Volatility. J Appl Econom 17:475–508
210. Meddahi N (2002) Moments of Continuous Time Stochastic Volatility Models. Working Paper, Imperial College, London
211. Melino A, Turnbull SM (1990) Pricing foreign currency options with stochastic volatility. J Econom 45:239–265
212. Merton RC (1969) Lifetime portfolio selection under uncertainty: the continuous-time case. Rev Econ Stat 51:247–257
213. Merton RC (1973) An Intertemporal Capital Asset Pricing Model. Econometrica 41:867–887
214. Merton RC (1976) Option pricing when underlying stock returns are discontinuous. J Finance Econs 3:125–144
215. Merton RC (1980) On estimating the expected return on the market: An exploratory investigation. J Finance Econ 8:323–361
216. Mizrach B (2006) The Enron Bankruptcy: When Did The Options Market Lose Its Smirk. Rev Quant Finance Acc 27:365–382
217. Mizrach B (2007) Recovering Probabilistic Information From Options Prices and the Underlying. In: Lee C, Lee AC (eds) Handbook of Quantitative Finance. Springer, New York
218. Nelson DB (1991) Conditional Heteroskedasticity in Asset Returns: A New Approach. Econometrica 59:347–370
219. Nelson DB, Foster DP (1994) Asymptotic Filtering Theory for Univariate ARCH Models. Econometrica 62:1–41
220. Pan J (2002) The jump-risk premia implicit in options: evidence from an integrated time-series study. J Finance Econ 63:3–50
221. Parkinson M (1980) The Extreme ValueMethod for Estimating the Variance of the Rate of Return. J Bus 53:61–65
222. Pasquariello P, Vega C (2007) Informed and Strategic Order Flow in the Bond Markets. Rev Finance Stud 20:1975–2019
223. Pearson ND, Sun TS (1994) Exploiting the Conditional Density in Estimating the Term Structure: An Application to the Cox, Ingersoll, and Ross Model. J Finance 49:1279–1304
224. Pedersen AR (1995) A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. Scand J Stat 22:55–71
225. Pennacchi GG (1991) Identifying the Dynamics of Real Interest Rates and Inflation: Evidence Using Survey Data. Rev Finance Stud 4:53–86
226. Piazzesi M (2003) Affine Term Structure Models. In: Hansen LP, Aït-Sahalia I (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming
227. Piazzesi M (2005) Bond Yields and the Federal Reserve. J Political Econ 113:311–344
228. Pitt MK, Shephard N (1999) Filtering via simulation: auxiliary particle filter. J Ammer Stat Assoc 94:590–599
229. Protter P (1992) Stochastic Integration and Differential Equations: A New Approach. Springer, New York
230. Renault E (1997) Econometric models of option pricing errors. In: Kreps D, Wallis K (eds) Advances in Economics and Econometrics, Seventh World Congress. Cambridge University Press, New York, pp 223–278
231. Richardson M, Smith T (1994) A Direct Test of the Mixture of Distributions Hypothesis: Measuring the Daily Flow of Information. J Financ Quant Anal 29:101–116
232. Rosenberg B (1972) The Behavior of Random Variables with Nonstationary Variance and the Distribution of Security Prices. working Paper, UCB, Berkley
233. Rubinstein M (1994) Implied Binomial Trees. J Financ 49:771–818
234. Santa-Clara P (1995) Simulated likelihood estimation of diffusions with an application to the short term interest rate. Dissertation, INSEAD
235. Schaumburg E (2005) Estimation of Markov processes with Levy type generators. Working Paper, KSM, Evanston
236. Schwert GW (1989) Why Does Stock Market Volatility Change Over Time? J Financ 44:1115–1153
237. Schwert GW (1990) Stock Volatility and the Crash of '87. Rev Financ Stud 3:77–102
238. Scott LO (1987) Option Pricing when the Variance Changes Randomly: Theory, Estimation and an Application. J Financ Quant Anal 22:419–438
239. Shephard N (1996) Statistical Aspects of ARCH and Stochastic Volatility Models. In: Cox DR, Hinkley DV, Barndorff-Nielsen OE (eds) Time Series Models in Econometrics, Finance and Other Fields. Chapman & Hall, London, pp 1–67
240. Shephard N (2004) Stochastic Volatility: Selected Readings. Oxford University Press, Oxford
241. Shiller RJ (1981) Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends? Ammer Econ Rev 71:421–436
242. Singleton KJ (2001) Estimation of affine asset pricing models using the empirical characteristic function. J Econom 102:111–141
243. Smith AA Jr (1993) Estimating Nonlinear Time-series Models using Simulated Vector Autoregressions. J Appl Econom 8:S63–S84
244. Stein JC (1989) Overreactions in the Options Market. J Finance 44:1011–1023
245. Stein EM, Stein JC (1991) Stock price distributions with stochastic volatility: an analytic approach. Rev Finance Stud 4:727–752
246. Tauchen GE (2005) Stochastic Volatility in General Equilibrium. Working Paper, Duke, University Durham
247. Tauchen GE, Pitts M (1983) The Price Variability-Volume Relationship on Speculative Markets. Econometrica 51:485–505
248. Tauchen GE, Zhou H (2007) Realized Jumps on Financial Markets and Predicting Credit Spreads. Working Paper, Duke University and Board of Governors, Washington D.C.
249. Taylor SJ (1986) Modeling Financial Time Series. Wiley, Chichester

250. Thompson S (2004) Identifying Term Structure Volatility from the LIBOR-Swap Curve. Working Paper, Harvard University, Boston
251. Todorov V (2006) Variance Risk Premium Dynamics. Working Paper, KSM, Evanston
252. Todorov V (2006) Estimation of Continuous-time Stochastic Volatility Models with Jumps using High-Frequency Data. Working Paper, KSM, Evanston
253. Trolle AB, Schwartz ES (2007) Unspanned stochastic volatility and the pricing of commodity derivatives. Working Paper, Copenhagen Business School and UCLA, Copenhagen
254. Trolle AB, Schwartz ES (2007) A general stochastic volatility model for the pricing of interest rate derivatives. Rev Finance Stud, forthcoming
255. Vasicek OA (1977) An equilibrium characterization of the term structure. J Finance Econ 5:177–188
256. Wiggins JB (1987) Option Values under Stochastic Volatility: Theory and Empirical Estimates. J Finance Econ 19:351–372
257. Whaley RE (1993) Derivatives on Market Volatility: Hedging Tools Long Overdue. J Deriv 1:71–84
258. Wright J, Zhou H (2007) Bond Risk Premia and Realized Jump Volatility. Working Paper, Board of Governors, Washington D.C.
259. Yang D, Zhang Q (2000) Drift-Independent Volatility Estimation Based on High, Low, Open, and Close Prices. J Bus 73:477–491
260. Zhang BY, Zhou H, Zhu H (2005) Explaining Credit Default Swap Spreads with the Equity Volatility and Jump Risks of Individual Firms. Working Paper, Fitch Ratings, Federal Reserve Board, BIS
261. Zhang L (2007) What you don't know cannot hurt you: On the detection of small jumps. Working Paper, UIC, Chicago
262. Zhang L, Mykland PA, Aït-Sahalia Y (2005) A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High Frequency Data. J Ammer Stat Assoc 100:1394–1411

Books and Reviews

Asai M, McAleer M, Yu J (2006) Multivariate Stochastic Volatility: A Review. Econom Rev 25:145–175
Bates DS (2003) Empirical option pricing: a retrospection. J Econom 116:387–404
Campbell JY, Lo AW, MacKinlay AC (1996) The Econometrics of Financial Markets. Princeton University Press, Princeton
Chib S, Omori Y, Asai M (2007) Multivariate Stochastic Volatility. Working Paper, Washington University
Duffie D (2001) Dynamic Asset Pricing Theory. Princeton University Press, Princeton
Gallant AR, Tauchen G (2002) Simulated Score Methods and Indirect Inference for Continuous-time Models. In: Hansen LP, Aït-Sahalia Y (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming
Garcia R, Ghysels E, Renault E (2003) The Econometrics of Option Pricing. In: Hansen LP, Aït-Sahalia Y (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming
Gouriéroux C, Jasiak J (2001) Financial Econometrics. Princeton University Press, Princeton
Johannes MS, Polson N (2006) Markov Chain Monte Carlo. In: Andersen TG, Davis RA, Kreiss JP, Mikosch T (eds) Handbook of Financial Time Series. Springer, Berlin
Jungbacker B, Koopman SJ (2007) Parameter Estimation and Practical Aspect of Modeling Stochastic Volatility. Working Paper, Vrije Universiteit, Amsterdam
Mykland PA (2003) Option Pricing Bounds and Statistical Uncertainty. In: Hansen LP, Aït-Sahalia Y (eds) Handbook of Financial Econometrics. North-Holland, Amsterdam, forthcoming
Renault E (2007) Moment-Based Estimation of Stochastic Volatility Models. In: Andersen TG, Davis RA, Kreiss JP, Mikosch T (eds) Handbook of Financial Time Series. Springer, Berlin
Singleton KJ (2006) Empirical Dynamic Asset Pricing: Model Specification and Econometric Assessment. Princeton University Press, Princeton

Structurally Dynamic Cellular Automata

Andrew Ilachinski
Center for Naval Analyses, Alexandria, USA

Article Outline

Glossary
Definition of the Subject
Introduction
The Basic Model
Emerging Patterns and Behaviors
SDCA as Models of Computation
Generalized SDCA Models
Related Graph Dynamical Systems
SDCA as Models of Fundamental Physics
Future Directions and Speculations
Bibliography

Glossary

Adjacency matrix The *adjacency matrix* of a graph with N sites is an $N \times N$ matrix $[a_{ij}]$ with entries $a_{ij} = 1$ if i and j are linked, and $a_{ij} = 0$ otherwise. The adjacency matrix is symmetric ($a_{ij} = a_{ji}$) if the links in the graph are undirected.

Coupler link rules *Coupler rules* are local rules that act on pairs of *next-nearest sites* of a graph at time t to decide whether they should be linked at $t + 1$. The decision rules fall into one of three basic classes – totalistic (T), outer-totalistic (OT) or restricted-totalistic (RT) – but can be as varied as those for conventional cellular automata.

Decoupler link rules *Decoupler rules* are local rules that act on pairs of *linked sites* of a graph at time t to decide whether they should be unlinked at $t + 1$. As for

coupler rules, the decision rules fall into one of three basic classes – totalistic (T), outer-totalistic (OT) or restricted-totalistic (RT) – but can be as varied as those for conventional cellular automata.

Degree The *degree* of a node (or site, i) of a graph is equal to the number of distinct nodes to which i is linked, and where the links are assumed to possess no directional information. In general graphs, the *in*-degree (= number of incoming links towards i) is distinguished from the *out*-degree (= number of outgoing links originating at i).

Effective dimension A quantity used to approximate the *dimensionality* of a graph. It is defined as the ratio between the average number of next-nearest neighbors to the average degree, both averaged over all nodes of the graph. The effective dimension equals the Euclidean dimension d, in cases where the graph is the familiar d-dimensional hypercubic lattice.

Graph A *graph* is a finite, nonempty set of nodes (referred to as "sites" throughout this article), together with (a possibly empty) set of edges (or links). The links may be either *directed* (in which case the edge from a site i, say, is directed away from i toward another site j, and is considered distinct from another directed edge originating at j and pointed toward i) or *undirected* (in which case if a link exists between sites i and j it carries no directional information).

Graph grammar *Graph grammars* (sometimes also referred to as *graph rewriting systems*) apply formal language theory to networks. Each language specifies the space of "valid structures", and the production (or "rewrite") rules by which given graphs may be transformed into other valid graphs.

Graph metric function The *graph metric function* defines the distance between any two nodes, i and j. It is equal to the length of the shortest path between i and j. If no path exists (such as when i and j are on two disconnected components of the same graph), the distance is assumed to be equal to ∞.

Graph-rewriting automata *Graph-rewriting automata* are generalized CA-like systems in which both (the number of) nodes and links are allowed to change.

Next-nearest neighbor Two sites i and j are *next-nearest neighbors* in a graph if (1) they are not directly linked (so that $a_{ij} = 0$; see *adjacency matrix*), and (2) there exists at least one other site k such that $k \notin \{i, j\}$, and i and j are both lined to k.

Random dynamics approximation The long-term behavior of structurally dynamic cellular automata may be approximated in certain cases (in which the structure and value configurations are both sufficiently random and uncorrelated) by a *random dynamics approximation*: values of sites are replaced by the probability p_σ of a site having value σ (and is assumed to be equal for all sites), and links between sites are replaced by the probability p_ℓ of being linked (and also assumed to be the same for all pairs of sites). The approximation often yields qualitatively correct predictions about how the real system evolves under a specific set of rules; for example, to predict whether one expects unbounded growth or that the lattice will eventually settle onto a low periodic state or simply decay.

Restricted totalistic rules *Restricted totalistic rules* are a generalized class of link rules (operating on pairs of sites, i and j), analogous to "outer totalistic" rules (that operate on site values) used in conventional CA. The local neighborhood around i and j is first partitioned into three sets: (1) the two sites, i and j; (2) sites connected to either i or j, but not both; and (3) sites connected to both i and j. The *restricted totalistic rule* is then completely defined by associating a specific action with each possible 3-tuple of site-value sums (where the individual components represent a unique sum in each of the three neighborhoods).

Structurally dynamic cellular automata *Structurally dynamic cellular automata* are generalizations of conventional cellular automata models in which the underlying lattice structure is dynamically coupled to the local site-value configurations.

SDCA model hierarchy The *SDCA model hierarchy* is a set of eight related structurally dynamic cellular automata models, defined explicitly for studying their formal computational capabilities. The hierarchy is ordered (from lowest to highest level) according to their relative computational strength. For example, the SDCA model at the top of the hierarchy is capable of simulating a conventional CA with a speedup factor of two.

Definition of the Subject

Structurally dynamic cellular automata (abbreviated, SDCA) are a generalized class of CA in which the topological structure of the (usually quiescent) underlying lattice is dynamically coupled to the local site value configuration. The coupling is defined to treat *geometry* and *value configurations* on an approximately equal footing: the lattice structure is altered locally as a function of individual site neighborhood value-states and geometries, while the underlying local topology supports site-value evolution precisely as in conventional nearest-neighbor CA models defined on random lattices.

SDCA provide a dynamical framework for a CA-like analysis of the generation, transmission and interaction of topological disturbances in a lattice. Moreover, they provide a natural testbed for studying self organized geometry; by which we mean true structural evolution, and not merely space-time patterns of value configurations that may be *interpreted* geometrically (but are really just "bits" of information overlayed on top of an otherwise static background lattice).

Introduction

SDCA were formally introduced in 1986 as part of a physics doctoral dissertation by Ilachinski [31], and developed further by Ilachinski and Halpern [29,30], Halpern [21,23], Halpern and Caltagirone [22], Majercik [39], and Alonso-Sanz and Martín [6,7,8]; in their original incarnation [28], and at least two subsequent papers [22,61], SDCA were called *topological automata*. Pedagogical discussions appear in Adamatzky [1] and Ilachinski [32]. Extensions of the basic SDCA model (all discussed in this article) include the addition of probabilistic rules, memory and reversibility.

Applications include the simulation of crystal growth [36], the study of pattern formation of random cellular structures [66], modeling synaptic plasticity in neural network models [19], phase transitions in chemical systems [62], chemical self-assembly [26], and gene-regulatory networks [22]. Majercik [39] has studied SDCA as generalized models of computation, and describes a CA-universal SDCA that can simulate any conventional CA of the same dimension.

More recently, O'Sullivan [56] and Saidani [63,64] have used graph-based CA models similar to SDCA to study urban dynamics and emergent behaviors of self-reconfigurable robots, respectively. Tomita et al. [67,68,69,70,71] have introduced *graph-rewriting automata* in which both links and (the number of) nodes are allowed to change; and show that these systems are capable of both self-replication and Turing universality (among with many other emergent behaviors). Since SDCA provide the basic formalism for describing locally induced topological changes within arbitrary graphs, they are a potentially powerful general tool for studying complex adaptive networks, such as communication and social networks [6]. The *concept* behind SDCA has also been used as a foundation for philosophical musings about computationally emergent artificiality [49].

More ambitious applications of SDCA encroach on fundamental physics. Because SDCA are inherently self-modifying systems – in which physical events are not just dynamically coupled to, but are an integral part of the spatio-temporal arena on which their transformations are defined – they are a potentially powerful methodological and ontological tool for exploring discrete pre-geometric theories of space-time [42]. Just as "value structure" solitons are ubiquitous in conventional CA models [32,74], "link structure" solitons might emerge in SDCA; physical particles would, in such a scheme, be viewed as geometrodynamic disturbances propagating within a dynamic lattice. Three SDCA-like theories of pregeometry have recently been proposed in which space-time is a self-organized emergent construct: Hillman [27], Nowotny and Requardt [55] and Wolfram [75].

Finally, we briefly comment on ostensible overlaps between SDCA and four other related fields of study: (1) *Lindenmeyer* (or L-) *systems*, (2) *graph grammars*, (3) *random graphs* (abbreviated, RG), and (4) *dynamic network analysis* (abbreviated, DNA). L-systems [57] are generalized CA systems in which the number of sites can grow with time, and consist of recursive rules for rewriting strings of symbols. If interpreted graphically, abstract symbol strings can be used to model growth processes of plants and evolving morphology of physical organisms. Graph grammars [20,35] apply formal language theory to networks, and consist of production rules that define the set of "valid structures" in a given graph language. The study of RG [15] was introduced by Erdos and Renyi in the late 1950s [16], and is a mathematical framework for exploring the general topological structures of computational systems and the behavior of certain random dynamical systems. Like SDCA, RG describes evolving graphs, but the dynamics are global and random. DNA [41,50] is an emerging field that fuses traditional social network theory with statistical analysis and modeling; part of its charter is to explore general properties of network generation and evolution.

While, conceptually speaking, there is a prima facie relationship between SDCA and all four fields of study, the elucidation of a more precise nature of the relationship between SDCA and these other systems awaits a future study. (The relationship appears particularly strong between SDCA and a generalized L-system called the *graph development system* (abbreviated, GDS), introduced by Doi [14], but not developed further since its original conception. Using incidence matrices to represent arbitrary topologies, GDS is essentially a grammar by which submatrices of the whole matrix are rewritten to describe topological changes. SDCA also formally falls under the broader rubrics of DNA and RG; however, there is no explicit reference to SDCA in the current literature of either field.)

The Basic Model

Conventional CA are defined on fixed, and typically regular, lattices (one-dimensional lines, two-dimensional Euclidean or hexagonal grids, etc.), the sites of which are populated with discrete-valued dynamic elements ($\equiv \sigma_i \in \{0, 1, \ldots, k\}$, where i labels a particular site on the lattice) that evolve according to local transition functions, $f\colon \sigma_i \to \sigma_i'$. We emphasize that the dynamics of conventional CA are confined to the temporal evolution of the σ_is.

SDCA generalize conventional CA in two ways: (1) they relax the assumption that the underlying lattice is uniform, allowing the local site $\leftrightarrow$ site connectivity pattern to vary throughout the lattice; and (2) they allow *both* the set $\{\sigma_i\}$ *and* the lattice to evolve according to local transition rules. The most obvious – also the most dramatic – conceptual change this entails over the dynamics of conventional CA, is that the meaning of "local" itself changes as a function of how the SDCA system evolves: previously far separated sites may become neighbors; and sites that are local at time t may become far separated at some later time, t'.

To properly define SDCA, we first generalize regular lattices to mathematical graphs $\equiv G$ () possessing arbitrary topology. Assuming G has N lattice sites, and that G is (for now) an *undirected* graph (meaning that none of G's links carry directional information), G is completely defined by the N-by-N *adjacency* matrix, ℓ_{ij}:

$$\ell_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are linked;} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Using the graph metric function,

$$D_{ij} = \underset{\text{Paths, } P_{ij}}{\text{Minimum}} \left[\#\text{links}, l_{rs} \,|\, \{r, s\} \in P_{ij}\right], \tag{2}$$

we can write a general r-neighborhood CA *value*-transition rule 'f' (which will from now on refer generically to as a σ-rule) in the form

$$\sigma_i^{t+1} = f\left[\{\sigma_j^t\} | j \in S_r^G(i)\right], \tag{3}$$

where $S_r^G(i) = \{j | D_{ij} \le r\}$ is the radius-r *graph sphere* about the site i. In words, the value of σ_i^{t+1} is some function, f, of the values σ_j^t in radius r graph sphere around the site i. With this distance measure, G becomes a *discrete metric space*. If G is a one-dimensional line, and $r = 1$, then $S_r^G(i) = \{i-1, i, i+1\}$; i. e., it is equal to the conventional three site local neighborhood of elementary CA.

We now formally extend a conventional CA's dynamic arena – limited to the values $\sigma_i^t \in \{0, 1, \ldots, k-1\}$, $i = 1, \ldots, N$ – to one that includes the components of the underlying lattice's adjacency matrix:

$$\begin{cases} \sigma^{t+1} = F_\sigma[\{\sigma^t\}, \{\ell^t\}] \\ \ell^{t+1} = F_\ell[\{\sigma^t\}, \{\ell^t\}] \end{cases}, \tag{4}$$

where F_σ and F_ℓ are some functions (to be defined explicitly below) that explicitly couple the changing *value* states and *geometries*. The complete system at time t is specified by the state-vector

$$|G\rangle_t = |\sigma_1^t, \ldots, \sigma_N^t;\ \{\ell_{ij}^t\}\rangle\,. \tag{5}$$

The time-evolution of $|G\rangle$ proceeds according to the following transition rules: (*i*) σ-rules of the general form given above and familiar from *CA* simulations and (*ii*) ℓ-rules, which are divided into site *couplers*, linking previously unconnected vertices and site *decouplers*, which disconnect linked points. Because the topology can be altered only by either a deletion of *existing links* or an addition of links between pairs of vertices 'i' and 'j' with $D_{ij} = 2$, the dynamics is *strictly local*.

To be more precise, we first restrict the general σ-rule F_1 to (maximally symmetric) *totalistic* (**T**) and *outer-totalistic* (**OT**) type. Since the underlying lattice is a fully dynamic object, $|G\rangle$ will, in general, tend towards having a complex local geometry with an unspecified local *directionality*. The most general rules which can therefore be applied are those which are completely invariant under all rotation and reflection symmetry transformations on local neighborhoods. **T** (**OT**) σ-rules are then specified by listing particular *sums* $\{\alpha\}$ (*outer-sums* $\{\alpha_0\}, \{\alpha_1\}$ corresponding to center site values '0' and '1' respectively) for which the value of the center site becomes '1'. Formally,

$$\sigma_i^{t+1} = \phi_{\{\alpha\}}\left(\sum_j \ell_{ij}^t \sigma_j^t, \sigma_i^t\right), \tag{6}$$

where

$$\phi_{\{\alpha\}}(x, a) = \begin{cases} \sum_\alpha \delta(x + a, \alpha) & \longleftrightarrow \mathbf{T} \\ a \sum_{\alpha_1} \delta(x, \alpha_1) + (1 - a) \sum_{\alpha_0} \delta(x, \alpha_0) & \longleftrightarrow \mathbf{OT}\,, \end{cases} \tag{7}$$

and $\delta(x, y)$ is the *Kronecker* delta. Note that $\sum_j \ell_{ij}^t \sigma_j^t$ sums the values of all sites 'j' linked to 'i' at time 't'. The action on the state $|G\rangle$ is represented by

$$\widehat{\phi}_{\{\alpha\}}^i |\sigma\rangle_t = \left|\sigma_1^t, \ldots, \sigma_i^{t+1} = \phi_{\{\alpha\}}\left(\sum \ell_{ij}^t \sigma_j^t, \sigma_i^t\right), \ldots, \sigma_N^t\right\rangle, \tag{8}$$

where we distinguish the *operator* $\widehat{\phi}^i$ acting on the global value state from the actual local transition *function* ϕ which transforms each site value.

Link Rules

Local geometry altering rules are constructed by direct analogy: for any two selected sites i and j we restrict attention to site values of vertices contained within a 1-*sphere* of either site; that is, to all $k \in S_1(i,j) = S_1(i) \cup S_1(j)$. Link operators, whose action on the state is represented by:

$$\begin{aligned} \textit{decouplers:}\quad & \widehat{\psi}^{ij}_{\{\beta\}} \left|\ell^t_{ij}\right\rangle = \left|\ell^t_{11},\ldots,\ell^{t+1}_{ij} = \psi^{ij},\ldots,\ell^t_{NN}\right\rangle \\ \textit{couplers:}\quad & \widehat{\omega}^{ij}_{\{\varepsilon\}} \left|\ell^t_{ij}\right\rangle = \left|\ell^t_{11},\ldots,\ell^{t+1}_{ij} = \omega^{ij},\ldots,\ell^t_{NN}\right\rangle , \end{aligned} \tag{9}$$

either *link* or *unlink* two sites 'i' and 'j' depending on whether the actual sum of values in $S_1(i,j)$ matches any of those given in the $\{\beta\}$ or $\{\varepsilon\}$ lists, which completely define *decouplers* and *couplers*, respectively.

In order to construct classes of rules analogous to the two types of σ-rules defined above, we partition the local neighborhood into 3 disjoint sets (see Fig. 1): $S_1(i,j) = V_{ij} \cup A_{ij} \cup B_{ij}$, where

$$\begin{cases} \vee_{ij} = \{i,j\}, \\ A_{ij} = \{k | k \in C_1(i) \cap C_1(j)\}, \text{ where } C_1(i) = S_1(i) - \{i\}, \\ B_{ij} = S_1(i) \cup S_1(j) - \vee_{ij} - A_{ij} . \end{cases} \tag{10}$$

The action of link operators is then conveniently expressed as a function of the sums within the individual partitions. Defining $\nu_{ij} = \sigma_i + \sigma_j$, $a_{ij} = \sum_{k \in A_{ij}} \sigma_k$, and $b_{ij} = \sum_{k \in B_{ij}} \sigma_k$, we get *decouplers*, $\psi^{ij}_{\{\beta\}} = \psi^{ij}_{\{\beta\}}(\nu_{ij}, a_{ij}, b_{ij})$, where

$$\psi^{ij}_{\{\beta\}}(x,y,z) = \begin{cases} \{1 - \sum_k \delta(x+y+z, \beta_k)\}\ell_{ij} & \leftrightarrow \mathbf{T} \\ \{1 - \sum_k \delta(x, \beta_{1,k})\delta(y+z, \beta_{2,k})\}\ell_{ij} & \leftrightarrow \mathbf{OT} \\ \{1 - \sum_k \delta(x, \beta_{1,k})\delta(y, \beta_{2,k})\delta(z, \beta_{3,k})\}\ell_{ij} & \leftrightarrow \mathbf{RT} , \end{cases} \tag{11}$$

and *couplers*, $\omega^{ij}_{\{\varepsilon\}} = \omega^{ij}_{\{\varepsilon\}}(\nu_{ij}, a_{ij}, b_{ij})$, where

$$\omega^{ij}_{\{\varepsilon\}}(x,y,z) = \begin{cases} \delta(D_{ij},2) \sum_k \delta(x+y+z, \varepsilon_k) & \leftrightarrow \mathbf{T} \\ \delta(D_{ij},2) \sum_k \delta(x, \varepsilon_{1,k})\delta(y+z, \varepsilon_{2,k}) & \leftrightarrow \mathbf{OT} \\ \delta(D_{ij},2) \sum_k \delta(x, \varepsilon_{1,k})\delta(y, \varepsilon_{2,k})\delta(z, \varepsilon_{3,k}) & \leftrightarrow \mathbf{RT} . \end{cases} \tag{12}$$

In the above expressions, **RT** stands for *restricted totalistic* rules which maximally subdivide the local neighborhood. The inclusion of an ℓ_{ij} in the expressions for ψ assures that only those sites already *linked* can be decoupled and the $\delta(D_{ij},2)$ in the equations defining ω are put in to make sure that only sites separated by distance = 2 may be dynamically coupled.

The three type-specific *sums* appearing above are indexed with the following conventions:

- **T** rules are defined by the 'k' overall sums of values in $S_1(i,j)$ for which the particular action is to be taken. For example, define 'ψ' by *unlinking* 'i' and 'j' if the total sum = 1 (= β_1), 3 (= β_2) or 5 (= β_3). Equation (11) then states that $\ell^{n+1}_{ij} = 0$ if and only if $\ell^n_{ij} = 1$ and $\nu^n_{ij} + a^n_{ij} + b^n_{ij} \in \{1,3,5\}$.
- **OT** rules are specified by giving 'k' 2-tuples $(\beta_{1,k}, \beta_{2,k})$, and $(\varepsilon_{1,k}, \varepsilon_{2,k})$, where $\{1,k\}$ labels the sum '$\sigma_i + \sigma_j$' and $\{2,k\}$ labels the corresponding *outer sum* $= \sum_{s \in S_1(i,j) - \{i,j\}} \sigma_s$. For example, *link* '$i$' and '$j$' if $\sigma_i + \sigma_j = 0$ and outer sum = $\{3,4\}$, so that 'ω' is defined by listing the two 2-tuples $(\varepsilon_{1,1}0, \varepsilon_{2,1} = 3)$ and $(\varepsilon_{1,2} = 0, \varepsilon_{2,2} = 4)$.
- **RT** rules are completely specified by giving the 'k' 3-tuples of values $(x\sigma_i + \sigma_j, y = \text{sum in } A, z = \text{sum in } B)$, for which the link operation between 'i' and 'j' is to be performed. For example, define 'ψ' by *unlinking* 'i' and 'j' for the following values of partitioned sums: $(0,0,1)$, $(0,0,2)$, $(0,1,1)$, $(1,1,1)$; we then have that $(\beta_{1,1} = 0, \beta_{2,1} = 0, \beta_{3,1} = 1)$, $(\beta_{1,2} = 0, \beta_{2,2} =$

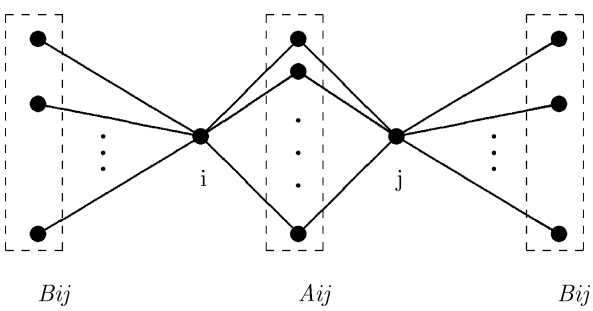

Structurally Dynamic Cellular Automata, Figure 1
Neighborhood partitioning. In the same way as *outer* sites can be considered separately for σ-transitions, we may, for topology transitions, distinguish between those sites belonging to *both i* and *j* ($\in A_{ij}$) and those belonging to *one* of the two sites but not both ($\in B_{ij}$). In this way we obtain the analogous *totalistic* (T), *outer-totalistic* (OT), and an additional type called *restricted totalistic* (RT)

$0, \beta_{3,2} = 2)$, $(\beta_{1,3} = 0, \beta_{2,3} = 1, \beta_{3,3} = 1)$, and $(\beta_{1,4} = 0, \beta_{2,4} = 1, \beta_{3,4} = 1)$.

Global transition operators are obtained by applying individual σ- and ℓ- operators to all *sites* and *site-pairs* in the graph G:

$$\begin{cases} \widehat{\Phi}_{\{\alpha\}} \, |\sigma\rangle = \prod_i \widehat{\phi}^{i}_{\{\alpha\}} \, |\sigma\rangle \ , \\ \widehat{\Psi}_{\{\beta\}} \, |\ell\rangle = \prod_{n\langle ij\rangle} \widehat{\psi}^{ij}_{\{\beta\}} \, |\ell\rangle \ , \\ \widehat{\Omega}_{\{\varepsilon\}} \, |\ell\rangle = \prod_{nn\langle ij\rangle} \widehat{\omega}^{ij}_{\{\varepsilon\}} \, |\ell\rangle \ , \end{cases} \tag{13}$$

where the products for $\widehat{\Psi}$ and $\widehat{\Omega}$ need to be taken only over *nearest* and *next nearest* pairs respectively. Given the full *value-topology* transition rule Γ, defined by

$$|G\rangle_{t+1} = (\widehat{\Omega}\widehat{\Psi}\widehat{\Phi})|G\rangle_t = \Gamma|G\rangle_t \ , \tag{14}$$

the fundamental problem is to understand the generic behavior of accessible graphs-G emerging from all possible initial structures and value configurations. We emphasize that the lattice *fully* participates in the dynamics and that, in general, no embedding is implied – it is the abstract *connectivity* itself whose evolution we are attempting to trace.

An Example

The application of the rather cumbersome expressions defining transition rules is in practice extremely straightforward, as we demonstrate with the following example: Consider a graph G defined as a (5×5) lattice with some distribution of values $\sigma = 1$ at time '$t = 1$' (see Fig. 2). We are interested in one *global* update of the system $|G\rangle_{t=1} \xrightarrow{\Gamma} |G\rangle_{t=2}$ with rules specified by

$$\begin{array}{ll} \text{(value)} & \\ \text{(topology)} & \end{array} \begin{cases} \Phi_{\{\alpha\}} : \{\alpha\}_{\mathrm{T}} = \{\alpha_1 = 1, \alpha_2 = 3, \alpha_3 = 5\} \ , \\ \Psi_{\{\beta\}} : \{\beta\}_{\mathrm{OT}} = \left\{ \begin{array}{l} (\beta_{1,1} = 1, \beta_{2,1} = 3) \\ (\beta_{1,2} = 1, \beta_{2,2} = 4) \end{array} \right\} , \\ \Omega_{\{\varepsilon\}} : \{\varepsilon\}_{\mathrm{OT}} = \{\varepsilon_{1,1} = 1, \varepsilon_{2,1} = 3\} \ . \end{cases} \tag{15}$$

We evolve the system by systematically sweeping through all *sites*, *linked* pairs, and *next-nearest neighbors*:

1. *All Sites:* … setting $\sigma_i = 1$ only at those 'i' for which the sum of the values at 'i' and its neighbors is equal to '2' at $t = 1$. By "neighbors" of any point 'i' we will always mean the set of vertices linked to 'i': (a, b), (h, m) and (x, y), for example, are all neighbors at $t = 1$. Writing out a few value-changing terms explicitly, we find that

$$\begin{aligned} \sigma_c^{t=2} &= \phi\left(\sigma_b^{t=1} + \sigma_c^{t=1} + \sigma_d^{t=1} + \sigma_h^{t=1}\right) \\ &= \phi(3) = 1, \quad \text{and} \\ \sigma_b^{t=2} &= \phi\left(\sigma_a^{t=1} + \sigma_b^{t=1} + \sigma_c^{t=1} + \sigma_g^{t=1}\right) \\ &= \phi(2) = 0. \end{aligned} \tag{16}$$

2. *All linked pairs of sites 'i' and 'j'*: … removing those links only if the 2-tuple $(\alpha, \beta) \in \{(1, 3), (1, 4)$, where $\alpha = \sigma_i + \sigma_j$ and 'β' is the sum of values of the neighbors of 'i' and 'j' at $t = 1$. For the points 'c' and 'h', for example, we have $(\alpha, \beta) = (1, 3)$, so that the link ℓ_{ch} is no longer present in $|G\rangle_{t=2}$:

$$\begin{aligned} \ell_{ch}^{t=2} &= \psi\left(\sigma_c^{t=1} + \sigma_h^{t=1}, \sigma_b^{t=1} + \sigma_d^{t=1} + \sigma_g^{t=1} \right. \\ &\qquad \left. + \sigma_i^{t=1} + \sigma_m^{t=1}\right) \ell_{ch}^{t=1} \\ &= \psi(1, 3)(1) = 0 \ . \end{aligned} \tag{17}$$

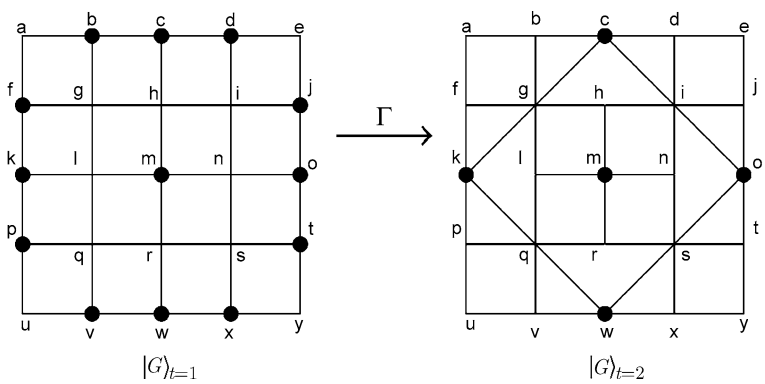

Structurally Dynamic Cellular Automata, Figure 2

Sample dynamic update of a (5×5) lattice from $t = 1$ to $t = 2$, obeying a T-type σ-rule with $\sigma \to \sigma'$ for local sums = 1, 3, 5 (i. e. $\alpha \in \{1, 3, 5\}$), and OT-type ℓ-rules: (i) *link* for $\{\varepsilon_{1,1} = 1, \varepsilon_{2,1} = 3\}$ and (ii) *unlink* for $\{\beta_{1,1} = 1, \beta_{2,1} = 3\}$ and $\{\beta_{1,2} = 1, \beta_{2,2} = 4\}$. *Solid* sites indicate that $\sigma = 1$

3. *All next-nearest neighbors 'i' and 'j'*: … linking them only if the 2-tuple $(\alpha, \beta) = \{(1, 3)\}$. By "next-nearest neighbor" we mean those pairs which are themselves unlinked but which share at least one other linked neighbor: (a, g), (h, r) and (w, y), for example, are all next-nearest neighbors at $t = 1$. For 'c' and 'g' we find

$$\begin{aligned}\ell_{cg}^{t=2} &= \omega\left(\sigma_c^{t=1} + \sigma_g^{t=1}, \sigma_b^{t=1} + \sigma_d^{t=1} + \sigma_f^{t=1} \right.\\ &\qquad \left. + \sigma_h^{t=1} + \sigma_l^{t=1}\right) \delta(D_{cg}, 2)\\ &= \omega(1,3)(1) = 1\,. \end{aligned} \tag{18}$$

Notice that although $\ell_{dn}^{t=1} = 0 \to \ell_{dn}^{t=2} = 1$, it is hidden by overlap with the remaining links $\ell_{di}^{t=2} = 1$ and $\ell_{in}^{t=2} = 1$. For this reason, not all link changes can always be observed directly in the following figures.

Other sites and links are updated in precisely the same manner. Had the link-rules been of **T**-type, only one sum would have to be considered: the sum of the values of the points in question along with their neighbors' values. Had they been, instead, of **RT**-type, three sums would have to be considered: the sum of the values of the sites in question, the sum of the values of their common neighbors (neighborhood A in Fig. 1) and the sum of the values of the points that are neighbors of one of the considered points, but not of the other (neighborhood B in Fig. 1). The final state $|G\rangle_{t=2}$ emerges after the above process has been applied concurrently to all pairs, neighbors and next-nearest neighbors in $|G\rangle_{t=1}$.

Comments

We conclude this section by making a few important general comments:

Comment 1. As defined above, Γ consists of three operators acting simultaneously on the state $|G\rangle$. More generally, one may prescribe any of 10 possible time-orderings to the operators Ω, Ψ and Φ. That is, specify certain intermediate state dependencies, so that, for example $\Gamma_1|G\rangle \equiv (\Omega\Psi)(\Phi|G\rangle)$ would in general be expected to yield results different from, say, $\Gamma_2|G\rangle \equiv \Omega(\Phi(\Psi|G\rangle))$. While we will be solely concerned with the synchronous time ordering defined above, we do not expect the qualitative results to depend critically on this choice.

Comment 2. A given rule Γ is completely defined by the set of sums $\{\alpha\}$, $\{\beta\}$ and $\{\varepsilon\}$. Alternatively, we can conveniently summarize a chosen transition rule by its vector-code $\vec{C} = (c[\phi], c[\psi], c[\omega])_{a,b}$, where

$$\begin{aligned} c[\phi] &= \begin{cases} \sum_\alpha 2^\alpha & \leftrightarrow \mathbf{T} \\ \sum_{\alpha_0} 2^{2\alpha_0} + \sum_{\alpha_1} 2^{(2\alpha_1+1)} & \leftrightarrow \mathbf{OT} \end{cases}\\ c[\psi] &= \begin{cases} \sum_k 2^{\beta_k} & \leftrightarrow \mathbf{T} \\ \sum_k 2^{3\beta_{2,k}+\beta_{1,k}} & \leftrightarrow \mathbf{OT} \\ \sum_k 2^{3(\beta_{2,k}+a\beta_{3,k})+\beta_{1,k}} & \leftrightarrow \mathbf{RT} \end{cases}\\ c[\omega] &= \begin{cases} \sum_k 2^{\varepsilon_k} & \leftrightarrow \mathbf{T} \\ \sum_k 2^{3\varepsilon_{2,k}+\varepsilon_{1,k}} & \leftrightarrow \mathbf{OT} \\ \sum_k 2^{3(\varepsilon_{2,k}+b\varepsilon_{3,k})+\varepsilon_{1,k}} & \leftrightarrow \mathbf{RT} \end{cases} \end{aligned} \tag{19}$$

where $a = \max\{\beta_{2,k}\} + 1$, $b = \max\{\varepsilon_{2,k}\} + 1$, and must be specified only for **RT**-type topology rules. The Γ appearing in the above example, therefore, can be summarized by $c[\phi] = 42$, $c[\psi] = 2^{3(3)+1} = 1024$ and $c[\omega] = 2^{3(4)+1} + 2^{3(3)+1} = 9216$. Note that '$\Psi$' and '$\Omega$' are chosen always to be of the same type.

Comment 3. Computer simulations of these systems require that some measures be taken to prevent possible memory overflows, such as would happen in cases either of pure coupling, where links are continually added and none deleted, or in isolated regions of a graph where for a few sites more neighbors are added than are allowed by memory. We thus introduce *working* link transition rules

$$\tilde{\psi}^{ij} \equiv \begin{cases} \psi^{ij} & \longleftrightarrow d_i' \text{ or } d_j' > \delta \equiv d_{\min} \\ 1 & \longleftrightarrow \text{else}\,, \end{cases} \tag{20}$$

$$\tilde{\omega}^{ij} \equiv \begin{cases} \omega^{ij} & \longleftrightarrow d_i' \text{ or } d_j' < \Delta \equiv d_{\max} \\ 0 & \longleftrightarrow \text{else}\,, \end{cases} \tag{21}$$

where $d_i = \text{degree}(i)$ (i. e. number of neighbors of i). In words: make a sweep of the lattice, temporarily storing the candidates to *add* and *delete* for each point. If, for any point i, the updated degree is greater than δ then proceed with deleting the stored deletion-candidates, otherwise do not delete; similarly, provided that the updated degree is less than Δ proceed with addition. Thus, it is sufficient that *one* of two points allow a dynamic link change between them for that change to be enacted. In the following, the complete constrained dynamics will be quoted as $\vec{C}_{(a,b)}^{[\delta,\Delta]}$. If constraints play no role in the actual evolution of specific examples, they will be left out of the definition.

Comment 4. Because each dynamic update involves three separate types of processing, the number of possible rules is extraordinarily large (see Table 1). Unlike pure σ-transitions, however, the fraction of the total number which yield interesting behavior (i. e. neither immediately explosive, where the number of links increases with-

Structurally Dynamic Cellular Automata, Table 1
Numbers of possible rules for each of the three types of transition rules. $d =$ *maximum allowable degree* and $a =$ *maximum sum to be used from partition* A_{ij}. *Example:* for $d = 5$, we have $N_\phi = 4096$, $N_\psi = 2^{24} \sim 2 \times 10^7$ and $N_\omega = 2^{21} \sim 2 \times 10^6$. We thus have $N_T = N_\phi N_\psi N_\omega \sim 10^{17}$ possible type OT Γ s

Rule type	ϕ	ψ	ω
T	2^{d+1}	2^{2d}	2^{2d-1}
OT	2^{2d+2}	$2^{6(d-1)}$	$2^{3(2d-3)}$
RT	—	$2^{3(a+1)(2d-1)}$	$2^{3(a+1)(2d+1)}$

out bound, nor immediately degenerative, where an initial graph rapidly dwindles to a few isolated links) appears to be manageably smaller.

Comment 5. Although it is the intrinsic geometrical patterning whose generic behavioral properties we are trying to deduce, one may approach SDCA from an alternative point of view: maintain the emphasis on unraveling the value configurational behavior, and interpret the presence of $[\Psi, \Omega]$ as background operators inducing nonlocal spatial connectivities. Whereas the systems defined above are completely abstract entities, in that locality is strictly defined by the link structure, the alternative scheme would be to embed the discrete networks in some specified manifold, and to study the effects of dynamically allocated nonlocal communication channels.

Emerging Patterns and Behaviors

Consider patterns that emerge from simple value seeds starting from ordered two dimensional Euclidean lattices. A single non-zero site may represent a small local disturbance that then propagates outward, restructuring the lattice. With appropriately chosen Γs one can induce a rich spectrum of different time evolutions only slightly perturbed by very few concurrent link changes to ones in which the initial geometry becomes radically altered. (The graphical representation of evolving one dimensional systems, in which link additions must be shown as *arcs* to avoid overlap with existing links, is needlessly confusing and is not considered.)

Figure 3 shows the first five iterations of a system starting from a four neighbor lattice with a single non-zero site at its center, the link structure is given explicitly and the solid circles represent sites with $\sigma = 1$. Notice how the link additions *follow* the emerging corrugated boundary surface of the value configuration. Remember that link additions are more than passive markers indicating particular correlations between local value configurations and structure; their presence directly influences all subsequent value development in their immediate vicinity.

Figure 4 (in which site values are suppressed for clarity) shows the continued development of this system. Though boundary effects begin to appear by $t = 25$, the

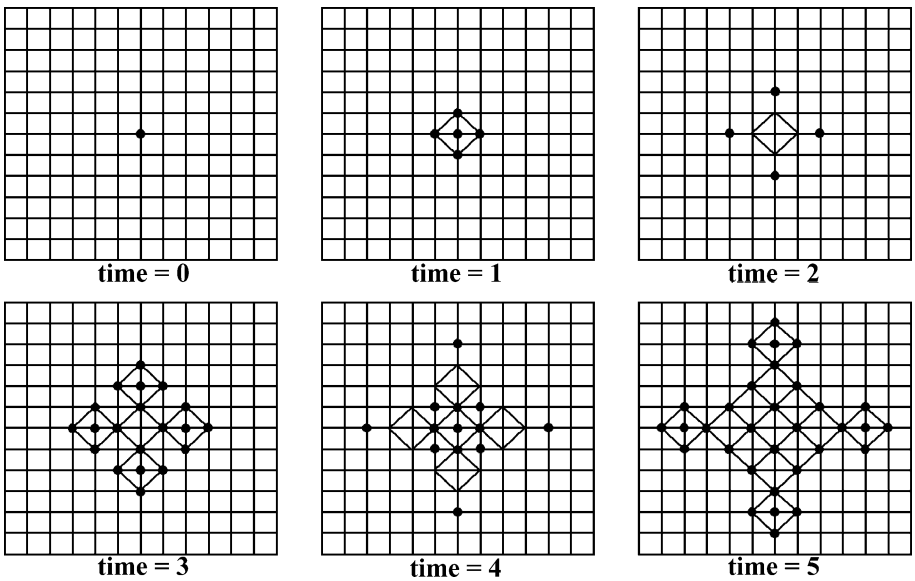

Structurally Dynamic Cellular Automata, Figure 3
First five iterations of an SDCA system starting from a 4-neighbor Euclidean lattice seeded with a single non-zero site at the center. The global transition rule Γ consists of T σ-rule and RT ℓ-rules: $\vec{C} = (26, 69648, 32904)_{[3,3]}$ (see text for rule definitions and code). Solid sites have $\sigma = 1$

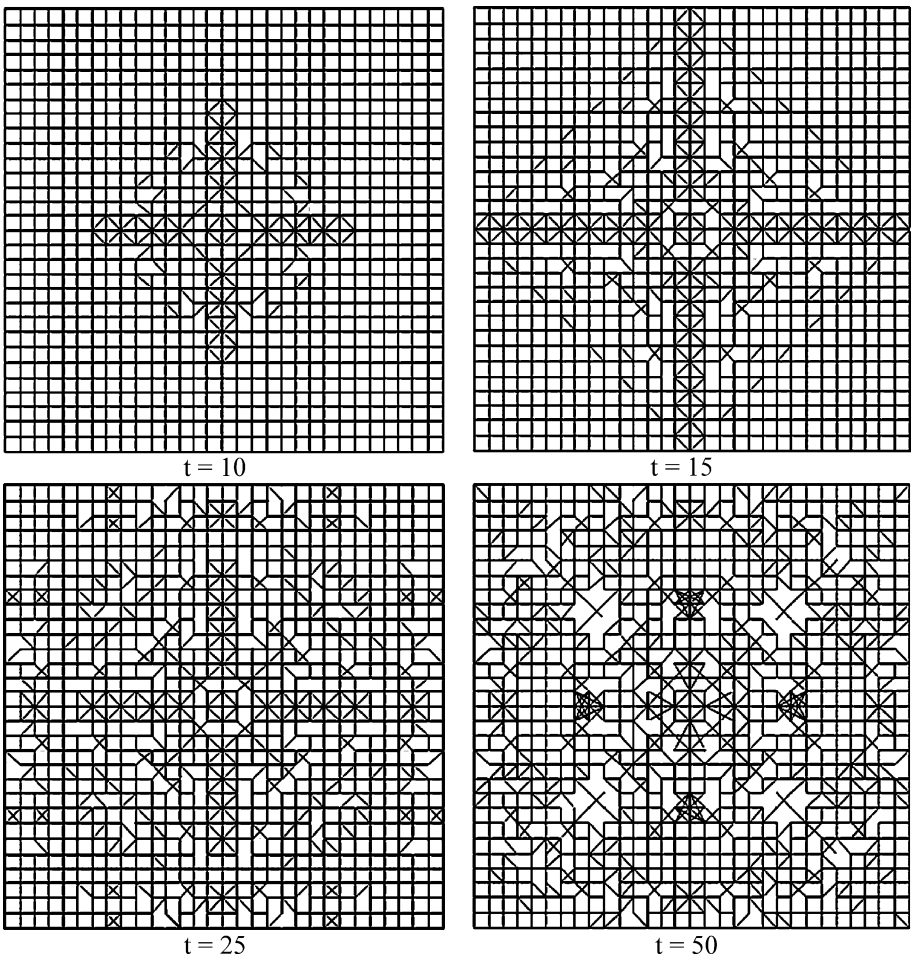

Structurally Dynamic Cellular Automata, Figure 4
Several further time frames in the structural evolution of the same system shown in the preceding figure. The values have been suppressed for clarity. The boundaries of the original lattice do not extend beyond the region shown so that the development is strictly confined to a 31 × 31 graph

characteristic manner in which this particular Γ restructures the initial graph is clear:

- There is a high degree of geometrical organization (the symmetry of the initial state is trivially preserved by the totally symmetric Γ).
- The lattice remains connected.
- The distribution of link changes made throughout the lattice remains fairly uniform (i. e. there is an approximate uniformity in the probability of appearance of particular local value states which induce a structural change.
- Link-lengths do not get arbitrarily large.

The last point implies that for a system embedded in the plane, communication channels remain approximately local. The global pattern emerges as a consequence of local ordering. On the other hand, Γs for which link-lengths get arbitrarily large are also easy to find.

Some other varieties of behavior are shown in Figs. 5 and 6. Figures 5a and b are representative of the class of ℓ-rules that only mildly perturb the underlying lattice (and for which σ states do not differ much from their conventional CA cousins). Other rules, of course, may have a stronger effect on the lattice, giving rise to associated σ states bearing little or no resemblance to their conventional CA counterparts.

Figure 5c shows an example of a link rule that *accelerates* the outward propagation of the value configuration. Compare the diameter of this pattern to that in the earlier figures, both shown at equal times. The outwardly oriented

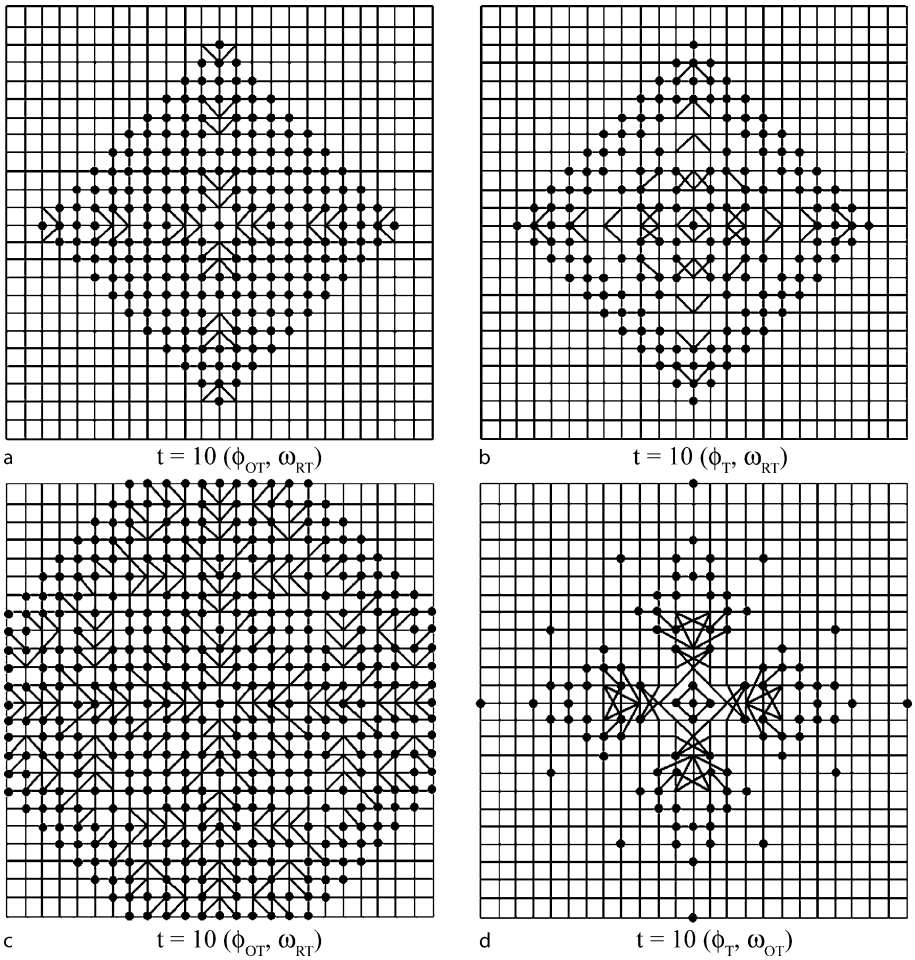

Structurally Dynamic Cellular Automata, Figure 5

Snapshot views of four typical developing states starting from a single non-zero site at the center of a 4-neighbor graph. Γs are as follows: a OT $c[\phi] = 1022$ and RT coupler $c[\omega] = 16$, $b = 1$; b T $c[\phi] = 22$ and RT coupler $c[\omega] = 32$, $b = 2$; c OT $c[\phi] = 1022$ and RT coupler $c[\omega] = 8$, $b = 1$; d T σ- and OT ℓ-rules $\vec{C} = (682, 19634061312, 133120)_{[2,8]}$

links that emerge from sites along the boundary surface become conduits by which non-zero values rapidly propagate. Had the underlying lattice topology been suppressed in this figure, and attention focused exclusively on the developing σ state, we could have interpreted the result as showing an effective increase in information propagation speed due to non-local connectivities (see comment 5 of the previous section).

Figure 5d, on the other hand, gives an example in which the link dynamics *lags behind* the σ development. The boundary proceeds outward essentially unaffected by changes in geometry, which are themselves confined to the interior parts of the lattice (at least at this early stage of this system's development).

Figure 6 shows snapshot views of a few system undergoing a slightly more complex evolution. Figure 6b, for example, shows a rule in which the outward σ propagation rapidly deletes most links from the original lattice but leaves a complex (though structurally stable) geometry at the origin of the initial disturbance. Figure 6c, on the other hand, shows a typical state of a system whose global connectivity becomes progressively more complicated.

A typical evolution starting from an initial state in which all sites are randomly assigned $\sigma = 1$ with probability $p = 1/2$ is shown in Fig. 7. Notice the rapid development of complex local connectivity patterns, the appearance of which points to a geometrical self-organization.

In general, structural behaviors emerging from random σ-states under typical Γs can be grouped into four basic classes (not to be confused with Wolfram's classification of elementary CA [74]):

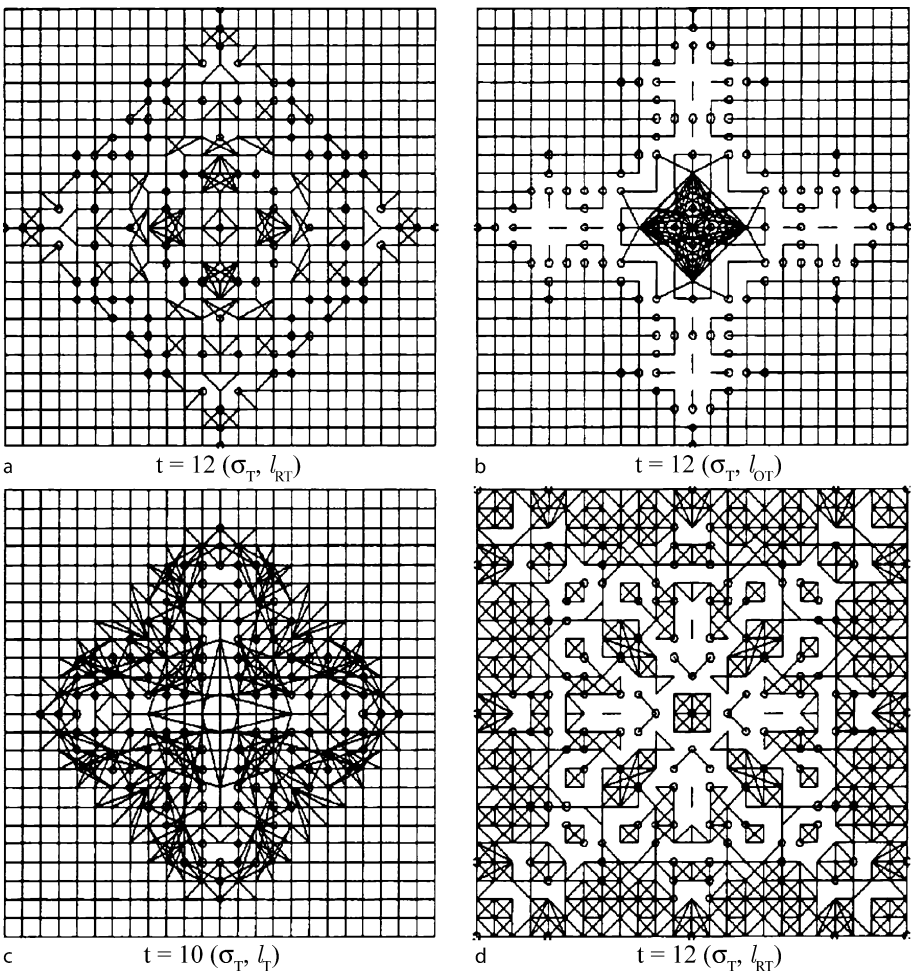

Structurally Dynamic Cellular Automata, Figure 6

Four more examples of states emerging from simple seeds. Figure a, b, c start from 4-neighbor graphs and d from an 8-neighbor graph (≡ 4-neighbor with diagonals). Γs are as follows: a T σ- and RT ℓ-rules $\vec{C} = (42, 69648, 32904)_{[3,3]}$; b T σ- and OT ℓ-rules $\vec{C} = (42, 589952, 8192)_{[2,8]}$; c T σ- and ℓ-rules $\vec{C} = (42, 128, 4)_{[0,10]}$; d T $c[\phi] = 682$ and RT ℓ-rules defined explicitly by $\Psi_{(104),(114),(124),(103),(113),(123)}$ and $\Omega_{(111),(215)}$

- *Class-1*, in which initial graphs decay into structurally much simpler final states: most links are destroyed, and graphs ℓ_{ij}^t, for sufficiently large t, consist essentially of a large number of small local subgraphs.
- *Class-2*, whose final states are characterized by periodic but globally connected geometries. SDCA typically arise in this class either because of a specific class-2 Φs remaining unchanged by the coupling to the lattice or class-3 Φs coupling with $\{\Psi, \Omega\}$ in such a way as to induce a lattice structure that supports a periodic state.
- *Class-3*, consisting of SDCA that tend to grow in size and complexity, at least as measured by two basic metrics: the *average degree*, $\langle \deg \rangle \equiv (1/N) \cdot \sum_i \left[|S_1(i)| - 1 \right]$, and *effective dimensionality*, $D_{\text{effec}} \equiv \langle N_{\text{nn}} \rangle / \langle \deg \rangle$, where $\langle N_{\text{nn}} \rangle$ is the average number of next-nearest neighbors. The values of both $\langle \deg \rangle$ and D_{effec} increase without bound for class-3 SDCA (unless an arbitrary upper constraint Δ is imposed on Γ).

Because the σ-density responds to the changing local neighborhood structure, it is possible that what at first appears to be an explosive growth in fact eventually leads to a more sedate, if not static, behavior at some larger $\langle \deg \rangle \gg \Delta$. Φs that yield $\langle \sigma \rangle_t \sim$ *constant* over a range of $\langle \deg \rangle$ (such as the *sum modulo*-2 rule; see below), when coupled with link rules that themselves become progressively less active with increasing $\langle \deg \rangle$, may induce evolutions leading to only mild

changes within specific ranges of the local structural parameters.

- *Class-4*, which is a provisional class (pending stronger evidence) that denotes a set of rules that yield open-ended σ- and ℓ changes, but during which the value of D_{effec} remains roughly constant. Ψs and Ωs belonging to this class effectively induce a *structural equilibrium*: despite the fact that large numbers of link changes continue to be made, so that the detailed structure of the evolving graph continually changes, the average ratio of the number of next-nearest to nearest neighbors stays approximately constant over long periods of time. While there is evidence to suggest this class is *real*, simulations have unfortunately been run for too short a time and on graphs containing too few sites to permit making any conclusive statements regarding the veracity of this class. Nonetheless, it is tempting to speculate that, for arbitrary values of D^*, there exists at least one set of SDCA rules for which $D_{\text{effec}} \approx D^*$ (within a desired $\epsilon > 0$) as the size of the graph $N \to \infty$. (*Pseudo* class-4 behavior, of course, can always be artificially induced either by imposing severe $[\delta, \Delta = \delta]$ constraints, or, as must typically be done for category-3 Γs, by deliberately impeding growth with some threshold Δ.)

Statistical Measures

As evidenced by Fig. 7, it is already nontrivial to meaningfully visualize the short-time evolution of (initially) *regu-*

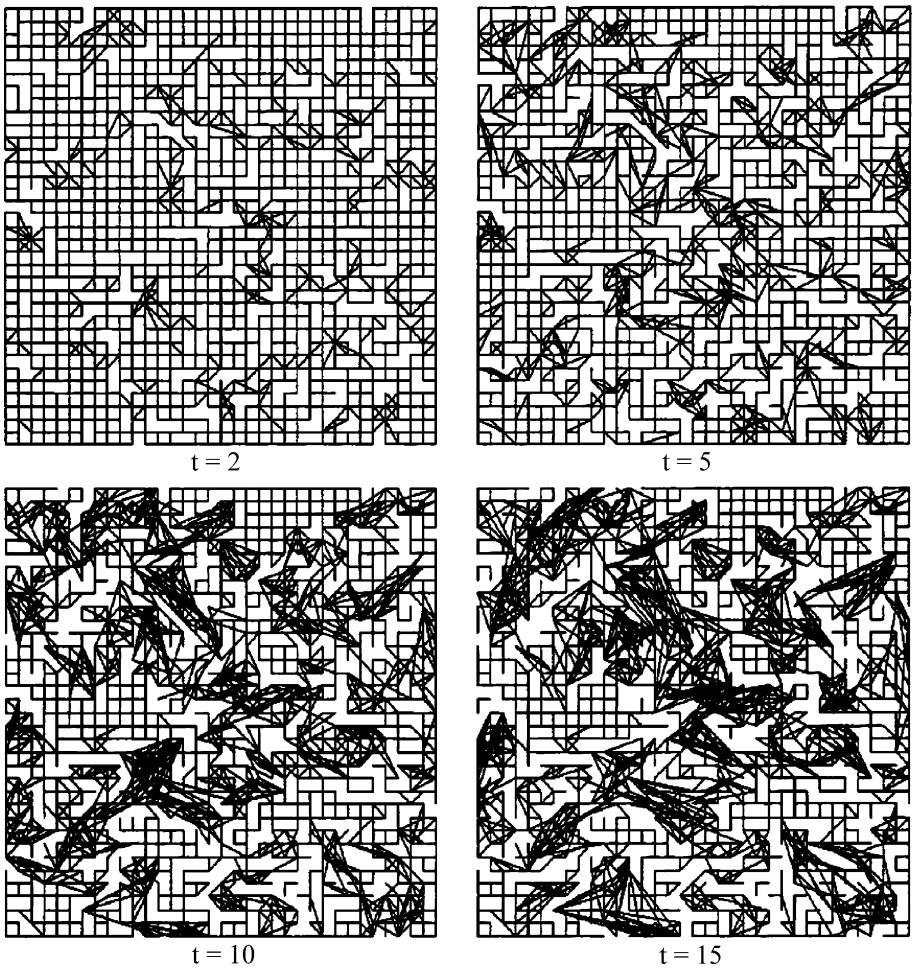

Structurally Dynamic Cellular Automata, Figure 7
Evolution of a 35 × 35 lattice, with randomly seeded sites. The development proceeds according to T σ- and OT ℓ-rules defined by code $\vec{C} = (84, 36864, 2048)$. The constraints are $[\delta = 0, \Delta = 10]$. The appearance of localized substructures is evidence of a geometrical self-organization

lar lattices that start with random initial value state. Visualizing the long-term dynamics of systems that start from a completely random state is even more difficult (although graph visualization algorithms may help). However, even in cases for which a direct visual inspection of the dynamics reveals little, one can always indirectly keep abreast of a given system's properties by monitoring its core structural and behavioral measures (a more detailed account is given in [31]).

Site value measures include the average density of sites with value $\sigma = 1$, $\langle\sigma\rangle_t \sim (1/N)\sum_{i=1}^{N}\sigma_i^t$; the local value correlation, $C^t \equiv \langle\sigma_i^t \cdot \sigma_j^t\rangle - (\rho^t)^2$, where $\langle\sigma_i^t \cdot \sigma_j^t\rangle$ is averaged over all pairs i and j with $\ell_{ij} = 1$; the fraction of sites whose value *changes* during one step of the evolution, $\Delta_t \equiv (1/N)\sum_{i=1}^{N}\{\sigma_i^{t-1} \oplus_2 \sigma_i^t\}$, where $\oplus_2$ is a sum *modulo*-2.

Geometry measures include the *average degree*, $\langle\text{deg}\rangle$; the *average number of next-nearest neighbors*, $\langle N_{\text{nn}}\rangle_t \equiv (1/N)\sum_i[|S_2(i)| - |S_1(i)| - 1]$; and D_{effec}. A measure of how the actual size of local neighborhoods changes with time may be obtained by embedding graphs into the two-dimensional plane and calculating the average pathlength at time t. Of course, global features that describe all complex networks – such as *connectivity*, *density*, *clustering*, and *path lengths* (2,6), are applicable to SDCA as well.

Link changes may be monitored by keeping track of (1) the total number of *link changes* (allowed under prescribed constraint conditions), $\Delta_t^{(l)} \equiv (1/2)\sum_{i=1}^{N}\cdot\sum_{j=1}^{N}\{l_{ij}^t \oplus_2 l_{ij}^{t-1}\}$; (2) the *constraint influence*, $f_l \equiv \Delta_t^{(l)}/N_t^{(l)}$, where $N_t^{(l)}$ is the total number of link changes that would have occurred in the absence of constraints ($f_l = 1$ indicates that the evolution is *pure*,

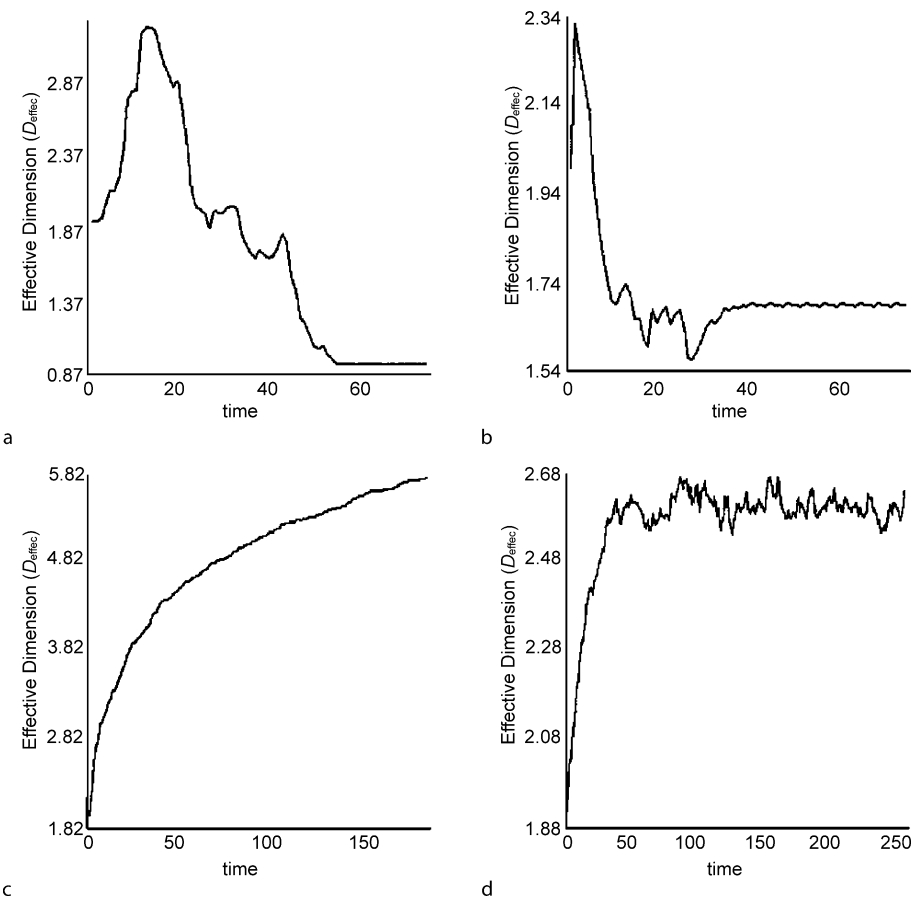

Structurally Dynamic Cellular Automata, Figure 8

Time development of the effective dimensionality D_{effec} for each of the four categories of behavior (see text): a T type Γ defined by $\vec{C} = (42, 128, 4)$; b T σ- and OT ℓ-rules $\vec{C} = (64, 9216, 1024)$; c T σ- and OT ℓ-rules $\vec{C} = (682, 512, 512)_{[0,10]}$; d T σ- and RT ℓ-rules defined explicitly by $\vec{\beta} \in \{(011), (110), (121), (233), (243)\}$ and $\vec{\epsilon} \in \{(120), (010), (021), (224)\}$

meaning it is unaffected by constraints; $f_l \sim$ *small* suggests that the imposed constraint window $[\delta, \Delta]$ has resulted in observed structures that are *impure*); (3) the link *creation-* and link *deletion-ratios*, $f_C \equiv N_C/\Delta_t^{(l)}$ and $f_D \equiv N_D/\Delta_t^{(l)}$, where N_C and N_D are the numbers of link created and destroyed, respectively; (4) the *activity levels*, $\gamma_C^t \equiv N_C/N_{\mathrm{nn}}^{t-1}$ and $\gamma_D^t \equiv N_D/N_{\mathrm{l}}^{t-1}$ (where N_l^{t-1} is the number of links at time $t-1$), which give the number of dynamic alterations relative to the corresponding spaces from which the candidates for alteration are selected; and (5) the *link evolution index*, $\gamma_L^n \equiv \left(1/N_l^{t=0}\right) \sum_i \sum_j \{l_{ij}^{t=0} \oplus_2 l_{ij}^n\}$, which gives the fraction of the initial lattice remaining after n iterations.

Figure 8 shows time series plots of D_{effec} for rules in each of the four behavioral classes defined above. The initial structure in each case is 35×35 4-neighbor Euclidean lattice, so that $D_{\mathrm{effec}}^{t=0} \sim 2$. Figure 8a gives an example of class-1 behavior, in which a short period of initial growth is followed by a decay into mostly disconnected clusters. The final state is characterized by $\langle \mathrm{deg} \rangle < 1$, and is stable. Figure 8b shows a system that starts from the same initial state as in Fig. 8a but whose Γ leads to a periodic geometry. Just the right number of links have been deleted to permit regions with isolated activity to emerge.

Figure 8c shows class-3 behavior in which D_{effec} steadily increases. The apparent leveling off seen toward the end of the run is due both to a decreased overall activity level and the increasingly effect of the $\Delta = 10$ constraint. The system in Fig. 8d exhibits class-4 behavior, characterized by a ongoing structural development within a relatively narrow interval of values of D_{effec}. Note that the structural changes here are essentially *pure*, and are not merely artifacts of any imposed constraints. Ilachinski (3) explores a wide range of emergent behaviors across all four classes, and examines the qualitative relationship between emergent behavior and initial σ- and ℓ-seeding.

Phase Plots

While it is of obvious interest to systematically explore every possible combination of β's and ϵ's that define ℓ-rules, Table 1 unfortunately suggests that the resulting rule space is simply too large. Nonetheless, we can learn much even by focusing our attention on a small subset of the complete rule space, keeping Φ, the initial σ-seeding, and all other factors constant. Specifically, consider the subset of all possible ℓ-rules that consists of **OT** link rules consisting of a single *coupler*, ω, and a single *decoupler*, ψ. Moreover, let $\phi \equiv \oplus_2$ (i. e., sum *modulo*-2 rule), demand that only pairs of $\sigma = 0$ sites be considered for a link change, and consider ℓ-rules belonging to the following set:

$$\begin{array}{lll} \textit{decouplers:} & \{\beta_{1,1} = 0, \beta_{2,1} = \mu\}, & 1 \le \mu \le 0, \\ \textit{couplers:} & \{\epsilon_{1,1} = 0, \epsilon_{2,1} = \nu\}, & 1 \le \nu \le 0. \end{array} \tag{22}$$

Figure 9 summarizes the behavior of a four neighbor, 25×25 lattice with periodic boundary conditions, starting from an initial σ-seed consisting of a single nonzero site. Four basic kinds of structural behaviors emerge:

1. *Static state:* this trivially occurs when the link rules are unable to take effect; namely, when $\mu \ge 7$ and $\nu \ge 8$.
2. *Rapid growth:* for an entire range of μ and ν, the average number of neighbors for each site of the lattice increases rapidly for 20–30 iterations.
   This number would likely continue to increase, were it not for the constraint conditions ($\equiv [0, 10]$). The "final state" is neither stable nor periodic. One sometimes also sees *delayed growth* in this class of behavior, in which case the link structure is initially relatively quiescent (and the behavior of the system as a whole mimics that of a conventional CA). As coupler rules are triggered by specific σ states, the average degree of the lattice rapidly increases (at least until the constraint conditions take effect).
3. *Spontaneous decay:* when decouplers are stronger than couplers, the average degree typically decreases. If this occurs too rapidly, the structure surrounding the single nonzero valued site may become isolated from other parts of the lattice. If a few non-zero values do not leak out into the outlying regions, link changes remain confined to the central subgraph, leading to either rapid stability or periodicity.
4. *Initial growth, followed by periodicity:* this is the least common behavior, and requires a delicate balance between coupler and decoupler rules.

It is interesting to compare these results with those obtained from a random σ seed. In this case, the sharp divisions between characteristic behaviors disappear, and there is a pronounced increase in the number of links for all μ and ν. However, the inclusion of an additional decoupler, may induce decay and periodicity. For example, consider the same initial lattice and Φ as used in Fig. 9, fix two **OT** ℓ-rules $\Psi_{(0,5)}$ and $\Omega_{(0,1)}$, and add the decoupler $\Psi_{(0,\mu)}$: $\{\beta_{1,1} = 0, \beta_{2,1} = \mu\}$, $1 \le \mu \le 9$. Surveying the emergent behaviors for this range of μ's, one now finds decaying lattices for $\mu \ge 2$. In each case, the initial graph succumbs to periodicity following a transient of between 50 and 100 iterations. The evolving lattice is also more prone to break up into small disconnected subgraphs.

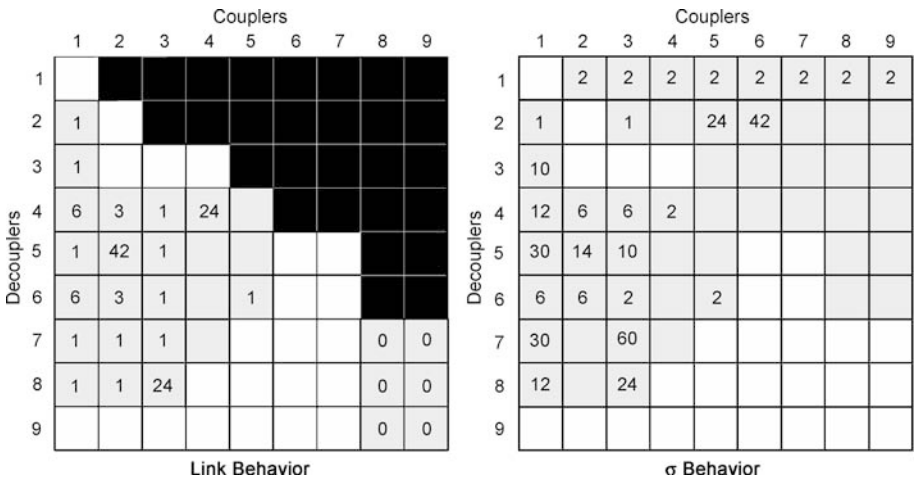

Structurally Dynamic Cellular Automata, Figure 9

Phase plot **that summarizes behavior of a four neighbor, 25 × 25 lattice with periodic boundary conditions, starting from an initial σ-seed consisting of a single nonzero site. Γ is defined by the sum *modulo*-2 σ-rule and ℓ-rules of the form: *decouplers*–$\{\beta_{1,1} = 0, \beta_{2,1} = \mu\}$, *couplers*–$\{\epsilon_{1,1} = 0, \epsilon_{2,1} = \nu\}$. *Grey areas* in both plots denote *periodic states*. *White areas* denote *growth* in the plot for link behavior, and a *nonperiodic* state for σ-behavior. The *black* area that appears in the link-bevavior plot denotes *decay*. Numbers that appear in *individual boxes* denote period lengths**

Although, just as in conventional CA, small changes to ℓ-rules can lead to large differences in emergent behavior, they generally appear to do so in a more predictable and patterned manner. Of course, particular classes of Γ may induce more complex phase plots; for example, isolated pockets of anomalous (and rapidly shifting) behavior may appear within larger surrounding regions undergoing otherwise mutually consistent and slowly changing dynamics. A better sense of the space of possible emergent behaviors, along with a deeper understanding of the relationship between Φ and ℓ-rules, awaits a future study.

SDCA as Models of Computation

The basic SDCA model, as outlined above (which we will denote as SDCA_0 to avoid possible confusion with the hierarchy of related SDCA models introduced in this section), was modified and generalized by Majercik [39] into a form more suitable for addressing its formal *computational capabilities* rather than as an exploratory toolkit for describing physical processes (which is the primary reason for which SDCA_0 were first conceived). Motivated primarily by finding models of human brain function (for which one intuitively expects nonlocal neural connections to play a fundamental role in the rewiring of neural tissue), Majercik shows that suitably generalized SDCA are not only capable of universal computation, but actually represent a more efficient class of computational models than conventional CA. Majercik also reports an SDCA that can solve the *firing squad* problem in $O(\log t)$ time (i. e., exponentially faster than the $O(t)$ in conventional CA), and a class of CA-*universal* SDCA models that can simulate any conventional CA with a speedup factor of two. (The *firing squad* problem [46] consists of finding a rule for which all sites in a CA evolve into a special state after the exactly the same number of steps.)

Majercik proceeds by first identifying five properties of SDCA_0 that, while reasonable from a physical modeling standpoint, make it difficult to rigorously formulate and prove theorems:

1. *Finiteness:* The requirement that SDCA_0 be strictly finite, both in time and space, is obviously necessary for computer experiments, but is unnecessarily restrictive for general theorem proving. Likewise, the assumption that the sets α, β and ϵ must be finite is questioned.
2. *Bidirectionality:* While SDCA_0 are defined with symmetric links, an obvious generalization that makes the basic model more readily applicable to neural dynamics (among other kinds of physical and biological systems) is to allow for unidirectional links.
3. *Link-rule Asymmetry:* While SDCA_0's link decoupler function (Eq. (11)) contains the factor ℓ_{ij} to explicitly prevent the system from inadvertently linking two unlinked sites, SDCA_0 does not include an analogous term for the coupler function (that is, a term to prevent an evolving system from inadvertently unlinking two linked sites).

4. *Inconsistency:* While σ-rules effectively ignore site positions, all three types of link rules assume that the various neighborhoods surroundings individual sites (A_{ij}, B_{ij}, and $C_{ij} \equiv \{k|D_{ik} = 1 \oplus D_{jk} = 1\}$, where '$\oplus$' denotes *exclusive or*) are all *recognized as such* by the dynamics. That is, the link rules effectively " know" the positions of a site's neighbors, while σ-rules possess no such information.
5. *Small Rule Set:* The class of σ- and ℓ-rules used by SDCA_0 may be generalized to include a far broader class of transition functions.

On the basis of these observations, Majercik [39] introduces a set of three core models to define a hierarchy of eight alternative SDCA computational systems, $\{\text{SDCA}^{(1)}, \text{SDCA}^{(2)}, \ldots, \text{SDCA}^{(8)}\}$. The three core models are (1) the *relative location* model ($= M_{\text{R}}$), (2) the *labeled links* model ($= M_{\text{L}}$), and (3) the *symmetric links* model ($= M_{\text{S}}$). They differ only in the degree to which their σ- and ℓ-transition functions depend on *specific* sites. For example, M_{R}'s transition functions depend on the state and exact relative position of each neighbor (and therefore "knows" the exact source of any state in a local neighborhood). In M_{L}, links are labeled and the transition functions know both neighbor states and the label of the links to given neighbors, but the exact neighbor locations remain unspecified. Finally, in M_{S}, it is assumed that no information about the source of the neighborhood states exists, and transition functions only know the number of neighbors in a particular state.

Each of the three core models may be defined in two versions: an *unbounded links* (abbreviated, UL) version, in which the number of neighbors a given site can have is unbounded, and a *bounded links* (abbreviated, BL) version, in which an explicit upper limit is imposed. In addition, there is also one *finite labels* version of M_{L}. Majercik imposes certain mild conditions on the local transition functions; for example, that local neighborhoods always remain strictly finite, σ-rules leave quiescent neighborhoods alone, and that links between sites with quiescent neighborhood remain unaltered.

***Relative Location* SDCA model** In the *Relative Location* model, the transition functions all have access to the exact relative *location* and *state* of each neighbor site. Define a *neighbor* of site i, $n_i \in S \times Z^d$, as a pair that specifies the state (by a single label) and relative location of the neighboring site (as a d-tuple of coordinates). Let $W = S \times Z^d$ be the set of all possible neighbors, and $\mathcal{F}_W$ (called the *neighborhood function*) be the set of all possible finite, nonempty, partial functions that map Z^d to W. The local state transition function $\sigma : \mathcal{F} \to W$ maps neighborhood functions to the state set of SDCA. The local link transition function $\lambda : \mathcal{F} \times \mathcal{F} \times \{0, 1, 2\} \to \{0, 1\}$ maps pairs of neighborhood functions (that define the neighborhoods of two sites, i and j and a number that specifies the status of the link between i and j: value *zero* meaning that i and j are neither direct neighbors nor next-nearest neighbors; value *one* meaning that i and j are immediate neighbors; and value *two* meaning i and j are next-nearest neighbors) to one of two link states: *zero*, meaning no link between i and j, and *one*, meaning a link exists.

***Labeled Links* SDCA model** The *Labeled Links* model removes from M_{R}'s transition functions any dependency on the exact relative location of a site's neighbors, but allows the links to still be *labeled* so that the transitions functions can distinguish one link from another. This ability to "label" links paves the way for us to define SDCA with unidirectional links, since the labels can be used to distinguish between the *input* and *output* links to a site. Consider, for example, the UL-version of M_{L}. Labeling the links by natural numbers, $\mathcal{N}$, we define a neighbor of site i, as a pair (q, n), where $q \in S$ labels the state of the neighboring site, and $n \in \mathcal{N}$ labels the link between i and its neighbor. Site i is defined as the direct neighbor linked via the 0th link, and the set of all possible neighbors, $W = S \times \mathcal{N}$. As for M_{R}, $\mathcal{F}_W$ is the set of neighborhood functions that map Z^d to W, the local state transition function $\sigma : \mathcal{F} \to S$ maps neighborhood functions to states in S, and the local link transition function $\lambda : \mathcal{F} \times \mathcal{F} \times \{0, 1, 2\} \to \{0, 1\}$ maps pairs of neighborhood functions and a number to either the values *zero* (unlinked) or *one* (linked).

***Symmetric Links* SDCA model** The *Symmetric Links* model imposes the strictest constraint of all by doing away with all means by which the local transition functions may distinguish different neighborhood orientations. Consider the unbounded link version of M_{S}. Assume the SDCA has a total of n states, and let $S = \{1, 2, \ldots, n\}$. Let $\vec{n}_i \in N^n$ be an n-dimensional vector such that $(n_i)_k$ is equal to the number of site i's neighbors in state k. Then the local state transition function $\sigma : N^n \to S$ maps vectors in N^n to states in S, and the local link transition function $\lambda : N^n \times \{0, 1, 2\} \to \{0, 1\}$ maps a vector in N^n and a link status label to either the values *zero* (unlinked) or *one* (linked). M_{S} can also be modified slightly to allow the local transition functions to retain knowledge of the state of site i: simply let $\sigma : S \times N^n \to S$ map the pair consisting of the *state* of site i and a vector that defines the distribution of states among i's immediate neighbors (excluding i). The local link function likewise assumes a simi-

lar form $\lambda\colon S^2 \times N^n \times \{0,1,2\} \to \{0,1\}$, where the first component of the 3-tuple input to *lambda* is a pair that defines the states of the two sites to which the link function is being applied.

SDCA as CA Simulators

What does it mean to say that one dynamical system *simulates* another? Heuristically, it means that, for certain initial states, one system behaves just like another [32,74]. Suppose we have two CA systems – CA and CA′ – defined by rules ϕ and ϕ', and initial states $\vec{\sigma} \in \Sigma$ and $\vec{\sigma}' \in \Sigma$, respectively. Then, loosely speaking, T iterations of CA are said to be "simulated" by nT ($n \geq 1$) iterations of CA′, provided there exists some invertible function, $f\colon \Sigma \to \Sigma$, by which $\vec{\sigma}'$ is replaced by $f(\vec{\sigma})$. Simulation is a transitive relationship: if system B simulates system A, and another system C simulates B, then C also simulates A.

For example, a single site with a particular value in CA may be simulated by a fixed *block of sites* in CA′. After n steps, the blocks in CA′ evolve to exactly the same final state as the single time-step evolution of individual sites in CA. As a concrete example, consider the elementary (one-dimensional, binary valued, conventional CA) rules ϕ_{18} and ϕ_{90}:

	111	110	101	100	011	010	001	000
	↓	↓	↓	↓	↓	↓	↓	↓
ϕ_{18}:	0	0	0	1	0	0	1	0
ϕ_{90}:	0	1	0	1	1	0	1	0

Provided that *two time steps* under ϕ_{18} are carried out for every time step of rule ϕ_{90}, it is easy to show that under the block transforms $0 \to f(0) = 00$ and $1 \to f(1) = 10$, the evolution of arbitrary starting configurations under ϕ_{90} is reproduced – or *simulated* – by ϕ_{18}. For example, the global state $\vec{\sigma} =$ '0011000' – which evolves into $\phi_{90}(\vec{\sigma}) =$ '0111100' under ϕ_{90} – yields the same state (after it is block-transformed) that results from two iterations of ϕ_{18} applied to ϕ_{90}'s block-transformed initial state, $f(\vec{\sigma}) =$ '00001010000000':

$$\phi_{18}[\phi_{18}[f(\vec{\sigma})]] = 00101010100000 = f[\phi_{90}(\vec{\sigma})]. \quad (23)$$

Now consider the specific case of SDCA simulating a conventional CA (we follow Majercik [39]). First, because SDCA cannot be expected to preserve the local topology of a simulated CA, it is necessary to define separate *encoding* ($= e$) and *decoding* ($= d$) functions – $e\colon \Sigma_{\mathrm{CA}} \to \Sigma_{\mathrm{SDCA}}$ transforms the initial configuration of the CA systems to configurations in the SDCA system being used to simulate it (where Σ_{CA} and Σ_{SDCA} are the configurations spaces of CA and SDCA, respectively); and $d\colon \Sigma_{\mathrm{SDCA}} \to \Sigma_{\mathrm{CA}}$ effectively performs the inverse transformation. Encoding (and decoding) functions are called *structurally defined* if they are recursive and use a finite amount of information to encode (or decode) a given configuration; and are otherwise expected to transform quiescent states to quiescent states. Majercik further assumes that (1) e has access to the rule table of the conventional CA system being simulated; (2) d does not have access to the rule tables of either system; and (3) e and d must together satisfy the relation: $e \times d = \mathrm{Identity}(\Sigma_{\mathrm{CA}})$.

Denoting the global transition functions of the CA and SDCA systems by Φ_{CA} and Φ_{SDCA}, respectively, Φ_{SDCA} is said to *simulate* Φ_{CA} if there exist $m \geq 1, n \geq 1$ and structurally-defined functions $e\colon \Sigma_{\mathrm{CA}} \to \Sigma_{\mathrm{SDCA}}$ and $d\colon \Sigma_{\mathrm{SDCA}} \to \Sigma_{\mathrm{CA}}$, such that for any configuration $\vec{\sigma} \in \Sigma_{CA}$ and any $k \geq 1$,

$$\Phi_{\mathrm{CA}}^{kn}(\vec{\sigma}) = d\left[\Phi_{\mathrm{SDCA}}^{km}\left[e\left(\vec{\sigma}\right)\right]\right]. \quad (24)$$

If $m > n$ then Φ_{SDCA} simulates Φ_{CA} with a *slowdown* factor of m/n. If $m < n$ then Φ_{SDCA} simulates Φ_{CA} with a *speedup* factor of n/m.

SDCA Hierarchy of Models

Majercik [39] uses the three generalized models introduced above ($M_{\mathrm{R}}, M_{\mathrm{L}}$, and M_{R}) to define a hierarchy of eight SDCA models of computation. At the top of his hierarchy (arranged from top-to-bottom in roughly, but not completely, decreasing order of computational strength; see discussion that follows) are the UL and BL versions of M_{R}: $\mathrm{SDCA}^{(8)}$ and $\mathrm{SDCA}^{(7)}$, respectively; followed by $\mathrm{SDCA}^{(6)} =$ UL version of M_{L}; $\mathrm{SDCA}^{(5)} =$ BL version of M_{L}; $\mathrm{SDCA}^{(4)} =$ a *finite labels* version of M_{L}; $\mathrm{SDCA}^{(3)} =$ UL version of M_{S}; $\mathrm{SDCA}^{(2)} =$ BL version of M_{S}; and, sitting on the lowest level (computationally speaking), is $\mathrm{SDCA}^{(1)} = \mathrm{SDCA}_0$.

A little thought suffices to establish certain relationships among the various classes. Give two classes, C_1 and C_2, let $C_1 \leq_{\mathrm{s}} C_2$ denote the fact that if, given any SDCA $S_1 \in C_1$ there exists an SDCA $S_2 \in C_2$ that simulates S_1. Then, for example, since any BL SDCA can be simulated by an unbounded links version of the same system, and a finite links version of M_{L} can be simulated by a bounded links version, we know immediately that $\mathrm{SDCA}^{(7)} \leq_{\mathrm{s}} \mathrm{SDCA}^{(8)}$, $\mathrm{SDCA}^{(4)} \leq_{\mathrm{s}} \mathrm{SDCA}^{(5)} \leq_{\mathrm{s}} \mathrm{SDCA}^{(6)}$, and $\mathrm{SDCA}^{(2)} \leq_{\mathrm{s}} \mathrm{SDCA}^{(3)}$. Similar reasoning [39] leads to the general relationship:

$$\begin{cases} \mathrm{SDCA}^{(3)} \leq_{\mathrm{s}} \mathrm{SDCA}^{(8)} \leq_{\mathrm{s}} \mathrm{SDCA}^{(6)}, & \text{and} \\ \mathrm{SDCA}^{(2)} \leq_{\mathrm{s}} \mathrm{SDCA}^{(7)} \leq_{\mathrm{s}} \mathrm{SDCA}^{(5)}. \end{cases} \quad (25)$$

Finally, since the unbounded links version of M_R has all the information necessary to construct the neighborhood partitions used by $SDCA_0$, and since $SDCA^{(8)} \leq_s SDCA^{(6)}$, we see that $SDCA_0 \leq_s SDCA^{(6)}$ and $SDCA_0 \leq_s SDCA^{(8)}$.

Majercik's two main results, which we state without proof, are:

> *Majercik Theorem 1:* Given an arbitrary 1-dimensional conventional CA with radius $r = 1$, there exists an unbounded links version of M_R ($= SDCA^{(8)}$ of the SDCA hierarchy) that can simulate it with a speedup factor of two.

> *Majercik Theorem 2:* There exists a 1-dimensional finite links version of M_L ($= SDCA^{(4)}$) that can simulate an arbitrary k-state 1-dimensional conventional CA with radius $r = 1$ with a slowdown factor $O\left(k^{2r}\sqrt{2r\log k}\right)$.

Detailed proofs of these two theorems appear in [39] (where they are called *Theorems 4.4* and *4.5*, respectively). In chapter 5 of his thesis [39], Majercik presents an explicit construction of a CA-universal $SDCA^{(4)}$ computational model, and compares it to Albert and Culik's [3] construction of a 1-dimensional CA-universal conventional CA that simulates any 1-dimensional, k-state, radius r CA with an $O\left(k^{8r}\right)$ slowdown. Although Majercik's CA-universal SDCA uses more states than Albert and Culik's universal CA, it is also markedly faster.

The reason why the SDCA is faster is at least intuitively clear. An SDCA's dynamic links effectively endow an otherwise conventional CA with a random access memory. Since SDCA can establish links between any two sites a distance d apart in $O(\log d)$ time, *any* site potentially has access to the state of any other site. While it may be argued that sites in conventional CA can also access the states of other cells, they *cannot do so permanently*. Once information is accessed once and used, the connection is lost, and must subsequently be re-established. Moreover, the links in SDCA can potentially connect sites that are arbitrarily far apart; so that, once a small number of links are dynamically created, they continue to provide long-range communication channels throughout the network. Since the propagation of information in a conventional CA is necessarily limited in being able to flow one site at a time, the overall computational speed is obviously limited.

However, it is worth pointing out that while the computational strength of Majercik's CA-universal SDCA model undoubtedly derives from its ability to forge long-range communication links, the results as quoted from [39] do not tap into what is potentially SDCA's greatest strength; namely, the ability to *adaptively create links, even as a given computation unfolds*. In Majercik's model, the links are dynamically coupled to an actual computation only insofar as they are initially fixed as a function of the initial state. While the local structure certainly *evolves* (as it does in all SDCA systems, as the computation itself unfolds), it does so purely as a consequence of the SDCA rules, and not adaptively to the evolution.

Majercik concludes his thesis by speculating on how an *adaptive* variant of his CA-universal SDCA may be used to explore certain aspects of evolutionary learning. (Working from a different set of assumptions, Halpern [23,24] applies evolutionary programming techniques to $SDCA_0$ to explore what happens when the structure is allowed to play an *explicit* dynamic role in the computation; see next section.) The question of whether there exist SDCA-universal SDCA models – that are able to simulate certain classes of the SDCA hierarchy, for example – remains open.

SDCA & Genetic Algorithms

Genetic algorithms (abbreviated, GA) are a class of heuristic search algorithms and computational models of adaptation and evolution based on natural selection. In nature, the search for beneficial adaptations to a continually changing environment (i. e., evolution) is fostered by the cumulative evolutionary knowledge that each species possesses of its forebears. This knowledge, which is encoded in the chromosomes of each member of a species, is passed on from one generation to the next by a mating process in which the chromosomes of "parents" produce "offspring" chromosomes. A comprehensive review of GA is given by Mitchell [45].

While GAs may be effectively used to search for "interesting" topological structures (but for which the structures themselves do not play any dynamic role; see, for example, Lehmann [37]), Halpern [23] is the first to explore a novel hybrid algorithm between GA and SDCA, in which SDCA rules are used to *evolve a GA*. Weinert et al. [72] explore a related "structurally dynamic" GA model, in which links between adjacent individuals of a population are dynamically chosen according to deterministic or probabilistic rules. In this section, we follow Halpern [23,24].

Formally, GAs are defined by (1) an ensemble of "candidate solution" vectors, $\{\vec{s}_i\}: \vec{s}_i \in M_{\mathcal{P}} \subseteq R^n$, where M is the set of all possible solutions to a given "problem" $\mathcal{P}$ (the $\vec{s}_i$ are usually, but not always, defined as a string of binary numbers [45]), and (2) a "fitness function", $f(\vec{s})$, that represents how well a given $\vec{s}$ "solves" $\mathcal{P}$. The goal of the GA is to find the global optimal solution,

$\vec{s}^*$ such that (from the point of view of *maximizing* fitness, $f(\vec{s}) \leq f(\vec{s}^*) \equiv f^*, \forall \vec{s} \in M$. Optimization proceeds through the combined processes of *selection*, *breeding*, *mutation*, and *crossover replacement* [45]; to which – in the hybrid SDCA↔GA algorithm, Halpern adds the new feature of *self-selective neighborhood structure*.

It should be immediately noted that this is not an ad-hoc addition. Muhlenbein [47] points out that if each generation of a GA searches over the entire possible solution space, the algorithm may – depending on the fitness function – converge prematurely to a sub-optimal solution. To reduce the likelihood of this happening, Muhlenbein introduces a spatial population structure; restricting fitness and mating to neighborhoods called *demes*. Demes are geographically separate subpopulations in which candidate solutions evolve along disparate trajectories; though occasional mixing still occurs through the process of migration.

In Halpern's variant [23], an otherwise conventional GA is placed within the structure of SDCA_0 (i. e., the basic model defined by Eqs. (11)–(14)). Heuristically, this allows each candidate solution to "choose a community" with which to mate, during each generation. The choice of neighborhoods thus becomes an integral component of the GA, and is determined dynamically by the evolving solutions.

Halpern's algorithm proceeds as follows [24]: *(Step 1)* an initially random lattice (defined by adjacency matrix $l_{ij}^{(t=0)}$) is seeded with single-chromosome candidate solutions of fixed length, one per site; *(Step 2)* a fitness function, $f_i = \sum_{i=1}^{N} \delta_{ij}$, is defined to assign a numerical measure of "optimality" to each site (N is the number of sites, and δ_{ij} is the value – equal to 0 or 1 – of the jth gene of the ith chromosome; *(Step 3)* each site i ranks each of its nearest and *next*-nearest neighbors according to f_i; *(Step 4)* each site disconnects with a fraction, f_D, of its least-fit neighbors, and connects with a fraction, f_C, of its fittest next-nearest neighbors; *(Step 5)* each site randomly mates with one of its nearest neighbors (i. e., the usual processes of mutation and crossover operations are applied [45]); *(Step 6)* the least fit members of the population are replaced by the offsprings from *Step 5*; and *(Step 7)* loop through steps *5–7*, until some suitable "optimality" threshold (or some other convergence criterion) is satisfied.

Halpern [23,24] reports a wide range of resulting behaviors, collectively suggesting a clear relationship between the parameters defining the GA optimization and lattice connectivity. Of particular interest are the dynamic conditions for which the fitness-based creation and deletion of links increases the rate of growth of overall fitness. The fastest convergence occurs when lattice connectivity first increases, then decreases, then eventually levels off. In the first stage, the fittest possible communities are first established; in the second stage, connections with poorer candidate solution are deleted; finally, in the third stage, the system essentially "fine-tunes" its optimal solutions. Halpern [24] finds two different evolutionary paths toward high connectivity: (1) monotonic growth over time (for low mutation rates, p_μ), and (2) a *phase transition* between low and high degrees of connectivity (for some p_μ^*). Using SDCA↔GA hybrid model parameters $N = 100$ and $f_D = f_C = 0.1$, $p_\mu^* \approx 0.05$, for which Halpern [24] finds a sharp increase in the number of links per site between generations 350 and 450.

Despite the novelty of the approach, and the promising link between optimization rates and dynamic structure established in [23], concrete applications of the algorithm – except for Weinert et al. [72] work on a related hybrid GA algorithm – have yet to be developed. One suggestion, from Halpern [24], is to use the SDCA↔GA hybrid model for finding "optimal" connectivity patterns in parallel computers. The search algorithm may be used to directly model how component processors are connected, and decide to keep or sever existing links, or establish new ones, adaptively as a function of local fitness criteria.

Generalized SDCA Models

Despite SDCA being obviously more "complex" than conventional CA (and certainly more complex to formally define, if only because one must specify *both* σ *and* ℓ rules), the SDCA model nonetheless has more in common with *elementary* CA than with any of its brethren's more "complicated" variants. By "elementary" CA we mean the simplest one-dimensional CA with $\sigma \in \{0, 1\}$ and local neighborhoods consisting only of left and right (i. e., *nearest*) neighbors. Just as there are many generalizations of elementary CA – for example, increasing the state space to include σ's that take on one of N values, larger-sized neighborhoods, and memory, among many other possibilities – so too there are natural extensions of basic SDCA. In this section we discuss three generalizations: (1) rules that are *reversible in time*, (2) rules that retain a *memory of past states*, and (3) *probabilistic* rules.

Reversible SDCA

The first generalization of the basic SDCA model, explored extensively by Alonso-Sanz [7], is to apply the *Fredkin reversible-rule construction* to ℓ rules to render them reversible in time. Consider a conventional CA system that is first-order in time, $\sigma_i^{t+1} = \phi[\sigma_j^t \in \mathcal{N}_i]$, where $\mathcal{N}_i$ is the

neighborhood around site i and, generally, $\sigma_i \in \mathcal{Z}_k$. The *Fredkin* construction converts this system into an explicitly invertible one that is *second-order* in time by subtracting the value of the center site at time $t-1$:

$$\sigma_i^{t+1} = \phi\left[\sigma_j^t \in \mathcal{N}_i\right] \ominus_k \sigma_i^{t-1}, \tag{26}$$

where '$\ominus_k$' is subtraction *modulo-k*. Since Eq. (26) can be trivially solved for σ_i^{t-1} $\left(= \phi[\sigma_j^t \in \mathcal{N}_i] \ominus_k \sigma_i^{t+1}\right)$, we see that any *pair of consecutive configurations uniquely specifies the backwards trajectory of the system*. Moreover, this is true for arbitrary (and, in particular, *irreversible*) functions ϕ.

Now, exactly the same procedure may be applied to link functions:

$$\begin{cases} l_{ij}^{t+1} = \psi\left(\{\sigma_k^t\}, \{l_{ij}^t\}\right) \ominus_2 l_{ij}^{t-1}, \\ l_{ij}^{t+1} = \omega\left(\{\sigma_k^t\}, \{l_{ij}^t\}\right) \ominus_2 l_{ij}^{t-1}, \end{cases} \tag{27}$$

where $\ominus_2$ is subtraction *modulo*-2 (since links are obviously binary valued).

Following Alonso-Sanz [7,8], we consider these two specific SDCA link rules (which will also be used in a later example):

$$\begin{cases} \psi\left(\sigma_i^t, \sigma_j^t, l_{ij}^t\right) = 0 \text{ iff } l_{ij}^t = 1 \text{ and } \sigma_i^t + \sigma_j^t = 0\,, \\ \omega\left(\sigma_i^t, \sigma_j^t, l_{ij}^t\right) = 1 \text{ iff } l_{ij}^t = 0, \sigma_i^t > 0, \sigma_j^t > 0, \text{ and } D_{ij} = 2. \end{cases} \tag{28}$$

Figure 10 compares the evolution of the *Fredkin reversible* version of these rules to their memoryless counterpart. Both evolutions start on a two dimensional hexagonal lattice, and values evolve according to the three-state (i. e., $\sigma \in \{0, 1, 2\}$), next-nearest neighborhood **T** *beehive* rule. The *beehive* rule is defined explicitly by assigning one of three values (0, 1, or 2), to each possible 3-tuple, (N_0, N_1, N_2), that gives the number of local sites with N_0 0s, N_1 1s, and N_2 2s [7]: $(0,0,6) \to 0$, $(0,1,5) \to 1$, $(0,2,4) \to 2$, $(0,3,3) \to 1$, $(0,4,2) \to 2$, $(0,5,1) \to 0$, $(0,6,0) \to 0$, $(1,0,5) \to 0$, $(1,1,4) \to 2$, $(1,2,3) \to 2$, $(1,3,2) \to 2$, $(1,4,1) \to 1$, $(1,5,0) \to 1$, $(2,0,4) \to 0$, $(2,1,3) \to 0$, $(2,2,2) \to 2$, $(2,3,1) \to 2$, $(2,4,0) \to 0$, $(3,0,3) \to 0$, $(3,1,2) \to 2$, $(3,2,1) \to 2$, $(3,3,0) \to 0$, $(4,0,2) \to 0$, $(4,1,1) \to 0$, $(4,2,0) \to 2$, $(5,0,1) \to 2$, $(5,1,0) \to 0$, $(6,0,0) \to 0$.

The top row of Fig. 10 shows the first four steps ($t = 1, 2, 3$, and 4) in the memoryless evolution of the initial "ring" of sites that appears at $t = 1$. The link rules used for this run are those defined in Eq. (28). Since the decoupler removes links between pairs of sites whose values are equal to zero, most of the lattice disappears after a single time step, and both value and link activity is confined to a small region. After two more steps of changes, the system quickly attains a fixed point: $\{\sigma^t, \ell_{ij}^t\} = \{\sigma^{t=4}, \ell_{ij}^{t=4}\}$ for all $t \geq 5$. While the frequency of states is not constrained to total six for a dynamic lattice, the *beehive* rule is unchanged; if the sum of frequencies at a given site exceeds six, the site value remains the same.

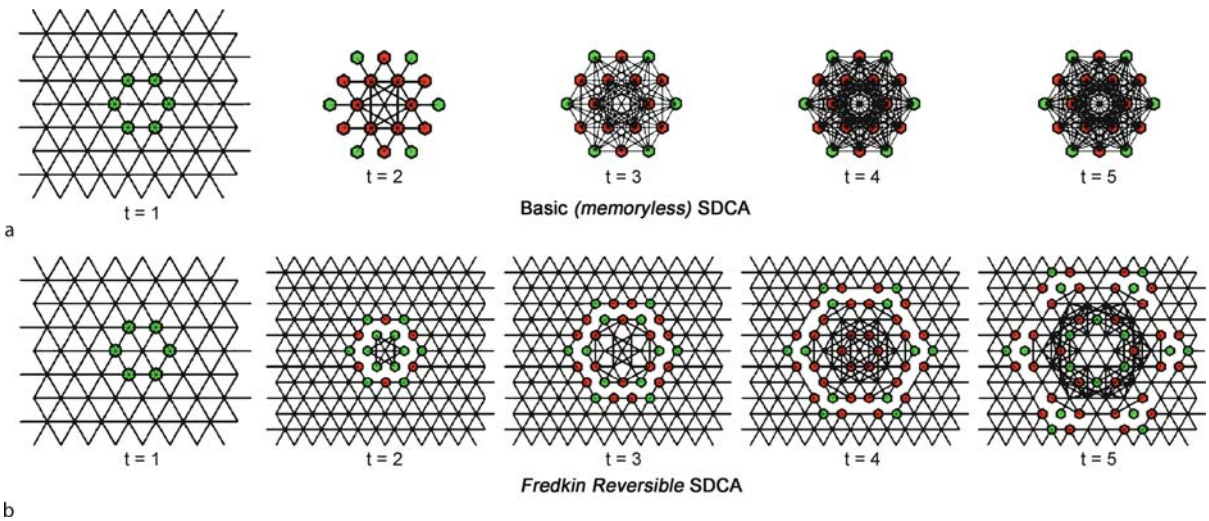

Structurally Dynamic Cellular Automata, Figure 10

Comparison between first few time steps of a a *memoryless* SDCA, evolving according to link rules defined in Eq. (28), and b the *Fredkin reversible* versions of these rules (obtained by applying Eq. (27) to Eq. (28)). In both cases, σ's evolve according to the *beehive* rule defined in the text. (Reproduced with permission from [7])

The bottom row shows the evolution of the *Fredkin reversible* versions of the rules defined in Eq. (28) (to simplify the visualization, links along the border sites are not shown). In contrast to the basic SDCA version, the initial lattice in this case *does not* decay. Since, according to Eq. (27) (which assumes that $\ell_{ij}^{t=0} = \ell_{ij}^{t=1}$), the initial hexagonal lattice is subtracted from the evolved structure at $t = 1(modulo\text{-}2)$, the original graph is effectively restored, and the outlying regions appear undisturbed.

SDCA with Memory

A second generalization to the basic SDCA model, introduced and studied by Alonso-Sanz and Martín [6,7,8], is to endow both σ-rules and ℓ-rules with *memory*. The rules for conventional memoryless CA and SDCA, depend only on neighborhood configurations that appear on the immediately preceding time step. Therefore, rules may be said to possess a "memory" of depth m if they depend explicitly on values (in the case of CA), or on both values and link states (in the case of SDCA), that existed on m previous time steps. We note, in passing, that since the *Fredkin construction* couples states at times $t+1$, t and $t-1$, reversibility may be considered a specific form of memory that extends backwards a single step.

Of course, there is no unique prescription for introducing a dependency on past values; and a variety of alternative memory mechanisms have been proposed in the literature (for example, see page 43 in [32] and page 118 in [74]). We focus our discussion on the approach proposed by Alonso-Sanz (14), and for the moment confine our attention to value rules, $\phi : \sigma \to \sigma'$. Alonso-Sanz's approach is to preserve the form of the transition rule, but have it act on an effective site value that is a weighted function of its m prior values.

This is done by introducing a memory-endowed value rule, ϕ_m, that – in contrast to its memoryless version, ϕ – is not, in general, a function of a given site's current value, σ_i, *alone*, but is instead a function of the transformed value, $s = \mathcal{M}_\phi(\sigma; m, \alpha)$, obtained from σ_i's past m values: $\phi_m : s \to \sigma'$, where $0 \le \alpha \le 1$ is a numerical *memory factor*. The value transforming memory function, $\mathcal{M}$, assumes the following specific form (to avoid confusion, note that in Eqs. (29) and (30), σ_i^x means *the value of* σ_i *at time* $t = x$, and α^x means *the numerical quantity* α *raised to the power* x):

$$s_i^t = M_\phi\left(\sigma_i^t; m, \alpha\right) = \begin{cases} 1 & \text{if } \left(\hat{\sigma}_i^t\right)_m > 1/2, \\ \sigma_i^t & \text{if } \left(\hat{\sigma}_i^t\right)_m = 1/2, \\ 0 & \text{if } \left(\hat{\sigma}_i^t\right)_m < 1/2, \end{cases} \tag{29}$$

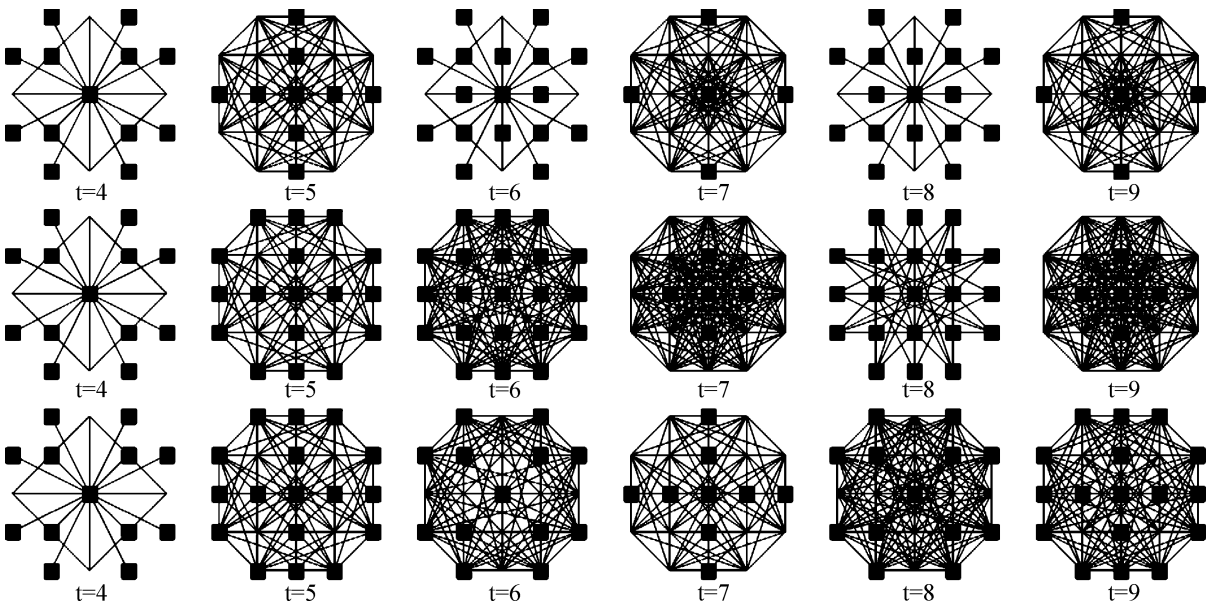

Structurally Dynamic Cellular Automata, Figure 11

Sample runs of a SDCA with memory for memory weighting $\alpha = 0.6$. The SDCA is initialized as a Euclidean four-neighbor lattice, and evolves according to the *parity* T σ-rule and the two ℓ rules defined in Eq. (28). The first row of evolving patterns applies memory only to *values*; the second row applies memory only to *links*, and the third row shows the evolution when memory is applied to both. (Reproduced by permission from [8])

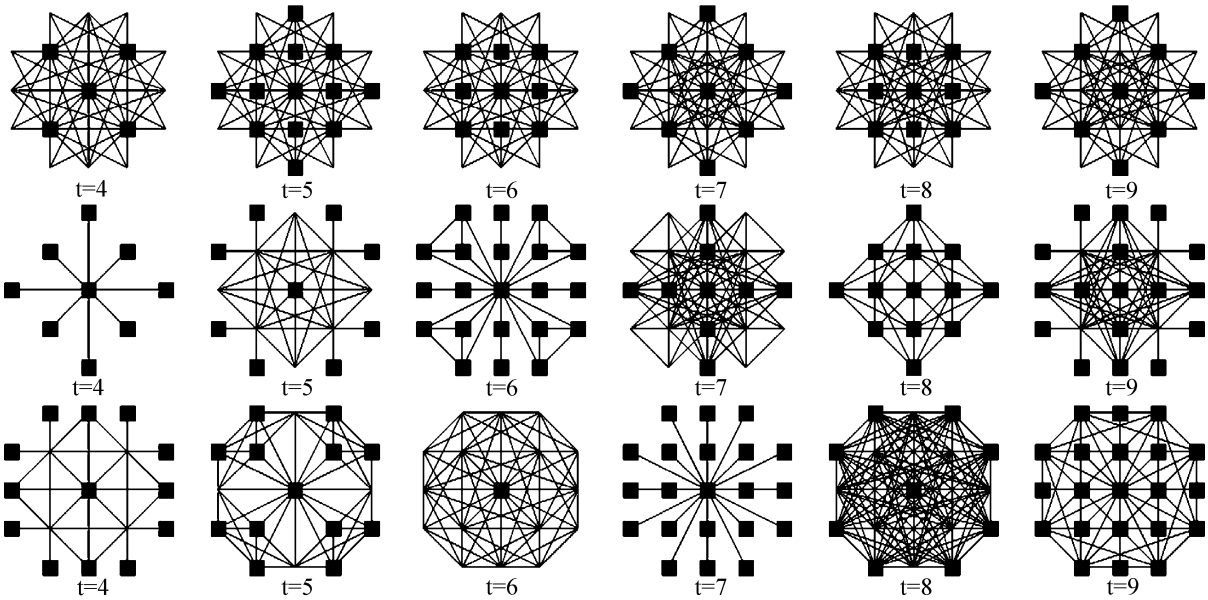

Structurally Dynamic Cellular Automata, Figure 12
Sample runs of the same SDCA shown in Fig. 11, but with memory weighting $\alpha = 1.0$. (Reproduced by permission from [8])

where

$$\left(\hat{\sigma}_i^t\right)_m = \frac{\sigma_i^t + \sum_{\Delta t=1}^{m} \alpha^{\Delta t} \cdot \sigma_i^{t-\Delta t}}{1 + \sum_{\Delta t=1}^{m} \alpha^{\Delta t}} \,. \tag{30}$$

At any given time, t, the depth m can never exceed $t-1$. Our discussion follows Alonso-Sanz [8], and sets $m(t) \equiv t-1$ for all t; i. e., we assume that $\mathcal{M}_\phi(\sigma; m, \alpha)$ yields a weighted mean value of *all* the previous values of a given site. In practice, memory becomes active only after a certain number of initialization steps, here taken to be three; with seeded values $s_i^1 = \sigma_i^1$ and $s_i^2 = \sigma_i^2$.

Memory can be added to link rules in a similar manner. The form of the link rules (ψ and ω) remains the same, but rather than acting on a graph that is defined by its adjacency matrix, ℓ_{ij}^t, ψ and ω instead act on the memory-transformed values, $L = \mathcal{M}_{(\psi,\omega)}(\ell; m, \alpha)$:

$$L_{ij}^t = M_{(\psi,\omega)}\left(l_{ij}^t; m, \alpha\right) = \begin{cases} 1 & \text{if } \left(\hat{l}_{ij}^t\right)_m > 1/2, \\ l_{ij}^t & \text{if } \left(\hat{l}_{ij}^t\right)_m = 1/2, \\ 0 & \text{if } \left(\hat{l}_{ij}^t\right)_m < 1/2, \end{cases} \tag{31}$$

where

$$\left(\hat{l}_{ij}^t\right)_m = \frac{l_{ij}^t + \sum_{\Delta t=1}^{m} \alpha^{\Delta t} \cdot l_{ij}^{t-\Delta t}}{1 + \sum_{\Delta t=1}^{m} \alpha^{\Delta t}} \,. \tag{32}$$

As for memory-endowed σ-rules, the memory for link rules is activated only on the third iteration step, and the system is initialized by setting $L_i^1 = \sigma_i^1$ and $L_i^2 = \sigma_i^2$.

Figures 11 and 12 show the effects of applying partial memory weighting ($\alpha = 0.6$) and full memory ($\alpha = 0.6$), respectively, to a SDCA that starts with a Euclidean four-neighbor lattice, and evolves according to the *parity* **T** σ-rule (that assigns a value *zero* to a site if the sum of the values in its neighborhood is *even*, and assigns the value *one* if the sum is *odd*) and the ℓ rules defined above in Eq. (28). The first row of evolving patterns (for each α) applies memory only to *values*; the second applies memory only to *links*, and the third appliers memory to both. Figure 13 shows the reversible *beehive* SDCA shown in Fig. 10, but with full memory ($\alpha = 1.0$).

Probabilistic SDCA

Another natural extension of the basic SDCA model is to replace the set of explicit σ- and/or ℓ-rules with probabilities. In this way one can study the evolution of a system that undergoes random but σ-dependent lattice changes. For example, this may be useful for studying genetic networks in which new links are forged (with a given probability) only if both genes are active, and existing connections are broken if both sites are inactive.

Following Halpern and Caltagirone [22], consider the *parity* **T** σ-rule and the following probabilistic versions of

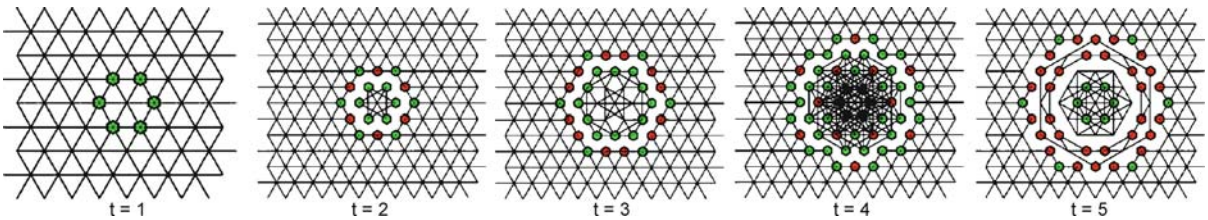

Structurally Dynamic Cellular Automata, Figure 13
Sample runs of a reversible *beehive* SDCA with full memory ($\alpha = 1.0$); compare to Fig. 10. (Reproduced by permission from [8])

decoupler (ψ_p) and coupler rules (ω_p):

$$(decoupler): \begin{cases} l_{ij}^{t+1} = \psi_p\left(l_{ij}^t; \sigma_i^t, \sigma_j^t, p_D\right), \\ \psi_p \equiv 1 - \delta\left(\sigma_i^t + \sigma_j^t, 0\right) \cdot \delta\left(p_D > r\right), \end{cases}$$
$$(coupler): \begin{cases} l_{ij}^{t+1} = \omega_p\left(l_{ij}^t; \sigma_i^t, \sigma_j^t, p_D\right), \\ \omega_p \equiv \delta\left(D_{ij}, 2\right) \cdot \delta\left(\sigma_i^t + \sigma_j^t, 2\right) \cdot \delta\left(p_C > r\right), \end{cases} \tag{33}$$

where P_D and P_C are the *decoupler* and *coupler* probabilities, respectively, and r is a random number between 0 and 1.

Thus, ψ_p unlinks two previously linked sites with probability P_D *if and only if* the sum of their site values is *zero*; and ω_p links two previously unlinked sites with probability P_C *if and only if* they are next-nearest neighbors and the sum of their site values is *two*.

Figure 14 shows time series plots of $\langle\sigma\rangle$ as a function of time for three different cases: (1) $P_D = 0$ (no decoupling at all), (2) $P_D = 1/2$, and (3) $P_D = 1$ (decoupler rule applied 100% of the time (consistent with *non*-probabilistic SDCA rules). We see that changing P_D induces qualitatively different σ behavior, that ranges from small fluctuations around $\langle\sigma\rangle \sim 0.5$ (for $P_D = 0$), to decay to small static values ($\langle\sigma\rangle = 0.05$ for $P_D = 1/2$, and $\langle\sigma\rangle = 0.12$ for $P_D = 1$).

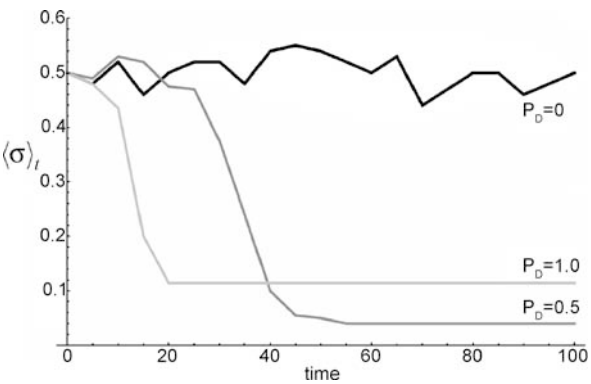

Structurally Dynamic Cellular Automata, Figure 14
Time series of average σ value, $\langle\sigma\rangle_t$, for the Halpern-Caltagirone rules (defined in Eq. (33)) and for three values of *decoupler* probability: $P_D = 0$, $P_D = 1/2$, and $P_D = 1$. (Reproduced with permission from [22])

Halpern and Caltagirone [22] have studied a wide range of probabilistic SDCA, using random initial σ configurations, *step-function*, *parity*, and Conway's *life* σ-rules, Cartesian and random initial lattice structures, and various probabilities $0 \le P_D \le 1$ and $0 \le P_C \le 1$. Some of their results are reproduced (with permission) in the *behavioral phase* plots shown in Fig. 15. (The *step-function* rule is defined by $\sigma_i^{t+1} = 0$ if and only if $\sum_j \ell_{ij}^t \sigma_i^t > 2$ and $\sigma_i^{t+1} = 1$ if and only if $\sum_j \ell_{ij}^t \sigma_i^t \le 2$; Conway's *life* rule assigns $\sigma^{t+1} = 1$ to a site if and only if $\sigma^{(t)} = 0$ and the sum of values in its neighborhood at time t is equal to 3 or $\sigma^t = 1$ and the sum of values is equal to 2 or 3; otherwise $\sigma^{t+1} = 0$.)

Figure 15 shows a wide range of possible behaviors. Consider, for example, the number of *links per site* for the case where the lattice is updated with probabilistic ℓ-rules and the σ's are all random (shown at the top left of the figure). Four distinct classes of behavior appear, with *growth* dominant for most values of P_D and P_C.

Pure decoupling (or pure coupling) leads to complete decay (or growth to a stable state); a mixed state of coupling/decoupling generally yields slow growth. Periodic behavior occurs only for $P_D \sim P_C \sim 1$. Compare this behavior with the cases where the σ-rule is either the *parity* value rule (shown in the middle of the top row of Fig. 15) or the *step-function* rule (shown at left bottom of the figure). While the *parity* rule also displays four similar phases (growth to stability, decay to stability, incomplete growth, and incomplete decay), decaying structures eventually reach a stable (not null) final state. The *step-function* rule shows an even greater variety of possible behaviors, and appears more sensitive to small changes in link probabilities.

The probabilistic SDCA system discussed in this section adds a stochastic element *specifically to SDCA*. Of course, there are other ways of injecting stochasticity into a CA with dynamic topology. For example, Makowiec [40] combines the deterministic evolution of a conventional

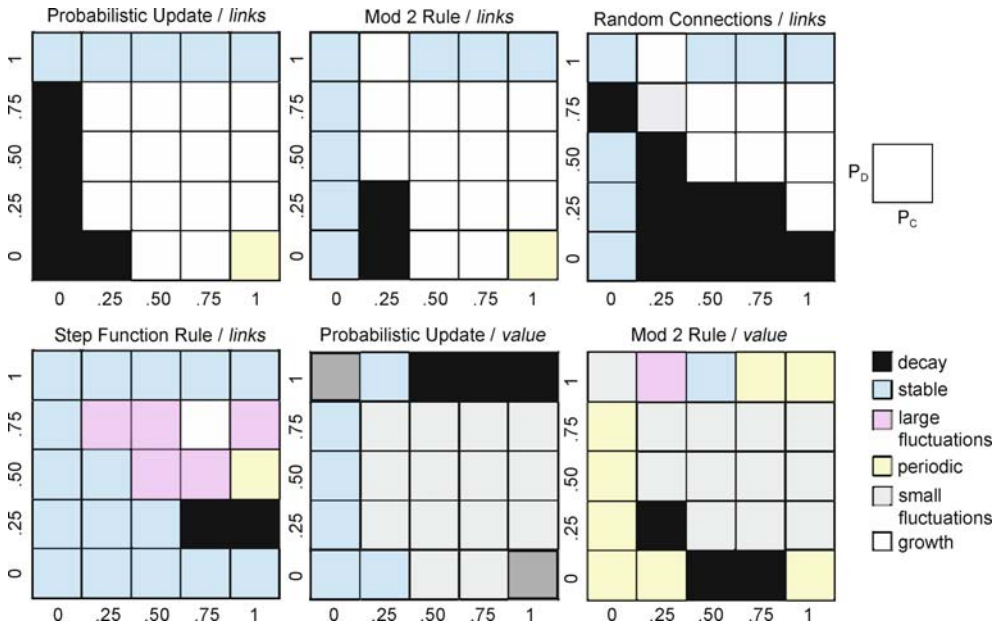

Structurally Dynamic Cellular Automata, Figure 15

Behavioral phase **plots summarizing the long term evolution for several different σ and ℓ-rules defined in Eq. (33). The *x* and *y* axes for each plot depict values ($\in \{0, .25, .5, .75, 1\}$) of P_C and P_D, respectively. There are six classes of behavior: *growth*, *decay*, *stability*, large and small *fluctuations* (around a stable lattice), and *periodicity*. The initial graph is a Cartesian four-neighbor lattice in each case except for the *top-right* plot (labeled *Random connections/links*) for which the initial graph is random. (Reproduced with permission from [22])**

CA with an asynchronous stochastic evolution of its underlying lattice (patterned after the Barabasi–Albert [9] model of degree distributions in small-world networks), to explore the influence of dynamic topology on the zero-temperature limit of ferromagnetic transitions.

Random Dynamics Approximation

For cases in which the structure and value configurations are both sufficiently random and uncorrelated, a *random dynamics approximation* (abbreviated, RDA) may suffice to qualitatively predict how the system will tend to evolve under a specific rule set; for example, to predict whether a given rule is more (or less) likely to yield unbounded growth, to eventually settle into a low periodic state, or to simply decay. The idea is to approximate the real SDCA as a *mean-field*; that is, assume all local value and structural correlations are close to zero (and can thus be ignored), and replace all specific site values and local link geometries with average, or effective, values.

More precisely, assuming that (1) the probability $p_n^{(\sigma_i)}$ of a site 'i' having value $\sigma = 1$ at time $t = n$ is the same for all sites – so that $p_n^{(\sigma_i)} = p_n^{(\sigma)}$ for all i – and (2) that the probability $p_n^{(l_{ij})}$ of two sites 'i' and 'j' being linked at $t = n$ is the same for all pairs of sites – so that $p_n^{(l_{ij})} = p_n^{(l)}$ for all i and j – the RDA evolution equations may be written formally as follows:

$$\begin{cases} p_{n+1}^{(\sigma)} = F_{\mathrm{RDA}}[p_n^{(\sigma)}, p_n^{(l)}; \Gamma_{\mathrm{SDCA}}], \\ p_{n+1}^{(l)} = G_{\mathrm{RDA}}[p_n^{(\sigma)}, p_n^{(l)}; \Gamma_{\mathrm{SDCA}}], \end{cases} \tag{34}$$

where SDCA's rule Γ_{SDCA} (defined in Eq. (14)) is included, formally, to remind us that the functional forms assumed by F_{RDA} and G_{RDA} will be different for different Γ_{SDCA}s.

The first function, F_{RDA}, is the easier of the two to calculate. For any given site with degree d we simply count the total number of ways to distribute the local σ-values among the d possible neighboring sites to obtain the desired sums that define a given rule. In this way we find the average expected σ density at $t = n + 1$, assuming all sites in the lattice have the same degree d at time $t = n$:

$$p_{n+1}^{(\sigma)}(d, p_n^{(\sigma)}) = \begin{cases} \sum_{\{\alpha\}} \binom{d+1}{\alpha} \left[p_n^{(\sigma)}\right]^{\alpha} \left(1 - p_n^{(\sigma)}\right)^{d+1-\alpha} & \leftrightarrow \mathbf{T} \\ \sum_{\{\alpha_0\}} \binom{d}{\alpha_0} \left[p_n^{(\sigma)}\right]^{\alpha_0+1} \left(1 - p_n^{(\sigma)}\right)^{d-\alpha_0} & \\ \quad + \sum_{\{\alpha_1\}} \binom{d+1}{\alpha_1} \left[p_n^{(\sigma)}\right]^{\alpha_1} \left(1 - p_n^{(\sigma)}\right)^{d+1-\alpha_1} & \leftrightarrow \mathbf{OT} \end{cases} \tag{35}$$

We then get $F_{\mathrm{RDA}} \to p_{n+1}^{(\sigma)} = \sum_d P(d; p_n^{(l)}) \cdot p_n^{(\sigma)}(d, p_n^{(\sigma)})$ as an average over all possible degrees, where $P(d; p_n^{(l)})$ is the probability that any site has exactly d neighbors. Since this means that, out of a total of $N-1$ possible neighbors, a given site must have exactly d links, and not be connected to any of the remaining $(N-1-d)$ sites, we have by inspection:

$$P\left(d; p_n^{(l)}\right) = \binom{N-1}{d}\left[p_n^{(l)}\right]^d \left(1-p_n^{(l)}\right)^{N-1-d}. \quad (36)$$

To calculate the second function in Eq. (34) ($= G_{\mathrm{RDA}}$), we first define the local transition functions

$$\begin{cases} p_n^a(d_1, d_2, \lambda) \\ = \mathrm{Prob}\left(l=1 \to l'=0 \mid d_i=d_1, d_j=d_2, \left|A_{ij}\right|=\lambda\right), \\ p_n^b(d_1, d_2, \lambda) \\ = \mathrm{Prob}\left(D=2 \to l'=1 \mid d_i=d_1, d_j=d_2, \left|A_{ij}\right|=\lambda\right), \end{cases} \quad (37)$$

which give the probabilities that any two sites – i and j – will be *disconnected* (p_n^a) or *connected* (p_n^b) if they have prescribed degrees $d_i = d_1$ and $d_j = d_2$, and are each linked to the *same* λ sites in the shared neighbor set, A_{ij} (see Fig. 1). In the case of type-**T** σ- and ℓ-rules, p_n^a and p_n^b are given explicitly by (**OT** versions of σ- and ℓ-rules, and **RT** versions of ℓ-rules are defined by similar, but slightly more complicated, expressions):

$$\begin{cases} p_n^a(d_1, d_2, \lambda) = \sum_k \binom{d_1+d_2-\lambda}{\beta_k}\left[p_n^{(\sigma)}\right]^{\beta_k} \\ \qquad \cdot \left(1-p_n^{(\sigma)}\right)^{d_1+d_2-\lambda-\beta_k}, \\ p_n^b(d_1, d_2, \lambda) = \sum_k \binom{d_1+d_2+2-\lambda}{\varepsilon_k}\left[p_n^{(\sigma)}\right]^{\varepsilon_k} \\ \qquad \cdot \left(1-p_n^{(\sigma)}\right)^{d_1+d_2+2-\lambda-\varepsilon_k}, \end{cases} \quad (38)$$

where β_k and ϵ_k refer to the sums that appear in Eqs. (11) and (12). The total probability that any two sites will be disconnected ($l = 1 \to l' = 0$) or connected ($D = 2 \to l' = 1$) – P_n^a and P_n^b, respectively – may then be obtained by summing over all possible local topologies:

$$\begin{cases} P_n^a \equiv \left\langle \mathrm{Prob}\left(l=1 \to l'=0\right)\right\rangle \\ \quad = \sum_{d_1}\sum_{d_2}\sum_{\lambda} P_1(d_1, d_2, \lambda) \cdot p_n^a(d_1, d_2, \lambda), \\ P_n^b \equiv \left\langle \mathrm{Prob}\left(D=2 \to l'=1\right)\right\rangle \\ \quad = \sum_{d_1}\sum_{d_2}\sum_{\lambda} P_2(d_1, d_2, \lambda) \cdot p_n^b(d_1, d_2, \lambda), \end{cases} \quad (39)$$

where

$$\begin{cases} P_1(d_1, d_2, \lambda) = \mathrm{Prob}(\text{sites } i, j \mid l_{ij}=1 \text{ have } d_i = d_1, \\ \qquad d_j = d_2, \left|A_{ij}\right| = \lambda), \\ P_2(d_1, d_2, \lambda) = \mathrm{Prob}(\text{sites } i, j \mid l_{ij}=0 \text{ have } d_i = d_1, \\ \qquad d_j = d_2, \left|A_{ij}\right| = \lambda). \end{cases} \quad (40)$$

To find P_1 we need to count, from among the remaining $N-2$ sites, the number of ways of selecting disjoint sets S_1, containing $d_1 - 1 - \lambda$ sites linked only to i; S_2, consisting of $d_2 - 1 - \lambda$ sites connected only to j; and S_3, with λ sites linked to *both* i and j. But this is simply a multinomial coefficient, so we can write:

$$P_1(d_1, d_2, \lambda) = \frac{(N-2)_{d_1+d_2-\lambda-2}}{(d_1-1-\lambda)!\,(d_2-1-\lambda)!\lambda!} \cdot \left[p^{(l)}\right]^{d_1+d_2-2}\left(1-p^{(l)}\right)^{2(N-1)-d_1-d_2}, \quad (41)$$

where $(n)_k \equiv n(n-1)\cdots(n-k+1)$. Similarly, for P_2, we need to count the number of ways of choosing $d_1 - \lambda$ sites from i, $d_2 - \lambda$ sites from j, and λ sites from both:

$$P_2(d_1, d_2, \lambda) = \frac{(N-3)_{d_1+d_2-\lambda-2}}{(d_1-\lambda)!\,(d_2-\lambda)!\lambda!} \cdot \left[p^{(l)}\right]^{d_1+d_2}\left(1-p^{(l)}\right)^{2(N+\lambda-3)-d_1-d_2}. \quad (42)$$

The second (link-update) function of the pair of functions in Eq. (34) is thus given by

$$G_{\mathrm{RDA}} \to p_{n+1}^{(l)} = p_n^{(l)} \cdot \left(1-P_n^a\right) + \left(1-p_n^{(l)}\right) \cdot P_{D=2} \cdot P_n^a, \quad (43)$$

where, assuming that two sites, i and j, are not themselves connected, $P_{D=2}$ = probability that there exists at least one site k, such that $D_{ik} = D_{jk} = 1$, which implies that $P_{D=2} = 1 - \mathrm{Prob}\,(\text{there is no such } k) = 1 - (1 - [p_n^{(l)}]^2)^{N-2}$; and P_1 and P_2 are defined in Eqs. (41) and (42).

A *structural equilibrium* is established when $p_{n+1}^{(l)} \approx p_n^{(l)}$, which happens when the average number of new connections is equal to the average number of link deletions: $P_n^b \cdot \langle N_{\mathrm{nn}}\rangle = P_n^a \cdot \langle \mathrm{deg}\rangle$, where $\langle \mathrm{deg}\rangle = p_n^{(l)}(N-1)$ is the average *degree*, and $\langle N_{\mathrm{nn}}\rangle$ is the average number of *next-nearest neighbors*. For SDCA rules that naturally tend to produce graphs with minimal site value and structural

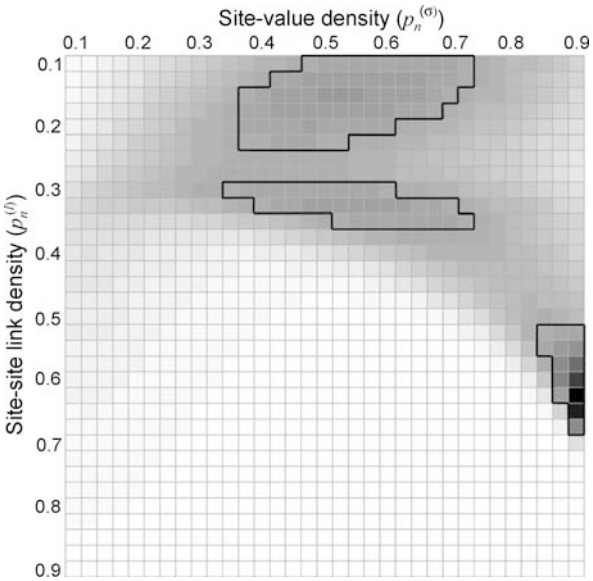

Structurally Dynamic Cellular Automata, Figure 16
Density-plot of $\gamma_{\text{c::d}}$ for an OT *decoupler* rule: {(0, 0), (1, 1), (1, 2), (2, 2)}; an OT *coupler* rule: {(1, 1)}; $0.1 \leq p_n^{(\sigma)} \leq 0.9$; and $0.1 \leq p_n^{(l)} \leq 0.9$; the rectangular area highlighted in black denotes the "equilibrium boundary" that separates regions of growth and decay

correlations, the predicted ratio of RDA link creations to deletions, $\gamma_{c:d} \equiv P_n^b \cdot \langle N_{\text{nn}} \rangle / P_n^a \cdot \langle \text{deg} \rangle$, may be used to predict qualitatively how the graphs will evolve. Since the average number of pairs of sites a distance $D = 2$ apart $= \binom{N}{2} \cdot P_{D=2} = \langle N_{\text{nn}} \rangle \cdot N/2$, we find that:

$$\gamma_{\text{c::d}} = \frac{P_n^b}{P_n^a} \cdot \frac{\left(1 - p_n^{(l)}\right)}{p_n^{(l)}} \cdot \left\{ 1 - \left(1 - \left[p_n^{(l)} \right]^2 \right)^{N-2} \right\} . \quad (44)$$

$\gamma_{\text{c::d}}$ is also implicitly a function of site-value density, since $p_n^{(\sigma)}$ appears in both P_n^a and P_n^b, defined in Eq. (39).

Figure 16 shows a grayscale density-plot of $\gamma_{\text{c::d}}$ for an **OT** *decoupler* rule: $\{(0, 0), (1, 1), (1, 2), (2, 2)\}$; an **OT** *coupler* rule: $\{(1, 1)\}$; $0.1 \leq p_n^{(\sigma)} \leq 0.9$; and $0.1 \leq p_n^{(l)} \leq 0.9$. Areas that are close to *white* represent combinations of $(p_n^{(\sigma)}, p_n^{(l)})$ for which $\gamma_{\text{c::d}} \ll 1$, and which therefore predict "decay"; areas that are close to *black* represent combinations of $(p_n^{(\sigma)}, p_n^{(l)})$ for which $\gamma_{\text{c::d}} \gg 1$, and predict "growth"; the rectangular area highlighted in black denotes the "equilibrium boundary" that separates regions of growth and decay.

Related Graph Dynamical Systems

The original SDCA model [28] represents *one* (albeit not entirely arbitrary) approach to dynamically coupling site values ($\{\sigma_i\}$) and topology ($\{l_{ij}\}$), of the normally quiescent lattice. Since this model was primarily introduced as a general tool to explore self-organized emergent *geometries*, σ values are an integral dynamic component only because SDCA's original rules were conceived to generalize conventional CA rules, not replace them. Moreover, SDCA's link rules are, by design, close analogs of their conventional-CA brethren; this is the reason why SDCA's ψ and ω rules assume the familiar **T** and **OT** (and related **RT**) forms, as defined in Sect. "The Basic Model". Indeed, while the preceding sections of this article have introduced several generalizations – such as the addition of probabilistic rules, reversibility and memory – in each case, the basic *form* of the rules (as defined in Eqs. (11), (12), and (13)) has remained essentially the same. However, just as for conventional CA, an almost endless variety of different *kinds* of rules can in principle be defined; including rules that alter the geometry but are *not* functions of the σ states. In this section, we look at two illustrative examples of SDCA-like dynamical systems: one that uses coupled σ-ℓ rules, and another whose rules depend only on *topology*.

Graph Rewriting Automata

Tomita, Kurokawa, and Murata [67,68,69,70,71] have recently introduced *graph rewriting automata* (abbreviated, GRA), in which both links and (the number of) sites are allowed to change. Motivated by CA models of self-reproduction, Tomita et al suggest that fixed, two-dimensional lattices – used as static backdrops to most conventional models – are unnecessarily restrictive for describing self-reproductive processes. They cite, as an example, the inability of conventional CA to describe biological processes (such as embryonic development) that must unfold in a finite closed space; once the underlying space of the CA is defined at the start, however large (and sometimes deliberately assumed *infinite*), its size remains the same throughout the development. This not only makes it hard to model the typically growing need that developing organisms have for *space*, but makes it impractical even to provide some room for avoiding overlaps between the original and daughter patterns [67].

Motivated by these, and other issues related to computation, Tomita et al. [67,69] introduce GRA, which is a form of graph grammar [20]. At first glance, GRA appear superficially similar to SDCA, at least in the sense that they both dynamically couple site values with topology. However, the transition rules are very different, and – in GRA's case – two properties hold that are not true for SDCA systems: (1) *all sites have exactly three neighbors at all times* (which is the minimum number of neighbors that yield nontrivial graphs [67]), and (2) *multiple links*

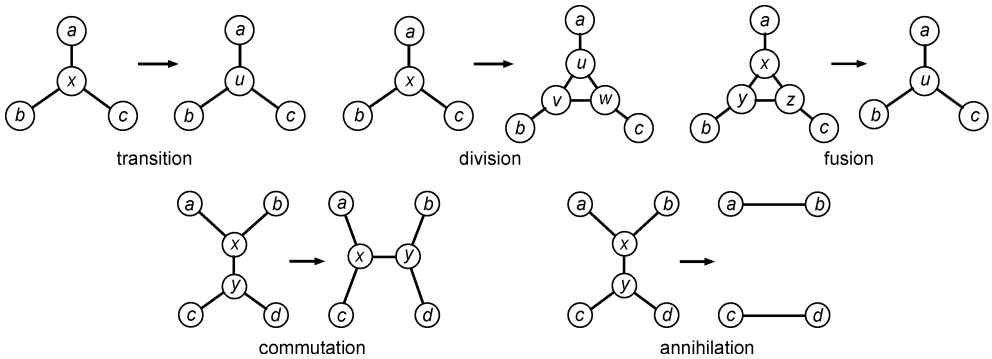

Structurally Dynamic Cellular Automata, Figure 17
Graphical representations of the actions of the GRA rules defined in Eq. (45). (Reproduced from [69] with permission)

are allowed to exist between any two sites. The authors claim that the 3-neighbor restriction not only does *not* constrain the space of emergent geometries (an observation that is echoed by Wolfram [75]; see Subsect. "Network Automata" below) but has the added benefit of allowing the rules to be expressed in a regular form: each rule is defined by a rule *name* and, at most, six *symbols* for its argument:

$$
\begin{aligned}
&(\sigma\ \textit{rules}):\\
&\quad \left\{ \textit{transition}\,(x,a,b,c) \quad \to (u,a,b,c), \right.\\
&(\textit{site rules}):\\
&\quad \left\{ \begin{array}{ll} \textit{division}\,(x,a,b,c) & \to (u,v,w,a,b,c),\\ \textit{fusion}\,(x,y,z,a,b,c) & \to (u,a,b,c), \end{array} \right.\\
&(\textit{link rules}):\\
&\quad \left\{ \begin{array}{ll} \textit{commutation}\,(x,y,a,b,c,d) & \to (x,y,a,b,c,d),\\ \textit{annihilation}\,(x,y,a,b,c,d) & \to (a,b,c,d), \end{array} \right.
\end{aligned}
\tag{45}
$$

where x, y and z denote the σ values of the center sites *before* undergoing a structural change; u, v and w denote the σ values of the center sites *after* the structural change; and a, b, c, and d denote the states of the *neighboring sites*. The ordering is unimportant, so long as a given string can be obtained from another by cyclic permutation, otherwise the strings are different; i. e., (a,b,c) is both topologically and functionally equivalent to (b,c,a), but (c,b,a) is different. The action of σ, value, and links is graphically illustrated in Figure 17.

By convention, the GRA algorithm is applied in two steps: (1) site rules (*transition*, *division* and *fusion*) are executed first, and at all subsequent even time steps, followed by (2) link rules (*commutation* and *annihilation*), executed at odd steps.

In the event that multiple rules are simultaneously applicable – such as might happen, for example, if the rules include more than one division, or fusion, for the same lefthandside argument in their expressions (in Eq. (45)) – the order in which the rules are applied is determined by an a priori priority ranking. Also, since applying either *commutation* or *annihilation* rules to adjacent links yields inconsistency, whenever a local context arises in which this might happen, the application of these rules is temporarily suppressed. (This is done by sweeping through the link set *twice*: on the first pass, a temporary flag is set for each link that satisfies a rule condition; on the second pass, the link rule is applied *if and only if* the four neighboring links did not raise flags during the first pass.)

Figure 18 shows the first few steps in applying one *division* and two *commutation* rules to a simple initial graph. (Kohji Tomita provides several movies of GRA evolutions on his website: http://staff.aist.go.jp/k.tomita/ga/)

Tomita, Kurokawa, and Murata [67,68,69,70,71] report a variety of emergent behaviors, including (1) *arbitrary resolution* (because GRA rules effectively allow an arbitrary number of sites to "grow" out of any initial structure, these systems define their own "boundary conditions" and graphs with arbitrary resolution are possible); (2) *repetitive structures*, in which some geometrical subset of an initial graph is reproduced, indefinitely, and continu-

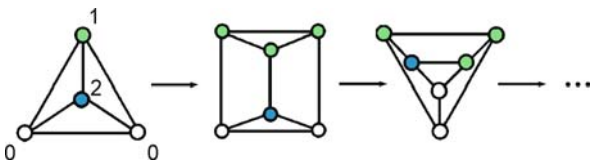

Structurally Dynamic Cellular Automata, Figure 18
Sample GRA evolution starting from the graph on the left. The rules are (see Eq. (45)): (1) *division*(1, 0, 0, 2) → (1, 1, 1, 0, 0, 2), (2) *commutation*(1, 2) → (1, 2), and (3) *commutation*(0, 0) → (0, 0). (Reproduced from [69] with permission)

ously grafted onto the original structure; and (3) *self-replication*, in which both site-value and structure is replicated after N steps. In [70], Tomita et al. describe how genetic algorithms [45] may be used for automating the search for self-replicating patterns.

In [67], Tomita et al. also present the design of a self-reproducing *Turing Machine*. *Turing machines* are abstract symbol manipulating devices that mimic the basic operations of a computer. Formally, they consist of a "tape" (of indefinite length, to record data), a "head" (that reads/writes symbols on the tape, and that can move left or right), and "state transition rules" (that tell the head which new symbols to write given the current state of the tape). The tape is analogous to "memory" in a modern computer; the head is analogous to the microprocessor. A Turing machine is called "universal" if it can simulate any other Turing machine.

Tomita et al.'s [67] *Turing machine* is modeled as a ladder structure: the upper sites constitute the "tape" mechanism; the lower sites form the "tape head" that reads the tape; both ends of the ladder are single sites that define "end of tape"; and the two ends are joined to form a loop. Although the tape is initially finite, the ladder can grow to arbitrary length, as required, by using appropriate GRA rules. Tomita et al. [67] self-replicating Turing GRA consists of 20 states and 257 (2-symbol) rules. They also introduce a design for a universal Turing machine [69] that consists of 30 states and 955 rules for reproduction, and 23 states and 745 rules for computation. While self-reproducing universal Turing machines can be described using conventional CA, their expression using GRA rules are considerably more compact.

Dynamic Graphs as Models of Self-Reconfigurable Robots

In the context of looking for self-reconfiguration algorithms that may be used to manufacture modular robots for industry, Saidani [63,64] has recently introduced a *dynamic graph calculus* that includes rules similar to those that define SDCA; but which depend only on the *topology* of (but not the σ-values living on) the lattice. Saidani and Piel [65] have also introduced an interactive programming environment for studying dynamic graph simulations called *Dynagraph*, and implemented in Smalltalk.

There are two basic approaches to designing modular robots: (1) to develop a set of elementary generic modules that can be rapidly assembled by humans to form robots that solve a specific problem, and (2) to design a set of (otherwise identical) primitive components that can adaptively reconfigure themselves. Focusing on the latter approach, Saidani [64] formally reinterprets modular "robots" to mean modular *networks*; and proceeds to model adaptive robotic self-reconfigurations as a class of recursive graph dynamical systems. In contrast to other related dynamic graph models [18,25], the "modules" (or subgraphs) of Saidani's model use local knowledge of their neighborhood topology to collectively evolve to some goal configuration. Although the dynamics transforms the global state, the evolution remains strictly decentralized, and individual modules do not know the (desired) final state.

Apart from restricting the dynamics to topology *alone* (indeed, none of the sites harbor information states of any kind), Saidani [63,64,65] further assumes that (1) connections between sites are *directional* (both *to*- and *from*-links may coexist between the same two modular components); (2) "active" sites reconfigure their local neighborhood by accepting, keeping, or removing their adjacent links according to rules that are functions of their current topology (defined as a given sites' current local neighborhood and the current neighborhood of its neighbors: a site only knows about its own *in*- and *out*-degree, which can obviously be computed from its local topology, and the *in*- and *out*-degrees of its nearest neighbors); (3) a site controls its *outgoing* links (and can connect or disconnect any outgoing links), but cannot sever incoming connections; (4) sites must maintain at least one link throughout an evolution (so that the graph remains connected); and (5) all sites are equipped with the same set of rules.

As in conventional CA and the basic SDCA model, the "reconfiguration" proceeds synchronously throughout the graph. The decision process includes an innate stochastic element: in the event that there is a rule that specifies that a site is to establish a link to a neighbor of one of its neighbors, but all neighboring sites have the same degree (which is the only dynamical discriminant), the neighbor with which a new link will be forged is selected at random.

As a concrete example, Saidani [64] presents a *tree-to-chain* algorithm that evolves an initial "tree" graph to a linear chain of linked sites (see Fig. 19). While we do not reproduce the full algorithm here, it is essentially a case-driven list of rules of the form *if condition* C_1 (*and condition* $C_2, \ldots$ *and condition* C_n) *then connect (or disconnect) site* i *to (from) the* n*th neighbor of* i*'s neighbor*, v. For example, an explicit "rule" might be: *if* $1 \leq \deg^-(i) \leq 2$ *and* $\deg^+(i) = 1$ *and* $|\tau(i)| = 2$ *then link* i *to a neighboring site* j *that has* $\deg^-(j) = 0$, where $\deg^-(i)$ and $\deg^+(i)$ are the *in*- and *out*-degrees of site i, and $\tau(i)$ is the total number of sites to which i is currently linked (with either incoming or outgoing links).

Conceptually, the details of Saidani's rules are less important than what the unfolding process represents as

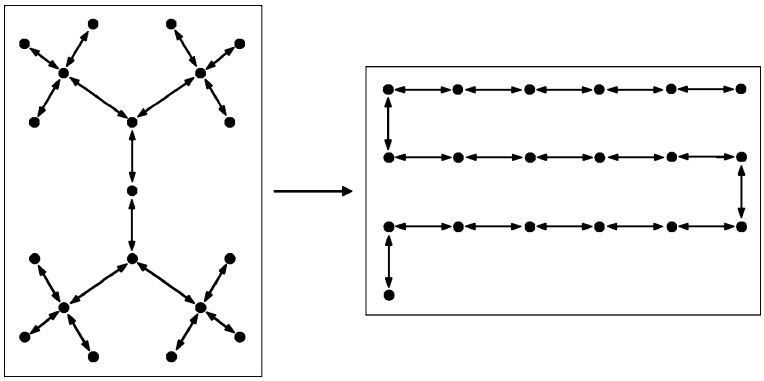

Structurally Dynamic Cellular Automata, Figure 19
Schematic illustration of a *tree* topology reconfiguring itself into a linear *chain* using a set of case-based *"if–then"* topology rules defined in [64]; see text for details

a whole. An initial graph – which we recall is to be viewed as a distillation of a "modular robot" – is transformed, by the individual sites (or parts of the robot), into another *desired* structure; i. e., the graph is entirely *self-reconfigured*. Though the broader reverse-engineering problem (which includes asking such fundamental questions as *"How can a desired final state be mapped onto a specific cased-based list of graphical rules?"*) remains, as yet, unanswered, and the *Dynagraph* work environment [65] is currently limited to experimenting only with graphs that have less than 30 sites, the basic model already represents a viable new approach to using dynamic graphs to describe self-reconfigurable robots; and is potentially more far-reaching as a general model of topologically-reconfigurable dynamical systems.

SDCA as Models of Fundamental Physics

Pregeometric Theories of Emergent Space-Time

Although SDCA are a natural formal extension of conventional CA – and serve as general-purpose modeling tools – their conception was originally motivated by fundamental physics; specifically, by a search for models of self-organized emergent discrete space-time [42]. *"Space acts on matter, telling it how to move; … matter reacts back on space, telling it how to curve"*, which is the central lesson of Einstein's *geometrodynamics*, as explained by Misner, Thorne and Wheeler in their classic text on Gravitation [44]. Wheeler [73] has been a particularly eloquent spokesman for the need to search for what he calls a *pregeometry*, or a set of basic elements out of which what we normally think of geometry is built, but which are themselves devoid of a specific dimensionality: *"Space-time … often considered to be the ultimate continuum of physics, evidences nowhere more clearly than at big bang and at collapse that it cannot be a continuum. Obliterated in those events is not only matter, but the space and time that envelope that matter … we are led to ask out of what 'pregeometry' the geometry of space and spacetime are built"*. Wheeler has also proposed the idea that particles be viewed as geometric disturbances of spacetime, called *geometrodynamic excitons*.

A priori, SDCA appear tailor-made for describing pregeometric theories of space-time. Since in SDCA, lattice and local σ-values are explicitly coupled, and geometry and value configurations are treated on an approximately equal footing, SDCA is certainly at least formally consistent with Einstein's geometrodynamic credo. The structure is altered locally as a function of individual site neighborhood value-states and geometries, while local site-connectivity supports the site-value evolution in exactly the same way as in conventional CA models defined on random lattices. The microphysical view of physics that emerges from this construction is one in which a fundamentally discrete pregeometry continually evolves in time as an amorphous structure but with a globally well-defined dimensionality. Particles are constructs of that amorphous structure and can be viewed as locally persistent substructures – i. e. geometrical or topological solitons – with dimensions that differ from the surrounding value. Just as "value structure" solitons are ubiquitous in conventional CA models [32,74], "link structure" solitons might emerge in SDCA; physical particles would, in such a scheme, be viewed as geometrodynamic disturbances propagating within a dynamic lattice.

Of course, speculation regarding the ultimate constituents of matter and space-time date back at least as far as 500 BC when the philosopher Democritus mused on the idea that matter is made of indivisible units separated by void. Since then there have been countless attempts, with

varying degrees of success, to fashion an entirely discrete theory of nature. We limit our discussion to a short survey of some recent work that centers on ideas that are either direct outgrowths of, or are otherwise conceptually related to, SDCA models. (A short history of pregeometric theories appears in chapter twelve of Ilachinski [32]).

One of the earliest proponents of pregeometry is Zuse [76], who speculated on what it would take for a CA-like universe to sustain "digital particles" on a cellular lattice. He focused on two main problems: (1) *How does the universe's observed isotropy arise from a CA's (Euclidean, hexagonal, etc.) anisotropy?*, and (2) *What is the information content of a physical particle?* As an answer to the first question, Zuse suggests …

" … variable and growing automata. Irregularities of the grid structure are a function of moving patterns, which is represented by digital particles. Now, not only certain values are assigned to the single crosspoints of the grid in the concept of the cellular automaton which are interrelated and sequencing each other, but also the irregularities of the grid are itself functions of these values of the just existing interlinking network. One can imagine rather easily that in such a way the interdependence of mass, energy, and curvature of space may logically result from the behavior of the grid structure."

Jourjine [33] generalizes Euclidean lattice field theory on a d-dimensional lattice to a cell complex. Using homology theory to replace points by cells of various dimensions and fields by functions on cells, he develops a formalism that treats space-time as a dynamical variable and describes the change in the dimension of space-time as a phase transition.

Kaplunovsky and Weinstein [34] develop a field-theoretic formalism that treats the topology and dimension of the spacetime continuum as dynamically generated variables. Dimensionality is introduced out of the characteristic behavior of the energy spectrum of a system of a large number of coupled oscillators.

Dadic and Pisk [13] introduce a self-generating discrete-space model that is based on the local quantum-mechanics of graphs. Just as in SDCA, Dadic and Pisk's spatial structure is discrete but not static; it is fundamentally amorphous and evolves in time. Though the metric is essentially the same one used to define SDCA (i. e., D_{effec}), it is generalized to unlabeled graphs by referring to the topological description of the node positions rather than their arbitrary labels. Though their "graph dynamics" differs from what is used by SDCA (and uses a symmetrized Fock space that is local in terms of their graph metric, where "Fock space" is a Hilbert space used to describe quantum states with a variable, or unspecified, number of particles, and is made from the direct sum of tensor products of single-particle; or, in this case, single-*graph*, Hilbert spaces) it shares two important properties with SDCA: (1) interactions depend only on the local properties of the graph, and (2) interactions induce only minimal changes to the local metric function. An important consequence of their theory is that the dimension of a graph is a scale dependent quantity that is generated by the dynamics.

Combinatorial Space-Time Hillman [27] introduces a *combinatorial space-time*, which he defines as a class of dynamical systems in which finite pieces of spacetime contain finite amounts of information. Spacetime is modeled as a combinatorial object, constructed by dynamically coupling copies of finitely many types of certain allowed neighborhoods. There is no *a priori* metric, and no concept of continuity, which is expected to emerge on the macroscale.

The construction (and evolution) of spaces proceeds in three steps: (1) define a set X of combinatorial n-dimensional spaces (examples are conventional CA graphs, graphs with directional links, or some other kind of embedded symmetry); (2) define a set of local, invertible *primitive* maps $T\colon X \leftrightarrow Y$ between pairs of space sets, such that the maps do not all commute with one another (for example, a simple renaming of the sites or links gives an invertible, local map); (3) generate an arbitrary set of local invertible graph transformations by composing primitive maps with one another. Since the primitive maps are deliberately chosen so that they do not all commute, the act of composition yields infinitely many nontrivial transformations. The *orbits* $\{T^z(x) \mid z \in Z\}$ (for each space x in X) are $(n+1)$-dimensional combinatorial spacetimes; which include reversible CA and SDCA-like networks in which geometry evolves locally over time. Formally, Hillman uses matrices of nonnegative integers, directed graphs, and symmetric tensors to describe these systems, so that local equivalences between space sets are generated by simple matrix transformations. Concrete examples of dynamic combinatorial space-time graphs are given in [27].

Structurally Dynamic Disordered Cellular Networks As an explicit example of how dynamic graphs can be used to model pregeometry, consider *structurally dynamic disordered cellular networks* (abbreviated, SDDCN), recently introduced by Nowotny and Requardt [53,54,55,58,59,60]. SDDCN are a class of models closely related to SDCA but developed explicitly to describe a discrete, dynamic space-time fundamental physics. The main difference between the two models is that whereas link connections in SDCA

are strictly local, SDDCN are capable of generating both local and *translocal* links.

In contrast to more mainstream high-energy theories of fundamental physics (which are dominated by string theory and/or loop quantum gravity, both of which assume a certain level discretization at the Planck scale, but assume that a discrete space-time emerges from an underlying *continuum* physics), SDDCN takes a *bottom-up* approach. SDDCN assumes that there is underlying dynamic, discrete and highly erratic network substratum that consists of (on a given scale) irreducible mutually interacting agents exchanging information via primordial channels (links). The known continuum structures are expected to emerge on a macroscopic (or, mesoscopic) scale, via a sequence of coarse graining and/or renormalization steps.

Like SDCA, SDDCN are defined on arbitrary graphs, G, initially defined by a specified set of sites and links. Both sites and links are allowed to take on values. Site values, σ_i (which represent a primitive "charge"), are taken from some discrete set, $q \cdot \mathcal{Z}$, where q is a discrete *quantum of information*; link states assume the values $J_{ij} \in \{-1, 0, +1\}$, and represent an elementary coupling. The J_{ij} are equivalent to SDCA's l_{ij}, but take on *three* values rather than two. Heuristically, J_{ij} represent directed edges pointing either from site i to j (if $J_{ij} = 1$), or from j to i (if $J_{ij} = -1$); or, in the case of $J_{ij} = 0$, the absence of a link. At each time step (representing an elementary quantum of time), an elementary quantum q is transported along each existing directed link in the indicated direction. As for SDCA, SDDCN dynamically couples site values to links.

Nowotny and Requardt [53] introduce two network models: one in which connected sites that have very different internal states typically lead to large local fluctuations (= SDDCN$_1$), and another in which sites with similar internal states are connected (= SDDCN$_2$):

$$
\begin{aligned}
&SDDCN_1\\
&\leftrightarrow \begin{cases} \sigma_i^{t+1} = \sigma_i^t + \sum_j J_{ji}^t, & \\ J_{ij}^{t+1} = \operatorname{sign}\left(\Delta\sigma_{ij}\right) & \text{for} \begin{cases} \left|\Delta\sigma_{ij}\right| \geq \lambda_2, \text{ or} \\ \left|\Delta\sigma_{ij}\right| \geq \lambda_1 \wedge J_{ij}^t \neq 0, \end{cases} \\ J_{ij}^{t+1} = 0, & \text{otherwise}, \end{cases}\\
&SDDCN_2\\
&\leftrightarrow \begin{cases} \sigma_i^{t+1} = \sigma_i^t + \sum_j J_{ji}^t, & \\ J_{ij}^{t+1} = \operatorname{sign}\left(\Delta\sigma_{ij}\right) & \text{for} \begin{cases} 0 < \left|\Delta\sigma_{ij}\right| < \lambda_1, \text{ or} \\ 0 < \left|\Delta\sigma_{ij}\right| < \lambda_2 \wedge J_{ij}^t \neq 0, \end{cases} \\ J_{ij}^{t+1} = J_{ij}^t, & \text{for } \Delta\sigma_{ij} = 0, \\ J_{ij}^{t+1} = 0, & \text{otherwise}, \end{cases}
\end{aligned}
\tag{46}
$$

where $\Delta\sigma_{ij} = \sigma_i^t - \sigma_j^t$, and $\lambda_2 \geq \lambda_1 \geq 0$. Since SDDCN is intended to model pregeometric dynamics, Nowotny and Requardt [53] caution that the t parameter that appears in these equations must not to be confused with the "true time" that (they expect) emerges on coarser scales. In keeping with its physics-based motivation, SDDCN's dynamical laws depend only on the relative differences in site values, not on their absolute values. Indeed, charge is nowhere either created or destroyed, so that SDDCN conserves global "charge": $\sum_i \sigma_i^t =$ constant, where the arbitrary constant can be set to *zero*.

Both models start out initially on a simplex graph with $N \sim 200$ nodes, so that the maximum number of possible links is $N(N-1)/2$. The initial σ-seed consists of a uniform random distribution of values scattered over the interval $\{-k, -k+1, \ldots, k-1, k\}$, where $k \sim 100$. The initial values for link states, $J_{ij}^{t=0}$, are selected from $\{-1, 1\}$ with equal probability; i. e., the initial state is a maximally entangled nucleus of nodes and links. Nowotny and Requardt [55] state that "*... in a sense, this is a scenario which tries to imitate the big bang scenario. The hope is, that from this nucleus some large-scale patterns may ultimately emerge for large clock-time*". For most properties (other than the $\langle\sigma\rangle_t$ and $\sum_i |\sigma_i^{t+1} - \sigma_i^t|$, which are both equal to *zero* by construction), the average over the width of the initial vertex state distribution, taken over λ_1 and λ_2, specific realizations of initial conditions, and time, depend linearly on network size.

We summarize Nowotny's and Requardt's [53,55] findings, culled from extensive numerical experiments: (1) the appearance of very short limit cycles in SDDCN$_1$ (period 6 and multiples of 6, with the longest having period 36 on a network of size $N = 800$), (2) Much longer limit cycles and transients in SDDCN$_2$, both of which appear to grow approximately exponentially, (3) structurally, SDDCN$_1$ evolve from almost fully connected simplex networks to more sparse connectivities with increasing $\lambda_{1/2}$; there is a regime in which few vertices with very high degree coexist with many vertices with a low degree; for large (around $\lambda_1 \approx 60$; for large $\lambda_{1/2}$, the graph eventually breaks apart and all nodes become isolated; (4) for SDDCN$_2$, nodes typically have zero degree small $\lambda_{1/2}$, and links become increasingly dense as $\lambda_{1/2}$ increase; the degree distribution is generally broad and remains so for large $\lambda_{1/2}$ (the authors also note observing multiple local maxima of the distributions in a wide range of $\lambda_{1/2}$ values); (5) for SDDCN$_1$, there is an abrupt phase-transition in the temporal fluctuations of vertex degrees (defined as $\deg_i(t+1) - \deg_i(t)$) from a state in which there are essentially no fluctuations ("frozen network") to one with strong fluctuations ("liquid network"); (6) the distribu-

tion of site values is strongly bimodal for $62 \leq \lambda_1 \leq 85$ for SDDCN_1 (while SDDCN_1 distributions are not bimodal, the width of the site value distributions for different values of λ_1 appears modulated.

From a fundamental physics perspective, the most interesting class of behaviors of SDDCN involves emergent dimensionality. Nowotny and Requardt [55] argue that since the continuum is a self-organized dynamic structure that emerges in the limit of large N and t, the most useful measure of "dimension" cannot be purely local (as in the case of *effective dimensionality*, D_{effec}, used for describing SDCA systems). Rather, it must be an intrinsically *global property*; one that is independent of any arbitrary embedding dimension, and one that can take on relatively stable values in the *whole* (to characterize effective system-wide characteristics), while simultaneously being relatively impervious to otherwise rapidly changing structural changing taking place on the microscale. Toward this end, Nowotny and Requardt [53] define the upper (and lower) scaling dimensions, $D_S^U(i)$ (and $D_S^L(i)$), with respect to site i:

$$D_S^U(i) = \limsup_{r\to\infty} \frac{\ln \beta(i,r)}{\ln r}\,, \qquad D_S^L(i) = \liminf_{r\to\infty} \frac{\ln \beta(i,r)}{\ln r}\,, \tag{47}$$

and the upper (and lower) connectivity dimensions, $D_C^U(i)$ (and $D_C^L(i)$), with respect to site i:

$$D_C^U(i) = \limsup_{r\to\infty} \frac{\ln \partial\beta(i,r)}{\ln r}\,, \qquad D_C^L(i) = \liminf_{r\to\infty} \frac{\ln \partial\beta(i,r)}{\ln r}\,, \tag{48}$$

where $\beta(i,r) = \#\,\text{sites}\, j \mid D_{ij} \leq r$, and $\partial\beta(i,r)$ is the number of sites on the *surface* of the r-sphere. When the upper and lower limits coincide, we have the *scaling dimension* ($= D_S$) and the *connectivity dimension* ($= D_C$), respectively. D_S is related to well known dimensional concepts in fractal geometry; D_C is a more physical measure that describes how the graph is connected, and thus how sites may potentially influence one another [55]. Preliminary research [53] suggests that under certain conditions, behavior resembling a structural phase transition to states with stable internal (and/or connectivity) dimensions is possible.

Network Automata

Stephen Wolfram devotes chapter nine of his Opus – *A new kind of science* (abbreviated, NKS) [75] – to applying CA to fundamental physics; and speculates on ways in which space may be described using a dynamic network. The central, overarching theme of NKS is that "simple" programs often suffice to capture complex behaviors.

The bold claim made in chapter nine of NKS is that, on an even more fundamental level, what underlies *all the laws of physics*, as we currently understand them, is a simple CA-like program, from which, ultimately, all the phenomenologically observed complexity in the universe naturally emerges. As for the specific forms such a "program" may take, Wolfram's intellectual point of departure echoes that of other proponents of a discrete dynamic pregeometric theory:

"*… cellular automata … cells are always arranged in a rigid array in space. I strongly suspect that in the underlying rule for our universe there will be no such built-in structure. Rather … my guess is that at the lowest level there will just be certain patterns of connectivity that tend to exist, and that space as we know it will then emerge from these patterns as a kind of large-scale limit*".

Wolfram introduces his *network automata* (abbreviated, NA) with these basic assumptions (see additional notes in NKS [75] on the evolution of networks: pp. 1037–1040): (1) features of our universe emerge solely from properties of space, (2) the underlying model (and/or "rules") must contain only a minimal underlying geometric structure, (3) the individual sites of emergent graphs must not be assigned any intrinsic position, (4) sites are limited to possessing purely *topological* information (that defines the set of sites to which a given site is connected), (5) incoming and outgoing connections need not be distinguished, and (6) all sites have exactly the same total number of links to other sites (which is assumed equal to *three*). This last assumption – which is essentially the same one made by Nowotny and Requardt [53] as the basis of their SDDCN model; see Subsect. "Structurally Dynamic Disordered Cellular Networks" above – does not lead to any loss of generality. With two connections, only very trivial graphs are possible; and it is easy to show that any site with more than three links can always be redefined, locally, as a collection of sites with exactly three links each (see Fig. 20).

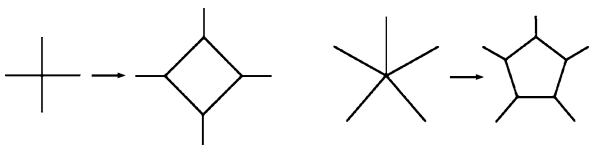

Structurally Dynamic Cellular Automata, Figure 20
Illustration of how sites that have more than three links can always be redefined as a set of sites with exactly three links each

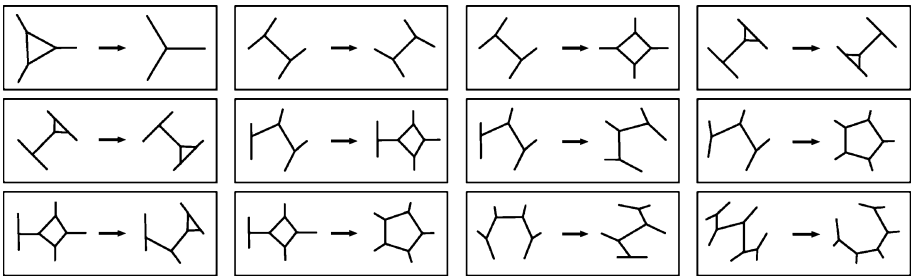

Structurally Dynamic Cellular Automata, Figure 21
Examples of planarity-preserving network substitution rules. (Reproduced from [75] with permission)

Wolfram [75] gives several concrete examples of evolving graphs (as models of pregeometry), the dynamics of which are prescribed by a set of *substitution* rules rules; i. e., explicit lists of the topological configurations (of sites and links) that are used to replace (at time $t+1$) specific local configurations (as they appear at time t). However, in contrast to SDCA rules, Wolfram's substitution rules are strictly *topological*; no site-value information is used. Also, the number of sites in the graph can change as the graph evolves; where, in SDCA, the number remains constant.

Figure 21 shows examples of rules in which specific clusters of sites are replaced with other clusters of sites. While the rules shown in the figure share the property that they all preserve planarity, there is no particular reason for imposing such a restriction; in fact, rules that generate non-planarity are just as easy to define. Wolfram speculates (pp. 526–530 in [75]) that "particle states" may be defined as mobile *non*-planar subgraphs that persist on an otherwise planar, but *randomly* fluctuating topology. Reversible versions of these rules may also be constructed, by associating a "backward" version with each "forward" transformation.

Some care must be taken while both *defining* and *applying* these rules consistently. For example, if a cluster of sites contains a certain number of links at t, one is not permitted to define a rule that replaces that cluster with another one that has a different number of connections. Another restriction is that rules must be independent of orientation; that is, if a candidate rule requires identifying the specific links (of, say, an otherwise topologically symmetric n-link local subgraph) before activating a desired substitution, that rule is likewise forbidden. However, even with these restrictions, a large number of rules are still possible. For example, 419 distinct rules may be defined for clusters with no more than five sites.

In applying network rules, one cannot simply simultaneously replace all pertinent subgraphs with their replacements, since, in general, two or more subgraphs with the same topology may overlap somewhere within the network. Since there is no *priori*, or universally consistent, way of ordering the subgraphs, *meta*-rules must be imposed to eliminate any possible ambiguities. For example, one method (*m1*) is to restrict replacements to a single subgraph per time step, selecting the subgraph whose replacement entails the minimal change to all recently updated sites. Another method (*m2*) is to allow all possible nonoverlapping replacements, while ignoring those that overlap. Wolfram reports that, although the second method obviously produces larger graphs in fewer steps, the two methods generally produce qualitatively similar structures.

Figure 22 traces the first few steps in the evolution of a simple graph under the action of a single substitution rule (defined at the center of the figure). Figures 22a and 22b show the results of applying this rule using methods *m1* and *m2*, respectively. In each case, the top row shows the form of the network before the substitution takes place at that step, and the bottom row shows the network that results from the substitution. The subgraph (or subgraphs, in Fig. 22a) involved in the replacement is highlighted at both top and bottom.

Wolfram also suggests that analogs of *mobile automata* [43] can be defined for evolving networks. By tagging a site i, say, with a "charge", $\sigma_i \equiv 1$, substitution rules may be defined to replace clusters of sites around the charged site. The effect is that the charge itself appears to move, as its effective (relative) position within the network changes as the geometric dynamics unfolds. (However, Wolfram also notes – on page 1040 in [75] – that *"despite looking at several hundred cases I have not been able to find network mobile automata with especially complicated behavior"*).

Future Directions and Speculations

Although SDCA were first introduced over two decades ago [28], much of their behavior remains unexplored. Of course, this is due largely to the difficulty of studying dy-

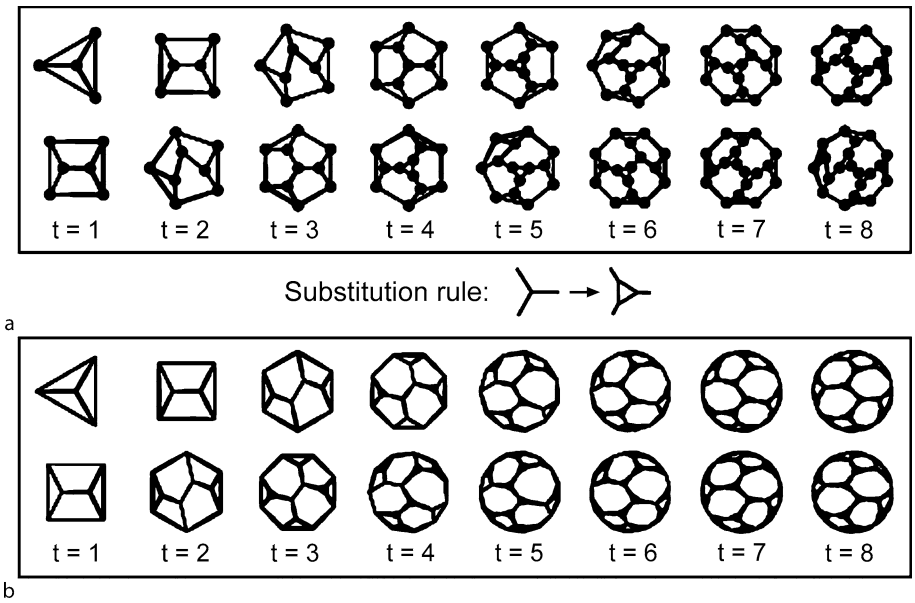

Structurally Dynamic Cellular Automata, Figure 22
Examples of network evolutions using the substitution rule shown at center. See text for explanation. (Reproduced from [75] with permission)

namical systems that harbor an a priori vastly larger coupled value-geometry space than the "merely" spatially-confined behavioral space of conventional CA. Only relatively recently have desktop computers become sufficiently powerful, and visualization programs adept enough at rendering multidimensional graphs [12], to make a serious study of SDCA behaviors possible. For example, the general-purpose math programs *Mathematica* (http://www.wri.com) and *Maple* (http://www.maplesoft.com) both provide powerful built-in graph-rendering algorithms to help visualize complex graphs. Standalone public-domain packages are also available; for example, AGNA [2], NetDraw [11], and Pajek [52]. In this final section, we list several open questions and briefly speculate on possible future directions.

Because of the relative paucity of studies dedicated purely to exploring the space of emergent structures (such as Wolfram's [74] pioneering studies of conventional CA), many (even very fundamental) questions remain open: *What kinds of geometries can arise?, Which subspace of the space of all possible graphs corresponds to those that are actually attainable using SDCA (and SDCA-like) rules?, What are the conditions for which certain geometries do, and do not, form?, What combinations of σ- and ℓ-rules give rise to specific kinds of graphs?*

Other open problems include: (1) determining whether the (provisionally defined) set of class-4 rules, for which effective dimension appears to remain constant, is *genuine*, rather than being either a long-term transient or an unintentional artifact of imposed run-time constraints; and, if this class is "real", we obviously need to ask, *How large is it?*, and *Under what conditions does it arise?*; (2) developing SDCA as formal mathematical models, perhaps as members of a broader class of graph grammars [20,35]; and (3) finding purely geometric analogs of the solitons known to exist in conventional CA models [32,74].

This article has introduced several generalizations of the basic SDCA model, including *memory* effects (Subsect. "SDCA With Memory"), *reversibility* (Subsect. "Reversible SDCA"), *probabilistic transitions* (Subsect. "Probabilistic SDCA"), and a class of SDCA-like dynamical systems that evolve according to rules that depend only on *topology* (Subsects. "Dynamic Graphs as Models of Self-Reconfigurable Robots" and "Network Automata"). However, other possibilities abound: (1) σ site-variables may take on a larger range of values, $\sigma \in \{0, 1, \ldots, k-1\}$; (2) link variables, ℓ_{ij}, may similarly take on a larger range of values, $\ell_{ij} \in \{0, \pm 1, \pm 2, \ldots, \pm m\}$ (where, say, $\pm$ determines "directionality", and absolute value, $|\ell_{ij}|$, represents either channel capacity for information flow or some other innate property); and (3) both sites and links may take on richer, and more explicitly "active", roles of *agent-actors* [17].

Apart from these formal extensions, some obvious future applications include modeling communication and social network dynamics, studying the dynamics of plas-

ticity in artificial neural networks, designing adaptive self-reconfiguring parallel-computer networks (as well as "amorphous" computer chips), studying behaviors of gene-regulatory networks, and providing the conceptual core for fundamental pregeometric physical theories of discrete, emergent space-times.

Bibliography

Primary Literature

1. Adamatzky A (1995) Identification of Cellular Automata. Taylor and Francis, London
2. Applied Graph & Network Analysis software. http://benta.addr.com/agna/. Accessed 14 Oct 2008
3. Albert J, Culik II K (1987) A simple universal cellular automaton and its one-way and totalistic version. Complex Syst 1:1–16
4. Ali SM, Zimmer RM (1995) Games of Proto-Life in Masked Cellular Automata. Complexity International, vol 2. http://www.complexity.org.au
5. Ali SM, Zimmer RM (2000) A Formal Framework for Emergent Panpsychism. In: Tucson 2000: Consciousness Research Abstracts. http://www.consciousness.arizona.edu/tucson2000/. Accessed 14 Oct 2008
6. Alonso-Sanz R, Martin M (2006) A structurally dynamic cellular automaton with memory in the hexagonal tessellation. In: El Yacoubi S, Chopard B, Bandini S (eds) Lecture Notes in Computer Science, vol 4173. Springer, New York, pp 30–40
7. Alonso-Sanz R (2006) The Beehive Cellular Automaton with Memory. J Cell Autom 1(3):195–211
8. Alonso-Sanz R (2007) A structurally dynamic cellular automaton with memory. Chaos, Solitons, Fractals 32(4):1285–1304
9. Barabasi AL, Albert R (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
10. Bollobas B (2002) Modern Graph Theory. Springer, New York
11. Borgatti SP (2002) NetDraw 1.0: Network visualization software. Analytic Technologies, Harvard
12. Chen C (2004) Graph Drawing Algorithms. In: Information Visualization. Springer, New York
13. Dadic I, Pisk K (1979) Dynamics of discrete-space structure. Int J Theor Phys 18:345–358
14. Doi H (1984) Graph theoretical analysis of cleavage pattern: graph developmental system and its application to cleavage pattern in ascidian egg. Dev Growth Differ 26(1):49–60
15. Durrett R (2006) Random Graph Dynamics. Cambridge University Press, New York
16. Erdos P, Renyi A (1960) On the evolution of random graphs. Publ Math Inst Hung Acad Scie 5:11–61
17. Ferber J (1999) Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence. Addison-Wesley, New York
18. Ferreira A (2002) On models and algorithms for dynamic communication networks: the case for evolving graphs. In: 4th Recontres Francophones sur les Aspects Algorithmiques des Télécommunications (ALGOTEL 2002), Meze, France
19. Gerstner W, Kistler WM (2002) Spiking Neuron Models. Single Neurons, Populations, Plasticity. Cambridge University Press, New York
20. Grzegorz R (1997) Handbook of Graph Grammars and Computing by Graph Transformation. World Scientific, Singapore
21. Halpern P (1989) Sticks and stones: a guide to structurally dynamic cellular automata. Amer J Phys 57(5):405–408
22. Halpern P, Caltagirone G (1990) Behavior of topological cellular automata. Complex Syst 4:623–651
23. Halpern P (1996) Genetic algorithms on structurally dynamic lattices. In: Toffo T, Biafore M, Leao J (eds) PhysComp96. New England Complex Systems Institute, Cambridge, pp 135–136
24. Halpern P (2003) Evolutionary Algorithms on a Self-Organized, Dynamic Lattice. In: Bar-Yam Y, Minai A (eds) Unifying Themes in Complex Systems, vol 2. Proceedings of the Second International Conference on Complex Systems. Westview Press, Cambridge
25. Harary F, Gupta G (1997) Dynamic graph models. Math Comp Model 25(7):79–87
26. Hasslacher B, Meyer D (1998) Modeling dynamical geometry with lattice gas automata. Int J Mod Phys C 9:1597
27. Hillman D (1995) Combinatorial Spacetimes. Ph D Dissertation, University of Pittsburg
28. Ilachinski A (1986) Topological life-games I. Preprint. State University of New York at Stony Brook
29. Ilachinski A, Halpern P (1987) Structurally dynamic cellular automata. Preprint. State University of New York at Stony Brook
30. Ilachinski A, Halpern P (1987) Structurally dynamic cellular automata. Complex Syst 1(3):503–527
31. Ilachinski A (1988) Computer Explorations of Self Organization in Discrete Complex Systems. Diss Abstr Int B 49(12):5349
32. Ilachinski A (2001) Cellular Automata: A Discrete Universe. World Scientific, Singapore
33. Jourjine AN (1985) Dimensional phase transitions: coupling of matter to the cell complex. Phys Rev D 31:1443
34. Kaplunovsky V, Weinstein M (1985) Space-time: arena or illusion? Phys Rev D 31:1879–1898
35. Kniemeyer O, Buck-Sorlin GH, Kurth W (2004) A Graph Grammar Approach to Artificial Life. Artif Life 10(4):413–431
36. Krivovichev SV (2004) Crystal structures and cellular automata. Acta Crystallogr A 60(3):257–262
37. Lehmann KA, Kaufmann M (2005) Evolutionary algorithms for the self-organized evolution of networks. In: Proceedings of the 2005 Conference on Genetic and Evolutionary Computation, Washington DC. ACM Press, New York
38. Love P, Bruce M, Meyer D (2004) Lattice gas simulations of dynamical geometry in one dimension. Phil Trans Royal Soc A: Math Phys Eng Sci 362(1821):1667–1675
39. Majercik S (1994) Structurally dynamic cellular automata. Master's Thesis, Department of Computer Science, University of Southern Maine
40. Makowiec D (2004) Cellular Automata with Majority Rule on Evolving Network. In: Lecture Notes in Computer Science, vol 3305. Springer, Berlin, pp 141–150
41. Mendes RV (2004) Tools for network dynamics. Int J Bifurc Chaos 15(4):1185–1213
42. Meschini D, Lehto M, Piilonen J (2005) Geometry, pregeometry and beyond. Stud Hist Phil Mod Phys 36:435–464
43. Miramontes O, Solé R, Goodwin B (1993) Collective behavior of random-activated mobile cellular automata. Physica D 63:145–160
44. Misner CW, Thorne KS, Wheeler JA (1973) Gravitation. W.H. Freeman, New York
45. Mitchell M (1998) An Introduction to Genetic Algorithms. MIT Press, Boston

46. Moore EF (1962) Sequential Machines: Selected Papers. Addison-Wesley, New York
47. Muhlenbein H (1991) Parallel genetic algorithm, population dynamics and combinatorial optimization. In: Schaffer H (ed) Third International Conference on Genetic Algorithms. Morgan Kauffman, San Francisco
48. Murata S, Tomita K, Kurokawa H (2002) System generation by graph automata. In: Ueda K (ed) Proceedings of the 4th International Workshop on Emergent Synthesis (IWES '02), Kobe University, Japan, pp 47–52
49. Mustafa S (1999) The Concept of Poiesis and Its Application in a Heideggerian Critique of Computationally Emergent Artificiality. Ph D Thesis, Brunel University, London
50. Newman M, Barabasi A, Watts DJ (2006) The Structure and Dynamics of Networks. Princeton University Press, New Jersey
51. Nochella J (2006) Cellular automata on networks. Talk given at the Wolfram Science Conference (NKS2006), Washington DC, USA, 16–18 June
52. Nooy W, Mrvar A, Batagelj V (2005) Exploratory Social Network Analysis with Pajek. Cambridge University Press, New York
53. Nowotny T, Requardt M (1998) Dimension Theory of Graphs and Networks. J Phys A 31:2447–2463
54. Nowotny T, Requardt M (1999) Pregeometric Concepts on Graphs and Cellular Networks as Possible Models of Space-Time at the Planck-Scale. Chaos, Solitons, Fractals 10:469–486
55. Nowotny T, Requardt M (2006) Emergent Properties in Structurally Dynamic Disordered Cellular Networks. arXiv:cond-mat/0611427. Accessed 14 Oct 2008
56. O'Sullivan D (2001) Graph-cellular automata: a generalized discrete urban and regional model. Environ Plan B: Plan Des 28(5):687–705
57. Prusinkiewicz P, Lindenmayer (1990) The Algorithmic Beauty of Plants. Springer, New York
58. Requardt M (1998) Cellular Networks as Models for Planck-Scale Physics. J Phys A 31:7997–8021
59. Requardt M (2003) A geometric renormalisation group in discrete quantum space-time. J Math Phys 44:5588–5615
60. Requardt M (2003) Scale free small world networks and the structure of quantum space-time. arXiv.org:gr-qc/0308089
61. Rose H (1993) Topologische Zellulaere Automaten. Master's Thesis, Humboldt University of Berlin
62. Rose H, Hempel H, Schimansky-Geier L (1994) Stochastic dynamics of catalytic CO oxidation on Pt(100). Physica A 206:421–440
63. Saidani S (2003) Topodynamique de Graphe. Les Journées Graphes, Réseaux et Modélisation. ESPCI, Paris
64. Saidani S (2004) Self-reconfigurable robots topodynamic. In: IEEE International Conference on Robotics and Automation, vol 3. IEEE Press, New York, pp 2883–2887
65. Saidani S, Piel M (2004) DynaGraph: a Smalltalk Environment for Self-Reconfigurable Robots Simulation. European Smalltalk User Group Conference. http://www.esug.org/
66. Schliecker G (1998) Binary random cellular structures. Phys Rev E 57:R1219–R1222
67. Tomita K, Kurokawa H, Murata S (2002) Graph Automata: Natural Expression of Self-Reproduction. Physica D 171(4):197–210
68. Tomita K, Kurokawa H, Murata S (2005) Self-description for construction and execution in graph rewriting automata. In: Lecture Notes in Computer Science, vol 3630. Springer, Berlin, pp 705–715
69. Tomita K, Kurokawa H, Murata S (2006) Two-state graph-rewriting automata. NKS 2006 Conference, Washington, DC
70. Tomita K, Kurokawa H, Murata S (2006) Automatic generation of self-replicating patterns in graph automata. Int J Bifurc Chaos 16(4):1011–1018
71. Tomita K, Kurokawa H, Murata S (2006) Self-Description for Construction and Computation on Graph-Rewriting Automata. Artif Life 13(4):383–396
72. Weinert K, Mehnen J, Rudolph G (2002) Dynamic Neighborhood Structures in Parallel Evolution Strategies. Complex Syst 13(3):227–244
73. Wheeler JA (1982) The computer and the universe. Int J Theor Phys 21:557
74. Wolfram S (1984) Universality and complexity in cellular automata. Physica D 10:1–35
75. Wolfram S (2002) A New Kind of Science. Wolfram Media, Champaign, pp 508–545
76. Zuse K (1982) The Computing Universe. Int J Theor Phys 21:589–600

Books and Reviews

Battista G, Eades P, Tamassia R, Tollis IG (1999) Graph Drawing: Algorithms for the Visualization of Graphs. Prentice Hall, New Jersey
Bornholdt S, Schuster HG (eds) (2003) Handbook of Graphs and Networks. Wiley-VCH
Breiger R, Carley K, Pattison P (2003) Dynamical Social Network Modeling and Analysis. The National Academy Press, Washington DC
Dogogovtsev SN, Mendes JF (2003) Evolution of Networks. Oxford University Press, New York
Durrett R (2006) Random Graph Dynamics. Cambridge University Press, New York
Gross JL, Yellen J (eds) (2004) Handbook of Graph Theory. CRC Press, Boca Raton

Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy

Benjamin A. Brooks[1], James H. Foster[1], Jeffrey J. McGuire[2], Mark Behn[2]
[1] School of Ocean and Earth Science and Technology, University of Hawaii, Honolulu, USA
[2] Department of Geology and Geophysics, Woods Hole Oceanographic Institution, Woods Hole, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Monitoring Motion:
Subaerial and Submarine Geodetic Methods

Data Analysis and Inversion
Discussion: Slow Earthquake and Submarine Landslide Process
Future Directions: Slow Earthquakes and Submarine Landslide Monitoring
Acknowledgments
Bibliography

Glossary

Submarine landslide A gravitational mass failure feature on the seafloor.

Slow earthquake A discrete slip event that produces millimeter to meter-scale displacements identical to those produced during earthquakes but without the associated seismic shaking.

GPS The Global Positioning System consists of a constellation of at least 24 medium earth orbiting satellites transmitting two or more microwave frequencies for use in precise positioning.

Seafloor geodesy The application of geodetic methods (studies of the change in the shape of the earth's surface) applied to a submarine environment.

Definition of the Subject

The term 'submarine landslide' encompasses a multitude of gravitational mass failure features at areal scales from square meters to thousands of square kilometers. Here, we concentrate on the large end of that spectrum, namely the submarine landslides that, when they move either in contained slip events or catastrophically, can generate surface displacements equivalent to $> M6$ earthquakes and/or hazardous tsunami.

The term 'slow earthquake' describes a discrete slip event that produces millimeter to meter-scale displacements identical to those produced during earthquakes but without the associated seismic shaking. Slow earthquakes, primarily associated with tectonic fault zones, have been recognized and studied with increasing frequency in the past decade largely due to the decreasing cost and proliferation of Global Positioning System (GPS) geodetic networks capable of detecting the ground motion [1,2,3]. Recently, one such GPS network on the south flank of Kilauea volcano, has recorded multiple slow earthquakes on the subaerial portion of a large landslide system that occurs primarily in the submarine environment [4,5,6]. Because the bathymetric charts surrounding the Hawaiian islands are littered with the remnants of massive, catastrophically emplaced submarine landslides (Fig. 1) it is natural to wonder if a slow-slipping submarine landslide is a precursory stage of one that will ultimately fail catastrophically.

We see two principal reasons why monitoring submarine landslides and slow earthquakes associated with them is important. First, because catastrophic failure of submarine landslides can cause tsunami they represent significant hazards to coastal zones. Understanding and monitoring how slow slip may lead to accelerated slip and catastrophic failure is, therefore, very important in terms of hazard mitigation. Second, submarine landslide systems can be some of the most active as well as spatially confined deforming areas on earth and so they represent excellent targets of study for furthering our understanding of the general fault failure process. For instance a pertinent question for which we do not yet have an answer is: are fault frictional properties homogeneous enough that the occurrence of slow earthquakes on a detachment fault plane underlying a landslide could relieve stress on the fault or do the slow earthquakes in one region load a neighboring seismogenic patch bringing it closer to a large sudden failure (i. e. an earthquake)?

While installation and operation of GPS networks on land is now relatively routine and somewhat inexpensive, the in situ monitoring of submarine landslide motion represents a significant technical challenge with accordingly higher costs. Submarine geodesy e. g. [7], however, is a nascent and rapidly evolving field with relative and absolute positioning techniques being intensely studied and developed. The near future is sure to see many advances in our monitoring and understanding of submarine landslides and slow earthquakes due to the application of submarine geodetic techniques.

Introduction

Submarine Landslides

The last 30 years have seen a dramatic increase in the recognition of submarine landslides world-wide, due largely to the increased prevalence and capability of swath and side-looking sonar mapping systems and systematic submarine mapping programs. For instance, the side-scan sonar mapping of the Hawaiian exclusive economic zone in the 1980s resulted in the discovery that massive submarine landslides are spatially distributed along the entire Hawaiian Ridge [8] (Fig. 1). These features are some of the largest landslides on the planet, with more than 70 attaining lengths greater than 20 km and some having lengths greater than 200 km and total volumes exceeding 5000 km^3. Since then, submarine landslides associated with other volcanic islands e. g. [9,10,11], mid-ocean ridges [12], and continental margins [13] have also

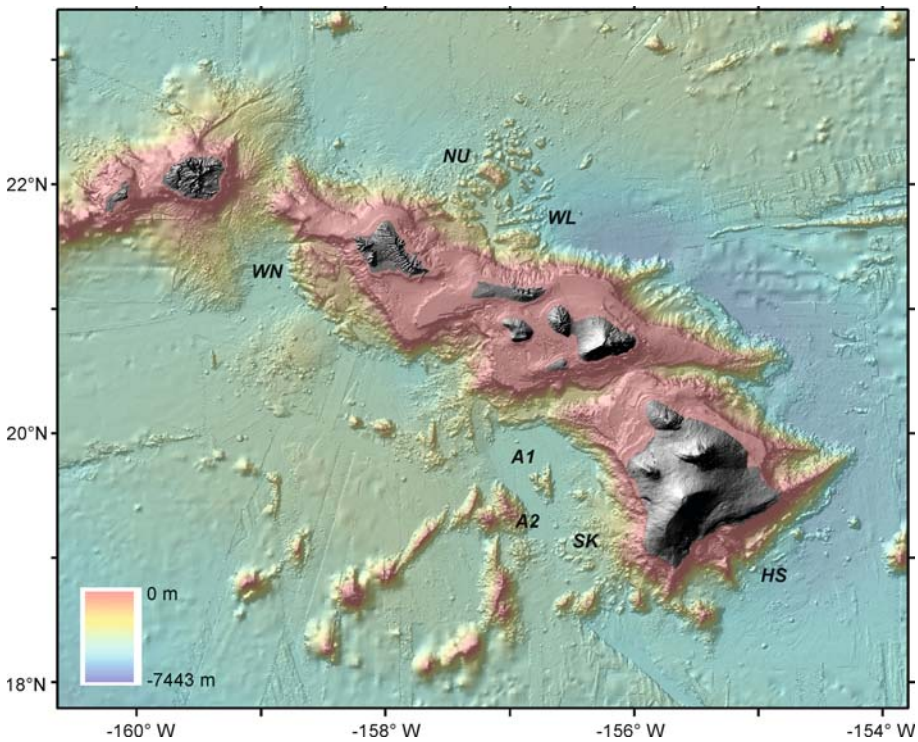

Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 1
Topographic and bathymetric map of the Hawaiian Islands (data from http://geopubs.wr.usgs.gov/i-map/i2809/). Studied submarine landslides are indicated: HS, Hilina Slump; SK, South Kona; A1, Alika 1; A2, Alika 2; WL, Wailau; NU, Nu'uanu; WN, Wai'anae

been studied. Of these, the slopes flanking volcanic islands, especially when they are in their steeper-sloped shield-building stage [14,15], tend to be particularly susceptible to landslide instability and so they have been the focus of much recent research in the Canary Islands [16,17] and especially in the Hawaiian Islands [18,19,20,21,22,23]. For instance, most of the Hawaiian submarine landslides are thought to be inactive, except for those on the flanks of the Big Island e. g. [21].

The morphology of submarine landslide features is similar to their subaerial counterparts. In map view they generally exhibit lobate and hummocky bathymetry. In cross-section, a wedge-shaped region of deformed material thins down-slope and is underlain by a gently-sloping planar dislocation surface (sometimes referred to as a basal detachment or 'decollement') separating the deformed carapace from the underlying undeformed substratum (Fig. 2). An upslope extensional head-scarp region transitions into a contractional fold belt towards the toe. The normal faults in the upslope region typically intersect the surface at high angles (> 45 degrees) and are manifested as scalloped-shaped scarps at the head of the slide that separate regions of differentially tilted fault blocks; in the sub-surface they may continue at high angles until they intersect the basal decollement, or they may sole with depth either into the decollement or into another sub-horizontal slip-surface [24]. The contractional regions are characterized by folded and bulging layers, closed depressions, and steep toes [20,21,25]. Moore et al. [8] separated the Hawaiian submarine landslides into two principal types: 'slumps' and 'debris avalanches'. The slumps are wide (up to ∼ 100 km), deep-seated (∼ 10 km thick), and have surface inclinations of up to 3 degrees while the debris avalanches are long (up to ∼ 230 km), shallowly-seated (50 m–2 km thick), and have surface slopes generally less than 3 degrees.

Concomitant with the mapping of the landslide features has been the increasing recognition that sudden submarine landslide movement can cause tsunami with destructive implications for coastal societies e. g. [26]. A particularly well-known example of this scenario is the 1929 Grand Banks failure [27]. Moreover, in the 1990s alone workers have attributed at least 5 tsunami events to catastrophic landslide failure sources: (1) 1992 Flores Island, Indonesia [28]; (2) 1994 Mindoro, Phillipines [29]; (3) 1998 Papua New Guinea e. g. [30]; (4) Kocaeli, Turkey [31]; (5) 1999 Pentecost Island, Vanuatu [32]. In Hawaii, the M_w7.7 1975 Kalapana earthquake, most likely

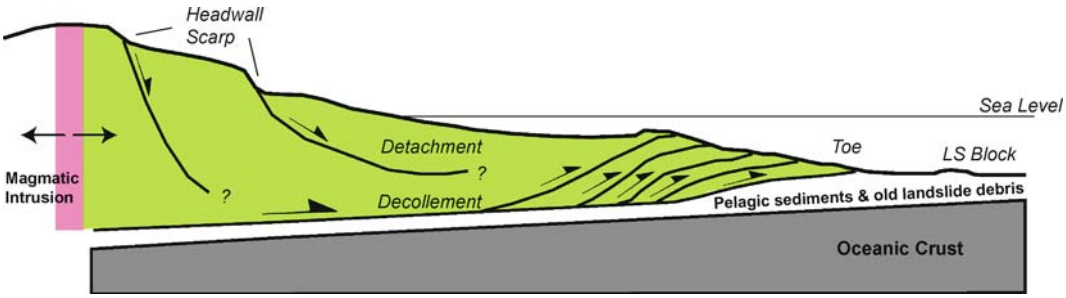

Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 2
Schematic cross-section of a submarine landslide flanking an active ocean-island volcano (after [21])

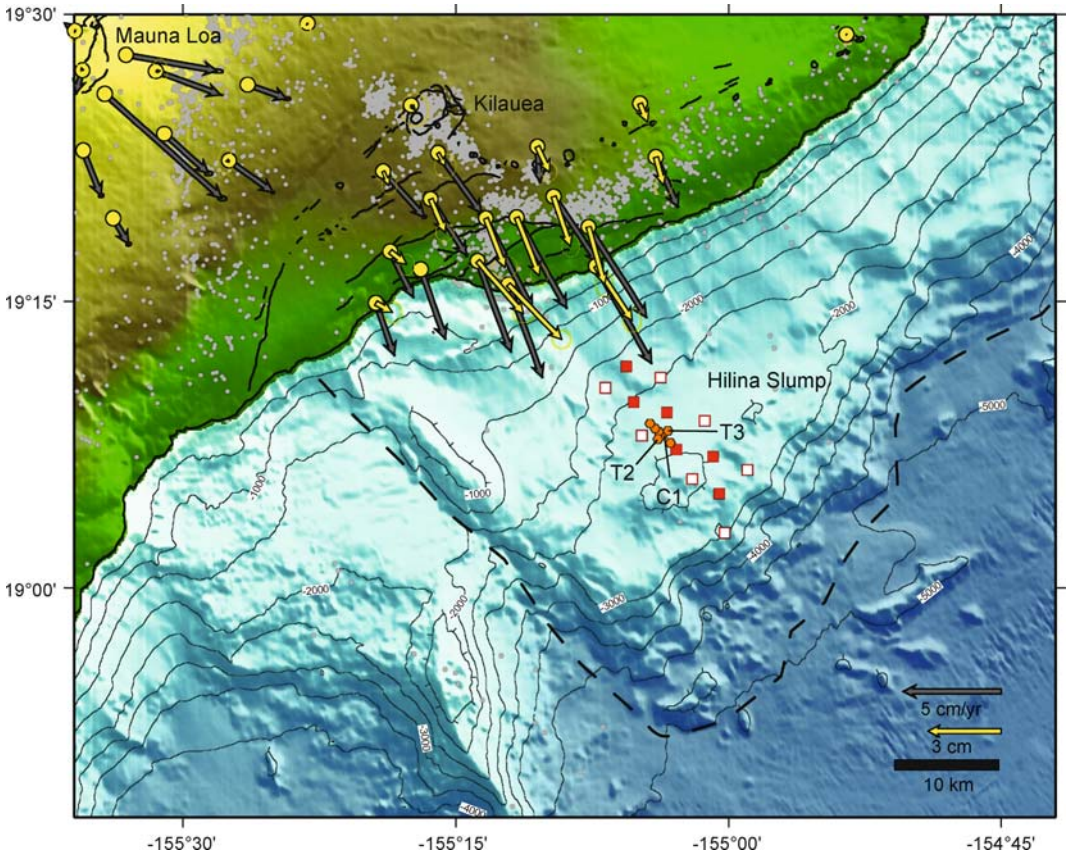

Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 3
Map of geodetic networks on the Hilina Slump overlain on topographic/bathymetric map. *Yellow circles* on land are CGPS sites operated jointly by the USGS Hawaiian Volcano Observatory, University of Hawaii, and Stanford University. *Red squares* offshore (*open and filled*) are seafloor geodetic sites operated by Scripps Institution of Oceanography. *Orange circles* offshore are acoustic extensometer sites deployed by our group with locations of transponders (T2,T3) and transceiver (C1) indicated. *Grey vectors* are average horizontal velocities from 1997–2005. *Yellow vectors* are horizontal motions from the January 2005 slow earthquake. *Grey dots* are earthquakes from the HVO catalog for the period May 2004–2005

due to slip of the fault surface underlying an active submarine landslide [33,34,35], caused local loss of life and damage in Southern California.

From a hazards standpoint it is particularly important to understand how tsunami are generated by submarine landslides because the propagation time between tsunami generation to runup is typically on the order of minutes. For instance, a catastrophic failure of the west side of the island of Hawaii would likely send tsunami waves around the Hawaiian islands that would reach the densely popu-

lated areas of Oahu's Waikiki beaches in less than an hour and more likely ∼ 30 min (G. Fryer, personal communication, 2007). Simulating waves generated by a sudden, chaotic disturbance of the seafloor, as expected from a submarine landslide, is quite complicated e. g. [36,37] and not all workers agree on approaches or results. Murty [38], however, stressed that parameters such as slide angle, water depth, density, speed, duration of the slide are second order, while instantaneously displaced volume is likely the most important parameter controlling tsunami generation.

Despite the increasing awareness of their hazard, little is known about how submarine landslides actually move, largely because of the challenge of installing instruments and retrieving data from the submarine environment. Moore et al. [8] recognized that while slumps more likely move relatively slowly, debris avalanches could be deposited very rapidly based on, for instance, uphill flow of material in the distal portions of the landslide deposits. In agreement with this, the estimated downhill velocities from the 1929 Grand Banks event was 60–100 km/h [39] whereas many studies have documented ∼ 6–10 cm/yr horizontal and vertical velocities associated with a submarine landslide flanking the Island of Hawaii's Kilauea volcano, the Hilina Slump [4,5,6,40,41,42]. Recently, GPS data from the Hilina Slump have elucidated that not only does the slump move at the above-stated, fairly smooth, background velocities, but also that the slump will occasionally deform in discrete accelerated cm-scale motions equivalent to *M*6 earthquakes but without the shaking [4,5,6]. These 'slow earthquakes' last hours and accommodate cm's of ground displacement (Fig. 3). It is not currently known, however, how the occurrence of a slow-earthquake in a submarine landslide system will affect its future movement by either making it more or less probable of failing catastrophically.

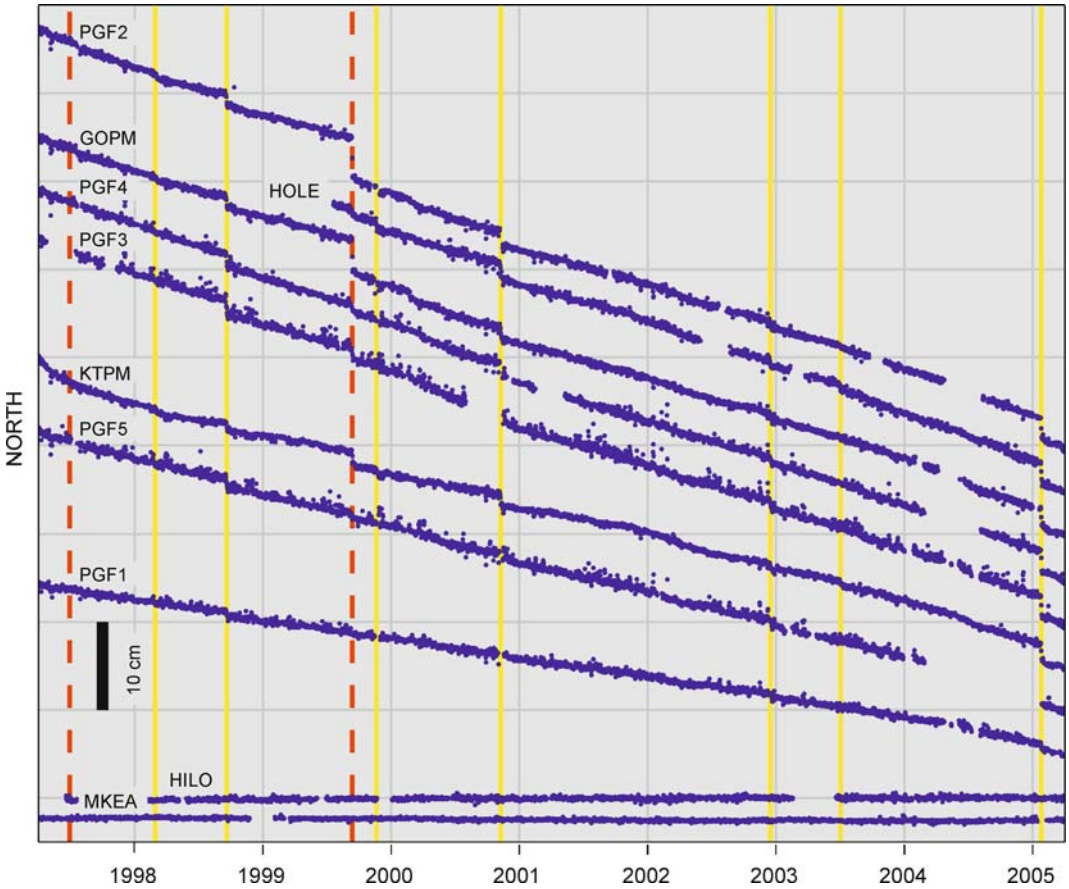

Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 4
Time series of north component of motion for selected GPS stations. Offsets of 7 slow earthquakes are identified by *yellow lines*. *Red dashed lines* are magmatic diking events

Slow Earthquakes

The term 'slow earthquake' has been used to describe a variety of transient aseismic deformation phenomena including slow precursor events preceding large earthquakes [43,44,45,46,47,48], afterslip following earthquakes [49,50,51], variable fault creep rates [52,53,54,55], certain subduction zone thrust events with unusually long durations and large amplitude tsunamis for their size [56], and discrete fault-slip events that do not produce detectable seismic shaking but are accompanied by ground displacements very similar to those produced during earthquakes [1,2,3,4,5,6,57]. Hereafter, when we use 'slow earthquake' we will refer to this latter description although the term was first used in the modern literature to describe a slow precursor to the great Chilean 1960 earthquake [43].

In contrast to slow precursor events whose ground motions, like traditional earthquakes, are measured in seconds or minutes, slow earthquake (SE) displacements usually accrue over time periods ranging from hours to days and so they have been typically sensed with geodetic methods. In the last decade, the proliferation of continuous GPS (CGPS) networks has led to numerous SE observations and the discovery of some very rich behavior. In certain regions SEs have occurred with very regular periods [2,4,58], they are often associated with non-volcanic tremor [59,60] and, apparently, SEs follow very different scaling laws (moment vs. duration) than traditional earthquakes [61]. Explanations for SE slip behavior has varied to date. For subduction zones, the combination of deep ($>$ 35 km) SE sources and their association with tremor (a phenomenon initially thought to be caused by forced fluid flow [62] but more recently also explained in terms of shear failure [63,64]) led to one current hypothesis that SE mechanics are controlled by water released during metamorphic phase changes at the interface between a subducting and overriding plate [65,66]. Other explanations have invoked rate- and state-variable frictional behavior to suggest that SEs occur preferentially at transitions between velocity strengthening and weakening regimes on a fault plane [67,68,69] and that temporally varying climatic load changes could help explain SE periodicity [58].

Due, in part, to their large magnitude deformation signal and the high concentration of CGPS networks focused on them, subduction zones have dominantly been the location of the most SEs to date [1,2,57,59,70,71,72]. Recently, the CGPS network on Kilauea volcano's mobile south flank has recorded multiple SEs [4,5,6] (Fig. 4) and it is through the Kilauea events that SEs have come to be associated with submarine landslides.

Monitoring Motion: Subaerial and Submarine Geodetic Methods

Not surprisingly, much more is known about the motion of subaerial than submarine landslides. Geodetic measurements on land combined with contemporaneous measurements of other properties (pore-water pressures, strength of materials, etc.) have allowed, in some cases, a very thorough understanding of how landslide motion is related to driving forces such as gravitational stresses and rainfall. For instance, Baum and Reid [73] instrumented a slow-moving submarine landslide in Honolulu's Manoa valley with extensometers recording at 15 minute intervals and rain gauges and found a direct correlation between rain fall and deformation events. Similarly, Malet et al. [74] showed that GPS-measured surface velocities increased to as high as 20 cm/day following periods of higher rainfall during May 1999 at the Super-Sauze earthflow in the French Alps. At a slightly different scale, Hilley et al. [75] used InSAR to simultaneously map deformation of multiple landslides in California's Berkely Hills at $\sim$ monthly intervals and found that landslide motion correlated with times of high precipitation and that during the 1997–1998 El Nino event displacement rates doubled from the background rate of $\sim$ 27–38 mm/year, albeit with a $\sim$ 3 month time lag between the onset of motion and the high precipitation.

Much of the current knowledge that we have about the motion of submarine landslides comes from their easier-to-monitor subaerial portions, such as at Kilauea's Hilina slump (Fig. 3). In the case of the Hilina Slump, fully 3/4 of the feature resides offshore at depths greater than 2000 m and the deformation monitoring network is necessarily concentrated on the down-dropped blocks near the headwall scarp of the entire system. Recently, however, submarine methods have started to provide geodetic information from the ocean floor itself e. g. [7], e. g. [76,77,78,79]. As these methods become more cost-effective and widespread they will surely yield much insight into submarine landslide kinematics and, eventually, be employed in operational hazard monitoring/mitigation scenarios.

Whether a network is solely subaerial, submarine, or some combination of the two, it is important to consider that monitoring strategies can vary substantially depending on the time duration of the expected signal, the desired threshold of detection, and the desired time latency for individual solutions. For instance, simple detection and warning of catastrophic landslide failure needs the most rapid solution latency but coarse detection thresholds (meter rather than millimeter level, for instance) are appropriate. Conversely, if the goal is to detect small, potentially precursory motions such as slow earthquakes,

then detection threshold must be as sensitive as possible along with solution latency being low.

Subaerial Geodesy: GPS

A multitude of geodetic techniques, from mechanically- to electromagnetically-based, are currently employed on landslides to measure motion caused by a range of deformational phenomena spanning opening of small surface cracks to the motion of kilometer sized blocks e. g. [74]. Data from Global Positioning System (GPS) networks has increasingly contributed to the library of observations associated with landslide motion. In particular, for submarine landslides that are large enough so that their motion causes earthquakes or slow earthquakes [4,6,80], a technique which is suitable for inter-station distances measured in kilometers is most appropriate and so below, we concentrate on the use of GPS with submarine landslides.

GPS GPS networks capable of sub-cm to mm-scale 3-dimensional ground motion detection are now deployed in many of Earth's most actively deforming zones (see for example, http://sps.unavco.org/crustal_motion/dxdt/) and readers are referred to thorough reviews of the general technique and its application for geodynamic studies [81,82,83].

Crustal motion GPS networks are usually divided into two types: those that record data continuously (CGPS) and those whose individual monuments are occupied less frequently in survey mode (SGPS). Depending on a variety of factors including the modernity of the receiver, the bandwidth of telemetry networks (should they exist), and the storage capacity of the archival center, CGPS sampling rate generally varies between once every 30 and 1 s (though most modern receivers are capable of sampling at frequencies higher than 1 Hz). SGPS sampling is more varied though it usually comprises re-occupation of sites at intervals ranging from months to years with occupation times of hours to days and sampling rates similar to CGPS. Accordingly, SGPS networks are more useful for wider ranging spatial characterization of deformation phenomena rather than for the rapid detection of motion or for tracking temporal evolution during transient events.

High rates of sampling alone, however, do not guarantee that CGPS network positional solutions will achieve their highest precision and/or accuracy. Assuming that all sites within a CGPS network have stable monuments and high-grade geodetic antennae and dual frequency receivers, the most important components of its error budget for deformation monitoring are: (1) integer ambiguity resolution; (2) orbital estimation; (3) atmospheric delay estimation; (4) antenna multipath; (5) satellite constellation geometry, and (6) intra-network baseline length. For networks monitoring landslides with spatial scales on the order of kms or tens of kms, however, the baselines are short enough that errors scaling with baseline lengths are small contributors to the error budget. In addition, as absolute positioning in a global reference frame is not essential, precise orbital estimation is less important. For the other error sources, freely available software packages such as GAMIT, GIPSY, and BERNESE (http://facility.unavco.org/software/processing/processing.html), and precise orbit processing centers such as the IGS (International GNSS Service, http://igscb.jpl.nasa.gov/) or the Scripps Orbit and Permanent Array Center (SOPAC, http://sopac.ucsd.edu/) usually allow good enough mitigation of these errors so that daily GPS solutions have resolution on the order of 2–5 mm in the horizontal and $\sim$ 2–3 times worse in the vertical e. g. [83]. This resolution rule-of-thumb, however, generally applies to post-processed data with occupation times exceeding $\sim$ 6–8 h, use of precise orbits, and the best atmospheric estimation techniques e. g. [84]. Because of the hours-to-days delay needed for estimating various grades of precise orbits and the computational time needed to estimate all the parameters for all sites with this much data this standard rule-of-thumb cannot necessarily be applied to a real-time or near real-time solution.

Real- and Near-Real Time GPS Processing There are a variety of processing techniques that may be suitable for real-time or near-real time monitoring applications for the subaerial counterparts to submarine landslides. For instance, the well-known real time kinematic (RTK) positioning technique frequently employed by the surveying community applies differential corrections sent via radio link between stations to yield site position estimates with cm-scale precision over baselines up to $\sim$ 10 km in real-time [85]. The technique works best for baseline distances less than $\sim$ 10 km typically, because the assumption of correlated errors between stations is not necessarily valid and differential corrections cannot be accurately applied for larger baselines. Thus, RTK may be a suitable monitoring technique if a more spatially contained portion of a landslide, such as a fault zone, is a particularly good indicator of motion for the overall larger unit.

It is the number of measurement epochs required to resolve the integer-cycle phase ambiguity inherent with GPS data, however, that is the principal factor limiting the temporal latency of high resolution GPS positioning solutions. Kinematic techniques typically require minutes worth of GPS data (when sampled at 1–30 s) in order to re-

solve integer ambiguities for initialization and reinitialization if a cycle slip or loss of phase lock occurs during measurement, although progress is being made on mitigating ionospheric effects and reducing time of integer ambiguity resolution e. g. [86]. Regardless, if GPS station locations are chosen with relatively good sky-view then loss of lock due to poor satellite visibility should be minimal.

Recently, Bock et al. [87] developed a method of resolving integer ambiguities from a single epoch of dual frequency GPS phase and pseudo range data that provides independent epoch-by-epoch position estimates over baselines as large as ~ 40 km. Their method allowed instantaneous positioning resolution of 1.5 cm in the horizontal and 7–8 times worse in the vertical coordinates. Langbein and Bock [88] applied the technique to determine offsets in positional time series due to slip events recorded by the Parkfield, California CGPS network which has similar spatial scales to a typical large submarine landslide. They found that offset sensitivity was ~ 5 mm for a 2 s sampling window and this decreased to ~ 2 mm when a 60 s window was used [88]. To the best of our knowledge this represents the current state-of-the art in terms of real-time detection of motion over spatial and temporal scales typical of submarine landslides.

At Kilauea, we have developed an automated *near* real-time processing strategy for our monitoring efforts at the Hilina Slump (Figs. 3 and 5) [89]. The CGPS data are telemetered hourly by radio modem back to the USGS Hawaii Volcano Observatory (HVO) where they are retrieved via FTP. We use a sliding window approach, collecting all data available within the most current two-hour period and performing a network solution using the PAGES [90] processing software. We use the 30 s, ionosphere free phase combination as the observable and use and hold fixed the IGS ultra-rapid orbits and apply the NGS antenna phase calibration patterns for each site [91]. We also apply the IERS standard solid Earth tide [92] and Schwiderski ocean tide loading corrections [93,94]. Every half hour we estimate a piece-wise, linear neutral atmosphere (troposphere) correction and one set of N–S and E–W neutral atmospheric gradient corrections for each site. We apply a 'weak' constraint of 10 cm to the atmospheric corrections based on examination of previous adjustments.

Figure 5a shows an example of baseline change time series from indicative stations derived from this approach and re-run in a simulated near real-time mode for the time period bracketing the slow earthquake from January 26–28, 2005 at Kilauea. Generally, the hourly baseline change noise levels are on the order of ± 10 mm, although because of some particularly large outliers, the standard deviation of the baseline changes (not including the time period of the SE) is closer to 20 mm (Fig. 5b). Some of the large excursions in the time series not associated with the SE are due to the high amplitude, strongly spatially and temporally varying atmospheric water vapor gradients often associated with tropical islands such as Hawaii e. g. [95]. For instance for the excursion in the MANE-PGF3 baseline in the middle of day 18 (Fig. 5a), the time series takes a sharp upward bend until the start of day 19 when it again begins to oscillate about its mean value from present days. This baseline change gradient is very similar to the onset of the SE near the beginning of day 26. It is clear from the remainder of the time series after day 27 that the ~ 25 mm offset is permanent, however, and so indicative of a real deformation event.

In a real-time monitoring scenario, because of the 2 h temporal latency and the atmospheric noise levels on the order of 1/3 the maximum signal levels in the baseline change plots, it would probably require on the order of 12–24 h before this event could be definitively classified as a deformational event. This could certainly be useful to hazard mitigators for being aware that a slow event such as an SE was occurring, but not in the event of a rapidly accelerating catastrophic collapse where the detection and warning time must be on the order of minutes. For more rapid warning, more sophisticated filtering techniques could be employed (such as monument motion, and network-wide coherence assessment as described below for the Network Inverse Filter [96]) although regions of large atmospheric gradient will always be hampered by low signal-to-noise ratios unless more sophisticated atmospheric mitigation techniques such as tomographic mapping [97] are employed in a real-time manner. Accordingly, the small detection threshold levels of the epoch-by-epoch processing at Parkfield described above [88] must be taken in the context of the relatively low atmospheric delay environment present there.

Submarine Geodesy

Subaerial geodetic methods have, until now, provided essentially the entirety of geodetic evidence for submarine landslide-related deformation including slow earthquakes. Our understanding of submarine landslide motion and slow earthquake mechanics stands to increase dramatically in the coming decade, however, owing to the rapidly advancing field of submarine geodesy which can now reliably provide in situ measurements of seafloor deformation. The submarine environment is particularly challenging for geodesy techniques especially if they require the propagation of energy across a medium (ocean water) that

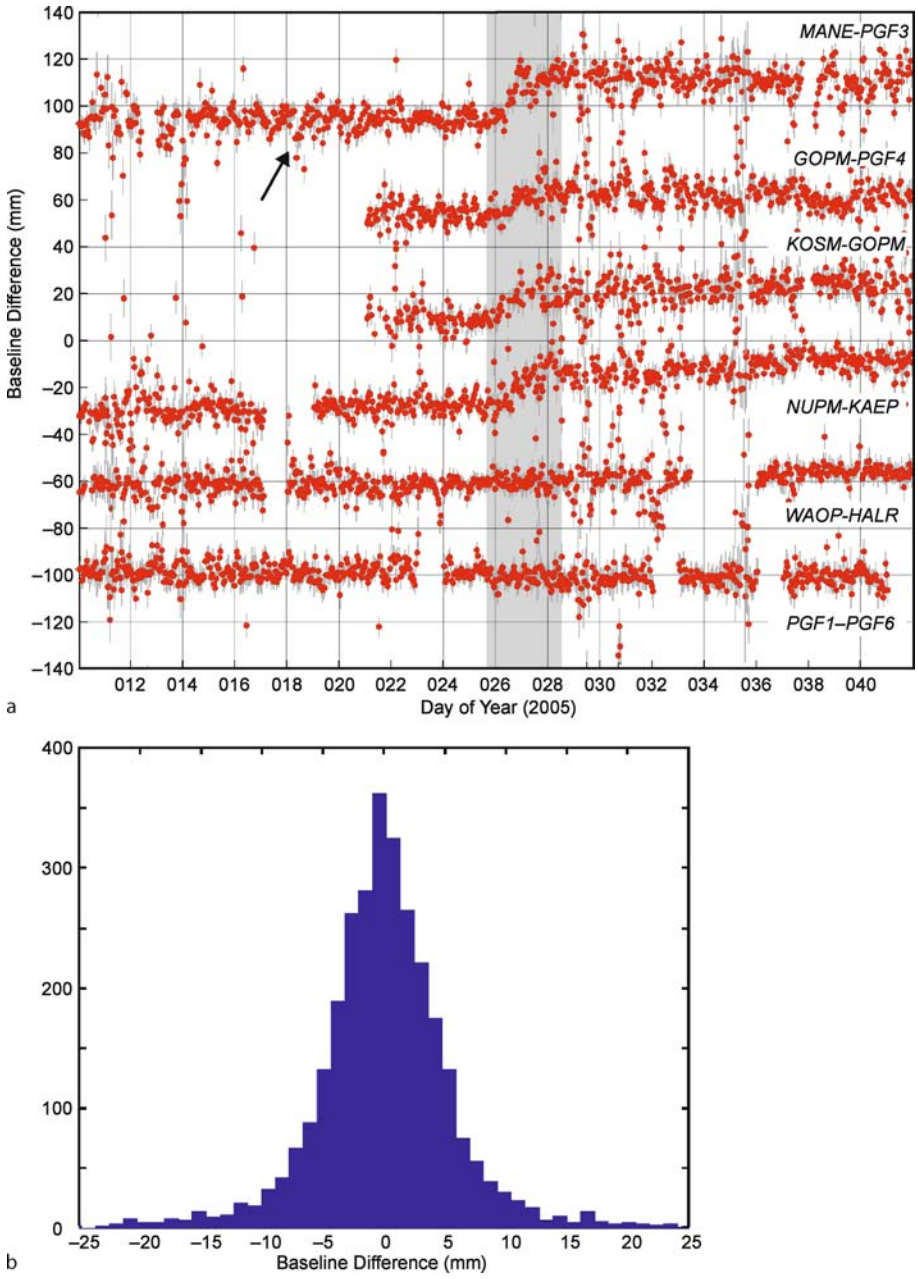

Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 5
a Baseline difference (position – median position) and 2σ errors for near-real time processing of selected sites at the Hilina Slump during early 2005. See text for description of processing. *Grey shaded region* indicates the time interval of the slow earthquake. *Black arrow* indicates an example of an anomalous trend due to atmospheric delays discussed in text. b *Histogram* of baseline difference values for all of the sites in a, excluding the time period of the slow earthquake

can exhibit highly spatially and temporally variable material properties.

Direct Path and Indirect Acoustic Approach One of the submarine geodesy techniques most frequently employed to date is direct path acoustic measurement of baseline length changes using acoustic transponders. This method has been used to make in situ observations of tectonic motions on the seafloor at the Juan de Fuca ridge where one study found cm-scale motion occurring

over a period of a few days associated with an on-axis eruption [77] and another found no detectable motion over a period of a few years at a quiescent ridge segment [76]. By interrogating each other, pairs of instruments use two-way travel time measurements to constrain the inter-station distance. After correcting for variations in sound speed that result from changes in water temperature, salinity, pressure, and local tilt of the monument itself, these systems yield sub-cm precision for individual measurements. Operating frequencies typically range from 7.5 to 108 kHz and in deeper isothermal waters, the upward refraction of sound waves prevents signals from one transponder being received by the other at distances greater than ~ 1 km [79]. Travel-time is determined simply by correlating the transmitted and received signals and picking the peak of the resultant correlogram. In quiet operating environments this can usually be done at the scale of $\sim 5\,\mu$s, resulting in a ~ 4 mm range error [7].

For the greater depth ranges, Sweeney et al. [79] devised an approach that allowed ± 2 cm resolution measurements for baselines up to 10 km at 2500–2600 m depths at a stable site on the Juan de Fuca plate. Sweeney et al. [79] suspended an acoustic interrogator hundreds of meters above the seafloor in a position acoustically visible to an array of seafloor-mounted instruments. In a manner similar to Spiess [98] they estimated the relative positions of the stations by moving the interrogator and collecting acoustic range data multiple locations.

The focus of these initial deployments of the direct and indirect acoustic approach systems was on measuring annual tectonic rates, and not necessarily on capturing transient events such as SEs that occur over hours or days. Accordingly sampling rates in current seafloor geodesy projects typically do not exceed a few times per day, largely because providing adequate power to seafloor instruments is still financially prohibitive. However with cabled oceanographic observatories scheduled to come online in 2007–2010 (http://www.neptunecanada.ca/; http://www.orionprogram.org/) the power delivery problem could be solved at high priority seafloor sites.

From October, 2005 through June, 2006 we deployed seven 10 kHz Linkquest transponders mounted ~ 3 m above steel tripods and spaced over a distance of ~ 3 km on the Hilina Slump (Fig. 3). One of the goals of this pilot project was to evaluate the performance of the transponders for use in submarine landslide monitoring at significant depth; each unit operated at a depth between 2640 and 2690 m. Battery power restrictions for the ~ 8 month duration of the project meant that the instruments ranged to one another 12 times per day.

In Fig. 6 we show range change time series from two transponder-pair baselines (~ 530 and 683 m respectively) for which data recovery was complete over the experiment's duration. For the other transponder pairs data recovery was not as complete because either: (1) the unit was knocked over by local mass-wasting events or (2) the baseline distance was too long and data dropouts occurred when the transponders lost sync with one another. The two way travel times were picked from the peak of the correlation function between the outgoing and received waveforms at one end of the baseline. The black dots in Fig. 6a and b show the raw measurements, while the red dots show the distance measurements corrected for the variations in sound speed due to temperature changes, which were measured by an external conductivity and temperature sensor mounted on the transponder frames. The raw time series show significant long-term trends due to temperature variations that are effectively removed using just the temperature measurements at the transponders. We did not correct for salinity variations because all of our conductivity sensors provided contaminated data due to clogging by the local mass wasting events.

For these baseline pairs which were oriented approximately perpendicular to the maximum expected motion of the Hilina Slump we expect essentially no motion for such short baselines and during such a short measurement period. At the 1σ level individual measurement noise is ~ 4.1 and 5.7 cm respectively, though it is clear that smaller, cm-scale changes would be detectable given a long enough time period. For instance, daily estimates of baseline length have approximately a 1 cm standard deviation. For rapid event detection, baseline changes would need to be on the order of 10 or more cm or measurement frequency would need to be increased.

GPS-Acoustic Method The acoustic measurements described above provide only relative measurements of baseline length changes; however, for submarine landslide monitoring efforts it may be desirable to place submarine and subaerial measurements in a single regional or global reference frame. Largely with the aim of furthering seafloor active tectonics studies, a research group spearheaded by the late Fred Spiess conceived of, and have begun implementing integrated GPS and acoustic seafloor geodesy (GPS-A) studies, described in great detail by Spiess et al. [7] and references therein.

Briefly, GPS-A positions seafloor geodetic markers in a global reference frame by combining kinematic GPS positioning of a sea-surface platform (ship or buoy), precision underwater sound travel time measurement over km-scale path lengths, and a strategy for eliminating the large

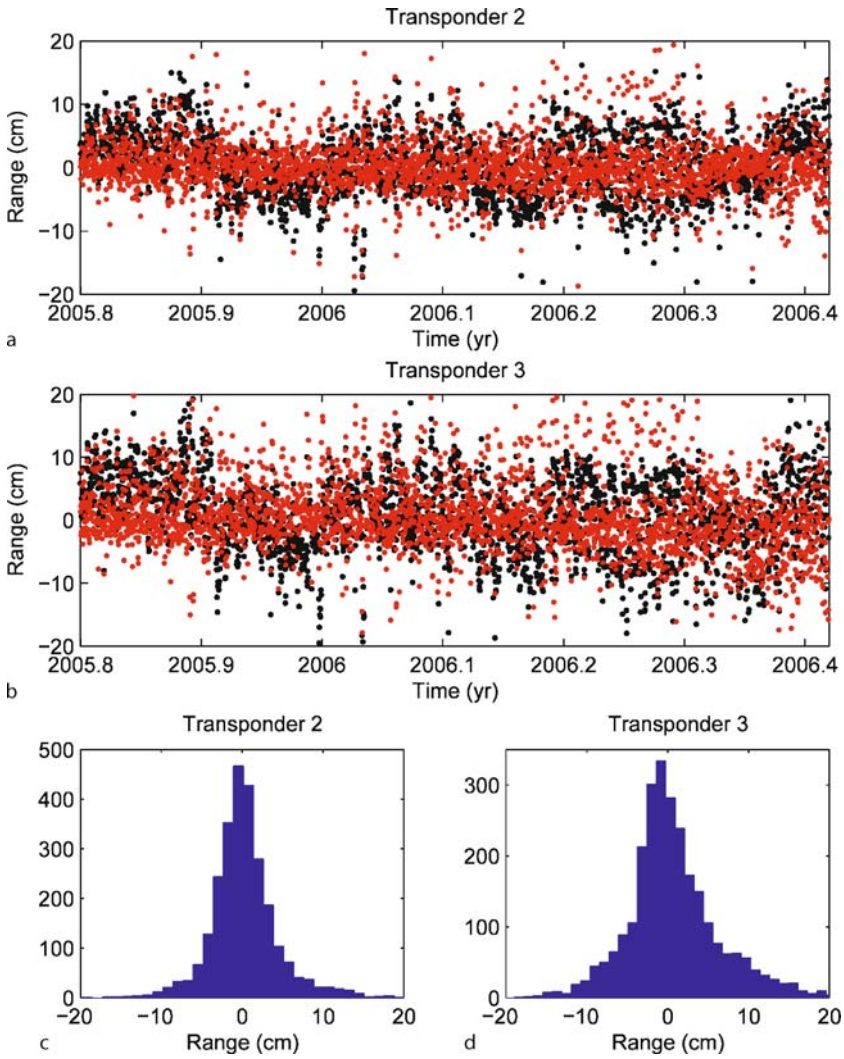

Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 6
a,b Baseline difference (position – median position) for transponder-transceiver pairs (T2-C1, T3-C1 in Fig. 3) pairs from October 2005 – June 2006. See text for processing description. *Black dots*, raw baseline difference uncorrected for temperature. *Red dots*, baseline differences corrected for temperature. c,d *Histogram* of baseline difference values from a and b, respectively

errors in travel time measurement arising from a spatially and temporally varying sound speed structure in the near surface portion of the water column. The error mitigation strategy is based on the recognition that the location of the sea surface midpoint above the center of a triangle of seafloor transponders (themselves positioned via a near-bottom acoustic survey) is the point at which the three sea surface-to-seafloor travel times are identical and so, independent of sound speed.

Spiess et al. [7] reported repeatabilities of ± 0.8 cm and ± 3.9 cm in the north and east components, respectively, of seafloor measurements on the Juan de Fuca plate from 1994–1995. More recently, Gagnon et al. [78] reported velocities with ~ 5.5 cm magnitudes and 0.6–0.8 cm 1σ errors from surveys between 2001–2003 on the updip portion of the Nazca-South America subduction zone in the Peru trench.

Optical Path Length Zumberge [99] combined commercial subaerial surveying technology (an EDM) with an optical fiber strainmeter in order to devise a low-power, cost-effective system for the harsh seafloor operating environment. EDMs (electronic distance meters) typically can measure distances with 1–2 mm precision over several km

by transmitting intensity modulated infra-red energy and measuring travel time from far away reflective surfaces. Zumberge [99] modified the EDM for seafloor geodesy by focusing the transmitted beam into the fiber's core and gathering the reflected light from the fiber's far-end into the EDM receiver optics.

The composite instrument comprises an optical fiber (enclosed in a hermetically sealed stainless steel casing) that stretches between two anchors, one 'active' and one 'passive', separated by several hundred meters on the seafloor. The active anchor houses the EDM and all supporting electronics and the passive anchor's purpose is to fix the fiber to the seafloor. In a test of their instrument, Zumberge et al. [100], report 1mm scatter of distance measurements over a 500-m-long fiber during a 50 day long period.

One of the obvious advantages of the optical path length technique is the high precision, isolation from sea water-borne errors, and the continuous nature of the measurement. The disadvantage, however, is that the technique is limited to short baselines ($< \sim 1$ km) and so the precise location of straining regions must be known a priori. Nonetheless, for monitoring purposes with very well-defined targets such as the headwall portion of an active landslide, the technique holds much promise.

Pressure Sensor Vertical Deformation Phillips et al. [40] recently developed a new technique that used pressure sensors in campaign-mode repeat surveys to measure vertical deformation rates of the seafloor at depths exceeding 2000 m at the offshore portion of the Hilina Slump. Depth and pressure are related to one another in a straightforward manner through the hydrostatic equation, though for measuring seafloor deformation rates at the cm/yr level many other reductions must be performed (Phillips et al. [40] – also her thesis). The pressure sensor method is technologically challenging, requires much ship time, and requires during each site visit both a geodetic monument and a pressure sensor (lowered from a ship) to be operating in close proximity to one another at depth on the seafloor. Because, in each revisit it was not practical to collocate the pressure sensor in a repeatable fashion on the benchmark, Phillips et al. [40] found it necessary to measure the vertical offset between the pressure sensor and the benchmark via acoustic ranging.

Important error sources for this technique are poor knowledge of α the specific volume of the water column (the reciprocal of density) and the speed of sound in depth. Due to changing tides, secular and seasonal barometric changes it is also necessary to utilize a common, stable reference, such as mean sea level (MSL) which, in turn, requires estimation of the geopotential anomaly.

Ultimately, Phillips et al. [40] conclude that their data show significant vertical deformation (9 ± 2.4 cm/yr) in the mid-section of the HS and negligible deformation towards the outer bench (just inboard of the toe of the entire landslide feature). These data are consistent with dislocation models delimiting the potential amount (25.0–28.1 $\pm$ 7.3 cm/yr), spatial extent (24.8–27.0 $\pm$ 0.5 km seaward of Kilauea's East rift zone), and depth (7 km) of the principal slipping surface below the landslide. This is a significant advance not only because it represents the first data set that allows a glimpse of how strain is partitioned in a massive submarine landslide such as the Hilina Slump, but also because it represents an ongoing monitoring project at a submarine landslide, albeit at temporally sparse sampling rate.

Data Analysis and Inversion

In addition to the challenge of acquiring geodetic data from submarine or even subaerial landslides is the added challenge of inferring sub-surface fault geometry and/or deformation processes, with attendant realistic error bounds on estimated parameters, from a data set of earth surface displacements. We identify two principal types of analysis modes: (1) process-based and (2) hazards-based. Process-based analysis focuses on deriving the most accurate and, if possible, complete assessment of the factors involved in the observed landslide motion. Hazards-based analysis has two primary components: (a) rapid warning and (b) long-term hazards estimation. Rapid warning comprises, with the smallest temporal latency possible, providing information regarding the current state of the landslide system and how it relates to current, short- and long-term dangers to life and infrastructure. Long-term estimation comprises making probabilistic statements about the components and potential future behavior of the system and so temporal latency is not a limiting consideration.

For either of the two types of analysis modes, the first objective is to usually try to relate the measured displacements d, of monuments at the surface of the earth to the source parameters, m, of the buried feature causing the deformation (usually considered a fault) through $G(m)$, a model of the deformation process. For instance, one very common model formalism, the dislocation in an elastic half-space, is a non-linear problem whose model vector m contains 9 parameters describing the location, orientation, and relative displacement of the dislocation approximating a faulting source [101,102].

Determining the most likely components of m and associated errors falls under the wide category of geophysical inversion for which there are a multitude of techniques e. g. [103]. In addition to the computational power available and the total amount of time allotted to the inversion, the choice of inversion method depends upon the nature of the model (for instance, linear vs. non-linear), the number of parameters in m, and the computational cost associated with individual model realizations and misfit assessments. Furthermore, it must be decided if the goal of an inversion is to obtain a rapid solution (a solution that satisfies some predetermined misfit level), an optimal solution (i. e. one 'best-fitting' solution), or rather to probe parameter space as thoroughly as possible with the goal of estimating posterior probability densities for each of the parameters for their value as indicators of resolution and uncertainty e. g. [104]. For instance, the most robust inversion method possible is the direct or 'grid' search, where the misfit for every possible permutation of m in parameter space is calculated. For more complicated or non-linear $G(m)$ however the computational cost associated with each forward model run can make a grid-search time prohibitive, even for a process-based analysis. For these types of more difficult problems, inversion based on Monte Carlo sampling. which collects pseudo-random samples from multidimensional parameter space as a proxy for the problem's true posterior probability density, $\sigma(m)$, has been found to be quite successful [105,106]. Indeed, it is common practice in the tectonic geodesy community to use the Okada model combined with some type of Monte-Carlo method to derive parameters of earthquakes and slow earthquakes from surface displacements observed with GPS [80,104,107].

The recent evolution of the work on the SEs at Kilauea demonstrates how important the inversion results are to the overall analysis either from a process-based or long-term hazards-based perspective. In their initial recognition of the November 2000 SE, Cervelli et al. [5] used a simulated annealing optimization routine [80] and the Okada model to invert the GPS observations and conclude that the most likely dislocation source for the SE occurred on a gently landward-dipping thrust fault plane at 5–6 km depth. Combined with the fact that the SE post-dated by 9 days a burst of local rainfall of nearly 1 m and reasonable estimates of local hydrologic parameters, Cervelli et al. [5] suggested that the increased pore pressure due to deeply percolating rainwater triggered SE motion by inducing a ~ 2 MPa pressure decrease of the effective normal stress on their preferred fault plane. Brooks et al. [4], however, showed from Gibbs Sampling inversion [104] results of GPS data on three additional SEs that posterior distributions of estimated fault parameters allow a wide family of equivalently plausible solutions, ranging from deeper seated decollement solutions at ~ 8 km depth to the more shallow fault plane favored by Cervelli et al. [5] (Fig. 7). Moreover, they showed that the other SEs were not associated with anomalous rainfall. Segall et al. [6] then used the same inversion method of Cervelli et al. [5] for the GPS data of the additional SEs relocations of high-frequency earthquakes triggered by the January 2005 SE, seismicity rate theory [108], and Coulomb stress modeling [109] to conclude that the SE (and other similar events) likely occurred at a depth of $\sim 8 \pm 1$ km on the main decollement plane below the Hilina Slump.

In this case, the crucial addition to the source inversion was the added constraint of the triggered microearthquakes which were relocated to depths near the decollement [6]. In their earthquake relocations, however, Segall et al. [6] did not use the full waveform data; rather, they used a double-difference-derived mapping with manual picks, assuming a 1D velocity model, between triggered events and previous high-precision relocations and tomography from elsewhere at Kilauea [110]. Two other studies, [111,112] performed high precision relocations using waveform cross correlation data and found different depths for the same cluster of events as Segall et al. [6]. Got and Okubo [111] suggested that their relocated events (including those triggered by the 1998 SE) do not illuminate a sub-horizontal fault plane but rather a deeper, steeply south-dipping reverse fault. Wolfe et al. [112] found that the triggered seismicity from the four SEs identified by Brooks et al. [4] consistently relocates on distinct map-view clusters aligned in the direction of the SE displacements themselves and in a subhorizontal band with depths of ~ 5 km. This is a solution consistent with Morgan et al. [21] who used seismic reflection data to identify moderately landward-dipping fault planes at similar depths. While the epicenters are well constrained in the Wolfe et al. [112] study, however, they were concerned that poor station geometry as well as near source velocity heterogeneity may bias the absolute depth of the relocations and the analyses may not be capable of distinguishing between shallow and deep fault zones. Thus, while the decollement is certainly the prime candidate on which a SE would occur (in agreement with the Segall et al. [6] suggestion), the analyses remain somewhat equivocal. This, in turn affects our best understanding of the most important hazard issue, the propensity of the decollement for a catastrophic failure. In order to more definitively answer these questions the above-mentioned research groups in conjunction with the Hawaiian Volcano Observatory of the US Geological Survey (the group charged with hazards analysis and

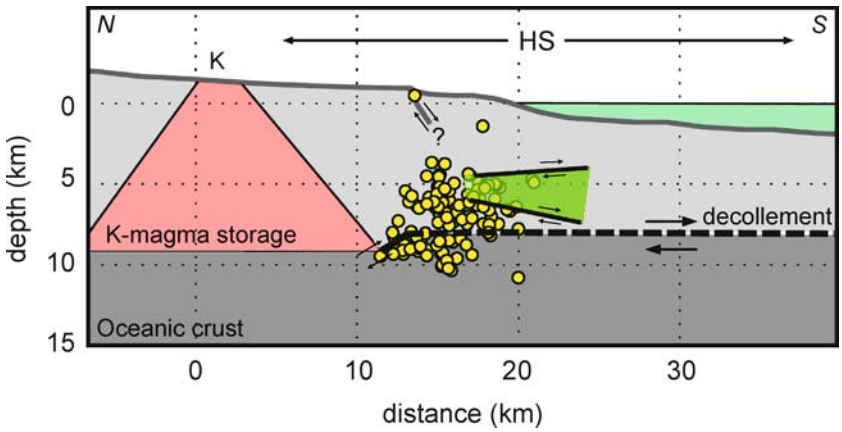

Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 7
North-south cross-section of the Hilina Slump after Got and Okubo [111]. K, Kilauea. HS, Hilina slump. *Black line* below K is the seismic (*solid*) and aseismic (*dashed*) position of the decollement from Got and Okubo [111]. Two *thicker black lines* with *green shaded area*, are the family of geologically plausible dislocation solutions equivalently supported by inversion of GPS data. *Yellow colored circles*, locations of microearthquakes from HVO catalog for ±5 days around each slow earthquake event

mitigation for the region) have teamed up in a joint seismological and geodetic research project aimed at definitively constraining the depths of triggered seismicity, and, hopefully, the depth of SE sources in the region.

While the above example focuses mostly on process and long-term hazards assessment, recent developments focused on deriving the time-dependent history of slip during a deformation event could also potentially be used for automated real-time inversion and event detection [96,113]. The Network Inverse Filter (NIF) [96] is a recursive Kalman filter algorithm that operates on either processed position estimates or raw phase data (rather than derived displacement rates) from an entire network of geodetic stations, includes a stochastic description of local benchmark motion, and finds a non-parametric description of slip rate on a fault plane as a function of time. In current implementations, the NIF employs Green's functions relating slip to surface displacement computed from the analytical solutions for a dislocation in an elastic half-space [101] although there is no reason other deformation models could not be employed [113]. Application of the NIF to the 1999 Cascadia SE allowed McGuire and Segall [113], for instance, to determine that slip rate on the slipping fault plane took up to $\sim$ 20 days to reach its peak and that the southern portion of the fault had finished slipping before the northern portion began to slip. Cervelli et al. [5] and Segall et al. [6] also applied the NIF to the Kilauea SEs to derive source-time functions for the events. One caveat, however, is that the fault plane is not solved for by the NIF, rather, it must be known and held fixed a priori. For subduction zones with greater spatial geodetic coverage above the slipping portion of the fault this condition may be satisfied more satisfactorily than at submarine landslides.

Discussion: Slow Earthquake and Submarine Landslide Process

Largely because of CGPS subaerial measurements at Kilauea there are now a suite of observations of slow earthquakes related to submarine landslide motion. As it seems that SEs are a fairly general fault slip phenomenon, occurring at a variety of subduction zone locales globally, it stands to reason that SEs may be common to submarine landslide-related deformation, at least on the flanks of ocean island volcanoes. Clearly, as seafloor geodesy projects become more common we will learn if this is the case or not. It is also clear that the search for a theoretical understanding of the slow earthquake phenomena will be coupled to further understanding the motion of submarine landslides.

Currently, explanations for slow earthquake slip behavior focus on rock mechanic theory that follows an empirically derived rate- and state-variable frictional constitutive law based on laboratory experiments e. g. [114]. (The description below follows the summary and notation in Scholz [114]). The Dieterich–Ruina or 'slowness' law is expressed as:

$$\tau = [\mu_0 + a\ln(V/V_0) + b\ln(V_0\theta/\zeta)]\sigma \ ,$$

where τ is shear stress, μ_0 is the steady-state friction, σ is effective normal stress, V is slip velocity, V_0 is a reference

velocity, a and b are frictional material properties, ζ is the critical slip distance and θ is a state variable that evolves according to:

$$d\theta/dt = 1 - \theta V/\zeta \,.$$

The frictional stability of the system, then, depends on $(a - b)$, the velocity dependence of steady-state friction, defined:

$$a - b = \delta\mu/\delta(\ln(V)) \,.$$

In the context of a simple spring-slider model approximating a slipping fault, the boundary between the stable and unstable frictional regimes will occur at a critical value of effective normal stress, σ_c, given by:

$$\sigma_c = k\zeta/-(a - b) \,,$$

where k is the stiffness. When $(a - b) > 0$ the material is said to follow 'velocity strengthening' behavior and the system is stable – earthquakes cannot nucleate in this regime and earthquakes propagating into such regions will be abruptly terminated. When $(a - b) < 0$ the material is said to follow 'velocity weakening behavior and the system is unstable for $\sigma \geqslant \sigma_c$ – earthquakes will nucleate in this regime. When $\sigma \leqslant \sigma_c$ the system exhibits oscillatory behavior and is said to be 'conditionally stable', it is stable under quasi-static loading but requires a discrete velocity perturbation in order for earthquakes to nucleate. Others interpret these relations slightly differently suggesting that nucleation occurs when matrix stiffness (scale and stress-rate dependent) drops to a critical value over a spatial scale large enough to promote rupture [68,115].

It may be in the boundary of this stability transition where slow earthquake slip behavior arises. For instance, a common observation of most of the subduction zone SEs is that they occur down-dip of the 'locked' zone or near the base of the seismogenic zone where earthquakes nucleate [1,57,58,69,71]. At these depths in subduction zones temperatures are close to the 450°C temperature at which feldspar starts to exhibit plastic behavior and so conditionally stable behavior would be expected [114], although it has recently been shown that transient oscillatory behavior may also arise naturally from system dynamics alone [116]. Faults in the conditionally stable regime, under steady-state loading, slip aseismically unless a perturbation to the system is large enough to push the fault across the stability boundary triggering an earthquake [117]. If the perturbation, however, is not quite large enough to push the fault across the stability regime then a period of stable sliding at increased velocity will occur (i. e. a slow earthquake) as the fault evolves back to steady-state [117]. It is not yet clear, however, how the explanations for subduction-zone related SEs may translate to the submarine landslide environment. While many subduction zone SEs apparently occur down-dip of the 'locked' zone where earthquakes nucleate along the subduction interface, at Kilauea, SEs apparently occur up-dip of the zone where the majority of earthquakes occur [4,6].

Recent studies have more explicitly focused on helping to explain SEs in terms of the rate- and state-variable formalism. For instance, Kato [67] used rate- and state friction to simulate the effect that 'asperities', velocity weakening regions surrounded by velocity strengthening regions, have on slip behavior and found that episodic SEs occur when the velocity-weakening patch size is close to the critical size of earthquake nucleation. Similarly, Liu and Rice [68] found that when they applied along-strike variations in the frictional $(a - b)$ parameter in models of subduction zone processes, that SE-type transient deformation events emerged spontaneously near the down-dip end of the seismogenic zone. They suggested that the downdip end of the seismogenic zone is likely to be in the conditionally stable boundary between unstable- and stable regimes and that this could allow the SE behavior. More recently, Lowry [58] employed further theoretical implications of rate- and state-variable friction under resonant loading conditions [115,118] to suggest that Earth's response to climatic redistribution of atmospheric, hydrospheric, and cryospheric loads could lead to resonant fault slip behavior that could explain observed periodic slow earthquakes on the Cocos-North America plate boundary at Guerrero, Mexico.

Shibazaki and Shimamoto [119] recently proposed an alternate model of slow earthquakes that was motivated by laboratory experiments on the velocity dependence of frictional stability. They introduced a cutoff velocity to the rate-state formulation to mimic laboratory results where fault surfaces transition from velocity weakening at low sliding velocities to velocity strengthening at higher velocities. Implementing the laboratory values for these transitions reproduced the slip (10^{-7} m/s) and rupture propagation velocities (km/day) seen in subduction zone environments. Their model predicts a linear relationship between these two velocities that can in principle be inferred from high quality continuous geodetic measurements of slow earthquakes. This relationship may be an important way to relate observations of slow earthquakes to mechanical models, particularly in regions such as Kilauea where the fault planes are shallow enough (< 10 km) to allow the details of the rupture to be resolved with high quality instrumentation. We note additionally that other recent

model parameterizations [116] yield recurrence intervals and propagation velocities similar to observed events and Lowry [58] also suggested means of relating geodetic observations and model parameters.

Two additional factors that may modulate the frictional properties of faults in submarine landslide environments include local hydrologic and magmatic forces (in regions of active volcanism) [23]. For instance, the Nov. 2000 SE at Kilauea post-dated by 9 days an intense rain storm of nearly 1 m at the southeastern Big Island [5]. Cervelli et al. [5] estimated permeability using regionally appropriate hydrologic parameters for porosity and fault zone diffusivity and suggested that the 1 m of rain could have triggered SE motion by inducing a ~ 2 MPa pressure decrease of the effective normal stress on a gently landward dipping fault at ~ 5 km depth. Brooks et al. [4], however, showed that 3 other SEs at Kilauea were not preceded by anomalous rainfall and that other periods of anomalous rainfall were not accompanied by SEs. Additionally, prior to these events Iverson [23] was skeptical of rainfall triggers as he found from a rigid wedge analysis that, in order for ground water head gradients to be large enough to destabilize the Hilina Slump, implausibly large clay layers (~ 200 m thick) or very low hydraulic diffusivity ($\sim 10^{-11}$ m^2/s) needed to be present. Others, however, have found that magmatic injection-induced mechanical and thermal pressurization of fluids may help to explain Canary and Cape Verdes Islands flank instabilities [120]. Excess shear stresses exerted on a basal decollement because of rift zone magma injection have also been shown to be of sufficient magnitude to potentially cause slip [22,23]. Although others have suggested from geodetic observations of discrete dike events at Kilauea's rift zones that the dikes are injected passively, as a response to decollement slip, rather than as a trigger for it [121,122].

Future Directions: Slow Earthquakes and Submarine Landslide Monitoring

It is clear that expanded seafloor geodetic monitoring of submarine landslides would be extremely important in bettering our understanding of both the slow earthquake process and the hazards associated with submarine landslides. For instance, at the Hilina Slump, it is not even known how much of submarine portion of the landslide actually displaces the seafloor during a slow earthquake. Given the logistical and financial challenges associated with seafloor geodesy, however, it is reasonable to ask if it is worth it to society to monitor submarine landslides. In the Hawaiian Islands, for instance, one compilation suggests that there have been at least 6 tsunami-generating landslide events in the past 300 000 years [123]. For comparison, the average 50 000 year recurrence interval for such events is 1 to 2 orders of magnitude larger than typical earthquake recurrence intervals in Southern California where substantial resources are focused on earthquake-cycle related monitoring (http://www.wgcep.org/). As discussed above, however, landslide-generated tsunami can be quite damaging both locally and regionally. One study simulated ~ 30 m wave heights reaching the California coast within 6 h of a massive collapse of the Hilina Slump at Kilauea [124]. Clearly minimizing the impact of such an event would be of societal benefit.

Continuous submarine landslide deformation monitoring, although expensive, is not out of the realm of possibility, as costs for seafloor geodetic instrumentation are decreasing. In the coming years, in order to make continuous monitoring efforts more feasible, research will likely focus on two major technical challenges: (1) instrument power delivery and (2) data transmission. While cabling a network via seafloor pathways is certainly one way of satisfying both requirements, cabling may not be the best long-term solution for a number of reasons. First, the costs of large cable lengths and laying them on the seafloor is often measured in the millions of dollars. Second, especially for submarine landslides, the seafloor across which the cable need be lain can be very rough and cable failure can be a quite common occurrence. Third, if cable runs are long then annual maintenance efforts and cost can also be quite high. Accordingly, techniques such as acoustic data transmission [125] and local power generation via a buoy, for instance, will be critical.

Acknowledgments

We thank the Hawaiian Volcano Observatory of the US Geological Survey, in particular Asta Miklius, Maurice Sako, and Kevan Kamibayashi, for their collaboration and efforts in operating the CGPS network at Kilauea. We thank Paul Segall of Stanford University for sharing GPS data from this network. We thank Matt Gould and Nick Witzell of WHOI for engineering support. We thank Capt. Rick Myer and the crew of the R/V Kilo Moana for their help in deploying and retrieving the extensometers. We thank Mark Schenewerk for help with implementing PAGES. We thank Katie Phillips and Dave Chadwell for helpful discussion and sharing a manuscript in preparation. We thank Cecily Wolfe and Gerard Fryer for stimulating discussion. In particular we thank J. Moore, T. Lowry, and D. Chadwell for thorough and insightful reviews which improved this contribution. This research was supported by the Geophysics and Geochemistry pro-

grams of the US National Science Foundation, grants from School of Ocean and Earth Science and Technology at the University of Hawaii, and a grant from the Deep Ocean Exploration Institute of the Woods Hole Oceanographic Institution.

Bibliography

1. Dragert H, Wang K, James TS (2001) A silent slip event on the deeper Cascadia subduction interface. Science 292:1525–1528
2. Miller MM, Melbourne TI, Johnson DJ, Sumner WQ (2002) Periodic slow earthquakes from the Cascadia subduction zone. Science 295:2423
3. Ozakawa S, Murakami M, Tada T (2001) Time-dependent inversion study of the slow thrust event in the Nankai trough subduction zone, southwestern Japan. J Geophys Res 106:782–802
4. Brooks BA, Foster JH, Bevis MF, Frazer LN, Wolfe CJ, Behn M (2006) Periodic slow earthquakes on the flank of Kilauea volcano, Hawai'i Earth and Planet Sci Lett 246:207–216
5. Cervelli P, Segall P, Johnson K, Lisowski M, Miklius A (2002) Sudden aseismic fault slip on the south flank of Kilauea volcano. Nature 415:1014–1018
6. Segall P, Desmarais EK, Shelly D, Miklius A, Cervelli P (2006) Earthquakes Triggered by Silent Slip Events on Kilauea Volcano, Hawaii. Nature 442:71–74
7. Spiess FN et al (1998) Precise GPS/Acoustic positioning of seafloor reference points for tectonic studies. Phys Earth Planet Inter 108:101–112
8. Moore JG et al (1989) Prodigious Submarine Landslides on the Hawaiian Ridge. J Geophys Res 94:17465–17484
9. Krastel S et al (2001) Submarine landslides around the Canary Islands. J Geophys Res 106:3977–3997
10. Carracedo JC, Day SJ, Guillou H, Perez Torrado FJ (1999) Giant Quaternary landslides in the evolution of La Palma and El Hierro, Canary Islands. J Volcan Geotherm Res 94:169–190
11. Day SJ, Heleno da Silva SIN, Fonseca JFBD (1999) A past giant lateral collapse and present-day flank instability of Fogo, Cape Verde Islands. J Volcan Geotherm Res 94:191–218
12. Tucholke B (1992) Massive submarine rockslide in the rift-valley wall of the Mid-Atlantic Ridge. Geology 20:129–132
13. Driscoll N, Weissel JK, Goff JA (2000) Potential for large-scale submarine slope failure and tsunami generation along the US Mid-Atlantic coast. Geology 28:407–410
14. Moore JG (1964) Giant submarine landslides on the Hawaiian Ridge. US Geol Survey Prof Paper D 501:95–98
15. Moore JG, Fiske RS (1969) Volcanic substructure inferred from dredge samples and ocean-bottom photographs, Hawaii. Geol Soc Am Bull 80:1191–1202
16. Marti J, Hurlimann M, Ablay G, Gudmundsson A (1997) Vertical and lateral collapses on Tenerife (Canary Islands) and other volcanic ocean islands. Geology 25:879–882
17. Ward SN, Day S (2001) Cumbre Vieja Volcano – Potential collapse and tsunami at La Palma, Canary Islands. Geophys Res Lett 28:3397–3400
18. Clague DA, Moore JG (2002) The proximal part of the giant submarine Wailau landslide, Molokai, Hawaii. J Volcan Geotherm Res 113:259–287
19. Coombs ML, Clague DA, Moore GF, Cousens BL (2004) Growth and collapse of Waianae Volcano, Hawaii, as revealed by exploration of its submarine flanks. Geochem Geophys Geosyst 5: doi:10.1029/2004GC000717
20. Morgan JK, Clague DA, Borchers DC, Davis AS, Milliken KL (2007) Mauna Loa's submarine western flank: Landsliding, deep volcanic spreading, and hydrothermal alteration. Geochem Geophys Geosyst 8:Q05002; doi:10.1029/2006GC001420
21. Morgan JK, Moore GF, Clague DA (2003) Slope failure and volcanic spreading along the submarine south flank of Kilauea volcano, Hawaii. J Geophys Res 108:2415, doi:10.1029/2003JB002411
22. Dieterich J (1988) Growth and Persistence of Hawaiian Volcanic Rift Zones. J Geophys Res 93:4258–4270
23. Iverson RM (1995) Can magma-injection and groundwater forces cause massive landslides on Hawaiian volcanoes? J Volcan Geotherm Res 66:295–308
24. Cannon EC, Bürgmann R, Owen SE (2001) Shallow normal faulting and block rotation associated with the 1975 Kalapana earthquake, Kilauea volcano, Hawaii. Bull Seismol Soc Am 91:1553–1562
25. Moore JG, Normark WR, Holcomb RT (1994) Giant Hawaiian landslides. Ann Rev Earth Planet Sci 22:119–144
26. Bardet J-P, Synolakis CE, Davies HL, Imamura F, Okal EA (2003) Landslide Tsunamis: Recent Findings and Research Directions. Pure Appl Geophys 160:1793–1809
27. Hasegawa HS, Kanamori H (1987) Source Mechanism of the Magnitude 7.2 Grand Banks Earthquake of November 18, 1929: Double-couple or Submarine Landslide? Bull Seismol Soc Am 77:1984–2004
28. Yeh H et al (1993) The Flores Island Tsunami. Eos, Transactions, Am Geophys Union 74:371–373
29. Imamura F, Synolakis CE, Titov V, Lee S (1995) Field Survey of the 1994 Mindoro Island, Phillipines Tsunami. Pure Appl Geophys 144:875–890
30. Geist EL (2000) Origin of the 17 July 1998 Papua New Guinea Tsunami: Earthquake or Landslide? Seismol Res Lett 71:344–351
31. Yalciner AC et al (1999) Field Survey of the 1999 Izmit Tsunami and Modeling Effort of New Tsunami Generation Mechanism. Eos, Transactions, Am Geophys Union F751(abstract):80
32. Caminade JP et al (2001) Vanuatu Earthquake and Tsunami Caused Much Damage, Few Casualties. Eos Trans Am Geophys Union 81:641,646–647
33. Nettles M, Ekstrom G (2004) Long-Period Source Characteristics of the 1975 Kalapana, Hawaii, Earthquake. Bull Seismol Soc Am 94:422–429
34. Tilling RI et al (1976) Earthquake and Related Catastrophic Events, Island of Hawaii, November 29, 1975: A preliminary report. US Geol Surv Circ 740:33
35. Day SJ, Watts P, Grilli ST, Kirby JT (2004) Mechanical models of the 1975 Kalapana, Hawaii earthquake and tsunami. Mar Geol 215:59–92, doi:10.1016/j.margeo.2004.11.008
36. Okal EA, Synolakis CE (2003) A Theoretical Comparison of Tsunamis from Dislocations and Landslides. Pure Appl Geophys 160:2177–2188
37. Ward S (2001) Landslide tsunami. J Geophys Res 106:11201–11215
38. Murty TS (2003) Tsunami wave height dependence on landslide volume. Pure Appl Geophys 160(10–11):2147–2153

39. Fine IV, Rabinovich AB, Bornhold BD, Thomson RE, Kulikov EA (2005) The Grand Banks landslide-generated tsunami of November 18, 1929, preliminary analysis and numerical modeling. Mar Geol 215:45–57
40. Phillips KA, Chadwell CD, Hildebrand JA (2008) Vertical deformation measurements on the submerged south flank of Kilauea volcano, Hawai'i reveal seafloor motion associated with volcanic collapse. J Geophys Res 113:B05106; doi:10.1029/2007JB005124
41. Owen S et al (2000) Rapid deformation of Kilauea Volcano: Global Positioning System measurements between 1990 and 1996. J Geophys Res B: Solid Earth 105:18983–18998
42. Delaney PT et al (1998) Volcanic spreading at Kilauea, 1976–1996. J Geophys Res 103:18003–18023
43. Kanamori H, Stewart GS (1979) A slow earthquake. Phys Earth Planet Inter 18:167–175
44. Sacks IS, Suyehiro S, Linde AT, Snoke JA (1978) Slow earthquakes and stress redistribution. Nature 275:599–602
45. Ihmle PF, Jordan TH (1994) Teleseismic search for slow precursors to large earthquakes. Science 266:1547–1551
46. Kedar S, Watada S, Tanimoto T (1994) The 1989 Macquarie Ridge earthquake: Seismic moment estimation from long-period free oscillations. J Geophys Res 99:17893–17908
47. McGuire JJ, Ihmle PF, Jordan TH (1996) Time-domain observations of a slow precursor to the 1994 Romanche transform earthquake. Science 274:82–85
48. Kanamori H, Cipar JJ (1974) Focal process of the great Chilean earthquake May 22:1960. Phys Earth Planet Inter 9:128–136
49. Burgmann R et al (2001) Rapid aseismic moment release following the 5 December, 1997 Kronotsky, Kamchatka, earhquake. Geophys Res Lett 28:1331–1334
50. Heki K, Miyazaki S-I, Tsuji H (1997) Silent fault slip following an interplate thrust earthquake at the Japan Trench. Nature 386:595–598
51. Segall P, Burgmann R, Matthews M (2000) Time-dependent triggered afterslip following the 1989 Loma Prieta earthquake. J Geophys Res B: Solid Earth 105:5615–5634
52. Linde AT, Gladwin MT, Johnston MJS, Gwyther RL, Bilham RG (1996) A slow earthquake sequence on the San Andreas Fault. Nature (London) 383:65–68
53. Gwyther RL, Gladwin MT, Mee M, Hart RHG (1996) Anomalous shear strain at Parkfield during 1993–94. Geophys Res Letters 23:2425–2428
54. Gao SS, Silver PG, Linde AT (2000) Analysis of deformation data at Parkfield, California: Detection of a long-term strain transient. J Geophys Res 105:2955–2967
55. Lienkaemper JL, Galehouse JS, Simpson RW (1997) Creep Response of the Hayward Fault to Stress Changes Caused by the Loma Prieta Earthquake. Science 276:2014–2016
56. Kanamori H, Kikuchi M (1993) The 1992 Nicaragua earthquake: a slow tsunami earthquake associate with subducted sediments. Nature 361:714–716
57. Lowry AR, Larson KM, Kostoglodov V, Bilham R (2001) Transient fault slip in Guerrero, southern Mexico. Geophys Res Lett 28:3753–3756
58. Lowry AR (2006) Resonant slow fault slip in subduction zones forced by climatic load stress. Nature 442(7104):802–805
59. Ito Y, Obara K, Shiomi K, Sekine S, Hirose H (2006) Slow earthquakes coincident with episodic tremors and slow slip events. Science 26:503–506
60. Rogers G, Dragert H (2003) Episodic Tremor and slip on the Cascadia Subduction Zone: The chatter of silent slip. Science 300:1942–1943
61. Ide S, Beroza GC, Shelly DR, Uchide T (2007) A scaling law for slow earthquakes. Nature 447:76–79; doi:10.1038/nature05780
62. Aki K, Fehler M, Das S (1977) Source mechanism of volcanic tremors: fluid driven crack models and their application to the 1963 Kilauea eruption. J Volcan Geotherm Res 141:259–287
63. Rubenstein JL et al (2007) Non-volcanic tremor driven by large transient shear stresses. Nature 448:579–582
64. Shelly DR, Beroza GC, Ide S (2007)Non-volcanic tremor and low-frequency earthquake swarms. Nature 446:305–307; doi:10.1038/nature05666
65. Peacock S, Wang K (1999) Seismic Consequences of Warm Versus Cool Subduction Metamorphism: Examples from Southwest and Northeast Japan. Science 286:937–939
66. Julian B (2002) Seismological detection of slab metamorphism. Science 296:1625–1626
67. Kato N (2004) Interaction of slip on asperities: Numerical simulation of seismic cycles on a two-dimensional planar fault with nonuniform frictional property. J Geophys Res 109:B12306; doi:10.1029/2004JB003001
68. Liu Y, Rice JR (2005) Aseismic slip transients emerge spontaneously in three-dimensional rate and state modeling of subduction earthquake sequences. J Geophys Res 110:B08307; doi:10.1029/2004JB003424
69. Miyazaki S, McGuire JJ, Segall P (2003) A transient subduction zone slip episode in southwest Japan observed by the nationwide GPS array. J Geophys Res 108(B2):2087; doi:10.1029/2001JB000456
70. Kostoglodov V et al (2003) A large silent earthquake in the Guerrero seismic gap, Mexico. Geophys Res Lett 30:1807, doi:10.1029/2003GL017219
71. Freymueller J et al (2001) The great alaska 'earthquake' of 1998–2001. EOS Trans Am Geophys Un 82(47)
72. Hirose H, Hirahara K, Kimata F, Fujii F, Miyazaki S (1999) A slow thrust slip event following the two 1996 Hyuganada earthquakes beneath the Bungo Channel, southwest Japan. Geophys Res Lett 26:3237–3240
73. Baum RL, Reid ME (1995) Geology, hydrology, and mechanics of a slow-moving, clay-rich landslide, Honolulu, Hawaii. In: Haneberg WC, Anderson SA (eds) Clay and Shale Slope Instability. Geol. Soc. of Am, Boulder, Colorado
74. Malet JP, Maquaire O, Calais E (2002) The use of Global Positioning System techniques for the continuous monitoring of landslides: application to the Super-Sauze earthflow (Alpes-de-Haute-Provence, France). Geomorphology 43:33–54
75. Hilley GE et al (2004) Dynamics of slow-moving landslides from permanent scatterer analysis. Science 304(5679):1952–1955
76. Chadwell CD et al (1999) No spreading across the southern Juan de Fuca Ridge axial cleft during 1994–1996. Geophys Res Lett 26:2525–2528
77. Chadwick WW, Embley RW, Milburn HR, Meining C, Stapp M (1999) Evidence for deformation associated with the 1998 eruption of Axial Volcano, Juan de Fuca Ridge, from acoustic extensometer measurements. Geophys Res Lett 26:3441–3444

78. Gagnon K, Chadwell CD, Norabuena E (2005) Measuring the onset of locking in the Peru-Chile trench with GPS and acoustic measurements. Nature 434:205–208
79. Sweeney AD, Chadwell CD, Hildebrand JA, Spiess FN (2005) Centimeter-level positioning of seafloor acoustic transponders from a deeply-towed interrogator. Mar Geol 28:39–70, doi:10.1080/01490410590884502
80. Cervelli P, Murray MH, Segall P, Aoki Y, Kato T (2001) Estimating source parameters from deformation data, with an application to the March 1997 earthquake swarm off the Izu Peninsula, Japan J Geophys Res 106:11217–11237
81. Dixon TH (1991) An Introduction to the Global Positioning System and Some Geological Applications. Rev Geophys 29:249–276
82. Hager BH, King RW, Murray MH (1991) Measurement of Crustal Deformation Using the Global Positioning System. Ann Rev Earth Planet Sci 19:351–382
83. Segall P, Davis JL (1997) GPS applications for geodynamics and earthquake studies. Ann Rev Earth Planet Sci 25:301–336
84. Eckl MC, Snay RA, Soler T, Cline MW, Mader GL (2001) Accuracy of GPS-derived relative positions as a function of interstation distance and observing-session duration. J Geodesy 75:633–640
85. Leick A (2004) GPS satellite surveying. Wiley, Hoboken
86. Warnant R, Kutiev I, Marinov P, Bavier M, Lejeune S (2007) Ionospheric and geomagnetic conditions during periods of degraded GPS position accuracy: 1. Monitoring variability in TEC which degrades the accuarcy of Real-Time Kinematic GPS applications. Adv Space Res 39:875–880
87. Bock Y, Nikolaidis RM, de Jonge PJ, Bevis M (2000) Instantaneous geodetic positioning at medium distances with the Global Positioning System. J Geophys Res 105:28233–28253
88. Langbein, J (2004) Noise in two-color electronic distance meter measurements revisited. J Geophys Res 109:B04406; doi:10.1029/2003JB002819
89. Brooks BA, Foster JF, Miklius A, Schenewerk M (2005) Extended GPS Network and Near Real-Time processing, Mauna Loa Volcano, Hawai'i. Eos Trans AGU, 86:Fall Meet Suppl
90. Schenewerk M, Dillinger W, Hilla S (2000) http://www.ngs.noaa.gov/GRD/GPS/DOC/toc.html
91. Mader G, MacKay JR (1996) Geoscience Laboratory, Office of Ocean and Earth Sciences NOS, NOAA. Silver Spring, Maryland
92. McCarthy D, Petit G (eds) (2003) IERS Technical Note 32
93. Cartwright DE, Edden AC (1973) Corrected tables of tidal harmonics. Geophys J R Astron Soc 33:253–264
94. Cartwright DE, Taylor RJ (1971) New computation in the tide-generating potential. Geophys J R Astron Soc 23:45–74
95. Foster J, Bevis M, Chen Y-L, Businger S, Zhang Y (2003) The Ka'u Storm (Nov 2000): Imaging precipitable water using GPS. J Geophys Res 108:4585
96. Segall P, Matthews M (1997) Time dependent inversion of geodetic data. J Geohys Res 102:22391–22409
97. Flores A, Ruffini G, Rius A (2000) 4D tropospheric tomography using GPS slant wet delays. Ann Geophys 18:223–234
98. Spiess FN (1985) Analysis of a possible sea floor strain measurement system. Mar Geodesy 9:385–398
99. Zumberge M (1997) Precise Optical Path Length Measurement Through an Optical Fiber: Application to Seafloor Strain Monitoring. Ocean Eng 24:532–542
100. Zumberge M et al (2006) In: Nadim F, Pottler R, Einstein H, Klapperich H, Kramer S (eds) 2006 ECI Conference on Geohazards. Lillehammer, Norway
101. Okada Y (1985) Surface deformation due to shear and tensile faults in a half-space. Bull Seismol Soc Am 75:1135–1154
102. Steketee JA (1958) Some Geophysical Applications of the Elasticity Theory of Dislocations. Can J Phys 36:1168–1198
103. Menke W (1984) Geophysical Data Analysis: Discrete Inverse Theory. Academic Press Inc, San Diego
104. Brooks BA, Frazer LN (2005) Importance reweighting reduces dependence on temperature in Gibbs samplers: an application to the coseismic geodetic inverse problem. Geophys J Int 161:12–21
105. Mosegaard K, Sambridge M (2002) Monte Carlo analysis of inverse problems. Inverse Probl 18:29–54
106. Kirkpatrick S, Gelatt CDJ, Vecchi MP (1983) Optimization by simulated annealing. Science 220:671–680
107. Hudnut KW et al (1996) Co-Seismic Displacements of the 1994 Northridge, California, Earthquake. Bull Seismol Soc Am 86:19–36
108. Dieterich J, Cayol V, Okubo PG (2000) The use of earthquake rate changes as a stress meter at Kilauea volcano. Nature 408:457–460
109. Stein RS (1999) The role of stress transfer in earthquake occurrence. Nature 402:605–609
110. Hansen S, Thurber CH, Mandernach MJ, Haslinger F, Doran C (2004) Seismic velocity and attenuation structure of the East Rift Zone and South Flank of Kilauea Volcano, Hawaii. Bull Seismol Soc Am 94:1430–1440
111. Got J-L, Okubo PG (2003) New insights into Kilauea's volcano dynamics brought by large-scale relative relocation of microearthquakes. J Geophys Res 108(B7):2337; doi:10.1029/2002JB002060
112. Wolfe CJ, Brooks BA, Foster JH, Okubo PG (2007) Microearthquake streaks and seismicity triggered by slow earthquakes on the mobile south flank of Kilauea Volcano, Hawai'i. Geophys Res Lett (In press)
113. McGuire JJ, Segall P (2003) Imaging of aseismic fault slip transients recorded by dense geodetic networks. Geophys J Int 155:778–788
114. Scholz CJ (1998) Earthquakes and friction laws. Nature 391:37–42
115. Perfettini H, Schmittbuhl J, Rice JR, Cocco M (2001) Frictional response induced by time-dependent fluctuations of the normal loading. J Geophys Res 106:435–438
116. Liu Y, Rice JR (2007) Spontaneous and triggered aseismic deformation transients in a subduction fault model. J Geophys Res 112:B09404; doi:10.1029/2007JB004930
117. Rice JR, Gu JC (1983) Earthquake aftereffects and triggered seismic phenomena. Pure Appl Geophys 121:187–219
118. Perfettini H, Schmittbuhl J (2001) Periodic loading on a creeping fault: Implications for tides. Geophys Res Lett 28:435–438
119. Shibazaki B, Shimamoto T (2007) Modelling of short-interval silent slip events in deeper subduction interfaces considering the frictional properties at the unstable-stable transition regime. Geophys J Int 171(1):191–205
120. Elsworth D, Day SJ (1999) Flank collapse triggered by intrusion: the Canarian and Cape Verde Archipelagoes. J Volcan Geotherm Res 94:323–340

121. Cervelli P et al (2002) The 12 September 1999 upper east rift zone dike intrusion at Kilauea Volcano, Hawaii. J Geophys Res B: Solid Earth 107:3-1–3-13
122. Owen S et al (2000) January 30, 1997 eruptive event on Kilauea Volcano, Hawaii, as monitored by continuous GPS. Geophys Res Lett 27:2757–2760
123. McMurtry GM, Watts P, Fryer GJ, Smith JR, Imamura F (2004) Giant landslides, mega-tsunamis, and paleo-sea level in the Hawaiian Islands. Mar Geol 203:219–233
124. Ward S (2002) Slip-sliding away. Nature 415:973–974
125. Frye D et al (2006) An Acoustically Linked Moored–Buoy Ocean Observatory. EOS Trans AGU 87:213–218

Swarm Intelligence

Gerardo Beni
University of California Riverside, Riverside, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Biological Systems
Robotic Systems
Artificial Life Systems
Definition of Swarm
Standard-Mathematics Methods
Swarm Optimization
Non-Linear Differential Equations Methods
Limitations of Standard-Mathematics Methods
Cellular-Computing Methods
Randomness in Swarm Intelligence
Asynchronous Swarms
The Realization of Asynchronous Swarms
Characteristics of Swarm Intelligence
Swarms of Intelligent Units
Future Directions
Bibliography

Glossary

Swarm Intelligence (Definition 1, Sect. "Definition of the Subject"). The *intuitive* notion of "Swarm Intelligence" is that of *a "swarm" of agents (biological or artificial) which, without central control, collectively (and only collectively) carry out (unknowingly, and in a somewhat-random way) tasks normally requiring some form of "intelligence"*. (Definition 5, Sect. "Swarms of Intelligent Units"). The capability of universal computation carried out with natural asynchrony by a dynamic cellular computing system, none of whose cells can predict the computation done by the swarm.

Swarm Robotics The technology of robotic systems capable of Swarm Intelligence.

Stigmergy Indirect communication through modification of the environment.

Pheromone A chemical that triggers an innate behavioral response in another member of the same animal species

Elementary Swarm An ordered set of N units described by the N components v_i $(i = 1, 2, \ldots, N)$ of a vector $\boldsymbol{v}$; any unit i may update the vector, at any time t_i, using a function f of K_i vector components.

$$\forall i \in N\colon v_i(t+1) = f(\mathbf{v}_{k\in K(i)}(t))$$

Cellular automaton A system, evolving in discrete time steps, with four properties: a grid of cells, a set of possible states of the cells, a neighborhood, and a function which assigns a new state to a cell given the state of the cell and of its neighborhood.

von Neumann architecture Computer design that uses one processing unit and one storage unit holding both instructions and data.

Cellular-computing architecture Computer design that uses cellular automata, and related machines, as processors and as storage of instruction and data.

Intelligence (working definition for Swarm Intelligence) Ability to carry out universal computation.

Natural asynchrony Asynchronous updating characterized by three properties: more than one unit may update at each time step; any unit may update more than once in each updating cycle; and the updating order varies randomly for every updating cycle.

Optimization algorithms Algorithms to satisfy a set of constraints and/or optimize (e. g., minimize) a function by systematically choosing the values of the variables from an allowed set.

Swarm optimization Ant colony optimization, Particle swarm optimization, and related probabilistic optimization algorithms.

Ant colony optimization Probabilistic optimization algorithm where a colony of artificial ants cooperate in finding solutions to optimization problems.

Particle swarm optimization Probabilistic optimization algorithm where a swarm of potential solutions (particles) cooperate in finding solutions to discrete optimization problems.

Game of Life A cellular automaton designed to simulate life-like phenomena.

Dynamic cellular computing system Cellular computing system whose cells are mobile.

Unpredictable system A system such that complete knowledge of its state and operation at any given time

is insufficient to compute the system's future state before the system reaches it.

Definition of the Subject

The research area identified as "Swarm Intelligence" has evolved rapidly. The term "Swarm Intelligence" first appeared in 1989 [1,2]; by 2007 "Swarm Intelligence" was in the title of four books [3,4,5,6], two series of conference proceedings [7,8] and in a new technical journal [9], without mentioning other areas in which the term swarm itself has become popular. As the use of the term "Swarm Intelligence" has increased, its meaning has broadened to a point in which it is often understood to encompass almost any type of collective behavior. And since the term "Swarm Intelligence" has popular appeal, it is also sometimes used in contexts which have limited scientific or technological content. Other meanings, however, refer rigorously to precise concepts. The following treatment of Swarm Intelligence is based only on concepts that can be clearly defined and quantified. Hence it is more restricted than some broader Swarm Intelligence presentations but, even so, it describes an interrelated scientific/technical core which forms a solid basis for a well-defined multidisciplinary research area.

Definition 1 The intuitive notion of "Swarm Intelligence" is that of a "swarm" of agents (biological or artificial) which, without central control, collectively (and only collectively) carry out (unknowingly, and in a somewhat-random way) tasks normally requiring some form of "intelligence".

A more specific definition requires a detailed discussion and so it is given at the end of the article. (Sects. "Characteristics of Swarm Intelligence" and "Swarms of Intelligent Units").

Although this notion of Swarm Intelligence might seem vague, we will see in the course of this article that in fact it has many specific implications. Note that the notion is broad, which partly explains its widespread use, but not so broad as to include any type of collective action of groups of simple entities, as will become clear later.

These characteristics of Swarm Intelligence are also those of several biological systems, e. g., some insect societies or some components of the immune system, so that Swarm Intelligence has become important for understanding certain mechanisms in biology.

Technologically, the importance of 'swarms' is mainly based on the potential advantages over centralized systems. The potential advantages are: (1) *economy*; the swarm components (units) are simple, hence, (in principle) mass producible, modularizable, interchangeable, and disposable; (2) *reliability*, due to the redundancy of the components; destruction/death of some units has negligible effect on the accomplishment of the task, as the swarm adapts to the loss of few units; (3) ability to perform *tasks beyond those of centralized systems*, e. g., escaping enemy detection.

From this initial perspective on potential advantages, the actual application of Swarm Intelligence has extended to many areas, described in the body of this article, and its potential for future applications remains high, as discussed in the concluding section.

Some current, proposed and/or potential applications are in defense and space technologies, (e. g., control of groups of unmanned vehicles in land, water, or air), flexible manufacturing systems, advanced computer technologies (biocomputing), medical technologies, and telecommunications.

Introduction

Swarm Intelligence (SI) investigations were initially motivated by studies of groups of simple robotic units which offered promise for technology. These 'swarms' are modeled as collections of simple quasi-identical units with decentralized control and independent clocks [10]. The number of units is intermediate between those of typical systems investigated in physics and other traditional science fields; in fact, the swarm is of the order of 10^2 to 10^s units, where $s \ll 23$, i. e., the swarm is not composed of so many units that statistical physics methods can be applied to it, nor of such a few units that its dynamic can be solved exactly (or numerically to high precision).

These features are typical also of many biological systems, such as insect societies. They are also the features of robotic systems potentially economic, reliable and capable of tasks beyond the capabilities of centralized systems.

There are various established fields of science and technology that deal to a certain extent with the intuitive notion of SI: Artificial Life, Ethology, Robotics, Artificial Intelligence, Computation, Self-Organization, Complexity, Economics, and Sociology. In all these fields there is at some level, or in some application, the need to understand, model, and predict the operation of groups of units which only by working together (in a not very structured way, and 'unaware' of the evolution of the group) carry out "intelligent" tasks. A simple illustration is provided by classic economics. In a free market economy, people trade with each other in an unstructured way: each makes independent decisions at unpredictable times and unaware of the global results. But the outcome is the solution to

a complex problem – the problem of correct pricing. Such a 'swarm' solves a problem not (or poorly) solvable by centralized control economies, thus exhibiting, in a certain sense, a high form of "intelligence".

From this example it is clear that the intuitive notion of SI is easy to grasp. But it is not so easy to make it less vague and more precise and quantitative, i. e., to make it a useful working concept for science and technology. That the concept of SI is not easy to quantify follows from the difficulty of defining several of the key components of the intuitive notion of SI (Definition 1). First, "intelligence" is a notoriously ambiguous concept. Second, "randomly" is also not easily defined and quantified. Third, "only collectively" must be specified in terms of the critical number of agents required for the emergence of SI. In what sense is a unit 'simple' or, 'un-intelligent', and the task carried out 'complex' or, 'intelligent'? Fourth, "unknowingly" implies that the global status and the goal of the swarm are, at least to some extent, unknown to the single agents. Which algorithms and communication schemes result in tasks carried out 'unknowingly' by the agents?

Because of these difficulties, in this article we first use the aforementioned (Definition 1) intuitive notion of SI to describe the current main areas of studies considered to be SI. This will provide an overview of the current status of the field; it will make it possible to quantify the four vague concepts ('intelligence', 'randomly', 'collectively', 'unknowingly') and to reach a more sharply defined concept of SI. From this, we will be able to see more clearly the limitations of SI, and so, its realistic potential for future applications.

In this article, the main areas of SI studies are described by making three very broad distinctions: (1) scientific interest vs. technological interest (Sects. "Biological Systems"–"Definition of Swarm"); (2) standard mathematics vs. cellular computational mathematics (Sects. "Standard-Mathematics Methods"–"Cellular-Computing Methods"); (3) synchronous operation vs. asynchronous operation (Sects. "Randomness in Swarm Intelligence"– Swarms of Intelligent Units"). These distinctions in turn will provide a guide to clarifying the four vague concepts in the intuitive definition of SI (Definition 1) and, thus, a conceptual orientation for future studies and applications; they will also provide criteria for evaluating the promise of SI to solve complex problems that traditional approaches cannot.

Focusing on the first distinction (scientific vs. technological interest), the main scientific interest in SI originated with the work of biologists studying insect societies [3]. The main technological interest originated with roboticists trying to design distributed robotic systems [1,2]. A valuable reference on the development of SI is [3], dealing, in parallel, with these two interrelated interests.

Biological Systems

Probably the best known and seminal biology experiment in SI is the 'double bridge' experiment by Goss et al. [11]. While studying the foraging of ants they observed that if ants, starting from a point *S*, could reach food at a point *F* via two paths of different lengths, the ants would at first choose one of the two paths randomly; but after some ants had returned from *F* to *S*, more ants would choose to go from *S* to *F* via the shortest path; and eventually practically all the ants would choose the shortest path (see Fig. 1).

The key insight was the realization that the ants were finding the best path via *stigmergy*, that is, by *communicating through modification of the environment*. Ants are blind but they communicate chemically via pheromones. By laying pheromones along the path when returning from the food source F, the ants effectively marked the shortest path by laying more pheromones on it. After that, the ants that would start from *S*, would choose the path marked by more pheromones, i. e., the shortest path.

Thus it was observed and understood a method of self-organization and a method to solve a nontrivial problem by a form of collective intelligence, with many of the elements of the intuitive definition of SI given above.

Later, Dorigo [12] realized that this method could be abstracted and generalized to design algorithms that rely on 'artificial ants' to solve much more complex problems (see Sect. "Swarm Optimization"). Thus the close connection between biological studies of SI and its potential for technological application was first clearly demonstrated.

Many other experiments in ants and other social insects have confirmed the potential for developing bio-inspired algorithms [3,13,14,15]. For actual insect societies, various ant algorithms have been applied to model

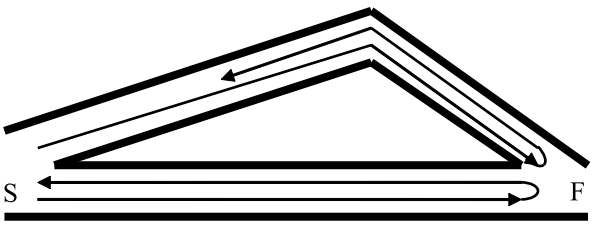

Swarm Intelligence, Figure 1
Illustration of the double bridge experiments. The ants following the shorter path (the lower path) return to the source before the ants which have taken the longer path. In this way the shorter path has a higher density of pheromones; as a result, ants starting at *S* will now prefer the shorter path

tasks such as: division of labor, cemetery organization, brood care, carrying of large objects, constructing bridges, foraging, patrolling, chaining, sorting eggs, nest building and nest brooming. Social insects constitute 2% of insects, half being ants. Besides ants, termites, bees, and wasps, have been observed to exhibit some forms of SI behavior as in the aforementioned tasks.

Apart from insects, many other biological groups exhibit behavior with some of the features of SI, such as flocks of birds and schools of fish. A seminal model of artificial flocks and schools of fish was proposed by Craig Reynolds in 1987 [16].

It is a computational model for simulating the animation of a group of entities called "boids", i. e., it is intended to represent the group movement of flocks of birds and fish schools. In this model, each boid makes its own decisions on its movement according to a small number of simple rules that react to the neighboring members in the flock and the environment it can sense. The simple local rules of each boid generate complex global behaviors of the entire flock. In addition to being used to simulate group motion in a number of movies and games, this flocking behavior has been used, e. g., for time-varying data visualization [17].

In studying these biological systems several concepts of relevance to SI were recognized. They can be summarized as:

(1) Multiple communication (of various types) among units;
(2) Randomness (random fluctuations);
(3) Positive feedback, to reinforce random fluctuations;
(4) Negative feedback for stabilization.

Of the various types of communication, we have already noted *stigmergy*, i. e., *indirect communication* by modification of the environment. On the other hand, *direct communication* may occur *unit-to-unit*, contact being a special case, (e. g., via antennae or mandibles in insects) or by *broadcasting* within a certain range (e. g., acoustically or chemically). The type and specific mode of communication has been found to be critical to the task performed, as e. g., in what types of patterns are formed [18].

Finally, the most basic lesson from biological studies of SI is that biology has found solutions to hard computational problems, and that the design principles used in doing this, can be imitated.

Robotic Systems

The actual realization of SI systems as collections of robots is a very hard problem; in fact, it is quite difficult to make even small groups of robots perform useful tasks [19,20]. Making even a single mobile, autonomous robot work in a reliable way (even in simplified environments) is a complex project. Often the technical problems with small groups of robots are quite far from the goal of SI, so there is not much reason to use the term 'swarm'. Terms such as "collective robotics", "multi-robot systems", and "distributed autonomous robotic systems" are generally, and more appropriately, used. But, whenever the tasks carried out by these robotic systems become scalable to large numbers, the term "swarm robotics" is appropriate and, in fact, it has come into use. More typically, "swarm robotics" simply describes the design of groups of robotic units performing a collective task. Each robotic unit cannot solve the task alone; and collectively the robotic units try to accomplish a common task without centralized control.

As for any robotic system in general, each robotic unit, and the group as a whole, require design of: mechanics, control, and communications. The emphasis of current research, in relation to swarm robotics, is primarily on the latter two: (1) effective *communication* among the robot units [21], and (2) effective *control* via decentralized algorithms and robustness [22].

(1) Research in robotic *communication* has become important with the growth of wireless communication networking and the lower cost of building robotic units, thus opening a new range of applications for multi-robot systems with networking capabilities, including swarm robotics. In fact swarm robotics provides the common ground for convergence of information processing, communication theory and control theory [21].
(2) Research in *control* of robotic swarms is particularly important to guarantee the stability of the swarm since the swarm does not have a centralized control. The stability of a swarm is a special case of the general problem of distributed control. In fact, after swarm robotics algorithms for task implementation have been devised, the practical realization requires stability, and robustness, i. e., proper control. Swarm control presents new challenges to robotics and control engineers: various types of controllers for swarms are currently being investigated, e. g. neural controllers [23].

The control theory example coming closest to the problem of swarm control is perhaps that of 'formation' control, e. g., the control of multi-robot teams or autonomous aircrafts or land or water vehicles. These studies, when ex-

tended to decentralized systems, lead to consider problems of asynchronous stability of distributed robotic systems and swarms. [24]

Although much progress has been made in swarm robotics, the application of SI algorithms is still underdeveloped; one reason is that often the SI behavior emerges only above a critical number which is too large to make the construction of a robotic swarm practical, because too complex or expensive. Investigations of this type are thus generally carried out by simulation [22,25].

These simulations are specialized methods of swarm robotics. An example is 'executable models' [25] which can run in simulation or on a mobile robotic unit and can execute all aspects of a robotic unit behavior (sensing, information processing, actuation, motion), i. e., they fully represent how perception is translated into action by each robotic unit. They can be used to test either experimentally or in simulation how a group of robotic units behaves. Executable models are an evolution of early protocols (so called 'behavior-based' protocols) designed around the subsumption architecture [26]. Behavior-based protocols have now been generalized into Markov-type methods, i. e., protocols where the transitions between the possible states of a robotic unit are specified by a probability transition matrix as in Markov processes [27].

Looking at applications, swarm robotics has by now accumulated a collection of standard problems which recur often in the literature. One group of problems is based on *pattern formation*: aggregation, self-organization into a lattice, deployment of distributed antennas or distributed arrays of sensors, covering of areas, mapping of the environment, deployment of maps, creation of gradients etc. A second group of problems focuses on *some specific entity in the environment*: finding the source of a chemical plume, homing, goal searching, foraging, prey retrieval, etc. And a third group of problems deals with more *complex group behavior*: cooperative transport, mining (stick picking), shepherding, flocking, containment of oil spills, etc. This is not an exhaustive list: other generic robotic tasks, such as obstacle avoidance and all terrain navigation, are also swarm robotics tasks.

One envisioned application of swarm robotics which has received media attention is the ANTS (autonomic nanotechnology swarm) project by NASA [28,29]. This project envisions *nanobots* (i. e., a swarm of microscopic robots) operating autonomously to form structures for space exploration. The idea is inspired by the example of insect societies; it envisions a technology of self-similar, reconfigurable, miniaturized robotic units with a software strategy to endow the swarm with 'intelligence'. This ANTS 'intelligence' is at an intermediate level between traditional Artificial Intelligence (i. e., highly symbolic) and reactive responses 'intelligence', i. e., intelligence without internal representation [30]. Its basis is a software construct called a 'neural basis function' to bridge the gap between lower and higher level functions and to be capable of autonomous behavior. In one potential implementation, a Saturn autonomous ring array would launch 1,000 spacecraft with specialized instruments—organized as 10 subswarms—to perform in situ exploration of Saturn's rings to understand their constitution and formation.

The European Union sponsored swarm robotics project [31,32] was completed in 2005 after demonstrating several critical tasks, such as: autonomous self-assembly, cooperative obstacle avoidance, and group transport. For this project a new type of robot, called an s-bot was developed. A swarm-bot is any device composed of more than one s-bot, which is a mobile robot unit capable of connecting or disconnecting from another s-bot. S-bots have relatively simple sensors and motors and limited computational capabilities. Using their grippers, s-bots can assemble into a swarm-bot that is able to solve problems too difficult for a single s-bot. For example, a swarm-bot could transport an object too heavy for a single s-bot. [33]

Although swarm robotics could be defined as the robotic implementation of SI (Definition 1), so far, as noted, this implementation remains a distant goal. Meanwhile, concepts from SI can be usefully applied to collections of cooperating robots. Thus, referring to the intuitive notion of SI (Definition 1), the robotic swarm, can be characterized by the type of algorithm and of (decentralized) control, the number of units above which new behavior emerges, the communication method (range, topology, bandwidth), the processing and memory capability of each unit, and the heterogeneity of the group.

Swarm robotics, besides the implementation of SI algorithms, includes the material (mechanical and electronic) realization of the units comprising the swarm. This is, as noted, an arduous task which often becomes the emphasis of research in swarm robotics. But, as it was emphasized in the early years of SI, even if the material construction of the swarm were accomplished, SI algorithms would remain as the most difficult challenge for swarm robotics. This can be easily seen from the fact that a 'robot' swarm with very advanced hardware is already available for experimentation: it is a group of human beings. Each person could be limited in a controlled way, e. g., by allowing each person to handle only a specific device according to specific rules. Algorithms to make such a swarm doing intelligent tasks are in the province of SI, but they are not simple to devise, as common experience shows.

Artificial Life Systems

The areas of *Self-organization*, *Complexity* and *Artificial Life (or A-life)* are all older and broader fields than SI and overlap with it to various extents.

A-life is conceptually placed somewhere between science and technology, and between biology and robotics. During the mid-1980s attempts at imitating living systems with machines grew rapidly and resulted in the formation of the research field of "Artificial Life" [34]. A-life investigates phenomena characteristic of living systems primarily through computational and (to a lesser extent) robotic methods.

Its scope is wide, ranging from investigations of how life-like properties develop from inorganic components to how cognitive processes emerge in natural or artificial systems. It includes research on any man-made systems that mimic the characteristics of natural living systems. By this criterion it includes SI, but actual, current A-life research is not much focused on SI; rather it focuses on origin and synthesis of life; evolutionary robotics; morphogenesis, learning, etc.

The basic theories at the foundation of A-life, and of relevance to SI, are the theories of *self-organization* and *complexity*. A-life studies systems which are typically characterized by many strongly coupled degrees of freedom. Systems of this type are more generally investigated within the science of *complexity* which began to be an active field of research in the early '80s. It is multidisciplinary and it investigates physical, biological, computational, and social science problems, including a vast range of topics [35] from environmental sciences to economics as it is clear from the content of this Encyclopedia. One basic feature that these systems have in common is the emergence of complex behavior from simple components, a notion we also find in SI.

In regard to self-organization, we note that, as many systems in nature, A-life systems may start disordered and featureless, and then spontaneously organize themselves to produce ordered structures, i. e., they self-organize. The theory of self-organization, going back to the 1950's [36], grew out of a variety of disciplines, but mainly from thermodynamics, non-linear dynamics and control theory. Self-organization can be defined as the spontaneous creation of a globally coherent (i. e., entropy lowering) pattern out of local interactions – a concept also relevant to SI.

Because of its distributed character, self-organization tends to be robust, resisting perturbations. The dynamics of a self-organizing system is typically non-linear, because of feedback relations between the components. Positive feedback leads to fast growth, which ends when all components have been absorbed into the new configuration, leaving the system in a stable, negative feedback state. Non-linear systems have in general several stable states, and this number tends to increase (bifurcate) as an increasing input of energy forces the system away from its thermodynamic equilibrium. To adapt to a changing environment, the system needs a variety of stable states that is large enough to react to perturbations but not so large as to make its evolution uncontrollably chaotic. The most adequate states are selected according to their fitness, either directly by the environment, or by subsystems that have adapted to the environment at an earlier stage.

Formally, the basic mechanism underlying self-organization is the (often driven by randomness) variation which explores different regions in the system's state space until it enters an attractor. This precludes further variation outside the attractor, and thus restricts the freedom of the system's components to behave independently. It is equivalent to the decrease of statistical entropy that defines self-organization.

It is useful to keep this brief sketch of self-organization theory in mind as we proceed in describing SI, since the concepts in the theory of SI are evolved from a combination of concepts of self-organization and computation.

Definition of Swarm

After having looked, in the previous three sections, at actual robotic, biological, and A-life systems and ideas of complexity and self-organization related to SI, we can return to the intuitive definition of SI (Definition 1) and make it more quantitative.

The intuitive notion consists of four elements: SI is "intelligence" achieved "collectively", "randomly", and "unknowingly". An elementary swarm retaining these four elements can be defined as

Definition 2 Elementary Swarm An ordered set of N units described by the N components v_i $(i = 1, 2, \ldots, N)$ of a vector $\mathbf{v}$; any unit i may update the vector, at any time t_i, using a function f of K_i vector components. $\forall i \in N\colon \mathbf{v}_i(t+1) = f(v_{k\in K(i)}(t))$

The Elementary Swarm describes an internally driven "collective" action. External input may be added in f. "Randomness" is built in the updating times. The evolution occurs "unknowingly" since the units have no processing capability. The Elementary Swarm can be generalized so that randomness appears also in the parameters of the function f. A further generalization is obtained by letting each vector component to be not just one number but a set of parameters.

Hereinafter we call '*Swarm*' (capital S) any system capable of SI. It is worth noting that even in Swarms more general than the Elementary Swarm, the modeling is assumed restricted in such a way that *no unit is capable of computing the Swarm's next global state,* (see also Sect. "Swarms of Intelligent Units"). Finally, "intelligence" is expected to be achieved by running appropriate algorithms via the updating function f. If and how this is going to be possible requires a more mathematical discussion, which is the subject of Sects. "Standard-Mathematics Methods"–"Cellular-Computing Methods".

Standard-Mathematics Methods

The science of biological swarms and the engineering of robotic swarms, as well as research in A-life relevant to SI, have progressed by using a broad range of mathematical techniques. All these techniques can be classified in two main groups: (1) '*standard-mathematics' methods* and (2) *cellular-computational methods*.

By *standard-mathematics methods* (SMm) we mean any method that is based on the standard tools of applied mathematics and computations based on standard (Von Neumann) computer architectures. Examples are methods in differential equations, stochastic techniques, linear systems, and optimization. By *cellular-computing methods* (CCm) we mean highly parallel and local computational methods, with simple cells as the basic units of computation, typically carried out on ► Mathematical Basis of Cellular Automata, Introduction to (CA) [37].

These two mathematical approaches reflect two distinct trends in the evolution of SI research, as described below. We consider first (Sects. "Swarm Optimization"–"Limitations of Standard-Mathematics Methods") the approach to SI based on SMm, since the greatest number of significant results in the area of SI has been obtained, so far, by standard-mathematics methods; specifically in the areas of *optimization* and *non-linear dynamics*. We consider them in turn in the next two sections.

Swarm Optimization

Optimization is by far the largest research area associated with SI. This is due to two extremely successful optimization methods, whose origin is related to models of SI. The two methods are the Ant Colony Optimization (ACO) [12,13] and the Particle Swarm Optimization (PSO) [4,38]. Both ACO and PSO, originated in the early nineties and have resulted in hundreds of applications based on variations of the original algorithms. So much so that the field of "Swarm Optimization" could stand alone, apart from its relation to SI with which it is sometimes even identified. A thorough and recent description of swarm optimization techniques is in [6]. Here only the key concepts of swarm optimization are reviewed, as they relate to SI.

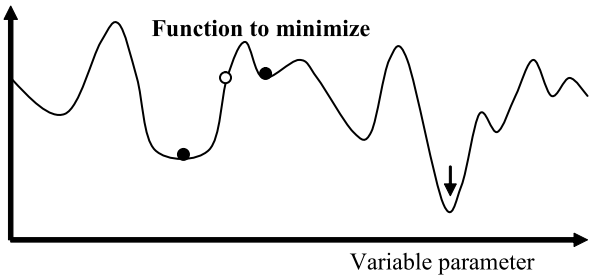

Swarm Intelligence, Figure 2
Simplified illustration of the typical problem encountered in optimization. Starting from the value represented by the *open circle* and varying the parameter continuously the algorithm will find one of the nearest local minima (*black circles*) rather than the global minimum (indicated by the *arrow*)

In PSO and ACO, as in any optimization methods, a function must be optimized, e. g., minimized. To find the minimum of the function, the variable is changed in a systematic way – the optimization method. Generally, the variable spans a multidimensional space. The search for the global minimum is nontrivial since the function may have many local minima, and the search could end into one of them (see Fig. 2). Various techniques to avoid this trapping have been developed by using some degree of randomness in the search strategy. For example, simulated annealing [39] was developed to overcome the limitations of non-random methods, e. g., the gradient descent [40]. PSO and ACO belong to this class of optimization techniques that make use of randomized searches.

Particle Swarm Optimization (PSO)

PSO was developed by Kennedy and Eberhart in 1995 [38] inspired by the social behavior of bird flocking and fish schooling [16].

In PSO, the position of each unit of the swarm is a point in the variable space of the function to be minimized. Every unit tries to reach the position corresponding to the minimum of the function. Each unit is assumed to know the global minimum value of the function, and to detect the value of the function at its location as well as the value of the function at the locations of a group of neighbors. The size of the group of neighbors is a parameter and could be the whole swarm.

The algorithm is, schematically, as follows. The units are initially in random locations. Every unit moves in

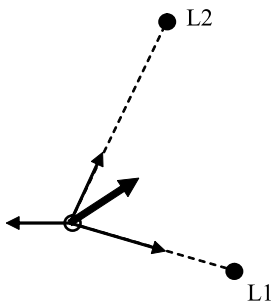

Swarm Intelligence, Figure 3
Illustration of the velocity update mechanism in PSO. A unit (*white circle*) originally moving to the left, changes its velocity by adding the two components in the direction of L1 (the unit's best location so far) and of L2 (the neighbors' best location so far). The magnitude of these components is determined by random weights. The new velocity (*heavy solid line*) generally tends to be in the direction of the global optimum

the variable space and remembers the location L1 where, among the locations visited so far, the function had minimum value. It also remembers the location L2 where, among the locations visited by all its neighbors, the function had minimum value.

At each time step, each unit calculates its distances from L1 and from L2; forms a weighted average of L1 and L2 using random weights; and changes its velocity in proportion to this weighted average (see Fig. 3). As a result every unit tends to move toward the location of the global minimum by taking advantage of its and its neighbors' knowledge. The process stops either when a unit is sufficiently close (by a chosen tolerance) to the location of the global minimum or when a chosen maximum number of iterations has been run.

PSO belongs to the category of stochastic, population-based algorithms, such as, e. g., genetic algorithms. The PSO's great merit is its simplicity. In many cases it outperforms genetic algorithms. Similarly to all optimization algorithms of this type, PSO convergence relies on the use of heuristics; and convergence does not mean convergence to the optimum. In fact, the basic PSO does not guarantee convergence even to a local minimum. The basic PSO is also inefficient in dynamic optimization problems, i. e., problems in which the optimum location changes. However, variations of the basic PSO have been proven to have improved performance in dynamic problems and to be capable of convergence to local minima.

Many improvements of the PSO basic algorithm have been developed and applied successfully to a wide range of optimization problems: continuous and discrete, constrained and unconstrained, single and multi-objective, static and dynamic. Specific applications cover just about all areas of applied optimization. The main classes of applications have been in areas such as:

(1) Neural networks (training, supervised and unsupervised learning, architecture selection, etc.);
(2) Game learning;
(3) Clustering;
(4) Design (aircraft wings, antennas, circuits);
(5) Scheduling & Planning (maintenance, traveling salesman, power transmission, etc.);
(6) Controllers (flight path, air temperature, power stabilizers, etc.);
(7) Data mining. For more detail see, e. g., [6].

Ant Colony Optimization (ACO)

The key idea of ACO [12,13] is an abstraction and generalization of the Goss et al. [11] two-path experiment with ants. A first generalization of the two-path problem is finding the shortest path between the starting point S and the final point F when between S and F there are many possible paths.

The two-path problem can be represented by a graph with 3 vertices (S, A, F) and three arcs $(S \to A,\ A \to F,\ S \to F)$. The short path is $S \to F$ and the long path is $S \to A \to F$.

If we add another vertex B, and make the graph complete (i. e. we join with an arc each vertex to every other vertex) we obtain five possible paths: $S \to F$, $S \to A \to F$, $S \to B \to F$, $S \to A \to B \to F$, $S \to B \to A \to F$ (see Fig. 4). The idea is easily generalized to a complete graph with more vertices. The number of paths increases exponentially so checking the length of all the possible paths becomes computationally unfeasible for a large enough number of vertices.

A number N of 'ants' start at vertex S choosing randomly which vertex to go next. At every new iteration, each ant decides which arc to traverse next, with probability proportional to the amount of pheromone on the arc relative to the total amount of pheromone on the possible arcs that the ants could choose.

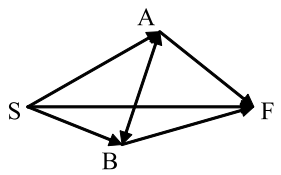

Swarm Intelligence, Figure 4
The five possible paths from S to F: $S \to F$; $S \to A \to F$; $S \to B \to F$; $S \to A \to B \to F$; $S \to B \to A \to F$

After an ant, k, reaches the destination F, the length of its path L_k is remembered (if an ant reaches the destination via loops, L_k is calculated after removing the loops). The ant retraces exactly the path (without loops) and deposits pheromones on the arcs in proportion to $1/L_k$.

In this way, the marking (with pheromones) of the paths by all the ants modifies the graph so that the probability of any ant taking the shortest path at the next trip from S to F increases. Eventually all the ants will follow the same path, if the algorithm converges, and the path will be the shortest if the convergence is to the global minimum. As in most optimization methods, this is never guaranteed.

The idea of the original ACO algorithm [12] adds to the foregoing sketch of the basic model three more elements. First, each arc has an a priori propensity to be traversed (regardless of pheromone content); second, each ant keeps in memory a tabu list of arcs not to be traversed, to avoid loops; and third, the pheromones evaporate at a given rate.

ACO key insight is the application of the concept of stigmergy to stochastic optimization. The ants communicate by modifying the environments (graph) and act probabilistically on the basis of the modified environment.

Many variations of the basic ACO algorithm have been proposed and implemented. The many variations take advantage of specific knowledge about the specific problem, i. e., they use heuristics, e. g., by setting the a priori propensity of traversing an arc, or by setting the evaporation rate.

ACO eventually resulted in a meta-heuristic which is a strategy for designing ACO heuristics. Various ACO-based metaheuristics have been developed. Similarly to PSO, ACO algorithms have been applied to all the basic types of optimization problems: continuous and discrete, constrained and unconstrained, single and multi-objective, static and dynamic. The first application of ACO was to the Traveling Salesman Problem, which is an NP -hard combinatorial optimization problem, and it is the most frequently attacked problem using various ACO heuristics. The main classes of other applications are to problems of

(1) Ordering (scheduling, routing);
(2) Assignment (Neural network training, image segmentation, design);
(3) Subsets finding (maximum independent set);
(4) Grouping (clustering, bin packing).

Clearly " swarm optimization" successfully uses concepts from the general notion of SI, but optimization is not in itself a necessary characteristic of SI. In fact, many tasks actually or potentially carried out by swarms are not-optimal in any sense.

Non-Linear Differential Equations Methods

One fruitful approach to modeling swarms has been to treat each individual as a discrete particle. These "individual-based" models have been employed in quite a few biological and mathematical studies. They are based on simple rules of motion for each individual, involving some combination of self-propulsion, random movement, and interaction with neighboring organisms. The models typically take the form of coupled ***non-linear*** difference or differential equations, which may be stochastic or deterministic, depending on the particular features of each model. Numerical simulations have revealed collective behavior. But a main disadvantage of such models is that, for realistic numbers of individuals, analytical results for the collective motion are difficult or impossible to obtain. It is worth mentioning that some progress has been made in obtaining analytical results for stationary groups. In [41], a discrete model was formulated, and a Lyapunov functional was used to successfully predict an equilibrium state of equally spaced organisms. However, analytical (nonstatistical) descriptions of nonequilibrium states in discrete swarm models are few.

Other investigations of swarming have been carried out in a continuum setting, in which relevant quantities are described as scalar or vector fields. This approach goes back to 1980; reviews are provided in [42]. Continuum models may be constructed a priori or by coarse-graining a particle model. In general, continuum models provide a convenient setting in which to study large populations, since one may apply machinery from the analysis of *partial differential equations*. In the context of swarms, the focus has generally been on models in which the population density satisfies a convection-diffusion equation ensuring that the population density is conserved while individuals travel with a set average velocity. Recent models of this type [43] can predict, e. g., whether a population aggregates or disperses, the regions of aggregation, and length scales of the density patterns.

Limitations of Standard-Mathematics Methods

In describing the SI investigations in the previous two sections, we have encountered the concepts of *optimization* and *non-linearity* (and earlier, in Sect. "Artificial Life Systems", *complexity* and *self-organization* arising from nonlinearity). To see more precisely the relation of these four concepts to SI we refer back to the definition of Elementary Swarm (Definition 2). Note that, by this definition, the Swarm is in principle capable of optimization, complexity and self-organization. In fact, it is clear that a Swarm is a *self-organizing* system, by definition, and that, depend-

ing on the choice of the updating function f, the pattern formed by the swarm components might, in principle, achieve high *complexity*.

As for non-linearity, it is convenient to think of f as a function of a function,

$$f = f(g(K_i)$$

where g represents the dependence from the neighbors. While g maybe linear or non-linear, the function f, representing the mode of updating, will typically be non-linear and/or probabilistic. In spite of this, the evolution of the Swarm only in special cases can be modeled by standard *non-linear dynamics* as studied via nonlinear differential equations (as we have seen in the previous section).

In regard to *optimization*, the Elementary Swarm can be easily designed to be an optimizer. In fact, if the updating is sequential, and f and v_i are chosen appropriately, one obtains a simple PSO system.

Thus, the Elementary Swarm describes a simple but powerful system capable, in principle, of self-organizing, and to produce complex structures and optimal solutions. On the other hand, even the Elementary Swarm is more general than these properties; a Swarm is not restricted by the notions of optimization or non-linear dynamics (and self-organization or complexity tied to non-linear dynamics). All these can be properties of the swarm but none is a requirement for SI. To find out what SI can do that is beyond what we have described so far, we must look at the computational capabilities of swarms. And for this we need to look at cellular-computing methods since standard-mathematics methods are ill suited to deal with computation.

Cellular-Computing Methods

Some of the first studies in SI were based on computational models, and, more precisely, on distributed algorithms applied to robotic units [44,45,46].

In these computational models, the swarm was developed as an evolution from a distributed system of processors, as follows. In *distributed computing* the algorithms are designed for a 'static' set of processing units, where, 'static' is meant literally as 'not moving'. For illustration, if a set of CPU's, computing in a distributed way, via wireless communication, started moving around, this system would look very much like a robotic swarm. In fact, referring to the intuitive notion of SI (Definition 1), all points of the SI definition would be satisfied by such a dynamic, distributed computing system provided the CPU's had, in some sense, limited capabilities.

The main point is that this distributed computing swarm differs from the robotic, and non-robotic, swarms described in the previous sections (Sects. "Standard-Mathematics Methods"–"Limitations of Standard-Mathematics Methods") in that the *intelligent task of the swarm is now seen as a 'computation'*. And this focus on computation leads us now to consider the other broad set of techniques used in SI research, i. e. techniques based not on standard-mathematics but on *cellular computing*.

Cellular computing differs qualitatively from the standard Von Neumann computing architecture. The latter is based on one complex processor that sequentially performs, at each time step, a single complex task. In contrast, in *cellular computing*, a very large number of simple processors (cells) are the units of computation. They compute (typically) in parallel with local connections between cells. The qualification 'simple' can be made precise by requiring, e. g., each cell to be a 'finite state' machine. Cellular automata are the most obvious examples of cellular computing systems but cellular computing applies to many other systems as well [47].

By definition, then, cellular computing contains several of the features of SI. The Elementary Swarm of Definition 2 can be regarded as performing a cellular computation. And in fact, cellular computing has been used extensively in A-Life studies, including systems with strong relation to SI [34]. Cellular computing systems offer SI something that the SI systems described in the 'standard-mathematics' Sects. "Standard-Mathematics Methods"–"Limitations of Standard-Mathematics Methods") lack, i. e., a clear characterization of *intelligent task*.

Intelligence as Universal Computation

Intelligence is an ambiguous concept, escaping a unique definition [48]. By identifying 'intelligence' with 'computation', the concept is restricted, but, at the same time, it can be made precise. In fact, in SI we define intelligence unambiguously as the '*ability to carry out universal computation*'.

Universal computation (or universality) is the property of a computer system (or language) which, with appropriate programming, can be made to perform exactly the same set of tasks as any other computer system (or language).

Universal computation (i. e., the ability to emulate a universal computer), is essentially the limit of any model of computation (Church–Turing thesis) [49]. It was first proven by Turing in 1936 that no system can ever carry out explicit computations more sophisticated than those carried out by a Turing machine. Subsequently, universal-

ity has been found to be a widespread property of many cellular computing systems [50].

One of the first cellular computing systems shown to be capable of universal computation, is Conway's game of life [51]. This CA is also the prototypical example of A-life systems. And it is also an example of the strong connection between universal cellular computing and bio-inspired systems.

More recently, a large number of simple cellular computing systems have been found to be capable of universal computation [50]. Many of these systems are CA, or related systems, using very simple rules of evolution with local interactions. And so they are useful starting points for modeling SI.

In particular, cellular computing is the most appropriate to endow the swarm with the property of *unpredictability*. The latter property was an original motivation for SI [1,2] and it is crucial in the task of escaping detection by a predator; it is also of importance in engineering swarms for strategic defense applications.

Unpredictability is almost a built-in property of cellular computing systems because, if one observes the rules of evolution in their raw form, it is usually almost impossible to tell much about the overall behavior they will produce.

Relations to Standard-Mathematics Methods

SMm cannot provide the swarm with the element of universal computation, which we have taken as the working definition of 'intelligence'. The only way would be to make each unit a von Neumann (i. e. standard) computing system. In a sense, this violates the notion of Swarm, since in a Swarm, by definition each unit must be 'simple'. (This point will be further clarified in Sect. "Swarms of Intelligent Units"). On the contrary, the main advantage of cellular computing systems over standard mathematics systems is the possibility of universal computation by simple units.

For this reason CCm are the natural paradigm for the understanding and designing SI systems, in spite of the fact that the approach to SI based on cellular computing has so far produced fewer so-called SI applications than the approach based on SMm. Indeed, SMm have basic limitations for modeling SI. This is because the use of SMm tends to restrict the range of tasks performable by the Swarm. And this happens because SMm typically solve problems by specifying *constraints*, i. e., conditions to be satisfied by the solution, e. g., by specifying equations. But most computational problems cannot be solved in this way.

The optimization methods described in the previous sections (PSO, ACO, etc.), illustrate the point. In these iterative methods, the key issue is what kind of changes should be made at each iteration step. Starting from a random pattern, at each step a change is made to get the pattern closer to satisfying the constraint(s). Since direct methods (e. g., gradient descent) rarely work as the pattern gets stuck into local minima, randomness in updating is added. In this way, larger portions of the solution space are sampled. The larger the changes made, the faster one can potentially approach a global minimum, but the greater the chance of overshooting. The result is that no iteration technique of this type can guarantee a solution to general combinatorial optimization problems. As we have seen, the swarm optimization methods (ACO, PSO) rely on heuristics to adjust the search and obtain often (non-optimal but) satisfactory solutions. But, in general, for the great majority of combinatorial optimization problems (e. g., the Traveling Salesman Problem [52]), no polynomial upper bound on the time complexity has been found so far. And this happens in many problems whose solution is sought by using randomness to satisfy the imposed constraints. As an example, a set of identical balls cannot be shaken into an ordered, closed-packed configuration. With extremely high probability, they lock into some configuration or another, not the optimal (close-packing) one.

This fact has *important implications* for SI. What it says is that no matter how much randomness is added to the system, it may never evolve to reach the solution specified by the constraints. Although, ultimately, constraints can be set up as a way of specifying algorithms, and hence computing, it is far simpler to specify algorithms via rules of evolution, as it is done in cellular computing.

The conclusion is that methods based on constraints and other SMm are not ideally suited for systems evolving with great complexity, and in particular they are not suitable for universal computation. Thus, if SI is to be a framework for (biological or engineered) swarms to carry out 'intelligent tasks' with the greatest generality, a methodology that allows for the swarm to carry out universal computation is necessary. To this aim, CCm are the most suitable.

Unfortunately, although CCm have many advantages over SMm for modeling SI, they address only three of the four key elements of the notion (Definition 1) of SI ('intelligence', 'collectively', 'randomly', 'unknowingly'), leaving out the element of 'randomness'. Generally, CCm operate deterministically and do not include 'randomness', as, e. g., 'swarm optimization' systems do. But this does not have necessarily to be the case. The issue is addressed in the next section.

Randomness in Swarm Intelligence

Randomness is a key element in the notion of SI (cfr. Definition 1 and Definition 2). Examples from biology justify this requirement. Randomness is not easily quantified precisely but, whatever the form and measure chosen, the point is that for swarms some form of randomness is necessary – otherwise they would fail to be models for analyzing a large class of biological systems. But what kind of randomness is essential to model these biological systems?

Randomness in the number and type of agents is not important – the agents could be strictly identical and remain in the same number. Randomness in the initial conditions is not essential either. Many swarms evolve from regular initial conditions into highly complex and random patterns. Randomness of external input from the environment is not always present, and it is certainly not a requirement for biological swarm behavior.

What about the randomness artificially added to the units, as in swarm optimization?

The randomness added to the units in PSO or ACO algorithms is modeled as originating from the random behavior of each unit. This is a plausible assumption in relation to biological systems. But the swarms in PSO and ACO are updated in an orderly (non-random) way, typically sequentially (there are also some parallel implementations [13]) whereas, in biological systems, the units update in a disordered, random fashion.

And it is this type of randomness that is both necessary in any biologically relevant model of swarms and sufficient to provide many (but not all) of the advantages of randomness in solving swarm engineering problems.

The conclusion is that the only randomness that is truly essential for SI is randomness in the times of operation of the units. Each unit has its own clock, not synchronized with other units' clocks. Other types of randomness in the behavior of the units or the environment may be required to solve specific problems, but randomness in times or operation is necessary for any biologically realistic model.

Interestingly though, many applications so far considered in the area of SI do not yet include this randomness in the models. We have already mentioned that typical optimizing swarms update sequentially; and CA systems operate largely in parallel, i. e. synchronously. Synchronous or sequential operations are by far the most common updating modes in either SMm or CCm.

The Implicit Assumption of Asynchrony Irrelevance

As noted, it is a basic fact that biological agents, apart from exceptional cases, do not operate synchronously (or sequentially) in groups. It is also a fact that people in social groups do not operate synchronously, or sequentially. If SI is supposed to model biological and social swarms, SI must be based on models that do not operate synchronously or sequentially [53].

And if biological swarms are capable of solving problems (including optimization) without synchrony (nor sequentially), as they do, then models that imitate those swarms should operate asynchronously (not sequentially).

But, as noted, the main modeling paradigms for bio-inspired algorithms, standard-mathematics and cellular computing, are either essentially sequential or synchronous.

An example from SMm is the solution of partial differential equations: they operate synchronously on every point (clearly seen in solving them numerically and iteratively). This unrealistic use of differential equations in biological processes has been pointed out, e. g., in the problem of morphogenesis [54]. The Turing diffusion-reaction model [55], being based on differential equations, implies synchronicity and central control, hence it is physically not realistic for a scale of the order of 100 cells.

In fact, synchronicity leads to realistic models only whenever the spatio-temporal resolution is high, as, e. g., for phenomena typically studied in physics. But when the units studied are complex or few enough to have a less fine spatio-temporal resolution, as in biology or human societies, synchronicity is not realistic, as it is obvious by observation.

Thus, in using synchronous methods for biological or human societies, implicitly a strong assumption is being made, i. e., that the synchronously (or sequentially) and non-synchronously (and not sequentially) obtained solutions would coincide.

But this assumption has no validity. In fact, it has been shown, for example, that CA, when running in synchronous and non-synchronous ways, normally produce totally different results [56]. This has been noted already in the nineties [57] in A-Life studies. In [57] two well-known CA were compared: Conway's "game of life" [52], and the Immune Network model. The former is a 2-dimensional CA capable of universal computation when run synchronously. But the behavior is totally different when run without synchrony: the Game of Life stops producing complex patterns and converges to a fixed point.

The Immune Network model is asynchronous and the Game of Life synchronous. The crucial factor in the different behavior of the two systems was identified as the synchronous vs. asynchronous updating. In fact, it was concluded that, in this case, asynchrony induces stability

in CA. This agrees qualitatively with studies in standard-mathematics [58].

In conclusion, the assumption that asynchrony makes no difference has been found not to be valid. Hence, asynchronous systems must be studied as such, not by using synchronous models. Moreover, different types of asynchrony yield different results, as discussed in Sect. "Asynchronous Swarms".

Asynchronous Swarms

Several cellular computing studies in the nineties [37,59] led to a variety of results emphasizing the role that different types of asynchrony play in the results. Studying asynchronous systems is complicated because, among other things, deviation from synchronicity, i. e., from the mode of updating all units in parallel at each time step, may occur in several different ways. For example, sequential updating and random updating are both asynchronous but very different.

Types of Asynchrony

Unfortunately, there is no standard vocabulary for the various types of asynchrony. Thus we use the following classification to describe the possible types of asynchrony.

Consider an updating cycle (UC), i. e., the time interval at the end of which all units have been updated at least once. Eight types of UC can be identified by the presence or absence of any of following three properties: (*S*) *Synchronicity*: more than one unit may update at each time step; (*M*) *Multiplicity*: any unit may update more than once in each UC; (*R*) *Randomness*: the updating order varies randomly for every UC (see Fig. 5).

These eight basic types of asynchronous updating, can be further specialized. For example, if all three properties are absent ($\sim S, \sim M, \sim R$), the updating is sequential. But the sequential updating order of the units can be fixed in different ways. Studies of CA have proven that the behavior differs markedly not only for the eight types of asynchrony, but even among different sequential ordering [37,56].

In [56] the ($S, M, \sim R$) form of updating has been applied to describe processes where each unit has independent clocks but the clocks have a fixed, non random frequency. This type of asynchrony is considered a good model for forest ecosystems, fire spread, and other natural and artificial systems. The results are very different when updating of the type ($\sim S, M, R$), ($\sim S, \sim M, R$), or sequential ($\sim S, \sim M, \sim R$) are applied to the same system.

In conclusion, the crucial point is that [56] the exact manner of updating can have a profound effect on overall system behavior. The implication of this is that when comparing models of natural systems or artificial multi-agent systems it must be stated which updating scheme has been used, otherwise meaningful comparison between different studies may not be possible.

In particular, returning to swarm optimization, one may ask whether in tasks such as finding the shortest path, it is realistic to apply to natural systems (such as in-

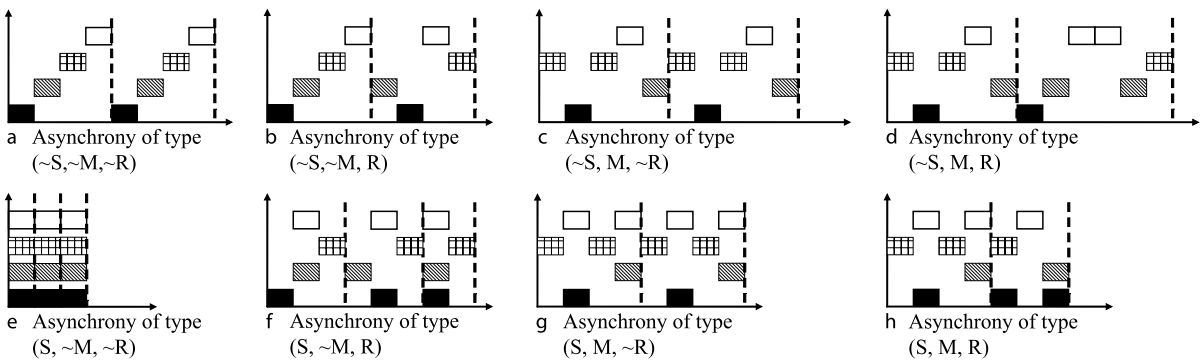

Swarm Intelligence, Figure 5

a. Asynchrony of type ($\sim S, \sim M, \sim R$); b. Asynchrony of type ($\sim S, \sim M, R$); c. Asynchrony of type ($\sim S, M, \sim R$); d. Asynchrony of type ($\sim S, M, R$); e. Asynchrony of type ($S, \sim M, \sim R$); f. Asynchrony of type ($S, \sim M, R$); g. Asynchrony of type ($S, M, \sim R$); h. Asynchrony of type (S, M, R); Illustration of the 8 types of updating, according to Synchronicity, Multiplicity, and Randomness. Four units are represented by rectangles with different patterns, from black (*bottom*) to white (*top*). The horizontal axis measures time steps in units equal to the base of a rectangle. The vertical dashed line indicates the end of an updating cycle (i. e., all units have updated at least once). The label below the horizontal axis specifies the type of updating (~ means 'not'). The standard 'parallel' and 'sequential' updating are, respectively, (S,~M,~R), Fig. 5e, and ($\sim S, \sim M, \sim R$), Fig. 5a

sect societies) swarm models which are sequential (such as swarm optimization models) or synchronous (such as models based on differential equations). While these models work effectively as artificial swarms there is no proof that they apply to natural systems, which are asynchronous.

Modeling Asynchrony by Synchronous Swarms

Because of the widespread use of synchronous methods in simulations of SI, one might wonder under what conditions a synchronous but stochastic model could be equivalent to an asynchronous one.

To answer this question, let us consider the two types of stochastic models most commonly used to model randomness in synchronously updated systems. The randomness may be included in (1) the possible outcomes of the updating function or (2) in the choice of the function applied to the updating.

Referring to the definition of Elementary Swarm (Definition 2) the two cases correspond to generalizing the updating function as follows:

$$\text{Case (1)} \quad \forall i \in N\colon v_i(t+1) = f(\mathbf{v}_{k\in K(i)}(t)\,\zeta)$$

where ζ is a random variable.

$$\text{Case (2)} \quad \forall i \in N\colon v_i(t+1) = f_{(t)}(\mathbf{v}_{k\in K(i)}(t))\,;$$

where $P[f_{(t)} = f_\gamma]$ is the probability mass function of choosing $f_{(t)} = f_\gamma$ out of a set of N_f possible functions $\{f_\gamma; \gamma = 1, \ldots, N_f\}$.

Case (1) is typical of probabilistic CA, and it is also the method used in PSO. In these systems the state vector, at each time step, evolves according to a fixed rule which produce a new state vector from the previous one. The rule is based on the state of the neighbors of each unit and does not change from step to step but the outcome of the rule is probabilistic.

Case (2) is what is done for example in probabilistic Iterated Function Systems [23]. In probabilistic Iterated Function Systems a vector evolves via a set of maps (a map is a function whose domain and range coincide); at each time step a map is chosen, probabilistically, from a set of possible maps.

In either cases (1) or (2), the updating scheme fails to model the actual time evolution of natural systems not so much because the updating are applied synchronously but because the randomness is applied *collectively*, i. e. to all the units in the same way. On the other hand a synchronous algorithm realistically simulating independent random updating can be run as case (2) applied *individually* to each unit, as follows:

$$\text{Case (3)} \quad \forall i \in N\colon v_i(t+1) = f_{(t)i}(\mathbf{v}_{k\in K(i)}(t))$$

where $P[f_{(t)1} = f_{\gamma 1},\ f_{(t)2} = f_{\gamma 2}, \ldots, f_{(t)N} = f_{\gamma N}]$ with $f_{(t)i} \in \{f_\gamma; \gamma = 1, \ldots, N_f\}$ is the joint probability mass function of each unit i updating, at time t, according to the function $f_{(t)i}$.

In the simplest embodiment of case (3), the set of possible updating functions consists only of the identity and of another function f, with probabilities p and $(1-p)$ respectively (i. e., a Bernoulli process). In such a case, every unit, at each time step, either does not update, with probability p, or updates according to the function f, with probability $(1-p)$. Running this algorithm synchronously is equivalent to asynchronous independent updating of the units in a random way – a realistic description of a random swarm. So, under these independently stochastic conditions, running a simulation synchronously, represents correctly the physical asynchronous updating of the swarm units. On the other hand, this does not change the fact that different results are obtained when using this random updating (whether simulated with stochastic synchrony or not) instead of synchronous or sequential updating.

Local Synchrony and Self-Synchronization

Another approach to dealing with the problem of random updating by the swarm units, is to explore the possibility of self-synchronization. If the swarm can self-synchronize, then all the results for synchronous swarms could be applied.

To look into this, let us return to the classification of the types of asynchrony, i. e., the SMR classification above. If the SMR properties are applied to blocks of units rather than individual units, the resulting updating orders are referred to as '*locally synchronous*'.

CA with cells organized into blocks have been investigated [60]. These CA relax the normal requirement of all cells having the same update rule. Cells within a block are updated synchronously, but blocks are updated asynchronously. They experimented with different SMR types of asynchrony and concluded that synchronous and asynchronous CA can be evolved with equivalent computational properties, but CA of the asynchronous type may require a larger number of cells [60]. Another study [61] has shown cases in which local synchronization can lead to the same outcome as with global synchronization. But how can local synchronization be achieved?

A number of schemes have appeared in which the order of updating depends on local interactions and leads

to local synchronization. In effect, what local interactions (or constraints) can do is to force a unit to wait to update until others are ready, and so this creates a local synchronization. An asynchronous CA model that can behave as a synchronous CA has been demonstrated [62]; it functions by the addition of extra constraints on the order of updating, effectively providing a type of local synchronization. Whether these methods of self-synchronization may in some cases result in realistic models of natural systems of SI remains an open question.

The Natural Asynchrony of Swarms

We have seen that the implicit assumption of asynchrony-synchrony equivalence must be rejected and that different types of asynchrony give different results. But what type of asynchrony is most relevant to SI? There is no easy answer. For example, ants work and rest; active and resting periods have an aperiodic pattern for individual ants, but for the whole colony there are synchronized periodic patterns of active and resting periods.

In spite of the difficulty of finding a clear cut answer to the question of the natural mode of SI updating, from observations of biological systems, and from local synchronization models, it may be plausible to assume that the essential form of asynchrony in SI is the randomness in the working of the individual clocks, as argued in Sect. "Randomness in Swarm Intelligence"; hence the SI asynchrony must be characterized by the presence of all three asynchrony properties, i. e., (SMR).

In conclusion, at this stage of our discourse, the Swarm remains defined as in Definition 2, qualified by SMR asynchronous updating, which hereinafter we call *'natural'* asynchrony. Note that stochastic synchronous simulations of this model can also be carried out as, e. g., in case (3) above.

The Realization of Asynchronous Swarms

So far we have established the importance and type of asynchronous models in SI, but what SI investigations using asynchronous swarms are there?

As noted in Sect. "Asynchronous Swarms", research in asynchronous models is still very limited, relative to synchronous models, and this in spite of the fact that the very first models of SI were all asynchronous, using SMm based on finite differences [46]. Explicit updating schemes in finite difference methods can also be regarded as parallel CA, thus belonging to both SMm and CCm. Investigations of asynchronicity in finite difference methods are not common [58]. Examples include a non-linear updating rule was based on a linear relation between two neighboring units [45]. A gradient type of swarm updating, was also proposed in modeling morphogenesis [54].

For swarms updating with 'natural' asynchrony, i. e., according to (SMR), a recent study [58] gives a proof of convergence to the same state as by using synchronous or sequential iterations. It was also shown that, under certain conditions, the (SMR) asynchronous updating leads to convergence while synchronous updating does not. This is another example of the advantages of randomness in allowing the swarm to reach a fixed state.

At the end of Sect. "Cellular-Computing Methods" we concluded that CCm have, for SI modeling, many advantages over SMm. The most crucial advantage is the possibility of universal computation which we took as the definition of intelligence for SI. We also noted, however, that studies based on CCm which include randomness are scarce. We described a few in Sect. "Randomness in Swarm Intelligence", especially in discussing the qualitative differences with synchronous CA and in relation to mechanisms of local and self-synchronization.

Generally these studies model relatively trivial phenomena but cannot model nontrivial phenomena such as universal computation. In fact, there are very few studies of universal computation in asynchronous CA. Significant advances have been made only recently. The first attempts were made by simulating a synchronous CA on an asynchronous CA [63] after which a synchronous model, as a Turing Machine, was simulated on the synchronous CA. However, this asynchronous CA is, in practical realization, synchronous.

Improved asynchronous CA do not rely on global synchronization but conduct asynchronous computation directly by simulating Delay-Insensitive circuits, i. e., circuits in which delays of signals do not affect the correctness of the circuit operation [64]. This method essentially uses local synchronization with undetermined exact timing between transitions. In this way an asynchronous CA, with a hexagonal cell structure, capable of universal computing has been realized [65]. Although relying on local synchronization, this type of asynchronous CA can mimic natural phenomena as, e. g., phenomena that rely on chemical reactions which occur only when the right molecules are available in the right positions at the right times.

More recently a computation-universal and construction-universal asynchronous CA has been designed [66] and used to implement self-reproducing [67,68] machines. Besides computational universality, construction universality is important in SI because it allows the swarm to be hardware reconfigurable, an important characteristic of many biological systems.

We note that the recent interest in asynchronous CA stems not directly from SI but from nanotechnology. In fact nanocomputer architectures with asynchronous updating may reduce heat dissipation, an important limiting factor in scaling down the size of computing chips [64]. In this respect, it is likely that SI concepts will play a major role in nanoscale systems.

In conclusion, the recent realization [66] of universal asynchronous CA is a major step toward the realization of true SI.

Characteristics of Swarm Intelligence

The demonstration of universal computation in asynchronous CA, amounts to a validation of the concept of SI. In fact we can now combine universal computation with the Elementary Swarm definition (Definition 2) to quantify the intuitive definition of SI (Definition 1), as

Definition 3 SI is the study of *universal cellular-computing systems updating with natural asynchrony*.

Here 'natural' means SMR-asynchrony (see Sect. "Asynchronous Swarms") or updating randomly in parallel, as in case (3) of Sect. "Asynchronous Swarms". We call it 'natural' since it models the natural mode of updating of typical biological swarms and human societies.

A few remarks about Definition 3. The four elements of the intuitive notions (Definition 1) are made precise by Definition 3: 'collectively' and 'unknowingly' are inherent in the structure of cellular computing; 'intelligence' is in universal computation; and 'randomness' is in the natural asynchronous operation.

Definition 3 deals with CCm and may appear to exclude SMm in SI; but this is not the case. In fact, many SMm, as used in SI, can be regarded as special cases of cellular computing methods, as, e. g., are swarm optimization and iterative methods in finite differences.

Also Definition 3 does imply that every SI system must be capable of universal computation; what Definition 3 does is to establish a focus of attention for the SI area of studies and at the same time give a precise and realistic meaning to the kind of 'intelligence' aimed at in SI, rather than the often vague and exaggerated meanings given in the popular literature.

Definition 3 also indicates how SI becomes of relevance beyond biological systems and robotics. In fact SI will likely be an important concept in the future of computation. At very small scales, time delays between computational components cannot guarantee synchrony; the various components must have independent clocks, thus beginning to resemble the operation of a biological swarm. So, swarms are likely to be studied extensively in connections with nanocomputing.

The above arguments bring up the question as to whether SI is nothing more than asynchronous cellular computing. The answer is that designs of asynchronous CA, as investigated in computer engineering, are generally not models of SI. A first, basic reason is that, with one recent exception [66], asynchronous CA, do not update 'naturally'. A second, more fundamental reason, is that in most SI studies (both in natural and technological systems) **the units are dynamic**. When the units of an asynchronous CA are made mobile, a different, and more complex set of problems need to be solved due to the changing neighborhoods of each cell. These issues of dynamic reconfigurations of cells have not been addressed in asynchronous cellular computing designs and are likely to remain outside the scope of research aimed at improved computer architectures. The computational problems arising from dynamically reconfiguring cells are central in SI. We address this issue next.

Dynamics in Swarm Intelligence

In the definitions of SI and Swarm given so far (Def. 1., Def. 2., Def. 3.), there has been no mention of the dynamics of the units. But a general characteristic of the units of a swarm is that almost invariably they are mobile. In fact we have already discussed the dynamic nature of swarms in Sects. "Biological Systems" and "Robotic Systems" in relation to biological and robotic swarms. The reason we have so far omitted this dynamic character of the units from the progressively more precise definitions of SI is simply for clarity of exposition: if the dynamics is introduced after all the other elements of SI have been defined and quantified it is easier to single out its real importance.

From Definition 3, with appropriate specializations, all aspects of SI considered so far can be included in a common core of studies. The most general notion of SI is in fact that of *universal computation carried out with 'natural asynchrony' by a cellular computing system*. But we should add, *whose cells are, in general, mobile units*.

The latter qualification would be unnecessary if the description of the dynamic state could be included among the state variables of the cell. But this is not possible in computing cells since the computation depends on neighbors that change their locations. The fact that *dynamic cellular computing systems* (also called *cellular robotic systems*) are not equivalent to (and very hard to simulate by) cellular computers has been emphasized since the very beginning of SI studies [44]. Thus, we conclude by stating a definition of SI which, while remaining grounded in the

intuitive ideas (Definition 1), includes all the concepts discussed quantitatively in this article.

Definition 4 SI is the capability of universal computation carried out with 'natural' asynchrony by a dynamic cellular computing system.

As we have seen, studies in the area of SI so far have been concerned with models of collective behavior which, to some limited degree, approach SI as defined above. Even though, to date, no system with SI (Def. 4.) has been built (or designed), significant progress has been made and, from what we have seen, it is reasonable to expect that it can be done.

In fact, although there is not yet proof of universal computation carried out with 'natural' asynchrony by a *cellular robotic system* (i. e., a dynamic cellular computing system), the recent proof [66] for 'static' cellular computers indicates that this may be possible in the near future. Other future perspectives for SI are discussed in Sect. "Future Directions".

Unpredictability in Swarm Intelligence

We are now in a position to consider an aspect of SI which has been inherent to the concept of SI from its inception, i. e., the unpredictability [1,2] of the Swarm.

The unpredictability of the Swarm agrees with the common intuition that it is usually difficult to predict what a program will do by reading its code, and the more so the lower the level of the language used by the program. More precisely, a Swarm, like other universal computers, may be impossible to predict in the sense that even if one knows the rules of evolution and an initial state, it can still take an irreducible [50,69] amount of computation to actually predict future states. Furthermore, the unpredictability of the Swarm is of a more general character than that of any universal computer because of the randomness inherent in its evolution and because of its dynamics.

The unpredictability of a Swarm by a Von Neumann universal computer has been argued in [1,10] on the basis of its dynamics. The unpredictability of a Swarm by a cellular automaton has also been discussed [1,10]. Although unpredictability is difficult to quantify, it is generally engineered by adding randomness to the system as in camouflage and cryptography. Also, in animals, randomness and dynamics are used by a herd to avoid predators by becoming unpredictable. And so in team sports, such as soccer, unpredictability by the opponent is usually achieved by a combination of randomness and dynamics.

Therefore, intuitively we may conjecture that among systems capable of universal computation, a Swarm, because it operates with randomness and dynamically, would be *the least predictable*.

Swarms of Intelligent Units

We can now consider collections of intelligent units, i. e., such that each unit is capable of universal computation. These seem excluded from the definitions of SI given so far – a key characteristic of SI is that intelligence is an emergent property, happening only above a certain critical number of units, and not a property of any of the individual units. On the other hand, some of these groups are often included in broad considerations of SI [4] as applied to human societies. Under what conditions can these groups be regarded as swarms?

A simple answer to this question runs as follows. Consider the special case of Definition 4, when the cells are not finite state machines but universal computing units. As long as the task at hand cannot be accomplished by a single unit, but only by more than a critical number of units, the system operates as a swarm. That this can be the case is supported intuitively and from computational considerations, as follows.

Intuitively, we may refer back to the example of the free market model mentioned as illustration of SI at the beginning of the article. Each individual contributes only as a trader; the 'computation' of the market price is done by the swarm collectively and could not be done by any individual agent. The point is that each unit, albeit 'intelligent', uses only a fraction of its capability, i. e., the trading ability, thus operating, effectively, in a restricted, 'non-intelligent' capacity.

Computationally, we have noted in discussing unpredictability that, in spite of theoretical computational equivalence among universal computers, the capability for one universal computer to predict another is limited. And, in fact, at the end of Sect. "Swarms of Intelligent Units", we put forth the conjecture that a Swarm is the least predictable universal computer.

But even if this conjecture were not true, it still makes sense to think of a case when a Swarm of universal computers is unpredictable by any of the units comprising it, as it has been argued for the case of units capable of universal computation with Von Neumann architecture [1,10]. In this sense we may think of a swarm of universal computers as being more capable than any one of its units, in spite of the fact that the Swarm and any of its units are computationally equivalent. For these reasons, it makes sense to apply, under appropriate conditions, the notion of SI also to human societies.

We may give then a more complete definition of Swarm by adding the characteristic of unpredictability of the Swarm by its units.

Definition 5 SI is the capability of universal computation carried out with 'natural' asynchrony by a dynamic cellular computing system, none of whose cells can predict the computation done by the Swarm.

The latter specification is obviously redundant for common cellular computing systems but it is useful to exclude from SI trivial cases of human social activities; for example, activities of human groups whose association cannot be proven to solve a problem that could not have been solved by any individual alone.

Many popular interpretations of SI have appropriated the label SI to refer to almost any trivial human group activity, such as brainstorming. In such activities often there is no way of proving that the final output of the group could not have been predicted by one of the members of the group. Ultimately, however, these considerations require an understanding of the relation between human thinking and computation, and thus fall beyond the scope of this article.

Similarly, beyond the scope of this article and SI, fall studies of multiagent systems in Artificial Intelligence [70]. In the Artificial Intelligence area, the emphasis on multi-agent systems is in finding decision algorithms, i. e. 'agents', for open environments in which these agents must operate robustly and rapidly, i. e. 'intelligently'. Generally, however, the problem of collective decision making, organization theory, distributed reasoning and distributed Artificial Intelligence is typically beyond the scope of SI.

SI deals with human groups only when they operate at a low level of intelligence. Definition 5 includes human groups of individuals that operate under restrictions which the Swarm can overcome. An example is that of groups of individuals each with a limited computation device connected wirelessly to neighbors as in a cellular computing system. With appropriate algorithms such a Swarm could compute universally while the individual device cannot. This concept is applicable in, e. g., defense operations or emergency mass evacuations strategies.

In conclusion, Definition 5 embodies the intuitive definition of SI (Definition 1) and indicates why SI methods can be used to solve problems not solvable by traditional methods. Besides having all the properties of universal CA, the Swarm operates, as natural systems do, by independent clocks with no centralization and can be designed to be dynamic and unpredictable by any system including any its own units.

Future Directions

The field of SI is still less than 20 years old and it is in a formative phase, with SI researchers engaged in a broad range of disciplines. During these formative years the 'conversation' about SI has followed several strands around some common themes with various emphases ranging from speculative inquiries to practical interests, as we have seen in this article.

The meaning of the term SI has tended to broaden, covering now many areas, apparently only weakly related by the some intuitive notions. We have seen, however, how all these SI ideas have a 'center of attraction' in a basic concept of SI that can be made precise and quantified (Definition 4 or 5), and thus used to provide unity, continuity, and boundaries, thus preventing the area from broadening to the point of being unable to sustain an effective research community.

With this perspective, SI can also be seen as having an ultimate theoretical goal for the practical realization of engineered Swarms, whether robotic, biological or simply computational.

Practically, the goal of SI will remain two-fold: to provide models to explain biological societies and to engineer algorithms and devices with capability beyond those of traditional technologies. It will continue to include Swarm Robotics and bio-inspired algorithms such as swarm optimization methods.

But although commonly regarded as a typical example of bio-inspired technology, SI applications are likely to go beyond bio-inspired systems. We can see this if we consider how nature-inspired technologies have evolved.

Science discovers laws of nature, and technology makes inventions using those laws, often together with design ideas also derived from nature. Thus, for example, laws of physics and designs inspired by crystal structures are now applied to nanotechnologies; similarly, laws of biology and designs inspired by genetic configurations are applied to make artificial organisms in biotechnology. More recently a new kind of science [50] is discovering the laws and designs of computing machines as though they were natural systems, and these discoveries are likely to be used to invent new software algorithms and hardware implementations of those algorithms. SI, besides being bio-inspired, can be said to be inspired also by this new science, which, in some respects, can be more general than biology.

In this perspective, SI will evolve into the study of what amounts to be very powerful computing systems. Designing the simplest of these, i. e., the simplest universal dynamic cellular-computing system updating with natural

asynchrony, is an example of a future theoretical and practical challenge for SI. More immediate, future applications can be extrapolated from the examples given throughout the article.

Bibliography

Primary Literature

1. Beni G, Wang J (1989) Swarm Intelligence in Cellular Robotic Systems, Proceed. NATO Advanced Workshop on Robots and Biological Systems, Tuscany, Italy, June 26–30
2. Beni G, Wang J (1989) Swarm Intelligence. Proc. 7th Ann Meeting of the Robotics Society of Japan, pp 425–428 (in Japanese)
3. Bonabeau E, Dorigo M, Theraulaz G (1999) Swarm Intelligence: From Natural to Artificial Systems. Oxford Univ Press, New York
4. Kennedy J, Eberhart RC, Shi Y (2001) Swarm Intelligence. Morgan Kauffman, San Mateo
5. Abraham A, Grosan C, Ramos V (2006) Swarm Intelligence in Data Mining (Studies in Computational Intelligence). Springer, vol 34
6. Engelbrecht AP (2006) Fundamentals of Computational Swarm Intelligence. Wiley, New York
7. Dorigo M, Gambardella LM, Birattari M, Martinoli A (eds) (2006) Ant Colony Optimization and Swarm Intelligence: 5th International Workshop, ANTS 2006, Brussels, Belgium, September 4–7, (2006), Proceedings (Lecture Notes in Computer Science). Springer, Berlin
8. IEEE Swarm Intelligence Symposium, Honolulu, Hawaii, 1–5 April 2007. http://www.computelligence.org/sis/2007/?q=node/2. Accessed 6 Mar 2008
9. Dorigo M (ed) Swarm Intelligence, ISSN: 1935-3812. Springer US (to begin publication Summer 07)
10. Beni G (2004) From Swarm Intelligence to Swarm Robotics. In: Sahin E, Spear WM (eds.) Swarm Robotics March, Revised Selected Papers, SAB 2004 International Workshop, Santa Monica, CA, 17 July 2004. Lecture Notes in Computer Science vol 3342. Springer, Berlin, pp 1–9
11. Goss S, Aron S, Deneubourg JL, Pasteel JM (1989) Self-organized shortcuts in the Argentine ant. Naturwissenschaften 76:579–581
12. Dorigo M (1992) Optimization, Learning and Natural Algorithms. Ph D Thesis, Dipartimento di Elettronica, Politecnico di Milano, Milan (in Italian)
13. Dorigo M, Stutzle T (2004) Ant Colony Optimization. MIT Press, Cambridge
14. Olariu S, Zomaya AY (2005) Handbook of Bioinspired Algorithms and Applications. Chapman Hall/Crc Computer Information Science, Boca Raton
15. Passino K (2004) Biomimicry for Optimization, Control, and Automation. Springer, London
16. Reynolds C (1987) Flocks, Herds, and Schools: a distributed behavioral model. Computer Graphics 21(4):25–34
17. Moere AV (2004) Information Flocking: Time-Varying Data Visualization using Boid Behaviors. Proc Eighth Int Conf Inform Visual, pp 409–414
18. Eftimie R, de Vries G, Lewis MA (2007) Complex spatial group patterns result from different animal communication mechanisms. Proc Nat Acad Sci 104(17)
19. Parker LE, Schneider FE, Schultz AC (2005) Multi-Robot Systems. From Swarms to Intelligent Automata, vol 3: Proceedings from the 2005 International Workshop on Multi-Robot Systems. Springer, Dordrecht
20. Sahin E, Spears WM (2005) Swarm Robotics. SAB 2004 International Workshop, Santa Monica, CA, July 17, 2004, Revised Selected Papers (Lecture Notes in Computer Science) Springer, Berlin
21. Winfield A, Redi J http://www.robocomm.org/2007/index.shtml. Accessed 6 Mar 2008
22. Sahin EL, Spears WM, Winfield AFT (Eds.) (2007) Swarm Robotics. Second SAB 2006 International Workshop, Rome, Italy, September 30-October 1, 2006 Revised Selected Papers. Series: Lecture Notes in Computer Science vol 4433. Springer, Berlin
23. Baldassarre G, Trianni V, Bonani M, Mondada F, Dorigo M, Nolfi S (2007) Self-organized coordinated motion in groups of physically connected robots. IEEE Trans Syst Man Cybern Part B: Cybernetics 37(1):224–239
24. Gazi V, Passino KM (2004) Stability Analysis of Social Foraging Swarms. IEEE Trans Syst Man Cybern B 34:539–557
25. Balch T, Dellaert F, Feldman A, Guillory A, Isbell CL Jr, Khan Z, Pratt SC, Stein AN, Wilde H (2006) How AI and Multi-Robot Systems research will accelerate our understanding of social animal behavior. Proc IEEE 94(7):1445–1463
26. Brooks R (1986) A robust layered control system for a mobile robot. IEEE J Robot Autom RA-2(1):14
27. Johnson N, Galata A, Hogg DB (1998) The acquisition and use of interaction behavior models. In: Proc IEEE Computer Society Conf Computer Vision and Pattern Recognition, pp 866–871, IEEE Computer Society Press, Santa Barbara
28. Curtis SA, Mica J, Nuth J, Marr G, Rilee ML, Bhat M (2000) Autonomous Nano-Technology Swarm. Proceedings of the 51st International Aeronautical Congress, IAF-00-Q5.08
29. ANTS website: http://ants.gsfc.nasa.gov, Goddard Ants Team. Accessed 6 Apr 2008
30. Brooks RA (1991) Intelligence without representation. Artif Intell 47:139–159
31. Dorigo M, Tuci E, Groß R, Trianni V, Labella TH, Nouyan S, Ampatzis C, Deneubourg J-L, Baldassarre G, Nolfi S, Mondada F, Floreano D, Gambardella LM (2004) The SWARM-BOTS project. In: Sahin E, Spears WM (eds) Proc. of the 1st Int. Workshop on Swarm Robotics. LNCS vol 3342. Springer, Berlin, pp 26–40
32. Mondada F, Pettinaro GC, Guignard A, Kwee IV, Floreano D, Deneubourg J-L, Nolfi S, Gambardella LM, Dorigo M (2004) SWARM-BOT: A new distributed robotic concept. Auton Robots 17(2–3):193–221
33. Mondada F, Gambardella LM, Floreano D, Nolfi S, Deneubourg J-L, Dorigo M (2005) The cooperation of swarm-bots: Physical interactions in collective robotics. IEEE Robot Autom Mag 12(2):21–28
34. Rocha LM, Yaeger LS, Bedau MA, Floreano D, Goldstone RL, Vespignani A (2006) Artificial Life X Proceedings of the Tenth International Conference on the Simulation and Synthesis of Living Systems. MIT Press, Cambridge
35. Traub J (ed) Journal of Complexity. Elsevier http://www.elsevier.com/wps/find/journaldescription.cws_home/622865/description#description
36. Nicolis G, Prigogine I (1977) Self-Organization in Non-Equilibrium Systems. Wiley, New York

37. Sipper M (1999) The Emergence of Cellular Computing. IEEE Comput 32(7):18–26
38. Kennedy J, Eberhart RC (1995) Particle swarm optimization. In: Proc. IEEE int'l conf. on neural networks vol IV, pp 1942–1948. IEEE service center, Piscataway
39. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by Simulated Annealing. Science 220(4598):671–680
40. Snyman JA (2005) Practical Mathematical Optimization: An Introduction to Basic Optimization Theory and Classical and New Gradient-Based Algorithms. Springer Science-Business Media, New York
41. Mogilner A, Edelstein-Keshet L, Bent L, Spiros A (2003) Mutual interactions, potentials, and individual distance in a social aggregation. J Math Biol 47:353–389
42. Murray JD (2002) Mathematical Biology I: An Introduction, 3rd edn., Interdiscip Appl Math 17. Springer, New York
43. Topaz CM, Bertozzi A (2004) Swarming Patterns in two-dimensional kinematic model for biological groups. SIAM J Appl Math 65(1):152–174
44. Beni G (1988) The concept of cellular robot. Proc 3rd IEEE Symposium on Intelligent Control, pp 57–61. Arlington
45. Beni G, Hackwood S (1992) Stationary waves in cyclic swarms. In: Proc IEEE International Symposium on Intelligent Control. August 10–13, 1992. Institute of Electrical and Electronics Engineers (IEEE), Glasgow
46. Beni G (1992) Distributed Robotic Systems and Swarm Intelligence. J Robot Soc Japan (in Japanese) 10:31–37
47. Sipper M (1997) Evolution of Parallel Cellular Machines: The Cellular Programming Approach. Lecture Notes in Computer Science. Springer
48. Gottfredson LS (1997) Mainstream Science on Intelligence: An Editorial with 52 Signatories, History, and Bibliography. Intelligence 24(1):13–23
49. Cooper SB (2003) Computability Theory. Chapman Hall/CRC, Boca Raton
50. Wolfram S (2002) A New Kind of Science. Wolfram Media, Champaign
51. Gardner M (1970) The fantastic combinations of John Conway's new solitaire game 'life'. Sci Am 223:120–123
52. Johnson DS, McGeoch LA (2003) The traveling salesman problem: A case study in local optimization. In: Aarts EHL, Lenstra JK (eds) Local Search in Combinatorial Optimization, pp 215–310. Wiley, Chichester
53. Huberman BA, Glance NS (1993) Evolutionary Games and computer simulations. Proc Nati Acad Sci USA 90:7716–7718
54. Liang P, Beni G (1995) Robotic Morphogenesis. Proc Int Conf Robot Automat 2:2175–2180
55. Turing AM (1952) The Chemical Basis for Morphogenesis. Phil Trans Roy Soc London B 237:37–72
56. Cornforth D, Green D, Newth D (2005) Ordered Asynchronous Processes in Multi-Agent Systems. Physica D 204(1–2):70–82
57. Bersini H, Detour V (1994) Asynchrony induces stability in CA based models. In: Brooks RA, Maes P (eds) Artificial Life IV. MIT Press, Cambridge, pp 382–387
58. Beni G (2004) Order by Disordered Action in Swarms. In: Sahin E, Spear WM (eds.) (2005) Swarm Robotics, SAB 2004 International Workshop. LNCS, vol 3342. Springer, Berlin, pp 153–171
59. Schonfisch B, deRoos A (1999) Synchronous and Asynchronous Updating in Cellular Automata. Biosyst 51:123–143
60. Sipper M, Tomassini M, Capcarrere MS (1997) Evolving asynchronous and scalable non-uniform cellular automata. In: Proceedings of International Conference on Artificial Neural Networks and Genetic Algorithms (ICANNGA97). Springer, New York
61. Clapham N (2002) Emergent synchrony: simple asynchronous update rules can produce synchronous behavior. In: Sarker R, McKay RI, Gen M, Namatame A (eds) Proceedings of the Sixth Australia–Japan Joint Workshop on Intelligent and Evolutionary Systems. Australian National University, Canberra, pp 41–46
62. Nehaniv CL (2002) Evolution in asynchronous cellular automata. In: Standish RK, Abbass HA, Bedau MA (eds) Proceedings of the Eighth Conference on Artificial Life. MIT Press, Cambridge, pp 65–74
63. Nakamura K (1974) Asynchronous cellular automata and their computational ability. Syst Comput Controls 5(5):58–66
64. Lee J, Peper F, Adachi S, Morita K (2004) Universal delay-insensitive circuits with bi-directional and buffering lines. IEEE Trans Comput 53(8):1034–1046
65. Adachi S, Peper F, Lee J (2004) Universality of hexagonal asynchronous totalistic cellular automata. In: Chopard B, Hoekstra AG, Sloot PMA (eds) (2004) Cellular Automata, 6th International Conference on Cellular Automata for Research and Industry. LNCS, vol 3305. Springer, Berlin, pp 91–100
66. Takada Y, Isokawa T, Peper F, Matsui N (2006) Construction Universality in purely asynchronous cellular automata. J Comput Syst Sci 72:1368–1385
67. von Neumann J (1966) Theory of Self-Reproducing Automata. Univ Illinois Press, Champaign (edited and completed by Burks AW)
68. Langton CG (1984) Self-reproduction in cellular automata. Phys D 10:135–144
69. Wolfram S (1985) Undecidability and Intractability in Theoretical Physics. Phys Rev Lett 54:735–738
70. Weiss G (2000) Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence. MIT Press, Cambridge

Books and Reviews

References [3, 4, 6, 10, 13, 14, 15, 48, 51] are the main books and reviews. The journal Swarm Intelligence [9] is the main source for new research results. The following provide additional general material related to Swarm Intelligence.

Sahin E, Spears WM (2005) Swarm Robotics: SAB 2004 International Workshop, Santa Monica, CA, July 17, 2004, Revised Selected Papers (Lecture Notes in Computer Science). Springer, Berlin

Dorigo M, Sahin E (2004) Swarm Robotics– special issue editorial. Autonom Robot 17(2–3)111–113

Sipper M (2002) Machine Nature: The Coming Age of Bio-Inspired Computing. McGraw-Hill, New York

Camazine S, Deneubourg J-L, Franks NR, Sneyd J, Theraulaz G, Bonabeau E (2001) Self-Organization in Biological Systems. Princeton Univ Press, Princeton

Symbolic Dynamics

Brian Marcus
Department of Mathematics,
University of British Columbia, Vancouver, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
Origins of Symbolic Dynamics: Modeling of Dynamical Systems
Shift Spaces and Sliding Block Codes
Shifts of Finite Type and Sofic Shifts
Entropy and Periodic Points
The Conjugacy Problem
Other Coding Problems
Coding for Data Recording Channels
Connections with Information Theory and Ergodic Theory
Higher Dimensional Shift Spaces
Future Directions
Bibliography

Glossary

In this glossary, we give only brief descriptions of key terms. We refer to specific sections in the text for more precise definitions.

Almost conjugacy (Sect. "Other Coding Problems") A common extension of two shift spaces given by factor codes that are one-to-one almost everywhere.

Automorphism (Sect. "The Conjugacy Problem") An invertible sliding block code from a shift space to itself; equivalently, a shift-commuting homeomorphism from a shift space to itself; equivalently, a topological conjugacy from a shift space to itself.

Dimension group (Sect. "The Conjugacy Problem") A particular group associated to a shift of finite type. This group, together with a distinguished sub-semigroup and an automorphism, captures many invariants of topological conjugacy for shifts of finite type.

Embedding (Sect. "Shift Spaces and Sliding Block Codes") A one-to-one sliding block code from one shift space to another; equivalently, a one-to-one continuous shift-commuting mapping from one shift space to another.

Factor map (Sect. "Shift Spaces and Sliding Block Codes") An onto sliding block code from one shift space to another; equivalently, an onto continuous shift-commuting mapping from one shift space to another. Sometimes called *Factor Code*.

Finite equivalence (Sect. "Other Coding Problems") A common extension of two shift spaces given by finite-to-one factor codes.

Full shift (Sect. "Shift Spaces and Sliding Block Codes") The set of all bi-infinite sequences over an alphabet (together with the shift mapping). Typically, the alphabet is finite.

Higher dimensional shift space (Sect. "Higher Dimensional Shift Spaces") A set of bi-infinite arrays of a given dimension, determined by a collection of finite forbidden arrays. Typically, the alphabet is finite.

Markov partition (Sect. "Origins of Symbolic Dynamics: Modeling of Dynamical Systems") A finite cover of the underlying phase space of a dynamical system, which allows the system to be modeled by a shift of finite type. The elements of the cover are closed sets, which are allowed to intersect only on their boundaries.

Measure of maximal entropy (Sect. "Connections with Information Theory and Ergodic Theory") A shift-invariant measure of maximal measure-theoretic entropy on a shift space. Its measure-theoretic entropy coincides with the topological entropy of the shift space.

Road problem (Sect. "Other Coding Problems") A recently-solved classical problem in symbolic dynamics, graph theory and automata theory.

Run-length limited shift (Sect. "Coding for Data Recording Channels") The set of all bi-infinite binary sequences whose runs of zeros, between two successive ones, are bounded below and above by specific numbers.

Shift equivalence (Sect. "The Conjugacy Problem") An equivalence relation on defining matrices for shifts of finite type. This relation characterizes the corresponding shifts of finite type, up to an eventual notion of topological conjugacy.

Shift space (Sect. "Shift Spaces and Sliding Block Codes") A set of bi-infinite sequences determined by a collection of finite forbidden words; equivalently, a closed shift-invariant subset of a full shift.

Shift of finite type (Sect. "Shifts of Finite Type and Sofic Shifts") A set of bi-infinite sequences determined by a *finite* collection of finite forbidden words.

Sliding block code (Sect. "Shift Spaces and Sliding Block Codes") A mapping from one shift space to another determined by a finite sliding block window; equivalently, a continuous shift-commuting mapping from one shift space to another.

Sofic shift (Sect. "Shifts of Finite Type and Sofic Shifts") A shift space which is a factor of a shift of finite type; equivalently, a set of bi-infinite sequences determined by a finite directed labeled graph.

State splitting (Sect. "The Conjugacy Problem") A splitting of states in a finite directed graph that creates

a new graph, whose vertices are the split states. The operation that creates the new graph from the original graph is a basic building block for all topological conjugacies between shifts of finite type.

Strong shift equivalence (Sect. "The Conjugacy Problem") An equivalence relation on defining matrices for shifts of finite type. In principle, this relation characterizes the corresponding shifts of finite type, up to topological conjugacy.

Topological conjugacy (Sect. "Shift Spaces and Sliding Block Codes") A bijective sliding block code from one shift space to another; equivalently, a shift-commuting homeomorphism from one shift space to another. Sometimes called *conjugacy*.

Topological entropy (Sect. "Entropy and Periodic Points") The asymptotic growth rate of the number of finite sequences of given length in a shift space (as the length goes to infinity).

Zeta function (Sect. "Entropy and Periodic Points") An expression for the number of periodic points of each given period in a shift space.

Definition of the Subject

Symbolic dynamics is the study of shift spaces, which consist of infinite or bi-infinite sequences defined by a shift-invariant constraint on the finite-length sub-words. Mappings between two such spaces can be regarded as codes or encodings. Shift spaces are classified, up to various kinds of invertible encodings, by combinatorial, algebraic, topological and measure-theoretic invariants.

The subject is intimately related to many other areas of research, including dynamical systems, ergodic theory, automata theory and information theory. Shift spaces and their associated shift mappings are used to model a rich and important class of smooth dynamical systems and ergodic measure-preserving transformations. These models have provided a valuable tool for classifying and understanding fundamental properties of dynamical systems. In addition, techniques from symbolic dynamics have had profound applications for data recording applications, such as algorithms and analysis of invertible encodings, and problems in matrix theory, such as characterization of the set of eigenvalues of a nonnegative matrix.

Introduction

This article is intended to give a picture of major topics in symbolic dynamics. Section "Origins of Symbolic Dynamics: Modeling of Dynamical Systems" reviews the roots of symbolic dynamics in modeling of dynamical systems. Section "Shift Spaces and Sliding Block Codes" lays the foundation by defining the kinds of spaces and mappings considered in the subject. Section "Shifts of Finite Type and Sofic Shifts" focuses on distinguished special classes of spaces, known as shifts of finite type and sofic shifts. Section "Entropy and Periodic Points" introduces the most fundamental invariants, periodic points and topological entropy. Sections "The Conjugacy Problem" and "Other Coding Problems" survey progress on the conjugacy problem and other classification/coding problems for shifts of finite type and sofic shifts. In Sect. "Coding for Data Recording Channels", we present applications to coding for data recording. Section "Connections with Information Theory and Ergodic Theory" provides a link with information theory and ergodic theory. Finally, Sect. "Higher Dimensional Shift Spaces" treats higher dimensional symbolic dynamics.

While this article covers many of the most important topics in the subject, others have been omitted or treated lightly, due to space limitations. These include one-sided shift spaces, countable state symbolic systems, orbit equivalence, flow equivalence, the automorphism group, cellular automata, and substitution systems. References to work in these sub-areas can be found in the sources mentioned below.

For introductory reading on symbolic dynamics and its applications, beyond this article, one can consult the textbooks Kitchens [65] and Lind and Marcus [78]. There are also excellent introductory survey articles, such as Boyle [23], Lind and Schmidt [79], and S. Williams [132]. In addition there are very good expositions which focus on other aspects of the subject. These include Beal [11], which focuses on connections between symbolic dynamics and automata theory, the lecture notes Marcus–Roth–Siegel [83], which focuses on constrained coding applications, and Immink [54], which focuses on applications to data storage. There are also several excellent collections of articles on special areas of the subject, such as [17,133], and [128]. The book [15] by Berstel and Perrin treats the subject of variable length codes and contains many ideas related to symbolic dynamics. Finally, there is an excellent recent survey by Boyle on Open Problems in Symbolic Dynamics [24].

Origins of Symbolic Dynamics: Modeling of Dynamical Systems

Symbolic dynamics began as an effort to model dynamical systems using sequences of symbols. A *dynamical system* is a pair (X, T) where X is a set and T is a transformation from X to itself. For definiteness, we assume that T is invertible, although this is not a necessary restriction.

Since T maps X to itself, we can iterate T: $T^2 = T \circ T$, $T^3 = T \circ T \circ T$, etc. The *orbit* of a point $x \in X$ is the sequence of points: $\ldots, T^{-2}(x), T^{-1}(x), x, T(x), T^2(x), \ldots$ In the theory of dynamical systems, one asks questions about orbits such as the following: Are there periodic orbits (i. e., x such that $T^n(x) = x$ for some $n > 0$)? Are there dense orbits (the orbit of x is dense if for any point y in X, $T^n(x)$ is "close" to y for some n)? How does the behavior of an orbit vary with x? How can we describe the collection of all orbits of the dynamical system? When is the dynamical system "chaotic"? For more information on dynamical systems, we refer the reader to [16,38,47].

The subject of dynamical systems has its roots in Classical Mechanics; in that setting, X is the set of all possible states of a system (e. g., the positions, momenta of all particles in a physical system), and the transformation T is the time evolution map, which maps the state of the system at one time to the state of the system at one time unit later.

Symbolic dynamics provides a model for the orbits of a dynamical system (X, T) via a space of sequences. This is done by "quantizing" X into cells, associating symbols to the cells and representing points as bi-infinite sequences of symbols. For instance in Fig. 1, X is a square, and T is some transformation of the square.

We have drawn a portion of the orbit of a point $x \in X$ for the dynamical system (X, T). We have also quantized X into two cells: the left half, called '0', and the right half, called '1'. Then the point x is represented by the bi-infinite sequence $s(x) = \ldots s_{-2}s_{-1}.s_0s_1s_2 \ldots$ where s_n is the label of the cell to which $T^n(x)$ belongs (here, we use the decimal point to separate coordinates s_i, $i < 0$ from s_i, $i \geq 0$). So, for x as given in Fig. 1, we see that

$$s(x) = \ldots 11.001 \ldots$$

for instance $s_0 = 0$ because x belongs to the left half of the square, and $s_2 = 1$ because $T^2(x)$ belongs to the right half of the square. Now, if x is represented by the sequence $s(x)$, then $T(x)$ is represented by the shift of $s(x)$:

$$s(T(x)) = \ldots 110.01 \ldots .$$

So, T is represented 'symbolically' as the shift transformation.

By representing all points of X as bi-infinite sequences, we obtain a *symbolic dynamical system* (Y, σ) where Y is a set of sequences (representing X) and σ is the shift transformation (representing T). The "symbolic" refers to the symbols, and the "dynamical" refers to the action of the shift transformation.

For this representation to be faithful, distinct points should be represented by distinct sequences, and this imposes extra conditions on how X is quantized into cells. Also, Y is typically a set of sequences constrained by certain rules, such as a certain symbol may only be followed by certain other symbols.

In this way, one can use symbolic dynamics to study dynamical systems. Properties of orbits of the original dynamical system are reflected in properties of the resulting sequences. For instance, a point whose orbit is periodic becomes a periodic sequence, and the distribution of the orbit of a point x in X is reflected in the distribution of finite strings within $s(x)$. Beginning with Hadamard [46] in 1898 and followed by Hedlund, Morse and others in the 1920s, 1930s and 1940s [48,49,92,93], this method was used to prove the existence of periodic, almost periodic and other interesting motions in classical dynamical systems, such as geodesic flows on surfaces of negative curvature; this was done by finding interesting sequences satisfying the constraints defined by the corresponding symbolic dynamical system. Later on, this was extended to general hyperbolic systems, where the symbolic dynamics is constructed using a *Markov Partition*, which is a disjoint collection of open sets whose closures cover X, each of which looks like a "rectangle", with vertical (resp., horizontal) fibers contracted (resp., expanded) by T. Markov Partitions were developed by Adler and Weiss [5], Sinai [120] and Bowen [18,19]; see also [13] and Sect. 6.5 in [78].

In more recent years, symbolic dynamics has been used as a tool in classification problems for dynamical systems. Here, the problem of determining when one dynamical system is 'equivalent' to another becomes, via symbolic dynamics, a coding problem. Roughly speaking, two dynamical systems, (X_1, T_1) and (X_2, T_2), are *equivalent* if there is an invertible mapping from X_1 to X_2 which makes T_1 "look like" T_2. If the corresponding symbolic dynamical systems are denoted (Y_1, σ) and (Y_2, σ), then an equivalence between (X_1, T_1) and (X_2, T_2) becomes a time-in-

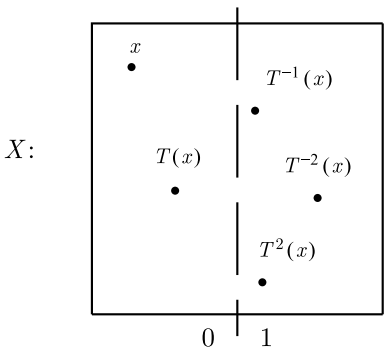

Symbolic Dynamics, Figure 1
Representing points symbolically

variant, invertible encoding from Y_1 to Y_2 (time-invariant because the shift transformation represents the dynamics). Thus, the classification problem in dynamical systems leads to a coding problem between constrained sets of sequences.

We have described dynamical systems as the discrete-time iteration of a single mapping. However, continuous-time iterations have been studied since the inception of the subject. These are known as continuous-time flows, with the main example being the set of solutions to a system of ordinary differential equations. Indeed, the work of Hedlund and Morse mentioned above was done in this context.

Shift Spaces and Sliding Block Codes

Let $\mathcal{A}$ be an alphabet of symbols, which we assume to be finite. The principal objects of study in symbolic dynamics are certain kinds of collections of sequences of symbols from $\mathcal{A}$. Typically, these sequences are infinite $x = x_0x_1x_2\ldots$, but it is often more convenient to deal with *bi-infinite sequences* $x = \ldots x_{-2}x_{-1}x_0x_1x_2\ldots$ For some problems, the results are similar in the infinite and bi-infinite categories, while for other problems, they are quite different. In this article, we focus on the bi-infinite setting.

The symbol x_i is the ith *coordinate* of x. When writing a specific sequence, we need to specify which is the 0th coordinate. As suggested in Sect. "Origins of Symbolic Dynamics: Modeling of Dynamical Systems", this is done with a decimal point to separate the x_i with $i \geq 0$ from those with $i < 0$: $x = \ldots x_{-2}x_{-1}.x_0x_1x_2\ldots$. A *block* or *word* over $\mathcal{A}$ is a finite sequence of symbols from $\mathcal{A}$. A block of length N is called an *N-block*. For blocks u, v, the block uv is the concatenation of u and v, and for a block w, the concatenation of N copies of w is denoted w^N.

The *full $\mathcal{A}$-shift* $\mathcal{A}^{\mathbf{Z}}$ is the set of all bi-infinite sequences of symbols from $\mathcal{A}$. The *full r-shift* is the full shift over the alphabet $\{0, 1, \ldots, r-1\}$. The *shift map* σ on a full shift maps a point x to the point $y = \sigma(x)$ whose ith coordinate is $y_i = x_{i+1}$.

The *orbit* of a point in a full shift is its orbit under the shift map. The full shift contains many different types of orbits. For instance, it contains a dense orbit (namely, any sequence which contains every block in the alphabet) and periodic orbits (namely, any sequence which is periodic).

We are interested in sets that can be specified by a list (finite or infinite) of forbidden blocks. Namely, given a collection $\mathcal{F}$ of "forbidden blocks" over $\mathcal{A}$, the subset X consisting of all sequences in $\mathcal{A}^{\mathbf{Z}}$, none of whose subwords belong to $\mathcal{F}$, is called a *shift space* (or simply *shift*), and we write $X = X_{\mathcal{F}}$. When a shift space X is contained in a shift space Y, we say that X is a *subshift* of Y.

Example 1 X is the set of all binary sequences with no two 1's next to each other. Here $X = X_{\mathcal{F}}$, where $\mathcal{F} = \{11\}$. This shift is called the *golden mean shift*, for reasons that will become apparent later.

Example 2 X is the set of all binary sequences so that between any two 1's there are an even number of 0's. We can take for $\mathcal{F}$ the collection

$$\{10^{2n+1}1 : n \geq 0\} .$$

This example is naturally called the *even shift*.

Example 3 X is the set of all binary sequences such that between any two successive 1's number of 0's is prime. We can take for $\mathcal{F}$ the collection

$$\{10^n1 : n \text{ is composite}\} .$$

This example is naturally called the *prime shift*.

Alternatively (and equivalently), shift spaces can be defined as closed, shift-invariant subsets of full shifts. Here, "closed" means with respect to a metric, for which two points are close if they agree in a large "central block"; one such metric is $\rho(x, y) = 2^{-k}$ if $x \neq y$, with k maximal such that $x_{[-k,k]} = y_{[-k,k]}$ (with the conventions that $\rho(x, y) = 0$ if $x = y$ and $\rho(x, y) = 2$ if $x_0 \neq y_0$).

Let X be a subset of a full shift, and let $\mathcal{B}_N(X)$ denote the set of all N-blocks that occur in elements of X. The *language of* X is $\mathcal{B}(X) = \cup_N \mathcal{B}_N(X)$. It can be shown that the language of a shift space determines the shift space uniquely, and so we can equally well describe a shift space by specifying the "occurring" or "allowed" blocks, rather than the forbidden blocks. For example, the golden mean shift is specified by the language of blocks in which 1's are isolated.

This establishes a connection with automata theory [6,11], which studies collections of blocks, rather than infinite or bi-infinite sequences. The languages that occur in symbolic dynamics, i. e. as $\mathcal{B}(X)$ for some shift space X, are simply those sets $\mathcal{L}$ of blocks that satisfy two simple properties: every sub-block of an element of $\mathcal{L}$ belongs to $\mathcal{L}$, and every element $w \in \mathcal{L}$ is extendable to a larger block $awb \in \mathcal{L}$, with $a, b \in \mathcal{A}$.

Some examples are best described by allowed blocks. One example is the famous *Morse shift*.

Example 4 Let $A_0 = 0$ and inductively define blocks $A_{n+1} = A_n\overline{A_n}$, where $\overline{A_n}$ denotes bitwise complement. The shift space whose allowed blocks are the sublocks of the A_n is called the *Morse shift*.

This shift space has an alternative description as follows. Since each A_n is a prefix of A_{n+1}, the sequence of blocks A_n determines a unique right-infinite sequence x^+ (with each A_n as a a prefix). If we denote x^- as the left-infinite sequence obtained by writing x^+ backwards, then the Morse shift is the closure of the orbit of the point $x^-.x^+$. The sequence x^+ is known as the Prouhet–Thue–Morse (PTM) sequence, since Prouhet and Thue introduced it earlier, but for different purposes.

While it is not immediately obvious, it can be shown that the PTM sequence is not periodic; moreover, the Morse shift is a *minimal shift*, which means that all its orbits are dense [91]. In contrast, while the full, golden mean, even and prime shifts all have dense orbits, each also has a dense set of periodic orbits.

There are two simple, but very important, constructions in symbolic dynamics that construct from a given shift space a new version which in some sense looks deeper into the space at the cost of a larger alphabet and more complex description.

Let X be a shift space over the alphabet $\mathcal{A}$, and $\mathcal{A}^{(N)} = \mathcal{B}_N(X)$. We can consider $\mathcal{A}^{(N)}$ as an alphabet in its own right, and form the full shift $(\mathcal{A}^{(N)})^{\mathbf{Z}}$. Define the *Nth higher block code* by

$$(\beta_N(x))_i = x_{[i,i+N-1]} \,.$$

Then the Nth *higher block shift* or *higher block presentation* of a shift space X is the image $X^{[N]} = \beta_N(X)$ in the full shift over $\mathcal{A}^{(N)}$.

Similarly, define the *Nth higher power code* $\gamma_N \colon X \to (\mathcal{A}^{(N)})^{\mathbf{Z}}$ by

$$(\gamma_N(x))_i = x_{[iN,iN+N-1]} \,.$$

The Nth *higher power shift* X^N of X is the image $X^N = \gamma_N(X)$ of X. The difference between $X^{[N]}$ and X^N is that the former is constructed by considering overlapping blocks and the latter by non-overlapping blocks.

Next, we turn to mappings between shift spaces. Suppose that $x = \ldots x_{-1}.x_0x_1\ldots$ is a sequence in a shift space X over $\mathcal{A}$. We can transform x into a new sequence $y = \ldots y_{-1}.y_0y_1\ldots$ over another alphabet C as follows. Fix integers m and n with $-m \le n$. To compute the ith coordinate y_i of the transformed sequence, we use a function Φ that depends on the "window" of coordinates of x from position $i-m$ to position $i+n$. Here $\Phi \colon \mathcal{B}_{m+n+1}(X) \to C$ is a fixed *block map*, called an $(m+n+1)$-*block map* from allowed $(m+n+1)$-blocks in X to symbols in C, and so

$$y_i = \Phi(x_{i-m}x_{i-m+1}\ldots x_{i+n}) = \Phi(x_{[i-m,i+n]}) \,.$$

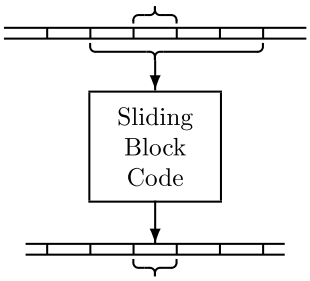

Symbolic Dynamics, Figure 2
Sliding block code

This is illustrated in Fig. 2.

Let X be a shift space over $\mathcal{A}$, and $\Phi \colon \mathcal{B}_{m+n+1}(X) \to C$ be a block map. Then the map $\phi \colon X \to C^{\mathbf{Z}}$ defined by $y = \phi(x)$, with y_i given by Φ above, is called the *sliding block code* with *memory* m and *anticipation* n *induced by* Φ. We will denote the formation of ϕ from Φ by $\phi = \Phi_\infty^{[-m,n]}$, or more simply by $\phi = \Phi_\infty$ if the memory and anticipation of ϕ are understood. If not specified, the memory is taken to be 0. If Y is a shift space contained in $C^{\mathbf{Z}}$ and $\phi(X) \subseteq Y$, we write $\phi \colon X \to Y$.

In analogy with the characterization of shift spaces as closed shift-invariant sets, sliding block codes can be characterized in a topological manner: namely, as the maps between shift spaces that are continuous and commute with the shift. This result is known as the Curtis–Hedlund–Lyndon theorem [50].

Example 5 Let $\mathcal{A} = \{0,1\} = C$, $X = \mathcal{A}^{\mathbf{Z}}$, $m = 0$, $n = 1$, and $\Phi(a_0a_1) = a_0 + a_1 \pmod 2$. Let $\phi = \Phi_\infty \colon X \to X$.

Example 6 The sliding block code, generated by $\Phi(00) = 1, \Phi(01) = 0 = \Phi(10)$, maps the golden mean shift onto the even shift.

Example 7 There is a trivial sliding block code from the full 2-shift into the full 3-shift, generated by $\Phi(0) = 0$, $\Phi(1) = 1$.

If a sliding block code $\phi \colon X \to Y$ is onto, then ϕ is called a *factor code* or *factor map*, and Y is a *factor* of X. If $\phi \colon X \to Y$ is one-to-one, then ϕ is called an *embedding* of X into Y. The sliding block code in Example 7 is an embedding but not a factor code, while the codes in Examples 5 and 6 are factor maps, but not embeddings.

A major (and unrealistic) goal of symbolic dynamics is to classify in an explicit way shift spaces up to the following natural notion of equivalence. A sliding block code $\phi \colon X \to Y$ is a *conjugacy* (or *topological conjugacy*) if it is invertible with sliding block inverse. Equivalently, a conjugacy is a bijective sliding block code and therefore simulta-

neously a factor code and an embedding. If there is a conjugacy from one shift space X to another Y, we say that X and Y are *conjugate*, denoted $X \cong Y$.

As an example, the higher block map β_N is a conjugacy between a shift space X and its higher block shift $X^{[N]}$. Via this code, we can "re-code" any sliding block code as a 1-block code (though typically a conjugacy and its inverse cannot, by this artifice, be simultaneously re-coded to 1-block codes).

In this section, we have given examples of relatively simple sliding block codes. But the typical conjugacy, as well as factor code and embedding, can be much more complicated.

Shifts of Finite Type and Sofic Shifts

A *shift of finite type (SFT)* is a shift space that can be described by a finite set of forbidden blocks, i. e., a shift space X having the form $X_{\mathcal{F}}$ for some finite set $\mathcal{F}$ of blocks. The terminology shift of finite type (or *subshift of finite type*) comes from dynamical systems (Smale [121]).

An SFT is *M-step* (or has *memory M*) if it can be described by a collection of forbidden blocks all of which have length $M+1$. It is easy to see that any SFT is M-step for some M.

Since any shift space can be defined by many different collections of forbidden blocks, it is useful to have the following equivalent condition expressed in terms of allowed blocks: an SFT is M-step if and only if whenever u is an allowed block of length at least M, u' is the suffix of u with length M and a is a symbol, then ua is allowed if and only if $u'a$ is allowed. In other words, in order to tell whether a symbol can be allowably concatenated to the end of an allowed word u, one need only look at the last M symbols of u. This is analogous to the "finite memory" property of M-step Markov chains.

The golden mean shift X is is a 1-step SFT, since it was defined by a forbidden list consisting of exactly one block: $\mathcal{F} = \{11\}$. Equivalently, it is only the last symbol of an allowed block that determines whether a given symbol can be concatenated at the end. In contrast, the even shift is not an SFT: for any M, the symbol 1 can be concatenated to the end of exactly one of the (allowed) words 10^M and 10^{M+1}.

Recall that the higher block code β_M is a conjugacy from X to $X^{[M]}$. Via this code, any SFT can be recoded to a 1-step SFT. And so any sliding block code on a shift space can be recoded to a 1-block code on a 1-step SFT. It is useful to have a concrete description of 1-step SFT's. In fact, these are precisely the shift spaces consisting of all bi-infinite sequences of vertices along paths on a finite directed graph. These are called *vertex shifts*. We find it more convenient to work with sequences of edges instead. To be precise:

Let G be a finite directed graph (or simply graph) with vertices (or *states*) $\mathcal{V} = \mathcal{V}(G)$ and *edges* $\mathcal{E} = \mathcal{E}(G)$. For an edge e, $i(e)$ denotes the initial state and $t(e)$ the terminal state. A *path* in G is a finite sequence of edges in G such that the terminal state of an edge coincides with the initial state of the following edge; a *cycle* in G is path that begins and ends at the same state. We will assume that G is *essential*, i. e., that every state has at least one outgoing edge and one incoming edge.

The *adjacency matrix* $A = A(G)$ is the matrix indexed by $\mathcal{V}$ with A_{IJ} equal to the the number of edges in G with initial state I and terminal state J. Since a graph and its adjacency matrix essentially determine the same information, we will frequently associate a graph G with its adjacency matrix A and a nonnegative integer matrix A with a graph G.

The *edge shift* X_G or X_A is the shift space over the alphabet $\mathcal{A} = \mathcal{E}$ defined by

$$\begin{aligned} X_G &= X_A \\ &= \left\{\xi = (\xi_i)_{i\in\mathbf{Z}} \in \mathcal{E}^{\mathbf{Z}}\colon \text{ each } \xi_{i+1} \text{ follows } \xi_i\right\}. \end{aligned}$$

It can be readily verified that edge shifts are 1-step SFT's. While edge shifts do not include all 1-step SFT's, any 1-step SFT can be recoded to an edge shift, and, compared with vertex shifts, edge shifts offer the advantage of a more compact description.

For many purposes, one can study a general shift space X by breaking it into smaller, more well-behaved pieces. A shift space is *irreducible* if whenever u and w are allowed blocks, there is a "connecting" block v such that uvw is allowed. While shift spaces do not always decompose into disjoint unions of irreducible shifts, every SFT can be written as a finite disjoint union of irreducible SFT's X_i together with "transient" one-way connections from one X_i to another. And irreducible edge shifts can be characterized in a particularly concrete form: namely, X_G is irreducible if and only if G is *irreducible*, i. e., for every ordered pair of vertices I and J there is a path in G starting at I and ending at J.

There is a stronger notion which is defined by a uniformity condition on the length of the connecting block. A shift space is *mixing* if whenever u and w are allowed blocks, there is an N, possibly depending on u and w, such that for all $n \geq N$, there is block v of length n such that uvw is allowed. And an edge shift X_G is mixing if and only if G is *primitive*, i. e., there is an integer N such that for any $n \geq N$ and any ordered pair of vertices I and J there is

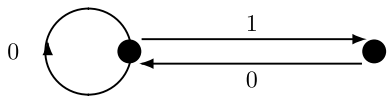

Symbolic Dynamics, Figure 3
Presentation of golden mean shift

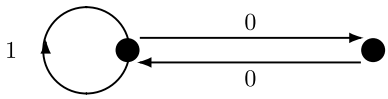

Symbolic Dynamics, Figure 4
Presentation of even shift

a path in G of length n starting at I and terminating at J. It follows that for SFT's in the definition of mixing, the uniform connecting length N can be chosen independent of the allowed blocks u and w.

It can be shown that, in some sense, any irreducible SFT X can be broken down into a union of disjoint maximal mixing shifts; namely, X can be written as the disjoint union of finitely many sets X_i, $i = 0, \ldots, p-1$ such that $\sigma(X_i) = X_{i+1} \bmod p$ and for each i, σ^p restricted to X_i can be regarded as a mixing SFT. This is a consequence of Perron–Frobenius theory, upon which symbolic dynamics relies heavily; see Seneta [118] for an introduction to this theory.

A *sofic shift* is the set of bi–infinite sequences obtained from a finite labeled directed graph $\mathcal{G} = (G, \mathcal{L})$; here, G is a finite directed graph and $\mathcal{L}$ is a labeling of the edges of G. The labeled graph is often called a *presentation* of the sofic shift. The golden mean shift and even shift are sofic, with presentations given in Figs. 3 and 4.

SFT's are sofic because any M-step SFT can be presented by a graph whose states are allowed M-blocks. Note also that any sofic shift is a factor of an SFT, namely via a (1-block) factor code $\mathcal{L}_\infty$ on the edge shift X_G based on a presentation $(G, \mathcal{L})$. In fact, the converse is true, and so the sofic shifts are precisely the shift spaces that are factors of SFT's. This was the original definition of sofic shifts given by Weiss [130].

Typically, the labeling is *right resolving*, which means that at any given state, all outgoing edges have distinct labels (as in Figs. 3 and 4).

Theorem 1

(a) Any sofic shift has a right resolving presentation.
(b) Any irreducible sofic shift has a unique minimal right resolving presentation.

Part (a) is a direct consequence of the subset construction in automata theory [6,11] which constructs a right resolving presentation from an arbitrary presentation; see also Coven and Paul [33,34]. Part (b) makes use of the the state-minimization algorithm from automata theory [6,11], but requires an idea beyond that found in automata theory (Fischer [39,40]). The unique presentation in part (b) may be regarded as a canonical presentation. We remark that for an irreducible (resp. mixing) sofic shift, the underlying graph of the unique minimal right resolving presentation is irreducible (resp., primitive).

Sometimes it is useful to weaken the concept of right resolving to *right closing*, which means "right resolving with delay"; more precisely a labeling is right closing, with delay D if all paths of length $D+1$ with the same initial state and the same label have the same initial edge. Also, we sometimes consider *left resolving* and *left closing* labellings (replace "outgoing" with "incoming" in the definition of right resolving, and replace "initial" with "terminal" in the definition of right closing).

While the class of SFT's is defined by a "finite-memory" property, the more general class of sofic shifts is defined by a "finite-state" property, in that the possible symbols that can occur at time 0 are determined by the past via one of finitely many states. In a presentation, the vertices can be viewed as state information which connects sequences in the past with sequences in the future.

It is not difficult to show that neither the prime shift (Example 3) nor the Morse shift (Example 4) are sofic and therefore also not SFT. For the prime shift, this can be done by an application of the pumping lemma from automata theory [6] (or p. 68 of [78]).

There are uncountably many shift spaces, but only countably many sofic shifts. So, it is not surprising that the behavior of sofic shifts is very special. However, they are very useful in modeling smooth dynamical systems (Sect. "Origins of Symbolic Dynamics: Modeling of Dynamical Systems") and in information theory and applications to data recording (Sect. "Coding for Data Recording Channels"), where they arise as *constrained systems*, although this term is usually reserved for the set of (finite) blocks obtained from a finite directed labeled graph [83].

Entropy and Periodic Points

An invariant of conjugacy is an object associated to a shift space that is preserved under conjugacy. It can be shown that many of the concepts that we have already introduced are invariants: irreducibility, mixing, as well as the properties of being a shift of finite type or sofic shift. Beyond these qualitative invariants, there are many quantitative invariants that can, in many cases, be computed explicitly. Foremost among these is topological entropy.

The *(topological) entropy* (or simply *entropy*) of X is:

$$h(X) = \lim_{N\to\infty} \frac{\log|\mathcal{B}_N(X)|}{N} ;$$

here, $|\cdot|$ denotes cardinality, and for definiteness the log is taken to mean $\log_2$ but any base will do. A subadditivity argument shows that the limit does indeed exist [127]. It should be evident that $h(X)$ is a measure of the "size" or "complexity" of X, as it is simply the asymptotic growth rate of the number of blocks that occur in X. Topological entropy for continuous dynamical systems was defined by Adler, Konheim and McAndrew [3], in analogy with measure-theoretic entropy.

Since a k-block sliding block code from a shift space X to a shift space Y maps $\mathcal{B}_{N+k-1}(X)$ into $\mathcal{B}_N(Y)$, it is not hard to see that entropy is an invariant of conjugacy and that it cannot increase under factors and cannot decrease under embeddings.

In many cases, one can explicitly compute entropy. For example, for the full r-shift X, $|\mathcal{B}_N(X)| = r^N$, and so $h(X) = \log r$. And from the defining sequence of the Morse shift, we see that the number of distinct 2^N-blocks is at most $4(2^N)$, it follows that the growth rate cannot be exponential and so the entropy of the Morse shift is zero.

For SFT's and more generally for sofic shifts, entropy can be computed explicitly. The key to this computation is the following result, which is based on the Perron–Frobenius theorem.

Theorem 2 [99,119] *For any graph G, $h(X_G) = \log\lambda_{A(G)}$, where $\lambda_{A(G)}$ is the largest eigenvalue of $A(G)$.*

The rough idea is that the number of N-blocks in X_G is the number of paths of length N and thus also the sum of the entries of $A(G)^N$, which is controlled by the largest eigenvalue. This is proven first for primitive graphs, whose adjacency matrices have a unique eigenvalue of maximum modulus and this eigenvalue is positive; in fact, for a primitive graph and a pair of states I, J, the number of paths of length N from I to J grows like $\lambda_{A(G)}^N = 2^{Nh(X_G)}$. For general graphs, one uses the decomposition into primitive, and then irreducible, graphs.

To extend this to a sofic shift Y, one uses a right resolving presentation $(G, \mathcal{L})$ of Y. Since every block of Y is the label of at most most $\mathcal{V}(G)$ paths in G, it follows that:

Theorem 3 *Let $\mathcal{G} = (G, \mathcal{L})$ be a right-resolving labeled graph presenting a sofic shift Y. Then $h(Y) = h(X_G)$.*

From Fig. 3, we see that the golden mean shift is obtained as a right resolving presentation of the graph with adjacency matrix:

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} .$$

A computation shows that λ_A is the golden mean, and so the entropy of the golden mean shift is the log of the golden mean; this is one explanation of the meaning of the term golden mean shift. From Fig. 4, we see that the even shift has the same entropy.

One cannot overstate the importance of entropy as an invariant. Yet, it is somewhat crude; it is perhaps not surprising that a single numerical invariant would not be sufficient to completely capture the many intricacies of shift spaces, even of sofic shifts or SFT's.

An invariant finer than entropy is the zeta function, which combines information regarding the numbers of periodic sequences of all periods, described as follows.

For a shift space X, let $p_n(X)$ denote the number of points in X of period n (i. e., the number of $x \in X$ such that $\sigma^n(x) = x$). It is straightforward to show that each $p_n(X)$ is an invariant. Since distinct periodic sequences define distinct blocks, it follows that

$$\limsup_{n\to\infty} \frac{1}{n} \log p_n(X) \le h(X) .$$

The inequality can be strict; for example, there are shift spaces (such as the direct product of the Morse shift with the full 2-shift) with positive entropy but no periodic points at all.

However, for irreducible SFT's and sofic shifts, the entropy $h(X)$ can be recovered from the sequence $p_n(X)$. The key to understanding this is the fact that for an edge shift $X = X_A$, $p_n(X) = \mathrm{tr}(A^n)$ and thus is the sum of the nth powers of the (non-zero) eigenvalues of A; in the case that A is primitive, the largest eigenvalue λ_A strictly dominates the other eigenvalues, and thus for large n,

$$\log p_n(X) = \log \mathrm{tr}(A^n) \sim n \log\lambda_A \sim nh(X) .$$

This shows that for a mixing SFT, the entropy equals the growth rate of numbers of periodic points. In fact, this result applies to all SFT's and sofic shifts.

Theorem 4 *For a sofic shift X,*

$$\limsup_{n\to\infty} \frac{1}{n} \log p_n(X) = h(X) .$$

The lim sup turns out be a limit in the case that X is a mixing sofic shift.

Most of what we have stated for $p_n(X)$ here applies equally well to $q_n(X)$, the number of points of *least* period n in X. This follows from the fact that "most" periodic points of period n have least period n.

The periodic point information can be conveniently combined into a single invariant, known as the *zeta function*. For a shift space X,

$$\zeta_X(t) = \exp\left(\sum_{n=1}^{\infty} \frac{p_n(X)}{n}\, t^n\right) .$$

For an edge shift X_A, one computes the zeta function to be the reciprocal of a polynomial:

$$\zeta_{X_A}(t) = \frac{1}{t^r\, \chi_A(t^{-1})} = \frac{1}{\det(I - tA)} ,$$

which is completely determined by the non-zero eigenvalues (with multiplicity) of A (Bowen and Lanford [21]). As an example, the zeta function of the golden mean shift is $1/(1-t-t^2)$.

The discussion above applies equally well to SFT's since they are conjugate to edge shifts. For a sofic shift, the zeta function turns out to be a rational function, i. e., quotient of two polynomials. This can be shown by analyzing properties of a right resolving presentation of the sofic shift. From this it turns out that the zeta function of the even shift is $(1-t)/(1-t-t^2)$. The technique for computing zeta functions of sofic shifts was developed by Manning [80] (actually, Manning developed the technique to compute zeta functions of hyperbolic dynamical systems).

So, for sofic shifts, all of the periodic point information is determined by a finite collection of complex numbers, namely the zeros and poles of the zeta function.

Finally, we mention another simple invariant obtained from the periodic points. The *period*, $\mathrm{per}(X)$, of a shift space X is the gcd of lengths of periodic points in X, i. e., the gcd of the set of n such that $p_n(X) \neq 0$. If $X = X_G$ is an edge shift and G is irreducible, then $\mathrm{per}(X_G) = \mathrm{per}(G)$, which is defined to be the gcd of cycle lengths in G and coincides with the gcd of the lengths of all cycles in G based at any given state. For an irreducible graph with period p and any state I, the number of cycles of length N, a multiple of p, at I grows like $\lambda_{A(G)}^N = 2^{Nh(X_G)}$. Also, an irreducible graph G is primitive if $\mathrm{per}(G) = 1$.

The Conjugacy Problem

The conjugacy problem for SFT's and sofic shifts is a major open problem. After much effort, it remains unsolved today. Much of what we know goes back to R. Williams [131] in the 1970s. One of Williams' main results was that any conjugacy can be decomposed into simple building blocks, as follows.

Let A and B be nonnegative integral matrices, with associated graphs G and H. An *elementary equivalence from A to B* is a pair (R, S) of rectangular nonnegative integral matrices satisfying

$$A = RS\,, \quad B = SR\,. \tag{1}$$

In this case we write $(R, S)\colon A \equiv B$. A *strong shift equivalence of lag ℓ from A to B* is a sequence of ℓ elementary equivalences

$$\begin{aligned}
&(R_1, S_1)\colon A = A_0 \equiv A_1\,,\\
&(R_2, S_2)\colon A_1 \equiv A_2\,,\\
&\ldots\,,\\
&(R_\ell, S_\ell)\colon A_{\ell-1} \equiv A_\ell = B\,.
\end{aligned}$$

In this case we write $A \approx B(\mathrm{lag}\,\ell)$. We say that A is *strong shift equivalent to B* (and write $A \approx B$) if there is a strong shift equivalence of some lag from A to B.

Via the matrix Equation (1), one creates a graph K whose state set is the disjoint union of $\mathcal{V}(G)$ and $\mathcal{V}(H)$; for each $I \in V_G$ and $J \in V_H$, the graph has R_{IJ} edges from I to J and S_{JI} edges from J to I. The equation $A = RS$ allows one to associate each edge e of G with a unique path $r(e)s(e)$ of length two running from $\mathcal{V}(G)$ to $\mathcal{V}(H)$ to $\mathcal{V}(G)$. Similarly, the equation $B = SR$ allows one to associate each edge e of H with a path $s(e)r(e)$ of length two running from $\mathcal{V}(H)$ to $\mathcal{V}(G)$ to $\mathcal{V}(H)$. One can show that the 2-block sliding block code defined by $\Phi(ef) = s(e)\,r(f)$ defines a conjugacy from X_A to X_B, and so whenever $A \approx B$, $X_A \cong X_B$. Williams proved this and its converse:

Theorem 5 *(R. Williams [131]) The edge shifts X_A and X_B are conjugate if and only if A and B are strong shift equivalent.*

In fact, Williams showed that any conjugacy can be decomposed into a composition of conjugacies defined by elementary equivalences. This can be interpreted in a way that shows that X_A and X_B are conjugate if and only if we can pass from G to H by a sequence of state splittings and amalgamations, defined as follows.

Let G be a graph with states $\mathcal{V}$ and edges $\mathcal{E}$. For each $I \in \mathcal{V}$, partition the outgoing edges from I into disjoint nonempty sets $\mathcal{E}_I^1, \mathcal{E}_I^2, \ldots, \mathcal{E}_I^{m(I)}$. Let $\mathcal{P}$ denote the resulting partition of $\mathcal{E}$, and let $\mathcal{P}_I$ denote the partition $\mathcal{P}$ restricted to $\mathcal{E}_I$. The *out-split graph H formed from G using $\mathcal{P}$* has states $I^1, I^2, \ldots, I^{m(I)}$, where I ranges over the states in $\mathcal{V}$, and edges e^j, where e is any edge in $\mathcal{E}$ and

$1 \le j \le m(t(e))$. If $e \in \mathcal{E}$ goes from I to J, then $e \in \mathcal{E}_I^i$ for some i, and we define the e^j to have initial state I^i and terminal state J^j, The 2-block code generated by $\Phi(ef) = e^j$, where $f \in \mathcal{E}_{t(e)}^j$, defines a conjugacy, called an out-splitting code, from X_G to X_H. Similarly, one defines an in-splitting code. Inverses of these conjugacies are called out-amalgamation and in-amalgamation codes.

Figure 5 depicts an out-splitting. The graph (a) on the left has three states I, J, K and the partition elements that define the splitting are $\mathcal{E}_I^1 = \{a\}, \mathcal{E}_I^2 = \{b, c\}, \mathcal{E}_J^1 = \{d\}$, $\mathcal{E}_K^1 = \{e\}, \mathcal{E}_K^2 = \{f\}$; the graph (b) is the resulting split graph.

By interpreting state splitting and amalgamation in terms of adjacency matrices one can show that such operations generate elementary equivalences. And one can decompose elementary equivalences into splittings and amalgamations. It follows that:

Theorem 6 *(R. Williams [131]) Every conjugacy from one edge shift to another is the composition of splitting codes and amalgamation codes.*

This classification for edge shifts naturally extends to SFT's since every SFT is conjugate to an edge shift. It also extends to sofic shifts, and we describe this in the context of irreducible sofic shifts.

Recall that an irreducible sofic shift has a unique minimal right resolving presentation. Any labeled graph can be completely described by a symbolic adjacency matrix, which records the transitions (edges) in the underlying physical graph, as well as the labels of the edges. Namely, the symbolic adjacency matrix is indexed by the states of the underlying graph and the (I, J)-entry is the formal sum of the labels of edges from I to J. It turns out that the notions of elementary equivalence, and hence strong shift equivalence, can be extended to more general categories, in particular to symbolic adjacency matrices.

Theorem 7 *(Krieger [72], Nasu [97]) Let X and Y be irreducible sofic shifts. Let A and B be the symbolic adjacency matrices of the minimal right-resolving presentations of X and Y, respectively. Then X and Y are conjugate if and only if A and B are strong shift equivalent.*

The classification, provided by these results, would be of limited use if the story ended here. Fortunately, Williams showed that strong shift equivalence yields a strong, delicate and somewhat computable necessary condition for conjugacy.

Let A and B be nonnegative integral matrices and $\ell \ge 1$. A *shift equivalence* of lag ℓ is a pair (R, S) of rectangular nonnegative integral matrices satisfying the shift equivalence equations

$$AR = RB\,, \quad SA = BS\,, \quad A^\ell = RS\,, \quad B^\ell = SR\,.$$

We denote this situation by $(R, S)\colon A \sim B(\text{lag } l)$. We say that A is *shift equivalent* to B, written $A \sim B$, if there is a shift equivalence from A to B of some lag. It is not hard to see that an elementary equivalence is a shift equivalence of lag 1 and that shift equivalence is an equivalence relation. It follows that:

Theorem 8 *(Williams [131]) Strong shift equivalence implies shift equivalence. More precisely, if $A \approx B(\text{lag } \ell)$, then $A \sim B(\text{lag } \ell)$.*

Recall from Sect. "Entropy and Periodic Points" that for an edge shift X_A, the set of nonzero eigenvalues, with multiplicity, of A determines the zeta function and hence this set is an invariant of conjugacy. Using the shift equivalence equations, one can show more: the entire Jordan form corresponding to the nonzero eigenvalues is an invariant. This information depends only on properties of the adjacency matrix considered as a linear transformation (over $\mathbf{R}$ or $\mathbf{Q}$). However, A is a nonnegative, integral matrix and both nonnegativity and integrality provide substantially more information. One such invariant, that follows from shift equivalence and makes use of integrality is the Bowen–Franks group [20], $BF(A) = Z^r/Z^r(I - A)$.

Until recently all of the information contained in known conjugacy invariants, such as those above, was subsumed in shift equivalence. And Kim and Roush showed that shift equivalence is decidable [60,61], meaning that there is a finite decision procedure via a Turing machine that decides whether two given edge shifts, and therefore two given SFT's, are conjugate (they also showed that a notion of shift equivalence for sofic shifts, formulated by Boyle and Krieger [26], is decidable [62]). So, a central focus of the subject was the question: is shift equivalence a complete invariant of conjugacy? The answer turns out to be No, as proven by Kim and Roush [63]; see also the survey article [125]. However, it is a complete invariant of a weaker form of conjugacy; we say that X_A and X_B are *eventually conjugate* if all sufficiently large powers are conjugate. It is not hard to show that if $A \sim B$, then X_A and X_B are eventually conjugate, and the converse is true as well:

Theorem 9 *(Kim and Roush [60], Williams [131]) Edge shifts X_A and X_B are eventually conjugate if and only if A and B are shift equivalent.*

Also the Kim–Roush counterexamples and subsequent work do not bear on a special case of the conjugacy problem: if A is shift equivalent to the 1×1 matrix $[n]$, is

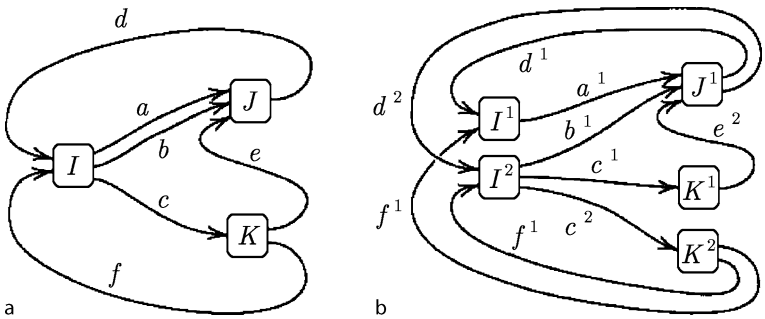

Symbolic Dynamics, Figure 5
A state splitting

X_A conjugate to the full n-shift? This question, known as the *little shift equivalence problem*, is particularly intriguing because in this case the condition $A \sim [n]$ is simply the statement that A has exactly one non-zero eigenvalue, namely n.

Shift equivalence can be characterized in another way that has turned out to be very useful. Let A be an $r \times r$ integral matrix. Let R_A denote the real eventual range of A, i.e, $R_A = R^r A^r$. The *dimension group* of A is defined:

$$\Delta_A = \{v \in R_A : vA^k \in \mathbf{Z}^r \text{ for some } k \geq 0\} .$$

The *dimension group automorphism* δ_A of A is the restriction of A to Δ_A, so that $\delta_A(\mathbf{v}) = \mathbf{v}A$ for $\mathbf{v} \in \Delta_A$. The *dimension pair* of A is (Δ_A, δ_A).

If A is also nonnegative, then we define the *dimension semigroup* of A to be

$$\Delta_A^+ = \{v \in R_A : vA^k \in (Z^+)^r \text{ for some } k \geq 0\} .$$

The *dimension triple* of A is $(\Delta_A, \Delta_A^+, \delta_A)$.

It can be shown that the dimension triple completely characterizes shift equivalence, i. e., two nonnegative integral matrices are shift equivalent if and only if their dimension groups are isomorphic by an isomorphism that preserves the dimension semigroup and intertwines the dimension group automorphisms. Also, by associating equivalence classes of certain subsets of the shift space X_A to elements of Δ_A^+, one can interpret the dimension triple in terms of the action of the shift map on X_A [27,68]. And the dimension triple arises prominently in the study of the automorphism group of an SFT, which we now briefly describe. The dimension group for SFT's was developed by Krieger [68,69].

In many areas of mathematics, objects are studied by means of their symmetries. This holds true in symbolic dynamics, where symmetries are expressed by automorphisms. An *automorphism* of a shift space X is a conjugacy from X to itself. The set of all automorphisms of a shift space X is a group under composition, and is naturally called the *automorphism group*, denoted aut(X).

The goals are to understand aut(X) as a group (What kinds of subgroups does it contain? How "big" is it?) and how it acts on X, e. g., given shift-invariant subsets U, V, such as finite sets of periodic points, when is there an automorphism of X that maps U to V?. One might hope that the automorphism group would shed new light on the conjugacy problem for SFT's. Indeed, tools developed to study the automorphism group eventually paved the way for Kim and Roush to find examples of shift equivalent matrices that are not strong shift equivalent. On the other hand, the automorphism group cannot tell the entire story. For instance, aut(X_A) and aut($X_{A^\top}$) are isomorphic, since any automorphism read backwards can be viewed as an automorphism of the transposed shift, yet X_A and $X_{A^\top}$ may fail to be conjugate (for an example due to Kollmer, see p. 81 of [105]). It is not even known if the automorphism groups of the full 2-shift and the full 3-shift are isomorphic.

A good deal of our understanding of the action of the automorphism group on an SFT comes from understanding its induced representation as an action of the dimension group; this action is known as the *dimension representation*. For a much more thorough exposition on aut(X), we refer the reader to Wagoner [125,126].

Other Coding Problems

The difficulties encountered in attempts to solve the conjugacy problem motivated the formulation and study of weaker, but meaningful, notions of equivalence. For instance, we might say that two shift spaces are equivalent if one can be invertibly encoded to the other by some kind of "finite-state machine". A precise version of this is as follows.

Shift spaces X and Y are *finitely equivalent* if there is an SFT W together with finite-to-one factor codes $\phi_X\colon W \to X$ and $\phi_Y\colon W \to Y$. We call W a *common extension*, and ϕ_X, ϕ_Y the *legs*.

Here, by "finite-to-one" we mean merely that each point has a finite number of inverse images. It can be shown that any finite-to-one factor code from one shift space to another must preserve entropy, and so entropy is an invariant of finite equivalence. For irreducible sofic shifts, entropy is a complete invariant:

Theorem 10 *(Parry [100]) Two irreducible sofic shifts are finitely equivalent if and only if they have the same entropy.*

Note that from this result and the fact that finite-to-one codes between general shift spaces preserve entropy, for irreducible sofic shifts, we could have just as well defined finite equivalence with the common extension W merely being a shift space. However, if W is an SFT, we get a more concrete coding interpretation as follows. First, we recode W to an edge shift X_G and recode the legs, $\phi_X = (\Phi_X)_\infty$ and $\phi_Y = (\Phi_Y)_\infty$, to one-block codes, and (with a bit more argument) we can assume that G is irreducible. In this set-up, the finite-to-one condition translates to the so-called "no-diamond" condition, which means that for any given pair of states I, J and finite sequence w, there is at most one path from I to J with label w [33,34]. Since, for any fixed state I, the number of cycles of length n, a multiple of $p = \mathrm{per}(G)$, at I grows like $2^{nh(W)}$, we have, in this set-up a means to invertibly encode a "large" set of allowed blocks in X to allowed blocks in Y: namely, fix state I, and a large n, which is a multiple of p; for any cycle γ of length n at state I encode the Φ_X-label of γ to the Φ_Y-label of γ.

For encoding and decoding, one can dispense with state information if the legs are "almost invertible". A factor code ϕ is *almost invertible* if it is one-to-one on sequences that are typical in the following sense: x is typical if every allowed block appears infinitely often in x both to the left and the right. We then say that shift spaces X and Y are *almost conjugate* if there is an SFT W and almost invertible factor codes $\phi_X\colon W \to X$, $\phi_Y\colon W \to Y$. We call (W, ϕ_X, ϕ_Y) an *almost conjugacy* between X and Y.

For irreducible sofic shifts, it can be shown that any almost invertible factor code is finite-to-one, and so almost conjugacy implies finite equivalence. Thus, entropy is again an invariant, and together with a second very mild invariant, it is complete:

Theorem 11 *(Adler–Marcus [4]) Let X and Y be irreducible sofic shifts with minimal right resolving presentations $(G, \mathcal{L})$ and $(H, \mathcal{M})$. Then X and Y are almost conjugate if and only if $h(X) = h(Y)$ and $\mathrm{per}(G) = \mathrm{per}(H)$.*

In particular, if X and Y are mixing, then $\mathrm{per}(G) = 1 = \mathrm{per}(H)$ and entropy itself is a complete invariant.

Thus, with an an almost conjugacy, one can invertibly encode most sequences in X to those of Y without the need for auxiliary state information. Moreover, if X and Y are almost conjugate, then there is an almost conjugacy of X and Y in which one leg is right-resolving and the other leg is left-resolving (and the common extension is irreducible [4]). This gives an even more concrete interpretation to the encoding.

The proofs of Theorems 10 and 11 are actually quite constructive. For illustration we consider a very special, but historically important, case.

Let G be a graph with constant out-degree n. A *road coloring* Φ is a labeling of G such that at each state of G, each symbol $0, \ldots, n-1$ appears exactly once as the label of an outgoing edge. An n-ary word w is *synchronizing* if all paths that are labeled w end at the same state. Figure 6 gives examples of road-colorings, with $n = 2$.

For a road-coloring, a binary word may be viewed as a sequence of instructions given to drivers starting at each of the states. A synchronizing word is a word that drives everybody to the same state. For instance, in Fig. 6a, the word 11 drives everybody to the state in the lower-left corner. But Fig. 6b does not have a synchronizing word because whenever a driver takes a '0' road he stays where he is and whenever a driver takes a '1' road he rotates by 120 degrees. The road-coloring in Fig. 6c is essentially the only road-coloring of its underlying graph, and there is no synchronizing word because drivers must always oscillate between the two states.

Now, let X be the full n-shift and Y be an irreducible SFT with entropy $\log n$. Suppose that we could find a presentation $(G, \mathcal{L})$ of Y with G having constant out-degree n and $\mathcal{L}_\infty$ finite-to-one. Then, define Φ to be any road coloring of G. Then Φ_∞ would be a finite-to-one (in fact, right resolving!) factor code from X_G to X. And we would obtain a finite equivalence with $X = X_G$, $\phi_Y = \mathcal{L}_\infty$ and $\phi_X = \Phi_\infty$. If, moreover, we could choose $\mathcal{L}$ and Φ such that ϕ_Y and ϕ_X are almost invertible, then we would have an almost conjugacy.

It turned that this could be arranged for ϕ_Y via a construction related to state splitting [2]. And if Y were mixing, then G could be chosen to be primitive and ϕ_Y almost invertible. In this setting, $\phi_X = \Phi_\infty$ would be almost invertible iff Φ has a synchronizing word; the sufficiency follows from the fact that every bi-infinite binary sequence which contains w infinitely often to the left would be the label of exactly one bi-infinite sequence of edges (to see this, use the synchronizing and road-coloring properties).

The construction of such a labeling Φ became known

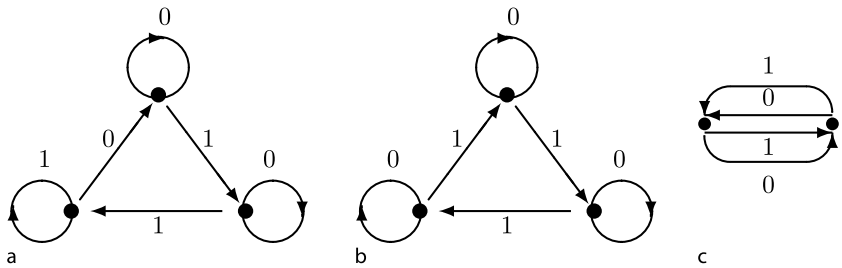

Symbolic Dynamics, Figure 6
Some road-colorings

as the Road Problem, which remained open for thirty years. In the meantime, a weaker version of the the road problem was solved and applied to yield this special case of Theorem 11 (see [2]). Nevertheless, the problem remained an important problem in graph/automata theory and was only recently solved:

Theorem 12 *(Road theorem (Trachtman [122])) If G is a finite directed primitive graph with constant out-degree n, there a road-coloring of G which has a synchronizing word.*

Trachtman's approach relies heavily on earlier work of Friedman [44] and Kari [57].

The primitivity assumption above is close to necessary. Clearly some kind of connectivity is required and in the presence of irreducibility, primitivity would be necessary since otherwise there would be a "phase" introduced in the graph that would never allow a word to synchronize, as in Fig. 6c.

So far, we have focused on equivalences between symbolic systems. There has also been considerable attention paid to problems of embedding one system into another and factoring one onto another.

One of the most striking results of this type is the Krieger embedding theorem. It is not hard to show that any proper subshift of an irreducible SFT must have strictly smaller entropy. Thus, a necessary condition for a proper embedding of a shift space into an irreducible SFT is that it have strictly smaller entropy. This condition, together with a trivially necessary condition on periodic points, turns out to be sufficient. Recall that $q_n(X)$ denotes the number of points of least period n in X.

Theorem 13 *(Embedding Theorem (Krieger [70])) Let X and Y be irreducible shifts of finite type. Then there is a proper embedding of X into Y if and only if $h(X) < h(Y)$ and for each $n \geq 1$, $q_n(X) \leq q_n(Y)$.*

In fact, Krieger's theorem shows that these conditions are necessary and sufficient for a proper embedding of any shift space into a mixing shift space. The analogous problems for embedding into irreducible or mixing sofic shifts are still open, though there are partial results [22].

Using the embedding theorem and other tools from symbolic dynamics, Boyle and Handelman [25] obtained a stunning application to linear algebra: namely, a complete characterization of the non-zero spectra of primitive matrices over **R**. In fact, they obtained characterizations of non-zero spectra for primitive matrices over many other subrings of **R**. While they did not obtain a complete characterization over **Z**, they formulated a conjecture for **Z** and obtained many partial results towards that conjecture, which was later proven using other tools. The result, stated below, shows that three simple necessary conditions on a set of nonzero complex numbers are actually sufficient. In order to state these conditions, we need the following notation:

Let $\Lambda = \{\lambda_1, \ldots, \lambda_k\}$ be a list of nonzero complex numbers (with multiplicity). Let $f_\Lambda(t) = \prod_{i=1}^k (t - \lambda_i)$, and $\mathrm{tr}_n(\Lambda) = \sum_{d/n} \mu(n/d) \sum_{i=1}^k \lambda_i^k$, with μ being the Mobius Inversion function.

- Integrality Condition: $f_\Lambda(t)$ is a monic polynomial (with integer coefficients).
- Perron Condition: There is a positive entry in Λ, occurring just once, that strictly dominates in absolute value all other entries. We denote this entry by λ_Λ.
- Net Trace Condition: $\mathrm{tr}_n(\Lambda) \geq 0$ for all $n \geq 1$.

Theorem 14 *(Kim–Ormes–Roush [64]) Let Λ be a list of nonzero complex numbers satisfying the Integrality, Perron, and Net Trace Conditions. Then there is a primitive integral matrix A for which Λ is the non-zero spectrum of A.*

These conditions are all indeed necessary for Λ to be the non-zero spectrum of a primitive integral matrix. The integrality condition is that Λ forms a complete set of algebraic conjugates; the Perron condition states that Λ must satisfy the conditions of the Perron–Frobenius theorem for primitive matrices; and the Net Trace condition assures that the number of periodic points of least period n would be nonnegative.

We now turn from embeddings to factors. One special case, which is somewhat related to the Road Problem above and also important for data recording applications (Sect. "Coding for Data Recording Channels") is:

Theorem 15 [1,81] *An SFT X factors onto the full n-shift iff $h(X) \geq \log(n)$.*

While this special case treats both the equal entropy case ($h(X) = \log(n)$) and unequal entropy case ($h(X) > \log(n)$), in general, the factor problem naturally divides into two cases: lower entropy factors and equal entropy factors. In either case, a trivial necessary condition for Y to be a factor of X is that whenever $q_n(X) \neq 0$, there exists a d/n such that $q_d(Y) \neq 0$. This condition is denoted $P(X) \searrow P(Y)$. Building on ideas from Krieger's embedding theorem, this necessary condition was shown to be sufficient.

Theorem 16 *(Lower entropy factor theorem (Boyle [22])) Let X and Y be irreducible SFT's with $h(X) > h(Y)$. Then there is a factor code from X to Y if and only if $P(X) \searrow P(Y)$.*

As with the embedding theorem, the lower entropy factors problem for irreducible sofic shifts is still open.

The equal entropy factors problem for SFT's is quite different. Clearly, $P(X) \searrow P(Y)$ is a necessary condition. A second necessary condition involves the dimension group and is simplest to state in the case of mixing edge shifts X_A and X_B.

We say that a subgroup Δ of the dimension group Δ_A is *pure* if whenever an integer multiple of an element $v \in \Delta_A$ is in Δ, then so is v; intuitively Δ does not have any "rational holes" in Δ_A. The condition is that there is a pure δ_A-invariant subgroup Δ of Δ_A such that (Δ_B, δ_B) is a quotient of $(\Delta, \delta_A|_\Delta)$.

In the equal entropy case, this condition and the trivial periodic point condition, $P(X) \searrow P(Y)$, subsume all known necessary conditions for the existence of a factor code from one mixing edge shift to another. It is not known if these two conditions are sufficient. For references on this problem, see [9,27,66].

Coding for Data Recording Channels

In magnetic recording, within any given clock cell, a '1' is represented as a change in magnetic polarity, while a '0' is represented as an absence of such a change. Two successive 1's (separated by some number, $m \geq 0$, of 0's) are read as a voltage peak followed by a voltage trough (or vice versa). If the peak and trough occur too close together, *intersymbol interference* can occur: the amplitudes of the peak and trough are degraded, and the positions at which they occur are distorted. In order to control intersymbol interference, it is desirable that 1's not be too close together, or equivalently that runs of 0's not be too short. On the other hand, for timing control, it is desirable that runs of 0's not be too long; this is a consequence of the fact that only 1's are observed: the length of a run of 0's is inferred by connection to a clock via a feedback loop, and a long run of 0's could cause the clock to drift more than one clock cell.

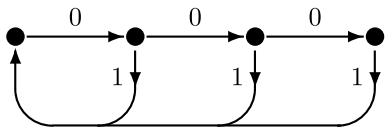

Symbolic Dynamics, Figure 7
***X*(1, 3)**

This gives rise to *run-length-limited shift spaces*, $X(d, k)$, where runs of 0's are constrained to be bounded below by some positive integer d and bounded above by some positive $k \geq d$. More precisely, $X(d, k)$ is defined by the constraints that 1's occur infinitely often in each direction, and there are at least d 0's, but no more than k 0's, between successive 1's. Note that $X(d, k)$ is an SFT with forbidden list $\mathcal{F} = \{0^{k+1}, 10^i1, 0 \leq i < d\}$. Figure 7 depicts a labeled graph presentation of $X(1, 3)$. The SFT's $X(1, 3)$, $X(2, 7)$, and $X(2, 10)$ have been used in floppy disks, hard disks, and the compact audio disk, respectively.

Now in order to record completely arbitrary information, we need to build an encoder which encodes arbitrary binary sequences into sequences that satisfy a given constraint (such as $X(d, k)$). The encoder is a finite-state machine, as depicted in Fig. 8. It maps arbitrary binary data sequences, grouped into blocks of length p (called p-blocks), into constrained sequences, grouped into blocks of length q (called q-blocks). The encoded q-block is a function of the p-block as well as an internal state. When concatenated together, the sequence of encoded q-blocks must satisfy the given constraint. Also, the encoded sequences should be decodable, meaning that given the initial encoder state, a string of p-blocks can be

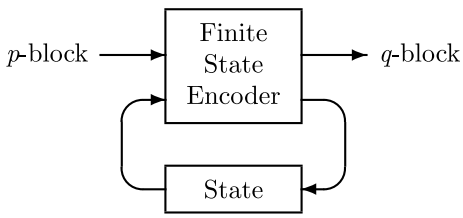

Symbolic Dynamics, Figure 8
Encoder

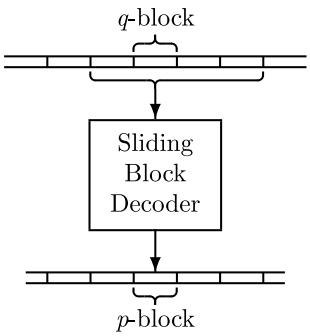

Symbolic Dynamics, Figure 9
Sliding block decoder

recovered from its encoded string of q-blocks, possibly allowing a fixed delay in time.

If the constrained sequences satisfy the constraints of a sofic shift X, we say that such a code is a *rate p:q finite-state code into X*. In terms of symbolic dynamics, such a code consists of an edge shift X_G, a right resolving factor code ϕ_1 from X_G onto the full 2^p-shift and a right closing sliding block code ϕ_2 into X^q (the right closing condition expresses the decodability condition).

In most applications, it is important that a stronger decoding condition be imposed. Namely, a *sliding block decodable* rate p:q finite-state code into X consists of a finite-state code given by (X_G, ϕ_1, ϕ_2) and a sliding block code $\psi : X^q \to X_{2^p}$ such that $\psi \circ \phi_2 = \phi_1$. This means that the decoded p-block depends only upon the local context of the received q-block – that is, decoding is accomplished by applying a time-invariant function to a window consisting of a bounded amount of memory and/or anticipation, but otherwise is state-independent (see Fig. 9, which depicts the situation where the memory is 1 and the anticipation is 2). The point is that whenever the window of the decoder passes beyond a raw channel error, that error cannot possibly affect future decoding; thus, sliding block decoders control error propagation.

Symbolic dynamics has played an important role in providing a framework for constructing such codes as well as for establishing bounds on various figures of merit for such codes. Specifically, by modifying the constructions used in the proofs of Theorems 10 and 11, in the early 1980's, Adler, Coppersmith and Hassner (ACH) established the following results:

Theorem 17 *(Finite-state coding theorem [1]) Let X be a sofic shift and p, q be positive integers. Then there is a rate p:q finite-state code into X if and only if $p/q \le h(X)$.*

Theorem 18 *(Sliding block decoding theorem [1]) Let X be an SFT and p, q be positive integers. Then there is a rate p:q sliding-block decodable finite-state code into X if and only if $p/q \le h(X)$.*

We have described all of this in the context of the binary alphabet for data sequences. In fact, it works just as well for any finite alphabet, and the method used to prove Theorem 18 solved, at the same time, a special case of the factor problem for SFT's: namely, Theorem 15 above. Sometimes, Theorems 17 and 18 are stated only for rate 1 : 1 codes but in the context of arbitrary finite alphabets (e. g. see Chap. 5 in [78]), this easily extends to the general $p : q$ case by passing to powers.

The ACH paper was the beginning of a rigorous theory of constrained coding. Theorem 18 has been extended to a large class of sofic shifts, and bounds have been established on such figures of merit as number of encoder states as well as the size of the decoding window of the sliding block decoder. While the state-splitting algorithm does construct codes with "relatively small" decoding windows, it is not yet understood how to construct codes with the smallest such windows. This is of substantial engineering interest since the smaller the decoding window, the smaller the error propagation. There is now a substantial literature on the construction of these types of codes, including the state-splitting algorithm as well as many other methods of encoder/decoder design and a wealth of examples that go well beyond the run length limited constraints. See for example the expositions [11,54], and [83] as well as the papers [7,8,10,12,31,41,42,43,53,56].

Connections with Information Theory and Ergodic Theory

The concept of entropy was developed by Shannon [119] in information theory in the 1940's and was adapted to ergodic theory in the 1950's and to dynamical systems, and in particular symbolic dynamics, in the 1960's. Shannon focused on entropy for random variables and finite sequences of random variables, but the concept naturally extends to stationary stochastic processes, in particular stationary Markov chains. Roughly speaking, for a stationary process μ, the entropy $h(\mu)$ is the asymptotic growth rate of the number of allowed sequences, weighted by the joint stationary probabilities of μ (for background on entropy and information theory see Cover and Thomas [35]).

A Markov chain on a graph G is defined by assigning transition probabilities to the edges. Let P denote the stochastic matrix indexed by states of G, with P_{IJ} equal to the sum of the transition probabilities of all edges from I to J. This, together with a stationary vector for P, completely defines the (joint distributions of the) Markov chain.

So, a graph itself can be viewed as specifying only which transitions are possible. For that reason, one could view G as an "intrinsic Markov chain", and Parry [99] used this terminology when he introduced SFT's based on graphs and matrices.

More generally, a stationary process on a shift space X is any stationary process which assigns positive probability only to allowed blocks in X.

For an irreducible graph G with strictly positive transition probabilities on all edges, by Perron–Frobenius theory, P will always have a unique stationary vector. And there is a particular such Markov chain μ_G on G that in some sense distributes probabilities on paths as uniformly as possible. This Markov chain is the most "random" possible stationary process (not just among Markov chains) on X_G in the sense that it has maximal entropy; moreover, its entropy coincides with the topological entropy of X_G.

The Markov chain μ_G is defined as follows: let λ denote the largest eigenvalue of $A(G)$ and w, v denote corresponding left, right eigenvectors, normalized such that $w \cdot v = 1$; for any path γ of length n from state I to state J, we define the stationary probability of γ:

$$\mu_G(\gamma) = \frac{w_I v_J}{\lambda^n} .$$

It is clear from the formula that this distribution is fairly uniform, since all paths with the same initial state, terminal state and length have the same stationary probability.

This defines a joint stationary distribution on allowed blocks of X_G, and it is not hard to show that it is consistent and Markov, given by assigning transition probability $v_J/(v_I\,\lambda)$ to each edge from I to J. To summarize:

Theorem 19 *Let G be an irreducible graph.*

- *Any stationary process μ on G satisfies $h(\mu) \le \log \lambda$, and*
- *The Markov chain μ_G is the unique stationary process on G such that $h(\mu_G) = \log \lambda$.*

The construction of μ_G is effectively due to Shannon [119] and uniqueness is due to Parry [99] The unique entropy-maximizing stationary process on irreducible edge shifts naturally extends to irreducible SFT's and irreducible sofic shifts (this process will be M-step Markov for an M-step SFT).

Many results in symbolic dynamics were originally proved using μ_G. One example is the fact that any factor code from one irreducible SFT to another of the same entropy must be finite-to-one [32]. This result, and many others, were later proven using methods that rely only on the basic combinatorial structure of G and X_G. Nevertheless, μ_G provides much motivation and insight into symbolic dynamics problems, and it illustrates a connection with ergodic theory, which we now discuss.

For background on ergodic theory (such as the concepts of measure-preserving transformations, homomorphisms, isomorphisms, ergodicity, mixing, and measure-theoretic entropy), we refer the reader to [107,113,127]. Observe that a stationary process on a shift space X can be viewed as a measure-preserving transformation: the transformation is the shift mapping and the measure on "cylinder sets", consisting of $x \in X$ with prescribed coordinate values $x_i = a_i, \ldots, x_j = a_j$, is defined as the probability of the word $a_i \ldots a_j$; the stationarity of the process translates directly into preservation of the measure. The symbolic dynamical concepts of irreducibility and mixing correspond naturally to the concepts of ergodicity and (measure-theoretic) mixing in ergodic theory.

It is well-known [107,127] that the measure-preserving transformation (MPT) defined by a stationary Markov chain μ on an irreducible (resp., primitive) graph G is ergodic (resp., mixing) if μ assigns strictly positive conditional probabilities to all edges of G.

Now, suppose that G and H are irreducible graphs and $\phi\colon X_G \to X_H$ is a factor code. For a stationary measure μ on G, we define a stationary measure $\nu = \phi(\mu)$ on X_H by transporting μ to X_H: for a measurable set A in X_H, define

$$\nu(A) = \mu\left(\phi^{-1}(A)\right) .$$

Then ϕ defines a measure-preserving homomorphism from the MPT defined by μ to the MPT defined by ν. Since measure-preserving homomorphisms between MPT's cannot reduce measure-theoretic entropy, we have $h(\nu) \le h(\mu)$.

Suppose that now ϕ is actually a conjugacy. Then it defines a measure-preserving isomorphism, and so $h(\mu) = h(\nu)$. If $\mu = \mu_G$, then by uniqueness, we have $\nu = \mu_H$. Thus ϕ defines a measure-theoretic isomorphism between the MPT defined by μ_G and the MPT defined by μ_H. In fact, this holds whenever ϕ is merely an almost invertible factor code. This establishes the following result:

Theorem 20 *Let G, H be irreducible graphs, and let μ_G, μ_H be the stationary Markov chains of maximal entropy on G, H. If X_G, X_H are almost conjugate (in particular, if they are conjugate), then the measure-preserving transformations defined by μ_G and μ_H are isomorphic.*

Hence conjugacies and almost conjugacies yield isomorphisms between measure-preserving transformations defined by stationary Markov chains of maximal entropy.

In fact, the isomorphisms obtained in this way have some very desirable properties compared to the run-of-the-mill isomorphism. For instance, an isomorphism ϕ between stationary processes typically has an infinite window; i. e., to know $\phi(x)_0$, you typically need to know all of x, not just a central block $x_{[-n,n]}$ (these are the kinds of isomorphisms that appear in general ergodic theory and in particular in Ornstein's celebrated isomorphism theory [98]). In contrast, by definition, a conjugacy always has a finite window of uniform size. It turns out that an isomorphism obtained from an almost conjugacy, as well as its inverse, has finite expected coding length in the sense that to know $\phi(x)_0$, you need to know only a central block $x_{[-n(x),n(x)]}$, where the function $n(x)$ has finite expectation [4]. In particular, by Theorem 11, whenever X_G and X_H are mixing edge shifts with the same entropy, the measure-preserving transformations defined by μ_G and μ_H are isomorphic via an isomorphism with finite expected coding length.

The notions of conjugacy, finite equivalence, almost conjugacy, embedding, factor code, and so on can all be generalized to the context of stationary measures, in particular to stationary Markov chains. For instance, a conjugacy between two stationary measures is a map that is simultaneously a conjugacy of the underlying shift spaces and an isomorphism of the associated measure-preserving transformations. Many results in symbolic dynamics have been generalized to the context of stationary Markov chains. There is a substantial literature on this, in particular on finitary isomorphisms with finite expected coding time, e. g., [71,84,101,114], and [90]. The expositions [102,105] give a nice introduction to the subject of strong finitary codings between stationary Markov chains. See also the research papers [45,85,86,103,104,123,124].

Higher Dimensional Shift Spaces

In this section we introduce higher dimensional shift spaces. For a more thorough introduction, we refer the reader to Lind [76]. For the related subject of tiling systems, see Robinson [112], Radin [110], and Mozes [94,95].

The *d-dimensional full $\mathcal{A}$-shift* is defined to be $\mathcal{A}^{\mathbf{Z}^d}$. Ordinarily, $\mathcal{A}$ is a finite alphabet, and here we restrict ourselves to this case. An element x of the full shift may be regarded as a function $x\colon \mathbf{Z}^d \to \mathcal{A}$, or, more informally, as a "configuration" of alphabet choices at the sites of the integer lattice $\mathbf{Z}^d$.

For $x \in \mathcal{A}^{\mathbf{Z}^d}$ and $F \subseteq \mathbf{Z}^d$, let x_F denote the restriction of x to F. The usual metric on the one-dimensional full shift naturally generalizes to a metric on $\mathcal{A}^{\mathbf{Z}^d}$ given by $\rho(x,y) = 2^{-k}$, where k is the largest integer such that $x_{[-k,k]^d} = y_{[-k,k]^d}$ (with the usual conventions when $x = y$ and $x_{\mathbf{0}} \neq y_{\mathbf{0}}$). In analogy with one dimension, according to this definition, two points are "close" if they agree on a large cube $[-k,k]^d$.

We define higher dimensional shift spaces, with the following terminology. A *shape* is a finite subset F of $\mathbf{Z}^d$, and a *pattern* f on a shape F is a function $f\colon F \to \mathcal{A}$. We say that X is a *d-dimensional shift space* (or *d-dimensional shift*) if it can be represented by a list $\mathcal{F}$ (finite or infinite) of "forbidden" patterns

$$X = X_{\mathcal{F}} = \left\{x \in \mathcal{A}^{\mathbf{Z}^d} : \sigma^n(x)_F \notin \mathcal{F} \text{ for all } n \in \mathbf{Z}^d \text{ and all shapes } F\right\}.$$

Just as in one dimension, we can equivalently define a shift space to be a closed (with respect to the metric ρ) translation-invariant subset of $\mathcal{A}^{\mathbf{Z}^d}$. Here "translation-invariance" means that $\sigma^n(X) = X$ for all $n \in \mathbf{Z}^d$, where σ^n is the translation in direction n defined by $(\sigma^n(x))_m = x_{m+n}$.

We say that a pattern f on a shape F *occurs* in a shift space X if there is an $x \in X$ such that $x_F = f$. Hence the analogue of the language of a shift space is the set of all occurring patterns.

A *d-dimensional shift of finite type* X is a subset of $\mathcal{A}^{\mathbf{Z}^d}$ defined by a finite list $\mathcal{F}$ of forbidden patterns. Just as in one dimension, a d-dimensional shift of finite type X can also be defined by specifying allowed patterns instead of forbidden patterns, and there is no loss of generality in requiring the shapes of the patterns to be the same. Thus we can specify a finite list $\mathcal{L}$ of patterns on a fixed shape F, and set

$$X = X_{\mathcal{L}^c} = \left\{x \in \mathcal{A}^{\mathbf{Z}^d} : \text{ for all } n \in \mathbf{Z}^d,\ \sigma^n(x)_F \in \mathcal{L}\right\}.$$

In fact, there is no loss in generality in assuming that F is a d-dimensional cube $F = [0,k]^d$.

Given a finite list $\mathcal{L}$ of patterns on a shape F, we say that a pattern f' on a shape F' is *$\mathcal{L}$-admissible* (in the shift of finite type $X = X_{\mathcal{L}^c}$) if each of its sub-patterns, whose shape is a translate of F, belongs to $\mathcal{L}$. Of course, any pattern which occurs in X is $\mathcal{L}$-admissible. But an $\mathcal{L}$-admissible pattern need not occur in X.

The analogue of vertex shift (or 1-step shift of finite type) in higher dimensions is defined by a collection of d transition matrices $A_1, \ldots, A_d$ all indexed by the same set of symbols $\mathcal{A} = \{1, \ldots, m\}$. We set

$$\Omega(A_1, \ldots, A_d) = \left\{x \in \{1, \ldots, m\}^{\mathbf{Z}^d} : A_i(x_n, x_{n+e_i}) = 1 \text{ for all } n, i\right\},$$

where e_i is as usual the ith standard basis vector and $A_i(a,b)$ denotes the (a,b)-entry of A_i. Such a shift space

is called a *matrix subshift*. When $d = 2$, this amounts to a pair of transition matrices A_1 and A_2 with identical vertex sets. The matrix A_1 controls transitions in the horizontal direction and the matrix A_2 controls transitions in the vertical direction. Note that any matrix subshift is a shift of finite type and, in particular, can be specified by a list $\mathcal{L}$ of patterns on the unit cube $F = \{(a_1, \ldots, a_n) \colon a_i \in \{0, 1\}\}$; specifically, $\Omega(A_1, \ldots, A_n) = X_{\mathcal{L}^c}$ where $\mathcal{L}$ is the set of all patterns $f \colon F \to \{1, \ldots, m\}$ such that if $n, n + e_i \in F$, then $A_i(f(n), f(n + e_i)) = 1$. When we speak of admissible patterns for a matrix subshift, we mean $\mathcal{L}$-admissible patterns with this particular $\mathcal{L}$. Just as in one dimension, we can recode any shift of finite type to a matrix subshift.

Higher dimensional SFT's can behave very differently from one dimension. For example, there is a simple method to determine if a one-dimensional edge shift, and therefore a one-dimensional shift of finite type, is nonempty, and there is an algorithm to tell, for a given finite list $\mathcal{L}$, whether a given block occurs in $X = X_{\mathcal{L}^c}$. The corresponding problems in higher dimensions, called the *nonemptiness problem* and the *extension problem*, turn out to be undecidable [14,111]; see also [67]. Even for two-dimensional matrix subshifts X, these decision problems are undecidable. On the other hand, there are some special classes where these problems are decidable. This class includes any two-dimensional matrix subshift such that A_1 commutes with A_2 and $A_2^{\top}$. For this class, any admissible pattern on a cube must occur, and so the nonemptiness and extension problems are decidable; see [87,88].

A point x in a d-dimensional shift X is *periodic* if its orbit $\{\sigma^n(x) \colon n \in \mathbf{Z}^d\}$ is finite. Observe that this reduces to the usual notion of periodic point in one dimension. Now an ordinary (one-dimensional) nonempty shift of finite type is conjugate to an edge shift X_G, where G has at least one cycle. Hence a one-dimensional shift of finite type is nonempty if and only if it has a periodic point. This turns out to be false in higher dimensions [14,111], and this fact is intimately related to the undecidability results mentioned above. While one can formulate a notion of zeta function for keeping track of numbers of periodic points, the zeta function is hard to compute, even for very special and explicit matrix subshifts, and, even in this setting, it is not a rational function[77].

In higher dimensions, the entropy of a shift is defined as the asymptotic growth rate of the number of occurring patterns in arbitrarily large cubes. In particular, for two-dimensional shifts it is defined by

$$h(X) = \lim_{n\to\infty} \frac{1}{n^2} \log |X_{[0,n-1]\times[0,n-1]}| \,,$$

where $X_{[0,n-1]\times[0,n-1]}$ denotes the set of occurring patterns on the square

$$[0, n-1] \times [0, n-1]$$

that occur in X. Recall from Sect. "Entropy and Periodic Points" that it is easy to compute the entropy of a (one-dimensional) shift of finite type using linear algebra. But in higher dimensions, there is no analogous formula and, in fact, other than the group shifts mentioned below, the entropies of only a very few higher dimensional shifts of finite type have been computed explicitly. Even for the two-dimensional "golden mean" matrix subshift defined by the horizontal and vertical transition matrices

$$A_1 = A_2 = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$$

an explicit formula for the entropy is not known. However, there are good numerical approximations to the entropy of some matrix subshifts (e. g., [30,96]). And recently the set of numbers that can occur as entropies of SFT's in higher dimensions has been characterized [51,52]; this characterization turns out to be remarkably different from the analogous characterization in one dimension [74].

One of the few 2-dimensional SFT's for which entropy has been computed is the domino tiling system (see [58], Chap. 5 in [115]), which consists of all possible tilings of the plane using the 1×2 and 2×1 dominoes. This can be translated into an SFT with four symbols L, R, T, B, subject to the constraints:

- $x_{i,j} = L \Rightarrow x_{i+1,j} = R$,
- $x_{i,j} = R \Rightarrow x_{i-1,j} = L$,
- $x_{i,j} = T \Rightarrow x_{i,j-1} = B$,
- $x_{i,j} = B \Rightarrow x_{i,j+1} = T$.

The entropy of this SFT is given by a remarkable integral formula:

$$\frac{1}{4} \int_0^1 \int_0^1 (4 - 2\cos(2\pi s) - 2\cos(2\pi t))\, \mathrm{d}s\, \mathrm{d}t$$

Further work along these lines can be found in [59,117].

One can formulate notions of irreducibility and mixing for higher dimensional shift spaces. It turns out that for SFT's there are several notions of mixing that all coincide in one dimension, but are vastly different in higher dimensions. For instance, a higher dimensional shift space is *strongly irreducible* if there is an integer $R > 0$ such that for any two shapes F, F' of distance at least R, any occuring configurations on F and F' can be combined to form an occuring configuration on $F \cup F'$ [29,129]. For one-dimensional SFT's, this is equivalent to mixing, but much stronger than mixing for two-dimensional SFT's.

Just as in one dimension, we have the notion of sliding block code for higher dimensional shifts. For finite alpha-

bets $\mathcal{A}, \mathcal{B}$, a cube $F \subset \mathbf{Z}^d$ and a function $\Phi\colon \mathcal{A}^F \to \mathcal{B}$, the mapping $\phi = \Phi_\infty\colon \mathcal{A}^{Z^d} \to \mathcal{B}^{Z^d}$ defined by

$$\Phi_\infty(x)_n = \Phi(x_{n+F})$$

is called a sliding block code. By restriction we have the notion of sliding block code from one d-dimensional shift space to another. As expected, for d-dimensional shifts X and Y, the sliding block codes $\phi\colon X \to Y$ coincide exactly with the continuous translation-commuting maps from X to Y, i. e., the maps which are continuous with respect to the metric ρ, defined above, and which satisfy $\phi \circ \sigma^n = \sigma^n \circ \phi$ for all $n \in \mathbf{Z}^d$. Thus it makes sense to consider the various coding problems, in particular the conjugacy, factor and embedding problems, in the higher dimensional setting, but these are very difficult.

Even the question of determining when a higher dimensional SFT of entropy at least $\log n$ factors onto the full n-shift seems very difficult (in contrast to Theorem 15). However, there are some positive results for strongly irreducible SFT's. For instance, it is known that any strongly irreducible SFT of entropy strictly larger than $\log n$ factors onto the full n-shift [37,55]. In fact, that result requires only a weaker assumption than strong irreducibility, but still much stronger than mixing; recent examples [28] show, among other things, that one cannot weaken that assumption to mere mixing.

There are results on other coding problems as well. For instance, a version of the Embedding theorem in one dimension (Theorem 13) holds for SFT's in two dimensions with a strong mixing property [73]; however, it is required that the shift to be embedded contains no points that are periodic in any single direction. And some other results on entropy of proper subshifts of strongly irreducible SFT's carry over from one dimension to higher dimensions [106,109]. But in many cases where versions of the result carry over, the proofs are much different from those in one dimension.

Measures of maximal entropy for two-dimensional SFT's behave very differently from the one dimensional case. For instance, even with very strong mixing properties, such as strong irreducibility, there can be more than one measure of maximal entropy [29], and the relationships among entropy-preserving, finite-to-one, and almost invertibility for factor codes discussed in Sect. "Other Coding Problems" can be very different in higher dimensions [89]. Other differences with respect to entropy can be found in [108].

There is also the natural notion of higher dimensional sofic shifts, which can be defined as those shift spaces that are factors of SFT's. Recall that every one-dimensional sofic shift is a right-resolving, and hence entropy-preserving, factor of an SFT. It is not known if there is an analogue to this fact in higher dimensions, although recently there has been some progress: every sofic shift Y is a factor of an SFT with entropy arbitrarily close to $h(Y)$ [36].

Finally, there is a subclass of d-dimensional SFT's which is somewhat tractable, namely d-dimensional shifts with group structure in the following sense. Let $\mathcal{A}$ be a (finite) group. Then the full d-dimensional shift over $\mathcal{A}$ is also a group with respect to the coordinate-wise group structure. A (higher-dimensional) *group shift* is a subshift of $\mathcal{A}^{\mathbf{Z}^d}$ which is also a subgroup. For a survey on results for this class, we refer the reader to [79].

Future Directions

The future directions of the subject will likely be determined by progress on solutions to open problems. In the course of describing topics in this article, we have mentioned many open problems along the way. For a much more complete list on a wealth of sub-areas of symbolic dynamics, we refer the reader to the article [24]. While it is difficult to single out the most important challenges, certainly the problem of understanding multi-dimensional shift spaces, especially of finite type, is one of the most important.

Bibliography

1. Adler RL, Coppersmith D, Hassner M (1983) Algorithms for sliding block codes – an application, of symbolic dynamics to information theory. Trans IEEE Inf Theory 29:5–22
2. Adler RL, Goodwyn LW, Weiss B (1977) Equivalence of topological Markov shifts. Isr J Math 27:48–63
3. Adler R, Konheim A, McAndrew M (1965) Topological entropy. Trans Amer Math Soc 114:309–319
4. Adler R, Marcus B (1979) Topological entropy and equivalence of dynamical systems. Mem Amer Math Soc 219. AMS, Providence
5. Adler R, Weiss B (1970) Similarity of automorphisms of the torus. Mem Amer Math Soc 98. AMS, Providence
6. Aho AV, Hopcroft JE, Ullman JD (1974) The design and analysis of computer algorithms. Addison-Wesley, Reading
7. Ashley J (1988) A linear bound for sliding block decoder window size. Trans IEEE Inf Theory 34:389–399
8. Ashley J (1996) A linear bound for sliding block decoder window size, II. Trans IEEE Inf Theory 42:1913–1924
9. Ashley J (1991) Resolving factor maps for shifts of finite type with equal entropy. Ergod Theory Dynam Syst 11:219–240
10. Béal M-P (1990) The method of poles: a coding method for constrained channels. Trans IEEE Inf Theory 36:763–772
11. Béal M-P (1993) Codage symbolique. Masson, Paris
12. Béal M-P (2003) Extensions of the method of poles for code construction. Trans IEEE Inf Theory 49:1516–1523
13. Bedford T (1986) Generating special Markov partitions for hyperbolic toral automorphisms using fractals. Ergod Theory Dynam Syst 6:325–333

14. Berger R (1966) The undecidability of the Domino problem. Mem Amer Math Soc. AMS, Providence
15. Berstel J, Perrin D (1985) Theory of codes. Academic Press, New York
16. Blanchard P, Devaney R, Keen L (2004) Complex dynamics and symbolic dynamics. In: Williams S (ed) Symbolic dynamics and its applications. Proc Symp Appl Math. AMS, Providence, pp 37–59
17. Blanchard F, Maass A, Nogueira A (2000) Topics in symbolic dynamics and applications. In: Blanchard F, Maass A, Nogueira A (eds) LMS Lecture Notes, vol 279. Cambridge University Press, Cambridge
18. Bowen R (1970) Markov partitions for Axiom A diffeomorphisms. Amer J Math 92:725–747
19. Bowen R (1973) Symbolic dynamics for hyperbolic flows. Amer J Math 95:429–460
20. Bowen R, Franks J (1977) Homology for zero-dimensional basic sets. Ann Math 106:73–92
21. Bowen R, Lanford OE (1970) Zeta functions of restrictions of the shift transformation. Proc Symp Pure Math AMS 14:43–50
22. Boyle M (1983) Lower entropy factors of sofic systems. Ergod Theory Dynam Syst 3:541–557
23. Boyle M (1993) Symbolic dynamics and matrices. In: Brualdi R et al (eds) Combinatorial and graph theoretic problems in linear algebra. IMA Vol Math Appl 50:1–38
24. Boyle M (2007) Open problems in symbolic dynamics. Contemporary Math (to appear)
25. Boyle M, Handelman D (1991) The spectra of nonnegative matrices via symbolic dynamics. Ann Math 133:249–316
26. Boyle M, Krieger W (1986) Almost Markov and shift equivalent sofic systems. In: Aleixander J (ed) Proceedings of Maryland Special Year in Dynamics 1986–87. Lecture Notes in Math, vol 1342. Springer, Berlin, pp 33–93
27. Boyle M, Marcus BH, Trow P (1987) Resolving maps and the dimension group for shifts of finite type. Mem Amer Math Soc 377. AMS, Providence
28. Boyle M, Pavlov R, Schraudner M (2008) preprint
29. Burton R, Steif J (1994) Nonuniqueness of measures of maximal entropy for subshifts of finite type. Ergod Theory Dynam Syst 14(2):213–235
30. Calkin N, Wilf H (1998) The number of independent sets in a grid graph. SIAM J Discret Math 11:54–60
31. Cidecyian R, Evangelos E, Marcus B, Modha M (2001) Maximum transition run codes for generalized partial response channels. IEEE J Sel Area Commun 19:619–634
32. Coven E, Paul M (1974) Endomorphisms of irreducible shifts of finite type. Math Syst Theory 8:167–175
33. Coven E, Paul M (1975) Sofic systems. Isr J Math 20:165–177
34. Coven E, Paul M (1977) Finite procedures for sofic systems. Monats Math 83:265–278
35. Cover T, Thomas J (1991) Elements of information theory. Wiley, New York
36. Desai A (2006) Subsystem entropy for Z^d sofic shifts. Indag Math 17:353–360
37. Desai A (2008) A class of Z^d-subshifts which factor onto lower entropy full shifts. Proc Amer Math Soc, to appear
38. Devaney R (1987) An introduction to chaotic dynamical systems. Addison-Wesley, Reading
39. Fischer R (1975) Sofic systems and graphs. Monats Math 80:179–186
40. Fischer R (1975) Graphs and symbolic dynamics. Colloq Math Soc János Bólyai: Top Inf Theory 16:229–243
41. Franaszek PA (1968) Sequence-state coding for digital transmission. Bell Syst Tech J 47:143–155
42. Franaszek PA (1982) Construction of bounded delay codes for discrete noiseless channels. J IBM Res Dev 26:506–514
43. Franaszek PA (1989) Coding for constrained channels: a comparison of two approaches. J IBM Res Dev 33:602–607
44. Friedman J (1990) On the road coloring problem. Proc Amer Math Soc 110:1133–1135
45. Gomez R (2003) Positive K-theory for finitary isomoprhisms of Markov chains. Ergod Theory Dynam Syst 23:1485–1504
46. Hadamard J (1898) Les surfaces a courbures opposées et leurs lignes geodesiques. J Math Pure Appl 4:27–73
47. Hassellblatt B, Katok A (1995) Introduction to the modern theory of dynamical systems. Cambridge University Press, Cambridge
48. Hedlund GA (1939) The dynamics of geodesic flows. Bull Amer Math Soc 45:241–260
49. Hedlund GA (1944) Sturmian minimal sets. Amer J Math 66:605–620
50. Hedlund GA (1969) Endomorphisms and automorphisms of the shift dynamical system. Math Syst Theory 3:320–375
51. Hochman M, Meyerovitch T (2007) A characterization of the entropies of multidimensional shifts of finite type. Ann Math, to appear
52. Hochman M (2007) On the dynamics and recursive properties of multidimensional symbolic systems. Preprint
53. Hollmann HDL (1995) On the construction of bounded-delay encodable codes for constrained systems. Trans IEEE Inf Theory 41:1354–1378
54. Immink KAS (2004) Codes for mass data storage, 2nd edn. Shannon Foundation Press, Eindhoven
55. Johnson A, Madden K (2005) Factoring higher-dimensional shifts of finite type onto full shifts. Ergod Theory Dynam Syst 25:811–822
56. Karabed R, Siegel P, Soljanin E (1999) Constrained coding for binary channels with high intersymbol intereference. Trans IEEE Inf Theory 45:1777–1797
57. Kari J (2001) Synchronizing finite automata on Eulerian digraphs. Springer Lect Notes Comput Sci 2136:432–438
58. Kastelyn PW (1961) The statistics of dimers on a lattice. Physica A 27:1209–1225
59. Kenyon R (2008) Lectures on dimers. http://www.math.brown.edu/~rkenyon/papers/dimerlecturenotes.pdf
60. Kim KH, Roush FW (1979) Some results on decidability of shift equivalence. J Comb Inf Syst Sci 4:123–146
61. Kim KH, Roush FW (1988) Decidability of shift equivalence. In: Alexander J (ed) Proceedings of Maryland special year in dynamics 1986–87. Lecture Notes in Math, vol 1342. Springer, Berlin, pp 374–424
62. Kim KH, Roush FW (1990) An algorithm for sofic shift equivalence. Ergod Theory Dynam Syst 10:381–393
63. Kim KH, Roush FW (1999) Williams conjecture is false for irreducible subshifts. Ann Math 149:545–558
64. Kim KH, Ormes N, Roush F (2000) The spectra of nonnegative integer matrices via formal power series. Amer J Math Soc 13:773–806
65. Kitchens B (1998) Symbolic dynamics: One-sided, two-sided and countable state Markov chains. Springer, Berlin
66. Kitchens B, Marcus B, Trow P (1991) Eventual factor maps

and compositions of closing maps. Ergod Theory Dynam Syst 11:85–113
67. Kitchens B, Schmidt K (1988) Periodic points, decidability and Markov subgroups, dynamical systems. In: Alexander JC (ed) Proceedings of the special year. Springer Lect Notes Math 1342:440–454
68. Krieger W (1980) On a dimension for a class of homeomorphism groups. Math Ann 252:87–95
69. Krieger W (1980) On dimension functions and topological Markov chains. Invent Math 56:239–250
70. Krieger W (1982) On the subsystems of topological Markov chains. Ergod Theory Dynam Syst 2:195–202
71. Krieger W (1983) On the finitary isomorphisms of Markov shifts that have finite expected coding time. Wahrscheinlichkeitstheorie Z 65:323–328
72. Krieger W (1984) On sofic systems I. Isr J Math 48:305–330
73. Lightwood S (2003/04) Morphisms form non-periodic Z^2 subshifts I and II. Ergod Theory Dynam Syst 23:587–609, 24: 1227–1260
74. Lind D (1984) The entropies of topological Markov shifts and a related class of algebraic integers. Ergod Theory Dynam Syst 4:283–300
75. Lind D (1989) Perturbations of shifts of finite type. SIAM J Discret Math 2:350–365
76. Lind D (2004) Multi-dimensional symbolic dynamics. In: Williams S (ed) Symbolic dynamics and its applications. Proc Symp Appl Math 60:81–120
77. Lind D (1996) A zeta function for Z^d-actions. In: Pollicott M, Schmidt K (eds) Proceedings of Warwick Symposium on Z^d-actions. LMS Lecture Notes, vol 228. Cambridge University Press, Cambridge, pp 433–450
78. Lind D, Marcus B (1995) An introduction to symbolic dynamics and coding. Cambridge University Press, Cambridge
79. Lind D, Schmidt K (2002) Symbolic and algebraic dynamical systems. In: Hasselblatt B, Katok A (eds) Handbook of Dynamics Systems. Elsevier, Amsterdam, pp 765–812
80. Manning A (1971) Axiom A diffeomorphisms have rational zeta functions. Bull Lond Math Soc 3:215–220
81. Marcus B (1979) Factors and extensions of full shifts. Monats Math 88:239–247
82. Marcus BH, Roth RM (1991) Bounds on the number of states in encoder graphs for input-constrained channels. Trans IEEE Inf Theory 37:742–758
83. Marcus BH, Roth RM, Siegel PH (1998) Constrained systems and coding for recording chapter. In: Brualdi R, Huffman C, Pless V (eds) Handbook on coding theory. Elsevier, New York; updated version at http://www.math.ubc.ca/~marcus/Handbook/
84. Marcus B, Tuncel S (1990) Entropy at a weight-per-symbol and embeddings of Markov chains. Invent Math 102:235–266
85. Marcus B, Tuncel S (1991) The weight-per-symbol polytope and scaffolds of invariants associated with Markov chains. Ergod Theory Dynam Syst 11:129–180
86. Marcus B, Tuncel S (1993) Matrices of polynomials, positivity, and finite equivalence of Markov chains. J Amer Math Soc 6:131–147
87. Markley N, Paul M (1981) Matrix subshifts for $\mathbf{Z}^\nu$ symbolic dynamics. Proc Lond Math Soc 43:251–272
88. Markley N, Paul M (1981) Maximal measures and entropy for $\mathbf{Z}^\nu$ subshifts of finite type. In: Devaney R, Nitecki Z (eds) Classical mechanics and dynamical systems. Dekker Notes 70:135–157
89. Meester R, Steif J (2001) Higher-dimensional subshifts of finite type, factor maps and measures of maximal entropy. Pac Math J 200:497–510
90. Mouat R, Tuncel S (2002) Constructing finitary isomorphisms with finite expected coding time. Isr J Math 132:359–372
91. Morse M (1921) Recurrent geodesics on a surface of negative curvature. Trans Amer Math Soc 22:84–100
92. Morse M, Hedlund GA (1938) Symbolic dynamics. Amer J Math 60:815–866
93. Morse M, Hedlund GA (1940) Symbolic dynamics II, Sturmian trajectories. Amer J Math 62:1–42
94. Mozes S (1989) Tilings, substitutions and the dynamical systems generated by them. J Anal Math 53:139–186
95. Mozes S (1992) A zero entropy, mixing of all orders tiling system. In: Walters P (ed) Symbolic dynamics and its applications. Contemp Math 135:319–326
96. Nagy Z, Zeger K (2000) Capacity bounds for the three-dimensional (0, 1) run length limited channel. Trans IEEE Inf Theory 46:1030–1033
97. Nasu M (1986) Topological conjugacy for sofic systems. Ergod Theory Dynam Syst 6:265–280
98. Ornstein D (1970) Bernoulli shifts with the same entropy are isomorphic. Adv Math 4:337–352
99. Parry W (1964) Intrinsic Markov chains. Trans Amer Math Soc 112:55–66
100. Parry W (1977) A finitary classification of topological Markov chains and sofic systems. Bull Lond Math Soc 9:86–92
101. Parry W (1979) Finitary isomorphisms with finite expected code-lengths. Bull Lond Math Soc 11:170–176
102. Parry W (1991) Notes on coding problems for finite state processes. Bull Lond Math Soc 23:1–33
103. Parry W, Schmidt K (1984) Natural coefficients and invariants for Markov shifts. Invent Math 76:15–32
104. Parry W, Tuncel S (1981) On the classification of Markov chains by finite equivalence. Ergod Theory Dynam Syst 1:303–335
105. Parry W, Tuncel S (1982) Classification problems in ergodic theory. In: LMS Lecture Notes, vol 67. Cambridge University Press, Cambridge
106. Pavlov R (2007) Perturbations of multi-dimensional shifts of finite type. Preprint
107. Petersen K (1989) Ergodic theory. Cambridge University Press, Cambridge
108. Quas A, Sahin A (2003) Entropy gaps and locally maximal entropy in Z^d-subshifts. Ergod Theory Dynam Syst 23: 1227–1245
109. Quas A, Trow P (2000) Subshifts of multidimensional shifts of finite type. Ergod Theory Dynam Syst 20:859–874
110. Radin C (1996) Miles of tiles. In: Pollicott M, Schmidt K (eds) Ergodic Theory of Z^d-actions. LMS Lecture Notes, vol 228. Cambridge University Press, Cambridge, pp 237–258
111. Robinson RM (1971) Undecidability and nonperiodicity for tilings of the plane. Invent Math 12:177–209
112. Robinson EA (2004) Symbolic dynamics and tilings of $\mathbf{R}^d$. In: Williams S (ed) Symbolic dynamics and its applications. Proc Symp Appl Math 60:81–120
113. Rudolph D (1990) Fundamentals of measurable dynamics. Oxford University Press, Oxford

114. Schmidt K (1984) Invariants for finitary isomorphisms with finite expected code lengths. Invent Math 76:33–40
115. Schmidt K (1990) Algebraic ideas in ergodic theory. AMS-CBMS Reg Conf 76
116. Schmidt K (1995) Dynamical systems of algebraic origin. Birkhauser, Basel
117. Schwartz M, Bruck S (2008) Constrained codeds as networks of relations. IEEE Trans Inf Theory 54:2179–2195
118. Seneta E (1980) Non-negative matrices and Markov chains, 2nd edn. Springer, Berlin
119. Shannon C (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423,623–656
120. Sinai YG (1968) Markov partitions and C-diffeomorphisms. Funct Anal Appl 2:64–89
121. Smale S (1967) Differentiable dynamical systems. Bull Amer Math Soc 73:747–817
122. Trachtman A (2007) The road coloring problem. Israel J Math, to appear
123. Tuncel S (1981) Conditional pressure and coding. Isr J Math 39:101–112
124. Tuncel S (1983) A dimension, dimension modules and Markov chains. Proc Lond Math Soc 46:100–116
125. Wagoner J (1992) Classification of subshifts of finite type revisited. In: Walters P (ed) Symbolic dynamics and its applications. Contemp Math 135:423–444
126. Wagoner J (2004) Strong shift equivalence theory. In: Walters P (ed) Symbolic dynamics and its applications. Proc Symp Appl Math 60:121–154
127. Walters P (1982) An introduction to ergodic theory. Springer Grad Text Math 79. Springer, Berlin
128. Walters P (1992) Symbolic dynamics and its applications. In: Walter P (ed) Contemp Math 135. AMS, Providence
129. Ward T (1994) Automorphisms of Z^d-subshifts of finite type. Indag Math 5:495–504
130. Weiss B (1973) Subshifts of finite type and sofic systems. Monats Math 77:462–474
131. Williams RF (1973/74) Classification of subshifts of finite type. Ann Math 98:120–153; Erratum: Ann Math 99:380–381
132. Williams S (2004) Introduction to symbolic dynamics. In: Williams S (ed) Symbolic dynamics and its applications. Proc Symp Appl Math 60:1–12
133. Williams S (2004) Symbolic dynamics and its applications. In: Williams S (ed) Proc Symp Appl Math 60. AMS, Providence

Synchronization Phenomena on Networks

Guanrong Chen[1], Ming Zhao[2], Tao Zhou[2], Bing-Hong Wang[2,3]
[1] Department of Electronic Engineering, City University of Hong Kong, Hong Kong, China
[2] Department of Modern Physics and Nonlinear Science Center, University of Science and Technology of China, Hefei Anhui, China
[3] Institute of Complex Adaptive Systems, Shanghai Academy of System Science, Shanghai, China

Article Outline

Glossary
Definition of the Subject
Introduction
Basic Concepts of Network Synchronization
Synchronizability Versus Structure
Enhancing Network Synchronizability
Future Research Outlook
Acknowledgments
Bibliography

Glossary

Synchronization A problem in time-keeping, requiring the coordination of events to operate a system or a task in unison.

Distance A measure between two nodes, defined as the number of edges connecting them through the shortest paths.

Average distance The mean distance, averaged over all pairs of nodes on the network.

Clustering coefficient The probability that two randomly-selected neighboring nodes of a node are directly connected each other.

Node-degree The number of edges incident from a node.

Random-graph network A type of graph obtained by starting with a set of nodes and then adding edges between them at random.

Small-world network A type of graph in which most nodes are not neighbors of each other, but most nodes can be reached from any other node by a small number of connection steps; thus, a small-world network is highly clustered like a regular graph, and yet with a small average distance, just like a random graph.

Scale-free network A type of graph in which a small number of nodes have a large number of connections while a large number of nodes have a small number of connections, whose node-degree distribution typically follows a power-law form, with both structure and dynamics being independent of the network size.

Node-betweenness A measure of the extent to which a given node is occupied by the amount of information passing through it via shortest paths between other nodes, namely, the portion of shortest paths between all pairs of nodes which have data traffic going through this particular node in the network.

Definition of the Subject

The subject under consideration is synchronization on complex networks, with respect to the phenomena and

particularly the ability of achieving synchrony of a network of dynamical systems. The subject of synchronization is quite old, but it is a significant one continuously calling for serious and systematic investigation. Ever since the careful study of two synchronous pendulum clocks by the great Dutch scientist Christian Huygens in 1665, the subject has evolved to be an independent and indispensable field of scientific research. The current study of complex networks, on the other hand, is pervading all kinds of sciences, ranging from physical to biological, even to social sciences. Its impact on modern engineering and technology is prominent and will be far-reaching. Typical complex dynamical networks include the Internet, the World Wide Web, various wireless communication networks, metabolic networks, biological neural networks, social relationship networks, financial and economic networks, and so on. As it has turned out today, the study of synchronization phenomena and synchronous behaviors of dynamical systems such as oscillators on complex networks has become overwhelming. This article offers an overview of the state-of-the-art advances and developments of the subject of synchronization on various complex networks, with emphasis on network synchronizability and performance.

Introduction

Many biological, social and technological systems can be properly described by complex networks with nodes representing individuals or organizations and edges characterizing the interactions among them [1,2,3,4,5]. One of the goals in the current studies on complex networks is to understand and explain how the topological properties of a network affect the behaviors of dynamical systems built upon the network. Typical examples include understanding how the topology of the Internet affects the spread of the computer viruses [6,7,8,9,10], how the structure of a power grid affects the cascading failures over time [11,12,13,14,15], how the connecting patterns of an intercommunication network affect its data traffic and dynamics [16,17,18,19,20], and so on.

Synchronous behaviors have been observed in various complex networks in nature and human society [23,24,25,26], and they have been studied for hundreds of years since the systematic investigation of pendulum synchrony by the great Dutch scientist Christian Huygens in 1665 [27].

To understand how network structure affects the synchronizability of a network not only has broad theoretical interest [28], but also has important practical value [29]. One typical case in point is the synchronicity of sensors in biological neural networks, where neurons communicate with each other through synaptic junctions for which a mechanism called asynchronous release is important [30]. There are many careful studies about collective synchronization in the earlier literature, with a basic assumption that dynamical systems of coupled oscillators evolve either on regular networks [31,33,34] or on random networks [35,36]. However, the structures of most real-world networks are neither completely regular nor completely random, but rather, somewhere in between. Thus, it becomes important and even necessary to consider how network structure affects the synchronization process and the synchronizability of the dynamical systems on such networks. Recently, it has been found that networks with small-world effects and scale-free properties are quite different from, and oftentimes achieve synchronization more easily than, regular networks such as lattices [37,38,39,40,41,42,43,44].

The study of synchronization on complex networks has gone through several stages in the past decade, encompassing several important aspects of the subject: various synchronization phenomena on complex networks and their stability analysis, the relationships between structural ingredients and a network's synchronizability, the enhancement or reduction of network synchronizability, etc. The first two are quite well understood today while the last one will be further addressed in this article. First, some basic concepts about synchronization of networked dynamical systems and the associated stability analysis are introduced. Second, some intrinsic relations between network structure and synchronizability are discussed. Third, three types of methods, namely, regulating coupling patterns, modifying network structures, and designing output functions, are introduced for enhancing network synchronizability. Finally, some open questions are posed which are deemed significant for further studies of the important subject of complex network synchronization.

To proceed, some notations are introduced [5], among which three are most significant with respect to network synchronization: average distance L, clustering coefficient C, node-degree k_i of node i, and the corresponding probability density function of degree distribution $p(k)$.

In the past few years, by taking advantage of both high-speed computing power and the huge amount of real data available on the web, scientists were able to search and find some common statistical characteristics shared by many real-world networks. It is found that most real networks have a very small average distance, scaled approximately as $L \sim \ln N$, where N is the size of the network (i. e., the total number of its nodes), while their clustering coefficient is rather large, as compared with *random-graph networks*

(Erdös and Rényi [28]). A network having both of these two characteristics is referred to as a *small-world network*, described by Watts and Strogatz [21]. Moreover, the degree distributions of many real networks obey a power-law form $p(k) \sim k^{-\gamma}$, where $p(k)$ is the probability density function for the corresponding degree distribution, and γ is the power-law exponent (typically $2 < \gamma < 3$) [1,2,3]. The power-law distribution falls off much more gradually than an exponential one, allowing for a few nodes with very large degrees to exist. Networks with power-law degree distributions usually belong to the class of *scale-free networks*, characterized by Barabási and Albert [22].

Basic Concepts of Network Synchronization

A general model of coupled identical oscillators on a network can be described by [40,41]

$$\dot{\boldsymbol{x}}_i = \boldsymbol{F}(\boldsymbol{x}_i) - \sigma \sum_{j=1}^{N} G_{ij} \boldsymbol{H}(\boldsymbol{x}_j) , \quad i = 1, \ldots, N , \qquad (1)$$

where $\dot{\boldsymbol{x}}_i = \boldsymbol{F}(\boldsymbol{x}_i)$ governs the dynamics of the ith oscillator, with state vector $\boldsymbol{x}_i$; $\boldsymbol{H}(\boldsymbol{x}_j)$ is the output function; σ is the coupling strength; $G = [G_{ij}]$ is an $N \times N$ coupling matrix determined by the given coupling pattern among the N oscillators.

In the typical situation when the oscillators are symmetrically coupled, the coupling matrix G has the same form as the graph Laplacian L, i. e., $G = L$, with

$$L_{ij} = \begin{cases} k_i & \text{for } i = j \\ -1 & \text{for } j \in \Lambda_i \\ 0 & \text{otherwise} , \end{cases} \qquad (2)$$

where k_i is the degree of node i and Λ_i is the set of its neighboring nodes. In this setting, L is symmetrical and semi-positive definite, and all the rows of L have a zero sum, so that its smallest eigenvalue λ_1 is always a single zero and all the other eigenvalues are strictly positive. Thus, the eigenvalues of L can be ranked as

$$0 = \lambda_1 < \lambda_2 \leq \lambda_3 \leq \cdots \leq \lambda_N .$$

For network (1), the synchronization manifold is an invariant manifold: $\boldsymbol{x}_1 = \boldsymbol{x}_2 = \cdots = \boldsymbol{x}_N = \boldsymbol{s}$, typically satisfies $\dot{\boldsymbol{s}} = \boldsymbol{F}(\boldsymbol{s})$ in engineering applications.

For a dynamical system, the so-called master stability function is usually defined to be the ratio of the largest Lyapunov exponent versus a connectivity parameter of the system [42,45]. For some dynamical systems, the master stability function is negative when $\lambda_2 > \alpha_1/\sigma$ for some constant α_1. In this case, the largest Lyapunov exponent is negative, and consequently the network is synchronizable; moreover, the larger the λ_2 is, the better the network synchronizability will be [40,41].

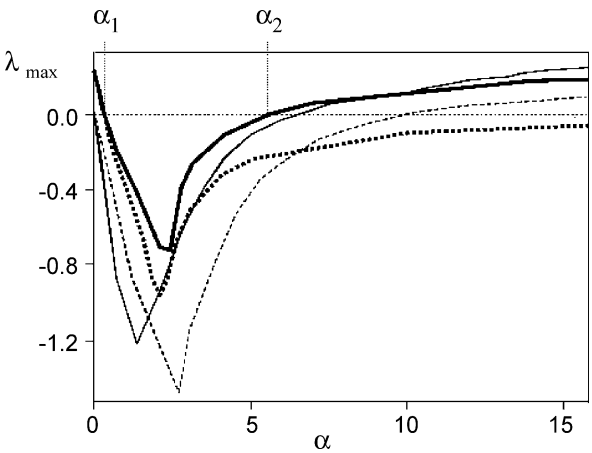

Synchronization Phenomena on Networks, Figure 1
Four typical master stability functions for coupled Rössler oscillators: chaotic (*bold curve*) and periodic (*regular curve*); with *y*-coupling (*dashed curve*) and *x*-coupling (*dotted curve*). The *vertical ordinate* shows the change of the largest Lyapunov exponent. Curves are all scaled for clearer visualization (after [42])

For some other dynamical systems, the master stability function is negative only within a finite interval (α_1, α_2) [46], over which the largest Lyapunov exponent is negative [42,45], where α_1 and α_2 are constants. In this case, the network is synchronizable for some σ when the eigenratio $R = \lambda_N/\lambda_2$ satisfies $R < \alpha_2/\alpha_1$; moreover, a smaller R indicates a better network synchronizability.

The former case corresponds to networks for which the synchronized region is unbounded (the bold-dashed curve in Fig. 1), and the latter, bounded (the bold-solid curve and the two regular lines in Fig. 1) [42]. In both cases, the right-hand side of the above two inequalities depends only on the dynamics of each individual oscillator and the output function of the network, while the eigenvalue λ_2 and eigenratio R depend only on the Laplacian L. Therefore, the problem of synchronization can be divided into two parts: choosing suitable dynamics (including the aforementioned parameters and output function) and analyzing the eigenvalues of the Laplacian. In fact, these two cases can co-exist [32,101].

The same stability analysis can also be applied to some more complicated coupling patterns [40,41,42,45,47], including the case where G is non-diagonalizable (see Fig. 2 and [48]).

Network (1) has only identical oscillators, while in the real world parameter mismatch between oscillators is very

common, so that both the amplitudes and phases of different oscillators become different. However, quite often, only the frequencies of oscillations are of concern in some applications, while the amplitudes are not important. In such cases, phase synchronization is the topic for study, for which the *Kuramoto model* [49,50,51,52,53] is a representative platform.

In the Kuramoto model, oscillators run at arbitrary frequencies and they are coupled through a periodic (e. g., sine) function of their phase differences. More precisely, the model consists of a population of N coupled phase-oscillators $\theta_i(t)$ having natural frequencies ω_i distributed with a given probability density $g(\omega)$, governed by

$$\dot{\theta}_i = \omega_i - \sigma \sum_{j=1}^{N} G_{ij} \sin(\theta_i - \theta_j) , \quad i = 1, \ldots, N . \quad (3)$$

To measure the synchronization phenomena, an order parameter M is introduced:

$$M \equiv \left[\left\langle \left| N^{-1} \sum_{j=1}^{N} e^{i\phi_j} \right| \right\rangle \right] , \quad (4)$$

where ϕ is a function of θ, and $\langle \cdot \rangle$ and $[\cdot]$ denote the average over time and over different configurations, respectively.

Initially, each node is assigned a random phase. Without coupling, all the oscillators run independently and, at any time, the phases of the oscillators are distributed almost uniformly on the interval $[0, 2\pi]$, yielding $M = O(1/\sqrt{N})$. In this situation, the oscillators are generally not synchronized. With coupling, as the coupling strength gradually increases to beyond a certain threshold, interactions among oscillators become stronger and more inter-influential, which gradually dominate the individual self-oscillations. Eventually, collective synchronization of all oscillators emerges spontaneously. During this transition process, the order parameter M increases from 0 to 1.

Synchronizability Versus Structure

Previous studies have demonstrated that both scale-free and small-world networks are much easier to synchronize than regular lattices [37,38,39,40,41,42,43,44]. At this point, a natural question arises: what makes them easier to synchronize? An intuitive answer might be their average distance, which is much shorter than that of a regular network with the same size. However, after some systematic investigations on the relation between structural ingredients and the network synchronizability, Nishikawa et al. [54] found that as the network becomes more heterogeneous, i. e., the degree distribution becomes wider,

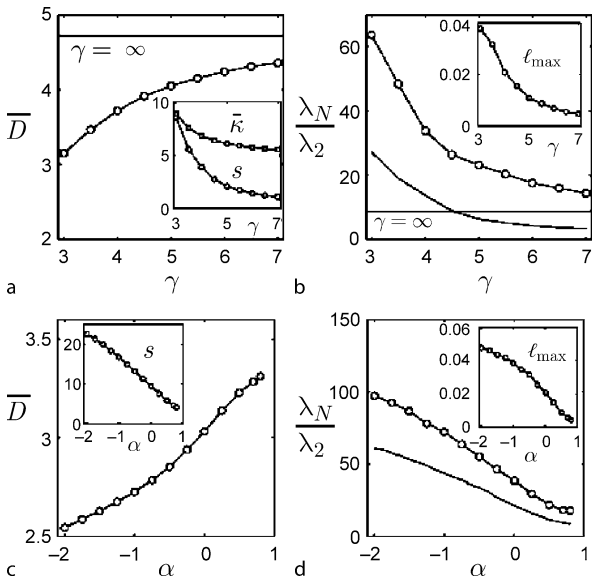

Synchronization Phenomena on Networks, Figure 2
Synchronization of scale-free networks. a, b the semi-random model; c, d the growing model with aging of nodes. The small *insets* are the responses of the indicated parameters with respect to the changing parameters γ or α under the same conditions (after [54])

a network can become less synchronizable even though its average distance becomes much shorter. Figure 2 gives two examples of this phenomenon. In a semi-random model [55], with the increase of the power-law exponent γ, which makes the network more homogeneous, the network average distance $\bar{D}$ becomes longer and the standard deviation of the degree distribution reduces (Fig. 2a and inset); meanwhile, the eigenratio λ_N/λ_2 of its Laplacian becomes smaller (Fig. 2b), indicating improvement of the network synchronizability. In a growing model of scale-free networks with aging nodes [56], it is also observed that as the average distance increases and the degree distribution becomes more homogeneous, the network gains a better synchronizability (Fig. 2c,d).

A heuristic exploration may be given: in a network with a heterogeneous degree distribution, a few "central" oscillators, which interact with a large number of other oscillators, tend to be overloaded by the traffic passing through them. When too many independent traffic signals with different phases and frequencies are traversing through a node at the same time, they cause congestion, leading to the reduction of network synchronizability. The same also happens to overloaded edges [54].

On the other hand, based on experience with WS small-world networks, Hong et al. [57] concluded that the

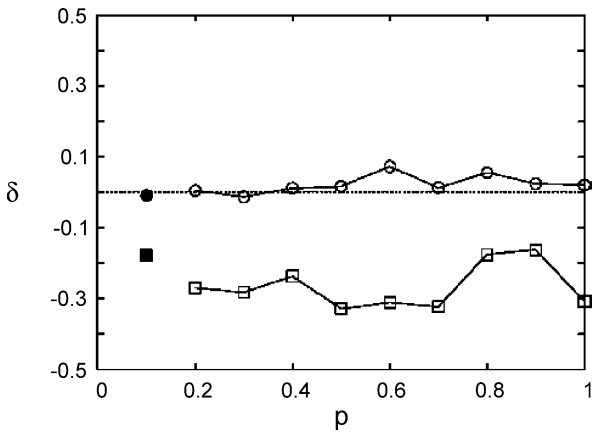

Synchronization Phenomena on Networks, Figure 3
Behavior of the difference δ of the eigenratio in a WS network with rewiring probability p (after [57])

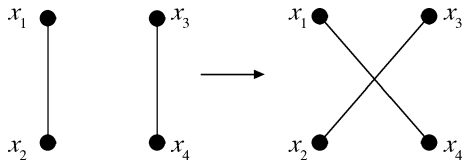

Synchronization Phenomena on Networks, Figure 4
Sketch of maps of the random interchanging algorithm (after [62])

maximal node betweenness [58,59,60] is a good indicator for network synchronizability: the smaller, the better, and vice versa. To confirm their observation, they calculated the difference of the eigenratio before and after the removal of a node from a WS network [21]. Figure 3 plots the difference $\delta \equiv (\lambda_N/\lambda_2)_{\text{after}} - (\lambda_N/\lambda_2)_{\text{befor}}$. The reduction of the ratio is brought about by the removal of the node with the maximal betweenness (empty squares in the figure). In comparison, random removal of a node makes the eigenratio almost unchanged (empty circles in the figure). This implies that the node with the maximal betweenness plays an important role in determining the synchronizability of the network. However, for scale-free networks, this "maximal betweenness indicator" may not work, as pointed out in [61] with a counterexample given in [62].

In the above studies, a network is usually modified in order to see how the synchronizability changes as the network structure varies. It is worth emphasizing that during the modification process all the topological ingredients [5] have been changed at the same time, therefore it is impossible to obtain any accurate relation between one particular ingredient and the network synchronizability. Knowing this problem, by using the edge-exchange operation [63,64], Zhao et al. [62] derived some fairly accurate relations between the synchronizability and the average distance as well as the heterogeneity of the degree distribution, on small-world and scale-free network models. Figure 4 presents a sketch of maps of their random interchanging algorithms. The algorithmic operations will change only the network average distance while keeping the degree of each node unchanged. Thus, the relations between the two concerned ingredients can be investigated separately. Extensive simulations have verified that either shortening the average distance or lowering the heterogeneity may lead to a better synchronizability, but only their combination can always ensure that the network will synchronize easily.

McGraw and Menzinger [65] investigated the relations between the clustering coefficient and network synchronizability, and concluded that for both random-graph and scale-free networks, increasing the clustering coefficient hinders global synchronization if the coupling strength is strong, but it promotes the synchronization of scale-free networks when the coupling strength is weak. Figure 5 shows this phenomenon. The main reason is that the clusters around the hub-nodes promote the formation of frequency-synchronized clusters, but they will inhibit the synchronization of the network as a whole. The early hub synchronization accounts for the slightly enhanced order parameter when the coupling is weak [65,66]. This analysis is based on non-identical oscillators in the Kuramoto model. On the other hand, by means of master stability analysis, Wu et al. [67] reported a negative correlation between the clustering coefficient and synchronizability through a scale-free network model with a tunable clustering coefficient [68].

Besides the main focus on small-world effects and scale-free properties, as described by the clustering coefficient, average distance and degree distribution, some further studies on the effects of other topological ingredients on network synchronization have also been reported, particularly the degree-degree correlation. A network is said to show *assortative* (or *disassortative*) *mixing*, if the nodes having many connections tend to connect to other nodes with many (or few) connections. The extent of this degree-degree correlation can be measured by the Pearson coefficient [69]: its positive (or negative) value indicates assortative (or disassortative) mixing. Di Bernardo et al. found that disassortative networks generally have a better synchronizability than the assortative ones [70,71]. However, later works [72,73] show that the degree distribution, coupling pattern, and degree-degree correlation among the nodes compete with each other in an intrinsic manner,

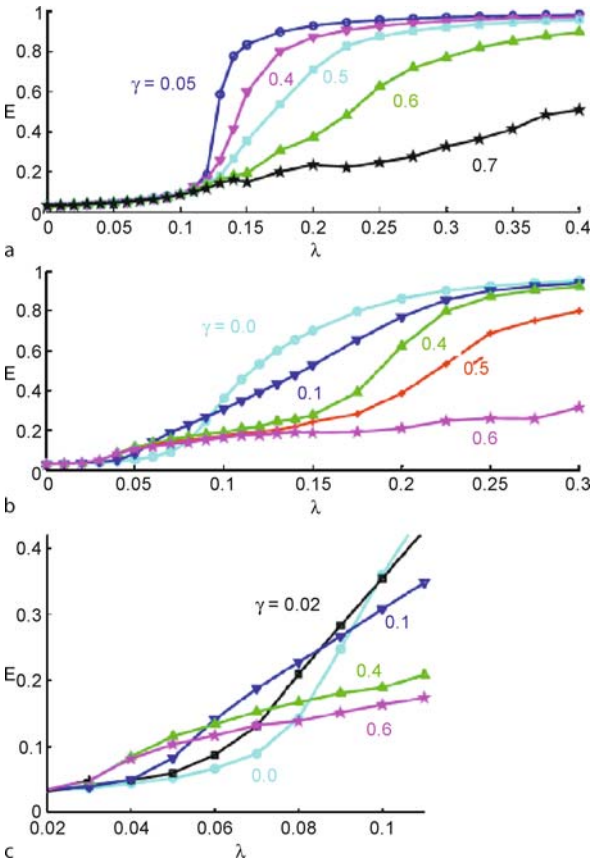

Synchronization Phenomena on Networks, Figure 5
Order parameter *M* vs coupling strength λ, for different values of the clustering coefficient γ. a Poisson degree distribution. b, c Power-law degree distribution. c A close-up of the transition region, showing that increase of the clustering coefficient leads to an advanced (lower-λ) transition (after [65])

thereby together determining the network synchronizability. That is, for one coupling pattern, disassortative mixing may predict better synchronizability, while for another coupling pattern, the result can be the opposite.

As we gain more knowledge of various network structures, more attention is paid to the effects of local structures of complex networks on their global behaviors and dynamics. Huang et al. [74] found that in complex networks with prominent clusters, the synchronizability is determined by the interplay between intercluster and intracluster edges: a network is mostly synchronizable when the numbers of the two types of edges are approximately equal. If not equal, for example as the number of intracluster edges increases, an abnormal synchronization phenomenon appears: although the network average distance becomes smaller, the network synchrony is weakened or even destroyed.

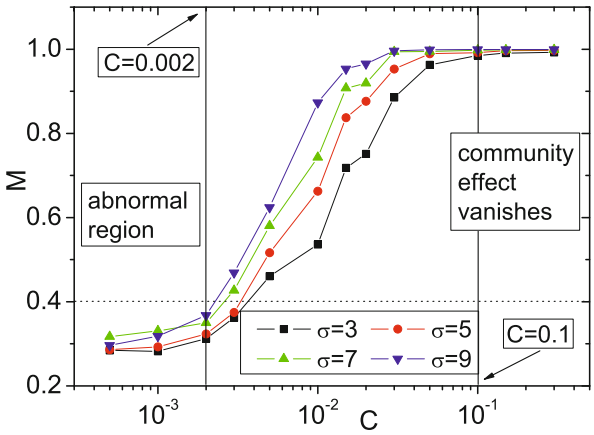

Synchronization Phenomena on Networks, Figure 6
Order parameter *M* vs. community strength *C* for different values of the coupling strength σ (after [76])

Furthermore, the synchronization phenomenon of a complex network with a community structure has also been discussed. Qualitatively, a *community* is defined as a subset of nodes within a network with the property that the connections among the nodes therein are denser than those within the other parts of the network [75]. Zhou et al. studied phase synchronization in a network with a community structure [76]. Defining the edges connecting two nodes in one community as *internal edges*, and those connecting nodes between two communities as *external edges*, the ratio of the number of external edges to the number of internal edges can be used to characterize the strength of the community structure, denoted by *C*. Clearly, a smaller *C* corresponds to sparser external edges thus a more prominent community structure. Figure 6 shows the relationship between the order parameter *M* and the community strength *C* for different coupling strengths σ. It is found from Fig. 6 that a strong community structure will hinder global synchronization no matter what the coupling strength is, but this effect will vanish when the fraction of external connections exceeds 0.1.

Using a modified simulated annealing algorithm, Donetti et al. [77] generated an entangled network with optimal synchronizability. These kinds of networks are shown to have an extremely homogeneous structure: distributions of node degrees, distances, betweenness, and loops are all very uniform. Also, these networks are characterized by short average distances and large loops, with no well-defined community structures. In the approach of [77], rewiring is applied, i. e., at each time step, the number of rewiring trials is randomly extracted from an exponential distribution. Except for rewiring which reduces or increases the eigenratio, and except for operations that

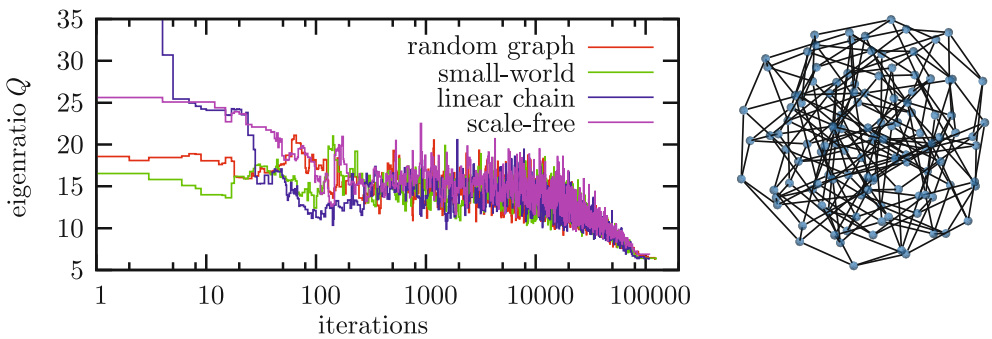

Synchronization Phenomena on Networks, Figure 7
Eigenratio *Q* as a function of the number of algorithmic iterations. Starting from different initial configurations, all the networks are converted via iterations to *entangled networks* (after [77])

disconnect the network, for different initial configurations the optimization process will always lead to the same optimal result. Figure 7 shows the changes of the eigenratio in the optimizing process and the resultant network configuration.

Enhancing Network Synchronizability

With a clearer understanding of the relations between the network structure and synchronizability, a natural question about how to enhance the network synchronizability is in order. Some effective synchronizability-enhancement methods are introduced in this section.

Coupling Pattern Regulation

In general, scale-free networks are much harder to synchronize than random networks with the same size and the same average degree. One reason is that in scale-free networks, there are some "central" oscillators that interact with a large number of other nodes [54]. Thus, when too many independent signals with different phases and frequencies are traversing through a "central" oscillator at the same time they may have conflicts, thereby causing traffic congestion. Hence, generally speaking, the more heterogeneous the degree distribution, the more difficult for the network to synchronize. It is also known that in scale-free networks, when the oscillators are coupled symmetrically, oscillators with larger degrees usually approach the final synchronized state first, and then the others with smaller degrees synchronize to them gradually [65]. Therefore, when the oscillators are coupled asymmetrically, if the coupling strength from the "central" oscillators to the other nodes are stronger than the reverse, the network will synchronize much easier and faster.

Based on this idea, Motter, Zhou and Kurths [78,79,80] proposed a new coupling pattern, which we will call the MZK pattern, which can sharply improve network synchronizability. After that, quite a few methods for regulating coupling patterns are brought forward to improve network synchronizability, some static and some dynamic.

Static Coupling Patterns

In static coupling patterns, the elements of the coupling matrix are formulated based on the MZK pattern, as

$$G_{ij} = L_{ij}/k_i^{\beta} \,, \tag{5}$$

where β is a tunable parameter. The coupling is weighted when $\beta \neq 0$, and unweighted when $\beta = 0$. In spite of the asymmetry of this coupling matrix G, it can be proved that all the eigenvalues of G are nonnegative reals with only one eigenvalue being zero if the network is connected. Rewrite Eq. (5) as

$$G = D^{-\beta} L \,, \tag{6}$$

where $D = \mathrm{diag}(k_1, \ldots, k_N)$ is a diagonal matrix and L is the Laplacian. From the identity

$$\det(D^{-\beta} L - \lambda I) = \det(D^{-\beta/2} L D^{-\beta/2} - \lambda I) \,, \tag{7}$$

where I is the $N \times N$ identity matrix, one can prove that the spectrum of G is the same as that of the following symmetric matrix:

$$H = D^{-\beta/2} L D^{-\beta/2} \,. \tag{8}$$

Similarly to the case of matrix G, if the network is connected then all eigenvalues of H other than the single $\lambda_1 = 0$, are positive. With $\beta = 1$, the matrix H is a normalized Laplacian. Thus, if the network is connected and $N \geq 2$, then

$$0 < \lambda_2 \leq N/(N-1) \,, \quad 2 \leq \lambda_N \leq N/(N-1) \,. \tag{9}$$

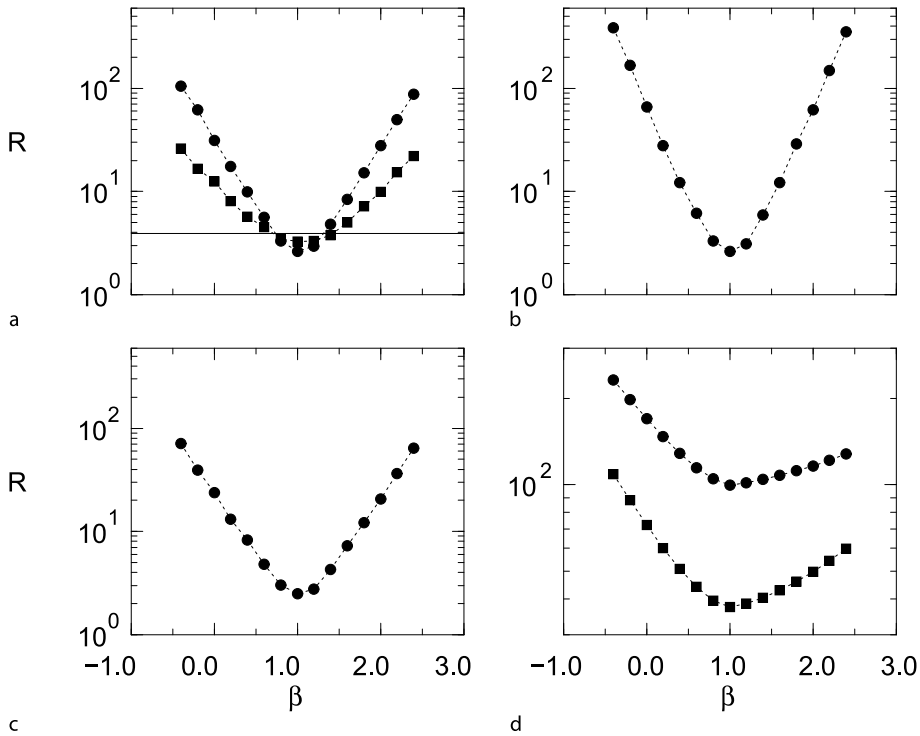

Synchronization Phenomena on Networks, Figure 8

Eigenratio R as a function of β for four kinds of complex networks specified in [78]. For each model, the synchronizability peaks at $\beta = 1.0$ (after [78])

Figure 8 shows the changes of the eigenratio R with the parameter β in four kinds of complex networks specified in [78]. It can be seen that the eigenratio R has a well-defined minimum at $\beta = 1$ in all cases. Mathematically, this means that the best results are obtained when the matrix D has a square-root. It is also clear that the more heterogeneous the network is, the more prominent the minimum of the eigenratio R becomes.

By explicitly relating the asymmetry in the connections to an age order among different nodes, Hwang et al. [81] found that age-ordered networks provide a better propensity for synchronization. The main reason is that an older node becomes weaker, therefore more easily influenced by other nodes. In this coupling pattern, the off-diagonal entries of the zero-row-sum coupling matrix G are

$$G_{ij} = -a_{ij}\frac{\Theta_{ij}}{\sum_{j\in\Lambda_i}\Theta_{ij}}\,, \tag{10}$$

where a_{ij} are the elements of the adjacency matrix A ($a_{ij} = 1$ if nodes i and j are connected, and $a_{ij} = 0$ otherwise), and $\Theta_{ij} = (1-\theta)/2$ (or $\Theta_{ij} = (1+\theta)/2$) for $i > j$ (or $i < j$). The parameter $\theta \in (-1, 1)$ governs the coupling asymmetry in the network: the limit $\theta \to -1$ (or $\theta \to 1$) gives a unidirectional coupling, where the old (or young) nodes drive the young (or old) ones. When $\theta = 0$, the coupling pattern degenerates to the MZK pattern at $\beta = 1$.

For a generic θ, the spectrum of the coupling matrix G is in the complex plane and the complex eigenvalues appear in pairs of complex conjugates ($\lambda_1 = 0$; $\lambda_\ell = \lambda_\ell^r + j\lambda_\ell^i$, $\ell = 2, \ldots, N$). It can be proved that (i) $0 < \lambda_2^r \le \cdots \le \lambda_N^r \le 2$, and (ii) $|\lambda_\ell^i| \le 1, \forall \ell$. The best propensity for synchronization is then ensured when both the ratio $\lambda_N^r/\lambda_\ell^r$ and $M \equiv \max_\ell\{|\lambda_\ell^i|\}$ are simultaneously made as small as possible.

In scale-free network models, the age of a node can be denoted by the time when it is being added to the network. The class of scale-free networks under study is generated from the Barabasí–Albert model [82,83]. For comparison, a highly homogeneous random network with an arbitrary initial age ordering is considered, with the average degree being equal to that of the scale-free network. Figure 9 shows the variation of the synchronizability of the two networks versus the parameter θ. For the random network, symmetric coupling makes the ratio λ_N^r/λ_2^r smallest, while for the scale-free model, the propensity for synchronization is better (or worse) when $\theta \to -1$ (or $\theta \to 1$). As for M, there are only very small differences between

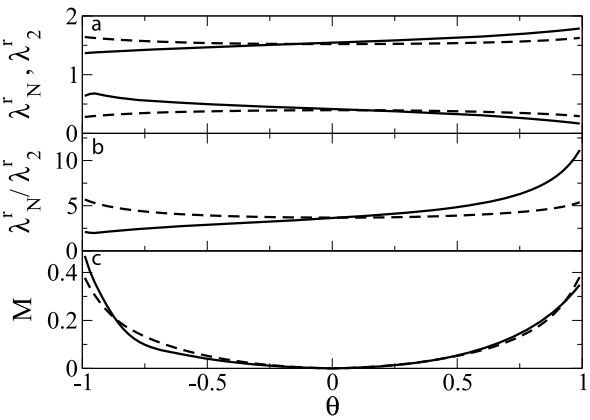

Synchronization Phenomena on Networks, Figure 9
a λ_N^r and λ_2^r, b λ_N^r/λ_2^r, and c M, vs. θ, for a scale-free network ($m = 5$ and $B = 0$, *solid curves*) and random network (*dashed curves*) (after [81])

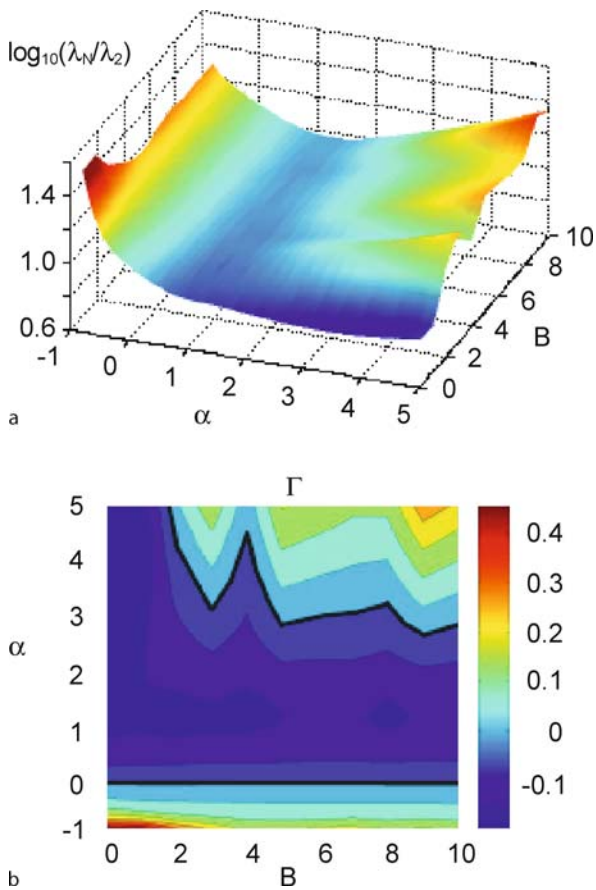

Synchronization Phenomena on Networks, Figure 10
a λ_N/λ_2 for scale-free networks vs. the parameter space (α, B). b $\Gamma = \log(\lambda_N/\lambda_2) - [\log(\lambda_N/\lambda_2)]_{\alpha=0}$ vs. (α, B). The domain with $\Gamma < 0$ is outlined by the *black contours* drawn (after [85])

the scale-free and the random-network models. Thus, it is concluded that in scale-free networks, the network synchronizability is enhanced when the dominant coupling direction is from older to younger nodes [84].

Taking the edge-weights into account, Chavez et al. [85,86] investigated the propensity for synchronization of some weighted complex networks, where the weight in an arbitrary edge, ℓ_{ij}, is defined as its traffic load [87], which quantifies the traffic of shortest paths which make use of that edge. In this coupling pattern, the off-diagonal entries of the zero-row-sum coupling matrix G are

$$G_{ij} = -\frac{\ell_{ij}^{\alpha}}{\sum_{j\in\Lambda_i}} \ell_{ij}^{\alpha}\,, \tag{11}$$

where α is a tunable parameter, and ℓ_{ij} is the load of the edge connecting nodes i and j.

Although G is asymmetric for all α, just like the MZK pattern, it can be proved that all its eigenvalues are non-negative reals with only one zero eigenvalue if the network is connected. The case of $\alpha = 0$ corresponds to the best synchronizability condition for the MZK pattern. From Eq. (11), it can be seen that in the limit of $\alpha = +\infty$ (or $\alpha = -\infty$) only the edges with the largest (or smallest) loads ℓ_{ij} are selected as the incoming edges for each node i. Therefore, this generates a network with at least N directed edges, which can be either connected or disconnected. In the connected (or disconnected) case, the ratio λ_N/λ_2 will be equal to 2 (or $+\infty$), thus yielding a very strong (or weak) condition for synchronization.

Figure 10a shows the logarithm of λ_N/λ_2 in the parameter space (α, B) for the above-discussed model [82,83]. Parameter B is used to regulate the heterogeneity of the degree distribution. It can be observed that the surface of λ_N/λ_2 has a prominent minimum when $\alpha \simeq 1$ for all values of B above a given threshold $B_c > 0$, which means that the weighting procedure based on edge loads always enhances the network propensity for synchronization. The quantity $\Gamma = \log(\lambda_N/\lambda_2) - [\log(\lambda_N/\lambda_2)]_{\alpha=0}$ shown in Fig. 10b may be used to quantify the synchronizability enhancement.

The coupling patterns proposed by both Hwang et al. [81] and Chavez et al. [85,86] can enhance the propensity for network synchronization. The former works well only for age-ordered networks, while the latter requires the knowledge of the load on each edge of the whole network. Therefore, a general coupling pattern using only local information would be very desirable. Based on the idea that different nodes should play different roles in a network, Zhao et al. [73] proposed a coupling pattern which

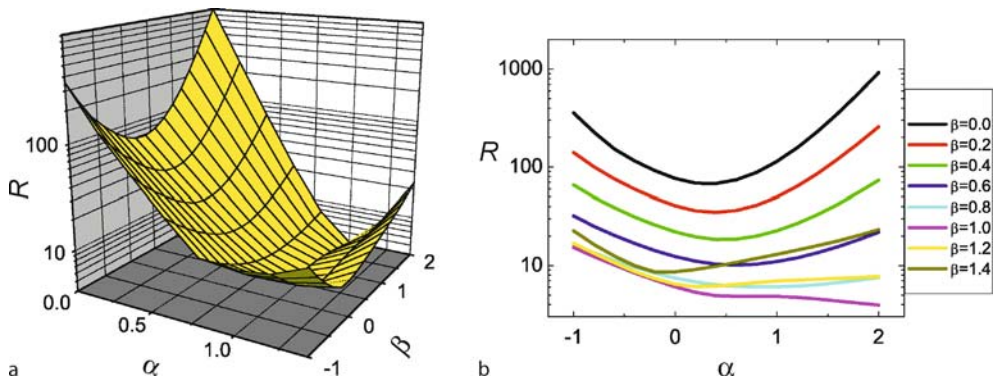

Synchronization Phenomena on Networks, Figure 11
a Eigenratio R in the parameter plane (α, β). b R vs. α for different values of parameter β (after [73])

requires only the degrees of neighboring nodes. The coupling matrix G of this pattern is given by

$$G_{ij} = \begin{cases} -k_j^{\alpha}/S_i^{\beta} & \text{for } j \in \Lambda_i \\ S_i/S_i^{\beta} & \text{for } i = j \\ 0 & \text{otherwise}\,, \end{cases} \tag{12}$$

where $S_i = \sum_{j\in\Lambda_i} k_j^{\alpha}$. When $\alpha = \beta = 0$, this coupling pattern degenerates to the symmetric coupling pattern [45], where the case of $\alpha = 0$ corresponds to the MZK pattern [78] and the case of $\beta = 1$ is equivalent to the one introduced in [80] (see Eq. (15) in [80] for more details). Although this G is asymmetric for all α with $\beta \neq 0$, it can also be proved that all its eigenvalues are non-negative reals with only one zero eigenvalue, if the network is connected. Figure 11 shows some simulation results. From the figure, it can be concluded that there is always some parameter region in which the eigenratio R is smaller than that of the symmetrically coupled case ($\alpha = \beta = 0$) and that of the optimal case with the MZK pattern ($\alpha = 0$ and $\beta = 1$).

From the viewpoint of gradient fields, Wang et al. [88] also derived a coupling pattern that has the same configuration as Eq. (12) with $\beta = 1$.

Dynamic Coupling Patterns

The coupling patterns discussed above are all based on a network having a fixed structure which remains unchanged throughout the synchronizing process.

Zhou et al. [89] investigated synchronization in a scale-free network of chaotic oscillators, where the coupling strength of a node develops adaptively according to the local synchronizing property between the node and its neighbors. In this coupling pattern, the off-diagonal entries of the zero-row-sum coupling matrix G are

$$G_{ij} = -a_{ij}W_{ij}\,, \tag{13}$$

where $W_{ij} > 0$ is the coupling strength from node j to node i if they are connected. Here, suppose that the strength between node i and all its k_i neighbors increases uniformly among the k_i connections, in order to suppress its difference Δ_i from the mean activity of its neighbors; namely,

$$G_{ij}(t) = -a_{ij}V_i(t)\,, \quad \dot{V}_i = \gamma\Delta_i/(1+\Delta_i)\,, \tag{14}$$

where $\Delta_i = |\boldsymbol{H}(\boldsymbol{x}_i) - (1/k_i)\sum_j a_{ij}\boldsymbol{H}(\boldsymbol{x}_j)|$, and $\gamma > 0$ is the adaptation parameter. It is clear that, in this adaptive coupling scheme, the input weight ($W_{ij} = V_i$) and the output weight ($W_{ji} = V_j$) of node i are generally asymmetrical.

Next, synchronization of a network of coupled Rössler oscillators and a chaotic foodweb model on Barabasí–Albert scale-free networks are considered, and two cases of unbounded and bounded stability zones are investigated, respectively. When the stability zone is unbounded, the transition to synchronization is shown in Fig. 12a. Starting from random initial conditions on the chaotic attractors, the local synchronization difference $\Delta \gg 1$, and the input weights of each node, both increase uniformly on the whole network, i. e., $W_{ij} = V_i(t) \approx \gamma t$ (Fig. 12a, inset). After a short period of time, the weights V_i of different nodes develop at different rates and then converge to different values $\tilde{V}_i$. The input weight is smaller on average for nodes with larger degrees k_i (Fig. 12b). Here, the synchronization error is measured by averaging all local errors over the nodes: $E(t) = \langle|\boldsymbol{x}_i - \langle\boldsymbol{x}_i\rangle|\rangle$.

The dependence of the input weight of a node on its degree follows a power law,

$$V(k) \sim k^{-\theta}\,, \tag{15}$$

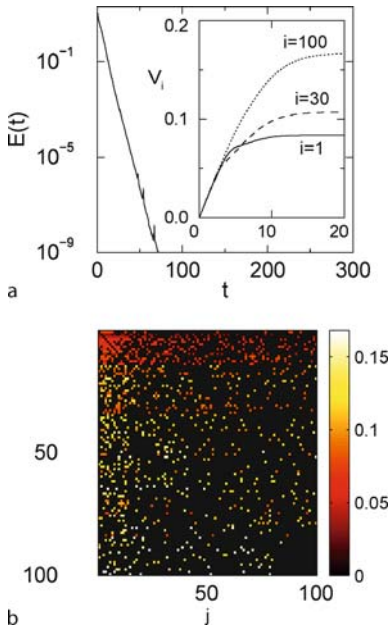

Synchronization Phenomena on Networks, Figure 12
a Transition to synchronization in an adaptive network of Rössler oscillators, indicated by the (averaged) synchronization error $E(t) = \langle |x_i - \langle x_i \rangle| \rangle$. *Inset*: the input strength $V_i(t)$ vs. time over three nodes. b The weighted coupling matrix $\tilde{G}$ crystallized after the adaptation (for the foodweb model) (after [89])

with exponent $\theta = 0.48 \pm 0.01$ for both oscillator models. Importantly, this scaling is also robust to the variation of network parameters, such as the minimal degree M (Fig. 13b), which should not be confused with the order parameter M elsewhere, the system size N (Fig. 13c), and the orders of magnitudes of the adaptation parameter γ (Fig. 13d).

When the stability zone is bounded, synchronization can always be achieved by the adaption mechanism of Eq. (14) if $\gamma \leq \gamma_c$ for a threshold γ_c somewhat depending on N and the oscillator dynamics. The two resulting weighted networks display the same power-law behavior as in Eq. (15), but with different exponents: $\theta = 0.54 \pm 0.01$ (Rössler oscillator) or $\theta = 0.36 \pm 0.01$ (foodweb). The eigenratio R for the weighted networks, after the adaptation, and for the unweighted networks (symmetric coupling) is calculated as a function of N (Fig. 14a), and as a function of the ratio S_{max}/S_{min} (Fig. 14b), where S_{max} and S_{min} are the maximum and minimum intensities of the variable coupling strengths of the model. Clearly, this adaptive coupling scheme is more effective than symmetric coupling for network synchronization.

Huang [90] investigated another adaptive coupling pattern, in which a node is coupled with its neighbors non-uniformly through different coupling strengths, and showed that they have better synchronizability than other networks with symmetric coupling patterns.

In all the coupling patterns discussed above, whether the coupling pattern is static or dynamic, only the coupling strength is tunable while the connectivity matrix always remains unchanged. However, as is intuitively clear, network synchronizability can also be significantly improved by evolving the graph topology giving rise to a time-varying connectivity matrix. This has been recently confirmed by Boccaletti et al. [91].

It has been shown [91] that to make a network synchronizable, either the coupling matrix $G(t) = G$ remains unchanged, or if starting from an initial wiring condition $G(0) = G_0$, the coupling matrix $G(t)$ commutes at any time with G_0, i. e., $G_0 G(t) = G(t) G_0, \forall t$. At any time, a zero-row-sum symmetric commuting matrix $G(t)$ can be constructed, as

$$G(t) = V \Lambda(t) V^{\mathrm{T}} , \tag{16}$$

where $V = \{\boldsymbol{v}_i, \ldots, \boldsymbol{v}_N\}$ is an orthogonal matrix with columns being the eigenvectors of G_0, and $\Lambda(t) = \mathrm{diag}[0, \lambda_2(t), \ldots, \lambda_N(t)]$ with $\lambda_i(t) > 0, \forall i > 1$. This set of matrices is referred to as the dissipative commuting set of $G(0)$. A condition to ensure the network synchronization will be stable is

$$S_i = \lim_{T \to \infty} \frac{1}{T} \int_0^T \Lambda_{\max}(\sigma \lambda_i(t')) \mathrm{d}t' < 0 \quad \forall i \neq 1 , \tag{17}$$

where $\Lambda_{\max}(\sigma \lambda_i)$ is the maximal transversal (conditional) Lyapunov exponent along the direction of the ith eigenvector, and S_i is its time average. Hence, it does not require $\Lambda_{\max}(\sigma \lambda_i(t)) < 0$ at all times. One can even construct a commutative evolution such that at each time there exists one eigenvalue λ_i for which $\Lambda_{\max}(\sigma \lambda_i(t)) > 0$, and yet obtain a stable synchronization manifold. Thus, interestingly, synchronization in a dynamical network can be achieved even in the case where each individual commutative graph does not give rise to synchronized behavior.

Modifications of Network Structures

It is well known that the synchronizability of a dynamical network is determined simultaneously by the network coupling pattern, the dynamical characteristics of the oscillators on its nodes, and the network structure. In the above, several cases with variable coupling patterns have been discussed. For some real-world networks, however, the coupling pattern cannot be modified at will. Thus, if the dynamics of the oscillators are given and fixed, and

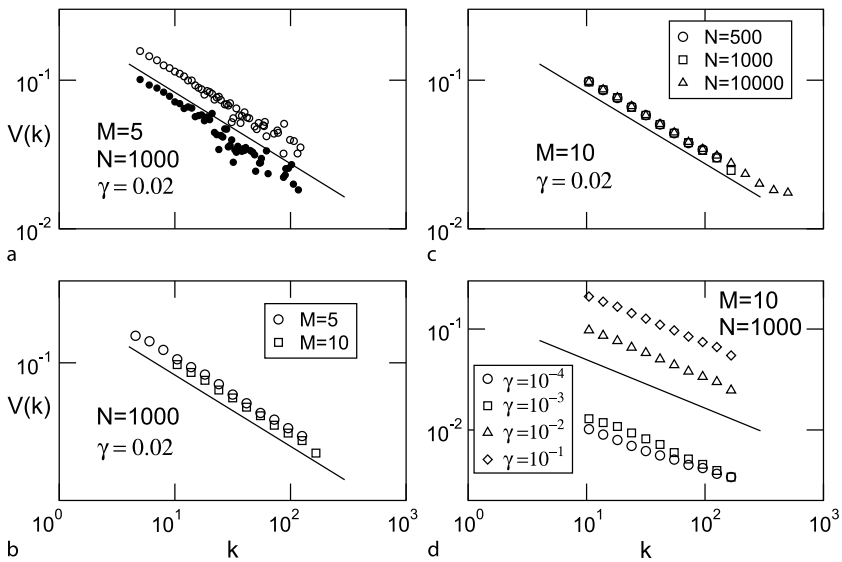

Synchronization Phenomena on Networks, Figure 13
Average input weight $V(k)$ of nodes with degree k as a function of k for a network of Rössler oscillators (*empty circles*) and the food-web model (*filled circles*) (a), and its dependence on various parameters, M (b), N (c), and γ (d), where the M should not be confused with the order parameter M elsewhere (after [89])

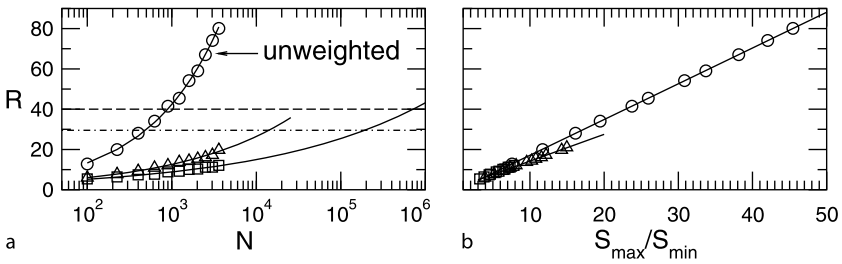

Synchronization Phenomena on Networks, Figure 14
Eigenratio R as a function of N (a), and S_{max}/S_{min} (b). The networks are synchronizable if $R < R_\epsilon$ in a, Rössler oscillators (*squares*), $R_\epsilon = 40$ (*dashed curve*), foodweb model (*triangles*), $R_\epsilon = 29$ (*dashed-dotted curve*) (after [89])

if the coupling patterns cannot be changed, then the only way to enhance the network synchronizability is to make a change to the network structure.

There are some effective techniques to enhance the network synchronizability by modifying the network structure, as further discussed below in the rest of this section.

Reducing Maximal Betweenness

In scale-free networks, the average distance is often very short while the node-degree and node-betweenness distributions are both quite broad. The bottleneck for the network synchronizability seems to be the maximal node betweenness [57]. In order to reduce the node betweenness of the hubs, Zhao et al. [92] suggested a method of structural perturbations. Specifically, for a hub x_0, $m-1$ auxiliary nodes, labeled as $x_1, \ldots, x_{m-1}$, are added around it. These m nodes are fully connected together. Then, all the edges incident from x_0 are re-distributed to all the nodes x_i (including x_0 itself), $i = 0, 1, \ldots, m-1$. After this process, the betweenness of x_0 is divided into m almost equal parts associating with these m nodes. This process is called m-division. A sketch map of a 3-division process on node x_0 is shown in Fig. 15.

Due to the huge sizes of many real-life networks, it is usually impossible to obtain the node betweenness from a complex network. Fortunately, studies have shown that there exists a strongly positive correlation between the node-degree and the node-betweenness in Barabasí–Albert networks and some other heterogeneous networks [87,93]. That is, a node with larger degree has

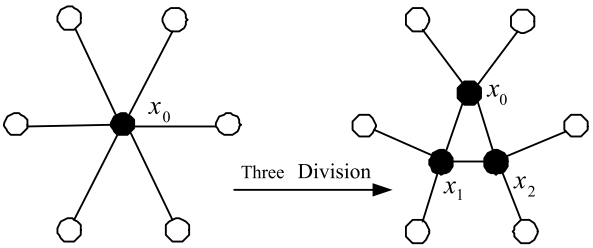

Synchronization Phenomena on Networks, Figure 15
Sketch map for the 3-division process on x_0. The *solid circle* on the *left* is the node x_0 with degree 6. After the 3-division process, this x_0 is divided into 3 nodes, x_0, x_1 and x_2, which are fully connected. The six edges incident from x_0 are then re-distributed to all the three nodes (after [92])

higher node-betweenness statistically. Therefore, for practical reasons, it can be assumed that nodes with higher betweenness are those with larger degrees in Barabasí–Albert networks.

To further explore how the structural perturbations affect the network synchronizability, the eigenratios before and after the m-division process were compared in [92] for a Barabasí–Albert scale-free network with the coupling matrix being Laplacian. For use in the rest of the article, we define a characteristic value $R = r'/r$, in which r and r' are the eigenratios before and after the division, respectively. Figure 16 shows the correlation between R and the probability ρ of the divided nodes. It is clear that even the m-division of a tiny fraction of nodes can sharply enhance network synchronizability.

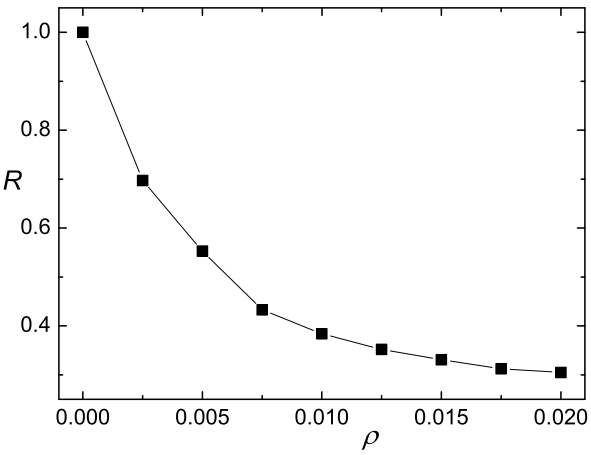

Synchronization Phenomena on Networks, Figure 16
Behavior of value R vs. the fraction of divided nodes ρ. As the number of divided nodes increases, R is reduced, leading to better synchronization (after [92])

Shortening the Average Distance

Zhou et al. [94] investigated the synchronizability of a network model named *crossed double cycles* (CDCs). They not only clarified the relationship between average distance and network synchronizability, but also provided a possible way to make a network more synchronizable.

In the language of graph theory [95,96,97], a cycle C_N denotes a network consisting of N nodes (vertices) $x_1, \ldots, x_N$. These N nodes are arranged in a ring, and the nearest two nodes are connected to each other. Thus, C_N has N edges connecting the nodes $x_1x_2, x_2x_3, \ldots, x_{N-1}x_N, x_Nx_1$. The set of all such CDCs, denoted by $G(N, m)$, can be constructed by adding two edges, called crossed edges, to each node in C_N. The two nodes connecting by a crossed edge have distance m within C_N. For example, the network $G(N, 3)$ can be constructed from C_N by connecting $x_1x_4, x_2x_5, \ldots, x_{N-1}x_2, x_Nx_3$ together. A sketch map of $G(20, 4)$ is shown in Fig. 17 for illustration.

Figure 18 shows how the average distance L affects the network synchronizability (measured by the characteristic value R). It is clear that the network synchronizability is very sensitive to the average distance: as L increases, R sharply spans more than three magnitudes. And the network synchronizability is remarkably enhanced by reducing L. When the crossed length m is not too small or too large (compared to N), networks with the same average distance have approximately the same synchronizability,

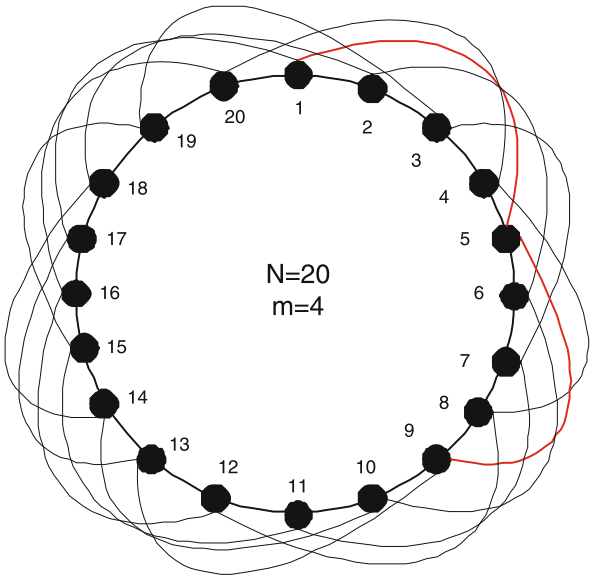

Synchronization Phenomena on Networks, Figure 17
Sketch map of $G(20, 4)$ (after [94])

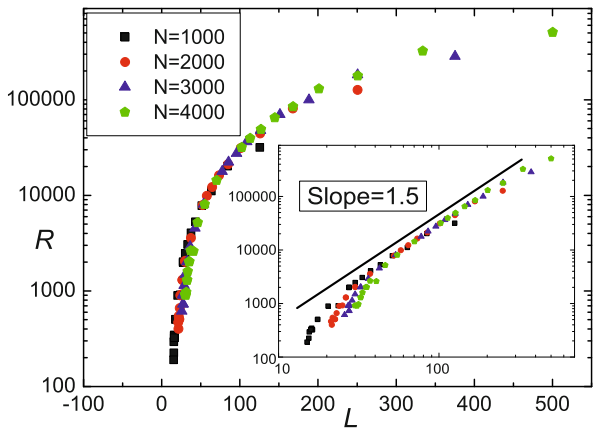

Synchronization Phenomena on Networks, Figure 18
Characteristic value *R* vs. average distance *L* of CDCs. The *black squares*, *red circles*, *blue triangles* and *green pentagons* represent the cases of $N = 1000$, 2000, 3000 and 4000, respectively. The *inset* shows the same data in log-log plot, indicating that the characteristic value *R* approximately obeys a power-law form $R \sim L^{1.5}$. The *solid line* has slope 1.5, for comparison (after [94])

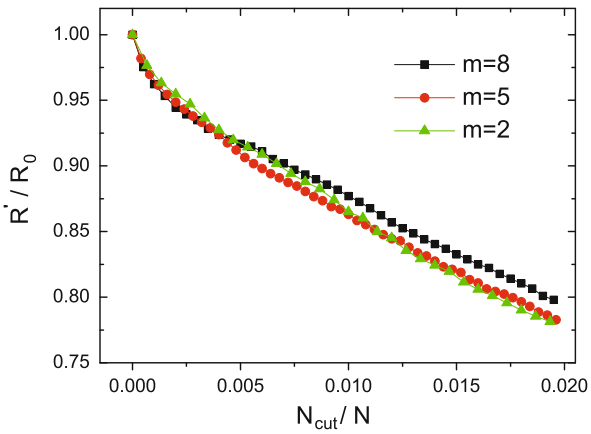

Synchronization Phenomena on Networks, Figure 19
Changes of the synchronizability as a function of the proportion of cut edges N_{cut}/N for different values of the average distance (after [99])

regardless of the network sizes. More interestingly, the numerical results show that the characteristic value R approximately obeys a power-law form, as $R \sim L^{1.5}$ (inset of Fig. 18).

Decoupling Nodes by Removing Heavily-Loaded Edges

In the synchronization process, not only hubs may be the bottlenecks but some edges with large loads may also limit the network synchronizability. Yin et al. [99] found that a scale-free network can become more synchronizable after some of its heavily-loaded edges have been removed. To reduce the computational cost, they used local information to approximately rank the edges, according to the values of $k_i \times k_j$, where i and j denote two adjacent nodes connected by an edge. Subsequently, at each time step, an edge with the highest rank is removed, i. e., the two nodes are decoupled at both sides of their connecting heavily-loaded edge. After this operation, the characteristic value is decreased, as shown by Fig. 19.

Designing the Output Function

Very recently, the relationship between graph theory and network synchronizability received some special attention [100]. For example, Duan et al. [101,102,103] found that for networks with disconnected complementary graphs, adding edges will often increase their synchronizability. The complementary graph of a given graph G is defined to be the graph consisting of all the nodes of G and all the edges that are not in G.

In addition, they found [101,102] that when the couplings between nodes are symmetric, an unbounded synchronized region is always easier to analyze than a bounded synchronized region (see Sect. "Basic Concepts of Network Synchronization" to recall their definitions). Therefore, to effectively enhance network synchronizability, they presented a design method for the output function (i. e., $\boldsymbol{H}$ in network (1), or the inner linking matrix in the linear coupling case), such that the resultant network has an unbounded synchronized region, for the case where the synchronous state is an equilibrium of the network.

If the synchronous state is an equilibrium, then both $D\boldsymbol{F}(\boldsymbol{s}(t))$ and $D\boldsymbol{H}(\boldsymbol{s}(t))$ in network (1), as discussed in Sect. "Enhancing Network Synchronizability" (part B), reduce to constant matrices, denoted by F and H, respectively. The synchronized region is the stability region of the matric pencil $F + \alpha H$ with respect to parameter α. It can be proved that there exists a matrix H of rank 1 (meaning that only one component in each state vector is used for coupling), such that the stability region is unbounded. The method for obtaining the desired output function is outlined below: first, take a column vector b such that (F, b) is stabilizable [104]; then, find a matrix $P = P^{\mathrm{T}}$ such that $FP + PF^{\mathrm{T}} - 2bb^{\mathrm{T}} < 0$; consequently, taking $k = b^{\mathrm{T}} P^{-1}$ leads to the stability of $F - \alpha bk$ for all α in the unbounded region; finally, $H = bk$ is the matrix to be found.

For illustration, synchronization of a simple 6-node network (shown in Fig. 20) is studied, where each node

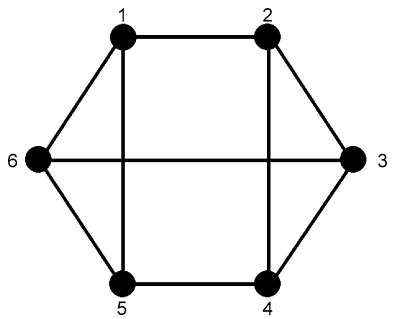

Synchronization Phenomena on Networks, Figure 20
A network of 6 nodes (after [101])

is located with a third-order smooth Chua's circuit [105]. At first, arbitrarily take the output function

$$H = \begin{pmatrix} 0.8348 & 9.6619 & 2.6591 \\ 0.1002 & 0.0694 & 0.1005 \\ -0.3254 & -8.5837 & -0.9042 \end{pmatrix}. \quad (18)$$

But the network does not synchronize. The states of node 1 are shown in Fig. 21a. Then, let $b = (0, 0, 1)^{\mathrm{T}}$ and $k = (0.0708, -0.15590, 0.4296)$, and then set $H = bk$, so synchronization is achieved as guaranteed by the theory. Figure 21b shows that the states of node 1 quickly reach the equilibrium.

Future Research Outlook

Complex network synchronization is a rapidly growing subject attracting increasing attention from various fields of physics, engineering, mathematics, and biology alike. Despite the current great advances and progress, there are still many important open questions.

In the studies of static coupling, Nishikawa and Motter [48] once pointed out that optimal global synchronizability, with eigenratio being equal to 1, can be obtained from a directed network structure without loops. Even if adding one loop of length 2 (in a directed network, two opposite edges between node i and node j can be considered as a loop of length 2), the eigenratio will be doubled [73,85]. Another scenario is shown by extending the conclusion in [48] to the case of non-identical oscillators [106]. Some further works in this direction will be helpful for in-depth understanding about the role of loops in network synchronization.

In the studies of dynamic coupling, the cost of coupling has not been taken into account. However, cost is usually very significant in some self-driven systems (for example, in wireless sensor networks [107,108] and in distributed autonomous robotic systems [109]). For each node to report its current state to the neighbors (or to detect the states of all its neighbors) requires a certain amount of power, while the total power assigned to each node is often limited, even if such communications are possible. Yet, as found in collective behaviors of biological swarms, a few effective leaders can well organize the whole population [110]. And a recent study has pointed out that partial coupling is more than enough to keep the coherence of self-propelled particles [111]. Therefore, it is very natural to expect to synchronize a complex network with a very low cost, which is an important issue for further investigation.

Very recently, there are some attempts at detecting the network structures with the help of the synchronization phenomenon on complex networks [112,113,114]: to discover the hierarchical community structure by the dynamic time scales of the network synchronization process [112,113], or to infer the complete connectivity of a network from its stable response dynamics [114], etc. These seem quite useful for optimal network design, analy-

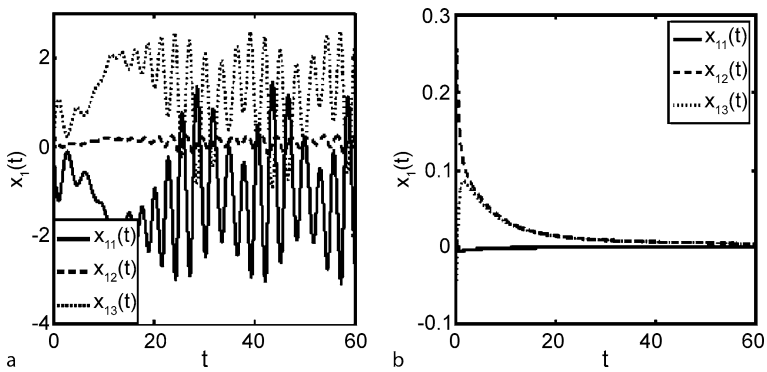

Synchronization Phenomena on Networks, Figure 21
States of node 1 (after [101])

sis, and utilization in general, therefore should be pursued with special efforts.

Similar to the aforementioned open questions, many theoretically attractive and practically important problems about various aspects of synchronization on complex networks can be posted and described. As the network research further evolves in different fields, many new dynamical phenomena and analytic issues will also emerge. Importance notwithstanding, the subject of "Synchronization Phenomenon of Networks" will continue to prove itself an theoretically interesting and technically challenging subject for scientific research in the years to come.

Acknowledgments

The authors thank the American Physics Society and the Elsevier Publisher for their permission to use some of their published simulation figures, which have all been referenced (in the figure captions). The authors also thank Dr. Jian-Guo Liu for his helpful comments and suggestions. G.R.C. was supported by the NSFC/RGC joint research scheme under grant N-CityU107/07. B.H.W. was supported by the NNSFC under Grants 10472116 and A0524701, and the Specialized Program under the Presidential Funds of CAS. T.Z. was supported by NNSFC under Grants 10635040 and 70471033.

Bibliography

1. Albert R, Barabási AL (2002) Rev Mod Phys 74:47
2. Dorogovtsev SN, Mendes JFF (2002) Adv Phys 51:1079
3. Newman MEJ (2003) SIAM Rev 45:167
4. Boccaletti S, Latora V, Moreno Y, Chaves M, Hwang DU (2006) Phys Rep 424:175
5. da F Costa L, Rodrigues FA, Travieso G, Boas PRV (2007) Adv Phys 56:167
6. Pastor-Satorras R, Vespignani A (2001) Phys Rev Lett 86:3200
7. Zhu CP, Xiong SJ, Tian YJ, Li N, Jiang KS (2004) Phys Rev Lett 92:218702
8. Zhou T, Yan G, Wang BH (2005) Phys Rev E 71:046141
9. Zhou T, Fu ZQ, Wang BH (2006) Prog Nat Sci 16:452
10. Zhou T, Liu JG, Bai WJ, Chen GR, Wang BH (2006) Phys Rev E 74:056109
11. Motter AE, Lai YC (2002) Phys Rev E 66:065102
12. Goh KI, Lee DS, Kahng B, Kim D (2003) Phys Rev Lett 91:148701
13. Motter AE (2004) Phys Rev Lett 93:098701
14. Zhou T, Wang BH (2005) Chin Phys Lett 22:1072
15. Galstyan A, Cohen P (2007) Phys Rev E 75:036109
16. Guimerá R, Díaz-Guilera A, Vega-Redondo F, Cabrales A, Arenas A (2002) Phys Rev Lett 89:248701
17. Tadić B, Thurner S, Rodgers GJ (2004) Phys Rev E 69:036102
18. Zhao L, Lai YC, Park K, Ye N (2005) Phys Rev E 71:026125
19. Yan G, Zhou T, Hu B, Fu ZQ, Wang BH (2006) Phys Rev E 73:046108
20. Wang BH, Zhou T (2007) J Korean Phys Soc 50:134
21. Watts DJ, Strogatz SH (1998) Nature 393:440
22. Barabási AL, Albert R (1999) Science 286:509
23. Strogatz SH, Stewart I (1993) Sci Am 269:102
24. Gray CM (1994) J Comput Neurosci 1:11
25. Néda Z, Ravasz E, Vicsek T, Barabási AL (2000) Phys Rev E 61:6987
26. Glass L (2001) Nature 410:277
27. Chen G, Dong X (1998) From Chaos Order. World Scientific, Singapore
28. Erdös P, Rényi A (1960) Pub Math Ins Hung Acad Sci 5:17
29. Wang XF (2002) Int J Bifurc Chaos 12:885
30. Heidelberger R (2007) Nature 450:625
31. Heagy JF, Carroll TL, Pecora LM (1994) Phys Rev E 50:1874
32. Stefański A, Perlikowski P, Kapitaniak T (2007) Phys Rev E 75:016210
33. Wu CW, Chua LO (1995) IEEE Trans Circuits Syst I 42:430
34. Jost J, Joy MP (2001) Phys Rev E 65:016201
35. Gade PM (1996) Phys Rev E 54:64
36. Manrubia SC, Mikhailov AS (1999) Phys Rev E 60:1579
37. Gade PM, Hu CK (1999) Phys Rev E 60:4966
38. Gade PM, Hu CK (2000) Phys Rev E 62:6409
39. Lago-Fernández LF, Huerta R, Corbacho F, Sigüenza JA (2000) Phys Rev Lett 84:2758
40. Wang XF, Chen G (2002) Int J Bifurc Chaos Appl Sci Eng 12:187
41. Wang XF, Chen G (2002) IEEE Trans Circuits Syst I 49:54
42. Barahona M, Pecora LM (2002) Phys Rev Lett 89:054101
43. Jiang PQ, Wang BH, Bu SL, Xia QH, Luo XS (2004) Int J Mod Phys B 18:2674
44. Lind PG, Gallas JAC, Herrmann HJ (2004) Phys Rev E 70:056207
45. Pecora LM, Carrol TL (1998) Phys Rev Lett 80:2109
46. Hu G, Yang J, Liu W (1998) Phys Rev E 58:4440
47. Pecora LM, Barahona M (2005) Chaos Complex Lett 1:61
48. Nishikawa T, Motter AE (2006) Phys Rev E 73:065106
49. Kuramoto Y (1975) In: Araki H (ed) Internaltional Symposium on Mathematical Problems in Theoretical Physics. Lecture Notes in Physics, vol 30. Springer, New York
50. Kuramoto Y (1984) Chemical Oscillations, Wave and Turbulence. Springer, Berlin
51. Kuramoto Y, Nishikawa I (1987) J Stat Phys 49:569
52. Pikovsky A (2001) Synchronization. Cambridge University Press, Cambridge
53. Acebrón JA, Bonilla LL, Vicente CJP, Ritort F, Spigler R (2005) Rev Mod Phys 77:137
54. Nishikawa T, Motter AE, Lai YC, Hoppensteadt FC (2003) Phys Rev Lett 91:014101
55. Newman MEJ, Strogatz SH, Watts DJ (2001) Phys Rev E 64:026118
56. Dorogovtsev SN, Mendes JFF (2000) Phys Rev E 62:1842
57. Hong H, Kim BJ, Choi MY, Park H (2004) Phys Rev E 69:067105
58. Freeman L (1979) Soc Netw 1:215
59. Newman MEJ (2001) Phys Rev E 64:016132
60. Zhou T, Liu JG, Wang BH (2006) Chin Phys Lett 23:2327
61. Fan ZP (2006) Complex Networks: From Topology to Dynamics. Ph D Thesis, City University of Hong Kong
62. Zhao M, Zhou T, Wang BH, Yan G, Yang HJ, Bai WJ (2006) Physica A 371:773
63. Maslv S, Sneppen K (2002) Science 296:910
64. Kim BJ (2004) Phys Rev E 69:045101(R)
65. McGraw PN, Menzinger M (2005) Phys Rev E 72:015101
66. Gómez-Gardeñes J, Moreno Y, Arenas A (2007) Phys Rev Lett 98:034101

67. Wu X, Wang BH, Zhou T, Wang WX, Zhao M, Yang HJ (2006) Chin Phys Lett 23:1046
68. Holme P, Kim BJ (2002) Phys Rev E 65: 026107
69. Newman MEJ (2002) Phys Rev Lett 89:208701
70. di Bernardo M, Garofalo F, Sorrentino F (2007) Int J Bifurc Chaos 17:3499
71. Sorrentino F, di Bernardo M, Cuéllar GH, Boccaletti S (2006) Physica D 224:123
72. Chavez M, Hwang DU, Martinerie J, Boccaletti S (2006) Phys Rev E 74:066107
73. Zhao M, Zhou T, Wang BH, Ou Q, Ren J (2006) Eru Phys J B 53:375
74. Huang L, Park K, Lai YC, Yang L, Yang K (2006) Phys Rev Lett 97:164101
75. Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D (2004) Proc Natl Acad Sci USA 101:2658
76. Zhou T, Zhao M, Chen G, Yan G, Wang BH (2007) Phys Lett A 368:431
77. Donetti L, Hurtado PI, Muñoz MA (2005) Phys Rev Lett 95:188701
78. Motter AE, Zhou C, Kurths J (2005) Phys Rev E 71:016116
79. Motter AE, Zhou C, Kurths J (2005) Europhys Lett 69:334
80. Motter AE, Zhou C, Kurths J (2005) AIP Conf Proc 776:201
81. Hwang DU, Chavez M, Amann A, Boccaletti S (2005) Phys Rev Lett 94:138701
82. Dorogovtsev SN, Mendes JFF, Samukhin AN (2000) Phys Rev Lett 85:4633
83. Krapivsky PL, Redner S (2001) Phys Rev E 63:066123
84. Zou Y, Zhu J, Chen G (2006) Phys Rev E 74:046107
85. Chavez M, Hwang DU, Amann A, Hentschel HGE, Boccaletti S (2005) Phys Rev Lett 94:218701
86. Chavez M, Hwang DU, Amann A, Boccaletti S (2006) Chaos 16:015106
87. Goh KI, Kahng B, Kim D (2001) Phys Rev Lett 87:278701
88. Wang X, Lai YC, Lai CH (2007) Phys Rev E 75:056205
89. Zhou C, Kurths J (2006) Phys Rev Lett 96:164102
90. Huang D (2006) Phys Rev E 74:046208
91. Boccaletti S, Hwang DU, Chavez M, Amann A, Kurths J, Pecora LM (2006) Phys Rev E 74:016102
92. Zhao M, Zhou T, Wang BH, Wang WX (2005) Phys Rev E 72:057102
93. Barthélemy M (2004) Eur Phys J B 38:163
94. Zhou T, Zhao M, Wang BH (2006) Phys Rev E 73:037101
95. Bondy JA, Murty USR (1976) Graph Theory with Applications. MacMillan, London
96. Bollobás B (1998) Modern Graph Theory. Springer, New York
97. Xu JM (2003) Theory and Application of Graphs. Kluwer, Dordrecht
98. Newman MEJ, Watts DJ (1999) Phys Rev E 60:7332
99. Yin CY, Wang WX, Chen G, Wang BH (2006) Phys Rev E 74:047102
100. Comellas F, Gago S (2007) J Phys A Math Theor 40:4483
101. Duan Z, Chen G, Huang L (2007) Phys Rev E 76:056103
102. Duan Z, Chen G, Huang L (2007) Phys Lett A 372:3741
103. Duan Z, Liu C, Chen G (2008) Physica D 237:1006
104. Friedland B (1986) Control System Design. McGraw-Hill, New York
105. Tsuneda A (2005) Int J Bifurc Chaos 15:1
106. Um J, Han SG, Kim BJ, Lee SI (2008) (unpublished)
107. Akyildiz IF, Su W, Sankarasubramaniam Y, Cayirci E (2002) Comput Netw 38:393
108. Oqren P, Fiorelli E, Leonard NE (2004) IEEE Trans Automat Contr 49:1292
109. Arai T, Pagello E, Parker LE (2002) IEEE Trans Robot Automat 18:655
110. Couzin LD, Krause J, Franks NR, Levin SA (2005) Nature 433:513
111. Zhang HT, Chen M, Zhou T (2007) arXiv:0707.3402
112. Arenas A, Díaz-Guilera A, Pérez-Vicente CJ (2006) Phys Rev Lett 96:114102
113. Boccaletti S, Ivanchenko M, Latora V, Pluchino A, Rapisarda A (2007) Phys Rev E 75:045102(R)
114. Timme M (2007) Phys Rev Lett 98:224101

Synergetics: Basic Concepts

Hermann Haken
Institut für Theoretische Physik, Universität Stuttgart, Stuttgart, Germany

Article Outline

Glossary
The Role of Synergetics in Science
The Laser Paradigm
The Hierarchical Structure of Synergetics
Basic Equations
Method of Solution
Quantum Theoretical Formulation
Regular Spatial and Spatio-Temporal Patterns
A Further Mathematical Tool: Shannon Information and the Maximum (Information) Entropy Principle
Phenomenological Synergetics
Semantic Synergetics
Some Selected Examples
History and Relations to Other Fields
Future Directions
Bibliography

Glossary

Synergetics Science of cooperation.

Pattern A pattern is essentially an arrangement. It is characterized by the order of the elements of which it is made rather than by the intrinsic nature of these elements (Norbert Wiener).

Self-organization Formation of spatio-temporal patterns (structures) and/or performance of functions without an "ordering hand".

State vector Set of time- or time-independent variables that characterize the state of a system.

Evolution equations Determine the temporal evolution of the state vector. May be deterministic, stochastic or both.

Control parameter One or a set of (mostly externally) fixed parameters in the evolution equations.

Spectrum Set of eigenvalues belonging to linear stability equations with boundary conditions.

Stability of a system System returns after a (small) perturbation of its state vector into original state.

Instability Loss of stability.

Order parameters Collective variables that determine the macroscopic behavior of systems.

Slaving principle A general theorem that allows the reduction of the variables of a system to order parameters (close to instability).

Trajectory Smooth curve $q(t)$ of solution of evolution equation in q-space.

Attractor Region in the state vector space ("q-space") to which all neighboring states are attracted in the course of time.

Fixed point, stable Point in q space to which all neighboring trajectories converge in course of time.

Limit cycle, stable A closed trajectory to which all neighboring trajectories converge.

Probability distribution function Function that determines the probability of a random variable r to have fixed value $r = r_0$.

Fokker Planck equation Evolution equation for probability density function, based on drift and diffusion.

Normal form Especially simple polynomial expression that still captures the essential features, e. g. of the right hand side of deterministic evolution equations.

Schrödinger picture of quantum mechanics In it operators are time-independent, while the wave-function ("state vector") is time-dependent and determined by the Schrödinger equation.

Heisenberg picture in quantum mechanics The state vector is time-independent, while the operators are time-dependent and determined by Heisenberg equations of motion.

Fluctuating forces Stochastic (random) forces appearing in evolution equations.

Quantum classical correspondence Establishes relation between quantum mechanical density matrix and classical quasi-probability distribution.

Symmetry Invariance of a system against specific transformations (e. g. mirror symmetry).

Group Set of elements with specific multiplication rules (axioms).

Dynamical system System whose state vector changes in the course of time deterministically.

Langevin equation Originally: evolution equation for velocity of a Brownian particle subject to damping and fluctuating force.

Generalized Langevin equation General evolution equations that contain both a deterministic and a stochastic part ("fluctuating forces").

Hamilton operator Classical Hamilton function, in which variables, e. g. position x and momentum p, are replaced by quantum mechanical operators.

Spatial coordinate (vector $\boldsymbol{x}$) in one, two or three dimensions.

Δ Laplace operator (in 1,2 or 3 dimensions) .

∇ Vector $\left(\frac{\mathrm{d}}{\mathrm{d}x_1}, \frac{\mathrm{d}}{\mathrm{d}x_2}, \frac{\mathrm{d}}{\mathrm{d}x_3}\right)$ in 1,2 or 3 dimensions .

The Role of Synergetics in Science

In science, we may essentially distinguish between two trends:

1. The accumulation of knowledge
2. Information reduction in the sense of finding general principles, common features.

In physics, such unifying approaches are well known: the unification of magnetism, electricity and, later on, weak and other interactions leading eventually to a unified field theory. General relativity unifies concepts of space, time and gravitation. While these unifications take place at a fundamental level, one may ask whether it is worthwhile to look also for unifications at say more macroscopic or phenomenological levels. One example is thermodynamics, another the theory of phase transitions of systems in thermal equilibrium by means of the renormalization group approach, or the concept of fractals, etc.

The main goal of Synergetics is the search for unifying principles for systems that are composed of many individual parts or components, and that may show the phenomenon of self-organization, i. e. the spontaneous formation of spatial, temporal, spatial-temporal or functional structures. The systems under discussion are, in the widest sense of the word, open physical systems whose states are maintained by an in- and outflux of energy, matter and /or information. A typical and well known example is that of a fluid in a pan that is uniformly heated from below. When the temperature difference between the lower and upper surface exceeds a critical value, the formerly homogeneous fluid develops roll or hexagonal patterns in which the fluid moves in a specific manner (Fig. 1).

As it turned out, the general principles originally elaborated in physics, can also be applied to many other systems, such as in biology, economy, ecology, sociology,

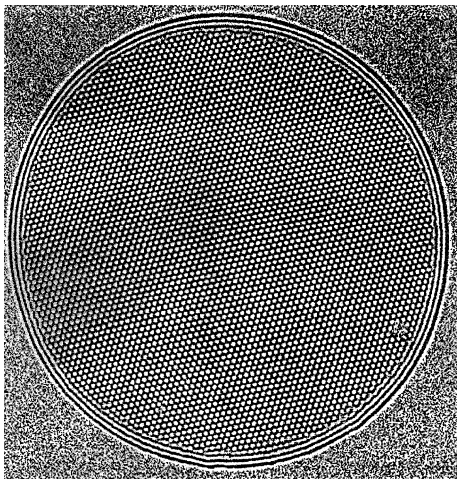

Synergetics: Basic Concepts, Figure 1
Hexagonal pattern of a fluid (liquid helium) uniformly heated from below [12]

management theory, psychology etc. In spite of the great variety of the individual systems with their components quite different in nature, such principles apply to large classes of phenomena. This is achieved by restricting the study to situations where the systems undergo qualitative changes at macroscopic scales. Here macroscopic means "with time and length scales large compared to those of the individual components".

This leads to the definition of Synergetics as given in the preamble of the Springer Series in Synergetics: "An ever increasing number of scientific disciplines deal with complex systems. These are systems that are composed of many parts which interact with one another in a more or less complicated manner. One of the most striking features of many such systems is their ability to spontaneously form spatial or temporal structures. A great variety of these structures are found, in both the inanimate and the living world. In the inanimate world of physics and chemistry, examples include the growth of crystals, coherent oscillations of laser light, and the spiral structures formed in fluids and chemical reactions. In biology we encounter the growth of plants and animals (morphogenesis) and the evolution of species. In medicine we observe, for instance, the electromagnetic activity of the brain with its pronounced spatio-temporal structures. Psychology deals with characteristic features of human behavior ranging from simple pattern recognition tasks to complex patterns of social behavior. Examples from sociology include the formation of public opinion and cooperation or competition between social groups."

In recent decades, it has become increasingly evident that all these seemingly quite different kinds of structure formation have a number of important features in common. The task of studying analogies as well as differences between structure formation in these different fields has proved to be an ambitious but highly rewarding endeavor. The Springer Series in Synergetics provides a forum for interdisciplinary research and discussions on this fascinating new scientific challenge. It deals with both experimental and theoretical aspects. The scientific community and the interested layman are becoming ever more conscious of concepts such as self-organization, instabilities, deterministic chaos, nonlinearity, dynamical systems, stochastic processes, and complexity. All of these concepts are facets of a field that tackles complex systems, namely Synergetics.

The Laser Paradigm

This example elucidates central concepts used in Synergetics in a qualitative fashion. An example for the laser device (an acronym for light amplification by stimulated emission of radiation, originally called optical maser [121]) is the gas laser in which gas atoms are enclosed in a tube at the end-faces of which mirrors are mounted. The mirrors serve the purpose of reflecting light running in axial direction sufficiently often so that the corresponding light wave stays for an extended period in this device and can interact intensely with the atoms. The atoms are excited from the outside, e. g. by a pump light source. After having been excited, each atom can spontaneously emit a light wave track. In the usual case of a *lamp*, these wave tracks are emitted independently of each other and the amplitudes are Gaussian distributed. When the pump intensity is increased beyond a critical value, the present state gives way to a single wave with *stable amplitude* on which small *amplitude fluctuations* and *phase diffusion* are superimposed [53]. The pump intensity serves as *control parameter*. At its critical value, the old state becomes *unstable*. The emerging coherent wave acts as *order parameter* that via stimulated emission forces the electrons of the gas molecules to emit light waves in a coherent fashion. This action of the order parameter on the individual parts of the system is called *slaving principle*. If the pump power is increased further, more instabilities can appear, and a variety of temporal but also spatio-temporal patterns of light waves may appear, such as laser light chaos [55] or ultrashort laser pulses. The first laser threshold shows the typical features of a phase transition of a system in thermal equilibrium, namely critical slowing down, critical fluctuations and symmetry breaking [25,46,53,58,119], as well as the emergence of a c-number amplitude of the quantized light field (Fig. 2).

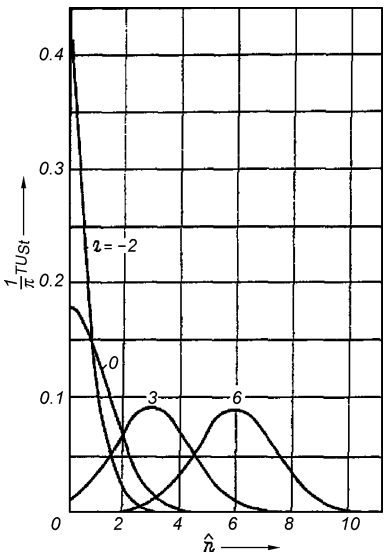

Synergetics: Basic Concepts, Figure 2
The stationary distribution function of the laser light intensity as a function of the normalized intensity $\hat{n}$. The individual *curves* refer to different normalized pump power values a, where $a < 0$ below threshold, $a = 0$ at threshold, $a > 0$ above threshold (after [116])

The Hierarchical Structure of Synergetics

Before I discuss the mathematical approach in detail and to provide the ground for farther reaching applications, I hint at the three levels of Synergetics:

1. *The microscopic theory*, based either on microscopic equations, such as in the laser example, those of quantum mechanics and quantum field theory, or in biology on mathematical models on the behavior of individual parts of a system. At this level, concepts, such as order parameters and enslavement (cf. Sect. "The Laser Paradigm"), can be mathematically derived.
2. *Phenomenological Synergetics* directly starts from concepts, such as order parameters and enslavement, which then may be cast into mathematical relations.
3. *Semantic Synergetics* deals with cases where a mathematical formulation is (at present or in principle) not possible, but still formulations using concepts and relationships unearthed in Synergetics are applicable.

A general goal of Synergetics consists in elaborating relationships between levels 1, 2, 3.

In the present article I will mainly focus my attention on the mathematical formulation dealing with 1. and 2.

Basic Equations

The basic equations are classical or quantum mechanical evolution equations, in which the temporal evolution of the microscopic quantities under consideration is described by ordinary or partial differential equations. Since the systems are open, the inputs and outputs of energy, matter and/or information must be taken care of, which, quite often, appears in the form of coupling to heat baths in the sense of thermodynamics. In open systems, these heat baths must be kept at different temperatures, in order to maintain the non-equilibrium state of the system. The heat bath variables can be eliminated which gives rise to differential equations which contain "pumping" and "damping" terms as well as fluctuating (stochastic) forces. In the case of quantum mechanical equations the stochastic forces are operators. With the inclusion of stochastic forces, the classical or quantum mechanical equations acquire the character of stochastic differential equations which may be called "generalized Langevin equations".

Depending on the definition of the random forces, we may distinguish between the $\hat{I}$ to, the Statonovich and the Klimontovich approach [62,72,134]. As is well known in statistical physics, Langevin equations can be converted into equations for distribution functions, such as e. g. the Fokker–Planck equation. A further approach, mainly used in quantum mechanics, but also in models on sociodynamics, is the master equation.

In order not to overload this article, I will focus my attention on the treatment of evolution equations.

This approach seems to be particularly suited for the treatment of phase transition- like phenomena, i. e. the transitions between qualitatively different states of a system. If noise is neglected and transients are not treated, these transitions are called bifurcations [5,22,49,71,79,84,91].

At the microscopic level the systems are described by a state vector q with components $q_1, \ldots, q_n$ which may also be space dependent, $q_j = q_j(x, t)$, where x is a one, two or three dimensional vector. The time dependence is described by evolution equations of the form of a vector equation.

$$\dot{q} = N(q, \nabla, \alpha) + F(q, \nabla, \alpha) \,. \tag{1}$$

The dot $\dot{}$ means time-derivative. N is a vector valued function that depends on q in a nonlinear fashion. ∇ indicates spatial derivatives (of any order) or non-local integrations e. g. of the form

$$\int K(x, x')q(x')\,\mathrm{d}x' \tag{2}$$

where K is a matrix.

α represents a set of fixed control parameters. If not otherwise stated, we explicitly treat only one control parameter. Equation (1) must be supplemented by appropriate boundary and initial conditions. F is a vector valued stochastic function of time with vanishing mean.

Method of Solution

We assume that for a certain control parameter value α_0 the state vector as solution of Eq. (1) is known, $q = q_0$. The following cases have been considered, see e. g. [62]:

a) q_0 is a stable fixed point (Section "Instability of a Fixed Point")
b) q_0 is a stable limit cycle (Section "Instability of a Limit Cycle, $q_0(t)$ [62]")
c) q_0 is a stable n-dimensional torus. (Section "Instability of Tori [62]")

Now the control parameter value is changed and the stability of the system is checked by means of linear stability analysis [52].

Instability of a Fixed Point We first elucidate our general procedure by means of the instability of an originally stable fixed point. This procedure differs from the classical approach of bifurcation theory [90,124] in two important aspects:

1. The role of the fluctuating forces is fully taken into account in order to be able to make contact with the theory of phase transitions in the Landau sense [86].
2. The approach covers the surrounding of the fixed point in order to deal with relaxation processes towards the newly evolving stable states.

The hypothesis

$$q(t) = q_0 + W(t) \tag{3}$$

is inserted into (1) and the Eq. (1) with $F \equiv 0$ linearized with respect to $W(t)$,

$$\dot{W} = LW \tag{4}$$

where L may be a linear differential (or integral) linear operator.

The solutions are of the form

$$W(x,t) = \mathrm{e}^{\lambda_k t} \sum_{d=0}^{D} t^d v_{k,d}(x) \tag{5}$$

where $D > 0$ may happen if the corresponding eigenvalue λ_k is degenerate. In the following we consider $D = 0$ and $v_{k,d} = v_k$. The unstable modes $v_k \equiv v_u$ are connected with

$$\mathrm{Re}\,\lambda_k \geq 0\,, \tag{6}$$

the stable modes $v_k \equiv v_s$ with

$$\mathrm{Re}\,\lambda_k < 0\,. \tag{7}$$

It is assumed that $\mathrm{Re}\,\lambda_k < A < 0$, A fixed, if the eigenvalues are discrete.

We decompose the wanted solution to the original non-linear and stochastic equations into a super position of modes determined by the instability analysis whereby we distinguish between the unstable and stable modes. The amplitudes of the unstable modes are the order parameters. Inserting

$$q(t) = q_0 + \sum_u \xi_u(t) v_u(x) + \sum_s \xi_s(t) v_s(x) \tag{8}$$

into the Eqs. (1) and projecting both sides of the resulting equation on the stable and unstable modes, we obtain equations of the form

$$\dot{\xi}_u = \lambda_u \xi_u + \hat{N}_u\left(\{\xi_u\}, \{\xi_s\}\right) + \hat{F}_u\left(\{\xi_u\}, \{\xi_s\}\right) \tag{9}$$

$$\dot{\xi}_u = \lambda_s \xi_s + \hat{N}_s\left(\{\xi_u\}, \{\xi_s\}\right) + \hat{F}_s\left(\{\xi_u\}, \{\xi_s\}\right)\,. \tag{10}$$

λ_u, λ_s are the eigenvalues (6), (7), which are assumed to be discrete. By a suitable, in general nonlinear, transformation to new variables, $\tilde{N}(\{\xi_u\})$ can be cast into a particularly simple form ("normal form" theory [101,103], initiated by Poincaré [113]).

If the eigenvalues $\lambda_u, 0 > \mathrm{Re}\,\lambda_s > -|B|$ are a continuous function of an index, e. g. a wave number k, wave packets of $\xi_u(t)$ are used as new order parameter variables Ξ and $\lambda_u(k)$ is replaced by an operator $\Lambda_u(-i\frac{\mathrm{d}}{\mathrm{d}x})$ in one space-dimension or, more generally, $\Lambda_u = (-i\nabla)$ [62]. For a related approach in fluid dynamics cf. [105].

The central idea of further procedure consists in eliminating the amplitudes of the stable modes. This is achieved by the *slaving principle* [56,62,65,148] which allows us to express the amplitudes of the stable modes in terms of the unstable modes

$$\xi_s(t) = f_s\left(\{\xi_u(t)\}, t\right)\,, \tag{11}$$

where ξ_s, ξ_u are taken at the same time t. The explicit time-dependence of f_s stems exclusively from that of the fluctuating forces. f_s can be explicitly calculated in terms of a series expansion in powers of the order parameters. For practical purposes, in general only a few terms are needed.

For a general discussion of the convergence of this series see [62]. When noise is neglected, contact can be made with center manifold theory [76,112], which originally was a mere existence theory and was not constructive. For more recent developments, see books on bifurcation theory. A related approach is based on time-scale separation: The slowly damped or undamped modes serve as order parameters, which enslave the rapidly damped modes. A special case is adiabatic elimination.

Resulting Langevin Equations The enslaved mode amplitudes can be expressed by the order parameters and inserted in (9), so that closed equations for the order parameters alone result.

$$\dot{\xi}_u = \lambda_u \xi_u + \tilde{N}_u(\{\xi_u\}) + \tilde{F}_u(\{\xi_u\}, t) \tag{12}$$

where $\tilde{N}$ is a polynominal of $\xi(x,t)$ starting with at least second order. $\tilde{F}$ is a stochastic force. A simple, yet prototypical example is (with a single order parameter $\xi = \xi_u$)

$$\dot{\xi} = \lambda\xi + a\xi^2 - b\xi^3 + F(t)\,, \quad b > 0 \tag{13}$$

or

$$\dot{\xi} = -\frac{\partial V(\xi)}{\partial \xi} + F(t)\,, \tag{14}$$

with the potential

$$V = -\frac{\lambda}{2}\xi^2 - \frac{a}{3}\xi^3 + \frac{b}{4}\xi^4\,. \tag{15}$$

If λ_u, λ_s (6), (7) represent a continuous spectrum, (generalized) Ginzburg–Landau equations result [62]. For example, the complex Ginzburg–Landau equation with fluctuating force reads [6].

$$\begin{aligned} &(\xi(x,t) \equiv \xi_u\,,\ \text{complex order parameter}) \\ &\dot{\xi} = \lambda\xi + a\Delta\xi - c\,|\xi|^2\,\xi + F(t)\,. \end{aligned} \tag{16}$$

A further example is given by the Swift–Hohenberg equation [135], see also [24] (which was derived differently, however)

$$\dot{\xi}(x,t) = (a - b\Delta)^2\xi(x,t) + c\xi(x,t) - d\xi(x,t)^3\,. \tag{17}$$

The Eqs. (12,13,16,17) allow for a great variety of solutions. In the case of real λ and a single order parameter, a nonequilibrium phase transition occurs (see below). In case of λ complex, and (at least) one complex order parameter, Landau–Hopf bifurcation [67,68], i. e. formation of a limit cycle may happen. In case of (at least) three order parameters and no noise, deterministic chaos may occur [89,118,132] (in the presence of noise, mixed effects may occur).

Fokker–Planck Equation Below and above the instability point in control parameters space, in a first step the fluctuations can be neglected and then, in the next step, taken care of by means of lowest order perturbation theory. In order to cover the transition region, under well defined conditions a Fokker–Planck equation for the probability density function $f(\{\xi_u\})$ of the order parameters can be derived. For details see [62,63a}] and the article by T. Frank, this volume.

The Fokker–Planck equation is of the general form

$$\dot{f}(\{\xi_u\}) = -\sum_u \frac{\partial}{\partial \xi_u}\left(\tilde{N}_u f\right) + \frac{1}{2}\sum_{uv}\frac{\partial^2}{\partial\xi_u\partial\xi_v}\left(Q_{uv} f\right)\,. \tag{18}$$

It is assumed that $\tilde{F}_u$ in (12) is δ correlated in time,

$$\langle \tilde{F}_u(t)\tilde{F}_v(t')\rangle = Q_{uv}\delta(t-t')\,. \tag{19}$$

If $\tilde{F}_u$ depends on ξ_u, the $\hat{I}$ to, Stratonovich or Klimontovich procedure must be applied.

In the case of a single order parameter, where the Langevin equation [88], originally with $\tilde{N} = -\alpha\xi$) is given by

$$\dot{\xi} = \tilde{N}(\xi) + F(t)\,, \quad \langle F(t)F(t')\rangle = Q\delta(t-t')\,. \tag{20}$$

The steady state distribution function of (18) is given by [62,116]

$$f(\xi) = N\exp\left(-2\int^{\xi}(\tilde{N}(\xi')/Q\mathrm{d}\xi') \equiv N\exp(-2V(\xi)/Q\right) \tag{21}$$

provided the boundary conditions are

$$f(\xi) \to 0 \quad \text{for } |\xi| \to \infty\,. \tag{22}$$

In the second Eq. (21), $Q = \text{const.}$ is assumed. N is a normalization constant. A generalization of (18) to continuous variables, $\xi_u(x,t)$, gives rise to a functional Fokker–Planck equation. An explicit solution of the Fokker–Planck equation in the case of several discrete or continuous order parameters can be found if the drift and diffusion coefficients obey the rules of detailed balance [45,47].

Nonequilibrium Phase Transition. Connection with Landau Theory The explicit form of the solution of the Fokker–Planck Eq. (21) allows us to make contact with the theory of phase transitions in the sense of the Landau theory [86] where

$$\begin{aligned} f(\xi) &= N\exp\left(-F(\xi,T)/(kT)\right)\,, \\ F(\xi,T) &= F(0,T) + a(T-T_c)\xi^2 + \frac{\beta}{4}\xi^4\,. \end{aligned} \tag{23}$$

In (21), V corresponds to the free energy F and the noise strength Q corresponds to *absolute* temperature T. T_c is the critical temperature, and (23) refers to a second order phase transition. In case of a first order phase transition, an additional term $\gamma\xi^3$ appears in (23).

An important difference between phase transitions at thermal equilibrium and in the present case of non-equilibrium should be mentioned, however. The decisive constants in the case of non-equilibrium [62] phase transitions are rate constants in contrast to thermodynamic quantities in (23). While non-equilibrium phase transitions described by (21) were experimentally very well verified for instance in the case of lasers [116] (Fig. 1), in the case of thermal equilibrium the Landau theory can not be considered as a good approximation and had been replaced by the concept of critical exponents etc. as dealt with by renormalization group theory [75,146]. For a treatment of the time dependent Fokker–Planck equations see Risken [117].

In a number of cases the drift- and diffusion coefficients of the Fokker–Planck equation are by themselves expectation values, defined on the probability density function so that the Fokker–Planck equation becomes non-linear. For more details see the article by T.D. Frank in this volume.

Instability of a Limit Cycle, $q_0(t)$ [62] The instability is checked by linear stability analysis by means of the hypothesis

$$q(t) = q_0(t) + W(t)\,, \tag{24}$$

where $q_0(t)$ is a time-periodic solution to (1) with $\alpha = \alpha_0$, $W(t)$ a small deviation.

Inserting (24) into (1) with $F \equiv 0$ and linearization leads to an equation of the form (4), where L because of $q_0(t)$ has become also a time-periodic function with the same period as $q_o(t)$. According to Floquet theory [35], the solutions to (4) with periodic $L(t)$ are given by

$$W(t) = \mathrm{e}^{\lambda_f t} v_j(t) \tag{25}$$

(in the case of nondegeneracy), where $v_j(t)$ has the same period as q_0, i. e. L.

Depending on $\mathrm{Re}\,\lambda_j \geq 0$ or < 0 we distinguish between unstable and stable modes (6, 7), respectively. One eigenvalue is $= 0$ and corresponds to an indeterminate phase shift, which in nonlinear analysis is taken care of by a phase $\phi(t)$ that acts as additional order parameter. In order to solve the fully nonlinear and stochastic equations, the hypothesis

$$q(t) = q_0(t + \phi(t)) + \sum_u \xi_u(t) v_u\,(t + \phi(t)) + \sum_s \xi_s(t) v_s\,(t + \phi(t)) \tag{26}$$

is inserted in the Eqs. (1). The subsequent procedure follows the lines outlined above and leads to order parameter equations of the form

$$\dot{\xi}_u = \lambda_u \xi_u + \hat{N}_u\,(\{\xi_u\}\,, \phi) + \hat{F}_u\,(\{\xi_u\}\,, \phi) \tag{27}$$

$$\dot{\phi} = M\,(\{\xi_u\}\,, \phi) + G\,(\{\xi_u\}\,, \phi) \tag{28}$$

where $\hat{N}, \hat{F}, M, G$ are polynominals in $\{\xi_u\}$ and periodic functions of ϕ.

The novelty as compared to the case of an unstable fixed point consists in the introduction of a phase as order parameter.

When noise is neglected, the newly evolving, i. e. bifurcating solutions are either two (or several) limit cycles or tori. Also basically, depending on the system, also a "back bifurcation" to a stable focus can happen.

Instability of Tori [62] The corresponding theory is rather complex so that a few words must suffice here. The basic idea [62] is based on an extension of (24,26) where q_0 is chosen as a quasi periodic function

$$q_0 = q_0(\omega_1 t, \omega_2 t, \ldots, \omega_M t) \tag{29}$$

where the $\omega' s$ must be sufficiently irrational in the sense of the KAM (Kolmogorov [80], Arnold [3], Moser [100]) theorem. Besides amplitudes as order parameters, also phases $\phi_1(t), \ldots, \phi_M(t)$ are introduced. For details cf. [62], and for alternative approaches [21,128].

A Remark on the Method of Solution of Evolution Eqs. (1)

In this article the central role of order parameters is stressed because this allows us to establish profound analogies between quite different systems. In practical applications it may be preferable, however, to apply other methods of solution, analytical, numerical or mixed, in order to derive the spatial, temporal or spatio-temporal patterns. In this way, the Springer Series in Synergetics have developed a "tool box" of models [98].

Quantum Theoretical Formulation

In a quantum theoretical treatment one deals with quantum mechanical Langevin equations which are Heisenberg equations of motion for operators to which pumping

and damping terms as well as random noise sources are added. Here, according to quantum theory, the system's observables are represented by time-dependent quantum mechanical operators, Ω_j. For instance, by the position operator $\hat{x}$ and the momentum operator $\hat{p}$ of a particle, or, in quantum field theory, by creation and annihilation operators $\hat{b}^+, \hat{b}$, respectively. The quantum mechanical Langevin equations read (see, for instance [54,58]):

$$\dot{\Omega}_j = \frac{i}{\hbar}\left[H, \Omega_j\right] + \text{damping} + F_j(t)\,, \tag{30}$$

where H is the Hamilton operator, and $F_j(t)$ are stochastic operators which usually are assumed to be δ-correlated in time. The quantum mechanical properties can be determined by the postulate of quantum mechanical consistency of Ω_j, (cf. [54], appendix).

If the non-commutativity of operators is taken care of, the procedure to derive order parameter equations is formally the same as in the case of classical Langevin equations as indicated above. The Fokker–Planck equation, however, must be replaced by a density matrix equation, originally introduced as master equation [109]. For nonequilibrium systems, such as the laser, see [50,126,144], also [54,119]. Using methods of quantum classical correspondence, this density matrix equation can be converted into a Fokker–Planck equation under specific conditions. The basic idea is this:

Quantum-Classical Correspondence

There are several ways to define quantum classical correspondence. In the case of position operator $\hat{x}$ and momentum operator $\hat{p}$ with the commutator $[\hat{p}, \hat{x}] = \frac{\hbar}{i}$ and the density matrix ρ, the Wigner distribution function $W(x, p)$ [145] is defined by

$$W(x,p) = \frac{1}{(2\pi)^2}\cdot\int\int_{-\infty}^{\infty} \mathrm{e}^{-ikx-ilp} \cdot \mathrm{tr}\left(\mathrm{e}^{ik\hat{x}+il\hat{p}}\rho\right)\mathrm{d}k\mathrm{d}l \tag{31}$$

where "tr" means trace.

Thus a relation is established between the quantum mechanical density matrix and a classical quasi-density $W(x, p)$. Based on (31) or (34,35,36), a density matrix equation can be converted into a generalized Fokker–Planck equation [54].

By the transformation of $\hat{x}, \hat{p}$ to creation and annihilation operators b^+, b by means of

$$\hat{b}^+ = \frac{1}{\sqrt{2\hbar}}(\hat{x} + i\hat{p}) \tag{32}$$

$$\hat{b} = \frac{1}{\sqrt{2\hbar}}(\hat{x} - i\hat{p}) \tag{33}$$

an alternative form to (31) is given by

$$P(\beta, \beta^*) = \frac{1}{\pi^2}\int\int_{-\infty}^{\infty} \mathrm{e}^{-i\beta k - i\beta^* l}\cdot\mathrm{tr}(\mathrm{e}^{ik\hat{b}^+ + il\hat{b}}\rho)\mathrm{d}k\mathrm{d}l\,. \tag{34}$$

Because $\hat{b}^+, \hat{b}$ are noncommuting operators, $[\hat{b}^+, \hat{b}] = 1$, different "quasiprobability" distributions P result, if

$$\mathrm{e}^{ik\hat{b}^+ + il\hat{b}}$$

is replaced by

$$\mathrm{e}^{ik\hat{b}^+}\mathrm{e}^{il\hat{b}} \tag{35}$$

or

$$\mathrm{e}^{ik\hat{b}}\mathrm{e}^{il\hat{b}^+}\,. \tag{36}$$

(35) gives rise to the Glauber–Sudarshan representation. For details and references see [54].

Regular Spatial and Spatio-Temporal Patterns

One of the most striking features of nonequilibrium systems in physics, chemistry and biology is their capability of forming (more or less) regular spatial pattern (for explicit examples see below). (There is a rich literature on pattern formation in physics, especially fluids [20,24,92,136], but also semiconductors [125] and nonlinear optics [133], chemistry [28,33,83] and biology [7,95,96,102] and general [69,70,98,104,110,111,115,140]. Furthermore, the patterns exhibit striking similarities in spite of the fact that the individual parts are quite different. The methodology of Synergetics (e. g. [62]) provides us with a basic insight into the causes of such analogies.

Pattern formation is determined by at least three causes:

1. internal mechanism, such as e. g. the interplay between reactions and diffusion in large scale chemical processes,
2. the influence of boundaries,
3. initial conditions.

Concerning 1) and 2) between two (limiting) cases can be distinguished.

1. dimensions of the internally evolving patterns are of the same or larger order as those of the boundaries. Here a strong influence of the boundaries must be expected.

2. dimensions of evolving patterns are small compared to those of the boundaries (boundaries $\to \infty$).

To bring out the essential features we consider that originally for a control parameter value α_0 the system is homogeneous and quiescent. The approach can, however, be extended to a space dependent reference state (which, e. g. resulted from a first bifurcation leading to $q_0 = q_0(x)$) and the cases of a limit cycle or torus. The space may be 1, 2 or three dimensional Euclidian or, e. g., a 2 or 3 sphere.

Infinite Boundaries

We start with 2) infinite boundaries, the medium is homogeneous and isotropic. We assume a continuous transition from the homogeneous to the "bifurcating" state. The evolving patterns are determined by the leading terms in (8) that we call the "mode skeleton"

$$q(x,t) = q_0 + \sum \xi_u(t) v_u(x) \tag{37}$$

and the order parameter Eq. (12). The functions $v_u(x)$ are *the space-dependent part of the* solutions to (4) where L is a differential (or integral) operator which is invariant against translation and rotation. Thus, e. g., L commutes with the displacement operator

$$\Omega_a : \; x \to x + a \,, \quad a \text{ constant vector} \,.$$

Thus $v_u(x)$ can be chosen as eigenfunction to Ω_a,

$$\Omega_a v_u(x) = \Lambda v_u(x) \tag{38}$$

with

$$v_u = \mathrm{e}^{ikx} \tag{39}$$

$$\Lambda = \mathrm{e}^{ika} \tag{40}$$

i. e. plane waves. Which waves must be considered in (37) is determined by λ_u in (6) as well as by the order parameter Eq. (12).

The condition $\mathrm{Re}\,\lambda_u(k) = 0$ defines $k = k_{\mathrm{crit}}$. As was shown by means of many examples $k \approx k_{\mathrm{crit}} \neq 0$. If the boundaries are finite, such a discrete k must be chosen which comes closest to k_{crit}. If the boundaries tend to infinity, a continuous set k is taken care of by (generalized) Ginzburg–Landau equation (see above). If the boundaries are "narrow" in 1 or 2 dimensions, but large in the remaining dimensions, the wave vector k must be split into k_{II} and $k_\perp$ where k_{II} is practically continuous and $k_\perp$ discrete. Quite often only one $k_\perp$ (the most critical) needs to be considered. This leads to practically 2 (or 1) dimensional patterns connected with k_{II}. In the 2-dimensional case, the modes with $|k_{\mathrm{II}}| = k_{\mathrm{crit.}}$ are degenerate. This degeneracy can be lifted by a weak influence of boundaries (leading to roll patterns), by specific initial condition which (by chance) prefers a specific roll pattern, or by terms in the order parameter-equations that lead to specific combinations, e. g.

$$k_1 + k_2 + k_3 = 0 \,, \tag{41}$$

where k_j, $j = 1, 2, 3$ belong to k_{II}.

This gives rise to the formation of hexagons. This is the case if the leading term of $\tilde{N}$ contains

$$\int v_{k_1} v_{k_2} v_{k_3} \mathrm{d}^2 x \neq 0 \,. \tag{42}$$

In three dimensions this mechanism may lead to plane wave fronts stabilizing each other which gives rise to icosaeders, as observed in diatomea.

An important class of spatio-temporal patterns (in 2 dimensions) results when the system utilizes *rotation symmetry*. This can best be explained by the following example:

In many cases of practical interest, N in (1) and thus L in (4) contain the Laplace operator Δ. When written in planar polar coordinates r, ϑ, solutions to (4) are of the general form

$$v \propto \mathrm{e}^{i(m\vartheta - kr - \omega t)} \tag{43}$$

(times a rotation symmetric function $g(r)$) which represents *spirals*. $m = 0$ represents concentric rings, while an integer $m > 0$ represents the number of spiral arms. $\omega = 0$ represents standing spirals, $\omega \neq 0$ rotating spirals.

The mode skeleton (37) is composed of functions of the form (43). Which of the functions (43) appear in (37) depends on the competition Eqs. (12) for order parameters, which may also allow for a super position of counter rotating spirals (such as in the sunflower head). As group theory shows (see below), solutions (43) with different m's belong to different irreducible representations, and do not coexist in (37). This does not exclude the coexistence of differently rotating spirals in *different* regions of space, however.

The above results can be cast into the isomorphy principle:

While the "true" q is represented by (we omit the homogeneous q_0)

$$q = \sum_u \xi_u v_u(x) + \text{enslaved modes, with same symmetry.} \tag{44}$$

and v_u "true modes", its symmetry features can be replaced by a "representative" q':

$$q' = \sum_k \xi_k R_k(x) \,, \tag{45}$$

where R_k represent the "elementary" functions showing the symmetry under consideration. While the material significance and explicit form of q according to (44) may be quite different for different material substrates, q' (45) shows the *same* patterns for different systems.

These results can be deepened by invoking group theory, in which also the effect of the boundaries is taken into account.

Symmetries, Group Theory, Representation Theory, Finite Boundaries

Consider a set of transformations G_j of space variables $x \to x'$ so that

$$G_j q \to q' \tag{46}$$

Example 1 G_j induces the translation

$$x \to x + a \quad \text{so that} \quad G_j q(x) = q(x + a) \,. \tag{47}$$

The transformations must be so that they are compatible with the internal properties of the system (1) and the boundary conditions. Example: when dealing with a problem on a 2-dimensional sphere, the transformed coordinates x must not leave the sphere.

Because of the symmetry of the problem, the transformations G_j form a group defined by

1. existence of unity E such

$$G_j E = G_j \quad \text{for all } j \tag{48}$$

2. the product of two group elements is again an element of the group,

$$G_j G_k = G_l \quad \text{for all } j, k \tag{49}$$

3. existence of an inverse G_j^{-1} for all j so that

$$G_j^{-1} G_j = E \,, \tag{50}$$

4. associative law

$$(G_k G_l) G_j = G_k (G_l G_j) \tag{51}$$

for all group elements.

In the following we first ignore random forces, i. e. we consider (1) with $F \equiv 0$.

$$\dot{q}(x, t) = N(q, \Delta, \alpha) \,. \tag{52}$$

Jointly with the boundary conditions, (52) defines a function space S in which all functions to be considered must lie (i. e. can be represented by linear combinations of a complete set of (vector valued) basic functions of S; example: S is a Hilbert space)

Definition 1 The system is invariant against G_j if for all $f \varepsilon S$

$$G_j N\left(G_j^{-1} f\right) = N(f) \,. \tag{53}$$

Example 2

$$G_j \colon\ x \to x + a \,, \tag{54}$$

$$N(f) = \Delta f + V(x) f + f^2 \,. \tag{55}$$

Then

$$G_j \cdot N\left(G_j^{-1} f\right) = \Delta G_j^{-1} f(x+a) + V(x+a) G_j^{-1} f(x+a) + \left(G_j^{-1} f(x+a)\right)^2 \tag{56}$$

$$= \Delta f(x) + V(x+a) f(x) + f(x)^2 \tag{57}$$

$$\neq N(f) = \Delta f(x) + V(x) f(x) + f(x)^2 \tag{58}$$

unless $V(x + a) = V(x)$. If a in (54) is arbitrary, N is not invariant against (54).

Application of G_j to q in (52) leads to

$$\frac{\mathrm{d}}{\mathrm{d}t}(G_j q) = N(G_j q) \tag{59}$$

or because of (53), (with $f = G_j q$), to

$$\frac{\mathrm{d}}{\mathrm{d}t}(G_j q) = G_j N(q) \,. \tag{60}$$

In the spirit of representation theory of groups the action of G_j on f can be understood as an abstract operation, but also as a matrix acting on the vector f in S- space.

By appropriate transformation of basis of q, and using the representation theory of symmetry groups, all matrices U_j belonging to all group elements j can simultaneously be decomposed into "irreducible" representations so that (in the example of 3 irreducible representations)

$$U_j = \begin{pmatrix} \square & \bigcirc & \bigcirc \\ \bigcirc & \square & \bigcirc \\ \bigcirc & \bigcirc & \square \end{pmatrix} \tag{61}$$

Each box □ is a matrix $U_j^{(k)}$ with dimension Dk, so that

$$D1 + D2 + \cdots + Dk = \text{dimension } U_j \,.$$

Example 3 Rotation group applied to 2-sphere (e. g. earth surface). Basis functions are spherical harmonics Y_m^l with "quantum numbers" l, m. Subspace l fixed, $m = 0, \ldots, l-1$. As a consequence, the mode skeleton reduces to (q_0 dropped)

$$q^l = \sum_m \xi_m(t) Y_m^l \,. \tag{62}$$

There is no coupling between different ls, which implies a low dimensional dynamics of ξ_m.

Generally, the original function space S is decomposed into subspaces forming the basis of each irreducible representation. This implies a symmetry reduction beyond bifurcation point, compared to the situation below bifurcation point, where

$$G_j q = q \quad \text{for all } j \,, \tag{63}$$

i. e. q fully symmetric under G.

In our example beyond the bifurcation point q is given by q^l where Y_m^l transforms according to the subgroup G^l, which leaves the space spanned by Y_m^l invariant. If, however, group elements not belonging to G^l are applied to q^l, this space is left. In other words, q^l is connected with a lower symmetry than q (63). By bifurcations, the symmetry of q is lowered and one speaks of "symmetry breaking instability". If fluctuating forces in (1), i. e. in (52) are taken into account, the full symmetry can be restored (under specific conditions on the fluctuating forces).

While group theory has found important and widespread applications to quantum theory, it is less frequently used in problems of Synergetics, though there it may lead to deep insights as pointed out above. (For an in-depth approach see [42,43,120].)

On top of, or jointly with, regular patterns, a variety of defects as well as boundaries between different patterns may occur (cf. contribution by Pismen, this volume and [110,111]).

A Further Mathematical Tool: Shannon Information and the Maximum (Information) Entropy Principle

While evolution equations are the backbone of Synergetics, also other tools are invoked to deal with complex systems. Such a tool is Shannon information [129] which is defined by

$$i = -\sum_j p_j \log_2 p_j \tag{64}$$

where p_j is the relative frequency of the event j or, in a different interpretation, the probability of finding the realization j in an experiment. The maximum (information) entropy principle as formulated by Jaynes [73,74], for an earlier proposal see [27]), allows one to make unbiased guesses on systems on which only incomplete data are known by maximizing the informations, i. e. (64) = max! or = extremum! under given constraints.

A simple example is provided by a gas composed of N particles, where the total kinetic energy $\mathrm{E}_{\mathrm{kin}}^{\mathrm{tot}}$ is fixed. Denoting the kinetic energy of a particle with mass m and velocity v_i by $f_i = (m/2)v_i^2$, the mean kinetic energy per particle is

$$\sum_i p_i f_i = \mathrm{E}_{\mathrm{kin}}^{\mathrm{tot}}/N \tag{65}$$

To fix p_i, (64) must be maximized under the normalization condition

$$\sum_i p_i = 1 \tag{66}$$

and the constraint (65).

Using Lagrange multipliers, λ, λ_1, the result reads

$$p_i = \exp\left(-\lambda - \lambda_1 m v_i^2/2\right) \tag{67}$$

i. e. the Maxwell–Boltzmann distribution function. Also relations between the Lagrange multipliers λ, λ_1 can be established which, evidently, have fundamental thermodynamic significance.

This approach has been extended to the treatment of nonequilibrium phase transitions, i. e. determination of order parameters, enslaved modes and emerging patterns [60]. The crucial idea consists in the proper choice of constraints, as which the moments of the variables q_i are chosen:

$< \ldots >$ means average over the joint distribution function $f(q_1, q_2, \ldots, q_n)$ which replaces p_j and the vector $(q_1, \ldots, q_N)$ replaces j. The variables q_j may be discrete or continuous.

$$f_i = < q_i >, \quad i = 1, 2, \ldots, N \,. \tag{68}$$

$$f_{ij} = < q_i q_j > \,. \tag{69}$$

$$f_{ijkl} = < q_i q_j q_k q_l >, \quad i, j, k, l = 1, 2, \ldots, N \,. \tag{70}$$

The resulting distribution function is given by

$$q = \exp V(\lambda, q) \tag{71}$$

with

$$V(\lambda, q) = \lambda + \sum_i \lambda_i q_i + \cdots + \sum_{ijkl} \lambda_{ijkl} q_i q_j q_k q_l \,. \quad (72)$$

(71) is a starting point to make contact with the Landau or Ginzburg–Landau theory of phase transitions [86], and to guessing Fokker–Planck equations. The approach allows one to calculate the efficiency of self-organizing systems close to their instability points.

The method has been extended to the "unbiased modeling" of stochastic processes: how to guess path integrals, Fokker–Planck equations and Langevin-Îto equations [60]. The central quantity to be searched for is the probability density P_n of paths.

Let $q(t)$ be the state vector $q = (q_1, \ldots, q_N)$ at time t, then

$$P_n(t_n, t_{n-1}, \ldots, t_0) = P_n\big(q(t_n), t_n; q(t_{n-1}), t_{n-1}; \ldots; q(t_0), t_0\big) \,, \quad t_n > t_{n-1} > \ldots > t_0 \,. \quad (73)$$

This task is simplified if the Markov hypothesis on the process holds, i. e.

$$P_n(t_n, t_{n-1}, \ldots, t_0) = \hat{P}\left(q(t_n), t_n \middle| q(t_{n-1}) t_{n-1}\right) \cdot P_{n-1} \quad (74)$$

where $\hat{P}$ is the transition probability so that only transition probabilities between subsequent states (with $\Delta t \to 0$) must be guessed in addition to P_0. In the frame of the present approach, this task is fulfilled by use of the maximum information principle. The constraints to be used are essentially conditional first order moments and two-time correlation functions of the state vectors $q(t), q(t')$.

Phenomenological Synergetics

In many fields of science, including medicine, the microscopic variables and their dynamics are not well-known or not known at all. Nevertheless, in quite a number of cases, namely where dramatic macroscopic changes of the system's behavior take place, general insights, gained by Synergetics, can be invoked. A paradigm for this procedure is the modeling of Kelso's finger experiments [77,78] (Fig. 3). He instructed subjects to move their index fingers in parallel which was accordingly performed. However, when the speed of the fingers was increased, the parallel movement was replaced by a symmetric movement quite involuntarily and spontaneously. In other words, a transition from a parallel to an anti-parallel phase takes place. In terms of Synergetics, the interpretation is simple: the control parameter consists in the prescribed frequency ω of the finger movement, whereas the macroscopic quantity, i. e. the order parameter that changes dramatically is provided by the relative phase of the two index fingers. According to the experience made in Synergetics, the order parameter, here called ϕ obeys a typical order parameter equations of the form [64]

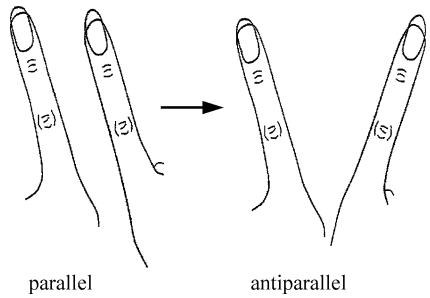

Synergetics: Basic Concepts, Figure 3
Transition between finger movements from parallel to symmetric in Kelso's experiment [64]

$$\dot{\phi} = -\frac{\partial V}{\partial \phi} + F(t) \,, \quad (75)$$

where $V(\phi, \omega)$is a potential function and F a fluctuating force. When the control parameter ω is changed, the potential runs through a series of forms as depicted in Fig. 4. As was shown in detail, at a critical value of ω, the transition from one potential minimum to another one occurs, as related to the change of the kind of finger movement. The mathematical analysis shows hysteresis, critical slowing down and critical fluctuations [59] which reject the idea that the brain acts like a computer via a motor program but rather via self-organization.

Another application is made by the Synergetic computer [61] (Figs. 5, 6), where to each pattern to be recognized a specific order parameter is attached. Pattern recognition is then achieved via a competition between order parameters. The competition equations are given by

$$\dot{\xi}_k = \frac{\partial V}{\partial \xi_k} V(\xi_1, \ldots, \xi_M) = -\frac{1}{2} \sum_k \lambda_k \xi_k^2 + \beta \sum_{k,k'} \xi_k^2 \xi_{k'}^2 - C \sum_k \xi_k^4 \,. \quad (76)$$

This approach may serve also for modeling of brain functions: both recognition as well as movements are governed by the establishing of order parameters which may wander from one quasi attractor to another one. Quasi attractors are defined as attractors that vanish after the task has been accomplished, e. g. after a pattern has been recognized or movement performed.

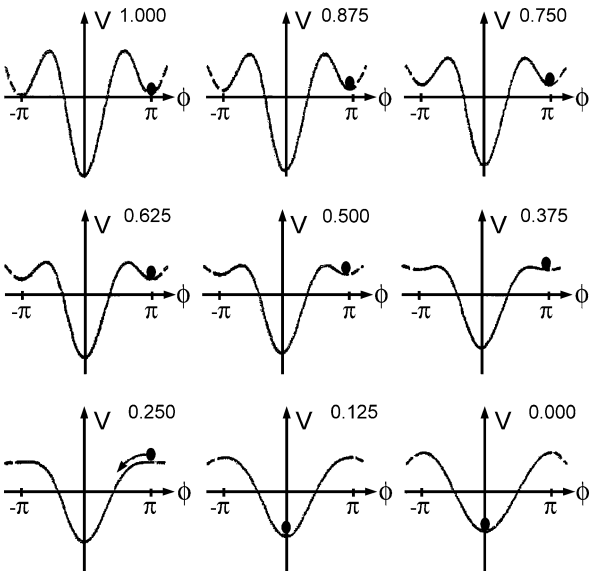

Synergetics: Basic Concepts, Figure 4
Sequence of potential curves of the Haken–Kelso–Bunz model of Kelso's experiment [64]

Based on the concept of order parameters, a learning procedure for Synergetic computers has been developed [61]. Here the number of patterns to be recognized is prescribed and then a special functional must be minimized. In the case of the Synergetic computer, it is possible to make contact between the microscopic and the mesoscopic description, i. e. the microscopic variables are pixel values q_j, j pixel index, whereas the mesoscopic (or macroscopic) quantities are the order parameters ξ_k.

The relation between ξ_k, q_j is given by

$$\xi_k = \sum_j v_j^{k+} q_j \,, \tag{77}$$

where v_j^{k+} are adjoint prototype patterns, with k pattern index, j pixel index.

The relation between prototype patterns v_j^k and their adjoints is given by

$$\sum_j v_j^{k+} v_j^{k'} = \delta_{kk'} \tag{78}$$

At the phenomenological level the order parameter concept allows us to interpret and model complex movement patterns, e. g. learning to ride on a pedalo [59]. In the experiments, LED's are fixed at the joints of the subject and their positions measured which gives rise to a series of time-dependent tracks. Then, in a first step, a principle component analysis is performed, in the next step, by means of a variational principle, the best fit is searched in

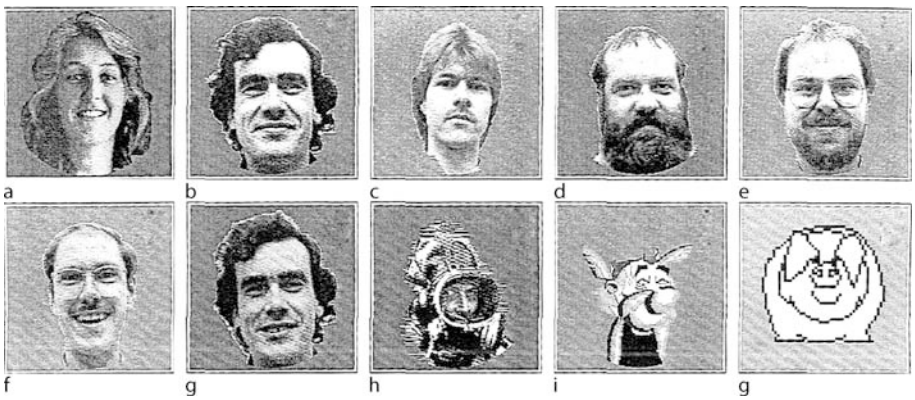

Synergetics: Basic Concepts, Figure 5
Recognition of faces by the synergetic computer: stored or learned prototype patterns [61]

Synergetics: Basic Concepts, Figure 6
Pattern recognition by the synergetic computer: recognition of a specific face of which initially only a subset of pixels is presented [61]

Synergetics: Basic Concepts, Figure 7
Example of an ambivalent figure: young / or old woman? [34]

terms of order parameters and their equations of motion, in order to mimic the actual tracks. While in the learning phase several order parameters are needed, at the end the whole movement is governed by a rather simple equation for a complex order parameter.

During the development of Synergetics it turned out that there are strong relations to gestalt theory [85] as well as to psycho physics. A typical example is provided by ambivalent figures where (Fig. 7) [34] shows an example. An observer may either perceive a young woman or an old woman, but not both simultaneously, rather the perception switches between these two percepts. In the mathematical modeling to each percept an order parameter is attached [61], which obeys the typical equations of Synergetics. The control parameter invoked here is attention. According to an early suggestion by Wolfgang Köhler [85], when a pattern is recognized, the corresponding attention fades away. This has been modeled mathematically based on a competition dynamics between two order parameters, when the control parameter (attention) of one pattern fades away, the other pattern gets the possibility of being perceived. Then in the next step the corresponding attention parameter fades away and the first pattern may re-appear (Fig. 8) [61]. This model describes details of the observed phenomena, such as the dependence of the duration of the perception of one face as compared to that of the other face, dependent on the bias which face is recognized first. Also, one may distinguish between slow, medium and fast observers, depending on the individual parameters.

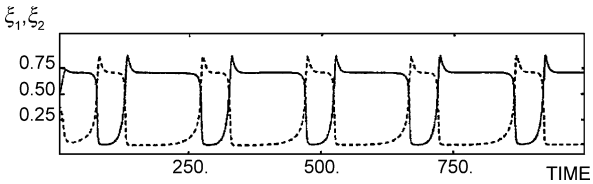

Synergetics: Basic Concepts, Figure 8
Order parameter oscillations belonging to the recognition young woman / old woman with bias towards the young woman [59]

Quite generally, order parameters may have properties of gestalt in the sense that they are invariant against size, orientation and perception of objects in space.

In medicine, a syndrome has the characteristic features of an order parameter. On the one hand it is generated by the co-operation, or at least by the simultaneous presence of specific features, on the other hand once the syndrome (order parameter) is established, it acts on the individual parts of the system, where the slaving principle induces specific phenomena at the level of individual parts. Clearly, the concept of circular causality plays an important role here. It shows that the syndrome, at least in general, can not be cured by curing an individual symptom, but rather by curing a decisive majority of individual causes.

Semantic Synergetics

In soft sciences, but also in medicine and other fields, a mathematical modeling, even at the level of order parameters may not be possible. Nevertheless, Synergetics may provide us with qualitative insights into basic mechanisms. In psychology and psychiatry [122], quite often specific mental states can be ascribed to a patient. For instance in bipolar patients a depressive phase or a manic phase may appear or in depressive patients a normal phase and a depressive phase. Another example is provided by patients with a compulsory action. In the spirit of Synergetics, as a theory of indirect control, one may ask, whether there are appropriate control parameters by means of which the behavior of a person can be changed. Let the two states be represented by the positions of a ball in a landscape with two valleys. In this situation, direct control means to push the ball from the unwanted position to the wanted. Indirect control means to lower the potential hill between the two valleys so that the wanted transition may occur via self-organization. This may happen through interventions used in cognitive psychology, a change of environmental conditions, or/and by specific medication. The central issue here is that the patient is not directly influenced, e. g. by saying you must do this or that, but rather by a soft

changing of his/her point of view. A number of successes have been reported about this method which is, to some extent, well known in psychiatry, but finds here a scientific theoretical basis. For more details see the article by G. Schiepek and V. Perlitz, this section, and in a somewhat related form [66].

Some Selected Examples

The study of nonlinear, self-sustained oscillations [1,2,13] be it in radio-engineering, mechanics or other fields, has a long tradition. In the context of bifurcation theory, their origin was unearthed by Hopf [67,68].

Nonlinear optics [99] and, when quantum effects are important, quantum optics [57,97,123,142] provide us with a wealth of phenomena, in particular of the formation of coherent oscillations. A device, closely related to the laser, is the parametric oscillator [44], in which, within a nonlinear crystal, incoming pumplight is split into a signal and an idler. Then, similar to the laser light, the signal light becomes amplified, and its generation can be described as that of a nonlinear quantum-mechanical oscillator. Fluid dynamics is rich of pattern formations (including chaos) [12,16,17,18,31,41,89,92,105,115,118,127,135,136], to mention just a few. In a fluid heated uniformly from below, with increasing temperature difference, several instabilities may occur for instance giving rise to stationary patterns, such as rolls, hexagons (Fig. 9) or squares. In the next step the rolls may start to show oscillations, and still more complex patterns may occur (Fig. 3). In the case of the Taylor instability [137], a liquid is placed in between two coaxial cylinders, where the outer one is rotating. With increasing rotation speed, a hierarchy of instabilities is reached, first the formation of roles, then oscillating rolls at one frequency, then oscillation of rolls at two frequencies, and finally weak turbulence, i. e. chaos occurs (Fig. 10) [31,93]. Important phenomena are the establishing of boundaries and of defects as described in the article by Pismen [110,111] and other articles of this Encyclopedia. A rich variety of pattern formation may occur in semiconductors [125], where electrons and holes as well as currents form specific spatio-temporal patterns. In meteorology, atmospheric convection patterns and other instabilities are treated [38]. In chemistry, oscillations and large scale patterns arise by means of the interplay of chemical reactions and diffusions [9,15,28,32,33,149], e. g. concentric ring patterns, each starting from a center, which then annihilate each other when colliding. An important class is provided by spiral patterns which may have one to several arms (Fig. 11). In biology, specific models on morphogenesis were treated, such as the formation of stripe or spot

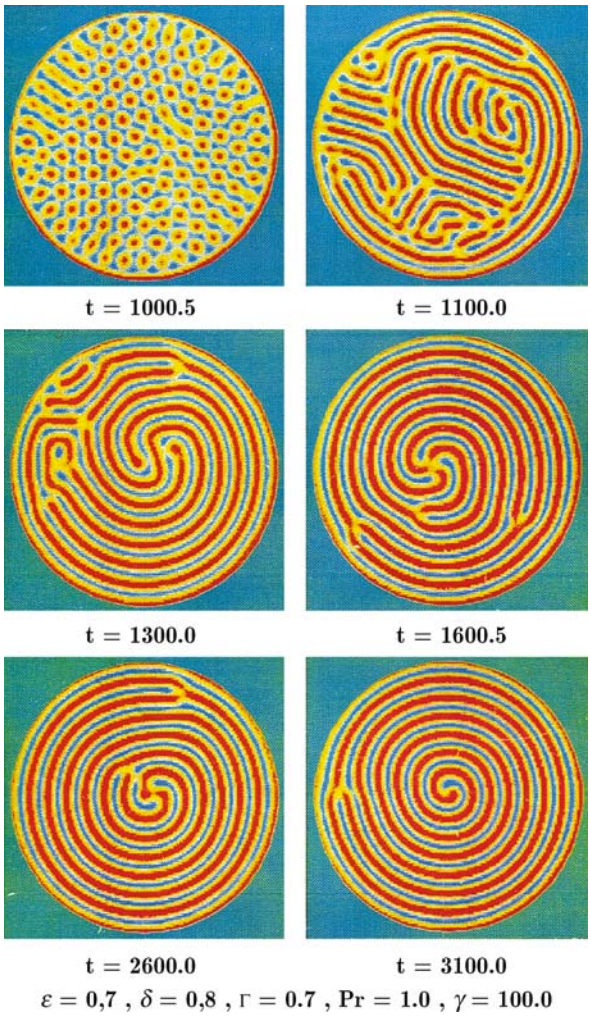

Synergetics: Basic Concepts, Figure 9
Model calculation of the motion of a fluid in a circular pan uniformly heated from below (after [29]). *Upper left corner*: above a critical temperature difference between lower and upper surface of the fluid layer, a hexagonal pattern appears. If the boundary is also heated uniformly, a transition to the spiral pattern with one or several arms can be found (*lower right corner*)

patterns on animal furs or skins of fish (Fig. 12) or still more complicated patterns on sea shells [39,62,95,96,102]. The basic idea which can be traced back to Turing [139] is this: originally unspecialized cells produce activator and inhibitor molecules which by reaction and diffusion form a prepattern, a morphogenetic field [147]. At positions of high activator concentration, genes are switched on which then leads to cell differention producing e. g. pigments. In aggregating slime mold, spiral or concentric ring patterns are observed [14,37]. Mathematical models on prebiotic evolution [26] study the competition between species of

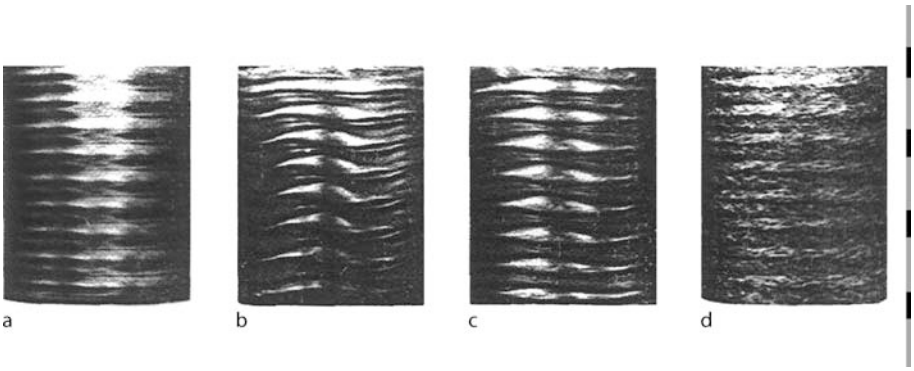

Synergetics: Basic Concepts, Figure 10
Pattern hierarchy in the Taylor–Couette instability. A fluid in between two vertical coaxial cylinders of which the outer one rotates, shows no macroscopic movement pattern, if the movement of the outer cylinder is slow. When the rotation speed is increased, first a role pattern appears in which the fluid moves outwards at one height, and then inwards at another height. This movement pattern is periodic with respect to height [137]. At a further critical rotation speed, the pattern shows oscillations which at a further speed transform into a motion with two frequencies until eventually chaotic motion appears. The experiments were done by [31], the modeling was done for the first transition (homogeneous to roles) and especially the second transition (roles to oscillating roles) by [93]

Synergetics: Basic Concepts, Figure 11
Belousov–Shabotinsky reaction: the occurrence of spirals (courtesy A.T. Winfree). They may show one to several arms. The centers of the spirals may occur at different positions. Spirals hitting each other, annihilate each other

Synergetics: Basic Concepts, Figure 12
Stripe pattern on a tropical fish

biomolecules and the "survival" of the fittest, where pronounced analogies with the dynamics of laser photons can be unearthed, fully in line with Synergetics [62]. In the understanding of brain function, for instance, steering of movements, pattern recognition or decision making, the reduction of degrees of freedom of the numerous neurons to few order parameters is central [59].

The concepts and principles of Synergetics shed new light on important relationships in economy, such as cooperation and competition between companies, the important role of indirect steering by means of control parameters, such as taxes, interest rates. It can be shown, that a fusion of companies does not necessarily lead to so called synergy effects, but rather critically depends on initial conditions and details of the cooperation between the previously separated firms. Important insights are also gained into fundamental processes of climatology, as well as in ecology such as the by now well-known and publicly discussed effects that even small concentrations of chemicals in the atmosphere can change the climate dramatically. The same is true for lakes, in which beyond a critical pollution, fish population dies out entirely.

In this way, the numerous examples collected in the field of Synergetics, provide not only scientists but also the public with impressive examples of dramatic changes (in-

stabilities) provoked by even a slight change of control parameters. Clearly, an important research subject of Synergetics is a detailed study of which control parameters are critical and to which control parameters a system is rather insensitive. Sociology is an important field for the application of stochastic models [8]. In particular, basic concepts of Synergetics have proven useful in the developing field of sociodynamics, where e. g. phase transition-like phenomena may occur [143].

History and Relations to Other Fields

The term Synergetics was coined by H. Haken in 1969 in a lecture at University of Stuttgart. A first description of the goals of this field was given by H. Haken and R. Graham in 1971 [63] where the unifying role of the concept of order parameters is outlined. A relationship exists to the general system theory due to von Bertalanffi [10], which also aims at the exploration of analogies between different systems, but on the level of the individual elements rather than on the level of order parameters. Von Bertalanffi coined the term flux equilibrium (Fließgleichgewicht) in order to characterize homeostasis in active systems [11]. A general mathematical frame for Synergetics is provided by dynamic systems theory (see, for instance, [49]) which, however, in the traditional approach ignores stochastic processes (mainly chance events) which are also of great relevance for Synergetics. Here the theory of Markov processes with their typical equations, such as Langevin equations, Fokker–Planck equations, Chapman–Kolmogorov equations, the Kramers–Moyal expansion etc. is important (see for instance [134] and ► Linear and Non-linear Fokker–Planck Equations by T. Frank).

A basic feature of Synergetics consists in dealing with nonlinearities in complex systems and studying, mainly quantitatively, qualitative changes at macroscopic scales. Qualitative changes of systems at macroscopic levels are studied also by catastrophe theory [5,138], which may be interpreted as a study of the surfaces of equilibrium points of few order parameters, where different cases are classified according to the (low) number of control and order parameters. Chaos theory studies the mostly irregular dynamics of deterministic low dimensional continuous [89,106,118,132] or discrete dynamic systems [23,30,48,94,130], where the behavior is mainly characterized by so called Lyapounov exponents, various kinds of fractal dimensions and chaotic attractors. The slaving principle of Synergetics provides a basis for an application of chaos theory to multi-component systems in that Synergetics shows the possibility of reducing the degrees of freedom. Synergetics shares some of its topics with singularity theory [4,42,43], which applies to bifurcation points and their surrounding. Another point of contact is bifurcation theory (see the quotations in previous chapters), in which the branching of solutions of the dynamic system close to instability points is studied. The term dissipative structure was coined by Prigogine [40] to characterize evolving structures in systems away from thermal equilibrium where as in all such non-equilibrium systems dissipation occurs. A typical example is that of the convection instability. Prigogine tried to base his approach on thermodynamics, introducing concepts of entropy production and excess entropy production. As we now know, these concepts are, however, insufficient to deal with structure formation in such systems [87]. Based on a fundamental idea of A. Turing [139], Prigogine and Nicolis [114], see also [108], treated macroscopic pattern formation in a specific chemical reaction model. For more recent work see [107].

Because of the fundamental importance of thermodynamics, we elucidate its relationship to Synergetics more closely.

Thermodynamics (see for instance [19]) deals with systems in and out of thermal equilibrium. A central concept is entropy. In a closed system, it tends to its maximum value. Thermal equilibrium is characterized by the equipartition theorem: each degree of freedom has an average energy of $1/2\ kT,\ k =$ Boltzmann constant, T absolute temperature. This may refer e. g. to gas atoms as well as to collective excitations in crystals. These systems are in thermal equilibrium with their surrounding (heatbaths, reservoirs).

Irreversible thermodynamics [51] treats systems which are not in thermal equilibrium but close to it. It mainly deals with transport and relaxation processes. A central concept is entropy production.

In the domains of physics, chemistry, biology, Synergetics deals with systems far from (thermal) equilibrium . This state is caused and maintained by an in- and outflow of matter, energy and / or information. This is achieved by a coupling of the "proper system" to heat baths (reservoirs) at different temperatures. The former concepts of thermodynamics, in particular the first and second law, are still valid for the total system ("proper" plus reservoirs), but no more sufficient to deal with the kinetics of the proper system. Now the central concept is growth and decay rates. In systems far from thermal equilibrium, collective modes are formed. One or several of them compete best for the external supply of matter, energy, information and grow at the expense of all other degrees of freedom (or modes). Thus the equipartion theorem is no more valid. In general, the behavior of the system is governed by few degrees of freedom (order parameters). Inci-

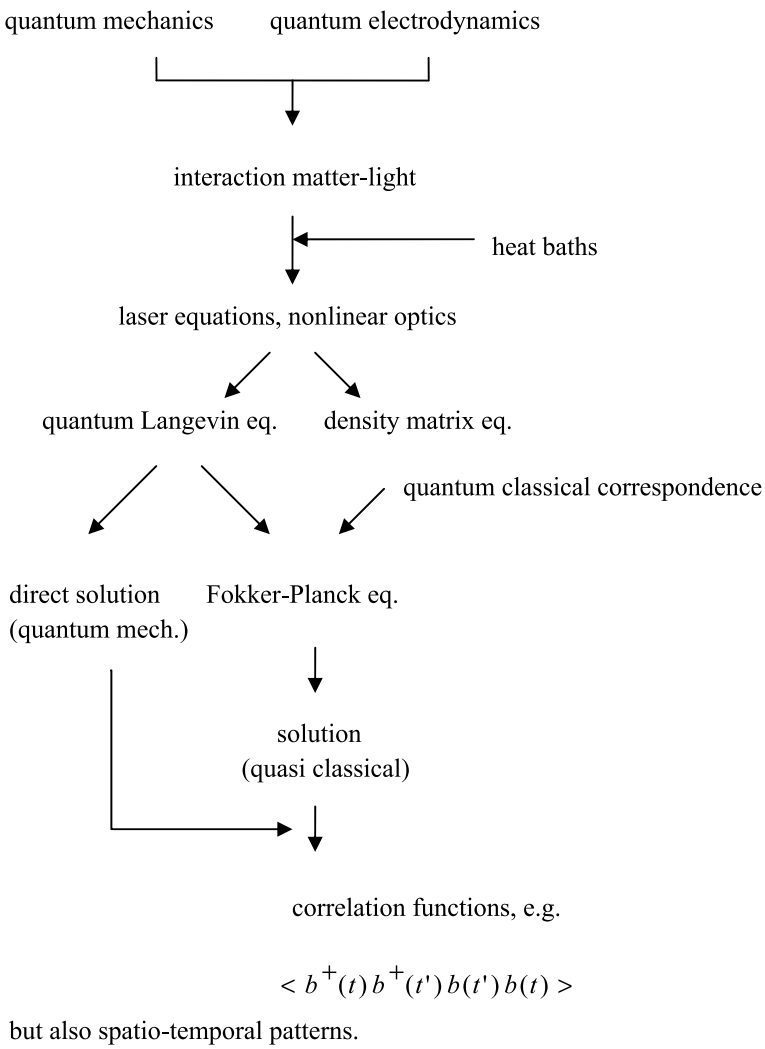

Synergetics: Basic Concepts, Scheme 1
Quantum optics, example laser (after [54])

dentally, this "growth and competition" principle applies to a great variety of fields out of physics, chemistry and biology, where "modes" may not only be special physical structures, but may mean behavioral patterns, special functions etc. Quite often, a "mode" is initiated by a chance event (fluctuation). Clearly, a generalized Darwinian principle can be seen: The interplay between *mutations* (microscopic chance events) and *selection* (competition between mascropic modes) leads to macroscopic patterns (structures) in the widest sense of the word.

In present days research, a new name is spreading, namely complexity or complexity theory. There seems to be no precise definition of this field available in the scientific community. Of what is known so far, we may conclude that this field has strong ties to the original field of Synergetics in that it searches also for general principles but, in addition, it allows the collection or accumulation of knowledge on all kinds of complex systems, as is witnessed in the excellent Complexity Digest, weekly edited by Gottfried Mayer. What "complexity" eventually might mean is reflected by the present encyclopedia.

Future Directions

Synergetics is surely not a closed scientific discipline but quite open to further research. On the one hand we may think of further applications of the principles of Synergetics that have been hitherto elaborated on, such as order parameters etc. Here a wide field of application is provided by robotics, construction of prostheses, automatic steering

of cars etc. On the other hand, new ideas to endow systems with self-organizing properties are needed, e. g. groups of mobile agents for the execution of specific tasks. First steps have been done for instance by Kornienko [81].

Bibliography

Basically, the individual volumes of the Springer Series in Synergetics and selected volumes of the Springer Series on Complexity provide further information. Here, because of lack of space, only a few basic papers can be quoted.

The number of books, reviews and original papers on the topic of the present article is enormous. I have tried to achieve a fair balance between original contributions and reviews / books. Nevertheless, my selection must remain – to some extent – arbitrary and incomplete.

Primary Literature

1. Abraham R, Marsden JE (1978) Foundations of mechanics. Benjamin / Cummings, Reading
2. Andronov A, Vitt A, Khaikin SE (1966) Theory of oscillators. Pergamon Press, London-Paris
3. Arnold VI (1963) Russ Math Surv 18:9
4. Arnold VI (1993) Dynamical systems VI singularity theory 1 v 6 (Encyclopaedia of Mathematical Sciences). Springer, Berlin
5. Arnold VI, Afrajmovich VS, Il'yashenko, Yu S, Shil'nikov Lf (1999) Bifurcation theory and catastrophe theory. Springer, Berlin
6. Aronson IS, Kramer L (2002) The world of the complex Ginzburg–Landau equation. Rev Mod Phys 74:99–143
7. Babloyantz A (1986) Molecules, dynamics and life: An introduction to Self-Organization of matter. Wiley, Indianapolis
8. Bartholomew DJ (1967) Stochastic models for social processes. Wiley, London
9. Belousov BP (1959) Sb Ref radats Med, Moscow
10. Bertalanffi L von (1950) Brit J Phil Sci 1:134
11. Bertalanffi von L (1953) Biophysik des Fließgleichgewichts. Vieweg, Braunschweig
12. Bodenschatz E, Pesch W, Ahlers G (2000) Annu Rev Fluid Mech 32:709
13. Bogoliubov NN, Mitropolsky YA (1961) Asymptotic methods in the theory of nonlinear oscillations. Hindustan Publ Corp, Delhi
14. Bonner JT, Barkley PS, Hall EM, Konjin TM, Mason JW, O'Keefe G, Wolfe PB (1972) Devel Biol 20:72
15. Bray CH (1921) J Am Chem Soc 43:1262
16. Busse FH (1972) J Fluid Mech 52:1–97
17. Bénard H (1900) Ann Chim Phys 7(23):62
18. Bénard H (1900) Rev Gén Sci Pures Appl 11:1261–1309
19. Callen HB (1960) Thermodynamics. Wiley, New York
20. Chandrasekhar S (1961) Hydrodynamic and hydromagnetic stability. Clarendon Press, Oxford
21. Chenciner A, Iooss G (1979) Arch Ration Mech Anal 69:109
22. Chow S-N, Hale JK (1982) Methods of bifurcation theory. Springer, Berlin
23. Collet P, Eckmann JP (1980) Iterated maps on the interval as dynamical system. Birkhäuser, Boston
24. Cross MC, Hohenberg P (1993) Rev Mod Phys 65:851
25. DeGiorgio V, Scully MO (1970) Phys Rev A 2:1170
26. Eigen M, Schuster P (1977) Naturwissenschaften 64:541, (1978) 65:7, 65:341
27. Elsasser W (1937) Phys Rev 52:987, (1968) Z Phys 171
28. Epstein IR, Pojman JA (1998) Introduction to nonlinear chemical dynamics. Oxford University Press, New York
29. Fantz M, Bestehorn M Friedrich R, Haken H (1993) Phys Lett A 174:48–52
30. Feigenbaum MJ (1978) J Stat Phys 19:25
31. Fenstermacher RP, Swinney HL, Gollub JP (1979) J Fluid Mech 94:103
32. Field RJ, Korös E, Noyes RM (1972) Am Chem Soc 49:8649
33. Fife PC (1979) Mathematical aspects of reacting and diffusing systems. Springer, Berlin
34. Fisher G (1967) Measuring ambiguity. Am J Psychol 80:541–547
35. Floquet G (1883) Sur les équations différentielles linéaires à coefficients périodiques. Ann Ècole Norm 2(12):47
36. Gardiner CW (1994) Handbook of stochastic methods. Springer Series in Synergetics, vol 13. Springer, Berlin
37. Gerisch G, Hess B (1974) Proc Nat Acad Sci (Wash) 71:2118
38. Giaiotti DB, Steinacker R, Stel F (2007) Atmospheric convection. Research and Operational Forecasting Aspects, Cism International Centre for Mechanical Sciences Courses and Lectures. Springer, Wien
39. Gierer A, Meinhard H (1972) Kybernetik 12:30
40. Glansdorff P, Prigogine I (1971) Thermodynamic theory of structure, stability, and fluctuations. Willey, New York
41. Gollup J, Benson SV (1979) In: Haken H (ed) Pattern Formation by Dynamic Systems and Pattern Recognition. Springer Series in Synergetics, vol 5. Springer, Berlin
42. Golubitsky M, Schaeffer D (1988) Singularities and groups in bifurcation theory I, vol 1. Springer, Berlin
43. Golubitsky M, Stewart I, Schaeffer D (1988) Singularities and groups in bifurcation theory, vol 2. Springer, Berlin
44. Graham R (1970) Quantum statistics of optical parametric oscillation. In: Kay SM, Maitland A (eds) Quantum Optics. Academic Press, New York
45. Graham R (1981) Z Phys B 40:149
46. Graham R, Haken H (1968) Z Phys 213:420 (1970) 235, 237:31,166
47. Graham R, Haken H (1971) Z Phys 248:289
48. Grossmann S, Thomae S (1977) Z Naturforsch 32A:1353
49. Guckenheimer J, Holmes P (1983) Nonlinear oscillations, dynamical systems, and bifurcations of vector fields. Springer, Berlin
50. Haake F (1973) In: Springer Tracts in Modern Physics, vol 66. Springer, Berlin, p 98
51. Haase R (1969) Thermodynamics of irreversible processes. Addison-Wesley, Reading
52. Hahn W (1967) Stability of motion. In: Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen, Bd 138. Springer, Berlin
53. Haken H (1964) Z Phys 181:96
54. Haken H (1970) In: Encyclopedia of Physics, vol XXV/2c: Laser Theory. Springer, Berlin
55. Haken H (1975) Phys Lett 53A:77
56. Haken H (1975) Z Phys B 21:105, B 22:69, B 23:388
57. Haken H (1979) Light, vol 1, Elements of Quantum Optics. North-Holland Physics Publishing, Amsterdam, New York

58. Haken H (1985) Light, vol 2, Laser Light Dynamics. North-Holland Physics Publishing, Amsterdam, New York
59. Haken H (1996) Principles of brain functioning. Springer, Berlin
60. Haken H (2000) Information and Self-Organization, 2nd edn. Springer, Berlin
61. Haken H (2004) Synergetic computers and cognition, 2nd edn. Springer, Berlin
62. Haken H (2004) Synergetics: An introduction and advanced topics. Springer, Berlin
63. Haken H, Graham R (1971) Umschau 6:191
64. Haken H, Kelso S, Bunz H (1985) Biol Cybern 51:347
65. Haken H, Wunderlin A (1982) Z Phys B 47:179
66. Hansch D (2002) Evolution und lebenskunst. Grundlagen der psychosynenergetik. Ein Selbstmanagement-Lehrbuch. Vandenhoeck und Ruprecht, Göttingen
67. Hopf E (1942) Abzweigung einer periodischen lösung eines differentialsystems. Berichte der Mathematisch-Physikalischen Klasse der Sächsischen Akademie der Wissenschaften zu Leipzig XCIV 1, Leipzig
68. Hopf E (1948) Commun Pure Appl Math 1:303
69. Horsthemke W, Lefever R (1983) Noise-Induced transitions, Springer Series in Synergetics, vol 15. Springer, Berlin
70. Hoyle RG (2006) Pattern formation. Cambridge UP, Cambridge
71. Iooss G, Joseph DD (1980) Elementary stability and bifurcation theory. Springer, Berlin
72. Îto K (1969) Stochastic processes. Universitet Matematisk Institut, Aarhus
73. Jaynes ET (1957) Phys Rev 106:4, 620, Phys Rev 108:171
74. Jaynes ET (1967) In: Delaware Seminar in the Foundations of Physics. Springer, Berlin
75. Kadanoff LP, Götze W, Hamblen D, Hecht R Lewis EAS, Palcanskas VV, Rayl M, Swift J, Aspnes D, Kane J (1967) Rev Mod Phys 39:395
76. Kelley A (1967) In: Abraham R, Robbin J (eds) Transversal Mappings and Flows. Benjamin, New York
77. Kelso JAS (1981) Bull Psyconomic Soc 18:63
78. Kelso JAS (1995) Dynamic patterns: The self-organization of brain and behavior. MIT Press, Cambridge
81. Kernbach S (2008) Structural self-organization in multi-agent and multi-robotic systems, Thesis. Stuttgart University, Logos, Berlin
79. Kielhöfer, Hansjörg (2004) Bifurcation theory, an introduction with applications to PDES. Springer, Berlin
80. Kolmogorov AN (1954) Dokl Akad Nauk USSR 98:527
82. Kuhn TS (1996) The structure of scientific revolutions, 3rd edn. University of Chicago Press, Chicago
83. Kuramoto Y (1984) Chemical oscillations, waves and turbulence. Springer, Berlin
84. Kuznetsov, Yuri A (1995) Elements of applied bifurcation theory. Springer, Berlin
85. Köhler W (1920) Die physischen gestalten in ruhe und im stationären zustand. Vieweg, Braunschweig
86. Landau LD, Lifshitz IM (1959) In: Course of Theoretical Physics, vol 5, Statistical Physics. Pergamon Press, London-Paris
87. Landauer R (1975) Phys Rev A 12:636
88. Langevin P (1908) Sur la théorie du movement brownien. CR Acad Sci Paris 146:530
89. Lorenz EN (1963) J Atmospheric Sci 20:130, 20:448
90. Lyapunov AM (1906) Sur la masse liquide homogène donnée d'un movement de rotation. Zap Acad Nauk St. Petersburg 1:1
91. Ma, Tian, Wang, Shouhong (2005) Bifurcation theory and applications. World Scientific, Singapore
92. Manneville P (1990) Dissipative structures and weak turbulence. Academic Press, San Diego
93. Marx K (1987) Analytische und numerische behandlung der zweiten instabilität beim Taylor-Problem der flüssigkeitsdynamik, Thesis. Stuttgart University, Shaker, Aachen
94. May RM (1976) Nature 261:459
95. Meinhardt H (1982) Models of biological pattern formation. Academic, London
96. Meinhardt H (1990) The beauty of sea shells. Springer, Berlin
97. Meystre P, Sargent M (1990) Elements of quantum optics. Springer, Berlin
98. Mikhailov AS (1993) Foundations of synergetics. In: Distributed Active Systems, II (with AY Loskutov): Complex Patterns. Springer, Berlin
99. Mills DL (1991) Nonlinear optics basic concepts. Springer, Berlin
100. Moser J (1967) Math Ann 169:136
101. Murdock J (2002) Normal forms and unfoldings for local dynamical systems. Springer, New York
102. Murray JD (1989) Mathematical biology, 2nd edn 1993, 3rd edn 2002/2003. Springer, Berlin
103. Nayfeh, Ali H (1993) Method of normal forms. Wiley, New York
104. Nekorkin VI, Velarde MG (2002) Synergetic phenomena in active lattices. Patterns, waves, solitons, chaos. Springer, Berlin
105. Newell AC, Whitehead JA (1969) J Fluid Mech 38:279
106. Newhouse S, Ruelle D, Takens F (1978) Commun Math Phys 64:35
107. Nicolis G (1995) Introduction to nonlinear science. Cambridge University Press, Cambridge
108. Nicolis G, Prigogine I (1977) Self-organization in non-equilibrium systems. Wiley, New York
109. Pauli H (1928) Probleme der modernen physik. In: P Debye (ed) Festschrift zum 60. Geburtstag A Sommerfelds. Hirzel, Leipzig
110. Pismen LM (1999) Vortices in nonlinear fields. Clarendon Press, Oxford
111. Pismen LM (2006) Patterns and interfaces in dissipative dynamics. Springer, Berlin
112. Pliss VA (1964) Izv Akad Nauk SSSR, Mat Ser 28:1297
113. Poincaré H (1960) Les methods nouvelles de la méchanique céleste. Gauthier-Villars, Paris. Reprint 1892/99 Dover Publ, New York
114. Prigogine I, Nicolis G (1967) J Chem Phys 46:3542
115. Rabinovich MI, Ezersky AB, Weidmann PD (2000) The dynamics of patterns. World Scientific, Singapore
116. Risken H (1965) Z Phys 186:85
117. Risken H (1989) The Fokker Planck eq. Springer, Berlin
118. Ruelle D, Takens F (1971) Commun Math Phys 20:167
119. Sargent M, Scully MO, Lamb WE (1974) Laser physics. Addison-Wesley, Reading
120. Sattinger DH (1980) Group theoretic methods in bifurcation theory. Lecture Notes Math, vol 762. Springer, Berlin
121. Schawlow AL, Townes CH (1958) Phys Rev 112:1940
122. Schiepek G (1999) Die grundlagen der systemischen therapie. Theorie-Praxis-Forschung. Vandenhoeck und Ruprecht, Göttingen

123. Schleich WP (2001) Quantum optics in phase space. Wiley-VCH, Weinheim
124. Schmidt E (1908) Zur theorie der linearen und nichtlinearen integralgleichungen, Teil 3. Math Annalen 65:370
125. Schöll E (2001) Nonlinear spatio-temporal dynamics and chaos in semiconductors. Cambridge University Press, Cambridge
126. Scully M, Lamb WE (1967) Phys Rev 159:208 (1968) 166:246
127. Segel LA (1969) J Fluid Mech 38:203
128. Sell GR (1979) Arch Ration Mech Anal 69:199
129. Shannon CE, Weaver W (1949) The mathematical theory of communication. Univ of Illin Press, Urbana
130. Smale S (1967) Bull AMS 73:747
131. Sounders PT (1980) An introduction to catastrophe theory. Cambridge University Press, Cambridge
132. Sparrow CT (1982) The Lorenz equations: bifurcations, chaos and strange attractors. Springer, Berlin
133. Staliunas K, Sanchez-Morcillo V, Gaul LJ (2003) Transverse patterns in nonlinear optical resonators. Springer, Berlin
134. Stratonovich RL (1963) Topics in the theory of random noise, vol 1, 1967 vol II. Gordon Breach, New York-London
135. Swift J, Hohenberg PC (1977) Phys Rev A 15:319
136. Swinney HL, Gollub JP (eds) (1981) Hydrodynamic instabilities and the transition to turbulence, Topics Appl Phys, vol 45. Springer, Berlin
137. Taylor GI (1923) Phil Trans R Soc Lond A 223:289
138. Thom R (1975) Structural stability and morphogenesis. Benjamin, Reading
139. Turing AM (1952) Phil Trans R Soc Lond B 237:37
140. Vavilin VA, Zhabotinsky AM, Yaguzhinsky LS (1967) Oscillatory processes in biological and chemical systems. Moscow Science Publ, Moscow, p 181
141. Walgraef D (1997) Spatio-temporal pattern formation. Springer, New York
142. Walls DF, Milburn GJ (1994) Quantum optics. Springer, Berlin
143. Weidlich W (2000) Sociodynamics. A systematic approach to mathematical modelling in the social sciences. Harwood Academic Publishers, Amsterdam
144. Weidlich W, Haake F (1965) Z Physik 186:203
145. Wigner EP (1932) Phys Rev 40:749
146. Wilson KG, Kogut J (1974) Phys Rep 12 C:75
147. Wolpert L (1969) J Theor Biol 25:1
148. Wunderlin A, Haken (1981) Z Phys B 44:135
149. Zaikin AN, Zhabotinsky AM (1970) Nature 225:535

Books and Reviews

Springer Series in Synergetics. Founded by Haken H (1977) vols 1–ca 100. Springer, Berlin

Synergetics, Introduction to

Hermann Haken
Institut für Theoretische Physik, Universität Stuttgart, Stuttgart, Germany

Synergetics (Greek **synergeon**: science of cooperation) is an interdisciplinary field of research that deals with the behavior of complex systems, i. e. systems composed of many individual parts that in general may produce complex behavior. The systems considered may belong to a great variety of fields, ranging from natural sciences with physics, chemistry, biology, through medicine and psychology to economy, ecology, management theory etc. The central topic of the synergetics enterprise is this: By means of their cooperation, the individual parts of a system can spontaneously produce structures, i. e. special arrangements between the elements or specially coordinated behavior without specific steering from the outside, i. e. without the interference of an external agent. The physical systems, studied by synergetics, are away from thermal equilibrium. The structures may be spatial patterns, be it regular or irregular, temporal patterns, i. e. all kinds of self-sustained oscillations, ranging from harmonic oscillations to chaotic behavior, spatio-temporal patterns, or functional patterns such as produced, for instance, by the brain of humans or animals, giving rise to specific behavior. The basic question synergetics asks is this: Are there general principles that govern the self-organization of systems, irrespective of the nature of the individual parts? When this question was posed some 40 years ago by H. Haken, the problem put forward seemed to be absurd in view of the great variety of possible elements. However, this research program has turned out to be quite successful as witnessed, for instance by the Springer Series in Synergetics, as well as by an independent book series in Russian. The price to be paid for approaching this goal is as follows:

1. Focus the study on those situations where the **macroscopic state** of a system changes **qualitatively**.
2. Start from comparatively simple systems that are provided either by systems in physics (e. g. lasers or fluids) or model systems dealing with self-organization processes in a variety of fields, such as biology.

As has become evident over the past few decades, self-organization phenomena occur in a vast variety of fields whose presentation actually would fill a whole encyclopedia, which would have led to an encyclopedia within an encyclopedia. Indeed the reader him- or herself may find numerous examples in his/her own field of research. In view of this situation, I felt it would be wise to present to the reader a few basic concepts and only a few prototypical examples. It is, indeed, an outstanding fact that few concepts allow us to cover a great variety of self-organization phenomena from a unifying point of view. Among these concepts are stability, instability, control parameters, order parameters and the slaving principle. Future research

will have to add more concepts as has already happened, for instance, in chaos theory.

As the reader will note, from the theoretical point of view, synergetics has become a meeting place between bifurcation theory and mathematical theories related to it, the theory of stochastic processes, and phase transition theory, actually in the sense of the Landau theory. The mathematical tools employed are generalized Langevin equations and the Fokker–Planck equation, and to some extent also the density matrix equation. In physics, it has become possible to start from first principles. For instance in quantum optics, based on Heisenberg equations of motion for operators, the coherence properties of laser light (and other non-linear optical devices) were derived in every detail (its explicit presentation would require, however, a whole handbook article, so that the reader is referred to the original literature). Fluid dynamics was chosen as a highly illustrative and prototypical example of pattern formation in systems away from thermal equilibrium. When a fluid is energetically excited, with an increasing degree of excitation it may run through a hierarchy of spatio-temporal patterns. At comparatively low excitation levels, well defined patterns evolve. While regular patterns stand in the foreground of the article ► Fluid Dynamics, Pattern Formation by Bestehorn, the article by Pismen ► Patterns and Interfaces in Dissipative Dynamics emphasizes defects and interfaces. At higher excitation levels, we reach the field of turbulence, which even after the ground-breaking work by Kolmogorov is still a hot subject of research. Important steps are done here, e. g. by the study of the dynamics of vortices. The work by Friedrich and Peinke, authors of the article ► Fluid Dynamics, Turbulence comprises also the Fokker–Planck equation approach. Actually, in the introductory article ► Synergetics: Basic Concepts by Haken, a method using the concept of information (entropy) is outlined how to derive a Fokker–Planck equation from measured data. Friedrich (private communication) and Peinke found a more direct access, the results of which are briefly outlined in the article by Friedrich and Peinke.

A quite modern line of research on complex systems follows up the method of the Fokker–Planck equation and especially of **non-linear** Fokker–Planck equations. This will be outlined in the article ► Linear and Non-linear Fokker–Planck Equations by Frank.

So far, all these articles deal with "hard" science, especially physics. As it has turned out over past decades, the general concepts of synergetics have great potentialities in fields often called "soft" science such as movement science, medicine, and even psychology and psychiatry. Actually, as the following articles will show, basic concepts of synergetics help to convert soft science into hard science. The article by Fuchs and Kelso ► Movement Coordination deals with movements of humans (and animals). But though these "systems" are highly complex, large classes of transitions between movement coordination can be theoretically and experimentally treated in great detail, even of a prototypical character.

While these approaches might be called "macroscopic", the article by Tass, Popovych and Hauptmann ► Brain Pacemaker penetrates into the microscopic level of brain functions, namely by modeling the collective behavior of neurons and their reactions to external interventions. This paper is remarkable, because it shows a new aspect of the research on coupled nonlinear oscillators. Whereas so far the problem of how such systems synchronize, was in the foreground of research, now for medical reasons, namely to fight Parkinsons's disease, the problem arises how to desynchronize such a system by appropriate interventions.

The mathematical theory of synergetics or perhaps in a wider sense dynamical systems theory with concepts such as attractors, allows one to cast earlier concepts developed by Gestalt theory into clear-cut mathematical models. The order parameters just play the role of Gestalt. Synergetics may be considered as a theory of indirect control of systems, in that it provides psychiatrists with insights for how to intervene with their clients. A comprehensive article, dealing with these aspects, is given in ► Self-Organization in Clinical Psychology by Schiepek and Perlitz. This article shows how methods of mathematical analysis are penetrating more and more into the field of psychiatry, and allowing doctors to monitor the mental and behavioral state of a patient.

In conclusion, in particular with respect to biological systems (but not exclusively), the following remarks are in order:

1. While in physical systems, at least those considered in the article on synergetics, the control parameters are fixed from the outside, in biological systems the control parameters are in general produced by the system itself. They are the slowly varying variables in this case. When their dynamics is admitted, specific phenomena may occur, such as the recognition of ambivalent figures (switching between two or several percepts).
2. In biological and some physical processes, pattern formation (morphogenesis) is a two- or multi-step process. In the first step **dynamic structures** are formed. In a second step these dynamic patterns are transformed into solid patterns. A nice example is provided by radiolarians in which eventually dynamic concentration

patterns of chemicals are transferred into skeletons by means of calcification. The exploration of such two- or multi-step processes is, as it seems, just at its beginning.

In the article ► Intentionality: A Naturalization Proposal on the Basis of Complex Dynamical Systems, W. Tschacher shows how mental processes can be linked to material processes where he tackles the longstanding problem of intentionality and relates it to a gradient dynamics which can be formalized by concepts of synergetics.

Last but not least, Synergetics as a theory dealing with systems composed of many individual parts deals with the interrelation between human individuals and their ability to form specific social groups, viewed as a process of self-organization. An outstanding example is provided by Juval Portugali's article ► Self-Organization and the City.

System Dynamics, Analytical Methods for Structural Dominance Analysis in

Christian Erik Kampmann[1], Rogelio Oliva[2]
[1] Department of Innovation and Organizational Economics, Copenhagen Business School, Copenhagen, Denmark
[2] Mays Business School, Texas A&M University, College Station, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Characterizing Linear and Nonlinear Systems
Traditional Control Theory Approaches
Pathway Participation Metrics
Eigenvalue Elasticity Analysis
Eigenvectors and Dynamic Decomposition Weights (DDW)
Future Directions
Bibliography

Glossary

Behavior mode The traditional meaning of the term is the qualitative nature of the observed system behavior, such as damped or expanding oscillations, overshoot and collapse, exponential growth or adjustment to equilibrium, or limit cycles. In linear systems theory, the term has a more specific meaning, cf. the explanation for eigenvalues.

Bode plot (phase and gain plot) A tool used in classical control theory to characterize the frequency response, i. e., the amplification A and phase shift ϕ in the system output variable of interest $x(t) = A\sin(\omega t + \phi)$ compared to the input variable $u(t) = \sin(\omega t)$, as a function of the frequency ω of the input.

Chaos A type of behavior exhibited by nonlinear systems that appears to be approximately periodic but with a seemingly random element. A hallmark of chaotic behavior is that it is sensitive to initial conditions.

Dominant structure A general term for the feedback loops (or possibly external driving forces) that are "most important" in generating a behavior pattern of interest. In nonlinear models, particularly single-transient models, there is frequently a shift in structural dominance, i. e. in the strength and significance of certain feedback loops.

Dynamic decomposition weights (DDW) An application of Eigenvector Elasticity Analysis (EVA) that focuses on how parameter changes influence the relative weights (DDW's) of the system behavior modes in a particular variable.

Eigenvalue An eigenvalue for a square matrix $\boldsymbol{A}$ is a value λ for which the equation $\boldsymbol{Ar} = \lambda \boldsymbol{r}$ has a non-zero solution $\boldsymbol{r} \neq \boldsymbol{0}$. The column vector $\boldsymbol{r}$ is called the (right) eigenvector corresponding to the eigenvalue λ. The eigenvalues and eigenvectors determine the behavior modes (components) in the solution to the linear dynamical system $\dot{\boldsymbol{x}} = \boldsymbol{Ax}$. A real eigenvalue λ leads to an exponential behavior mode $\exp(\lambda t)$ while a complex eigenvalue $\lambda = \tau \pm i\omega$ leads to oscillatory behavior modes $\exp(\tau\, t)\sin(\omega t + \phi)$. The eigenvectors determine the weight, or the degree to which a particular behavior mode is expressed in a particular system variable.

Eigenvalue elasticity analysis (EEA) A method of analyzing the significance of a structural element, say a loop or a link in the model with a gain g, in terms of its marginal effect upon the eigenvalues λ of the system. There are several such measures, such as the *influence measure* $\partial\lambda/\partial g \cdot g$, the *elasticity* $\partial\lambda/\partial g \cdot (g/\lambda)$, or, in the case of complex-valued eigenvalues, the effect upon the damping ratio, natural frequency, damping time, etc., as illustrated in Fig. 7. See also *Loop Eigenvalue Elasticity Analysis (LEEA)*.

Eigenvector See explanation for *Eigenvalue*.

Eigenvector elasticity analysis (EVA) A complement to Eigenvalue Elasticity Analysis (EEA) that looks explicitly at the expression or relative weight of each behavior mode in each system variable. These weights are related to the eigenvectors of the system matrix.

Frequency domain A term used to describe the analysis of signals with respect to frequency. While a time domain graph shows the behavior of the signal over time, the frequency domain graphs shows how much of the signalvariance lies within each given frequency band.

Independent loop set (ILS) Although the number of feedback loops in a model can be very large (theoretically astronomically large), there is a much smaller *independent loop set* that can be considered independent structural elements. For a strongly connected system (where any pair of variables are connected via causal chain in both directions) with N links and n variables, there are exactly $N - n + 1$ independent loops. Simple algorithms exist for constructing independent loop sets, in particular *Shortest Independent Loop Sets (SILS)*. See also explanation for *Loop Eigenvalue Elasticity Analysis (LEEA)*.

Linear dynamical system A system where the rates $\dot{\boldsymbol{x}} = (\mathrm{d}x_1/\mathrm{d}t, \ldots, \mathrm{d}x_n/\mathrm{d}t)$ are a linear function of the state variables $\boldsymbol{x} = (x_1, \ldots, x_n)$ and exogenous or control variables $\boldsymbol{u} = (u_1, \ldots, u_p)$, expressed by the equation $\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}$ where $\boldsymbol{A}$ is an $n \times n$ matrix and $\boldsymbol{B}$ is an $n \times p$ matrix. Unlike nonlinear systems of the general form $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u})$, linear systems have analytical solutions based on the eigenvalues and eigenvectors of the matrix $\boldsymbol{A}$ (cf. explanation for *Eigenvalues*).

Linear systems theory The mathematical theory of *linear dynamical systems*.

Loop eigenvalue elasticity analysis (LEEA) A form of eigenvalue elasticity analysis (EEA) that uses graph theory to express structural changes in terms of change in the strength of individual feedback loops. *Independent* loops can be assigned individual (loop) eigenvalue elasticities or influence measures just like other structural elements (see explanation for *Eigenvalue Elasticity Analysis (EEA)* and *Independent Loop Set (ILS)*).

Model simplification approach A way of attributing dynamic behavior to particular pieces of structure by replacing the full model with a simplified structure. See also *Structure contribution approach*.

Nonlinear systems Systems of the form $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u})$ where $\boldsymbol{f}$ is a nonlinear function. See explanation for *Linear dynamical systems*.

Pathway participation metric A measure that decomposes the curvature ($\ddot{x} = \mathrm{d}^2x/\mathrm{d}t^2$) of a variable x_i into the individual driving components, $\ddot{x}_i = \sum_j \partial\dot{x}_i/\partial x_j \cdot \dot{x}_j$. By considering the sign of the curvature relative to the slope, i. e., $\ddot{x}/\dot{x}$, one may define behavior as (apparently) dominated by positive $\ddot{x}/\dot{x} > 0$ or negative $\ddot{x}/\dot{x} < 0$ feedback loops. The component (pathway) with the largest absolute value and the same sign as $\ddot{x}/\dot{x}$ is then defined as the dominant structure.

Quasilinear models Models that are almost linear in structure around the operating point of interest so that they may be well approximated by a linear model.

Quasiperiodic behavior A behavior that is a sum of oscillations of incommensurate frequencies so that the system never returns to exactly the same point (which would be the case for periodic behavior).

Shortest independent loop set (SILS) An *Independent Loop Set (ILS)* that consists of the shortest possible loops (in terms of the number of nodes and links in each loop). Since the choice of ILS is far from unique, an SILS provides a more focused choice of loops, which are typically also easier to interpret due to their short length.

Single-transient models Models where the behavior of interest is the transition toward an equilibrium or constant growth rate. Models are typically nonlinear, exhibit patterns such as smooth transition, overshoot and collapse, growth, or stagnation.

Structure contribution approach A way of linking model structure to dynamic behavior by considering how individual pieces of structure (feedback loops or subsystems) contribute to the behavior pattern of interest by turning the structure on or off (in traditional simulation experiments) or by considering the marginal effect of small changes in structure (the eigenvalue approach). See also *Model simplification approach*.

Definition of the Subject

The link between system structure and dynamic behavior is one of the defining elements in the system dynamics paradigm, yet it is only recently that systematic, mathematically rigorous methods for exploring this link have started to become available. In a sense, a simulation model can be viewed as an explicit and consistent theory of the behavior it exhibits. Although this point of view has certain merits, not least the fact that it lifts the discussion from outcomes to causes of these outcomes and from events to underlying structure [11,59], we are concerned here with a more compact explanation of the system's behavior. In fact, most system dynamics modeling projects report their results in terms of simpler explanations of the observed results, typically in terms of dominant feedback loops that produce the salient features of the behavior.

Most often, dominant structure is thought of in terms of feedback loops and, occasionally, external driving forces to the system. For simple systems with relatively few vari-

ables it is usually easy to use intuition and trial and error simulation experiments to explain the dynamic behavior as resulting from particular feedback loops. In larger systems, this method becomes increasingly difficult and the risk of incorrect explanations rises accordingly. There is a need, therefore, for analytical methods that provide some consistency and rigor to this process.

These analytical tools are important to the practitioner because the structure-behavior link is the key to finding leverage points for policy initiatives. And they are important to the theorist because a system dynamics theory of a particular phenomenon is an account of how certain feedback loops cause certain dynamic patterns of behavior to appear. The qualitative understanding of the model behavior is often at least as important as the particular numerical predictions obtained, even in applied studies. Yet the rigor of such an account depends directly on the rigor with which structure-behavior link can be made in a given model.

The classical disciplines of linear systems theory and control engineering have provided a set of concepts and tools, particularly system eigenvalues and eigenvectors, that can also be applied under many circumstances to the nonlinear models found in system dynamics, not as a complete theory but as a pragmatic aid. This article reviews the recent advances in analytical tools based on linear systems theory and discusses its future potential for the both the system dynamics practitioner and the theorist.

Though we strongly believe in the utility of these methods, it is important to realize that advances in nonlinear dynamics and complexity theory in recent decades have shown that it is not possible to construct a complete theory of dominant structure because nonlinear systems are capable of exceedingly complex and intricate behavior that is impossible to predict without actually simulating the system. Furthermore, applications of graph theory to system dynamics models have revealed that the concept of feedback loops has some inherent problems and limitations because there are potentially many different loop descriptions of the same system (see [28,40]). Thus, the analytical tools should be viewed as pragmatic aids to model analysis that can guide the modeler's intuition, rather than universal methods that provide automatic answers.

We first provide a brief historical introduction to the different ways scholars have thought about the notion of dominant structure, including an example of the traditional approach to structural analysis. In the next section we present the formal mathematical representation of linear and nonlinear systems and how one may describe the dynamic behavior in terms of behavior modes and system eigenvalues. In the four following sections we present alternative approaches to performing this analysis. We conclude with a summary of the current state of research and a discussion of future directions.

Introduction

Understanding model behavior is closely related to the process of model testing and validation, for which there is a well-established tradition and an extensive literature in the field (e. g., [2,10,17,36,46,47]). Indeed there is no sharp line between model building, testing, validation, and analysis – in practice, the analyst undertakes all these processes simultaneously [17].

Of particular concern is whether one can identify pieces of structure that are in some sense "important" in generating the observed behavior of interest. Traditionally, system dynamics analysts have relied on trial-and-error simulation to discover these structures, by changing parameter values or switching individual links and feedback loops on and off. The tradition is well developed and includes a set of principles for partial model formulation and testing based the organizational theory of bounded rationality [27,36].

The intuition guiding this effort often relies on simple feedback systems with one or a few state variables, where the behavior is fully documented and understood. In particular, the modeler uses well-understood "generic structures" that seem to appear again and again in system dynamics models, such as "overshoot and carrying capacity collapse", "drifting goal structure", etc. (see [30,56,60] for an account of these structures). Clearly such structures can be a useful aid to understanding if the model is sufficiently simple to allow such simple structures to be identified.

A simple example of a generic structure is the classical model of diffusion, sometimes known as the Bass model ([3], see also Chapter 9 in [59]). The model structure is illustrated in Fig. 1, and the resulting behavior, an s-shaped growth curve, is illustrated in Fig. 2. The idea behind the model is that the adoption of a new technology is driven by the number of users that have already adopted it, through a word-of-mouth effect. One may interpret the s-shaped behavior as the interaction of two feedback loops, namely loop no. 2, the positive "word-of-mouth", and loop no. 1, the negative "exhaustion" loop (see Fig. 1). In the beginning, the positive loop dominates, leading to exponential growth in the number of adopters. Later, however, the negative loop gains strength, and the behavior shifts to an exponential adjustment toward the eventual market saturation. Thus, the traditional feedback loop analysis helps give an intuitive understanding of the dynamics of the model.

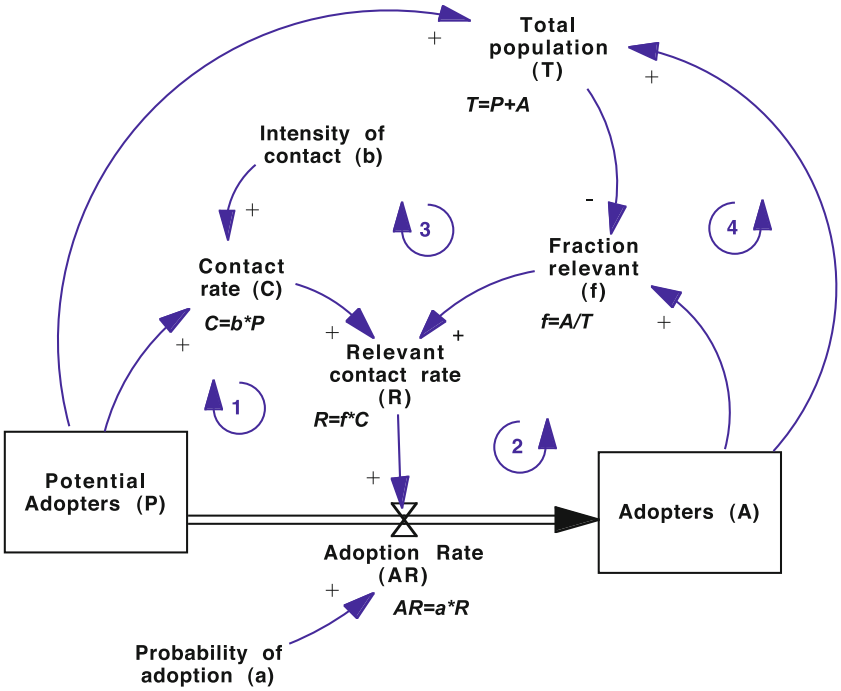

System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 1
The Bass model of diffusion

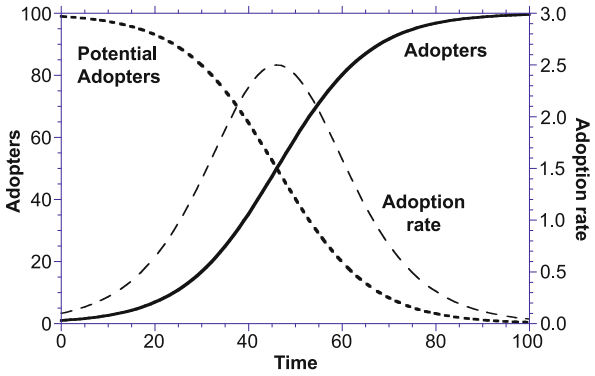

System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 2
Behavior of the Bass model

In large-scale models with perhaps hundreds of state variables, however, the traditional approach shows significant limitations. In practice, model building and analysis is often done using a "nested" partial model testing approach where one goes from the level of small pieces of structure to entire subsystems of the model, with frequent re-use of known formulations and partial models. Although this approach does carry a long way, it can be very difficult to discover feedback mechanisms that transcend model substructures in ways not anticipated by the modeler in the original dynamic hypothesis. Thus, there is a danger that observed behavior is falsely attributed to certain feedback mechanisms when in fact another set of feedbacks is driving the outcome. Likewise, one may make false inferences about how a particular feedback mechanism modifies the behavior, e. g., whether it attenuates or amplifies a particular oscillation.

Modern software packages can run extensive tests for sensitivity and "reality checks" where a large number of parameters are varied simultaneously [44]. This is clearly a significant improvement over "manual" trial and error methods, particularly when these methods are combined with statistical inference methods such as Kalman Filtering or Monte Carlo maximum likelihood estimation [6,8,39,43,45,55]. A variant of this approach involves using statistical experimental design and correlation methods to screen for significant model structure (parameters), as suggested by Ford and Flynn [9]. Indeed, the prospects of marrying such methods with modern search and optimization methods like classifier systems [26] or genetic algorithms [19] seem very promising. However, these methods are more addressing issues in estimation, validation and testing than inferences about or understanding how (dominant) structure is causing behavior.

Richardson [47] suggested a taxonomy of approaches to the notion of dominant structure, where he distin-

guishes along three dimensions, namely linear vs. nonlinear systems, model reduction vs. loop contribution, and the characterization of behavior in terms of time graphs vs. eigenvalues or frequency response. Of these, the distinction between model reduction and loop contribution is the most important.

In the model reduction approach, the idea is to replace a large complicated model with a simplified smaller model that captures the "essence" of the dynamics. A good example of this is Sterman's simple model of the economic long wave [58], which was distilled from the much larger System Dynamics National Model [18]. Eberlein [5,7] attempted to tackle model simplification in a systematic way in linear systems by focusing on retaining specific behavior modes. In large part his results were negative: it is generally not possible to build simpler models that reproduce the salient behavior without sacrificing either the accuracy of the behavior or the ability to relate the simplified model variables to those in the full model. It is fair to say that this line of inquiry has largely been abandoned as a result. Extracting the "essence" of a model remains an art more than a science.

The focus here will be on Richardson's second category, the loop contribution or, more generally, the *structure contribution* approach. It reflects the intuitive idea that if one removes the element under consideration, e. g. by weakening a link or switching off a feedback loop, and the behavior then "disappears", one would say that the element in some sense "causes" the observed behavior.

This notion underlies the traditional trial-and-error simulation approach, sometimes supplemented with methods from the classical control engineering, which focuses on how structural elements modify the behavior of the system, viewed in terms of the frequency response. Typically, the method works "backwards" by starting with simple feedback systems of single loops and then considering the marginal effect of adding links and loops. We discuss this approach in Sect. "Traditional Control Theory Approaches" below.

If, instead, one considers marginal (infinitesimal) changes in structure, e. g. in the strength of a particular link, it is possible to derive rigorous analytical results for the resulting change in behavior expressed as the eigenvalues of the linearized model. One would then say that if a change in a system element has a relatively large effect upon the behavior pattern of interest, this element is "significant" in "causing" the behavior. This is what underlies the *eigenvalue elasticity* and *eigenvector* approaches discussed in Sects. "Eigenvalue Elasticity Analysis", "Eigenvectors and Dynamic Decomposition Weights (DDW)". The marginal and experimental approaches may supplement each other well, where a marginal analysis may identify elements that can then be tested experimentally for their significance.

Unlike the traditional control method and the eigenvalue method that work in the structural and *frequency domain*, the *pathway participation* method (PPM) relates directly to the time path of particular system variables and is more concerned with the qualitative nature of the time path, expressed in terms of signs of the slope (whether growing or declining) and curvature (whether convex or concave) than with numerical measures of degree of influence. Briefly stated, the PPM traces the causal links in the variables influencing the system variable in question and then identifies the most important chain of links. We discuss this method in Sect. "Pathway Participation Metrics".

Common to the approaches discussed here is that they all build upon a precise mathematical characterization of the system behavior. In the next section, we demonstrate how the concepts from linear systems theory may be used to give a precise characterization of behavior in terms of component *behavior modes*.

Characterizing Linear and Nonlinear Systems

A system dynamics model can be represented mathematically as a set of ordinary differential equations

$$\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} \equiv \dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{u}(t)) \ , \tag{1}$$

where $\boldsymbol{x}(t)$ is a (column) vector of n state variables (levels) $(x_1(t), \ldots, x_n(t))$, $\boldsymbol{u}(t)$ is a column vector of p exogenous variables or control variables $\big(u_1(t), \ldots, u_p(t)\big)$, $\boldsymbol{f}(\,)$ is a corresponding vector function, and t is simulated time. In this paper, we restrict our attention to the state variables (levels) of the model for notational convenience, ignoring the auxiliary variables. Mathematically, a model can always be brought to the *reduced* form (1), but in practice, the auxiliary variables give a more intuitive account of the analysis. Likewise, we do not consider time-varying systems (where time t enters as an explicit argument in the function $\boldsymbol{f}$), since these can usually be accommodated by an appropriate definition of the exogenous variables $\boldsymbol{u}$. In general, $\boldsymbol{f}$ is a nonlinear function of its arguments, and we speak of a *nonlinear* system. Conversely, if $\boldsymbol{f}$ is a linear function, we speak of a *linear* system.

Figure 3 and Table 1 show a well-known example, the inventory–workforce model. It has three state variables, Inventory (INV), Workforce (WF), and Expected Demand

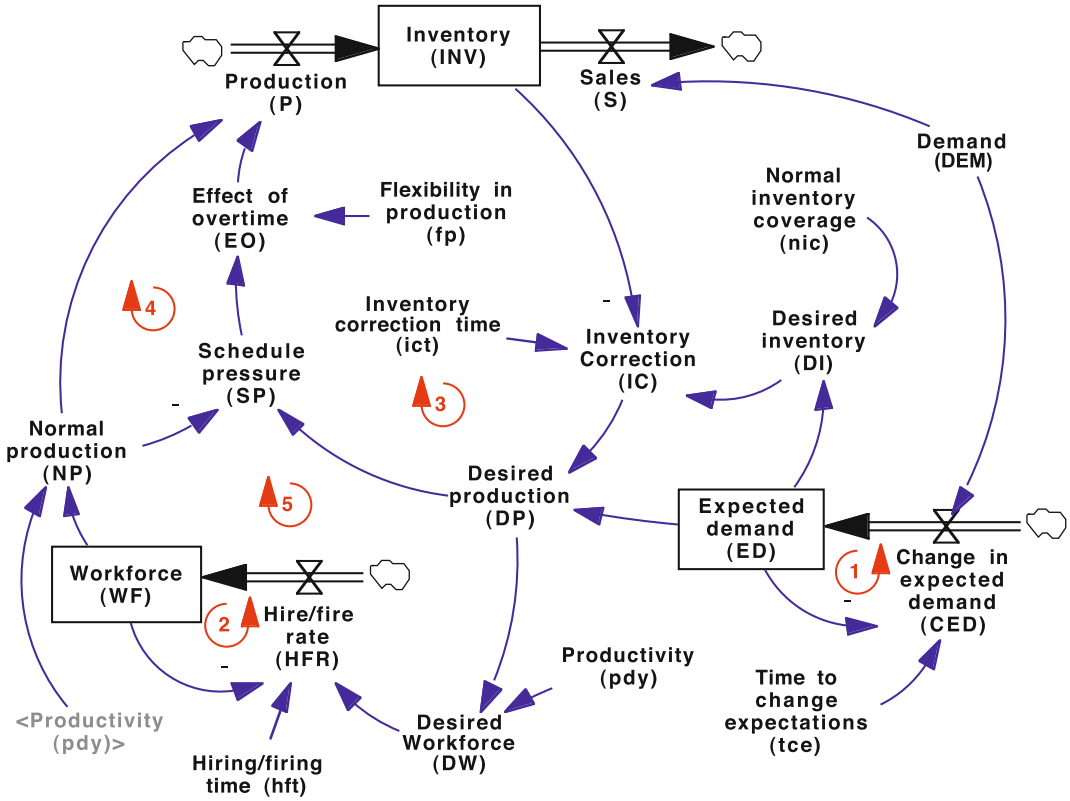

System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 3
Flow diagram of the inventory workforce model

(ED), and one exogenous variable, Demand (DEM), i. e.,

$$\boldsymbol{x}(t) = \begin{pmatrix} \text{INV} \\ \text{WF} \\ \text{ED} \end{pmatrix}; \quad \boldsymbol{u}(t) = (\text{DEM}) , \tag{2}$$

and the function $\boldsymbol{f}$ is determined by the equations in Table 1.

Given the model structure (1), knowledge of the initial conditions $\boldsymbol{x}(0)$, and the path of the input variables $\boldsymbol{u}(t)$, the behavior of the model is completely determined. It is in this sense that the model structure (1) constitutes a "theory" of the time behavior $\boldsymbol{x}(t)$, as mentioned in the introduction. Yet, we are interested in methods that yield a more compact explanation, short of having to simulate the entire model structure.

It turns out that in its ultimate form, this dream is beyond reach: Since the days of Henri Poincaré, mathematicians have known that it is impossible to find general analytical solutions to nonlinear systems. Furthermore, the development of nonlinear dynamics and chaos theory has proven that such systems, even when they have very few state variables, can produce highly complex and intricate behavior that goes beyond general analytic methods (e. g., [42,48]). Thus, we will never find a final general theory where we can infer the behavior of the system directly from its structure; instead, we will always have to rely on simulation to discover the dynamics implied by the structure. (This is not to say that no general analytical results exist in nonlinear systems. The field of chaos theory has uncovered a number of universal features, e. g., relating to the transition from periodic or quasi-periodic behavior to chaos, where the transitions show both qualitative and quantitative similarities that are independent of the specific forms of the model equations (see, e. g., [42]). However, these universal features relate to specific situations such as period-doubling or intermittency routes to chaos).

The best we can hope for, therefore, is a set of tools that will guide intuition and help identify *dominant structure* in the model. By dominant structure we mean particular feedback loops that are in some sense "important" in shaping the behavior of interest. To the extent that we can identify such dominant structures, we may say that we have found a "theory" of the observed behavior.

Although the term "behavior" may appear rather loose, experience and reflection tells us that there is a limited number, perhaps a dozen or so, of relevant behav-

System Dynamics, Analytical Methods for Structural Dominance Analysis in, Table 1
Equations of the inventory workforce model

Equation	Name	Units
d/dt(ED) = (DEM − ED)/tce, ED_0 = DEM ∗ df	Expected demand	[Units/Month]
d/dt(INV) = P − S, INV_0 = DI	Inventory	[Units]
d/dt(WF) = HFR, WF_0 = DWF	Workforce	[Workers]
S = DEM	Shipments	[Units/Month]
P = NP ∗ EO	Production	[Units/Month]
EO = fp ∗ (1 − (1 − 1/fp)SP)	Effect of overtime	[Dimensionless]
NP = WF ∗ pdy	Normal production	[Units/Month]
DI = ED ∗ nic	Desired inventory	[Units]
SP = DP/NP	Schedule pressure	[Dimensionless]
DEM = 1(Exogenous)	Demand	[Units/Month]
DP = ED + IC	Desired production	[Units/Month]
IC = (DI − INV)/tci	Inventory correction	[Units/Month]
HFR = (DWF − WF)/hft	Hire/fire rate	[Workers/Month]
DWF = ED/pdy	Desired workforce	[Workers]
hft = 5	Hire/fire time	[Month]
pdy = 1	Productivity	[Units/Month/Worker]
tce = 4	Time to change expectations	[Month]
fp = 1.05	Flexibility in production	[Dimensionless]
nic = 3	Normal inventory coverage	[Months]
ict = 2	Inventory correction time	[Months]
df = 0.5	Disequilibrium fraction	[Dimensionless]

ior patterns that dynamical systems can exhibit. Some of these behaviors, like exponential growth, exponential adjustment, and damped or expanding oscillations, are typical of linear systems. Others, like limit cycles, quasiperiodic motion, mode-locking, and chaos, can only be exhibited by nonlinear systems.

Common to the approaches considered in this paper is that they are based on tools from linear systems theory, i. e., they approximate the nonlinear model (1) with a linearized version, using first-order Taylor expansion around some operating point $\boldsymbol{x}_0, \boldsymbol{u}_0$, i. e.,

$$\dot{\boldsymbol{x}}(t) \approx \boldsymbol{f}(\boldsymbol{x}_0, \boldsymbol{u}_0) + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}(\boldsymbol{u} - \boldsymbol{u}_0) \ , \quad (3)$$

or, by redefinition of the variables $\boldsymbol{x} \to \boldsymbol{x} - \boldsymbol{x}_0 - \boldsymbol{f}(\boldsymbol{x}_0, \boldsymbol{u}_0) \times (t - t_0)$ and $\boldsymbol{u} \to \boldsymbol{u} - \boldsymbol{u}_0$,

$$\dot{\boldsymbol{x}}(t) \approx \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t) \ , \quad (4)$$

where $\boldsymbol{A}$ is constant $n \times n$ matrix of partial derivatives $\partial f_i / \partial x_j$ and $\boldsymbol{B}$ is constant $n \times p$ matrix of partial derivatives $\partial f_i / \partial u_j$, and all partial derivatives are evaluated at the operating point.

For the linear system (4), there is a well-developed and extensive theory of the system behavior as a function of its structure, expressed in the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. One may broadly distinguish two parts of the theory, named classical control theory (e. g., [38]) and modern linear systems theory (e. g., [4,31]). We return to the classical control theory in the next section.

Modern control theory or linear systems theory (LST) is concerned with the dynamical properties of the system as a direct function of the system matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. A key element in this theory is the notion of the system *eigenvalues*, i. e., the eigenvalues of the matrix $\boldsymbol{A}$. If, for simplicity, we restrict ourselves to the endogenous dynamics of the system (set $\boldsymbol{u} = 0$), we can write the solution to (4) as

$$x_i(t) = c_{i,1} \exp(\lambda_1 t) + c_{i,2} \exp(\lambda_2 t) + \cdots + c_{i,n} \exp(\lambda_n t) \ , \quad i = 1, \ldots, n \ , \quad (5)$$

where $\lambda_1, \ldots, \lambda_n$ are the n eigenvalues of the matrix $\boldsymbol{A}$ and $c_{i.j}$ are constants that depend upon the eigenvectors and the initial condition of the system. In other words, the resulting behavior is a weighted sum of distinct *behavior modes*, $\exp(\lambda t)$. If an eigenvalue is real, the corresponding behavior mode is exponential growth (if $\lambda > 0$) or exponential decay (if $\lambda < 0$). Complex-valued eigenvalues come in complex conjugate pairs $\lambda = \tau \pm i\omega$ which give rise to oscillations $\exp(\tau t) \sin(\omega t + \phi)$ of frequency ω that are either expanding (if $\tau > 0$) or damped (if $\tau < 0$).

In this manner, the eigenvalues serve as a compact and rigorous characterization of the behavior (of linear systems).

At any point in time, any system, linear or nonlinear, may be approximated by the expression (5). Whether it remains a good approximation depends upon how much and how quickly the eigenvalues change due to the nonlinearities in the function $\boldsymbol{f}$. If they are more or less constant for significant periods of time, we may speak of *quasilinear systems* that are well approximated by the linear system. In some cases, however, the eigenvalues change so rapidly that it makes little sense to characterize the behavior by equation (5). (See [29] for further discussion).

Traditional Control Theory Approaches

The first set of methods, which we call the traditional approach, has been used for decades and is part of the standard curriculum in system dynamics teaching at the graduate level. It involves using the concepts from classical control theory [38] to very simple systems with only a few state variables.

The starting point is the simple first- and second-order positive and negative feedback loops found in any introductory treatment of system dynamics. The advantage of the approach is its simplicity. Although it serves at a guide to intuition, however, the obvious shortage is that it applies rigorously only to simple systems. There have been some attempts to treat higher-order systems by adding a few feedback loops [23], but the step to large-scale models is beyond this method given its inherent limitations.

Graham [23] distills a number of principles that are based on the metaphor of a "disturbance" traveling along the chain of causal links in a feedback loop and getting amplified, damped, and possibly delayed in the process. For major negative feedback loops, which are known to tend to produce oscillation, adding minor negative loops and cross-links, or shortening the delay times increases the damping. Conversely, adding positive loops in to the oscillatory system tends to lengthen the period of oscillation whereas the effect on the damping depends upon the delays in the positive loop. Using the metaphor of pushing a child on a swing, it becomes clear that the timing of the propagation of a disturbance has as much importance for its effect on the damping as its strength.

For analyzing the behavior of positive feedback loops, Graham suggested calculating the Open-loop steady-state gain (OLSSG), a measure of the amplification around the loop. A gain greater than unity will result in exponential growth while gains less than 1 will give exponential adjustment (leveling off or decay). The intuition is perhaps best illustrated by an example: sales-driven growth. Suppose a salesperson can eventually pull in \$100,000 per month in orders (probably with a several-month long delay), and assume that the company allocates 10% of revenue to marketing. Then this eventually leads to \$10,000 per month for sales efforts. If the cost of a salesperson (salary, overhead, expenses etc.) is, say, \$8,000 dollars per month, then the efforts of the current sales force will provide enough revenue to support $10,000/8,000 = 1.25$ persons per current person. Thus, the OLSSG of the positive loop from salespersons → orders → revenues → marketing budget → salespersons is 1.25, and the system will grow exponentially (until other factors limit the growth). Conversely, if the gain is less than 1, one salesperson will not sell enough to support their own cost, and the loop will lead to exponential decay. Graham showed how the actual rate of growth is partly determined by the OLSSG, and partly by the time constants (delays etc.) involved. (See also Subsect. 15.3 in [59]).

In the context of oscillating systems, system dynamics has also employed a concept from classical control theory, frequency response. The frequency response is determined from the transfer function of the system, $G(i\omega)$, which is a complex-valued function that specifies how an input signal $u(t)$ with frequency ω results in an output signal $x(t)$ that may be phase shifted (delayed), and either amplified or attenuated. For linear systems, G can be calculated directly from the system matrices in (4) – the transfer function (matrix) is $\boldsymbol{G}(i\omega) = \boldsymbol{B}(i\omega\boldsymbol{I} - \boldsymbol{A})^{-1}$, where $\boldsymbol{I}$ is the identity matrix (see e. g. [4]). For nonlinear systems, G may be found through simulation experiments.

Usually, G is represented in a *Bode* or *phase-and-gain* diagram. For instance, Fig. 4 shows a Bode diagram of the

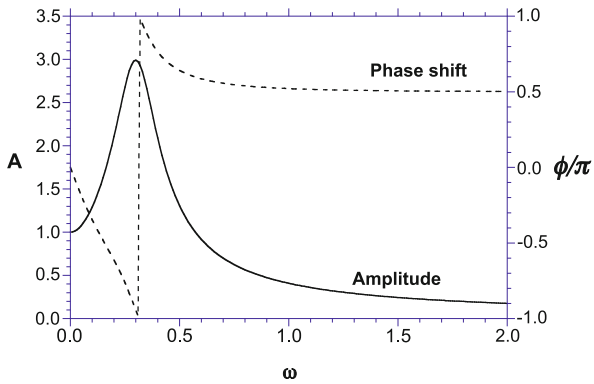

System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 4

Phase-and-gain diagram (Bode diagram) showing the inventory $A\sin(\omega t + \phi)$ with amplitude A and phase shift ϕ, relative to a sinusoidal demand $\sin(\omega t)$, for varying values of the frequency ω of the demand fluctuation

inventory variable INV (t) relative to the exogenous demand input variable DEM (t) in the inventory workforce model in Fig. 3. The diagram shows how the relative amplitude of the oscillation and the relative phase shift (in radians) between input and output varies as a function of the frequency of the input.

It is clear from the diagram that there is a certain frequency range, around the system's own natural frequency, where fluctuations in demand are greatly amplified compared to other frequencies. Indeed, it is a general phenomenon in systems that they will tend to amplify certain frequencies while attenuating other frequencies. This may be used to explain or understand the role of particular structures in the model in generating oscillation at certain frequencies, even when there are no oscillations coming in from the outside world. (External random noise is enough to produce oscillations in the system because random noise contains fluctuations at all frequencies). In this manner, the approach nicely demonstrates the "endogenous viewpoint" that behavior (oscillations) is generated internally by the system. As an analytic tool for large scale systems, however, the method does not seem to produce any additional insights. Thus, we may conclude that the classical approaches serve mostly as intuitive metaphors to guide the analyst rather than as full analytical tools.

Pathway Participation Metrics

The *pathway participation* method [34,35] represents a further development of an original suggestion by Richardson [46] to provide a rigorous definition of loop polarity and loop dominance. Richardson motivated this with the common confusion associated with positive feedback loops, which may exhibit a wide range of behaviors [23], as Barry Richmond noted with wonderful humor:

"Positive loops are … er, well, they give rise to exponential growth … or collapse … but only under certain conditions … Under other conditions they behave like negative feedback loops …" [49].

Richardson proposed that the polarity of a loop be defined as the sign of the expression

$$\frac{\partial \dot{x}_i}{\partial x_i} = \frac{\partial f_i(\boldsymbol{x}, \boldsymbol{u})}{\partial x_i}, \tag{6}$$

in the model (1), with a positive sign indicating a positive loop and vice-versa. When several loops operate simultaneously, the sign of the expression indicates whether the positive or negative loops dominate. Note, however, that the definition only applies to minor loops (i. e. loops involving a single level). Put differently, it only considers the diagonal elements of the matrix $\boldsymbol{A}$ in the linearized system (4). Richardson [46] demonstrates how even with this limitation, analyzing the system with this metric can (sometimes) yield insights into behavior of higher-order systems.

The expression (6) hints that it is relevant to consider the curvature, i. e., the second time derivative, $\ddot{x}$, of a variable when looking for dominant structure. Although he does not say so explicitly, this is effectively the focus of Mojtahedzadeh's pathway method. Figure 5 shows how one may classify behavior by comparing the first and second time derivatives of a variable. As seen in the figure, the sign of the expression $\ddot{x}/\dot{x}$, which Mojtahedzadeh denotes the total *pathway participation metric* or *PPM*, indicates whether the behavior appears dominated by positive or negative loops, much in line with Richardson's definition of dominant polarity. A zero curvature indicates a shift in loop dominance (cf. the middle column in the figure). Note, however, that the interpretation of the middle row in the figure where the slope $\dot{x}$ is zero has no clear interpretation in terms of loop dominance. Indeed this hints at one of the weaknesses of the approach that we will return to below.

Mojtahedzadeh's method proceeds by decomposing the PPM into its constituent terms as follows,

$$\mathrm{PPM}_i = \frac{\ddot{x}_i}{\dot{x}_i} = \sum_{j=1}^{n} \frac{\partial f_i}{\partial x_j} \frac{\dot{x}_j}{\dot{x}_i}, \tag{7}$$

where, for brevity, we have chosen to ignore the exogenous variables u. One might say that each of the terms in the sum in (7) represents the separate influence of each of the systems' state variables on the behavior of x_i. Mojtahedzadeh in fact uses a normalized measure for the terms,

$$\frac{\left(\partial f_i/\partial x_j\right) \dot{x}_j}{\sum_{k=1}^{n} \left|\left(\partial f_i/\partial x_k\right) \dot{x}_k\right|}, \tag{8}$$

which can vary between -1 and $+1$, to measure the relative importance of the pathway from variable j. By explicitly considering auxiliary variables y in the model, one may further decompose each term $\partial f_i/\partial x_j$ into a sum of terms

$$\frac{\partial f_i^k}{\partial x_j} = \frac{\partial f_i}{\partial y_1} \cdot \frac{\partial y_1}{\partial y_2} \cdot \ldots \cdot \frac{\partial y_{m-1}}{\partial y_m} \cdot \frac{\partial y_m}{\partial x_j}, \tag{9}$$

corresponding to a causal chain or pathway $\pi_k = \{x_j \to y_m \to \cdots y_2 \to y_1 \to \dot{x}_i\}$. Mojtahedzadeh now considers each possible pathway (9) and defines the "dominant" pathway as the one with the largest numerical value

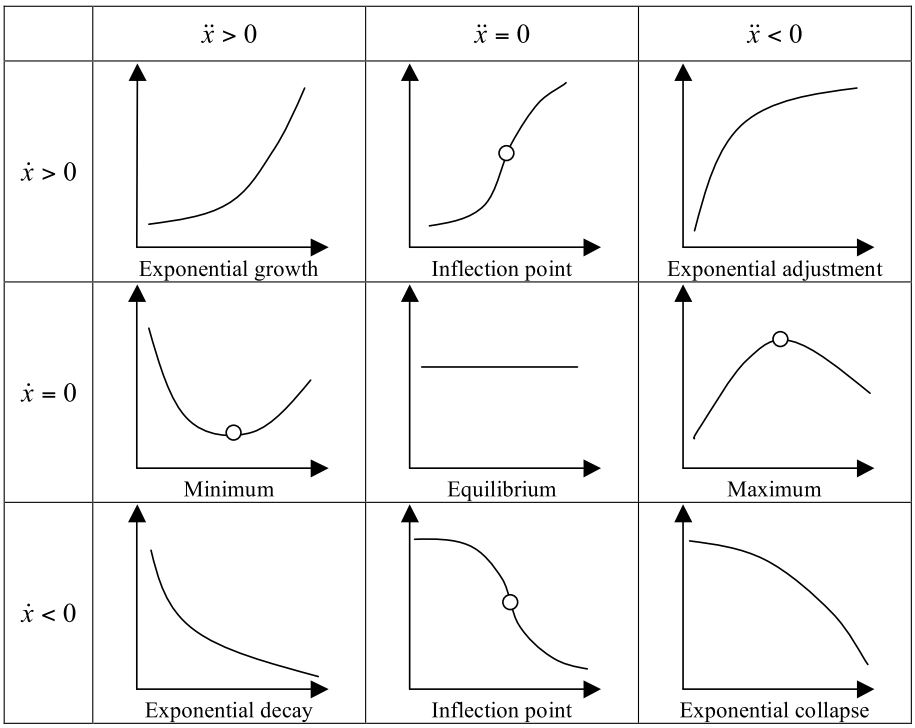

System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 5
Characteristic behavior patterns based on the first and second time derivatives

and the same sign as PPM_i. Having selected this dominant pathway, $\pi^*_{ij} = \{x_j \to y_m \to \cdots y_2 \to y_1 \to \dot{x}_i\}$, which originates in the state variable x_j the procedure is repeated for that state variable x_j, and so forth, until one either reaches one of the already "visited" state variables (in which case a loop has been found) or an exogenous variable (in which case an external driving force has been found). Thus, the procedure may result in three alternative forms of dominant structure illustrated in Fig. 6, namely a "pure" minor or major feedback loop, a pathway from a feedback loop elsewhere in the system, or a pathway from an exogenous variable.

By dividing the observed model behavior into different phases according to the taxonomy in Fig. 5 and then applying the method just described at different points in during these phases, one can reveal how the dominant structure changes over time. For illustration, the PPM method is applied to the Bass model and the results are presented in Fig. 7. The figure shows the metrics of four alternative pathways (four feedback loops) and the results accord nicely with the informal analysis done earlier: The method identifies two phases, exponential growth, exponential adjustment, and identifies the "word-of-mouth" positive loop (loop 1) as dominant in the first phase and the "exhaustion" loop (loop 2) as dominant in the second phase.

The PPM method is still mostly used at an early explorative stage on rather simple models, where it does appear to aid insight into the dynamics (e. g. [41]), and has been implemented in a software package, *Digest*, [35].

From the studies performed so far, it is clear that the main strength of this method is its relative computational simplicity (it does not require computing eigenvalues, which is a numerically demanding task), and the intuitive and direct connection it makes between the observed behavior and the influencing structural elements. Unlike the other approaches which operate in the "frequency domain", the method considers the time path of a specific variable directly.

There are, however, some important outstanding issues that remain to be clarified. First, the method is not suitable for oscillatory systems. The problem is easy to recognize when one considers how the PPM measure will vary over the course of a sinusoidal outcome: The sign of the PPM will shift twice during each cycle, indicating that the behavior is alternately dominated by positive and negative loops, even though the system structure, and hence the loop dominance, may remain unchanged all the time.

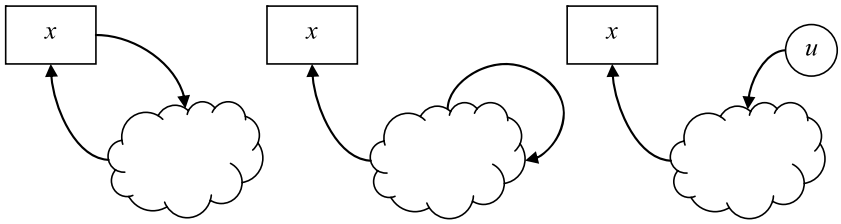

System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 6
Three alternative forms of dominant structure in the PPM method

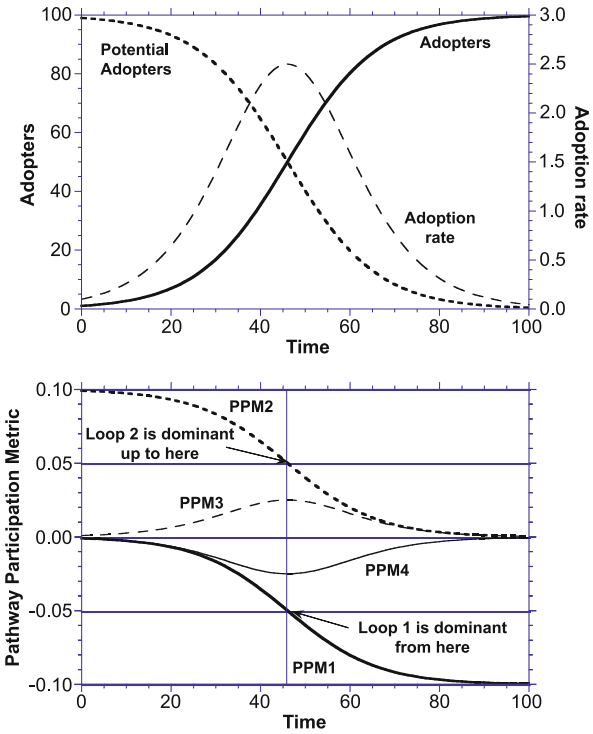

System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 7
Pathway participation measures in the Bass model

Richardson [46] already alluded to this problem by noting that the measure only considers the diagonal elements in the system matrix in (4), yet we know that the structure causing oscillation is the major negative loop that involve the off-diagonal elements. This is a significant limitation, given the prevalence and importance of oscillation in system dynamics analysis.

A second limitation of the current implementation of PPM is that it uses a depth-first search for the single most influential pathway for a variable. This strategy does not capture the situation where more than one structure may contribute significantly to the model behavior and, through the depth-first algorithm, may miss alternative paths that could prove to yield a larger total value of the metric. This problem could be addressed by modifying the search algorithm and is most likely of minor importance.

Another issue is how to treat the case when $\dot{x} = 0$ since it appears in the denominator of the terms in (6). However, it is not clear that it is necessary to do this division, given that it is easy to identify the nine cases in the figure by simply examining its sign. Thus, the issue is probably not of much significance.

The fourth issue, on the other hand, is more significant, namely the emphasis on identifying a single "dominant" structure. In reality, of course, the behavior of a variable is influenced by many loops and pathways at once. Reducing the consideration to a single one of these may miss important features of the structure-behavior relationships. For instance, a variable may be influenced by two negative loops and one positive, with the sum of the two negative loops dominating the influence of the positive loop, even though that loop by itself has the strongest influence on the behavior. It is more appropriate to consider the relative importance of alternative pathways, yet the method does not address how one would partition the behavior among pathways (the three structures in Fig. 6) – only among individual links.

Thus, while the notion of pathways seems an interesting and useful idea, it may be that it will ultimately be more effective to use a list, ranked in order of magnitude, of the pathways that influence a variable.

Finally, the method shares a weakness with the traditional method in that it considers primarily partial system structures rather than global system properties. In contrast, the two eigenvalue methods to which we now turn are based on a rigorous characterization of the entire system (at a given point in time).

Eigenvalue Elasticity Analysis

The third method may be termed *eigenvalue elasticity analysis* (or EEA for short) and builds upon the tools

from modern linear systems theory (LST), applied to the linearized model (4). The method is concerned with the structural elements that significantly affect the system eigenvalues or behavior modes – the values λ in (5). Specifically, it measures influence by the elasticity of an eigenvalue λ with respect to some parameter g in the model, defined as $\varepsilon = (\partial\lambda/\partial g)(g/\lambda)$, i. e. the fractional change in the eigenvalue relative to the fractional change in the parameter. The advantage of this fractional measure is that it is dimensionless, i. e., independent upon the choice of units, including the time scale unit. Sometimes, the influence measure is used instead, defined as $\mu = (\partial\lambda/\partial g)g$. This measure has dimension $[1/time]$ and so depends upon the choice of is time unit, but it is generally easier to interpret for complex-valued eigenvalues and avoids numerical problems with very small or zero eigenvalues (see [29,54]).

The idea behind EEA was first introduced in system dynamics by Forrester [14] in the context of economic stabilization policy. For purposes of policy analysis in oscillating systems, one may define a number of criteria from engineering control theory, all of which relate to the eigenvalues of the system, as summarized in Table 2. Figure 8 provides a graphical characterization of the eigenvalues and policy criteria in the complex plane. Though these measures are not new, the EEA method is unique in its attempt to use them to gain qualitative intuitive understanding of the system. A significant step in this direction was first suggested by Forrester [15] with the notion that the elasticities of any links in the model (corresponding to elements of the matrix $\boldsymbol{A}$ in the linearized system (4)), can be interpreted as the sum of elasticities of all feedback loops containing that link. We have chosen to name this approach *loop eigenvalue elasticity analysis (LEEA)*.

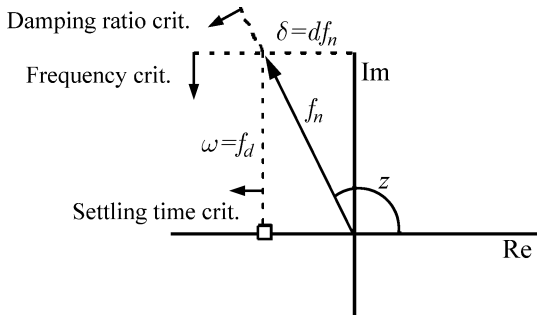

System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 8
Characterization of eigenvalues plotted in the complex plane

Kampmann [28] provided a rigorous definition of LEEA and also pointed to the fact that feedback loops are not independent. In other words, given the possibly very large number of loops in a given model (Kampmann demonstrated how the theoretical maximum number of loops grows combinatorically with the number of variables), it only makes sense to speak of individual contributions of a limited set of *independent* loops. He proved that a fully connected system (where there is a feedback loop between any pair of variables – the typical case in system dynamics models) with N links and n variables has a total of $N - n + 1$ independent loops and provided a procedure for constructing this set and calculating the loop elasticities.

Kampmann's analysis points to a fundamental issue relating to the notion of feedback loops as a way to explain

System Dynamics, Analytical Methods for Structural Dominance Analysis in, Table 2
Stabilization policy criteria and corresponding effects on eigenvalues and BDW of a policy change in a system element *g*

Policy Criterion	Description	Change in eigenvalue $\lambda = \delta \pm i\omega, \omega > 0$	Change in BDW w	Appropriate measure in time path
Damping	Increases the rate of decay of oscillation (or decreases the rate of expansion)	$\frac{\partial\delta}{\partial g}\frac{g}{\delta} < 0$	N/A	$\frac{x(t+T)}{x(t)}$
Frequency	Decreases the frequency of oscillation (or lengthens the period T)	$\frac{\partial\omega}{\partial g}\frac{g}{\omega} < 0$	N/A	T
Variance	Reduces the variance of a target variable (or the weighted average variances of several variables)	No simple relation	$\frac{\partial w}{\partial g}\frac{g}{w} < 0$	$\int x(t)^2\,dt$
Auto-spectrum	Reduces variance of target variable(s) within a target frequency range	No simple relation	$\frac{\partial w}{\partial g}\frac{g}{w} < 0$	Filter in frequency domain
Frequency response gain	Reduces the gain (amplification) in the target frequency range for a particular combination of disturbance exogenous and output variables.	Based upon transfer function $G(i\omega)$		

System Dynamics, Analytical Methods for Structural Dominance Analysis in, Table 3
Loops and their influences in the inventory workforce model. Values are measured at time $t = 0$. The model contains three eigenvalues, $\lambda_1 = -0.250$ and $\lambda_2, \lambda_3 = -0.138 \pm i\,0.285$. The influence measure is defined as $g \cdot \partial\lambda/\partial g$. For the imaginary part, a positive influence measure means that the frequency is increased

Loop	Nodes	Gain	Influence on Re[λ_1]	Influence on Re[λ_2]	Influence on Im[λ_2]
1	ED > CED	−0.250	−0.250	0.000	0.000
2	W > HFR	−0.200	0.000	−0.100	−0.022
3	INV > IC > DP > SP > EO > P	−0.076	0.000	−0.038	0.008
4	INV > IC > DP > DW > HFR > W > NP > P	−0.100	0.000	0.000	0.176
5	INV > IC > DP > DW > HFR > W > NP > SP > EO > P	0.015	0.000	0.000	−0.027

behavior: the significance assigned to a particular loop depends upon the context (the chosen independent loop set). In other words, feedback loops are derived and relative concepts rather than fundamental independent building blocks of systems. Oliva [40] further refined the definition of independent loop sets by introducing the *Shortest independent loop set (SILS)* along with a procedure for constructing the set. Although a SILS is not generally unique, experience seems to suggest that it is easier to interpret [41]. Yet the issue remains that independent feedback loop sets are relative concepts.

In Table 3, we show how the LEEA analysis applies to the simple inventory–workforce model in Fig. 3. The model contains a total of 5 feedback loops, all of which are independent. The loops are listed in Table 3, including their constituent variables (nodes), and the gain of the loop (defined in a similar manner to the pathway participation metrics above). We see that there are three minor negative loops, related to the exponential smoothing of expected demand (loop 1) and the adjustment of workforce to desired workforce (loop 2). The minor loop 3 is the "overtime shortcut" that allows production to adjust part way to desired production immediately so one does not have to wait for the workforce to adjust. Loop 4 is the main major negative loop that adjusts inventory to desired levels via workforce adjustment. Finally, loop 5 (the only positive loop) is a fairly weak loop that moderates the effect of loop 4 by adjusting the overtime effect "back to normal" when the workforce is brought in line with desired production.

Although the model is nonlinear (due to the overtime function), the eigenvalues do not change very much over the course of its behavior. The model contains one real eigenvalue ($\lambda_1 = -0.250$) and one pair of complex conjugate eigenvalues ($\lambda_2, \lambda_3 = -0.138 \pm i\,0.285$). The first eigenvalue corresponds to the adjustment of expected demand (ED). The other pair produces a damped oscillation in inventory and workforce.

Table 3 also shows the loop influences upon the three eigenvalues. Note how there is a one-to-one correspondence between loop 1 (the adjustment of expected demand) and the first eigenvalue. This is due to fact that the ED level constitutes a single strongly connected component of the model (see Fig. 3), i. e. there is no feedback between this level and the rest of the model. We also note that the workforce adjustment and the overtime loops have a stabilizing influence upon the behavior (they make the real part of the oscillatory eigenvalues more negative and have relative little effect upon the frequency of oscillation). Conversely, the major negative loop 4 has a destabilizing influence, since strengthening it will increase the frequency of oscillation and not increase the damping. The effects of loop 5 are fairly weak.

From this analysis, one would therefore expect parameters that strengthen loop 2 (shortening hire/fire time) or loop 3 (increase overtime effect) would stabilize the system while strengthening loop 4 (shorter inventory adjustment time) will destabilize the system. Indeed this is what happens, as illustrated in the simulations in Fig. 9.

The EEA/LEEA method has been applied in a number of contexts (e. g. [1,20,22,24,29,51,52,54]), but remains a tool employed only by specialists in fundamental research, not least because it has not been incorporated into standard software packages. Thus, the potential of the method for widespread practice remains unexplored.

One might be skeptical that a method derived from linear systems theory may have any use for the nonlinear models found in system dynamics. Kampmann and Oliva [29] considered what types of models the method would be particularly suited for. They defined three categories of models, based upon the behavior they are designed to exhibit: 1) linear and quasilinear models, 2) nonlinear single-transient models, and 3) nonlinear periodic models. The first category encompasses models of oscillations, possibly combined with growth trends, with relatively stable equilibrium points, (e. g., the classical in-

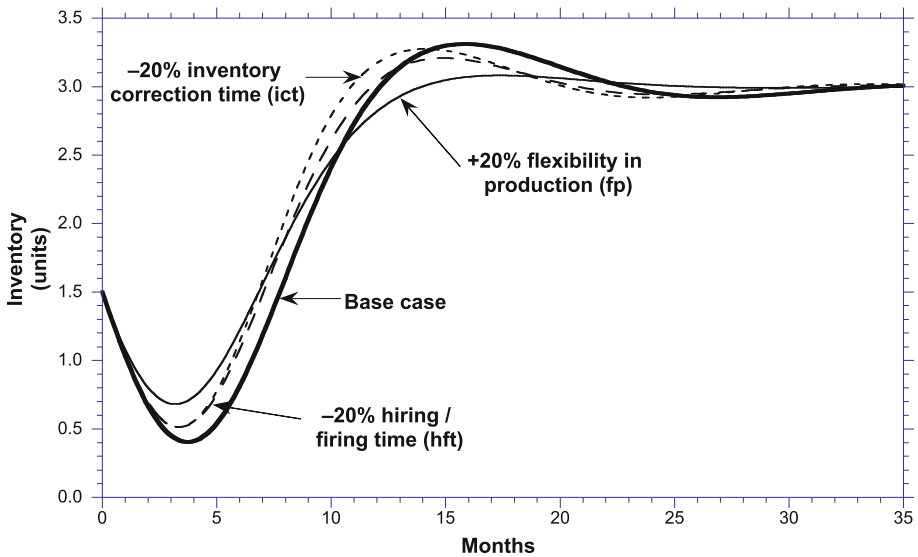

System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 9
Simulated behavior of inventory–workforce model, showing the effect on inventory of parameter changes for overtime (flexibility of production), inventory adjustment time (ict), and labor hiring/firing time (hft), respectively

dustrial dynamics models [11]). Nonlinearities may modify behavior (particularly responses to extreme shocks) but the instabilities and growth trends can be analyzed in terms of linear relationships. Kampmann and Oliva concluded that LEEA showed the most promise and potential for this class of models because the analytical foundations are solid and valid, and because the method has the ability to find high-elasticity loops even in large models very quickly without much intervention on the part of the analyst.

The second class is typical of scenario models like the World Model [13,33], the Urban Dynamics Model [12], or the energy transition model in [57], to name a few, that show a single transient behavior pattern, like overshoot and collapse or a turbulent transition to a new equilibrium. In these models, nonlinearities usually play an essential role in the dynamics. Yet it is possible to divide the behavior into distinct phases where certain loops tend to dominate the behavior. In this class of models LEEA also shows promise by measuring shifts in structural dominance by the change in elasticities. But it requires more input from the analyst (e. g. in defining the different phases of the transition) and it has no obvious advantage over other methods, like PPM.

The third class, nonlinear periodic models, are those that exhibit fluctuating behavior in which nonlinearities play an essential role, such as like limit cycles, quasiperiodic behavior, or chaos, (see, e. g. [48]). Here the utility of the method is much less clear and depends upon the specifics of the model in question. For example, the classic Lorenz model that exhibits limit cycles, period doubling and deterministic chaos does not lend itself to any insight using LEEA [29]. This is particularly the case in systems with strong nonlinearities such as min and max functions. In these systems, the behavior may change abruptly (eigenvalues suddenly shift) in what is called border-collision bifurcations [37,61]. In other cases, the method of breaking the behavior into phases with different dominant structures may yield significant insight from LEEA. For instance, Sterman's simple long wave model [58] lends itself well to this approach (e. g. [24,28]).

In the present paper, we add a fourth category of models or behavior for which the method has not been explored yet. We name this category nonlinear multi-modal models. These encompass the cases where one behavior mode interacts with and therefore modifies another behavior mode – something that can only happen in nonlinear systems. The most common example is mode-locking or entrainment, in which oscillations become synchronized (e. g. [25]). Another example is mode modification, where one behavior mode (growth or oscillation) affects the character of another (typically oscillation). An example of this is the interaction of the business cycle with the economic long wave, where the former tends to get more severe during long wave downturns [16]. Whether LEEA can contribute to this class of models remains to be seen.

Compared to the former two methods, the EEA/LEEA is mathematically more general and rigorous, though

many of the mathematical issues in the method remain to be addressed, as we summarize below. This rigor is also the main strength of the method, since it provides an unambiguous and complete measure of the influence of the entire feedback structure on all behavior modes.

A weakness or challenge that is starting to show up is the computational intensity in calculating eigenvalues and elasticities. This is not so much an issue of computer time and memory space as of the stability of numerical methods. Kampmann and Oliva [29] found that the numerical method used sometimes proved unstable, yielding meaningless results. Clearly, there is a need to explore this issue further, possibly building upon the developments in control engineering.

A more serious weakness is the difficulty in interpreting the results: Eigenvalues do not directly relate to the observed behavior of a particular variable. The concepts of eigenvalues and elasticities are rather abstract and unintuitive [10]. There is a need for tools and methods that can translate them into visible, visceral, and salient measures. Here, the measures in Table 2 may provide a guide. In particular, it is possible to use (linear) filtering in the frequency domain to define a behavior of interest. For example, an analyst may be concerned with structures causing a typical business cycle (3–4-year oscillation) and, by specifying a filter that "picks out" that range of fluctuation, could obtain measures for structures that have elasticities in that range. Because filters are typically linear operators, all the analytical machinery of the LEEA method will also apply in this case – a significant advantage.

Using filters will also solve an issue that appears in large-scale models, namely the presence of several identical or nearly identical behavior modes. Saleh et al. [54] do consider the analytical problems associated with repeated eigenvalues, where it becomes necessary to use generalized eigenvectors, and where other behavior modes appear involving power functions of time. A filter essentially constitutes a weighted average of behavior modes and in this fashion avoids the "identity problem" of non-distinct eigenvalues.

The most serious theoretical issue, in our view, is how the results are interpreted using the feedback loop concept. As mentioned, the concept is relative (to a choice of an independent loop set). Moreover, practice reveals that the number of loops to consider is rather large and that the loops elasticities often do not have an easy or intuitive explanation. A lot of care must be taken when interpreting the results. For instance, Kampmann and Oliva [29] found that "phantom loops" – loops that cancel each other by logical necessity and are essentially artifacts of the equation formulations used in the model – could nonetheless have large elasticities and thus seriously distort the interpretation of the results. An example of "phantom loops" is found in the Bass model in Fig. 1, where loops 3 and 4 are artifacts of the way the model is formulated. If the variable *Total population* (T) was eliminated from the equations, the loops would disappear and in fact they exactly cancel each other out (since T is constant). Nonetheless, they appear on the list of loops and appear to have a separate influence on behavior. These kinds of problems may not be intractable, but their resolution will require careful mathematical analysis.

Finally, a problem with EEA and LEEA is that it only considers changes to behavior modes, not the degree to which these modes are expressed in a system variable of interest. This issue is addressed by also considering the eigen*vectors* of the system, which is the foundation for the analysis in the next section.

Eigenvectors and Dynamic Decomposition Weights (DDW)

The last set of methods, which are still in early development, we have termed the *eigenvector-based* approach (EVA). EVA attempts to improve the EEA/LEEA method by considering how much an eigenvalue or behavior mode is expressed in a particular system variable. The logic of the method and how EEA and EVA complement each other is shown in Fig. 10. As shown by Kampmann [28], in a sense there is a one-to-one correspondence between eigenvalues and loop gains whereas the eigenvectors arise from the remaining "degrees of freedom" in the system. The observed behavior of the state variables in the model is then the combined outcome of the behavior modes (from the loop gains) and the weights for each mode (from the eigenvectors) in the respective state variable.

A number of researchers have attempted to develop EVA methods. Some emphasize the curvature (second time derivative) of the behavior, similar to the starting point of the PPM method [24,50,51,52]. The slope or rate of change $\dot{x}(t)$ of a given variable x in the linearized system may be written by

$$\dot{x}(t-t_0) = w_1 \exp(\lambda_1 (t-t_0)) + \cdots + w_n \exp(\lambda_n (t-t_0)) \,, \quad (10)$$

where the weights w_i are related to the eigenvectors. Then the curvature at time t_0 is

$$\ddot{x}(t_0) = w_1\lambda_1 + \cdots + w_n\lambda_n \,. \quad (11)$$

One may therefore interpret (11) has the sum of contribution from individual behavior modes. Güneralp [24]

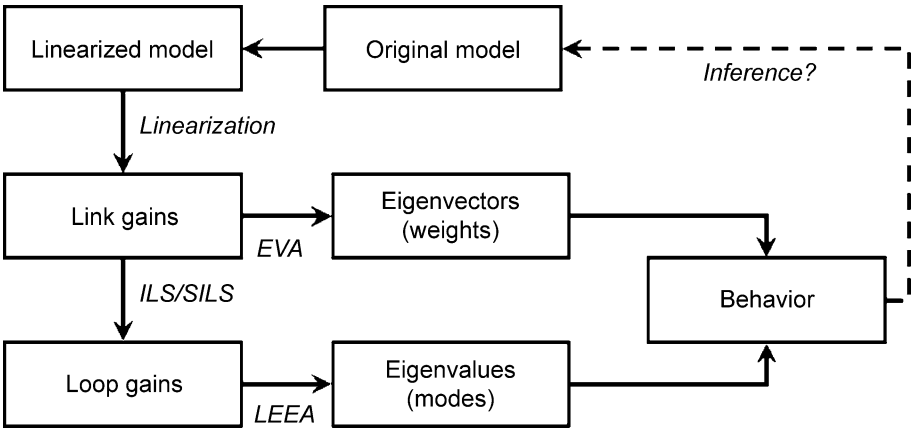

System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 10
Schematic view of eigenvalue and eigenvector analysis approach

suggested using the terms on the right-hand side of (11) as weights to combine elasticities of individual behavior modes ε_i with respect to some system element (like a link gain or a loop gain) into a weighted sum

$$\bar{\varepsilon} = \frac{\sum_{i=1}^{n} w_i \lambda_i \varepsilon_i}{\sum_{i=1}^{n} |w_i \lambda_i|} , \tag{12}$$

as a measure of the overall significance of that system element. He further normalized the elasticity measure by the elasticity measure for other system elements, i. e., assuming there are K such elements (loops or links), the relative importance ρ_k of the kth element is defined as

$$\rho_k = \frac{\bar{\varepsilon}_k}{\sum_{j=1}^{K} |\bar{\varepsilon}_j|} , \tag{13}$$

with the motivation that elasticities may vary greatly in numerical values, making comparisons at different points in time difficult, whereas ρ_k is a relative measure varying between +1 and −1. His results shed an alternative light on the behavior of these models, but the mathematical meaning, consistency and significance of the doubly normalized measure (13) remains to be clarified. It is still too early to tell what the most useful approach will be, but one may note that the emphasis on the curvature shares the basic weakness in the PPM approach in dealing with oscillations.

Other researchers have looked directly at the *dynamic decomposition weights (DDW)* w_i in (10), i. e., the relative weight of the modes for a particular variable, from a policy criterion perspective, similar to Forrester's original focus and the starting point for the EEA analysis [21,53,54].

For instance, Saleh et al. [54] look at how alternative stabilization policies affect the behavior of business cycle models, using both a simple inventory–workforce model [59], and a more extensive model based on Mass [32] and used in the LEEA analysis of Kampmann and Oliva [29]. Using the procedure in Fig. 9, they decompose the net stabilizing effect of a policy into its effect on the behavior mode itself (LEEA) and its effect on the expression of that mode in the variable of interest, measured the dynamic decomposition weights (EVA or DDW).

To illustrate the approach we perform the computations for the inventory–workforce model (Fig. 3 and Table 1). We find that the following equations describe the behavior of the state variables

$$\begin{aligned} \mathrm{ED} &= 1 - 0.500\mathrm{e}^{-0.250t} \\ \mathrm{INV} &= 3 - 2.167\mathrm{e}^{-0.250t} \\ &\quad + 1.134\mathrm{e}^{-0.138t} \sin(2.945 + 0.285t) \\ \mathrm{WF} &= 1 + 0.669\mathrm{e}^{-0.250t} \\ &\quad - 1.169\mathrm{e}^{-0.138t} \sin(1.553 - 0.285t) \ . \end{aligned} \tag{14}$$

As expected from the structure of the model, the behavior of *Expected Demand* does not have an oscillatory component and only shows a short transient exponential adjustment for the stock to match *Demand*. On the other hand, *Inventory* and *Workforce*, in addition to having the transient behavior to reach equilibrium captured by the first eigenvalue, have an oscillatory component represented by the second eigenvalue. Note that each state variable has a different *Dynamic Decomposition Weight* (w) for each

System Dynamics, Analytical Methods for Structural Dominance Analysis in, Table 4
Elasticity to parameters of weight of eigenvalue 2 (−0.138 + *i* 0.85) on inventory and influence of parameters on eigenvalue 2 – inventory–workforce model

Parameter	w_2 on INV Elasticity	Influence on Re[λ_2]	Influence on Im[λ_2]
Demand	2.000	0.000	0.000
Inventory correction time	0.656	0.038	−0.157
Flexibility in production	0.364	−0.549	−0.266
Productivity	−0.353	0.000	0.000
Time to change expectations	−0.240	0.000	0.000
Normal inventory coverage	0.239	0.000	0.000
Hiring/Firing time	0.238	0.100	−0.127
Disequilibrium fraction	0.000	0.000	0.000

reference mode, i. e., each eigenvalue contributes differently to the overall behavior of each state variable.

An exploration of the policy design space can be achieved by assessing the influence of model parameters on the dynamic decomposition weight. By focusing on the weights of the behavior modes for the variable of interest we can identify leverage points to increase or decrease the presence of a behavior mode in the variable. The weight elasticity column in Table 4 reports the parameter elasticity of w_2 (the weight of eigenvalue 2, the oscillatory behavior mode) on *Inventory* ($\varepsilon_w = (\mathrm{d}w/\mathrm{d}p)(p/w)$). The magnitude of the elasticity quantifies the impact that changes in the parameter value have on the weight of the oscillatory behavior model on *Inventory*. The table is sorted in descending order of absolute value of elasticity.

Changes in parameters, however, not only impact the behavior decomposition weights, but also change the eigenvalues themselves. This dual impact of parameter changes introduces a challenge in developing policy recommendations. The last two columns of Table 4 report the influence on the eigenvalue (real and imaginary part) for each parameter. These measures of influence should be interpreted in a similar way as the weight elasticities. The influence measure is defined as $\mu_\lambda = (\partial\lambda/\partial p)\, p$. A positive real-part measure indicates that increasing the parameter will destabilize the system by lengthening the settling time and vice-versa. A positive imaginary-part measure indicates that increasing the parameter will increase the frequency of oscillation – normally considered a destabilizing influence – and vice-versa.

Five parameters, *demand, disequilibrium fraction*, *productivity, normal inventory coverage*, and *time to change expectations*, have no influence on the oscillatory behavior mode. *Demand* and *disequilibrium fraction* are initialization constants that do not participate in any of the feedback loops in the model. *Productivity* is essentially a scaling measure having to do with the definition of units of labor and goods in the model. Redefining units should not affect the dynamics of the model. While *time to change expectations* is involved in loop 1, it does not participate in the oscillatory behavior observed in the model since, as discuss above, *Expected Demand* is in a separate strongly connected component of the model.

In accordance with LEEA, the *flexibility in production* parameter, which strengthens overtime loop 3 (cf. Fig. 3), has a strong stabilizing influence, by both increasing the damping and lowering the frequency of the oscillatory mode. Likewise, as predicted by LEEA, a shorter *hiring/firing time* will increase damping by strengthening the labor adjustment loop 5 but, again in accordance with LEEA, also increases the frequency of adjustment because it also strengthens the major loop 4. Finally, lowering the *inventory correction time* will strengthen the link from *inventory* to *desired production*, and consequently the three loops 3, 4 and 5, with the net effect that although the adjustment is a little faster (a more negative real part), the frequency is also increased significantly, i. e., it is a less effective way of stabilizing the system (cf. Fig. 9).

As an alternative approach, Fig. 11 shows what happens to the frequency response of the state variables (*Inventory* INV, *Workforce* WF, and *Expected Demand* ED) when the parameter Hiring/Firing Time (hft) is reduced by 2% from 5 to 4.9. There are a number of things to notice in the figure. First, there is no effect whatsoever on the ED variable, which should not be surprising, given that there is no feedback to this variable from the rest of the system. Second, the effect on the amplitude, like the amplitude itself, is strongly dependent upon the frequency of variation. We see that there is a significant amount of dampening on the *Inventory* fluctuation around the resonant frequencies in the range 0.1 to 0.3. On the other hand, there is a small amount of amplification of inventory in the higher frequency ranges. The effect on *Workforce* is very different: though there is a small attenuation in the reso-

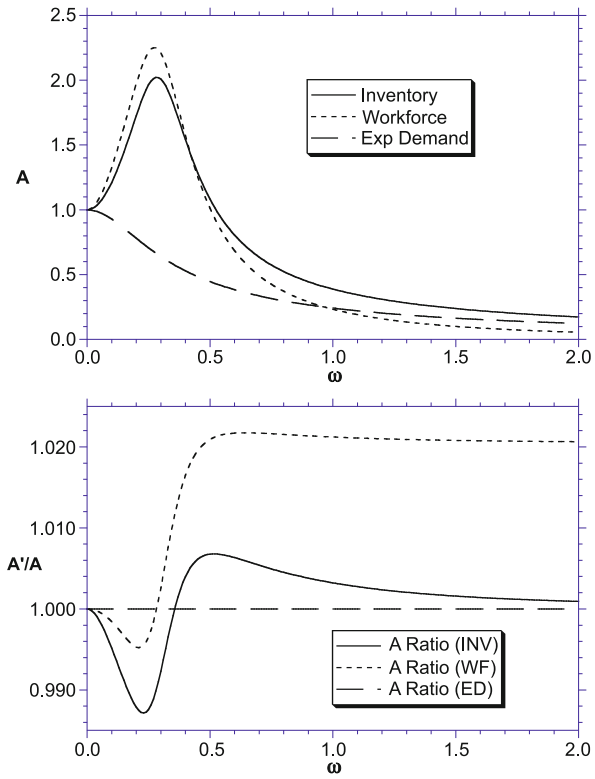

System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 11

Effect on frequency response of the inventory workforce model of reducing the parameter Hiring/Firing Time (hft) from 5.0 to 4.9. The diagram shows the gain of the base case (*upper graph*) for the three state variables, and the resulting change in the gain, measured as the ratio (*A′/A*) from the parameter change (*lower graph*)

nant frequencies, there is a significant increase in variance in the higher frequency range. In other words, although the LEEA analysis showed a faster hiring policy to be stabilizing (by strengthening loop 2, cf. Table 3), the DDW analysis shows that it depends – both upon the variable in question and the context (frequency of variation).

Future Directions

As mentioned above, it is not possible to construct a complete theory that will automatically provide modelers with "the" dominant structure. Given the analytical intractability of nonlinear high-order systems found in our field, the most we can hope for is a set of tools that will guide the analysis and aid the development of the modeler's intuition.

That said, however, we are left with an impression that the analytical foundation for these tools is in need of further development before one rushes into implementing them into software packages. We are quite satisfied with the current state of affairs in this regard, where code, models, and documentation are made freely to download (most of the cited papers provide a URL to their code and models). Understanding *how* and *why* the tools work the way they do is crucial, and this will require that a number of puzzles, uncertainties, and technical problems be addressed. Only then will the time come to submit the methods for wider application to test their real-world utility.

While the classical method remains a useful intuitive guide and teaching tool for graduate students, there are no signs that it may be developed further. (That said, it is possible that the classical control transfer function method may be employed in the eigensystem approaches to explore nested canonical systems, though this is purely speculative). The pathway method would benefit from a firmer mathematical foundation. In particular, it would be important to compare how its results and conclusions compare to those found in the LST. It is possible that the pathway method may eventually be merged with the LST approaches as a subset of a general analytical toolbox. We believe that there is a great deal of promise in combining the eigenvalue and eigenvector analysis in the LST approaches. This combination will yield a complete system characterization and an understanding of both how particular feedback loops are involved in generating a behavior mode, and how system elements determine the expression of that behavior mode in a particular variable. A unified LST approach along the lines suggested in Fig. 10 thus seems within reach.

It will probably be a while, however, before these methods will find their way into widely available and use-friendly software packages. Apart from the theoretical issues alluded to above, a number of technical issues related to numerical calculations, various "pathological cases" (such as non-distinct eigenvalues), and special cases of feedback loops ("figure-eight" loops, for instance), will need to be addressed.

On the more creative side, it would be interesting to explore alternative forms of visualizing the various influence measures developed. For instance, one could imagine that links between variables in a model diagram "glow" in different colors and intensities depending upon their effect on a behavior pattern in question. This is not just a question of fancy user interfaces: as mentioned in the introduction, the function of these tools will be as intuitive consistent aids to understanding, not analytical "answering machines". In this light, the visualization is as important as the analytical principles behind it. Given the power of the human eye in finding patterns in visual data, this could be a significant next step.

Bibliography

1. Abdel-Gawad A, Abdel-Aleem B, Saleh M, Davidsen P (2005) Identifying dominant behavior patterns, links and loops: Automated eigenvalue analysis of system dynamics models. Proceedings of the Int System Dynamics Conference, Boston, July 2005. System Dynamics Society, Albany
2. Barlas Y (1989) Multiple Tests for Validation of System Dynamics Type of Simulation Models. Eur J Oper Res 42(1):59–87
3. Bass FM (1969) A new product growth for model consumer durables. Manag Sci 15(5):215–227
4. Chen CT (1970) Introduction to Linear System Theory. Holt, Rinnehart and Winston, New York
5. Eberlein R (1984) Simplifying Models by Retaining Selected Behavior Modes. Ph D Thesis, Sloan School of Management, MIT, Cambridge
6. Eberlein RL (1986) Full Feedback Parameter Estimation. In: Proceedings of the Int Systems Dynamics Conference, Sevilla, Spain. System Dynamics Society, Albany, pp 69–83
7. Eberlein RL (1989) Simplification and understanding of models. Syst Dyn Rev 5(1):51–68
8. Eberlein RL, Wang Q (1985) Statistical Estimation and System Dynamics Models. Proceedings of the Int Systems Dynamics Conference, Keystone, USA. System Dynamics Society, Albany, pp 206–222
9. Ford A, Flynn H (2005) Statistical screening of system dynamics models. Syst Dyn Rev 21(4):273–303
10. Ford DN (1999) A Behavioral Approach to Feedback Loop Dominance Analysis. Syst Dyn Rev 15(1):3–36
11. Forrester JW (1961) Industrial Dynamics. Productivity Press, Cambridge
12. Forrester JW (1969) Urban Dynamics. Productivity Press, Cambridge
13. Forrester JW (1971) World Dynamics. Productivity Press, Cambridge
14. Forrester N (1982) A Dynamic Synthesis of Basic Macroeconomic Policy: Implications for Stabilization Policy Analysis. Ph D Thesis, Sloan School of Management, MIT, Cambridge
15. Forrester N (1983) Eigenvalue analysis of dominant feedback loops. Proceedings of the Int System Dynamics Conference, Chestnut Hill, USA. System Dynamics Society, Albany
16. Forrester JW (1993) System Dynamics and the Lessons of 35 Years. In: DeGreene KB (ed) Systems-Based Approach to Policymaking. Kluwer, Norwell, pp 199–240
17. Forrester JW, Senge PM (1980) Tests for Building Confidence in System Dynamics Models. TIMS Stud Manag Sci 14:209–228
18. Forrester JW, Mass NJ, Ryan CJ (1976) The System Dynamics National Model: Understanding Socio-Economic Behavior and Policy Alternatives. Technol Forecast Soc Chang 9(1–2):51–68
19. Goldberg DE (1989) Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading
20. Gonçalves P (2003) Demand bubbles and phantom orders in supply chains. Ph D Thesis, Sloan School of Management, MIT, Cambridge
21. Gonçalves P (2008) Behavior modes, pathways and overall trajectories: Eigenvalue and eigenvector analysis in system dynamics. Syst Dyn Rev, forthcoming
22. Gonçalves P, Lerpattarapong C, Hines JH (2000) Implementing formal model analysis. Proceedings of the Int System Dynamics Conference, Bergen, Norway, August 2000. System Dynamics Society, Albany
23. Graham AK (1977) Principles on the Relationship Between Structure and Behavior of Dynamic Systems. Ph D Thesis, Sloan School of Management, MIT, Cambridge
24. Güneralp B (2006) Towards Coherent Loop Dominance Analysis: Progress in Eigenvalue Elasticity Analysis. Syst Dyn Rev 22(3):263–289
25. Haxholdt C, Kampmann CE, Mosekilde E, Sterman JD (1995) Mode Locking and Entrainment of Endogenous Economic Cycles. Syst Dyn Rev 11(3):177–198
26. Holland JH (1992) Adaptation in Natural and Artificial Systems. MIT Press, Cambridge
27. Homer JB (1983) Partial-Model Testing as a Validation Tool for System Dynamics. In: Proceedings of the Int System Dynamics Conference, Chestnut Hill, USA, July 1983. System Dynamics Society, Albany, pp 920–932
28. Kampmann CE (1996) Feedback Loop Gains and System Behavior (unpublished manuscript). In: Proceedings of the Int System Dynamics Conference, Cambridge, USA, July 1996. System Dynamics Society, Albany, pp 260–263
29. Kampmann CE, Oliva R (2006) Loop Eigenvalue Elasticity Analysis: Three Case Studies. Syst Dyn Rev 22(2):146–162
30. Lane DC, Smart C (1996) Reinterpreting 'generic structure': evolution, application and limitations of a concept. Syst Dyn Rev 12(2):87–120
31. Luenberger DG (1979) Introduction to Dynamic Systems: Theory, Models and Applications. Wiley, New York
32. Mass NJ (1975) Economic Cycles: An Analysis of Underlying Causes. Productivity Press, Cambridge
33. Meadows DH, Meadows DL, Randers J, Behrens III WW (1972) The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind. Universe Books, New York
34. Mojtahedzadeh MT (1996) A path taken: Computer-assisted heuristics for understanding dynamic systems. Ph D Thesis, Rockefeller College of Pubic Affairs and Policy, State University of New York at Albany, Albany
35. Mojtahedzadeh MT, Andersen D, Richardson GP (2004) Using Digest to implement the pathway participation method for detecting influential system structure. Syst Dyn Rev 20(1):1–20
36. Morecroft JDW (1985) Rationality in the Analysis of Behavioral Simulation Models. Manag Sci 31(7):900–916
37. Mosekilde E, Laugesen JL (2006) Nonlinear Dynamic Phenomena in the BEER Model. Department of Physics, The Technical University of Denmark, Kongens Lyngby
38. Ogata K (1990) Modern Control Engineering, 2nd edn. Prentice Hall, Englewood Cliffs
39. Oliva R (2003) Model Calibration as a Testing Strategy for System Dynamics Models. Eur J Operat Res 151(3):552–568
40. Oliva R (2004) Model Structure Analysis Through Graph Theory: Partition Heuristics and Feedback Structure Decomposition. Syst Dyn Rev 20(4):313–336
41. Oliva R, Mojtahedzadeh M (2004) Keep it simple: Dominance assessment of short feedback loops. Proceedings of the Int System Dynamics Conference, Oxford, UK, July 2004. System Dynamics Society, Albany
42. Ott E (1993) Chaos in Dynamical Systems. Cambridge University Press, New York

43. Peterson DW (1980) Statistical Tools for System Dynamics. In: Randers J (ed) Elements of the System Dynamics Method. Productivity Press, Cambridge, pp 224–241
44. Peterson DW, Eberlein RL (1994) Reality Checks: A Bridge Between Systems Thinking and System Dynamics. Syst Dyn Rev 10(2/3):159–174
45. Radzicki MJ (2004) Expectation Formation and Parameter Estimation in Uncertain Dynamical Systems: The System Dynamics Approach to Post Keynesian-Institutional Economics. Proceedings of the Int System Dynamics Conference, Oxford, UK, July 2004. System Dynamics Society, Albany
46. Richardson GP (1984/1995) Loop Polarity, Loop Dominance, and the Concept of Dominant Polarity. Syst Dyn Rev 11(1):67–88
47. Richardson GP (1986) Dominant structure. Syst Dyn Rev 2(1):68–75
48. Richardson GP (ed) (1988) System Dynamics Review. Chaos Special Issue 4:1–2
49. Richmond B (1980) A new look at an old friend. Plexus, Resource Policy Center, Thayer School of Engineering, Dartmouth College, Hanover
50. Saleh M (2002) The characterization of model behavior and its causal foundation. Ph D Thesis, Dept of Information Science, University of Bergen, Bergen
51. Saleh M, Davidsen P (2001) The origins of behavior patterns. Proceedings of the Int System Dynamics Conference, Atlanta, July 2001. System Dynamics Society, Albany
52. Saleh M, Davidsen P (2001) The origins of business cycles. Proceedings of the Int System Dynamics Conference, Atlanta, July 2001. System Dynamics Society, Albany
53. Saleh M, Oliva R, Davidsen P, Kampmann CE (2006) Eigenvalue Analysis of System Dynamics Models: Another Perspective. Proceedings of the Int System Dynamics Conference, Neijmegen, The Netherlands, July 2006. System Dynamics Society, Albany
54. Saleh M, Oliva R, Davidsen P, Kampmann CE (2008) A comprehensive analytical approach for policy analysis of system dynamics models. Mays Business School, Texas A&M University, Working Paper
55. Schweppe F (1973) Uncertain Dynamical Systems. Prentice-Hall, Englewood Cliffs
56. Senge PM (1990) The Fifth Discipline: The Art & Practice of the Learning Organization. Doubleday Currency, New York
57. Sterman JD (1981) The Energy Transition and the Economy: A System Dynamics Approach. Ph D Thesis, Sloan School of Management, MIT, Cambridge
58. Sterman JD (1985) A Behavioral Model of the Economic Long Wave. J Econ Behav Org 6(1):17–53
59. Sterman JD (2000) Business dynamics: Systems thinking and modeling for a complex world. Irwin McGraw-Hill, Boston
60. Wolstenholme E (2004) Using generic system archetypes to support thinking and modelling. Syst Dyn Rev 20(4):341–356
61. Zhusubaliyev ZT, Mosekilde E (2003) Bifurcation and Chaos in Piecewise-Smooth Dynamical Systems. World Scientific, Singapore

System Dynamics, The Basic Elements of

George P. Richardson
Rockefeller College of Public Affairs and Policy, University at Albany, State University of New York, Albany, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Feedback Thinking
Loop Dominance and Nonlinearity
The Endogenous Point of View
System Structure
Suggestions for Further Reading on the Core of System Dynamics
Bibliography

Glossary

Endogenous Generated from within. Contrasting with "exogenous," meaning generated by forces external to a system or point of view.

Feedback loop A closed path of causal influences and information, forming a circular-causal loop of information and action.

System dynamics System dynamics is a computer-aided approach to theory-building, policy analysis and strategic decision support emerging from an endogenous point of view.

Definition of the Subject

System dynamics is a computer-aided approach to theory-building, policy analysis, and strategic decision support emerging from an endogenous point of view [18,20]. It applies to dynamic problems arising in complex social, managerial, economic, or ecological systems – literally any dynamic systems characterized by interdependence, mutual interaction, information feedback, and circular causality.

Introduction

The field of system dynamics developed initially from the work of Jay W. Forrester. His seminal book *Industrial Dynamics* [7] is still a significant statement of philosophy and methodology in the field. Within ten years of its publication, the span of applications grew from corporate and industrial problems to include the management of research

and development, urban stagnation and decay, commodity cycles, and the dynamics of growth in a finite world. It is now applied in economics, public policy, environmental studies, defense, theory-building in social science, and other areas, as well as its home field, management. The name industrial dynamics no longer does justice to the breadth of the field (for extensive examples, see [20,28], so it has become generalized to system dynamics. The modern name suggests links to other systems methodologies, but the links are weak and misleading. System dynamics emerges out of servomechanisms engineering, not general systems theory or cybernetics [18].

The system dynamics approach involves:

- Defining problems dynamically, in terms of graphs over time.
- Striving for an endogenous, behavioral view of the significant dynamics of a system, a focus inward on the characteristics of a system that themselves generate or exacerbate the perceived problem.
- Thinking of all concepts in the real system as continuous quantities interconnected in loops of information feedback and circular causality.
- Identifying independent stocks or accumulations (levels) in the system and their inflows and outflows (rates).
- Formulating a behavioral model capable of reproducing, by itself, the dynamic problem of concern. The model is usually a computer simulation model expressed in nonlinear equations, but is occasionally left unquantified as a diagram capturing the stock-and-flow/causal feedback structure of the system.
- Deriving understandings and applicable policy insights from the resulting model.
- Implementing changes resulting from model-based understandings and insights.

Mathematically, the basic structure of a formal system dynamics computer simulation model is a system of coupled, nonlinear, first-order differential (or integral) equations,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}(t) = \mathbf{f}(\mathbf{x}, \mathbf{p}) ,$$

where $\mathbf{x}$ is a vector of levels (stocks or state variables), $\mathbf{p}$ is a set of parameters, and $\mathbf{f}$ is a nonlinear vector-valued function. Such a system has been variously called a *state-determined system* in the engineering literature, an *absolute system* [3], an *equifinal system* [32], and a *dynamical system* [16].

Simulation of such systems is easily accomplished by partitioning simulated time into discrete intervals of length $\mathrm{d}t$ and stepping the system through time one $\mathrm{d}t$ at a time. Each state variable is computed from its previous value and its net rate of change $x'(t)$: $x(t) = x(t - \mathrm{d}t) + \mathrm{d}t \cdot x'(t - \mathrm{d}t)$. In the earliest simulation language in the field (DYNAMO) this equation was written with time scripts K (the current moment), J (the previous moment), and JK (the interval between time J and K): $X_\mathrm{K} = X_\mathrm{J} + \mathrm{DT} \cdot \mathrm{XRATE}_\mathrm{JK}$ (see, e. g., [22]). The computation interval $\mathrm{d}t$ is selected small enough to have no discernible effect on the patterns of dynamic behavior exhibited by the model. In more recent simulation environments, more sophisticated integration schemes are available (although the equation written by the user may look like this simple Euler integration scheme), and time scripts may not be in evidence. Important current simulation environments include STELLA and iThink (isee Systems, http://www.iseesystems.com/), Vensim (Ventana Systems, http://www.vensim.com/), and Powersim (http://www.powersim.com/).

Forrester's original work stressed a continuous approach, but increasingly modern applications of system dynamics contain a mix of discrete difference equations and continuous differential or integral equations. Some practitioners associated with the field of system dynamics work on the mathematics of such structures, including the theory and mechanics of computer simulation, analysis and simplification of dynamic systems, policy optimization, dynamical systems theory, and complex nonlinear dynamics and deterministic chaos.

The main applied work in the field, however, focuses on understanding the dynamics of complex systems for the purpose of policy analysis and design. The conceptual tools and concepts of the field – including feedback thinking, stocks and flows, the concept of feedback loop dominance, and an endogenous point of view – are as important to the field as its simulation methods.

Feedback Thinking

Conceptually, the feedback concept is at the heart of the system dynamics approach. Diagrams of loops of information feedback and circular causality are tools for conceptualizing the structure of a complex system and for communicating model-based insights. Intuitively, a feedback loops exists when information resulting from some action travels through a system and eventually returns in some form to its point of origin, potentially influencing future action. If the tendency in the loop is to reinforce the initial action, the loop is called a *positive* or *reinforcing* feedback loop; if the tendency is to oppose the initial action, the loop is called a *negative, counteracting*, or *balancing* feedback loop. The sign of the loop is called its *po-*

larity. Balancing loops can be variously characterized as goal-seeking, equilibrating, or stabilizing processes. They can sometimes generate oscillations, as when a pendulum seeking its equilibrium goal gathers momentum and overshoots it. Reinforcing loops are sources of growth or accelerating collapse; they are disequilibrating and destabilizing. Combined, balancing and reinforcing circular causal feedback loops can generate all manner of dynamic patterns.

Feedback loops are ubiquitous in human and natural systems and, under various names and representations, have been widely recognized in popular and scholarly literature. Feedback thought has been present implicitly or explicitly for hundreds of years in the social sciences and literally thousands of years in recorded history [9]. We have the vicious circle originating in classical logic and morphing into common usage, the bandwagon effect, the invisible hand of Adam Smith, Malthus's correct observation of population growth as a self-reinforcing process, Keynes's consumption multiplier, the investment accelerator of Hicks and Samuelson, compound interest or inflation, the biological concepts of proprioception and homeostasis, Festinger's cognitive dissonance, Myrdal's principle of cumulative causation, Venn's idea of a suicidal prophecy, Merton's related notion of a self-fulfilling prophecy, and so on. Each of these ideas can be concisely and insightfully represented as one or more loops of causal influences with positive or negative polarities. Great social scientists and feedback thinkers; great social theories are feedback thoughts. (For a full exposition of the evolution of the feedback concept see [19].)

Loop Dominance and Nonlinearity

The loop concept underlying feedback and circular causality by itself is not enough, however. The explanatory power and insightfulness of feedback understandings also rest on the notions of active structure and loop dominance. Complex systems change over time. A crucial requirement for a powerful view of a dynamic system is the ability of a mental or formal model to change the strengths of influences as conditions change, that is to say, the ability to shift *active* or *dominant structure*.

In a system of equations, this ability to shift loop dominance comes about endogenously from nonlinearities in the system. For example, the S-shaped dynamic behavior of the classic logistic growth model ($\mathrm{d}P/\mathrm{d}t = aP - bP^2$) or similar structures like the Gompertz curve ($\mathrm{d}P/\mathrm{d}t = aP - bP\ln(P)$) can be seen as the consequence of a shift in loop dominance from a positive, self-reinforcing feedback loop (aP) producing exponential-like growth, to a negative feedback loop ($-bP^2$ or $-bP\ln(P)$) that brings the system to its eventual goal. The shift in loop dominance in these models comes about from the nonlinearity in the second term, which grows faster than the first term and eventually overtakes it. Only nonlinear models can endogenously alter their active or dominant structure and shift loop dominance.

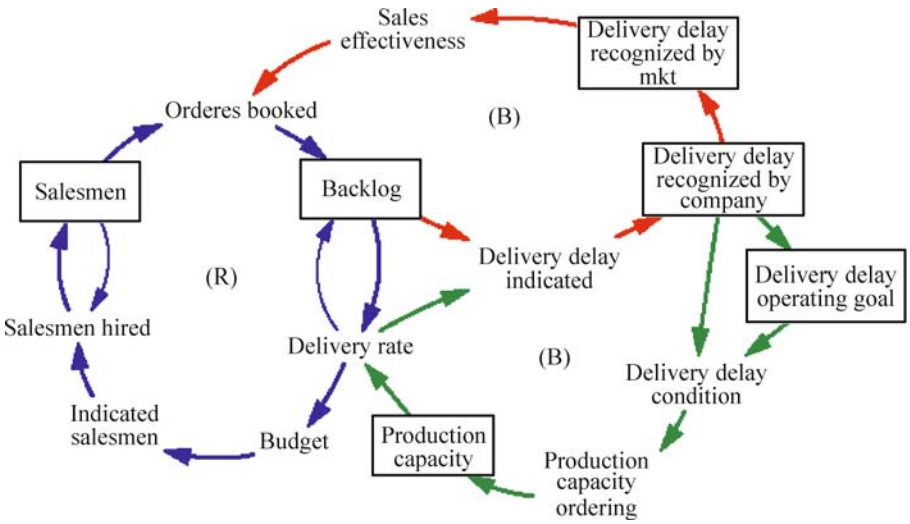

System Dynamics, The Basic Elements of, Figure 1
Core structure of Forrester's market growth model [8], showing a *blue reinforcing loop* underlying the growth (or reinforcing decline) of Salesmen, Orders, and Revenue, a *red balancing loop* containing various delayed recognitions of the company's delivery delay, and a *green balancing loop* responsible for capacity ordering if the delivery delay drops too far below its operating goal

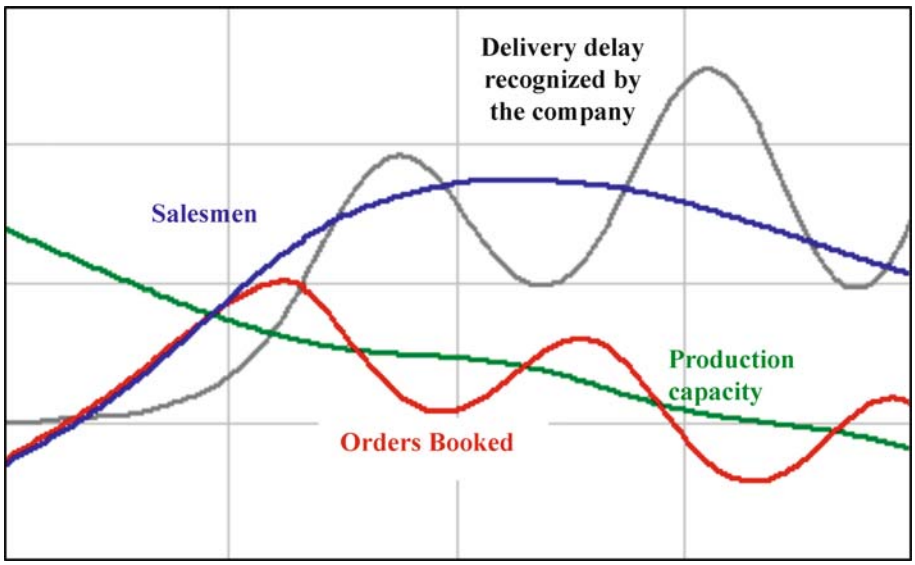

System Dynamics, The Basic Elements of, Figure 2
The dynamic behavior of the model shown in Fig. 1, illustrating an early growth phase, which turns into an oscillatory phase as the feedback loop dominance shifts to the *red balancing delivery delay loop*, and results in a long term corporate decline as the *green capacity ordering loop* responds to a sliding operating goal for the acceptable delivery delay

Real systems are perceived to change their active or dominant structure over time, often because of the build-up of internal forces. Thus from a feedback perspective, the ability of nonlinearities to generate shifts in loop dominance is the fundamental reason for advocating nonlinear models of social system behavior.

Figures 1 and 2, abstracted from an early, classic paper [8] illustrate these ideas. In Fig. 1 salesmen (in the blue reinforcing loop) book orders for the company; if enough revenue is generated, there is enough budget to hire more salesmen and corporate growth ensues. Whether salesmen (in this simplified picture) book enough orders depends on the company's delivery delay for the product, as perceived by the market (red balancing loop). The company builds production capacity according to its perceived need, as indicated by its perceived delivery delay and its target for that (green balancing loop).

Figure 2 shows the dynamics this feedback structure endogenously generates. In the early phase, salesmen grow as orders and revenue grow; the system's exponential growth behavior in that phase is generated by the reinforcing salesmen loop. But then the feedback loop dominance soon shifts to the balancing delivery delay loop, which constrains sales effectiveness and brings a halt to growth. The system moves into an oscillatory phase generated by the various monitoring and perception delays around the now dominant red balancing loop. Salesmen eventual peak and decline, as the green production capacity ordering loop fails to keep production capacity sufficient to hold the delivery delays in check.

Thus the dynamic behavior of this system is a consequence of its feedback structure and the nonlinearities that shift loop dominance endogenously over time. The particular decline scenario shown in Fig. 2 illustrates one of the deep insights of the model: the adaptive goal structure, in which the delivery delay operating goal moves slowly to accommodate changes in the company's delivery delay, weakens the green balancing loop trying to bring on capacity. The company never perceives its delivery delay is sufficiently higher than its (sliding) target, so it fails to order sufficient capacity to sustain growth. A fixed goal for the acceptable delivery delay sends a stronger signal, which can turn this corporate decline into oscillating growth [8].

Thus, nonlinearity is crucial to the system dynamics approach. However, it is crucial not merely because of its mathematical properties but because it enables the formalization of a profoundly powerful perspective on theory and policy – the *endogenous point of view*.

The Endogenous Point of View

The concept of endogenous change is fundamental to the system dynamics approach. It has both philosophical and engineering origins. A deep and lasting insight of the earliest attempts at servomechanisms control is the realization that *the attempt to control a system generates dynamics of*

its own, complicating the dynamics trying to be controlled. A governor mechanism imposed to control the speed of a steam engine can generate oscillatory "hunting behavior," as the control system overshoots and undershoots the set point. As it becomes part of the system, the governing mechanism thus generates dynamics of its own.

The insight transfers readily, but with added significance, from engineering systems to people systems: Attempts to control complex human systems – coercing, guiding, managing, governing – generate dynamics of their own. Moreover, some of these endogenously generated dynamics are created by the control mechanisms themselves (like the governor of a steam engine) and some are created by human creative responses to the management efforts (e. g., principal-agent interactions). These natural and human forces, creating counteracting and compensating pressures in response to system control efforts, emerge as complicated circular-causal feedback structures. The often complex, difficult-to-understand dynamics of such management systems are to a great degree a consequence of their internal structures.

To capture and analyze such management complexities, one must look inward to see the ways a complex system naturally responds to system pressures. The endogenous point of view is thus central to the system dynamics approach. It dictates aspects of model formulation: exogenous disturbances are seen at most as *triggers* of system behavior (like displacing a pendulum); the *causes* are contained within the structure of the system itself (like the interaction of a pendulum's position and momentum that produces oscillations). Corrective responses are also not modeled as functions of time, but are dependent on conditions within the system. Time by itself is not seen as a cause in the endogenous point of view.

Theory building and policy analysis are significantly affected by this endogenous perspective. Taking an endogenous view exposes the natural *compensating* tendencies in social systems that conspire to defeat many policy initiatives. Feedback and circular causality are delayed, devious, and deceptive. For understanding, system dynamics practitioners strive for an *endogenous point of view*. The effort is to uncover the sources of system behavior that exist within the structure of the system itself.

System Structure

These ideas are captured almost explicitly in Forrester's [9] organizing framework for system structure:

- Closed boundary
- Feedback loops
- Levels
- Rates
- Goal
- Observed condition
- Discrepancy
- Desired action.

The *closed boundary* signals the endogenous point of view. The word *closed* here does not refer to open and closed systems in the general system sense, but rather refers to the effort to view a system as *causally* closed. The modeler's goal is to assemble a formal structure that can, *by itself*, without exogenous explanations, reproduce the essential characteristics of a dynamic problem.

The causally closed system boundary at the head of this organizing framework identifies the endogenous point of view as the feedback view pressed to an extreme. Feedback thinking can be seen as a *consequence* of the effort to capture dynamics within a closed causal boundary. Without causal loops, all variables must trace the sources of their variation ultimately outside a system. Assuming instead that the causes of all significant behavior in the system are contained within some closed causal boundary forces causal influences to feed back upon themselves, forming causal loops. Feedback loops enable the endogenous point of view and give it structure.

Levels and Rates

Stocks (accumulations, or "levels" in early system dynamics literature) and the flows ("rates") that affect them are essential components of system structure. A map of causal influences and feedback loops is not enough to determine the dynamic behavior of a system. A constant inflow yields a linearly rising stock; a linearly rising inflow yields a stock rising along a parabolic path; a stock with inflow proportional to itself grows exponentially; two stocks in a balancing loop have a tendency to generate oscillations; and so on. For example, the boxes in Fig. 1 represent accumulations in the company and its market; the three stocks in the red balancing loop (the order backlog and the two perceptions of the company's delivery delay) give that loop its tendency to generate oscillations which propagate through out the system. Accumulations are the memory of a dynamic system and contribute to its disequilibrium and dynamic behavior.

Forrester [7] placed the operating policies of a system among its rates, the inflows and outflows governing change in the system. Many of these rates of change assume the classic structure of a negative feedback loop striving to take action to reduce the discrepancy between the observed condition of the system and a goal. The simplest

such rate structure results in an equation of the form

$$\text{RATE} = \frac{\text{GOAL} - \text{LEVEL}}{\text{ADJUSTMENT TIME}} ,$$

where ADJUSTMENT TIME is the time over which the level adjusts to reach the goal. This simple formulation reflects Forrester's more general statement about rates in his hierarchy of system structure (above) which can be richly thought of as

$$\text{RATE} = f(\text{DESIRED ACTION})$$

$$\text{DESIRED ACTION} = g(\text{DESIRED CONDITION}, \text{OBSERVED CONDITION})$$

$$\text{OBSERVED CONDITION} = h(\text{LEVELS}) ,$$

for some functions f, g, and h representing particular system characteristics.

Operating policies in a management system can influence the *flows* of information, material, and resources, which are the only means of changing the accumulations in the system. While flows can be changed quickly, as a matter of relatively quick decision making, stocks change slowly – they rise when inflows are great than outflows, and decline when inflows are less than outflows.

The simple "tub dynamics" of stocks are clear even to children, yet can be befuddling in complex systems. The accumulation of green house gases in the atmosphere, for example, affects the *flow* of heat energy radiated from the earth. To turn around global warming, the *accumulation* of green house gases must drop far enough to raise radiant energy above the inflow of solar energy, a simple stock-and-flow insight. But to cause the accumulation of green house gases to drop, their generation must fall below their natural absorption rate (another simple stock-and-flow observation). So turning around global warming is a process involving a chain of at least two significant accumulations, and people have trouble thinking it through reliably. The accumulations can only be changed by managing their associated flows. They will change only slowly even if we manage the technical and political pitfalls involved in lowering green house gas production (see [29]).

The significance of stocks in complex systems is vivid in a resource-based view of strategy and policy. Resources that enable a corporation or government to function or flourish are stocks, usually accumulated over long periods of time with significant investment of time, energy, and money. Reputations are also stocks, built over similarly long periods of time. While inadequate by themselves to give a full picture of the dynamics of a complex system, stocks and flows are vital components of system structure, without which fundamental understandings of dynamics are impossible [33].

Behavior is a Consequence of System Structure

The importance of stocks and flows appears most clearly when one takes a *continuous* view of structure and dynamics. Although a discrete view, focusing on separate events and decisions, is entirely compatible with an endogenous feedback perspective, the system dynamics approach emphasizes a continuous view [7]. The continuous view strives to look beyond events to see the dynamic patterns underlying them: model not the appearance of a discrete new housing unit in a city, but focus instead on the rise and fall of aggregate numbers of housing units. Moreover, the continuous view focuses not on discrete decisions but on the *policy structure* underlying decisions: not why this particular apartment building was constructed but what persistent pressures exist in the urban system that produce decisions that change housing availability in the city. Events and decisions are seen as surface phenomena that ride on an underlying tide of system structure and behavior. It is that underlying tide of policy structure and continuous behavior that is the system dynamicist's focus.

There is thus a *distancing* inherent in the system dynamics approach – not so close as to be confused by discrete decisions and myriad operational details, but not so far away as to miss the critical elements of policy structure and behavior. Events are deliberately blurred into dynamic behavior. Decisions are deliberately blurred into perceived policy structures. Insights into the connections between system structure and dynamic behavior, which are the goal of the system dynamics approach, come from this particular distance of perspective.

Suggestions for Further Reading on the Core of System Dynamics

The *System Dynamics Review*, the journal of the System Dynamics Society, published by Wiley, is the best source of current activity in the field, including methodological advances and applications.

The core of a vibrant field is difficult to discern in the flow of current work. However, the works that the field itself singles out as exemplary can give some reliable hints about what is considered vital to the core. In this sense two edited volumes are noteworthy: An early, interesting collection of applications is Roberts [24]; Richardson [21] is a more recent two-volume edited collection in

the same spirit, containing prize-winning work in philosophical background, dynamic decision making, applications in the private and public sectors, and techniques for modeling with management.

In addition, the following works, selected from among winners of the System Dynamics Society's *Jay Wright Forrester Award* (see www.systemdynamics.org/Society_Awards.htm), can be considered insightful although implicit exemplars of the core of system dynamics. (Publications are listed beginning with the most recent; see the bibliography for full citations):

- Thomas S. Fiddaman, "Exploring policy options with a behavioral climate-economy model"
- Kim D. Warren, *Competitive Strategy Dynamics*
- Eric F. Wolstenholme, "Towards the Definition and Use of a Core Set of Archetypal Structures in System Dynamics"
- Nelson P. Repenning, "Understanding Fire Fighting in New Product Development"
- John D. Sterman, *Business Dynamics, Systems Thinking and Modeling for a Complex World*
- Peter Milling, "Modeling innovation processes for decision support and management simulation."
- Erling Moxnes, "Not Only the Tragedy of the Commons: Misperceptions of Bioeconomics."
- Jac A. M. Vennix, *Group Model Building: Facilitating Team Learning Using System Dynamics*
- Jack B. Homer, "A System Dynamics Model of National Cocaine Prevalence."
- Andrew Ford, "Estimating the Impact of Efficiency Standards on Uncertainty of the Northwest Electric System."
- Khalid Saeed, *Towards Sustainable Development: Essays on System Analysis of National Policy*
- Tarek Abdul-Hamid and Stuart Madnick, *Software Project Dynamics: An Integrated Approach*
- George P. Richardson, *Feedback Thought in Social Science and Systems Theory*
- Peter M. Senge, *The Fifth Discipline*
- John D. W. Morecroft, "Rationality in the Analysis of Behavioral Simulation Models."
- John D. Sterman, "Modeling Managerial Behavior: Misperceptions of Feedback in a Dynamic Decision Making Experiment."

For texts on the system dynamics approach, see Alfeld and Graham [2], Richardson and Pugh [22], Wolstenholme [34], Ford [6], Maani and Cavana [11], and the most comprehensive text to date, Sterman [28].

Bibliography

1. Abdul-Hamid T, Madnick S (1991) Software project dynamics: an integrated approach. Prentice Hall, Englewood Cliffs
2. Alfeld LE, Graham AK (1976) Introduction to urban dynamics. Pegasus Comunications, Waltham
3. Ashby H (1956) An Introduction to cybernetics. Chapman & Hall, London
4. Fiddaman TS (2002) Exploring policy options with a behavioral climate-economy model. Syst Dyn Rev 18(2):243–267
5. Ford A (1990) Estimating the impact of efficiency standards on uncertainty of the northwest electric system. Oper Res 38(4):580–597
6. Ford A (1999) Modeling the environment: an introduction to system dynamics of environmental systems. Island Press, Washington
7. Forrester JW (1961) Industrial dynamics. MIT Press, Cambridge. Reprinted by Pegasus Communications, Waltham
8. Forrester JW (1968) Market growth as influenced by capital investment. Ind Manag Rev (MIT, now Sloan Manag Rev) 9(2):83–105. Reprinted widely, e. g., Richardson [21]
9. Forrester JW (1969) Urban dynamics. MIT Press, Cambridge. Reprinted by Pegasus Communications, Waltham
10. Homer JB (1992) A system dynamics model of national cocaine prevalence. Syst Dyn Rev 9(1):49–78
11. Maani KE, Cavana RY (2000) Systems thinking and modelling: understanding change and complexity. Pearson Education, New Zealand
12. Milling P (1996) Modeling innovation processes for decision support and management simulation. Syst Dyn Rev 12(3):211–234
13. John DW, Morecroft JDW (1985) Rationality in the analysis of behavioral simulation models. Manag Sci 31(7):900–916
14. Morecroft JDW, Sterman JD (eds) (1994) Modeling for learning organizations. System dynamics series. Pegasus Communications, Waltham
15. Moxnes E (1998) Not only the tragedy of the commons: misperceptions of bioeconomics. Manag Sci 44(9):1234–1248
16. Nicholis G, Prigogine I (1977) Self-organization in nonequilibrium systems: from dissipative structures to order through fluctuations. Wiley, New York
17. Repenning NR (2001) Understanding fire fighting in new product development. J Prod Innov Manag 18(5):285–300
18. Richardson GP (1991) System dynamics: simulation for policy analysis from a feedback perspective. In: Fishwick PA, Luker PA (eds) Qualitative simulation modeling and analysis. Springer, New York
19. Richardson GP (1991) Feedback thought in social science and systems theory. University of Pennsylvania Press, Philadelphia. Reprinted by Pegasus Communications, 1999
20. Richardson GP (1996) System dynamics. In: Gass S, Harris C (eds) The encyclopedia of operations research and management science. Kluwer, New York
21. Richardson GP (ed) (1996) Modelling for management: simulation in support of systems thinking. International library of management. Dartmouth, Aldershot
22. Richardson GP, Pugh AL III. (1981) Introduction to system dynamics modeling with DYNAMO. MIT Press, Cambridge. Reprinted by Pegasus Communications, Waltham
23. Richmond B (1993) Systems thinking: critical thinking skills for the 1990s and beyond. Syst Dyn Rev 9:(2)113–133

24. Roberts EB (ed) (1978) Managerial applications of system dynamics. MIT Press, Cambridge. Reprinted by Pegasus Communications, Waltham
25. Saeed K (1991) Towards sustainable development: essays on system analysis of national policy. Progressive Publishers, Lahore
26. Senge PM (1990) The fifth discipline: the art and practice of the learning organization. Doubleday/Currency, New York
27. Sterman JD (1988) Modeling managerial behavior: misperceptions of feedback in a dynamic decision making experiment. Manag Sci 35(3):321–339
28. Sterman JD (2001) Business dynamics: systems thinking and modeling for a complex world. Irwin McGraw-Hill, Boston
29. Sterman JD, Sweeney LB (2002) Cloudy skies: assessing public understanding of global warming. Syst Dyn Rev 18(2):207–240
30. System Dynamics Review. 1985–present. Wiley, Chichester
31. Vennix JAM (1996) Group model building: facilitating team learning using system dynamics. Wiley, Chichester
32. von Bertalanffy L (1968) General systems theory: foundations, development, applications. George Braziller, New York
33. Warren KD (2002) Competitive strategy dynamics. Wiley, United Kingdom
34. Wolstenholme EF (1990) System enquiry: a system dynamics approach. Wiley, Chichester
35. Wolstenholme EF (2003) Towards the definition and use of a core set of archetypal structures in system dynamics. Syst Dyn Rev 19(1):7–26

System Dynamics in the Evolution of the Systems Approach

Markus Schwaninger
Institute of Management, University of St. Gallen,
St. Gallen, Switzerland

Article Outline

Glossary
Definition of the Subject
Introduction
Emergence of the Systems Approach
Common Grounds and Differences
The Variety of Systems Methodologies
System Dynamics –
Its Features, Strengths and Limitations
Actual and Potential Relationships
Outlook
Appendix
Bibliography

Glossary

Cybernetics The science of communication and control in complex, dynamical systems. The core objects of study are information, communication, feedback and adaptation. In the newer versions of cybernetics, the emphasis is on observation, self-organization, self-reference and learning.

Dynamical system The dynamical system concept is a mathematical formalization of time-dependent processes. Examples include the mathematical models that describe the swinging of a clock pendulum, the flow of water in a river, and the evolution of a population of fish in a lake.

Law of requisite variety Ashby's law of requisite variety says: "Only variety can destroy variety". It implies that the varieties of two interacting systems must be in balance, if stability is to be achieved.

Organizational cybernetics The science which applies cybernetic principles to organization. Synonyms are *Management Cybernetics* and *Managerial Cybernetics*.

System There are many definitions of *system*. Two examples: A portion of the world sufficiently well defined to be the subject of study; something characterized by a structure, for example, a social system (Anatol Rapoport). A system is a family of relationships between its members acting as a whole (International Society for the Systems Sciences).

System dynamics A methodology and discipline for the modeling, simulation and control of dynamic systems. The main emphasis falls on the role of structure and its relationship with the dynamic behavior of systems, which are modeled as networks of informationally closed feedback loops between stock and flow variables.

Systems approach A perspective of inquiry, education and management, which is based on system theory and cybernetics.

System theory A formal science of the structure, behavior, and development of systems. In fact there are different system theories. General system theory is a transdisciplinary framework for the description and analysis of any kind of system. System theories have been developed in many domains, e. g., mathematics, computer science, engineering, sociology, psychotherapy, biology and ecology.

Variety A technical term for *complexity* which denotes the number of (potential) states of a system.

Definition of the Subject

The purpose of this chapter is to give an overview of the role of system dynamics (SD) in the context of the evolution of the systems movement. This is necessary because SD is often erroneously taken as the systems approach as such, not as part of it. It is also requisite to show that the

processes of the evolution of both SD in particular and the systems movement as a whole are intimately linked and intertwined. Finally, in view of the purpose of the chapter the actual and potential relationships between system dynamics and the other strands of the systems movement are evaluated. This way, complementarities and synergies are identified.

Introduction

The purpose of this contribution is to give an overview of the role of system dynamics in the context of the evolution of the systems movement. "Systems movement" – often referred to briefly as "systemics" – is a broad term, which takes into account the fact that there is no single system approach, but a range of different ones. The common denominator of the different system approaches in our day is that they share a worldview focused on complex dynamic systems, and an interest in describing, explaining and designing or at least influencing them. Therefore, most of the system approaches offer not only a theory but also a way of thinking ("systems thinking" or "systemic thinking") and a methodology for dealing with systemic issues or problems.

System dynamics (SD) is a discipline and a methodology for the modeling, simulation and control of complex, dynamic systems. SD was developed by MIT professor Jay W. Forrester (e. g. [20,21]) and has been propagated by his students and associates. SD has grown to a school of numerous academics and practitioners all over the world. The particular approach of SD lies in representing the issues or systems-in-focus as meshes of closed feedback loops made up of stocks and flows, in continuous time and subject to delays.

The development of the system dynamics methodology and the worldwide community that applies SD to modeling and simulation in radically different contexts suggest that it is a "systems approach" on its own. Nevertheless, taking "system dynamics" as the (one and only) synonym for "systemic thinking" would be going too far, given the other approaches to systemic thinking as well as a variety of system theories and methodologies, many of which are complementary to SD. In any case, however, the SD community has become the strongest "school" of the Systems approach, if one takes the numbers of members in organizations representing the different schools as a measure (by 2006, the System Dynamics Society had more than 1000 members).

The rationale and structure of this contribution is as follows. Starting with the emergence of the systems approach, the multiple roots and theoretical streams of systemics are outlined. Next, the common grounds and differences among different strands of the systems approach are highlighted, and the various systems methodologies are explored. Then the distinctive features of SD are analyzed. Finally comes a reflection on the relationships of SD with the rest of the systems movement as well as with potential complementarities and synergies.

In Table 1, a time-line overview of some milestones in the evolution of the systems approach in general and System Dynamics in particular is given. Elaborating on each of the sources quoted therein would reach beyond the purpose of this chapter. However, to convey a synoptic view, a diagram showing the different systems approaches and their interrelationships is provided in the Appendix "Systems Approaches – An Overview".

Emergence of the Systems Approach

The systems movement has many roots and facets, with some of its concepts going back as far as ancient Greece. What we name as "the systems approach" today materialized in the first half of the twentieth century. At least two important components should be mentioned: those proposed by von Bertalanffy and by Wiener.

Ludwig von Bertalanffy, an American biologist of Austrian origin, developed the idea that organized wholes of any kind should be describable and, to a certain extent, explainable, by means of the same categories, and ultimately by one and the same formal apparatus. His *general systems theory* triggered a whole movement which has tried to identify invariant structures and mechanisms across different kinds of organized wholes (for example, hierarchy, teleology, purposefulness, differentiation, morphogenesis, stability, ultrastability, emergence, and evolution).

In 1948 Norbert Wiener, an American mathematician at the Massachusetts Institute of Technology, published his seminal book on *Cybernetics*, building upon interdisciplinary work carried out in cooperation with Bigelow, an IBM engineer, and Rosenblueth, a physiologist. Wiener's opus became the transdisciplinary foundation for a new science of capturing as well as designing control and communication mechanisms in all kinds of dynamic systems [81]. Cyberneticists have been interested in concepts such as information, communication, complexity, autonomy, interdependence, cooperation and conflict, self-production ("autopoiesis"), self-organization, (self-) control, self-reference and (self-) transformation of complex dynamic systems.

Along the genetic line of the tradition which led to the evolution of General Systems Theory (von Berta-

lanffy, Boulding, Gerard, Miller, Rapoport) and Cybernetics (Wiener, McCulloch, Ashby, Powers, Pask, Beer), a number of roots can be identified, in particular:

- Mathematics (for example, Newton, Poincaré, Lyapunov, Lotka, Volterra, Rashevsky)
- Logic (for example, Epimenides, Leibniz, Boole, Russell and Whitehead, Goedel, Spencer-Brown)
- Biology, including general physiology and neurophysiology (for example, Hippocrates, Cannon, Rosenblueth, McCulloch, Rosen)
- Engineering and computer science, including the respective physical and mathematical foundations (for example, Heron, Kepler, Watt, Euler, Fourier, Maxwell, Hertz, Turing, Shannon and Weaver, von Neumann, Walsh)
- Social and human sciences, including economics (for example, Hume, Adam Smith, Adam Ferguson, John Stuart Mill, Dewey, Bateson, Merton, Simon, Piaget).

In this last-mentioned strand of the systems movement, one focus of inquiry is on the role of feedback in communication and control in (and between) organizations and society, as well as in technical systems. The other focus of interest is on the multidimensional nature and the multilevel structures of complex systems. Specific theory building, methodological developments and pertinent applications have occurred at the following levels:

- Individual and family levels (for example, systemic psychotherapy, family therapy, holistic medicine, cognitive therapy, reality therapy)
- Organizational and societal levels (for example, managerial cybernetics, organizational cybernetics, sociocybernetics, social systems design, social ecology, learning organizations)
- The level of complex (socio-)technical systems (systems engineering)

The notion of "socio-technical systems" has become widely used in the context of the design of organized wholes involving interactions of people and technology (for instance, Linstone's multi-perspectives-framework, known by way of the mnemonic TOP (*T*echnical, *O*rganizational, *P*ersonal/individual).

As can be noted from these preliminaries, different kinds of system theory and methodology have evolved over time. One of these is a theory of dynamic systems by Jay W. Forrester, which serves as a basis for the methodology of system dynamics. Two eminent titles are [20] and [21]. In SD, the main emphasis falls on the role of structure and its relationship with the dynamic behavior of systems, modeled as networks of informationally closed feedback loops between stock and flow variables. Several other mathematical systems theories have been elaborated, for example, mathematical general systems theory (Klir, Pestel, Mesarovic and Takahara), as well as a whole stream of theoretical developments which can be subsumed under the terms "dynamic systems theory" or "theories of nonlinear dynamics" (for example, catastrophe theory, chaos theory and complexity theory). Under the latter, branches such as the theory of fractals (Mandelbrot), geometry of behavior (Abraham), self-organized criticality (Bak), and network theory (Barabasi, Watts) are subsumed. In this context, the term "sciences of complexity" is used.

In addition, a number of mathematical theories, which can be called "system theories," have emerged in different application contexts, examples of which are discernible in the following fields:

- Engineering, namely information and communication theory (Shannon and Weaver), technology and computer-aided systems theory (for example, control theory, automata, cellular automata, agent-based modeling, artificial intelligence, cybernetic machines, neural nets)
- Operations research (for example, modeling theory and simulation methodologies, Markov chains, genetic algorithms, fuzzy control, orthogonal sets, rough sets)
- Social sciences, economics in particular (for example, game theory, decision theory)
- Biology (for example, Sabelli's Bios theory of creation)
- Ecology (for example, E. and H. Odum's systems ecology).

Most of these theories are transdisciplinary in nature, i. e., they can be applied across disciplines. The Bios theory, for example is applicable to clinical, social, ecological and personal settings [54]. Examples of essentially non-mathematical system theories can be found in many different areas of study, e. g.:

- Economics, namely its institutional/evolutionist strand (Veblen, Myrdal, Boulding, Dopfer)
- Sociology (for example, Parsons' and Luhmann's social system theories, Hall's cultural systems theory)
- Political sciences (for example, Easton, Deutsch, Wallerstein)
- Anthropology (for example, Levi Strauss's structuralist-functionalist anthropology, Margaret Mead)
- Semiotics (for example, general semantics (Korzybski, Hayakawa, Rapoport), cybersemiotics (Brier))
- Psychology and psychotherapy (for example, systemic intervention (Bateson, Watzlawick, F. Simon), and fractal affect logic (Ciompi))

- Ethics and epistemology (for example, Vickers, Churchman, von Foerster, van Gigch)

Several system-theoretic contributions have merged the quantitative and the qualitative in new ways. This is the case for example in Rapoport's works in game theory as well as general systems theory, Pask's conversation theory, von Foerster's cybernetics of cybernetics (second-order cybernetics), and Stafford Beer's opus in managerial cybernetics. In all four cases, mathematical expression is virtuously connected to ethical, philosophical, and epistemological reflection. Further examples are Prigogine's theory of dissipative structures, Mandelbrot's theory of fractals, complex adaptive systems (Holland et al.), Kauffman's complexity theory, and Haken's synergetics, all of which combine mathematical analysis and a strong component of qualitative interpretation.

A large number of systems methodologies, with the pertinent threads of systems practice, have emanated from these theoretical developments. Many of them are expounded in detail in specialized encyclopedias (e. g., [27] and, under a specific theme, named *Systems Science and Cybernetics*, of the Encyclopedia of Life Support Systems [18]). In this chapter, only some of these will be addressed explicitly, in order to shed light on the role of SD as part of the systems movement.

Common Grounds and Differences

Even though the spectrum of system theories and methodologies outlined in the preceding section may seem multifarious, all of them have a strong common denominator: They build on the idea of systems as organized wholes. An objectivist working definition of a system is that of a whole, the organization of which is made up by interrelationships. A subjectivist definition is that of a set of interdependent variables in the mind of an observer, or, a mental construct of a whole, an aspect that has been emphasized by the position of constructivism. *Constructivism* is a synonym for *second-order cybernetics*. While first-order cybernetics concentrates on regulation, information and feedback, second-order cybernetics focuses on observation, self-organization and self-reference. Heinz von Foerster established the distinction between 'observed systems' for the former and 'observing systems' for the latter [74].

From the standpoint of operational philosophy, a system is, as Rapoport says, "a part of the world, which is sufficiently well defined to be the object of an inquiry or also something, which is characterized by a structure, for example, a production system" [50].

In recent systems theory, the aspect of relationships has been emphasized as the main building block of a system, as one can see from a definition published by the International Society for the Systems Sciences (ISSS): "A system is a family of relationships between its members acting as a whole" [63]. Also, purpose and interaction have played an important part in reflections on systems: Systems are conceived, in the words of Forrester [21], as "wholes of elements, which cooperate towards a common goal." Purposeful behavior is driven by internal goals, while purposive behavior rests on a function assigned from the outside. Finally, the aspects of open and closed functioning have been emphasized. Open systems are characterized by the import and export of matter, energy and information. A variant of particular relevance in the case of social systems is the operationally closed system, that is, a system which is self-referential in the sense that its self-production (autopoiesis) is a function of production rules and processes by which order and identity are maintained, and which cannot be modified directly from outside. As we shall see, this concept of operational closure is very much in line with the concept of circularity used in SD.

At this point, it is worth elaborating on the specific differences between two major threads of the systems movement, which are of special interest because they are grounded in "feedback thought" [52]: The cybernetic thread, from which organizational cybernetics has emanated, and the servomechanic thread in which SD is grounded. As Richardson's detailed study shows, the strongest influence on cybernetics came from biologists and physiologists, while the thinking of economists and engineers essentially shaped the servomechanic thread. Consequently, the concepts of the former are more focused on the adaptation and control of complex systems for the purpose of maintaining stability under exogenous disturbances. Servomechanics, on the other hand, and SD in particular, take an endogenous view, being mainly interested in understanding circular causality as the principal source of a system's behavior. Cybernetics is more connected with communication theory, the general concern of which can be summarized as how to deal with randomly varying input. SD, on the other hand, shows a stronger link with engineering control theory, which is primarily concerned with behavior generated by the control system itself, and by the role of nonlinearities. Managerial cybernetics and SD both share the concern of contributing to management science, but with different emphases and with instruments that are different but in principle complementary. Finally, the mathematical foundations are generally more evident in the basic literature on SD than in the writings on organizational cybernetics, in which the formal apparatus underlying model formulation is confined to a small number of publications (e. g., [7,10]),

which are less known than the qualitative treatises. The terms *management cybernetics* and *managerial cybernetics* are used as synonyms for *organizational cybernetics*.

The Variety of Systems Methodologies

The methodologies that have evolved as part of the systems movement cannot be expounded in detail here. The two epistemological strands in which they are grounded, however, can be identified – the positivist tradition and the interpretivist tradition.

Positivist tradition denotes those methodological approaches that focus on the generation of "positive knowledge," that is, a knowledge based on "positively" ascertained facts. *Interpretivist tradition* denotes those methodological approaches that emphasize the importance of subjective interpretations of phenomena. This stream goes back to Greek art and science of the interpretation and understanding of texts.

Some systems methodologies have been rooted in the positivist tradition, and others in the interpretivist tradition. The differences between the two can be described along the following set of polarities:

- An objectivist versus a subjectivist position
- A conceptual–instrumental versus a communicational/cultural/political rationality
- An inclination to quantitative versus qualitative modeling
- A structuralist versus a discursive orientation.

A positivistic methodological position tends toward the objectivistic, conceptual–instrumental, quantitative and structuralist–functionalist in its approach. An interpretive position, on the other hand, tends to emphasize the subjectivist, communicational, cultural, political, ethical and esthetic—that is, the qualitative and discursive aspects. It would be too simplistic to classify a specific methodology in itself as being "positivistic" or "interpretative". Despite the traditions they have grown out of, several methodologies have evolved and been reinterpreted or opened to new aspects (see below).

In the following, a sample of systems methodologies will be characterized and positioned in relation to these two traditions, beginning with those in the positivistic strand:

- *"Hard" OR methods*. Operations research (OR) uses a wide variety of mathematical and statistical methods and techniques—for example of optimization, queuing, dynamic programming, graph theory, time series analysis—to provide solutions for organizational and managerial problems, mainly in the operational domains of production and logistics, and in finance.
- *Living systems theory*. In his LST, James Grier Miller [44] identifies a set of 20 necessary components that can be discerned in living systems of any kind. These structural features are specified on the basis of a huge empirical study and proposed as the "critical subsystems" that "make up a living system." LST has been used as a device for diagnosis and design in the domains of engineering and the social sciences.
- *Viable system model*. To date, Stafford Beer's VSM is probably the most important product of organizational cybernetics. It specifies a set of management functions and their interrelationships as the sufficient conditions for the viability of any human or social system (see [10]). These are applicable in a recursive mode, for example, to the different levels of an organization. The VSM has been widely applied in the diagnostic mode, but also to support the design of all kinds of social systems. Specific methodologies for these purposes have been developed, for instance for use in consultancy. The term viable system diagnosis (VSD) is also used.

The methodologies and models addressed up to this point have by and large been created in the positivistic tradition of science. Other strands in this tradition do exist, e. g., systems analysis and systems engineering, which together with OR have been called "hard systems thinking" (p. 127 in [31]). Also, more recent developments such as mathematical complexity and network theories, agent-based modeling and most versions of game theory can be classified as hard systems approaches.

The respective approaches have not altogether been excluded from fertile contacts with the interpretivist strand of inquiry. In principle, all of them can be considered as instruments for supporting discourses about different interpretations of an organizational reality or alternative futures studied in concrete cases. In our time, most applications of the VSM, for example, are constructivist in nature. To put it in a nutshell, these applications are (usually collective) constructions of a (new) reality, in which observation and interpretation play a crucial part. In this process, the actors involved make sense of the system under study, i. e., the organization in focus, by mapping it on the VSM. At the same time they bring forth "multiple realities rather than striving for a fit with one reality" (p. 299 in [29]).

The second group of methodologies is part of the interpretive strand:

- *Interactive Planning*. IP is a methodology, designed by Russell Ackoff [1], and developed further by Jamshid

Gharajedaghi [28], for the purpose of dealing with "messes" and enabling actors to design their desired futures, as well as to bring them about. It is grounded in theoretical work on purposeful systems, reverts to the principles of continuous, participative and holistic planning, and centers on the idea of an "idealized design."

- *Soft Systems Methodology*. SSM is a heuristic designed by Peter Checkland [13,14] for dealing with complex situations. Checkland suggests a process of inquiry constituted by two aspects: A conceptual one, which is logic based, and a sociopolitical one, which is concerned with the cultural feasibility, desirability and implementation of change.
- *Critical Systems Heuristics*. CSH is a methodology, which Werner Ulrich [67,68] proposed for the purpose of scientifically informing planning and design in order to lead to an improvement in the human condition. The process aims at uncovering the interests that the system under study serves. The legitimacy and expertise of actors, and particularly the impacts of decisions and behaviors of the system on others – the "affected" – are elicited by means of a set of boundary questions. CSH can be seen as part of a wider movement known as the "Emancipatory Systems Approach" which embraces, e. g., Freire's Critical Pedagogy, Interpretive Systemology, and Community OR (see pp. 291ff in [31]).

All three of these methodologies (IP, SSM, and CSH) are positioned in the interpretive tradition. Other methodologies and concepts which can be subsumed under the interpretive systems approach are, e. g., Warfield's science of generic design, Churchman's social system design, Senge's soft systems thinking, Mason and Mitroff's strategic assumptions surfacing and testing (SAST), Eden and Ackermann's strategic options in development and analysis (SODA), and other methodologies of soft operational research (for details, see pp. 211ff in [31]). The interpretive methodologies were designed to deal with qualitative aspects in the analysis and design of complex systems, emphasizing the communicational, social, political and ethical dimensions of problem solving. Several authors mention explicitly that they do not preclude the use of quantitative techniques or include such techniques in their repertoire (e. g., the biocyberneticist Frederic Vester).

In an advanced understanding of system dynamics both of these traditions—positivist and interpretivist—are synthesized. The adherents of SD conceive of model building and validation as a semi-formal, relativistic, holistic social process. Validity is understood as usefulness or fitness in relation to the purpose of the model, and validation as an elaborate set of procedures – including logico-structural, heuristic, algorithmic, statistical, and also discursive components – by which the quality of and the confidence in a model are gradually improved (see [4,5,59]).

System Dynamics – Its Features, Strengths and Limitations

The features, strengths and limitations of the SD methodology are a consequence of its specific characteristics. In the context of the multiple theories and methodologies of the systems movement, some of the distinctive features of SD are (for an overview, see [52], pp. 142ff in [31]):

- *Feedback as conceptual basis*. SD model systems are high-order, multiple-loop networks of closed loops of information. Concomitantly, an interest in non-linearities, long-term patterns and internal structure rather than external disturbances is characteristic of SD (p. 31 in [40]). However, SD models are not "closed systems", as sometimes is claimed, in the sense that (a) flows can originate from outside the system's boundaries, (b) representations of exogenous factors or systems can be incorporated into any model as parameters or special modules, and (c) new information can be accommodated via changes to a model. In other words, the SD view hinges on a view of systems which are closed in a causal sense but not materially (p. 297 in [52]).
- *Focus on internally generated dynamics*. SD models are conceived as closed systems. The interest of users is in the dynamics generated inside those systems. Given the nature of closed feedback loops and the fact that delays occur within them, the dynamic behavior of these systems is essentially non-linear.
- *Emphasis on understanding*. For system dynamicists the understanding of the dynamics of a system is the first goal to be achieved by means of modeling and simulation. Conceptually, they try to understand events as embedded in patterns of behavior, which in turn are generated by underlying structures. Such understanding is enabled by SD as it "shows how present policies lead to future consequences" (Sect. VIII in [23]). Thereby, the feedback loops are "a major source of puzzling behavior and policy difficulties" (p. 300 in [52]). SD models purport to test mental models, hone intuition and improve learning (see [65]).
- *High degree of operationality*. SD relies on formal modeling. This fosters disciplined thinking; assumptions, underlying equations and quantifications must be clarified. Feedback loops and delays are visualized and formalized; therewith the causal logic inherent in a model

is made more transparent and discussable than in most other methodologies [53]. Also, a high level of realism in the models can be achieved. SD is therefore apt to support decision-making processes effectively.

- *Far-reaching requirements (and possibilities) for the combination of qualitative and quantitative aspects of modeling and simulation*. This is a consequence of the emphasis on understanding. The focus is not on point-precise prediction, but on the generation of insights into the patterns generated by the systems under study.
- *High level of generality and scale robustness*. The representation of dynamic systems in terms of stocks and flows is a generic form, which is adequate for a wide spectrum of potential applications. This spectrum is both broad as to the potential subjects under study, and deep as to the possible degrees of resolution and detail [38]. In addition, the SD methodology enables one to deal with large numbers of variables within multiple interacting feedback loops (p. 9 in [22]). SD has been applied to the most diverse subject areas, e. g., global modeling, environmental issues, social and economic policy, corporate and public management, regional planning, medicine, psychology and education in mathematics, physics and biology.

The features of SD just sketched out result in both strengths and limitations. We start with the strengths.

Strengths of SD

1. Its *specific modeling approach* makes SD particularly helpful in gaining insights into the patterns exhibited by dynamic systems, as well as the structures underlying them. Closed-loop modeling has been found most useful in fostering understanding of the dynamic functioning of complex systems. Such understanding is especially facilitated by the principle of modeling the systems or issues under study in a continuous mode and at rather high aggregation levels [20,38]. With the help of relatively small but insightful models, and by means of sensitivity analyses as well as optimization heuristics incorporated in the application software packages, decision-spaces can be thoroughly explored. Vulnerabilities and the consequences of different system designs can be examined with relative ease.
2. The *generality of the methodology* and its power to crystallize operational thinking in realistic models have triggered applications in the most varied contexts. Easy-to-use software and the features of screen-driven modeling via graphic user interfaces provide a strong lever for collaborative model-building in teams (cf. [2,69]).
3. Another strong point is the *momentum of the SD movement*. Due to the strengths commented above this point, the community of users has grown steadily, being probably the largest community within the systems movement. Lane (p. 484 in [36]) has termed SD "one of the most widely used systems approaches in the world."
4. Its specific features make SD an exceptionally effective tool for *conveying systemic thinking* to anybody. Therefore, it also has an outstanding track-record of classroom applications for which "learner-directed learning" [24] or "learner-centered learning" is advocated [25,26]. Pertinent audiences range from schoolchildren at the levels of secondary and primary schools to managers and scientists.

Given these strengths, the community of users has not only grown significantly, but has also transcended disciplinary boundaries, ranging from the formal and natural sciences to the humanities, and covering multiple uses from theory building and education to the tackling of real-world problems at almost any conceivable level. Applications to organizational, societal and ecological issues have seen a particularly strong growth. This feeds back on the availability and growth of the knowledge upon which the individual modeler can draw.

The flip side of most of the strengths outlined here embodies the limitations of SD; we concentrate on those which can be relevant to a possible complementarity of SD with other systems methodologies.

Limitations of SD

1. The main point here is that SD does not provide a framework or methodology for the *diagnosis and design* of organizational structures in the sense of interrelationships among organizational actors. This makes SD susceptible to completion from without – a completion which organizational cybernetics (OC), and the VSM in particular, but also living system theory (LST), especially can provide. The choice falls on these two approaches because of their strong heuristic power and their complementary strengths in relation to SD (cf. [57,61]).
2. Another limitation of SD is related to the *absorption of variety* (complexity) by an organization. *Variety* is a technical term for *complexity*, which denotes a (high) number of potential states or behaviors of a system (based on [3,8]). SD offers an approach to the handling of variety which allows modeling at different scales of a problem or system [47]. It focuses on the identification, at a certain resolution level or possibly several resolution levels, of the main stock variables which will be

affected by the respective flows. These, in turn, will be influenced by parameters and auxiliary variables. This approach, even though it enables thinking and modeling at different scales, does not provide a formal procedure for an organization to cope with the external complexity it faces, namely, for designing a structure which can absorb that complexity. In contrast, OC and LST offer elaborate models to enable the absorption of variety, in the case of the VSM based explicitly on Ashby's *Law of Requisite Variety*. It says "Only variety can destroy variety", which implies that the varieties of two interacting systems must be in balance, if stability is to be achieved [3]. The VSM has two salient features in this respect. Firstly, it helps design an organizational unit for viability, by enabling it to attenuate the complexity of its environment, and also to enhance its eigenvariety, so that the two are in balance. The term *variety engineering* has been used in this context [9]. Secondly, the recursive structure of the VSM ensures that an organization with several levels will develop sufficient eigenvariety along the fronts on which the complexity it faces unfolds. Similarly, LST offers the conditions for social systems to survive, by maintaining thermodynamically highly improbable energy states via continuous interaction with their environments. The difference between the two approaches is that the VSM functions more in the strategic and informational domains, while the LST model essentially focuses on the operational domain. In sum, both can make a strong contribution related to coping with the external complexity faced by organizations, and therefore can deliver a strong complement to SD.

3. Finally, the design of *modeling processes* confronts SD with specific challenges. The original SD methodology of modeling and simulation was to a large extent functionally and technically oriented. This made it strong in the domain of logical analysis, while the socio-cultural and political dimensions of the modeling process were, if not completely out of consideration, at least not a significant concern in methodological developments. The SD community – also under the influence of the soft systems approaches – has become aware of this limitation and has worked on incorporating features of the social sciences into its repertoire. The following examples, which document this effort to close the gap, stand for many. Extensive work on group model building has been achieved, which explores the potential of collaborative model building [69]. A new schema for the modeling process has been proposed, which complements logic-based analysis by cultural analysis [37]. The social dimension of system dynamics-based modeling has become subject to intensive discussion ([77]; and other contributions to the special issue of *Systems Research and Behavioral Science*, Vol. 51, No. 4, 2006). Finally, in relation to consultancy methodology, modeling has been framed as a learning process [34] and as second-order intervention [60].

As has been shown, there is a need to complement classical SD with other methodologies, when issues are at stake which it cannot handle by itself. VSM and LST are excellent choices when issues of organizational diagnosis or design are to be tackled.

The limitations addressed here call attention to other methodologies which exhibit certain features that traditionally were not incorporated, or at least not explicit, in SD methodology. One aspect concerns the features that explicitly address the subjectivity of purposes and meanings ascribed to systems. In this context, support for problem formulation, model construction and strategy design by individuals on the one hand and groups on the other are relevant issues. Also, techniques for an enhancement of creativity (e. g., the generation and the reframing of options) in both individuals and groups are a matter of concern. Two further aspects relate to methodological arrangements for coping with the specific issues of negotiation and alignment in pluralist and coercive settings.

As far as the modeling processes are concerned, group model building has proven to be a valuable complement to pure modeling and simulation. However, there are other systems methodologies which should be considered as potentially apt to enrich SD analysis, namely the soft approaches commented upon earlier, e. g., interactive planning, soft system methodology and critical system heuristics.

On the other hand, SD can be a powerful complement to other methodologies which are more abstract or more static in nature. This potential refers essentially to all systems approaches which stand in the interpretive ("soft") tradition, but also to approaches which stand in the positivist traditions, such as the VSM and LST. These should revert to the support of SD in the event that tradeoffs between different goals must be handled, or if implications of long-term decisions on short-term outcomes (and vice versa) have to be ascertained, and whenever contingencies or vulnerabilities must be assessed.

Actual and Potential Relationships

It should be clear by now that the systems movement has bred a number of theories and methodologies, none of which can be considered all-embracing or complete. All of

them have their strengths and weaknesses, and their specific potentials and limitations.

Since Burrell and Morgan [12] adverted to incommensurability between different paradigms of social theory, several authors have acknowledged or even advocated methodological complementarism. They argue that there is a potential complementarity between different methods, and, one may add, models, even if they come from distinct paradigms. Among these authors are, e. g., Brocklesby [11], Jackson [30], Midgley [43], Mingers [45], Schwaninger [55] and Yolles [83]. These authors have opened up a new perspective in comparison with the non-complementaristic state-of-the-art.

In the past, the different methodologies have led to the formation of their own traditions and "schools," with boundaries across which not much dialogue has evolved. The methodologies have kept their protagonists busy testing them and developing them further. Also, the differences between different language games and epistemological traditions have often suggested incommensurability, and therewith have impaired communication. Prejudices and a lack of knowledge of the respective other side have accentuated this problem: Typically, "hard" systems scientists are suspicious of "soft" systems scientists. For example, many members of the OR community, not unlike orthodox quantitatively oriented economists, adhere to the opinion that "SD is too soft." On the other hand the protagonists of "soft" systems approaches, even though many of them have adopted feedback diagrams (causal loop diagrams) for the sake of visualization, are all too often convinced that "SD is too hard." Both of these judgments indicate a lack of knowledge, in particular of the SD validation and testing methods available, on the one hand, and the technical advancements achieved in modeling and simulation, on the other (see [5,59,66]).

In principle, both approaches are complementary. The qualitative view can enrich quantitative models, and it is connected to their philosophical, ethical and esthetical foundations. However, qualitative reasoning tends to be misleading if applied to causal network structures without being complemented by formalization and quantification of relationships and variables. Furthermore, the quantitative simulation fosters insights into qualitative patterns and principles. It is thus a most valuable device for validating and honing the intuition of decision makers, via corroboration and falsification.

Proposals that advocate mutual learning between the different "schools" have been formulated inside the SD community (e. g., [35]). The International System Dynamics Conference of 1994 in Stirling, held under the banner of "Transcending the Boundaries," was dedicated to the dialogue between different streams of the systems movement.

Also, from the 1990s onwards, there were vigorous efforts to deal with methodological challenges, which traditionally had not been an important matter of scientific interest within the SD community. Some of the progress made in these areas is documented in a special edition of *Systems Research and Behavioral Science* (Vol. 21, No. 4, July-August 2004). The main point is that much of the available potential is based on the complementarity, not the mutual exclusiveness, of the different systems approaches.

In the future, much can be gained from leveraging these complementarities. Here are two examples of methodological developments in this direction, which appear to be achievable and potentially fertile: The enhancement of qualitative components in "soft" systems methodologies in the process of knowledge elicitation and model building (cf. [69]), and the combination of cybernetics-based organizational design with SD-based modeling and simulation (cf. [61]). Potential complementarities exist not only across the qualities – quantities boundary, but also within each one of the domains. For example, with the help of advanced software, SD modeling ("top-down") and agent-based modeling ("bottom-up") can be used in combination.

From a meta-methodological stance, generalist frameworks have been elaborated which contain blueprints for combining different methodologies where this is indicated. Two examples are:

- *Total systems intervention* (TSI) is a framework proposed by Flood and Jackson [19], which furnishes a number of heuristic schemes and principles for the purpose of selecting and combining systems methods/methodologies in a customized way, according to the issue to be tackled. SD is among the recommended "tools".
- *Integrative systems methodology* (ISM) is a heuristic for providing actors in organizations with requisite variety, developed by Schwaninger [55,56]. It advocates (a) dealing with both content– and context-related issues during the process, and (b) placing a stronger emphasis on the validation of qualitative and quantitative models as well as strategies, in both dimensions of the content of the issue under study and the organizational context into which that issue is embedded. For this purpose, the tools of SD (to model content) and organizational cybernetics – the VSM (to model context) – are cogently integrated.

These are only two examples. In principle, SD could make an important contribution in the context of most of

the methodological frameworks, far beyond the extent to which this has been the case. Systems methodologists and practitioners can potentially benefit enormously from including SD methodology in their repertoires.

Outlook

There have recently been calls for an eclectic "mixing and matching" of methodologies. In light of the epistemological tendencies of our time towards radical relativism, it is necessary to warn against taking a course in which "anything goes". It is most important to emphasize that the desirable methodological progress can only be achieved on the grounds of scientific rigor. This postulate of "rigor" is not to be confused with an encouragement of "rigidity." The necessary methodological principles advocated here are disciplined thinking, a permanent quest for better models (that is, thorough validation), and the highest achievable levels of transparency in the formalizations as well as of the underlying assumptions and sources used. Scientific rigor, in this context, also implies that combinations of methodologies reach beyond merely eclectic add-ons from different methodologies, so that genuine integration towards better adequacy to the issues at hand is achieved.

The contribution of system dynamics can come in the realms of the following:

- Fostering disciplined thinking
- Understanding dynamic behaviors of systems and the structures that generate them
- Exploring paths into the future and the concrete implications of decisions
- Assessing strategies as to their robustness and vulnerabilities, in ways precluded by other, more philosophical, and generally "soft" systems approaches

These latter streams can contribute to reflecting and tackling the meaning- and value-laden dimensions of complex human, social and ecological systems. Some of their features should and can be combined synergistically with system dynamics, particularly by being incorporated into the repertoires of system dynamicists. From the reverse perspective, incorporating system dynamics as a standard tool will be of great benefit for the broad methodological frameworks. Model formalization and dynamic simulation may even be considered necessary components for the study of the concrete dynamics of complex systems.

Finally, there are also many developments in the "hard", i. e., mathematics-, statistics-, logic-, and informatics-based methods and technologies, which are apt to enrich the system dynamics methodology, namely in terms of modeling and decision support. For example, the constantly evolving techniques of time-series analysis, filtering, neural networks and control theory can improve the design of system-dynamics-based systems of (self-)control. Also, a bridge across the divide between the top-down modeling approach of SD and the bottom-up approach of agent-based modeling appears to be feasible. Furthermore, a promising perspective for the design of genuinely "intelligent organizations" emerges if one combines SD with advanced database-management, cooperative model building software, and the qualitative features of the "soft" systems methodologies.

The approaches of integrating complementary methodologies outlined in this contribution definitely mark a new phase in the history of the systems movement.

Appendix

Milestones in the Evolution of the Systems Approach in General and System Dynamics in Particular

The table gives an overview of the systems movement's evolution, as shown in its main literature; and that overview is not exhaustive.

Systems Approaches – An Overview

Note: This diagram shows three streams of the systems approach in the context of their antecedents. The general systems thread has its origins in philosophical roots from antiquity: The term *system* derives from the old Greek σύστημα (systēma), while, *cybernetics* stems from the Greek κυβερνήτης (kybernētēs). The arrows between the threads stand for interrelationships and efforts to synthesize the connected approaches. For example, integrated systems methodology is an integrative attempt to leverage the complementarities of system dynamics and organizational cybernetics. Enumerated to the left and right of the scheme are the fields of application. The big arrows in the upper region of the diagram indicate that the roots of the systems approach continue influencing the different threads and the fields of application even if the path via general systems theory is not pursued.

The diagram is not a complete representation, but the result of an attempt to map the major threads of the systems movement and some of their interrelations. Hence, the schema does not cover all schools or protagonists of the movement. Why does the diagram show a dynamic and evolutionary systems thread and a cybernetics thread, if cybernetics is about dynamic systems? The latter embraces all the approaches that are explicitly grounded in cybernetics. The former relates to all other approaches

System Dynamics in the Evolution of the Systems Approach, Table 1
Milestones in the evolution of the systems approach in general and system dynamics in particular

Foundations of general system theory		
Von Bertalanffy	Zu einer allgemeinen Systemlehre	1945
	An Outline of General System Theory	1950
	General System Theory	1968
Bertalanffy, Boulding, Gerard, Rapoport	Foundation of the Society for General Systems Research	1953
Klir	An Approach to General System Theory	1968
Simon	The Sciences of the Artificial	1969
Pichler	Mathematische Systemtheorie	1975
Miller	Living Systems	1978
Mesarovic & Takahara	Abstract Systems Theory	1985
Rapoport	General System Theory	1986
Foundations of cybernetics		
Macy Conferences (Josiah Macy, Jr. Foundation)	Cybernetics. Circular Causal, and Feedback Mechanisms in Biological and Social Systems	1946–1951
Wiener	Cybernetics or Control and Communication in the Animal and in the Machine	1948
Ashby	An Introduction to Cybernetics	1956
Pask	An Approach to Cybernetics	1961
Von Foerster, Zopf	Principles of Self-Organization	1962
McCulloch	Embodiments of Mind	1965
Foundations of organizational cybernetics		
Beer	Cybernetics and Management	1959
	Towards the Cybernetic Factory	1962
	Decision and Control	1966
	Brain of the Firm	1972
Von Foerster	Cybernetics of Cybernetics	1974
Foundations of system dynamics		
Forrester	Industrial Dynamics	1961
	Principles of Systems	1968
	Urban Dynamics	1969
	World Dynamics	1971
Meadows et al.	Limits to Growth	1972
Richardson	Feedback Thought in Social Science and Systems Theory	1991
Systems methodology		
Churchman	Challenge to Reason	1968
	The Systems Approach	1968
Vester & von Hesler	Sensitivitätsmodell	1980
Checkland	Systems Thinking, Systems Practice	1981
Ackoff	Creating the Corporate Future	1981
Ulrich	Critical Heuristics of Social Planning	1983
Warfield	A Science of Generic Design	1994
Schwaninger	Integrative Systems Methodology	1997
Gharajedaghi	Systems Thinking	1999
Sabelli	Bios – A Study of Creation	2005
Selected recent works in system dynamics		
Senge	The Fifth Discipline	1990
Barlas & Carpenter	Model Validity	1990
Vennix	Group Model Building	1996
Lane & Oliva	Synthesis of System Dynamics and Soft Systems Methodology	1998
Sterman	Business Dynamics	2000
Warren	Strategy Dynamics	2002, 2008
Wolstenholme	Archetypal Structures	2003
Morecroft	Strategic Modelling	2007
Schwaninger & Grösser	Theory-building with System Dynamics & Model Validation	2008, 2009

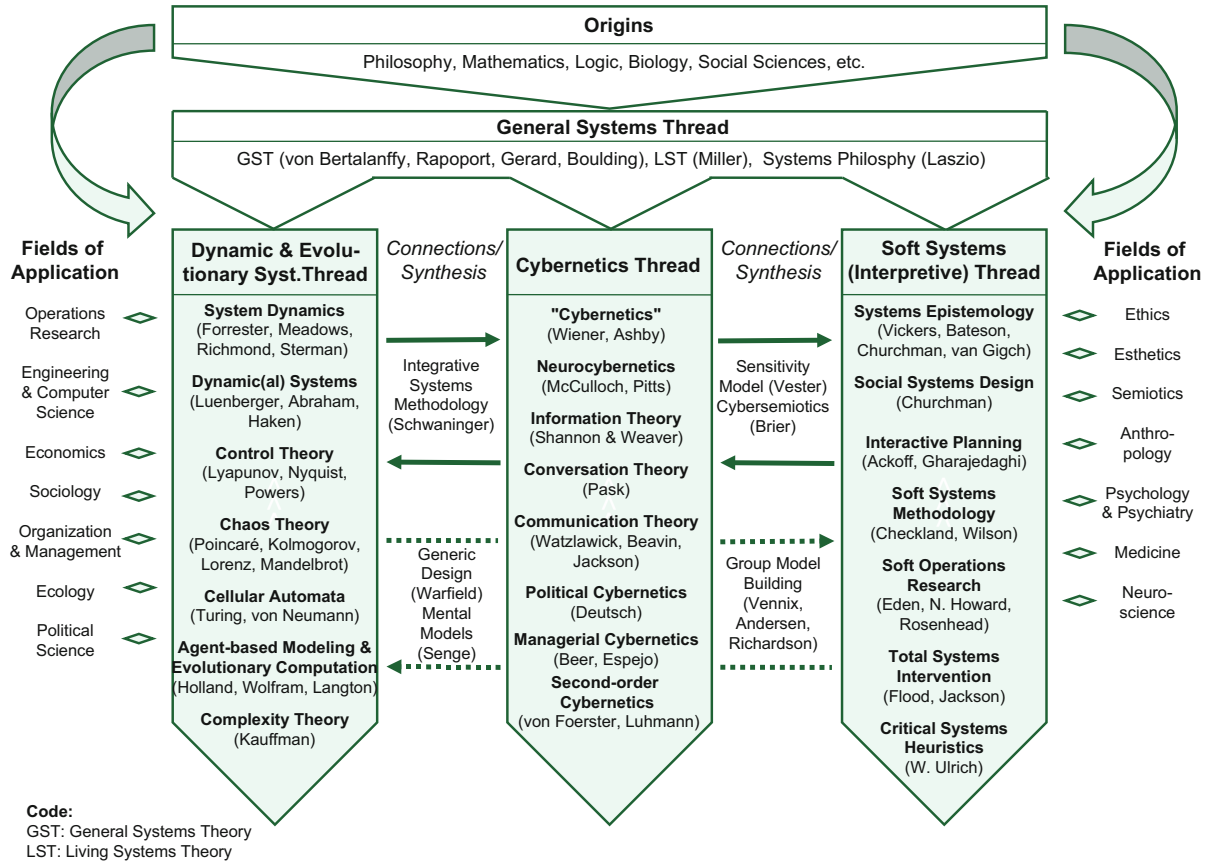

System Dynamics in the Evolution of the Systems Approach, Figure 1

concerned with dynamic or evolutionary systems. The simplification made it necessary to somewhat curtail logical perfection for the sake of conveying a synoptic view of the different systems approaches, in a language that uses the categories common in current scientific and professional discourse. Overlaps exist, e. g., between dynamic systems and chaos theory, cellular automata and agent-based modeling.

Bibliography

Primary Literature

1. Ackoff RL (1981) Creating the Corporate Future. Wiley, New York
2. Andersen DF, Richardson GP (1997) Scripts for Group Model Building. Syst Dyn Rev 13(2):107–129
3. Ashby WR (1956) An Introduction to Cybernetics. Chapman & Hall, London
4. Barlas Y (1996) Formal aspects of model validity and validation in system dynamics. Syst Dyn Rev 12(3):183–210
5. Barlas Y, Carpenter S (1990) Philosophical roots of model validity: Two paradigms. Syst Dyn Rev 6(2):148–166
6. Beer S (1959) Cybernetics and Management. English Universities Press, London
7. Beer S (1994) Towards the Cybernetic Factory. In: Harnden R, Leonard A (eds) How Many Grapes Went into the Wine. Stafford Beer on the Art and Science of Holistic Management. Wiley, Chichester, pp 163–225 (reprint, originally published in 1962)
8. Beer S (1966) Decision and Control. Wiley, Chichester
9. Beer S (1979) The Heart of Enterprise. Wiley, Chichester
10. Beer S (1981) Brain of the Firm, 2nd edn. Wiley, Chichester
11. Brocklesby J (1993) Methodological complementarism or separate paradigm development – Examining the options for enhanced operational research. Aust J Manag 18(2):133–157
12. Burrell G, Morgan G (1979) Sociological Paradigms and Organisational Analysis. Hants, Gower
13. Checkland PB (1981) Systems Thinking, Systems Practice. Wiley, Chichester
14. Checkland PB, Poulter J (2006) Learning for Action: A Short Definitive Account of Soft Systems Methodology, and its Use Practitioners, Teachers and Students. Wiley, Chichester
15. Churchman CW (1968) Challenge to Reason. McGraw-Hill, New York
16. Churchman CW (1968) The Systems Approach. Delacorte Press, New York
17. Churchman CW (1979) The Systems Approach and its Enemies. Basic Books, New York
18. Encyclopedia of Life Support Systems (2002) published under: http://www.eolss.net/

19. Flood RL, Jackson MC (1991) Creative Problem Solving. Total Systems Intervention. Wiley, Chichester
20. Forrester JW (1961) Industrial Dynamics. MIT Press, Cambridge
21. Forrester JW (1968) Principles of Systems. MIT Press, Cambridge
22. Forrester JW (1969) Urban Dynamics. MIT Press, Cambridge
23. Forrester JW (1971) World Dynamics. Pegasus Communications, Waltham
24. Forrester JW (1993) System Dynamics and the Lessons of 35 Years. In: DeGreene KB (ed) Systems-Based Approach to Policy Making. Kluwer, Boston
25. Forrester JW (1993) System Dynamics as an Organizing Framework for Pre-college Education. Syst Dyn Rev 9(2):183–194
26. Forrester JW (1997) System Dynamics and K-12 Teachers. A Lecture at the University of Virginia School of Education, Massachusetts Institute of Technology. System Dynamics Group Paper D-4665–4
27. François C (2004) International Encyclopedia of Systems and Cybernetics, 2nd edn. Saur, München
28. Gharajedaghi J (1999) Systems Thinking. Managing Chaos and Complexity. Butterworth-Heinemann, Boston
29. Harnden RJ (1989) Technology for Enabling: The Implications for Management Science of a Hermeneutics of Distinction. The University of Aston, Birmingham
30. Jackson MC (1991) Systems Methodology for the Management Sciences. Plenum Press, New York
31. Jackson MC (2000) Systems Approaches to Management. Kluwer Academic/Plenum, New York
32. Kauffman SA (1993) The Origins of Order. Self-Organization and Selection in Evolution. Oxford University Press, New York
33. Klir GJ (1969) An Approach to General Systems Theory. Nostrand, New York
34. Lane DC (1994) Modeling as Learning: A Consultancy Methodology for Enhancing Learning in Management in Management Teams. In: Morecroft J, Sterman JD (eds) Modeling for Learning Organizations. Productivity Press, Portland, pp 205–240
35. Lane DC (1994) With a little help from our friends: How system dynamics and soft OR can learn from each other. Syst Dyn Rev 10(2–3):101–134
36. Lane DC (2006) IFORS' Operational Research Hall of Fame. Jay Wright Forrester. Int Trans Oper Res 13:483–492
37. Lane DC, Oliva R (1998) The Greater Whole: towards a synthesis of system dynamics and soft system methodology. Eur J Oper Res 107(1):214–235
38. La Roche U, Simon M (2000) Geschäftsprozesse simulieren: flexibel und zielorientiert führen mit Fliessmodellen. Orell Füssli, Zürich
39. McCulloch WS (1965) Embodiments of Mind. MIT Press, Cambridge
40. Meadows DH (1980) The Unavoidable A Priori. In: Randers J (ed) Elements of the System Dynamics Method. MIT Press, Cambridge, pp 23–57
41. Meadows DH, Meadows DL, Randers J, Behrens III WW (1972) Limits to Growth. Universe Books, New York
42. Mesarovic MD, Takahara Y (1985) Abstract Systems Theory. Springer, Berlin
43. Midgley G (2000) Systemic Intervention. Philosophy, Methodology, and Practice. Kluwer, New York
44. Miller JG (1978) Living Systems. McGraw-Hill, New York
45. Mingers J (1997) Multi-paradigm Multimethodology. In: Mingers J, Gill A (eds) Multimethodology. Wiley, Chichester
46. Morecroft J (2007) Strategic Modelling and Business Dynamics: a Feedback Systems Approach. Wiley, Chichester
47. Odum HT, Odum EC (2000) Modeling for all Scales: An Introduction to System Simulation. Academic Press, San Diego
48. Pask G (1961) An Approach to Cybernetics. Hutchinson, London
49. Pichler F (1975) Mathematische Systemtheorie. de Gruyter, Berlin
50. Rapoport A (1953) Operational philosophy: Integrating Knowledge and Action. Harper, New York
51. Rapoport A (1986) General System Theory. Essential Concepts and Applications Abacus Press, Turnbridge Wells
52. Richardson GP (1999) Feedback Thought in Social Science and Systems Theory. Pegasus Communications, Waltham (Originally published in 1991)
53. Richmond B (1997) The "Thinking" in systems thinking: How can we make it easier to master? Syst Think 8(2):1–5
54. Sabelli H (2005) Bios: a Study of Creation. World Scientific, Hackensack
55. Schwaninger M (1997) Integrative systems methodology: Heuristic for requisite variety. Int Trans Oper Res 4(4):109–123
56. Schwaninger M (2004) Methodologies in conflict: Achieving synergies between system dynamics and organizational cybernetics. Syst Res Behav Sci 21(4):1–21
57. Schwaninger M (2006) Theories of viability. A comparison. Syst Res Behav Sci 23:337–347
58. Schwaninger M, Groesser S (2008) System dynamics as model-based theory building. Syst Res Behav Sci 25:1–19
59. Schwaninger M, Groesser S (2009) Model Validation: The Quest for Quality in System Dynamics Modeling. Encyclopaedia of Complexity and Systems Science. Springer, New York
60. Schwaninger M, Janovjak M, Ambroz K (2006) Second-order intervention: Enhancing organizational competence and performance. Syst Res Behav Sci 23:529–545
61. Schwaninger M, Pérez Ríos J (2008) System dynamics and cybernetics: A synergetic pair. Syst Dyn Rev 24(2):145–174
62. Senge PM (1990) The Fifth Discipline. The Art and Practice of the Learning Organization. Doubleday, New York
63. Shapiro M, Mandel T, Schwaninger M et al (1996) The Primer Toolbox. International Society for the Systems Sciences, http://www.isss.org/primer/toolbox.htm
64. Simon HA (1969) The Sciences of the Artificial. MIT Press, Cambridge
65. Sterman JD (1994) Learning in and about complex systems. Syst Dyn Rev 10(2–3):291–330
66. Sterman JD (2000) Business Dynamics. Systems Thinking and Modeling for a Complex World. Irwin/McGraw-Hill, Boston
67. Ulrich W (1983) Critical Heuristics of Social Planning. Haupt, Bern
68. Ulrich W (1996) A Primer to Critical Systems Heuristics for Action Researchers. The Centre of Systems Studies, University of Hull, Hull
69. Vennix JAM (1996) Group Model Building. Facilitating Team Learning Using System Dynamics. Wiley, Chichester
70. Vester F, Von Hesler A (1980) Sensitivitätsmodell. Regionale Planungsgemeinschaft Untermain, Frankfurt am Main
71. Von Bertalanffy L (1949) Zu einer allgemeinen Systemlehre. Bl Dtsch Philos 18(3/4) (Excerpts: in Biol Gen 19(1):114–129 and in General System Theory, 1968, Chapter III)
72. Von Bertalanffy L (1950) An outline of general system theory. Br J Philos Sci 1:139–164

73. Von Bertalanffy L (1968) General System Theory. Braziller, New York
74. Von Foerster H (1984) Observing Systems, 2nd edn. Intersystems Publications, Seaside
75. Von Foerster H (ed) (1995) Cybernetics of Cybernetics, 2nd edn. Future Systems, Minneapolis (Originally published in 1974)
76. Von Foerster H, Zopf GW (eds) (1962) Principles of Self-Organization. Pergamon Press, Oxford
77. Vriens D, Achtenbergh J (2006) The social dimension of system dynamics-based modelling. Syst Res Behav Sci 23(4): 553–563
78. Warfield JN (1994) A Science of Generic Design: Managing Complexity through Systems Design, 2nd edn. Iowa State University Press, Ames
79. Warren K (2002) Competitive Strategy Dynamics. Wiley, Chichester
80. Warren K (2008) Strategic Management Dynamics. Wiley, Chichester
81. Wiener N (1948) Cybernetics: Control and Communication in the Animal and in the Machine. MIT Press, Cambridge
82. Wolstenholme E (2003) Towards the definition and use of a core set of archetypal structures in system dynamics. Syst Dyn Rev 19(1):7–26
83. Yolles MA (1998) Cybernetic exploration of methodological complement. Kybern 27(5):527–542

Books and Reviews

Jackson MC (2003) Systems Thinking: Creative Holism for Managers. Wiley, Chichester
Klir GJ (2001) Facets of Systems Science, 2nd edn. Kluwer Academic/Plenum, New York
Midgley G (ed) (2003) Systems Thinking, vol 4. Sage, London
Ragsdell G, Wilby J (eds) (2001) Understanding Complexity. Kluwer Academic/Plenum, New York
Richardson GP (ed) (1996) Modelling for Management. Simulation in Support of Systems Thinking, 2 Volumes. Aldershot, Dartmouth
Schwaninger M (2006) Intelligent Organizations. Powerful Model for Systemic Management. Springer, Berlin
Van Gigch JP (2003) Metadecisions. Rehabilitating Epistemology. Kluwer Academic/Plenum, New York

System Dynamics, Introduction to

Brian Dangerfield
Centre for OR & Applied Statistics, Salford Business School, University of Salford, Salford, UK

When Jay Wright Forrester published his first paper in 1958 he subtitled it *"a major breakthrough for decision-makers"*. At the time some thought this rather an exaggeration if not pompous. Now that 50 years of system dynamics (SD) has elapsed we can at least point to the achievements made and re-state continuing progress in the pages of this section. Was it a 'major breakthrough'? It certainly has the potential to raise the standards in evidence-based policy making to warrant this description and some startlingly good examples of such work will be mentioned here. But after 50 years perhaps one might expect more than has surfaced heretofore.

The key might be connected to the skills required to formulate good SD models – those which address a real-world problem with devastating simplicity and insight. It is deceptively easy to produce an SD model but there are subtleties involved in producing a really effective model for policy purposes. An uplift in modeling skills is something which a subset of the (now significant) amount of published material on SD is aimed at and this section will add to that corpus of work. In addition it will illustrate the extent to which SD applications have spread from its genesis in business to embrace health care, environmental, energy and climate issues, project management, some aspects of biological science and human physiology, governmental and public policy generally, economics (mainly macro), the diffusion of innovations and finally social and economic development. Other applications are being encountered as the power of the methodology is becoming appreciated. It has long since justified the change of title from ***Industrial*** Dynamics (1958) to ***System*** Dynamics (1970 onwards).

Richardson contributes an overview of the basics of SD modeling (see ► System Dynamics, The Basic Elements of). The underlying conceptual framework is that of the information feedback loop together with resource stocks and flows and an endogenous perspective on causation. The simplicity of the loop concept is apt to contribute to the apparent ease with which SD models can be created (along with the icon-based suites of SD software). But the novice reader should appreciate that it can take time to assimilate the modeling skills necessary to execute well an SD model-based application. Practice is essential and the references included will lead to further published material to assist the steep climb up the learning curve. So-called experts are still being confronted with the subtleties of SD modeling after years of involvement.

To place the SD methodology in context, the contribution by Schwaninger (see ► System Dynamics in the Evolution of the Systems Approach) profiles it alongside various others 'systems' based approaches which have emerged in the management and social sciences. Those professing to become experts in SD need to know about the other range of approaches which co-exist in the field of systems science. All these other methodologies have their own enthusiasts and this may even extend to the formation of societies with annual conferences. His Appendix B

shows a diagram of the different systems approaches and their interrelationships.

The foundations of the SD methodology can be characterized by certain philosophical issues. Olaya's text (see ► System Dynamics Philosophical Background and Underpinnings) defines a central one as presentationalism, associated with the notion of 'mental models'. A number of other philosophical issues which relate to SD are introduced, including those of positivism and social theory.

The practice of SD when applied to real-world applications essentially involves managerial learning and will often involve an interaction with client teams rather than one individual. How best to organize such structured approaches to participative model building is described by Rouwette and Vennix (see ► Group Model Building). Client participation is required for successful modeling.

If the promotion of learning and understanding is the primary *raison d'etre* of SD, then achievement of this goal in an individual can be a significant accomplishment, especially if that person is the most senior in the client team. But there is a further goal to be pursued should the study fully reap the benefits of the SD methodology: How can we foster *organizational* learning? Maani tackles this head on (see ► System Dynamics and Organizational Learning). He defines the core capabilities of a learning organization and goes on to list the developing literature on organizational learning and, most importantly, how SD can aid the process through learning laboratories and microworlds.

Running an SD model creates a time-path of output behavior covering all the variables it is deemed necessary to include in the model. The various runs of the model are, most frequently, addressed in comparative fashion rather than taken in isolation. They can therefore be described as computer-based scenarios each of which charts a possible but not assured future. Georgantzas (see ► Scenario-Driven Planning with System Dynamics) describes environmental (traditional) scenario generation for which there is a considerable body of literature. But he emphasizes that successful strategy design involves the integration of three things: a knowledge of the business environment; the effects of unstated assumptions about change in the environment and strategy on performance; and finally the need to *compute* the effects on organizational performance. These three facets are accomplished by the process of SD modeling.

Thus far this introductory roadmap has covered all the background for contextualizing and creating an SD model. We now turn to various tasks associated with ex post modeling activities. Three such aspects are covered: model validation; analytical methods to explain behavior and determine dominant loops; and model optimization.

Schwaninger and Groesser (see ► System Dynamics Modeling: Validation for Quality Assurance) range over the various aspects of model validation, beginning with its epistemological foundations. In real-world modeling studies testing and validation is a sine qua non of the process. The range of tests made available and the attention given to the task of validation in the literature mark out SD as unique in the field of management science. Few other methodologies get near to the variety of tests which can be applied to an SD model. The authors consider the range of tests under three headings: model-related context; model structure; and model behavior.

Kampmann and Oliva deal with the behavioral analysis issue (see ► System Dynamics, Analytical Methods for Structural Dominance Analysis in). This activity tries to shed light on the model's dynamic behavior: Why does it behave as it does? What loop structures are responsible for the dominant behavior – and indeed shifts in that behavior where it occurs? In other words, they explore the link between system structure and dynamic behavior. Early methods used eigenvalue analysis but, since then, more sophisticated approaches have been put forward. A major advance will occur when one or more of these is refined enough to be included in an SD software package. This is likely to take some time although an improved user interface showing links glowing with differing degrees of intensity, reflecting their relative importance, is possible in the not-too-distant future.

Dangerfield describes the methods for improving model performance (see ► System Dynamics Models, Optimization of). The task can be categorized under two headings: calibration and policy optimization. The former relates to the determination of optimal parameter sets which deliver the best fit of the model to past time series data. Policy optimization on the other hand seeks to establish policies which deliver the 'best' performance against a suitable metric, such as minimum cost or maximum revenue. Using such an approach can accelerate the learning which comes from repeated runs of the model. Sadly, in the existing SD literature, there is scant evidence of its use in real-world studies.

The methodology of SD exists for no other reason than to offer a quantum leap in the standards of policy analysis. Therefore, any review must include a range through the landscape which defines areas of application. There are eight such areas covered in this section and the choice has been made in the knowledge that there are others which may also have been included and some new areas which are only just being opened up to the tools of SD modeling.

Business Strategy was the genesis of SD applications and rightly takes pride of place. This is the field in which

the most numerous SD applications occur. Lyneis (see ► Business Policy and Strategy, System Dynamics Applications to) concentrates on the process of how SD models are used in the task of strategy formulation. He goes on to consider the various drivers of business dynamics such as oscillations in supply chains and boom and bust life cycles. Detailed references are provided for a wide range of business application case studies.

Health care is consuming a higher share of GDP in many Western industrialized countries. This is due to the age profile of the population and advances in pharmacological and medical technologies. It is unsurprising that SD methods have been applied in tackling some of the most high-profile issues in health care and the relatively recent literature is testimony to the success of SD-based analysis. Indeed, it is arguable that some of the best modeling applications have surfaced in this sector. To do justice to the field of health care two contributions were solicited, in part because of the different funding systems which exist on either side of the Atlantic: Wolstenholme surveys the work done by UK and European authors (see ► Health Care in the United Kingdom and Europe, System Dynamics Applications to), whilst Hirsch and Homer concentrate on work published by US authors (see ► Health Care in the United States, System Dynamics Applications to).

Wolstenholme describes work carried out in the UK and Continental Europe but gives particular emphasis to three areas where models have been deployed. He starts with the problem of delayed hospital discharge which generates hospital capacity problems. Epidemiology is also reviewed, in particular research on the epidemiology of HIV/AIDS. Finally, recent work on mental health reform in the UK is described.

Hirsch and Homer note that the system in the USA is comparatively difficult to manage because of its free-market approach and relative lack of regulation. They concentrate on three main areas: disease epidemiology including heart disease and diabetes; substance abuse; and health care capacity and delivery.

Along with health care, the depletion of environmental resources and its effects has consumed many thousands of column inches in printed news media. SD has been employed in the pursuit of more compelling applications in this sector and the efforts go back to the well-known *Limits to Growth* study in 1971–72. Ford charts the most notable efforts which have emerged (see ► System Dynamics Models of Environment, Energy and Climate Change). He ranges over environmental resource problems in the western USA, models for greater understanding of climate change and global warming and concludes with studies in energy, specifically two applications to the electric power industry.

The field of economics is one where SD has received a mostly hostile reception. The statistical economic modeling tool of econometrics has an extensive history and as a preferred modeling methodology seems hard to dislodge. However, there are an increasing number of heterodox economists who are prepared to embrace SD concepts and Radzicki (see ► System Dynamics and Its Contribution to Economics and Economic Modeling) describes the advances taking place. Whilst some of the literature embodies the translation of existing economic models into an SD format (which is a laudable objective) he calls for more economic dynamics models to be built from scratch embodying the best practice in SD modeling. Economic policy is too important to be informed by a single, seemingly unassailable, modeling methodology and it is to be hoped that in the future SD will become even more accepted as a viable tool for use in this field.

In a similar vein comes the contribution of Saeed (see ► Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Allevation). He takes an economic modeling perspective and describes an SD model which explains resource allocation, production and entitlements in a market economy. Its purpose is to understand better how poverty might be reduced in the context of the redistribution of income. A comprehensive listing of the model is provided in an appendix.

The application of SD to public policy generally is dealt with by Andersen, Rich and MacDonald (see ► Public Policy, System Dynamics Applications to). They emphasize how public policy issues are complex, cross organizational boundaries, involve stakeholders with widely different perspectives and evolve over time, such that longer term results may be wholly different from short-term outcomes. Detail is provided for one public policy case involving the Governor's Office of Regulatory Assistance in New York State. They conclude with coverage of studies in a range of public domains such as defense, health care, education and the environment.

One area of SD application has brought the methodology into the legal arena. Disruption and delay in the execution of complex projects invariably finds two parties in dispute. Such disputes often center upon time delays and use of resources on projects – and what might have happened if things had been managed differently. SD models have been employed by parties to such disputes to attempt to justify the occurrence of these events. Howick, Ackermann, Eden and Williams (see ► Delay and Disruption in Complex Projects) report on how cognitive mapping, cause mapping and SD can be fused into what they de-

scribe as a cascade model building process. The result is a rigorous process for explaining why a project behaved in a certain way.

New products and processes are emerging at an ever-increasing rate in modern times. We need to understand the myriad mechanisms which are the basis for their rate of adoption. Milling and Maier range over various SD models which have been created to understand and improve the management of the diffusion of innovations (see ► Diffusion of Innovations, System Dynamics Analysis of the). From the often-cited Bass diffusion model (1969) the authors develop a series of additional features in a modular fashion. These features include competition, network externalities, dynamic pricing and research and development. They conclude by stressing how it is not possible to offer general recommendations for strategies in dynamic and complex environments; such recommendations can only be given in the context of the specific case under scrutiny.

System Dynamics and Its Contribution to Economics and Economic Modeling

Michael J. Radzicki
Worcester Polytechnic Institute, Worcester, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Types of Dynamic Simulation
Translating Existing Economic Models into a System Dynamics Format
Improving Existing Economic Models with System Dynamics
Creating Economic Dynamics Models from Scratch
Model Validity
Controversies
Future Directions
Bibliography

Glossary

Stock Stocks, which are sometimes referred to as "levels" or "states", accumulate (i. e., sum up) the information or material that flows into and out of them. Stocks are thus responsible for decoupling flows, creating delays, preserving system memory, and altering the time shape of flows.

Flow Flows of information or material enter and exit a system's stocks and, in so doing, create a system's dynamics. Stated differently, the net flow into or out of a stock is the stock's rate of change. When human decision making is represented in a system dynamics model, it appears in the system's flow equations. Mathematically, a system's flow equations are ordinary differential equations and their format determines whether or not a system is linear or nonlinear.

Feedback Feedback is the transmission and return of information about the amount of information or material that has accumulated in a system's stocks. When the return of this information reinforces a system's behavior, the loop is said to be positive. Positive loops are responsible for the exponential growth of a system over time. Negative feedback loops represent goal seeking behavior in complex systems. When a negative loop detects a gap between the amount of information or material in a system's stock and the desired amount of information or material, it initiates corrective action. If this corrective action is not significantly delayed, the system will smoothly adjust to its goal. If the corrective action is delayed, however, the system can overshoot or undershoot its goal and the system can oscillate.

Full information maximum likelihood with optimal filtering FIMLOF is a sophisticated technique for estimating the parameters of a system dynamics model, while simultaneously fitting its output to numerical data. Its intellectual origins can be traced to control engineering and the work of Fred Schwepe. David Peterson pioneered a method for adapting FIMLOF for use in system dynamics modeling.

Definition of the Subject

System dynamics is a computer modeling method that has its intellectual origins in control engineering, management science, and digital computing. It was originally created as a tool to help managers better understand and control corporate systems. Today it is applied to problems in a wide variety of academic disciplines, including economics. Of note is that system dynamics models often generate behavior that is both counterintuitive and at odds with traditional economic theory. Historically, this has caused many system dynamics models to be evaluated critically, especially by some economists. However, today economists from several schools of economic thought are beginning to

use system dynamics, as they have found it useful for incorporating their nontraditional ideas into formal models.

Introduction

System dynamics is a computer simulation modeling methodology that is used to analyze complex nonlinear dynamic feedback systems for the purposes of generating insight and designing policies that will improve system performance. It was originally created in 1957 by Jay W. Forrester of the Massachusetts Institute of Technology as a method for building computer simulation models of problematic behavior within corporations. The models were used to design and test policies aimed at altering a corporation's structure so that its behavior would improve and become more robust. Today, system dynamics is applied to a large variety of problems in a multitude of academic disciplines, including economics.

System dynamics models are created by identifying and linking the relevant pieces of a system's structure and simulating the behavior generated by that structure. Through an iterative process of structure identification, mapping, and simulation a model emerges that can explain (mimic) a system's problematic behavior and serve as a vehicle for policy design and testing.

From a system dynamics perspective a system's structure consists of stocks, flows, feedback loops, and limiting factors. Stocks can be thought of as bathtubs that accumulate/de-cumulate a system's flows over time. Flows can be thought of as pipe and faucet assemblies that fill or drain the stocks. Mathematically, the process of flows accumulating/de-cumulating in stocks is called integration. The integration process creates all dynamic behavior in the world be it in a physical system, a biological system, or a socioeconomic system. Examples of stocks and flows in economic systems include a stock of inventory and its inflow of production and its outflow of sales, a stock of the book value of a firm's capital and its inflow of investment spending and its outflow of depreciation, and a stock of employed labor and its inflow of hiring and its outflow of labor separations.

Feedback is the transmission and return of information about the amount of information or material that has accumulated in a system's stocks. Information travels from a stock back to its flow(s) either directly or indirectly, and this movement of information causes the system's faucets to open more, close a bit, close all the way, or stay in the same place. Every feedback loop has to contain at least one stock so that a simultaneous equation situation can be avoided and a model's behavior can be revealed recursively. Loops with a single stock are termed minor, while loops containing more than one stock are termed major.

Two types of feedback loops exist in system dynamics modeling: positive loops and negative loops. Generally speaking, positive loops generate self-reinforcing behavior and are responsible for the growth or decline of a system. Any relationship that can be termed a virtuous or vicious circle is thus a positive feedback loop. Examples of positive loops in economic systems include path dependent processes, increasing returns, speculative bubbles, learning-by-doing, and many of the relationships found in macroeconomic growth theory. Forrester [12], Radzicki and Sterman [46], Moxnes [32], Sterman (Chap. 10 in [55]), Radzicki [44], Ryzhenkov [49], and Weber [58] describe system dynamics models of economic systems that possess dominant positive feedback processes.

Negative feedback loops generate goal-seeking behavior and are responsible for both stabilizing systems and causing them to oscillate. When a negative loop detects a gap between a stock and its goal it initiates corrective action aimed at closing the gap. When this is accomplished without a significant time delay, a system will adjust smoothly to its goal. On the other hand, if there are significant time lags in the corrective actions of a negative loop, it can overshoot or undershoot its goal and cause the system to oscillate. Examples of negative feedback processes in economic systems include equilibrating mechanisms ("auto-pilots") such as simple supply and demand relationships, stock adjustment models for inventory control, any purposeful behavior, and many of the relationships found in macroeconomic business cycle theory. Meadows [27], Mass [26], Low [23], Forrester [12], and Sterman [54] provide examples of system dynamics models that generate cyclical behavior at the macro-economic and micro-economic levels.

From a system dynamics point of view, positive and negative feedback loops fight for control of a system's behavior. The loops that are dominant at any given time determine a system's time path and, if the system is nonlinear, the dominance of the loops can change over time as the system's stocks fill and drain. From this perspective, the dynamic behavior of any economy – that is, the interactions between the trend and the cycle in an economy over time – can be explained as a fight for dominance between the economy's most significant positive and negative feedback loops.

In system dynamics modeling, stocks are usually conceptualized as having limits. That is, stocks are usually seen as being unable to exceed or fall below certain maximum and minimum levels. Indeed, an economic model that can generate, say, either an infinite and/or a negative work-

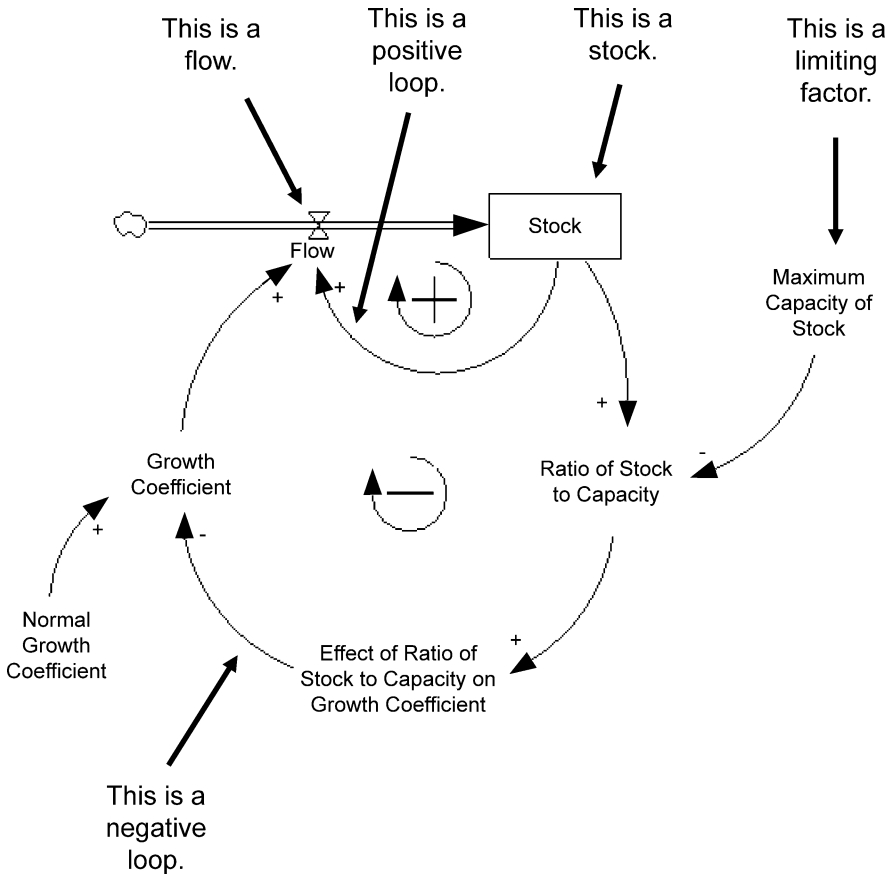

System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 1
Simple system dynamics model containing examples of all components of system structure

force would be seen as severely flawed by a system dynamicist. As such, when building a model system dynamicists search for factors that may limit the amount of material or information that the model's stocks can accumulate. Actual socioeconomic systems possess many limiting factors including physical limits (e. g., the number of widgets a machine can produce per unit of time), cognitive limits (e. g., the amount of information an economic agent can remember and act upon), and financial limits (e. g., the maximum balance allowed on a credit card). When limiting factors are included in a system dynamics model, the system's approach to these factors must be described. Generally speaking, this is accomplished with nonlinear relationships. Figure 1 presents a simple system dynamics model that contains examples of all of the components of system structure described above.

Types of Dynamic Simulation

From a system dynamics point of view, solving a dynamic model – any dynamic model – means determining how much material or information has accumulated in each of a system's stocks at every point in time. This can be accomplished in one of two ways – analytically or via simulation. Linear dynamic models can be solved either way. Nonlinear models, except for a few special cases, can only be solved via simulation.

Simulated solutions to dynamic systems can be attained from either a continuous (analog) computer or a discrete (digital) computer. Understanding the basic ideas behind the two approaches is necessary for understanding how economic modeling is undertaken with system dynamics.

In the real world, of course, time unfolds continuously. Yet, devising a way to mimic this process on a machine is a bit tricky. On an analog computer, the continuous flow of economic variables in and out of stocks over time is mimicked by the continuous flow of some physical substance such as electricity or water. A wonderful example of the later case is the Phillips Machine, which simulates an orthodox Keynesian economy (essentially the IS-LM model) with flows of colored water moving through pipes and ac-

cumulating in tanks. Barr [2] provides a vivid description of the history and restoration of the Phillips Machine.

On a digital computer, the continuous flow of economic variables in and out of stocks over time is approximated by specifying the initial amount of material or information in a system's stocks, breaking simulated time into small increments, inching simulated time forward by one of these small increments, calculating the amount of material or information that flowed into and out of the system's stocks during this small interval, and then repeating. The solution to the system will always be approximate because the increment of time cannot be made infinitesimally small and thus simulated time cannot be made perfectly continuous. In fact, on a digital computer a trade-off exists between round-off error and integration error. If the increment of time is made too large, the approximate solution can be poor due to integration error. If the increment of time is made too small, the approximate solution can be ruined due to round-off error.

In system dynamics modeling the "true" behavior of the underlying system is conceptualized to unfold over continuous time. As such, mathematically, a system dynamics model is an ordinary differential equation model. To approximate the solution to a continuous time ordinary differential equation model on a digital (discrete) computer, however, difference equations are used. Unlike traditional difference equation modeling in economics, in which the increment of time is chosen to match economic data (typically a quarter or a year), the increment of time in system dynamics modeling is chosen to yield a solution that is accurate enough for the problem at hand, yet avoids the problems associated with significant round-off and integration error.

The use of difference equations to approximate the underlying differential equations represented by a system dynamics model provides another interesting option when it comes to economic modeling. Since many well known dynamic economic models have been created with difference equations, they can be recast in a system dynamics format by using the difference equations in the system dynamics software literally as difference equations, and not as a tool to approximate the underlying continuous time system. Although doing this deviates from the original ideas embodied in the system dynamics paradigm, it is occasionally done when a modeler feels that analyzing a difference equation model in a system dynamics format will yield some additional insight.

Translating Existing Economic Models into a System Dynamics Format

There are three principle ways that system dynamics is used for economic modeling. The first involves translating an existing economic model into a system dynamics format, while the second involves creating an economic model from scratch by following the rules and guidelines of the system dynamics paradigm. Forrester [7], Richardson and Pugh [47], Radzicki [42], and Sterman [55] provide extensive details about these rules and guidelines. The former approach is valuable because it enables well-known economic models to be represented in a common format, which makes comparing and contrasting their as-

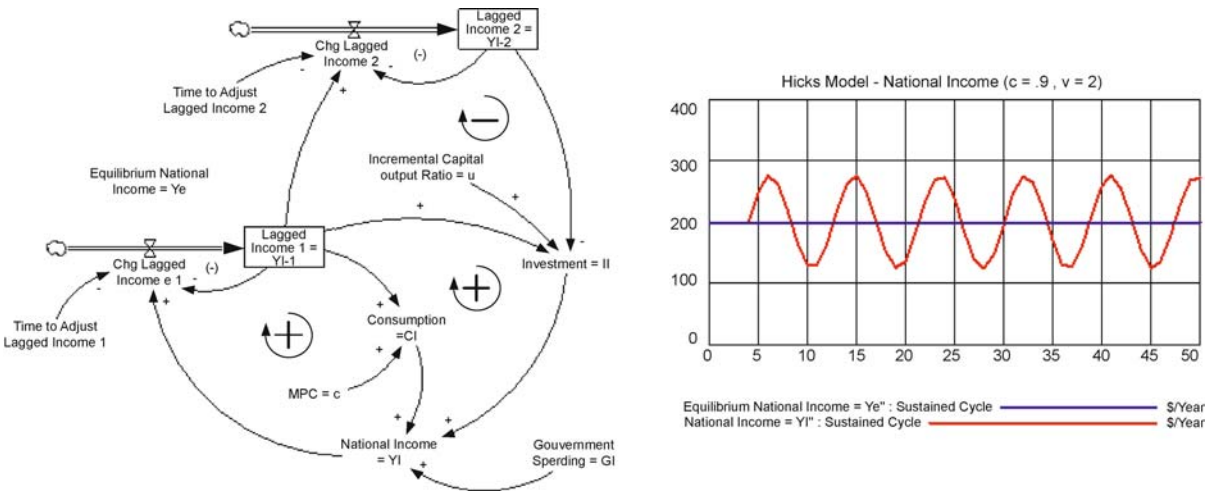

System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 2
System dynamics representation of John Hicks' multiplier-accelerator difference equation model

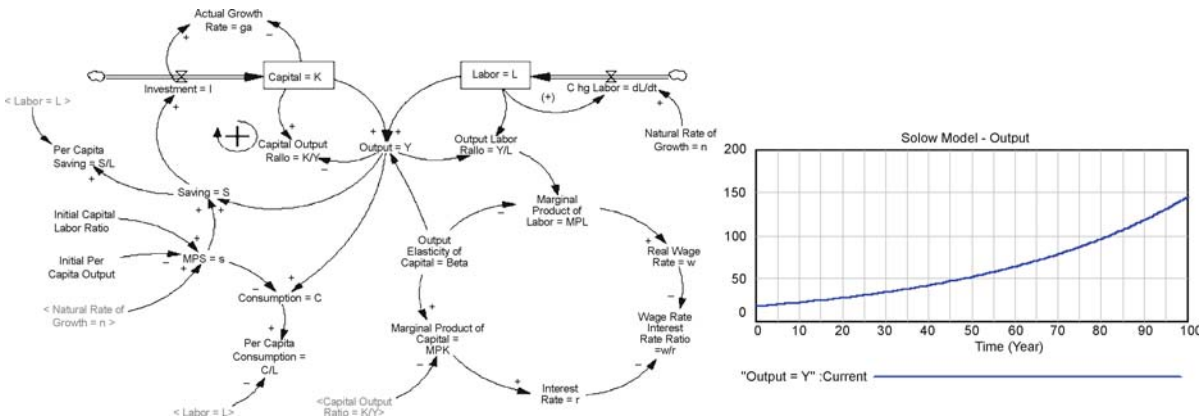

System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 3
System dynamics representation of Robert Solow's ordinary differential equation growth model

sumptions, concepts, structures, behaviors, etc., fairly easy. The latter approach is valuable because it usually yields models that are more realistic and that produce results that are "counterintuitive" [11] and thus thought-provoking.

The third way that system dynamics can be used for economic modeling is a "hybrid" approach in which a well known economic model is translated into a system dynamics format, critiqued, and then improved by modifying it so that it more closely adheres to the principles of system dynamics modeling. This approach attempts to blend the advantages of the first two approaches, although it is more closely related to the former.

Generally speaking, existing economic models that can be translated into a system dynamics format can be divided into four categories: written, static (mathematical), difference equation, and ordinary differential equation. Existing economic models that have been created in either a difference equation or an ordinary differential equation format can be translated into system dynamics in a fairly straight-forward manner. For example, Fig. 2 presents Sir John Hicks' [21] Multiplier-Accelerator difference equation model in a system dynamics format and Fig. 3 presents the Robert Solow's [52] ordinary differential equation growth model in a system dynamics format.

Translating existing static and written economic models and theories into a system dynamics format is a more formidable task. Written models and theories are often dynamic, yet are described without mathematics. Static models and theories are often presented with mathematics, but lack equations that describe the dynamics of any adjustment processes they may undergo. As such, system dynamicists must devise equations that capture the dynamics being described by the written word or that reveal the adjustment processes that take place when a static system moves from one equilibrium point to another.

An interesting example of a system dynamics model that was created from a written economic model is Barry Richmond's [48] model of Adam Smith's *Wealth of Nations*. This model was created principally from Robert Heilbronner's [20] written description of Smith's economic system. A classic example of a static model that has been translated into a system dynamics format is a simple two sector Keynesian cross model, as is shown in Fig. 4.

Improving Existing Economic Models with System Dynamics

The simple two sector Keynesian cross model presented in Fig. 4 is an example of a well known economic model that can be improved after it has been translated into a system dynamics format. More specifically, in this example the flow of investment spending in the model does not accumulate anywhere. This violates good system dynamics modeling practice and can be fixed. Figure 5 presents the improved version of the Keynesian Cross model, which now more closely adheres to the system dynamics paradigm. Other well known examples of classic economics models that have been improved after they have been translated into a system dynamics format and made to conform more closely with good system dynamics modeling practice include the cobweb model [27], Sir John Hicks' multiplier-accelerator model [23], the IS-LM/AD-AS model [13,59], Dale Jorgenson's investment model [51], William Nordhaus' [34] DICE climate change model [4,5], and basic micro economic supply and demand mechanisms [24]. Low's improvement of Hicks' model is particularly interesting because it results

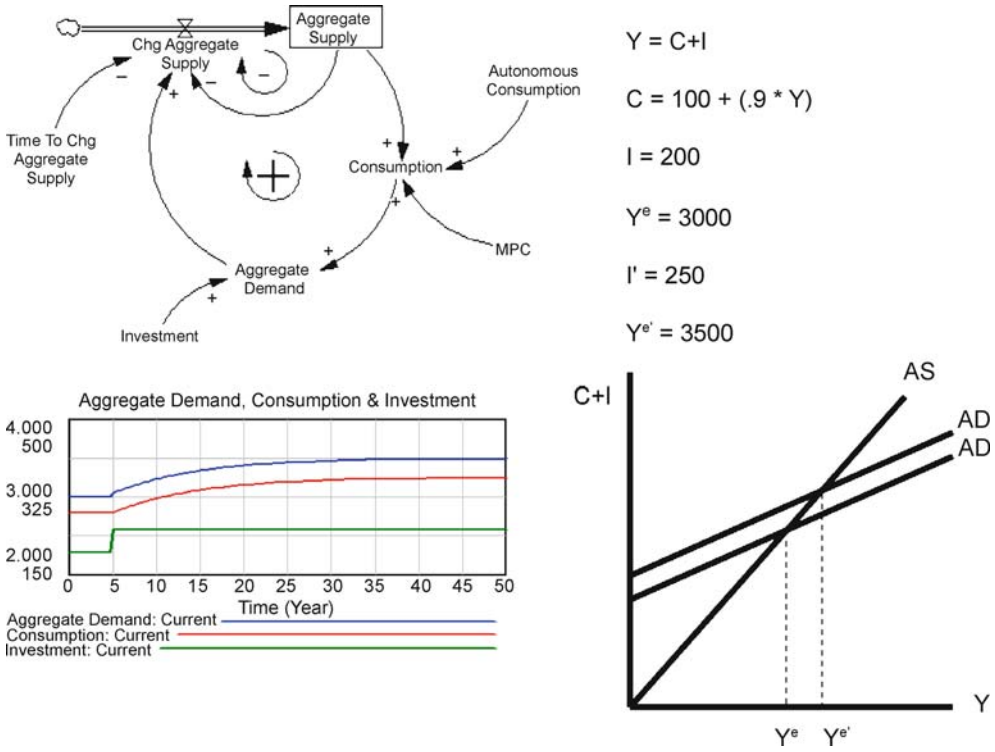

System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 4
Simple two sector Keynesian cross model in a system dynamics format

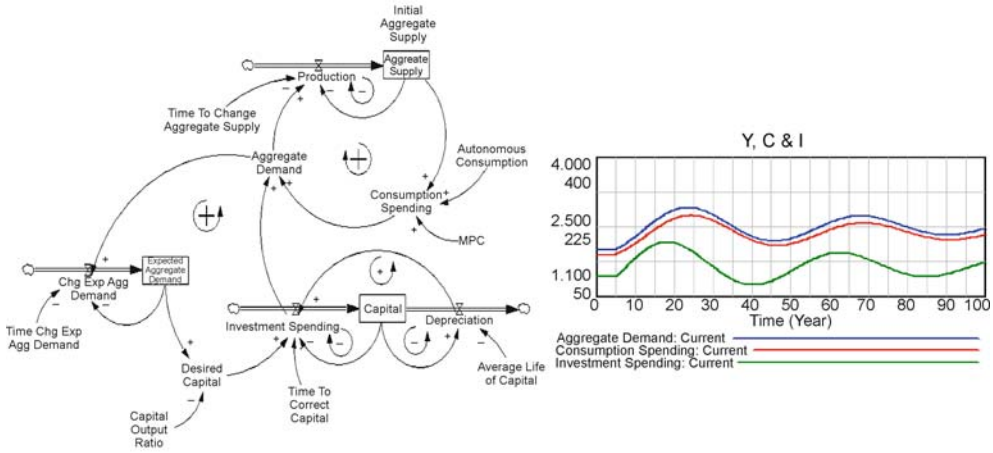

System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 5
Improved simple two sector Keynesian cross model

in a model that closely resembles Bill Phillips' [40] multiplier-accelerator model. Senge and Fiddaman's contributions are also very interesting because they demonstrate how the original economic models are special cases of their more general system dynamics formulations.

Creating Economic Dynamics Models from Scratch

Although translating well known economic models into a system dynamics format can arguably make them easier to understand and use, system dynamicists believe that the

"proper" way to model an economic system that is experiencing a problem is to do so from scratch while following good system dynamics modeling practice. Unlike orthodox economists who generally follow a deductive, logical positivist approach to modeling, system dynamicists follow an inductive pattern modeling or case study process. More specifically, a system dynamicist approaches an economic problem like a detective who is iteratively piecing together an explanation at a crime scene. All types of data that are deemed relevant to the problem are considered including numerical, written, and mental information. The system dynamicist is guided in the pattern modeling process by the perceived facts of the case, as well as by real typologies (termed "generic structures" in system dynamics) and principles of systems. Real typologies are commonalities that have been found to exist in different pattern models and principles of systems are commonalities that have been found to exist in different real typologies. Paich [36] discusses generic structures at length and Forrester [8] lays out a set of principles of systems.

Examples of a real typologies in economics include Forrester's [9] *Urban Dynamics* model, which can reproduce the behavior of many different cities when properly parametrized for those cities, and Homer's [22] model of the diffusion of new medical technologies into the market place, which can explain the behavior of a wide variety of medical technologies when properly parametrized for those technologies. Examples of fundamental principles of systems include the principle of accumulation, which states that the dynamic behavior of any system is due to flows accumulating in stocks, and the notion of stocks and flows being components of feedback loops. The parallels for these principles in economics can be found in modern Post Keynesian economics, in which modelers try to build "stock-flow consistent models," and in institutional economics, in which the principle of "circular and cumulative causation" is deemed to be a fundamental cause of economic dynamics. Radzicki [41,43,45] lays out the case for the parallels that exist between methodological concepts in system dynamics and methodological concepts in various schools of economic thought.

The economic models that have been historically created from scratch by following the system dynamics paradigm have tended to be fairly large in scale. Forrester's [12] national economic model is a classic example, as are the macroeconomic models created by Sterman [53], the Millennium Institute [31], Radzicki [45], Wheat [59], and Yamaguchi [60]. Dangerfield [3] has developed a model of Sarawak (E. Malaysia) to analyze and plan for economic transition from a production economy to a knowledge-based one. With the exception of Radzicki [45], whose model is based on ideas from Post Keynesian and institutional economics, these models, by and large, embody orthodox economic relationships.

Model Validity

When a system dynamics model of an economic system that is experiencing a problem is built from scratch, the modeling process is typically quite different from that which is undertaken in traditional economics. As such, the question is raised as to whether or not an original system dynamics model is in any sense "valid".

System dynamicists follow a "pattern modeling" approach [41] and do not believe that models should be judged in a binary fashion as either "valid" or "invalid". Rather, they argue that confidence in models can be generated along multiple dimensions. More specifically, system dynamicists such as Peterson [38], Forrester and Senge [16] and Barlas [1] have developed a comprehensive series of tests that can be applied to a model's structure and behavior and they argue that the more tests a model can pass, the more confidence a model builder or user should place in its results. Even more fundamentally, however, Forrester [13] has argued that the real value generated through the use of system dynamics comes, not from any particular model, but from the modeling *process* itself. In other words, it is through the iterative *process* of model conceptualization, creation, simulation, and revision that true learning and insight are generated, and *not* through interaction with the resulting model.

Another issue that lies under the umbrella of model validity involves fitting models to time series data so that parameters can be estimated and confidence in model results can be raised. In orthodox economics, of course, econometric modeling is almost universally employed when doing empirical research. Orthodox economic theory dictates the structure of the econometric model and powerful statistical techniques are used to tease out parameter values from numerical data.

System dynamicists, on the other hand, have traditionally argued that it is not necessary to tightly fit models to time series data for the purposes of parameter estimation and confidence building. This is because:

1. the battery of tests that are used to build confidence in system dynamics models go well beyond basic econometric analysis;
2. the particular (measured) time path that an actual economic system happened to take is merely one of an infinite number of paths that it could have taken and is a result of the particular stream of random shocks that

happened to be historically processed by its structure. As such, it is more important for a model to mimic the basic character of the data, rather than fit it point-by-point [14];
3. utilizing the pattern modeling/case study approach enables the modeler to obtain parameter values via observation below the level of aggregation in the model, rather than via statistical analysis [18];
4. the result of a system dynamics modeling intervention is typically a set of policies that improve system performance and increase system robustness. Such policies are usually feedback-based rules (i. e., changes to institutional structure) that do not require the accurate point prediction of system variables.

Although the arguments against the need to fit models to time series data are well known in system dynamics, many system dynamicists feel that it is still a worthwhile activity because it adds credibility to a modeling study. Moreover, in modern times, advances in software technology have made this process relatively easy and inexpensive. Although several techniques for estimating the parameters of a system dynamics model from numerical data have been devised, perhaps the most interesting is David Peterson's [38,39] Full Information Maximum Likelihood with Optimal Filtering (FIMLOF). Figure 5 presents a run from the Harrod growth model, to which an adaptive expectations structure has been added, after it has been fit via FIMLOF to real GDP and labor supply data for the United States economy for the years 1929–2002. The fit is excellent and the estimated parameter values are consistent with those from more traditional econometric studies. See Radzicki [44] for a detailed description of the model and its parameter estimates.

Controversies

Since system dynamics modeling is undertaken in a way that is significantly different from traditional economic modeling, it should come as no surprise that many economists have been extremely critical of some system dynamics models of economic systems. For example, Forrester's [9] *Urban Dynamics* and [10] *World Dynamics* models have come under severe attack by economists, as has (to a lesser degree) his national economic model. On the other hand, the first paper in the field of system dynamics is Forrester [6], which is essentially a critique of traditional economic modeling.

Greenberger et al. [19] present a nice overview of the controversies surrounding the *Urban Dynamics* and *World Dynamics* models. Forrester and his colleagues' replies to criticisms of the *Urban Dynamics* model are contained in Mass [25] and Schroeder et al. [50].

One of the harshest critics of the *World Dynamics* (WORLD2) model has been Nordhaus [33]. Nordhaus [35] has also very critical of the well known follow-up study to *World Dynamics* known as *The Limits to Growth* [28]. Meadows et al. [29,30] contain updates to the original *Limits to Growth* (WORLD3) model, as well as replies to the world modeling critics.

Forrester [12] presents a nice overview of his national economic model, and the critiques by Stolwijk [57] and Zellner [61] are typical of the attitude of the professional economists toward macroeconomic modeling that is undertaken by following the traditional system dynamics paradigm. The criticism of Forrester's national economic model by the economics profession has probably been less severe, relative to the criticisms of the *Urban Dynamics* and world models, because most of its details are still largely unpublished at the time of this writing.

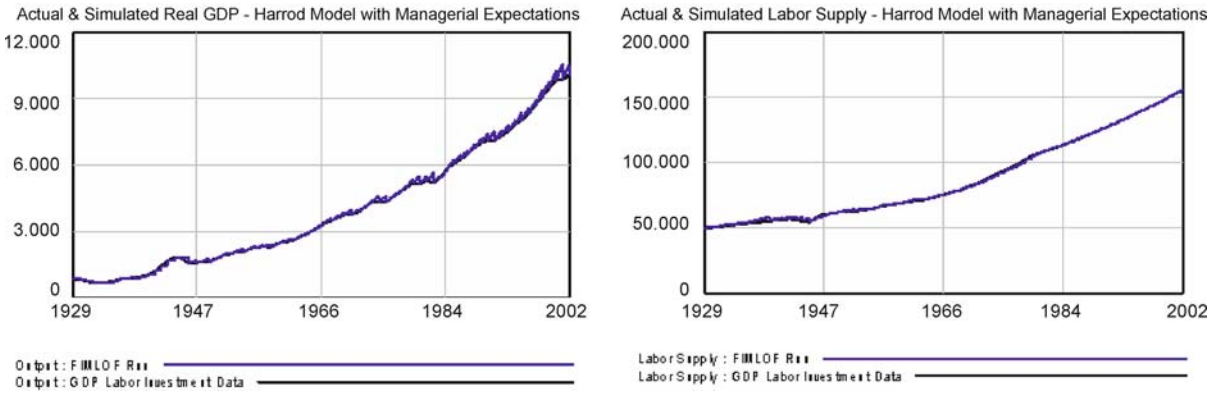

System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 6
Fit of the Harrod growth model to US macroeconomic data for the years 1929–2002

Another interesting and timely example of the sort of controversy surrounding system dynamics modeling in economics is provided by Sterman and Richardson [56]. In this paper they present a technique for testing whether Hubbert's lifecycle method or the geologic analogy method yields superior estimates of the ultimately recoverable amount of petroleum resources. This study was motivated by a disagreement with a traditionally trained economist over the proper way to conceptualize this issue. Sterman and Richardson devised a clever synthetic data experiment in which a system dynamics model serves as the "real world" with a known ultimately recoverable amount of oil. Hubbert's method and the geologic analogy method are then programmed into the model so they can "watch" the data being generated by the "real world" and provide dynamic estimates of the "known" ultimately recoverable stock of oil. The results showed that Hubbert's method was quite accurate, although it had a tendency to somewhat underestimate the ultimately recoverable amount of oil, while the geologic analogy method tended to overshoot the resource base quite substantially.

Future Directions

Historically, system dynamicists who have engaged in economic modeling have almost never been trained as professional economists. As such, they have had the advantage of being able to think about economic problems differently from those who have been trained along traditional lines, but have also suffered the cost of being seen as "amateurs" or "boy economists" [41] by members of the economics profession. The good news is that there are currently several schools of economic thought, populated by professional economists, in which system dynamics fits quite harmoniously. These include Post Keynesian economics, institutional economics, ecological economics, and behavioral economics. Historically, the economists in these schools have rejected many of the tenets of traditional economics, including most of its formal modeling methods, yet have failed to embrace alternative modeling techniques because they were all seen as inadequate for representing the concepts they felt were important. However in the modern era, with computers having become ubiquitous and simulation having become in some sense routine, system dynamics is increasingly being accepted as an appropriate tool for use in these schools of economic thought. The future of economics and system dynamics will most probably be defined by the economists who work within these schools of thought, as well as by their students. The diffusion of system dynamics models of economic systems through their translation into user-friendly interactive "learning environments" that are available over the world wide web will most likely also be of great importance (see [24,59]).

Bibliography

Primary Literature

1. Barlas Y (1989) Multiple Tests for Validation of System Dynamics Type of Simulation Models. Eur J Operat Res 42(1):59–87
2. Barr N (1988) The Phillips Machine. LSE Q Winter 2(4):305–337
3. Dangerfield BC (2007) System dynamics advances strategic economic transition planning in a developing nation. In: Qudrat-Ullah H, Spector M, Davidsen P (eds) Complex decision-making: Theory & practice. Springer, New York, pp 185–209
4. Fiddaman T (1997) Feedback complexity in integrated climate-economy models. Ph D Dissertation, Sloan School of Management, Massachusetts Institute of Technology. Available from http://www.systemdynamics.org/
5. Fiddaman T (2002) Exploring policy options with a behavioral climate-economy model. Syst Dyn Rev 18(2):243–267
6. Forrester J (1957) Dynamic models of economic systems and industrial organizations. System Dynamics Group Memo D-0. Massachusetts Institute of Technology. Available from http://www.systemdynamics.org/
7. Forrester J (1961) Industrial dynamics. Pegasus Communications, Inc., Waltham
8. Forrester J (1968) Principles of systems. MIT Press, Cambridge
9. Forrester J (1969) Urban dynamics. Pegasus Communications, Inc., Waltham
10. Forrester J (1971) World dynamics. Pegasus Communications, Inc., Waltham
11. Forrester J (1975) Counterintuitive behavior of social systems. In: Forrester J (ed) Collected papers of Jay W Forrester, Pegasus Communications, Inc., Waltham, pp 211–244
12. Forrester J (1980) Information sources for modeling the national economy. J Am Statist Assoc 75(371):555–567
13. Forrester J (1985) 'The' model versus a modeling 'process'. Syst Dyn Rev 1(1 and 2):133–134
14. Forrester J (2003) Economic theory for the new millennium. In: Eberlein R, Diker V, Langer R, Rowe J (eds) Proceedings of the Twenty-First Annual Conference of the System Dynamics Society. Available from http://www.systemdynamics.org/
15. Forrester J, Low G, Mass N (1974) The debate on world dynamics: A response to Nordhaus. Policy Sci 5:169–190
16. Forrester J, Senge P (1980) Tests for building confidence in system dynamics models. In: Legasto Jr AA, Forrester JW, Lyneis JM (eds) TIMS Studies in the Management Sciences: System Dynamics, vol 14. North Holland Publishing Company, Amsterdam, pp 209–228
17. Forrester N (1982) A dynamic synthesis of basic macroeconomic theory: Implications for stabilization policy analysis, Ph D Dissertation, Alfred P Sloan School of Management, Massachusetts Institute of Technology. Available from http://www.systemdynamics.org/
18. Graham A (1980) Parameter estimation in system dynamics modeling. In: Randers J (ed) Elements of the system dynamics method. Pegasus Communications, Inc., Waltham, pp 143–161
19. Greenberger M, Crenson M, Crissey B (1976) Models in the policy process: Public decision making in the computer era, Russell Sage Foundation, New York

20. Heilbroner R (1980) The worldly philosophers, 5th edn. Simon & Schuster, New York
21. Hicks J (1950) A contribution to the theory of the trade cycle. Oxford University Press, London
22. Homer J (1987) A diffusion model with application to evolving medical technologies. Technol Forecast Soc Change 31(3):197–218
23. Low G (1980) The multiplier-accelerator model of business cycles interpreted from a system dynamics perspective. In: Randers J (ed) Elements of the system dynamics method. Pegasus Communications, Inc., Waltham, pp 76–94
24. Mashayekhi A, Vakili K, Foroughi H, Hadavandi M (2006) Supply demand world: An interactive learning environment for teaching microeconomics. In: Grosler A, Rouwette A, Langer R, Rowe J, Yanni J (eds) Proceedings of the of the Twenty-Fourth International Conference of the System Dynamics Society. Available at http://www.systemdynamics.org/
25. Mass N (1974) Readings in urban dynamics, vol I. Wright-Allen Press, Cambridge
26. Mass N (1975) Economic cycles: An analysis of the underlying causes. MIT Press, Cambridge
27. Meadows D (1970) Dynamics of commodity production cycles. Massachusetts MIT Press, Cambridge
28. Meadows D, Meadows D, Randers J, Behrens III W (1972) The limits to growth: A report for the Club of Rome's project on the predicament of mankind. Universe Books, New York
29. Meadows D, Meadows D, Randers J (1992) Beyond the limits: Confronting global collapse, envisioning a sustainable future. Chelsea Green Publishing, White River Junction
30. Meadows D, Meadows D, Randers J (2002) Limits to growth: The 30-year update. Chelsea Green Publishing, White River Junction
31. Millennium Institute (2007) Introduction and Purpose of Threshold 21, http://www.millennium-institute.org/resources/elibrary/papers/T21Overview.pdf
32. Moxnes E (1992) Positive feedback economics and the competition between 'hard' and 'soft' energy supplies. J Sci Ind Res 51(March):257–265
33. Nordhaus W (1972) World dynamics: Measurement without data. Econom J 83(332):1156–1183
34. Nordhaus W (1992) The "DICE" model: Background and structure of a dynamic integrated climate-economy model of the economics of global warming. Cowles Foundation for Research in Economics at Yale University, Discussion Paper No. 1009
35. Nordhaus W (1992) Lethal model 2: The limits to growth revisited. Brookings Pap Econo Activity 2:1–59
36. Paich M (1985) Generic structures. Syst Dyn Rev 1(1 and 2):126–132
37. Paich M (1994) Managing the global commons. MIT Press, Cambridge
38. Peterson D (1975) Hypothesis, estimation, and validation of dynamic social models – energy demand modeling. Ph D Dissertation, Department of Electrical Engineering, Massachusetts Institute of Technology. Available from http://www.systemdynamics.org/
39. Peterson D (1980) Statistical tools for system dynamics. In: Randers J (ed) Elements of the system dynamics method. Pegasus Communications, Inc., Waltham, pp 226–245
40. Phillips W (1954) Stabilization policy in a closed economy. Econom J 64(254):290–323
41. Radzicki M (1990) Methodologia oeconomiae et systematis dynamis. Syst Dyn Rev 6(2):123–147
42. Radzicki M (1997) Introduction to system dynamics. Free web-based system dynamics tutorial. Available at http://www.systemdynamics.org/DL-IntroSysDyn/index.html
43. Radzicki M (2003) Mr. Hamilton, Mr. Forrester and a foundation for evolutionary economics. J Econom Issues 37(1):133–173
44. Radzicki M (2004) Expectation formation and parameter estimation in nonergodic systems: the system dynamics approach to post Keynesian-institutional economics. In: Kennedy M, Winch G, Langer R, Rowe J, Yanni J (eds) Proceedings of the Twenty-Second International Conference of the System Dynamics Society. Available at http://www.systemdynamics.org/
45. Radzicki M (2007) Institutional economics, post keynesian economics, and system dynamics: Three strands of a heterodox economics braid. In: Harvey JT, Garnett Jr. RF (eds) The future of heterodox economics. University of Michigan Press, Ann Arbor
46. Radzicki M, Sterman J (1994) Evolutionary economics and system dynamics. In: Englund R (ed) Evolutionary concepts in contemporary economics. University of Michigan Press, Ann Arbor, pp 61–89
47. Richardson G, Pugh A (1981) Introduction to system dynamics modeling with DYNAMO. Pegasus Communications, Inc., Waltham
48. Richmond B (1985) Conversing with a classic thinker: An illustration from economics. Users Guide to STELLA, Chapt 7. High Performance Systems, Inc., Lyme, New Hampshire, pp 75–94
49. Ryzhenkov A (2007) Controlling employment, profitability and proved non-renewable reserves in a theoretical model of the US economy. In: Sterman J, Oliva R, Langer R, Rowe J, Yanni J (eds) Proceedings of the of the Twenty-Fifth International Conference of the System Dynamics Society. Available at http://www.systemdynamics.org/
50. Schroeder W, Sweeney R, Alfeld L (1975) Readings in Urban Dynamics, vol II. Wright-Allen Press, Cambridge
51. Senge P (1980) A system dynamics approach to investment function formulation and testing. Soc-Econ Plan Sci 14:269–280
52. Solow R (1956) A contribution to the theory of economic growth. Q J Econom 70:65–94
53. Sterman J (1981) The energy transition and the economy: A system dynamics approach. Ph D Dissertation, Alfred P Sloan School of Management, Massachusetts Institute of Technology. Available at http://www.systemdynamics.org/
54. Sterman J (1985) A behavioral model of the economic long wave. J Econom Behav Organ 6:17–53
55. Sterman J (2000) Business dynamics: Systems thinking and modeling for a complex world. Irwin-McGraw-Hill, New York
56. Sterman J, Richardson G (1985) An experiment to evaluate methods for estimating fossil fuel resources. J Forecast 4(2):197–226
57. Stolwijk J (1980) Comment on 'information sources for modeling the national economy' by Jay W Forrester. J Am Stat Assoc 75(371):569–572
58. Weber L (2007) Understanding recent developments in growth theory. In: Sterman J, Oliva R, Langer R, Rowe J, Yanni J (eds) Proceedings of the of the Twenty-Fifth International Conference of the System Dynamics Society. Available at http://www.systemdynamics.org/

59. Wheat D (2007) The feedback method of teaching macroeconomics: Is it effective? Syst Dyn Rev 23(4):391–413
60. Yamaguchi K (2007) Balance of payments and foreign exchange dynamics – S D Macroeconomic Modeling (4). In: Sterman J, Oliva R, Langer R, Rowe J, Yanni J (eds) Proceedings of the of the Twenty-Fifth International Conference of the System Dynamics Society. Available at http://www.systemdynamics.org/
61. Zellner A(1980) Comment on 'information sources for modeling the national economy' by Jay W Forrester. J Am Stat Assoc 75(371):567–569

Books and Reviews

Alfeld L, Graham A (1976) Introduction to urban dynamics. Wright-Allen Press, Cambridge
Lyneis J (1980) Corporate planning and policy design: A system dynamics approach. PA Consulting, Cambridge
Meadows D, Meadows D (1973) Toward global equilibrium: Collected papers. Wright-Allen Press, Cambridge
Meadows D, Behrens III W, Meadows D, Naill R, Randers J, Zahn E (1974) Dynamics of growth in a finite world. Wright-Allen Press, Cambridge
Meadows D, Robinson J (1985) The electronic oracle: Computer models and social decisions. Wiley, New York
Richardson G (1999) Feedback thought in social science and systems theory. Pegasus Communications, Inc., Waltham
Sterman J (1988) A skeptic's guide to computer models. In: Grant L (ed) Foresight and National Decisions. University Press of America, Lanham, pp 133–169, Revised and Reprinted in: Barney G, Kreutzer W, Garrett M (1991) Managing a Nation. Westview Press, Boulder, pp 209–230

System Dynamics Modeling: Validation for Quality Assurance

Markus Schwaninger, Stefan Groesser
Institute of Management, University of St. Gallen, St. Gallen, Switzerland

Article Outline

Glossary
Definition of the Subject
Introduction
Epistemological Foundations
Validation Methods
Validation Process
Synopsis and Outlook
Appendix: Overview of the Tests Described in This Chapter
Bibliography

Glossary

Model/model system A model is a simplified representation of a real system. Models can be descriptive or prescriptive (normative). Their functions can be to enable explanation, anticipation or design. A distinction used in this contribution is between causal and non-causal models, with System Dynamics models being of the former type. The term *model system* is used to stress the systemic character of a model; this serves to identify it as an organized whole of variables and relationships on the one hand, and to distinguish it from the *real system* which is to be modeled, on the other.

Model validity A model's property of adequately reflecting the system modeled. Validity is the primary measure of model quality. It is a matter of degree, not a dichotomized property.

Model purpose The goal for which a model is designed or the function it is intended to fulfill. The model purpose is closely linked to the end-model user or model owner. Model purpose is the criterion for the choice of a model's boundary and design.

Modeling process The process involving phases such as problem articulation, boundary selection, development of a dynamic hypothesis, model formulation, model testing, policy formulation and policy evaluation [28]. The modeling process is followed by model use and implementation, i. e., the realization of actions designed or facilitated by the use of the model.

Validation process Validation is the process by which model validity is enhanced systematically. It consists in gradually building confidence in the usefulness of a model by applying validation tests as outlined in this chapter. In principle, validation pervades all phases of the modeling process, and, in addition, extends into the phases of model use and implementation.

Definition of the Subject

The present chapter addresses the question of building better models. This is crucial for coping with complexity in general, and in particular for the management of dynamic systems. Both the epistemological and the methodological-technological aspects of model validation for the achievement of high-quality models are discussed. The focus is on formal models, i. e. those formulated in a stringent, logical, and mostly mathematical language.

Introduction

The etymological root of *valid* is the Latin word *validus*, which denotes attributes such as strong, powerful and firm. A valid model, then, is well-founded and difficult to reject because it accurately represents the perceived real system which it is supposed to reflect. This system can

be either one that already exists or one that is being constructed, or even anticipated, by a modeler or a group of modelers.

Validation standards in System Dynamics are more rigorous than those of many other methodologies. Let us distinguish between two types of mathematical models, which are fundamentally different: Causal, theory-like models and non-causal, statistical (correlational) models [4]. The former are explanatory, i. e., they embody theory about the functioning of a real system. The latter are descriptive and express observed associations among different elements of a real system. System Dynamics models are causal models.

Non-causal models are tested globally, in that the statistical fit between model and data series from the real system under study is assessed. If the fit is satisfactory, the model is considered to be accurate ("valid", "true"). In contrast, system dynamicists postulate that models be not only right, but right for the right reasons. As the models are made up of causal interdependencies, accuracy is required for each and every variable and relationship. The following principle applies: if only one component of the model is shown to be wrong, the whole model is rejected even if the overall model output fits the data [4]. This strict standard is conducive to high-quality modeling practice.

A model is an abstract version of a perceived reality. Simulation is a way of experimenting with mathematical models to gain insights and to employ these to improve the real system under study. It is often said that System Dynamics models should portray problems or issues, not systems. This statement must be interpreted in the sense that one should not try to set the boundaries of the model too widely, but rather give the model a focus by concentrating on an object in accordance with the specific purpose of the model. In a narrower definition, even an issue or problem can be conceived of as a "system", i. e., "a portion of the world sufficiently well defined to be the subject of study" [21]. Validity then consists in a stringent correspondence between model system and real system.

We will treat the issue of model validation as a means of assuring high-quality models. We interject that validity is not the only criterion of model quality, other criteria including parsimony, ease-of-use, practicality, importance, etc. [22].

In the following, the epistemological foundations of model validity are reviewed (Sect. "Epistemological Foundations"). Then, an overview of the methods for assuring model validity is given (Sect. "Validation Methods"). Further, the survey includes an overview of the validation process (Sect. "Validation Process") and our final conclusions (Sect. "Synopsis and Outlook").

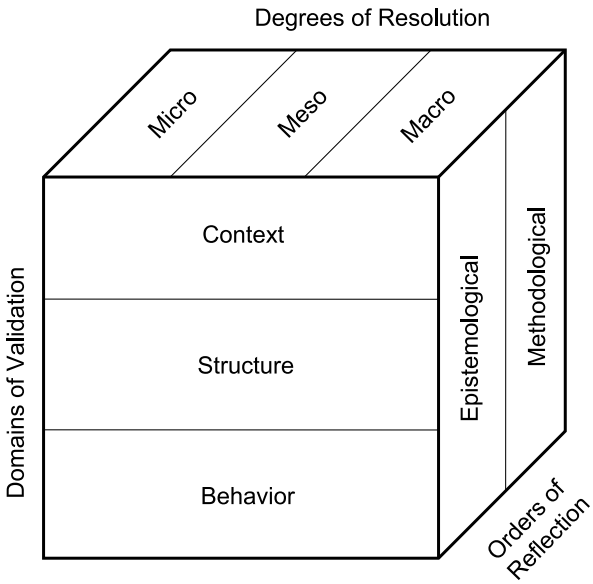

System Dynamics Modeling: Validation for Quality Assurance, Figure 1
The Validation Cube – A frame of reference showing three dimensions of the validation topic

The substance of this article will be made more palpable by means of the following frame of reference. We call it the Validation Cube. The diagram in Fig. 1 shows three dimensions of the validation topic:

- *Orders of Reflection:* We distinguish between an *epistemological* and a *methodological* layer. These define the objects of the next two Sects. "Epistemological Foundations" and "Validation Methods".
- *Domains of Validation:* The three domains, *context*, *structure* and *behavior* refer to the groups of validation methods as described in Sect. "Validation Methods".
- *Degrees of Resolution:* We address the different granularities of models. *Micro* refers to the smallest building blocks of models (e. g., variables or small sets of variables), *meso* to modules which constitute a model, and *macro* to the model as a whole.

Epistemological Foundations

Epistemology is the theory that enquires into the nature and grounds of knowledge: "What can we know and how do we know it?" [13]. These questions are of utmost importance when dealing with models and their validity, because a method of validation is only as good as its epistemological basis.

We can only briefly refer to the antecedents of the epistemological perspective inherent in the idea of model validation as commonly held today in the community of sys-

tem dynamicists. One could go back to Socrates who, in Plato's *Republic* (fourth century BC), addressed the problematic relationship between reality, image and knowledge. One could also refer to John Locke (seventeenth century), the first British empiricist who maintained that ideas could come only from experience, while admitting that our knowledge about external objects is uncertain. We will address the philosophical movements of the nineteenth and twentieth centuries, which are direct sources of the epistemology which is important for model validation. The reader may kindly excuse us for certain massive simplifications that we are obliged to make.

What will be said here about theories applies equally to formal models. In System Dynamics, models either embody theories or they are considered essential components of theories. In addition, processes of modeling and theory-building are of the same nature; a model, like any theory, is built and improved in a dialectic of propositions and refutations [22].

Positivism and Critique

Positivism is a scientific doctrine founded by Auguste Comte (nineteenth century) which raises the *positive* to the principle of all scientific knowledge. "Positive", in this context, is not meant to be the opposite of negative, but the given, factual, or indubitably existent. The positive is associated with features such as being real, useful, certain, and precise. Positivism confines science to the observable and manipulable, drawing on the mathematical, empirical orientation of the natural sciences as its paragon. The objectivist claim of positivism is that things exist independently of the mind and that truths are detached from human values and beliefs. This stance calls for models that approximate an objective reality.

A younger development in this vein is the school of logical positivism, also logical empiricism (with Schlick, Neurath, Hempel, etc.), which concentrates on the problem of meaning and has developed the verifiability principle: Something is meaningful only if verifiable empirically, i. e., ultimately by observation through the senses. To verify here means to show to be true [13]. For the logical positivists, the method of verification is the essence of theory-building. Tests of theories hinge on their confirmation by facts. In System Dynamics, testing models on real-world data is a core component of validation.

Positivism has been criticized for being reductionist, i. e., for its tendency to reduce concepts to simpler or empirically more accessible ones, and to conceive of learning as an accumulation of particular details. The critique has also asserted that there is no theory-independent identification of facts, and therefore different theories cannot be tested by means of the same data [6]. Another objection maintains that social facts are not merely given, but produced by human action, and that they are subject to interpretation [23]. These arguments introduce the principle of relativity, which is of crucial importance for the field of model validation: A model is a subjective construction by an observer.

Pragmatism – A Challenge to Positivism

Pragmatism, which arose in the second half of the nineteenth century, emphasizes action and the practical consequences of thinking. Its founder, Charles Sanders Peirce, was interested in the effects that the meaning of scientific concepts could have on human experience and action. He defined truth as "the opinion which is fated to be ultimately agreed to by all who investigate" [13], whereby truth is linked to consensual validation. For pragmatists, truth is in what works (Ferdinand Schiller) or satisfies us (John Dewey), and what we find believable and consistent: " 'The true' … is only the expedient in the way of our thinking", and "truth is *made* … in the course of experience." (see p. 581 and p. 583 in [11]).

Pragmatism is often erroneously disdained for supposedly being a crass variety of utilitarianism and embodying a crude instrumentalist rationality. A more accurate view considers the fact that pragmatists are not satisfied with a mere ascertainment of truth; instead they ask: "If an idea or assumption is true, does this make a concrete difference to the life of people? How can this truth be actualized?" In other words, pragmatism does not crudely equate truth and utility. It rather postulates that those truths which are useful to people ought to be put into practice [23].

Pragmatism introduces the criteria of confidence and usefulness, which are more operational as guides to the evaluation of experiments than is the notion of an absolute truth, which is unattainable in the realm of human affairs. At the same time, pragmatism triggers a crucial insight for the context of model-building: The validity of a model depends not only on the absolute quality of that model but also hinges on its suitability with respect to a purpose [7]. In the context of model validation, then, truth is a relative property; more exactly, a *truth* holds for a limited domain only.

More Challenges to Positivism

We discuss three more challenges to positivism in the twentieth century. First, Thomas Kuhn's theory of scientific revolutions [12]: Kuhn shows, by means of historical cases, that in the sphere of science, generally accepted

ways of looking at the world ("paradigms") change over time through fundamental shifts. Therefore, the activities of a scientist are largely shaped by the dominant scientific worldview. Second, Willard Van Orman Quine and Wilfrid Sellars argue that knowledge creation and theory-building is a holistic, conversational process, as opposed to the reductionist and confrontational views [4].

Both of these movements contribute to our understanding of how real systems are to be modeled and validated: as organized wholes, and consciously with respect to the values and beliefs underlying a given modeling process. This approach adheres to the spirit of models themselves, by means of which the behavior of whole systems can be simulated and tested on their inherent assumptions.

A third challenge is presented by the interpretive streams of epistemology (for an overview, see [9]). Among them, a main force which expands the possibilities of scientific methodologies is the strand of hermeneutics. Derived from the Greek *hermeneuein* – to interpret or to explain – the term *hermeneutics* stands for a school, mainly associated with Hans-Georg Gadamer, which pursues the ideal of a human science of understanding. The emphasis is on interpretation in an interplay between a subject-matter and the interpreter's position. This emphasis introduces the subjective into scientific methodology. Hermeneutics denies both that a single "objective true interpretation" can transcend all individual viewpoints, and that humans are forever confined within their own ken [13]. This epistemology offers a necessary complement to a scientific stance, which exclusively hinges on "hard", quantitative methods in order supposedly to achieve absolute objectivity. The implication of hermeneutics for model validation is that it recognizes the pertinence of subjective judgment. In this connection, interpretive discourses play a crucial role in group model-building and validation. Such discourses lead beyond the subjective, entailing the creation of inter-subjective, shared realities. We will return to this factor in Sect. "Validation Process".

Critical Rationalism

Critical rationalism is a philosophical position founded by Karl R. Popper [19,20]. It grew out of positivism but rejected its verificationist stance. Critical rationalism posits that, in the social domain, theories can never be definitely proved, but can only reach greater or lesser levels of truth. Scientific proofs are confined to the realm of the formal sciences, namely logic and mathematics.

As Popper demonstrates, all theories are provisional. As a consequence, the main criterion for the assessment of a theory's truth status is *falsification* [19]. A theory holds as long as it is not refuted. Consequently, any theory can be upheld as long as it passes the test of falsification. In other words, the fertile approaches to science are not those of corroboration, but the falsificationist efforts to test if theories can be upheld. In the context of modeling this means that validation must undertake attempts to falsify a model, thereby testing its robustness.

Even Popper's theory of science is not unchallenged. For example, Kuhn has made the point that its principles are applicable only to normal science, which operates incrementally within a given paradigm, but not to anomalous science, which uncovers unsuspected phenomena in periods of scientific revolution [12]. This observation has an implication for model validation: Alternative and even multiple model designs should be assessed for their ability to account for fundamental change.

On the Meaning of Validity and Validation

One of the predominant convictions about science is the obsessive idea that proofs are the touchstone of the validity of both theories and models. We follow a different rationale, reverting to the philosophy of science as embodied in critical rationalism.

Popper's refutationist concept (as opposed to a verificationist concept) of theory-testing implies both an evolutionist perspective and an empiricist stance. The evolutionist perspective is primary because it welcomes the challenges posed to a theory, since these attempts at falsification lead to an evolutionary process: successful falsification efforts result in revisions and improvements of the theory. Correspondingly, empiricism is paramount in the social sciences, because the main source for the refutation of a theory is empirical evidence. However, falsification can also be grounded in logical arguments where empirical evidence cannot be obtained. In this sense, a structuralist approach as used in System Dynamics validation transcends the bounds of logical empiricism.

As a consequence of the evolutionist perspective, there is no such thing as absolute validity. Validity is always imperfect, but it can be improved over time. The empiricist aspect of theory-building implies that theories must be validated by means of empirical data. However, logical assay, estimation and judgment are complementary to this empiricist component (see below).

A validation process is about gradually building confidence in the model under study [2]. This is both analytical and synthetic. It is directed at the model as a whole as much as it is at the components of the model. The touchstone of validity is less whether the model is right or wrong: as Sterman states, "… all models are wrong." [28].

Some models, however, fulfill the purpose ascribed to them, i. e., they are useful. Models are inherently incomplete; they cannot claim to be true in an absolute sense, but only to be relatively true [4]. In this sense, *validation* is a *goal-oriented* activity and *validity* a *relative* concept.

Finally, the validation process often involves several people because the necessary knowledge is distributed. In these cases, the dialectics of propositions and refutations, as well as the interaction of different subjective viewpoints, and consensus-building, are integral. Validation processes, then, are semiformal, discursive social procedures with a holistic as opposed to a fragmentary orientation [ibidem].

On Objectivity

If subjective views and judgments are as prominent as alleged above, does objectivity play a role at all? Operational philosophy shows a way out of this dilemma: Rapoport defines objectivity as "invariance with respect to different observers." [21]. Popper has a similar stance in proposing that general statements must be formulated in a way that they can be criticized and, where applicable, falsified [20]. This concept of objectivity is a challenge to model validation: When defining concepts and functions, one must first of all strive for falsifiable statements. In principle, formal models meet this criterion: each variable and every function or relationship can be challenged. And they must be challenged, so that their robustness can be tested. The duty, then, is in finding the invariances that are inter-subjectively accepted as the best approximations to truth. Frequently this is best achieved in group model-building processes [30]. Finally, truth is something we search for but do not possess [20], i. e., even an accepted model cannot guarantee truth with final certainty.

Validation Methods

A considerable set of qualitative and quantitative tests has been developed for the enhancement of model validity. The state-of-the-art has been documented in seminal publications [2,4,7,8,14,17,28]. Our purpose here is to present and exemplify the different tests to encourage and help those who strive to develop high-quality System Dynamics models.

In the following, an overview of the types of tests developed for System Dynamics models is given, without any claim to completeness. Most of these tests have been documented extensively in [2,7,8,28]. The descriptions of the tests adhere closely to the specifications of these authors (mainly Forrester and Senge). In addition, we have developed a new category for tests that concentrate on the context in which the model is to be developed. High-quality models can be created only if the relevant context is taken into consideration. To facilitate orientation, we have attached an overview of all described tests in the Appendix.

In this section we describe three groups of tests: those related to model-related context, tests of model structure and tests of model behavior. Many of the tests described in the following can be utilized for explanatory analysis which aims at an understanding of the problematic behavior of the issue under study. Others are suitable for normative ends, in analyzes targeted on improvements of system performance with regard to a specified objective of the reference system. Also known as policy tests, or policy analyses, these "tests of policy implications differ from other tests in their explicit focus on comparing changes in a model and in the corresponding reality. Policy … tests attempt to verify that response of a real system to a policy change would correspond to the response predicted by the model" [8]. Policy testing can show the risk involved in adopting the model for policy making.

Tests About the Model-Related Context

These tests deal with aspects related to the situation in which the model is to be developed and embedded. They imply metalevel decisions which have to be taken in the first place, before engaging in model-building. Applied ex-post-facto, i. e., after modeling, they allow for assessing the utility of the modeling endeavor as such.

Issue Identification Test. The *raison d'être* of a System Dynamics model is its ability to adequately address an issue and to enhance stakeholders' understanding, an ability which may lead to policy insights and system improvements. The issue identification test examines whether or not the identified issue or problem is indeed meaningful. Has the "right" problem been identified? Does the problem statement address the origins of an issue or only superficial *symptoms*? Whenever complex issues are addressed by a model, different perspectives (e. g. professional, economic, political) must be integrated for accurate problem identification and modeling. This is not a "one-shot-only" test; it must be applied recurrently during the modeling procedure. By reflecting regularly on the correctness of the identified issue, the modeler can increase the likelihood of capturing the origins of suboptimal system behavior.

Adequacy of Methodology Test. Simulation models respond to the limitations of humans' mental ability to comprehend complex, dynamic feedback systems [27]. The adequacy of methodology test scrutinizes whether the

System Dynamics methodology is best-suited for dealing with the issue under study. One needs to clearly ascertain if that issue is characterized by dynamic complexity, feedback mechanisms, nonlinear interdependency of structural elements and delays between causes and effects. One needs to ask also if the issue under study could be better addressed by another methodology. For example, in a case where the question is to understand the difference in numerical outcomes between two configurations of a production system, it lets one determine whether discrete event simulation would fulfill this requirement more accurately than System Dynamics.

System Configuration Test. This test asks the fundamental question about whether the structural configuration chosen can be accepted. It challenges the assumption that the model represents the actual working of the system under study. The applicability of a different design would be suggested by its ability to capture new conditions, such as different system configurations, phenomena or rules of the game. Even revolutionary changes might be considered. Such an outlook may require a totally new model, or an alternative model designed from a different vantage point. This would at least feasibly approximate the need to take paradigmatic change into account.

System Improvement Test. The purpose of modeling is to understand a part of reality and to resolve an issue. The system improvement test can be performed only after the modeling project (an ex-post-facto test), once the insights derived from the model have already been implemented in the real system. This test reestablishes the connection between the abstract mathematical model and the real system. The system improvement test helps to evaluate whether or not model development was successful. In operational terms, any improvements of the real system under study must be compared with explicit objectives. In practice, the test might assess the impact of the modeling process or the model use either on the mental models of decision makers or on changes in organization structures. In principle, assessing the impact of a modeling endeavor is very difficult (one preliminary example is provided by Snabe and Grössler [25]).

Tests of Model Structure

Tests of model structure refer to the "nuts and bolts" of System Dynamics modeling, i. e., to the formal concepts and interrelationships which represent the real system. Model structure tests aim to increase confidence in the structure of the created theory about the behavior mode of interest. The model structure can be assessed by means of either direct or indirect inspection. Tests of model structure assess whether the logic of the model is attuned to the corresponding structure in the real world. They do not yet compare the model behavior with time series data from the real system.

Direct Structure Tests Direct structure tests assess whether or not the model structure conforms to relevant descriptive knowledge about the real system or class of systems under study. By means of direct comparison, they qualitatively assess any disparities between the original system structure and the model structure.

Structure Examination Test. Examination in this case means comparison in the sense just outlined. Qualitative or quantitative information about the real system structure can be obtained either empirically or theoretically. Empirically based tests include reviews of model assumptions about system elements and their interdependencies, e. g., reviews made by highly knowledgeable experts of the real system. Theory-based tests compare the model structure with theoretical knowledge from literature about the type of system being studied. Thereby, a preference for theoretical knowledge specific to the modeled situation over more abstract and general knowledge is usually the case.

To pass the structure examination test, a model must not contradict either the evidence or knowledge about the structure of the real system. This test ensures that the model contains only those structural elements and interconnections that are most likely extant in the real system. In this context, formal inspections of the model's equations, reviews of the syntax for the stock and flow diagram, and walkthroughs along the causal loop diagrams and their embodied causal explanations may be indicated. The experienced reader might recommend the use of statistical tests to identify and validate model structure. As Forrester and Senge [8] indicate, a long-standing discussion exists about the application of inferential statistical tests for structure examination. After a series of experiments, Forrester and Senge conclude "that conventional statistical tests of model structure are not sufficient grounds for rejecting the causal hypotheses in a system dynamics model." [8]. In the future, however, new statistical approaches might enrich the testing procedures.

Parameter Examination Test. A parameter is a quantity that characterizes a system and is held constant in a case under study, but may be varied in different cases (e. g., energy consumption per capita per day). The aim of parameter examination is to evaluate a model's parameters

against evidence or knowledge about the real system. The test can utilize both empirical and theoretical information. Furthermore, the test can be conceptual or numerical. The conceptual parameter examination test is about construct validity; it identifies elements in the real system that correspond to the parameters of the model. Conceptual correspondence means that the parameters match elements of the real system's structure. Numerical parameter examination checks to see if the quantities of the conceptually confirmed parameters are estimated accurately. Techniques for the estimation of parameters are described in [9].

Direct Extreme Condition Test. Extreme conditions do not often occur in reality; they are exceptions. The validity of a model's equations under extreme conditions is evaluated by assessing the plausibility of the results generated by the model equations against the knowledge about what would happen under a similar condition in reality. Direct extreme condition testing is a mental process and does not involve computer simulation. Ideally, it is applied to each equation separately. It consists of assigning extreme values to the input variables of each equation. The values of the output variables are then interpreted in terms of what would happen in the real system under these extreme conditions. For example, if a population is zero, then neither births, deaths, nor consumption of resources can occur.

Boundary Adequacy Structure Test. Boundary adequacy is given if the model contains the relevant structural relationships that are necessary and sufficient to satisfy a model's purpose. Consequently, the boundary adequacy test inquires whether the chosen level of aggregation is appropriate and if the model includes all relevant aspects of structure. It should ensure that the model contains the concepts that are important for addressing the problem endogenously. For instance, if parameters are likely to change over time, they should be endogenized [8]. The pertinent validation question is: "Should this parameter be endogenized or not?" That question must be decided in view of the model's purpose.

The boundary adequacy test can be applied in three ways: as a structural test, as a behavioral test, and as a policy test. The names are correspondingly: boundary adequacy structure test, boundary adequacy behavior test, and boundary adequacy policy test.

As a test of model structure, the boundary adequacy test involves developing a convincing hypothesis relating the proposed model structure to the particular issue addressed by the model. The boundary adequacy behavior/policy test (explained in Subsect. "Indirect Structure Tests") continues this line of thinking.

Dimensional Consistency Test. This test checks the dimensional consistency of measurement units of the expressions on both sides of an equation. The test is performed only at the equation level. When all tests of the individual equations are passed, a large system of dimensionally consistent equations results. This test is passed only if consistency is achieved without the use of parameters that have no meaning in respect to the real world. The dimensional consistency test is a powerful test to establish the internal validity of a model.

Indirect Structure Tests Indirect structure tests assess the validity of the model structure indirectly by examining model-generated outcome behaviors. These tests require computer simulation. The comparative activities in these tests are based on logical plausibility considerations which in turn are based on the mental models of the analyst. Comparisons of model generated data and time series about the real system are not yet involved. The tests can be applied to different degrees of model completeness, i. e., to the smallest "atomic" model components, to sub-models, as well as to the entire model.

Indirect Extreme Condition Test. For this test, the modeler assigns extreme values to selected model parameters and compares the generated model behavior to the observed or expected behavior of the real system under the same extreme conditions. This test is the logical continuation of the direct extreme condition test, i. e., many of the extreme conditions mentally developed in the previous stage can now be deployed to evaluate the simulated behavioral consequences. This test can be used for the explanatory analysis phase of modeling, but also for the normative phase of policy development. In the first instance, indirect extreme conditions are used to develop a structure that can reproduce the system behavior of interest and guard against developments impossible in reality. In the latter instance, the introduction of policies aims to improve the system's performance. The indirect extreme policy test introduces extreme policies to the model and compares the simulated consequences to what would be the most likely outcome of the real system if the same extreme policies would have been implemented.

Behavior Sensitivity Test. Sensitivity analysis assesses changes of model outcome behavior given a systematic variation of input parameters. This test reveals those parameters to which the model behavior is highly sensitive, and asks if the real system would exhibit a similar sensitivity to changes in the corresponding parameters. "The behavior sensitivity test examines whether or not plausi-

ble shifts in model parameters can cause a model to fail behavior tests previously passed. To the extent that such alternative parameter values are not found, confidence in the model is enhanced." [8]. A model can be numerically sensitive, i. e., the numerical values of variables change significantly, but the behavioral patterns are conserved. It can also exhibit behavioral sensitivity, i. e., the modes of model behavior change remarkably based on systematic parameter variations (Barlas [3] defines several distinct patterns of model behavior).

As the test for indirect extreme conditions, the behavior sensitivity test can also be deployed to assess policy sensitivity. It can reveal the degree of robustness of model behavior and hence indicate to what degree model-based policy recommendations might be influenced by uncertainty in parameter values. If the same policies would be recommended regardless of parameter changes over a plausible range, risk in using the model would be lower than if two plausible sets of parameters lead to distinct policy recommendations.

Integration Error Test. Integration error is the deviation between the analytical solution of differential equations and the numerical solution of difference equations. This test ascertains whether the model behavior is sensitive to changes in either the applied integration method or the chosen integration interval (often referred to as simulation time step). Euler's method is the simplest numerical technique for solving ordinary differential and difference equations. For models that require more precise integration processes, the more elaborated Runge–Kutta integration methods can produce more accurate results, but they require more computational resources.

Boundary Adequacy Behavior Test/Boundary Adequacy Policy Test. The logic for testing boundary adequacy has already been developed under the aspect of direct structure testing in the preceding section. The indirect structure version of this test asks whether model behavior would change significantly if the boundary were extended or reduced; i. e., the test involves conceptualizing additional structure or canceling unnecessary structure with regard to the purpose of the study. As one example of expanding the model boundary, this version of the test allows one to detail the treatment of model assumptions considered as unrealistically simple but still important for the model's purpose. On the other hand, simplifying the model is also a way to reduce the model boundary. The loop-knockout analysis is a useful method to implement this two-sided test. Knockout analysis checks behavior changes induced by the connection and disconnection of a portion of the model structure, and helps the modeler to evaluate the usefulness of those changes with respect to the model's purpose.

The other version of this test is the boundary adequacy policy test. It examines whether policy recommendations would change significantly if the boundary were extended (or restricted): That is, what would happen if the boundary assumptions were relaxed (or confined)?

Loop Dominance Test. Loop dominance analysis studies the internal mechanisms of a dynamic model and their temporal, relative contribution to the outcome behavior of the model. The relative contribution of a mechanism is a complex quantitative statement that explains the fraction of the analyzed behavior mode caused by the mechanism considered in ► System Dynamics, Analytical Methods for Structural Dominance Analysis in. The analysis reveals the relative strengths of the feedback loops in the model. The loop dominance test compares these results with the modeler's or client's assumption about which are the dominant feedback loops in the real system. Since the results are analytical statements, interpretation and comparison with the real system requires profound knowledge about the system under study.

Loop dominance analysis reveals insights about a model on a different level of analysis than the other validation tests discussed so far: It works not on the level of individual concepts or behaviors of variables but on the level of causal structure, and compares the temporal significance of the different structures to each other. The use of this test for model validation is a novelty. If the relative loop dominances of the model map the relative loop dominances of the real system, confidence in the model is enhanced. If the relative loop dominances of the real system are not known, it is still possible to evaluate whether or not the loop dominance logic in the model is reasonable.

Tests of Model Behavior

Tests of model behavior are empirical and compare simulation outcomes with data from the real system under study. On that basis, inferences about the adequacy of the model can be made. The empirical data can either be historical or refer to reasonable expectations about possible future developments.

Behavior Reproduction Tests The family of behavior reproduction tests examines how well model-generated behavior matches the observed historical behavior of the real system. As a principle, models should be tested against

data not only from periods of stability but also from unstable phases. Policies should not be designed or tested on the premise of normality, but rather should be validated with a view toward robustness and adaptiveness.

Symptom Generation Test. This test indicates whether or not a model produces the symptom of difficulty that motivated the construction of the model. To pass the symptom generation test is a prerequisite for considering policy changes, because "unless one can show how internal policies and structures cause the symptoms, one is in a poor position to alter those causes" [8].

Summary statistics, which measure and enable the interpretation of quantitative deviations, provide the means to operationalize the symptom generation test.

One known example is Theil inequality statistics, which measures the mean square-error (MSE) between the model-generated behavior and the historical time series data. It breaks down the deviation into three sources of error: Bias (U_m), unequal variation (U_s), and unequal covariation (U_c) [26].

An example taken from Schwaninger and Groesser [22] illustrates the interpretation of the error sources.

This example from an industrial firm concerns the design of a model that replicates the observed, historical product life-cycle pattern with high accuracy (Fig. 2). "Product Revenue" is the main variable of interest and specifies the symptom (growth phase followed by rapid decay). The mean square-error for revenues is 0.35. The individual components of the inequality statistics are: $U_m = 0.01, U_s = 0.01, U_c = 0.98$. The break down of the statistics shows that the major part of the error is in the U_c component, while the other two sourees of error are small. This signifies that the point-by-point values of the simulated and historical data do not match, even though the model captures the dominant trend and the average values in the historical data. Such a situation indicates that a major part of the error is probably unsystematic, and therefore the model should not be rejected for failing to match the noise component of the data. The residuals of the historic and simulated time series show no significant trend. This strengthens the assessment that the model comprises a structure that captures the fundamental dynamics of the issue under consideration.

Frequency Generation and Phase Relationship Tests. These tests focus on the frequencies of time series and phase relationships between variables. An example is the pattern of investment cycles in an industry. These tests are superior to point-by-point comparisons between model-generated and observed behavior (cf. [7]).

Frequency refers to periodicities of fluctuation in a time series. Phase relationship is the relationship between the time series of at least two variables. In principle, three phase relations are possible: Preceding, simultaneous, and successive. The frequency generation test evaluates whether or not the periodicity of a variable is in accordance with the real system. The phase relationship test assesses the phase shifts of at least two variables by comparing their trajectories.

If the phase shift between the selected simulation variables contradicts the phase shift between the same variables as observed or expected in the real system, a structural flaw in the model might be diagnosed. The test can

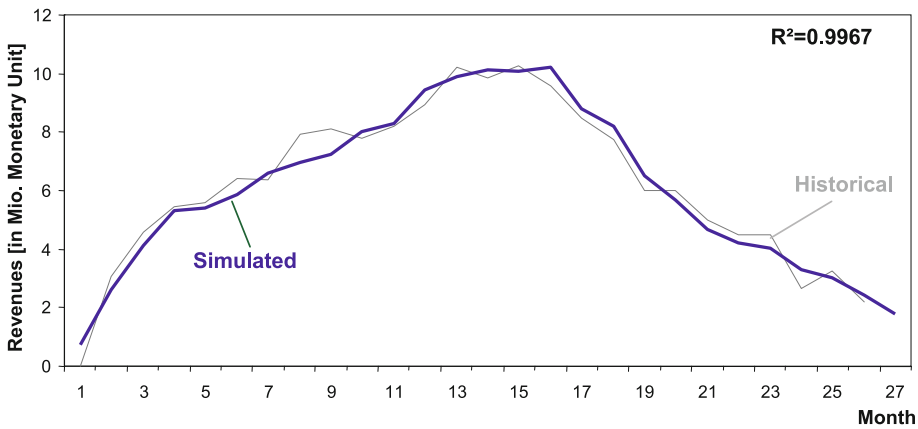

System Dynamics Modeling: Validation for Quality Assurance, Figure 2
An example comparison of historical and simulated time series for product revenues. The explained variance is close to 100% ($R^2 = 0.9967$)

uncover failures in the model, but offers only little guidance as to where the erroneous part of the model might be. The autocorrelation function test is one way to operationalize the frequency generation test [1]. The function test consists in comparing the autocorrelation functions of the observed and the model-generated behavior outputs, and can detect if significant errors between them exist.

Modified Behavior Test. Modified behavior can arise from a modified model structure or changes in parameter values. This test concerns changes in the model structure. It can be performed if data about the behavior of a structurally modified version of the real system are available. "The model passes this test if it can generate similar modified behavior, when simulated with structural modifications that reflect the structure of the "modified" real system" [2]. The applicability of this test is rather limited since it requires specific data about the modified real system which must be similar in kind to the original real system. Only under this condition can additional insights into the suitability of the original model structure be obtained. If the modified real system deviates strongly from the original real system, the test does not result in any additional insights, because no stringent conclusions about the validity of the original system can be derived from a model that is dissimilar in its structure.

Multiple Modes Test. A mode is a pattern of observed behavior. The multiple mode test considers whether a model is able to generate more than one mode of observed behavior, for instance, if a model about the production sector of an economy generates distinct patterns of fluctuations for the short-term (production, employment, inventories, and prices) and for the long term (investment, capital stock) [15]. " A model able to generate two distinct periodicities of fluctuation observed in a real system provides the possibility for studying possible interaction of the modes and how policies differentially affect each mode" [8].

Behavior Characteristic Test. Characteristics of a behavior are features of historical data that are clearly distinguishable, e. g., the peculiar shape of an oscillating time series, sharp peaks, long troughs, or such unusual events as an oil crisis. Since System Dynamics modeling is not about point prediction, the behavior characteristic test evaluates whether or not the model can generate the circumstances and behavior leading to the event. The creation of the exact time of the behavior is not part of the test.

Behavior Anticipation Tests System Dynamics models do not strive to forecast future states of system variables. Nevertheless, given that the fundamental system structure is not subject to rapid and fundamental change, dynamic models might provide insights about the possible range of future behaviors. Hence, behavior anticipation tests are similar to behavior reproduction tests but possess a higher level of uncertainty.

Pattern Anticipation Test. This test examines whether a model generates patterns of future behavior which are assumed to be qualitatively correct. The limits of anticipation reside in the fact that that the structure of the system may change over time. The pattern anticipation test entails evaluation of periods, phase relationships, shape, or other characteristics of behavior anticipated by the model. One possibility for implementing this test is to split the historical time series into two data sets and introduce an artificial present time at the end of the first data series. The first set is then used for model development and calibration. The second data series is employed to perform the behavior anticipation test, i. e., to evaluate whether the model is able to anticipate possible future behavior.

This test can also be used for policy considerations, in which case it is called "Changed Behavior Anticipation Test". It determines whether the model correctly anticipates how the behavior of the real system will change if a governing policy is altered.

Event Anticipation Test. In respect to System Dynamics, the anticipation of events does not imply knowing the exact time at which the events occur; it rather means understanding the dynamic nature of events and being able to identify the antecedents leading to them. For instance, the event anticipation test is passed if a model has the ability to anticipate a steep peak in food prices based on the development of the conditioning factors.

Behavior Anomaly Test In constructing and analyzing a System Dynamics model, one strives to make it behave like the real system under study. However, the analyst may detect anomalous features of the model's behavior which conflict with the behavior of the real system. Once the behavioral anomaly is traced to components of the model structure responsible for the anomaly, one often finds flaws in model assumptions. The test for recognizing behavioral anomalies is sporadically applied throughout the modeling process.

Family Member Test A System Dynamics model often represents a family of social systems. Whenever possible, a model should be a general representation of the class of that system to which the particular case belongs. One

should ask if the model can generate the behavior in other instances of the same class. "The family-member test permits a repeat of the other tests of the model in the context of different special cases that fall within the general theory covered by the model. The general theory is embodied in the structure of the model. The special cases are embodied in the parameters. To perform this test, one uses the particular member of the general family for picking parameter values. Then one examines the newly parametrized model in terms of the various model tests to see if the model has withstood transplantation to the special case" [8]. The model should be calibrated so as to be applicable to the widest range of related systems. For the family member test, only the parameter values of the model are subject to alterations; changes in the model structure are part of the modified behavior test, as discussed in the preceding section.

Surprise Behavior Test A surprising model behavior is a behavior that is not expected by the analysts. When such an unexpected behavior appears, the model analysts must first understand the causes of the unexpected behavior within the model. They then compare the behavior and its causes with those of the real system. In many cases, the surprising behavior turns out to be due to a formulation flaw in the model. However, if this procedure leads to the identification of behavior previously unrecognized in the real system, the confidence in the model's usefulness is strongly enhanced. Such a situation may signify a model-based identification of a counter-intuitive behavior in a social system.

Turing Test The Turing test is a qualitative test which uses the intuitive knowledge of system experts to evaluate model behavior. Experts are presented with a shuffled collection of real and simulated output behavior patterns. They are asked if they can distinguish between these two types of patterns. If they are unable to discern which pattern belongs to the real system and which to the simulation output, the Turing test is passed. Similar to the phase relationship test, the Turing test is powerful in its ability to indicate structural flaws, but offers only little guidance for locating them in the model.

Validation Process

The validation process pervades all phases of model-building and reaches even beyond, into the phases of model implementation and use. The diagram in Fig. 3 visualizes the function of validation in the process of model-building.

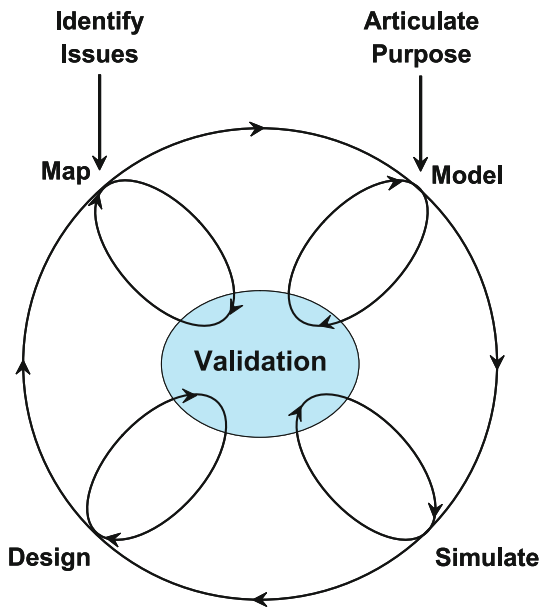

System Dynamics Modeling: Validation for Quality Assurance, Figure 3
Validation in the context of the System Dynamics modeling procedure

For the purposes of this contribution, validation is placed at the center of the scheme. From there it is dispersed through all steps of the modeling process, Map (high-level model creation), Model (build the formal model), Simulate (explore scenarios, etc.) and Design (articulation of policies). We have limited the differentiation of these steps in order to highlight the structure of the process – a recursive structure drawn as a nested loop line. After the initial identification of issues and the articulation of model purpose, the simplified diagram denotes the four phases, of mapping to modeling to simulation and design. The small loops symbolize micro-processes in which, for example, a model is submitted to validation, e. g., a direct structure test, which may lead to its modification (two small arrows). The larger loops illustrate more comprehensive processes. For example, an indirect structure test of the model is carried out, in which the behavior is tested by means of simulation. Or a policy test by simulation leads to implications for design (large loop), and the design is validated in detail thereafter (small loop).

Now, we should note that the process scheme reminds us of a further aspect which is quite fundamental. If the results of the model's operation, e. g., a "prediction", diverge from the results of a test, then either the model is wrong or the test is inadequate (see p. 168 in [24]). This meta-perspective lets us keep an eye on the adequacy of the tests:

is the logic of the test flawless? Are the data sources in order? (see adequacy of methodology test in Sect. "Validation Methods").

Model-building is a process of knowledge-creation, and model validation is an integral part of it. As the model is validated using the methods described in the former chapter, insights emerge, and a better understanding of the system under study keeps growing. But model-building is also a construction of a reality in the minds of observers [31,32] concerned with an issue. In this procedure, validation is supposed to be a "guarantor" for the realism of the model, a control function for preventing gross aberrations in individual and collective perceptions. Validation should encompass precautions against cognitive limitations and modeler blindness. The set of tests presented above is a system of heuristic devices for enhancing such provisions. A question not yet answered is how these tests should be ordered along the timeline. We have fleshed out three structural principles, which are illustrated in Fig. 4:

1. *Validation is a parallel process:* Validation in all three domains – context, structure and behavior – is carried out in a synchronized fashion, as shown in Fig. 4. Context validation is continuous, while the other two components show alternations.
2. *Parts of the validation process have a sequential structure:* This refers to the alternations between the components of structure and behavior validation. In principle, they occur alternately, with structural validation taking the lead and behavior validation following. After that, one might revert to structural validation again and so forth.
3. *Validation processes are polyrhythmic:* The length and accentuation of validation activities vary among the three levels. This fact is symbolized by the frequency of the vertical lines in the blocks of the chronogram.

A further important factor affecting the validation process is the degree of resolution: micro, meso or macro (as visualized in Fig. 1). The focus of validation is primarily on micro-objects, the smallest building blocks of a model, for example, a stock or a subsystem containing a stock with its flows. One could call them metaphorically *atoms* or *molecules*. Each building-block should be validated individually, before it is integrated into the overall model structure. The reason is that at this atomic level disfunctionalities or errors of thinking are discovered immediately, while at higher levels of resolution the identification of structural flaws is more difficult and cumbersome. The same holds for the relation between modules (meso) and the whole model (macro). Before adding a module, it should be validated in itself. This way, errors at the level of the whole system can be minimized and, it is very important to add, counterintuitive behavior of the model can be understood with more ease.

Until now we have examined what occurs in a validation process and how the process is structured. Finally, we raise the issue of who the actors are and why. In this context, we will concentrate on group processes in model validation.

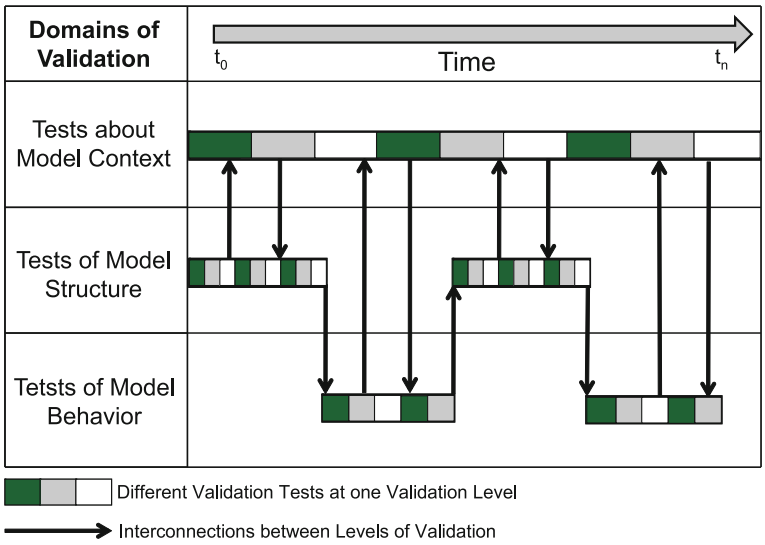

System Dynamics Modeling: Validation for Quality Assurance, Figure 4
The interplay of validation activities

Different observers associate diverse contents with a system, and they might even conceive the system distinctly, as far as its boundaries, goals and structures are concerned. They might also succumb to erroneous inferences and therefore adhere to defective propositions. Consequently, error-correcting devices are needed. A powerful mechanism for this purpose is the practice of model-building and validation in groups. We have already referred to that concept in respect to several of the methods discussed in Sect. "Validation Methods", and now we will briefly expand on it.

Group Model-building (GMB) is a methodology to facilitate team learning with the help of System Dynamics [30]. The methodology consists of a set of methods and instruments as well as heuristic principles. These are meant to facilitate the elicitation of knowledge, the negotiation of meanings, the creation of a shared understanding of a problem in a team, as well as the joint construction and validation of models. The process of GMB is essentially a dialog in which different interpretations of the real system under study are exposed, transformed, aligned and translated into the concepts and relationships which make up the model system. This is mainly a matter of structural validation, of qualitative mapping and the elaboration of the formal model.

Given its transdisciplinary approach, GMB enables an integration of different perspectives into one shared image of the system-in-focus. GMB is an important provision for attaining higher model quality: it can broaden the available knowledge base, inhibit errors and show itself to be a cohesive force in the quest for consensual model validation. The opportunity for validation inheres in the broad knowledge base normally available in a modeling group. Much of this knowledge can be leveraged for validation purposes. Most validation tests are carried out in coordination with model-building activities. Often the tests become a task to be accomplished between workshops. However, the members of the model-building group can, in principle, be made available for knowledge input into and monitoring of validation activities.

A functioning GMB process requires a number of necessary elements [18]: commitment of key players (e. g., attendance of workshops), impartial facilitation, on-the-spot modeling at conversational pace, with continuous display of the developing model as well as an interactive and iterative group process.

Let us not forget that there are many situations in which one single person is in charge of building and validating a model. In these cases the modeler must constantly challenge his or her own position. Normally, it is preferred that one should also call for external judgment in reviews, walkthroughs and the like. The same holds for knowledge supply. One-person modelers can find a lot of material in the media, libraries, the internet, etc., but it is also usually beneficial to find experienced persons from whom to elicit relevant knowledge, or even persons who join the modeling and validation venture.

Synopsis and Outlook

Models should be relevant for coping with the complexity of the real world. At the same time, the methods by which they are constructed must be rigorous; otherwise the quality of the model suffers. Rigor and relevance are not entirely dichotomous, but given resource constraints they are in competition to a certain extent. Lack of rigor in building a model is often worse than limitations to the model's relevance. One may say, *cum grano salis*: incomplete validation entails complete irrelevance. Modelers must find a way to ensure both rigor and relevance, as both are necessary conditions for achieving the model purpose. Neither alone is sufficient, but one may assume that, taken together, rigor and relevance are sufficient conditions. The relative importance of these two dimensions of model building may vary over time as a function of the model quality achieved. At the beginning, relevance might be more important, while at high levels of model accomplishment rigor might become prevalent.

Investing in high model quality is indeed both worthwhile and imperative. It is impressive to register the fact that model validation has achieved higher levels of rigor not only in the academic field but also in the world of affairs: According to Coyle and Exelby, the need for orientating decisions about "real-world" affairs has also fueled strong efforts among commercial modelers and consultants for ensuring model validity [5].

We have discussed two essential aspects of model validation, the epistemological foundations and methodological procedures for ensuring model validity. The main conclusion we have reached on epistemology is that crude positivism has been superseded by newer philosophical orientations that provide guidance for an adequate concept of validation in System Dynamics. Validation has been defined as a rich and well-defined process by which the confidence in a model is gradually enhanced. Validity, then, is always a matter of degree, never an absolute property.

Well-defined here is not meant in the sense of a rigid algorithm, but as the rigorous application of a battery of validation methods which we have described in some detail. We have included a number of new validation tests by which modelers' understanding of the relevant con-

text can be scrutinized. These additional tests are rightly supposed to prevent wrong methodological choices. They should also trigger innovative approaches to the issues under study and foster the ability to think in terms of contingencies. Finally, they should liberate modelers from tunnel vision and open avenues to creativity. The imperative here is to cultivate a "sense of the possible" (Robert Musil's *Möglichkeitssinn*) and a skepticism against the supposedly impossible (see also [29]).

Simulation based on formal dynamic models is likely to become ever more important for both private and public organizations. It will continue to support managers at all levels in decision-making and policy design. The more that models are relied upon, the greater the importance of their high quality. Therefore, model validation is one of the big issues lying ahead in System Dynamics modeling.

Appendix: Overview of the Tests Described in This Chapter

1. **Tests of the Model-Related Context**
   - 1.1 Issue Identification Test
   - 1.2 Adequacy of Methodology Test
   - 1.3 System Configuration Test
   - 1.4 System Improvement Test
2. **Tests of Model Structure**
   - 2.1 Direct Structure Tests
     - 2.1.1 Structure Examination Test
     - 2.1.2 Parameter Examination Test
     - 2.1.3 Direct Extreme Condition Test
     - 2.1.4 Boundary Adequacy Structure Test
     - 2.1.5 Dimensional Consistency Test
   - 2.2 Indirect Structure Tests
     - 2.2.1 Indirect Extreme Condition Test
     - 2.2.2 Behavior Sensitivity Test
     - 2.2.3 Integration Error Test
     - 2.2.4 Boundary Adequacy Behavior Test/Boundary Adequacy Policy Test
     - 2.2.5 Loop Dominance Test
3. **Tests of Model Behavior**
   - 3.1 Behavior Reproduction Tests
     - 3.1.1 Symptom Generation Test
     - 3.1.2 Frequency Generation and Phase Relationship Test
     - 3.1.3 Modified Behavior Test
     - 3.1.4 Multiple Modes Test
     - 3.1.5 Behavior Characteristic Test
   - 3.2 Behavior Anticipation Tests
     - 3.2.1 Pattern Anticipation Test
     - 3.2.2 Event Anticipation Test
   - 3.3 Behavior Anomaly Test
   - 3.4 Family Member Test
   - 3.5 Surprise Behavior Test
   - 3.6 Turing Test

Bibliography

Primary Literature

1. Barlas Y (1990) An autocorrelation function test for output validation. Simulation 55(1):7–16
2. Barlas Y (1996) Formal aspects of model validity and validation in system dynamics. Syst Dyn Rev 12(3):183–210
3. Barlas Y (2006) Model validity and testing in System Dynamics: Two specific tools. Paper presented at the 24th International Conference of the System Dynamics Society, Nijmegen
4. Barlas Y, Carpenter S (1990) Philosophical roots of model validity – two paradigms. Syst Dyn Rev 6(2):148–166
5. Coyle G, Exelby D (2000) The validation of commercial system dynamics models. Syst Dyn Rev 16(1):27–41
6. Feyerabend P (1993) Against method, 3rd edn. Verso, London
7. Forrester JW (1961) Industrial dynamics. MIT Press, Cambridge
8. Forrester JW, Senge PM (1980) Test for building confidence in System Dynamics models. In: Legasto AA Jr, Forrester JW, Lyneis JM (eds) System Dynamics. North-Holland Publishing Company, Amsterdam, pp 209–228
9. Graham AK (1980) Parameter estimation in system dynamics modeling. TIMS Studies in the Management Sciences 14:125–142
10. Heracleous L (2006) Discourse, interpretation, organization. Cambridge University Press, Cambridge
11. James W (1987) Writings 1902–1910. Library of America, New York
12. Kuhn T (1996) The structure of scientific revolutions, 3rd edn. University of Chicago Press, Chicago
13. Lacey AR (1996) A dictionary of philosophy, 3rd revised edn. Barnes and Noble, New York
14. Lane DC (1995) The folding star: A comparative reframing and extension of validity concepts in System Dynamics. In: Simada T, Saeed K (eds) Proceedings of 1995 international System Dynamics conference, 30 July–4 Aug, vol I. System Dynamics Society, Lincoln, pp 111–130
15. Mass NJ (1975) Economic cycles: An analysis of underlying causes. Productivity Press, Cambridge
16. Mattheij RMM, Rienstra SW, Boonkkamp JH MtT (2005) Partial differential equations: Modeling, analysis, computation. Society for Industrial and Applied Mathematics (SIAM), Eindhoven
17. Petersen DW, Eberlein RL (1994) Understanding models with vensim. In: Morecroft JDW, Sterman JD (eds) Modeling for learning organiziations. Productivity Press, Portland, pp 339–358
18. Phillips LD (2007) Decision conferencing. In: Edwards W, Miles RF, von Winterfeldt D (eds) Advances in decision analysis. From foundations to applications. Cambridge University Press, Cambridge, pp 375–399
19. Popper KR (1959) The logic of scientific discovery. Basic Books, New York (latest edition: 2002, Routledge, London)
20. Popper KR (1972) Objective knowledge: An evolutionary approach. Clarendon Press, Oxford
21. Rapoport A (1954) Operational philosophy. Integrating knowledge and action. Harper, New York

22. Schwaninger M, Groesser SN (2008) Model-based theory-building with system dynamics. Syst Res Behav Sci 25:1–19
23. Seiffert H, Radnitzky G (1994) Handlexikon der Wissenschaftstheorie, 2nd edn. DTV Wissenschaft, Munich
24. Smith VL (2008) Rationality in economics: Constructivist and ecological forms. Cambridge University Press, Cambridge
25. Snabe B, Grössler A (2006) System dynamics modelling for strategy implementation – case study and issues. Syst Res Behav Sci 23(4):467–481
26. Sterman JD (1984) Appropriate summary statistics for evaluating the historical fit of system dynamics models. Dynamica 10(2):51–66
27. Sterman JD (1989) Misperceptions of Feedback in Dynamic Decision Making. Organ Behav Human Decis Process 43(3):301–335
28. Sterman JD (2000) Business dynamics. Systems thinking and modeling for a complex world. Irwin/McGraw-Hill, Boston
29. Taleb NN (2007) The black swan. The impact of the highly improbable. Random House, New York
30. Vennix JAM (1996) Group model building: Facilitating team learning using System Dynamics. Wiley, Chichester
31. von Foerster H (1984) Observing systems, 2nd edn. Intersystems Publications, Seaside
32. von Glasersfeld E (1991) Abschied von der Objektivität. In: Watzlawick P, Krieg P (eds) Das Auge des Betrachters. Piper, Munich, pp 17–30

Books and Reviews

Finlay PN (1997) Validity of decision support systems: Towards a validation methodology. Syst Res Behav Sci 14(3):169–182
Forrester JW (1961) Industrial dynamics. MIT Press, Cambridge
Law AM (2007) Simulation modeling and analysis, 4th edn. McGraw-Hill, New York
Legasto AA Jr, Forrester JW, Lyneis JM (eds) (1980) System Dynamics. North-Holland, Amsterdam
Morecroft J (2007) Strategic modelling and business dynamics: A feedback systems approach. Wiley, Chichester
Sargent RG (2004) Validation and verification of simulation models. In: Ingalls RG, Rossetti MD, Smith JS, Peters BA (eds) Proceedings of the 2004 winter simulation conference. ACM-Association for Computing Machinery, Washington DC, pp 17–28
Sterman JD (2001) Business dynamics. Systems thinking and modeling for a complex world. Irwin/McGraw-Hill, Boston
Warren K (2008) Strategic management dynamics. Wiley, Chichester

System Dynamics Models of Environment, Energy and Climate Change

Andrew Ford
School of Earth and Environmental Sciences,
Washington State University, Pullman, Washington, USA

Article Outline

Glossary
Definition of the Subject
Introduction
The Model of Mono Lake
The Model of the Salmon in the Tucannon River
Models of Climate Change
System Dynamics Models of the Carbon Cycle
Lessons from the Regulated Power Industry in the 1970s
Simulating the Power Industry Response to a Carbon Market
Conditions for Effective Interdisciplinary Modeling
Future Directions
Bibliography

Glossary

CO_2 Carbon dioxide is the predominant greenhouse gas. Anthropogenic CO_2 emissions are created largely by the combustion of fossil fuels.

CGCM Coupled general circulation model, a climate model which combines the atmospheric and oceanic systems.

GCM General circulation model, a term commonly used to describe climate models maintained at large research centers.

GHG GHG is a greenhouse gas such as CO_2 and methane. These gases contribute to global warming by capturing some of the outgoing infrared radiation before it leaves the atmosphere.

GT Gigaton, a common measure of carbon storage in the global carbon cycle. A GT is a billion metric tons.

IPCC The Intergovernmental Panel on Climate Change was formed in 1988 by the World Meteorological Organization and the United Nations Environmental Program. It reports research on climate change. Their assessments are closely watched because of the requirement for unanimous approval by all participating delegates.

Definition of the Subject

System dynamics is a methodology for studying and managing complex systems which change over time. The method uses computer modeling to focus our attention on the information feedback loops that give rise to the dynamic behavior. Computer simulation is particularly useful when it helps us understand the impact of time delays and nonlinearities in the system. A variety of modeling methods can aid the manager of complex systems. Coyle (p. 2 in [3]) puts the system dynamics approach in perspective when he describes it as that "branch of control theory which deals with socio-economic systems, and that

branch of management science which deals with problems of controllability." The emphasis on controllability can be traced to the early work of Jay Forrester [9] and his background in control engineering [10]. Coyle highlighted controllability again in the following, highly pragmatic definition:

> *System dynamics is a method of analyzing problems in which time is an important factor, and which involve the study of how a system can be defended against, or made to benefit from, the shocks which fall upon it from the out-side world.*

The emphasis on controllability is important as it directs our attention to understanding and managing the system, not to the goal of forecasting the future state of the system. Making point predictions is the objective of some modeling methods, but system dynamics models are used to improve our understanding of the general patterns of dynamic behavior. System dynamics has been widely used in business, public policy and energy and environmental policy making. This article describes applications to energy and environmental systems.

Introduction

System dynamics has been used extensively in the study of environmental and energy systems. This article describes some of these applications, paying particular attention to the problem of global climate change. The applications were selected to illustrate the power of the method in promoting an interdisciplinary understanding of complex problems.

The applications to environmental and energy systems are similar to applications to other systems described in this encyclopedia. They usually begin with the recognition of a dynamic pattern that represents a problem. System dynamics is based on the premise that we can improve our understanding of the dynamic behavior by the construction and testing of computer simulation models. The models are especially helpful when they illuminate the key feedbacks that give rise to the problematic behavior.

System dynamics is explained in the core article in this volume, in the early texts by Forrester [9], Coyle [3] and Richardson [18] and in more recent texts on strategy by Warren [22] and by Morecroft [17]. The most comprehensive explanation is provided in the text on business dynamics by Sterman [19]. Applications to environmental systems are explained in the text by Ford [7]. The most widely read application to the environment is undoubtedly *The Limits to Growth* [16]. Collections of environmental applications appear in special issues of the *System Dynamics Review* [11,20].

The models are normally implemented with visual software such as Stella (http://www.iseesystems.com), Vensim (http://www.vensim.com/) or Powersim (http://www.powersim.com/). These programs use stock and flow icons to help one see where the accumulations of the system take place. They also help one to see the information feedback in the simulated system. The programs use numerical methods to show the dynamic behavior of the simulated system. The examples selected for this article make use of the Stella and Vensim software.

This article begins with textbook examples of environmental resources in the western US. The management of water levels at Mono Lake in Northern California is the first example. It shows a hydrological model to simulate the decline in lake levels due to water exported out of the basin. The second example involves the declining salmon population in the Tucannon River in Eastern Washington. These examples demonstrate the clarity of the approach, and they illustrate the potential for interdisciplinary modeling.

The article then turns to the topic of climate change and global warming. The focus is on the global carbon cycle and the growing concentration of carbon dioxide (CO_2) in the atmosphere. A wide variety of models have been used to improve our understanding of the climate system and the importance of anthropogenic CO_2 emissions. Examples of system dynamics models are presented to show how they can improve our understanding and provide a platform for interdisciplinary analysis.

System dynamics has also been widely applied in the study of energy problems, especially problems in the electric power industry. The final section describes two applications to electric power. The first involved the financial problems of regulated electric utilities in the US during the 1970s. It demonstrates the usefulness of the method in promoting an interdisciplinary understanding of the utilities' financial problems. The second study dealt with the CO_2 emissions in the large electricity system in the Western USA and Canada. It demonstrated how the power industry could lead the way in reducing CO_2 emissions in the decades following the implementation of a market in carbon allowances.

The Model of Mono Lake

Mono Lake is an ancient inland sea on the east side of the Sierra Nevada Mountains in California. Microscopic algae thrive in its saline waters, and the algae support huge populations of brine flies and brine shrimp which can,

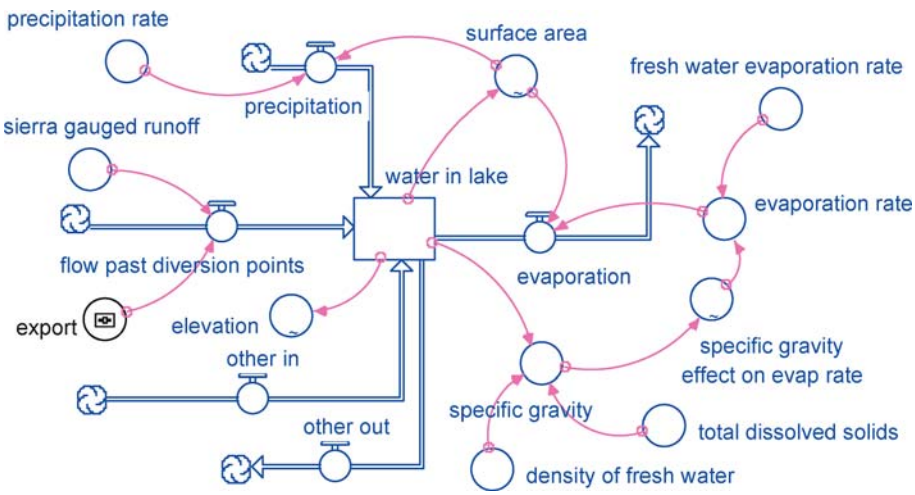

System Dynamics Models of Environment, Energy and Climate Change, Figure 1
Stella diagram of the model of Mono Lake

under the right conditions, provide a virtually limitless food supply for migratory and nesting birds. Starting in 1941, stream flows toward Mono Lake were diverted into the aqueduct for export to Los Angeles. The large export deprived the lake of the historical flows, and the volume shrunk over the next four decades. By 1980, the lake's volume was cut approximately in half, and its salinity nearly doubled. Higher salinity levels posed risks to the ecosystem, and environmental scientists feared for the future of the lake ecosystem. Various groups filed suit in the 1970s to limit exports, and the California Supreme Court ruled in 1983 that public trust doctrine mandated a reconsideration of the management of the waters of the Mono Basin. That reconsideration led to a long-term plan to limit exports until the lake's elevation would return to safer levels.

Figure 1 shows a system dynamics model to simulate water flows and storage in the Mono Basin. The goal was to understand the pattern of decline over four decades and to study the responsiveness of the lake to a change in export policy. The model is implemented with the Stella software, and Fig. 1 shows how the model appears when using the software. A single stock variable is used to represent the storage in the basin. The main flow into the lake is the flow from gauged streams that bring runoff from the Sierra to the lake. The aqueduct system diverts a portion of this flow south to Los Angeles, and the flow allowed past the diversion points is the main flow into the lake. The main outflow is the evaporation. It depends on the surface area of the lake and the evaporation rate. The surface area depends in a nonlinear way on the volume of water in the lake. Figure 1 shows that this model follows the standard, system dynamics practice of using familiar names to convey the meaning of the variables in the model. (These particular names match the terms used by water managers and hydrological models of the basin.)

Figure 2 shows the simulated decline in the lake if exports were allowed to continue at high levels for 50 years. The lake would decline from 6374 to around 6342 feet above sea level, a value which is designated as a hypothetical danger level for this simulation. The long, gradual decline is a match of projections by the other hydrological models used in the management plan for the basin. The lake will continue to fall until the area has been reduced sufficiently to create an evaporation which will lead to a balance of the flows in and out of the basin.

Figure 3 shows the simulated responsiveness of the lake to a change in export. The export is cut to zero midway through the simulation, and the elevation increases rapidly in the ensuing decade. The simulation reveals an immediate and rapid response, indicating that there is little downward momentum associated with the hydrology of the basin. This responsiveness is highly relevant to the management plan. When the lake falls to a dangerous level, the export could be reduced, and the lake would climb to higher elevations within a few years after the change in policy. This rapid response supports the "wait and see" argument by those who advocated waiting for full signs of a dangerous salinity before changing export policy.[1] But there is far more than hydrology at work in this system. The waters of Mono Lake support a com-

[1] "Wait and see" may be supported by an analysis of the hydrology of the basin, but it does not necessarily make sense when considering the long delays in the political and managerial process to change water export.

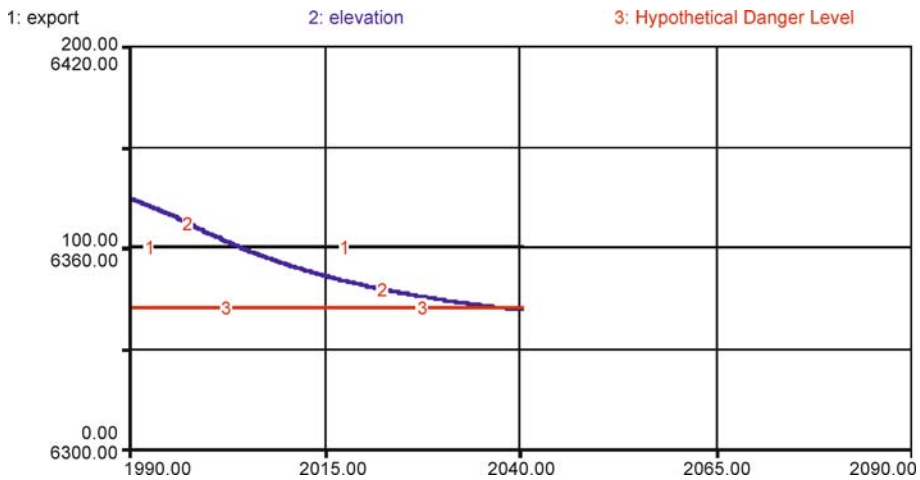

System Dynamics Models of Environment, Energy and Climate Change, Figure 2
Simulated decline in Mono Lake elevation if historical export were allowed to continue until the year 2040

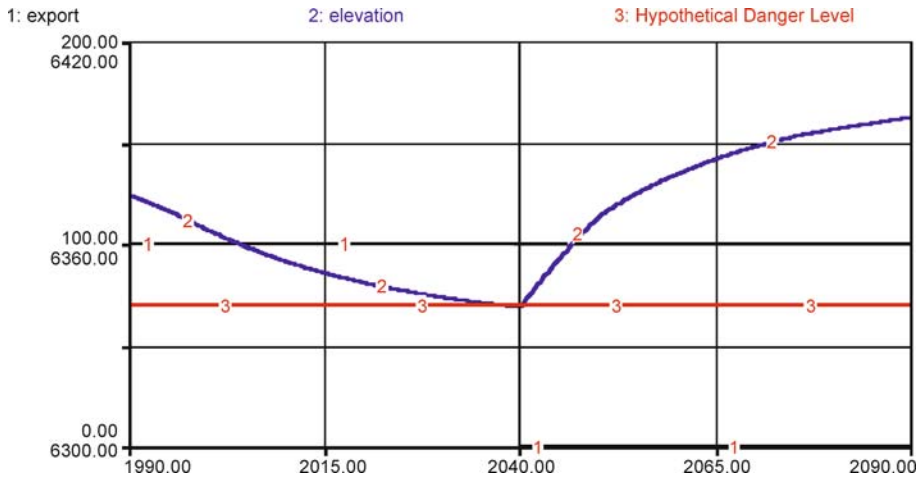

System Dynamics Models of Environment, Energy and Climate Change, Figure 3
Simulated recovery of Mono Lake elevation if export is set to zero for the second half of the simulation

plex ecosystem which may or may not recover as quickly as the lake elevation. To explore the larger system requires an interdisciplinary model, one that looks at both hydrology and population biology.

Figure 4 shows a model of the population of brine shrimp that live in Mono Lake. The life cycle begins when the adult females deposit cysts in the summer. A stock is assigned to the over wintering cysts. The nauplii and juvenile phases are combined into a second stock, and the maturation leads to a new population of adults in the following summer. The model operates in months and is simulated over a long time interval to show the population response to long-term changes in elevation and in salinity. The model shows the population's response to changes in lake elevation, so one can learn about the delays in the population's response to the changes in lake elevation. Since the shrimp life cycle is 12 months, one would expect the population to rebound rapidly after the increase in elevation and the reduction in salinity. The model confirms that the shrimp population would increase rapidly in the years following the elimination of water export from the basin.

The Mono Lake models are textbook models [7]. They demonstrate the clarity that the system dynamics approach brings to the modeling of environmental systems. The stock and flow icons help one see the structure of the system, and the long variable names help one appreciate the individual relationships. The simulation results help one understand the downward momentum in the system. In this particular case, there is no significant down-

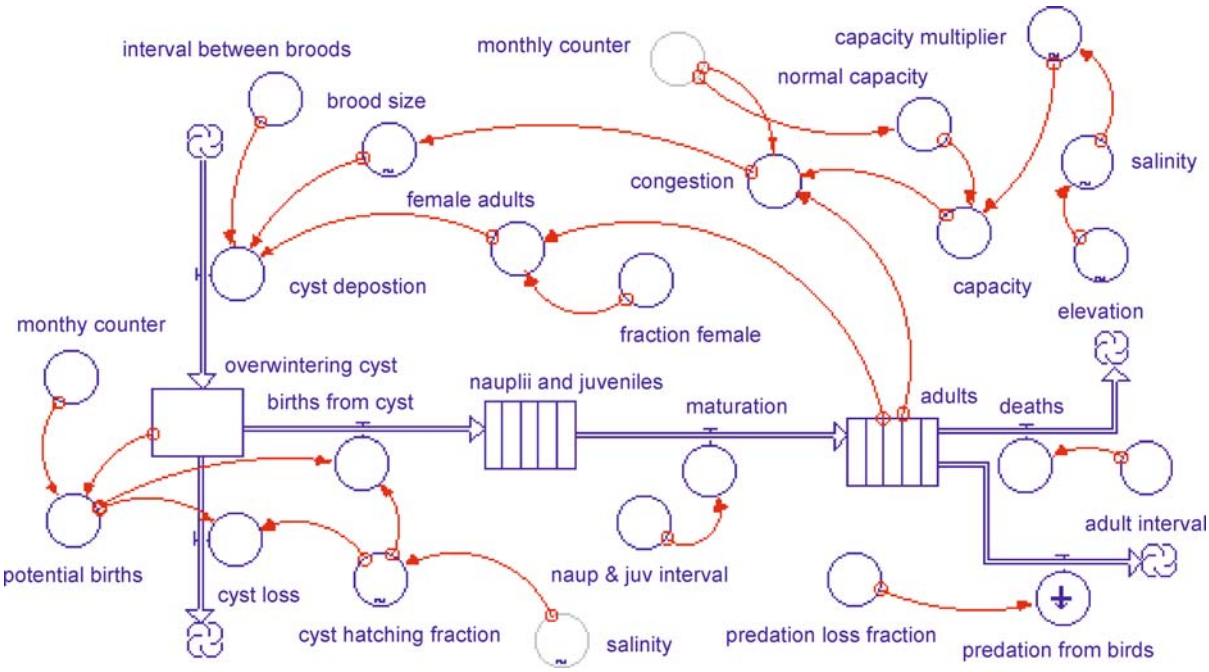

System Dynamics Models of Environment, Energy and Climate Change, Figure 4
Stella model of the brine shrimp population of Mono Lake

ward momentum associated with either the hydrological dynamics or the population dynamics.

The model in Fig. 1 allows for a system dynamics portrayal of the type of calculations commonly performed by hydrologists. Compared to the previous methods in hydrology, system dynamics adds clarity and ease of experimentation. The population model in Fig. 4 is a system dynamics version of the type of modeling commonly performed by population biologists. System dynamics adds clarity and ease of experimentation in this discipline as well.

The main theme of this article is that system dynamics offers the opportunity for interdisciplinary modeling and exploration. The Mono Lake case illustrates this opportunity with the combination of the hydrological and biological models that allows one to simulate management policies that control export based on the size of the brine shrimp population. The new model is no longer strictly hydrology nor strictly population biology; it is an interdisciplinary combination of both. And by using stock and flow symbols that are easily recognized by experts from many fields of study, the system dynamics enables quick transfer of knowledge. The ability to combine perspectives from different disciplines is one of the most useful aspects of the system dynamics approach to environmental and energy systems. This point is illustrated further with each of the remaining examples in the article.

The Model of the Salmon in the Tucannon River

The next example involves the decline in salmon populations in the Snake and Columbia River system of the Pacific Northwest. By the end of the 1990s, the salmon had disappeared from 40% of their historical breeding ranges despite a public and private investment of more than $1 billion. The annual salmon and steelhead runs had dwindled to less than a quarter of the runs from one hundred years ago. Figure 5 shows a system dynamics model one of the salmon runs, the population of Spring Chinook that spawn in the Tucannon River. The river rises in the Blue Mountains of Oregon and flows 50 miles toward the Snake River in Eastern Washington. It is estimated that the river originally supported runs of 20 thousand adults. But the number of returning adults has declined substantially due to many changes in the past sixty years. These changes include agricultural development in the Tucannon watershed, hydro-electric development on the Snake and Columbia, and harvesting in the ocean.

Each of the stocks in Fig. 5 correspond to a different phase in the salmon life cycle (see Table 1), with a total lifecycle of 48 months. The parameters represent predevelopment conditions, the conditions prior to agricultural development in the Tucannon watershed and hydro-electric development on the Snake and Columbia. Each of these parameters is fixed regardless of the size of the salmon

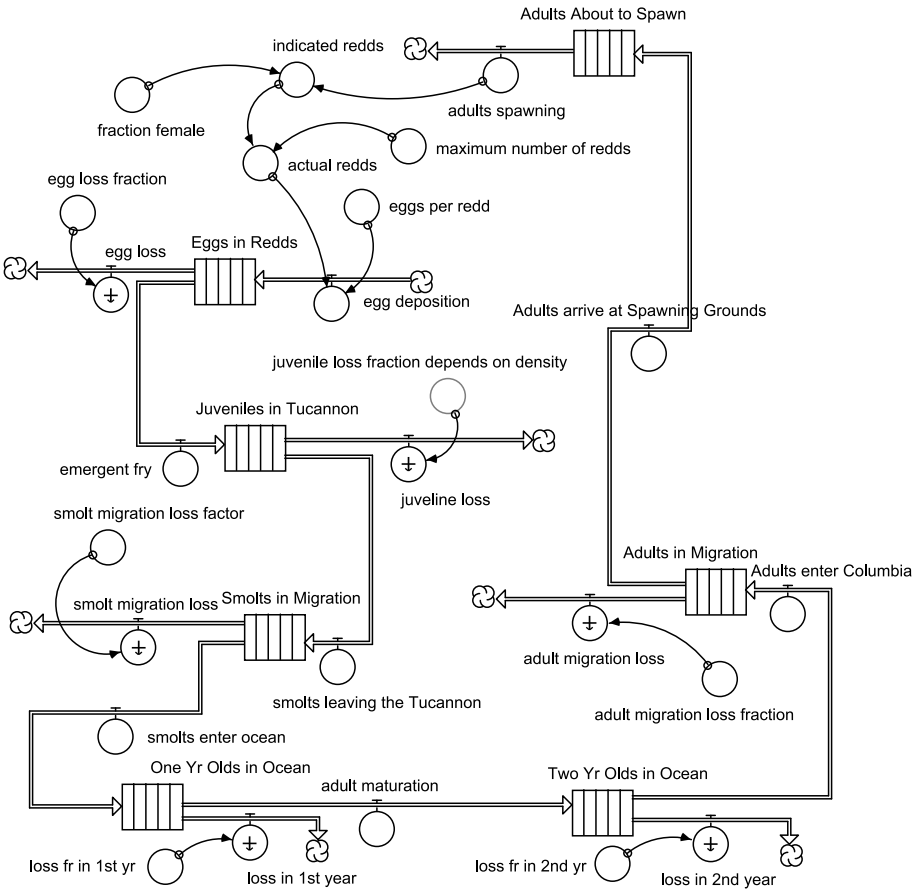

System Dynamics Models of Environment, Energy and Climate Change, Figure 5
Stella diagram of the model of the salmon life cycle

populations. One of the most important variables is the "juvenile loss fraction depends on density." It can be as low as 50% when there are only a few emergent fry each spring. With higher densities, however, juvenile survival becomes more difficult due to crowding in the cool and safe portions of the river.

Figure 6 shows the model results over a 480 month period with the population parameters in Table 1. The simulation begins with a small number to see if the population will grow to the 20 thousand adults that were thought to have returned to the river in earlier times. The time graph shows a rapid rise to around 20 thousand adults within the first 120 months of the simulation. The remainder of the simulation tests the population response to variability in environmental conditions, as represented by random variations in the smolt migration loss fraction. (This loss tends to be high in years with low runoff and low in years with high runoff.) Figure 6 confirms that the model simulates the major swings in returning adults due to environmental variability. The runs can vary from a low of ten thousand to a high of thirty thousand.

System dynamics models are especially useful when they help us to understand the key feedbacks in the system. Positive feedback loops are essential to our under-

System Dynamics Models of Environment, Energy and Climate Change, Table 1
Inputs to simulate the salmon population under pre-development conditions

Months in each phase		Population parameters	
Adults ready to spawn	1	fraction female	50%
eggs in redds	6	eggs per redd	3,900
juveniles in Tucannon	12	egg loss fraction	50%
smolts in migration	1	smolt migration loss factor	90%
one yr olds in ocean	12	loss fr for first yr	35%
two yr olds in ocean	12	loss fr for second yr	10%
adults in migration	4	adult migration loss fraction	25%

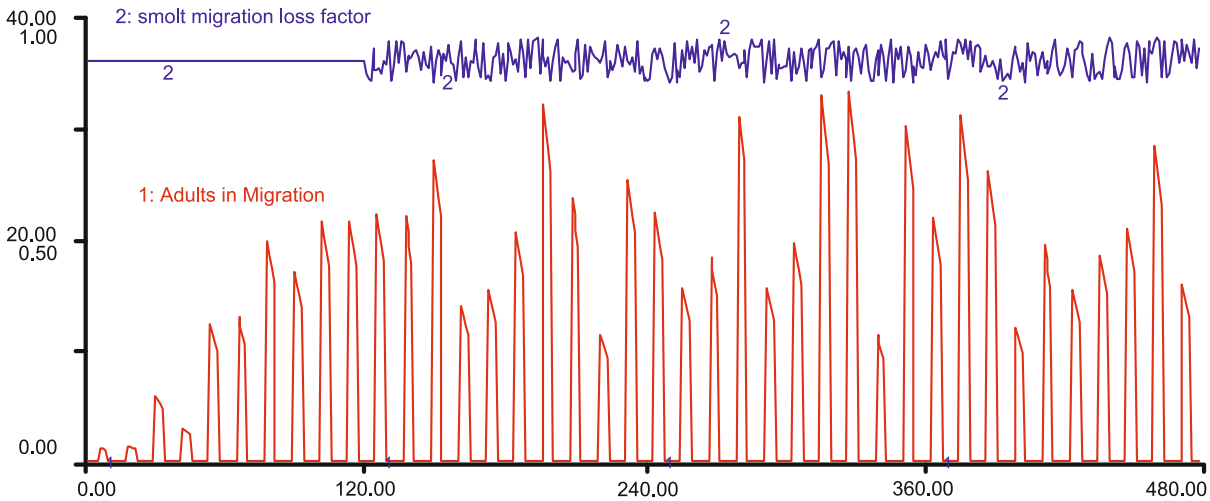

System Dynamics Models of Environment, Energy and Climate Change, Figure 6
Test of the salmon model with random variations in the smolt migration losses

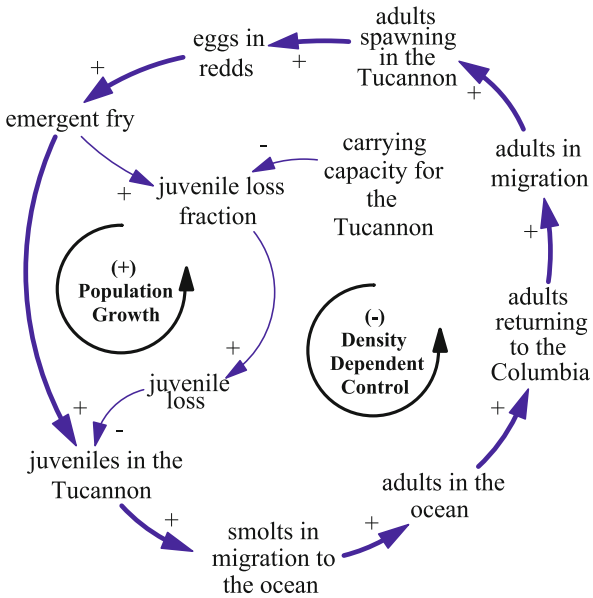

System Dynamics Models of Environment, Energy and Climate Change, Figure 7
Key feedback loops in the salmon model

standing of rapid, exponential growth; negative feedbacks are essential to our understanding of the controllability of the system. Causal loop diagrams are often used to depict the feedback loops at work in the simulated system. Figure 7 shows an example by emphasizing the most important feedback loops in the salmon model.

Most readers will immediately recognize the importance of the outer loop which is highlighted by bold arrows in the diagram. Starting near the top, imagine that there are more spawning adults and more eggs in redds. We would then expect to see more emergent fry, more juveniles, more smolts in migration, more salmon in the ocean, more adults entering the Columbia, and a subsequent increase in the number of spawning adults. This is the positive feedback loop that gives the salmon population the opportunity to grow rapidly under favorable conditions.

An equally important feedback works its way around the inner loop in the diagram. If we begin at the top with more spawners, we would expect to see more eggs, more fry and a greater juvenile loss fraction as the fry compete for space in the river. With a higher loss fraction, we expect to see fewer juveniles survive to be smolts, fewer smolts in migration, and fewer adults in the ocean. This means we would see fewer returning adults and less egg deposition. This "density dependent feedback" becomes increasingly strong with larger populations, and it turns out to be crucial to the eventual size of the population. Simulating density dependent feedback is also essential to our understanding of the recovery potential of the salmon population. Suppose, for example, that the salmon experience high losses during the adult migration, This will mean that fewer adults reach the spawning grounds. There will be less egg deposition and fewer emergent fry in the following spring. The new cohort of juveniles will then experience more favorable conditions, and a larger fraction will survive the juvenile stage and migrate to the ocean. The density dependent feedback is crucial to the population's ability to withstand shocks from external conditions.[2]

[2]The shocks could take the form of changes in ocean mortalities, changes in harvesting and changes in the migration mortalities. These

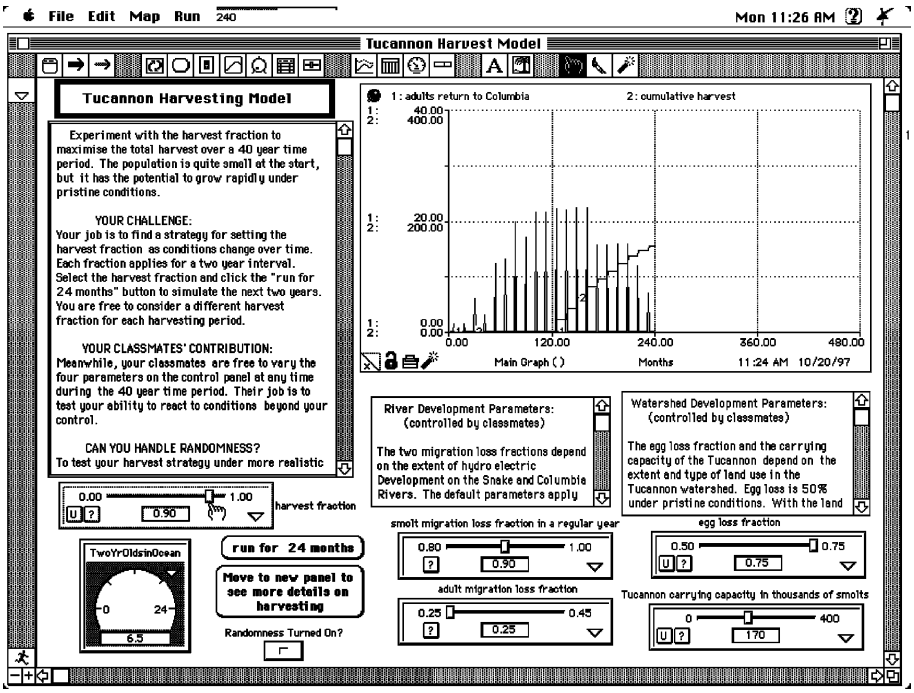

System Dynamics Models of Environment, Energy and Climate Change, Figure 8
Salmon harvesting model to encourage student experimentation

Figure 8 shows a version of the model to encourage student experimentation with harvesting policies. The information fields instruct the students to work in groups of three with one student playing the role of "the harvest manager". The harvest manager's goal is to achieve a large, sustainable harvest through control of the harvest fraction. The other students are given control of the parameters that describe conditions on the Snake and Columbia and in the Tucannon watershed. These students are encouraged to make major and unpredictable changes to test the instincts of the harvest manager.

Models designed for highly interactive simulations of this kind are sometimes called "management flight simulators" because they serve the same function as actual flight simulators. With a pilot simulator, the trainee takes the controls of an electro-mechanical model and tests his instincts for managing the simulated airplane under difficult conditions. The Tucannon harvesting model provides a similar opportunity for environmental students. They can learn the challenge of managing open access fisheries that are vulnerable to over harvesting and the tragedy of the commons [12]. In this particular exercise, students learn that they can achieve a sustainable harvest under a wide variety of difficult and unpredictable conditions. The key to sustainability is harvest manager's freedom to change the harvest fraction in response to recent trends in number of returning adults. This is an important finding for fishery management because it reveals that the population dynamics are not the main obstacle to sustainability. Rather, unsustainable harvesting is more likely to occur when the managers find it difficult to change the harvest fraction in response to recent trends. This is the fundamental challenge of an open-access fishery.

The salmon model is a system dynamics version of the type of modeling commonly performed by population biologists. System dynamics adds clarity and ease of experimentation compared to these models. It also provides a launching point for model expansions that can go beyond population biology. Figure 9 shows an example. This is a student expansion to change the carrying capacity from a user input to a variable that responds to the user's river restoration strategy. The student was trained in geomorphology and was an expert on restoring degraded rivers in the west. The Tucannon began the simulation with 25 miles of river in degraded condition and the remaining 25 miles in a mature, fully restored river with a much higher carrying capacity. The new model per-

shocks are external to the boundary of this model, so one is reminded of Coyle's definition of system dynamics. That is, the model helps us understand how the salmon population could withstand the shocks which fall upon it from the out-side world.

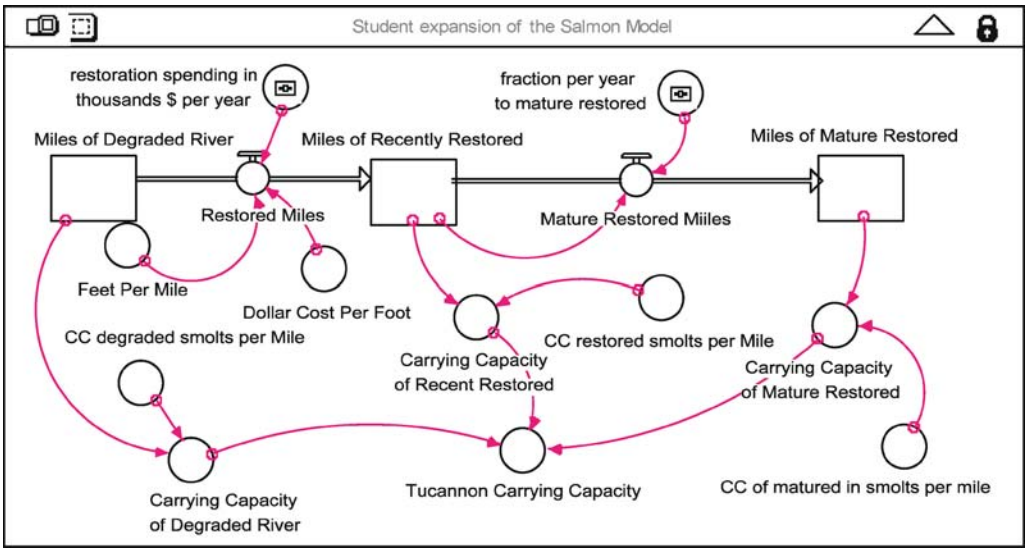

System Dynamics Models of Environment, Energy and Climate Change, Figure 9
Student addition to simulate river restoration

mits one to experiment with the timing of river restoration spending and to learn the impact on the management of the salmon fishery.

The student's model provides another example of interdisciplinary modeling that aids our understanding of environmental systems. In this particular case, the modeling of river restoration is normally the domain of the geomorphologist. The model of the salmon population is the domain of the population biologist. Their work is often conducted separately, and their models are seldom connected. This is unfortunate as the experts working in their separate domains miss out on the insights that arise when two perspectives are combined within a single model. In the student's case, surprising insights emerged when the combined model was used to study the economic value of the harvesting that could be sustained in the decades following the restoration of the river. To the student's surprise, the new harvesting could "pay back" the entire cost of the river restoration in less than a decade.

Models of Climate Change

Scientists use a variety of models to keep track of the greenhouse gasses and their impact on the climate. Some of the models combine simulations of the atmosphere, soils, biomass and ocean response to anthropogenic emissions. The more developed models include CO_2, methane, nitrous oxides and other greenhouse gas (GHG) emissions and their changing concentrations in the atmosphere. Claussen [2] classifies climate models as simple, intermediate and comprehensive. The simple models are sometimes called "box models" since they represent the storage in the system by highly aggregated stocks. The parameters are usually selected to match the results from more complicated models. The simple models can be simulated faster on the computer, and the results are easier to interpret. This makes them valuable for sensitivity studies and in scenario analysis [13].

The comprehensive models are maintained by large research centers, such as the Hadley Center in the UK. The term "comprehensive" refers to the goal of capturing all the important processes and simulating them in a highly detailed manner. The models are sometimes called GCMs (general circulation models). They can be used to describe circulation in the atmosphere or the ocean. Some simulate both the ocean and atmospheric circulation in a simultaneous, interacting fashion. They are said to be coupled general circulation models (CGCMs) and are considered to be the "most comprehensive" of the models available [2]. They are particularly useful when a high spatial resolution is required. However, a disadvantage of the CGCMs is that only a limited number of multi-decadal experiments can be performed even when using the most powerful computers.

Intermediate models help scientists bridge the gap between the simple and the comprehensive models. Claussen [2] describes eleven models of intermediate complexity. These models aim to "preserve the geographic integrity of the Earth system" while still providing the opportunity for multiple simulations to "explore the param-

eter space with some completeness. Thus, they are more suitable for assessing uncertainty". Figure 10 characterizes the different categories of models based on their relative emphasis on:

- number of processes (right axis)
- detailed treatment of the each process (left axis), and the
- extent of integration among the different processes (top axis).

Regardless of the methodology, climate modeling teams must make some judgments on where to concentrate their attention. No model can achieve maximum performance along all three dimensions. (Figure 10 uses the dashed lines to draw our attention to the impossible task of doing every thing within a single model.)

The comprehensive models strive to simulate as many processes as possible with a high degree of detail. This approach provides greater realism, but the models often fail to simulate the key feedback loops the link that atmospheric system with the terrestrial and oceanic systems. (An example is the feedback between CO_2 emissions, temperatures and the decomposition of soil carbon. If higher temperatures lead to accelerated decomposition, the soils could change from a net sink to a net source of carbon [15].) The simple models sacrifice detail and the number of processes in order to focus on the feedback effects between the processes. Using Claussen's terminology, one would say that such models aim for a high degree of "integration". However, the increased integration is achieved by limiting the number of processes and the degree of detail in representing each of the processes.

System dynamics has been used in a few applications to climate change. These applications fit in the category of simple models whose goal is to provide a highly integrated representation of the system. Two examples are described here; both deal with the complexities of the global carbon cycle.

System Dynamics Models of the Carbon Cycle

Figures 11 and 12 depict the global carbon cycle. Figure 11 shows the carbon flows in a visual manner. Figure 12 uses the Vensim stock and flow icons to summarize carbon storage and flux in the current system. The storage is measured in GT, gigatons of carbon, (where carbon is the C in CO_2). The flows are in GT/year of carbon with values rounded off for clarity.

The left side of Fig. 12 shows the flows to the terrestrial system. The primary production removes 121 GT/yr from the atmosphere. This outflow exceeds the return flows by

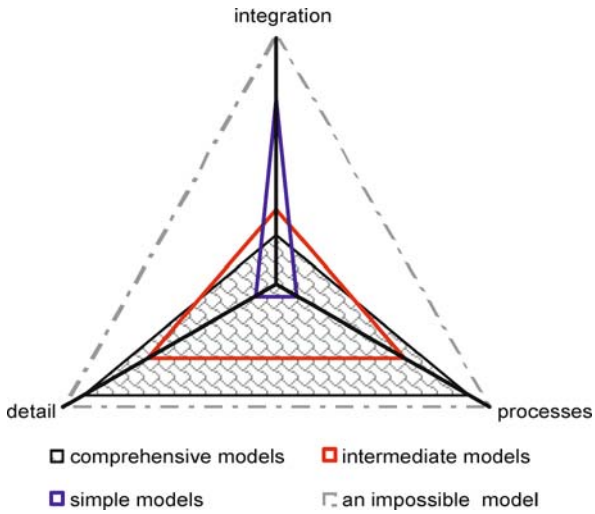

System Dynamics Models of Environment, Energy and Climate Change, Figure 10
Classification of climate models

1 GT/year. This imbalance suggests that around 1 GT of carbon is added to the stocks of biomass and soil each year. So the carbon stored in the terrestrial system would grow over time (perhaps due to extensive reforestation of previously cleared land.) The right side of Fig. 12 shows the flows from the atmosphere to the ocean. The CO_2 dissolved in the ocean each year exceeds the annual release back to the atmosphere by 2 GT. The total, net-flow out of the atmosphere is 3 GT/year which means that natural processes are acting to negate approximately half of the current anthropogenic load.

As the use of fossil fuels grows over time, the anthropogenic load will increase. But scientists do not think that natural processes can continue to negate 50% of an ever increasing anthropogenic load. On the terrestrial side of the system, there are limits on the net flow associated with reforestation of previously cleared land. And there are limits to the carbon sequestration in plants and soils due to nitrogen constraints. On the ocean side of the system, the current absorption of 2 GT/year is already sufficiently high to disrupt the chemistry of the ocean's upper layer. Higher CO_2 can reduce the concentration of carbonate, the ocean's main buffering agent, thus affecting the ocean's ability to absorb CO_2 over long time periods.

Almost of the intermediate and comprehensive climate models may be used to estimate CO_2 accumulation in the atmosphere in the future. For this article, it is useful to draw on the mean estimate published in *Climatic Change* by Webster [23]. He used the climate model developed at the Massachusetts Institute of Technology, one of the eleven models of "intermediate complexity" in the

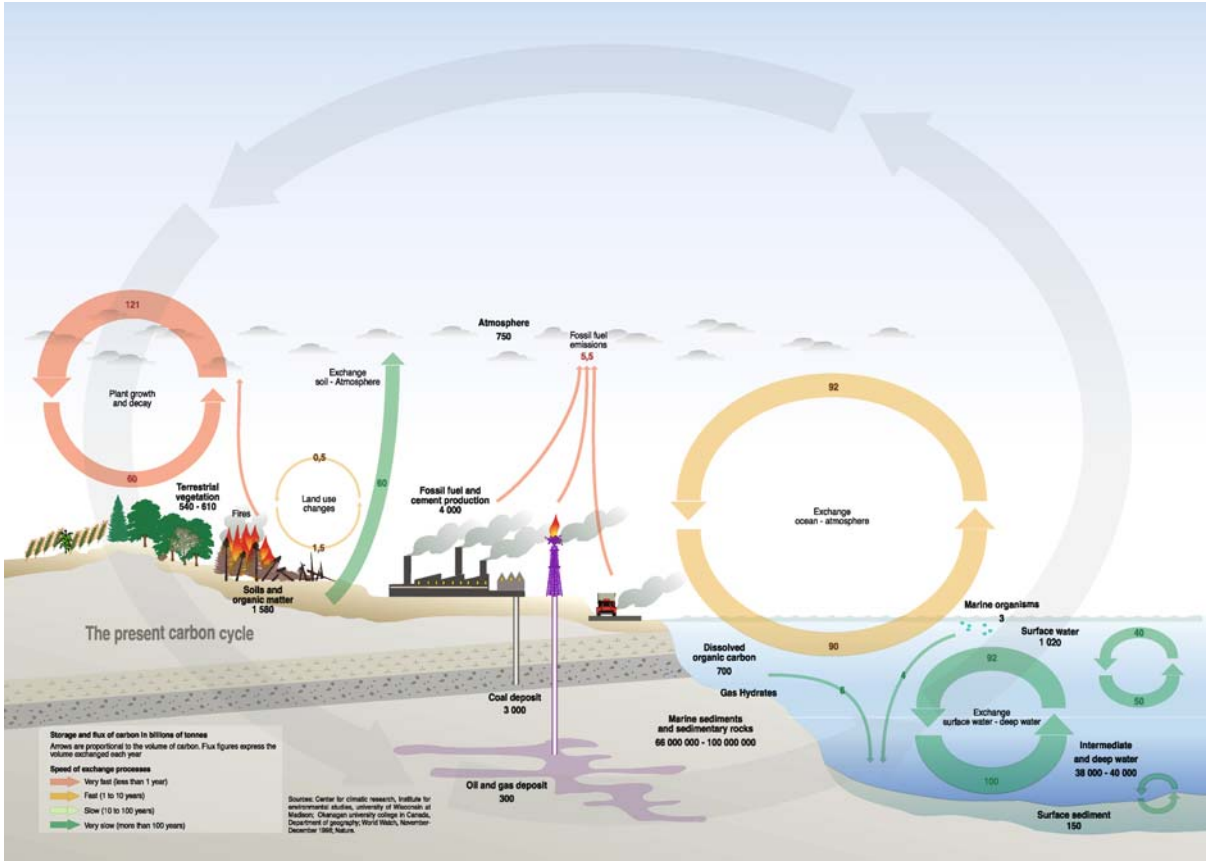

System Dynamics Models of Environment, Energy and Climate Change, Figure 11
The global carbon cycle. (Source: United Nations Environmental Program (UNEP) http://www.unep.org/)

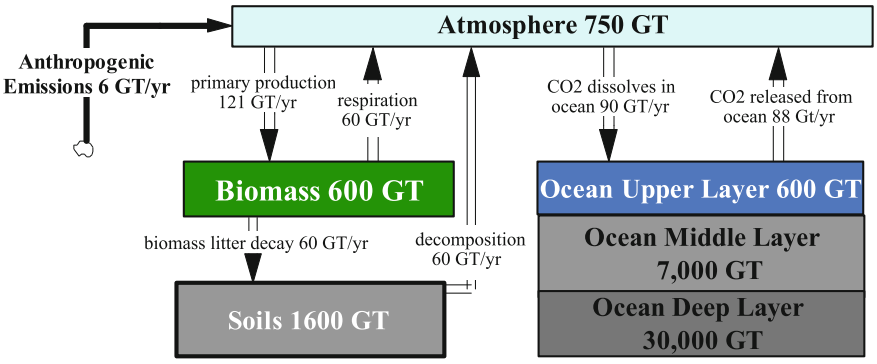

System Dynamics Models of Environment, Energy and Climate Change, Figure 12
Diagram of the stocks and flows in the carbon cycle

review by Claussen [2]. The model began the simulation in the year 2000 with an atmospheric CO_2 concentration of 350 parts per million (ppm). (This concentration corresponds to around 750 GT of carbon in the atmosphere.) The mean projection assumed that anthropogenic emissions would grow to around 19 GT/year by 2100. The mean projection of atmospheric CO_2 was around 700 ppm by 2100. The amount of CO_2 in the atmosphere would be twice as high at the end of the century.

Figure 13 shows the simplest possible model to explain the doubling of atmospheric CO_2. The stock accumulates the effect of three flows, each of which is specified by the

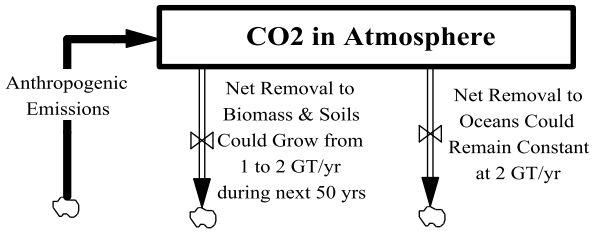

System Dynamics Models of Environment, Energy and Climate Change, Figure 13
Simple model to understand accumulation of CO_2 in the atmosphere

user. Anthropogenic emissions are set to match Webster's assumption. They grow to 19 GT/year by the end of the century. Net removal to oceans is assumed to remain constant at 2 GT/year for the reasons given previously. Net removal to biomass and soils is then subject to experimentation to allow this simple model to match Webster's results. A close match is provided if the net removal increases from 1 to 2 GT/year during the first half of the century and then remains at 2 GT/year for the next fifty years. With these assumptions, the CO_2 in the atmosphere would double from 750 to 1500 GT during the century. This means that the atmospheric concentration would double from 350 to 700 ppm, the same result published by Webster [23].

The model in Fig. 13 is no more than an accumulator. This is the simplest of possible models to add insight on the dynamics of CO_2 accumulation in the atmosphere. It includes a single stock and only three flows, with all of the flows specified by the user. There are no feedback relationships which are normally at the core of system dynamics models. This extreme simplification is intended to make the point that simple models may provide perspective on the dynamics of a system. In this case, a simple accumulator can teach one about the sluggish response of atmospheric CO_2 in the wake of reductions in the anthropogenic emissions. As an example, suppose carbon policies were to succeed in cutting global emissions dramatically in the year 2050. By this year, emissions would have reached 10 GT/yr, so the supposed policy would reduce emissions to 5 GT/yr. What might then happen to CO_2 concentrations in the atmosphere for the remainder of the century? Experiments with highly educated adults [21] suggest that some subjects would answer this question with "pattern matching" reasoning. For example, if emissions are cut in half, it might make sense that CO_2 concentrations would be cut in half as well. But pattern matching leads one astray since the accumulation of CO_2 in the atmosphere responds to the total effect of the flows in Fig. 13. Were anthropogenic emissions to be reduced to 5 GT/year and net removals were to remain at 4 GT/year, the CO_2 concentration would continue to grow, and atmospheric CO_2 would reach 470 ppm by the end of the century.

The model in Fig. 13 is an extreme example to make a point about the usefulness of simple models. The next example is by Fiddaman [6]. It was selected as illustrative of the type of model that would emerge after a system dynamics study. Figure 14 shows the view of the carbon cycle, one of 30 views in the model. The model simulates the climate system within a larger system that includes growth in human population, growth in the economy, and changes in the production of energy. The model was organized conceptually as nine interacting sectors with a high degree of coupling between the energy, economic and the climate sectors.

Fiddaman focused on policy making, particularly the best way to put a price on carbon. In the current debate, this question comes down to a choice between a carbon tax and a carbon market. His simulations add support to those who argue that the carbon tax is the preferred method of putting a price on carbon. The simulations also provide another example of the usefulness of system dynamics models that cross disciplinary boundaries. By representing the economy, the energy system and the climate system within a single, tightly coupled model, he provides another example of the power of system dynamics to promote interdisciplinary exploration of complex problems.

System dynamics has also been applied to a wide variety of energy problems [1,7]. Indeed, a key word frequency count in 2004 revealed nearly 400 energy entries in the System dynamics bibliography [11]. Many of these applications deal with the electric power industry, and I have selected two electric studies to illustrate the usefulness of the approach. The first involves the regulatory and financial challenges of the investor owned electric utilities in the United States.

Lessons from the Regulated Power Industry in the 1970s

The 1970s was a difficult decade for the regulated power companies in the United States. The price of oil and gas was increasing rapidly, and the power companies were frequently calling on their regulators to increase retail rates to cover the growing cost of fuel. The demand for electricity had been growing rapidly during previous decades, often at 7 %/year. At this rate, the demand doubled every decade, and the power companies faced the challenge of doubling the amount of generating capacity to ensure that demand would be satisfied. The power companies dealt with this challenge in previous decades by building ever

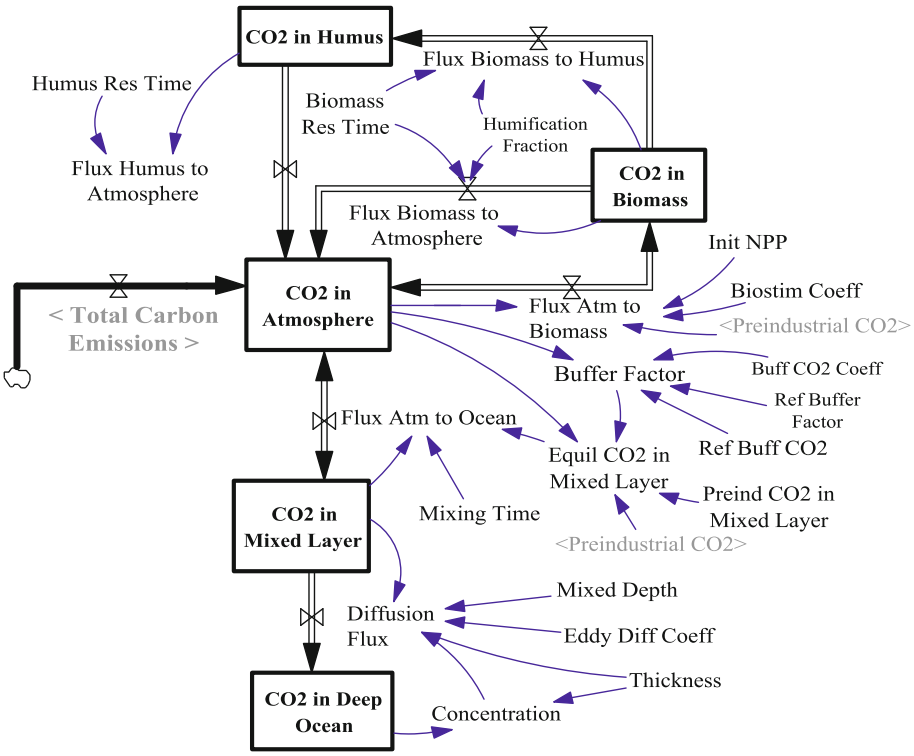

System Dynamics Models of Environment, Energy and Climate Change, Figure 14
Representation of the carbon cycle in the model by Fiddaman [6]

larger power plants (whose unit construction costs declined due to economies of scale). But the economies of scale were exhausted by the 1970s, and the power companies found themselves with less internal funds and poor financial indicators. Utilities worried that the construction of new power plants would not keep pace with demand, and the newspapers warned of curtailments and blackouts.

Figure 15 puts the financial problems in perspective by showing the forecasting, planning and construction processes. The side by side charts allows one to compare the difficult conditions of the 1970s with conditions in previous decades. Figure 15a shows the situation in the 1950s and 1960s. Construction lead times were around 5 years, so forecasts would extend 5 years into the future. Given the costs at the time, the power company would need to finance $3 billion in construction. This was a substantial, but manageable task for a company with $10 billion in assets.

Figure 15b shows the dramatic change in the 1970s. Construction lead times had grown to around 10 years, and construction costs had increased as well. The power company faced the challenge of financing $10 billion in construction with an asset base of $10 billion. The utility executives turned to the regulators for help. They asked for higher electricity rates in order to increase annual revenues and improve their ability to attract external financing. The regulators responded with substantial rate increases, but they began to wonder whether further rate increases would pose a problem with consumer demand. If consumers were to lower electricity consumption, the utility would have less sales and less revenues. The executives might then be forced to request another round of rate increases. Regulators wondered if they were setting loose a "death spiral" of ever increasing rates, declining sales and inadequate financing.

Figure 16 puts the problem in perspective by showing the consumer response to higher electricity rates along side of the other key feedback loops in the system. Higher electricity rates do pose the problem which came to be called "the death spiral". But the death spiral does not act in isolation. Figure 16 reminds us that higher rates lead to lower consumption and to a subsequent reduction in the demand forecast and in construction. After delays for the new power plants to come on line, the power companies experiences a reduction in its "rate base" and the "allowed revenues". When the causal relationships are traced around the outer loop, one sees a negative feedback loop that could act to stabilize the situation. The prob-

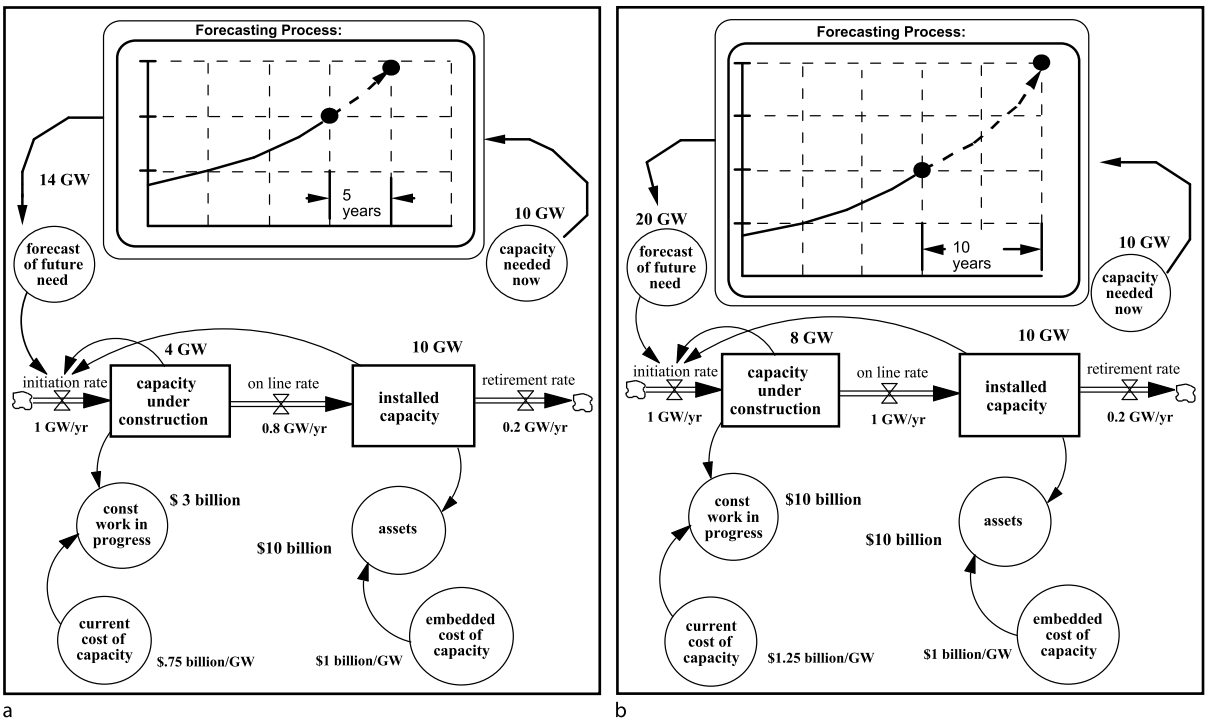

System Dynamics Models of Environment, Energy and Climate Change, Figure 15
a The electric utility's financial challenge during the 1950s and 1960s. b The electric utility's financial challenge during the 1970s

lem, however, is that the delays around the outer loop are substantially longer than the delay for the death spiral.

The utility companies financial challenge was the subject of several system dynamics studies in the 1970s and 1980s [7]. The studies revealed that the downward spiral could pose difficult problems, especially if consumers reacted quickly while utilities were stuck with long-lead time, capital intensive power plants under construction. The studies showed that utility executives needed to do more than rely on regulators to grant rate increases; they needed to take steps on their own to soften the impact of the death spiral. The best strategy was to shift the investments to technologies with shorter lead times. (As an example, a power company in coal region would do better to switch from large to smaller coal plants because of the small plants' shorter lead time.) The studies also revealed that the company's financial situation would improve markedly with slower growth in demand. By the late 1970s and early 1980s, many power companies began to provide direct financial incentives to their customers to slow the growth in demand. System dynamics studies showed that the company-sponsored efficiency programs would be beneficial to the both the customers (lower electric bills) and to the power companies (improved financial performance).

An essential feature of the utility modeling was the inclusion of power operations along side of consumer behavior, company forecasting, power plant construction, regulatory decision making and company financing. This interdisciplinary approach is common within the system dynamics community because practitioners believe that insights will emerge from simulating the key feedback loops. (This belief leads one to follow the cause and effect connections around the key loops regardless of the disciplinary boundaries that are crossed along the way.) This approach contrasts strongly with the customary modeling framework of large power companies who were not familiar with system dynamics. Their approach was to assign models to different departments (i. e., operations, accounting and forecasting) and string the models together to provide a view of the entire corporation over the long-term planning interval.

Figure 17 shows what can happen when models within separate departments are strung together. A large corporation might use 30 models, but this diagram makes the point by describing three models. The analysis would begin with an assumption on future electricity prices over the

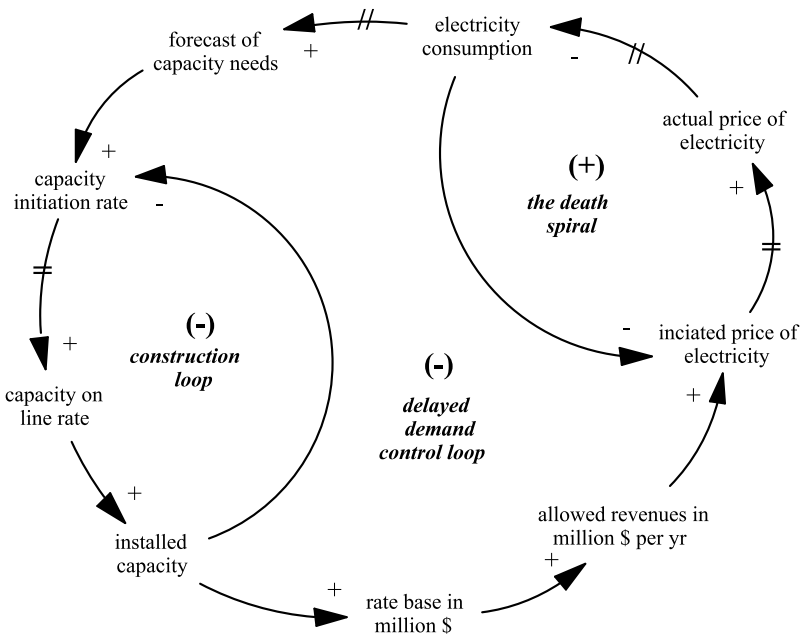

System Dynamics Models of Environment, Energy and Climate Change, Figure 16
Key feedbacks and delays faced by power companies in the 1970s

20-year interval. These are needed to prepare a forecast of the growth in electricity load. The forecast is then given to the planning department which may run a variety of models to select the number power plants to construct in the future. The construction results are then handed to the accounting and rate making departments to prepare a forecast of electricity prices. When the company finally completes the many calculations, the prices that emerge may not agree with the prices that were assumed at the start. The company must then choose whether to ignore the contradiction or to repeat the entire process with a new estimate of the prices at the top of the diagram. This was not an easy choice. Ignoring the price discrepancy was problematic because it was equivalent to ignoring the "death spiral," one of the foremost problems of the 1970s. Repeating the analysis was also problematic. The new round of calculations would be time consuming, and there was no guarantee that consistent results would be obtained at the end of the next iteration.

The power companies' dilemma from the 1970s is described here to make an important point about the usefulness of system dynamics. System dynamics modeling is ideally suited for the analysis of dynamic problems that require a feedback perspective. The method allows one to "close the loop", as long as one is willing to cross the necessarily disciplinary boundaries. In contrast, other modeling methods are likely to be extremely time consuming or fall short in simulating the key feedbacks that tie the system together.

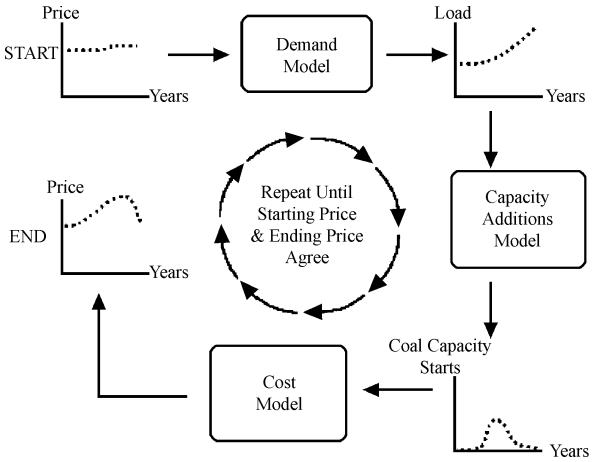

System Dynamics Models of Environment, Energy and Climate Change, Figure 17
The iterative approach often used by large power companies in the 1970s

Simulating the Power Industry Response to a Carbon Market

The world is getting warmer, both in the atmosphere and in the oceans. The clearest and most emphatic description

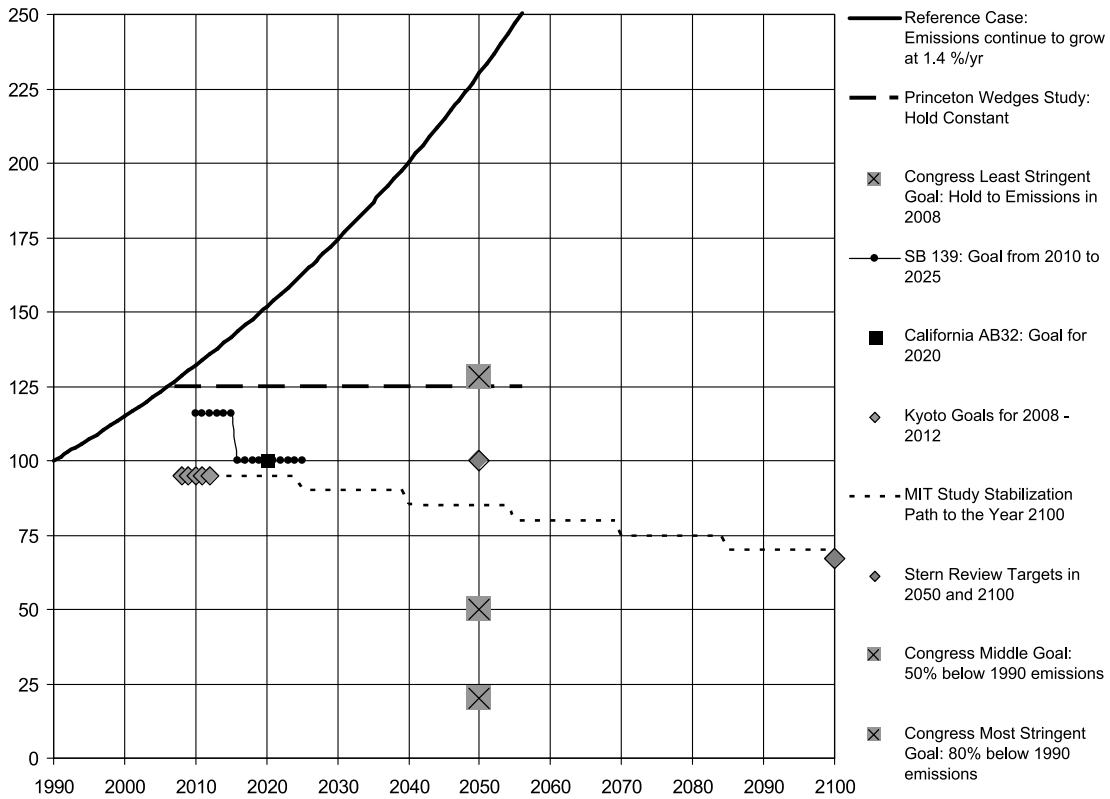

System Dynamics Models of Environment, Energy and Climate Change, Figure 18
Comparison of goals for emissions (100 on the vertical axis represent emissions in the year 1990)

of global warming was issued by the intergovernmental panel on climate change (IPCC) in February of 2007. Their summary for policymakers (p. 4 in [14]) reported that the "Warming of the climate system is unequivocal, as is now evident from observations of increases in global average air and ocean temperatures, widespread melting of snow and ice and rising global mean sea level". The IPCC concluded that "most of the observed increase is very likely due to the observed increase in anthropogenic greenhouse gas concentrations". As a consequence of the IPCC and other warnings, policymakers around the world are calling for massive reductions in CO_2 and other greenhouse gas (GHG) emissions to reduce the risks of global warming.

Figure 18 summarizes some of the targets for emission reductions that have been adopted or proposed around the world. In many cases, the targets are specified relative to a country's emissions in the year 1990. So, for ease of comparison, the chart uses 100 to denote emissions in the year 1990. Emissions have been growing at around 1.4%/year. The upward curve shows the future emissions if this trend continues: emissions would reach 200 by 2040 and 400 by 2090. The chart shows the great differences in the stringency of the targets. Some call for holding emissions constant; others call for dramatic reductions over time. Some targets apply to the next two decades; many extend to the year 2050; and some extend to the year 2100. However, when compared to the upward trend, all targets require major reductions relative to business as usual.

The targets from the Kyoto treaty are probably the best known of the goals in Fig. 18. The treaty became effective in February of 2005 and called for the Annex I countries to reduce emissions, on average, by 5% below 1990 emissions by the year 2008 and to maintain this limit through 2012. The extension of the Kyoto protocol beyond 2012 is the subject of ongoing discussions. The solid line from 2010 to 2050 represents the "stabilization path" used in the climate modeling by Webster [23]. The limit on emissions was imposed in modeling calculations designed to stabilize atmospheric CO_2 at 550 ppmv or lower. The scenario assumed that the Kyoto emissions caps are adopted by all countries by 2010. The policy assumed that the caps would be extended and then further lowered by 5% every 15 years. By the end of the century, the emissions would be 35% below the value in 1990.

System Dynamics Models of Environment, Energy and Climate Change, Figure 19
Map of the western electricity system

This article concentrates on Senate Bill 139, The Climate Stewardship Act of 2003. Figure 19 shows the S139 targets over the interval from 2010 to 2025. The bill called for an initial cap on emissions from 2010 to 2016. The cap would be reduced to a more challenging level in 2016, when the goal was to limit emissions to no more than the emissions from 1990. S139 was introduced by Senators McCain and Lieberman in January of 2003. It did not pass, but it was the subject of several studies including a highly detailed study by the Energy Information Administration [5]. The EIA used a wide variety of models to search for the carbon market prices that would induce industries to lower emissions to come into compliance with the cap. The carbon prices were estimated at $22 per metric ton of CO_2 when the market was to open in 2010. They were projected to grow to $60 by the year 2025.

The EIA study showed that the electric power sector would lead the way in reducing emissions. By the year 2025, power sector emissions would be reduced 75% below the reference case. This reduction was far beyond the reductions to be achieved by other sectors of the economy. This dramatic response was possible given the large use of coal in power generation and the power industry's wide range of choices for cleaner generation.

A system dynamics study of S139 was conducted at Washington State University (WSU) to learn if S139 could lead to similar reductions in the west. Electricity generation in the western system is provided in a large, interconnected power system shown in Fig. 19. This region has considerably more hydro resources, and it makes less use of coal-fired generation than the nation as a whole. The goal was to learn if dramatic reductions in CO_2 emissions could be possible in the west and to learn if they could be achieved with generating technologies that are commercially available today.

The opening view of the WSU model is shown in Fig. 20. The model deals with generation, transmission and distribution to end use customers, with price feedback on the demand for electricity. The model is much larger than the textbook models described earlier in this article. Fifty

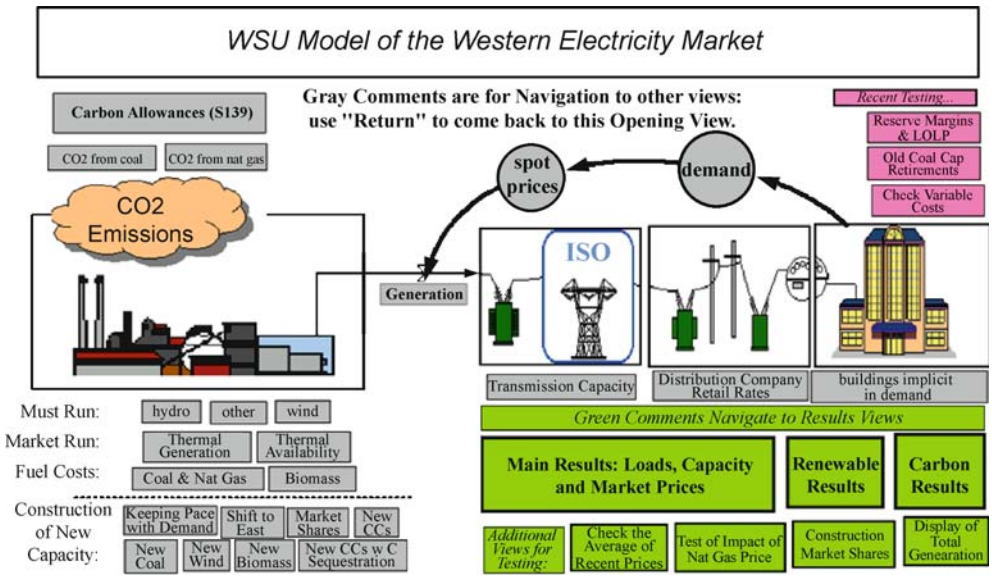

System Dynamics Models of Environment, Energy and Climate Change, Figure 20
Opening view of the model of the western electricity system

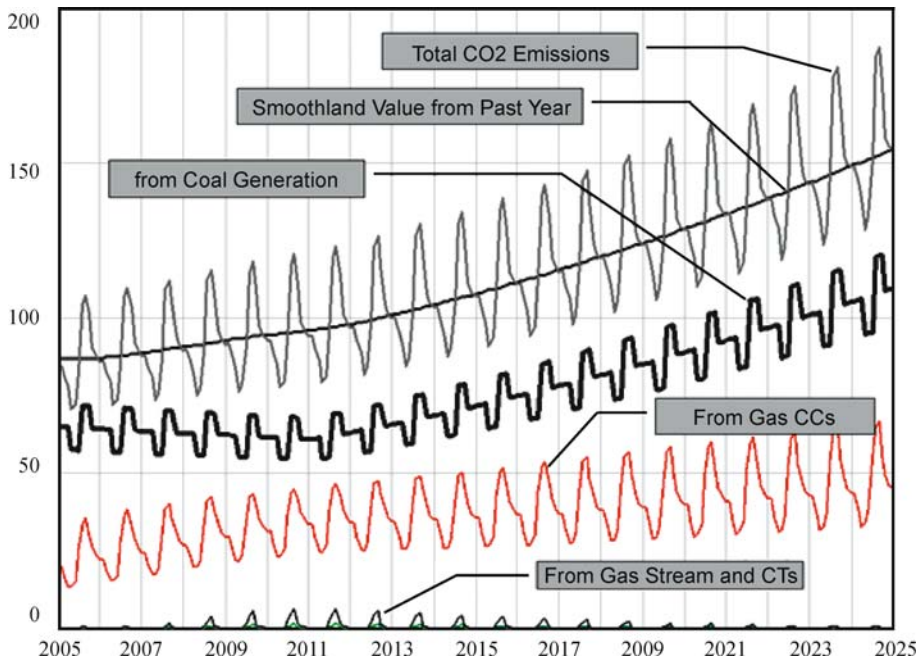

System Dynamics Models of Environment, Energy and Climate Change, Figure 21
Annual emissions in a base case simulation (annual emissions are in million metric tons of carbon)

views are required to show the all the diagrams and the simulation results. The opening view serves as a central hub to connect with all the other views.

The opening view uses Vensim's comment icons to draw attention to the CO_2 emissions in the model. The emissions arise mainly from coal-fired power plants, as shown in Fig. 21. A smaller, but still significant fraction of the emissions is caused by burning natural gas in combined cycle power plants. Total emissions vary with the seasons of the year, with the peak normally appearing in the summer when almost all of the fossil-fueled plants are needed to satisfy peak demand. The base case shows annual emissions growing by over 75% by the year 2025.

A major challenge for the system dynamics model is representing power flows across a transmission grid. Finding the flows on each transmission line and the prices in each area is difficult with the standard tools of system dynamics. It simply doesn't make sense to represent the power flows with a combination of stocks, flows and feedback processes to explain the flows. It makes more sense to calculate the flows and prices using traditional power systems methods, as explained by Dimitrovski [4]. The power flows were estimated using an algebraic approach which power engineers label as a reduced version of a direct-current optimal power flow calculation. The solution to the algebraic constraints were developed with the Matlab software and then transferred to user-defined functions to operate within the Vensim software. The Vensim simulations were set to run over twenty years with time in months. (A typical simulation required 240 months with changes during a typical day handled by carrying along separate calculations for each of 24 h in a typical day.) These are extensive calculations compared to many system dynamics models, so there was concern that we would lose the rapid simulation speed that helps to promote interactive exploration and model testing. The important methodological accomplishment of this project was the inclusion of network and hourly results within a long-term model without losing the rapid simulation response that encourages users to experiment with the model.

One of the model experiments called for a new simulation with carbon prices set to follow the \$20 to \$60 trajectory projected by the EIA for S139. These prices were specified as a user input, and the model responded with a change in both short-term operations and long-term investments. The important result was a 75% reduction in CO_2 emissions by the end of the simulation. This dramatic reduction corresponds almost exactly to the EIA estimate of CO_2 reduction for the power industry in the entire US.

Figure 22 helps one understand how CO_2 emissions could be reduced by such a large amount. These diagrams show the operation of generating units across the Western US and Canada for a typical day in the summer of the final year of the simulation. Figure 22a shows the reference

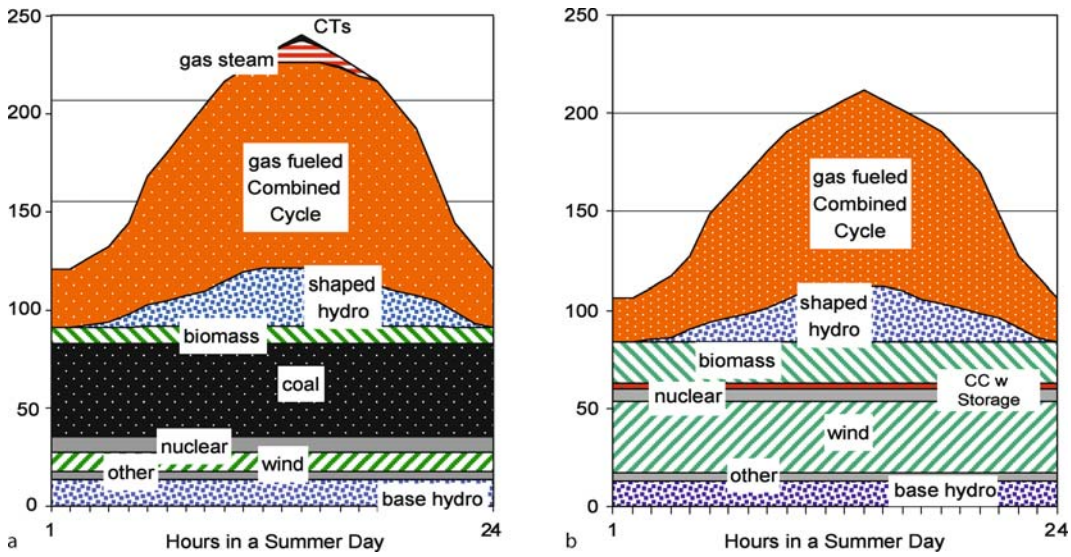

System Dynamics Models of Environment, Energy and Climate Change, Figure 22
a Projected generation for a peak summer day in 2024 in the reference case. b Projected generation for a peak summer day in 2024 in the S139 case

case; Figure 22b shows the case with S139. The side by side comparison helps one visualize the change in system operation. A comparison of the peak loads shows that the demand for electricity would be reduced. The reduction is 9%, which is due entirely to the consumers' reaction to higher retail electric prices over time.

Figure 22b shows large contributions from wind and biomass generation. Wind generation is carbon free, and biomass generation is judged to be carbon neutral, so these generating units make an important contribution by the end of the simulation. Both of these generating technologies are competitive with today's fuel prices and tax credits. The model includes combined cycle gas generation equipped with carbon capture and storage, a technology that is not commercially available today. The model assumes that advances in carbon sequestration over the next two decades would allow this technology to capture a small share of investment near the end of the simulation. By the year 2025, the combined cycle plants with sequestration equipment would provide 2% of the generation.

The most important observation from Fig. 22 is the complete elimination of coal-fired generation in the S139 case. Coal-fired units are shown to operate in a base load mode in the reference simulation. They provide around 28% of the annual generation, but they account for around two/thirds of the CO_2 emissions in the western system. The carbon prices from S139 make investment in new coal-fired capacity unprofitable at the very start of the simulated market in 2010. As the carbon prices increase, utilities to cut back on coal-fired generation and compensate with increased generation from gas-fired CC capacity. In the simulations reported here, this fuel switching would push the coal units into the difficult position of operating fewer and fewer hours in a day. Eventually this short duration operation is no longer feasible, and coal generation is eliminated completely by the end of the simulation.

The WSU study of the western electric system was selected as the concluding example because of its novel treatment of network flows inside a system dynamics model [4]. The model is also interesting for its treatment of daily price changes within a long-term model. (Such changes are important to the simulation of revenues in the wholesale market.) From a policy perspective, the study confirms previous modeling of the pivotal role of the electric power industry in responding to carbon markets. The study indicated that the western electricity system could achieve dramatic reductions in CO_2 emissions within 15 years after the opening of a carbon market, and it could do so with technologies that are commercially available today [8].

Conditions for Effective Interdisciplinary Modeling

All of the applications demonstrate the usefulness of system dynamics in promoting interdisciplinary modeling. The article concludes with comments on the level of effort and the conditions needed for effective, interdisciplinary modeling.

The examples in this article differ substantially in the level of effort required, from several weeks for the classroom examples to several years for the energy studies. The textbook examples involved student expansions of models of Mono Lake and the Tucannon salmon. The expansions were completed by undergraduate students in projects lasting two or three weeks. The key was the students' previous education (classes from many different departments) and their receptiveness to an interdisciplinary approach.

Fiddaman's model of the climate and energy system [6] was a more ambitious exercise, requiring several years of effort as part of his doctoral research. Bringing multi-year interdisciplinary modeling projects to a successful conclusion requires one to invest the time to master several disciplines and to maintain a belief that there are potential insights at the end of the effort.

The electric power industry examples were also ambitious projects that required several years of effort. The modeling of the western electricity system was a four-year project with support from the National Science Foundation. The long research period was crucial for it allowed the researchers from power systems engineering, system dynamics and environmental science to take the time to learn from one another. The modeling of the electric company problems in the 1970s was also spread over several years of effort. The success of this modeling was aided by utility planners, managers and modelers who were looking for a systems view of their agency and its problems. They saw system dynamics as a way to tie existing ideas together within an integrated portrayal of their system. Their existing ideas were implemented in models maintained by separate functional areas (i. e., forecasting, accounting, operations). The existing models often provided a foundation for the system dynamics models (i. e., in the same way that the comprehensive climate models in Fig. 11 provide support for the development of the more integrated models). The key to effective, interdisciplinary modeling within such large organizations is support from a client with a strong interest in learning and with managerial responsibility for the larger system.

Future Directions

This article concludes with future directions for system dynamics applications to climate change. People often talk of mitigation and adaptation. Mitigation refers to the challenge of lowering greenhouse gas emissions to avoid dangerous anthropogenic interference with the climate system. Adaptation refers to the challenge of living in a changing world.

Mitigation: The challenge of lowering CO_2 and other GHG emissions is the fundamental challenge of the coming century. The next two decades will probably see various forms of carbon markets, and system dynamics can aid in learning about market design. It is important that we learn how to make these markets work well. And if they don't work well, it's important to speed the transition to a carbon tax policy with better prospects for success. System dynamics can aid in learning about markets, especially if it is coupled with simulating gaming to allow market participants and regulators to "experience" and better understand market dynamics.

Adaptation: The world will continue to warm, and sea levels will continue to rise. These trends will dominate the first half of this century even with major reductions in CO_2 emissions. These and other climate changes will bring a wide variety of problems for management of water resource, public health planning, control of invasive species, preservation of endangered species, control of wildfire, and coastal zone management, just to name a few. Our understanding of the adaptation challenges can be improved through system dynamics modeling. The prospects for insight are best if the models provide an interdisciplinary perspective on adapting to a changing world.

Bibliography

Primary Literature

1. Bunn D, Larsen E (1997) Systems modelling for energy policy. Wiley, Chichester
2. Claussen M et al (2002) Earth system models of intermediate complexity: closing the gap in the spectrum of climate system models. Climate Dyn 18:579–586
3. Coyle G (1977) Management system dynamics. Wiley, Chichester
4. Dimitrovski A, Ford A, Tomsovic K (2007) An interdisciplinary approach to long term modeling for power system expansion. Int J Crit Infrastruct 3(1–2):235–264
5. EIA (2003) United States Department of Energy, Energy Information Administration, Analysis of S139, the Climate Stewardship Act of 2003
6. Fiddaman T (2002) Exploring policy options with a behavioral climate-economy model. Syst Dyn Rev 18(2):243–264
7. Ford A (1999) Modeling the environment. Island Press, Washington
8. Ford A (2008) Simulation scenarios for rapid reduction in carbon dioxide emissions in the western electricity system. Energy Policy 36:443–455
9. Forrester J (1961) Industrial dynamics. Pegasus Communications
10. Forrester J (2000) From the ranch to system dynamics: An autobiography, in management laureates. JAI Press
11. Ford A, Cavana R (eds) (2004) Special Issue of the Syst Dyn Rev
12. Hardin G (1968) The tragedy of the commons. Science 162:1243–1248

13. IPCC (1997) An introduction to simple climate models used in the IPCC second assessment report. ISBN 92-9169-101-1
14. IPCC (2007) Climate change 2007: The physical science basis, summary for policymakers. www.ipcc.ch/
15. Kump L (2002) Reducing uncertainty about carbon dioxide as a climate driver. Nature 419:188–190
16. Meadows DH, Meadows DL, Randers J, Behrens W (1972) The limits to growth. Universe Books
17. Morecroft J (2007) Strategic modelling and business dynamics. Wiley, Chichester
18. Richardson J, Pugh A (1981) Introduction to system dynamics modeling with dynamo. Pegasus Communications
19. Sterman J (2000) Business dynamics. McGraw-Hill, Irwin
20. Sterman J (ed) (2002) Special Issue of the Syst Dyn Rev
21. Sterman J, Sweeney L (2007) Understanding public complacency about climate change. Clim Chang 80(3–4):213–238
22. Warren K (2002) Competitive strategy dynamics. Wiley, Chichester
23. Webster M et al (2003) Uncertainty analysis of climate change and policy response. Climat Chang 61:295–320

Books and Review

Houghton J (2004) Global warming: The complete briefing, 3rd edn. Cambridge University Press, Cambridge

System Dynamics Models, Optimization of

Brian Dangerfield
Centre for OR & Applied Statistics, Salford Business School, University of Salford, Salford, UK

Article Outline

Glossary
Definition of the Subject
Optimization as Calibration
Optimization of Performance (Policy Optimization)
Examples of SD Optimization Reported in the Literature
Future Directions
Bibliography

Glossary

Econometrics A statistical approach to economic modeling in which all the parameters in the structural equations are estimated according to a 'best fit' to historical data.

Maximum likelihood A statistical concept which underpins calibration optimization and which generates the most likely parameter values; it is equivalent to the parameter set which minimizes the chi-square value.

Objective function See **Payoff** below.

Optimization The process of improving a model's results in terms of either an aspect of its performance or by calibrating it to fit reported time series data.

Payoff A formula which expresses the objective, say, maximization of profits, minimization of costs or minimization of the differences between a model variable and historical data on that variable.

Zero-one parameter A parameter which is used as a multiplier in a policy equation and serves the effect of bringing in or removing a particular influence in determining the optimal policy.

Definition of the Subject

The term 'optimization' when related to system dynamics (SD) models has a special significance. It relates to the mechanism used to improve the model vis-à-vis a criterion. This collapses into two fundamentally different intentions. Firstly one may wish to improve the model in terms of its performance. For instance, it may be desired to minimize overall costs of inventory whilst still offering a satisfactory level of service to the downstream customer. So the criterion here is cost, and this would be minimized after searching the parameter space related to service level. The direction of need may be reversed and maximization may be desired as, for instance, if one had a model of a firm and wished to maximize profit subject to an acceptable level of payroll and advertising costs. Here the parameter space being explored would involve both payroll and advertising parameters. This type of optimization might be described generically as *policy optimization*.

Optimization of performance is also the *raison d'etre* of other management science tools, most notably mathematical programming. But such tools are usually static: they offer the 'optimum' resource allocation given a set of constraints and a performance function to either maximize or minimize. These models normally relate to a single time point and may then need to be re-run on a weekly or monthly basis to determine a new optimal resource allocation. In addition, these models are often linear (certainly so in the case of linear programming), whereas SD models are usually non-linear. So the essential differences are that SD model optimization for performance involves both a dynamic and a non-linear model.

A separate improvement to the model may be sought where it is required to fit the model to past time series data. Optimization here involves minimizing a statistical function which expresses how well the model fits a time-series of data pertaining to an important model variable. In other words a vector of parameters are explored with a view to determining the particular parameter combination which

offers the best fit between the chosen important model variable and a past time series data set of this variable. This type of optimization might be generically termed *model calibration*. If *all* the parameters in the SD model are determined in this fashion then the process is equivalent to the technique of econometric modeling. A good comparison between system dynamics and econometric modeling can be found in Meadows and Robinson [12].

Optimization as Calibration

In these circumstances we wish to determine optimal parameters, those which, following a search of the parameter space, offer the best fit of a particular model variable to a time series dataset on that variable taken from real world reporting.

As an example consider a variation of the one of the epidemic models which are made available with the Vensim™ software. The stock-flow diagram is presented as Fig. 1.

In this epidemiological system members of a susceptible population become infected and join the infected population. Epidemiologists call this an S–I model. It is a simpler variation of the S–I–R model which includes recovered (R) individuals.

Suppose some data on new infections (at intervals of five days) are available covering 25 days of a real-world epidemic. The model is set with a time horizon of 50 days which is consistent with, say, a flu epidemic or an infectious outbreak of dysentery in a closed population such as a cruise ship. The 'current' run of the model is shown in Fig. 2, with the real-world data included for comparison.

Clearly there is not a very good correspondence between the actual data and the model variable for the infection rate (infections). We wish to achieve a better calibration, and so there is a need to select relevant parameters through which the calibration optimization can be performed over. Referring back to Fig. 1, we can see that the *fraction infected from contact* and the *rate that people contact other people* are two possible parameters to consider. The initial infected and initial susceptible are also parameters of the model in the strict sense of the term, but we will ignore them on this occasion. In this model the *initial infected* is 10 persons and *initial susceptibles* number 750,000 persons.

The chosen value for the *fraction infected from contact* is 0.1, while that for the *rate that people contact other people* is 5.0. The former is a dimensionless number while the latter is measured as a fraction per day (1/day). This is obtained from consideration of the *rate of potential infectious contacts* (persons/day) as a proportion of the *susceptible population* (persons).

The optimization process for calibration involves reading into the model the time series data, in this case on new infections, and, secondly, determining the range for the search in parameter space. There is usually some basic background knowledge which allows a sensible range to be entered. For instance, a probability can only be specified between 0 and 1.0. In this case we have chosen to specify

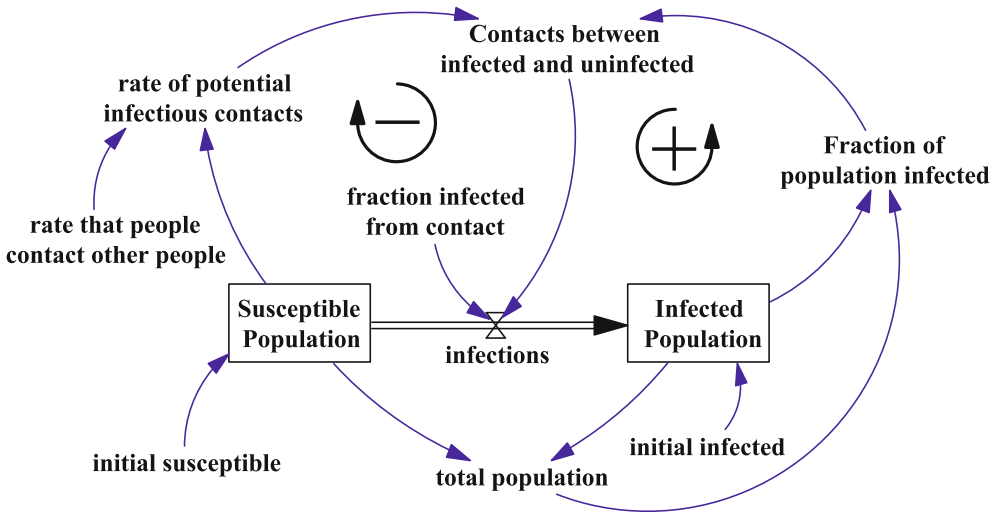

System Dynamics Models, Optimization of, Figure 1
Stock-flow diagram for a simple epidemic model

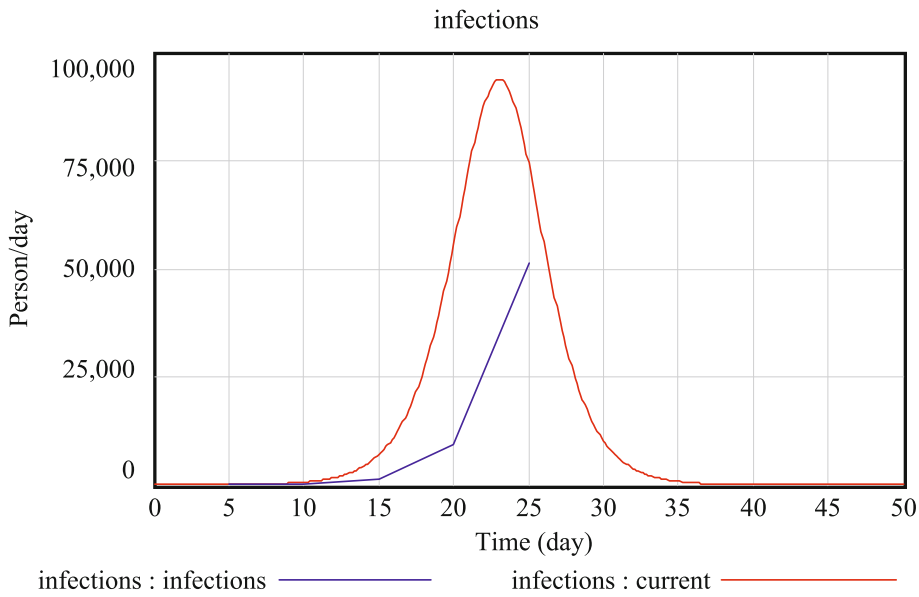

System Dynamics Models, Optimization of, Figure 2
Current (base) run of the model and reported data on infections

the ranges as follows:

$$0.03 \leq \text{fraction infected from contact} \leq 0.7$$

$$2 \leq \text{rate that people contact other people} \leq 10\,.$$

A word of warning is necessary in respect of optimizing delay parameters. Because there is a risk of mathematical instability in the model if the value of DT (the TIME STEP) is too large relative to the smallest first-order delay constant, it is important to ensure the TIME STEP employed in the model is sufficiently small to cope with delay constant values which may be reached during the search of the delay parameter space. In other words ensure the minimum number for the search range on the delay parameter is at least double the value of the TIME STEP.

Maximum Likelihood Estimation and the Payoff Function

The optimization process involves a determination of what are termed statistically as maximum likelihood estimates. In Vensim™ this is achieved by maximizing a payoff function. Initially this is negative and the optimization process should ensure this becomes less negative. An ideal payoff value, after optimization, would be zero. A weighting is needed in the payoff function too, but for calibration optimization this is normally 1.0. Driving the payoff value to be larger by making it less negative has parallels with the operation with the simplex algorithm common in linear programming. This algorithm was conceived initially for problems where the objective function was to be minimized. Its use on maximization problems is achieved by minimizing the negative of the objective function.

During the calibration search, Vensim™ takes the difference between the model variable and the data value, multiplies it by the weight, squares it and adds it to the error sum. This error sum is minimized. Usually data points will not exist at every time point in the model. Here the model TIME STEP is 0.125 (1/8th), but let us assume that reported data on new infections have been made available only at times $t = 5, 10, 15, 20$ and 25 so the sum of squares operation is performed only at these five time points.

The data are shown as Table 1.

System Dynamics Models, Optimization of, Table 1
Data used for calibration experiment

Time	5	10	15	20	25
Infections	30	230	1400	9500	51 400

The Recording Point for Reported Data

System dynamics models differentiate between stock and flow variables and the software used for simulating such models advances by a small constant TIME STEP (also known as DT). This has implications for the task of fitting real-world reported data to each type of system dynamics model variable. The following is the issue: at what point

in a continuum of time steps should the reported data be recorded at? This is important because the reported data has to be read into the model to be compared with the simulated data. The answer will be different for stock and flow variables.

Where the reported data relate to a stock variable the appropriate time point for recording will be known. If it is recorded at the end of the day (say a closing bank balance) then the appropriate point for data entry in the model will be the beginning of the next day. Thus the first data point above is at time $t = 5$ (5.00) and would, if it were a stock, correspond to a record taken at the very end of time period 4.

However, if the data relate to a flow variable, as in the case of new infections here, the number is the total new infections which have occurred over the entire time unit (day, week, month etc.) and so there is a decision to be reached as to which time point the data are entered at. This is because the TIME STEP (DT) is hardly ever as large as the basic time unit which the model is calibrated in. The use of 5 (10, 15 etc.) above implies that the data on new infections over the period of time $t = 0$ to $t = 5$ are compared with the corresponding model variable at time $5 + 1 * \mathrm{DT}$ (and the new infections over the period $t = 5$ to $t = 10$ at time $10 + 1 * \mathrm{DT}$ etc.). A more appropriate selection might be towards the end of the 5-day time period. Following the example above using a TIME STEP $= 0.125$, this might be at time $4 + 7 * \mathrm{DT}$ (that is at 4.875).

Calibration Optimization Results

Based upon the data on new infections shown above and the chosen ranges for the parameter search, the following output is obtained (Table 2). After 114 simulations the optimized values for our two parameters are shown to be 0.08 and 5.12 and the payoff is over 2500 times larger (less negative). Replacing the original parameters with the optimized values reveals the result shown in Fig. 3. To take things further we may wish to put confidence intervals on the estimated parameters. One way of accomplishing this is by profiling the likelihood and is described in Dangerfield and Roberts [3].

System Dynamics Models, Optimization of, Table 2
Results from the calibration optimization

Initial point of search fraction infected from contact = 0.1 rate that people contact other people = 5 Simulations = 1 Pass = 0 Payoff = −2.67655e + 009
Maximum payoff found at: fraction infected from contact = 0.0794332 *rate that people contact other people = 5.11568 Simulations = 114 Pass = 6 Payoff = −1.06161e + 006

Avoid Cumulated Data

There might be a temptation to optimize parameters against cumulated data when the data are reported essentially as a flow, as is the case here. Were the data to be cumulated we would obtain as shown in Table 3.

The results from this optimization are shown in Table 4. The ranges for the parameter space search are kept the same but the payoff function now involves a comparison of the model variable *infected population* with the corresponding cumulated data. Figure 4 shows the resultant fit to infected population is good, but that is manifestly not borne out when we consider the plot of infections obtained from the same optimization run (Fig. 5).

The reason for this is rooted in statistics. The maximum likelihood estimator is equivalent to the chi-squared statistic. This is turn assumes that each expected data value is independent. A cumulated data series would not exhibit this property of independence.

As an aside it is worth pointing out that this model, with suitable changes to the variable names and the time constants involved, could equally represent the diffusion

System Dynamics Models, Optimization of, Table 3
Cumulated reported data for the infected population

Time	5	10	15	20	25
Infected population	30	260	1660	11 160	62 560

System Dynamics Models, Optimization of, Table 4
Results from the calibration using cumulated data

Initial point of search fraction infected from contact = 0.1 rate that people contact other people = 5 Simulations = 1 Pass = 0 Payoff = −2.48206e + 011
Maximum payoff found at: fraction infected from contact = 0.0726811 *rate that people contact other people = 4.96546 Simulations = 145 Pass = 6 Payoff = −212 645 The final payoff is −212 645

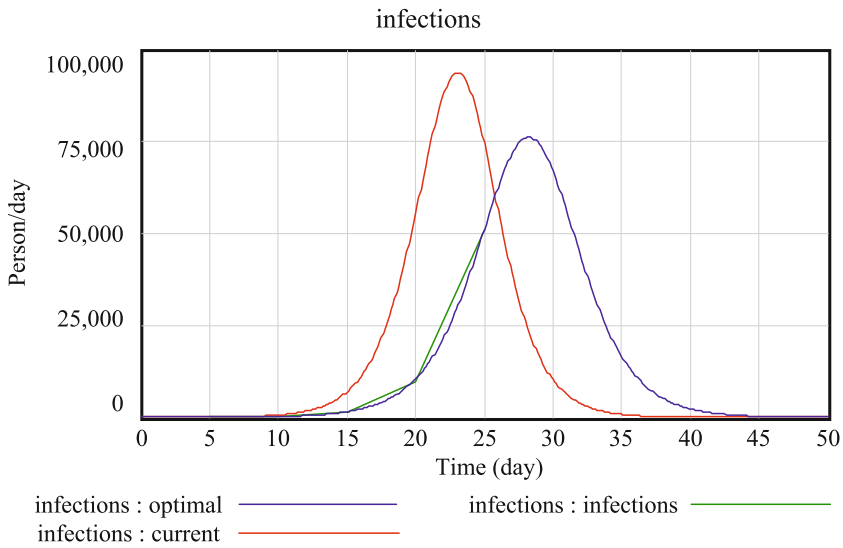

System Dynamics Models, Optimization of, Figure 3
Reported data on infections and optimized (calibrated) model; the base case (current) is reproduced for reference

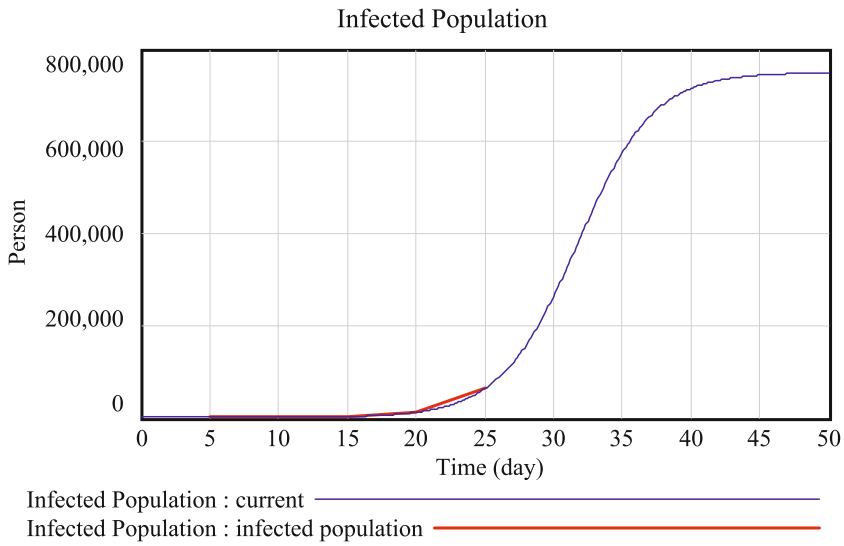

System Dynamics Models, Optimization of, Figure 4
The cumulative model variable (infected population) together with reported data

of a new product into a virgin market. In systems terms the structures are equivalent. The *fraction infected from contact* is the same as, say, the fraction reached by word of mouth or advertising and the *rate that people contact other people* is a measure of the potential interactions at which new products might be mentioned amongst the members of the relevant market segment. An *infected population* is equivalent to a customer base, the number of adopters of the relevant product. So it is possible to shed light on important real-world marketing parameters through a calibration optimization of models of this general structure.

Optimization of Performance (Policy Optimization)

An example model is to be used to illustrate the process of optimization to improve the performance of the system, and this is illustrated in Fig. 6. It concerns the service re-

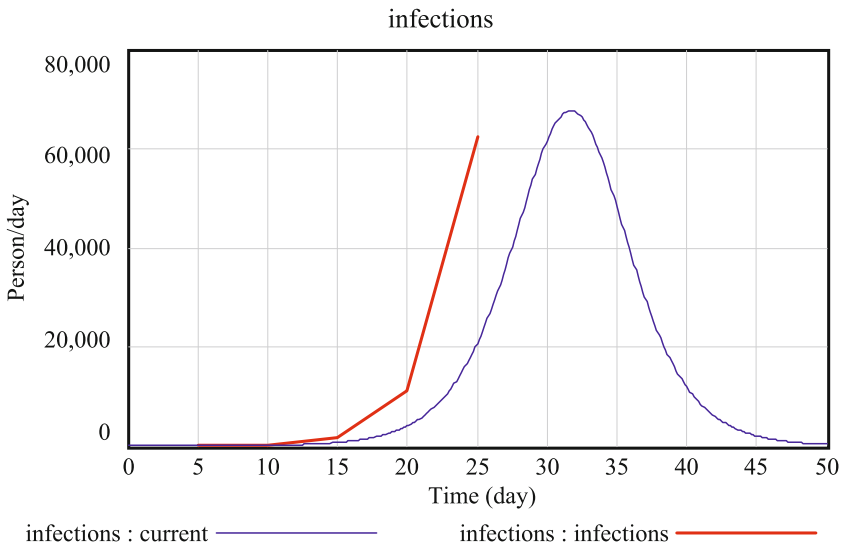

System Dynamics Models, Optimization of, Figure 5
The corresponding fit to infections is poor

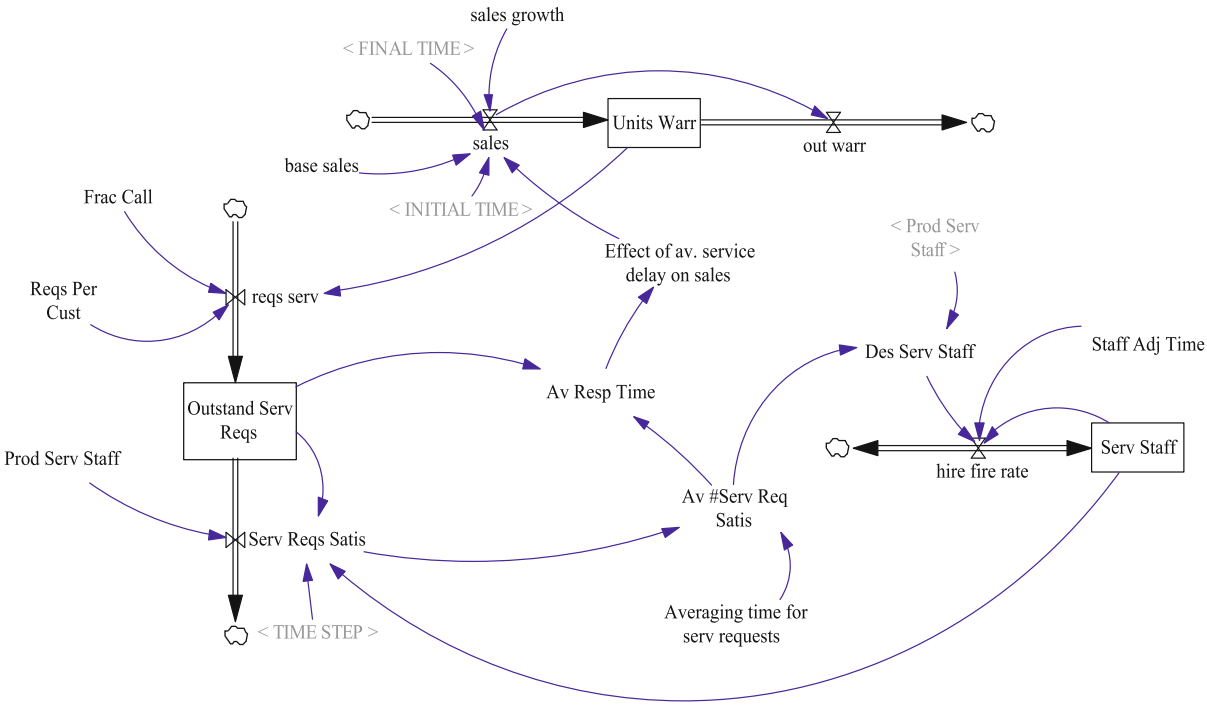

System Dynamics Models, Optimization of, Figure 6
Model of service delays for durable goods under warranty

quirements which can arise following the sale of a durable good. These items are typically sold with a 12-month warranty and during this time the vendor is obliged to offer service if a customer calls for it. In this particular case the vendor is not being responsive in terms of staffing the service section. The result is that as sales grow the increasing number of service requests is putting pressure on the service personnel. The delay in responding to service calls also increases and the effect of this is that future sales are depressed because of the vendor's acquired reputation for

poor service response. The basic behavior mode is overshoot and collapse.

In the model depicted in Fig. 6, the growth process is achieved by a RAMP function which causes sales of the good to increase linearly by 20 units per month from a base of 500 units per month.

The payoff function is restricted to the variable *Sales*. However, this need not be the case. Where a number of variables might be options in a payoff function, it is possible to assign weights to each such that the sum of the weights is 1.0 (or 100). The optimization process will then proceed with the software accumulating a weighted payoff which it will attempt to maximize. Weights are positive when more is better and negative when less is better.

Policy Experiment No. 1

Here it is decided to try to improve the productivity of the service staff. Currently they manage, on average, to respond to 120 calls per operative per month. It may be an option to improve their productivity by, say, providing them with hand-held devices which direct each operative from one call to the next – calls which may have arisen since setting out from their base. In this way their call routing is improved.

The optimization parameter is *Prod Serv Staff*, and we select an upper limit for the search range of 240 calls per person per month. The chosen performance variable is *Sales*, since we wish to maximize this – or at least not have it overly depressed by poor response times. The results are shown in Table 5. We see that the payoff is increased and that the optimum productivity is a modest increase of 2.6 requests per month, on average. This should be easily achievable and perhaps without expenditure on high-tech devices. The graphical output for sales is shown in Fig. 7.

For comparison, the effect of increasing the productivity to as high as 150 calls per month, on average, is also shown. This would represent an increase of 25% and would be much more difficult to accomplish. Here the benefit of optimization is highlighted. A modest increase in productivity returns a visibly improved sales performance (although the basic behavior mode is unchanged), whilst a much greater productivity increase offers little extra benefit for the effort and cost involved in improving productivity.

Policy Experiment No. 2

Another approach to policy optimization involves the use of a zero-one parameter which has the effect of either including or excluding an influence on policy. Suppose it was thought that the quantity of product units in warranty should exert an influence on the numbers of service personnel hired (or fired). The equation for the desired number of service staff (*Des Serv Staff*) can be expressed as:

Des Serv Staff = "Av #Serv Req Satis"/Prod Serv Staff ∗ trigger + ("Av #Serv Req Satis"/Prod Serv Staff)∗(Units Warr/initial units in warranty)∗(1-trigger). (Units: Persons).

The *trigger* variable is initially set to 1.0 and so the more sophisticated policy is not active. The optimization run results are shown in Table 6. Clearly there is benefit from including the more sophisticated policy which takes into account the current numbers of product units in warranty.

The graphical output is unequivocal (Fig. 8). Sales are continuously increasing when the recruitment policy for service personnel takes into account the number of product units in warranty. The depressive effect on sales of poor service performance is non-existent.

Whilst this might seem an obvious policy, it is surprising how easily the naïve alternative might be accepted without question. The number of calls a typical operative can manage each month is well known along with the (his-

System Dynamics Models, Optimization of, Table 5
Optimization results for the productivity of the service staff

Initial point of search Prod Serv Staff = 120 Simulations = 1 Pass = 0 Payoff = 27 743.5
Maximum payoff found at: *Prod Serv Staff = 122.647 Simulations = 27 Pass = 3 Payoff = 29 915

System Dynamics Models, Optimization of, Table 6
Optimization results from selection of policy drivers

Initial point of search trigger = 1 Simulations = 1 Pass = 0 Payoff = 27 743.5
Maximum payoff found at: ∗trigger = 0 Simulations = 13 Pass = 3 Payoff = 47 090.9

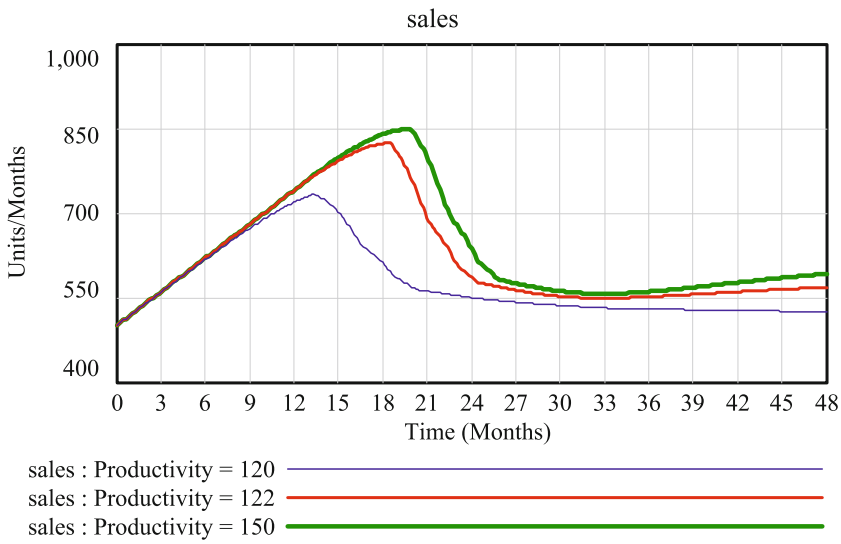

System Dynamics Models, Optimization of, Figure 7
Plots of sales achieved for differing productivities

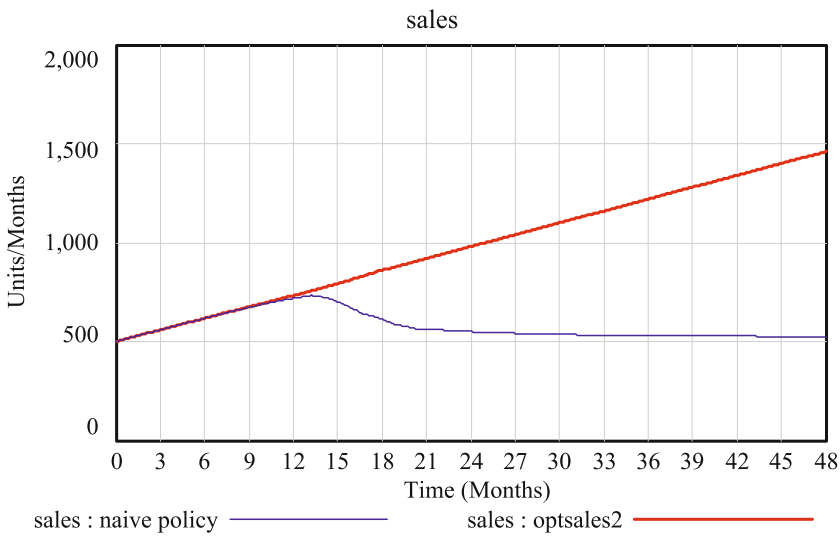

System Dynamics Models, Optimization of, Figure 8
Comparison of sales from two different policy drivers

torical) number of service requests satisfied. Hence, the desired number of staff is more or less fixed. This comes undone when there is a growth in the number of products sold. In this different environment such a simplistic policy can, as shown, lead to overshoot and collapse. Notice needs to be taken of the changing number of product units in warranty in order that a more effective system performance is achieved.

The above experiments are illustrative only, and there is no intention of over-working a simple teaching model in order to uncover an ideal policy. In the case of policy optimization a wide range of possible alternatives exists. Indeed, a process of learning naturally arises through carrying out repeated optimization experiments with the model [1].

Examples of SD Optimization Reported in the Literature

Amongst the earliest work in this area the writings of Keloharju are worthy of mention. He contributed a number of papers on the topic in the pages of *Dynamica*. See, for

example, Keloharju [9]. His work brought the concept to prominence but he did not employ the method on anything other than problems described in text books or postulated by himself. For instance, an application of optimization to the project model contained in Richardson and Pugh's [13] text is contained in Keloharju and Wolstenholme [11]. A statement of the method together with some textbook examples is also available [10]. Additionally, an overview of the methods and their deployment on textbook examples has been contributed by the current author [3]. Finally, there is an example of optimization applied to defence analysis. Again though it is a standard defence model – the armored advance model – rather than any real-world study [14].

Retaining the emphasis on textbook problems for the moment, Duggan [6] employs Coyle's model [1] of the *Domestic Manufacturing Company* to illustrate the methods of multi-objective optimization – an advance over standard SD optimization with its single objective function. The concept of multiple objectives arises from multi-criteria decision-making where a situation can be judged on more than one performance metric. While a multi-objective payoff function can be formulated using a set of weights, it is argued that the selection of the weights is very individual-specific. The multi-objective approach – underpinned by the methods of genetic algorithms – rests upon determining a Pareto-optimal situation, defined as one where no improvement is possible without making some other aspect worse. In other words the method strives for an optimal solution which is not dominated by any other solution. The author demonstrates the approach combining two objectives in the model: one for the differences between desired stock and actual and another between desired backlog and actual.

In terms of applications to real-world problems, the current author has also used the methods of optimization in research conducted in connection with modeling the epidemiology of HIV/AIDS. Fitting a model of AIDS spread to data was carried out for a number of European countries [2,4]. The optimized parameters furnished support for some of the features of AIDS epidemiology which, at the time, were being uncovered by other branches of science. For example, the optimized output revealed that a U-shaped profile of infectiousness in a host was necessary in order to achieve a best fit to data on new AIDS cases. This infectiousness profile was also evidenced by virologists who had analyzed patients' blood and other secretions on a longitudinal basis.

Within this strand of research, a much more complex optimization was performed using American data on transfusion-associated AIDS cases [5]. The purpose here was to estimate the parameters for a number of plausible statistical HIV incubation distributions. Given the nature of the data, the point of infection could be quite accurately determined, but two difficulties were evident: the data were right-censored and the number receiving infected transfusions in each quarter was unknown. However, the SD optimization could estimate this number as part of the process, in addition to estimating parameters of the incubation distribution. The best fit was found to be a three stage distribution similar to the gamma and one which accorded with the high-low-high U-shaped infectiousness profile which was receiving support from a number of sources.

In the marketing domain Graham and Ariza [8] carried out an optimization on a system dynamics model which was designed to shed light on the allocations to make from a marketing budget in a high-tech client firm. Assuming the budget was fixed, the task was to optimize the allocations across more than 90 'buckets' – combinations of product lines, marketing channels and types of marketing. However, these were not discrete: advertising on one product line might have crossover effects on another and the impacts could propagate over a period of time. One major conclusion for this firm was that the advertising allocation should be increased markedly. In general intuitive allocations were shown to fall short of the ideal: they were directionally correct but magnitudes fell short often by factors of three or four.

Future Directions

A primary aim must be to see more published work which describes optimization studies carried out on real-world SD applications. There may be frequent use of optimization in consulting assignments but such activities are rarely published. The references herein suggest that, thus far, outside of unpublished work, the number may be three at most. Whilst software requirements may have inhibited use of SD optimization in the past, there are now no computational barriers to its use and it is to be hoped that in future this quite powerful analytical tool in SD will feature in more application studies.

An advance in the methodology itself has been developed by Duggan [7] and this is a promising pointer for the future. Based on genetic algorithms, it is best suited to the class of SD problems that are agent-based and this highlights a slight limitation. Traditional optimization takes the policy equations as given and explores the parameter space to determine an optimal policy. Instead he has offered an approach which searches over both parameter space and policy (strategy). Theoretically there is no limit

to the number of strategies which can be evaluated in this approach, although the user has to define a set in advance of the runs. Under a conventional optimization approach a limited tilt at this is possible using the zero-one parameter method suggested above, although this would restrict the enumerated strategies to two only. Duggan demonstrates the new approach using a classic SD problem: the four agent beer-game. We await its use in a real-world application.

Bibliography

1. Coyle RG (1996) System dynamics modelling: a practical approach. Chapman & Hall, London
2. Dangerfield BC, Roberts CA (1994) Fitting a model of the spread of AIDS to data from five European countries. In: O.R. Work in HIV/AIDS, 2nd edn. Operational Research Society, Birmingham, pp 7–13
3. Dangerfield BC, Roberts CA (1996) An overview of strategy and tactics in system dynamics optimisation. J Oper Res Soc 47(3):405–423
4. Dangerfield BC, Roberts CA (1996) Relating a transmission model of AIDS spread to data: some international comparisons. In: Isham V, Medley G (eds) Models for infectious human diseases: Their structure and relation to data. Cambridge University Press, Cambridge, pp 473–476
5. Dangerfield BC, Roberts CA (1999) Optimisation as a statistical estimation tool: an example in estimating the AIDS treatment-free incubation period distribution. Syst Dyn Rev 15(3):273–291
6. Duggan J (2005) Using multiple objective optimisation to generate policy insights for system dynamics models. In: Proceedings of the international system dynamics conference, Boston. System Dynamics Society. (CD-ROM)
7. Duggan J (2008) Equation-based policy optimisation for agent-oriented system dynamics models. Syst Dyn Rev 24(1):97–118
8. Graham AK, Ariza CA (2003) Dynamic, hard and strategic questions: using optimisation to answer a marketing resource allocation question. Syst Dyn Rev 19(1)27–46
9. Keloharju R (1977) Multi-objective decision models in system dynamics. Dynamica 3(1)3–13 and 3(2)45–55
10. Keloharju R, Wolstenholme EF (1988) The basic concepts of system dynamics optimisation. Syst Pract 1:65–86
11. Keloharju R, Wolstenholme EF (1989) A case study in system dynamics optimisation. J Oper Res Soc 40(3):221–230
12. Meadows DM, Robinson JM (1985) The electronic oracle. Wiley, Chichester (Now available from the System Dynamics Society, Albany NY)
13. Richardson GP, Pugh AL (1981) An introduction to system dynamics modelling with DYNAMO. MIT Press, Cambridge (Now available from Pegasus Communications, Waltham, MA)
14. Wolstenholme EF, Al-Alusi AS (1987) System dynamics and heuristic optimisation in defence analysis, Syst Dyn Rev 3(2):102–115

System Dynamics and Organizational Learning

Kambiz Maani
Chair in Systems Thinking and Practice,
The University of Queensland, Brisbane, Australia

Article Outline

Glossary
Definition of the Subject
Introduction
Systems Thinking and Modeling Methodology
Organizational Learning
Modeling for Organizational Learning
Future Directions
Bibliography

Glossary

Stock In system dynamics stock is a concept representing accumulation and the state of a variable, such as, assets, inventory, capacity, reputation, morale etc. Stock can be measured at any point of time. In mathematical terms, stock is the sum over time (integral) of one or more flows.

Flow Flow or rate represents change or movement in a stock such as, buying assets, building inventories, adding capacity, losing reputation or morale, etc. Flow is measured as "per unit of time" like hiring rate (employees hired per year, production rate (units made per day), or rainfall (inches of rain per month).

Causal loops Causal loops (model) are visual maps that connect a group of variables with known or hypothesized cause and effect relationships. A causal loop can be open or closed. Causal loops can be used for complex problem solving/decision making, consensus building, conflict resolution, priority setting and group learning.

Feedback In a cause and effect chain (system), feedback is a signal from the effect/s to cause/s as to its/their influence on downstream effect/s. Feedback can be information, decision or action. For example, if X causes or changes Y, Y in turn could influence or change X directly or through other intervening variables. This creates a *closed* "causal loop" with either a positive or amplifying feedback (Reinforcing – R) or a negative feedback with damping, counteracting or (Balancing – B) effect.

Delay Cause and effect relationships are often not close in time or space. The lapse time between a cause and its effect is called a systems delay or simply delay. Because

some delays in physical, natural and social systems are rather long they mask the underlying or earlier causes when effects become evident. This provides confusion and unintended consequences, especially in social systems, such as economics, education, immigration, judicial systems, etc.

Reference Mode Reference mode is the actual/observed pattern of a key variable of interest to decision makers or policy analysts. It represents the actual behavior of a variable over time which is used to compare with the simulated pattern of the same variable generated by a simulation model to validate the accuracy of the model.

Simulation A computer tool and methodology for modeling complex situations and challenging problems where mathematical tools fail to operate.

Microworld Microworlds are simulation models of real systems such as a firm, a hospital, a market, or a production system. They provide a "virtual" world where decision makers can test and experiment their policies and strategies in a laboratory environment before implementation. Microworlds are constructed using system dynamic software with user friendly interfaces.

Leverage Leverage refers to decisions and actions for change and intervention which have the highest likelihood of lasting and sustainable outcomes. Leverage decisions are best reached by open discussion after the group develops a deep understanding of system dynamics through a causal loop or stock & flow modeling process.

Systems thinking Systems thinking is a paradigm for viewing reality based on the primacy of the whole and relationships. It is one of the key capabilities (disciplines) for organizational learning [30]. Systems Thinking consists of a series of conceptual and modeling tools such as behavior over time, causal loop diagrams and systems archetypes. These tools reveal cause and effect dynamics over time and assist understanding of complex, non-linear, and counter-intuitive behaviors in all systems – physical, natural and social.

Definition of the Subject

System dynamics (SD) is "a methodology for studying and managing complex feedback systems… While the word system has been applied to all sorts of situations, feedback is the differentiating descriptor here. Feedback refers to the situation of *X* affecting *Y* and *Y* in turn affecting *X* perhaps through a chain of causes and effects… Only the study of the whole system as a feedback system will lead to correct results." [36]

Sterman ([35], p 4) defines System Dynamics as "a method to enhance learning in complex systems". "System dynamics is fundamentally interdisciplinary… It is grounded in the theory of nonlinear dynamics and feedback control developed in mathematics, physics, and engineering. Because we apply these tools to the behavior of human as well as physical and technical systems, system dynamics draws on cognitive and social psychology, economics, and other social sciences."

Wolstenholme's [40] offers the following description for system dynamics and its scope:

> A rigorous way to help thinking, visualizing, sharing, and communication of the future evolution of complex organizations and issues over time; for the purpose of solving problems and creating more robust designs, which minimize the likelihood of unpleasant surprises and unintended consequences; by creating operational maps and simulation models which externalize mental models and capture the interrelationships of physical and behavioral processes, organizational boundaries, policies, information feedback and time delays; and by using these architectures to test the holistic outcomes of alternative plans and ideas; within a framework which respects and fosters the needs and values of awareness, openness, responsibility and equality of individuals and teams.

Organizational Learning

Organizational learning is the ability of organizations to enhance their collective capacity to learn and to act, harmoniously. According to Senge [30] "Real learning gets to the heart of what it means to be human. Through learning we re-create ourselves. Through learning we become able to do something we never were able to do. Through learning we re-perceive the world and our relationship to it. Through learning we extend our capacity to create, to be part of the generative process of life. There is within each of us a deep hunger for this type of learning." Organizational learning extends this learning to the organization and its members.

Introduction

History of System Dynamics

(This section is due to US Department of Energy website) "System dynamics was created during the mid-1950s by Professor Jay W. Forrester of the Massachusetts Institute of Technology. Forrester arrived at MIT in 1939 for graduate study in electrical engineering. His first research assistantship put him under the tutelage of Professor Gordon

Brown, the founder of MIT's Servomechanism Laboratory. Members of the MIT Servomechanism Laboratory, at the time, conducted pioneering research in feedback control mechanisms for military equipment. Forrester's work for the Laboratory included traveling to the Pacific Theatre during World War II to repair a hydraulically controlled radar system installed aboard the aircraft carrier Lexington. The Lexington was torpedoed while Forrester was on board, but not sunk.

At the end of World War II, Jay Forrester turned his attention to the creation of an aircraft flight simulator for the US Navy. The design of the simulator was cast around the idea, untested at the time, of a digital computer. As the brainstorming surrounding the digital aircraft simulator proceeded, however, it became apparent that a better application of the emerging technology was the testing of computerized combat information systems. In 1947, the MIT Digital Computer Laboratory was founded and placed under the direction of Jay Forrester. The Laboratory's first task was the creation of WHIRLWIND I, MIT's first general-purpose digital computer, and an environment for testing whether digital computers could be effectively used for the control of combat information systems. As part of the WHIRLWIND I project, Forrester invented and patented coincident-current random-access magnetic computer memory. This became the industry standard for computer memory for approximately twenty years. The WHIRLWIND I project also motivated Forrester to create the technology that first facilitated the practical digital control of machine tools.

After the WHIRLWIND I project, Forrester agreed to lead a division of MIT's Lincoln Laboratory in its efforts to create computers for the North American SAGE (Semi-Automatic Ground Environment) air defense system. The computers created by Forrester's team during the SAGE project were installed in the late 1950s, remained in service for approximately twenty-five years, and had a remarkable "up time" of 99.8%.

Forrester's seminal book Industrial Dynamics [11] "is still a significant statement of philosophy and methodology in the field. Since its publication, the span of applications has grown extensively and now encompasses work in

- corporate planning and policy design
- public management and policy
- biological and medical modeling
- energy and the environment
- theory development in the natural and social sciences
- dynamic decision making
- complex nonlinear dynamics" [36]

Systems Thinking and Modeling Methodology

System Dynamics is one of the five phases of systems thinking and modeling intervention methodology [6,21]. These distinct but related phases are as follows:

1. Problem structuring;
2. Causal loop modeling;
3. System dynamics modeling;
4. Scenario planning and modeling;
5. Implementation and organizational learning (learning lab).

These phases follow a process, each involving a number of steps, as outlined in Table 1. This process does not require all phases to be undertaken, nor does each phase require all the steps listed. Which phases and steps are included in a particular project or intervention depends on the issues or problems that have generated the systems enquiry and the degree of effort that the organization is prepared to commit to.

System Dynamics Modeling

This phase follows the causal modeling phase. Although it is possible to go into this phase directly after problem structuring, performing the causal modeling phase first will enhance the conceptual rigor and learning power of the systems approach. The completeness and wider insights of systems thinking is generally absent from other simulation modeling approaches, where causal modeling does not play a part. The following steps are generally followed in the system dynamics modeling phase.

1. Develop a high-level map or systems diagram showing the main sectors of a potential simulation model, or a 'rich picture' of the main variables and issues involved in the system of interest.
2. Define variable types (e. g. stocks, flows, converters, etc.) and construct stock flow diagrams for different sectors of the model.
3. Collect detailed, relevant data including media reports, historical and statistical records, policy documents, previous studies, and stakeholder interviews.
4. Construct a computer simulation model based on the causal loop diagrams or stock-flow diagrams. Identify the initial values for the stocks (levels), parameter values for the relationships, and the structural relationships between the variables using constants, graphical relationships and mathematical functions where appropriate. This stage involves using specialized computer packages like STELLA, *ithink*, VENSIM, POWERSIM, DYSMAP, COSMIC and Consideo.

System Dynamics and Organizational Learning, Table 1
The five phase process of systems thinking and modeling (Source: [6])

Phases	Steps	
1	Problem structuring	Identify problems or issues of concern to management, Collect preliminary information and data
2	Causal loop modeling	Identify main variables, Prepare behavior over time graphs (reference mode), Develop causal loop diagram (influence diagram), Analyze loop behavior over time and identify loop types, Identify system archetypes, Identify key leverage points, Develop intervention strategies
3	**System dynamic modeling**	Develop a systems map or rich picture, Define variable types and construct stock-flow diagrams, Collect detailed information and data, Develop a simulation model, Simulate steady -state/stability conditions, Reproduce reference mode behavior (base case), Validate the model, Perform sensitivity analysis, Design and analyze policies, Develop and test strategies
4	Scenario planning and modeling	Plan general scope of scenarios, Identify key drivers of change and keynote uncertainties, Construct forced and learning scenarios, Simulate scenarios with the model, Evaluate robustness of the policies and strategies
5	Implementation and organizational learning	Prepare a report and presentation to management team, Communicate results and insights of proposed intervention to stakeholders, Develop a microworld and learning lab based on the simulation model, Use learning lab to examine mental models and facilitate

5. Simulate the model over time. Select the initial value for the beginning of the simulation run, specify the unit of time for the simulation (e. g. hour, day, week, month, year, etc.). Select the simulation interval (DT) (e. g. 0.25, 0.5, 1.0) and the time horizon for the simulation run (i. e. the length of the simulation). Simulate model stability by generating steady state conditions.
6. Produce graphical and tabular output for the base case of the model. This can be produced using any of the computer packages mentioned above. Compare model behavior with historical trends or hypothesized reference modes (behavior over time charts).
7. Verify model equations, parameters and boundaries, and validate the model's behavior over time. Carefully inspect the graphical and tabular output generated by the model.
8. Perform sensitivity tests to gauge the sensitivity of model parameters and initial values. Identify areas of greatest improvement (key leverage points) in the system.
9. Design and test policies with the model to address the issues of concern to management and to look for system improvement.
10. Develop and test strategies (i. e. combinations of functional policies, for example operations, marketing, finance, human resources, etc.).

Organizational Learning

(This section is adapted from [21])

Peter Senge, who popularized the concept through his seminal book: The Fifth Discipline [30], describes a learning organization as one 'which is continually expanding its ability to create its future'. He identifies five core capabilities (disciplines) of the learning organization that are derived from three "higher orientations": creative orientation; generative conversation; and systems perspective. "The reality each of us sees and understands depend on what we believe is there. By learning the principles of the five disciplines, teams begin to understand how they can think and inquire that reality, so that they can collaborate in discussions and in working together create the results that matter [to them]."

As Fig. 1 shows the learning organization capabilities are dynamically interrelated, and collectively they lead to organizational learning.

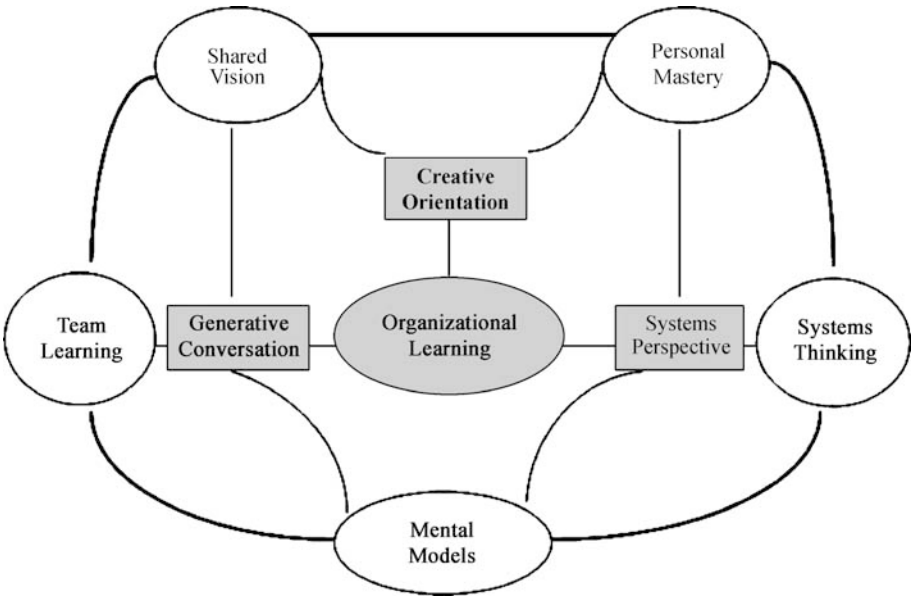

System Dynamics and Organizational Learning, Figure 1
The core capabilities of a learning organization (Source: [19])

Senge maintains that *Creative orientation* is the source of a genuine desire to excel. It is the source of an intrinsic motivation and drive to achieve. It relinquishes personal gains in favor of the common good. *Generative conversation* refers to a deep and meaningful dialog to create unity of thought and action. *Systems perspective* is the ability to see things holistically by understanding the interconnectedness of the parts. The foregoing elements give rise to the five core capabilities of learning organizations, namely: personal mastery; shared vision; mental models; team learning and dialog; and systems thinking. These five disciplines are described below. Figure 1 below shows the core capabilities and their relationships.

Personal Mastery

Senge [30] describes that personal mastery is the cornerstone and 'spiritual' foundation of the learning organization. It is born out of a creative orientation and systemic perspective. Personal mastery instils a genuine desire to do well and to serve a noble purpose. People exhibiting high levels of personal mastery focus "on the desired result itself, not the process or the means they assume necessary to achieve that result" [30]. These people can "successfully focus on their ultimate intrinsic desires, not on secondary goals. This is a cornerstone of Personal Mastery". Personal mastery also requires a commitment to truth, which means to continually challenge "theories of why things are the way they are". Without committing to the truth, people all too quickly revert to old communication routines which can distort reality and prevent them from knowing where they really stand.

Shared Vision

It is commonly assumed that in contemporary organizations senior management can develop a vision which employees will follow with genuine commitment. This is a fallacy. Simply promoting a 'vision statement' could result in a sense of apathy, complacency and resentment. Instead, there needs to be a genuine endeavor to understand what people will commit to. The overriding vision of the group must build on the personal visions of its members. Shared vision should align diverse views and feelings into a unified focus.

This is emphasized by Arie de Geus [9] when he describes what makes a truly extraordinary organization. "The feeling of belonging to an organization and identifying with its achievements is often dismissed as soft. But case histories repeatedly show that a sense of community is essential for long term survival". For example, when Apple Corporation challenged IBM, it was in its 'adolescent' years, characterized by creativity, confidence and even defiance. This is similar to the spirit in Team New Zealand when it competed against the bigger-budget syndicates! Within these organizations there is a real passion for the outcome; a common vision for success [19].

Creating a shared vision is the most fundamental job of a leader [26]. By creating a vision, the leader provides a vehicle for people to develop commitment, a common goal around which people can rally, and a way for people to *feel* successful. The leader must appeal to people's emotions if they are to be energized towards achieving the goal. Emotional acceptance of, and belief in, a vision is far more powerful in energizing team members than is intellectual recognition that the vision is simply a 'good idea'. One of the most powerful ways of communicating a vision is through a leader's personal example and actions, demonstrating behavior that symbolizes and furthers that vision.

Mental Model and Leadership

Mental models reflect beliefs, assumptions and feelings that shape one's world views and actions. They are formed through family, education, professional and social learning based, on the most part, on cultural and social norms. Mental models, however, can be altered and aligned.

Organizations are often constrained by deep-seated belief systems, resulting in preconceived ideas on how things ought to perform. Goodstein and Burke ([14] p. 10), pioneers in the field of social psychology of organizations, observed that 'the first step in any change process is to unfreeze the present patterns of behavior as a way of managing resistance to change'. The leader has a pivotal role in dismantling negative mental models and shaping new ones.

In order to get people to engage in open discussions of issues that affect the organization, a leader must appeal to their emotions and must get beyond the superficial level of communication. In the 1970s Shell Oil undertook major changes in its leadership approach and communications style. According to a manager at Shell, "When I tried to talk personally about an issue rather than say 'here's the answer', it was powerful. It caused me to engage in dialog with others that resulted in mutual learning on all sides" ([7] p. 71).

The leader is a 'designer', and part of that role is designing the governing ideas of purpose and core values by which people will live [30,31]. In this role, the leader must propose and model the manner in which the group has to operate internally. This provides ample opportunities for leaders to examine their deeply held assumptions about the task, the means to accomplish it, the uniqueness of the people and the kinds of relationship that should be fostered among the people. Only after people have observed and *experienced* the organizational values in practice would these values become the basis for prolonged group behavior. These values should be manifested first and should be most visible in the leader's own behavior.

Leadership, especially in knowledge-based organizations, must be distributed and shared to a far greater extent than it was in the past. For example, in the Chicago Bulls basketball team, Michael Jordan changed his role: it became not only that of an individually brilliant player but *also* that of a leader whose job it was to raise the level of play of other team members. After this transition, the Bulls began their record run of championship seasons [7].

Team Learning and Dialog

The word 'dialog' comes from the Greek words *dia* and *logos*. It implies that when people engage in dialog, the meaning *moves through* them – Thus, it enables them to 'see through words' [16]. Dialog is an essential requirement for organizational learning. It results from generative conversation, shared vision, and transparent mental models. Dialog creates a deep sense of listening and suspending one's own views. Feedback is an integral aspect of dialog.

Communication routines in organizations are generally anti-learning and promote mediocrity. They include 'defensive routines' [2] – statement that can stifle dialog and innovative thinking. Exposing and unlearning such routines, and understanding the powerful detrimental impact they have on learning, are serious challenges many organizations face if they are to create effective learning environments.

Many leaders are charismatic and are highly eloquent when it comes to presenting their ideas; that's often why they get to the top of the organization. However, many appear to lack the ability to extract the very best from employees in a non-threatening manner. Without this ability, leaders may miss many good ideas, or might act on many bad ones.

In a group context, encouragement from the leader and mutual encouragement among group members is essential. Furthermore, personal differences must be put aside in order for effective dialog to ensue.

How Organizations Learn

(This section is edited from Wikipedia: http://en.wikipedia.org/wiki/Organizational_learning)
"Argyris and Schon were the first to propose concepts and models that facilitate organizational learning, the following literatures have followed in the tradition of their work:

- March and Olsen [23] attempt to link up individual and organizational learning. In their model, individual be-

liefs lead to individual action, which in turn may lead to an organizational action and a response from the environment which may induce improved individual beliefs and the cycle then repeats over and over. Learning occurs as better beliefs produce better actions.

- Argyris and Schon [3] distinguish between single-loop and double-loop learning, related to Gregory Bateson's concepts of first and second order learning. In single-loop learning, individuals, groups, or organizations modify their actions according to the difference between expected and obtained outcomes. In double-loop learning, the entities (individuals, groups or organization) question the values, assumptions and policies that led to the actions in the first place; if they are able to view and modify those, then second-order or double-loop learning has taken place. Double loop learning is the learning about single-loop learning.
- Kim [17], as well, in an article titled "The link between individual and organizational learning", integrates Argyris, March and Olsen and another model by Kofman into a single comprehensive model; Further, he analyzes all the possible breakdowns in the information flows in the model, leading to failures in organizational learning; For instance, what happens if an individual action is rejected by the organization for political or other reasons and therefore no organizational action takes place?
- Nonaka and Takeuchi [27] developed a four stage spiral model of organizational learning. They started by differentiating Polanyi's concept of "tacit knowledge" from "explicit knowledge" and describe a process of alternating between the two. Tacit knowledge is personal, context specific, subjective knowledge, whereas explicit knowledge is codified, systematic, formal, and easy to communicate. The tacit knowledge of key personnel within the organization can be made explicit, codified in manuals, and incorporated into new products and processes. This process they called "externalization". The reverse process (from explicit to implicit) they call "internalization" because it involves employees internalizing an organization's formal rules, procedures, and other forms of explicit knowledge. They also use the term "socialization" to denote the sharing of tacit knowledge, and the term "combination" to denote the dissemination of codified knowledge. According to this model, knowledge creation and organizational learning take a path of socialization, externalization, combination, internalization, socialization, externalization, combination... etc. in an infinite spiral.
- Flood [10] discusses the concept of organizational learning from Peter Senge and the origins of the theory from Argyris and Schon. The author aims to "re-think" Senge's *The Fifth Discipline* through systems theory. The author develops the concepts by integrating them with key theorists such as Bertalanffy, Churchman, Beer, Checkland and Ackoff. Conceptualizing organizational learning in terms of structure, process, meaning, ideology and knowledge, the author provides insights into Senge within the context of the philosophy of science and the way in which systems theorists were influenced by twentieth-century advances from the classical assumptions of science.
- Nick Bontis et al. [4] empirically tested a model of organizational learning that encompassed both stocks and flows of knowledge across three levels of analysis: individual, team and organization. Results showed a negative and statistically significant relationship between the misalignment of stocks and flows and organizational performance.
- Imants [15] provides theory development for organizational learning in schools within the context of teachers' professional communities as learning communities, which is compared and contrasted to teaching communities of practice. Detailed with an analysis of the paradoxes for organizational learning in schools, two mechanisms for professional development and organizational learning, (1) steering information about teaching and learning and (2) encouraging interaction among teachers and workers, are defined as critical for effective organizational learning.
- Common [8] discusses the concept of organizational learning in a political environment to improve public policy-making. The author details the initial uncontroversial reception of organizational learning in the public sector and the development of the concept with the learning organization. Definitional problems in applying the concept to public policy are addressed, noting research in UK local government that concludes on the obstacles for organizational learning in the public sector: (1) overemphasis of the individual, (2) resistance to change and politics, (3) social learning is self-limiting, i. e. individualism, and (4) political "blame culture". The concepts of *policy learning* and *policy transfer* are then defined with detail on the conditions for realizing organizational learning in the public sector."

Modeling for Organizational Learning

In general, the *process* of model building can be an effective conduit for collective learning. System Dynamics modeling, in particular, can be used to enhance organizational

learning [35] through rapid feedback and experimentation and its facility to test assumptions and mental models. As we have discussed, dealing effectively with mental models is one of the core competencies for organizational learning

Ackoff [1] likens complex problems to "messes". "Messy problems are defined as situations in which there are large differences of opinion about the problem or even on the question of whether there is a problem. Messy situations make it difficult for a management team to reach agreement. System Dynamics modeling with groups known as Group Model Building (GMB) is a powerful tool for dealing with these. SD and GMB are especially effective in dealing with semi-structured and ill-structured decision situations."

GMB offers an opportunity to align and share piecemeal mental models and create the possibility of assimilating and integrating partial mental models into a holistic system description [38,39]. GMB and SD can help uncover 'illusions' that may occur due to the fact that the definition of a problem may be a socially constructed phenomenon that has not been put to test ([18] p. 84).

Learning Laboratory

(This section is adapted from [21], Chapter 6)
Learning laboratory is a setting as well as a process in which a group can learn together. The purpose of the learning lab is to enable managers to test their long held assumptions and to experiment and 'see' the consequences of their actions, policies and strategies. This often results in finding inconsistencies and the discovery of *unintended* consequences of actions and decisions, *before* they are implemented. System Dynamics models known as Microworlds or Management Flight Simulators (MFS) are the 'engine' behind the learning lab. "Just as an airline uses flight simulators to help pilots learn, system dynamics is, partly, a method for developing management flight simulators, often computer simulation models, to help us learn about dynamic complexity, understand the sources of policy resistance, and design more effective policies." ([35], p. 4)

A learning lab is distinct from so-called management games. In management games, the players are required to compete – design the 'best' strategy and 'beat' other players or teams. The competitive nature of management games often encourages aggressive and individualistic behavior with scant regard for group learning and gaining deep insights. The learning lab, in contrast, aims to enhance *learning*: To test individual and group mental models and to provide deeper understanding and insights into why systems behave the way they do. This will help the participants to test their theories and discover inconsistencies and 'blind spots' in policies and strategies *before* they are implemented.

A significant benefit of the learning lab stems from the process in which participants examine, reveal and test their mental models and those of their organization. The learning lab can also help participants

- To align strategic thinking with operational decisions;
- To connect short-term and long-term measures;
- To facilitate integration within and outside the organization;
- To undertake experimentation and learning;
- To balance competition with collaboration.

Managerial Practice Field

Team and teamwork are parts of the lexicons of numerous organizations today. Company after company has reorganized work around a variety of team concepts. From factories to hospitals, *titles* like 'manager' and 'supervisor' have been replaced by *roles* such as 'facilitator' and 'team leader'. Despite this level of attention to team and teamwork the expected benefits have been marginal at best.

But when we examine real teams, such as sporting teams, orchestras or ballet companies more closely, they all share one key characteristic. That is they *practice* a lot more than they 'perform'. Practice involves allowing time and space to experiment with new ways, try different approaches and most importantly, make mistakes without the fear of failure. In fact, making mistakes is indispensable to learning. One cannot learn from doing things right all the time! Yet a great deal of organizational energy and attention is devoted to the prevention and masking of mistakes.

But, what is the *practice field* for management teams? The fact is that the practice field is, by and large, absent from the managerial world. In other words, there is no time and no space for management to 'practice' in the true sense of the word – to experiment, make mistakes and learn together. In this era of restructuring and downsizing, lack of time is the greatest impediment to managerial and organizational learning. As a recent advertisement by IBM reads, "Innovative Thinking! We don't even have time for bad thinking". The pace in the modern work environment is so unrelenting that there is virtually no room for managers to slow down, to practice, to reflect and learn. The consequence of this lack of practice and learning space is grave, in that most organizations only achieve a small fraction of their potential – about 5%, according to Jay Forrester, the father of System Dynamics [13].

In order to fill this gap, the concept of learning laboratory has been developed to provide practice fields for managers. The learning lab allows learning to become an integral part of managerial work and helps learning to become institutionalized [17].

Aligning Mental Models Through the Learning Laboratory

Mental models are formed throughout one's life. Family, school, culture, religion, profession and social norms play important roles in this formation. Therefore, modifying one's mental model is not a small matter. The most effective way to check one's mental models is to *experience* alternative realities at first hand and see their implications with a new 'lens' [5].

There are rarely any opportunities in the course of a manager's daily work for him/her to engage in lengthy, drawn-out experimentation. Learning in a 'laboratory' setting is a viable and powerful alternative. Fortunately, advanced computers and sophisticated system dynamics software have enabled the creation of managerial learning labs where managers can experiment, test their theories and learn rapidly. Thus, learning labs can play a significant role in clarifying and changing mental models. Learning lab deals with mental models at three levels [33], as described below.

- *Mapping* mental models. This step begins at the conceptualization phase. Here, the learning lab participants articulate and clarify their assumptions, views, opinions, and biases regarding the issue at hand.
- *Challenging* mental models. The participants identify and discuss inconsistencies and contradictions in their assumptions. This step will begin at the conceptualizations phase and will continue to the experimentation phase.
- *Improving* mental models. Having conducted experimentation and testing, the participants reflect on the outcomes. This may cause them to alter, adjust, improve and harmonize their mental models.

The laboratory setting provides a neutral and 'safe' space for the participants to create a shared understanding of complex and endemic issues. The following characteristics of the learning lab provide a powerful catalyst for alignment of divergent mental models in the organization.

- The laboratory environment is neutral and non-threatening. The emphasis is on learning and theory building (what we *don't* know), not on winning or display of knowledge.
- Lack of hierarchy. Managers and staff are equal in this environment. The traditional hierarchy is minimized in the laboratory setting.
- The response time is fast. Hence, the feedback cycle is short, which leads to rapid learning.
- There is no cost or 'loss of face' attached to failure. Hence, it is safe to make mistakes. In fact, mistakes provide opportunities for learning.
- People can see the consequences of their actions first hand. No one attempts to convince or teach anyone else or force his or her preconceived views on others. People learn by themselves and through group interactions.

Implications for Management

The practice field and the learning lab concepts offer fresh and challenging implications for managers and their role. They suggest that a leader/manager should think as a *scientist*, be open to and welcome hard questions, experiment with new ideas, and be prepared to be *wrong*. This requires managers to learn systems thinking skills and use them not just for 'solving' problems but as powerful tools for communication, team building and organizational learning. This means that an effective leader should be the 'designer' of the ship and not its captain [31]. Once they have designed a new structure, strategy, policy or procedure then the managers/leaders should allow (i. e. create a practice field for) the staff to experience the new design, and experiment with it and learn for themselves – the desired outcome is shared understanding leading to alignment of thoughts and actions. This is the essence of organizational learning.

Future Directions

Agent-Based Modeling (ABM)

Agent based modeling (ABM) is an emerging modeling technology which draws its theories and techniques from complexity science [29]. While System Dynamics and Agent-Based Modeling (ABM) use different modeling philosophies and approaches, they can be used complementarily and synergistically.

System Dynamics focuses on modeling structures (i. e. relationships, policies, strategies) that underlie behavior of systems. This may be viewed as a weakness of system dynamics approach in that behavior is assumed to be solely a function of structure (model relationships defined a priori). In contrast, in ABM, organizations are modeled as a system of semi-autonomous decision-making elements – purposeful individuals called *agents*. Each agent individually assesses its situation and makes de-

cisions based upon value hierarchies representing goals, preferences, and standards for behavior. Thus, macro-behavior is not modeled separately but *emerges* from the micro-decisions of individual agents. In other words, in agent based modeling; "emergent" behavior is expected as a result of agents' interactions. This is a key difference between the two approaches.

While system dynamics acknowledges the critical role of individual and organizational mental models (e. g., motivations, values, norms, biases, etc.) it does not explicitly model them. SD utilizes factual data or "cold knowledge" and does not take into account decision makers 'mood'. In contrast, ABM attempts to capture "warm knowledge", representing emotional and human context of decision-making.

Recent advances in video game technology allow the development of multi-agent, artificial 'society' simulators with capabilities for modeling physiology, stress and emotion in decision-making [34]. At the simplest level, an agent-based model consists of a system of agents and their relationships. This new approach enables superior understanding of the complexity in organizations and their relevant business environments. This in turn provides an opportunity for new sophistications in game-play that enhances decision-making. Experience with agent-based modeling shows that even a simple agent-based model can exhibit complex behavior patterns and provide valuable information about the dynamics of the real world system that emulates them.

Despite their differences, SD and ABM can be used in a complementary fashion. Both ABM and SD are powerful tools for transforming information into knowledge and understanding leading to individual and group learning. However, the transition from knowledge to understanding may not be immediate or transparent. This requires a deep shift in mental models through experimentation and group learning.

Systems Thinking and Sustainability

Systems Thinking has a natural affinity with sustainability modeling and management. Sustainability issues are complex; cut across several disciplines; involve multiple stakeholders and require a long term integrated approach. Thus, the systems paradigm and tools have direct and powerful applications in sustainability issues and management.

The applications of system dynamics in sustainability go back to the early 1970s with Jay Forrester's "World2 and World3 analyzes against 30 years of history", followed by "World Dynamics" and "Limits to Growth" [24] and "Beyond the Limits" [25]. "The politics of the environment has also evolved dramatically since 1970. Public awareness of the reality of the environmental challenge has risen; Ministries of Environment have become commonplace" ([28] p. 220). As an example today concern over carbon emissions has already become an international currency. As a result, sustainability has brought a fresh challenge for governments, business and industry, scientists, farmers and all the citizens of the world collectively to find systemic solutions that are mutually and globally agreeable. Systems Thinking and System dynamics can make real and valuable contributions to addressing this challenge.

Bibliography

Primary Literature

1. Ackoff RA (1999) Re-creating the corporation – A design of organizations for the 21st century. Oxford University Press, Oxford
2. Argyris C (1992) The next challenge for TQM: Overcoming organisational defences. J Qual Particip 15:26–29
3. Argyris C, Schon D (1978) Organizational learning: A theory of action perspective. Addison-Wesley, Reading
4. Bontis N, Crossan M, Hulland J (2002) Managing an organizational learning system by aligning stocks and flows. J Manag Stud 39(4):437–469
5. Brown JS (1991) Research that reinvents the corporation. Harv Bus Rev 68:102–111
6. Cavana R, Maani K (2004) A methodological framework for integrating systems thinking and system dynamics. In: System Dynamics Society Proceedings. Oxford
7. Cohen E, Tichy N (1997) How leaders develop leaders. Training and Development, May
8. Common R (2004) Organisational learning in a political environment: Improving policy-making in UK government. Policy Stud 25(1):35–49
9. De Geus A (1997) The living company. Harv Bus Rev 75(2):51–59
10. Flood RL (1999) Rethinking the fifth discipline: Learning within the unknowable. Routledge, London
11. Forrester JW (1961) Industrial dynamics. Productivity Press, Cambridge
12. Forrester JW (1971) World dynamics. Wright-Allen (Subsequently re-published by Productivity Press, and Pegasus Communications)
13. Forrester JW (1994) Building a foundation for tomorrow's organizations. In: Systems thinking in action video collection, vol 1. Pegasus Communications, Cambridge
14. Goodstein L, Burke W (1991) Creating successful organisation change. Organ Dyn 19(4):5–17
15. Imants J (2003) Two basic mechanisms for organizational learning in schools. Europ J Teach Educ 26(3):293–311
16. Isaacs W (1993) Taking flight: Dialogue, collective thinking and organisational learning. Organ Dyn 22(2):24–39
17. Kim DH (1993) The link between individual and organizational learning. Sloan Manag Rev 35(1):37–50

18. Maani K (2002) Consensus building through systems thinking – the case of policy and planning in healthcare. Aust J Inform Syst 9(2):84–93
19. Maani K, Benton C (1999) Rapid team learning. Lessons from team New Zealand's America's cup campaign. Organ Dyn 27(4)
20. Maani K, Cavana R (2007) Systems methodology. Syst Think 18(8):2–7
21. Maani K, Cavana R (2007) Systems thinking, system dynamics – Managing change and complexity, 2nd edn. Prentice Hall, Pearson Education, Auckland
22. Maani K, Pourdehnad J, Sedehi H (2003) Integrating system dynamics and intelligent agent-based modelling – theory and case study. Euro INFORMS, Istanbul
23. March JG, Olsen JP (1975) The uncertainty of the past; Organizational ambiguous learning. Europ J Polit Res 3:147–171
24. Meadows DH, Meadows DL, Randers J, Behren W (1972) The limits to growth. Universe Press, New York
25. Meadows DH, Meadows DL, Randers J (1992) Beyond the limits. Chelsey Green, Post Mills
26. Nadler DA, Tushman ML (1990) Beyond the charismatic leader: Leadership and organisational change. Calif Manag Rev
27. Nonaka I, Takeuchi H (1995) The knowledge creating company. Oxford University Press, New York
28. Randers J (2000) From limits to growth to sustainable development or SD (sustainable development) in a SD (system dynamics) perspective. Syst Dyn Rev 16(3):213–224
29. Rothfeder J (2003) Expert voices: Icosystem's Eric Bonabeau. CIO Insights
30. Senge P (1990) The fifth discipline: The art and practice of the learning organisation. Currency
31. Senge P (1990) The leader's New Work: Building learning organisation's. Sloan Manag Rev:7–23
32. Senge P (1992) Building learning organisation's. J Qual Particip:1–8
33. Senge P, Sterman JD (1991) Systems thinking and organizational learning: Acting locally and thinking globally in the organization of the future. In: Kochan T, Useem M (eds) Transforming organizations. Oxford University Press, Oxford
34. Silverman BG et al (2002) Using human models to improve the realism of synthetic agents. Cogn Sci Q 3
35. Sterman JD (2000) Business dynamics, systems thinking and modeling for a complex world. McGraw-Hill, Irwin
36. System Dynamics Society website http://www.systemdynamics.org/
37. US Department of Energy Introduction to system dynamics, A systems approach to understanding complex policy issues, US Department of Energy. http://www.systemdynamics.org/DL-IntroSysDyn/inside.htm
38. Vennix JAM (1995) Building consensus in strategic decision-making: System Dynamics As A Support System. Group Decis Negot 4(4):335–355
39. Vennix JAM (1996) Group model-building: Facilitating team learning using system dynamics. Wiley, Chichester, chapt 5
40. Wolstenholme E (1997) System dynamics in the elevator (SD1163), e-mail communication, 24 Oct 1997 system-dynamics@world.std.com

Books and Reviews

(This section is due to M. Anjali Sastry and John D. Sterman, "An Annotated Survey of the Essential System Dynamics Literature System Dynamics Group", Sloan School of Management, MIT)

Industrial and Economic Dynamics: The Foundations

Forrester JW (1961) Industrial dynamics. Productivity Press, Cambridge (Presents dynamic analysis of a business problem through a model of a production-distribution system that shows oscillatory behavior. Policies to improve system performance are discussed, and numerous policy experiments are demonstrated. Includes full equation listing.)

Forrester JW (1968) Principles of systems. Productivity Press, Cambridge (System structure and behavior are differentiated, with examples showing how structure determines behavior. Rates and levels are described. Inventory model shows effects of delivery delay and resulting production cycles.)

Forrester JW (1975) Collected papers of Jay W. Forrester. Productivity Press, Cambridge (Includes many seminal papers, such as Industrial Dynamics: A Major Breakthrough for Decision Makers; Common Foundations Underlying Engineering and Management; A New Corporate Design; Market Growth as Influenced by Capital Investment; and Counterintuitive Behavior of Social Systems.)

Forrester JW (1989) The beginnings of system dynamics (Working Paper No. D-4165). System Dynamics Group, Sloan School of Management, MIT, Cambridge (A personal history beginning on the high plains of western Nebraska. Describes the early projects that shaped the field.)

Mass NJ (1975) Economic cycles: An analysis of underlying causes. Productivity Press, Cambridge (Shows how production scheduling and work force management policies generate the 3–5 year business cycle. Economic cycles, in turn, are caused by capital investment policies that fail to account for delays in acquiring long-lead time plant and equipment.)

Meadows DL (1970) Dynamics of commodity production cycles. Productivity Press, Cambridge (Develops a simple generic model of commodity supply and demand with explicit production capacity and delays, prices and markets. Applies the model to hogs, chicken and cattle.)

Urban and Public Policy Dynamics

Alfeld LE, Graham AK (1976) Introduction to urban dynamics. Productivity Press, Cambridge (A very readable introductory text. Uses the urban system as an example to teach general points about modeling methods, formulation and analysis.)

Forrester JW (1969) Urban dynamics. Productivity Press, Cambridge (Seminal model of urban growth and decay, controversial then and vindicated now. Chapter 6 describes general characteristics of complex systems such as compensating feedback and shifting the burden to the intervener.)

Mass NJ (ed) (1974) Readings in urban dynamics, vol I. Productivity Press, Cambridge (Extensions, modification, and responses to criticisms of the Urban Dynamics model.)

Schroeder WW, III, Sweeney RE, Alfeld LE (eds) (1975) Readings in urban dynamics, vol II. Productivity Press, Cambridge (Further extends and explores the Urban Dynamics model.)

Limits to Growth and Other Global Models

Forrester JW (1973) World Dynamics, 2nd edn. Productivity Press, Cambridge (The first global model, on which Limits to Growth was based. The extreme simplicity of the model allowed it to be presented to a wide audience.)

Meadows DL, Meadows DH (eds) (1974) Toward global equilibrium: Collected papers. Productivity Press, Cambridge (Describes and explores, through system dynamics models, poli-

cies for sustainability designed to avoid the collapse shown in the 'business as usual' WORLD3 scenarios.)

Meadows DH, Meadows DL, Randers J, Behrens WW III (1972) The limits to growth: A report for the club of Rome's project on the predicament of mankind. Universe Books, New York (Classic controversial study of the human future. Nontechnical presentation of structure, assumptions, and results of the WORLD3 model. Concluded that present policies were unsustainable; shows how alternate policies could stabilize population at a high standard of living.)

Meadows DL, Behrens WW III, Meadows DH, Naill RF, Randers J, Zahn EKO (1974) Dynamics of growth in a finite world. Productivity Press, Cambridge (Full documentation and data for the WORLD3 model used in the Limits to Growth. Describes the structure and assumptions; includes all data needed for complete replication of all runs in the popular book. Formulations described here may be useful to all system dynamics modelers.)

Meadows D, Richardson J, Bruckmann G (1982) Groping in the dark. Wiley, New York (Describes a range of global models built under different approaches and discusses the strengths, weaknesses, and implications of each. Presented in an engaging, personal style.)

Meadows DH, Meadows DL Randers J (1992) Beyond the limits: Confronting global collapse, envisioning a sustainable future. Chelsea Green, Post Mills (Follows up on Limits to Growth. Shows that many problems described in 1972 have worsened, as predicted by the model. Argues for a shift in values necessary to create a sustainable and equitable future.)

SD for Management: Firm and Market Models

Coyle RG (1977) Management System Dynamics. Wiley, New York (Text emphasizing managerial modeling, with a focus on operations and examples including discrete elements.)

Hall RI (1976) A system pathology of an organization: The rise and fall of the Old Saturday Evening Post. Adm Sci Q 21(2):185–211 (A case-study using a system dynamics model to explain how failure to understand the feedbacks among policies governing ad rates, ad and editorial pages, marketing, and pricing lead to the failure of the Post just as circulation reached an all-time high.)

Lyneis JM (1980) Corporate planning and policy design. Productivity Press, Cambridge (Begins with a simple model of inventory management in a manufacturing firm and gradually extends the model to one of the entire firm.)

Merten PP (1991) Loop-based strategic decision support systems. Strat Manag J 12:371–382 (Describes a model of a multinational firm establishing new markets in less-developed countries. Captures qualitative shifts in firm structure and organization endogenously as the firm evolves.)

Morecroft JDW (1984) Strategy support models. Strat Manag J 5(3):215–229 (Describes the use of models as participants in the ongoing dialogue among managers regarding strategy formation and evaluation. Emphasizes the processes for model development and use that enhance the utility of modeling in design of high-level corporate strategy.)

Morecroft JDW, Lane DC, Viita PS (1991) Modelling growth strategy in a biotechnology startup firm. Syst Dyn Rev 7(2):93–116 (Describes a case-study of a start-up in which system dynamics modeling helps to define a desirable growth strategy for the firm. The integrated model generated strategies that allowed different parts of the firm to choose consistent approaches.)

Roberts EB (ed) (1978) Managerial applications of system dynamics. Productivity Press, Cambridge (Extensive collection of early corporate models, including history and commentary by practitioners. Covers R&D management, production and operations, human resources, and other applications areas.)

Economic Models

Forrester JW (1989) The system dynamics national model: Macrobehavior from microstructure. In: Milling PM, Zahn EOK (eds) Computer-based management of complex systems: International System Dynamics Conference. Springer, Berlin (Provides an overview of the national modeling project in which both micro- and macro-economic factors are included. Model generates business cycles, inflation, stagflation, the economic long wave, and growth.)

Saeed K (1986) The dynamics of economic growth and political instability in the developing countries. Syst Dyn Rev 2(1):20–35 (Shows how rapid economic development can generate social and political instability through a model that links socio-political factors to economic development.)

Sterman JD (1985) A behavioral model of the economic long wave. J Econ Behav Organ 6(1):17–53 (Proposes and tests a simple model of the long wave. The intended rationality of each decision rule is tested and the long wave is explained as the unintended result of the interaction of locally rational decision processes. The model is the basis for the STRATAGEM-2 game, and can exhibit chaos.)

Sterman JD (1989) Deterministic chaos in an experimental economic system. J Econ Behav Organ 12:1–28 (Sterman's 1985 model of the long wave is converted into a management flight simulator and used as an experiment in which subjects make the capital investment decision. Simple decision rules capturing subject's policies are estimated and explain their behavior well. Simulation of these rules yields deterministic chaos for about 25% of the subjects.)

Sterman JD (1986) The economic long wave: Theory and evidence. Syst Dyn Rev 2(2):87–125 (Comprehensive overview of the theory of long waves arising from the System Dynamics National Model. Reviews the feedback structures responsible for the long wave and empirical evidence supporting the dynamic hypotheses. Discusses the role of innovation and political value change.)

Conceptualizing, Formulating and Validating Models

Barlas Y (1989) Multiple tests for validation of system dynamics type of simulation models. Europ J Operat Res 42(1):59–87 (Discusses a variety of tests to validate SD models, including structural and statistical tests.)

Barlas Y, Carpenter S (1990) Philosophical roots of model validation: Two paradigms. Syst Dyn Rev 6(2):148–166 (Contrasts the system dynamics approach to validity with the traditional, logical empiricist view of science. Finds that the relativist philosophy is consistent with SD and discusses the practical implications for modelers and their critics.)

Forrester JW (1980) Information sources for modeling the national economy. J Am Stat Assoc 75(371):555–574 (Argues that modeling the dynamics of firms, industries, or the economy requires use of multiple data sources, not just numerical data and statistical techniques. Stresses the role of the mental and descriptive data base; emphasizes the need for first-hand field study of decision making.)

Forrester JW (1985) The model versus a modeling process. Syst Dyn Rev 1(1):133–134 (The value of a model lies not in its predictive ability alone but primarily in the learning generated during the modeling process.)

Forrester JW (1987) Fourteen 'Obvious Truths'. Syst Dyn Rev 3(2):156–159 (The core of the system dynamics paradigm, as seen by the founder of the field.)

Forrester JW (1987) Nonlinearity in high-order models of social systems. Europ J Operat Res 30(2):104–109 (Nonlinearity is pervasive, unavoidable, and essential to the functioning of natural and human systems. Modeling methods must embrace nonlinearity to yield realistic and useful models. Linear and nearly-linear methods are likely to obscure understanding or lead to erroneous conclusions.)

Homer JB (1983) Partial-model testing as a validation tool for system dynamics. In: International System Dynamics Conference, pp 920–932 (How model validity can be improved through partial model testing when data for the full model are lacking.)

Legasto AA Jr, Forrester JW, Lyneis JM (eds) (1980) System dynamics. In: TIMS studies in the management sciences, vol 14. North-Holland, Amsterdam (Collection of papers focused on methodology. Includes Forrester and Senge on Tests for Building Confidence in System Dynamics Models and Gardiner & Ford's discussion on Which Policy Run is Best, and Who Says So?)

Mass N (1991) Diagnosing surprise model behavior: A tool for evolving behavioral and policy insights. Syst Dyn Rev 7(1):68–86 (Guidelines for learning from surprise model behavior with tests to resolve anomalous behavior.)

Morecroft JDW (1982) A critical review of diagramming tools for conceptualizing feedback system models. Dynamica 8(1):20–29 (Critiques causal-loop diagrams and proposes subsystem and policy structure diagrams as superior tools for representing the structure of decisions in feedback models.)

Randers J (ed) (1980) Elements of the system dynamics method. Productivity Press, Cambridge (Includes Mass on Stock and Flow Variables and the Dynamics of Supply and Demand; Mass & Senge on Alternative Tests for Selecting Model Variables; and Randers' very useful Guidelines for Model Conceptualization.)

Richardson GP (1986) Problems with causal-loop diagrams. Syst Dyn Rev 2(2):158–170 (Causal-loop diagrams cannot show stock-and-flow structure explicitly and can obscure important dynamics. Offers guidelines for proper use and interpretation of CLDs.)

Richardson GP, Pugh AL III (1981) Introduction to system dynamics modeling with DYNAMO. Productivity Press, Cambridge (Introductory text with excellent treatment of conceptualization, stocks and flows, formulation, and analysis. A good way to learn the DYNAMO simulation language as well.)

Roberts N, Andersen DF, Deal RM, Grant MS, Shaffer WA (1983) Introduction to computer simulation: A system dynamics modeling approach. Addison-Wesley, Reading (Easy-to-understand introductory text, complete with exercises.)

Sterman JD (1984) Appropriate Summary Statistics for Evaluating the Historical Fit of System Dynamics Models. Dynamica 10(2):51–66 (Describes the use of rigorous statistical tools for establishing model validity. Shows how Theil statistics can be used to assess goodness-of-fit in dynamic models.)

Wolstenholme EF (1990) System enquiry – A system dynamics approach. Wiley, Chichester (Describes a research methodology for building a system dynamics analysis. Emphasizes causal-loop diagramming, mapping of mental models, and other tools for qualitative system dynamics.)

Modeling for Learning: Systems Thinking and Organizational Learning

Kim D (1989) Learning laboratories: Designing a reflective learning environment. In: Milling PM, Zahn EOK (eds) Computer-based management of complex systems: International system dynamics conference. Springer, Berlin (A case-study of a process designed to convey dynamic insights to participants in a workshop setting designed around a management flight simulator game.)

Morecroft JDW (1988) System dynamics and microworlds for policymakers. Europ J Operat Res 35(3):301–320 (Describes the model-building tools available to managers and policymakers.)

Morecroft JDW, Sterman JD (eds) (1992) Modelling for Learning. Eur J Operat Res Special Issue 59(1) (17 papers describing models and methods to enhance learning, both for individuals and organizations. Covers elicitation and group process techniques, management flight simulators, and tools for capturing, representing, and simulating mental and formal models.)

Richmond B (1990) Systems thinking: A critical set of critical thinking skills for the 90's and beyond. In: Andersen DF, Richardson GP, Sterman JD (eds) International System Dynamics Conference, 1990 (Proposes a process and skill set to teach systems thinking. The process relies on learner-directed learning. The skill set includes general scientific reasoning and SD, supported by simulation.)

Senge PM (1990) Catalyzing systems thinking within organizations. In: Masarik F (ed) Advances in organization development. Ablex, Norwood (Presents a case study in which the use of system dynamics generated insights into a chronic business problem. Steps in generating, testing and disseminating a system dynamics model are described.)

Senge PM (1990) The fifth discipline: The art and practice of the learning organization. Doubleday Currency, New York (Introduces systems thinking as part of a wider approach to organizational learning. Conveys basic system structures to a non-technical business audience by means of anecdotes and archetypes.)

Decision Making

Morecroft JDW (1983) System dynamics: Portraying bounded rationality. Omega 11(2):131–142 (SD models represent decision making as boundedly rational. Reviews and contrasts the concept of bounded rationality as developed by Herbert Simon. Uses Forrester's Market Growth model to show how locally rational decision rules can interact to yield globally dysfunctional outcomes.)

Morecroft JDW (1985) Rationality in the Analysis of Behavioral Simulation Models. Manag Sci 31(7):900–916 (Shows how the intended rationality of decision rules in SD models can be assessed, and how one analyzes a simulation model and output to understand the assumed bounds on rationality in dynamic models. A model of salesforce effort allocation is used to illustrate.)

Sterman JD (1987) Expectation formation in behavioral simulation models. Behav Sci 32:190–211 (Proposes and tests a simple dynamic model of expectation formation in dynamic models (the TREND function). Shows how the TREND function explains a forty year history of inflation forecasts and several different types of long-term energy demand forecasts.)

Sterman JD (1989) Misperceptions of feedback in dynamic decision making. Organ Behav Hum Decis Process 43(3):301–335 (Describes an experiment with a simple economic system in which subjects systematically generate costly oscillations. Estimates decision rules to characterize subject behavior. Finds that people systematically ignore feedbacks, time delays, accumulations, and nonlinearities. These misperceptions of feedback lead to poor quality decisions when dynamic complexity is high.)

Sterman JD (1989) Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. Manag Sci 35(3):321–339 (Analyzes the results of the Beer Distribution Game. Misperceptions of feedback are found to cause poor performance in the beer game, as in other experiments. Estimates of the subjects' decision rules show they ignore time delays, accumulations, feedbacks, and nonlinearities.)

Selected Applications of SD

Abdel-Hamid TK, Madnick SE (1991) Software project dynamics: An integrated approach. Prentice Hall, Englewood Cliffs (Integrated SD model of the software development process. The model covers design, coding, reviewing, and quality assurance; these are integrated with resource planning, scheduling, and management of software projects. Includes full documentation, validation, and policy tests.)

Cooper KG (1980) Naval ship production: A claim settled and a framework built. Interfaces 10(6) (An SD model was used to quantify the causes of cost overruns in a large military shipbuilding project. One of the first and most successful applications of system dynamics to large-scale project management; initiated a long line of related project modeling work.)

Ford A, Bull M (1989) Using system dynamics for conservation policy analysis in the pacific northwest. Syst Dyn Rev 5(1):1–15 (Describes the use of an extensive SD model of electric power generation with endogenous demand. The model is used to evaluate strategies for conservation and new generation capacity. Includes discussion of implementation and integration of the SD model with other existing planning tools.)

Gardiner LK, Shreckengost RC (1987) A system dynamics model for estimating heroin imports into the United States. Syst Dyn Rev 3(1):8–27 (Describes how the CIA used SD to estimate the illegal importation of drugs to the US.)

Homer JB (1985) Worker burnout: A dynamic model with implications for prevention and control. Syst Dyn Rev 1(1):42–62 (Explains how knowledge workers can experience cycles of burnout through a simple system dynamics model. Avoiding burnout requires that one work at less than maximum capacity.)

Homer JB (1987) A diffusion model with application to evolving medical technologies. Technol Forecast Soc Chang 31(3):197–218 (Presents a generic model of the diffusion of new medical technologies. Case studies of the cardiac pacemaker and an antibiotic illustrate how the same model can explain the different diffusion dynamics of successful and unsuccessful technologies.)

Homer JB (1993) A system dynamics model of national cocaine prevalence. Syst Dyn Rev 9(1):49–78 (An excellent model of the interacting dynamics of addiction, policy-setting, and enforcement.)

Jensen KS, Mosekilde E, Holstein-Rathlou N (1985) Self-sustained oscillations and chaotic behaviour in kidney pressure regulation. In: Prigogine I, Sanglier M (eds) Laws of nature and human conduct. Taskforce of Research Information and Study on Science, Brussels (Presents a system dynamics model of the dynamics of rat kidneys. Experimental data show previously unexplained oscillations, sometimes chaotic. The model explains how these fluctuations arise. Excellent example of SD applied to physiology.)

Levin G, Hirsch GB, Roberts EB (1975) The persistent poppy: A computer-aided search for heroin policy. Ballinger, Cambridge (Examines the interactions within a community among drug users, the police and justice system, treatment agencies, and the citizens. Analyzes policies designed to restore the community's health.)

Levin G, Roberts EB, Hirsch GB, Kligler DS, Roberts N, Wilder JF (1976) The Dynamics of Human Service Delivery. Ballinger, Cambridge (Presents a generic theory of human service delivery, with case studies and examples drawn from mental health care, dental planning, elementary education, and outpatient care.)

Naill RF (1992) A system dynamics model for national energy policy planning. Syst Dyn Rev 8(1):1–19

Naill RF, Belanger S, Klinger A, Peterson E (1992) An analysis of the cost effectiveness of US energy policies to mitigate global warming. Syst Dyn Rev 8(2):111–128 (Reviews the 20 year history of the SD energy models used by the US Dept. of Energy to forecast and analyze policy options for national energy security, including the impact of US policies on global climate change.)

Sklar Reichelt K (1990) Halter marine: A case study of the dangers of litigation. (Working Paper No. D-4179). System Dynamics Group, Sloan School of Management, MIT, Cambridge (A case-study illustrating the use of system dynamics in litigation. Suitable for classroom teaching.)

Sturis J, Polonsky KS, Mosekilde E, Van Cauter E (1991) Computer model for mechanisms underlying ultradian oscillations of insulin and glucose. Am J Physiol 260(Endocrinol. Metab. 23):E801–E809 (New experimental data show that the human glucose/insulin system is inherently oscillatory. An SD model explains these dynamics. The model is validated against detailed physiological data.)

Cross-Fertilization and Comparative Methodology

Allen PM (1988) Dynamic models of evolving systems. Syst Dyn Rev 4(1–2):109–130 (Reviews approaches to nonlinear dynamics, self-organization, and evolution developed in the Brussels school by Prigogine, Allen, and others. Provides illustrations and examples.)

Kim DH (1990) Toward learning organizations: Integrating total quality control and systems thinking. (Working Paper No. D-4036). System Dynamics Group, Sloan School of Management, MIT, Cambridge (Argues that SD and Total Quality Management are complementary approaches to improvement and organizational learning. Systems thinking and modeling are needed to speed the improvement cycle for processes with long time delays.)

Meadows DH, Robinson JM (1985) The electronic oracle: Computer models and social decisions. Wiley (Comparative assessment of the underlying assumptions, boundary, limitations, and uses of different models, including optimization, simulation, and econometrics. Offers guidelines for assessing model assumptions, including ways to recognize the implicit biases of each modeling paradigm.)

Powers WT (1990) Control theory: A model of organisms. Syst Dyn

Rev 6(1):1–20 (An explicit feedback control perspective on perception and decision making in living organisms. Argues the behaviorist and cognitive paradigms have fundamentally misunderstood the concept of feedback. For Powers, feedback allows organisms to control perceptions by altering behavior.)

Radzicki MJ (1990) Methodologia oeconomiae et systematis dynamis. Syst Dyn Rev 6(2):123–147 (Surveys the institutionalist paradigm in economics and argues that system dynamics is compatible with the institutionalist perspective. The SD approach offers a means by which institutional theories can be formalized and tested.)

Sterman JD (1985) The growth of knowledge: Testing a theory of scientific revolutions with a formal model. Technol Forecast Soc Chang 28(2):93–122 (Presents a formal dynamic model of TS Kuhn's theory of scientific revolutions.)

Sterman JD (1988) A skeptic's guide to computer models. In: Grant L, Lanham MD (eds) Foresight and national decisions. University Press of America (Reviews different modeling methods and their underlying assumptions in nontechnical language. Provides a list of questions model users should ask to assess whether a model or method are appropriate to the problem.)

Other Themes: Pulling the Threads Together

Cooper K, Steinhurst W (eds) (1992) The system dynamics society bibliography. System Dynamics Society. Available from Julie Pugh, 49 Bedford Rd., Lincoln MA, USA 01773. (Lists over 3,000 system dynamics journal articles, books, conference proceedings and working papers. Available in computer-readable format and compatible with bibliographic software)

Meadows DH (1989) System dynamics meets the press. Syst Dyn Rev 5(1):68–80 (Reviews the history of encounters between SD and the media. Offers guidelines for effective communication to the public at large. Stresses the importance of communicating even the simplest system concepts.)

Meadows DH (1991) The global citizen. Island Press, Washington (A collection of Dana's syndicated newspaper columns applying system dynamics principles to problems of everyday life, from organic farming to the fall of the Soviet Union. Emphasizes environmental issues.)

Richardson GP (1991) Feedback thought in social science. University of Pennsylvania Press (Traces the history of the concept of feedback in the social sciences through two threads of thought – the cybernetic and feedback threads. System dynamics is placed in context in a readable and scholarly manner.)

Software

DYNAMO. Pugh-Roberts Associates, Cambridge MA. (The first widely-used computer language developed to simulate system dynamics models, DYNAMO is still in use, available for mainframes and PCs. Many of the models in the system dynamics literature were simulated in DYNAMO)

DYSMAP. University of Salford, UK (PC-based simulation language with syntax similar to DYNAMO. Includes optimization capability based on hill-climbing.)

Microworld Creator and S^4. Microworlds Inc., Cambridge MA (Easy to use environment for simulation and gaming. S^4, the 'industrial strength' version, supports arrays and includes diagnostics for analyzing behavior. Both Creator and S^4 support user-defined information displays and facilitate rapid development of management flight simulators.)

STELLA and ithink. High Performance Systems, Hanover NH. (User-friendly modeling software with full graphical interface. Models are entered graphically, at the level of the stock and flow diagram. Widely used in education from elementary school up; also used in research and practice.)

Vensim. Ventana Systems, Harvard MA. (Powerful simulation environment for SD models. Runs on workstations and PCs. Includes array capability and a wide range of features for analyzing model behavior.)

System Dynamics Philosophical Background and Underpinnings

Camilo Olaya
Universidad de Los Andes, Bogotá, Colombia

Article Outline

Glossary
Definition of the Subject
Introduction
System Dynamics
"Real" World and Presentationalism
The Discussions on Positivism: Presentationalism and Knowledge
A Brief Note Regarding Social Theory
Simulation and Method
Future Directions
Bibliography

Glossary

Philosophy The reflection and study of our most basic assumptions – or the assumptions themselves.

Mental model A mental image of selected concepts and relationships of the world around us which we consider relevant for explaining the behavior of a particular system.

Presentationalism Synonymous of idealism. The view that material objects or external realities do not exist apart from our knowledge or consciousness of them.

Definition of the Subject

We all tend to take things for granted. Indeed it is a common place to judge formal models exclusively based on the technical grounds and on the logic with which those models were built without a proper reflection on the assumptions underlying those models. This omission is even more pressing in complexity and system science, since these areas represent a novel challenge for philosophers of science – e. g. see an overview in [34].

What is the idea of reality with which we work? What do we assume about human nature? What kind of knowledge do we pursue? What kind of knowledge do we obtain?

What is the scope of rational inquiry? What are the basis and the implications of our own reasoning methods? The identification of how philosophy has shaped the work of scientists – on a conscious or unconscious level – is essential for comprehending the implications, the limitations, and the scope of our very scientific practice. The lack of concern by scientists for these issues may explain many of their failures which has produced just a sort of inertial blindness that is easy to recognize in current scientific debates.

One of the strengths of system dynamics (abbreviated, SD) is that it leads us to make explicit our assumptions about the systems we deal with. This attitude, i. e. the importance of reflection upon our own assumptions, is also fundamental for the very development and practice of system dynamics. Many of the debates on different issues of every day scientific practice such as model conceptualization, formal model building, validation, policy design, etc. are informed and can be enlightened by the reflection on the philosophical background behind those processes. There are also various fundamental aspects of SD that are yet to be demarcated, e. g. the characterization of SD explanations. Furthermore, the ambiguity of the discussions found in large part of related literature, characterized by superficiality, confusion of terms, misdirected arguments, etc. only adds noise and it complicates the advance of a discipline. This article sketches and overview on some basic assumptions regarding the development and the practice of system dynamics. Various suggestions that help to integrate various debates are introduced and important clarifications are also indicated.

Introduction

The philosophical background and underpinnings of a discipline should have to do with its most basic universal assumptions. Such a discussion becomes difficult if we bear in mind that those assumptions are not necessarily shared by practitioners and researchers. Nevertheless, central premises can be identified which in turn can be related to important questions regarding philosophical concerns such as reality and knowledge.

The article is organized as follows. After this short introduction, the second sections develops an overview of the origins of system dynamics underlining fundamental aspects that formed what can be called the core of the discipline. This historical review highlights the initial interest, purposes and initial assumptions around the foundation of SD. With these elements the following sections introduce various philosophical issues that can be identified underlying system dynamics. Perhaps the central aspect is presentationalism, a stance associated with the notion of "mental models" which is central in SD; this is the topic of the fourth section. The following section makes a clarification on the controversial issue of positivism and relates presentationalism with knowledge. The sixth section summarizes the position of system dynamics regarding social theory. The seventh section presents the inquiry of explanation clarifying that in spite that SD involves causal models the nature of its kind of explanation can be found in the notion of mechanism. The eighth section introduces the implication of the use of computer simulation as a distinctive epistemology which is different from the traditional discourse in philosophy of science. The ninth section outlines future directions.

Before starting, a brief warning should be made: given the scope of this review and the limited space for covering very wide subjects then this article should be viewed as a broad introductory overview of the different topics. The cited literature has been selected as possible starting points for further inquiry.

System Dynamics

In order to address a "philosophical background" the first question naturally would be: What is system dynamics? Already this inquiry can be a matter of debate, e. g. [102]. Indeed SD has been labeled as a theory [23,49], a method [18,56,63,98,108], a methodology [81], a field of study [17,78], a tool [61], a paradigm, among other nouns. A natural starting point is the work of Jay Forrester, the founder of system dynamics. A brief historical review should help to grasp the very core we are looking for since it can show the initial motivations, assumptions, and purposes behind the development of SD.

Genesis of System Dynamics

Jay W. Forrester, member of the Sloan School of Management at MIT, was looking for linkages between engineering and management education given his background in feedback control systems and computers [28]. In 1956 he wrote a "note" to the Faculty Research Seminar, the first ever MIT "D-memo"; in this communication he sketched the worldview of what would be known as "system dynamics" [32].

He started with a strong criticism of economic models. The following were the central aspects: (i) their failure to reflect adequately the loop structures that make up economic systems; in particular this neglect leads to exclude inherent properties of closed loops such as resistance to change, accumulations and delays; (ii) The incapacity for including flows of goods, money, information, and la-

bor, in one single interrelated model; (iii) The exclusion of changing mental attitudes that affect and explain economic processes; (iv) The use of linear equations for describing systems; (v) The restriction of building models constrained by the capacity for manipulating numerical data and solving the equations; (vi) The overconfidence in multiple regression analysis for obtaining coefficients for equations that define economic behavior; (vii) The lack of reflection and discussion on the very assumptions underlying every model preferring an emphasis on the logic with which the model is developed.

After delineating these points, Forrester then proceeded in the same note to highlight techniques that were largely underused at that time: servomechanisms, differential equations, and what he called "the art of simulation". Anchored on the mentioned assessment and on these developments Forrester conceived "a new avenue of attack for understanding the firm and the economy" ([28] p. 336) envisaging *a new kind of models* that would include aspects such as:

Dynamic structure: Detailed attention to the *sequences of actions* which occur in the system being studied and to the *forces* which trigger or temper such actions, with a particular concern on the controlling influences of lags and delays.

Information flows: Explicit recognition of information flow channels and information transformation with time and transmission.

Decision criteria: Re-examination of the proper decision criteria which must not be defined as depending only on current values of gross economic variables; instead, such criteria must be traced to the motivations, hopes, objectives and optimism of the people involved, including as well what he calls business man's intuition which "represents a disordered accumulation of basic insights into how people and social systems react. The hope for the future lies in generating an orderly arrangement of basic insights" ([28] p. 342).

Non-linear systems: Economic systems present most – if not all – of the time highly non-linear characteristics.

Differential equations: The behavior of economic systems should be better described by non-linear differential equations since they have been developed to describe delays, momentum, elasticity, reservoirs, and accelerations, which are better suited quantities for describing the economic world. In practice these equations would be handled as incremental difference equations in order to obtain numerical solutions.

Incremental changes in variables: To prefer the formulation of a model in terms of the motivations that cause incremental changes in a variable since the new value of a variable "can be found by solving the equations for its incremental change and then adding the change to the preceding full value of the variable" ([28] p. 344).

Model complexity: Much complex and complete models can be developed with these techniques.

Empirical solutions: It is useless to look for explicit unique or "correct" solutions; instead, these models provide diverse solutions according to the different assumptions about the model structure and the initial values of the variables.

Symbolism and correspondence with real counterparts: The possibility of having a pictorial representation – a flow diagram – whose processes of information, money, goods, and people, are moved, i. e. simulated, time-step-by-time-step from place to place.

Structure over coefficient accuracy: To prefer a structure in which we have confidence using intuitively estimated coefficients instead of using unlikely structures with accurately derived coefficients from statistical data.

A subsequent advance came in 1958 with an article entitled "Industrial Dynamics: A Major Breakthrough for Decision Makers" published in the *Harvard Business Review* [27]. In this article Forrester shaped the previous ideas with the concern that management should evolve from a highly fragmentized art to a profession capable of recognizing unified systems given that the task of management is to interrelate the flows of information, materials, labor, money, and capital equipment. He again emphasized features such as electronic data processing, decision making, simulation, feedback control, and information flows. These elements were presented as the cornerstones of the innovative industrial dynamics program at MIT.

Industrial Dynamics

The definitive breakthrough came in 1961 with his *magnus opus* "Industrial Dynamics" [24]. The main motivation behind was the development of a *science* for designing and controlling industrial systems, the quest for a management science. In particular he conceived it as "a method of system analysis for management." (p. 9). He stated:

> Industrial dynamics is the study of the information-feedback characteristics of industrial activity to show how organizational structure, amplification (in policies), and time delays (in decisions and actions) interact to influence the success of the enterprise (p. 13).

Forrester underlined four pillars for this new science: information-feedback control theory anchored on the con-

cept of servomechanisms, the study of decision-making processes, an experimental approach to system analysis based on simulation, and the use of computers.

Coming from engineering, he had in mind models that deal with nonlinear dynamic systems whose purpose is to *design* new systems – as opposed to just *explain* systems; the models should show how changes in policies or structure will produce better or worse behavior. In order to accomplish this aim Forrester indicated that we should focus on understanding the characteristics of the system in hands (instead of looking for specific predictions) and on our assumptions about them. "We then have a means for tracing the implications of our assumptions" ([24] p. 55).

What should be included in a model? Forrester underlined that "there will be no such thing as *the* model of a social system, any more than there is *the* model of an aircraft … the factors that must be included arise directly from the questions that are to be answered" ([24] p. 60). It is expected that these factors will include closed-loop information-feedback structures that give rise to so much of the interesting behavior. An important aspect of this new kind of models is symbolism and pictorial representations by means of flow diagrams with a special emphasis on correspondence: the model variables should correspond to those in the system being represented. In this book Forrester also demarcated the network structure of this new kind of models as made of four basic components: accumulations (levels), flows (rates), decision functions, and information channels; these networks trace cause-effect relationships which are described via mathematical formalization. He also discussed in detail how to represent delays and how to model decision processes which are particularly defined by general policies, i. e. rules that state how operation decisions are made converting information into action; this study of general policies explains the importance for SD of the examination of human decision-making processes. Another fundamental characteristics are continuous flows and aggregation: "grouping of individual events into classes … Our interest in the model … is from the viewpoint above the separate individual transactions" ([24] p. 65); it is assumed that different individual items are controlled by the same identical decision-function; this notion leads to aggregation which is a distinctive aspect of SD: "items controlled by sufficiently similar policies that depend on sufficiently equivalent information sources may be combined into a single channel" ([24] p. 109); the central criterion for such aggregation is the purpose of the model. Finally, model significance (or validity) rests on its suitability for a particular purpose which is motivated by the design of improved industrial and economic systems.

Principles of System Dynamics

The main concern in these initial steps were business operations. However, Forrester sketched a glance to a broader view in the final part of his book; there, he speaks of "system" dynamics since "the study of systems can provide a framework to unite subjects … The dynamic model represents a system as broad as one chooses to describe" ([24] pp. 344, 346). Indeed, looking for a more general view he presents a section of "principles of systems structure":

> The principles to be discussed here all arise in the context of information-feedback systems. They are systems principles. They are not the principles of the management art such as have been taught in organization, production, and human relations courses. Because the principles apply to systems behavior, they do not fall into neat separate packages … The concept of a system implies interaction and interdependence. In attempting to identify factors that are common to all systems, we must keep the essential indivisibility in mind ([24] pp. 347, 348).

Indeed, Forrester reaffirmed this general *systems* view in a follow-up book entitled *Principles of Systems* [25] published in 1968. He gives a basic definition of system as "a grouping of parts that operate together for a common purpose" ([25] p. 1-1). Moreover, he suggests that the way to organize knowledge is with this idea of system which are represented by the models we develop:

> A structure (theory) is essential if we are to effectively interrelate and interpret our observations in any field of knowledge … Without an organizing structure, knowledge is a mere collection of observations, practices and conflicting incidents … A model is s substitute for an object or system … Any set of rules and relationships that describe something is a model of that thing. In this sense, all of our thinking depends on models. Our mental processes use concepts which we manipulate into new arrangements. These concepts are not, in fact, the real system that they represent. The mental concepts are abstractions based on our experience. This experience has been filtered and modified by our individual perception and organization processes to produce our mental models that represent the world around us ([25] pp. 1-2, 3-1).

The previous statement summarizes core assumptions for SD. It points at the central notion of *mental model* and it emphasizes that these models are models of *systems*. In fact, in an article published in 1968 in *Management*

Science, Forrester [28] underlined that this application of feedback concepts to social systems was evolving toward a theory of structure in systems with a particular goal of policy design. Specifically, "industrial dynamics is a philosophy of structure in systems. It is also gradually becoming a body of principles that relate structure to behavior" ([28] p. 141). Two fundamental variables are identified: levels and rates, and the basic structural element is the feedback loop: "every decision is responsive to the existing condition of the system and influences that condition" ([28] p. 143). And, as stated above, the structure is an aid to organize knowledge in a particular situation given a particular purpose which is motivated by the pursue of explanation and policy design.

This historical review has accentuated central aspects of the foundation of SD. As a summary, this science initially envisaged by Forrester for designing systems is known nowadays as system dynamics. The models developed in SD have distinctive characteristics: dynamic structures, information flows, study of decision criteria, non-linearity, difference equations, symbolism and correspondence, and emphasis on confidence based on the structure of the model. The practice of this science is anchored on: the concept of servomechanism, the study of decision-making processes, the embrace of an experimental approach, and the use of computers for building formal models. These models are models of systems and the main goal is to help to organize our knowledge – which can be seen as arranged in mental models – so as to enhance learning processes and systems design on concrete settings and under specific purposes.

With these elements in mind, it is now possible to proceed with a brief discussion on important aspects regarding assumptions on reality and knowledge that can be recognized behind these premises and the practice of SD.

"Real" World and Presentationalism

The traditional distinctions of *ontology*, *epistemology*, and *methodology*, form the habitual framework to drive a philosophical discussion. But the isolation of these issues may not be the most adequate or clear strategy. Given the interrelated nature of such categories and the misleading discussion on those terms which is present in a large part of organization and management science literature – part of this confusion will be exposed and clarified below – then a different plan will be used to develop the rest of this article. An instinctive option is to pick up significant issues and to relate them with what we can identify as part of the core of SD.

Where to start? Assumptions concerning a "real" world can be a first step to take since the traditional debate developed through the years around the claim of building models of "social systems" has fueled part of the discussions in systems science. As most of examinations on philosophical matters the debate has been permeated by a confusion originated in terms and words without the proper examination of the topic. However, this short revision is useful for opening the assessment of the premises of SD regarding a real world.

A prominent example is the criticism made by influential commentators who state that SD models represent an assumed "objective" real world [49]. This kind of critique usually labels SD as a "hard" approach, e. g. [48], meaning with such a term models of an assumed objective machine-like world – and habitually including and inverting the meaning of the term "positivism" – a mistaken assessment that still can be seen nowadays, e. g. [16]. This type of comments can be illustrated with the following quote from the work of Flood and Jackson [23]:

> System dynamics models still center on capturing the structure of the "real world" … the underlying assumption of SD that there is an external world made up of systems the structure of which can be grasped using models built upon feedback processes … Because intentions derive from inside social systems, from the conscious human actors which constitute them, many possible appreciations of the nature and purpose of particular social systems are possible. SD simply does not deal with the innate subjectivity of human beings … In essence the argument is that social systems cannot be studied, in the way of system dynamics, objectively from the outside (pp. 78, 79–80).

However, that is not what SD looks for. The mentioned emphasis on the examination of human decision-making processes and on the assumptions behind, the notion of mental model, the fact that model significance rests on its suitability for a particular purpose, among other aspects, should suffice to illustrate the point. Already Forrester in 1961 [24] emphasized: "a model can be useful if it represents only what we *believe* to be the nature of the system under study … we are forced to commit ourselves on what we believe is the relative importance of various factors. We shall discover inconsistencies in our basic assumptions … Thorough any of these we learn" (pp. 57–58, emphasis original). This should be enough to discard the assessment made by Flood and Jackson, and similar critics, who mistakenly placed SD in the terrain of a sort of naive realism as depicted in the quote above; this point is extensively clar-

ified by Lane [55,56]. More importantly, this argument is helpful to introduce the discussion about the nature of SD models, the assumptions behind these models, and their relation with a "reality". The notion that drives these matters has been labeled in SD literature as "mental model" and even though this concept is not free of debate [21] it is placed at the core of the discipline.

The idea of mental model was already addressed by Forrester, as it has been indicated; he demarcated it in 1970 as the mental image of selected concepts and relationships of the world around us that we carry in our heads [26]; furthermore, "the mental model is fuzzy. It is incomplete. It is imprecisely stated … [It] changes with time" (p. 213). Doyle and Ford [22], looking for a consensus on this subject, propose as a definition: "a relatively enduring and accessible [conscious], but limited, internal conceptual representation of an external system (historical, existing or projected) whose structure is analogous to the perceived structure of that system" (p. 411). Sterman [98] summarizes what the expression "mental model" refers to: "our beliefs about the networks of causes and effects that describe how a system operates, along with the boundary of the model (which variables are included and which are excluded) and the time horizon we consider relevant … Most of us do not appreciate the ubiquity and invisibility of mental models, instead believing naively that our senses reveal the world as it is. On the contrary, our world is actively constructed (modeled) by our senses and brain" (p. 16–17). It should be noticed that these mental models may refer as well to planned or desired systems existing in the mind of the modeler [54]. The ultimate goal of system dynamics is to enhance our learning processes by testing and improving our mental models in a way that becomes consistent with the complexity of the systems that we face and design everyday [98].

Then, what is a SD model as related to some "real" world? Or how is it related to the notion of mental models? The usual option to answer these questions is to frame the discussion in the debate realism/anti-realism. This examination is even more relevant considering that some criticism labels SD as anchored on "realism". A brief clarification follows.

Realism? Anti-Realism?

Typical of latest debates in history of science concerns the dispute between the so-called *realism*, i. e. theories are true *or* false as descriptions of the world, and *instrumentalism*, i. e. theories are more or less adequate. This latter position is closer to pragmatism: theories are just instruments to systematize – and for many philosophers predict – observations, but theories are not claims about the world; or they can be also subjected to linguistic frameworks but still always truth-value-less. An overview of this debate is made by Leplin [60]. Yet, here it should be clarified that both positions and those definitions – habitually taken by historians of science – are two sides of the same coin. In short, both are forms of idealism. This assessment will be commented next.

On the one hand, scientific "realism" is usually defended using the method of abduction as source of knowledge, i. e. inference to the best explanation, which is just a form of induction and hence it is confirmationist, ultimately relativistic. It is not realism at all; it is exactly the opposite, i. e. idealism. For instance, take the influential ideas of Sellars [91] who defends an illustrative traditional position of scientific realism holding the source of knowledge on observation and relying on justification by induction for building theoretical frameworks; he emphasizes: "Laws in question are stipulated to be inductively established in the observation framework" (p. 313), the theories are then refined by empirical generalizations and what he calls further "injections" of images of the theory into the observational framework, in a typical process of instructional correction. A further refined and formal model of such a framework (proposed by Friedmann, holding a model–submodel relation) is commented by Morrison [66] who still holds, nevertheless, the search for truth and justification as support for his criticism and the necessity of confirmation [67]; see also for example the paper of Kukla [52] with a criticism to the ideas of Friedmann, yet also supported also on confirmation. In short, this "realism" is just idealism as we know it. Similar "realist" positions abound on the literature of the history of science. E.g. Smith [95] postulates a realist stance based on "common sense" but, since for him "science begins with observations" (p. 53), his realism ends up, indeed, trapped in sensations, i. e. idealism. Quantum mechanics does not escape the debate, e. g. arguing in favor of realism within the subjective theory of probability (the Copenhagen interpretation, Bohr, Heisenberg, etc.), Dickson [20] stands on the verification criterion for discussing on what he calls "quantum realism"; naturally such base can not succeed, which is anyway best self-explanatorily reflected in the first part of the title of his paper "An Empirical Reply to Empiricism"; simply there is no such reply. Another case is the criticism of Brown [13] to the deterministic notion of "realism" that Cherniak discards (who supports it on computer simulation of finite agents); the criticism of Brown shows a Galilean notion of realism, i. e. achievable true descriptions of the world via laws, but in a very "complex world" and thus, for him, inaccessible to agents with lim-

ited cognitive capacity; this Galilean view is found in large part of complexity science. In other attempts, for instance Schlagel [90] defends "contextualistic realism" which ends up in relativism: elements of the world have a conditional status relative to contexts and conditions, the existents are real relative to the particular structures and contexts on which they depend; his proposal is just idealism and can be better described as a sort of multi-phenomenalism, indeed relativism. The diverse "realist" positions *actually rooted on non-negotiable empiricism* seem endless; e. g. further examples are found in [1,59,75], and in most papers with the expression "scientific realism" in the abstract.

On the other hand, let us consider a supposedly opposite position, for instance instrumentalism. Instrumentalism actually ends up in relativism as well, e. g. with respect to the points of view where instruments can be applied; at the end the notions of "applicability" or "adequacy" become reference frameworks for establishing partial truths. Newton-Smith [68] proposes a conciliatory position, he calls it "modest realism" which ends up indeed in a sort of "moderate" relativism; a catalog of various scientific realisms can be found in that paper too, all of them addressing still the question of how such truths about a real world can be sustained (justified). In short, instrumentalists declare the dependence on observations, i. e. idealism, and thus this position is in fact sharing assumptions with the alleged "realists". Leplin [60] summarizes: "some theory can be reduced to observation by defining or translating theoretical terms into terms that describe observable conditions. The remainder must be construed instrumentally" (p. 394).

The "realism" vs. instrumentalism debate is futile and yet it is perhaps one of the core discussions in philosophy of science. But unlike these influential historians of science, the presented dispute tends to be dismissed by several professional philosophers who see it as self-serving and unsophisticated; Fuller [33] underlines this assessment though he also illustrates the consequences of leaving such pointlessness discussions to endure. In this case, two seemingly unaware idealist factions argue about the best way to establish positive knowledge, e. g. either with a supposedly "true" description of the world (more precisely: *phenomena*, subjectivism) or by pragmatism (and again: *phenomena*, relativism). This shared *idealism* is the matter of the next section.

Presentationalism

A broader debate can be assessed framed in the opposition between *realism* and *idealism*, see e. g. [72] and [77]. It will be shown that SD rests on the broad stance known as idealism, i. e. presentationalism – this latter term is preferred here just for clarity [10].

Firstly, it should be commented the widely inverted and misleading use of both terms. And there is good company. Blackmore [10] shows various celebrities that became misusers of the expressions in question such as the former president of the American History of Science Society and prominent Harvard University professor, Erwin Hiebert, and Sir Russell Brain, outstanding neurologist and former president of the British Association for the Advancement of Science. Add the seemingly customary tendency to quote references in second hand without inspecting direct sources and a few decades later we have the terms used in exactly their opposite original sense in journals and books. Blackmore pictures the situation:

> Like-minded 'empiricists' have restricted what they understand by the term to what idealistic philosophers of science *have wanted them to understand by it*. And since the term 'realism' sounds good to 'tough-minded empiricists' and since idealistic philosophers such as Hume, Comte, Schuppe, and Mach and their recent successors *scarcely if ever have admitted their idealism*, many scientists and historians of science have let themselves be seduced into reversing the normal epistemological definitions of 'realism' and 'idealism'. Even worse, respected scholars such as Stillman Drake, perhaps our most outstanding authority on the manuscripts of Galileo, and Larry Laudan, a young, energetic, and much published commentator on Mach and 'empiricism', have allowed themselves to become advocates of hopelessly naive 'non-philosophical' positions. Drake is sure that Galileo held no philosophical position, or that if he did, it had no effect on his scientific work. Laudan is equally positive on the basis of Mach's written comments that his phenomenalistic epistemology had no influence on Mach's 'empirical' methodology of science. The simplicity of their views can party be explained by their tendency to understand by the term 'philosophy', not one's most basic universal assumptions, *but expressed talk about speculative matters*. Similarly, many 'materialists' who feel sure that the physical world is directly given in experience, and who accept the idealist Kant's distinction between 'science' and 'metaphysics', are convinced that anyone who identifies the physical world with what is *beyond* immediate experience is an 'idealist' and 'metaphysician' and (following Wittgenstein) 'is merely uttering nonsense' (p. 131 in [10]).

In order to clarify, it is appropriate to underscore the attitude behind an empiricist epistemology. A major defining posture is what has been labeled as *idealism*, given the natural disbelief of a world beyond senses, which is the pillar of an empiricist epistemology. The term "idealism" comes from the "idea" of Bishop Berkeley, who took physical objects as "ideas" which included sensations and thoughts: "It is evident to any one who takes a survey of the objects of human knowledge, that they are either ideas actually imprinted on the senses, or else such as are perceived by attending to the passions and operations of the mind, or lastly ideas formed by help of memory and imagination, either compounding, dividing, or barely representing those originally perceived in the aforesaid ways." [8]. This idealism relies entirely on senses and mind-dependent worlds since consequently sense-data were the only things of whose existence our perceptions could assure us, and that to be known is to be 'in' a mind, and therefore to be mental. Berkeley, therefore, concluded that nothing can ever be known except what is in some mind, and that whatever is known without being in my mind must be in some other mind [85]. Hunter [47] summarizes:

> As a result of their constraints on knowledge and meaning, empiricists tend to be skeptical of necessary truths that are independent of mind and language, and of putative eternal abstract entities (p. 110).

This idealism can be equally identified with terms such as 'phenomenalism', 'neutral monism', or 'subjective idealism' (e. g. [10]), or presentationalism. In other words, "anything in time or space, anything than can be known by the human mind, is phenomenal" (p. 146 in [15]).

Taking this posture to the context of science, Bartley [4] provides the implications: "Presentationalists see the subject matter of science not as an external reality independent of sensation. The subject matter of science is our sensory perceptions. The collectivity of these sensations is renamed 'nature' … The aim of science is seen not as the description and explanation of that independent external reality but as the efficient computation of perceptions … [It] became the dominant twentieth-century *philosophy* of physics" (p. 11, 16, emphasis original). In general this position is the pillar of the prevalent conception of science which has been fueled by physics (for instance in the interpretation of quantum mechanics of Bohr and Heisenberg) and backed by influential names like Mach, Russell, Wittgenstein, Ayer, Lewis, Carnap, etc.

To appreciate the contrast, a standard definition of *realism* can be:

> 'Realism' … is used for the view that material objects exist externally to us and independently of our sense experience. Realism is thus opposed to idealism, which holds that no such material objects or external realities exist apart from our knowledge or consciousness of them, the whole universe thus being dependent on the mind or in some sense mental. It also clashes with phenomenalism, which, while avoiding much idealist metaphysics, would deny that material objects exist except as groups or sequences of sensa, actual and possible (p. 126 in [10]).

Apart from the particular emphasis on material objects (as opposed to Berkeley's *ideas*), another point is that realism defends a cosmocentric thesis opposed to the anthropocentric view in the discussions of the alleged "realists" of science presented earlier; in the latter case the observer-centered learning process is fundamental, and it is carried on via induction looking for acquiring positive, verifiable, and true knowledge – or justified true belief. Moreover, let us recall that historians of science denote with the term "realism" just the concern with supposed true descriptions of the world. But in fact *realism* does not imply that knowledge is achievable, it does not imply that the world is a perfect clock, it does not imply determinism, it does not imply that there can be correspondence between theories and such real world; these are different affirmations that unfortunately seem to be muddled inside the same bag. One thing is to assume a real world beyond senses. But a different inquiry is the character we ascribe to it. Another different issue is the role we assume for our senses. Another very different concern is the question of knowability, etc. Rescher [77] illustrates typical examples of the confusing use of terms in literature: "The three positions to the effect that real things just exactly are things as *philosophy* or as *science* or as '*commonsense*' takes them to be – positions generally designated as *scholastic*, *scientific* and *naive* realism, respectively – are in fact versions of epistemic idealism exactly because they see reals as inherently knowable and do not contemplate mind-transcendence for the real" (p. 187).

Coming back to presentationalism, this position then assumes that we are imprinted by the environment, and we call to this *impressions* "knowledge". Given the limitation of our senses then presentationalism postulates that nothing more can be known; and thus, such assumption is used to construct the world in our minds: for a presentationalist, strictly speaking, the world is not re-presented since we do not have access to it, the world is just what is *presented* to our senses: the world as we experience it happens to be the world itself; and, since anything that can be known by

the human mind is, then, phenomenal (sensations, etc.), therefore knowledge strictly depends on – and is source in – what is sensed, e. g. observed. Hence knowledge needs to be justified in order to avoid error; and yet, the only existent knowledge is the imperfect evidence sourced in sensation and, nevertheless, a foundation that can be justified is pursued. This is the popular plan we have come up with, so far, to try to avoid the destruction of empiricism made by Hume. The picture can be summarized:

> Almost all traditional epistemologies are Lamarckian in their accounts of the growth of knowledge. This is conspicuously true of presentationalism, almost all adherents to which maintain an inductivist, justificationalist account of knowledge growth, according to which knowledge is constructed out of sensations (as building blocks or elements) by a relative passive process of combination, accumulation, repetition, and induction (p. 25 in [4]).

This position is identified with *idealism* and most popularly associated with *epistemological empiricism*, with all its assumptions and its consequent scientific method.

The last point to underline is that this epistemology has subordinated ontology. A remarkable illustration of this type of problems was already made by Bowman [11]: "The result … is the more or less deliberate abnegation of a genuine epistemology and the substitution for it of a highly formal logic. Hence the paradox illustrated equally in the case of Plato and, recently, of Mr Russell, of a radical empiricism (expressed in Plato's Protagorean theory of sensation and in Russell's subjectivism) subsisting side by side with the extremist rationalism. Such a dualistic position is the despair of philosophy, and indicates a failure in the synthetic work of thought" (p. 485). This despair is easy to recognize in current science which presents a contradictory ontological position whose difficulty is found in its subsumption under a radical epistemology. Indeed extreme empiricism, i. e. presentationalism, has become the metaphysics, i. e. the theory of reality, of our science. The debates on "scientific realism" presented earlier picture this failure. On this particular subject the different use of terms in literature is a source of confusion; but here the terms have been inverted by historians of science and scientists; and beyond a semantic confusion this has brought a narrow conception and a very restricted examination of epistemological assumptions. In short, the subjectivism of Descartes and Kant – or more precisely, Kantian idealism – is what now is labeled as "realism", e. g. *everything* has become "phenomena". As a matter of fact common expressions like "objective phenomena" or "real phenomena" uncover an idealistic position where the "objective" or "real" are just *phenomena*. Indeed, the so-called scientific "realism" commented above is nothing more than *a sort of empiricism* driven by Hume and Kant. This "realism" is just ontology overlapped by epistemology (*idealism*). More precisely, regarding Kant, let us recall his "Transcendental Idealism" which was the common answer of Kant when he was accused of idealist, e. g. see [93], denying to be a "dogmatic idealist" (in the Berkeley sense; see e. g. [101]); a full discussion of this failed defence of Kant is made by Guyer [37]. In particular Turbayne [101] defends Kant when he was accused of having misinterpreted (or even completely having misunderstood) Berkeley's idealism; yet, Turbayne's conclusion summarizes the Kantian idealism (and its ambiguity) unmistakably: "The Kantian antidote to this is not the a priori nature of space, but *its reality or subjectivity*, which assimilates space and its contents into the realm of ideas, and thus *prevents* illusion" (p. 243, emphases added). Regarding consequences, perhaps the best summary is the radical idealist position of Mach – who denied even the existence of atoms since they cannot be observed. Kant seems to be taken for granted without a proper reflection on his position.

These few points were commented since this debate is the major informer of the method and the assumptions of management science, organization science, and social science in general, places where SD has its roots. More important, by making this clarifications then a clear ground for SD can be envisaged. It can be affirmed that SD has been mistakenly labeled as "realist" by many commentators alluding the alleged aim of building "true" descriptive theories of the world, and using the misleading definitions of historians of science. But also from the discussion above it should be clear that SD safely rests on presentationalism. Moreover, the identified presentationalist stance known as "instrumentalism", and in general the so-called anti-realists positions (independent of the inverted use of the term), fit to the SD worldview: the abstractions from our experience are arranged in mental models which form knowledge that we want to improve in order to make better decisions. The SD models built for achieving this goal are judged against their adequacy and suitability for a particular purpose; these models are not claims about the world but instruments for systematizing observations and for boosting learning processes using experimentation via simulation.

The Discussions on Positivism: Presentationalism and Knowledge

An issue previously mentioned which is closely connected with presentationalism is relativism and positivism. Since

positivism usually – and mistakenly – is pejoratively associated with a supposed objective representation of reality, then an important clarification is needed. In short: positivism is consistent with presentationalism and with relativism as well.

Presentationalism Brings Positivism

Blackmore [10] reminds that strictly speaking neither "rationalism" nor "empiricism" are properly epistemological terms at all; the entrenched idealism has led just to this narrowed identification. For instance, usually the term "empiricism" is synonymous of knowledge sourced in observation, i. e. in a restricted epistemological context, but such popular narrowness is inaccurate. Blackmore remarks: "Granted, that if one *means* by 'empiricism' not just an extensive and careful concern with empirical evidence but *restricting* reference or knowledge or both to sensory appearances, then there are indeed epistemological implications. One has become an epistemological phenomenalist or subjective idealist, or if you will, a positivist" (p. 130, emphases original). This clarification is needed for two reasons; on the one hand, as it was stated, SD has been labeled as "positivist" but the critics take this term as a sort of naive realism; the confusion is patent once we realize that positivism actually is a consequence of idealism, the opposite doctrine of realism. On the other hand, since SD is better identified with idealism then a sort of positivism can be also associated with it, but not the sort of "positivism" that the critics have in mind but the authentic positivism; and yet we will see that with the use of simulation positivism does not necessarily fit either.

The case of management science is a good example regarding the discussion on relativism and positivism. Let us consider the traditional and unfortunate sharp division between "hard vs. soft" which also takes the form "quantitative vs. qualitative". However, this discussion is misleading as well. It is not difficult to find researchers that claim to be anti-positivists but being themselves grounded on positivism (e. g. empirical observation, verification, induction, etc.) without noticing the contradiction. A good example is the claimed opposition between positivism and phenomenology. Yet, phenomenology is authentic positivism when is committed to evidence – Husserl himself underlines this aspect – see e. g. [94]. Indeed it is easy to appreciate an inverted use of the term "positivism" in literature; in management science this is a favored and widespread practice where so-called anti-positivists do not notice their positivism. The fact is that anthropocentrism is the ground in our most influential epistemologies that recognize the obvious imperfections of our sensorial apparatus but, nevertheless, rely knowledge on sensation (observation, etc.), that is, positivism, which is nothing more that our anxiety to confirm and validate, i. e. justify, our "imperfect" knowledge, e. g. empirically. This is a simplification of highly loaded terms; clarifications and further discussions can be found elsewhere, e. g. regarding positivism see [9,97,100].

Positivism is Anchored on Justification

Within a presentationalist worldview the search for confirmation and verification is nothing less than the search for justification of knowledge where the intellectual authority lies in sense experience. From a presentationalist account it is straightforward to have a justificationist approach for confirming and verifying theories. Following Bartley [4]:

> Preoccupied with the avoidance of error, they suppose that, in order to avoid error, they must make no utterances that cannot be justified by – i. e., derived from – the evidence available. Yet sense perception seems to be the only available evidence … The claim that there is an external world *in addition* to the evidence is a claim going *beyond* the evidence. Hence, claims about such realms are unjustifiable. Crucial to the presentationalist argument are, then, two things: the desire to give a firm foundation or justification to the tenets of science, and the construal of sense experience as the incorrigible source of all knowledge (pp. 12–13, emphases original).

In fact justification philosophy taken as the search for epistemic 'authorities' has been the dominant style of western philosophy looking for "well-grounded" knowledge. For instance in the customary view of knowledge as *justified true belief*, e. g. in the sense of Russell [86]) – as the result of systematic analysis "of our sensory experience of a knowable external reality" (p. 47 in [96]). Within this popular position the central problem of epistemology – as succinctly formulated by Radnitzky [76] – becomes:

> "When is it rational or, so to speak consistent with one's intellectual integrity, to *accept* a particular position?" The formulation suggests the direction in which the answer is to be sought: "When concerned with a statement, a theory, etc., accept those and only those statements, theories, etc., which not only are true but whose truth has been established" (p. 282, emphasis original).

The goal of justification is usually entrenched within the method of induction where every new repeated observation is a confirmation that validates – justifies – the theoretical statement. Even with weaker conditions the way

of reasoning is the same, for instance within the ideas of Ayer where strict verifiability is seen as a too rigid criterion – he introduces *confirmability to some degree*, instead of complete and conclusive verifiability (see [88]); justification is still pursued. The appeal of justification can be explained because it looks for avoiding relativism (inherently attached to presentationalism) since not all positions are equally good or bad and it suggests to look for something beyond blind belief [76]. Though it is not the only option, nevertheless, it is the most common view of science; the concerns on validation, justification, verification, confirmation, and generalization, are part of this popular and influential view. Here the observer is the fixed point of reference. In short, within a justificationist logic, it is rational to accept only those positions that have been justified according to the rational authority which in the case of presentationalism is sense experience, consistent then with the highly influential ideas of Locke, Berkeley, Hume, Mach, Carnap that have shaped our prevalent view of science [4].

Justification in System Dynamics

Turning back to SD, it must be recalled the role of mental models whose characteristic nature of "abstractions based on experience" can be better assessed within a presentationalist stance. Here justification has also a place. How is this knowledge justified? Already Forrester [24] emphasized, within the debate of model validation, that, "knowledge of all forms can be brought to bear on forming an opinion of whether or not a model is suitable to its particular purpose" (p. 129). Therefore, Forrester [25] also emphasized that "we can never prove that any model is an exact representation of 'reality' … Models are then to be judged, not on an absolute scale that condemns them for failure to be perfect, but on a relative scale that approves them if they succeed in clarifying our knowledge" (p. 3–4). This sort of relativism will be addressed next.

With the aim of placing this stance within the discussion of history of science, Barlas and Carpenter [3] addressed the "philosophical roots of model validation" associating SD with what they called a "relativist philosophy of science". In this view justification is pursued: such a knowledge is seen as socially, culturally and historically dependent and it becomes socially justified belief. Here "a valid model is assumed to be only one of many possible ways of describing a real situation … for every model carries in it the modeler's world view … validation is a matter of social conversation" (p. 157). Hence, confirmation and verification are pursued through a social process relative to a frame of reference. This sort of moderate relativism was later criticized by Vásquez, Liz and Aracil [103] for whom such relativism is unacceptable in spite of their recognition that there is no privileged single model (or set of models); since they are also concerned with epistemological justification these authors present Putnam's internal "realism" as a more adequate way to conceptualize the type of knowledge consistent with the assumptions about reality held by SD practitioners; in short, these authors underline that mental models are the source of knowledge – and its justification – helping to select the structures that must be assumed as working in real systems; here knowledge is taken as internal to the conceptual scheme of SD. Since there can be many models for a given situation, the authors argue that this framework gives the possibility of convergence as a result of "the strong interactive character of mental models" (p. 34) recognizing that in any case SD modeling is a process of revision and adjustment. With this proposal these authors seek to achieve justification and some realistic representational content in spite of the plurality of alternative SD models available for a specific situation. It is easy to see that this proposal is still relativistic, in this case knowledge is relative to the mental models and the conceptual scheme – though the mentioned authors would not agree since for them there is a "reality" given by the internal representational schema, in this case the mental models of the modelers.

The fail to recognize presentationalism as the epistemology which is driving ontology is at the root of the presented discussions. This is a distinctive trait present in large part of the philosophy of science literature. However, the characteristic problem of mistaking positivism with a sort of supposed "objectivism" is also present in this discussion – in fact, Barlas and Carpenter, following the misdirecting literature on the subject, argue that the relativist philosophy that they defend rejects positivism; Vásquez, Liz and Aracil also follow the same inertia. However, these theses, which seek for confirmation, verification, justification, reflect the search for *positive* knowledge within an idealist epistemology. In the first case, knowledge is confirmed and accepted through social interaction and it is relative to a context. In the second case, knowledge is justified on mental models. To appreciate a genuine contrasting position see an anti-justificationist approach to validation in computer simulation in [51]; the core of anti-justificationism can be found in the work of Popper, e. g. [73,74].

System Dynamics as Presentationalist

It should be clear that the assumptions on SD reject naive realism; the models are not supposed to be accurate and

corresponding descriptions of an external true reality – and furthermore, this is not what the term "positivism" refers to. In any case validation in SD does not mean a supposed "positive proof" or to assess the development of "true models" of the world. On the contrary, SD aims at enhancing our ways of reasoning, it emphasizes a process of learning so as to consistently improve our mental models which are product of our experience and the operations of our mind. Hence SD is closer to presentationalism. Knowledge can be socially justified and our mental models can be enhanced. Moreover, a particular emphasis on the modeling *process* is also underlined – see e. g. [29,44,89] and it will be commented in the eight section devoted to simulation.

A Brief Note Regarding Social Theory

Another important discussion is related to the theories about the "social world" that are supposedly held in SD, i. e. the social theory behind SD if any. Part of the misguided debate is explained by the widespread use of the traditional framework of Burrell and Morgan [14] whose oversimplification of social science in four paradigms has deviated major issues covering important topics under a too practical and inadequate schema – e. g. see a criticism in Deetz [19]. Jackson [49] provides a summary of such usual misconstruction:

> System dynamics … is essentially functionalist in nature. It sees system structure as the determining force behind system behavior … If humans are free to construct social systems as they wish, what determining influence does system structure have? … This tension between determinism and free will is unresolved (p. 39).

It should suffice to recall from the discussion above that the *aggregate* approach of SD is not a theory of human behavior; SD is not concerned with *individual* action. Furthermore, it does not assume that a structure, of any kind, determines human behavior either, i. e. the sort of determinism that Burrell and Morgan [14] oppose to "free will" and in the lines of the already vague term known as "structuralism" – see an early clarification of the problems of such type of oversimplification in [84]. In any case, this sort of criticism has been answered and clarified by Lane [56] who has underlined the main point: SD is concerned with aggregate social phenomena and not with individual meaningful actions [55]. Moreover, system dynamics does not propose invariant causal laws, as Lane [56] also concludes: "The only universal law/theory on offer is a grand methodological, or structural theory, associated with a representation scheme … it does not attract the determinism-related criticisms attached to grand theory in the sense of Parsons and Mills" (p. 111). In [57] Lane proposes to link system dynamics with a different framework: agency/structure theories.

Servomechanism

Part of the confusion is because of the misunderstanding by various commentators of the notion of feedback that underlies SD. This point has been also a source of misconception given the use of feedback in theories of control applied to social systems, e. g. cybernetics. Consider for example the following comment by Flood and Jackson [23]:

> The attempt of SD thinkers to model external reality is misguided … The emphasis placed on "structure" as the means of revealing knowledge about the optimal behavior of systems cannot be accepted … SD modelers using feedforward control appear to believe that there are optimal future states that we should steer systems towards" (pp. 80, 81).

Such criticism apparently is associated with the notion of feedback used in cybernetics. However SD does not pursue optimization, let alone by studying "knowledge revealing" structures in order to achieve supposed optimal behavior patterns. Instead, the goal is to have a better understanding of feedback structures in order to enhance decision-making and policy design. This point has been also addressed by Lane [53,55]. The central clarification of this issue has been made by Richardson [78] who distinguishes two different threads in the development of the concept of feedback in the social sciences: the cybernetics thread and the servomechanisms thread. The failure in noticing these two different lines of thought has produced various misconceptions regarding the notion of feedback in SD, a concept that has been shown as one of its building blocks. On the one hand, the cybernetic conception of feedback is defined in terms of input and output, it is limited only to loops of negative polarity which in turn are conceived as the mechanisms of control – and hence there is a particular interest in goal seeking and goal formulation given the concern in cybernetics for achieving adaptive behavior via directed processes and homeostatic mechanisms; feedback mechanisms guide this pursue of viable behavior which is carried by goal-seeking processes. On the other hand, in SD, coming from the servomechanisms thread, feedback loops are taken as intrinsic parts of the system (and not just as mechanisms of control), it includes also loops of positive polarity, and such feedback structures are seen as responsible for counterintuitive behaviors and policy resistance in

social systems; here the analysis is directed toward policy design.

Explanation and Mechanism

The next interesting question would be how we can achieve better understanding and better policy design by enhancing our mental models. How can we characterize this type of knowledge?

A solid account of *explanation* should be placed at the heart of any scientific activity. The general inquiry about [scientific] explanation has to do with "learning how the process of doing science facilitates understanding, and what type(s) of understanding science provides" ([7] p. 307). In a very intuitive way, a first approach to explanation might be associated plainly with removing puzzlement [6]. It is also common to affirm that an explanation aims to answer queries of *why* in order to provide understanding [87]. Yet, to characterize such idea is a major and open unresolved question in philosophy of science [69]. The notion of causality has traditionally played a central role; this view has pervaded most of scientific research where theory development and explanations are essentially conceived as the search for causes, e. g. [50,87]. However, several explanations are not essentially based on simple causal relations but on other approaches such as identification, models, analogies, formal linguistics, laws of association, laws of co-existence, variational principles, among others [83]. Berger [7] underlines the prominence of this question when attempting to characterize the explanations provided by nonlinear dynamical modeling:

> Mathematical modeling [is recognized] as one of the central activities of science, and it is reasonable to say that modeling explanations dramatically increase our understanding of the world. But the modeling explanations found in contemporary scientific research show that the interesting claims of causal accounts are untenable … An adequate account of scientific explanation must accommodate modeling explanations, because they are simply too central to ignore (pp. 329–330).

The main goal of this section is to explore the position and the characterization of the kind of explanation pursued in system dynamics. This characterization fits with the presented core of SD and the presentationalist stance. And again various clarifications will be needed along the way.

Causality

The difficult issue of causality can be treated in several senses. In the first place, a possible relationship associated with the term "determinism" on human behavior is dismissed with a previous argument: SD causal models does not point at supposed laws of causality governing human action [55,56]. What is more interesting is to investigate the concept of causality as such in SD models; after all, a large part of SD modeling relies on what are known as "causal"-loop diagrams. Forrester emphasized the term *interrelationships* [24] where feedback loops are understood as closed informational paths connecting in a sequence decisions that control actions [25]; he labeled it as a "circular cause-and-effect structure" (p. 1–9). In fact the development of *causal*-loop diagrams has become important in SD practice; in particular, flow diagrams were initially recognized as useful pictorial representations that help to formulate and communicate the structure of a dynamic model [24]. These models are ultimately theories of behavior, surely causal theories of behavior. But consistent with presentationalism, these theories are sourced in the mental models of the modeler, there is no direct connection to an alleged causation in a real world. Sterman summarizes a definition: "a causal diagram consists of variables connected by arrows denoting the causal influences among the variables" [98]; here, every link represents *what the modeler believes is* a causal relationship between two variables; this causal attribution is seen as a central feature of mental models, as Sterman also stresses "we all create and update cognitive maps of causal connections among entities and actors … Within a causal field, people use various cues to causality including temporal and spatial proximity of cause and effect, temporal precedence of causes, covariation, and similarity of cause and effect" [98]. Again, the core of the discussion should be driven by the concept of mental model in order to deliver a clear discussion on this view of causality held in SD. This section outlines a framework.

Most of the time we seem to hold a strong causal view of the world. In particular, the causal relationship – whatever that could be – tends to be the source of explanatory power, i. e. the *explanans*, and one usual source of validity, that is, *for having a relevant valid explanation we must have a causal relationship*. That is the usual principle, and it is usually associated to the term "determinism". Recalling the discussion on presentationalism, one should note that – as Hesslow [42] underlines – if we are going to retain a Humean view of the world, i. e. consistent with idealism, then it seems that we have two different paths, probabilistic or deterministic. The latter one is of interest here (for the probability account of causation see e. g. [43]). Within a deterministic approach, a cause is always sufficient condition or a part of a sufficient condition for the effect, that is, if A_t is a nonsufficient cause of $B_{t'}$, then there must be

some auxiliary condition $C_{t''}$, [with $t'' < t'$], such that A_t in combination with $C_{t''}$ is sufficient for $B_{t'}$. Hesslow clarifies this "sufficiency principle" [42]:

> The deterministic approach has been something of a received view of causation. This view, which we may call the 'sufficiency principle' is also common among scientists. The sufficiency principle is not in itself strictly deterministic. It does not mean that every event has a sufficient cause, only that if an event has a cause, then it has a sufficient one. However, it seems that the popularity of the sufficiency principle is a reflection of a widely spread, though usually implicit, commitment to the stronger thesis, that every event has a sufficient cause (p. 592).

The cited paper of Hesslow aims to show that the deterministic approach is superior to the probabilistic one, that is, the idea that the probabilistic account presupposes determinism. Indeed what is happening is the trap of Hume – so to speak. The issue in hands is illustrated with the *Humean fork*: based on observation of constant conjunction of events – altogether with temporal priority, i. e. the cause is observed prior to its effect – we *suppose* a causal connection between them – see summaries in e. g. [46,70] and the original work of Hume [45]. These *suppositions* are arranged in our mental models, that is, in our theories about what we assume as the relevant causal connections that we suppose so as to explain the world.

With this framework in mind, the network of causal connections that the SD modeler believes to be relevant indicates a sort of "sufficiency principle", but of a special kind. It is sufficient relative to the purpose of the model, as it has been indicated above. And it is sourced in the mental model since SD is seen as a vehicle for learning and not as a device for operating a "real world". How are these causal relationships portrayed? In a generic form, a feedback loop based stance is based on the fact that decisions in a time t affect the environment which affects again the future assessment of the situation in a time $t'(t' > t)$ which usually is the base for new upcoming decisions and actions taken in a time t'' $(t'' > t')$ and so on. These relationships are not necessarily close in space or time. Furthermore, a related issue is what can be labeled as "multiple causality". The complexity of social systems is a current concern for social scientists; the multiple interactions among several agents, actors, or entities, is what has been distinguished as the key to study and to understand complex systems because of the recognition of our inability to deal with them based on traditional incomplete simple-causality thinking – the assumption that explanation of phenomena can be satisfactory or even sufficient based in simple unidirectional causal relationships between variables or constructs. What is more, feedback loops have important implications associated with counterintuitive behavior that usually we do not consider easily or that we misunderstand; they have a key role in complex settings; in fact they are, to a large extent, responsible for the arising of complex behaviors; this is a central affirmation in SD [26,98]. The simplest example might be the tendency that we have to infer linear growth from single first order positive feedback loops. But the behavior here actually is exponential. Let S be the state of a system and g the constant fractional growth rate, the linear first order positive loop and its solution are:

$$\frac{\mathrm{d}S}{\mathrm{d}T} = gS$$

and it has as solution:

$$S = S_0 \mathrm{e}^{gt} .$$

The central question is: are these causal theories, portrayed in system dynamics, claims about the world? i. e. Are these models assumed true descriptions of the world? Clearly no. From the presentationalist stance of SD, the causal-loop diagram is essentially what the modeler *believes* is the relevant causal network for the problem in hands. It constitutes his theory about it. The source of knowledge is the mental model and causality is only a supposition of the modeler and a way to arrange knowledge, it is not an affirmation of truth about a supposed causal world. And furthermore, causation as such is not the source of explanation provided by SD. In order to clarify this we should take a look to the notion of explanation held in system dynamics.

Mechanism

System dynamics aims to answer *why* questions. This is done generally via the development of *dynamic hypotheses*. A core premise of SD has always been to enhance learning and to provide understanding [30]. This aim has been stated from the very beginning; for instance Forrester asserted in *Industrial Dynamics*: "Our objective is to enhance understanding and to clarify our thinking about the system" p. 57 in [24]).

How is this goal pursued? A SD model should be able to account for a specific problematic behavior which is explained in terms of the structure of the model; here the term "structure" refers to the stock and flow organization, the feedback loops and the rules of interaction [98]. This approach to explanation is known as a dynamic hypothesis and is the core concept in order to provide understanding from a system dynamics point of view; its endogenous

character is the chief feature that makes it intelligible; for instance: "One key task in this search for insightful, system level understanding is the telling of 'system stories' – coherent, dynamically correct explanations of how influential pieces of system structure give rise to important patterns of system behavior" (p. 1 in [65]). In fact this task represents one of the more significant research lines [79].

How can we characterize this particular kind of explanation? As a first point, the notion of organized social complexity helps to drive this discussion. A quote borrowed from Hayek illustrates it:

> Where we have to deal with such social wholes we cannot, as we do in the natural sciences, start from the observation of a number of instances which we recognize spontaneously by their common sense attributes as instances of 'societies' or 'economies' … What we group together as instances of the same collective or whole are different complexes of individual events, in themselves perhaps quite dissimilar, but believed by us to be related to each other in a similar manner: they are classifications or selections of certain elements of a complex picture on the basis of a theory about their coherence (p. 43 in [40]).

Based on the quote above of Hayek, he suggests conceiving the *explanation* as modeling [104], and for social sciences he rejects the usual prediction and control aspirations and asks the reader to focus more on models to explain typical processes [64], he depicts it with biology: "It deals with pattern-building forces, the knowledge of which is useful for creating conditions favorable to the production of certain kinds of results, while it will only in comparatively few cases be possible to control all the relevant circumstances" (as cited in p. 202 in [104]). Hayek calls it "explanation of the principle". Essentially he means the explanation of a *kind* of phenomena instead of particular events. As another example consider mathematics: "A set of equations which shows merely the form of a system of relationships but does not give the values of the constants contained in it, is perhaps the best general illustration of an explanation merely of the principle on which any phenomenon is produced" (p. 291 in [39]). This analogy illustrates the notion of abstract relations that would build an "explanation of the principle" which can be associated with *mechanism* as the source of explanatory power – instead of causality.

Again, a clarification is needed given the widespread identification of the term "mechanism" with ontic commitments. Fundamentally mechanism is a kind of *explanation*. It should be noticed that "mechanism" refers to *epistemological* issues. However, the term is habitually associated with assumptions about reality. But Hogben already clarified in 1930:

> In any discussion between the two [mechanist and holist or vitalist], the combatants are generally at cross purposes. The mechanist is primarily concerned with an epistemological issue. His critic has always an ontological axe to grind. The mechanist is concerned with how to proceed to a construction which will represent as much about the universe as human beings with their limited range of receptor organs can agree to accept. The vitalist or holist has an incorrigible urge to get behind the limitations of our receptor organs and discover what the universe is really like (1930, p. 100, as cited in [12], p. 347).

The explanatory notion of mechanism is well underlined by Grene: "Let us look for a mechanism which might underlie the phenomena we hope to understand, seeking wherever we may relevant sources from which to derive … an analogue of a possible mechanism … [Such an explanation is of value because it tells us] *how in fact those phenomena are produced*" (as cited in [12], p. 346, emphasis original).

However, there is still no agreement about what a mechanism is and how it appears to succeed in science as a way to provide understanding. Perhaps the most complete account is the one of Tabery [99] who proposes integrating two complementary points of view. These two aspects are (i) the interactions among several parts and, (ii) the activities associated with these interactions. Both characteristic are taken as necessary for having a mechanism-based explanation.

On the one hand there is the emphasis on interactions, a thesis supported by Glennan [36] with a central concern on the nature of *complex systems* since the role of parts and its interactions are conceived as indispensable. The work of Bechtel and Richardson [5] develops the association of mechanism and complex systems (in biology and psychology) and emphasizes the tasks of decomposition and localization as the heuristics in order to uncover mechanisms. This position replies to the conventional view of a mechanism as merely the interactions between causal processes as the essential *explanans*; Glennan [36] stresses: "A mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations" (p. 344). Glennan clarifies also that he avoids using the term "causal-law" and instead uses "change-relating generalizations" because these relations are not exception-less as traditionally a law is understood. In addi-

tion, he emphasizes the very different character of such account which is in opposition to the traditional causal view in which mechanisms are sequences or chains of events leading up to a particular event – which is often associated in systems theory literature with "linear thinking".

On the other hand we have activities. Machamer, Darden and Craver [62] emphasize this aspect adding that mechanisms are not only inter-connected entities but also activities producing regular changes from initial to finish conditions; they call themselves dualists since for them both notions – entities and activities – are necessary to constitute a mechanism: "The organization of these entities and activities determines the ways in which they produce the phenomenon. Entities often must be appropriately located, structured, and oriented, and the activities in which they engage must have a temporal order, rate, and duration" (p. 3). It is important to underline their critique to Glennan's view arguing that the concept of *activity* is fundamental to understand the changes produced (because of the activities) through the process and not only as the black-box view of change of states or change of properties of the inter-connected entities; they picture it clearly with the following statement: "it is not the penicillin that causes the pneumonia to disappear, but what the penicillin does" (p. 6). Furthermore, in order to account for a mechanism they emphasize three distinctions: set-up conditions (as *part* of the mechanism, not as a sort of input; this includes relevant entities and their properties and initial states), intermediate activities (including also relevant entities, properties, and an intelligible account of the activities that link them) and termination conditions (such as privileged endpoints, equilibrium states or the final stage of some unitary integral process). They also draw attention to the fact that mechanisms take place in nested multi-level hierarchies and that they usually are not full pictures but truncated abstract accounts – *a mechanism schema* – depending on the required level of detail or aggregation.

System dynamics modeling is perhaps one of the best ways to picture this kind of explanation. One can distinguish two main components in the structure of SD models: the physical and institutional assumptions – including the chosen parts/variables and the interconnections between them, and the decision rules of the agents [98]. The interplay between the physical structure and the associated decision rules as the explanation for behavior is a foundational aspect. Indeed the interconnections and the activities needed to account for a mechanism are included in these system dynamic models structures. Specifically, the activities producing change can be referenced in the links of the models and the decision rules describe how the interactions produce certain activities. The whole set of initial and final conditions and the inter-connected parts engaged in producing activities characterize a mechanism. For example, the simplest mechanism is perhaps a single feedback loop. The set-up conditions are the initial values of the variables involved, the termination condition is the endpoint of the loop which can be accounted in a mechanism as "the final stage of what is identified as a unitary, integral process" (p. 12 in [62]). The intermediate activities are depicted by the links and the application of the decision rules. For instance, a simple positive first order loop can produce exponential behavior beyond the particular values of the variables involved. More intricate structures are the source of different behaviors, i. e. the change in the values or patterns of variables through time.

This stance fits an explanation of an abstract principle in the sense of Hayek. The structure is the source of explanation of patterns of behavior, i. e. the change in the values or patterns of variables of interest through time. This is known in SD as a dynamic hypothesis [98]. With this focus on aggregate patterns – instead of individual events – as consequence of the structure, it can be said that the "causal" mechanism is indeed the loop structure of the system, or the particular and relevant feedback substructures of the model that may explain the behavior; for instance Richardson [78] illustrates it in this way: "The 'cause' of an arms race is viewed not as a given event or even a given sequence of events, but as a feedback structure dominated by self-reinforcing positive loops, within which events take place" (p. 338). These types of explanations are based on *mechanisms* as explanatory power and not in simple causal relationships as the source of explanation, even less in (substantivalist) causality, i. e. change in singular properties/entities. Furthermore, these hypotheses are developed for each problem consistent with the mental models of the modelers. This is why system dynamics is not committed to specific theories and only to the explanation of problematic behaviors in terms of structure of the model in order to enhance learning and decision-making.

A particular remark must be made. It can be noticed a natural link between mechanism and the idea of "generic structure". This latter expression has been used in different senses in SD literature. Lane and Smart [58] trace the evolution of this concept – see also [71], they identify three different interpretations. One of these view cannot be connected with mechanism, the one popularized by Senge, e. g. [92,109,110], usually known under the expression "system archetypes". The lack of computer simulation within this interpretation points at a problem of validity in its scope and claims, as Lane and Smart dis-

cuss [58], e. g. since this perspective skips the possibility of formal computer model building then the relation between structure and behavior is weak and the mentioned approach becomes just a hasty shortcut from problematic behavior to insights and principles, and without the experimental spirit of SD for enhancing learning. But two other notions are relevant. On the one hand, generic structures can be conceived as general models (theories of behavior) of a class of systems that are associated with a domain of application, e. g. urban development, supply chain, economic growth. Lane and Smart label these structures as "canonical situation models". On the other hand, a generic structure may refer to theories of mathematical structures (feedback loops, levels, rate equations, etc.) that generate corresponding dynamic behaviors, i. e. "systems belong to the same class if they can be represented by the same structure … This dynamic structure when abstracted from any application domain data defines the class of system" (p. 93 in [58]); therefore they offer transferability of structure across diverse domains. These models can be labeled as "abstracted micro-structures", e. g. patterns of exponential growth, goal seeking, oscillation, etc. [58,98]. Both interpretations look for establishing a general class of models that formally link structure with behavior and they constitute an important line of research. These developments contribute to different aspects of SD practice; in particular they directly fuel the processes of conceptualization and formal model construction (e. g. see [31,58,98]), and more important, they enhance understanding and the improvement of our mental models as long as we can exploit the powerful idea of having general classes of models, either within a domain of application or across different domains by transferring structures across them.

Consequently, system dynamics explanations can be characterized as *mechanisms*, since there can be found the source of explanatory power. In spite of its causal diagrams, the *explanans*, i. e. that which does the explaining, is based on mechanisms – dynamic hypotheses based on structures; and the problematic behavior is the *explanandum*, i. e. that which is explained. Following Glymour: "Remains, however, a considerable bit of science that sounds very much like explaining, and which perhaps has causal implications, but which does not seem to derive its point, its force, or its interest from the fact that it has something to do with causal relations (or their absence)" (as cited in [p. 212 in 83]). The theories built with SD are essentially structure-based and not content-based (substantivalist) explanations i. e. they are not associated with intrinsic properties of objects or entities but with the consequences of processes and activities entrenched in relevant parts of the structure of the model. Recalling Hayek who identifies explanation with modeling, it is interesting to notice the range of his thought expressed half a century ago and that accurately illustrates SD explanations:

> Any model defines a certain range of phenomena which can be produced by the type of situation which it represents. We may not be able directly to confirm that the causal mechanism determining the phenomenon in question is the same as that of the model. But we know that, if the mechanism is the same, the observed structures must be capable of showing some kinds of action and unable to show others (p. 221 in [38]).

A further reminder follows. Since SD was previously identified with instrumentalism, then a mechanism is not to be taken as a description; here a mechanism is an instrument for arranging observations. It should be kept in mind that mechanism is a kind of explanation which refers to epistemological issues.

The popularity of simple causality as the way to characterize the explanation of phenomena contrasts with the assumptions made in SD: structures that generate processes responsible for behavior. This is consistent with the purpose of system dynamics simulation which might be oriented to activities such as theoretical-representations building, articulation and testing in order to learn in and about complex systems [98]. System dynamics uses simulation as a method which is different from the traditional inductive logic of research that deals with single instances which attempts to confirm theories via repeated observation. However, SD does not dismiss presentationalism as it was shown. This should be highlighted as an important and distinctive characteristic of SD. Though there is a commitment with a real world, justification is rooted in social processes and on mental models, and it is also relative to the purpose of the model. Furthermore, the goal is to enhance our decision-making processes by improving our mental models. How can we characterize such method? A short comment follows.

Simulation and Method

In a plain sense "simulation means driving a model of a system with suitable inputs and observing the corresponding outputs" (p. 23 in [2]). But simulation actually is not just a matter of number crunching. Its scope is broader. And it represents another challenge for philosophy of science. Winsberg [107] illustrates it:

> Typically, to a philosopher of science, epistemological issues arise when we try to justify high level the-

> oretical claims based on low level data or specific observational reports. But simulation is about starting with theory and working your way down. This kind of epistemology is, to the philosopher of science, a curious beast. It is an epistemology that is concerned with justifying inferences from a theory to its application – an inference that most philosophy of science has assumed is deductive and consequently not in need of justification (p. S447).

What is simulation in SD? It can be affirmed that it is a technique able to represent and test theoretical concepts and not only – in a narrow sense – a tool to just solve mathematical problems. Besides, the emphasis on processes, on patterns of collective action and on the relations between components and its dynamic consequences can be better addressed with simulation because of its capacities to represent these issues with fewer restrictions than other approaches [35]. But there is more. Simulation reflects a very different attitude. This way of inquiry suggests a whole different and new scientific methodology [82,106,107]. Winsberg emphasizes that "simulation represents an entirely new mode of scientific activity – one that lies between theory and experiment … a form of theory articulation or 'model building' (pp. 117, 119 in [106]). Axelrod [2] indeed suggests "a third way" of doing science:

> Simulation as a way of doing science can be contrasted with the two standard methods of induction and deduction … Simulation is a third way of doing science. Like deduction, it starts with a set of explicit assumptions. But unlike deduction, it does not prove theorems. Instead, a simulation generates data that can be analyzed inductively. Unlike typical induction, however, the simulated data comes from a rigorously specified set of rules rather than direct measurement of the real world (p. 24).

Moreover, its strength rests on the capacity for conducting *experiments* [82]. This emphasis on experimentation is the key to understand why this approach is different. Our mental models are nothing more than theoretical models that attribute properties and relations to the systems they represent; the relevance of these theoretical models depends on the purpose of the model. And computer simulation simply permits experiments of these (theoretical) models. This is where the novelty and the power of this methodology are to be found, in the very iterative process of model building and experimentation via simulation. This position contrasts with the traditional prominence of assumed representational capacities of *theories* and *models* where usually the emphasis has been placed, see [107].

But computer simulation has a distinct epistemology [105] that emphasizes *the process of modeling*. The method was demarcated by Forrester [24] in *Industrial Dynamics*:

> Simulation consists of tracing through, step by step, the actual flows of orders, goods, and information, and observing the series of new decisions that take place … This is the counterpart of trying a new policy or organizational structure in the real system… After a simulation run comes interpretation of the results. Did it turn out as expected? If not, why? As the experiment is examined, new questions arise … This is a process of invention and trial … Each simulation result teaches, and it also prompts additional questions … Such experimentation will yield new insights into the characteristics of the system that the model represents (pp. 23, 44–45, 55).

Hence, the method of simulation through continued experimentation is aimed at providing better understanding of the modeled system. As it was mentioned, the method calls attention to the *process* – see also [44,89]. Indeed SD aims at developing a *modeling* culture (consistent with [80]) that gives emphasis to model building as an ongoing dialectic between stakeholders instead of a mapping exercise concerned on the efficacy of the model itself.

Why is fundamental the use of the computer? Perhaps the best answer is provided again by Forrester [26]:

> We stress the importance of being explicit about assumptions and interrelating them in a computer model … The most important difference between the properly conceived computer model and the mental model is in the ability to determine the dynamic consequences when the assumptions within the model interact with one another. The human mind is not adapted to sensing correctly the consequences of a mental model … The computer model … is a statement of system structure. It contains the assumptions being made about the system … Generally, the consequences are unexpected (pp. 213–215).

The shortcomings of our mental models coupled with the complexity of the systems we model lead to the use of the computer. Sterman [101] summarizes these drawback with aspects such our flawed cognitive maps and our erroneous inferences about dynamics. As it was shown, the strength of explanation and understanding in SD is not in the causal models as such; the heart lies in the development of dynamic hypotheses with the use of simulation in order to enhance our understanding and our decision-mak-

ing processes. Explanations are posed under the notion of mechanism; but this is an iterative process that seeks a central aim: to improve our own theories about the world.

Future Directions

There are several aspects to develop based on this reflection on various assumptions behind system dynamics. The central argument was built around the position known as presentationalism. This stance integrates and informs many of the debates on related subjects, e. g. validation, and it characterizes the initial purposes and assumptions held in the field. However, there are a number of lines to emphasize and to develop.

We have focused on the role of the idea of "mental model" for the practice of SD. Yet, it can be seen utilization of SD models under assumptions which are closer to a naive realism that seem to ignore the purposes of enhancing understanding and learning processes. This article should help to underscore that system dynamics is less naive – and hence more powerful – when we recognize a presentationalist stance which means that our theories about the world are just that, *theories* based on our experience and on the operations of our own mind. We can improve these theories with the use of system dynamics, that is, making our assumptions about systems explicit and using simulation as a method for enhancing understanding and for developing explanations that guide our processes of systems design. This recognition includes the relative nature of justification of knowledge held in SD and the emphasis on mechanism as a powerful way to develop explanations about complex systems.

It has been introduced mechanism as the way to characterize the particular type of explanation pursued in SD. The explanatory force does not rest on causal relations as such but on the structures – physical and decision rules aspects – and on the dynamic processes and activities that explain change. The idea of mechanism is shaped in SD under the expression "generic structure". The focus on understanding behavior in terms of abstract structures is a central line of inquiry. This article underlines the importance of developing such a line of research since it has been located at the core of the kind of knowledge that SD provides. The issue of unification to provide understanding of diverse phenomena is a definitive step in the way to assert that the field progresses as long as broader range of phenomena may be explained with the same mechanism. Should be the advance of system dynamics assessed by the progress in this type of study? Behind this discussion there are major and provocative issues that arise to be developed.

The long debate of qualitative and quantitative modeling is informed by this characterization. It is clear that the issue of explanation compels theorists and practitioners to ask themselves what is the kind of explanation they are pursuing and if such explanations are enough and satisfying; it is worthy to ask for qualitative modeling what kind of understanding it gives and how it can be characterized, in other words, to give an account of *explanans* and *explanandum*. Another line to develop might be oriented around the following question: what would be *essential* criteria for comparing different arranges or modes of organization in order to identifying them as belonging to the same type of mechanism? The identification and search of mechanisms becomes a powerful heuristic for guiding the modeling process.

There are promising suggestions for philosophy of science as well. SD offers guidelines, e. g. can the ways in which system dynamicists work provide meaningful insights, or even concrete accounts, for the philosophical unresolved issue of explanation? The mechanism depicted in system dynamics proposes a kind of explanation that goes beyond the received view based on causation. A related question is whether explanation must always follow a deductive path; the classic models of Hempel, e. g. [41], emphasized the condition of deduction and general laws for having an explanation; however, the explanation in SD is not framed under a deductive schema from universal covering laws, instead it can be conceived as a sort of abductive reasoning based on the understanding of the dynamics of the model as a way to understand the actual behavior it accounts for. A further issue is that in spite of the lack of universality, i. e. no universal laws, SD models aim to provide understanding for a diverse range of phenomena that might share relevant influential structures and similar associated behaviors, that is, it accounts for regularities in order to unify them in a certain kind of explanation under the same explanation. This situation insinuates a flavor of paradox because of the traditional rigid association of universal laws with the explanation of regularities, but in SD there are no general laws though the aim is to explain general regularities. The study of *models* and computer simulation – instead of abstract theories and traditional methodologies – is an additional indication that SD suggests for philosophy of science, including the emphasis on the modeling process.

Bibliography

1. Aronson JL (1988) Testing for Convergent Realism. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol One: Contributed Papers. University of Chicago Press, Chicago, pp 188–193

2. Axelrod R (1997) Advancing the art of simulation in the social science. In: Conte R, Hegselmann R, Terna P (eds) Simulating Social Phenomena. Springer, Berlin, pp 21–40
3. Barlas Y, Carpenter S (1990) Philosophical roots of model validation: two paradigms. Syst Dyn Rev 6:148–166
4. Bartley III WW (1987) Philosophy of biology versus philosophy of physics. In: Radnitzky G, Bartley III WW (eds) Evolutionary epistemology, rationality, and the sociology of knowledge. Open Court, La Salle, pp 7–45
5. Bechtel W, Richardson RC (1993) Discovering complexity: Decomposition and localization as strategies in scientific research. Princeton University Press, Princeton
6. Benjamin AC (1941) Modes of scientific explanation. Philos Sci 8:486–492
7. Berger R (1998) Understanding science: Why causes are not enough. Philos Sci 65:306–332
8. Berkeley G (1948–1957) The works of George Berkeley, Bishop of Cloyne. Thomas Nelson and Sons, London
9. Black M (1934) The principle of verifiability. Analysis 2:1–6
10. Blackmore J (1979) On the inverted use of the terms 'Realism' and 'Idealism' among scientists and historians of science. Br J Philos Sci 30:125–134
11. Bowman AA (1916) Kant's phenomenalism in its relation to subsequent metaphysics. Mind, New Ser 25:461–489
12. Brandon RN (1984) Grene on mechanism and reductionism: more than just a side issue. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol II: Symposia and Invited Papers. University of Chicago Press, Chicago, pp 345–353
13. Brown HI (1990) Cherniak on scientific realism. Br J Philos Sci 41:415–427
14. Burrell G, Morgan G (1979) Sociological paradigms and organizational analysis. Heinemann, London
15. Chapin JP (1941) Idealism and its relation to science. Philosophy of Science 8:142–146
16. Checkland P, Pidd M, Morecroft J (2004) Working ideas, insights for systems modeling: The Broader community of systems thinkers. In: Kennedy M, Winch G, Langer R, Rowe J, Yanni J (eds) Proceedings of the 22nd International Conference of the System Dynamics Society, Keble College, University of Oxford, England. System Dynamics Society, Albany
17. Coyle G (2000) Qualitative and quantitative modelling in system dynamics: some research questions. Syst Dyn Rev 16:225–244
18. Coyle RG (1979) Management system dynamics. Wiley, Chichester
19. Deetz S (1996) Describing differences in approaches to organization science: Rethinking Burrell and Morgan and their legacy. Organ Sci 7:191–207
20. Dickson M (1995) An empirical reply to empiricism: Protective measurements opens the door for quantum realism. Philos Sci 62:122–140
21. Doyle JK, Ford DN (1998) Mental models concepts for system dynamics research. Syst Dyn Rev 14:3–29
22. Doyle JK, Ford DN (1999) Mental models concepts revisited: some clarifications and a reply to Lane. Syst Dyn Rev 15:411–415
23. Flood R, Jackson M (1991) Creative problem solving. Wiley, Chichester
24. Forrester JW (1961) Industrial Dynamics. Press MIT, Cambridge
25. Forrester JW (1971) Principles of Systems. Wright-Allen Press, Cambridge
26. Forrester JW (1975) Counterintuitive behavior of social systems. In: Collected papers of Jay W. Forrester. Wright-Allen Press, Cambridge, pp 211–244
27. Forrester JW (1975) Industrial Dynamics: A Major breakthrough for decision makers. In: Collected papers of Jay W Forrester. Wright-Allen Press, Cambridge, pp 1–29
28. Forrester JW (1975) Industrial Dynamics – After the first decade. In: Collected papers of Jay W. Forrester. Wright-Allen Press, Cambridge, pp 133–150
29. Forrester JW (1985) "The" model versus a modeling "process". Syst Dyn Rev 1:133–134
30. Forrester JW (1987) Lessons from system dynamics modeling. Syst Dyn Rev 3:136–149
31. Forrester JW (1994) System dynamics, systems thinking, and soft OR. Syst Dyn Rev 10:245–256
32. Forrester JW (2003) Dynamic models of economic systems and industrial organizations. Syst Dyn Rev 19:331–345
33. Fuller S (1994) Retrieving the point of the realism-instrumentalism debate: Mach vs. Planck on science education policy. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol One: Contributed Papers. University of Chicago Press, Chicago, pp 200–208
34. Gershenson C, Aerts D, Edmonds B (2007) Worldviews, science and us. Philosophy and complexity. World Scientific, Singapore
35. Gilbert N, Troitzsch KG (1999) Simulation for the social scientist. Open University Press, Buckingham
36. Glennan SS (2002) Rethinking mechanistic explanation. Philos Sci 69:342–353
37. Guyer P (1983) Kant's intentions in the refutation of idealism. Philos Rev 92:329–383
38. Hayek FA (1955) Degrees of explanation. Br J Philos Sci 6:209–225
39. Hayek FA (1942) Scientism and the study of society, Part I. Economica, New Ser 9:267–291
40. Hayek FA (1943) Scientism and the study of society, Part II. Economica, New Ser 10:34–63
41. Hempel CG, Oppenheim P (1948) Studies in the logic of explanation. Philos Sci 15:135–175
42. Hesslow G (1981) Causality and determinism. Philos Sci 48:591–605
43. Hitchcock CR, Salmon WC (2000) Statistical explanation. In: Newton-Smith WH (ed) A Companion to the philosophy of science. Blackwell Publishers, Malden, pp 470–479
44. Homer JB (1996) Why we iterate: scientific modeling in theory and practice. Syst Dyn Rev 12:1–19
45. Hume D (1740) A treatise of human nature. Oxford University Press, Oxford
46. Humphreys P (2000) Causation. In: Newton-Smith WH (ed) A companion to the philosophy of science. Blackwell Publishers, Malden, pp 31–40
47. Hunter B (1992) Empiricism. In: Dancy J, Sosa E (eds) A Companion to Epistemology. Blackwell Publishers, Oxford, pp 110–115
48. Jackson M (1991) Systems methodology for the management sciences. Plenum Press, New York
49. Jackson M (2003) Systems thinking: Creative holism for managers. Wiley, Chichester

50. Jobe EK (1985) Explanation, causality, and counterfactuals. Philos Sci 52:357–389
51. Kleindorfer GB, Ganeshan R (1993) The philosophy of science and validation in simulation. In: Evans GW, Mollaghasemi M, Russell EC, Biles WE (eds) Proceedings of the 1993 Winter Simulation Conference. IEEE, Piscataway, pp 50–57
52. Kukla A (1995) Scientific realism and theoretical unification. Analysis 55:230–238
53. Lane D (1994) With a little help from our friends: How system dynamics and soft OR can learn from each other. Syst Dyn Rev 10:101–134
54. Lane D (1999) Friendly amendment: A commentary on Doyle and Ford's proposed re-definition of 'mental model'. Syst Dyn Rev 15:185–194
55. Lane D (2000) Should system dynamics be described as a 'Hard' or 'Deterministic' systems approach? Syst Res Behav Sci 17:3–22
56. Lane D (2001) Rerum cognoscere causas: Part I – How do the ideas of system dynamics relate to traditional social theories and the voluntarism/determinism debate? Syst Dyn Rev 17:97–118
57. Lane D (2001) Rerum cognoscere causas: Part II – Opportunities generated by the agency/structure debate and suggestions for clarifying the social theoretic position of system dynamics. Syst Dyn Rev 17:293–309
58. Lane D, Smart C (1996) Reinterpreting 'generic structure': evolution, application and limitations of a concept. Syst Dyn Rev 12:87–120
59. Leplin J (1992) Realism and Methodological Change. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol Two: Symposia and Invited Papers. University of Chicago Press, Chicago, pp 435–445
60. Leplin J (2000) Realism and Instrumentalism. In: Newton-Smith WH (ed) A Companion to the Philosophy of Science. Blackwell Publishers, Malden, pp 393–401
61. Luna-Reyes LF, Andersen DL (2003) Collecting and analyzing qualitative data for system dynamics: methods and models. Syst Dyn Rev 19:271–296
62. Machamer P, Darden L, Craver CF (2000) Thinking about mechanisms. Philos Sci 67:1–25
63. Meadows DH (1980) The Unavoidable A Priori. In: Randers J (ed) Elements of the system dynamics method. Productivity Press, Cambridge, pp 23–57
64. Milford K (1994) In pursuit of rationality. A note on Hayek's The Counter-Revolution of Science. In: Birner J, van Zijp R (eds) Hayek, Co-ordination and Evolution. Routledge, London, pp 323–340
65. Mojtahedzadeh M, Andersen D, Richardson GP (2004) Using digest to implement the pathway participation method for detecting influential system structure. Syst Dyn Rev 20:1–20
66. Morrison M (1988) Reduction and realism. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol One: Contributed Papers. University of Chicago Press, Chicago, pp 286–293
67. Morrison M (1990) Unification, realism and inference. Br J Philos Sci 41:305–332
68. Newton-Smith WH (1988) Modest realism. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol Two: Symposia and Invited Papers. University of Chicago Press, Chicago, pp 179–189
69. Newton-Smith WH (2000) Explanation. In: Newton-Smith WH (ed) A companion to the philosophy of science. Blackwell Publishers, Malden/Oxford, pp 127–133
70. Newton-Smith WH (2000) Hume. In: Newton-Smith WH (ed) A companion to the philosophy of science. Blackwell Publishers, Malden, pp 165–168
71. Paich M (1985) Generic structures. Syst Dyn Rev 1:126–132
72. Pettit P (1992) Realism. In: Dancy J, Sosa E (eds) A companion to epistemology. Blackwell Publishers, Oxford, pp 420–424
73. Popper K (1963) Conjectures and refutations. The growth of scientific knowledge. Routledge and Kegan Paul, London
74. Popper K (1968) The Logic of Scientific Discovery. Hutchinson, London
75. Psillos S (1996) Scientific realism and the "pessimistic induction". Proceedings of the Biennial Meeting of the Philosophy of Science Association, Part I: Contributed Papers. University of Chicago Press, Chicago, pp S306-S314
76. Radnitzky G (1987) In defense of self-applicable critical rationalism. In: Radnitzky G, Bartley III WW (eds) Evolutionary epistemology, rationality, and the sociology of knowledge. Open Court, La Salle, pp 279–312
77. Rescher N (1992) Idealism. In: Dancy J, Sosa E (eds) A companion to epistemology. Blackwell Publishers, Oxford, pp 187–191
78. Richardson GP (1991) Feedback thought in social science and systems theory. Pegasus Communications, Waltham
79. Richardson GP (1996) Problems for the future of system dynamics. Syst Dyn Rev 12:141–157
80. Richardson KA (2002) On the limits of bottom-up computer simulation: Towards a nonlinear modeling culture. In: Sprague RH Jr (ed) Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03). IEEE
81. Roberts EB (1978) System dynamics – An introduction. In: Roberts EB (ed) Managerial applications of system dynamics. Pegasus Communications Inc., Waltham
82. Rohrlich F (1990) Computer simulation in the physical sciences. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol Two: Symposia and Invited Papers. University of Chicago Press, Chicago, pp 507–518
83. Ruben D (1990) Explaining explanation. Routledge, London
84. Runciman WG (1969) What is structuralism? Br J Sociol 20:253–265
85. Russell B (1912) The problems of philosophy. Oxford University Press, Oxford
86. Russell B (1948) Human knowledge: Its scope and limits. Simon and Schuster, New York
87. Salmon W (1992) Explanation. In: Dancy J, Sosa E (eds) A companion to epistemology. Blackwell Publishers, Oxford, pp 129–132
88. Salmon WC (2000) Logical Empiricism. In: Newton-Smith WH (ed) A companion to the philosophy of science. Blackwell Publishers, Malden, pp 233–242
89. Schaffernicht M (2006) Detecting and monitoring change in models. Syst Dyn Rev 22:73–88
90. Schlagel RH (1981) Contextualistic realism. Philos Phenomenol Res 41:437–451
91. Sellars W (1976) Is scientific realism tenable? PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol Two: Symposia and Invited Papers. University of Chicago Press, Chicago, pp 307–334

92. Senge P (1990) The fifth discipline: The art and practice of the learning organization. Doubleday, New York
93. Sidgwick H, Caird E (1880) Kant's refutation of idealism. Mind, New Ser 5:111–115
94. Sinha D (1963) Phenomenology and positivism. Philos Phenomenol Res 23:562–577
95. Smith DW (1982) The realism in perception. Noûs 16:42–55
96. Spender J-C (1996) Making knowledge the basis of a dynamic theory of the firm. Strateg Manag J 17:45–62
97. Stace WT (1944) Positivism. Mind, New Ser 53:215–237
98. Sterman J (2000) Business dynamics. Systems thinking and modeling for a complex world. McGraw-Hill, Boston
99. Tabery JG (2004) Synthesizing activities and interactions in the concept of a mechanism. Philos Sci 71:1–15
100. Taube M (1937) Positivism, science, and history. J Philos 34:205–210
101. Turbayne CM (1955) Kant's refutation of dogmatic idealism. Philos Q 5:225–244
102. Vanderminden P (2006) System dynamics – A field of study, a methodology or both. In: Größler A, Rouwette E, Langer R, Rowe J, Yanni J (eds) 24th International Conference of The System Dynamics Society. Radboud University Nijmegen, System Dynamics Society, Albany
103. Vásquez M, Liz M, Aracil J (1996) Knowledge and reality: some conceptual issues in system dynamics modeling. Syst Dyn Rev 12:21–37
104. Weimer W (1999) Hayek's approach to the problems of complex phenomena: an introduction to the theoretical psychology of The Sensory Order. In: Boettke P (ed) The Legacy of Friedrich von Hayek, vol II. Edward Elgar Publishing Limited, Cheltenham, pp 200–244
105. Winsberg E (1999) Sanctioning models: The epistemology of simulation. Sci Context 12:275–293
106. Winsberg E (2003) Simulated experiments: Methodology for a virtual world. Philos Sci 70:105–125
107. Winsberg E (2001) Simulations, models, and theories: Complex physical systems and their representations. Philos Sci 68:S442–S454
108. Wolstenholme EF (1990) System enquiry. Wiley, Chichester
109. Wolstenholme EF (2003) Towards the definition and use of a core set of archetypal structures in system dynamics. Syst Dyn Rev 19:7–26
110. Wolstenholme EF (2004) Using generic system archetypes to support thinking and modelling. Syst Dyn Rev 20:341–356

System Regulation and Design, Geometric and Algebraic Methods in

Alberto Isidori[1], Lorenzo Marconi[2]
[1] Department of Informatica e Sistemistica, University of Rome, La Sapienza, Italy
[2] Center for Research on Complex Automated Systems (CASY), University of Bologna, Bologna, Italy

Article Outline

Glossary
Definition
Introduction
The Generalized Tracking Problem
The Steady-State Behavior of a System
Necessary Conditions for Output Regulation
Sufficient Conditions for Output Regulation
Future Directions
Bibliography

Glossary

Exosystem A dynamical system modeling the set of all exogenous inputs (command/disturbances) affecting a controlled plant.

Internal model A model of the exogenous inputs (command/disturbances) affecting a controlled plant, embedded in the interior of the controller.

Generalized tracking problem The problem of designing a controller able to asymptotically track/reject any exogenous command/disturbance in a fixed set of functions.

Observer A device designed to asymptotically track the state of a dynamical system on the basis of measured observations.

Steady state A family of behaviors, in a dynamical system, that are asymptotically approached, as actual time tends to infinity or as initial time tends to minus infinity.

Definition

A central problem in control theory is the design of feedback controllers so as to have certain outputs of a given plant *to track* prescribed reference trajectories. In any realistic scenario, this control goal has to be achieved in spite of a good number of phenomena which would cause the system to behave differently than expected. These phenomena could be endogenous, for instance parameter variations, or exogenous, such as additional undesired inputs affecting the behavior of the plant. In numerous design problems, exogenous inputs are not available for measurement, nor are known ahead of time, but rather can only be seen as unspecified members of a given family of functions. Embedding a suitable "internal model" of such a family in the controller is a design strategy that has proven to be quite successful in handling uncertainties in the controlled plant as well as in the exogenous inputs.

Introduction

The problem of controlling the output of a system so as to achieve asymptotic tracking of prescribed trajectories and/or asymptotic rejection of disturbances is a central

problem in control theory. There are essentially three different possibilities to approach the problem: tracking by dynamic inversion, adaptive tracking, tracking via internal models. Tracking by *dynamic inversion* consists in computing a precise initial state and a precise control input (or equivalently a reference trajectory of the state), such that, if the system is accordingly initialized and driven, its output exactly reproduces the reference signal. The computation of such control input, however, requires "perfect knowledge" of the entire trajectory which is to be tracked as well as "perfect knowledge" of the model of the plant to be controlled. Thus, this type of approach is not suitable in the presence of large uncertainties on plant parameters as well as on the reference signal. *Adaptive* tracking consists in tuning the parameters of a control input computed via dynamic inversion in such a way as to guarantee asymptotic convergence to zero of a tracking error. This method can successfully handle parameter uncertainties, but still presupposes the knowledge of the entire trajectory which is to be tracked (to be used in the design of the adaptation algorithm) and therefore an approach of this kind is not suited to the problem of tracking unknown trajectories. Of course, one might consider the problem of tracking a slowly varying reference trajectory as a stabilization problem in the presence of a slowly varying unknown parameter, but this would, in most cases, yield a very conservative solution.

In most cases of practical interest, the trajectory to be tracked (or the disturbance to be rejected) is not available for measurement. Rather, it is only known that this trajectory is simply an (undefined) member in a set of functions, for instance the set of all possible solutions of an ordinary differential equation. These cases include the classical problem of the set point control, the problem of active suppression of harmonic disturbances of unknown amplitude, phase and even frequency, the synchronization of nonlinear oscillations, and similar others. It is in these cases that *tracking via internal models* proves particularly efficient, in its ability to handle simultaneously uncertainties in plant parameters as well as in the trajectory which is to be tracked.

For linear multivariable systems, tracking problems of this kind (those in which the exogenous commands and/or disturbances are only known to be members in the set of solutions of a given ordinary differential equations) have been addressed in very elegant geometric terms by Davison, Francis, Wonham [8,10,11] and others. In particular, one of the most relevant contributions of [11] was a clear delineation of what is known as *internal model principle*, i. e. the fact that the property of perfect tracking is insensitive to plant parameter variations "only if the controller utilizes feedback of the regulated variable, and incorporates in the feedback path a suitably reduplicated model of the dynamic structure of the exogenous signals which the regulator is required to process". Conversely, in a stable-closed loop system, if the controller utilizes feedback of the regulated variable and incorporates an internal model of the exogenous signals, the output regulation property is insensitive to plant parameter variations.

A nonlinear enhancement of this theory, which uses a combination of geometry and nonlinear dynamical systems theory, was initiated in [15,16,19] in the context of solving the problem near an equilibrium, in the presence of exogenous signals which were produced by a Poisson stable system. In particular, in [19] it was shown how the use center manifold theory near an equilibrium determines the necessity of the existence of an internal model whenever one can solve an output regulation problem in spite of (small) parameter uncertainties (see also [3]). Since these pioneering contributions, the theory has experienced a tremendous growth, culminating in the development of design methods able to handle the case of parametric uncertainties affecting the autonomous (linear) system which generates the exogenous signals (such as in [9,23]), the case of nonlinear exogenous systems (such as in [5]), or a combination thereof (as in [22]). The purpose of these notes is to present a self-contained exposition of the fundamentals of these design methods.

The Generalized Tracking Problem

In this article, we address tracking problems that can be cast in the following terms. Consider a finite-dimensional, time-invariant, nonlinear system described by equations of the form

$$\begin{aligned} \dot{x} &= f(w, x, u) \\ e &= h(w, x) \\ y &= k(w, x)\,, \end{aligned} \tag{1}$$

in which $x \in \mathbb{R}^n$ is a vector of state variables, $u \in \mathbb{R}^m$ is a vector of inputs used for *control* purposes, $w \in \mathbb{R}^s$ is a vector of inputs which cannot be controlled and include *exogenous* commands, exogenous disturbances and model uncertainties, $e \in \mathbb{R}^p$ is a vector of *regulated* outputs which include tracking errors and any other variable that needs to be steered to 0, $y \in \mathbb{R}^q$ is a vector of outputs that are available for *measurement* and hence used to feed the device that supplies the control action. The problem is to design a controller, which receives $y(t)$ as input and produces $u(t)$ as output, able to guarantee that, in the re-

sulting closed-loop system, $x(t)$ remains bounded and

$$\lim_{t\to\infty} e(t) = 0\,, \tag{2}$$

regardless of what the exogenous input $w(t)$ actually is.

The ability to successfully address this problem very much depends on how much the controller is allowed to know about the exogenous disturbance $w(t)$. In the ideal situation in which $w(t)$ is available to the controller in real-time, the design problem indeed looks much simpler. This is, though, only an extremely optimistic situation which does not represent, in any circumstance, a realistic scenario. The other extreme situation is the one in which nothing is known about $w(t)$. In this, pessimistic, scenario the best result one could hope for is the fulfillment of some prescribed ultimate bound for $|e(t)|$, but certainly not a sharp goal such as (2). A more comfortable, intermediate, situation is the one in which $w(t)$ is only known *to belong to a fixed family* of functions of time, for instance the family of all solutions obtained from a fixed ordinary differential equation of the form

$$\dot{w} = s(w) \tag{3}$$

as the corresponding initial condition $w(0)$ is allowed to vary on a prescribed set. This situation is in fact sufficiently distant from the ideal but unrealistic case of perfect knowledge of $w(t)$ and from the realistic but conservative case of totally unknown $w(t)$. But, above all, this way of thinking at the exogenous inputs covers a number of cases of major practical relevance. There is, in fact, abundance of design problems in which parameter uncertainties, reference commands and/or exogenous disturbances can be modeled as functions of time that satisfy an ordinary differential equation.

The control law is to be provided by a system modeled by equations of the form

$$\begin{aligned} \dot{\xi} &= \varphi(\xi, y) \\ u &= \gamma(\xi, y) \end{aligned} \tag{4}$$

with state $\xi \in \mathbb{R}^\nu$. The initial conditions $x(0)$ of the *plant* (1), $w(0)$ of the *exosystem* (3) and $\xi(0)$ of the *controller* (4) are allowed to range over fixed *compact* sets $X \subset \mathbb{R}^n$, $W \subset \mathbb{R}^s$ and, respectively, $\Xi \subset \mathbb{R}^\nu$. All maps characterizing the model of the controlled plant, of the exosystem and of the controller are assumed to be sufficiently differentiable.

The problem which will be studied, known as the *generalized tracking problem* (or *problem of output regulation* or also *generalized servomechanism problem*) is to design a feedback controller of the form (4) so as to obtain a closed loop system in which all trajectories are bounded and the regulated output $e(t)$ asymptotically decays to 0 as $t \to \infty$. More precisely, it is required that the composition of (1), (3) and (4), that is the *autonomous* system

$$\begin{aligned} \dot{w} &= s(w) \\ \dot{x} &= f(w, x, \gamma(\xi, k(w, x))) \\ \dot{\xi} &= \varphi(\xi, k(w, x)) \end{aligned} \tag{5}$$

with output

$$e = h(w, x)\,,$$

be such that:

- The positive orbit of $W \times X \times \Xi$ is bounded, i. e. there exists a bounded subset S of $\mathbb{R}^s \times \mathbb{R}^n \times \mathbb{R}^\nu$ such that, for any $(w_0, x_0, \xi_0) \in W \times X \times \Xi$, the integral curve $(w(t), x(t), \xi(t))$ of (5) passing through (w_0, x_0, ξ_0) at time $t = 0$ remains in S for all $t \geq 0$.
- $\lim_{t\to\infty} e(t) = 0$, uniformly in the initial condition, i. e., for every $\varepsilon > 0$ there exists a time $\bar{t}$, depending only on ε and *not on* (w_0, x_0, ξ_0), such that the integral curve $(w(t), x(t), \xi(t))$ of (5) passing through (w_0, x_0, ξ_0) at time $t = 0$ satisfies $\|e(t)\| \leq \varepsilon$ for all $t \geq \bar{t}$.

The Steady-State Behavior of a System

Limit Sets

The generalized tracking problem can be seen as the problem of forcing in the plant, by means of an appropriate control input $u(t)$, a response $x(t)$ that asymptotically compensates the effect, on the regulated variable $e(t)$, of the exogenous input $w(t)$. The classical way in which the problem is addressed for linear, time-invariant systems, when the exosystem is a neutrally stable linear system, is to seek a controller forcing in the associated closed-loop system (5) a (stable) "steady state" behavior entirely contained in the kernel of the map defining the tracking error e. Thus, it is natural to expect that a similar tool should also be effective in the more general setting considered here. It appears, though, that a rigorous investigation of the concept of "steady state", beyond the classical domain of linear system theory, had never been fully pursued.

Motivated by the current practice in linear system theory, the "steady state" behavior of a dynamical system can be viewed as a kind *limit* behavior, approached either as the *actual* time t tends to $+\infty$ or, alternatively, as the *initial* time t_0 tends to $-\infty$. Relevant, in this regard, are certain concepts introduced by G.D. Birkhoff in his classical

1927 essay, where he asserts that "with an arbitrary dynamical system ... there is associated always a closed set of 'central motions' which do possess this property of regional recurrence, towards which all other motions of the system in general tend asymptotically" (see p. 190 in [2]). In particular, a fundamental role is played by the concept of ω-limit set of a given point, which is defined as follows. Consider an *autonomous* ordinary differential equation

$$\dot{x} = f(x) \tag{6}$$

with $x \in \mathbb{R}^n$, $t \in \mathbb{R}$. It is well known that, if $f\colon \mathbb{R}^n \to \mathbb{R}^n$ is locally Lipschitz, for any $x_0 \in \mathbb{R}^n$, the solution of (6) with initial condition $x(0) = x_0$, denoted by $x(t, x_0)$, exists on some open interval of the point $t = 0$ and is unique.

Assume, in particular, that $x(t, x_0)$ is defined for all $t \geq 0$. A point x is said to be an ω-limit *point* of the motion $x(t, x_0)$ if there exists a sequence of times $\{t_k\}$, with $\lim_{k\to\infty} t_k = \infty$, such that

$$\lim_{k\to\infty} x(t_k, x_0) = x\,.$$

The ω-limit *set* of a point x_0, denoted $\omega(x_0)$, is *the union* of all ω-limit points of the motion $x(t, x_0)$.

It is obvious from this definition that an ω-limit point *is not* necessarily a limit of $x(t, x_0)$ as $t \to \infty$, as the solution in question may not admit any limit as $t \to \infty$. It happens though, that if the motion $x(t, x_0)$ is *bounded*, then $x(t, x_0)$ asymptotically approaches *the set* $\omega(x_0)$. This property is precisely described in what follows [2].

Lemma 1 *Suppose there is a number M such that $\|x(t, x_0)\| \leq M$ for all $t \geq 0$. Then, $\omega(x_0)$ is a nonempty compact connected set, invariant under (6). Moreover, the distance of $x(t, x_0)$ from $\omega(x_0)$ tends to 0 as $t \to \infty$.*

One of the remarkable features of $\omega(x_0)$, as indicated in this Lemma, is the fact that this set is *invariant* for (6). Invariance means that for any initial condition $\bar{x}_0 \in \omega(x_0)$ the solution $x(t, \bar{x}_0)$ of (6) exists *for all* $t \in (-\infty, +\infty)$ and that $x(t, \bar{x}_0) \in \omega(x_0)$ for all such t. Put in different terms, the set $\omega(x_0)$ is filled by motions of (6) which are bounded backward and forward in time. The other remarkable feature is that $x(t, x_0)$ *approaches* $\omega(x_0)$ as $t \to \infty$, in the sense that the distance *of the point* $x(t, x_0)$ (the value at time t of the solution of (6) starting in x_0 at time $t = 0$) *from the set* $\omega(x_0)$ tends to 0 as $t \to \infty$.

Since any motion $x(t, x_0)$ which is bounded in positive time asymptotically approaches the ω-limit set $\omega(x_0)$ as $t \to \infty$, one may be tempted to look, for a system (6) in which *all* motions are bounded in positive time, at the *union* of the limit sets of all points x_0, i. e. at the set

$$\Omega = \bigcup_{x_0 \in \mathbb{R}^n} \omega(x_0)$$

and to say that the system is in steady state if its state $x(t)$ evolves in the (invariant) set Ω. There is a major drawback, though, in taking this as definition of "steady state" behavior of a nonlinear system: the convergence of $x(t, x_0)$ to Ω is not guaranteed to be *uniform* in x_0, even if the latter ranges over a compact set (see, e. g [7]).

One of the main motivations for looking into the concept of steady state is the aim *to shape* the steady state response of a system to a given (or to a given family of) forcing inputs. But this motivation looses much of its meaning if the time needed to get within an ε-distance from the steady state may grow unbounded as the initial state changes (even when the latter is picked within a fixed *bounded* set). In other words, *uniform* convergence to the steady state (which is automatically guaranteed in the case of linear systems) is an indispensable feature to be required in a nonlinear version of this notion. The set Ω, the union of all ω-limit points of all points in the state space does not have this property of uniform convergence, but there is a larger set which does have this property. This larger set, known as the ω limit set *of a set*, is precisely defined as follows.

Consider again system (6), let B be a subset of $\mathbb{R}^n$ and suppose $x(t, x_0)$ is defined for all $t \geq 0$ and all $x_0 \in B$. The ω-limit set of B, denoted $\omega(B)$, is the set of all points x for which there exists a sequence of pairs $\{x_k, t_k\}$, with $x_k \in B$ and $\lim_{k\to\infty} t_k = \infty$ such that

$$\lim_{k\to\infty} x(t_k, x_k) = x\,.$$

It is clear from the definition that if B consists of only one single point x_0, all x_k's in the definition above are necessarily equal to x_0 and the definition in question reduces to the definition of ω-limit set of a point, given earlier. It is also clear form this definition that, if for some $x_0 \in B$ the set $\omega(x_0)$ is nonempty, all points of $\omega(x_0)$ are points of $\omega(B)$. In fact, all such points have the property indicated in the definition, if all the x_k's are taken equal to x_0. Thus, in particular, if all motions with $x_0 \in B$ are bounded in positive time,

$$\bigcup_{x_0 \in B} \omega(x_0) \subset \omega(B)\,.$$

However, the converse inclusion is not true in general.

The relevant properties of the ω-limit set of a set, which extend those presented earlier in Lemma 1, can be summarized as follows [14].

Lemma 2 *Let B be a nonempty bounded subset of $\mathbb{R}^n$ and suppose there is a number M such that $\|x(t, x_0)\| \leq M$ for all $t \geq 0$ and all $x_0 \in B$. Then $\omega(B)$ is a nonempty compact*

set, invariant under (6). Moreover, the distance of $x(t, x_0)$ from $\omega(B)$ tends to 0 as $t \to \infty$, uniformly in $x_0 \in B$. If B is connected, so is $\omega(B)$.

Thus, as it is the case for the ω-limit set of a point, we see that the ω-limit set of a bounded set, being compact and invariant, is filled with motions which exist for all $t \in (-\infty, +\infty)$ and are bounded backward and forward in time. But, above all, we see that the set in question is *uniformly* approached by motions with initial state $x_0 \in B$, a property that the set Ω does not have. Note also that the set of all such trajectories is a "behavior", in the sense of J.C. Willems [25].

The set $\omega(B)$, as shown in the previous Lemma, asymptotically attracts, as $t \to \infty$, all motions that start in B. Since the convergence to $\omega(B)$ is uniform in x_0, it is also true that, whenever $\omega(B)$ is contained in the interior of B, the set $\omega(B)$ is *asymptotically stable*, in the sense of Lyapunov (see [14]).

The Steady State Behavior of a Nonlinear System

Consider now again system (6), with initial conditions in a closed subset $X \subset \mathbb{R}^n$. Suppose the set X is *positively invariant*, which means that for any initial condition $x_0 \in X$, the solution $x(t, x_0)$ exists for all $t \geq 0$ and $x(t, x_0) \in X$ for all $t \geq 0$. The motions of this system are said to be *ultimately bounded* if there is a bounded subset B with the property that, for every compact subset X_0 of X, there is a time $T > 0$ such that $\|x(t, x_0)\| \in B$ for all $t \geq T$ and all $x_0 \in X_0$. In other words, if the motions of the system are ultimately bounded, every motion eventually enters and remains in the bounded set B.

Suppose the motions of (6) are ultimately bounded and let $B' \neq B$ be any other bounded subset with the property that, for every compact subset X_0 of X, there is a time $T > 0$ such that $\|x(t, x_0)\| \in B'$ for all $t \geq T$ and all $x_0 \in X_0$. Then, it is easy to check that $\omega(B') = \omega(B)$. Thus, in view of the properties described in Lemma 2 above, the following definition can be adopted (see [7]).

Definition Suppose the motions of system (6), with initial conditions in a closed and positively invariant set X, are ultimately bounded. A *steady state* motion is any motion with initial condition $x(0) \in \omega(B)$. The set $\omega(B)$ is the *steady state locus* of (6) and the *restriction* of (6) to $\omega(B)$ is the *steady state behavior* of (6).

The notion introduced in this way recaptures the classical notion of steady state for linear systems and provides a new powerful tool to deal with similar issues in the case of nonlinear systems.

Example In order to see how this notion includes the classical viewpoint, consider an n-dimensional, single-input, *asymptotically stable* linear system

$$\dot{z} = Fz + Gu \tag{7}$$

forced by the harmonic input $u(t) = u_0 \sin(\omega t + \phi_0)$. A simple method to determine the periodic motion of (7) consists in viewing the forcing input $u(t)$ as provided by an autonomous "signal generator" of the form

$$\dot{w} = Sw \qquad u = Qw$$

in which

$$S = \begin{pmatrix} 0 & \omega \\ -\omega & 0 \end{pmatrix} \qquad Q = (1 \quad 0)$$

and in analyzing the state state behavior of the associated "augmented" system

$$\begin{aligned} \dot{w} &= Sw \\ \dot{z} &= Fz + GQw\,. \end{aligned} \tag{8}$$

As a matter of fact, let Π be the unique solution of the Sylvester equation $\Pi S = F\Pi + GQ$ and observe that the graph of the linear map

$$\begin{aligned} \pi : \mathbb{R}^2 &\to \mathbb{R}^n \\ w &\mapsto \Pi w \end{aligned}$$

is an invariant subspace for the system (8). Since all trajectories of (8) approach this subspace as $t \to \infty$, the limit behavior of (8) is determined by the restriction of its motion to this invariant subspace.

Revisiting this analysis from the viewpoint of the more general notion of steady state introduce above, let $W \subset \mathbb{R}^2$ be a set of the form

$$W = \{w \in \mathbb{R}^2 : \|w\| \leq c\} \tag{9}$$

in which c is a fixed number, and suppose the set of initial conditions for (8) is $W \times \mathbb{R}^n$. This is in fact the case when the problem of evaluating the periodic response of (7) to harmonic inputs whose amplitude does not exceed a fixed number c is addressed. The set W is compact and invariant for the upper subsystem of (8) and, as it is easy to check, the ω-limit set of W under the motion of the upper subsystem of (8) is the subset W itself.

The set $W \times \mathbb{R}^n$ is closed and positively invariant for the full system (8) and, moreover, since the lower subsystem of (8) is a linear asymptotically stable system driven by a bounded input, it is immediate to check that the motions

of system (8), with initial conditions taken in $W \times \mathbb{R}^n$, are ultimately bounded. As a matter of fact, any bounded set B of the form

$$B = \{(w, z) \in \mathbb{R}^2 \times \mathbb{R}^n : w \in W, \|z - \Pi w\| \leq d\}$$

in which d is any positive number, has the property indicated in the definition of ultimate boundedness. It is easy to check that

$$\omega(B) = \{(w, z) \in \mathbb{R}^2 \times \mathbb{R}^n : w \in W, z = \Pi w\} ,$$

i. e. $\omega(B)$ is the graph of the restriction of the map π to the set W. The restriction of (8) to the invariant set $\omega(B)$ characterizes the steady state behavior of (7) under the family of all harmonic inputs of fixed angular frequency ω, and amplitude not exceeding c.

Example A similar result, namely the fact that the *steady state locus* is the *graph* of a map, can be reached if the "signal generator" is any nonlinear system, with initial conditions chosen in a compact invariant set W. More precisely, consider an augmented system of the form

$$\begin{aligned} \dot{w} &= s(w) \\ \dot{z} &= Fz + Gq(w) , \end{aligned} \tag{10}$$

in which $w \in W \subset \mathbb{R}^r$, $x \in \mathbb{R}^n$, and assume that: (i) all eigenvalues of F have negative real part, (ii) the set W is a compact set, invariant for the the upper subsystem of (10).

As in the previous example, the ω-limit set of W under the motion of the upper subsystem of (10) is the subset W itself. Moreover, since the lower subsystem of (10) is a linear asymptotically stable system driven by the bounded input $u(t) = q(w(t, w_0))$, the motions of system (10), with initial conditions taken in $W \times \mathbb{R}^n$, are ultimately bounded.

It is easy to check that the steady state locus of (10) is the graph of the map

$$\begin{aligned} \pi : W &\to \mathbb{R}^n \\ w &\mapsto \pi(w) , \end{aligned}$$

defined by

$$\pi(w) = \lim_{T \to \infty} \int_{-T}^{0} \mathrm{e}^{-F\tau} Gq(w(\tau, w))\, \mathrm{d}\tau . \tag{11}$$

To see why this is the case, pick any initial condition (w_0, z_0) for (10) on the graph of π and compute the solution $z(t)$ of the lower equation of (10) by means of the classical variation of constants formula, to obtain

$$z(t) = \mathrm{e}^{Ft} z_0 + \int_0^t \mathrm{e}^{F(t-\tau)} Gq(w(\tau, w_0))\, \mathrm{d}\tau$$

Since by hypothesis $z_0 = \pi(w_0)$, using (10) one obtains

$$\begin{aligned} z(t) =& \mathrm{e}^{Ft} \int_{-\infty}^{0} \mathrm{e}^{-F\tau} Gq(w(\tau, w_0))\, \mathrm{d}\tau \\ &+ \int_0^t \mathrm{e}^{F(t-\tau)} Gq(w(\tau, w_0))\, \mathrm{d}\tau \\ =& \int_{-\infty}^{t} \mathrm{e}^{F(t-\tau)} Gq(w(\tau, w_0))\, \mathrm{d}\tau \\ =& \int_{-\infty}^{0} \mathrm{e}^{-F\theta} Gq(w(\theta + t, w_0))\, \mathrm{d}\theta \\ =& \int_{-\infty}^{0} \mathrm{e}^{-F\theta} Gq(w(\theta, w(t, w_0)))\, \mathrm{d}\theta \\ =& \pi(w(t, w_0)) = \pi(w(t)) \end{aligned}$$

which proves the invariance of the graph of π for (10). It is deduced from this that that any point of the graph of π is necessarily a point of the steady state locus of (10). To complete the proof of the claim it remains to show that no other point of $W \times \mathbb{R}^n$ can be a point of the steady state locus. But this is a straightforward consequence of the fact that F has eigenvalues with negative real part.

There are various ways in which the result discussed in the previous example can be generalized. For instance, it can be extended to describe the steady state response of a nonlinear system

$$\dot{z} = f(z, u) \tag{12}$$

in the neighborhood of a locally exponentially stable equilibrium point. To this end, suppose that $f(0, 0) = 0$ and that the matrix

$$F = \left[\frac{\partial f}{\partial z} \right](0, 0)$$

has all eigenvalues with negative real part. Then, it is well known (see e. g. p. 275 [13]) that it is always possible to find a compact subset $Z \subset \mathbb{R}^n$, which contains $z = 0$ in its interior and a number $\sigma > 0$ such that, if $\|z_0\| \in Z$ and $\|u(t)\| \leq \sigma$ for all $t \geq 0$, the solution of (12) with initial condition $z(0) = z_0$ satisfies $\|z(t)\| \in Z$ for all $t \geq 0$. Suppose that the input u to (12) is produced, as before, by a signal generator of the form

$$\begin{aligned} \dot{w} &= s(w) \\ u &= q(w) \end{aligned} \tag{13}$$

with initial conditions chosen in a compact invariant set W and, moreover, suppose that, $\|q(w)\| \leq \sigma$ for all $w \in W$. If this is the case, the set $W \times Z$ is positively invariant for

$$\begin{aligned} \dot{w} &= s(w) \\ \dot{z} &= f(z, q(w)) , \end{aligned} \tag{14}$$

and the motions of the latter are ultimately bounded, with $B = W \times Z$. The set $\omega(B)$ may have a complicated structure but it is possible to show, by means arguments similar to those which are used in the proof of the Center Manifold theorem, that if Z and B are small enough the set in question can still be expressed as the graph of a map $z = \pi(w)$. In particular, the graph in question is precisely the center manifold of (14) at $(0, 0)$ if $s(0) = 0$ and the matrix

$$S = \left[\frac{\partial s}{\partial w}\right](0)$$

has all eigenvalues on the imaginary axis.

A common feature of the examples discussed above is the fact that the steady state locus of a system of the form (14) can be expressed as the graph of a map $z = \pi(w)$. This means that, so long as this is the case, a system of this form has a *unique* well defined *steady state response* to the input $u(t) = q(w(t))$. As a matter of fact, the response in question is precisely $z(t) = \pi(w(t))$. Of course, this may not always be the case and *multiple* steady state responses to a given input may occur. In general, the following property holds.

Lemma 3 *Let W be a compact set, invariant under the flow of (13). Let Z be a closed set and suppose that the motions of (14) with initial conditions in $W \times Z$ are ultimately bounded. Then, the steady state locus of (14) is the graph of a set-valued map defined on the whole of W.*

Necessary Conditions for Output Regulation

Taking advantage of the notions introduced in the previous section, we are now in a position to highlight some general properties that any controller that solves a problem of output regulation must necessarily have. Recall that, as defined earlier, the problem of output regulation is solved if, in the composite system (5):

- the positive orbit of $W \times X \times \Xi$ is bounded,
- $\lim_{t\to\infty} e(t) = 0$, uniformly in the initial condition.

The notions introduced in the previous section are instrumental to prove the following, elementary – but fundamental – result, which is a nonlinear enhancement of a Lemma of [10] on which all the theory of output regulation for linear systems is based.

Lemma 4 *Suppose the positive orbit of $W \times X \times \Xi$ is bounded. Then*

$$\lim_{t\to\infty} e(t) = 0$$

if and only if

$$\omega(W \times X \times \Xi) \subset \{(w, x, \xi) \colon h(w, x) = 0\}. \tag{15}$$

It is seen from this simple result that the problem of output regulation can be simply cast as the problem of *shaping the steady state locus of the closed loop system*, in such a way that property (15) holds.

To proceed with the analysis in a more concrete fashion, we consider from now on the special case in which the controlled plant (4) is modeled by equations *in normal form*

$$\begin{aligned}
\dot{z} &= f_0(w, z) + f_1(w, z, e_1)e_1 \\
\dot{e}_1 &= e_2 \\
&\vdots \\
\dot{e}_{r-1} &= e_r \\
\dot{e}_r &= q(w, z, e_1, \ldots, e_r) + b(w, z, e_1, \ldots, e_r)u \\
e &= e_1 \\
y &= col(e_1, \ldots, e_r)\,,
\end{aligned} \tag{16}$$

with state $(z, e_1, \ldots, e_r) \in \mathbb{R}^{n-r} \times \mathbb{R}^r$, control input $u \in \mathbb{R}$, regulated output $e \in \mathbb{R}$, measured output $y \in \mathbb{R}^r$. The functions $f_0(\cdot), f_1(\cdot), q(\cdot), b(\cdot), s(\cdot)$ in (16) and (3) are assumed to be at least continuously differentiable. It is also assumed that

$$b(w, z, e_1, \ldots, e_r) \neq 0 \qquad \forall (w, z, e_1, \ldots, e_r)\,.$$

The initial conditions of (16) range on a set $Z \times E$, in which Z is a fixed *compact* subset of $\mathbb{R}^{n-r}$ and $E = \{(e_1, \ldots, e_r) \in \mathbb{R}^r \colon |e_i| \leq c\}$, with c a fixed number.

Suppose that a controller of the form (4) solves the problem of output regulation. Then Lemma 4 applies and, since $e = e_1$, we deduce that the steady state locus of the closed loop system (5) is necessarily a subset of the set of all states in which $e_1 = 0$. This being the case, it is seen from the form of the equations (16) that, when the closed loop system (5) is in steady state, necessarily also

$$e_2 = e_3 = \cdots = e_r = 0\,.$$

As a consequence, the following conclusions hold:

- *The steady state locus $\omega(W \times Z \times E \times \Xi)$ of the closed-loop system is a subset of the set $\mathbb{R}^s \times \mathbb{R}^{n-r} \times \{0\} \times \mathbb{R}^\nu$.*
- *The restriction of the closed-loop system to its steady state locus $\omega(W \times Z \times E \times \Xi)$ reduces to*

$$\begin{aligned}
\dot{w} &= s(w) \\
\dot{z} &= f_0(w, z) \\
\dot{\xi} &= \varphi(\xi, 0)\,.
\end{aligned} \tag{17}$$

- *For each* $(w, z, 0, \ldots, 0, \xi) \in \omega(W \times Z \times E \times \Xi)$

$$0 = q(w, z, 0, \ldots, 0) + b(w, z, 0, \ldots, 0)\gamma(\xi, 0) . \quad (18)$$

The prior analysis implicitly assumes that the positive orbit of W under the flow of exosystem is bounded, i. e. that the motions of the exosystem asymptotically approach the its own steady state locus $\omega(W)$. In principle, $\omega(W)$ may differ from W but there is no loss of generality in assuming from the very beginning that the two sets coincide. After all, the problem in question is a problem concerning how the closed-loop system behaves in steady state and there is no special interest in considering exosystems that are not "in steady state". We make this assumption precise as follows.

Assumption (i) *The compact set W is invariant for (3).*

With this in mind we observe that, by Lemma 3, if the positive orbit of $W \times Z \times E \times \Xi$ under the flow of (5) is bounded, then $\omega(W \times Z \times E \times \Xi)$ is the graph of a (possibly set-valued) map defined on the whole of W. Consider now the set

$$\mathcal{A}_{ss} = \{(w, z) \colon (w, z, 0, \ldots, 0, \xi) \in \omega(W \times Z \times E \times \Xi), \text{ for some } \xi \in \mathbb{R}\}$$

and define the map

$$\begin{aligned} u_{ss} : \mathcal{A}_{ss} &\to \mathbb{R} \\ (w, z) &\mapsto -\frac{q(w, z, 0, \ldots, 0)}{b(w, z, 0, \ldots, 0)} . \end{aligned}$$

By construction, the set $\mathcal{A}_{ss}$ is the graph of a (possibly set-valued) map defined on the whole of W, which is invariant for the dynamics of

$$\begin{aligned} \dot{w} &= s(w) \\ \dot{z} &= f_0(w, z) , \end{aligned} \quad (19)$$

that are precisely *the zero dynamics of the "augmented system"* (3)–(16), while the map $u_{ss}(\cdot)$ is the control that forces the motion of (3) –(16) to evolve on $\mathcal{A}_{ss}$.

With this in mind, the conclusions reached above can be rephrased in the following terms. Suppose that a controller of the form (4) solves the problem of output regulation for (16) with exosystem (3). Then, there exists a (possibly set-valued) map defined on the whole of W whose graph $\mathcal{A}_{ss}$ is invariant for the autonomous system (19). Moreover, for each $(w_0, z_0) \in \mathcal{A}_{ss}$ there is a point $\xi_0 \in \mathbb{R}^\nu$ such that the integral curve of (19) issued from (w_0, z_0) and the integral curve of

$$\dot{\xi} = \varphi(\xi, 0)$$

issued from ξ_0 satisfy

$$u_{ss}(w(t), z(t)) = \gamma(\xi(t)) , \qquad \forall t \in \mathbb{R} .$$

This is a nonlinear version of the celebrated *internal model principle* of [11].

Sufficient Conditions for Output Regulation

The Control Structure

On the basis of the ideas presented in the previous section we proceed now with the construction of a controller that solves the problem of output regulation. The "steady state" features of this controller are those identified at the end of the section, namely this controller has to be able to "generate" all controls of the form $u_{ss}(w(t), z(t))$ for any "steady state" trajectory $w(t), z(t)$ of (19). The controller should incorporate a device that generates all such trajectories (the *internal model*), thus making sure that the "appropriate" state-state behavior takes place, and a device guaranteeing that convergence to this specific steady state behavior occurs. It is here that additional assumptions are needed.

Note that, since W is invariant for $\dot{w} = s(w)$, the closed cylinder

$$C := W \times \mathbb{R}^{n-r}$$

is locally invariant for (19). Hence, it is natural regard (19) as a system defined on C and endow the latter with the subset topology.

Assumption (ii) *There exists a bounded subset B of C which contains the positive orbit of the set $W \times Z$ under the flow of (19) and the resulting omega-limit set*

$$\mathcal{A} := \omega(W \times Z)$$

satisfies

$$(w, z) \in C , \quad |(w, z)|_{\mathcal{A}} \le d_0 \quad \Rightarrow \quad z \in Z \quad (20)$$

where d_0 is a positive number.

While in the analysis of the necessity we have only identified the existence of a compact set (actually, the graph of a map defined on W) which is invariant for (19), the new assumption (ii) implies, in its first part, the existence of a compact set $\mathcal{A}$ (still the graph of a map defined on W) which is not only invariant but also uniformly attractive of all trajectories of (19) issued from points of $W \times Z$. The second part of the assumption, in turn, guarantees that this set is also stable in the sense of Lyapunov. In the next assumption we strengthen this property by also requiring the

set $\mathcal{A}$ to be locally exponentially stable (this assumption is useful to straighten the subsequent analysis, but is not essential).

Assumption (iii) *There exist $M \geq 1$, $\lambda > 0$ such that*

$$(w_0, z_0) \in C\,, \quad |(w_0, z_0)|_{\mathcal{A}} \leq d_0 \quad \Rightarrow$$
$$|(w(t), z(t))|_{\mathcal{A}} \leq Me^{-\lambda t}|(w_0, z_0)|_{\mathcal{A}}, \qquad \forall t \geq 0$$

in which $(w(t), z(t))$ denotes the solution of (19) passing through (w_0, z_0) at time $t = 0$.

To simplify the exposition, we address the special case in which the controlled system (16) has relative degree 1, and in which the coefficient $b(w, z, e_1)$ is identically equal to 1. In other words, we consider a system modeled by equations of the form

$$\begin{aligned} \dot{z} &= f_0(w, z) + f_1(w, z, e)e \\ \dot{e} &= q(w, z, e) + u \\ y &= e\,. \end{aligned} \tag{21}$$

There is no loss of generality in considering a system having this simple form (21) because, as shown for instance in [9,22], the case of a more general system of the form (16) can easily be reduced, by appropriate manipulations, to this one.

For convenience, rewrite the augmented system (3)–(21) as

$$\begin{aligned} \dot{z} &= f_0(z) + f_1(z, e)e \\ \dot{e} &= q_0(z) + q_1(z, e)e + u \end{aligned} \tag{22}$$

having set $z = (w, z)$. Consistently let $Z := W \times Z$ denote the compact set where the initial condition $z(0)$ is supposed to range. In these notations, assumptions (i) – (ii) - (iii) express the property that, in the autonomous system

$$\dot{z} = f_0(z)\,, \tag{23}$$

the set $\mathcal{A}$ is asymptotically and locally exponentially stable, with a domain of attraction that contains the set Z.

Suppose now that the control u is chosen as $u = -ke$. The closed-loop system thus obtained can be regarded as a feedback interconnection of

$$\dot{z} = f_0(z) + f_1(z, e)e \tag{24}$$

viewed as a system with input e and state z, and

$$\dot{e} = q_0(z) + q_1(z, e)e - ke \tag{25}$$

viewed as a system with input z and state e.

By assumption, system (24) possesses, when $e = 0$, an invariant set $\mathcal{A}$ which is asymptotically and locally exponentially stable, with a domain of attraction that contains the set Z of all admissible initial conditions. Thus, standard arguments (see, e. g [6].) can be invoked to claim that, if k is large enough, all trajectories of the interconnection (24)–(25) with initial conditions in $Z \times E$ remain bounded and the state (z, e) can be steered to an arbitrary small neighborhood of the set $\mathcal{A} \times \{0\}$. This does not solve the problem at issue, though, because the variable $e(t)$ is not guaranteed to converge to zero (but only to converge to a neighborhood of zero, whose size can be made arbitrarily small by increasing the gain coefficient k). The condition for having $e(t) \to 0$ as $t \to \infty$ is simply that the "coupling" term $q_0(z)$ vanishes on the set $\mathcal{A}$, but there is no reason for this to occur (see again, e. g.[6]). This is why a more elaborate, internal-model-based, controller is needed.

System (16) being affine in the control input u, it seems natural to look for a controller having a similar structure, namely a controller of the form

$$\begin{aligned} \dot{\xi} &= \varphi(\xi) + Gv \\ u &= \gamma(\xi) + v \end{aligned} \tag{26}$$

with state $\xi \in \mathbb{R}^{\nu}$, in which v is a residual control input, to be eventually chosen as a function of the measured output y. Here $\varphi(\cdot)$, G and $\gamma(\cdot)$ are functions to be determined. We will show in what follows that, if the triplet $\{\varphi(\xi), G, \gamma(\xi)\}$ possesses what we will define as *asymptotic internal model* property, the choice of the residual control v in (26) as

$$v = -ke$$

solves the problem of output regulation, provided that the gain coefficient k is sufficiently high.

The Internal Model

Controlling this system by means of (26) yields a closed-loop system

$$\begin{aligned} \dot{z} &= f_0(z) + f_1(z, e)e \\ \dot{e} &= q_0(z) + q_1(z, e)e + \gamma(\xi) + v \\ \dot{\xi} &= \varphi(\xi) + Gv \end{aligned} \tag{27}$$

which, regarded as a system with input v and output e, has relative degree 1 and zero dynamics given by

$$\begin{aligned} \dot{z} &= f_0(z) \\ \dot{\xi} &= \varphi(\xi) - G[\gamma(\xi) + q_0(z)]\,. \end{aligned} \tag{28}$$

System (27) can be put in normal form by means of the change of variables

$$\chi = \xi - Ge$$

which yields

$$\begin{aligned} \dot{z} &= f_0(z) + f_1(z,e)e \\ \dot{\chi} &= \varphi(\chi + Ge) - G\gamma(\chi + Ge) - Gq_0(z) - Gq_1(z,e)e \\ \dot{e} &= q_0(z) + q_1(z,e)e + \gamma(\chi + Ge) + v\,. \end{aligned} \tag{29}$$

Setting $x = (z, \chi)$, this system can be further rewritten in the form

$$\begin{aligned} \dot{x} &= f(x) + \ell(x,e)e \\ \dot{e} &= q(x) + r(x,e)e + v \end{aligned} \tag{30}$$

in which

$$f(x) = \begin{pmatrix} f_0(z) \\ \varphi(x) - G[\gamma(x) + q_0(z)] \end{pmatrix}$$

$$q(x) = q_0(z) + \gamma(x)$$

and $\ell(x,e)$, $r(x,e)$ are suitable continuous functions.

Suppose now that the residual control v is chosen as $v = -ke$. This yields a closed-loop system having a structure similar to the one considered in the previous subsection, namely a feedback interconnection of

$$\dot{x} = f(x) + \ell(x,e)e \tag{31}$$

viewed as a system with input e and state x, and

$$\dot{e} = q(x) + r(x,e)e - ke \tag{32}$$

viewed as a system with input x and state e. As claimed earlier, a high-gain control on e, namely a control $v = -ke$ with large k, would succeed in steering $e(t)$ to zero if two conditions are fulfilled:

(P1) the dynamics (28) possesses a compact invariant set which is asymptotically (and locally exponentially) stable, with a domain of attraction that contains the set $Z \times \Xi$ of all admissible initial conditions, and

(P2) the function $q_0(z) + \gamma(\xi)$ vanishes on this invariant set.

These are the properties that will be sought in what follows. Note that the fulfillment of these is determined only by properties of the autonomous system (23) and of the function

$$\rho = q_0(z) \tag{33}$$

which, in the composite system (28), can be viewed as the output of (23) driving a system of the form

$$\dot{\xi} = \varphi(\xi) - G[\gamma(\xi) + \rho]\,. \tag{34}$$

For convenience, we will say that triplet $\{\varphi(\xi), G, \gamma(\xi)\}$ *is an asymptotic internal model of the pair* (23) – (33) if properties (P1) and (P2) are satisfied. In this terminology, we can summarize as follows the conclusion obtained so far.

Proposition 1 *Pick compact sets Z, E and Ξ for the initial conditions of the closed-loop system (3), (21), (26). Assume that (i)-(ii)-(iii) hold and that the triplet $\{\varphi(\xi), G, \gamma(\xi)\}$ is an asymptotic internal model of (23) – (33). Then there exists $k^\star > 0$ such that for all $k \geq k^\star$ the controller (26) with $v = -ke$ solves the generalized tracking problem.*

The notion of steady state provides a useful interpretation of the properties in question. In fact, recall that, by assumption, all trajectories of system (23) with initial conditions in Z asymptotically converge to the compact invariant set $\mathcal{A}$, and the latter is also locally exponentially stable. If property (P1) holds, all trajectories of the composite system (28) with initial conditions in $Z \times \Xi$ asymptotically converge to the limit set $\omega(Z \times \Xi)$. Since (28) is a triangular system, it is readily seen that the set $\omega(Z \times \Xi)$ is the graph of a set-valued map defined on $\mathcal{A}$, i. e. that there exists a map

$$\tau : z \in \mathcal{A} \mapsto \tau(z) \subset \mathbb{R}^\nu\,,$$

such that

$$\omega(Z \times \Xi) = \{(z,\xi) \colon z \in \mathcal{A}, \xi \in \tau(z)\} \colon = \mathrm{gr}(\tau)\,.$$

The set $\mathrm{gr}(\tau)$ is the steady state locus of (28) and the restriction of the latter to this invariant set characterizes its steady state behavior. Property (P2), on the other hand, expresses the property that at each point of $(z,\xi) \in \mathrm{gr}(\tau)$

$$q_0(z) = -\gamma(\xi)\,. \tag{35}$$

Thus, looking again at system (28), it is realized that $\mathrm{gr}(\tau)$ is in fact invariant for

$$\begin{aligned} \dot{z} &= f_0(z) \\ \dot{\xi} &= \varphi(\xi)\,. \end{aligned} \tag{36}$$

Note that, if the map $\tau(z)$ is single-valued and C^1, its invariance for (36) is expressed by the property that

$$\frac{\partial \tau(z)}{\partial z} f_0(z) = \varphi(\tau(z)) \qquad \forall\, z \in \mathcal{A}\,, \tag{37}$$

while the fact that (35) holds at each point of $(z,\xi) \in \mathrm{gr}(\tau)$ is expressed by the property that

$$q_0(z) = -\gamma(\tau(z)) \qquad \forall\, z \in \mathcal{A}\,. \tag{38}$$

The Design of an Internal Model

As we have seen in the earlier sections, the proposed controller, if the asymptotic internal model property holds, is able to force – in the closed loop system – convergence to a steady state in which the regulated variable is identically zero. As a consequence, the controller solves the generalized tracking problem. It remains to be shown, therefore, how the asymptotic internal model property can be obtained. To this end, it is convenient to observe that the properties required in (P1) and (P2) are quite similar to properties that are usually sought in the design of *state observers*. As a matter of fact it is seen from (37) and (38) that, for each $z_0 \in \mathcal{A}$, the function of time

$$\hat{\xi}(t) = \tau(z(t, z_0))$$

which is defined (and bounded) for all $t \in \mathbb{R}$ satisfies

$$\frac{\mathrm{d}\hat{\xi}(t)}{\mathrm{d}t} = \varphi(\hat{\xi}(t)) \tag{39}$$

and, moreover

$$\gamma(\hat{\xi}(t)) = -q_0(z(t, z_0)) \,.$$

In view of the latter, system (34) can be rewritten in the form

$$\dot{\xi} = \varphi(\xi) + G[\gamma(\hat{\xi}) - \gamma(\xi)] \tag{40}$$

and interpreted as *a copy of the dynamics* (39) *of* $\hat{\xi}$ *corrected by an "innovation term"* $[\gamma(\hat{\xi}) - \gamma(\xi)]$ *weighted by an "output injection gain"* G. This is the classical structure on an *observer* and the requirement in (P1) expresses the property that the difference $\xi(t) - \hat{\xi}(t)$ (the "observation error", in our interpretation) should asymptotically decay to zero (with ultimate exponential decay).

This interpretation is at the basis of a number of major recent advances in the design of regulators. In fact, in a number of recent papers, this interpretation has been pursued and, taking into consideration various approaches to the design of nonlinear observers, has lead to effective design methods (see [5,9,22]). Two of such design methods are highlighted in the remaining part of this section.

The High-Gain Observer as an Internal Model (see [5])

The construction summarized in this section relies upon the following additional hypothesis.

Assumption (iv) *Suppose there exist an integer $d > 0$ and a locally Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ such that, for any $z_0 \in \mathcal{A}$, the solution $z(t)$ of passing through z_0 at time $t = 0$ is such that the function $\rho(t) := q_0(z(t))$ satisfies*

$$\rho^{(d)}(t) = f(\rho(t), \rho^{(1)}(t), \ldots, \rho^{(d-1)}(t))$$

for all $t \in \mathbb{R}$.

Let $\tau : W \times \mathbb{R}^{n-1} \to \mathbb{R}^d$ be the map defined as

$$\tau(z) := \mathrm{col}(q_0(z), L_{f_0} q_0(z), \ldots, L_{f_0}^{d-1} q_0(z)) \tag{41}$$

and let $f_c : \mathbb{R}^d \to \mathbb{R}$ be a C^1 function with compact support which agrees with $f(\cdot)$ on $\tau(\mathcal{A})$. Then, it easy to check that the properties indicated in (37) and (38) are fulfilled by choosing

$$\varphi(\xi) = \begin{pmatrix} \xi_2 \\ \vdots \\ \xi_d \\ f_c(\xi_1, \xi_2, \ldots, \xi_d) \end{pmatrix}, \qquad \gamma(\xi) = \xi_1 \,. \tag{42}$$

Comparing this construction with the earlier remarks we observe, in particular, that system

$$\begin{aligned} \dot{z} &= f_0(z) \\ \rho &= q_0(z) \end{aligned} \tag{43}$$

is *immersed into a system which is uniformly observable*, in the sense of [12] (even though system (43) might not have had such a property). It is precisely this that makes it possible to choose G in such a way that the property indicated in (P1) can be achieved.

As a matter of fact, the property in question is achieved by choosing

$$G = D_g \begin{pmatrix} c_0 \\ \vdots \\ c_{d-1} \end{pmatrix}$$

where $D_g = diag(g, g^2, \cdots, g^d)$, g is a design parameter, and the c_i's are such that the polynomial $\lambda^d + c_0\lambda^{d-1} + \cdots + c_{d-1} = 0$ is Hurwitz, as formally proved in Lemmas 1 and 2 of [5] to which the interested reader is referred for details.

It is worth noting that the assumption in question clearly covers the interesting (and widely addressed in the recent past literature, see [15]) case in which the function $f(\cdot)$ is linear, namely the case in which (43) is immersed into a linear observable system. In this case, although the choice indicated above is clearly still valid, a more direct way of designing the regulator is to use $f(\cdot)$ instead of $f_c(\cdot)$ in the definition of $\varphi(\xi)$, and simply choose G in such a way that $\dot{\xi} = \varphi(\xi) - G\gamma(\xi)$ is a stable linear system.

The Andrieu-Praly's Observer as an Internal Model (see [22]) In this subsection we exploit certain results of the theory presented in [1] to weaken (and, to some extent, suppress) the Assumption (iv) presented at the beginning of the earlier subsection.

Let $(F, G) \in \mathbb{R}^{d\times d} \times \mathbb{R}^{d\times 1}$ be a controllable pair and set

$$\varphi(\xi) = F\xi + G\gamma(\xi), \tag{44}$$

with $\gamma : \mathbb{R}^d \to \mathbb{R}$ a continuous function to be determined later. If this is the case, the composite system (28) becomes

$$\begin{aligned} \dot{z} &= f_0(z) \\ \dot{\xi} &= F\xi - Gq_0(z)\,, \end{aligned} \tag{45}$$

which is precisely a system of the form (9), considered in the Example. If the matrix F is Hurwitz, and we restrict z to belong to the set $\mathcal{A}$, this system has a well defined steady state behavior, which is the graph of the map

$$\tau(z) = \int_{-\infty}^{0} \mathrm{e}^{-Fs} Gq_0(z(s, z))\mathrm{d}s\,. \tag{46}$$

As shown in Example, the graph in question is invariant for system (45), is asymptotically (and locally exponentially) stable and with a domain of attraction that coincides with $W \times \mathbb{R}^d$. Moreover, it can also be shown that there exists a number $\ell > 0$ such that, if the eigenvalues of F have real part which is less ℓ, the map (46) is C^1 (see, e. g. [22]). If this is the case, to say that the graph of (46) is invariant for (45) is equivalent to say that

$$\frac{\partial \tau}{\partial z} f_0(z) = F\tau(z) - Gq_0(z) \qquad \forall\, z \in \mathcal{A}\,. \tag{47}$$

This being the case, it is immediate to check that properties (P1) and (P2) will be fulfilled if a function $\gamma(\xi)$ can be found that renders (38) satisfied. As a matter of fact, bearing in mind (44), condition (37) becomes

$$\frac{\partial \tau}{\partial z} f_0(z) = F\tau(z) - G\gamma(\tau(z)) \qquad \forall\, z \in \mathcal{A}$$

which, if condition (38) holds, reduces to (47). This, shows that a triplet having the asymptotic internal model property can be found if a function $\gamma(\cdot)$ exists which satisfies (38). It is here that the dimension d of the pair (F, G) plays a role, as formalized in the next proposition whose proof can be found in [22].

Proposition 2 *Suppose*

$$d \geq 2(s + n - r) + 2\,.$$

Then for almost all choices (see [22] for details) of a controllable pair (F, G), with F a Hurwitz matrix whose eigenvalues have real part which is less than ℓ, the map (46) satisfies

$$\tau(z_1) = \tau(z_2) \quad \Rightarrow \quad q_0(z_1) = q_0(z_2)\,.$$

As a consequence there exist a continuous map $\gamma : \tau(\mathcal{A}) \to \mathbb{R}$ such that

$$q_0(z) = -\gamma(\tau(z)) \qquad \forall\, z \in \mathcal{A}\,. \tag{48}$$

The map $\tau(\cdot)$ in (46) is defined only on $\mathcal{A}$, but is not difficult to extend it to a C^1 map defined on the whole set $W \times \mathbb{R}^{n-r}$, as shown in [22]. Also the map $\gamma(\cdot)$ that makes (48) true can be extended to the whole $\mathbb{R}^d$, but this extension is only known to be continuous.

We have shown in this way that the existence of triplet $\psi(\xi), G, \gamma(\xi)$ which has the internal model property can always be achieved, so long as the integer d is large enough. This result shows that the general design procedure outlined earlier in the article is always applicable (so long as the standing hypotheses (i)–(ii)–(iii) are applicable). From the constructive viewpoint, though, it must be observed that the result indicated in Proposition PR1 is only an existence result and that the function $\gamma(\xi)$, whose existence is guaranteed, is only known to be continuous. Obtaining continuous differentiability of such $\gamma(\xi)$ and a constructive procedure are likely to require further hypotheses, which should be in any case weaker than Assumption (iv) considered earlier, whose study is subject of current investigation.

Future Directions

One of the basic hypotheses of the theory described in the previous sections is that the zero dynamics of the augmented system consisting of the controlled plant and of the exosystem possess a compact attractor which is also locally asymptotically stable. This assumption, with an acceptable abuse of terminology, is usually referred to as the "minimum-phase" property. Another standing hypothesis is that the regulated variable coincides with the measured variable. The future directions of the research in this area are aimed at the removal of these assumptions. There are several directions in which the problem can be tackled. One way is the use of control structures by means of which the controlled system (plus a part of the controller) can be interpreted with systems having a different zero dynamics (in particular a zero dynamics possessing a compact attractor). This is the nonlinear equivalent of certain design procedures for linear systems based on the assignment of zeros. This technique has proven to be power-

ful in the stabilization of certain classes of nonlinear systems (see [18]) and is expected to be successful, with appropriate enhancements, in the design of regulators. The analysis of the necessary conditions for output regulation has also shown that, if the generalized tracking problem is solvable, the inverse dynamics of the augmented system consisting of the controlled plant and of the exosystem possesses a compact invariant set to which all initial conditions are asymptotically controllable (by means of the regulated variable viewed as a control). This set is not necessarily asymptotically stable (as it would be under the "minimum-phase" hypothesis) but could be asymptotically stabilizable (by either full-state or measurement-based feedback). Thus, another direction in which the research will evolve, in the development of control schemes for possibly non "minimum-phase" nonlinear systems, is based on the exploitation of the (weaker) assumption that in the zero dynamics of the augmented system there is a compact set that can be made invariant and asymptotically stable by means of feedback. This will yield a design procedure in which a virtual control (either full-state or measurements-based) is designed to stabilize that compact invariant set. The (possibly dynamic) controller obtained in this way will then be embedded into a regulator designed according to the principles developed in this article.

Bibliography

1. Andrieu V, Praly L (2006) On the existence of a Kazantis-Kravaris/Luenberger observer. SIAM J Contr Optim 45:432–456
2. Birkhoff GD (1927) Dynamical systems. American Mathematical Society, Providence
3. Byrnes CI, Delli Priscoli F, Isidori A, Kang W (1997) Structurally stable output regulation of nonlinear systems. Automatica 33:369–385
4. Byrnes CI, Isidori A (2003) Limit Sets, Zero dynamics and internal models in the problem of nonlinear output regulation. IEEE Trans Autom Contr 48:1712–1723
5. Byrnes CI, Isidori A (2004) Nonlinear internal models for output regulation. IEEE Trans Autom Contr 49:2244–2247
6. Byrnes CI, Isidori A, Praly L (2003) On the asymptotic properties of a system Arising in non-equilibrium theory of output regulation, preprint of the Mittag-Leffler Institute. 18, Stockholm
7. Isidori A, Byrnes CI (2007) The steady-state response of a nonlinear system: ideas, tools and applications, preprint
8. Davison EJ (1976) The robust control of a servomechanism problem for linear time-invariant multivariable systems. IEEE Trans Autom Contr AC-21:25–34
9. Delli Priscoli F, Marconi L, Isidori A (2006) A new approach to adaptive nonlinear regulation. SIAM J Contr Optim 45:829–855
10. Francis BA (1977) The linear multivariable regulator problem. SIAM J Contr Optim 14:486–505
11. Francis BA, Wonham WM (1976) The internal model principle of control theory. Automatica 12:457–465
12. Gauthier JP, Kupka I (2001) Deterministic observation theory and applications. Cambridge University Press, Cambridge
13. Hahn W (1967) Stability of motions. Springer, New York
14. Hale JK, Magalhães LT, Oliva WM (2002) Dynamics in infinite dimensions. Springer, New York
15. Huang J, Lin CF (1994) On a robust nonlinear multivariable servomechanism problem. IEEE Trans Autom Contr 39:1510–1513
16. Huang J, Rugh WJ (1990) On a nonlinear multivariable servomechanism problem. Automatica 26:963–972
17. Isidori A (1995) Nonlinear Control Systems. Springer, London
18. Isidori A (2000), A tool for semiglobal stabilization of uncertain non-minimum-phase nonlinear systems via output feedback. IEEE Trans Autom Contr AC-45:1817–1827
19. Isidori A, Byrnes CI (1990) Output regulation of nonlinear systems. IEEE Trans Autom Contr 25:131–140
20. Isidori A, Marconi L, Serrani A (2003) Robust autonomous guidance: An internal model-based approach. Springer, London
21. Khalil H (1994) Robust servomechanism output feedback controllers for feedback linearizable systems. Automatica 30:587–1599
22. Marconi L, Praly L, Isidori A (2006) Output stabilization via nonlinear luenberger observers. SIAM J Contr Optim 45:2277–2298
23. Serrani A, Isidori A, Marconi L (2001) Semiglobal nonlinear output regulation with adaptive internal model. IEEE Trans Autom Contr 46:1178–1194
24. Teel AR, Praly L (1995) Tools for semiglobal stabilization by partial state and output feedback. SIAM J Control Optim 33:1443–1485
25. Willems JC (1991) Paradigms and puzzles in the theory of dynamical systems, IEEE Transaction on Automatic Control 36:259–294

Systems Biology of Human Immunity and Disease

Jared C. Roach
Seattle Childrens Hospital, Seattle, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Challenges and Solutions for Human Systems Biology
Human-Specific Problems
Examples
Future Directions
Bibliography

Glossary

Complex disease A disease that has an etiology inconsistent with the simple models of genetic inheritance proposed by Gregor Mendel. A synonym for "complex disease" is "non-Mendelian disease." Type 2 diabetes is

an example of a complex disease, as are most common diseases. The term "complex phenotype" is employed analogously to describe non-disease properties of individuals, such as height. Typically, complex diseases are influenced by multiple genetic and environmental factors. For most diseases, little is known about either of these sets of factors.

Nuclear regulation Many signals are integrated and processed on segments of DNA that are near or adjacent to genes. These segments of DNA are called cis-regulatory elements. Proteins that bind cis-regulatory elements are called transcription factors. The informational output of cis-regulatory signal integration is the rate of production of messenger RNA (mRNA) encoded by that gene. A major goal for many human systems biology projects is to understand key nuclear regulatory networks.

Regulatory network A paradigm for modeling information flow within a biological system, such as a cell. Elements of the system that can have multiple states are encoded as nodes; modes of communication between these elements are encoded as edges. A typical node would be a protein that might have multiple states related to its level of phosphorylation or cellular location. A typical edge would represent the catalytic effect of one protein upon another's state of phosphorylation.

Signaling pathway Signals typically reach cells through the binding of molecules to receptors on the cell surface. Over the course of seconds to minutes, the receptor interacts with other proteins or small molecules, changing their structure, concentration, or intracellular location. These changes can in turn cause other changes. One common ultimate effect is an alteration in transcription factor binding to target genes, in turn causing a change in mRNA levels. The term "signaling pathway" pre-dates modern systems biology, and represents a paradigm for information flow that is largely unidirectional and linear. Systems biology replaces this paradigm with that of a "regulatory network."

Definition of the Subject

Systems biology is the derivation of emergent properties of a multicomponent biological system through the construction of quantitative predictive models. These predictive models are typically formulated as networks. A portion of such a model is shown in Fig. 1. Human systems biology is the practice of systems biology to elucidate emergent properties of humans or any subcomponent of the human body, such as a cell. Human systems present challenges and opportunities not found in other organisms.

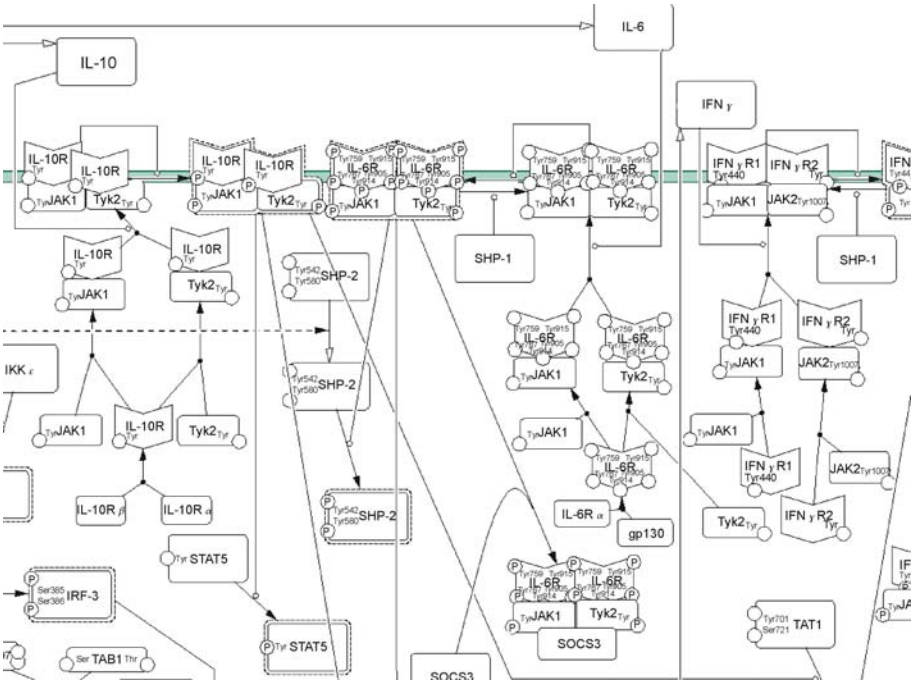

Systems Biology of Human Immunity and Disease, Figure 1
Example of a portion of a network model of a macrophage signaling network. This network was drafted with aid of CellDesigner software [34], and is reproduced with permission of The Systems Biology Institute (period)

Most problems in human systems biology are motivated by a desire to understand, predict, prevent, ameliorate, or cure a human disease. The human immune system is a mediator for most human diseases, and is the key system mediating autoimmune and infectious diseases. Many problems in human systems biology focus on the immune system.

As defined, human systems biology is a subset of systems biology. The subject merits special consideration largely because of medical importance. The human subset of systems biology is not an exceptionally distinct subset compared to other reasonable subsets, such as mammalian systems biology. However, it is fairly distinct from systems biology as applied to unicellular and invertebrate organisms.

Introduction

The difference between human systems biology and the practice of systems biology in other organisms is one of philosophy and emphasis [1]. The need to improve human health drives human systems biology. The emphasis of general systems biology is the development of technological, methodological, and algorithmic approaches to science. Human systems biology exploits these developments, but tends to be more focused on clinical endpoints. It is acceptable to approach these endpoints with incomplete, partial, hybrid, or modified versions of systems biology. Because these approaches include elements of systems biology, such as the analysis of high-throughput data, they are labeled as systems biology, even though they may not include all elements standard to general systems biology, such as the use of multiple orthologous data sets [5].

The needs of modern medicine set the context for human systems biology. From a data-driven point of view, there have been three major epochs of medicine: (1) observational medicine, (2) evidence-based medicine, and (3) predictive and personalized medicine [51]. These categories overlap to some extent, but they correspond fairly well to distinct periods of time. Observational medicine was prevalent from the dawn of civilization until early in the twentieth century. It is characterized by personal experience and oral tradition. An example and highlight of observational medicine was the discovery of digitalis as a treatment for dropsy, now known as edema. The discovery is credited in 1776 to William Withering, an English doctor who lived outside of London. Dr. Withering noticed that a patient with dropsy had been cured by a folk practitioner using a preparation from the foxglove plant. It is likely that this discovery had been made many times prior to the mid-eighteenth century, but lack of organized methods for communication and data handling kept the knowledge fragmented and incomplete. Dr. Withering was rich and connected. He was a member of the Lunar Society, a group of scientists founded by Matthew Boulton and Erasmus Darwin, the grandfather of Charles. These connections, foreshadowing those available to modern scientists through the internet, enabled the knowledge of foxglove to be preserved, distributed, and improved.

By the early twentieth century, incremental improvements to medical knowledge became increasingly common, and evidence-based medicine flourished. Advancements were enabled primarily by: (1) advances in communication, such as inexpensive publishing by printing presses, (2) chemistry, permitting the purification of active ingredients from botanical medicines, and (3) the nascent field of statistics. Statistics enabled legitimacy and falsifiability of medical facts. The first major application of statistics to medicine was the work of John Snow in 1843 to correlate water sources with cholera outbreaks. This correlation, and many others since, enabled preventive medicine. Chemistry, and later other fields of science and engineering, enabled reductionism. For example, reductionism enabled the identification of digitalis as the component of foxglove responsible for its effects on the cardiovascular system. The culmination of evidence-based medicine was the Framingham Heart Study circa 1948. Thousands of people from the town of Framingham and their descendants were observed longitudinally. Many health parameters were recorded. The Framingham study provided much of our modern knowledge concerning heart disease risk, including effects of diet, exercise, and aspirin. The Framingham Heart Study gave birth to the term "risk factor" and bolstered the recognition of prevention as the most cost-effective form of medicine. Prior to preventive medicine as a philosophy, a heart attack would be viewed as the beginning of a relationship between patient and doctor. With preventive medicine, the relationship would begin at birth or even before, and focus on modifying risk factors to prevent an attack from ever occurring. Data processing and analysis for such studies became increasingly sophisticated and required computers. Progress in evidence-based medicine was linked to the simultaneous revolution in computational power. In particular, the Framingham Heart Study benefited from scientific and sociological momentum generated in Word War II. Operations Research was largely born in the Battle of the Atlantic. The US population was willing to participate in large studies with society as the primary beneficiary. Command-and-control bureaucracies existed to coordinate large projects. Projects requiring the collaboration of hundreds of medi-

cal professionals were favorably viewed by researchers and their funding agencies.

By the end of the Human Genome Project at the beginning of the twenty-first century, limitations of the previous century's approach to evidence-based medicine were beginning to be recognized and solutions to these limitations were envisioned. Statistics, exemplified by linear regression models and chi-squared tests, required increasingly large cohorts to identify weaker risk factors. Knowledge gained from large cohorts was applicable only to populations (such as "all white males") and not individuals (such as the person reading this article). Studies became very expensive and could take decades or even centuries to elucidate correlations between causes and effects. The stage was set for the era of predictive and personalized medicine. A major approach to developing knowledge for this era of data-driven medicine has been and will be human systems biology.

In 1892, William Osler observed, "If it were not for the great variability among individuals medicine might be a science, not an art." Historically, the art of medicine was to personalize by trial and error. In the absence of personalization, a patient was prescribed a treatment based on population-based statistics. This is still the norm. If the treatment fails, the next best treatment from a population-inferred list will be prescribed. This process will repeat until the patient or doctor tires of it or dies, money becomes limiting, or effective treatment is achieved. Personalized medicine aims to match the best treatment for an individual without such perilous trial-and-error—and to turn Dr. Osler's art into a science. Previously, data for personalized models was based on orally conveyed family history, together gender and a crude estimate of genetic heritage, or race. Such personalization was tailored to subpopulations of people. 21st-century personalized medicine will develop predictions from an individual's genes and personal developmental and exposure history. Interventions will be tailored to that individual, not to a subpopulation to which the individual belongs. Systems-biology research creates the data and knowledge required by this approach. In particular, models are created based on the understanding of molecular mechanisms. Emergent properties arise from holistic analysis of the interactions in the modeled system. Each individual can be modeled based on their genetic risks, environmental exposures and their responses, biomarkers, and health assessments. Recommendations for particular medications, doses, interventions, and lifestyle changes will be personalized [12].

Expectations for modern research may be portrayed very highly in the popular press, with hopes for miracle cures. These may arise, but the most profound changes in medicine will most likely come from steady incremental advances. Research in all epochs of medicine has produced incremental advances. Cumulatively, these advances have enabled vast gains in human health (Fig. 2). Future incremental advances are expected throughout the epoch of personalized and preventive medicine begins.

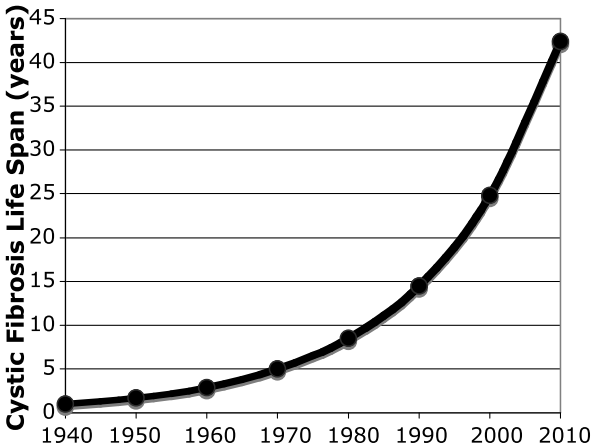

Systems Biology of Human Immunity and Disease, Figure 2
Exponential rise in expected life span for cystic fibrosis. A fairly low exponent, 5.5% in this case, can lead to dramatic advances in human health over time. The goals of human systems biology are to drive such exponential advances

The phrase "systems biology" was first used early in the twentieth century to describe analyses of ecosystems, but the phrase never became popular until the turn of 21st century with the establishment of several institutes and research groups focused on systems-biology research and using the term "systems biology" to describe that research. The timing of this surge in interest was almost entirely due to the completion of the Human Genome Project and the need to develop research methodologies appropriate for the post-genomic era. The complete genome sequence provided a near-comprehensive "parts list" of genes. Coinciding approximately with the Human Genome Project, high-throughput technologies for measuring other important parts of human systems were developed, including technologies for measuring proteins and transcripts [18]. Neologisms with the suffix "omics" describing these new fields of analysis have become popular, including "proteomics" and "transcriptomics." Other factors driving modern systems biology have been (1) the tremendous boost to human collaboration and information sharing provided by the internet and (2) the conceptualization of biology as an information science. The digital nature of the genome has led to a belief that all of the individual components and interactions of biological

systems are ultimately knowable, can be represented in data structures, and are amenable to computational analysis. The elucidation of the digital genome is a triumph of reductionism that has energized the pursuit of emergent properties. In the near future, complete sequences of the genomes of many humans will be known, further facilitating correlations between genotype and phenotype.

Modern systems biology focuses on systems at the cellular or subcellular level because systems of these complexities are considered to be neither too trivial to study nor too complex to comprehend. Some systems biology projects involve systems that are either more or less complex. In particular, human systems biology tends to focus on larger, more complex systems. For most problems of medical interest, it is generally anticipated that clinically useful predictions will require modeling at the multi-cellular level or at the level of the entire organism. The need to focus on very complex systems and to make clinically useful predictions underlies fundamental differences between the conduct of systems biology in humans and in other organisms.

There are tens of thousands of genes, small molecules, and proteins in a typical cellular system. Properties of these components and of the entire system need to be studied under many different conditions and dynamically over time periods ranging from seconds to years. There are several important generalizations that can be made about the techniques required to study such systems. Systems biology requires: (1) the analysis of high-throughput data sets, (2) the analysis of multiple qualitatively distinct different types of data ("orthologous data"), (3) a multidisciplinary collaborative team capable of both data generation ("bench work") and data analysis ("computational biology"). Also, because systems biology makes predictions, its practice is necessarily iterative: a model is built, predictions are tested, discrepancies between predictions and measurements are used to refine the model, predictions of the new model are tested, and so on until the incremental utility of the model refinements becomes small [47].

The application of the above approach to science is not new; it is merely new to biology. Other disciplines have used collaborative research employing multiple high-throughput data sources to produce models and emergent predictions. For example, in meteorology, a daily weather report is a simple prediction derived from such a process. One of the utilities of the phrase "systems biology" is that it brings attention to the methodological differences between this approach and the more prevalent reductionist approaches in biology during the twentieth century. The coining of the phrase thus helps catalyze a sociological change in biology.

A number of disciplines that might semantically be considered to fit the definition of systems biology are not usually implied by users of the phrase because these disciplines were well established prior to the completion of the Human Genome Project. These include ecology, physiology, and biochemistry. In these fields, research may involve formulation of hypotheses as networks, and prediction of emergent properties of multicomponent systems [26].

Challenges and Solutions for Human Systems Biology

Human systems biology requires asking questions that are distinct from typical questions asked of prokaryotic systems. These distinct questions require distinct methodologies to answer [30].

There are many challenges for human systems biology. Underlying network models are far more complex than those of prokaryotes [43]. There are more components, or network nodes, of nearly every type. For example, there are more genes, as discussed in the accompanying chapter on genome complexity. Each gene has more cis-regulatory elements. On average, any given gene has more functions and products. These products are primarily mRNAs. On average, any given mRNA has more functions and products. These products are primarily proteins. Eukaryotic cells have more cellular compartments, including mitochondria, lysomomes, phagosomes, a nucleus, endoplasmic reticulum, and secretory granules. Humans have many more cells, and more cell types, and they are organized into different tissues [6]. Furthermore, there is no guarantee that insights gained from one human tissue will be easily extrapolatable to other human tissues; the same protein may play a very different role. In addition to many more nodes in human networks, there are many more interactions. In particular, the number and complexity of interactions operating on cis-regulatory elements of eukaryotic genes is greater (Fig. 3) [42,48].

Not only are human systems much more complex than other systems, but it is much harder to acquire data from these systems. Many experiments possible in other organisms are impossible in humans due to ethical concerns. Such concerns prevent the intentional creation of humans with identical genomes. Humans have very long life spans that essentially rule out acquiring data over the life span of particular individuals, particularly for systems biology, which requires iteration. Projects that might take weeks in a prokaryotic context would take many centuries if the experimental strategies were directly applied to human problems. The cost of performing experiments on human

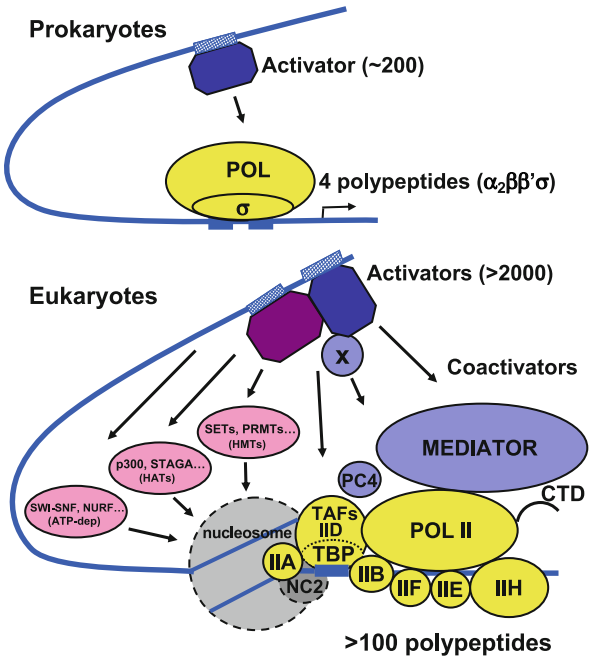

Systems Biology of Human Immunity and Disease, Figure 3
Eukaryotic versus prokaryotic cis-regulation. The number of factors involved in regulating expression of a eukaryotic gene is at least an order of magnitude greater than for a prokaryotic gene. The complexity of the interactions is also greater. Reprinted from [42]

systems also tends to be much higher than in other systems.

As a consequence of the difficulty of research in mammals, whether that research is reductionist or holistic, on average for any given node or edge of a human network there is less prior information available than there would be for a simple model organism. In many cases, a literature or database search for information pertaining to a human gene will yield either no information or information restricted to a few cell types under only a few conditions [11]. For example, the target genes and sequences of most human transcription factors are completely unknown, leading to substantial uncertainties in models of human nuclear regulatory networks.

There are many solutions to these challenges. Some solutions remain visionary or require technologies that are in development, such as the ability to sequence an individuals genome for less than $1,000. If the sole goal of human research was knowledge for the sake of knowledge, then the most economically efficient approach to most human systems biology problems would be to wait for these technologies to be fully developed before applying them directly to disease research. However, medicine carries a personal urgency not found in most other areas of research. Delays in research can be measured in human life and well being. So in addition to investing in technology and methodology, human systems biology must work with the tools at hand.

Because so much less is known about the basic elements of human systems, a continued focus on the individual nodes and edges is important. Human systems biology projects do not displace traditional reductionist biology projects, but rather rely on them. A key fundamental question to be asked of all possible nodes in a network is that of membership: "Does this node belong in the network?" The same question can be asked of each possible edge. The various qualities of the network components also need to be determined. These include directionality of directed edges and rate constants for parts of the network modeled dynamically. If the most important nodes and edges in a network can be determined, future reductionist research can be appropriately directed to expand information on these network components. The complexity of human systems, the difficulty of automatically parsing the literature, and the high false positive rate in many human high-throughput data sets demands increased expert curation at all steps in the process. The sheer volume of the data makes such a call for human curation seemingly futile, but currently without such curation, most models will not make useful medical predictions. Appropriately directed projects will employ automated tools to reduce the burden of curation. For this purpose some of the most useful tools will be interfaces that present curators with information in forms that are easy to comprehend and integrate manually and that also permit decisions and rationales to be quickly recorded. The visual computer interfaces imagined for the 2002 science fiction movie "Minority Report" serve to stimulate ideas of what such interfaces might become.

Human systems biologists must also model simpler, or smaller, networks than would be appropriate for a simple model organism. Although somewhat paradoxical, this need arises from the intense need for data curation and the difficulty of acquiring high-quality data. The more complex the systems, the greater the number of free parameters. For larger networks, the lack of experimental data prevents these parameters from being constrained, and so no statistical confidence can be placed in predictions. Therefore, a reasonable approach for building a predictive network in a human system is to start very small. In some cases, networks with as few as two nodes and a single connecting edge can form a reasonable nucleus for the iterative process of building a systems biology network. Although such a network would be too trivial to discuss in

a prokaryotic setting, it might be capable of making a clinically useful prediction and so would merit consideration in a human disease setting.

One possible approach to reducing the network complexity is to focus on only a few key nodes. Another is to model systems at grosser levels [9]. Thus rather than representing each protein as a node, one might group a set of proteins as a conceptual "module," and only represent the module as a node in the network. However, this approach requires exceptional care and curation, as it becomes hard to define and even harder to measure quantitative properties of modules [30]. One approach to developing hypotheses related to the functionality of modules is to perform comparative network analysis between different vertebrates [8]. Such analysis requires at least a moderate understanding of a network in at least two systems, which is currently rare, but may become more available over the next few decades.

It is easy for researchers to be distracted by the vast quantity of high-throughput data available, and to yield to reporting analyzes that are the result of automated in silico predictions of algorithms that often cannot be tested against gold standards. Human curation is often viewed as subjective, as it is hard to provide numerical statistics associated with such curation. Numerical statistics, on the other hand, are fairly easy to provide with output from automated algorithms, and these outputs are considered objective. From a medical point of view, however, the criterion of utility must trump evaluation of research based on simple ratio of subjectivity to objectivity. Medical objectivity must be evaluated at a deeper level of objectivity: quantitative measures of improved health care. Frequently, the output of automated algorithms is indeed objective, but is also nonsense. Examples of such nonsense are often manifested as very large networks consisting of hundreds or thousands of nodes and edges. Such networks are said to resemble hairballs and have been termed "ridiculomes." These networks have little if any predictive capability [32]. A weather report, as published in a daily newspaper, offers an example of a prediction in a useful form. Systems biology predictions should aim to be more like weather reports and less like hairballs.

Despite the large number of obstacles facing human research, there are a few paths to knowledge that work very well in humans. These derive largely from the ability of human research subjects to intellectually contribute to a research project. This is especially true for phenotypes of higher cognitive function. The most successful examples of human-specific research projects are family and genome-wide association studies. Volunteers with specific phenotypes bring themselves to the attention of researchers, permitting the genetic difference between affected and unaffected individuals to be determined. This is the most powerful and productive method that human systems biologists currently have for identifying the genes that are key nodes in human disease systems.

Given the fairly large number of differences in the state of knowledge and in the constraints on how research might be performed, the genre of questions asked by human systems biologists differs markedly from questions asked in other systems. Whereas design of a synthetic microorganism from the ground up might be a reasonable experimental agenda for a microbial biofuel project, other questions are more natural in human medicine.

Human-Specific Problems

It became fairly clear by the height of the era of evidence-based medicine that prevention has the highest ratio of benefit versus cost to society. Prevention is most effective when coupled to prediction. Interventions have risks and costs, so should be targeted to those predicted to have sufficient risk to benefit from a preventive intervention. Therefore a major goal of systems biology is to develop personalized predictive risk assessments. These risk assessments can be absolute (e. g., an individual's risk at the time of birth or at the time of a clinic visit) or conditional (e. g., risk if the individual stopped smoking). The types of outcomes that are useful to predict for an individual include whether or not a vaccination is protective, and whether a disease might occur and when it is likely to do so. A particular world health challenge today is the unpredictability of the effectiveness of the BCG vaccine for tuberculosis. Individuals who receive the vaccine have little way of knowing whether or not they will develop protective immunity [3]. The identification of biomarkers enabling such a prediction would be useful for counseling individuals as well as developing society-wide measures for preventing tuberculosis. Predictions at the level of a cell or tissue include what developmental programs these cells might follow or specific outcomes such as apoptosis, endocytosis, or mobility.

Other types of questions that can be asked in human systems include:

1. The identification of key information-transducing nodes in a system, as well as the inference of nodes that must be part of the system but are currently unsuspected or unknown. Such inference would be based on experimental results that differed from predicted results.
2. The identification of therapeutic drug targets [19].

3. Prediction of the effects of a gene knockout or knockdown, or overexpression of that gene.
4. Most generally, prediction of the effects of any perturbation to a system, including chemical and environmental perturbations.
5. Prediction of a particular property of a node or edge, such as Gene Ontology category, the presence of negative regulation, or the presence of positive or negative feedback.
6. Prediction of the concentration or flux or production of a specific gene, mRNA, or protein under a specific set of conditions at a specific time in a system.

The most successful projects will define exactly what questions will be addressed by the project before the project begins, and will define measurable quantitative variables that will be predicted by models [23]. The ability of a research group to precisely and accurately make quantitative predictions is one useful measure of whether a particular research methodology and approach is working effectively. Some questions lend themselves more naturally to quantitative predictions, such as those related to flux or concentration of small molecules. Others, such as the identification of drug targets, are very difficult to evaluate on any timescale less than a few decades.

Systems biology must be applied to all human diseases, as the urgency of clinical discovery is paramount. However, from the point of view of leveraging human systems biology research to gain fundamental understanding of a system, some human diseases may lend themselves more readily to knowledge discovery. These will be the diseases for which the etiology lies contained in a fairly small network and can be understood by understanding the interactions of a relatively few nodes. Thus, one might predict that research with Noonan Syndrome, which appears to be related to defects within the Ras signaling network may lead more immediately to fundamental understanding than research in type 1 diabetes, which appears to have etiologic effects spread throughout many cell types that arise from both genetic and environmental influences, and that act at disparate timepoints throughout the life of the individual.

Examples

Research in Noonan Syndrome represents an elegant example of systems biology as applied to human disease. Phenotypes of Noonan Syndrome include congenital heart malformation, short stature, and learning problems. Approximately 1 in 2,000 children are born with Noonan Syndrome. The range and severity of features can vary greatly between individuals, suggesting that there are multiple causes of the syndrome that affect common etiologic pathways, but in subtly different manners. Frequent transmission from parent to child suggested a genetic defect with autosomal dominant inheritance. Traditional genetic approaches mapped Noonan Syndrome to chromosome 12q24. By 2001, it was shown that about half of all individuals with Noonan syndrome were caused by a mutation of the PTPN11 gene at position 12q24. The PTPN11 gene encodes a tyrosine phosphatase. This discovery provided the first insight into the formation of a predictive model for Noonan Syndrome. With an underlying hypothesis that Noonan Syndrome was the emergent result of the output of an unknown dysfunctional network, the discovery that PTPN11 was part of that network allowed orthologous data sets and analysis to be leveraged [28,29]. PTPN11 is a node in an intracellular signaling network, including the traditional Ras/MAP kinase signaling pathway. This network governs cell division and differentiation. This network was fairly well characterized prior to the discovery of the association of PTP11 with Noonan Syndrome, so the network knowledge could be immediately applied to predict other nodes that, when perturbed, might also result in a Noonan Syndrome phenotype [13]. When tested, using an iterative network-refinement approach, several of these genes were also found to be fundamental to Noonan Syndrome etiology. These genes include RAF1, SOS1, HRAS, KRAS, and RAF1 [36,39,44,46,52].

Genetic association studies show great promise for breaking into etiologic network models for other human diseases. The technology for association studies for autoimmune diseases has been progressing steadily since the 1970s. However, until 2006, the only major genetic association with these diseases had been the MHC locus, which encodes the molecular complexes that train the immune system to recognize particular peptides as activators. The accumulated progress in technology moved research across a discovery threshold early in the twenty-first century and a large number of additional gene associations in a variety of autoimmune diseases have been discovered, as well in other complex human diseases. These include the discovery of IFIH1 for type 2 diabetes [45], ITPR3 for type 1 diabetes [40], and IL7R for multiple sclerosis [15,22]. Further progress in these diseases will be somewhat slower than for Noonan Syndrome, but there is every reason to believe that these larger and more complex networks can be modeled in clinically useful ways by applying the techniques discussed in the previous section.

In addition to genetic association studies, data analysis from high-throughput sources can also help enumerate key nodes in human networks. Carefully curated analyzes of orthologous high-throughput data sets can yield lists of

nodes that are likely to be fundamental to predictive network models. One of the more promising human systems for this systems biology approach is modeling macrophage activation through innate immune receptors [2]. Non-predictive network diagrams based on extensive literature analysis are available for the macrophage [34]. Such diagrams facilitate the development of predictive network hypotheses [35].

Rather than breaking into an unknown network, Bergholdt et al. [7] have approached the type 1 diabetes network by assuming that it is a subnetwork of current iterations of the global human protein-protein interaction network. They take key-node analysis a step further by developing a statistic for key subnetworks. Significance is computed from genetic association data for all genes in the subnetwork, as well as the interactions between these genes. This approach is amenable to further refinement by the incorporation of additional orthologous data sets, such as function and expression, into the significance metric. Bergholdt et al. succeed in identifying a number of candidate subnetworks that may transduce signals that cause diabetes.

Human macrophages are tractable experimental systems even when taken out of the context of the human body, and much of the macrophage regulatory network operates in a near native fashion even when isolated from other human cells. Because systems biology of the human immune system can in cases be reduced to the analysis of single cells, it shares some simplicities with prokaryotic systems biology. Systems biology of the human immune system is therefore a natural initial focus for human systems biology.

A large number of high-throughput data sets have been acquired from macrophages, particularly time-series following stimulation through toll-like receptors. Several preliminary comparative network analyzes are available [21,37]. Reasonable near-term expectations from such studies, which may involve dozens of researchers and years of research, are the addition of one or several nodes and key interactions to a basic network such as the identification of the role of ATF3 as a important negative regulator of macrophage activation [14] or the bulk identification of many key nodes without their interactions such as the identification of all transcriptionally active transcription factors [33,41]. Studies of signaling networks in T-cells have also resulted in predictive models, but are impeded by the resistance of T-cells to some forms of genetic perturbations, such as RNAi interference [25].

A lot of clinical value is currently being generated through analyzes of high-throughput data sets that result in predictive models for disease [27]. In some cases these analyzes are done on only one type of data, such as microarray data, and are not orthologous data sets in the truest sense. Furthermore, the results are frequently statements of correlation, rather than causation, and thus are a fairly preliminary step towards a true understanding of the system, or a true predictive etiologic model. Nevertheless these predictions are amenable to iterative refinement with a good prospect that such refinements will eventually elevate these models to the level of understanding. Examples of such efforts include diagnosis and stratification.

Statistical methodologies such as discriminant or principle components analysis can extract variables that are highly informative for classifying systems. Quantitative values can be obtained from many thousand nodes in systems with known properties. For example, one can measure the expression of thousands of genes in a set of malignant tumors and controls. Simplified classifiers can be extracted from these many thousand variables. Clinically useful stratification of breast cancers into subtypes that respond preferentially to different therapeutic interventions has been one example of this approach [16,49,50]. These multivariate approaches also show promise for diagnosing disease, including acute infections [38], arthritis [4], and diabetes [24]. These diagnostic techniques show promise for detecting presymptomatic disease, in which cases the semantic line between prediction and diagnosis blurs considerably. These methodologies can also stratify disease, allowing for the tailoring of specific therapies to the disease. Additionally, responses to diseases, medications, therapies, or the environment can themselves be stratified, permitting the early identification of adverse reactions and changes towards a more positive therapeutic direction.

Future Directions

The most productive element of the systems biology paradigm is collaboration. Collaboration between scientists with similar skills allows projects of greater size and complexity to be tackled. Collaboration between scientists with different knowledge and skill sets permits knowledge to be combined in manners previously unimaginable and opens previously inaccessible vistas of knowledge. For decades, both types of collaboration have been widely used in other branches of science, such as meteorology, astronomy, and particle physics. Therefore the most promising future directions for system biology are those that maximize collaborations.

These collaborations will be enabled by new methods of data sharing, most likely leveraging the worldwide web. Interactive and updateable sources of information such as

wikis and blogs will be key. Many online scientific journals have recently begun to encourage online commenting of papers. New methods of collaboration may evolve. These may involve virtual collaboration and might leverage peer-to-peer networks such as Facebook. The Nature publishing group has initiated such a tool, Nature Network.

New societal methods of career recognition and funding will likely stimulate systems biology. Traditional methods have recognized single-author or first author papers and tend to focus on journal publishing as a measure of productivity and utility. New methods will recognize collaboration and non-traditional measures of productivity, such as the creation of online resources. Recently, the NIH has altered its tenure track rules in recognition of these needs. Many of these changes will require a new generation of researchers willing to embrace them. Progress in systems biology will be increasingly driven by distributed worldwide collaborations that may involve researchers unaware of each other, with the collaborative element being contributions to virtual online projects. Projects will involve both specialists and generalists. In many cases data will be generated by a specialist in an experimental technique, and analyzed by distinct specialists in analytical techniques. Generalists will need to have skills in multiple disciplines and will be needed to direct projects and integrate data. Among the traditional twentieth-century fields of biology, immunology required extensive multidisciplinary knowledge from its practitioners; it is not surprising that many of the pioneers of systems biology were immunologists. Investigators trained in the clinical practice of medicine will be needed to set clinical priorities and recognize opportunities for the transfer of information from bench to bedside. Walls between disciplines will need to erode.

New and improved technologies will increasingly yield lower-costs, higher-throughputs, better sensitivity, better reproducibility and precision, better accuracy, and more comprehensivity [17,20]. Sequencing technologies continue to show promise for all of these characteristics, and dropping costs may permit them to underlie many measurements. Transcriptomics is currently performed mostly with arrays which may have fundamental upper limits on accuracy. Future transcriptomics may breach the accuracy barrier by employing transcript enumeration technologies that aim to sequence, for example, every mRNA in a single cell. Improvement in computational prediction and direct measurement of transcription factor binding to targets will permit these interactions to be confidently added to networks, and will vastly improve the predictive power of models of nuclear regulation. New methods of creating, storing, and sharing medical information electronically must be adopted that allow efficient use of aggregate information for research while preserving patient confidentiality and their rights to control their own information.

Effective improvements to human health cannot be made solely by a small group of individuals. Society must participate. Individuals must be informed, must support research, and must be willing to participate in and receive therapeutic interventions. Systems biology must therefore be alert to the broad needs of society [10]. These needs can help focus priorities. A society willing to modify its diet but reluctant to take pharmaceuticals would inform systems biology to prioritize models that predict the effects of altered diets over models of drug metabolism, and vice versa. Because human systems biology relies on human participation in research, such as in the Framingham Study and in gene association studies, efforts must be made to continue to educate all people in a deep inquiry-based scientific knowledge and to encourage them to share their personal medical histories and experience with the common knowledge pool. Current cultural shifts in the US towards increased guarding of personal information and mistrust of scientific information will create substantial challenges for the next generation of systems biologists.

Bibliography

Primary Literature

1. Aderem A (2007) Systems biology. Curr Opin Biotechnol 18(4):331–332
2. Aderem A, Hood L (2001) Immunology in the post-genomic era. Nat Immunol 2(5):373–5
3. Aderem A, Smith KD (2004) A systems approach to dissecting immunity and inflammation. Semin Immunol 16(1):55–67
4. Allantaz F, Chaussabel D, Stichweh D, Bennett L, Allman W, Mejias A, Ardura M, Chung W, Wise C, Palucka K, Ramilo O, Punaro M, Banchereau J, Pascual V (2007) Blood leukocyte microarrays to diagnose systemic onset juvenile idiopathic arthritis and follow the response to IL-1 blockade. J Exp Med 204(9):2131–44
5. Auffray C, Imbeaud S, Roux-Rouquie M, Hood L (2003) From functional genomics to systems biology: concepts and practices. C R Biol 326(10–11):879–92
6. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A (2005) Reverse engineering of regulatory networks in human B cells. Nat Genet 37(4):382–90
7. Bergholdt R, Størling ZM, Lage K, Karlberg EO, Olason PI, Aalund M, Nerup J, Brunak S, Workman CT, Pociot F (2007) Integrative analysis for finding genes and networks involved in diabetes and other complex diseases. Genome Biol 8(11):R253
8. Beyer A, Bandyopadhyay S, Ideker T (2007) Integrating physical and genetic maps: from genomes to interaction networks. Nat Rev Genet 8(9):699–710
9. Bornholdt S (2005) Systems biology. Less is more in modeling large genetic networks. Science 310(5747):449–51

10. Buchanan A, Califano A, Kahn J, McPherson E, Robertson J, Brody B (2002) Pharmacogenetics: ethical issues and policy options. Kennedy Inst Ethics J 12(1):1–15
11. Chaussabel D (2004) Biomedical literature mining: challenges and solutions in the 'omics' era. Am J Pharmacogenomics 4(6):383–93
12. Daly AK (2007) Individualized drug therapy. Curr Opin Drug Discov Devel 10(1):29–36
13. Gelb BD, Tartaglia M (2006) Noonan syndrome and related disorders: dysregulated RAS-mitogen activated protein kinase signal transduction. Hum Mol Genet 15, Spec No 2:R220–6
14. Gilchrist M, Thorsson V, Li B, Rust AG, Korb M, Roach JC, Kennedy K, Hai T, Bolouri H, Aderem A (2006) Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4. Nature 441(7090):173–8
15. Gregory SG, Schmidt S, Seth P, Oksenberg JR, Hart J, Prokop A, Caillier SJ, Ban M, Goris A, Barcellos LF, Lincoln R, McCauley JL, Sawcer SJ, Compston DA, Dubois B, Hauser SL, Garcia-Blanco MA, Pericak-Vance MA, Haines JL; for the Multiple Sclerosis Genetics Group (2007) Interleukin 7 receptor alpha chain (IL7R) shows allelic and functional association with multiple sclerosis. Nat Genet 39(9):1083–91
16. Hoadley KA, Weigman VJ, Fan C, Sawyer LR, He X, Troester MA, Sartor CI, Rieger-House T, Bernard PS, Carey LA, Perou CM (2007) EGFR associated expression profiles vary with breast tumor subtype. BMC Genomics 8(1):258
17. Hood L (2002) A personal view of molecular technology and how it has changed biology. J Proteome Res 1(5):399–409
18. Hood L, Galas D (2003) The digital code of DNA. Nature 421(6921):444–8
19. Hood L, Perlmutter RM (2004) The impact of systems approaches on biological problems in drug discovery. Nat Biotechnol 22(10):1215–7
20. Hood L, Heath JR, Phelps ME, Lin B (2004) Systems biology and new technologies enable predictive and preventative medicine. Science 306(5696):640–3
21. Iliev DB, Roach JC, Mackenzie S, Planas JV, Goetz FW (2005) Endotoxin recognition: In fish or not in fish? FEBS Lett 579(29):6519–6528
22. International Multiple Sclerosis Genetics Consortium, Hafler DA, Compston A, Sawcer S, Lander ES, Daly MJ, De Jager PL, de Bakker PI, Gabriel SB, Mirel DB, Ivinson AJ, Pericak-Vance MA, Gregory SG, Rioux JD, McCauley JL, Haines JL, Barcellos LF, Cree B, Oksenberg JR, Hauser SL (2007) Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med 357(9):851–62
23. Janes KA, Yaffe MB (2006) Data-driven modelling of signal-transduction networks. Nat Rev Mol Cell Biol 7(11):820–8
24. Kaizer EC, Glaser CL, Chaussabel D, Banchereau J, Pascual V, White PC (2007) Gene expression in peripheral blood mononuclear cells from children with diabetes. J Clin Endocrinol Metab 92(9):3705–11
25. Kemp ML, Wille L, Lewis CL, Nicholson LB, Lauffenburger DA (2007) Quantitative network signal combinations downstream of TCR activation can predict IL-2 production response. J Immunol 178(8):4984–92
26. Kitano H (2002) Computational systems biology. Nature 420(6912):206–10
27. Kitano H (2007) The theory of biological robustness and its implication in cancer. Ernst Schering Res Found Workshop. (61):69–88
28. Kratz CP, Schubbert S, Bollag G, Niemeyer CM, Shannon KM, Zenker M (2006) Germline mutations in components of the Ras signaling pathway in Noonan syndrome and related disorders. Cell Cycle 5(15):1607–11
29. Lee ST, Ki CS, Lee HJ (2007) Mutation analysis of the genes involved in the Ras-mitogen-activated protein kinase (MAPK) pathway in Korean patients with Noonan syndrome. Clin Genet 72(2):150–5
30. Ma'ayan A, Jenkins SL, Neves S, Hasseldine A, Grace E, Dubin-Thaler B, Eungdamrong NJ, Weng G, Ram PT, Rice JJ, Kershenbaum A, Stolovitzky GA, Blitzer RD, Iyengar R (2005) Formation of regulatory patterns during signal propagation in a Mammalian cellular network. Science 309(5737):1078–83
31. Ma'ayan A, Jenkins SL, Goldfarb J, Iyengar R (2007) Network analysis of FDA approved drugs and their targets. Mt Sinai J Med 74(1):27–32. Ma'ayan A, Blitzer RD, Iyengar R (2005) Toward predictive models of mammalian cells. Annu Rev Biophys Biomol Struct 34:319–49
32. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7 Suppl 1:S7
33. Natarajan M, Lin KM, Hsueh RC, Sternweis PC, Ranganathan R (2006) A global analysis of cross-talk in a mammalian cellular signalling network. Nat Cell Biol 8(6):571–80
34. Oda K, Kitano H (2006) A comprehensive map of the toll-like receptor signaling network. Mol Syst Biol 2:2006.0015
35. Oda K, Matsuoka Y, Funahashi A, Kitano H (2005) A comprehensive pathway map of epidermal growth factor receptor signaling. Mol Syst Biol 1:2005.0010
36. Pandit B, Sarkozy A, Pennacchio LA, Carta C, Oishi K, Martinelli S, Pogna EA, Schackwitz W, Ustaszewska A, Landstrom A, Bos JM, Ommen SR, Esposito G, Lepri F, Faul C, Mundel P, Lopez Siguero JP, Tenconi R, Selicorni A, Rossi C, Mazzanti L, Torrente I, Marino B, Digilio MC, Zampino G, Ackerman MJ, Dallapiccola B, Tartaglia M, Gelb BD (2007) Gain-of-function RAF1 mutations cause Noonan and LEOPARD syndromes with hypertrophic cardiomyopathy. Nat Genet 39(8):1007–12
37. Purcell MK, Smith KD, Hood L, Winton JR, Roach JC (2006) Conservation of toll-like receptor signaling pathways in teleost fish. Comp Biochem Physiol Part D Genomics Proteomics 1(1):77–88
38. Ramilo O, Allman W, Chung W, Mejias A, Ardura M, Glaser C, Wittkowski KM, Piqueras B, Banchereau J, Palucka AK, Chaussabel D (2007) Gene expression patterns in blood leukocytes discriminate patients with acute infections. Blood 109(5):2066–77
39. Razzaque MA, Nishizawa T, Komoike Y, Yagi H, Furutani M, Amo R, Kamisago M, Momma K, Katayama H, Nakagawa M, Fujiwara Y, Matsushima M, Mizuno K, Tokuyama M, Hirota H, Muneuchi J, Higashinakagawa T, Matsuoka R (2007) Germline gain-of-function mutations in RAF1 cause Noonan syndrome. Nat Genet 39(8):1013–7
40. Roach JC, Deutsch K, Li S, Siegel AF, Bekris LM, Einhaus DC, Sheridan CM, Glusman G, Hood L, Lernmark AA, Janer M on behalf of the Swedish Childhood Diabetes Study Group and the Diabetes Incidence in Sweden Study Group. (2006) Genetic mapping at 3-kilobase resolution reveals inositol 1,4,5-triphosphate receptor 3 as a risk factor for type 1 diabetes in Sweden. Am J Hum Genet 79(4):614–627

41. Roach JC, Smith KD, Strobe KL, Nissen SM, Haudenschild CD, Zhou D, Vasicek TJ, Held GA, Stolovitzky GA, Hood L, Aderem A (2007) Transcription factor expression in lipopolysaccharide-activated peripheral blood derived mononuclear cells. PNAS 104(41):16245–50
42. Roeder RG (2003) The eukaryotic transcriptional machinery: complexities and mechanisms unforeseen. Nat Med 9(10):1239–44
43. Schlitt T, Brazma A (2007) Current approaches to gene regulatory network modelling. BMC Bioinformatics. 8 Suppl 6:S9
44. Schubbert S, Zenker M, Rowe SL, Boll S, Klein C, Bollag G, van der Burgt I, Musante L, Kalscheuer V, Wehner LE, Nguyen H, West B, Zhang KY, Sistermans E, Rauch A, Niemeyer CM, Shannon K, Kratz CP (2006) Germline KRAS mutations cause Noonan syndrome. Nat Genet 38(3):331–6
45. Smyth DJ, Cooper JD, Bailey R, Field S, Burren O, Smink LJ, Guja C, Ionescu-Tirgoviste C, Widmer B, Dunger DB, Savage DA, Walker NM, Clayton DG, Todd JA (2006) A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. Nat Genet 38:617–9
46. Sovik O, Schubbert S, Houge G, Steine SJ, Norgard G, Engelsen B, Njolstad PR, Shannon K, Molven A (2007) De novo HRAS and KRAS mutations in two siblings with short stature and neuro-cardio-facio-cutaneous features. J Med Genet 44(7):e84
47. Tian Q, Stepaniants SB, Mao M, Weng L, Feetham MC, Doyle MJ, Yi EC, Dai H, Thorsson V, Eng J, Goodlett D, Berger JP, Gunter B, Linseley PS, Stoughton RB, Aebersold R, Collins SJ, Hanlon WA, Hood LE (2004) Integrated genomic and proteomic analyses of gene expression in Mammalian cells. Mol Cell Proteomics 3(10):960–9
48. Tijan R (1995) Molecular machines that control genes. Scientific American. 272(2):54–61
49. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415(6871):530–6
50. Weigelt B, Hu Z, He X, Livasy C, Carey LA, Ewend MG, Glas AM, Perou CM, Van't Veer LJ (2005) Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer. Cancer Res 65(20):9155–8
51. Weston AD, Hood L (2004) Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. J Proteome Res 3(2):179–96
52. Zenker M, Horn D, Wieczorek D, Allanson J, Pauli S, van der Burgt I, Doerr HG, Gaspar H, Hofbeck M, Gillessen-Kaesbach G, Koch A, Meinecke P, Nowak A, Rauch A, Reif S, von Schnakenburg C, Seidel H, Wehner LE, Zweier C, Bauhuber S, Matejas V, Kratz CP, Thomas C, Kutsche K (2007) SOS1 is the second most common Noonan gene but plays no major role in cardio-facio-cutaneous syndrome. J Med Genet 44(10):651–6

Books and Reviews

Alon U (2006) An introduction to systems biology: Design principles of biological circuits. Chapman and Hall, Norwell
Bower JM, Bolouri H (2001) Computational modeling of genetic and biochemical networks. MIT Press, Cambridge
Capra F (1996) The web of life. Doubleday, New York
Carroll SB (2004) From DNA to diversity: Molecular genetics and the evolution of animal design. Blackwell Publishing Limited, Boston
Davidson EH (2006) The regulatory genome: Gene regulatory networks in development and evolution. Academic Press, San Diego
Fall CP et al (2002) Computational cell biology. Springer, New York
Goodsell DS (1992) The machinery of life. Springer, New York
Kaneko K (2006) Life: An introduction to complex systems biology. Springer, New York
Kauffman SA (1993) The origins of order: Self-organization and selection in evolution. Oxford University Press, Oxford
Kitano H (2001) Foundations of systems biology. MIT Press, Cambridge
Latchman DS (2004) Eukaryotic transcription factors. Elsevier, Amsterdam
Michal G (1999) Biochemical pathways. Wiley, New York
Palsson BO (2006) Systems biology: Properties of reconstructed networks. Cambridge University Press, Cambridge
Voit EO (2000) Computational analysis of biochemical systems: A practical guide for biochemists and molecular biologists. Cambridge University Press, Cambridge

Systems Biology, Introduction to

Robert A. Meyers
Ramtech Limited, Larkspur, USA

Systems biology is a relatively new branch of life sciences that is highly interdisciplinary and aims to provide a quantitative analysis and understanding of life systems within a mathematical framework. Systems biology is the study of the interconnected aspect of molecular, cellular, tissue, whole animal and ecological processes, and comprises mathematical and mechanistic studies of dynamical, mesoscopic, open, spatiotemporally defined, nonlinear, complex systems that are far from thermodynamic equilibrium. The animating principle of systems biology is that the most important behaviors and networks of interacting elements (e. g.,molecules, cells, etc.), rather than by individual elements and interactions. This really necessitates the integrated application of high-throughput measurement technologies and advanced computational methods to model and predict biological responses. The 11 articles comprising this section describe the basic aspects and processes of systems biology and the relationship of systems biology to complexity and systems science. The basis and content of each of the 11 articles in this section are described below. There are six additional articles that were nominated and accepted into other sections which are connected to complexity in Systems Biology, and supplement the 11 articles in this section. These six are listed at the end of this Introduction.

With rare exceptions, all known living organisms encode their genetic material in the form of double-stranded DNA, in one or more chromosomes, collectively referred to as the "genome" (▶ Genome Organization). The genome includes most of the information needed by the cells to stay alive, to differentiate into new cell types, and to perform their functions in the context of the organism. As such, it is the ultimate resource for identifying the full set of components in the living system. Eukaryotic genomes are much larger than strictly needed to encode the relatively modest set of genes in them, but several mechanisms give rise to a very complex transcriptome. It is necessary to understand how the genome is organized and how it evolved and be able to build suitable null hypotheses for testing whether predictions are likely to be real, and thereby understand whether specific observations are likely to be biologically meaningful.

Sequence data alone is currently of limited use for identifying the functional elements of a genome and for elucidating how these elements interact to control physiological processes.

The field of functional genomics aims to meet these challenges using the sequence data as a blueprint (▶ Functional Genomics for Characterization of Genome Sequences). In a broad sense, functional genomics is defined as the large-scale experimental study of gene function and interactions. In the above article, the techniques and challenges in functional genomics are illustrated in the context of an ever more common case scenario: given a complete genome sequence, how does one figure out what the sequence means? The current state of the art is interpreted to arrive at the function or functions of a given genome sequence.

Among the most fundamental problems in biology is deciphering the relationship between genotype and phenotype in the complex system of life forms. Understanding a complex system broadly requires that one (1) identify the elements, (2) determine the function of each element, (3) identify and characterize the interactions between elements, and (4) assemble all of this information into a mathematical model that accurately simulates the system and predicts its responses to novel perturbations (▶ Systems Genetics and Complex Traits). In the context of systems biology, this process begins with sequencing an organism's genome and identifying the functional elements, e. g. the genes. There are numerous methods for determining the function(s) of a given gene product, where "function" can describe both the specific biochemical activity carried out by the gene product and the role that activity plays in the organism's response to an environmental or developmental stimulus. Understanding complex genetics is of increasing relevance to the study of human health and is essential to the development of predictive, preventive, and personalized medicine.

Systems biology is the derivation of emergent properties of a multicomponent biological system through the construction of quantitative predictive models. These predictive models are typically formulated as networks. Human systems biology is the practice of systems biology to elucidate emergent properties of humans or any subcomponent of the human body, such as a cell (▶ Systems Biology of Human Immunity and Disease). Human systems present challenges and opportunities not found in other organisms. Most problems in human systems biology are motivated by a desire to understand, predict, prevent, ameliorate, or cure a human disease. The human immune system is a mediator for most human diseases, and is the key system mediating autoimmune and infectious diseases.

Understanding the operation of cellular networks is probably one of the most challenging and intellectually exciting scientific fields today (▶ Biological Models of Molecular Network Dynamics). Cellular networks are some of the most complex natural systems we know. Even in a "simple" organism such as E. coli, there are at least four thousand genes with many thousands of interactions between molecules of many different sizes. With the availability of new experimental and theoretical techniques, our understanding of the operation of cellular networks has made great strides in the last few decades. An important outcome of this work is the development of predictive quantitative models. Such models of cellular function will have a profound impact on our ability to manipulate living systems which will lead to new opportunities for generating energy, mitigating our impact on the biosphere and last but not least, opening up new approaches and understanding of important disease states such as cancer and aging.

Biological research over the past century or so has been dominated by reductionism – identifying and characterizing individual biomolecules – and has enjoyed enormous success. Throughout this history, however, it has become increasingly clear that an individual biomolecule can rarely account for a discrete biological function on its own. A biological process is almost always the result of a complex interplay of relationships amongst biomolecules, and the treatment of these relationships as a graph is a natural and useful abstraction (▶ Biomolecular Network Structure and Function).

Broadly speaking, a biomolecular network is a graph representation of relationships (of which there are many types) amongst a group of biomolecules. Vertices or nodes

represent biomolecules, including macromolecules such as genes, proteins, and RNAs, or small biomolecules like amino acids, sugars, and nucleic acids. An edge or link between two vertices indicates a relationship between the corresponding biomolecules, which could include physical interaction, genetic interaction, or a regulatory relationship (e. g., the protein product of gene A regulates the expression of gene B). This abstraction, although simplifying, converts a complex web of biological relationships into a mathematical graph, from which we can study its structural features as well as their implications on biological functions.

Data integration and model building have become essential activities in biological research as technological advancements continue to empower the measurement of biological data of increasing diversity and scale (► Biological Data Integration and Model Building). High-throughput technologies provide a wealth of global data sets (e. g. genomics, transcriptomics, proteomics, metabolomics), and the challenge becomes how to integrate this data to maximize the amount of useful biological information that can be extracted. Integrating biological data is important and challenging because of the nature of biology. Biological systems have evolved over the course of billions of years, and in that time biological mechanisms have become very diverse, with molecular machines of intricate detail. Thus, while there are certainly great general scientific principles to be distilled – such as the foundational evolutionary theory – much of biology is found in the details of these evolved systems. This emphasis on the details of systems and the history by which they came into being (i. e. evolution) are distinct features of biology as a science, and influence the need for large-scale data integration. Also, biological systems are responsive to varying environments, with potential system states influenced by the combinatorics of all possible molecular and environmental perturbations. Thus, data space in this realm is extraordinarily large. There is no shortage of possibilities to explore, and vast amounts of data will be needed to reengineer biological systems and understand their workings at a (near) complete level of detail or at a level where accurate prediction can be achieved.

Another reason data integration is essential, is that biology arises through the complex interworking of many components. Thus, to understand biology we must integrate biological data in a way that will allow us, not only to access all our acquired data efficiently, but even more importantly allow us to study, predict, modify, and even engineer the emergent properties of biological systems, where the whole is more than the sum of the parts.

Boolean Networks are a class of discrete dynamical systems that can be characterized by the interaction of a set of Boolean variables. Random Boolean Networks, which are ensembles of random network structures, are a simple model class for studying dynamical properties of gene regulatory networks (► Boolean Modeling of Biological Networks). Boolean Networks have been used as generic models for dynamics of complex systems of interacting entities, such as social and economic networks, neural networks, as well as gene and protein interaction networks. Despite their conceptual simplicity, Boolean Networks exhibit complex nonlinear behaviors that are, to this day, a challenging object of investigation for theoretical physicists and mathematicians. Further, a discretization of gene expression is often regarded as an experimentally justifiable simplification making Boolean network models attractive tools for the biologist.

Many processes in cell biology, such as those that carry out metabolism, the cell cycle, and various types of signaling, are comprised of biochemical reaction networks. It has proven useful to study these networks using computer simulations because they allow us to quantitatively investigate hypotheses about the networks. Deterministic simulations are sufficient to predict average behaviors at the population level, but they cannot address questions about noise, random switching between stable states of the system, or the behaviors of systems with very few molecules of key species. These topics are investigated with stochastic simulations (► Stochastic Models of Biological Processes). The dominant types of stochastic simulation methods that are used to investigate biochemical reaction networks are covered, as well as some of the results that have been found with them. As new biological experiments continue to reveal more detail about biological systems, and as computers continue to become more powerful, researchers will increasingly turn to simulation methods that can address stochastic and spatial details.

Systems biology has various definitions. Common features among accepted definitions generally involve the description and analysis of interacting biomolecular components. Systems analysis of a biological network is quickly demonstrating its utility as it helps to characterize biomolecular behavior that could not otherwise be produced by the individual components alone. Three areas in which systems analysis has been implemented in biology include: (1) the generation and statistical analysis of high-throughput data in an effort to catalog and characterize cellular components; (2) the construction and analysis of computational models for various biological systems (e. g., metabolism, signaling, and transcriptional regulation); and (3) the integration of the knowledge of parts and

computational models to predict and engineer biological systems (synthetic biology). Metabolism, as a system, has played an important role in the development of systems biology, especially in the modeling sense (► Metabolic Systems Biology). This is because the network components (e. g., enzymes and metabolites) have been studied in detail for decades, and many links between components have been experimentally characterized. Metabolic systems biology, compared to systems biology in general, entails the computational analysis of these enzymes and metabolites and the metabolic pathways in which they participate. Metabolic systems biology, using genome-scale metabolic network reconstructions and their models, has helped (1) to elucidate biomolecular function; (2) to predict phenotypic behavior; (3) to discover new biological knowledge; and (4) to design experiments for engineering applications. Constraint-based methods have played a pivotal role in the analysis of large and genome-scale metabolic networks. The structure, mathematical formulation, and analytical techniques of constraints-based methods have also paved the way for the successful modeling of other complex biological networks, such as transcriptional regulation and signaling networks.

Finally, ecological systems are paradigmatic examples of complex systems. For example, consider the thousands of species interacting in complex ways within rich communities such as tropical rainforests or coral reefs. The most pressing questions ecologists face deal with concepts such as stability, resilience, thresholds and non-linearities which are at the core of the sciences of complexity. How robust are these cathedrals of biodiversity? At which rate will they disassemble as a consequence of global change. For example, one of the long-standing questions in ecology is the relationship between complexity and stability. A review of some of the applications of the complexity sciences in the realm of ecological systems is presented (► Ecological Systems). Predicting the consequences of global change on biodiversity requires an interdisciplinary approach in which complexity approaches may be very useful. Information theory and diversity, networks, complex dynamics and spatiotemporal dynamics are all discussed relative to ecology as a complex system.

Additional articles connected to Systems Biology:

- ► Consciousness and Complexity
- ► Molecular Evolution, Networks in
- ► Complexity in Systems Level Biology and Genetics: Statistical Perspectives
- ► Fractals in Biology
- ► Biochemistry, Chaotic Dynamics, Noise, and Fractal Space in
- ► Exobiology and Complexity

Systems and Control, Introduction to

Matthias Kawski
Department of Mathematics and Statistics,
Arizona State University, Tempe, USA

Control of dynamical systems has a long history: Watt's automatic centrifugal governor designed to regulate steam engines in the 1780s is considered a precursor of modern feedback control. Similarly, Bernoulli's work in the 1690s on the brachystochrone problem is a progenitor of optimal control. A distinguishing external feature of *controlled* dynamical systems is the presence of inputs that interact with the system in the form of deliberate controls, or unavoidable perturbations. Commonly control systems also have outputs (observations) that typically provide only incomplete information about the state of the system. Systems and control theory utilizes a broad array of mathematical disciplines – but a uniting, characteristic feature is the kind of questions being asked. In the 1950s the field became more formalized, it began to develop its own distinctive identity, and it has rapidly evolved ever since. Naturally, the linear theory developed first and quickly became a mature subject, with controllers now pervading everyday life, from thermostats, dozens of controllers in every automobile, to cell phones, attitude control of satellites, electric power networks, highway traffic control, to name just a few.

The articles in this section focus on the more recent developments, articulating how the fundamental problems and questions of systems and control theory are addressed in ever new settings, on new tools, and new applications.

Whereas in classical dynamical systems one tries to predict the uniquely determined future, a basic distinguishing question in control theory asks whether it is possible to steer the system to any desired target. This question about *controllability* has a simple and elegant answer in the case of linear systems. While much also is known for nonlinear systems, many challenges remain in both the finite dimensional setting [see ► Finite Dimensional Controllability] and in the setting of *distributed systems* modeled by partial differential equations [see ► Control of Non-linear Partial Differential Equations].

Assuming controllability using controls that are functions of time, the next question asks whether one can *automate* the control using feedback, that is, by making it a function of the state or of a measured output. Whereas for linear systems the theory is straightforward, unavoidable topological obstacles make nonlinear feedback stabilization problems much more challenging [see ► Stability and Feedback Stabilization]. A notion dual to control-

lability is that of observability: Rather than asking about the map from inputs to the state, a system is observable if the history of the outputs (measurements) determines the state of the system [see ► Observability (Deterministic Systems) and Realization Theory]. The typical engineering problem involves much more complex multi-tasking, dealing with multiple objectives in the presence of diverse constraints. Much of this well developed theory of output regulation, disturbance rejection, while attending to performance criteria, has recently been generalized to nonlinear systems [see ► System Regulation and Design, Geometric and Algebraic Methods in]. Generally, if there is one, then there are many control strategies that achieve a certain task – naturally one likes to single out a strategy that is the best. This leads to optimal control theory, which supersedes much of the classical calculus of variations, and which has rich interfaces with Lagrangian and Hamiltonian mechanics [see ► Maximum Principle in Optimal Control].

In many applications all one has to work with are sampled data for inputs and measured outputs. Moreover, commonly these are also corrupted by measurement errors and noises. A fundamental problem is to devise algorithms to *identify* the best model system that could generate these input-output pairs. Characteristically such algorithm should work in real time and improve the model upon newly available measurements. This well developed broad field includes geometric approaches to stochastic systems [see ► Stochastic Noises, Observation, Identification and Realization with] as well as more formal learning theories [see ► Learning, System Identification, and Complexity].

Systems and control theory is unified by a common core of questions, but employs a diverse array of models. Consequently it interfaces a wide range of mathematical disciplines, utilizes many different mathematical tools, but also fosters new development in diverse subdisciplines. An area that has evolved closely together with control is nonsmooth analysis with its rich theory of tangent objects to nonsmooth sets. Not only are controlled dynamical systems naturally modeled by differential inclusions, but control intrinsically is full of objects that are not differentiable in a classical sense. Most important is the value function in optimal control which encodes the minimal cost (time) to the target. Closely related to classical dynamic programming, this function ought to satisfy the Hamilton–Jacobi–Bellmann equation, a first order partial differential equation, but the value function generally is not differentiable in a traditional sense [see ► Nonsmooth Analysis in Systems and Control Theory]. Complementary to the nonsmooth approach are differential geometric approaches which include Lie theory and symplectic geometry. Closely related is a functional analytic operator calculus introduced into control by Agrachëv and Gamkrelidze which linearizes problems by imbedding them into infinite dimensional settings. This dramatically facilitates formal manipulations of nonlinear objects, and also gives new ways of looking at the underlying geometry [see ► Chronological Calculus in Systems and Control Theory].

Many of the early successful engineering implementations of systems and control theory used models with both continuous time and states (differential equations). But with the advent of ubiquitous digital controllers as well as ever broadening fields of applications the theories for increasingly diverse models are becoming well developed, too. These include, systems with or mix of continuous and discrete states (suitable, for example, to model systems with hysteresis) [see ► Hybrid Control Systems] and discrete time Hamiltonian and Lagrangian systems [see ► Discrete Control Systems] There are many new applications with quickly developing theories such as control of biological and biomedical applications, social dynamics, manufacturing systems, financial markets, and many more. Some of the most intense research in recent years has focused on systems with distributed intelligent agents (controllers) each of which has access to only local information [see ► Robotic Networks, Distributed Algorithms for]. Compared to these emerging areas the geometric theory of mechanical systems is one of the most mature and thriving research areas with scores of exciting new problems. It serves as a role model for the depth of geometric insight and for its well-understood close interconnections with many other disciplines [see ► Mechanical Systems: Symmetries and Reduction].

Acknowledgment

This work was partially supported by the National Science Foundation through the grant DMS 05-09030.

Systems Genetics and Complex Traits

Gregory W. Carter, Aimée M. Dudley
Institute for Systems Biology, Seattle, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Measuring Genetic Interactions Genome-Wide

Constructing Genetic Interaction Networks
Interpreting Genetic Interactions
Future Directions
Acknowledgments
Bibliography

Glossary

Allele A copy or alternate version of a gene.

Complex phenotype A trait (or phenotype) caused by polymorphisms in multiple genes.

Diploid An organism with two copies of each chromosome.

Epistasis A genetic interaction in which the double-mutant phenotype is identical to one of the single-mutants (masking epistasis); sometimes used as a general term for genetic interaction.

Genetic interaction A relationship that characterizes how two (or more) genetic perturbations or alleles combine to affect a phenotype.

Genetic network A representation of functional relationships between genes, gene products, or gene perturbations, often displayed as a graph.

Genetic perturbation The modification of a gene's presence, structure, or activity by a change in the DNA sequence in or near a gene.

Genetic transformation Genetic modification of a cell by the introduction of exogenous genetic material.

Haploid An organism with one copy of each chromosome.

Haplotype A set of polymorphisms that are linked and thus inherited as a unit.

Homolog One of two or more genes with similar DNA sequence due to shared ancestry.

Hybrid A diploid offspring from the mating of individuals from two distinct populations.

Intercross The mating of individuals from two distinct populations.

Isogenic A term describing strains of model organisms that are derived from the same individual or inbred line, e. g. the yeast deletion collection, in which each strain differs only by the presence of a specifically engineered mutation (e. g. a gene knock-out).

Masking epistasis A genetic interaction in which the double-mutant phenotype is identical to one of the single-mutants.

Mendelian genetics The study of traits caused by variations of a single gene.

Microarray hybridization Sequence-specific binding of nucleic acid (DNA or RNA) to short nucleic acid sequences tethered to a microarray chip.

Molecular barcode A short, unique sequence inserted into DNA that can be used to identify or quantify a gene, gene product, or strain.

Network motif A small subnetwork repeated in a larger network, suggesting repeated organization for a specific function.

Phenotype An observable property of an organism; sometimes implies an aberrant property, e. g. the mutant has a phenotype.

Pleiotropy The phenomenon in which a mutation in a single gene causes multiple phenotypes.

Quantitative trait locus (QTL) A chromosomal region linked to a measurable phenotype.

RNA interference (RNAi) A method of silencing a gene's expression by introducing a small RNA molecule with sequences complementary to a portion of the gene's mRNA transcript.

Reciprocal-hemizygote A hybrid in which one copy of a gene or chromosomal region has been removed.

Sporulation The formation of spores in fungi, produced by meiosis.

Synthetic sickness or lethality (SSL) A genetic interaction in which two genetic perturbations without individual fitness effects combine to cause a fitness defect or inviability.

3′ Untranslated region (UTR) The RNA sequence that follows the protein-coding region of a messenger RNA.

Definition of the Subject

Among the most fundamental problems in biology is deciphering the relationship between genotype and phenotype. Genetics, the study of how the DNA sequence of an individual (genotype) affects an observable characteristic (phenotype), has many properties of a systems science. Organisms contain large number of protein-coding genes, with ~500 predicted in the simple bacterium *Mycoplasma genitalium*, ~25,000 in the human genome, and over 50,000 in some plants. The relationship between genotype and phenotype can be complex, with multiple traits per gene (pleiotropy) and multiple genes per trait (complex phenotypes). Genes are the blueprints for both RNA transcripts and proteins that interact physically and functionally to produce the emergent behaviors that characterize life. Furthermore, biological systems can be systematically studied by perturbing elements of the system, such as genes or environmental factors, and quantitatively measuring the effects. This research strategy is generally considered a central component of systems bi-

ology [39,40], and here we refer to its application to the genotype-phenotype problem as *systems genetics*.

Understanding a complex system broadly requires one to (1) identify the elements, (2) determine the function of each element, (3) identify and characterize the interactions between elements, and (4) assemble all of this information into a mathematical model that accurately simulates the system and predicts its responses to novel perturbations. In the context of biology, this process usually begins with sequencing an organism's genome and identifying the functional elements, e. g. the genes. There are numerous methods for determining the function(s) of a given gene product, where "function" can describe both the specific activity carried out by the gene product and the role that activity plays in the organism's response to an environmental or developmental stimulus. For example, molecular or biochemical methods may reveal a protein to be enzyme that converts a substrate to a product, while genetic analysis may reveal that this activity is required for growth in the presence if a specific drug. Similarly, there are both molecular and genetic means of determining which gene products interact with each other, by detecting physical interactions (e. g. protein binding) and genetic interactions (e. g. a mutation in one gene modifies the effect of a mutation in a second gene). These interactions vary with respect to the nature, strength, and directionality of the interaction. Integrating these data into a coherent computational model of the system is a major challenge of systems biology, with early examples including a comprehensive analysis of the galactose utilization pathway in yeast [40] and bacterial chemotaxis [70].

Systems genetics approached to this problem use large-scale, high-throughput methods to decipher the network of gene functions and genetic interactions in an organism. Apart from their genome-wide scale, the data and methods used often mimic those of conventional genetic analysis. In classical Mendelian genetics, genes are defined as the entities by which traits are inherited. A specific trait exhibited by an organism thus depends on the version(s) of the corresponding gene, known as an allele, carried by that individual. Modern molecular biology has narrowed the concept of a gene to a distinct DNA sequence that is transcribed to encode an RNA or protein product. However, few phenotypes are entirely determined by a single gene. In most cases, multiple genes interact to confer a phenotype. These interactions can vary from simple modifier relationships, in which an allele of one gene alters the phenotype of another gene, to multiple genes with complicated interdependencies. The study of how multiple genes interact to influence a trait is generally known as complex genetics.

Understanding complex genetics is of increasing relevance to the study of human health and is essential to the development of predictive, preventive, and personalized medicine. Many diseases, such as asthma, diabetes, heart disease, and cancer, show a degree of heritability that cannot be traced to a single gene [3]. Moreover, risks inherited from identified disease genes are often modified by background gene variation [11,22,33,34]. Genetic risk in these cases is determined by an individual's allelic profile across many genes, combined with multiple environmental factors. Many of these allelic variations consistently appear together, as haplotypes, which are being mapped in the human HapMap project [1]. Multiple genes involved in cancer are being cataloged in The Cancer Genome Atlas [17]. This data must be combined with new modeling techniques to develop models for understanding the multigenic, molecular basis of human diseases. Developing new treatments for diseases with genetic susceptibilities will require not only the ability to genotype and classify patients on the basis of molecular fingerprints in tissues, but also an understanding of how genetic variants interact to affect clinical outcomes.

Introduction

Genetic interaction refers to the phenotypic effect of combining two or more genes with allelic variations. Genetic interactions are ideally observed under constant environmental conditions and in isogenic strains of model organisms, so that the lone variables are the genetic perturbations (i. e. the allele forms) carried by the strains being compared. Since genetic variants often combine in complex ways to affect phenotypes, it is necessary to infer and model the functional interactions between trait genes rather than viewing each gene as an independent factor. These functional interactions include all modes of activation or repression of one gene's activity by another. Take as an example, a measurable phenotype and a wild-type strain that defines the genetic background. Next consider two genes, A and B, which can be varied from that of the wild-type strain by mutations such that the (isogenic) genetic background is otherwise unchanged. Taking cell growth in the presence of the drug caffeine as an example phenotype, imagine that a mutation in A inhibits growth in the presence of caffeine. The presence of an additional mutation (B) in the strain (i. e. an AB double mutant) could have one of many different effects. B could of suppress the phenotype of A (restoring wild-type growth on caffeine), enhance the phenotype (decreasing growth on caffeine further than that observed in the A single mutant), or show no effect at all (growth equal to that of the A

mutant). Interactions such as these have historically been used to map functional pathways [36].

In addition to the direction of the effect, the magnitude of phenotypic suppression or enhancement also provides important information about the nature of the genetic interaction. Additive effects, where the magnitude of an AB double mutant is equal to the combined effect of each individual's phenotype, are consistent with independent (non-interacting) genes in the same genetic pathway (e. g. where A and B both confer a phenotype) or in different genetic pathways (e. g. where A displays the phenotype, but B does not). Synergistic effects describe an AB double mutant trait that is quantitatively more extreme than would be expected from the linear combination of both. Perhaps the best studied class of synergistic interaction is synthetic lethality [21], in which the combination of two non-lethal mutations, A and B, combine to confer lethality in the AB double mutant. The concept has been appreciated for many years in both fruit fly [75] and yeast [30], and synthetic lethal interactions have identified redundant functions in a common (essential) biological pathways and parallel functions in separate pathways that are able to buffer each other [32]. Such buffering against genetic defects has long been believed to confer a fitness advantage [84]. The buffering of disease-related genes is of particular interest in identifying alleles that either alleviate or aggravate effects within individuals across a genetically diverse population.

In addition to assigning genes to functional pathways and uncovering functional relationships between pathways, the direction and magnitude of genetic interactions can also be used to determine the fine-structure of a genetic network and information flow through the system. Analysis of synthetic and epistatic interactions (Fig. 1) can be used to determine functional topology and the direction of information flow, commonly referred to by geneticists as "ordering genes in a pathway" [4]. Taking again the example of a mutation in gene A that inhibits growth in the presence of caffeine, now consider a mutation in another gene, C, that enhances growth under that condition. If the phenotype of an AC double mutant is a caffeine growth defect equal to that of the A single mutant, A is said to be epistatic to C. Such an interaction is often interpreted as evidence that A acts downstream of C, in that its mutation masks any modifications that occur further upstream of the phenotypic output [4]. This hypothesis is an example of serial information flow through the system, with the perturbed genes influencing each other and the growth phenotype in a linear sequence.

While genetic interaction analysis cannot, when taken alone, decipher the biochemical activity of the gene products, it can establish functional relationships between gene pairs. One can thus view a genetic interaction as providing a comprehensive view of the coordinated activity of two genes, encompassing every step in cellular processing between genotype and phenotype. A broad picture of complex genetic control emerges when these interactions are assembled into a network.

Measuring Genetic Interactions Genome-Wide

Since assaying a genetic interaction requires measuring the phenotype of interest in multiple strains, inferring genetic interactions on a genome-wide scale has historically been intractable. The application of high-throughput assays in model organisms with sequenced genomes has made the systematic study of complex genetics possible. To date, these studies have primarily involved the yeast *Saccharomyces cerevisiae*, which is easy to genetically modify and has a large community of researchers committed to the public dissemination of genome-wide reagent sets, and the nematode worm *Caenorhabditis elegans*, which has proven amenable to perturbations by RNAi. The large-scale study of how a genotype contributes to the control of a phenotype has been greatly aided by advances in methods to both systematically perturb genes and measure phenotypes on a genomic scale. These high-throughput methods produce large, quantitative data sets, requiring parallel development of computational and numerical modeling methods to interpret the output in terms of biological function. The primary objective of these methods is to identify functionally relevant genes and describe how they influence one another to generate cellular activity. Using a systems biology approach to integrate this knowledge with other data types has the potential to generate models capable of validating biological insights and identifying high-priority candidate molecules for targeted therapeutic intervention.

Synthetic Sickness and Lethality

Functional profiling of individual genes is most directly carried out by deleting the gene and observing phenotypic consequences. Through advances in genomics [87], a library of yeast mutants [27] now exists with deletions for every gene in the background of the laboratory strain derived from S288c [86], which was sequenced in the yeast genome project [29]. The yeast genome contains ~6000 protein coding genes (also know as open reading frames or ORFs). Approximately 5000 of these genes are nonessential and were deleted by replacing the gene of interest with a "marker gene" that confers drug resistance and two molecular barcodes, short DNA sequences that uniquely identify each deletion mutant. The remain-

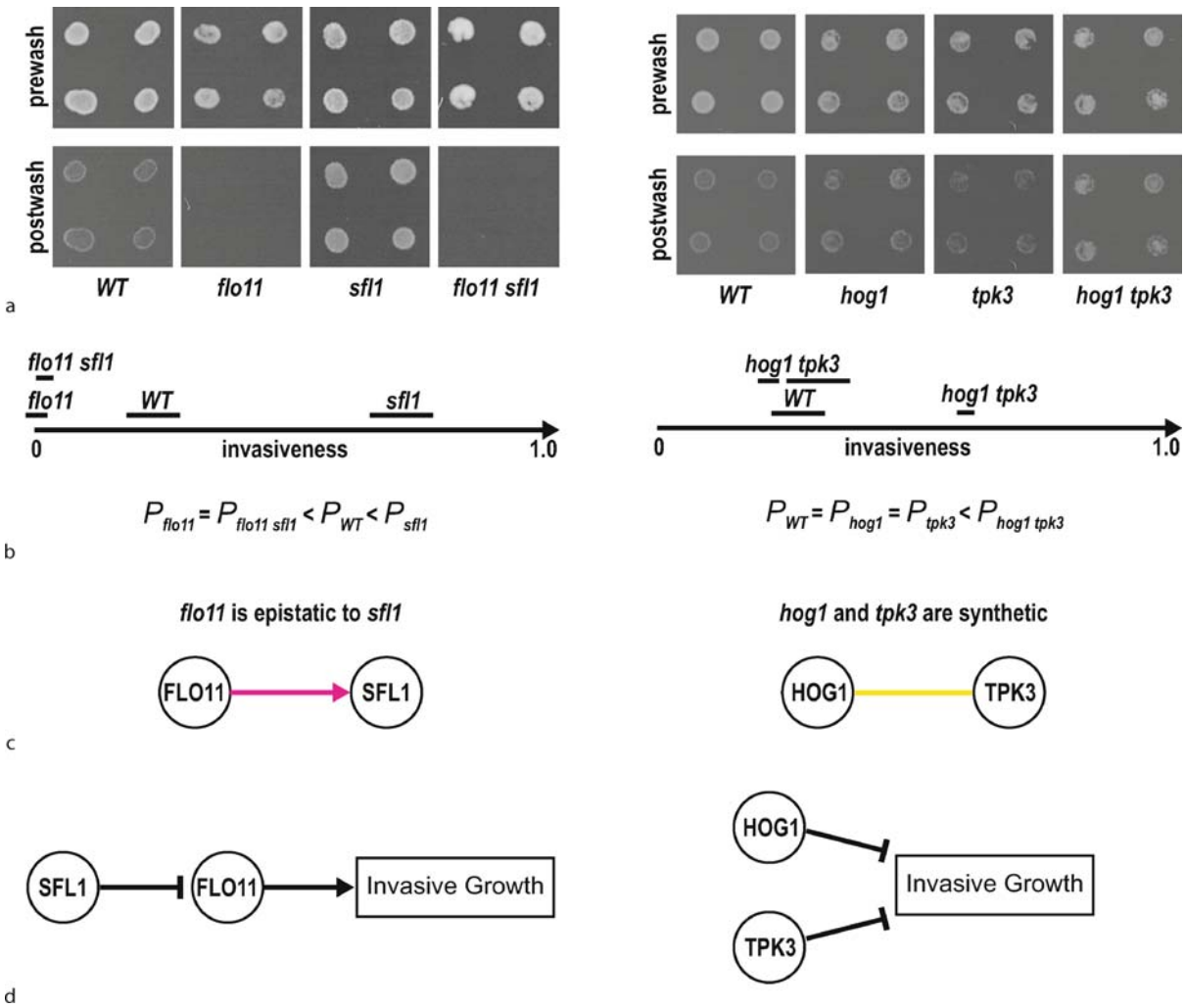

Systems Genetics and Complex Traits, Figure 1

Two examples of genetic interactions for yeast invasion. *Left* panels show the interaction between deletions of *SFL1* and *FLO11*, and the *right* panels show the interaction between deletions of *HOG1* and *TPK3*. a Yeast invasion assays for relevant strains. Yeast colonies (*light spots*) were grown for two days on nitrogen-limiting media (prewash) and cells were scraped from the surface (postwash). b Quantification of invasive growth, as the ratio of photo-assay signal before and after wash. Experimental ranges were determined by the variation among replicate colonies. c Genetic interactions as classified by Drees et al. [23]. d Classical genetics interpretations of pathway ordering. Epistasis (*left*) implies the genes are sequential in a pathway, and synthesis (*right*) implies the genes lie in parallel pathways. (Adapted from [23])

ing ~ 1000 genes are essential for cell viability under standard laboratory conditions and thus must be maintained in strains that contain another functional copy of the gene. The lists of essential and non-essential genes are continually updated [44] as ORF annotation and verification improves [46] and closely-related strains of *S. cerevisiae* that have slight variations in essential genes are discovered [63]. Nonetheless, the observation that only 15–20% of yeast genes are essential, at least under optimal growth conditions, suggests that the genetic network governing yeast growth buffers the cell against genetic variation [32] and implies that multiple genetic mutations may compromise cell growth to a much greater degree than that observed for any single mutation.

Using the yeast deletion library as a starting point, three methods – synthetic genetic array (SGA), diploid-based synthetic lethality analysis with microarrays (dSLAM), and epistatic miniarray profiles (E-MAP) – have been developed as comprehensive, high-throughput assays for detecting pair-wise genetic interactions that

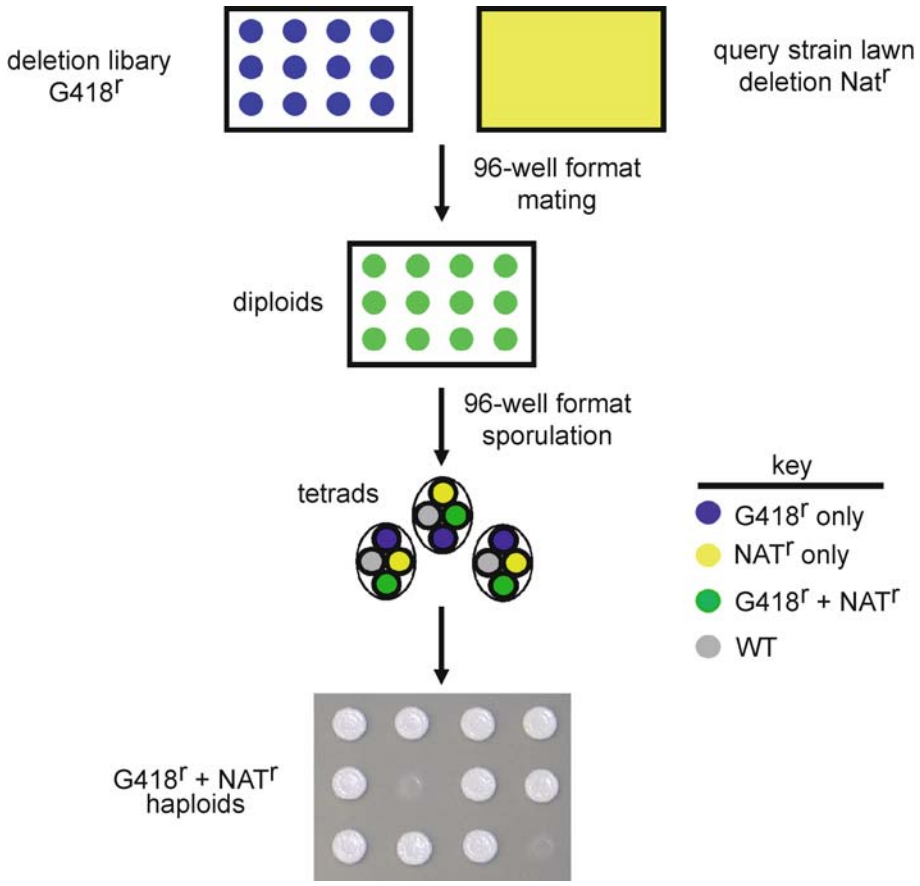

Systems Genetics and Complex Traits, Figure 2

Overview of the synthetic genetic array (SGA) and diploid-based synthetic lethality with microarrays (dSLAM) methods. Many of the experimental details have been simplified, more detailed descriptions are presented in [6]. In the SGA method a haploid strain (*yellow*) is constructed that contains a deletion of the "query gene" marked by a selectable drug marker (Nat^r) and a specialized reporter gene that facilitates high-throughput selection of haploid progeny that have mated and undergone sporulation 79. The query strain is mated to an ordered grid, haploid yeast deletion library (*blue*) marked by the selectable drug marker G418^r, forming diploids (*green*) that contain both deletion mutants, but which also contain a functional copy of each gene. Those diploids are then sporulated, allowing random shuffling of the genes and produces haploid progeny, which will only contain one copy of each gene of interest, i. e. the deleted or functional copy of the query gene. In the final step (photograph of actual yeast grids), haploid progeny that contain both the query gene (Nat^r) and the deletion strain present in that position of the grid (G418^r) are selected. The ability of the haploid progeny to grow in the presence of both drugs indicates that both deletions are present in the same strain, i. e. no interaction. Lack of growth indicates the inability to recover the double mutant combination, i. e. synthetic lethality

confer synthetic growth defects. Experimental details of these methods have been extensively reviewed [6,48,56]. We briefly summarize them here, as data generated by these methods dominates the genetic interaction data currently available. We then discuss other methods now being applied to comprehensively measure genetic interactions in yeast, worm, and mouse.

The synthetic genetic array (SGA) method (Fig. 2) was developed by Boone and colleagues [79,80] and extended to incorporate essential genes by transcriptional control via an artificial promoter [20]. The availability of a variety of drug-resistance markers enables a simple strategy for efficient construction of double gene deletion strains in yeast. Each strain in the haploid yeast deletion library harbors a gene (Kan) that confers resistance to the drug G418. In the SGA method a second haploid strain is constructed. This strain is a different mating type, and contains a deletion of the gene of interest marked by a gene (Nat) encoding resistance to a second drug (nourseothricin) and a reporter gene that enables selection for haploid progeny strains. This strain is mated to the deletion library, the resulting diploid strains are sporulated, and haploid progeny

resistant to both drugs are selected. Synthetic lethality is indicated by the inability to recover haploid strains with both drug markers, suggesting that this combination of mutant alleles is inviable. Double mutants which show significant growth defects (so called "synthetic sickness") can also be identified by this method. The throughput of the method is increased by robotic manipulation of strains and phenotype scoring by image analysis software. To date, over 400,000 interactions have been tested and tens of thousands have been discovered [61], providing sufficient data for sophisticated network analysis and a valuable source of functional relationships between hundreds of yeast genes [80].

The issue of precise growth quantification of double-mutants has been addressed by a modification of the SGA method. In this approach, called epistatic miniarray profiles (E-MAP), growth rates are quantified and genetic interactions are identified as cases in which a double mutant strain exhibits a growth rate significantly above or below the expectation based on the two single mutants [19,64]. Essential genes were tested in this study by decreasing mRNA abundance via destabilization of the 3' UTR. Precise quantification of growth rates was performed by photo-assays of colony size and, based on the assumption that genetic interactions are rare, identifying the most significant deviations from the expectation that a double-mutant growth rate will be the product of the two single-mutant growth rates. Interactions with expected growth rates were classified as neutral, those with growth rates above expectation were labeled alleviating, and those with growth rates below expectation were labeled aggravating. To date, pair-wise interactions between 424 genes that localized to the endoplasmic reticulum [64] and 743 *Saccharomyces cerevisiae* genes involved in various aspects of chromosome biology (e. g. DNA replication/repair, chromosome segregation and transcriptional regulation) [18] have been tested.

Diploid-based synthetic lethality analysis with microarrays (dSLAM) (Fig. 3) is an alternate, transformation-based approach to assaying synthetic lethal interactions [57,58]. The method relies on molecular barcodes that uniquely mark each deletion strain to identify double-mutant fitness defects by DNA microarray hybridization. Heterozygous diploid knockouts [27] are initially transformed with a reporter gene that enables their conversion to haploid mutants following sporulation. Then, the gene of interest is deleted via high-efficiency integrative transformation, and the pool of yeast strains is sporulated and separated into single and double mutants using the drug resistance markers. Genomic DNA containing identifying bar codes from each pool is isolated. Bar codes are am-

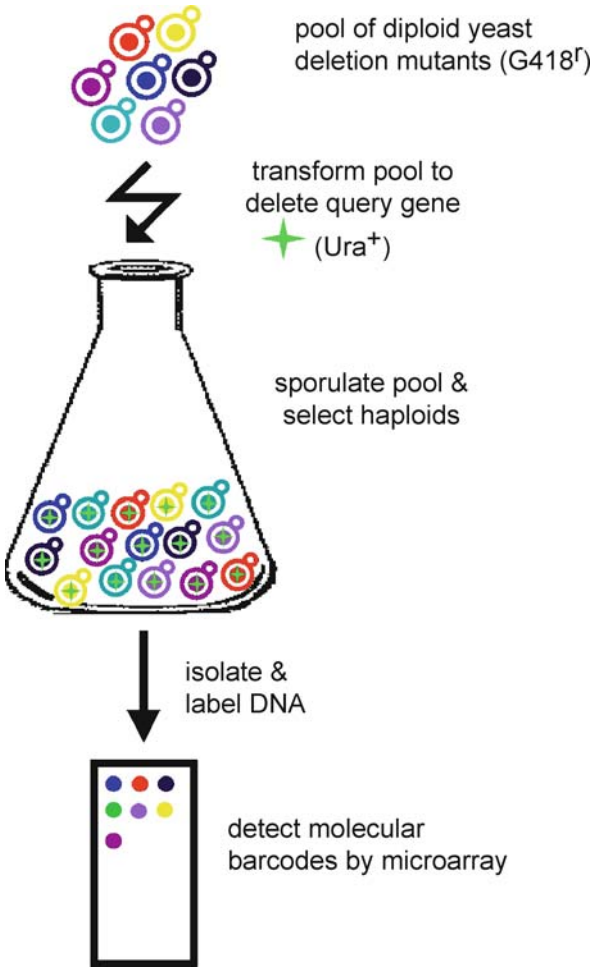

Systems Genetics and Complex Traits, Figure 3
The dSLAM method takes advantage of the molecular barcodes present in each deletion strain to facilitate the manipulation of strains in a pooled format. This method uses a diploid version of the yeast deletion set (heterozygous diploids) in which each strain contains both the gene deletion (G418^r) and a functional copy of the same gene. Prior to the analysis, the same reporter gene that facilitates high-throughput selection of haploid progeny used in SGA is introduced into all of the strains (not shown). The first step is the deletion of the query gene in (ideally) all the strains in the pool by transforming in a construct that will delete the gene of interest and contains a marker that will allow successfully transformed cells to grow in the absence of the nutrient uracil (Ura$^+$). The pool is then sporulated and haploid progeny containing both the G418^r deletion from the collection and Ura$^+$ query deletion are selected. DNA is then isolated from the pool of strains and the sequences containing the molecular barcodes are fluorescently labeled and hybridized to a DNA microarray that can detect the barcode sequences. Thus, microarray intensity provides a quantitative readout of the amount of each mutant recovered in the selection, with extremely low microarray intensity indicating synthetic lethality

plified, labeled, and hybridized to microarrays to quantify the ratio of single to double mutant growth rates. SSL interactions, also referred to as synthetic fitness or lethality (SFL) defects, are identified by high ratios that correspond to reduced double-mutant growth. Initial comparisons provide evidence that the dSLAM method achieves higher data consistency and lower false-negative rates than the SGA approach. dSLAM can also be modified to study dosage dependent SSL, genetic suppression, and haplo-insufficiency. Although the microarray based method potentially allows a more quantitative measure of synthetic sickness, the low signal intensity observed for many barcodes has hindered precise quantification. Because it relies on growth in mutant pools, the dSLAM method may be the most efficient method for genome-wide SSL screens.

The relatively recent introduction of RNA interference (RNAi) technology for gene knock-down allows multiple loss-of-function genetic perturbations to be carried out at high throughput [43,50] in model organisms such as the worm *C. elegans* and the fruit fly *D. melanogaster*. Large-scale RNAi based synthetic lethal screens have been performed in *C. elegans* [49]. Approximately 350 synthetic interactions between 160 genes in the worm EGF/Ras, Notch, and Wnt signaling pathways were discovered among 65,000 tested gene pairs. The highly-connected "hub" genes encode chromatin regulation proteins that are conserved across multiple species, which were hypothesized to be general buffers of genetic variation. Another study used the same experimental techniques to test functional redundancy between duplicated genes in *C. elegans* [78]. Synthetic interactions were detected in eleven percent of the duplicated pairs tested (16/143), suggesting that in some cases duplicated genes have a degree of functional redundancy. Interestingly, the majority of the interacting gene pairs were found to be duplicated over 80 million years ago, strongly suggesting that this functional redundancy can be conserved by positive selection.

Genetic Interactions for Quantitative Phenotypes

The vast majority of genetic interactions detected by the large-scale methods described above have been limited to the phenotype space of growth in rich yeast medium. However, phenotypes can be derived from any observable trait, including growth under a variety of environmental conditions, cell morphology, metabolite production, or gene expression pattern. To date, the most comprehensive analysis of genetic interactions detected using a phenotype other than growth in rich media include an analysis of interactions in nitrogen-poor media [23], or in the presence of a DNA damaging agent [71].

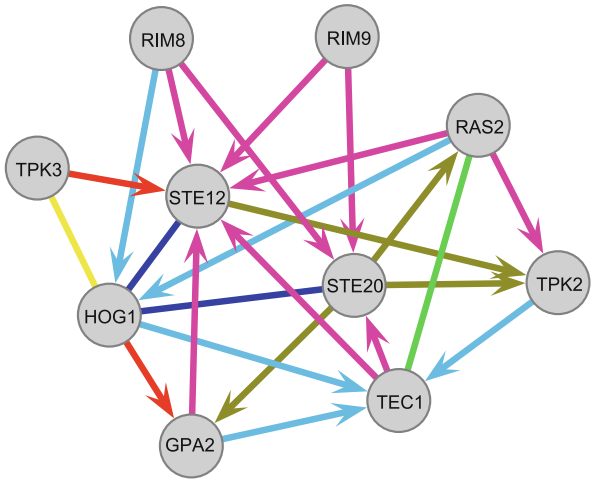

Systems Genetics and Complex Traits, Figure 4
Subsection of the yeast invasion network showing genes from various pathways and the variety of genetic interactions observed. Edge colors correspond to different interaction modes, e. g. *violet* is epistasis, *yellow* is synthesis, and *dark blue* is additivity. Subset of data from [23]

A mixture of gene deletions and high-copy, gain-of-function gene mutations were combined to map interactions in the yeast invasion network for nitrogen-poor media (Fig. 4) [23]. When starved for nitrogen, many strains of *S. cerevisiae* undergo cell differentiation into a pathogen-like, filamentous growth form and colonies subsequently invade solid media [28]. Drees and co-workers measured invasive growth by photographing colonies before and after vigorously washing cells from the agar surface (Fig. 1a). Multiple replicates of each genotype were assessed using image analysis software to quantitatively determine a mean invasion score and its variance (Fig. 1b). Since the phenotypes of the original single mutants ranged from non-invasive to strongly hyper-invasive, a wide variety of genetic interactions were possible. These interactions were first translated into mathematical inequalities that defined the relationship between the wild-type strain, the two single mutants, and the double mutant in terms of "less than", "greater than", and "equal to" relationships. For example, a synthetic interaction in which a double mutant (AB) exhibits a phenotype not shown by either single mutant (A or B) would be either "AB < A = B = WT" or "A = B = WT < AB". After systematically enumerating all 45 possible inequalities, the equations were grouped into nine "modes" of genetic interaction. Many of these interaction modes, such as synthesis, epistasis, or suppression, are well known to geneticists, but less familiar modes also appear. For example, the mode labeled "asynthesis"

was defined as a double mutant phenotype which matched that of both single mutants but was distinct from the wild type. This general analysis was applied to over 130 genes and over 1800 pair-wise interactions were tested, revealing a rich spectrum of functional interactions that included examples of all nine modes.

A similarly multi-modal genetic interaction network was more recently constructed for genes that impart resistance to the DNA-damaging agent methyl methanesulfonate (MMS) [71]. Twenty-six such gene deletions (out of over 4700 tested) were identified by reduced growth rate in the presence of MMS, and double-mutant strains were constructed and assayed for all 325 pair-wise combinations. Although the selection method precluded the detection of synthetic interactions, this study is a quantitative analysis of genetic interactions among a set of genes implicated in a specific biological process.

Analysis of Complex Phenotypes that Result from Natural Variation

Although the majority of systems biology approaches to complex genetics have involved engineered genetic perturbations, genetic interaction studies are being extended to models of natural populations. A series of papers by Kruglyak and colleagues have attempted to infer genetic interactions in recombinant progeny strains [7,8,74]. In this work, two parent yeast strains, the sequenced "laboratory strain" (BY) and a "wild strain" isolated from a vineyard (RM) were mated and 112 haploid progeny strains were isolated. Each strain was genotyped at approximately 3000 markers and RNA expression levels were measured using DNA microarrays [9]. Statistical linkages were found between expression variations and quantitative trait loci (QTLs), and these loci were then assessed for interactions [74]. Although the data suggest that most transcripts were influenced by multiple loci [7], statistically significant pair-wise interactions have only been found for 225 transcripts [8]. Another series of yeast studies, taking high-temperature growth (HTG) as a quantitative phenotype, used reciprocal-hemizygosity as a tool to dissect QTLs in yeast [72]. The contribution of each allele to HTG was inferred by constructing hybrids of two yeast strains and then systematically deleting one copy of each gene within a candidate QTL. The QTL architecture was found to be more complex than expected, featuring both *in cis* and *in trans* linkages and suggesting that QTL regions cannot always be narrowed to a single local gene. The method was also used to explore how the influences of three trait genes were affected by variation in genetic backgrounds [69]. Reciprocal-hemizygosity analysis with ten other yeast strains demonstrated that the trait genes identified in one background did not necessarily contribute in the same way in other genetic backgrounds. These results demonstrate that, in addition to complexity within a QTL region, there is additional complexity due to genetic interactions. Together these studies illustrate some of the substantial complications that will arise in the analysis of complex genetics in the context of natural populations, such as the identification of trait loci in a background of genetic random variance, the separation of phenotypic effects from *cis*-acting and *trans*-acting polymorphisms, and the limitations of the pair-wise interaction analysis to complex genetics.

Furthermore, the logic of genetic interactions could be applied to functional interactions between different types of biomolecules. For example, recent work has studied the genetic basis of drug response in yeast [59]. Multiple polymorphisms were found that linked genetic loci to certain small molecule drugs, constituting a set of 124 gene-drug interactions. This approach provides a model for understanding variation in drug response within a genetically diverse population. Another study directly applied genetic interaction techniques to pair-wise interactions between drugs by mapping a set of compounds into classes that correspond to the cellular functions they affect [93]. These works begin to address how the understanding of complex genetics can be extended to involve additional system elements and, ultimately, generate a view of the organism as a complex system.

Overall, the study of genetic interactions is a field that has seen substantial development in the past decade. High-throughput methods will continue to open new avenues of research and expand established approaches to genome-wide scales. Data generated by these experiments will require formal and systematic approaches for analysis. Currently, network methods are the dominant framework.

Constructing Genetic Interaction Networks

Prior to high-throughput experimentation, genetic studies mapped interactions between a few genes in isolated functional pathways. These pathways represent information flow through a small number of elements that regulate a specific cell behavior, such as response to osmotic stress. Today, high-throughput techniques enable the assembly of large-scale genetic interaction data into networks and hold the potential for constructing semi-global maps of multiple, inter-connected pathways.

This wealth of data requires a formalism to efficiently analyze relationships between system elements. Complex systems are often best visualized as networks in which in-

dividual points, called *nodes*, are connected by lines or arrows, called *edges*. In genetic interaction networks, the nodes commonly represent genes, alleles, or gene perturbations (e. g. a gene deletion). Edges that connect nodes can take on different meanings, ranging from classic regulatory influences of one gene's activation or repression of another's activity to edges that signify a direct observation, such as synthetic lethality. The central challenge is to represent the richness of complex genetics in an informative way. Because the construction of each study is informed by a different biological problem, diverse network approaches have been pursued in the literature.

Since they are binary, synthetic lethal interactions are easily represented as a network of gene deletions connected by SSL edges. Although the formal information content of the edges is low, SSL networks have the advantage of relatively high coverage. Pair-wise combination of about 1000 yeast genes have been tested, and although this represents only 3% of all possible yeast pairs it encompasses 500,000 experiments. Four thousand of these pairs were SSL positive, producing a sparse but topologically interesting network. Non-essential genes averaged 34 synthetic lethal partners, and essential genes were found to have more interaction partners on average [80]. Extrapolating these results genome-wide, the authors estimate 200,000 essential pair-wise combinations in yeast, a number of lethal outcomes that is much larger than the 1000 genes that are essential individually [6]. The current network of 4000 genetic interactions demonstrates that SSL interactions are more likely to occur between non-essential genes of the same or related function, unlike SSL partners of essential genes that tend to show a broad spectrum of functions. This suggests that network location can be a predictor of gene function for non-essential genes, enabling novel functional hypotheses for genes based on a common function of interaction partners.

The known SSL network exhibits two interesting topological properties. First, it appears to be a *small world* network, defined as a network in which the length of the shortest path between two nodes is smaller than random expectation. Nodes are densely grouped into local clusters in which the interaction partners of any gene also tend to interact. Thus, the most likely candidates for SSL interactions for a given gene are the interaction partners of its known SSL partners, greatly increasing the likelihood of finding additional interaction partners in a targeted screen [80]. The second network property observed is that the number of SSL partners for each gene follows a power-law or *scale-free* distribution in both the yeast [80] and worm [49] data. As a result, most genes have relatively few interactions, and a small number of genes have relatively large number of interactions. This latter set, referred to as hubs, can be viewed as more important for fitness since the organism is most sensitive to pair-wise perturbations involving them. If this property is also a feature of human networks, the hub concept may inform the treatment of disease. Taking cancer as an example, hubs of cancer cell networks might be targeted for perturbation to increase the chance of a SSL interaction with a gene already genetically perturbed in the cancerous cell.

Nearly 5000 additional, and mostly novel, synthetic fitness or SSL interactions were discovered by studying genes involved in processes related to DNA integrity in yeast using dSLAM [57]. A total of 16 gene modules were identified in the network by clustering genes with a high congruence in SSL interactions, and each module was assigned a function based on its member genes. The resulting network mapped compensatory pathways among multiple biological processes, including DNA replication, DNA repair, checkpoint signaling, chromatin structure maintenance, and the response to oxidative stress. The modular architecture was consistent with prior models, and involved many genes with human homologs linked to cancer and aging.

The study of SSL interaction networks has the practical advantages of a simple, rapid assay and the direct mapping of each positive result to an edge in the interaction network. This narrow definition of genetic interaction, however, cannot explore other types of genetic interaction that have proved informative, such as suppression and masking epistasis. Encompassing all types of genetic interaction requires a systematic formalism such as the inequality relationships developed by Drees et al. [23]. As described above, these inequalities can be grouped into rules of genetic interaction corresponding to familiar and newly defined interaction modes. While many of these modes, such as synthesis, were symmetric with respect to the interacting genes, four of the modes produce directional edges. For example, instances of suppression were rendered with a colored arrow from the suppressor gene to its target gene. The resulting 1800 genetic interactions between 133 genes compose a genetic interaction network, derived with the *PhenotypeGenetics* plug-in for the *Cytoscape* software platform [67]. The nodes of this network included both gene deletions and plasmid-borne over-expressers, and the edges were mapped using nine colors corresponding to the nine modes of genetic interaction based on classical genetic interaction rules. This produced a richly multichromatic network suitable for informatic analysis (Fig. 4).

The multiplicity of genetic interaction modes facilitates the inference of relationships between individual

genes and specific biological functions. In the Drees study, individual alleles were often found to interact in a particular mode, or *monochromatically*, with partner genes of coherent biological function, leading to hypotheses for regulatory and pathway organization [23]. Large-scale patterns of mutual information were also extracted from the data set, and groups of genes with significant mutual information between them formed network cliques corresponding to physical pathways. The genetic interaction patterns define a map of information flow from specific genetic perturbations to quantifiable phenotype effects.

This network was further analyzed by Taylor and Galitski [76], who searched for repeated subnetworks or *motifs* of three or four nodes, in which repetition suggested functional organization. Statistically enriched motifs were then assessed for functional coherence and many instances of functional monochromaticity were observed involving multiple genes. For example, deletions of multiple protein kinase genes were found to suppress gain-of-function perturbations of both signaling and transcription factor genes. This suggests that the kinases mediate information flow from the signaling proteins to transcriptional factors, and these pathways are broken by kinase deletions.

Multi-modal genetic interaction analysis was also performed on the MMS-sensitivity network of St. Onge et al. [71]. Genetic interactions were initially classified as aggravating, alleviating, or neutral interactions based on the double-mutant growth rates relative to that of the two single mutants. Alleviating interactions were of particular interest since there are many ways genes can mask or mimic another's effect. Thus, these interactions were sub-classified into five rules of interaction based on models of pathway ordering and protein co-function, four of which were assigned directionality. The resulting network model uncovered functional relationships and pathway order. The principle of genetic congruence, that genes with similar functional roles will share interaction patterns with functionally related genes in the system, was used to implicate a new regulator in DNA repair.

Both the yeast invasion and MMS networks classified genetic interactions into multiple modes based on interaction inequalities as described above. However, the authors used slightly different classification schemes. For example, in the invasion network "epistasis" included any interaction in which one single-mutant phenotype masks the phenotype of another. In the MMS network a further distinction was drawn between "masking epistasis", in which the single-mutant with a stronger effect masks the other, and "suppression", in which a weakly defective single-mutant masks the strongly defective one. While both networks proved informative, it is unclear which choice of mode classifications are the most meaningful for a given data set. This issue can be analyzed from the perspective of information theory. A measure of network complexity that places information in the context of the system under study could be used for the unsupervised classification of interaction modes for these two data sets. Network classification schemes that produce greater complexity might also generate more biologically meaningful information, i. e. more "biological statements" linking genes to functions were obtained in the more complex networks. Such complexity-based methods represent a powerful and general approach to genetic interaction analysis, with potential for the study of mammalian systems in which interactions are complex and gene annotation data is sparse.

Fully quantitative studies of genetic interactions have been hindered by the limited amount of numeric phenotype data required for more sophisticated mathematical techniques that might decipher the principles underlying genetic interaction. To overcome this limitation many researchers have adopted gene expression microarrays as a quantitative phenotype [9,41,42,47,62,85]. For each strain and environmental condition, genome-wide transcript levels constitute thousands of quantitative phenotype measurements that provide much greater molecular detail than a whole-cell phenotype, such as growth rate. This approach was combined with classical epistasis analysis by Van Driessche et al. [82]. By clustering mutants based on genome-wide expression profiles and applying the principle of masking epistasis, the authors accurately reconstructed the protein kinase A signaling pathway. In a more recent study, gene expression data was combined with whole-cell phenotype data and mathematical modeling to infer and predict genetic interactions relevant for yeast cell differentiation [12]. This analysis used combinatorial perturbations to infer how a set of regulatory genes quantitatively affect each other's activity as well as the expression of thousands of downstream genes. This quantitative model of genetic interaction in terms of genetic influences allowed the prediction of the effects of novel pair-wise genetic perturbations. The decreasing costs of microarrays and advances in genome-wide data analysis make gene expression data an increasingly attractive option for genetic interaction analysis.

Interpreting Genetic Interactions

The functional relationships described by genetic interaction networks reveal pathway organization and gene function. For predictive and therapeutic purposes, however, it is useful to translate these functional relationships into biomolecular hypotheses. Such hypotheses can sug-

gest additional genetic perturbations to test and validate models, establishing the applicability of network analysis techniques and enabling an iterative process of model refinement. Moreover, the implication of specific alleles, molecules, and interactions in health-related processes can identify candidate genes and gene products for targeted therapeutic intervention with drug compounds or RNA interference. Thus, the interpretation of genetic interaction networks is necessary to improve the basic understanding of complex genetics as well as facilitate the development of personalized and predictive medicine.

Formulating such hypotheses generally involves both molecular data integration and computational modeling. While genetic interactions define functional relationships and uncover pathways of information flow, molecular interaction databases provide a wiring diagram of possible mechanisms for the transmission of functional information. These complementary data types can be integrated to generate models of phenotype regulation [13,31,83,88]. Various computational methods enable this data integration. Kinetic modeling, for example, uses differential equations to predict the functional outcomes when a subset of physical interactions is active. Constraint based approaches use data analysis techniques to narrow the possible mechanisms that may have generated an observation, until the best hypothesis is identified [5,65,89,91]. Finally, whole-cell phenotypes can be dissected at the molecular level by integrating RNA expression levels and proteomic data [12,35,40]. These systems biology approaches have recently been applied to interpret complex genetic networks.

One particularly model-driven approach involves the use of computational models that simulate cellular processes to predict genetic interactions *in silico*. This requires a model, often highly detailed, that captures the relevant activity of every gene product involved in the process being simulated. Metabolic models based on Flux-Balance Analysis (FBA) have been successful in predicting growth rates for yeast [26] and bacteria [25] under a range of environmental conditions and genetic perturbations. Segré and collaborators used an FBA model to explore genetic interactions among metabolic genes in yeast by performing single and double gene knockouts *in silico* [66]. The model predicted that genetic interactions for growth rates fall into three distinct modes: cases in which the double deletion effect is significantly less than the product of the two single mutants (alleviating); cases in which the double deletion effect is significantly greater than the product of the two single mutants (aggravating); and cases in which the double deletion is not significantly different than the product of the two single mutants (neutral). The study also predicted similarity in interaction modes, or monochromaticity, between co-functional genes, a result that echoes the finding that genes with common functions preferentially interact in SSL networks. This work illustrates the power of using a system-wide computational simulation to predict complex genetic interactions using sensitive, quantitative whole-cell phenotypes. One of the major disadvantages of such an approach is the extensive amount of information necessary to formulate an FBA model. Furthermore, these predictions have not been empirically verified, and this would involve a major experimental undertaking.

As an alternative to model-based approaches, data-driven computational analysis has also been used to predict interacting gene pairs. Multiple types of relationships between gene products, such as common localization, function, and patterns of gene expression, were used to create a decision tree for predicting SSL interactions in yeast [88]. Genetic interactions were also predicted on a genome-wide scale in the nematode worm *Caenorhabditis elegans* by computationally integrating physical interactions, genetic interaction, gene expression, phenotype, and functional annotation data from yeast, worm, and fly [95]. The authors of this study used logistic regression to determine the predictive power of each type of data, or *feature*, based on a training set of 1816 known genetic interactions in *C. elegans*. Each feature was thereby weighted according to its predictive importance and the probability of interaction between every pair of *C. elegans* genes was calculated. In total, this procedure generated over 18,000 likely genetic interactions. As with previous studies, co-functional genes proved more likely to exhibit genetic interactions (although this is not an entirely independent result given that known genetic interactions between co-functional homologs were the predictors with greatest weight in the logistic regression). A handful of interaction candidates were experimentally tested using RNAi, and about half of the pair-wise perturbations exhibited some degree of genetic interaction in either a vulval development or pharyngeal pumping phenotypes. Interestingly, the vast majority of newly implicated genes did not have vulval development phenotypes when perturbed individually, supporting the hypothesis that the observed double-perturbation phenotypes were the result of genetic interaction. The trade-off inherent in this approach to predicting genetic interactions is that although the analysis can be performed on a genome-wide scale, the nature of each interaction (e. g. additivity, suppression) cannot be predicted. Nevertheless, the approach enables the computational prioritization of potential interaction pairs and greatly reduces the space of likely genetic interactions.

Targeted strategies of data integration have also been used to provide biological context to observed genetic in-

teractions. Kelley and Ideker [45] demonstrated a use for protein-protein and protein-DNA binding networks in understanding the synthetic lethal interaction network in yeast. The study evaluated three standard pathway models for SSL interaction [30,81]: the *between-pathway model*, in which SSL interaction partners are situated in parallel pathways with redundant functions; the *within-pathway model*, in which SSL gene pairs occur for protein subunits of a single complex or pathway that does not lose function until multiple components are removed; and *indirect effects*, in which many pathways are affected and the SSL deletions cannot be associated with specific molecular mechanisms. Based on maximum likelihood, the authors estimated that 40% of yeast SSL interactions could be explained by either the within-pathway or between-pathway models. These models suggest specific patterns of genetic interaction. For the between-pathway model, near-bipartite SSL subnetworks suggest that the missing bipartite links represent undetected interactions. Alternatively, the within-pathway model suggests that genes with many common interaction partners should themselves interact. These predictions were cross-validated by removing one-fifth of the interactions from the data set and then tested to see how well the withheld interactions could be predicted. Both models showed enhanced prediction accuracy relative to randomly predicted interactions, although the between-pathway model was a significantly better predictor, a finding consistent with those suggested by a decision tree approach [88] and an analysis of dSLAM modularity [57,90]. In the worm, integration of protein data also reinforces the compensatory pathway interpretation for SSL interactions [49]. A concurrent study by Zhang et al. [94] integrated molecular and genetic interaction data using the concept of network motifs [53] to guide biological interpretation [92]. In this work, a global network containing protein-protein binding, protein-DNA binding, gene co-expression, sequence homology, and SSL relationships was systematically searched for all possible three-node motifs. A motif in which two proteins linked by protein-protein interaction both have SSL interactions with a third protein was detected in the network ten times more often than random expectation. This was interpreted as a sign of higher-order organization: two proteins that bind in the same complex or pathway share genetic interactions with other proteins in a separate complex or pathway. This interpretation was supported by a search for a four-protein motif in which two parallel pairs of co-binding proteins are linked by all four possible SSL interactions. This motif was found over two hundred times more than random expectation, further supporting the interpretation that SSL interactions are often hallmarks of parallel or compensatory pathways. The goals of integrative analysis are the systematic interpretation of SSL interactions in terms of molecular interactions and the association of specific pathways with observed genetic interactions.

For networks that include multiple modes of genetic interactions, such as suppression, additivity, and genetic synthesis, finding biologically meaningful, non-random network topologies in an integrated network is hindered by the enormous number of possible motifs and the large datasets required for adequate statistics. One strategy for interpreting these rich genetic interaction networks is a quantitative generalization of the classical genetic interaction approach of observing how genetic perturbations interact to affect phenotypes [4]. This method has historically been used to infer how genes influence one another and thereby affect a downstream phenotype, revealing functional relationships such as activation, repression, and pathway order. Using filamentous growth in yeast as a model system (a precursor to invasive growth due to nitrogen starvation), a data decomposition technique was developed to separate genetically "direct" (not necessarily molecularly direct) effects of regulator genes from the genetically "indirect" effects that involve genetic interactions between regulator genes (Fig. 5) [12]. Molecular interaction data were then integrated with the decomposition results to construct regulatory network models of information flow through paths of genes (and gene products) linked by physical interactions. These networks represented quantitative hypotheses of influence from the causal perturbation, through a series of molecular interactions, to the expression of affected genes. Since additional genes were implicated in specific influence-mediator roles in the network, quantitative predictions could be made for gene expression effects arising from their deletion and combinations of deletions. This was done by removing the quantitative influence corresponding to the molecular pathways involving the deleted gene(s). A second round of gene expression data was collected to test a set of predictions and the results demonstrated genome-wide success. Next, by using additional decomposition methods to link gene expression patterns with the filamentous growth phenotype, a network model of whole-cell phenotype regulation was constructed. Predictions for additional perturbations were formulated and experimentally tested. The model successfully implicated new regulators of the phenotype and was used to accurately predict phenotypes for novel combinatorial deletions. While this initial work used laboratory-engineered mutant strains, the techniques can be readily extended to dissect and map complex genetic systems in which a handful of trait genes have been identified. With the increasing availability of human interaction

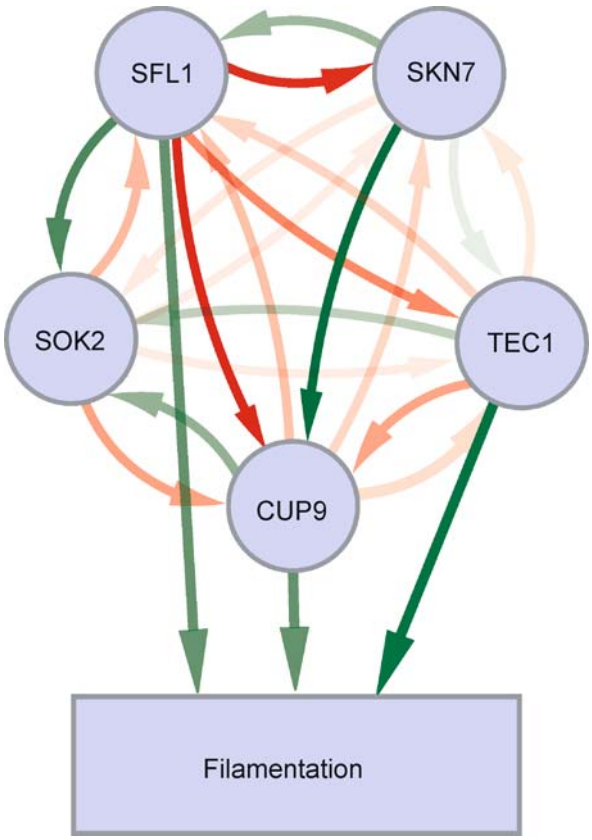

Systems Genetics and Complex Traits, Figure 5
Influence network of five transcription factors that are trait genes for the filamentous growth phenotype in yeast. *Green* and *red* edges map positive and negative influences, respectively, between genes and from genes to the filamentation phenotype. Edge intensity denotes influence strength. (Adapted from [12])

data [73] and further modeling developments to address allelic variation in outbred populations, similar quantitative and integrative techniques may ultimately be applied to disease-related models.

Future Directions

Although genetic complexity is central to all life, systems biology approaches to complex genetics have to this point focused on simpler model organisms. The genetic manipulability, rapid reproduction, and genome-wide interaction databases for yeast, worm, and model prokaryotes have enabled efficient development of experimental and analytical methods. However, the possibility of understanding human disease through the study of complex genetics has catalyzed substantial advances in genetic association studies in mammalian models and human populations [54,55]. The current status of systems approaches to human heath are discussed elsewhere in this encyclopedia (see ► Systems Biology, Introduction to and ► Systems Biology of Human Immunity and Disease), and here we will limit discussion to research opportunities in model systems that are candidates for the systems approaches discussed in this chapter. Future developments can be roughly split into developments in two categories: laboratory resources and computational methods.

Recent programs in rodent breeding have engineered controlled yet genetically diverse populations. These experiments conceptually resemble the analysis of recombinant meiotic progeny generated by crosses in yeast (Sect. "Measuring Genetic Interactions Genome-Wide"). In mouse, two inbred, homozygous parent strains are crossed and two of the resulting F_1 offspring are mated (Fig. 6). The resulting F_2 progeny are then segregated into breeding pairs, and each pair begins a line of repeated sibling inbreeding that continues until a homozygous inbred strain is obtained. Unlike yeast, in which homozygous diploids can be obtained in a matter of weeks, producing recombinant inbred (RI) lines in mice typically requires 20 generations, or 7–8 years, to produce the desired RI lines. Nevertheless, RI genetics have been integrated with gene expression and phenotype assays in mice [10,15,51] and rats [37]. While each of these studies was able to associate individual loci with gene expression patterns, implicating new genes in metabolic, nervous system, and stem cell function, the relatively small numbers of RI strains (approximately 30 in each study) was insufficient for systematically inferring polygenic effects. Inferring such effects would require additional model parameters to encompass the quantitative interaction between each gene pair, leading to overfitting for small data sets. Nevertheless, these works provide prototypes for further studies using RI strains.

While constructing RI strains is expensive and laborious, they have significant advantages over simpler intercross strains in the study of complex genetics. Multiple mice from each RI line will allow biological replication and the observation of isogenic strains under a variety of environmental conditions. Strains can be assayed over long time scales and multiple generations. A collection of shared RI strains would be especially valuable as a community asset both in terms of labor saved, since the strains need only be constructed and genotyped once, and data sharing, since the results from multiple researchers can be integrated in novel analysis. With this in mind, the Complex Trait Consortium is pursuing the Collaborative Cross project [16], which aims to create approximately 1000 RI strains derived from eight inbred founder strains. This set of RI strains will contain a degree of genetic diversity suf-

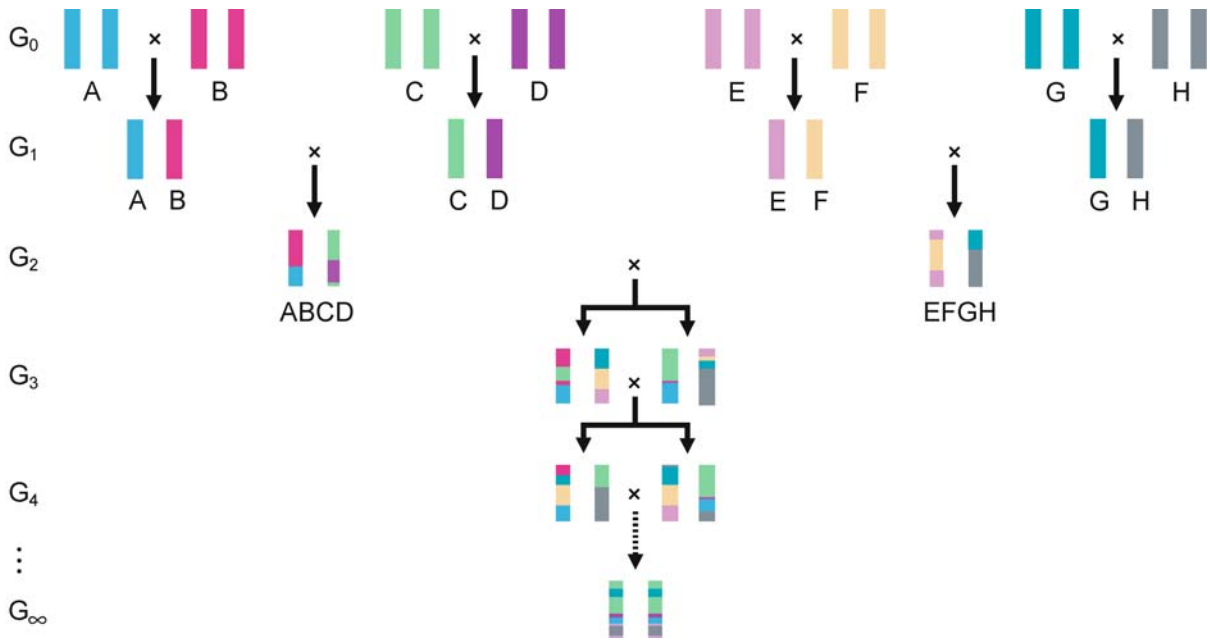

Systems Genetics and Complex Traits, Figure 6

The collaborative cross strategy for breeding recombinant inbred (RI) strains. The starting parents of the cross (Generation 0, G_0) consists of eight inbred laboratory mouse strains. The goal of the breeding strategy is to create 1000 new RI mouse lines that contain different combinations of alleles from the original parents. A theoretical example of the generation of a single chromosome for a single RI strain is shown. In the first generation of the cross (G_1) progeny receive an unaltered copy of a chromosome from each parent. In all subsequent generations chromosomal material is exchanged between chromosomes generating chromosomes that contain genomic regions from a variety of parents. The lines are ultimately converted to recombinant inbred (RI) strains, in which an individual contains identical copies of these hybrid chromosomes ($G_{infinity}$) by multiple generations of inbreeding, in this case mating brother-sister pairs. (Reprinted with permission from The Complex Trait Consortium [16])

ficient to uncover gene-gene, gene-drug, and gene-environment interactions. Furthermore, the resulting 1000 inbred mouse lines could be crossed to produce hundreds of thousands of intercross (RIX) strains. Inbred homozygosity ensures that these genotypes will be reproducible and the genetic differences between RI strains will produce low coefficients of inbreeding. Because the RIX strains are not themselves inbred, they will more closely resemble the human population in terms of heterozygosity, genetic mixing, and hybrid vigor. When combined with the proliferation of gene annotation and molecular interaction data, resources like the mice derived from the Collaborative Cross will allow systems approaches to mammalian genetics and provide a platform for adaption to human disease.

In addition to improvements in laboratory resources, a number of analytical advances are required to better understand large-scale genetic interaction networks. These include further developments in data integration, quantitative prediction, data analysis, ontologies, and frameworks for classifying complex genetics.

Strategies of data integration have taken diverse forms. These range from all-inclusive, nearly universal strategies of data pooling [88,95], to integrating only a handful of physical interactions implicated by phenotype modeling [13,45], to modeling constrained by incomplete data resources [52]. The accuracy and reliability of data varies greatly both across data types (e. g. protein-protein versus SSL interactions) and within data types (e. g. protein-protein data from yeast two-hybrid versus co-immunoprecipitation assays). Some methods have used machine learning techniques to weigh data types based on predictive utility [38,94]. Through the repeated application of such methods to multiple data sets and biological systems a consistent picture of data reliability might emerge, providing valuable information for modeling and interpretation of complex genetic systems. Finally, in many cases the integration of published data appears to be performed at a late stage in the work as researchers seek more complete explanations for their results. While this is to some extent an inevitable result of large-scale screens or discovery-driven science, it would be advantageous for investigators to con-

sider from the outset the types of data that may be most complementary and form a coherent plan of data integration [24]. The use of successful mathematical models and biological simulations at such an early stage of research would aid in this integrative research design.

Genetic interaction analysis is often showcased for its predictive modeling potential. Since genetic interactions inherently reveal functional information, under-characterized genes can be assigned to biological pathways [23,71] or functions [80], and these implicit hypotheses about gene function can be experimentally tested. However, physical interaction networks organized into pathways of information flow can also explicitly predict the consequences of additional perturbation [12,66]. The predictions are often quantitative and represent a potentially powerful tool for predictive medicine when applied to the development of therapeutic interventions. Expanded use of computational modeling has the potential to generate more accurate and sophisticated predictions of how systems respond to genetic or environmental perturbations. Established modeling strategies include kinetic modeling, which has been used to understand complex cell circuitry [14,60] and to elucidate subtle stochastic effects [2], and Boolean network approaches, which have been used to model sophisticated systems such as host-parasite interactions [77] and uncover fundamental properties of biological dynamics [68]. Combined with time-series assays of responses to environmental stimuli in genetically diverse organisms, these quantitative modeling techniques can be a powerful tool in understanding the dynamics of polygenic traits.

In the analysis of genetic interaction data, the field of systems genetics now is primarily focused on synthetic lethal interactions, to the extent that many authors use the terms genetic interaction and synthetic lethality synonymously. However, synthetic lethal interactions may be less relevant in the complex genetics of natural populations than in laboratory-based screens. Their broad application may also be limited by reliance on technology specific to model organisms and networks composed of binary interactions. For instance, although synthetic screens can reveal compensatory modules and classify genes into functional pathways, the scale-free architecture of these networks often features many pathways connected to a common hub [49]. Since newly discovered synthetic interactions are most likely to include a hub and a non-hub gene, many pathways will be connected to the hub, and thus be equally likely candidates. It may therefore be difficult to place the latter gene in the correct pathway without additional information. Alternatively, considering a broader spectrum of genetic interactions can both resolve pathway membership, via mutually informative partners, and provide information about pathway structure, e. g. the placement of a factor in a signaling or transcriptional cascade. A network encoding a more general set of genetic interaction rules can thus enable the modeling of specific mechanisms in detail rather than congruence methods based on guilt-by-association. Examples of such rules are directional, targeted interactions like epistasis and suppression (Fig. 4) [23,71] or quantifiable influences describing one gene's influence on another (Fig. 5) [12].

To this point, the classification and interpretation of genetic interactions has been inconsistent due to the paucity of quantitative phenotypes that can be widely screened paired with a limited knowledge of the nature and prevalence of genetic interactions. As the field of complex genetics evolves, so too will the meaning of genetic interaction. To date, genetic interaction studies have primarily focused on pair-wise interactions that are interpreted on a network level. This is often a result of exploring complex genetics with engineered strains, in which single and double genetic perturbations are systematically performed. The combinatorial expansion makes systematic construction of higher-order perturbations (such as triple-mutants) intractable. Indeed, a comprehensive set of double perturbations is daunting, and would be made even more complicated when one considers the corresponding increase in the number and richness of possible phenotypes. In contrast, populations derived from genetically diverse strains, such as yeast and mouse crosses, will sample all genetic combinations of arbitrarily high order. It is possible that network analysis based on pair-wise interactions will be inadequate for experiments involving these strains. Modules or densely interacting "gene complexes" may better describe the space of polygenic effects. With these issues being compounded by the limited capacity to assay many of the phenotypes most relevant to disease, it is currently unknown which approaches will yield the most information about the underlying biology. It is likely that the optimal approach will vary by organism and experiment.

Mature network analyses developed in a broad and systematic framework could be readily expanded beyond genotype-phenotype modeling. To model effects of environment and drugs, additional factors that affect phenotype might be integrated. Potential factors include metabolites, exogenous molecules such as drugs or environmental compounds, and external conditions such as temperature or cell density. A model encompassing multiple factors could quantify how each factor interacts or influences each other factor, in order to produce a map of how effectors interact to generate a phenotype or outcome. Predictive network models focused on specific phe-

notypes would allow the development of highly specific drugs and aid in the prediction of side effects by observing how a particular perturbation affects connected systems *in silico*. These models could also aid in the identification of allelic combinations that cause genetic disease, the prediction of desired drug-allele interactions for personalized medicine, and the development of multi-drug treatments. Creating such models will require substantial advances in integration of mathematical modeling and genetics, with the focus modest-sized systems rather than genome-scale screens.

In addition to advancing biomedicine, predictive models could also inform evolutionary biology and bioengineering. Rare mutations that confer fitness advantages could be identified by systematically predicting the fitness consequences of perturbations. Perturbations predicted to have effects on multiple phenotypes may uncover evolutionary trade-offs, in which the fitness balance might vary as the environment changes. Fitness assays, such as the E-MAP method in yeast [64], could be used to experimentally test these predictions in model organisms. These effects could in turn be used as input for population modeling and predict evolutionary trends suitable for laboratory testing with fast-replicating yeast or bacteria. As a more practical application, the ability to predict the behavior of re-engineered and synthetic biological networks is a key ingredient in the design and development of novel bioengineered systems. Models that predict complex phenotypes quantitatively and precisely will be a necessary component of elaborate modeling schemes that link genotype, phenotype, and ecosystem.

High throughput data acquisition and the development of computational approaches have enabled the field of systems genetics to progress rapidly. Far-sighted experimental planning and continued technological advances can be expected to fuel continued progress. Current studies, primarily focused on yeast as a model system, are laying the groundwork for network analysis and modeling methods that may be adapted to higher organisms. A target for early adaption may lie in the field of metazoan development to study complex genetic networks involved in multicellular biology, possibly expanding the use of *C. elegans* as a model due to its ease of manipulation by RNAi technology [43,48]. Knowledge accrued from these studies coupled with advances in statistical genetics may form a solid basis for applications in more genetically diverse populations, including the RI mouse strains generated by the Collaborative Cross and their RIX intercross counterparts. The result will be new tools for complex human genetics. Meanwhile, the mapping of the human HapMap, The Cancer Genome Atlas, and human interaction networks will begin to provide data necessary to make use of these tools in personalized and predictive medicine.

Acknowledgments

G.W.C. is supported by grants K25GM079404 and P50GM076547 from NIH and FIBR EF-0527023 from NSF. A.M.D. is supported by a Genome Scholar and Faculty Transition award from the NIH/NHGRI.

Bibliography

1. International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–320
2. Acar M, Becskei A, van Oudenaarden A (2005) Enhancement of cellular memory by reducing stochastic transitions. Nature 435:228–32
3. Altmuller J, Palmer LJ, Fischer G, Scherb H, Wjst M (2001) Genomewide scans of complex human diseases: true linkage is hard to find. Am J Hum Genet 69:936–50
4. Avery L, Wasserman S (1992) Ordering gene function: the interpretation of epistasis in regulatory hierarchies. Trends Genet 8:312–6
5. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK (2003) Computational discovery of gene modules and regulatory networks. Nat Biotechnol 21:1337–42
6. Boone C, Bussey H, Andrews BJ (2007) Exploring genetic interactions and networks with yeast. Nat Rev Genet 8:437–49
7. Brem RB, Kruglyak L (2005) The landscape of genetic complexity across 5700 gene expression traits in yeast. Proc Natl Acad Sci USA 102:1572–7
8. Brem RB, Storey JD, Whittle J, Kruglyak L (2005) Genetic interactions between polymorphisms that affect gene expression in yeast. Nature 436:701–3
9. Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. Science 296:752–5
10. Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T, Su AI, Vellenga E, Wang J, Manly KF, Lu L, Chesler EJ, Alberts R, Jansen RC, Williams RW, Cooke MP, de Haan G (2005) Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. Nat Genet 37:225–32
11. Carlborg O, Haley CS (2004) Epistasis: too often neglected in complex trait studies? Nat Rev Genet 5:618–25
12. Carter GW, Prinz S, Neou C, Shelby JP, Marzolf B, Thorsson V, Galitski T (2007) Prediction of phenotype and gene expression for combinations of mutations. Mol Syst Biol 3:96
13. Carter GW, Rupp S, Fink GR, Galitski T (2006) Disentangling information flow in the Ras-cAMP signaling network. Genome Res 16:520–6
14. Chen KC, Calzone L, Csikasz-Nagy A, Cross FR, Novak B, Tyson JJ (2004) Integrative analysis of cell cycle control in budding yeast. Mol Biol Cell 15:3841–62
15. Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA, Threadgill DW, Manly KF, Williams RW (2005) Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nat Genet 37:233–42

16. Churchill GA, Airey DC, Allayee H, Angel JM, Attie AD, Beatty J, Beavis WD, Belknap JK, Bennett B, Berrettini W, Bleich A, Bogue M, Broman KW, Buck KJ, Buckler E, Burmeister M, Chesler EJ, Cheverud JM, Clapcote S, Cook MN, Cox RD, Crabbe JC, Crusio WE, Darvasi A, Deschepper CF, Doerge RW, Farber CR, Forejt J, Gaile D, Garlow SJ, Geiger H, Gershenfeld H, Gordon T, Gu J, Gu W, de Haan G, Hayes NL, Heller C, Himmelbauer H, Hitzemann R, Hunter K, Hsu HC, Iraqi FA, Ivandic B, Jacob HJ, Jansen RC, Jepsen KJ, Johnson DK, Johnson TE, Kempermann G, Kendziorski C, Kotb M, Kooy RF, Llamas B, Lammert F, Lassalle JM, Lowenstein PR, Lu L, Lusis A, Manly KF, Marcucio R, Matthews D, Medrano JF, Miller DR, Mittleman G, Mock BA, Mogil JS, Montagutelli X, Morahan G, Morris DG, Mott R, Nadeau JH, Nagase H, Nowakowski RS, O'Hara BF, Osadchuk AV, Page GP, Paigen B, Paigen K, Palmer AA, Pan HJ, Peltonen-Palotie L, Peirce J, Pomp D, Pravenec M, Prows DR, Qi Z, Reeves RH, Roder J, Rosen GD, Schadt EE, Schalkwyk LC, Seltzer Z, Shimomura K, Shou S, Sillanpaa MJ, Siracusa LD, Snoeck HW, Spearow JL, Svenson K, Tarantino LM, Threadgill D, Toth LA, Valdar W, de Villena FP, Warden C, Whatley S, Williams RW, Wiltshire T, Yi N, Zhang D, Zhang M, Zou F (2004) The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat Genet 36:1133–7
17. Collins FS, Barker AD (2007) Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. Sci Am 296:50–7
18. Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, Chu CS, Schuldiner M, Gebbia M, Recht J, Shales M, Ding H, Xu H, Han J, Ingvarsdottir K, Cheng B, Andrews B, Boone C, Berger SL, Hieter P, Zhang Z, Brown GW, Ingles CJ, Emili A, Allis CD, Toczyski DP, Weissman JS, Greenblatt JF, Krogan NJ (2007) Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. Nature 446:806–10
19. Collins SR, Schuldiner M, Krogan NJ, Weissman JS (2006) A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. Genome Biol 7:R63
20. Davierwala AP, Haynes J, Li Z, Brost RL, Robinson MD, Yu L, Mnaimneh S, Ding H, Zhu H, Chen Y, Cheng X, Brown GW, Boone C, Andrews BJ, Hughes TR (2005) The synthetic genetic interaction spectrum of essential genes. Nat Genet 37:1147–52
21. Dobzhansky T (1946) Genetics of Natural Populations. Xiii. Recombination and Variability in Populations of Drosophila Pseudoobscura. Genetics 31:269–90
22. Donehower LA, Harvey M, Slagle BL, McArthur MJ, Montgomery CA Jr, Butel JS, Bradley A (1992) Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumours. Nature 356:215–21
23. Drees BL, Thorsson V, Carter GW, Rives AW, Raymond MZ, Avila-Campillo I, Shannon P, Galitski T (2005) Derivation of genetic interaction networks from quantitative phenotype data. Genome Biol 6:R38
24. Facciotti MT, Bonneau R, Hood L, Baliga NS (2004) Systems Biology Experimental Design – Considerations for Building Predictive Gene Regulatory Network Models for Prokaryotic Systems. Current Genomics 5:527–544
25. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol 3:121
26. Forster J, Famili I, Fu P, Palsson BO, Nielsen J (2003) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Res 13:244–53
27. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, Arkin AP, Astromoff A, El-Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian KD, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Guldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kotter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo C, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Ross-Macdonald P, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sookhai-Mahadeo S, Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, Wang CY, Ward TR, Wilhelmy J, Winzeler EA, Yang Y, Yen G, Youngman E, Yu K, Bussey H, Boeke JD, Snyder M, Philippsen P, Davis RW, Johnston M (2002) Functional profiling of the Saccharomyces cerevisiae genome. Nature 418:387–91
28. Gimeno CJ, Ljungdahl PO, Styles CA, Fink GR (1992) Unipolar cell divisions in the yeast S. cerevisiae lead to filamentous growth: regulation by starvation and RAS. Cell 68:1077–90
29. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996) Life with 6000 genes. Science 274:546, 563–7
30. Guarente L (1993) Synthetic enhancement in gene interaction: a genetic tool come of age. Trends Genet 9:362–6
31. Gunsalus KC, Ge H, Schetter AJ, Goldberg DS, Han JD, Hao T, Berriz GF, Bertin N, Huang J, Chuang LS, Li N, Mani R, Hyman AA, Sonnichsen B, Echeverri CJ, Roth FP, Vidal M, Piano F (2005) Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis. Nature 436: 861–5
32. Hartman JL, Garvik B, Hartwell L (2001) Principles for the buffering of genetic variation. Science 291:1001–1004
33. Harvey M, McArthur MJ, Montgomery CA Jr, Bradley A, Donehower LA (1993) Genetic background alters the spectrum of tumors that develop in p53-deficient mice. Faseb J 7:938–43
34. Harvey M, Vogel H, Morris D, Bradley A, Bernstein A, Donehower LA (1995) A mutant p53 transgene accelerates tumour development in heterozygous but not nullizygous p53-deficient mice. Nat Genet 9:305–11
35. Haugen AC, Kelley R, Collins JB, Tucker CJ, Deng C, Afshari CA, Brown JM, Ideker T, van Houten B (2004) Integrating phenotypic and expression profiles to map arsenic-response networks. Genome Biol 5:R95
36. Huang LS, Sternberg PW (1995) Genetic dissection of developmental pathways. Methods Cell Biol 48:97–122
37. Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, Maciver F, Mueller M, Hummel O, Monti J, Zidek V, Musilova A, Kren V, Causton H, Game L, Born G, Schmidt S, Muller A, Cook SA, Kurtz TW, Whittaker J, Pravenec M, Aitman TJ (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat Genet 37:243–53
38. Hwang D, Rust AG, Ramsey S, Smith JJ, Leslie DM, Weston AD, de Atauri P, Aitchison JD, Hood L, Siegel AF, Bolouri H (2005) A data integration methodology for systems biology. Proc Natl Acad Sci USA 102:17296–301
39. Ideker T, Galitski T, Hood L (2001) A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet 2:343–72

40. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292:929–34
41. Jansen RC (2003) Studying complex biological systems using multifactorial perturbation. Nat Rev Genet 4:145–51
42. Jansen RC, Nap JP (2001) Genetical genomics: the added value from segregation. Trends Genet 17:388–91
43. Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, Welchman DP, Zipperlen P, Ahringer J (2003) Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature 421:231–7
44. Kastenmayer JP, Ni L, Chu A, Kitchen LE, Au WC, Yang H, Carter CD, Wheeler D, Davis RW, Boeke JD, Snyder MA, Basrai MA (2006) Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. Genome Res 16:365–73
45. Kelley R, Ideker T (2005) Systematic interpretation of genetic interactions using protein networks. Nat Biotechnol 23:561–6
46. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241–54
47. Klose J, Nock C, Herrmann M, Stuhler K, Marcus K, Bluggel M, Krause E, Schalkwyk LC, Rastan S, Brown SD, Bussow K, Himmelbauer H, Lehrach H (2002) Genetic analysis of the mouse brain proteome. Nat Genet 30:385–93
48. Lehner B (2007) Modelling genotype-phenotype relationships and human disease with genetic interaction networks. J Exp Biol 210:1559–66
49. Lehner B, Crombie C, Tischler J, Fortunato A, Fraser AG (2006) Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. Nat Genet 38:896–903
50. Lehner B, Tischler J, Fraser AG (2006) RNAi screens in Caenorhabditis elegans in a 96-well liquid format and their application to the systematic identification of genetic interactions. Nat Protoc 1:1617–20
51. Li H, Chen H, Bao L, Manly KF, Chesler EJ, Lu L, Wang J, Zhou M, Williams RW, Cui Y (2006) Integrative genetic analysis of transcription modules: towards filling the gap between genetic loci and inherited traits. Hum Mol Genet 15:481–92
52. Maslov S, Sneppen K, Eriksen KA, Yan KK (2004) Upstream plasticity and downstream robustness in evolution of molecular networks. BMC Evol Biol 4:9
53. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298:824–7
54. Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, Edwards S, Phillips JW, Sachs A, Schadt EE (2004) Genetic inheritance of gene expression in human cell lines. Am J Hum Genet 75:1094–1105
55. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430:743–7
56. Ooi SL, Pan X, Peyser BD, Ye P, Meluh PB, Yuan DS, Irizarry RA, Bader JS, Spencer FA, Boeke JD (2006) Global synthetic-lethality analysis and yeast functional profiling. Trends Genet 22:56–63
57. Pan X, Ye P, Yuan DS, Wang X, Bader JS, Boeke JD (2006) A DNA integrity network in the yeast Saccharomyces cerevisiae. Cell 124:1069–81
58. Pan X, Yuan DS, Xiang D, Wang X, Sookhai-Mahadeo S, Bader JS, Hieter P, Spencer F, Boeke JD (2004) A robust toolkit for functional profiling of the yeast genome. Mol Cell 16:487–96
59. Perlstein EO, Ruderfer DM, Roberts DC, Schreiber SL, Kruglyak L (2007) Genetic basis of individual differences in the response to small-molecule drugs in yeast. Nat Genet 39:496–502
60. Ramsey SA, Smith JJ, Orrell D, Marelli M, Petersen TW, de Atauri P, Bolouri H, Aitchison JD (2006) Dual feedback loops in the GAL regulon suppress cellular heterogeneity in yeast. Nat Genet 38:1082–1087
61. Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hon GC, Myers CL, Parsons A, Friesen H, Oughtred R, Tong A, Stark C, Ho Y, Botstein D, Andrews B, Boone C, Troyanskya OG, Ideker T, Dolinski K, Batada NN, Tyers M (2006) Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. J Biol 5:11
62. Rockman MV, Kruglyak L (2006) Genetics of global gene expression. Nat Rev Genet 7:862–72
63. Schacherer J, Ruderfer DM, Gresham D, Dolinski K, Botstein D, Kruglyak L (2007) Genome-wide analysis of nucleotide-level variation in commonly used Saccharomyces cerevisiae strains. PLoS ONE 2:e322
64. Schuldiner M, Collins SR, Thompson NJ, Denic V, Bhamidipati A, Punna T, Ihmels J, Andrews B, Boone C, Greenblatt JF, Weissman JS, Krogan NJ (2005) Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. Cell 123:507–19
65. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 34:166–76
66. Segre D, Deluna A, Church GM, Kishony R (2005) Modular epistasis in yeast metabolism. Nat Genet 37:77–83
67. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–504
68. Shmulevich I, Kauffman SA, Aldana M (2005) Eukaryotic cells are dynamically ordered or critical but not chaotic. Proc Natl Acad Sci USA 102:13439–44
69. Sinha H, Nicholson BP, Steinmetz LM, McCusker JH (2006) Complex genetic interactions in a quantitative trait locus. PLoS Genet 2:e13
70. Spiro PA, Parkinson JS, Othmer HG (1997) A model of excitation and adaptation in bacterial chemotaxis. Proc Natl Acad Sci USA 94:7263–8
71. St Onge RP, Mani R, Oh J, Proctor M, Fung E, Davis RW, Nislow C, Roth FP, Giaever G (2007) Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. Nat Genet 39:199–206
72. Steinmetz LM, Sinha H, Richards DR, Spiegelman JI, Oefner PJ, McCusker JH, Davis RW (2002) Dissecting the architecture of a quantitative trait locus in yeast. Nature 416:326–30
73. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122:957–68

74. Storey JD, Akey JM, Kruglyak L (2005) Multiple locus linkage analysis of genomewide expression in yeast. PLoS Biol 3:e267
75. Sturtevant AH (1956) A Highly Specific Complementary Lethal System in Drosophila Melanogaster. Genetics 41:118–23
76. Taylor RJ, Siegel AF, Galitski T (2007) Network motif analysis of a multi-mode genetic-interaction network. Genome Biol 8:R160
77. Thakar J, Pilione M, Kirimanjeswara G, Harvill ET, Albert R (2007) Modeling systems-level regulation of host immune responses. PLoS Comput Biol 3:e109
78. Tischler J, Lehner B, Chen N, Fraser AG (2006) Combinatorial RNA interference in Caenorhabditis elegans reveals that redundancy between gene duplicates can be maintained for more than 80 million years of evolution. Genome Biol 7:R69
79. Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, Andrews B, Tyers M, Boone C (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science 294:2364–2368
80. Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Menard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu AM, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C (2004) Global mapping of the yeast genetic interaction network. Science 303:808–13
81. Tucker CL, Fields S (2003) Lethal combinations. Nat Genet 35:204–5
82. Van Driessche N, Demsar J, Booth EO, Hill P, Juvan P, Zupan B, Kuspa A, Shaulsky G (2005) Epistasis analysis with global transcriptional phenotypes. Nat Genet 37:471–7
83. Vidal M (2005) Interactome modeling. FEBS Lett 579:1834–8
84. Waddington CH (1942) Canalization of Development and the Inheritance of Acquired Characters. Nature 150:563
85. Wayne ML, McIntyre LM (2002) Combining mapping and arraying: An approach to candidate gene identification. Proc Natl Acad Sci USA 99:14903–6
86. Winston F, Dollard C, Ricupero-Hovasse SL (1995) Construction of a set of convenient Saccharomyces cerevisiae strains that are isogenic to S288C. Yeast 11:53–5
87. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, Chu AM, Connelly C, Davis K, Dietrich F, Dow SW, El Bakkoury M, Foury F, Friend SH, Gentalen E, Giaever G, Hegemann JH, Jones T, Laub M, Liao H, Liebundguth N, Lockhart DJ, Lucau-Danila A, Lussier M, M'Rabet N, Menard P, Mittmann M, Pai C, Rebischung C, Revuelta JL, Riles L, Roberts CJ, Ross-MacDonald P, Scherens B, Snyder M, Sookhai-Mahadeo S, Storms RK, Veronneau S, Voet M, Volckaert G, Ward TR, Wysocki R, Yen GS, Yu K, Zimmermann K, Philippsen P, Johnston M, Davis RW (1999) Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science 285:901–6
88. Wong SL, Zhang LV, Tong AH, Li Z, Goldberg DS, King OD, Lesage G, Vidal M, Andrews B, Bussey H, Boone C, Roth FP (2004) Combining biological networks to predict genetic interactions. Proc Natl Acad Sci USA 101:15682–7
89. Workman CT, Mak HC, McCuine S, Tagne JB, Agarwal M, Ozier O, Begley TJ, Samson LD, Ideker T (2006) A systems approach to mapping DNA damage response pathways. Science 312:1054–1059
90. Ye P, Peyser BD, Pan X, Boeke JD, Spencer FA, Bader JS (2005) Gene function prediction from congruent synthetic lethal interactions in yeast. Mol Syst Biol 1:0026
91. Yeang CH, Mak HC, McCuine S, Workman C, Jaakkola T, Ideker T (2005) Validation and refinement of gene-regulatory pathways on a network of physical interactions. Genome Biol 6:R62
92. Yeger-Lotem E, Sattath S, Kashtan N, Itzkovitz S, Milo R, Pinter RY, Alon U, Margalit H (2004) Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. Proc Natl Acad Sci USA 101:5934–9
93. Yeh P, Tschumi AI, Kishony R (2006) Functional classification of drugs by properties of their pairwise interactions. Nat Genet 38:489–94
94. Zhang LV, King OD, Wong SL, Goldberg DS, Tong AH, Lesage G, Andrews B, Bussey H, Boone C, Roth FP (2005) Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network. J Biol 4:6
95. Zhong W, Sternberg PW (2006) Genome-wide prediction of C. elegans genetic interactions. Science 311:1481–4

T

Thermodynamics of Computation

H. John Caulfield, Lei Qian
Fisk University, Nashville, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Thermodynamics
Computer Equivalents of the First and Second Laws
The Thermodynamics of Digital Computers
Analog and Digital Computers
Natural Computing
Quantum Computing
Optical Computing
Thermodynamically Inspired Computing
Cellular Array Processors
Conclusions
Future Directions
Bibliography

Glossary

Thermodynamics Literally accounting for the impact of temperature and energy flow. It is a branch of physics that describes how temperature, energy, and related properties affect behavior of an object or event.

First law of thermodynamics Also known as the law of *energy conservation*. It states that the energy remains constant in an isolated system.

Second law of thermodynamics The second law of thermodynamics asserts that the *entropy* of an isolated system never decreases with time.

Entropy A measure of the disorder or unavailability of energy within a closed system.

Computing Any activity with input/output patterns mapped onto real problems to be solved.

Turing machine An abstract computing model that was invented by Alan Turing in 1936, long before the first electronic computer was invented, to serve as an idealized model for computing. A Turning machine has a tape that is unbounded in both directions, a read-write head and a finite set of instructions. At each step, the head may modify the symbol on the tape right under the head, change the state of the head, and then move on the tape one unit to the left or right. Although extremely simple, Turing machines can solve any problem that can be solved by any computers that could possibly be constructed (see **Church–Turing thesis**).

The Church–Turing thesis A combined hypothesis about the nature of computable functions. It states that *any function that is naturally regarded as computable is computable by a Turing Machine*. In early 20th century, various computing models such as Turing Machine, λ-calculus and recursive functions are invented. It was proved that Turing Machine, λ-calculus and recursive functions are equally powerful. Alonzo Church and Alan Turing independently raised the thesis that any *naturally computable function* must be a recursive function or, equivalently, be computed by a Turing Machine, or be a λ-definable function. In other words, it is not possible to build a computing device that is more powerful than those machines with simplest computing mechanisms. Note that Church–Turing thesis is not a provable or refutable conjecture because "*naturally computable function*" is not a rigorous definition. There is no way to prove or refute it. But it has been accepted by nearly all mathematicians today.

Analog An analog computer is often negatively defined as a computer that is not digital. More properly, it is a computer that uses quantities are that can be made proportional to the amount of signal detected. That analogy between a real number and a physical property gives the name "analog." Many problems arise in analog computing, because it is so difficult to obtain

a large number of distinguishable levels and because noise builds up in cascaded computations. But, being unencoded, it can always be run faster than its digital counterpart.

Digital The name derives from digits (the fingers). Digital computing works with discrete items like fingers. Most digital computing is binary – using 0 and 1. Because signals are restored to a 1 or a 0 after each operation, noise accumulation is not much of a problem. And, of course, digital computers are much more flexible than analog computers that tend to be rather specialized.

Quantum computer A quantum computer is a computing device that makes direct use of quantum mechanical phenomena such as superposition and entanglement. Theoretically, quantum computers can achieve much faster speed than traditional computers. It can solve some NP hard or even exponentially hard problem within linear time. Due to many technical difficulties, no practical quantum computer using entanglement has been built up to now.

Definition of the Subject

Thermodynamics is a very general field that helps to define the physics limits that impose on all stuff. When people study an abstract computing model such as a Turning Machine, thermodynamics does not impact their behavior because nothing physical is been done. However, real computers must be made of some physical material, material that has a temperature. The physical stuff must have at least two readily distinguishable states that can be interpreted to two different values. In the process of computing, the stuff should be able to change the state under control. The change of the state must obey the laws of thermodynamics. For example, there is a minimum amount of energy associated with a bit of information. Information is limited by the entropy change $k \log 2$ and the energy is $kT \log 2$, where k is the Boltzmann's constant and T is the absolute temperature of the device and its surroundings. Mysteriously, information and entropy seem linked even though one is physical and the other is conceptual. Those limits are vital to know, as it allows designers to stop seeking improvements as they reach or near those limits. At the same time, thermodynamics indicates where computers do not come near what is possible. Such aspects of computers represent potentially important fields to explore.

Introduction

Thermodynamics – the science of the changes in energy over time – is one of the most fundamental sciences. Arguably, the first and second laws of thermodynamics are the scientific principles most scientists would label as the most fundamental of all the laws of science. No one ever expects them to be violated. We will explore the following aspects of the relationship between thermodynamics and computing

- The energetics and corresponding temporal properties of classical digital computers.
- Possible modifications such as conservative computing, analog computing, and quantum computing.
- Thermodynamically-inspired computing.

Unsurprisingly, we will find that those topics are all related.

Before we can do any of those things, however, we must review thermodynamics itself and do so in terms suited for this discussion

Thermodynamics

Thermodynamics originated as a set of general rules characterizing the effects of system parameters such as energy and temperature over time. Somewhat later, thermodynamics came to be partitioned into two broad classes – the classical thermodynamics (now called equilibriums thermodynamics as it predicts the final system parameters that result from some perturbing effect) and nonequilibrium thermodynamics (what happens if energy input to a system continues over time so equilibrium is never obtained).

The fact that thermodynamics applies to quantities averaged over the system and deals only with probabilistic results, means that no physical law is broken if entropy within the system decreases due to either of two reasons: the occurrence of an improbable event that is allowable but fleeting or the occurrence of a state that is not reachable by equilibrium thermodynamics but is reachable with continuing input of energy. The latter is an example of nonequilibrium thermodynamics. This also connects nonequilibrium thermodynamics with topics such as nonlinear dynamics, chaos, emergence, and so forth. Two concepts from nonlinear dynamics are especially interesting: attractor states and emergence. Attractor states are system states such that if the state of the system is near a given attractor (within its "attractor basin"), successive states tend to move toward that attractor. Some of the attractor states in systems far from equilibrium can be highly organized. The properties of highly organized nonequilibrium thermodynamic states (states can be fixed conditions, periodic events, or even more complex situations called strange attractors that are somewhat predictable and necessarily chaotic in the technical sense of that word. There is generally a control parameter such as power input which de-

termines the system performance. For some values systems often develop excessive sensitivity to the values of the control parameter leading to long term unpredictability of the deterministic result – a condition called chaos. Systems that are too unorganized like an ideal gas, are not very interesting form a behavior viewpoint. Systems that are fully organized like a crystal also have limited interest. But between those extremes lie regions of high complexity and great interest. One of the favorite and well justified sayings of complexity theorists is that life arises on the edge of chaos. You are a system that shows spectacular organization if kept away from equilibrium with continuing input of energy. The properties that arise far from equilibrium are said to be emergent and are generally unpredictable until discovered experimentally. A general feature of such organized complexity in systems maintained well away form equilibrium is self organization. That is the order observed does not have to be created by a scientist. Instead it arises spontaneously under the correct circumstances.

Turing machines can be accounted for by conventional thermodynamics, but some computer such as neural networks and cellular array processors can become complex enough to exhibit interesting and useful emergent properties that must be discussed in terms of nonlinear thermodynamics. Such self organized complex processor behavior has also been called "synergistic."

The most complex computer we know is the human brain. It not only arises on the edge of chaos, but also uses and consumes chaos. Over-regular temporal behavior of the brain is called epilepsy and is not a desirable state.

Two of the first and most influential people to explore nonequilibrium thermodynamics were Warren Weaver and Ilya Priogogine.

Weaver [1] discussed three ways people can understand the world: simple systems, statistical methods (conventional thermodynamics), and nonequilibrium organized complexity [thermodynamics]. Organized complexity is compatible with classical thermodynamics from an overall perspective but can be achieved locally in space and time.

Prigogine's Nobel-Prize-winning work on nonequilibrium thermodynamics is explained beautifully in his popular books [2,3,4].

Let us now review very briefly what thermodynamics deals with and why it is important to computers.

Temperature, T, is a measure of the driving force for energy. There can be no systematic energy flow from a lower temperature region to a higher temperature region.

Entropy, S, is a measure of the disorder. The second law of thermodynamics, perhaps science's most sacred law, says that entropy never decreases. Indeed it nearly always increases. Note that entropy is essentially an abstract quantity. Indeed, it is the negative of information as defined by Shannon. Therefore information is sometimes called negentropy – negative entropy. There is a metaphysical puzzle here that has been more ignored than solved. How does a mathematical concept like information come to govern physics as does entropy?

Boltzmann's constant k symbolizes the relationship between information and physics. It appears in purely physical equations such as the ideal gas law. It also appears in equations about information. Leo Szilard suggested that writing a memory does not produce entropy; it is possible to store information into a memory without increasing entropy. However, entropy is produced in every case that the memory is *erased*. It turns out that the (minimum) entropy created by erasing one bit is given by

$$S \text{ per erased bit} = k \ln 2 \,,$$

and the number $\ln 2 \approx 0.69$ is the natural logarithm of 2. What's going on here? How are information and the physical world related?

Information relates to the answers to yes or no questions. Entropy is physical entity not necessarily related to questions and answers. Yet the mathematics of both is the same. Information is the negative of entropy. What are we to make of this metaphysically? One view (originating in the mind and writing of John Archibold Wheeler) is that information is primary. The slogan is "It from Bit." Wheeler's metaphysics is extremely appealing, because it requires no external scaffolding. But it strikes many as not quite convincing. There is an opposite approach that begins with stuff and energy. The physical world is made of relationships between parts. The universe is self organized and self defined. Stuff is differentiation (yes and no) within the whole. Another view is that physics (stuff, energy, motion) is primary. Information in a brain or a computer is present only if some stuff is used and some energy is expended in generating, transmitting, or storing it. In that case, information relates to the probability that a given record would appear by chance if all records were equally probable. The less likely the chance appearance, the more information is present. In this way, information ties back to Bayes's Theorem. Neither caricature of a position is complete enough to be fair, but perhaps this footnote will stimulate some readers to dig deeper. This is not a metaphysics chapter, so we take no side in this dispute. But, some of us are physicists and physicists are drawn to physics by the hope of finding answers not just equations. We cannot help asking such questions, so we think noting this unresolved metaphysical problem is appropriate.

Note also, that no metaphysical problem can ever be fully resolved, as the criteria for truth of metaphysics lie in the domain of metametaphysics (Ockham's razor, etc.) and so on ad infinitum.

There is a fundamental relationship between the change ΔE in the physical quantity energy and the change ΔS in the mathematical quantity entropy, namely

$$\Delta E = T\Delta S\,.$$

What is remarkable about thermodynamics is that it tells us nothing about how things actually work. It tells us instead about what cannot be done. It is the science of fundamental limits. It is, in that sense, a negative field. Physicists sometimes have to work very hard to find the detailed ways nature enforces these limits in any particular problem.

When energy ceases to have any net flow, the system is said to be at equilibrium. Until the latter part of the 20th century, thermodynamics meant near-equilibrium behavior. Far from equilibrium, very interesting things can occur. Order can arise from disorder at least locally. The second law is satisfied by creating even more disorder somewhere in the universe. Life exists far from equilibrium and seems to increase order. Indeed it does – locally. But it must take energy from its environment and cause a global increase in energy.

The laws of thermodynamics define and thus forbid magic – something from nothing (the first law) or effects without causes (the second law). We will interpret all magic forbidding laws as "thermodynamic" even if the apply to nonphysical things such as information

Computer Equivalents of the First and Second Laws

We propose two laws of information inspired by the two laws of thermodynamics just described.

First law: Information is not increased in a computer. A computer cannot create information. The output result is strictly determined by the data and the instructions. If we were smart enough and had unlimited resources (time and memory), we would know that $1 + 2 + 3 + 4 + 5 = 15$. We only need a computer, because computer can compute faster and more reliably. We may not have known the answer now, but we could if we had enough time, memory and patience. The result is not new but derived information.

Second law: The energy required to destroy one bit of information is not smaller than the temperature T multiplied by some minimum amount of information ΔS. This second law applies for conservative as well as dissipative logic – terms to be described later. This second law, like the second law of thermodynamics, deals with irreversible acts.

The Thermodynamics of Digital Computers

Just as all digital computers are equivalent to a Turing machine, all computations in digital computers can be thought of as being performed by Boolean logic gates. If we understand the thermodynamics of a Boolean logic gate, we understand the thermodynamics of any digital computer.

All readers of this chapter will recognize that Boolean logic gates such as *AND*, *OR*, *NOR*, and so forth convert two one-bit inputs into a single one-bit output. Immediately, we notice

- A Boolean logic gate has less information out than we put in.
- Such a logic gate is irreversible.

From the two laws of computing, we conclude.

- Since information must be conserved, one bit of information must somehow have been converted into some noninformative state.
- The amount of energy required is related to the information content of a single bit multiplied by the operating temperature T.

The person who first realized and determined the amount of energy needed to destroy one bit of information is

$$\Delta E = kT \log 2$$

was an IBM scientist named Landauer. Here k is the Boltzmann constant. The average energy of a molecule at temperature T is about kT, so this makes some intuitive sense as well.

This is really a very small amount of energy – molecules of that energy are striking you right now and you can't even feel them. Two things make this energy remark important. First, most current gates require millions of kTs per operation. This is no accident. It is required as we will see later when we discuss analog computing. Second, the number of operations per second is now in the billions. So trillions of kTs are dissipated in current computers each second. They get very hot and consume a great amount of power.

Landauer and fellow IBM physicist Bennett then took a totally unexpected and brilliant direction of research. What would happen, they asked, if we could use logic gates that do not destroy information? Then there would be no thermodynamic minimum energy requirement for such

a gate's operation. Perhaps the energy cost of computing could be reduced or even eliminated.

A trivial example from arithmetic will illustrate these concepts for readers not familiar with them.

Here is a conventional (dissipative) arithmetic expression:

$$2 + 3 = 5 .$$

It would be more informative to write

$$2 + 3 \Rightarrow 5 .$$

If we know the data (2,3) and the instructions (+), the result (5) is determined. But this is irreversible. If we know the result (5) and the instruction to be inverted (+), we cannot recover the data (2, 3). The data could have been (3, 2), (1, 4), etc. Here is an information conserving arithmetic expression:

$$2, 3 \circledR 3, 2 .$$

The operator ® means reverse the order of the data. Such an operation is clearly conservative of information. And, clearly, it is reversible:

$$3, 2 \circledR 2, 3 .$$

An ® gate would have no inevitable energy price. It destroys no information (our first law is satisfied).

Here is a not such trivial example. Consider the operation:

$$a, b \circledcirc a + b, a - b$$

This operation is also conservative of information because with $a + b$ and $a - b$, one can easily recover the original a and b.

Since that time, a number of conservative, reversible logic gates have been proposed and implemented. We can say some things in general about such gates.

- They will have equal numbers of inputs and outputs (usually three).
- The changes effected by the gates must either rearrange the data (like the ® operator in our arithmetic example) or mix them (like the © in the second example). All electronic gates so far proposed involve rearrangements. In optical logic, reversible mixing is common.

Stringing conservative logic gates together can allow us to produce sought after results. Inevitably, however, we will also have computed things we do not care about – called garbage.

Garbage is inevitable in reversible computing, because the components must retain the information normally destroyed by a conventional Boolean logic gate. Reversibility eliminates one problem (the inevitable dissipation of information) but creates another (the inevitable appearance of garbage bits).

Nevertheless, we must detect the desired answer, and that must cost at least $kT \log 2$. The zero energy possibility applies only to the logic operations, not the entering of data into the system or the extraction of data from the system. Failure to understand this has caused more than a few to proclaim that zero-energy logic operations are impossible.

We enter data at the input ports and wait for an answer to reach the output ports. As the system is reversible, not all occurrences lead in the "right direction." Temperature-dependent random walk will eventually produce the answer. But with a little energy bias in the forward direction, we will get the answer faster at the cost of expending energy – the more energy, the faster the results are likely to arrive. One might expect that something akin to the energy-time uncertainty principle might apply here, so very low energy gates might require a very long time to operate. Indeed that has usually been the case. An exception will be noted in the case of certain optical elements.

Analog and Digital Computers

> Because digital computers and computation have been so successful, they have influenced how we think about both computers as machines and computation as a process – so much so, it is difficult today to reconstruct what analog computing was all about… It is a history in which digital machines can do things 'better' and 'faster' than other machines… However, what is at stake here are not matters of speed or precision. Rather, it is an argument about what can be rendered and understood through a machine that does computation [5].

All computers involve the operation of some physical system and the interpretation of physically measurable quantities in numerical terms. The most obvious interpretation is

$$N = cQ .$$

The number N is the measured quantity Q multiplied by some conversion coefficient. In classical physics Q is a continuous variable. Computers using such numbers are called analog computers.

Analog computers have several profound disadvantages relative to digital computers.

- **Sensitivity**. By this we mean the ability to distinguish small difference in large numbers. Analog computers tend to have trouble being able to distinguish, say, 47 from 48. Digital computers only have to distinguish between 0 and 1. It encodes bigger numbers in terms of 0s and 1s.
- **Inability to compute complex or even real numbers in most cases**, because most useful quantities Q are usually nonnegative. This is usually "solved" by encryption of real or complex numbers as multiple nonnegative numbers.
- **Cumulative errors**. When the output from one analog computation is used as an input to another, the errors are carried forward and accumulated. This effect is made even worse by the fact that some problems amplify the input errors significantly. In linear algebra problems, do instance, that error amplification factor is called the condition number of the matrix. That number is never less than 1 and often exceedingly large (ill conditioned matrices) or even infinite (singular matrices).
- **Inflexibility**. This chapter was written using a word processor in a digital computer. Making an analog word processor would require almost inconceivably complex.

Some distinct advantages come from being continuous in space or time [6].

In addition, it is widely believed that anything an analog computer can compute can also be computed by a digital computer. In fact, that belief is a corollary of what is called the strong Church–Turing thesis. It asserts that anything that can be computed in any way (including analog computations) can also be computed by a Turing machine. The more widely accepted form of the Church–Turing thesis simply asserts that anything that is sensibly computable by any computer can be computed by a Turing machine. There are some who believe they have discovered exceptions to the Church–Turing thesis and many more who doubt that. This is not the place to examine that controversy. It is mentioned here for two reasons.

First, it suggests that analog computers cannot do anything a digital computer cannot do. There is no advantage in principle. There may be practical advantages to analog computing when it suffices as we have already suggested, though. But there is much more to computing than simply getting the right answer. Speed, cost, and power consumption are examples. Sometimes, analog computing can win on those criteria.

Second, the Church–Turing thesis is thermodynamic in the generalized sense that it sets presumably immutable constraints but does not provide information on how those constraints are enforced.

So why would anyone bother with analog computers? There are many reasons. Here are some obvious ones.

- A binary gate is an analog gate followed by a nonlinear operator (a quantizer) that sets low signals to 0 and high signals to 1 or vice versa. So the difference between analog and binary computing lies in two things – the number of quantization levels (2 for binary and many more for analog) and how often measurements and corrections are made (after each operation for binary computers and after the whole computation for analog). So, we can argue that all computers are analog at their hearts.
- But encryption in terms of many binary operations, they are inevitably slower and more power hungry than digital computers. We could digitize to more than two values. It takes far fewer ternary values than binary values to represent a large number, but the more quantization levels, the more probable it is that we will make an error. It is generally assumed that two is the optimum number overall.

But there are more subtle effects that sometimes favor analog over digital.

Consider the representation of physically measured time signal in a digital computer. The sampling theorem tells us how frequently we must sample the signal IF we know its bandwidth. But suppose we do not know. We can do either of two dangerous things: apply a bandpass filter and risk losing information or guess the bandwidth and risk producing aliasing. So discretization of an analog signal has a significant probability of giving misleading results.

Iterative digital solution of linear algebra problems converge only for condition numbers limited by the digitization itself, while analog solutions have no such limit because the condition number itself is undefined.

A more general treatment can be found in [7].

For chaotic systems, discretization can lead to wildly different results with different levels of discretization [8,9,10].

So what can we say about the thermodynamics of analog computers? The operation occurs in three steps:

1. **Preparation**. The apparatus is set up so that the input parameters are the ones we choose and the physics will carry out the desired transformations.
2. **Time Evolution**. The physical system evolves through time.

3. **Readout**. The desired physical parameters are measured.

Preparation is necessary in digital computing as well and has not been treated in any thermodynamic discussion. Nor will we treat it here. The last two steps must somehow be governed by the two laws of computing propounded earlier. That is, the system cannot produce what we interpret as information beyond that we inserted in the preparation. The underlying physics may be exceedingly complex, but the informatics must be as described by the first law. The readout requires some minimum energy. In a digital computer with L levels the Landauer analysis says we must dissipate $kT \log L$ in energy. Here is at least a qualitative explanation of the fact that we need far more than kT to operate a binary logic gate. We need to measure to far more than just two levels in order to make the binary decisions with sufficiently low error rate.

Analog computing is an emerging field, in part because of interest in what is called "natural computing."

Natural Computing

In serial non Von Neumann machines the interconnections between processors have negligible impact on system space and energy consumption. But concurrency (e. g. parallelism) requires more interconnections and more attention to their space and energy consumption. And those problems become worse as the number of concurrent processors increases. For large enough numbers of processors, interconnections are no longer negligible and can become limiting.

Two approaches have been proposed to ameliorate such problems: optical interconnection and quantum computing. Optical interconnection may be useful, because optical paths can cross in space without effect. That is many optical paths can share the same space. Quantum computing is of interest because N entangled states can probe 2^N paths. Both avoid massive interconnection with wires.

Richard Feynman anticipated both approaches. One of Feynman's goals in suggesting quantum computing was to solve the wiring problem. Another motivation also guided him. He knew that the accuracy with which the results of a quantum experiment could be predicted depended on the number of path integrals evaluated, so (in a sense) each natural event in quantum mechanics is equivalent to infinitely many calculations. The only way to calculate quantum events with perfect accuracy would be to use a quantum computer.

Let us describe a natural computation. We begin with a physical system P that operates on an input I_N to produce an output O_N. It does this naturally without human assistance.

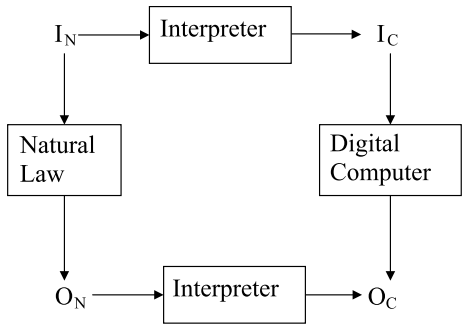

Thermodynamics of Computation, Figure 1
Nature converts input I_N into output O_N by natural law. The digital computer performing some fixed algorithm converts input I_C into output O_C. If, for all allowable I_N, we have a rule that concerts it to an equivalent I_C and if the output of the natural computer I_N can be converted to the output I_C, then we have in the physical operator a "natural computer"

Now suppose we wish to use that physical process P to calculate some result. If there is a computer operation C that operates on an input I_C to produce an output O_C, and if there are mappings

$$\text{M1} : I_N \leftrightarrow I_C$$
$$\text{M2} : O_N \leftrightarrow O_C$$

Then we can declare that the P and C are congruent and thus compute the same results if the diagram in Fig. 1 commutes.

That is if nature in doing what it will do automatically allows us to establish the starting state I_N (state preparation) in a way that can be viewed using M1 as representing the computer input I_C and if the output O_N of the natural process and if M2 carries that to the very output that a conventional computer produces with input, then the two processes are congruent and we have a natural computer.

Certainly, the most widely studied form of natural computing at this moment is the quantum computer.

Quantum Computing

The behaviors of matter in very small scale obey laws of quantum mechanics, which are so different from what we usually experience as to be essentially beyond our understanding. We can write the equations and perform the experiments but the results inevitably seem weird. Despite this, it is possible to harness some of these effects for computing. Quantum computers usually utilize superposition

(a principal of quantum mechanism which claims that while the state of any object is unknown, it is actually in all possible states simultaneously, as long as we don't measure it) and quantum entanglement (a quantum mechanical phenomenon in which the quantum states of two or more objects have to be described with reference to each other, even though the individual objects may be spatially separated) to perform computation [11,12].

In traditional digital computers, information is carried by numerous binary digits, which are called bits. In quantum computing, information is carried by particles which carry quantum states. Such quantum logic bit is called qubit. A qubit can be expressed by

$$|i\rangle = a\,|0\rangle + b\,|1\rangle$$

where $|0\rangle$ and $|1\rangle$ are two orthonormal states. The coefficients a and b are the normalized probability amplitudes (in general, complex-valued numbers) of states $|0\rangle$ and $|1\rangle$ that satisfy the condition $|a|^2 + |b|^2 = 1$. These two orthonormal states can be polarization in orthogonal directions, two oppositely directed spins of atoms, etc. It seems that a qubit carries partially information of $|0\rangle$ and partially information of $|1\rangle$. However, when a qubit is measured (which is necessary if we want to read the output), the qubit "collapses" to either state 0 or state 1 with the probabilities $|a|^2$ or $|b|^2$. This process is irreversible. Whence a qubit collapses to a state, there is no way to recover the original state. On the other hand, it is also not possible to clone a quantum state without destroying the original one.

A very weird but extremely useful phenomenon in quantum world is **entanglement**. It is the coherence among two or more objects even though these objects are spatially separated. Consider a 2-qubit state: $(|0\rangle|0\rangle + |0\rangle|1\rangle + |1\rangle|0\rangle + |1\rangle|1\rangle)/2$. It can be decomposed to $((|0\rangle + |1\rangle)/\sqrt{2})((|0\rangle + |1\rangle)/\sqrt{2})$. That means we have 50% probability to read $|0\rangle$ and 50% probability to read $|1\rangle$ when we measure the first qubit. We will get same probability for the second qubit. The measured value of the first qubit is independent of that of the second qubit. However, consider the state: $1/\sqrt{2}(|0\rangle|1\rangle - |1\rangle|0\rangle)$. It is easy to prove that one cannot decompose the state into a form $(a\,|0\rangle + b\,|1\rangle)(c\,|0\rangle + d\,|1\rangle)$. Actually, if we measure the first qubit, we have 50% probability to read $|1\rangle$ and 50% chance to read $|0\rangle$. However, if we read $|0\rangle$ when the first qubit is measured, we **always** read $|1\rangle$ for the second qubit; if we read $|1\rangle$ for the first qubit, we **always** read $|0\rangle$ for the second qubit. The measurement to one of them will affect the state of the other qubit immediately no matter how far they are from each other.

Consider a quantum system with n qubits. If there is no entanglement, it can be decomposed into the product of n 1-qubit systems, or $2n$ independent numbers. However, if entanglement is allowed, then there are 2^n independent coefficients and therefore can hold 2^n independent values. So it can operate tremendous number of values in parallel. It is possible to solve exponentially hard or NP complete problems in polynomial or even linear time. In 1994, Peter Shor of AT&T devised an algorithm for factoring an integer N in $O((\log N)^3)$ time and $O(\log N)$ space with a quantum computer. The security of RSA, the most widely used public-key cryptography method is based on the assumption that factoring a large integer is not computationally feasible. A quantum computer can break it easily.

Even though quantum computers are much more powerful than traditional digital computers in terms of speed, it still obeys Church–Turing thesis. In other words, no quantum computer can solve a problem that a traditional computer cannot. Quantum computers can still be simulated by traditional computers. The difference is that a traditional computer may have to spend millions of years to solve a problem that a quantum computer can solve in seconds. Therefore, the first law of computing – information cannot increase in a computer – is still true because a quantum computer has the same power as a traditional computer in terms of computability.

Just as Boolean logic gates play a central role in traditional computers, quantum logic gates are the heart of quantum computers. Unlike Boolean logic gates, all quantum logic gates are reversible. More precisely, a quantum logic gate can be represented by a unitary matrix M (an $n \times n$ matrix M is unitary if $M^+M = I_n$ where M^+ is the conjugate transpose of M). So if M is the matrix that represents the quantum logic gate, M^+ represents the reverse operation of the gate.

While it is possible to define quantum logic gates for any number of qubits, the most common quantum logic gates operate on spaces of one or two qubits. For example, Hadamard gate and phase shifter gates are single-qubit gate and CNOT (controlled-NOT) gate is a two-qubit gate. With Hadamard gates, phase shifter gates and CNOT gates, one can do any operation that can be done by a quantum computer. We call sets of quantum gates like this by **universal quantum gates**.

Because quantum logic gates are all reversible, quantum computers do not destroy any information until outputs are measured. Therefore, quantum computers may consume less energy than traditional computers in theory.

While quantum computers have tremendous power and may consume much less energy, they are extremely difficult to be built. A quantum computer must maintain

coherence among its qubits long enough to perform an algorithm. Preventing qubits from interacting with the environment (which will cause decoherence) is extremely hard. We can get around the difficulty by giving up the entanglements. However, the quantum computer is no longer as powerful if we do so. Without entanglements, the state space of n-qubits only has $2n$ dimensions rather than 2^n dimensions. Exponentially hard or NP complete problems cannot be computed by these computers within polynomial time (assume $\mathrm{P} \neq \mathrm{NP}$) [13]. However, these computers use reversible logic gates and therefore have advantage in energy consuming.

In conclusion, even though quantum computers use a quite different model, two laws of information still apply to them.

Optical Computing

Optical computing has been studied for many decades [14]. This work has led to conclusions that seem obvious in retrospect; namely that

- Optical computing can never compete with electronic computing for general purpose computing.
- Optical computing has legitimate roles in niches defined by the characteristic that the information being processed is already in the optical domain, e. g. in optical communication or optical storage.
- Two types of complexity arise. One is the complexity of the optical effects themselves, e. g. optics often has three or more activities going on at each location, so if the processes are nonlinear (e. g. photorefractives or cross talk among active elements) then complexity factors are likely to arise [15].

This is nonequilibrium thermodynamics at work. Not to be confused with this is the savings in temporal coherence accomplished by use of space, time, and fanin complexity to consume computational complexity [16,17].

Optical computing is limited by thermodynamics in a unique sense. Let us suppose that we wish to measure an output signal that is either 1 or 0 in amplitude. Either for speed or power consumption, we seek to measure as few photons as possible to decide which of those two states (0 or 1) is present. Because the photons do not (without encouragement) come evenly spaced in time of arrival, we must detect many photons to be sure that no detected photon means a logical zero rather than a statistical extreme when a logical 1 is intended. With normal statistics, it takes about 88 detected photons to distinguish between what was intended as a logical 1 and what was intended as a zero. It is possible to make a light source with substantially uniform emission rates. This is called squeezed light. Naively, squeezing can overcome that problem to an amazing extent, but the squeezed state is very fragile and many of the central activities in an optical computer destroy the order squeezing put in before it can be detected [18].

Thermodynamically Inspired Computing

The approach of systems to equilibrium may become a model for computing. The most prominent of these are called simulated annealing, the Hopfield net, and the Boltzmann machine. Basically, they are ways to encode a quantity to be minimized as a sort of pseudo energy and then simulating some energy minimization activity. And, of course hybridization of these methods make the clean division among them impossible.

Annealing of materials is an ancient art. Domains in a metal sword, for instance, may be randomly disposed to one another, making them brittle. The goal of annealing is to get them aligned, so they behave more like a single crystal and thus hardened. This is done by a method sometimes called "heat and beat." The material is heated to a high temperature where domains changes are fairly easy and then beat with a hammer to encourage such movement. Then the material is cooled to a little lower temperature where one hopes the gains so far are not rerandomized and beat again. The art is in the cooling schedule and in the beating. In thermodynamic terms, we are trying to get the sword in its minimum energy state where no beating against the opponent will cause it to relax into a lower energy state. Unlike more familiar optimization methods such as steepest descent or Newton–Raphson that find only a local extremum (one near the starting state), simulated annealing aims at finding the global extremum independently of starting state. It does this by pure imitation of annealing. Some monotonic figure of merit for the problem is chosen. It might be the RMS difference between the output vector of a matrix-vector multiplication and a target vector. Minimizing it is analogous to minimizing the sword's energy. Driving it to or near zero by adjusting the input vector is a way of solving the ubiquitous $A\mathbf{x} = \mathbf{b}$ problem for $\mathbf{x}$. In early stages, big random steps (chosen from a thermodynamic expression governing probabilities as a function of energy and "temperature") are used. The perturbed system is then "cooled" in the sense of allowing the starting fictitious T value to decrease slight, and the process is repeated. This is very computer intensive but very effective. More information can be found in many places including [19,20,21].

A Hopfield net [22] is very similar to simulated annealing in that it uses guided stochastic methods to minimize a user-defined energy (Figure of merit). It is based on adjusting connections of a rather simple binary connected neural network by picking a starting state and adjusting one interconnection at a time.

A Boltzmann machine is a more thermodynamically inspired version of a Hopfield network [23].

One more thermodynamically-inspired means for computation needs to be mentioned, namely stochastic resonance [24]. It is a creative use of noise. Noise is usually thought of as the enemy of signal. But if in some nonlinear operation the signal is so small that it is not useful, adding a certain amount of noise to the input may help. Indeed, there is an optimal amount of noise for each case. Too little or too much noise are both bad, but there is a noise value in resonance with the problem which give optimal results.

This is not the place for detailed discussions of these thermodynamically-inspired methods. Rather, our intent is to show that thermodynamics has impacted not only physical computers but also the way those computers are used to achieve certain goals.

Cellular Array Processors

Cellular array processors (some prefer "cellular automata") are discrete binary valued elements in fixed locations or arrays [25]. Usually started with random inputs, they can be observed to produce quite interesting space-time histories as each cell is updated according to a fixed rule using its current value and those of its immediate neighbors. Under some circumstances, those space-time patterns simply terminate or generate chaos. Often, however, the system self organizes to produce highly organized patterns. One "application" is so-called "Artificial Life" which began with the immensely popular set of evolution rules for a two-dimensional array called "the Game of Life." For the few who have not played with it, we urge you to do so. An online version is available free at http://www.bitstorm.org/gameoflife/. Many brilliant computer scientists have become enthralled with it. For instance, Stephen Wolfram analyzed it in terms of thermodynamics [26].

Conclusions

Thermodynamics

- Governs all computing.
- Has inspired conservative, reversible computing.
- Is not easy to apply to non Turing computers such as analog and quantum.
- Has directly inspired some optimization methods.

In the end, computers are at least partially physical in the classical sense, so they are ultimately limited by thermodynamics. The presence of many virtual operations in some forms of quantum computing can reduce the price of physicality dramatically but cannot eliminate it [27].

Future Directions

Classical equilibrium thermodynamics places limits on what can be accomplished under different conditions. More importantly, it tells us when we are approaching a limit sate. Things often get exceedingly difficult in such cases, so effort might be better spent where there is greater room for improvement.

Optical computing may allow the reaching of the energy required for reversible computing soon. As there is no fundamental limit but negative energy is not defined, we can say that zero-energy computing is the limit. Doing the computation by interferometry of beams passing through passive optics requires no energy. At last, we will have an example of how the choice of medium can be critical. That kind of zero-energy computation is not possible for electronics. Early versions of zero-energy gates have already been described and made [28,29]. That work might not have been undertaken except for the "permission" for zero-energy systems granted by thermodynamics.

Bibliography

Primary Literature

1. Weaver W (1948) Science and complexity. Am Sci 36:536–541
2. Nicolis G, Prigogine I (1977) Self-organization in nonequilibrium systems. Wiley, New York
3. Prigogine I, Stengers I (1984) Order out of chaos. Bantam Books, New York
4. Prigogine I, Stengers I, Toffler A (1986) Order out of chaos: Man's new dialogue with nature. Flamingo, London
5. Nyce JM (1996) Guest editor's introduction. IEEE Ann Hist Comput 18:3–4
6. Orponen P (1997) A survey of continous-time computation theory. In: Du DZ, Ko KI (eds) Advances in Algorithms, Languages, and Complexity. Kluwer, Dordrecht, pp 209–224
7. Pour-El MB, Richards JI (1989) Computability in Analysis and Physics. Springer, Berlin
8. Grantham W, Amitm A (1990) Discretization chaos – Feedback control and transition to chaos. In: Control and dynamic systems, vol 34. Advances in control mechanics. Pt. 1 (A91–50601 21–63). Academic Press, San Diego, pp 205–277
9. Herbst BM, Ablowitz MJ (1989) Numerically induced chaos in the nonlinear Schrödinger equation. Phys Rev Lett 62: 2065–2068
10. Ogorzalek MJ (1997) Chaos and complexity in nonlinear electronic circuits. World Sci Ser Nonlinear Sci Ser A 22
11. Nielsen MA, Chuang IL (2000) Quantum Computation and Quantum Information. Cambridge University Press, New York

12. Bouwmeester D, Ekert A, Zeilinger A (2001) The Physics of Quantum Information. Springer, Berlin
13. Caulfield HJ, Qian L (2006) The other kind of quantum computing. Int J Unconv Comput 2(3)281–290
14. HJ Caulfield, Vikram CS, Zavalin A (2006) Optical logic redux. Opt Int J Light Electron Opt 117:199–209
15. Caulfield HJ, Kukhtarev N, Kukhtareva T, Schamschula MP, Banarjee P (1999) One, two, and three-beam optical chaos and self organization effects in photorefractive materials. Mat Res Innov 3:194–199
16. Caulfield HJ, Brasher JD, Hester CF (1991) Complexity issues in optical computing. Opt Comput Process 1:109–113
17. Caulfield HJ (1992) Space – time complexity in optical computing. Multidimens Syst Signal Process 2:373–378
18. Shamir J, Caulfield HJ, Crowe DG (1991) Role of photon statistics in energy-efficient optical computers. Appl Opt 30: 3697–3701
19. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220(4598):671–680
20. Cerny V (1985) A thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm. J Optim Theory Appl 45:41–51
21. Das A, Chakrabarti BK (eds) (2005) Quantum Annealing and Related Optimization Methods. Lecture Notes in Physics, vol 679. Springer, Heidelberg
22. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci USA 79:2554–2558
23. Hinton GE, Sejnowski TJ (1986) Learning and Relearning in Boltzmann Machines. In: Rumelhart DE, McClelland JL (eds) and the PDP Research Group. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. vol 1. Foundations. Cambridge MIT Press, Cambridge, pp 282–317
24. Benzi R, Parisi G, Sutera A, Vulpiani A (1983) A theory os stochastic resonance in climatic change. SIAM J Appl Math 43:565–578
25. Ilachinski A (2001) Cellular Automata: A Discrete Universe. World Scientific, Singapore
26. Wolfram S (1983) Statistical mechanics of cellular automata. Rev Mod Phys 55:601–644
27. Kupiec SA, Caulfield HJ (1991) Massively parallel optical PLA. Int J Opt Comput 2:49–62
28. Caulfield HJ, Soref RA, Qian L, Zavalin A, Hardy J (2007) Generalized optical logic elements – GOLEs. Opt Commun 271: 365–376
29. Caulfield HJ, Soref RA, Vikram CS (2007) Universal reconfigurable optical logic with silicon-on-insulator resonant structures. Photonics Nanostruct 5:14–20

Books and Reviews

Bennett CH (1982) The thermodynamics of computation a review. Int J Theor Phys 21:905–940
Bernard W, Callen HB (1959) Irreversible thermodynamics of nonlinear processes and noise in driven systems. Rev Mod Phys 31:1017–1044
Bub J (2002) Maxwell's Demon and the thermodynamics of computation. arXiv:quant-ph/0203017
Casti JL (1992) Reality Rules. Wiley, New York
Deutsch D (1985) Quantum Theory, the Church-Turing Principle and the Universal Quantum Computer. Proc Roy Soc Lond Ser A Math Phys Sci 400:97–117
Karplus WJ, Soroka WW (1958) Analog Methods: Computation and Simulation. McGraw Hill, New York
Leff HS, Rex AF (eds) (1990) Maxwell's Demon: Entropy, Information, Computing. Princeton University Press, Princeton
Zurek WH (1989) Algorithmic randomness and physical entropy. Phys Rev Abstr 40:4731–4751

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation

Cai-Zhuang Wang[1], Gun-Do Lee[2], Kai-Ming Ho[1]
[1] Ames Laboratory and Department of Physics and Astronomy, Iowa State University, Ames, USA
[2] School of Materials Science and Engineering and Inter-university Semiconductor Research Center (ISRC), Seoul National University, Seoul, Korea

Article Outline

Glossary
Definition of the Subject
Introduction
Tight-Binding Potentials for Carbon
TBMD Simulations of Cage and Tube Formation
Future Directions
Acknowledgments
Bibliography

Glossary

Tight-binding method A method to calculate electronic structure of a condensed matter system (solid or liquid) through modeling the interactions (overlap and hopping matrix elements) between the atomic orbitals on each atom in the system.

Molecular dynamics An atomistic simulation method for studying the structure and dynamics of a condensed system (solid or liquid) based on given interatomic potentials and on classical equations of motion (Newton or Langevin).

Tight-binding molecular dynamics Molecular dynamics simulations using interatomic potentials described by the tight-binding method.

Environment-dependent tight-binding potential A tight-binding potential model in which the interatomic hopping matrix elements between a pair of atoms in the solid or liquid phases depend not only on

the distance between the two atoms, but also on the geometry of neighboring atoms.

Stone–Wales transformation A common topological transformation in two-dimensional carbon networks (e. g., carbon nanotube, carbon fullerene). The Stone–Wales transformation proceeds by rotating a bond between two carbon atoms by 90 degree which changes the topology of the network.

Definition of the Subject

Tight-binding molecular dynamics has emerged as a useful method for studying the structural, dynamical, and electronic properties of materials. In this article, we will review the development of tight-binding potentials for carbon systems and the accuracy of the tight-binding potentials for studying nanostructures such as carbon fullerenes and carbon nanotubes. We will also review atomistic simulations using tight-binding molecular dynamics to study the nucleation and formation of carbon fullerenes and single-walled carbon nanotubes. Several formation pathways and nucleation mechanisms have been discussed including nucleation from the gas phase, formation by zipping graphitic patches, growth via coalescence, junction formation by self-healing of vacancies, and transformation from nanodiamonds.

Introduction

Atomistic modeling and simulation of materials has attracted continuous interests in physics, chemistry, and materials science. In the past four decades, first-principles simulation methods based on density functional theory and formalism [23,37] (e. g., Car–Parrinello method [5], VASP [39,40,41,42,43]) have been well developed and widely used in computer simulation studies of materials [7,8]. However, due to heavy computational workload, first-principles methods are prohibitive for studying systems with increasing complexity. Most of molecular dynamics (MD) simulations using the first-principles density functional formalism for force calculations are still limited to a small number of atoms (a few hundred atoms) and a short time period (a few hundred ps) on commonly available computers. On the other hand, quite a number of empirical interatomic potentials suitable for large scale simulations have been developed for covalent materials such as silicon and carbon [3,4,9,13,32,77,81,82,83]. These potentials include three-body interactions and are usually fitted to energy-volume phase diagrams and other properties of the crystalline structures obtained by first-principles calculations or from experiments. Although simulations with classical potentials are fast, these empirical potentials do not always give correct descriptions for structures and properties that are not explicitly included in the fitting database [1,11]. Electronic structure information cannot be obtained, nor can we expect these classical potentials to give accurate descriptions of phenomena where quantum mechanical interference effects are essential (e. g., Jahn–Teller distortions around vacancies, conjugated π-state effects in carbon systems, electronic entropy at high temperatures). There is a large class of problems in the emerging area of nano science and technology that requires more atoms than first-principles techniques can handle and demands more accuracy than classical potentials can provide. Thus it is imperative to have a scheme that is powerful enough to treat several thousand atoms (tens of thousand to million atoms with massively parallel computers) while being accurate enough so that we can trust the results.

In the past two decades, a scheme for molecular dynamics simulation based on a simplified quantum mechanical description of interatomic forces, i. e., tight-binding molecular dynamics (TBMD), has been developed [31,46,62,71,74,86,87]. This method bridges the gap between classical-potential simulations and the first-principles Car–Parrinello method. In the same spirit as the Car–Parrinello scheme [5], TBMD incorporates electronic structure into molecular dynamics simulation through a tight-binding Hamiltonian H_{TB}. Using the Hellmann–Feynman theorem, the quantum mechanical many-body nature of the interatomic forces are taken into account naturally by calculating the quantum state of the electrons at each MD time step. Since the TBMD scheme usually uses a minimal basis set for the electronic structure calculations and the Hamiltonian matrix elements are parametrized or tabulated, larger numbers of atoms can be tackled within the present computer capabilities. One of the distinctive features of this scheme in comparison with other empirical schemes is that all the parameters in the model can be obtained theoretically. It is therefore very useful for studying novel nano-materials where experimental data may not be readily available. The scheme has been demonstrated to be a powerful method for studying various structural, dynamical and electronic properties of covalent systems as well as metallic systems.

In this article, we will review the development of tight-binding models for carbon and applications of tight-binding molecular dynamics to the study of the formation of carbon nanostructures such as fullerenes and carbon nanotubes. The discovery of carbon fullerenes [38,44] and carbon nanotubes [24] have opened a new field in the science and technology of nano-scale materials. Carbon fullerenes and nanotubes have been shown to have var-

ious novel properties due to their size and dimensionality. Carbon fullerenes are zero-dimensional and carbon nanotubes are one-dimensional nano scale materials. Among many techniques used in this very dynamic research field, tight-binding molecular dynamics simulations have made notable contributions to the understanding of various aspects of the structures, formation mechanism, as well as the mechanical and electronic properties of these novel nanomaterials. In this article, we will focus ourselves on the simulation studies of fullerene and nanotube formation.

This article is organized as follows: In the next section, we will review the development of two popular tight-binding potentials for carbon, i. e., the transferable two-center tight-binding potential and the environment-dependent tight-binding potential developed by the authors and co-workers. The accuracy of the potentials for simulation studies of carbon-based materials will also be discussed. In Sect. "TBMD Simulations of Cage and Tube Formation", we will review tight-binding molecular dynamics simulations of various atomic processes of the formations of fullerenes and single-walled carbon nanotubes. Finally, concluding remarks are given in Sect. "Future Directions".

Tight-Binding Potentials for Carbon

The expression for binding energy of a system with M atoms and N valence electrons in tight-binding molecular dynamics is given by

$$E_{\text{binding}} = E_{\text{bs}} + E_{\text{rep}} \tag{1}$$

The first term on the right hand side of Eq. (1) is the band structure energy which is equal to the sum of the one-electron eigenvalues ε_i of the occupied states given by a tight-binding Hamiltonian H_{TB},

$$E_{\text{bs}} = \sum_i f_i \varepsilon_i \tag{2}$$

where f_i is the electron occupation (Fermi–Dirac) function and $\sum_i f_i = N$. The tight-binding Hamiltonian H_{TB} is constructed following the scheme proposed by Slater and Koster [75]. For systems containing s and p orbitals, such as carbon, the tight-binding Hamiltonian H_{TB} consists of four types of hopping integrals $h_{ss\sigma}, h_{sp\sigma}, h_{pp\sigma}$, and $h_{pp\pi}$ as well as two on-site atomic energies e_s and e_p. The hopping integrals are dependent on the interatomic distances and in general also on the binding environment of the structures. The on-site atomic energies are also dependent on the bonding environment of the atoms. These hopping integrals and on-site energies are treated as parameters in the tight-binding scheme. In the following subsections, we will discuss how these parameters are determined.

The second term on the right hand site of Eq. (1) is a repulsive energy usually expressed as the sum of short-ranged pairwise interactions

$$E_{\text{rep}} = \sum_{i,j} \phi(r_{i,j}) \tag{3}$$

or a functional of pairwise interactions

$$E_{\text{rep}} = \sum_i F_i\Big(\sum_j \phi(r_{i,j})\Big) \tag{4}$$

where F is a functional which, for example, can be a 4th order polynomial as will be discussed in the following subsections.

Two-Center Tight-Binding Potential

The most commonly used two-center tight-binding potential for carbon was developed by Xu, Wang, Chan, and Ho (XWCH) in the early 1990s [96] following the transferable scaling function proposed by Goodwin, Skinner, and Pettifor (GSP) for Si [16]. The dependence of the TB hopping parameters and the pairwise potential on the interatomic separation is given by:

$$h_\alpha(r) = h_\alpha(r_0)\left(\frac{r_0}{r}\right)^n \exp\left[n\left\{-\left(\frac{r}{r_c}\right)^{n_c} + \left(\frac{r_0}{r_c}\right)^{n_c}\right\}\right] \tag{5}$$

and

$$\phi(r) = \phi_0\left(\frac{d_0}{r}\right)^m \exp\left[m\left\{-\left(\frac{r}{d_c}\right)^{m_c} + \left(\frac{d_0}{d_c}\right)^{m_c}\right\}\right] \tag{6}$$

where r_0 denotes the nearest-neighbor atomic separations in diamond and $h_\alpha(r_0)$, n, n_c, r_c, ϕ_0, m, d_0, d_c, and m_c are parameters. Unlike the GSP model, the scaling parameters can be different for different hopping integrals and the pairwise repulsion term $\phi(r)$. Moreover, for the convenience of molecular-dynamics simulation, the scaling function $h(r)$ and the pair potential $\phi(r)$ are also required to go smoothly to zero at some designated cutoff distance. This was achieved by replacing the tail of $h(r)$ with a third order polynomial $t_h(r - r_1)$ whose coefficients are determined by requiring that the connection of $h(r)$ and $t_h(r)$ at r_1, (the matching point, $r_1 \le r_m$), be smooth up to the first derivative, and that $t_h(r)$ and its first derivative be zero

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Table 1
Parameters for the functions $h(r)$ and $\phi(r)$ of the XWCH tight-binding model

n	n_c	r_c(Å)	r_0(Å)	r_1(Å)	
2.0	6.5	2.18	1.536329	2.45	
ϕ_0(eV)	m	m_c	d_c(Å)	d_0(Å)	d_1(Å)
8.18555	3.30304	8.6655	2.1052	1.64	2.57

at r_m. The same procedure is used to determine $t_\phi(r-d_1)$, which replaces the tail of $\phi(r)$. The expression for the repulsive energy E_{rep} in this model takes the functional form of Eq. (4) with $F(x)$ being a 4th order polynomial in terms of $x = \sum_j \phi(r_{ij})$.

The potential is given a relatively short cutoff distance (r_m and d_m) of 2.6 Å. The parameters in the model are chosen primarily by fitting first-principles density functional calculation results of energy-volume curves of different carbon polytypes: diamond, graphite, linear chain, simple cubic, and face-centered cubic structures, with special emphasis on the diamond, graphite, and linear chain structures. Additional checks were made to ensure that the model gives reasonable results for the electronic band structure, elastic moduli, and phonon frequencies in the diamond and graphite structures, although these properties do not enter explicitly into the fitting procedure.

The resulting sp^3 tight-binding parameters were: $E_s = -2.99\,\text{eV}$, $E_p = 3.71\,\text{eV}$, $h_{ss\sigma}(r_0) = -5.0\,\text{eV}$, $h_{sp\sigma}(r_0) = 4.7\,\text{eV}$, $h_{pp\sigma}(r_0) = 5.5\,\text{eV}$, and $h_{pp\pi}(r_0) = -1.55\,\text{eV}$. The scaling parameters for $h(r)$ and $\phi(r)$, the coefficients for the tail functions $t_h(r-r_1)$ and $t_\phi(r-d_1)$, and the coefficients for the repulsive energy polynomial function $F(x) = \sum_{n=0}^{4} c_n x^n$, with $x = \sum_j \phi(r_{ij})$ are given in Table 1 and Table 2 respectively.

To correctly handle the effects of charge transfer in disordered systems, particularly in the presence of dangling bonds, a Hubbard-like term,

$$H_u = \frac{1}{2} u \left(q_i - q_i^0\right)^2 \tag{7}$$

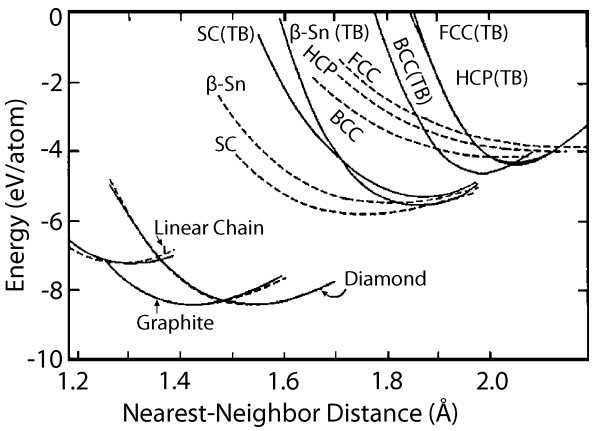

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 1
Binding energy versus nearest-neighbor atomic distance for carbon in different crystalline structures calculated using the XWCH tight-binding potential (*solid lines*) are compared with the results from first-principles density functional (LDA) calculations (*dashed lines*). (From Ref. [96])

was added to the tight-binding Hamiltonian H_{TB}, where q_i is the Mulliken population at atomic site i and q_i^0 is the number of valence electrons of atom i. The parameter u is taken to be 4 eV for carbon.

The XWCH carbon potential has been shown to have good transferability when applied to a variety of crystal structures with low coordinations (i. e., carbon chain, graphene, and diamond). This can be seen from Fig. 1 and Tables 3 and 4 where the energies, vibrational and elastic properties for different coordinated crystalline structures obtained from this model are compared with first-principles calculations and experimental data. Description of higher-coordinated metallic structures is only qualitative with this potential. Applications in the molecular-dynamics study of liquid and amorphous phases of carbon [64,89,90,91,92] indicated that the potential is fairly good for describing lower-coordinated carbon systems over a wide range of bonding environments.

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Table 2
Coefficients of the polynomial functions $t_h(r-r_1)$, $t_\phi(r-d_1)$, and $F(x)$ in the XWCH TB model. See also the text for details

	$t_s(r-r_1)$	$t_\phi(r-d_1)$	$F(x)$
c_0	$6.7392620074314 \times 10^{-3}$	$2.2504290109 \times 10^{-8}$	-2.5909765118191
c_1	$-8.1885359517898 \times 10^{-2}$	$-1.4408640561 \times 10^{-6}$	0.5721151498619
c_2	0.1932365259144	$2.10433033744 \times 10^{-5}$	$-1.7896349903996 \times 10^{-3}$
c_3	0.3542874332380	$6.60243902262 \times 10^{-5}$	$2.3539221516757 \times 10^{-5}$
c_4			$-1.24251169551587 \times 10^{-7}$

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Table 3

Elastic constants, phonon frequencies and Grüneisen parameters of diamond calculated from the XWCH-TB model [96] and the environment-dependent TB (EDTB) model [78] are compared with experimental results [59,60]. Elastic constants are in units of 10^{12}dyn/cm^2 and the phonon frequencies are in terahertz

	XWCH	EDTB	Experiment
a(Å)	3.555	3.585	3.567
B	4.56	4.19	4.42
$c_{11} - c_{12}$	6.22	9.25	9.51
c_{44}	4.75	5.55	5.76
$\nu_{LTO(\Gamma)}$	37.80	41.61	39.90
$\nu_{TA(X)}$	22.42	25.73	24.20
$\nu_{TO(X)}$	33.75	32.60	32.0
$\nu_{LA(X)}$	34.75	36.16	35.5
$\gamma_{LTO(\Gamma)}$	1.03	0.93	0.96
$\gamma_{TA(X)}$	−0.16	0.30	
$\gamma_{TO(X)}$	1.10	1.50	
$\gamma_{LA(X)}$	0.62	0.98	

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Table 4

Elastic constants, phonon frequencies and Grüneisen parameters of graphite calculated from the XWCH-TB model [96] and the environment-dependent TB (EDTB) model [78] are compared with experimental results [14]. Elastic constants are in units of 10^{12}dyn/cm^2 and the phonon frequencies are in terahertz

	XWCH	EDTB	Experiment
$c_{11} - c_{12}$	8.40	8.94	8.80
E_{2g_2}	49.92	48.99	47.46
A_{2u}	29.19	26.07	26.04
$\gamma(E_{2g_2})$	2.00	1.73	1.63
$\gamma(A_{2u})$	0.10	0.05	

The reliability of this potential for fullerene calculations has been further tested by studying the ground-state geometries and energies of carbon fullerenes. The fully relaxed C_{60} molecule obtained by the XWCH tight-binding potential [88] has icosahedral symmetry and bond lengths of 1.40 Å and 1.46 Å respectively for the double and single bonds, which agree very well with the first-principles density functional calculation results of 1.40 Å and 1.45 Å. The tight-binding calculation for C_{60} yields a cohesive energy of 0.41 eV per atom relative to graphite and HOMO–LUMO energy separation of 1.61 eV, agree very well with the LDA results of 0.40 eV per atom and 1.71 eV respectively. The potential also describes well the vibrational modes of C_{60} and C_{70} fullerenes [88]. Furthermore, the potential has been employed successfully to determine the ground state geometries of every even-numbered carbon

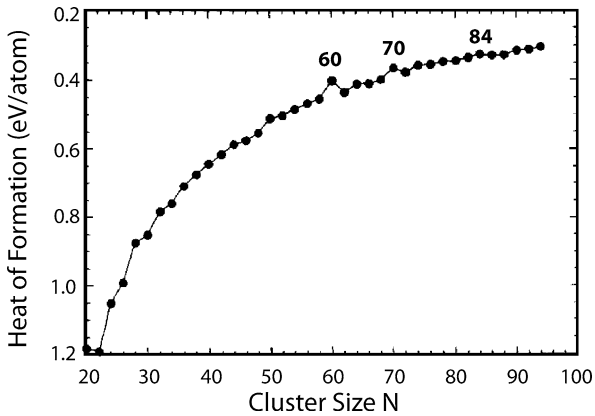

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 2

Heat of formation of carbon fullerenes relative to that of graphite as a function of fullerene size obtained from the TBMD calculations. (From Ref. [102])

fullerenes ranging from C_{20} to C_{102} [99,100,101,102]. The heats of formation as a function of the cluster size obtained from the calculations are plotted in Fig. 2. While the energies of small fullerenes grow rapidly as the cluster size increases, the energies of large fullerenes which have no adjacent pentagons are found to increase at a slower rate. The most interesting feature as shown in Fig. 2 is that C_{60}, C_{70}, and C_{84}, which correspond to "magic" numbers in the abundance peaks in the carbon cluster beam mass spectrum, are energetically more favorable than their neighbors. Typical accuracy of the XWCH tight-binding potential in describing the energetics of large fullerenes can also be seen from Table 5 where the relative energies of some C_{78}, C_{82}, and C_{84} isomers obtained from the tight-binding calculations are compared with the results from ab-initio calculations using the density functional method within the local density approximation (LDA). Although the energy ordering obtained by tight-binding calculations are not in perfect agreement with the results of LDA calculations (see the case of C_{82}), tight-binding calculations predict the lower-energy isomers correctly.

The XWCH tight-binding potential has also been demonstrated to have good accuracy for single-walled carbon nanotube calculations. Ozaki et al. have shown that the Young's moduli of the zigzag (17,0) and armchair (10,10) nanotubes are 988 and 973 Gpa, respectively, as calculated by this potential [69]. These values are very close to the first-principles results of 1 TPa [67]. The formation energy and activation energy of the topological pentagon-heptagon-heptagon-pentagon (5-7-7-5) defect in a single-walled carbon nanotube under tensile strain

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Table 5
The relative energies of some C_{78}, C_{82}, and C_{84} fullerene isomers obtained by the XWCH tight-binding potential are compared with the results from the LDA calculations. The energies are in the unit of eV/molecule

Cluster Size	Isomer Symmetry	ΔE (TB)	ΔE (LDA)
C_{78}			
	$C_{2v'}$	0.000	0.000
	$D_{3h'}$	0.087	0.160
	C_{2v}	0.284	0.210
	D_3	0.324	0.361
	D_{3h}	0.913	1.200
C_{82}			
	C_2	0.000	0.000
	C_s	0.076	0.123
	$C_{2'}$	0.166	0.312
	$C_{2''}$	0.191	0.418
	$C_{s'}$	0.226	0.385
	$C_{s'}$	0.229	0.467
	C_{2v}	0.285	0.656
	C_{3v}	0.566	1.287
	$C_{3v'}$	0.731	1.025
C_{84}			
	D_2-22	0.000	0.000
	D_{2d}-23	0.033	0.025
	C_2-11	0.277	0.370
	D_2-1	1.915	2.150

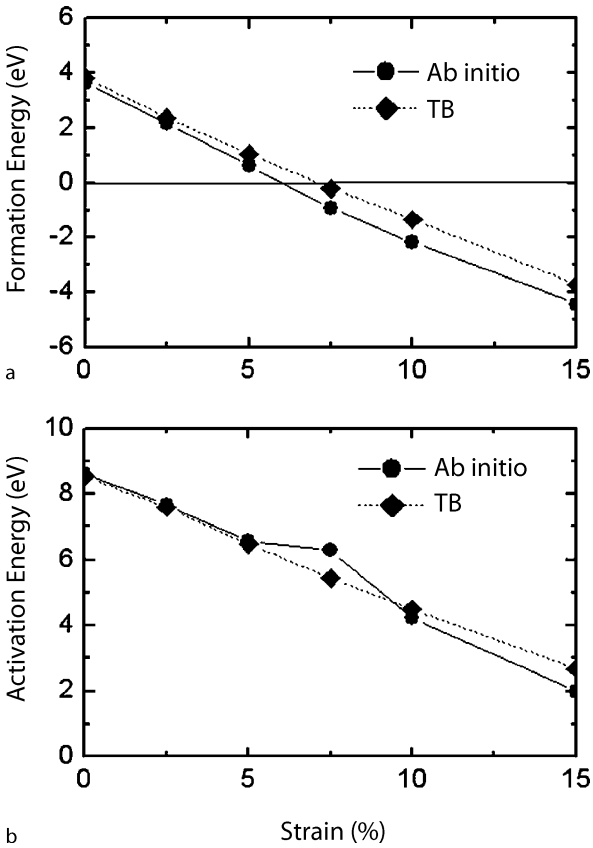

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 3
Comparison between ab initio and tight-binding (XWCH) results on the strain-dependent of a formation energies and b activation energies of a (5-7-7-5) topological defect in a (5,5) carbon nanotube. Ab initio results are in *solid circles* and tight-binding results are in *solid diamonds*. (From Ref. [104])

have been studied by Zhao et al. [104] using the tight-binding potential and first-principles calculations. Fig. 3 shows that the results from the tight-binding calculations are in very good agreement with the first-principles results. It should be noted that the potential has very short cutoff distance of 2.6 Å, which makes it inaccurate for describing the interaction between graphite layers. Therefore, the potential is not suitable for studying multi-walled carbon nanotubes.

Environment-Dependent Tight-Binding Potential

Although the XWCH potential is accurate for the structures and properties of lower-coordinated covalent carbon systems, it fails to give good descriptions for the higher-coordinated carbon structures. We noted that the accuracy and transferability of the XWCH model are limited by the approximations inherent in the Slater and Koster formulation of the tight-binding theory. One severe approximation is the assumption of a fixed minimal basis set independent of the bonding environment of the atom. Experience from first principles calculations showed that a fixed minimal basis set optimized for a given atomic geometry usually will not give accurate results for total energies when the atomic geometry of the system is changed. Minimal basis sets need to have the flexibility to deform according to the bonding environment of the atom on which they are based so that they can describe accurately the electronic bonding behavior for different atomic structures. Another major limitation in the Slater and Koster theory is the use of the two-center approximation. Such an approximation is more justified when the system are strongly covalent bonded (e. g., lower-coordinated structures of carbon). The approximation becomes poorer for metallic systems. Contributions beyond pairwise interactions need to be included. Furthermore, the Lowdin procedure used to construct the orthogonal basis set may also result in additional structure-dependent contributions to the two-

center hopping integrals because the overlap matrices are different for different structures. Molecular dynamics simulation of complex carbon structures (e. g., liquid, amorphous, clusters, surfaces) requires that tight-binding potential models must be able to describe the system accurately in various coordination and bonding environments.

A tight-binding potential for carbon that include the environment dependence for off-diagonal as well as diagonal matrix elements have been developed by Tang, Wang, Chan, and Ho [78]. An orthogonal sp^3 basis set is used and the hopping parameters and the pairwise repulsive potential are expressed as

$$h(r_{ij}) = \alpha_1 R_{ij}^{-\alpha_2} \exp[-\alpha_3 R_{ij}^{\alpha_4}](1 - S_{ij}) \tag{8}$$

In this expression, $h(r_{ij})$ denotes the possible types of interatomic hopping parameters $h_{ss\sigma}$, $h_{sp\sigma}$, $h_{pp\sigma}$, $h_{pp\pi}$, and the pairwise repulsive potential $\phi(r_{ij})$ between atoms i and j. r_{ij} is the real distance and R_{ij} is a scaled distance between atoms i and j. S_{ij} is a screening function. The parameters α_1, α_2, α_3, and α_4 and parameters for the bond-length scaling function R_{ij} and the screening function S_{ij} can be different for different hopping parameters and the pairwise repulsive potential. Note that expression Eq. (8) reduces to the the two-center form of the previous subsection if we set $R_{ij} = r_{ij}$ and S_{ij}=0.

The screening function S_{ij} is expressed as a hyperbolic tangent (tanh) function with argument ξ_{ij} given by

$$\xi_{ij} = \beta_1 \sum_l \exp[-\beta_2 (\frac{r_{il} + r_{jl}}{r_{ij}})^{\beta_3}] \tag{9}$$

where β_1, β_2, and β_3 are adjustable parameters. Note that ξ_{ij} depends not only on the distance between atoms i and j, but also on the positions of the neighbors of atoms i and j. Maximum screening effect occurs when the atom l is sitting on the line connecting the atoms i and j (i. e., $r_{il} + r_{lj}$ is minimum). This approach allows us to distinguish between the first and further neighbor interactions without explicit specification. This is well-suited for molecular dynamics simulations when one encounters configurations where it may be difficult to define exactly which atoms are first neighbors and which atoms are second neighbors.

The bond-length scaling function scales the distance between two atoms according to their effective coordination numbers. Longer effective bond lengths are assumed for higher coordinated atom pairs. The strength of the hopping parameters between atoms i and j are therefore dependent on the coordination number of the atoms: weaker interaction strength for higher-coordinated structures. The scaling between the real and effective interatomic distance is given by

$$R_{ij} = r_{ij}(1 + \delta\Delta) \tag{10}$$

where $\Delta = \frac{1}{2}[(\frac{n_i - n_0}{n_0}) + (\frac{n_j - n_0}{n_0})]$ is the fractional coordination number relative to the coordination number of the diamond structure n_0, averaging between the coordination numbers n_i and n_j of atoms i and j.

The coordination number is modeled by a smooth function, $n_i = \sum_j (1 - S_{ij})$ with S_{ij} has the form of the screening function described above. By choosing the parameters β_1, β_2 and β_3 to be 2.0, 0.0478, and 7.16 respectively, the coordination number calculated from this smooth function is 2.08639, 3.17678, 4.41022, 6.23620, 10.38529, and 11.89829 for the linear-chain, graphite, diamond, simple cubic, bcc, and fcc structures, respectively. These values give a reasonable representation of the effective coordinations in these structures.

Besides the hopping parameters, the diagonal matrix elements are also dependent on the bonding environments. The expression for the diagonal matrix elements is

$$e_{\lambda,i} = e_{\lambda,0} + \sum_j \Delta e_\lambda(r_{ij}) \tag{11}$$

where $\Delta e_\lambda(r_{ij})$ takes the same expression as Eq. (8), λ denotes the two types of orbitals (s or p). $e_{s,0}$ and $e_{p,0}$ are the on-site energies of a free atom which were chosen to be −6.041 and 1.024 eV, respectively.

Finally, the repulsive energy term E_{rep} is expressed in a form similar to the XWCH model [96] discussed above.

The parameters in the model are determined by first fitting to the self-consistent first-principles density functional calculations results of electronic band structures and then the cohesive energy versus volume curves of diamond, β-tin, simple cubic, bcc, and fcc structures respectively. These parameters are listed in Tables 6 and 7. The cutoff distance for the interatomic interaction in this potential is 3.3 Å.

As shown in Fig. 4, the environment-dependent tight-binding (EDTB) potential model describes very well not only the covalent structures, but also the higher-coordinated metallic structures. The model also describes the electronic band structures of various carbon crystals with a reasonable accuracy as one can see from Fig. 5. The elastic constants and phonon frequencies in the diamond and graphite structures from the environment-dependent tight-binding model are also improved over the XWCH two-center model as one can see from Tables 3 and 4.

The EDTB carbon potential has been applied successfully to the studies of structures and properties of amorphous carbon [18,19,72], surfaces [18,93], nanodia-

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Table 6
The parameters of the EDTB model for carbon. The TB hopping integrals are in the unit of eV and the interatomic distances are in the unit of Å. ϕ is dimensionless

	α_1	α_2	α_3	α_4	β_1	β_2	β_3	δ
$V_{ss\sigma}$	−8.9491	0.8910	0.1580	2.7008	2.0200	0.2274	4.7940	0.0310
$V_{sp\sigma}$	8.3183	0.6170	0.1654	2.4692	1.3000	0.2274	4.7940	0.0310
$V_{pp\sigma}$	11.7955	0.7620	0.1624	2.3509	1.0400	0.2274	4.7940	0.0310
$V_{pp\pi}$	−5.4860	1.2785	0.1383	3.4490	0.2000	8.5000	4.3800	0.0310
ϕ	30.0000	3.4905	0.00423	6.1270	1.5035	0.205325	4.1625	0.002168
Δe_s, Δe_p	0.1995275	0.029681	0.19667	2.2423	0.055034	0.10143	3.09355	0.272375

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Table 7
The coefficients (in unit of eV) of the polynomial function $F(x)$

c_0	c_1	c_2	c_3	c_4
12.201499972	0.583770664	$0.336418901 \times 10^{-3}$	$-0.5334093735 \times 10^{-4}$	$0.7650717197 \times 10^{-6}$

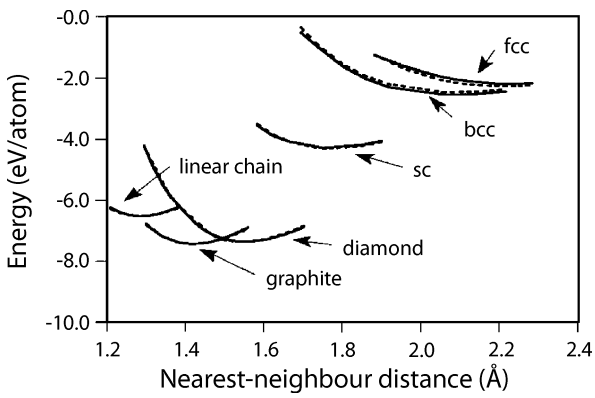

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 4
Binding energy versus nearest-neighbor atomic distance for carbon in different crystalline structures calculated using the environment-dependent tight-binding potential (*solid lines*) are compared with the results from first-principles density functional (GGA) calculations (*dashed lines*). (From Ref. [78])

mond [52,72], as well as nanotubes [34,35,67]. In particular, Wang and Ho have applied the potential to simulate the atomic process of the structural transformation at diamond surfaces under intense ultrafast laser irradiation [93]. Interesting insights have been gained from the simulation study which shows that the transformation pathways from the diamond surface into graphite sheets are different depending on whether the driving force of the transformation is "thermal" or "non-thermal". Such different behaviors are illustrated in Figs. 6 and 7. Very recently, Lee et al. [52] also used the EDTB to study the atomic process of the transformation of nanodiamond to tube-like fullerene cages and formation of carbon nanotube junctions. This simulation will be discussed in more detail later in this review.

TBMD Simulations of Cage and Tube Formation

Although carbon fullerenes and carbon nanotubes of various sizes, lengths, topologies and chiralities have been observed in experiments, the nucleation mechanism and formation process of such remarkable nano-scale materials are still not well understood. On the other hand, it has been well demonstrated that the mechanical and electronic properties of these nanomaterials are sensitive to the size and chirality of the structures. For example, it is well known that single-walled carbon nanotube can either be metallic or semiconducting depending on the size and helicity of the tubes [20,63,66,73,95]. It has also been shown that the band-gap of the nanotubes can be modified by elastic deformation of the tubes [33,54,67,70,97] and by the introduction of topological defects that make metal-semiconductor or semiconductor-semiconductor junctions [10,12]. Recently, "X", "Y", and "T"-shaped molecular junctions between the single-walled carbon nanotubes (SWCNT) have been created experimentally by electron beam irradiation of crossed SWCNTs [79]. These junctions of SWCNT could act as nano scale multiterminal electronic devices. Understanding the formation process of the carbon nanostructures at the atomic-scale is therefore highly desirable and is important for the structural control and mass production of these nanomaterials. In this section, we will review tight-binding molecular dynamics simulation studies that investigate several pathways to control catalyst-free growth of these nanoscale materials.

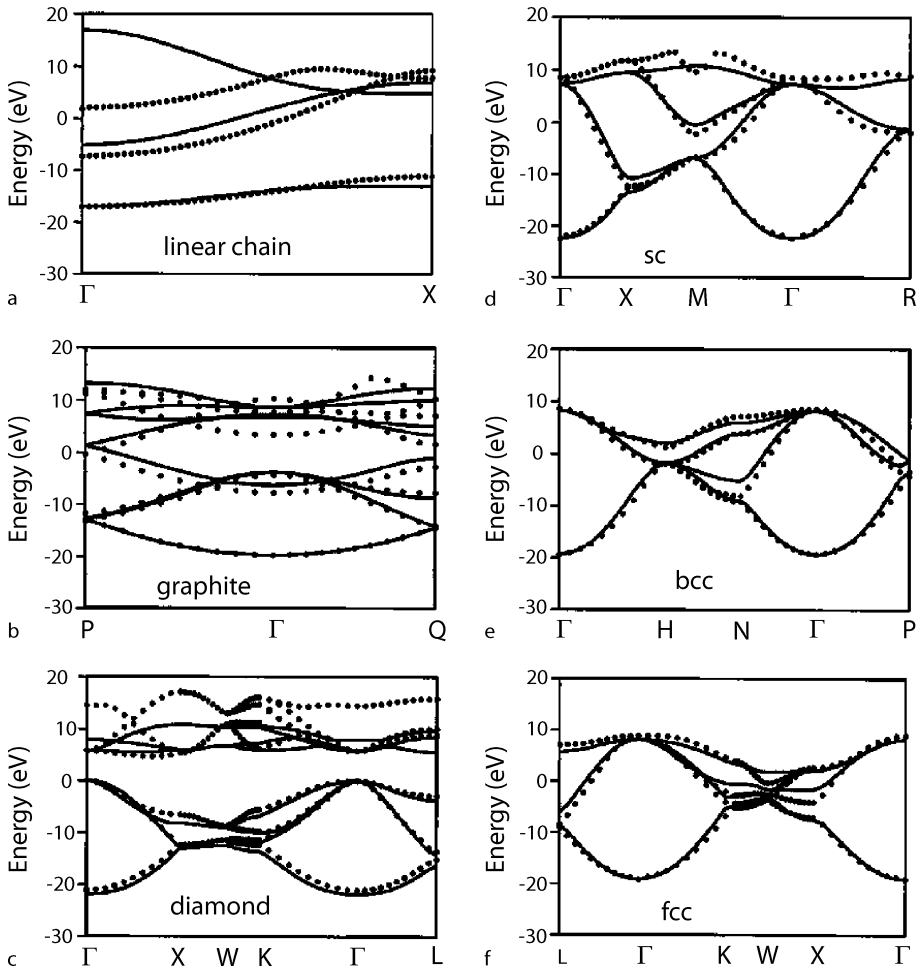

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 5
The electronic energy bands of various crystalline structures of carbon calculated using the environment-dependent tight-binding potential (*solid curves*) are compared with the first-principles LDA calculation results (*dots*). The Fermi levels are located at E = 0 eV. (From Ref. [78])

Formation by Gas Phase Condensation

In order to understand the nucleation mechanism and growth condition of C_{60} buckyball, Wang et al. have performed molecular dynamics simulations using the XWCH carbon tight-binding potential to study the formation of C_{60} from gaseous carbon atoms [94]. They enclosed 60 carbon atoms in a hollow sphere of radius R with specular reflection imposed when the atoms hit the inner surface of the sphere. They start the simulation by heating the carbon atoms to very high temperatures (10000 K) in a larger sphere ($R = 9.22$Å). The carbon atoms under this condition are found to be gas like. Then they gradually reduce the temperature and the radius of the sphere. When the temperature is reduced to 6000 K within a sphere of radius 5.3 Å, they found that polygonal rings nucleate rather rapidly from the originally loose linear-chained cluster (Fig. 8a–c). By reducing the sphere radius gradually from 5.3 to 3.832 Å while keeping the temperature at 6000 K, they found that a close cage structure is forming in the process of simulation as one can see from Fig. 8d–f. The structure of the C_{60} cages obtained from this simulation are similar to that of buckyball but with many defects due to the rapidity of the compression and cooling in the simulation. These defects include seven- and four-membered rings and adjacent pentagons. In a later study, Laszlo has performed TBMD simulation to investigate the formation of C_{60} in helium atmosphere using the same XWCH carbon tight-binding potential and obtained similar results [48]. Action-derived molecular dy-

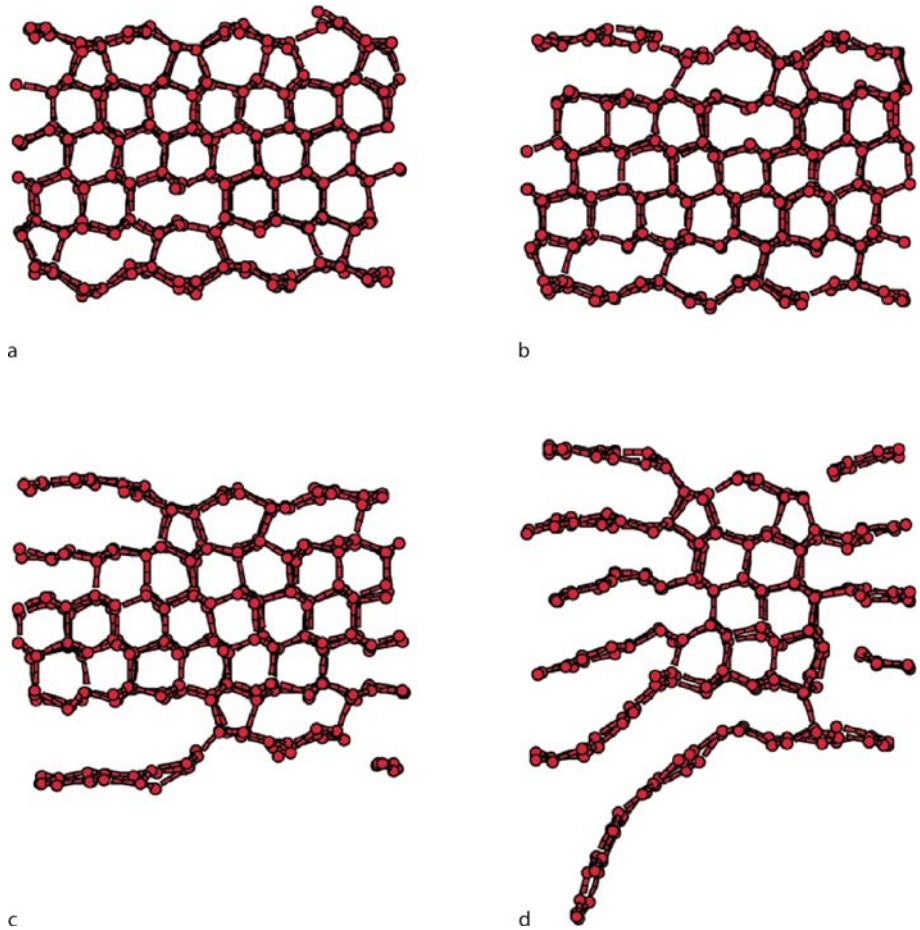

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 6
Graphitization of the diamond (1 1 1) surface via the thermal process. The snapshot pictures are taken from the tight-binding molecular dynamics simulation in which the electrons and the ions are thermal equilibrated at 2700 K. The plots show the side view of the simulation unit cell which is a 12-layer slab with two (1 1 1) surfaces (the top and bottom layers). Periodic boundary conditions are imposed in the plans perpendicular to the surface. Graphitization is found to occur through the formation of graphite-diamond interfaces (see d). The whole process takes about 3 ps. (From Ref. [93])

namics simulation using the XWCH carbon tight-binding potential suggested that existence of chains in the precursor models of tangled polycyclics and open cages is beneficially for the formation of C_{60} molecule [28]. The formation mechanism of C_{60} by C_2 assembly has also been investigated extensively by Irle et al. using density functional tight binding (DFTB) and QM/MD simulations [76,78,105].

Tight-binding molecular dynamics simulations were also performed by Oh and Lee [68] to understand how the carbon atoms from the gas phase are adsorbed on the carbon nanotube edge to either grow further or close up the tube by forming a dome-like cap. They start the simulations with a carbon nanotube that is capped at the bottom but open at the top. A number of carbon atoms in a gas phase are initially placed above the open end of the tube. The gaseous carbon atoms are confined to stay within a sphere of diameter 14 Å. The center of the confining sphere is located at the axis of the tube and is about 7 Å above the open tube edge. The carbon atoms are randomly placed in the sphere and the simulations were performed at a temperature of 2000 K. The simulations showed that once a pentagon is formed at the open edge of the tube by the deposited carbon atoms, the top of the tube closes rapidly. Some of the simulation results can be seen from Fig. 9. Irle et al. have also investigated the formation of carbon cages from the tube precursor using DFTB molecular dynamics simulations [25,106].

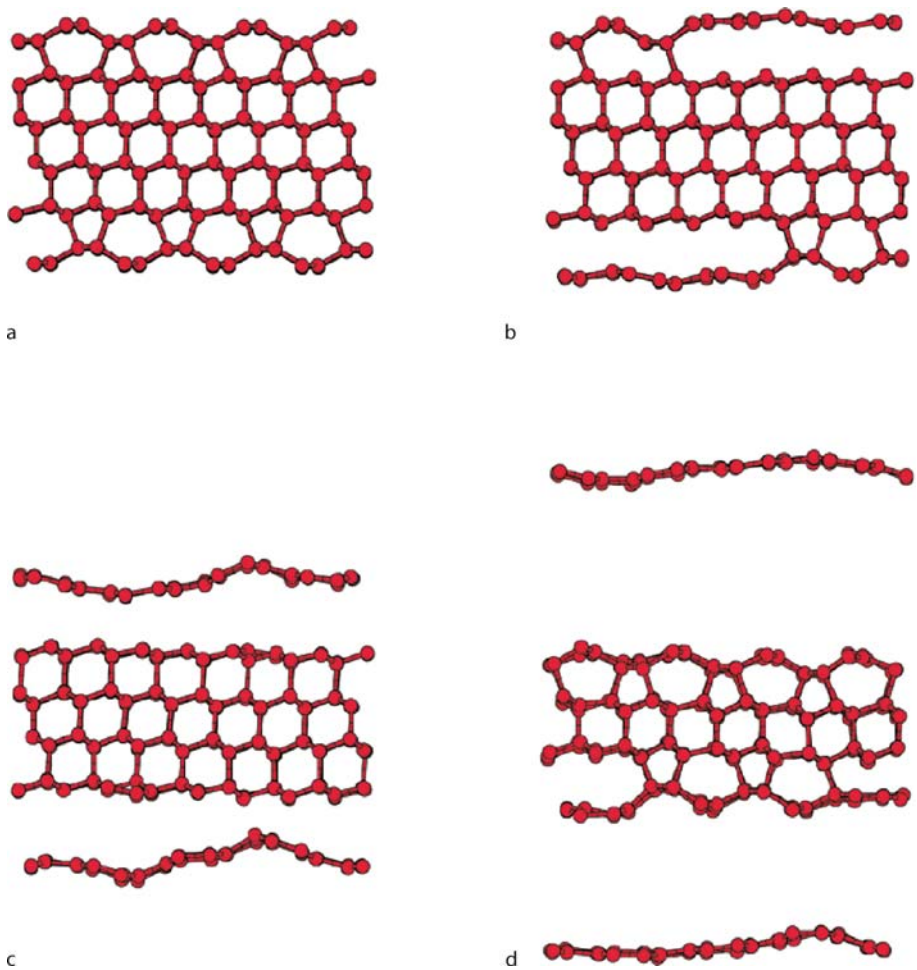

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 7
Graphitization of the diamond (1 1 1) surface due to the effects of hot electron plasma (no-thermal process). The snapshot pictures are taken from the present tight-binding molecular dynamics simulation in which the electronic temperature is raised to 15,000 K and the ions are evolved freely. The orientation of the simulation unit cell is the same as specified in Fig. 6. Note that the graphitization takes place in a layer-by-layer fashion. The slab is graphitized completely within 500 fs of simulation time. (From Ref. [93])

Nucleation from Graphitic Patches

Carbon nanotubes can be viewed as graphitic sheets with hexagonal lattice that have been wrapped up into a seamless cylinder. Although the energy barrier for rolling a big piece of graphitic sheet into a tube is very high, two small graphitic patches that zip along their edges and open up into a tubular segment could be energetically favorable and kinetically possible. Such tube-like rings can serve as a nucleus for further growth of nanotubes. Using the XWCH potential, Zhang and Crespi [103] have performed tight-binding molecular dynamics simulations on double-layered graphitic patches as illustrated in Fig. 10 and showed that a nucleus of nanotube can be formed from the simple planar patches once their edges are connected to each other by additional bridging carbon atoms. Their simulations suggested that there is no energy barrier for nucleation of a nanotube of diameter less than 23 Å (see Fig. 11). For a tube of diameter of 34 Å, the energy barrier for the tube formation would be large. Therefore, nucleation of a large tube is energetically disfavored via the graphitic patches connection mechanism. Kawai et al. [29] have studied the graphitic patch mechanism in great detail by performing extensive tight-binding molecular dynamics simulations with different initial stacking of graphitic patches and with different temperatures and velocities for collision between the graphitic patches. From hundreds of independent simulations, they found that the nuclei of

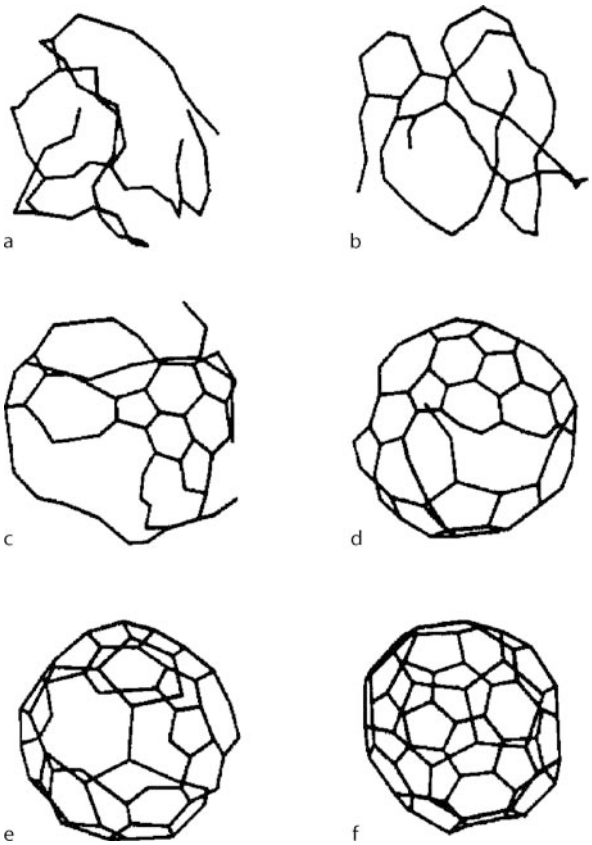

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 8
Perspective view of C_{60} cage formation process at T = 6000 K. The carbon atoms are connected by *straight lines* when the interatomic distances are less than 1.8 Å. a, b, and c are snapshots at 2.8ps, 4.2ps, and 5.6ps respectively and with the sphere radius R = 5.32 Å. d, e, and f are typical snapshot when the sphere radius is reduced to 4.61 Å, 3.90 Å, 3.832 Å respectively. Note a closed cage forms when R = 3.832 Å. The total simulation time is around 30ps. (From Ref. [94])

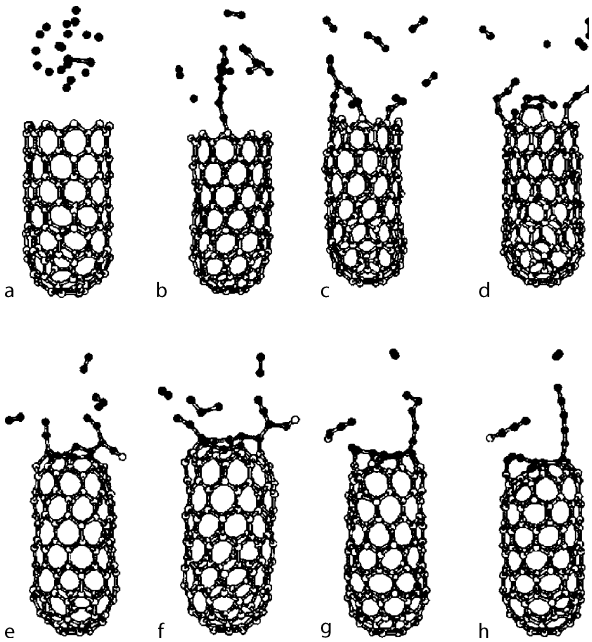

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 9
Typical snapshot (side view) of gaseous carbon adsorption on the open edge of a zigzag nanotube during the TBMD simulations at T=2000k. a initial configuration t=0, b t = 0.84 ps, c t = 1.96 ps, d t = 2.52 ps, e t = 3.64 ps, f t = 4.2 ps, g optimized structure at T=0K from the simulation geometry at 5.6 ps. The bonds are drawn when the distance is within 1.9 Å. (From Ref. [68])

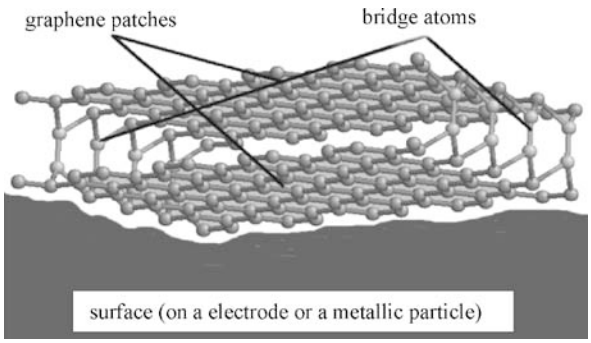

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 10
A schematic diagram of the nucleation model in which a nanotube forms via edge-mediated opening of a double-layered graphitic patch. (From Ref. [103])

nanotubes, nanohorns, and nanocages can be generated depending on the initial stacking manners and interlayer distances, temperatures, as well as the initial velocities for the collision. Some of their simulation results are shown in Fig. 12.

Growth via Coalescence

Another possible pathway for forming large fullerene cages and nanotubes is due to coalescence of fullerene cages. In 1992, Zhang et al. [98] performed tight-binding molecular dynamics to study the collision of two C_{60} buckyballs. They showed that with a proper choice of initial velocities, two C_{60} buckyball can be fused together to form a elongated C_{120} fullerene cage. Coalescence of C_{60} buckyballs to form nanotubes have been recently observed in experiments. Luzzi and Smith [58] showed that electron-beam irradiation on an array of

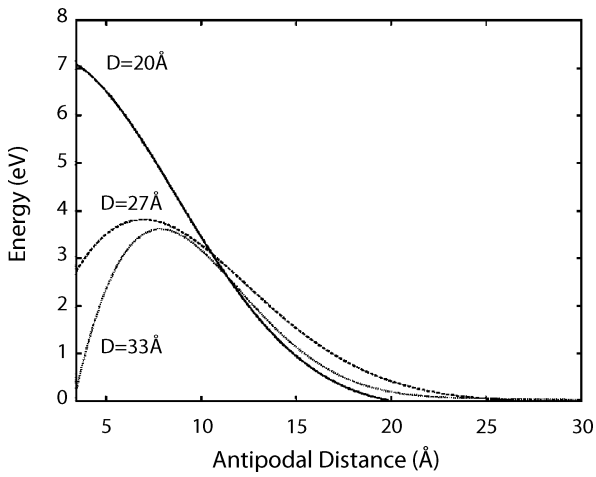

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 11

The energy as a function of a reaction coordinate (i. e., the distance between opposite inner surface as the tube nucleus opens) for nucleation graphitic patches of varying diameter when opened and of length 1 nm. The smallest patches (D = 20 Å) have no barrier to opening. (From Ref. [103])

fullerene molecules inside a single-walled carbon nanotube [76] called nanopeapods causes the coalescence of the fullerene to make a carbon nanotube inside the bigger tube. Iijima and co-workers baked nanopeapods up to 1473 K for 14 hours. They also found that the inside fullerenes start to coalesce at 1073 K and complete transformation to a single-wall nanotube at 1473 K [2]. Kim et al. have recently performed tight-binding molecular dynamics to study the dynamics of fullerene coalescence. They showed that the dynamical process of forming a C_{120} nanocapsule with various chiral indexes of (5,5), (10,0), (6,6), and (12,0) through coalescence of two C_{60} buckyball as shown in Fig. 13 consists of successive Stone–Wales-type bond rotations with a transition energy barrier of about 8–12 eV [36]. Hemandez et al. have also performed TBMD simulation using the XWCH potential to study fullerene coalescence inside a (10, 10) nanotube (diameter ca. 1.38 nm) at 1500 K [22]. Their study have demonstrated that fullerene molecules can easily coalesce inside SWNTs via a process driven by either thermal annealing or electron irradiation. The resulting structure consists of a corrugated tubule nested inside the original SWNT.

The coalescence of two smaller nanotubes to make a bigger nanotube has also been observed to occur by Nikolaev et al. and Terrones et al. when irradiated by electrons or ions at elevated temperatures [65,80]. The coalescence was considered to be induced by vacancies or other defects. Terrones et al. have performed tight-binding molecular dynamics simulations to study the coalescence and showed that two (10,10) nanotubes coalesce to make a (20,20) nanotube when defects are introduced [80]. The simulation was performed at 1273 K in order to accelerate the creation of interlinks and surface reconstruction between the two smaller tubes. The total simulation time is 150 ps and the coalescence of the two smaller tubes was observed to occur after 100 ps via the "zipping" mechanism. They also suggested that coalescence would not occur if the two nanotubes had the same chirality, because of the required global rearrangement of the bonding network. Recently, Kawai et al. [30] performed tight-binding molecular dynamics to investigate the coalescence of small carbon nanotubes, that is, combinations of the armchair (3,3), zigzag (5,0), and chiral (4,2) nanotubes. They found that two small nanotubes having the same or different chirality can coalesce without initially introducing atomic defects to enhance the reaction. They also found that the chiral index of the coalesced nanotube can be expressed as a vector sum of the indexes of the original nanotubes. These simulation results also suggested that the chirality of nanotube can be changed through chemical reaction. The simulations were performed with initial temperature ranging from 500–2500 K but a thermostat was not used during the simulation. Corresponding kinetic energy of translational velocity (center-of-mass velocity) ranged from 0.09 to 0.52 eV/atom is introduced to facilitate the collision and coalescence of the small tubes. Some of their simulations on the coalescence process are shown in Fig. 14.

Local coalescence between two nanotubes is believed to play an important role in making nanotube junctions. It has been shown by experiment that "X", "Y", and "T" nanotube junctions can be created by controlled electron beam exposure of crossed tubes at elevated temperatures [79]. Some of these junctions are shown in Fig. 15. Terrones et al. have performed tight-binding molecular dynamics simulation to study the formation mechanism of the junctions [79]. They showed that vacancies and interstitials created under electron beam exposure can trigger the local coalescence of the tubes by forming the various junctions between the two nanotubes. The simulation was performed at 1273 K with two (8,8) nanotubes crossing each other perpendicularly. In order to include the irradiation effects, 20 atoms were removed randomly from the nanotubes to create vacancies in the crossing region. The coalescence process during the simulation is shown in Fig. 16. Similar simulations were also performed by Menon et al. using a different tight binding carbon potential [61].

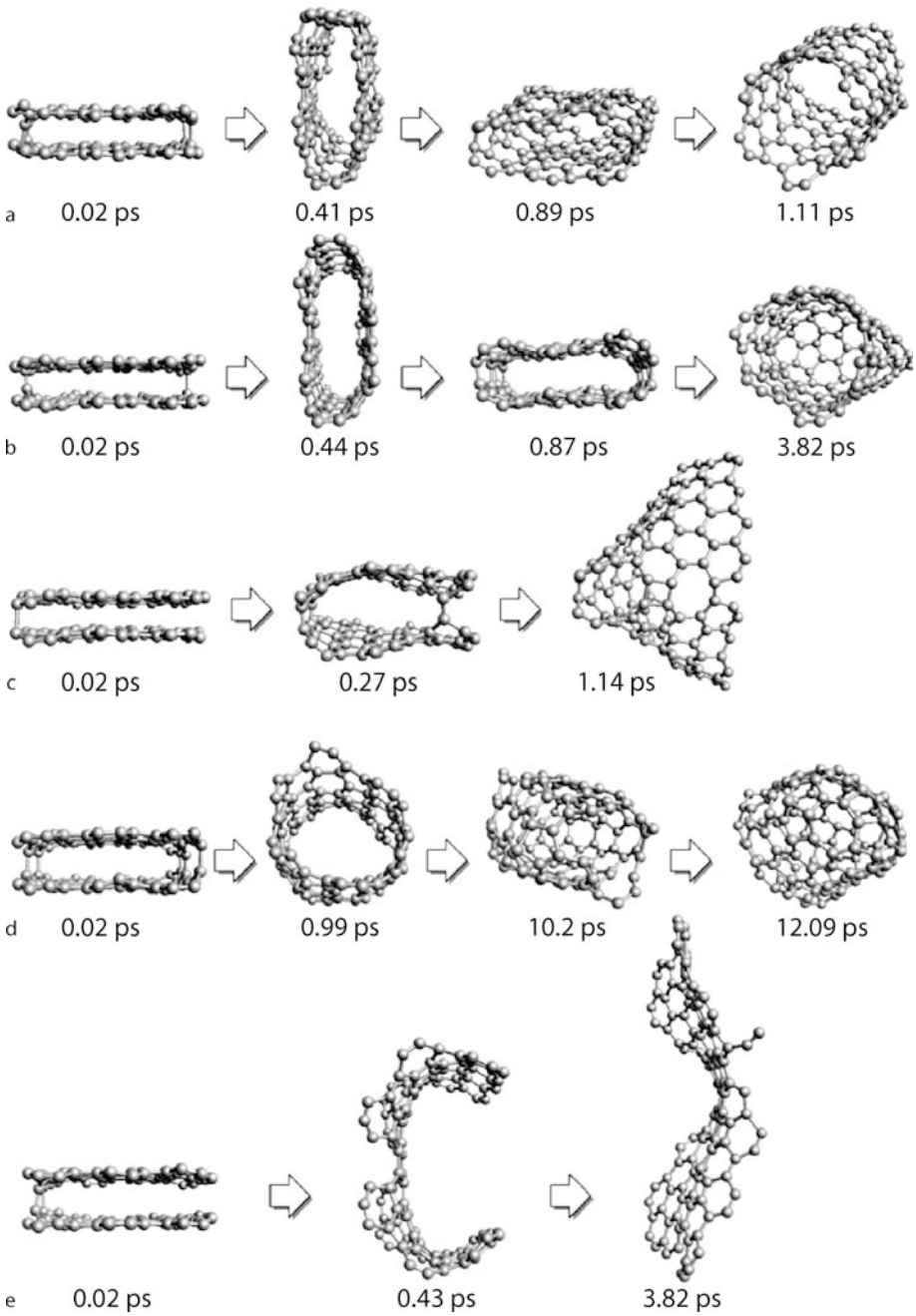

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 12
The time evolution of two reacting graphitic patches starting with A-A staking but with different interlayer distances D_0 and different temperatures: a D_0=2.1 Å, T_0=2000K, b D_0=2.1 Å, T_0=1000K, c D_0=2.3 Å, T_0=2000K, d D_0=1.9 Å, T_0=1000K, e D_0=1.9 Å, T_0=1000K. Note that d and e have the same starting condition but different end product. (From Ref. [29])

Junction Formation by Self-Healing of Vacancies

In the previous subsection, we noted that the formation of carbon nanotube junctions usually involves vacancy defects induced by irradiations. Very recently, the detailed atomistic processes of vacancy reconstruction which lead to the formation of nanotube junctions have been investigated by tight-binding molecular dynamics simulations

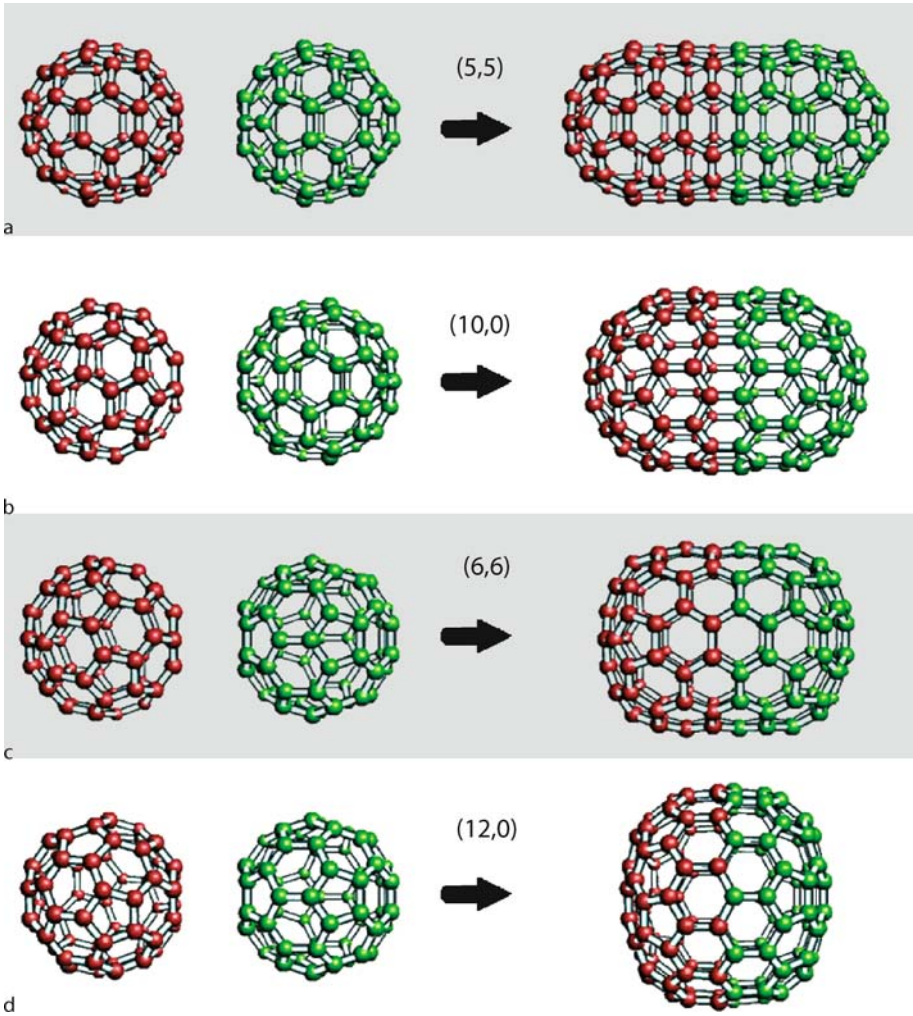

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 13
Initial and final states for forming various C_{120} nanocapsules as precursors for fullerene-based nanotube growth. (From Ref. [36])

based on the environment-dependent tight-binding carbon potential developed by Tang et al. [49,50,51,78]. The original EDTB carbon potential [78] is modified by incorporating an angle dependence factor into the repulsive energy to describe correctly the diffusion of an adatom and a vacancy in carbon nanotubes and graphene [49,50,51]. It has been shown that a carbon nanotube semiconductor-metal intramolecular junction can be formed by self-assembly of vacancy defects after the tube has been subjected to electron or ion irradiation [51].

Figure 17 shows the atomic details of vacancy reconstruction in a (16,0) SWCNT with a six-vacancy hole by the TBMD simulation [51]. The TBMD simulation is performed starting from a relaxed six-vacancy hole geometry as shown in Fig. 17a. In the early stage of the simulation, the SWCNT is heated up to high temperature through a constant-temperature molecular dynamics simulation. It was fund that rearrangement of carbon atoms around the vacancy hole starts to occur near 4500K at the simulation time of 18 ps through the rotation of carbon dimers, i. e., Stone–Wales transformation. After 19 ps of the simulation time, three hexagons at the lower left corner of the vacancy hole (those containing the atoms 1–4 in Fig. 17b) are recombined into a pentagon-octagon-pentagon defect by successive Stone–Wales transformations of the dimers 1–2 and 3–4 as shown in Fig. 17b. After the simulation time of 20 ps, another two hexagons (containing the atoms 5–7) on the other side of the vacancy hole are also reconstructed into one pentagon and one heptagon by the Stone–Wales transformation of the dimer 5–6. In order to prevent the

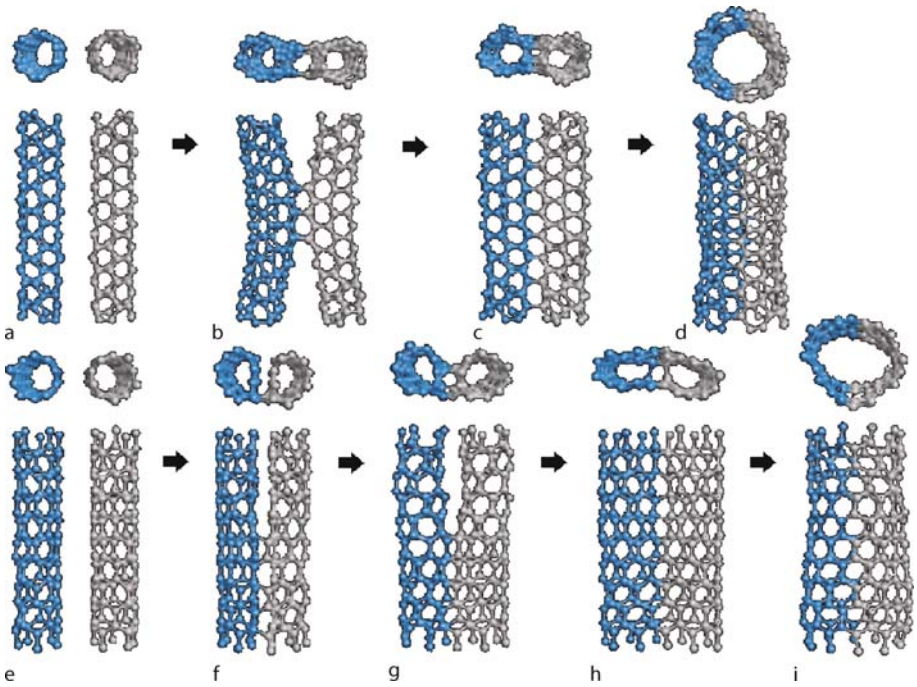

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 14
Coalescence process of the same two achiral nanotubes by colliding with each other (top and side views). The figures a–d are the snapshots of the reaction for (3,3) + (3,3), and the figures e–i are those of reaction process for (5,0) + (5,0). The resulting structures were identified as armchair (6,6) and zigzag (10,0) nanotube respectively. No defects are introduced before the collision. (From Ref. [30])

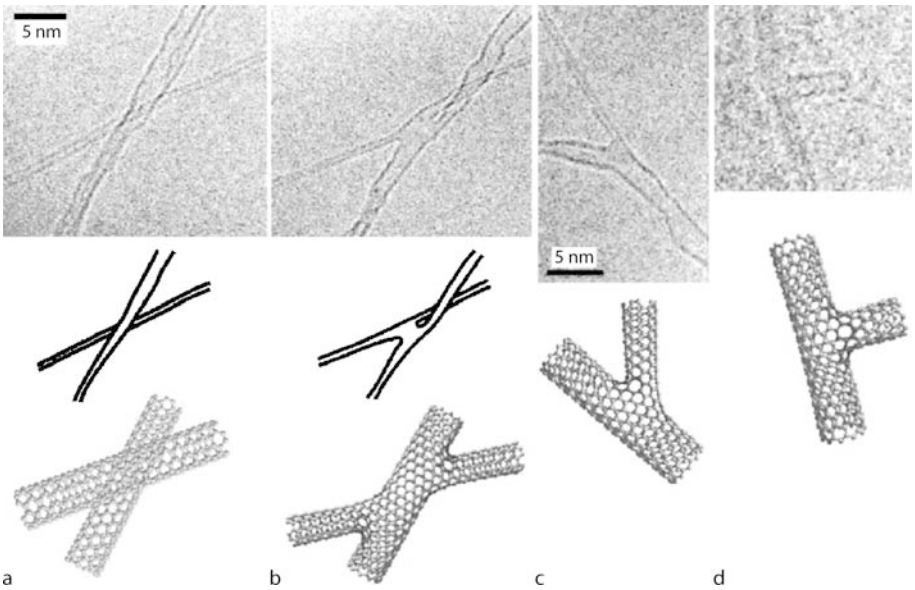

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 15
Various carbon nanotube junctions created by experiment. (From Ref. [79])

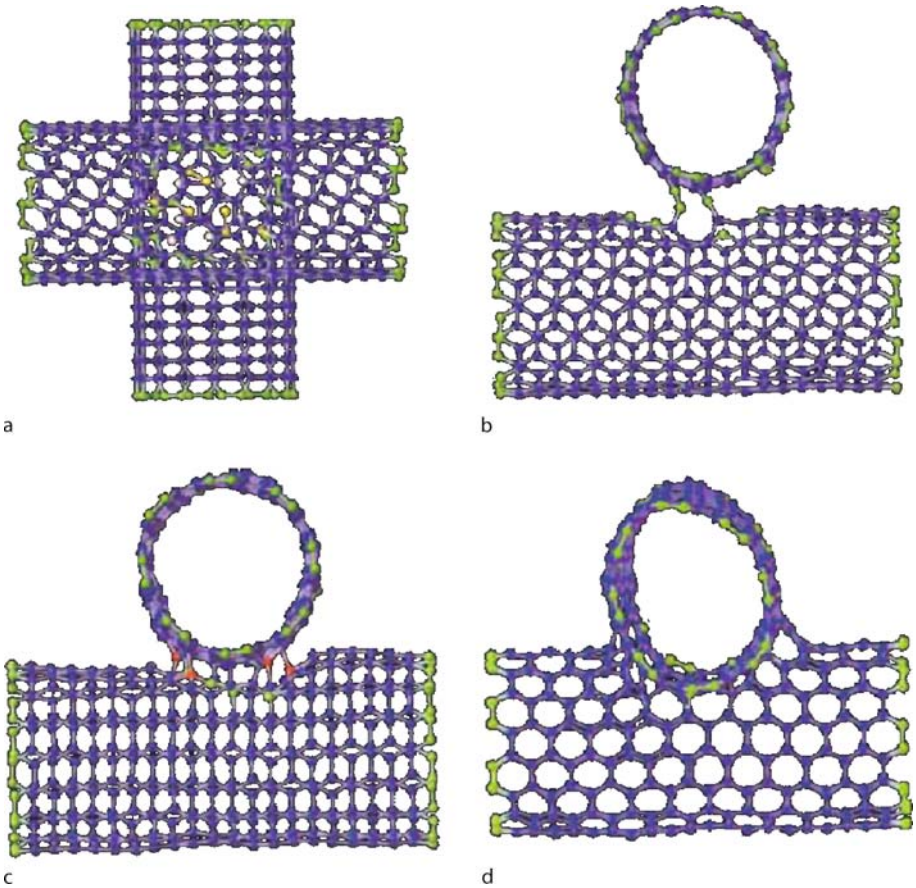

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 16
Sequences of merging between two crossing (8,8) carbon nanotubes into a unique X-like junction. a TBMD simulation starts with the random creation of 20 vacancies in the lattice of the two tubes in the localized neighboring region (top view). b After 10 ps, two links between the two defective tubes are formed via carbon chain (side view). c After 100 ps, the connection between the two tubes is established, although some sp^3-carbon atoms (*red*) and dangling bonds (*green*) are still remaining. d After 220 ps, surface reconstruction occurs and the carbon system approach an X junction. The reconstructed surface contains six heptagons, one octagon, one pentagon, and two dangling bonds. (From Ref. [79])

evaporation of carbon atoms, the system is then cooled down to 3,000 K for 4 ps and the vacancy hole is healed during this simulation period as shown in Fig. 17c. The structure immediately after the healing process consists of four pentagons and four heptagons with a two-fold rotation symmetry. The pentagon a and b and the heptagon c and d are related to the pentagon e and f and the heptagon g and h, respectively through the 2-fold axis which goes through the center of the carbon bond between atom 2 and 6. After the simulation time of 24 ps, the system is heated up again to 4,500 K for 7 ps and another structural reconstruction among the defects is observed. As shown in Fig. 17c and Fig. 17d, as the result of a Stone–Wales transformation of the dimer 1–3, the two heptagons (c and d in Fig. 17c) and one pentagon (b in Fig. 17c) on the left side of the 2-fold symmetry axis are transformed into three hexagons while one hexagonal ring containing the carbon atom 3 is transformed into a heptagonal ring. Finally a pentagon-heptagon pair defect, which has been observed in the experiment after the irradiation [21], is emerged through the reconstruction process. Since the dimer 5–7 is equivalent to the dimer 1–3 due to the 2-fold symmetry at the stage of Fig. 17c, the dimer 5–7 is expected to undergo a similar Stone–Wales transformation. Indeed, after 41 ps of simulation time, the Stone–Wales transformation happens to the carbon dimer 5–7. Consequently another pentagon-heptagon pair defect is formed at the right side of the 2-fold axis in the same way as the formation of the previous pentagon-heptagon pair on the left side of the 2-fold axis. The structure with two pentagon-heptagon pairs

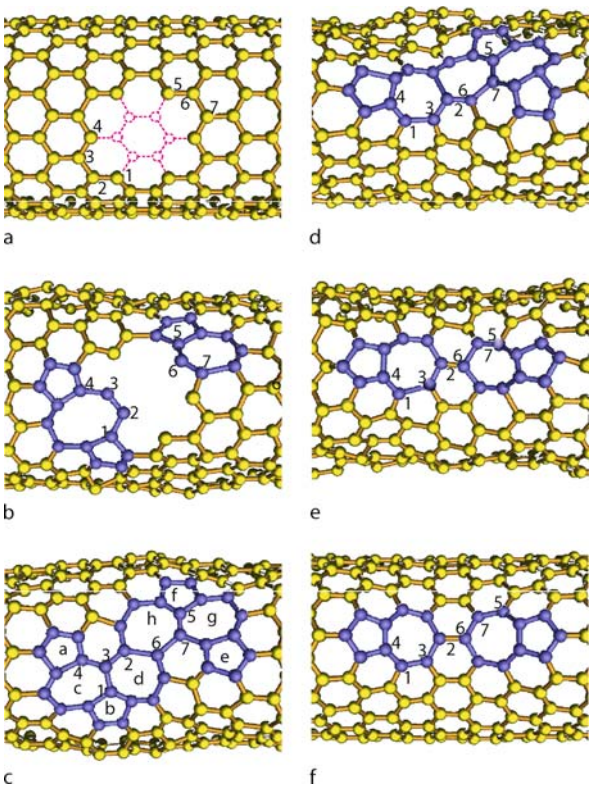

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 17

Atomic processes in the TBMD simulation for the (16, 0) SWCNT with six vacancies. a 0 K (at time $t = 0$ ps), b $\sim$ 4, 500 K ($t \approx 20.2$ ps), c $\sim$ 3, 100 K ($t \approx 23.2$ ps), d $\sim$ 4, 400 K ($t \approx 32.3$ ps), e $\sim$ 4, 700 K ($t \approx 41.5$ ps), f $\sim$ 90 K ($t \approx 53.9$ ps). The carbon atoms on the rear side of the tube are concealed in figures in order to see the reconstruction of vacancies more clearly. *Dotted circles* in a indicate the positions of the six carbon vacancies in the perfect (16, 0) SWCNT. *Yellow colors* indicate carbon atoms and bonds in hexagonal rings. *Blue colors* indicate carbon atoms and bonds in non-hexagonal rings. See the text for small letters in c and numbers. (From Ref. [51])

in Fig. 17e is very stable energetically and can sustain its shape without any changes for more than 20 ps in the simulation even at a temperature 4,500 K. At the final stage of the simulation, the system is gradually cooled down to 0 K in 12.5 ps and the structure with two pentagon-heptagon pair defects is found to maintain without any additional reconstruction as shown in Fig. 17f.

Figure 18 shows the front view of the initial and final structure from the TBMD simulation. The vacancy hole in the initial structure is healed up in the final structure and the radius of the tube in the middle section is reduced. The diameter and chirality in the center part of the final structure is found to be (15, 0), which is one of the metallic

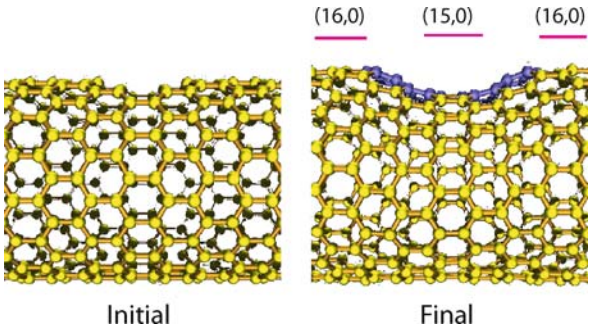

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 18

Front views of initial and final structure from TBMD simulation for (16, 0) SWCNT with six vacancies. The initial structure corresponds to Fig. 17a. The final structure corresponds to Fig. 17f. (From Ref. [51])

SWCNTs. In order to understand the effects of the vacancy cluster size on the formation of junctions, they have also performed the TBMD simulation to study the junction formation dynamics of a (16, 0) SWCNT containing a hole of ten vacancies. The formation of two pentagon-heptagon pair defects is also observed, with the mechanism similar to that in the simulation of the (16, 0) SWCNT with six vacancies discussed earlier in this subsection. The most interesting difference between the simulation results of the ten and six vacancies is that the length of the (15, 0) tube section is longer with ten vacancies. These simulation results demonstrate that intramolecular semiconductor-metal junctions of SWCNTs can be produced by irradiation followed by a proper annealing which allow various vacancy defects generated by the irradiation to reconstruct into the pentagon-heptagon pairs at the junction. These simulations also suggest a mechanism for synthesis of carbon nanotube semiconductor-metal intramolecular junctions with specific locations and controlled sizes and show the possibility of application to nanoelectronic devices.

Transformation from Nanodiamond

Nanometer-sized diamonds have been found in interstellar dust [53], solid detonation products [17], and diamond-like films [6]. Recently, Raty et al. [72] performed ab initio calculations and tight-binding molecular dynamics simulations to study the structure of a nanodiamond and found that the carbon nanoparticle consist of a diamond core and a reconstructed fullerene-like surface. Experiments have shown that diamond nanoparticles of diameter $\sim$ 5 nm can be transformed into spherical and polyhedron carbon onions at high temperatures [45,84,85].

Using the environment-dependent carbon tight-binding potential developed by Tang et al. [78], Lee et al. have performed tight-binding molecular dynamics simulations to study the structural transformation of nanodiamond at high temperature. The simulations show that upon annealing up to 2500 K, a 1.4 nm-diameter nanodiamond is transformed into a cage structure that looks like a single-walled capped nanotube [52].

The simulation was started with a bulk-terminated carbon cluster of 275 atoms within a sphere of diameter of 1.4 nm cut from bulk diamond. This cluster is relaxed using the steepest descent method with the environment-dependent tight-binding carbon potential. The cluster structure after the relaxation is similar to that of the previous ab initio calculation [72]. Due to the surface reconstruction of nanodiamonds, there are graphite-like fragments present at the first atomic layer of the (1 1 1) facets, together with the formation of pentagons which link the graphene fragments to the underneath atoms (Fig. 19a). The graphite-like fragments at the (1 1 1), $(\bar{1}\,\bar{1}\,1)$, $(\bar{1}\,1\,\bar{1})$, and $(1\,\bar{1}\,\bar{1})$ surfaces consist of 3 pentagons and 3 hexagons and those at the $(\bar{1}\,1\,1)$, (1 1 1), $(1\,1\,\bar{1})$, and $(\bar{1}\,\bar{1}\,\bar{1})$ surfaces consist of only 3 pentagons. Starting from this relaxed cluster geometry, tight-binding molecular dynamics simulation was performed to investigate the structural transformation of the nanodiamond at high temperatures.

Figure 19 displays snapshots of the system during the structural transformation into a capped nanotube. First the nanodiamond cluster was heated up to about 2,500 K by constant-temperature molecular dynamics simulations. Near 2,500 K, as shown in Fig. 19b, the (1 1 1) surface layer of the nanodiamond begins to graphitize after a simulation time of 3 picoseconds (ps), the exfoliation of the graphitized (1 1 1) layer occurs by breaking the bonds between graphene fragments and the underneath "core" atoms. This resembles the graphitization process of the (1 1 1) surface of bulk diamond induced by nanosecond laser pulses [93]. As the simulation continues, the graphitized layer evaporates, breaking down into carbon dimers one-by-one from the end of layer. Similarly, the $(\bar{1}\,\bar{1}\,\bar{1})$ surface layer consisting of three pentagons undergoes the same exfoliation and evaporation process as that of the (1 1 1) surface. At about 18 ps, as shown in Fig. 19c, the graphitization process extends to the entire cluster surface. The "core" atoms and the surface "shell" atoms start to separate at the bottom side of the cluster. As the bonds between the "core" and "shell" atoms start to break up, the cluster begins to inflate like a bubble. At this stage, if the thermostat is maintained to keep the system at the constant temperature of 2,500 K, the whole cluster would completely evaporate within a simulation time of 45 ps. To prevent the full vaporization, the system is cooled by decreasing the temperature from 2,500 K to 2,000 K in 10 ps (during the simulation time between 25 ps and 35 ps). The cluster is then further cooled down to a temperature of $\sim$ 1, 500 K

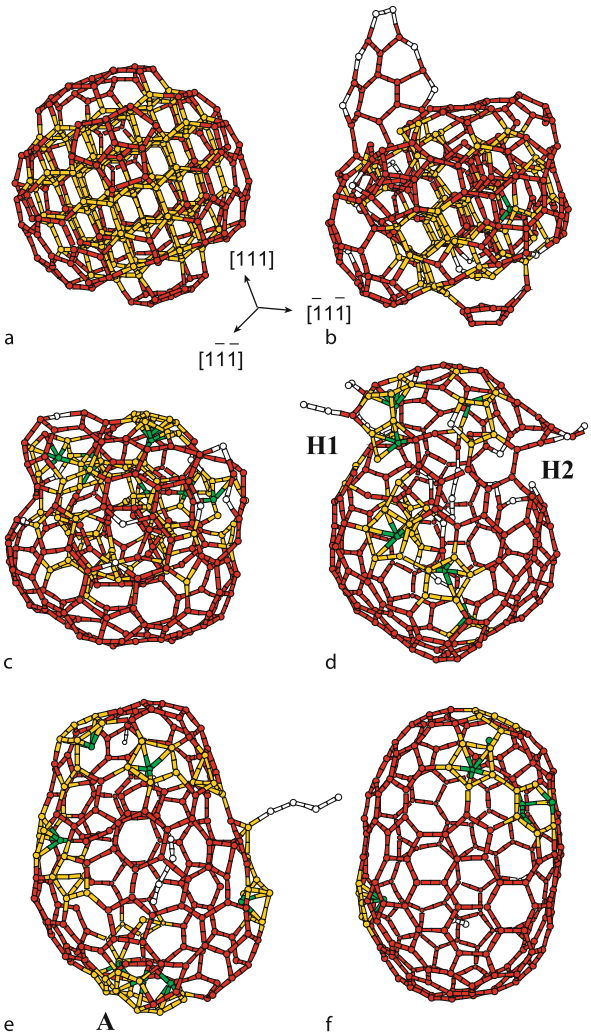

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 19

Atomic processes of structural transformation of nanodiamond to capped nanotube by successive annealings. a 0 K (at time $t = 0$ ps), b $\sim$ 2, 500 K ($t \approx 3$ ps), c $\sim$ 2, 500 K ($t \approx 19$ ps), d $\sim$ 2, 100 K ($t \approx 35$ ps), e $\sim$ 1, 900 K ($t \approx 50$ ps), f $\sim$ 20 sK ($t \approx 120$ ps). Simulated annealings with temperatures up to 3,000 K are performed during the process e→f. *White color* indicates atoms and bonds of 2 and less-fold coordination. *Red* and *yellow colors* indicate atoms and bonds of 3-fold coordination and 4-fold coordination, respectively. Green colors indicate atoms and bonds of 5 and higher-fold coordination. Note that two holes H1 and H2 are created in d which serve as exits for inner atoms to "flow-out" to the surface of the structure. (From Ref. [52])

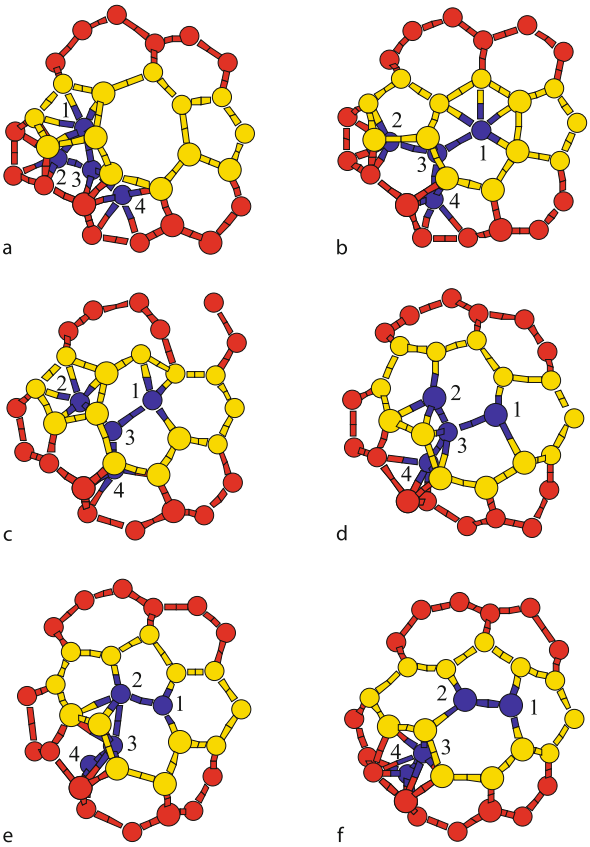

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 20

Conversion of inner residual carbon atoms into surface by the 'direct adsorption' mechanism. The figure shows the part of the cluster marked by 'A' in Fig. 19e. Blue atoms indicate the inner atoms of the cluster at the initial configuration a. Red and yellow atoms indicate surface atoms. a Four inner atoms (labels 1, 2, 3, and 4) are seen to adhere to the surface of the nanotube from inside. b–c Atom '1' then moves to the center of a heptagonal ring, and breaks the bond at the boundary of the heptagonal and pentagonal ring. d A sp^2 bonding configuration for atom "1" is formed and at the same time atom '2' also breaks the bond at the boundary of the heptagonal and the other pentagonal ring. e–f eventually, atoms '1' and '2' are combined and converted into surface atoms. (From Ref. [52])

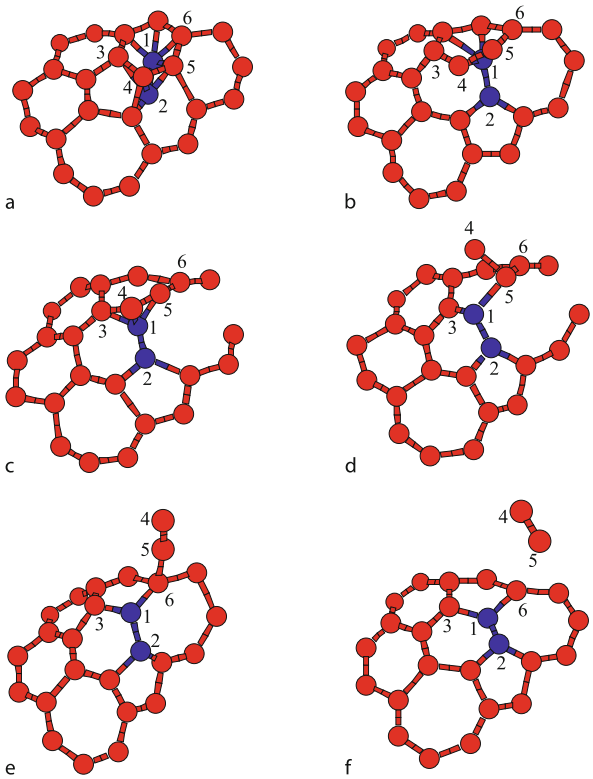

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 21

Conversion of residual inner carbon atoms into surface by the 'push out' mechanism. Blue and red colors indicate inner and surface atoms respectively at the initial configuration of a. a–b The inner atoms '1' and '2' are seen to move from one place to another, breaking their bonds with some surface atoms. c–d Atom '3' becomes a 4-fold coordinated atom by bonding to the inner atom '1', and then breaks the bond with atom '4' to keep its 3-fold coordination. d–e After the inner atom '1' forms a bond with atom '6', it breaks the bond with atom '5'. f Atom '6' also breaks the bond with atom '5' to keep its 3-fold coordination and the carbon dimer formed by atom '4' and '5' is evaporated and the inner atoms '1' and '2' finally convert into surface atoms. (From Ref. [52])

in 20 ps when a stable cage structure is found to form. The two holes ('H1' and 'H2' in Fig. 19d), generated by the successive breaking of bonds among surface atoms play an important role in pumping the inner carbon atoms out onto the surface to form the graphitic layer. The process is referred to as the 'flow-out' mechanism by Lee et al. The 'flow-out' mechanism is a unique process in the transformation of a nanodiamond structure into a cage structure. This mechanism is distinguished from other transformation mechanisms observed on bulk diamond surfaces.

During the annealing process (stage e–f in Fig. 19), two other interesting atomic processes, namely the 'direct absorption' and 'push-out' mechanisms, have been identified from the simulation to play a crucial role in the conversion of the residual inner carbon atoms of the nanodiamond into the surface atoms of the nanotube. These mechanisms are illustrated in Figs. 20 and 21 respectively.

Future Directions

In this review, we have shown that tight-binding molecular dynamics is an accurate and efficient method for studying

the structures and properties of carbon fullerenes and carbon nanotubes. Tight-binding molecular dynamics simulation studies of various pathways of the fullerene and nanotube formations have provided useful insights into the formation mechanism of these nanostructures at the atomic scale.

Nevertheless, our understanding of the formation of carbon nanostructures is still very far from being completed. For example, it is a common practice in experiment to use small amount of transition metals (e. g., Fe) to catalyze the nucleation and growth of single-walled carbon nanotubes [15]. The atomic processes under such growth condition have not yet been well studied. Multi-walled carbon nanotubes and buckyonions are also not well studied even though such nanostructures are frequently seen in experiments. The bottleneck is the development of accurate and transferable tight-binding potentials. So far, there are no reliable potentials for carbon/transition metal interactions suitable for tight-binding molecular dynamics simulations. In addition, the tight-binding carbon potentials we discussed in this paper (i. e., the XWCH and EDTB potentials) do not describe the interaction between graphite layers accurately. They are therefore not suitable for studying multi-walled carbon nanotubes and buckyonions. Besides the carbon nanostructures, there is increasing interest in using tight-binding molecular dynamics to study other nanostructures (e. g. Si-based nanostructures and metal nanostructures). The success of such future studies depends on the development of more refined tight-binding potentials.

Very recently, the authors and co-worker have developed a method to extract environment-dependent minimal-basis-set orbitals from ab initio wavefunctions [55,56,57]. These orbitals give an exact description of the occupied electronic states and are highly localized on the individual atoms, making them ideal for use in an accurate tight-binding description of the system. This scheme is easily generalizable to systems involving different atomic species and will simplify the task of generating accurate tight-binding potentials for complex systems. We anticipate that more accurate and transferable tight-binding potentials for nanostructure simulations will be generated based on these highly localized environment-dependent minimal-basis-set orbitals.

Acknowledgments

We would like to thank Dr. Wencai Lu for her help in preparing the figures. Ames Laboratory is operated for the US Department of Energy by Iowa State University under Contract No. DE-AC02-07CH11358. This work was supported by the Director for Energy Research, Office of Basic Energy Sciences including a grant of computer time at the National Energy Research Supercomputing Center (NERSC) in Berkeley.

Bibliography

1. Ballone P, Milani P (1990) Phys Rev B 42:3201
2. Bandow S, Takizawa M, Hirahara K, Yudasaka M, Iijima S (2001) Chem Phys Lett 337:48
3. Baskes MI (1987) Phys Rev Lett 59:2666
4. Biswas R, Hamann DR (1985) Phys Rev Lett 55:2001
5. Car R, Parrinello M (1985) Phys Rev Lett 55:2471
6. Chang YK, Hsieh HH, Pong WF, Tsai MH, Chien FZ, Tseng PK, Chen LC, Wang TY, Chen KH, Bhusari DM (1999) Phys Rev Lett 82:5377
7. Chelikowsky JR (2000) J Phys D 33:R33
8. Chelikowsky JR, Louie SG (eds) (1996) Quantum Theory of Real Materials. Kluwer Press, Boston
9. Chelikowsky JR, Phillips JC, Kamal M, Strauss M (1989) Phys Rev Lett 62:292
10. Chico L, Crespi VH, Benedict LX, Louie SG, Cohen ML (1996) Phys Rev Lett 76:971
11. Cowley ER (1988) Phys Rev Lett 60:2379
12. Crespi VH, Cohen ML (1997) Phys Rev Lett 79:2093
13. Dodson BW (1987) Phys Rev B 35:2795
14. Dresselhaus MS, Dresselhaus G (1982) In: Cardona M, Guntherodt G (ed) Light Scattering in Solids III. Springer, Berlin, p 8
15. Dresselhaus MS, Dresselhaus G, Avouris P (eds) (2001) Carbon Nanotubes: Synthesis, Structure, Properties, Applications. Springer, Berlin
16. Goodwin L, Skinner AJ, Pettifor DG (1989) Europhys Lett 9:701
17. Greiner NR, Phillios DS, Johnson JD, Volk F (1988) Nature 333:440
18. Haerle R, Galli G, Baldereschi A (1999) Appl Phys Lett 75:1718
19. Haerle R, Riedo E, Pasquarello A, Baldereschi A (2001) Phys Rev B 65:045101
20. Hamada N, Sawada S-I, Oshiyama A (1992) Phys Rev Lett 68:579
21. Hashimoto A, Suenaga K, Gloter A, Urita K, Iijima S (2004) Nature 430:870
22. Hernandez E, Meunier V, Smith BW, Rurali R, Terrones H, Buongiorno M Nardelli, Terrones M, Luzzi DE, Charlierr J-C (2003) Nanolett 3:1037
23. Hohenberg P, Kohn W (1964) Phys Rev 136:B864
24. Iijima S (1991) Nature 354:56
25. Irle S, Zheng G, Elstner M, Morokuma K (2003) Nanolett 3:1657
26. Irle S, Zheng G, Elstner M, Morokuma K (2003) Nanolett 3:465
27. Irle S, Zheng G, Wang Z, Morokuma K (2006) J Phys Chem. B 110:14531
28. Lee I-H, Kim H, Lee J (2004) J Chem Phys 120:4672
29. Kawai T, Miyamoto Y, Sugino O, Koga Y (2002) Phys Rev B 66:033404
30. Kawai T, Miyamoto Y, Sugino O, Koga Y (2002) Phys Rev Lett 89:085901
31. Khan FS, Broughton JQ (1989) Phys Rev B 39:3688
32. Khor KE, Das S (1988) Phys Rev B 38:3318
33. Kilic C, Ciraci S, Gulseren O, Yildirim T (2000) Phys Rev B 62:R16345

34. Kim D-H, Chang KJ (2002) Phys Rev B 65:155402
35. Kim D-H, Sim H-S, Chang KJ (2001) Phys Rev B 64:115409
36. Kim Y-H, Lee I-H, Chang KJ, Lee S (2003) Phys Rev Lett 90:065501
37. Kohn W, Sham LJ (1965) Phys Rev 140:A1133
38. Krätschmer W, Lamb LD, Fostiropoulos K, Huffman DR (1990) Nature 347:354
39. Kresse G (1993) Ph.D Technische University of Wien
40. Kresse G, Furthmuller J (1996) Comput Mater Sci 6:15
41. Kresse G, Furthmuller J (1996) Phys Rev B 54:11 169
42. Kresse G, Hafner J (1993) Phys Rev B i47:558
43. Kresse G, Joubert J (1999) Phys Rev B 59:1758
44. Kroto HW, Heath JR, O'Brien SC, Curl RF, Smalley RE (1985) Nature 318:162
45. Kuznetsov VL, Chuvilin AL, Butenko YV, Mal'kov IY, Titov VM (1994) Chem Phys Lett 222:343
46. Laasonen K, Nieminen RM (1991) J Phys Condens Matter 2:1509
47. Laasonen K, Nieminen RM (1991) ibid 3:7455
48. Laszlo I (1998) Europhys Lett 44:741
49. Lee G-D, Wang CZ, Yoon E, Hwang N-M, Ho KM (2006) Phys Rev B 74:245411
50. Lee G-D, Wang CZ, Yoon E, Hwang N-M, Kim D-Y, Ho KM (2005) Phys Rev Lett 95:205501
51. Lee G-D, Wang CZ, Yu J, Yoon E, Hwang N-M, Ho K-M (2007) Phys Rev B 76:165413
52. Lee GD, Wang CZ, Yu J, Yoon E, Ho KM (2003) Phys Rev Lett 91:265701
53. Lewis RS, Anders E, Draine BT (1989) Nature 339:117
54. Lu JQ, Wu J, Duan WH, Liu F, Zhu BF, Gu BL (2003) Phys Rev Lett 90:156601
55. Lu WC, Wang CZ, Ruedenberg K, Ho KM (2004) Phys Rev B 70:041101
56. Lu WC, Wang CZ, Schmidt MW, Bytautas L, Ho KM, Ruedenberg K (2004) J Chem Phys 120:2629
57. Lu WC, Wang CZ, Schmidt MW, Bytautas L, Ho KM, Ruedenberg K (2004) J Chem Phys 120:2638
58. Luzzi DE, Smith BW (2000) Carbon 38:1751
59. Madelung O, Schulz M (eds) (1987) Semiconductors: Intrinsic Properties of Group IV Elements and III-V, II-VI and I-VII Compounds. Landolt-Börnstein New Series III/22a. Springer, Berlin
60. Madelung O, Schulz M, Weiss H (eds) (1982) Semiconductors: Physics of Group IV Elements and III-V Compounds. Landolt-Börnstein New Series III/17a. Springer, Berlin
61. Menon M, Andriotis AN, Srivastava D, Ponomareva I, Chernozatonskii LA (2003) Phys Rev Lett 91:144501
62. Menon M, Subbaswamy KR (1991) Phys Rev Lett 67:3487
63. Mintmire JW, Dunlap BI, White CT (1992) Phys Rev Lett 68:631
64. Morris JR, Wang CZ, Ho KM (1995) Phys Rev B 52:4138
65. Nikolaev P, Thess A, Rinzler AG, Colbert DT, Smalley RE (1997) Chem Phys Lett 266:422
66. Odom TW, Huang J-L, Kim P, Lieber CM (1998) Nature 391:62
67. Ogata S, Shibutani Y (2003) Phys Rev B 68:165409
68. Oh D-H, Lee YH (1998) Phys Rev B 58:7407
69. Ozaki T, Iwasa Y, Mitani T (2000) Phys Rev Lett 84:1712
70. Park C-J, Kim Y-H, Chang KJ (1999) Phys Rev B 60:10656
71. Porezag D, Frauenheim T, Kohler T, Seifert G, Kaschner R (1995) Phys Rev B 51:12947
72. Raty JY, Galli G, Bostedt C, Van Buuren TW, Terminello LJ (2003) Phys Rev Lett 90:037401
73. Saito R, Fujita M, Dresselhaus G, Dresselhaus MS (1992) Appl Phys Lett 60:2204
74. Sankey OF, Niklewski DJ (1989) Phys Rev B 40:3979
75. Slater JC, Koster GF (1954) Phys Rev 94:1498
76. Smith BW, Monthioux M, Luzzi DE (1998) Nature 396:323
77. Stillinger FH, Weber TA (1985) Phys Rev B 31:5262
78. Tang MS, Wang CZ, Chan CT, Ho KM (1996) Phys Rev B 53:979
79. Terrones M, Banhart F, Grobert N, Charlier J-C, Terrones H, Ajayan PM (2002) Phys Rev Lett 89:75505
80. Terrones M, Terrones H, Banhart F, Charlier J-C, Ajayan PM (2000) Science 288:1226
81. Tersoff J (1986) Phys Rev Lett 56:632
82. Tersoff J (1988) Phys Rev B 37:6991
83. Tersoff J (1988) Phys Rev Lett 61:2879
84. Tomita S, Fujii M, Hayashi S (2002) Phys Rev B 66:245424
85. Tomita S, Fujii M, Hayashi S, Yamamoto K (1999) Chem Phys Lett 305:225
86. Virkkunen R, Laasonen K, Nieminen RM (1991) ibid 3:7455
87. Wang CZ, Chan CT, Ho KM (1989) Phys Rev B 39:8592
88. Wang CZ, Chan CT, Ho KM (1992) Phys Rev B 46:9761
89. Wang CZ, Ho KM (1993) Phys Rev Lett 71:1184
90. Wang CZ, Ho KM (1994) Phys Rev B 50:12429
91. Wang CZ, Ho KM, Chan CT (1993) Phys Rev B 47:14835
92. Wang CZ, Ho KM, Chan CT (1993) Phys Rev Lett 70:611
93. Wang CZ, Ho KM, Shirk M, Molian P (2000) Phys Rev Lett 85:4092
94. Wang CZ, Xu CH, Chan CT, Ho KM (1992) J Phys Chem 96:3563
95. Wildoer JWG, Venema LC, Rinzler AG, Smalley RE, Dekker C (1998) Nature 391:59
96. Xu CH, Wang CZ, Chan CT, Ho KM (1992) J Phys Condens Matter 4:6047
97. Yang L, Han J (2000) Phys Rev Lett 85:154
98. Zhang BL, Wang CZ, Chan CT, Ho KM (1993) J Phys Chem 97:3134
99. Zhang BL, Wang CZ, Ho KM (1992) Chem Phys Lett 193:225
100. Zhang BL, Wang CZ, Ho KM (1992) J Chem Phys 96:7183
101. Zhang BL, Wang CZ, Ho KM, Xu CH, Chan CT (1993) J Chem Phys 97:5007(1992); 98:3095
102. Zhang BL, Xu CH, Wang CZ, Chan CT, Ho KM (1992) Phys Rev B 46:7333
103. Zhang P, Crespi VH (1999) Phys Rev Lett 83:1791
104. Zhao QZ, Nardelli MB, Bernholc J (2003) Phys Rev B 65:144105
105. Zheng G, Irle S, Morokuma K (2005) J Chem Phys 122:014708
106. Zheng G, Irle S, Elstner M, Morokuma K (2004) J Phys Chem A 108:3182

Tiling Problem and Undecidability in Cellular Automata

Jarkko Kari
Department of Mathematics,
University of Turku, Turku, Finland

Article Outline

Glossary
Definition of the Subject

Introduction
The Tiling Problem and Its Variants
Undecidability in Cellular Automata
Future Directions
Acknowledgments
Bibliography

Glossary

Cellular automata (CA) A d-dimensional cellular automaton consists of an infinite d-dimensional grid of cells, indexed by $\mathbb{Z}^d$. Each cell stores an element of a finite state set S. Configuration $c\colon \mathbb{Z}^d \to S$ specifies the states of all cells. The set of all configurations is $S^{\mathbb{Z}^d}$. The neighborhood vector $N = (\vec{n}_1, \vec{n}_2, \ldots, \vec{n}_m)$ is a sequence of m distinct elements of $\mathbb{Z}^d$ specifying the relative locations of the neighbors of the cells: A cell located at $\vec{x} \in \mathbb{Z}^d$ has m neighbors, in positions $\vec{x} + \vec{n}_1, \vec{x} + \vec{n}_2, \ldots, \vec{x} + \vec{n}_m$. Finally, the local update rule $f\colon S^m \to S$ specifies the new state of a cell, based on the old states of its neighbors. In one step, configuration c is transformed into configuration e where, for all $\vec{x} \in \mathbb{Z}^d$,

$$e(\vec{x}) = f\left[c(\vec{x} + \vec{n}_1), c(\vec{x} + \vec{n}_2), \ldots, c(\vec{x} + \vec{n}_m)\right] .$$

The mapping $c \mapsto e$ is the global transition function, or the CA-function, $G\colon S^{\mathbb{Z}^d} \to S^{\mathbb{Z}^d}$ specified by the CA $\mathcal{A} = (d, S, N, f)$.

Tiles A Wang tile is a unit square tile with colored edges. Tiles have an orientation, i. e. they may not be rotated or reflected. The colors give a local matching rule that specifies which tiles may be placed next to each other: Two adjacent tiles must have identical colors on the abutting edges. A Wang tile set consists of a finite number of Wang tiles.

A more general definition: a d-dimensional tile set is a quadruple $\mathcal{T} = (d, T, N, R)$ where T is a finite set whose elements are called tiles, $N = (\vec{n}_1, \vec{n}_2, \ldots, \vec{n}_m)$ is a neighborhood vector of m distinct elements of $\mathbb{Z}^d$ and $R \subseteq T^m$ is a relation of allowed patterns. The neighborhood vector has the same interpretation as in the definition of cellular automata: it gives the relative locations of the neighbors of cells.

Tiling A covering of the plane using tiles. A valid Wang tiling by a Wang tile set T is an assignment $t\colon \mathbb{Z}^2 \to T$ of tiles to cells such that the local matching rule is satisfied between all adjacent tiles. We say that T admits tiling t.

More general definition: A d-dimensional tiling using tile set $\mathcal{T} = (d, T, N, R)$ is a mapping $t\colon \mathbb{Z}^d \to T$. Tiling t is valid at cell $\vec{x} \in \mathbb{Z}^d$ if

$$t(\vec{x} + \vec{n}_1, \vec{x} + \vec{n}_2, \ldots, \vec{x} + \vec{n}_m) \in R .$$

Tiling t is called valid if it is valid at every cell $\vec{x} \in \mathbb{Z}^d$.

Periodic tiling A tiling that is invariant under some nonzero translation. A two-dimensional tiling is called totally periodic if it is invariant under two linearly independent translations. A totally periodic tiling is automatically periodic in horizontal and vertical directions, which means that it consists of a rectangular pattern that is repeated horizontally and vertically to fill the plane. A two-dimensional tile set that admits a valid periodic tiling automatically admits also totally periodic tiling.

More generally, a d-dimensional tiling $t\colon \mathbb{Z}^d \to T$ is periodic with period $\vec{p} \in \mathbb{Z}^d, \vec{p} \neq \vec{0}$, if $t(\vec{x}) = t(\vec{x} + \vec{p})$ for all $\vec{x} \in \mathbb{Z}^d$. It is totally periodic if it is periodic with d linearly independent periods $\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_d$. Note that when $d > 2$ it is possible that a tile set admits a periodic tiling but does not admit any totally periodic tiling.

Aperiodic tile set A two-dimensional tile set that admits a valid tiling of the plane, but does not admit any valid periodic tilings. Smallest known aperiodic set of Wang tiles contain 13 tiles [3,13]

Turing machine (TM) Turing machines are computation devices commonly used to formally define the concept of an algorithm. They also provide us with the most basic undecidable decision problems. A Turing machine consists of a finite state control unit that moves along an infinite tape. The tape has symbols written in cells that are indexed by $\mathbb{Z}$. Depending on the state of the control unit and the symbol currently scanned on the tape the machine may overwrite the tape symbol, change the internal state and move along the tape one cell to the left or right. We formally define a Turing machine as a 6-tuple $\mathcal{M} = (Q, \Gamma, \delta, q_0, q_h, b)$ where Q and Γ are finite sets (the state alphabet and the tape alphabet, respectively), $q_0, q_h \in Q$ are the initial and the halting states, respectively, $b \in \Gamma$ is the blank symbol and $\delta\colon Q \times \Gamma \to Q \times \Gamma \times \{-1, 1\}$ is the transition function that specifies the moves of the machine. A configuration (or instantaneous description) of the machine is a triplet (q, i, t) where $q \in Q$ is the current state, $i \in \mathbb{Z}$ is the position of the machine on the tape and $t\colon \mathbb{Z} \to \Gamma$ describes the content of the tape. In one time step configuration (q, i, t) becomes $(q', i + d, t')$ if $\delta(q, t(i)) = (q', \gamma, d)$ and $t'(i) = \gamma$ and $t'(j) = t(j)$ for all $j \neq i$. We denote

this move by

$$(q, i, t) \vdash (q', i + d, t') .$$

The reflexive, transitive closure of $\vdash$ is denoted by $\vdash^*$, that is,

$$(q, i, t) \vdash^* (q', i', t')$$

if and only if (q', i', t') can be reached from (q, i, t) by executing zero or more Turing machine moves.

Decision problem A decision problem is an algorithmic question with a yes/no -answer. The problem has an input (called the instance of the problem) and a well defined answer "yes" or "no" associated to each instance.

The halting problem TURING MACHINE HALTING ON BLANK TAPE is the decision problem whose input is a Turing machine $\mathcal{M} = (Q, \Gamma, \delta, q_0, q_h, b)$ and the answer is positive if and only if the Turing machine eventually enters its halting state q_h when started in the initial state q_0 on a totally blank tape, i. e. initially every tape location has the blank symbol b.

(Un)decidability Some decision problems can not be solved by any algorithm. Such problems are called undecidable. In contrast, decidable decision problems are solved by some algorithm. An example of an undecidable problem is the TURING MACHINE HALTING ON BLANK TAPE.

Semi-algorithm An algorithm-like procedure for a decision problem that correctly returns a positive answer on positive input instances, but on negative instances runs for ever without ever returning an answer.

Semi-decidability A decision problem is called semi-decidable if there is a semi-algorithm for it. For example, the decision problem TURING MACHINE HALTING ON BLANK TAPE is semi-decidable since one can simulate any given Turing machine step-by-step until (if ever) it halts.

Recursive and recursively enumerable (re) A formal language L is called recursive if it is decidable whether a given word belongs to L. The language is called recursively enumerable (re for short) if this membership problem is semi-decidable.

The tiling problem The decision problem that gets as input a tile set $\mathcal{T}$, and asks whether there exists a valid tiling by $\mathcal{T}$. The tiling problem was proved undecidable for Wang tiles by R. Berger [2]. Its complement (i. e., "Does there not exist a valid tiling ?") is semi-decidable.

The tiling problem with a seed tile The decision problem that gets as input a tile set $\mathcal{T}$ and one tile s, and asks whether $\mathcal{T}$ admits a tiling that contains tile s at least once. The problem is undecidable for Wang tiles, but its complement is semi-decidable [24].

The periodic tiling problem The decision problem to determine if a given set of Wang tiles admits a periodic tiling. The problem is undecidable, but it is semi-decidable [7].

The finite tiling problem A decision problem where we are given a set T of Wang tiles and a specific blank tile $B \in T$ whose all four edges are colored with the same color. A tiling $t\colon \mathbb{Z} \to T$ is called finite if

$$\{\vec{x} \in \mathbb{Z}^2 | t(\vec{x}) \neq B\}$$

is a finite set. If $t(\vec{x}) = B$ for all $\vec{x} \in \mathbb{Z}^2$ then t is called trivial. The finite tiling problem asks whether there exist non-trivial valid finite tilings. The problem is undecidable but semi-decidable.

Surjective cellular automata A cellular automaton (CA) is called surjective if every configuration has a pre-image, that is, if its global transition function $G\colon S^{\mathbb{Z}^d} \to S^{\mathbb{Z}^d}$ is surjective.

Injective (reversible) CA A CA is injective if every configuration has at most one pre-image, that is, the global transition function is one-to-one. It is well known that a CA is injective if and only if it is bijective (every configuration has a unique pre-image), which in turn is equivalent to reversibility (=there exists an inverse CA that traces the CA back in time.)

Limit set The limit set of a CA is its maximal attractor. In other words, it is the compact and translation invariant set

$$\bigcap_{i=0}^{\infty} G(S^{\mathbb{Z}^d})$$

where G is the global transition function and S is the state set.

Nilpotent cellular automata Nilpotent CA have trivial dynamics. A CA is called nilpotent if its limit set contains only one configuration. The unique element of the limit set is the quiescent configuration. This is equivalent to every initial configuration eventually becoming the quiescent configuration.

Equicontinuous CA Cellular automaton G is called equicontinuous if for every finite $A \subseteq \mathbb{Z}^d$ there exists a finite $B \subseteq \mathbb{Z}^d$ such that any two initial configurations that agree inside B will agree inside A for all subsequent steps. In other words,

$$\forall \vec{x} \in B\colon c(\vec{x}) = e(\vec{x}) \\ \implies \forall t \in \mathbb{N} \text{ and } \forall \vec{x} \in A\colon G^t(c)(\vec{x}) = G^t(e)(\vec{x}).$$

This means that equicontinuous CA can be reliably simulated in finite windows. It is known that a CA is equicontinuous if and only if it is ultimately periodic [17]: $\exists n, p \in \mathbb{N}: G^n = G^{n+p}$. In this sense the dynamics of equicontinuous CA is trivial.

Sensitive CA Cellular automaton G is called sensitive to initial conditions if there exists a finite set $B \subseteq \mathbb{Z}^d$ of cells such that for every configuration c and every finite set $A \subseteq \mathbb{Z}$ of cells there exists a configuration e and time $t \geq 0$ such that $e(\vec{x}) = d(\vec{x})$ for all $\vec{x} \in A$ but $G^t(e)(\vec{x}) \neq G^t(c)(\vec{x})$ for some $\vec{x} \in B$. This means that arbitrarily distant modifications to any configuration c may propagate to a fixed observation window B.

Topological entropy The topological entropy $h(G)$ of a CA G measures the complexity of its dynamics. For any finite $A \subseteq \mathbb{Z}^d$ and positive integer n we define the equivalence relation $\equiv_{A,n}$ among initial configurations as follows: For all $c, e \in S^{\mathbb{Z}^d}$

$$c \equiv_{A,n} e \Leftrightarrow \forall \vec{x} \in A, 0 \leq t < n: G^t(c)(\vec{x}) = G^t(e)(\vec{x}) .$$

In other words, two configurations are equivalent if we can not observe any difference in their orbits in region A within the first n time instances. Let us denote by $N_G(A, n)$ the number of equivalence classes of $\equiv_{A,n}$. Then the topological entropy is

$$h(G) = \sup_A \lim_{n\to\infty} \frac{\log N_G(A, n)}{n}$$

where the supremum is over all finite $A \subseteq \mathbb{Z}^d$. The entropy always exists. In the one-dimensional case the entropy is always a finite, non-negative number. If $d \geq 2$ then the entropy can also be infinite.

Definition of the Subject

We consider the following algorithmic questions concerning cellular automata. All problems are decision problems, that is, the answer for each input instance is either yes or no. All problems considered are undecidable, i. e. no algorithm can solve them. We only consider problems whose undecidability is proved using a reduction from the tiling problem or its variant.

Injectivity

Input: Cellular Automaton A
Question: Is A injective (i. e. reversible)?

There is an algorithm that solves Injectivity for one-dimensional CA [1]. But the problem is undecidable among two-dimensional CA [9,11]. The problem is semi-decidable in any dimension.

Surjectivity

Input: Cellular Automaton A
Question: Is A surjective?

Also Surjectivity is decidable among one-dimensional CA. The two-dimensional question is, however, undecidable [11]. The complement of the problem (i. e. non-surjectivity) is semi-decidable in any dimension.

Nilpotency

Input: Cellular Automaton A
Question: Is A nilpotent?

Nilpotency is undecidable even among one-dimensional CA [10]. It is undecidable even if A has a spreading state, i. e. a state q such that any cell whose neighborhood contains q becomes q. However, Nilpotency is semi-decidable in any dimension. Based on problem Nilpotency we can prove a Rice's theorem for Cellular Automata limit sets: any non-trivial decision problem concerning limit sets is undecidable [12].

Topological entropy

Input: Cellular Automaton A.
Question: Is the topological entropy of A less than constant $c > 0$?

Problem Topological entropy is undecidable for every constant $c > 0$, even in the one-dimensional case. This can be proved proved using a direct reduction from Nilpotency [8]. Also, direct reductions from Nilpotency prove the undecidability of the following two problems [5,14]:

Equicontinuity

Input: Cellular Automaton A.
Question: Is A equicontinuous?

Sensitivity to initial conditions

Input: Cellular Automaton A.
Question: Is A sensitive to initial conditions?

Introduction

Several decision problems concerning cellular automata are known to be undecidable, that is, no algorithm exists that solves them. Some undecidability results easily follow from the universal computation capabilities of cellular automata, while others require more elaborate proofs.

Reductions from the tiling problem and its variants turn out to be useful in proving various questions concerning CA undecidable. We consider the problems of determining if a given two-dimensional CA is surjective or injective, whether a one-dimensional CA is nilpotent or equicontinuous, and whether the topological entropy is less than some constant.

The fact that the tiling problem is closely related to cellular automata is not surprising considering their apparent similarity: both involve assignments of symbols over a finite alphabet onto integer lattice points. The difference is that tilings are static while cellular automata change the assignments dynamically according to the local rule. Undecidability of the tiling problem on the two-dimensional plane naturally leads to undecidability results concerning single step properties of two- and higher dimensional cellular automata. But also asymptotic properties of one-dimensional cellular automata can be related to tiling problems by viewing the space-time diagram of the CA as a tiling. This naturally leads to the definition of deterministic tile sets: Wang tiles where tiles are uniquely determined by some of their neighbors.

We start by discussing the tiling problem and its variants. We do not proof the undecidability of all the variants. Rather, literature references for the proofs are provided. We then define a particular tile set that has an interesting plane-filling property. This tile set is a useful tool in the reduction to prove that it is undecidable to tell whether a given two-dimensional CA is injective (reversible). We then provide reductions that show several questions concerning cellular automata undecidable.

The Tiling Problem and Its Variants

In this section we discuss the tiling problem and several of its variants.

Definition of Tiles

For our purposes it is convenient to define tiles in a way that most closely resembles cellular automata. In the d-dimensional cellular space the cells are indexed by $\mathbb{Z}^d$. A neighborhood vector

$$N = (\vec{n}_1, \vec{n}_2, \ldots, \vec{n}_m)$$

consists of m distinct elements $\vec{n}_i \in \mathbb{Z}^d$. Each $\vec{n}_i$ specifies the relative location of a neighbor of each cell. More precisely, the ith neighbor of the cell in position $\vec{x} \in \mathbb{Z}^d$ is located at $\vec{x} + \vec{n}_i$.

A tile set is a finite set T whose elements are called tiles. A local matching rule tells which patterns of tiles are allowed in valid tilings. The matching rule is given as an m-ary relation $R \subseteq T^m$ where m is the size of the neighborhood. Tilings are assignments

$$t\colon \mathbb{Z}^d \to T$$

of tiles into cells. Tiling t is valid at $\vec{x} \in \mathbb{Z}^d$ if

$$t(\vec{x} + \vec{n}_1, \vec{x} + \vec{n}_2, \ldots, \vec{x} + \vec{n}_m) \in R\,.$$

Tiling t is called valid if it is valid at every position $\vec{x} \in \mathbb{Z}^d$.

A convenient – and historically earlier – way of defining tiles is in terms of edge labelings. A Wang tile is a two-dimensional unit square with colored edges. The local matching rule is determined by these colors: A tiling is valid at position $\vec{x} \in \mathbb{Z}^2$ iff each of the four edges of the tile in position $\vec{x}$ have the same color as the abutting edge in the adjacent tile. Clearly this is a two-dimensional tile set with the neighborhood vector $[(-1,0),(1,0),(0,-1),(0,1)]$ and a particular way of defining the local relation R.

Computations and Tilings

The basic observation in establishing undecidability results concerning tilings is the fact that valid tilings can be forced to contain a complete simulation of a computation by a given Turing machine. To any given Turing machine $\mathcal{M} = (Q, \Gamma, \delta, q_0, q_h, b)$ we associate the Wang tiles shown in Fig. 1, and we call these tiles the *machine tiles* of $\mathcal{M}$. Note that in the illustrations, instead of colors, we use labeled arrows on the sides of the tiles. Two adjacent tiles match if and only if an arrow head meets an arrow tail with the same label. Such arrow representation can be converted into the usual coloring representation of Wang tiles by assigning to each arrow direction and label a unique color.

The machine tiles of $\mathcal{M}$ contain the following tiles:

(i) For every tape letter $a \in \Gamma$ a *tape tile* of Fig. 1a,

(ii) For every tape letter $a \in \Gamma$ and every state $q \in Q$ an *action tile* of Fig. 1b or c. Tile (b) is used if

$$\delta(q, a) = (q', a', -1)$$

and tile (c) is used if

$$\delta(q, a) = (q', a', +1).$$

(iii) For every tape letter $a \in \Gamma$ and non-halting state $q \in Q \setminus \{q_h\}$ two merging tiles shown in Fig. 1d.

The idea of the tiles is that a configuration of the Turing machine $\mathcal{M}$ is represented as a row of tiles in such a way that the cell currently scanned by $\mathcal{M}$ is represented by an

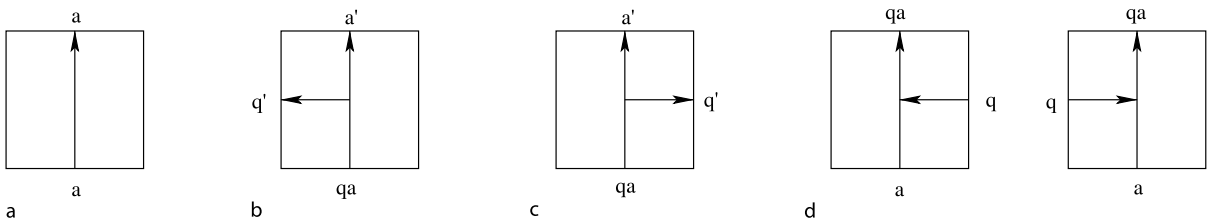

Tiling Problem and Undecidability in Cellular Automata, Figure 1
Machine tiles associated to a Turing machine

action tile, its neighbor where the machine moves into has a merging tile and all other tiles on the row are tape tiles. If this row is part of a valid tiling then it is clear that the rows above must be similar representations of subsequent configurations in the Turing machine computation, until the machine halts.

The machine tiles above are the basic tiles associated to Turing machine $\mathcal{M}$. Additional tiles will be added depending on the actual variant of the tiling problem.

The Tiling Problem

The tiling problem is the decision problem of determining if at least one valid tiling is admitted by the given set of tiles.

TILING PROBLEM
Input: Tile set $\mathcal{T}$.
Question: Does $\mathcal{T}$ admit a valid tiling?

The tiling problem is easily seen decidable if the input is restricted to one-dimensional tile sets. It is classical result by R. Berger that the tiling problem of two-dimensional tiles is undecidable, even if the input consists of Wang tiles [2,21]:

Theorem 1 TILING PROBLEM *is undecidable for Wang tile sets $\mathcal{T}$. The complement problem (non-existence of valid tilings) is semi-decidable.*

We do not prove this result here. The undecidability proofs in [2,21] are based on an explicit construction of an aperiodic tile set such that additional tiles implementing Turing machine simulations can be embedded in valid tilings. The aperiodic set is needed to force the presence of tiles that initiate Turing machine simulation in arbitrarily large regions.

Note that semi-decidability of the complement problem is apparent: a semi-algorithm simply tries to tile larger and larger regions until (if ever) a region is found that can not be properly tiled. Note also that a semi-algorithm exist for those tile sets that admit a valid, totally periodic tiling: All totally periodic tilings can be effectively enumerated and it is a simple matter to test each for validity of the tiling constraint. Combining the two semi-algorithms above yields a semi-algorithm that correctly identifies tile sets that (i) do not admit any valid tiling, or (ii) admit a valid periodic tiling. Only aperiodic tile sets fail to satisfy either (i) or (ii), so we see that the existence of aperiodic tile sets is implied by Theorem 1.

In the following sections we consider some variants of the tiling problem whose undecidability is easier to establish.

Variants of the Tiling Problem

TILING PROBLEM WITH A SEED TILE
Input: Tile set $\mathcal{T}$ and one tile s.
Question: Does $\mathcal{T}$ admit a valid tiling such that tile s is used at least once?

The seeded version was shown undecidable by H. Wang [24]. We present the proof here because the proof is quite simple and shows the general idea of how Turing machine halting problem can be reduced to problems concerning tiles:

Theorem 2 TILING PROBLEM WITH A SEED TILE *is undecidable for Wang tile sets. The complement problem is semi-decidable.*

Proof The semi-decidability of the complement problem follows from the following semi-algorithm: For $r = 1, 2, 3, \ldots$ try all tilings of the radius r square around the origin to see if there is a valid tiling of the square such that the origin contains the seed tile s. If for some r such a tiling is not found then halt and report that there is no tiling containing the seed tile.

Consider then undecidability. We reduce the decision problem TURING MACHINE HALTING ON BLANK TAPE, a problem that is well known to be undecidable. For any given Turing machine $\mathcal{M}$ we can effectively construct a tile set and a seed tile in such a way that they form a positive in-

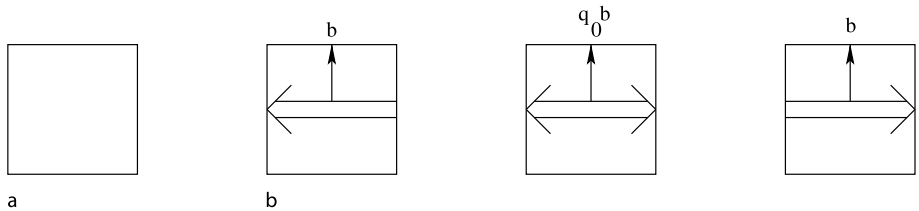

Tiling Problem and Undecidability in Cellular Automata, Figure 2
a the blank tile, and b three initialization tiles

stance of TILING PROBLEM WITH A SEED TILE if and only if $\mathcal{M}$ is a negative instance of TURING MACHINE HALTING ON BLANK TAPE. For the given Turing machine $\mathcal{M}$ we construct the machine tiles of Fig. 1 as well as the four tiles shown in Fig. 2. These are the blank tile and three initialization tiles. They initialize all tape symbols to be equal to blank b, and the Turing machine to be in the initial state q_0. The middle initialization tile is chosen as the seed tile s.

Let us prove that a valid tiling containing a copy of the seed tile exists if and only if the Turing machine M does not halt when started on the blank tape:
"$\Longleftarrow$": Suppose that the Turing machine $\mathcal{M}$ does not halt on the blank tape. Then a valid tiling exists where one horizontal row is formed with the initialization tiles, all tiles below this row are blank, and the rows above the initialization row contain consecutive configurations of the Turing machine.
"$\Longrightarrow$": Suppose that a valid tiling containing the middle initialization tile exists. The seed tile forces its row to be formed by the initialization tiles, representing the initial configuration of the Turing machine on the blank tape. The machine tiles force the following horizontal rows above the seed row to contain the consecutive configurations of the Turing machine. There is no merge tile containing a halting state so the Turing machine does not halt – otherwise a valid tiling could not be formed.

Conclusion: Suppose we had an algorithm that solves TILING PROBLEM WITH A SEED TILE. Then we also have an algorithm (which simply constructs the tile set as above and determines if a tiling with seed tile exists) that solves TURING MACHINE HALTING ON BLANK TAPE. This contradicts the fact that this problem is known to be undecidable. □

In the following tiling problem variant we are given a Wang tile set T and specify one tile $B \in T$ as the *blank tile*. The blank tile has all four sides colored by the same color. A *finite tiling* is a tiling where only a finite number of tiles are non-blank. A finite tiling where all tiles are blank is called *trivial*.

FINITE TILING PROBLEM
Instance: A finite set T of Wang tiles and a blank tile $B \in T$
Problem: Does there exist a valid finite tiling that is not trivial?

Theorem 3 *The* FINITE TILING PROBLEM *is undecidable. It is semi-decidable while its complement is not semi-decidable.*

Proof For semi-decidability notice that we can try all valid tilings of larger and larger squares until we find a tiling of a square where all tiles on the boundary are blank, while some interior tile is different from the blank tile. If such a tiling is found then the semi-algorithm halts, indicating that a valid, finite, non-trivial tiling exists.

To prove the undecidability we reduce the problem TURING MACHINE HALTING ON BLANK TAPE. For any given Turing machine $\mathcal{M}$ we construct the machine tiles of Fig. 1 as well as the blank tile, boundary tiles and the halting tiles shown in Fig. 3.

The halting tiles of Fig. 3b are constructed for all tape letters $a \in \Gamma$ and the halting state q_h. The purpose of the halting tiles is to erase the Turing machine from the configuration once it halts. The lower border tiles of Fig. 3c initialize the configuration to consist of the blank tape symbol b and the initial state q_0. The top border tiles are made for every tape symbol $a \in \Gamma$. They allow the absorption of the configuration as long as the Turing machine has been erased. The border tiles on the sides are labeled with symbols L and R to identify the left and the right border of the computation area.

Let us prove that the tile set admits a valid, finite, non-trivial tiling if and only if the Turing machine halts on the empty tape.
"$\Longleftarrow$": Suppose that the Turing machine halts on the blank tape. Then a tiling exists where the boundary tiles isolate a finite portion of the plane (a "board") for the simulation of the Turing machine, the bottom tiles of the board initialize the Turing machine on the blank tape, and

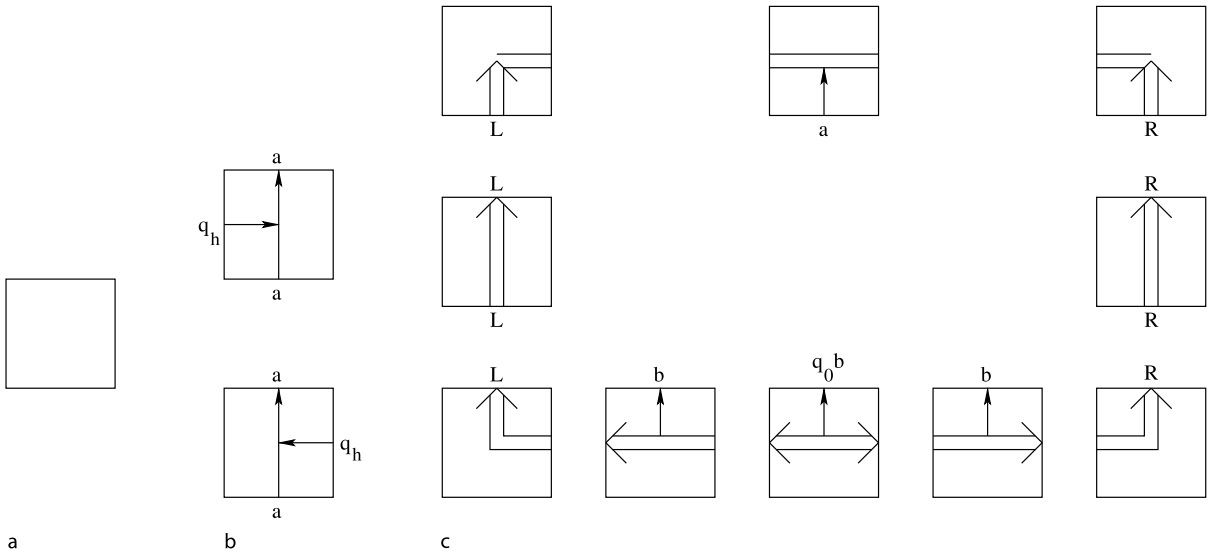

Tiling Problem and Undecidability in Cellular Automata, Figure 3
a the blank tile *B*, b halting tiles, and c border tiles

inside the board the Turing machine is simulated until it halts. After halting only tape tiles are used until they are absorbed by the topmost row of the board. If the board is made sufficiently large the entire computation fits inside the board, so the tiling is valid. All tiles outside the board are blank so the tiling is finite.

"$\Longrightarrow$": Suppose then that a finite, non-trivial tiling exists. The only non-blank tiles with a blank bottom edge are the lower border tiles of Fig. 3c, so the tiling must contain a lower border tile. Horizontal neighbors of lower border tiles are lower border tiles, so we see that the only way to have a finite tiling is to have a contiguous lower border that ends at both sides in a corner tile where the border turns upwards. The vertical borders must again – due to the finiteness of the tiling – end at corners where the top border starts. All in all we see that the boundary tiles are forced to form a rectangular board.

The lower boundary of the board initializes the Turing machine configuration on the blank tape, and the rows above it are forced by the machine tiles to simulate consecutive configurations of the Turing machine. Because the Turing machine state symbol is not allowed to touch the side or the upper boundary of the board, the Turing machine must be erased by a halting tile, i. e. the Turing machine must halt. □

The third variation of the tiling problem we consider is the PERIODIC TILING PROBLEM where we ask whether a given set of tiles admits a valid periodic tiling.

PERIODIC TILING PROBLEM
Input: Tile set $\mathcal{T}$.
Question: Does $\mathcal{T}$ admit a valid periodic tiling?

Theorem 4 *The* PERIODIC TILING PROBLEM *is undecidable for Wang tile sets. It is semi-decidable while its complement is not semi-decidable.*

For a proof, see [7].

Deterministic Tiles

The tiling problem of one-dimensional tiles is decidable. However, tiles can provide undecidability results for one-dimensional CA when we use the trick that we view space-time diagrams as two-dimensional tilings. But not every tiling can be a space-time diagram of a CA: the tiling must be locally deterministic in the direction that corresponds to time. This leads to the consideration of determinism in Wang tiles.

Consider a set T of Wang tiles, i. e. squares with colored edges. We say that T is *NW-deterministic* if for all $a, b \in T$, $a \neq b$, either the upper (=northern) edges of a and b or the left (=western) edges of a and b have different colors. See Fig. 4a for an illustration.

Consider now a valid tiling of the plane by NW-deterministic tiles. Each tile is uniquely determined by its left and upper neighbor. Then tiles on each diagonal in the NE-SW direction locally determine the tiles on the next diagonal below it. If we interpret these diagonals as con-

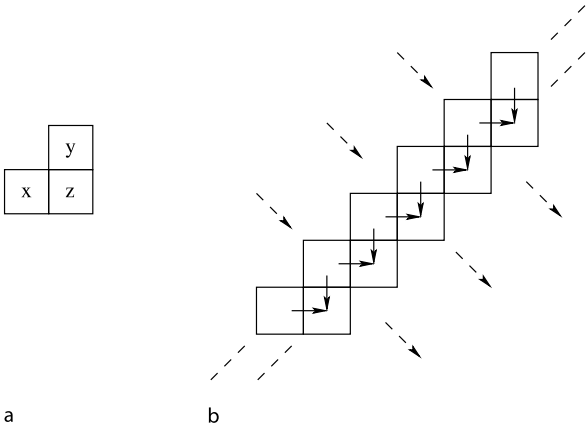

Tiling Problem and Undecidability in Cellular Automata, Figure 4
NW-deterministic sets of Wang tiles: a there is at most one matching tile *z* for any *x* and *y*, b diagonals of NW-deterministic tilings interpreted as configurations of one-dimensional CA

figurations of a CA then there is a local rule such that valid tilings are space-time diagrams of the CA, see Fig. 4b.

We define analogously *NE-*, *SW-* and *SE-deterministic* tile sets. Finally, we call a tile set 4-way deterministic if it is deterministic in all four directions simultaneously.

The tiling problem is undecidable among NW-deterministic tile sets [10], even among 4-way deterministic tile sets [18]:

Theorem 5 *The decision problem* TILING PROBLEM *is undecidable among 4-way deterministic sets of Wang tiles.*

As discussed at the end of Sect. "The Tiling Problem", the theorem also means that 4-way deterministic aperiodic tile sets exist. In fact, the proof of Theorem 5 in [18] uses one such aperiodic set that was reported in [16].

Plane Filling Directed Tiles

A d-dimensional *directed tile* is a tile that is associated a *follower vector* $\vec{f} \in \mathbb{Z}^d$. Let $\mathcal{T} = (d, T, N, R)$ be a tile set, and let $F\colon T \to \mathbb{Z}^d$ be a function that assigns tiles their follower vectors. We call $\mathcal{D} = (d, T, N, R, F)$ a set of directed tiles. Let $t \in T^{\mathbb{Z}^d}$ be an assignment of tiles to cells. For every $\vec{p} \in \mathbb{Z}^d$ we call $\vec{p} + F(t(\vec{p}))$ the follower of $\vec{p}$ in t. In other words, the follower of $\vec{p}$ is the cell whose position relative to $\vec{p}$ is given by the follower vector of the tile in cell $\vec{p}$.

Sequence $\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_k$ where all $\vec{p}_i \in \mathbb{Z}^d$ is a (finite) *path* in t if

$$\vec{p}_{i+1} = \vec{p}_i + F(t(\vec{p}_i))$$

for all $1 \le i < k$. In other words, a path is a sequence of cells such that the next cell is always the follower of the previous cell. One-way infinite and two-way infinite paths are defined analogously.

In the following we only discuss the two-dimensional case ($d = 2$) and the follower of each tile is one of the four adjacent positions:

$$F(a) \in \{(\pm 1, 0), (0, \pm 1)\} \quad \text{for all } a \in T\,.$$

In this case the follower is indicated in drawings as a horizontal or vertical arrow over the tile.

A set of two-dimensional directed tiles is said to have the *plane-filling property* if it satisfies the following two conditions:

(a) There exists $t \in T^{\mathbb{Z}^2}$ and a one-way infinite path $\vec{p}_1, \vec{p}_2, \vec{p}_3, \ldots$ such that the tiling in t is valid at $\vec{p}_i$ for all $i = 1, 2, 3, \ldots$.
(b) For all t and $\vec{p}_1, \vec{p}_2, \vec{p}_3, \ldots$ as in (a), there are arbitrarily large $n \times n$ squares of cells such that all cells of the squares are on the path.

Intuitively the plane-filling property means that the simple device that moves over tiling t, repeatedly verifies the correctness of the tiling it its present location and moves on to the follower, necessarily eventually either finds a tiling error or covers arbitrarily large squares. Note that the plane-filling property does not assume that the tiling t is correct everywhere: as long as it is correct along a path the path must snake through larger and larger squares.

Note that conditions (a) and (b) imply that the tile set is aperiodic. There exist tile sets that satisfy the plane filling-filling property, as proved in [11]:

Theorem 6 *There exists a set of directed Wang tiles that has the plane-filling property.*

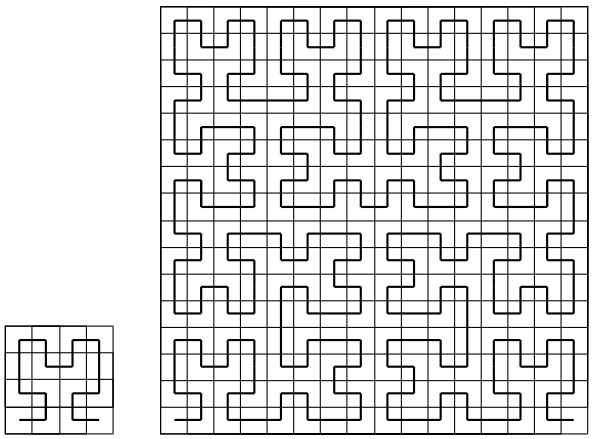

Tiling Problem and Undecidability in Cellular Automata, Figure 5
Fractions of the plane-filling Hilbert curve through 4 × 4 and 16 × 16 squares

The proof of Theorem 6 in [11] constructs a set of Wang tiles such that the path that does not find any tiling errors is forced to follow the well known Hilbert-curve shown in Fig. 5.

Undecidability in Cellular Automata

Let us begin with one-step properties of two-dimensional CA.

Theorem 7 INJECTIVITY *is undecidable among two-dimensional CA. It is semi-decidable in any dimension.*

Proof The semi-decidability follows from the fact that injective CA have an inverse CA. One can effectively enumerate all CA and check them one-by-one until (if ever) the inverse CA is found.

Let us next prove INJECTIVITY undecidable by reducing the TILING PROBLEM into it. In the reduction a set D of directed tiles that has the plane filling property is used. The existence of such D was stated in Theorem 6.

Let T be a given set of Wang tiles that is an instance of the TILING PROBLEM. One can effectively construct a two-dimensional CA whose state set is

$$S = T \times D \times \{0, 1\}$$

and the local rule updates the bit component of a cell as follows:

- If either the T- or the D-components contain a tiling error at the cell then the state of the cell is not changed, but
- if the tilings according to both T- and D-components are valid at the cell then the bit of the follower cell (according to the direction in the D-component) is added to the present bit value modulo 2.

The tile components are not changed. Let us prove that this CA G is not injective if and only if T admits a valid tiling.

"$\Longleftarrow$": Suppose a valid tiling exists. Construct two configurations c_0 and c_1 where the T- and D-components form the same valid tilings $t \in T^{\mathbb{Z}^2}$ and $d \in D^{\mathbb{Z}^2}$, respectively. In c_0 all bits are 0 while in c_1 they are all 1. Since the tilings are everywhere valid, every cell performs modulo 2 addition of two bits, which means that every bit becomes 0. Hence $G(c_0) = G(c_1) = c_0$, and G is not injective.

"$\Longrightarrow$": Suppose then that G is not injective. There are two different configurations c_0 and c_1 such that $G(c_0) = G(c_1)$. Tile components are not modified by the CA so they are identical in c_0 and c_1. There is a cell $\vec{p}_1$ such that c_0 and c_1 have different bits at cell $\vec{p}_1$. Since these bits become identical in the next configuration, the D-tiling must be correct at $\vec{p}_1$ and c_0 and c_1 must have different bits in the follower position $\vec{p}_2$. We repeat the reasoning and obtain an infinite sequence of positions $\vec{p}_1, \vec{p}_2, \vec{p}_3, \ldots$ such that each $\vec{p}_{i+1}$ is the follower of $\vec{p}_i$, and the D tiling is correct at each $\vec{p}_i$. It follows from the plane filling property of D that path $\vec{p}_1, \vec{p}_2, \vec{p}_3, \ldots$ covers arbitrarily large squares. Also the tiling according to the T-components must be valid at each cell of the path. Hence tile set T admits valid tilings of arbitrarily large squares, and therefore it admits a valid tiling of the entire plane. □

Analogously we can prove the undecidability of SURJECTIVITY. It is convenient to use the well known Garden-of-Eden theorem of Moore and Myhill to convert the surjectivity property into injectivity on finite configurations:

Theorem 8 (Garden-of-Eden theorem) *A cellular automaton is non-surjective if and only if there are two distinct configurations that differ in a finite number of cells and that have the same successor* [19,20].

Theorem 9 SURJECTIVITY *is undecidable among two-dimensional CA. It's complement is semi-decidable in any dimension.*

Proof A semi-algorithm for non-surjectivity enumerates all finite patterns one-by-one until a pattern is found that can not appear in $G(c)$ for any configuration c.

To prove undecidability we reduce the FINITE TILING PROBLEM, using the set D of 23 directed tiles shown in Fig. 6. These directed tiles are used in an analogous way as in the proof of Theorem 7.

The topmost tile in Fig. 6 is called blank. All other tiles have a unique incoming and outgoing arrow. In valid tilings arrows and labels must match. The non-blank tiles are considered directed: the follower of a tile is the neighbor directed to by the outgoing arrow on the tile. Since each non-blank tile has exactly one incoming arrow, it is clear that if the tiling is valid at a tile then the tile is the follower of exactly one of its four neighbors.

The tile at the center in Fig. 6 where the dark and light thick horizontal lines meet is called the *cross*. It has a special role in the forthcoming proof. A *rectangular loop* is a valid tiling of a rectangle using tiles in D where the follower path forms a loop that visits every tile of the rectangle, and the outside border of the rectangle is colored blank. See Fig. 7 for an example of a rectangular loop through a rectangle of size 12×7. (The edge labels are not shown for the sake of clarity of the figure.) It is easy to see that a rectangular loop of size $2n \times m$ exist for all $n \geq 2$ and $m \geq 3$. Any tile in an even column in the interior of

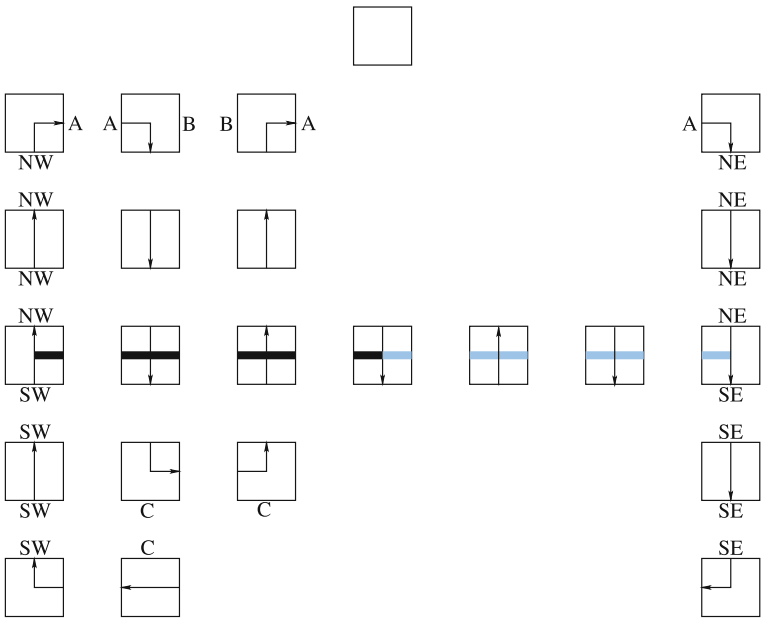

Tiling Problem and Undecidability in Cellular Automata, Figure 6
Tiles used in the proof of the undecidability of SURJECTIVITY

the rectangle can be made to contain the unique cross of the rectangular loop.

It is easy to see that the tile set D has the following property:
Finite plane-filling property: Let $t \in D^{\mathbb{Z}^2}$ be a tiling, and $\vec{p}_1, \vec{p}_2, \vec{p}_3, \ldots$ a path in t such that the tiling t is valid at $\vec{p}_i$ for all $i = 1, 2, 3, \ldots$. If the path covers only a finite number of different cells then the cells on the path form a rectangular loop.

Let b and c be the blank and the cross of set D. For any given tile set T with blank tile B we construct the following two-dimensional cellular automaton. The state set S contains triplets

$$(d, t, x) \in D \times T \times \{0, 1\}$$

under the following constraints:

- If $d = c$ then $t \neq B$, and
- if $d = b$ or d is any tile containing label SW, SE, NW, NE, A, B or C, then $t = B$.

In other words, the cross must be associated with a non-blank tile in T while the blank of D as well as all the tiles on the boundary of a rectangular loop are forced to be associated with the blank tile of T. The triplet $(b, B, 0)$ where both tile components are blank and the bit is 0 is the quiescent state of the CA. The local rule is as follows: Let (d, t, x) be the current state of a cell.

- If $d = b$ then the state is not changed.
- If $d \neq b$ then the cell verifies the validity of the tilings according to both D and T at the cell. If either tile component has a tiling error then the state is not changed. If both tilings are valid then the cell modifies its bit component by adding the bit of its follower modulo 2.

Let us prove that this CA is not surjective if and only if T admits a valid, finite, non-trivial tiling.
"$\Longleftarrow$": Suppose a valid, finite, non-trivial tiling $t \in T^{\mathbb{Z}^2}$ exists. Consider a configuration of the CA whose T-components form the valid tiling t and the D-components form a rectangular loop whose interior covers all non-blank elements of t. Tiles outside the rectangle are all blank and have bit 0. The cross can be positioned so that it is in the same cell as some non-blank tile in t. In such a configuration both T and D tilings are everywhere valid. The CA updates the bits of all tiles in the rectangular loop by performing modulo 2 addition with their followers, while the bits outside the rectangle remain 0. We get two different configurations that have the same image: In c_0 all bits in the rectangle are equal to 0 while in c_1 they are all equal to 1. The local rule updates the bits so that $G(c_0) = G(c_1) = c_0$. Configurations c_0 and c_1 only differ in a finite number of cells, so it follows from the Garden-of-Eden theorem that G is not surjective.
"$\Longrightarrow$": Suppose then that the CA is not surjective. According to the Garden-of-Eden theorem there are two finitely

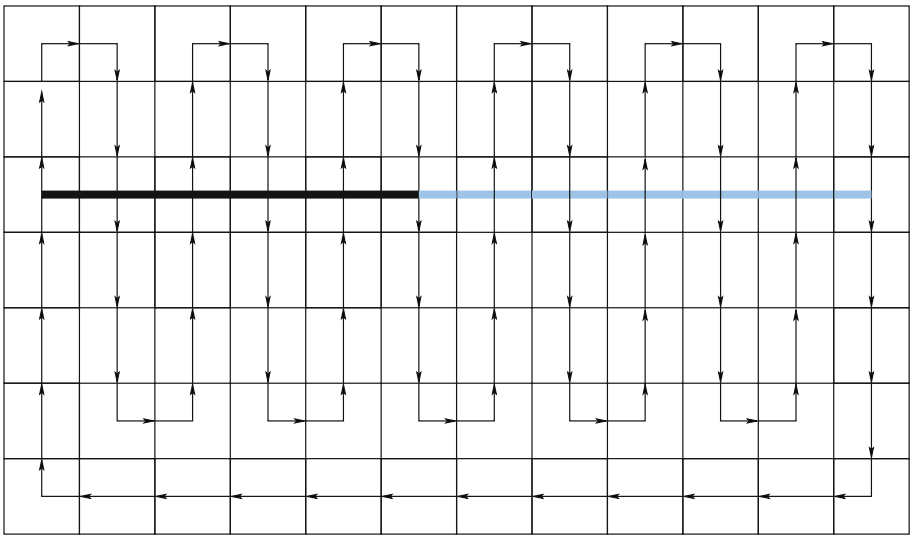

Tiling Problem and Undecidability in Cellular Automata, Figure 7
A rectangular loop of size 12 × 7

different configurations c_0 and c_1 such that $G(c_0) = G(c_1)$. Since only bit components of states are changed, the tilings in c_0 and c_1 according to D- and T-components of the states are identical. There is a cell $\vec{p}_1$ such that c_0 and c_1 have different bits at cell $\vec{p}_1$. Since these bits become identical in the next configuration, the D-tiling must be correct at $\vec{p}_1$ and c_0 and c_1 must have different bits in the follower position $\vec{p}_2$. We repeat the reasoning and obtain an infinite sequence of positions $\vec{p}_1, \vec{p}_2, \vec{p}_3, \ldots$ such that each $\vec{p}_{i+1}$ is the follower of $\vec{p}_i$, and the D tiling is correct at each $\vec{p}_i$. Moreover, c_0 and c_1 have different bits in each position $\vec{p}_i$. Because configurations c_0 and c_1 only differ in a finite number of cells, we see that the path can only contain a finite number of distinct cells. It follows then from the finite plane-filling property of D that the path must form a valid rectangular loop.

Also the tiling according to the T-components must be valid at each cell of the path. Because of the constraints on the allowed triplets, the T-components on the boundary of the rectangle are the blank B, while the cross in the interior contains a non-blank element of T. Hence there is a valid tiling of a rectangle according to T that contains a non-blank tile and has a blank boundary. This means that a finite, valid and non-trivial tiling is possible. □

Undecidable Properties of One-Dimensional CA

Using deterministic Wang tiles and interpreting space-time diagrams as tilings one obtains undecidability results for long-term behavior of one-dimensional CA.

Theorem 10 NILPOTENCY *is undecidable among one-dimensional CA. It is undecidable even among one-dimensional CA that have a spreading state q, i. e. a state that spreads to all neighbors.* NILPOTENCY *is semi-decidable in any dimension.*

Proof For semi-decidability notice that, for $n = 1, 2, 3 \ldots$, we can effectively construct a cellular automaton whose global function is G^n and check whether the local rule of the CA maps everything into the same state. If that happens for some n then we halt and report that the CA is nilpotent.

To prove undecidability we reduce the TILING PROBLEM of NW-deterministic Wang tiles. Let T be a given NW-deterministic tile set. One can effectively construct a one-dimensional CA whose state set is $S = T \cup \{q\}$ and the local rule turns a cell into the quiescent state q except in the case that the cell and its right neighbor are in states $x, y \in T$, respectively, and tile $z \in T$ exists so that tiles $x\,y, z$ match as in Fig. 4a. In this case z is the new state of the cell. Note that state q is a spreading state.

Let us prove that the CA is not nilpotent if and only if T admits a valid tiling.

"$\Longleftarrow$": Suppose a valid tiling exists. If $c \in T^{\mathbb{Z}}$ is a diagonal of this tiling then the configurations $G^n(c)$ in its orbit are subsequent diagonals of the same tiling, for all $n = 1, 2, \ldots$. This means that c never becomes quiescent, and the CA is not nilpotent.

"$\Longrightarrow$": Suppose no valid tiling exists. Then there is number n such that no valid tiling of an $n \times n$ square exists.

This means that for every initial configuration $c \in S^{\mathbb{Z}}$ the configuration $G^{2n}(c)$ is quiescent: If it is not quiescent then a valid tiling of an $n \times n$ square can be read from the space time diagram of configurations $c, G(c), \ldots, G^{2n}(c)$. We conclude that the CA is nilpotent. □

Undecidability of NILPOTENCY has some interesting corollaries. First, it implies that the topological entropy of a one-dimensional CA cannot be calculated, not even approximated [8].

Theorem 11 TOPOLOGICAL ENTROPY *is undecidable.*

Proof Let us reduce NILPOTENCY. Let $c > 0$ be any constant, and let $n > 2^c$ be an integer. For any given one-dimensional CA G with state set S and a spreading state $q \in S$ construct a new CA whose state set is $S \times \{1, 2, \ldots, n\}$, and the local rule applies G in the first components of the states and shifts the second components one cell to the left. In addition, state (q, i) is turned into state $(q, 1)$.

If G is nilpotent then also the new CA is nilpotent and its topological entropy is 0. If G is not nilpotent then there is a configuration $c \in S^{\mathbb{Z}}$ such that no cell ever turns into the spreading state q. But then the second components form a left shift over the alphabet $\{1, 2, \ldots, n\}$ so the topological entropy is at least $\log_2 n > c$. □

It also follows that is undecidable to determine if a given one-dimensional CA is ultimately periodic [5]:

Theorem 12 EQUICONTINUITY *is undecidable among one-dimensional CA.*

Proof Among one-dimensional CA with a spreading state EQUICONTINUITY is equivalnet to NILPOTENCY. □

Theorem 13 SENSITIVITY TO INITIAL CONDITIONS *is undecidable among one-dimensional CA.*

Proof Originally the result was proved in [5] using an elaborate reduction of the Turing machine halting problem. However, undecidability of NILPOTENCY provides the result directly, as pointed out in [14]. Namely, a one-dimensional cellular automaton whose neighborhood vector contains only strictly positive numbers is either nilpotent or sensitive. Adding a constant to all elements of the neighborhood vector does not affect the nilpotency status of a CA. So for any given one-dimensional CA we proceed as follows: add a positive constant to the elements of the neighborhood vector so that they all become positive. The new CA is sensitive if and only if the original CA was not nilpotent. The result then follows from Theorem 10. □

As a final application of undecidability of NILPOTENCY consider other questions concerning the limit set (=maximal attractor) of one-dimensional CA. One can show that NILPOTENCY can be reduced to any non-trivial question [12]. More precisely, let PROB be a decision problem that takes arbitrary one-dimensional CA as input. Suppose that PROB always has the same answer for any two CA that have the same limit set. Then we say that PROB is a decision problem concerning the limit sets of CA. We call PROB non-trivial if there exist both positive and negative instances.

Theorem 14 *Let* PROB *be any non-trivial decision problem concerning the limit sets of CA. Then* PROB *is undecidable* [12].

Other Undecidability Results

In the previous sections we only considered decision problems that have been proved undecidable using reductions from the tiling problem or its variant. There are many other decision problems that have been proved undecidable using other techniques. Below are a few, with literature references.

We call a CA G *periodic* if there is number n such G^n is the identity function. This is equivalent to saying that every configuration is periodic, that is, every configuration returns back to itself. Clearly a periodic CA is necessarily injective. In fact, periodic CA are exactly those CA that are injective and equicontinuous.

PERIODICITY
Input: Cellular Automaton A
Question: Is A periodic?

Theorem 15 PERIODICITY *is undecidable among one-dimensional CA* [15].

A CA is called sensitive to initial conditions if there exists a finite set $B \subseteq \mathbb{Z}^d$ of cells such that for every configuration c and every finite set $A \subseteq \mathbb{Z}$ of cells there exists a configuration e and time $t \geq 0$ such that $e(\vec{x}) = d(\vec{x})$ for all $\vec{x} \in A$ but $G^t(e)(\vec{x}) \neq G^t(c)(\vec{x})$ for some $\vec{x} \in B$.

SENSITIVITY TO INITIAL CONDITIONS
Input: Cellular Automaton A
Question: Is A sensitive to initial conditions?

Theorem 16 SENSITIVITY TO INITIAL CONDITIONS *is undecidable among one-dimensional CA* [5].

The following problems deal with dynamics on finite configuration. We hence suppose that the given CA has a quiescent state, i.e. a state q such that $f(q, q, \ldots, q) = q$ where f is the local update rule of the CA. A configuration $c \in S^{\mathbb{Z}^d}$ is called finite (w.r.t. q) if all but a finite number

of cells are in state q. Questions similar to NILPOTENCY and EQUICONTINUITY can be asked in the space of finite configurations:

NILPOTENCY ON FINITE CONFIGURATIONS
Input: Cellular Automaton A with a quiescent state
Question: Does every finite configuration evolve into the quiescent configuration?

EVENTUAL PERIODICITY ON FINITE CONFIGURATIONS
Input: Cellular Automaton A with a quiescent state
Question: Does every finite configuration evolve into a temporally periodic configuration?

Theorem 17 NILPOTENCY ON FINITE CONFIGURATIONS *and* EVENTUAL PERIODICITY ON FINITE CONFIGURATIONS *are undecidable for one-dimensional CA* [4,23].

Future Directions

Several interesting and challenging open questions remain. In particular, the decidability statuses of the following decision problems concerning basic dynamical properties are unknown.

We call a CA G *periodic* if there is number n such G^n is the identity function. This is equivalent to saying that every configuration is periodic, that is, every configuration returns back to itself. Clearly a periodic CA is necessarily injective. In fact, periodic CA are exactly those CA that are injective and equicontinuous.

PERIODICITY
Input: Cellular Automaton A
Question: Is A periodic?

The question is undecidable for two-dimensional CA (the construction in the proof of Theorem 7 shows it) but the decidability status is unknown for one-dimensional CA.

We call a one-dimensional CA G *positively expansive* if there exists a finite set $A \subseteq \mathbb{Z}$ of cells such that for any two distinct configurations c and e there exists $t \geq 0$ such that $G^t(c)$ and $G^t(e)$ differ in some cell in A. We call an injective CA G *expansive* for some finite A holds the following: for any two distinct configurations c and e there exists $t \in \mathbb{Z}$ such that $G^t(c)$ and $G^t(e)$ differ in some cell in A. It is known that a two- and higher dimensional CA can not be expansive or positively expansive [6,22], so the following decision problems are only asked for one-dimensional CA:

POSITIVE EXPANSIVITY
Input: One-dimensional cellular Automaton A
Question: Is A positively expansive?

EXPANSIVITY
Input: One-dimensional cellular Automaton A
Question: Is A expansive?

The decidability status of both POSITIVE EXPANSIVITY and EXPANSIVITY is unknown.

Acknowledgments

Research supported by the Academy of Finland grant 211967.

Bibliography

Primary Literature

1. Amoroso S, Patt Y (1972) Decision Procedures for Surjectivity and Injectivity of Parallel Maps for Tessellation Structures. J Comput Syst Sci 6:448–464
2. Berger R (1966) The Undecidability of the Domino Problem. Mem Am Math Soc 66:1–72
3. Culik K II (1996) An aperiodic set of 13 Wang tiles. Discret Math 160: 245–251
4. Culik K II, Yu S (1988) Undecidability of CA Classification Schemes. Complex Syst 2:177–190
5. Durand B, Formenti E, Varouchas G (2003) On undecidability of equicontinuity classification for cellular automata. In: Proceedings of Discrete Models for Complex Systems, Lyon, France, 16–19 June 2003, pp. 117–128
6. Finelli M, Manzini G, Margara L (1998) Lyapunov Exponents versus Expansivity and Sensitivity in Cellular Automata. J Complex 14:210–233
7. Gurevich YS, Koryakov IO (1972) Remarks on Berger's paper on the domino problem. J Sib Math J 13:319–321
8. Hurd LP, Kari J, Culik K (1992) The topological Entropy of Cellular Automata is Uncomputable. Ercodic Theor Dyn Syst 12:255–265
9. Kari J (1990) Reversibility of 2D cellular automata is undecidable. Physica D 45:379–385
10. Kari J (1992) The nilpotency problem of one-dimensional cellular automata. SIAM J Comput 21:571–586
11. Kari J (1994) Reversibility and surjectivity problems of cellular automata. J Comput Syst Sci 48:149–182
12. Kari J (1994) Rice's theorem for the Limit Sets of Cellular Automata. Theoret Comput Sci 127:229–254
13. Kari J (1996) A small aperiodic set of Wang tiles. Discret Math 160:259–264
14. Kari J (2008) Undecidable properties on the dynamics of reversible one-dimensional cellular automata. In: Proceedings of Journées Automates Cellulaires, Uzés, France, 21–25 April 2008
15. Kari J, Ollinger N Periodicity and immortality in reversible computing. (in press)
16. Kari J, Papasoglu P (1999) Deterministic aperiodic tile sets. J Geom Funct Anal 9:353–369
17. Kurka P (1997) Languages, equicontinuity and attractors in cellular automata. Ergod Theory Dyn Syst 17:417–433
18. Lukkarila V (2007) The 4-way deterministic tiling problem is undecidable. (in press)

19. Moore EF (1962) Machine models of self-reproduction. In: Proceedings of the Symposia in Applied Mathematics 14:17–33
20. Myhill J (1963) The Converse to Moore's Garden-of-Eden Theorem. In: Proceedings of the American Mathematical Society 14:685–686
21. Robinson RM (1971) Undecidability and Nonperiodicity for Tilings of the plane. Invent Math 12:177–209
22. Shereshevsky MA (1993) Expansiveness, entropy and polynomial growth for groups acting on subshifts by automorphisms. Indag Math 4:203–210
23. Sutner K (1989) A note on the Culik-Yu classes. Complex Syst 3:107–115
24. Wang H (1961) Proving theorems by pattern recognition – II. Bell Syst Techn J 40:1–42

Books and Reviews

Codd EF (1968) Cellular Automata. Academic Press, New York
Garzon M (1995) Models of massive parallelism: analysis of cellular automata and neural networks. Springer, New York
Hedlund G (1969) Endomorphisms and automorphisms of shift dynamical systems. Math Syst Theory 3:320–375
Kari J (2005) Theory of cellular automata: a survey. Theoret Comput Sci 334:3–33
Toffoli T, Margolus N (1987) Cellular Automata Machines. MIT Press, Cambridge
Wolfram S (ed) (1986) Theory and Applications of Cellular Automata. World Scientific Press, Singapore
Wolfram S (2002) A New Kind of Science. Wolfram Media, Canada

Tomography, Seismic

Jose Pujol
Dept. of Earth Sciences, The University of Memphis, Memphis, USA

Article Outline

Definition of the Subject
Introduction
Fundamentals of X-ray Computerized Tomography
Arrival-Time Seismic Tomography
Solution of Ill-Posed Linear Problems
Examples
Future Directions
Bibliography

Definition of the Subject

Seismic tomography refers to a number of techniques designed to investigate the interior of the earth using arrival times and/or waveforms from natural and artificial sources. The most common product of a tomographic study is a velocity model, although other parameters, such as attenuation, are also studied. The importance of seismic tomography stems from two facts. One, it generally has higher resolution than that provided by other geophysical methods. Two, it provides information that (a) can help solve fundamental problems concerning the internal structure of the earth at a global scale, and (b) has been used in tectonic and seismic hazards studies at a local scale. Seismic tomography has also been applied to data collected in boreholes, but because of the high expenses associated with drilling, borehole tomography is relatively little used.

Introduction

In the most general terms, seismic tomography problems are inverse problems, and before the word "tomography" entered the seismological literature the term inversion was used. This change occurred as a consequence of the revolution in medical imaging caused by computerized (or computed) X-ray tomography (CT) (also known as computer assisted or axial, tomography, CAT), introduced in the 1970s. The importance of CT on seismic tomography was that it provided efficient numerical techniques for the solution of systems of equations with extremely large number of unknowns (10^5 or more), which allowed a great expansion of the seismological inverse problems that could be tackled. In addition, according to [1], changing the name inversion to tomography enhanced the "believability" of the inverse method.

Seismic tomography covers a wide range of scales, from the very small (e. g., borehole tomography, involving distances of a few kilometers at most) to the very large (i. e., whole earth tomography). Initially, seismic tomography involved body-wave arrival times, but later it was extended to include waveform and surface wave information. Arrival-time tomography is simpler and for this reason it is the most popular. Waveform tomography requires software for the computation of synthetic seismograms for comparison with the observed ones, which makes the whole inversion process more difficult both theoretically and practically. Surface wave tomography involves the determination of phase velocities, which is more complicated than picking arrival times. The three approaches, however, have one thing in common, namely, they all end up requiring the solution of a linear system of equations. This task might sound simple, but in practice it is not because the system generally does not have a unique solution, which means that solving it requires making decisions (either implicitly or explicitly) regarding the nature of the solution. In addition, given the complexity of wave propagation in heterogeneous media (such as the earth), the in-

verse problems solved by seismologists involve highly simplified models of the earth. These simplifications enter into the linear system to be solved, which adds another layer of uncertainty to the solution. This is in sharp contrast with X-ray tomography, which is comparatively unaffected by this kind of uncertainties. For this reason it might be argued that the name tomography does not do justice to the kind of problems that seismologists solve, and this fact should be born in mind by readers from other disciplines. Other differences between X-ray and seismic tomography, of more practical nature, are given below.

The goal of most of the seismic tomography work is to derive 3-D velocity models of portions of the earth. Currently, most of the research concentrates on two scales, global and local. At the global scale, the tomographic models generally have higher resolution than that provided by other geophysical methods, and for this reason it has the potential to provide constraints on the fate of the subducted slabs, on models of mantle convection, on petrological and geochemical models, on studies of the geomagnetic and gravity fields, on mineral physics, and on the core-mantle boundary, among others (see, e. g., [41,96,129,161]). Local tomography usually involves the simultaneous determination of a 3-D velocity model and the relocation of the seismic events used to determine the model. Traditionally, local velocity models have been used in structural and tectonic interpretations, but more recently they have become important in seismic hazard studies, as the ground motion caused by earthquakes are amplified by low-velocity materials, which increases the hazard in the case of large events (see, e. g., [160]). As discussed below, the standard practice of locating earthquakes with layered velocity models may lead to significant location errors when the lateral velocity variations are significant. Therefore, the simultaneous velocity-inversion and event relocation has the potential to produce improved event locations, which is also important in the context of seismic hazards studies.

Fundamentals of X-ray Computerized Tomography

Historical Overview

Computerized tomography began in the early 1970s with the introduction of the X-ray scanner developed by G. Hounsfield [65,66], who was able to combine data generated using X-ray techniques with computerized data acquisition and processing techniques. This resulted in the first commercially viable CT machine, with clinical applications presented in [6]. The history of the development of CT is extremely interesting and will be summarized here, but to put it into a broader perspective we will review briefly the X-rays acquisition and processing techniques in use before CT was introduced.

Röntgen's discovery of X-rays in 1895 revolutionized the practice of medicine by allowing a view of the interior of the human body. However, an X-ray image is the 2-D projection of a 3-D body, which means that its interpretation is subject to ambiguity. This fact was noticed soon after the radiographic technique was introduced, and a number of researchers began to develop techniques to produce 3-D views (see, e. g., [155]). The goal was to generate X-ray images of thin slices of a patient. This was achieved by moving the film and the X-ray source in such a way that the objects in a particular plane were emphasized while others were blurred by the motion. This process is discussed in, e. g., [138]. Work on this goal began in 1914 in several European countries (Germany, France, Italy, Holland), and was motivated, at least in part, by World War I. Different approaches were tried and different names were assigned to some of them, but only one, tomography, remained. This word comes from the Greek word tomos, which means cut or slice, and was introduced by the Berliner physician G. Grossmann, whose tomograph was commercially available in 1935 [155,156].

Conventional tomography represented a considerable improvement over the original radiographic techniques, but the advent of computers opened new research avenues based on the digital processing of the X-ray images, or shadowgrams. A shadowgram is a photographic plate developed after it has been illuminated by X-rays that passed through an object and gives a measure of the absorptivity of the rays by the object, which in many materials is roughly proportional to its density [12]. The availability of computers allowed shadowgrams, as well as other images, to be scanned and digitized for further processing using Fourier transform techniques as well as numerical solution of matrix equations. This work will be referred to below, but it is important to realize that they were in place when CT was introduced, and were major contributors to the improvement of the quality of the early CT results.

One of the earliest papers on medical computerized tomography is by A. Cormack [26], who received the Nobel Prize in Physiology or Medicine in 1979 together with G. Hounsfield for their contributions to the development of the technique, which were carried out independently. Cormack was a South African physicist working at the University of Cape Town. In 1955 the physicist at the Cape Town hospital resigned and Cormack replaced him for six months in 1956. During this time he became interested in the determination of the absorption of X or gamma rays passing through an inhomogeneous medium. His motivation was medical, in the context of radiotherapy, which

required a good knowledge of the absorption coefficient of the bones and tissues of a patient. Cormack's approach was to consider a 2-D problem, as a 3-D one can be solved in terms of a succession of 2-D layers. The 2-D problem was well known and can be formulated as follows. A beam of monoenergetic rays of intensity I_o traverses a finite 2-D domain $\mathcal{D}$ along a straight line L, and the intensity of the ray emerging from $\mathcal{D}$ is I, given by

$$I = I_o \exp \int_L f(l)\mathrm{d}l \tag{1}$$

where f is the absorption coefficient, which is a function of position within $\mathcal{D}$, and $\mathrm{d}l$ denotes a length element along L. Dividing both sides of Eq. (1) by I_o and taking the natural logarithm gives

$$g_L \equiv \ln \frac{I}{I_o} = \int_L f(l)\mathrm{d}l \ . \tag{2}$$

The quantity g_L is known, and the question posed by Cormack was whether f could be computed using g_L determined for a number of lines. Cormack solved the problem in terms of a series expansion and demonstrated the feasibility of the method with two simple samples made of aluminum and either wood or Lucite [26,27]. His work, however, went essentially unnoticed until the publication of a paper [28] where the connection between the CT problem and the Radon transform was discussed. That paper in turn, was stimulated by Hounsfield's work, which became known in 1971 [29]. Cormack moved to Tufts University (United States) in 1957.

The actual implementation of the CT technique as a viable commercial enterprise is due to Hounsfield [65], who was an engineer working at the Central Research Laboratories of Electrical and Musical Industries, the English company that eventually became the well known musical records company EMI. As recounted by Hounsfield [67], after finding that a project he had been working on would not be commercially viable, he was given the opportunity to think about other areas of research that he thought might be more fruitful. This rare opportunity was probably the result of the considerable amount of money that the group *The Beatles* had brought to EMI [107]. One of the projects Hounsfield had been working on was connected with automatic pattern recognition, which in 1967 led to the idea of what eventually became the technique of computerized tomography. To materialize this idea Hounsfield built the CT machine and developed the numerical technique to process the data (see Subsect. "Iterative Solutions"). According to some, the second task may have been the more fundamental of the two [107].

We close this summary with two notes of historical interest. First, B. Korenblyum, S. Tetel'baum, and A. Tyutin worked on tomography at the Kiev Polytechnic Institute and published their results in obscure Russian journals in 1957 and 1958. These authors formulated the tomography problem using line integrals, solved it exactly in terms of the inverse Radon transform, and discussed a reconstruction algorithm. However, additional references on their work could not be found [11]. Second, W. Oldendorf, a neurologist at the University of California, Los Angeles, developed a scanning instrument, although he did not solve the problem mathematically [106]. Oldendorf patented his instrument in 1963 but was not able to attract companies interested in manufacturing a commercial version of it. Because of this contribution, Oldendorf was considered for the Nobel Price together with Cormack and Hounsfield, but eventually was excluded. Possible reasons for the decision of the Nobel committee and the politics involved can be found in [19].

Solving the CT Problem

The computerized tomography problem is to solve Eq. (2) when g_L has been determined for a large number of lines having different positions and directions. To make this problem more precise, the definition of projection will be introduced. First consider the following line integral of a function $f(x, y)$

$$p(r, \vartheta) = \int_L f(x, y)\mathrm{d}l \tag{3}$$

where the line L is defined in terms of its angle ϑ with the y axis and its distance r to the origin (Fig. 1). Alternatively, it can be said that L passes through point P with polar coordinates $(r\cos\vartheta, r\sin\vartheta)$ and is perpendicular to $\overline{OP}$. The collection of line integrals along parallel lines is known as a *parallel projection* of $f(x, y)$. Because this is the only type of projection considered here, the qualifier parallel will be dropped. In terms of Eq. (3), the projection of f is obtained by letting r be a variable while keeping the value of ϑ fixed (i. e., ϑ must be considered a parameter). Note that some authors call a particular line integral a projection along that line (see, e. g., [132]).

Equation (3) was solved using two different approaches that broadly speaking can be referred to as analytical and iterative. The analytical approach followed different paths, but in each case the result was a closed-form solution, which had to be solved numerically. This approach is relevant to seismology for at least two reasons. First, it provides the theoretical underpinnings of the concept of backprojection, which has entered the seismic to-

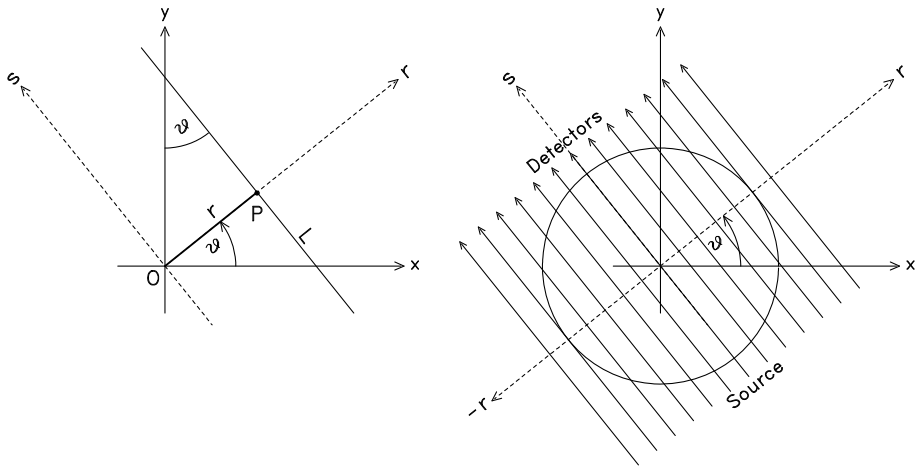

Tomography, Seismic, Figure 1
Left: **Geometry for the definition of projection. *L* is the projection line and *r* is the distance from the line to the origin. The axis *s* is parallel to *L*. *Right:* Geometry for parallel X-ray tomography. The circle bounds an object to be projected. Each line is equivalent to the line *L* on the left. There is a source of X-rays on one end of the lines and detectors on the opposite end**

mography literature. Second, it is directly related to the concepts of slant-stack and Radon transform popular in the reflection seismology literature. The iterative approach was to write the CT problem in matrix form and to solve it iteratively. This approach was widely applied to the solution of seismological inverse problems. Because of their close relation to seismology, the two approaches will be discussed here.

Analytical Solutions The most popular solutions are based on the use of the Fourier transform, but before considering them we will briefly mention the work of Radon, the Bohemian mathematician that investigated Eq. (3) in a paper published it 1917. It was written in German, and an English translation (by R. Lohner) can be found in [35].

Radon considered the function

$$p(r,\theta) = p(-r,\theta+\pi) = \int_{-\infty}^{\infty} f(r\cos\vartheta - s\sin\vartheta, r\sin\vartheta + s\cos\vartheta)\mathrm{d}s \tag{4}$$

(see Eq. (11) below; his symbols were somewhat different), and the mean value of $p(r,\theta)$ for the lines tangent to a circle with center at a point $Q = (x, y)$ and radius a

$$\bar{p}_Q(a) = \frac{1}{2\pi}\int_0^{2\pi} p(x\cos\vartheta + y\sin\vartheta + a, \theta)\,\mathrm{d}\theta \tag{5}$$

and proved that

$$f(Q) = -\frac{1}{\pi}\int_0^{\infty}\frac{\mathrm{d}\bar{p}_Q(a)}{a} = -\frac{1}{\pi}\int_0^{\infty}\frac{\mathrm{d}\bar{p}_Q}{\mathrm{d}a}\frac{\mathrm{d}a}{a}, \tag{6}$$

where the last equality is given in [28]. A proof of Radon's result can be found in [62]. Equations (4) (or equivalent expressions) and (6) are known as the Radon and inverse Radon transform, respectively. Although Eq. (6) looks simple, its practical implementation is not [132] and probably for this reason it did not receive as much attention (after Cormack's 1973 paper, [28]) as the other methods developed to solve the problem. As a matter of historical interest we note that Radon's problem arises in a number of scientific fields and had been solved independently more than once in the early part of the twentieth century [30]. Yet, these results were not widely known and had to be derived again.

The analytical approach developed along two different lines and produced results formally different from that of Radon. One was pioneered by Cormack [26,27], but, as noted earlier, it went essentially unnoticed. A second line originated in radio astronomy [17,18], electron microscopy (e. g., [34,36]) and X-ray radiography (e. g., [12,124]), and was based on the use of the Fourier transform. This is the approach that will be taken here and its description will be based mainly on [20,76,88] and [132].

The early development of the CT technique was based on the use of a number of parallel projections determined for different values of ϑ in Eq. (3). This is the case that will be analyzed here. To find the equation that represents L we will use that fact that its slope is equal to $-1/\tan\vartheta$ and its intercept with the y axis is $r/\sin\vartheta$. Thus

$$y = -\frac{\cos\vartheta}{\sin\vartheta}x + \frac{r}{\sin\vartheta} \tag{7}$$

so that

$$x \cos\vartheta + y \sin\vartheta = r\,. \tag{8}$$

To tie this definition to the concept of *slant stack* in seismology we note that it is defined using Eq. (3) with f replaced by a function $u(x, t)$, where u, x, and t represent seismic wave amplitude, distance, and time, respectively, y is replaced by t, and t is written in terms of the intercept and slope of the line L (generally indicated with τ and p, so that $t = px + \tau$) (see, e. g., [40,127]).

Now we will introduce a new coordinate system (r, s) obtained by a rotation of angle ϑ of the (x, y) system (Fig. 1). In the (r, s) system the points on the line L have constant r and variable s. The two systems are related by the following transformations of coordinates

$$\begin{pmatrix} r \\ s \end{pmatrix} = \begin{pmatrix} \cos\vartheta & \sin\vartheta \\ -\sin\vartheta & \cos\vartheta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x\cos\vartheta + y\sin\vartheta \\ -x\sin\vartheta + y\cos\vartheta \end{pmatrix} \tag{9}$$

and

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{pmatrix} \begin{pmatrix} r \\ s \end{pmatrix} = \begin{pmatrix} r\cos\vartheta - s\sin\vartheta \\ r\sin\vartheta + s\cos\vartheta \end{pmatrix}\,. \tag{10}$$

Using Eq. (10), Eq. (3) can be rewritten as

$$\begin{aligned} p(r,\theta) &= \int_{-\infty}^{\infty} f(r\cos\vartheta - s\sin\vartheta, r\sin\vartheta + s\cos\vartheta)\mathrm{d}s \\ &\equiv \int_{-\infty}^{\infty} \hat{f}_\vartheta(r,s)\mathrm{d}s \end{aligned} \tag{11}$$

where $\hat{f}_\vartheta$ represents the function f when it is written in terms of r and s, and the subscript ϑ is used to emphasize that it enters in the computations as a parameter. In Eq. (11) r is allowed to vary continuously, although in practical applications (such as CT) r is a discrete variable, as sketched in Fig. 1. Also note that r is allowed to be negative, which means that $0 \le \vartheta < \pi$, and that although $|r|$ is allowed to extend to infinity, for functions defined over a finite domain in the (x, y) plane the projections for lines outside of the domain will be zero.

To proceed further we will work in the wavenumber domain. Let the one-dimensional Fourier transform (or 1-D F.T.) of $p(r, \theta)$ with respect to r be

$$P(k,\theta) = \int_{-\infty}^{\infty} p(r,\theta)\mathrm{e}^{-\mathrm{i}2\pi kr}\mathrm{d}r \tag{12}$$

where k indicates wavenumber (equivalent to the frequency in the time domain). Introducing Eq. (11) into this expression and then going back to the (x, y) coordinate system gives

$$\begin{aligned} P(k,\theta) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \hat{f}_\vartheta(r,s)\mathrm{e}^{-\mathrm{i}2\pi kr}\mathrm{d}r\mathrm{d}s \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\mathrm{e}^{-\mathrm{i}2\pi k(x\cos\vartheta + y\sin\vartheta)}\mathrm{d}x\mathrm{d}y \\ &= F_\vartheta(k_x, k_y) \end{aligned} \tag{13}$$

where r has been written using Eq. (9), the Jacobian of the coordinates transformation is equal to one, F_ϑ is the two-dimensional Fourier transform of f and

$$k_x = k\cos\vartheta\,; \quad k_y = k\sin\vartheta\,. \tag{14}$$

Note that

$$k^2 = k_x^2 + k_y^2\,; \quad k_y/k_x = \tan\theta\,. \tag{15}$$

In summary,

$$P(k,\theta) = F_\vartheta(k_x, k_y)\,. \tag{16}$$

In words, the 1-D F.T. of the projection $p(r, \theta)$ of $f(x, y)$ is equal to the 2-D F.T. of $f(x, y)$. This relation is known as the Fourier (central) slice theorem. This name comes from the fact that the 2-D F.T. is known in a slice (i. e., a line) through the origin in the (k_x, k_y) space, as Eq. (15) shows. Moreover, the angle of this line with the k_x axis is ϑ, which is equal to the angle between the x and r axes. Equation (16) suggests one approach to the determination of $f(x, y)$. Find the 1-D F.T. of the projections for discrete values of θ between 0 and π and then use the 2-D inverse F.T. to recover $f(x, y)$ numerically. However, because the values of the 2-D F.T will be defined on a polar coordinates grid, it must be interpolated to a Cartesian grid in the (k_x, k_y) space.

A different approach is as follows. Let $F(k_x, k_y)$ be the 2-D F.T. of $f(x, y)$. Then, $f(x, y)$ is given by the inverse F.T.

$$f(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(k_x,k_y)\mathrm{e}^{\mathrm{i}2\pi(k_x x + k_y y)}\mathrm{d}k_x\mathrm{d}k_y. \tag{17}$$

In this expression k_x and k_y are defined on a Cartesian coordinate system, which will be converted to polar coordinates using Eq. (14). This gives

$$f(x,y) = \int_0^{\pi}\int_{-\infty}^{\infty} P(k,\vartheta)\mathrm{e}^{\mathrm{i}2\pi k(x\cos\vartheta + y\sin\vartheta)}|k|\,\mathrm{d}\vartheta\,\mathrm{d}k \tag{18}$$

where Eq. (16) was used and $|k|\,\mathrm{d}\vartheta\,\mathrm{d}k$ is the Jacobian for the transformation of coordinates. Because the Jacobian relates the elements of area in the two coordinate systems, the absolute value of k is needed because the area is positive. Two comments are in order here. First, ϑ is no longer a parameter; now it is used as a variable. Second, the polar coordinates have been defined using $0 \le \theta < \pi$ and $-\infty < k < \infty$. This convention is equivalent to the standard one (i. e., $0 \le \theta < 2\pi$, $0 \le k < \infty$.)

Now Eq. (18) will be rewritten as follows

$$f(x, y) = \int_0^{\pi} p^*(x\cos\vartheta + y\sin\vartheta, \vartheta)\,\mathrm{d}\theta \tag{19}$$

where

$$p^*(r, \vartheta) = \int_{-\infty}^{\infty} P(k, \vartheta)|k|\mathrm{e}^{\mathrm{i}2\pi kr}\mathrm{d}k\,. \tag{20}$$

The integral in Eq. (20) involves the product of two 1-D F.T.s, $P(k, \theta)$ and $|k|$, so, in principle, one way to solve it would be to use the following property. Given two functions $g(r)$ and $h(r)$ with F.T.s $G(k)$ and $H(k)$, then

$$\int_{-\infty}^{\infty} G(k)H(k)\mathrm{e}^{\mathrm{i}2\pi kr}\mathrm{d}k = \int_{-\infty}^{\infty} g(\rho)h(r-\rho)\mathrm{d}\rho \tag{21}$$

where the integral on the right-hand side is the convolution of g and h (see, e. g., [94]). A problem with this formulation is that $|k|$ does not have an inverse F.T., as can be seen from the fact that it does not go to zero as k goes to infinity, which is one of the properties of the F.T. [149]. There are two ways to get around this problem. One is to solve Eq. (21) in terms of distributions. A basic yet rigorous introduction to distributions, which can be considered an extension of the concept of function, can be found in [120]. Let us rewrite Eq. (20) as follows

$$p^*(r, \theta) = \frac{1}{2\pi^2}\int_{-\infty}^{\infty} P(k, \theta)(\mathrm{i}2\pi k)(-\mathrm{i}\pi\,\mathrm{sgn}\,k)\mathrm{e}^{\mathrm{i}2\pi kr}\mathrm{d}k \tag{22}$$

where

$$\mathrm{sgn}\,k = \begin{cases} 1 & k > 0 \\ -1 & k < 0\,. \end{cases} \tag{23}$$

Then, using Eq. (21) we can write

$$p^*(r, \theta) = \frac{1}{2\pi^2}\mathcal{F}^{-1}\left\{\mathrm{i}2\pi kP(k, \theta)\right\} * \mathcal{F}^{-1}\left\{-\mathrm{i}\pi\,\mathrm{sgn}\,k\right\} \tag{24}$$

where $\mathcal{F}^{-1}$ represents the inverse F.T. of the function in braces and the $*$ stands for convolution. The first inverse is just $\partial p(r, \theta)/\partial r$ and the second inverse is equal to $1/r$ [120]. Combining these two results gives

$$\begin{aligned} p^*(r, \theta) &= -\frac{1}{2\pi}\left(-\frac{1}{\pi r} * \frac{\partial p(r, \theta)}{\partial r}\right) \\ &\equiv -\frac{1}{2\pi}\mathcal{H}\left\{\frac{\partial p(r, \theta)}{\partial r}\right\} \end{aligned} \tag{25}$$

where $\mathcal{H}$ stands for the Hilbert transform of the function in braces [120]. The inverse expression for the slant stack can be derived using expressions similar to those presented above [25].

The importance of Eq. (25) is mostly theoretical, and from a practical point of view a different method of solution was found to be more useful. Before proceeding, however, note that Eq. (20) without $|k|$ on the right-hand side corresponds to the 1-D inverse F.T. of $P(k, \theta)$, equal to $p(r, \theta)$ (see Eq. (12)). Therefore, p^* is a *filtered projection*, with $|k|$ the filter response. Clearly, the effect of the filter is to amplify the components of p corresponding to the higher wavenumbers. The approach used here is also based on the use of Eq. (21), with $g(r) = p(r, \theta)$ (as before) and $h(r)$ a function whose F.T. is an approximation to $|k|$ in the sense that

$$H(k) \approx \begin{cases} |k|\,; & |k| < K \\ 0\,; & \text{elsewhere} \end{cases} \tag{26}$$

where K can be taken as the spatial Nyquist frequency k_N (equal to $1/2a$, where a is the spacing between projection lines). To avoid aliasing, $P(k, \theta)$ must be zero for $|k| > k_N$. Clearly, there is some freedom in the selection of $h(r)$. For example, in [18] and [124] $H(k)$ is defined by an expression similar to Eq. (26) with an equal sign in place of the less than sign. This work was followed by the introduction of an improved function [133]. For our purposes, however, what is important is not the particular choice of h but the fact that we can write the following approximation for $f(x, y)$

$$f(x, y) \approx \tilde{f}(x, y) \equiv \int_0^{\pi}\int_{-\infty}^{\infty} p(\rho, \theta)\, h(x\cos\vartheta + y\sin\vartheta - \rho)\mathrm{d}\rho\mathrm{d}\theta\,. \tag{27}$$

Finally, because the projections are determined for discrete and equispaced values of r and θ we can write

$$\tilde{f}(x, y) \approx \frac{\pi a}{N}\sum_{i=1}^{N}\sum_{j=-J}^{J} p(ja\,, \theta_i)\, h(x\cos\vartheta_i + y\sin\vartheta_i - ja) \tag{28}$$

where $\theta_i = (i-1)\pi/N$, N is fixed, and J is a finite limit that takes into account the finite size of the object (after [132]).

To end this discussion of the transform methods of solution we will go back to Eq. (19). The integration there is an example of the operation known as *backprojection*, to be indicated with $\mathcal{B}$ and defined by

$$\mathcal{B}\{q\}(x,y) = \int_0^{\pi} q(x\cos\vartheta + y\sin\vartheta, \vartheta)\,\mathrm{d}\theta \qquad (29)$$

where q is an arbitrary function of r and ϑ. An example of this operation is provided in Fig. 2a, which shows two projections of a rectangular object corresponding to ϑ equal to 0 and $\pi/2$. The two projections, identified by $p(x,0)$ and $p(y,\pi/2)$, are nonzero for limited ranges of the y and x axes. To compute the backprojection let us consider a particular value y' of y and assign the value $p(y',\pi/2)$ to every point in the line defined by the points (x,y'). The object may be part of a larger body, and for this reason the line will extend beyond the x values that bound the object. Repeating this operation for all the values y' and then proceeding in a similar way for the $p(x,0)$ projection produces an image such as that shown in Fig. 2b. The two backprojections overlap in the area occupied by the rectangular object, but also contribute values to other areas of the (x,y) plane, which is clearly not correct. The problem is not related to the use of only two projections; even when a larger number of projections is used, the contributions outside of the object do not disappear. The reason for that is that the determination of $f(x,y)$ requires p^*, not p, as shown by Eq. (19) (see, e. g., [23]).

Iterative Solutions As was the case with the analytical solutions, the iterative solutions were developed before the advent of CT. The earliest one was the *algebraic reconstruction technique* (ART, [51]) and was developed for the solution of electron microscopy problems as an alternative to analytical solutions based on the Fourier transform. These problems are represented by an equation similar to Eq. (2). The basic idea behind ART was the discretization of a plane object in terms of a square grid of points. The goal was to find the optical density ρ_{ij} at each point (i,j) of the grid. A ray of a projection at an angle θ was defined as a band of width w across the plane at the same angle, and corresponds to each of our lines L. One possible choice for w is to make it equal to the grid spacing. The basis of ART is to start with an average value of the density (computed from the observations) and to use it to compute the projections for all the rays for all the angles. Then, two updating methods were used, one additive and the other multiplicative. In the original version of the additive method, for each ray the difference between observed and computed projections was used to update the density values in the cells along the ray. If an updated value of density became negative it was set to zero. This was done for all the rays, one ray at a time. This completed one iteration. Then a new one was started with the updated densities used as starting values. Originally, the difference between the observed and computed projections was divided equally among all the cells in the ray, but in later applications each cell received a weight representing its contribution to a given ray. The weight was used when the cell densities were updated (see, e. g., [50]). In the multiplicative version of ART the density value of a cell was updated by multiplying it by the ratio of computed to observed projections. Again, a weighting scheme was introduced later. A technique similar to additive ART was developed independently by Hounsfield and applied to his X-ray scanner (see, e. g., [20,133]).

A second technique, simultaneous iterative reconstruction technique (SIRT) was introduced in [48], which was highly critical of the performance of ART. SIRT also

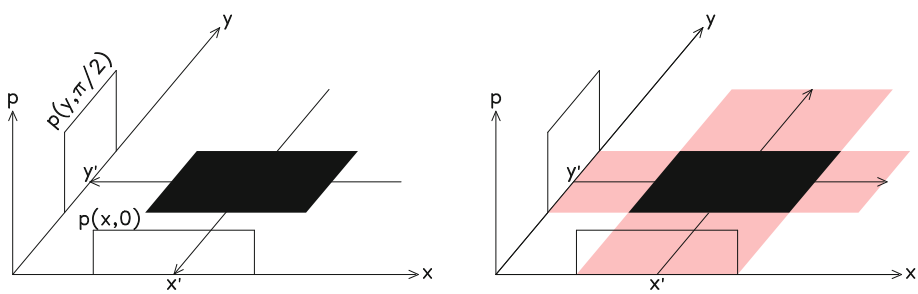

Tomography, Seismic, Figure 2
Illustration of the concept of back projection. *Left:* projection of a rectangular box using two sets of *parallel lines* (two representative *lines* are shown). *Right:* backprojection of the two projections shown on the *left*. Note the change in the direction of the *arrows*. For each *line*, the value of its projection is assigned to all the *points* along the *line* and are added to the values that come from all the other projections. See text for details. After [20]

has additive and multiplicative versions, and the main difference with the ART counterparts is that at each iteration the density of each cell is updated using all the projections passing through that cell. Subsequent work showed that although ART is computationally more efficient than SIRT, it has the problem that is more affected by errors in the data (see, e. g., [20,39,63]). SIRT, on the other hand, has some undesirable properties, referred to in Subsect. "Regularization Approach".

Arrival-Time Seismic Tomography

As noted earlier, this type of tomography is much simpler than surface wave and waveform tomography, and lends itself to a fairly self-contained discussion, to be presented here. The other two types of tomography will not be discussed here because they require a knowledge of seismic theory and data processing beyond the scope of this article.

This section is organized as follows. First, the differences between seismic and X-ray tomography will be discussed. The early applications of CT techniques to seismic problems ignored these differences, and although the results thus obtained opened the way to a new research approach in seismology, they may have been affected by several types of errors, the sources of which will become clear here. Second, tomography using local data will be addressed. In this case both the location of the events as well as the velocity model must be determined. In addition, a distinction between velocity and slowness tomography must be made. As noted earlier, before seismic tomography there was velocity inversion. Then, after the introduction of CT techniques the inversion parameter became slowness, not velocity. Therefore, the equations to be solved in the two cases are different, and because some of the velocity inversion programs are still in use, the two parameterizations, and their relationship, will be presented here. In addition, regardless of the parameterization used, it is necessary to decouple (or separate) the location part from the tomographic part of the problem, which requires introducing new analytical tools. Another topic related to local tomography is the computation of travel times, which will be briefly considered here. Finally, the case of tomography using teleseismic arrival times is analyzed.

Comparison with X-ray Tomography

Seismic tomography differs from X-ray tomography in two fundamental ways because, first, the former is nonlinear, and second, the locations of the seismic sources are generally unknown (i. e., when earthquakes are used). Another major difference is that in X-ray tomography there is control over position and number of sources and receivers, which is not the case for seismic tomography. The first two differences will be considered in more detail.

The expression for travel time, t, along a ray is given by

$$t = \int_R \frac{1}{v} \mathrm{d}s = \int_R u \mathrm{d}s \tag{30}$$

where R denotes the raypath between two fixed points within an elastic medium with velocity $v(\mathbf{x})$ that depends on position (indicated by the vector $\mathbf{x}$), $\mathrm{d}s$ is a line element along the raypath, and u is the slowness, equal to the inverse of velocity

$$u(\mathbf{x}) = \frac{1}{v(\mathbf{x})} \,. \tag{31}$$

An advantage of writing t in terms of u instead of v is that in terms of u, t has an expression similar to Eq. (2), which constitutes the basis of X-ray tomography. This similarity, however, is more apparent than real. To see that we will introduce the concept of linear operation. Given an operation $\mathcal{O}$, it is said to be linear if

$$\mathcal{O}(f + h) = \mathcal{O}(f) + \mathcal{O}(h) \tag{32}$$

and

$$\mathcal{O}(\alpha f) = \alpha \mathcal{O}(f) \tag{33}$$

where f and h are functions and α is a scalar. A simple example of linear operation is the integration over a given interval

$$\int_a^b \left[f(x) + h(x)\right] \mathrm{d}x = \int_a^b f(x)\mathrm{d}x + \int_a^b h(x)\mathrm{d}x \,. \tag{34}$$

In particular, the operation defined by Eq. (2) is linear because L is always a straight line that does not depend on f. Therefore, if we let $f = f_1 + f_2$ we can write

$$\begin{aligned} g_L(f) &= \int_L [f_1(s) + f_2(s)]\mathrm{d}s \\ &= \int_L f_1(s)\mathrm{d}s + \int_L f_2(s)\mathrm{d}s \\ &= g_L(f_1) + g_L(f_2) \,. \end{aligned} \tag{35}$$

Now let us consider the travel time problem. Let $R(u_1)$ and $R(u_2)$ denote the raypaths between two fixed points within an elastic medium assuming that it had slownesses u_1 and $u_2 \neq u_1$ at two different times. This situation is generally not possible in the earth but can be simulated on a computer. The corresponding travel times are given by

$$t_1 = \int_{R(u_1)} u_1(s)\mathrm{d}s \,, \quad t_2 = \int_{R(u_2)} u_2(s)\mathrm{d}s \,. \tag{36}$$

Note that the two raypaths will be different because u_1 and u_2 have been assumed to be different. The trivial case of $u_1 = ku_2$, where k is a constant, is ignored. Therefore, it will not be possible, in general, to write $t_1 + t_2$ as an integral over a common path involving $u_1 + u_2$. Moreover, if the medium has slowness $u_1 + u_2$ the raypath $R(u_1 + u_2)$ between the two points will be different from $R(u_1)$ and $R(u_2)$ and, in general,

$$t_1 + t_2 \neq \int_{R(u_1+u_2)} \left[u_1(s) + u_2(s)\right] \mathrm{d}s \tag{37}$$

which in turn means that the travel time problem cannot be expressed in terms of a linear operation. Therefore, the seismic tomography problem is nonlinear, and the CT solution techniques cannot be applied directly except when the 3-D velocity variations in the medium are small, a condition that severely limits its application to situations of seismic interest.

Another major difference between seismic tomography and CT is that the locations of the sources (mostly earthquakes) are generally not known. Therefore, the determination of a velocity model requires the simultaneous determination of the source locations. This is particularly important when the events are locally recorded. To see that we will consider a very simple example. Let us assume that a vertical fault separates rocks with low- and high-velocities, denoted v_1 and v_2, respectively, and that earthquakes occur along a vertical line within the fault. Let $v_1 = 3$ km/s and $v_2 = 4$ km/s. The origin of the coordinate system will be the fault location at the surface and the event depths will be 1, 2, 3, 4, and 5 km. The earthquakes will be assumed to be recorded by a local network of eight stations on the surface. The theoretical arrival times computed for this geometry were used with a single-event location program and two constant velocity models with velocities equal to 3 and 3.5 km/s and with a 1-D model with five layers and velocities between 2.5 and 4 km/s. The locations obtained with these models (Fig. 3) are affected by significant errors, which cannot be inferred by the small root-mean-square errors, which range between 0.02 and 0.09 s, with the larger value corresponding to the deeper event. These results show that the effect of the lateral juxtaposition of high- and low-velocity rocks when the events are located with 1-D velocity models is a mislocation of the events in a direction away from the low-velocity zone. A well-documented example of this situation is provided by the aftershocks of the 1994 $M = 6.7$ Northridge, California, earthquake, which occurred within the sedimentary rocks of the San Fernando basin. These events and the velocity model determined using them are considered in

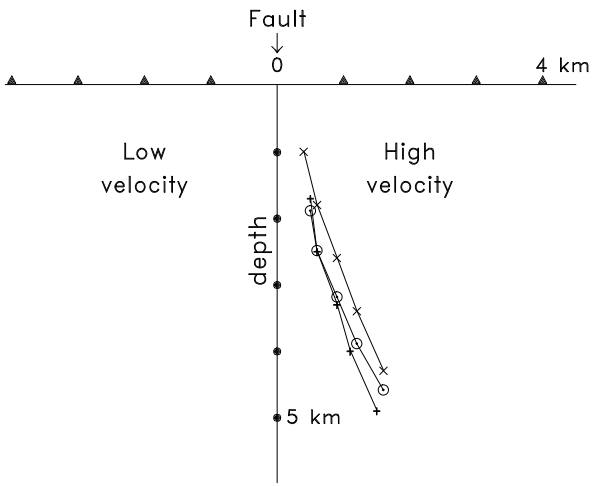

Tomography, Seismic, Figure 3
Simple example (2-D) of the effect of lateral velocity variations on earthquake location when those variations are ignored. A fault divides the medium into two quarter spaces with velocities equal to 3 km/s *(left)* and 4 km/s *(right)*. The *triangles* and *dots* represent stations and hypocenters, respectively. Synthetic arrival times computed for these events, stations, and velocity model are used with a single-event location program and three 1-D velocity models. Two models have constant velocities, equal to 3 km/s and 3.5 km/s, and the other has five layers with velocities between 2.5 and 4 km/s. The locations determined using these models are indicated by pluses, crosses and circles

Subsect. "*P* Wave Velocity Model for the San Fernando, California, Area".

This example has a counterpart. What happens when mislocated events are kept at their erroneous locations and are used to determine a 2-D velocity model around the fault? Clearly, such a velocity model will be affected by the location errors and its reliability will be questionable. Examples showing the significant bias that these errors may have in the computed velocities are given in [54,145], and [146] for geometries similar to that in Fig. 3. In [146], $v_1 = 5$ km/s and $v_2 = 6$ km/s and there was only one earthquake on the fault. In this case the velocity determined by inversion was 5.44 km/s to the left of the fault and 5.56 km/s to the right of it. This example is clearly an oversimplification, but the result is useful because it shows that ignoring the mislocation introduced by lateral velocity variations will result in a velocity model showing variations smaller than the actual ones. For the case of teleseismic tomography, event mislocation may not be a serious problem except when earthquakes from subducting slabs are used, in which case the effect on the computed velocities may be important (see, e. g., [32,153]).

Another difference between seismic tomography and X-ray tomography is that in the later the position of the

sources and the receivers can be chosen so that the computed images have the desired resolution. In contrast, in seismic tomography there is no control over the position of the earthquakes, while the number and position of stations is dictated by financial considerations as well as practical constraints imposed by the nature of the terrain. For example, mountains frequently constitute a severe obstacle to station deployments. Even more serious is the presence of oceans, which cover two-thirds of the earth's surface and essentially prevent the routine deployment of permanent seismic networks similar to those on land. As a consequence, seismic tomographic models, particularly at a global scale, are affected to some extent by low-resolution problems.

Finally, we note that soon after X-ray tomography began, several competing methods of solution were proposed (see Subsect. "Solving the CT Problem"), and to compare their performances synthetic data generated for a model of a cross section of the skull were used [133]. With this approach it was possible to detect artifacts in some of the solutions and to investigate the effect of noise in the data. In contrast, seismic tomography methods were rarely subject to a similar validation analysis in spite of the approximations and simplifying assumptions made. These differences stem, in part, on the complicated nature of wave propagation in the earth, which is difficult to replicate in a computer on account of theoretical difficulties and computational cost. On the other hand, the propagation of X-rays is relatively simple and easy to simulate on a computer. In addition, medical tomography, in general, is well funded, while seismic tomography is only poorly funded [84].

Local Velocity Tomography

The simultaneous determination of the locations of a group of earthquakes and a 3-D velocity model can be considered an extension of the standard method of earthquake location. Usually, the velocity structure is modeled in terms of constant-velocity blocks or by velocity values on a grid. The basic ideas were introduced by Aki and Lee [2] and Crosson [33], although the latter only considered layered velocity models. Because this inverse problem is nonlinear in both the earthquake location and the velocity determination, it is solved by linearizing it about initial estimates of the locations and origin times of the events and model velocities. With this approach the original problem is replaced by other in which the unknowns are adjustments to the initial estimates. Then these adjustments are added to the initial estimates, the corresponding values are used as new initial estimates, and the process is repeated iteratively until some stopping criterion is met. This process will be described mathematically in the following.

Let us use subscripts i and j to identify the stations and earthquakes used, respectively, and let M and N_j be the number of events and the number of stations with arrival times for the jth event, respectively. The blocks or grid points used to parameterize the velocity structure constitute a 3-D array, but for computational purposes each block or point will be identified with a single subscript, say k. This requires establishing an ordering scheme that assigns a single index to a triplet of indices. For example, we may use the following scheme: $(1,1,1) \to 1$, $(2,1,1) \to 2$, $(n,1,1) \to n$, $(n,2,1) \to n+1$, and so on. Let K represent the total number of velocity parameters. The linearized problem can be written as

$$w_{ij} r_{ij} = w_{ij}\left(\mathrm{d}T_j + \frac{\partial t}{\partial x}\mathrm{d}x_j + \frac{\partial t}{\partial y}\mathrm{d}y_j + \frac{\partial t}{\partial z}\mathrm{d}z_j + \sum_k \frac{\partial t}{\partial v_k}\mathrm{d}v_k\right) \qquad i = 1,\ldots,N_j \quad j = 1,\ldots,M \tag{38}$$

where

$$r_{ij} = \mathcal{T}_{ij}^{\mathrm{obs}} - \mathcal{T}_{ij}^{\mathrm{comp}} \equiv \mathcal{T}_{ij}^{\mathrm{obs}} - (T_j + t_{ij}) \tag{39}$$

is the arrival time residual for the jth earthquake and the ith station. The subscript k labels the blocks or grid points associated with the raypath from the jth hypocenter to the ith station. Therefore, k is a function of i and j, but this dependence is left implicit to simplify the notation. Also note that the values of k associated with a given ray do not have any particular ordering. The meaning of the other variables is as follows:

$\mathrm{d}T_j, (\mathrm{d}x_j, \mathrm{d}y_j, \mathrm{d}z_j), \mathrm{d}v_k$: adjustments to origin time, hypocentral coordinates, and velocities, respectively,
$\mathcal{T}_{ij}^{\mathrm{obs}}$: observed arrival time,
$\mathcal{T}_{ij}^{\mathrm{comp}}$: computed arrival time,
T_j: origin time,
t_{ij}: computed travel time,
t, v: travel time and velocity,
w_{ij}: quality weight.

In Eqs. (38) and (39) the only unknowns are the adjustments. All the other quantities are computed using initial estimates of the velocities (v_k) and the origin time (T_j) and hypocentral coordinates (x_j, y_j, z_j) of each earthquake. The expressions for the derivatives can be found in [85].

Local Slowness Tomography

As noted earlier, the travel time problem cannot be expressed in terms of a linear operation on slowness, and for this reason the problem is linearized assuming that the difference between the initial and actual velocity models is small. Let us go back to Eq. (36) with $u_2 = u_1 + \mathrm{d}u$, $\mathrm{d}u \ll u_1$. Then the difference $\mathrm{d}t$ in travel times is given by

$$\mathrm{d}t = t_2 - t_1 = \int_{R(u_1+\mathrm{d}u)} [u_1(s) + \mathrm{d}u(s)]\mathrm{d}s - \int_{R(u_1)} u_1(s)\mathrm{d}s \approx \int_{R(u_1)} \mathrm{d}u(s)\mathrm{d}s\,. \quad (40)$$

Here we have invoked *Fermat's principle*, which allows replacing $R(u_1 + \mathrm{d}u)$ with $R(u_1)$. Recall that this principle states that raypaths are paths of stationary travel time (see, e. g., [3,120]), which in turn means that the variation of travel time along a raypath is zero. Two points must be noted here, however. First, the principle applies to small variations in raypaths, which in turn requires a small $\mathrm{d}u$, a condition which may not be valid in reality. This question was tested in the context of the propagation of rays through subducting slabs [31] and it was found that arrival times at teleseismic distances calculated using Fermat's principle were the same or later than those determined using exact ray tracing. The difference between exact and approximate times depended on event depth, with the larger errors corresponding to intermediate-depth events. A consequence of these errors was an underestimation of the velocity anomaly. Second, Fermat's principle applies to curves with the same end points, but in practice the end point of $R(u_1)$ corresponding to the event location is not the true location because it has been determined with an incorrect velocity model. Therefore, $R(u_1)$ is not necessarily close to the true ray path. In practice, however, the approximations involved in the use of Fermat's principle in local tomography improve as the iterations proceed, and can essentially be ignored.

For completeness, the relation between small variations $\mathrm{d}v$ and $\mathrm{d}u$ in velocity and slowness will be considered. Because travel time is equal to distance divided by velocity, here we are interested in $1/(v + \mathrm{d}v)$, with $\mathrm{d}v \ll v$, which can be approximated as follows

$$\frac{1}{v+\mathrm{d}v} = \frac{1}{v}\left(1+\frac{\mathrm{d}v}{v}\right)^{-1} \approx \frac{1}{v} - \frac{1}{v^2}\mathrm{d}v \equiv u + \mathrm{d}u\,. \quad (41)$$

Therefore, it does not make any theoretical difference whether the inverse problem is formulated in terms of slowness or velocity as long as $\mathrm{d}u$ and $\mathrm{d}v$ are such that the approximations introduced are valid.

Now let us go back to the joint hypocentral location and determination of a 3-D slowness model. The only difference with the expression for velocity tomography (Eq. (38)) is in the last term, which is replaced by the right-hand side of Eq. (40). This gives

$$w_{ij} r_{ij} = w_{ij}\left(\mathrm{d}T_j + \frac{\partial t}{\partial x}\mathrm{d}x_j + \frac{\partial t}{\partial y}\mathrm{d}y_j + \frac{\partial t}{\partial z}\mathrm{d}z_j + \int_{R_{ij}} \mathrm{d}u\mathrm{d}s\right) \quad (42)$$
$$i = 1, \ldots N_j \quad j = 1, \ldots M$$

where R_{ij} indicates the raypath between the corresponding station-event pair. To proceed further the integral must be discretized, which can be done using a block model. After that Eq. (42) becomes

$$w_{ij} r_{ij} = w_{ij}\left(\mathrm{d}T_j + \frac{\partial t}{\partial x}\mathrm{d}x_j + \frac{\partial t}{\partial y}\mathrm{d}y_j + \frac{\partial t}{\partial z}\mathrm{d}z_j + \sum_k l_{ijk}\mathrm{d}u_k\right) \quad (43)$$
$$i = 1, \ldots N_j \quad j = 1, \ldots M$$

where l_{ijk} and $\mathrm{d}u_k$ are the length of the ray in the kth block and the corresponding slowness perturbation.

Note that although the last terms in Eqs. (38) and (43) are different, in principle, the velocity and slowness models derived using the two formulations should satisfy $u(\mathbf{x}) = 1/v(\mathbf{x})$. Let us consider this question for the block parameterization. To compute the derivative of travel time with respect to velocity we will use the following approximation

$$\frac{\partial t_{ij}}{\partial v_k} \approx \frac{\partial t_{ijk}}{\partial v_k} \quad (44)$$

[85], where the subscripts in t are used to identify the rays and blocks involved. The meaning of this relation is that the change in travel time along a ray path due to a change in the velocity of a given block is approximately equal to the change in travel time in that block. Note that

$$t_{ijk} = \frac{l_{ijk}}{v_k} \quad (45)$$

which allows us to write

$$\frac{\partial t_{ij}}{\partial v_k}\mathrm{d}v_k \approx l_{ijk}\left(-\frac{1}{v_k^2}\mathrm{d}v_k\right) \equiv l_{ijk}\mathrm{d}u_k\,. \quad (46)$$

Using this result we see that Eqs. (38) and (43) are identical, which means that the two formulations should give the

same results as long as the block structure is the same in the two cases and the approximations introduced are valid. Of course, the numerical implementation of the software corresponding to the two approaches should be the same.

Decoupling of the Earthquake Location and Tomography Problems

Equation (38) can be written in matrix form as:

$$\mathbf{W}_j\mathbf{A}_j\mathrm{d}\mathbf{x}_j + \mathbf{W}_j\mathbf{B}_j\mathrm{d}\mathbf{v} = \mathbf{W}_j\mathbf{r}_j\,; \quad j = 1, M \tag{47}$$

where $\mathbf{r}_j$ is the vector of residuals r_{ij}, $\mathbf{A}_j$ is an $N_j \times 4$ matrix of partial derivatives of time with respect to origin time and hypocentral coordinates, $\mathrm{d}\mathbf{x}_j$ is the vector of origin time and hypocenter adjustments, $\mathbf{W}_j$ is an $N_j \times N_j$ matrix of weights, $\mathbf{B}_j$ is an $N_j \times K$ matrix of partial derivatives of travel time with respect to velocities, and $\mathrm{d}\mathbf{v}$ is a vector of K velocity adjustments. The matrix $\mathbf{B}_j$ has zero entries for the blocks not traversed by any ray. The matrix $\mathbf{W}_j$ has only one nonzero entry per row and column and is not necessarily diagonal because the ordering of stations for different earthquakes may not be always the same.

For Eq. (43) the matrix form is

$$\mathbf{W}_j\mathbf{A}_j\mathrm{d}\mathbf{x}_j + \mathbf{W}_j\mathbf{C}_j\mathrm{d}\mathbf{u} = \mathbf{W}_j\mathbf{r}_j\,; \quad j = 1, M \tag{48}$$

with $\mathbf{C}_j$ an $N_j \times K$ matrix whose entries are raypath lengths in individual blocks and $\mathrm{d}\mathbf{u}$ is a vector of K slowness adjustments. All the other quantities are as in Eq. (47).

Clearly, Eqs. (47) and (48) are formally equivalent, which means that for the following discussion on how to solve them it is unnecessary to distinguish between velocity and slowness. For this reason, from now on we will refer exclusively to Eq. (47), with the understanding that the results below apply to Eq. (48) as well.

Equation (47) represents M systems of equations coupled through the common vector $\mathrm{d}\mathbf{v}$. Therefore, these systems could be combined, in principle, into a single system, but because its size could become computationally unmanageable, it is necessary to decouple the earthquake location part from the inversion part. A very efficient approach to do that is based on the singular value decomposition (SVD) of a matrix [110] and is known as the method of parameter separation. The following presentation is based on [119]. The SVD of an arbitrary $n \times m$ matrix $\mathbf{G}$ is given by

$$\mathbf{G} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}} \tag{49}$$

(see, e. g., [3,46]) where the superscript T indicates matrix transposition, $\mathbf{U}$ and $\mathbf{V}$ are $n \times n$ and $m \times m$ matrices with columns given by the eigenvectors of $\mathbf{G}\mathbf{G}^{\mathrm{T}}$ and $\mathbf{G}^{\mathrm{T}}\mathbf{G}$, respectively, and $\boldsymbol{\Lambda}$ is an $n \times m$ a matrix with diagonal elements equal to the singular values of $\mathbf{G}$ and off-diagonal elements equal to zero. The singular values are positive and equal to the square roots of the eigenvalues of $\mathbf{G}\mathbf{G}^{\mathrm{T}}$ and $\mathbf{G}^{\mathrm{T}}\mathbf{G}$. The number of nonzero singular values, say p, cannot exceed the minimum of m and n. Matrices $\mathbf{U}$ and $\mathbf{V}$ are orthogonal, i. e.,

$$\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}_n\,, \quad \mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{I}_m\,, \tag{50}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix. Let us assume that the singular values are sorted in nonincreasing order (i. e., largest first). Then, matrices $\mathbf{U}$ and $\mathbf{V}$ can be partitioned as follows:

$$\mathbf{U} = (\mathbf{U}_p \quad \mathbf{U}_0)\,, \quad \mathbf{V} = (\mathbf{V}_p \quad \mathbf{V}_0) \tag{51}$$

where the subscripts p and 0 indicate that the columns of the matrices come from the eigenvectors corresponding to the nonzero and zero singular values, respectively. Partitioned matrices are discussed in, e. g., [103]. Matrix $\boldsymbol{\Lambda}$ can be partitioned in a similar way. Then, writing Eq. (49) in terms of these partitioned matrices gives

$$\mathbf{G} = \begin{pmatrix}\mathbf{U}_p & \mathbf{U}_0\end{pmatrix}\begin{pmatrix}\boldsymbol{\Lambda}_p & \mathbf{O}\\ \mathbf{O} & \mathbf{O}\end{pmatrix}\begin{pmatrix}\mathbf{V}_p^{\mathrm{T}}\\ \mathbf{V}_0^{\mathrm{T}}\end{pmatrix} = \mathbf{U}_p\boldsymbol{\Lambda}_p\mathbf{V}_p^{\mathrm{T}} \tag{52}$$

where $\boldsymbol{\Lambda}_p$ is the $p \times p$ diagonal matrix having as elements the p nonzero singular values and $\mathbf{O}$ represents zero matrices of appropriate sizes. Using this result and

$$\mathbf{U}_0^{\mathrm{T}}\mathbf{U}_p = \mathbf{O} \tag{53}$$

which is a consequence of the fact that the columns of $\mathbf{U}$ are orthogonal to each other, we get the following important result

$$\mathbf{U}_0^{\mathrm{T}}\mathbf{G} = \mathbf{U}_0^{\mathrm{T}}\mathbf{U}_p\boldsymbol{\Lambda}_p\mathbf{V}_p^{\mathrm{T}} = \mathbf{O}\,. \tag{54}$$

To apply Eq. (54) to our problem we will use

$$\begin{aligned}\mathbf{G} &= \mathbf{W}_j\mathbf{A}_j = \mathbf{U}_j\boldsymbol{\Lambda}_j\mathbf{V}_j^{\mathrm{T}}\\ &= \begin{pmatrix}\mathbf{U}_{jp} & \mathbf{U}_{j0}\end{pmatrix}\boldsymbol{\Lambda}_j\begin{pmatrix}\mathbf{V}_{jp} & \mathbf{V}_{j0}\end{pmatrix}^{\mathrm{T}} = \mathbf{U}_{jp}\boldsymbol{\Lambda}_{jp}\mathbf{V}_{jp}^{\mathrm{T}}\,.\end{aligned} \tag{55}$$

Because $\mathbf{W}_j\mathbf{A}_j$ has four columns and at least as many rows, $p \le 4$. If $p < 4$, this means that one or more of the columns of the matrix is a linear combination of the others and the location of the corresponding event cannot be determined uniquely. Events with $p < 4$ should not be used in the inversion. Now multiply Eq. (47) by $\mathbf{U}_{j0}^{\mathrm{T}}$ on the left and use the equivalent of Eq. (54) for the jth event. This gives

$$\mathbf{U}_{j0}^{\mathrm{T}}\mathbf{W}_j\mathbf{B}_j\mathrm{d}\mathbf{v} = \mathbf{U}_{j0}^{\mathrm{T}}\mathbf{W}_j\mathbf{r}_j\,; \quad j = 1, M\,. \tag{56}$$

Because $\mathbf{A}_j$ does not appear in Eq. (56), we have been able to decouple the velocity parameters from the earthquake parameters. Note, however, that this equation depends on $\mathbf{A}_j$ implicitly via $\mathbf{U}_{j0}$, which means that location errors will translate into errors in $\mathbf{U}_{j0}$ and, thus, in $\mathbf{dv}$.

To simplify the following discussion Eq. (56) will be rewritten as

$$\underbrace{\mathbf{B}'_j}_{(N_j-4)\times K}\ \underbrace{\mathbf{dv}}_{K\times 1} = \underbrace{\mathbf{r}'_j}_{(N_j-4)\times 1}\ ; \quad j = 1, M \tag{57}$$

with obvious expressions for the primed quantities. The sizes of the matrix and vectors involved are also indicated (assuming $p = 4$). Equation (57) represents M systems of equations with the common unknown vector $\mathbf{dv}$ and can be written in compact form in terms of a partitioned matrix and vector

$$\mathcal{B}\,\mathbf{dv} = \boldsymbol{\rho} \tag{58}$$

where

$$\mathcal{B} = \left(\mathbf{B}_1'^T\ \mathbf{B}_2'^T\ \ldots\ \mathbf{B}_M'^T\right)^{\mathrm{T}} \tag{59}$$

and

$$\boldsymbol{\rho} = \left(\mathbf{r}_1'^T\ \mathbf{r}_2'^T\ \ldots\ \mathbf{r}_M'^T\right)^{\mathrm{T}} . \tag{60}$$

Let N be the largest of the N_j. Then, matrix $\mathcal{B}$ has K columns and up to $M \times (N-4)$ rows, which means that it may become very large. For example, there may be thousands of earthquakes, thousands of velocity parameters, and tens of stations, which means that how to solve Eq. (60) becomes an issue. If the size of the matrix and the computer resources allow it, a solution based on the use of the SVD would be convenient (see Sect. "Solution of Ill-Posed Linear Problems"). If this approach is not feasible, one can solve Eq. (60) by least squares, which requires solving

$$\mathcal{B}^{\mathrm{T}}\mathcal{B}\,\mathbf{dv} = \mathcal{B}^{\mathrm{T}}\boldsymbol{\rho} \tag{61}$$

which on account of Eqs. (59) and (60) becomes

$$\left(\sum_{j=1}^{M}\mathbf{B}_j'^T\,\mathbf{B}_j'\right)\mathbf{dv} = \sum_{j=1}^{M}\mathbf{B}_j'^T\mathbf{r}_j\,. \tag{62}$$

This approach however, may have two problems. One is that the resulting matrix may still be too large for the computer facilities available and the other is the possibility of numerical loss of precision due to the matrix multiplications. This second problem can be alleviated through the use of double precision, although at the expense of longer computer times. For these reasons, for large tomographic problems the ensuing linear systems are solved using iterative matrix solvers (see Subsect. "Regularization Approach). Also note that Eq. (62) is not the result of the "accumulation" of individual equations $\mathbf{B}_j'^{\mathrm{T}}\mathbf{B}_j'\mathbf{dv} = \mathbf{B}_j'^{\mathrm{T}}\mathbf{r}_j$, as it is sometimes stated.

Once Eq. (58) has been solved, the earthquakes must be relocated. To do that Eq. (47) will be used, but to make the analysis more general it will be assumed that both P and S wave arrivals are available. Because the P wave arrivals are independent of the S wave velocity, and vice versa, we can write two pairs of equations similar to Eqs. (47) and (56), and to solve for $\mathbf{dv}^P$ and $\mathbf{dv}^S$, where the superscripts identify the type of arrivals. After that is done we can write

$$\mathbf{W}_j^P\mathbf{A}_j^P\,\mathbf{dx}_j = \mathbf{W}_j^P\mathbf{r}_j^P - \mathbf{W}_j^P\mathbf{B}_j^P\,\mathbf{dv}^P\,; \quad j = 1, M \tag{63}$$

and

$$\mathbf{W}_j^S\mathbf{A}_j^S\,\mathbf{dx}_j = \mathbf{W}_j^S\mathbf{r}_j^S - \mathbf{W}_j^S\mathbf{B}_j^S\mathbf{dv}^S\,; \quad j = 1, M\,. \tag{64}$$

These two equations are coupled through the common vector $\mathbf{dx}_j$ and can be written as a single equation as follows

$$\begin{pmatrix}\mathbf{W}_j^P\mathbf{A}_j^P\\ \mathbf{W}_j^S\mathbf{A}_j^S\end{pmatrix}\mathbf{dx}_j = \begin{pmatrix}\mathbf{W}_j^P\mathbf{r}_j^P - \mathbf{W}_j^P\mathbf{B}_j^P\mathbf{dv}^P\\ \mathbf{W}_j^S\mathbf{r}_j^S - \mathbf{W}_j^S\mathbf{B}_j^S\mathbf{dv}^S\end{pmatrix}\,; \quad j = 1, M\,. \tag{65}$$

[122]. The matrix on the left-hand side of this equation is small ($2N \times 4$ at most) and can be solved using any of the standard methods. Once each $\mathbf{dx}_j$ has been found, it is used to get new initial estimates of origin time and hypocentral estimates (equal to $T_j + \mathrm{d}T_j\,, x_j + \mathrm{d}x_j$ and so on). After updating the velocities (i. e., they become $v_k + \mathrm{d}v_k$) a new iteration is started. This iterative process is repeated until some stopping criterion is met. An early criterion [144] was based on the use the statistical F test. With this test it is possible to determine if the decrease in the sum of arrival time residuals squared from a given iteration to the next is statistically significant ([80] and references therein). A simpler, yet effective approach is to stop when the root-mean square of all the travel-time residuals reaches the expected value of the error in the data.

Computation of Local Travel Times

The iterative determination of a 3-D velocity model and simultaneous determination of earthquake locations requires the availability of software for the computation of

travel times in that type of models. The theoretical solution to this problem using ray theory was well known (see, e. g., [22,44,85,112], and references therein) at the time the seismic tomographic method began to be developed, but because exact ray tracing is a computationally time-consuming task, in most of the earlier tomographic codes approximate ray tracing methods were used. One of them was introduced by Thurber [144], and was based on the use of arcs of circles to approximate the ray paths connecting the source and station. Although the method was fast, its accuracy was questionable. For example, in [42] the weights of arrivals with epicentral distances between 20 and 45 km were decreased and arrivals with distances larger than 45 km were not used. Because this approximate method was popular, the readers of the earlier literature should be aware of the limitations of the method, which can result in errors in the computed velocities and in the locations determined with them.

A significant improvement to Thurber's [144] method was introduced by Um and Thurber [152], who developed a method based on the perturbation of an initial raypath between the source and the station using a ray-theoretical equation. Ray tracing methods that find the raypath between two points by iterative perturbation of an initial estimate are known as *bending methods* (see, e. g., [75] and references therein). The Um and Thurber method is based on an approximate computation of the perturbations, and for this reason it is commonly referred to as a *pseudo-bending method*. However, one of the disadvantages of this method, noted by the authors, is that it is not appropriate for use in media with velocity discontinuities and where nearly constant velocities are present. This method is included in a popular tomographic package, and these limitations should be taken into account when considering the results obtained using it. Pseudo-bending solutions that do not have the limitations of the Um and Thurber method exist [99,115], but their applications to seismic tomography appear to be limited. The Um and Thurber's method was used in combination with Snell's law to account for velocity discontinuities at the Conrad and Moho boundaries and the upper boundary of subducting plates [163], but these boundaries must be known from other studies.

An early example of tomography with exact ray tracing can be found in [61]. In this case the ray tracing problem was solved by Runge–Kutta integration of six simultaneous differential equations that can be derived from the eikonal equation. This approach corresponds to the *shooting method*, which requires the specification of the two angles at the source that define a ray. These two angles must be changed until the end point is within a specified distance from the station. This approach does not introduce any approximations (beyond those inherent to ray theory), but, as already noted, is more time consuming than the approximate methods. The software based on this approach was later improved [14] by incorporation of the method of parameter separation described in Subsect. "Decoupling of the Earthquake Location and Tomography Problems". This software was applied to the 1989, $M = 7.1$, Loma Prieta, California, mainshock-aftershock sequence [116]. Another example of tomography with exact ray tracing [79] is based on the division of the earth into constant-velocity blocks and ray tracing based on the use of Snell's law.

A different kind of approach to the computation of travel times, which does not require ray tracing, was introduced by Vidale [157,158]. In his method, the velocity model consists of a set of values assigned to points in a regular equispaced 2-D or 3-D grid and travel times are computed using a finite-difference approximation to the partial derivatives of travel times with respect to the spatial coordinates and plane or circular wavefront approximations. It must be noted, however, that the eikonal equation is not solved using the classical finite-difference method, which is not a simple task and is seldomly used [21]. The output of Vidale's method is a series of wavefronts of minimum travel times. This method is more time consuming than the approximate methods referred to above and has some problems when large velocity contrasts are present. Moreover, the approximations involved may not be adequate in the vicinity of the source, where the wavefronts are highly curved [21]. Vidale's method has been implemented in software for earthquake location in media with 3-D variations [102].

Vidale's [157,158] approach motivated a method developed by Podvin and Lecomte [113], who used Huygens's principle for the computation of travel times. The application of the principle is equivalent to the propagation of local wavefronts, which is done as follows. Let us assume that the four vertices of one of the sides of a unit cube in the grid have known arrival times. These points can be combined into four groups of three adjacent points. Each group defines a plane wavefront, which is used to compute the arrival time at each of the other four corners of the same cube. The selection of the appropriate local wavefront follows a set of pre-established rules. In addition, each point within the grid is the common vertex of eight unit cubes around it. These cubes form a larger cube with the grid point at its center and each side contributes sixteen local wavefronts. Therefore, for each grid point within the grid, ninety-six wavefronts must be considered. The advantage of this approach is that it accounts for the existence of transmitted, diffracted, and head waves and performs well even in the presence of large velocity

contrasts. The computations can be carried out in parallel in a multiprocessor machine or sequentially when only one processor is used. In the latter case the method becomes computationally very time consuming, but its implementation in a tomographic software package [15] has resulted in highly detailed velocity models. Examples are given in Sect. "Examples". Podvin and Lecomte's software is becoming very popular and for this reason it is worth noting that the software has a minor flaw that results in time differences for rays moving equal distances to the left and right of the source [151]. For example, for a 0.5 km grid size and source-receiver offsets up to 30 km the difference is between about 0.03 and 0.04 s, with the larger values for offsets less than 10 km. Decreasing the grid size reduces the error.

The two wavefront methods referred to above do not compute raypaths directly. However, in seismic tomography they are needed to associate arrival time residuals with specific velocity grid points. In Podvin and Lecomte's software, rays are traced from the receiver to the source following a direction opposite to the time gradient, with the endpoint of the ray at most a distance h from the source, where h is the grid spacing. Recall that in isotropic media rays are defined as curves whose tangents are everywhere perpendicular to a wavefront (see, e. g., [69,120]).

There is yet another approach to the computation of travel times, namely, the shortest-path method, which is based on concepts borrowed from network theory. In this approach the velocity model is also based on a grid of points and connections are established between close points. Each connection is given a weight, equal to the travel time between them. The shortest path between any two points is that along the connections for which the sum of the weights is smallest. This path is an approximation to the seismic raypath. This method was introduced in seismology in [101] and was further developed in [98]. Additional references can be found in, e. g., [9,21], and [147]. Applications to seismic tomography and 3-D earthquake location can be found in [7,8] and [100], respectively.

To end this section we note that the methods considered here assume a Cartesian coordinate system, so that the curvature of the earth is ignored, i. e., the earth is assumed to be flat. This assumption, however, has limitations (see, e. g., [125,134]), which should be taken into account when dealing with epicentral distances that exceed about 150–200 km. This range is provided as guide only. For a given epicentral distance, the error in travel time generally increases with hypocentral depth. For example, for a depth of 200 km and an epicentral distance of 2° the error is about 0.25 s [125]. When the errors introduced by ignoring the earth's curvature become important (e. g., larger than the arrival-time picking errors) it is advisable to introduce a cut-off epicentral distance beyond which the arrival times are either not used or weighted down (see, e. g. [78] and Subsect. "*P* and *S* Wave Velocity Models for Taiwan").

Teleseismic Tomography

The data most commonly used in teleseismic tomographic studies are *P* wave arrivals for the investigation of the mantle and *PcP* and *PKP* arrivals for the investigation of the core-mantle boundary (see, e. g. [96]), although other arrivals have also been used. A major source for the data is the International Seismological Center, which has collected millions of arrivals contributed by hundreds of stations around the world. Although it is possible to perform a simultaneous event location and velocity determination following a formulation similar to that described for local tomography (see, e. g. [137]), the most common approach is to keep the event locations fixed. This is the approach to be describe here.

The use of teleseismic data to determine a 3-D velocity model of the crust and upper mantle underneath a seismic array was introduced by Aki et al. [4]. They divided the portion of the earth being investigated into homogeneous horizontal layers subdivided into blocks, computed the theoretical arrival times for earthquakes having published locations, formed arrival time residuals, and solved for slowness perturbations in each of the blocks with respect to a model with constant slowness in each layer. A simplified version of the arguments presented in [4] follows. Let u_k and u_k^o be the actual and reference slowness of block k and let m_k be the fractional slowness perturbation, equal to

$$m_k = \frac{u_k - u_k^o}{u_k^o} \equiv \frac{\mathrm{d}u_k}{u_k^o} \tag{66}$$

where $\mathrm{d}u_k$ is equivalent to the $\mathrm{d}u$ in Eq. (40) and is assumed to be much smaller than u_k. Solving for u_k we get

$$u_k = u_k^o(1 + m_k)\,. \tag{67}$$

Given a particular earthquake-station pair, the actual travel time in the kth block traveled by the corresponding ray is given by

$$t_k = l_k u_k = l_k u_k^o(1 + m_k) \approx t_k^o + t_k^o m_k \tag{68}$$

where l_k is the actual length of the ray in the kth block and the approximate sign arises because the actual path and the path in the reference medium are assumed to be close on account of Fermat's principle. Then, the difference in

travel time introduced by the slowness perturbation m_k is given by

$$\mathrm{d}t_k = t_k - t_k^o \approx t_k^o m_k \,. \tag{69}$$

Introducing the right-hand side of Eq. (66) in this expression gives

$$t_k \approx \frac{t_k^o}{u_k^o}\mathrm{d}u_k = l_k^o \mathrm{d}u_k \tag{70}$$

where l_k^o is the raypath length in the kth block for the reference model. The right-hand side of this expression is equivalent to $l_{ijk}\mathrm{d}u_k$ in Eq. (43).

Let us consider the arrival time residual that arises when using teleseismic data and the earthquake locations used are those determined with the reference slowness model. Using the notation introduced earlier we can write

$$r_{ij} = \delta t_j + \sum_k l_{ijk}^o \mathrm{d}u_k \tag{71}$$

where the subscripts i and j denote station and event, and the sum is over all the blocks. A block not traversed by any ray will have the corresponding l_{ijk}^o equal to zero. The term δt_j includes the contributions to the residuals not accounted for by the second term. For example, δt_j includes errors in event origin time and location, picking errors, possible errors in the reference slowness model outside of the volume investigated, and possible errors in the model parameterization. This δt_j is unknown, but because it is assumed to be effectively constant for a given earthquake, it can be eliminated by averaging r_{ij} over all the stations that recorded the jth event and subtracting the result from Eq. (71). The average residual is given by

$$\overline{r_{ij}} = \frac{1}{N_j}\sum_{i=1}^{N_j} r_{ij} = \delta t_j + \sum_k \overline{l_{ijk}^o}\mathrm{d}u_k \tag{72}$$

where the overbar indicates average, as defined by the first equality. Next, subtracting Eq. (72) from Eq. (71) gives the relative residual

$$r'_{ij} = r_{ij} - \overline{r_{ij}} = \sum_k \left(l_{ijk}^o - \overline{l_{ijk}^o}\right)\mathrm{d}u_k \,. \tag{73}$$

A similar result can be found in [3]. This equation shows that the slowness perturbations $\mathrm{d}u_k$ are obtained by solving a linear system, as expected, and that the main difference with Eq. (43) is the presence of source terms in the latter (aside from a weight factor). Once Eq. (73) is solved, the perturbations are added to the reference slownesses and the tomographic problem is solved. It is important to note however, that the resulting model is not well resolved vertically. In fact, the effect of a uniform perturbation over a layer cannot be distinguished from a change in δt_j. This problem is intrinsic to the method and cannot be removed [3]. Examples showing the capabilities and limitations of the method can be found in [45]. A potentially severe limitation is the effect introduced by the presence of large sedimentary basins, which usually have significantly smaller velocities than the surrounding rocks. These basins may introduce travel time anomalies as large as the upper mantle residuals [93]. Clearly, the tomographic results will be incorrect if the arrival times are not corrected for large crustal effects, which requires information derived independently (see, e. g., [93,131]).

Solution of Ill-Posed Linear Problems

Regardless of the type of arrivals considered, the tomographic problem reduces to the solution of equations of the form

$$\mathbf{Ax} = \mathbf{b} \tag{74}$$

where $\mathbf{A}$ is a known $m \times n$ matrix with information about the model, $\mathbf{x}$ is an n vector of unknowns, and $\mathbf{b}$ is a known m vector. Solving this problem requires consideration of the question of the existence of a solution, but before proceeding it is convenient to introduce the following definition, due to the French mathematician Hadamard (see, e. g., [47,53], and references therein). The problem represented by Eq. (74) is said to be *well posed* if it has a solution, it is unique, and depends continuously on $\mathbf{b}$. The latest condition means that small changes in $\mathbf{b}$ cause small changes in $\mathbf{x}$. When any of these conditions is not satisfied the problem is said to be *ill posed*. For a unique solution of Eq. (74) to exist a necessary (but not sufficient) condition is that $m \geq n$. If $m > n$, the system is said to be *overdetermined*. It must be noted, however that this condition (to be assumed here) does not imply that a solution exists, or that it is unique (if it exists). For example, if some of the rows of $\mathbf{A}$ constitute a linearly dependent set (e. g., some of them are linear combinations of other rows) a solution may exist but it will not be unique. On the other hand, a solution may not exist at all. In this case the system is said to be *inconsistent*, a condition that in seismic tomography arises because of errors in both the data and the model. The magnitude of the data errors depend on arrival type. Teleseismic data generally have larger errors than local data, while local S wave arrivals have larger errors than the corresponding P arrivals because they are more difficult to identify. Model errors arise because the parameterization of the velocity (or slowness) field does not reproduce

the actual variations appropriately. Although a more faithful representation of the earth would be advantageous, an obvious problem is that the data may not be adequate to resolve the parameters of a more detailed model. Clearly, if the block size or grid spacing used to parameterize the velocity (or slowness) field are too small with respect to the interstation spacing, and the spatial distribution of the seismic sources is not favorable, then a large number of blocks may not be traversed by any ray, and matrix $\mathbf{A}$ will be very large with a large number of zero entries (i. e., the matrix is said to be *sparse*). For a given set of arrival times, an increase in the number of unknown parameters results in a relative decrease in the number of constraints, and in the resolution of the solution (see below).

A very powerful tool for the solution of the system (74) is provided by the Moore–Penrose generalized inverse $\mathbf{A}^\dagger$ of $\mathbf{A}$. Let $\mathbf{A}$ have p nonzero singular values, sorted by decreasing order. Then, using the SVD (see Eq. (49))

$$\mathbf{A}^\dagger = \mathbf{V}\boldsymbol{\Lambda}^\dagger\mathbf{U}^\mathrm{T} = \mathbf{V}_p\boldsymbol{\Lambda}_p^{-1}\mathbf{U}_p^\mathrm{T} \tag{75}$$

where $(\boldsymbol{\Lambda}^\dagger)_{ii}$ is equal to $1/\lambda_i$ if $\lambda_i \neq 0$ and zero otherwise [3,111]. The second equality follows from the product of partitioned matrices. The generalized inverse solution is given by

$$\mathbf{x}^\dagger = \mathbf{A}^\dagger\mathbf{b} \tag{76}$$

which has the property that is a minimum-length solution (see, e. g., [3]). A more general solution is of the form

$$\mathbf{x} = \mathbf{x}^\dagger + \mathbf{z} \tag{77}$$

where

$$\mathbf{A}\mathbf{z} = \mathbf{0}\,. \tag{78}$$

The vectors $\mathbf{z}$ that satisfy Eq. (78) constitute the *null space* of $\mathbf{A}$. The column vectors of $\mathbf{A}$ corresponding to the zero singular value satisfy an equation similar to Eq. (78) and constitute a base for the null space of $\mathbf{A}$. Therefore, $\mathbf{z}$ is a linear combination of the columns of the matrix $\mathbf{V}_o$.

The definition of $\boldsymbol{\Lambda}^\dagger$ is based on a sharp distinction between zero and nonzero singular values. In practice, however, this clear-cut situation does not occur, with some of them nonzero yet much smaller than the largest one. In this case the solution of Eq. (76) may be strongly affected by errors in the data. This question is made more precise when the condition number, κ, of $\mathbf{A}$ is introduced

$$\kappa = \frac{\lambda_{\text{largest}}}{\lambda_{\text{smallest}}} \geq 1\,. \tag{79}$$

Then, a perturbation $\mathrm{d}\mathbf{b}$ (such as errors) in the data introduces a perturbation $\mathrm{d}\mathbf{x}$ in the solution that satisfies

$$\frac{|\,\mathrm{d}\mathbf{x}|}{|\mathbf{x} + \mathrm{d}\mathbf{x}|} \leq \kappa\frac{|\mathrm{d}\mathbf{b}|}{|\mathbf{b} + \mathrm{d}\mathbf{b}|} \tag{80}$$

(e. g., [46,49]). Therefore, when κ is large a small change in the data may cause a significant change in the solution. If κ is large the matrix is said to *be ill-conditioned*; otherwise it is said to be *well-conditioned*. Although the terms "large" and "small" are obviously vague, the idea is not. For example, for a given data set and two velocity models leading to matrices $\mathbf{A}$ having condition numbers that differ by a factor of say 10, the matrix with the larger κ may lead to a solution more affected by the errors in the data.

In the following, ill-conditioned problems will be considered a form of ill-posed problems and the distinction between the two will no longer be made. The question that must be addressed now is how to solve this type of problems. Two approaches are available, one based on the so-called regularization of the problem and the other based on the Bayesian statistics. The two approaches are discussed in detail below.

Regularization Approach

As noted above, tomographic problems are likely to be ill-posed, and to get a solution it is necessary to recourse to mathematical techniques that turn them well-posed. This process is known as regularization (see, e. g., [53,55,80,148,164]). The reader must be aware, however, that there is a price to be paid; namely one ends up with a family of solutions and to objectively select the most appropriate among them may not be easy or even possible. For a system with a matrix for which the computation of its singular value decomposition is practically feasible, the generalized inverse solution has several advantages, but for most tomographic problems the size of the matrices involved is so large that this option is not practical and will not be pursued here. Instead, we will discuss techniques that lead to systems that can be solved with computationally efficient methods.

Before proceeding we will introduce the following definitions. A square symmetric matrix $\mathbf{C}$ is said to be *positive semidefinite* if

$$\mathbf{y}^\mathrm{T}\mathbf{C}\mathbf{y} \geq 0\,; \quad \mathbf{y} \neq \mathbf{0}\,. \tag{81}$$

The matrix $\mathbf{C}$ is *positive definite* if the $\geq$ sign in Eq. (81) is replaced by $>$.

Let us summarize some of the properties of these matrices (see, e. g., [121]).

(1) Let $\mathbf{v}_i$ be an eigenvector of $\mathbf{C}$ and λ_i its corresponding eigenvalue. Then

$$\mathbf{v}_i^{\mathrm{T}}\mathbf{C}\mathbf{v}_i = \lambda_i \mathbf{v}_i^{\mathrm{T}}\mathbf{v}_i = \lambda_i |\mathbf{v}_i|^2 \,. \tag{82}$$

Because $|\mathbf{v}_i| > 0$, $\lambda_i \geq 0$ if $\mathbf{C}$ is positive semidefinite, and $\lambda_i > 0$ if $\mathbf{C}$ is positive definite. In the second case the inverse of $\mathbf{C}$ exists because $\mathbf{C} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}$, with $\mathbf{U}$ and $\boldsymbol{\Lambda}$ similar to those in Eq. (49) and Eq. (50) (see, e. g., [103]), and $\mathbf{C}^{-1} = \mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^{\mathrm{T}}$.

(2) Any matrix of the form $\mathbf{B}^{\mathrm{T}}\mathbf{B}$ is either positive definite or semidefinite. Consider

$$\mathbf{y}^{\mathrm{T}}(\mathbf{B}^{\mathrm{T}}\mathbf{B})\mathbf{y} = (\mathbf{B}\mathbf{y})^{\mathrm{T}}\mathbf{B}\mathbf{y} = |\mathbf{B}\mathbf{y}|^2 \geq 0\,; \quad \mathbf{y} \neq \mathbf{0}\,. \tag{83}$$

Therefore, $\mathbf{B}^{\mathrm{T}}\mathbf{B}$ is at least positive semidefinite. In addition, if the inverse of this matrix exits, all of its eigenvalues will be positive and $\mathbf{B}^{\mathrm{T}}\mathbf{B}$ will be positive definite. If the inverse does not exist, the matrix will be positive semidefinite.

(3) Let $\mathbf{C}$ be a diagonal matrix with positive diagonal elements. Then $\mathbf{C}$ is positive definite because

$$\mathbf{y}^{\mathrm{T}}\mathbf{C}\mathbf{y} = \sum_i c_i y_i^2 > 0\,; \quad \mathbf{y} \neq \mathbf{0}\,, \quad c_i = (\mathbf{C})_{ii} > 0\,. \tag{84}$$

(4) Let matrices $\mathbf{B}$ and $\mathbf{P}$ be positive semidefinite and definite, respectively, and let λ^2 be a scalar. Then $\mathbf{B} + \lambda^2\mathbf{P}$ is positive definite because

$$\mathbf{y}^{\mathrm{T}}(\mathbf{B}+\lambda^2\mathbf{P})\mathbf{y} = \mathbf{y}^{\mathrm{T}}\mathbf{B}\mathbf{y}+\lambda^2\mathbf{y}^{\mathrm{T}}\mathbf{P}\mathbf{y} > 0\,; \quad \mathbf{y} \neq \mathbf{0}\,. \tag{85}$$

These results will be used to show that the typical methods of solution of Eq. (74) when $\mathbf{A}^{-1}$ does not exist or is ill-conditioned are based on the solution of a new well-posed problem.

The simplest regularization approach is to constrain (in some sense) the length of the solution vector $\mathbf{x}$. Let us consider the minimization of the function

$$\bar{S}(\mathbf{x}) = w|\mathbf{A}\mathbf{x} - \mathbf{b}|^2 + \mathbf{x}^{\mathrm{T}}\mathbf{D}\mathbf{x} \equiv wS(\mathbf{x}) + Q(\mathbf{x}) \tag{86}$$

with respect to $\mathbf{x}$. Here w is a positive weighting factor, $\mathbf{D}$ is a symmetric positive definite matrix, and S and Q are defined by the identity. This problem was solved by Levenberg [87] and a similar one by Marquardt (who used $\mathbf{D} = \mathbf{I}$) [92] in the context of non-linear problems. The minimization of $\bar{S}$ leads to the following equation

$$\left(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \lambda^2\mathbf{D}\right)\mathbf{x} = \mathbf{A}^{\mathrm{T}}\mathbf{b}\,; \quad \lambda^2 = \frac{1}{w} \tag{87}$$

(see, e. g., [121]). A similar equation was introduced by Tihonov [148], who investigated the regularization of ill-posed linear problems in the context of integral equations. The $\mathbf{x}$ that solves Eq. (87) is known as the *damped least-squares estimator*. When $\lambda^2 = 0$, Eq. (87) corresponds to the *ordinary* least-squares solution.

Note that Eq. (87) is a special case of Eq. (85) with $\mathbf{B} = \mathbf{A}^{\mathrm{T}}\mathbf{A}$ and $\mathbf{P} = \mathbf{D}$, which means that there is a value of λ^2 that makes the matrix on the left side of Eq. (87) well conditioned. Therefore, for a given $\mathbf{D}$, Eq. (87) will have a family of solutions, which will depend on λ^2. A problem with this equation, however, is that it may be too large for tomographic problems. For this reason, Eq. (87) will be derived by consideration of the following system

$$\begin{pmatrix} \mathbf{A} \\ \lambda\mathbf{D}^{1/2} \end{pmatrix} \mathbf{x} = \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix} . \tag{88}$$

This new system will be solved using ordinary least squares, which requires solving

$$\begin{pmatrix} \mathbf{A}^{\mathrm{T}} & \lambda\mathbf{D}^{1/2} \end{pmatrix} \begin{pmatrix} \mathbf{A} \\ \lambda\mathbf{D}^{1/2} \end{pmatrix} \mathbf{x} = \begin{pmatrix} \mathbf{A}^{\mathrm{T}} & \lambda\mathbf{D}^{1/2} \end{pmatrix} \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix} \tag{89}$$

where $\mathbf{D}^{1/2}$ is known as the square root of $\mathbf{D}$ (see, e. g., [121]). After performing the matrix multiplications we obtain Eq. (87), which shows that the minimization of $\bar{S}(\mathbf{x})$ is equivalent to solving Eq. (74) under the constraint

$$\lambda\mathbf{D}^{1/2}\mathbf{x} = \mathbf{0}\,. \tag{90}$$

Clearly, the λ in this equation is unnecessary, but it is introduced to indicate that it is a weighting factor. We also note that in many applications $\mathbf{D} = \mathbf{I}$.

Let us mention some of the properties of the damped least-squares solution in the context of nonlinear linearized inverse problems. This type of problems require an initial estimate of the solution, say $\mathbf{x}_0$. If this estimate is poorly chosen, it may lead to a non-convergent iterative process, and for this reason Levenberg [87] decided to minimize $\bar{S}$. Moreover, when dealing with linearized problems, S is the linearized version of the true residuals. Let s represent the sum of residuals whose linearization leads to S. A question investigated by Levenberg [87] regards the relations between $\bar{S}$, S and s, and between the damped and ordinary least-squares solutions, indicated by $\mathbf{x}_\lambda$ and $\mathbf{x}^{\mathrm{ls}}$, respectively. Some of his results are as follows:

$$S(\mathbf{x}_\lambda) < S(\mathbf{x}_0) \tag{91}$$

$$Q(\mathbf{x}_\lambda) < Q(\mathbf{x}^{\mathrm{ls}}) \tag{92}$$

and

$$s(\mathbf{x}_\lambda) < s(\mathbf{x}_0) \quad \text{for some} \quad \lambda^2 \tag{93}$$

as long as $\mathbf{x}_0$ is not a stationary point of s [87,121]. Equations (91) and (92) show that it is possible to simultaneously minimize S and Q for all λ^2, while Eq. (93) shows that there are values of λ^2 that minimize s, which is the quantity of actual interest to us. Although none of these results is obvious, they are implicitly assumed when solving linearized problems. In addition, if λ^2 is large enough, Eq. (87) becomes

$$\mathbf{x} \approx \frac{1}{\lambda^2}\mathbf{D}^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{b} \tag{94}$$

with $\mathbf{x}$ going to zero as λ^2 goes to infinity. Therefore, when solving linearized nonlinear problems λ^2 should be relatively large in the early iterations, so that $\mathbf{x}$ is small enough to assure convergence to a solution. Then, as the iterations proceed the value of λ^2 should be decreased gradually until it reaches its regularization value.

Another regularization approach is to limit the "roughness" of the solution by constraining its Laplacian to be equal to zero ([86] and references therein). This constraint can be introduced by writing the Laplacian using a finite-difference approximation

$$\begin{aligned} 6x_{i,j,k} - \big(&x_{i-1,j,k} + x_{i+1,j,k} + x_{i,j-1,k} \\ &+ x_{i,j+1,k} + x_{i,j,k-1}, + x_{i,j,k+1}\big) = 0 \end{aligned} \tag{95}$$

where $x_{i,j,k}$ is the value of $\mathbf{x}$ at a grid point (identified by the three indices) and $x_{i\pm1,j,k}$, $x_{i,j\pm1,k}$, and $x_{i,j,k\pm1}$ are the values of $\mathbf{x}$ at adjacent points. This 3-D constraint is used in [15]. Equation (95) must be translated into a matrix form, which will depend on the number of grid points in the three dimensions. Let $\mathbf{L}$ indicate this matrix. Then, the constrained system becomes

$$\begin{pmatrix} \mathbf{A} \\ \lambda\mathbf{L} \end{pmatrix} \mathbf{x} = \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix} . \tag{96}$$

To solve this new system by ordinary least squares we must write an expression similar to Eq. (89), which in turn gives

$$\left(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \lambda^2\mathbf{L}^{\mathrm{T}}\mathbf{L}\right)\mathbf{x} = \mathbf{A}^{\mathrm{T}}\mathbf{b} . \tag{97}$$

In general, $\mathbf{L}$ is positive definite (see, e. g., [5,114]) and so is $\mathbf{L}^{\mathrm{T}}\mathbf{L}$, which means that this approach also leads to a well-conditioned system.

In practice, Eqs. (88) and (96) can be solved using methods that compute the solution iteratively, such as the ART, SIRT, and LSQR methods. As noted earlier, ART is affected by noise in the data, while SIRT has the disadvantage that it introduces an unwanted scaling of the problem. This question will not be discussed here, but it should be noted that the term SIRT does not refer to a single technique; rather it refers to a family of iterative methods based on ideas similar to those described in [48]. The earlier application to geophysical problems is described in [39], but the algorithm used there was equivalent to the solution of a problem that was a scaled version of the original one, i. e., an equation similar to Eq. (58) is multiplied by a certain diagonal matrix which is not controlled by the user [73]. A similar problem affects the original SIRT formulation [63]. A general discussion of this question can be found in [154]. The LSQR method [108] was introduced in seismology by Nolet ([104] and references therein), does not have the problems that affect ART and SIRT, and is very well suited for the tomographic method because the matrices involved are highly sparse, which helps speed the computations considerably. For these reasons, the LSQR method is used widely used in seismic tomography (see, e. g., [15,86,163]).

So far we have not discussed how the parameter λ^2 should be chosen. This question does not have a definite answer. In their discussion of the damped least-squares method, Lawson and Hanson [83] plotted the norm (i. e. length) of the solution vector versus the norm of the vector of residuals obtained for different values of the damping parameter and noted (for specific examples) that the resulting curve had an approximate L shape. The optimal value of the damping parameter was that corresponding to the corner of the curve at which it goes from nearly vertical to nearly horizontal. The idea here is to minimize both the residual and the length of the solution. This approach is discussed further in [55] and [56], where the logarithms of the two norms, rather than the norms themselves, are used. A somewhat related approach is based on the plot of the data variance versus the solution variance [42]. A recent example of choice of damping parameter motivated by this latter approach is provided in [78].

We end this section with the definition of *resolution*. Regardless of how Eq. (74) is solved, we can write

$$\hat{\mathbf{x}} = \hat{\mathbf{A}}\mathbf{b} \tag{98}$$

where $\hat{\mathbf{A}}$ can be considered the inverse, in some sense, of $\mathbf{A}$. Now write $\mathbf{b}$ using Eq. (74). This gives

$$\hat{\mathbf{x}} = \hat{\mathbf{A}}\mathbf{A}\mathbf{x} \equiv \mathbf{R}\mathbf{x} \tag{99}$$

where $\mathbf{R}$ is known as the resolution matrix (see, e. g., [3]) and is defined by the identity. If $\mathbf{R} = \mathbf{I}$, $\hat{\mathbf{x}} = \mathbf{x}$. If $\mathbf{A}$ were known perfectly well, this result would imply that the solution vector $\hat{\mathbf{x}}$ is exactly equal to the true solution. In practice, in seismic tomography $\mathbf{A}$ is only poorly known, and

even if $\mathbf{R} = \mathbf{I}$ there is no reason to say that $\hat{\mathbf{x}}$ represents the true solution, as sometimes stated. To see that, consider an extreme case. Let us suppose that a velocity model consists of just one block. In this case $\mathbf{A}$ will have only one column and will be well conditioned, which means that the corresponding Eq. (74) can be solved by least squares. The resulting resolution matrix will be equal to $\mathbf{I}$, but it is obvious that the computed solution will not correspond to the true velocity of the earth (except in the unlikely case that the velocity is a constant). If the number of blocks is gradually increased to make the model more realistic, at some point $\mathbf{A}$ will become ill-posed and a regularized solution must be introduced. In this case from Eq. (87) we get

$$\hat{\mathbf{A}} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \lambda^2\mathbf{D}\right)^{-1}\mathbf{A}^{\mathrm{T}} \tag{100}$$

and

$$\mathbf{R} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \lambda^2\mathbf{D}\right)^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{A} = \frac{\mathbf{A}^{\mathrm{T}}\mathbf{A}}{\mathbf{A}^{\mathrm{T}}\mathbf{A} + \lambda^2\mathbf{D}} \neq \mathbf{I} \quad \text{for} \quad \lambda^2 \neq 0\,. \tag{101}$$

Here $\mathbf{R}$ was written as a ratio for visual purposes only. Note that $\mathbf{R}$ is not equal to $\mathbf{I}$ when $\lambda^2 = 0$ unless $(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}$ exists. On the other hand, from Eq. (85) we know that the inverse of the matrix in parentheses always exists because $\mathbf{D}$ is assumed to be positive definite.

Now consider the implications of a resolution matrix not equal to the identity. Using Eq. (98) we can write the ith component of the solution vector as

$$\hat{x}_i = \sum_k R_{ik} x_k\,, \tag{102}$$

so that $\hat{x}_i$ is equal to a linear combination of some of the components of $\mathbf{x}$, not to x_i (as would be the case if $\mathbf{R} = \mathbf{I}$). This result can be interpreted as follows. As the number of parameters in a velocity model is increased (e. g., the number of blocks is increased), it is expected that it will become a more realistic representation of the actual velocity variations in the earth. However, the available data do not allow the determination of individual values of the parameters; only average values can be computed. We may summarize this situation by saying that in geophysics, as in economics, "there is no such thing as a free lunch". Note that $\mathbf{R}$ does not depend on the the data. Rather, it is controlled by the distribution of stations and seismic sources. Seismologists have some control on the former (within financial and logistic limits), but not on the latter, which places a strong constraint on the resolution that can be achieved in any given case. The Bayesian approach discussed below is an attempt to circumvent this intrinsic limit.

The computation of the resolution matrix when the number of velocity parameters is very large is computationally too expensive, and in such cases the resolution is estimated using synthetic arrival times using velocity models with simple patterns. An early approach [52] was based on the use of several high velocity areas extending through the whole model depth embedded in a constant velocity medium. The synthetic data were computed using the data raypaths and inverted as the actual data. Comparison of the model obtained by inversion with the model used to generate the synthetic data gives an idea of the resolving capabilities of the distribution of sources and receivers used. In a similar approach, synthetic data were generated for a checkerboard model with alternating positive and negative velocities in three dimensions [72]. In another variation only two blocks of the model had velocities different from those of the rest of the blocks [70]. The checkerboard approach has become very popular, but in principle, other synthetic data sets can be used (see Sect. "Examples").

Finally, let us note that in addition to its mathematical definition, the term resolution is generally used (i. e., not only in seismology) to indicate level of detail. For example, given two images of the same physical object acquired using different instruments, the one showing more detail is said to have more resolution. The same idea applies to velocity models, and it it is not uncommon to see references in the literature to high- or low-resolution models. These terms are useful in a qualitative way, and may not be directly related to what one would call high or low in a mathematical sense. In fact, $\mathbf{R}$ can be made closer (and even equal) to $\mathbf{I}$ by decreasing the number of parameters (see, e. g., [74]), which by itself will decrease the level of detail in the model. For example, if a number of blocks in a model is replaced by a single block occupying the same volume, the velocity in the larger block will be the average (in some sense) of the velocities in the smaller blocks. As a consequence, if the small-block velocities are significantly different, the model based on the larger blocks would have lost detail and, thus, will have less resolution, although mathematically it may be larger.

Bayesian Approach

This approach is based on probability considerations, and to discuss it we need a few basic definitions and results, based on [130]. Let X and Y be random variables and x and y be elements of a discrete set of real values. Also let $P_X(x)$ be the probability that the event "$X = x$" occurs and $P_{X,Y}(x, y)$ be the probability that the event "$X = x$ and $Y = y$" occurs. Then the *conditional probability* that

the event "$Y = y$" occurs given that the event "$X = x$" has occurred is given by

$$P_{Y|X}(y|x) = \frac{P_{X,Y}(x, y)}{P_X(x)} . \tag{103}$$

Similarly,

$$P_{X|Y}(x|y) = \frac{P_{X,Y}(x, y)}{P_Y(y)} . \tag{104}$$

These two equations can be combined by elimination of the common factor, giving

$$P_{X|Y}(x|y) = \frac{P_{Y|X}(y|x)P_X(x)}{P_Y(y)} . \tag{105}$$

This result is known as *Bayes' theorem* (or rule).

When the random variables can take on values defined over a continuous interval, which is the case of interest to us, the following definitions are introduced.

(a) *Probability distribution function* $F_X(x)$:

$$F_X(x) = P(-\infty < X \leq x) \tag{106}$$

where P indicates the probability of the event in parentheses.

(b) *Probability density function* $p_X(x)$:

$$p_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x} . \tag{107}$$

From Eq. (107) it follows that

$$F_X(x) = \int_{-\infty}^{x} p_X(a)\mathrm{d}a . \tag{108}$$

A well known example of probability density function (or pdf, for short) is the Gaussian pdf, introduced below.

The definitions given above are for a single random function. When dealing with a number of them it is convenient to introduce a vector notation. For the case of n random variables $X_1, X_2, \ldots, X_n$, we have the following two definitions

$$p_{\mathbf{X}}(\mathbf{x}) = p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots , x_n) \tag{109}$$

and

$$F_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{\mathbf{x}} p_{\mathbf{X}}(\mathbf{a})\mathrm{d}\mathbf{a} \tag{110}$$

where the subscript in $p_{\mathbf{X}}$ and $F_{\mathbf{X}}$ is the vector $\mathbf{X} = (X_1\, X_2\, \ldots\, X_n)^{\mathrm{T}}$, the integral symbol represents an n-fold integral, and $\mathrm{d}\mathbf{a} = \mathrm{d}a_1\mathrm{d}a_2 \ldots \mathrm{d}a_n$. Then

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^n F_{\mathbf{X}}(\mathbf{x})}{\partial x_1 \partial x_2 \ldots \partial x_n} \tag{111}$$

and

$$\begin{aligned} p_{\mathbf{X}}(\mathbf{x})\,\mathrm{d}\mathbf{x} &= F_{\mathbf{X}}(\mathbf{x} + \mathrm{d}\mathbf{x}) - F_{\mathbf{X}}(\mathbf{x}) \\ &= P(\mathbf{x} < \mathbf{X} \leq \mathbf{x} + \mathrm{d}\mathbf{x}) \end{aligned} \tag{112}$$

where $\mathrm{d}\mathbf{x}$ is the vector with ith component given by $\mathrm{d}x_i$.

For vector random variables Bayes' theorem becomes

$$p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = \frac{p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\, p_{\mathbf{X}}(\mathbf{x})}{p_{\mathbf{Y}}(\mathbf{y})} . \tag{113}$$

In the following we will adopt the common practice of dropping the subscripts of the probability density functions.

The application of Bayes' theorem to inverse problems is based on the following interpretation of the theorem. Before an experiment is conducted, a random variable $\mathbf{X}$ has a given a priori pdf $p(\mathbf{x})$. During the experiment data $\mathbf{y}$ are collected, and as a result $\mathbf{X}$ has an a posteriori pdf given by $p(\mathbf{x}|\mathbf{y})$ (see, e. g. p. 36 in [10]). The pdf $p(\mathbf{y}|\mathbf{x})$ represents the probability that an experiment would have produced the result $\mathbf{y}$ if the value of the variable had been $\mathbf{x}$. To apply Bayes' theorem to our problem, we need to rewrite Eq. (74) adding a random error vector $\mathbf{e}$

$$\mathbf{A}\mathbf{x} = \mathbf{b} + \mathbf{e} \tag{114}$$

and to assume that $\mathbf{x}$ and $\mathbf{b}$ are random variables. In this formulation the data are represented by $\mathbf{b}$ (which plays the role of $\mathbf{y}$), and $p(\mathbf{b}|\mathbf{x})$ for a given $\mathbf{x}$ is the probability of the error $\mathbf{e}$ [71]. The denominator in Eq. (113) does not depend on $\mathbf{x}$ and can be ignored. The question that remains is the selection of the form of the probability density functions on the right side of the equation. The standard choice is a *Gaussian* pdf, given, in general, by

$$\begin{aligned} p(\mathbf{x}) = {} & (2\pi)^{-n/2}|\mathbf{C}_{\mathbf{x}}|^{-1/2} \\ & \cdot \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^{\mathrm{T}}\mathbf{C}_{\mathbf{x}}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})\right] \end{aligned} \tag{115}$$

where n is as in Eq. (109), $\mathbf{C}_{\mathbf{x}}$ and $\boldsymbol{\mu}_{\mathbf{x}}$ are the covariance matrix and mean of $\mathbf{x}$, respectively, and the vertical bars indicate determinant. Introducing this expression and a similar one for $p(\mathbf{e})$ in Eq. (113), and then taking logarithms on both sides of the resulting equation gives

$$\begin{aligned} \ln[p(\mathbf{x}|\mathbf{b})] = {} & -\frac{1}{2}\left[(m+n)\ln(2\pi) + \ln|\mathbf{C}_{\mathbf{e}}| + \ln|\mathbf{C}_{\mathbf{x}}| + S\right] \\ & - \ln(p(\mathbf{y})) \end{aligned} \tag{116}$$

where m is the number of components of $\mathbf{b}$ and

$$S = (\mathbf{e}-\boldsymbol{\mu}_{\mathbf{e}})^{\mathrm{T}}\mathbf{C}_{\mathbf{e}}^{-1}(\mathbf{e}-\boldsymbol{\mu}_{\mathbf{e}}) + (\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}})^{\mathrm{T}}\mathbf{C}_{\mathbf{x}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{x}}). \tag{117}$$

Now we will make the additional assumption that $\boldsymbol{\mu}_\mathbf{e} = \mathbf{0}$ and will write $\mathbf{e} = \mathbf{A}\mathbf{x} - \mathbf{b}$ (using Eq. (114)). Under these conditions Eq. (117) becomes

$$S = (\mathbf{A}\mathbf{x}-\mathbf{b})^{\mathrm{T}}\mathbf{C}_\mathbf{e}^{-1}(\mathbf{A}\mathbf{x}-\mathbf{b})+(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})^{\mathrm{T}}\mathbf{C}_\mathbf{x}^{-1}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x}). \quad (118)$$

The last step is to find the value of $\mathbf{x}$ that maximizes $\ln p(\mathbf{x}|\mathbf{b})$ (and thus, $p(\mathbf{x}|\mathbf{b})$) and, therefore, minimizes S. Taking the derivative of S with respect to $\mathbf{x}$ and setting it equal to zero leads to the *maximum a posteriori estimator* $\mathbf{x}_{\mathrm{MAP}}$, given by

$$\mathbf{x}_{\mathrm{MAP}} = \mathbf{P}\left(\mathbf{A}^{\mathrm{T}}\mathbf{C}_\mathbf{e}^{-1}\mathbf{b} + \mathbf{C}_\mathbf{x}^{-1}\boldsymbol{\mu}_\mathbf{x}\right) \quad (119)$$

where

$$\mathbf{P} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{C}_\mathbf{e}^{-1}\mathbf{A} + \mathbf{C}_\mathbf{x}^{-1}\right)^{-1} . \quad (120)$$

An alternative expression for $\mathbf{x}_{\mathrm{MAP}}$ is

$$\mathbf{x}_{\mathrm{MAP}} = \boldsymbol{\mu}_\mathbf{x} + \mathbf{P}\mathbf{A}^{\mathrm{T}}\mathbf{C}_\mathbf{e}^{-1}(\mathbf{b} - \mathbf{A}\boldsymbol{\mu}_\mathbf{x}) . \quad (121)$$

These results are based on [13]. An expression similar to Eq. (119) is given in [130]. The Bayesian approach presented here was introduced in the context of X-ray tomography under the assumption that $\boldsymbol{\mu}_\mathbf{x} = \mathbf{0}$ [71]. An application can be found in [64]. In geophysics a similar approach was introduced in [141] and [142].

Equation (121) is important for two reasons. First, $\mathbf{x}_{\mathrm{MAP}}$ can be interpreted as the sum of two terms: one corresponding to the prior information about the value of $\mathbf{x}$ (given by $\boldsymbol{\mu}_\mathbf{x}$) and another one that contains the new information provided by the data collected in the experiment. Second, using Eq. (121) and the fact that the mean values of $\mathbf{x}$ and $\mathbf{e}$ are $\boldsymbol{\mu}_\mathbf{x}$ and $\mathbf{0}$, respectively, it can be shown that the expected value of $\mathbf{x}_{\mathrm{MAP}}$ is $\boldsymbol{\mu}_\mathbf{x}$. Therefore, the estimator is *biased* (i. e., its expected value is not $\mathbf{x}$). The significance of this result is that the expected value of the estimator is independent of the data, which is not a desirable feature. Finally, we also note that the covariance matrix of the difference $\mathbf{x}_{\mathrm{MAP}} - \mathbf{x}$ is equal to the matrix $\mathbf{P}$ [13].

Now let us consider two special cases. First, no a priori information is used. In this case S in Eq. (118) is equal to the first term on the right hand side and its minimization leads to the *generalized least-squares estimator*, given by

$$\hat{\mathbf{x}} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{C}_\mathbf{e}^{-1}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{C}_\mathbf{e}^{-1}\mathbf{b} \quad (122)$$

(if the inverse exists). The *ordinary least-squares estimator* is obtained from Eq. (122) with $\mathbf{C}_\mathbf{e} = \sigma_e^2\mathbf{I}$, where σ_e is a measure of the standard deviation of the errors. The estimator $\hat{\mathbf{x}}$ has the properties that it is unbiased and minimizes the determinant of its covariance matrix, and does not require a Gaussian pdf (see, e. g., [10]).

Second, let us assume that the a priori information is $\boldsymbol{\mu}_\mathbf{x} = \mathbf{0}$ and $\mathbf{C}_\mathbf{x} \neq \mathbf{O}$. Then, using Eqs. (119) and (120) we get

$$\begin{aligned}\mathbf{x}_{\mathrm{MAP}} &= \left(\mathbf{A}^{\mathrm{T}}\mathbf{C}_\mathbf{e}^{-1}\mathbf{A} + \mathbf{C}_\mathbf{x}^{-1}\right)^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{C}_\mathbf{e}^{-1}\mathbf{b} \\ &= \mathbf{C}_\mathbf{x}\mathbf{A}^{\mathrm{T}}\left(\mathbf{A}\mathbf{C}_\mathbf{x}\mathbf{A}^{\mathrm{T}} + \mathbf{C}_\mathbf{e}\right)^{-1}\mathbf{b} . \end{aligned} \quad (123)$$

A derivation of the second equality is given in [13]. This result agrees with similar results derived in [47] and [74] using different approaches. An alternative derivation can be found in [3]. If we further simplify the problem by letting $\mathbf{C}_\mathbf{e} = \sigma_e^2\mathbf{I}$ and $\mathbf{C}_\mathbf{x} = \sigma_x^2\mathbf{I}$, the first equality in Eq. (123) gives

$$\begin{aligned}\mathbf{x}_{\mathrm{MAP}} &= \left(\frac{1}{\sigma_e^2}\mathbf{A}^{\mathrm{T}}\mathbf{A} + \frac{1}{\sigma_x^2}\mathbf{I}\right)^{-1}\frac{1}{\sigma_e^2}\mathbf{A}^{\mathrm{T}}\mathbf{b} \\ &= \left(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \frac{\sigma_e^2}{\sigma_x^2}\mathbf{I}\right)^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{b} . \end{aligned} \quad (124)$$

This particular form of $\mathbf{x}_{\mathrm{MAP}}$ has been referred to as the *stochastic inverse* [2,4]. If we now put $\lambda = \sigma_e/\sigma_x$ we see that this solution is formally similar to the damped least-squares solution obtained from Eq. (87) with $\mathbf{D} = \mathbf{I}$.

Another important result concerning $\mathbf{x}_{\mathrm{MAP}}$, as given in Eq. (119), is that it can be obtained by simply treating the prior information as a constraint, i. e., $\mathbf{x} = \boldsymbol{\mu}_\mathbf{x}$, which then can be added to Eq. (114) in a way similar to what was done in Eqs. (88) and (96), and then finding the generalized least-squares estimator, defined by Eq. (122) [74,136]. When proceeding in this way there is no need to make the Gaussianity assumption, although without it (or some other assumption) it is not possible to derive statistical properties of the estimator.

Finally, we note that although the Bayesian approach seems to be a good way to handle the problem of low resolution that affects many seismic tomography problems, it has some serious drawbacks that should not be overlooked. As noted earlier, the Bayesian solution is biased by the prior information, and if this information is incorrect, the solution will be affected by some amount of error that cannot to be quantified objectively. The assumption that the velocity adjustments can be described by a Gaussian pdf (or any other simple pdf) is introduced by mathematical convenience, not because it is likely to represent the actual 3-D variations of the velocity adjustments within the earth. This applies, particularly, to crustal areas within complicated tectonic settings. Finally, the covariance matrix is also needed, but because it is essentially unknowable, it is frequently assumed that it is diagonal. Again, this is an assumption based on convenience, rather than on scientific facts. A critique of the Bayesian approach as applied

to geophysical inverse problems can be found in [109]. For a somewhat different perspective see [104]. A good overview of the Bayesian method can be found in [57]. An example of the problems that erroneous prior information may introduce is provided in Subsect. "Effect of Inaccurate Prior Information on the Location of Aftershocks".

Examples

Here we present two examples of local tomography, one from the San Fernando basin in southern California and one from Taiwan, and also give an example of the problems that can be created when erroneous prior information is used. The tomography software used was written by H. Benz [15] and solves for earthquake locations and slownesses using the formulation described in Subsect. "Local Slowness Tomography" – Subsect. "Decoupling of the Earthquake Location and Tomography Problems". The slowness model is parameterized in terms of constant velocity blocks and the computation of arrival times is performed using the software of Podvin and Lecomte [113]. As noted in Subsect. "Computation of Local Travel Times", this software has become popular because of its ability to handle sharp velocity variations accurately. As a consequence, the resulting velocity models show high resolution (i. e. a high level of detail). For the examples presented here, none of the other existing models show such resolution. Additional results derived using Benz's software can be found in, e. g., [95,105] and [159]. The original software relocated the events with a constant v_p/v_s ratio but was modified according to Eq. (65). To regularize the solution Eq. (96) is used.

P Wave Velocity Model for the San Fernando, California, Area

The Los Angeles, California, region is one of the most seismically active in the United States, and because of its extremely large population (over ten million), a large earthquake there may be catastrophic in terms of human and/or economic costs. A major source of seismic hazard there is the presence of large sedimentary basins, which have much lower velocity than their host rocks. As a consequence, they can amplify significantly the ground motion caused by earthquakes. For example, a study published in 1989 [160] showed that the ground motion in the San Fernando and Los Angeles basins can be amplified by factors of about three and four with respect to rock sites, and that the 3-D velocity models available were not capable of generating this amplification. To address this and other problems, the Southern California Earthquake Center (SCEC) supported the development of reference of 3-D *P* and *S* wave velocity models for the major basins of southern California [91]. These models were constructed using depth and age of sediments data compiled as part of oil and water exploration studies and geological studies. Empirical relations between these two types of data were used to estimate *P* wave velocities. Additional empirical relations between *P* wave velocities, density, and Poisson's ratio were used to calculate *S* wave velocity. For depths less than 300 m, the *P* and *S* wave velocities were constrained using borehole velocity data. Deep borehole information was used for calibration purposes. For the San Fernando basin four boreholes were available, but the deepest one was only one 3.5 km deep. Therefore, below that depth the model does not have hard constraints. For the rocks outside of the basins an existing 3-D tomographic velocity model [58] was used. This model assigns velocities to points on a grid, which has a 15 km x 15 km horizontal spacing at depths of 1, 4, 6, 10, 15, 17, 22, 31 and 33 km.

The *P* wave velocities in the SCEC model for the Los Angeles basin were compared to sonic log velocities in [140]. This comparison shows that the standard deviation of the velocity differences is about 440 m/s, which is up to 20% of the model velocities. An underestimation of the velocities near the center of the basin and an overestimation near the border was also observed. An additional comparison [139] produced results that are consistent with those in [140].

The 1994, $M = 6.7$, Northridge, California, earthquake occurred on a previously unknown blind thrust fault in the San Fernando valley, to the NW of Los Angeles. This was the costliest earthquake in the United States (about US$ 20,000 million), although the number of deaths was small (58, [143]), thanks to building code provisions and the fact that the event occurred early in the morning (4:30 AM local time). Because of its importance, this earthquake has been extensively studied, but more than ten years after its occurrence, a number of important questions still remain partially or totally unanswered (see, e. g., [123]). One of them is the exact nature of the 3-D velocity variations in the area, as the existing velocity models (see, e. g., [59,97,118,162]) have low resolution.

The velocity model issue has been addressed in [123]. The Benz's tomography software was applied to events in the Northridge mainshock-aftershock sequence and to aftershocks of the nearby 1971, $M = 6.6$, San Fernando earthquake. The data used were 192,421 *P* wave first arrivals from 12,656 events recorded during 1981–2000 by 81 permanent and temporary stations and 799 aftershocks of the San Fernando earthquake recorded by a portable network of 20 stations (Fig. 4). The velocity model covers a volume with a surface area of 80 km by 80 km and a depth

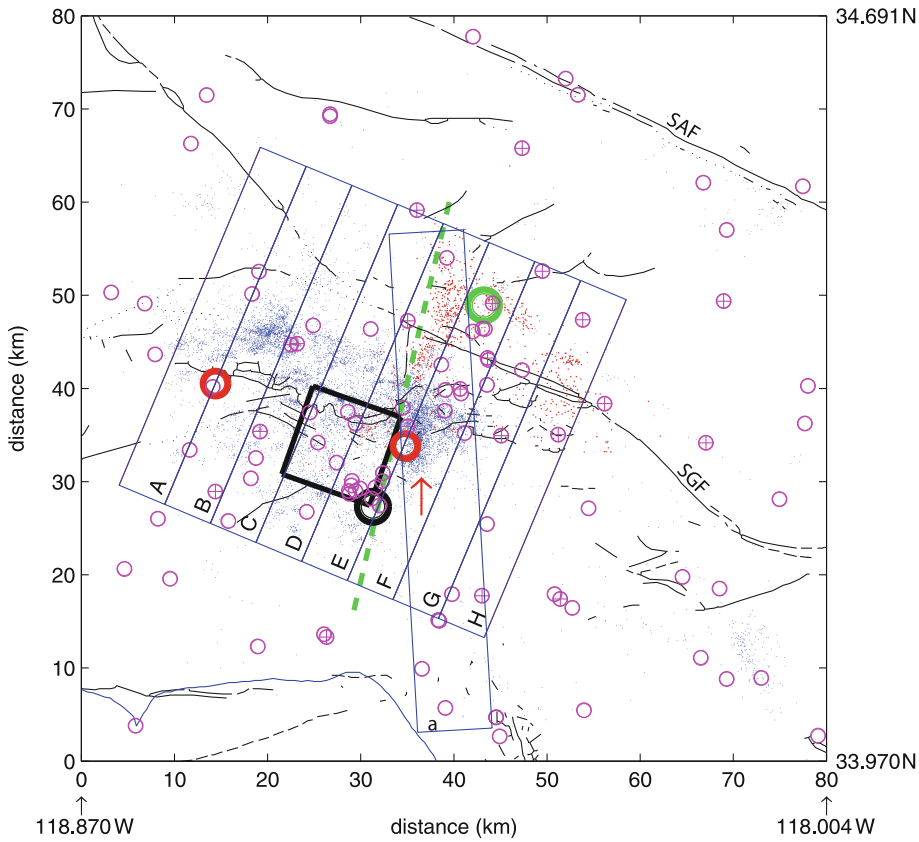

Tomography, Seismic, Figure 4

Area around the San Fernando valley, California, for which a 3-D *P* wave velocity model has been determined. *Blue dots* indicate the epicenters of 12,656 seismic events recorded during 1981–2000 (mostly aftershocks of the 1994, $M = 6.7$, Northridge earthquake). *Red dots* indicate epicenters of 799 aftershocks of the 1971, $M = 6.6$, San Fernando earthquake recorded during February-April 1971. The locations shown here were determined as part of the velocity determination and earthquake relocation. The large *black* and *green circles* indicate the epicenters of the Northridge and San Fernando mainshocks, respectively, the latter provided by the U.S. Geological Survey (U.S.G.S.). *Magenta open* and *crossed circles* indicate the stations that recorded the Northridge and San Fernando events, respectively. The large *red circles* indicate the epicenters of the two largest Northridge aftershocks (the location of the eastern one is from the U.S.G.S. and is poorly constrained). *Black lines* indicate faults. The San Andreas and San Gabriel faults are labeled SAF and SGF. The bold box is the projection of the Northridge earthquake fault plane determined in [68] using geodetic data. The events and velocities in boxes A through H are shown in cross section form in Fig. 5. Most of the San Fernando events are to the east of the *green dashed line* while the Northridge mainshock rupture occurred to the west of that line. The prominent band of Northridge seismicity in a roughly N-S direction (identified by the red arrow in box F) did not occur within the area that ruptured during the main shock. The events and velocity within the box labeled a are shown in cross section form in Fig. 5. From [123]

of 24 km and was divided into cubic blocks with sides of 2 km. For the computation of travel times cubic blocks with sides of 1 km were used. The initial locations were computed using a standard 1-D velocity model with three layers having thicknesses of 5.5, 10.5 and 16 km, and corresponding velocities of 5.5, 6.3, and 6.7 km/s. For the inversion the initial velocity model was the 1-D model described in [60] with a small modification (the velocity in the upper 2 km was reduced from 4.8 to 4.5 km/s). A cross section of the model is shown in a subsequent figure. The number of iterations was ten, and the value of λ in Eq. (96) ranged between 64 (first four iterations) and 20 (last two iterations) with intermediate values in between. Because some of the stations are in the mountains, a depth of −2 km was used as a reference depth for the model. The root-mean square residual for all the events was 0.17 s and 0.07 s for the first and last iterations. Representative velocity cross sections (Fig. 5) show that the resulting model has much more detail than any of the other published models. To make sure that this detail was not artificial the initial model was

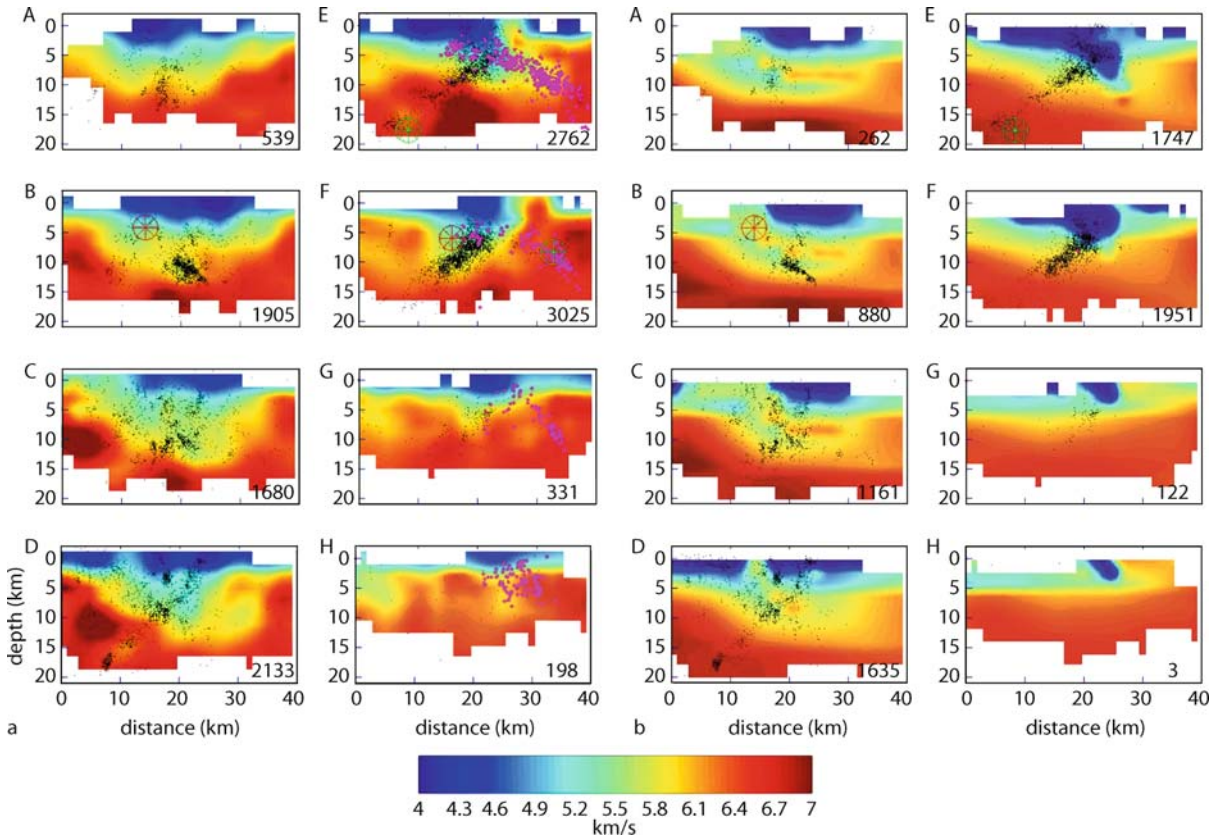

Tomography, Seismic, Figure 5

a Depth cross sections for the 3-D *P* wave velocity model determined by inversion and the events in the boxes A through H in Fig. 4. The width of the cross sections is 5.3 km. The letters are on the southern ends of the cross sections. The velocities are assigned to the centers of the model blocks and are interpolated along planes passing through the centers of the cross sections. *Black* and *magenta dots* indicate Northridge and San Fernando aftershocks. The large circled asterisks in cross sections E, B and F indicate the Northridge mainshock, its two largest aftershocks and the San Fernando mainshock (see Fig. 4). The number in the right lower corner of each cross section denotes the number of events. Only the velocity blocks covered by a combined ray length of 0.1 km or more are shown. Note the correlation between seismicity and velocity. The events between about 10 and 20 km in D and E are within high-velocity, basement rocks, and form narrow and well-defined lineations. These events span about 10 km horizontally and basically define the width of the fault that slipped during the main shock. The Northridge aftershocks in F correspond mostly to those indicated by the *red arrow* in Fig. 4. Most of these events are shallower than about 14 km and form a band of seismicity within and near the edge of the basin. The north-dipping events in B below about 10 km probably occurred on the Santa Susana fault. b Corresponding cross sections for the SCEC 3-D velocity model [91] and the events that occurred during 1994. From [123]

changed and different data subsets were used, with the result that the most important aspects of the model were robust. In addition, realistic synthetic arrival times were generated with the velocity model and event locations from the last iteration. This involved using the stations that had recorded the observed data (for each event) and applying the original weights. Then the synthetic arrivals times were inverted as the actual data (i. e., the same initial locations and velocity model were used). The corresponding velocity model is very close to the input model (Fig. 6). A few areas have velocity differences of up to ±0.5 km/s, but this does not detract from an overall good agreement. Regarding the hypocentral locations, the average difference between the true and computed values is 0.15 km in epicenter and 0.23 km in depth.

The inversion model is also supported by two additional pieces of evidence. One is the SCEC model, which was sampled at the centers of the blocks in the inversion model. However, because the SCEC model has the surface as zero depth and the elevation in the area ranges between 0 and 1.6 km, a direct depth comparison is not strictly possible and for comparison purposes it was re-

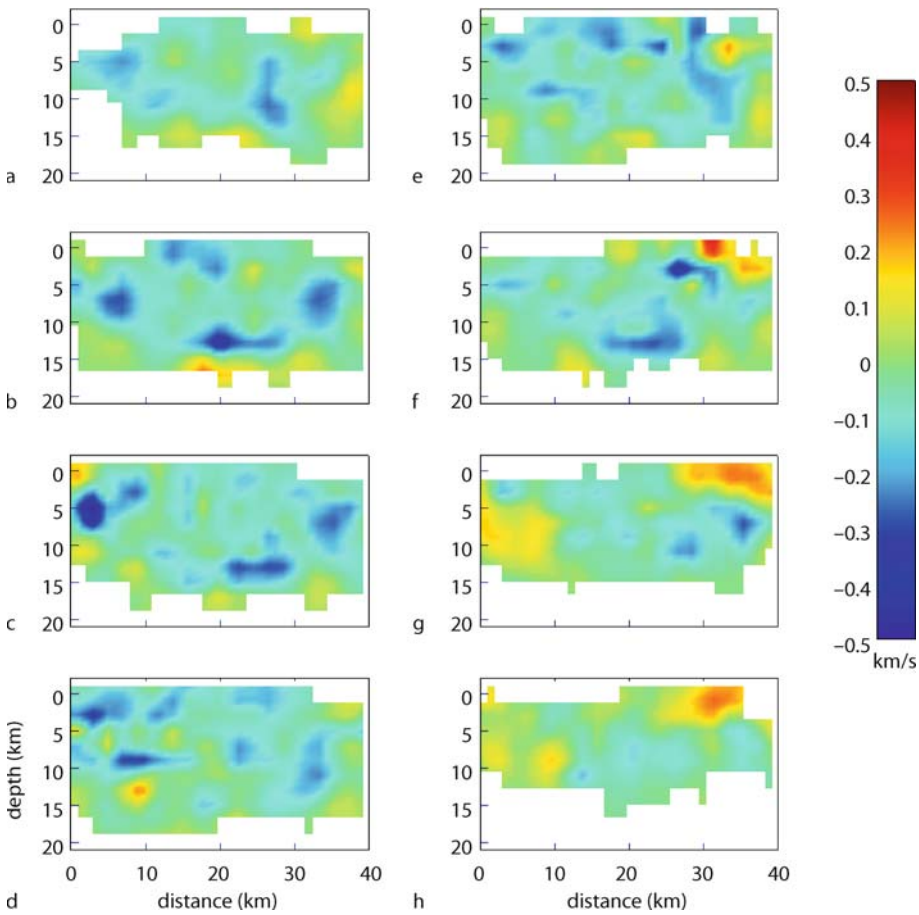

Tomography, Seismic, Figure 6

a Depth cross sections of the difference between the 3-D model shown in Fig. 5a and the 3-D model determined by inversion of synthetic arrival times generated using the model in Fig. 5a and the event locations determined during the inversion. See text for further details. From [123]

ferred to a base elevation of −0.6 km. As Fig. 5 shows, the inversion model has the main features of the SCEC model, and although there are some obvious differences, some of them come from deficiencies in the SCEC model, as discussed next. A second piece of evidence is the excellent agreement (Fig. 7) with a roughly 55 km long density model derived from the analysis of gravity data [82]. In particular, the low-velocity areas in the model correlate with the Los Angeles, San Fernando and Soledad basins, while high velocities underlie the Santa Monica and San Gabriel mountains. In contrast, the SCEC model does not match the density model equally well. Figure 7 also shows another tomographic model [59], computed for a grid with a 10 km × 10 km horizontal spacing at depths of 1, 4, 6, 8, 12, 16 and 20 km. Clearly, this model has a poor resolution and may be affected by significant artifacts, which may be a direct consequence of the large horizontal dimensions of the blocks. Therefore, for a more direct comparison the inversion with Benz's software was repeated with blocks having horizontal sides of 10 km. As Fig. 8 shows, the corresponding results closely resembles those in Fig. 7b, which seems to indicate that the lack of resolution of the model in Fig. 7d may be due to either the software or the inversion parameters used, or both.

The results of the event relocation are also important. There are two types of differences between the initial and final locations. One is quasi-systematic, with the initial locations on average 1.0 km to the east and 0.5 km to the south of the inversion locations, and 1.5 km deeper. These epicentral differences are a consequence of the low velocities within the basin, which bias the event locations as noted earlier (see Fig. 3). These results are in agreement with a similar shift found using the joint hypocentral determination (JHD) technique [118], and have not been re-

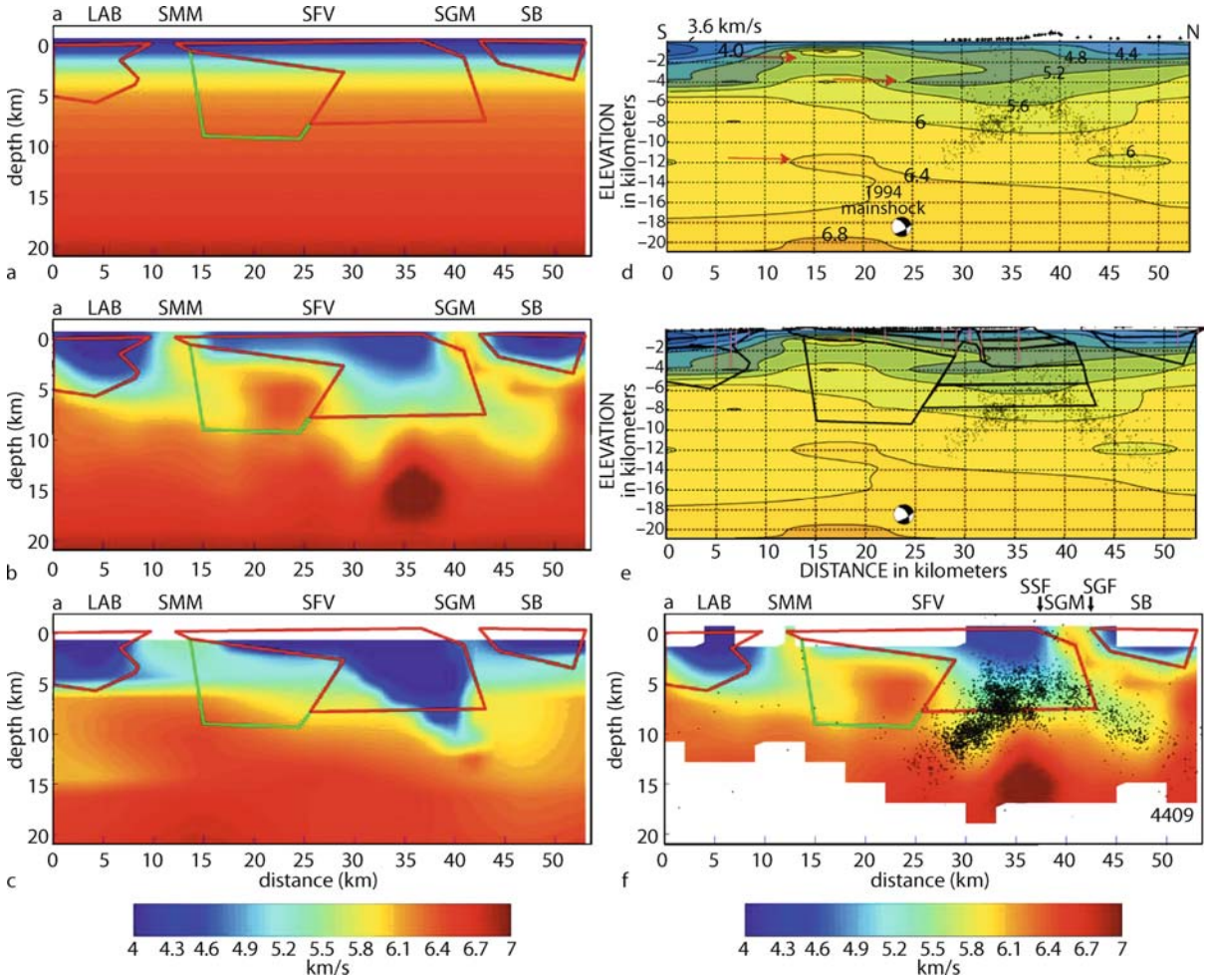

Tomography, Seismic, Figure 7

Velocity depth cross section along the center of box a in Fig. 5 for several velocity models. a Initial model. b Inversion model. All the blocks, regardless of the ray coverage, are shown. The polygons represent the bodies (simplified) used in [82] to match a gravity profile. The *red lines* bound materials with densities ranging between 2.00 and 2.55 g/cc. The area bounded by *red* and *green lines* corresponds to low-density basement (2.65 g/cc). Elsewhere in the figure the density is 2.71 g/cc. Note the excellent agreement between the extent of the low velocities in the Los Angeles (LAB), San Fernando (SFV) and Soledad (SB) basins and the low-density bodies, as well as the presence of high velocities in the areas of the Santa Monica (SMM) and the San Gabriel (SGM) mountains. c SCEC 3-D velocity model. This model does not fit the density model as well as the inversion model. d and e the 3-D velocity model described in [59] and original density model (from [82]). The contour labels in d indicate velocities in km/s. The *red arrows* indicate possible artifacts in the model. *Crosses* indicate the locations of gravity stations. f Inversion velocity model for blocks having ray coverage of 0.1 km or larger. The fact that the initial model is one-dimensional indicates that the 3-D velocity variations seen in a for the blocks not shown here were determined in earlier iterations. Also shown are the events within box a in Fig. 4. The box width is 8 km. Note that the Northridge and San Fernando aftershocks are underlain by a wedge of basement, with the seismicity occurring where rocks with lower velocities are present. The *arrows* labeled SSF and SGF indicate the positions of the Santa Susana and San Gabriel faults. From [123]

ported in other published tomographic inversion papers. A second difference between the initial and final locations is a considerable reduction in the epicentral scatter seen in the single-event locations. On average, there is a 1.5 km difference between the single-event and inversion epicentral locations.

The combination of a high-resolution velocity model and improved event locations has important tectonic implications. For example, Fig. 5a shows that the most of the seismicity occurred within the sedimentary rocks of the basin. There is little activity within the basement, mostly confined to an area around the fault plane of

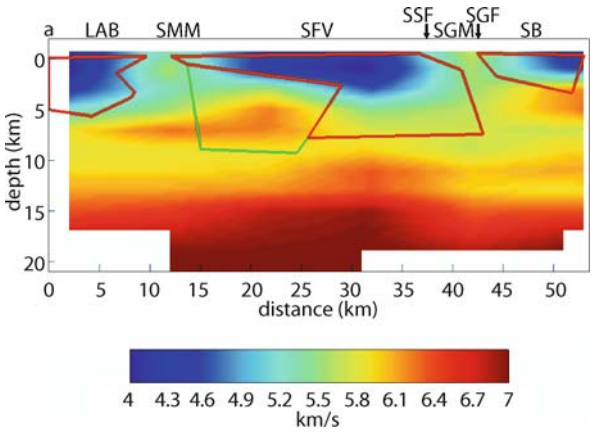

Tomography, Seismic, Figure 8

Similar to Fig. 7f for a velocity model with blocks having the two horizontal sides and the vertical side 10 km and 2 km long respectively. Comparison with Fig. 7b shows that even though the block size is horizontally much larger, the major features of the model are recovered

the Northridge earthquake. The relation between the Northridge earthquake aftershocks to the east of the main shock and San Fernando aftershocks has also been clarified; they occur on the flanks of a common high velocity block (Fig. 7f). Other significant results, not discussed here, can be found in [123].

P and *S* Wave Velocity Models for Taiwan

The seismic hazard in Taiwan is also very high and for this reason a large network of seismic stations covers the island (Fig. 9). In addition, the collision and subduction processes involving the Eurasian and Philippine plates in the Taiwan region have resulted in very strong ongoing tectonic and orogenic activities. As a consequence, Taiwan is the focus of numerous studies, past and present, aimed at getting a better understanding of the nature of the deformation processes there. Some of those studies involve the determination of the crustal an upper mantle velocity structure, which has been investigated by a number of researchers (see, e. g., [90,126,128]). However, although the resulting information has been highly valuable, the resolution of the velocity models was relatively low. This situation was improved by the work described in [77], which resulted in *P* and *S* wave velocity models with much higher resolutions. The main results will be described here (see also [78]). As noted above, the inversion was carried out using Benz's software. The block size was 8 km × 8 km (horizontally) and 2 km in depth for velocity and 2 km × 2 km × 2 km for travel time computation. The dataset included 69,758 *P* wave arrivals and 42,733 *S* wave arrivals from 6285 events distributed as uniformly as possible recorded by 78 permanent stations as well as 18,502 and 10,789 *P* and *S* wave arrivals, respectively, from 1298 events recorded during two 30-station portable deployments (Fig. 10). The initial and final values of λ in Eq. (96) for the *P* and *S* wave inversions were 128 and 50 and 128 and 20, respectively. The number of iterations was 20 and the initial and final average root mean square residuals were 0.58 s and 0.15 s and 0.67 s and 0.21 s for the *P* and *S* waves, respectively. To avoid exceeding the limitations of the flat-earth approximation (see Subsect. "Computation of Local Travel Times"), station-event pairs with epicentral distances greater than 140 km were not used in the inversion. The number of events used in the inversion was less than 5% of the total number of events recorded. Once the velocity models were determined, they were used to relocate all the locatable events using a 3-D location program based on the tomographic software [24].

Representative cross-sectional views of the computed 3-D *P*- and *S*-wave velocity models are shown in Fig. 11. As in the previous example, these models show higher resolution than other published models. The resolution of these models was investigated with a checkerboard test, which shows that most of the original velocity pattern is well recovered to depths of 25 to 30 km for most blocks (Figs. 12, 13). Below that resolution decreases. Further evidence for the reliability of the velocity models comes from the comparison of observed station corrections and those determined using synthetic data generated with those models. This approach was introduced in [116] and [118]. The station corrections were computed using the JHD technique (see, e. g., [116,119]). The JHD corrections carry information on the lateral velocity under an array and are usually much larger than the corresponding station residuals. For example, in Taiwan they can be up to about ±1 s and ±2 s for *P* and *S* waves (Fig. 14), with the positive corrections corresponding to low-velocity areas (such as sedimentary basins) and the negative corrections associated with high-velocity areas. In general, if the observed and synthetic station corrections do not agree well with each other, it can be stated with confidence that the velocity model used to generate the synthetic data cannot be correct. On the other hand, a good agreement indicates that the inversion model is able to reproduce, at least, the actual velocity variations in an average sense along raypaths. For the Taiwan 3-D velocity model the agreement between the actual and synthetic *P* and *S* wave JHD corrections for two subsets of events is good (see Fig. 14 for an example), which gives confidence to the overall quality of the model.

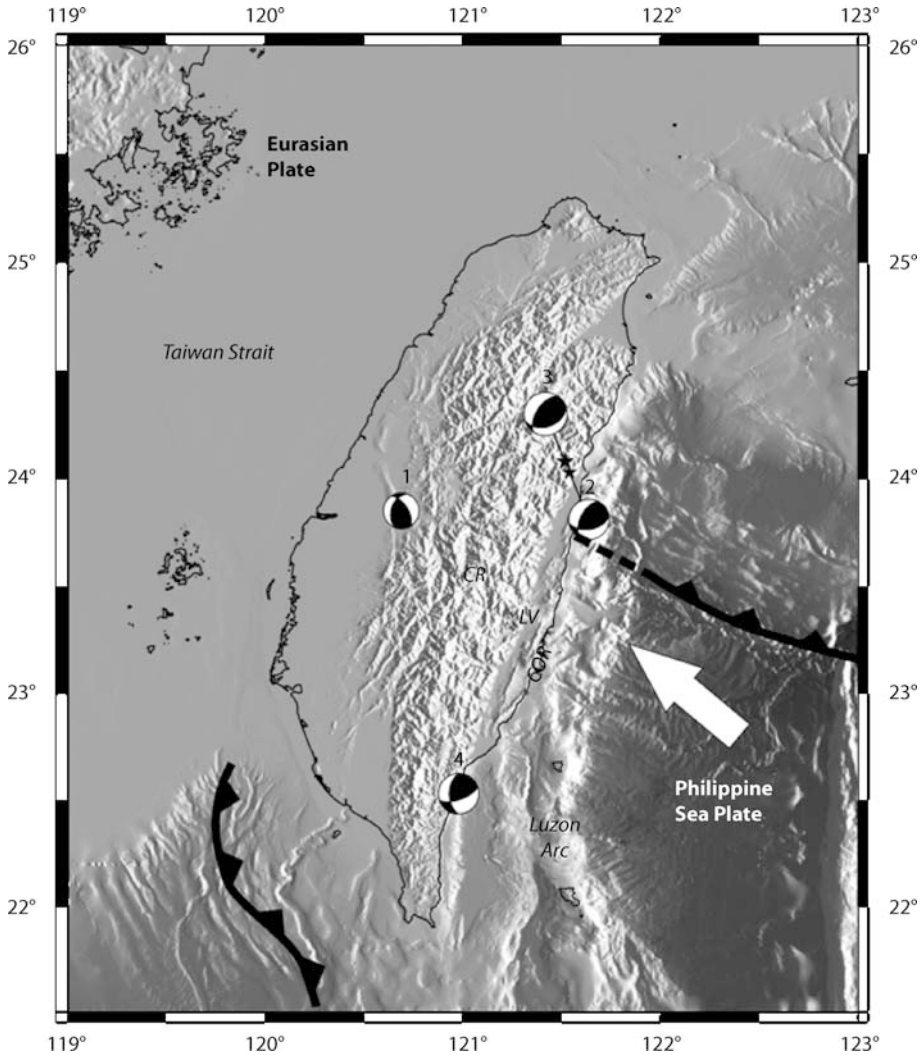

Tomography, Seismic, Figure 9

Tectonic setting of Taiwan and surrounding area. The Philippine Sea plate both subducts beneath, and collides with, the Eurasian plate, while the South China Sea sub-plate subducts beneath the Philippine Sea plate in southern Taiwan. The Longitudinal Valley (LV) is the suture zone corresponding to the collision and separates the Coastal Range (COR, part of the Philippine Sea plate) from the Central Mountain Range (CR). The beach ball symbols numbered 2–4 correspond to earthquakes with abnormal *Pn* waves recorded by stations along the eastern coast. Number 1 corresponds to an earthquake that does not show those *Pn* waves. See text for details. From [89]

Additional evidence in support of the *P* wave model is provided by the analysis of anomalous *Pn* waves recorded by stations along the collision suture zone in eastern Taiwan and generated by shallow eastern Taiwan events [89]. These waves can be observed at stations with epicentral distances as small as 60 km. This critical distances is much smaller than that for stations elsewhere in Taiwan and indicates the presence of an elevated Moho (i. e., a thinner crust) along the suture zone. Figure 15 shows travel times for three eastern earthquakes showing the normal and anomalous *Pn* waves and the corresponding Moho depths (having 36–38 km and 22–23 km depth ranges, respectively). In contrast, an earthquake in western Taiwan only shows the normal *Pn* waves. The presence of thinner crust along the suture zone agrees well with the 3-D velocity described here [77,78], as Fig. 16 shows. Because this zone is on the edge of the network, the resolution, although still acceptable, is not as good as under the network and the details of the model are not well resolved. However, an inversion of synthetic data similar to that de-

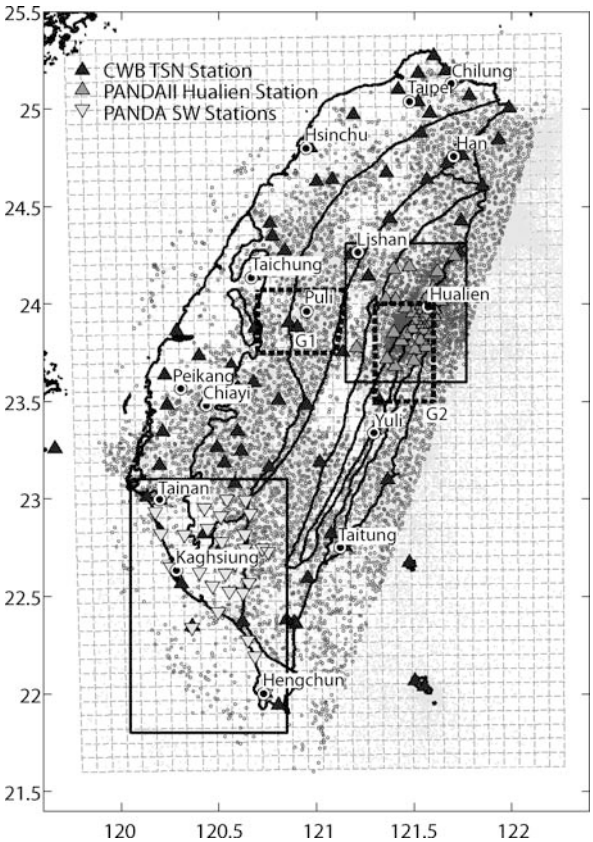

Tomography, Seismic, Figure 10

Seismic stations and events used in the 3-D tomographic study of Taiwan. *Solid triangles* indicate stations of the Taiwan Seismic Network (TSN), operated by the Central Weather Bureau (CWB). *Grey* and *inverted triangles* indicate stations of a portable network (PANDA) deployed in the Hualien and Pingtung areas (*solid rectangles*), respectively. The *small circles* correspond to the epicenters of the events used in the 3-D tomography. The *grey background* corresponds to the epicenters determined by the CWB between 1991 and 2002. The events within the *dashed rectangles* (labeled G1 and G2) were used with the joint hypocentral determination (JHD) technique. The 8 km × 8 km grid of the 3-D velocity is shown by the *dashed lines*. From [78]

scribed for the Northridge data showed that the main features of the velocity model on the eastern side are well recovered [77] and confirms that the profiles in Figs. 11 and 16 are representative of the actual velocity variations.

Effect of Inaccurate Prior Information on the Location of Aftershocks

The relocation of the mainshock and aftershocks of the 1989, $M = 7.1$, Loma Prieta, California, earthquake, offers a good example of the pitfalls that inaccurate prior information may introduce. This sequence occurred along the San Andres fault and was located using two 1-D *P* wave velocity models, one for stations to the southwest of the fault and one for stations to the northeast of it [37]. This division was based on differences in surface geology across the fault. The velocity of the NE model was up to about 0.1 km/s lower in the upper 1 km, and between 0.2 and 0.5 km/s higher between 1 and 9 km depth, with respect to the SE model. Below that depth the velocity differences did not exceed 0.1 km/s (Fig. 17). The two models were derived using a 1-D velocity inversion program that also included the computation of station corrections. According to [37], these models reflected the presence of elevated basement to the NE of the fault and Tertiary and younger sediments above 9 km depth to the SW of the fault. Note that although the prior information was not used as a quantitative constraint, it was essential to establish the boundary between the two velocity models and was implicitly used to justify the main features of the models. The seismicity was located with a single-event location program and these two velocity models and the corresponding station corrections. The mainshock fault plane inferred from the seismicity, however, had a considerable discrepancy with the fault plane determined using geodetic data, with the former consistently to the northeast of the latter by as much as 3 km. A possible explanation for this discrepancy was the presence of lateral variations in the values of the elastic constants in the vicinity of the fault zone, which were not taken into account when the geodetic fault plane was determined [43]. A more likely explanation, however, is that the discrepancy was the result of a systematic mislocation of the events introduced by the use of incorrect velocity models. This problem was identified in [117], where the geodetic fault plane was compared to the event locations computed with the joint hypocentral determination (JHD) technique and a single velocity model, equal to the average of the two models described above [116]. With these new locations, the discrepancy was greatly reduced except in the southern end of the rupture zone, where the difference was about 1.2 m. This residual discrepancy is probably due to a significant lateral velocity contrast there. Subsequently, the analysis described in [37] was repeated with two modified velocity models [38], with the result that the new locations became similar to those in [116]. Interestingly, the new NE and SW velocity models in [38] are significantly different from their earlier counterparts (Fig. 17), with higher velocities to the SW of the San Andreas fault, rather than to the NE. Thus, these new models contradict the geological arguments used as supporting evidence for the earlier models, which in turn means that the prior information was not quite correct. Also worth noting is the fact that

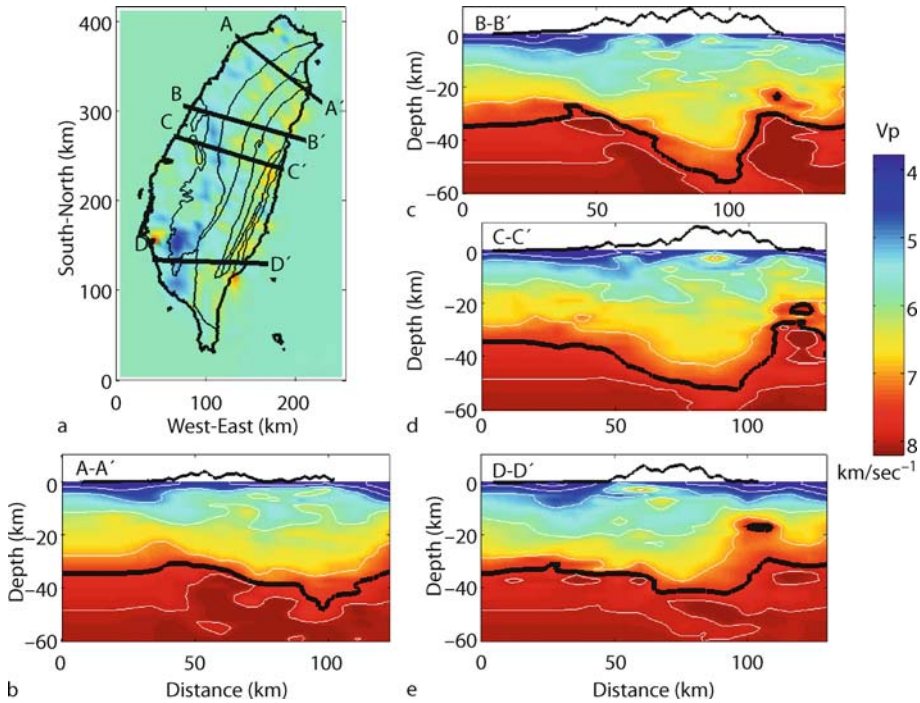

Tomography, Seismic, Figure 11

Cross-sectional views of the Taiwan 3-D *P* wave velocity model. The 7.8 km/s contour line (*black line*) marks the depth to the Moho (approximately). The cross section locations are shown in a. The *bold lines* below 0 depth indicate the 7.8 km/s contour line. The *bold lines* above 0 depth indicate elevation (with a vertical exaggeration of 3). From [78]

the locations determined using a 3-D velocity model also showed the discrepancy with the geodetic fault plane (see, e. g., [43]). Clearly, this result casts doubts on the reliability of the 3-D model used.

Future Directions

Seismic tomography began about thirty years ago [2,4,33] and since then the field has grown steadily. However, because of the intrinsic difficulties of wave propagation in the earth and its computer simulation, the ill-posedness of the inverse problem that must be solved (particularly at the global scale), and very limited funding, progress has been slow. Fortunately, in spite of these obstacles the point has been reached where it is possible to establish with some certainty which features of the internal composition of the earth are well resolved and which ones are not, and what needs to be done to improve the existing knowledge (see, e. g. [16,41,129,135,150]). It is clear that progress will come from several fronts and will be fueled by the steady decline in the price of computers and their increased power as well as the availability of relatively inexpensive PC clusters. This development will allow the generation of more realistic synthetic seismograms (see, e. g. [81]) and the computation of more accurate inverse solutions as well as the computation of resolution and covariance matrices. The integration of tomographic, geodynamic, and mineral physics models will lead, hopefully, to a better understanding of the earth's interior and the processes therein. However, because oceans cover two-thirds of the earth-surface, it will be necessary to make progress in the deployment of ocean-bottom seismographs before these goals can be achieved fully.

Bibliography

1. Aki K (1993) Overview. In: Iyer H, Hirahara K (eds) Seismic tomography. Chapman, London, pp 1–8
2. Aki K, Lee W (1976) Determination of three-dimensional velocity anomalies under a seismic array using first P arrival times from local earthquakes, 1. A homogeneous initial model. J Geophys Res 81:4381–4399
3. Aki K, Richards P (1980) Quantitative seismology, vol 2. Freeman, San Francisco
4. Aki K, Christoffersson A, Husebye E (1977) Determination of the three-dimensional seismic structure of the lithosphere. J Geophys Res 82:277–296
5. Allen M, Isaacson E (1998) Numerical analysis for applied science. Wiley, New York

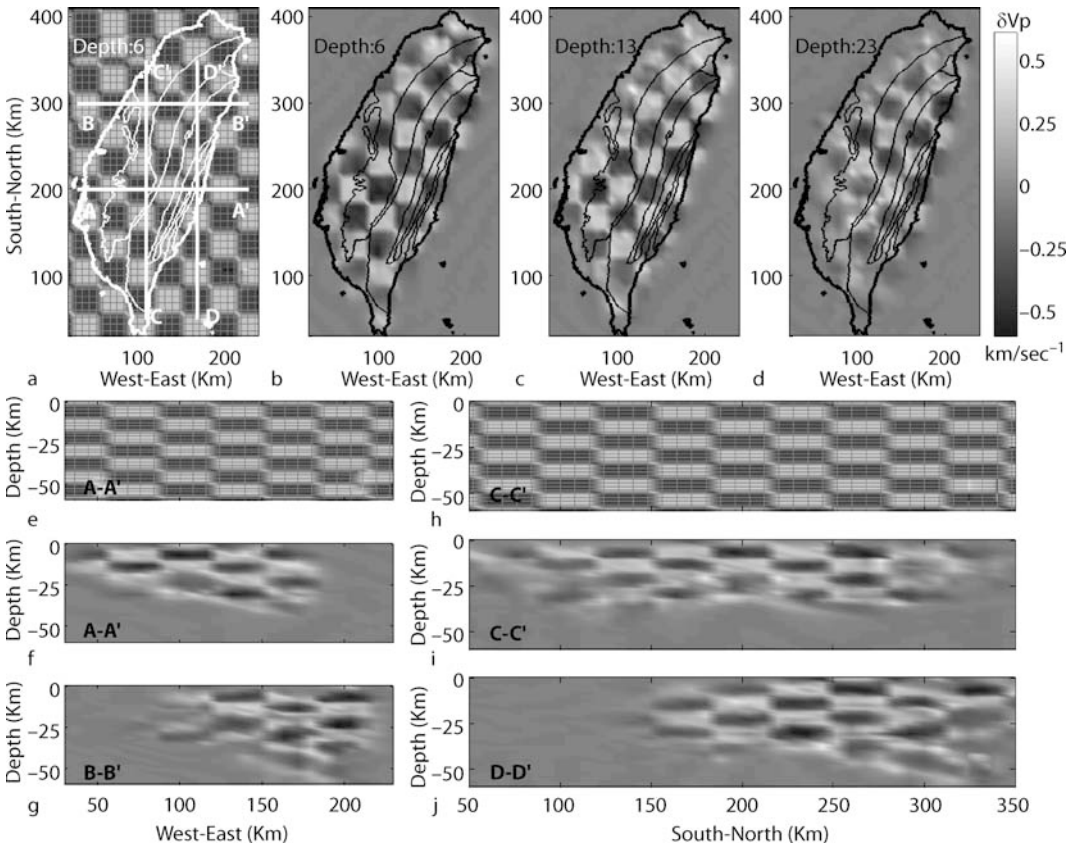

Tomography, Seismic, Figure 12

Results of the checkerboard test for the Taiwan *P* wave velocity model. The checkerboard pattern is shown in a, e and h. The velocity variations across the block boundaries is 0.6 km/s. Synthetic arrival times were generated with this model and the earthquake and station locations used in the inversion of the actual data. The results of the inversion of the synthetic data are shown in map view for different depths in b–d. Cross sectional views are shown in f, g, i, and j. From [78]

6. Ambrose J (1973) Computerized transverse axial scanning (tomography): Part 2. Clin Appl Br J Radiol 46:1023–1047
7. Bai C-Y, Greenhalgh S (2005) 3-D multi-step travel time tomography: imaging the local, deep velocity structure of Rabaul volcano, Papua New Guinea. Phys Earth Planet Inter 151:259–275
8. Bai C-Y, Greenhalgh S (2006) 3D local earthquake hypocenter determination with an irregular shortest-path method. Bull Seism Soc Am 96:2257–2268
9. Bai C-Y, Greenhalgh S, Zhou B (2007) 3D ray tracing using a modified shortest-path method. Geophysics 72(4):T27–T36
10. Bard Y (1974) Nonlinear parameter estimation. Academic Press, New York
11. Barret H, Hawkins W, Joy M (1983) Historical note on computed tomography. Radiology 147:172
12. Bates R, Peters T (1971) Towards improvements in tomography. NZ J Sci 14:883–896
13. Beck J, Arnold K (1977) Parameter estimation in engineering and science. Wiley, New York
14. Benz H, Smith R (1984) Simultaneous inversion for lateral velocity variations and hypocenters in the Yellowstone region using earthquake and refraction data. J Geophys Res 89:1208–1220
15. Benz H, Chouet B, Dawson P, Lahr J, Page R, Hole J (1996) Three-dimensional *P* and *S* wave velocity structure of Redoubt Volcano, Alaska. J Geophys Res 101:8111–8128
16. Boschi L, Ampuero J-P, Peter D, Mai P, Soldati G, Giardini D (2007) Petascale computing and resolution in global seismic tomography. Phys Earth Planet Inter 163:245–250
17. Bracewell R (1956) Strip integration in radio astronomy. Aust J Phys 9:198–217
18. Bracewell R, Riddle A (1967) Inversion of fan-beam scans in radio astronomy. J Astrophys 150:427–434
19. Broad W (1980) Riddle of the Nobel debate. Science 207: 37–38
20. Brooks R, Di Chiro G (1976) Principles of computer assisted tomography (CAT) in radiographic and radioisotopic imaging. Phys Med Biol 5:689–732
21. Červený V (2001) Seismic ray theory. Cambridge University Press, Cambridge
22. Červený V, Molotkov I, Pšenčik I (1977) Ray method in seismology. Charles University, Prague

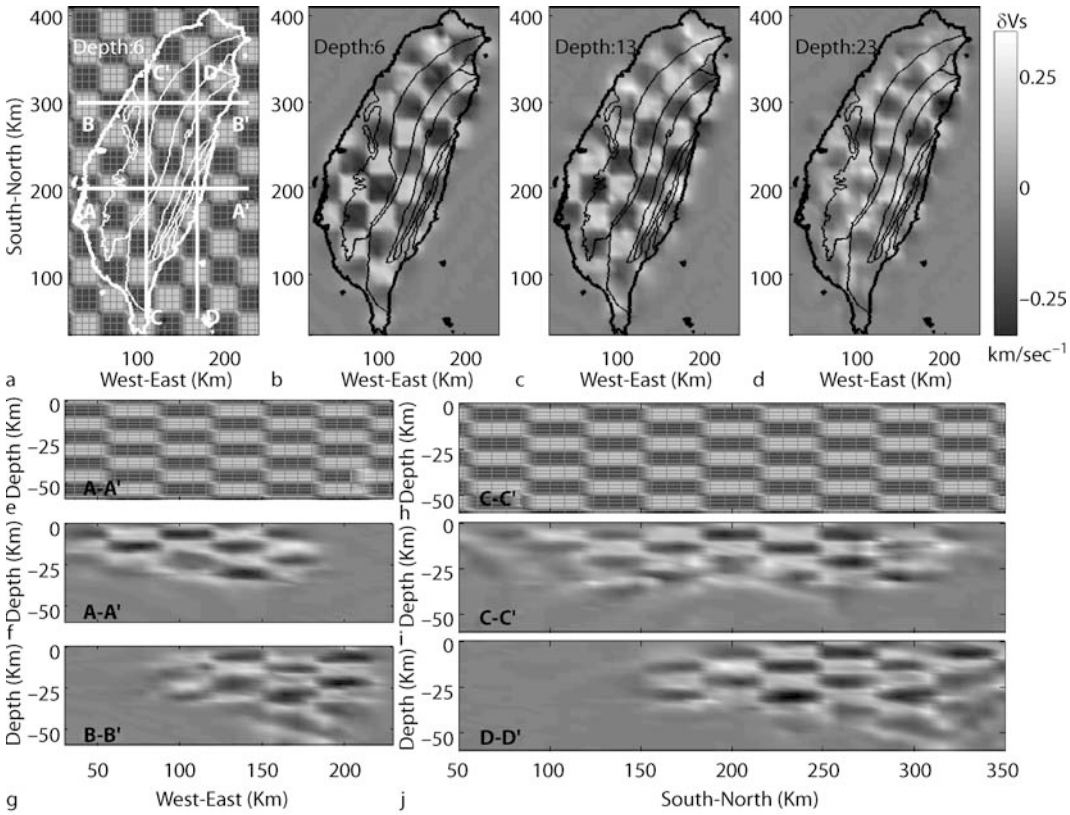

Tomography, Seismic, Figure 13

Similar to Fig. 12 for the *S* wave velocity model. The velocity variations across box boundaries is 0.346 km/s. From [78]

23. Chapman C (1987) The Radon transform and seismic tomography. In: Nolet G (ed) Seismic tomography. Reidel, Dordrecht, pp 25–47
24. Chen H, Chiu J-M, Pujol J, Kim K, Chen K-C, Huang B-S, Yeh Y-H, Chiu S-C (2006) A simple algorithm for local earthquake location using 3D V_P and V_S models: test examples in the central United States and in central eastern Taiwan. Bull Seis Soc Am 96:288–305
25. Claerbout J (1985) Imaging the Earth's Interior. Blackwell Scientific Publications, Boston
26. Cormack A (1963) Representation of a function by its line integrals, with some radiological applications. J Appl Phys 34:2722–2727
27. Cormack A (1964) Representation of a function by its line integrals, with some radiological applications, II. J Appl Phys 35:2908–2913
28. Cormack A (1973) Reconstruction of densities from their projections, with applications in radiological physics. Phys Med Biol 18:195–207
29. Cormack A (1980) Recollections of my work with computer assisted tomography. Mol Cell Biochem 32:57–61
30. Cormack A (1982) Computed tomography: some history and recent developments. Proc Symp Appl Math 27:35–42
31. Creager K (1984) Geometry, velocity structure, and penetration depths of descending slabs in the western Pacific. Ph D dissertation, University of California, San Diego
32. Creager K, Boyd T (1992) Effects of earthquake mislocation on estimates of velocity structure. Phys Earth Planet Inter 75: 63–76
33. Crosson R (1976) Crustal structure modeling of earthquake data. 1. Simultaneous least squares estimation of hypocenter and velocity parameters. J Geophys Res 81:3036–3046
34. Crowther R, DeRosier D, Klug A (1970) The reconstruction of a three-dimensional structure from projections and its application to electron microscopy. Proc R Soc Lond Ser A 317:319–340
35. Deans S (1983) The Radon transform and some of its applications. Wiley, New York
36. DeRosier D, Klug A (1968) Reconstruction of three dimensional structures from electron micrographs. Nature 217: 130–134
37. Dietz L, Ellsworth W (1990) The October 17, 1989 Loma Prieta, California, earthquake and its aftershocks: Geometry of the sequence from high-resolution locations. Geophys Res Lett 17:1417–1420
38. Dietz L, Ellsworth W (1997) Aftershocks of the Loma Prieta earthquake and their tectonic implications. In: P Reasenberg (ed) The Loma Prieta, California, earthquake of October 17, 1989 – Aftershocks and postseismic effects. US Geol Surv Prof Pap 1550-D, D5-D47
39. Dines K, Lytle R (1979) Computerized geophysical tomography. Proc Inst Electr Electron Eng 67:1065–1073

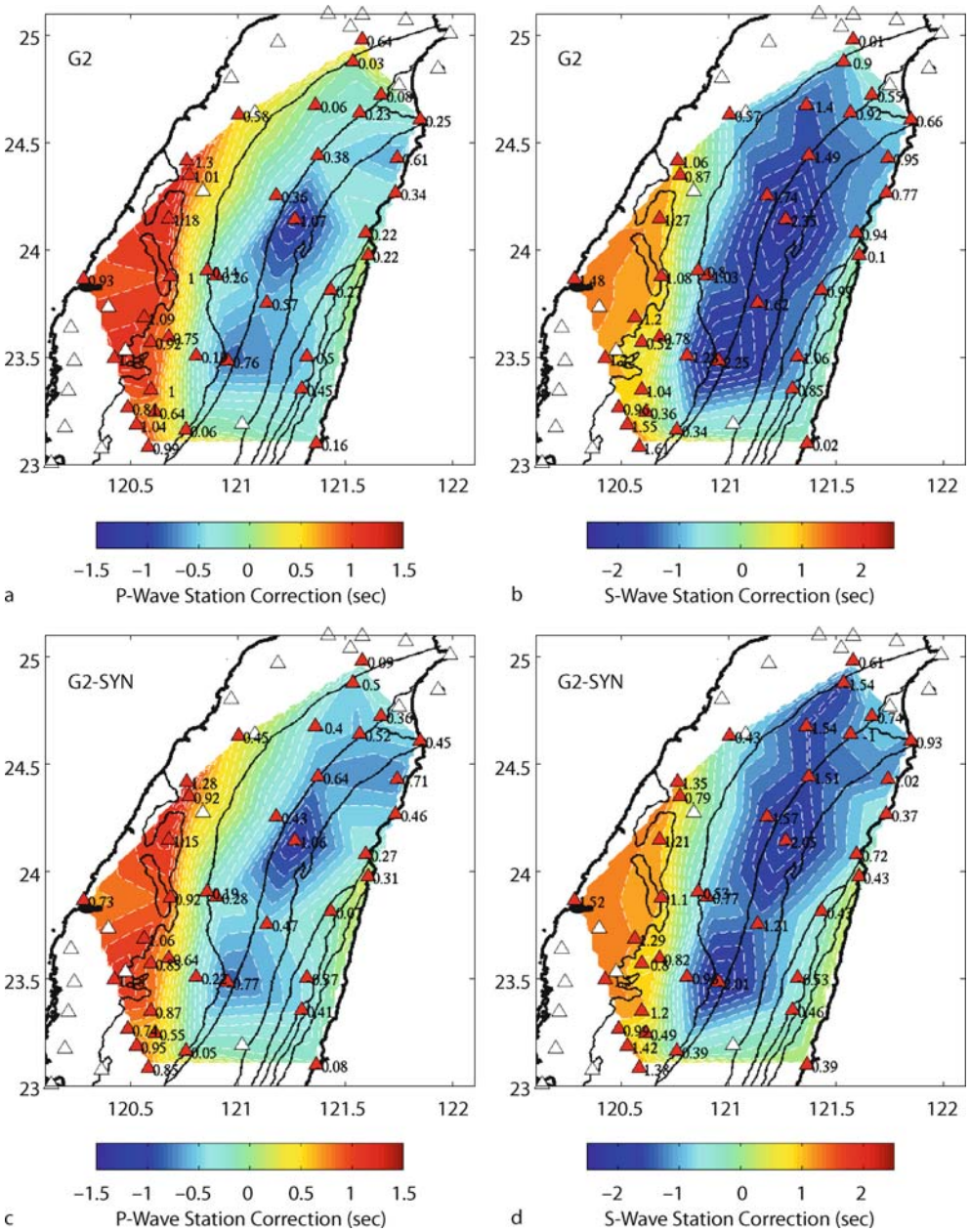

Tomography, Seismic, Figure 14

Top: ***P*** **and** ***S*** **wave station corrections determined using the joint hypocentral determination technique (JHD) and the data in box G2 in Fig. 10.** ***Bottom:*** **JHD station corrections obtained using synthetic data generated with the 3-D velocity model and locations determined by tomographic inversion. From [78]**

40. Durrani T, Bisset D (1984) The Radon transform and its properties. Geophysics 49:1180–1187; Errata, 1985, 50:884–886
41. Dziewonski A (2003) Global seismic tomography: What we really can say and what we make up. Geol Soc Am Penrose Conference, Plume IV: Beyond the Plume Hypothesis, Abstracts (available at: www.mantleplumes.org/Penrose/PenPDFAbstracts/Dziewonski_Adam_abs.pdf)
42. Eberhart-Phillips D (1986) Three-dimensional velocity structure in northern California Coast Ranges from inversion of local earthquake arrival times. Bull Seism Soc Am 76: 1025–1052
43. Eberhart-Phillips D, Stuart W (1992) Material heterogeneity simplifies the picture: Loma Prieta. Bull Seism Soc Am 82:1964–1968

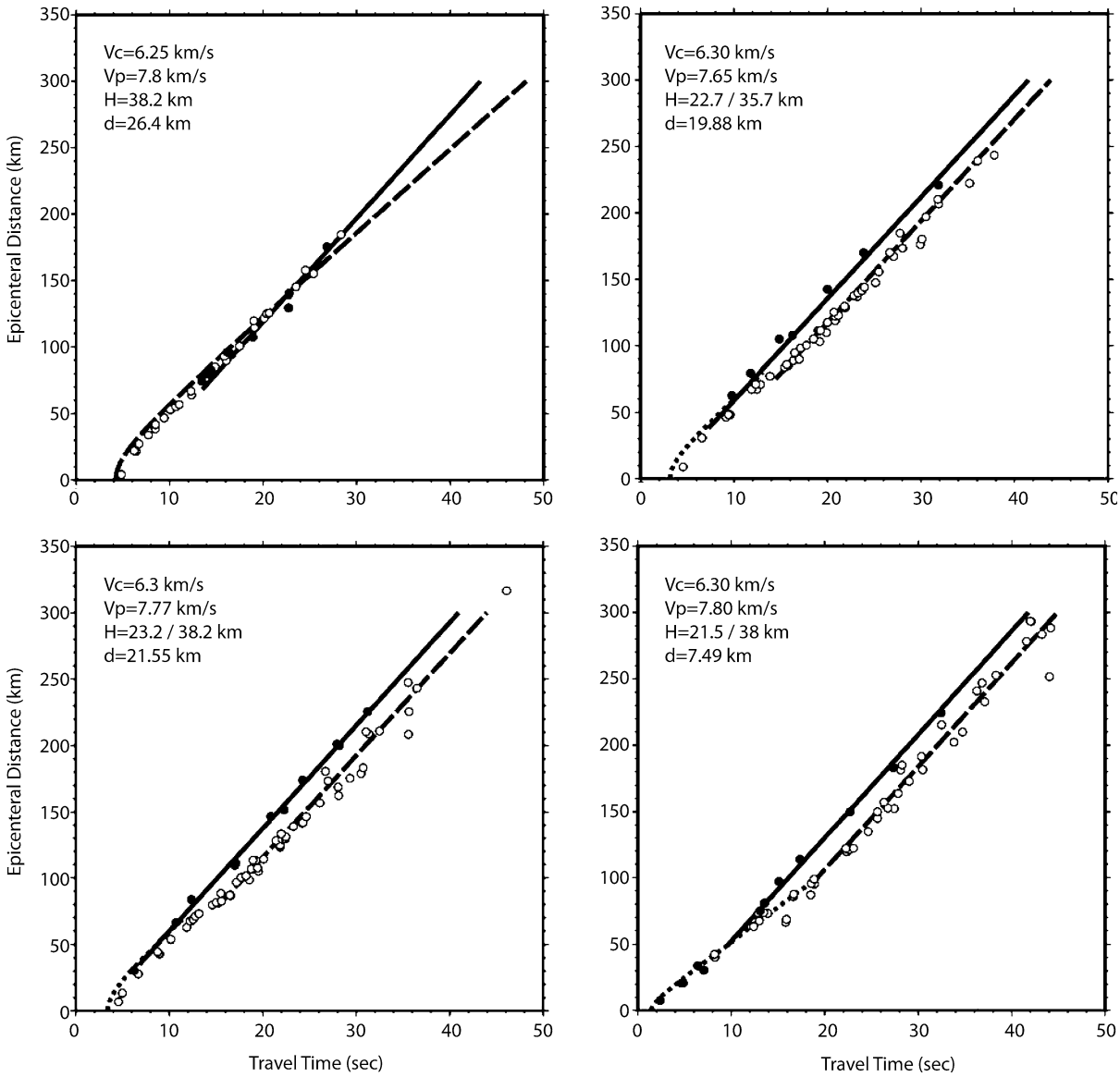

Tomography, Seismic, Figure 15

Epicentral distance versus travel time plots for the four earthquakes in Fig. 9. The plot in the *upper left corner* corresponds to the event numbered 1. The *solid* and *dashed lines* are best-fit lines corresponding to the anomalous and normal *Pn* waves. For events 2–4 (along Taiwan's east coast) the anomalous *Pn* waves are seen in the eastern stations. Vc: crustal velocity; Vp: mantle velocity; H: depth(s) to the Moho; d: event depth. From [89]

44. Eliseevnin V (1965) Analysis of rays propagating in an inhomogeneous medium. Sov Phys Acoust 10:242–245
45. Evans J, Achauer U (1993) Teleseismic velocity tomography using the ACH method: theory and application to continental-scale studies. In: Iyer H, Hirahara K (eds) Seismic tomography. Chapman, London, pp 319–360
46. Forsythe G, Malcolm M, Moler C (1977) Computer methods for mathematical computations. Prentice-Hall, Englewood Cliffs
47. Franklin J (1970) Well-posed extensions of ill-posed linear problems. J Math Anal Appl 31:682–716
48. Gilbert P (1972) Iterative methods for the three-dimensional reconstruction of an object from projections. J Theor Biol 36:105–117
49. Gill P, Murray W, Wright M (1981) Practical optimization. Academic Press, London
50. Gordon R (1974) A tutorial on ART. Inst Electr Electron Eng Trans Nucl Sci NS-21:78–93

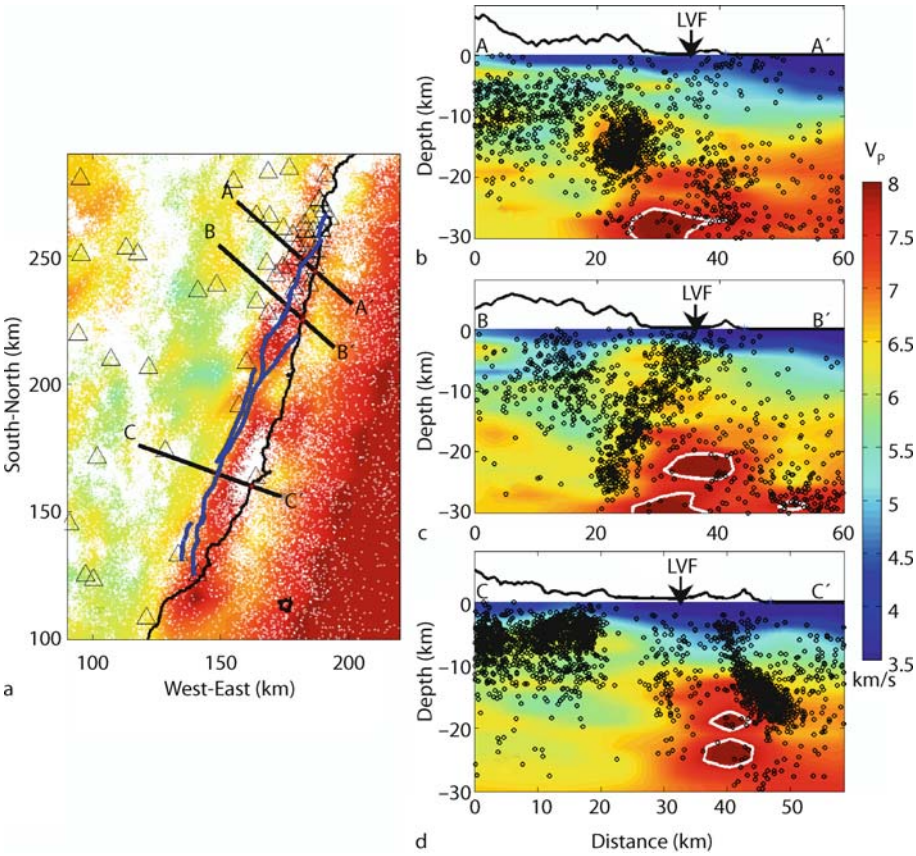

Tomography, Seismic, Figure 16

Left: **Map view of the Taiwan *P* wave velocity model at a depth of 20 km (from [78]). The coordinate system is as in Fig. 11a. *Blue lines*: surface trace of active faults. *White dots*: epicentral locations. *Right:* Cross-sectional views of *P* wave velocity across the Longitudinal Valley (LVF) and seismicity. The locations of the cross sections are shown on the *left*. Note the elevated high-velocity oceanic upper mantle along the entire collision suture zone. The 7.8 km/s contour is indicated by the *white line* around the *dark-red areas*. From [89]**

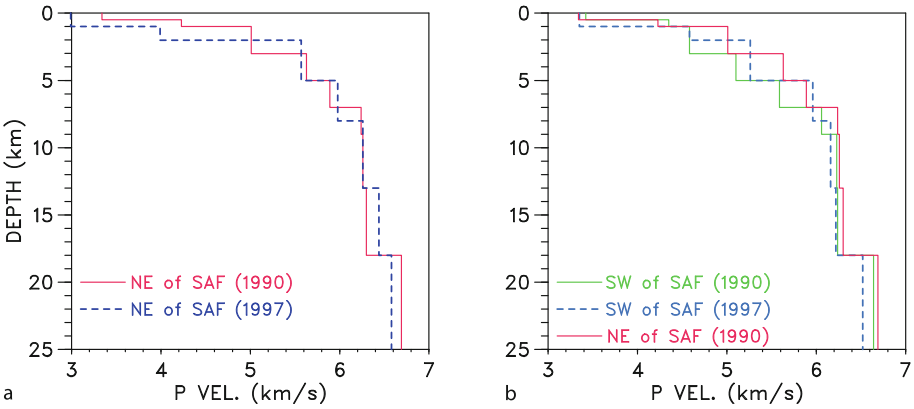

Tomography, Seismic, Figure 17

P **wave velocity models used in [37] and [38] to locate the 1989, *M* = 7.1, Loma Prieta, California, earthquake and its aftershocks. *Left:* Models for stations to the NE of the San Andreas fault. *Right:* Models for stations to the SW of the San Andreas fault. For a comparison, the NE 1990 model is also shown. After [119]**

51. Gordon R, Bender R, Herman G (1970) Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography. J Theor Biol 29:471–481
52. Grand S (1987) Tomographic inversion for shear velocity beneath the North American plate. J Geophys Res 92:14065–14090
53. Groetsch C (1993) Inverse problems in the mathematical sciences. Vieweg, Braunschweig
54. Gubbins D (1981) Source location in laterally varying media. In: Husebye E, Mykkeltveit S (eds) Identification of seismic sources – Earthquake or underground explosion. Reidel, Dordrecht, pp 543–573
55. Hansen P (1992) Analysis of discrete ill-posed problems by means of the L-curve. SIAM Rev 34:561–580
56. Hansen P (1994) Regularization Tools: A Matlab package for analysis and solution of discrete ill-posed problems. Num Algorithms 6:1–35; (Software available at: http://www2.imm.dtu.dk/~pch/Regutools/)
57. Hanson K (1987) Bayesian and related methods in image reconstruction from incomplete data. In: Stark H (ed) Image recovery: theory and applications. Academic, Orlando, pp 79–125
58. Hauksson E (2000) Crustal structure and seismicity distribution adjacent to the Pacific and North America plate boundary in southern California. J Geophys Res 105:13875–13903
59. Hauksson E, Haase J (1997) Three-dimensional V_P and V_P/V_S velocity models of the Los Angeles basin and central Transverse Ranges, California. J Geophys Res 102:5423–5453
60. Hauksson E, Jones L, Hutton K (1995) The 1994 Northridge earthquake sequence in California: seismological and tectonic aspects. J Geophys Res 100:12335–12355
61. Hawley B, Zandt G, Smith R (1981) Simultaneous inversion for hypocenters and lateral velocity variations: an iterative solution with a layered model. J Geophys Res 86:7073–7086
62. Herman G (1980) Image reconstruction from projections. Academic Press, New York
63. Herman G, Lent A (1976) Iterative reconstruction algorithms. Comput Biol Med 6:273–294
64. Herman G, Hurwitz H, Lent A, Lung H-P (1979) On the Bayesian approach to image reconstruction. Inform Contr 42:60–71
65. Hounsfield G (1973) Computerized transverse axial scanning (tomography): Part 1. Description of system. J Br Radiol 46:1016–1022
66. Hounsfield G (1979) Computed medical imaging, Nobel Lecture. (available at: nobelprize.org/nobel_prizes/medicine/laureates/1979/hounsfield-lecture.pdf)
67. Hounsfield G (1980) Autobiography. In: Wilhelm O (ed) The Nobel Prizes 1979. The Nobel Foundation, Stockholm (available at: nobelprize.org/nobel_prizes/medicine/laureates/1979/hounsfield-autobio.html)
68. Hudnut K et al (1996) Co-seismic displacements of the 1994 Northridge, California, earthquake. Bull Seism Soc Am 86(1B):S19–S36
69. Hudson J (1980) The excitation and propagation of elastic waves. Cambridge University Press, Cambridge
70. Humphreys E, Clayton R (1988) Adaptation of back projection tomography to seismic travel times problems. J Geophys Res 93:1073–1085
71. Hurwitz H (1975) Entropy reduction in Bayesian analysis of measurements. Phys Rev A 12:698–706
72. Inoue H, Fukao Y, Tanabe K, Ogata Y (1990) Whole mantle P-wave travel time tomography. Phys Earth Planet Inter 59:294–328
73. Ivansson S (1983) Remark on an earlier proposed iterative tomographic algorithm. J Geophys R Astr Soc 75:855–860
74. Jackson D (1979) The use of a priori data to resolve nonuniqueness in linear inversion. J Geophys R Astr Soc 57:137–157
75. Julian B, Gubbins D (1977) Three-dimensional seismic ray tracing. J Geophys 43:95–113
76. Kak A, Slaney M (1988) Principles of computerized tomographic imaging. Inst Electr Electron Eng Press, New York
77. Kim K-H (2003) Subsurface structure, seismicity patterns, and their implication to tectonic evolution in Taiwan. Ph D dissertation, University of Memphis, Memphis
78. Kim K-H, Chiu J-M, Pujol J, Chen K-C, Huang B-S, Yeh Y-H, Shen P (2005) Three-dimensional V_P and V_S structural models associated with the active subduction and collision tectonics in the Taiwan region. J Geophys Int 162:204–220
79. Koch M (1985) Non-linear inversion of local seismic travel times for the simultaneous determination of 3D-velocity structure and hypocenters – application to the seismic zone Vrancea. J Geophys 56:160–173
80. Koch M (1993) Simultaneous inversion for 3-D crustal structure and hypocenters including direct, refracted and reflected phases – I Development, validation and optimal regularization of the method. J Geophys Int 112:385–412
81. Komatitsch D, Tsuboi S, Tromp J (2005) The spectral-element method in seismology. In: Levander A, Nolet G (eds) Seismic earth: array analysis of broadband seismograms. Geophysical Monograph Series, vol 157. Am Geophys Union, Washington DC, pp 205–227
82. Langenheim V, Griscom A, Jachens R, Hildenbrand T (2000) Preliminary potential-field constraints on the geometry of the San Fernando basin, southern California. US Geol Survey Open-File Report 00–219
83. Lawson C, Hanson R (1974) Solving least squares problems. Prentice-Hall, Englewood Cliffs
84. Lee W, Pereyra V (1993) Mathematical introduction to seismic tomography. In: Iyer H, Hirahara K (eds) Seismic tomography. Chapman, London, pp 9–22
85. Lee W, Stewart S (1981) Principles and applications of microearthquake networks. Academic Press, New York
86. Lees J, Crosson R (1989) Tomographic inversion for three-dimensional velocity structure at Mount St. Helens using earthquake data. J Geophys Res 94:5716–5728
87. Levenberg K (1944) A method for the solution of certain nonlinear problems in least squares. Quart Appl Math 2:164–168
88. Lewitt R (1983) Reconstruction algorithms: transform methods. Proc Inst Electr Electron Eng 71:390–408
89. Liang W-T, Chiu J-M, Kim K (2007) Anomalous Pn waves observed in eastern Taiwan: implications of a thin crust and elevated oceanic upper mantle beneath the active collision-zone suture. Bull Seism Soc Am 97:1370–1377
90. Ma K-F, Wang J-H, Zhao D (1996) Three-dimensional seismic velocity structure of the crust and uppermost mantle beneath Taiwan. J Phys Earth 44:85–105
91. Magistrale H, Day S, Clayton R, Graves R (2000) The SCEC southern California reference three-dimensional seismic velocity model version 2. Bull Seism Soc Am 90(6B):S65–S76

92. Marquardt D (1963) An algorithm for least-squares estimation of nonlinear parameters. J Soc Ind Appl Math 11:431–441
93. Martin M, Ritter J, CALIXTO Working Group (2005) High-resolution teleseismic body-wave tomography beneath SE Romania – I. Implications for three-dimensional versus one-dimensional crustal correction strategies with a new crustal velocity model. J Geophys Int 162:448–460
94. Meskò A (1984) Digital filtering: applications in geophysical exploration for oil. Wiley, New York
95. Monna S, Filippi L, Beranzoli L, Favali P (2003) Rock properties of the upper-crust in Central Apennines (Italy) derived from high-resolution 3-D tomography. Geophys Res Lett 30(61): 1–4 doi:10.1029/2002GL016780
96. Morelli A (1993) Teleseismic tomography: core-mantle boundary. In: Iyer H, Hirahara K (eds) Seismic tomography. Chapman, London, pp 163–189
97. Mori J, Wald D, Wesson R (1995) Overlapping fault planes of the (1971) San Fernando and 1994 Northridge, California earthquakes. Geophys Res Lett 22:1033–1036
98. Moser T (1991) Shortest path calculation of seismic rays. Geophysics 56:59–67
99. Moser T, Nolet G, Snieder R (1992) Ray bending revisited. Bull Seismol Soc Am 82:259–288
100. Moser T, Van Eck T, Nolet G (1992) Hypocenter determination in strongly heterogeneous earth models using the shortest path method. J Geophys Res 97:6563–6572
101. Nakanishi I, Yamaguchi K (1986) A numerical experiment on nonlinear image reconstruction from first-arrival times for two-dimensional island arc structure. J Phys Earth 34:195–201
102. Nelson G, Vidale J (1990) Earthquake locations by 3-D finite-difference travel times. Bull Seism Soc Am 80:395–410
103. Noble B, Daniel J (1977) Applied linear algebra. Prentice-Hall, Englewood Cliffs
104. Nolet G (1993) Solving large linearized tomographic problems. In: Iyer H, Hirahara K (eds) Seismic tomography. Chapman, London, pp 227–247
105. Okubo P, Benz H, Chouet B (1997) Imaging the crustal magma sources beneath Mauna Loa and Kilauea Volcanoes, Hawaii. Geology 25:867–870
106. Oldendorf W (1961) Isolated flying spot detection of radio density discontinuities – Displaying the internal structural pattern of a complex object. IRE Trans Biomed Elec BME-8: 68–72
107. Oransky I (2004) Obituary. Sir Godfrey N Hounsfield. Lancet 364:1032
108. Paige C, Saunders M (1982) LSQR: An algorithm for sparse linear equations and sparse least square problems. ACM Trans Math Softw 8:43–71
109. Parker R (1994) Geophysical inverse theory. Princeton University Press, Princeton
110. Pavlis G, Booker J (1980) The mixed discrete-continuous inverse problem: application to the simultaneous determination of earthquake hypocenters and velocity structure. J Geophys Res 85:4801–4810
111. Penrose R (1955) A generalized inverse for matrices. Proc Camb Phil Soc 51:406–413
112. Pereyra V, Lee W, Keller H (1980) Solving two-point seismic-ray tracing problems in a heterogeneous medium. Bull Seism Soc Am 70:79–99
113. Podvin P, Lecomte I (1991) Finite difference computation of traveltimes in very contrasted velocity models: a massively parallel approach and its associated tools. J Geophys Int 105:271–284
114. Press W, Teukolsky S, Vetterling W, Flannery B (1992) Numerical Recipes. Cambridge University Press, Cambridge
115. Prothero W, Taylor W, Eickemeyer J (1988) A fast, two-point, three-dimensional raytracing algorithm using a simple step search method. Bull Seism Soc Am 78:1190–1198
116. Pujol J (1995) Application of the JHD technique to the Loma Prieta, California, mainshock-aftershock sequence and implications for earthquake location. Bull Seism Soc Am 85: 129–150
117. Pujol J (1996) Comment on: "The 1989 Loma Prieta earthquake imaged from inversion of geodetic data" by Thora Árnadóttir and Paul Segall. J Geophys Res 101:20133–20136
118. Pujol J (1996) An integrated 3D velocity inversion – joint hypocentral determination relocation analysis of events in the Northridge area. Bull Seism Soc Am 86(1B):S138–S155
119. Pujol J (2000) Joint event location – The JHD technique and applications to data from local seismic networks. In: Thurber C, Rabinowitz N (eds) Advances in seismic event location. Kluwer, Dordrecht, pp 163–204
120. Pujol J (2003) Elastic wave propagation and generation in seismology. Cambridge University Press, Cambridge
121. Pujol J (2007) The solution of nonlinear inverse problems and the Levenberg-Marquardt method. Geophysics 72(4): W1–W16
122. Pujol J et al (1989) 3-D P- and S-wave velocity structure of the Andean foreland in San Juan, Argentina, from local earthquakes. Eos Trans Am Geoph Union 70(43):1213
123. Pujol J, Mueller K, Peng S, Chitupolu V (2006) High-resolution 3D *P*-wave velocity model for the East Ventura–San Fernando basin, California, and relocation of events in the Northridge and San Fernando aftershock sequences. Bull Seism Soc Am 96:2269–2280
124. Ramachandran G, Lakshminarayanan A (1971) Three-dimensional reconstruction from radiographs and electron micrographs: application of convolutions instead of Fourier transforms. Proc Natl Acad Sci USA 68:2236–2240
125. Ratchkovsky N, Pujol J, Biswas N (1997) Relocation of earthquakes in the Cook Inlet area, south central Alaska, using the joint hypocenter determination method. Bull Seism Soc Am 87:620–636
126. Rau R-J, Wu F (1995) Tomographic imaging of lithospheric structures under Taiwan. Earth Planet Lett 133:517–532
127. Robinson E (1982) Spectral approach to geophysical inversion by Lorentz, Fourier, and Radon transforms. Proc Inst Electr Electron Eng 70:1039–1054
128. Roecker S, Yeh Y, Tsai Y (1987) Three-dimensional P and S wave velocity structures beneath Taiwan: deep structure beneath an arc-continent collision. J Geophys Res 92:10547–10570
129. Romanowicz B (2003) Global mantle tomography: Progress status in the past 10 years. Annu Rev Earth Planet Sci 31: 303–328
130. Sage A, Melsa J (1971) Estimation theory with applications to communications and control. McGraw-Hill, New York
131. Sandoval S, Kissling E, Ansorge J, Svekalapko Seismic Tomography Working Group (2003) High-resolution body wave tomography beneath the Svekalapko array: I, A priori three-dimensional crustal model and associated traveltime effects on teleseismic wave fronts. J Geophys Int 153:75–87

132. Shepp L, Kruskal J (1978) Computerized tomography: the new medical X-ray technology. Am Math Mon 85:420–439
133. Shepp L, Logan B (1974) The Fourier reconstruction of a head section. Inst Electr Electron Eng Trans Nucl Sci NS-21:21–43
134. Snoke J, Lahr J (2001) Locating earthquakes: at what distances can the Earth no longer be treated as flat? Seism Res Lett 72:538–541
135. Soldati G, Boschi L, Piersanti A (2006) Global seismic tomography and modern parallel computers. Ann Geophys 49: 977–986
136. Sorenson H (1980) Parameter estimation: principles and problems. Dekker, New York
137. Spakman W (1993) Iterative strategies for non-linear travel time tomography using global earthquake data. In: Iyer H, Hirahara K (eds) Seismic tomography. Chapman, London, pp 190–226
138. Stanton L (1969) Basic medical radiation physics. Appleton-Century-Crofts, New York
139. Stewart J, Choi Y, Graves R, Shaw J (2005) Uncertainty of southern California basin depth parameters. Bull Seism Soc Am 95:1988–1993
140. Süss M, Shaw J (2003) *P* wave seismic velocity structure derived from sonic logs and industry reflection data in the Los Angeles basin, California. J Geophys Res 108(13):1–18 doi:10.1029/2001JB001628
141. Tarantola A, Valette B (1982) Inverse problems = quest for information. J Geophys 50:159–170
142. Tarantola A, Valette B (1982) Generalized nonlinear inverse problems solved using the least squares criterion. Rev Geophys Space Phys 20:219–232
143. Teng T-L, Aki K (1996) Preface to the 1994 Northridge earthquake special issue. Bull Seism Soc Am 86(1B):S1–S2
144. Thurber C (1983) Earthquake locations and three-dimensional crustal structure in the Coyote Lake area, central California. J Geophys Res 88:8226–8236
145. Thurber C (1992) Hypocenter-velocity structure coupling in local earthquake tomography. Phys Earth Planet Inter 75: 55–62
146. Thurber C (1993) Local earthquake tomography: velocities and V_P/V_S – theory. In: Iyer H, Hirahara K (eds) Seismic tomography. Chapman, London, pp 563–583
147. Thurber C, Kissling E (2000) Advances in travel-time calculations for three-dimensional structures. In: Thurber C, Rabinowitz N (eds) Advances in seismic event location. Kluwer, Dordrecht, pp 71–99
148. Tihonov A (1963) Solution of incorrectly formulated problems and the regularization method. Sov Math 4:1035–1038; (Note: a more common transliteration of this author's Russian name is Tikhonov.)
149. Titchmarsh W (1948) Introduction to the theory of Fourier integrals. Oxford University Press, Oxford
150. Trampert J, Van der Hilst R (2005) Towards a quantitative interpretation of global seismic tomography. In: Van der Hilst R, Bass J, Matas J, Trampert J (eds) Earth's deep mantle: structure, composition, and evolution. Geophysical Monograph Series, vol 160. Am Geophys Union, Washington DC, pp 47–62
151. Tryggvason A, Bergman B (2006) A traveltime reciprocity discrepancy in the Podvin & Lecomte time3d finite difference algorithm. J Geophys Int 165:432–435
152. Um J, Thurber C (1987) A fast algorithm for two-point seismic ray tracing. Bull Seism Soc Am 77:972–986
153. Van der Hilst R, Engdahl E (1992) Step-wise relocation of ISC earthquake hypocenters for linearized tomographic imaging of slab structure. Phys Earth Planet Inter 75:39–53
154. Van der Sluis A, Van der Vorst H (1987) Numerical solution of large, sparse linear algebraic systems arising from tomographic problems. In: Nolet G (ed) Seismic tomography. D Reidel, Dordrecht, pp 49–83
155. Van Tiggelen R (2002) In search for the third dimension: from radiostereoscopy to three-dimensional imaging. JBR-BTR 85:266–270
156. Van Tiggelen R, Pouders E (2003) Ultrasound and computed tomography: spin-offs of the World Wars. JBR-BTR 86: 235–241
157. Vidale J (1988) Finite-difference calculation of travel times. Bull Seism Soc Am 78:2062–2076
158. Vidale J (1990) Finite-difference calculation of travel times in three dimensions. Geophysics 55:521–526
159. Villaseñor A, Benz H, Filippi L, De Luca G, Scarpa R, Patanè G, Vinciguerra S (1998) Three-dimensional *P*-wave velocity structure of Mt. Etna, Italy. Geophys Res Lett 25: 1975–1978
160. Wald D, Graves R (1998) The seismic response of the Los Angeles basin, California. Bull Seism Soc Am 88:337–356
161. Woodhouse J, Dziewonski A (1989) Seismic modelling of the Earth's large-scale three-dimensional structure. Phil Trans Roy Soc Lond A 328:291–308
162. Zhao D, Kanamori H (1995) The 1994 Northridge earthquake: 3-D crustal structure in the rupture zone and its relation to the aftershock locations and mechanisms. Geophys Res Lett 22:763–766
163. Zhao D, Hasegawa A, Horiuchi S (1992) Tomographic imaging of *P* and *S* wave velocity structure beneath northeastern Japan. J Geophys Res 97:19909–19928
164. Zhdanov M (2002) Geophysical inverse theory and regularization problems. Elsevier, Amsterdam

Topological Complexity of Molecules

Dušanka Janežič[1], Ante Miličević[2], Sonja Nikolić[3], Nenad Trinajstić[3]

[1] National Institute of Chemistry, Ljubljana, Slovenia

[2] The Institute of Medical Research and Occupational Health, Zagreb, Croatia

[3] The Rugjer Bošković Institute, Zagreb, Croatia

Article Outline

Glossary
Definition of the Subject
Introduction
The Hierarchical Approaches to Molecular Complexity
Criteria for Topological Complexity Indices
Examples of Topological Complexity Indices

The Usefulness of the Reviewed Topological Complexity Indices to Assess the Complexity of Molecular Graphs
Future Directions
Bibliography

Glossary

Complexity The word 'complexity' is made up from the Latin roots – 'com' meaning 'together' and 'plectere' meaning 'to plait'. Complexity is a difficult concept to define. One way to define it is as follows. A system is complex if it consists of a number of components interacting with each other in many different ways, so that these interactions sometimes lead to unexpected collective (emergent) properties. Hence, increasing complexity is associated with an increasing number of components in a system and with increasing versatility of their interactions.

Complexity (descriptor) index A complexity (descriptor) index is a number that is used to assess the complexity of a system.

Graph A graph, sometimes called a non-directed graph and, usually denoted by G, is a mathematical object which consists of two non-empty sets: one set, denoted usually by V, is a set of elements called vertices and the other, usually denoted by E, is a set of unordered pairs of distinct elements of V called edges. Thus, $G = (V, E)$. In directed graphs, E is a set of ordered pairs of elements of V. In multigraphs more than one edge can join two vertices. A graph G is connected if every pair of vertices is joined by a path. If there is no path between two vertices in G, then G is the disconnected graph having two or more components. A graph G is a planar graph if it can be drawn in a plane in such a way that no two edges intersect. A dual G^* of a planar graph G can be constructed in this way: Place a vertex in each region of G, including the exterior region, and if two regions have an edge e in common, join the corresponding vertices by an edge e^* crossing only e. The inner dual is a subgraph of dual which does not contain the vertex corresponding to the exterior region. A graph is complete if each pair of its vertices is adjacent.

Graph invariant An invariant of a graph G is a number associated with G which has the same value for any graph isomorphic to G.

Graph-theoretical distance The length of the shortest path between two vertices in a graph G is the graph-theoretical distance.

Laplacian matrix The Laplacian matrix $\mathbf{L}$ of a graph G is a real symmetric matrix whose diagonal elements are the vertex-degrees of G and off-diagonal elements are -1 if vertices are connected in G, otherwise zero.

Molecular graph A molecular graph is a 2-dimensional representation of a molecule and is generated by replacing atoms and bonds with vertices and edges, respectively. All molecular graphs in this article are connected graphs.

Path A path is a sequence of edges, each edge sharing one vertex with the sequence-adjacent edges and sharing no vertices with any other edge. The length of a path is the number of edges it contains.

Subgraph A subgraph of a graph G is a graph with all of its vertices and edges in G. A spanning subgraph is a subgraph containing all the vertices of G.

Topological complexity Topological complexity is the complexity of graphs and is determined completely by the particular adjacency of a graph's vertices – the issues of metrics and geometry are not relevant.

Tree A tree is an acyclic graph – one which has no cycles. A spanning tree of a graph G is a connected, acyclic subgraph containing all the vertices of G.

Vertex-adjacency matrix The vertex-adjacency matrix $\mathbf{A}$ is a 0-1 matrix representing graph. Entry 1 denotes adjacent vertices, all other entries are zero.

Walk A walk in a graph G is an alternating sequence of vertices and edges of G, such that each edge e begins and ends with the vertices immediately preceding and following e in the sequence. The length of a walk is the number of edges it contains. The number of walks of length l beginning at vertex i and ending at vertex j is given by the i, j-element of the lth power of the vertex-adjacency matrix, whilst the number of the self-returning walks of length l is given by the i, i-element of the lth power of the vertex-adjacency matrix.

Definition of the Subject

The concept of complexity has intrigued people from the beginning of history, but only in our times have attempts to quantify it begun to appear with a new science emerging – the science of complexity. Manifestation of complexity can be found everywhere in nature and life [1] and different levels of complexity are encountered in arts, humanities and sciences [2].

Like many concepts in chemistry, the concept of complexity appears to be a fuzzy, but useful concept [3]. The fuzziness of this concept has not prevented chemists from attempting to quantify it [4].

Here we are concerned with topological complexity of molecules. It should be stated at the outset that there are

different levels of complexity regarding the structure of molecules [5], i. e., elemental or compositional or one-dimensional complexity (which is determined by the partition of a graph's vertices into classes of different type), topological or two-dimensional complexity and dynamical or three-dimensional complexity (a complexity feature of fluctuating molecules).

The topological complexity index will be abbreviated as TCI, and this symbol will be used throughout the article.

Introduction

Initial work on the topological complexity of molecules was based on use of information theory in studies of the complexity of living systems and the most of the pioneering papers appeared in the *Bulletin of Mathematical Biophysics* (this journal changed the title in 1972 to the *Bulletin of Mathematical Biology*). The idea behind this approach was reductionistic: the complexity of living systems is largely determined by the complexity of the constituent organic molecules. More recently, Bonchev and Buck [6] carried out a comparative analysis of the topological structure of molecules and biological networks and pointed out both similarities and differences.

The first attempt to quantify the complexity of biological molecules *via* an index (descriptor) was carried out in 1953 by Dancoff and Quastler [7] within the framework provided by the Shannon information theory [8]. The Dancoff–Quastler complexity index was based on the information of the atomic composition of a molecule. Morowitz [9] in 1955 produced another complexity index that combined the information on the atomic composition with that on type and multiplicity of bonds in the molecule. In the same year, Rashevsky [10] pointed out that there are molecules with the same atomic composition and the same atomic valencies but differ in the way in which atoms are connected, that is, they differ by their topology. Rashevsky called his complexity measure as the topological information content – this was the first structural complexity measure devised for molecules. He was also first to use graphs to compute the information content of organic molecules. Karreman [11] immediately pointed out that in living processes chemical reactions between the molecules rather than the molecules themselves are responsible for the complexity of living systems. Hence, he proposed the topological information content of reactions to be used in assessing the complexity in living systems. Karreman did not use the term graph, but graphs appear in his paper under the term the topology of a molecule. A year later (1956), Trucco [12,13] reformulated the Rashevsky complexity index in terms of the automorphism group of the molecule, since the topology of a molecule is not always sufficient. He used graph-theoretical and group-theoretical arguments in his work. Mowshowitz [14] in 1968 used Rashevsky's and Trucco's ideas to propose structural information content as an index of the relative complexity of graphs. His paper is based on concepts from graph theory, group theory and combinatorics.

The first to study the topological complexity of graphs without using information theory was Minoli [15]. Bonchev and Trinajstić [16], however, again employed information-theoretical arguments to derive the branching index. Though they did not specify that their index is a complexity index, their approach was soon after extended by Bertz [17] who introduced an information theoretic complexity index that became known as the Bertz (complexity) index [18]. These three initial papers and Bertz's two subsequent papers [19,20] stimulated a number of authors to search for a qualified index that will account for the complexity of molecules; many following the path set by Minoli, deriving complexity indices from consideration of graphs representing molecules [4] without depending upon information theory.

In the text that follows, first we present several hierarchical approaches to molecular complexity, then various criteria for TCIs and a selection of TCIs. The last part of the text contains a comparison between the selected TCIs for trees, cycles, isomeric trees on six vertices and planar polycyclic graphs on four vertices, thoughts on the possible development of the subject and a number of references pertinent to the material discussed.

The Hierarchical Approaches to Molecular Complexity

The first hierarchical approach to molecular complexity was proposed by Bertz in 1983 [19]. His approach, consisting of four hierarchical levels of molecular complexity, is presented in Fig. 1.

As seen from this figure, the Bertz approach contains both topological (branching, rings, multiple bonds) and non-topological (molecular size, symmetry, functionality, elemental composition) structural features. However, the Bertz approach has really only three hierarchical levels of molecular complexity since branched structures are not more complex than rings and multiple bonds and cannot be derived from them.

Four years later (1987), Bonchev and Polansky [21] proposed the hierarchical organization of the total complexity of chemical systems consisting of five levels (see

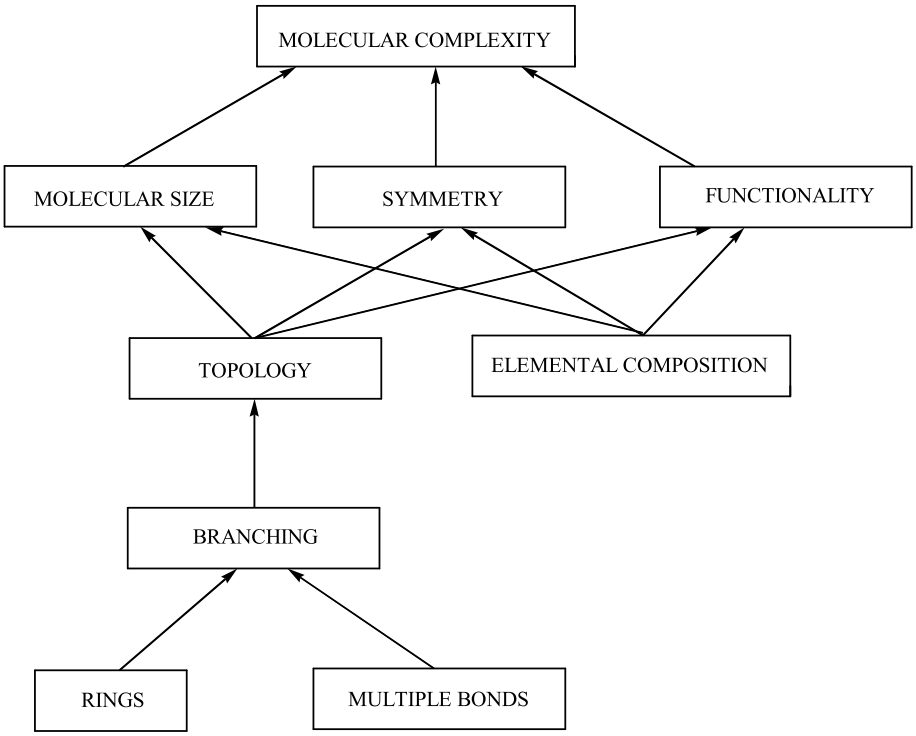

Topological Complexity of Molecules, Figure 1
A hierarchical organization of molecular complexity as proposed by Bertz [19]

Fig. 2) and then the hierarchical levels of topological complexity of chemical systems.

The basic level in this scheme is the size of a chemical system and is followed by its topology. Topology discriminates between systems of the same size. The physical nature of the system is placed on the third level. The physical nature of the system elements and pairwise relations which in chemistry stand for the elemental composition and bond types, discriminate between the systems of the same size and topology. If systems remain in the same complexity class after the first three levels have been considered, the next level to be employed is the specific metric of a system, e. g., the geometric characteristic of a system. The last level is specific symmetry, e. g., the point groups of a system symmetry, which takes care of uniformity or diversity of element distribution. The mathematical model based on the Bonchev–Polansky hierarchical concept of total complexity, denoted by $C_{\text{total complexity}}$, is a multicomponent vector whose components are the hierarchical levels:

$$C_{\text{total complexity}} = (C_{\text{size}}, C_{\text{topology}}, C_{\text{physical nature}}, C_{\text{metric}}, C_{\text{symmetry}}) \,. \quad (1)$$

The Bonchev–Polansky proposal for the topological complexity of chemical systems was slightly condensed in 1997 by Bonchev and Seitz [5] who used graph-theoretical terminology. The difference between the Bonchev–Polansky and the Bonchev–Seitz proposals is exclusion of symmetry from the latter proposal. The Bonchev–Seitz proposal in modified form is shown in Fig. 3.

The hierarchical levels of topological complexity start with the system connectedness which allows differentiation between connected, disconnected and non-planar graphs. The next level, the system adjacency, consists of three classes: directed graphs, non-directed graphs and multigraphs. The connectivity patterns are placed in the third level. Four such patterns were considered: linearity, bridging, branching and cyclicity. In the fourth level metric properties of graphs, e. g., distances, paths, walks are placed in one branch, and the other contains subgraphs of increasing size and number. The mathematical model based on the Bonchev–Seitz hierarchical concept of complexity can be summarized as:

$$C_{\text{topological complexity}} = (C_{\text{connectedness}}, C_{\text{adjacency}}, C_{\text{connectivity pattern}}, C_{\text{metric}}, C_{\text{subgraphs}}) \,. \quad (2)$$

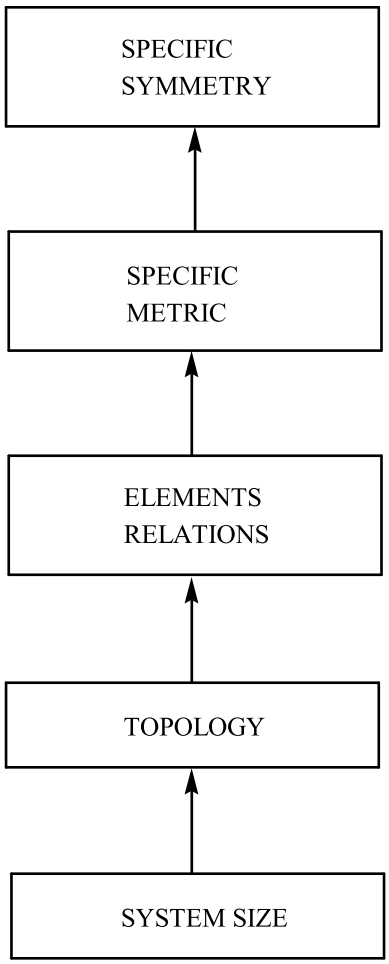

Topological Complexity of Molecules, Figure 2
A hierarchical organization of total complexity of chemical systems as proposed by Bonchev and Polansky [21]

Criteria for Topological Complexity Indices

TCIs arose from the need to compare molecules according to their structural complexity. However, to be useful, TCIs must fulfill certain criteria. Several authors proposed criteria that TCIs should satisfy. A first proposal is due to Minoli [15] in 1975. He proposed the following four criteria for the TCI:

1. TCI should monotonically increase with the number of vertices of the graph.
2. TCI should monotonically increase with the number of edges of the graph.
3. TCI should reflect the degree of connectedness of the graph.
4. TCI should satisfy one's intuition regarding the complexity of objects by assigning a higher value to a graph which looks complicated and vice versa.

Twelve years (1987) after Minoli, Bonchev and Polansky [21] proposed the following nine criteria for the TCI:

1. TCI should be independent of the nature of the systems.
2. TCI should be specified within a unique theoretical concept.
3. TCI should take into account different complexity levels and their hierarchy.
4. TCI should exhibit a stronger dependence on the relations between the elements of a system than on the number of elements.
5. TCI should increase monotonically with the number of different complexity features.
6. TCI should agree with the intuitive ideas of complexity.
7. TCI should differentiate non-isomorphic systems.
8. TCI should not be too sophisticated.
9. TCI should be applicable to practical purposes.

Eleven years later (1998), Bertz and Wright [22] analyzed the Minoli and the Bonchev–Polansky criteria and rejected intuition as a criterion which had been included in both sets of criteria. The following five criteria can be extracted from their work:

1. TCI should increase monotonically with factors that increase complexity as embodied in homologous series.
2. TCI should increase with increasing branching and cyclicity.
3. TCI should increase with the number of multiple edges.
4. TCI should take into account symmetry as a simplifying feature.
5. TCI should be sensitive to heteroatoms.

Intuition should not, however, be treated lightly, since many results achieved in chemistry as well as in other natural sciences were based on guidelines provided by the researcher's intuition.

Two years after Bertz and Wright, Rücker and Rücker [23] discussed the following desirable properties of the TCI:

1. TCI should increase with increasing graph size.
2. TCI should increase with increasing branching.
3. TCI should increase with increasing cyclicity.
4. TCI should increase with increasing number of multiple edges.
5. TCI should increase with increasing number of weighted vertices.
6. TCI should increase with decreasing symmetry.

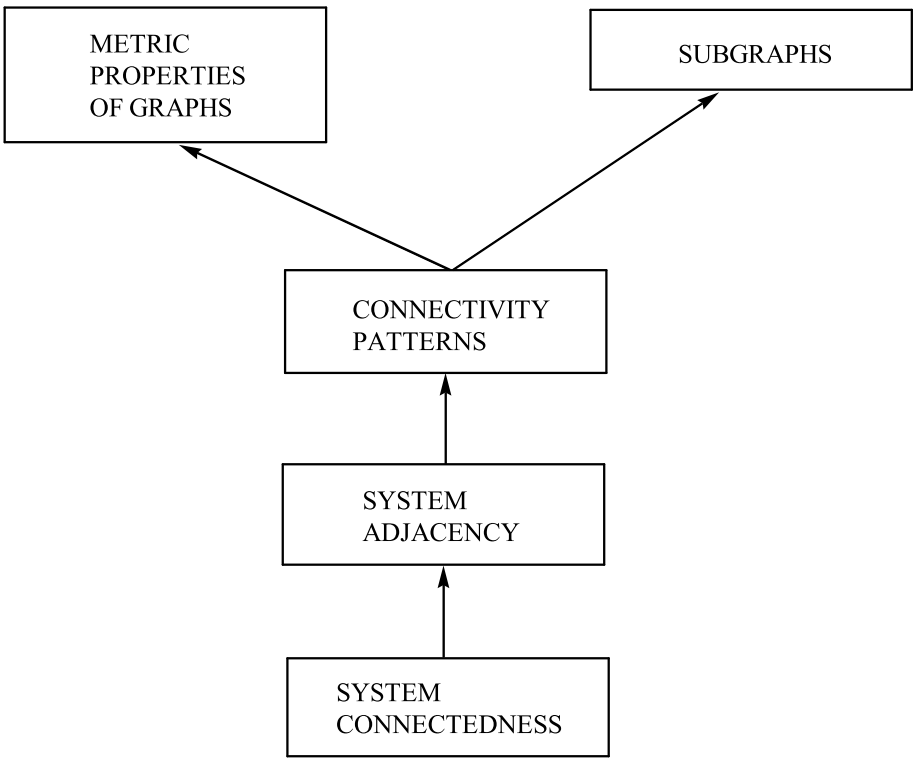

Topological Complexity of Molecules, Figure 3
A hierarchical organization of topological complexity of chemical systems as proposed by Bonchev and Seitz [5]

All of the above criteria point out that useful TCI should increase with the structural characteristics such as the graph size, branching, cyclicity, the number of weighted vertices and multiple edges but decrease with symmetry. These criteria will be used here to judge the applicability of TCIs.

Examples of Topological Complexity Indices

In this section, beginning with the Minoli index [15], we present chronologically several TCIs that were specifically designed to study the complexity of graphs and molecules. It should be noted that we use the term "graph" to mean both mathematical graphs and molecular graphs.

The Minoli Index

In 1975, Minoli proposed an index with aim to assess the combinatorial complexity of graphs. The Minoli index MI is defined as:

$$\mathrm{MI} = [VE/(V+E)] \sum_l P_l , \tag{3}$$

where V and E are the number of vertices and edges, respectively, in a graph G, while P_l is the number of paths of length l. The Minoli index is a highly degenerate TCI – it does not distinguish between isomeric graphs. In order to increase the discriminatory ability of MI, Bonchev [24] replaced the sum of the number of paths with the total length of all paths L_p. This variant of the Minoli index, called the Minoli–Bonchev index MBI, is defined as:

$$\mathrm{MBI} = [VE/(V+E)] L_\mathrm{p} \tag{4}$$

where L_p is computed by:

$$L_\mathrm{p} = \sum_l l\, P_l \, . \tag{5}$$

The Bonchev–Trinajstić Index

Two years after Minoli, Bonchev and Trinajstić [16] proposed an information-theoretic index based on the graph-theoretical distances. The Bonchev–Trinajstić index BTI is defined as:

$$\mathrm{BTI} = n \log_2 n - \sum_l n_l \log_2 n_l \tag{6}$$

where n is the total number of distances in a graph G, n_l is the number of graph-theoretical distances of length l and $n = \sum_l n_l$. The logarithm is taken at basis 2 for measuring the information content in bits. BTI was originally used to study molecular branching and was only later used to study the complexity of graphs and molecules.

The Bertz Index

Another information-theoretic index was proposed by Bertz [17]. The Bertz index BI is defined as:

$$\mathrm{BI} = 2n \log_2 n - \sum_i n_i \log_2 n_i \tag{7}$$

where n is the number of pairs of adjacent edges in a graph G and n_i is the number of pairs of adjacent edges in the ith class by symmetry. Equation (7) represents a pragmatic modification of Eq. (6). Bertz added to Eq. (7) the extra term $n \log_2 n$ to prevent $\mathrm{BI} = 0$ when all pairs of adjacent edges in G are equivalent. The first part of Eq. (7) takes into account structural characteristics of G, such as size, branching and cyclicity, and the second part deals with the symmetry of G in terms of equivalent pairs of adjacent edges. Bertz introduced his index for use in synthesis planning and analysis of synthetic strategies [19,20,22,25,26,27].

The Number of Spanning Trees

The first authors who used spanning trees explicitly as TCIs were Bonchev, Kamenski and Temkin in 1987 [28] in their work on the complexity of the linear mechanisms of chemical reactions. These authors also suggested an unwieldy method for computing spanning trees. However, in 1983, Gutman, Mallion and Essam [29] produced an elegant method for computing spanning trees of labeled graphs. These authors stated in passing "… the number of spanning trees of a labeled molecular-graph, sometimes referred to as the complexity of the graph", but they did not explore this use of the spanning trees.

The number of spanning trees of a planar graph G, $t(G)$, based on the matrix-tree theorem, is given by:

$$t(G) = \det \mathbf{L}^* \tag{8}$$

where $\mathbf{L}^*$ is the following difference matrix:

$$\mathbf{L}^* = \boldsymbol{\Delta}^* - \mathbf{A}^* \,. \tag{9}$$

$\mathbf{A}^*$ is the vertex-adjacency matrix of an inner dual G^* of G and $\boldsymbol{\Delta}^*$ is a diagonal matrix of G^* with elements equal to the size of each cycle making up the polycyclic graph. Matrix $\mathbf{L}^*$ may also be regarded as a Laplacian matrix of the vertex-weighted inner dual G^*, the weights of vertices in G^* being equal to the sizes of cycles in the parent graph G. The spanning trees can also be computed using eigenvalues λ_i of the Laplacian matrix of G [30,31]:

$$t(G) = (1/V) \prod_{i=2}^{V} \lambda_i(\mathbf{L}) \,. \tag{10}$$

It should be noted that the smallest eigenvalue of the Laplacian matrix λ_1 is always zero, thus the multiplication of eigenvalues in Eq. (10) starts with λ_2. In d-regular graphs, that is graphs in which every vertex has the same degree d, the eigenvalues of the Laplacian matrix $\lambda_i(\mathbf{L})$ and the vertex-adjacency matrix $\lambda_i(\mathbf{A})$ are related by:

$$\lambda_i(\mathbf{L}) = d - \lambda_i(\mathbf{A}) \,. \tag{11}$$

In fullerenes whose graphs are three regular graphs (d=3), Eq. (10) transforms by means of Eq. (11) into [32]:

$$t(G) = (1/V) \prod_{i=2}^{V} \left[3 - \lambda_i(\mathbf{A})\right] . \tag{12}$$

This formula was used by Fowler [33] to study the complexity of isomeric fullerenes.

Indices Based on Connected Subgraphs

Bertz and Sommer [34] and Bertz and Wright [22] advocated the use of the number of kinds of connected subgraphs N_s and the total number of connected subgraphs N_t of a graph G as TCIs. The set of all connected subgraphs starts with vertices and is followed by edges, 3-vertex subgraphs, etc. and ends with G itself, since in formal graph theory G is its own subgraph. The generation and computation of the subgraph-counting numbers N_s and N_t is a rather difficult problem. But, Rücker and Rücker [35,36] solved it by designing an efficient computer program for generating and counting all connected subgraphs of G. These authors named their program BERTZ in recognition of Bertz's work on the complexity of molecules and reactions.

Independently, Bonchev [37] also used N_t to evaluate the complexity of molecules and in the same paper he also proposed two novel TCIs based on the connected subgraphs. One of these, he called simply the topological complexity index, denoted by TC. TC is equal to the sum of vertex degrees in all connected subgraphs (or the subgraph total adjacency) of G:

$$\mathrm{TC} = \sum_s \sum_i d_i(s) \tag{13}$$

where the first sum is over all connected subgraphs s in G, the second sum is over all vertices in a subgraph i and d_i is the degree of vertices in i. It should be noted that in computing TC the vertex-degrees in subgraphs are taken as they are in G. If the vertex-degrees are taken as they are in each subgraph, the corresponding index is denoted by TC1. Thus, TC is always greater than TC1, except for a single-vertex graph, where $\mathrm{TC} = \mathrm{TC1} = 0$. TC and TC1 were the first overall TCIs.

Bonchev later used other graph invariants to generate overall indices, such as the overall Wiener index [38], based on the Wiener indices [39] of subgraphs and overall Zagreb indices [40], based on the Zagreb indices [41,42].

The Randić Complexity Index

The Randić complexity index RCI [43] is a symmetry-dependent complexity index based on the concept of augmented vertex-degree [44,45]. The augmented degree of a vertex i in a graph G is the sum of its degree and degrees of all other vertices j with weight $1/2^{l(ij)}$ depending on their distance l from the vertex i. Then the RCI is the sum of all augmented degrees of vertices that are not symmetrically equivalent:

$$\mathrm{RCI} = \sum_{i=1}^{N^*} \sum_{j=1}^{N} d_j / 2^{l(i,j)} \tag{14}$$

where N^* is the number of vertices that are not symmetrically equivalent and N is the number of all vertices in a graph G. The first sum is over only the symmetrically non-equivalent vertices and the second sum is over all vertices in G.

The Total Walk Count

Rücker and Rücker [23,46] proposed the total walk count (twc) as a measure of the complexity of graphs and molecules. The twc's can be computed from atomic walk count sums (awcs's), molecular walk counts (mwc's) and the vertex-adjacency matrices of (molecular) structures considered.

The atomic walk count of order l of atom i $(\mathrm{awc})_l(i)$ is the number of all possible walks of length l which start at the vertex i and end at any vertex j:

$$(\mathrm{awc})_l(i) = \sum_{j=1}^{V} (\mathbf{A}^l)_{ij} \,. \tag{15}$$

The molecular walk count of order l $(\mathrm{mwc})_l$ is equal to the sum of all atomic walk counts of order l:

$$(\mathrm{mwc})_l = \sum_{i=1}^{V} (\mathrm{awc})_l(i) \,. \tag{16}$$

Finally, the twc is equal to the sum of all molecular walk counts for l from 1 to $V-1$:

$$\mathrm{twc} = \sum_{l=1}^{V-1} (\mathrm{mwc})_l \,. \tag{17}$$

There is another way to compute the total walk count by utilizing a relationship between the Morgan extended connectivities MEC [47] and powers of the vertex-adjacency matrix. This relationship was first noticed by Razinger [48]; Rücker and Rücker [46] proved that MECs and $(\mathrm{awc})_l$'s are identical. Thus, the MEC of vertex i of the lth order can be computed by the iterative summation of the $(l-1)$th order contributions of the neighbors of i and is equal to $(\mathrm{awc})_l(i)$. Consequently, the calculation of twc for reasonably sized (molecular) structures requires only a sequence of addition steps. Rücker and Rücker [46] termed this procedure the Morgan summation procedure, and based on this procedure they prepared the program MORGAN, so named as a tribute to Morgan's work, which is now almost forgotten.

The twc index does not take care of the graph's symmetry. In order to do that the twc needs to be modified by symmetry. The symmetry-modified twc is called the walk complexity (wcx) [23,49]. It can be obtained from the atomic walk count sum awcs. The awcs of atom i can be computed by summing up the $(\mathrm{awc})_l(i)$ over all l from 1 to $V-1$:

$$\mathrm{awcs}(i) = \sum_{l=1}^{V-1} (\mathrm{awc})_l(i) \,. \tag{18}$$

Then, the wcx is given by the following:

$$\mathrm{wcx} = \sum_{i} \mathrm{awcs}(i) \tag{19}$$

where i is the first integer from each set of integers denoting vertices in an equivalence class by symmetry.

The Usefulness of the Reviewed Topological Complexity Indices to Assess the Complexity of Molecular Graphs

Comparison of the reviewed TCIs is carried out on four sets of simple molecular graphs, that is, on linear trees

representing carbon skeletons of *n*-alkanes containing from two to eight carbon atoms, cycles representing carbon skeletons of cycloalkanes from three to eight carbon atoms, trees on six vertices representing carbon skeletons of isomeric hexanes and three planar polycyclic graphs on four vertices representing carbon skeletons of cyclobutane, bicyclo[1.1.0]butane and tetrahedrane. The idea behind this choice of graphs is that if the TCIs fail to assess the complexity of simple graphs they certainly would not be able to handle graphs with much more complicated structures.

Complexity of Linear Trees

We considered linear trees from two to eight vertices to study the behavior of TCIs with the size increase. Linear trees are convenient to use because the structural factor strongly influencing their complexity is their size. However, there is another (silent) structural feature present in trees that cannot be avoided – the symmetry. The symmetry is identified through the number of vertices not equivalent by symmetry. It will influence the symmetry-dependent TCIs, but not dramatically. The studied linear trees are depicted in Fig. 4 and their TCIs are given in Table 1.

From Table 1, it is evident that all of considered TCIs increase with increasing linear alkane tree size. The reason for such behavior is due to the regularity of the linear alkane tree structures that allows most of the indices considered to be defined trivially in terms of the number of vertices.

MI and MB Indices

$$\mathrm{MI} = V^2(V-1)^2/(4V-2) \tag{20}$$

$$\mathrm{MB} = V^2(V+1)(V-1)^2/6(2V-1)\,. \tag{21}$$

BT and BI Indices

$$\mathrm{BT} = \binom{V}{2}\log_2\binom{V}{2} - \sum_{i=2}^{V-1} i\log_2 i \tag{22}$$

$$\mathrm{BI} = \begin{cases} 2(V-2)\log_2(V-2)-(V-2) & \text{if } V = \text{even} \\ 2(V-2)\log_2(V-2)-(V-3) & \text{if } V = \text{odd}\,. \end{cases} \tag{23}$$

t(G) Index

The $t(G)$ index is not given because for all acyclic structures; it is always equal to 1, that is, the spanning tree of a tree is identical to the tree itself.

N_s and N_t Indices

$$N_\mathrm{s} = V \tag{24}$$

$$N_\mathrm{t} = V(V+1)/2\,. \tag{25}$$

TC and TC1 Indices

$$\mathrm{TC} = V(V-1)(V+4)/3 \tag{26}$$

$$\mathrm{TC1} = V(V^2-1)/3\,. \tag{27}$$

RCI Index

$$\mathrm{RCI} = \begin{cases} 3V-6+6/2^V & \text{if } V = \text{even} \\ 3V-3-3(2^{(V-1)/2}-1)/2^{V-1} & \text{if } V = \text{odd}\,. \end{cases} \tag{28}$$

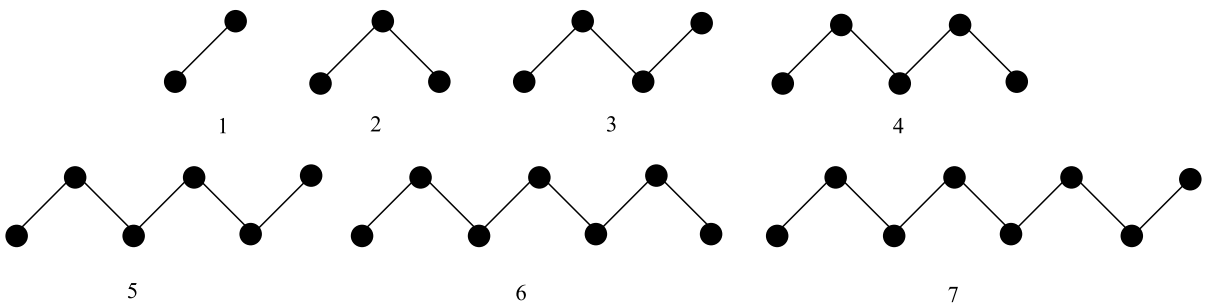

Topological Complexity of Molecules, Figure 4
Linear trees representing carbon skeletons of *n*-alkanes from two to eight carbon atoms

Topological Complexity of Molecules, Table 1
Topological complexity indices of smaller linear trees representing carbon skeletons of linear alkanes from two to eight carbon atoms

Linear alkane tree[a]	Topological complexity index[b]										
	MI	MB	BT	BI	N_s	N_t	TC	TC1	RCI	twc	wcx
1	0.7	0.7	0	–	2	3	4	2	1.5	2	1
2	3.6	4.8	2.8	0.0	3	6	14	8	5.3	10	7
3	10.3	17.1	8.8	2.0	4	10	32	20	6.4	32	16
4	22.2	44.4	18.5	7.5	5	15	60	40	11.4	88	56
5	40.9	95.5	32.2	12.0	6	21	100	70	12.1	222	111
6	67.8	180.9	50.4	19.2	7	28	154	112	17.7	536	320
7	104.5	313.6	73.1	25.0	8	36	224	168	18.0	1254	617

[a]Considered linear alkane trees, given in Fig. 4
[b]MI = Minoli index; MB = Minoli–Bonchev index; BT = Bonchev–Trinajstić index, BI = Bertz index; N_s, N_t = the number of kinds and the total number of connected subgraphs; TC, TC1 = Bonchev indices; RCI = Randić complexity index; twc = the total walk count; wcx = the walk complexity

twc Index

$$\mathrm{twc} = \mathrm{mwc}_V - 2V + \sum_{i=0}^{V-1} s_i \tag{29}$$

where

$$\mathrm{mwc}_l = 2^l V - \sum_{i=0}^{l-1} 2^{l-1-i} s_i \quad (l \leqslant V) \tag{30}$$

and

$$s_i = \begin{cases} 2\binom{i}{i/2} & \text{if } i = \text{even} \\ 2\binom{i}{(i+1)/2} & \text{if } i = \text{odd}\,. \end{cases} \tag{31}$$

wcx Index

The closed form of the wcx index is available only for V = even:

$$\mathrm{wcx} = \left(\mathrm{mwc}_V - 2V + \sum_{i=0}^{V-1} s_i\right)/2\,. \tag{32}$$

Complexity of Monocycles The next set of graphs considered is the set of lower monocycles from three to eight vertices. In their case, besides the cycle size, a more visible role is played by symmetry. The studied monocycles are depicted in Fig. 5 and their TCIs are given in Table 2.

All considered TCIs increase with the size of the cycle, thus indicating the larger the cycle the greater complexity of the cycle. The regular structures of the monocycles allows one to express the employed TCIs simply in terms of the number of vertices.

MI and MB Indices

$$\mathrm{MI} = [V^2(V-1)]/2 \tag{33}$$

$$\mathrm{MB} = V^3(V-1)/4\,. \tag{34}$$

BT and BI Indices

$$\mathrm{BT} = \begin{cases} \binom{V}{2}\log_2(V-1) - V(V-2)/2 & \text{if } V = \text{even} \\ \binom{V}{2}\log_2\left[(V-1)/2\right] & \text{if } V = \text{odd} \end{cases} \tag{35}$$

$$\mathrm{BI} = V\log_2 V\,. \tag{36}$$

t(G) Index

$$t(G) = V\,. \tag{37}$$

N_s and N_t Indices

$$N_s = V + 1 \tag{38}$$

$$N_t = V^2 + 1\,. \tag{39}$$

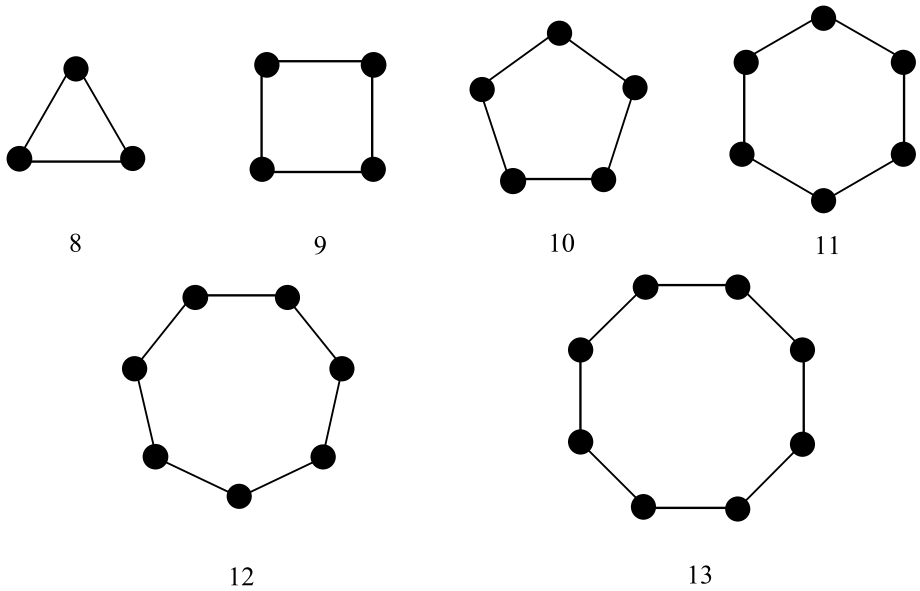

Topological Complexity of Molecules, Figure 5
Cycles representing carbon skeletons of cycloalkanes from three to eight carbon atoms

Topological Complexity of Molecules, Table 2
Topological complexity indices of smaller monocycles representing carbon skeletons of cycloalkanes from three to eight carbon atoms

Cycle[a]	Topological complexity index[b]											
	MI	MB	BT	BI	$t(G)$	N_s	N_t	TC	TC1	RCI	twc	wcx
8	9	13.5	0	4.8	3	4	10	42	24	4.00	18	6
9	24	48.0	5.5	8.0	4	5	17	88	56	4.50	56	14
10	50	125.0	10.0	11.6	5	6	26	160	110	5.00	150	30
11	90	270.0	22.8	15.5	6	7	37	264	192	5.25	372	62
12	147	514.5	33.3	19.7	7	8	50	406	308	5.50	882	126
13	224	896.0	54.6	24.0	8	9	65	592	464	5.63	2032	254

[a]Considered monocycles, given in Fig. 5
[b]MI = Minoli index; MB = Minoli–Bonchev index; BT = Bonchev–Trinajstić index, BI = Bertz index; $t(G)$ = the number of spanning trees; N_s, N_t = the number of kinds and the total number of connected subgraphs; TC, TC1 = Bonchev indices; RCI = Randić complexity index; twc = the total walk count; wcx = the walk complexity

TC and TC1 Indices

$$\mathrm{TC} = V[V(V+1)+2] \tag{40}$$

$$\mathrm{TC1} = V[V(V-1)+2]\,. \tag{41}$$

RCI Index

$$\mathrm{RCI} = \begin{cases} 6(2^{V/2}-1)/2^{V/2} & \text{if } V = \text{even} \\ 6 - 1/2^{(V-5)/2} & \text{if } V = \text{odd}\,. \end{cases} \tag{42}$$

twc and wcx Indices

$$\mathrm{twc} = 2V(2^{V-1}-1) \tag{43}$$

$$\mathrm{wcx} = 2(2^{V-1}-1)\,. \tag{44}$$

Branching and Complexity

To illustrate the impact of branching on the complexity, we selected isomeric trees on six vertices representing the carbon skeletons of hexanes. It should also be pointed out

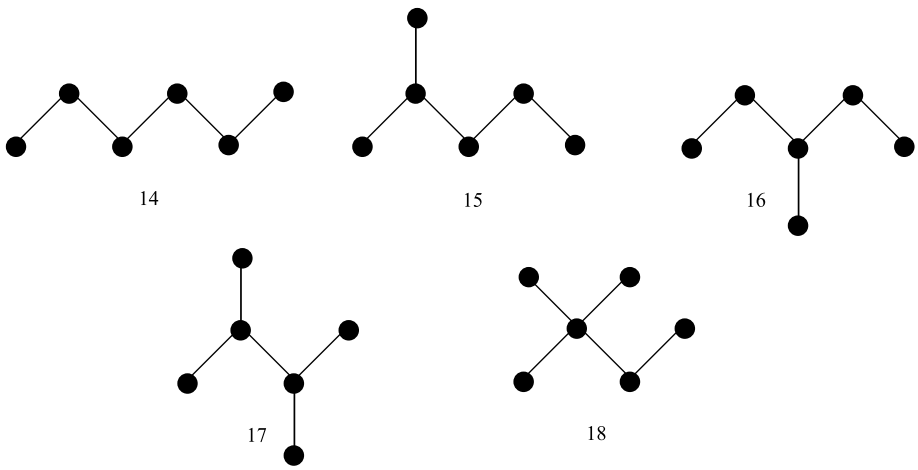

Topological Complexity of Molecules, Figure 6
Hexane trees representing carbon skeletons of isomeric hexanes

that, as before, the impact of branching cannot be isolated from the impact of symmetry on the complexity. Branching is intuitively identified through the appearance of vertices with degrees higher than two – trees containing only vertices with degrees one and two are not branched. The symmetry impact will be strongly reflected on the symmetry-dependent TCIs. The isomeric hexane trees are depicted in Fig. 6 and their TCIs are given in Table 3. The Minoli index is not included since it possesses the same value (40.9) for all hexane trees.

As can be seen from Table 3, four of the symmetry-independent TCIs (N_t, TC, TC1, twc) indicate that the increasing branching of a hexane tree *via* the number of branched vertices increases the tree complexity as one would expect intuitively. The symmetry-dependent TCIs (MB, BT) produced the anti-intuitive prediction, that is, increasing branching decreases the tree complexity. However, if we consider −MB and −BT, then these indices agree in their prediction with the above four symmetry-independent TCIs. The four remaining TCIs (BI, N_s, RCI, wcx) behave erratically, but two of them (BI, wcx) agree at least in predicting that **14** will be the least and **18** the most complex structure among the five isomeric hexane trees. N_s and RCI predict **14** to be the least complex structure, but the complexity-ordering of the remaining four isomeric trees is dubious.

Topological Complexity of Molecules, Table 3
Trees on six vertices representing carbon skeletons of isomeric hexanes

Hexane tree[a]	Topological complexity index[b]									
	MB	BT	BI	N_s	N_t	TC	TC1	RCI	twc	wcx
14	95.5	32.2	12.0	6	21	100	70	12.1	222	111
15	87.3	28.6	21.2	8	24	127	88	21.9	268	234
16	84.5	27.4	19.2	8	25	136	94	17.8	284	201
17	79.1	23.5	21.0	7	28	164	112	9.8	330	124
18	76.4	22.6	29.8	8	30	181	122	19.5	370	272

[a]Hexane trees are given in Fig. 6
[b]MB = Minoli–Bonchev index; BT = Bonchev–Trinajstić index, BI = Bertz index; N_s, N_t = the number of kinds and the total number of connected subgraphs; TC, TC1 = Bonchev indices; RCI = Randić complexity index; twc = the total walk count; wcx = the walk complexity

Cyclicity and Complexity

Cyclicity can be considered either in terms of closing a cycle or increasing the number of cycles. If we consider the former case, then we can compare TCIs in Tables 1 and 2. Comparison between the linear trees and monocycles with the same number of vertices shows that seven TCIs (MI, MB, N_s, N_t, TC, TC1, twc) always possess greater values for monocycles than for linear alkanes, thus indicating that the monocyclic structures are more complex than linear structures. This agrees with conclusion reached by Bertz and Zamfirescu [50] that C_V is more complex than P_V for all V, where C_V and P_V denote monocycles and linear trees, respectively, with V vertices. However, the four symmetry-dominated TCIs (BT, BI, RCI, wcx) violate this. This is expected since monocycles are more symmetry-rich than linear alkanes. However, it should be noted that whilst BT, RCI and wcx are always smaller for monocycles than for the corresponding linear trees, the BI index ex-

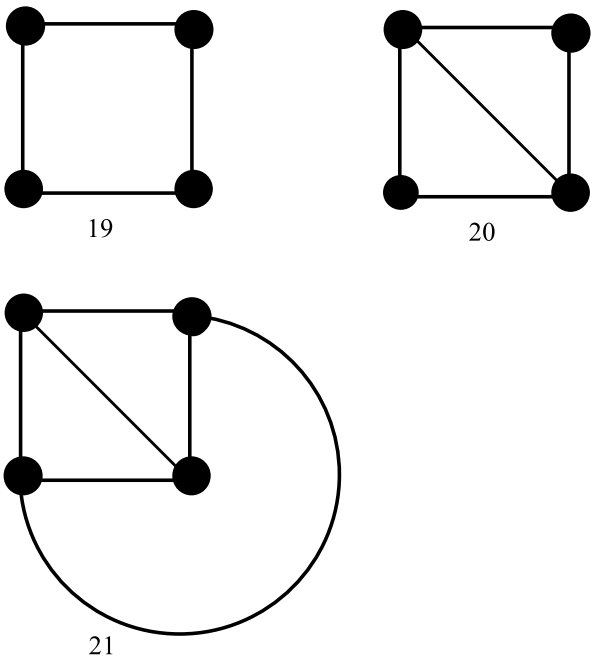

Topological Complexity of Molecules, Figure 7
Planar cyclic graphs on four vertices

hibits such behavior only for linear trees and monocycles with eight or more vertices.

As stated above cyclicity also manifests itself through the increasing number of cycles. To illustrate this, we considered three planar cyclic graphs on four vertices depicting carbon skeletons of well-known organic molecules: cyclobutane (one cycle), bicyclo[1.1.0]butane (two cycles) and tetrahedrane (three cycles). These three cyclic graphs are depicted in Fig. 7 and their TCIs are given in Table 4.

Cyclicity ordering as implied by consideration of cyclic graphs *via* the numbers of cycles is followed by ten TCIs (MI, MB, BI, $t(G)$, N_s, N_t, TC, TC1, twc, wcx). BT predicts the inverse order, but its prediction could be amended if −BT is considered. Only in the case of RCI, a strong symmetry-dependent TCI predicts counter-intuitively that **20** is more complex than **21** since **21** is a symmetry-richer structure than **20**.

Summary

We applied 12 TCIs on four sets of simple molecular graphs. They exhibit the following characteristics:

(i) MI increases with increasing size of linear trees and monocycles and with increasing cyclicity. It is not applicable to isomeric structures.

(ii) MB and BT increase with increasing size of linear trees and monocycles and decrease with increasing branching. However, if one considers −MB and −BT, then these indices increase with increasing branching. MB increases, but BT decreases with increasing cyclicity. Again if one considers −BT, it increases with increasing cyclicity.

(iii) $t(G)$ can be used to assess the complexity only of graphs containing cycles.

(iv) N_s increases with increasing size of linear trees and monocycles and with increasing cyclicity, but fails in the case of branching where it predicts only the least branched structure.

(v) N_t increases with increasing size of linear trees and monocycles and with increasing branching and cyclicity. This is a good TCI, but the procedure of getting N_t increases exponentially with problem size.

(vi) TC and TC1 behave similarly to N_t, as expected, since they depend on the knowledge of all connected subgraphs. Thus, they suffer from the same computational problem as does N_t. However, for small to

Topological Complexity of Molecules, Table 4
Planar cyclic graphs on four vertices representing carbon skeletons of cyclobutane, bicyclo[1.1.0]butane and tetrahedrane

Cyclic graph[a]	Topological complexity index[b]											
	MI	MB	BT	BI	$t(G)$	N_s	N_t	TC	TC1	RCI	twc	wcx
19	24.0	48.0	5.5	8.0	4	5	17	88	56	4.5	56	14
20	42.2	86.7	3.9	36.0	8	9	33	254	152	12.0	102	51
21	72.0	158.4	0.0	43.0	16	10	64	648	372	7.5	156	39

[a]Planar cyclic graphs, given in Fig. 7
[b]MI = Minoli index; MB = Minoli–Bonchev index; BT = Bonchev–Trinajstić index, BI = Bertz index; $t(G)$ = the number of spanning trees; N_s, N_t = the number of kinds and the total number of connected subgraphs; TC, TC1 = Bonchev indices; RCI = Randić complexity index; twc = the total walk count; wcx = the walk complexity

medium size molecular graphs all these three indices appear to be useful TCIs.

(vii) The twc increases with increasing size of linear trees and monocycles and with increasing branching and cyclicity. Therefore, the twc is also a useful TCI regarding the graphs considered.

(viii) BI increases with increasing size of linear trees and monocycles, but decreases with increasing cyclicity. With respect to branching, BI predicts only the least and the most branched structures. This index also possesses a strongly symmetry-dependent component and hence it does not produce a consistent ordering between monocycles and the corresponding linear trees.

(ix) RCI and wcx suffer from their overemphasis on the role of symmetry. Therefore, the use of these two TCIs and BI require explicit input of symmetry information *via* a reliable approach for symmetry recognition e. g., [51,52,53].

Future Directions

Assessment of complexity apparently requires complex approaches. Perhaps a way to proceed is to consider that complexity is not a strictly numerical quantity. Some authors [50,54,55] already pointed out that the complexity is a partially ordered quantity. The Hasse diagram reflecting the partial order of indices appears to be a very useful device with which to appraise the topological complexity of (molecular) graphs. One way to construct the Hasse diagram for a set of molecular graphs is as follows: The graph **A** in a diagram is placed above graph **B** if all TCIs considered have a smaller value for **A** than for **B**. Besides, two graphs **A** and **B** are directly linked by an edge downward from **A** to **B** if and only if no third graph is placed by this partial ordering between them.

Bibliography

Primary Literature

1. Waldrop MM (1992) Complexity: The emerging science at the edge of order and chaos. Touchstone, New York
2. Rouvray DH (2003) An introduction to complexity. In: Bonchev D, Rouvray DH (eds) Complexity in chemistry. Taylor & Francis, London, pp 1–27
3. Rouvray DH (1997) Are the concepts in chemistry all fuzzy? In: Rouvray DH (ed) Concepts in chemistry: A contemporary challenge. Research Studies Press, New York, pp 1–15
4. Nikolić S, Trinajstić N, Tolić IM, Rücker G, Rücker C (2003) On molecular complexity indices. In: Bonchev D, Rouvray DH (eds) Complexity in chemistry. Taylor & Francis, London, pp 29–89
5. Bonchev D, Seitz WA (1997) The concept of complexity in chemistry. In: Rouvray DH (ed) Concepts in chemistry: A contemporary challenge. Research Studies Press, Taunton, pp 353–381
6. Bonchev D, Buck GA (2007) From molecular to biological structure and back. J Chem Inf Model 47:909–917
7. Dancoff SM, Quastler H (1953) The information content and error rate in living things. In: Quastler H (ed) Essays on the Use of Information Theory in Biology. University of Illinois Press, Urbana
8. Shannon C, Weaver W (1949) Mathematical theory of communications. University of Illinois Press, Urbana
9. Morowitz H (1955) Some order-disorder considerations in living systems. Bull Math Biophys 17:81–86
10. Rashevsky N (1955) Life, information theory and topology. Bull Math Biophys 17:229–235
11. Karreman G (1955) Topological information content and chemical reactions. Bull Math Biophys 17:279–285
12. Trucco E (1956) A note on the information content of graphs. Bull Math Biophys 18:129–135
13. Trucco E (1956) On the information content of graphs: Compound symbols. Different states for each point. Bull Math Biophys 18:237–253
14. Mowshovitz A (1968) Entropy and the complexity of graphs: I. An index of the relative complexity of a graph. Bull Math Biophys 30:175–204
15. Minoli D (1975) Combinatorial graph complexity. Atti della Accademia Nazionale dei Lincei – Classe di Scienze fisiche, mathematiche e naturali (Serie 8) 59:651–661
16. Bonchev D, Trinajstić N (1977) Information theory, distance matrix and molecular branching. J Chem Phys 67: 4517–4533
17. Bertz SH (1983) The first general index of molecular complexity. J Am Chem Soc 103:3599–3601
18. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley, Weinheim, pp 300
19. Bertz SH (1983) A mathematical model of complexity. In: King RB (ed) Chemical applications of topology and graph theory. Elsevir, Amsterdam, pp 206–221
20. Bertz SH (1983) On the complexity of graphs and molecules. Bull Math Biol 45:849–855
21. Bonchev D, Polansky OE (1987) On the topological complexity of chemical systems. In: King RB, Rouvray DH (eds) Graph Theory and Topology in Chemistry. Elsevier, Amsterdam, pp 125–158
22. Bertz SH, Wright WF (1998) The graph theory approach to synthetic analysis: Definition and application of molecular complexity and synthetic complexity. Graph Theory Notes New York 35:32–48
23. Rücker G, Rücker C (2000) Walk counts, labirinthicity and complexity of acyclic and cyclic graphs and molecules. J Chem Inf Comput Sci 40:99–106
24. Bonchev D (1990) Problems of computing molecular complexity. In: Rouvray DH (ed) Computational Chemical Graph Theory. Nova Science Publishers, NY, pp 33–63
25. Bertz SH (1982) Convergence, molecular complexity and synthetic analysis. J Am Chem Soc 104:5801–5803
26. Bertz SH (2003) Complexity of molecules and their synthesis. In: Bonchev D, Rouvray DH (eds) Complexity in Chemistry. Taylor & Francis, London, pp 91–156
27. Bertz SH, Rücker C (2004) In search of simplification: the use of topological complexity indices to guide retrosynthetic analysis. Croat Chem Acta 77:221–235

28. Bonchev D, Kamenski D, Temkin ON (1987) Complexity index of the linear reaction mechanisms of chemical reactions. J Math Chem 1:345–388
29. Gutman I, Mallion RG, Essam JW (1983) Counting the spanning trees of a labelled molecular-graph. Mol Phys 50:859–877
30. Mohar B (1989) Laplacian matrices of graphs. In: Graovac A (ed) MATH/CHEM/COMP 1989. Elsevier, Amsterdam, pp 1–8
31. Trinajstić N, Babić D, Plavšić, Amić D, Mihalić Z (1994) The Laplacian matrix in chemistry. J Chem Inf Comput Sci 34: 368–376
32. Mihalić Z, Trinajstić N (1994) On the number of spanning trees in fullerenes. Fullerene Sci Technol 2:89–95
33. Fowler PW (2003) Complexity, spanning trees and relative energies of fullerene isomers. MATCH Comun Math Comput Chem 48:87–96
34. Bertz SH, Sommer TJ (1997) Rigorous mathematical approaches to strategic bonds and synthetic analysis based on conceptually simple new complexity indices. Chem Com 2409–2410
35. Rücker G, Rücker C (2001) On finding non-isomorphic connected subgraphs and distinct molecular substructures. J Chem Inf Comput Sci 41:314–320; erratum 825
36. Rücker G, Rücker C (2001) Substructure, subgraph and walk counts as measures of the complexity of graphs and molecules. J Chem Inf Comput Sci 41:1457–1462
37. Bonchev D (1997) Novel indices for the topological complexity of molecules. SAR QSAR Environ Res 7:23–43
38. Bonchev D (2001) The Overall Wiener index – A new tool for characterizing molecular topology. J Chem Inf Comput Sci 41:582–592
39. Wiener H (1947) Structural determination of paraffin boiling points. J Am Chem Soc 69:17–20
40. Bonchev D, Trinajstić N (2001) Overall molecular descriptors. 3. Overall Zagreb indices. Sar & QSAR Environ Res 12:213–235
41. Gutman I, Trinajstić N (1972) Graph theory and molecular orbitals. III. Total π-Electron Energy of Alternant Hydrocarbons. Chem Phys Lett 17:535–538
42. Gutman I, Ruščić B, Trinajstić N, Wilcox Jr CF (1975) Graph theory and molecular orbitals. XII. Acyclic polyenes. J Chem Phys 62:3399–3405
43. Randić M (2001) On complexity of transitive graphs representing degenerate rearrangments. Croat Chem Acta 74:683–705
44. Randić M, Plavšić D (2002) On the concept of molecular complexity. Croat Chem Acta 75:107–116
45. Randić M, Plavšić D (2003) On characterization of molecular complexity. Int J Quantum Chem 91:20–31
46. Rücker G, Rücker C (1993) Counts of all walks as atomic and molecular descriptors. J Chem Inf Comput Sci 33:683–695
47. Morgan HL (1965) The generation of a unique machine description for chemical structures – A technique developed at Chemical Abstracts Service. J Chem Doc 5:107–113
48. Razinger M (1982) Extended connectivity of chemical graphs. Theor Chim Acta 61:581–586
49. Gutman I, Rücker C, Rücker G (2001) On walks in molecular graphs. J Chem Inf Comput Sci 41:739–745
50. Nikolić S, Tolić IM, Trinajstić N, Baučić I (2000) On the Zagreb indices as complexity indices. Croat Chem Acta 73:909–921; see Note 28 and remarks by DJ Klein on complexity partial ordering
51. Rücker G, Rücker C (1990) Computer perception of constitutional (topological) symmetry: TOPSYM, a fast algorithm for partitioning atoms and pairwise relations among atoms into equivalence classes. J Chem Inf Comput Sci 30:187–191
52. Rücker G, Rücker C (1991) On using the adjacency matrix power method for perception of symmetry and for isomorphism testing of highly intricate graphs. J Chem Inf Comput Sci 31:123–126
53. Rücker G, Rücker C (1991) Isocodal and isospectral points, edges and pairs in graphs and how to cope with them in computerized symmetry recognition, J Chem Inf Comput Sci 31:422–427
54. Bertz SH, Zamfirescu C (2000) New complexity indices based on edge covers. MATCH Com Math Comput Chem 42:39–70
55. Rajtmajer SM, Miličević A, Trinajstić N, Randić M, Vukičević D (2006) On the complexity of Archimedean solids. J Math Chem 39:119–132

Books and Reviews

Cvetković DM, Doob M, Sachs H (1995) Spectra of graphs – Theory and application, 3rd edn. Barth, Heidelberg

Devillers J, Balaban AT (eds) (1999) Topological descriptors and related descriptors in QSAR and QSPR. Gordon & Breach, Amsterdam

Harary F (1971) Graph theory, 2nd print. Addison-Wesley, Reading

Janežič D, Miličević A, Nikolić S, Trinajstić N (2007) Graph-Theoretical Matrices in Chemistry. University of Kragujevac, Kragujevac

Trinajstić N (1992) Chemical graph theory, 2nd edn. CRC, Boca Raton

Wilson RJ (1972) Introduction to graph theory. Oliver & Boyd, Oxford

Topological Dynamics

Ethan Akin
Mathematics Department, The City College,
New York City, USA

Article Outline

Glossary
Definition of the Subject
Introduction and History
Dynamic Relations, Invariant Sets and Lyapunov Functions
Attractors and Chain Recurrence
Chaos and Equicontinuity
Minimality and Multiple Recurrence
Future Directions
Cross References
Bibliography

Glossary

Attractor An attractor is an invariant subset for a dynamical system such that points sufficiently close to the set

remain close and approach the set in the limit as time tends to infinity.

Equilibrium An equilibrium, or a fixed point, is a point which remains at rest for all time.

Invariant set A subset is invariant if the orbit of each point of the set remains in the set at all times, both positive and negative. The set is + invariant, or forward invariant, if the forward orbit of each such point remains in the set.

Lyapunov function A continuous, real-valued function on the state space is a Lyapunov function when it is non-decreasing on each orbit as time moves forward.

Orbit The orbit of an initial position is the set of points through which the system moves as time varies positively and negatively through all values. The forward orbit is the subset associated with positive times.

Recurrence A point is recurrent if it is in its own future. Different concepts of recurrence are obtained from different interpretations of this vague description.

Repellor A repellor is an attractor for the reverse system obtained by changing the sign of the time variable.

Transitivity A system is transitive if every point is in the future of every other point. Periodicity, minimality, topological transitivity and chain transitivity are different concepts of transitivity obtained from different interpretations of this vague description.

Definition of the Subject

A *dynamical system* is model for the motion of a system through time. The time variable is either discrete, varying over the integers, or continuous, taking real values. The systems considered in topological dynamics are primarily deterministic, rather than stochastic, so the the future states of the system are functions of the past. As the name suggests, topological dynamics concentrates on those aspects of dynamical systems theory which can be grasped by using topological methods.

Introduction and History

The many branches of dynamical systems theory are outgrowths of the study of differential equations and their applications to physics, especially celestial mechanics. The classical subject of ordinary differential equations remains active as a topic in analysis, see, e. g. the older texts Coddington and Levinson [28] and Hartman [39] as well as Murdock [50]. One can observe the distinct fields of differentiable dynamics, measurable dynamics (that is, ergodic theory) and topological dynamics all emerging in the work of Poincaré on the Three Body Problem (for a history, see Barrow-Green [23]).

The transition from the differential equations to the dynamical systems viewpoint can be seen in the two parts of the great book Nemitskii and Stepanov [52]. We can illustrate the difference by considering the *initial value problem* in ordinary differential equations:

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= \xi(x) \\ x(0) &= p\,. \end{aligned} \tag{1}$$

Here x is a vector variable in a Euclidean space $X = \mathbb{R}^n$ or in a manifold X, and the initial point p lies in X. The infinitesimal change $\xi(x)$ is thought of as a vector attached to the point x so that ξ is a vector field on X.

The associated *solution path* is the function ϕ such that as time t varies, $x = \phi(t, p)$ moves in X according to the above equation and with $p = \phi(0, p)$ so that p is associated with the initial time $t = 0$. The solution is a curve in the space X along which x moves beginning at the point p. A theorem of differential equations asserts that the function ϕ exists and is unique, given mild smoothness conditions, e. g. Lipschitz conditions, on the function ξ.

Because the equation is *autonomous*, i. e. ξ may vary with x, but is assumed independent of t, the solutions satisfy the following *semigroup identities*, sometimes also called the *Kolmogorov equations*:

$$\phi(t, \phi(s, p)) = \phi(t + s, p)\,. \tag{2}$$

Suppose we solve Eq. (1), beginning at p, and after s units of time, we arrive at $q = \phi(s, p)$. If we again solve the equation, beginning now at q, then the identity (2) says that we continue to move along the old curve at the same speed. Thus, after t units of time we are where we would have been on the old solution at the same time, $t + s$ units after time 0.

The initial point p is a parameter here. For each solution path it remains constant, the fixed base point of the path. The solution path based at p is also called the *orbit* of p when we want to emphasize the role of the initial point. It follows from the semigroup identities that distinct solution curves, regarded as subsets of X, do not intersect and so X is subdivided, *foliated*, by these curves. Changing p may shift us from one curve to another, but the motion given by the original differential equation is always on one of these curves.

The gestalt switch to the dynamical systems viewpoint occurs when we reverse the emphasis between t and p. Above we thought of p as a fixed parameter and t as the time variable along the solution path. Instead, we now think of the initial point, relabeled x, as our variable and the time value as the parameter.

For each fixed t value we define the time-t map $\phi^t\colon X \to X$ by $\phi^t(x) = \phi(t,x)$. For each point $x \in X$ we ask whither it has moved in t units of time. The function $\phi\colon T \times X \to X$ is called the *flow* of the system and the semigroup identities can be rewritten:

$$\phi^t \circ \phi^s = \phi^{t+s} \text{ for all } t,s \in T\,. \tag{3}$$

These simply say that the association $t \mapsto \phi^t$ is a group homomorphism from the additive group T of real numbers to the automorphism group of X. In particular, observe that the time-0 map ϕ^0 is the identity map 1_X.

We originally obtained the flow ϕ by solving a differential equation. This requires differentiable structure on the underlying space which is why we specified that X be a Euclidean space or a manifold. The automorphism group is then the group of *diffeomorphisms* on X.

For the subject of topological dynamics we begin with a flow, a continuous map ϕ subject to the condition (3). In modern parlance a flow is just a continuous group action of the group T of additive reals on the topological space X. The automorphism group is the group of *homeomorphisms* on X. If we replace the group of reals by letting T be the group of integers then the action is entirely determined by the generator $f =_{\text{def}} \phi^1$, the time 1 homeomorphism with ϕ^n obtained by iterating f n times if n is positive and iterating the inverse $f^{-1}|n|$ times if n is negative.

Above I mentioned the requirement that the vector field ξ be smooth in order that the flow function ϕ be defined. However, I neglected to point out that in the absence of some sort of boundedness condition the solution path might go to infinity in a finite time. In such a case the function ϕ would not be defined on the entire domain $T \times X$ but only on some open subset containing $\{0\} \times X$. The problems related to this issue are handled in the general theory by assuming that the space X is compact. Of course, Euclidean space is only locally compact. In applying the general theory to systems on a noncompact space one usually restricts to some compact invariant subset or else compactifies, i. e. embeds the system in one on a larger, compact space. Already in Nemytskii and Stepanov [52] much attention is devoted to conditions so that a solution path has a compact closure, see also Bahtia and Szego [27].

As topological dynamics has matured the theory has been extended to cover the action of more general topological groups T, usually countable and discrete, or locally compact, or *Polish* (admits a complete, separable metric). This was already emphasized in the first treatise which explicitly concerned topological dynamics, Gottschalk and Hedlund [38].

In differentiable dynamics we return to the case where the space X is a manifold and the flow is smooth. The breadth and depth of the results then obtained make it much more than a subfield of topological dynamics, see, for example, Katok and Hasselblatt [45]. For measurable dynamics we weaken the assumption of continuity to mere measurability but assume that the space carries a measure invariant with respect to the flow, see, for example, Peterson [54] and Rudolph [56]. The measurable and topological theories are especially closely linked with a number of parallel results, see Glasner [36] as well as Alpern and Prasad [12]. The reader should also take note of Oxtoby [53], a beautiful little book which explicitly describes this parallelism using a great variety of applications.

The current relationship between topological dynamics and dynamical systems theory in general is best understood by analogy with that between point-set, or general, topology and analysis.

General topology proper, even excluding algebraic topology and homotopy theory, is a large specialty with a rich history and considerable current research (for some representative surveys see the Russian Encyclopedia volumes [13,14,15]). But much of this work is little known to nonspecialists. On the other hand, the fundamentals of point-set topology are part of the foundation upon which modern analysis is built. Compactness was a rather new idea when it was used by Jesse Douglas in his solution of the Plateau Problem, Douglas [32] (see Almgren [11]). Nowadays continuity, compactness and connectedness are in the vocabulary of every analyst. In addition, there recur unexpected applications of hitherto specialized topics. For example, indecomposable continua, examined by Bing and his students, see Bing [25], are now widely recognized and used in the study of strange attractors, see Brown [26].

Similarly, the area of topological dynamics proper is large and some of the more technical results have found application. We will touch on some of these in the end, but for the most part this article will concentrate on those basic aspects which provide a foundation for dynamical systems theory in general. We will focus on chain recurrence and the theory of attractors, following the exposition of Akin [1] and Akin, Hurley and Kennedy [8]).

To describe what we want to look for, let us begin with the simplest qualitative situation: the differential equation model (1) where the vector field ξ is the gradient of some smooth real-valued potential function U on X. Think of X as the Euclidean plane and the graph of U as a surface in space over X. The motion in X can be visualized on the surface above. On the surface it is always upward, perpendicular to the contour curves of constant height. For simplicity we will assume that U has isolated critical points.

These critical points: local maxima, minima and saddles, are *equilibria* for the system, points at which the the gradient field ξ vanishes. We observe two kinds of behavior. The orbit of a critical point is constant, resting at equilibrium. The other kind exhibits what engineers call *transient* behavior. A non-equilibrium solution path moves asymptotically toward a critical point (or towards infinity). As it approaches its limit, the motion slows, becoming imperceptible, indistinguishable from rest at the limit point. The set of points whose orbits tend to a particular critical point e is called the *stable set* for e.

Each local maximum e is an *attractor* or *sink*. The stable set for e is an open set containing e which is called the *domain of attraction* for e. Such a state is called *asymptotically stable* illustrated by the rest state of a cone on its base.

The local minima are *repellors* or *sources* which are attractors for the system with time reversed. Solution paths near a repellor move away from it. The stable set for a repellor e consists of e alone. Consider a cone balanced on its point.

A saddle point between two local maxima is like the highest point of a pass between two mountains. Separating the domains of attraction for the two peaks are solution paths which have limit the saddle point equilibrium.

There is a kind of knife, used for cutting bread dough, which has a semi-circular blade. The saddle point equilibrium is a state like this knife balanced on the midpoint of its blade. A slight perturbation will cause it to fall down on one side or the other. But if you start the knife balanced elsewhere along its blade there remains the possibility – not achievable in practice – of its rolling back along the blade toward balance at the midpoint equilibrium. Notice that these are first-order systems with no momentum. Imagine everything going on in thick, clear molasses. As this model illustrates, it is usually true that the stable set of a saddle point is a lower dimensional set in X and the union of the domains of attraction of the local maxima is a dense open subset of X.

There do exist examples where the stable set of a saddle has nonempty interior. This is a pathology which we will hope to exclude by imposing various conditions. For example, the potential function U is called a *Morse function* when all of its critical points are nondegenerate. That is, the *Hessian* matrix of second partials is nonsingular at each critical point. For the gradient system of a Morse function each equilibrium is of a type called *hyperbolic*. For the saddle points of such a system the stable sets are manifolds of lower dimension.

From the cone example, we omitted what physicists call *neutral stability*, the cone resting on its side. From our point of view this is another sort of pathology: an infinite, connected set of equilibria. Each of these equilibria is *stable* but not asymptotically stable. If we perturb the cone by lifting its point and turning it a bit, then it drops back toward an equilibrium near to but not necessarily identical with the original state (Remember, no momentum).

We obtain a similar classification into sinks, saddles, etc. and complementary transient behavior when we remove the assumption that the system comes from the gradient of U and retain only the condition that U increases along nonequilibrium solution paths. The function U is then called a *strict Lyapunov function* for the system. Instead of the steepest path ascent of a mountain goat, we may observe the spiraling upward of a car on a mountain road.

However, these gradient-like systems are too simple to represent a typical dynamical system. Lacking is the general behavior complementary to transience, namely *recurrence*. A point is recurrent – in some sense – if it is "in its own future" – in the appropriate sense. Beyond equilibrium the simplest kind of recurrence occurs on a periodic orbit. A periodic orbit returns infinitely often to each point on the orbit. As we will see, there are increasingly broad concepts of recurrence obtained by extending the notion of the "future" of a point. Clearly, a real-valued function cannot be strictly increasing along a periodic orbit. When we consider Lyapunov functions in general we will see that they remain constant along the orbit of each recurrent point.

Nonetheless, the picture we are looking for can be related to the gradient landscape by replacing the critical points by blobs of various sizes. Each blob is a closed, invariant set of a special type. A subset A is an *invariant set* for the system when A contains the entire orbit of each of its points. The special condition on each blob A which replaces an equilibrium is a kind of *transitivity*. This means that if p and q are points of A then each point is in the "future" of the other and so, in particular, each point is recurrent in the appropriate sense. Here again it remains to provide a meaning – or actually several different meanings – for this vague notion of "future".

In this more general situation the transient orbits need not converge to a point. Instead, each accumulates on a closed subset of one of these blobs. If there are only finitely many of the blobs then there is a classification of them as attractor, repellor, or saddle analogous to the description for equilibria in the gradient system.

Within each blob the motion may be quite complicated. It is in attempting to describe such motions that the concept of *chaos* arises.

For some applied fields this sort of thinking is relatively new. When I learned population genetics – admit-

tedly that was over thirty years ago – most of the analysis consisted of identifying and classifying the equilibria, tricky enough in several variables. This is perfectly appropriate for gradient-like systems and is a good first step in any case. However, it has become apparent that more complicated recurrence may occur and so requires attention.

While most applications use differential equations and the associated flows, it is more convenient to develop the discrete time theory and then to derive from it the results for flows. In what follows we will describe the results for a *cascade*, a homeomorphism f on a compact metric space X with the dynamics introduced by iteration. Focusing on this case, we will not discuss further real flows or noninvertible functions, and we will omit as well the extensions to noncompact state spaces and to compact, nonmetrizable spaces.

Dynamic Relations, Invariant Sets and Lyapunov Functions

It is convenient to assume that our state spaces are nonempty and metrizable. It is essential to assume that they are compact. Recall that if A is a subset of a compact metric space X then A is closed if and only if it is compact. Also, the continuous image of a compact set is compact and so for continuous maps between compact metric spaces the image of a closed set is closed. Perhaps less familiar are the following important results:

Proposition 1 *Let X be a compact metric space and $\{A_n\}$ be a decreasing sequence of closed subsets of X with intersection A.*

(a) If U is an open subset of X with $A \subseteq U$ then for sufficiently large n, $A_n \subseteq U$.
(b) If A_n is nonempty for every n, then the intersection A is nonempty.
(c) If $h\colon X \to Y$ is a continuous map with Y a metric space then

$$\bigcap_n h(A_n) = h(A)\,. \tag{4}$$

Proof (a): We are assuming $A_{n+1} \subseteq A_n$ for all n and $A = \bigcap_n A_n$. The complementary open sets $V_n = X \setminus A_n$ are increasing and, together with U, they cover X. By compactness, $\{U, V_1, \ldots, V_N\}$ covers X for some N and so $\{U, V_N\}$ suffice to cover X. Hence, A_N is a subset of U as is A_n for any $n \geq N$.

(b): If A is empty then $U = \emptyset$ is an open set containing A and so by (a), A_n is empty for sufficiently large n.

(c): Since $A \subseteq A_n$ for all n, it is clear that $h(A)$ is contained in $\bigcap_n h(A_n)$. On the other hand, if y is a point of the latter intersection then $\{h^{-1}(y) \cap A_n\}$ is a decreasing sequence of nonempty compact sets. By (b) the intersection $h^{-1}(y) \cap A$ is nonempty. □

If $\{B_n\}$ is any sequence of closed subsets of X then we define the *lim sup* :

$$\limsup_n\ B_n =_{\text{def}} \bigcap_n \overline{\bigcup_{k \geq n} B_k} \tag{5}$$

where $\overline{Q}$ denotes the closure of Q. It follows from (b) above that the lim sup of a sequence of nonempty sets is nonempty. It is easy to check that

$$\overline{\bigcup_n B_n} = \left(\bigcup_n B_n\right) \cup \left(\limsup_n\ B_n\right)\,. \tag{6}$$

We want to study a homeomorphism on a space. By compactness this is just a bijective (= one-to-one and onto) continuous function from the space to itself. It will be convenient to use the more general language of relations.

A function $f\colon X \to Y$ is usually described as a rule associating to every point x in X a unique point $y = f(x)$ in Y. In set theory the function f is defined to be the set of ordered pairs $\{(x, f(x))\colon x \in X\}$. Thus, the function f is a subset of the product $X \times Y$. It is what other people call the graph of the function. We will use this language so that, for example, the identity map 1_X on X is the diagonal subset $\{(x, x)\colon x \in X\}$. The notation is extended by defining a *relation from X to Y*, written $F\colon X \to Y$, to be an arbitrary subset of $X \times Y$. Then $F(x) = \{y\colon (x, y) \in F\}$. Thus, a relation is a function exactly when the set $F(x)$ contains a single point for every $x \in X$. In the function case, we will use the symbol $F(x)$ for both the set and the single point it contains, the latter being the usual meaning of $F(x)$.

As they are arbitrary subsets of $X \times Y$ we can perform set operations like union, intersection, closure and interior on relations. In addition, for $F\colon X \to Y$ we define the *inverse* $F^{-1}\colon Y \to X$ by

$$F^{-1} =_{\text{def}} \{(y, x)\colon (x, y) \in F\}\,. \tag{7}$$

If $A \subseteq X$ then its *image* is

$$\begin{aligned} F(A) &=_{\text{def}} \{y\colon (x, y) \in F \text{ for some } x \in A\} \\ &= \bigcup_{x \in A} F(x) = \pi_2((A \times Y) \cap F)\,, \end{aligned} \tag{8}$$

where $\pi_2\colon X \times Y \to Y$ is the projection to the second coordinate.

If $G\colon Y \to Z$ is another relation then the *composition* $G \circ F\colon X \to Z$ is the relation given by

$$\begin{aligned} G \circ F =_{\text{def}} \{(x,z)\colon\ & \text{there exists } y \in Y \\ & \text{such that } (x,y) \in F \text{ and } (y,z) \in G\} \\ &= \pi_{13}((X \times G) \cap (F \times Z))\,, \end{aligned} \tag{9}$$

where $\pi_{13}\colon X \times Y \times Z \to X \times Z$ is the projection map. This generalizes composition of functions and, as with functions, composition is associative. Clearly, $(G \circ F)^{-1} = F^{-1} \circ G^{-1}$.

We call F a *closed relation* when it is a closed subset of $X \times Y$. Clearly, the inverse of a closed relation is closed and by compactness, the composition of closed relations is closed. If A is a closed subset of X and F is a closed relation then the image $F(A)$ is a closed subset of Y. For relations being closed is analogous to being continuous for functions. In fact, a function is continuous if and only if, regarded as a relation, it is closed. This is another application of compactness.

If $Y = X$, so that $F\colon X \to X$, then we call F a *relation on* X. For a positive integer n we define F^n to be the n-fold composition of F with $F^0 =_{\text{def}} 1_X$ and $F^{-n} =_{\text{def}} (F^{-1})^n = (F^n)^{-1}$. This is well-defined because composition is associative. Clearly, $F^m \circ F^n = F^{m+n}$ when m and n have the same sign, i. e. when $mn \geq 0$. On the other hand, the equations $F \circ F^{-1} = F^{-1} \circ F = 1_X = F^0$ all hold if and only if the relation F is a bijective function.

The utility of this relation-speak, once one gets used to it, is that it allows us to extend to this more general situation our intuitions about a function as a way of moving from input here to output there. For example, if $\epsilon \geq 0$ then we can use the metric d on X to define the relations on X

$$\begin{aligned} V_\epsilon &=_{\text{def}} \{(x,y)\colon d(x,y) < \epsilon\}\,, \\ \bar{V}_\epsilon &=_{\text{def}} \{(x,y)\colon d(x,y) \leq \epsilon\}\,. \end{aligned} \tag{10}$$

Thus, $V_\epsilon(x)$ and $\bar{V}_\epsilon(x)$ are the open ball and the closed ball centered at x with radius ϵ. We can think of these relations as ways of moving from a point x to a nearby point.

Each $\bar{V}_\epsilon$ is a closed, symmetric and reflexive relation. The triangle inequality is equivalent to the inclusion $\bar{V}_\epsilon \circ \bar{V}_\delta \subseteq \bar{V}_{\epsilon+\delta}$.

In general, a relation F on X is *reflexive* if $1_X \subseteq F$, *symmetric* if $F^{-1} = F$ and *transitive* if $F \circ F \subseteq F$.

Now we apply this relation notation to a homeomorphism f on X.

For $x \in X$ the *orbit sequence* of x is the bi-infinite sequence $\{\ldots, f^{-2}(x), f^{-1}(x), x, f(x), f^2(x), \ldots\}$. This is just the discrete time analogue of the solution path discussed in the Introduction. We are thinking of it as a sequence and so as a function from the set $\mathbb{Z}$ of discrete times to the state space X with parameter the initial point x.

Now we define the *orbit relation*

$$\mathcal{O}f =_{\text{def}} \bigcup_{n=1}^{\infty} f^n\,. \tag{11}$$

Thus, $\mathcal{O}f(x) = \{f(x), f^2(x), \ldots\}$ is a set, not a sequence, consisting of the states which follow the initial point x in time. Notice that for reasons which will be clear when we consider the cyclic sets below, we begin the union with $n = 1$ rather than $n = 0$ and so the initial point x itself need not be included in $\mathcal{O}f(x)$.

If with think of the point $f(x)$ as the immediate temporal successor of x then $\mathcal{O}f(x)$ is the set of points which occur on the orbit of x at some positive time. This is the first – and simplest – interpretation of the "future" of x with respect to the dynamical system obtained by iterating f.

It is convenient to extend this notion by including the limit points of the positive orbit sequence. For $x \in X$ define

$$\begin{aligned} \omega f(x) &= \limsup_n \{f^n(x)\}\,, \\ \mathcal{R}f(x) &=_{\text{def}} \overline{\mathcal{O}f(x)} = \mathcal{O}f(x) \cup \omega f(x)\,. \end{aligned} \tag{12}$$

Thus, from f we have defined the orbit relation $\mathcal{O}f$ and the *orbit closure relation* $\mathcal{R}f$ with $\mathcal{R}f = \mathcal{O}f \cup \omega f$.

While $\mathcal{R}f(x)$ is closed for each x, the relation $\mathcal{R}f$ itself is usually not closed. As was mentioned above, among relations the closed relations are the analogues of continuous functions. We obtain closed relations by defining

$$\begin{aligned} \Omega f &= \limsup_n f^n\,, \\ \mathcal{N}f(x) &=_{\text{def}} \overline{\mathcal{O}f} = \mathcal{O}f \cup \Omega f\,. \end{aligned} \tag{13}$$

Here we are taking the closure in $X \times X$ and so we obtain closed relations. The relation $\mathcal{N}f$, defined by Auslander et al. [18,19,20] (see also Ura [63,64]), is called the *prolongation* of f. $\mathcal{N}f(x)$ is our next, broader, notion of the "future" of x.

To compare all these suppose that $x, y \in X$. Then $y \in \mathcal{O}f(x)$ when there exists a positive integer n such that $y = f^n(x)$ while $y \in \omega f(x)$ if there is a sequence of positive integers $n_i \to \infty$ such that $f^{n_i}(x) \to y$. On the other hand, $y \in \Omega f(x)$ iff there are sequences $x_i \to x$ and $n_i \to \infty$ such that $f^{n_i}(x_i) \to y$. Thus, $y \in \mathcal{R}f(x)$ if for every $\epsilon > 0$ we can run along the orbit of x and at some positive time make a smaller than ϵ jump to y. Similarly, $y \in \mathcal{N}f(x)$ if for every $\epsilon > 0$ we can make an initial ϵ

small jump to a point x_1, run along the orbit of x_1 and at some positive time make an ϵ small jump to y.

The relations $\mathcal{O}f$ and $\mathcal{R}f$ are transitive but usually not closed. In general, when we pass to the closure, obtaining $\mathcal{N}f$, we lose transitivity. When we consider Lyapunov functions we will see why it is natural to want both of these properties, closure and transitivity.

The intersection of any collection of closed, transitive relations is a closed, transitive relation. Notice that $X \times X$ is such a relation. Thus, we obtain $\mathcal{G}f$, the smallest closed, transitive relation which contains f by intersecting:

$$\mathcal{G}f =_{\text{def}} \bigcap \{Q \subseteq X \times X : \overline{Q} = Q \text{ and } f, Q \circ Q \subseteq Q\} . \tag{14}$$

There is an alternative procedure, due to Conley, see [29], which constructs a closed, transitive relation, generally larger than $\mathcal{G}f$, in a simple and direct fashion.

A *chain* or 0-chain is a finite or infinite sequence $\{x_n\}$ such that $x_{n+1} = f(x_n)$ along the way, i. e. a piece of the orbit sequence. Given $\epsilon \geq 0$ an ϵ-chain is a finite or infinite sequence $\{x_n\}$ such that each x_{n+1} at most ϵ distance away from the point $f(x_n)$, i. e. $x_{n+1} \in \bar{V}_\epsilon(f(x_n))$. If the chain has at least two terms, but only finitely many, then the first and last terms are called the beginning and the end of the chain. The number of terms minus 1 is then called the *length* of the chain. We say that x *chains to* y, written $y \in Cf(x)$ if for every $\epsilon > 0$ there is an ϵ chain which begins at x and ends at y. That is,

$$Cf =_{\text{def}} \bigcap_{\epsilon > 0} \mathcal{O}(\bar{V}_\epsilon \circ f) . \tag{15}$$

Compare this with (12) and (13). If $y \in \mathcal{R}f(x)$ then we can get to y by moving along the orbit of x and then taking an arbitrarily small jump at the end. If $y \in \mathcal{N}f(x)$ then we are allowed a small jump at the beginning as well as the end. Finally, if $y \in Cf(x)$ then we are allowed a small jump at each iterative step.

The chain relation is of great importance for applications. Suppose we are computing the orbit of a point on a computer. At each step there is usually some round-off error. Thus, what we take to be an orbit sequence is in reality an ϵ chain for some positive ϵ. It follows that, in general, what we can expect to compute directly about f is only that level of information which is contained in Cf.

As the intersection of transitive relations, Cf is transitive. It is not hard to show directly that the chain relation Cf is also closed. Since $x_1 = x,\ x_2 = f(x)$ is a 0-chain beginning at x and ending at $f(x)$, we have $f \subseteq Cf$. It follows that $\mathcal{G}f \subseteq Cf$.

This inclusion may be strict. The identity map $f = 1_X$ is already a closed equivalence relation. Hence, $\mathcal{G}1_X = 1_X$. On the other hand, for any $\epsilon > 0$ the relation $\mathcal{O}V_\epsilon$ is an open equivalence relation, and so each equivalence class is clopen (= closed and open). If X is connected then the entire space is a single equivalence class. Since this is true for every positive ϵ we have

$$X \text{ connected} \implies C1_X = X \times X . \tag{16}$$

Since Cf is transitive, the composites $\{(Cf)^n\}$ form a decreasing sequence of closed transitive relations. We denote by ΩCf the intersection of this sequence. That is,

$$\Omega Cf =_{\text{def}} \bigcap_{n=0}^{\infty} (Cf)^n . \tag{17}$$

One can show that $y \in \Omega Cf(x)$ if and only if for every $\epsilon > 0$ and positive integer N there is an ϵ chain of length greater than N which begins at x and ends at y. In addition, the following identity holds (compare (12) and (13)):

$$Cf = \mathcal{O}f \cup \Omega Cf . \tag{18}$$

Thus, built upon f we have a tower of relations:

$$f \subseteq \mathcal{O}f \subseteq \mathcal{R}f \subseteq \mathcal{N}f \subseteq \mathcal{G}f \subseteq Cf . \tag{19}$$

These are the relations which capture the successively broader notions of the "future" of an input x.

A useful identity which holds for $\mathcal{A} = \mathcal{O}, \mathcal{R}, \mathcal{N}, \mathcal{G}$ and C is

$$\mathcal{A}f = f \cup (\mathcal{A}f) \circ f = f \cup f \circ (\mathcal{A}f) . \tag{20}$$

These are easy to check directly for all but $\mathcal{G}$. For that one, observe that $f \cup (\mathcal{G}f) \circ f$ and the other composite are closed and transitive and so each contains $\mathcal{G}f$.

It is also easy to show that for $\mathcal{A} = \mathcal{O}, \mathcal{N}, \mathcal{G}$ and C: $\mathcal{A}(f^{-1}) = (\mathcal{A}f)^{-1}$ and so we can omit the parentheses in these cases. The analogue for $\mathcal{R}$ is usually false and we define

$$\alpha f =_{\text{def}} \omega(f^{-1}) . \tag{21}$$

Thus, $\alpha f(x)$ is the set of limit points of the negative time orbit sequence $\{x, f^{-1}(x), f^{-2}(x), \dots\}$ and it is usually not true that αf equals $(\omega f)^{-1}$.

For $\Theta = \alpha, \omega, \Omega$ and ΩC it is true that

$$\Theta f = f \circ \Theta f = f^{-1} \circ \Theta f = \Theta f \circ f = \Theta f \circ f^{-1} . \tag{22}$$

Now we are ready to consider the variety of recurrence concepts. Recall that a point x is recurrent – in some

sense – if it lies in its own "future". Thus, for any relation F on X we define the *cyclic set*

$$|F| =_{\text{def}} \{x\colon (x,x) \in F\}\,. \tag{23}$$

Clearly, if F is a closed relation then $|F|$ is a closed subset of X.

A point x lies in $|f|$ when $x = f(x)$ and so $|f|$ is the set of *fixed points* for f, while $x \in |\mathcal{O}f|$ when $x = f^n(x)$ for some positive integer n and so $|\mathcal{O}f|$ is the set of *periodic points*. It is easy to check that every periodic point is contained in $|\omega f|$ and so $|\omega f| = |\mathcal{R}f|$. These are called *recurrent points* or sometimes the *positive recurrent points* to distinguish them from $|\alpha f|$, the set of *negative recurrent points*. Similarly, we have $|\Omega f| = |\mathcal{N}f|$, called the set of *nonwandering points*. The points of $|\mathcal{G}f|$ are called *generalized nonwandering* and those of $|Cf| = |\Omega Cf|$ are called *chain recurrent*. The set of periodic points and the sets of recurrent points need not be closed. The rest, associated with closed relations, are closed subsets.

For an illustration of these ideas, observe that for $x, y, z \in X$

$$y, z \in \omega f(x) \implies z \in \Omega f(y)\,. \tag{24}$$

Just hop from y to a nearby point on the orbit of x, moving arbitrarily far along the orbit, you repeatedly arrive nearby z and then can hop to it. In particular, with $y = z$ we see that every point y of $\omega f(x)$ is non-wandering. However, the points of $\omega f(x)$ need not be recurrent. That is, while $y \in \Omega f(y)$ for all $y \in \omega f(x)$ it need not be true that $y \in \omega f(y)$. In particular, $\mathcal{R}f(y)$ can be a proper subset of $\mathcal{N}f(y)$.

On the other hand, it is true that for most points x in X the orbit closure $\mathcal{R}f(x)$ is equal to the prolongation set $\mathcal{N}f(x)$. Recall that a subset of a complete metric space is called *residual* when it is the countable intersection of dense, open subsets. By the Baire Category Theorem a residual subset is dense.

Theorem 1 *If f is a homeomorphism on X then $\{x \in X\colon \omega f(x) = \Omega f(x)\} = \{x \in X\colon \mathcal{R}f(x) = \mathcal{N}f(x)\}$ is a residual subset of X.*

In particular, if every point is nonwandering, i. e. $x \in \Omega f(x)$ for all x, then the set of recurrent points is residual in X. However, if the set of nonwandering points is a proper subset of X, it need not be true that most of these points are recurrent. The closure of $|\omega f|$ can be a proper subset of the closed set $|\Omega f|$.

From recurrence we turn to the notion of invariance.

Let F be a relation on X and A be a closed subset of X. We call A + *invariant* for F if $F(A) \subseteq A$ and *invariant* for F if $F(A) = A$. For a homeomorphism f on X, the set A is invariant for f if and only if it is + invariant for f and for f^{-1}. For example, (20) implies that for $\mathcal{A} = \mathcal{O}, \mathcal{R}, \mathcal{N}, \mathcal{G}$ and C each of the sets $\mathcal{A}f(x)$ is + invariant for f and (22) implies that for $\Theta = \alpha, \omega, \Omega$ and ΩC each of the sets $\Theta f(x)$ is invariant for f. Finally, for $\mathcal{A} = \mathcal{O}, \mathcal{R}, \mathcal{N}, \mathcal{G}$ and C each of the cyclic sets $|\mathcal{A}f|$ is invariant for f.

If A is a nonempty, closed invariant subset for a homeomorphism f then the restriction $f|A$ is a homeomorphism on A and we call this dynamical system the *subsystem* determined by A. In general, if F is a relation on X and A is any subset of X then we call the relation $F \cap (A \times A)$ on A the *restriction* of F to A.

For a homeomorphism f the families of + invariant subsets and of invariant subsets are each closed under the operations of closure and interior and under arbitrary unions and intersections. If A is + invariant then the sequence $\{f^n(A)\}$ is decreasing and the intersection is f invariant. Furthermore, this intersection contains every other f invariant subset of A and so is the maximum invariant subset of A.

If A is + invariant for f then it is + invariant for $\mathcal{O}f$. If, in addition, A is closed then it is + invariant for $\mathcal{R}f$. However, + invariance with respect to the later relations in the tower (19) are successively stronger conditions, and the relations of (19) provide convenient tools for studying these conditions.

We call a closed + invariant subset A a *stable* subset, or a *Lyapunov stable* subset, if it has a neighborhood basis of + invariant neighborhoods. That is, if G is open and $A \subseteq G$ then there exists a + invariant open set U such that $A \subseteq U \subseteq G$.

Theorem 2 *A closed subset A is + invariant for $\mathcal{N}f$ if and only if it is a stable + invariant set for f.*

Proof If G is an open set which contains A and $\mathcal{N}f(A) \subseteq A$ then $U = \{x\colon \mathcal{N}f(x) \subseteq G\}$ is an open set which contains A and which is + invariant by (20). The reverse implication is easy to check directly. □

Invariance with respect to $\mathcal{G}f$ is characterized by using Lyapunov functions, which generalize the strict Lyapunov functions described in the introduction.

For a closed relation F on X, a Lyapunov function L for F is a continuous, real-valued function on X such that

$$(x, y) \in F \implies L(x) \le L(y)\,, \tag{25}$$

(some authors, e. g. Lyapunov, use the reverse inequality).

For any continuous, real-valued function L on X the set $\{(x, y)\colon L(x) \le L(y)\}$ is a closed, transitive relation on X. To say that L is a Lyapunov function for F is exactly to say that this relation contains F. It follows that if

L is a Lyapunov function for a homeomorphism f then it is automatically a Lyapunov function for the closed, transitive relation $\mathcal{G}f$. That L be a Lyapunov function for $\mathcal{C}f$ is usually a stronger condition. For example, any continuous, real-valued function is a Lyapunov function for 1_X, but by (16) if X is connected then constant functions are the only Lyapunov functions for $\mathcal{C}1_X$.

The following result is a dynamic analogue of Urysohn's Lemma in general topology and it has a similar proof.

Theorem 3 *A closed subset A is + invariant for $\mathcal{G}f$ if and only if there exists a Lyapunov function $L\colon X \to [0,1]$ for f such that $A = L^{-1}(1)$.*

A Lyapunov function for a homeomorphism f is non-decreasing on each orbit sequence $x, f(x), f^2(x), \ldots$. Suppose that x is a periodic point, i. e. $x \in |\mathcal{O}f|$. Then $x = f^n(x)$ for some positive integer n, and it follows that L must be constant on the orbit of x.

Now suppose, more generally, that x is a generalized recurrent point, i. e. $x \in |\mathcal{G}f|$. By (20) $(f(x), x) \in \mathcal{G}f$ and so $L(f(x)) = L(x)$ whenever L is a Lyapunov function for f (and hence for $\mathcal{G}f$). Recall that the set $|\mathcal{G}f|$ of generalized recurrent points is f invariant. It follows that $f^n(x) \in |\mathcal{G}f|$ for every integer n and so $L(f^{n+1}(x)) = L(f^n(x))$ for all n. Thus, a Lyapunov function for a homeomorphism f is constant on the orbit of each generalized recurrent point x.

Similarly, if x is a chain recurrent point, i. e. $x \in |\mathcal{C}f|$, then $L(f(x)) = L(x)$ whenever L is a Lyapunov function for $\mathcal{C}f$ and so a Lyapunov function for $\mathcal{C}f$ is constant on the orbit of each chain recurrent point.

Of fundamental importance is the observation that one can construct Lyapunov functions for f (and for $\mathcal{C}f$) which are increasing on all orbit sequences which are not generalized recurrent (respectively, chain recurrent).

Theorem 4 *For a homeomorphism f on a compact metric space X there exist continuous functions $L_1, L_2\colon X \to [0,1]$ such that L_1 is a Lyapunov function for f, and hence for $\mathcal{G}f$, and L_2 is a Lyapunov function for $\mathcal{C}f$ and, in addition,*

$$\begin{aligned} x \in |\mathcal{G}f| &\iff L_1(x) = L_1(f(x)) \,, \\ x \in |\mathcal{C}f| &\iff L_2(x) = L_2(f(x)) \,. \end{aligned} \tag{26}$$

Lyapunov functions which satisfy the conditions of (26) are called *complete Lyapunov functions* for f and for $\mathcal{C}f$, respectively.

If L is a Lyapunov function for f then we define the set of *critical points* for L

$$|L| =_{\text{def}} \big\{x \in X\colon L(x) = L(f(x))\big\} \,. \tag{27}$$

This language is a bit abusive because here criticality describes a relationship between L and f. It does not depend only upon L. However, we adopt this language to compare the general situation with the simpler strict Lyapunov function case in the introduction. Similarly, we call $L(|L|) \subseteq \mathbb{R}$ the set of *critical values* for L. The complementary points of X and $\mathbb{R}$ respectively are called *regular points* and *regular values* for L. Thus, the generalized recurrent points are always critical points for a Lyapunov function L and for a complete Lyapunov function these are the only critical points.

For invariance with respect to $\mathcal{C}f$ we turn to the study of attractors.

Attractors and Chain Recurrence

For a homeomorphism f on X we say that a closed set U is *inward* if $f(U)$ is contained in U°, the interior of U. By compactness this implies that $\bar{V}_\epsilon \circ f(U) \subseteq U$ for some $\epsilon > 0$. That is, U is + invariant for the relation $\bar{V}_\epsilon \circ f$. Hence, any ϵ chain for f which begins in U remains in U. It follows that an inward set for f is $\mathcal{C}f$ + invariant.

For example, assume that L is a Lyapunov function for f. Then for any $s \in \mathbb{R}$ the closed set $U_s = \{x\colon L(x) \geq s\}$ is + invariant for f. Suppose now that s is a regular value for L. This means that for all x such that $L(x) = s$ we have $L(f(x)) > L(x) = s$. On the other hand, for the remaining points x of U_s we have $L(f(x)) \geq L(x) > s$. Thus, $f(U_s)$ is contained in the open set $\{x\colon L(x) > s\} \subseteq U_s$ and so U_s is inward. It easily follows that L is a Lyapunov function for $\mathcal{C}f$ if the set of critical values is nowhere dense.

If U is inward for f then we define $A = \bigcap_{n=0}^{\infty}\{f^n(U)\}$ to be the *attractor* associated with U. Since an inward set U is + invariant for f, the sequence $\{f^n(U)\}$ is decreasing. In fact, for U inward and $n, m \in \mathbb{Z}$ we have

$$n > m \implies f^n(U) \subseteq f^m(U)^\circ \,. \tag{28}$$

The associated attractor is the maximum f invariant subset of U and $\{f^n(U)\colon n \in \mathbb{Z}\}$ is a sequence of inward neighborhoods of A, forming a neighborhood basis for the set A. That is, if G is any open which contains A then by Proposition 1(a), $f^n(U) \subseteq G$ for sufficiently large n.

For example, the entire space X is an inward set and is its own associated attractor. In general, a set A is inward and equal to its own attractor if and only if A is a clopen, f invariant set.

The power of the attractor idea comes from the equivalence of a number of descriptions of different apparent

strength. We use the "weak" ones to test for an attractor and then apply the "strong" conditions. The following collects these alternative descriptions.

Theorem 5 *Let f be a homeomorphism on X and A be a closed f invariant subset of X. The following conditions are equivalent.*

(i) *A is an attractor. That is, there exists an inward set U such that $\bigcap_{n=0}^{\infty} f^n(U) = A$.*
(ii) *There exists a neighborhood G of A such that $\bigcap_{n=0}^{\infty} f^n(G) \subseteq A$.*
(iii) *A is $\mathcal{N}f$ + invariant and the set $\{x\colon \omega f(x) \subseteq A\}$ is a neighborhood of A.*
(iv) *The set $\{x\colon \Omega f(x) \subseteq A\}$ is a neighborhood of A.*
(v) *The set $\{x\colon \Omega Cf(x) \subseteq A\}$ is a neighborhood of A.*
(vi) *A is $\mathcal{G}f$ + invariant and the set $A \cap |\mathcal{G}f|$ is clopen in the relative topology of the closed set $|\mathcal{G}f|$.*
(vii) *A is Cf + invariant and the set $A \cap |Cf|$ is clopen in the relative topology of the closed set $|Cf|$.*

Applying Theorem 2 to condition (iii) of Theorem 5 we see that if A is a stable set for f and, in addition, the orbit of every point in some neighborhood of x tends asymptotically toward A then A is an attractor. The latter condition alone does not suffice, although the strengthening in (iv) is sufficient. For example, the homeomorphism of $[0,1]$ defined by $t \mapsto t^2$ has $\{0\}$ as an attractor. If we identify the two fixed points $0, 1 \in [0,1]$ by mapping t to $z = e^{2\pi i t}$ then we obtain a homeomorphism f on the unit circle X. For every $z \in X$, we have $\omega f(z) = \{1\}$, the unique fixed point. However, $\{1\}$ is not an attractor for f. In fact, $\Omega f(1) = X$.

The class of attractors is closed under finite union and finite intersection. Using infinite intersections we can characterize Cf invariance.

Theorem 6 *Let f be a homeomorphism on X and A be a closed f invariant subset of X. The following conditions are equivalent. When they hold we call A a quasi-attractor for f.*

(i) *A is Cf + invariant.*
(ii) *A is the intersection of a (possibly infinite) set of attractors.*
(iii) *The set of inward neighborhoods of A form a basis for the neighborhood system of A.*

From Theorem 2 again it follows that a quasi-attractor is stable.

If A is a closed invariant set for a homeomorphism f then A is called an *isolated invariant set* if it is the maximum invariant subset of some neighborhood U of A. That is,

$$A = \bigcap_{n=-\infty}^{+\infty} f^n(U)\,. \tag{29}$$

In that case, U is called an *isolating neighborhood* for A. Notice that if U is a closed isolating neighborhood for A and the positive orbit of x remains in U, i. e. $f^n(x) \in U$ for $n = 0, 1, \ldots$ then $\omega f(x) \subseteq A$ because $\omega f(x)$ is an invariant subset of U.

Since an attractor is the maximum invariant subset of some inward set, it follows that an attractor is isolated. Conversely, by condition (iii) of Theorem 5 an invariant set is an attractor precisely when it is isolated and stable. In particular, a quasi-attractor is an attractor exactly when it is an isolated invariant set.

An attractor for f^{-1} is called a *repellor* for f. If U is an inward set for f then $\overline{X \setminus U} = X \setminus (U^\circ)$ is an inward set for f^{-1} and $B = \bigcap_{n=0}^{\infty} f^{-n}(\overline{X \setminus U})$ is the associated repellor for f. We call B the repellor *dual* to the attractor $A = \bigcap_{n=0}^{\infty} f^n(U)$. Recall that an f invariant set is f^{-1} invariant. In particular, attractors and repellors are both f and f^{-1} invariant. The open set $\bigcup_{n=0}^{\infty} f^{-n}(U) = X \setminus B$ is called the *domain of attraction* for A. The name comes from the implication:

$$\begin{aligned} x \in X \setminus B &\implies \omega(x) \subseteq \Omega Cf(x) \subseteq A\,, \\ x \in X \setminus A &\implies \alpha(x) \subseteq \Omega Cf^{-1}(x) \subseteq B\,. \end{aligned} \tag{30}$$

For example, the entire space X is both an attractor and a repellor with dual $\emptyset$.

If A is an attractor then we call the set of chain recurrent points in A, i. e. $A \cap |Cf|$, the *trace* of the attractor A. An attractor is determined by its trace via the equation

$$Cf(A \cap |Cf|) = A\,. \tag{31}$$

The trace of an attractor is a clopen subset of $|Cf|$ by part (vii) of Theorem 5 Conversely, suppose A_0 is a subset of $|Cf|$ which is + invariant for the restriction of Cf to $|Cf|$. That is, $x \in A_0$ and $y \in Cf(x) \cap |Cf|$ implies $y \in A_0$. If A_0 is clopen in $|Cf|$, then $Cf(A_0)$ is the attractor with trace A_0 and $Cf^{-1}(|Cf| \setminus A_0)$ is the dual repellor. By Theorem 6 if A_0 is merely closed then $Cf(A_0)$ is a quasi-attractor.

With the relative topology the set $|Cf| = |Cf^{-1}|$ of chain recurrent points is a compact metric space and so has only countably many clopen subsets (Every clopen subset is a finite union of members of a countable basis for the topology). It follows that, while there are often

uncountably many inward sets, there are only countably many attractors.

When restricted to $|Cf| = |Cf^{-1}|$ the closed relation $Cf \cap Cf^{-1}$ is reflexive as well as symmetric and transitive. The individual equivalence classes are closed f invariant subsets of X called the *chain components* of f, or the *basic sets* of f. These chain components are the analogues of the individual critical points in gradient case described in the introduction.

Any two points of a chain component are related by Cf. This is a type of transitivity condition. As with recurrence, there are several – increasingly broad – notions of dynamic transitivity and these can be associated with the relations of (19). First, we consider when the entire system f on X is transitive in the some way. Then we say that a closed f invariant subset A is a transitive subset in this way, when the subsystem $f|A$ on A is transitive in the appropriate way.

A group action on a set is called transitive when one can move from any element of the set to any other by some element of the group. That is, the entire set is a single orbit of the group action. Notice that this use of the word is unrelated to transitivity of a relation. Recall that a relation F on a set X is transitive when $F \circ F \subseteq F$. We are now considering when it happens that any two points of X are related by F. This just says that $F = X \times X$ which we will call the *total relation* on X.

First, what does it mean to say that $\mathcal{O}f$ is total, that is to say all of X lies in a single orbit of f? First, compactness implies that X is then finite and so the homeomorphism f is a permutation of the finite set X. Such permutation is a product of disjoint cycles and $\mathcal{O}f = X \times X$ exactly when all of X consists of a single cycle. The associated invariant subsets are the periodic orbits of f, including the fixed points.

Next, we consider when the orbit closure relation $\mathcal{R}f$ is total, that is, when every point is in the orbit closure of every other.

Theorem 7 *Let f be a homeomorphism on X. The following conditions are equivalent and when they hold we call f minimal.*

(i) For all $x \in X$, $\mathcal{O}f(x)$ is dense in X.
(ii) For all $x \in X$, $\mathcal{R}f(x) = X$, i. e. $\mathcal{R}f = X \times X$.
(iii) For all $x \in X$, $\omega f(x) = X$, i. e. $\omega f = X \times X$.
(iv) X is the only nonempty, closed f + invariant subset of X.
(v) X is the only nonempty, closed f invariant subset of X.

Recall our convention that the state space of a dynamical system is nonempty, although we do allow the empty set as an invariant subset. For example, the empty set is the repellor/attractor dual to the attractor/repellor which is the entire space.

Thus, a closed f invariant subset A of X is *minimal* when it is nonempty but contains no nonempty, proper f invariant subset. From compactness it follows via the usual Zorn's Lemma argument that every nonempty, closed f + invariant subset of X contains a minimal, nonempty, closed f invariant subset.

Theorem 8 *Let f be a homeomorphism on X. The following conditions are equivalent and when they hold we call f topologically transitive.*

(i) For some $x \in X$, $\mathcal{O}f(x)$ is dense in X, i. e. $\mathcal{R}f(x) = X$.
(ii) For all $x \in X$, $\mathcal{N}f(x) = X$, i. e. $\mathcal{N}f = X \times X$.
(iii) For all $x \in X$, $\Omega f(x) = X$, i. e. $\Omega f = X \times X$.
(iv) X is the only closed f + invariant subset with a nonempty interior.

If f is topologically transitive then the set $\{x : \omega f(x) = X\}$ is residual, i. e. it is the countable intersection of dense, open subsets of X.

Every point in a minimal system has a dense orbit. If f is topologically transitive then the set of *transitive points* for f,

$$Trans_f =_{\text{def}} \{x : \omega f(x) = X\} = \{x : \mathcal{R}f(x) = X\}, \quad (32)$$

is dense by the Theorem 8. However, its complement is either empty, which is the minimal case, or else is dense as well. Also, there is a usually rich variety of invariant subsets. It may happen, for instance, that the set of periodic points is dense. Most well-studied examples of chaotic dynamical systems are non-minimal, topologically transitive systems. In fact, Devaney used the conjunction of topological transitivity and density of periodic points in an infinite system as a definition of chaos.

The broadest notion of transitivity is associated with the chain relation Cf.

Theorem 9 *Let f be a homeomorphism on X. The following conditions are equivalent and when they hold we call f chain transitive.*

(i) For all $x \in X$, $Cf(x) = X$, i. e. $Cf = X \times X$.
(ii) For all $x \in X$, $\Omega Cf(x) = X$, i. e. $\Omega Cf = X \times X$.
(iii) X is the only nonempty inward set.
(iv) X is the only nonempty attractor.

Before proceeding further, we pause to observe that each of these three concepts is the same for f and for its inverse.

Because the f invariant subsets are the same as the f^{-1} invariant subsets, we see that f^{-1} is minimal when f is. Since $\mathcal{N}(f^{-1}) = (\mathcal{N}f)^{-1}$ and $C(f^{-1}) = (Cf)^{-1}$, it follows as well that f^{-1} is topologically transitive or chain transitive when f satisfies the corresponding property.

There are many naturally occurring chain transitive subsets. For every $x \in X$ the limit sets $\alpha f(x)$ and $\omega f(x)$ are chain transitive subsets as are each of the chain components. Notice that if x and y are points of some chain component B then by definition of the equivalence relation $Cf \cap Cf^{-1}$, x and y can be connected by ϵ chains in X for any positive ϵ. To say that B is a chain transitive subset is to make the stronger statement that they can be connected via ϵ chains which remain in B.

Contrast this positive result with the trap set by implication (24) which suggests that $B = \omega f(x)$ is a topologically transitive subset. However, (24), which says $B \times B \subseteq \mathcal{N}f$, does not imply that the restriction $f|B$ is topologically transitive, i. e. $B \times B = \mathcal{N}(f|B)$. It may not be possible to get from near y to near z without a hop which takes you outside B.

In fact, if f is any chain transitive homeomorphism on a space X then it is possible to embed X as a closed subset of a space Y and extend f to a homeomorphism g on Y in such a way that $X = \omega g(y)$ for some point $y \in Y$.

Any chain transitive subset for f is contained in a unique chain component of $|Cf|$. In fact the chain components are precisely the maximal chain transitive subsets. In particular, each subset $\alpha f(x)$ and $\omega f(x)$ is contained in some chain component of f.

By using (16) one can show that each connected component of the closed subset $|Cf|$ is contained in some chain component as well. It follows that the space of chain components, that is, the space of $Cf \cap Cf^{-1}$ equivalence classes with the quotient topology from $|Cf|$, is zero-dimensional. So this space is either countable or is the union of a Cantor set with a countable set.

An individual chain component B is called *isolated* when it is a clopen subset of the chain recurrent set $|Cf|$ or, equivalently, if it is an isolated point in the space of chain components. Thus, B is isolated when it has a neighborhood U in X such that $U \cap |Cf| = B$. It can be proved that B is an isolated chain component precisely when it is an isolated invariant set, i. e. it admits a neighborhood U satisfying (29).

The individual chain components generalize the role played by the critical points in the gradient case. They are the blobs we described at the end of the introduction. We complete the analogy by identifying which chain components are like the relative maxima and minima among the critical points.

Theorem 10 *Let B be a nonempty, closed, f invariant set for a homeomorphism f on X. The following conditions are equivalent and when they hold we call B a terminal chain component.*

(i) B is a chain transitive quasi-attractor.
(ii) B is a Cf + invariant, chain transitive subset of X.
(iii) B is a Cf + invariant chain component.
(iv) In the set of nonempty, closed Cf + invariant subsets of X the set B is an element which is minimal with respect to inclusion.

In particular, B is a chain transitive attractor if and only if it is an isolated, terminal chain component.

From condition (iv) and Zorn's Lemma it follows that every nonempty, closed Cf + invariant subset of X contains a terminal chain component. In particular, $\Omega Cf(x)$ contains a terminal chain component for every $x \in X$.

Clearly, if there are only finitely many chain components then each terminal chain component is isolated and so is an attractor.

Corollary 1 *Let f be a homeomorphism on X and let B be a nonempty, closed, Cf + invariant subset of X. If for some $x \in X$, $B \subseteq \omega f(x)$ then B is a terminal chain component and $B = \omega f(x)$.*

Proof B contains some terminal chain component B_1 and $\omega f(x)$ is contained in some chain component B_2. Thus, $B_1 \subseteq B \subseteq \omega f(x) \subseteq B_2$. Since distinct chain components are disjoint, $B_1 = B_2$. □

Because $C(f^{-1}) = (Cf)^{-1}$, the chain components for f and f^{-1} are the same. A chain component is called an *initial chain component* for f when it is a terminal chain component for f^{-1}.

Let L be a Lyapunov function for Cf. L is constant on each chain component and Theorem 4 says that L can be constructed to be strictly increasing on the orbit of every point of $X \setminus |Cf|$. That is, for such a complete Lyapunov function the set $|L|$ of critical points equals $|Cf|$. In that case, the local maxima occur at terminal chain components and, as a partial converse, an isolated terminal chain component is a local maximum for any complete Cf Lyapunov function.

Theorem 11 *For a homeomorphism f on X let L be a Cf Lyapunov function and let B be a chain component with r the constant value of L on B.*

(a) If there exists a neighborhood U of B such that $L(x) < r$ for all $x \in U \setminus B$, then B is a terminal chain component.

(b) *Assume L is a complete Lyapunov function. If B is an attractor then it is a terminal chain component and the domain of attraction U is an open set containing B such that $L(x) < r$ for all $x \in U \setminus B$.*

In particular, if there are only finitely many chain components for f then by using L, a complete Cf Lyapunov function which distinguishes the chain components, we obtain the picture promised at the end of the introduction. The local maxima of L occur at the terminal chain components which are attractors. The local minima are at the initial chain components which are repellors. The remaining chain components play the role of saddles. As in the gradient case, it can happen that there is an open set of points x such that $\omega f(x)$ is contained in one of these saddle chain components. There is a natural topological condition which, when it holds, excludes this pathology.

Theorem 12 *For a homeomorphism f on X, assume that $\mathcal{N}f = Cf$ or, equivalently, that $\Omega f = \Omega Cf$.*

If A is a stable closed f invariant subset of X then A is a quasi-attractor.

For a residual set of points x in X, $\omega f(x)$ is a terminal chain component and $\alpha f(x)$ is an initial chain component.

Proof The result for a stable set A follows from the characterizations in Theorems 2 and 6

By Theorem 1 the set of x such that $\omega f(x) = \Omega f(x)$ is always residual. By assumption this agrees with the set of x such that $\omega f(x) = \Omega Cf(x)$. Corollary 1 implies that for such x the set $\omega f(x)$ is a terminal chain component. For $\alpha f(x)$ apply this result to f^{-1}. □

If X is a compact manifold of dimension at least two, then the condition $\mathcal{N}f = Cf$ holds for a residual set in the Polish group of homeomorphisms on X with the uniform topology. If, in addition, f has only finitely many chain components (and this is not a residual condition on f) then it follows that the points which are in the domain of attraction of some terminal chain component and in the domain of repulsion of some initial chain component form a dense, open subset of X.

Chaos and Equicontinuity

Any chain transitive system can occur as a chain component, but in the most interesting cases the chain components are topologically transitive subsets. In this section we consider the antithetical phenomena of equicontinuity and chaos in topologically transitive systems.

If f is a homeomorphism on X and $x \in X$ then x is called a *transitive point* when its orbit $\mathcal{O}f(x)$ is dense in X. As in (32) we denote by $Trans_f$ the set of transitive points for f. The system is minimal exactly when every point is transitive, i. e. when $Trans_f = X$.

To study equicontinuity we introduce a new metric d_f defined using the original metric d on X.

$$d_f(x, y) =_{\text{def}} \sup \{d(f^n(x), f^n(y)) \colon n = 0, 1, 2, \dots\} . \tag{33}$$

It is easy to check that d_f is a metric, i. e. it satisfies the conditions of positivity and symmetry as well as the triangle inequality. However, it is usually not topologically equivalent to the original metric d, generating a topology which is usually strictly finer (= more open sets). We call a point $x \in X$ an *equicontinuity point* for f when for every $\epsilon > 0$ there exists a $\delta > 0$ so that for all $y \in X$

$$d(x, y) < \delta \implies d_f(x, y) \leq \epsilon , \tag{34}$$

or, equivalently, if for every $\epsilon > 0$ there exists a neighborhood U of x with d_f diameter at most ϵ (the terms "neighborhood", "open set", etc. will always refer to the original topology given by d). Here the d_f diameter of a subset $A \subseteq X$ is

$$\text{diam}_{\text{f}}(\text{A}) =_{\text{def}} \sup \{d_f(x, y) \colon x, y \in A\} . \tag{35}$$

We denote the, possibly empty, set of equicontinuity points for f by Eq_f.

When every point is an equicontinuity point, the system associated with f is called *equicontinuous*, or we just say that f is equicontinuous. This coincides with the concept of equicontinuity of the set of functions $\{f^n \colon n = 1, 2, \dots\}$. Recall that any finite set of continuous functions is equicontinuous, but equicontinuity for an infinite set is often, as in this case, a strong condition. Equicontinuity of f says exactly that the metrics d and d_f are topologically equivalent. For compact spaces topologically equivalent metrics are uniformly equivalent. Hence, if f is equicontinuous, then for every $\epsilon > 0$ there exists a $\delta > 0$ such that for all $x, y \in X$ implication (34) holds.

When f is equicontinuous then we can replace d by d_f since in the equicontinuous case the latter is a metric giving the correct topology. The homeomorphism f is then an *isometry* of the metric, i. e. $d_f(x, y) = d_f(f(x), f(y))$ for all $x, y \in X$. This equality uses a famous little problem which keeps being rediscovered: If f is a surjective map on a compact metric space X with metric d such that $d(f(x), f(y)) \leq d(x, y)$ for every $x, y \in X$ then f is an isometry on X, see e. g. Alexopoulos [9] or Akin [3], Proposition 2.4(c).

Conversely, if f is an isometry of a metric d on X (with the correct topology) then f is clearly equicontinuous.

When X has a dense set of equicontinuity points we call the system *almost equicontinuous*. If f is almost equicontinuous but not equicontinuous then the δ in (34) will depend upon $x \in Eq_f$ as well as upon ϵ.

On the other hand, we say that the system *has sensitive dependence upon initial conditions*, or simply that f is *sensitive* when there exists $\epsilon > 0$ such that $\mathrm{diam}_f(U) > \epsilon$ for every nonempty open subset U of X, or, equivalently, if there exists $\epsilon > 0$ such that for any $x \in X$ and $\delta > 0$, there exists $y \in X$ such that $d(x, y) < \delta$ but for some $n > 0 d(f^n(x), f^n(y)) > \epsilon$. Here the important issue is that ϵ is independent of the choice of x and δ.

Sensitivity is a popular candidate for a definition of chaos in the topological context. Suppose for a sensitive homeomorphism f you are attempting to estimate analyze the orbit of x, but may make an – arbitrarily small – positive error for initial point input. Even if you are able to compute the iterates exactly, you still cannot be certain that you always remain ϵ close to the orbit you want. If you have chosen a bad point y as input then at some time n you will be at $f^n(y)$ more than distance ϵ away from the point $f^n(x)$ that you want.

For transitive systems, the *Auslander–Yorke Dichotomy Theorem* holds (see [21]):

Theorem 13 *Let f be a topologically transitive homeomorphism on a compact metric space X. Exactly one of the following alternatives is true.*

- *(Sensitivity) The homeomorphism f is sensitive and there are no equicontinuity points, i. e. $Eq_f = \emptyset$.*
- *(Almost Equicontinuity) The set of equicontinuity points coincides with the set of transitive points, i. e. $Eq_f = Trans_f$, and so the set of equicontinuity points is residual in X.*

Proof For $\epsilon > 0$ let $Eq_{f,\epsilon}$ be the union of all open sets with d_f diameter at most ϵ. So $x \in Eq_{f,\epsilon}$ if and only if it has a neighborhood with d_f diameter at most ϵ. It is then easy to check that $x \in Eq_{f,\epsilon}$ if and only if $f(x) \in Eq_{f,\epsilon}$. Thus, $Eq_{f,\epsilon}$ is an f invariant open set. If $Eq_{f,\epsilon}$ is nonempty and x is a transitive point then the orbit eventually enters $Eq_{f,\epsilon}$. Since $f^n(x) \in Eq_{f,\epsilon}$ for some positive n, it follows that $x \in Eq_{f,\epsilon}$ because the set is invariant. This shows that

$$Eq_{f,\epsilon} \neq \emptyset \implies Trans_f \subseteq Eq_{f,\epsilon} . \tag{36}$$

If for every $\epsilon > 0$ we have $Eq_{f,\epsilon} \neq \emptyset$ then $Trans_f \subseteq \bigcap_{\epsilon>0} Eq_{f,\epsilon} = Eq_f$. We omit the proof that only the transitive points can be equicontinuous in a topologically transitive system.

If, instead, for some $\epsilon > 0$ the set $Eq_{f,\epsilon}$ is empty then by definition f is sensitive. □

The system is minimal if and only if $Trans_f = X$ and so a topologically transitive system is equicontinuous if and only if it is both almost equicontinuous and minimal. In particular, we have:

Corollary 2 *Let f be a minimal homeomorphism on X. Either f is sensitive or f is equicontinuous.*

Banks et al. [22] showed that a topologically transitive homeomorphism on an infinite space with dense periodic points is always sensitive. On the other hand, there do exist topologically transitive homeomorphisms which are almost equicontinuous but not minimal and hence not equicontinuous. If f is an almost equicontinuous, topologically transitive homeomorphism then there is a sequence of positive integers $n_k \to \infty$ such that the sequence $\{f^{n_k}\}$ converges uniformly to the identity 1_X, see Glasner and Weiss [37] and Akin, Auslander and Berg [6]. In general, when such a convergent sequence of iterates exists the homeomorphism f is called *uniformly rigid*.

Other notions of chaos are defined using ideas related to proximality. A pair $(x, y) \in X \times X$ is called *proximal* for f if

$$\lim\ \inf_{n\to\infty} d(f^n(x), f^n(y)) = 0 , \tag{37}$$

or, equivalently, if $1_X \cap \omega(f \times f)(x, y) \neq \emptyset$ where $f \times f$ is the homeomorphism on $X \times X$ defined by $(f \times f)(x, y) = (f(x), f(y))$. If a pair is not proximal then it is called *distal*. The pair is called *asymptotic* if

$$\lim_{n\to\infty} d(f^n(x), f^n(y)) = 0 . \tag{38}$$

We will call (x, y) a *Li–Yorke pair* if it is proximal but not asymptotic. That is, (37) holds but

$$\lim\ sup_{n\to\infty} d(f^n(x), f^n(y)) > 0 . \tag{39}$$

Li and Yorke [47] called a subset A of X a *scrambled set* if every non-diagonal pair in $A \times A$ is a Li–Yorke pair. Following their definition f is called *Li–Yorke chaotic* when there is an uncountable scrambled subset.

The homeomorphism f is called *distal* when every nondiagonal pair in $X \times X$ is a distal pair. If f is an isometry then clearly f is distal and so every equicontinuous homeomorphism is distal. There exist homeomorphisms f which are minimal and distal but not equicontinuous, as described in e. g. Auslander [17]. Since such an f is minimal but not equicontinuous, it is sensitive. On the other hand, a distal homeomorphism has no proximal pairs and so is certainly not Li–Yorke chaotic.

There exists an almost equicontinuous, topologically transitive, non-minimal homeomorphism f such that a fixed point $e \in X$ is the unique minimal subset of X and so the pair (e, e) is the unique subset in $X \times X$ which is minimal for $f \times f$. It follows that $(e, e) \in \omega(f \times f)(x, y)$ for every pair $(x, y) \in X \times X$ and so every pair is proximal. On the other hand, because f is uniformly rigid, every pair (x, y) is recurrent for $f \times f$ and it follows that no non-diagonal pair is asymptotic. Thus, the entire space X is scrambled and f is Li–Yorke chaotic but not sensitive.

There is a sharpening of topological transitivity which always implies sensitivity as well as most other topological conditions associated with chaos. A homeomorphism f on X is called *weak mixing* if the homeomorphism $f \times f$ on $X \times X$ is topologically transitive. If (x, y) is a transitive point for $f \times f$ then since $\omega(f \times f)(x, y) = X \times X$ it follows that $d_f(x, y) = M$, where M is the diameter of the entire space X. Suppose f is weak mixing and U is any nonempty open subset of X. Since $Trans_{f \times f}$ is dense in $X \times X$, there is a transitive pair (x, y) in $U \times U$. Hence, $diam_f(U) = M$. When f is weak mixing, there exists an uncountable $A \subseteq X$ such that every nondiagonal pair in $A \times A$ is a transitive point for $f \times f$, Iwanik [44] and Huang and Ye [42], see also Akin [5]. Since such a set A is clearly scrambled, it follows that f is Li–Yorke chaotic.

For any two subsets $U, V \subseteq X$ we define the *hitting time set* to be the set integers given by:

$$N(U, V) =_{\text{def}} \{n \geq 0 \colon f^n(U) \cap V \neq \emptyset\} . \tag{40}$$

The topological transitivity condition $\Omega f = X \times X$ says that whenever U and V are nonempty open subsets the hitting time set $N(U, V)$ is infinite. We call f *mixing* if for every pair of nonempty open subsets U, V there exists k such that $n \in N(U, V)$ for all $n \geq k$. As the names suggest, mixing implies weak mixing.

Example The most important example of a chaotic homeomorphism is the ubiquitous *shift homeomorphism*. We think a of fixed finite set A as an *alphabet* and for any positive integer k the sequences in A of length k, i. e. the elements of A^k, are called *words of length k*. When A is equipped with the discrete topology it is compact and so by the Tychonoff Product Theorem any product of infinitely many copies of A is compact when given the product topology. Using the group of integers as our index set, we let $X = A^{\mathbb{Z}}$. There is a metric compatible with this topology. For $x, y \in X$ let

$$d(x, y) =_{\text{def}} \text{infimum } \{2^{-n} \colon x_i = y_i \text{ for all } i \text{ with } |i| < n\} . \tag{41}$$

The metric d is an *ultrametric*. That is, it satisfies a strengthening of the triangle inequality. For all $x, y, z \in X$:

$$d(x, z) \leq \max(d(x, y), d(y, z)) . \tag{42}$$

The ultrametric inequality is equivalent to the condition that for every positive ϵ the open relation V_ϵ is an equivalence relation.

For any word $a \in A^k$ and any integer j, the set $\{x \in X \colon x_{i+j} = a_i \colon i = 1, \ldots, k\}$ is a clopen subset of X called a *cylinder set*, which we will denote $U_{a,j}$. For example, with $k = 2n - 1$ and $j = -n$ the cylinder sets are precisely the open balls of radius ϵ when $2^{-n} < \epsilon \leq 2^{-n+1}$. From this we see that X is a Cantor set with cylinder sets as a countable basis of clopen sets.

On X the shift homeomorphism s is defined by the equation

$$s(x)_i = x_{i+1} \text{ for all } i \in \mathbb{Z} . \tag{43}$$

The fixed points $|s|$ are exactly the constant sequences in X and the periodic points $|\mathcal{O}s|$ are the periodic sequences in X. That is, $s^t(x) = x$ for some positive integer t exactly when $x_{i+t} = x_i$ for all $i \in \mathbb{Z}$.

It is easy to see that s is mixing. For example, if $a, b \in A^{2n-1}$ and $j = -n$, then the hitting time set $N(U_{a,j}, U_{b,j})$ contains every integer t larger than $2n$. For if $x_{i-n} = a_i$ and $x_{i-n+t} = b_i$ for $i = 1, \ldots, 2n - 1$ then $x \in U_{a,j}$ and $s^t(x) \in U_{b,j}$. This illustrates why s is chaotic in the sense of unpredictable. Given $x \in X$, if $y \in X$ satisfies $d(x, y) = 2^{-n}$, then moving right from position 0 the first n entries of y are known exactly. They agree with the entries of x. But after position n, x provides no information about y. The remaining entries on the right can be chosen arbitrarily.

Since s is mixing it is certainly topologically transitive, but it is useful to characterize the transitive points of s. For any point $x \in X$ and positive integer n we can scan from the central position x_0, left and right $n - 1$ steps and observe a word of length $2n - 1$. As we apply s the central position shifts right. When we have applied s^t it has shifted t steps. Scanning left and right we observe a new word. A point is a transitive point when for every n we can observe every word in A^{2n-1} in this way by varying t. To construct a transitive point x we need only list the – countably many – finite words of every length and lay them out end to end to get the right side of x. The left side of x can be arbitrary.

The shift homeomorphism is also expansive. In general, a homeomorphism f on a compact metric space X is called *expansive* when the diagonal 1_X is an isolated

invariant set for the homeomorphism $f \times f$ on $X \times X$ and so there exists $\epsilon > 0$ such that $\bar{V}_\epsilon$ is an isolating neighborhood in the sense of (29). That is, if $x, y \in X$ and $d(f^n(x), f^n(y)) \leq \epsilon$ for all $n \in \mathbb{Z}$ then $x = y$. Such a number $\epsilon > 0$ is called an *expansivity constant*. With respect to the metric given by (41) it is easy to check that $\frac{1}{2}$ is an expansivity constant for the shift homeomorphism s.

The shift is interesting in its own right. In addition, the chaotic behavior of other important examples, especially expansive homeomorphisms, are studied by comparing them with the shift.

For a compact metric space X let $H(X)$ denote the automorphism group of X, i. e. the group of homeomorphisms on X, with the topology of uniform convergence. The metric on $H(X)$ is defined by the equation, for $f, g \in H(X)$:

$$d(f, g) =_{\text{def}} \text{supremum } \{d(f(x), g(x)) \colon x \in X\} . \quad (44)$$

We obtain a topologically equivalent, but complete, metric by using $\max(d(f, g), d(f^{-1}.g^{-1}))$. The automorphism group is a Polish (= admits a complete, separable metric) topological group. For $f \in H(X)$ we define the *translation* homeomorphisms ℓ_f and r_f on $H(X)$ by

$$\ell_f(g) =_{\text{def}} f \circ g \qquad \text{and} \qquad r_f(g) =_{\text{def}} g \circ f . \quad (45)$$

For any $x \in X$, the *evaluation map* $ev_x \colon H(X) \to X$ is the continuous map defined by $ev_x(f) = f(x)$.

For any $f \in H(X)$ let G_f denote the closure in $H(X)$ of the cyclic subgroup generated by f. That is,

$$G_f = \overline{\{f^n \colon n \in \mathbb{Z}\}} \subseteq H(X) . \quad (46)$$

Thus, G_f is a closed, abelian subgroup of $H(X)$.

Let $Iso(X)$ denote the closed subgroup of isometries in $H(X)$ (In contrast with $H(X)$ this varies with the choice of metric). It follows from the Arzela–Ascoli Theorem that $Iso(X)$ is compact in $H(X)$. If f is an isometry on X then G_f is a compact, abelian subgroup of $Iso(X)$. Furthermore,the translation homeomorphisms ℓ_f and r_f restrict to isometries on the compact space $Iso(X)$ and G_f is a closed invariant subset. In fact, under either ℓ_f or r_f, G_f is just the orbit closure of f regarded as a point of $Iso(X)$.

Recall that if f is an equicontinuous homeomorphism, then we can replace the original metric by a topologically equivalent one, e. g. replace d by d_f, to get one for which f is an isometry.

If f is a homeomorphism on X and g is a homeomorphism on Y then we say that a continuous function $\pi \colon Y \to X$ *maps* g to f when $\pi \circ g = f \circ \pi \colon Y \to X$. If, in addition, π is a homeomorphism then we call π an *isomorphism* from g to f.

Theorem 14 *Let f be a homeomorphism on X. Fix $x \in X$. The evaluation map $ev_x \colon H(X) \to X$ maps ℓ_f on $H(X)$ to f on X.*

If f is an almost equicontinuous, topologically transitive homeomorphism then ev_x restricts to a homeomorphism from the closed subgroup G_f of $H(X)$ onto $Trans_f$ the residual subset of X consisting of the transitive points for f.

If f is a minimal isometry then ev_x restricts to a homeomorphism from the compact subgroup G_f of $H(X)$ onto X and so is an isomorphism from ℓ_f on G_f to f on X.

The isometry result is classical, see e. g. Gottschalk and Hedlund [38]. It shows that minimal, equicontinuous homeomorphisms are just translations on compact, monothetic groups, where a topological group is *monothetic* when it has a dense cyclic subgroup. Similarly a topologically transitive, almost equicontinuous homeomorphism which is not minimal is obtained by "compactifying" a translation on a noncompact monothetic group, see Akin and Glasner [7]. In fact, the group cannot even be locally compact and so, for example, is not the discrete group of integers. While examples of such topologically transitive, almost equicontinuous but not equicontinuous systems are known, it is not known whether there are any finite dimensional examples. And it is known that they cannot occur on a zero-dimensional space, i. e. the Cantor set. If finite dimensional examples do not exist then every topologically transitive homeomorphism on a compact manifold is sensitive except for the equicontinuous ones which we now describe.

Example We can identify the circle S with the quotient topological group $\mathbb{R}/\mathbb{Z}$. For $a \in \mathbb{R}$ the translation $L_a = R_a$ on $\mathbb{R}$ induces the rotation τ_a on the circle $\mathbb{R}/\mathbb{Z}$. If $a \in \mathbb{Z}$ then this is the identity map. If a is rational then τ_a is periodic. But if a is irrational then τ_a is a minimal isometry on S. More generally, if $\{1, a_1, \ldots, a_n\}$ is linearly independent over the field of rationals $\mathbb{Q}$ then on the torus $X = S^n$ the product homeomorphism $\tau_{a_1} \times \cdots \times \tau_{a_n}$ is a minimal isometry. Such systems are sometimes called *quasi-periodic*.

Of course, the translation by 1 on the finite cyclic group $\mathbb{Z}/k\mathbb{Z}$ of integers modulo k is just a version of a single periodic orbit, the unique minimal map on a finite space of cardinality k.

Recall that if $\{X_1, X_2, \ldots\}$ is a sequence of topological spaces and $\{p_n \colon X_{n+1} \to X_n\}$ is a sequence of continuous maps, then the *inverse limit* is the closed subset LIM of the product space $\Pi_{n=1}^{\infty} X_n$:

$$LIM = \lim \{X_n, p_n\} =_{\text{def}} \{x \colon p_n(x_{n+1}) = x_n \text{ for } n = 1, 2, \ldots\} . \quad (47)$$

If the spaces are compact and the maps are surjective then the *nth* coordinate projection π_n maps the compact space *LIM* onto X_n for every n (Hint: use Proposition 1). Also, if the spaces are topological groups and the maps are homomorphisms, then *LIM* is a closed subgroup of the product topological group. Finally, if $\{f_n : X_n \to X_n\}$ is a sequence of homeomorphisms such that p_n maps f_{n+1} to f_n for every n then the product homeomorphism $\Pi_{n=1}^{\infty} f_n$ restricts to a homeomorphism f on *LIM* and π_n maps f to f_n for every n.

Now let $\{k_n\}$ be an increasing sequence of positive integers such that k_n divides k_{n+1} for every n. Let X_n denote the finite cyclic group $\mathbb{Z}/k_n\mathbb{Z}$ and let $p_n : X_{n+1} \to X_n$ be the quotient homomorphism induced by the inclusion map $k_{n+1}\mathbb{Z} \to k_n\mathbb{Z}$. Let f_n denote the translation by 1 on X_n. Define X to be the inverse limit of this system with f the restriction to X of the product homeomorphism. X is a topological group whose underlying space is zero-dimensional and perfect, i. e. the Cantor set. The product homeomorphism is an isometry when we use the metric analogous to the one defined by (41). Furthermore, the restriction f to X is a minimal isometry. These systems are usually called *adding machines* or *odometers*. It can be proved that every equicontinuous minimal homeomorphism on a Cantor space is isomorphic to one of these. In general, every equicontinuous minimal homeomorphism is isomorphic to (1) a periodic orbit, (2) an adding machine, (3) a quasi-periodic motion on a torus, or (4) a product with each factor either an irrational rotation on a circle, an adding machine or a periodic orbit. The informal expression *strange attractor* is perhaps best defined as a topologically transitive attractor which is not equicontinuous, i. e. which is not one of these.

The word "chaos" suggests instability and unpredictability. However, for many examples what is most apparent is stability. For the Henon attractor or the Lorenz attractor the word "the" is used because in simulations one begins with virtually any initial point, performs the iterations and, after discarding an initial segment, one observes a particular, repeatable picture. The set as a whole is a predictable feature, stable under perturbation of initial point, an varying continuously with the defining parameters.

However, once the orbit is close enough to the attractor, essentially moving within the attractor itself, the motion is unpredictable, sensitive to arbitrarily small perturbations. What remains is statistical prediction. We cannot exactly predict when the orbit will enter some small subset of the attractor, but we can describe approximately the amount of time it spends in the subset. Such analysis requires an invariant measure and this takes us to the boundary between topological and measurable dynamics.

By a measure μ on a compact metric space X we will mean a Borel probability measure on the space. Such a measure acts, via integration, on the Banach algebra $C(X)$ of continuous real-valued functions on X. The set $\mathcal{P}(X)$ of such measures can thus be regarded as a convex subset of the dual space of $C(X)$. Inheriting the weak* topology on the dual space, $\mathcal{P}(X)$ becomes a compact, metrizable space. There is a natural inclusion map $\delta : X \to \mathcal{P}(X)$ which associates to $x \in X$ the point mass at x, denoted δ_x.

The *support* of a measure μ on X is the smallest closed set with measure 1. Denoted $|\mu|$ its complement can be obtained by taking the union out of a countable base for the topology of those members with measure 0. The measure has *full support* (or simply μ is *full*) when $|\mu| = X$. A measure is full when every nonempty open set has positive measure.

A continuous map $h : X \to Y$ induces a map $h_* : \mathcal{P}(X) \to \mathcal{P}(Y)$ which associates to μ the measure $h_*\mu$ defined by $h_*\mu(A) = \mu(h^{-1}(A))$ for every Borel subset A of Y. The continuous linear operator h_* is the dual of $h^* : C(Y) \to C(X)$ given by $u \mapsto u \circ h$. Furthermore, h_* is an extension of h. That is, $h_*(\delta_x) = \delta_{h(x)}$. The supports are related by

$$h(|\mu|) = |h_*\mu| \,. \tag{48}$$

In particular, if f is a homeomorphism on X then f_* is a linear homeomorphism on $\mathcal{P}(X)$. extending f on X. A measure μ such that $f_*\mu = \mu$ is called an *invariant measure* for f. Thus, $|f_*|$, the set of fixed points for f_*, is the set of invariant measures for f. Clearly, $|f_*|$ is compact, convex subset of $\mathcal{P}(X)$. The classical theorem of Krylov and Bogolubov says that this set is nonempty. To prove it, one begins with an arbitrary point $x \in X$ and considers the sequence of Cesaro averages

$$\sigma_n(f, x) =_{\text{def}} \frac{1}{n+1} \sum_{i=0}^{n} \delta_{f^n(x)} \,. \tag{49}$$

It can be shown that every measure in the nonempty set of limit points of this sequence lies in $|f_*|$. If the set of limit points consists of a single measure μ, i. e. the sequence converges to μ, then x is called a *convergence point* for the invariant measure μ. For $\mu \in |f_*|$ we denote by $\mathrm{Con}(\mu)$ the – possibly empty – set of convergence points for μ.

The extreme points of the compact convex set $|f_*|$ are called the *ergodic* measures for f. That is, μ is ergodic if it is not in the interior of some line segment connecting a pair of distinct invariant measures. Equivalently, μ is ergodic if a Borel subset A of X is invariant, i. e. $f^{-1}(A) = A$, only when the measure $\mu(A)$ is either 0 or 1. It follows that if

$u\colon X \to \mathbb{R}$ is a Borel measurable, invariant function, then for any ergodic measure μ there is a set of measure 1 on which u is constant. Or more simply, $u \circ f = u$ implies u is constant almost everywhere with respect to μ.

The central result in measurable dynamics is the Birkhoff Pointwise Ergodic Theorem which says, in this context:

Theorem 15 *Given a homeomorphism f on a compact metric space X, let $u\colon X \to \mathbb{R}$ be a bounded, Borel measurable function. There exists a bounded, invariant, Borel measurable function. $\hat{u}\colon X \to \mathbb{R}$ such that for every f invariant measure μ*

$$\int u \, \mathrm{d}\mu = \int \hat{u} \, \mathrm{d}\mu \quad \text{and} \lim_{n\to\infty} \frac{1}{n+1} \sum_{i=0}^{n} u(f^n(x)) = \hat{u}(x) \tag{50}$$

almost everywhere with respect to μ.

In particular, if μ is ergodic then

$$\hat{u}(x) = \int u \, \mathrm{d}\mu \tag{51}$$

almost everywhere with respect to μ.

It is a consequence of the ergodic theorem that

$$\mu \text{ ergodic} \implies \mu(\mathrm{Con}(\mu)) = 1 \,. \tag{52}$$

Thus, with respect to an ergodic measure μ for almost every point x,

$$\lim_{n\to\infty} \frac{1}{n+1} \sum_{i=0}^{n} u(f^n(x)) = \int u \, \mathrm{d}\mu \,, \tag{53}$$

for every $u \in C(X)$. The left hand side is the time-average of the function u along the orbit with initial point x and it equals the space-average on the right.

For any f invariant measure μ it follows from (48) that the support, $|\mu|$, is an f invariant subset of X. The Poincaré Recurrence Theorem says that if U is any open set with $U \cap |\mu| \neq \emptyset$ then U is non-wandering. That is, for some positive integer n, $U \cap f^{-n}(U) \neq \emptyset$. This is clear from the observation that the open sets $f^{-n}(U)$ all have the same, positive measure and so they cannot all be pairwise disjoint. Applying this to the subsystem obtained by restricting to $|\mu|$ it follows from Theorem 1 that the set of recurrent points in $|\mu|$ form a dense G_δ subset of $|\mu|$.

If $x \in \mathrm{Con}(\mu)$ then the orbit of x is dense in $|\mu|$ and so the subsystem on $|\mu|$ is topologically transitive whenever $\mathrm{Con}(\mu)$ is nonempty. By the Birkhoff Ergodic Theorem this applies whenever μ is ergodic.

Since the support is a nonempty, closed, invariant subspace it follows that if f is minimal then every invariant measure has full support. The homeomorphism f is called *strictly ergodic* when it is minimal and has a unique invariant measure μ, which is necessarily ergodic. In that case, it can be shown that every point is a convergence point for μ, that is, $\mathrm{Con}(\mu) = X$.

We note that in the dynamical systems context the topological notion of residual (that is, a dense G_δ subset) is quite different from the measure theoretic idea (a set of full measure). For example,

$$Con =_{\mathrm{def}} \bigcup_{\mu \in |f_*|} \mathrm{Con}(\mu) \tag{54}$$

is the set of points whose associated Cesaro average sequence is a Cauchy sequence. It follows that *Con* is a Borel set. By (52) $\mu(Con) = 1$ for every ergodic measure μ. Since every invariant measure is a limit of convex combinations of ergodic measures it follows that *Con* has measure 1 for every invariant measure μ.

On the other hand, it can be shown that if f is the shift homeomorphism on $X = A^{\mathbb{Z}}$ then for x in the dense G_δ subset of X the set of limit points of the sequence $\{\sigma_n(f, x)\}$ is all of $|f_*|$, Denker et al. [30] or Akin (Chap. 9 in [1]). Since the shift has many different invariant measures it follows that *Con* is disjoint from this residual subset. In general, if a homeomorphism f has more than one full, ergodic measure then *Con* is of first category, Akin [1, Theorem 8.11]. In particular, if f is minimal but not strictly ergodic then *Con* is of first category.

The plethora of invariant measures undercuts somewhat their utility for statistical analysis. Suppose that there are two different ergodic measures μ and ν with common support, some invariant subset A of X. By restricting to A, we can reduce to the case when $A = X$ and so μ and ν are distinct full, ergodic measures. The sets $\mathrm{Con}(\mu)$ and $\mathrm{Con}(\nu)$ are disjoint and of measure 1 with respect to μ and ν, respectively. For an open set U the average frequency with which an orbit of $x \in \mathrm{Con}(\mu)$ lies is U is given by $\mu(U)$ and similarly for ν. These mutually singular measures lead to different statistics.

One solution to this difficulty is to select what seems to be the best invariant measure in some sense, e. g. the measure of maximum entropy (see the article by King) if it should happen to be unique. However, as our introductory discussion illustrates, this somewhat misses the point.

Return to the case of a chaotic attractor A, a closed invariant subset of X. It often happens that the state space X comes equipped with a natural measure λ or at least a Radon–Nikodym equivalence class of measures, all with

the same sets of measure 0. For example, if X is a manifold then λ is locally Lebesgue measure. The measure λ is usually not f invariant and it often happens that the set A of interest has λ measure 0. What we want, an *appropriate* measure μ for this situation, would be an invariant measure with support A, i. e.

$$\mu \in |f_*| \text{ and } |\mu| = A\,, \tag{55}$$

and an open set U containing A such that with respect to λ almost every point of U is a convergence point for μ. That is,

$$\lambda(U \setminus \mathrm{Con}(\mu)) = 0\,. \tag{56}$$

Notice that for $x \in \mathrm{Con}(\mu)$, $\omega f(x) = |\mu| = A$ and so by Corollary 1 A is a terminal chain component. That is, for such a measure to exist, the attractor A must be at least chain transitive. If, in addition, $\mathrm{Con}(\mu) \cap A \neq \emptyset$ then A is topologically transitive.

At least when strong hyperbolicity conditions hold this program can be carried out with μ the Bowen measure for the invariant set, see Katok–Hasselblatt (Chapter 20 in [45]).

Minimality and Multiple Recurrence

In this section we provide a sketch of some important topics which were neglected in the above exposition. We first consider the study of minimal systems.

In Theorem 7 we defined a homeomorphism f on a compact space X to be *minimal* when every point has a dense orbit. A subset A of X is a minimal subset when it is a nonempty, closed, invariant subset such that the restriction $f|A$ defines a minimal homeomorphism on A. The term "minimal" is used because $f|A$ is minimal precisely when A is minimal, with respect to inclusion, in the family of nonempty, closed, invariant subsets. By Zorn's lemma every such subset contains a minimal subset. Since every system contains minimal systems, the classification of minimal systems provides a foundation upon which to build an understanding of dynamical systems in general.

On the other hand, if you start with the space, homeomorphisms which are minimal – on the whole space – are rather hard to construct. Some spaces have the fixed-point property, i. e. every homeomorphism on the space has a fixed point, and so admit no minimal homeomorphisms. The tori, which are monothetic groups, admit the equicontinuous minimal homeomorphisms described in the previous section. Fathi and Herman [34] constructed a minimal homeomorphism on the 3-sphere. For most other connected, compact manifolds it is not known whether they admit minimal homeomorphisms or not. It is even difficult to construct topologically transitive homeomorphisms, but for these a beautiful – but nonconstructive – argument due to Oxtoby shows that every such manifold admits topologically transitive homeomorphisms. He uses that Baire Category Theorem to show that if the dimension is at least two then the topologically transitive homeomorphisms are residual in the class of volume preserving homeomorphisms, see Chap. 18 in [53].

Since we understand the equicontinuous minimal systems, we begin by building upon them. This requires a change in out point of view. Up to now we have mostly considered the behavior of a single dynamical system (X,f) consisting of a homeomorphism f on a compact metric space X. Regarding these as our objects of study we turn now to the maps between such systems. A *homomorphism* of dynamical systems, also called an *action map*, $\pi : (X, f) \to (Y, g)$ is a continuous map $\pi : X \to Y$ such that $g \circ \pi = \pi \circ f$ and so, inductively, $g^n \circ \pi = \pi \circ f^n$ for all $n \in \mathbb{Z}$. Thus, π maps the orbit of a point $x \in X$ to the orbit of $\pi(x) \in Y$. In general, for the map $\pi \times \pi : X \times X \to Y \times Y$ we have that

$$\begin{aligned} \pi \times \pi(\mathcal{A}f) &\subseteq \mathcal{A}g \quad \text{for } \mathcal{A} = \mathcal{O}, \mathcal{R}, \mathcal{N}, \mathcal{G}, C \\ \pi \times \pi(\Theta f) &\subseteq \Theta g \quad \text{for } \Theta = \omega, \alpha, \Omega, \Omega C\,. \end{aligned} \tag{57}$$

That is, π maps the various dynamic relations associated with f to the corresponding relations for g.

If π is bijective then it is called an *isomorphism* between the two systems and the inverse map π^{-1} is an action map, continuous by compactness.

If X is a closed invariant subset of Y with $f = g|X$ then the inclusion map π is an action map and then (X,f) is called the *subsystem* of (Y,g) determined by the invariant set X.

On the other hand, if π is surjective then (Y,g) is called a *factor*, or *quotient system*, of (X,f) and (X,f) is called a *lift* of (Y,g). A surjective action map is called a *factor map*.

For a factor map $\pi : (X, f) \to (Y, g)$ we define an important subset $R(\pi)$ of $X \times X$:

$$\begin{aligned} R(\pi) &=_{\mathrm{def}} \{(x_1, x_2) \in X \times X : \pi(x_1) = \pi(x_2)\} \\ &= (\pi \times \pi)^{-1}(1_Y)\,. \end{aligned} \tag{58}$$

The subset $R(\pi)$ is an ICER on X. That is, it is an invariant, closed equivalence relation. In general, if R is any ICER on X then on the space of equivalence classes the homeomorphism f induces a homeomorphism and the natural quotient map is a factor map of dynamical systems. The original factor system (Y,g) is isomorphic to the quotient system obtained from the ICER $R(\pi)$.

Notice that if (Y, g) is the *trivial system*, meaning that Y consisting of a single point, then $R(\pi)$ is the total relation $X \times X$ on X.

We use the ICER $R(\pi)$ to extend various definitions from dynamical systems to action maps between dynamical systems. For example, a factor map $\pi : (X, f) \to (Y, g)$ is called *equicontinuous* if for every $\epsilon > 0$ there exists $\delta > 0$ such that

$$(x_1, x_2) \in R(\pi) \text{ and } d(x_1, x_2) < \delta \implies d_f(x_1, x_2) < \epsilon \,. \quad (59)$$

Comparing this with (34) we see that (X, f) is equicontinuous if and only if the factor map to the trivial system is equicontinuous.

Similarly, recall that $(x_1, x_2) \in X$ is a *distal pair* for (X, f) if $\omega(f \times f)(x_1, x_2)$ is disjoint from the diagonal 1_X, and the system (X, f) is *distal* when every nondiagonal pair is distal. A factor map $\pi : (X, f) \to (Y, g)$ is distal when every nondiagonal pair in $R(\pi)$ is distal. Again (X, f) is a distal system if and only if the factor map to the trivial system is distal.

It is easy to check that a distal lift of a distal system is distal. Since an equicontinuous factor map is distal, it follows that an equicontinuous lift of an equicontinuous system is distal. However, it need not be equicontinuous.

Example If a is irrational then the rotation τ_a on the circle $Y = \mathbb{R}/\mathbb{Z}$, defined by $x \mapsto a + x$, is an equicontinuous minimal map. On the torus $X = \mathbb{R}/\mathbb{Z} \times \mathbb{R}/\mathbb{Z}$ we define f by

$$f(x, y) =_{\text{def}} (a + x, x + y) \,. \quad (60)$$

It can be shown that (X, f) is minimal. The projection π to the first coordinate defines an equicontinuous factor map to $Y, \tau_a)$ and so (X, f) is distal. It is not, however, equicontinuous.

The above projection map is an example of a *group extension*. Let G be a compact topological group, like $\mathbb{R}/\mathbb{Z}$. Given any dynamical system (Y, g) and any continuous map $q : Y \to G$ we let $X = Y \times G$ and define f on X by:

$$f(x, y) =_{\text{def}} (g(x), L_{q(x)}(y)) \,. \quad (61)$$

The homeomorphism commutes with $1_Y \times R_z$ for any group element z and from this it easily follows that the projection π to the first coordinate is an equicontinuous factor map. If (Y, g) is minimal then the restriction of π to any minimal subset of X defines an equicontinuous factor map from the associated minimal subsystem. It can be shown that any equicontinuous factor map between minimal systems can be obtained via a factor from such a group extension.

The Furstenberg Structure Theorem says that any distal minimal system can be obtained by a – possibly transfinite – inverse limit construction, beginning with an equicontinuous system and such that each lift is an equicontinuous factor map. For the details of this and for the structure theorem due to Veech for more general minimal systems, we refer the reader to Auslander (Chaps. 7, 14 in [17]) respectively.

The factor maps described above are projections from products. There exist examples which are not product projections even locally. For example, in Auslander (Chap. 1 in [17]) the author uses a construction due to Floyd to build an action map π between minimal systems which is not an isomorphism but which is *almost one-to-one*. That is, for a residual set of points x in the domain the set of preimages $\pi^{-1}(\pi(x))$ is a singleton. Such a factor map is the opposite of distal. It is a *proximal* map, meaning that every pair $(x_1, x_2) \in R(\pi)$ is a proximal pair.

Also, one cannot base all one's constructions upon equicontinuous minimal systems. A dynamical system (X, f) is called *weak mixing* when the product $(X \times X, f \times f)$ is topologically transitive. Inclusion (57) implies that if a dynamical system is minimal, topologically transitive or chain transitive then any factor satisfies the corresponding property. It follows that any factor of a weak mixing system is weak mixing. It is clear that only the trivial system is both weak mixing and equicontinuous. Hence for a weak mixing system the trivial system is the only equicontinuous factor. For a minimal system the converse is true: if the trivial system is the only equicontinuous factor then the system is weak mixing. Furthermore, nontrivial weak mixing, minimal systems do exist.

Gottschalk and Hedlund in [38] introduced the idea using various special families of subsets of $\mathbb{N}$, the set of nonnegative integers, in order to distinguish different sorts of recurrence. A *family* $\mathcal{F}$ is a collection of subsets of $\mathbb{N}$ which is hereditary upwards. That is, if $A \subseteq B$ and $A \in \mathcal{F}$ then $B \in \mathcal{F}$. The family is called *proper* when it is a proper subset of the entire power set of $\mathbb{N}$, i. e. it is neither empty nor the entire power set. The heredity property implies that a family $\mathcal{F}$ is proper iff $\mathbb{N} \in \mathcal{F}$ and $\emptyset \notin \mathcal{F}$. For a family $\mathcal{F}$ the dual family, denoted $\mathcal{F}^*$ (or sometimes $k\mathcal{F}$) is defined by:

$$\begin{aligned} \mathcal{F}^* &=_{\text{def}} \{B \subseteq \mathbb{N} : B \cap A \neq \emptyset \text{ for all } A \in \mathcal{F}\} \\ &= \{B \subseteq \mathbb{N} : \mathbb{N} \setminus B \notin \mathcal{F}\} \,. \end{aligned} \quad (62)$$

It is easy to check that $\mathcal{F}^{**} = \mathcal{F}$ and that $\mathcal{F}^*$ is proper if and only if $\mathcal{F}$ is.

A *filter* is a proper family which is closed under pairwise intersection. A family $\mathcal{F}$ is the dual of a filter when it satisfies what Furstenberg calls the *Ramsey Property*:

$$A \cup B \in \mathcal{F} \implies A \in \mathcal{F} \text{ or } B \in \mathcal{F} \,. \tag{63}$$

For example, a set is in the dual of the family of infinite sets if and only if it is cofinite, i. e. its complement is finite. The family of cofinite sets is a filter.

A subset $A \subseteq \mathbb{N}$ is called *thick* when it contains arbitrarily long runs. That is, for every positive integer L there exists i such that $i, i+1, \ldots, i+L \in A$. Dual to the thick sets are the syndetic sets. A subset B is called *syndetic* or *relatively dense* if there exists a positive integer L such that every run of length L meets B.

In (40) we defined the hitting time set $N(U,V)$ for (X,f) when $U, V \subseteq X$. When U is a singleton $\{x\}$ we omit the braces and so have

$$N(x,V) = \{n \geq 0 \colon f^n(x) \in V\} \,. \tag{64}$$

It is clear that (X,f) is topologically transitive when $N(U,V)$ is nonempty for every pair of nonempty open sets U, V. Furstenberg showed that the stronger property that (X,f) be weak mixing is characterized by the condition that each such $N(U,V)$ is thick.

Recall that a point $x \in X$ is recurrent when $x \in \omega f(x)$. The point is called a *minimal point* when it is an element of a minimal subset of X in which case this minimal subset is $\omega f(x)$. Clearly, point $x \in X$ is recurrent when $N(x,U)$ is nonempty for every neighborhood U of x. Gottschalk and Hedlund proved that x is a minimal point if and only if every such $N(x,U)$ is relatively dense. For this reason, minimal points are also called *almost periodic* points.

Furstenberg reversed this procedure by using dynamical systems arguments to derive properties about families of sets and more generally to derive results in combinatorial number theory.

If $f_1, \ldots, f_k$ are homeomorphisms on a space X then $x \in X$ is a *multiple recurrent point* for $f_1, \ldots, f_k$ if there exists a sequence of positive integers $n_i \to \infty$ such that the k sequences $\{f_1^{n_i}(x)\}, \ldots, \{f_k^{n_i}(x)\}$ all have limit x. That is, the point $(x, \ldots, x)$ is a recurrent point for the homeomorphism $f_1 \times \cdots \times f_k$ on X^k. The Furstenberg Multiple Recurrence Theorem says:

Theorem 16 *If $f_1, \ldots, f_k$ are commuting homeomorphisms on a compact metric space X, i. e. $f_i \circ f_j = f_j \circ f_i$ for $i, j = 1, \ldots, k$, then there exists a multiple recurrent point for $f_1, \ldots, f_k$.*

Corollary 3 *If f is a homeomorphism on a compact metric space X and $x \in X$ then for every positive integer k and every $\epsilon > 0$ there exist positive integers m, n such that with $y = f^m(x)$ the distance between any two of the points $y, f^n(y), f^{2n}(y), \ldots, f^{kn}(y)$ is less than ϵ.*

Proof This follows easily by applying the Multiple Recurrence Theorem to the restrictions of the homeomorphisms $f, f^2, \ldots, f^k$ to the closed invariant subset $\omega f(x)$. Obtain a multiple recurrent point $y' \in \omega f(x)$ and then given $\epsilon > 0$ choose $y = f^m(x)$ to approximate y' sufficiently closely. □

These results are of great interest in themselves and they have been extended in various directions, see Bergelson–Leibman [24] for example. In addition, Furstenberg used the corollary to obtain a proof of Van der Waerden's Theorem:

Theorem 17 *If $B_1, \ldots, B_p$ is a partition of $\mathbb{N}$ then at least one of these sets contains arithmetic progressions of arbitrary length.*

Proof It suffices to show that for each $k = 1, 2, \ldots$ and some $a = 1, \ldots, p$ the subset B_a contains an arithmetic progression of length $k+1$. For this will then apply to some fixed B_a for infinitely many k. Let $A = \{1, \ldots, p\}$ and on $X = A^{\mathbb{Z}}$ define the shift homeomorphism defined by (43). Using the metric given by (41) and $\epsilon \leq \frac{1}{2}$ we observe that if $x, y \in X$ with $d(x,y) < \epsilon$ then $x_0 = y_0$.

Choose $x \in X$ such that $x_i = a$ if and only if $i \in B_a$ for $i \in \mathbb{N}$. Apply Corollary 3 to x and ϵ. Choosing positive integers m, n such that the points $f^m(x), f^{m+n}(x), f^{m+2n}(x), \ldots, f^{m+kn}(x)$ all lie within ϵ of each other we have that $m, m+n, \ldots, m+kn$ all lie in B_a where a is the common value of the 0 coordinate of these points. □

One of the great triumphs of modern dynamical systems theory is Furstenberg's use of the ergodic theory version of these arguments to prove Szemerédi's Theorem [35, Theorem 3.21].

Theorem 18 *Let B be a subset of $\mathbb{N}$ with positive upper Banach density, that is*

$$\limsup_{|I| \to \infty} |B \cap I| / |I| > 0 \,, \tag{65}$$

where I varies over bounded subintervals of $\mathbb{N}$ and $|A|$ denotes the cardinality of $A \subset \mathbb{N}$. The set B contains arithmetic progressions of arbitrary length.

For details and further results, we refer to Furstenberg's beautiful book [35].

Future Directions

As far as applied work is concerned, the resolution of the statistical problems, described at the end in the Sect. "Chaos and Equicontinuity", concerning chaotic attractors would be most useful. In addition, it would be nice to know which compact manifolds admit minimal homeomorphisms.

Cross References

► Entropy in Ergodic Theory

Bibliography

The approach to topological dynamics which was described in the initial sections is presented in detail in Akin [1] and [2]. For the deeper, more specialized work in topological dynamics see Auslander [17], Ellis [33] and Akin [4]. A general survey of the field is given in de Vries [65]. Furstenberg [35] is a classic illustration of the applicability of topological dynamics to other fields.

Clark Robinson's text [55] is an excellent introduction to dynamical systems in general. More elementary introductions are Devaney [31] and Alligood et al. [10]. Hirsch et al. [40] provides a nice transition from differential equations to the general theory.

Much of modern differentiable dynamics grows out of the work of Smale and his students, see especially the classic Smale [60], included in the collection [61]. The seminal paper Shub and Smale [58] provides a bridge between this work and the purely topological aspects of attractor theory, see also Shub [57].

The fashionable topic of chaos has generated a large sample of writing whose quality exhibits extremely high variance. For expository surveys I recommend Lorenz [48] and Stewart [62]. An excellent collection of relatively readable, classic papers is Hunt et al. [43].

For applications in biology see Hofbauer and Sigmund [41] and May [49]. Also, don't miss Sigmund's delightful book [59].

1. Akin E (1993) The General Topology of Dynamical Systems. Am Math Soc
2. Akin E (1996) Dynamical systems: the topological foundations. In: Aulbach B, Colonius F (eds) Six Lectures on Dynamical Systems. World Scientific, Singapore, pp 1–43
3. Akin E (1996) On chain continuity. Discret Contin Dyn Syst 2:111–120
4. Akin E (1997) Recurrence in Topological Dynamics: Furstenberg Families and Ellis Actions. Plenum Press, New York
5. Akin E (2004) Lectures on Cantor and Mycielski sets for dynamical systems. In: Assani I (ed) Chapel Hill Ergodic Theory Workshops. Am Math Soc, Providence, pp 21–80
6. Akin E, Auslander J, Berg K (1996) When is a transitive map chaotic? In: Bergelson V, March K, Rosenblatt J (eds) Conference in Ergodic Theory and Probability. Walter de Gruyter, Berlin, pp 25–40
7. Akin E, Glasner E (2001) Residual properties and almost equicontinuity. J d'Analyse Math 84:243–286
8. Akin E, Hurley M, Kennedy JA (2003) Dynamics of Topologically Generic Homeomorphisms. Memoir, vol 783. Am Math Soc, Providence
9. Alexopoulos J (1991) Contraction mappings in a compact metric space. Solution No.6611. Math Assoc Am Mon 98:450
10. Alligood KT, Sauer TD, Yorke JA (1996) Chaos: An Introduction to Dynamical Systems. Spring Science Business Media, New York
11. Almgren FJ (1966) Plateau's Problem. Benjamin WA, Inc., New York
12. Alpern SR, Prasad VS (2001) Typical dynamics of volume preserving homeomorphisms. Cambridge University Press, Cambridge
13. Arhangel'skii AV, Pontryagin LS (eds) (1990) General topology I. Springer, Berlin
14. Arhangel'skii AV (ed) (1995) General Topology III. Springer, Berlin
15. Arhangel'skii AV (ed) (1996) General Topology II. Springer, Berlin
16. Auslander J (1964) Generalized recurrence in dynamical systems. In: Contributions to Differential Equations, vol 3. John Wiley, New York, pp 55–74
17. Auslander J (1988) Minimal Flows and Their Extensions. North-Holland, Amsterdam
18. Auslander J, Bhatia N, Siebert P (1964) Attractors in dynamical systems. Bol Soc Mat Mex 9:55–66
19. Auslander J, Siebert P (1963) Prolongations and generalized Liapunov functions. In: International Symposium on Nonlinear Differential Equations and Nonlinear Mechanics. Academic Press, New York, pp 454–462
20. Auslander J, Siebert P (1964) Prolongations and stability in dynamical systems. Ann Inst Fourier 14(2):237–268
21. Auslander J, Yorke J (1980) Interval maps, factors of maps and chaos. Tohoku Math J 32:177–188
22. Banks J, Brooks J, Cairns G, Davis G, Stacey P (1992) On Devaney's Definition of Chaos. Am Math Mon 99:332–333
23. Barrow-Green J (1997) Poincaré and the Three Body Problem. American Mathematical Society, Providence
24. Bergelson V, Leibman A (1996) Polynomial extensions of Van der Waerden's and Szemeredi's theorems. J Am Math Soc 9:725–753
25. Bing RH (1988) The Collected Papers of R. H. Bing. American Mathematical Society, Providence
26. Brown M (ed) (1991) Continuum Theory and Dynamical Systems. American Mathematical Society, Providence
27. Bhatia N, Szego G (1970) Stability Theory of Dynamical Systems. Springer, Berlin
28. Coddington E, Levinson N (1955) Theory of Ordinary Differential Equations. McGraw-Hill, New York
29. Conley C (1978) Isolated Invariant Sets and the Morse Index. American Mathematical Society, Providence
30. Denker M, Grillenberger C, Sigmund K (1978) Ergodic Theory on Compact Spaces. Lect. Notes in Math, vol 527. Springer, Berlin
31. Devaney RL (1989) An Introduction to Chaotic Dynamical Sys-

tems, 2nd edn. Addison-Wesley Publishing Company, Redwood City
32. Douglas J (1939) Minimal surfaces of higher topological structure. Ann Math 40:205–298
33. Ellis R (1969) Lectures on Topological Dynamics. WA Benjamin Inc, New York
34. Fathi A, Herman M (1977) Existence de difféomorphismes minimaux. Soc Math de France, Astérique 49:37–59
35. Furstenberg H (1981) Recurrence in Ergodic Theory and Combinatorial Number Theory. Princeton Univ. Press, Princeton
36. Glasner E (2003) Ergodic Theory Via Joinings. American Mathematical Society, Providence
37. Glasner E, Weiss B (1993) Sensitive dependence on initial conditions. Nonlinearity 6:1067–1075
38. Gottschalk W, Hedlund G (1955) Topological Dynamics. American Mathematical Society, Providence
39. Hartman P (1964) Ordinary Differential Equations. John Wiley, New York
40. Hirsch MW, Smale S, Devaney RL (2004) Differential Equations, Dynamical Systems and an Introduction to Chaos, 2nd edn. Elsevier Academic Press, Amsterdam
41. Hofbauer J, Sigmund K (1988) The Theory of Evolution and Dynamical Systems. Cambridge Univ. Press, Cambridge
42. Huang W, Ye X (2002) Devaney's chaos or 2-scattering implies Li–Yorke's chaos. Topol Appl 117:259–272
43. Hunt B, Kennedy JA, Li T-Y, Nusse H (eds) (2004) The Theory of Chaotic Attractors. Springer, Berlin
44. Iwanik A (1989) Independent sets of transitive points. In: Dynamical Systems and Ergodic Theory. Banach Cent Publ 23:277–282
45. Katok A, Hasselblatt B (1995) Introduction to the Modern Theory of Dynamical Systems. Cambridge Univ. Press, Cambridge
46. Kuratowski K (1973) Applications of the Baire-category method to the problem of independent sets. Fundam Math 81:65–72
47. Li TY, Yorke J (1975) Period three implies chaos. Am Math Mon 82:985–992
48. Lorenz E (1993) The Essence of Chaos. University of Washington Press, Seattle
49. May RM (1973) Stability and Complexity in Model Ecosystems. Princeton Univ. Press, Princeton
50. Murdock J (1991) Perturbations. John Wiley and Sons, New York
51. Mycielski J (1964) Independent sets in topological algebras. Fundam Math 55:139–147
52. Nemytskii V, Stepanov V (1960) Qualitative Theory of Differential Equations. Princeton U Press, Princeton
53. Oxtoby J (1980) Measure and Category, 2nd edn. Springer, Berlin
54. Peterson K (1983) Ergodic Theory. Cambridge Univ. Press, Cambridge
55. Robinson C (1995) Dynamical Systems: Stability, Symbolic Dynamics and Chaos. CRC Press, Boca Raton
56. Rudolph D (1990) Fundamentals of Measurable Dynamics: Ergodic Theory on Lebesgue Spaces. Oxford Univ. Press, Oxford
57. Shub M (1987) Global Stability of Dynamical Systems. Springer, New York
58. Shub M, Smale S (1972) Beyond hyperbolicity. Ann Math 96:587–591
59. Sigmund K (1993) Games of Life: Explorations in Ecology, Evolution and Behavior. Oxford Univ. Press, Oxford
60. Smale S (1967) Differentiable dynamical systems. Bull Am Math Soc 73:747–817
61. Smale S (1980) The Mathematics of Time. Springer, Berlin
62. Stewart I (1989) Does God Play Dice? Basil Blackwell, Oxford
63. Ura T (1953) Sur les courbes défines par les équations différentielles dans l'espace à m dimensions. Ann Sci Ecole Norm Sup [3] 70:287–360
64. Ura T (1959) Sur le courant extérieur à une région invariante. Funk Ehv 2:143–200
65. de Vries J (1993) Elements of Topological Dynamics. Kluwer, Dordecht

Topological Dynamics of Cellular Automata

Petr Kůrka[1,2]
[1] Département d'Informatique, Université de Nice Sophia Antipolis, Nice, France
[2] Center for Theoretical Study, Academy of Sciences and Charles University, Prague, Czechia

Article Outline

Glossary
Definition of the Subject
Introduction
Topological Dynamics
Symbolic Dynamics
Equicontinuity
Surjectivity
Permutive and Closing Cellular Automata
Expansive Cellular Automata
Attractors
Subshifts and Entropy
Examples
Future Directions
Acknowledgments
Bibliography

Glossary

Almost equicontinuous CA has an equicontinuous configuration.

Attractor omega-limit of a clopen invariant set.

Blocking word interrupts information flow.

Closing CA distinct asymptotic configurations have distinct images.

Column subshift columns in space-time diagrams.

Cross section one-sided inverse map.

Directional dynamics dynamics along a direction in the space-time diagram.

Equicontinuous configuration nearby configurations remain close.
Equicontinuous CA all configurations are equicontinuous.
Expansive CA distinct configurations get away.
Finite time attractor is attained in finite time from its neighborhood.
Jointly periodic configuration is periodic both for the shift and the CA.
Lyapunov exponents asymptotic speed of information propagation.
Maximal attractor omega-limit of the full space.
Nilpotent CA maximal attractor is a singleton.
Open CA image of an open set is open.
Permutive CA local rule permutes an extremal coordinate.
Quasi-attractor a countable intersection of attractors.
Signal subshift weakly periodic configurations of a given period.
Spreading set clopen invariant set which propagates in both directions.
Subshift attractor limit set of a spreading set.

Definition of the Subject

A **topological dynamical system** is a continuous selfmap $F\colon X \to X$ of a topological space X. Topological dynamics studies iterations $F^n\colon X \to X$, or trajectories $(F^n(x))_{n\geq 0}$. Basic questions are how trajectories depend on initial conditions, whether they are dense in the state space X, whether they have limits, or what are their accumulation points. While cellular automata were introduced in the late 1940s by Neumann [36] as regular infinite networks of finite automata, topological dynamics of cellular automata began in 1969 with Hedlund [19] who viewed one-dimensional cellular automata in the context of symbolic dynamics as endomorphisms of the shift dynamical systems. In fact, the term "cellular automaton" never appears in his paper. Hedlund's main results are the characterizations of surjective and open cellular automata. In the early 1980s Wolfram [41] produced space-time representations of one-dimensional cellular automata and classified them informally into four classes using dynamical concepts like periodicity, stability and chaos. Wolfram's classification stimulated mathematical research involving all the concepts of topological and measure-theoretical dynamics, and several formal classifications were introduced using dynamical concepts.

There are two well-understood classes of cellular automata with remarkably different stability properties. Equicontinuous cellular automata settle into a fixed or periodic configuration depending on the initial condition and cannot be perturbed by fluctuations. This is a paradigm of stability. Positively expansive cellular automata, on the other hand, are conjugated (isomorphic) to one-sided full shifts. They have dense orbits, dense periodic configurations, positive topological entropy, and sensitive dependence on the initial conditions. This is a paradigm of chaotic behavior. Between these two extreme classes there are many distinct types of dynamical behavior which are understood much less. Only some specific classes or particular examples have been elucidated and a general theory is still lacking.

Introduction

Dynamical properties of CA are usually studied in the context of symbolic dynamics. Other possibilities are measurable dynamics (see Pivato ► Ergodic Theory of Cellular Automata) or non-compact dynamics in Besicovitch or Weyl spaces (see Formenti and Kůrka ► Dynamics of Cellular Automata in Non-compact Spaces). In symbolic dynamics, the state space is the Cantor space of symbolic sequences. The Cantor space has distinguished topological properties which simplify some concepts of dynamics. This is the case of attractor and topological entropy. Cellular automata can be defined in the context of symbolic dynamics as continuous mappings which commute with the shift map.

Equicontinuous and almost equicontinuous CA can be characterized using the concept of blocking words. While equicontinuous CA are eventually periodic, closely related almost equicontinuous automata, are periodic on a large (residual) subset of the state space. Outside of this subset, however, their behavior can be arbitrarily complex.

A property which strongly constrains the dynamics of cellular automata is surjectivity. Surjective automata preserve the uniform Bernoulli measure, they are bounded-to-one, and their unique subshift attractor is the full space. An important subclass of surjective CA are (left- or right-) closing automata. They have dense sets of periodic configurations. Cellular automata which are both left- and right-closing are open: They map open sets to open sets, are n-to-one and have cross-sections. A fairly well-understood class is that of positively expansive automata. A positively expansive cellular automaton is conjugated (isomorphic) to a one-sided full shift. Closely related are bijective expansive automata which are conjugated to two-sided subshifts, usually of finite type.

Another important concept elucidating the dynamics of CA is that of an attractor. With respect to attractors, CA classify into two basic classes. In one class there are CA

which have disjoint attractors. They have then countably infinite numbers of attractors and uncountable numbers of quasi-attractors i. e., countable intersections of attractors. In the other class there are CA that have either a minimal attractor or a minimal quasi-attractor which is then contained in any attractor. An important class of attractors are subshift attractors—subsets which are both attractors and subshifts. They have always non-empty intersection, so they form a lattice with maximal element.

Factors of CA that are subshifts are useful because factor maps preserve many dynamical properties while they simplify the dynamics. In a special case of column subshifts, they are formed by sequences of words occurring in a column of a space-time diagram. Factor subshifts are instrumental in evaluating the entropy of CA and in characterizing CA with the shadowing property.

A finer classification of CA is provided by quantitative characteristics. Topological entropy measures the quantity of information available to an observer who can see a finite part of a configuration. Lyapunov exponents measure the speed of information propagation. The minimum preimage number provides a finer classification of surjective cellular automata. Sets of left- and right-expansivity directions provide finer classification for left- and right-closing cellular automata.

Topological Dynamics

We review basic concepts of topological dynamics as exposed in Kůrka [24]. A **Cantor space** is any metric space which is **compact** (any sequence has a convergent subsequence), **totally disconnected** (distinct points are separated by disjoint **clopen**, i. e., closed and open sets), and **perfect** (no point is isolated). Any two Cantor spaces are homeomorphic. A **symbolic space** is any compact, totally disconnected metric space, i. e., any closed subspace of a Cantor space. A **symbolic dynamical system** (SDS) is a pair (X, F) where X is a symbolic space and $F\colon X \to X$ is a continuous map. The nth **iteration** of F is denoted by F^n. If F is bijective (invertible), the negative iterations are defined by $F^{-n} = (F^{-1})^n$. A set $Y \subseteq X$ is **invariant**, if $F(Y) \subseteq Y$ and **strongly invariant** if $F(Y) = Y$.

A **homomorphism** $\varphi\colon (X, F) \to (Y, G)$ of SDS is a continuous map $\varphi\colon X \to Y$ such that $\varphi \circ F = G \circ \varphi$. A **conjugacy** is a bijective homomorphism. The systems (X, F) and (Y, G) are **conjugated**, if there exists a conjugacy between them. If φ is surjective, we say that (Y, G) is a **factor** of (X, F). If φ is injective, (X, F) is a **subsystem** of (Y, G). In this case $\varphi(X) \subseteq Y$ is a closed invariant set. Conversely, if $Y \subseteq X$ is a closed invariant set, then (Y, F) is a subsystem of (X, F).

We denote by $d\colon X \times X \to [0, \infty)$ the metric and by $B_\delta(x) = \{y \in X\colon d(y, x) < \delta\}$ the ball with center x and radius δ. A finite sequence $(x_i \in X)_{0 \le i \le n}$ is a **δ-chain** from x_0 to x_n, if $d(F(x_i), x_{i+1}) < \delta$ for all $i < n$. A point $x \in X$ **ε-shadows** a sequence $(x_i)_{0 \le i \le n}$, if $d(F^i(x), x_i) < \varepsilon$ for all $0 \le i \le n$. A SDS (X, F) has the **shadowing property**, if for every $\varepsilon > 0$ there exists $\delta > 0$, such that every δ-chain is ε-shadowed by some point.

Definition 1 Let (X, F) be a SDS. The **orbit relation** $\mathcal{O}_F$, the **recurrence relation** $\mathcal{R}_F$, the **non-wandering relation** $\mathcal{N}_F$, and the **chain relation** $\mathcal{C}_F$ are defined by

$$
\begin{aligned}
(x, y) \in \mathcal{O}_F &\iff \exists n > 0, y = F^n(x) \\
(x, y) \in \mathcal{R}_F &\iff \forall \varepsilon > 0, \exists n > 0, d(y, F^n(x)) < \varepsilon \\
(x, y) \in \mathcal{N}_F &\iff \forall \varepsilon, \delta > 0, \exists n > 0, \exists z \in B_\delta(x), \\
&\qquad\qquad d(F^n(z), y) < \varepsilon \\
(x, y) \in \mathcal{C}_F &\iff \forall \varepsilon > 0, \exists \varepsilon - \text{ chain from } x \text{ to } y\,.
\end{aligned}
$$

We have $\mathcal{O}_F \subseteq \mathcal{R}_F \subseteq \mathcal{N}_F \subseteq \mathcal{C}_F$. The diagonal of a relation $S \subseteq X \times X$ is $|S| := \{x \in X\colon (x, x) \in S\}$. We denote by $S(x) := \{y \in X\colon (x, y) \in S\}$ the S-image of a point $x \in X$. The orbit of a point $x \in X$ is $\mathcal{O}_F(x) := \{F^n(x)\colon n > 0\}$. It is an invariant set, so its closure $(\overline{\mathcal{O}_F(x)}, F)$ is a subsystem of (X, F). A point $x \in X$ is **periodic** with period $n > 0$, if $F^n(x) = x$, i. e., if $x \in |\mathcal{O}_F|$. A point $x \in X$ is **eventually periodic**, if $F^m(x)$ is periodic for some **preperiod** $m \ge 0$. The points in $|\mathcal{R}_F|$ are called **recurrent**, the points in $|\mathcal{N}_F|$ are called **non-wandering** and the points in $|\mathcal{C}_F|$ are called **chain-recurrent**. The sets $|\mathcal{N}_F|$ and $|\mathcal{C}_F|$ are closed and invariant, so $(|\mathcal{N}_F|, F)$ and $(|\mathcal{C}_F|, F)$ are subsystems of (X, F). The set of **transitive** points is $\mathcal{T}_F := \{x \in X\colon \overline{\mathcal{O}(x)} = X\}$. A system (X, F) is **minimal**, if $\mathcal{R}_F = X \times X$. This happens if each point has a dense orbit, i. e., if $\mathcal{T}_F = X$. A system is **transitive**, if $\mathcal{N}_F = X \times X$, i. e., if for any non-empty open sets $U, V \subseteq X$ there exists $n > 0$ such that $F^n(U) \cap V \neq \emptyset$. A system is transitive if it has a transitive point, i. e., if $\mathcal{T}_F \neq \emptyset$. In this case the set of transitive points $\mathcal{T}_F$ is **residual**, i. e., it contains a countable intersection of dense open sets. An infinite system is **chaotic**, if it is transitive and has a dense set of periodic points. A system (X, F) is **weakly mixing**, if $(X \times X, F \times F)$ is transitive. It is **strongly transitive**, if (X, F^n) is transitive for any $n > 0$. A system (X, F) is **mixing**, if for every non-empty open sets $U, V \subseteq X$, $F^n(U) \cap V \neq \emptyset$ for all sufficiently large n. A system is **chain-transitive**, if $\mathcal{C}_F = X \times X$, and **chain-mixing**, if for any $x, y \in X$ and any $\varepsilon > 0$ there exist chains from x to y of arbitrary, large enough length. If a system (X, F) has the shadowing property, then $\mathcal{N}_F = \mathcal{C}_F$. It follows that a chain-transitive system with the shadowing property is

transitive, and a chain-mixing system with the shadowing property is mixing.

A **clopen partition** of a symbolic space X is a finite system of disjoint clopen sets whose union is X. The **join** of clopen partitions $\mathcal{U}$ and $\mathcal{V}$ is $\mathcal{U} \vee \mathcal{V} = \{U \cap V : U \in \mathcal{U}, V \in \mathcal{V}\}$. The inverse image of a clopen partition $\mathcal{U}$ by F is $F^{-1}(\mathcal{U}) = \{F^{-1}(U) : U \in \mathcal{U}\}$. The entropy $H(X, F, \mathcal{U})$ of a partition and the **entropy** $\mathbf{h}(X, F)$ of a system are defined by

$$H(X, F, \mathcal{U}) = \lim_{n\to\infty} \frac{\ln |\mathcal{U} \vee F^{-1}(\mathcal{U}) \vee \cdots \vee F^{-(n-1)}(\mathcal{U})|}{n} ,$$

$$\mathbf{h}(X, F) = \sup\{H(X, F, \mathcal{U}) : \mathcal{U} \text{ is a clopen partition of } X\} .$$

Symbolic Dynamics

An **alphabet** is any finite set with at least two elements. The cardinality of a finite set A is denoted by $|A|$. We frequently use alphabet $\mathbf{2} = \{0, 1\}$ and more generally $\mathbf{n} = \{0, \ldots, n-1\}$. A word over A is any finite sequence $u = u_0 \ldots u_{n-1}$ of elements of A. The length of u is denoted by $|u| := n$ and the word of zero length is denoted by λ. The set of all words of length n is denoted by A^n. The set of all non-zero words and the set of all words are

$$A^+ = \bigcup_{n>0} A^n, \quad A^* = \bigcup_{n\geq 0} A^n .$$

We denote by $\mathbb{Z}$ the set of integers, by $\mathbb{N}$ the set of non-negative integers, by $\mathbb{N}^+$ the set of positive integers, by $\mathbb{Q}$ the set of rational numbers, and by $\mathbb{R}$ the set of real numbers. The set of one-sided configurations (infinite words) is $A^{\mathbb{N}}$ and the set of two-sided configurations (biinfinite words) is $A^{\mathbb{Z}}$. If u is a finite or infinite word and $I = [i, j]$ is an interval of integers on which u is defined, put $u_{[i,j]} = u_i \ldots u_j$. Similarly for open or half-open intervals $u_{[i,j)} = u_i \ldots u_{j-1}$. We say that v is a **subword** of u and write $v \sqsubseteq u$, if $v = u_I$ for some interval $I \subseteq \mathbb{Z}$. If $u \in A^n$, then $u^\infty \in A^{\mathbb{Z}}$ is the infinite repetition of u defined by $(u^\infty)_{kn+i} = u_i$. Similarly $x = u^\infty.v^\infty$ is the configuration satisfying $x_{i+k|u|} = u_i$ for $k < 0$, $0 \leq i < |u|$ and $x_{i+k|v|} = v_i$ for $k \geq 0$, $0 \leq i < |v|$. On $A^{\mathbb{N}}$ and $A^{\mathbb{Z}}$ we have metrics

$$d(x, y) = 2^{-n} \quad \text{where } n = \min\{i \geq 0 : x_i \neq y_i\}, \quad x, y \in A^{\mathbb{N}}$$

$$d(x, y) = 2^{-n} \quad \text{where } n = \min\{i \geq 0 : x_i \neq y_i \text{ or } x_{-i} \neq y_{-i}\}, \quad x, y \in A^{\mathbb{Z}} .$$

Both $A^{\mathbb{N}}$ and $A^{\mathbb{Z}}$ are Cantor spaces. In $A^{\mathbb{N}}$ and $A^{\mathbb{Z}}$ the **cylinder sets** of a word $u \in A^n$ are $[u] := \{x \in A^{\mathbb{N}} : x_{[0,n)} = u\}$, and $[u]_k := \{x \in A^{\mathbb{Z}} : x_{[k,k+n)} = u\}$, where $k \in \mathbb{Z}$. Cylinder sets are clopen and every clopen set is a finite union of cylinders. The shift maps $\sigma : A^{\mathbb{N}} \to A^{\mathbb{N}}$ and $\sigma : A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ defined by $\sigma(x)_i = x_{i+1}$ are continuous. While the two-sided shift is bijective, the one-sided shift is not: every configuration has $|A|$ preimages. A **one-sided subshift** is any non-empty closed set $\Sigma \subseteq A^{\mathbb{N}}$ which is shift-invariant, i. e., $\sigma(\Sigma) \subseteq \Sigma$. A **two-sided subshift** is any non-empty closed set $\Sigma \subseteq A^{\mathbb{Z}}$ which is strongly shift-invariant, i. e., $\sigma(\Sigma) = \Sigma$. Thus, a subshift Σ represents a SDS (Σ, σ). Systems $(A^{\mathbb{Z}}, \sigma)$ and $(A^{\mathbb{N}}, \sigma)$ are called **full shifts**. Given a set $D \subseteq A^*$ of **forbidden words**, the set $S_D := \{x \in A^{\mathbb{N}} : \forall u \in D, u \not\sqsubseteq x\}$ is a one-sided subshift, provided it is non-empty. Any one-sided subshift has this form. Similarly, $S_D := \{x \in A^{\mathbb{Z}} : \forall u \in D, u \not\sqsubseteq x\}$ is a two-sided subshift, and any two-sided subshift has this form. A (one- or two-sided) subshift is of **finite type** (SFT), if the set D of forbidden words is finite. The **language of a subshift** Σ is the set of all subwords of configurations of Σ,

$$\mathcal{L}^n(\Sigma) = \{u \in A^n : \exists x \in \Sigma, u \sqsubseteq x\},$$

$$\mathcal{L}(\Sigma) = \bigcup_{n\geq 0} \mathcal{L}^n(\Sigma) = \{u \in A^* : \exists x \in \Sigma, u \sqsubseteq x\} .$$

The entropy of a subshift Σ is $\mathbf{h}(\Sigma, \sigma) = \lim_{n\to\infty} \ln |\mathcal{L}^n(\Sigma)|/n$. A word $w \in \mathcal{L}(\Sigma)$ is **intrinsically synchronizing**, if for any $u, v \in A^*$ such that $uw, wv \in \mathcal{L}(\Sigma)$ we have $uwv \in \mathcal{L}(\Sigma)$. A subshift is of finite type if all sufficiently long words are intrinsically synchronizing (see Lind and Marcus [29]).

A subshift Σ is **sofic**, if $\mathcal{L}(\Sigma)$ is a regular language, i. e., if $\Sigma = \Sigma_{\mathcal{G}}$ is the set of labels of paths of a **labelled graph** $\mathcal{G} = (V, E, s, t, l)$, where V is a finite set of vertices, E is a finite set of edges, $s, t : E \to V$ are the source and target map, and $l : E \to A$ is a labelling function. The labelling function extends to a function $\ell : E^{\mathbb{Z}} \to A^{\mathbb{Z}}$ defined by $\ell(x)_i = l(x_i)$. A graph $\mathcal{G} = (V, E, s, t, l)$ determines a SFT $\Sigma_{|\mathcal{G}|} = \{u \in E^{\mathbb{Z}}, \forall i \in \mathbb{Z}, t(u_i) = s(u_{i+1})\}$ and $\Sigma_{\mathcal{G}} = \{\ell(u) : u \in \Sigma_{|\mathcal{G}|}\}$ so that $\ell : (\Sigma_{|\mathcal{G}|}, \sigma) \to (\Sigma_{\mathcal{G}}, \sigma)$ is a factor map. If $\Sigma = \Sigma_{\mathcal{G}}$, we say that $\mathcal{G}$ is a **presentation** of Σ. A graph $\mathcal{G}$ is **right-resolving**, if different outgoing edges of a vertex are labelled differently, i. e., if $l(e) \neq l(e')$ whenever $e \neq e'$ and $s(e) = s(e')$. A word w is **synchronizing** in $\mathcal{G}$, if all paths with label w have the same target, i. e., if $t(u) = t(u')$ whenever $\ell(u) = \ell(u') = w$. If w is synchronizing in $\mathcal{G}$, then w is intrinsically synchronizing in $\Sigma_{\mathcal{G}}$. Any transitive sofic subshift Σ has a unique **minimal right-resolving presentation** $\mathcal{G}$ which has the

smallest number of vertices. Any word can be extended to a word which is synchronizing in $\mathcal{G}$ (see Lind and Marcus [29]).

A **deterministic finite automaton (DFA)** over an alphabet A is a system $\mathcal{A} = (Q, \delta, q_0, q_1)$, where Q is a finite set of states, $\delta\colon Q \times A \to Q$ is a transition function and q_0, q_1 are the initial and rejecting states. The transition function extends to $\delta\colon Q \times A^* \to Q$ by $\delta(q, \lambda) = q$, $\delta(q, ua) = \delta(\delta(q, u), a)$. The language accepted by $\mathcal{A}$ is $\mathcal{L}(\mathcal{A}) := \{u \in A^*\colon \delta(q_0, u) \neq q_1\}$ (see e. g., Hopcroft and Ullmann [20]). The DFA of a labelled graph $\mathcal{G} = (V, E, s, t, l)$ is $\mathcal{A}(\mathcal{G}) = (\mathcal{P}(V), \delta, V, \emptyset)$, where $\mathcal{P}(V)$ is the set of all subsets of V and $\delta(q, a) = \{v \in V\colon \exists u \in q, u \xrightarrow{a} v\}$. Then $\mathcal{L}(\mathcal{A}(\mathcal{G})) = \mathcal{L}(\Sigma_G)$. We can reduce the size of $\mathcal{A}(\mathcal{G})$ by taking only those states that are accessible from the initial state V.

A **periodic structure** $\mathbf{n} = (n_i)_{i\geq 0}$ is a sequence of integers greater than 1. For a given periodic structure $\mathbf{n}$, let $X_{\mathbf{n}} := \prod_{i\geq 0}\{0, \ldots, n_i - 1\}$ be the product space with metric $d(x, y) = 2^{-n}$ where $n = \min\{i \geq 0 : x_i \neq y_i\}$. Then $X_{\mathbf{n}}$ is a Cantor space. The **adding machine** (odometer) of $\mathbf{n}$ is a SDS $(X_{\mathbf{n}}, F)$ given by the formula

$$F(x)_i = \begin{cases} (x_i + 1) \bmod n_i & \text{if } \forall j < i, x_j = n_j - 1 \\ x_i & \text{if } \exists j < i, x_j < n_j - 1\,. \end{cases}$$

Each adding machine is minimal and has zero topological entropy.

Definition 2 A map $F\colon A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ is a **cellular automaton (CA)** if there exist integers $m \leq a$ (**memory** and **anticipation**) and a **local rule** $f\colon A^{a-m+1} \to A$ such that for any $x \in A^{\mathbb{Z}}$ and any $i \in \mathbb{Z}$, $F(x)_i = f(x_{[i+m,i+a]})$. Call $r = \max\{|m|, |a|\} \geq 0$ the **radius** of F and $d = a - m \geq 0$ its **diameter**.

By a theorem of Hedlund [19], a map $F\colon A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ is a cellular automaton if it is continuous and commutes with the shift, i. e., $\sigma \circ F = F \circ \sigma$. This means that $F\colon (A^{\mathbb{Z}}, \sigma) \to (A^{\mathbb{Z}}, \sigma)$ is a homomorphism of the full shift and $(A^{\mathbb{Z}}, F)$ is a SDS. We can assume that the local rule acts on a symmetric neighborhood of 0, so $F(x)_i = f(x_{[i-r,i+r]})$, where $f\colon A^{2r+1} \to A$. There is a trade-off between the radius and the size of the alphabet. Any CA is conjugated to a CA with radius 1. Any σ-periodic configuration of a CA $(A^{\mathbb{Z}}, F)$ is F-eventually periodic. Hence the set of F-eventually periodic configurations is dense. Thus, a cellular automaton is never minimal, because it has always an F-periodic configuration. A configuration $x \in A^{\mathbb{Z}}$ is **weakly periodic**, if there exists $p \in \mathbb{Z}$ and $q \in \mathbb{N}^+$ such that $F^q\sigma^p(x) = x$. A configuration $x \in A^{\mathbb{Z}}$ is **jointly periodic**, if it is both F-periodic and σ-periodic. A CA $(A^{\mathbb{Z}}, F)$ is **nilpotent**, if $F^n(A^{\mathbb{Z}})$ is a singleton for some $n > 0$.

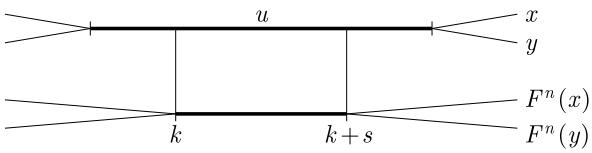

Topological Dynamics of Cellular Automata, Figure 1
A blocking word

Equicontinuity

A point $x \in X$ of a SDS (X, F) is **equicontinuous**, if

$$\forall \varepsilon > 0, \exists \delta > 0, \forall y \in B_\delta(x), \forall n \geq 0, \quad d(F^n(y), F^n(x)) < \varepsilon\,.$$

The set of equicontinuous points is denoted by $\mathcal{E}_F$. A system is **equicontinuous**, if $\mathcal{E}_F = X$. In this case it is uniformly equicontinuous, i. e.,

$$\forall \varepsilon > 0, \exists \delta > 0, \forall x, y \in X, (d(x, y) < \delta \Rightarrow \forall n \geq 0, d(F^n(x), F^n(y)) < \varepsilon)\,.$$

A system (X, F) is **almost equicontinuous**, if $\mathcal{E}_F \neq \emptyset$. A system is **sensitive**, if

$$\exists \varepsilon > 0, \forall x \in X, \forall \delta > 0, \exists y \in B_\delta(x), \exists n \geq 0, \quad d(F^n(y), F^n(x)) \geq \varepsilon\,.$$

Clearly, a sensitive system has no equicontinuous points. The converse is not true in general but holds for transitive systems (Akin et al. [2]).

Definition 3 A word $u \in A^+$ with $|u| \geq s \geq 0$ is ***s*-blocking** for a CA $(A^{\mathbb{Z}}, F)$, if there exists an **offset** $k \in [0, |u|-s]$ such that

$$\forall x, y \in [u]_0, \forall n \geq 0, F^n(x)_{[k,k+s)} = F^n(y)_{[k,k+s)}\,.$$

Theorem 4 (Kůrka [27]) *Let $(A^{\mathbb{Z}}, F)$ be a CA with radius $r \geq 0$. The following conditions are equivalent.*

(1) $(A^{\mathbb{Z}}, F)$ is not sensitive.
(2) $(A^{\mathbb{Z}}, F)$ has an r-blocking word.
*(3) $\mathcal{E}_F$ is **residual**, i. e., a countable intersection of dense open sets.*
(4) $\mathcal{E}_F \neq \emptyset$.

For a non-empty set $B \subseteq A^*$ define

$$\mathcal{T}_\sigma^n(B) := \{x \in A^{\mathbb{Z}} : (\exists j > i > n, x_{[i,j)} \in B) \text{ and } (\exists j < i < -n, x_{[j,i)} \in B)\},$$

$$\mathcal{T}_\sigma(B) := \bigcap_{n \geq 0} \mathcal{T}_\sigma^n(B) .$$

Each $\mathcal{T}_\sigma^n(B)$ is open and dense, so the set $\mathcal{T}_\sigma(B)$ of ***B*-recurrent** configurations is residual. If B is the set of r-blocking words, then $\mathcal{E}_F = \mathcal{T}_\sigma(B)$.

Theorem 5 (Kůrka [27]) *Let $(A^{\mathbb{Z}}, F)$ be a CA with radius $r \geq 0$. The following conditions are equivalent.*

(1) $(A^{\mathbb{Z}}, F)$ is equicontinuous, i. e., $\mathcal{E}_F = A^{\mathbb{Z}}$.
(2) There exists $k > 0$ such that any $u \in A^k$ is r-blocking.
(3) There exists a preperiod $q \geq 0$ and a period $p > 0$, such that $F^{q+p} = F^q$.

In particular every CA with radius $r = 0$ is equicontinuous. A configuration is equicontinuous for F if it is equicontinuous for F^n, i. e., $\mathcal{E}_F = \mathcal{E}_{F^n}$. This fact enables us to consider equicontinuity along rational directions $\alpha = \frac{p}{q}$.

Definition 6 The sets of **equicontinuous directions** and **almost equicontinuous directions** of a CA $(A^{\mathbb{Z}}, F)$ are defined by

$$\mathfrak{E}(F) = \left\{ \frac{p}{q} : p \in \mathbb{Z}, q \in \mathbb{N}^+, \mathcal{E}_{F^q \sigma^p} = A^{\mathbb{Z}} \right\} ,$$

$$\mathfrak{A}(F) = \left\{ \frac{p}{q} : p \in \mathbb{Z}, q \in \mathbb{N}^+, \mathcal{E}_{F^q \sigma^p} \neq \emptyset \right\} .$$

Clearly, $\mathfrak{E}(F) \subseteq \mathfrak{A}(F)$, and both sets $\mathfrak{E}(F)$ and $\mathfrak{A}(F)$ are convex (Sablik [37]): If $\alpha_0 < \alpha_1 < \alpha_2$ and $\alpha_0, \alpha_2 \in \mathfrak{A}(F)$, then $\alpha_1 \in \mathfrak{A}(F)$. Sablik [37] considers also equicontinuity along irrational directions.

Proposition 7 *Let $(A^{\mathbb{Z}}, F)$ be an equicontinuous CA such that there exists $0 \neq \alpha \in \mathfrak{A}(F)$. Then $(A^{\mathbb{Z}}, F)$ is nilpotent.*

Proof We can assume $\alpha < 0$. There exist $0 \leq k < m$ and $w \in A^m$, such that for all $x, y \in A^{\mathbb{Z}}$ and for all $i \in \mathbb{Z}$ we have

$$x_{[i,i+m)} = y_{[i,i+m)} \implies \forall n \geq 0, F^n(x)_{i+k} = F^n(y)_{i+k},$$

$$w = x_{[i,i+m)} = y_{[i,i+m)} \implies \forall n \geq 0, F^n \sigma^{\lfloor n\alpha \rfloor}(x)_{i+k} = F^n \sigma^{\lfloor n\alpha \rfloor}(y)_{i+k} .$$

Take n such that $l := \lfloor n\alpha \rfloor + m \leq 0$. There exists $a \in A$ such that $F^n \sigma^{\lfloor n\alpha \rfloor}(z)_k = a$ for every $z \in [w]_0$. Let $x \in A^{\mathbb{Z}}$ be arbitrary. For a given $i \in \mathbb{Z}$, take a configuration $y \in [x_{[i,i+m)}]_i \cap [w]_{i-l+m}$. Then $z := \sigma^{i-\lfloor n\alpha \rfloor}(y) = \sigma^{i-l+m}(y) \in [w]_0$ and $F^n(x)_{i+k} = F^n(y)_{i+k} = F^n \sigma^{\lfloor n\alpha \rfloor}(z)_k = a$. Thus, $F^n(x) = a^\infty$ for every $x \in A^{\mathbb{Z}}$, and $F^{n+t}(x) = F^n(F^t(x)) = a^\infty$ for every $t \geq 0$, so $(A^{\mathbb{Z}}, F)$ is nilpotent (see also Sablik [38]). □

Theorem 8 *Let $(A^{\mathbb{Z}}, F)$ be a CA with memory m and anticipation a, i. e., $F(x)_i = f(x_{[i+m,i+a]})$. Then exactly one of the following conditions is satisfied.*

(1) $\mathfrak{E}(F) = \mathfrak{A}(F) = \mathbb{Q}$. This happens if $(A^{\mathbb{Z}}, F)$ is nilpotent.
(2) $\mathfrak{E}(F) = \emptyset$ and there exist real numbers $\alpha_0 < \alpha_1$ such that

$$(\alpha_0, \alpha_1) \subseteq \mathfrak{A}(F) \subseteq [\alpha_0, \alpha_1] \subseteq [-a, -m] .$$

(3) There exists $-a \leq \alpha \leq -m$ such that $\mathfrak{A}(F) = \mathfrak{E}(F) = \{\alpha\}$.
(4) There exists $-a \leq \alpha \leq -m$ such that $\mathfrak{A}(F) = \{\alpha\}$ and $\mathfrak{E}(F) = \emptyset$.
(5) $\mathfrak{A}(F) = \mathfrak{E}(F) = \emptyset$.

This follows from Theorems II.2 and II.5 in Sablik [37] and from Proposition 7. The zero CA of Example 1 belongs to class (1). The product CA of Example 4 belongs to class (2). The identity CA of Example 2 belongs to class (3). The Coven CA of Example 18 belongs to class (4). The sum CA of Example 11 belongs to class (5).

Sensitivity can be expressed quantitatively by **Lyapunov exponents** which measure the speed of information propagation. Let $(A^{\mathbb{Z}}, F)$ be a CA. The left and right **perturbation sets** of $x \in A^{\mathbb{Z}}$ are

$$W_s^-(x) = \{y \in A^{\mathbb{Z}} : \forall i \leq s, y_i = x_i\} ,$$

$$W_s^+(x) = \{y \in A^{\mathbb{Z}} : \forall i \geq s, y_i = x_i\} .$$

The left and right **perturbation speeds** of $x \in A^{\mathbb{Z}}$ are

$$I_n^-(x) = \min\{s \geq 0 : \forall i \leq n, F^i(W_s^-(x)) \subseteq W_0^-(F^i(x))\} ,$$

$$I_n^+(x) = \min\{s \geq 0 : \forall i \leq n, F^i(W_{-s}^+(x)) \subseteq W_0^+(F^i(x))\} .$$

Thus, $I_n^-(x)$ is the minimum distance of a perturbation of the left part of x which cannot reach the zero site by time n. Both $I_n^-(x)$ and $I_n^+(x)$ are nondecreasing. If $0 < s < t$, and if $x_{[s,t]}$ is an r-blocking word (where r is the radius), then $\lim_{n\to\infty} I_n^-(x) \leq t$. Similarly, if $s < t < 0$ and if $x_{[s,t]}$ is an r-blocking word, then $\lim_{n\to\infty} I_n^+(x) \leq |s|$. In particular, if $x \in \mathcal{E}_F$, then both

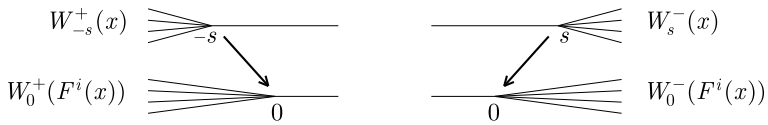

Topological Dynamics of Cellular Automata, Figure 2
Perturbation speeds

$I_n^-(x)$ and $I_n^+(x)$ have finite limit. If $(A^{\mathbb{Z}}, F)$ is sensitive, then $\lim_{n\to\infty}(I_n^-(x) + I_n^+(x)) = \infty$ for every $x \in A^{\mathbb{Z}}$.

Definition 9 The left and right **Lyapunov exponents** of a CA $(A^{\mathbb{Z}}, F)$ and $x \in A^{\mathbb{Z}}$ are

$$\lambda_F^-(x) = \liminf_{n\to\infty} \frac{I_n^-(x)}{n}\,, \quad \lambda_F^+(x) = \liminf_{n\to\infty} \frac{I_n^+(x)}{n}\,.$$

If F has memory m and anticipation a, then $\lambda_F^-(x) \le \max\{a, 0\}$ and $\lambda_F^+(x) \le \max\{-m, 0\}$ for all $x \in A^{\mathbb{Z}}$. If $x \in \mathcal{E}_F$, then $\lambda_F^+(x) = \lambda_F^-(x) = 0$. If F is right-permutive (see Sect. "Permutive and Closing Cellular Automata") with $a > 0$, then $\lambda_F^-(x) = a$ for every $x \in A^{\mathbb{Z}}$. If F is left-permutive with $m < 0$, then $\lambda_F^+(x) = -m$ for every $x \in A^{\mathbb{Z}}$.

Theorem 10 (Bressaud and Tisseur [9]) *For a positively expansive CA (see Sect. "Expansive Cellular Automata") there exists a constant $c > 0$, such that for all $x \in A^{\mathbb{Z}}$, $\lambda^-(x) \ge c$ and $\lambda^+(x) \ge c$.*

Conjecture 11 (Bressaud and Tisseur [9]) Any sensitive CA has a configuration x such that $\lambda^-(x) > 0$ or $\lambda^+(x) > 0$.

Surjectivity

Let $(A^{\mathbb{Z}}, F)$ be a CA with diameter $d \ge 0$ and a local rule $f\colon A^{d+1} \to A$. We extend the local rule to a function $f\colon A^* \to A^*$ by $f(u)_i = f(u_{[i,i+d]})$ for $i < |u| - d$, so $|f(u)| = \max\{|u| - d, 0\}$. A **diamond** for f (Fig. 3 left) consists of words $u, v \in A^d$ and distinct $w, z \in A^+$ of the same length, such that $f(uwv) = f(uzw)$.

Theorem 12 (Hedlund [19], Moothathu [33]) *Let $(A^{\mathbb{Z}}, F)$ be a CA with local rule $f\colon A^{d+1} \to A$. The following conditions are equivalent.*

(1) $F\colon A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ is surjective.
(2) For each $x \in A^{\mathbb{Z}}$, $F^{-1}(x)$ is finite or countable.
(3) For each $x \in A^{\mathbb{Z}}$, $F^{-1}(x)$ is a finite set.
(4) For each $x \in A^{\mathbb{Z}}$, $|F^{-1}(x)| \le |A|^d$.
(5) $f\colon A^* \to A^*$ is surjective.
(6) For each $u \in A^+$, $|f^{-1}(u)| = |A|^d$.
(7) For each $u \in A^+$ with $|u| \le d \cdot \log_2 |A| \cdot (2d + |A|^{2d})$, $|f^{-1}(u)| = |A|^d$.
(8) There exists no diamond for f.

It follows that any injective CA is surjective and hence bijective. Although (6) asserts equality, the inequality in (4) may be strict. Another equivalent condition states that the uniform Bernoulli measure is invariant for F. In this form, Theorem 12 has been generalized to CA on mixing SFT (see Theorem 2B.1 in Pivato ▶ Ergodic Theory of Cellular Automata).

Theorem 13 (Blanchard and Tisseur [5])

(1) Any configuration of a surjective CA is non-wandering, i. e., $|\mathcal{N}_F| = A^{\mathbb{Z}}$.
(2) Any surjective almost equicontinuous CA has a dense set of F-periodic configurations.
(3) If $(A^{\mathbb{Z}}, F)$ is an equicontinuous and surjective CA, then there exists $p > 0$ such that $F^p = \mathrm{Id}$. In particular, F is bijective.

Theorem 14 (Moothathu [32]) *Let $(A^{\mathbb{Z}}, F)$ be a surjective CA.*

(1) $|\mathcal{R}_F|$ is dense in $A^{\mathbb{Z}}$.
*(2) F is **semi-open**, i. e., $F(U)$ has non-empty interior for any open $U \ne \emptyset$.*
(3) If $(A^{\mathbb{Z}}, F)$ is transitive, then it is weakly mixing, and hence totally transitive and sensitive.

Conjecture 15 Every surjective CA has a dense set of F-periodic configurations.

Proposition 16 (Acerbi et al. [1]) *If every mixing CA has a dense set of F-periodic configurations, then every surjective CA has a dense set of jointly periodic configurations.*

Definition 17 Let $(A^{\mathbb{Z}}, F)$ be a CA with local rule $f\colon A^{d+1} \to A$.

(1) The **minimum preimage number** (Fig. 3 right) $\mathbf{p}(F)$ is defined by

$$p(F, w) = \min_{t \le |w|} |\{u \in A^d : \exists v \in f^{-1}(w), v_{[t,t+d)} = u\}|,$$
$$\mathbf{p}(F) = \min\{p(F, w) : w \in A^+\}\,.$$

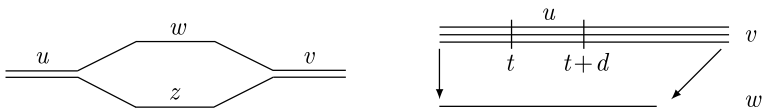

Topological Dynamics of Cellular Automata, Figure 3
A diamond (*left*) and a magic word (*right*)

(2) A word $w \in A^+$ is **magic**, if $p(F, w) = \mathbf{p}(F)$.

Recall that $\mathcal{T}_\sigma(w)$ is the (residual) set of configurations which contain an infinite number of occurrences of w both in $x_{(-\infty,0)}$ and in $x_{(0,\infty)}$. Configurations $x, y \in A^{\mathbb{Z}}$ are ***d*-separated**, if $x_{[i,i+d)} \neq y_{[i,i+d)}$ for all $i \in \mathbb{Z}$.

Theorem 18 (Hedlund [19], Kitchens [23]) *Let $(A^{\mathbb{Z}}, F)$ be a surjective CA with diameter d and minimum preimage number $\mathbf{p}(F)$.*

(1) If $w \in A^+$ is a magic word, then any $z \in \mathcal{T}_\sigma(w)$ has exactly $\mathbf{p}(F)$ preimages. These preimages are pairwise d-separated.
(2) Any configuration $z \in A^{\mathbb{Z}}$ has at least $\mathbf{p}(F)$ pairwise d-separated preimages.
(3) If every $y \in A^{\mathbb{Z}}$ has exactly $\mathbf{p}(F)$ preimages, then all long enough words are magic.

Theorem 19 *Let $(A^{\mathbb{Z}}, F)$ be a CA and $\Sigma \subseteq A^{\mathbb{Z}}$ a sofic subshift. Then both $F(\Sigma)$ and $F^{-1}(\Sigma)$ are sofic subshifts. In particular, the first image subshift $F(A^{\mathbb{Z}})$ is sofic.*

See e. g., Formenti and Kůrka [16] for a proof. The **first image graph** of a local rule $f\colon A^{d+1} \to A$ is $\mathcal{G}(f) = (A^d, A^{d+1}, s, t, f)$, where $s(u) = u_{[0,d)}$ and $t(u) = u_{[1,d]}$. Then $F(A^{\mathbb{Z}}) = \Sigma_{\mathcal{G}(f)}$. It is algorithmically decidable whether a given CA is surjective. One decision procedure is based on the Moothathu result in Theorem 12(7). Another procedure is based on the construction of the DFA $\mathcal{A}(\mathcal{G}(f))$ (see Sect. "Symbolic Dynamics"). A CA with local rule $f\colon A^{d+1} \to A$ is surjective if the rejecting state $\emptyset$ cannot be reached from the initial state A^d in $\mathcal{A}(\mathcal{G}(f))$. See Morita ► Reversible Cellular Automata for further information on bijective CA.

Permutive and Closing Cellular Automata

Definition 20 Let $(A^{\mathbb{Z}}, F)$ be a CA, and let $f\colon A^{d+1} \to A$ be the local rule for F with smallest diameter.

(1) F is **left-permutive** if $\forall u \in A^d, \forall b \in A, \exists! a \in A$, $f(au) = b$.
(2) F is **right-permutive** if $\forall u \in A^d, \forall b \in A, \exists! a \in A$, $f(ua) = b$.
(3) F is **permutive** if it is either left-permutive or right-permutive.
(4) F is **bipermutive** if it is both left- and right-permutive.

Permutive CA can be seen in Examples 8, 10, 11, 18.

Definition 21 Let $(A^{\mathbb{Z}}, F)$ be a CA.

(1) Configurations $x, y \in A^{\mathbb{Z}}$ are **left-asymptotic**, if $\exists n$, $x_{(-\infty,n)} = y_{(-\infty,n)}$.
(2) Configurations $x, y \in A^{\mathbb{Z}}$ are **right-asymptotic**, if $\exists n$, $x_{(n,\infty)} = y_{(n,\infty)}$.
(3) $(A^{\mathbb{Z}}, F)$ is **right-closing** if $F(x) \neq F(y)$ for any left-asymptotic $x \neq y \in A^{\mathbb{Z}}$.
(4) $(A^{\mathbb{Z}}, F)$ is **left-closing** if $F(x) \neq F(y)$ for any right-asymptotic $x \neq y \in A^{\mathbb{Z}}$.
(5) A CA is **closing** if it is either left- or right-closing.

Proposition 22

(1) Any right-permutive CA is right-closing.
(2) Any right-closing CA is surjective.
(3) A CA $(A^{\mathbb{Z}}, F)$ is right-closing if there exists $m > 0$ such that for any $x, y \in A^{\mathbb{Z}}$, $x_{[-m,0)} = y_{[-m,0)}$ and $F(x)_{[-m,m]} = F(y)_{[-m,m]} \implies x_0 = y_0$ (see Fig. 4 left).

See e. g., Kůrka [24] for a proof. The proposition holds with obvious modification for left-permutive and left-closing CA. The multiplication CA from Example 14 is both left- and right-closing but neither left-permutive nor right-permutive. The CA from Example 15 is surjective but not closing.

Proposition 23 *Let $(A^{\mathbb{Z}}, F)$ be a right-closing CA. For all sufficiently large $m > 0$, if $u \in A^m$, $v \in A^{2m}$, and if $F([u]_{-m}) \cap [v]_{-m} \neq \emptyset$, then (Fig. 4 right)*

$$\forall b \in A, \exists! a \in A, F([ua]_{-m}) \cap [vb]_{-m} \neq \emptyset\,.$$

See e. g., Kůrka [24] for a proof.

Theorem 24 (Boyle and Kitchens [6]) *Any closing CA $(A^{\mathbb{Z}}, F)$ has a dense set of jointly periodic configurations.*

Theorem 25 (Coven, Pivato and Yassawi [14]) *Let F be a left-permutive CA with memory 0.*

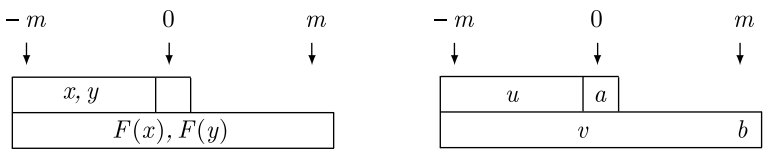

Topological Dynamics of Cellular Automata, Figure 4
Closingness

(1) If $\mathcal{O}(x)$ is infinite and $x_{[0,\infty)}$ is fixed, i. e., if $F(x)_{[0,\infty)} = x_{[0,\infty)}$, then $(\overline{\mathcal{O}(x)}, F)$ is conjugated to an adding machine.
(2) If F is not bijective, then the set of configurations such that $(\overline{\mathcal{O}(x)}, F)$ is conjugated to an adding machine is dense.

A SDS (X, F) is **open**, if $F(U)$ is open for any open $U \subseteq X$. A **cross section** of a SDS (X, F) is any continuous map $G\colon X \to X$ such that $F \circ G = \mathrm{Id}$. If F has a cross section, it is surjective. In particular, any bijective SDS has a cross section.

Theorem 26 (Hedlund [19]) *Let $(A^{\mathbb{Z}}, F)$ be a CA. The following conditions are equivalent.*

(1) $(A^{\mathbb{Z}}, F)$ is open.
(2) $(A^{\mathbb{Z}}, F)$ is both left- and right-closing.
(3) For any $x \in A^{\mathbb{Z}}$, $|F^{-1}(x)| = \mathbf{p}(F)$
(4) There exist cross sections $G_1, \ldots, G_{\mathbf{p}(F)}\colon A^{\mathbb{Z}} \to A^{\mathbb{Z}}$, such that for any $x \in A^{\mathbb{Z}}$, $F^{-1}(x) = \{G_1(x), \ldots, G_{\mathbf{p}(F)}(x)\}$ and $G_i(x) \neq G_j(x)$ for $i \neq j$.

In general, the cross sections G_i are not CA as they need not commute with the shift. Only when $\mathbf{p}(F) = 1$, i. e., when F is bijective, the inverse map F^{-1} is a CA. Any CA which is open and almost equicontinuous is bijective (Kůrka [24]).

Expansive Cellular Automata

Definition 27 Let $(A^{\mathbb{Z}}, F)$ be a CA.

(1) F is **left-expansive**, if there exists $\varepsilon > 0$ such that if $x_{(-\infty,0]} \neq y_{(-\infty,0]}$, then $d(F^n(x), F^n(y)) \geq \varepsilon$ for some $n \geq 0$.
(2) F is **right-expansive**, if there exists $\varepsilon > 0$ such that if $x_{[0,\infty)} \neq y_{[0,\infty)}$, then $d(F^n(x), F^n(y)) \geq \varepsilon$ for some $n \geq 0$.
(3) F is **positively expansive**, if it is both left- and right-expansive, i. e., if there exists $\varepsilon > 0$ such that for all $x \neq y \in A^{\mathbb{Z}}$, $d(F^n(x), F^n(y)) \geq \varepsilon$ for some $n > 0$.

Any left-expansive or right-expansive CA is sensitive and (by Theorem 12) surjective, because it cannot contain a diamond. A bijective CA is **expansive**, if

$$\exists \varepsilon > 0, \forall x \neq y \in A^{\mathbb{Z}}, \exists n \in \mathbb{Z}, d(F^n(x), F^n(y)) \geq \varepsilon\,.$$

Proposition 28 *Let $(A^{\mathbb{Z}}, F)$ be a CA with memory m and anticipation a.*

(1) If $m < 0$ and if F is left-permutive, then F is left-expansive.
(2) If $a > 0$ and if F is right-permutive, then F is right-expansive.
(3) If $m < 0 < a$ and if F is bipermutive, then F is positively expansive.

See e. g., Kůrka [24] for a proof.

Theorem 29 (Nasu [34,35])

(1) Any positively expansive CA is conjugated to a one-sided full shift.
(2) A bijective expansive CA with memory 0 is conjugated to a two-sided SFT.

Conjecture 30 Every bijective expansive CA is conjugated to a two-sided SFT.

Definition 31 Let $(A^{\mathbb{Z}}, F)$ be a CA. The left- and right-**expansivity direction sets** are defined by

$$\begin{aligned}
&\mathfrak{X}^-(F) \\
&\quad = \left\{ \frac{p}{q} \colon p \in \mathbb{Z}, q \in \mathbb{N}^+,\ F^q \sigma^p \text{ is left-expansive} \right\}, \\
&\mathfrak{X}^+(F) \\
&\quad = \left\{ \frac{p}{q} \colon p \in \mathbb{Z}, q \in \mathbb{N}^+,\ F^q \sigma^p \text{ is right-expansive} \right\}, \\
&\mathfrak{X}(F) = \mathfrak{X}^-(F) \cap \mathfrak{X}^+(F)\,.
\end{aligned}$$

All these sets are convex and open. Moreover, $\mathfrak{X}^-(F) \cap \mathfrak{A}(F) = \mathfrak{X}^+(F) \cap \mathfrak{A}(F) = \emptyset$ (Sablik [37]).

Theorem 32 (Sablik [37]) *Let $(A^{\mathbb{Z}}, F)$ be a CA with memory m and anticipation a.*

(1) If F is left-permutive, then $\mathfrak{X}^-(F) = (-\infty, -m)$.
(2) If F is right-permutive, then $\mathfrak{X}^+(F) = (-a, \infty)$.
(3) If $\mathfrak{X}^-(F) \neq \emptyset$ then there exists $\alpha \in \mathbb{R}$ such that $\mathfrak{X}^-(F) = (-\infty, \alpha) \subseteq (-\infty, -m)$.
(4) If $\mathfrak{X}^+(F) \neq \emptyset$ then there exists $\alpha \in \mathbb{R}$ such that $\mathfrak{X}^+(F) = (\alpha, \infty) \subseteq (-a, \infty)$.
(5) If $\mathfrak{X}(F) \neq \emptyset$ then there exists $\alpha_0, \alpha_1 \in \mathbb{R}$ such that $\mathfrak{X}(F) = (\alpha_0, \alpha_1) \subseteq (-a, -m)$.

Theorem 33 *Let $(A^{\mathbb{Z}}, F)$ be a cellular automaton.*

(1) F is left-closing if $\mathfrak{X}^-(F) \neq \emptyset$.
(2) F is right-closing if $\mathfrak{X}^+(F) \neq \emptyset$.
(3) If $\mathfrak{A}(F)$ is an interval, then F is not surjective and $\mathfrak{X}^-(F) = \mathfrak{X}^+(F) = \emptyset$.

Proof (1) The proof is the same as the following proof of (2).

($2 \Leftarrow$) If F is not right-closing and $\varepsilon = 2^{-n}$, then there exist distinct left-asymptotic configurations such that $x_{(-\infty,n]} = y_{(-\infty,n]}$ and $F(x) = F(y)$. It follows that $d(F^i(x), F^i(y)) < \varepsilon$ for all $i \geq 0$, so F is not right-expansive. The same argument works for any $F^q\sigma^p$, so $\mathfrak{X}^+(F) = \emptyset$.

($2 \Rightarrow$) Let F be right-closing, and let $m > 0$ be the constant from Proposition 23. Assume that

$$F^n(x)_{[-m+(m+1)n, m+(m+1)n]} = F^n(y)_{[-m+(m+1)n, m+(m+1)n]}$$

for all $n \geq 0$. By Proposition 23,

$$F^{n-1}(x)_{m+(m+1)(n-1)+1} = F^{n-1}(y)_{m+(m+1)(n-1)+1} .$$

By induction we get $x_{[-m,m+n]} = y_{[-m,m+n]}$. This holds for every $n > 0$, so $x_{[0,\infty)} = y_{[0,\infty)}$. Thus, $F\sigma^{m+1}$ is right-expansive, and therefore $\mathfrak{X}^+(F) \neq \emptyset$.

(3) If there are blocking words for two different directions, then the CA has a diamond and therefore is not surjective by Theorem 12. □

Corollary 34 *Let $(A^{\mathbb{Z}}, F)$ be an equicontinuous CA. There are three possibilities.*

(1) If F is surjective, then $\mathfrak{A}(F) = \mathfrak{E}(F) = \{0\}$, $\mathfrak{X}^-(F) = (-\infty, 0)$, $\mathfrak{X}^+(F) = (0, \infty)$.
(2) If F is neither surjective nor nilpotent, then $\mathfrak{A}(F) = \mathfrak{E}(F) = \{0\}$, $\mathfrak{X}^-(F) = \mathfrak{X}^+(F) = \emptyset$.
(3) If F is nilpotent, then $\mathfrak{A}(F) = \mathfrak{E}(F) = \mathfrak{R}$, $\mathfrak{X}^-(F) = \mathfrak{X}^+(F) = \emptyset$.

The proof follows from Proposition 7 and Theorem 13 (see also Sablik [38]). The identity CA is in class (1). The product CA of Example 3 is in class (2). The zero CA of Example 1 is in class (3).

Attractors

Let (X, F) be a SDS. The **limit set** of a clopen invariant set $V \subseteq X$ is $\Omega_F(V) := \bigcap_{n \geq 0} F^n(V)$. A set $Y \subseteq X$ is an **attractor**, if there exists a non-empty clopen invariant set V such that $Y = \Omega_F(V)$. We say that Y is a **finite time attractor**, if $Y = \Omega_F(V) = F^n(V)$ for some $n > 0$ (and a clopen invariant set V). There exists always the largest attractor $\Omega_F := \Omega_F(X)$. Finite time maximal attractors are also called **stable limit sets** in the literature. The number of attractors is at most countable. The union of two attractors is an attractor. If the intersection of two attractors is non-empty, it contains an attractor. The **basin** of an attractor $Y \subseteq X$ is the set $\mathcal{B}(Y) = \{x \in X;\ \lim_{n\to\infty} d(F^n(x), Y) = 0\}$. An attractor $Y \subseteq X$ is a **minimal attractor**, if no proper subset of Y is an attractor. An attractor is a minimal attractor if it is chain-transitive. A periodic point $x \in X$ is **attracting** if its orbit $\mathcal{O}(x)$ is an attractor. Any attracting periodic point is equicontinuous. A **quasi-attractor** is a non-empty set which is an intersection of a countable number of attractors.

Theorem 35 (Hurley [22])

(1) If a CA has two disjoint attractors, then any attractor contains two disjoint attractors and an uncountably infinite number of quasi-attractors.
(2) If a CA has a minimal attractor, then it is a subshift, it is contained in any other attractor, and its basin of attraction is a dense open set.
(3) If $x \in A^{\mathbb{Z}}$ is an attracting F-periodic configuration, then $\sigma(x) = x$ and $F(x) = x$.

Corollary 36 *For any CA, exactly one of the following statements holds.*

(1) There exist two disjoint attractors and a continuum of quasi-attractors.
(2) There exists a unique quasi-attractor. It is a subshift and it is contained in any attractor.
(3) There exists a unique minimal attractor contained in any other attractor.

Both equicontinuity and surjectivity yield strong constraints on attractors.

Theorem 37 (Kůrka [24])

(1) A surjective CA has either a unique attractor or a pair of disjoint attractors.
(2) An equicontinuous CA has either two disjoint attractors or a unique attractor which is an attracting fixed configuration.
(3) If a CA has an attracting fixed configuration which is a unique attractor, then it is equicontinuous.

We consider now subshift attractors of CA, i. e., those attractors which are subshifts. Let $(A^{\mathbb{Z}}, F)$ be a CA. A clopen F-invariant set $U \subseteq A^{\mathbb{Z}}$ is **spreading**, if there exists $k > 0$ such that $F^k(U) \subseteq \sigma^{-1}(U) \cap \sigma(U)$. If U is a clopen invariant set, then $\Omega_F(U)$ is a subshift iff U is spreading (Kůrka [25]). Recall that a language is **recursively enumerable**, if it is a domain (or a range) of a recursive function (see e. g., Hopcroft and Ullmann [20]).

Theorem 38 (Formenti and Kůrka [17]) *Let $\Sigma \subseteq A^{\mathbb{Z}}$ be a subshift attractor of a CA $(A^{\mathbb{Z}}, F)$.*

(1) $A^* \setminus \mathcal{L}(\Sigma)$ is a recursively enumerable language.
(2) Σ contains a jointly periodic configuration.
(3) (Σ, σ) is chain-mixing.

Theorem 39 (Formenti and Kůrka [17])

(1) The only subshift attractor of a surjective CA is the full space.
(2) A subshift of finite type is an attractor of a CA if it is mixing.
*(3) Given a CA $(A^{\mathbb{Z}}, F)$, the intersection of all subshift attractors of all $F^q\sigma^p$, where $q \in \mathbb{N}^+$ and $p \in \mathbb{Z}$, is a non-empty F-invariant subshift called the **small quasi-attractor** $\mathcal{Q}_F$.$(\mathcal{Q}_F, \sigma)$ is chain-mixing and $F\colon \mathcal{Q}_F \to \mathcal{Q}_F$ is surjective.*

The system of all subshift attractors of a given CA forms a lattice with join $\Sigma_0 \cup \Sigma_1$ and meet $\Sigma_0 \wedge \Sigma_1 := \Omega_F(\Sigma_0 \cap \Sigma_1)$. There exist CA with infinite number of subshift attractors (Kůrka [26]).

Proposition 40 (Di Lena [28]) *The basin of a subshift attractor is a dense open set.*

By a theorem of Hurd [21], if Ω_F is SFT, then it is stable, i. e., $\Omega_F = F^n(A^{\mathbb{Z}})$ for some $n > 0$. We generalize this theorem to subshift attractors.

Theorem 41 *Let U be a spreading set for a CA $(A^{\mathbb{Z}}, F)$.*

(1) There exists a spreading set $W \subseteq U$ such that $\Omega_F(W) = \Omega_F(U)$ and $\widetilde{\Omega}_\sigma(W) := \bigcap_{i\in\mathbb{Z}} \sigma^i(W)$ is a mixing subshift of finite type.
(2) If $\Omega_F(W)$ is a SFT, then $\Omega_F(W) = F^n(\widetilde{\Omega}_\sigma(W))$ for some $n \geq 0$.

Proof

(1) See Formenti and Kůrka [17].
(2) Let D be a finite set of forbidden words for $\Omega_F(W)$. For each $u \in D$ there exists $n_u > 0$ such that $u \notin \mathcal{L}(F^{n_u}(\widetilde{\Omega}_\sigma))$. Take $n := \max\{n_u : u \in D\}$. □

By Theorem 5, every equicontinuous CA has a finite time maximal attractor.

Definition 42 Let $\Sigma \subseteq A^{\mathbb{Z}}$ be a mixing sofic subshift, let $\mathcal{G} = (V, E, s, t, l)$ be its minimal right-resolving presentation with factor map $\ell\colon (\Sigma_{|\mathcal{G}|}, \sigma) \to (\Sigma, \sigma)$.

(1) A homogenous configuration $a^\infty \in \Sigma$ is **receptive**, if there exist intrinsically synchronizing words $u, v \in \mathcal{L}(\Sigma)$ and $n \in \mathbb{N}$ such that $ua^m v \in \mathcal{L}(\Sigma)$ for all $m > n$.
(2) Σ is **almost of finite type (AFT)**, if $\ell\colon \Sigma_{|\mathcal{G}|} \to \Sigma$ is one-to-one on a dense open set of $\Sigma_{|G|}$.
(3) Σ is **near-Markov**, if $\{x \in \Sigma : |\ell^{-1}(x)| > 1\}$ is a finite set of σ-periodic configurations.

(3) is equivalent to the condition that ℓ is left-closing, i. e., that $\ell(u) \neq \ell(v)$ for distinct right-asymptotic paths $u, v \in E^{\mathbb{Z}}$. Each near-Markov subshift is AFT.

Theorem 43 (Maass [30]) *Let $\Sigma \subseteq A^{\mathbb{Z}}$ be a mixing sofic subshift with a receptive configuration $a^\infty \in \Sigma$.*

(1) If Σ is either SFT or AFT, then there exists a CA $(A^{\mathbb{Z}}, F)$ such that $\Sigma = \Omega_F = F(A^{\mathbb{Z}})$.
(2) A near-Markov subshift cannot be an infinite time maximal attractor of a CA.

On the other hand, a near-Markov subshift can be a finite time maximal attractor (see Example 21). The language $\mathcal{L}(\Omega_F)$ can have arbitrary complexity (see Culik et al. [15]). A CA with non-sofic mixing maximal attractor has been constructed in Formenti and Kůrka [17].

Definition 44 Let $f\colon A^{d+1} \to A$ be a local rule of a cellular automaton. We say that a subshift $\Sigma \subseteq A^{\mathbb{Z}}$ has **decreasing preimages**, if there exists $m > 0$ such that for each $u \in A^* \setminus \mathcal{L}(\Sigma)$, each $v \in f^{-m}(u)$ contains as a subword a word $w \in A^* \setminus \mathcal{L}(\Sigma)$ such that $|w| < |u|$.

Proposition 45 (Formenti and Kůrka [16]) *If $(A^{\mathbb{Z}}, F)$ is a CA and $\Sigma \subseteq A^{\mathbb{Z}}$ has decreasing preimages, then $\Omega_F \subseteq \Sigma$.*

For more information about attractor-like objects in CA see Sect. "Invariance of Maxentropy Measures" of Pivato ► Ergodic Theory of Cellular Automata.

Subshifts and Entropy

Definition 46 Let $F\colon A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ be a cellular automaton.

(1) Denote by $\mathcal{S}_{(p,q)}(F) := \{x \in A^{\mathbb{Z}} : F^q\sigma^p(x) = x\}$ the set of all weakly periodic configurations of F with period $(p, q) \in \mathbb{Z} \times \mathbb{N}^+$.
(2) A **signal subshift** is any non-empty $\mathcal{S}_{(p,q)}(F)$.
(3) The **speed subshift** of F with speed $\alpha = \frac{p}{q} \in \mathbb{Q}$ is $\mathcal{S}_\alpha(F) = \overline{\bigcup_{n>0} \mathcal{S}_{(np,nq)}(F)}$.

Note that both $\mathcal{S}_{(p,q)}(F)$ and $\mathcal{S}_\alpha(F)$ are closed and σ-invariant. However, $\mathcal{S}_{(p,q)}(F)$ can be empty, so it need not be a subshift.

Theorem 47 *Let $(A^{\mathbb{Z}}, F)$ be a cellular automaton with memory m and anticipation a, so $F(x)_i = f(x_{[i+m,i+a]})$.*

(1) If $\mathcal{S}_{(p,q)}(F)$ is non-empty, then it is a subshift of finite type.
(2) If $\mathcal{S}_{(p,q)}(F)$ is infinite, then $-a \le p/q \le -m$.
(3) If $p_0/q_0 < p_1/q_1$, then $\mathcal{S}_{(p_0,q_0)}(F) \cap \mathcal{S}_{(p_1,q_1)}(F) \subseteq \{x \in A^{\mathbb{Z}} : \sigma^p(x) = x\}$, where $p = q(p_1/q_1 - p_0/q_0)$ and $q = \mathbf{lcm}(q_0, q_1)$ (the least common multiple).
(4) $\mathcal{S}_{(p,q)}(F) \subseteq \mathcal{S}_{\frac{p}{q}}(F) \subseteq \Omega_F$ and $\mathcal{S}_{\frac{p}{q}}(F) \neq \emptyset$.
(5) If $\mathcal{X}(F) \neq \emptyset$ or if $(A^{\mathbb{Z}}, F)$ is nilpotent, then F has no infinite signal subshifts.

Proof (1), (2) and (3) have been proved in Formenti and Kůrka [16].

(4) Since F is bijective on each signal subshift, we get $\mathcal{S}_{(p,q)}(F) \subseteq \Omega_F$, and therefore $\mathcal{S}_{p/q}(F) \subseteq \Omega_F$. Since every $F^q\sigma^p$ has a periodic point, we get $\mathcal{S}_{p/q}(F) \neq \emptyset$.

(5) It has been proved in Kůrka [25], that a positively expansive CA has no signal subshifts. This property is preserved when we compose F with a power of the shift map. If $(A^{\mathbb{Z}}, F)$ is nilpotent, then each $\mathcal{S}_{(p,q)}(F)$ contains at most one element. □

The Identity CA has a unique infinite signal subshift $\mathcal{S}_{(0,1)}(\mathrm{Id}) = A^{\mathbb{Z}}$. The CA of Example 17 has an infinite number of infinite signal subshifts of the same speed. A CA with infinitely many infinite signal subshifts with infinitely many speeds has been constructed in Kůrka [25]. In some cases, the maximal attractor can be constructed from signal subshifts (see Theorem 49).

Definition 48 Given an integer $c \ge 0$, the ***c*-join** $\Sigma_0 \overset{c}{\vee} \Sigma_1$ of subshifts $\Sigma_0, \Sigma_1 \subseteq A^{\mathbb{Z}}$ consists of all configurations $x \in A^{\mathbb{Z}}$ such that either $x \in \Sigma_0 \cup \Sigma_1$, or there exist integers b, a such that $b - a \ge c$, $x_{(-\infty,b)} \in \mathcal{L}(\Sigma_0)$, and $x_{[a,\infty)} \in \mathcal{L}(\Sigma_1)$.

The operation of join is associative, and the c-join of sofic subshifts is sofic.

Theorem 49 (Formenti, Kůrka [16]) *Let $(A^{\mathbb{Z}}, F)$ be a CA and let $\mathcal{S}_{(p_1,q_1)}(F), \ldots, \mathcal{S}_{(p_n,q_n)}(F)$ be signal subshifts with decreasing speeds, i. e., $p_i/q_i > p_j/q_j$ for $i < j$. Set $q := \mathbf{lcm}\{q_1, \ldots q_n\}$ (the least common multiple). There exists $c \ge 0$ such that for $\Sigma := \mathcal{S}_{(p_1,q_1)}(F) \overset{c}{\vee} \cdots \overset{c}{\vee} \mathcal{S}_{(p_n,q_n)}(F)$ we have $\Sigma \subseteq F^q(\Sigma)$ and therefore $\Sigma \subseteq \Omega_F$. If moreover $F^{nq}(\Sigma)$ has decreasing preimages for some $n \ge 0$, then $F^{nq}(\Sigma) = \Omega_F$.*

Definition 50 Let $(A^{\mathbb{Z}}, F)$ be a CA.

(1) The ***k*-column homomorphism** $\varphi_k\colon (A^{\mathbb{Z}}, F) \to ((A^k)^{\mathbb{N}}, \sigma)$ is defined by $\varphi_k(x)_i = F^i(x)_{[0,k)}$.
(2) The ***k*th column subshift** is $\Sigma_k(F) = \varphi_k(A^{\mathbb{Z}}) \subseteq (A^k)^{\mathbb{N}}$.
(3) If $\psi\colon (A^{\mathbb{Z}}, F) \to (\Sigma, \sigma)$ is a factor map, where Σ is a one-sided subshift, we say that Σ is a **factor subshift** of $(A^{\mathbb{Z}}, F)$.

Thus, each $(\Sigma_k(F), \sigma)$ is a factor of $(A^{\mathbb{Z}}, F)$ and each factor subshift is a factor of some $\Sigma_k(F)$. Any positively expansive CA $(A^{\mathbb{Z}}, F)$ with radius $r > 0$ is conjugated to $(\Sigma_{2r+1}(F), \sigma)$. This is an SFT which by Theorem 29 is conjugated to a full shift.

Proposition 51 (Shereshevski and Afraimovich [39]) *Let $(A^{\mathbb{Z}}, F)$ be a CA with negative memory and positive anticipation $m < 0 < a$. Then $(A^{\mathbb{Z}}, F)$ is bipermutive if it is positively expansive and $\Sigma_{a-m+1}(F) = (A^{a-m+1})^{\mathbb{N}}$ is the full shift.*

Theorem 52 (Blanchard and Maass [4], Di Lena [28]) *Let $(A^{\mathbb{Z}}, F)$ be a CA with radius r and memory m.*

(1) If $m \ge 0$ and $\Sigma_r(F)$ is sofic, then any factor subshift of $(A^{\mathbb{Z}}, F)$ is sofic.
(2) If $\Sigma_{2r+1}(F)$ is sofic, then any factor subshift of $(A^{\mathbb{Z}}, F)$ is sofic.

If $(x_i)_{i\ge0}$ is a 2^{-m}-chain in a CA $(A^{\mathbb{Z}}, F)$, then for all i, $F(x_i)_{[-m,m]} = (x_{i+1})_{[-m,m]}$, so $u_i = (x_i)_{[-m,m]}$ satisfy $F([u_i]_{-m}) \cap [u_{i+1}]_{-m} \neq \emptyset$. Conversely, if a sequence $(u_i \in A^{2m+1})_{i\ge0}$ satisfies this property and $x_i \in [u_i]_{-m}$, then $(x_i)_{i\ge0}$ is a 2^{-m}-chain.

Theorem 53 (Kůrka [27]) *Let $(A^{\mathbb{Z}}, F)$ be a CA.*

(1) If $\Sigma_k(F)$ is an SFT for any $k > 0$, then $(A^{\mathbb{Z}}, F)$ has the shadowing property.
(2) If $(A^{\mathbb{Z}}, F)$ has the shadowing property, then any factor subshift is sofic.

Any factor subshift of the Coven CA from Example 18 is sofic, but the CA does not have the shadowing property (Blanchard and Maass [4]). A CA with shadowing property whose factor subshift is not SFT has been constructed in Kůrka [24].

Proposition 54 *Let $(A^{\mathbb{Z}}, F)$ be a CA.*

(1) $\mathbf{h}(A^{\mathbb{Z}}, F) = \lim_{k\to\infty} \mathbf{h}(\Sigma_k(F), \sigma)$.
(2) If F has radius r, then

$$\mathbf{h}(A^{\mathbb{Z}}, F) \le 2 \cdot \mathbf{h}(\Sigma_r(F), \sigma) \le 2r \cdot \mathbf{h}(\Sigma_1(F), \sigma) \le 2r \cdot \ln|A|.$$

(3) If $0 \le m \le a$, then $\mathbf{h}(A^{\mathbb{Z}}, F) = \mathbf{h}(\Sigma_a(F), \sigma)$.

See e. g., Kůrka [24] for a proof.

Conjecture 55 If $(A^{\mathbb{Z}}, F)$ is a CA with radius r, then $\mathbf{h}(A^{\mathbb{Z}}, F) = \mathbf{h}(\Sigma_{2r+1}(F), \sigma)$.

Conjecture 56 (Moothathu [32]) Any transitive CA has positive topological entropy.

Definition 57 The **directional entropy** of a CA $(A^{\mathbb{Z}}, F)$ along a rational direction $\alpha = p/q$ is $\mathbf{h}_\alpha(A^{\mathbb{Z}}, F) := \mathbf{h}(A^{\mathbb{Z}}, F^q\sigma^p)/q$.

The definition is based on the equality $\mathbf{h}(X, F^n) = n \cdot \mathbf{h}(X, F)$ which holds for every SDS. Directional entropies along irrational directions have been introduced in Milnor [31].

Proposition 58 (Courbage and Kamiński [11], Sablik [37]) *Let $(A^{\mathbb{Z}}, F)$ be a CA with memory m and anticipation a.*

(1) If $\alpha \in \mathcal{E}(F)$, then $\mathbf{h}_\alpha(F) = 0$.
(2) If $\alpha \in \mathcal{X}^-(F) \cup \mathcal{X}^+(F)$, then $\mathbf{h}_\alpha(F) > 0$.
(3) $\mathbf{h}_\alpha(F) \le (\max(a+\alpha, 0) - \min(m+\alpha, 0)) \cdot \ln|A|$.
(4) If F is bipermutive, then $\mathbf{h}_\alpha(F) = (\max\{a+\alpha, 0\} - \min\{m+\alpha, 0\}) \cdot \ln|A|$.
(5) If F is left-permutive, and $\alpha < -a$, then $\mathbf{h}_\alpha(F) = |m+\alpha| \cdot \ln|A|$.
(6) If F is right-permutive, and $\alpha > -m$, then $\mathbf{h}_\alpha(F) = (a+\alpha) \cdot \ln|A|$.

The directional entropy is not necessarily continuous (see Smillie [40]).

Theorem 59 (Boyle and Lind [8]) *The function $\alpha \mapsto \mathbf{h}_\alpha(A^{\mathbb{Z}}, F)$ is convex and continuous on $\mathcal{X}^-(F) \cup \mathcal{X}^+(F)$.*

Examples

Cellular automata with binary alphabet $\mathbf{2} = \{0, 1\}$ and radius $r = 1$ are called **elementary** (Wolfram [42]). Their local rules are coded by numbers between 0 and 255 by

$$f(000) + 2 \cdot f(001) + 4 \cdot f(010) + \cdots + 32 \cdot f(101) + 64 \cdot f(110) + 128 \cdot f(111)\,.$$

Example 1 (The zero rule ECA0) $F(x) = 0^\infty$.

The zero CA is an equicontinuous nilpotent CA. Its equicontinuity directions are $\mathfrak{E}(F) = \mathfrak{A}(F) = (-\infty, \infty)$.

Example 2 (The identity rule ECA204) $\mathrm{Id}(x) = x$.

The identity is an equicontinuous surjective CA which is not transitive. Every clopen set is an attractor and every configuration is a quasi-attractor. The equicontinuity and expansivity directions are $\mathfrak{A}(\mathrm{Id}) = \mathfrak{E}(\mathrm{Id}) = \{0\}$, $\mathcal{X}^-(\mathrm{Id}) = (-\infty, 0)$, $\mathcal{X}^+(\mathrm{Id}) = (0, \infty)$. The directional entropy is $h_\alpha(\mathrm{Id}) = |\alpha|$.

Example 3 (An equicontinuous rule ECA12) $F(x)_i = (1 - x_{i-1})x_i$.

$$000:0,\quad 001:0,\quad 010:1,\quad 011:1,\quad 100:0,\quad 101:0,\quad 110:0,\quad 111:0.$$

The ECA12 is equicontinuous: the preperiod and period are $m = p = 1$. The automaton has finite time maximal attractor $\Omega_F = F(A^{\mathbb{Z}}) = \Sigma_{\{11\}} = \mathcal{S}_{(0,1)}(F)$ which is called the **golden mean subshift**.

Example 4 (A product rule ECA128) $F(x)_i = x_{i-1}x_ix_{i+1}$.

The ECA128 is almost equicontinuous and 0 is a 1-blocking word. The first column subshift is $\Sigma_1(F) = S_{\{01\}}$. Each column subshift $\Sigma_k(F)$ is an SFT with zero entropy, so F has the shadowing property and zero entropy. The nth image

$$F^n(\mathbf{2}^{\mathbb{Z}}) = \{x \in \mathbf{2}^{\mathbb{Z}} : \forall m \in [1, 2n], 10^m1 \not\sqsubseteq x\}$$

is a SFT. The first image graph can be seen in Fig. 10 left. The maximal attractor $\Omega_F = S_{\{10^n1:n>0\}}$ is a sofic subshift and has decreasing preimages. The only other attractor is the minimal attractor $\{0^\infty\} = \Omega_F([0]_0)$, which is also the minimal quasi-attractor. The equicontinuity directions are

Topological Dynamics of Cellular Automata, Figure 5
ECA12

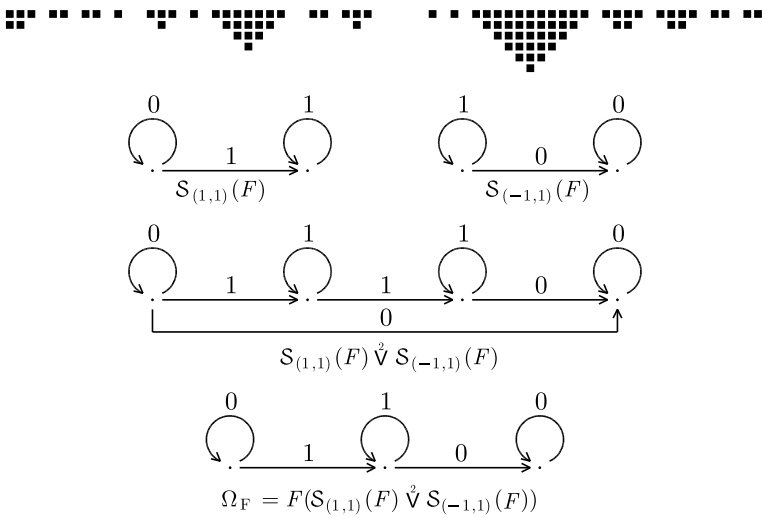

Topological Dynamics of Cellular Automata, Figure 6
Signal subshifts of ECA128

$\mathfrak{E}(F) = \emptyset$ and $\mathfrak{A}(F) = [-1, 1]$. For Lyapunov exponents we have $\lambda_F^-(0^\infty) = \lambda_F^+(0^\infty) = 0$ and $\lambda_F^-(1^\infty) = \lambda_F^+(1^\infty) = 1$. The only infinite signal subshifts are non-transitive subshifts $\mathcal{S}_{(1,1)}(F) = S_{\{10\}}$ and $\mathcal{S}_{(-1,1)}(F) = S_{\{01\}}$. The maximal attractor can be constructed using the join construction $\Omega_F = F(\mathcal{S}_{(1,1)}(F) \overset{2}{\vee} \mathcal{S}_{(-1,1)}(F))$ (Fig. 6).

Example 5 (A product rule ECA136) $F(x)_i = x_i x_{i+1}$.

The ECA136 is almost equicontinuous since 0 is a 1-blocking word. As in Example 4, we have $\Omega_F = S_{\{10^k 1 : k>0\}}$. For any $m \in \mathbb{Z}$, $[0]_m$ is a clopen invariant set, which is spreading to the left but not to the right. Thus, $Y_m = \Omega_F([0]_m) = \{x \in \Omega_F : \forall i \leq m, x_i = 0\}$ is an attractor but not a subshift. We have $Y_{m+1} \subset Y_m$ and $\bigcap_{m \geq 0} Y_m = \{0^\infty\}$ is the unique minimal quasi-attractor. Since $F^2\sigma^{-1}(x)_i = x_{i-1}x_i x_{i+1}$ is the ECA128 which has a minimal subshift attractor $\{0^\infty\}$, F has the small quasi-attractor $\mathcal{Q}_F = \{0^\infty\}$. The almost equicontinuity directions are $\mathfrak{A}(F) = [-1, 0]$.

Example 6 (A unique attractor) $(\mathbf{2}^{\mathbb{Z}}, F)$ where $F(x)_i = x_{i+1}x_{i+2}$.

The system is sensitive and has a unique attractor $\Omega_F = S_{\{10^k 1 : k>0\}}$ which is not F-transitive. If $x \in [10]_0 \cap \Omega_F(\mathbf{2}^{\mathbb{Z}})$, then $x_{[0,\infty)} = 10^\infty$, so for any $n > 0$, $F^n(x) \notin [11]_0$. However, $(A^{\mathbb{Z}}, F)$ is chain-transitive, so it does not have the shadowing property. The small quasi-attractor is $\mathcal{Q}_F = \{0^\infty\}$. The topological entropy is zero. The factor subshift $\Sigma_1(F) = \{x \in \mathbf{2}^{\mathbb{N}} : \forall n \geq 0, (x_{[n,n+1]} = 10 \implies x_{[n,2n+1]} = 10^{n+1})\}$ is not sofic (Gilman [18]).

Example 7 (The majority rule ECA232) $F(x)_i = \lfloor (x_{i-1} + x_i + x_{i+1})/2 \rfloor$.

$$000 : 0, \quad 001 : 0, \quad 010 : 0, \quad 011 : 1, \quad 100 : 0, \quad 101 : 1, \quad 110 : 1, \quad 111 : 1.$$

The majority rule has 2-blocking words 00 and 11, so it is almost equicontinuous. More generally, let $E = \{u \in \mathbf{2}^* : |u| \geq 2, u_0 = u_1, u_{|u|-2} = u_{|u|-1}, 010 \not\sqsubseteq u, 101 \not\sqsubseteq u\}$. Then for any $u \in E$ and for any $i \in \mathbb{Z}$, $[u]_i$ is a clopen invariant set, so its limit set $\Omega_F([u]_i)$ is an attractor. These attractors are not subshifts. There exists a subshift attractor given by the spreading set $U := \mathbf{2}^{\mathbb{Z}} \setminus ([010]_0 \cup [101]_0)$. We have $\Omega_F(U) = \mathcal{S}_{(0,1)}(F) = S_{\{010,101\}}$. There are two more infinite signal subshifts $\mathcal{S}_{(-1,1)}(F) = S_{\{001,110\}}$ and $\mathcal{S}_{(1,1)}(F) = S_{\{011,100\}}$. The maximal attractor is $\Omega_F = \mathcal{S}_{(1,0)}(F) \cup (\mathcal{S}_{(1,1)}(F) \overset{3}{\vee} \mathcal{S}_{(-1,1)}(F)) = S_{\{010^k 1, 10^k 10, 01^k 01, 101^k 0, : k>1\}}$. All column subshifts are SFT, for example $\Sigma_1(F) = S_{\{001,110\}}$ and the entropy is zero. The equicontinuity directions are $\mathfrak{E}(F) = \emptyset$, $\mathfrak{A}(F) = \{0\}$.

Topological Dynamics of Cellular Automata, Figure 7
ECA136

Topological Dynamics of Cellular Automata, Figure 8
A unique attractor $F(x)_i = x_{i+1}x_{i+2}$

Topological Dynamics of Cellular Automata, Figure 9
The majority rule ECA232

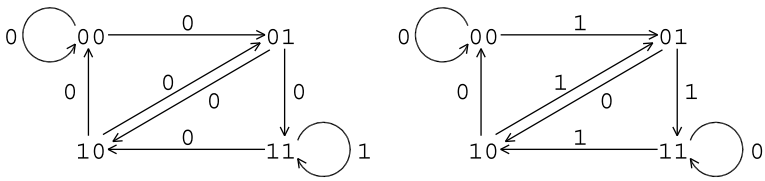

Topological Dynamics of Cellular Automata, Figure 10
First image subshift of ECA128 (*left*) and ECA106 (*right*)

Example 8 (A right-permutive rule ECA106) $F(x)_i = (x_{i-1}x_i + x_{i+1}) \bmod 2$.

$$000:0,\quad 001:1,\quad 010:0,\quad 011:1,\quad 100:0,\quad 101:1,\quad 110:1,\quad 111:0.$$

The ECA106 is transitive (see Kůrka [24]). The first image graph is in Fig. 10 right. The minimum preimage number is $\mathbf{p}_F = 1$ and the word $u = 0101$ is magic. Its preimages are $f^{-1}(0101) = \{010101, 100101, 000101, 111001\}$ and for every $v \in f^{-1}(u)$ we have $v_{[4,5]} = 01$. This can be seen in Fig. 11 bottom left, where all paths in the first image graph with label 0101 are displayed. Accordingly, $(01)^\infty$ has a unique preimage $F^{-1}((01)^\infty) = \{(10)^\infty\}$. On the other hand 0^∞ has two preimages $F^{-1}(0^\infty) = \{0^\infty, 1^\infty\}$ (Fig. 11 bottom right) and 1^∞ has three preimages $F^{-1}(1^\infty) = \{(011)^\infty, (110)^\infty, (101)^\infty\}$. We have $\mathfrak{X}^-(F) = \emptyset$ and $\mathfrak{X}^+(F) = (-1, \infty)$ and there are no equicontinuity directions. For every x we have $\lambda_F^-(x) = 1$. On the other hand the right Lyapunov exponents are not constant. For example, $\lambda_F^+(0^\infty) = 0$ while $\lambda_F^+((01)^\infty) = 1$. The only infinite signal subshift is the golden mean subshift $\mathcal{S}_{(-1,1)}(F) = S_{\{11\}}$.

Example 9 (The shift rule ECA170) $\sigma(x)_i = x_{i+1}$.

The shift rule is bijective, expansive and transitive. It has a dense set of periodic configurations, so it is chaotic. Its only signal subshift is $\mathcal{S}_{(-1,1)}(\sigma) = \mathbf{2}^{\mathbb{Z}}$. The equicontinuity and expansivity directions are $\mathfrak{E}(\sigma) = \mathfrak{A}(\sigma) = \{-1\}$, $\mathfrak{X}^-(\sigma) = (-\infty, -1)$, $\mathfrak{X}^+(\sigma) = (-1, \infty)$. For any $x \in \mathbf{2}^{\mathbb{Z}}$ we have $\lambda_\sigma^-(x) = 0$, $\lambda_\sigma^+(x) = 1$.

Example 10 (A bipermutive rule ECA102) $F(x)_i = (x_i + x_{i+1}) \bmod 2$.

$$000:0,\quad 001:1,\quad 010:1,\quad 011:0,\quad 100:0,\quad 101:1,\quad 110:1,\quad 111:0.$$

The ECA102 is bipermutive with memory 0, so it is open but not positively expansive. The expansivity directions are $\mathfrak{X}^-(F) = (-\infty, 0)$, $\mathfrak{X}^+(F) = (-1, \infty)$, $\mathfrak{X}(F) = (-1, 0)$. If $x \in Y := \overline{W_0^+(0^\infty.10^\infty)}$, i. e., if $x_0 = 1$ and $x_i = 0$ for all $i > 0$, then $(\overline{\mathcal{O}(x)}, F) = (Y, F)$ is conjugated to the adding machine with periodic structure $\mathbf{n} = (2, 2, 2, \dots)$. If $-2^n < i \le -2^{n-1}$, then $(F^m(x)_i)_{m \ge 0}$ is periodic with period 2^n. There are no signal subshifts. The minimum preimage number is $\mathbf{p}_F = 2$ and the two cross sections G_0, G_1 are uniquely determined by

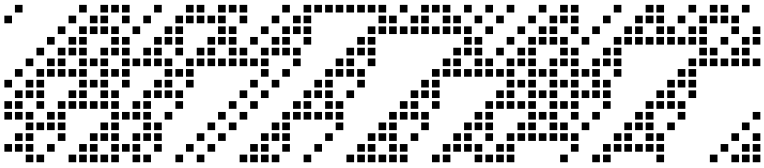

```
11    01 ---> 11    01 ---> 11      01    01    01    01    01
00 \     11 \    00 \     11    00   10 \    10 \    10 \    10 \    10
10 ---> 00    10 ---> 00    10      00 ---> 00 ---> 00 ---> 00 ---> 00
01 ---> 10 ---> 01 ---> 10 ---> 01   11 ---> 11 ---> 11 ---> 11 ---> 11
   0     1     0     1              0     0     0     0
```

Topological Dynamics of Cellular Automata, Figure 11
Preimages in ECA106

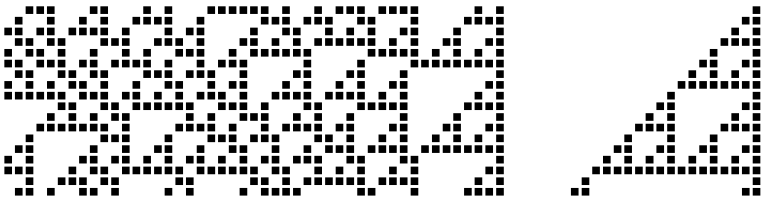

Topological Dynamics of Cellular Automata, Figure 12
ECA102

the conditions $G_0(x)_0 = 0$, $G_1(x)_0 = 1$. The entropy is $\mathbf{h}(A^{\mathbb{Z}}, F) = \ln 2$.

Example 11 (The sum rule ECA90) $F(x)_i = (x_{i-1} + x_{i+1}) \bmod 2$.

$$000:0,\quad 001:1,\quad 010:0,\quad 011:1,\quad 100:1,\quad 101:0,\quad 110:1,\quad 111:0.$$

The sum rule is bipermutive with negative memory and positive anticipation. Thus it is open, positively expansive and mixing. It is conjugated to the full shift on four symbols $\Sigma_2(F) = \{00, 01, 10, 11\}^{\mathbb{N}}$. It has four cross-sections G_0, G_1, G_2, G_3 which are uniquely determined by the conditions $G_0(x)_{[0,1]} = 00$, $G_1(x)_{[0,1]} = 01$, $G_2(x)_{[0,1]} = 10$, and $G_3(x)_{[0,1]} = 11$. For every $x \in \mathbf{2}^{\mathbb{Z}}$ we have $\lambda_F^-(x) = \lambda_F^+(x) = 1$. The system has no almost equicontinuous directions and $\mathfrak{X}^-(F) = (-\infty, 1)$, $\mathfrak{X}^+(F) = (-1, \infty)$. The directional entropy is continuous and piecewise linear (Fig. 13).

Example 12 (The traffic rule ECA184) $F(x)_i = 1$ if $x_{[i-1,i]} = 10$ or $x_{[i,i+1]} = 11$.

$$000:0\quad 001:0\quad 010:0\quad 011:1\quad 100:1\quad 101:1\quad 110:0\quad 111:0\,.$$

The ECA184 has three infinite signal subshifts

$$\mathcal{S}_{(1,1)}(F) = S_{\{11\}} \cup \{1^\infty\},\quad \mathcal{S}_{(0,1)}(F) = S_{\{10\}},\quad \mathcal{S}_{(-1,1)}(F) = S_{\{00\}} \cup \{0^\infty\}$$

and a unique F-transitive attractor $\Omega_F = \mathcal{S}_{(1,1)}(F) \overset{2}{\vee} \mathcal{S}_{(-1,1)}(F) = S_{\{1(10)^n 0 : n>0\}}$ which is sofic. The system has neither almost equicontinuous nor expansive directions. The directional entropy is continuous, but neither piecewise linear nor convex (Smillie [40]).

Example 13 (ECA62) $F(x)_i = x_{i-1}(x_i + 1) + x_i x_{i+1}$.

$$000:0,\quad 001:1,\quad 010:1,\quad 011:1,\quad 100:1,\quad 101:1,\quad 110:0,\quad 111:0.$$

There exists a spreading set $U = A^{\mathbb{Z}} \setminus ([0^6]_2 \cup [1^7]_1 \cup \bigcup_{v \in f^{-1}(1^7)} [v]_0)$ and $\Omega_F(U)$ is a subshift attractor which contains σ-transitive infinite signal subshifts $\mathcal{S}_{(1,2)}(F)$ and $\mathcal{S}_{(0,3)}(F)$ as well as their join. It follows $\mathcal{Q}_F = \Omega_F(U) = F^2(\mathcal{S}_{(1,2)}(F) \overset{3}{\vee} \mathcal{S}_{(0,3)}(F))$. The only other infinite signal subshifts are $\mathcal{S}_{(4,4)}(F)$ and $\mathcal{S}_{(-1,1)}(F)$ and

$$\Omega_F = F^2(\mathcal{S}_{(4,4)}(F) \overset{3}{\vee} \mathcal{S}_{(1,2)}(F) \overset{3}{\vee} \mathcal{S}_{(0,3)}(F) \overset{3}{\vee} \mathcal{S}_{(-1,1)}(F)).$$

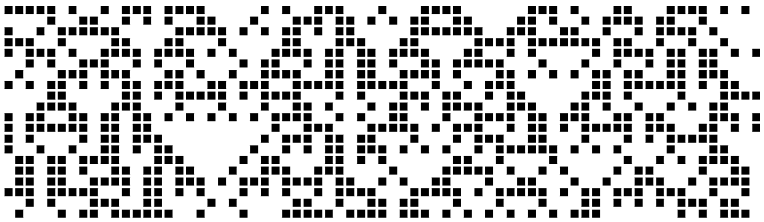

$$\mathbf{h}_\alpha(F) = \begin{cases} (1-\alpha)\ln 2 & \text{for} \quad \alpha \le -1 \\ 2\ln 2 & \text{for} \quad -1 \le \alpha \le 1 \\ (1+\alpha)\ln 2 & \text{for} \quad \alpha \ge 1 \end{cases}$$

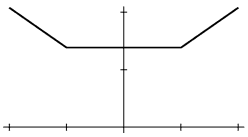

Topological Dynamics of Cellular Automata, Figure 13
Directional entropy of ECA90

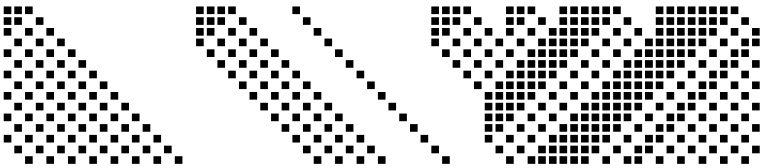

Topological Dynamics of Cellular Automata, Figure 14
The traffic rule ECA184

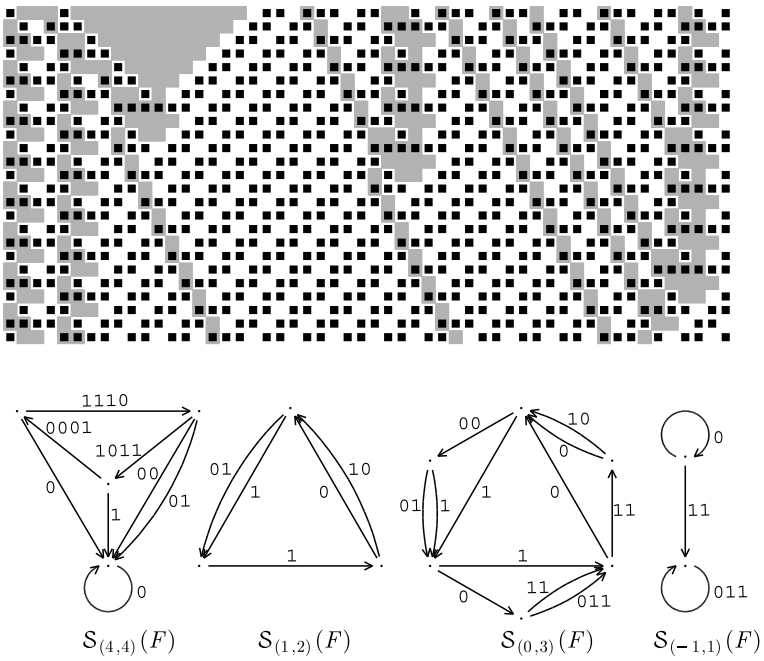

Topological Dynamics of Cellular Automata, Figure 15
ECA 62 and its signal subshifts

00001000030000123000 0
00002000120000312000 0
00010000300001230000 0
00020001200003120000

Topological Dynamics of Cellular Automata, Figure 16
The multiplication CA

In the space-time diagram in Fig. 15, the words 00, 111 and 010, which do not occur in the intersection $\mathcal{S}_{(1,2)}(F) \cap \mathcal{S}_{(0,3)}(F) = \{(110)^\infty, (101)^\infty, (011)^\infty\}$ are displayed in grey (Kůrka [25]).

Example 14 (A multiplication rule) $(\mathbf{4}^{\mathbb{Z}}, F)$, where

$$F(x)_i = \left(2x_i + \left\lfloor \frac{x_{i+1}}{2} \right\rfloor\right) \bmod 4$$
$$= 2x_i + \left\lfloor \frac{x_{i+1}}{2} \right\rfloor - 4\left\lfloor \frac{x_i}{2} \right\rfloor .$$

00	01	02	03	10	11	12	13	20	21	22	23	30	31	32	33
0	0	1	1	2	2	3	3	0	0	1	1	2	2	3	3

We have

$$F^2(x)_i = \left(4x_i + 2 \cdot \left\lfloor \frac{x_{i+1}}{2} \right\rfloor + (x_{i+1}) \bmod 4\right) \bmod 2$$
$$= x_{i+1} = \sigma(x)_i .$$

Thus, the CA is a "square root" of the shift map. It is bijective and expansive, and its entropy is $\ln 2$. The system expresses multiplication by two in base four. If $x \in A^{\mathbb{Z}}$ is left-asymptotic with 0^∞, then $\varphi(x) = \sum_{i=-\infty}^{i=\infty} x_i 4^{-i}$ is finite and $\varphi(F(x)) = 2\varphi(x)$.

Example 15 (A surjective rule) $(\mathbf{4}^{\mathbb{Z}}, F)$, $m = 0$, $a = 1$, and the local rule is

00	11	22	33	01	02	12	21	03	10	13	30	20	23	31	32
0	0	0	0	1	1	1	1	2	2	2	2	3	3	3	3

The system is surjective but not closing. The first image automaton is in Fig. 17 top. We see that $F(\mathbf{4}^{\mathbb{Z}})$ is the full shift, so F is surjective. The configuration $0^\infty.1^\infty$ has left-asymptotic preimages $0^\infty.(21)^\infty$ and $0^\infty.(12)^\infty$, so F is not right-closing. This configuration has also right-asymptotic preimages $0^\infty.(12)^\infty$ and $2^\infty.(12)^\infty$, so F is not left-closing (Fig. 17 bottom). Therefore, $\mathcal{X}^-(F) = \mathcal{X}^+(F) = \emptyset$.

Example 16 $(\mathbf{2}^{\mathbb{Z}} \times \mathbf{2}^{\mathbb{Z}}, \mathrm{Id} \times \sigma)$, *i. e.*, $F(x, y)_i = (x_i, y_{i+1})$.

The system is bijective and sensitive but not transitive. $\mathfrak{A}(F) = \mathfrak{E}(F) = \emptyset$, $\mathcal{X}^-(F) = (-\infty, 1)$, $\mathcal{X}^+(F) = (0, \infty)$. There are infinite signal subshifts

$$\mathcal{S}_{(0,n)} = \mathbf{2}^{\mathbb{Z}} \times |\mathcal{O}_\sigma^n|, \quad \mathcal{S}_{(-n,n)} = |\mathcal{O}_\sigma^n| \times \mathbf{2}^{\mathbb{Z}}$$

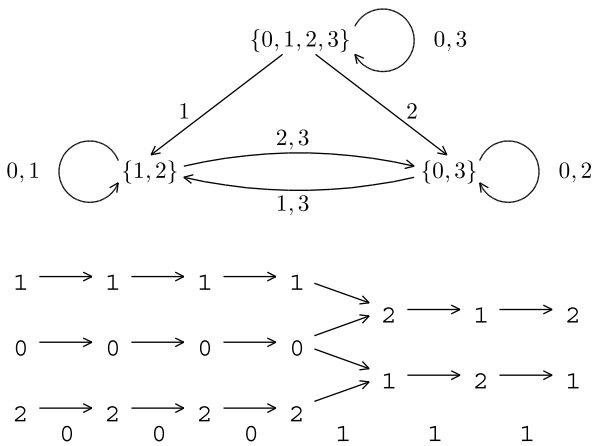

Topological Dynamics of Cellular Automata, Figure 17
Asymptotic configurations

where $|\mathcal{O}_\sigma^n| = \{x \in \mathbf{2}^{\mathbb{Z}} : \sigma^n(x) = x\}$, so the speed subshifts are $\mathcal{S}_0(F) = \mathcal{S}_{-1}(F) = A^{\mathbb{Z}}$.

Example 17 (A bijective CA) $(A^{\mathbb{Z}}, F)$, where $A = \{000, 001, 010, 011, 100\}$, and

$$F(x, y, z)_i = (x_i, (1 + x_i)y_{i+1} + x_{i-1}z_i,$$
$$(1 + x_i)z_{i-1} + x_{i+1}y_i) \bmod 2 .$$

The dynamics is conveniently described as movement of three types of particles, $1 = 001$, $2 = 010$ and $4 = 100$. Letter $0 = 000$ corresponds to an empty cell and $3 = 011$ corresponds to a cell occupied by both $1 = 001$ and $2 = 010$. The particle $4 = 100$ is a wall which neither moves nor changes. Particle 1 goes to the left and when it hits a wall 4, it changes to 2. Particle 2 goes to the right and when it hits a wall, it changes to 1. Clearly 4 is a 1-blocking word, so the system is almost equicontinuous. It is bijective and its inverse is

$$F^{-1}(x, y, z)_i = (x_i, (1 + x_i)y_{i-1} + x_{i+1}z_i,$$
$$(1 + x_i)z_{i+1} + x_{i-1}y_i) \bmod 2 .$$

The first column subshift is $\Sigma_1(F) = \{0, 1, 2, 3\}^{\mathbb{N}} \cup \{4^\infty\}$. We have infinite signal subshifts $\mathcal{S}_{(-1,1)}(F) = \{0, 1\}^{\mathbb{Z}}$, $\mathcal{S}_{(1,1)}(F) = \{0, 2\}^{\mathbb{Z}}$, $\mathcal{S}_{(0,1)}(F) = \{0, 4\}^{\mathbb{Z}}$. For $q > 0$ we get

$$\mathcal{S}_{(0,q)}(F) = \{x \in A^{\mathbb{Z}} : \forall u \in \mathbf{4}^*, (4u4 \sqsubseteq x \Longrightarrow$$
$$(\exists m, 2m|u| = q) \text{ or } u \in \{0\}^*)\}$$

so the speed subshift is $\mathcal{S}_0(F) = A^{\mathbb{Z}}$. The equicontinuity directions are $\mathfrak{A}(F) = \{0\}$, $\mathfrak{E}(F) = \emptyset$. The expansivity directions are $\mathcal{X}^-(F) = (-\infty, -1)$, $\mathcal{X}^+(F) = (1, \infty)$.

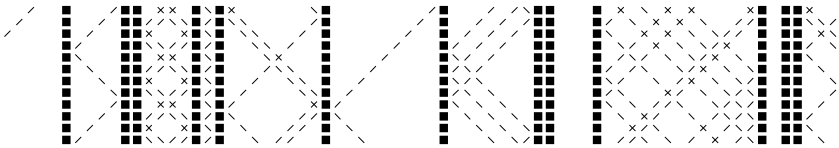

Topological Dynamics of Cellular Automata, Figure 18
A bijective almost equicontinuous CA

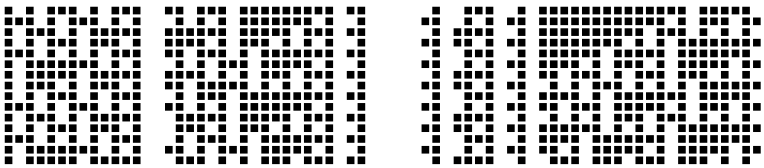

Topological Dynamics of Cellular Automata, Figure 19
The Coven CA

Example 18 (The Coven CA, Coven and Hedlund [13], Coven [12]) $(\mathbf{2}^{\mathbb{Z}}, F)$ where $F(x)_i = x_i + x_{i+1}(x_{i+2} + 1) \bmod 2$.

The CA is left-permutive with zero memory. It is not right-closing, since it does not have a constant number of preimages: $F^{-1}(0^\infty) = \{0^\infty\}$, $F^{-1}(1^\infty) = \{(01)^\infty, (10)^\infty\}$. It is almost equicontinuous with 2-blocking word 000 with offset 0. It is not transitive but it is chain-transitive and its unique attractor is the full space (Blanchard and Maass [4]). While $\Sigma_1(F) = \mathbf{2}^{\mathbb{Z}}$, the two-column factor subshift

$$\Sigma_2(F) = \{10, 11\}^{\mathbb{N}} \cup \{11, 01\}^{\mathbb{N}} \cup \{01, 00\}^{\mathbb{N}}$$

is sofic but not SFT and the entropy is $\mathbf{h}(A^{\mathbb{Z}}, F) = \ln 2$. For any $a, b \in \mathbf{2}$ we have $f(1a1b) = 1c$ where $c = a + b + 1$ (here f is the local rule and the addition is modulo 2). Define a CA $(\mathbf{2}^{\mathbb{Z}}, G)$ by $G(x)_i = (x_i + x_{i+1} + 1) \bmod 2$ and a map $\varphi\colon \mathbf{2}^{\mathbb{Z}} \to \mathbf{2}^{\mathbb{Z}}$ by $\varphi(x)_{2i} = 1$, $\varphi(x)_{2i+1} = x_i$. Then $\varphi\colon (\mathbf{2}^{\mathbb{Z}}, G) \to (\mathbf{2}^{\mathbb{Z}}, F)$ is an injective morphism and $(\mathbf{2}^{\mathbb{Z}}, G)$ is a transitive subsystem of $(\mathbf{2}^{\mathbb{Z}}, F)$. If $x_i = 0$ for all $i > 0$ and $x_{2i} = 1$ for all $i \le 0$, then $(\overline{\mathcal{O}(x)}, F)$ is conjugated to the adding machine with periodic structure $\mathbf{n} = (2, 2, 2, \dots)$, and $I_n^+((10)^\infty) = 2$. We have $\mathfrak{E}(F) = \emptyset$, $\mathfrak{A}(F) = \{0\}$, $\mathfrak{X}^-(F) = (-\infty, 0)$ and $\mathfrak{X}^+(F) = \emptyset$. We have $\mathcal{S}_0(F) = \mathbf{2}^{\mathbb{Z}}$ and there exists an increasing sequence of non-transitive infinite signal subshifts $\mathcal{S}_{(0,2^n)}(F)$:

$$\begin{aligned}\mathcal{S}_{(0,1)}(F) &= S_{\{10\}} \subset \mathcal{S}_{(0,2)}(F)\\ &= S_{\{1010,1110\}} \subset \mathcal{S}_{(0,4)}(F) \subset \cdots .\end{aligned}$$

Example 19 (Gliders and Walls, Milnor [31]) The alphabet is $A = \{0, 1, 2, 3\}$, $F(x)_i = f(x_{[i-1,i+1]})$, where the local rule is

$$\begin{aligned}&x3x : 3, \quad 12x : 3, \quad 1x2 : 3, \quad x12 : 3, \quad 1xx : 1,\\ &\qquad x1x : 0, \quad xx2 : 2, \quad x2x0.\end{aligned}$$

Directional entropy has discontinuity at $\alpha = 0$ (see Fig. 20).

Example 20 (Golden mean subshift attractor: ECA132)

$$\begin{aligned}&000 : 0, \quad 001 : 0, \quad 010 : 1, \quad 011 : 0, \quad 100 : 0,\\ &\qquad 101 : 0, \quad 110 : 0, \quad 111 : 1.\end{aligned}$$

While the golden mean subshift $S_{\{11\}}$ is the finite time maximal attractor of ECA12 (see Example 3), it is also an infinite time subshift attractor of ECA132. The clopen set $U := [00]_0 \cup [01]_0 \cup [10]_0$ is spreading and $\Omega_F = S_{\{11\}} = \mathcal{S}_{(0,1)}$. There exist infinite signal subshifts $\mathcal{S}_{(1,1)}(F) = S_{\{10\}}$, $\mathcal{S}_{(-1,1)}(F) = S_{\{01\}}$ and the maximal attractor is their join $\Omega_F = \mathcal{S}_{(1,1)}(F) \overset{2}{\vee} \mathcal{S}_{(0,1)}(F) \overset{2}{\vee} \mathcal{S}_{(-1,1)}(F) = S_{\{11\}} \cup (\mathcal{S}_{(1,1)}(F) \overset{2}{\vee} \mathcal{S}_{(-1,1)}(F))$.

Example 21 (A finite time sofic maximal attractor) $(\mathbf{2}^{\mathbb{Z}}, F)$, where $m = -1$, $a = 2$ and the local rule is

$$\begin{aligned}&0000 : 0, \quad 0001 : 0, \quad 0010 : 0, \quad 0011 : 1,\\ &0100 : 1, \quad 0101 : 0, \quad 0110 : 1, \quad 0111 : 1,\\ &1000 : 1, \quad 1001 : 1, \quad 1010 : 0, \quad 1011 : 1,\\ &1100 : 1, \quad 1101 : 0, \quad 1110 : 0, \quad 1111 : 0.\end{aligned}$$

The system has finite time maximal attractor $\Omega_F = F(\mathbf{2}^{\mathbb{Z}}) = S_{\{01^{2n+1}0:\, n \ge 0\}}$ (Fig. 22 left). This is the **even**

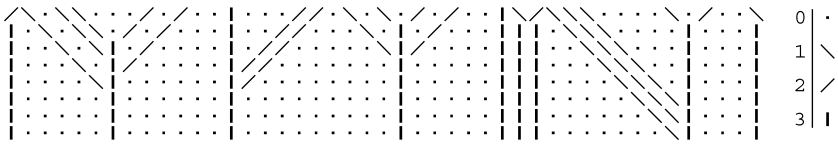

$$\mathbf{h}_\alpha(A^{\mathbb{Z}}, F) = \begin{cases} (1+|\alpha|)\ln 2 & \text{for} \quad \alpha \neq 0 \\ 2\ln 2 & \text{for} \quad \alpha = 0 \end{cases}$$

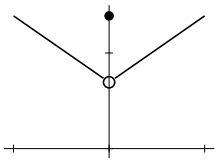

Topological Dynamics of Cellular Automata, Figure 20
Directional entropy of Gliders and walls

Topological Dynamics of Cellular Automata, Figure 21
ECA132

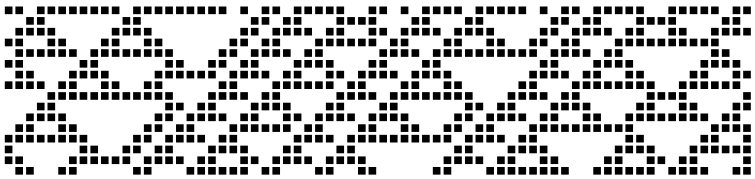

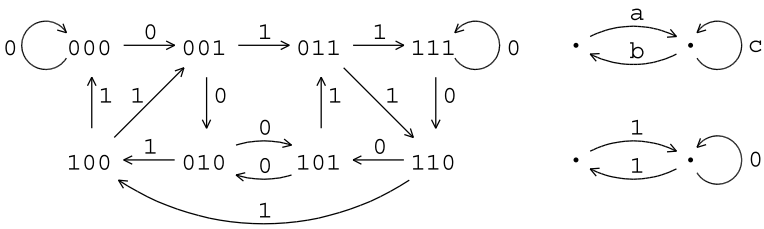

Topological Dynamics of Cellular Automata, Figure 22
The first image graph and the even subshift

| | | |

```
3 0 3 0 0 3 0 0 0 3 0 0 0 0 3 1 1 0 2 3 0 0 0 0 0 3 0 0 0 0 0 3 0 3
3 1 3 0 1 3 0 0 1 3 0 0 0 1 3 1 1 1 1 3 0 0 0 0 1 3 0 0 0 0 1 3 1 3
3 2 3 0 2 3 0 0 2 3 0 0 0 2 3 1 1 1 2 3 0 0 0 0 2 3 0 0 0 0 2 3 2 3
3 1 3 1 1 3 0 1 1 3 0 0 1 1 3 1 1 2 1 3 0 0 0 1 1 3 0 0 0 1 0 3 1 3
3 2 3 1 2 3 0 1 2 3 0 0 1 2 3 1 2 0 2 3 0 0 0 1 2 3 0 0 0 1 1 3 2 3
3 1 3 2 1 3 0 2 1 3 0 0 2 1 3 2 0 1 1 3 0 0 0 2 1 3 0 0 0 1 1 3 1 3
3 1 3 0 2 3 1 0 2 3 0 1 0 1 3 0 0 1 2 3 0 0 1 0 2 3 0 0 0 1 2 3 2 3
3 2 3 1 1 3 1 1 1 3 0 1 0 2 3 0 0 2 1 3 0 0 1 1 1 3 0 0 0 2 0 3 1 3
3 1 3 1 2 3 1 1 2 3 0 1 1 1 3 0 1 0 2 3 0 0 1 1 2 3 0 0 1 0 1 3 2 3
3 2 3 2 1 3 1 2 1 3 0 1 1 2 3 0 1 1 1 3 0 0 1 2 1 3 0 0 1 0 1 3 1 3
3 0 3 0 2 3 2 0 2 3 0 1 2 1 3 0 1 1 2 3 0 0 2 0 2 3 0 0 1 0 2 3 2 3
3 1 3 1 0 3 0 1 1 3 0 2 0 2 3 0 1 2 1 3 0 1 0 1 1 3 0 0 1 1 0 3 1 3
3 2 3 1 1 3 0 1 2 3 1 0 1 1 3 0 2 0 2 3 0 1 0 1 2 3 0 0 1 1 1 3 2 3
3 1 3 1 2 3 0 2 1 3 1 0 1 2 3 1 0 1 1 3 0 1 0 2 1 3 0 0 1 1 1 3 1 3
3 2 3 2 1 3 1 0 2 3 1 0 2 1 3 1 0 1 2 3 0 1 1 0 2 3 0 0 1 1 2 3 2 3
3 0 3 0 2 3 1 1 1 3 1 1 0 2 3 1 0 2 1 3 0 1 1 1 1 3 0 0 1 2 0 3 1 3
```

Topological Dynamics of Cellular Automata, Figure 24
Sensitive system with logarithmic perturbation speeds

subshift whose minimal right-resolving presentation is in Fig. 22 right. We have $E = \{a, b, c\}$, $l(a) = l(b) = 1$, $l(c) = 0$. A word is synchronizing in $\mathcal{G}$ (and intrinsically synchronizing) if it contains 0. The factor map ℓ is right-resolving and also left-resolving. Thus ℓ is left-closing and the even subshift is AFT. We have $\ell^{-1}(1^\infty) = \{(ab)^\infty, (ba)^\infty\}$, and $|\ell^{-1}(x)| = 1$ for each $x \neq 1^\infty$. Thus, the even subshift is near-Markov, and it cannot be an infinite time maximal attractor.

Example 22 (Logarithmic perturbation speeds) $(\mathbf{4}^{\mathbb{Z}}, F)$ where $m = 0$, $a = 1$, and the local rule is

$$00:0,\quad 01:0,\quad 02:1,\quad 03:1,\quad 10:1,\quad 11:1,\quad 12:2,\quad 13:2,$$

$$20:0,\quad 21:0,\quad 22:1,\quad 23:1,\quad 30:3,\quad 31:3,\quad 32:3,\quad 33:3\,.$$

The letter 3 is a 1-blocking word. Assume $x_i = 3$ for $i > 0$ and $x_i \neq 3$ for $i \leq 0$. If $\varphi(x) = \sum_{i=1}^{\infty} x_{-i} \cdot 2^i$ is finite, then $\varphi(F(x)) = \varphi(x) + 1$. Thus, $(\mathcal{O}(x), F)$ is conjugated to the adding machine with periodic structure $\mathbf{n} = (2, 2, 2, \dots)$, although the system is not left-permutive. If $x = 0^\infty.3^\infty$, then for any $i < 0$, $(F^n(x)_i)_{n\geq 0}$ has preperiod $-i$ and period 2^{-i}. For the zero configuration we have

$$n < 2^s + s \Longrightarrow F^n(W_s^-(0^\infty)) \subseteq W_0^-(0^\infty)$$
$$2^{s-1} + s - 1 \leq n < 2^s + s \Longrightarrow I_n^-(0^\infty) = s$$

and therefore $\log_2 n - 1 < I_n^-(0^\infty) < \log_2 n + 1$. More generally, for any $x \in \{0, 1, 2\}^{\mathbb{Z}}$ we have $\lim_{n\to\infty} I_n^-(x)/\log_2 n = 1$.

Example 23 (A sensitive CA with logarithmic perturbation speeds) $(\mathbf{4}^{\mathbb{Z}}, F)$ where $m = 0$, $a = 2$ and the local rule is

$$33x:0,\quad 032:0,\quad 132:1,\quad 232:0,\quad 02x:1,\quad 03x:1,$$

$$12x:2,\quad 13x:2,\quad 20x:0,\quad 21x:0,\quad 22x:1,\quad 23x:1\,.$$

A similar system is constructed in Bressaud and Tisseur [9]. If $i < j$ are consecutive sites with $x_i = x_j = 3$, then $F^n(x)_{(i,j)}$ acts as a counter machine whose binary value is increased by one unless $x_{j+1} = 2$:

$$B_{ij}(x) = \sum_{k=1}^{j-i-1} x_{i+k} \cdot 2^{j-i-1-k}$$
$$B_{ij}(F(x)) = B_{ij}(x) + 1 - \xi_2(x_{j+1}) - 2^{j-i-1} \cdot \xi_2(x_{i+1})$$

Here $\xi_2(2) = 1$ and $\xi_2(x) = 0$ for $x \neq 2$. If $x \in \{0, 1, 2\}$, then $\lim_{n\to\infty} I_n^-(x)/\log_2(n) = 1$. For periodic configurations which contain 3, Lyapunov exponents are positive. We have $\lambda^-((30^n)^\infty) \approx 2^{-n}$.

Future Directions

There are two long-standing problems in topological dynamics of cellular automata. The first is concerned with expansivity. A positively expansive CA is conjugated to a one-sided full shift (Theorem 29). Thus, the dynamics of this class is completely understood. An analogous assertion claims that bijective expansive CA are conjugated to two-sided full shifts or at least to two-sided subshifts of finite type (Conjecture 30). Some partial results have been obtained in Nasu [35].

Another open problem is concerned with chaotic systems. A dynamical system is called chaotic, if it is topologically transitive, sensitive, and has a dense set of periodic

points. Banks et al. [3] proved that topological transitivity and density of periodic points imply sensitivity, provided the state space is infinite. In the case of cellular automata, transitivity alone implies sensitivity (Codenotti and Margara [10] or a stronger result in Moothathu [32]). By Conjecture 15, every transitive CA (or even every surjective CA) has a dense set of periodic points. Partial results have been obtained by Boyle and Kitchens [6], Blanchard and Tisseur [5], and Acerbi et al. [1]. See Boyle and Lee [7] for further discussion.

Interesting open problems are concerned with topological entropy. For CA with non-negative memory, the entropy of a CA can be obtained as the entropy of the column subshift whose width is the radius (Proposition 54). For the case of negative memory and positive anticipation, an analogous assertion would say that the entropy of the CA is the entropy of the column subshift of width $2r + 1$ (Conjecture 55). Some partial results have been obtained in Di Lena [28]. Another aspect of information flow in CA is provided by Lyapunov exponents which have many connections with both topological and measure-theoretical entropies (see Bressaud and Tisseur [5] or Pivato ► Ergodic Theory of Cellular Automata). Conjecture 11 states that each sensitive CA has a configuration with a positive Lyapunov exponent.

Acknowledgments

I thank Marcus Pivato and Mathieu Sablik for careful reading of the paper and many valuable suggestions. The research was partially supported by the Research Program CTS MSM 0021620845.

Bibliography

Primary Literature

1. Acerbi L, Dennunzio A, Formenti E (2007) Shifting and lifting of a cellular automaton. In: Computational logic in the real world. Lecture Notes in Computer Sciences, vol 4497. Springer, Berlin, pp 1–10
2. Akin E, Auslander J, Berg K (1996) When is a transitive map chaotic? In: Bergelson, March, Rosenblatt (eds) Conference in Ergodic Theory and Probability. de Gruyter, Berlin, pp 25–40
3. Banks J, Brook J, Cains G, Davis G, Stacey P (1992) On Devaney's definition of chaos. Am Math Monthly 99:332–334
4. Blanchard F, Maass A (1996) Dynamical behaviour of Coven's aperiodic cellular automata. Theor Comput Sci 291:302
5. Blanchard F, Tisseur P (2000) Some properties of cellular automata with equicontinuous points. Ann Inst Henri Poincaré 36:569–582
6. Boyle M, Kitchens B (1999) Periodic points for onto cellular automata. Indag Math 10(4):483–493
7. Boyle M, Lee B (2007) Jointly periodic points in cellular automata: computer explorations and conjectures. (manuscript)
8. Boyle M, Lind D (1997) Expansive subdynamics. Trans Am Math Soc 349(1):55–102
9. Bressaud X, Tisseur P (2007) On a zero speed cellular automaton. Nonlinearity 20:1–19
10. Codenotti B, Margara L (1996) Transitive cellular automata are sensitive. Am Math Mon 103:58–62
11. Courbage M, Kamiński B (2006) On the directional entropy of Z^2-actions generated by cellular automata. J Stat Phys 124(6):1499–1509
12. Coven EM (1980) Topological entropy of block maps. Proc Am Math Soc 78:590–594
13. Coven EM, Hedlund GA (1979) Periods of some non-linear shift registers. J Comb Theor A 27:186–197
14. Coven EM, Pivato M, Yassawi R (2007) Prevalence of odometers in cellular automata. Proc Am Math Soc 815:821
15. Culik K, Hurd L, Yu S (1990) Computation theoretic aspects of cellular automata. Phys D 45:357–378
16. Formenti E, Kůrka P (2007) A search algorithm for the maximal attractor of a cellular automaton. In: Thomas W, Weil P (eds) STACS Lecture Notes in Computer Science, vol 4393. Springer, Berlin, pp 356–366
17. Formenti E, Kůrka P (2007) Subshift attractors of cellular automata. Nonlinearity 20:105–117
18. Gilman RH (1987) Classes of cellular automata. Ergod Theor Dynam Syst 7:105–118
19. Hedlund GA (1969) Endomorphisms and automorphisms of the shift dynamical system. Math Syst Theory 3:320–375
20. Hopcroft JE, Ullmann JD (1979) Introduction to Automata Theory, Languages and Computation. Addison-Wesley, London
21. Hurd LP (1990) Recursive cellular automata invariant sets. Complex Syst 4:119–129
22. Hurley M (1990) Attractors in cellular automata. Ergod Theor Dynam Syst 10:131–140
23. Kitchens BP (1998) Symbolic Dynamics. Springer, Berlin
24. Kůrka P (2003) Topological and symbolic dynamics, Cours spécialisés, vol 11. Société Mathématique de France, Paris
25. Kůrka P (2005) On the measure attractor of a cellular automaton. Discret Contin Dyn Syst 2005:524–535; supplement
26. Kůrka P (2007) Cellular automata with infinite number of subshift attractors. Complex Syst 17:219–230
27. Kůrka P (1997) Languages, equicontinuity and attractors in cellular automata. Ergod Theor Dynam Syst 17:417–433
28. Lena PD (2007) Decidable and computational properties of cellular automata, Ph D thesis. Universita de Bologna e Padova, Bologna
29. Lind D, Marcus B (1995) An Introduction to Symbolic Dynamics and Coding. Cambridge University Press, Cambridge
30. Maass A (1995) On the sofic limit set of cellular automata. Ergod Theor Dyn Syst 15:663–684
31. Milnor J (1988) On the entropy geometry of cellular automata. Complex Syst 2(3):357–385
32. Moothathu TKS (2005) Homogenity of surjective cellular automata. Discret Contin Dyn Syst 13(1):195–202
33. Moothathu TKS (2006) Studies in topological dynamics with emphasis on cellular automata, Ph D thesis. University of Hyderabad, Hyderabad
34. Nasu M (1995) Textile Systems for Endomorphisms and Automorphisms of the Shift. Mem Am Math Soc 546
35. Nasu M (2006) Textile systems and one-sided resolving automorphisms and endomorphisms of the shift. American Mathematical Society, Providence

36. von Neumann J (1951) The general and logical theory of automata. In: Jefferss LA (ed) Cerebral Mechanics of Behaviour. Wiley, New York
37. Sablik M (2006) Étude de l'action conjointe d'un automate cellulaire et du décalage: une approche topologique et ergodique, Ph D thesis. Université de la Mediterranée, Marseille
38. Sablik M (2007) Directional dynamics of cellular automata: a sensitivity to initial conditions approach. Theor Comput Sci 400(1–3):1–18
39. Shereshevski MA, Afraimovich VS (1992) Bipermutive cellular automata are topologically conjugated to the one-sided Bernoulli shift. Random Comput Dynam 1(1):91–98
40. Smillie J (1988) Properties of the directional entropy function for cellular automata. In: Dynamical systems, vol 1342 of Lecture Notes in Mathematics. Springer, Berlin, pp 689–705
41. Wolfram S (1984) Computation theory of cellular automata. Comm Math Phys 96:15–57
42. Wolfram S (1986) Theory and Applications of Cellular Automata. World Scientific, Singapore

Books and Reviews

Burks AW (1970) Essays on Cellular automata. University of Illinois Press, Chicago
Delorme M, Mazoyer J (1998) Cellular automata: A parallel Model. Kluwer, Amsterdam
Demongeot J, Goles E, Tchuente M (1985) Dynamical systems and cellular automata. Academic Press, New York
Farmer JD, Toffoli T, Wolfram S (1984) Cellular automata. North-Holland, Amsterdam
Garzon M (1995) Models of Massive Parallelism: Analysis of Cellular automata and Neural Networks. Springer, Berlin
Goles E, Martinez S (1994) Cellular automata, Dynamical Systems and Neural Networks. Kluwer, Amsterdam
Goles E, Martinez S (1996) Dynamics of Complex Interacting Systems. Kluwer, Amsterdam
Gutowitz H (1991) Cellular Automata: Theory and Experiment. MIT Press/Bradford Books, Cambridge Mass. ISBN 0-262-57086-6
Kitchens BP (1998) Symbolic Dynamics. Springer, Berlin
Kůrka P (2003) Topological and symbolic dynamics. Cours spécialisés, vol 11. Société Mathématique de France, Paris
Lind D, Marcus B (1995) An Introduction to Symbolic Dynamics and Coding. Cambridge University Press, Cambridge
Macucci M (2006) Quantum Cellular automata: Theory, Experimentation and Prospects. World Scientific, London
Manneville P, Boccara N, Vichniac G, Bidaux R (1989) Cellular automata and the modeling of complex physical systems. Springer, Berlin
Moore C (2003) New Constructions in Cellular automata. Oxford University Press, Oxford
Toffoli T, Margolus N (1987) Cellular Automata Machines: A New Environment for Modeling. Mass MIT Press, Cambridge
von Neumann J (1951) The general and logical theory of automata. In: Jefferss LA (ed) Cerebral Mechanics of Behaviour. Wiley, New York
Wolfram S (1986) Theory and applications of cellular automata. World Scientific, Singapore
Wolfram S (2002) A new kind of science. Media Inc, Champaign
Wuensche A, Lesser M (1992) The Global Dynamics of Cellular Automata. In: Santa Fe Institute Studies in the Sciences of Complexity, vol 1. Addison-Wesley, London

Topological Magnetohydrodynamics and Astrophysics

Mitchell A. Berger[1,2]
[1] Mathematics, University of Exeter, Devon, UK
[2] Mullard Space Science Laboratory, UCL, London, UK

Article Outline

Glossary
Definition of the Subject
Introduction
The Topology of Curves
Magnetohydrodynamics
Magnetic Helicity
Applications to Astrophysical Fields
Future Directions
Bibliography

Glossary

Corona The atmosphere of the sun or a star. The solar corona is generally much hotter (at temperatures up to 2 million K) than the photosphere (solar surface), due to magnetic heating.

Crossing number The projection of a three-dimensional curve onto a plane will exhibit a certain number of crossings (where the projected curve passes over itself). This is the *crossing number* of the curve; in general the crossing number depends on the orientation of the plane as well as the original three-dimensional curve. The *average crossing number* averages over all possible planar projections. This removes the dependence on orientation, but there is still a dependence on the geometry of the curve. Given arbitrary distortions of a curve, without letting the curve pass through itself, there will be a minimum number of crossings, the *minimum crossing number*. For closed curves (*knots*), and for collections of curves (*links*) the minimum crossing number provides a measure of topological complexity. Crossing numbers can also be defined for vector fields in terms of the crossings seen between all pairs of field lines.

Force-free field A force-free magnetic field does not impart any magnetic forces. In the absence of other forces, such as pressure or gravity, a magnetized fluid in equilibrium will possess a force-free field.

Helicity integral The *helicity integral* $H(\mathbf{V}, \mathbf{W})$ of two vector fields $\mathbf{V}$ and $\mathbf{W}$ measures the net linking of the field lines of $\mathbf{V}$ with those of $\mathbf{W}$. Common exam-

ples in a fluid with magnetic field $\mathbf{B}$ and vorticity $\boldsymbol{\omega}$ include the *kinetic helicity* $H(\boldsymbol{\omega}, \boldsymbol{\omega})$, which measures the self linking of the vorticity ω; the *magnetic helicity* $H(\mathbf{B}, \mathbf{B})$, which measures the self-linking of magnetic field $\mathbf{B}$ with itself; and the *cross-helicity* $H(\mathbf{B}, \boldsymbol{\omega})$, the linking between the magnetic field with the vortex field.

Linking number The *linking number* L of two oriented closed curves measures how much they intertwine about each other. If the two curves are projected onto a plane, L equals one half the number of (signed) crossings. The linking number is invariant to arbitrary deformations of the curves, as long as the two curves do not pass through each other.

Magnetohydrodynamics (MHD) Magnetohydrodynamics studies the evolution of a fluid carrying electrical currents and subjected to magnetic forces. Typical examples are liquid metals and ionized plasmas. In the latter case, MHD neglects small-scale effects, such as those arising from fluctuations about charge neutrality or from details of particle trajectories. *Ideal MHD* assumes a perfectly conducting plasma. The magnetic lines of force in ideal MHD are convected by the fluid motions without slipping through the fluid, breaking, or passing through each other.

Photosphere The surface of the sun. Near the photosphere there is a steep change in pressure, density, and optical depth. Also the temperature reaches a minimum (on average 5 800 K).

Reconnection Reconnection occurs when field lines change their connectivity. In a simple reconnection event two bundles of field lines meet in a small region. The field lines are cut in this region, allowing the two pieces of a field line from the first bundle to connect to two corresponding pieces from the second bundle. Reconnection events in nature may involve several simple events.

Twist number The *twist number* $\mathcal{T}_{\mathrm{w}}$ applies to ribbons and tubes. For ribbons, it measures the extent to which one side of the ribbon rotates about the other. For magnetic flux tubes, it measures how much field lines within the tube rotate about the axis of the tube.

Winding number Consider two curves extending between parallel planes. The *winding number* w measures the angle (divided by 2π) through which the two curves rotate about each other.

Writhe The *writhe number* $\mathcal{W}_{\mathrm{r}}$ is a property of a single curve. It measures three-dimensional geometrical structure, and can change if the curve is distorted. For a ribbon, the writhe of the axis of the ribbon plus the twist of an edge of the ribbon about the axis equals the linking number between the edge and the axis (Călugăreanu theorem).

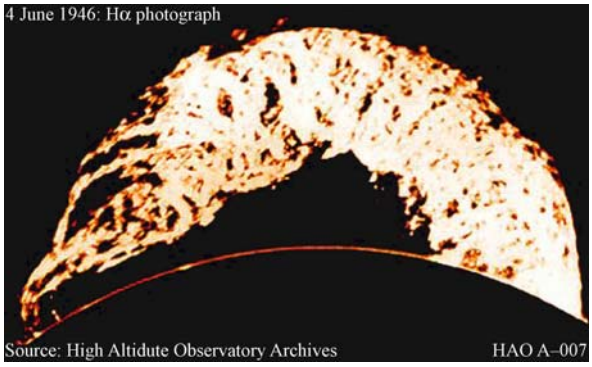

Topological Magnetohydrodynamics and Astrophysics, Figure 1
The "Grand Daddy" erupting prominence of June 1946. The hot plasma in the prominence traces out coiled magnetic field lines. The plasma is photographed in the *Hα* line, sensitive to plasma of temperatures about 20 000 K. The structure extends some 200 000 km above the solar surface (the arc at the bottom)

Definition of the Subject

Many important processes in astrophysics involve magnetic fields, from solar flares and eruptions (Fig. 1) to the formation of galactic jets. Magnetic fields are often highly structured by their *field lines*; a particularly striking example of this can be seen in the fibrous appearance of clouds in the solar atmosphere (Fig. 10). These linear structures can be twisted, kinked, and interlinked. Several geometrical quantities exist which measure the amounts of such structural features. When some of these geometrical quantities are left unchanged by some set of deformations of the field, they are called *topological invariants*. Astrophysicists have paid particular attention to the magnetic helicity integral, because it is approximately conserved in highly conducting plasmas. In our own solar system, magnetic helicity is transported from the interior of the sun, through the solar surface and atmosphere and into interplanetary space. Measurements of this transport assist in the understanding of basic physical processes in these regions.

Introduction

Recent advances in numerical simulations, analytic theory, and astrophysical observations have allowed us to reveal nature in its full three-dimensional glory (plus time as well). Geometrical and topological measures help us understand the complicated three-dimensional structures that we find. Some quantities, such as linking and helicity integrals, can be positive or negative. The sign changes for

a mirror or parity transformation, so these quantities measure handedness as well as topological aspects of a structure. Other quantities, like minimum energy and crossing number, are positive definite, and measure topological complexity without regard to handedness.

Section "The Topology of Curves" will briefly review the kinds of topological invariants relevant to magnetic field studies. Emphasis will be placed on the Gauss linking integral, which provides the template for the magnetic helicity invariant. The next section provides a brief introduction to magnetohydrodynamics, especially concerning the structure of magnetic equilibria. Section "Magnetic Helicity" covers magnetic helicity in some detail. Astrophysical applications of helicity and other topological concepts will be found in Sect. "Applications to Astrophysical Fields". The review ends with a brief discussion of future directions.

The Topology of Curves

Types of Topological Invariants

Figure 2 shows three configurations. In the first, two curves link each other. The Gauss linking integral (described below) detects this linking. For example it evaluates to -3 for the curves shown but will always give zero for unlinked curves. The Gauss integral is an example of an *isotopic* invariant: it does not change no matter how much we distort the curves (as long as the curves are not allowed to pass through each other).

The middle configuration demonstrates a more subtle topological structure. The figure shows the *Borromean rings*, where any two rings are unlinked but the configuration as a whole cannot be taken apart. The Gauss integral is useless here. Isotopic invariants such as 'higher order linking numbers' do exist which detect this topology [10,13,36,42,47,53,64].

The configuration on the right shows two curves extending between parallel planes. Unlike the previous two examples, these are not closed curves: they have end points on boundaries. Nevertheless, we can still find isotopic invariants and complexity measures. For the two curves shown, the *winding number* w gives the net angle (divided by 2π) through which the two curves rotate about each other. The winding number is an isotopic invariant, given the restriction of fixed boundaries (i. e. the endpoints of the curves are fixed on the boundary planes, and the rest of the curves are not allowed to pass through the planes).

We can also look for positive definite numbers which can be minimized over all possible geometrical configurations. In particular, we can count the number of crossings of the curves as seen in a plane projection. If we distort the curves by introducing extra twist and kinks the number of crossings may go up. For the Borromean rings, for example, over the set of all configurations (reachable by arbitrary distortions without curves passing through each other) the minimum number of crossings is six. The *minimum crossing number* is not an isotopic invariant; instead it gives a measure of complexity.

The Gauss Linking Number

Here we give more detail on perhaps the simplest but most powerful isotopic invariant, the Gauss linking number. For more detail and proofs of the properties of the linking number describe below, see [1,15,25,80]. Early in the 19th century Gauss discovered an integral formula that measures how many times two closed curves link each other [28]. Let $\mathbf{x}(\sigma)$ be a point on curve 1 and $\mathbf{y}(\tau)$ a point on curve 2, where σ and τ parametrize curves 1 and 2. Also let $\mathbf{r} = \mathbf{y} - \mathbf{x}$ be the relative position vector.

Definition 1 The Gauss linking integral L_{12} is given by

$$L_{12} = \frac{1}{4\pi} \oint_1 \oint_2 \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\sigma} \cdot \frac{\mathbf{r}}{r^3} \times \frac{\mathrm{d}\mathbf{y}}{\mathrm{d}\tau} \, \mathrm{d}\tau \, \mathrm{d}\sigma \, . \tag{1}$$

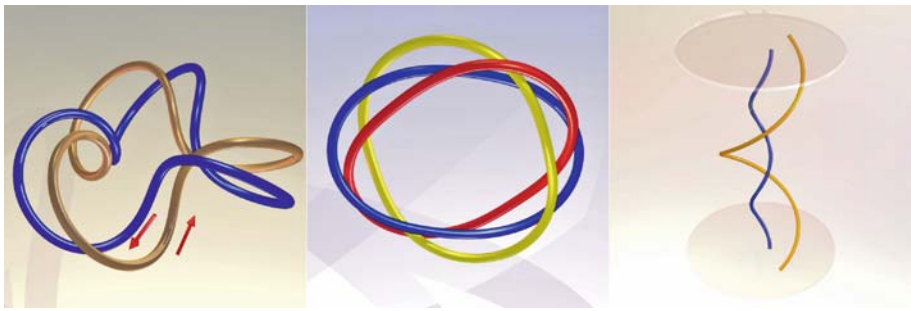

Topological Magnetohydrodynamics and Astrophysics, Figure 2
Left: **two linked curves.** ***Middle:*** **Borromean rings.** ***Right:*** **two braided curves**

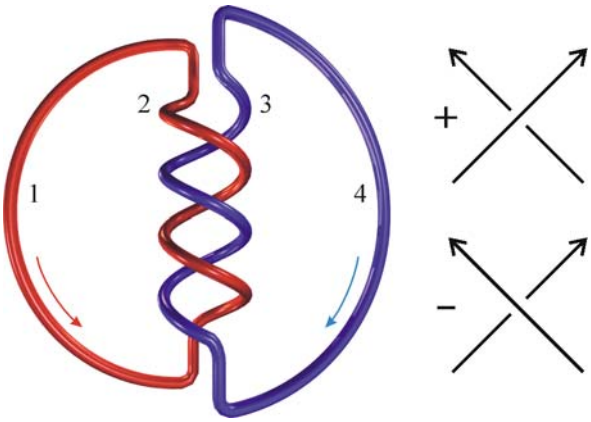

Topological Magnetohydrodynamics and Astrophysics, Figure 3
Two linked curves with four negative crossings. The linking number is $L = -2$

Theorem 1 *The Gauss linking integral gives an integer for any two closed curves. This integer measures how many times (in a right-handed sense) curve 1 crosses a surface bounded by curve 2.*

The Gauss linking integral is invariant to deformations of the two curves, as long as the two curves do not pass through each other. This follows from the fact that the integral counts points of intersection of one curve with a surface bounded by the other. Outside those points of intersection, the curves are free to move without changing the sum. Even if they do move, the sum will not change as long as the curves do not pass through through each other.

A direct evaluation of the Gauss linking integral would be difficult (although, as it always gives an integer a numerical evaluation does not need to be very precise). Fortunately one can evaluate the integral simply by projecting the curves onto a plane and counting the number of crossings (see Fig. 3).

Theorem 2 *Let N_+ and N_- be the numbers of positive and negative crossings of a link as seen in a plane projection. Then*

$$L_{12} = (N_+ - N_-)/2 \,. \tag{2}$$

This theorem shows that linking number can be calculated directly from a picture of the two curves; often we do not need to deal directly with the formidable looking linking integral. Similarly, helicity integrals for vector fields (Sect. "Magnetic Helicity") often can be calculated directly from the morphology of the field, without ever evaluating the integral itself.

Twist and Writhe

The self-helicity of a single flux tube can be decomposed into two contributions called *twist* $\mathcal{T}_\mathrm{w}$ and *writhe* $\mathcal{W}_\mathrm{r}$. First we define twist and writhe for closed curves, then generalize to magnetic flux tubes.

Recall that the Gauss linking integral consists of a double line integral, integrated along two separate curves. Suppose we try doing both line integrals along the same curve?

Definition 2 The *writhe* of a closed curve γ is given by

$$\mathcal{W}_\mathrm{r} = \frac{1}{4\pi} \oint_\gamma \oint_\gamma \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\sigma} \cdot \frac{\mathbf{r}}{r^3} \times \frac{\mathrm{d}\mathbf{y}}{\mathrm{d}\tau} \, \mathrm{d}\tau \, \mathrm{d}\sigma \,. \tag{3}$$

Unlike the linking integral L, however, the writhe $\mathcal{W}_\mathrm{r}$ is not a topological invariant – it changes if we distort the curve.

Suppose we consider a *ribbon*, which can be described mathematically by specifying two almost parallel curves: the axis curve running along the center of the ribbon, and a secondary curve (one of the edges of the ribbon). A DNA molecule provides an important application of the theory of ribbons [1,85]. The two curves may twist about each other; in addition the axis may itself be coiled.

We can define the twist $\mathcal{T}_\mathrm{w}$ as the net amount that the secondary curve twists about the direction of the axis:

Definition 3 Let $\mathbf{T}(s)$ be the tangent vector to the axis curve at arclength s, and let $\hat{\mathbf{v}}$ be a unit vector at s pointing to the secondary curve. Then the *twist number* is

$$\mathcal{T}_\mathrm{w} = \frac{1}{2\pi} \oint \mathbf{T}(s) \cdot \hat{\mathbf{v}}(s) \times \frac{\mathrm{d}\hat{\mathbf{v}}(s)}{\mathrm{d}s} \, \mathrm{d}s \,. \tag{4}$$

If the ribbon closes upon itself, we may calculate the linking number of the two curves. In 1961 Călugăreanu [21] proved a remarkable theorem:

Theorem 3

$$L = \mathcal{T}_\mathrm{w} + \mathcal{W}_\mathrm{r} \,. \tag{5}$$

We can also define twist, writhe, and linking for a tube filled with curves (such as magnetic field lines). For every field line within the tube, we can define a twist number with respect to the central axis (see Fig. 4).

Magnetohydrodynamics

Field Lines

We briefly review magnetohydrodynamics (MHD) [81, 83,84,87], with an emphasis on the geometrical structure

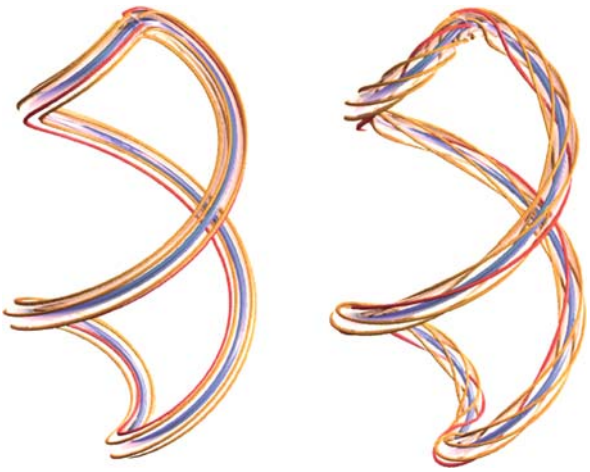

Topological Magnetohydrodynamics and Astrophysics, Figure 4
Two tubes with the same axis geometry. The central axes have a writhe of $\mathcal{W}_r = -0.72$. The left tube has zero internal twist; i. e. $\mathcal{T}_w = 0$. On the right $\mathcal{T}_w = 5$

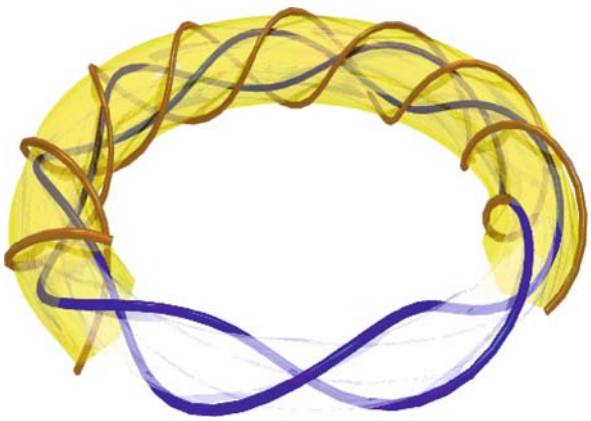

Topological Magnetohydrodynamics and Astrophysics, Figure 5
A magnetic field where the field lines form nested toroidal surfaces. Suppose a field line closes upon itself after encircling the torus *n* times the short way around and *m* times the long way around. Then it has the (rational) periodicity ratio $T = n/m$. Here *T* gives the number of times the field line twists about the toroidal axis in one circuit the long way around

of the *magnetic field lines*. A magnetic field line is a curve which follows the direction of the magnetic field vector **B**: if s denotes arclength along the curve and $\mathbf{x}(s)$ gives the position vector on the curve at s, then

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}s} = \frac{\mathbf{B}}{|\mathbf{B}|} \,. \tag{6}$$

Of course, field lines are infinitely thin, and there are infinitely many of them. Often people prefer to consider bundles of field lines, thin enough to be more or less coherent over their length, called flux elements or *flux tubes*. In many cases a field can be approximated as a finite set of flux tubes [80]. Often topological quantities like magnetic helicity can be calculated directly from an examination of the field line or flux tube structure, without ever directly seeing the equations for **B**.

The Maxwell Equations

Let us start with the Maxwell equations (e. g. [86]). The internal equations involve the electric and magnetic fields **B** and **E** only; matter terms involving charges and currents do not appear.

$$\nabla \cdot \mathbf{B} = 0 \,, \tag{7}$$

$$\nabla \times \mathbf{E} + \partial_t \mathbf{B} = 0 \,. \tag{8}$$

Equation (7) says that there are no magnetic monopoles. This lack of source charges has a profound topological consequence: lines of magnetic flux have no endpoints. Magnetic field have four options: they can either form closed curves, extend to infinity, wander ergodically within a volume, or wrap around a toroidal surface. The last option is illustrated in Fig. 5. Generally, a magnetic field may fill a torus with a family of nested surfaces, with smoothly varying twist T. When T is irrational, field lines never exactly close upon themselves (although they will come arbitrarily close).

The Maxwell source equations in vacuum with charge density ρ and electrical current **J** (in SI units with permittivity ϵ_0 and permeability μ_0) are

$$\nabla \cdot \epsilon_0 \mathbf{E} = \rho \,, \tag{9}$$

$$\nabla \times \frac{1}{\mu_0}\mathbf{B} - \epsilon_0 \partial_t \mathbf{E} = \mathbf{J} \,. \tag{10}$$

Equation (9) implies that electric field lines start and stop at electric charges. For non-relativistic applications, $\partial_t \mathbf{E}$ is small, and Eq. (10) gives

$$\nabla \times \frac{1}{\mu_0}\mathbf{B} \approx \mathbf{J} \tag{11}$$

i. e. magnetic field lines circle electric currents.

Magnetized Fluids

In magnetohydrodynamics (MHD), we consider a neutral fluid which can carry electrical currents. Physically, this may be liquid metal or a neutral plasma. For basic references emphasizing applications in Tokomaks and other

fusion devices, see [81]. For solar physics applications, see [82] and [78].

Electrical currents tend to dissipate in physical media. In MHD we assume some functional relation between the electric field $\mathbf{E}$ and the electrical current $\mathbf{J}$. The simplest relation is the linear Ohm's law

$$\mathbf{E} = \eta \mathbf{J} \tag{12}$$

where η is the resistivity. Some MHD studies employ a *generalized Ohm's law* with more terms on the right-hand side [79]; for example, the Hall effect can be important for collisionless plasmas in the Earth's magnetosphere [24].

Consider a small region of fluid moving at speed $\mathbf{V}$ with respect to some fixed or laboratory coordinate frame. Because of the assumption of charge neutrality, the electrostatic electric field vanishes in the rest frame of this fluid element. Thus $\mathbf{E} = \eta \mathbf{J}$ in this frame. However, if we Lorentz transform to the fixed frame, the electric field becomes

$$\mathbf{E} = \mathbf{B} \times \mathbf{V} + \eta \mathbf{J} \,. \tag{13}$$

If we assume uniform resistivity η, by the Maxwell equation Eq. (8) we obtain the *magnetic induction equation*

$$\partial_t \mathbf{B} = \nabla \times (\mathbf{V} \times \mathbf{B}) + \eta \nabla^2 \mathbf{B}. \tag{14}$$

The last (dissipation) term shows that a concentration of magnetic field will diffuse into the fluid. Often in astrophysical plasmas this term is small, except near intense electrical currents. If we set this term to zero, then we have the situation of *ideal MHD*. In ideal MHD, magnetic field lines are convected by the fluid like material strings. Field lines cannot be cut or cross through each other. This preserves the topology of the field lines; for example if two closed field lines are linked, then they will stay linked no matter how much the velocity $\mathbf{V}$ moves them around. Thus we have a magnetic analogy to the Helmholtz circulation theorem: consider a closed curve C (which may or may not be a field line) moving with the flow. Then the net magnetic flux encircled by C is a constant of the motion.

Ideal MHD allows us to follow the evolution of field lines in time; in a relativistic description field lines trace out surfaces in space-time [35,37]. When $\eta \neq 0$ then strictly speaking magnetic field lines cannot be unambiguously followed in time; however some understanding of field line evolution is still possible [70]. In regions of intense currents $\eta \nabla^2 \mathbf{B}$ may be large even if η is small. In this case field lines can break and reconnect with each other, thus changing the field topology. This phenomenon of *reconnection* is an active area of research (e. g. [41,70,83]).

In addition to the magnetic induction equation, we need to describe the evolution of the velocity field. The Navier–Stokes equation in the presence of magnetic forces gives

$$\rho \left(\partial_t \mathbf{V} + \mathbf{V} \cdot \nabla \mathbf{V}\right) = -\nabla p + \mathbf{J} \times \mathbf{B} + \nu \nabla^2 \mathbf{V} \tag{15}$$

where p and ρ are the pressure and density of the flow, and ν is the kinematic viscosity.

Magnetostatic Equilibria

Of particular interest are equilibrium solutions where $\mathbf{V} = \partial_t \mathbf{V} = 0$. In this case

$$\mathbf{J} \times \mathbf{B} = \nabla p \,. \tag{16}$$

As $\mu_0 \mathbf{J} = \nabla \times \mathbf{B}$, this is a nonlinear equation for $\mathbf{B}$. The theory of magnetic equilibria is of particular importance in studies of the solar atmosphere, where magnetic structures can last on timescales up to months (in contrast, dynamic timescales such as the wave travel time across a coronal loop may only be seconds) [3,45].

First we consider pressure-balanced equilibria. Note that solutions with non-uniform pressure satisfy $\mathbf{B} \cdot \nabla p = \mathbf{J} \cdot \nabla p = 0$, so field lines and current lines lie within level surfaces of the pressure. A *magnetic surface* is a surface S which no field lines cross, i. e. $\mathbf{B} \cdot \hat{n} = 0$, where $\hat{n}$ is the unit normal. For fields completely contained within a finite magnetic surface, the pressure surfaces must lie on nested torii [23,49]. (Note that a sphere can be filled with nested torii, as in a Hill's vortex; for the outermost torus the central hole degenerates into a line segment extending from the top of the sphere to the bottom.)

Pressure gradients can be neglected in diffuse plasmas (for example in the solar corona typical magnetic forces are three orders of magnitude higher than pressure forces). In this case equilibria take the form of force-free fields [44,50,71]. If pressure $p = 0$ then the electrical current is parallel to the field, so that

$$\nabla \times \mathbf{B} = \lambda(\mathbf{x}) \mathbf{B} \tag{17}$$

for some scalar function $\lambda(\mathbf{x})$ of position $\mathbf{x}$. Taking the divergence of this equation gives $\mathbf{B} \cdot \nabla \lambda = 0$. Hence field lines (and current lines too) lie within level surfaces, this time of the function λ rather than p. Again inside a finite magnetic surface field lines lie on nested torii.

The one exception occurs when λ is uniform in space. In this case field lines have the possibility of ergodically filling a volume. The topology of such fields is of great interest [31,80]. These fields are eigenfunctions of the curl operator, $\nabla \times \mathbf{B} = \lambda \mathbf{B}$. They are sometimes called *linear*

force–free fields because this is a linear equation, and hence much easier to solve than in the case of variable λ.

Force–free solutions minimize the magnetic energy subject to topological constraints: given the set of all magnetic configurations with the same field line topology, local minima in the magnetic energy will be free of magnetic forces. It is usually difficult to determine whether a particular solution is a local or global (true) minimum of the energy. The energetics of magnetic equilibria have application in knot theory. The minimum energy of a magnetic field in the form of a knotted flux tube provides, like minimum crossing number (Sect. "Types of Topological Invariants") a measure of topological complexity [49,50].

Linear force–free fields minimize the energy for all fields with the same magnetic helicity (see Sect. "Magnetic Helicity" below), and the same boundary conditions [8,27,71]. Note that fields with different field line topologies can have the same magnetic helicity. Thus reconnection is generally needed to enable a field to relax to the linear force–free state [34,67].

Magnetic Helicity

Overview

Helicity integrals measure the net interlinking of the field lines of a vector field. Physicists, astronomers, and mathematicians calculate the helicity of magnetic fields in a variety of situations, from plasma experiments to descriptions of galactic magnetic fields. Key properties of helicity include its relations to field structure and its conservation in highly conducting plasmas. Magnetic energy remains the most popular integral quantity employed in analysis of magnetic fields. However, magnetic helicity has some advantages compared to energy. First, energy can convert between several physical forms – ideal (non-resistive) forces exchange energy between the magnetic field and the fluid flow, while non-ideal effects create thermal energy or accelerate non-thermal cosmic ray particles. In contrast, magnetic helicity only lives in the magnetic field; it cannot convert into other quantities. Secondly, magnetic energy can dissipate rapidly during reconnection, whereas helicity is approximately conserved, with rigorous estimates on what 'approximately' means. Helicity conservation will be described in detail in Sects. "Evolution of Helicity" and "Magnetic Helicity and Reconnection". Third, the input of magnetic helicity through the boundaries of a volume can be simpler to observe than energy input. This applies particularly to flow of helicity from the interior of the sun into the solar atmosphere.

First we describe how magnetic helicity measures some simple structural features of the field. Figure 6 shows two linked flux tubes. As in calculations of magnetic induction, we can express the magnetic helicity as a sum of a mutual helicity between the two tubes, and the self helicities of each tube. The mutual helicity expresses the linking of the tubes (Sect. "The Gauss Linking Number"), while the self helicities measure the twist of field lines within the tubes (Sect, "Twist and Writhe"). The self helicity of a tube actually involves not only simple twisting of field lines, but also the writhe of the tube, which measures knotting and kinking of the tube axis [1,15,52].

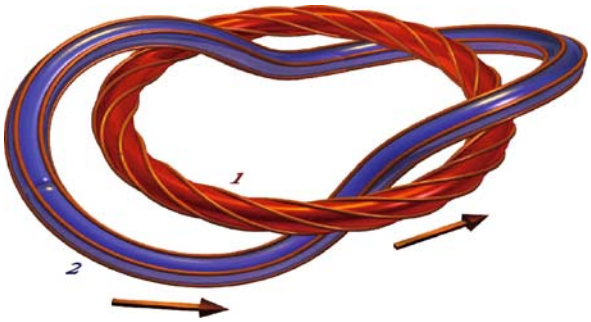

Topological Magnetohydrodynamics and Astrophysics, Figure 6 Two linked flux tubes. Tube 1 has axial flux Φ_1, while tube 2 has flux Φ_2. The tubes link each other twice in a right handed sense, so $H_{12} = H_{21} = 2\Phi_1\Phi_2$, for a total mutual helicity of $4\Phi_1\Phi_2$. Tube 1 has twist $T_1 = 4$ so its self helicity is $4\Phi_1^2$; tube 2 has zero net self helicity. Thus the total helicity of the configuration is $H = 4\Phi_1\Phi_2 + 4\Phi_1^2$

Many applications concern a field residing in some subregion of space. In general, the field lines or flux tubes will cross the boundary of the region. Figure 10 shows hot plasma loops in the solar corona. The plasma traces out the shape of the magnetic field. (Transport of plasma across field lines is much slower than transport along a line. Also, thermal conductivity is high along the field direction. Thus density and temperature can vary from field line to field line, but are roughly uniform along each line.) Each loop has ends (called footpoints) in the surface of the sun (the photosphere). The helicity of fields which cross a boundary such as the photosphere will be discussed in Sect. "Helicity in Open Volumes". Again we can describe the helicity as a sum of mutual and self helicities, but linking number gives way to a more general measure of entanglement.

Helicity Integrals

We will consider the mutual linking of two vector fields $\mathbf{V}$ and $\mathbf{W}$ which reside in a volume $\mathcal{D}$. Any $\mathbf{V}$ or $\mathbf{W}$ lines that cross the boundary $\partial\mathcal{D}$ of the volume form loops outside $\mathcal{D}$. If we are only given information about the field structure inside $\mathcal{D}$, we will not know the shape of these

loops. The loops outside $\partial\mathcal{D}$ may be highly linked! The Gauss integral, if restricted to the parts of the field lines within $\mathcal{D}$, will not detect this external linking. In fact, it will not give a topological invariant. Section "Helicity in Open Volumes" gives a general definition for helicity integrals which eliminates this problem. For now we simply assume all field lines close within $\mathcal{D}$.

The helicity integral $H(\mathbf{V},\mathbf{W})$ measures how much the flux of $\mathbf{V}$ links the flux of $\mathbf{W}$ inside $\mathcal{D}$. The form of the integral mirrors that of the Gauss integral Eq. (1):

Definition 4 (Closed field regions) Consider two closed divergence-free vector fields $\mathbf{V}$ and $\mathbf{W}$ inside a domain $\mathcal{D}$ with boundary $\partial\mathcal{D}$. We assume that the field lines of $\mathbf{V}$ and $\mathbf{W}$ close within $\mathcal{D}$:

$$\mathbf{V}\cdot\hat{\mathbf{n}}|_{\partial\mathcal{D}} = \mathbf{W}\cdot\hat{\mathbf{n}}|_{\partial\mathcal{D}} = 0\,. \tag{18}$$

Then the *helicity integral* $H(\mathbf{V},\mathbf{W})$ is given by

$$H(\mathbf{V},\mathbf{W}) \equiv \frac{1}{4\pi}\int_{\mathcal{D}}\int_{\mathcal{D}} \mathbf{V}(\mathbf{x})\cdot\frac{\mathbf{r}}{r^3}\times\mathbf{W}(\mathbf{y})\mathrm{d}^3x\mathrm{d}^3y\,. \tag{19}$$

We have taken Eq. (1), replaced the field line tangent vectors $\mathrm{d}\mathbf{x}/\mathrm{d}\sigma$ and $\mathrm{d}\mathbf{y}/\mathrm{d}\tau$ with $\mathbf{V}$ and $\mathbf{W}$, and then replaced the line integrals with volume integrals. In effect, the volume integrals sum over all pairs of field lines of $\mathbf{V}$ and $\mathbf{W}$ (weighted by flux) [48,80]. The volume element d^3x can be regarded as the product of a differential line element $\mathrm{d}\sigma$ along a $\mathbf{V}$ field line and an area element d^2x perpendicular to that line. The area element combines with the $\mathbf{V}$ field strength to give the flux weighting.

We will be principally interested in the magnetic helicity $H(\mathbf{B},\mathbf{B})$, which measures the self linking of magnetic flux. But several other quadratic quantities appearing in physics and fluid mechanics can be interpreted as helicity integrals. The kinetic helicity $H(\boldsymbol{\omega},\boldsymbol{\omega})$ measures the self linking of vorticity, while $H(\boldsymbol{\omega},\mathbf{B})$, the cross helicity, measures the mutual linking between vorticity and magnetic flux.

The definition for $H(\mathbf{V},\mathbf{W})$ above is closest in form to the Gauss linking integral, but in practice would be difficult to calculate directly. Fortunately, we can simplify the integral through the use of vector potentials.

Vector Potentials

Definition 5 For a divergence free field $\mathbf{V}$ inside a closed field region $\mathcal{D}$, we define an inverse curl operation by

$$\mathrm{curl}^{-1}\mathbf{W}(\mathbf{x}) \equiv -\frac{1}{4\pi}\int_{\mathcal{D}}\frac{\mathbf{r}}{r^3}\times\mathbf{W}(\mathbf{y})\mathrm{d}^3y\,. \tag{20}$$

A simple calculation shows $\nabla\times\mathrm{curl}^{-1}\mathbf{W} = \mathbf{W}$. The operator inverting the curl within $\mathcal{D}$ is called the modified Biot–Savart operator [85] (the unmodified operator being integrated over all $\mathbb{R}^3$). Then

$$H(\mathbf{V},\mathbf{W}) = \int_{\mathcal{D}}\mathbf{V}\cdot\mathrm{curl}^{-1}\mathbf{W}\mathrm{d}^3x = \int_{\mathcal{D}}\mathrm{curl}^{-1}\mathbf{V}\cdot\mathbf{W}\mathrm{d}^3y\,. \tag{21}$$

For example, when integrated over all space, the magnetic energy $E = \frac{1}{2\mu_0}\int B^2\mathrm{d}^3x$ can be written as a helicity integral, as soon as we realize that $\mathbf{B}$ is a vector potential for the electric current $\mathbf{J}$, $\nabla\times B = \mu_0\mathbf{J}$. Thus we can write

$$E = \tfrac{1}{2}H(\mathbf{B},\mathbf{J}) \tag{22}$$

i. e. the linking of magnetic flux and electric current gives twice the magnetic energy.

Gauge–Invariance of Closed Helicity

Equation (20) gives a unique expression for the inversion of the curl operator. However, for any divergence free vector field $\mathbf{W}$, there exist an infinite number of vector fields $\mathbf{A}_W$ whose curl gives $\nabla\times\mathbf{A}_W = \mathbf{W}$. Suppose we let

$$\mathbf{A}_W = \mathrm{curl}^{-1}\mathbf{W} + \nabla\psi \tag{23}$$

for some gauge function ψ. Transforming a vector field by adding a zero-curl field is called a *gauge transformation*. For closed fields the helicity can readily be shown to be invariant to gauge transformations.

Helicity in Open Volumes

So far we have considered helicity integrals inside closed volumes, where the field lines never cross the boundary. However, this requirement is far too restrictive for many physical and mathematical applications. Solar magnetic fields, for example, cross a natural boundary at the photosphere. Scientists studying the interior of the sun will see this as an outer boundary, while those who study the solar atmosphere regard the photosphere as an inner boundary. We will need a form of helicity integral which works just for the interior, or just for the exterior. Furthermore, the sum of interior and exterior helicities should sensibly relate to the helicity integral of all space. Helicity should be allowed to flow across boundaries as well, just like energy. Laboratory physicists studying confined fusion energy devices also deal with magnetic fields not completely enclosed inside the plasma. In fact, magnetic helicity can be injected into a plasma (to improve its stability) via the field lines that cross the boundary [62]. Mathematicians

also interest themselves in topological objects which do not close upon themselves (unlike knots and links). Tangles and braids, for example, involve curves with fixed endpoints on a boundary [18].

Here we will consider the magnetic helicity $H(\mathbf{B}) = H(\mathbf{B}, \mathbf{B})$ inside an arbitrary volume $\mathcal{D}$. We give a definition of helicity which retains topological meaning and is gauge invariant. First we divide space, and our magnetic field, into pieces inside and outside our domain of interest $\mathcal{D}$.

Definition 6 Let space be divided into domains $\mathcal{D}$ and $\mathcal{D}'$, containing magnetic fields $\mathbf{B}$ and $\mathbf{B}'$. At the boundary $\partial\mathcal{D}$, $\mathbf{B} \cdot \hat{n} = \mathbf{B}' \cdot \hat{n}$. The magnetic field defined in all space is

$$\{\mathbf{B}, \mathbf{B}'\}(\mathbf{x}) = \begin{cases} \mathbf{B}\,, & \mathbf{x} \in \mathcal{D}\,; \\ \mathbf{B}'\,, & \mathbf{x} \in \mathcal{D}'\,. \end{cases} \tag{24}$$

Unfortunately, $H(\{\mathbf{B}, \mathbf{B}'\})$ includes information about all the helical structure in the exterior region $\mathcal{D}'$. We need to subtract this extra information. Simply integrating Eq. (19) or Eq. (21) over $\mathcal{D}$ as before will not do, as $\mathcal{D}$ is no longer a closed field region. The integral will no longer be gauge invariant or topologically meaningful. Instead, we measure the helicity relative to a minimal base state. This procedure is similar to measuring voltage relative to ground, or potential gravitational energy relative to sea level. Thus we will look for some simple vector field $\mathbf{P}$ inside $\mathcal{D}$ for which we can calculate the reference helicity $H(\{\mathbf{P}, \mathbf{B}'\})$. Once we subtract this reference helicity, the dependence on the external field will vanish [14].

The boundary information $\mathbf{B} \cdot \hat{n}$ tells us the distribution of flux crossing the boundary $\partial\mathcal{D}$. It also determines a unique vector field, the vacuum (or potential) field $\mathbf{P}$:

Definition 7 The vacuum (potential) field $\mathbf{P}$ in $\mathcal{D}$ satisfies

$$\mathbf{P} \cdot \hat{\mathbf{n}}|_S = \mathbf{B} \cdot \hat{n}\,; \tag{25}$$

$$\nabla \times \mathbf{P}(\mathbf{x}) = 0\,, \quad \mathbf{x} \in \mathcal{D}\,. \tag{26}$$

If $\mathcal{D}$ is multiply connected, the net flux of $\mathbf{P}$ through any closed curve on $\partial\mathcal{D}$ should also be the same for $\mathbf{B}$ and $\mathbf{P}$.

A simple variational calculation will show that the vacuum field is the minimum energy state consistent with the boundary data $\mathbf{B} \cdot \hat{n}$. Magnetic field lines spiral about current lines. As the vacuum field has zero current, it has minimum helical structure. It also requires a minimum of information for its specification. These qualities make it the ideal choice for the reference field.

Definition 8 The magnetic helicity inside an arbitrary volume $\mathcal{D}$ is given by

$$H_{\mathcal{D}} = H(\{\mathbf{B}, \mathbf{B}'\}) - H(\{\mathbf{P}, \mathbf{B}'\})\,. \tag{27}$$

Theorem 4 [14,29]

1. *$H_{\mathcal{D}}$ is gauge invariant;*
2. *$H_{\mathcal{D}}$ can be expressed as an integral over $\mathcal{D}$ alone:*

$$H_{\mathcal{D}} = \int_{\mathcal{V}} (\mathbf{A} + \mathbf{A_P}) \cdot (\mathbf{B} - \mathbf{P}) d^3x \tag{28}$$

where $\nabla \times \mathbf{A}_P = \mathbf{P}$.

3. *$H_{\mathcal{D}}$ is independent of the field $\mathbf{B}'$ outside of $\mathcal{D}$.*

Self and Mutual Helicity

The total helicity of a collection of magnetic structures can be decomposed into sums of self helicities and mutual helicities. The self helicity of a single structure measures properties of field lines within the structure like twist and kinking. The mutual helicity between two structures measures their linking.

As an example, consider two solar coronal loops 1 and 2 with fluxes Φ_1, Φ_2. We will assume that the loops are uniformly twisted through T_1 and T_2 turns (e. g. a field line rotates about the axis of loop 1 $2\pi T_1$ times). The two loops will then have self-helicities due to their twist

$$H_{11} = T_1 \Phi_1^2\,; \qquad H_{22} = T_2 \Phi_2^2\,. \tag{29}$$

However, two loops will in general also share a mutual helicity if they cross over each other or are misaligned. The total helicity of the two loops is the sum of the self and mutual helicities

$$H = H_{11} + H_{22} + 2H_{12}\,. \tag{30}$$

For the moment, let the corona be the upper half space $\{z > 0\}$. The photospheric ($z = 0$) ends of the loops 1 and 2 are assumed to be small and located at the points $1^+, 1^-, 2^+$, and 2^-. Here $B_z > 0$ at 1^+ and 2^+. If the loops cross as seen from above, we assume that loop 1 is the upper loop. Consider the quadrilateral $1_+2_+1_-2_-$. Let α and β be the angles of at vertices 1_+ and 1_-, respectively (as defined in Fig. 7). Then the helicity integral gives [7]

$$H_{12} = H_{21} = \frac{\Phi_1 \Phi_2}{2\pi}(\alpha + \beta)\,. \tag{31}$$

Next consider N loops. In this case there are N self helicities and $N(N-1)$ mutual helicities:

$$H = \sum_{i=1}^{N} H_{ii} + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} H_{ij}\,. \tag{32}$$

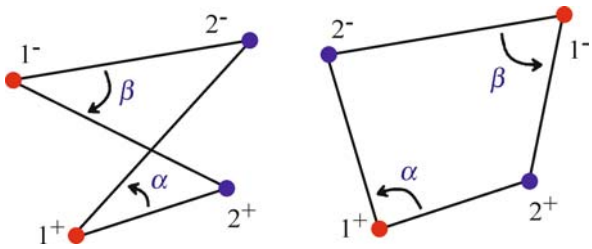

Topological Magnetohydrodynamics and Astrophysics, Figure 7
Definition of the angles in Eq. (31)

Thus the total helicity H equals the sum of the entries in a matrix H_{ij}. If N is large then there will be many more mutual helicity terms. In this case ignoring the self helicities (if they are difficult to observe) may only give a small error.

Helicity of Toroidal and Poloidal Fields

In Cartesian or spherical geometries it is often useful to decompose a magnetic field into toroidal and poloidal components. Let

$$\mathcal{L} \equiv \begin{cases} -\hat{z} \times \nabla, & \text{Cartesian geometry;} \\ -\mathbf{r} \times \nabla, & \text{Spherical geometry.} \end{cases} \tag{33}$$

Then we can write

$$\mathbf{B} = \mathcal{L}T + \nabla \times \mathcal{L}P \,, \tag{34}$$

where T is the toroidal function and P is the poloidal function.

Theorem 5 [8] *Consider a magnetic field in a region $\mathcal{D}$. Assume that $\mathcal{D}$ is either 1) all space, 2) a half space bounded by a plane, 3) a layer bounded by two planes, 4) the interior or exterior of a sphere, or 5) a spherical shell bounded by two concentric spheres. Then inside $\mathcal{D}$*

1. *A purely poloidal field ($T = 0$) has $H = 0$.*
2. *A purely toroidal field ($P = 0$) has $H = 0$.*
3. *In general,*

$$H = 2 \int_{\mathcal{D}} \mathcal{L}T \cdot \mathcal{L}P d^3x \,. \tag{35}$$

Evolution of Helicity

Suppose the plasma inside $\mathcal{D}$ has magnetic diffusivity η and a velocity field $\mathbf{V}$. Then one can derive a Poynting theorem for helicity [7]. First we will need a vector potential for $\mathbf{P}$ which simplifies the expressions (as everything is gauge–invariant we are allowed to choose a convenient gauge to do our calculations). Here $\hat{\mathbf{n}}$ points out of $\mathcal{D}$ at the boundary $\partial\mathcal{D}$.

Definition 9 Let $\mathbf{A_P}$ be the unique vector potential satisfying

$$\nabla \times \mathbf{A_P} = \mathbf{P} \,, \tag{36}$$

$$\nabla \cdot \mathbf{A_P} = 0 \,, \tag{37}$$

$$\hat{\mathbf{n}} \cdot \nabla \times \mathbf{A_P} = B_n \qquad \text{at} \quad \partial\mathcal{D} \,, \tag{38}$$

$$\mathbf{A_P} \cdot \hat{\mathbf{n}} = 0 \qquad \text{at} \quad \partial\mathcal{D} \,. \tag{39}$$

Theorem 6 [6,7,27]

$$\frac{dH}{dt} = -2 \int_{\mathcal{V}} \eta \mathbf{J} \cdot \mathbf{B} d^3x + 2 \oint_{\partial\mathcal{D}} \big((\mathbf{A_P} \cdot \mathbf{V})\mathbf{B} - (\mathbf{A_P} \cdot \mathbf{B})\mathbf{V} \big) \cdot \hat{\mathbf{n}} d^2x \,. \tag{40}$$

The first term gives dissipation of magnetic helicity, while the second gives flow of helicity across the boundary (see Fig. 8).

Magnetic Helicity and Reconnection

The helicity dissipation term in Theorem 6 is generally very small. We can readily compare it to magnetic energy dissipation [7]: The magnetic energy is $E = 1/(2\mu_0) \int B^2 \mathrm{d}^3x$, while magnetic energy dissipates at the rate $\mathrm{d}E/\mathrm{d}t = \int \eta J^2 \mathrm{d}^3x$. Thus a Schwartz inequality gives

$$\left(\frac{\mathrm{d}H}{\mathrm{d}t} \right)^2 \leq 8\mu_0 \eta E \frac{\mathrm{d}E}{\mathrm{d}t} \,. \tag{41}$$

For astrophysical conditions this is a very sharp bound. For example, when integrated over the lifetime of a solar flare, it gives a change in helicity less than one part in 10^5 compared to the change in magnetic energy. While this is a rigorous inequality, in practice it may overestimate helicity loss. The inequality assumes the helicity dissipation occurs over the entire integration volume. If the dissipation is in a small subvolume (as frequently occurs in reconnection events), then the helicity loss may be substantially less than the upper bound. The relative conservation of helicity compared to energy is of importance in both fusion plasmas [67] and solar coronal fields [34]. Almost exact helicity conservation during reconnection implies that

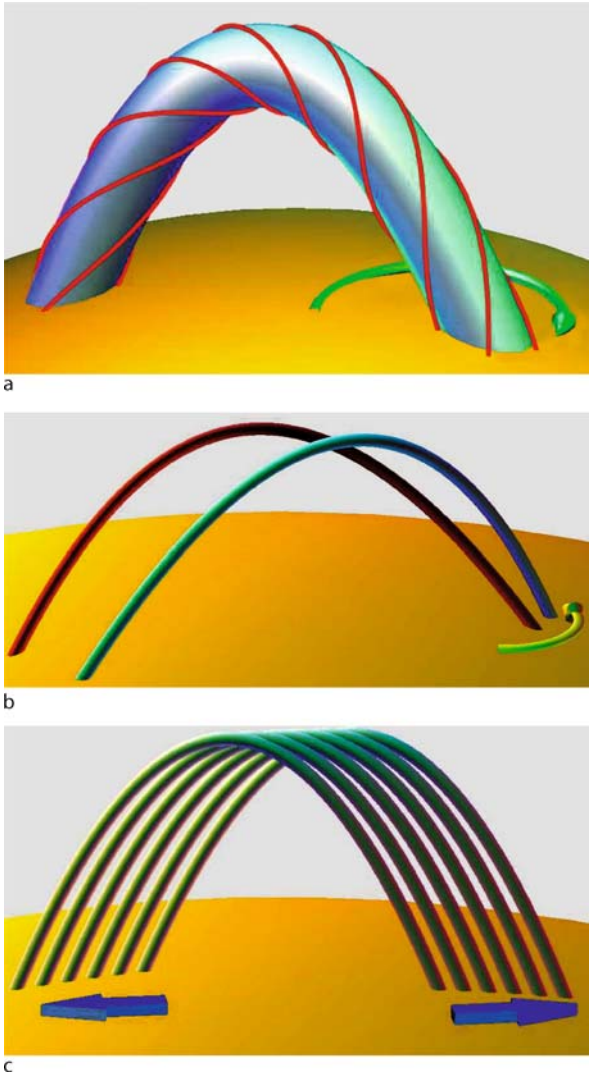

Topological Magnetohydrodynamics and Astrophysics, Figure 8
Three ways in which helicity can flow through a boundary. A single flux tube can become twisted due to rotational motions at the boundary. Two flux tubes can braid if their end points move around each other. An arcade can acquire helicity due to shearing motions at the boundary

mutual helicity can be converted into self helicity, or vice-versa [7]. For example, flux transfer events at the Earth's magnetopause transfer the mutual helicity of interplanetary and terrestrial magnetic fields into internal twist of magnetic flux ropes [72].

The concept of writhe can be extended to curves with endpoints [15,30,65]. For example, tubes of magnetic flux in the atmosphere of the sun can become highly twisted, leading to a kinking of the flux tube geometry (see Fig. 9). In this situation the twist number of magnetic lines of force about the central axis of the tube decreases, with a corresponding increase in the writhe of the axis [15,33].

Applications to Astrophysical Fields

Magnetic fields are ubiquitous in the universe. Stars and the interstellar medium consist predominantly of highly conducting ionized plasma. On all but the smallest scales the fluid MHD approximation works well (exceptions include the fine structure of reconnection events, the acceleration of cosmic ray particles, and the effects of neutral atoms on the plasma). Magnetic fields generate a wide range of activity [73,78,79]: the atmospheres of the sun and stars are subject to violent magnetic storms known as *flares*; large parts of the solar atmosphere can simply lift off into space as *coronal mass ejections*; magnetic fields structure stellar winds, accretion disks and galactic jets.

The solar magnetic field is thought to be generated (and regenerated every 11 years in the solar cycle) by a *dynamo* converting the kinetic energy of convection into magnetic energy [76,87]. Magnetic helicity $H(\mathbf{B},\mathbf{B})$ and electrical current helicity $H(\mathbf{J},\mathbf{J})$ moderate the growth of the magnetic field in a dynamo [19,39]. Dynamo models in which the magnetic field keeps its structure but grows in magnitude must have zero magnetic helicity; otherwise magnetic helicity would not be conserved [32,51].

After new magnetic flux is generated in the interior of the sun, it can reach the surface (*photosphere*) via the effects of convection and buoyancy. Measurements of helicity flow across the photosphere (as illustrated in Fig. 8) can give us clues about the generation process. Present day measurements involve direct measurement of the helicity Poynting flux [22,54,59], matching best-fit linear force–free fields with observations [4], and calculating the effects of solar differential rotation on the solar magnetic field [16,26]. Structures in the corona tend to have negative magnetic helicity in the Northern hemisphere and positive helicity in the southern hemisphere [60]. See [77] for a collection of helicity studies relevant to the sun.

Magnetic field lines rising from the photosphere can either form loops in the corona (see Fig. 10) and reenter the photosphere, or wander off into the interplanetary medium. The loops provide a mathematical mapping of the surface to itself. This connectivity is invariant to ideal motions in the atmosphere, but slowly changes due to motions of field line endpoints in the photosphere. Reconnection events such as flares can rapidly change the connectivity, of course. Discontinuities in the connectivity are called *separatrices*. Two separatrix surfaces can interest in a *separator* line. Separators in turn can inter-

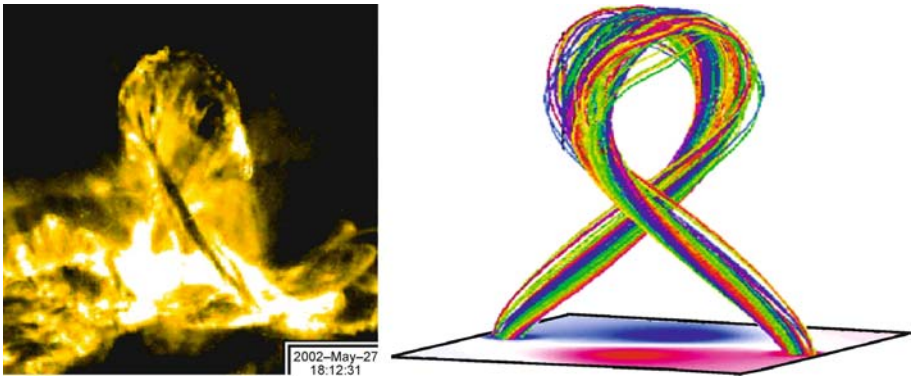

Topological Magnetohydrodynamics and Astrophysics, Figure 9
An erupting magnetic structure in the solar corona. *On the left* is an image from the TRACE mission (195 Å line). The *right picture* is from a numerical simulation of this event by Török and Kliem 2005 [69]

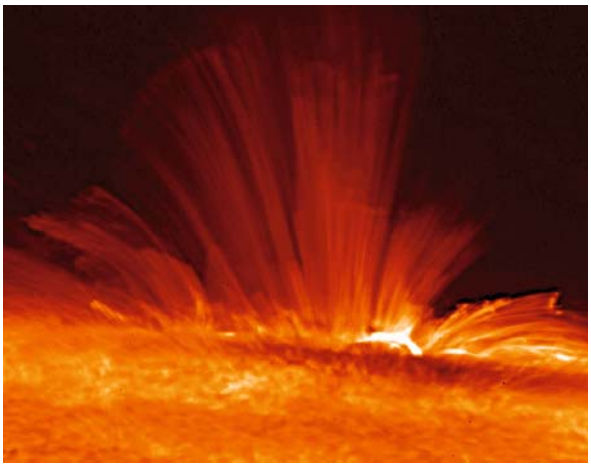

Topological Magnetohydrodynamics and Astrophysics, Figure 10
Loops of hot plasma in the solar atmosphere (*corona*), as seen by the Hinode mission solar optical telescope. As the thermal conductivity is much greater parallel to the magnetic field than across the field, individual field lines tend to have an almost uniform temperature. As a consequence, the plasma loops seen at particular temperatures trace the magnetic field lines. Dark areas also have magnetic fields, but with the wrong plasma temperature for the telescope's wavelength band. Image credit: Hinode JAXA/NASA

sect in a *magnetic neutral point*. The complex of separatrices, separators, and neutral points is called the *skeleton* of the magnetic field. Recent years have seen extensive research into the structure of the magnetic skeleton in solar fields, and its relevance to solar activity such as flares and mass ejections [20,43,68] (a more extensive review appears in [75]). Sometimes the connectivity of the field lines can have sharp gradients without actual discontinuities. The regions of sharp gradients may still have enhanced activity [5,9,61].

The solar corona is heated to over 1 million degrees, even though the surface is only 5000 K. The likely mechanism for this heating involves many small tiny reconnection events called microflares or nanoflares. Parker [56, 57,73,74] has developed a theory of *topological dissipation* to account for this. Photospheric motions continually increase the complexity of the coronal field topology by moving the endpoints of the field lines (as in Fig. 8). This leads to an increase in minimum crossing numbers and minimum magnetic energy [11,12]. The corona can only relieve these stresses by reconnecting the field lines in flare events.

As flares only change the connectivity, but do not destroy magnetic helicity, the helicity of a coronal region can build up over time. It is possible that coronal mass ejections provide the mechanism of allowing the sun to shed excess helicity into outer space [40,45,63].

Future Directions

Astrophysicists have developed a great body of knowledge and experience concerning the topology of potential fields (e. g. [17,20,75]). However, the topology of force-free fields [3,8,38] needs further exploration. Global field line descriptions in terms of the magnetic skeleton, topological invariants like helicity, and complexity measures like minimum crossing number have evolved separately; their interrelationships require further study.

Magnetic equilibrium studies have concentrated mostly on smooth solutions to the equilibrium equations. However, smooth solutions may not always exist or be available to an evolving astrophysical plasma [2,46,55,56,

57,58,74]. In this case current sheets will form, leading to activity like flares and coronal mass ejections. The mathematical properties of equilibrium solutions with current sheets could be a direction of fruitful future work.

Another future direction concerns detailed analyzes of the flow of helicity throughout the heliosphere, from the interior of the sun through the photosphere, into the corona, and out to magnetic clouds in the interplanetary medium, using the helicity Poynting Theorem 6. Helicity flow may also be of particular importance in dynamo studies [66].

Bibliography

Primary Literature

1. Aldinger J, Klapper I, Tabor M (1995) Formulae for the calculation and estimation ofwrithe. J Knot Theory Ram 4:343–372
2. Aly JJ (2005) A uniqueness result for a simple force-free magnetic field submitted to a topological constraint. Astronom Astrophys 429:15–18
3. Antiochos SK (1987) The topology of force-free magnetic fields and its implications for coronal activity. Astrophys J 312: 886–894
4. Aulanier G, Srivastava N, Martin SF (2000) Model Prediction for an Observed Filament. Astrophys J 543:447–456
5. Aulanier G, Pariat E, Demoulin P (2005) Current sheet formation in quasiseparatrix layers and hyperbolic flux tubes. Astronom Astrophys 444:961–976
6. Barnes DC (1988) Mechanical injection of magnetic helicity. Phys Fluids 31:2214–2220
7. Berger MA (1984) Rigorous new limits on magnetic helicity dissipation in the solar corona. Geophys Astrophys Fluid Dyn 30:79–104
8. Berger MA (1985) Structure and stability of constant α force-free fields. Astrophys J 59:433–444
9. Berger MA (1989) Three-dimensional reconnection from a global viewpoint. In: Guyenne TD, Hunt JJ (eds) Reconnection in Space Plasma. ESA Paris SP-285 II:83–86
10. Berger MA (1990) Third order link integrals. J Phys A: Math Gen 23:2787–2793
11. Berger MA (1993) Energy-crossing number relations for braided magnetic fields. Phys Rev Lett 70:705–708
12. Berger MA (1994) Coronal Heating by Dissipation of Magnetic Structure. Space Sci Rev 68:3–14
13. Berger MA (2001) Topological invariants in braid theory. Lett Math Phys 55:181–192
14. Berger MA, Field GB (1984) The topological properties of magnetic helicity. J Fluid Mech 147:133–148
15. Berger MA, Prior P (2006) The writhe of open and closed curves. J Phys A: Math Gen 39:8321–8348
16. Berger MA, Ruzmaikin A (2000) Magnetic helicity production by differential rotation. J Geophys Res 105:10481–10490
17. Beveridge C, Longcope DW (2005) On Three-Dimensional Magnetic Skeleton Elements due to Discrete Flux Sources. Solar Phys 227:193–206
18. Birman JS (1991) Recent Developments In Braid And Link Theory Math Intell 13:52–60
19. Blackman EG, Field GB (2000) Constraints on the Magnitude of α in Dynamo Theory. Astrophys J 534:984–988
20. Brown DS, Priest ER (2001) The topological behaviour of 3D null points in the Sun's corona. Astronom Astrophys 367: 339–346
21. Călugăreanu (1961) On isotopy classes of three-dimensional knots and their invariants. Czechoslovak Math J T11:588–598
22. Chae J (2001) Observational Determination of the Rate of Magnetic Helicity Transport through the Solar Surface via the Horizontal Motion of Field Line Footpoints. Astrophys J 560: L95–L98
23. Chui AYK, Moffatt HK (1995) The energy and helicity of knotted magnetic flux tubes. Phil Trans R Soc London A 451:609–618
24. Deng XH, Matsumoto H (2001) Rapid magnetic reconnection in the Earth's magnetosphere mediated by whistler waves. Nature 410:557–560
25. Dennis MR, Hannay JH (2005) Geometry of Calugareanu's theorem. Proc Roy Soc A 2062:3245–3254
26. DeVore CR (2000) Magnetic helicity generation by solar differential rotation. Astrophys J 539:944–953
27. Dixon A, Berger MA, Browning P, Priest ER (1989) A generalization of the Woltjer minimum energy principle. Astronom Astrophys 225:156–166
28. Epple M (1998) Orbits of asteroids, a braid, and the first link invariant. Math Intell 20:45–52
29. Finn J, Antonsen TM (1985) Magnetic helicity: what is it, and what is it good for? Comments Plasma Phys Contr Fusion 9:111–120
30. Fuller FB (1978) Decomposition of the linking of a ribbon: a problem from molecular biology. Proc Natl Acad Sci USA 75:3557–3561
31. Ghrist R, Komendarczyk R (2006) Overtwisted energy-minimizing curl eigenfields. Nonlinearity 19:41–52
32. Gilbert AD (2002) Magnetic Helicity in Fast Dynamos. Geophys Astrophys Fluid Dyn 96:135–151
33. Green LM, Kliem B, Török T, van Driel-Gesztelyi L, Attrill GDR (2007) Transient Coronal Sigmoids and Rotating Erupting Flux Ropes. Sol Phys 246:365–391
34. Heyvaerts J, Priest ER (1984) Coronal heating by reconnection in DC current systems – a theory based on Taylor hypothesis. Astronom Astrophys 137:63–78
35. Hornig G (1997) The covariant transport of electromagnetic fields and its relation to magnetohydrodynamics. Phys Plasmas 4:646–654
36. Hornig G, Mayer C (2002) Towards a third-order topological invariant for magnetic fields. J Phys A: Math Gen 35:3945–3959
37. Hornig G, Schindler K (1996) Magnetic topology and the problem of its invariant definition. Phys Plasmas 3:781–791
38. Hudson TS, Wheatland MS (1999) Topological Differences Between Force-Free Field Models. Sol Phys 186:301–310
39. Ji HT (1999) Turbulent dynamos and magnetic helicity. Phys Rev Lett 83:3198–3201
40. Kumar A, Rust DM (1996) Interplanetary magnetic clouds, helicity conservation, and current-core flux ropes. J Geophys Res 101:15667–15684
41. Lau YT, Finn JM (1990) Three-dimensional kinematic reconnection in the presence of field nulls and closed field lines. Astrophys J 350:672–691
42. Laurence P, Stredulinsky E (2000) Asymptotic Massey products, induced currents and Borromean torus links. J Math Phys 41:3170–3191

43. Longcope DW, Klapper I (2002) A general theory of connectivity and current sheets in coronal magnetic fields. Astrophys J 579:468–481
44. Low BC (1990) Equilibrium and dynamics of coronal magnetic fields. Ann Rev Astron Astrophys 28:491–524
45. Low BC (1994) Magnetohydrodynamic processes in the solar corona – flares, coronal mass ejections, and magnetic helicity. Phys Plasmas 1:1684–1690
46. Low BC, Wolfson R (1988) Spontaneous formation of electric current sheets and the origin of solar flares. Astrophys J 324:574–581
47. Massey WS (1998) Higher order linking numbers. J Knot Theory Ram 7:393–414
48. Moffatt HK (1969) The degree of knottedness of tangled vortex lines. J Fluid Mech 35:117–129
49. Moffatt HK (1985) Magnetostatic equilibria and analogous Euler flows of arbitrarily complex topology, Part 1: Fundamentals. J Fluid Mech 159:359–378
50. Moffatt HK (1990) The energy spectrum of knots and links. Nature 347:367–369
51. Moffatt HK, Proctor MRE (1985) Topological constraints associated with fast dynamo action. J Fluid Mech 154:493–507
52. Moffatt HK, Ricca RL (1992) Helicity and the Călugăreanu invariant. Proc Royal Society London A 439:411–429
53. Monastyrski MI, Sasarov PV (1987) Topological invariants in magnetohydrodynamics. Sov Phys JETP 66:683–688
54. Moon YJ, Chae J, Choe GS et al. (2002) Flare activity and magnetic helicity injection by photospheric horizontal motions. Astrophys J 574:1066–1073
55. Ng CS, Bhattacharjee A (1998) Nonequilibrium and current sheet formation in line-tied magnetic fields. Phys Plasmas 5:4028–4040
56. Parker EN (1972) Topological dissipation and the small-scale fields in turbulent gases. Astrophys J 174:499–510
57. Parker EN (1983) Magnetic neutral sheets in evolving fields, II – Formation of the solar corona. Astrophys J 264:642–647
58. Parker EN (2004) Tangential discontinuities in untidy magnetic topologies. Phys Plasmas 11:2328–2332
59. Pariat E, Nindos A, Démoulin P, Berger MA (2006) What is the spatial distribution of magnetic helicity injected in a solar active region? Astronom Astrophys 52:623–630
60. Pevtsov AA, Canfield RC, Metcalf TR (1995) Latitudinal variation of helicity of photospheric magnetic fields. Astrophys J 440:L109–L112
61. Priest ER, Demoulin P (1995) Three-dimensional magnetic reconnection without null points, 1: Basic theory of magnetic flipping. J Geophys Res 100(A9):23443–23464
62. Redd AJ, Jarboe TR, Hamp WT et al (2007) Overview of the Helicity Injected Torus (HIT) program. J Fusion Energy 26: 163–168
63. Rust DM (1994) Spawning and shedding helical magnetic fields in the solar atmosphere. Geophys Res Lett 21:241–244
64. Ruzmaikin A, Akhmetiev P (1994) Topological invariants of magnetic fields, and the effects of reconnection. Phys Plasmas 1:331–336
65. Starostin EL (2005) On the writhing number of a non-closed curve. In: Calvo J, Millett K, Rawdon E, Stasiak A (eds) Physical and numerical models in knot theory including applications to the life sciences. Series on Knots and Everything. World Scientific Publishing, Singapore, pp 525–545
66. Subramanian K, Brandenburg A (2006) Magnetic Helicity Density And Its Flux In Weakly Inhomogeneous Turbulence. Astrophys J 648:L71–L74
67. Taylor JB (1974) Relaxation of toroidal plasma and generation of reverse magnetic fields. Phys Rev L 33:1139–1141
68. Titov VS, Hornig G, Demoulin P (2002) Theory of magnetic connectivity in the solar corona. J Geophys Res Space Phys 107:SSH3-1
69. Török T, Kliem B (2005) Confined and ejective eruptions of kink-unstable flux ropes. Astrophys J 630:L97–L100
70. Wilmot-Smith AL, Priest ER, Hornig G (2005) Magnetic diffusion and the motion of field lines. Geophys Astrophys Fluid Dyn 99:177–197
71. Woltjer L (1958) A theorem on force-free magnetic fields. Proc Natl Acad Sci USA 44:489–491
72. Wright A, Berger MA (1989) The effect of reconnection upon the linkage and interior structure of magnetic flux tubes. J Geophys Res 94:1295–1302
73. Parker EN (1979) Cosmical Magnetic Fields: Their Origin and Their Activity. International Series of Monographs on Physics, Clarendon Press. Oxford University Press, Oxford
74. Parker EN (1994) Spontaneous Current Sheets in Magnetic Fields: With Applications to Stellar X-Rays. International Series on Astronomy and Astrophysics, vol 1. Oxford University Press, New York
75. Longcope DW (2005) Topological Methods for the Analysis of Solar Magnetic Fields. Living Reviews of Solar Physics 2:7 Max-Planck-Gesellschaft, Katlenburg-Lindau
76. Brandenburg A, Subramanian K (2005) Astrophysical magnetic fields and nonlinear dynamo theory. Phys Rep 417:1–205
77. Brown MR, Canfield RC, Pevtsov AA (eds) (1999) Magnetic Helicity in Space and Laboratory Plasmas. Geophysical Monograph 111, American Geophysical Union, Washington
78. Aschwanden M (2006) Physics of the Solar Corona. Springer, New York
79. Choudhuri AR (1998) The Physics of Fluids and Plasmas: An Introduction for Astrophysicists. Cambridge University Press, Cambridge
80. Arnold VI, Khesin BA (1998) Topological Methods in Hydrodynamics. Springer, New York
81. Bateman G (1978) MHD Instabilities. MIT Press, Cambridge
82. Priest ER (1984) Solar Magnetohydrodynamics. Springer, Berlin
83. Priest ER, Forbes TG (1999) Magnetic Reconnection: MHD Theory and Applications. Cambridge University Press, Cambridge
84. Biskamp D (1997) Nonlinear Magnetohydrodynamics. Cambridge Monographs on Plasma Physics 1, Cambridge
85. Cantarella J, DeTurck D, Gluck H (2001) The Biot-Savart operator for application to knot theory, fluid dynamics, and plasma physics. J Math Phys 42:876–905
86. Jackson JD (1998) Classical Electrodynamics. Wiley, New York
87. Moffatt HK (1978) Magnetic Field Generation in Electrically Conducting Fluids. Cambridge Univ Press, Cambridge

Books and Reviews

Kulsrud RM (2004) Plasma Physics for Astrophysics. Princeton University Press, Princeton

Moffatt HK (1983) Transport effects associated with turbulence, with particular attention to the influence of helicity. Rep Prog Phys 46:621–664

Moffatt HK, Tsinober A (1992) Helicity in laminar and turbulent flow. Ann Rev Fluid Mech 24:281–312
Parker EN (2007) Conversations on Electric and Magnetic Fields in the Cosmos. Princeton University Press, Princeton

Traffic Breakdown, Probabilistic Theory of

Boris S. Kerner[1], Sergey L. Klenov[2]
[1] Group Research GR/ETI, HPC: G021, Daimler AG, Sindelfingen, Germany
[2] Moscow Institute of Physics and Technology, Department of Physics, Dolgoprudny, Russia

Article Outline

Glossary
Definition of the Subject
Introduction
Traffic Breakdown – First-Order Phase Transition from Free Flow to Synchronized Flow
Probabilistic Description of Traffic Breakdown with Cellular Automata (CA) Traffic Flow Model
Probabilistic Description of Traffic Breakdown Based on Master Equation
Capacity of Free Flow at Bottlenecks
Conclusions
Future Directions
Acknowledgments
Bibliography

Glossary

Traffic breakdown Traffic breakdown is the onset of congested traffic in an initial free traffic flow. Traffic breakdown occurs mostly at effectual road bottlenecks like on- and off-ramps, roadworks, road gradients, reduction of road lanes, etc. Traffic breakdown results in the emergence of the synchronized flow phase of congested traffic, i. e., traffic breakdown is a phase transition from the free flow traffic phase to synchronized flow traffic phase ($F \to S$ transition for short). Thus the terms *traffic breakdown* and an $F \to S$ *transition* are synonyms related to the same phenomenon of the onset of congestion in free flow.

Effectual bottleneck An effectual bottleneck is a freeway bottleneck at which an $F \to S$ transition (traffic breakdown) occurs during many days and years of observations. Because only effectual bottlenecks are considered in the article, the term *bottleneck* is used below for an effectual bottleneck.

Nucleation and probabilistic features of traffic breakdown Traffic breakdown is a *local first-order* $F \to S$ transition, which exhibits a nucleation feature:

(i) At the same bottleneck, traffic breakdown can be either *spontaneous* or *induced*.
(ii) Onset and dissolution of congestion are accompanied by a *hysteresis* effect.
(iii) Traffic breakdown exhibits the following *probabilistic* features: (a) at the same flow rates in some of the realizations (days) traffic breakdown occurs but in other realizations it does not occur; (b) empirical probability of the breakdown is a growing function of the flow rate downstream of the bottleneck;
(iv) In a realization, the flow rate at which traffic breakdown is observed can be considerably smaller than flow rates at which no traffic breakdown has occurred before.

Nucleation model for traffic breakdown In three-phase traffic theory, it is suggested that a bottleneck introduces a speed and density disturbance that is spatially limited: at the same flow rates, the average speed is lower and vehicle density is greater within the disturbance. The disturbance is on average motionless and it is localized in a neighborhood of the bottleneck. The total number of vehicles within the disturbance can be considered a local vehicle cluster (cluster for short). Traffic breakdown occurs when a critical cluster appears whose subsequent growth leads to synchronized flow emergence.

Definition of the Subject

Traffic breakdown is the onset of congested traffic in an initial free traffic flow. Thus traffic breakdown restricts free flow conditions in vehicular traffic. In observations, traffic breakdown exhibits the probabilistic nature. For this reason, the development of the probabilistic theory of traffic breakdown is of a great important for all kind of traffic management and control in transportation networks and for other traffic engineering applications.

Introduction

The onset of congestion in an initial free flow is accompanied by a sharp decrease in average vehicle speed in the free flow to a considerably lower speed in congested traffic. This speed breakdown occurs mostly at freeway bottlenecks and is called the breakdown phenomenon or traffic breakdown (see, e. g., [5,10,11,31,35]). The traffic breakdown occurs mostly at effectual road bottlenecks like

on- and off-ramps, roadworks, road gradients, reduction of road lanes, etc. An effectual bottleneck is a bottleneck where traffic breakdown most frequently occurs on many different days.

It has been found that the breakdown phenomenon has a probabilistic nature [5,35]. The probabilistic nature of traffic breakdown means the following: at a given flow rate traffic breakdown can occur but it should not necessarily occur. This means that on one day the onset of congestion occurs, however, on another day at the same flow rate and at the same traffic conditions the onset of congestion is not observed.

There are a huge number of publications devoted to theoretical investigations of traffic breakdown (see references in [11,17,30,31]). However, the puzzle of empirical features of traffic breakdown has been understood only recently [14,16]. Consequently, as explained in ► Traffic Congestion, Modeling Approaches to in this Encyclopedia, earlier traffic flow theories and models reviewed in [2,3,4,8,9,12,26,30,31,32,34,36,37,38] cannot explain and predict the set of the fundamental features of traffic breakdown.

For these reasons, Kerner introduced an alternative traffic flow theory called three-phase traffic theory, which overcomes the disadvantages of other traffic flow models and theories in the explanation of spatiotemporal empirical features of traffic breakdown [17]. For this reason, the three-phase traffic theory is the basis for the probabilistic theory of traffic breakdown, which explains empirical spatiotemporal probabilistic features of traffic breakdown.

In three-phase traffic theory, besides the free flow traffic phase there are two traffic phases in congested traffic: synchronized flow and wide moving jam. Thus there are three traffic phases in this theory: 1. Free flow. 2. Synchronized flow. 3. Wide moving jam. The synchronized flow and wide moving jam phases in congested traffic are defined through spatiotemporal empirical criteria. The definition of the wide moving jam phase [J]: A wide moving jam is a moving jam that maintains the mean velocity of the downstream jam front, even when the jam propagates through any other traffic states or bottlenecks. The definition of the synchronized flow phase [S]: In contrast with the wide moving jam phase, the downstream front of the synchronized flow phase does not exhibit the wide moving jam characteristic feature; in particular, the downstream front of the synchronized flow phase is often fixed at a bottleneck.

It turns out that traffic breakdown is a local phase transition from the free flow traffic phase to synchronized flow traffic phase ($F \rightarrow S$ transition) [17]. Thus the terms traffic breakdown, breakdown phenomenon, speed breakdown and an $F \rightarrow S$ transition are synonyms that are related to the same phenomenon of the onset of congestion in free flow.

Reviews of empirical features of traffic breakdown and resulting traffic congested patterns as well as modeling approaches to traffic congestion have already been made in this Encyclopedia (► Traffic Congestion, Modeling Approaches to, ► Traffic Congestion, Spatiotemporal Features of). In this article, we discuss approaches to probabilistic description of traffic breakdown.

The article is organized as follows. In Sect. "Traffic Breakdown – First-Order Phase Transition from Free Flow to Synchronized Flow", a qualitative discussion of a nucleation model for traffic breakdown, which can explain fundamental empirical features of traffic breakdown, is made. Cellular automata probabilistic description of traffic breakdown is reviewed in Sect. "Probabilistic Description of Traffic Breakdown with Cellular Automata (CA) Traffic Flow Model". In Sect. "Probabilistic Description of Traffic Breakdown Based on Master Equation" based on the nucleation model of Sect. "Traffic Breakdown – First-Order Phase Transition from Free Flow to Synchronized Flow", we consider a probabilistic description of traffic breakdown with master equation. A link between traffic breakdown, freeway capacity, and the diagram of congested patterns at bottlenecks is considered in Sect. "Capacity of Free Flow at Bottlenecks".

Traffic Breakdown – First-Order Phase Transition from Free Flow to Synchronized Flow

Fundamental Empirical Features of Traffic Breakdown

Empirical features of traffic breakdown ($F \rightarrow S$ transition) and their explanation have already been considered in ► Traffic Congestion, Modeling Approaches to in this Encyclopedia. For this reason, we make here only a brief discussion of the fundamental empirical features of this phenomenon, which are as follows (Fig. 1):

A. Traffic breakdown is a local phase transition from the free flow phase to the synchronized flow phase ($F \rightarrow S$ transition).

B. At the same bottleneck, traffic breakdown can be either *spontaneous* or *induced*.

C. Onset and dissolution of congestion are accompanied by a *hysteresis* effect.

D. Traffic breakdown exhibits the *probabilistic* nature.

The fundamental feature **A** has already been explained above. Feature **B**. In empirical observations, the spontaneous traffic breakdown (spontaneous $F \rightarrow S$ transition) occurs within an initial free flow at a bottleneck due to an *internal* speed disturbance in the free flow in a neighbor-

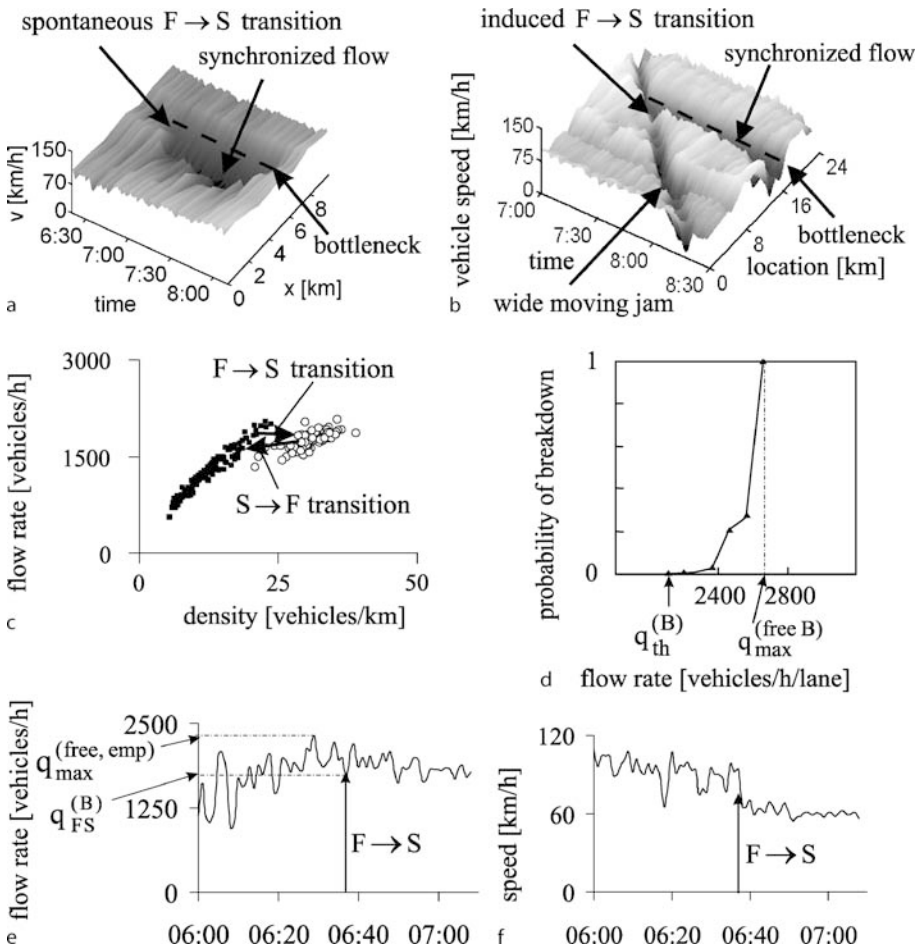

Traffic Breakdown, Probabilistic Theory of, Figure 1

Fundamental empirical features of traffic breakdown (*F* → *S* transition): a, b Spontaneous (a) and induced (b) traffic breakdowns; vehicle speed in space and time; taken from [17]. c Hysteresis effect associated with a (*arrows* labeled *F* → *S* and *S* → *F* show traffic breakdown and *S* → *F* transition at bottleneck, respectively). d Probability of traffic breakdown as function of flow rate downstream of bottleneck; 10 min average data; taken from [35]. e, f Flow rate (e) and speed (f) at bottleneck as time-function before and after breakdown (arrows labeled *F* → *S* show the time instant of the breakdown). Taken from [17]

hood of the bottleneck (Fig. 1a). In contrast, the induced traffic breakdown (induced $F \to S$ transition) is caused by a short-time *external* speed disturbance in traffic flow in a neighborhood of the bottleneck. This external disturbance is usually related to the propagation of a moving spatiotemporal congested pattern that initially occurs at a *different* freeway location than that of the induced $F \to S$ transition. The induced $F \to S$ transition can occur, after the congested pattern has reached the bottleneck. In particular, the breakdown can be induced by a wide moving jam propagating upstream through a bottleneck (Fig. 1b). Induced $F \to S$ transition at a bottleneck can also occur when a region of synchronized flow first occurs downstream of this bottleneck, and the region later reaches the bottleneck due to the upstream propagation of synchronized flow (see an empirical example in ► Traffic Congestion, Modeling Approaches to).

Feature **C**. The $F \to S$ transition, which leads to congested pattern emergence, and the $S \to F$ transition, which leads to the dissolution of this congested pattern, are accompanied by a well-known *hysteresis effect* and hysteresis loop in the flow-density plane (see references in [10,11]): a congested pattern emerges usually at a greater flow rate downstream of the bottleneck than this flow rate is at which the congested pattern dissolves (Fig. 1a,c).

Feature **D**. The probability of traffic breakdown at a bottleneck is an increasing function of flow rate (Fig. 1d)

[35]. At given traffic parameters (weather, etc.), the maximum flow rate downstream of an on-ramp bottleneck (denoted $q_{\max}^{(\text{free,emp})}$), which was measured on a specific day before congestion occurred, can be greater than the flow rate (denoted $q_{\text{FS}}^{(\text{B})}$) at which traffic breakdown occurs (Fig. 1e,f).

Model of Traffic Breakdown at Highway Bottleneck

In three-phase traffic theory [17], the fundamental empirical features of traffic breakdown are explained as follows. The possibility of induced and spontaneous traffic breakdowns at the bottleneck (Fig. 1a,b) can be explained by the *nucleation* nature of breakdown phenomenon. This means that there should be some critical speed in free flow required for traffic breakdown. The breakdown occurs, if due to a speed disturbance in free flow in the neighborhood of a bottleneck the speed decreases below the critical one. Otherwise, if free flow speed within the disturbance is greater than the critical one, the breakdown does not occur.

The critical speed within a speed disturbance required for the breakdown should depend on the flow rate downstream of the bottleneck. The smaller the flow rate, the lower the critical speed required for the breakdown in free flow. In contrast, when the flow rate increases, the critical speed at the bottleneck required for the breakdown increases. Obviously, the lower the critical speed, the smaller the probability for the breakdown. This explains the increasing character of the probability of traffic breakdown on the flow rate (Fig. 1d).

In accordance with other systems of natural science, the disturbance in free flow at the bottleneck within which the critical speed is reached can be called a critical speed disturbance or a *nucleus* for traffic breakdown. Only if a disturbance appears randomly at the bottleneck within which the speed is equal or lower than the speed within the critical disturbance, traffic breakdown occurs. In other words, the probability for traffic breakdown is the probability of random critical speed disturbance appearance at the bottleneck. These nucleation features of traffic breakdown are general ones for first-order phase transitions observed in a diverse variety of physical, chemical, biological and other complex spatiotemporal systems.

Thus, the empirical features **A-D** of traffic breakdown mean that the breakdown is a *first-order* $F \to S$ transition. The flow rate in free flow downstream of the bottleneck should be a control parameter for this phase transition: at the same other traffic characteristics, the greater the flow rate, the greater the probability for the $F \to S$ transition. For this reason, there is the maximum flow rate (denoted $q_{\max}^{(\text{free B})}$ and that is shown in Fig. 1d) downstream of the bottleneck at which breakdown probability denoted by $P_{\text{FS}}^{(\text{B})}$ reaches one, i. e., $P_{\text{FS}}^{(\text{B})} = 1$. There is also a threshold flow rate (denoted $q_{\text{th}}^{(\text{B})}$). At the flow rate in free flow downstream of the bottleneck that is smaller than $q_{\text{th}}^{(\text{B})}$, breakdown probability $P_{\text{FS}}^{(\text{B})} = 0$. Both $q_{\max}^{(\text{free B})}$ and $q_{\text{th}}^{(\text{B})}$ are mean values, which found in many different realizations (days) at which traffic breakdowns have occurred.

Model of Local Vehicle Cluster at Bottleneck In a nucleation model of traffic breakdown based on three-phase traffic theory [15,17,20,21,22], we assume that the breakdown is associated with the existence of an *initial non-homogeneity* of free flow at the bottleneck. Due to this non-homogeneity, the speed is lower and density is greater in a neighborhood of the bottleneck than outside the bottleneck. This local speed and density disturbance at the bottleneck can be considered a local *vehicle cluster* in free flow at the bottleneck. Traffic breakdown probability within the cluster should be considerably greater than outside the cluster. This should explain why traffic breakdown is observed mostly at bottlenecks.

The non-homogeneity in free flow at the bottleneck, i. e., the vehicle cluster at the bottleneck exists even in a hypothetical case in which there were *no random fluctuations* in traffic flow. In this hypothetical case, this permanent non-homogeneity can be considered a *deterministic* vehicle cluster (a *deterministic* local disturbance) localized at the bottleneck.

Let us firstly discuss an on-ramp bottleneck. In this case, the non-homogeneity of an initial free flow at the bottleneck is caused by two flows, which merge at the bottleneck: (i) An on-ramp inflow with the rate q_{on}. (ii) A flow on the main road upstream of the bottleneck with the rate q_{in}. This flow merging occurs permanent and on the same freeway location (within an on-ramp merging region). For this reason, the non-homogeneity of free flow at the bottleneck is on average motionless and permanent (Fig. 2).

To explain features of this non-homogeneity, note that at a given high enough flow rate q_{in} in free flow on the main road upstream of the bottleneck, vehicles that merge from the on-ramp onto the main road force the vehicles on the main road to decelerate in the vicinity of an on-ramp merging region. This leads to a local decrease in speed and consequently to a local increase in density, i. e. to vehicle cluster occurrence and existence in free flow in the vicinity of the bottleneck. One of the characteristics of the cluster is the cluster size, i. e., the *total* number of vehicles (denoted by N) within the cluster location (Fig. 2a).

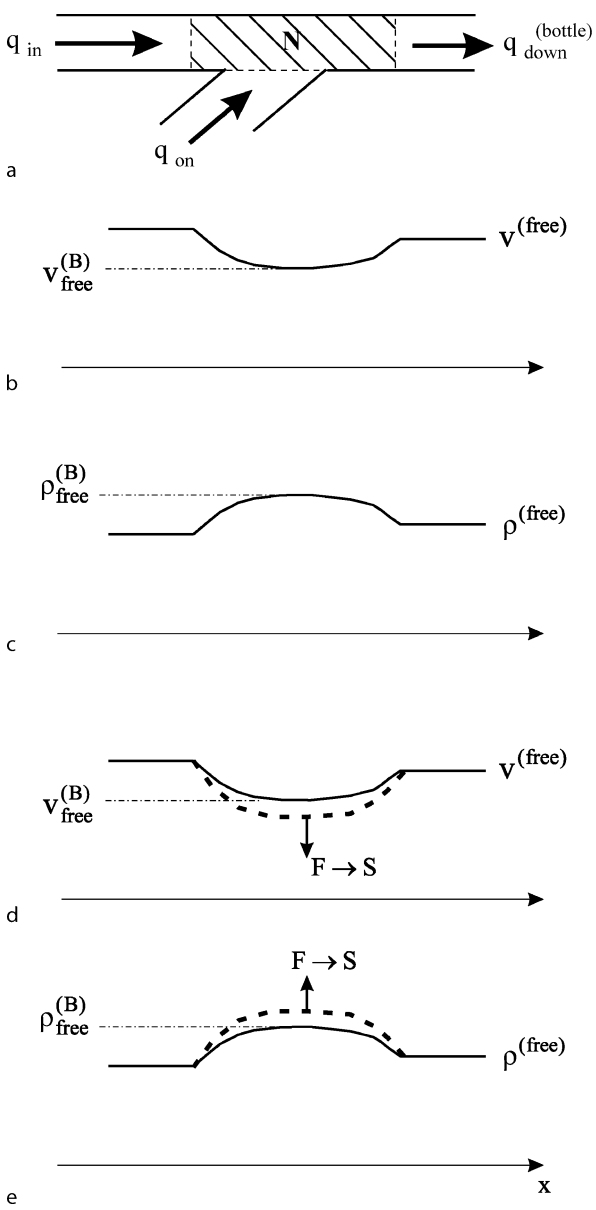

Traffic Breakdown, Probabilistic Theory of, Figure 2
Qualitative explanations of vehicle cluster in free flow at on-ramp bottleneck: a Qualitative schema of on-ramp bottleneck with a vehicle cluster (*dashed region*) localized at the bottleneck. b, c Qualitative spatial speed (b) and density (c) distributions within deterministic vehicle cluster (deterministic local disturbance). d, e *Dashed curves* show qualitative spatial speed (d) and density (e) distributions within a vehicle cluster (local disturbance) at a fixed time instant for the case when fluctuations in free flow lead to an increase in the cluster size *N* in comparison with the size of the deterministic cluster $N^{(\mathrm{determ})}$. *Arrows* labeled *F* → *S* symbolize traffic breakdown that occurs when a critical cluster appears

It must be noted that here and below, the flow rates q_{in}, and q_{on}, as well as all other flow rates used in this article below, are total flow rates across associated road locations, which are divided by the number of lanes on the main road downstream of the bottleneck. As the flow rates, the cluster size N is also considered per a highway lane: the total vehicle number of vehicles within the cluster is divided by the same number of lanes on the main road downstream of the bottleneck as that used for flow rate unit definition.

As mentioned above, in the hypothetical free flow at a bottleneck in which there were no random fluctuations, a deterministic cluster occurs at the bottleneck. We denote the size of this deterministic cluster by $N = N^{(\mathrm{determ})}$. The speed $v_{\mathrm{free}}^{(\mathrm{B})}$ and density $\rho_{\mathrm{free}}^{(\mathrm{B})}$ within this deterministic cluster (Fig. 2b,c) correspond to the conditions

$$v_{\mathrm{free}}^{(\mathrm{B})} < v^{(\mathrm{free})},\ \rho_{\mathrm{free}}^{(\mathrm{B})} > \rho^{(\mathrm{free})}\ , \tag{1}$$

where $v^{(\mathrm{free})}$ and $\rho^{(\mathrm{free})}$ are the average vehicle speed and density in homogeneous free flow on the main road downstream of the cluster (Fig. 2). Because the deterministic cluster is permanent and motionless, for the on-ramp bottleneck at given q_{in} and q_{on}, the flow rate

$$q_{\mathrm{sum}} = q_{\mathrm{in}} + q_{\mathrm{on}} \tag{2}$$

does not depend on the spatial co-ordinate in the *hypothetical* state of the free flow. For this reason, the speed and density on the main road satisfy the following condition:

$$q_{\mathrm{sum}} = v^{(\mathrm{free})}\rho^{(\mathrm{free})} = v_{\mathrm{free}}^{(\mathrm{B})}\rho_{\mathrm{free}}^{(\mathrm{B})}\ . \tag{3}$$

The flow rate q_{sum} downstream of the bottleneck in the state of free flow under consideration depends on the bottleneck type. In the cases of off-ramp and merge bottlenecks (a merge bottleneck is a bottleneck caused by a decrease in the number of highway lanes in the flow direction), we get

$$q_{\mathrm{sum}} = q_{\mathrm{in}}\ . \tag{4}$$

Under condition (4), the formula (1) and (3) are also valid.

Explanation of Traffic Breakdown Within Vehicle Cluster To explain traffic breakdown within the vehicle cluster, we use the hypothesis of three-phase traffic theory about a double Z-speed-density characteristic (Fig. 3) [17], ► Traffic Congestion, Modeling Approaches to. In this theory, the first Z-speed-density characteristic between states F and S explains traffic breakdown (arrow $F \to S$ in Fig. 3). In accordance with the double Z-characteristic (Fig. 3), at a given value q_{on} there should be non-

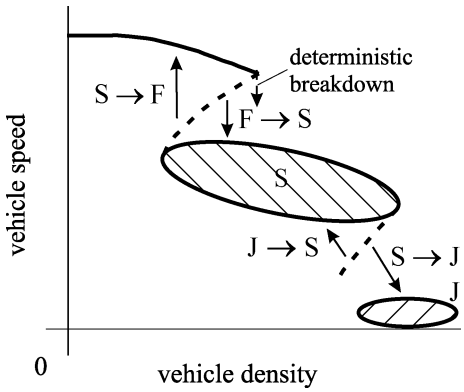

Traffic Breakdown, Probabilistic Theory of, Figure 3
Qualitative double Z speed-density characteristic for phase transitions of three-phase traffic theory. Taken from [17], ► Traffic Congestion, Modeling Approaches to

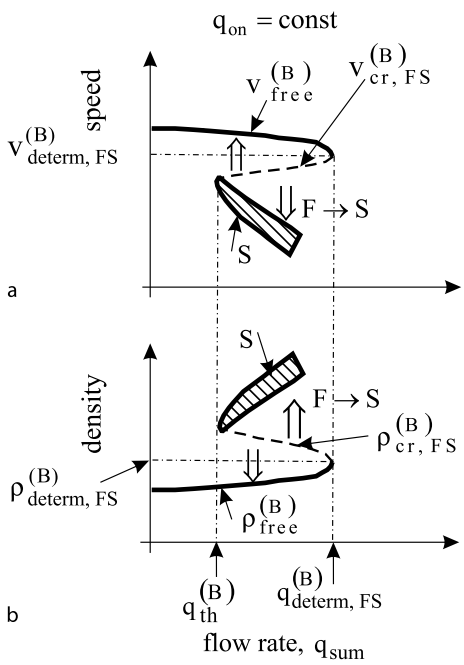

Traffic Breakdown, Probabilistic Theory of, Figure 4
Qualitative shapes of speed (a) and density (b) dependences on the flow rate q_{sum} associated with the location of vehicle cluster

monotonous dependence of speed on the flow rate q_{sum} at the location of vehicle cluster (Fig. 4a); this speed-flow characteristic leads to a *S*-shaped density dependence on this flow rate (Fig. 4b).

When q_{on} is a given value and the flow rate q_{sum} increases due to an increase in q_{in}, the vehicle speed in free flow within the deterministic cluster $v^{(B)}_{free}$ decreases and in accordance with (3) the associated density $\rho^{(B)}_{free}$ increases. However, this increase is limited by some critical density $\rho^{(B)}_{free} = \rho^{(B)}_{determ,FS}$ (Fig. 4b) within the deterministic cluster associated with a critical flow rate

$$q_{sum} = q^{(B)}_{determ,FS} \,. \tag{5}$$

After this critical deterministic cluster is reached, the further increase in q_{sum} should lead to *deterministic traffic breakdown* at the bottleneck causing spontaneous synchronized flow emergence at the bottleneck. The critical deterministic cluster can be considered *deterministic nucleus* for traffic breakdown. After the critical deterministic cluster is reached, deterministic traffic breakdown occurs at the bottleneck even if there were no random fluctuations in free flow at the bottleneck.

Due to fluctuations in real free flow, spontaneous traffic breakdown ($F \to S$ transition) can occur within the flow rate range

$$q^{(B)}_{th} \le q_{sum} < q^{(B)}_{determ,FS} \,, \tag{6}$$

i. e., before the deterministic nucleus for traffic breakdown is reached (arrows labeled $F \to S$ in Fig. 4). In (6), the characteristic flow rate $q^{(B)}_{th}$ is the threshold flow rate for traffic breakdown, which is considerably less than $q^{(B)}_{determ,FS}$: as mentioned above, at $q_{sum} < q^{(B)}_{th}$, the breakdown probability is equal to zero.

To explain the physics of this traffic breakdown, we note that due to fluctuations in real free flow at the bottleneck, the cluster size N can exceed the deterministic cluster size $N^{(determ)}$. In this case, the vehicle density increases and speed decreases within the cluster in free flow at the bottleneck (Fig. 2d,e) in comparison with the density and speed within the deterministic cluster (Fig. 2b,c). In other words, due to fluctuations in real free flow, the speed and density within the vehicle cluster at the bottleneck and therefore the cluster size N can differ considerably from, respectively, the values $v^{(B)}_{free}$, $\rho^{(B)}_{free}$, and $N^{(determ)}$ for the deterministic cluster. In general, the vehicle cluster localized at the bottleneck with the size N can be considered consisting of two components: (a) the deterministic cluster with the size $N^{(determ)}$ and (b) a random cluster component with the size $N - N^{(determ)}$.

There are a critical speed denoted by $v^{(B)}_{cr,FS}$ and the associated critical density $\rho^{(B)}_{cr,FS}$, which depend on the flow rate (Fig. 4). When within the cluster the speed is equal to or lower than $v^{(B)}_{cr,FS}$, respectively the density is equal to or greater than $\rho^{(B)}_{cr,FS}$, then traffic breakdown occurs spontaneously. Thus when the cluster size N exceeds some critical cluster size associated with a critical density $\rho^{(B)}_{cr,FS}$ within the cluster (Fig. 4b), then traffic breakdown occurs.

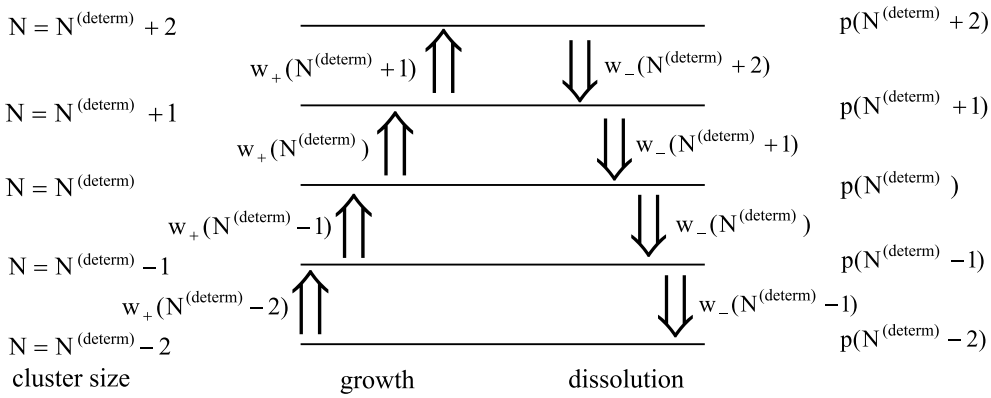

Traffic Breakdown, Probabilistic Theory of, Figure 5
Schematic illustration of vehicle cluster transformation due to fluctuations in an initially non-homogeneous free flow at bottleneck. Taken from [20,21,22]

The cluster with the critical size and therefore critical density within the cluster can be considered a critical vehicle cluster or a nucleus for traffic breakdown at the bottleneck. Otherwise, if the cluster size N is smaller than the critical one, the cluster decays towards the initial deterministic cluster with the cluster size $N^{(\text{determ})}$. Thus fluctuations in the cluster size in the neighborhood of $N = N^{(\text{determ})}$ are responsable for traffic breakdown under condition (6) (Fig. 5; a more detailed explanation of this figure appears in Subsect. "Probabilistic Traffic Breakdown Model Based on Three-Phase Traffic Theory").

Probabilistic Description of Traffic Breakdown with Cellular Automata (CA) Traffic Flow Model

The Kerner–Klenov–Wolf (KKW) cellular automata (CA) model is the first CA traffic flow model developed in the framework of three-phase traffic theory, which can show and predict empirical features of traffic breakdown and resulting congested patterns [23]. Simulations of this model show that in accordance with the nucleation model of traffic breakdown considered above there is a vehicle cluster, i. e., a local density (and speed) disturbance in free flow, which is localized at the bottleneck. In the KKW-model, as in the nucleation model of Subsect. "Model of Traffic Breakdown at Highway Bottleneck", traffic breakdown in free flow at the bottleneck occurs mostly at the location of the vehicle cluster.

Moreover, in the KKW-model traffic breakdown is also a first-order $F \to S$ transition: (i) in a finite flow rate range, there can be spontaneous and induced breakdowns at the same bottleneck, (ii) there is a hysteresis effect associated with traffic breakdown ($F \to S$ transition) and a return $S \to F$ transition (iii) there is a random time delay of traffic breakdown, and (iv) the breakdown probability is an increasing function of the flow rate q_{sum} (see below). For these reasons, we briefly discuss this traffic flow model as well as a probabilistic theory of traffic breakdown associated with the model.

KKW Three-Phase Traffic Flow CA Model

Basic driver behavioral assumptions of the KKW CA model are similar to those of the our stochastic three-phase traffic flow model [18,19] discussed in ► Traffic Congestion, Modeling Approaches to:

(i) *Fundamental hypothesis of three-phase traffic theory.* In synchronized flow, a driver accepts a range of different hypothetical steady state speeds at the same space gap (space gap is the net distance between vehicles) to the preceding vehicle. This means that hypothetical steady model states of synchronized flow cover a 2D-region in the flow-density plane (Fig. 6a). The boundaries of this 2D region F, L, and U are respectively associated with free flow, a synchronization space gap, and a safe space gap determined through a safe speed. The 2D region of the steady states is associated with a driver behavioral assumption that in synchronized flow a driver is able to recognize whether the space gap is increasing or decreasing regardless of the speed difference to the preceding vehicle.

(ii) *Line J and 2D region of steady states.* In the model, the line J, which represents the motion of the downstream jam front in the flow-density plane (the line J

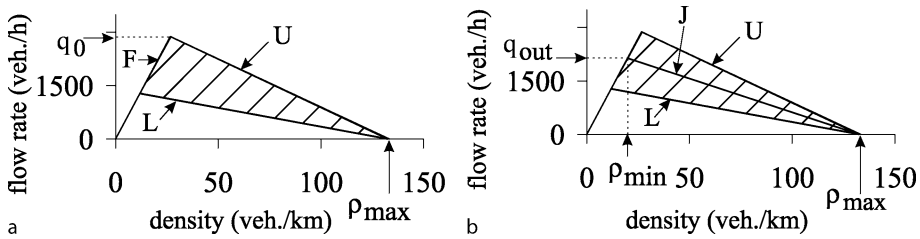

Traffic Breakdown, Probabilistic Theory of, Figure 6
Steady states of the KKW CA model and the line *J*. Taken from [23]

is explained in ► Traffic Congestion, Modeling Approaches to), is between the boundaries L and U (Fig. 6b), i. e., the line J divides the 2D region of steady states of synchronized flow onto two classes: the states on and above the line J and the states below the line J, which are metastable and stable states with respect to wide moving jam formation, respectively. Thus at a given steady speed, a driver behavioral assumption is that the space gap in synchronized flow associated with the line J, i. e., between the jams is greater than the safe one and it is smaller than the synchronization gap.

(iii) *Speed adaptation effect in synchronized flow*. The speed adaptation effect takes place, when the vehicle cannot pass the preceding vehicle, within the space gap range:

$$g_{s,n} \le g_n \le G_n \,, \tag{7}$$

where $g_n = x_{\ell,n} - x_n - d$ is the space gap, x_n is the vehicle co-ordinate, the lower index ℓ marks functions and values related to the preceding vehicle; G_n is a synchronization gap; $g_{s,n}$ is a safe gap determined from the equation $v_n = v_{s,n}$, in which v_n is the vehicle speed, $v_{s,n} = g_n/\tau$ is a safe speed taken as in the Nagel–Schreckenberg (NaSch) CA model [33]; all vehicles have the same length d, which includes the minimum space gap between vehicles within a wide moving jam; index n corresponds to the discrete time $t = n\tau, n = 0, 1, 2, \ldots$; and τ is time step. Under condition (7), the vehicle tends to adjust its speed to the preceding vehicle without caring, what the precise space gap is, as long as it is safe. For example, at a given time-independent speed of the preceding vehicle $v_{\ell,n} = v_\ell =$ const, this speed adaptation leads to car following with $v_n = v = v_\ell$ at a time-independent space gap $g_n = g$. There is a multitude of these gaps associated with the same speed $v = v_\ell$. These gaps lie between the synchronization gap and safe gap, i. e., there is no desired (or optimal) space gap in synchronized flow.

(iv) *Over-acceleration effect*. In synchronized flow of a lower density, a driver searches for the opportunity to accelerate and to pass. A competition between the speed adaptation (item (iii)) and over-acceleration effects simulates traffic breakdown ($F \to S$ transition). The over-acceleration is simulated as a collective effect, which occurs on average in traffic flow, through the use of lane changing to a faster lane as well as through the use of random vehicle acceleration. In KKW CA model, both random acceleration and random deceleration are simulated by adding of the same random term $a\tau\eta_n$ to vehicle speed, where

$$\eta_n = \begin{cases} -1 & \text{if } r < p_b, \\ 1 & \text{if } p_b \le r < p_b + p_a, \\ 0 & \text{otherwise}\,, \end{cases} \tag{8}$$

p_a is probability of random acceleration, p_b is probability of random deceleration, $p_a + p_b \le 1$, $r = \text{rand}(0, 1)$ is a random number uniformly distributed between 0 and 1, a is the vehicle acceleration. The random vehicle acceleration is described by the second line in (8).

(v) *Pinch effect in synchronized flow*. Moving in synchronized flow, a driver comes on average closer to the preceding vehicle over time that should explain the pinch effect. This effect is simulated through the increase in random vehicle acceleration at low vehicle speeds. For this purpose, the probability p_a is chosen to be a decreasing speed function

$$p_a(v_n) = \begin{cases} p_{a1} & \text{if } v_n < v_p \\ p_{a2} & \text{if } v_n \ge v_p, \end{cases} \tag{9}$$

where p_{a1}, p_{a2}, and v_p are constants; $p_{a1} > p_{a2}$.

(vi) *Over-deceleration effect*. In the model, a competition between a well-known over-deceleration associated with driver reaction time and the speed adaptation

effect determines moving jam emergence in synchronized flow. As in the NaSch CA model [33], the over-deceleration is also simulated as a collective effect through the use of random fluctuations in vehicle deceleration. The random vehicle deceleration is described by the first line in formula (8).

(vii) *Driver time delay in acceleration*. In the model, this well-known effect should describe driver delay in acceleration at the downstream front of synchronized flow or wide moving jam after the preceding vehicle has begun to accelerate. A driver time delay in acceleration is simulated as a collective effect through the use of a random vehicle deceleration (8). Indeed, a vehicle that has to accelerate at the current time step can remain its speed due to applying of random vehicle deceleration. In the KKW CA model, as in a version of the NaSch CA model made in [1], the slow-and-start rules are used to describe a longer driver time delay in acceleration when the vehicles begins to accelerate from a standstill:

$$p_b(v_n) = \begin{cases} p_0 & \text{if } v_n = 0 \\ p & \text{if } v_n > 0 \,, \end{cases} \tag{10}$$

where p, $p_0 > p$ are constants [1]; consequently, the mean time in vehicle acceleration is

$$\tau_{\text{del}}^{(a)} = \frac{\tau}{1 - p_0} \,. \tag{11}$$

Thus the basis of the KKW three-phase CA model are driver behavioral assumptions made in three-phase traffic theory (items (i)-(v)). In addition, over-deceleration (item vi) and driver time delay in acceleration (item vii) introduced in the NaSch-model in the framework of the fundamental diagram approach [1,33] (see review ▶ Traffic Congestion, Modeling Approaches to in this Encyclopedia) have also been incorporated. Finally, the KKW model reads as follows:

$$v_{n+1} = \max(0, \min(\tilde{v}_{n+1} + a\tau\eta_n,\ v_n + a\tau,\ v_{\text{free}},\ v_{s,n})) \,, \tag{12}$$

$$x_{n+1} = x_n + v_{n+1}\tau \,, \tag{13}$$

$$\tilde{v}_{n+1} = \max(0, \min(v_{\text{free}}, v_{c,n}, v_{s,n})) \,, \tag{14}$$

$$v_{c,n} = \begin{cases} v_n + a\tau & \text{at } g_n > G_n \,, \\ v_n + a\tau\,\text{sign}(v_{\ell,n} - v_n) & \text{at } g_n \le G_n \,, \end{cases} \tag{15}$$

where $\tilde{v}_n$ is the speed at time step n without fluctuating part, v_{free} is the maximum speed in free flow. The synchronization gap G_n is taken as speed function:

$$G(v_n) = kv_n\tau \,, \tag{16}$$

where k is constant.

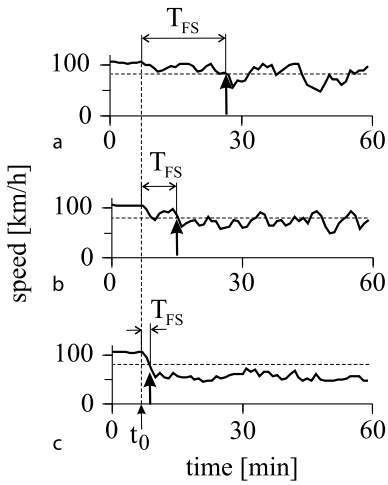

Traffic Breakdown, Probabilistic Theory of, Figure 7
Time delays T_{FS} of traffic breakdown ($F \to S$-transition) at on-ramp bottleneck at the same q_{in} for different flow rates to the on-ramp q_{on}: a 70, b 90, c 120 vehicles/h. *Up-arrows* mark the time $t_0 + T_{FS}$ at which traffic breakdown has occurred. At time $t = t_0$ on-ramp inflow is switched on. Taken from [23].

Time Delay and Probability of Traffic Breakdown

Simulations of the KKW CA model [23] show that at a given flow rate in free flow on the main road upstream of the bottleneck q_{in}, free flow is metastable with respect to a local first-order $F \to S$ transition, which occurs spontaneously at the bottleneck, when the flow rate q_{sum} in free flow downstream of the bottleneck is satisfied condition (6).

Traffic breakdown is characterized by a random time delay $T_{FS}^{(B)}$ (Fig. 7): in different realizations, at the same q_{on} and q_{in} usually different $T_{FS}^{(B)}$ are found. The smaller q_{on} and/or q_{in} are, the longer the mean time delay $T_{FS}^{(B,mean)}$ of traffic breakdown is.

Theoretical probability for an $F \to S$ transition at the bottleneck (Fig. 8) shows the same features as empirical probability for $F \to S$ transition [35] (see ▶ Traffic Congestion, Modeling Approaches to). To study the $F \to S$ transition probability (Fig. 8), a large number of realizations (runs) have been calculated for given flow rates $q_{sum} = q_{on} + q_{in}$ and q_{on} during a given time interval T_0. At the beginning of each run there was free flow at the on-ramp. For each run it was checked, whether the $F \to S$

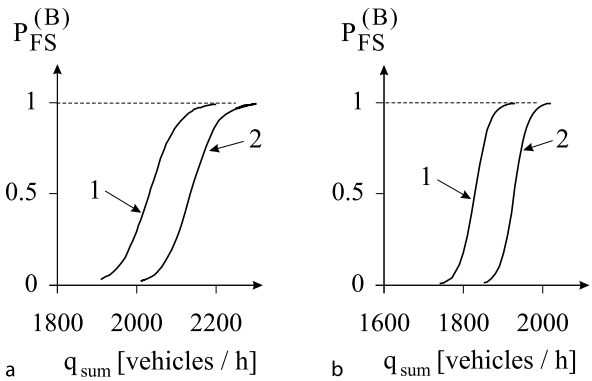

Traffic Breakdown, Probabilistic Theory of, Figure 8
Probability $P_{FS}^{(B)}$ that an $F \to S$ transition occurs at the bottleneck within $T_0 = 30$ min (curve 1) or already within $T_0 = 15$ min (curve 2), after the on-ramp inflow was switched on, versus the traffic demand upstream of the on-ramp, $q_{sum} = q_{in} + q_{on}$. Results are shown for two different flow rates to the on-ramp, $q_{on} = 60$ vehicles/h in (a) and $q_{on} = 200$ vehicles/h in (b). Taken from [23]

transition at the on-ramp occurred during the time interval T_0 or not. The result of these simulations is the number of realizations n_P where the $F \to S$ transition at the on-ramp had occurred in comparison with the number of all realizations N_P. Then the probability for an $F \to S$ transition is $P_{FS}^{(B)} = n_P/N_P$.

The flow rate q_{sum} was changed and the procedure with all realizations was repeated at the same flow rate to the on-ramp q_{on}. The flow rate q_{sum} at which the $F \to S$ transition at the on-ramp occurred in all realization is therefore related to the probability of the $F \to S$ transition $P_{FS}^{(B)} = 1$. We found that lower flow rates q_{sum} correspond to $P_{FS}^{(B)} < 1$. As expected, we found an increase of the probability $P_{FS}^{(B)}$ as a function of the flow rate q_{sum} at a given flow rate to the on-ramp q_{on}; the related function found numerically can be approximated by the formula [23]

$$P_{FS}^{(B)} = \frac{1}{1 + \exp[\alpha(q_P - q_{sum})]}\,, \tag{17}$$

where α and q_P are functions of q_{on} and T_0.

Probabilistic Description of Traffic Breakdown Based on Master Equation

The nucleation model for traffic breakdown of Subsect. "Model of Traffic Breakdown at Highway Bottleneck", which is confirmed by numerical simulations of the KKW CA model discussed above, can be used for an analytical study of the nucleation rate, breakdown probability, and mean time delay of traffic breakdown at a bottleneck based on the well-known master equation approach for probabilistic description of first-order phase transitions in complex systems of diverse nature [7].

Master Equation and One-Step Processes

Traffic flow variables (speed, density, occupancy, flow rate) within the vehicle cluster localized at a bottleneck (Subsect. "Model of Traffic Breakdown at Highway Bottleneck") are time dependent random variables. A time dependent random variable $X(t)$ can be characterized by a set of joint probabilities [7]

$$p(x_1, t_1; x_2, t_2; \ldots; x_n, t_n), \quad n = 1, 2, \ldots, \tag{18}$$

where $x_1, x_2, \ldots, x_n$ are random values of variable X at time instants $t_1 \le t_2 \le \cdots \le t_n$. The joint probabilities $p(x_1, t_1; \ldots; x_n, t_n)$ and $p(x_1, t_1; \ldots; x_{n-1}, t_{n-1})$ are connected each other by a conditional probability

$$p(x_n, t_n | x_1, t_1; \ldots; x_{n-1}, t_{n-1}), \quad n = 2, 3, \ldots, \tag{19}$$

as follows [7]

$$\begin{aligned} &p(x_1, t_1; \ldots; x_n, t_n) \\ &\quad = p(x_n, t_n | x_1, t_1; \ldots; x_{n-1}, t_{n-1}) \\ &\qquad \cdot p(x_1, t_1; \ldots; x_{n-1}, t_{n-1})\,. \end{aligned} \tag{20}$$

The basic assumption of *Marcovian* stochastic process is that the conditional probability $p(x_n, t_n | x_1, t_1; \ldots; x_{n-1}, t_{n-1})$ depends on the value of random variable X at the last previous time moment t_{n-1} only [7]

$$p(x_n, t_n | x_1, t_1; \ldots; x_{n-1}, t_{n-1}) = p(x_n, t_n | x_{n-1}, t_{n-1}). \tag{21}$$

In this case, according to (20) all joint probabilities $p(x_1, t_1; \ldots; x_n, t_n)$ (18) at $n > 1$ can be expressed in terms of the conditional probability (21) and the probability $p(x_1, t_1)$ as

$$\begin{aligned} &p(x_1, t_1; \ldots; x_n, t_n) \\ &\quad = p(x_1, t_1) \prod_{k=2}^{n} p(x_k, t_k | x_{k-1}, t_{k-1})\,. \end{aligned} \tag{22}$$

In the nucleation model of Subsect. "Model of Traffic Breakdown at Highway Bottleneck" is assumed that there is a vehicle cluster localized at a bottleneck and this cluster exists in free flow even if there were no fluctuations in the flow. In the model, the cluster dynamics describes traffic breakdown ($F \to S$ transition) at the bottleneck. For

an analytical study of this nucleation model, we will further make the following additional assumptions: (i) The cluster dynamics is associated with a Marcovian stochastic process; (ii) the random variable X of this process is associated with the total number of vehicles in the cluster N only (see Subsect. "Outflow Rate from Cluster" for more detail). Then X is a discrete variable. In this case, the Chapman–Kolmogorov equation for the conditional probability $p(x, t|x', t')$ reads [7]:

$$p(x, t|x', t') = \sum_{x''} p(x, t|x'', t'')p(x'', t''|x', t') , \quad (23)$$

where $t \geq t'' \geq t'$, the summation is taken over all possible values of x''. In turn, the conditional probability $p(x, t|x', t')$ determines the time evolution of the probability $p(x, t)$ as

$$p(x, t) = \sum_{x'} p(x, t|x', t')p(x', t') . \quad (24)$$

The differential equation for the conditional probability $p(x, t|x', t')$ that follows from (23) can be written as [7]

$$\frac{\partial p(x, t|x', t')}{\partial t} = \sum_{y \neq x} \Big[W(x, y, t)p(y, t|x', t') - W(y, x, t)p(x, t|x', t') \Big] , \quad (25)$$

where

$$W(x, y, t) = \lim_{\Delta t \to 0} p(x, t + \Delta t|y, t)/\Delta t \quad \text{at } x \neq y \quad (26)$$

is the transition rate, i. e., the probability per unit time for the transition from y to $x \neq y$ at time t. The Eq. (25) is usually called *master equation*. Multiplying (25) by $p(x', t')$ and using (24), one can write (25) as the equation for probability $p(x, t)$ [7]

$$\frac{\partial p(x, t)}{\partial t} = \sum_{y \neq x} \big[W(x, y, t)p(y, t) - W(y, x, t)p(x, t) \big] . \quad (27)$$

In the model of traffic breakdown (Subsect. "Model of Traffic Breakdown at Highway Bottleneck" and Subsect. "Outflow Rate from Cluster"), we will make once more assumption that at each time instant the total number of vehicles within the cluster N can be changed by one vehicle only. This assumption is associated with a *one-step process*, also called *birth-and-death process*, when transitions between neighboring states $X = x$ and $X = x \pm 1$ can occur only [7]. The transition rate $W(x, y, t)$ for one-step processes is [7]

$$W(x, y, t) = w_+(y)\delta_{x,y+1} + w_-(y)\delta_{x,y-1} , \quad (28)$$

where $w_+(y)$ is the rate of transition from y to $x = y + 1$, $w_-(y)$ is the rate of transition from y to $x = y - 1$; $\delta_{x,y}$ is Kroneker symbol.

Substituting (28) into the master Eq. (27), one finds the master equation for one-step processes [7]

$$\begin{aligned} \frac{\partial p(x, t)}{\partial t} &= w_+(x - 1)p(x - 1, t) \\ &\quad + w_-(x + 1)p(x + 1, t) \\ &\quad - [w_+(x) + w_-(x)]p(x, t) . \end{aligned} \quad (29)$$

For the master Eq. (29), the mean first passage time T for transition from an initial point $x = a$ to a finite point $x = b$ is given by formula [7]

$$T = \sum_{x=a}^{b} \left[(w_+(x)p_s(x))^{-1} \sum_{y=0}^{x} p_s(y) \right] , \quad (30)$$

where $p_s(x)$ is a steady solution of (29):

$$p_s(x) = p_s(0) \prod_{y=1}^{x} \frac{w_+(y - 1)}{w_-(y)} \quad \text{at } x > 0 , \quad (31)$$

the reflecting boundary condition

$$w_-(0) = 0 \quad (32)$$

is assumed to be satisfied at the point $x = 0$.

If the points a and b are separated by a potential barrier and therefore $p_s^{-1}(x)$ has a strong maximum at a point $x = c$ between a and b, the formula (30) is reduced to [7]:

$$T = \frac{1}{w_+(c)} \sum_{x=0}^{c} p_s(x) \sum_{x=a}^{b} p_s^{-1}(x) . \quad (33)$$

We will use this formula for the evaluation of the mean time delay for traffic breakdown $T_{FS}^{(B,mean)}$ in Subsect. "Nucleation Rate and Mean Time Delay of Traffic Breakdown at Bottlenecks".

Critical Discussion of Earlier Probabilistic Traffic Breakdown Models

It must be noted that models for vehicle cluster nucleation in traffic flow based on the master Eq. (29), which should describe traffic breakdown, have firstly been in-

troduced by Mahnke et al. and Kühne et al. [24,25,28,29] (see review [30]). However, in [21,22] we have explained in detail that and why the models of Ref. [24,25,28,29,30] do not explain the fundamental empirical features of traffic breakdown, i. e., they cannot describe the probabilistic traffic breakdown in real traffic flow. In particular, the nucleation models of Ref. [24,28,29,30] describe the nucleation of a wide moving jam in free flow ($F \to J$ transition). This is in contrast with the fundamental empirical feature **A** of traffic breakdown (Subsect. "Fundamental Empirical Features of Traffic Breakdown").

Briefly, the most crucial differences of the nucleation models of Ref. [24,25,28,29,30] with the model of traffic breakdown of Subsect. "Model of Traffic Breakdown at Highway Bottleneck" [15,17,20,21,22] are as follows.

The basic assumption of the nucleation models of Ref. [24,25,28,29,30] is that there are *no* vehicle clusters on a road without fluctuations. This basic assumption is made both for a homogeneous road [24,28,29,30] and for a road with a bottleneck [25,30]. It is suggested that for vehicle cluster emergence firstly a random pre-cluster should occur from fluctuations in free flow. This pre-cluster should forego subsequent cluster evolution towards a critical cluster (nucleus) whose growth should lead to traffic breakdown. In other words, it is suggested that without random pre-cluster the number of vehicles within the cluster, i. e., the cluster size is exactly equal to *zero* and, consequently, without random fluctuations no traffic breakdown is possible.

In contrast with the nucleation models of Ref. [24,25,28,29,30], in the nucleation model of Subsect. "Model of Traffic Breakdown at Highway Bottleneck" [20,21,22] the cluster size N is equal to the *total* number of vehicles within the cluster. Therefore, even if there were no random fluctuations in free flow, rather than the cluster size is equal to zero, the cluster size $N = N^{(\text{determ})}$, i. e., this is equal to the size of the deterministic cluster (Fig. 5) and, consequently, the deterministic traffic breakdown is possible [20,21,22]. The nucleation model of Subsect. "Model of Traffic Breakdown at Highway Bottleneck" results from the model assumption that an initial free flow at a bottleneck is non-homogeneous, i. e., a vehicle cluster exists in free flow even if random fluctuations in traffic flow were not taken into account.

In the model of Subsect. "Model of Traffic Breakdown at Highway Bottleneck" [20,21,22], due to fluctuations in real free flow, the cluster size N changes in a neighborhood of $N = N^{(\text{determ})}$ (Fig. 5). These fluctuations in the cluster size lead to traffic breakdown before the condition for the deterministic breakdown is satisfied. Below we will see that this nucleation model explains all fundamental empirical features of traffic breakdown (Subsect. "Fundamental Empirical Features of Traffic Breakdown").

Probabilistic Traffic Breakdown Model Based on Three-Phase Traffic Theory

Outflow Rate from Cluster An important characteristic of the nucleation model of Subsect. "Model of Traffic Breakdown at Highway Bottleneck" is the outflow rate from the vehicle cluster $q_{\text{down}}^{(\text{bottle})}$ (Figs. 2a and 9). In contrast with the on-ramp and merge bottlenecks for which there is only one outflow from the cluster, for the off-ramp bottleneck, there are two different outflows from the cluster, which are related to the flow rate on the main road downstream of the merging region of the off-ramp denoted by $q_{\text{down}}^{(\text{main})}$ and to the flow rate to the off-ramp road denoted by q_{off}. This means that for the off-ramp bottleneck (Fig. 9b)

$$q_{\text{down}}^{(\text{bottle})} = q_{\text{down}}^{(\text{main})} + q_{\text{off}} \,. \tag{34}$$

In the nucleation model for traffic breakdown discussed in Subsect. "Model of Traffic Breakdown at Highway Bottleneck", we make further an assumption that the outflow rate from the cluster depends on the total number of vehicles within the cluster, i. e., on the cluster size N (Fig. 9) [20,21,22]:

$$q_{\text{down}}^{(\text{bottle})} = q_{\text{down}}^{(\text{bottle})}(N) \,. \tag{35}$$

To determine $q_{\text{down}}^{(\text{bottle})}(N)$, we can use the flow rate dependence of the density (Fig. 4b). However, in the nucleation model for traffic breakdown [20,21,22], which we discuss below, we make an approximation in which we average the infinite synchronized flow states for each density

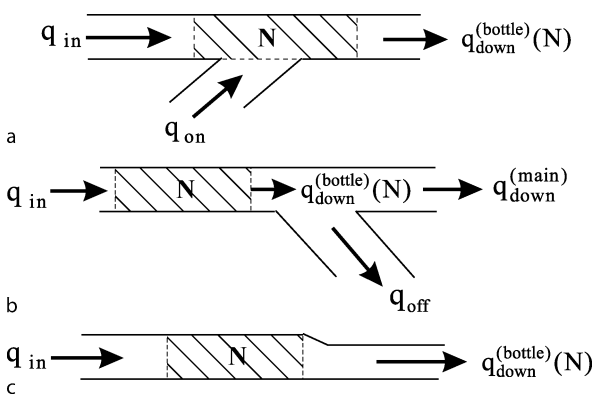

Traffic Breakdown, Probabilistic Theory of, Figure 9
Qualitative explanation of vehicle cluster (dashed regions) at bottlenecks: a On-ramp bottleneck. b Off-ramp bottleneck. c Merge bottleneck

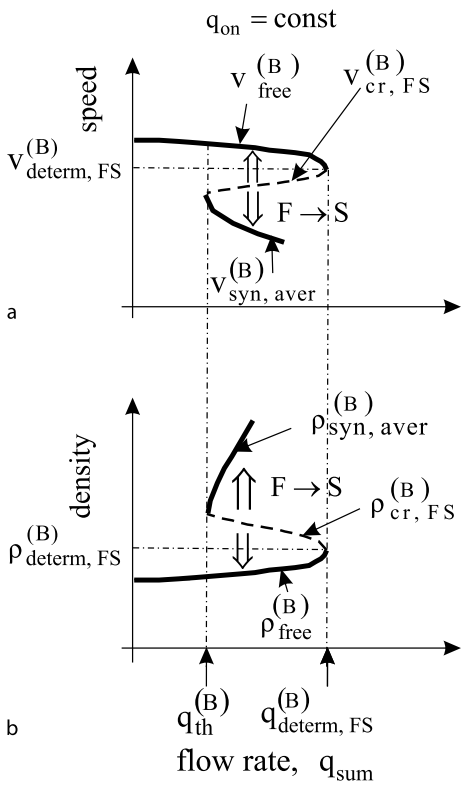

Traffic Breakdown, Probabilistic Theory of, Figure 10
Simplified qualitative shapes of speed (a) and density (b) dependences on the flow rate q_{sum} associated with location of vehicle cluster. Taken from [20,21,22]

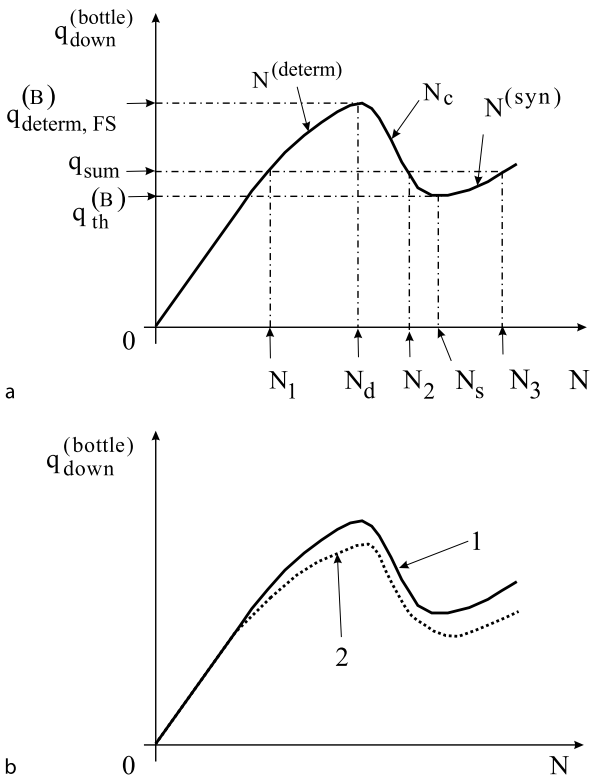

Traffic Breakdown, Probabilistic Theory of, Figure 11
Qualitative dependencies of the outflow rate $q^{(bottle)}_{down}$ on the total vehicle number N within the cluster localized at the bottleneck (a), and possible dependencies of the N-shaped function $q^{(bottle)}_{down}(N)$ on q_{on} for two different values q_{on} (b); curve 1 for $q_{on} = q^{(1)}_{on}$, curve 2 for $q_{on} = q^{(2)}_{on} > q^{(1)}_{on}$

to one average speed. Thus in this approximation, rather than Fig. 4, we use the related simplified non-monotonous dependences shown in Fig. 10.

In the model, it is further assumed that the shape of the characteristic $q^{(bottle)}_{down}(N)$ (Fig. 11) follows from the S-shaped density-flow characteristic shown in Fig. 10b. In this nucleation model, the characteristic $q^{(bottle)}_{down}(N)$ has at least two different branches $q^{(bottle)}_{down}(N)$ labeled $N^{(determ)}$ and N_c in Fig. 11a. These branches are related to the vehicle number ranges, respectively, given by the conditions

$$0 \le N \le N_d \tag{36}$$

and

$$N_d < N \le N_s \,. \tag{37}$$

The branches $N^{(determ)}$ and N_c in Fig. 11a are associated with the density branches $\rho^{(B)}_{free}$ and $\rho^{(B)}_{cr,FS}$ of the S-shaped density-flow characteristic in Fig. 10b, respectively. The branch $N^{(determ)}$ is associated with the case in which at a high enough flow rate q_{sum} and the on-ramp flow rate $q_{on} > 0$ the deterministic cluster exists at the bottleneck under free flow conditions. The branch N_c is associated with the case in which the critical cluster whose growth leads to an $F \to S$ transition occurs at the bottleneck.

In addition, from the S-shaped density-flow characteristic (Fig. 10b) can be seen that at

$$N > N_s \,, \tag{38}$$

there can be a third branch $N^{(syn)}$ on the characteristic $q^{(bottle)}_{down}(N)$ (Fig. 11a) associated with the branch $\rho^{(B)}_{syn,aver}$ for averaged synchronized flow states in Fig. 10b. In this case, $q^{(bottle)}_{down}(N)$ (35) is a N-shaped flow-vehicle-number characteristic. It should be noted that the branch for synchronized flow $v^{(B)}_{syn,aver}$ in Fig. 10b and the associated branch $N^{(syn)}$ on the characteristic $q^{(bottle)}_{down}(N)$ (Fig. 11a) follow from the microscopic three-phase traffic theory, rather than from the nucleation model. This branch, which

can be drawn for the case only in which a localized SP (LSP) occurs as a result of the breakdown, is shown with the aim of a qualitative illustration of a possible traffic flow state after synchronized flow nucleation. This branch has no influence on the nucleation rate and probability of an $F \to S$ transition discussed below. In other words, the nucleation effect leading to traffic breakdown and its characteristics are fully independent of possible congested patterns resulting from this $F \to S$ transition. These congested patterns, which cannot be described by the nucleation model, are discussed in [17], ► Traffic Congestion, Spatiotemporal Features of.

At the critical point $N = N_d$ at which the branches $N^{(\text{determ})}$ and N_c merges, the function $q_{\text{down}}^{(\text{bottle})}(N)$ has its maximum point. At the threshold point $N = N_s$ at which the branches N_c and $N^{(\text{syn})}$ merges, the function $q_{\text{down}}^{(\text{bottle})}(N)$ has its minimum point.

In the case of the on-ramp bottleneck, quantitative characteristics of the N-shaped function $q_{\text{down}}^{(\text{bottle})}(N)$ (e. g., values N_d and N_s) can depend on the flow rate to the on-ramp q_{on}. This is because the on-ramp inflow and the flow upstream of the bottleneck can make a considerable different influence on the cluster size and the outflow rate from the cluster $q_{\text{down}}^{(\text{bottle})}$. In particular, it can turn out that at the same N the greater q_{on}, the more difficult for vehicles to escape from the cluster, i. e., the less $q_{\text{down}}^{(\text{bottle})}$ is. This is confirmed by microscopic simulations [23] and reflected in Fig. 11b in which it is assumed that the greater q_{on}, the less $q_{\text{down}}^{(\text{bottle})}$ and the greater N_d and N_s are. Thus, in a case of the on-ramp bottleneck instead of (35) we should use

$$q_{\text{down}}^{(\text{bottle})} = q_{\text{down}}^{(\text{bottle})}(N, q_{\text{on}}) , \tag{39}$$

where q_{on} is a parameter.

In accordance with the nucleation model (Fig. 11a), critical cluster occurrence describes an $F \to S$ transition at the bottleneck. There are two reasons for this statement: (i) A vehicle cluster is on average motionless, i. e., fixed at the bottleneck. This is also related to the definition of synchronized flow whose downstream front is usually fixed at the bottleneck, whereas the downstream from of a wide moving jam propagates through any bottleneck while maintaining the mean downstream jam velocity [17]. (ii) The shape of the chosen function (35) in the nucleation model associated with this motionless cluster (Fig. 11a) follows from a Z-shaped speed-flow characteristic for traffic breakdown (Fig. 10b) found in a microscopic traffic flow theory. In this theory has been shown that if these two requirements are satisfied, then rather than an $F \to J$ transition, an $F \to S$ transition occurs at the bottleneck, as found in empirical observations.

Steady States Steady states of the vehicle number N within the cluster at a given q_{sum} are associated with solutions of the equation

$$q_{\text{sum}} = q_{\text{down}}^{(\text{bottle})}(N, q_{\text{on}}) , \tag{40}$$

for the on-ramp bottleneck and

$$q_{\text{sum}} = q_{\text{down}}^{(\text{bottle})}(N) , \tag{41}$$

for the off-ramp and merge bottlenecks. As can be seen from Fig. 11a, at given flow rates q_{on} and q_{in} that satisfy condition (6) there can be at least two steady states: $N = N_1$ associated with the deterministic cluster and $N = N_2$ associated with the critical cluster. These steady states are the roots of Eq. (40) (or Eq. (41)), i. e., they are associated with the intersection points of the horizontal line $q = q_{\text{sum}}$ with the branches $N^{(\text{determ})}$ and N_c of the characteristic $q_{\text{down}}^{(\text{bottle})}(N, q_{\text{on}})$ (Fig. 11a), respectively. In addition, as mentioned above if an LSP occurs as a result of an F→S transition, then there is a third root of Eq. (40) (or Eq. (41)), $N = N_3$, associated with the intersection point of the horizontal line $q = q_{\text{sum}}$ with the branch $N^{(\text{syn})}$ of the characteristic $q_{\text{down}}^{(\text{bottle})}(N, q_{\text{on}})$.

If the flow rate q_{sum} increases, then the critical vehicle number difference within the cluster

$$\Delta N_c = N_2 - N_1 \tag{42}$$

decreases. This critical vehicle number difference is associated with the vehicle number difference within the critical cluster and within the deterministic cluster at the bottleneck. The growth of the critical cluster leads to traffic breakdown at the bottleneck.

At the critical flow rate (5), we get $\Delta N_c = 0$: the steady states N_1 and N_2 merge into one point with the critical vehicle number $N = N_d$ at which

$$q_{\text{determ,FS}}^{(\text{B})} = q_{\text{down}}^{(\text{bottle})}(N_d, q_{\text{on}}) . \tag{43}$$

At $q_{\text{sum}} \geq q_{\text{determ,FS}}^{(\text{B})}$ the deterministic traffic breakdown should occur even if there were no random increase in the vehicle number within the deterministic cluster at the bottleneck.

If the flow rate q_{sum} decreases gradually, then the threshold flow rate

$$q_{\text{sum}} = q_{\text{th}}^{(\text{B})} \tag{44}$$

is reached at which the steady states N_2 and N_3 merge into one threshold steady state $N = N_s$ at which

$$q_{\text{th}}^{(\text{B})} = q_{\text{down}}^{(\text{bottle})}(N_s, q_{\text{on}}) . \tag{45}$$

Based on the nucleation model discussed above and on the master Eq. (29), we consider the dynamics of the vehicle cluster at a bottleneck (dashed regions in Fig. 9) [20,21,22]. The probability $p(N,t)$ that N vehicles occur within the cluster at time t is given by the master Eq. (29) in which the variable x is the total vehicle number within the cluster N:

$$\begin{aligned}\frac{\partial p(N,t)}{\partial t} &= w_+(N-1)p(N-1,t)\\ &+ w_-(N+1)p(N+1,t)\\ &- [w_+(N)+w_-(N)]p(N,t), \quad \text{at } N>0\,.\end{aligned} \tag{46}$$

The master Eq. (46) describes probability p for the total vehicle number within the cluster, i. e., the cluster size N, which randomly changes due to fluctuations in a neighborhood of the size $N^{(\text{determ})}$ of the deterministic cluster (Fig. 5) [20,21,22]. The size of the deterministic cluster $N^{(\text{determ})}$ does not depend on fluctuations. Random fluctuations cause either a decrease or an increase in the cluster size N in comparison with $N^{(\text{determ})}$. The fluctuations, which decrease the cluster size N, are associated with an increase in speed within the cluster. Therefore, these fluctuations can prevent the breakdown. In contrast, fluctuations, which increase the cluster size N, i. e., decrease the speed within the cluster, can cause traffic breakdown before the deterministic nucleus for traffic breakdown associated with condition (5) is reached.

In the master Eq. (46), w_+ and w_- are the attachment rate onto and detachment rate from the cluster, respectively. w_+ does not depend on N, i. e., regardless of N

$$w_+ = q_{\text{sum}}\,. \tag{47}$$

In contrast, w_- is given by the formula

$$w_- = q_{\text{down}}^{(\text{bottle})}\,, \tag{48}$$

i. e., w_- depends on N in accordance with (35) for off-ramp and merge bottlenecks, or else in accordance with (39) for the case of an on-ramp bottleneck; in the latter case, w_- depends on N and q_{on}. When $N=0$, the following boundary condition should be satisfied in (46)

$$w_-(0) = 0\,, \tag{49}$$

i. e., no vehicles can leave the cluster. In this case,

$$\frac{\partial p(0,t)}{\partial t} = w_-(1)p(1,t) - w_+(0)p(0,t), \quad \text{at } N=0\,. \tag{50}$$

Nucleation Rate and Mean Time Delay of Traffic Breakdown at Bottlenecks

As follows from the analysis of the model (46)-(50) and formula (33) made in [20,21,22], within the flow rate range (6) the mean time delay of an $F\to S$ transition at the bottleneck is

$$T_{\text{FS}}^{(\text{B,mean})} = 2\pi\tau_b \exp\{\Delta\Phi\}\,, \tag{51}$$

where a potential barrier for the $F\to S$ transition

$$\Delta\Phi = \Phi(N_2) - \Phi(N_1)\,, \tag{52}$$

the potential $\Phi(N)$

$$\Phi(N) = \begin{cases}\sum_{n=1}^{N} \ln\frac{w_-(n)}{w_+} & \text{at } N>0\,,\\ 0 & \text{at } N=0\,,\end{cases} \tag{53}$$

$$\tau_b = \left(w'_-(N_1)\mid w'_-(N_2)\mid\right)^{-\frac{1}{2}}\,, \tag{54}$$

$w'_-(N) = dw_-/dN$. Respectively, the nucleation rate for traffic breakdown is

$$G_{\text{FS}}^{(\text{B})} = \frac{1}{T_{\text{FS}}^{(\text{B,mean})}} = \frac{1}{2\pi\tau_b}\exp\{-\Delta\Phi\}\,. \tag{55}$$

It can be seen from (51) that the mean time delay for traffic breakdown decreases exponentionally with increase in potential barrier $\Delta\Phi$ (52). If in Fig. 12 the total flow rate increases from $q_{\text{sum}} = q_{\text{sum}}^{(1)}$ to $q_{\text{sum}} = q_{\text{sum}}^{(2)}$, which is close to the critical flow rate (5) for deterministic traffic breakdown, then the potential barrier $\Delta\Phi$ (52) decreases from $\Delta\Phi_1$ to $\Delta\Phi_2$.

In contrast, if the total flow rate decreases from $q_{\text{sum}} = q_{\text{sum}}^{(1)}$ to $q_{\text{sum}} = q_{\text{sum}}^{(3)}$, which is close to the threshold flow rate $q_{\text{th}}^{(B)}$ (44) for random traffic breakdown, then the potential barrier $\Delta\Phi$ (52) increases from $\Delta\Phi_1$ to $\Delta\Phi_3$ (Fig. 12). At the threshold point $q_{\text{sum}} = q_{\text{th}}^{(\text{B})}$ (44), the potential barrier $\Delta\Phi(N)$ reaches the maximum value

$$\Delta\Phi = \Phi(N_s) - \Phi(N_{\text{th}}), \tag{56}$$

where $N_{\text{th}} = N_1$ at $q_{\text{sum}} = q_{\text{th}}^{(\text{B})}$. As a result, the mean time delay $T_{\text{FS}}^{(\text{B,mean})}$ (51) strongly increases as q_{sum} approaches the threshold point $q_{\text{th}}^{(\text{B})}$. Under the condition

$$q_{\text{sum}} < q_{\text{th}}^{(\text{B})} \tag{57}$$

no traffic breakdown at the bottleneck regardless of a random increase in the vehicle number within the cluster is possible.

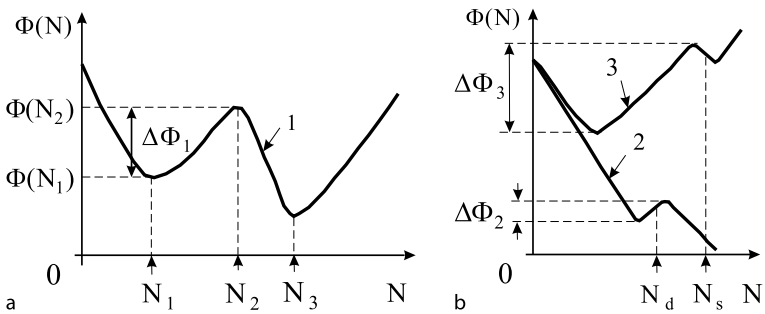

Traffic Breakdown, Probabilistic Theory of, Figure 12
Qualitative shape of the potential $\Phi(N)$ (53) for different flow rates q_{sum}: curves 1, 2, and 3 are related to the corresponding flow rates $q_{\text{sum}}^{(1)}$, $q_{\text{sum}}^{(2)}$, and $q_{\text{sum}}^{(3)}$ satisfying the condition $q_{\text{sum}}^{(3)} < q_{\text{sum}}^{(1)} < q_{\text{sum}}^{(2)}$. Characteristic values of the vehicle number within the cluster $N = N_i$, $i = 1, 2, 3$, $N = N_s$ and $N = N_d$ are discussed in Subsect. "Probabilistic Traffic Breakdown Model Based on Three-Phase Traffic Theory" (Fig. 11). Taken from [20,21,22]

If in the vicinity of the critical vehicle number N_d the function $w_-(N)$ can be approximated by a parabolic function of N, then the following approximate formula can be derived from (51)-(54):

$$T_{\text{FS}}^{(\text{B,mean})} = \frac{\sqrt{2}\pi N_d}{q_{\text{determ,FS}}^{(\text{B})}(\xi_d \Delta_c)^{1/2}} \left(\frac{1+\Delta_c^{1/2}}{1-\Delta_c^{1/2}}\right)^{2\sqrt{2/\xi_d}N_d} \times \exp\left(-\frac{4\sqrt{2}N_d\Delta_c^{1/2}}{\sqrt{\xi_d}}\right), \tag{58}$$

where

$$\xi_d = -(N^2 d^2 \ln w_-/dN^2)\big|_{N=N_d} \tag{59}$$

is a dimensionless value of the order of 1,

$$\Delta_c = \frac{q_{\text{determ,FS}}^{(\text{B})} - q_{\text{sum}}}{q_{\text{determ,FS}}^{(\text{B})}}, \tag{60}$$

i. e., Δ_c is the relative difference between the critical flow rate $q_{\text{determ,FS}}^{(\text{B})}$ for the deterministic $F \to S$ transition and the total flow rate q_{sum} (2). If in (58) $\Delta_c \ll 1$, then we get

$$T_{\text{FS}}^{(\text{B,mean})} = \frac{\sqrt{2}\pi N_d}{q_{\text{determ,FS}}^{(\text{B})}(\xi_d \Delta_c)^{1/2}} \exp\left(\frac{8N_d\Delta_c^{3/2}}{3\sqrt{2\xi_d}}\right). \tag{61}$$

Respectively, the nucleation rate $G_{\text{FS}}^{(\text{B})} = 1/T_{\text{FS}}^{(\text{B,mean})}$ for traffic breakdown at the bottleneck associated with (61) is

$$G_{\text{FS}}^{(\text{B})} = \frac{q_{\text{determ,FS}}^{(\text{B})}(\xi_d \Delta_c)^{1/2}}{\sqrt{2}\pi N_d} \exp\left(-\frac{8N_d\Delta_c^{3/2}}{3\sqrt{2\xi_d}}\right). \tag{62}$$

Note that $q_{\text{determ,FS}}^{(\text{B})}$, N_d and ξ_d for the on-ramp bottleneck depend q_{on}; in this case, $T_{\text{FS}}^{(\text{B,mean})}$ (61) and $G_{\text{FS}}^{(\text{B})}$ (62) are functions of q_{sum} and q_{on}.

As usual for each first-order phase transition observed in many other systems in natural science [7], the nucleation rate for traffic breakdown (62) is an exponentional function of Δ_c (60). For traffic flow, in accordance with (62) and (60) the exponential growth of the nucleation rate with Δ_c (60) is very sensible to the critical value for the deterministic breakdown phenomenon $q_{\text{determ,FS}}^{(\text{B})}$. This emphasizes the important impact of the deterministic cluster, which occurs at the on-ramp bottleneck at $q_{\text{on}} > 0$, on the nucleation rate for traffic breakdown (62) at a given flow rate q_{sum}.

Probability of Breakdown

The mean time delay $T_{\text{FS}}^{(\text{B,mean})}$ of traffic breakdown at the bottleneck enables us to find probability $P_{\text{FS}}^{(\text{B})}$ that traffic breakdown occurs at the bottleneck during a time interval for observing traffic flow T_{ob}. Let us consider probability

$$P_{\text{C}}^{(\text{B})} = 1 - P_{\text{FS}}^{(\text{B})} \tag{63}$$

that during a chosen time interval T_{ob} there is *no* $F \to S$ transition, i. e., free flow remains at the bottleneck. In accordance with a stochastic theory of dynamic systems with first-order phase transitions [7], a dependence of the probability $P_{\text{C}}^{(\text{B})}$ on T_{ob} governs by the equation

$$\frac{dP_{\text{C}}^{(\text{B})}}{dT_{\text{ob}}} = -\frac{P_{\text{C}}^{(\text{B})}}{T_{\text{FS}}^{(\text{B,mean})}}, \tag{64}$$

where $T_{\text{FS}}^{(\text{B,mean})}$ is constant at given values of q_{sum} and q_{on}. This equation derived with Kramer's method [7] is valid under condition that $T_{\text{FS}}^{(\text{B,mean})}$ is long enough in comparison with a relaxation time towards a metastable steady state with the deterministic cluster of the size

$N^{(\text{determ})} = N_1$. This relaxation time is of the order of τ_b (54), i. e., Eq. (64) is valid, if

$$\tau_b \ll T_{\text{FS}}^{(\text{B,mean})} . \tag{65}$$

Note that as follows from (51), condition (65) can be satisfied only if the potential barrier $\Delta\Phi$ in (51) is relatively high, i. e., $\Delta\Phi \gg 1$ (52).

In accordance with [7], the solution of (64) is

$$P_{\text{C}}^{(\text{B})}(T_{\text{ob}}) = P_0 \exp\left(-T_{\text{ob}}/T_{\text{FS}}^{(\text{B,mean})}\right) , \tag{66}$$

where P_0 is constant. If the time interval T_{ob} is short enough, i. e., $T_{\text{ob}} \ll T_{\text{FS}}^{(\text{B,mean})}$, but the condition

$$\tau_b \ll T_{\text{ob}} \tag{67}$$

is also satisfied, then probability $P_{\text{C}}^{(\text{B})}$ should be close to one, therefore, $P_0 \approx 1$ in (66). Then under conditions (65), (67) probability $P_{\text{C}}^{(\text{B})}$ is given by the approximate formula

$$P_{\text{C}}^{(\text{B})}(T_{\text{ob}}) = \exp\left(-T_{\text{ob}}/T_{\text{FS}}^{(\text{B,mean})}\right) . \tag{68}$$

In this case, in accordance with (63) probability $P_{\text{FS}}^{(\text{B})}$ of traffic breakdown at the bottleneck during the time interval T_{ob} is

$$P_{\text{FS}}^{(\text{B})}(T_{\text{ob}}) = 1 - \exp\left(-T_{\text{ob}}/T_{\text{FS}}^{(\text{B,mean})}\right) . \tag{69}$$

Since $T_{\text{FS}}^{(\text{B,mean})}$ in (69) is a function of q_{sum} and q_{on}, breakdown probability $P_{\text{FS}}^{(\text{B})} = P_{\text{FS}}^{(\text{B})}(T_{\text{ob}}, q_{\text{sum}}, q_{\text{on}})$. Probability of traffic breakdown $P_{\text{FS}}^{(\text{B})}$ for two different time intervals T_{ob} as functions of the flow rate q_{sum} calculated through formula (58), (60), and (69) is shown in Fig. 13. These dependences of breakdown probability on the flow rate are qualitatively the same as those found with the KKW-model (see Fig. 8).

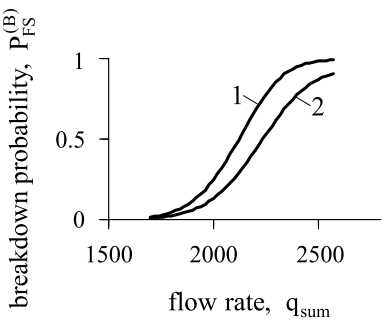

Traffic Breakdown, Probabilistic Theory of, Figure 13
Probability of traffic breakdown $P_{\text{FS}}^{(\text{B})}(T_{\text{ob}}, q_{\text{sum}}, q_{\text{on}})$ for different time intervals for observing traffic flow $T_{\text{ob}} = 30$ (curve 1) and 15 min (curve 2). $q_{\text{determ,FS}}^{(\text{B})} = 2800$ vehicles/h, $N_{\text{d}} = 16$, $\xi_{\text{d}} = 2$

Capacity of Free Flow at Bottlenecks

Definition of Free Flow Capacity

Freeway capacity of free flow at a bottleneck is limited by traffic breakdown at the bottleneck. During a given time interval T_{ob}, traffic breakdown can occur with probability $P_{\text{FS}}^{(\text{B})}$. In other words, an attribute of this *probabilistic freeway capacity at the bottleneck* is the probability $P_{\text{C}}^{(\text{B})}$ (63) that free flow remains at the bottleneck during the time interval T_{ob} [17]. Freeway capacity is reached when

$$P_{\text{C}}^{(\text{B})} < 1 . \tag{70}$$

Thus we define freeway capacity in free flow at a bottleneck as follows [17]:

- Freeway capacity in free flow is equal to the flow rate downstream of the bottleneck at which free flow remains at the bottleneck with the probability $P_{\text{C}}^{(\text{B})} < 1$ (70) during a given time interval T_{ob} for observing traffic flow. This means that at this flow rate with the probability $P_{\text{FS}}^{(\text{B})} > 0$ an $F \to S$ transition (traffic breakdown) occurs at the bottleneck during the time interval T_{ob}.

Because there can be an infinite number of flow rates downstream of the bottleneck for which the condition (70) is satisfied, there can also be an infinite number of freeway capacities in free flow at the bottleneck. Each of the freeway capacities has two attributes:

(1) The probability $P_{\text{C}}^{(\text{B})}$ (70) that free flow remains at the bottleneck during a given time interval T_{ob} for observing traffic flow.
(2) The time interval T_{ob}.

When for free flow at the bottleneck rather than the condition (70) the condition

$$P_{\text{C}}^{(\text{B})} = 1 \tag{71}$$

is satisfied, then the flow rate downstream of the bottleneck in this free flow is lower than any of the freeway capacities.

It must be noted that flow rates are usually complicated time-functions in real traffic. For this reason, in empirical observations measured data is usually averaged for a time interval T_{av}. Thus empirical definition of freeway capacity at a bottleneck is as follows (for more detail, see Sect. 10.3.1 of the book [17] and ► Traffic Congestion, Modeling Approaches to):

- Freeway capacity in free flow is equal to the flow rate downstream of the bottleneck at which free flow remains at the bottleneck with the probability $P_C^{(B)} < 1$ (70) during a given averaging time interval T_{av}.

Infinite Free Flow Capacities and Diagram of Congested Patterns

A diagram of congested patterns determines regions of the emergence and existence of various congested patterns occurring at a bottleneck as functions of flow rates upstream of the bottleneck. In particular, a part of the diagram for synchronized flow patterns (SP), which can occur at an on-ramp bottleneck due to traffic breakdown, in the flow-flow plane whose co-ordinates are q_{in} and q_{on} is presented in Fig. 14 found in a microscopic three-phase traffic theory. Let us show that boundaries on this diagram associated with emergence and dissolution of the SPs determine infinite free flow capacities at the bottleneck [17].

At the boundary $F_S^{(B)}$ of the diagram an $F \to S$ transition occurs during the time T_{ob} with breakdown probability

$$P_{FS}^{(B)} = 1 \,. \tag{72}$$

This means that the boundary $F_S^{(B)}$ determines an infinite number of *maximum* freeway capacities of free flow at the bottleneck $q_{max}^{(free\ B)}$ for different flow rates q_{in} and q_{on} at the given T_{ob}.

There is also a region in the diagram of congested patterns at a bottleneck below and left of the boundary $F_S^{(B)}$

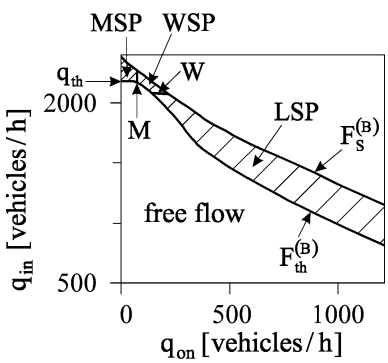

Traffic Breakdown, Probabilistic Theory of, Figure 14
Freeway capacity and metastable region (hatched region) with respect to SP emergence and existence in the diagram of congested patterns at the on-ramp bottleneck found in a microscopic three-phase traffic theory. MSP is a moving SP, WSP is a widening SP, LSP is a localized SP (see explanations of SPs in ► Traffic Congestion, Spatiotemporal Features of). Taken from [17]

within which an $F \to S$ transition occurs during the time T_{ob} with probability $0 < P_{FS}^{(B)} < 1$. The more distant a point within this region in the diagram from the boundary $F_S^{(B)}$, the smaller the probability $P_{FS}^{(B)}$ of traffic breakdown.

For this reason, there is a *threshold* boundary $F_{th}^{(B)}$ in the diagram (Fig. 14), which restricts the above region of the diagram within which $0 < P_{FS}^{(B)} < 1$. Below and left of the threshold boundary $F_{th}^{(B)}$ in the diagram, breakdown probability

$$P_{FS}^{(B)} = 0 \,, \tag{73}$$

i. e., no traffic breakdown is possible during the time T_{ob}. There are an infinite amount of *threshold* flow rates downstream of an on-ramp bottleneck $q_{th}^{(B)} = q_{on} + q_{in}$ for traffic breakdown at the bottleneck associated with different points (q_{on}, q_{in}) in the diagram of congested patterns at the threshold boundary $F_{th}^{(B)}$. As above mentioned, at the flow rate $q_{sum} < q_{th}^{(B)}$ (57) no traffic breakdown is possible, i. e., the flow rate is smaller than any of the freeway capacities in free flow at the bottleneck. This means that the treshold boundary $F_{th}^{(B)}$ determines the infinite number of *minimum* freeway capacities in free flow at the bottleneck $q_{th}^{(B)}$ for different q_{in} and q_{on} at the given T_{ob}.

Different points (q_{on}, q_{in}) for an on-ramp bottleneck between the boundaries $F_S^{(B)}$ and $F_{th}^{(B)}$ in the diagram of congested patterns are related to the infinite amount of the flow rates in free flow downstream of the bottleneck associated with the infinite number of freeway capacities in free flow at the bottleneck. This is because for each of these flow rates the probability $P_C^{(B)}$ (63) satisfies the condition

$$0 < P_C^{(B)} < 1 \,. \tag{74}$$

Thus free flow, which is observed during the time interval T_{ob}, is in a metastable state with respect to traffic breakdown on and between the boundaries $F_S^{(B)}$ and $F_{th}^{(B)}$ in the diagram of congested patterns. This metastability of free flow means that either spontaneous or induced traffic breakdown is possible. In the case of a spontaneous breakdown, a critical cluster (nucleus) for the breakdown can occur spontaneously at the bottleneck leading to traffic breakdown. In the case of an induced breakdown, the role of the cluster plays a congested pattern that has initially occurred at another location as the bottleneck location; due to pattern propagation, the induced traffic breakdown occurs after the pattern reaches the bottleneck (see a more detailed explanation of spontaneous and induced traffic breakdowns in ► Traffic Congestion, Modeling Approaches to).

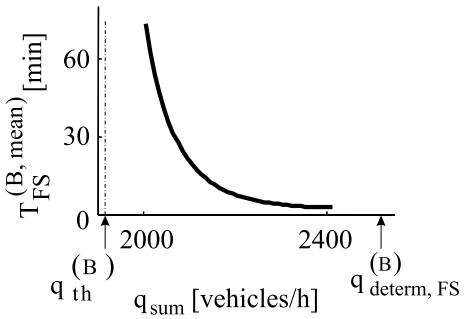

Traffic Breakdown, Probabilistic Theory of, Figure 15
Mean time delay for traffic breakdown $T_{FS}^{(B,mean)}$ (51) as functions of the total flow rate q_{sum} for $q_{on} = 300$ vehicles/h

Probability of Breakdown and Diagram of Congested Patterns

A link between the mean time delay $T_{FS}^{(B,mean)}$ in traffic breakdown, traffic breakdown probability $P_{FS}^{(B)}$, and the diagram of congestion patterns (Fig. 14) discussed above is confirmed by numerical simulations presented in Figs. 15 and 16 [20,21,22]. For the numerical simulations, we used an example of the function $w_{-}(N)$ (48)

$$w_{-}(N) = N\left[\frac{a(q_{on})}{1 + (N/N_0(q_{on}))^4} + b(q_{on})\right], \qquad (75)$$

which exhibits qualitatively the same shape as that shown in Fig. 11. The analytical function (75) allows us to perform a numerical analysis of the mean time delay (51) and the associated nucleation rate (55) for the breakdown phenomenon ($F \to S$ transition). For the analysis of (51) and (55), only branches $N^{(determ)}$ and N_c (Fig. 11) of the function (75) associated with the deterministic and critical clusters within which $N \le N_s$ are relevant. This is because the maximum possible value of $N = N_2$ in the potential barrier $\Delta\Phi$ (52) that determines the nucleation rate (55) is equal to N_s. The nucleation model cannot be applied for $q_{on} = 0$; therefore, in Fig. 16 points in the vicinity of $q_{on} = 0$ show only the tendency of the boundaries and the flow rates for the limiting case of small values q_{on} in which, however, the on-ramp inflow rate $q_{on} > 0$, specifically, it is assumed that the deterministic cluster still exists at the bottleneck.

A numerical study shows that the flow rate dependence of the mean time delay $T_{FS}^{(B,mean)}$ (51) (Fig. 15) exhibits qualitative features found in a microscopic three-phase traffic theory [19,23].

In the diagram of congested patterns at the bottleneck (Fig. 16a), there are boundaries $F_{S,\zeta}^{(B)}$ (curves 1 and 2) each of them is found from the condition that the breakdown occurs during the time T_{ob} with the probability $P_{FS}^{(B)}$ that is equal to a given value ζ:

$$P_{FS}^{(B)}(T_{ob}, q_{sum}, q_{on}) = \zeta, \quad \zeta = \text{const}\,. \qquad (76)$$

At the boundary $F_{S,\zeta}^{(B)}$, the flow rate

$$q_{sum} = q_G^{(B)}(q_{on}, T_{ob}) \qquad (77)$$

depends on q_{on} and T_{ob}. The boundary $F_{S,\zeta}^{(B)}$ at $\zeta \approx 1$ (curve 1 in Fig. 16a) is qualitatively similar with the boundary $F_S^{(B)}$ found in numerical simulation of a three-phase microscopic traffic flow model shown in Fig. 14. In the diagram, there is also the threshold boundary $F_{th}^{(B)}$ (curve $F_{th}^{(B)}$ in Fig. 16a) at which (44) is satisfied. The boundary $F_{determ,S}^{(B)}$ determines the on-ramp inflow dependence of the critical flow rate for the deterministic breakdown.

Boundaries for traffic breakdown can also be calculated for another congested pattern diagram whose co-

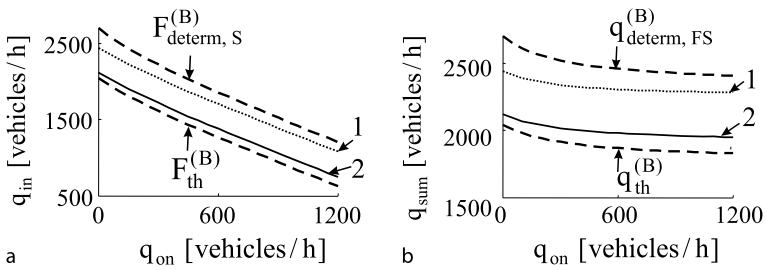

Traffic Breakdown, Probabilistic Theory of, Figure 16
Boundaries of constant values of breakdown probability $P_{FS}^{(B)}$ in diagrams of congested patterns in the flow-flow planes $q_{in}(q_{on})$ (a) and $q_{sum}(q_{on})$ (b). $F_{determ,S}^{(B)}$ and $q_{determ,FS}^{(B)}$ (5) are the critical boundary for deterministic traffic breakdown and the critical flow rate for deterministic traffic breakdown, respectively. $F_{th}^{(B)}$ and $q_{th}^{(B)}$ (44) are the threshold boundary and the threshold flow rate for traffic breakdown, respectively. Curves 1, 2 are related to different chosen values $P_{FS}^{(B)}$ in (76): $\zeta = 0.986$ (curves 1) and $\zeta = 0.22$ (curves 2) at $T_{ob} = 15$ min

ordinates are q_{sum} and q_{on} (Fig. 16b). Curves 1 and 2 in Fig. 16b show on-ramp inflow dependences of the critical flow rate for traffic breakdown associated with two different values ζ of breakdown probability (76). In this diagram, the boundaries $q^{(\text{B})}_{\text{determ,FS}}$ and $q^{(\text{B})}_{\text{th}}$ determine the on-ramp inflow dependences of the critical flow rate for the deterministic breakdown and the threshold flow rate for traffic breakdown, respectively.

Conclusions

Probabilistic theory of traffic breakdown [15,17,20,21,22] discussed in this review article explains fundamental empirical features of traffic breakdown **A-D** (Subsect. "Fundamental Empirical Features of Traffic Breakdown"). The nucleation theory of traffic breakdown allows us to conclude that in accordance with measured traffic data traffic breakdown is a local phase transition from free flow to synchronized flow ($F \to S$ transition). The $F \to S$ transition exhibits features of first-order phase transitions observed in diverse complex systems of natural science: (i) an $F \to S$ transition can be either spontaneous or induced; (ii) emergence and dissolution of synchronized flow is accompanied by a hysteresis effect; (iii) breakdown probability is an increasing function of control parameters (flow rates).

Future Directions

Whereas empirical *macroscopic* features of traffic breakdown have been found (Subsect. "Fundamental Empirical Features of Traffic Breakdown"), there are almost no empirical investigations of *microscopic* empirical characteristics of the breakdown. In particular, behavior of speed fluctuations in real free flow at bottlenecks, appearance and growth of critical fluctuations (critical vehicle clusters) leading to the breakdown have not been understood. In general, empirical and theoretical analyzes of various sources of traffic flow fluctuations and their behavior near traffic breakdown should be a very important direction in the field of traffic flow dynamics.

Although some *spatiotemporal* features of nuclei for traffic breakdown and synchronized flow are incorporated into the KKW CA traffic flow model (Sect. "Probabilistic Description of Traffic Breakdown with Cellular Automata (CA) Traffic Flow Model"), this is not made for the analytical study of traffic breakdown made in Sect. "Probabilistic Description of Traffic Breakdown Based on Master Equation". In the latter case, to use the well-known probabilistic theory of first-order phase transitions based on a master equation and one step processes [7], possible various spatial shapes of nuclei have been ignored and suggested that the cluster dynamics can be described by the dynamic behavior of the vehicle number within the cluster only. Therefore, we believe that there is a great potential for further analytical probabilistic theories of traffic breakdown in which spatiotemporal character of vehicle clusters at bottlenecks is taken into account.

Acknowledgments

We would like to thank Andreas Hiller, Hubert Rehborn, Mario Aleksić, Ines Maiwald-Hiller and Olivia Brickley for help and useful suggestions.

Bibliography

1. Barlović R, Santen L, Schadschneider A, Schreckenberg M (1998) Metastable states in cellular automata for traffic flow. Eur Phys J B 5:793–800
2. Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. Phys Rep 329:199
3. Cremer M (1979) Der Verkehrsfluss auf Schnellstrassen. Springer, Berlin
4. Daganzo CF (1997) Fundamentals of transportation and traffic operations. Elsevier, New York
5. Elefteriadou L, Roess RP, McShane WR (1995) Probabilistic nature of breakdown at freeway merge junctions. Transp Res Rec 1484:80–89
6. Fukui M, Sugiyama Y, Schreckenberg M, Wolf DE (eds) (2003) Traffic and granular flow' 01. Springer, Heidelberg
7. Gardiner CW (1994) Handbook of stochastic methods for physics, chemistry, and the natural sciences. Springer, Berlin
8. Gartner NH, Messer CJ, Rathi A (eds) (1997) Special report 165:Revised monograph on traffic flow theory. Trans Res Board, Washington DC
9. Haight FA (1963) Mathematical theories of traffic flow. Academic Press, New York
10. Hall FL, Agyemang-Duah K (1991) Freeway capacity drop and the definition of capacity. Trans Res Rec 1320:91–98
11. Hall FL, Hurdle VF, Banks JH (1992) Synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relationships on freeways. Transp Res Rec 1365:12–18
12. Helbing D (2001) Traffic and related self-driven many-particle systems. Rev Mod Phys 73:1067–1141
13. Helbing D, Herrmann HJ, Schreckenberg M, Wolf DE (eds) (2000) Traffic and granular flow' 99. Springer, Heidelberg
14. Kerner BS (1998) Empirical features of self-organization in traffic flow. Phys Rev Lett 81:3797–3400
15. Kerner BS (2000) Theory of breakdown phenomenon at highway bottlenecks. Trans Res Rec 1710:136–144
16. Kerner BS (2002) Empirical macroscopic features of spatial-temporal traffic patterns at highway bottlenecks. Phys Rev E 65:046138
17. Kerner BS (2004) The physics of traffic. Springer, Berlin, New York
18. Kerner BS, Klenov SL (2002) A microscopic model for phase transitions in traffic flow. J Phys A Math Gen 35:L31–L43

19. Kerner BS, Klenov SL (2003) Microscopic theory of spatial-temporal congested traffic patterns at highway bottlenecks. Phys Rev E 68:036130
20. Kerner BS, Klenov SL (2005) Probabilistic breakdown phenomenon at on-ramps bottlenecks in three-phase traffic theory. arXiv:cond-mat/0502281
21. Kerner BS, Klenov SL (2006) Probabilistic breakdown phenomenon at on-ramp bottlenecks in three-phase traffic theory: Congestion nucleation in spatially non-homogeneous traffic. Physica A 364:473–492
22. Kerner BS, Klenov SL (2006) Probabilistic breakdown phenomenon at on-ramp bottlenecks in three-phase traffic theory. Transp Res Rec 1965:70–78
23. Kerner BS, Klenov SL, Wolf DE (2002) Cellular automata approach to three-phase traffic theory. J Phys A Math Gen 35:9971–10013
24. Kühne R, Mahnke R, Lubashevsky I, Kaupužs J (2002) Probabilistic description of traffic breakdown. Phys Rev E 65:066125
25. Kühne R, Mahnke R, Lubashevsky I, Kaupužs J (2004) Probabilistic description of traffic breakdown caused by on-ramp. arXiv:cond-mat/0405163
26. Leutzbach W (1988) Introduction to the theory of traffic flow. Springer, Berlin
27. Lorenz M, Elefteriadou L (2000) A probabilistic approach to defining freeway capacity and breakdown. Trans Res Cir E-C018:84–95
28. Mahnke R, Kaupužs J (1999) Stochastic theory of freeway traffic. Phys Rev E 59:117–125
29. Mahnke R, Pieret N (1997) Stochastic master-equation approach to aggregation in freeway traffic. Phys Rev E 56: 2666–2671
30. Mahnke R, Kaupužs J, Lubashevsky I (2005) Probabilistic description of traffic flow. Phys Rep 408:1–130
31. May AD (1990) Traffic flow fundamentals. Prentice-Hall, New Jersey
32. Nagatani T (2002) The physics of traffic jams. Rep Prog Phys 65:1331–1386
33. Nagel K, Schreckenberg M (1992) A cellular automaton model for freeway traffic. J Phys France I 2:2221–2229
34. Nagel K, Wagner P, Woesler R (2003) Still flowing: Approaches to traffic flow and traffic jam modeling. Oper Res 51:681–710
35. Persaud BN, Yagar S, Brownlee R (1998) Exploration of the breakdown phenomenon in freeway traffic. Trans Res Rec 1634:64–69
36. Whitham GB (1974) Linear and nonlinear waves. Wiley, New York
37. Wiedemann R (1974) Simulation des Verkehrsflusses. University of Karlsruhe, Karlsruhe
38. Wolf DE (1999) Cellular automata for traffic simulations. Physica A 263:438–451

Traffic Congestion, Modeling Approaches to

Boris S. Kerner
GR/ETI, HPC: G021, Daimler AG,
Sindelfingen, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Free and Congested Traffic
Three Traffic Phases
Characteristic Parameters of Wide Moving Jam Propagation and Line *J*
Traffic Breakdown and Highway Capacity
Moving Jam Emergence in Synchronized Flow (S → J Transition)
Empirical Double *Z*-Characteristic for Phase Transitions in Traffic Flow
Hypotheses of Three-Phase Traffic Theory as the Result of Traffic Phase Definitions
Critical Discussion of Fundamental Diagram Modeling Approach to Traffic Congestion
Three-Phase Traffic Flow Models
Link Between Three-Phase Traffic Theory and Fundamental Diagram Approach
Conclusions
Future Directions
Cross References
Acknowledgments
Bibliography

Glossary

Free flow Free flow is usually observed, when the vehicle density in traffic is small enough. The flow rate increases in free flow with increase in vehicle density, whereas the average vehicle speed is a decreasing density function.

Traffic breakdown In empirical observations, when density in free flow increases and becomes great enough, the phenomenon of the onset of congestion is observed in this free flow: The average speed decreases sharply to a lower speed in congested traffic. This speed breakdown observed during the onset of congestion is called the breakdown phenomenon or traffic breakdown.

Congested traffic Congested traffic is defined as a state of traffic in which the average speed is *lower* than the minimum average speed that is possible in free flow.

Bottleneck Traffic breakdown occurs mostly at highway bottlenecks. Just as defects and impurities are important for phase transitions in complex spatially distributed systems of various nature, so are freeway bottlenecks in freeway traffic. The bottleneck can be a result of roadworks, on- and off-ramps, a decrease in the number of freeway lanes, road curves and road gradients, etc.

Moving jams A moving jam is a localized structure of great vehicle density spatially limited by two jam fronts. Within the downstream jam front vehicles accelerate escaping from the jam; within the upstream jam front, vehicles slow down approaching the jam. The jam as a whole structure propagates upstream in traffic flow. Within the jam (between the jam fronts) vehicle density is great and speed is very low (sometimes as low as zero). A sequence of moving jams is often called "stop-and-go" traffic.

Traffic flow model A traffic flow model is devoted to the explanation and simulation of traffic flow phenomena, which are observed in measured data of real traffic flow, and to the prediction of new traffic flow phenomena that could be found in real traffic flow. First of all, a traffic flow model should explain and predict empirical (measured) features of traffic breakdown.

Fundamental diagram of traffic flow The fundamental diagram is a relationship between the flow rate and density in vehicle traffic. Because the flow rate is the product of the average speed and density, the fundamental diagram is associated with a relationship between these traffic variables. In accordance with an obvious result of traffic measurements, on average the speed decreases when the density increases. Thus in the flow-density plane, the fundamental diagram should pass through the origin (when the density is zero so is the flow rate) and should have at least one maximum. The fundamental diagram gives also a connection between the average space gap (net distance) between vehicles and the average speed.

Steady states of traffic flow Steady states of traffic flow are hypothetical states of homogeneous (in time and space) traffic flow of identical vehicles (and identical drivers) in which all vehicles move with the same time-independent speed and have the same space gaps.

Fundamental diagram approach to traffic flow theory and modeling The fundamental hypothesis of the fundamental diagram approach to traffic flow theory and modeling suggests that steady states of both free flow and congested traffic lie on a one-dimension curve(s) (i. e., on a theoretical fundamental diagram of traffic flow) in the flow-density plane. At each given time-independent speed of the preceding vehicle, the theoretical fundamental diagram determines a single desired space gap at which a vehicle moves with this time-independent speed while following the preceding vehicle. This model vehicle behavior is related to a steady state of traffic flow associated with a hypothetical noiseless and acceleration less (deceleration less) model limit.

Three-phase traffic theory In the author's three-phase traffic theory, besides the free flow phase there are two phases in congested traffic, the synchronized flow and wide moving jam phases. The synchronized flow and wide moving jam phases in congested traffic are defined through empirical spatiotemporal criteria (definitions) [S] and [J]. In contrast with the fundamental diagram approach, the fundamental hypothesis of three-phase traffic theory suggests that in steady states of *synchronized flow*, at each given time-independent speed of the preceding vehicle, there is an infinite number of space gaps at which the vehicle can move with this speed, while following the preceding vehicle. Thus hypothetical steady states of synchronized flow cover a two-dimensional (2D) region in the flow-density plane, i. e., there is no desired space gap in hypothetical steady states of synchronized flow in this theory.

Wide moving jam traffic phase In three-phase traffic theory, the following definition of the wide moving jam phase [J] in congested traffic based on measured traffic data is made. A wide moving jam is a moving jam that maintains the mean velocity of the downstream jam front, even when the jam propagates through any other traffic states or bottlenecks. This is the characteristic feature of the wide moving jam phase.

Synchronized flow traffic phase In three-phase traffic theory, the following definition of the synchronized flow phase [S] in congested traffic based on measured traffic data is made. The downstream front of synchronized flow does not exhibit the above mentioned characteristic feature of wide moving jams; specifically, the latter front is often fixed at a bottleneck.

Microscopic criterion for traffic phases Within wide moving jams, there are regions in which traffic flow is interrupted: the inflow into the jam has no influence on the jam outflow. A sufficient condition for flow interruption determines the microscopic criterion for the wide moving jam phase. Based on this criterion both the synchronized flow and wide moving jam phases can be identified in congested traffic from single vehicle data measured even at one freeway location. This is because if in measured data congested traffic states associated with the wide moving jam phase have been identified, then with certainty all remaining congested states in the data set are related to the synchronized flow phase.

F → S transition In all known observations, traffic breakdown is a phase transition from the free flow phase to synchronized flow phase (F → S transition). Thus the

terms *traffic breakdown* and an $F \rightarrow S$ *transition* are synonyms related to the same phenomenon of the onset of congestion in free flow.

Highway capacity of free flow at bottleneck Highway (freeway) capacity of free flow at a bottleneck (called also bottleneck capacity) is limited by traffic breakdown, i. e., $F \rightarrow S$ transition at the bottleneck, which occurs with probability (denoted by $P_{FS}^{(B)}$) during a given averaging time interval (denoted by T_{av}) for traffic variables. Highway capacity in free flow is equal to the flow rate downstream of the bottleneck at which free flow remains at the bottleneck during the time interval T_{av} with probability $P_C^{(B)} = 1 - P_{FS}^{(B)}$, which is less than one; thus highway capacity has two attributes $P_C^{(B)}$ and T_{av}. At each given T_{av}, there are infinite highway capacities in a limited range associated with different probabilities $P_C^{(B)}$.

Spontaneous traffic breakdown (spontaneous $F \rightarrow S$ transition) at bottleneck If before traffic breakdown free flow has been at a bottleneck as well as upstream and downstream in a neighborhood of the bottleneck, the breakdown is caused by occurrence and subsequent growth of speed disturbances (fluctuations) within the free flow at the bottleneck. Such traffic breakdown is called a *spontaneous* traffic breakdown at the bottleneck. A speed disturbance begins to grow, i. e., the speed within the disturbance begins to decrease over time with the subsequent breakdown at the bottleneck, if within the disturbance the initial speed is equal to or lower than a critical speed for an $F \rightarrow S$ transition. Such a critical speed disturbance can be considered a nucleus for the breakdown at the bottleneck. There can be various sources for speed disturbances whose occurrence and subsequent growth lead to spontaneous traffic breakdown: unexpected vehicle braking and/or lane changing in free flow, fluctuations in flow rates upstream of the bottleneck, vehicle merging from other roads in the bottleneck neighborhood (e. g., at on-ramp bottlenecks), etc.

Induced traffic breakdown (induced $F \rightarrow S$ transition) at Bottleneck In contrast with spontaneous traffic breakdown, which occurs when before the breakdown free flow is in a neighborhood of the bottleneck, an induced traffic breakdown at the bottleneck is caused by the propagation of a moving spatiotemporal *congested* traffic pattern, which has initially occurred at a *different* road location (e. g., at another bottleneck). When this congested pattern reaches the bottleneck, the pattern induces traffic breakdown at the bottleneck. The congested pattern whose propagation causes the breakdown at the bottleneck can be considered an external speed disturbance at the bottleneck.

Spontaneous wide moving jam emergence in synchronized flow Within synchronized flow, there can be speed disturbances (fluctuations) whose growth leads to wide moving jam emergence. Such wide moving jam emergence is called *spontaneous* wide moving jam emergence in synchronized flow (spontaneous $S \rightarrow J$ transition). A speed disturbance begins to grow, if within the disturbance the initial speed is equal to or lower than a critical speed for an $S \rightarrow J$ transition. Such a critical speed disturbance can be considered a nucleus for the $S \rightarrow J$ transition. The growing speed disturbance in synchronized flow propagates upstream; for this reason, in contrast with spontaneous traffic breakdown at a bottleneck, the $S \rightarrow J$ transition occurs usually upstream of the road location at which the critical speed disturbance has initially appeared. There can be various sources for speed disturbances whose occurrence and subsequent growth lead to spontaneous $S \rightarrow J$ transition: unexpected vehicle braking and/or lane changing within synchronized flow, fluctuations in upstream flow rates, vehicle merging from other roads, etc.

Definition of the Subject

Understanding of vehicular traffic congestion is the key for effective traffic management, traffic control, organization, and other engineering applications, which should improve traffic safety and result in high-quality mobility. In empirical observations, traffic congestion occurs as the result of a so-called traffic breakdown: the vehicle speed decreases sharply and vehicle density increases in an initially free traffic flow. The subsequent development of congested traffic exhibits a very complex spatiotemporal behavior. To explain traffic breakdown and resulting traffic congested patterns a huge number of traffic theories and models have been developed. It is clear that if a traffic flow model cannot show empirical features of traffic breakdown, the model cannot also show and predict many other traffic phenomena observed in real congested traffic. Thus an assessment of modeling approaches to show and predict traffic breakdown in vehicle traffic is the starting point for a further consideration of the reliability of these modeling approaches to show and predict traffic flow phenomena observed in real congested traffic. Furthermore, the development of modeling approaches to traffic congestion, which are consistent with empirical results, is of great importance for development of traffic control meth-

ods as well as management strategies, which should either prevent traffic breakdown or lead to dissolution of existing congested patterns.

Modeling approaches for the explanation of traffic breakdown include traffic flow models based on a diverse variety of methods of complexity and systems science. This is explained by an extremely complex non-linear character of real vehicular traffic. Moreover, there are links between spatiotemporal phenomena associated with vehicular traffic congestion and traffic phenomena observed in many other complex systems like pedestrian traffic, which is responsible for the complexity of crowd and evacuation dynamics, and air traffic. For this reason, the understanding of vehicular traffic congestion is also very important for the further progress in understanding of many other many-particle complex systems.

Introduction

Earlier traffic flow models, which claim to explain and predict the onset of traffic congestion, are reviewed in [16,18,21,33,35,37,40,84,92,93,95,98,102,126,127,128]. All these models are associated with the fundamental diagram approach to traffic flow modeling.

In 1955 Lighthill and Whitham [86] wrote in their classic work (see p. 319 in [86]): "…The fundamental hypothesis of the theory is that at any point of the road the flow (vehicles per hour) is a function of the concentration (vehicles per mile)…" (Fig. 1a). This hypothesis is also the fundamental one for all traffic flow theories and models reviewed in [7,16,18,21,33,35,37,40,84,92,93,95,98,102,126,127,128]. The fundamental diagram for a traffic flow model means that in all these traffic flow models it is suggested that for a given time-independent speed of the preceding vehicle there is a single desired (or optimal) space gap for the following vehicle that the vehicle chooses in a hypothetical acceleration less (deceleration less) and noiseless model limit case.

The models within the framework of the fundamental diagram can be classifed into two main classes. The first model class refers to the classic LWR (Lighthill–Whitham–Richards) model introduced in 1955–1956 [86,109] (see also references in [16,21,33,37,40,84,93,126]). Examples for this model class are the cell-transmission models of Daganzo [20]. The basic idea of the LWR-model is that maximum flow rate (denoted q_0 in Fig. 1a), associated with the maximum point at the fundamental diagram, determines capacity of free flow at a bottleneck, i. e., the bottleneck capacity. Thus if the flow rate upstream of the bottleneck exceeds the bottleneck capacity, then traffic breakdown should occur.

The second model class refers to the classic GM (General Motors) model of Herman, Gazis, Rothery, Montrool and Potts introduced in 1959–1961 [34,43]. The basic idea of the GM-model approach is as follows. Beginning at a critical density, there is instability of steady model states at the fundamental diagram. This model instability that should explain traffic breakdown is associated with a finite value of driver reaction time. The instability can qualitatively be explained, if we suggest that in a homogeneous traffic flow a vehicle decelerates unexpectedly, which causes a local speed decrease (speed disturbance) in the flow. Due to the driver reaction time, the following vehicle can decelerate stronger than it would be necessary to avoid collision. This effect is called the over-deceleration effect (or human over-reaction). As a result of over-deceleration, the speed of this vehicle can become lower than the speed of the preceding vehicle. The same over-deceleration effect can occur for many other following vehicles that can lead to the growth of the initial speed disturbance, i. e., to traffic flow instability. Examples for the GM model class are the optimal velocity (OV) models by Newell, by Whitham, and by Bando et al., the Payne macroscopic model and its variants, the Wiedemann psychophysical traffic flow model and its variants, the Gipps-model, the NaSch (Nagel–Schreckenberg) cellular automata (CA) model and its variants, the Krauß-model, the IDM-model (Intelligent Driver Model) of Treiber et al. as well as a huge number of other traffic flow models (see below and references in [16,18,33,37,40,84,92,93,95,98,102,126,127,128]).

Naturally, traffic flow models should explain empirical spatiotemporal features of traffic congestion. However, it is only recently that spatiotemporal features of congested traffic patterns have been adequately understood [52]. Consequently, these earlier traffic flow theories and models cannot explain and predict many of these empirical spatiotemporal traffic pattern features. When these traffic flow models are used for simulations of freeway control and management strategies, the related simulations cannot predict many of the freeway traffic phenomena that would occur through the use of a simulated control strategy.

To understand this critical conclusion, let us consider the empirical basis of a theoretical fundamental diagram. In all observations of traffic flow the greater the vehicle density, the lower the average vehicle speed (Fig. 1b). This should prove the theoretical fundamental diagram as the basis for traffic flow theories and models. However, this obvious and correct empirical result is associated with the *averaging* of measured traffic data. In particular, if traffic data is associated with measurements of the speed and flow rate at a freeway location(s) during some long enough av-

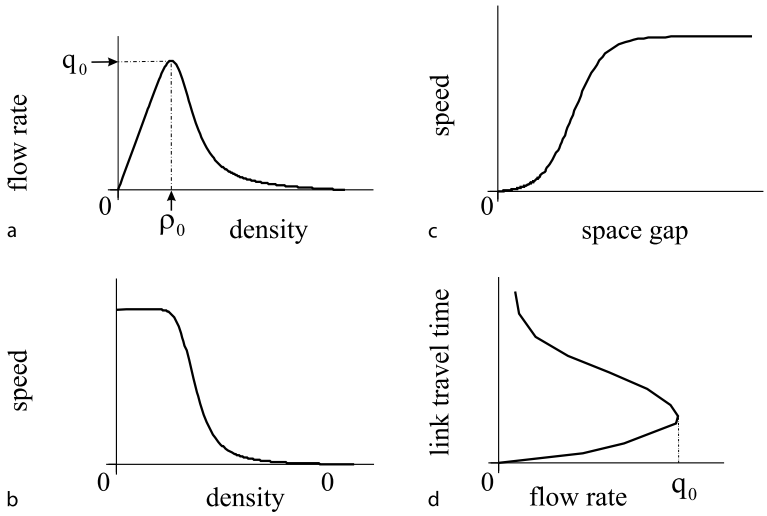

Traffic Congestion, Modeling Approaches to, Figure 1
Qualitative example of fundamental diagram: a Flow-density relationship (fundamental diagram). b,c The speed-density (b), speed-space-gap (c) and link-travel-time-flow (d) relationships associated with a

eraging time interval, then to find a point at the fundamental diagram related to a chosen density, all measured data for different flow rates and speeds associated with densities within a small density range around the chosen density should be averaged to one value of the flow rate (Fig. 1a) and associated average speed (Fig. 1b). By this data averaging almost all information about the *spatiotemporal* structure of traffic patterns propagating through the freeway location(s) at which these measurements have been made is *lost*. The remaining information that determines a fundamental diagram shape is associated with averaging characteristics of traffic patterns, which are most frequently observed during the averaging time interval at the related freeway location(s).

Because of the failure of earlier traffic flow theories and models in explanation of traffic congestion, in 1996–1999 the author introduced a three-phase traffic theory [52], which explains empirical features of traffic breakdown and resulting congested patterns.

In this article, we will try to explain this dramatic development of traffic flow theories to traffic congestion. The article is organized as follows. In Sects. "Free and Congested Traffic"–"Empirical Double *Z*-Characteristic for Phase Transitions in Traffic Flow", we consider empirical features of traffic breakdown and moving jam emergence in real freeway traffic; here we give also a brief explanation of these empirical traffic flow features based on three-phase traffic theory. In Sect. "Hypotheses of Three-Phase Traffic Theory as The Result of Traffic Phase Definitions", we show that hypotheses of three-phase traffic theory, which explain the empirical features of traffic congestion, are the result of the traffic phase definitions made in this theory. In Sect. "Critical Discussion of Fundamental Diagram Modeling Approach to Traffic Congestion", based on a consideration of traffic flow theories and models in the context of the fundamental diagram approach we explain that and why these models cannot show features of real traffic breakdown. We discuss also achievements of the fundamental diagram approach in explanation of empirical wide moving jam propagation and characteristic parameters of wide moving jams. In Sect. "Three-Phase Traffic Flow Models", we briefly discuss a three-phase traffic flow model, which has been used to derive simulation results explaining empirical features of traffic congestion in Sects. "Free and Congested Traffic"–"Empirical Double *Z*-Characteristic for Phase Transitions in Traffic Flow". A link between three-phase traffic theory and the fundamental diagram approach is discussed in Sect. "Link Between Three-Phase Traffic Theory and Fundamental Diagram Approach".

Free and Congested Traffic

At small enough vehicle density, interactions between vehicles in free flow are negligible. Therefore, vehicles have an opportunity move with their desired maximum speeds (if this speed is not restricted by road conditions or traffic regulations). When the density increases, vehicle interaction cannot be neglected any more. As a result of vehicle interaction in free flow, the average speed decreases with

increase in density (measured data left of the dashed line *FC* in Fig. 2a).

This common vehicle behavior in free flow leads to the *fundamental diagram for free flow*, i. e., a certain curve with a positive slope between the flow rate and density associated with averaging of measured data shown left of the dashed line *FC* in Fig. 2a to one average flow rate for each density (curve *F* in Fig. 2b) (e. g., [16,18,21,40,42,44,89,93,101,108,126,127]).

The fundamental diagram for free flow is cut off at some empirical limit (maximum) point of free flow (denoted $(\rho_{\max}^{(\text{free,emp})}, q_{\max}^{(\text{free,emp})})$ in Fig. 2b) (e. g., [39,93]). At this limit point of free flow, the average vehicle speed has a minimum possible value for free flow. Thus empirical points of free flow as well as the associated fundamental diagram lie to the left of the dashed line *FC* in the flow-density plane whose slope equals this minimum possible speed in free flow $v_{\min}^{(\text{free,emp})}$.

When the vehicle density is great enough, traffic is usually in a congested state (e. g., [16,18,21,40,42,44,89,93,101,108,126,127]). We define traffic phenomena in terms of their empirical features found in measured data rather than their genesis. Often, congested traffic is defined as a traffic state, which merely results from a freeway bottleneck as the inevitable consequence of an upstream flow that exceeds the bottleneck capacity. In this definition it seems obvious to suggest that congested flow is not self-organized, but rigidly controlled by external constraints. However, it has been shown that a variety self-organized processes are responsible for features of real traffic congestion [52]. For these reasons, we use the following *definition* of congested traffic.

Congested traffic states will be defined as complementary to states of free flow (e. g., [76]). It has already been noted that empirical points related to free flow can be approximately presented by a curve with a positive slope in the flow-density plane (Fig. 2b). Congested traffic will be defined, therefore, as a state of traffic where the average vehicle speed is *lower* than the minimum possible speed in free flow, which is related to the limit point $(\rho_{\max}^{(\text{free, emp})}, q_{\max}^{(\text{free, emp})})$ in free flow (Fig. 2b). Thus empirical points of congested traffic lie to the right of the dashed line *FC* in the flow-density plane whose slope equals the minimum possible speed in free flow.

In this definition of congested traffic, nothing is said about the origin of the congestion. It is related to empirical facts that a congested state can occur spontaneously also away from bottlenecks and that many diverse self-organizing effects are responsible for traffic congestion [52], some of them will be considered in the article.

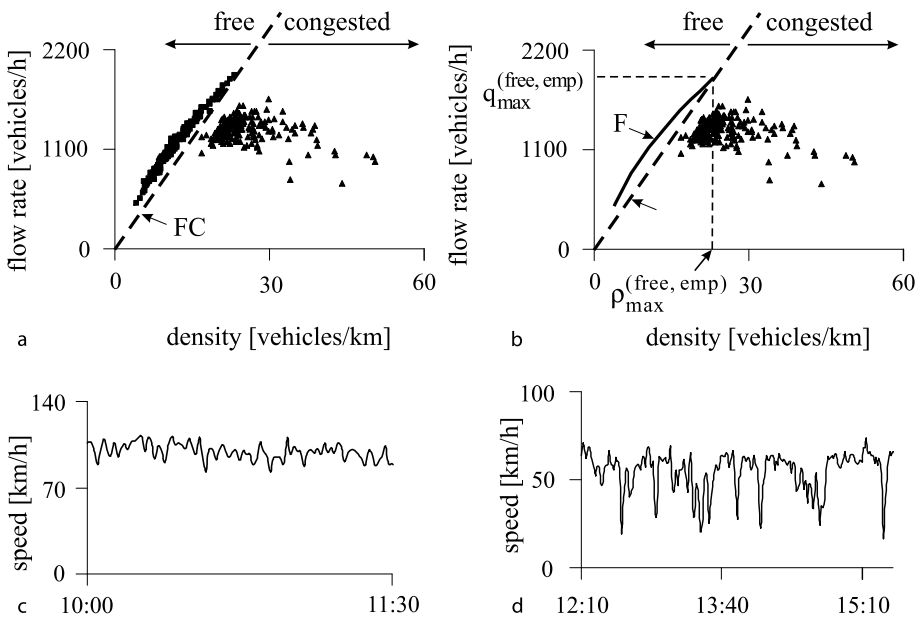

Traffic Congestion, Modeling Approaches to, Figure 2

Free flow and congested traffic (e. g., [76]). a Data for free flow (*points left of the dashed line FC*) and for congested traffic (*points right of the dashed line FC*) measured at a road location. b The fundamental diagram for free flow (*curve F*) and the same measured data for congested traffic as those in a. c,d Vehicle speed in free flow (c) and congested traffic (d), related to *points left* and *right of the line FC* in a, respectively. 1-min average data

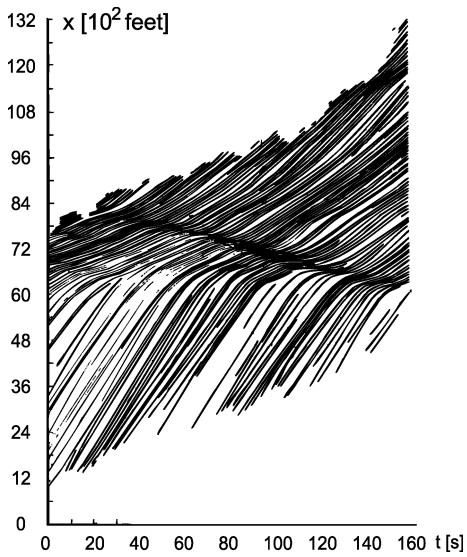

Traffic Congestion, Modeling Approaches to, Figure 3
An empirical moving jam: traffic dynamics derived from aerial photography. Taken from Treiterer [121]

In congested traffic, the "stop-and-go" phenomenon is observed, i. e., a sequence of different moving traffic jams can appear [26,27,28,29,75,76,121,122,123]. Moving jams have been studied empirically by many authors, in particular, in classic empirical works by Edie et al. [26,27,28,29], Treiterer et al. [121,122,123] (Fig. 3) and Koshi et al. [75,76].

Three Traffic Phases

A traffic phase is a traffic state considered in space and time that possesses specific *empirical spatiotemporal* features. These features are specific (unique) only to this traffic phase. Note that a traffic state is characterized by a certain set of statistical properties of traffic variables. Examples of traffic variables are the flow rate q (vehicles/h), vehicle speed v (km/h), vehicle space gap (vehicle space gap is also called as net distance or space headway) g (m), time headway (s) that is also called time space or time gap between vehicles, and vehicle density ρ (vehicles/km).

Based on investigations of congested spatiotemporal patterns measured on different freeways over many days and years (see references in [52]), the three-phase traffic theory introduced by the author suggests that besides the free flow traffic phase there are two other traffic phases in congested traffic: synchronized flow and wide moving jam.

Traffic flow consists of free flow and congested traffic. Congested traffic consists of the synchronized flow phase and the wide moving jam phase. Thus, there are three traffic phases in three-phase traffic theory:

- Free flow
- Synchronized flow
- Wide moving jam

Empirical Macroscopic Criteria for Defining Phases in Congested Traffic

The synchronized flow and wide moving jam phases in congested traffic are defined through the use of the following *empirical (objective) criteria for the traffic phases in congested traffic*. These empirical criteria are related to some characteristic spatiotemporal empirical features of the traffic phases [52]:

[J] The wide moving jam traffic phase is defined as follows. A wide moving jam is a moving jam that maintains the mean velocity of the downstream *front* of the jam, v_g, as the jam propagates. Vehicles accelerate within the downstream jam front from low speed states (sometimes as low as zero) inside the jam to higher speeds downstream of the jam. On average a wide moving jam maintains the mean velocity of the downstream jam front, even as it propagates through other (possibly very complex) traffic states or freeway bottlenecks. This is a characteristic feature of the wide moving jam traffic phase.

[S] The synchronized flow traffic phase is defined as follows. In contrast to the wide moving jam traffic phase, the downstream front of the synchronized flow traffic phase does *not* maintain the mean velocity of the downstream front. In particular, the downstream front of synchronized flow is often *fixed* at a freeway bottleneck. In other words, the synchronized flow traffic phase does not show aforementioned characteristic feature of the wide moving jam traffic phase.

The downstream front of synchronized flow separates synchronized flow upstream from free flow downstream. Within the downstream front of synchronized flow vehicles accelerate from lower speeds in synchronized flow upstream of the front to higher speeds in free flow downstream of the front.

Thus the definitions of the traffic phases in congested traffic are made via the spatiotemporal empirical criteria [J] and [S]. The definitions [J] and [S] are associated with dynamic behavior of the downstream front of these phases, while a wide moving jam or synchronized flow propagates through other traffic states or bottlenecks.

The definitions [J] and [S] of the traffic phases in congested traffic mean that if a congested traffic state is not related to the wide moving jam phase, then with certainty the state is associated with the synchronized flow phase. This is because congested traffic can be either within the synchronized flow phase or within the wide moving jam phase. In other words, if in measured data congested traffic states associated with the wide moving jam phase have been identified, then with certainty all remaining congested states in the data set are related to the synchronized flow phase. In particular, it must be stressed that a moving jam can be associated either with a wide moving jam or with synchronized flow. In accordance with the phase definitions [J] and [S], this depends on the dynamic behavior of the downstream jam front, while the jam propagates through other traffic states or bottlenecks.

One example of application of the objective criteria (definitions) for the traffic phases in congested traffic [J] and [S] is shown in Fig. 4. In Fig. 4, a moving jam propagating through the location of bottleneck B_2 *induces* the synchronized flow phase at this bottleneck (this induced phase transition is labeled "induced F → S transition" in Fig. 4a). Indeed, in accordance with the definition [S], the downstream front of synchronized flow is fixed at the bottleneck B_2. In contrast, the moving jam propagates through this bottleneck with the mean velocity of the downstream jam front remaining unchanged. Thus in accordance with the definition [J], this moving jam is associated with the wide moving jam phase. Synchronized flow is self-sustaining for a very long time (more than an hour) upstream of bottleneck B_2. The possibility of the induced F → S transition means that an F → S transition exhibits nucleation character that is a characteristic feature of first-order phase transitions.

Another empirical example is shown in Fig. 5. It can be seen that a sequence of two moving jams propagates through different states of traffic flow and through a bottleneck while maintaining the downstream jam front velocity. Therefore, these moving jams belong to the wide moving jam phase. In contrast, there is a congested traffic in which speed is much lower than in free flow (compare vehicle speeds in Fig. 5c, d). The downstream front of this congested traffic flow, where vehicles accelerate to free flow, is fixed at the bottleneck (dashed line in Fig. 5a). Therefore, this congested traffic belongs to the synchronized flow phase.

We can see that whereas in wide moving jams both the speed and flow rate are very low (sometimes as low as zero), in synchronized flow the flow rate is high (compare the flow rates within the wide moving jams and within synchronized flow in Fig. 5d). The vehicle speed in synchronized flow is considerably lower than in free flow. However, the flow rates in the free flow and synchronized flow phases can be close to one another (Fig. 5c, d).

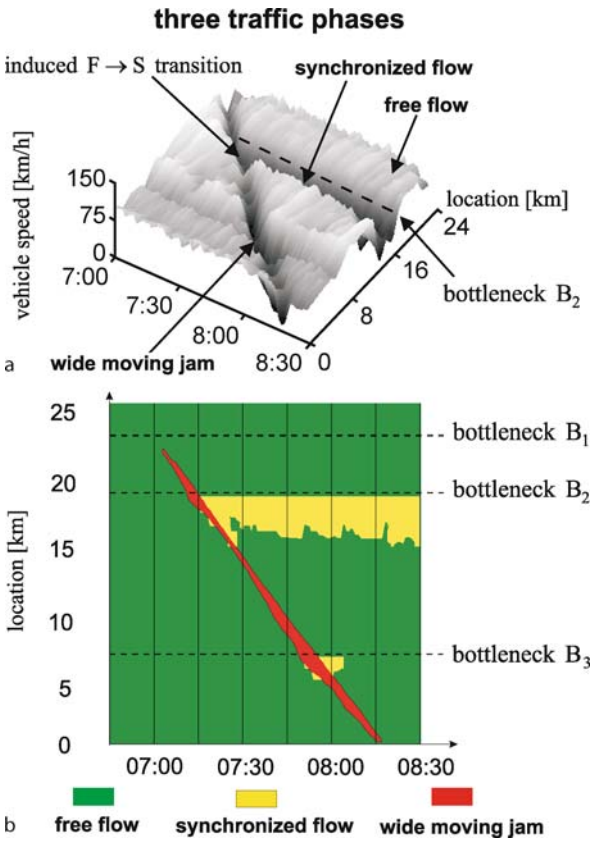

Traffic Congestion, Modeling Approaches to, Figure 4
Measured three traffic phases. a Vehicle speed in space and time. b A graph of a with the free flow phase (*green*), the synchronized flow phase (*yellow*), and the wide moving jam phase (*red*). B_1, B_2, and B_3 are locations of bottlenecks. Taken from [52]

Microscopic Criterion for Traffic Phases in Congested Traffic

The spatiotemporal criteria for a wide moving jam [J], which distinguish the wide moving jam and synchronized flow phases, can be explained by a traffic flow discontinuity within a wide moving jam. Traffic flow is interrupted by the wide moving jam: there is no influence of the inflow into the jam on the jam outflow [52]. A difference between the jam inflow and the jam outflow changes the jam width only. This *traffic flow interruption effect* is a general effect for each wide moving jam.

The jam outflow becomes independent of the jam inflow, when the traffic flow interruption effect within

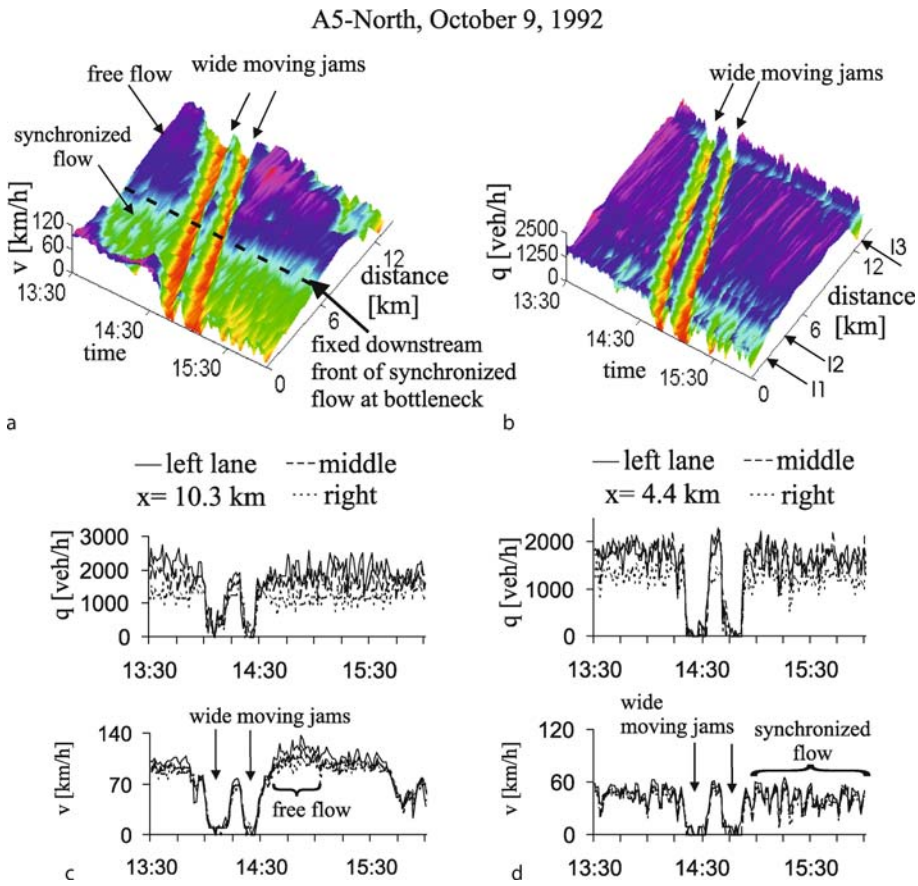

Traffic Congestion, Modeling Approaches to, Figure 5

Measured three traffic phases. a,b Vehicle speed (a) and flow rate averaged across all freeway lanes (b) as functions of time and space. c,d Flow rate and average vehicle speed at two different freeway locations c and d in different freeway lanes. In b, three intersections with other freeways are labeled by I1–I3 that show approximate locations of possible freeway bottlenecks. Taken from [52]

a moving jam occurs. In a hypothetical case, when all vehicles within a moving jam do not move, the criterion for the traffic flow interruption effect within the jam is

$$I = \frac{\tau_J}{\tau_{del}^{(a)}} \gg 1 , \qquad (1)$$

where τ_J is the jam duration, i. e., the time interval between the upstream and downstream jam fronts passing a road location and $\tau_{del}^{(a)}$ is the mean time delay in vehicle acceleration at the downstream jam front from a vehicle standstill; $\tau_{del}^{(a)}$ determines the jam outflow; I is approximately equal to the vehicle number stopped within the jam. Note that corresponding to empirical results $\tau_{del}^{(a)} \approx 1.5 - 2$ sec [52].

In real traffic, vehicles come to a stop at very different blank spaces to each other when they meet a wide moving jam. We call these blank spaces as *blanks* within the jam. Later the vehicles begin to cover these blanks within the jam. This vehicle motion within the jam occurs at low vehicle speeds and it creates new blanks upstream. Thus low speed states that cover blanks within the jam are associated with upstream moving of the blanks within the jam (moving blanks for short; for more details see [67]). A *sufficient* criterion for flow interruption within the jam is

$$I_s = \frac{\tau_{max}}{\tau_{del}^{(a)}} \gg 1 , \qquad (2)$$

where τ_{max} is the maximum time headway between two vehicles within the jam ($\tau_{max} \leq \tau_J$).

For this and the following consideration, we should define the term *time headway* between vehicles. At a given time instant $t = t_1$, the time headway (time gap) $\tau(t_1)$ is *defined* as a time it takes for a vehicle to reach a freeway location at which the bumper of the preceding vehicle is at the time instant t_1. In single vehile measurements

shown in Fig. 6, t_1 is the time at which the preceding vehicle leaves the detector whose location is therefore related to the location of the bumper of the preceding vehicle in the time headway definition; the time headway is equal to $\tau(t_1) = t_2 - t_1$, where t_2 is the time at which the vehicle front has been recorded at the detector.

Under condition (2), there are at least several vehicles within the jam that are in a standstill or if they are still moving, it is only with a negligible low speed in comparison with the speed in the jam inflow and outflow. These vehicles separate vehicles accelerating at the downstream jam front from vehicles decelerating at the upstream jam front. For this reason, the jam outflow does not depend on the jam inflow. In other words, this jam propagates through a bottleneck while maintaining the mean downstream jam front velocity. In accordance with the macroscopic objective spatiotemporal criteria [J] and [S], this jam is a wide moving jam. Thus, we can assume that the traffic flow interruption effect can be used as a criterion to distinguish the synchronized flow and wide moving jam phases in single vehicle data. This is possible even if data is measured at a single freeway location.

The interruption of traffic flow within a moving jam can also be found in *empirical* single vehicle data. In an

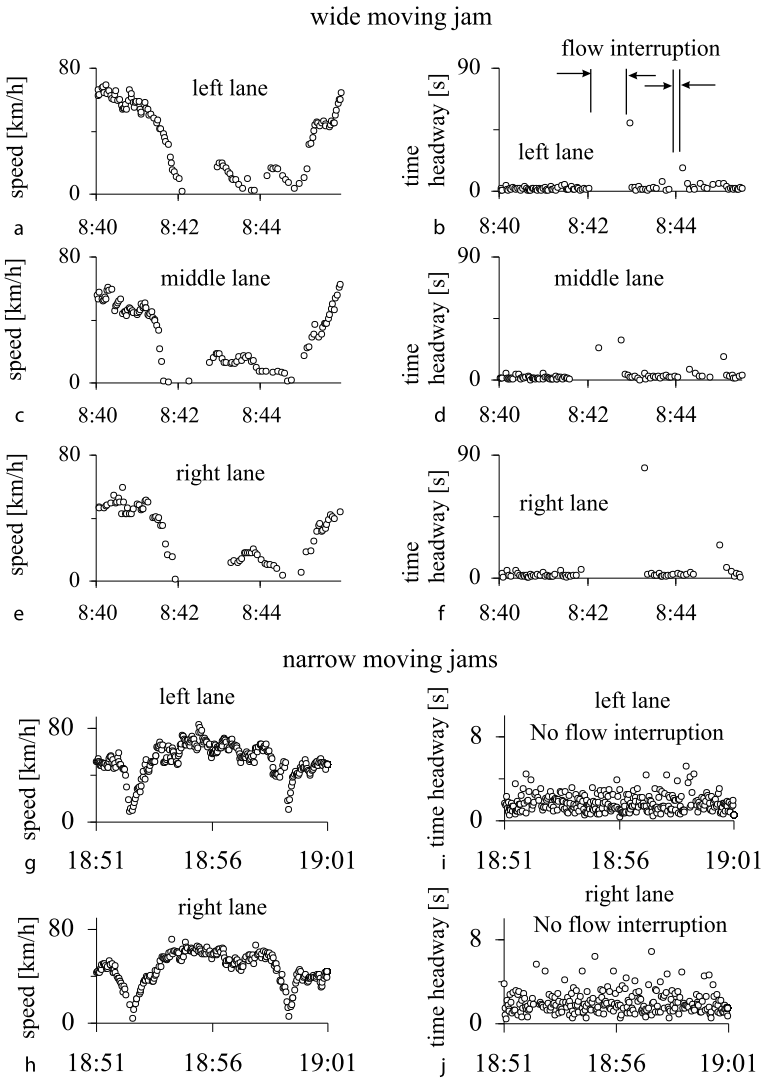

Traffic Congestion, Modeling Approaches to, Figure 6
Measured single vehicle data analysis: Single vehicle data for speed in different freeway lanes for moving jams (*left figures*) and the related time headways (*right figures*). Taken from [67]

example shown in Fig. 6a–f, flow interruption effect occurs two times during upstream jam propagation through a road detector (these time intervals are labeled "flow interruption" in Fig. 6b). The values τ_{max} for the first flow interruption intervals within the wide moving jam are equal to approximately 50 s in the left lane, 24 s in the middle line, and 80 s in the right line. These values τ_{max} satisfy the criterion (2). This means that traffic flow is discontinuous within the moving jam, i. e., this moving jam can be associated with the wide moving jam phase. Later, some vehicles within the jam have time headways of about 4 s or longer associated with moving blanks [67,68].

In another example shown in Fig. 6g–j, there are two moving jams. However, rather than wide moving jams these moving jams can be classified as narrow moving jams. A narrow moving jam is a moving jam, which consists of the jam fronts only. Narrow moving jams are associated with the synchronized flow phase. This is because there are no traffic flow interruption intervals within a narrow moving jam. Indeed, in empirical observations upstream and downstream of the jam, as well as within the jam there are many vehicles that traverse the induction loop detector. There is no qualitative difference in time headways for different time intervals associated with these narrow moving jams and in traffic flow upstream or downstream of the jams (Fig. 6i, j).

This can be explained if it is assumed that each vehicle, which meets a narrow moving jam, which consists of the jam fronts only, can nevertheless accelerate later almost without any time delay within the jam. Within the upstream front vehicles must decelerate to a very low speed. However, the vehicles can accelerate almost immediately at the downstream jam front. These assumptions are confirmed by single vehicle data shown in Fig. 6i, j,

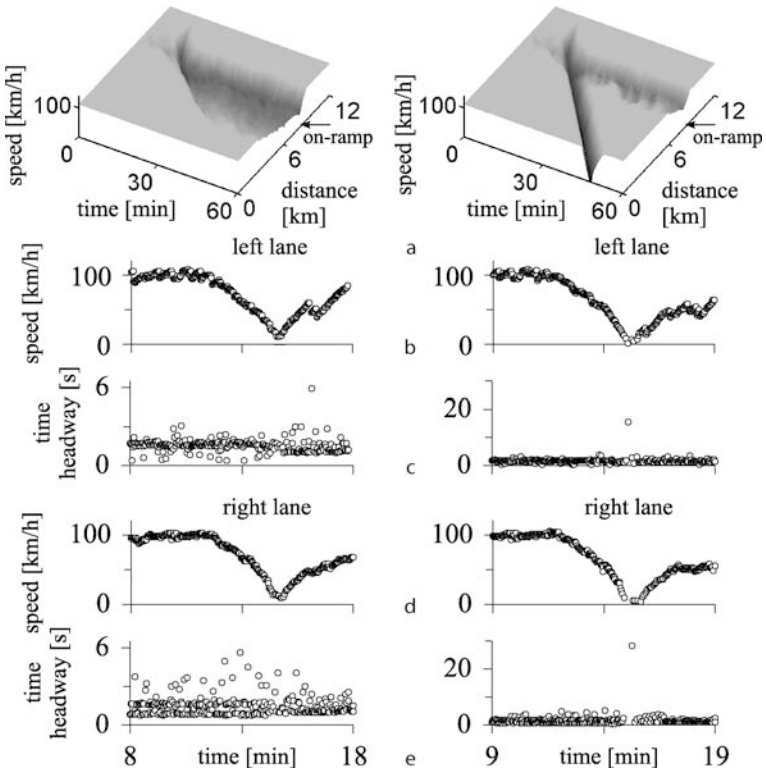

Traffic Congestion, Modeling Approaches to, Figure 7

Comparison of microscopic criterion with macroscopic spatiotemporal objective criteria [J] and [S] for the phases in congested traffic of identical vehicles (results of model simulations). *Left figures* are related to a narrow moving jam. *Right figures* are related to a wide moving jam. a The narrow moving jam is caught at an on-ramp bottleneck that is associated with the synchronized flow definition [S] (*left*); in contrast, the wide moving jam propagates through the bottleneck that is associated with the wide moving jam definition [J] (*right*). Vehicle speed on the main road in space and time that is averaged across two lanes. b,d Single vehicle data for speed in the *left* (b) and in *right lanes* (d) at location 50 m downstream of the end of the on-ramp merging region. c,e Time headways associated with b and d, respectively. Taken from [68]

in which time intervals between different measurements of time headways for different vehicles exhibit the same behavior away and within the jams. Even if within a narrow moving jam there are vehicles that are stopped, the condition

$$I_s \sim 1 \tag{3}$$

can be satisfied. Under this condition, there is no flow interruption within this jam. Thus regardless of these narrow moving jams, traffic flow is not discontinuous, i. e., the narrow moving jams are associated with the synchronized flow phase. This single vehicle analysis enables us to assume that congested traffic in Fig. 6a–f is associated with a wide moving jam. In contrast, congested traffic in Fig. 6g–j is associated with the synchronized flow phase.

It must be stressed that for a proof of the microscopic criterion for the traffic phases (2), we should compare results of phase identification based on this criterion in measured single vehicle data with the traffic phase definitions, i. e., with the macroscopic criteria for the traffic phases [J] and [S] (Sect. "Empirical Macroscopic Criteria for Defining Phases in Congested Traffic"). However, such a empirical proof can be possible, if the single vehicle data associated with a congested pattern in the neighborhood of a bottleneck is measured at many freeway locations including locations upstream and downstream of this bottleneck; this is because only in this case can we prove the behavior of the downstream front of the congested pattern at the bottleneck. Unfortunately, we do not have such measured single vehicle data. For this reason, we can make a proof of the microscopic criterion for the traffic phases based on numerical simulations in the context of three-phase traffic flow theory only. This proof, which is based on numerical simulations of the Kerner–Klenov stochastic three-phase traffic flow model (Sect. "Kerner–Klenov Stochastic Three-Phase Traffic Flow Model"), has been performed in [68] (Fig. 7).

The results of numerical simulations (Fig. 7) allow us to conclude that the microscopic criterion (2) enables us indeed to distinguish the wide moving jam and synchronized flow phases in single vehicle data measured at one road location.

Characteristic Parameters of Wide Moving Jam Propagation and Line *J*

The critical conclusion about earlier traffic flow models made in Sect. "Introduction" does not concern characteristic parameters of wide moving jam propagation, which was firstly discovered by Kerner and Konhäuser in 1994 [63] and later incorporated in a number of different

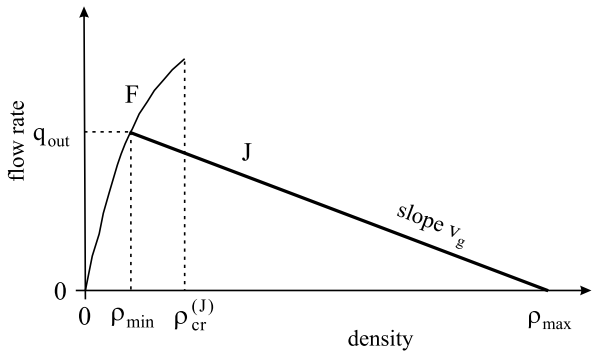

Traffic Congestion, Modeling Approaches to, Figure 8
Qualitative representation of the characteristic line for the downstream front of a wide moving jam (*line J*) whose slope is equal to the downstream jam front velocity v_g, and explanation of the metastability of states of free flow (*curve F*) with respect to moving jam formation [63]

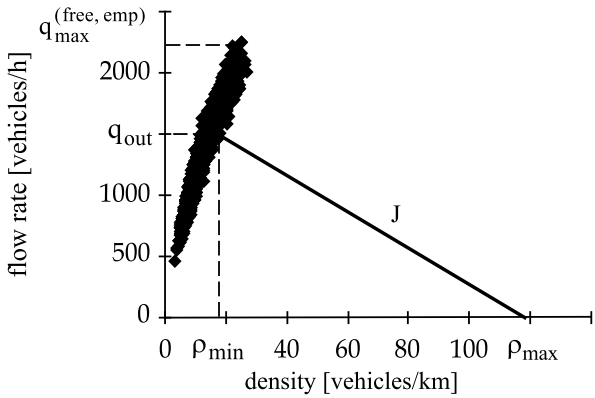

Traffic Congestion, Modeling Approaches to, Figure 9
Representation of the steady propagation of the downstream front of the downstream jam in the sequence of two wide moving jams in Fig. 5 by the line *J* in the flow-density plane and data for free flow (*black diamonds*). Taken from [52]

traffic flow models (see references in [40]). The Kerner–Konhäuser theory of wide moving jam propagation has been confirmed by empirical data in which wide moving jam propagation has been found (see references in [52]).

The characteristic feature of wide moving jam propagation [J], which defines the wide moving jam phase in congested traffic, i. e., a constancy of the mean velocity v_g of the downstream front of the jam while the jam propagates through bottlenecks can be presented by a line in the flow-density plane. This line is called the line *J* (Figs. 8 and 9). The slope of the line *J* is equal to the characteristic velocity v_g. If in the wide moving jam outflow free flow is formed, then the flow rate q_{out} in this jam outflow and the related density ρ_{min} give the left coordinates of

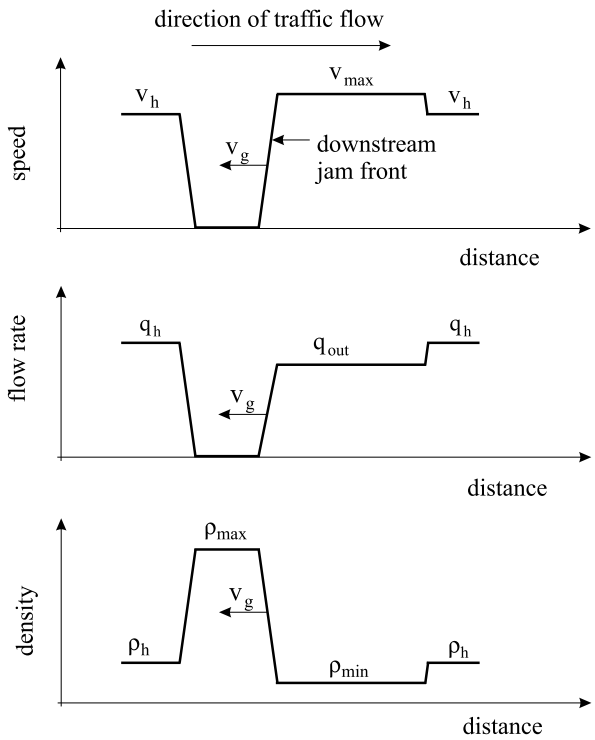

Traffic Congestion, Modeling Approaches to, Figure 10
Characteristic parameters of wide moving jams. Schematic representation of a wide moving jam at a fixed time [63]. Spatial distribution of vehicle speed v, flow rate q, and vehicle density ρ in the wide moving jam, which propagates through a homogeneous state of free flow with speed v_h, flow rate q_h, and density ρ_h

the line J; q_{out}, ρ_{min}, and the related mean vehicle speed v_{max} are characteristic parameters (Fig. 10). When control parameter of traffic (weather, percentage of long vehicles, etc.) do not change, the jam characteristic parameters do not depend on initial conditions, and they are the same for different wide moving jams. The right coordinates of the line J are related to the traffic variables within the jam, the density ρ_{max} and the average vehicle speed v_{min}. For simplicity we assume here that $v_{min} = 0$ (low speed states within wide moving jams associated with so-called moving blanks have been studied in ► Traffic Congestion, Spatiotemporal Features of and [67,68]).

Concerning the characteristic jam velocity v_g, some general assumptions can be made [52]. Each driver standing within a wide moving jam can start to accelerate from the standstill inside the jam to free flow downstream after two conditions have been satisfied:

(i) The preceding vehicle has already begun to move away from the jam.

(ii) Due to the preceding vehicle motion, after some time the net distance (space gap) between the two vehicles has exceeded a space gap g_{del}.

The mean time delay $\tau_{del}^{(a)}$ in vehicle acceleration (Sect. "Microscopic Criterion for Traffic Phases in Congested Traffic") corresponds to the space gap g_{del} between two successive drivers that accelerate from the standstill within the wide moving jam to free flow downstream.

The motion of the downstream front of a wide moving jam results from acceleration of drivers from the standstill inside the jam to flow downstream of the jam. Because the average distance between vehicles inside the jam, including average length of each vehicle, equals $1/\rho_{max}$, the velocity of the downstream front of the wide moving jam is

$$v_g = -\frac{1}{\rho_{max}\tau_{del}^{(a)}} . \tag{4}$$

In empirical observations, $v_g \sim -15$ km/h.

The line J and the characteristic parameters of wide moving jam propagation in free flow [63] are apparently the only results of earlier traffic flow theories to traffic congestion, which are confirmed in empirical observations of spatiotemporal features of congested patterns (see also Sect. "Classic General Motors (GM) Model Approach: Free Flow Instability due to Driver Reaction Time").

Traffic Breakdown and Highway Capacity

Spontaneous Traffic Breakdown (F → S Transition)

Traffic breakdown is accompanied by a sharp decrease in average vehicle speed in the free flow to a considerably lower speed in congested traffic (Figs. 11–13) (see, e. g., [30,38,39,93,105]).

Before traffic breakdown occurs at the bottleneck shown in Fig. 11a, there is free flow at the bottleneck as well as upstream and downstream in a neighborhood of the bottleneck. In this case, the breakdown is caused by occurrence and subsequent growth of speed disturbances (fluctuations) within the free flow at the bottleneck. In accordance with definitions of phase transitions in complex systems, such traffic breakdown is called *spontaneous* traffic breakdown. There can be various sources of speed disturbances whose growth lead to traffic breakdown, e. g., unexpected braking of a vehicle, lane changing, fluctuations in flow rates upstream of the bottleneck, vehicle merging onto the main road from other roads (e. g., at on-ramp bottlenecks), etc.

Traffic breakdown occurs usually at the same freeway bottlenecks of a freeway section. These bottlenecks are

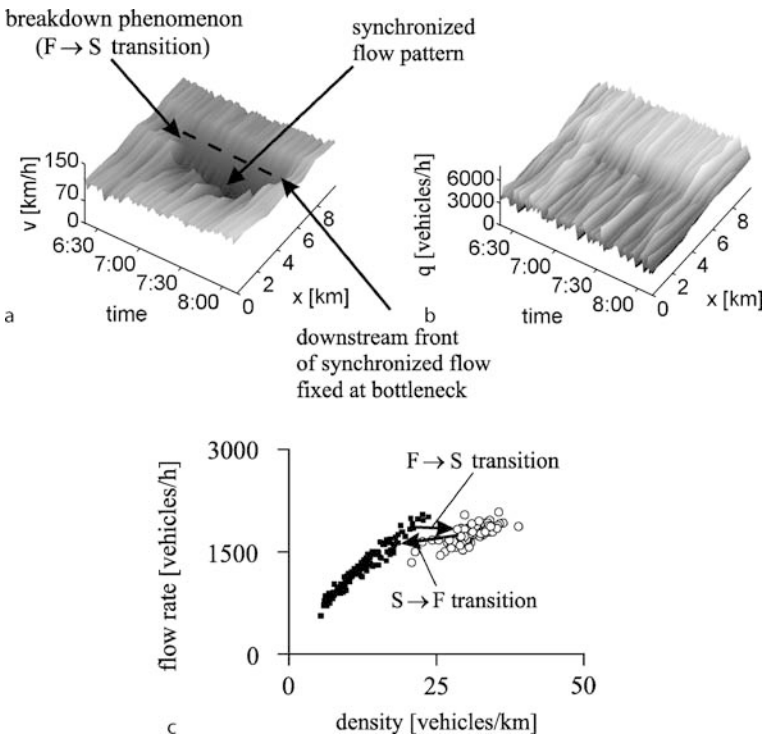

Traffic Congestion, Modeling Approaches to, Figure 11

Empirical example of traffic breakdown and hysteresis effect at on-ramp bottleneck: a,b Average speed (a) and flow rate (b) on the main road in space and time. c Hysteresis effect in the flow-density plane. 1-min average data. Taken from [52]. This example of traffic breakdown is qualitatively the same as many other examples observed in various countries (e. g., [30,38,39,105])

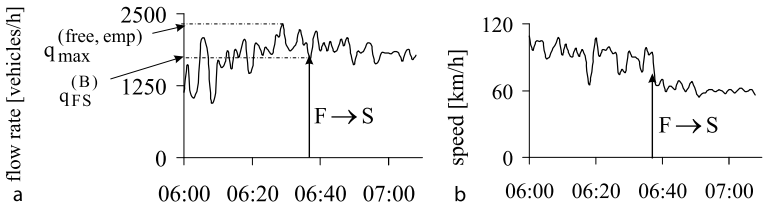

Traffic Congestion, Modeling Approaches to, Figure 12

Empirical example in which flow rate downstream of the bottleneck at which traffic breakdown occurs, $q_{FS}^{(B)}$, is smaller than $q_{max}^{(free, emp)}$. Taken from [52]

called *effectual bottlenecks*. An effectual bottleneck is a bottleneck where traffic breakdown most frequently occurs on many different days. Examples of effectual bottlenecks are adjacent bottlenecks B_1 and B_2 in Fig. 4b.

Elefteriadou et al. found that traffic breakdown has a probabilistic nature [30] that means the following: at a given flow rate traffic breakdown can occur but it should not necessarily occur. Thus on one day traffic breakdown occurs, however, on another day at the same flow rates and at the same traffic conditions traffic breakdown is not observed. In 1998, Persaud et al. found [105] that empirical probability of traffic breakdown at a bottleneck is an increasing flow rate function (Fig. 14). Later such an empirical probability of traffic breakdown was also found on different freeways in various countries [10,11,12,87,94,131].

Another empirical probabilistic characteristic of traffic breakdown is as follows. At given traffic parameters (weather, etc.), the flow rate downstream of an on-ramp bottleneck associated with the empirical maximum flow rate in free flow $q_{max}^{(free, emp)}$, which was measured on a specific day before congestion occurred, can be greater

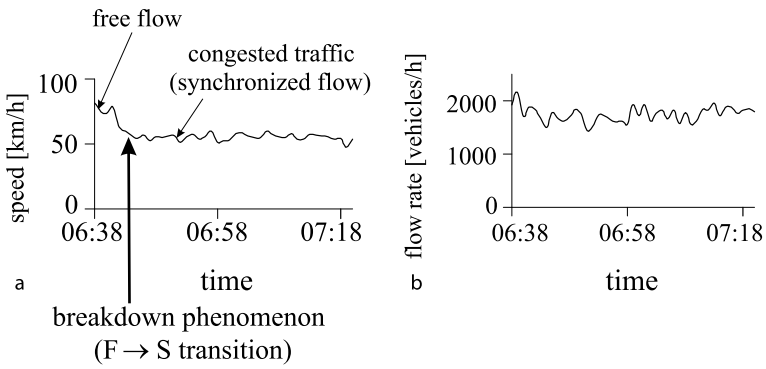

Traffic Congestion, Modeling Approaches to, Figure 13
Traffic breakdown at an on-ramp bottleneck. Vehicle speed (a) and flow rate downstream of the bottleneck (b) as functions of time related to Fig. 11 (1-min average data). This example of traffic breakdown is qualitatively the same as many other examples observed in various countries (e. g., [30,38,39,105])

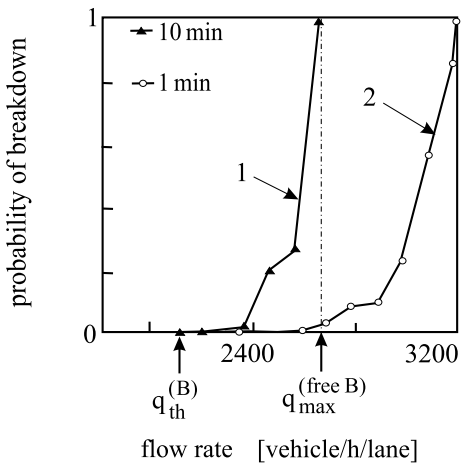

Traffic Congestion, Modeling Approaches to, Figure 14
Probability for traffic breakdown at an on-ramp bottleneck for two flow rate averaging time constants T_{av} (T_{av} =1 and 10 min). Data is measured on a freeway in Toronto, Canada. Taken from Persaud et al. [105] Note that the flow rates $q_{max}^{(free\ B)}$ and $q_{th}^{(B)}$ made on this figure are respectively the maximum and minimum free-way capacities of free flow at the bottleneck [52] shown for the averaging time interval for empirical data T_{av} = 10 min (see explanation in Sect. "Highway Capacity and First-Order F → S Transition")

than the flow rate $q_{FS}^{(B)}$ at which traffic breakdown occurs (Fig. 12).

Congested traffic in Fig. 13 that has occurred due to traffic breakdown shows a further development in space and time (Fig. 11a, b), with the following effects observed in different countries (e. g., [30,38,39,105]):

(i) The upstream front of congested traffic propagates upstream. The upstream front separates free flow upstream from congested traffic downstream. Thus the region of congestion broadens in the upstream direction.
(ii) The downstream front of congested traffic is fixed at the bottleneck (dashed line in Fig. 11).

Corresponding to the objective criteria for traffic phases in congested traffic [J] and [S] considered previously, the result of item (ii) means that the congested traffic resulting from traffic breakdown belongs to the synchronized flow phase. In all known observations, traffic breakdown is associated with an F → S transition. Thus the terms "F → S transition," "breakdown phenomenon," "traffic breakdown," and "speed breakdown" are synonyms related to the same effect: the onset of congestion in free flow.

The congested pattern (Fig. 11a, b) exists for about one hour at the bottleneck: at 7:40 free flow occurs at the bottleneck. This restoration of free flow is related to a reverse phase transition from synchronized flow to the free flow (S → F transition for short) at the bottleneck. The F → S and S → F transitions are accompanied by a well-known *hysteresis effect* and hysteresis loop in the flow-density plane: a congested pattern emerges usually at a greater flow rate downstream of the bottleneck than this flow rate is at which the congested pattern dissolves (see references in [38,39]) (Fig. 11c).

Traffic Breakdown Explanation Through Fundamental Hypothesis of Three-Phase Traffic Theory

In three-phase traffic theory, traffic breakdown (F → S transition) is explained by a competition between two opposing tendencies occurring within a random local speed

disturbance in which the speed is lower and vehicle density is greater than in an initial free flow [52]:

(i) A tendency towards the initial free flow due to vehicle acceleration associated with an *over-acceleration effect*;
(ii) A tendency towards synchronized flow due to vehicle deceleration associated with a *speed adaptation effect*.

To explain this mechanism of traffic breakdown, we use the fundamental hypothesis of the three-phase traffic theory formulated as follows [49,50,51]:

- Steady states of synchronized flow cover a two-dimensional (2D) region in the flow-density plane (Fig. 15a). The multitudes of free flow states overlap steady states of synchronized flow in the vehicle density. The free flow states on a multi-lane road and steady states of synchronized flow are separated by a gap in the flow rate and, therefore, by a gap in the speed at a given density: at each given density the synchronized flow speed is lower than the free flow speed.

This hypothesis and its consequences (Fig. 15b–d) contradict qualitatively with the fundamental diagram hypothesis and associated relationships of earlier traffic flow theories (Fig. 1). In particular, the fundamental hypothesis of the three-phase traffic theory states that for hypothetical steady states of congested traffic (synchronized flow) there is *no* link travel time-flow relationship (Fig. 15d). This is in contrast with a link travel time-flow relationship (Fig. 1d) whose existence is the basic assumption of all known theories and models of congested traffic networks (see references in ► Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment).

It should be noted that a well-known wide scattering of empirical data for congested traffic in the flow–density plane *cannot* be considered a proof of the fundamental hypothesis of three-phase traffic theory: a 2D-region for steady states of synchronized flow (Fig. 15a) is a *theoretical* assumption only. Indeed, the wide scattering of measured data for congested traffic in the flow–density plane can be associated for example with a variety of driver characteristics and vehicle parameters as well as with speed fluctuations, which are present in real traffic flow. As a proof of the fundamental hypothesis, mathematical results of traffic flow models can be considered in which this and other hypotheses of three-phase traffic theory have been used. This is because these results explain and predict empirical features of traffic breakdown and resulting congested patterns [52].

The low boundary of the 2D region for steady states of synchronized flow in the flow-density plane (labeled S_{low} in Fig. 15a) is associated with a *synchronization space gap* G between vehicles; at $g > G$ a vehicle accelerates. The upper boundary of the 2D region (labeled S_{upper}) is associ-

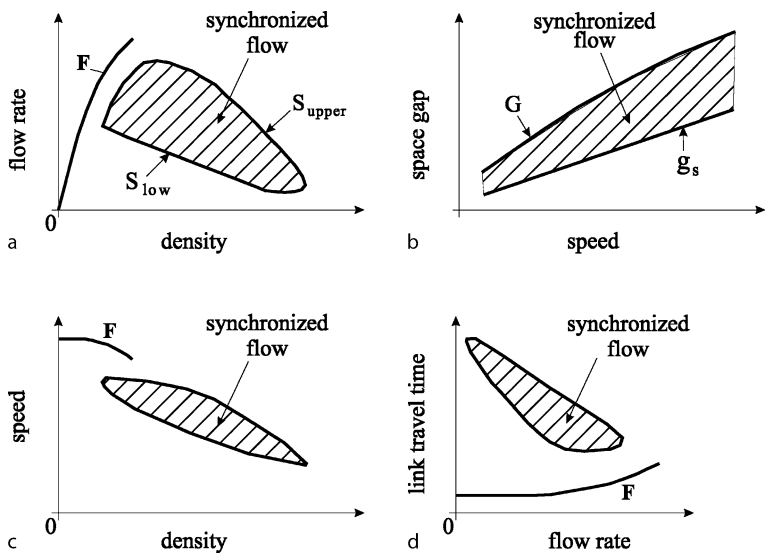

Traffic Congestion, Modeling Approaches to, Figure 15
Fundamental hypothesis of three-phase traffic theory [49,50,51]: a Qualitative representation of free flow states (F) and 2D steady states of synchronized flow (*dashed region*) in the flow-density plane. b Qualitative representation of a part of the 2D steady states of synchronized flow shown in a in the space-gap-speed plane (*dashed region*). c,d Qualitative representation of free flow states (F) and 2D steady states of synchronized flow in the speed-density (c) and link-travel-time-flow (d) planes associated with a

ated with a *safe space gap* g_s, i. e., at $g < g_s$ the vehicle decelerates. The synchronization gap is greater than the safe one at each given vehicle speed: $G(v) > g_s(v)$ (Fig. 15b). Both the synchronization gap G and safe gap g_s are usually speed increasing functions.

In hypothetical steady states of synchronized flow, the synchronization gap is defined through the condition that at each given steady speed v in synchronized flow there are *infinite* space gaps g within the range

$$g_s \leq g \leq G \tag{5}$$

at which a driver can move with this steady speed v. In other words, through condition (5) the synchronization gap determines the existence of the 2D region for steady states of synchronized flow (Fig. 15). The condition (5) is consistent with a driver behavioral assumption that at relatively small space gaps in synchronized flow a driver can recognize whether the space gap to the preceding vehicle increases or decreases over time. This is true even if the speed difference between the vehicle speed and the speed of the preceding vehicle is negligible.

The condition (5) is also consistent with a driver behavioral assumption that for comfortable driving in car-following a driver tends to adapt the speed to the speed of the preceding vehicle within a greater space gap than the safe one. This driver behavior about speed adaptation with the synchronized gap explains the 2D region for steady states of synchronized flow (Fig. 15) introduced in three-phase traffic theory [49,50,51]. To understand the statement about a strong connection between the 2D region for steady states of synchronized flow and the dynamic driver behavior about speed adaptation, let us discuss a general case of car-following in which speeds of the vehicle and the preceding vehicle are not time-independent.

Then rather than the synchronization gap that determines the 2D region of steady states of synchronized flow via the condition (5), we should consider a *dynamic* synchronization space gap, which can depend on the vehicle speed *and* the speed of the preceding vehicle; if speeds of the vehicle and the preceding vehicle are equal to each other and time-independent, then the dynamic synchronization gap is equal to the synchronization gap in (5).

As above mentioned, at $g > G$ a vehicle accelerates, whereas at $g < g_s$ the vehicle decelerates. We define the dynamic synchronization gap as a space gap within which a vehicle tends to adapt the speed to the speed of the preceding vehicle without caring, what the precise space gap is, as long as this space gap is not smaller than the safe gap. The space gap at which this speed adaptation has finished can be any space gap from the infinite number of space gaps within the space gap range

$$g_s \leq g \leq G(v, v_\ell) \tag{6}$$

in which it is suggested that the dynamic synchronization gap $G(v, v_\ell)$ can be a function of the vehicle speed v and the speed of the preceding vehicle v_ℓ. This speed adaptation within the dynamic synchronization gap is called the *speed adaptation effect* (Fig. 16a) [52].

For hypothetical steady states of synchronized flow, the dynamic synchronization gap is equal to the synchronization gap in (5), i. e., for the steady states the synchronization gaps in the conditions (5) and (6) are identical ones. For this reason, we use the term "synchronization gap" for any case of car-following behavior. Thus the speed adaptation effect within the synchronization gap is indeed associated with the 2D-region of steady states of synchronized flow given by the condition (5) (Fig. 15).

The condition (5) is in contradiction with the fundamental diagram hypothesis of all the earlier traffic flow theories (see Sect. "Introduction"): for congested traffic there is no fundamental diagram in three-phase traffic theory. To explain this statement, let us consider a hypothetical case when a vehicle approaches the preceding vehicle that moves with a slower time-independent speed. Then the vehicle adapts the speed to the speed of the preceding vehicle at a space gap, which is only one possible space gap from the infinite number of space gaps between the synchronization gap G and the safe gap g_s within the above 2D-region at this speed, i. e., the gaps that satisfy (5). There is no desired (or optimal) space gap, when a vehicle follows the preceding vehicle that moves with at a time-independent speed. This feature is associated with a speed adaptation effect within the synchronization gap.

Under condition (6), a vehicle accelerates when it is slower than the preceding vehicle, and decelerates usually when it is faster than the preceding vehicle. However, under condition (6) the vehicle can also accelerate, even if it is not slower than the preceding vehicle. This vehicle acceleration is called *vehicle over-acceleration* (over-acceleration for short) [52]. In general, the over-acceleration is defined as vehicle acceleration from car-following, in particular, when the space gap g is satisfied to condition (6), i. e., the vehicle is within the synchronization gap.

To understand the sense of the term *over-acceleration*, we consider a scenario in which a vehicle that moves in free flow on a multi-lane road approaches a slower moving preceding vehicle. If firstly the vehicle cannot pass the preceding vehicle, then the vehicle decelerates within the synchronization gap to the speed of the preceding vehicle, i. e., the speed adaptation effect is realized leading to

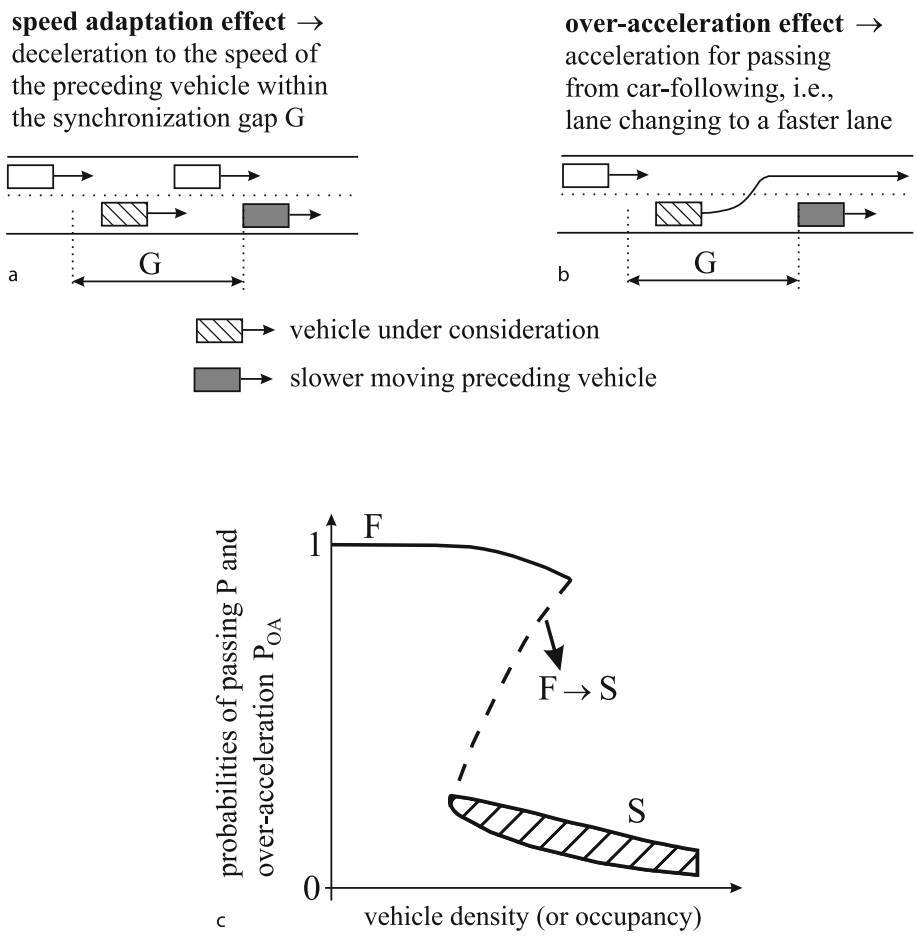

Traffic Congestion, Modeling Approaches to, Figure 16
Qualitative explanation of competition of vehicle deceleration due to speed adaptation (a) and vehicle acceleration due to over-acceleration (b) [51,52]. c Qualitative Z-shaped density function for passing probability and probability of over-acceleration. In c, *F* – free flow, *S* – synchronized flow, *dashed curves* between states *F* and *S* are related to critical probabilities for passing and over-acceleration associated with a critical speed required for traffic breakdown: When the speed within a speed disturbance is lower than the critical speed, then probabilities for passing and over-acceleration are smaller than the critical probabilities; as a result, traffic breakdown occurs (*arrows* labeled by F → S)

car-following of the slow preceding vehicle (Fig. 16a). We assume in this scenario that later the vehicle can pass this slow moving preceding vehicle. As well-known, to pass the preceding vehicle, the vehicle should change lane and accelerate. The vehicle acceleration takes place, even if the vehicle is *not* currently *slower* than the preceding vehicle; this explains why this vehicle acceleration is called *over-acceleration*. Thus in the case under consideration, over-acceleration is vehicle acceleration for passing from car-following, i. e., lane changing to a faster lane (Fig. 16b); as a result, the probability of the over-acceleration denoted by P_{OA} is equal to the passing probability denoted by P:

$$P_{OA} = P\,.$$

Passing in free flow is much more probable than in synchronized flow: passing probability P is greater in free flow than in synchronized flow. This obvious suggestion together with the fundamental hypothesis of three-phase traffic theory, specifically that at each given density the synchronized flow speed is lower than the free flow speed lead to the following hypothesis [51,52]:

- Passing probability P and, therefore, over-acceleration probability P_{OA} exhibit a *discontinuous* character; in particular, these probabilities are Z-shaped density functions (Fig. 16c): At a given density, there is a *finite drop* in passing and over-acceleration probabilities, when free flow transforms onto synchronized flow.

The Z-shape density function for probabilities of passing and over-acceleration (Fig. 16c) explain traffic breakdown as follows [52]. Free flow remains on a multi-lane road as long as the over-acceleration effect, which describes the tendency towards free flow, is stronger than the speed adaptation effect that describes the tendency towards synchronized flow. However, the greater the density in free flow, the smaller the probability of over-acceleration P_{OA}, i. e., the weaker the over-acceleration effect.

There is a critical probability of over-acceleration (dashed curve between states F and S in Fig. 16c) associated with a critical speed within a local speed disturbance required for traffic breakdown (explanation of the critical speed for traffic breakdown appears in Sects. "Highway Capacity and First-Order F → S Transition" and "Empirical Double Z-Characteristic for Phase Transitions in Traffic Flow"). This means that when probability of over-acceleration is equal to the critical probability of over-acceleration, then the tendency to free flow due to over-acceleration within the disturbance is on average exactly as strong as the tendency to synchronized flow due to speed adaptation.

When the speed within a local speed disturbance is lower than the critical speed, then probability of over-acceleration is smaller than the critical probability of over-acceleration. This means that the over-acceleration effect is weaker than the speed adaptation effect. In this case, the tendency to synchronized flow due to speed adaptation overcomes the tendency to free flow due to over-acceleration that results in traffic breakdown (arrow labeled by F → S in Fig. 16c).

In free flow at a bottleneck, there is a permanent local speed decrease called a permanent speed disturbance. This speed disturbance is localized in a neighborhood of the bottleneck. For example, for an on-ramp bottleneck the permanent speed disturbance is due to vehicles that merge from the on-ramp onto the main road. This speed disturbance exists even if there were no random speed disturbances in free flow. The permanent speed decrease within the disturbance increases the probability of traffic breakdown at the bottleneck considerably in comparison to the traffic breakdown probability at other freeway locations away from bottlenecks at which a speed disturbance should initially occur in free flow for traffic breakdown. This explains why in empirical observations traffic breakdown occurs mostly at bottlenecks.

Due to traffic breakdown, a variety of congested traffic patterns can occur at a bottleneck. There are congested patterns that consist of the synchronized flow traffic phase only. These patterns are called *synchronized flow patterns* (SP for short). The congested pattern, which emerges as a result of the F → S transition discussed above (Fig. 11) consists of synchronized flow only, i. e., this is an example of an SP. In more detail, various SPs and other congested patterns are discussed in ► Traffic Congestion, Spatiotemporal Features of.

Induced Traffic Breakdown

In empirical observation, besides spontaneous traffic breakdown at a bottleneck (Figs. 11 and 13) (spontaneous F → S transition) there can also be an induced traffic breakdown (induced F → S transition) [52].

In contrast with spontaneous traffic breakdown, which occurs when before the breakdown free flow is in a neighborhood of the bottleneck (Fig. 11a), the induced F → S transition at the bottleneck is caused by the propagation of a moving spatiotemporal *congested* traffic pattern, which has initially occurred at a *different* road location (e. g., at another bottleneck) than that of the bottleneck. When this congested pattern reaches the bottleneck, the pattern induces traffic breakdown at the bottleneck. The congested pattern whose propagation causes the breakdown at the bottleneck can be considered *external* disturbance at the bottleneck.

In particular, an F → S transition can be induced by a wide moving jam propagating upstream through a bottleneck. This case is shown in Fig. 4a. Propagating through bottleneck B_2, the wide moving jam causes synchronized flow emergence at the bottleneck. Synchronized flow remains at bottleneck B_2 even after the wide moving jam is far upstream of the bottleneck. The induced phase transition is labeled "induced F → S transition" in this figure. In this case, after the wide moving jam has passed the bottleneck a *synchronized flow pattern* (SP for short) is formed. The SP remains at the bottleneck for a long time (this SP is labeled "synchronized flow" in Fig. 4a).

Induced F → S transition at a bottleneck can also occur when a region of synchronized flow first occurs downstream of this bottleneck, and the region later reaches the bottleneck due to the upstream propagation of synchronized flow. However, in contrast to the above case of induced traffic breakdown caused by wide moving jam propagation, the initial synchronized flow is caught at the bottleneck [52].

To show this effect, let us consider propagation of a moving synchronized flow pattern (MSP) upstream of a bottleneck B_1 (labeled MSP in Fig. 17). The MSP propagates upstream. After the MSP reaches the upstream bottleneck B_2, the MSP is caught at the bottleneck. This catch effect is inconsistent with the traffic phase definition [J]. In

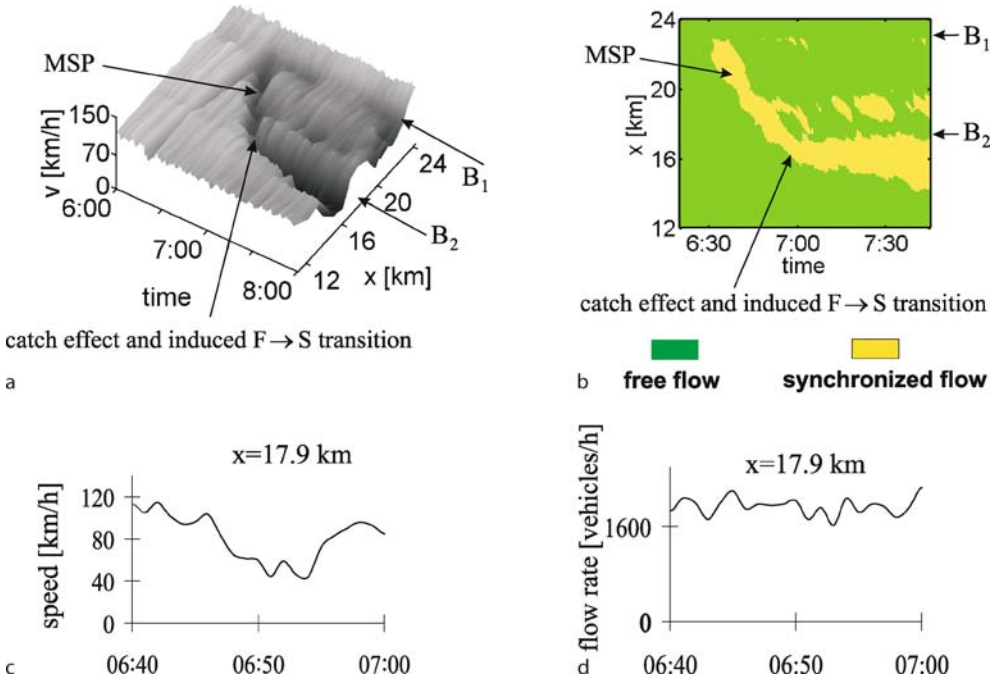

Traffic Congestion, Modeling Approaches to, Figure 17
Empirical example of induced traffic breakdown at bottleneck. a Vehicle speed in space and time. b A graph of a with the free flow phase (*green*) and the synchronized flow phase (*yellow*). c,d Average (across the road) speed (b) and flow rate (c) within the MSP at the location $x = 17.9$ km (about 0.8 km downstream of the bottleneck location B_2). Bottlenecks B_2 and B_1 are the same as those in Fig. 4. Taken from [52]

other words, the MSP satisfies the traffic phase definition [S], i. e., this pattern is indeed an SP.

Highway Capacity and First-Order F → S Transition

The possibility of both induced and spontaneous traffic breakdowns at the same bottleneck can be explained by the *nucleation* nature of traffic breakdown. This means that there should be some critical speed required for traffic breakdown. The breakdown occurs only, if due to a speed disturbance in free flow in the neighborhood of a bottleneck the speed decreases below the critical one. Otherwise, if free flow speed within the disturbance is greater than the critical one, breakdown does not occur.

The critical speed should depend on the flow rate downstream of the bottleneck; the smaller the flow rate, the lower the critical speed should be required for the breakdown in free flow. In contrast, when the flow rate increases, the critical speed at the bottleneck required for the breakdown increases. This means that the greater the flow rate downstream of the bottleneck, the smaller should be a local speed decrease in free flow at the bottleneck to cause the breakdown. Obviously, the lower this speed decrease should be, the smaller the probability for the speed decrease. This explains the increasing character of the probability of traffic breakdown (Fig. 14).

Thus traffic breakdown at a bottleneck exhibits the following **fundamental empirical features**:

A. Traffic breakdown is an F → S transition.
B. At the same bottleneck, traffic breakdown can be either *spontaneous* or *induced*.
C. Onset and dissolution of congestion are accompanied by a *hysteresis* effect.
D. Traffic breakdown exhibits the *probabilistic* nature.

In accordance with other systems of natural science, when the speed within a speed disturbance in free flow at the bottleneck is equal to or lower than the critical speed, this speed disturbance can be called critical speed disturbance or a *nucleus* for traffic breakdown. Thus the probabilistic feature of traffic breakdown is associated with the nucleation character of the breakdown. Only if a critical speed disturbance appears at the bottleneck, traffic breakdown occurs. In other words, the probability for traffic breakdown is the probability of critical speed disturbance appearance at the bottleneck.

The empirical features of traffic breakdown **A-D** mean that the breakdown is a first-order F → S transition. In accordance with first-order F → S transition features, there is the maximum flow rate (denoted $q_{\max}^{(\text{free B})}$ and that is shown in Fig. 14 for the averaging time interval for empirical data $T_{\text{av}} = 10$ min only) downstream of the bottleneck at which breakdown probability reaches one, i. e.,

$$P_{\text{FS}}^{(\text{B})} = 1 \,. \tag{7}$$

There is also the threshold flow rate (denoted $q_{\text{th}}^{(\text{B})}$). At the flow rate in free flow downstream of the bottleneck that is smaller than $q_{\text{th}}^{(\text{B})}$, breakdown probability

$$P_{\text{FS}}^{(\text{B})} = 0 \,. \tag{8}$$

Highway (freeway) capacity of free flow at a bottleneck is limited by traffic breakdown at the bottleneck. Thus the fundamental empirical features of traffic breakdown **A-D** are also the fundamental empirical features for highway capacity.

For a given averaging time interval for traffic variables T_{av}, traffic breakdown can occur with probability $P_{\text{FS}}^{(\text{B})}$. In other words, an attribute of this *probabilistic highway capacity at the bottleneck* (bottleneck capacity) is the probability

$$P_{\text{C}}^{(\text{B})} = 1 - P_{\text{FS}}^{(\text{B})} \tag{9}$$

that free flow remains at the bottleneck during the time interval T_{av} [52]. Highway capacity is reached when

$$P_{\text{C}}^{(\text{B})} < 1 \,. \tag{10}$$

Thus the definition of highway capacity made in three-phase traffic theory, which satisfies the fundamental empirical features of traffic breakdown **A-D**, reads as follows [52]:

- Highway capacity of free flow is equal to the flow rate downstream of the bottleneck at which free flow remains at the bottleneck with the probability $P_{\text{C}}^{(\text{B})} < 1$ (10) during a given averaging time interval T_{av}.

Because there can be an infinite number of flow rates downstream of the bottleneck for which the condition (10) is satisfied, there can also be an infinite number of capacities of free flow at the bottleneck. Each of the freeway capacities has two attributes:

(1) The probability $P_{\text{C}}^{(\text{B})}$ (10) that free flow remains at the bottleneck during a given averaging time interval for traffic variables T_{av}.

(2) The time interval T_{av}.

When for free flow at the bottleneck rather than the condition (10) the condition

$$P_{\text{C}}^{(\text{B})} = 1 \tag{11}$$

is satisfied, then the flow rate downstream of the bottleneck in this free flow is smaller than any of the freeway capacities. In accordance with (9), condition (11) is equivalent to (8). Thus the flow rate in free flow downstream of the bottleneck denoted by q_{sum} that is smaller than any of the capacities satisfy

$$q_{\text{sum}} < q_{\text{th}}^{(\text{B})} \,. \tag{12}$$

For this reason, the threshold flow rate $q_{\text{th}}^{(\text{B})}$ is the minimum capacity.

In accordance with (9), condition (7) is equivalent to $P_{\text{C}}^{(\text{B})} = 0$. Thus the maximum flow rate $q_{\max}^{(\text{free B})}$ determines the maximum capacity. Because $q_{\max}^{(\text{free B})}$ is considerably greater than $q_{\text{th}}^{(\text{B})}$, there are infinite freeway capacities denoted by $q_{\text{C}}^{(\text{B})}$, which satisfy

$$q_{\text{th}}^{(\text{B})} \le q_{\text{C}}^{(\text{B})} \le q_{\max}^{(\text{free B})} . \tag{13}$$

Note that for an on-ramp bottleneck, $q_{\max}^{(\text{free B})}$ and $q_{\text{th}}^{(\text{B})}$ depend on q_{on} and q_{in}; for this reason, there are also infinite maximum and infinite minimum capacities associated with different values $q_{\max}^{(\text{free B})}$ and $q_{\text{th}}^{(\text{B})}$, respectively.

The maximum and minimum capacities can depend considerably on "control" parameters of traffic (weather, percentage of long vehicles, etc.). In empirical observations, they are found from a study of a flow rate dependence of breakdown probability $P_{\text{FS}}^{(\text{B})}(q_{\text{sum}})$ (Fig. 14). For this reason, $q_{\max}^{(\text{free B})}$ and $q_{\text{th}}^{(\text{B})}$ are mean values found in many different realizations (days) at which traffic breakdowns have occurred.

To explain the meaning of the averaging time interval T_{av} in the capacity definition, we should note that flow rates are usually complicated time-functions in real traffic. For this reason, in empirical observations measured data is usually averaged for a time interval T_{av}. The longer T_{av}, the greater the probability of traffic breakdown at the same average flow rate. For this reason, T_{av} is an important attribute of highway capacity.

The sense of the infinite highway capacities (13) is as follows. At each flow rate q_{sum} downstream of a bottleneck, which satisfies the condition

$$q_{\text{th}}^{(\text{B})} \le q_{\text{sum}} < q_{\max}^{(\text{free B})} \,, \tag{14}$$

traffic breakdown at the bottleneck is possible: traffic breakdown occurs spontaneously with a probability $P_{FS}^{(B)}$, which is within the range $0 < P_{FS}^{(B)} < 1$, or traffic breakdown can be induced. Thus each of these flow rates is highway capacity. Breakdown probability $P_{FS}^{(B)}$ depends on q_{sum} strongly. Thus probability $P_{C}^{(B)}$ (9) that traffic breakdown does not occur spontaneously, i. e., that free flow remains at the bottleneck is an attribute of highway capacity. This capacity attribute distinguishes different highway capacities. The probability function $P_{C}^{(B)}(q_{sum})$ depends on T_{av}. Thus T_{av} is another attribute of highway capacity. Traffic breakdown cannot be possible at the bottleneck only then, when the flow rate q_{sum} satisfies the condition (12). Thus only the flow rates that satisfy the condition (12) are smaller than highway capacity.

In contrast with real traffic, in simulations of traffic with traffic flow models flow rates upstream of a bottleneck can be chosen as time-independent values. In this case, T_{av} in the above capacity definition should be replaced by a time interval for observing traffic flow T_{ob}. The related capacity definition and its relation to diagram of congested patterns at bottlenecks as well as other results of a probabilistic theory of traffic breakdown and highway capacity appears in ► Traffic Breakdown, Probabilistic Theory of.

Relation Between Speed and Flow Rate in Synchronized Flow Resulting from Traffic Breakdown

It can be seen in Fig. 13 that whereas there is a sharp decrease in average vehicle speed, the flow rate does not necessarily abruptly decrease during an F → S transition and within an emergent SP. In most observations, during traffic breakdown the flow rate in the synchronized flow at the location of traffic breakdown is almost as great as in free flow. In some cases, the flow rate in synchronized flow can be even greater than the initial flow rate in free flow [52].

Another example of this feature is shown in Fig. 17c, d in which average speed within an MSP is considerably lower than the one in free flow, whereas the flow rate within the MSP is on average the same as that in free flow. In contrast with this very important feature of many other synchronized flow patterns observed in real traffic, the flow rate within a wide moving jam is considerably smaller than in free and synchronized flows (see Fig. 5c, d).

The empirical feature of synchronized flow, resulting from traffic breakdown that the flow rate within the SP can be as great as in free flow, is the important one for feedback on-ramp metering control applications (see Sect. 23.3 in [52] and a recent review [56]).

Moving Jam Emergence in Synchronized Flow (S → J Transition)

We have already mentioned that wide moving jams do not emerge spontaneously in free flow: no spontaneous phase transition from the free flow phase to the wide moving jam phase (F → J transition for short) has been observed [52]. Wide moving jams can emerge *spontaneously only* in the synchronized flow phase (S → J transition).

In synchronized flow at higher speeds, wide moving jams should not necessarily emerge spontaneously. Observations show that the greater the density in synchronized flow, the more likely is spontaneous moving jam emergence in that synchronized flow. Thus a wide moving jam emerges in an initial free flow due to a sequence of two phase transitions: Firstly, an F → S transition occurs and synchronized flow emerges. Later and usually at other freeway locations than the location of the F → S transition, an S → J transition occurs spontaneously leading to wide moving jam emergence. This sequence of phase transitions is called the F → S → J transitions.

Pinch Effect

In empirical observations, S → J transition development is as follows. Narrow moving jams emerge spontaneously with subsequent self-compression of synchronized flow. This self-compression of synchronized flow with narrow moving jam emergence is called the pinch effect in synchronized flow. In the related pinch region of synchronized flow, the density is great and the speed is low, however, the flow rate can be great enough. Emergent narrow moving jams grow, propagating upstream within the pinch region. Some (or each) of these jams can transform into wide moving jams. Locations of these S → J transitions are the upstream locations of the pinch region.

A congested pattern at a hypothetical isolated bottleneck, which occurs due to the F → S → J transitions, is called a *general pattern* (GP). Thus the GP is a congested pattern, which consists of synchronized flow upstream of the bottleneck and wide moving jams that emerge spontaneously in that synchronized flow [52]. Within the GP, there are two traffic phases of congested traffic, synchronized flow and wide moving jam(s).

A scenario of GP emergence can involve the transformation of an initial SP into a GP (Fig. 18). Firstly, the SP occurs at an off-ramp bottleneck (labeled $B_{North\ 1}$). Over time, the density of synchronized flow increases. In the associated pinch region of synchronized flow, narrow moving jams emerge spontaneously. The narrow moving jams propagate upstream. Some of these narrow moving jams grow over time propagating upstream. Fi-

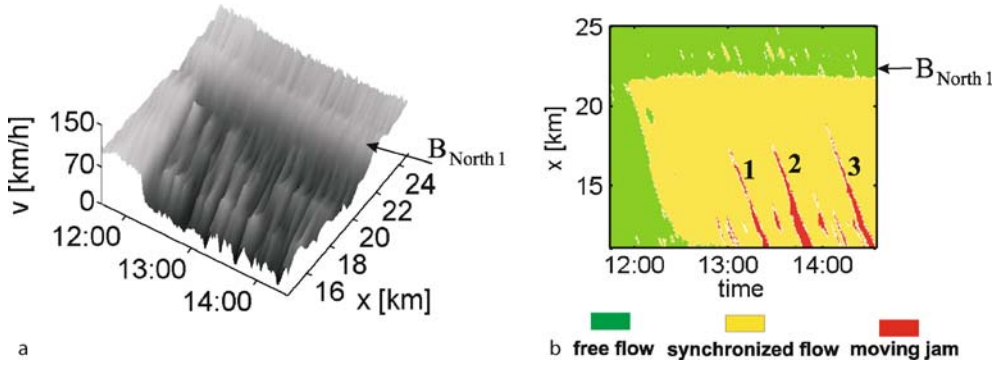

Traffic Congestion, Modeling Approaches to, Figure 18
Empirical example of general pattern (GP) emergence at an off-ramp bottleneck (labeled $B_{\text{North 1}}$). a Vehicle speed in space and time. b A graph of a with the free flow phase (*green*), the synchronized flow phase (*yellow*) and the wide moving jam phase (*red*) (the wide moving jams labeled 1, 2 and 3 in b). Taken from [52]

nally, the growing narrow moving jam can transform into a wide moving jam, i. e., an S → J transition occurs spontaneously. In ► Traffic Congestion, Spatiotemporal Features of, it is shown that the moving jams 1, 2 and 3 shown in Fig. 18b propagate further through an upstream on-ramp bottleneck while maintaining their mean velocities of the downstream jam fronts. Thus in accordance with the traffic phase definition [J], these moving jams are indeed wide moving jams. A more detailed consideration of empirical GPs and other congested traffic patterns appears in ► Traffic Congestion, Spatiotemporal Features of.

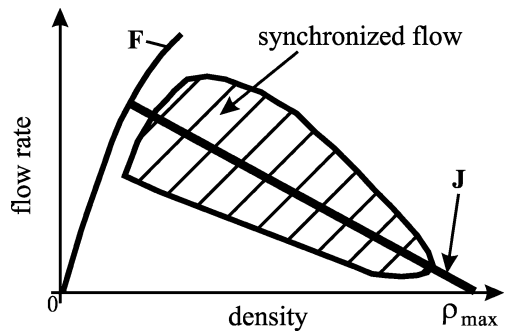

Traffic Congestion, Modeling Approaches to, Figure 19
Three-phase traffic theory in the flow-density plane [49,50,51]: Free flow (F), steady states of synchronized flow (2D *dashed region*) taken from Fig. 15a, and the line *J*

Explanations of S → J Transitions Through Three-Phase Traffic Theory

Empirical S → J transitions are explained in three-phase traffic theory by the following hypotheses [49,50,51]:

(a) All infinite steady states of traffic flow in the flow-density plane that lie on the line *J* are threshold states for wide moving jam existence and emergence (Fig. 19).
(b) The line *J* intersects the 2D-region of steady states of synchronized flow in the flow-density plane, i. e., there are synchronized flow steady states below and above the line *J* (Fig. 19).
(c) The line *J* separates all steady states of synchronized flow in the flow-density plane into two different classes. All states below the line *J* are stable with respect to wide moving jam emergence, i. e., no S → J transitions are possible within these states. All states on and above the line *J* are metastable states with respect to wide moving jam existence and emergence (Fig. 19).
(d) At a given synchronized flow speed, the greater the density in synchronized flow states that are above the line *J*, the greater the probability of an S → J transition.

For a more detailed explanation of S → J transitions (Fig. 20), let us assume that a state of synchronized flow directly upstream of a wide moving jam is associated with a point *k* in the flow-density plane. This point is below the line *J* (Fig. 20a, b). Because the velocity of the upstream front of the wide moving jam $v_g^{(up)}$ equals the slope of the line *K* (from a point *k* in free flow to the point $(\rho_{max}, 0)$), the absolute value $|v_g^{(up)}|$ is always less than that of the downstream front $|v_g|$ determined by the slope of the line *J*, i. e., the formula $|v_g^{(up)}| < |v_g|$ is valid. Therefore, the width of the wide moving jam gradually decreases and the jam dissolves. This means that no wide moving jams can persist continuously, i. e., all steady states of synchro-

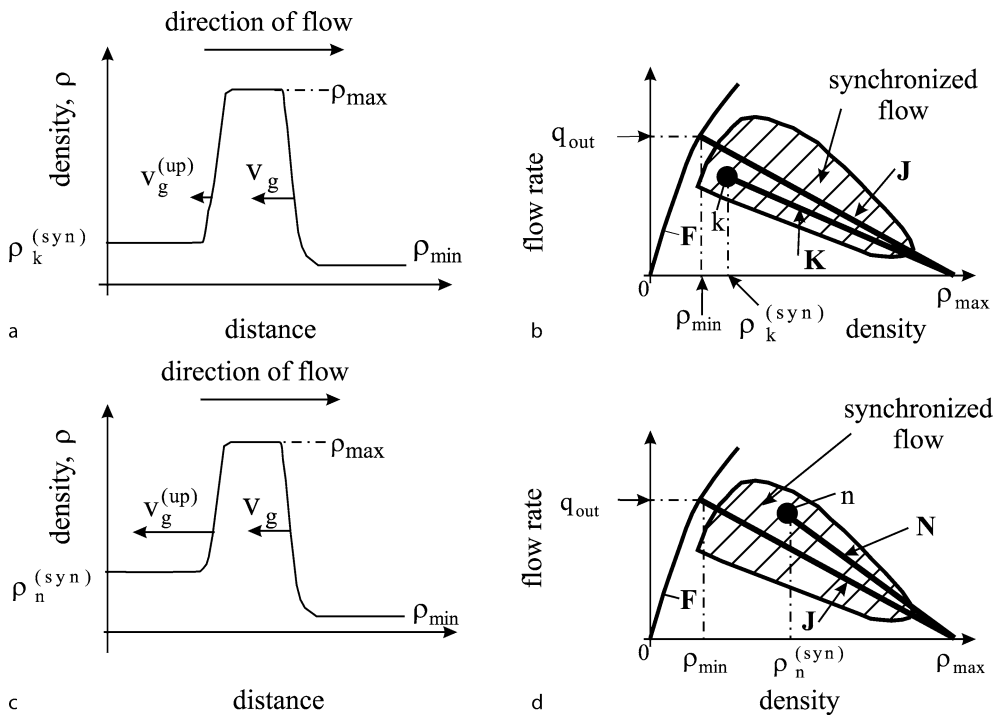

Traffic Congestion, Modeling Approaches to, Figure 20

Explanations of moving jam emergence in synchronized flow in three-phase traffic theory [49,50,51]: a,c Qualitative forms of wide moving jams at two different densities in synchronized flow upstream of the jams, $\rho^{(syn)} = \rho_k^{(syn)}$ (a) and $\rho^{(syn)} = \rho_n^{(syn)}$ (c). b,d Representation of the line *J* and upstream wide moving jam fronts (*lines K* (b) *and N* (d)) in the flow-density plane for the wide moving jams in a and c, respectively. In b,d states for free flow (*curve F*), steady states of synchronized flow (*dashed region*) are taken from Fig. 15a

nized flow below the line *J* are stable with respect to wide moving jam emergence.

In contrast, assume that a state of synchronized flow upstream of another wide moving jam is associated with a point *n* in the flow-density plane. This state is above the line *J* (Fig. 20c, d). In this case, the velocity of the upstream front of the wide moving jam $v_g^{(up)}$ equals the slope of the line *N* (from a point *n* in synchronized flow to the point $(\rho_{max}, 0)$), i. e., the absolute value $|v_g^{(up)}|$ is always higher than that of the downstream front $|v_g|$, i. e., the formula $|v_g^{(up)}| > |v_g|$ is valid. Therefore, the width of the wide moving jam in Fig. 20c should gradually increase. For these reasons, wide moving jams can be formed in states of synchronized flow that lie on or above the line J: these states are metastable with respect to S → J transitions, i. e., an S → J transition is a first-order phase transition. This analysis explains the hypothesis (c).

In synchronized flow states that are on the line *J*, the velocities of the downstream and upstream fronts of a wide moving jam are equal to each other; this explains the hypothesis (a) that these states are threshold ones for S → J transitions (see for more detail Sect. 6.3.1 in [52]).

Three-phase traffic theory explains S → J transitions in the metastable synchronized flow states by a competition between the speed adaptation effect (Sect. "Traffic Breakdown Explanation Through Fundamental Hypothesis of Three-Phase Traffic Theory") and over-deceleration. Over-deceleration introduced by Herman et al. [43] (Sect. "Classic General Motors (GM) Model Approach: Free Flow Instability due to Driver Reaction Time") occurs due to a finite driver reaction time: if the preceding vehicle begins to decelerate unexpectedly, then owing to the reaction time the vehicle starts deceleration with a delay; as a result, if the time delay is long enough the driver decelerates stronger than it is needed to avoid collisions. In earlier traffic flow models based on the fundamental diagram hypothesis, over-deceleration causes an instability of free flow beginning at a critical density [34,43]; the instability should explain traffic breakdown. As stressed in Sect. "Traffic Breakdown Explanation Through Fundamental Hypothesis of Three-Phase Traffic Theory", in

three-phase traffic theory rather than this instability associated with over-deceleration, a competition between the speed adaptation effect and over-acceleration explains empirical features of traffic breakdown.

To explain the competition between the speed adaptation effect and over-deceleration, we assume that a vehicle moving in a metastable synchronized flow state that is above the line *J* in Fig. 20d decelerates unexpectedly. If the following vehicle cannot pass this preceding vehicle, there are two possibilities: (i) The vehicle decelerates and due to the speed adaptation effect is able to adapt the speed to the speed of the preceding vehicle. Then rather than an S → J transition, a new synchronized flow state of lower speed is formed. (ii) Due to over-deceleration, the vehicle decelerates stronger than it would be needed for speed adaptation, i. e., its speed becomes lower than the speed of the preceding vehicle. If each of the following vehicles decelerates also to lower speed than the associated preceding vehicle, then finally the speed upstream decreases up to zero leading to an S → J transition. Thus if a local speed disturbance with a speed decrease occurs in synchronized flow, there is a tendency towards another state of synchronized flow associated with the vehicle adaptation effect. In contrast, due to over-deceleration there is also a tendency towards a wide moving jam. The greater the density, the more probable over-deceleration in comparison with speed adaptation. This explains the hypothesis (d) above.

Thus synchronized flow associated with a point in the flow-density plane that lies on or above the line *J* (Fig. 20d) is in a metastable state with respect to wide moving jam emergence. For this reason, there can be speed disturbances (fluctuations) within the synchronized flow whose growth lead to wide moving jam emergence. In accordance with definitions of phase transitions in complex systems, such wide moving jam emergence is called *spontaneous* wide moving jam emergence in synchronized flow (S → J transition). This explains the term *spontaneous* S → J transition used above [52].

A speed disturbance begins spontaneously to grow in a metastable state of synchronized flow [52], if within the initial speed disturbance the speed is equal or lower than a critical speed for an S → J transition; otherwise the disturbance decays over time. A critical speed disturbance associated with the critical speed within the disturbance can be considered a nucleus for the S → J transition.

There can be various sources for a speed disturbance whose occurrence and subsequent growth in metastable synchronized flow leads to a spontaneous S → J transition [52]: unexpected braking of a vehicle in synchronized flow, lane changing within synchronized flow on the main road, fluctuations in flow rates, vehicle merging from other roads (e. g., at bottlenecks), etc. Over time, a growing speed disturbance takes the form of a narrow moving jam (Sect. "Pinch Effect"), i. e., the growing speed disturbance of a great enough amplitude is synonym of a growing narrow moving jam in the pinch region of synchronized flow.

Note that lane changing as a source for the spontaneous occurrence of growing narrow moving jams in synchronized flow has recently been empirically studied in [1]. In numerical simulations of GP emergence at isolated bottlenecks based on three-phase traffic flow models, it has been found [62] that vehicle merging from other roads and lane changing in a neighborhood a bottleneck are the main sources for speed disturbances whose subsequent growth in metastable synchronized flow leads to spontaneous S → J transitions (see also simulation results of [54] associated with the effect of lane changing on spontaneous S → J transitions upstream of an off-ramp bottleneck).

It should be noted that in metastable synchronized flow there can also be an *induced* S → J transition. In contrast with a spontaneous S → J transition, induced wide moving jam emergence in the synchronized flow is caused by the upstream propagation of a moving jam, which has initially occurred within a *different* link of the road network connected with the road under consideration.

Comparison of F → S and S → J Transitions

Both a spontaneous F → S transition (Sect. "Spontaneous Traffic Breakdown (F → S Transition)") and a spontaneous S → J transition are first-order phase transitions that occur in metastable states of free flow (F) and synchronized flow (S), respectively. In both cases, critical speed disturbances should appear whose subsequent growth lead to the related phase transition. However, there is a qualitative difference between these two phase transitions. This difference is associated with the kinetics of the growth of the critical disturbances.

When a critical speed disturbance appears in free flow at a bottleneck, the growth of this disturbance and resulting traffic breakdown occurs usually also at the bottleneck. This is explained by a permanent speed disturbance, which is on average motionless and localized in free flow at the bottleneck. This disturbance increases probability of synchronized flow emergence, i. e., breakdown probability at the bottleneck considerably.

In contrast, a growing narrow moving jam in metastable synchronized flow jam propagates upstream. It takes some time delay before the growth of the jam leads to wide moving jam, i. e., an S → J transition occurs. For this reason, the S → J transition occurs upstream of the

road location at which a critical speed disturbance has initially appeared and begun to grow in synchronized flow. Thus, even if a critical speed disturbance has appeared in metastable synchronized flow at the location of a bottleneck, the S → J transition resulting from the growth of the disturbance occurs upstream of this bottleneck. The exclusion can be a very heavy bottleneck, which limits the flow rate in congested traffic considerably. A theory of traffic congestion at heavy bottlenecks appears in ► Traffic Congestion, Spatiotemporal Features of.

Note that there can be also nucleation-interruption effects, which play an important role for both spontaneous F → S and S → J transitions. A consideration of these effects is, however, out of the scope of the article (see Sects. 6.5.3 and 10.2 in [52]).

Empirical Double *Z*-Characteristic for Phase Transitions in Traffic Flow

As mentioned above, F → S and S → J transitions are first-order phase transitions. Both of these first-order phase transitions can be illustrated by a double *Z*-characteristic for the phase transitions. An empirical double *Z*-characteristic for the phase transitions, which is measured during GP formation at an on-ramp bottleneck (Fig. 21a), is presented in Fig. 21b.

Firstly, an F → S transition occurs at the bottleneck (arrows labeled by F → S in Figs. 21a,b). Synchronized flow propagates upstream. Later, in the synchronized flow the pinch effect is realized upstream of the bottleneck. In the pinch region of synchronized flow, narrow moving jams emerge. The jams propagate upstream growing in the jam amplitude. Some of these narrow moving jams transform into wide moving jams, i. e., S → J transitions occur (one of these transitions at approximately $x = 3.2$ km is labeled by arrow S → J in Fig. 21a). Thus S → J transitions occur later and upstream of the freeway location of the F → S transition. Within wide moving jams low speed states are observed that are probably associated with low speed within wide moving jams associated with so-called "moving blanks" that are discussed in ► Traffic Congestion, Spatiotemporal Features of.

Later the GP begins to dissolve. In particular, some of wide moving jams associated with the GP dissolve and either synchronized flow or free flow occur (J → S or J → F transitions). An example of an J → F transition associated with a wide moving jam that begins to dissolve between locations 1.5 and 0 km is shown in Fig. 21. Finally, an S → F transition occurs at the bottleneck and the GP disappears (arrows labeled S → F in Fig. 21).

We see that there are many diverse transitions between the three traffic phases as well as between diverse traffic states within the phases associated with congested pattern emergence and dissolution. The phase transitions and transitions within the same traffic phase could exhibit a variety of hysteresis effects. The phase transitions in traffic and transitions within the same traffic phase cannot be distinguished each other without knowledge of the whole *spatiotemporal dynamics* of the congested pattern (Fig. 21a).

As discussed in Sect. "Traffic Breakdown Explanation Through Fundamental Hypothesis of Three-Phase Traffic Theory", for F → S transition occurrence a local speed disturbance should occur in free flow at the bottleneck with the speed within the disturbance that is equal or lower than a critical speed. The greater the density (flow rate) downstream of the bottleneck, the higher this critical speed. Respectively, for S → J transition occurrence, a local speed disturbance should appear in synchronized flow upstream of the bottleneck with the speed within the disturbance that is equal or lower than another critical speed. For this reason, in a theoretical double *Z*-characteristic (Fig. 22) between states of free flow, synchronized flow, and low speed states within wide moving jams there are states associated with the critical speed disturbances (nuclei for

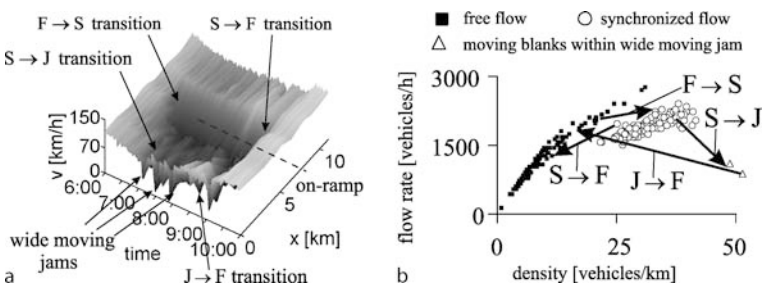

Traffic Congestion, Modeling Approaches to, Figure 21
Average speed in empirical GP at on-ramp bottleneck (a) and related empirical double *Z*-characteristic for phase transitions (b). Low speed states within wide moving jam are labeled as "moving blanks" in b

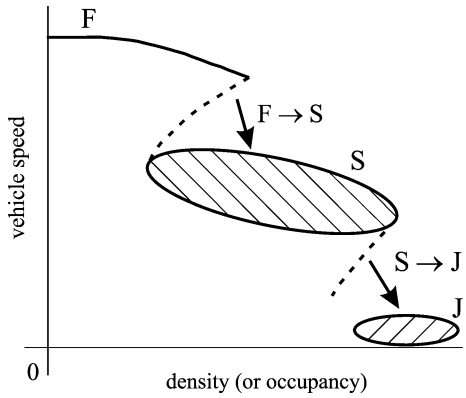

Traffic Congestion, Modeling Approaches to, Figure 22
Qualitative double *Z*-characteristic for F → S → J transitions at bottlenecks that includes states of free flow (F), synchronized flow (S) and low speed states (J) associated with moving blanks within wide moving jams [52]

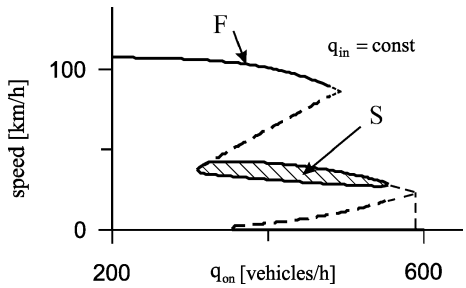

Traffic Congestion, Modeling Approaches to, Figure 23
Simulated double *Z*-characteristic for F → S → J transitions at on-ramp bottlenecks. Taken from [52]

phase transitions) needed for phase transition occurrence between the related traffic phases.

It should be noted that in addition to the critical speed (density) within a critical disturbance, the critical disturbance is also characterized by a *critical spatial shape*, i. e., by the critical speed (density) spatial distribution within the disturbance. This conclusion of three-phase traffic theory is valid for F → S and S → J transitions. However, for simplicity we neglect the critical spatial shape of the disturbances in Fig. 22, in which critical speed disturbances (nuclei for the phase transitions) are qualitatively presented by dashed curves between states *F* and *S* associated with critical speeds within the nuclei for F → S transitions and between states *S* and *J* associated with critical speeds within the nuclei for S → J transitions.

The double *Z*-characteristic features can also be seen on simulated double *Z*-characteristic for the F → S → J transitions shown in Fig. 23. For simplicity, it has been suggested that the speed within wide moving jams is equal to zero, i. e., states *J* shown in Fig. 22 have been neglected.

As mentioned, the states *J* are associated with moving blanks within the jams (Fig. 21b) [52]; empirical and simulation results associated with moving blanks are discussed in more detail in ► Traffic Congestion, Spatiotemporal Features of.

Hypotheses of Three-Phase Traffic Theory as the Result of Traffic Phase Definitions

In Sects. "Free and Congested Traffic"–"Empirical Double *Z*-Characteristic for Phase Transitions in Traffic Flow", we have shown that empirical features of traffic breakdown and congested pattern formation are explained in the framework of three-phase traffic theory. Here we explain that and why hypotheses of this theory [49,50,51] are the *result* of the traffic phase definitions [J] and [S]:

1. The definition of wide moving jam [J] determines the line *J* in the flow-density plane. Any point on the line *J* can be a final steady traffic state for vehicles accelerating at the downstream front of a wide moving jam. If the speed in this state is lower than the minimum speed in free flow, the state is a synchronized flow steady state that lies on the line *J*. Thus there should be infinite synchronized flow states lying on the line *J* (Fig. 20b, d).
2. The definition of synchronized flow [S] means that downstream fronts of synchronized flow regions do *not* exhibit the jam characteristic feature [J]. For this reason, in addition to synchronized flow steady states lying on the line *J* (item 1) there should be synchronized flow steady states that are away from the line *J* in the flow-density plane. Indeed, only in this case the downstream fronts between different states of synchronized flow do not exhibit the characteristic jam velocity given by the slope of the line *J*. Thus there should be also synchronized flow steady states away from the line *J*, i. e., that there is a 2D-region of the steady states in the flow-density plane. This explains the fundamental hypothesis of three-phase traffic theory (Fig. 15).
3. The hypothesis of three-phase traffic theory that an F → S transition is a first-order phase transition results from the 2D-region of synchronized flow steady states as already explained in Sect. "Traffic Breakdown Explanation Through Fundamental Hypothesis of Three-Phase Traffic Theory".
4. The definition [S] is associated with empirical results for moving jam behavior within synchronized flow [52]: depending on synchronized flow characteristics upstream of a moving jam, the jam can grow or dissolve propagating in the synchronized flow over time. We have noted (item 1) that there are infinite synchronized flow steady states that lie on the line *J*.

If these states were to build the upper boundary of the 2D-region of the synchronized flow steady states, i. e., there were no steady states above the line *J*, then, as explained in Sect. "Explanations of S → J Transitions Through Three-Phase Traffic Theory", all wide moving jams should dissolve propagating in the states of synchronized flow. This contradicts with mentioned empirical results about possible jam growth. In contrast, if the states on the line *J* were to build the low boundary of the 2D-region of the synchronized flow steady states, i. e., there were no steady states below the line *J*, then no wide moving jam can dissolve propagating in the states of synchronized flow. This also contradicts with mentioned empirical results about possible jam dissolution. Thus we should assume that there are synchronized flow steady states above *and* below the line *J*. This explains the hypothesis (b) of Sect. "Explanations of S → J Transitions Through Three-Phase Traffic Theory".

5. The hypotheses (a) and (c) of Sect. "Explanations of S → J Transitions Through Three-Phase Traffic Theory" about S → J transitions that result from item 2 and 4 have already been proven in Sect. "Explanations of S → J Transitions Through Three-Phase Traffic Theory".

Critical Discussion of Fundamental Diagram Modeling Approach to Traffic Congestion

In this section, we will explain and give reasons why all earlier traffic flow theories and models cannot show and explain *empirical* traffic breakdown discussed above.

Validation of Theoretical Fundamental Diagram for Traffic Flow

It has been mentioned in the introduction that the fundamental diagram (Fig. 1a) is the basis hypothesis for all earlier traffic flow theories models reviewed in [16,18,21,33,35,37,40,84,92,93,95,98,102,126,127,128]. However, empirical observations show that whereas fundamental diagrams for free flow are qualitatively the same at different freeway locations, in contrast, the shape of empirical fundamental diagrams for congested traffic depends qualitatively on the location at which data measurements have been performed within a congested pattern. An example for different fundamental diagram types associated with a GP at an on-ramp bottleneck is shown in Fig. 24 (see explanations in Chap. 15 in [52]).

This means that in a general case the validation of a theoretical fundamental diagram of a traffic flow model, which is very often used by researchers as an empirical basis for the fundamental diagram, cannot be performed. Depending on a road location various empirical fundamental diagram types exist. Moreover, if data, which are measured at many different locations within congested traffic, is used for averaging to one empirical fundamental diagram (this is often a case in the literature), then there is almost no possibility to find a relation of such a fundamental diagram to spatiotemporal features of real traffic patterns.

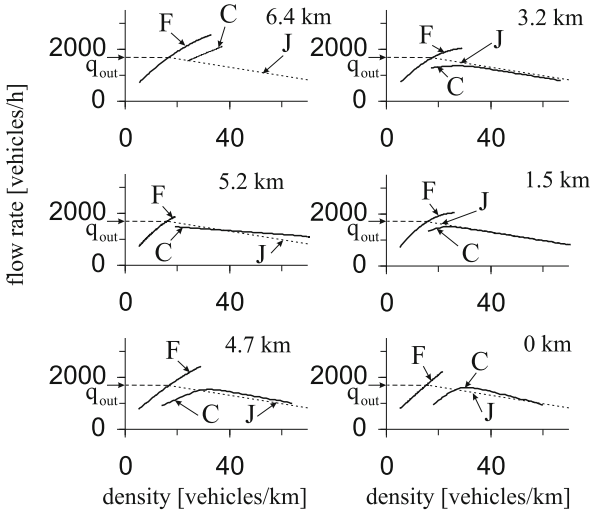

Traffic Congestion, Modeling Approaches to, Figure 24
Dependence of empirical fundamental diagram types for congested traffic at different locations within the GP shown in Fig. 21a. F free flow. C congested traffic. J line *J* (*dashed line*). On-ramp bottleneck is at the location about 6.4 km. See explanations of these fundamental diagram types in Chap. 15 in [52]

Thus there cannot be a choice of the theoretical fundamental diagram, which can alone give a correct relation between model solutions for spatiotemporal congested patterns and real measured spatiotemporal congested patterns. However, such a relation is very often suggested in a huge number of traffic flow models and theories within the fundamental diagram approach reviewed in [16,18,21,33,35,37,40,84,92,93,95,98,126,127,128] and briefly discussed in the reminder of this section. This may be one of the main reasons for failure of this well-accepted and widely used approach to traffic flow theory and modeling in describing of traffic breakdown and resulting spatiotemporal traffic congested patterns.

Classic Lighthill–Whitham–Richards (LWR) Theory of Onset of Traffic Congestion

The basic idea of the classic LWR traffic flow theory is as follows: the maximum flow rate q_0 associated with

the maximum point (ρ_0, q_0) at the fundamental diagram (Fig. 1a) determines free flow capacity at a bottleneck (often called "bottleneck capacity") denoted by q_{cap}:

$$q_{\text{cap}} = q_0 \,. \tag{15}$$

It is further suggested that due to a road non-homogeneity introduced by the bottleneck the maximum flow rate q_0 downstream of the bottleneck, which determines the capacity (15), is smaller than the capacity of a homogeneous road upstream of the bottleneck. For this reason, if the sum of the flow rates upstream of the bottleneck reaches the maximum flow rate q_0 at the fundamental diagram (Fig. 1a), i. e., it reaches the bottleneck capacity (15), then a further increase in the flow rates must lead to congestion (queue) formation and upstream congestion propagation upstream of the bottleneck.

These hypotheses of the LWR theory mean that congested traffic occurs only then, when the upstream flow rate exceeds the bottleneck capacity q_{cap} determined by (15). This conclusion of the LWR traffic flow theory about the reason for the onset of congestion is very often used as the definition of congested traffic. As we will see below, this hypothesis of the LWR theory about the onset of congestion as well as the associated congested traffic definition are in a very deep contradiction with empirical results.

In accordance with the LWR theory, traffic flow phenomena should be explained based on the law of conservation of the number of vehicles on the road

$$\frac{\partial \rho(x,t)}{\partial t} + \frac{\partial Q(\rho(x,t))}{\partial x} = 0 \,, \tag{16}$$

in which is assumed that there is a relationship between the flow rate Q and density ρ

$$Q = Q(\rho) \tag{17}$$

associated with the fundamental diagram for traffic flow. Here, x is a spatial coordinate in the direction of traffic flow, and t is time. The LWR-model (16), (17) has discontinuous solutions in the form of shock-waves (Fig. 25) with a shock-wave velocity

$$v_{\text{s}} = \frac{Q(\rho_2) - Q(\rho_1)}{\rho_2 - \rho_1} \,, \tag{18}$$

where the flow rates $Q(\rho_2)$, $Q(\rho_1)$ associated with (17) correspond to points in the flow-density plane that lie on the fundamental diagram; ρ_1 and ρ_2 are the densities upstream and downstream of the shock wave, respectively (Fig. 25). The shock wave can be represented in the flow-density plane by the line labeled "shock" whose slope is equal to the shock velocity v_{s} (Fig. 25b).

Because the LWR-model (16), (17) has discontinuous solutions in the form of shock waves, its numerical simulations are usually based on one of finite difference approximation methods for partial differential equations associated with the Godunov family methods. One of such discrete versions of the LWR-model (16) that is consistent with the LWR-hydrodynamic theory is the cell-transmission model of Daganzo [20].

In Daganzo's cell transmission model, a rectangular lattice with time spacing τ and space length (cell length) Δx is overlaid on the time-space plane; the x-coordinates represent the center of the cells into which a road has been discretized; t-coordinates are the times at which the cell vehicle densities are evaluated. Then the Daganzo model [20] for a single-lane road of length L with an on-ramp bottleneck can be written as follows:

$$\begin{aligned}\rho(x,t+\tau) = \rho(x,t) - \Big[& q\left(x + \frac{\Delta x}{2}, t + \frac{\tau}{2}\right) \\ & - q\left(x - \frac{\Delta x}{2}, t + \frac{\tau}{2}\right)\Big]\left(\frac{\tau}{\Delta x}\right) \\ & + q_{\text{on}}(t) f(x) \left(\frac{\tau}{\Delta x}\right) \,,\end{aligned} \tag{19}$$

$$\begin{aligned} & q\left(x + \frac{\Delta x}{2}, t + \frac{\tau}{2}\right) \\ & \qquad = \min(D(\rho(x,t)), R(\rho(x+\Delta x, t))) \,, \end{aligned} \tag{20}$$

$$f(x) = \begin{cases} \frac{1}{N_{\text{m}}} & \text{if } x_{\text{on}} \le x < x_{\text{on}} + L_{\text{m}} \,, \\ 0 & \text{otherwise,} \end{cases} \tag{21}$$

with boundary conditions

$$\rho(0,t) = \rho(\Delta x, t) \,, \quad \rho(L,t) = \rho(L - \Delta x, t) \,, \tag{22}$$

and initial conditions

$$\rho(x,0) = \rho_{\text{in}} \,, \tag{23}$$

where $R(\rho)$ and $D(\rho)$ are non-increasing so-called *receiving* and non-decreasing so-called *sending* curves, respectively; these two monotonic curves $R(\rho)$ and $D(\rho)$ take values in the interval $[0, q_0]$, as shown in Fig. 26a: $R(\rho) = Q(\rho)$ at $\rho \ge \rho_0$; $D(\rho) = Q(\rho)$ at $\rho \le \rho_0$; $x = x_{\text{on}}$ is the location of the farthest upstream cell of the on-ramp merging region of length $L_{\text{m}} = N_{\text{m}} \Delta x$ that includes the integer number N_{m} of cells (Fig. 26b); the flow rate $q_{\text{on}}(t)$ is zero at $t < t_0$ and $q_{\text{on}}(t) = q_{\text{on}}$ is constant at $t \ge t_0$; ρ_{in} is the density in an initial homogeneous free flow on the road, $q_{\text{in}} = Q(\rho_{\text{in}})$ is the corresponding flow rate.

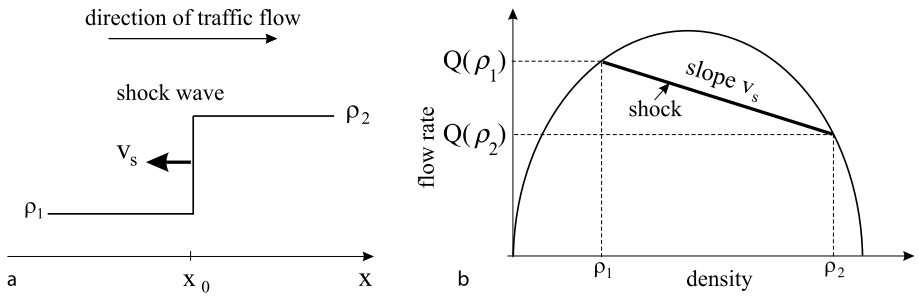

Traffic Congestion, Modeling Approaches to, Figure 25
Qualitative representation of a shock wave in space (a) and in the fundamental diagram with the line labeled "shock" whose slope is equal to the shock velocity v_s (b) [86]

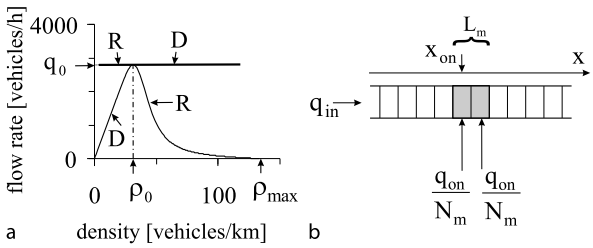

Traffic Congestion, Modeling Approaches to, Figure 26
Parameters of the Daganzo cell-transmission model (19)–(23) used for numerical simulations of traffic breakdown in Fig. 27: a Receiving (R) and sending (D) curves. b Model of on-ramp bottleneck. $Q(\rho) = \rho V(\rho)$, $V(\rho) = V_g(g)$, where $g = \rho^{-1} - \rho_{max}^{-1}$, $V_g(g) = V_0\big(\tanh((g - g_2)/g_1) + \tanh(g_2/g_1)\big)$ at $V_0 = 14$ m/s, $g_2 = 17$ m, $g_1 = 7$ m, $\tau = 0.1$ s, $\Delta x = 10$ m, $L = 20$ km, $N_m = 2$, $t_0 = 7$ min, $x_{on} = 16$ km, $q_0 = 2795$ and $q_{in} = 2676$ vehicles/h

Each of the simulations of congested traffic patterns with the Daganzo model (19)–(23) (Fig. 27) is performed during a limited time interval within which none of the patterns reach the road boundaries, i. e., free flow conditions remain at the boundaries. The simulations show common qualitative features of traffic breakdown and resulting congested traffic patterns that show the LWR theory at an isolated on-ramp bottleneck [62].

The onset of traffic congestion and resulting congested patterns within the LWR-theory are studied by an increase in the flow rate to the on-ramp q_{on}, while the flow rate in free flow upstream of the on-ramp bottleneck q_{in} is constant. Numerical simulations are made for various given values q_{on}. The difference in these values of q_{on} chosen for different simulations is small enough to ensure that all possible various congested patterns and their dynamic behavior characteristics can be found and studied.

If $q_{on} = 0$, the speed and density are spatially homogeneous. When the flow rate q_{on} increases beginning from zero (points 1–3 in Fig. 27a), the flow rate downstream of the bottleneck $q_{sum} = q_{on} + q_{in}$ increases too as long as the condition

$$q_{sum} = q_{on} + q_{in} < q_0 \tag{24}$$

is satisfied. In this case, a non-homogeneous motionless structure of traffic flow appears at the bottleneck related to different free flows downstream and upstream of the bottleneck (curves I in Fig. 27b, h). Speed decreases and density increases downstream of the bottleneck.

The differences in the speed and density upstream and downstream within this motionless free flow structure increase when q_{on} increases. However, this has a limit at a given large enough flow rate q_{in}. This limit is reached when the flow rate q_{on} reaches some critical value $q_{on} = q_{on}^{(d)}$ at which the flow rate q_{sum} is equal to the maximum flow rate on the fundamental diagram:

$$q_{sum} = q_{in} + q_{on}^{(d)} = q_0 \,. \tag{25}$$

When the flow rate q_{on} increases further, i. e.,

$$\Delta q = q_{in} + q_{on} - q_0 > 0 \,, \tag{26}$$

then the upstream front of the initial free flow structure, which is motionless at the condition $q_{sum} = q_{in} + q_{on} \leq q_0$ (curve I in Fig. 27b), begins to move upstream of the bottleneck, i. e., a wave (shock wave) of lower speed and greater density propagating upstream appears (Fig. 27c, d). As a result, a dense flow associated with the branch of the diagram with a negative slope occurs upstream of the bottleneck (points 4–6 in Fig. 27e, f).

At the critical point (25), the flow rates on the main road downstream and upstream of the bottleneck are continuous functions of q_{on} (points 3, 4 in Fig. 27a, f). In contrast with the flow rates, under condition (26) already at $\Delta q \to 0$ the minimum average speed v_{min} within the dense flow (point 4 in Fig. 27f) is lower than the speed

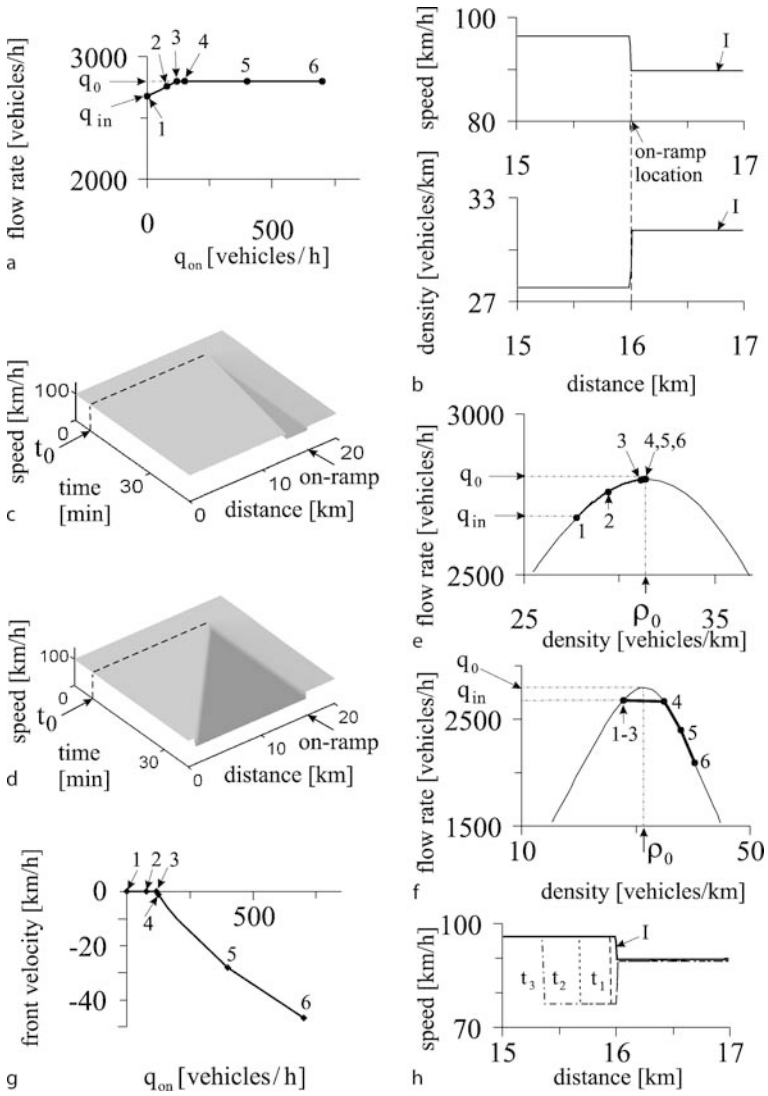

Traffic Congestion, Modeling Approaches to, Figure 27

Traffic breakdown in the LWR-theory [62]: a Flow rate downstream of bottleneck as a function of q_{on}. b Non-homogeneous motionless solution for free flow under condition (24). c,d Propagation of shock-waves under condition (26) for $q_{on} = 160$ (c) and 400 vehicles/h (d). e,f On-ramp flow rate dependences of the flow rate and density on main road downstream (e) and upstream of bottleneck (f). g $v_s(q_{on})$. h Shock-wave propagation associated with point 4 in a,e–g. Simulations of the Daganzo model (19)–(23) with parameters given in caption of Fig. 26

upstream of the dense flow (point 3 in Fig. 27f); however, if we choose a greater q_{in}, then this speed difference decreases and it tends to zero, when $q_{in} \to q_0$ (but $q_{in} < q_0$).

Regardless of the choice in q_{in} (but $q_{in} < q_0$), the absolute value of the shock-wave velocity $|v_s|$ increases continuously beginning from *zero*, when q_{on} first reaches and then exceeds the critical flow rate $q_{on}^{(d)}$ associated with (25). We find that if $\Delta q \to 0$, then $|v_s| \to 0$. For example, at $\Delta q = 3$ vehicles/h ($q_{on} = 120$ vehicles/h), the shock wave, within which the vehicle speed is associated with point 4 in Fig. 27f, has the velocity $v_s \approx -0.7$ km/h only; this very slow shock wave propagation can be seen in Fig. 27h in which speed distributions within the wave related to the times $t_1 = 25$, $t_2 = 55$ and $t_3 = 90$ min are shown.

The greater q_{on}, specifically, the greater Δq (26), the greater the absolute value of the shock-wave velocity $|v_s|$ (Fig. 27g). In accordance with (18), in numerical simula-

tions for each of the flow rates q_{on} (points 4–6 in Fig. 27a) the front velocity v_s is determined by a slope of the line between the point related to the flow rate q_{in} in free flow upstream of the wave front (shock) (points 1–3 in Fig. 27f) and the related point at the fundamental diagram for congested traffic associated with the flow rate within dense flow downstream of the wave front (points 4–6 in Fig. 27f). In addition, the flow rate downstream of the bottleneck, which is equal to q_0 under the condition (25), remains approximately to be equal to q_0, when q_{on} increases (points 4–6 in Fig. 27e).

The congested patterns in Fig. 27c, d at first glance resemble a widening SP. Indeed, in both cases a dense flow occurs upstream of the bottleneck whose downstream front is fixed at the bottleneck. Thus, this dense flow should satisfy the macroscopic spatiotemporal objective criteria for the synchronized flow phase [S] (Sect. "Introduction"). However, the criterion [S] defines the synchronized flow phase in *empirical* observations. This empirical synchronized flow occurs due to a first-order F → S transition. This F → S transition found in three-phase traffic theory explains the *fundamental empirical features* of traffic breakdown and bottleneck capacity (Sect. "Highway Capacity and First-Order F → S Transition"). In contrast, the LWR-theory cannot explain these features.

Moreover, the classic formula (18) of the LWR-theory for shock waves, which is the fundamental formula of traffic science (see e. g., the books [21,84,93,126]), is not consistent with empirical features of synchronized flow; therefore, the LWR-formula (18) cannot be applied for an adequate calculation of front (shock) velocities of SPs as well as of the front velocities of wide moving jams when the jams propagate through synchronized flow.

To explain these critical conclusions, we summarize below features of congestion emergence in the LWR-theory, which are in deep contradiction with empirical results:

1. In the LWR-theory, there is *no* induced traffic breakdown at a bottleneck observed in empirical observations.
2. In the LWR-theory, there is *no* hysteresis effect, which is observed by congested pattern emergence and dissolution in empirical observations.
3. In the LWR-theory there is *no* discontinuous change in the velocity v_s when due to an increase in q_{on} the condition (26) is satisfied: $|v_s|$ increases continuously beginning from zero, when q_{on} first reaches and then exceeds the critical flow rate $q_{on}^{(d)}$ associated with the condition (25). This is qualitatively different from those for the upstream front of congested patterns at the bottleneck in empirical data. In the latter case, when the flow rate q_{on} increases and an F → S transition occurs, a wave of synchronized flow occurs abruptly and propagates upstream with a *finite* velocity. This is associated with a first-order F → S transition, which cannot be found and explained in the LWR-theory.
4. In empirical observations, the flow rate downstream of the bottleneck at which breakdown occurs in a realization (day) is often smaller than flow rates observed in free flow downstream of the bottleneck before breakdown occurrence (Fig. 12). This cannot be explained by the LWR-theory even if one suggests that the capacity $q_{cap} = q_0$ (15) is a probabilistic value, which can be different in different realizations (days) at the same bottleneck (for a more detailed discussion of capacity definitions, see Sect. "Critical Discussion of Highway Capacity Definitions"). This is because for a given realization, within the framework of the LWR-theory there is a fixed bottleneck capacity at which traffic breakdown occurs. This contradicts with the above observations (Fig. 12).
5. In empirical data, after traffic breakdown has occurred at a bottleneck, wide moving jams can emerge spontaneously in synchronized flow upstream of the bottleneck, i. e., a GP emerges spontaneously. Regardless of the density, *no* spontaneous moving jam emergence can be found in the LWR-theory.
6. In empirical data, when a wide moving jam propagates through synchronized flow, there is no single relation between the density and flow rate in the synchronized flow. For this reason, in general case the classic LWR-formula for shock wave velocity (18) cannot be applied for a correct calculation of the jam front velocity while the jam propagates through synchronized flow.

Thus, both empirical features of traffic breakdown and resulting congested pattern formation discussed above are in a very deep contradiction with the LWR-theory. For these reasons, the LWR-theory and the associated traffic flow models (like cell-transmission models) cannot be used for reliable analysis of freeway traffic control and management strategies.

One of the reasons for drawbacks of the LWR-theory in explanation of empirical traffic breakdown is a single correspondence between the density and flow rate (17), i. e., the fundamental diagram for traffic flow used in this theory. In contrast, empirical measurements of synchronized flow show that there is no single correspondence between the density and flow rate in the synchronized flow. Thus instead of the LWR-formula (18), one should use the classic Stokes-formula for a shock-wave velocity derived

by Stokes in 1848 [114]:

$$v_s = \frac{q_2 - q_1}{\rho_2 - \rho_1}\,, \tag{27}$$

where ρ_1, q_1 and ρ_2, q_2 are the density and flow rate upstream and downstream of the shock wave, respectively. The fundamental difference between the LWR-formula (18) and the Stokes-formula (27) is that there is no given relationship between the density and flow rate in (27).

For these reasons, rather than the LWR-formula (18) in the model ASDA for tracking and prediction of wide moving jam propagation, which is successfully used in on-line applications on German freeways since 2000, the classic Stokes-formula (27) has been used. In the model ASDA, the density and flow rate ρ_1, q_1 upstream and ρ_2, q_2 downstream of the wide moving jam fronts in (27) are found based on measurements of these traffic flow variables (see a detailed consideration of the model ASDA and its application made in Chaps. 21 and 22 in [52]; for a brief review see ► Traffic Prediction of Congested Patterns).

Classic General Motors (GM) Model Approach: Free Flow Instability due to Driver Reaction Time

As long ago as 1958–1961, Herman, Montroll, Potts, Gazis, Rothery [34,43] and Komentani and Sasaki [73,74] suggested that traffic breakdown is associated with an instability of free flow. This instability is related to a finite reaction time of drivers. This driver reaction time is responsible for the over-deceleration effect discussed in Sect. "Introduction"; if the preceding vehicle begins to decelerate unexpectedly, then owing to the finite driver reaction time the following vehicle starts deceleration with a delay; as a result, if the time delay is long enough, the driver decelerates stronger than it is needed to avoid collisions. In this case, the driver speed can become lower than the speed of the preceding vehicle. If the over-deceleration effect is realized for next following drivers, the wave of vehicle speed reduction appears and increases in amplitude over time in traffic flow leading to a stop of some of the vehicles upstream. This instability should occur when the vehicle density on the fundamental diagram exceeds some critical density denoted $\rho_{cr}^{(J)}$.

In the car-following model of Herman, Montroll, Potts, Gazis, Rothery [34,43], which is often called the General Motors (GM) model, the driver reaction time denoted by τ_0 is explicitly used in the vehicle deceleration (acceleration) denoted by $a(t+\tau_0)$, i. e., the vehicle reacts with the time delay τ_0 on any changes in the space gap to the preceding vehicle $g(t)$ and the speed difference $\Delta v(t) = v_\ell(t) - v(t)$ (relative speed) between the speed of the preceding vehicle $v_\ell(t)$ and the vehicle speed $v(t)$. The GM model reads as follows [34]:

$$a(t+\tau_0) = \frac{\Delta v(t)[v(t+\tau_0)]^{m_1}}{T_0[g(t)+d]^{m_2}}\,, \tag{28}$$

$$a(t) = \frac{dv(t)}{dt}\,, \tag{29}$$

where d is the vehicle length; T_0, m_1, m_2 are constants. By integrating Eqs. (28), (29), one gets model solutions for steady states related to the fundamental diagram [34]

$$V(\rho) = V_0\left[1 - (\rho/\rho_{max})^{(m_2-1)}\right]^{(1-m_1)^{-1}}\,, \tag{30}$$

where V_0 is constant, $\rho_{max} = 1/d$.

The idea of the GM-model (28) with a driver reaction time is incorporated either explicitly or implicitly in many other traffic flow models reviewed in [16,18,33,35,37,40,84,92,93,95,98,102,126,127,128].

Newell Optimal Velocity (OV) Model and its Variants

At $m_1 = 0$ the GM-model is associated with the Newell optimal velocity (OV) model [100]

$$v(t+\tau_0) = V(g(t)+d)\,. \tag{31}$$

In this case, the OV-relation $V(g+d)$ and the associated fundamental diagram $Q(\rho) = \rho V(\rho)$ are determined from (31) at time-independent space gap g and speed v:

$$v = V(g+d)\,. \tag{32}$$

In the Bando et al. OV model [3,4,5]

$$\frac{dv(t)}{dt} = \frac{V(g(t)+d) - v(t)}{\tau_0^{(0)}} \tag{33}$$

it is suggested that

$$V(g+d) = 0.5V_0[\tanh(g-g_0) + \tanh(g+d)]\,, \tag{34}$$

$\tau_0^{(0)}$, g_0 are model parameters.

There are many other traffic flow models associated with a generalization of OV-models, which can be written in the form (e. g., [40])

$$\frac{dv(t)}{dt} = \phi(v(t), \Delta v(t), g(t))\,, \tag{35}$$

where the OV-relation $V(g)$ and the associated fundamental diagram $Q(\rho) = \rho V(\rho)$ are determined from (35) at time-independent space gap g and speeds v, v_ℓ, as well as at $v = v_\ell$:

$$\phi(V, 0, g) = 0\,. \tag{36}$$

In the models (35), a driver reaction time is used only implicitly in a traffic flow model. To these model belongs, for example, the Treiber et al. Intelligent Driver Model (IDM) [119] with

$$\phi = c_3\left[1-\left(\frac{v}{v_0}\right)^4-\left(\frac{g^*(v,\Delta v)}{g}\right)^2\right], \tag{37}$$

$$g^*(v,\Delta v) = g_1 + \tau_1 v + \frac{v\Delta v}{2}(a_0 b_0)^{\frac{1}{2}}, \tag{38}$$

where $c_3, g_1, v_0, \tau_1, a_0$ and b_0 are model parameters.

These models exhibit qualitatively the same features of free flow instability and resulting wide moving jam formation as those firstly found in [63] from an analysis of a version of Payne's macroscopic traffic flow model.

Payne-Macroscopic Traffic Flow Model and its Variants The Payne-model reads as follows [103,104]:

$$\frac{\partial\rho(x,t)}{\partial t}+\frac{\partial(\rho(x,t)v(x,t))}{\partial x}=0, \tag{39}$$

$$\frac{\partial v(x,t)}{\partial t}+v\frac{\partial v(x,t)}{\partial x}=\frac{V_0(\rho)-v(x,t)}{\tau^{(0)}}-\frac{c_0^2}{\rho}\frac{\partial\rho(x,t)}{\partial x}, \tag{40}$$

where the speed-density relationship $V_0(\rho)$ determines the fundamental diagram $Q(\rho)=\rho V_0(\rho)$; $\tau^{(0)}$, c_0 are model parameters. To avoid solutions with speed discontinuities, Kühne introduced [78] the viscosity term $\mu\partial^2 v(x,t)/\partial x^2$ in the Eq. (40) of the Payne-model. In [63], the Eq. (40) of the Payne-model has been rewritten as the Navier–Stokes equation:

$$\frac{\partial v(x,t)}{\partial t}+v\frac{\partial v(x,t)}{\partial x} = \frac{V_0(\rho)-v(x,t)}{\tau^{(0)}}-\frac{c_0^2}{\rho}\frac{\partial\rho(x,t)}{\partial x}+\frac{\mu}{\rho}\frac{\partial^2 v(x,t)}{\partial x^2} \tag{41}$$

with the speed-density relationship

$$V_0(\rho)=c_4\left[1+\left[\frac{\exp[(\frac{\rho}{\rho_{\max}})-c_5]}{c_6}\right]^{-1}-c_7\right], \tag{42}$$

c_i, $i=4,5,6,7$ are model parameters. There are a variety of other Navier–Stokes-like and gas-kinetic non-local traffic flow models, which, as the Payne-model, consist of the vehicle balance and velocity equations (see [18,102] and Sect. D of [40]). We may call all these macroscopic traffic flow models Payne-like models. This is because all of these macroscopic models show qualitatively the same features of free flow instability and resulting wide moving jam formation in free flow as those firstly found in [63,64] for a version of the Payne-model (39), (41), (42).

In the Aw–Rascle macroscopic model [2], instead of the Payne-Eq. (40), the following velocity equation has been introduced:

$$\frac{\partial v(x,t)}{\partial t}+v\frac{\partial v(x,t)}{\partial x}=\frac{V_0(\rho)-v(x,t)}{\tau^{(0)}}+\rho\frac{\mathrm{d}P}{\mathrm{d}\rho}\frac{\partial v(x,t)}{\partial x}, \tag{43}$$

where $P(\rho)$ is a density function. When the same fundamental diagram is chosen in this model as that in the Payne model, then the Aw–Rascle macroscopic model (39), (43) [2] exhibits also qualitatively the same features of free flow instability and resulting congested pattern formation [116] as those in Payne-like models and other GM-class models found earlier in [41,63,64,82].

Wiedemann Psychophysical Traffic Flow Model In Wiedemann's psychophysical model [127], a driver changes acceleration (deceleration) with a reaction time τ_0, if some thresholds (action points) are crossed where the driver changes his or her behavior. These thresholds are usually presented in the relative-speed-space-gap plane for a pair of preceding and follower vehicles. When the relative speed Δv is a large enough negative value and the associated threshold for vehicle deceleration is crossed, the driver decelerates to have the relative speed $\Delta v = 0$ at a desired space gap that depends on the vehicle speed. In the Wiedemann-model it is suggested that at small values of Δv a driver is not able to recognize whether he or she is slower and faster than the preceding vehicles. For this reason, when the speed of the preceding vehicle is a time-independent one, after reaching the desired space gap, the driver decelerates with a small deceleration before another threshold at a greater space gap is crossed. At this threshold the driver should recognize that he or she is slower than the preceding vehicle and so accelerates with a small acceleration to the desired space gap, then the driver decelerates with a small deceleration, and so on. Rather than this vehicle motion with small deceleration (acceleration), the dependence of the mentioned desired space gap on speed plays the role of the fundamental diagram for steady states in the GM-class models and together with a value of driver reaction time τ_0 determines the instability of model solutions such as in the GM-class models (see Sect. "Traffic Breakdown Resulting from Free Flow Instability in GM-class Models: Wide Moving Jam Emergence").

Nagel–Schreckenberg Cellular Automata (CA) Traffic Flow Model A different mathematical description for driver reaction time within the framework of the GM-class models has been introduced by Nagel and Schreckenberg [97]. In the Nagel–Schreckenberg (NaSch) stochastic

cellular automata (CA) model, in which the time and space are discrete values, the driver reaction time is described through the use of model driver fluctuations. The discrete time is $t = n\tau$, $n = 0, 1, 2, \ldots$; τ is the time step and the road is divided into cells of a finite length (see a history of CA traffic flow model development in review [88]). In the initial version of the NaSch CA model the cell length has been chosen to be equal the vehicle length d [97]. The update rules for vehicle motion in the NaSch CA model can be written as follows

$$v_{n+1} = \max(0, \min(v_{\text{free}}, v_n + 1, g_n) - \xi_n) \,, \tag{44}$$

$$x_{n+1} = x_n + v_{n+1} \,, \tag{45}$$

$$\xi_n = \begin{cases} 1 & \text{for } r \le p \,, \\ 0 & \text{otherwise} \,, \end{cases} \tag{46}$$

$r = \text{rand}(0, 1)$ denotes a random number uniformly distributed between 0 and 1; $p < 1$; time and space are in the units of τ and d, respectively.

To understand the NaSch-model, firstly note that g_n determines in (44) a safe speed: when $v_n > g_n$, then a driver decelerates at time step $n + 1$ because from (44) it follows that

$$v_{n+1} \le g_n \,. \tag{47}$$

This prevents vehicle collisions. Note that the safe speed determined through the space gap g was first introduced by Pipes [106].

Secondly, the condition

$$v_{n+1} = v_n + 1 \tag{48}$$

means that the vehicle accelerates at time step $n + 1$. However, in the NaSch-model due to model fluctuations with probability p this acceleration does not occur and instead of (48) from (44) it follows that the vehicle maintains its speed:

$$v_{n+1} = v_n \,. \tag{49}$$

Thus in this case, model fluctuations simulate a time delay in vehicle acceleration; this time delay is equal to

$$\tau_{\text{del}}^{(\text{a})} = \frac{\tau}{1 - p} \,. \tag{50}$$

Thirdly, let us assume that the vehicle should decelerate at time step $n + 1$ that occurs if $v_n > g_n$. Then without taking fluctuations into account from (44) the condition $v_{n+1} = g_n$ would be satisfied. However, due to model fluctuations with probability p we find $v_{n+1} = g_n - 1$, i. e., the vehicle decelerates stronger than is needed for safety conditions. This is the main idea of the GM-model for the over-deceleration effect discussed above, which should explain traffic flow instability.

In car-following models like the GM-model and different kinds of OV-models, *each* of the vehicles exhibit the same microscopic time delays, which are determined by deterministic rules of model motion (for the models of identical vehicles that are considered here only). In contrast, in the NaSch-model these driver time delays are simulated through the use of random model fluctuations. As a result, in the NaSch-model these driver time delays are described as "collective effects" that occur *on average* in traffic flow.

One of the advantages of the NaSch CA model (44) is the very fast computer simulation times of traffic flow in large networks. A disadvantage of this model in comparison with deterministic traffic flow models is the very great (non-realistic) fluctuations of vehicle speed. We know now that these large model fluctuations are necessary to simulate driver delay times in vehicle acceleration and deceleration. The problem of large speed fluctuations has been solved in the Krauß-stochastic microscopic model [77], which uses the same ideas for mathematical description of driver delay times in vehicle acceleration and deceleration as those in the NaSch CA model. However, rather than a discrete space of CA models, in the Krauß-model the continuum space co-ordinate is used. The Krauß-model can be written as follows

$$v_{n+1} = \max\left(0, \min\left(v_{\text{free}}, v_n + a\tau, v_n^{(\text{safe})}\right) - \xi_n\right) \,, \tag{51}$$

$$x_{n+1} = x_n + v_{n+1}\tau \,, \tag{52}$$

$\xi_n = a\tau r$, a is the maximum acceleration, the safe speed $v_n^{(\text{safe})} = v^{(\text{safe})}(g_n, v_{\ell,n})$ in (51) is a solution of the Gipps-equation [36]

$$v^{(\text{safe})}\tau^{(\text{safe})} + X_{\text{d}}(v^{(\text{safe})}) = g_n + X_{\text{d}}(v_{\ell,n}) \,, \tag{53}$$

where $X_{\text{d}}(u)$ is the distance traveled by the vehicle with an initial speed u at a time-independent deceleration b until it comes to a stop; $\tau^{(\text{safe})}$ is a safe time gap that can be individual for drivers; due to discrete model time $t = n\tau$, $n = 0, 1, 2, \ldots$ with time step $\tau = 1$ used in the model (51)

$$X_{\text{d}}(u) = b\left(\alpha\beta + \frac{\alpha(\alpha - 1)}{2}\right) \,, \tag{54}$$

where β and α are the fractional and integer parts of u/b, respectively; $\tau^{(\text{safe})} = \tau$.

Traffic Breakdown Resulting from Free Flow Instability in GM-class Models: Wide Moving Jam Emergence What types of congested traffic patterns should occur due to free flow instability in the GM-class models? This question has been answered by Kerner and Konhäuser in 1994 from their numerical study of a version of the Payne macroscopic model (39), (41), (42) [63]. As a result of the instability of free flow at a density that is greater than the critical one $\rho_{cr}^{(J)}$, wide moving jams emerge *spontaneously* in free flow (F → J transition for short).

The F → J transition is a first-order phase transition [63]. This means that at smaller densities $\rho < \rho_{cr}^{(J)}$ associated with metastable states of free flow a wide moving jam can be excited by a speed (or density) disturbance in an initial homogeneous free flow (Fig. 28a). If an initial amplitude of the growing disturbance is small enough, the F → J transition is associated with a "boomerang" behavior of the disturbance (Fig. 28a) [63]: firstly, the disturbance propagates downstream in free flow; then the disturbance comes to a stop strongly growing in its amplitude; as a result, the disturbance begins to propagate upstream; finally, a wide moving jam is forming that propagates upstream, i. e., the jam propagates through the location at which the initial disturbance has occurred.

The same boomerang effect associated with the F → J transition has been found, when small amplitude disturbances occur at an on-ramp bottleneck (Fig. 28b) [64]. Indeed, due to a speed disturbance that permanently exists at the bottleneck, a wide moving jam emerges spontaneously at the bottleneck in a metastable free flow [40, 41,64,82,119] (Figs. 28b and 29). At the same flow rate on the main road upstream of the bottleneck and a greater flow rate to the on-ramp a sequence of wide moving jams emerges spontaneously (Fig. 29c).

The conclusion in [63] that instability of free flow, which should explain traffic breakdown, leads to F → J transition is the *general* one for all models of the GM-class, which shows this instability (e. g., [40]). In particular, F → J transitions at the bottleneck for the NaSch CA model (Fig. 29e, f) and for the Wiedemann-model (Fig. 29g, h) show qualitatively the same features of spontaneous wide moving jam emergence as those in a version of the macroscopic Payne-like model in [63,64,82] (Fig. 29b, c).

In the macroscopic model (39), (41), (42), the safety gap g_{del} has been simulated through a choice of the fundamental diagram (42) with a very small absolute value of the derivative $|dV/d\rho|$ in the neighborhood of the jam density (Fig. 29a) [63]. Then due to this very small value $|dV/d\rho|$ at the jam density a vehicle could not almost accelerate from the initial speed $v = 0$ before the space gap to the preceding vehicle increases considerably. Thus only after a relatively long time delay in vehicle acceleration $\tau_{del}^{(a)}$ associated with the gap g_{del}, which considerably exceeds the initial space gap within the jam, could the vehicle escape from the jam. Thus this choice of the fundamental diagram shape simulates also implicitly a much longer time delay in vehicle acceleration $\tau_{del}^{(a)}$ at speeds that are close to zero $v \approx 0$ within the jam in comparison with considerably shorter time delay in vehicle acceleration at higher speeds. Different time delays in vehicle acceleration $\tau_{del}^{(a)}$ at $v \approx 0$ and at higher speeds is known as a slow-to-start rule in traffic flow modeling.

By introducing of the slow-to-start rule in the NaSch-model [6], the initial NaSch-model has been further developed to show qualitatively the same features of wide moving jam propagation as those in the Kerner–Konhäuser theory of wide moving jams and in empirical data [52]. It must be noted that rather than the implicit simulation of the slow-to-start rule through the use of a special form of the fundamental diagram discussed above, another mathematical idea of simulation of slow-to-start rule firstly introduced in [115] has been used in the NaSch-model [6]. Recall that in the NaSch-model the probability of fluctuations p simulates the time delay in acceleration in accordance with (50). Thus to simulate the slow-to-start rule,

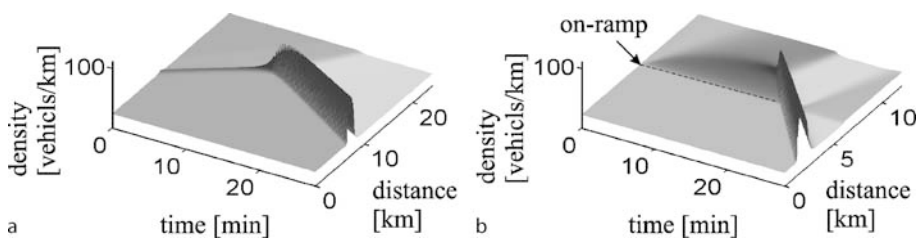

Traffic Congestion, Modeling Approaches to, Figure 28
"Boomerang" behavior of the growing disturbance in metastable free flow by wide moving jam formation. Vehicle density development in space and time. Simulations of the Payne-like model in [63]: a Homogeneous road [63]. b On-ramp bottleneck [64]

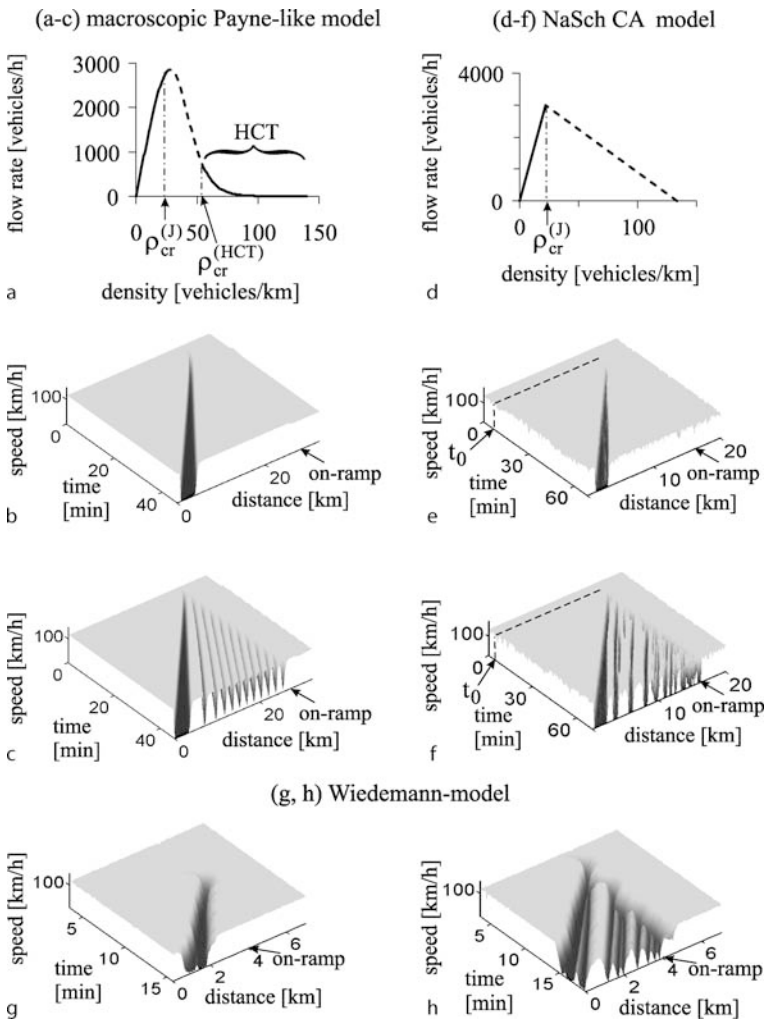

Traffic Congestion, Modeling Approaches to, Figure 29

Traffic breakdown at on-ramp bottleneck in the GM-model class with free flow instability [40,64,119]: Wide moving jam formation in free flow in a Payne-like model in [63] (a–c), in the NaSch CA model (d–f), and in the Wiedemann-model (g,h) (simulations with VISSIM 4.10-10 based on the Wiedemann-model; simulation parameters: flow rate in free flow on the main single-lane road upstream of the bottleneck $q_{in} = 2750$, flow rate to the on-ramp $q_{on} = 200$ vehicles/h for g and $q_{in} = 2500$, $q_{on} = 400$ vehicles/h for h, length of on-ramp merging region is 300 m, maximum speed in free flow is within the range 119–120 km/h, passenger vehicles only). a,d The fundamental diagrams of the Payne-like model (a) with parameters of Fig. 28 and the NaSch CA model (d); *dashed parts* of the diagrams are related to unstable states. b,c,e,f–h Average speed in space and time

this probability has been taken considerably greater at $v = 0$ than at $v > 0$ [6]:

$$p(v_n) = \begin{cases} p_1 & \text{for } v_n = 0\,, \\ p_2 & \text{for } v_n > 0\,, \end{cases} \tag{55}$$

where p_1, p_2 are constants, $p_1 > p_2$.

Thus, traffic breakdown is explained by an F → J transition in all models with free flow instability discussed in this subsection and a huge number of other traffic flow models of the GM-class models [16,40,95,98].

However, this general theoretical result [40,41,63,64,82,92,95,98,119] is in contradiction with empirical results; rather than the F → J transition, an F → S transition governs traffic breakdown in real traffic flow (Sect. "Highway Capacity and First-Order F → S Transition"). In other words, the traffic flow models with free flow instability of the GM-class [16,18,33,35,37,40,84,92,93,95,98,126,127,

128] cannot be used for a correct description of traffic breakdown.

The GM-class models reviewed in [16,18,33,35,37,40,84,92,93,95,98,102,126,127,128] cannot also been used for an adequate modeling of wide moving jam emergence; indeed, in the GM-class models wide moving jam emergence is explained by an F → J transition. In contrast, in empirical observations wide moving jams occur due to the F → S → J transitions (Sect. "Moving Jam Emergence in Synchronized Flow (S → J Transition)").

Why Wide Moving Jams do not Emerge Spontaneously in Empirical Free Flow To explain the critical conclusions about the GM-class models made above, we consider the following hypothesis of three-phase traffic theory [49,50]:

- At any density of free flow at which an F → J transition and an F → S transition are possible, a critical local speed (density) disturbance in free flow needed for the F → S transition is considerably smaller than that needed for the F → J transition.

To understand this hypothesis, we should compare two driver behavioral effects in free flow: the speed adaptation effect, which is responsible for an F → S transition ("Traffic Breakdown Explanation Through Fundamental Hypothesis of Three-Phase Traffic Theory"), and over-deceleration effect, which is responsible for an F → J transition.

As the speed adaptation effect, the over-deceleration effect occurs also within a local speed disturbance within which the speed is lower and the density is greater than in an initial free flow. In accordance with the over-deceleration effect, a driver approaching the slower moving preceding vehicle decelerates stronger than is required to avoid collisions. As a result, the speed of the driver becomes *lower* than the speed of the preceding vehicle. This can occur with a great probability if an initial space gap is small enough. If all following vehicles move initially also at small enough space gaps, then due to over-deceleration they decelerate stronger than is required to avoid collisions. Then the speed of each next following vehicle decreases continuously up to zero – wide moving jam emerges. Thus, the nucleation of wide moving jam in free flow is possible if each (or many) of the drivers following each other decelerate stronger than the associated preceding vehicles.

In real traffic flow, however, a driver, which moves initially in free flow, while approaching a preceding vehicle moving with a synchronized flow speed, begins to decelerate at a synchronization gap. This gap is great enough. For this reason, the driver will not necessarily decelerate to a lower speed than the synchronized flow speed of the preceding vehicle; this driver has enough time to compensate the driver's reaction time and to adjust the speed to that of the preceding vehicle. Thus the necessary condition for an F → J transition – the over-deceleration effect, which can occur only at small enough initial space gaps between vehicles following each other, is much harder to satisfy than the necessary condition for an F → S transition – the speed adaptation effect, which occurs within the synchronization gap. This driver behavior – speed adaptation within the synchronization gap explains the above hypothesis of three-phase traffic theory and empirical observations in which the F → S transition governs traffic breakdown.

Although a spontaneous F → J transition is not observed in real traffic, however, an induced F → J transition is possible. The induced F → J transition can be observed, if free flow on a road in a metastable state with respect to wide moving jam emergence, i. e., the flow rate in the free flow is equal to or greater than the outflow from a wide moving jam q_{out} (Sects. "Characteristic Parameters of Wide Moving Jam Propagation and Line *J*" and "Traffic Breakdown Resulting from Free Flow Instability in GM-class Models: Wide Moving Jam Emergence"). For example, a moving jam can initially occur within synchronized flow in an off-ramp lane of the road; if later this jam reaches the road, the jam can grow over time while propagating upstream on the road (see an empirical example of such an induced F → J transition in Sect. 12.6 in [52]).

Critical Discussion of Homogeneous Congested Traffic (HCT) In the Payne-like macroscopic models (e. g., [63,82]), the OV-models (e. g., [4,5]), the IDM [40,119] as well as some other traffic flow models (see references in [40]), the density region at the fundamental diagram, within which traffic flow is unstable, is limited at greater densities, i. e., flow states at the fundamental diagram are unstable *only* within the density range (dashed part of the fundamental diagram in Fig. 29a):

$$\rho_{\text{cr}}^{(\text{J})} < \rho < \rho_{\text{cr}}^{(\text{HCT})} \,. \tag{56}$$

In other words, within the density range

$$\rho_{\text{cr}}^{(\text{HCT})} < \rho < \rho_{\max} \tag{57}$$

homogeneous model states of congested traffic are stable with respect to small amplitude fluctuations. These states, in which the speed and flow rate are *homogeneous in space and time-independent*, have been called in [40,41] as *homogeneous congested traffic* (HCT) (Fig. 30a). The more the density exceeds $\rho_{\text{cr}}^{(\text{HCT})}$, the more stable is HCT with

respect to non-homogeneous speed disturbances in these models.

It must be noted that HCT is *not* a general result of traffic flow models of the GM-model class. No HCT appears regardless of the density, for example, in the NaSch CA model or in the Krauß-model: beginning from the critical density $\rho_{cr}^{(J)}$, all states of congested traffic at the fundamental diagram in these models are also unstable up to the jam density (Fig. 29d).

In simulations of traffic at an on-ramp bottleneck, HCT occurs at great enough flow rate to the on-ramp [8,40,41,82,119]. Thus, in simulations of the models satisfying (56) (Fig. 29a) the main features of congested patterns upstream of the on-ramp [8,40,41,82,95,119] may be illustrated with the following simplified *theoretical schemes*:

- high flow rate on the main road upstream of the on-ramp *and* low flow rate to the on-ramp → different kind of moving jams emerge spontaneously (Fig. 29b, c, e, f, g, h);
- high enough flow rate to the on-ramp → homogeneous congested traffic (HCT) occurs in which the density is high, the speed is very low and no moving jams and no other non-homogeneous traffic states can emerge spontaneously (Fig. 30a).

In empirical observations [52], however, the following *empirical schemes* are observed:

- Low flow rate to the on-ramp → synchronized flow in which the density is relatively low and the speed relatively high occurs; moving jams should not necessarily emerge;
- High enough flow rate to the on-ramp → moving jams emerge spontaneously in synchronized flow upstream of the on-ramp.

It can be seen that the theoretical schemes [8,40,41,82,95,119] and the empirical schemes [52] are inconsistent.

In contrast with this empirical scheme [52], in [111,119] "empirical" congested traffic states have been published, which should in particular prove the existence of HCT. Based on the measured data used in [111], let us show that this empirical proof is invalid.

In Fig. 10 in [111], spatiotemporal speed distributions within two congested patterns are shown to be homogeneous during congested pattern existence. The average speed within the associated patterns is very low, because these patterns have been caused by accidents. It must be stressed that these results in [111] have been derived with an adaptive smoothing method of data processing discussed in [111], i. e., with *processed data sets*. In contrast, our Fig. 31 shows *real unprocessed raw measured* data for one of these congested patterns related to Fig. 10a in [111].

To explain real measured data shown in Fig. 31, we should note that already in raw unprocessed data there is a large error in the average speed, when very low speeds are measured; if speeds, v, of all vehicles that have passed a detector during a 1-min interval are within the range $0 < v < 20$ km/h, then the road computer sets the average speed to 10 km/h. Only if no vehicle passes a detector during a 1-min interval the speed (and flow rate) is zero. This explains why in the speed data shown in Fig. 31 there are mostly two speed values, zero and 10 km/h. Only when average speeds are higher than 30 km/h, the speed can be used in deciding whether the speed distribution is really a homogeneous one or not.

We see that for low speeds based on the data, regardless of a method of further speed data processing, *no* conclusion about features of a spatiotemporal pattern can be made from an analysis of speeds *only*, as made in [111]. Rather than the average speed, the flow rate is measured with a sufficient accuracy at any density. We can see from the flow rate distribution shown in Fig. 31 that there are extremely complex spatiotemporal flow rate changes both in space and time between zero and 8 vehicles/min. This explains that in contrast with the statement in [111], the congested pattern is extremely non-homogeneous in space and time. Note that features of such complex spatiotemporal congested patterns occurring at heavy bottlenecks caused for example by accidents or bad weather condition

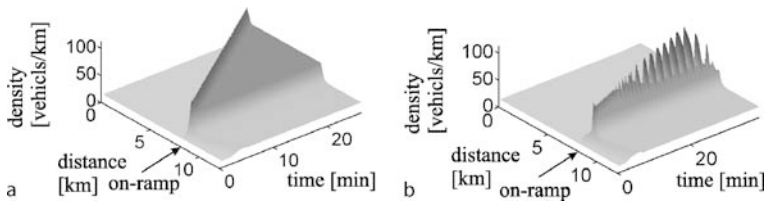

Traffic Congestion, Modeling Approaches to, Figure 30
Simulations of HCT (a) and OCT (b) at on-ramp bottleneck [40,41,82,119]

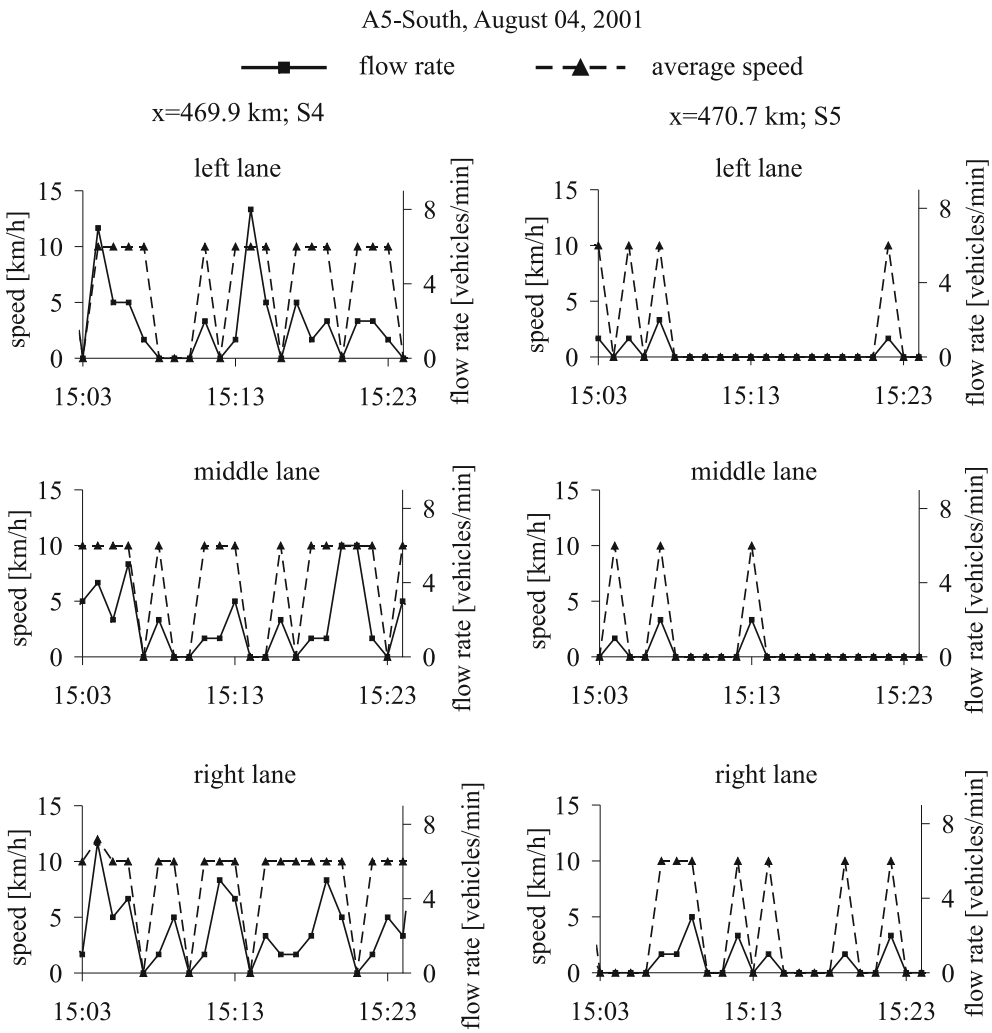

Traffic Congestion, Modeling Approaches to, Figure 31

Measured unprocessed flow rate (*squares*) and speed (*triangles*) at two detector locations within a congested pattern that has been presented in Fig. 10a in [111] as "homogeneous congested traffic" (HCT). The locations $x = 469.9$ km for the detector S4 and $x = 470.7$ km for the detector S5 are chosen in accordance with the freeway section sketch shown in Fig. 6 in [111]. The raw data is 1-minute averaged data

have recently been found in [55] and appears in ► Traffic Congestion, Spatiotemporal Features of.

Thus in reality the congested pattern shown in Fig. 10a in [111] has no relation with HCT. The same critical conclusion can be made about the congested pattern shown in Fig. 10b in [111], which should be another "empirical" example of HCT.

Due to existence of HCT model solutions, these traffic flow models exhibit also model solutions called oscillatory congested traffic (OCT) (Fig. 30b) [8,40,41,82,119]. Indeed, OCT appears in a neighborhood of the critical density $\rho_{cr}^{(HCT)}$ for instability of HCT: when the density in HCT decreases and it approaches the critical density $\rho_{cr}^{(HCT)}$, then due to instability of HCT, OCT occurs. Thus OCT model solutions result from the existence of HCT model solutions in these models, i. e., as HCT, OCT model solutions have no relation to real traffic flow.

In accordance with criticisms of the GM-class models made above, we should mention that congested pattern diagrams, which describe different types of congested traffic patterns in these models [8,40,41,64,65,82,95,119], have no relation to real traffic. This criticism includes also earlier results of the author et al. [64,65] made in the context of the fundamental diagram approach. The main point of

the criticism of these congested pattern diagrams is that at a great enough flow rate upstream of the bottleneck the onset of congestion in all these congested pattern diagrams is associated with F → J transitions [8,40,41,64,65,82,95,119] (Fig. 29); in contrast with this model result, in real traffic the onset of congestion in all observations is governed by an F → S transition (Sect. "Spontaneous Traffic Breakdown (F → S Transition)"). In addition, at great flow rates to the on-ramp, OCT and HCT model solutions appear in these diagrams, which, as explained above, have no relation to real traffic flow.

Models Combining LWR- and GM-Approaches

Considering wide moving jam emergence in the GM-class models, we have above suggested that the critical density $\rho_{cr}^{(J)}$ for the instability leading to F → J transitions is related to free flow, specifically, the condition

$$\rho_{cr}^{(J)} < \rho_0 \tag{58}$$

is satisfied, where the density ρ_0 is associated with the maximum flow rate at the fundamental diagram.

However, parameters of many traffic flow models of the GM-model class can also be chosen in the way that the critical density $\rho_{cr}^{(J)}$, at which steady speed states on the fundamental diagram becomes unstable as the density increases, is greater than the density ρ_0 (Fig. 32a,b):

$$\rho_{cr}^{(J)} > \rho_0 \,. \tag{59}$$

As in the LWR-model class, none of the fundamental empirical traffic breakdown features **B–D** of Sect. "Highway Capacity and First-Order F → S Transition" can be shown in the models. Dependent on the shape of the fundamental diagram and model parameters, there are two types of such models. In the first model type, all steady speed states are unstable when the density satisfies the condition $\rho > \rho_{cr}^{(J)}$ (Fig. 32a). In the second model type, when the density satisfies the condition $\rho \geq \rho_{cr}^{(HCT)}$ (Fig. 32b), there are also HCT model solutions criticized in Sect. "Critical Discussion of Homogeneous Congested Traffic (HCT)". Features of the GM-model class appear in these models when q_{on} increases and the density on the main road approaches the critical density $\rho_{cr}^{(J)}$. Then wide moving jams emerge spontaneously in the dense flow upstream of the bottleneck (Fig. 32d).

Thus by changing of model parameters in some of the models of the GM-class models, a model can exhibit qualitatively different conditions for congested traffic occurrence depending on whether the condition (58) (Fig. 29), or the condition (59) (Fig. 32) is satisfied, or else at any density there cannot be model instability leading to wide moving jam emergence at all, i. e., when the model exhibits qualitatively the same features as those of the LWR-model (Fig. 27).

In particular, when the shape of the fundamental diagram is chosen, then such transitions from condition (58) to condition (59) and, finally, to the LWR-model class occur in the OV model (33) by a gradual continuous decrease

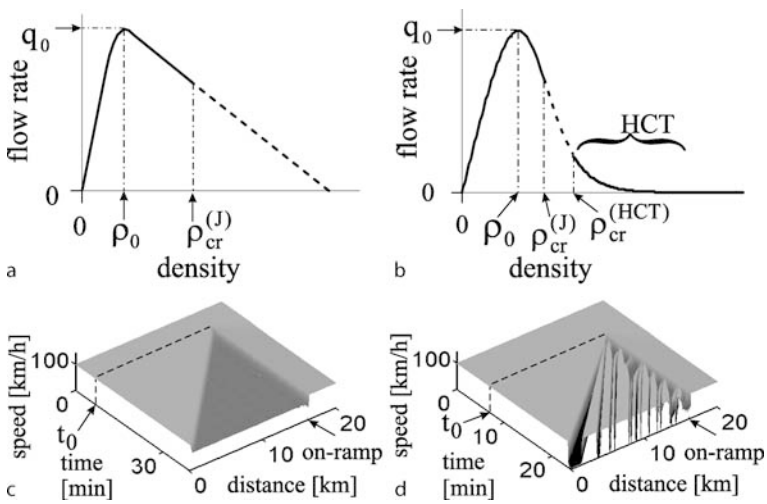

Traffic Congestion, Modeling Approaches to, Figure 32

Traffic breakdown and congested patterns at on-ramp bottleneck in the models combining LWR- and GM-approaches under condition (59): a,b Two types of fundamental diagrams; *dashed* parts show unstable flow states. c Propagation of (shock) waves under condition (26) that is the same as that in Fig. 27d. d Spontaneous wide moving jam emergence in the dense traffic in a when the flow rate to the on-ramp increases and the density downstream of the bottleneck approaches the critical density $\rho_{cr}^{(J)}$ [62]

in the model parameter $\tau_0^{(0)}$, in a version of the Payne-model (39), (41), (42) by a gradual continuous increase in the model parameter c_0, in the IDM (37), (38) by a gradual continuous increase in the model parameter c_3.

Critical Discussion of Empirical Tests of Traffic Flow Models

Obviously, traffic flow theories must be based on real behavior of drivers in traffic, and their solutions should show phenomena observed in traffic flow. For this reason, an empirical test of a traffic flow theory is of great importance. There are several approaches to perform such a test [69]:

(i) Vehicle speeds, time headways (net time gaps), and accelerations measured in a car-following experiment are compared with that, which a model shows (e. g., [84]).
(ii) Spatiotemporal evolution of speed (or density) associated with a processing of real empirical data made for example through the use of adaptive smoothing methods are compared with results of numerical simulations of congested patterns [111,119].
(iii) The empirical fundamental diagram for traffic flow (flow-density relationship) associated with measurements made at a freeway location are compared with a theoretical fundamental diagram for traffic flow used in a model or that results from the model (e. g., [15,31,42,44,71,72,83,99,112,117,129]).
(iv) Time headway distributions, optimal velocity (OV) functions, and some other single vehicle characteristics that are measured at a freeway location or result from data aggregation are compared with the associated model results (e. g., [71,72,99]).
(v) Traffic variables (e. g., flow rate and average speed) are measured on at least three different freeway locations. Measurements from two locations, which are the farthest upstream and downstream, are used as upstream and downstream boundary conditions for a model, respectively. The model calculates spatiotemporal distributions of traffic variables between these two locations. The distributions should correspond to empirical data measured at the locations, which have not been used as the boundary conditions for the modeling (e. g., [13,14,18,125]).

In all these cases, model parameters are chosen to have the best agreement with the associated empirical data.

As explained above in this section, past traffic flow theories and models reviewed in [16,33,40,95,98,128] cannot explain and reproduce empirical features of traffic breakdown. Nevertheless, these models can show a good correspondence with empirical data in the above empirical test approaches (i)–(v).

To explain this "phenomenon", we should note that empirical data used in the tests (i)–(iv) does not contain complete features of spatiotemporal traffic dynamics, which has been considered in part II of [52]. To study these fundamental features (traffic breakdown and resulting spatiotemporal congested patterns), real unprocessed measured data, which should include complete spatiotemporal distribution of synchronized flow within a congested pattern (see Sect. 2.4.11 in [52]), should be studied.

In particular, due to smoothing and/or selection of data in the empirical test (ii) [111,119], important characteristics of congested patterns get lost resulting in invalid analysis and classification of empirical congested traffic states. A characteristic example of such invalid analysis of empirical features of traffic congestion and the associated test of simulation results made in [111,119] has been considered in Sect. "Critical Discussion of Homogeneous Congested Traffic (HCT)" (Fig. 31).

The fundamental diagram, OV functions as classified in [71,72,99], time headway distributions, and other macroscopic and single vehicle (microscopic) characteristics of traffic flow used up to now in empirical tests (iii) and (iv) are associated with an averaging of spatiotemporal traffic pattern characteristics. For this reason, important features of spatiotemporal congested patterns and phase transitions in traffic flow are *lost* in these traffic flow characteristics. Thus in contrast with [71,72,99] it is not justified to use these macroscopic and microscopic characteristics as the *solely* empirical basis for a decision whether a traffic flow model can describe real traffic flow or not.

Moreover, as explained in ► Traffic Congestion, Spatiotemporal Features of, a criterion for the definitions of synchronized flow and wide moving jam in [71,72,99], which is based on a comparison of the flow-density correlations within empirical data associated with congested traffic, is incorrect. As shown in ► Traffic Congestion, Spatiotemporal Features of and [67], this is associated with a *large systematic error* in calculations of vehicle density made in [71,72,99] from empirical data (for a more detailed consideration see ► Traffic Congestion, Spatiotemporal Features of). For this reason, empirical tests of the CA traffic flow models, which are related to these traffic phase definitions, made in [71,72,99] are also invalid.

These critical conclusions explain why the NaSch CA models [6,97] and their further developments including a CA model with "comfortable driving" [70] show empirical fundamental diagrams, empirical OV functions and

time headway distributions satisfactory [72], even though these CA models [6,70,97] as found in [66] (see explanations to simulation results of the NaSch CA model with "comfortable driving" [70] presented in Figs. 20–23 in [66]) cannot show and predict the main empirical spatiotemporal features of phase transitions and synchronized flow.

In addition, it should also be noted that in empirical tests of the CA traffic flow models made in [71,72,99], simulations of spatiotemporal congested patterns shown by the CA traffic flow models are made for a spatially homogeneous road, i. e., the road without bottlenecks. In contrast, measured data used for these tests are related to congested traffic occurring due to bottlenecks. Spatiotemporal congested patterns occurring in simulations made on the homogeneous road depend fundamentally on initial conditions and on inflow conditions given in a model at the upstream boundary of the road (e. g., on time interval distributions for vehicles entering the road). However, as shown in [52], the vehicle dynamics at a bottleneck makes the greatest influence on spatiotemporal congested patterns occurring in simulations performed at the bottleneck. This means that for an adequate empirical test of traffic flow models rather than simulations on the homogeneous road [71,72,99], model simulations of spatiotemporal congested patterns should be made at bottlenecks [68,69].

In empirical test (v), firstly congested traffic is measured at a road location. Then measured traffic variables associated with this congested traffic are used at the downstream boundary of a traffic flow model. In other words, at this boundary traffic variables are *given* as time functions associated with congested traffic measured at a freeway location related to this model boundary. As a result, in simulations congested traffic given at the model boundary propagates further upstream. If the characteristic features of wide moving jam propagation can be shown by the model (as explained above, this is the case for many models of the GM-model class [40,95,98]), then a wave of congested traffic can propagate in the model upstream with velocities, which can be chosen close to empirical ones. For this reason, an approximate correspondence between model and some empirical traffic variable functions is possible [13,14,125], although the models [40,95,98] cannot show and cannot reproduce empirical traffic breakdown and many resulting empirical spatiotemporal congested patterns.

This is because the test (v) is inconsistent with the "open character" of non-linear dynamic process, "traffic". As mentioned, in this test *congested traffic* is often *given* at the farthest downstream boundary of a freeway network model. This is not the case for real traffic, in which initial congestion occurs *spontaneously within* a real freeway network, mostly at a bottleneck. This bottleneck cannot be the *farthest* downstream boundary of this network. To simulate the open traffic process adequately with real traffic, vehicles should leave *freely* the farthest downstream boundary of a network model. This means that *free flow conditions* should be given at the farthest downstream boundary of the network model.

Critical Discussion of Highway Capacity Definitions

There are two definitions of highway capacity, i. e., capacity of free flow at a bottleneck associated with the two model classes, the LWR-model and GM-model classes.

In the LWR-model, capacity of free flow at a bottleneck (bottleneck capacity) is equal to the flow rate q_{cap} downstream of the bottleneck (15): if the sum of flow rates upstream of the bottleneck exceeds q_{cap}, which is the maximum flow rate at the fundamental diagram, then traffic breakdown should occur leading to congested traffic upstream of the bottleneck.

To satisfy probabilistic features of traffic breakdown, in recent definitions of highway capacity is suggested that the capacity q_{cap} is a random value characterized by a capacity distribution function (e. g., [10]). However, even if is assumed that the highway capacity q_{cap} (15) is a random value whose distribution function is defined through an empirical breakdown probability [10], the fundamental empirical features of the capacity **A–C** (Sect. "Highway Capacity and First-Order F → S Transition") cannot be explained with the capacity determined by the maximum flow rate q_{cap} on the fundamental diagram (Fig. 24). This is because a random choice of q_{cap} changes *none* of qualitative features of the onset of congestion in the LWR-theory shown in Fig. 27; as explained in Sect. "Classic Lighthill–Whitham–Richards (LWR) Theory of Onset of Traffic Congestion", these LWR-theory features are, however, in a deep contradiction with the fundamental empirical features of traffic breakdown (Sect. "Highway Capacity and First-Order F → S Transition"). Thus even under assumption about a random character of q_{cap} in the capacity definition (15), the definition is not associated with empirical results. Therefore, traffic management methods based on this capacity definition cannot be used for a reliable traffic management [53].

In the GM-model class, the maximum bottleneck capacity can be defined to be equal to the flow rate downstream of the bottleneck at which free flow instability occurs in a traffic flow model. However, as shown in Sect. "Classic General Motors (GM) Model Approach:

Free Flow Instability due to Driver Reaction Time" in this model class the instability leads to an F → J transition. As mentioned above, this contradicts with the empirical fact that traffic breakdown is governed by an F → S transition. For this reason, the definition of free flow capacity made within the framework of the GM-model class does not satisfy the empirical feature of the bottleneck capacity **A** (Sect. "Highway Capacity and First-Order F → S Transition").

The capacity definition made in three-phase traffic theory that satisfies all fundamental empirical features of the bottleneck capacity **A–D** has been considered in Sect. "Highway Capacity and First-Order F → S Transition".

Common Features of Earlier Traffic Flow Modeling Approaches

The above consideration of earlier traffic flow models in the framework of the fundamental diagram hypothesis shows that stability features of steady state model solutions determine most of the qualitative features of the onset of congestion and resulting congested patterns. For this reason, it turns out that mathematically very different traffic flow models, which exhibit qualitatively the same stability features of steady speed states at the fundamental diagram, show also qualitatively the same features of traffic congestion. This leads to the traffic flow model classification made above and to the following conclusions about common features of earlier traffic flow modeling approaches:

(i) Most of the traffic flow models used by traffic researches are based on fundamental diagram hypothesis. These models can be classified into two main classes. The first model class refers to the classic LWR-theory. The basic idea of the LWR-theory is that maximum flow rate associated with the maximum point at the fundamental diagram determines free flow capacity at a bottleneck. Thus if the flow rate upstream of the bottleneck exceeds the capacity, then traffic breakdown should occur. The second model class refers to the basic idea of the classic GM-model. The basic idea of the GM-model approach is that beginning at a critical density there is instability of free flow. This free flow model instability associated with a finite value of a driver reaction time should explain traffic breakdown.
   (a) The LWR-theory and all traffic flow models based on this theory cannot show and cannot predict the fundamental empirical features of traffic breakdown (Sect. "Highway Capacity and First-Order F → S Transition") as well as empirical spontaneous moving jam emergence.
   (b) Free flow instability in models and theories based on the GM-model approach leads to an F → J transition. In contrast, in empirical observations, traffic breakdown is associated with an F → S transition. Thus, the models with free flow instability associated with a finite value of a driver reaction time cannot explain and cannot predict traffic breakdown observed in real traffic flow.

(ii) These traffic flow models are basic traffic flow models for simulations of freeway control and management strategies. However, we have to conclude that the related simulations of the control and management strategies cannot predict many of the freeway traffic phenomena that would occur through the use of a management strategy.

(iii) It must be stressed that the above criticism of all these earlier traffic flow modeling approaches does not diminish the following **achievements** of these approaches and associated models, which are also used in three-phase traffic theory and the related three-phase traffic flow models (see next Sect. "Three-Phase Traffic Flow Models"):
   (a) There are characteristic parameters and features of wide moving jam propagation [63]. One of these characteristic jam features is that the downstream front of wide moving jams propagates steadily along a road.
   (b) Classic ideas of traffic flow theories introduced and developed within the fundamental diagram approach about different driver time delays, various mathematical descriptions of driver acceleration and deceleration as well as safety conditions are also very important elements used in three-phase traffic flow models, which overcome the drawbacks of the earlier modeling approaches to traffic congestion (item (i) and (ii)). In particular, pertinent pioneering ideas, which were introduced in earlier models and theories by Herman, Montroll, Potts, Rothery, Gazis [34, 43], Komentani and Sasaki [73], Newell [100], Gipps [36], Payne [103,104], Nagel, Schreckenberg, Schadschneider, and co-workers [6,97], Bando, Sugiyama, and colleagues [4], Takayasu and Takayasu [115], Krauß et al. [77], Nagatani and Nakanishi [96] and by many other groups (see references in [16,18,21,33,35,37,40,84,88,92, 93,95,98,102,126,127,128]), are also very important elements of three-phase traffic models discussed below.

Three-Phase Traffic Flow Models

The first microscopic three-phase traffic theory that can reproduce empirical traffic breakdown and $F \to S \to J$ transitions (Sect. "Characteristic Parameters of Wide Moving Jam Propagation and Line *J*") as well as all known resulting empirical spatiotemporal congested patterns was developed in 2002 by Kerner and Klenov [57] (see below and for more detail Sect. 16.3 in [52]). They proposed a microscopic spatial continuum and discrete-time traffic flow model in which general rules of vehicle motion are in accordance with the fundamental hypothesis (Fig. 15) as well as with other hypotheses of three-phase traffic theory. Based on this Kerner–Klenov model, numerical spatiotemporal features of traffic breakdown and the pinch effect have firstly been found and studied. Some months later, Kerner, Klenov, and Wolf developed a CA approach to three-phase traffic theory (the KKW CA models) [66], which can also explain traffic breakdown and the pinch effect as observed in real traffic (see Sect. 16.2 and Chaps. 17–19 in [52]).

Later other various traffic flow models in the framework of the three-phase traffic theory have been developed: Davis [22] as well as Kerner and Klenov [62] have proposed different three-phase microscopic deterministic models; Jiang and Wu [45], Lee et al. [81], and Gao et al. [32] have developed various three-phase CA models; Siebel and Mauser [113] and Laval [80] have suggested macroscopic models; Jiang et al. [48] have suggested a macroscopic version of the Kerner–Klenov speed adaptation model in [62]. Features of phase transitions and congested patterns that these models exhibit are similar to those found in 2002 in the Kerner–Klenov stochastic and KKW CA models. Recent simulation results in the framework of the author's three-phase traffic theory can be found in [23,24,25,46,47,85,107,124].

It should be noted that the use of only some of the hypotheses of three-phase traffic theory in a traffic flow model does not ensure that the model can show empirical phase transitions in vehicular traffic. An example is Colombo's continuum model [17] that consists of Eq. (16) in which dependence $Q(\rho)$ is associated with the fundamental hypothesis of three-phase traffic theory (Fig. 19) [49,50,51]. However, solutions of Colombo's model [17] can show *none* of the empirical features of $F \to S$ and $S \to J$ transitions because in this model other hypotheses of three-phase traffic theory needed for the description of phase transitions in real traffic flow have not been taken into account.

For an example of three-phase traffic flow model development, we consider basic assumptions and formulation of the Kerner–Klenov stochastic three-phase traffic flow model for identical drivers and vehicles [57,58]. This are two reasons for this: (i) the model is the first three-phase traffic flow model and (ii) most of the simulation results discussed in Sects. "Free and Congested Traffic"–"Empirical Double *Z*-Characteristic for Phase Transitions in Traffic Flow" are based on this model.

Kerner–Klenov Stochastic Three-Phase Traffic Flow Model

Basic driver behavioral assumptions of the Kerner–Klenov model are as follows:

(i) *Fundamental hypothesis of three-phase traffic theory* (Sect. "Traffic Breakdown Explanation Through Fundamental Hypothesis of Three-Phase Traffic Theory"). In synchronized flow, a driver accepts a range of different hypothetical steady state speeds at the same space gap to the preceding vehicle. This means that hypothetical steady model states of synchronized flow cover a 2D-region in the flow-density plane (Fig. 33a). The boundaries of this 2D-region *F*, *L*, and *U* are respectively associated with free flow, a synchronization space gap, and a safe space gap determined through a safe speed. The 2D-region of steady states is associated with a driver behavioral assumption that in synchronized flow a driver is able to recognize whether the space gap is increasing or deceasing regardless of the speed difference to the preceding vehicle.

(ii) *Line J and 2D-region of steady states* (Sect. "Explanations of $S \to J$ Transitions Through Three-Phase Traffic Theory"). In the model, the line *J* is between the boundaries *L*, and *U* (Fig. 33b), i. e., the line *J* divides the 2D-region of steady states of synchronized flow onto two classes: the states on and above the line *J* and the states below the line *J*, which are metastable and stable states with respect to wide moving jam formation, respectively. Thus at a given steady speed, a driver behavioral assumption is that the space gap in synchronized flow associated with the line *J*, i. e., in the jam outflows is greater than the safe one and it is smaller than the synchronization gap.

(iii) *Speed adaptation effect in synchronized flow* (Sect. "Traffic Breakdown Explanation Through Fundamental Hypothesis of Three-Phase Traffic Theory"). The speed adaptation effect takes place when the vehicle cannot pass the preceding vehicle, within the space gap range:

$$g_{s,n} \le g_n \le G_n \,, \tag{60}$$

where $g_n = x_{\ell,n} - x_n - d$ is the space gap, x_n is the vehicle co-ordinate, the lower index ℓ marks functions and values related to the preceding vehicle; $g_{s,n}$ is the safe gap determined from the equation $v_n = v_{s,n}$, in which v_n is the vehicle speed, $v_{s,n}$ is a safe speed; G_n is a synchronization gap; all vehicles have the same length d, which includes the minimum space gap between vehicles within a wide moving jam; index n corresponds to the discrete time $t = n\tau$, $n = 0, 1, 2, \ldots$; τ is time step. Under condition (60), the vehicle tends to adjust its speed to the preceding vehicle without caring, what the precise space gap is, as long as it is safe. For example, at a given time-independent speed of the preceding vehicle $v_{\ell,n} = v_\ell = \text{const}$, this speed adaptation leads to car following with $v_n = v = v_\ell$ at a time-independent space gap $g_n = g$. There is an infinite number of these gaps associated with the same speed $v = v_\ell$. These gaps lie between the synchronization gap and safe gap, i. e., there is no desired (or optimal) space gap in synchronized flow.

(iv) *Over-acceleration effect* (Sect. "Traffic Breakdown Explanation Through Fundamental Hypothesis of Three-Phase Traffic Theory"). In synchronized flow of a lower density, a driver searches for the opportunity to accelerate and to pass. A competition between the speed adaptation (item (iii)) and over-acceleration effects simulates traffic breakdown (F → S transition). The over-acceleration is simulated as a collective effect, which occurs on average in traffic flow, through the use of lane changing to a faster lane as well as through the use of random vehicle acceleration

$$\xi_a = a\tau\Theta(p_a - r)\,, \tag{61}$$

where p_a is probability of random acceleration, a is the maximum acceleration, $r = \text{rand}\,(0,1)$, $\Theta(z) = 0$ at $z < 0$ and $\Theta(z) = 1$ at $z \geq 0$. The model fluctuations (61) are applied only if the vehicle should accelerate without model fluctuations.

(v) *Pinch effect in synchronized flow* (Sect. "Pinch Effect"). Moving in synchronized flow, a driver comes on average closer to the preceding vehicle over time that should explain the pinch effect. A driver time delay in deceleration simulates this effect through model fluctuations in deceleration

$$b_n = a\Theta(P_1 - r_1)\,, \tag{62}$$

applied under condition (60) only; P_1 is probability of random deceleration, $r_1 = \text{rand}\,(0,1)$.

(vi) *Over-deceleration effect*. As in the NaSch CA model, a well-known over-deceleration effect (human overreaction) associated with driver reaction time [34,43] is also simulated as a collective effect through the use of random fluctuations in vehicle deceleration

$$\xi_b = a\tau\Theta(p_b - r)\,, \tag{63}$$

which is applied only if the vehicle should decelerate without model fluctuations; p_b is the probability of random deceleration. In the Kerner–Klenov model, a competition between the over-deceleration and the speed adaptation effect (item iv) determines moving jam emergence in synchronized flow.

(vii) *Driver time delay in acceleration.* In the model, this well-known effect should describe driver delay in acceleration at the downstream front of synchronized flow or wide moving jam (in the latter case, the driver delay in acceleration is known as a slow-to-start rule [6,115] (Sect. "Traffic Breakdown Resulting from Free Flow Instability in GM-class Models: Wide Moving Jam Emergence")) after the preceding vehicle has begun to accelerate. A driver time delay in acceleration is simulated as a collective effect through the use of a random value of vehicle acceleration

$$a_n = a\Theta(P_0 - r_1) \tag{64}$$

applied under condition (60) and only then if the vehicle did not accelerate at the former time step; in the latter case $P_0 = p_0 > 0$, i. e., a vehicle accelerates with some probability p_0 that depends on the speed v_n; otherwise $P_0 = 0$. As in the NaSch CA model, the mean time in vehicle acceleration is

$$\tau_{\text{del}}^{(a)} = \frac{\tau}{1 - p_0}\,. \tag{65}$$

Thus, the basis of the Kerner–Klenov stochastic three-phase traffic flow model are driver behavioral assumptions made in three-phase traffic theory (items (i)–(v)). In addition, over-deceleration (item vi) and driver time delay in acceleration (item vii) introduced in earlier traffic flow models in the framework of the fundamental diagram approach have also been incorporated; in particular, as in the NaSch-model (44), these driver time delays appear through the use of model fluctuations. Finally, the Kerner–Klenov stochastic three-phase traffic flow model reads as follows:

$$v_{n+1} = \max(0, \min(v_{\text{free}}, \tilde{v}_{n+1} + \xi_n, v_n + a\tau, v_{s,n}))\,, \tag{66}$$

$$x_{n+1} = x_n + v_{n+1}\tau \,, \tag{67}$$

$$\tilde{v}_{n+1} = \max(0, \min(v_{\text{free}}, v_{c,n}, v_{s,n})) \,, \tag{68}$$

$$v_{c,n} = \begin{cases} v_n + \Delta_n & \text{at } g_n \le G_n \\ v_n + a_n\tau & \text{at } g_n > G_n \,, \end{cases} \tag{69}$$

where

$$\Delta_n = \max(-b_n\tau, \min(a_n\tau, v_{\ell,n} - v_n)) \,, \tag{70}$$

v_{free} is the maximum speed in free flow that is constant.

The synchronization gap G_n depends on the vehicle speed v_n and on the speed of the preceding vehicle $v_{\ell,n}$

$$G_n = G(v_n, v_{\ell,n}) \,, \tag{71}$$

where the function $G(u, w)$ is chosen as

$$G(u, w) = \max(0, k\tau u + \phi_0 a^{-1} u(u - w)) \,, \tag{72}$$

$k > 1$ and ϕ_0 are constants. If $v_n = v_{\ell,n}$, the synchronization gap G_n is $kv_n\tau$; this corresponds to a fixed time gap $k\tau$ that determines the line L in Fig. 33. If $v_n > v_{\ell,n}$, the gap G_n increases and vice versa.

As explained in items (iv) and (vi) above, random deceleration and acceleration ξ_n in (66) are applied depending on whether the vehicle decelerates or accelerates, or else maintains its speed:

$$\xi_n = \begin{cases} -\xi_b & \text{if } S_{n+1} = -1 \\ \xi_a & \text{if } S_{n+1} = 1 \\ 0 & \text{if } S_{n+1} = 0 \,, \end{cases} \tag{73}$$

where S in (73) denotes the state of motion ($S_{n+1} = -1$ represents deceleration, $S_{n+1} = 1$ acceleration, and $S_{n+1} = 0$ motion at nearly constant speed)

$$S_{n+1} = \begin{cases} -1 & \text{if } \tilde{v}_{n+1} < v_n - \delta \\ 1 & \text{if } \tilde{v}_{n+1} > v_n + \delta \\ 0 & \text{otherwise} \,, \end{cases} \tag{74}$$

where δ is constant ($\delta \ll a\tau$).

In (64) and (62), the probabilities P_0 and P_1 are

$$P_0 = \begin{cases} p_0(v_n) & \text{if } S_n \ne 1 \\ 1 & \text{if } S_n = 1 \,, \end{cases} \tag{75}$$

$$P_1 = \begin{cases} p_1 & \text{if } S_n \ne -1 \\ p_2(v_n) & \text{if } S_n = -1 \,, \end{cases} \tag{76}$$

where speed functions for probabilities $p_0(v_n)$ and $p_2(v_n)$ are considered in [52]; p_1 is constant.

In the model, the safe speed $v_{s,n}$ in (68) is chosen in the form

$$v_{s,n} = \min\left(v_n^{(\text{safe})}, \frac{g_n}{\tau} + v_\ell^{(a)}\right) \,, \tag{77}$$

where $v_n^{(\text{safe})} = v^{(\text{safe})}(g_n, v_{\ell,n})$ is taken from (51); $v_\ell^{(a)}$ is an "anticipation" speed of the preceding vehicle at the next time step:

$$v_\ell^{(a)} = \max\left(0, \min\left(v_{\ell,n}^{(\text{safe})} - a\tau, v_{\ell,n} - a\tau, \frac{g_{\ell,n}}{\tau}\right)\right) \,, \tag{78}$$

where $v_{\ell,n}^{(\text{safe})}$ is the safe speed for the preceding vehicle, $g_{\ell,n}$ is the space gap in front of the preceding vehicle.

Models for lane changing for a multi-lane freeway and of freeway bottlenecks associated with the Kerner–Klenov model as well as a further development of this model for heterogeneous flow consisting of various vehicles and drivers with different characteristics can be found in the book [52].

Methodology of Empirical Test

For an adequate comparison of empirical and theoretical dynamic nonlinear features of spatiotemporal congested patterns in freeway traffic, the following methodology of the empirical test is used [52].

As in real traffic, in a model of a freeway with on-ramp and off-ramp bottlenecks traffic occurs at the upstream road boundary of the main road and of an on-ramp(s). In

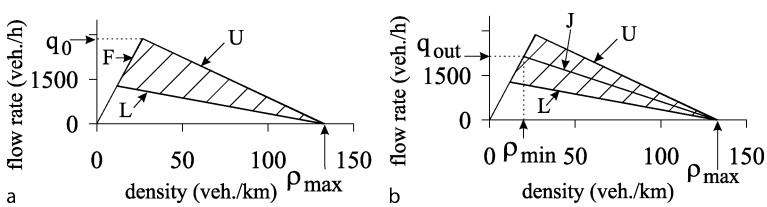

Traffic Congestion, Modeling Approaches to, Figure 33
Steady speed states for the Kerner–Klenov stochastic three-phase traffic flow in the flow-density plane (a) and the line *J* (b). Taken from [52]

this open traffic process, all traffic variables downstream result from traffic demand at the upstream road boundary and the on-ramp(s), as well as drivers' destinations (whether a vehicle leaves the main road to the off-ramp or it further follows the main road). At downstream model boundary conditions for vehicle freely leaving a modeling freeway section(s) are given. Spatiotemporal congested patterns emerge, develop, and dissolve in this open freeway model with the same types of bottlenecks as those in empirical observations.

The spatiotemporal dynamics of traffic breakdown and resulting congested patterns found in simulations should be compared with the associated spatiotemporal dynamics found in empirical observations. Only after the fundamental spatiotemporal features of traffic have been simulated and compared with the empirical data, *secondary characteristics* associated with these spatiotemporal patterns, which include for example fundamental diagrams, time headway distributions, speed and flow-density correlation functions and OV functions, can be compared with the associated empirical characteristics.

Such an empirical test made in [67,68,69] shows that three-phase traffic flow models can reproduce and predict all known microscopic and macroscopic empirical features of traffic breakdown and resulting congested traffic patterns.

Link Between Three-Phase Traffic Theory and Fundamental Diagram Approach

A link between three-phase traffic theory and the fundamental diagram approach to traffic flow modeling can be created through the use of the *averaging* of an infinite number of steady states of synchronized flow associated with a 2D region in the flow-density plane (Fig. 33) to one synchronized flow speed for each vehicle density. As a result, the averaged steady states of synchronized flow are related to 1D region, i. e., to a curve in the flow-density plane (curves S in Fig. 34). In this case, we should find a vehicle dynamics for a traffic flow model whose steady states are associated with such a fundamental diagram, however, the model is able to show and predict the existence of the free flow, synchronized flow, and wide moving jam phases, the $\mathrm{F} \to \mathrm{S} \to \mathrm{J}$ transitions between them including a double Z-characteristic for the phase transitions (Sects. "Traffic Breakdown and Highway Capacity" and "Moving Jam Emergence in Synchronized Flow ($\mathrm{S} \to \mathrm{J}$ Transition)").

Recall that three-phase traffic theory is a qualitative theory. For this reason, the fundamental hypothesis of this theory is associated with features of hypothetical steady states of synchronized flow that should cover a 2D region in the flow-density plane (Fig. 15a). Rather than a hypothesis about model steady states of synchronized flow, the fundamental hypothesis of a mathematical three-phase traffic flow model can be a hypothesis about some specific dynamic model features (for more detail, see footnote 4 of Sect. 4.3.4 in [52]).

Although such three-phase traffic flow models are possible to develop, however, through the replacing of a 2D-region for steady states of synchronized flow by a 1D-region, i. e., by a curve in the flow-density plane, these three-phase traffic flow models lose a possibility of description of very important features of synchronized flow found in empirical observations. This emphasizes the sense and importance of 2D steady states of synchronized flow for the development of a three-phase traffic flow model. We illustrate these general conclusions through a brief discussion of speed adaptation (SA) three-phase traffic flow models [62].

In contrast with the Kerner–Klenov model discussed in Sect. "Kerner–Klenov Stochastic Three-Phase Traffic Flow Model", which incorporates the fundamental hypothesis of three-phase traffic theory (Fig. 33a), in the SA models hypothetical steady states of synchronized flow are associated with a curve (curve S in Fig. 34), i. e., they cover a 1D region in the flow-density plane. As mentioned, the basis hypothesis of the SA models is associated with the hypothesis of three-phase traffic theory about a double Z-characteristic for the sequence of $\mathrm{F} \to \mathrm{S} \to \mathrm{J}$ transitions. Based on a mathematical implementation of this hypothesis onto the SA models, the models can show and explain both traffic breakdown ($\mathrm{F} \to \mathrm{S}$ transition) and moving jam emergence within synchronized flow (pinch effect and associated $\mathrm{S} \to \mathrm{J}$ transitions) [62] as found in empirical observations [50,52].

An $\mathrm{F} \to \mathrm{S}$ transition, i. e., traffic breakdown is simulated in the SA models through the use of one of the following features of steady speed states on the fundamental diagram:

(i) A *discontinuity* of steady speed solutions between free flow and synchronized flow states (Fig. 34a, b). This discontinuity determines the maximum (limit) point of free flow $(v_{\min}^{(\mathrm{free})}, \rho_{\max}^{(\mathrm{free})})$ on the fundamental diagram, i. e., at the densities $\rho \geq \rho_{\max}^{(\mathrm{free})}$ an $\mathrm{F} \to \mathrm{S}$ transition should occur.

(ii) An *instability* of steady speed solutions within a density range $(\rho_{\max}^{(\mathrm{free})}, \rho_{\min}^{(\mathrm{syn})})$ (curve FS in Fig. 34c). This instability determines also the maximum (limit) point of free flow $(v_{\min}^{(\mathrm{free})}, \rho_{\max}^{(\mathrm{free})})$ on the fundamental dia-

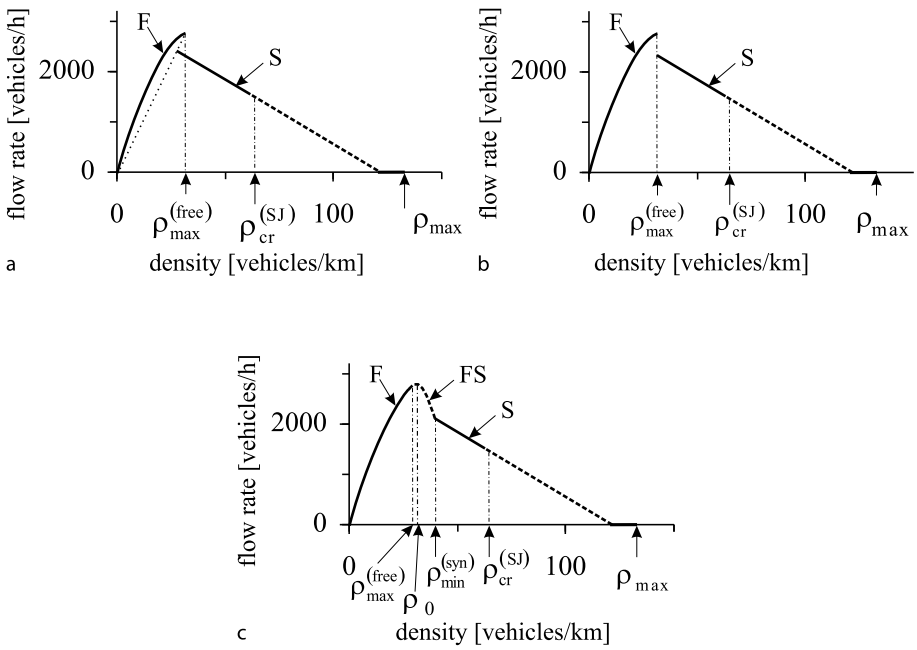

Traffic Congestion, Modeling Approaches to, Figure 34
Steady states for three different SA three-phase traffic flow models with discontinuity (a,b) and instability (c) of steady speed solutions between free flow (*curves F*) and synchronized flow states (*curves S*). Taken from [62]

gram, i. e., at the densities $\rho \geq \rho_{max}^{(free)}$ an F → S transition should occur.

A first-order F → S transition is simulated through the use of the above mentioned discontinuity or instability of steady speed solutions on the fundamental diagram (item (i) and (ii)) as well as a feature of the vehicle dynamics incorporated into the SA models that a vehicle tries to adjust the speed to the speed of the preceding vehicle while approaching synchronized flow states; this simulates the speed adaptation effect in synchronized flow (Sect. "Traffic Breakdown Explanation Through Fundamental Hypothesis of Three-Phase Traffic Theory").

The pinch effect and S → J transitions are simulated in the SA models through an instability of those synchronized flow model steady states whose speeds are lower than some critical speed $v_{cr}^{(SJ)}$, i. e., when synchronized flow density is greater than some critical density (denoted by $\rho_{cr}^{(SJ)}$ in Fig. 34): at each density $\rho > \rho_{cr}^{(SJ)}$ the synchronized flow state is unstable with respect to infinitesimal disturbances (dashed parts of curves *S* in Fig. 34).

It must be noted that rather than the fundamental diagram choice is associated with some empirical fundamental diagrams, in the SA models, to simulate the F → S and S → J transitions, which occur at *different* freeway locations, we chose some artificial features of steady states on fundamental diagrams (Fig. 34). Thus rather than validate these fundamental diagrams with empirical data measured at a single freeway location, the characteristic variables on these diagrams like $\rho_{max}^{(free)}$ and $\rho_{cr}^{(SJ)}$ should be validated based on spatiotemporal empirical features of the F → S → J transitions.

The SA models allow us to overcome drawbacks of earlier traffic flow models in the framework of the fundamental diagram approach in description of traffic breakdown and the pinch effect discussed in Sect. "Critical Discussion of Fundamental Diagram Modeling Approach to Traffic Congestion". This is because of the formulation of such a vehicle dynamics in the SA-models that leads to two *separated regions for phase transitions* and associated *two different critical densities* $\rho_{max}^{(free)}$ for traffic breakdown and $\rho_{cr}^{(SJ)}$ for S → J transitions on the fundamental diagram (Fig. 34). For these reasons, the SA models are able to show and predict the sequence of F → S → J transitions and the associated double *Z*-characteristic.

On the other hand, due to the averaging of an infinite number of steady states of synchronized flow to one synchronized flow speed, the SA models are not able to show important features of synchronized flow as well as many features of coexistence of the free flow, synchronized flow,

and wide moving jam traffic phases found in empirical observations [62]. These drawbacks of the SA models are associated with the ignoring of the fundamental hypothesis of three-phase traffic theory.

One of these drawbacks of the SA models is illustrated in Fig. 35. In accordance with empirical results [52], three-phase traffic flow models with 2D-region of steady states for synchronized flow can show and predict *both* free flow *and* synchronized flow between wide moving jams (Fig. 35c, d). In contrast, for the SA models, whose fundamental diagrams are shown in Fig. 34, *only* free flow can occur between wide moving jams, when the jams are moving at large enough distances each other (Fig. 35a, b).

To explain this, note that if a three-phase model exhibits a 2D region of steady states for synchronized flow (Fig. 36a), in the wide moving jam outflow either a state of free flow with the density $\rho_{\min}$ or any state of synchronized flow from infinite synchronized flow states that lie on the line J is possible. In contrast, there is only an unstable steady state of synchronized flow that lies on the line J in the case of the SA-models (Fig. 36b). For this reason, when distances between wide moving jams are great enough, free flow can be formed in the jam outflows only (Fig. 35a,b).

We can conclude that the fundamental hypothesis of three-phase traffic theory about a 2D-region for steady states of synchronized flow permits infinite synchronized flow states on, above, and below the line J. This allows us to simulate different synchronized flow states in a diverse variety of combinations with wide moving jams and free flows, as observed in empirical data. This is not possible to do based on the SA-models with a fundamental diagram for steady states of synchronized flow.

Conclusions

1. Freeway traffic flow can be free or congested. There are two phases in congested traffic, synchronized flow and wide moving jams. Thus there are three traffic phases: 1. Free flow. 2. Synchronized flow. 3. Wide moving jam. The synchronized flow and wide moving jam traffic phases are defined through spatiotemporal empiri-

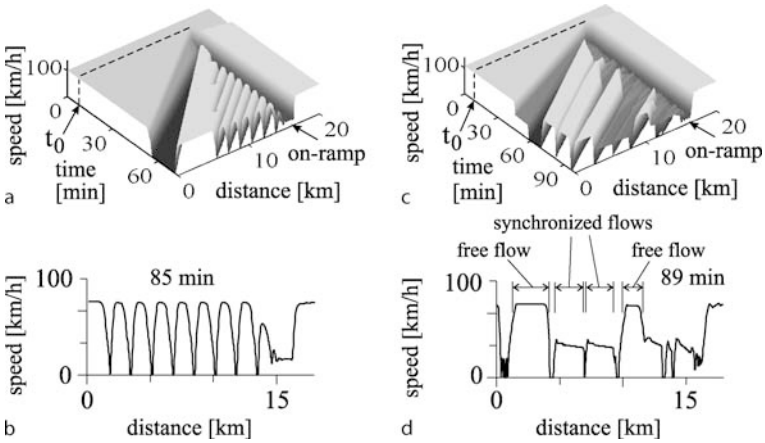

Traffic Congestion, Modeling Approaches to, Figure 35
Explanation of drawbacks of SA models: a,b In the SA models (Fig. 34), only free flow is formed between wide moving jams. c,d In a three-phase traffic flow model with 2D-region of steady states for synchronized flow both free and synchronized flows are possible between wide moving jams. Taken from [62]

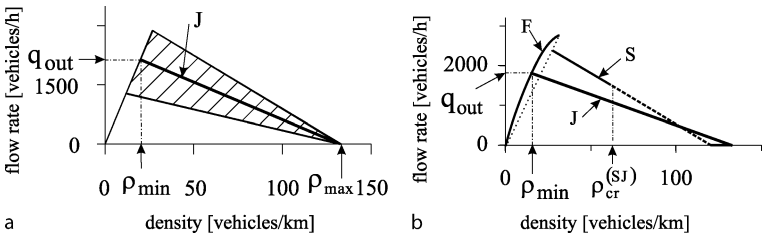

Traffic Congestion, Modeling Approaches to, Figure 36
Steady states for synchronized flow and line *J* for two classes of three-phase traffic flow models

cal criteria associated with propagation characteristics of the downstream fronts of these phases.

2. Empirical traffic breakdown at a freeway bottleneck, i. e., the onset of congestion in an initial free flow at the bottleneck is associated with a phase transition from free flow to synchronized flow ($F \rightarrow S$ transition). An $F \rightarrow S$ transition at the bottleneck can be spontaneous or induced.
3. Wide moving jams can emerge in synchronized flow ($S \rightarrow J$ transition). Thus wide moving jams emerge spontaneously due to a sequence of the $F \rightarrow S \rightarrow J$ transitions associated with a double *Z*-characteristic for these phase transitions. Wide moving jams do not emerge spontaneously in free flow, i. e., spontaneous $F \rightarrow J$ transitions are not observed in real traffic flow.
4. The most of traffic flow models used by traffic researches are based on the fundamental diagram hypothesis. These models can be classified into two main classes referred to the classic LWR-theory and to the classic GM-model. None of the traffic flow models reviewed in [16,18,21,33,35,37,40,84,92,93,95,98,126,127,128] can explain and predict traffic breakdown at the bottleneck as is observed in empirical data. These traffic flow models are basic traffic flow models for simulations of freeway control and management strategies. However, we have to conclude that the related simulations of the control and management strategies cannot predict many of the freeway traffic phenomena that would occur through the use of a simulated management strategy.
5. In contrast with the traffic flow models and theories reviewed in [16,18,21,33,35,37,40,84,92,93,95,98,126,127,128], three-phase traffic theory can explain the empirical features of traffic breakdown and resulting congested patterns. Thus traffic flow models in the framework of this theory can be used for simulations of control and management strategies in freeway traffic before they are introduced to the market.

Future Directions

There are still many "black areas" in understanding empirical *microscopic* features of synchronized flow and phase transitions in freeway traffic. Traffic flow models in the framework of three-phase traffic theory are only at the beginning of their development. There are almost no analytical results within the context of this theory (the exception is a probabilistic theory of traffic breakdown at a bottleneck (see ► Traffic Breakdown, Probabilistic Theory of)). These and many other unsolved problems are an interesting and important field of the future empirical and theoretical investigations.

Although the FOTO and ASDA models for tracking of spatiotemporal congested patterns based on three-phase traffic theory are already successfully used in on-line installations and there are ideas about applications of the theory for feedback on-ramp metering, speed limit control and advance driver assistant systems in vehicles, many other applications for control and management of traffic networks could be expected.

Cross References

► Freeway Traffic Management and Control

Acknowledgments

I would like to thank Sergey Klenov, Andreas Hiller, Hubert Rehborn, Mario Aleksić, Ines Maiwald–Hiller and Olivia Brickley for help and useful suggestions.

Bibliography

1. Ahn S, Cassidy MJ (2007) Freeway traffic oscillations and vehicle lane-change maneuvers. In: Allsop RE, Bell MGH, Hydecker BG (eds) Transportation and Traffic Theory 2007. Elsevier, Amsterdam, pp 691–710
2. Aw A, Rascle M (2000) Resurrection of "Second Order" models of traffic flow. SIAM J Appl Math 60:916–938
3. Bando M, Hasebe K, Nakayama A, Shibata A, Sugiyama Y (1994) Structure stability of congestion in traffic dynamics. Jpn J Appl Math 11:203–223
4. Bando M, Hasebe K, Nakayama A, Shibata A, Sugiyama Y (1995) Dynamical model of traffic congestion and numerical simulation. Phys Rev E 51:1035–1042
5. Bando M, Hasebe K, Nakayama A, Shibata A, Sugiyama Y (1995) Phenomenological study of dynamical model of traffic flow. J Phys I France 5:1389–1399
6. Barlović R, Santen L, Schadschneider A, Schreckenberg M (1998) Metastable states in cellular automata for traffic flow. Eur Phys J B 5:793–800
7. Bellomo N, Coscia V, Delitala M (2002) On the mathematical theory of vehicular traffic flow I. Fluid dynamic and kinetic modelling. Math Mod Meth App Sci 12:1801–1843
8. Berg P, Woods A (2001) On-ramp simulations and solitary waves of a car-following model. Phys Rev E 64:035602(R)
9. Bovy PHL (ed) (1998) Motorway analysis: new methodologies and recent empirical findings. Delft University Press, Delft
10. Brilon W, Geistefeld J, Regler M (2005) Reliability of freeway traffic flow: a stochastic concept of capacity. In: Mahamassani HS (ed) Proc of the 16th inter sym on transportation and traffic theory. Elsevier, Amsterdam, pp 125–144
11. Brilon W, Zurlinden H (2004) Kapazität von Straßen als Zufallsgröße. Straßenverkehrstechnik 4:164
12. Brilon W, Regler M, Geistefeld J (2005) Zufallscharakter der Kapazität von Autobahnen und praktische Konsequenzen – Teil 1. Straßenverkehrstechnik 3:136

13. Brockfeld E, Kühne RD, Skabardonis A, Wagner P (2003) Toward benchmarking of microscopic traffic flow models. Trans Res Rec 1852:124–129
14. Brockfeld E, Kühne RD, Wagner P (2005) Calibration and validation of simulation models. In: Proc of the transportation research board 84th annual meeting, TRB Paper No. 05-2152. TRB, Washington DC
15. Ceder A (ed) (1999) Transportation and traffic theory. Proc of the 14th international symposium on transportation and traffic theory, Elsevier, Oxford
16. Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. Phys Rep 329:199
17. Colombo RM (2003) Hyperbolic Phase Transitions in Traffic Flow. SIAM J Appl Math 63:708–721
18. Cremer M (1979) Der Verkehrsfluss auf Schnellstrassen. Springer, Berlin
19. Cowan RJ (1976) Useful headway models. Trans Rec 9:371–375
20. Daganzo CF (1993) The cell-transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. Trans Res B 28:269–287
21. Daganzo CF (1997) Fundamentals of transportation and traffic operations. Elsevier, New York
22. Davis LC (2004) Multilane simulations of traffic phases. Phys Rev E 69:016108
23. Davis LC (2006) Controlling traffic flow near the transition to the synchronous flow phase. Physica A 368:541–550
24. Davis LC (2006) Effect of cooperative merging on the synchronous flow phase of traffic. Physica A 361:606–618
25. Davis LC (2007) Effect of adaptive cruise control systems on mixed traffic flow near an on-ramp. Physica A 379:274–290
26. Edie LC, Foote RS (1958) Traffic flow in tunnels. Highw Res Board Proc Ann Meet 37:334–344
27. Edie LC, Foote RS (1960) Effect of shock waves on tunnel traffic flow. In: Highway Research Board Proceedings, vol 39. National Research Council, Washington DC, pp 492–505
28. Edie LC (1961) Car-following and steady state theory for non-congested traffic. Oper Res 9:66–77
29. Edie LC, Herman R, Lam TN (1980) Observed multilane speed distribution and the kinetic theory of vehicular traffic. Trans Sci 14:55–76
30. Elefteriadou L, Roess RP, McShane WR (1995) Probabilistic nature of breakdown at freeway merge junctions. Trans Res Rec 1484:80–89
31. Fukui M, Sugiyama Y, Schreckenberg M, Wolf DE (eds) (2003) Traffic and Granular Flow' 01. Springer, Heidelberg
32. Gao K, Jiang R, Hu S-X, Wang B-H, Wu Q-S (2007) Cellular-automaton model with velocity adaptation in the framework of Kerner's three-phase traffic theory. Phys Rev E 76:026105
33. Gartner NH, Messer CJ, Rathi A (eds) (1997) Special report 165: revised monograph on traffic flow theory. Transportation Research Board, Washington DC
34. Gazis DC, Herman R, Rothery RW (1961) Nonlinear follow-the-leader models of traffic flow. Oper Res 9:545–567
35. Gazis DC (2002) Traffic theory. Springer, Berlin
36. Gipps PG (1981) Behavioral car-following model for computer simulation. Trans Res B 15:105–111
37. Haight FA (1963) Mathematical theories of traffic flow. Academic Press, New York
38. Hall FL, Agyemang-Duah K (1991) Freeway capacity drop and the definition of capacity. Trans Res Rec 1320:91–98
39. Hall FL, Hurdle VF, Banks JH (1992) Synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relationships on freeways. Trans Res Rec 1365:12–18
40. Helbing D (2001) Traffic and related self-driven many-particle systems. Rev Mod Phys 73:1067–1141
41. Helbing D, Hennecke A, Treiber M (1999) Phase diagram of traffic states in the presence of inhomogeneities. Phys Rev Lett 82:4360–4363
42. Helbing D, Herrmann HJ, Schreckenberg M, Wolf DE (eds) (2000) Traffic and Granular Flow' 99. Springer, Heidelberg
43. Herman R, Montroll EW, Potts RB, Rothery RW (1959) Traffic dynamics: analysis of stability in car following. Oper Res 7: 86–106
44. Hoogendoorn SP, Luding S, Bovy PHL, Schreckenberg M, Wolf DE (eds) (2005) Traffic and Granular Flow' 03. Springer, Heidelberg
45. Jiang R, Wu QS (2004) Spatial-temporal patterns at an isolated on-ramp in a new cellular automata model based on three-phase traffic theory. J Phys A Math Gen 37:8197–8213
46. Jiang R, Wu QS (2005) Toward an improvement over Kerner–Klenov–Wolf three-phase cellular automaton model. Phys Rev E 72:067103
47. Jiang R, Wu QS (2007) Dangerous situations in a synchronized flow model. Physica A 377:633–640
48. Jiang R, Hua M-B, Wang R, Wu Q-S (2007) Spatiotemporal congested traffic patterns in macroscopic version of the Kerner–Klenov speed adaptation model. Phys Lett A 365:6–9
49. Kerner BS (1998) Theory of congested traffic flow. In: Rysgaard R (ed) Proc of the 3rd symposium on highway capacity and level of service, vol 2. Road Directorate, Ministry of Transport, Denmark, pp 621–642
50. Kerner BS (1998) Empirical features of self-organization in traffic flow. Phys Rev Lett 81:3797–3400
51. Kerner BS (1999) Congested traffic flow: observations and theory. Trans Res Rec 1678:160–167
52. Kerner BS (2004) The physics of traffic. Springer, Berlin
53. Kerner BS (2007) On-ramp metering based on three-phase traffic theory I. Traffic Eng Control 48:28–35
54. Kerner BS (2007) Study of freeway speed limit control based on three-phase traffic theory. Trans Res Rec 1999:30–39
55. Kerner BS (2008) A theory of traffic congestion at heavy bottlenecks. J Phys A Math Theor 41:215101
56. Kerner BS (2008) Three-phase traffic theory and its applications for freeway traffic control. In: Inweldi PO (ed) Transportation research trends. Nova Science Publishers, New York, pp 1–93
57. Kerner BS, Klenov SL (2002) A microscopic model for phase transitions in traffic flow. J Phys A Math Gen 35:L31–L43
58. Kerner BS, Klenov SL (2003) Microscopic theory of spatio-temporal congested traffic patterns at highway bottlenecks. Phys Rev E 68:036130
59. Kerner BS, Klenov SL (2005) Probabilistic breakdown phenomenon at on-ramps bottlenecks in three-phase traffic theory. cond-mat/0502281. e-print in http://arxiv.org/abs/cond-mat/0502281
60. Kerner BS, Klenov SL (2006) Probabilistic breakdown phenomenon at on-ramp bottlenecks in three-phase traffic theory: congestion nucleation in spatially non-homogeneous traffic. Physica A 364:473–492

61. Kerner BS, Klenov SL (2006) Probabilistic breakdown phenomenon at on-ramp bottlenecks in three-phase traffic theory. Trans Res Rec 1965:70–78
62. Kerner BS, Klenov SL (2006) Deterministic microscopic three-phase traffic flow models. J Phys A Math Gen 39:1775–1809
63. Kerner BS, Konhäuser P (1994) Structure and parameters of clusters in traffic flow. Phys Rev E 50:54–83
64. Kerner BS, Konhäuser P, Schilke M (1995) Deterministic spontaneous appearance of traffic jams in slightly inhomogeneous traffic flow. Phys Rev E 51:6243–6246
65. Kerner BS, Konhäuser P, Schilke M (1996) "Dipole-layer" effect in dense traffic flow. Phys Lett A 215:45–56
66. Kerner BS, Klenov SL, Wolf DE (2002) Cellular automata approach to three-phase traffic theory. J Phys A Math Gen 35:9971–10013
67. Kerner BS, Klenov SL, Hiller A, Rehborn H (2006) Microscopic features of moving traffic jams. Phys Rev E 73:046107
68. Kerner BS, Klenov SL, Hiller A (2006) Criterion for traffic phases in single vehicle data and empirical test of a microscopic three-phase traffic theory. J Phys A Math Gen 39:2001–2020
69. Kerner BS, Klenov SL, Hiller A (2007) Empirical test of a microscopic three-phase traffic theory. Non Dyn 49:525–553
70. Knospe W, Santen L, Schadschneider A, Schreckenberg M (2000) Towards a realistic microscopic description of highway traffic. J Phys A Math Gen 33:L477–L485
71. Knospe W, Santen L, Schadschneider A, Schreckenberg M (2002) Single-vehicle data of highway traffic: microscopic description of traffic phases. Phys Rev E 65:056133
72. Knospe W, Santen L, Schadschneider A, Schreckenberg M (2004) Empirical test for cellular automaton models of traffic flow. Phys Rev E 70:016115
73. Kometani E, Sasaki T (1958) J Oper Res Soc Jap 2:11
74. Kometani E, Sasaki T (1959) A safety index for traffic with linear spacing. Oper Res 7:704–720
75. Koshi M (2003) An interpretation of a traffic engineer on vehicular traffic flow. In: Fukui M, Sugiyama Y, Schreckenberg M, Wolf DE (eds) Traffic and Granular Flow' 01. Springer, Heidelberg, pp 199–210
76. Koshi M, Iwasaki M, Ohkura I (1983) Some findings and an overview on vehiclular flow characteristics. In: Hurdle VF (ed) Proc 8th international symposium on transportation and traffic theory. University of Toronto Press, Toronto, pp 403
77. Krauß S, Wagner P, Gawron C (1997) Metastable states in a microscopic model of traffic flow. Phys Rev E 55:5597–5602
78. Kühne R (1991) In: Brannolte U (ed) Highway capacity and level of service. A.A. Balkema, Rotterdam, pp 211
79. Kühne R, Mahnke R, Lubashevsky I, Kaupužs J (2002) Probabilistic description of traffic breakdown. Phys Rev E 65:066125
80. Laval JA (2007) Linking synchronized flow and kinematic waves. In: Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) Proc of the international workshop on traffic and granular flow. Springer, Berlin, pp 521–526
81. Lee HK, Barlović R, Schreckenberg M, Kim D (2004) Mechanical restriction versus human overreaction triggering congested traffic states. Phys Rev Lett 92:238702
82. Lee HY, Lee H-W, Kim D (1999) Dynamic states of a continuum traffic equation with on-ramp. Phys Rev E 59:5101–5111
83. Lesort J-B (ed) (1996) Transportation and traffic theory. Proc of the 13th international symposium on transportation and traffic theory. Elsevier, Oxford
84. Leutzbach W (1988) Introduction to the theory of traffic flow. Springer, Berlin
85. Li XG, Gao ZY, Li KP, Zhao XM (2007) Relationship between microscopic dynamics in traffic flow and complexity in networks. Phys Rev E 76:016110
86. Lighthill MJ, Whitham GB (1955) On kinematic waves. I Flow movement in long rives. II A theory of traffic flow on long crowded roads. Proc Roy Soc A 229:281–345
87. Lorenz M, Elefteriadou L (2000) A probabilistic approach to defining freeway capacity and breakdown. Trans Res Cir E-C018:84–95
88. Maerivoet S, De Moor B (2005) Cellular automata models of road traffic. Phys Rep 419:1–64
89. Mahmassani HS (ed) (2005) Transportation and traffic theory. Proc of the 16th inter sym on transportation and traffic theory. Elsevier, Amsterdam
90. Mahnke R, Kaupužs J (1999) Stochastic theory of freeway traffic. Phys Rev E 59:117–125
91. Mahnke R, Pieret N (1997) Stochastic master-equation approach to aggregation in freeway traffic. Phys Rev E 56:2666–2671
92. Mahnke R, Kaupužs J, Lubashevsky I (2005) Probabilistic description of traffic flow. Phys Rep 408:1–130
93. May AD (1990) Traffic flow fundamentals. Prentice-Hall, New Jersey
94. Okamura H, Watanabe S, Watanabe T (2000) An empirical study of the capacity of bottlenecks on the basic suburban Expressway sections in Japan. TRB Circular EC 018, Transportation Research Board, Washington DC
95. Nagatani T (2002) The physics of traffic jams. Rep Prog Phys 65:1331–1386
96. Nagatani T, Nakanishi K (1998) Delay effect on phase transitions in traffic dynamics. Phys Rev E 57:6415–6421
97. Nagel K, Schreckenberg M (1992) A cellular automaton model for freeway traffic. J Phys (France) I 2:2221–2229
98. Nagel K, Wagner P, Woesler R (2003) Still flowing: approaches to traffic flow and traffic jam modeling. Oper Res 51:681–716
99. Neubert L, Santen L, Schadschneider A, Schreckenberg M (1999) Single-vehicle data of highway traffic: a statistical analysis. Phys Rev E 60:6480–6490
100. Newell GF (1961) Nonlinear effects in the dynamics of car following. Oper Res 9:209–229
101. Newell GF (1982) Applications of queuing theory. Chapman Hall, London
102. Papageorgiou M (1983) Application of automatic control concepts in traffic flow modeling and control. Springer, Berlin
103. Payne HJ (1971) Models of freeway traffic and control. In: Bekey GA (ed) Mathematical models of public systems, vol 1. Simulation Council, La Jolla
104. Payne HJ (1979) Trans Res Rec 772:68
105. Persaud BN, Yagar S, Brownlee R (1998) Exploration of the breakdown phenomenon in freeway traffic. Trans Res Rec 1634:64–69
106. Pipes LA (1953) An operational analysis of traffic dynamics. J Appl Phys 24:274–287
107. Pottmeier A, Thiemann C, Schadschneider A, Schreckenberg M (2007) Mechanical restriction versus human overreaction: accident avoidance and two-lane simulations. In: Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) Proc of the international workshop on traffic and granular flow. Springer, Berlin, pp 503–508

108. Prigogine I, Herman R (1971) Kinetic theory of vehicular traffic. Elsevier, New York
109. Richards PI (1956) Shockwaves on the highway. Oper Res 4:42–51
110. Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) (2007) Traffic and Granular Flow' 05. In: Proc of the international workshop on traffic and granular flow. Springer, Berlin
111. Schönhof M, Helbing D (2007) Empirical features of congested traffic states and their implications for traffic modelling. Trans Sci 41:135–166
112. Schreckenberg M, Wolf DE (eds) (1998) Traffic and Granular Flow' 97. In: Proc of the international workshop on traffic and granular flow. Springer, Singapore
113. Siebel F, Mauser W (2006) Synchronized flow and wide moving jams from balanced vehicular traffic. Phys Rev E 73:066108
114. Stokes EE (1848) On a difficulty in the theory of sound. Phil Mag 33:349–356
115. Takayasu M, Takayasu H (1993) Phase transition and 1/f type noise in one dimensional asymmetric particle dynamics. Fractals 1:860–866
116. Tanga CF, Jiang R, Wu QS (2007) Phase diagram of speed gradient model with an on-ramp. Physica A 377:641–650
117. Taylor MAP (ed) (2002) Transportation and traffic theory in the 21st century. Proc of the 15th international symposium on transportation and traffic theory. Elsevier, Amsterdam
118. Tilch B, Helbing D (2000) Evaluation of single vehicle data in dependence of the vehicle-type, lane, and site. In: Helbing D, Herrmann HJ, Schreckenberg M, Wolf DE (eds) Traffic and Granular Flow' 99. Springer, Heidelberg, pp 333–338
119. Treiber M, Hennecke A, Helbing D (2000) Congested traffic states in empirical observations and microscopic simulations. Phys Rev E 62:1805–1824
120. Treiterer J (1967) Improvement of traffic flow and safety by longitudinal control. Trans Res 1:231–251
121. Treiterer J (1975) Investigation of traffic dynamics by aerial photogrammetry techniques. Ohio State University Technical Report PB 246 094. Columbus, Ohio
122. Treiterer J, Taylor JI (1966) Traffic flow investigations by photogrammetric techniques. Highw Res Rec 142:1–12
123. Treiterer J, Myers JA (1974) The hysteresis phenomenon in traffic flow. In: Buckley DJ (ed) Proc 6th international symposium on transportation and traffic theory. A.H. & AW Reed, London, pp 13–38
124. Wang R, Jiang R, Wu QS, Liu M (2007) Synchronized flow and phase separations in single-lane mixed traffic flow. Physica A 378:475–484
125. Wang Y, Papageorgiou M, Messmer A (2004) Predictive feedback routing control strategy for freeway network traffic [E-text type]. In: Proc of the 83rd Annual Transportation Research Board Meeting, TRB Paper No. 04-3429. TRB, Washington DC
126. Whitham GB (1974) Linear and nonlinear waves. Wiley, New York
127. Wiedemann R (1974) Simulation des Verkehrsflusses. University of Karlsruhe, Karlsruhe
128. Wolf DE (1999) Cellular automata for traffic simulations. Physica A 263:438–451
129. Wolf DE, Schreckenberg M, Bachem A (eds) (1995) Traffic and Granular Flow. Proc of the international workshop on traffic and granular flow. World Scientific, Singapore
130. Zhang P, Wong SC (2006) Essence of conservation forms in the traveling wave solutions of higher-order traffic flow models. Phys Rev E 74:026109
131. Zurlinden H (2003) Ganzjahresanalyse des Verkehrsflusses auf Straßen. In: Schriftenreihe des Lehrstuhls für Verkehrswesen der Ruhr-Universität Bochum, vol 26. Ruhr-Universität Bochum, Bochum

Traffic Congestion, Spatiotemporal Features of

Boris S. Kerner
GR/ETI, HPC: G021, Daimler AG, Sindelfingen, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Congested Patterns on Homogeneous Road
Congested Patterns at Isolated Bottlenecks
Diagram of Congested Patterns at Isolated Bottlenecks
Complex Congested Patterns and Pattern Interaction
Reproducible and Predictable Congested Patterns
Microscopic Features of Traffic Phases
Congested Patterns at Heavy Bottlenecks
Conclusions. Fundamental Empirical Features of Spatiotemporal Congested Freeway Traffic Patterns
Future Directions
Acknowledgments
Bibliography

Glossary

Free flow Free flow is usually observed, when the vehicle density in traffic is small enough. The flow rate increases in free flow with increase in vehicle density, whereas the average vehicle speed is a decreasing density function.

Congested traffic Congested traffic is defined as a state of traffic in which the average speed is *lower* than the minimum average speed that is possible in free flow. In empirical observations, congested traffic occurs mostly at a freeway bottleneck; the average speed decreases sharply in an initial free flow to a lower speed at the bottleneck. This speed breakdown observed during the onset of congestion is called the breakdown phenomenon or traffic breakdown. The bottleneck can be a result of road works, on- and off-ramps, a decrease

in the number of freeway lanes, road curves and road gradients, bad weather conditions, accidents, etc.

Effectual bottleneck An effectual bottleneck is a bottleneck at which traffic breakdown is observed.

Three traffic phases In three-phase traffic theory, there are the three traffic phases:

1. Free flow.
2. Synchronized flow.
3. Wide moving jam.

The wide moving jam and synchronized flow traffic phases associated with congested traffic are defined through the spatiotemporal macroscopic empirical (objective) criteria [J] and [S]. The definition [J]: A wide moving jam is a moving jam that maintains the mean velocity of the downstream jam front, even when the jam propagates through any other traffic states or freeway bottlenecks. The definition [S]: In contrast with the wide moving jam phase, the downstream front of the synchronized flow phase does not exhibit the wide moving jam characteristic feature; in particular, the downstream front of the synchronized flow phase is often fixed at a bottleneck.

Traffic breakdown is associated with a local phase transition from free flow to synchronized flow. Wide moving jams can emerge spontaneously in synchronized flow only; the jams result from growth of narrow moving jams that occur in the synchronized flow.

Synchronized flow patterns A synchronized flow pattern (SP) is a congested traffic pattern that consists of the synchronized flow phase only.

Pinch effect in synchronized flow The pinch effect in synchronized flow is the effect of spontaneous occurrence and growth of narrow moving jams in synchronized flow. Within the associated pinch region of synchronized flow a self-compression of synchronized flow is usually observed.

General congested patterns A general pattern (GP) is a congested traffic pattern at a hypothetical isolated bottleneck that consists of both traffic phases in congested traffic – synchronized flow and wide moving jam. Firstly, synchronized flow occurs at the bottleneck. Due to the pinch effect in this synchronized flow, later and upstream of the bottleneck narrow moving jams emerge spontaneously; jam amplitudes grow while the jams propagate upstream within the pinch region. Some of the jams (or each of them) transform into wide moving jams at the upstream boundary of the pinch region of synchronized flow. Thus GP structure consists of the pinch region of synchronized flow and a sequence of wide moving jams upstream.

Diagram of congested patterns The diagram of congested traffic patterns at a hypothetical isolated bottleneck shows regions of congested pattern existence and excitation depending on traffic demand and bottleneck characteristics.

Expanded congested patterns An expanded congested pattern (EP) is a congested traffic pattern whose synchronized flow affects two or more adjacent freeway bottlenecks.

Microscopic structure of wide moving jam The wide moving jam structure consists of alternations of regions in which traffic flow is interrupted and flow states of low speeds associated with moving blanks within the jam. The flow interruption intervals within the jam occur when some of the vehicles within the jam do not move.

Moving blanks within wide moving jam When vehicles meet a wide moving jam, they come to a stop at very different blank spaces (blanks for short) to each other. Later vehicles begin to move within the jam covering these blanks. Due to this low speed vehicle motion new blanks between vehicles occur upstream within the jam. A moving blank is a blank between vehicles within a wide moving jam, which moves upstream due to low speed vehicle motion within the jam.

Heavy bottleneck A heavy bottleneck is a bottleneck with a great bottleneck strength, i. e., a great bottleneck influence on traffic flow that limits the average flow rate within a congested pattern occurring upstream of the bottleneck to very small values. A heavy bottleneck can occur due to e. g., bad weather conditions, accidents, or heavy roadworks.

Traffic congestion at heavy bottleneck Traffic congestion at a heavy bottleneck can be associated with the phenomena of random disappearance and appearance of the pinch region over time and a complex non-regular dynamics of wide moving jams upstream of the pinch region as well as with the phenomenon of the merger of wide moving jams into a mega-wide moving jam (mega-jam). When the bottleneck strength increases strongly, the pinch region disappears and only the mega-jam survives and synchronized flow remains only within its downstream front separating free flow and congested traffic.

Definition of the Subject

Empirical observations show that traffic congestion exhibits extremely complex spatiotemporal features. Spatiotemporal features of traffic congestion are as follows:

(i) Diverse variety of spatiotemporal congested traffic patterns measured at a freeway bottleneck.
(ii) Complex evolution of these congested patterns in space and time that occurs when traffic demand or/and bottleneck characteristics change.
(iii) Complex spatiotemporal phenomena associated with congested patterns occurring at two or more adjacent freeway bottlenecks.
(iv) Transformations between various congested patterns that occurs due to phase transitions between different traffic phases within the pattern.
(v) Diverse microscopic characteristics of traffic congestion associated with complex driver behavior within congested patterns.
(vi) Complex non-regular dynamics of wide moving jams that can occur when a heavy bottleneck appears on a road.

Traffic congestion is a fact of life for many car drivers. For this reason, one of the aims of transportation and traffic research is to provide an understanding of freeway traffic congestion that can be used for effective traffic management, traffic control, organization, traffic assignment and other engineering applications, which should improve traffic safety and result in high-quality mobility.

Moreover, due to the very complex non-linear character of freeway traffic observed in empirical data, analysis of complexity in spatiotemporal traffic phenomena is needed for development of modeling approaches, which can explain these traffic phenomena and, therefore, are applicable for traffic prediction, traffic control, organization, traffic assignment and other traffic engineering applications.

Introduction

There are a huge number of publications devoted to empirical and experimental investigations of spatiotemporal features of vehicular traffic congestion (e. g., [9,10,11,12,51,68,69,70] and references in [2,17,20,38,54,57,58,59]). However, the puzzle of empirical spatiotemporal features of traffic congestion has been solved only recently [30,36]. Consequently, as has already explained in ► Traffic Congestion, Modeling Approaches to, earlier traffic flow theories and models reviewed in [1,2,3,4,16,17,20,54,56,57,58,59,72,73,74] cannot explain and predict most empirical features of the onset of congestion (traffic breakdown) and many important empirical features of resulting congested patterns (with the exclusion of explanation of wide moving jam propagation; see ► Traffic Congestion, Modeling Approaches to). These traffic flow models are standard ones for validation of freeway traffic control, management and assignment. Thus the related simulations of freeway control and management strategies cannot predict many of the freeway traffic phenomena that would occur through the use of a simulated control strategy.

Therefore, in 1996–1999 the author introduced an alternative traffic flow theory called three-phase traffic theory, which overcomes the disadvantages of all other traffic flow models and theories in the explanation of spatiotemporal empirical features of traffic congestion. For this reason, the three-phase traffic theory is the basis for an explanation of empirical spatiotemporal features of traffic congestion made in this article.

In three-phase traffic theory, besides the free flow traffic phase there are two traffic phases in congested traffic: synchronized flow and wide moving jam. Thus there are three traffic phases in this theory:

1. Free flow.
2. Synchronized flow.
3. Wide moving jam.

These phases are defined through spatiotemporal empirical criteria for these phases. Before we consider the phase definitions, we define the term "moving jam". A moving jam is a localized structure of great vehicle density spatially limited by two jam fronts. Within the downstream jam front vehicles accelerate escaping from the jam; within the upstream jam front, vehicles slow down approaching the jam. Both jam fronts (and, therefore, the jam as a whole structure) propagate upstream in traffic flow. Within the jam (i. e., between the jam fronts) vehicle density is great and speed is very low (sometimes as low as zero).

The definition of the wide moving jam phase [J]: A wide moving jam is a moving jam that maintains the mean velocity of the downstream jam front, even when the jam propagates through any other traffic states or bottlenecks.

The definition of the synchronized flow phase [S]: In contrast with the wide moving jam phase, the downstream front of the synchronized flow phase does not exhibit the wide moving jam characteristic feature; in particular, the downstream front of the synchronized flow phase is often fixed at a bottleneck.

Note that the downstream front of synchronized flow separates synchronized flow upstream from free flow downstream. Within the downstream front of synchronized flow vehicles accelerate from lower speeds in synchronized flow to higher speeds in free flow downstream of the synchronized flow. To apply the traffic phase definitions [J] and [S], spatiotemporal features of congested traffic should be known. The criteria [J] and [S] also mean that if a congested traffic state is not related to the wide moving jam phase, then with certainty the state is asso-

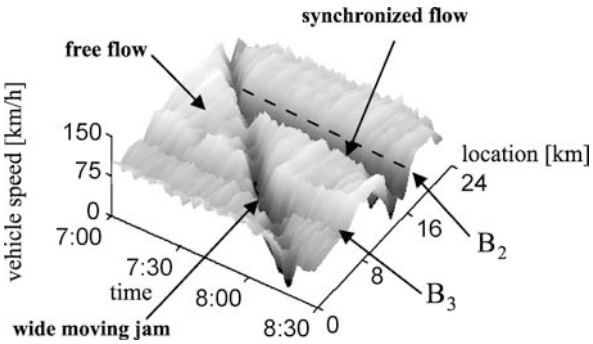

Traffic Congestion, Spatiotemporal Features of, Figure 1
Measured three traffic phases: Average vehicle speed in space and time. Taken from [38]

ciated with the synchronized flow phase. This is because in three-phase traffic theory congested traffic can be either within the synchronized flow phase or within the wide moving jam phase. In other words, if in measured data congested traffic states associated with the wide moving jam phase have been identified, then with certainty all remaining congested states in the data set are related to the synchronized flow phase. For example, the definition [S] means that moving jams, which are caught at the bottleneck, are related to the synchronized flow phase rather than to the wide moving jam phase.

The traffic phase definitions [J] and [S] are illustrated by measured data shown in Fig. 1. There are two qualitatively different types of congested patterns. The first congested pattern propagates through both on-ramp bottlenecks (labeled B_2 and B_3 in Fig. 1) while maintaining the mean velocity of the downstream front of the pattern. In accordance with the definition [J], this congested pattern is associated with the wide moving jam traffic phase in congested traffic. The downstream fronts of another congested pattern is fixed at bottleneck B_2. In accordance with the definition [S], these congested patterns are associated with the synchronized flow traffic phase.

In measured data, the onset of congestion in free flow is accompanied by a sharp decrease in the vehicle speed from a free flow speed to a considerably lower speed in congested traffic and it is called traffic breakdown, or speed breakdown, or else the breakdown phenomenon (see, e. g., [13,18,19,57,61]). As explained in ► Traffic Congestion, Modeling Approaches to, traffic breakdown is a first-order local phase transition from free flow to synchronized flow traffic phase (F → S transition) at the bottleneck, i. e., an F → S transition and traffic breakdown are synonyms associated with the same phenomenon: the onset of congestion in an initial free flow.

In synchronized flow resulting from an F → S transition, narrow moving jams can emerge spontaneously whose subsequent growth can lead to wide moving jam occurrence (S → J transition). The spontaneous emergence of narrow moving jams in synchronized flow with the subsequent jam growth, while the jams propagate upstream within the synchronized flow, is called the pinch effect in synchronized flow. S → J transitions occur usually at other road locations that the road location of the F → S transition resulting in the emergence of synchronized flow. Thus in real free flow, wide moving jams emerge due to a sequence of two phase transitions called F → S → J transitions: firstly, an F → S transition occurs; later and usually at another road location within the emergent synchronized flow an S → J transition is realized ([38], ► Traffic Congestion, Modeling Approaches to).

In ► Traffic Congestion, Modeling Approaches to of this Encyclopedia, the fundamental empirical features of traffic breakdown, a critical discussion of earlier traffic flow modeling approaches to traffic congestion, which cannot explain these features of traffic breakdown, a theory of the pinch effect in synchronized flow, as well as a stochastic traffic flow model in the framework of three-phase traffic theory have been reviewed. The latter model explains and predicts empirical features of traffic breakdown and wide moving jam emergence.

The main focus of this article is a review of empirical *spatiotemporal features* of traffic congested patterns resulting from traffic breakdown as well as an explanation of these empirical traffic patterns. The article is organized as follows. Congested patterns occurring on homogeneous roads (outside or else without highway bottlenecks) are discussed in Sect. "Congested Patterns on Homogeneous Road". In Sect. "Congested Patterns at Isolated Bottlenecks", we consider the main types of congested patterns that occur at isolated highway bottlenecks. A diagram of these patterns at an on-ramp bottleneck, i. e., regions of spontaneous occurrence of various types of congested patterns in the flow–flow plane whose coordinates are the flow rate to the on-ramp and the flow rate in free flow upstream of the bottleneck is discussed in Sect. "Diagram of Congested Patterns at Isolated Bottlenecks". In this section, we will also briefly discuss another congested pattern diagram associated with an off-ramp bottleneck. Complex congested patterns and their interaction, which are observed on real freeways with several adjacent bottlenecks, will be reviewed in Sect. "Complex Congested Patterns and Pattern Interaction". Reproducible and predictable features of spatiotemporal congested traffic patterns are briefly reviewed in Sect. "Reproducible and Pre-

dictable Congested Patterns". In Sect. "Microscopic Features of Traffic Phases", we consider some microscopic features of traffic congestion. In Sect. "Congested Patterns at Heavy Bottlenecks" we discuss features of congested patterns at heavy bottlenecks caused for example by bad weather conditions, accidents, or heavy roadworks. Finally, in Sect. "Conclusions. Fundamental Empirical Features of Spatiotemporal Congested Freeway Traffic Patterns" we make a summary of fundamental empirical spatiotemporal features of traffic congested patterns.

Congested Patterns on Homogeneous Road

As follows from empirical results and explained in three-phase traffic theory ([30,31,38], ► Traffic Congestion, Modeling Approaches to), a bottleneck increases the probability of traffic breakdown (F → S transition) in a neighborhood of this bottleneck considerably in comparison to other road locations. This explains why in real free flow the breakdown (F → S transition) occurs mostly at the bottleneck, whereas outside highway bottlenecks traffic breakdown is observed extremely seldom.

Although in real free flow the breakdown occurs mostly at a highway bottleneck, in this section we discuss spatiotemporal features of congested patterns occurring on homogeneous roads (outside or else without highway bottlenecks). This is because empirical, experimental, and theoretical studies of congested patterns occurring on homogeneous roads have a great *methodological* importance: these studies show that phase transitions and congested patterns in vehicular traffic result from complex driver interactions in traffic, rather than just as a result of highway bottlenecks [32,38].

Moving Synchronized Flow Patterns

An empirical example of the spontaneous emergence of synchronized flow in an initially free flow is shown in Fig. 2 [32]. The observation of this F → S transition outside bottlenecks was possible, because the distance between two adjacent bottlenecks (downstream off-ramp bottleneck at location $x = 23.4$ km, detectors D23 and upstream on-ramp bottleneck at $x = 17.1$ km, detectors D16) is great enough. We see that during the $F \to S$ transition at $x = 19$ km (detectors D18; up-arrow F → S at 06:36 in Fig. 2b) free flow was both at the off-ramp- and on-ramp bottlenecks (D23 and D16, D15).

After the F → S transition has occurred (labeled F → S in Figs. 2 and 3), the region of the synchronized flow propagates both downstream (up-arrow at location D20 in Fig. 3) and upstream (up-arrow at D17). A synchronized flow pattern, which propagates on a road as a whole localized structure is called a moving synchronized flow pattern (MSP).

A numerical simulation of an MSP on a homogeneous road without bottlenecks is shown in Fig. 4a. Due to a small speed disturbance, an F → S transition occurs spontaneously in an initially homogeneous free flow. This phase transition is a first-order one as this can be seen from Fig. 4b: there should be a speed disturbance of *finite* amplitude for the occurrence of the F → S transition. Thus there is a nucleation effect that is the usual one for other various first-order phase transitions in complex systems of different nature: there is a critical speed (density) disturbance in free flow, i. e., a nucleus required for the nucleation of an F → S transition. If the nucleus appears in free flow, the disturbance grows leading to the F → S transition. If in contrast the disturbance is smaller than the critical one, the disturbance decays. We see that the greater the density, the smaller the nucleus required for the F → S transition (curve F_S in Fig. 4b) [41].

Numerical simulations of F → S transitions (labeled F → S) with spontaneous emergence of many MSPs (labeled S) on a homogeneous circle road without bottlenecks is shown in Fig. 5. In this case, a KKW cellular automata three-phase traffic flow model has been studied [45] discussed in ► Traffic Breakdown, Probabilistic Theory of. There are model fluctuations of great amplitude in this model. For this reason, there are many local disturbances, which are nuclei for F → S transitions. As a result, many MSPs occur almost simultaneously on the road.

Pinch Effect and Moving Jam Emergence

Pinch effect conditions can be realized in synchronized flow outside bottlenecks [32]. An empirical example is shown in Fig. 6. In the synchronized flow that has initially occurred outside bottlenecks discussed in Subsect. "Moving Synchronized Flow Patterns" (Figs. 2 and 3), the region of synchronized flow with a greater density ($x = 20.9$ km, D20) occurs. Within this synchronized flow a narrow moving jam appears, which grows in its amplitude propagating upstream (down-arrows in Fig. 6). Finally, the jam transforms into a wide moving one. Both the spontaneous emergence of the jam and the resulting S → J transition occur outside the bottlenecks. The fact that the resulting jam is a wide moving one is confirmed by the empirical evidence that the jam propagates upstream through the on-ramp bottleneck (down-arrows at D16 and D15) while maintaining the characteristic velocity of the downstream front of the jam.

A numerical simulation of a spontaneous emergence of a moving jam within synchronized flow of an MSP with

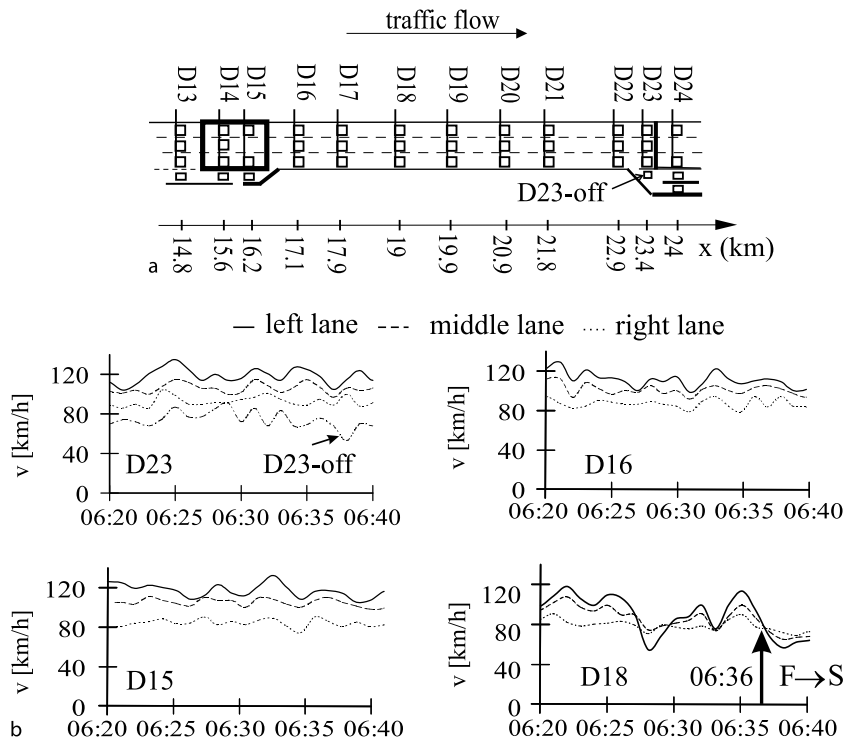

Traffic Congestion, Spatiotemporal Features of, Figure 2

Measured F → S transition outside bottlenecks: a Scheme of a part of freeway section of the freeway A5 South in Germany with road detectors labeled D13–D24 (see Fig. 2.1 of [38]). b Time-dependences of 1 min average speed in three road lanes at different locations. Taken from [32]

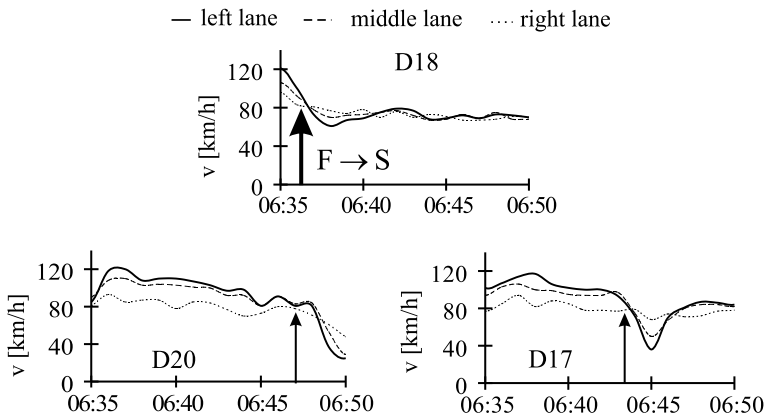

Traffic Congestion, Spatiotemporal Features of, Figure 3

Empirical propagation of synchronized flow outside bottlenecks from location of F → S transition (D18) downstream (D20) and upstream (D17) in data measured on the freeway section shown in Fig. 2a. Time-dependences of 1 min average speed in three road lanes at different locations. Taken from [32]

the subsequent S → J transition on a homogeneous road without bottlenecks is shown in Fig. 7 (labeled S → J) [41]. Another scenario of a sequence of F → S → J transitions, which occurs on a circular road is shown in Fig. 5: firstly, an F → S transition occurs and an MSP appears. Later and upstream of the location of the F → S transition a narrow moving jam emerges in synchronized flow of the MSP; the jam grows propagating upstream. Finally, a wide moving

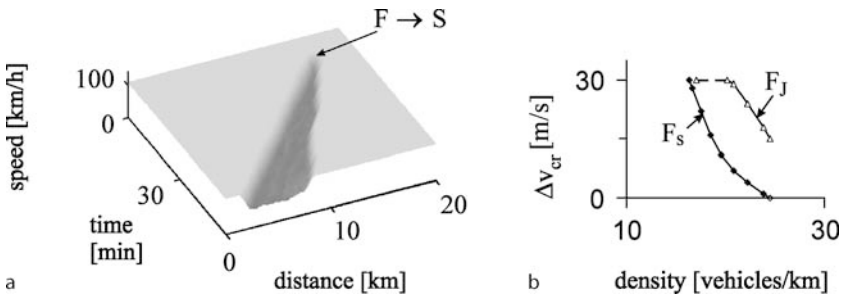

Traffic Congestion, Spatiotemporal Features of, Figure 4
Simulations of F → S transition on homogeneous road without bottlenecks: a MSP formation. b Critical amplitude (denoted by Δv_{cr}) of local speed disturbance, i. e., speed nucleus required for F → S transition (curve F_S) and F → J transition (curve F_J) as functions of vehicle density in initially homogeneous free flow. Taken from [41]

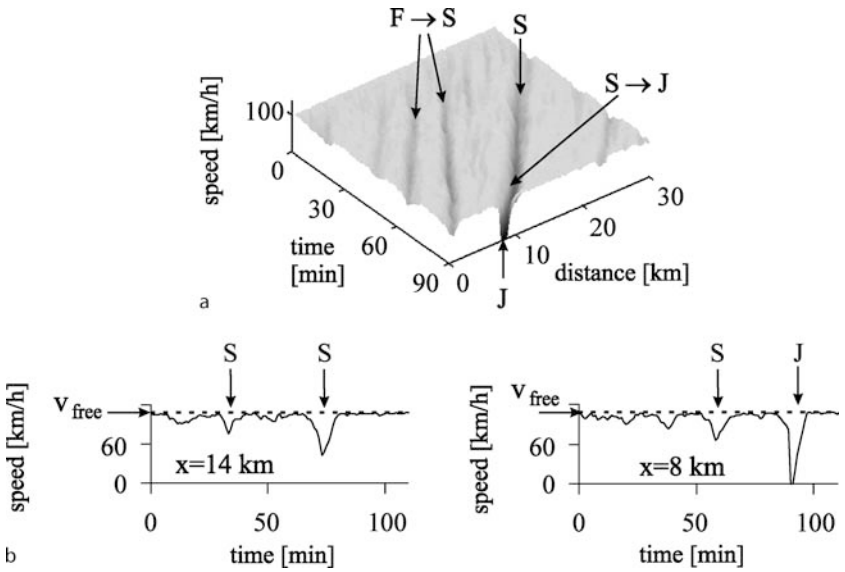

Traffic Congestion, Spatiotemporal Features of, Figure 5
Simulations of the spontaneous emergence of synchronized flow (F → S transition) in initially homogeneous free flow with the subsequent moving jam emergence in the emergent synchronized flow (S → J transition) on circle road without bottlenecks. Taken from [45]

jam emerges (labeled J). Results of three-phase traffic theory about F → S and S → J transitions on a homogeneous road [30,32,41,45] discussed above have also been confirmed in simulations of a cellular automata three-phase traffic flow model [24].

Pinch Effect Conditions in Homogeneous Synchronized Flow on Circle Road

In three-phase traffic theory has been found that an S → J transition is a first-order phase transition occurring in metastable states of synchronized flow. A detailed consideration of the multitude of metastable synchronized flow appears in Sect. "Moving Jam Emergence in Synchronized Flow (S → J Transition)" of ► Traffic Congestion, Modeling Approaches to in this Encyclopedia. In an initial homogeneous metastable synchronized flow, the S → J transition occurs, if a critical local speed (density) disturbance appears in this flow. This critical disturbance can be considered a nucleus required for the nucleation of an S → J transition. If the nucleus appears in the synchronized flow, the disturbance grows leading to the S → J transition. If in contrast the disturbance is smaller than the critical one, the disturbance decays. At a given speed, the greater the density in synchronized flow, the smaller the critical disturbance, i. e., the smaller the nucleus required for the

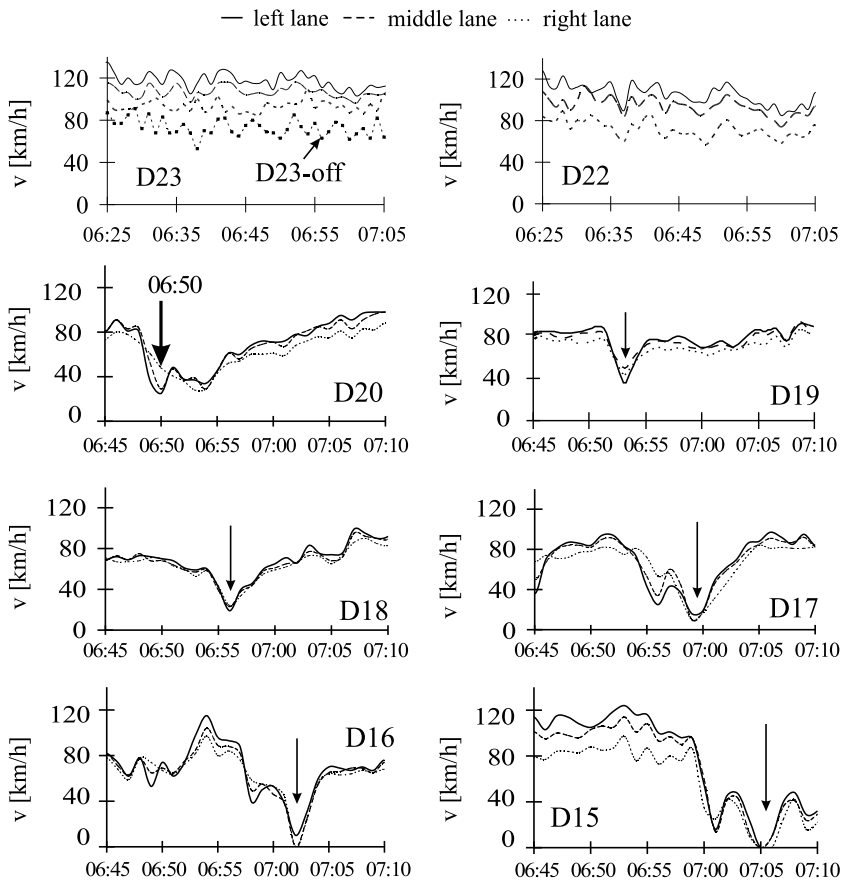

Traffic Congestion, Spatiotemporal Features of, Figure 6
Empirical spontaneous emergence of wide moving jam in synchronized flow (labeled by *down-arrows*) outside bottlenecks in data measured on the freeway section shown in Fig. 2a. Time-dependences of 1 min average speed in three road lanes at different locations. Taken from [32]

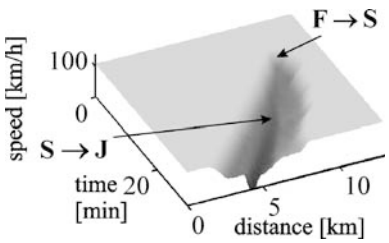

Traffic Congestion, Spatiotemporal Features of, Figure 7
Simulations of sequence of F → S → J transitions on homogeneous road without bottlenecks. Taken from [41]

S → J transition. Thus pinch effect conditions, i. e., synchronized flow in which speed disturbances grow spontaneously leading to the S → J transition can be realized on homogeneous road without bottlenecks ([30,31,32,38], ► Traffic Congestion, Modeling Approaches to).

Results of simulations of an S → J transition in synchronized flow on a circle road without bottlenecks under pinch effect conditions are presented in Fig. 8 [41]. In contrast with simulations of a sequence of F → S → J transitions shown in Fig. 5, in Fig. 8 rather than free flow, the initial traffic flow is associated with homogeneous synchronized flow that is in a metastable state with respect to an S → J transition. The initial synchronized flow speeds (72 km/h for the curve $S_J^{(1)}$ and 51 km/h for $S_J^{(2)}$) have been considerably lower than free flow speed ($v_{free} = 108$ km/h in Fig. 5). During a short time interval a local speed disturbance of an initial amplitude Δv has been applied in the synchronized flow: within the disturbance, the decrease in speed is equal to Δv. Then, the subsequent dynamic evolution of the initial speed disturbance in the initially homogeneous synchronized flow on the circle road is studied. It is found that there is a critical amplitude of the speed disturbance denoted by $\Delta v_{cr}^{(SJ)}$, i. e., there is a speed

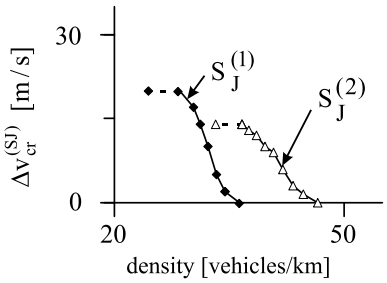

Traffic Congestion, Spatiotemporal Features of, Figure 8
Simulations of the spontaneous emergence of wide moving jams (S → J transition) in homogeneous synchronized flow on circle road without bottlenecks: Critical amplitude of local speed disturbance denoted by $\Delta v_{cr}^{(SJ)}$, i. e., speed nucleus required for S → J transition (curves $S_J^{(1)}$ and $S_J^{(2)}$ are related to the synchronized flow speeds 72 and 51 km/h, respectively). Taken from [41]

nucleus of a "size" $\Delta v_{cr}^{(SJ)}$ required for an S → J transition: if $\Delta v \geq \Delta v_{cr}^{(SJ)}$, then the disturbance grows leading to the spontaneous emergence of a wide moving jam in synchronized flow. If in contrast, $\Delta v < \Delta v_{cr}^{(SJ)}$, then the disturbance decays over time resulting in homogeneous synchronized flow. At a given synchronized flow speed, the study has been made for various initial densities of synchronized flow. We found that at each of the given speeds, the greater the density in synchronized flow, the smaller the nucleus required for the S → J transition (curves $S_J^{(i)}$, $i = 1, 2$ in Fig. 8). If the density of synchronized flow is great enough, then $\Delta v_{cr}^{(SJ)} \to 0$ (Fig. 8), i. e., a very small nucleus is required for the spontaneous emergence of a wide moving jam in the dense synchronized flow.

This means that in such an initially homogeneous dense synchronized flow on a homogeneous circle road without bottlenecks pinch effect conditions are fully satisfied: already very small speed (density) disturbances in this synchronized flow grow propagating upstream with the subsequent spontaneous moving jam emergence [32,41].

Pinch Effect Conditions in Driver Experiment: Spontaneous Emergence of Moving Jams in Synchronized Flow of Great Density on Circle Road

The result of three-phase traffic theory discussed in Subsect. "Pinch Effect Conditions in Homogeneous Synchronized Flow on Circle Road" [32,38,41,45] has recently been confirmed in a driver experiment [66]. In this experiment, wide moving jams emerge spontaneously within an initially homogeneous synchronized flow of great density. In the study, $N_{veh} = 22$ vehicles have initially homogeneously moved along a circle road of a length $L_{circle} = 230$ m with a speed of about 30 km/h. Over time, growing local speed (density) disturbances have occurred in this vehicle motion. The growth of these disturbances has led to wide moving jam formation, which propagate upstream with the velocity about 20 km/h that is associated with empirical observations of moving jams that emerge in the pinch region of synchronized flow [38].

The experiment [66] is a very interesting and important one for a deeper understanding of moving jam emergence in synchronized flow. However, we should note that both the statement of the authors that the spontaneous moving jam emergence occurs in *free flow* as well as the associated physical explanation of the jam emergence in traffic are *invalid*. Indeed, in the experiment [66] vehicles could not overtake each other and the vehicle speed (30 km/h) has been considerably lower than the speed in free flow (usually higher than 80 km/h for passenger cars). Moreover, the initial density of the initially homogeneous traffic flow of [66] has been $\rho_{hom} = N_{veh}/L_{circle} \approx 95$ vehicles/km that is considerably greater than the maximal density observed in free flow, which is usually appreciably smaller than 40 vehicles/km. The vehicle density 95 vehicles/km in the initially homogeneous flow in the experiment [66] is even considerably greater than the usual density in the pinch region of synchronized flow occurring upstream of on- and off-ramp bottlenecks; this density is usually within the range 35–60 vehicles/km [38].

Thus rather than free flow, the initially homogeneous flow in the experiment of [66] is associated with synchronized flow of a very great density. Rather than the moving jam emergence in free flow, in the experiment [66] the spontaneous moving jam emergence in an initially homogeneous synchronized flow of a very great density has been shown discussed in Subsect. "Pinch Effect Conditions in Homogeneous Synchronized Flow on Circle Road". This is in accordance with all other observations of wide moving jam emergence in synchronized flow occurring on homogeneous roads (Subsect. "Pinch Effect and Moving Jam Emergence") as well as upstream of bottlenecks (see for example Fig. 16 discussed in Subsect. "General Congested Patterns") [32,38].

In contrast with the invalid physical explanation for the spontaneous moving jam emergence in traffic made in [66], as follows from all known empirical observations of real traffic as well as from the characteristics of an initial traffic flow in the driver experiment [66] discussed above, moving jams do not emerge spontaneously in free flow ([32,38], ► Traffic Congestion, Modeling Approaches to). Three-phase traffic theory explains this result of empirical observations as follows. At the same density of free flow, the nucleus required for the synchronized flow emer-

gence in this free flow (curve F_S in Fig. 4) is considerably smaller than the nucleus required for the emergence of a moving jam in this free flow (curve F_J in Fig. 4). This can explain why rather than the moving jam emergence, the synchronized flow emergence governs the onset of congestion in real free flow. For this reason, all empirical observations show that moving jams spontaneously emerge in *free flow* due to a sequence of F → S → J transitions *only*. This is regardless of whether there are bottlenecks on a road or not. A more detailed explanation of this important empirical result can be found in [38] as well as in ► Traffic Congestion, Modeling Approaches to.

Congested Patterns at Isolated Bottlenecks

As mentioned above, the onset of congestion (F → S transition) in an initial free flow occurs mostly at a freeway bottleneck. A bottleneck at which traffic breakdown occurs is called the *effectual bottleneck*. After traffic breakdown has occurred, synchronized flow is formed at the effectual bottleneck. In many cases, the downstream front of the synchronized flow is fixed at the bottleneck (the exclusion is moving synchronized flow patterns). The location in a neighborhood of the bottleneck at which this downstream front of synchronized flow is fixed is called an *effective location of the effectual bottleneck* (the effective location of the bottleneck for short). It must be noted that the effective location of the bottleneck can be different from the location at which traffic breakdown (F → S transition) has occurred leading to congested pattern emergence. Moreover, both the location of traffic breakdown and the effective location of the bottleneck are probabilistic values in real traffic. Even for the same type of congested pattern the effective location of the bottleneck can randomly change over time [38].

On real freeways there are many effectual bottlenecks. When a congested pattern has occurred at a downstream effectual bottleneck, the pattern can propagate further upstream reaching an upstream effectual bottleneck. In this case, if free flow at the upstream bottleneck is in a metastable state with respect to an F → S transition (i. e., with respect to synchronized flow emergence), then congestion pattern propagation from the downstream bottleneck causes an induced F → S transition at the upstream bottleneck. This can lead to complex congested pattern transformation effects considered below.

Before we discuss these complex effects of congested pattern transformation, we would like to discuss congested patterns that emerge at a hypothetical isolated effectual bottleneck. As we have already mentioned, on real freeways there are many effectual bottlenecks, therefore, a consideration of the isolated bottleneck is an idealization of real traffic. Obviously, this idealization is easy to perform with a traffic flow model rather than with real traffic. However, on real freeways there can be freeway sections with adjacent effectual bottlenecks that are located far enough from each other. On such freeway sections during long enough time intervals before a congested pattern, which has emerged at a downstream bottleneck, reaches an upstream bottleneck, the congested pattern can be considered the pattern at an isolated bottleneck. This bottleneck is the downstream bottleneck at which the pattern has initially occurred. Such a simplification of the reality allows us to understand some important common spatiotemporal features of congested patterns. Nevertheless, although in this section only spatiotemporal features of congested patterns at isolated bottlenecks are discussed, all illustrations of *empirical* examples of congested patterns represent the real development of measured congested patterns, which sometimes cause induced traffic breakdowns and complex pattern transformations at upstream bottlenecks when the patterns reach them. To prevent confusions, these real induced pattern formation and complex congested pattern transformation will be further explained in Sect. "Complex Congested Patterns and Pattern Interaction" of this article. There are two main types of congested patterns at an isolated bottleneck: 1. Synchronized flow patterns (SP for short). 2. General patterns (GP for short).

Synchronized Flow Patterns

As a result of an F → S transition at an isolated bottleneck, various SPs can occur at the bottleneck. There are three main types of SPs [38]:

(1) Localized SP (LSP for short): The downstream front of an LSP is fixed at the bottleneck. The upstream front of the LSP does not continuously propagate upstream over time: this front is localized at some distance upstream of the bottleneck. In other words, the region of synchronized flow in the LSP is localized in space. However, the location of the upstream front of synchronized flow in the LSP and therefore the LSP width (in the longitudinal direction) can exhibit complex oscillations over time. Note that the upstream front of synchronized flow separates free flow upstream from synchronized flow downstream of this front. Vehicles must slow down within the upstream front of synchronized flow.
(2) Widening SP (WSP for short): As in the LSP, the downstream front of an WSP is fixed at the bottleneck. In contrast to the LSP, the upstream front of the WSP

propagates upstream continuously over time. In other words, the width of synchronized flow (in the longitudinal direction) in the WSP widens in the upstream direction. It must be noted that due to this widening, the upstream front of synchronized flow must eventually reach an upstream adjacent effectual bottleneck. At this upstream bottleneck another congested pattern either already exists or can be induced. In both cases, the WSP can be caught at this upstream bottleneck (catch effect; see below). Thus, in real traffic the WSP can usually exist only for a finite time, before the WSP reaches the upstream effectual bottleneck.

(3) Moving SP (MSP for short): In contrast to the LSP and WSP, an MSP is a localized pattern on a freeway that propagates within free flow over time between freeway bottlenecks; both downstream and upstream MSP fronts, which separate synchronized flow within the MSP and free flows away from the MSP, propagate on the freeway. Depending on flow rate distribution within the MSP and in free flows away from the MSP, each of the MSP fronts can propagate either upstream or downstream. In particular, downstream MSP propagation (i. e., when both MSP fronts propagate downstream) is possible; this occurs if the flow rate within the MSP is greater than the flow rates in free flows away from the MSP. Otherwise, upstream MSP propagation is realized. Due to MSP propagation, the MSP (or one of the MSP fronts) can reach an adjacent effectual bottleneck at which synchronized flow can be excited; then the MSP (or the associated MSP front) is caught at the bottleneck (catch effect). Thus, the MSP that propagates between bottlenecks can exist only for a finite time, before the MSP (or one the MSP fronts) reaches a nearest adjacent effectual bottleneck at which synchronized flow can be excited.

An empirical example of an LSP is shown in Fig. 9. It can be seen that there are no wide moving jams in the LSP. Moreover, we can see that whereas the speed is considerably lower within the LSP (Fig. 9a), the flow rate is close to the flow rate in free flow (see Fig. 9b, where the increase in flow rate downstream of the bottleneck is related to the on-ramp inflow).

The upstream front of synchronized flow in the LSP (Fig. 9a) does not continuously propagate upstream of the bottleneck over time: this upstream front is localized at some finite distance upstream of the bottleneck. This distance is a function of time. In other words, this SP is indeed an empirical example of an LSP.

In numerical simulations made in the framework of three-phase traffic theory, LSPs exhibit qualitatively the same spatiotemporal features as those found in empirical data (Fig. 10) [38]:

(i) The downstream front of the LSP is fixed at the on-ramp bottleneck. However, the upstream front of the LSP is localized on the main road at some distance L_{LSP} upstream of the on-ramp bottleneck (Fig. 10a,f).
(ii) Whereas the speed within the LSP is considerably lower than in free flow, the flow rate is approximately the same as that in free flow (Fig. 10b–e).
(iii) The LSP width L_{LSP} can be a complicated time function (Fig. 10f). The mean width of the LSP $L_{\mathrm{LSP}}^{(\mathrm{mean})}$ can depend strongly on q_{in} and q_{on}. In numerical simulations, $L_{\mathrm{LSP}}^{(\mathrm{mean})}$ varies between 0.5 km and 10 km.

An empirical example of an MSP is shown in Fig. 11. We consider two adjacent bottlenecks, a downstream bottleneck B_1 and an upstream bottleneck B_2. Synchronized flow first occurs at the downstream bottleneck B_1. The synchronized flow subsequently propagates upstream as a distinct localized structure. This is an example of an MSP (MSP is labeled "MSP" in Fig. 11a,c). When the MSP reaches the upstream bottleneck B_2, it does not propagate through the bottleneck as a localized structure: the MSP is caught at the bottleneck (catch effect). This is in contrast to wide moving jam propagation through an upstream bottleneck, as shown in Fig. 1 and for a wide moving jam labeled "wide moving jam" in Fig. 11b,c. It should be also noted that whereas the flow rate is very small within the latter wide moving jam, we cannot almost distinguish the MSP in the flow rate distribution in space and time shown in Fig. 11b. As already mentioned above, this is a very important features of many empirical (measured) SPs: The flow rate within the SP can be as great as in an initial free flow.

Numerical simulations of MSPs made in the framework of three-phase traffic theory show qualitatively the same spatiotemporal features as those found in empirical data [38]:

(i) The MSP occurs spontaneously on the main road at the off-ramp (Fig. 12a) or on-ramp (Fig. 12b) bottlenecks because the downstream front of synchronized flow can depart from the bottleneck. As a result, the MSP begins to propagate on the main road as an independent localized structure on a homogeneous road. As in empirical observations (Fig. 11b), whereas the speed within the MSP is considerably lower than in free flow, the flow rate is approximately the same as that in free flow (Fig. 12c,d).
(ii) When two adjacent bottlenecks exist on a freeway and an MSP appears at the downstream one, then the MSP

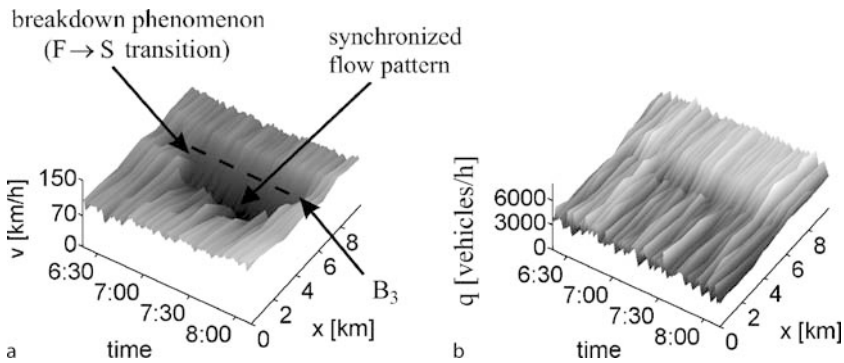

Traffic Congestion, Spatiotemporal Features of, Figure 9

Empirical example of a localized synchronized flow pattern (LSP) at on-ramp bottleneck B_3. a Averaged vehicle speed in the LSP in space and time. b Total flow rate across the freeway in space and time. The *dashed line* in a shows the location of the bottleneck. Taken from [38]

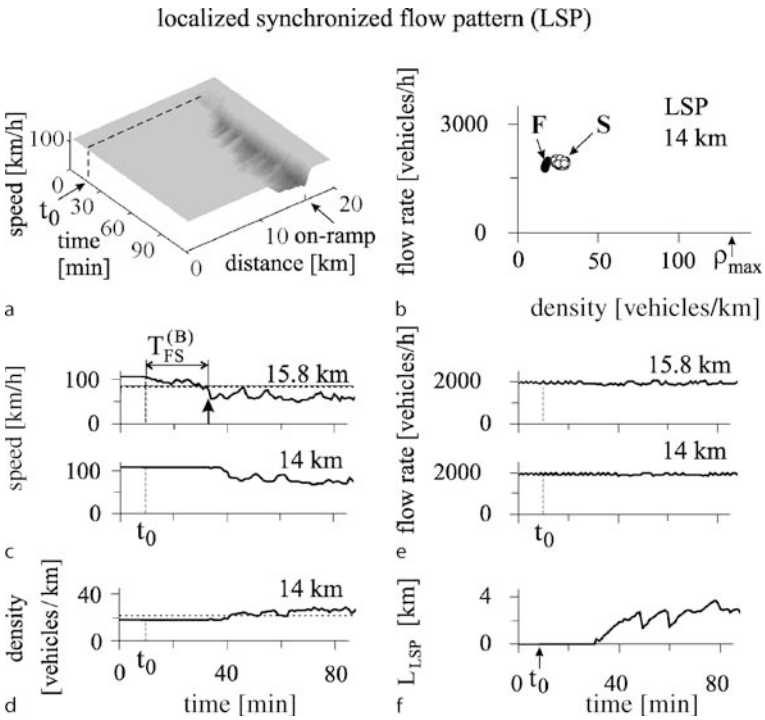

Traffic Congestion, Spatiotemporal Features of, Figure 10

Localized synchronized flow pattern (LSP) at on-ramp bottleneck in the framework of three-phase traffic theory (numerical simulations of the Kerner–Klenov stochastic three-phase traffic flow model considered in ► Traffic Congestion, Modeling Approaches to): a Averaged vehicle speed in the LSP in space and time. b Free flow (F) and synchronized flow within the LSP (S) in the flow–density plane. c–e Speed (c), density (d) and flow rate (e) as time-functions at fixed freeway locations. f The LSP width L_{LSP} as a time-function. Taken from [38]

can be caught at the downstream bottleneck with the subsequent SP formation.

An empirical example of an WSP is shown in Fig. 13. In this case, synchronized flow due to a spontaneous traffic breakdown is formed at an off-ramp bottleneck B_{North1}. The downstream front of this synchronized flow is fixed at this bottleneck. The upstream front of synchronized flow propagates continuously upstream over time. When this front reaches an upstream on-ramp bottleneck B_{North2}, the WSP is caught at the bottleneck (catch effect) causing an induced traffic breakdown at the bottleneck with the sub-

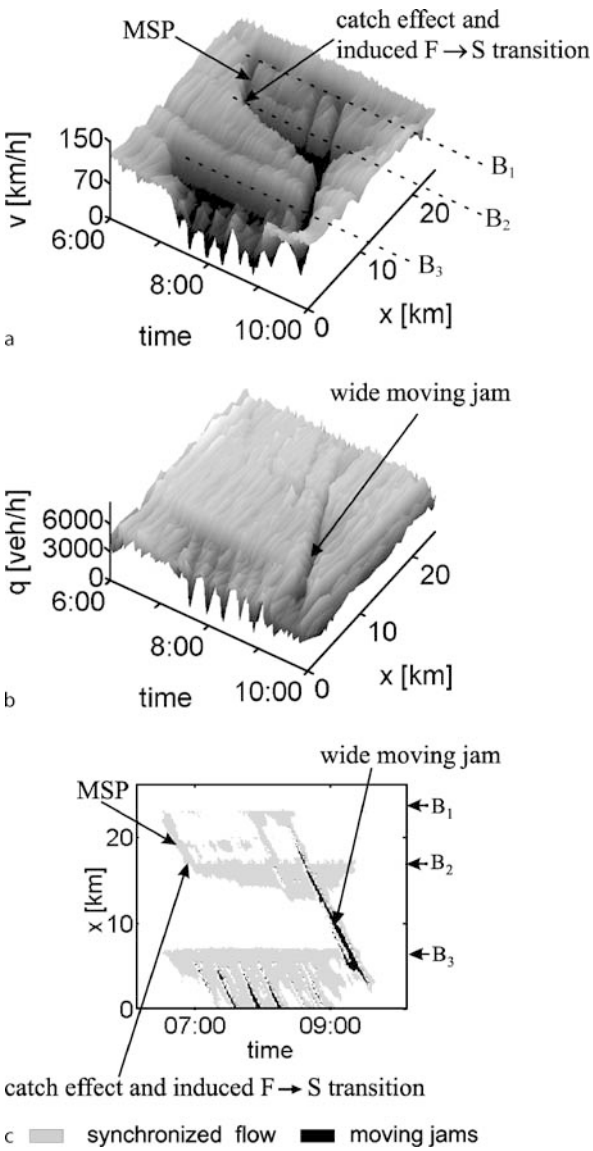

Traffic Congestion, Spatiotemporal Features of, Figure 11
Empirical example of MSP with subsequent catch effect and induced F → S transition at bottleneck B_2: a, b Average speed (a) and flow rate (b) in space and time. c Graph of a with overview of free flow (*white*), synchronized flow (*gray*), and moving jams (*black*). Bottlenecks B_2 and B_3 are the same as those in Fig. 1. Taken from [38]

sequent formation of another congested pattern upstream of the bottleneck.

Numerical simulations of WSPs made in the framework of three-phase traffic theory show qualitatively the same spatiotemporal features as those found in empirical data (Fig. 14) [38]. The WSP occurs spontaneously on the main road at the off-ramp (Fig. 14a) or on-ramp (Fig. 14b) bottlenecks. The downstream front of an WSP is fixed at the on-ramp bottleneck. The upstream front of the WSP is continuously widening upstream. As within other SPs, the speed within the WSP is considerably lower than in free flow and the flow rate is approximately the same as that in free flow (Figs. 13b and 14c,d).

General Congested Patterns

It must be noted that over time the average speed in synchronized flow of an WSP can decrease, and wide moving jams can begin to emerge spontaneously. In other words, the WSP can transform into a general pattern (GP). A GP is a congested pattern at an isolated bottleneck, which consists of synchronized flow whose downstream front is fixed at the bottleneck and wide moving jams that emerge spontaneously in this synchronized flow upstream of the bottleneck (Fig. 15a,b). Thus the GP consists of the both traffic phases of congested traffic, synchronized flow and wide moving jams.

In the empirical example shown in Fig. 15, WSP transformation into a GP occurs downstream of an upstream on-ramp bottleneck B_{North2}. For this reason, we can qualitatively consider the congested pattern that is upstream of the downstream off-ramp bottleneck B_{North1} and downstream of the upstream on-ramp bottleneck B_{North2} as an empirical example of a GP at the off-ramp bottleneck. Because the jams labeled 1, 2, and 3 propagate through the upstream bottleneck while maintaining their mean velocities of the jam downstream front (Fig. 15b), in accordance with the phase definition [J], these jams are indeed wide moving jams.

However, we must note that in many cases, upstream congestion (i. e., synchronized flow and wide moving jams in Fig. 15a,b upstream of the on-ramp bottleneck B_{North2}) can influence downstream congested pattern parameters considerably (Sect. "Complex Congested Patterns and Pattern Interaction"). For this reason, the above consideration of the empirical congested pattern between the adjacent bottlenecks B_{North1} and B_{North2} as a GP can only show some qualitative features of GPs at an isolated bottleneck.

Another empirical example of a GP, which is formed at an on-ramp bottleneck before wide moving jams or/and synchronized flow of this GP reach an eventual upstream bottleneck on the freeway, is shown in Fig. 16. We see that firstly, synchronized flow is formed upstream of the bottleneck (labeled by arrows F → S and S). Later, and at other freeway locations wide moving jams emerge in that synchronized flow. These wide moving jams propa-

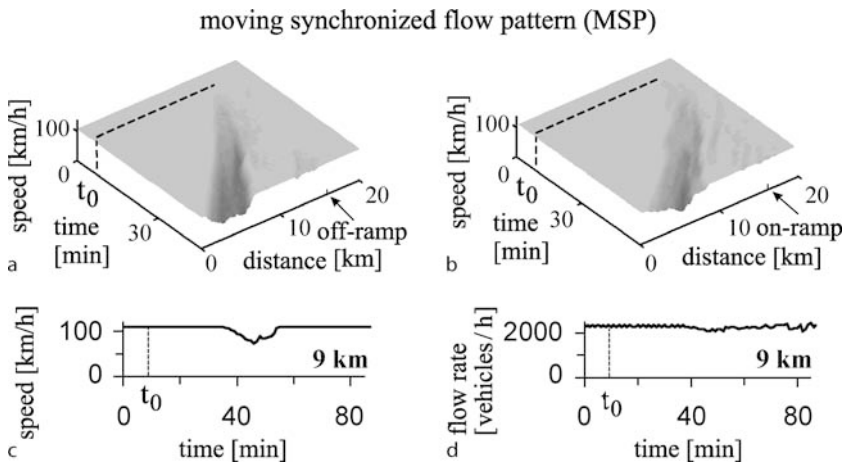

Traffic Congestion, Spatiotemporal Features of, Figure 12

Moving synchronized flow patterns (MSP) at off-ramp (a) and on-ramp (b–d) bottlenecks in the framework of three-phase traffic theory [38] (numerical simulations of the Kerner–Klenov stochastic three-phase traffic flow model considered in ► Traffic Congestion, Modeling Approaches to): a, b Averaged vehicle speed in the MSP in space and time. c, d Speed (c) and flow rate (d) as time-functions at a fixed freeway location. Taken from [38]

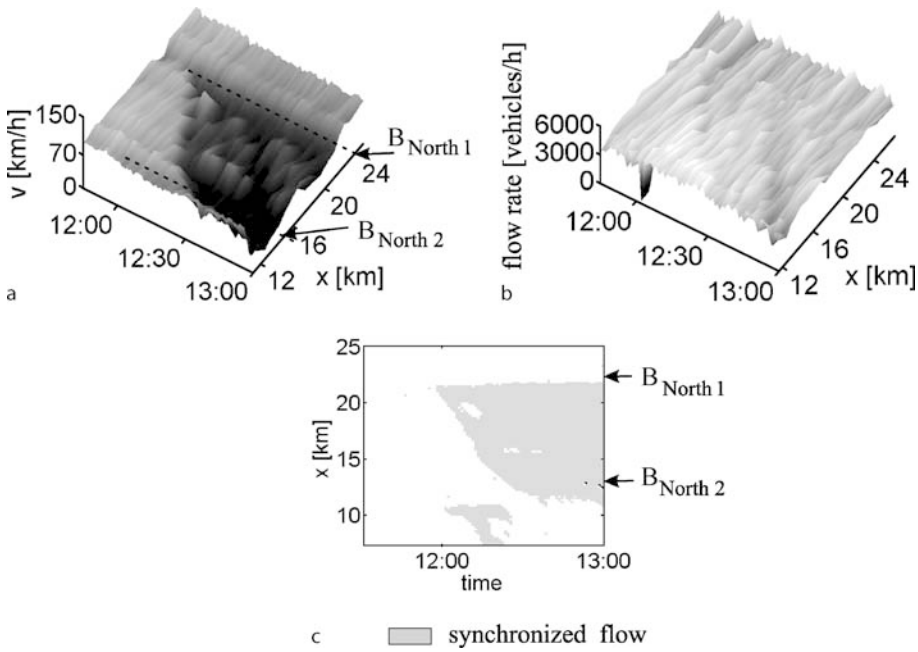

Traffic Congestion, Spatiotemporal Features of, Figure 13

Empirical example of an WSP and the catch effect: a, b Average speed (a) and flow rate (b) in space and time. c Graph of a with overview of free flow (*white*) and synchronized flow (*gray*). Taken from [38]

gate further upstream while maintaining the velocity of their downstream fronts. Free flow remains downstream of the bottleneck (location 7.8 km).

We can see that qualitative features of this GP are the same as those in Fig. 15. In both cases, firstly an F → S transition occurs at the bottleneck (arrow S in Figs. 15c and 16b). Synchronized flow propagates upstream. Later, the pinch effect is realized: moving jams emerge spontaneously in the synchronized flow upstream of the location of the initial traffic breakdown. These moving jams propagate upstream growing in amplitude. Some of the jams transform into wide moving jams, i. e., S → J transitions

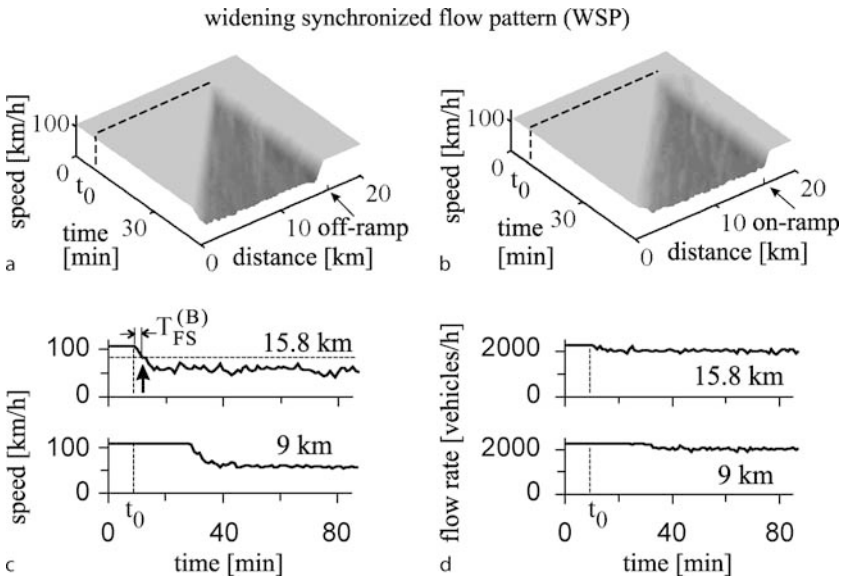

Traffic Congestion, Spatiotemporal Features of, Figure 14
Widening synchronized flow patterns (WSP) at off-ramp (a) and on-ramp (b–d) bottlenecks in the framework of three-phase traffic theory [38] (numerical simulations of the Kerner–Klenov stochastic three-phase traffic flow model considered in ► Traffic Congestion, Modeling Approaches to): a, b Averaged vehicle speed in the WSP in space and time. c, d Speed (c) and flow rate (d) as time-functions at a fixed freeway location. Taken from [38]

occur. In the example, the mean time interval between the emergence of a narrow moving jam and the jam transformation into a wide moving one is about 10 min, i. e., about ten times longer than the mean duration of F → S transitions, which are about 1 min in duration.

However, there are also some differences in GP formation in Figs. 15 and 16. In Fig. 15, WSP exists during a long enough time interval before the pinch effect occurs within the WSP, i. e., the WSP transforms into the GP. In contrast, in Fig. 16 the pinch effect occurs in synchronized flow after a short time interval. Secondly, in Fig. 15 the frequency of wide moving jam emergence in synchronized flow is considerably smaller than the one for the GP shown in Fig. 16.

The differences in GP formation that have been mentioned above result from different pinch effect features in these two cases (Figs. 15 and 16). It has been found that the stronger the self-compression of synchronized flow in the pinch region is (i. e., the more the increase in synchronized flow density), the more likely moving jams emerge and the greater the frequency of this moving jam emergence is. This is in accordance with pinch effect features discussed in Subsect. "Pinch Effect Conditions in Homogeneous Synchronized Flow on Circle Road". In the first case of the off-ramp bottleneck shown in Fig. 15, the pinch effect is weak (so-called "weak congestion" case), i. e., a relatively small increase in density of the synchronized flow within the initial WSP occurs. As a result, the frequency of wide moving emergence is small; the frequency depends on the percentage of vehicle leaving the main road to the off-ramp.

In contrast, in Fig. 16 the pinch effect is relatively strong, i. e., density increases considerably in the pinch region of synchronized flow; the frequency of wide moving jam emergence is greater than in the former case. It has been found that if the flow rates upstream of the on-ramp bottleneck increase, then a case of so-called "strong congestion" within the pinch region can be reached. In this case, regardless of further on-ramp flow rate increase, the flow rate within the pinch region on the main road reaches a limit (minimum) flow rate and wide moving jam frequency reaches the maximum value. This is valid, if bottleneck characteristics and traffic parameters like weather, percentage of long vehicles, and other road conditions do not change considerably.

These and other empirical spatiotemporal features of GP formation have been explained in three-phase traffic theory (Fig. 17). In particular, the following features of the pinch effect have been found:

(i) Firstly, an F → S transition occurs at the bottleneck. The emergent synchronized flow propagates

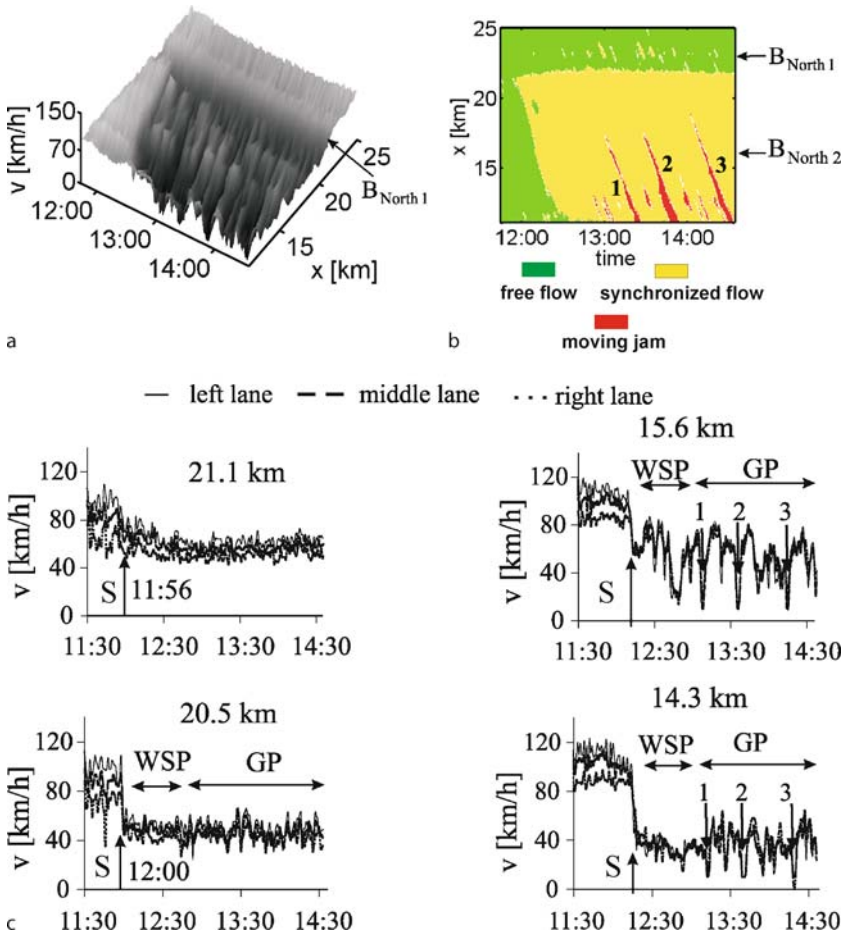

Traffic Congestion, Spatiotemporal Features of, Figure 15

Empirical example of WSP transformation into a GP over time: a Average speed in space and time. b Graph of a with overview of free flow (*green*), synchronized flow (*yellow*) and wide moving jams labeled 1, 2, and 3 (*red*). c Time dependence of speed at some fixed freeway locations. Taken from [38]

upstream. In the examples of GPs shown in Fig. 17, the upstream front of the congested patterns is the upstream front of the synchronized flow that has initially emerged at the bottleneck. There are also other GP examples, when the upstream front of the GP is associated with the upstream front of the farthest upstream wide moving jam of a GP (such GPs are not shown here; see the book [38]).

(ii) Later, the pinch effect occurs in synchronized flow upstream of the bottleneck. However, the flow rate does not necessarily decrease within the synchronized flow. Even if the flow rate decreases due to the pinch effect, the relative value of this decrease is considerably smaller than the decrease in average speed; this is explained by considerable increase in density, i. e., by self-compression of the synchronized flow. In the pinch region of a GP, narrow moving jams emerge spontaneously and grow propagating upstream (Fig. 17c), $x = 15.8$ km and $x = 14.5$ km.

(iii) Points related to the pinch region lie above the line J in the flow–density plane, i. e., these points are associated with metastable synchronized flow states with respect to wide moving jam emergence. This explains narrow moving jams emergence and growth in the pinch region; see explanations of metastable states of synchronized flow and of growing narrow moving jams, which are associated with the synchronized flow phase, in ([38], ► Traffic Congestion, Modeling Approaches to).

(iv) The location of an S → J transition, at which a narrow moving jam transforms into a wide moving jam, is related to the upstream boundary of the pinch re-

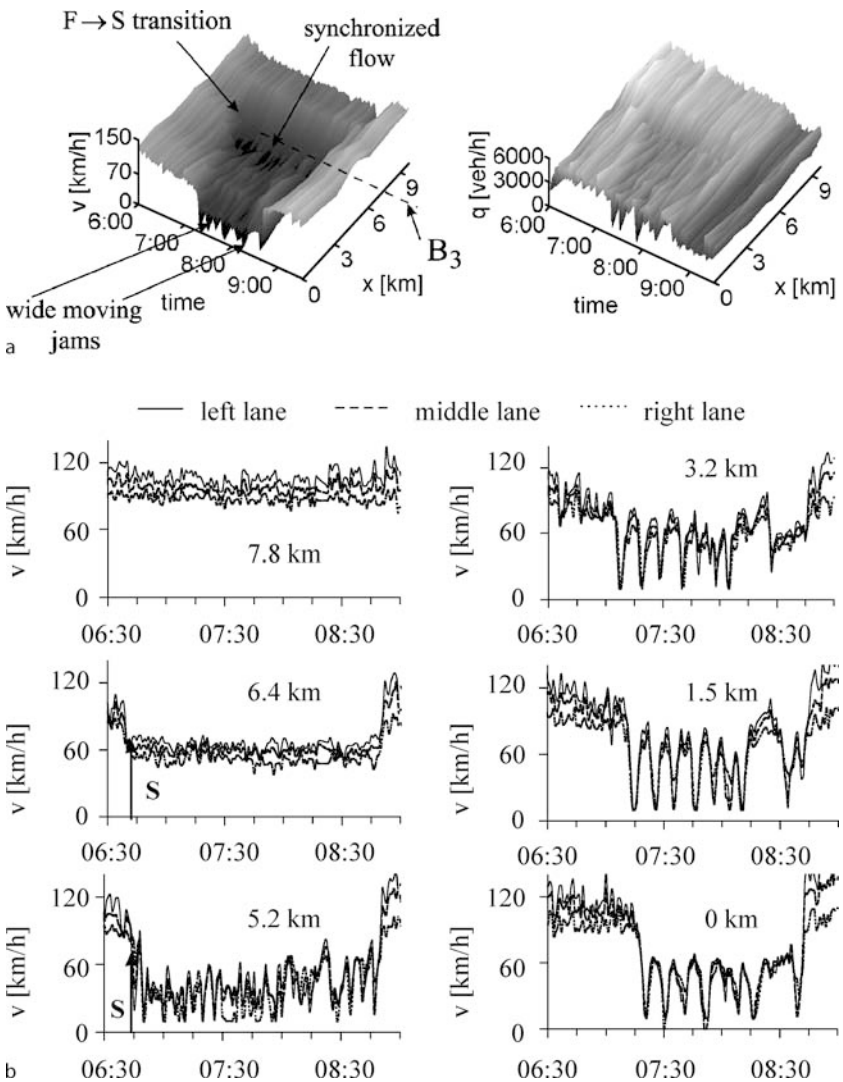

Traffic Congestion, Spatiotemporal Features of, Figure 16

Empirical example of GP at on-ramp bottleneck B_3. a Average vehicle speed (*left*) and total flow rate (*right*) across the freeway in space and time. b Speed in the three freeway lanes at different locations (1 min data). Taken from [38]

gion of the GP. Different narrow moving jams can transform into wide moving jams at different locations. For this reason, the upstream boundary of the pinch region can exhibit complicated oscillations over time.

(v) After a narrow moving jam has transformed into a wide moving jam, this wide moving jam can suppress the growth of the downstream narrow moving jam (Fig. 17c, $x = 13$ km). This jam suppression effect occurs only if the narrow moving jam is close to the downstream wide moving jam front.

(vi) Distances between different narrow moving jams in the pinch region can be very different to one another.

It must be noted that the well-known and very old term "stop-and-go" traffic, which is related to a sequence of moving traffic jams, will not be used in this article. For a traffic observer, both a sequence of narrow moving jams and a sequence of wide moving jams constitutes "stop-and-go" traffic. However, as already mentioned, narrow moving jams belong to the synchronized

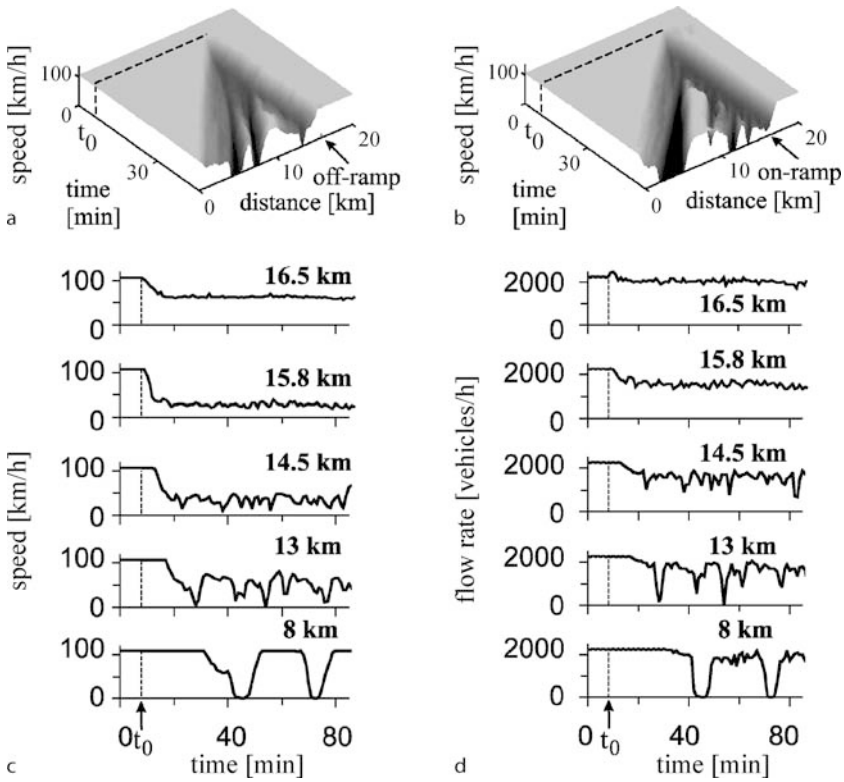

Traffic Congestion, Spatiotemporal Features of, Figure 17

Simulations of GP at off-ramp (a) and on-ramp bottlenecks (b–d). a, b Average vehicle speed in space and time. c, d Speed (c) and flow rate (d) at different locations (1 min data averaging) for the GP in b. Taken from [38]

flow phase, whereas wide moving jams belong to the qualitative different wide moving jam phase.

Diagram of Congested Patterns at Isolated Bottlenecks

A diagram of different congested patterns at an on-ramp bottleneck gives regions of congested pattern occurrence in the flow–flow plane whose coordinates are the flow rate in free flow on the main road upstream of the bottleneck q_{in} and the flow rate to the on-ramp q_{on} (Fig. 18a) [41,42,45]. There are two main boundaries in this diagram, $F_{\text{S}}^{(\text{B})}$ and $S_{\text{J}}^{(\text{B})}$.

Below and left of the boundary $F_{\text{S}}^{(\text{B})}$ free flow occurs. At the boundary $F_{\text{S}}^{(\text{B})}$ within a given time interval for observing free flow at the bottleneck an F → S transition occurs that leads to synchronized flow emergence.

Between the boundaries $F_{\text{S}}^{(\text{B})}$ and $S_{\text{J}}^{(\text{B})}$ different SPs (LSP, MSP, and WSP) emerge upstream of the on-ramp (Figs. 10, 12, and 14). Right of the boundary $S_{\text{J}}^{(\text{B})}$ wide moving jams occur spontaneously in synchronized flow, i. e., different GPs appear (Figs. 17 and 19). At the boundary $S_{\text{J}}^{(\text{B})}$ within a given time interval for observing synchronized flow an S → J transition occurs that leads to wide moving jam emergence.

Considering regions of SP emergence in the diagram, the following conclusions can be made. The WSP occurs above the boundary W in the diagram in Fig. 18a. Below the boundary W an LSP occurs. At greater q_{in} and a very small q_{on} an MSP can occur rather than the WSP. The MSP occurs spontaneously on the main road at the on-ramp bottleneck because at very low q_{on} the downstream front of synchronized flow can depart from the on-ramp bottleneck. The MSP begins to propagate on the main road as an independent localized structure (Fig. 12) on a homogeneous road. A sequence of MSPs can also occur on the main road at the on-ramp bottleneck.

When right of the boundary $S_{\text{J}}^{(\text{B})}$ at a given q_{in} the flow rate q_{on} increases, the average density within the pinch region of a GP increases too. There is a boundary labeled C in the diagram, which separates weak and strong conges-

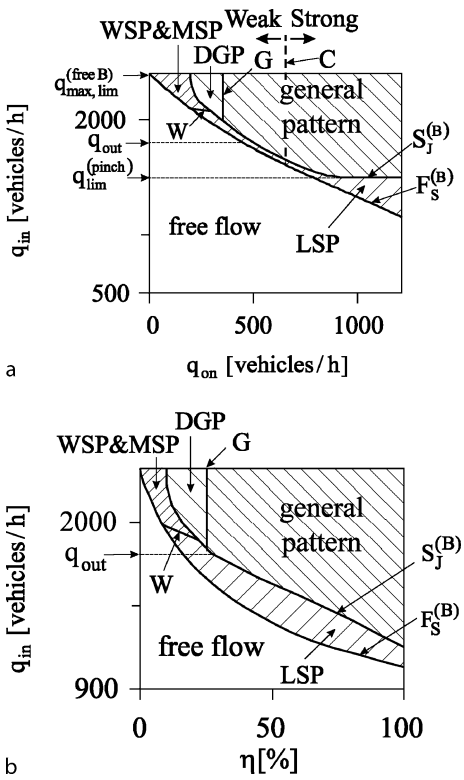

Traffic Congestion, Spatiotemporal Features of, Figure 18
Diagrams of congested patterns at on-ramp (a) and off-ramp bottlenecks (b). Results of numerical simulations. Taken from [38]

tion condition within the pinch region of GPs. Right of this boundary weak congestion at which the pinch region characteristic depends on q_{on} transforms into strong congestion at which the pinch region characteristics like the flow rate (this limit flow rate is denoted by $q_{\text{lim}}^{(\text{pinch})}$ in the diagram) and the frequency of wide moving jam emergence reach their limit values as observed in empirical data. As a result of strong congestion, there is saturation of the boundary $S_{\text{J}}^{(\text{B})}$ at greater q_{on} associated with the limit flow rate $q_{\text{lim}}^{(\text{pinch})}$ (Fig. 18a).

There is a region in the diagram that is between the boundary $S_{\text{J}}^{(\text{B})}$ and the line G within which a dissolving GP (DGP) occurs (Fig. 18a). In an DGP, after a wide moving jam has been formed this wide moving jam suppresses the pinch region in the initial GP. The suppression effect can occur when the initial flow rate q_{in} is high enough and the flow rate to the on-ramp q_{on} is lower than some characteristic value that determines the position of the boundary G in the diagram of congested patterns in Fig. 18a. To understand the physics of the DGP, let us consider flow rates q_{in} that satisfy the condition $q_{\text{in}} > q_{\text{out}}$($q_{\text{out}}$ is the flow rate in the wide moving jam outflow under condition that in this jam outflow free flow is formed). After the first wide moving jam has been formed in the GP, the flow rate in the jam outflow cannot be greater than q_{out}. This means that rather than an initial high flow rate q_{in} a smaller flow rate related to the jam outflow determines the inflow into the pinch region of the GP. At a low enough flow rate to the on-ramp q_{on} (left of the boundary G in Fig. 18a), the pinch region cannot exist at this lower inflow into the pinch region: the pinch region dissolves.

As a result of this dissolution effect, either free flow or an LSP appears at the on-ramp while the wide moving jam moves on the main road upstream of the on-ramp (Fig. 19). From the above consideration of the physics of a DGP, one can conclude that the boundary G, which separates DGPs and GPs in the diagram of congested patterns, should intersect the boundary $S_{\text{J}}^{(\text{B})}$ at the point $q_{\text{in}} = q_{\text{out}}$. This is correct if hysteresis effects are not taken into account. However, the hysteresis effect of GP formation leads to the result that the line G does not necessarily intersect the boundary $S_{\text{J}}^{(\text{B})}$ at the point $q_{\text{in}} = q_{\text{out}}$ as is shown in Fig. 18a.

In contrast to the on-ramp bottleneck, at an off-ramp bottleneck (Fig. 17a) a GP *only* under weak congestion can be formed. This is true for all values $\eta < 100\%$, where η is the percentage of vehicles that want to leave the main road to the off-ramp in the flow rate q_{in}.

The pattern diagram at the off-ramp bottleneck, i. e., the regions of pattern formation in the plane (η, q_{in}) (Fig. 18b) qualitatively resembles the diagram at the on-ramp bottleneck (Fig. 18a). However, there is no saturation of the boundary $S_{\text{J}}^{(\text{B})}$ at a higher η. This is associated with the mentioned fact that in GPs only the weak congestion condition occurs at all $\eta < 100\%$. Indeed, all pinch region characteristics in GPs do depend on η.

Note that GP parameters depend on the average flow rate $q^{(\text{cong})}$ within a GP at a bottleneck. The time interval of flow rate averaging of $q^{(\text{cong})}$ is suggested to be considerably longer than time distances between any moving jams within the GP. For the GP, the average flow rate $q^{(\text{cong})}$ is equal to the average flow rate within the pinch region of synchronized flow: $q^{(\text{cong})} = q^{(\text{pinch})}$. In the theory of GPs and the pinch effect [41,42,45] was found that the greater the bottleneck influence (called the bottleneck strength) on traffic is, the smaller $q^{(\text{pinch})}$. In turn, the smaller $q^{(\text{pinch})}$ is, the greater the mean frequency of wide moving jams in a GP, and the lower the average speed between the wide moving jams.

In particular, for an on-ramp bottleneck at a given flow rate q_{in} the bottleneck strength is the greater, the greater

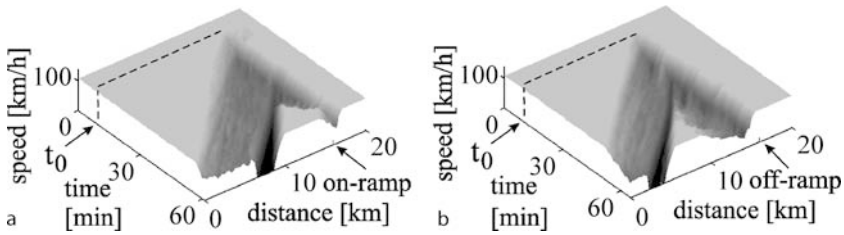

Traffic Congestion, Spatiotemporal Features of, Figure 19
Dissolving GP at on-ramp (a) and off-ramp (b) bottlenecks. Results of numerical simulations. Taken from [38]

the flow rate to the on-ramp q_{on}. For an Off-ramp bottleneck, at a given flow rate q_{in} the bottleneck strength is the greater, the greater the percentage of vehicles that leave the main road to the off-ramp η. However, as already mentioned for given on-ramp bottleneck (e. g., characteristics of vehicle merging from the on-ramp onto the main road) and traffic parameters (weather, percentage of long vehicles, etc.) and a given flow rate q_{in}, there is a saturation in decrease of $q^{(\text{pinch})}$ when q_{on} increases: beginning at a great enough q_{on}, the flow rate $q^{(\text{pinch})}$ reaches a limit (minimum) value $q_{\text{lim}}^{(\text{pinch})}$ [38,42]. In contrast, for the off-ramp bottleneck $q^{(\text{pinch})}$ decreases continuously when η increases up to $\eta \to 100\%$. At $\eta = 100\%$ all vehicles leave the main road to the off-ramp. Thus if the lane number on the main road is greater than the lane number of the off-ramp, then at $\eta = 100\%$ the off-ramp bottleneck can be considered a merge bottleneck caused by the lane reduction.

It can be seen from the diagrams of congested patterns (Fig. 18) that when the flow rates (and/or η for an off-ramp bottleneck) change, complex transformations between various SPs as well as between SPs and GPs occur. The latter is more probable under weak congestion conditions within the GPs. Moreover, there are also regions in these diagrams in which SPs and GPs are metastable patterns. This means that pattern transformation can also occur at given flow rates (and a given η) as a result of one of the local phase transitions (e. g., S → F, S → J, or J → S transitions) within an initial congested pattern (for more detail, see [38]).

Metastable free flow states with respect to SP formation as well as some of the metastable regions of SPs existence are considered in ► Traffic Breakdown, Probabilistic Theory of in connection with freeway capacity of free flow at a bottleneck, which is a very important traffic flow characteristic.

It should be noted that theoretical results discussed in Sects. "Congested Patterns at Isolated Bottlenecks" and "Diagram of Congested Patterns at Isolated Bottlenecks" are taken from simulations of three-phase traffic flow models of [41,42,45]. Recently other various traffic flow models in the framework of the three-phase traffic theory have been developed [5,15,23,27,44,52,53]. Features of congested patterns that these models exhibit [6,7,8,15,23,25,26,27,44,52,53,62,71] are similar to those found earlier in [41,42,45].

Complex Congested Patterns and Pattern Interaction

When there are two or more adjacent bottlenecks, there can be many new spatiotemporal phenomena. In particular, we consider the following empirical spatiotemporal features of congested patterns, which have been found out in empirical data [36,38]:

(i) Induced F → S transitions caused by upstream congested pattern propagation. A congested pattern, which occurs at the downstream bottleneck, can induce another congested pattern at the upstream bottleneck where free flow is realized before.
(ii) The catch effect. The catch of synchronized flow at an effectual bottleneck with the subsequent formation of a new congested pattern at this bottleneck.
(iii) An occurrence of expanded congested patterns (EP) where synchronized flow, which has initially appeared at the downstream bottleneck, affects the upstream bottleneck.
(iv) The influence of foreign wide moving jam propagation on a congested pattern at the upstream bottleneck. These foreign wide moving jams occur downstream of the congested pattern within a GP at a downstream bottleneck.
(v) An intensification of downstream congestion due to the onset of upstream congestion.

Induced F → S transition with catch effect. In an empirical example shown in Fig. 11, when the MSP has not yet reached bottleneck B_2, free flow is there. After the MSP is caught at the bottleneck, synchronized flow occurs at

bottleneck B_2. This synchronized flow continues for a long time (more than 2 hours) at bottleneck B_2. The upstream front of synchronized flow is now localized at some finite distance upstream of bottleneck B_2. This means that after the MSP has induced another SP at the bottleneck, instead of MSP propagation through the bottleneck, the MSP is caught at the bottleneck. Thus, the catch of the MSP at the bottleneck causes an induced speed breakdown at the bottleneck (these two effects are labeled "catch effect and induced F → S transition" in Fig. 11).

In the context of three-phase traffic theory, both induced traffic breakdown and catch effect can be found in accordance with empirical results. When two adjacent bottlenecks exist on a freeway and either an WSP or MSP appears at the downstream one, then the associated SP can be caught at the upstream bottleneck with the subsequent formation of a new congested pattern upstream of the upstream bottleneck (Fig. 20). For example, when the MSP shown in Fig. 20a reaches the upstream bottleneck, this synchronized flow causes an induced traffic breakdown at this upstream bottleneck. Synchronized flow is localized at some distance from the bottleneck, rather than the upstream front of this synchronized flow propagating further upstream continuously over time. This means that the initial WSP is caught at the upstream bottleneck. This catch effect causes the induced speed breakdown at the bottleneck. In another example, the upstream front of the WSP shown in Fig. 20b causes both induced traffic breakdown and catch effect after the upstream front of the WSP reaches the upstream bottleneck. Indeed, we can see that the initial WSP is caught at the upstream bottleneck. This catch effect causes also the induced speed breakdown at the bottleneck.

Expanded congested patterns (EP for short). An EP is *defined* as a congested traffic pattern whose synchronized flow affects at least two adjacent bottlenecks. A typical empirical example of EP is shown in Fig. 13a,c: the WSP that has occurred initially at the downstream off-ramp bottleneck B_{North1} exists only for a relatively short time, which is determined by the propagation of the upstream front of the WSP to the upstream on-ramp bottleneck B_{North2}. In this case, the lifetime of the WSP is about 20 min. After synchronized flow of the WSP is caught at the upstream bottleneck, a new congested pattern is formed. Synchronized flow in this pattern covers both bottlenecks B_{North1} and B_{North2}. Thus in accordance with EP definition, this congested pattern is an EP.

If the above empirical congested pattern (Fig. 13) is considered during a considerably longer time interval and on the whole freeway section on which measurements are available, we find a typical complex EP with several wide moving jams propagating through the EP and two upstream bottlenecks B_{North2} and B_{North3} while maintaining the mean velocity for the downstream jam front that are the same for these different jams (jams labeled 1,…,6 in Fig. 21).

What is typical in this EP? Firstly, synchronized flow affects several effectual bottlenecks. Secondly, additionally

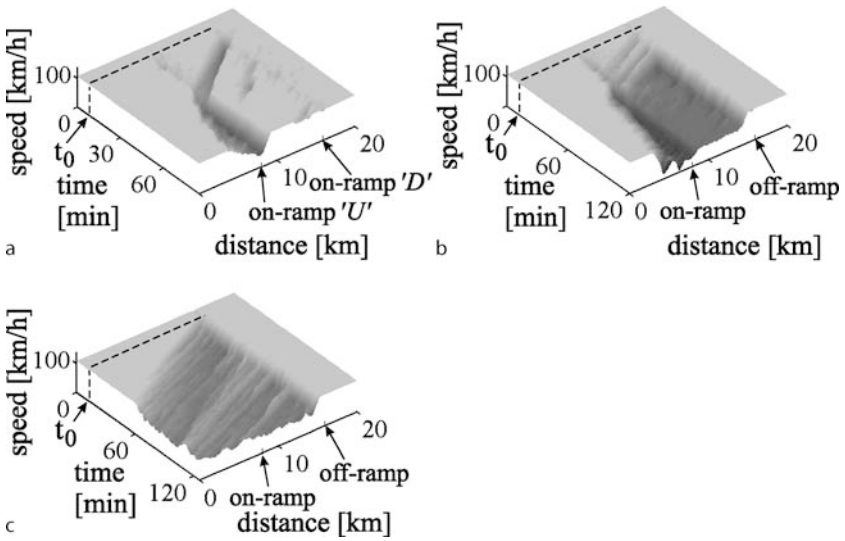

Traffic Congestion, Spatiotemporal Features of, Figure 20
Numerical simulations of induced pattern formation with the catch effect and expanded congested patterns (EPs). a, b Catch effect of MSP (a) and WSP (b) with induced formation of a new congested pattern at the upstream bottleneck. Taken from [38]

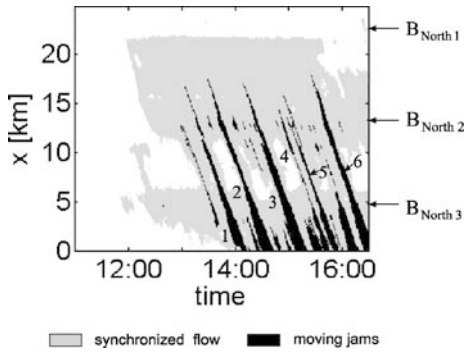

Traffic Congestion, Spatiotemporal Features of, Figure 21
Empirical example of an expanded congested pattern (EP). Free flow (*white*), synchronized flow (*gray*), moving jams (*black*). Taken from [38]

to wide moving jams that have emerged between bottlenecks B_{North1} and B_{North2}, new pinch regions in synchronized flow upstream of bottlenecks B_{North2} and B_{North3} are formed within which new moving jams emerge spontaneously. Many of these moving jams, however, are often suppressed by the initially wide moving jams propagating through the pinch region. An explanation of the wide moving jam suppression effect can be found in Sect. 7.6.3 of [38]. Thirdly, upstream and downstream of wide moving jams either synchronized flow or free flow can be formed within the same EP. Note that much more complex EPs can be observed, when effectual bottlenecks are located at short distances to each other (see Chap. 2 in [38]).

Simulations of a part of this EP (Fig. 22) made with the use of empirical road characteristics and flow rates are both qualitative and quantitative similar to the empirical EP (Figs. 15 and 21).

Foreign wide moving jams. A foreign wide moving jam is a wide moving jam, which has initially occurred downstream of an effectual bottleneck and propagates through synchronized flow of another congested pattern upstream of this bottleneck.

We mentioned above that wide moving jams in Fig. 21, which initially emerged upstream of the downstream bottleneck B_{North1}, propagate through the upstream bottleneck B_{North2} while maintaining the velocity of the downstream jam front. These wide moving jams subsequently propagate through the next upstream bottleneck B_{North3} as well.

Upstream of the bottlenecks B_{North2} and B_{North3} there are synchronized flows of other congested patterns formed. Thus the wide moving jams labeled 1, 2, ..., 6 in Fig. 21 that initially emerged downstream of the bottleneck B_{North2} are *foreign* wide moving jams when they propagate through synchronized flows upstream of the bottleneck B_{North2} or the upstream of bottleneck B_{North3} [38]. An example of numerical simulations of foreign wide moving jam propagation is shown in Fig. 22.

The identification of foreign wide moving jams is associated with the jam interaction effect; a foreign wide moving jam can exert a considerable influence on other moving jams that are just emerging in synchronized flow. In particular, the foreign wide moving jam can suppress the growth of a narrow moving jam that is close enough to the downstream front of the foreign wide moving jam.

It can turn out that in addition to foreign wide moving jams, new wide moving jams can emerge in the pinch region of synchronized flow upstream of the upstream bottleneck. As a result, a "united" sequence of wide moving jams is formed from the joining of the foreign wide moving jams and the wide moving jams that have emerged in synchronized flow upstream of the upstream bottleneck.

It should be noted that when the distances between adjacent effectual bottlenecks are great enough, spatially separated regions of synchronized flow can occur between adjacent effectual bottlenecks where the onset of congestion is realized. Often pinch regions are formed in these synchronized flows, leading to GP emergence. These GPs at different adjacent bottlenecks have spatially separated synchronized flow regions. We call these GPs "spatially separated" GPs [38].

An empirical example is shown in Fig. 23, where two spatially separated GPs occur at adjacent effectual bottlenecks B_2 and B_3. Synchronized flow upstream of the GP at bottleneck B_2 does not cover bottleneck B_3, i. e., no EP occurs between them. In other words, a congested pattern can consist of several spatially separated GPs, if there is enough distance between adjacent effectual bottlenecks (distances should usually be more than 4–5 km).

There can also be an intermediate case. An example can be seen between downstream bottleneck B_1 and upstream bottleneck B_2 in Fig. 23. We find that during some time intervals synchronized flow from the downstream GP at bottleneck B_1 covers bottleneck B_2, i. e., EPs appear. During other time intervals two spatially separated GPs at adjacent bottlenecks B_1 and B_2 are realized.

However, wide moving jams have emerged upstream of the downstream bottleneck B_1. The wide moving jams propagate through adjacent bottlenecks B_2 and B_3 while maintaining the velocity of the downstream jam fronts. These wide moving jams are foreign wide moving jams when they are upstream of adjacent bottlenecks B_2 and B_3, where other moving jams are emerging. Thus, these GPs have only spatially separated regions of synchronized flow. In contrast to the synchronized flow regions, wide moving jams of the GP at the downstream bottleneck propagate

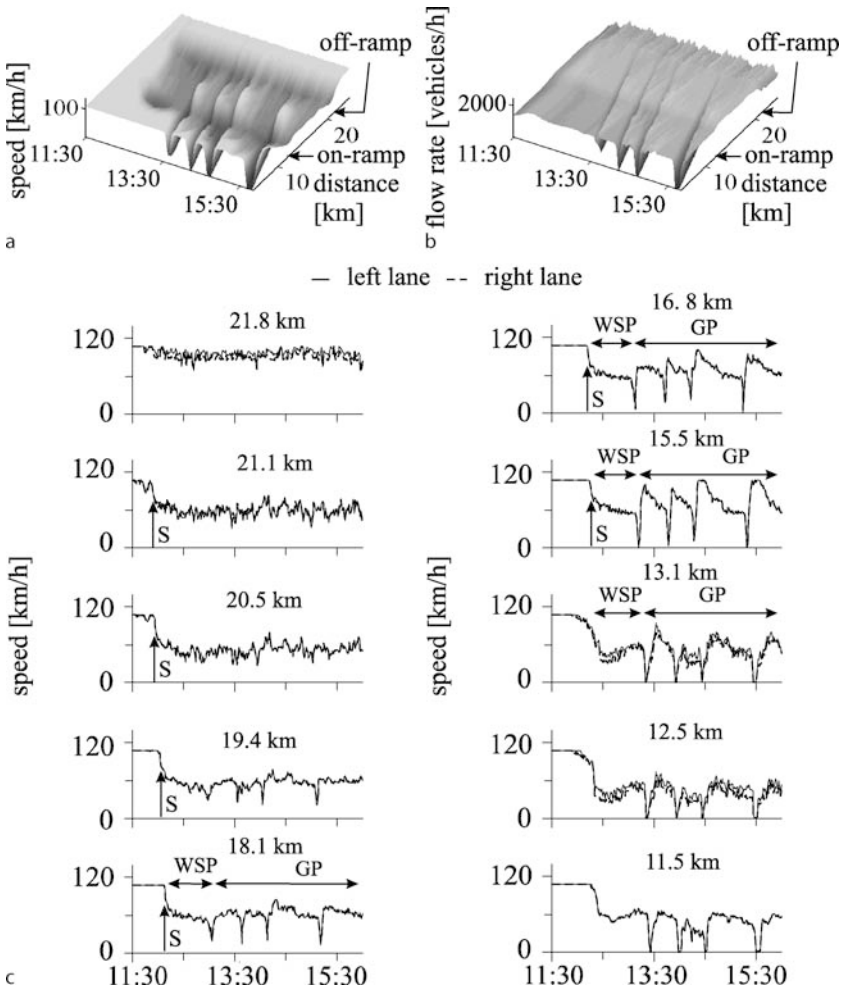

Traffic Congestion, Spatiotemporal Features of, Figure 22
Simulations of a part of empirical EP shown in Figs. 15 and 21: a, b Speed (a) and flow rate (b) in space and time. c Time-functions of speed at different locations. Taken from [46]

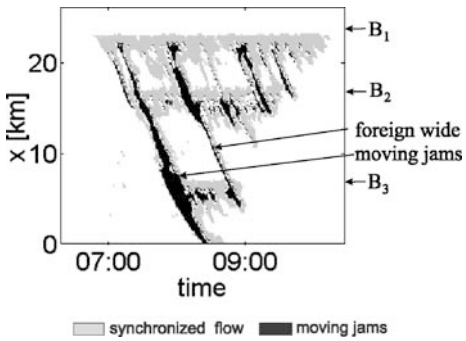

Traffic Congestion, Spatiotemporal Features of, Figure 23
Empirical example of foreign wide moving jam propagation. Free flow (*white*), synchronized flow (*gray*), and moving jams (*black*) in space and time. Taken from [38]

through the GP at the upstream bottlenecks. As already mentioned, this foreign jam propagation can considerably influence moving jam emergence at the upstream bottleneck. This usually occurs when spatially separated GPs appear at the freeway. Thus, due to foreign wide moving jam propagation, spatially separated GPs can form a complex congested pattern on the freeway (Fig. 23).

Intensification of downstream congestion due to upstream congestion. It can be expected that upstream congestion (congested pattern formation at an upstream bottleneck) should tend to a reduction in downstream congestion. This is because at the same traffic demand the flow rate in free flow downstream of the congested bottleneck (discharge flow rate) is usually lower than the initial flow rate in free flow at the same freeway location before

a congested pattern at the upstream bottleneck has been formed.

Nevertheless in some cases, we will see that rather than a *reduction* in downstream congestion, an *intensification* of downstream congestion can occur due to upstream congestion. This is the result of complex nonlinear interactions among congested patterns occurring at *different spatially separated* adjacent effectual bottlenecks.

To discuss an example of this effect, we consider a road with two on-ramp bottlenecks (Fig. 24). There are downstream and upstream on-ramp bottlenecks labeled '*D*' and '*U*', respectively. The initial flow rate in free flow upstream of the on-ramp '*D*'

$$q_{\text{in}}^{(\text{down})} = q_{\text{in}} + q_{\text{on}}^{(\text{up})} \tag{1}$$

in Fig. 24a are equal to each other in Fig. 24a,b. However, in Fig. 24a the flow rate

$$q_{\text{on}}^{(\text{up})} = 0\,, \tag{2}$$

whereas in Fig. 24b the flow rate

$$q_{\text{on}}^{(\text{up})} > 0\,. \tag{3}$$

The equality of the flow rate $q_{\text{in}}^{(\text{down})}$ (1) in Fig. 24a and b is realized due to greater flow rate q_{in} in free flow upstream of the upstream bottleneck in Fig. 24a in comparison with the flow rate q_{in} in Fig. 24b.

An initial DGP at the on-ramp '*D*', where *only one* wide moving jam emerges (Fig. 24a), transforms into a GP at this bottleneck where an *uninterrupted sequence* of wide moving jams emerges (Fig. 24b).

To explain this intensification of downstream congestion, note that after the wide moving jam in the DGP (Fig. 24a) emerges, the current flow rate upstream of the on-ramp '*D*', $q_{\text{in}}^{(\text{down})}$, has dropped to q_{out}:

$$q_{\text{in}}^{(\text{down})} = q_{\text{out}}\,. \tag{4}$$

This decrease in the flow rate $q_{\text{in}}^{(\text{down})}$ occurs because vehicles merging onto the main road from the on-ramp '*U*' must slow down when they reach the wide moving jam. In other words, when the wide moving jam is between the on-ramps '*D*' and '*U*' the flow rate $q_{\text{on}}^{(\text{up})}$ does not lead to an increase in the flow rate $q_{\text{in}}^{(\text{down})}$: due to the flow rate $q_{\text{on}}^{(\text{up})}$ only the jam inflow increases. In the example, the jam outflow $q_{\text{out}} = 1810$ vehicles/h. At the flow rate $q_{\text{in}}^{(\text{down})} = 1810$ vehicles/h the GP cannot exist upstream of the on-ramp '*D*' at the chosen flow rate to the on-ramp '*D*', $q_{\text{on}}^{(\text{down})}$ (Fig. 24a).

In Fig. 24b, an LSP is induced at the on-ramp '*U*' after the wide moving jam from a DGP (downstream congestion) has reached the on-ramp '*U*'. After the wide moving jam has passed the on-ramp '*U*', there are no wide moving jams between the on-ramps '*D*' and '*U*'. For this reason, vehicles merging onto the main road from the on-ramp '*U*' cause an increase in the flow rate $q_{\text{in}}^{(\text{down})}$ to the value $q_{\text{in}}^{(\text{down})} = 2150$ vehicles/h. This flow rate is much higher than 1810 vehicles/h. This leads to subsequent wide moving jam emergence at the on-ramp '*D*', and so on. As a result, a GP is formed at the on-ramp '*D*' (Fig. 24b).

The mean period of moving jam emergence in the GP is determined by the time of wide moving jam propagation to the on-ramp '*U*'. Before a wide moving jam has passed the on-ramp '*U*' the condition (4) is satisfied, and no moving jam can emerge at the given flow rate $q_{\text{on}}^{(\text{down})}$ at the on-ramp '*D*'.

Reproducible and Predictable Congested Patterns

It has been found that for each effectual bottleneck or each set of several adjacent effectual bottlenecks, where congested patterns occur, the spatiotemporal structure of congested patterns exhibits predictable, i. e., characteristic, unique, and reproducible features [38]. These predictable and reproducible pattern features can be almost *the same*

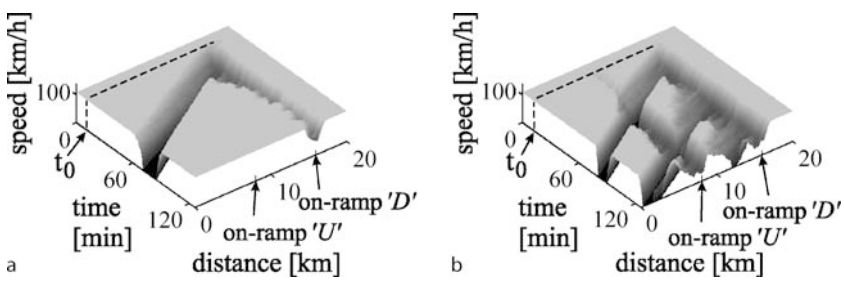

Traffic Congestion, Spatiotemporal Features of, Figure 24
Simulations of intensification of upstream congestion due to downstream congestion. Speed on the main road in space and time: a, b DGP (a) transforms into a GP (b). Taken from [38]

for different days and years. They can also persist over a large range of flow rates (traffic demand) at which the patterns exist. The predictable and reproducible pattern features remain in traffic flows with very different driver behavioral characteristics and different vehicle parameters. The predictable and reproducible pattern features can be used in the forecast of congested patterns at freeway bottlenecks. Overviews of some of the reproducible congested pattern features can be seen in Fig. 25.

There are pattern features that have a probabilistic nature. This means that in different realizations (days) congested pattern features can be different for the same traffic control parameters (the percentage of long vehicles, weather, other road conditions), traffic demand, and initial conditions. However, we can identify different "degrees of predictability" for different pattern features. Some "deterministic" pattern features and pattern features with a high "degree of predictability" are as follows:

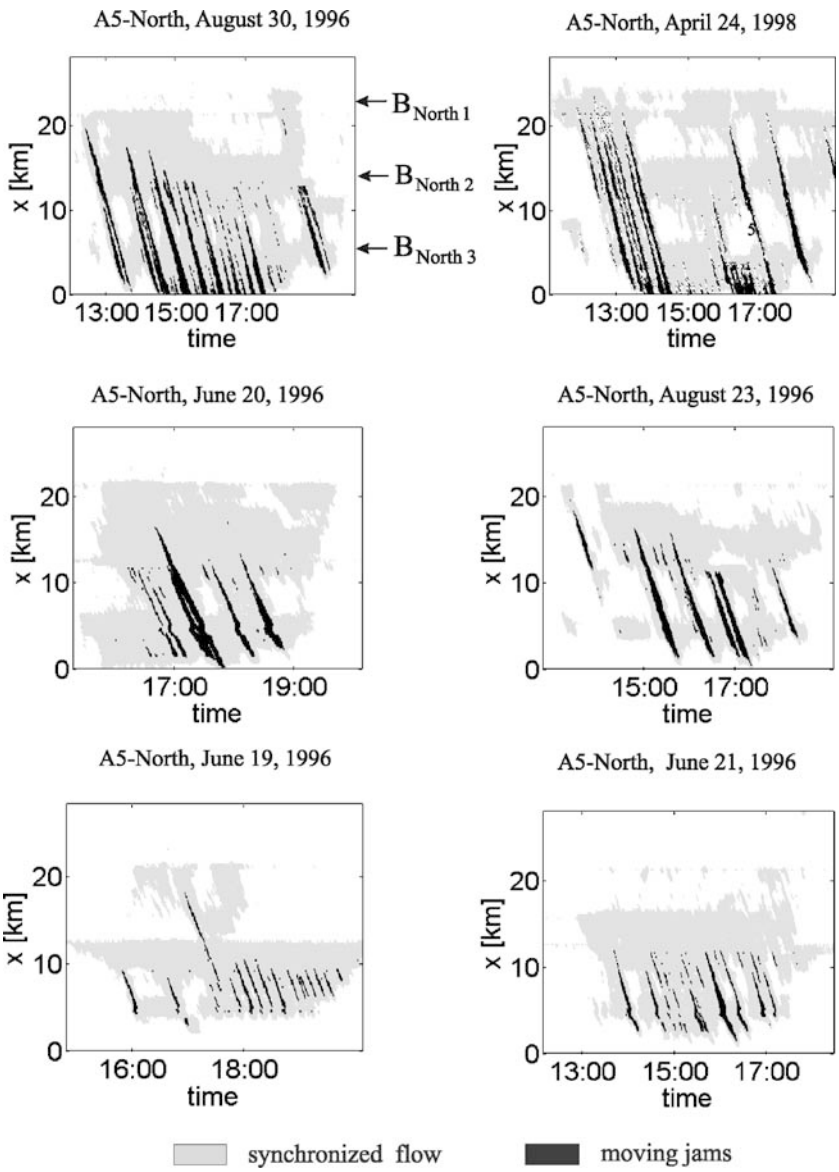

Traffic Congestion, Spatiotemporal Features of, Figure 25
Empirical examples of predictable and reproducible patterns. Overviews of free flow (*white*), synchronized flow (*gray*), and moving jams (*black*) in space and time. The effectual bottlenecks are the same as those in Fig. 21 [38]

(1) Traffic breakdown (F → S transition) occurs mostly at the same effectual freeway bottlenecks on different days and years.
(2) In contrast to synchronized flow, which occurs mostly at bottlenecks, wide moving jam emergence is usually observed in synchronized flow at a finite distance upstream of bottlenecks, if distances between the bottlenecks are large enough.
(3) Wide moving jams propagate through any complex state of traffic and through bottlenecks at the mean velocity of the downstream jam front.
(4) If free flow emerges in the wide moving jam outflow, then the flow rate in this jam outflow is considerably lower than the maximum possible flow rate in free flow.
(5) The catch effect: a local synchronized flow region reaching a bottleneck is caught at the bottleneck, when free flow at the bottleneck is in a metastable state to traffic breakdown.
(6) If a narrow moving jam is very close to the downstream front of a foreign wide moving jam, the foreign jam suppresses the growth of this narrow jam.

Some traffic pattern features with a middle "degree of predictability" are as follows:

(i) The instant of spontaneous traffic breakdown at an effectual bottleneck at the same time-dependences of upstream flow rates. This reveals the probabilistic nature of an F → S transition.
(ii) For a bottleneck or for a set of adjacent effectual bottlenecks that are close to one another there is a certain type of spatiotemporal congested pattern that appears with the highest probability, given the same traffic demand.

Some pattern features with a low "degree of predictability" are as follows:

(a) Whether a spontaneous or an induced traffic breakdown occurs, leading to synchronized flow at an effectual bottleneck.
(b) The instant of wide moving jam emergence in synchronized flow.

Microscopic Features of Traffic Phases

Moving Blanks Within Wide Moving Jams

The spatiotemporal criterion for a wide moving jam [J] of Sect. "Introduction" that defines the wide moving jam phase in congested traffic can be explained by a traffic flow interruption that occurs when vehicles are in a standstill within the jam. A sufficient criterion for this flow interruption within the jam is [47]

$$I_s = \frac{\tau_{\max}}{\tau_{\mathrm{del}}^{(a)}} \gg 1 \,, \tag{5}$$

where $\tau_{\max}$ is the maximum time headway between two vehicles within a wide moving jam and $\tau_{\mathrm{del}}^{(a)}$ is the mean time delay in vehicle acceleration at the downstream jam front from a vehicle standstill; $\tau_{\mathrm{del}}^{(a)}$ determines the jam outflow; corresponding to empirical results $\tau_{\mathrm{del}}^{(a)} \approx$ 1.5–2 sec [38].

The flow interruption effect is a general effect for each wide moving jam (see examples of flow interruption intervals labeled by "flow interruption" in Figs. 26c and 27a). For this reason, the criterion (5) can be considered as a microscopic criterion for the wide moving jam phase (more detailed explanations of flow interruption intervals and the criterion (5) for wide moving jams appear in ► Traffic Congestion, Modeling Approaches to and Sect. "Congested Patterns at Heavy Bottlenecks"). Between flow interruption intervals within the wide moving jam shown in Fig. 26b–g, vehicles within the jam exhibit time headways about 1.5–7 sec (Fig. 27b). These time headways are considerably shorter than flow interruption intervals (about 20 sec and longer) in Fig. 26c,e,g. The time headways are related to low speed states measured at detectors within the jam (Fig. 26b,d,f).

To understand the effect of these low speed states, note that when vehicles meet the wide moving jam, firstly they decelerate to a standstill at the upstream jam front. As a result, the first flow interruption interval in all lanes appears (Figs. 26c,e,g and 27a). Net distances (space gaps) between these vehicles can be very different and the mean space gap can exceed a minimum (safe) space gap considerably, i. e., regions with no vehicles can appear within the jam. These regions with no vehicles are called *blanks* within the wide moving jam.

Later vehicles move covering these blanks. This low speed vehicle motion is responsible for low speed states mentioned above (Fig. 26b,d,f). Consequently, due to this vehicle motion new blanks between vehicles occur upstream, i. e., the blanks move upstream within the jam. Then other vehicles within the jam that are upstream of these blanks begin also to move covering the latter blanks. This leads to moving blanks that propagate upstream within the jam.

Thus we define a *moving blank* as a blank between vehicles, which moves upstream due to vehicle motion within the jam. The vehicle motion occurs at a low vehi-

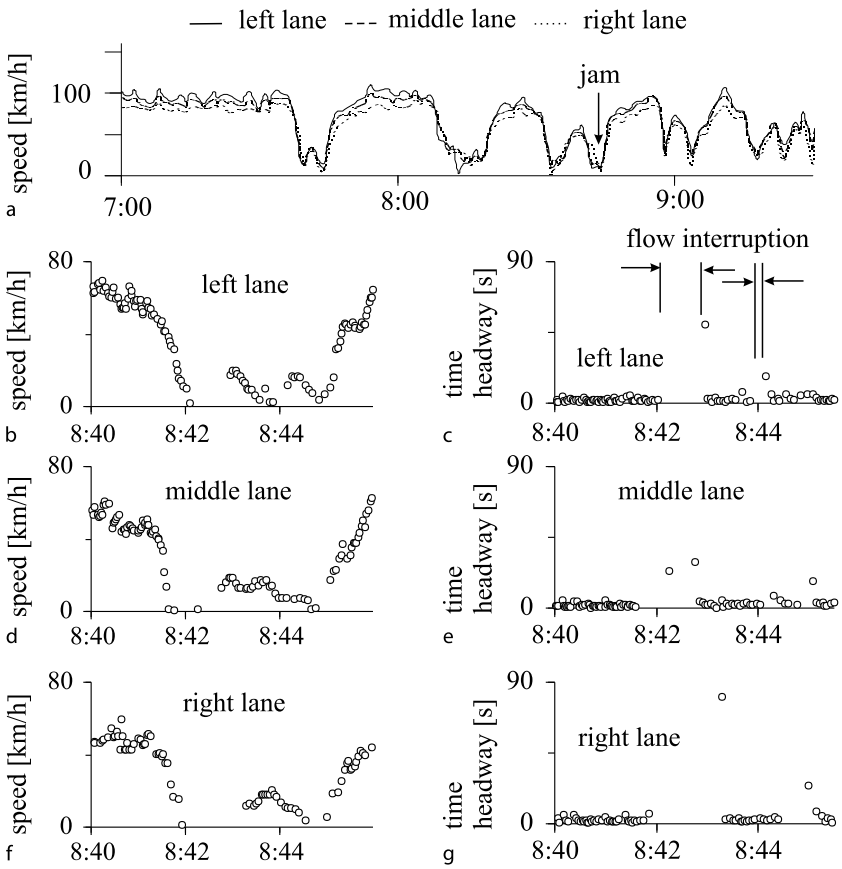

Traffic Congestion, Spatiotemporal Features of, Figure 26

Measured single vehicle data analysis: a Overview of measured data (1 min average data). b–g Single vehicle data for speed in three freeway lanes for a wide moving jam (*left figures*) labeled by "jam" in a and the related time headways (*right figures*). Taken from [47]

cle speed, i. e., the vehicle motion creating moving blanks within the jam is related to low speed states discussed above (Fig. 26b,d,f); these low speed states are the reason for moving blank emergence within the jam. This explains why moving blanks are associated with low speed states within the jam.

Time headways τ between vehicles in these low speed states with moving blanks are about 1.5–7 sec (Fig. 27b). Because these low speed states are the reason for moving blanks, we will refer the time headways in the low speed states to moving blanks. Thus the term *time headways for moving blanks* means the time headways in the low speed states causing upstream motion of blanks.

To explain possible short time headways associated with moving blanks observed in empirical data (about 1.5 sec), note that when vehicles begin to cover a blank within a wide moving jam, they start to move from a standstill within the jam. Thus, some of the vehicles can move with time headways between each other that are close to a time delay in vehicle acceleration from a standstill within the jam whose average value is $\tau_{\text{del}}^{(a)} \approx 1.5$–2 sec.

Time headways in low speed states within the jam, i. e., time headways associated with moving blanks (about 1.5–7 sec in Fig. 27b) must be distinguished from long time intervals of traffic flow interruption within the jam (about 20 sec and longer) (Figs. 26c,e,g and 27a). The flow interruption intervals within the jam occur when some of the vehicles within the jam do not move. This vehicle stoppage does not indicate whether there are some blanks between these vehicles or not. This is because during these long flow interruption intervals space gaps between vehicles that are stopped within the jam might be quite small.

Thus at a fixed freeway location, within a wide moving jam, long time intervals, in which no vehicle passes the location (flow interruption intervals), alternate with time intervals in which vehicles pass the location with low speeds and time headways of about 1.5–7 sec associated with moving blanks within the jam. This empiri-

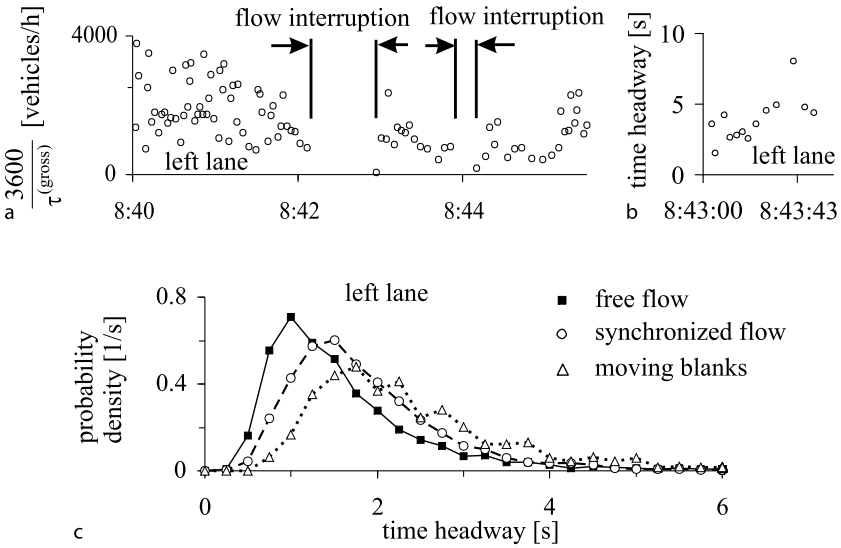

Traffic Congestion, Spatiotemporal Features of, Figure 27

Measured microscopic characteristics of traffic phases associated with single vehicle data shown in Fig. 26b–g: a Time distributions of the value $3600/\tau^{(gross)}$ in the left lane ($\tau^{(gross)}$ is the gross time headway). b Time headways associated with moving blanks in the left lane related to the jam shown in Fig. 26b. c Probability density of time headways in the left lane for free flow (*solid curve*), synchronized flow (*dashed curve*), and moving blanks within wide moving jams (*dotted curve*). Taken from [47]

cal result could be associated with the following microscopic jam space structure: freeway regions in which vehicles are in a standstill (traffic flow interruption) alternate with freeway regions in which blanks moving upstream (moving blanks) occur resulting from vehicle motion with low speeds.

Empirical data show that low speed states within wide moving jams associated with moving blanks are not necessarily synchronized between different lanes (Fig. 26b,d,f).

Moving blanks within a wide moving jam resemble electron holes in the valence band of semiconductors. As the moving blanks that propagate upstream appear due to downstream vehicle motion within the jam, so appearance of electron holes moving with the electric field results from electron motion against the electric field in the valence band of semiconductors. However, an electron hole is associated with a single vacancy for one electron in the valence band. In contrast, there can be wide (in the longitudinal direction) moving blanks associated with a "vacancy" for two or more vehicles within a wide moving jam. This moving blank can be split into two or more moving blanks during upstream blanks motion. On the other hand, several adjacent moving blanks can merge into one moving blank (a more detailed consideration of the spatiotemporal dynamics of moving blanks and flow interruption intervals appear in Sect. "Congested Patterns at Heavy Bottlenecks").

One of the microscopic characteristics of the traffic phases is time headway distributions. In particular, it has been found that the mean time headway for free flow is shorter than for congested traffic (see references in [20]). This can be explained at least by the following peculiarities of free flow and congested traffic:

(i) A vehicle can come very close to the preceding vehicle just before the vehicle passes it. However, a road detector cannot recognize whether a very short time headway that has been measured is related to car-following or to this vehicle passing effect.
(ii) A vehicle can easily change the lane, if the preceding vehicle would decelerate unexpectedly; thus, the vehicle can chose a shorter time headway in car-following than in congested traffic.
(iii) There is a frustration effect in congested traffic that can be responsible for some long time headways [20].

There can also be another reason for shorter mean time headways in free flow in comparison with mean time headways in congested traffic. It has been found that there are different distributions of time headways in synchronized flow and moving blanks associated with wide moving jams (Fig. 27c). It can be assumed that as long as vehicles are within a wide moving jam, the vehicles have no hurry while covering blanks with the jam.

Traffic Phases in Flow-Density Plane

There are a huge number of publications associated with an analysis of measured traffic data in the flow-density plane (see references in [2,20,57]). One of the aims of this analysis is to find criteria for a classification of qualitatively different traffic states and phases associated with congested traffic (see, e. g., [49,50,60]). However, as mentioned above traffic is a complex process in space and time, i. e., rather than a data analysis in the flow-density plane, to find qualitatively different traffic states and phases one should firstly study spatiotemporal features of traffic congested patterns. Otherwise, invalid conclusions about features of traffic states and phases can be made very easily.

To illustrate this, we consider single vehicle characteristics of low vehicle speeds within a wide moving jam associated with moving blanks (cross and triangle points) with free flow (black squares) and synchronized flow (circles) in the flow-density plane (Fig. 28a). Note that in contrast with single-vehicle data for a wide moving jam shown in Fig. 26b–g, which are used for a calculation of data for moving blanks shown in Fig. 28a, there are no flow interruption intervals in low speed data shown in Fig. 28c,d, i. e., the latter data are associated with the synchronized flow phase.

As well-known [2,20,57], in free flow the flow rate is on average an increasing density function in the flow-density plane (black squares in Fig. 28a). Due to small density states associated with low vehicle speeds within wide moving jams (cross points in Fig. 28a), the flow rate might also be considered on average an increasing density function in the flow-density plane for wide moving jams. For synchronized flow (Fig. 28a), measured points for synchronized flow exhibit the great scattering in the flow-density plane. These results have be used in [49,50,60] for the definition of a criterion for the traffic phases in measured data of congested traffic based on the flow-density correlation function

$$cc_{\rho,q}(\tau) = \frac{\langle \rho(t)q(t+\tau)\rangle - \langle \rho(t)\rangle\langle q(t+\tau)\rangle}{\sqrt{\langle \rho^2(t)\rangle - \langle \rho(t)\rangle^2}\sqrt{\langle q^2(t)\rangle - \langle q(t)\rangle^2}} \,. \quad (6)$$

In accordance with [49,50,60], small flow-density correlations in the data should be associated with synchronized flow, while large flow-density correlations in the data for congested traffic – with wide moving jams. Indeed, for data shown in Fig. 28 we find that the flow-density correlation coefficient (6) for synchronized flow is a very small value $cc_{\rho,q}^{(\mathrm{syn})}(0) \approx 0.06$, whereas for the wide moving jam $cc_{\rho,q}^{(\mathrm{jam})}(0) \approx 0.67$, i. e., it is comparable with the correlation coefficient for free flow $cc_{\rho,q}^{(\mathrm{free})}(0) \approx 0.97$.

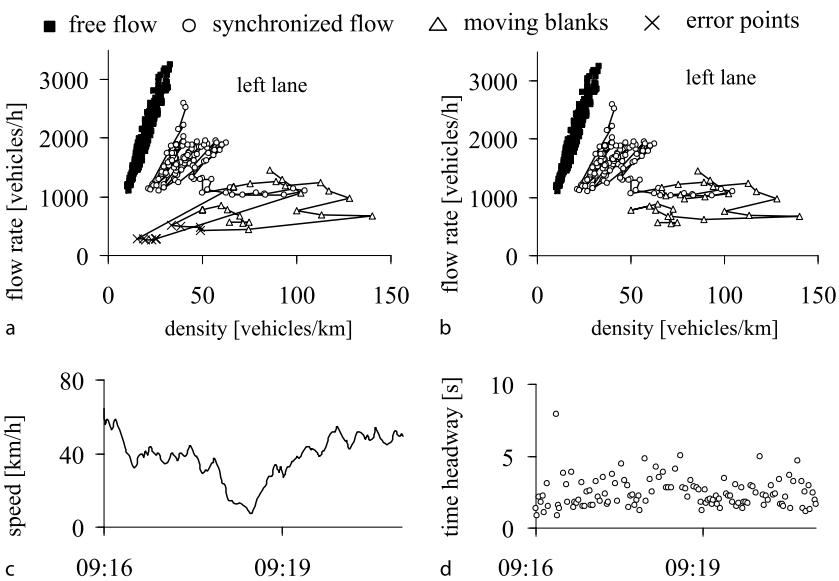

Traffic Congestion, Spatiotemporal Features of, Figure 28

Empirical characteristics of traffic phases: a Traffic phases in the flow-density plane for data in the left lane. b The same traffic phases as those in a, however, with *improved data* for moving blanks. Microscopic data for moving blanks, for synchronized flow and free flow used in a, b are associated with single-vehicle data, respectively, for wide moving jam shown in Fig. 26b,c, for synchronized flow shown in c, d, and for free flow shown in Fig. 26a. c, d Single-vehicle speeds (c) and time headways (d) for synchronized flow states used in a, b that are associated with data shown in Fig. 26a. Moving averaging over five vehicles is used. Taken from [47]

However, these conclusions of the definition for the traffic phases of [49,50,60] based on the flow-density correlation coefficient (6) are *invalid*. As explained in [47], the increasing density function (Fig. 28a) for wide moving jams in the flow-density plane as well as the associated great value of flow-density correlations are associated with a *large systematic error* in the data measured at road detectors. This error is due to incorrect density estimation within moving jams made in [49,50,60]. To understand this critical conclusion, recall that there are traffic flow interruption intervals within the wide moving jam (Fig. 26b–g). This leads to an incorrect calculation of the average speed v within the jam. The average speed measured by a detector can be related to a time interval, which includes at least one traffic flow interruption interval. In this case, the measured average speed is considerably higher than the real average speed within the jam. This is because vehicles do not move during traffic flow interruption intervals. As a result, the subsequent density estimation through the formula $\rho = q/v$ (q the flow rate) leads to very small densities within the jam. However, the real density of vehicles stopped within the jam during traffic flow interruption is very large. Thus the measured data associated with the traffic flow interruption effect within wide moving jams cannot be used for density estimation within the jams. To estimate real vehicle density within a wide moving jam during time intervals that include flow interruption intervals, the density definition (number of vehicles per freeway length at a given time moment), i. e., the *spatial averaging* should be used, rather than density estimation through the use of the formula $\rho = q/v$ in which the flow rate and average speed are related to the *time averaging* of vehicles passing a detector during a given time interval. The conclusion about the large systematic error in density estimation within wide moving jams is confirmed by a numerical analysis discussed in Subsect. "Numerical Simulations of Moving Blanks Within Wide Moving Jams".

This analysis shows also that within wide moving jams density estimation $\rho = q/v$ exhibits a relative small error during vehicle motion associated with moving blanks within the jams. This is because in this case the associated time headways are related to vehicles moving within the jams. If only these single vehicle data is used for density estimation, then points for moving blanks within the jam appear that are associated with random transformations in different directions in the flow-density plane (triangle points in Fig. 28b) rather than an increasing function of the flow rate with density (cross and triangle points in Fig. 28a). This is explained by low vehicle speed states that are associated with non-regular vehicle motion covering blanks within the wide moving jam. If error measurement points are removed (Fig. 28b), then the flow-density correlation coefficient for the wide moving jam $cc_{\rho,q}^{(\mathrm{jam})}(0) \approx 0.18$, i. e., it is small. This means that in contrast with [49,50,60] no valid traffic phase definition based on the flow-density correlations in congested traffic can be made.

In addition, we should mention that synchronized flow states can overlap states associated with moving blanks within a wide moving jam as this is shown in Fig. 28b. This is because the speed in synchronized flow states can be as low as the speed in states associated with moving blanks within the jam. Thus in a general case low speed states associated with moving blanks within wide moving jams cannot also be used for clearly distinguishing the synchronized flow and wide moving jam phases in congested traffic.

Moreover, the result mentioned above that synchronized flow exhibits a small flow-density correlation, as this is the case for measured data shown in Fig. 28b,c, is *not* a general one. An empirical example of synchronized flow, for which the flow-density correlation coefficient $cc_{\rho,q}^{(\mathrm{syn})}(0) \approx 0.91$ is as great as the one for free flow, is shown in Fig. 29. Thus we see that in contrast with [49,50,60] a great value of the flow-density correlation cannot be used as the criterion for distinguishing free flow from synchronized flow.

Numerical Simulations of Moving Blanks Within Wide Moving Jams

To understand features of propagation and evolution of moving blanks within a wide moving jam, we consider a moving jam that is induced downstream of an on-ramp bottleneck by applying of a short-time and high amplitude local speed disturbance in heterogeneous free flow consisting of fast and long vehicles (Fig. 30a,b). There are two time intervals in which flow interruption within the jam occurs. The maximum time headways (Fig. 30c) associated with these flow interruption effects satisfy the microscopic criterion (5) for wide moving jams ([47], ► Traffic Congestion, Modeling Approaches to) (time delays used in (5) for fast and long vehicles are $\tau_{\mathrm{del}}^{(\mathrm{a})} \approx 1.77$ and 3.33 sec, respectively). As a result, in accordance with the macroscopic spatiotemporal criterion [J] for the wide moving jam phase, this jam propagates through the bottleneck while maintaining the mean velocity of the downstream jam front.

Within the wide moving jam, between time intervals of flow interruptions there are intervals in which low speed states appear associated with moving blanks (Fig. 30b–d).

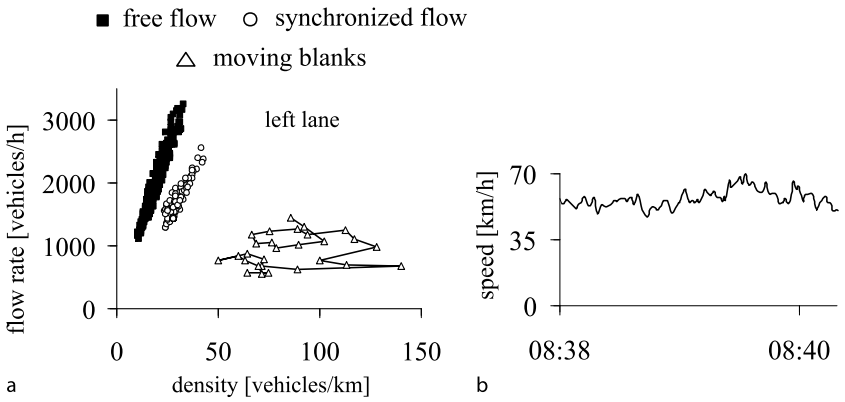

Traffic Congestion, Spatiotemporal Features of, Figure 29

Empirical characteristics of traffic phases: a Traffic phases in the flow-density plane for data in the left lane with improved data for moving blanks. Data for free flow and wide moving jam are the same as those in Fig. 28b. b Single-vehicle speeds for synchronized flow used in a are associated with data shown in Fig. 26a. Moving averaging over five vehicles is used

Simulated time headways for moving blanks that are about 2.5–7.5 sec in the left lane (Fig. 30d, left) correspond to empirical observed values (Fig. 27). As can be seen from vehicle trajectories within the jam, these low speed states associated with moving blanks are related to non-homogeneous vehicle motions that cover blanks between vehicles within the jam (Fig. 31). These moving blanks propagate upstream with a negative velocity that is on average equal to the characteristic speed of the downstream jam front. Spatial and temporal speed distributions in low speed states associated with moving blanks exhibit complex variations within the jam, which are different in the left and right lanes (Fig. 31b,c); these variations seem to correspond to a random vehicle speed behavior within the jam.

In accordance with empirical results of Subsect. "Traffic Phases in Flow-Density Plane", due to flow interruption within the wide moving jam shown in Fig. 30a–c, there is a large error in density distributions calculated through the formula $\rho = q/v$ at greater density (curves 2 in Fig. 30e) in comparison with density distributions found based on the density definition (vehicles per freeway length) (curves 1). Indeed, curves 2 in Fig. 30e show a significant density underestimation within moving jams. These error points (crosses in Fig. 32a) explain the systematic error in the empirical studies of states within wide moving jams made in [49,50,60] that has been illustrated in Fig. 28a (compare error points in Figs. 28a and 32a).

Within the regions of moving blanks, the difference between density calculated through the formula $\rho = q/v$ (curves 2 in Fig. 30e) and density found based on the density definition (curves 1) decreases considerably. For this reason, if error points associated with flow interruption within the jam are removed, then remaining points in the flow-density (Fig. 32b) exhibit a qualitative correspondence with the states within the jam calculated through the density definition (Fig. 32c). The latter states are related to the speed and density found at a fixed time moment 31.5 min (Fig. 30f,g). Nevertheless, there are some quantitative differences between speed and density distributions found through detector measurements (Fig. 32b) and through the use of the density definition in Fig. 32c. In particular, there are points in the flow-density plane found in the density calculation through the density definition (Fig. 32c), which are related to greater density up to the maximum jam density. These points cannot usually be found based on traffic measurements at a detector location (Fig. 32b).

To understand possible scenarios of emergence of moving blanks within a wide moving jam, spontaneous jam emergence in synchronized flow of a general pattern (GP) at an on-ramp bottleneck has been simulated in heterogeneous flow consisting of passenger cars and long vehicles (Fig. 33). Firstly, an F → S transition occurs spontaneously at the bottleneck. Synchronized flow that appears due to this F → S transition propagates upstream. Moving jams emerge spontaneously in that synchronized flow propagating upstream.

Simulations show that there are the following mechanisms for emergence of moving blanks within the jams: (i) Vehicle lane changing at the upstream jam front. (ii) Vehicles come to a stop at the upstream front of the wide moving jam at different space gaps to the related preceding vehicles. (iii) When vehicles move with low speeds cover-

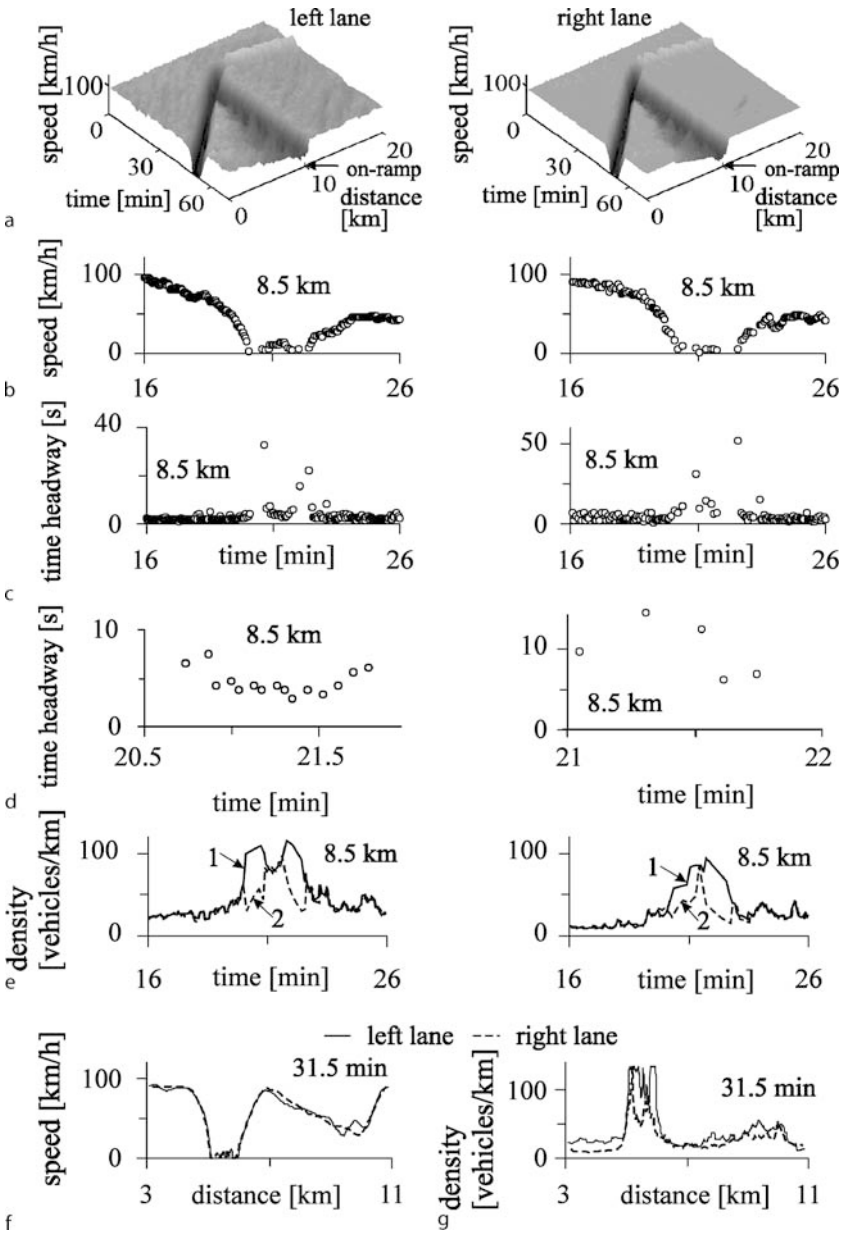

Traffic Congestion, Spatiotemporal Features of, Figure 30

Simulated characteristics of moving blanks within a wide moving jam propagating through an on-ramp bottleneck. a Speed in space and time on the main road. b–e Single vehicle data for time dependences of speed (b), time headways (c, d), and the associated density distributions calculated through the density definition (*solid curves* 1) and, as in empirical Fig. 28a, via the formula $\rho = q/v$ (*dashed curves* 2) (e) at location $x = 8.5$ km. *Left* and *right* figures (a–e) are related to the left and right lanes, respectively. f, g Speed (f) and density (g) related to a as functions of distance at time 31.5 min in the left and right lanes. Taken from [47]

ing downstream blanks within the jam, vehicle lane changing can result in new moving blanks.

In heterogeneous flow, moving blanks emerge most frequently due to vehicle lane changing (item (i) and (iii)). Specifically, a moving jam emerges firstly in the right lane only ($t = 23$ min in Figs. 33b,c and 34a). Then some vehicles change from the right lane to the left lane (arrow 1 in Fig. 34a). On the one hand, this lane changing decreases speed in the left lane. This leads to narrow jam formation in this lane. On the other hand, blanks between vehicles

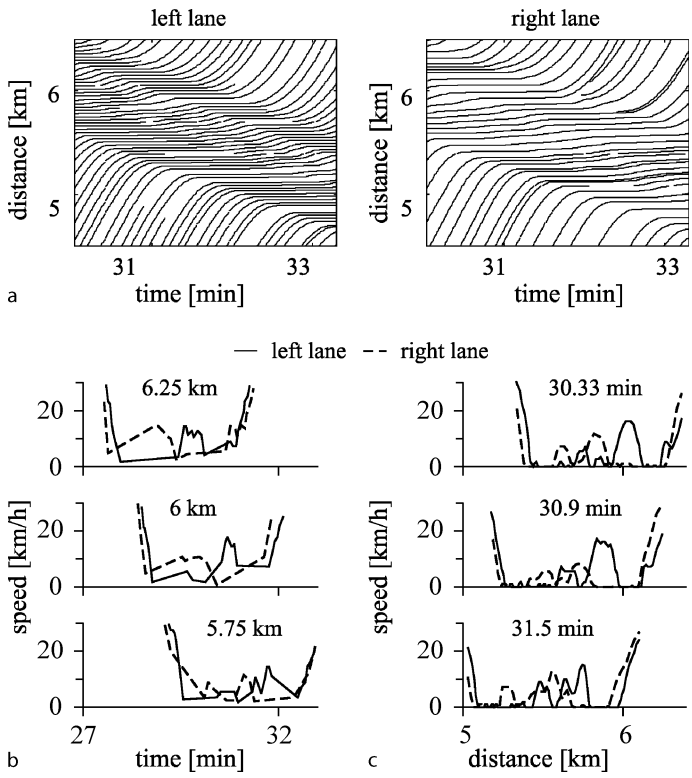

Traffic Congestion, Spatiotemporal Features of, Figure 31

Simulated microscopic characteristics of moving blanks within the wide moving jam shown in Fig. 30a: a Vehicle trajectories for each 4th vehicle in the left (*left*) and right (*right*) lanes. b, c Low speed states associated with moving blanks at three different locations (b) and three different time moments (c) in the left and right lanes. Taken from [47]

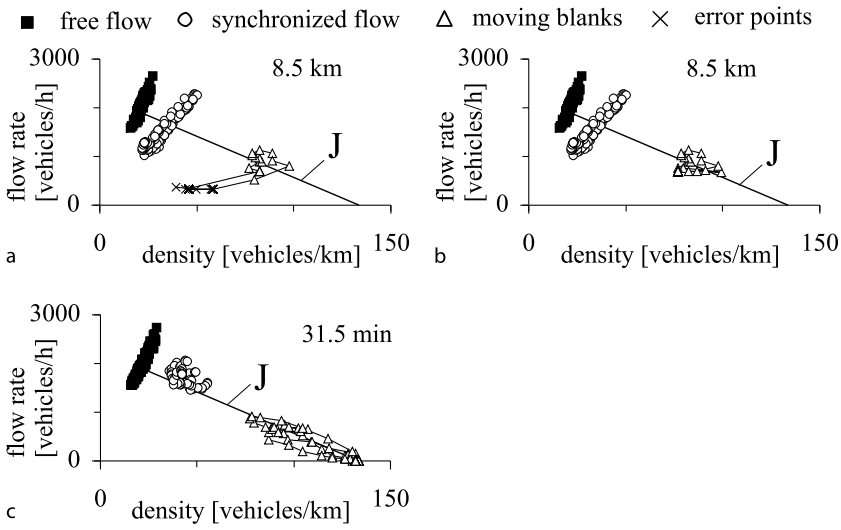

Traffic Congestion, Spatiotemporal Features of, Figure 32

Simulated characteristics of moving blanks within the wide moving jam in the left lane shown in Fig. 30a together with states of free flow and synchronized flow in the flow-density plane. a, b States within the jam that are determined at detector at $x = 8.5$ km through the formula $\rho = q/v$ in two cases in which all states within the jam measured at the detector are shown (a) and error states are removed (b). c States within the jam that are determined through the density definition (vehicles per freeway length) at time 31.5 min. J is the line J for the downstream jam front. Taken from [47]

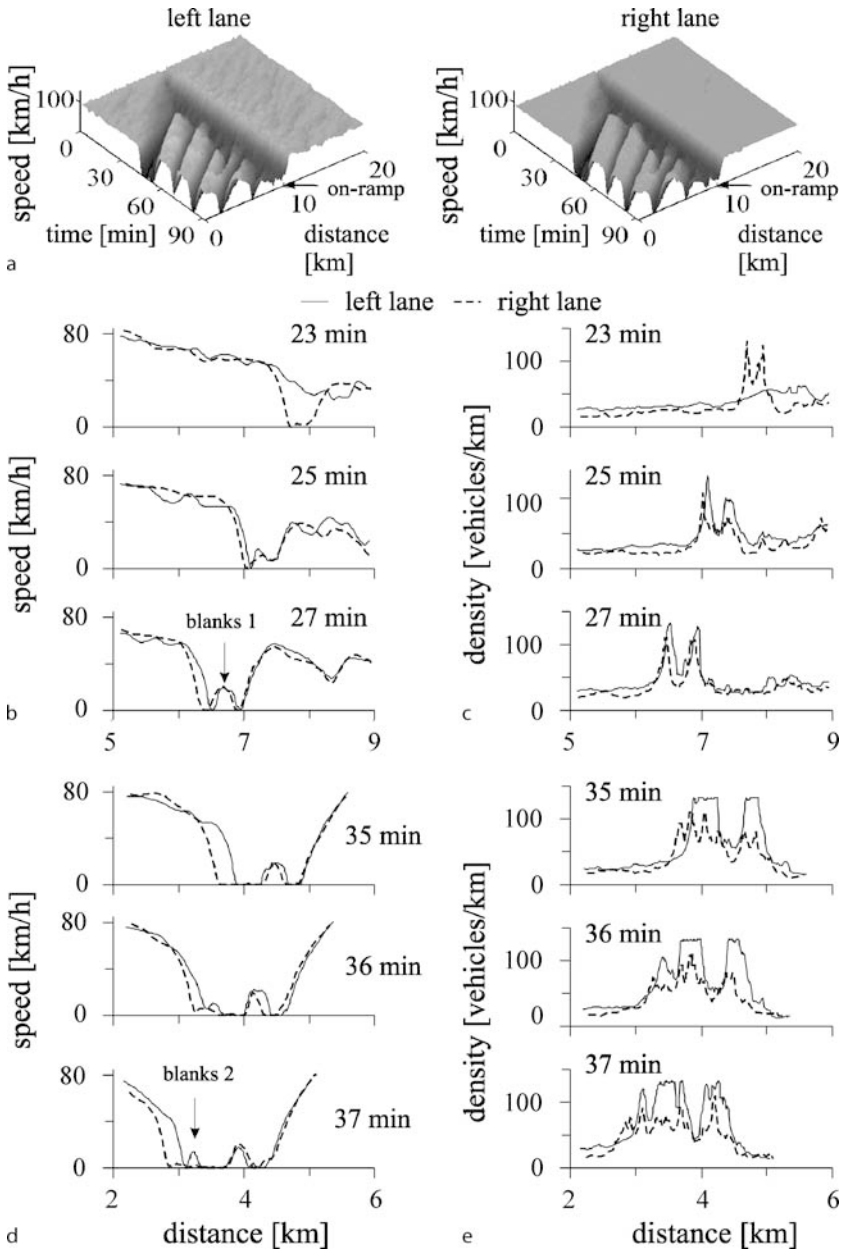

Traffic Congestion, Spatiotemporal Features of, Figure 33

Simulated characteristics of moving blanks emergence: a Speed in space and time within a general pattern (GP) at an on-ramp bottleneck in the left (*left*) and right lanes (*right*). b–e Speed (b, d) and density (c, e) at different time moments associated with the GP in a in the left and right lanes. In c, e, for density calculation the density definition (vehicles per freeway length) is used. Taken from [47]

in the right lane increase due to lane changing. Vehicles begin to cover these blanks ($t \approx 24$ min in Fig. 34a). As a result, low speed states within the jam appear associated with these moving blanks. Later, subsequent lane changing of vehicles from the right to the left lane results in an abrupt decrease in speed in the left lane upstream of the narrow moving jam ($t \approx 25$ min in Figs. 33b,c and 34a). Then this complex spatiotemporal speed distribution in the left lane transforms into a wide moving jam with moving blanks within the jam ($t \approx 25$ min in Fig. 33b,c). Thus

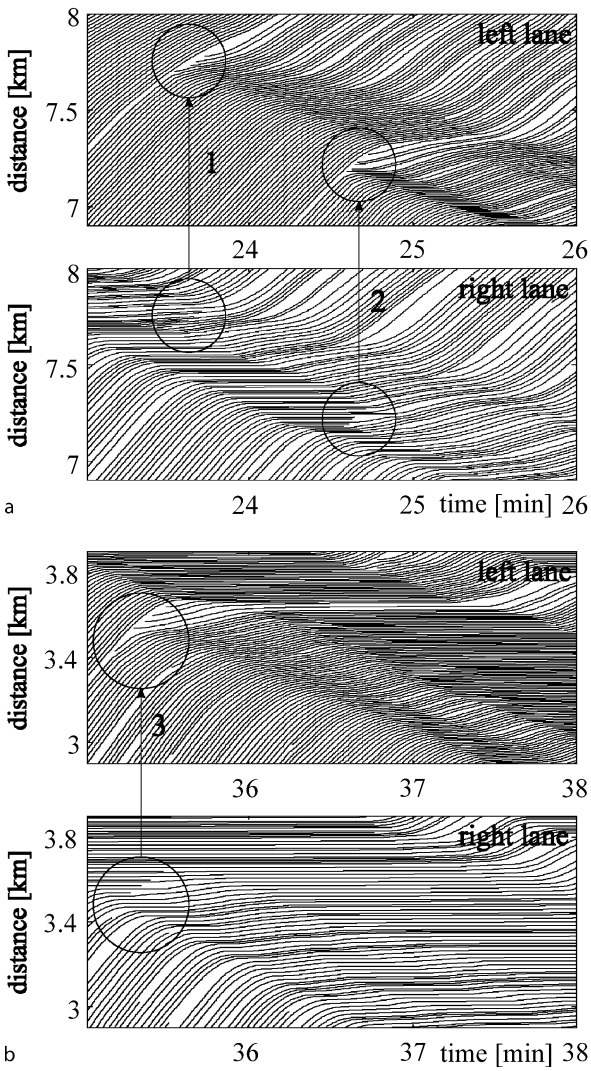

Traffic Congestion, Spatiotemporal Features of, Figure 34
Simulated characteristics of emergence of moving blanks in the GP shown in Fig. 33a: a, b Vehicle trajectories for different time intervals in the left and right lanes. Taken from [47]

in this scenario, vehicle lane changing from the right to the left lane at the upstream jam front is the main reason for emergence of moving blanks within the jam (moving blanks labeled "blanks 1" in Fig. 33b).

During subsequent jam propagation new moving blanks emerge (Fig. 33d,e). Firstly, the upstream jam front in the right lane is upstream of the upstream jam front in the left lane ($t = 35$ min, Fig. 33d,e). Due to lane changing of vehicles from the right to the left lane, the upstream jam front locations are synchronized with each other (arrow 3 in Fig. 34b). This lane changing causes emergence of moving blanks within the jam ($t = 37$ min, moving blanks labeled "blanks 2" in Fig. 33d).

Diverse Transitions Between Traffic Phases and States

Microscopic effects of moving jam emergence, propagation, and dissolution lead to diverse *Transitions* between traffic phases and traffic states. In a microscopic analysis of these transitions, we should distinguish the following three basic types of qualitatively different transitions measured at fixed freeway locations.

(i) Local phase transitions (denoted by F → S, S → F, S → J, and J → S transitions). Within a local freeway region a new traffic phase emerges instead of an initial traffic phase.
(ii) Transitions between existing traffic phases that are caused by propagation of a front separating free and synchronized flows or by wide moving jam propagation.
(iii) Transitions between different traffic states of the same traffic phase.

We consider the associated microscopic effects on an example of moving jam emergence within the DGP shown in Fig. 35a.

Firstly, an F → S transition occurs in a neighborhood of the bottleneck (location $x = 10.4$ km; transition of type (i) shown by arrows labeled "F → S" in Fig. 35b,c). As a result, synchronized flow appears at the bottleneck. Secondly, this synchronized flow propagates on the main road upstream of the bottleneck, i. e., a wave of upstream transitions from free flow to synchronized flow occurs (e. g., $x = 9.1$, 7.4, and 5.5 km; transitions of type (ii) shown by arrows 1–3 in Fig. 35d–i). Then a self-compression of this synchronized flow (pinch effect) within a localized region upstream of the bottleneck occurs. The pinch effect causes an abrupt transition between different states within the synchronized flow phase ($x = 9.1$ km; transition of type (iii) shown by arrows labeled "P" in Fig. 35d,e).

Within the related pinch region of synchronized flow, a narrow moving jam emerges spontaneously. During upstream jam propagation, the jam amplitude grows and the narrow moving jam transforms into a wide moving jam, i. e., an S → J transition occurs in the synchronized flow (location $x = 7.4$ km; transition of type (i) shown by arrows labeled "S → J" in Fig. 35f,g). The flow rate in the jam outflow q_{out} is smaller than the flow rate in the pinch region. As a result, the pinch region dissolves after the wide moving jam has been formed; synchronized flow of higher speed remains ($x = 9.1$ km; transition of type (iii) shown by arrows labeled "S" in Fig. 35d,e). For this reason,

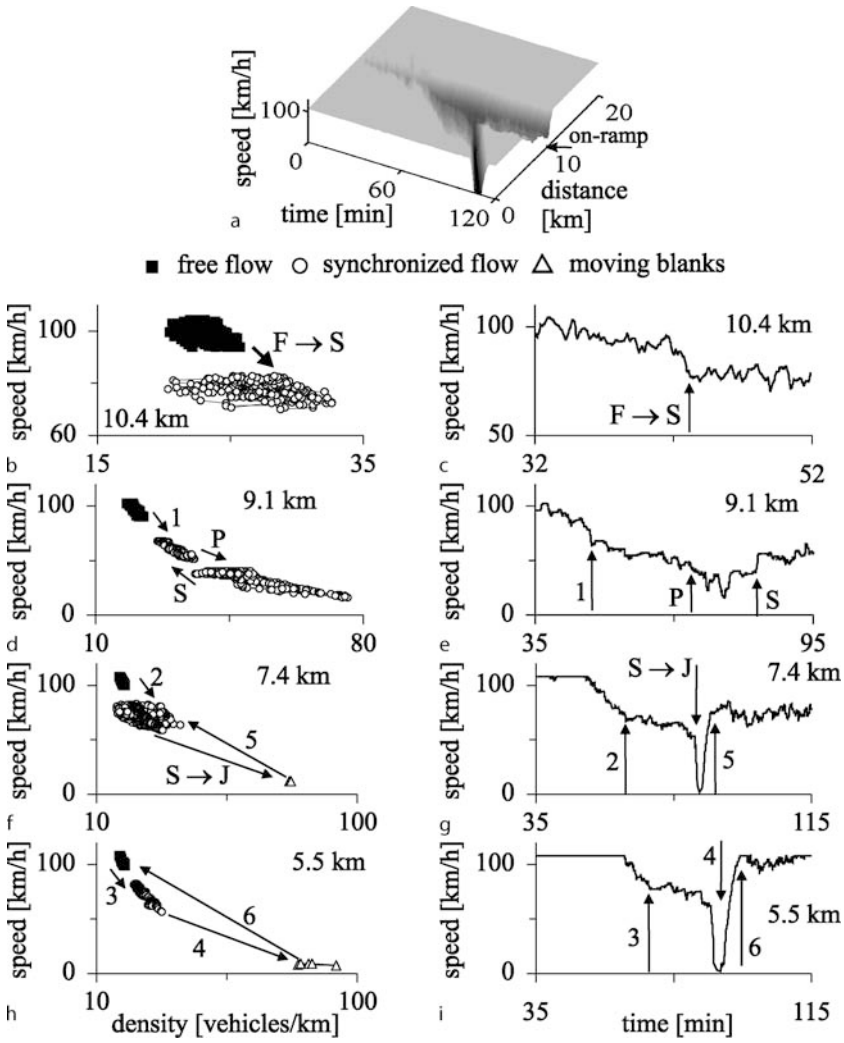

Traffic Congestion, Spatiotemporal Features of, Figure 35

Simulated characteristics of microscopic transitions between different traffic phases and states measured at fixed freeway locations as well as the associated hysteresis phenomena in traffic flow: a Speed in space and time within DGPs at on-ramp bottleneck. b–i Transitions in the speed-density plane (b, d, f, h) and in the related time-dependences of speed (c, g, e, i) at different road locations during wide moving jam emergence and propagation shown in a. Taken from [47]

in contrast with the GP in Fig. 33a the wide moving jam in Fig. 35a prevents subsequent moving jam emergence in synchronized flow upstream of the bottleneck, i. e., the DGP appears.

The resulting DGP consists of synchronized flow upstream of the wide moving jam, the jam propagating upstream, and synchronized flow that is downstream of the jam and upstream of the bottleneck (Fig. 35a). At each of the virtual detectors, jam propagation causes different transitions of type (ii):

1) From an initial traffic phase at a detector (free flow or synchronized flow) to low speed states associated with moving blanks within the wide moving jam (e. g., $x = 5.5$ km, transition from synchronized flow to the jam shown by arrows "4" in Fig. 35h,i.
2) From the jam to either free flow or synchronized flow after the jam has passed the detector ($x = 7.4$ km, transition from the jam to synchronized flow shown by arrows "5" in Fig. 35f,g and $x = 5.5$ km, transition from the jam to free flow shown by arrows "6" in Fig. 35h,i).

In Fig. 35, the flow rates to the on-ramp q_{on} and the flow rate in free flow upstream of the DGP q_{in} do not vary

over time. In contrast, if the flow rate q_{in} begins to decrease, then the DGP dissolves (Fig. 36). In this case, DGP dissolution starts upstream and propagates downstream. Firstly, a wave of synchronized flow dissolution, which is upstream of the jam, propagates downstream leading to transitions of type (ii) from synchronized flow to free flow (Fig. 36a). When this wave reaches the jam, the latter begins to dissolve. As a result of wide moving jam dissolution synchronized flow remains, i. e., an J → S transition occurs ($x = 7$ km; arrows labeled "J → S" in Fig. 36b). Then a new wave of synchronized flow dissolution propagating downstream is formed, which leads to transitions of type (ii) from synchronized flow to free flow ($x = 7$ km; arrow 7 in Fig. 36b). Due to the latter transitions, finally, an S → F transition (transition of type (i)) occurs at the bottleneck ($x = 10.4$ km; arrows labeled "S → F" in Fig. 36c).

These numerical simulations enable us to find states of free flow and synchronized flow as well as states within a wide moving jam on a double Z-characteristic for phase transitions (Fig. 37). Arrows between these states labeled "F → S" and "S → F" are associated with F → S and S → F transitions at the bottleneck ($x = 10.4$ km), whereas arrows labeled "S → J" and "J → S" are associated with S → J and J → S transitions, which have occurred at different locations, $x = 7.4$ km and 7 km, respectively (Figs. 35g and 36b). Thus, wide moving jam emergence (S → J transition) and dissolution (J → S transition) occur upstream of the bottleneck, whereas the F → S and S → F transitions occur at the bottleneck. As a result, states for the three traffic phases on the double Z-characteristic cannot be found through measurements of traffic variables at a single freeway location. Consequently, traffic

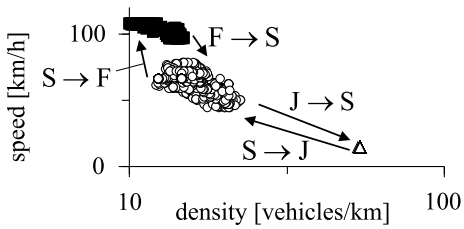

Traffic Congestion, Spatiotemporal Features of, Figure 37
States of the three traffic phases associated with double Z-characteristic in the speed-density plane for local phase transitions associated with Figs. 35 and 36

states and transitions shown by arrows in Fig. 37 are taken from different locations (i. e., from Fig. 35) at which the related local phase transitions have occurred.

We can conclude that during synchronized flow and wide moving jam emergence, propagation, and dissolution there are all three types of abrupt transitions (i), (ii), and (iii) between traffic variables. These transitions can cause a variety of "hysteresis" phenomena in the speed-density plane (Figs. 35, 36, and 37).

Congested Patterns at Heavy Bottlenecks

In general, the average flow rate $q^{(cong)}$ within a congested traffic pattern upstream of a bottleneck is the smaller, the greater the bottleneck influence of traffic (called the bottleneck strength). As in Sect. "Diagram of Congested Patterns at Isolated Bottlenecks", the time interval of the averaging of the flow rate $q^{(cong)}$ is suggested to be considerably longer than time distances between any moving jams

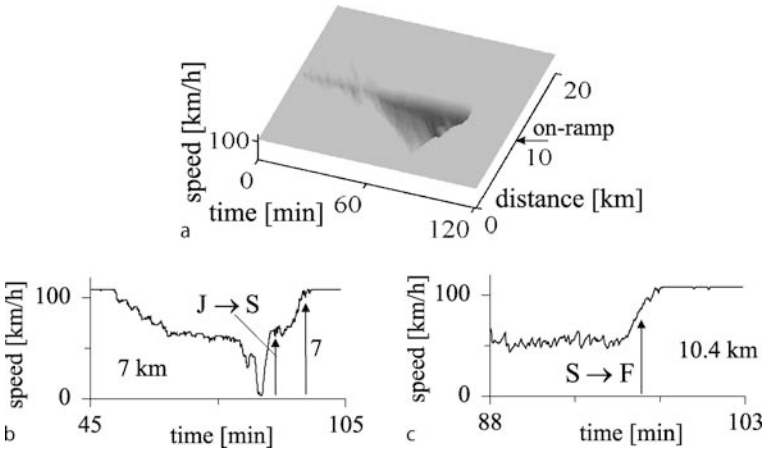

Traffic Congestion, Spatiotemporal Features of, Figure 36
Speed in time and space during a dissolution of the DGP in Fig. 35a occurring when the flow rate in free flow upstream of the DGP decreases beginning at $t = 75$ min (a) and the associated transitions in time-dependences of speed (b, c) at different road locations

within a congested pattern. Empirical data show that the flow rate $q^{(\mathrm{cong})}$ within GPs occurring at usual bottlenecks like on- and off-ramp bottlenecks discussed above in this article, which is equal to the average flow rate in the pinch region $q^{(\mathrm{pinch})}$, is approximately within a range

$$q^{(\mathrm{cong})} = q^{(\mathrm{pinch})} = 1100\text{–}1700\,\text{vehicles/h/lane}\,. \tag{7}$$

In contrast with the usual bottlenecks, due to for example bad weather conditions or accidents heavy bottlenecks can occur, which exhibit a much greater influence on traffic (greater bottleneck strength) that limits $q^{(\mathrm{cong})}$ to very small values, sometimes as low as zero. Results of a theory of traffic congestion at heavy bottlenecks developed recently [39,40] are discussed in this Sect. "Congested Patterns at Heavy Bottlenecks".

In this case, we can expect new interesting physical phenomena associated with complexity of traffic congestion at heavy bottlenecks. This expectation follows already from a qualitative consideration of the influence of the decrease in $q^{(\mathrm{cong})}$ on a sequence of wide moving jams. Indeed, from empirical single vehicle data studied in [46,47], we can conclude that the flow rate $q^{(\mathrm{blanks})}$ of low speed states associated with moving blanks within the jams satisfies an approximate condition

$$q^{(\mathrm{blanks})} \lesssim 600\,\text{vehicles/h/lane}\,. \tag{8}$$

The average flow rate within wide moving jams $q^{(\mathrm{blanks})}$ is defined as a number of vehicles $N^{(\mathrm{blanks})}$, which have passed a virtual road detector during time intervals of the propagation of n_{J} wide moving jams (it is assumed that $n_{\mathrm{J}} \gg 1$) through the detector, divided by the sum of all jam durations: $q^{(\mathrm{blanks})} = N^{(\mathrm{blanks})}/\sum_{i=1}^{n_{\mathrm{J}}} \tau_{\mathrm{J}}^{(i)}$, where $\tau_{\mathrm{J}}^{(i)}$ is a duration of a ith wide moving jam, i. e., the time interval between the downstream and upstream fronts of the jam i, while this jam propagates through the detector.

Between the wide moving jams, the average flow rate associated with non-interrupted flows in the jam outflows $q_{\mathrm{out}}^{(\mathrm{J})}$ is greater than $q^{(\mathrm{pinch})}$ (7), i. e., $q_{\mathrm{out}}^{(\mathrm{J})}$ is considerably greater than $q^{(\mathrm{blanks})}$ (8). Now we assume that due to bad weather conditions or an accident a heavy bottleneck occurs with a great strength for which

$$q^{(\mathrm{cong})} = q^{(\mathrm{blanks})}. \tag{9}$$

In this case, $q_{\mathrm{out}}^{(\mathrm{J})}$ must reduce to $q^{(\mathrm{blanks})}$ (8), i. e., the difference between flows within and outside wide moving jams disappears. As a result, at

$$q^{(\mathrm{cong})} \leq q^{(\mathrm{blanks})} \tag{10}$$

wide moving jams should merge into a mega-wide moving jam (mega-jam for short).

For an analysis of traffic congestion at heavy bottlenecks, the Kerner–Klenov stochastic three-phase traffic flow model of vehicular traffic on a two-lane road with identical drivers and vehicles [41,42] considered in ▶ Traffic Congestion, Modeling Approaches to has been used with the following model for a heavy bottleneck.

We suggest that there is a section of the road of a length L_{B} within which due to an accident or bad weather conditions drivers should increase a safety time headway $\tau^{(\mathrm{safe})}$ to the preceding vehicle as well as decrease the maximum speed to some v_{B} that is lower than the maximum speed in free flow: within the section, the safety time headway is equal to $\tau^{(\mathrm{safe})} = T_{\mathrm{B}} > 1\,\text{sec}$ that is longer than $\tau^{(\mathrm{safe})} = 1\,\text{sec}$ used in the model for vehicles moving outside this section. In according with the model, $\tau^{(\mathrm{safe})}$ determines a safe speed, which should not be exceeded by a driver; otherwise, the driver decelerates. As a result, within this section drivers move with the mean time headway that is longer than the mean time headway outside the bottleneck section and it is the longer, the longer $\tau^{(\mathrm{safe})} = T_{\mathrm{B}}$. Therefore the section with longer $\tau^{(\mathrm{safe})} = T_{\mathrm{B}}$ acts as a bottleneck on the road. In the bottleneck model, each chosen value T_{B} defines a specific bottleneck: the strength of this bottleneck is the greater, the longer T_{B}; in turn, the longer T_{B}, the smaller $q^{(\mathrm{cong})}$, i. e., the greater the flow rate limitation within congestion caused by the bottleneck.

This model feature allows us to simulate a heavy bottleneck caused by bad weather conditions or accidents leading to a long enough T_{B} within the bottleneck. Under bad weather conditions, a road section with a long value T_{B} can be caused, e. g., by a poor view due to fog on the section or a much longer deceleration way needed by snow and ice on the section. If an accident occurs on a road, a road section with a long value T_{B} can be caused, e. g., by much narrower lane widths allowed for driving on the road section; the same effect can occur under heavy roadworks.

In simulations discussed in below, we consider traffic phenomena occurring when the bottleneck strength increases gradually through the increase in T_{B}, when other bottleneck parameters are given constants (Fig. 38). However as simulations show, at a given T_{B} the increase in L_{B} or/and decrease in v_{B} lead also to an increase in the bottleneck strength with the subsequent decrease in the average flow rate $q^{(\mathrm{cong})}$ in congested traffic upstream of the bottleneck. For this reason, numerical values of T_{B} mentioned below, at which characteristic traffic phenomena occur at a heavy bottleneck, depend on L_{B} and v_{B}. In contrast, we find that numerical values of the flow rate $q^{(\mathrm{cong})}$ at which the traffic phenomena occur at the bottleneck do not change appreciably, when a greater value of L_{B}

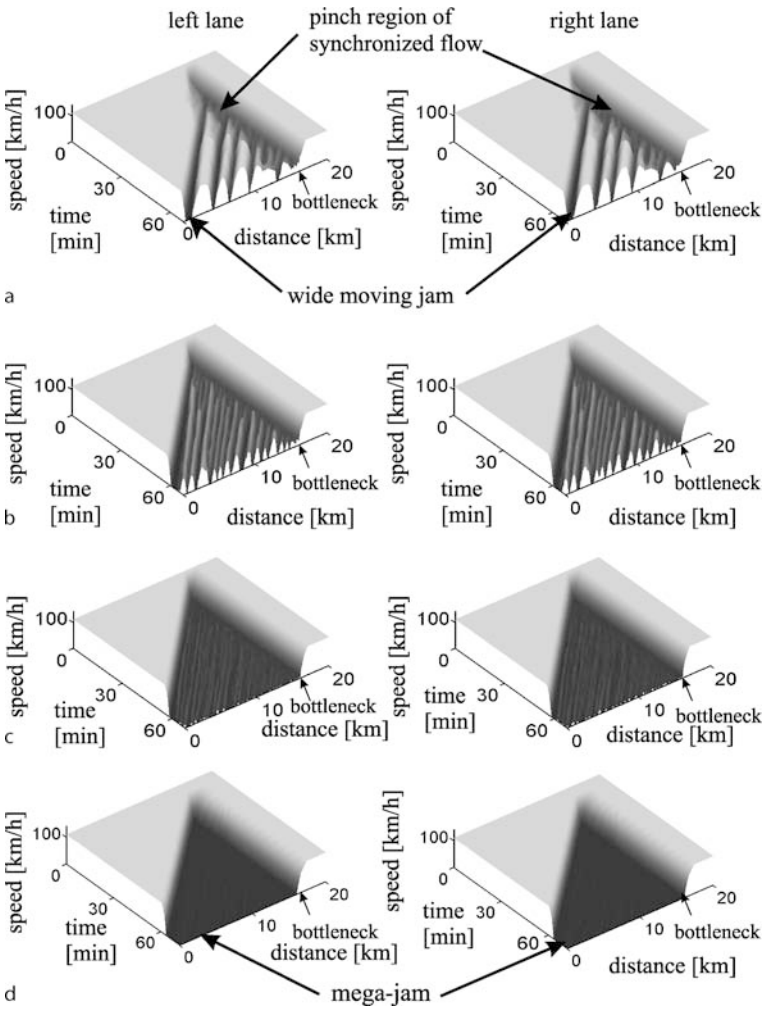

Traffic Congestion, Spatiotemporal Features of, Figure 38

Simulated speed in space and time in the left (*left*) and right (*right*) road lanes at different T_B: $T_B = 1.8$ (a), 2.4 (b), 12 (c), 60 sec (d). $q_{in} = 1946$ vehicles/h/lane. The upstream boundary of bottleneck region of the length $L_B = 300$ m is at $x = 16$ km; the maximum speed within this bottleneck region is $v_B = 60$ km/h. Resulting values of $q^{(cong)} = 1546$ (a), 1114 (b), 440 (c), 127 vehicles/h/lane (d)

or/and another value of v_B (in the range 50–80 km/h) are chosen. For this reason, the flow rates $q^{(cong)}$ can be considered characteristic values representing the bottleneck strength.

Evolution of Traffic Phases at Heavy Bottlenecks

When the bottleneck strength is not great, i. e., T_B is chosen to be not very long ($T_B = 1.8$ sec), then at a great enough flow rate q_{in} in free flow upstream of the bottleneck, firstly synchronized flow occurs spontaneously at the bottleneck. Then the pinch region with a relatively great flow rate $q^{(cong)} = q^{(pinch)}$ is formed (Figs. 38a and 39). At the pinch region upstream boundary wide moving jams emerge. Thus we found the phenomena of regular GP formation (Figs. 38a and 40a,b) discussed briefly in Subsect. "General Congested Patterns". The pinch region width $L^{(pinch)}(t)$ changes over time between about 1 and 2 km (Fig. 41a). $L^{(pinch)}$ is defined as the distance between the upstream boundary of the bottleneck ($x = 16$ km) and the road location upstream at which a wide moving jam has just been identified through the use of the jam microscopic criterion (5). There is also a nearly constant frequency of $L^{(pinch)}(t)$ oscillations associated with the maximum in the Fourier spectrum (Fig. 41b). Speed autocorrelation functions and associated Fourier spectra of speed time-dependences at shorter T_B show regular character of wide moving jam propagation (Fig. 40c,d).

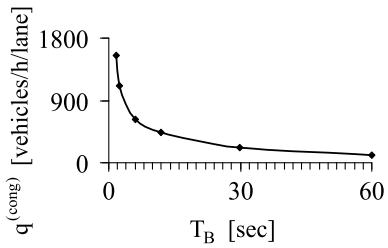

Traffic Congestion, Spatiotemporal Features of, Figure 39
Simulated average flow rate $q^{(cong)}$ within congested patterns shown in Fig. 38 as a function of T_B. The averaging time interval for $q^{(cong)}$ is 60 min

When T_B becomes longer and therefore the bottleneck strength increases, the flow rate $q^{(cong)} = q^{(pinch)}$ decreases (Fig. 39). However, if T_B remains a relatively small value ($1.8 < T_B < 3$ sec), then as for other bottleneck types, we found the following GP features some of them have already been discussed in Sect. "Diagram of Congested Patterns at Isolated Bottlenecks": the smaller the flow rate $q^{(pinch)}$, the greater the frequency of narrow moving jam emergence within the pinch region, the lower the maximum (and the average) speed between wide moving jams upstream of the pinch region, the smaller the mean pinch region width $L^{(pinch)}_{mean}$ (Figs. 38b and 41c,d,k,l). This can also be seen from a comparison of time-dependences of average speed within the region of wide moving jams for different T_B (Fig. 40a,e).

Qualitatively other phenomena are found when T_B further increases ($T_B > 3$ sec) and the average flow rate $q^{(cong)}$ decreases considerably (Figs. 38c,d and 39).

We find that there is a critical strength of a heavy bottleneck associated with $T_B = T_B^{(break)}$ that results in the critical flow rate $q^{(cong)} = q^{(cong)}_{break}$. When

$$q^{(cong)} \le q^{(cong)}_{break} \tag{11}$$

related to $T_B \ge T_B^{(break)}$, then there are random time intervals when the pinch region disappears, i. e., $L^{(pinch)} = 0$ (Fig. 41e,g). This means that there are time instants at which there is no pinch region and wide moving jams emerge directly at the upstream boundary of the bottleneck, whereas for other time intervals the pinch region appears again (Fig. 41e,g). We found that $q^{(cong)}_{break} \approx 920$ vehicles/h/lane ($T_B^{(break)} \approx 3$ sec at the chosen bottleneck parameters L_B and v_B). Under condition (11), $L^{(pinch)}(t)$ becomes a non-regular time-function (Fig. 41e,g,i) whose Fourier spectrum is broader, the longer T_B, i. e., the smaller $q^{(cong)}$ (Fig. 41f,h,j). Such a GP at an isolated heavy bottleneck can be considered the GP with a non-regular pinch region.

The more the bottleneck strength exceeds the critical bottleneck strength, the greater the difference $q^{(cong)}_{break} - q^{(cong)}$ and the longer the mean duration of time intervals within which the regular structure of the GP breaks and the pinch region disappears, and therefore, the smaller the mean length $L^{(pinch)}_{mean}$ of the pinch region (Fig. 41k,l). In contrast with regular time-dependences of the average speed within a sequence of wide moving jams (Fig. 40a–h), the time-functions of 1 min average speed at a greater bottleneck strength exhibit a non-regular behavior (Fig. 40i–l). This can be seen from speed autocorrelation functions and associated Fourier spectra of the average speed time-dependences (Fig. 40k,l).

When the bottleneck strength increases further, $L^{(pinch)}_{mean}$ decreases continuously (Fig. 41l). $L^{(pinch)}_{mean}$ reaches zero at a threshold bottleneck strength for the pinch region existence associated with $T_B = T_B^{(th)}$ that results in a threshold flow rate $q^{(cong)} = q^{(cong)}_{th}$. When

$$q^{(cong)} \le q^{(cong)}_{th} \tag{12}$$

related to $T_B \ge T_B^{(th)}$, then the pinch region of GPs does not exist. At model parameters, the pinch region of a GP disappears fully at $q^{(cong)}_{th} \approx 220$ vehicles/h/lane ($T_B^{(th)} \approx 30$ sec at the chosen L_B and v_B).

There is also another critical bottleneck strength, which we call the critical bottleneck strength for the mega-jam formation associated with $T_B = T_B^{(mega)}$ that results in the critical flow rate $q^{(cong)} = q^{(cong)}_{mega}$. When

$$q^{(cong)} \le q^{(cong)}_{mega} \tag{13}$$

related to $T_B \ge T_B^{(mega)}$, then wide moving jams *merge* onto a mega-jam. Numerical simulations show that (13) is equivalent to (10), i. e.,

$$q^{(cong)}_{mega} = q^{(blanks)} . \tag{14}$$

At model parameters, $q^{(cong)}_{mega} \approx 130$ vehicles/h/lane ($T_B^{(mega)} \approx 60$ sec at the chosen L_B and v_B).

We found also that the critical bottleneck strength for the mega-jam formation is greater than the threshold bottleneck strength for the pinch region existence that results in the condition

$$q^{(cong)}_{mega} < q^{(cong)}_{th} \tag{15}$$

related to $T_B^{(mega)} > T_B^{(th)}$. Under condition (13), i. e., when all wide moving jams merge onto a mega-jam, traffic congestion upstream of an isolated heavy bottleneck cannot

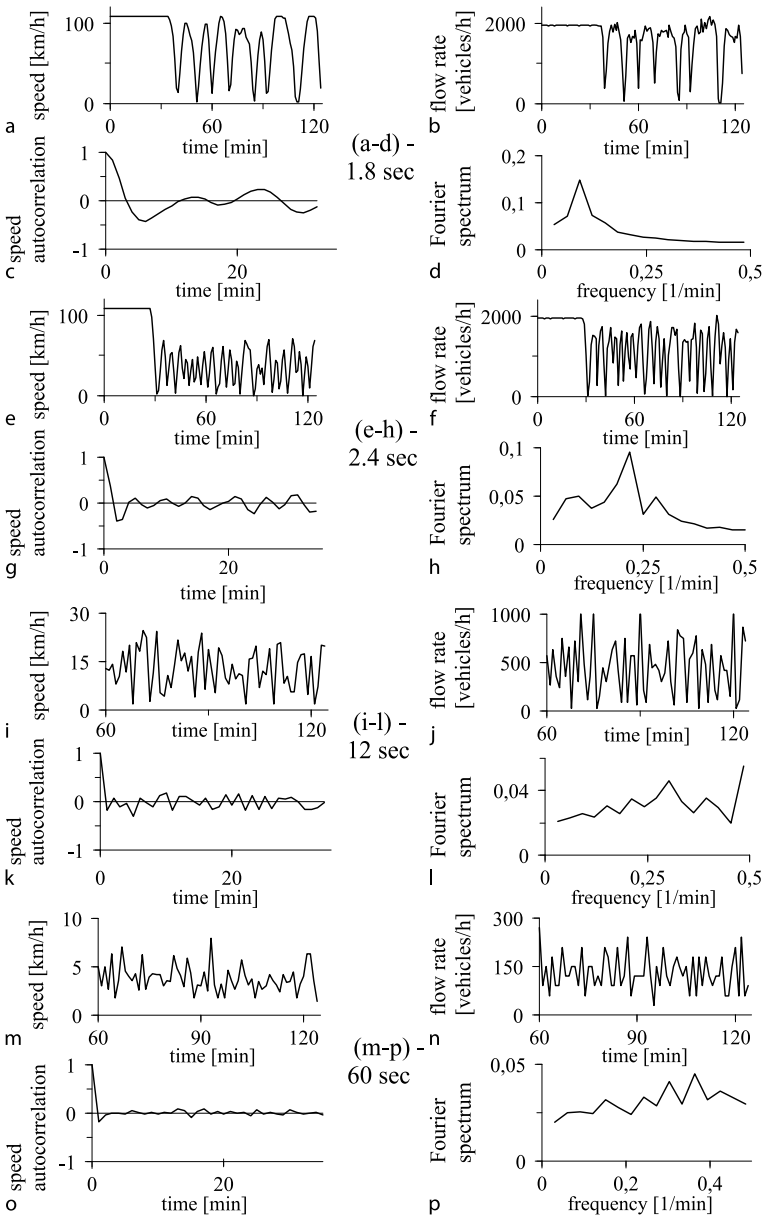

Traffic Congestion, Spatiotemporal Features of, Figure 40

Simulated characteristics of congested patterns shown in Fig. 38 related to location 10 km. Time-functions of speed (a, e, i, m), speed correlations (c, g, k, o), associated Fourier spectra (d, h, l, p), and flow rate (b, f, j, n) for different $T_B = 1.8$ (a–d), 2.4 (e–h), 12 (i–l), 60 sec (m–p). 1 min average data in the left lane

be considered as a GP any more. Thus in accordance with (13), (14), if traffic breakdown has occurred at a bottleneck, then the condition

$$q^{(\mathrm{cong})} > q^{(\mathrm{blanks})} \tag{16}$$

is a *necessary* condition for GP existence at this bottleneck.

As for GPs with a very non-regular pinch region (Fig. 40i–l), the time-functions of 1 min average speed within the mega-jam exhibit a non-regular behavior (Fig. 40m); this can be seen from speed autocorrelation functions and associated Fourier spectra of the average speed time-dependences (Fig. 40o,p).

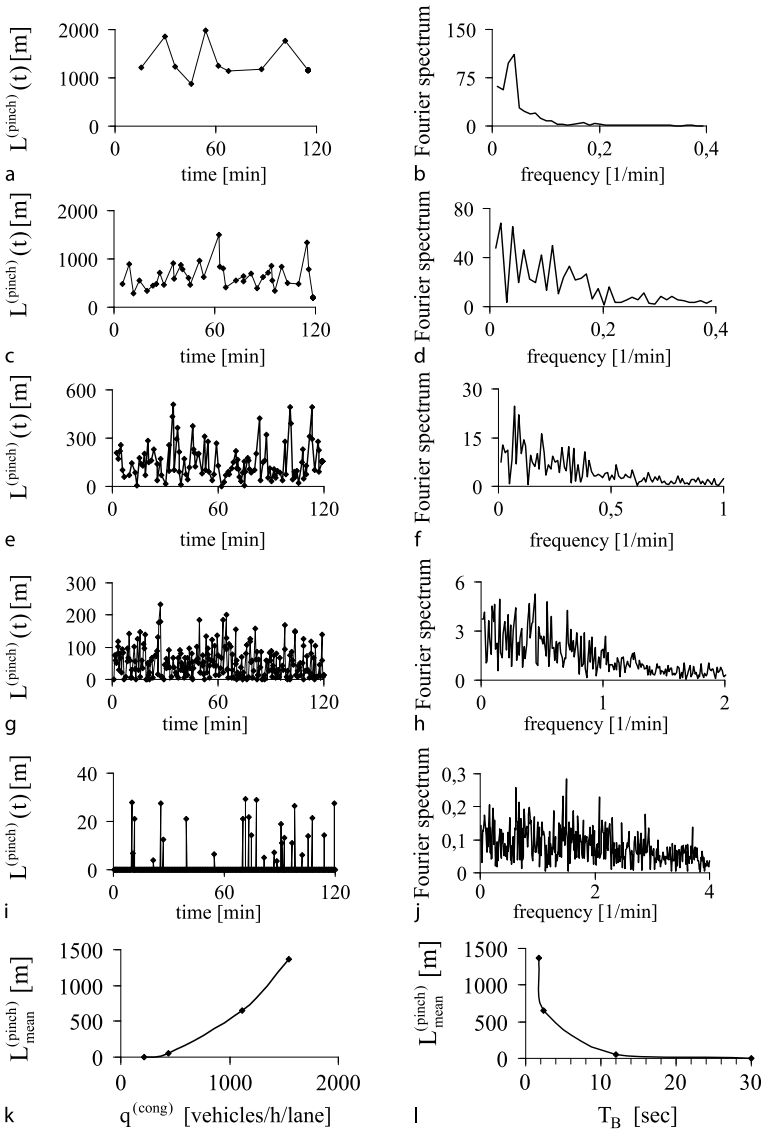

Traffic Congestion, Spatiotemporal Features of, Figure 41

Simulations of pinch effect: $L^{(\mathrm{pinch})}(t)$ (a, c, e, g, i), associated Fourier spectra (b, d, f, h, j) for congested patterns related to $T_B = 1.8$ (a, b), 2.4 (c, d), 6 (e, f), 12 (g, h), and 30 sec (i, j). k, l $L^{(\mathrm{pinch})}_{\mathrm{mean}}(q^{(\mathrm{cong})})$ (k) and $L^{(\mathrm{pinch})}_{\mathrm{mean}}(T_B)$ (l)

When $q^{(\mathrm{cong})}$ becomes zero, because behind a road location the road is closed, the mega-jam transforms into a queue of *motionless* vehicles, which therefore is not associated with vehicular traffic. Nevertheless, there is a link between the queue and the mega-jam. If at a time instant we allow several vehicles to escape from this queue, then simulations show that motion of these vehicles results in wide moving jam occurrence: the downstream front of the jam separates moving vehicles escaping from the queue and vehicles standing within the queue upstream of the front. When the number of vehicles escaping from the initial queue decreases, the downstream jam front transforms into a moving blank(s) subsequently covered by vehicles standing in the queue. When a vehicle per a long enough time interval is allowed to escape from the mega-jam, as simulations show, a sequence of such moving blanks within this jam occur; these moving blanks exhibit, however, the dynamics specifically associated with the mega-jam (Subsect. "Microscopic Spatiotemporal Structure of Mega-Jam").

Thus based on a study of three-phase traffic flow model we found that the complexity of traffic congestion at a heavy bottleneck caused for example by bad weather conditions or accidents is associated with the phenomenon of random disappearance and appearance of the pinch region over time as well as with the phenomenon of the merger of wide moving jams into a mega-jam. When the bottleneck strength increases continuously, the mean length of the pinch region decreases also continuously up to zero; at such a heavy bottleneck, the pinch region does not exist any more. Beginning from a greater bottleneck strength only the mega-jam survives in congested traffic upstream of the bottleneck and synchronized flow remains only within its downstream front separating free flow and congested traffic. In other words, in congested traffic occurring at such a very heavy bottleneck there are no continuous flows within the congested pattern any more with the one exception of the downstream front of synchronized flow that separates free flow downstream and the mega-jam upstream of the front. Synchronized flow remains *only* within this front. These phenomena reveal the evolution of the traffic phases in congested traffic when heavy bottlenecks occur in highway traffic.

Microscopic Spatiotemporal Features of Non-regular Moving Jam Dynamics

For GPs with a regular pinch region ((11) is not satisfied), we find results of Subsect. "General Congested Patterns" that at a small enough distance upstream of the bottleneck there is the pinch region of non-interrupted synchronized flow (Fig. 42a,b) with short enough time headways for which criterion (5) is satisfied for *none* of time headways (Fig. 42c,d) [38].

In contrast, under condition (11), when GPs with a non-regular pinch region occur, at the same small distance upstream of the bottleneck some long time headways appear in the pinch region of the GP for which criterion (5) is satisfied (Fig. 42g,h). This means that there are wide moving jams that emerge almost directly upstream of the bottleneck. This is related to the previous result that the pinch region width $L^{(\mathrm{pinch})}$ of such a GP is nearly as small as zero at some random time instants (Fig. 41e,g). During other time intervals, there is synchronized flow with short enough time headways for which criterion (5) is not satisfied (Fig. 42g,h). Thus at small enough distance upstream of the bottleneck, such GPs consists of a random alternation between synchronized flow and wide moving jams. This is related to the conclusion of Subsect. "Evolution of Traffic Phases at Heavy Bottlenecks" that for these GPs at some random time instants the pinch region disappears and appears again.

Single-vehicle data measured by a virtual road detector 6 km upstream of the bottleneck within the GPs with regular (Fig. 43a–d) and non-regular pinch regions (Fig. 43e–h) exhibits qualitatively the same and typical time-dependences of the speed and time headway for a sequence of wide moving jams of a GP [47]. We see that there are moving jams within which the maximum time headway $\tau_{\max}$ between two vehicles within the jams is very long, i. e., traffic flow is interrupted and criterion (5) is satisfied (Fig. 43c,d,g,h). Therefore, these jams are wide moving ones. There are time intervals between these jams within which the speed is high (Fig. 43a,b,e,f) and vehicle time headways are small (Fig. 43c,d,g,h). These regions between the wide moving jams are related to non-interrupted traffic flow. Thus at some distance upstream of the bottleneck there is a sequence of wide moving jams that is characteristic for GPs.

A wide moving jam consists of alternations of flow interruption intervals and moving blanks within the jam (Sect. "Microscopic Features of Traffic Phases"). In a general case, for a wide moving jam we can observe two or more flow interruption intervals, i. e., two or more long time headways $\tau^{(i)}$, $i = 1, 2, \ldots$ each of them satisfies criterion (5):

$$I_{\mathrm{s}}^{(i)} = \tau^{(i)}/\tau_{\mathrm{del}}^{(\mathrm{a})} \gg 1, \quad i = 1, 2, \ldots . \tag{17}$$

Between these flow interruption intervals, we find one or more vehicles that have passed the detector with speeds, which are considerably lower than the average speed both upstream and downstream of the wide moving jam. This low speed vehicle motion is associated with moving blanks (Subsect. "Numerical Simulations of Moving Blanks Within Wide Moving Jams"): the flow interruption intervals $\tau^{(i)}$, $i = 1, 2, \ldots$ are separated by one or more moving blanks within the wide moving jam.

At low scales in time and space, both GPs with a regular (Fig. 44a,b) and non-regular pinch regions (Fig. 44c,d) exhibit *spatiotemporal microscopic structures* in which regions of lower speed alternate with regions of higher speed; in both cases, we find that these regions propagate upstream. At the first glance, we can see only that the mean frequency of low speed regions increases with the increase in T_{B}, i. e., when the bottleneck strength increases; this result has already been mentioned in Subsect. "Evolution of Traffic Phases at Heavy Bottlenecks". However, if we consider the microscopic structures of GPs with a non-regular pinch region in larger scales in space and time (Figs. 45 and 46), we can see that there is a crucial qualitative differ-

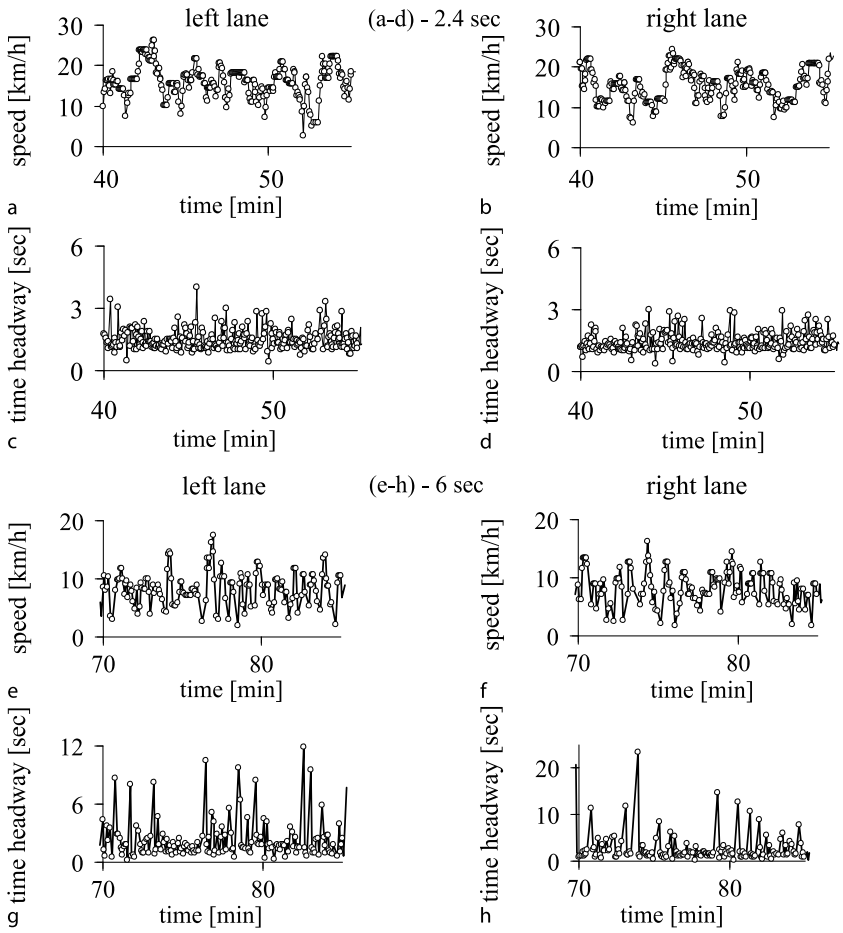

Traffic Congestion, Spatiotemporal Features of, Figure 42
Simulated single-vehicle data for speed (a, b, e, f) and time headway (c, d, g, h) measured by a virtual detector at location 15.8 km within congested traffic in the left figures (*left*) and right lanes (*right*) related to different $T_B = 2.4$ (a–d) and 6 sec (e–h). For e–h $q^{(cong)} = 626$ vehicles/h/lane

ence between them and the microscopic structures of GPs with a regular pinch region (Fig. 44a,b).

GPs with a regular pinch region exhibits a regular dynamics of wide moving jams discussed in Subsect. "General Congested Patterns", when the jams propagate upstream of the pinch region on a homogeneous road (Fig. 44a,b): some of the wide moving jams can dissolve during their upstream propagation (a so-called jam suppression effect) [38]. An example of the effect of the dissolution of wide moving jams can be seen in Fig. 44b (labeled "jam dissolution").

However, under condition (11), i. e., when a GP with a non-regular pinch region is realized, wide moving jams of the GP can exhibit complex and non-regular spatiotemporal dynamic behavior (Figs. 45 and 46). This non-regular jam dynamics is associated with the following effects:

(1) The effect of the *splitting* of a flow interruption interval onto two (or more) flow interruption intervals (Figs. 45a and 46c).
(2) The effect of the *emergence* of a new flow interruption interval (Figs. 45a and 46d).
(3) The effect of the *merging* of two (or more) traffic flow interruption intervals (Figs. 45a and 46a).
(4) The effect of the *dissolution* of a flow interruption interval (Figs. 45b,c and 46).

The dynamic effects (1)–(4) occur spontaneously during upstream propagation of wide moving jams. In different lanes, these effects occur often independently of each other and at different road locations. A spatiotemporal competition of these effects as well as a diverse variety of the *initiating* and *resulting* dynamic effects determine a non-regular

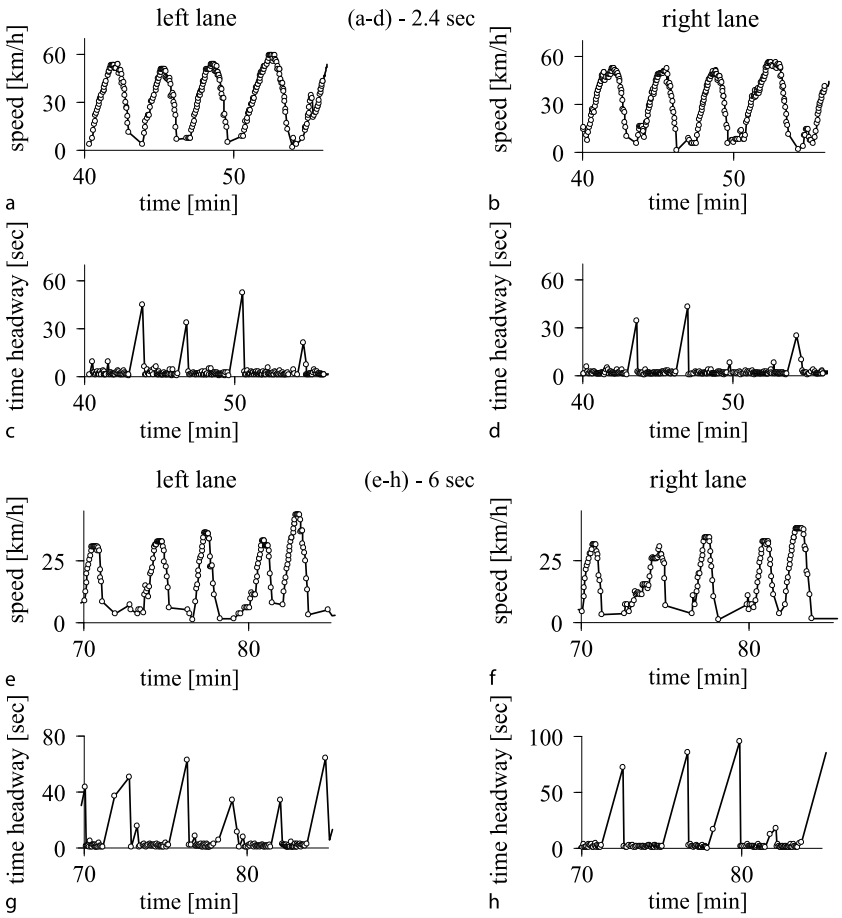

Traffic Congestion, Spatiotemporal Features of, Figure 43
Simulated single-vehicle data for speed (a, b, e, f) and time headway (c, d, g, h) measured by a virtual detector at location 10 km within congested traffic in the left lanes (*left*) and right lanes (right) related to different $T_B = 2.4$ (a–d) and 6 sec (e–h)

dynamic behavior of wide moving jams. The variety of the initiating and resulting effects associated with the effects (1)–(4) is as follows:

(i) The splitting of a flow interruption interval within a wide moving jam can result from the emergence of a new region of moving blanks within the jam.
(ii) The splitting of a flow interruption interval within an initial wide moving jam can result in the splitting of the jam onto two (or more) wide moving jams; in the latter case, synchronized flow(s) (or free flow(s)) occurs that separates the emergent wide moving jams. This is related to an J → S (or J → F) transition(s) occurring within the initial jam.
(iii) The effect of the emergence of a new flow interruption interval within a wide moving jam can result from the splitting of a region of moving blanks onto two (or more) regions of moving blanks separated by the emergent flow interruption interval.
(iv) The effect of the emergence of a new flow interruption interval occurring within synchronized flow between two wide moving jams can result in the emergence of a new wide moving jam. This is related to an S → J transition occurring in metastable synchronized flow between the jams.
(v) The effect of the merging of two (or more) traffic flow interruption intervals can result from the dissolution of two (or more) regions of moving blanks within a wide moving jam without wide moving jam dissolution.
(vi) The effect of the merging of two (or more) traffic flow interruption intervals can result in the merging of two (or more) wide moving jams.

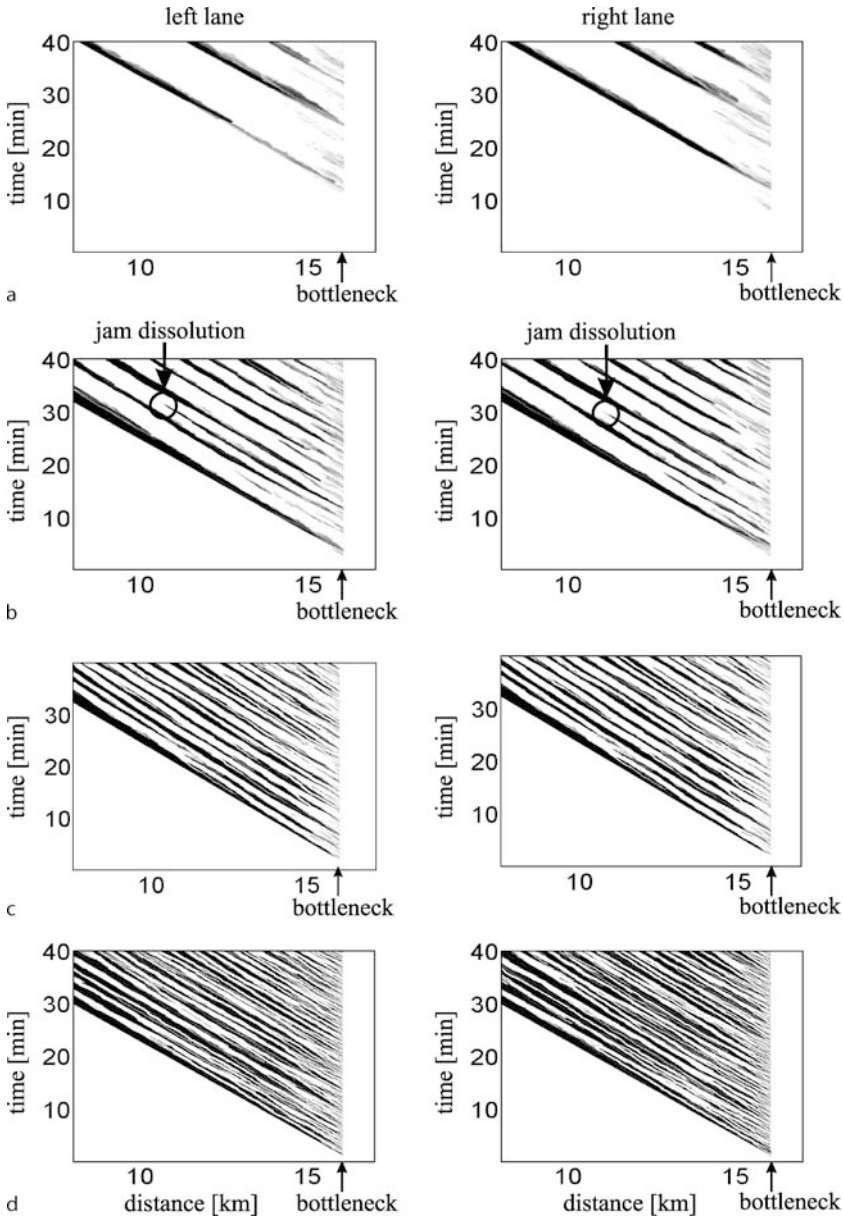

Traffic Congestion, Spatiotemporal Features of, Figure 44

Simulated spatiotemporal microscopic structures of GPs with regular (a, b) and non-regular pinch regions (c, d): Single-vehicle speeds within GPs presented in space and time by regions with variable darkness (the lower the speed, the darker the region) in the left (*left*) and right (*right*) road lanes associated with speed distributions at different $T_B = 1.8$ (a), 2.4 (b), 4 (c), 12 sec (d). For c $q^{(cong)} = 776$ vehicles/h/lane. The upstream boundary of the bottleneck region is at location $x = 16$ km (labeled "bottleneck")

(vii) The effect of the dissolution of a traffic flow interruption interval within a wide moving jam can result from the merging of two (or more) regions of moving blanks within the jam.

(viii) The effect of the dissolution of a traffic flow interruption interval can result in the dissolution of a wide moving jam.

The initiating and resulting effects (v)–(viii) decrease the mean frequency of regions of flow interruption intervals and regions of moving blanks within wide moving jams as well as the mean frequency of wide moving jams of a GP. In particular, the effects (vii) and (viii) occur very frequently close to the upstream boundary of the pinch region of GPs (see road locations between 16 and 15 km in

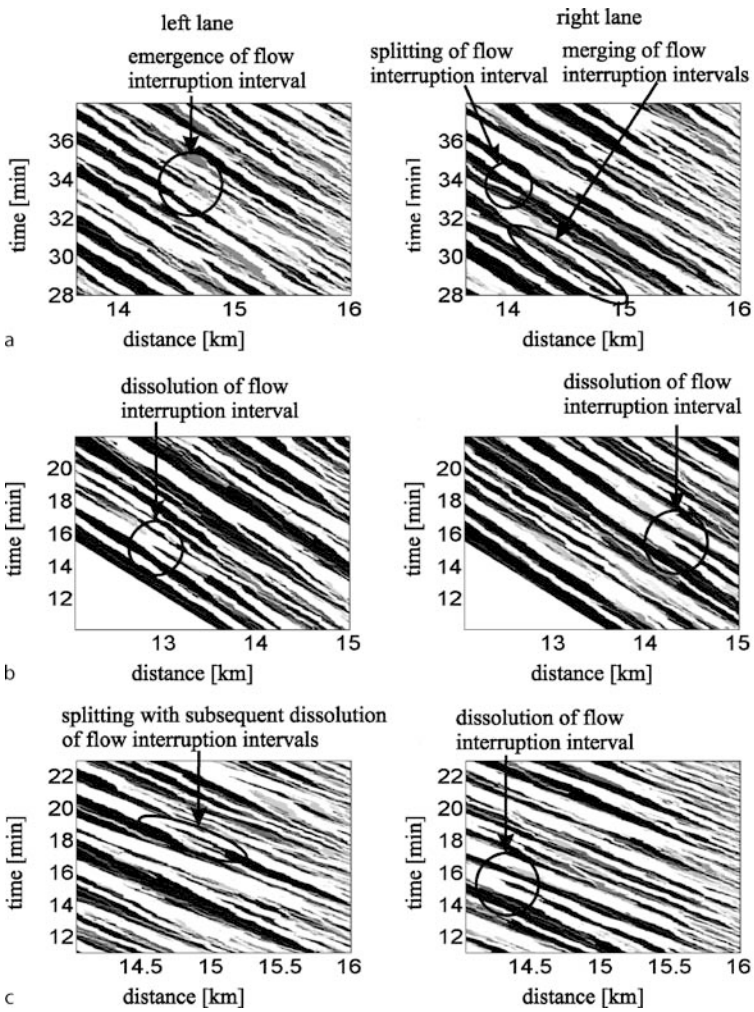

Traffic Congestion, Spatiotemporal Features of, Figure 45
Fragments of Fig. 44d in larger scales in time and space in the left lanes (*left*) and right lanes (*right*). *White regions* (single-vehicle speeds are equal to or higher than 5.4 km/h) are related to synchronized flows and moving blanks. *Black regions* (single-vehicle speeds are equal to zero) are related to flow interruption intervals. $T_B = 12$ sec

Figs. 45a,c and 46a,b). This leads to a decrease in the mean frequency of wide moving jams, when the jams propagate upstream.

There can be diverse sequences of the effects (1)–(4) over time with a random and arbitrarily sequence of the effects (i)–(viii). For example, the splitting of a flow interruption interval onto two intervals can be followed by the subsequent dissolution of these two intervals (Figs. 45c and 46b); the splitting of a flow interruption interval onto two intervals can be followed by the subsequent merging of these two intervals (Fig. 46c). The random and arbitrarily spatiotemporal sequence of the effects discussed can explain the non-regular spatiotemporal jam dynamics found in GPs with a non-regular pinch region. In turn, this non-regular spatiotemporal jam behavior can also explain the result of Subsect. "Evolution of Traffic Phases at Heavy Bottlenecks" that we cannot distinguish a sequence of wide moving jams in 1 min average data (Fig. 40i–l): the microscopic non-regular jam dynamics is averaged in this 1 min average data leading to non-regular and non-homogeneous speed distributions in which this real fine spatiotemporal jam dynamics *cannot be found*.

Microscopic Spatiotemporal Structure of Mega-Jam

The crucial difference between GPs and a mega-jam, which occurs under condition (13), is as follows: far enough upstream of the bottleneck, single-vehicle data of

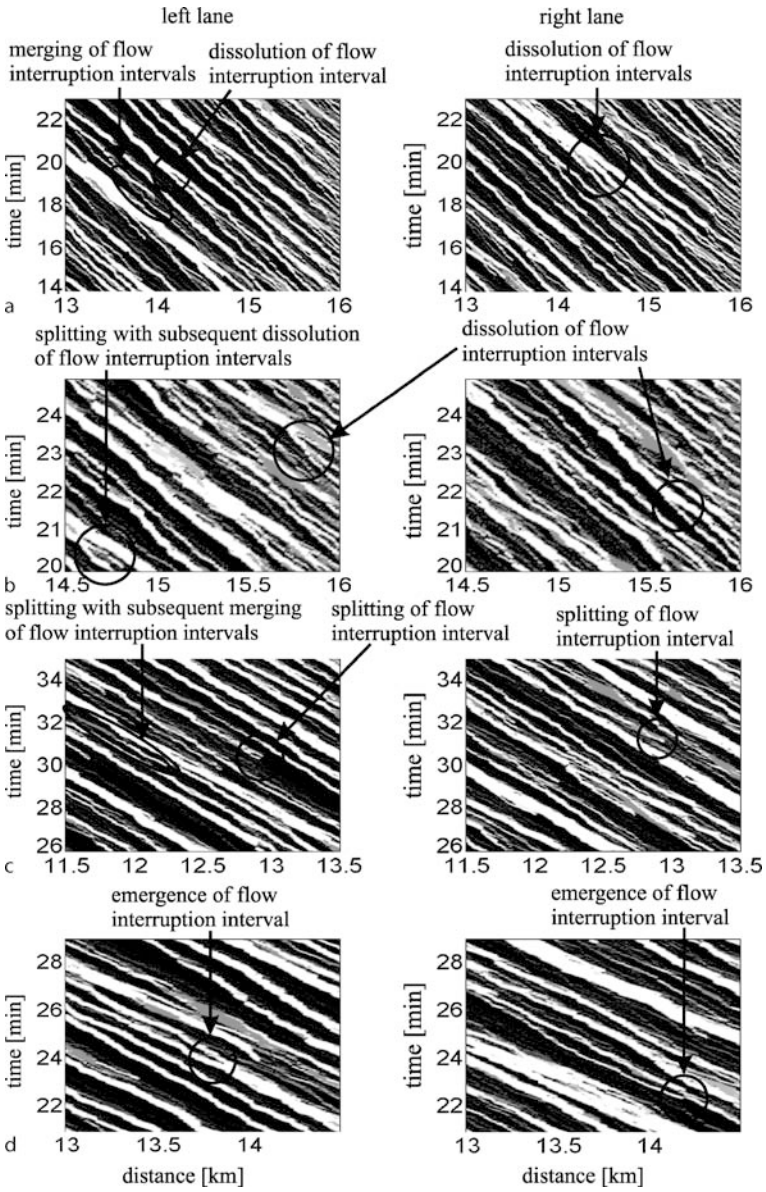

Traffic Congestion, Spatiotemporal Features of, Figure 46

Simulated single-vehicle speeds within GP presented in space and time by regions with variable darkness (the lower the speed, the *darker* the region) in the left (*left*) and right lanes (*right*). *White regions* (single-vehicle speeds higher than 3.6 km/h) are related to synchronized flows and moving blanks. *Black regions* (single vehicle speeds are equal to zero) are related to flow interruption intervals. $T_B = 30$ sec. $q^{(cong)} = 218$ vehicles/h/lane

a GP shows a sequence of wide moving jams separated by non-interrupted flows (Fig. 43); in contrast, within the mega-jam we can distinguish *no* sequence of wide moving jams (Fig. 47a,b). Rather than a sequence of wide moving jams, within the mega-jam there is a complex sequence of upstream moving flow interruption intervals (black regions in Fig. 48) for which criterion (17) is satisfied (Fig. 47c,d). These flow interruption intervals are separated by short time intervals within which vehicles move with low speeds (Fig. 47c,d). The latter time intervals are associated with moving blanks (white regions in Fig. 48).

Thus as the microscopic spatiotemporal structure of a wide moving jam, the microscopic structure of a mega-jam consists of an alternation of flow interruption inter-

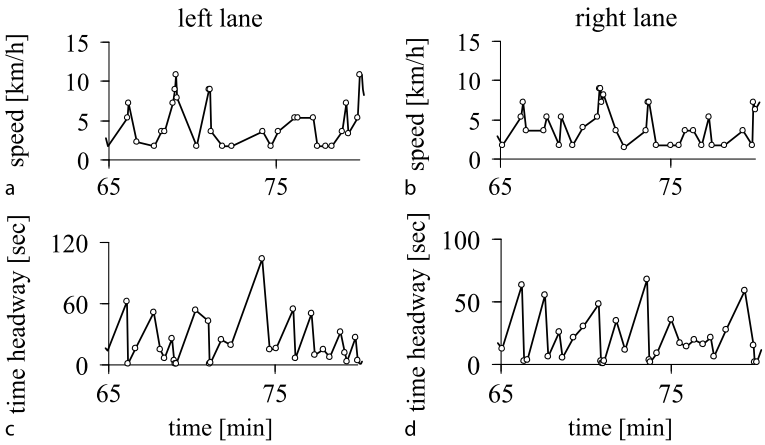

Traffic Congestion, Spatiotemporal Features of, Figure 47
Simulated single-vehicle data for speed (a, b) and time headway (c, d) measured by a virtual detector at location $x = 5$ km related to $T_B = 60$ sec. *Left* and *right* figures are related to the left and right lanes, respectively. The upstream boundary of the bottleneck is at location $x = 16$ km

vals and moving blanks (Fig. 48). This explain the term a *mega-jam* that is also a wide moving jam, however, with an extremely great width continuously growing over time. The continuous growth of the mega-jam width occurs as long as the flow rate upstream of the mega-jam q_{in} exceeds $q^{(cong)}$ (Fig. 38d). As for GPs with a non-regular pinch region, within the mega-jam we found a very complex and non-regular spatiotemporal dynamics of flow interruption intervals and moving blanks. However, because within the mega-jam there are no wide moving jams separated by non-interrupted flows, the non-regular mega-jam dynamics is associated with the dynamic effects (i), (iii), (v), and (vii) of Subsect. "Microscopic Spatiotemporal Features of Non-regular Moving Jam Dynamics" only.

Physics of Non-regular Spatiotemporal Dynamics of Traffic Congestion at Heavy Bottlenecks

A phase transition from synchronized flow to a wide moving jam is a first-order phase transition, which is characterized by a random time delay T_{SJ} [38]. T_{SJ} includes a random time delay of spontaneous nucleation of a narrow moving jam and a random time of the jam growth within the pinch region until the jam transforms into a wide moving jam at the upstream boundary of the pinch region. Thus the smaller T_{SJ}, the shorter $L^{(pinch)}$. The random character of T_{SJ} explains a time-dependence of $L^{(pinch)}(t)$ (Fig. 41c,e,g). We found that the greater the bottleneck strength, i. e., the smaller $q^{(cong)}$, the lower the speed and the greater the density within the pinch region. However, the lower the speed and the greater the density in synchronized flow, the smaller the nucleus required for the emergence of a flow interruption interval, i. e., the smaller the mean random time delay $T_{SJ}^{(mean)}$ ([38], ► Traffic Congestion, Modeling Approaches to).

Thus the greater the bottleneck strength, i. e., the smaller $q^{(cong)}$, the greater the probability for the occurrence of a negligibly small T_{SJ} and the emergence of wide moving jams directly upstream of the bottleneck; in the latter case $L^{(pinch)} = 0$. This explains why under condition (11) the regular structure of the GP breaks due to the random disappearance of the pinch region during some time intervals. This explains also why $L_{mean}^{(pinch)}$ decreases up to zero, when a strong increase in the bottleneck strength resulting in the subsequent strong decrease of $q^{(cong)}$ causes a decrease in $T_{SJ}^{(mean)}$ up to zero, i. e., when under condition (12) the pinch region does not exist.

As within the pinch region of synchronized flow of a GP, at greater bottleneck strengths resulting in smaller flow rates $q^{(cong)}$ within congested patterns, the synchronized flow speed between wide moving jams is also low (compare synchronized flow speeds between the jams in Fig. 43 for different T_B) and the density is great. Therefore, the probability of the emergence of flow interruption intervals between the jams, i. e., the emergence of new wide moving jams, increases rapidly with the bottleneck strength. As a result, many short flow interruption intervals appear at greater bottleneck strengths. On the other hand, short flow interruption intervals can dissolve easily, when the density decreases randomly upstream of these intervals. A spatiotemporal competition between the random emergence and dissolution of flow interruption inter-

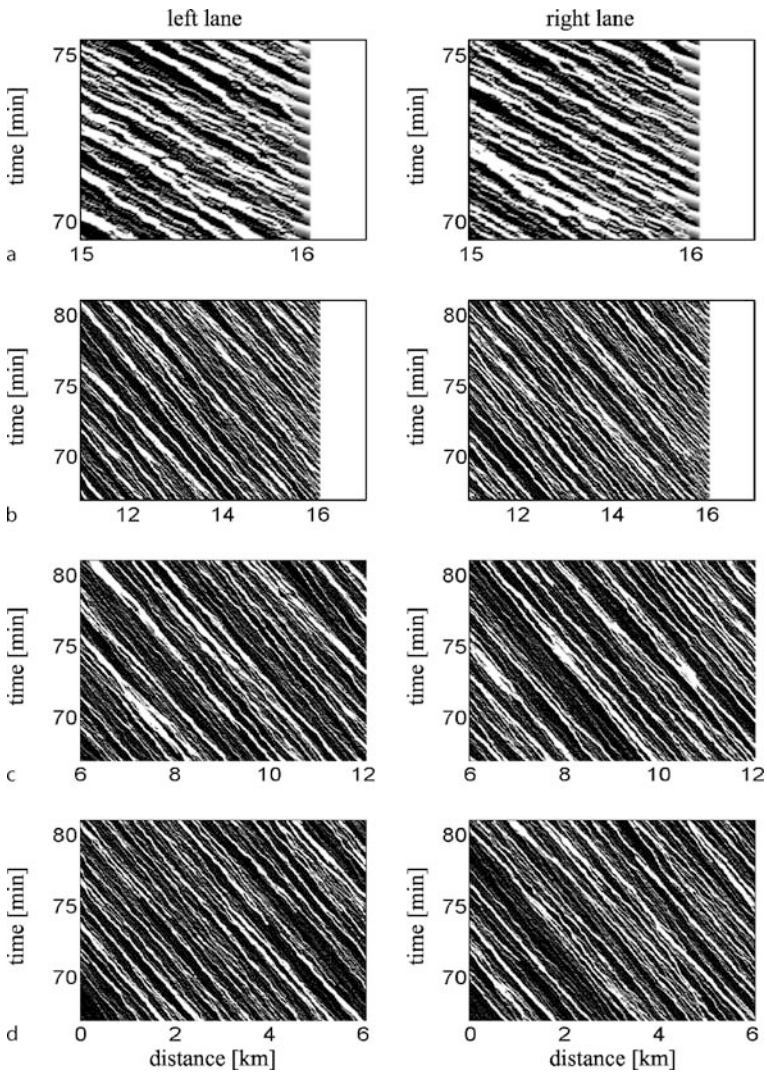

Traffic Congestion, Spatiotemporal Features of, Figure 48
Simulated single-vehicle speed data within mega-jam presented in space and time by regions with variable darkness (the lower the speed, the darker the region) in the left (*figures left*) and right lanes (*right*). *White regions* (single-vehicle speeds higher than 1.8 km/h) are related to moving blanks. *Black regions* (single vehicle speeds are equal to zero) are related to flow interruption intervals. $T_B = 60$ sec. The upstream boundary of the bottleneck is at location $x = 16$ km

vals can explain the non-regular spatiotemporal jam dynamics of Subsect. "Microscopic Spatiotemporal Features of Non-regular Moving Jam Dynamics".

There are several sources for random speed disturbances in the traffic flow model used for simulations considered in Subsect. "Evolution of Traffic Phases at Heavy Bottlenecks"–"Microscopic Spatiotemporal Structure of Mega-Jam", whose growth leads to the emergence of flow interruption intervals in synchronized flows: lane changing, the variety of the random delays in vehicle acceleration and decelerations ([38], ► Traffic Congestion, Modeling Approaches to). In addition, within the bottleneck vehicles are forced to move at much longer safe time headways $\tau^{(\mathrm{safe})} = T_B$ than away of the bottleneck $\tau^{(\mathrm{safe})} = 1$ sec. For this reason, the mean amplitude of speed disturbances in the pinch region is considerably greater than the one in synchronized flows between the jams on a homogeneous road. This explains why the non-regular jam dynamics is considerably visible only at greater bottleneck strengths than the critical bottleneck strength for the occurrence of GPs with a non-regular pinch region.

When the bottleneck strength increases strongly and condition (10), i. e., (13) is satisfied, wide moving jams merge into a mega-jam. The whole flow rate within any mega-jam is supplied by moving blanks *only*. As a result, under condition (10) we get

$$q^{(\mathrm{cong})} = q_{\mathrm{mega}}^{(\mathrm{blanks})} , \tag{18}$$

where $q_{\mathrm{mega}}^{(\mathrm{blanks})}$ denotes the average flow rate associated with moving blanks within the mega-jam, which satisfies the condition

$$q_{\mathrm{mega}}^{(\mathrm{blanks})} \leq q^{(\mathrm{blanks})} , \tag{19}$$

where, as defined above, $q^{(\mathrm{blanks})}$ is the average flow rate associated with moving blanks within wide moving jams separated by non-interrupted flows. The equality in (19) is related to (9). Thus if under condition (10) the bottleneck strength increases and therefore $q^{(\mathrm{cong})}$ decreases, then in accordance with (18), the flow rate $q_{\mathrm{mega}}^{(\mathrm{blanks})}$ decreases.

As mentioned in Subsect. "Evolution of Traffic Phases at Heavy Bottlenecks" (see(15)),the threshold bottleneck strength for the pinch region existence is smaller than the critical bottleneck strength for the mega-jam formation. To explain that this result has a general character, let us assume that the bottleneck strength is initially greater than the critical one for the mega-jam formation. In this case, $q^{(\mathrm{cong})} = q_{\mathrm{mega}}^{(\mathrm{blanks})} < q^{(\mathrm{blanks})}$. If now the bottleneck strength decreases gradually, then the flow rate $q^{(\mathrm{cong})}$ increases. The flow rate $q^{(\mathrm{cong})}$ can be supplied by moving blanks only up to the critical bottleneck strength satisfying condition (9), i. e., $q^{(\mathrm{cong})} = q^{(\mathrm{blanks})}$. When the bottleneck strength decreases further and becomes smaller than the critical one for the mega-jam formation, then $q^{(\mathrm{cong})} > q^{(\mathrm{blanks})}$. The difference $\Delta q_{\mathrm{blanks}} = q^{(\mathrm{cong})} - q^{(\mathrm{blanks})}$ should be supplied by vehicles accelerating from wide moving jams; this means that wide moving jams separated by synchronized flows must appear in congested traffic. In particular, such synchronized flows appear just upstream of the bottleneck as a result of the upstream propagation of wide moving jams that emerge directly upstream of the bottleneck. Nevertheless, a pinch region of the synchronized flow at the bottleneck does not still appear as long as the difference $\Delta q_{\mathrm{blanks}}$ is small enough. Indeed, the pinch region appears only, when at least one of the wide moving jams occurs at a *finite distance* upstream of the bottleneck in the synchronized flow due to a growing narrow moving jam that emerges initially in this flow. This is possible only, if the former wide moving jam due to its upstream propagation is at a great enough distance from the bottleneck. The latter can be realized, when the difference $\Delta q_{\mathrm{blanks}}$ exceeds also some *finite value*, i. e., the threshold bottleneck strength for the pinch region existence should be always smaller than the critical one for the mega-jam formation. Thus condition (15) is a general result for any heavy bottleneck.

Comparison with Empirical Results

To compare the above theory of traffic congestion at heavy bottlenecks [39,40] with empirical congested patterns, one should have measured data for traffic congestion at a bottleneck, whose strength should be manually continuously changeable from the one associated with usual bottlenecks like on- and off-ramps to great bottleneck strengths associated with heavy bottlenecks caused by bad weather conditions or accidents. Unfortunately, such measured data is not available. However, we can compare the theory with measured data related to two limiting cases:

(i) Traffic congestion at an usual on-ramp bottleneck (Fig. 16).
(ii) Traffic congestion at a heavy bottleneck caused by bad weather conditions – snow and ice on a road (Fig. 49).

The speed distributions with congested patterns found in simulations (Fig. 38a,b) for the range of $T_{\mathrm{B}} = 1.6$–2.4 sec associated with the flow rate range

$$q^{(\mathrm{cong})} = q^{(\mathrm{pinch})} = 1120\text{–}1800 \text{ vehicles/h/lane} \tag{20}$$

are qualitatively the same as those in an empirical GP shown in Fig. 16 and in all other known empirical GPs [38]. Moreover, quantitative values of empirical flow rates $q^{(\mathrm{pinch})}$ (7) are approximately associated with the theoretical result (20). In particular, within the pinch region of the GP shown in Fig. 16a, the flow rate averaged between 7:00 and 8:00 and across the road is $q^{(\mathrm{pinch})} = 1200$ vehicles/h/lane.

In contrast, for $T_{\mathrm{B}} > 6$ sec associated with the average flow rate $q^{(\mathrm{cong})} < 625$ vehicles/h/lane simulated congested patterns at heavy bottlenecks (Fig. 38c,d) are qualitatively the same as those we found in empirical data measured on many various days (and years) on different freeways, when very heavy bottlenecks caused by bad weather conditions or accidents occur for which empirical flow rates

$$q^{(\mathrm{cong})} \lesssim 600 \text{ vehicles/h/lane} . \tag{21}$$

In this case, rather than regular structure of traffic congestion of GPs (Figs. 38a,b and 16), both theoretical (Figs. 38c,d, 40i) and empirical traffic congested patterns (Fig. 49) exhibit non-regular spatiotemporal structure of

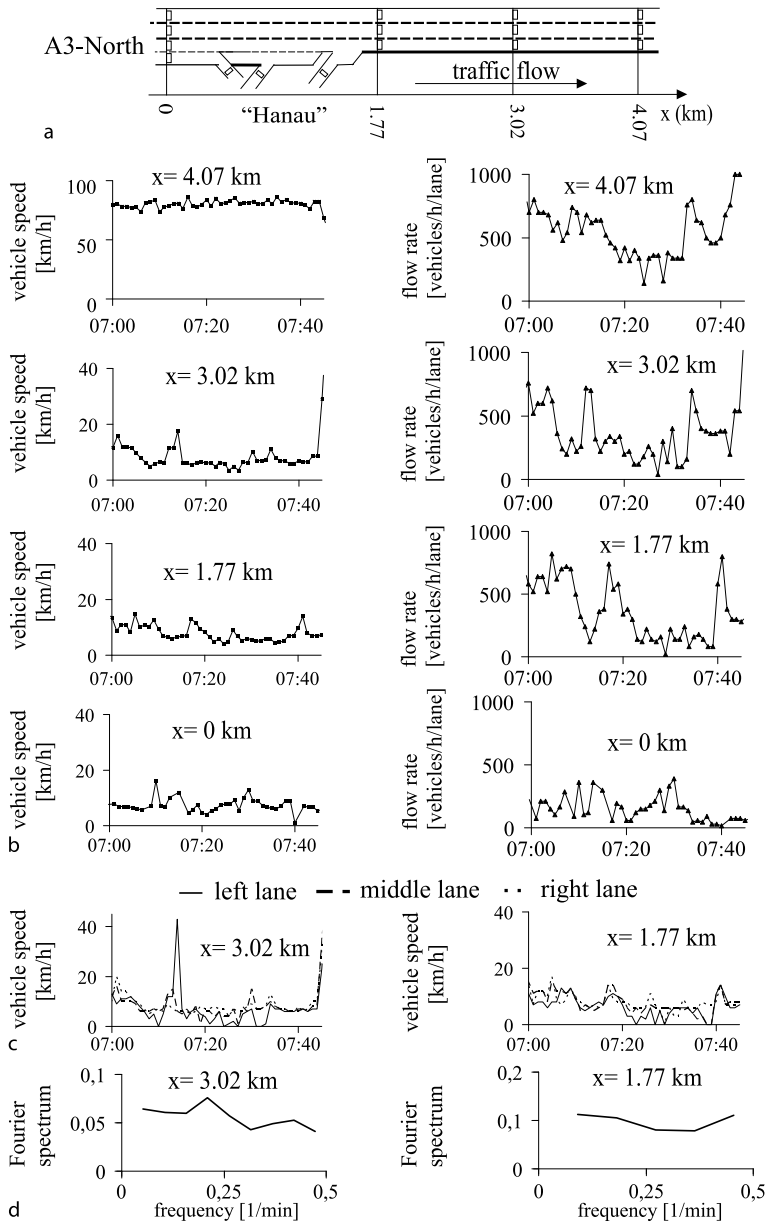

Traffic Congestion, Spatiotemporal Features of, Figure 49

Empirical structure of congestion caused by snow and ice: a Scheme of road detector arrangement on a section of the freeway A3-North in Germany near the intersection "Hanau". b Average speed and flow rate across the freeway at different locations. c Speed in different lanes at two locations. d Fourier spectra of time-dependences of the speed in the left lane at two locations. 1 min average data

congestion in which *no* sequences of wide moving jams can be distinguished in 1 min average data.

An example of such an empirical congested traffic pattern is shown in Fig. 49. Rather than the regular structure of congestion within the GP (Fig. 16), in measured data associated with bad weather conditions a non-regular spatiotemporal structure of congestion is observed (Fig. 49). A heavy bottleneck appears on February 02, 2006 between locations 4.07 and 3.02 km due to snow and ice. Upstream of the bottleneck, very low speed and flow rate patterns ($x \leq 3.02$ km in Fig. 49b) are observed. Downstream of the bottleneck ($x = 4.07$ km) vehicles have escaped from

the congestion (speed is high), however, the flow rate is very small because the bottleneck reduced the average flow rate within the congestion strongly. For example, at $x = 3.02$ km, the flow rate $q^{(\mathrm{cong})}$ averaged between 7:00 and 7:40 and across the road is 513 vehicles/h/lane. In contrast with the GP shown in Fig. 16, within traffic congestion shown in Fig. 49 non-regular low speed patterns are observed in which *no* sequence of wide moving jams can be distinguished ($x \leq 3.02$ km). This conclusion is regardless of the flow rates to on- and off-ramps, which in the data set lead to an increase in $q^{(\mathrm{cong})}$ at $x = 1.77$ km in comparison with $q^{(\mathrm{cong})}$ at $x = 0$ km (Fig. 49a,b). If speeds in different lanes are compared (Fig. 49c), we find even more non-regular speed time-dependences: whereas no vehicles pass a detector in one of the lanes, i. e., the speed is zero, during the same time interval the average speed in other lanes can be higher than zero. This explains why in time-dependences of speeds averaged across the road (Fig. 49b), this very low speed is seldom equal to exactly zero.

As congested traffic in measured data (Fig. 49b,c), simulated congested traffic at a heavy bottleneck is also non-homogeneous in space and time (Fig. 40i,j). In the measured and simulated data, as follows from Fourier spectra of speed time-dependences, these non-homogeneous congested patterns exhibit non-regular spatiotemporal behavior (Figs. 49d and 40l). This is in contrast with the regular structure of traffic congestion within the GP (Fig. 16).

As in measured data (Fig. 49b,c), in 1 min average data related to traffic congestion in simulations very low speed and flow rate patterns are realized upstream of the bottleneck in which *no* sequence of wide moving jams can be distinguished; downstream of the bottleneck ($x > 16.3$ km) vehicles have escaped from the congestion (speed is high) (Fig. 38c), however, the flow rate is very small because the bottleneck reduced the average flow rate within the congestion strongly. A summary of the above comparison of theory of traffic congestion with measured data is as follows:

(i) Spatiotemporal distributions of the average speed and flow rate within GPs occurring at usual bottlenecks are qualitatively the same in the above theory and measured data: in both simulated and measured data, GPs consist of the pinch region and wide moving jams upstream. Measured and simulated speed time-dependences show *regular* character of wide moving jam propagation.
(ii) In contrast with item (i), in 1 min average data related to simulated congested patterns and to measured data for traffic congestion at heavy bottlenecks *no* sequence of wide moving jams can be distinguished. In simulations and measured data, spatiotemporal distributions of the average speed and flow rate in these congested patterns are non-homogeneous in space and time. Fourier spectra of the associated non-homogeneous time-dependences of the speed show non-regular character of traffic congestion at the heavy bottlenecks both in simulations (Fig. 40l) and measured data (Fig. 49d).

As follows from Subsects. "Microscopic Spatiotemporal Features of Non-regular Moving Jam Dynamics" and "Microscopic Spatiotemporal Structure of Mega-Jam", the non-regular pattern behavior of item (ii) can be explained by two different reasons:

(1) The occurrence of a GP with the non-regular dynamics of wide moving jams.
(2) The occurrence of a mega-jam.

However, based on 1 min average data these two reasons of non-regular traffic congestion cannot be distinguished from each other. Indeed, theoretical results of Subsects. "Microscopic Spatiotemporal Features of Non-regular Moving Jam Dynamics" and "Microscopic Spatiotemporal Structure of Mega-Jam" allow us to suggest that at a very heavy bottleneck caused for example by bad weather conditions or accidents there should be a very fine microscopic spatiotemporal structure of traffic congestion associated with non-regular spatiotemporal dynamics of wide moving jams or/and flow interruption intervals as well as moving blanks. However, this theoretical fine structure of traffic congestion, even if it exists in real traffic, would be averaged in 1 min average measured data, i. e., they could not be found in this data. Thus to make a more detailed comparison of the theory of traffic congestion at heavy bottlenecks with measured data, single vehicle data measured within traffic congestion at heavy bottlenecks is required. Unfortunately, currently this measured single vehicle data is not available. Such a comparison of the theory of traffic congestion at heavy bottlenecks presented in Sect. "Congested Patterns at Heavy Bottlenecks" with measured single vehicle data is a separate and important task of further investigations.

Conclusions. Fundamental Empirical Features of Spatiotemporal Congested Freeway Traffic Patterns

There are empirical features of phase transitions and spatiotemporal congested patterns at freeway bottlenecks that are reproducible in traffic observations on numerous days and years on different freeways in various countries. Thus,

these qualitative empirical features remain in traffic flows with very different driver behavioral characteristics and vehicle parameters. These fundamental empirical macroscopic and microscopic features of the phase transitions and congested traffic are explained within the framework of three-phase traffic theory.

Fundamental empirical features of phase transitions and Congested patterns at freeway bottlenecks are as follows:

(i) Traffic can be either "free" or "congested." There are two types of the onset of congestion (traffic breakdown) in free flow at an effectual bottleneck, *spontaneous* and *induced*. The traffic breakdown and freeway capacity at the bottleneck possess a probabilistic nature.

(ii) An induced speed breakdown in free flow at the bottleneck can be caused either by moving jam propagation through the bottleneck or the upstream propagation of a synchronized flow pattern (SP) that has initially occurred downstream of the bottleneck.

(iii) In congested traffic, two different traffic phases can be distinguished: the synchronized flow phase and the wide moving jam phase. These traffic phases are defined through the use of the spatiotemporal empirical (objective) criteria [J] and [S]. Thus, there are three traffic phases in freeway traffic: (1) free flow; (2) synchronized flow; (3) wide moving jam.

(iv) Traffic breakdown in free flow at the freeway bottleneck is associated with a phase transition from free flow to synchronized flow ($F \rightarrow S$ transition).

(v) Wide moving jams do not emerge spontaneously in free flow. Wide moving jams can emerge spontaneously only in synchronized flow; wide moving jams emerge due to the sequence of $F \rightarrow S \rightarrow J$ transitions.

(vi) The higher the density in synchronized flow, the higher the frequency of spontaneous moving jam emergence in that synchronized flow. In synchronized flow of higher speeds and lower densities, wide moving jams should not necessarily emerge spontaneously.

(vii) During the dynamics of transformation of a narrow moving jam into a wide moving jam the mean velocity of the downstream jam front tends to the characteristic velocity. Wide moving jams possess *characteristic* parameters and features that do not depend on initial conditions and perturbations in traffic. The characteristic parameters are the same for different wide moving jams. The characteristic parameters can depend on control parameters of traffic (weather, road conditions, etc.). These characteristic parameters and features are as follows:

(a) The mean velocity of the downstream front of a wide moving jam. The wide moving jam propagates upstream through any traffic state and through any bottleneck while maintaining the mean velocity of the downstream jam front.

(b) When free flow occurs in the wide moving jam outflow, the flow rate and density in the jam outflow are also characteristic parameters.

(c) The flow rate in the jam outflow is lower than the maximum possible flow rate in free flow.

(viii) There are two main types of congested patterns at an isolated effectual freeway bottleneck: synchronized flow (SP) and general patterns (GP).

(ix) In contrast to wide moving jam propagation, an SP that propagates upstream is caught at the bottleneck at which free flow has been before (catch effect) when the SP reaches the bottleneck. Due to the catch effect, an induced $F \rightarrow S$ transition can occur.

(x) There are three types of SPs: a moving SP, a widening SP, and a localized SP.

(xi) There can be complex spontaneous transformations between various congested patterns at the bottleneck over time.

(xii) At the same traffic demand various congested patterns can be formed at the bottleneck (the probabilistic nature of spatiotemporal congested patterns at the bottleneck).

(xiii) If two or more adjacent effectual bottlenecks are close to one another, expanded congested patterns (EP) often emerge. In an EP, synchronized flow affects at least two adjacent effectual bottlenecks.

(xiv) When a wide moving jam that has initially emerged downstream of an effectual bottleneck propagates through synchronized flow upstream of the bottleneck, this foreign wide moving jam can considerably influence other moving jams, which are just emerging in that synchronized flow. In particular, the foreign wide moving jam can suppress the growth of a narrow moving jam that is close enough to the downstream front of the foreign wide moving jam.

(xv) For each effectual freeway bottleneck, or each set of several adjacent effectual bottlenecks where congested patterns occur, the spatiotemporal structure of a congested pattern possesses some predictability, i. e., characteristic, unique, and reproducible features.

(xvi) Rather than regular wide moving jam propagation observed upstream of usual bottlenecks like on- and

off-ramp bottlenecks (item iii-xv), no sequence of wide moving jams can be distinguished in 1 min average measured data for non-regular and non-homogeneous congested traffic, which is usually observed upstream of a heavy bottleneck caused for example by bad weather conditions or accidents. This empirical phenomenon can be explained by results of a theory of traffic congestion at heavy bottlenecks of Sect. "Congested Patterns at Heavy Bottlenecks", in particular by random disappearance and appearance of the pinch region of synchronized flow of a GP over time and a non-regular dynamics of wide moving jams within the GP, as well as by the merger of wide moving jams into a mega-wide moving jam (mega-jam) that should occur upstream of a very heavy bottleneck.

Future Directions

Although many empirical macroscopic spatiotemporal congested pattern features have already been understood, there are still many problems in understanding empirical microscopic features of these patterns as well as phase transitions within congested patterns, in particular occurring on freeway sections with many adjacent bottlenecks.

Microscopic features of phase transitions between the synchronized flow and wide moving jam phases have not still sufficiently been understood.

Spatiotemporal features of congested patterns associated with freeway networks, e. g., possible induced phase transitions and possible complex pattern transformations caused by upstream propagation of congested patterns onto neighborhood freeway sections of the network have almost not been found.

Traffic flow models in the framework of three-phase traffic theory are only at the beginning of their development. There are almost no analytical studies of the spatiotemporal characteristics of congested patterns within this theory.

An empirical microscopic spatiotemporal dynamics of traffic congestion at heavy bottlenecks has not still been studied.

These and many other unsolved problems are interesting and important fields of the future empirical and theoretical investigations of spatiotemporal congested traffic patterns.

We can also expect that features of spatiotemporal dynamics of traffic congestion reviewed in this article can be applied for a development of new effective methods of traffic control and management as well as driver assistant systems that increase vehicle safety in traffic.

Acknowledgments

I would like to thank Sergey Klenov, Andreas Hiller, Hubert Rehborn, Mario Aleksić, Ines Maiwald-Hiller and Olivia Brickley for help and useful suggestions.

Bibliography

1. Bellomo N, Coscia V, Delitala M (2002) On the Mathematical Theory of Vehicular Traffic Flow I. Fluid Dynamic and Kinetic Modelling. Math Mod Meth Appl Sci 12:1801–1843
2. Chowdhury D, Santen L, Schadschneider A (2000) Statistical Physics of Vehicular Traffic and Some Related Systems. Phys Rep 329:199
3. Cremer M (1979) Der Verkehrsfluss auf Schnellstrassen. Springer, Berlin
4. Daganzo CF (1997) Fundamentals of Transportation and Traffic Operations. Elsevier, New York
5. Davis LC (2004) Multilane simulations of traffic phases. Phys Rev E 69:016108
6. Davis LC (2006) Controlling traffic flow near the transition to the synchronous flow phase. Physica A 368:541–550
7. Davis LC (2006) Effect of cooperative merging on the synchronous flow phase of traffic. Physica A 361:606–618
8. Davis LC (2007) Effect of adaptive cruise control systems on mixed traffic flow near an on-ramp. Physica A 379:274–290
9. Edie LC (1961) Car-Following and Steady State Theory for Non-Congested Traffic. Oper Res 9:66–77
10. Edie LC, Foote RS (1958) Traffic Flow in Tunnels. Highw Res Board Proc Ann Meet 37:334–344
11. Edie LC, Foote RS (1960) Effect of Shock Waves on Tunnel Traffic Flow. In: Highway Research Board Proceedings, vol 39. HRB, National Research Council, Washington DC, pp 492–505
12. Edie LC, Herman R, Lam TN (1980) Observed Multilane Speed Distribution and the Kinetic Theory of Vehicular Traffic. Transp Sci 14:55–76
13. Elefteriadou L, Roess RP, McShane WR (1995) Probabilistic Nature of Breakdown at Freeway Merge Junctions. Transp Res Rec 1484:80–89
14. Fukui M, Sugiyama Y, Schreckenberg M, Wolf DE (eds) (2003) Traffic and Granular Flow' 01. Springer, Berlin
15. Gao K, Jiang R, Hu SX, Wang BH, Wu QS (2007) Cellular-automaton model with velocity adaptation in the framework of Kerner's three-phase traffic theory. Phys Rev E 76:026105
16. Gartner NH, Messer CJ, Rathi A (eds) (1997) Special Report 165: Revised Monograph on Traffic Flow Theory. Transportation Research Board, Washington DC
17. Haight FA (1963) Mathematical Theories of Traffic Flow. Academic Press, New York
18. Hall FL, Agyemang-Duah K (1991) Freeway capacity drop and the definition of capacity. Trans Res Rec 1320:91–98
19. Hall FL, Hurdle VF, Banks JH (1992) Synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relationships on freeways. Transp Res Rec 1365:12–18
20. Helbing D (2001) Traffic and related self-driven many-particle systems. Rev Mod Phys 73:1067–1141
21. Helbing D, Herrmann HJ, Schreckenberg M, Wolf DE (eds) (2000) Traffic and Granular Flow' 99. Springer, Berlin
22. Hoogendoorn SP, Luding S, Bovy PHL, Schreckenberg M, Wolf DE (eds) (2005) Traffic and Granular Flow' 03. Springer, Berlin

23. Jiang R, Wu QS (2004) Spatial-temporal patterns at an isolated on-ramp in a new cellular automata model based on three-phase traffic theory. J Phys A: Math Gen 37:8197–8213
24. Jiang R, Wu QS (2005) First order phase transition from free flow to synchronized flow in a cellular automata model. Eur Phys J B 46:581–584
25. Jiang R, Wu QS (2005) Toward an improvement over Kerner-Klenov-Wolf three-phase cellular automaton model. Phys Rev E 72:067103
26. Jiang R, Wu QS (2007) Dangerous situations in a synchronized flow model. Physica A 377:633–640
27. Jiang R, Hua MB, Wang R, Wu QS (2007) Spatiotemporal congested traffic patterns in macroscopic version of the Kerner-Klenov speed adaptation model. Phys Lett A 365:6–9
28. Kerner BS (1998) Theory of Congested Traffic Flow. In: Rysgaard R (ed) Proceedings of the 3rd Symposium on Highway Capacity and Level of Service, vol 2. Road Directorate, Ministry of Transport – Denmark, pp 621–642
29. Kerner BS (1998) Traffic Flow: Experiment and Theory. In: Traffic and Granular Flow' 97. Proceedings of the International Workshop on Traffic and Granular Flow. Springer, Singapore, pp 239–267
30. Kerner BS (1998) Empirical features of self-organization in traffic flow. Phys Rev Lett 81:3797–3400
31. Kerner BS (1999) Congested Traffic Flow: Observations and Theory. Trans Res Rec 1678:160–167
32. Kerner BS (1999) Theory of Congested Traffic Flow: Self-Organization without Bottlenecks. In: Ceder A (ed) Transportation and Traffic Theory. Elsevier, Amsterdam, pp 147–171
33. Kerner BS (1999) The Physics of Traffic. Phys World 12:25–30
34. Kerner BS (2000) Phase Transitions in Traffic Flow. In: Traffic and Granular Flow' 99. Springer, Berlin, pp 253–284
35. Kerner BS (2000) Experimental features of the emergence of moving jams in free traffic flow. J Phys A: Math Gen 33:L221–L228
36. Kerner BS (2002) Empirical macroscopic features of spatial-temporal traffic patterns at highway bottlenecks. Phys Rev E 65:046138
37. Kerner BS (2004) Three-phase traffic theory and highway capacity. Physica A 333:379–440
38. Kerner BS (2004) The Physics of Traffic. Springer, Berlin
39. Kerner BS (2007) Features of Traffic Congestion caused by bad Weather Conditions and Accidents. E-print arXiv:0712.1728
40. Kerner BS (2008) A Theory of Traffic Congestion at Heavy Bottlenecks. J Phys A Math Theor 41:215101
41. Kerner BS, Klenov SL (2002) A microscopic model for phase transitions in traffic flow. J Phys A: Math Gen 35:L31–L43
42. Kerner BS, Klenov SL (2003) Microscopic theory of spatio-temporal congested traffic patterns at highway bottlenecks. Phys Rev E 68:036130
43. Kerner BS, Klenov SL (2004) Spatialtemporal patterns in heterogeneous traffic flow with a variety of driver behavioural characteristics and vehicle parameters. J Phys A: Math Gen 37:8753–8788
44. Kerner BS, Klenov SL (2006) Deterministic microscopic three-phase traffic flow models. J Phys A: Math Gen 39:1775–1809
45. Kerner BS, Klenov SL, Wolf DE (2002) Cellular automata approach to three-phase traffic theory. J Phys A: Math Gen 35:9971–10013
46. Kerner BS, Klenov SL, Hiller A (2006) Criterion for traffic phases in single vehicle data and empirical test of a microscopic three-phase traffic theory. J Phys A: Math Gen 39:2001–2020
47. Kerner BS, Klenov SL, Hiller A, Rehborn H (2006) Microscopic features of moving traffic jams. Phys Rev E 73:046107
48. Kerner BS, Klenov SL, Hiller A (2007) Empirical test of a microscopic three-phase traffic theory. Non Dyn 49:525–553
49. Knospe W, Santen L, Schadschneider A, Schreckenberg M (2002) Single-vehicle data of highway traffic: Microscopic description of traffic phases. Phys Rev E 65:056133
50. Knospe W, Santen L, Schadschneider A, Schreckenberg M (2004) Empirical test for cellular automaton models of traffic flow. Phys Rev E 70:016115
51. Koshi M, Iwasaki M, Ohkura I (1983) Some Findings and an Overview on Vehiclular Flow Characteristics. In: Hurdle VF (ed) Proc. 8th International Symposium on Transportation and Traffic Theory. University of Toronto Press, Toronto, pp 403
52. Laval JA (2007) Linking Synchronized Flow and Kinematic Waves. In: Traffic and Granular Flow' 05. Proceedings of the International Workshop on Traffic and Granular Flow. Springer, Berlin, pp 521–526
53. Lee HK, Barlović R, Schreckenberg M, Kim D (2004) Mechanical Restriction versus Human Overreaction Triggering Congested Traffic States. Phys Rev Lett 92:238702
54. Leutzbach W (1988) Introduction to the Theory of Traffic Flow. Springer, Berlin
55. Mahmassani HS (ed) (2005) Transportation and Traffic Theory. Proceedings of the 16th Inter. Sym. on Transportation and Traffic Theory. Elsevier, Amsterdam
56. Mahnke R, Kaupužs J, Lubashevsky I (2005) Probabilistic description of traffic flow. Phys Rep 408:1–130
57. May AD (1990) Traffic Flow Fundamentals. Prentice-Hall, New Jersey
58. Nagatani T (2002) The physics of traffic jams. Rep Prog Phys 65:1331–1386
59. Nagel K, Wagner P, Woesler R (2003) Still Flowing: Approaches to Traffic Flow and Traffic Jam Modeling. Oper Res 51:681–716
60. Neubert L, Santen L, Schadschneider A, Schreckenberg M (1999) Single-vehicle data of highway traffic: A statistical analysis. Phys Rev E 60:6480–6490
61. Persaud BN, Yagar S, Brownlee R (1998) Exploration of the Breakdown Phenomenon in Freeway Traffic. Trans Res Rec 1634:64–69
62. Pottmeier A, Thiemann C, Schadschneider A, Schreckenberg M (2007) Mechanical Restriction versus Human Overreaction: Accident Avoidance and Two-Lane Simulations. In: Traffic and Granular Flow' 05. Proceedings of the International Workshop on Traffic and Granular Flow. Springer, Berlin, pp 503–508
63. Prigogine I, Herman R (1971) Kinetic Theory of Vehicular Traffic. Elsevier, New York
64. Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) (2007) Traffic and Granular Flow' 05. Proceedings of the International Workshop on Traffic and Granular Flow. Springer, Berlin
65. Schreckenberg M, Wolf DE (eds) (1998) Traffic and Granular Flow' 97. Proceedings of the International Workshop on Traffic and Granular Flow. Springer, Singapore
66. Sugiyama Y, Fukui M, Kikuchi M, Hasebe K, Nakayama A, Nishinari K, Tadaki S, Yukawa S (2008) Traffic jam without bottleneck – Experimental evidence for the physical mechanism of forming a jam. New J Phys 10:033001

67. Tilch B, Helbing D (2000) Evaluation of Single Vehicle Data in Dependence of the Vehicle-Type, Lane, and Site. In: Traffic and Granular Flow' 99. Springer, Berlin, pp 333–338
68. Treiterer J (1975) Investigation of Traffic Dynamics by Aerial Photogrammetry Techniques. Technical Report PB 246 094. Ohio State University, Columbus
69. Treiterer J, Myers JA (1974) The Hysteresis Phenomenon in Traffic Flow. In: Buckley DJ (ed) Procs. 6th International Symposium on Transportation and Traffic Theory. Reed, London, pp 13–38
70. Treiterer J, Taylor JI (1966) Traffic Flow Investigations by Photogrammetric Techniques. Highway Res Rec 142:1–12
71. Wang R, Jiang R, Wu QS, Liu M (2007) Synchronized flow and phase separations in single-lane mixed traffic flow. Physica A 378:475–484
72. Whitham GB (1974) Linear and Nonlinear Waves. Wiley, New York
73. Wiedemann R (1974) Simulation des Verkehrsflusses. University of Karlsruhe, Karlsruhe
74. Wolf DE (1999) Cellular automata for traffic simulations. Phys A 263:438–451

Traffic and Crowd Dynamics: The Physics of the City

Armando Bazzani, Bruno Giorgini, Sandro Rambaldi
"Fisica della Città" Laboratory, Center L. Galvani for Biocomplexity, Physics Department and INFN Sezione di Bologna, Bologna, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Automata Gas and Crowd Behavior
Automata Gas and Traffic Behavior
Simulation Results and Empirical Observations
Future Directions
Bibliography

Glossary

Automaton A physical particle with internal cognitive dynamics able to process information and to make decisions in order to achieve an aim.

Cognitive dynamics A dynamical system that mimics the individual decision mechanisms pointed out by neuroscience and behavioral studies.

Emergent properties Macroscopic properties of a statistical systems that are not the result of the superposition of the microscopic states of individual particles.

Global positioning system A system to detect the geographical coordinates of a receiver on the earth using a satellite network.

Mean field theory a technique introduced in statistical physics that studies the macroscopic dynamics using the dynamics of averaged quantities.

Self-organized dynamics Stationary macroscopic states that are the result of collective interactions among the microscopic components.

Utility function A function that measures the expected advantages or the disadvantages of future decisions.

Definition of the Subject

The physics of the city has been proposed as an interdisciplinary research field that applies the techniques of exact sciences to describe and understand the dynamical states of the new metropolis. The continuous growth and the social changes outlined by urban planners and sociologists, give to the actual cities a complex structure whose study requires a new generation of dynamical models. From this point of view, the traffic and crowd dynamics have a fundamental importance due to their impact on individual freedom and life quality. The study of the observed macroscopic phenomena using a complex systems approach, the search for the control parameters, and the understanding of the individual mobility demand are key points for future contributions of the complexity science to real human problems.

Introduction

Crowd and traffic dynamics are a crucial issue for modeling the complex urban mobility [5,8,11]. The idea that the city itself can be considered a whole laboratory to understand the mobility is at the base of a new research field: the physics of the city [13]. "Aggregate of beings that hold their biological history into its borders and model it within all their intentions proper to thinking creatures, the city results at the same time by the biological generation, the organic evolution, and the aesthetics creativity. The city is contemporary a natural object and a subject of culture". From these words of Claude Levi Strauss clearly emerges the complex nature of every urban system that today it is generally accepted, and considered in order to study the city, its development and evolution [8]. The city shows itself to be polymorphous, polysemic and polyglot [22], stratified in time and crossed by actors and objects whose dynamics are extremely different and can be conflictive to the point to engender a feeling of chaos, i. e., one can say that cities live at the edge of chaos and the problem is exactly to govern the system emerging complexity, and not

real chaos situations. To try to express the quality of this complexity in quantitative terms, using the instruments of the exact sciences, without losing its texture, one has to reduce the semantic, logical, syntactic and phenomenological field in which to articulate the possibilities of constructing models able to be descriptive, explanatory and, at least to some extent, predictive [78]. These models will never be completely isomorphic to reality, but can simulate some of its aspects and characteristics considered salient by the observer, thus playing a fundamental role in the choice of the significant observables and control parameters, which obviously depend from what one wants to control and govern. A simple consideration can help us in the reduction process: regardless of the variety and complexity of flows, forms and information, an urban system exists insofar as it is inhabited. Therefore our physics of the city will be essentially physics of an inhabited city, and given the large number of elementary components, this means non-equilibrium statistical physics [61,77], because a town is an open system. Moreover, for the elementary components moving in urban space-time, this also means the physics of dynamical systems, and since the individuals in the system have free will, probabilistic physics has to be considered (Pascal firstly modeled the human free will by the probability function, using the game of chance as paradigm). To finish, since the elementary components have memory and are capable of drawing information from the environment, processing it according to intentions, choices and decisions, the physics of the city must also be intentional, cognitive and decisional. As everyone knows, another basic ingredient of physics, besides the elementary components, is a space-time where the dynamics can develop (ubi materia, ibi geometria – Kepler). So one can ask if it is possible to identify a space-time structure proper to a generic urban system. Obviously we can describe the street network, and the different morphologies with a spatial metric that usually is not Euclidean, but is not sufficient to develop an urban dynamics. One needs also a clock, a time structure scale invariant. This structure can be modeled by the chronotopoi (literally, places of time), the primal agents of urban temporal dynamics able to generate time correlations that would not exist without them [10]. In the city planning language, they are defined as areas where are implanted, temporal scheduled activities, for example a hospital, the university, a shopping center and so on, that generates/attract mobility. Urban topology thus becomes chronotopic, and the interaction of the individual's agenda with the pulsation of the chronotopos produces complex urban mobility [36].

Assuming the mobility as a property of the individuals able to process information, the microscopic dynamics one generates both by properly physical interactions and by some decisional/cognitive mechanism [1]. Moreover a mesoscopic dynamics for individuals aggregate has been developed to study self-organization behaviors in the context of a non-equilibrium and non-linear evolution of the system (being the urban mobility a complex phenomenon that develops itself in an open complex system, the city) [39,45]. Using a computer different mobility situations are simulated and analyzed, in order to study the control parameters, phase transitions and shift from order to chaos, criticalities and the whole range of possible orbits, also those that are very difficult to reach in material experiments [24,25]. Obviously there is a problem of realism, because the simulations can be consistent and reasonable but far, sometime very far, from the effective mobility.

In general social systems, like urban traffic and crowd dynamics, consist of many autonomous intelligent entities, which are distributed over a structured space and interact each other to achieve certain goals [33,68]. A powerful method to study such systems are multi-agent models, which allow one introduce a cognitive behavior in the dynamics of individual entities. The concept of "agent" has been proposed in many research areas [17,67,87], from computer science and artificial life research to economy and sociology, so that its definition is quite generic: agent is anything able to act in a given environment. To suggest a more specific agent definition in the simulation of traffic and crowd dynamics [6,7,94], the crucial issue concerning the meaning and the utility of a social system model has to be firstly discussed [19,20,33]. Complexity science faces the problem of studying systems whose dynamics cannot be reduced to physical fundamental principles. The different scales of description turn out to be interconnected, not only by a bottom-up interaction structure typical of physical systems (the dynamics at microscopic levels determines the evolution of macroscopic variables), but also by top-down interactions, so that the microscopic behavior is influenced by the macroscopic states of the system [30]. A classical statistical physics approach is questionable in such a case. A further difficulty is that even the interactions among the elementary components of a complex system are often unknown, and they can only be described in a qualitative way according to some hypotheses, consequences of real phenomena direct observations [78]. Usually a model is a simplified representation of the reality, but not too simple in order to describe the interested phenomenon; moreover any model should have sufficiently generic features to be able to describe analogous phenomena observed in different systems from social to physical or biological ones [76,90]. In complexity science the role of the models is modified becoming not a specific

realization of a general theory (think for example to different cosmological models which correspond to different boundary conditions and solutions for the Einstein equations), but rather an instrument of knowledge to perform in silico experiments that allow to verify the consistency of the assumptions with the considered phenomena. In other words the models realize a "virtual reality" where the simulations point out the relevance of different hypotheses on the appearance and evolution of emergent properties identifying the control parameters. This modeling cannot be validated in the usual Galilean sense, since the models are themselves experiments. Moreover there is an intrinsic difficulty in defining Galilean experiments on complex systems, due to the great information needed for the reproducibility [9,34]. From a physical point of view, one has to look for statistical laws, that consider the minimal conditions for the existence of phase transitions toward particular self-organized states.

In our opinion, a paradigmatic physical model for complex systems can be the "automata gas" [95]: a statistical system whose elementary components perform both a physical dynamics and a cognitive behavior based on information exchanges [35,92]. In the sequel an automata gas model to describe the crowd and traffic dynamics is described. As a general convention the capital letters will denote scalar variables, whereas the lower letters will indicate vectors.

Automata Gas and Crowd Behavior

The pedestrian mobility is a basic human activity strictly linked with the detailed structure of urban environment and the social life [14,32,81]. Indeed due to the human being social nature, the pedestrian mobility will continue to play a fundamental role for the life quality of metropolis. The development of fast and comfortable transportation has reduced the spatial scale of walking activity to an average dimension of $\simeq$500 m. The direct observations on video films turns out to be the most useful tool to extract detailed information on pedestrian dynamics [50,99]. Various research group have pointed out the following items [39]:

I Pedestrians tend to walk at a desired velocity toward a local destination. The desired walking speed depends both on social characters (for example age, sex, job, ...) and on the cognitive internal state. Moreover individuals develop strategies in order to walk in the more comfortable way (for example avoiding collisions and least energy consuming), and following a best path according to their propensities and knowledge of the urban space.

II Pedestrians like to keep a certain distance from other pedestrians and borders as walls or obstacles. This distance can be interpreted as a social space preserving individuality, and it depends on various factors like crowding conditions, individual hurry and space geometry. Usually individuals don't want their social space violated by extraneous people. Moreover individuals who know each other, tend to form clusters that move as single entities performing a flocking dynamics [24,25].

III To avoid physical contact, individuals reduce the walking velocity and take local detours, but they feel aversion to move in opposite to their desired direction. In case of physical collisions, the incompressible nature of bodies play an important role and pedestrians move their body, so that they can slide on each other (if it is possible).

IV The best pedestrian paths usually are not the minimal time or space geodetic, because many other factors contribute to the choice of the trajectory, so a minimal action principle cannot model dynamics [15,54].

V The pedestrian paths seem to optimize some utility functions time or energy consuming [64], but depend on the information level of the individual on the urban space and single's propensities. They may be dynamically modified by decision mechanisms [40,46] in the presence of multiple choices, particular chronotopoi and geometrical or crowding conditions [11]. In normal crowding conditions, there are strategies automatically applied at particular points, like at a bottleneck, where the priority convention avoid criticalities. The decision mechanism reflects the individual free will, which is an unpredictable relevant character of pedestrian dynamics, but there are also some average behaviors, such as the herding effect or habit [40,49].

VI When panic phenomena occur, pedestrians lose their rationality and the dynamics are dominated by hard physical interactions [39]. People try to move considerably faster than normal and start pushing. As a consequence the pressure become extremely dangerous in jammed crowds, arching and clogging are observed at the exits so that even escape become very difficult or impossible [37].

In order to define an automata gas model for pedestrian dynamics that takes into account the previous observations, physical and information-based interactions are treated separately. Let us consider the following axioms, that define a common framework for several pedestrian dynamical models:

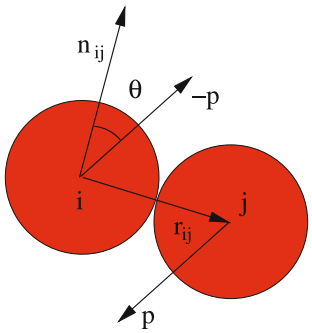

Traffic and Crowd Dynamics: The Physics of the City, Figure 1
Sketch of automata collision in the center of a mass system: $p = \mu(v_j - v_i)$ is the center of mass momentum and θ the collision angle with respect to the normal to the collision direction $r_j - r_i$

A1) each automaton i has an inertial mass m_i, an incompressible body-space of fixed radius R_b and a social space of radius R_s around its center r_i.

A2) Each automaton i has a desired velocity $\overline{v}_i$, and it continuously modifies its actual velocity v_i according to

$$\dot{v}_i = \gamma^{-1}(v_i - \overline{v}_i) \tag{1}$$

A3) When two automata i, j have distance $D_{ij} = \|r_j - r_i\| = 2R$ and their velocities satisfy the condition $(v_j - v_i) \cdot (r_j - r_i) < 0$, then a physical collision and the change of momenta $p_i = m_i v_i$ are computed according to

$$\begin{aligned} p'_i &= \alpha\left(p_i - (\mathcal{R}(\phi) - \mathcal{I})\mu(v_j - v_i)\right) \\ p'_j &= \alpha\left(p_j + (\mathcal{R}(\phi) - \mathcal{I})\mu(v_j - v_i)\right) \end{aligned} \tag{2}$$

where $\mu = m_i m_j/(m_i + m_j)$ is the reduced mass, $\alpha \leq 1$ the dissipation coefficient, and $\mathcal{R}(\phi)$ is a counterclockwise rotation of an angle $\phi > \theta$ (see Fig. 1), whereas $\mathcal{I}$ is the identity transformation.

A4) Each automaton has a local visual space given by a semicircle of radius R_v, symmetric with respect to the desired velocity direction. At a given time interval Δt, the automaton obtains information on positions and velocities of any other automaton or obstacle in its visual space.

A5) Each automaton i rotates and reduces its velocity

$$\dot{v}_i = \pm\omega \times v_i - \beta v_i \quad \beta > 0 \tag{3}$$

when it realizes that another automaton or an obstacle will enter into its social space, within a reaction time interval τ.

The axiom A1 concerns the physical properties of the automata gas: the incompressible body is usually considered by models for simulation of crowd dynamics and pedestrian panic [44]. In principle one has to consider different body radius for different automata to take into account the diversity among individuals in a crowd; this can be relevant studying the dynamics at bottlenecks [43,52]. The "inertial mass" m is also introduced to differentiate the automata behavior during collisions or interactions: an automaton with a great mass behaves more aggressively by forcing other automata to deviate from its trajectory. The existence of a social space has been proposed by sociologists who related this space with the distance between two individuals shaking hands with each other (in the West civilization). However this is not a fixed space, and it depends on the single characteristics (for example philanthropic or misanthropic one), and on crowding conditions and/or presence of attractive points in the space where people accept to be closer. One should also consider the possibility that individuals form compact groups when there is empathy among them.

The existence of a desired velocity (axiom A2) is almost obvious, but of course its determination is the result of a cognitive process where both determinism and individual free will have an important role. Eq. (1) is the most simple dynamics to simulate the automaton tendency to move at the desired velocity, and the same choice has been proposed by other models, where the only relevant parameter is γ: i. e., the "relaxation time scale" [45].

The collision dynamics defined by axiom A3 is inspired by granular flows interactions [18,38], and it allows one to simulate both the body repulsion, and a sliding effect that mimics the real effect due to the possibility of wheeling the body when two pedestrians collide. The relevant parameters are the rotation angle ϕ and the dissipation parameter $\alpha \in [0, 1]$: when $\phi = 2\theta$ (see Fig. 1) and $\alpha = 1$ one has a perfect elastic collision, whereas $\phi = \theta$ and $\alpha = \cos\theta$ gives an anelastic collision between two spheres. The direct observation of video films suggests the existence of a α dependence on θ, which has a maximal value at $\theta = \pi/2$ (i. e., two pedestrians tend to reduce suddenly their velocity in a head on collision). Conversely the choice of the ϕ within the interval $[\theta, 2\theta]$ seems to be suitable for the simulations.

The physical automata dynamics, which is the combination of Eqs. (1) and (2), can be defined in an equivalent way by using the social force model, where a second-order dynamics is used to simulate pedestrian movements with interaction "social forces", which depend both on the distance and on the reciprocal orientation of pedestrians (non-Newtonian character) [41]. The collisional approach

used in the automata gas model gives an advantage using essentially a first-order dynamics in the simulations, so that the numerical instability due to high force gradients are automatically solved and a longer integration time step can be used. Conversely the cellular automata models [3,48,70] solve the collisions problem by discretizing the space and are extremely convenient from a computational point of view, but they can introduce some bias in the microscopic dynamics whose macroscopic effect should be carefully considered [60].

The axioms A4 and A5 introduce a local vision mechanism: each automaton sees the others in front of it whose actual position is inside its visual space. For each of them, it computes the forecast position (according to the actual velocity) after a time τ. Then it selects the automata whose forecast positions are inside its social space, therefore it rotates and reduces its velocity, in order to avoid future collisions choosing the suitable sign in Eq. (3). The vision has some typical features of an information based interactions [84]:

a) It depends on the forecast positions of other automata, according to the available information;
b) It has a dichotomic nature, since only the automata that will eventually enter the social space are considered;
c) The sign choice of the rotation velocity is the result of an utility function evaluation.

Finally the interval Δt defines the decision timescale, when new information is achieved; by assumption, the following inequality holds

$$\Delta t \bar{v}_i \ll R_\mathrm{b} \ll \tau \bar{v}_i \simeq R_\mathrm{v} \,. \tag{4}$$

Therefore the automaton dynamics is the result of many decision processes, and the vision mechanism determines a collective interaction dynamics in the automata model, which is an essential feature for the appearance of emergent properties. According to the inequality (4), the average effect of local vision is to rotate the velocity in the direction where the "density" of counteracting automata is minimal, and to reduce the velocity proportionally to the surrounding automata density. From this point of view, the local vision is equivalent to a non-Newtonian repulsive long-range force with a spatial decaying scale $\simeq R_\mathrm{v}$, as proposed in the social force model. Indeed both the approaches give rise to the same macroscopical states in many cases of study. But the individual character of information-based interactions allows one to introduce the free will in the model, using the concept of cognitive state, without the artifact of external random perturbations in the dynamics. Using the idea of cognitive state to define the automata desired velocity, and the choice among multiple possibilities permits one to state two other axioms:

A6) Each automaton has a cognitive network representation of the physical surrounding space, dependent on its knowledge level. Each network node is related to a possible decision, whereas the arcs introduce connections between successive decisions (see for example Fig. 2a).
A7) The decision mechanism is defined by a subjective probability evaluation [27] based on the existence of an utility function and an internal cognitive dynamics [16,28]; the probability are computed taking into account the available information and the previous history.

The axiom A6 implies an internal dynamical structure for the particle automaton, which acquires a knowledge capability of urban space. The nodes are identified with particular regions of interest in the urban space (for example in Fig. 2 the nodes correspond to doors neighborhoods). The decision process is discontinuous, since the automaton gets new information at the node points. One can compare the cognitive network structure with a best paths origin-destination algorithm, based on the deterministic minimal action principle, which has been proposed to study the accessibility of urban spaces [15,54]. However, this approach gives a global and rigid description of the space that turns out to be unrealistic and, sometimes, useless, when one studies crowd dynamics where the automata paths are the results of a continuously evolving process, dominated by collective interactions and individual free-will. On the contrary, the space representation using a cognitive network that focuses the attention only on some particular points, it is both robust and flexible to allow a single automaton to creat its own individual path with optimization criteria based on local information.

The axiom A7 establishes a relation between an internal cognitive dynamics [93] and a probabilistic choice among different possibilities [62]. Let us consider the elementary dichotomic decision between the choices A and B, by associating A or B to the values of the variable x_i (for example A when $x_i > 0$) that defines the "cognitive state" of the i automaton. One assumes the existence of a stochastic dynamical system that represents the neuronal activities [83]

$$\dot{x}_i = \frac{\partial U}{\partial x}(x_i; I) + \sqrt{2T_i}\xi_t \tag{5}$$

where $U(x; I)$ is an utility function with two local maxima U_A and U_B. The "utility concept", introduced in econ-

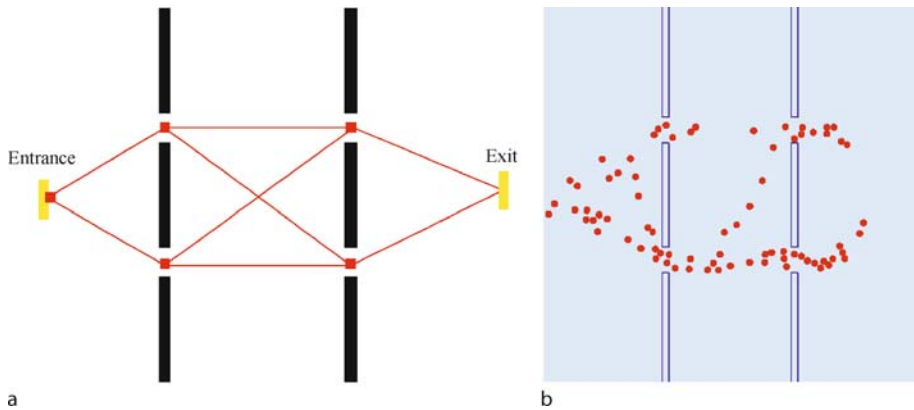

Traffic and Crowd Dynamics: The Physics of the City, Figure 2
a Cognitive network representation of a space with two parallel walls and four connecting doors; the automata enter from the left side and have to reach the exit on the right. The squares are the decision node where automata obtain new information and change the desired velocity. b Snapshot of the automata dynamics in the space shown in left picture. The cognitive network causes a flux organization along the shortest paths, but the subjective probabilistic decision implies the existence of "irrational" automata which follow longer paths

omy [31,96], is a measure of the perceived advantages of a certain decision with respect to the automaton final purpose, whereas the information I is any external variable that may influence the decision. Under the very general regularity hypotheses on U, the dynamics (5) evolves towards one of the local maxima in a stochastic way, due to the presence of the white noise ξ_t (unpredictable factors in the decision mechanism) and the "social temperature" T_{i}, which is a measure of the "irrational behavior" of the automaton. As a consequence even if all the automata have the same decision mechanism, the effective decision is subjective due to the different information I and the different temperature T_i. According to the Arrhenius' law, the transition probabilities from the choice A to B and vice versa read

$$P_{AB} \propto \exp\left(-\frac{U_A(I)}{T_i}\right) \quad P_{BA} \propto \exp\left(-\frac{U_B(I)}{T_i}\right) \tag{6}$$

By introducing an average balance equation for the stationary probabilities to choose A or B

$$P_A P_{AB} - P_B P_{BA} = 0$$

one obtains [16,46]

$$\begin{aligned} P_A &= \frac{\exp\left(\frac{U_A}{T_i}\right)}{\exp\left(\frac{U_A}{T_i}\right) + \exp\left(\frac{U_B}{T_i}\right)} \\ P_B &= \frac{\exp\left(\frac{U_B}{T_i}\right)}{\exp\left(\frac{U_A}{T_i}\right) + \exp\left(\frac{U_B}{T_i}\right)} \end{aligned} \tag{7}$$

The non-linear relation between utility and probability has two limit behaviors: for low temperature values, Eq. (7) defines a sigmoidal like function with a sharp transition between the two choices, whereas for high temperature values the relation is almost linear. These two regimes correspond to two different social behaviors of the automaton: in the first case there is a sudden change in the decision as utility overcomes a certain threshold (rational automaton), in the second case, the automaton choice becomes insensitive to utility changes (irrational behavior).

Automata Gas and Traffic Behavior

The traffic behavior in a urban network is the consequence of citizens mobility demand driven by the chronotopic activity of modern cities [21]. The application of physics to traffic modeling dates back more than 40 years [80]. From one hand the microscopic traffic dynamics is forced on a one-dimensional space (the road), from the other hand its spatial scale is extended to the whole city. For this reason various researches have proposed to simplify the microscopic physical dynamics, and to focus the attention on the cognitive behavior at the base of the individual mobility agenda and on the fluxes optimization in the road network [2,42,86,94]. Nevertheless the microscopic drivers behavior is responsible of local congestions formation that may propagate in the road network and cause a macroscopic phase transition of the whole system. The proposed models are divided in two main classes: the cellular automata models which discretize the road dynamics using discrete stochastic maps [63,75], and the continu-

ous models based on stochastic differential equations [39]. The auto-mobilis model [12] is based on a simple microscopic continuous dynamics, but able to describe different driving styles and behaviors at the crossing points (traffic lights, roundabouts and stop signals at insertions). The automata in the model move on a road using alternatively the two main behaviors that has been proposed by various researchers [51]: the so called "car-following" [4] and the "optimal velocity" behavior [74]. In the first case each driver adjusts his velocity to the velocity of the successive vehicle; in the second case each driver chooses his velocity according to the safety distance with the successive vehicle. Both models can be described by the differential equation

$$\begin{aligned} \dot{s}_j &= v_j \\ \dot{v}_j &= a(s_{j-1} - s_j, v_j, v_{j-1}(t-\tau); \Theta) \end{aligned} \tag{8}$$

where (s_j, v_j) are the position and the velocity of the j-automaton on a road (by convention the j-driver follows the $j-1$-driver), Θ is the parameter set that characterizes the individual behavior, τ is a reaction time to the velocity changes of the successive vehicle. The effect of line changes can be neglected since they are not frequent along urban street (of course this is not true at the intersections points that will be consider separately). To specify the acceleration in Eq. (8), the main ideas at the base of various models are summarized, by means of the following assumptions:

B1) An automaton-driver has a finite dimension d_0, it keeps a minimal distance $d_{\min}$ from the preceding automaton, it has a reaction time τ to the velocity changes of the other vehicles, and it has a desired velocity $v_{\max}$ (usually the maximal allowed velocity along the considered road, assuming that people respect the traffic rules)

$$\dot{v}_j = \alpha(v_j - v_{\max}) \tag{9}$$

where α is the acceleration capacity in a comfortable driving.
By measuring s_j at the center position, the distance between two successive drivers is $\Delta s_j = s_{j-1} - s_j - d_0$.

B2) Each automaton has a safety distance D_s and an attention distance D_{att} (i. e., a distance at which the driver considers the presence of the other drivers)

$$D_s = d_{\min} + 2\tau v_j \quad D_{att} = d_{\min} + 2.5\tau v_{\max} \tag{10}$$

Each time $\Delta s_j \le D_s$ it reduces the velocity according to

$$\dot{v}_j = -\beta(v_j - v_s) \tag{11}$$

where β is the breaking capacity.
Eq. (10) establishes a linear relation between reaction time τ and the safety distance D_s; this is consistent with direct observations for the typical urban velocity ranges. But to avoid accidents one requires a condition on the breaking capacity $\beta\tau \ge 1/2$ (cf. Eq. (11)).

B3) When the distance Δs_j is less than the attention distance D_{att}, the automaton adjusts its velocity to the velocity of the preceding vehicle

$$\dot{v}_j = -\alpha(v_j - v_{j-1}(t-\tau) - \gamma) \tag{12}$$

where the parameter $\gamma \ge 0$ simulates the behavior of impatient drivers.

To be more realistic, one takes also into account that the automata cannot correctly evaluate small velocities so that when $v_{j-1} \le v_{\min}$ one set $v_{j-1} = 0$ in Eq. (12).

The Eqs. (9), (11), (12) define the microscopic dynamics along a road. All the parameters are individual characters, and have a random distribution among the automata populations. Indeed the driver behavior can be correlated to citizens social characters (sex, age, birth place etc.), so that several classes of drivers can be identified. In some realistic situations the presence of different driver classes is crucial to understand urban traffic. The model can simulate different traffic regimes: at low density the free flow (Eq. (9)) is described, when the density increases the automata perform a synchronized flow, which becomes unstable when the local density overcome $1/D_s$ and gives rise to congestion effects. Indeed the model has infinite equilibria with automata moving at the same velocity $v \le v_s$. However a linear stability analysis using the Fourier transform shows that the dynamics (12) is always stable with a relaxation time

$$T \simeq \frac{2(1+\alpha\tau)^3}{\alpha(2\pi/\rho\lambda)^2} \tag{13}$$

in the limit $\rho\lambda \ll 1$, where ρ is the density and λ the wavelength of the perturbation, whereas the dynamics (11) become unstable when $\beta\tau < 1$ [37,69]. Therefore when the local density satisfies $\rho \le 1/D_s$, a congestion regime may rise; the instability can be driven by small percentage of "bad drivers" in the automata population, as shown in Fig. 3. This reflects the possibility that the different driving styles may cause local instability due to microscopic interactions, as suggested by direct observations [91].

To complete the dynamical model for urban traffic, one has to set the behavior at the crossing points in different situations assuming:

B4) At a traffic light in case of yellow colors, the automata compute the time t_c required to pass through the

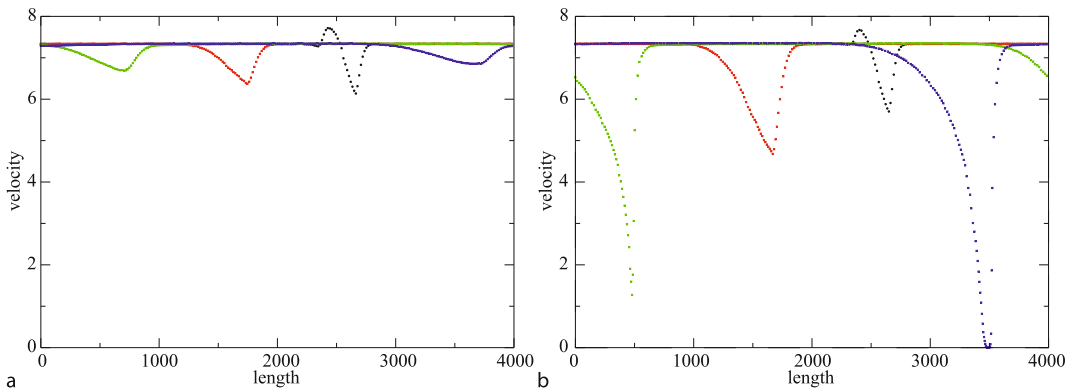

Traffic and Crowd Dynamics: The Physics of the City, Figure 3
a Stability study of the urban traffic micro-dynamics for an homogeneous automata population in a periodic road: the four curves (in the order: *black*, *red*, *green blues*) give the vehicle velocity (m/s) as a function of the position (m) on the road at different time separated by 5 minutes. The model uses "realistic" parameters: $d_0 = 4$ m, $d_{\min} = 1$ m, $v_{\max} = 16$ m/s, $v_{\min} = 1$ m/s, $\beta = 1.1\,\mathrm{s}^{-1}$, $\tau = .75$ s and $\alpha .5\,\mathrm{s}^{-1}$. The initial equilibrium density $1/16\,\mathrm{m}^{-1}$ is locally increased by a 10%. b The same as picture a, but the simulation considers a 20% of "bad drivers" randomly distributed on the road with reduced breaking capability $\beta = .9\,\mathrm{s}^{-1}$; the blue curve shows as a stop-and-go instability has been developed by the microscopic dynamics

cross at the actual velocity: if t_c is longer than the yellow time-interval, they stop.

B5) At a connection of two streets with priority rules, the vehicles in the secondary street are forced to reduce their velocity at a given value v_c ($v_c = 0$ in case of a stop signal); they compute both the distance D from the crossing point and the velocity v of the incoming vehicle in the main street, if the distance D is greater than $(v - v_c)/\beta$, the other vehicles can enter in the main street; otherwise it has to stop. In Fig. 4 a snapshot of a simulation in a crossing point with priority rules is shown. As a matter of fact, the emergent global properties of urban traffic cannot be understood, described, and eventually predicted, only using the automata micro-dynamics on the road network [86,94]. A global cognitive time-dependent field interacting with individuals is needed.

B6) In the urban space-time there exist the "chronotopoi", which are adaptive agents generating the individual mobility demand.
Examples of chronotopoi are historical centers, shopping centers, university campus, great social events as concerts or football matches etc.; in our sense a chronotopos is not a geometrical or morphological property, but an entity that introduces timing correlations between the city macroscopic level (social time) and the automata microscopic dynamics (individual time). The chronotopoi provide a cognitive map [79] to the automata by means of cognitive fields. A sketch of chronotopic cognitive fields overlapped to a Manhattan-like road network is plotted in Fig. 5, where the intensity of the attraction is measured by the well amplitudes.

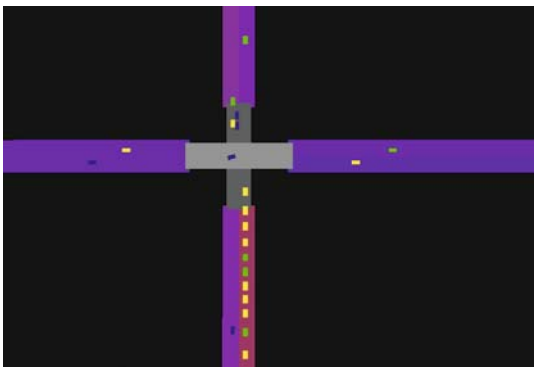

Traffic and Crowd Dynamics: The Physics of the City, Figure 4
Snapshot of vehicles dynamics simulation in a crossing point; the road in a lighter color at the cross has the priority. The different colors are related to the different vehicle directions

B7) Every automaton constructs a daily mobility agenda based on its propensities and the available chronotopic information.
The decision mechanism is modeled by de Finetti subjective probability [27], taking free will into account.

B8) The automaton trajectory is a path realization on a network, where the choice probabilities at the nodes depend on the chronotopic attraction.
The decision mechanism at a crossing point is outlined in Fig. 6, where the probabilities p_j depend on

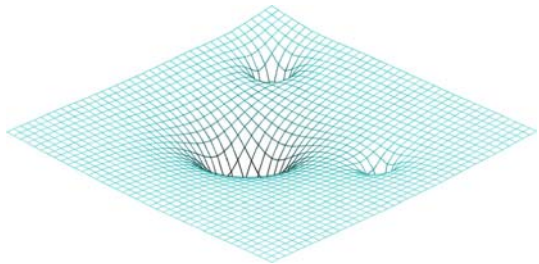

Traffic and Crowd Dynamics: The Physics of the City, Figure 5
Landscape of a abstract cognitive field superimposed to a road network

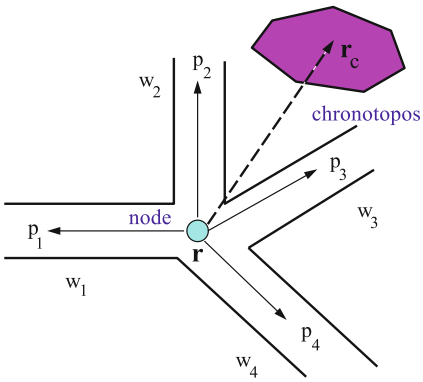

Traffic and Crowd Dynamics: The Physics of the City, Figure 6
Scheme of the automata decision mechanism at a crossing point: the choice probabilities P_j are moduled according to the chronotopic attraction

a best path algorithms to reach the chronotopos, and the weights w_j measure the street quality (typically accessibility, security and aestetics). The final probability choice is defined by

$$P_k = \frac{p_k w_k}{\sum_j p_j w_j} \,. \tag{14}$$

Simulation Results and Empirical Observations

In complexity sciences, the simulations and the empirical observations play the role respectively of theory and laboratory experiments in the traditional physics framework. The main focus is to observe the emergent properties of the system, and to understand the class of microscopic interactions that can explain such properties in order to detect the control parameters. Simulations and empirical observations have to be compared in this optic [9]. In the physics of the city, the real urban systems are the natural laboratories to observe and study crowd and traffic flow.

Crowd Behavior in Urban Spaces

An occasion to illustrate the crowd behavior in a whole pedestrian city, is Venice's Carnival, since about 100,000 tourists visit Venice each day during the last week. The observations are mainly based on video film analysis of particular regions of interest, like San Marco square and Riva degli Schiavoni. In principle the videos allow one to follow the single individual dynamics; however, the dimension of the considered regions is limited and a computer assisted analysis is difficult in crowded situations. Some quantitative information can be obtained by using software following the individual trajectories. The trajectories permit one to extract information both on the microscopic pedestrian dynamics, and on the individual cognitive behavior even if the noise level is still a serious problem [50]. Using a sample of $\simeq$600 trajectories in a moderately crowed area, we have computed for each individual the density of surroundings coming from the opposite direction and moving in a transverse direction within a distance of $\simeq$2 m. Then, after averaging all the local densities, the results are plotted in the Fig. 7. The observations indicate the existence of a low density area along the desired velocity direction, only if one focuses the attention on individuals moving in a counteracting way; conversely the individuals that move in other directions seems not to be considered. This is the sign for the existence of strategies to avoid collisions and to keep a certain area free along the direction of motion, based on a local vision mechanism. The automata gas dynamics can be illustrated performing some virtual experiments, where the emergent properties can be analyzed. The model parameters (cf. Eqs. (1), (2), (3)), have been fixed using R_b as space unit and measuring time in seconds: the desired velocity is set $\|\overline{v}\| = 2R_b$ with a spread of 10% among the automata and the relaxation time $\gamma = 1$, whereas the collision (2) is defined by $\phi = \theta + \pi/10$ (see Fig. 1) and $\alpha = .9$; finally the local vision has a radius $R_v = 2\overline{v}$, the social space is $2R_b$ and Eq. (3) uses $\omega = 3\pi$ and $\beta = .5$. Moreover all the automata have the same mass.

Let us consider two automata populations moving along opposite directions in a open space (for example, a square). The automata enter in the space from two sources located at the borders: the red ones move from the left to the right, whereas the yellow ones move from right to left. This system has two equilibrium regimes: the entropic regime where both the populations are uniformly spread in the space, and the organized regime where automata move along separated lines [47,89,98]. In the simulations, at low densities the automata relax to the entropic regime, whereas when the density increases one observes

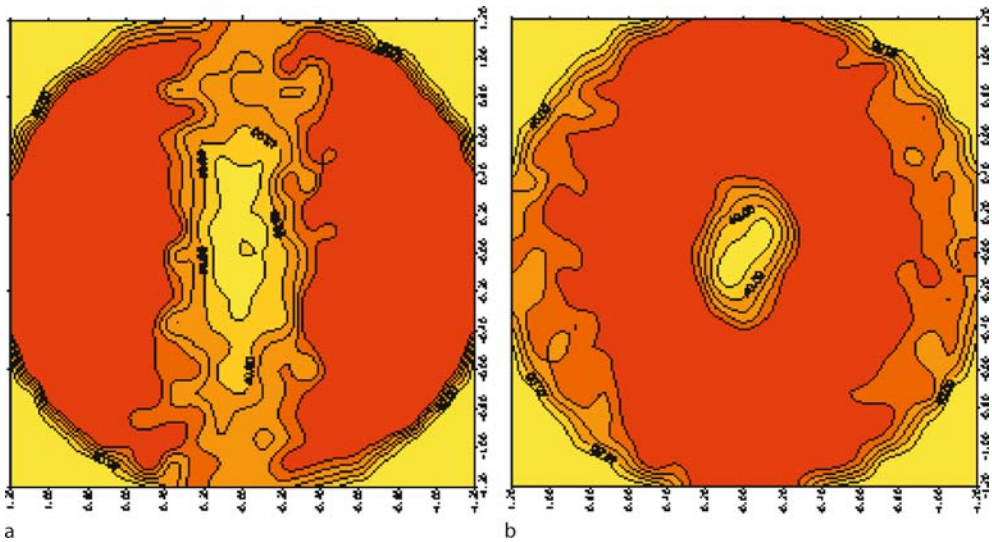

Traffic and Crowd Dynamics: The Physics of the City, Figure 7

The scale color from *red* to *yellow* defines the average local density around a pedestrian that move in the vertical up direction as measured from the film video data during the Venice's Carnival. The pedestrians were moving in San Marco's Square in a moderately crowded condition. The considered radius around the pedestrian is 2 m. The picture a refers to the density of people moving into the opposite direction with respect to the considered individual. The picture b refers to the density of people moving into a transverse direction

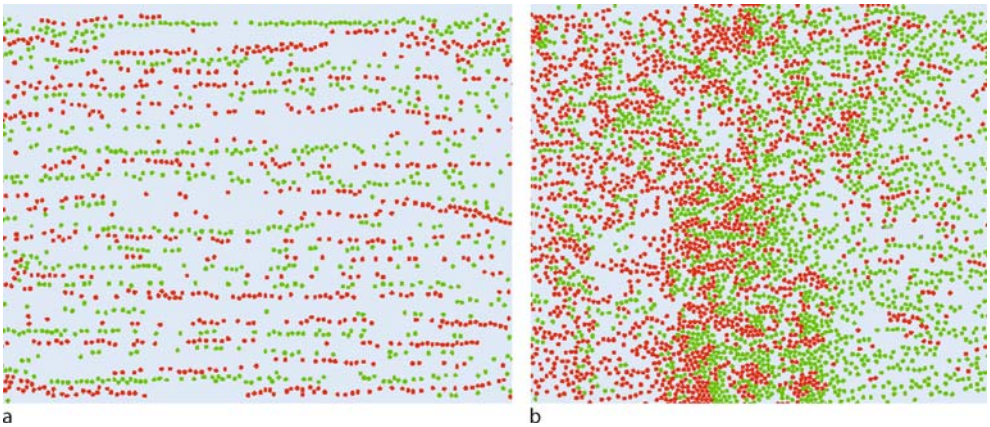

Traffic and Crowd Dynamics: The Physics of the City, Figure 8

a Simulation of two automata populations moving into opposite direction in a square. The parameters used in the model are (cf. previous section): $R_b = 1$ (the space unit), $\|\bar{v}\| = 2$, $\gamma = 1$, $\phi = \theta + \pi/10$, $R_v = 4$, $\omega = 3\pi$ and $\beta = .5$. The square dimensions are 120 × 120 and the automata enter from two sources located at the left border (*red automata*) and the right border (*yellow automata*). The appearance of an organized regime with automata moving along lines is clear. b The same as the picture a, but the incoming flux from the sources has been increased by a factor 2. The lines of people start to interact and a turbulent regime appears

the lines formation in the space as shown in Fig. 8a. The automata form narrow straight non-interacting lines, separated by empty spaces: sometimes the lines are continuous, in other cases they interrupt or join together. Moreover there are always empty regions, so that the space distribution is not uniform. This phenomenon can be interpreted by a mean field argument: every time two automata meet, they deviate to avoid a collision and then they recover the desired velocity according to Eq. (1). The final result is a transverse displacement with respect to the desired velocity for both the automata. As far as binary interactions are considered, the transverse displacements

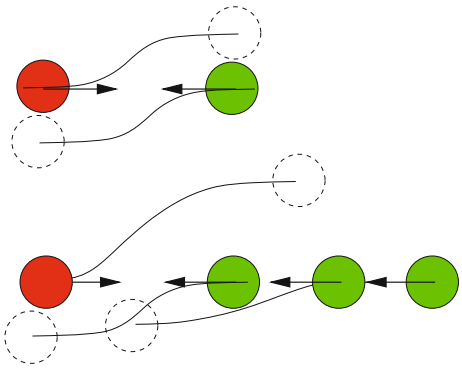

Traffic and Crowd Dynamics: The Physics of the City, Figure 9
Sketch of automata interaction: the difference between the binary interaction dynamics and a collective interaction dynamics is outlined

Traffic and Crowd Dynamics: The Physics of the City, Figure 10
Snapshot of a video film of San Marco's square during the Venetian Carnival 2007: the video analysis shows the existence of lines of people moving in opposite directions

are symmetric and only the entropic regime is stable. But the local vision introduces long-range collective interactions, and aligned automata moving in the same direction behave as a simple entity [57]: in other words when an automaton meets a line of people moving in the opposite direction, it performs a larger transverse displacement than the automata in the line as sketched in Fig. 9. Therefore the line structure becomes stable and the system relaxes towards an organized regime. This behavior has been observed by various research groups in real situations [47,55,65]: an accelerated video film in the crowding conditions shown in Fig. 10, points out clearly the existence of stable lines of people moving in opposite directions, and a non-homogenous distribution of pedestrians in the space. The lines survive also when the density increases, due probably to the slow down of the pedestrian desired velocity and the reduction of the social space. The simulations confirm this parametric dependence, and they also point out as to maintain the organized regime stability (for a fixed desired velocity and social space) it is essential to have some space among the automata lines, otherwise the interaction among the lines themselves will destroy the laminar character of the dynamics, and a new turbulent regime appears [45,71,72]. This implies the existence of a critical density value as it is shown in Fig. 8, where the simulation are repeated with a twice flux from the external sources. This turbulent regime has been observed in critical situations (like crowd stampede) when, under effect of panic, individuals behave irrationally, the velocity is increasing and the dynamics is mainly physical [50,59]. This was not the case during the Venetian Carnival and in the absence of panic, pedestrians take strategies that prevent critical situations. The automata gas model allows the study of the dynamics at a bottleneck, reducing the desired velocity when the density overcomes $1/R_b^2$ (i. e., when automata have to share the social space). A bottleneck is a narrow passage at which pedestrian interactions do not allow the laminar flow formation. The dynamics at bottlenecks is a crucial task for understanding the crowd behavior in urban spaces, and it has been studied by various researchers [43,52,88]. The simulation of the passage of two counteracting flows at a bottleneck is shown in Fig. 11a: the automata reduce the desired velocity when they see a counteracting flux inside the bottleneck area. As a consequence there appears a self-organized oscillatory regime for the flux across the bottleneck (see Fig. 11b), as it has been observed in other models and in laboratory experiments [66]. A mean field theory approach suggests a possible interpretation of the oscillating fluxes regime. Let n^r and n^g the density of automata moving in opposite directions in the bottleneck area; each automaton can be in two states: waiting state P_w and crossing state P_c. The balance equations according to the following transition probabilities read

$$\begin{aligned} P^r_{c,w} &\propto n^g P^g_c \left(1 + c\left(n^r P^r_c - n^g P^g_c\right)\right) \\ P^r_{w,c} &\propto \left(1 - c\left(n^r P^r_c - n^g P^g_c\right)\right) \end{aligned} \tag{15}$$

based on the hypothesis that the probability to cross the bottleneck in the positive direction (red automata) increases if the flux $n^r P^r_c - n^g P^g_c$ in the bottleneck is positive, and, on the contrary, the probability of returning in the waiting state increases when the flux is negative. Moreover if the flux is positive, the density n^r is decreased by a quantity $\Delta n(n^r P^r_c - n^g P^g_c))$, due to people passing

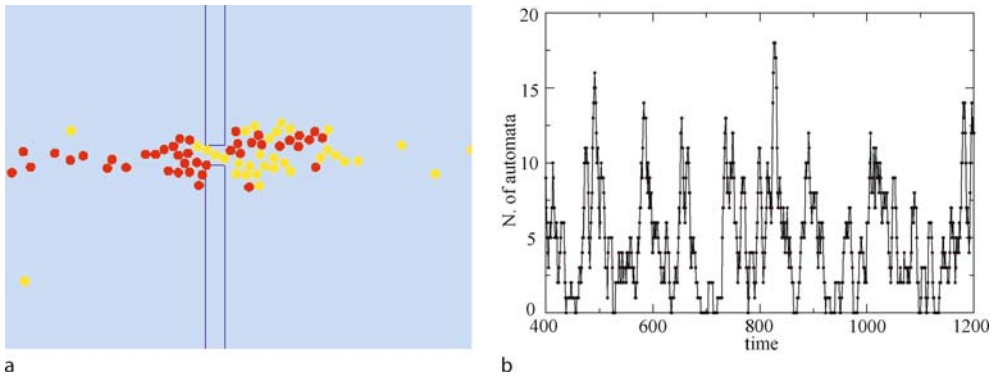

Traffic and Crowd Dynamics: The Physics of the City, Figure 11
a Dynamics of two counteracting flows at a bottleneck in automata gas simulation; the transverse bottleneck dimension is 2 time R_b with a constant incoming flux of one automaton each three time units. b Number of "*red*" automata that reach the right destination after crossing the bottleneck versus time: it is clear the appearance of an oscillating regime at the bottleneck

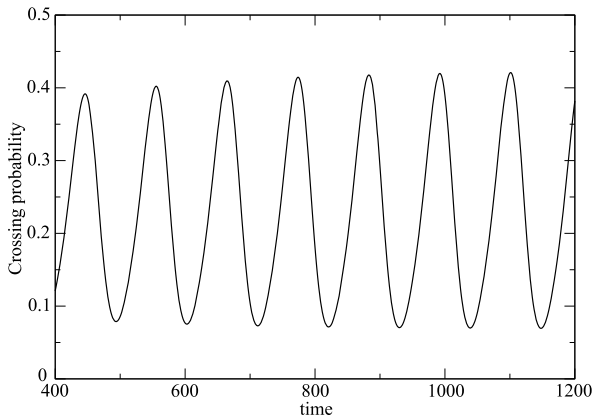

Traffic and Crowd Dynamics: The Physics of the City, Figure 12
Solution of the balance Eqs. (17) showing an oscillating behavior for the crossing probabilities at a bottleneck under the hypotheses (15); the parameter *c* is the control parameter and the oscillations disappear below a critical value

Traffic and Crowd Dynamics: The Physics of the City, Figure 13
Snapshot of a video film of high density crowd crossing a narrow bridge during the Carnival; there is a long queue in front of the bridge whereas the flux across the bridge is constant

trough the bottleneck, where Δn is a proportionality constant depending on the bottleneck geometry

$$\dot{n}^r = \Delta n(n^r P_c^r - n^g P_c^g) \quad \text{if} \quad (n^r P_c^r - n^g P_c^g) > 0 \quad (16)$$

The balance equations (analogous equations hold for the yellow automata) read

$$\begin{aligned} \dot{P}_w^r &= -P_{w,c}^r P_w^r + P_{c,w}^r P_c^r \\ \dot{P}_c^r &= P_{w,c}^r P_w^r - P_{c,w}^r P_c^r \end{aligned} \quad (17)$$

and the solution is plotted in Fig. 12 for a suitable set of parameters (the time unit is arbitrary), where the existence of a stationary oscillation regime in the crossing probability is evident. The parameter c (cf. Eq. (15)), which weights the flux effect on the transition probabilities, is a control parameter for the dynamics: if c is below a critical value, no oscillation is observed and the stationary solution is constant. Finally another interesting situation is illustrated by Fig. 13, where people had to cross a narrow bridge during the Carnival: in the movie the people stop nearby the bridge forming a funnel-like queue, whereas the flux across the bridge is constant. As a result, in the queue the people move in a discontinuous way (stop and go) using the empty spaces left by individuals that enter the bridge; in such a situation there are waves back-propagating along the queue [50]. The measure of the flux across a given area can be performed by simulating the situation plotted in

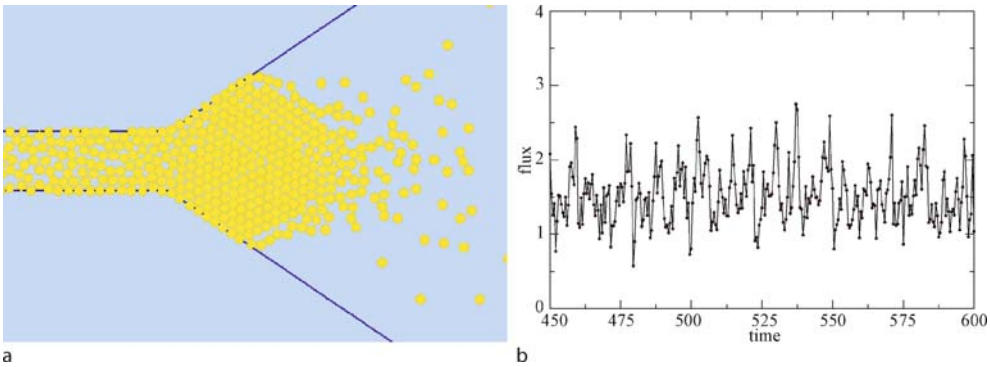

Traffic and Crowd Dynamics: The Physics of the City, Figure 14
a Scheme of the bridge crossing geometry used in the gas automata simulation. b Automata flux in across small area inside the queue (≃10 automata are considered simultaneously): fast periodic oscillations are detected due to short length back-propagating velocity waves

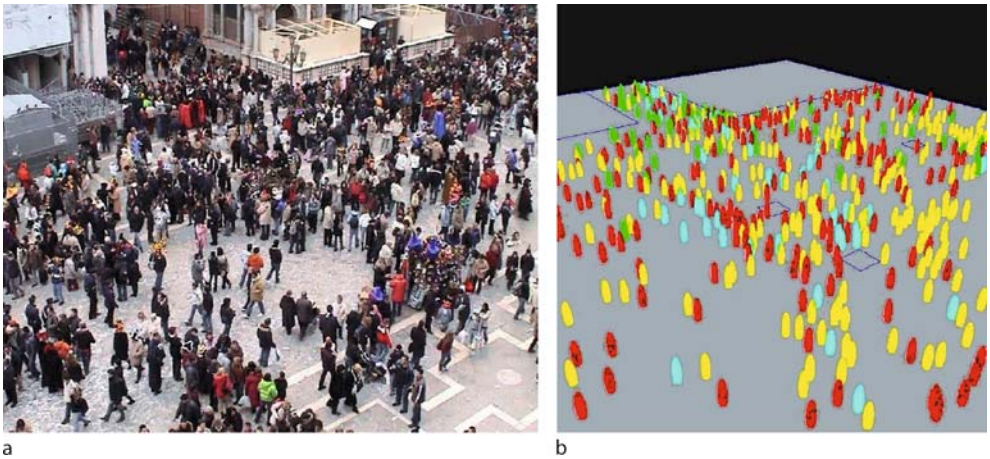

Traffic and Crowd Dynamics: The Physics of the City, Figure 15
a Mobility at the entrance of San Marco's square (snapshot from video movie). b Automata virtual experiment to study the mobility in a virtual space reproducing the urban space of the left picture

Fig. 14a; the results (see Fig. 14b) show a periodic structure that can be interpreted as short length back-propagating velocity waves. The finite automaton dimension has an essential role in the queue dynamics. As discussed previously, a quantitative comparison between empirical observations and virtual simulations have to face the intrinsic difficulty of reproducing the complexity of crowd behavior. However current computers can be used as virtual simulators that allow to study the crowd behavior in various realistic situations, becoming a helpful instrument for mobility planning [56]. Fig. 15 shows an example of crowd behavior in a real environment and in a virtual experiment by using the "Campus software" developed at the Physics of the City Laboratory of Bologna [99].

Cognitive Behavior and Experimental Laws in Traffic Dynamics

As discussed in the previous sections, the cognitive behavior of individuals due to information-based interactions, plays a fundamental role to determine the emergent properties of urban traffic and crowd dynamics [100]. A crucial theoretical issue is to understand when it is possible to apply a statistical physics approach to derive the evolution of macroscopic variables from the single automata behavior [23]. The main difficulty is that from a cognitive point of view, the automata are all "different particles", and the information-based interaction introduces long-range correlation among the automata dynamics, so that it is

very hard to justify a mean field approach. To illustrate a heuristic mean field approach, one can use the simple example of a dichotomic choice A and B for an automata gas, where two counteracting populations (red and yellow) share the same spatial environment [40,62]. An automaton i has an utility function $U(N^r, N^y)$ in the decision mechanisms, depending on the number of automata of the two populations that have made the same choice. According to Eq. (7), the probability of the choice A is given by

$$P_A^{i,r} = \frac{1}{1+\exp\left[\left(U(N_B^r, N_B^y) - U(N_A^r, N_A^y)\right)/T_i\right]} \quad \text{if } i \text{ is red}$$
$$P_A^{i,y} = \frac{1}{1+\exp\left[\left(U(N_B^y, N_B^r) - U(N_A^r, N_A^y)\right)/T_i\right]} \quad \text{if } i \text{ is yellow}\,. \tag{18}$$

By using a statistical approach, one set $N_A^r = N^r\langle P_A^i\rangle_r$ and $N_A^y = N^y\langle P_A^i\rangle_y$ where $\langle\ \ \rangle_{r,y}$ indicates the the average value on the population. Then the Eqs. (18) give a self-consistent system

$$N_A^r = \sum_i^r \frac{1}{1+\exp\left[\left(U(N^r - N_A^r, N^y - N_A^y) - U(N_A^r, N_A^y)\right)/T_i\right]}$$
$$N_A^y = \sum_i^y \frac{1}{1+\exp\left[\left(U(N^y - N_A^y, N^r - N_A^r) - U(N_A^r, N_A^y)\right)/T_i\right]} \tag{19}$$

where $\sum^{r,y}$ indicates that the sum consider only the red or yellow population. In a symmetric condition between the two populations, the solutions of (19) have the property $N_A^r = N_B^y$; therefore one reduces to the single equation

$$N_A = \sum_i \frac{1}{1+\exp\left[(U(N-N_A, N_A) - U(N_A, N-N_A))/T_i\right]}\,. \tag{20}$$

The obvious solution $N_A = N_B = N/2$ exists, but if the following condition holds

$$\frac{1}{4\overline{T}}\left.\frac{\partial V}{\partial N_A}\right|_{N_A=N/2}(P) > 1 \tag{21}$$

where $V(N_A) = U(N-N_A, N_A) - U(N_A, N-N_A)$ and $\overline{T}$ is the harmonic mean social temperature of the automata population, the symmetry is broken and two new solutions appear where the two populations make opposite choices. Eq. (21) defines a critical temperature $\overline{T}_*$ under which the phase transition occurs. This phenomenon allows one to simulate the herding cooperative effect, observed in crowd dynamics. In the Fig. 16, Eq. (20) is solved graphically for an utility function $U(P) = P$. The simulation results show that the statistical self-consistent approach (20), allowing to compute the correct value of critical temperature $\overline{T}_*$. In the simulation the automata use a utility function

$$U = \left(M + \frac{\phi\cdot\hat{v}}{N+1}\right)\,, \tag{22}$$

where M is a variable depending on the previous choice ($M = 0$ if there is not a previous choice), $\phi\cdot\hat{v}$ the flux at the door in the direction $\hat{v}$ of the desired velocity, and N is the number of automata in a neighborhood of the door, and the social temperature is uniformly spread in the same interval for both the population. The memory M is introduced to enforce the choice of two successive doors along the same horizontal line. It is remarkable that there are always automata that move according to a minority choice, simulating the existence of free-will. The effect of a finite population size gives a critical value for N, that has to be overcome for the formation of a self-organized dynamics.

An experimental observation of statistical laws in crowd dynamics is very difficult due to the limited space extension of the video movie, whereas most of the data for traffic dynamics are on the road fluxes and not on the individual vehicles trajectories [53,58]. Recently a GPS system has been set up on $\simeq$1% of the vehicle population in Italy for insurance reasons, measuring position, velocity and signal quality. The data registration starts every time the engine is switched on, and ends when the engine is switched off. The GPS system repetition rate is one second, but the data are recorded each time a vehicle has covered $\simeq$2 km. The accuracy of GPS data has a standard uncertainty in the position of $\simeq$10 m. The GPS data allow a reconstruction of the individual dynamics so that it is possible to look for the existence of statistical macroscopic laws [82]. Different size cities have been considered: a small town, (Senigallia, $\simeq 5\times 10^4$ inhabitants, near the Adriatic see), a medium city (Bologna, $\simeq 4\times 10^5$ inhabitants) and a metropolis (Rome, $\simeq 3\times 10^6$ inhabitants). The three cities have a very different topology: the Senigallia topology is substantially a strip of $2\times 12\,\text{km}^2$ bounded by the sea, with few accesses, whereas Bologna has a radial configuration around an historical center with a radius of $\simeq$6 km. The urban area of Rome is one order of magnitude larger. To understand the mobility demand [16,29], the distribution of inner vehicle trajectories has been stud-

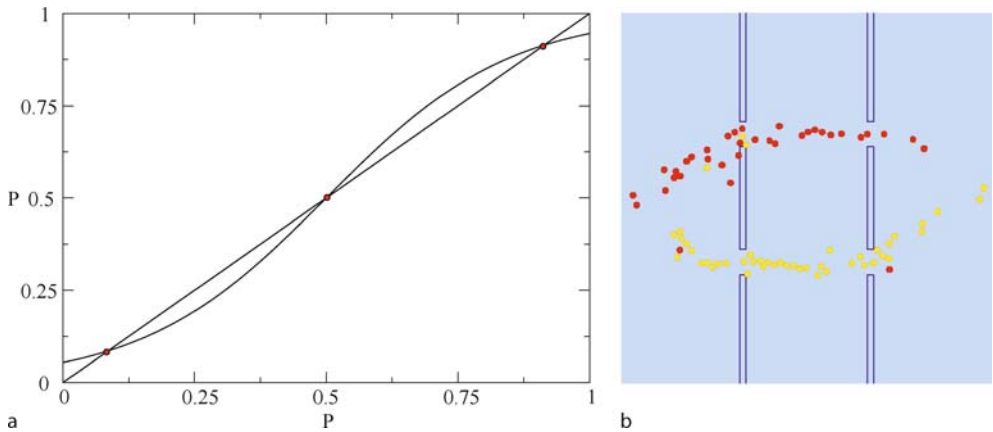

Traffic and Crowd Dynamics: The Physics of the City, Figure 16
a Graphical solution of the self-consistent Eq. (20) using an utility function $U = N_A/N$ for both choices and a temperature below the critical threshold 0.5. The solutions are indicated by red dots; beyond the trivial solution $P = N_A/N = 1/2$, two further solutions exist that breaks the system symmetry. b Self-organized dynamics by simulating two populations moving into opposite direction in the space sketched in Fig. 2

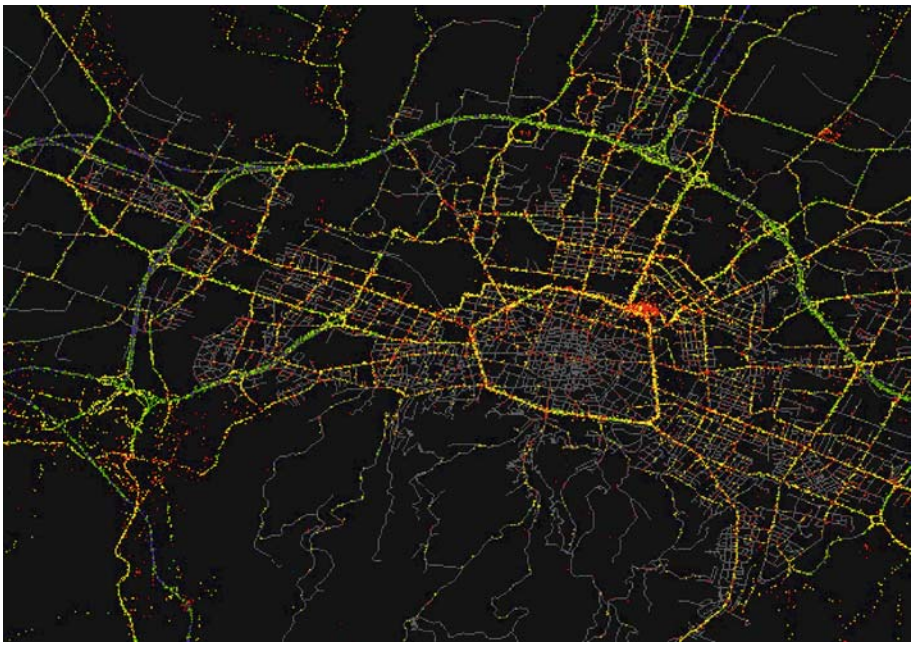

Traffic and Crowd Dynamics: The Physics of the City, Figure 17
The dots are the GPS recorded data in the Bologna urban area during the month of June 2006; the *dots* correspond to more than 1.5×10^6 data. The colors label the recorded velocity: *red* means $v < 30$ km/h, *yellow* $30 < v < 60$ km/h, *green* $60 < v < 90$ and *blue* $90 < v$ km/h

ied: i. e., the trajectories fully contained in the considered urban area. The area of the considered regions are 130 km^2 for Senigallia, 160 km^2 for Bologna and 400 km^2 (the area containing the "Grande Raccordo Anulare") for Rome. In Fig. 17 the plot the recorded data in the Bologna urban area for the whole month June 2006 is shown. In Fig. 18 the trajectories length distribution $\rho(s)$ is plotted for one month of data recording: the lengths are computed using the polygonal metric in the three considered cases: Rome (February 2007), Bologna (June 2006) and Senigallia (June

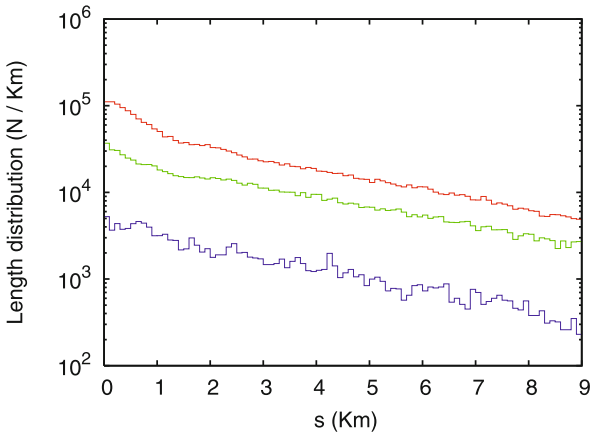

Traffic and Crowd Dynamics: The Physics of the City, Figure 18
Monthly length distribution of inner trajectories for the city of Rome (*top*) during February 2007, and for the cities of Bologna (*center*) and Senigallia (*bottom*) during June 2006. The vertical scale is logarithmic, and a polygonal metric for the lengths is used. The total number of trajectories is 1,176,763 for Rome, 79,154 for Bologna and 13,838 for Senigallia

2006). Even if there are almost two orders of magnitude between Senigallia and Rome's statistical samples, the path lengths distribution has the same behavior in the three cases. Moreover, it is really remarkable that the same exponential decay for the "long distance" trajectories is detected in all the considered cases, indicating a possible universal character at least in the class of Italian cities [82]. The exponential distribution law for individual trajectories has a statistical mechanics interpretation as the most probable distribution if each citizen has an independent mobility demand and there exists an average length for the urban trajectories. This means that the system behaves statistically as the citizens choose among randomly distributed destinations in the urban space, according to an "utility function" (proximity function) that increases linearly as a function of the trip length [16,29]. Let $\{l_k\}$ be the set of all possible lengths of origin-destination trajectories in the city, and let us suppose that N citizens choose randomly the trajectories, with the constraint that the total covered length is finite

$$L = \sum_k n_k l_k \,, \quad \sum_k n_k = N \,, \tag{23}$$

where n_k is the number of citizens choosing a trajectory of length l_k. A simple combinatorial argument show that the statistical weight of the distribution $\{n_k\}$ is

$$P(\{n_k\}) \propto \frac{1}{n_1! \cdot \dots \cdot n_k! \cdot \dots} \,. \tag{24}$$

In the limit $n_k \gg 1$ using the Stirling's approximation, one obtains

$$\log P(\{n_k\}) = -\sum_k (n_k \log n_k - n_k) + \text{const} \,. \tag{25}$$

The maximum of the logarithm (25) is computed using the Lagrangian multipliers method to take into account the constraints:

$$\frac{\partial}{\partial n_j}\left[\sum_k (n_k \log n_k - n_k) + \alpha n_k + \beta n_k l_k\right] = 0 \tag{26}$$

which implies

$$n_j = \exp(-\alpha - \beta l_j) \tag{27}$$

where the parameters α and β are determined by the conditions (23). In the continuous limit, a statistical distribution $\rho(s) \propto \exp(-\beta s)$ is achieved, consistently with experimental data. It is really remarkable that β^{-1} (i.e the average length of urban trajectories) is the same for very different cities (see Fig. 18) [73]; this can be an indication for the existence of an individual mobility demand that can be related to a universal energy law in human travel behavior, according to the law proposed by D. Helbing and R. Kölbl in [64], based on the data of the *UK National Travel Surveys* during the years 1972–1998. Moreover, it allows the conjecture that the universal behavior of the length distribution (Fig. 18) can be explained as the result of a specific cognitive behavior of the populations. In other words, the spatial distance is the norm for the individual driving car decision and the "activities sprawling" in modern cities observed by urban planners [7], introducing a decorrelation in the system, which allows to apply the statistical physics techniques.

In the case of Rome, the lengths distribution has been also computed using the arc-length metric. The experimental results are plotted in Fig. 19, confirming the exponential behavior shown in Fig. 18, even if the decay slope is slightly different. This is a remarkable fact since one expects a decreasing of the path-lengths distribution when the length approaches to zero: it should be not convenient to use the car for short travels (<1 km). A possible explanation is the presence of a relevant stochastic components in urban mobility (the so-called "asystematic mobility"), so that the individual mobility becomes the realization of a sequence of local destinations spread in the urban space. The experimental data in Fig. 19 suggest that the citizens use apparently two linear proximity functions, for the short and for the long length paths. Then the length distribution can be interpreted as an overlap of two exponential distributions

$$\rho(s) = A\left(\exp(-\beta_l s) + c\exp(-\beta_s s)\right) \,. \tag{28}$$

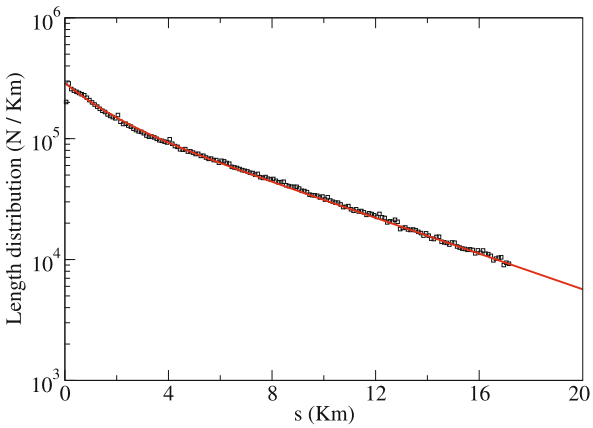

Traffic and Crowd Dynamics: The Physics of the City, Figure 19
***Squares*: length distribution of the inner trajectories in Rome during February 2007, computed using the arc-length metric. *Continuous curve*: interpolation of the experimental data with the analytical distribution (28). The interpolated formula has the following parameters: $\beta_l = .15\,\mathrm{km}^{-1}$, $\beta_s = .6\,\mathrm{km}^{-1}$, and $c = .7$. The vertical scale is logarithmic**

The continuous curve in Fig. 19 shows the interpolation results using the parameters: $\beta_l = .15\,\mathrm{km}^{-1}$, $\beta_s = .6\,\mathrm{km}^{-1}$ and $c = .7$.

However the existence of two exponentials curves for the length distribution should be studied more carefully, since the very short trajectories can be interpreted in various ways and the GPS signals have a low quality in such a case.

Future Directions

From a theoretical point of view, we think that the gas of automata can be a good paradigmatic framework to investigate complexity in a general sense. For crowd and traffic flow, the previous model simulation and empirical results permit to tackle the problem of predictability and governance of the mobility system [26]. Last but not least, the prediction, even if partial or probabilistic, is relevant not only for the pleasure knowledge, but also for the application in order to minimize the impact of provisional city users on the town, and in particular on the mobility of local populations to prevent critical and potentially dangerous crowding phenomena, to prefigure and plan safety and security saving trajectories.

Bibliography

1. Ball P (2003) The Physical Modelling of Human Social Systems. Complexus 1:190
2. Balmer M, Nagel K, Raney B (2004) Large-scale multi-agent simulations for transportation applications. J Intell Transp Syst 8(4):205
3. Bandini S, Manzoni S, Vizzari G (2006) Crowd Modeling and Simulation. The role of multi-agent simulation in design support systems. In: Van Leeuwen JP, Timmermans HJP (eds) Innovations in Design & Decision Support Systems in Architecture and Urban Planning. Springer, Netherlands, pp 161
4. Bando M et al (1995) Dynamical model of traffic congestion and numerical simulation. Phys Rev E 51(2):1035
5. Batty M (1971) Modelling Cities as Dynamic Systems. Nature 231:425–428
6. Batty M (2003) Agent-Based Pedestrian Modeling. In: Longley PA, Batty M (eds) Advanced Spatial Analysis. ESRI Press, Redlands, pp 81–105
7. Batty M (2005) Agents, Cells and Cities: New Representational Models for Simulating Multi-Scale Urban Dynamics. Environ Plan A 37(8):1373–1394
8. Batty M (2005) Cities and Complexity. Understanding Cities with Cellular Automata, Agent-Based Models, and Fractals. MIT Press, Cambridge
9. Batty M, Torrens P (2001) Modeling Complexity: The Limits to Prediction CASA. Working Paper Series, vol 36. Available online at www.casa.ucl.ac.uk
10. Bazzani A et al (2003) A Chronotopic Model of Mobility in Urban Spaces. Physica A 325:517–530
11. Bazzani A et al (2007) A Model for Asystematic Mobility in Urban Space. In: Albeverio S, Andrey D, Giordano P, Vancheri A (eds) The Dynamics of Complex Urban Systems: An Interdisciplinary Approach, vol 59. Physica, Heidelberg
12. Bazzani A, Rambaldi S, Giorgini B, Turchetti G (2007) Complexcity: modeling urban mobility. Adv Complex Syst 10(2):255
13. Bazzani A, Giorgini B, Rambaldi S (eds) (2008) Physics and the City. Adv Complex Syst 10(2):215–377
14. Bechtel R (1970) Human movement in architecture. In: Proshansky HM et al (eds) Environmental Psychology. Rinehart and Winston, New York
15. Bellman R (1957) Dynamics Programming. Princeton Univesity Press, Princeton
16. Ben-Akiva M, Lerman SR (1985) Discrete choice analysis theory and application to travel demand. The MIT University Press, Cambridge
17. Benenson I, Torrens PM (2004) Geosimulation: Automata-based modeling of urban phenomena. Wiley, London
18. Biroli G (2007) Jamming: A new kind of phase transition? Nat Phys 3:222
19. Brian WA, Durlauf NS, Lane AD (eds) (1997) The Economy as an Evolving Complex System II, Proceedings, vol XXVII. Addison-Wesley, Reading
20. Byrne DS (1998) Complexity Theory and the Social Sciences: An Introduction. Routledge, London
21. Cascetta E (2001) Transportation Systems Engineering: Theory and Methods. Kluwer, Boston
22. Chomsky N (2000) New Horizons in the Study of Language and Mind. Cambridge University Press, Cambridge
23. Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. Phys Rep 329:199
24. Chowdhury D, Nishinari K, Schadschneider K (2004) Self-organized patterns and traffic flow in colonies of organisms. Phase Transit 77:601624
25. Cucker F, Smale S (2005) Emergent Behavior in Flocks http://ttic.uchicago.edu/smale/papers.html

26. Daganzo FC (2007) Urban gridlock: Macroscopic modeling and mitigation approaches. Transp Res B 41:49
27. de Finetti B (1972) Probability, Induction and Statistics. Wiley, New York
28. De Martino B et al (2006) Frames, biases and rational decision making in the human being. Science 313:684–687
29. Domencich TA, Domencich VD (1975) Urban Travel Demand. A Behavioral Analysis. North Holland, Amsterdam
30. Ellis GRF (2006) Physics and the real world. Found Phys 36–2:227
31. Fishburn PC (1983) Transitive Measurable Utility. J Econ Theory 31:293
32. Fruin JJ (1971) Pedestrian planning and design. Metropolitan Association of Urban Designers and Environmental Planners, New York
33. Garnsey E, McGlade J (eds) (2006) Complexity and Co-evolution Continuity and Change in Socio-Economic Systems. Edward Elgard, Cheetenham
34. Giorgini B (2007) Philosophie naturelle de la causalité et du hasard dans un modèle de mobilité urbaine. In: Franceschelli S, Paty M, Roque T (eds) Chaos et systèmes dynamiques. Collection Visions des Sciences, vol 259. Editions Hermann, Paris
35. Giorgini B, Turchetti G (2005) From Newton-Boltzmann paradigms to complexity: a bridge to biosystems. In: The Science of Complexity: chimera or reality? Milan Research Centre for Industrial and Applied Mathematics, Esculapio Editore, Bologna, p 18
36. Giorgini B et al (2008) The Physics of the City: Modeling Complex Mobility. In: Diamantini D, Martinotti G (eds) Urban science forward look. Scripta Web, Napoli
37. Hayakawa H, Nakanishi K (1998) Universal behaviors in granular flows and traffic flows. Prog Theor Phys Suppl 130:57
38. Heinrich MJ, Nagel RS, Behringer RP (1996) Granular solids, liquids, and gases. Rev Mod Phys 68:1259
39. Helbing D (2001) Traffic and Related Self-Driven Many-Particle Systems. Rev Mod Phys 73:1067
40. Helbing D (2004) Dynamic decision behavior and optimal guidance through information services. In: Schreckenberg M, Selten R (eds) Models and experiments in Human Behaviour and Traffic Networks. Springer, Berlin, p 47
41. Helbing D, Molnár P (1995) Social force model for pedestrian dynamics. Phys Rev E 51:4282
42. Helbing D, Nagel K (2004) The physics of traffic and regional development. Contemp Phys 45(5):405–426
43. Helbing D et al (2006) Analytical approach to continuous and intermittent bottleneck flows. Phys Rev Lett 97:168001
44. Helbing D, Farkas I, Vicsek T (2000) Simulating dynamical features of escape panic. Nature 407:487
45. Helbing D, Farkas IJ, Vicsek T (2000) Freezing by Heating in a Driven Mesoscopic System. Phys Rev Lett 84:1240
46. Helbing D, Schönhof M, Kern D (2002) Volatile decision dynamics: experiments, stochastic description, intermittency control and traffic optimization. New J Phys 4:33
47. Helbing D, Molnár P, Farkas I, Bolay K (2003) Self-organizing pedestrian movement. Environ Plan B 28:361
48. Helbing D, Isobe M, Nagatani T, Takimoto K (2003) Lattice gas simulation of experimentally studied evacuation dynamics. Phys Rev E 67:067101
49. Helbing D, Buzna L, Johansson A, Werner T (2005) Self-organized pedestrian crowd dynamics: Experiments, simulations, and design solutions. Transp Sci 39–1:1
50. Helbing D, Johansson A, Al-Abideen HZ (2007) Dynamics of crowd disasters: an empirical study. Phys Rev E 75:046109
51. Hoogendoorn SP, Bovy PHL (2001) State-of-the-art of vehicular traffic flow modeling. In: Proceedings of the: Institution of Mechanical Engineers. Part I. J Syst Control Eng 215–4:283
52. Hoogendoorn SP, Daamen W (2005) Pedestrian Behavior at Bottlenecks. Transp Sci 39(2):147
53. Hoogendoorn SP et al (2003) Microscopic Traffic Data Collection by Remote Sensing. In: Transportation Research Board (TRB) 82nd Annual Meeting. Mira Digital Publishing, St. Louis, p 11
54. Hoogendoorn SP, Bovy PHL, Daamen W (2002) Microscopic Pedestrian Wayfinding and Dynamics Modelling. In: Schreckenberg M, Sharma SD (eds) Pedestrian and Evacuation Dynamics. Springer, Berlin, p 123
55. Isobe M, Adachi T, Nagatani T (2004) Experiment and simulation of pedestrian counter flow. Physica A 336:638
56. Jiang B (1998) SimPed: Simulating Pedestrian Flows in a Virtual Urban Environment. J Geogr Inf Decis Anal 3–1:21
57. John A, Schadschneider A, Chowdhury D, Nishinari K (2004) Collective effects in traffic on bi-directional ant trails. J Theor Biol 231:279
58. Kerner BS, Klenov SL (2006) Deterministic microscopic three-phase traffic flow models. J Phys A: Math Gen 39:1775
59. Kirchner A, Nishinari K, Schadschneider A (2003) Friction Effects and Clogging in a Cellular Automaton Model for Pedestrian Dynamics. Phys Rev E 67:056122
60. Kirchner A, Klüpfel H, Nishinari K, Schadschneider A, Schreckenberg M (2004) Discretization Effects and the Influence of Walking Speed in Cellular Automata Models for Pedestrian Dynamics. J Stat Mech P10011
61. Klimontovich YL (1999) Entropy, information, and criteria of order in open systems. Nonlinear Phenom Complex Syst 2–4:1
62. Kluegl F, Bazzan ALC (2004) Route Decision Behaviour in a Commuting Scenario: Simple Heuristics Adaptation. In: Schreckenberg M, Selten R (eds) Human Behaviour and Traffic Networks. Springer, Berlin, p 285
63. Knospe W et al (2000) CA Models for traffic flow: Comparison with empirical single-vehicle data. In: Helbing D, Herrmann HJ, Schreckenberg M, Wolf DE (eds) Traffic and Granular 1999: Social, Traffic, and Granular Dynamics. Springer, Berlin
64. Kölbl RR, Helbing D (2003) Energy laws in human travel behavior. New J Phys 5:48.1
65. Kretz T et al (2006) Experimental study of pedestrian counterflow in a corridor. J Stat Mech P10001
66. Kretz T, Wölki M, Schreckenberg M (2006) Characterizing correlations of flow oscillations at bottlenecks. J Stat Mech Theory Exp P02005
67. Miller JH, Page SE (2007) Complex Adaptive Systems. Princeton Studies in Complexity
68. Millonas MM (1994) Swarms, phase transitions, and collective intelligence. In: Langton CG (ed) Artificial Life III. Addison Wesley, Reading, p 418
69. Mitarai N, Nakanishi H (1999) Stability Analysis of Optimal Velocity Model for Traffic and Granular Flow under Open Boundary Condition. J Phys Soc Jpn 68–8:2475
70. Mizar LF (2007) Agent-Based Modelling and Micro-Simulations of Pedestrian Dynamics and Crowd Phenomena in Public Enviroments. PHD Thesis, Università degli Studi, Milano Bicocca

71. Muramatsu M, Nagatani T (2000) Jamming transition in two-dimensional pedestrian traffic. Phys A: Statist Mech Appl 275:281
72. Muramatsu M, Irie T, Nagatani T (1999) Jamming transition in pedestrian counter flow. Physica A 267:487
73. Naess P (2006) Are short daily trips compensated by higher leisure mobility? Environ Plan B: Plan Des 33:197
74. Nagatami T (2002) The physics of traffic jams. Report Prog Phys 65:1331
75. Nagel K, Schreckenberg M (1992) A cellular automaton model for freeway traffic. J Phys I 2:2221
76. Nicolis G, Prigogine I (1989) Exploring Complexity: An Introduction. Freeman, New York
77. Parisi G (1993) Statistical Physics and Biology. Princeton University Press, p 284
78. Parisi G (1999) Complex Systems: a Physicist's Viewpoint. Phys A 263–1:557
79. Portugali J (2005) Cognitive Maps Are over 60. COSIT 251
80. Prigogine I, Andrews FC (1960) A Boltzmann like approach for traffic flow. Oper Res 8:789
81. Pushkarev B, Zupan JM (1975) Urban space for pedestrians. The MIT Press, Cambridge
82. Rambaldi S et al (2007) Mobility in modern cities: looking for physical laws. Proceeding of the ECCS07 Conference, Dresden, 1–3 October 2007, paper n. 132
83. Rizzolatti G (1997) From Complexity to Creativity. Plenum Press, New York
84. Roederer GJ (2003) On the concept of information and its role in Nature. Entropy 5:3
85. Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) (2007) Traffic and Granular Flow '05. Springer, Heidelberg
86. Schreckenberg M, Neubert L, Wahle J (2001) Simulation of traffic in large road networks. Futur Gener Comput Syst 17–5:649
87. Schweitzer F (2003) Brownian Agents and Active Particles. Springer, Berlin
88. Tajima Y, Takimoto K, Nagatani T (2001) Scaling of pedestrian channel flow with a bottleneck. Physica A 294:257
89. Tajima Y, Takimoto K, Nagatani T (2002) Pattern formation and jamming transition in pedestrian counter flow. Physica A 313:709
90. Taylor M (2003) The Moment of Complexity. University of Chicago Press, Chicago
91. Treiber M, Hennecke A, Helbing D (2000) Congested Traffic States in Empirical Observations and Microscopic Simulations. Phys Rev E 62:1805
92. Turchetti G, Zanlungo F, Giorgini B (2007) Dynamics and thermodynamics of a gas of automata. Europhys Lett 78–5:58003
93. Van Gelder T (1998) The dynamical hypothesis in cognitive science. Behav Brain Sci 21:615
94. Vogel A, Nagel K (2005) Multi-agent based simulation of individual traffic in Berlin. Paper presented at CUPUM 2005 Conference, 29 June – 1 July 2005
95. Von Neumann J (1963) The general and logical theory of automata. In: Collected works, vol V. Pergamon Press, Oxford, p 288
96. Von Neumann J, Morgenstern O (1947) Theory of Games and Economic Behavior. Princeton University Press, Princeton
97. Wahle J et al (2000) Decision Dynamics in a Traffic Scenario. Adv Complex Syst 2:1
98. Weufeng F, Lizhong Y, Weicheng F (2003) Simulation of bi-directional pedestrian movement using a celluar automata model. Physica A 321:633
99. www.physicsofthecitylab.unibo.it
100. Yamori K (1998) Going with the flow: Micro-nacro dynamics in the macrobehavioral patterns of pedestrian crowds. Psychol Rev 105–3:530

Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment

Hesham Rakha, Aly Tawfik
Center for Sustainable Mobility, Virginia Tech Transportation Institute, Virginia Polytechnic Institute and State University, Blacksburg, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Driver Travel Decision Behavior Modeling
Static Traffic Routing and Assignment
Dynamic Traffic Routing
Traffic Modeling
Dynamic Travel Time Estimation
Dynamic or Time-Dependent Origin-Destination Estimation
Dynamic Estimation of Measures of Effectiveness
Use of Technology to Enhance System Performance
Related Transportation Areas
Future Directions
Appendix
Bibliography

Glossary

Link or arc A roadway segment with homogeneous traffic and roadway characteristics (e. g. same number of lanes, base lane capacity, free-flow speed, speed-at-capacity, and jam density). Typically networks are divided into links for traffic modeling purposes.

Route or path A sequence of roadway segments (links or arcs) used by a driver to travel from his/her point of origin to his/her destination.

Traffic routing The procedure that computes the sequence of roadways that minimize some utility objective function. This utility function could either be travel time or a generalized function that also includes road tolls.

Traffic assignment The procedure used to find the link flows from the Origin-Destination (O-D) demand. Traffic assignment involves two steps: (1) traffic routing and (2) traffic demand loading. Traffic assignment can be divided into static, time-dependent, and dynamic.

User equilibrium traffic assignment The assignment of traffic on a network such that it distributes itself in a way that the travel costs on all routes used from any origin to any destination are equal, while all unused routes have equal or greater travel costs.

System optimum traffic assignment The assignment of traffic such that the average journey travel times of all motorists is a minimum, which implies that the aggregate vehicle-hours spent in travel is also minimum.

Static traffic assignment Traffic assignment ignoring the temporal dimension of the problem.

Time-dependent traffic assignment An approximate approach to modeling the dynamic traffic assignment problem by dividing the time horizon into steady-state time intervals and applying a static assignment to each time interval.

Dynamic traffic assignment Traffic assignment considering the temporal dimension of the problem.

Traffic loading The procedure of assigning O-D demands to routes.

Synthetic O-D estimation The procedure that estimates O-D demands from measured link flow counts, which includes static, time-dependent, and dynamic.

Traffic stream motion model A mathematical representation (traffic flow model) for traffic stream motion behavior.

Car-following model A mathematical representation (traffic flow model) for driver longitudinal motion behavior.

Marginal link travel time The increase in a link's travel time resulting from an assignment of an additional vehicle to this link.

Road pricing Road pricing is an economic concept in which drivers are charged for the use of the road facility.

Definition of the Subject

The dynamic nature of traffic networks is manifested in both temporal and spatial changes in traffic demand, roadway capacities, and traffic control settings. Typically, the underlying network traffic demand builds up over time at the onset of a peak period, varies stochastically during the peak period, and decays at the conclusion of the peak period. As traffic congestion builds up within a transportation network, drivers may elect to either cancel their trip altogether, alter their travel departure time, change their mode of travel, or change their route of travel. Dynamic traffic routing is defined as the process of dynamically selecting the sequence of roadway segments from a trip origin to a trip destination. Dynamic routing entails using time-dependent roadway travel times to compute this sequence of roadway segments. Consequently, the modeling of driver routing behavior requires the estimation of roadway travel times into the near future, which may entail some form of traffic modeling.

In addition to dynamic changes in traffic demand, roadway capacities are both stochastic and vary dynamically as vehicles interact with one another along roadway segments. For example, the roadway capacity at a merge section varies dynamically as the composition of on-ramp and freeway demands vary [45,47,73,75,106,107,108,109,110,122,123,127,144,183]. To further complicate matters, traffic control settings (e. g. traffic signal timings) also vary both temporally and spatially, thus introducing another level of dynamics within transportation networks. All these factors make the dynamic assessment of traffic networks extremely complex, as shall be demonstrated in this article. The article is by no means comprehensive but does provide some insight into the various challenges and complexities that are associated with the assessment of dynamic networks.

Introduction

Studies have shown that even drivers familiar with a trip typically choose sub-optimal routes thus incurring extra travel time in the range of seven percent on average [103]. Furthermore, the occurrence of incidents and special events introduces other forms of variability that drivers are unable to anticipate and thus result in additional errors in a driver's route selection. Consequently, advanced traveler information systems (ATISs), which are an integral component of intelligent transportation systems (ITSs), can assist the public in their travel decisions by providing real-time travel information via route guidance systems; variable message signs (VMSs); the radio, or the web. It is envisioned that better travel information can enhance the efficiency of a transportation system by allowing travelers to make better decisions regarding their time of departure, mode of travel, and/or route of travel. An integral component of an ATIS is a dynamic traffic assignment (DTA) system. A DTA system predicts the transportation network state over a short time horizon (typically 15- to 60-min time horizon) by modeling complex demand and supply relationships through the use of

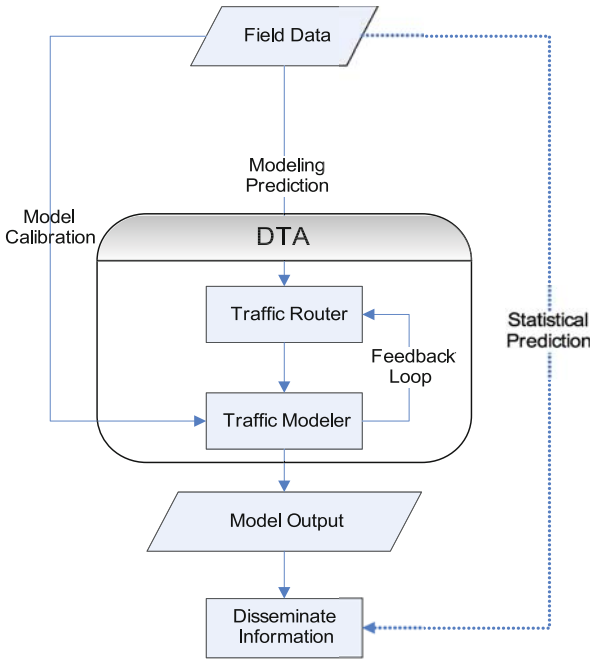

Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment, Figure 1
Schematic of an ATIS Framework

sophisticated models and algorithms. The DTA requires two sets of input, namely demand and supply data. Demand represents the demand for travel and is typically in the form of mode-specific time-dependent origin-destination (O-D) matrices. Alternatively, the supply component models the movement of individual vehicles along a roadway typically using roadway specific speed-flow-density relationships together with the explicit modeling of queue buildup and decay. Figure 1 illustrates schematically that an ATIS can utilize two approaches for the estimation of future traffic conditions, namely: statistical models or a DTA framework. This article focuses on the DTA approach and thus will be described in more detail. The DTA combines a traffic router and modeler, as illustrated in the figure. The traffic router estimates the optimum travel routes while the traffic modeler models traffic to evaluate the performance of traffic after assigning motorists to their routes. A feedback loop allows for the feedback of either travel times or marginal travel times, which in turn, are used by the traffic router to compute the optimum routes. This feedback continues until the travel times are consistent with the travel routes and there is no incentive for drivers to alter their routes.

A DTA can be applied off-line (in a laboratory) or on-line (in the field). An on-line application of a DTA entails gathering traffic data in real-time at any instant t and feeding these data to the DTA to predict short-term traffic conditions Δt temporal units into the future (i. e. at time $t + \Delta t$). As was mentioned earlier, the input to the DTA includes mode specific time dependent O-D matrices. Unfortunately, current surveillance equipment does not measure O-D matrices; instead they measure traffic volumes passing a specific point. Consequently, O-D estimation tools are required to estimate the O-D matrix from observed link counts, as illustrated in Fig. 2. However, the estimation of an O-D matrix requires identifying which O-D demands contribute to which roadway counts. The assigning of O-D demands to link counts involves what is commonly known in the field of traffic engineering as the traffic assignment problem. Traffic assignment in turn requires real-time O-D matrices and roadway travel times as input. Consequently, some form of feedback is required to solve this problem. A more detailed description of traffic assignment formulations and techniques is provided in Sect. "Static Traffic Routing and Assignment" and "Dynamic Traffic Routing", while the estimation of route travel times is described in Sect. "Dynamic Travel Time Estimation" and the estimation of O-D matrices is described in Sect. "Dynamic or Time-Dependent Origin-Destination Estimation".

The dynamic assessment of traffic networks using a DTA is both data driven (trapezoidal boxes) and model based (colored rectangular boxes), as illustrated in Fig. 2. This procedure involves: measuring raw field data, constructing model input data, executing a traffic model to predict future conditions, and advising a traveler in the case of control systems. The framework starts by measuring traffic states at instant "t" (roadway travel times and link flows) and subsequently estimating these traffic states Δt in the future. Procedures for the estimation of dynamic roadway travel times are provided in Sect. "Dynamic Travel Time Estimation" of this article. Using the measured link flows and travel times, an O-D matrix is constructed using a synthetic O-D estimator. Section "Dynamic or Time-Dependent Origin-Destination Estimation" describes the various formulations for estimating a dynamic O-D matrix together with some heuristic practical approaches to estimate this O-D matrix.

Once the O-D demands are estimated the future states are predicted using a traffic modeler. Section "Traffic Modeling" provides a brief overview of the various state-of-the-practice modeling approaches. The model also computes various measures of effectiveness (MOEs) including delay, fuel consumption, and emissions, as will be described in Sect. "Dynamic Estimation of Measures of Effectiveness". The traffic modeler can either combine traf-

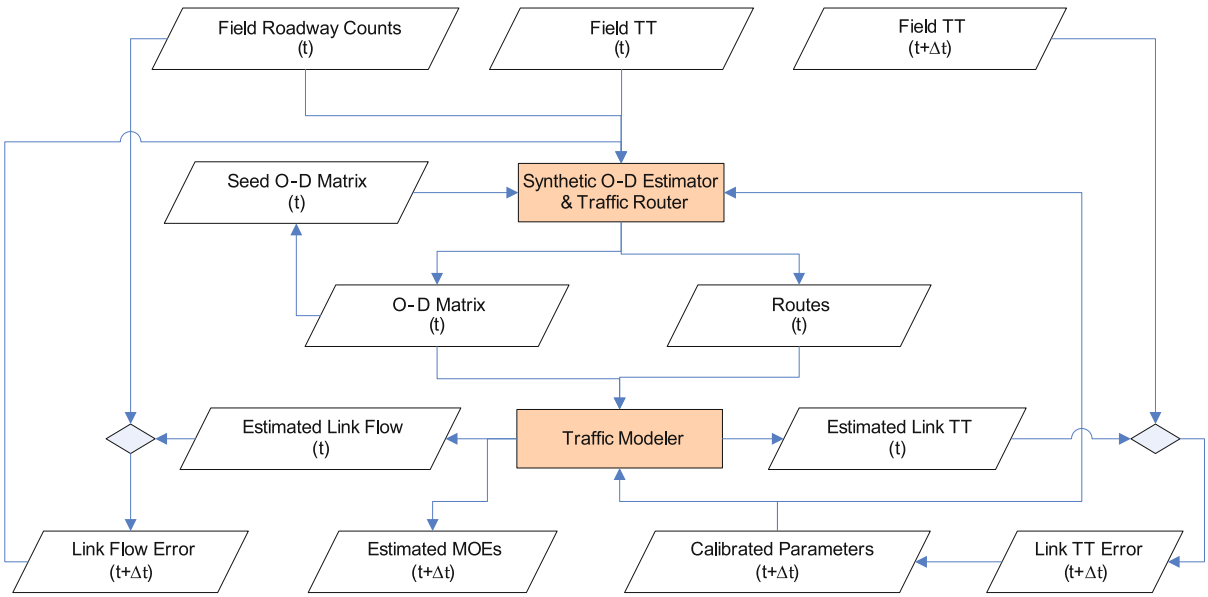

Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment, Figure 2
Dynamic Traffic Assessment and Routing Framework

fic modeling with traffic assignment or alternatively utilize the routes computed by the O-D estimator to route traffic. This closed loop optimal control framework can involve a single loop or in most cases may involve an iterative loop to attain equilibrium. The framework involves a feedback loop in which input model parameters are adjusted in real-time through the computation of an error between model predictions and actual measurements. This real-time calibration entails adjusting roadway parameters (e. g. capacity, free-flow speed, speed-at-capacity, and jam density) and traffic routes to reflect dynamic changes in traffic and network conditions. For example, the capacity of a roadway might vary because of changes in weather conditions and/or the occurrence of incidents. The system should be able to adapt itself dynamically without any user intervention.

This article attempts to synthesize the literature on the dynamic assessment and routing of traffic. The problem as will be demonstrated later in the paper is extremely complex because, after all, it deals with the human psychic, which not only varies from one person to another, but may also vary depending on the purpose of a trip, the level of urgency the driver has, and the psychic of the driver at the time the trip is made. This article is by no means comprehensive, given the massive literature on the topic, but does highlight some of the key aspects of the problem, how researchers have attempted to address this problem, and future research needs and directions.

The article discusses the various issues associated with the dynamic assessment of transportation systems. Initially, driver travel decision behavior modeling is presented and discussed. Subsequently, various traffic assignment formulations are presented together with the implementation issues associated with these formulations. Next, the mathematical formulations of these assignment techniques are discussed together with mathematical and numerical approaches to modeling dynamic traffic routing. Subsequently, the issues associated with the modeling of traffic stream behavior, the estimation of dynamic roadway travel times, and the estimation of dynamic O-D demands are discussed. Subsequently, the procedures for computation of various assessment measures are presented. Next, the use of technology to alter driver behavior is presented. Finally, directions for further research are presented.

Driver Travel Decision Behavior Modeling

As with the general case of modeling human behavior, modeling driver travel behavior has always been complicated, never accurate enough, and in constant demand for further research. Among the early attempts to model human choice behavior is the economic theory of the "economic man"; who in the course of being economic is also "rational" [222]. According to Simon's exact words, "*actual human rationality-striving can at best be an extremely*

crude and simplified approximation to the kind of global rationality that is implied, for example, by game-theoretical models".

In general, traffic assignment (static or dynamic assignment) has undoubtedly been among the most researched transportation problems, if not the most, for more than the past half of a century. However, DTA in particular has had the bigger share for almost one third of a century now. Since the early work of Merchant and Nemhauser [142,143], researchers have attempted to improve available DTA models, hence, providing a very rich and vastly wide literature.

As a result of the rapid technological evolution over the last decade of the previous century (the 20th century); manifested in the communications, information and computational technological advances; a worldwide initiative to add information and communications technology to transport infrastructure and vehicles, termed as the intelligent transportation systems (ITS) program, was introduced to the transportation science. According to the Wikipedia Encyclopedia, among the main objectives of ITS is to "*manage factors that are typically at odds with each other such as vehicles, loads, and routes to improve safety and reduce vehicle wear, transportation times and fuel consumption*". Needless to say, the ITS impact on route selection and roadway travel times has a direct effect on a DTA.

The main effect of ITS on DTA manifests itself within the area of advanced traveler information systems (ATIS). ATIS is primarily concerned with providing people, in general, and trip makers, in particular, with pre-trip and en-route trip-related information. According to the US Federal Highway Administration (FHWA), "*advanced traveler information includes static and real-time information on traffic conditions, and schedules, road and weather conditions, special events, and tourist information. ATIS is classified by how and when travelers receive their desired information (pre-trip or en-route) and is divided by user service categories. Operations essential to the success of these systems are the collection of traffic and traveler information, the processing and fusing of information - often at a central point, and the distribution of information to travelers. Important components of these systems include new technologies applied to the use and presentation of information and the communications used to effectively disseminate this information*" [157].

As will be discussed later, a significant amount of DTA research is directed towards developing data dissemination standards. These standards attempt to achieve the maximum possible benefits while complying with the ITS objectives. Although the provision of pre-trip information may influence traveler departure time and route of travel (and in extreme cases, might result in a person canceling his/her trip all together), thus requiring further complicated DTA models that capture forgone and induced demand, as will be discussed later. Moreover, probably the greatest dimension for DTA model complexity was introduced to research when the disseminated ATIS information was to be designed as a control factor to change the manner by which trips are distributed over the network, for example from user equilibrium to system optimum.

Although ITS and ATIS were practically introduced a little more than a decade ago, and in spite of the significant research funds and efforts that have been devoted to the topic, current available DTA models are, at least, relatively undeveloped, which necessitates new approaches that can capture the challenges from the application domains as well as for the fundamental questions related to tractability and realism [180]. This will be discussed briefly in the following section.

Driver travel decision theory is a complicated research area. Research within this area encompasses a very wide range of research efforts. Before going over a brief list of these possible research areas, it should be noted that most of these research areas overlap with one another. Therefore, for a valid driver behavior model, all of the following aspects should be efficiently covered in a practical and realistic manner. This been said, the following is a brief list of some of the main research areas that are highly related to driver travel decision theory:

- Human decision theory, which can be reflected in the trip maker's decision to make or cancel a scheduled trip, route and departure time selection, compliance with the pre-trip or en-route disseminated information, en-route path diversion and/or return, mode choice based on disseminated information, etc. Literature concerning human decision theory extends back to more than half a century ago and continues to be researched up to this date. Examples of the literature concerning the human decision theory include: administrative behavior [220,221], theory of choice [15], rational choice theory [222], game theory, and decision field theory [36]. Examples of the literature concerning driver decision theory include: decision field theory [229], approximate reasoning models [113], route choice utility models [89], inductive learning [151], effect of age on routing decisions [243], and rational learning [152].
- Design of disseminated information, which encompasses the criteria governing the dissemination of in-

formation, the structure and type of information to be disseminated, when data are disseminated, and identifying target drivers. This governs, to a large extent, the drivers' compliance rates in response to disseminated information. Hence, affecting the routes chosen by drivers, the traffic volumes on these routes and alternative routes, and different travel times, among others. Literature concerning the effect of ATIS and ATIS content on drivers behavior include: the required information that would reduce traffic congestion [14], the effect of ATIS on drivers route choice [1], commuters diversion propensity [214], the effect of traffic information disseminated through variable message signs on driver choices [172], drivers en-route routing decisions [111].

- Human perception based on experience and information provision, which is reflected in day-to-day variations in driver decisions. For example, given identical conditions on two separate days, the same person might select different routes and departure times; possibly due to different experiences on previous days. Examples of current literature include: models that include the incorporation of driver behavior dynamics under information provision [176], behavioral-based consistency seeking models [177], perception updating and day-to-day travel choice dynamics with information provision [104], the modeling of inertia and compliance mechanisms under real-time information [226], drivers psychological deliberation while making dynamic route choices [229], the effect of using in-vehicle navigational systems on diver behavior [11], the effect of network familiarity on routing decisions [128], and the effect of varying levels of cognitive loads on driver behavior [105].
- Among the challenges in modeling human decision theory are the possible data collection techniques. The current practice for data collection includes revealed and stated preference surveys. Research has demonstrated that surveyed stated preference results have significant biases; in comparison to real behavior. In addition to the research being performed to analyze, capture, and improve the reasons for such biases; other research directions are being performed to solve other survey problems. For example, the problems of low and slow survey participation rates, as well as under-represented groups in typical survey techniques. Examples of literature within this field include: stated preference for investigating commuters diversion propensity [214], using stated preference for studying the effect of advanced traffic information on drivers route choice [1], driver response to variable message sign-based traffic information according to stated preference data collected through three different survey administration methods, namely, an on-site survey, a mail-back survey and an internet-based survey [172], transferring insights into commuter behavior dynamics from laboratory experiments to field surveys [129] and [173], and the applicability of using driving simulators for data collection [113].
- Issues of uncertainty, which is a fundamental feature in most transportation phenomena. Research dealing with uncertainty has a wide application in DTA. It can be represented in the trip maker route travel time estimates, in the compliance rates of drivers to information, in the driver's trust in the disseminated information and its reliability, among others. Uncertainty-related research issues have been addressed through several approaches, like stochastic modeling, fuzzy control, and reliability indices. Examples of current literature include the works of Birge and Ho [28], Peeta and Zhou [178,179], Cantarell and Cascetta [38], Ziliaskopoulos and Waller [263], Waller and Ziliaskopoulos [245], Waller [244], Peeta and Jeong [177], Jha et al. [104], Peeta and Paz [171], Koutsopoulos et al. [113], and Hawas [89].

Static Traffic Routing and Assignment

Prior to describing the issues associated with dynamic routing, a description of static routing issues is first presented. This section describes two formulations for static traffic assignment, namely the User Equilibrium (UE) and System Optimum (SO) assignment. Traffic assignment is defined as the basic problem of finding the link flows given an origin-destination trip matrix and a set of link or marginal link travel times, as illustrated in Fig. 3. The solution of this problem can either be based on the assumption that each motorist travels on the path that minimizes his/her travel time – known as the UE assignment – or alternatively to minimize the system-wide travel time – known as the SO assignment. The traffic assignment initially computes the travel routes (paths) and then determines the unique link flows on the various network links. As will be discussed later, while the estimated link flows are unique the path flows that are derived from these link flows are not unique and thus require some computational tool to estimate the most-likely of these path flows (synthetic O-D estimator). If a time dimension is introduced to the assignment module the formulation is extended from a static to a dynamic context. However, as will be discussed later the addition of a time dimension deems the formulation non-convex and thus the mathematical program used to solve the problem becomes in-

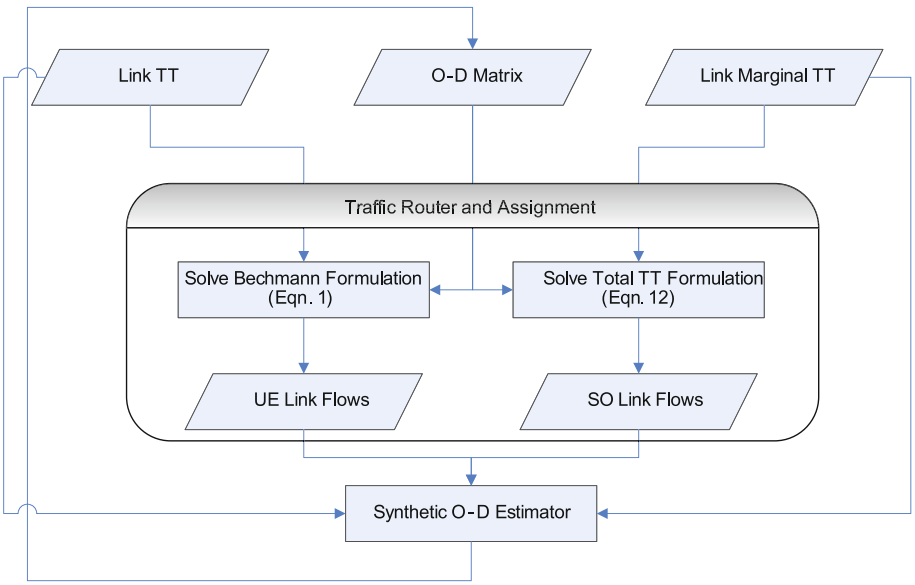

Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment, Figure 3
Traffic Assignment Framework

feasible and thus comes the need for a simulation-based solution approach.

Wardrop [246] was the first to explicitly differentiate between these two alternative traffic assignment methods or philosophies. Models based on Wardrop's first principle are referred to as UE, while those based on the second principle are deemed as SO. Wardrop's first principle states that "*traffic on a network distributes itself in such a way that the travel costs on all routes used from any origin to any destination are equal, while all unused routes have equal or greater travel costs*." Alternatively, Wardrop's second principle states that the average journey travel times of all motorists is a minimum, which implies that the aggregate vehicle-hours spent in travel is also minimum.

One of the most spectacular examples that illustrated that the UE flow in a network is in general different from the SO flow, is the Braess network [31]. In this network the system-optimal flow was obtained by completely suppressing the flow which would normally occur, on a certain link, at equilibrium. The Braess "paradox" was studied later in more detail [78,80,117,118,147,209,227,228]. For example, Stewart [228] illustrated three important facts using a very simple two-link network and the Braess paradox that included: (a) the equilibrium flow does not necessarily minimize the total cost; (b) adding a new link to a network may increase the total cost at equilibrium; (c) adding a new link to a network may increase the equilibrium travel cost for each individual motorist. Stewart also illustrated that a group of travelers having only one reasonable route may be seriously inconvenienced by another group of travelers who choose the same route in order to obtain a slight improvement in their personal cost of travel.

User Equilibrium vs. System Optimum Traffic Assignment

The differences between user and system optimum traffic assignment are best illustrated using an example illustration. The sample test network for this study is derived from an earlier study by Rakha [181]. The network consists of two one-way routes, numbered 1 and 2, from origin A to destination B. The travel time relationship for route 1 is characterized by the relationship $10 + 0.010\, v_1$ where v_1 is the traffic volume on route 1 (veh). Alternatively, the travel time along route 2 is characterized by the relationship $15 + 0.005\, v_2$ where v_2 is the traffic volume traveling along route 2 (veh). Considering at total demand of 1000 veh traveling between zones A and B, the travel time along routes 1 and 2 vary as a function of the volume on each of the routes, as illustrated in Fig. 4. The figure demonstrates that the travel times along routes 1 and 2 are equal at 16.5 min when 667 veh travel along route 1 and 333 veh travel along route 2. Alternatively, the system-optimum traffic assignment is achieved at a volume distribution of 500 veh on routes 1 and 2, respectively. From a traffic engineering point of view, the difference in total travel time between the system and user-optimum traffic assignment (16,250 versus 16,667 veh-min) is of interest. This difference rep-

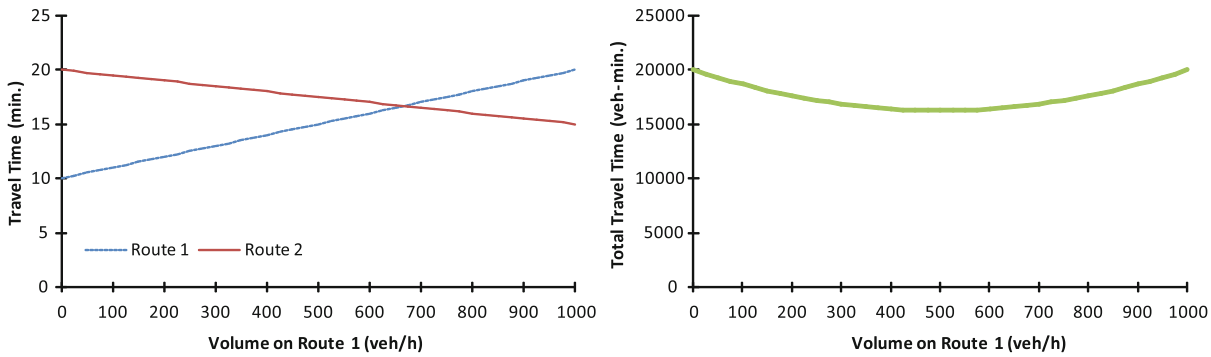

Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment, Figure 4
Variation in Route and System Travel Time for Test Network

resents the extent of possible benefits for a system versus user optimum routing for this particular network and traffic pattern. Figure 4 also illustrates how the average link travel times on routes 1 and 2 vary for the same range of possible routings of traffic between route 1 and 2. In this figure the difference between the travel times on route 1 and 2 (15.0 versus 17.5 min) represents the incentive that exists for vehicles on route 2 to change to route 1. When compared to the user equilibrium routing, the total difference in travel time is composed of two components, which represent the respective increases (route 1) and decreases (route 2) in average travel time that result from a shift from the system to user-optimum routing.

Implementation Issues

While the simple example illustrated the potential benefits of system optimized routings and the incentive that exists for drivers to switch back to the original user equilibrium routings, it is clear that neither an exhaustive enumeration nor an analytical approach (solving the differential equations of the system travel time) are satisfactory for finding the system optimized routings when more than just a few possible routes are available.

Different static traffic assignment algorithms have been developed over the past half century. These methods are broadly divided into non-equilibrium and equilibrium methods. Non-equilibrium methods include all-or-nothing assignment, where all traffic is assigned to a single minimum path between two zones (path that incurs the minimum travel time). Example algorithms for computing minimum paths include models developed by Dantzig [67] and Dijkstra [69]. Other non-equilibrium methods include incremental, iterative, diversion models, multipath assignment [68], and combined models. According to Van Vliet [240] the incremental assignment method (explained later) is capable of reaching an acceptable degree of convergence faster than an iterative method. With regards to diversion models, the most common diversion models include the California diversion curves [145] and the Detroit diversion curves [224]. Alternatively, multipath traffic assignment methods assign traffic stochastically. For example, the Dial method [68] stochastically diverts trips to alternate paths, but trips are not explicitly assigned to routes. Other multipath methods [34,35] assume that users do not know the actual travel times on each link, but a driver's estimate of link travel time is drawn randomly from a distribution of possible times. Finally, combined non-equilibrium models include combining capacity restraint models with probabilistic assignment [206], combining iterative with incremental assignment [253,254,255,256], or combining stochastic with equilibrium assignment [216].

Equilibrium assignment techniques are based on Wardrop's first principle [246]. These were classified by Matsoukis and Michalopoulos [138] into: assignments with fixed demand, assignments with elastic demands, and combined models. Only the first method will be discussed. The equilibrium assignment algorithm is a weighted combination of a sequence of all-or-nothing assignments. This produces a non-linear programming (NLP) problem which is subject to linear constraints. This NLP is very hard to solve and the approach seems to be of limited use for realistically sized equilibrium traffic assignment problems. The NLP problem can be replaced by a much simpler linear approximation and solved using the Frank–Wolfe algorithm [81]. This iterative linearization procedure still involves longer computational times than the iterative procedure. LeBlanc et al. [119] developed an iterative procedure solving one-dimensional searches and LP problems that minimize successively better linear approximations to the non-linear objective function. Nguyen [155] converted

the convex optimization problem into a set of simpler subproblems that could be solved with the convex-simplex method.

One of the most common approaches to implement a user equilibrium traffic assignment involves the use of an incremental traffic assignment technique [121,137,231], [236,237,253,255]. Such a technique breaks down the total traffic demand that is to be loaded onto the network into a number of increments that are each loaded onto the network in turn. Each increment is loaded onto what appears to be the shortest route, after all the previous increments have been loaded. The link travel times are then recalculated, in order to re-compute the fastest route for the next increment to be loaded. When more than one route are to be used for travel between a given origin and destination, the increments are automatically assigned alternatively to each route, when each becomes faster again after previous increments head along the other route. In the end, the extent to which the overall assignment approaches an equilibrium state depends upon the number of increments utilized, with the average final error being roughly proportional to the final increment size.

Van Aerde and Rakha [181,195] demonstrated that the system-optimum traffic assignment can also be solved considering an incremental traffic assignment. Specifically, Van Aerde and Rakha [181,195] recognized the fact that the increase in system travel time caused by the addition of one vehicle is composed of the additional travel time incurred by the subject vehicle and the increase in travel time that is imparted on all other vehicles which are already on the link. While the former quantity is usually already available as a direct or indirect measurement on the link, the derivation of the latter quantity is more subtle. It is a function of the rate of change of the average travel time, per additional vehicle, and the number of vehicles already on the link. In mathematical terms, this is simply the product of the derivative of the travel time versus volume relationship, with respect to volume, multiplied by the volume already present on the link. Consequently, the standard objective function that is utilized in any minimum path algorithm, which searches for the user equilibrium routes, can be replaced by a new objective function that minimizes the total travel time. This routing can be achieved using an incremental assignment of vehicles based on their marginal travel time as opposed to their actual travel time, which results in a system optimum as opposed to a user equilibrium routing, as was demonstrated earlier in Fig. 3. Stated differently at dynamic system optimum, the time-dependent marginal cost on all the paths actually used are equal and less than the marginal cost on any unused paths. In the static case, the path marginal cost (PMC) is the sum of the link marginal cost (LMC). However, in the dynamic case, the PMC evaluation is much more complicated since path flows are not assigned to links on the path simultaneously. However, within the dynamic context, most researchers [84,166] assume that the path flow perturbation travels along the path at the same speed as the additional flow unit. Shen et al. [217] demonstrated that this assumption is not necessarily correct. Furthermore, they presented a solution algorithm for path-based system optimum models based on a new PMC evaluation method. The approach was then tested and validated on a simple network.

Dynamic Traffic Routing

This section describes the mathematical formulations for the static routing problem together with some solution approaches to the problem. Subsequently, the extension of the problem for the dynamic context is presented together with state-of-the-art solution approaches.

The dynamic traffic assignment approach is summarized in Fig. 5 and involves three input variables, namely: dynamic link travel times (in the case of the UE assignment), dynamic marginal travel times (in the case of the SO assignment), and dynamic O-D matrices. In the case of the UE assignment the Bechmann formulation is solved (Eq. (1) if we use a time-dependent static (or quasi static) assignment as will be discussed in detail in the following sections, while in the case of the SO assignment Eq. (12) is solved. Within the static context these formulations are solved analytically using a mathematical program given that the objective function and feasible region are convex. Alternatively, in the dynamic context the objective function is non-convex and thus is more difficult to solve necessitating the use of a modeling approach to solve the problem.

After solving these two formulations the link flows are computed and input into an O-D estimator to provide an estimate of the O-D demand which is then compared to the initial solution. This feedback loop continues until the difference in either link flows or O-D flows is within a desired margin of error or the maximum number of iterations criteria is met.

Mathematical Formulations

Following the notation presented by Sheffi [215] we present the network notations that are used in the mathematical formulation of a static traffic assignment problem. Initially, the variable definitions are presented followed by the vector definitions (bold variables).

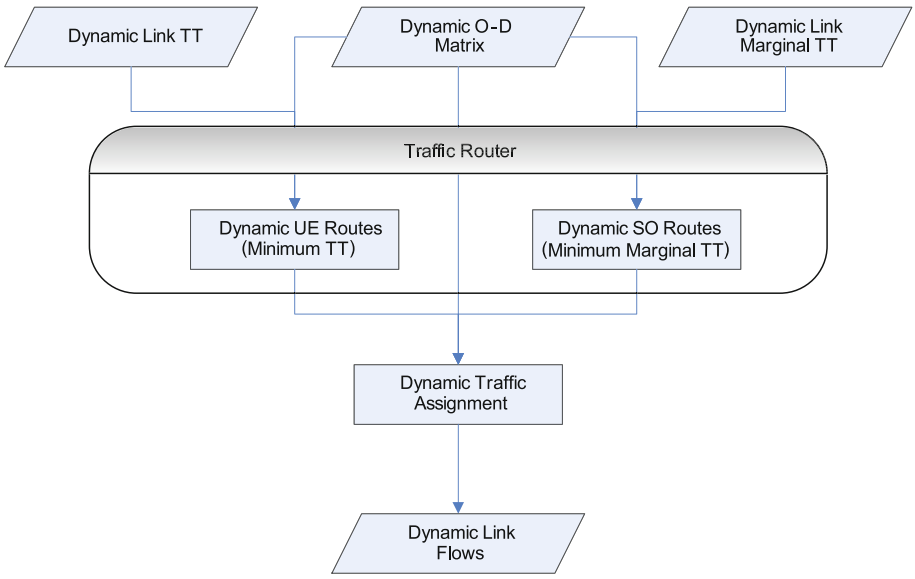

Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment, Figure 5
Dynamic Traffic Assignment Framework

$\boldsymbol{N}$ Set of network nodes
$\boldsymbol{A}$ Set of network arcs (links)
$\boldsymbol{R}$ Set of origin centroids
$\boldsymbol{S}$ Set of destination centroids
$\boldsymbol{k_{rs}}$ Set of paths connecting O-D pair $(r-s)$; $r \in R, s \in S$
$\boldsymbol{x_a}$ Flow on arc (a)
$\boldsymbol{t_a}$ Travel time on arc (a)
$\boldsymbol{f_k^{rs}}$ Flow on path (k) connecting O-D pair $(r-s)$
$\boldsymbol{c_k^{rs}}$ Travel time on path (k) connecting O-D pair $(r-s)$
$\boldsymbol{q_{rs}}$ Trip rate between origin (r) and destination (s)
$\boldsymbol{\delta_{a,k}^{rs}}$ Indicator variable:

$$\delta_{a,k}^{rs} = \begin{cases} 1, & \text{if arc } (a) \text{ is on path } (k) \text{ between} \\ & \text{O-D pair } (r-s) \\ 0, & \text{otherwise} \end{cases}$$

Using vector notations (bold variables) the variables are defined as,

$\boldsymbol{x}$ Vector of flows on all arcs, $= (\ldots, x_a, \ldots)$
$\boldsymbol{t}$ Vector of travel times on all arcs, $= (\ldots, t_a, \ldots)$
$\boldsymbol{f^{rs}}$ Vector of flows on all paths connecting O-D pair r-s, $= (\ldots, f_k^{rs}, \ldots)$
$\boldsymbol{f}$ Matrix of flows on all paths connecting all O-D pairs, $= (\ldots, \boldsymbol{f^{rs}}, \ldots)$
$\boldsymbol{c^{rs}}$ Vector of travel times on all paths connecting O-D pair r-s, $= (\ldots, c_k^{rs}, \ldots)$
$\boldsymbol{c}$ Matrix of travel times on all paths connecting all O-D pairs, $= (\ldots, \boldsymbol{c^{rs}}, \ldots)$
$\boldsymbol{q}$ Origin-destination matrix (with elements $= q_{rs}$)
$\boldsymbol{\Delta^{rs}}$ Link-path incidence matrix (with $\delta_{a,k}^{rs}$ elements) for O-D pair r-s, as discussed below
$\boldsymbol{\Delta}$ Matrix of link-path incidence matrices (for all O-D pairs), $= (\ldots, \Delta^{rs}, \ldots)$

The link-path incident matrix is of size equal to the number of links or arcs in the network (number of rows) and number of paths between origin (r) and destination (s). The element in the ath row, and kth column of $\boldsymbol{\Delta^{rs}}$ is $\delta_{a,k}^{rs}$. In other words, $(\Delta^{rs})_{a,k} = \delta_{a,k}^{rs}$.

The following basic relations are fundamental to the mathematical program formulation:

- A link performance function, which is also known as the volume-delay curve or the link congestion function, represents the relationship between flow and travel time on a link (a) $(t_a = t_a(x_a))$.
- The mathematical program formulations assume that travel time on a given link is only dependent on the flow on the subject link (the model does not capture the effect of opposing flows on the delay of opposed flows), or mathematically

$$\frac{\partial t_a(x_a)}{\partial x_b} = 0 \quad \forall a \neq b \quad \text{and} \quad \frac{\partial t_a(x_a)}{\partial x_a} > 0$$
$$\forall\, a \quad \text{where}\,,\ x_b \text{ is the flow on link } (b)\,.$$

- The travel time on a particular path equals the sum of the travel times on the links comprising that path as

$c_k^{rs} = \sum_a t_a \, \delta_{a,k}^{rs} \quad k \in k_{rs}, \ \forall \ r \in R, \ \forall \ s \in S$ or $\mathbf{c} = \mathbf{t}.\Delta$ considering the vector notation.

- The flow on each link equals the sum of the flows on all paths traversing the subject link as $x_a = \sum_r \sum_s \sum_k (f_k^{rs} \cdot \delta_{a,k}^{rs}) \quad \forall \ a \in A$ or $\mathbf{x} = \mathbf{f}.\Delta^T$.
- The above formula uses the incidence relationships to express link flows in term of path flows, i. e. $x = x(f)$. The incidence relationships also mean that the partial derivative of the link flow can be defined with respect to a particular path flow as follows,

$$\frac{\partial x_a(f)}{\partial f_l^{mn}} = \frac{\partial}{\partial f_l^{mn}} \sum_r \sum_s \sum_k (f_k^{rs} \cdot \delta_{a,k}^{rs}) = \delta_{a,l}^{mn} \ ,$$

$$\text{where } \frac{\partial f_k^{rs}}{\partial f_l^{mn}} = 0 \ \text{ if } k \neq l \ \text{ or } \ r - s \neq m - n \ .$$

Where, f_l^{mn} is the flow on path (l) connecting O-D pair ($m - n$). Since the function $x_a(f)$ includes a flow summation using the subscripts r, s, and k, the variable with respect to which the derivative is being taken is subscribed by m, n, and l, to avoid the confusion in differentiation.

User Equilibrium

As mentioned earlier, the UE model is based on the assumption that each traveler takes the path that minimizes his/her travel time from their origin to their destination, regardless of any effect this might have on the other network users. In other words, at equilibrium, none of the travelers will be able to reduce their travel times by unilaterally switching to another path. This implies that at equilibrium the link flow pattern is such that the travel times on all of the used paths connecting any given O-D pair will be equal. The travel time on all of these used paths will also be less than or equal to the travel time on any of the unused paths.

The mathematical program that represents this model can be cast using Bechmann's transformation as,

$$\begin{aligned} &\text{min.} \ z(x) = \sum_a \int_0^{x_a} t_a(w) \, \mathrm{d}w \\ &\text{s.t.} \quad \sum_k f_k^{rs} = q_{rs} \quad \forall r, s \\ &\qquad\qquad \text{(Flow conservation constraints)} \\ &\qquad f_k^{rs} \geq 0 \quad \forall k, r, s \\ &\qquad\qquad \text{(Non-negativety constraints)} \\ &\qquad x_a = \sum_r \sum_s \sum_k \left(f_k^{rs} \cdot \delta_{a,k}^{rs} \right) \quad \forall a \ . \end{aligned} \tag{1}$$

It is worth mentioning that this formulation "*has been evident in the transportation literature since the mid-1950's, but its usefulness became apparent only when solution algorithms for this program were developed in the late 1960's and early 1970's*" [215].

In order to prove that the solution of Beckmann's transformation program satisfies the user-equilibrium assignment, first the equivalence conditions will be discussed followed by the uniqueness conditions. In the equivalence conditions it will be shown that the first-order conditions for the minimization program are identical to the equilibrium conditions. Whereas, in the uniqueness conditions, it will be shown that the user-equilibrium equivalent minimization program has only one solution. Hence, proving that the solution of Beckmann's transformation program satisfies the user-equilibrium assignment problem.

Equivalency Conditions Beckmann's transformation program is a minimization program with linear equality and non-negativity constraints. In order to find the first-order conditions for such a program, the Lagrangian with respect to the equality constraints can be written as

$$L(f, u) = z[x(f)] + \sum_r \sum_s u_{rs} \left(q_{rs} - \sum_k f_k^{rs} \right) , \tag{2}$$

where u_{rs} denotes the dual variable associated with the flow conservation constraint for O-D pair (r-s). At the stationary point of the Lagrangian, the following first-order conditions have to hold with respect to the path-flow variables and the dual variables. First, with respect to the path-flow variables

$$\begin{aligned} &f_k^{rs} \frac{\partial L(f, u)}{\partial f_k^{rs}} = 0 \quad \forall \ k, r, s \\ &\text{and} \ \frac{\partial L(f, u)}{\partial f_k^{rs}} \geq 0 \quad \forall \ k, r, s \end{aligned} \tag{3}$$

must hold. Alternatively, with respect to the dual variables

$$\frac{\partial L(f, u)}{\partial u_{rs}} = 0 \quad \forall \ r, s \tag{4}$$

must hold. In addition to the following non-negativity constrains,

$$f_k^{rs} \geq 0 \quad \forall \ k, r, s \ . \tag{5}$$

Note that the formulation of this Lagrangian is given in

terms of path flow by using the incidence relationships, $x_a = x_a(f)$.

The partial derivative of $L(x, u)$ with respect to the flow variables f_l^{mn} can be given by

$$\frac{\partial L(f, y)}{\partial f_l^{mn}} = \frac{\partial}{\partial f_l^{mn}} z\left[x(f)\right] + \frac{\partial}{\partial f_l^{mn}} \sum_r \sum_s u_{rs} \left(q_{rs} - \sum_k f_k^{rs}\right) . \quad (6)$$

Using the chain rule the first term can be solved as

$$\begin{aligned} \frac{\partial}{\partial f_l^{mn}} z\left[x(f)\right] &= \sum_{b \in A} \frac{\partial z(x)}{\partial x_b} \cdot \frac{\partial x_b}{\partial f_l^{mn}} \\ &= \sum_b \left[\left(\frac{\partial}{\partial x_b} \sum_a \int_0^{x_a} t_a(w)\, \mathrm{d}w\right) \cdot \left(\frac{\partial x_b}{\partial f_l^{mn}}\right)\right] \\ &= \sum_b t_b \cdot \delta_{b,l}^{mn} = c_l^{mn} . \end{aligned} \quad (7)$$

The second term can be solved as

$$\frac{\partial}{\partial f_l^{mn}} \sum_r \sum_s u_{rs} \left(q_{rs} - \sum_k f_k^{rs}\right) = -u_{mn} , \quad (8)$$

because (a) u_{rs} is not a function of f_l^{mn}; (b) q_{rs} is constant; and

$$\frac{\partial f_k^{rs}}{\partial f_l^{mn}} = \begin{cases} 1, & \text{if } r = m, s = n, \text{ and } k = l \\ 0, & \text{otherwise} \end{cases} .$$

Consequently, Eq. (3) and (4) can be solved to derive

$$\frac{\partial L(f, u)}{\partial f_l^{mn}} = c_l^{mn} - u_{mn} .$$

Hence, we can derive the following first-order conditions,

$$\begin{aligned} & f_k^{rs}(c_k^{rs} - u_{rs}) = 0 && \forall\ k, r, s \\ & c_k^{rs} - u_{rs} \geq 0 && \forall\ k, r, s \\ & \sum_k f_k^{rs} = q_{rs} && \forall\ r, s \\ & f_k^{rs} \geq 0 && \forall\ k, r, s . \end{aligned} \quad (9)$$

We can imply the following from these conditions, (1) The first two conditions, for any path (k) connecting any O-D pair ($r - s$), either (a) The flow on that path, f_k^{rs}, equals zero, in which case, the travel time on this path, c_k^{rs}, will have a value that is greater than or equal to the value of the O-D specific Lagrange multiplier, u_{rs}, or, (b) the flow on that path will have a value (greater than zero), in which case, the travel time on this path will have a value equal to the value of the O-D specific Lagrange multiplier, u_{rs}. In both cases, the value of the O-D specific Lagrange multiplier is always less than or equal to the travel time on all other paths connecting the same O-D pair. Hence, this value of the Lagrange multiplier is the minimum path travel time between this O-D pair thus proving that the solution of Beckman's transformation program satisfies the user-equilibrium assignment.

The last two conditions satisfy the flow conversation and non-negativity constraints, respectively. The proof can further be explained as follows, paths connecting O-D pair ($r - s$) can be divided into two groups, (1) Paths with zero flow, and are characterized by a travel time which is either greater than or equal to the minimum travel time; and (2) Paths with non-negative flows, and are characterized by minimum travel times. Thus, confirming the user-equilibrium notion which states that no user can improve his/her travel times by unilaterally changing their routes.

The above proves that user-equilibrium conditions are satisfied at any stationary point of Beckman's transformation program. The following section proves that there is only one solution for Beckman's transformation program. It proves that Beckman's transformation program has only one stationary point, and that this point is a minimum.

Uniqueness Condition In order to prove that Beckmann's transformation program has only one solution, it is sufficient to prove that the objective function is strictly convex in the vicinity of the solution point, convex everywhere else (within the feasible solution region), and that the feasible region (defined by the constraints) is convex.

It is known that linear equality constraints ensure a convex feasible region, and that the addition of the non-negativity constrains does not alter this fact. The convexity of the objective function, with respect to link flows, can be proven in two different ways. The fist way can be achieved by the application of the properties of convex functions, on the link-performance functions. On the other hand, the second proof is achieved by proving that the Hessian matrix of the objective function is positive definite.

Link performance functions are known to be continuously increasing functions. Hence, link-performance functions are convex functions. The objective function equals the summation of the integral of the link-perfor-

mance functions of all links. Properties of convex functions state that integrals of convex functions are also convex functions, and that the summation of convex functions is also a convex function. Hence, proving that the objective function is convex everywhere. Subsequently proving that there is only one solution for Beckman's transformation program, with respect to link flows, and that solution is a minimum.

Recalling that

$$\frac{\partial^2 z(x)}{\partial x_m \partial x_n} = \frac{\partial t_m(x_m)}{\partial x_n} = \begin{cases} 1, & \text{for } m = n \\ 0, & \text{otherwise} \end{cases}, \quad (10)$$

the Hessian matrix for the objective function can be calculated to be as follows,

$$\nabla^2 z(x) = \begin{bmatrix} \frac{\partial^2 z(x)}{\partial x_1^2} & \frac{\partial^2 z(x)}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 z(x)}{\partial x_A \partial x_1} \\ \frac{\partial^2 z(x)}{\partial x_1 \partial x_2} & \frac{\partial^2 z(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 z(x)}{\partial x_A \partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 z(x)}{\partial x_1 \partial x_A} & \frac{\partial^2 z(x)}{\partial x_2 \partial x_A} & \cdots & \frac{\partial^2 z(x)}{\partial x_A^2} \end{bmatrix}$$
$$= \begin{bmatrix} \frac{dt_1(x_1)}{dx_1} & 0 & \cdots & 0 \\ 0 & \frac{dt_2(x_2)}{dx_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{dt_A(x_A)}{dx_A} \end{bmatrix}. \quad (11)$$

Obviously, the matrix is definite positive, proving that the objective function is strictly positive, and subsequently, has a unique minimum solution.

It is worth mentioning that the Beckman's transformation program is not convex with respect to path flows, and therefore, the equilibrium conditions themselves are not unique with respect to path flows. In other words, while there is actually only one unique solution for link flows, there are an infinite number of paths flows solutions that would produce this unique link flows solution, which raises the need to compute the most likely of these solutions using a synthetic O-D estimator as was described earlier and will be discussed later in more detail.

System Optimum

As mentioned earlier, the SO model attempts to minimize the total travel time spent in the network. Hence, it might assign certain trips to a slightly longer path (in terms of travel time), in order to reduce the travel time of other user trips by a value which is greater than the value of the increased travel time, and thus achieving a reduced total network travel time. Opposite to user equilibrium, in the system optimum state, users can reduce their travel times by unilaterally switching to alternative paths, which becomes a challenge to implement such a strategy. Therefore, the solution is not stable. SO network travel time mainly serves as a yardstick that measures the performance of a network.

The mathematical program that represents this model can be written as follows,

$$\begin{aligned} \text{min.} \quad & \tilde{z}(x) = \sum_a x_z \cdot t_z(x_a) \\ \text{s.t.} \quad & \sum_k f_k^{rs} = q_{rs} \quad \forall\, r, s \\ & \qquad \text{(Flow conservation constraints)} \\ & f_k^{rs} \geq 0 \quad \forall\, k, r, s \\ & \qquad \text{(Non-negativity constraints)} \\ & x_a = \sum_r \sum_s \sum_k \left(f_k^{rs} \cdot \delta_{a,k}^{rs} \right) \quad \forall\, a\,. \end{aligned} \quad (12)$$

As can be seen, the only difference between user-equilibrium and system optimum programs is the objective function. The SO optimum objective function equals the summation of the products of the travel time on each link times the traffic volume assigned to this link, for all links. Hence, it works on minimizing the total travel time experienced by all vehicles traveling on all links of the networks. On the other hand, the UE objective function equaled the summation of only the travel times of all links.

It can also be seen that the constraints in the SO model are exactly the same as in the UE model. Consequently, similar to the case with the user-equilibrium equivalent program, the solution of this program can be found by solving for the first-order conditions for a stationary point of the following Lagrangian

$$\tilde{L}(f, u) = \tilde{z}\left[x\left(f\right)\right] + \sum_r \sum_s \tilde{u}_{rs} \left(q_{rs} - \sum_k f_k^{rs} \right), \quad (13)$$

where $\tilde{u}_{rs}$ denotes the dual variable associated with the flow conservation constraint for O-D pair (r-s). At the stationary point of the Lagrangian, the following first-order conditions have to hold with respect to the path-flow variables

$$\begin{aligned} & f_k^{rs} \frac{\partial \tilde{L}(f, \tilde{u})}{\partial f_k^{rs}} = 0 \quad \forall\, k, r, s \\ \text{and} \quad & \frac{\partial \tilde{L}(f, \tilde{u})}{\partial f_k^{rs}} \geq 0 \quad \forall\, k, r, s\,. \end{aligned} \quad (14)$$

With respect to the dual variables

$$\frac{\partial \tilde{L}(f, u)}{\partial \tilde{u}_{rs}} = 0 \quad \forall\, r, s\,. \quad (15)$$

In addition to the non-negativity constraints

$$f_k^{rs} \geq 0 \quad \forall\ k, r, s\,. \tag{16}$$

Note that, the formulation of this Lagrangian is given in terms of path flow by using the incidence relationships, $x_a = x_a(f)$.

The partial derivative of $\tilde{L}(x, \tilde{u})$ with respect to the flow variables f_l^{mn} can be given by

$$\frac{\partial \tilde{L}(f, \tilde{u})}{\partial f_l^{mn}} = \frac{\partial}{\partial f_l^{mn}} \tilde{z}\left[x\left(f\right)\right] + \frac{\partial}{\partial f_l^{mn}} \sum_r \sum_s \tilde{u}_{rs} \left(q_{rs} - \sum_k f_k^{rs}\right)$$

$$\begin{aligned}
\frac{\partial}{\partial f_l^{mn}} \tilde{z}\left[x\left(f\right)\right] &= \sum_{b \in A} \frac{\partial \tilde{z}\left(x\right)}{\partial x_b} \cdot \frac{\partial x_b}{\partial f_l^{mn}} \\
&= \sum_b \left[\left(\frac{\partial}{\partial x_b} \sum_a x_a \cdot t_a\left(x_a\right)\right) \frac{\partial x_b}{\partial f_l^{mn}}\right] \\
&= \sum_b \left[t_b\left(x_b\right) + x_b \frac{\mathrm{d}t_b\left(x_b\right)}{\mathrm{d}x_b}\right] \delta_{b,l}^{mn} \\
&= \sum_b \tilde{t}_b \cdot \delta_{b,l}^{mn} = \tilde{c}_l^{mn}\,.
\end{aligned}$$

Assuming $\tilde{t}_b = t_b\left(x_b\right) + x_b \frac{\mathrm{d}t_b(x_b)}{\mathrm{d}x_b}$ and $\partial/(\partial f_l^{mn}) \sum_r \sum_s \tilde{u}_{rs}\left(q_{rs} - \sum_k f_k^{rs}\right) = -\tilde{u}_{mn}$ because (a) u_{rs} is not a function of f_l^{mn}; (b) q_{rs} is constant; and (c)

$$\frac{\partial f_k^{rs}}{\partial f_l^{mn}} = \begin{cases} 1, & \text{if } r = m, s = n, \text{ and } k = l \\ 0, & \text{otherwise}\,. \end{cases}$$

Therefore $(\partial \tilde{L}(f, \tilde{u}))/(\partial f_l^{mn}) = \tilde{c}_l^{mn} - \tilde{u}_{mn}$.

Where, $\tilde{t}_a$ is a summation of two terms, (1) $t_a(x_a)$, which is the travel time experienced by this additional driver when the total link flow is (x_a) and (2) $\frac{\mathrm{d}t_a(x_a)}{\mathrm{d}x_a}$, which is the additional travel time burden that this driver inflicts on each one of the other (x_a) travelers already using link a.

In summary, it can be interpreted as the marginal contribution of an additional traveler – or an infinitesimal flow unit – on the ath link to the total travel time on that link.

Substituting the above results into Eqs. (14) through (16), we get the following first-order conditions

$$\begin{aligned}
&f_k^{rs}(\tilde{c}_k^{rs} - \tilde{u}_{rs}) = 0 && \forall\ k, r, s \\
&\tilde{c}_k^{rs} - \tilde{u}_{rs} \geq 0 && \forall\ k, r, s \\
&\sum_k f_k^{rs} = q_{rs} && \forall\ r, s \\
&f_k^{rs} \geq 0 && \forall\ k, r, s
\end{aligned}$$

Similar to the interpretation of the user equilibrium conditions, the following can be implied from the above, (1) the first two Conditions, for any path (k) connecting any O-D pair (r-s), either (a) the flow on that path, f_k^{rs}, equals zero whenever the marginal total travel time on this path, $\tilde{c}_k^{rs}$, will have a value that is greater than or equal to the value of the O-D specific Lagrange multiplier, $\tilde{u}_{rs}$, or, (b) the flow on that path, f_k^{rs}, will have a value (greater than zero) whenever the marginal total travel time on this path, $\tilde{c}_k^{rs}$, will have a value equal to the value of the O-D specific Lagrange multiplier, $\tilde{u}_{rs}$. In both cases, the value of the O-D specific Lagrange multiplier is always less than or equal to the marginal total travel time on all other paths connecting the same O-D pair, i. e. the value of the Lagrange multiplier is the marginal travel time on the used paths between this O-D pair. (2) The last two conditions satisfy the flow conversation and non-negativity constraints, respectively.

The proof can further be explained as follows, paths connecting O-D pair (r-s) can be divided into two groups, (1) Paths with zero flow, and are characterized by a total marginal travel time which is either greater than or equal to the marginal travel time of the used networks (or the Lagrange multiplier). (2) Paths with non-negative flows, and are characterized by equal marginal travel times.

In order to prove that the SO program has only one solution, as was the case with the user equilibrium program, it is sufficient to prove that the objective function is strictly convex in the vicinity of the solution point, convex everywhere else (within the feasible solution region), and that the feasible region (defined by the constraints) is convex.

It is known that linear equality constraints assure a convex feasible region, and that the addition of the non-negativity constrains does not alter this fact. The convexity of the objective function, with respect to link flows, can be proven if the Hessian matrix of the objective function is positive definite.

Recalling that,

$$\begin{aligned}
\frac{\partial^2 \tilde{z}\left(x\right)}{\partial x_m \partial x_n} &= \frac{\partial}{\partial x_n} \cdot \left[\frac{\partial}{\partial x_m} \sum_a x_a \cdot t_a\left(x_a\right)\right] \\
&= \frac{\partial}{\partial x_n} \left[t_m\left(x_m\right) + x_m \frac{\mathrm{d}t_m\left(x_m\right)}{\mathrm{d}x_m}\right] \\
&= \begin{cases} 2\frac{\mathrm{d}t_n(x_n)}{\mathrm{d}x_n} + x_n \frac{\mathrm{d}^2 t_n(x_n)}{\mathrm{d}x_n^2}\,, & \text{for } m = n \\ 0\,, & \text{otherwise} \end{cases}.
\end{aligned}$$

As in the user equilibrium program, the Hessian matrix for the objective function can be calculated to be as

$$\nabla^2 z(x) = \begin{bmatrix} 2\frac{\mathrm{d}t_1(x_1)}{\mathrm{d}x_1} + x_1\frac{\mathrm{d}^2 t_1(x_1)}{\mathrm{d}x_1^2} & 0 & \cdots & 0 \\ 0 & 2\frac{\mathrm{d}t_2(x_2)}{\mathrm{d}x_2} + x_n\frac{\mathrm{d}^2 t_2(x_2)}{\mathrm{d}x_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 2\frac{\mathrm{d}t_A(x_A)}{\mathrm{d}x_A} + x_A\frac{\mathrm{d}^2 t_A(x_A)}{\mathrm{d}x_A^2} \end{bmatrix}.$$

This Hessian matrix is positive definite if all the diagonal terms are positive, which is manifested if the link performance functions are positive. Based on the earlier discussion in the user equilibrium section, it was demonstrated that link-performance functions are convex, and thus demonstrating that the objective function is strictly positive, and subsequently, has a unique minimum solution – with respect to link flows.

It is worth noting that user equilibrium and system optimum produce identical results in any of the following: (1) If congestion effects were ignored, i. e. $t_a(x_a) = t_a'$ (a constant value per arc) or (2) In case of minimal traffic volumes, that would have negligible effects on the arc specific travel times, $t_a(x_a)$.

Dynamic Traffic Assignment Solution Approach

The extension from a static to a dynamic formulation involves the introduction of two time indices into the formulation. The first time index identifies the time at which the path flow leaves its origin while the second time index identifies when the path flow is observed on a specific link. Unfortunately, the introduction of these time indices deems the objective function non-convex and thus two approaches are considered in solving this problem. The first approach is to divide the analysis period into time intervals while assuming that conditions are static within each time interval (time-dependent static or quasi static). The duration of these intervals are network dependent and should be sufficiently long enough to ensure that motorists can complete their trip within the time interval. The static UE and SO mathematical programs can then be solved for each time interval using the standard static formulations that were presented earlier. The mathematical solution approach requires a closed form solution using an analytical modeling approach. Analytical modeling of the network aims at finding the correct mathematical presentation of DTA models that would realistically reflect the real world problem with minimum compromises in the modeling of traffic behavior. The solution of such models should guarantee theoretical existence, uniqueness, and stability. Analytical models are valuable because theoretical insights can be analytically derived. Different analytical network modeling may include mathematical programming formulations, optimal control formulations, and variational inequality formulations [180]. Literature within this area of research is extensive. In general, models within the group may be classified into [180]: i) mathematical programming formulations, as the works of Merchant and Nemhauser [142,143], Ho [94], Carey [39,40,41], Janson [98,99], Birge and Ho [28], Ziliaskopoulos [262], Carey and Subrahmanian [42]; ii) optimal control formulations, as in the works of Friesz et al. [83], Ran and Shimazaki [204,205], Wie [249], Ran et al. [202], Boyce et al. [30]; and iii) variational inequality formulations, as with the works of Dafermos [64], Friesz et al. [82], Wie et al. [37], Ran and Boyce [203], Ran et al. [201], Chen and Hsueh [96].

Alternatively, the second approach involves the use of a simulation solution approach. Simulation models on the other hand, in spite of solving the DTA problem within a simulation environment, still use some form of mathematical abstraction of the problem. According to Peeta [180], "*the terminology simulation-based models may be a misnomer. This is because the mathematical abstraction of the problem is a typical analytical formulation, mostly of the mathematical programming variety in the current literature. However, the critical constraints that describe the traffic flow propagation, and the spatio-temporal interactions, such as the link-path incidence relationships, flow conservation, and vehicular movements are addressed through simulation instead of analytical evaluation while solving the problem. This is because analytical representations of traffic flows that adequately replicate traffic theoretic relationships and yield well-behaved mathematical formulations are currently unavailable. Hence, the term simulation-based primarily connotes the solution methodology rather than the problem formulation. A key issue with simulation-based models is that theoretical insights cannot be analytically derived as the complex traffic interactions are modeled using simulation. On the other hand, due to the inherently ill-behaved nature of the DTA problem, notions of convergence and uniqueness of the associated solution may not be particularly meaningful from a practical standpoint. In addition, due to their better fidelity vis-à-vis realistic traffic modeling, simulation-based models have gained greater acceptability in the context of real-world deployment*".

One of the early simulation DTA tools is the Simulation and Assignment in Urban Road Networks (SATURN) approach. The SATURN algorithm utilizes an equilibrium technique which optimally combines a succession of all-or-nothing assignments (i. e. it is an iterative equilibrium assignment based on iterative traffic loading) [29,87,241]. This model treats platoons of traffic rather than individual vehicles but delays vehicles but delays at intersections are treated in considerable detail. The model consists of two parts: a simulation component and a traffic assignment component. The traffic simulation component fits a delay-flow power curve to three points, namely: zero flow, current flow, and capacity. This delay-flow curve is used by the assignment model to route vehicles. For each traffic signal four cyclic flow profiles are considered: the IN pattern, the ARRIVE pattern, the ACCEPT pattern, and the OUT pattern. SATURN can account for delays caused by opposing flows, delays caused by vehicles on the same roadway, the shape of the arriving platoon, the effect of traffic signal phasing structure and offsets, and individual lane capacities. Arrival rates that exceed capacity are assumed to form queues that build up at constant rates. SATURN can model networks at two levels of detail, namely: inner and buffer. The model was used in studies in the U.K., Australia, and New Zealand. The limitations of the model include: (a) it assumes steady-state conditions for periods of 15–30 minutes and thus is a time-dependent dynamic assignment approach; (b) queues are modeled vertically and thus they cannot spillback to upstream intersections; (c) it is unsuitable for freeways; (d) it cannot model over-saturated conditions explicitly.

Another early simulation DTA that was developed in the late 1970s is the CONTRAM model (CONtinuous TRaffic Assignment Model). CONTRAM is similar to SATURN in that it combines traffic assignment with traffic simulation [121]. CONTRAM is a computer based time-varying assignment and queuing model. Unlike SATURN, vehicles are grouped within CONTRAM into packets where each packet is treated in the same way as a single vehicle when assigning it to its minimum path. Time varying flow conditions are modeled by dividing the simulation period into a number of consecutive time intervals, which need not be of the same length, and the packets leave each origin at a uniform rate through each such interval. The assignment is an incremental iterative technique where during the first iteration; packets are routed based on link-travel times of previous packets. However, in successive iterations, they are routed based on link travel times that reflect a weighting of travel times during previous iterations and previous packets. Prior to routing a packet, the packet volume is removed from its previously used links. An advantage of this assignment model is that it takes into account the effects of packets leaving later on the routing of packets which leave earlier. Thus, it decides upon the path based on a fully loaded network, rather than on one in which has only been loaded to the extent of any previous increments. This model is more dynamic than most models because vehicles are able to change their routing decisions while en-route, if traffic conditions alter. Satisfactory convergence is usually achieved in 5 to 10 iterations. The limitations of the model are: (a) introduction of signal optimization makes the model unable to converge; (b) vehicles queue vertically on a link; (c) no limitation of the storage capacity of a link is introduced; (d) it is unsuitable for freeway networks; and (e) it can only assign vehicles based on Wardrop's first principle.

A number of contemporary DTA models were developed using the basic CONTRAM concept, including the INTEGRATION [195,196,208,209,210,211,231,232,233,234,236,237], DYNASMART [2,3,4,5,53,100,101,102,170,225] and DYNAMIT [19,22,113,257] modeling approaches. In this section the INTEGRATION dynamic traffic assignment and modeling framework is briefly described as an example illustration of a microscopic traffic assignment and simulation approach. The INTEGRATION model is similar to the CONTRAM model in that it models individual vehicles (packets of unit size). Unlike, other traffic assignment models, the INTEGRATION traffic simulation logic is microscopic in that it models vehicles at a deci-second level of resolution. The software combines car-following, vehicle dynamics, lane-changing, energy, and emission models. Thus, mobile source emissions can be directly estimated from instantaneous speed and acceleration levels. Furthermore, the traffic and emission modeling modules have been tested and validated extensively. For example, the software, which was developed over the past two decades, has not only been validated against standard traffic flow theory [185,199], but has also been utilized for the evaluation of real-life applications [190,198]. Furthermore, the INTEGRATION software offers unique capability through the explicit modeling of vehicle dynamics by computing the tractive and resistance forces on the vehicle each deci-second [187,188,193].

The INTEGRATION software uses car-following models to capture the longitudinal interaction of a vehicle and its preceding vehicle in the same lane. The process of car-following is modeled as an equation of motion for steady-state conditions (also referred to as stationary conditions in some literature) plus a number of constraints that govern the behavior of vehicles while moving from one steady-state to another (decelerating and/or acceler-

ating). The first constraint governs the vehicle acceleration behavior, which is typically a function of the vehicle dynamics [185,192]. The second and final constraint ensures that vehicles maintain a safe position relative to the lead vehicle in order to ensure asymptotic stability within the traffic stream. A more detailed description of the longitudinal modeling of vehicle motion is provided by [194]. Alternatively, lane-changing behavior describes the lateral behavior of vehicles along a roadway segment. Lane changing behavior affects the vehicle car-following behavior especially at high intensity lane changing locations such as merge, diverge, and weaving sections.

The INTEGRATION model provides for 7 basic traffic assignment/ routing options: (a) Time-Dependent Method of Successive Averages (MSA); (b) Time-Dependent Sub-Population Feedback Assignment (SFA); (c) Time-Dependent Individual Feedback Assignment (IFA); (d) Time-Dependent Dynamic Traffic Assignment (DTA); (e) Time-Dependent Frank–Wolf Algorithm (FWA); (f) Time-Dependent External; and (g) Distance Based Routing.

The derivation of a time series of MSA traffic assignments involves analyzing each time slice in isolation of either prior or subsequent time slices (time-dependent static or quasi static). The link travel times, upon which the route computations are based, are estimated based on the prevailing O-D pattern and an approximate macroscopic travel time relationship for each link. Multiple paths are computed in an iterative fashion, where the tree for each subsequent iteration is based on the travel times estimated during the previous iterations. The weight assigned to each new tree is 1/N where N is the iteration number.

In the case of the feedback assignment vehicles base their routings on the experience of previous vehicle departures (incremental traffic assignment). In the case of the SFA assignment all drivers of a specific type are divided into 5 sub-populations each consisting of 20% of all drivers. The paths for each of these sub-populations are then updated every t seconds during the simulation based on real-time measurements of the link travel times for that specific vehicle class. The value of t is a user-specified value. Furthermore, the minimum path updates of each vehicle sub-population are staggered in time, in order to avoid having all vehicle sub populations update their paths at the same time. This results in 20% of the driver paths being updated every t/5 seconds. In the case of the IFA assignment all paths are customized to each individual driver and may therefore be unique relative to any other drivers. This incremental traffic assignment accounts the effect of earlier vehicle departures on the travel time of later which is very similar to the CONTRAM approach. However, unlike CONTRAM no iterations are made to reassign all the vehicles.

The INTEGRATION DTA computes the minimum path for every scheduled vehicle departure, in view of the link travel times anticipated in the network at the time the vehicle will reach these specific links. The anticipated travel time for each link is estimated based on anticipated link traffic volumes and queue sizes. This routing involves the execution of a complete mesoscopic DTA model prior to the simulation of the traffic. During this DTA, the routes of all vehicles are computed using the above procedure. Upon completion of this DTA, the actual simulation simply implements the routings computed as per the DTA.

Clearly the validity of any of these modeling approaches hinges on the ability of the traffic simulation model to reflect real-life behavior and capture all the complexities of traffic modeling. Clearly, no modeling approach can claim that it is capable of capturing every aspect of empirical traffic flow behavior and thus the output of such models should be interpreted within the context of how they model the spatio-temporal behavior of drivers.

It should also be noted that the models that were described in this section are heuristic approaches attempting to solve the mathematical formulations that were presented earlier and thus there is no guarantee that they converge to a single (unique) solution for UE and/or SO assignment problems for a complex dynamic network. Furthermore, it is not clear if drivers actually attain such an equilibrium state in such networks. Consequently, research is needed to study and develop models on how drivers select routes, how they respond to the dissemination of traffic information, and how their routing decisions vary temporally in the short- and long-term.

Traffic Modeling

A key component of a DTA is the modeling of traffic stream behavior in order to predict traffic states into the near future and compute link travel times and various measures of effectiveness, as was illustrated earlier in Fig. 2. This section briefly summarizes the various state-of-the-practice approaches to traffic modeling. Researchers have demonstrated that these approaches are unable to predict empirical spatio-temporal aspects [106] observed in the field. Conversely, others have argued that these approaches, while not perfect, capture the main aspects of empirical data. While our objective is not to argue either way, it is sufficient to note that these tools are being used by transportation professionals to assess dynamic networks and thus are presented in this section. These

approaches can be classified into three categories, which include: macroscopic, mesoscopic, and microscopic approaches. Each of these approaches is briefly described in this section. Again the description is by no means comprehensive but does provide a general overview of these approaches. The interested reader should consider reading the wealth of literature on this topic.

Prior to describing the specifics of the various modeling approaches it is important to note that with the exception to research conducted by Kerner [106,107], most existing approaches are based on the famous one-dimensional kinematic waves (KW) theory, which was proposed by Lighthill and Witham [126] and independently proposed by Richards [207]. The key postulate of the theory is that there exists a functional relationship between the traffic stream flow rate q and density k that might vary with location x but does not vary with time t (this contradicts the definition of dynamic given that within a dynamic process variables vary spatially and temporarily). It should be noted that in microscopic approaches, as will be described later, the fundamental diagram various temporally as a function of the traffic composition, thus overcoming some of the drawbacks o f this approach. The fundamental hypothesis of all traffic flow theories is the existence of a site-specific unique relationship between traffic stream flow and traffic stream density, commonly known as the fundamental diagram, the traffic stream motion model, or the car-following model at the microscopic level. The assumption is that all steady-state model solutions lie on the fundamental diagram and thus are referred to as fundamental diagram approaches [106]. Given that traffic stream space-mean speed can be related to traffic stream flow and density, a unique speed-flow-density relationship (in the macroscopic approach) is derived from the fundamental diagram for each roadway segment. This relationship can also be cast at the micro-level by relating the vehicle speed to its spacing, given that vehicle spacing is the inverse of traffic stream density. Some researchers have argued that the fundamental diagram approach cannot capture the spontaneous traffic stream failure that is observed in the field and thus these researchers have proposed other theories.

One of these theories is the three-phase traffic flow theory proposed by Kerner [106,107], which attempts to explain empirical spatiotemporal features of congested patterns. The theory divides traffic into three phases: free-flow, synchronized flow, and wide moving jams. The free-flow phase is consistent with the uncongested regime on a fundamental diagram and thus is not discussed further. The *synchronized flow* phase involves continuous traffic flow with no significant stoppage. The word "flow" reflects this feature. Within this phase there is a tendency towards synchronization of vehicle speeds and flows across the different lanes on a multilane roadway, and thus comes the name "synchronized." This synchronization of speeds is a result if the relatively low probability of passing within this phase. The third phase, *wide moving jam*, is a phase that involves traffic jams that propagate through other states of traffic flow and through any bottleneck while maintaining the velocity of the downstream jam front. The phrase "moving jam" reflects the propagation as a whole localized structure on a road. To distinguish wide moving jams from other moving jams, which do not characteristically maintain the mean velocity of the downstream jam front, Kerner uses the term wide moving jam. Kerner indicates that if a moving jam has a width (in the longitudinal direction) considerably greater than the widths of the jam fronts, and if vehicle speeds inside the jam are zero, the jam always exhibits the characteristic feature of maintaining the velocity of the downstream jam front.

Kerner distinguishes his three-phase traffic flow theory from fundamental diagram approaches in a number of aspects. He demonstrates that the fundamental diagram approach cannot capture two key empirically observed phenomena in traffic, namely: (a) the probabilistic nature of free-flow to synchronized flow transition (flow breakdown), and (b) the spontaneous formation of general patterns (GP), which include moving and wide moving jams. Alternatively, it could be hypothesized that by modeling individual driver behavior (micro or nano modeling), capturing vehicle acceleration constraints, and introducing stochastic differences between drivers that this may be sufficient to model these two key phenomena.

Macroscopic Modeling Approaches

In order to solve for the three traffic stream variables (q, k, and u) three equations are introduced. The first is the functional relationship between flow and density, or what is commonly known as the fundamental diagram. Typical functions include the Pipes triangular function (3 parameters), the Greenshields parabolic function [85] (2 parameters), or the Van Aerde [188] function (4 parameters). The second equation is the flow conservation equation (equation of continuity) that can be expressed as $\partial k(x,t)/\partial t + \partial q(x,t)/\partial x = 0$, considering no entering or exiting traffic. The third and final equation relates the traffic stream flow rate (q) to the traffic stream density (k) and space-mean speed (u) as $q = ku$. The numerical solution of the KW problem involves partitioning the network into small cells of length Δx and discretizing time into steps of duration Δt. For numerical stability $\Delta x = u\Delta t$.

The problem is solved by stepping through time and solving for the variables in every cell using the incremental transfer (IT) principle (essentially explicit finite difference method). Extensions to the standard KW solution have introduced IT solutions for each lane along a freeway where the freeway is modeled as a set of interacting streams linked by lane changes. Lane-changing vehicles can be treated as a fluid that can accelerate instantaneously, however this approach does not capture the reduction in capacity that is associated with lane changes. Consequently, further improvements have been introduced through the use of a hybrid approach [115] that combines microscopic and macroscopic models. Specifically, slow vehicles are treated as moving bottlenecks in a single KW stream, while lane changing vehicles are modeled as discrete particles with constrained motion. The model requires identifying a lane-change intensity parameter in addition to the functional relationship parameters that were described earlier. It is not clear how such a parameter is derived. The major drawbacks of this modeling approach are that it does not account for the dynamic changes in roadway capacity (e. g. the capacity of a weaving section varies as a function of the traffic composition), it cannot capture the spontaneous traffic stream failure that is observed in the field, it cannot capture the impact of opposing flows on the traffic behavior of an opposed flow (e. g. how the capacity of an opposed left turn movement is affected by the opposing through movement), and it ignores the stochastic nature of traffic.

Mesoscopic Modeling Approaches

The mesoscopic analysis tracks individual vehicles as they travel through the network along a sequence of links that are determined by the traffic assignment. The level of tracking involves computing the vehicle's travel speed on each link based upon the density on the link together with a user specified speed/density relationship. The vehicle is then held on the link for the duration of its travel time. At the vehicle's scheduled departure time, the vehicle is allowed to exit the link if the link privileges permit it to leave; otherwise, the vehicle is held on the link until the link privileges so permit. Link exit privileges may be controlled by traffic signals at the downstream end of the link or by any queues that may be present on the lane. Queues are stored for each lane separately to account for any queue length differentials that may occur (e. g., longer queues on left turn opposed lanes). The mesoscopic analysis captures the operational level of detail (e. g., the reduction in lane capacity as a result of an opposed flow) without having to track each vehicle's instantaneous speed profile. This means that the computational requirements for such a type of modeling are more than that required by a macroscopic analysis, but less than that required by a microscopic analysis. The INTEGRATION 1.50, DynaSMART, and DynaMIT models are examples of such modeling approaches. This approach suffers from similar drawbacks as identified in the macroscopic analysis procedures, namely an inability to capture correct spatiotemporal propagation of congestion, a failure to capture dynamic changes in capacity, a failure to capture for spontaneous breakdown in a traffic stream, and failure to capture the stochastic nature of traffic.

Microscopic Modeling Approaches

The third approach to modeling traffic is the microscopic analysis, which tracks each vehicle as it travels through the network on a second-by-second or deci-second level of resolution using detailed car-following and lane-changing models. Microscopic simulation software use car-following models to capture the longitudinal interaction of a vehicle and its preceding vehicle in the same lane. The process of car-following is modeled as an equation of motion for steady-state conditions (also referred to as stationary conditions in some literature) plus a number of constraints that govern the behavior of vehicles while moving from one steady-state to another (decelerating and/or accelerating). The first constraint governs the vehicle acceleration behavior, which is typically a function of the vehicle dynamics. The second and final constraint ensures that vehicles maintain a safe position relative to the lead vehicle in order to ensure asymptotic stability within the traffic stream. A more detailed description of the longitudinal modeling of vehicle motion is provided by Rakha et al. [194]. While there are a number of commercially available software packages that simulate traffic microscopically (CORSIM, Paramics, FREEVU, VISSIM, AIMSUN2, and INTEGRATION), these approaches are computationally intensive and cannot run in real-time. The INTEGRATION software has been able to capture the stochastic nature of traffic stream capacity by randomly modeling vehicle-specific car-following models. Furthermore, the model captures the capacity loss associated with recovery from breakdown through the vehicle acceleration constraints. The stochastic nature of car-following and lane-changing behavior may allow the model to capture spontaneous breakdown in traffic stream flow.

The amount of computation and memory necessary for simulating a large transportation network at a level of detail down to an individual traveler and an individual vehicle may be extensive. Hence a microscopic massively

parallel simulation approach entitled "cellular automata" (CA) is sometimes proposed to simulate large networks. The cellular automata approach essentially divides every link on the network into a finite number of cells. At a one second time step, each of these "cells" is scanned for a vehicle presence. If a vehicle is present, the vehicle position is advanced, either within the cell or to another cell, using a simple rule set [148,149,150]. The rule set is made simple to increase the computational speed necessary for a large simulation. Vehicles are moved from one grid cell to another based on the available gaps ahead, with modifications to support lane changing and plan following, until they reach the end of the grid. There, they wait for an acceptable gap in the traffic or for protection at a signal before moving through the intersection onto the next grid. This continues until each vehicle reaches its destination, where it is removed from the grid. Reducing the size of the "cell", expanding the rule set, and adding vehicle attributes increases the fidelity of the simulator, but also greatly affects the computational speed. The size of 7.5 meters in length and a traffic lane in width is often chosen as a default size for the "cell" as was applied with the TRANSIMS software [149]. This approach suffers from a number of drawbacks including the inability to capture the dynamic nature of roadway capacity, the inability to capture spontaneous breakdown of traffic stream, and the inability to capture opposing flow impacts on opposed flow saturation flow rates (e. g. the impact an opposing through movement flow has on the capacity of a permitted left turn movement that has to find a gap in this opposing flow).

Dynamic Travel Time Estimation

As was demonstrated earlier in the paper, the DTA requires arc (link) travel times in order to compute minimum paths. There are several systems commercially available that are capable of estimating real-time travel times. These can be broadly classified into spot speed measurement systems, spatial travel time systems, and probe vehicle technologies. Spot speed measurement systems, specifically inductance loop detectors, have been the main source of real-time traffic information for the past two decades. Other technologies for measuring spot speeds have also evolved, such as infrared and radar technologies. Regardless of the technology, the spot measurement approaches only measure traffic stream speeds over a short roadway segment at fixed locations along a roadway. These spot speed measurements are used to compute spatial travel times over an entire trip using space-mean-speed estimates. In addition, new approaches that match vehicles based on their lengths have also been developed [54,55,56,57]. However, these approaches require raw loop detector data as opposed to typical 20- or 30-second aggregated data. Alternatively, spatial travel time measurement systems use fixed location equipment to identify and track a subset of vehicles in the traffic stream. By matching the unique vehicle identifications at different reader locations, spatial estimates of travel times can be computed. Typical technologies include AVI and license-plate video detection systems. Finally, probe vehicle technologies track a sample of probe vehicles on a second-by-second basis as they travel within a transportation network. These emerging technologies include cellular geo-location, Global Positioning Systems (GPS), and Automatic Vehicle Location (AVL) systems.

Traffic routing strategies under recurring and non-recurring strategies should be based on forecasting of future traffic conditions rather than historical and/or current conditions. In general the traffic prediction approaches can be categorized into three broad areas: (i) statistical models, (ii) macroscopic models, and (iii) route choice models based on dynamic traffic assignment [24,25,26,28,168]. Time series models have been used in traffic forecasting mainly because of their strong potential for online implementation. Early examples of such approaches include [7] and more recently [120] and [97]. In addition, researchers have applied Artificial Neural Network (ANN) techniques for the prediction of roadway travel times [6,159,160,161,162,165]. These studies demonstrated that prediction errors were affected by a number of variables pertinent to traffic flow prediction such as spatial coverage of surveillance instrumentation, the extent of the loop-back interval, data resolution, and data accuracy.

An earlier publication [70] developed a low-pass adaptive filtering algorithm for predicting average roadway travel times using Automatic Vehicle Identification (AVI) data. The algorithm is unique in three aspects. First, it is designed to handle both stable (constant mean) and unstable (varying mean) traffic conditions. Second, the algorithm can be successfully applied for low levels of market penetration (less than 1 %). Third, the algorithm works for both freeway and signalized arterial roadways. The proposed algorithm utilizes a robust data-filtering procedure that identifies valid data within a dynamically varying validity window. The size of the validity window varies as a function of the number of observations within the current sampling interval, the number of observations in the previous intervals, and the number of consecutive observations outside the validity window. Applications of the algorithm to two AVI datasets from San Antonio, one from

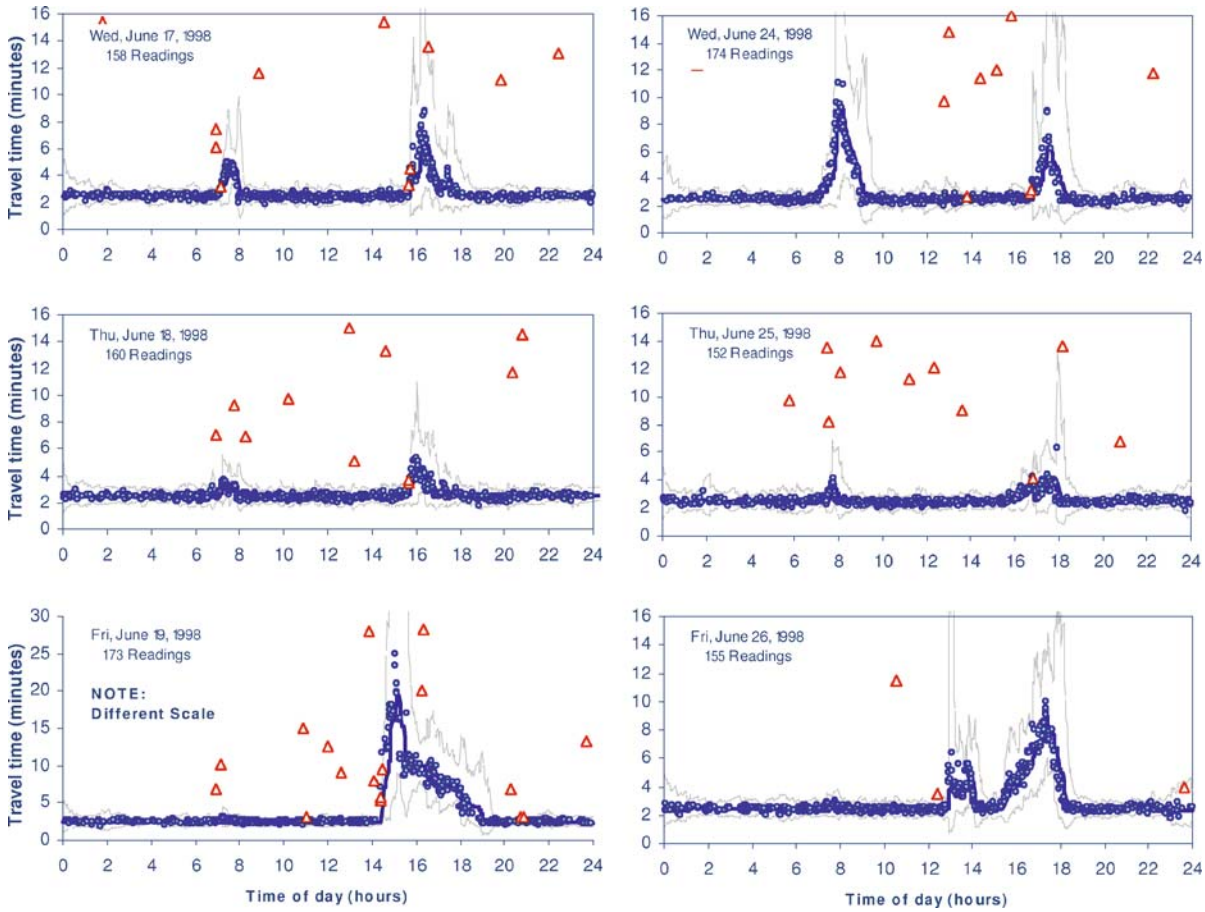

Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment, Figure 6
Sample Application of AVI Travel Time Estimation Algorithm [70]

a freeway link and the other from an arterial link, demonstrated the ability of the proposed algorithm to efficiently track typical variations in average link travel times while suppressing high frequency noise signals.

Within the filtering algorithm, the expected average travel time and travel time variance for a given sampling interval are computed using a moving average (MA) technique. As shown in Eqs. (17) and (18), it estimates the expected average travel time and expected travel time variance within a given sampling interval based on the set of valid travel time observations in the previous sampling interval and the corresponding previously smoothed moving average value using an adaptive exponential smoothing technique. In both equations, calculations of the smoothed average travel time and travel time variance are made using a lognormal distribution to reflect the fact that the distribution is right skewed (skewed towards longer travel times). Field data from the San Antonio AVI system demonstrated that this assumption is reasonable.

$$\tilde{t}_{i,k} = \begin{cases} e^{[\alpha \cdot \ln t_{i,k-1} + (1-\alpha) \cdot \ln \tilde{t}_{i,k-1}]}, & \text{if } n_{i,k} > 0\,, \\ \tilde{t}_{i,k-1}, & \text{if } n_{i,k} = 0\,, \end{cases} \tag{17}$$

$$\tilde{\sigma}^2_{i,k} = \begin{cases} \alpha \cdot \sigma^2_{i,k-1} + (1-\alpha) \cdot \tilde{\sigma}^2_{i,k-1}, & \text{if } n_{i,k} > 1\,, \\ \tilde{\sigma}^2_{i,k-1}, & \text{if } n_{i,k} \le 1\,. \end{cases} \tag{18}$$

It should be noted that $t_{i,k}$ is the observed average travel time along link i within the kth sampling interval (s), $\tilde{t}_{i,k}$ is the smoothed average travel time along link i in the kth sampling interval (s), $\sigma^2_{i,k}$ is the variance of the observed travel times relative to the observed average travel time in the kth sampling interval (s^2), $\tilde{\sigma}^2_{i,k}$ is the variance of the observed travel times relative to the smoothed travel time in the kth sampling interval (s^2), $n_{i,k}$ is the number of valid travel time readings on link i in the kth sampling interval, and $\alpha = 1 - (1-\beta)^{ni,k}$ for all i and k is an exponential

smoothing factor that varies as a function of the number of observations $n_{i,k}$ within the sampling interval, where β is a constant that varies between 0 and 1.

Figure 6 shows an example application of the algorithm using AVI data along I-35 in San Antonio, TX. The figure illustrates the average travel time estimate (thick line), the validity window bounds (thin lines), what are considered to be valid data (circular), and the observations that are considered to be outliers (triangles). The figure clearly illustrates the effectiveness of the algorithm in estimating roadway travel times for low levels of market penetration of AVI tags.

Once link travel times have been estimated, the expected trip or path travel times can be computed by summing the relevant smoothed link travel times. In addition, the trip travel time reliability, which is the probability that a trip can reach its destination within a given period at a given time of day, can be computed for use in traffic routing. Travel time reliability is a measure of the stability of travel time, and therefore is subject to fluctuations in flow [21]. Typically, when flow fluctuations are large, travel time is often longer than expected. As levels of congestion in transportation networks grow, generally the stability of travel time will have greater significance to transportation users. The trip travel time reliability can be computed as the probability $P(T \leq t)$ that the trip travel time (T) is less than some arbitrary travel time (t), using the cumulative distribution function estimated from an analysis of AVI field data. The current state-of-the-art in predicting trip travel time variability is to assume that the travel times on all the links along a path are generated by statistically independent normal distributions. Consequently, the trip variance can be computed as the summation of the link travel time variances for all links along a path. As part of the proposed research effort, different statistical techniques (not assuming independent normal variates) will be devised to estimate the trip travel time variance, as discussed in the Proposed Research Tasks section. These techniques will be tested using data from the video detection system that is currently implemented in the Blacksburg Area.

In addition, research has been conducted to estimate the optimum locations of surveillance equipment for the estimation of travel times. Specifically, an earlier publication developed an algorithm for optimally locating Automatic Vehicle Identification tag readers by maximizing the benefit that would accrue from measuring travel times on a transportation network [219]. The problem is formulated as a quadratic 0–1 optimization problem where the objective function parameters represent benefit factors that capture the relevance of measuring travel times as reflected by the demand and travel time variability along specified trips. An optimization approach based on the Reformulation-Linearization Technique coupled with semidefinite programming concepts was designed to solve the formulated reader location problem. Alternatively, a Genetic Algorithm (GA) approach was developed to optimally locate the AVI readers [13].

Dynamic or Time-Dependent Origin-Destination Estimation

As was demonstrated earlier the Bechmann UE and the SO formulations do not provide unique path flows and thus a synthetic O-D estimator is required to estimate the path flows from the unique link flows. The techniques used to estimate O-D demands can be categorized based on different factors, as will be discussed in detail. The first categorization, of the available O-D estimation techniques, relates to whether the O-D's to be estimated are static, and apply to only one observation time period, or whether estimates are required for a series of linked dynamic time periods. The next breakdown relates to whether the estimation is based on information about the magnitude of trip ends only, or whether information is available on additional links along the route of each trip. The former problem is commonly referred to as the trip distribution problem in demand forecasting, while the latter problem is commonly referred to as the synthetic O-D generation problem. Both problems are discussed, but this section will focus on the latter synthetic O-D generation problem. The former is viewed simply as a simpler subset of the latter.

Within the overall static synthetic O-D generation problem, there are two main flavors. The first exists when the routes that vehicles take through the network are known a priori. The second arises when these routes need to be estimated concurrently while the O-D is being estimated. A priori knowledge of routes can arise automatically when there is only one feasible route between each O-D pair, or when observed traffic volumes are only provided for the zone connectors at the origins and destinations in the network. The first condition is common when O-D's are estimated for a single intersection or arterial, or a single interchange or freeway. The second condition is the default for any trip distribution analysis. This section will initially focus on situations where the routes are known a priori. Subsequently, a solution to the more general problem which involves an iterative use of the solution approach when routes are not known a priori will be discussed.

Within the static/dynamic synthetic O-D generation problem, for scenarios where routes are known a priori

(or are assumed to be known a priori) there exist two sub-problems. The first of these problems relates to situations where flow continuity exists at each node in the network, and multiple O-D matrices can be shown to match these observed flows exactly. In this case, the most likely of these multiple O-D matrices needs to be identified. The second sub-problem relates to situations where flow continuity does not exist at either the node level or at the network level. In other words, the observed traffic flows are such that no matrix exists that will match the observed flows exactly. In this case, Van Aerde et al. [235] introduced a new set of complementary link flows that maintain flow continuity by introducing minimum alterations to the observed flows to solve the maximum likelihood problem.

The static synthetic O-D generation problem, for scenarios where flow continuity does exit, can be formulated in two different ways [242,250]. The first of these considers that the fundamental unit of measure is the individual trip, while the second considers that the fundamental unit of measure is the observation of a single vehicle on a particular link. The availability of a seed or target O-D matrix is implicit in the latter formulation, but can be dropped in the former formulation, as was demonstrated in an earlier publication [235]. However, only when a seed matrix is properly included in the former formulation is it guaranteed to yield consistent results with the latter formulation. In other words, the absence of a seed matrix in the trip based formulation can be shown to yield inconsistent results, at least for some networks in which the multiple solutions result in a different number of total trips.

An additional and related attribute, of the trip-based formulation of maximum likelihood, is the presence of a term in the objective function that is based on the total number of trips in the network. This term, referred to as T, is often dropped in some approximations. However, it can be shown that dropping this term can yield solutions that represent only a very poor approximation to the true solution [186]. In contrast, approximations involve the use of Stirling's approximation, for representing the logarithm of factorials, were shown to yield consistently very good approximations [186]. This finding is critical because use of Stirling's approximation is critical to being able to compute the derivatives that are needed to numerically solve the problem (it is difficult to take derivatives of terms that include factorials).

Other examples from literature include the works of Cremer and Keller [59], Cascetta et al. [43], Wu and Chang [252], Sherali et al. [218], Ashok and Ben-Akiva [18], Hu et al. [95], Chang and Tao [51], Pavlis and Papageorgiou [165], Peeta and Zhou [178,179], Peeta and Yang [174,175], Yang [258], Peeta and Bulusu [167].

Comparison of Synthetic O-D and Trip Distribution Formulations

Within the four-step planning process O-D matrices are estimated in the trip distribution step. Several methods are used for trip distribution including the gravity, growth factor, and intervening opportunities models. The gravity model is most utilized because it uses the attributes of the transportation system and land-use characteristics and has been calibrated and applied extensively to the modeling of numerous urban areas. The model assumes that the number of trips between two zones i and j (T_{ij}) is directly proportional to the number of trip productions from the origin zone (P_i) and the number of attractions to the destination zone (A_j) and inversely proportional to a function of travel time between the two zones (F_{ij}) as

$$T_{ij} = P_i \left(\frac{A_j F_{ij} K_{ij}}{\sum_j A_j F_{ij} K_{ij}} \right) . \tag{19}$$

Typically the values of trip productions and attractions are computed based on trip generation procedures. The values of F_{ij} are computed using a calibration procedure that involves matching modeled and field trip length distributions. The socio-economic adjustment factors (K_{ij}) values are used when the estimated trip interchange must be adjusted to ensure that it agrees with observed trips by attempting to account for factors other than travel time. The values of K are determined in the calibration process, but considered judiciously when a zone is considered to possess unique characteristics.

Because the O-D problem is under-specified, multiple O-D demands can generate identical link flows. For example, if one attempts to estimate an O-D matrix for a 100 zone network with, say 1000 links, one has more unknowns to solve for than there are constraints. In the case of the trip distribution process, there are 100x100 O-D cells to estimate, and only 2x100 trip end constraints. In the case of the synthetic O-D generation process, there are again 100x100 O-D cells to estimate, and only 1000 link constraints. Given the possibility of multiple solutions, both the trip distribution process and the synthetic O-D generation process invoke additional considerations to select a preferred matrix from among the multiple solutions.

In the case of synthetic O-D generation, the desire is to select from among all of the possible solutions, the most likely. This approach requires one to define a measure of the likelihood of each matrix. In general, there are two approaches to establish the likelihood of a matrix. One of them treats the trip as the basic unit of observa-

tion, while the other considers a volume count as the basic unit of observation. The first approach will be discussed in detail, while the interested reader might refer to the literature [235] for a more detailed description of the various formulations. It suffices to indicate that for any matrix with cells T_{ij}, the likelihood of the matrix can be estimated using a function $L = f(T_{ij}, t_{ij})$, where t_{ij} represents prior information. The prior information is often referred to as a seed matrix, and can be derived from a previous study or survey. In the absence of such prior information, all of the cells in this prior matrix should be set to a uniform set of values.

In the case of the trip distribution process, the additional information that is added is some form of impedance. For example, the original gravity model considered that the likelihood of trips between two zones was proportional to the inverse of the square of the distance between the two zones. Since that time, many more sophisticated forms of impedance have been considered, but for the purposes of this discussion, all of these variations can be generalized as being of the form F_{ij}, where $F_{ij} = f(c_{ij})$ or the generalized cost of inter-zonal travel. What is less obvious, however, is the fact that the use of this set of impedance factors F_{ij}, is essentially equivalent to the use of a seed matrix t_{ij}.

Van Aerde et al. [235] demonstrated that solving the trip distribution problem, using zonal trip productions and attractions as constraints, together with a trip impedance matrix, is essentially the same as solving the synthetic O-D problem using zone connector in and out flows as constraints, and utilizing a seed matrix based on Eq. (20).

$$t_{ij} = T\frac{F_{ij}}{\sum_{ij} F_{ij}} \tag{20}$$

Static Formulations

Entropy maximization and information minimization techniques have been used to solve a number of transportation problems [251]. The application of the entropy maximization principles to the static O-D estimation problem was first introduced by Willumsen [242,250]. Willumsen demonstrated that by maximizing the entropy, the most likely trip matrix could be estimated subject to a set of constraints.

The trip-based approach to defining maximum likelihood considers that the overall trip matrix is made up of uniquely identifiable individual trip makers. The most likely matrix is one where the likelihood function is maximized as

$$\text{max.} \qquad Z_1\left(T_{ij}\right) = \frac{T!}{\prod_{ij}\left(T_{ij}!\right)}\,. \tag{21}$$

The above formulation does not take into account any prior information, from for example a previous survey. While the seed matrix does not necessarily have to satisfy the observed link flows, the seed matrix can be utilized to expand the maximum likelihood function to

$$\text{max.} \quad Z_2\left(T_{ij}, t_{ij}\right) = \frac{T!}{\prod_{ij}\left(T_{ij}!\right)} \prod_{ij}\left(\frac{t_{ij}}{\sum_{ij} t_{ij}}\right)^{T_{ij}}\,. \tag{22}$$

It can be noted that the likelihood of an individual trip from i to j is $t_{ij}/\Sigma\Sigma t_{ij}$, based on the above seed matrix. Consequently, the probability of T_{ij} trips being drawn is $(t_{ij}/\Sigma\Sigma t_{ij})^{T_{ij}}$.

The above formulations of objective functions for expressing likelihood require additional constraints in order to be complete [242,250]. The simplest of these constraints indicate that the sum of all trips crossing a given link must be equal to the link flow on that link as

$$V_a = \sum_{ij} T_{ij} p_{ij}^a \qquad \forall\ a\,. \tag{23}$$

As will be shown later, the simplest mechanism, for including the above constraints in the earlier objective functions, is to utilize Lagrange multipliers. These multipliers permit an objective function with equality constraints to be transformed into an equivalent unconstrained objective function.

This simple set of equality constraints, while making the formulation complete, may at times also render the problem infeasible. A more general formulation that was proposed in the literature [235] is to minimize the link flow error, rather than eliminate the error. In other words, rather than finding the most likely O-D that exactly replicates the observed link flows, the problem is re-formulated as finding the most likely O-D matrix from among all of those that come equally close to matching the link flows. One expression that is proposed to capture the error to be minimized is shown in Eq. (24), and is subject to the flow continuity constraints in Eq. (25). The constraints in Eq. (25) can be introduced in Eq. (25) to yield an unconstrained objective function, yielding a set of complementary link flows V_a'. These complementary flows are those which deviate the least from the observed link flows, while satisfying link flow continuity. Given that these complementary link flows do satisfy flow continuity, they can now

be added in as rigid equality constraints to the objective function (21) or (22), and be guaranteed to yield a feasible solution.

$$\text{min.} \qquad Z_3\left(T_{ij}\right) = \sum_a \left(V_a - V_a'\right)^2 \tag{24}$$

$$V_a' = \sum_{ij} T_{ij} p_{ij}^a \qquad \forall\ a\ . \tag{25}$$

Alternatively, one can incorporate Eq. (25) into Eq. (24) to yield

$$\text{min.} \qquad Z_4\left(T_{ij}\right) = \sum_a \left(V_a - \sum_{ij} T_{ij} p_{ij}^a\right)^2 . \tag{26}$$

This equation should be minimized concurrently to maximizing the objective function (21) or (22). Unfortunately, it is not easy to combine one expression that desires to maximize likelihood with another that desires to minimize link flow error, as a Lagrangian can only add equality constraints to a constrained objective function. Van Aerde et al. proposed a solution to this problem which involves taking the partial derivatives of Eq. (26) with respect to each of the trip cells that are to be estimated as

$$\frac{\partial}{\partial T_{ij}} Z_4\left(T_{ij}\right) = \frac{\partial}{\partial T_{ij}} \sum_a \left(V_a - \sum_{ij} T_{ij} p_{ij}^a\right)^2 = 0 \qquad \forall\ i, j\ . \tag{27}$$

$$0 = 2\left(\sum_a \left(V_a \cdot p_{ij}^a\right) - \left(\sum_a p_{ij}^a \left(\sum_{xy} T_{xy} p_{xy}^a\right)\right)\right) \qquad \forall\ i, j \tag{28}$$

This yields as many equations as there are trip cells, as shown in Eq. (27). Furthermore, setting these derivatives equal to 0 is equivalent to minimizing Eq. (27). However, while Eq. (24) could not be added to the maximum likelihood objective function, the equalities in Eq. (28) can. This produces an unconstrained objective function that always yields a feasible solution computed as

$$\text{max.} \quad \frac{T!}{\prod_{ij} T_{ij}} \prod_{ij} \left(\frac{t_{ij}}{t}\right)^{T_{ij}} - \sum_{ij} \left(\lambda_{ij} \cdot 2\left(\sum_a \left(V_a \cdot p_{ij}^a\right) - \left(\sum_a p_{ij}^a \left(\sum_{xy} T_{xy} p_{xy}^a\right)\right)\right)\right), \tag{29}$$

where: $T = \sum_{ij} T_{ij}$ and $t = \sum_{ij} t_{ij}$.

The net result, of the above process, is to suggest that most synthetic O-D generation problems consist of two sub-problems. One of these involves finding a new set of complementary link flows that do produce link flow continuity, at which point the maximum likelihood problem can be solved as before. Alternatively, one can compute the partial derivatives, that will yield link flow continuity, while deviating by the least amount from the observed link flows, and then utilize them directly in the maximum likelihood formulation using Lagrange multipliers. Both solutions can be shown to yield identical results.

A first challenge with maximizing Eq. (29) is that it yields very large numbers that are difficult to work with. Further more, as it is common to maximize objective functions by taking their derivatives, and as it is more difficult to contemplate the derivative of a discontinuous expression, such as those including factorials, a simple approximation is made. This approximation involves taking the natural logarithm of either objective function Eq. (21) or (22). Taking the natural logarithm of the objective function both makes the output easier to handle and permits the use of Stirling's approximation as a convenient continuous equivalent to the term $\ln(x!)$ as

$$\ln(T!) = T \ln T - T \tag{30}$$

The resulting converted objective function using the Stirling approximation on the original objective function of Eq. (22) is computed as

$$\begin{aligned}\text{max.}\ T \ln\left(\frac{T}{t}\right) - T - \sum_{ij}\left(T_{ij} \ln\left(\frac{T_{ij}}{t_{ij}}\right) - T_{ij}\right) \\ - \sum_{ij}\left(\lambda_{ij} \cdot 2\left(\sum_a \left(V_a \cdot p_{ij}^a\right)\right.\right. \\ \left.\left. - \left(\sum_a p_{ij}^a \left(\sum_{xy} T_{xy} p_{xy}^a\right)\right)\right)\right) .\end{aligned} \tag{31}$$

Expanding and simplifying the various terms we derive

$$\begin{aligned}T \ln\left(\frac{T}{t}\right) - T - \sum_{ij}\left(T_{ij} \ln\left(\frac{T_{ij}}{t_{ij}}\right) - T_{ij}\right) \\ = T \ln\left(\frac{T}{t}\right) - \sum_{ij} T_{ij} \ln\left(\frac{T_{ij}}{t_{ij}}\right) .\end{aligned} \tag{32}$$

When Eq. (32) is augmented with the previously mentioned partial derivatives that minimize the link flow error

we derive

$$\begin{aligned}\max. \quad & T\ln\left(\frac{T}{t}\right) - \sum_{ij} T_{ij}\ln\left(\frac{T_{ij}}{t_{ij}}\right) \\ & - \sum_{ij}\left(\lambda_{ij}2\left(\sum_{a}\left(V_a \cdot p^a_{ij}\right)\right.\right. \\ & \left.\left. - \left(\sum_{a} p^a_{ij}\left(\sum_{xy} T_{xy}p^a_{xy}\right)\right)\right)\right). \quad (33)\end{aligned}$$

This equation, when solved, yields the most likely O-D matrix of all of those matrices that come equally close to matching the observed link flows.

It should be noted that the objective function of Eq. (33) is composed of two components. The first being the error between the field observed flows and the flows that satisfy flow continuity with minimum difference from observed flows. The second component represents the likelihood of an O-D matrix table. The objective is to find the O-D matrix with the maximum likelihood. In the case that the seed matrix is the optimum matrix the likelihood component resolves to zero.

Dynamic Formulations

The above formulations assume that the vehicles are assigned to all links simultaneously (i. e. a vehicle is present on all links along its path simultaneously). In order to address the dynamic nature of traffic, the analysis period can be divided into equally spaced time slices. Origin-destination demands are then indexed by the time slice they depart and the time slice they are observed on a link, as

$$\begin{aligned}\max. \quad & T_r\ln\left(\frac{T_r}{t_r}\right) - \sum_{rij} T_{rij}\ln\left(\frac{T_{rij}}{t_{rij}}\right) \\ & - \sum_{rij}\left(\lambda_{rij}2\left(\sum_{sa}\left(V_{sa} \cdot p^{sa}_{rij}\right)\right.\right. \\ & \left.\left. - \left(\sum_{sa} p^{sa}_{rij}\left(\sum_{rxy} T_{rxy}p^{sa}_{rxy}\right)\right)\right)\right). \quad (34)\end{aligned}$$

Where T_r is the total demand departing during time-slice r, t_r is the total seed matrix demand departing during time-slice r, T_{rij} is the traffic demand departing during time-slice r traveling between origin i and destination j, t_{rij} is the seed traffic demand departing during time-slice r traveling between origin i and destination j, λ_{rij} is the Lagrange multiplier for departure time-slice, origin, and destination combination rij, V_{sa} is the observed volume on link a during time slice s, and p^{sa}_{rij} is the probability of a demand between origin i and destination j during time-slice r is observed on link a during time-slice s. The solution of Eq. (34) is computationally extensive and has been demonstrated to not produce significantly better results than generating time-dependent O-D demands, as will be discussed.

Alternatively, the more common approach is to generate time-dependent O-D demands by solving Eq. (33) for each time-slice independently assuming that O-D demands can travel from the origin to destination zone within a time-slice (i. e. the trip travel time is less than the time-slice duration) without considering the interaction between time slices. This approach is computationally simpler and easier to implement and thus will be discussed in more detail. The formulation can written as

$$\begin{aligned}\max. \quad & T_r\ln\left(\frac{T_r}{t_r}\right) - \sum_{ij} T_{rij}\ln\left(\frac{T_{rij}}{t_{rij}}\right) \\ & - \sum_{ij}\left(\lambda_{rij}\cdot 2\left(\sum_{a}\left(V_a \cdot p^a_{rij}\right)\right.\right. \\ & \left.\left. - \left(\sum_{a} p^a_{rij}\left(\sum_{xy} T_{rxy}p^a_{rxy}\right)\right)\right)\right) \quad \forall\, r. \quad (35)\end{aligned}$$

Here the T_{rij} are solved for independent of other time-slices. It should be noted, that the approach ignores the interaction of demands across various time-slices which is a valid assumption if the network is not over-saturated. However, if the network is oversaturated the assumption of time slice independence may not be valid. The duration of the time-slice should be selected such that steady-state conditions are achieved within a time-slice.

Solution Algorithms

The solution of the set of equations presented in (35) is hard given that the objective function is nonconvex and that in many cases the p^a_{rij} are not available and thus the problem becomes to solve for T_{rij} and p^a_{rij} that maximize the objective function.

Here we present a numerical heuristic that solves the above formulation for large networks when the number of equations and unknowns becomes extremely computationally intensive. This special purpose equation solver has been developed and implemented in the QueensOD software. This solver fully optimizes the objective function of Eq. (35). The software has been shown to produce errors less than 1% for the range of values and derivatives being typically considered in the problem. A sample application of the QueensOD software is presented later in the paper, however, initially the heuristic approach is described.

The numerical solution begins by building a minimum path tree and performing an all-or-nothing traffic assignment of the seed matrix, as illustrated in Fig. 7. A relative or absolute link flow error is computed depending on user input. Using the link flow errors O-D adjustment factors are computed and utilized to modify the seed O-D matrix. The adjustment of the O-D matrix continues until one of two criteria are met, namely the change in O-D error reaches a user-specified minimum or the number of iterations criterion is met. If additional trees are to be considered, the model builds a new set of minimum path trees (loop 2) and shifts traffic gradually to the second minimum path tree. The minimum objective function for two trees is computed in a similar fashion as described for the single tree scenario. The process of building trees and finding the optimum solution continues until all possible trees have been explored. The proposed numerical solution ensures that in the case that the seed matrix is optimum no changes are made to the matrix. In addition, the use of the seed matrix as a starting point for the search algorithm ensures that the optimum solution resembles the seed matrix as closely as possible while minimizing the link flow error. In other words, the seed matrix biases the solution towards the seed matrix.

In order to demonstrate the applicability of the QueensOD software, a sample application to a 3500-link network of the Bellevue area in Seattle is presented. Other applications of the QueensOD software are described in detail in the literature [71,196]. The O-D demand for the Bellevue network was calibrated to AM peak Single Occupancy Vehicle (SOV) and High Occupancy Vehicle (HOV) flows. The seed matrix was created using the standard four-step transportation planning process by applying the EMME/2 model. The Seattle network was converted from EMME/2 format to INTEGRATION format.

The calibration of the O-D demand to tube and turning movement counts was conducted using the QueensOD software using the planning trip distribution O-D matrix as the seed solution. The calibration resulted in a high level of consistency between estimated and field observed link flow counts (coefficient of determination of 0.98 between the estimated and observed flows), as illustrated in Fig. 8. Figure 8 demonstrates that in calibrating the O-D matrix to observed traffic counts, the trip distribution O-D matrix (seed matrix) was modified significantly (coefficient of determination of 0.56 between trip distribution and synthetic O-D matrix). Consequently, it is evident that a modification of the trip distribution matrix was required in order to better match observed link and turning movement counts. It should be noted however, that the total number of trips was increased by only 4 % as a result of the synthetic O-D calibration effort. Consequently, the illustrated calibration effort resulted in a significant modification of the trip table with minor modification to the total number of trips.

In addition to the above mentioned research, a significant number of problem formulations and applications have been documented in the literature. To name a few (in chronological order) Cascetta et al. [43] tested a method based on two generalized least squares estimators on the

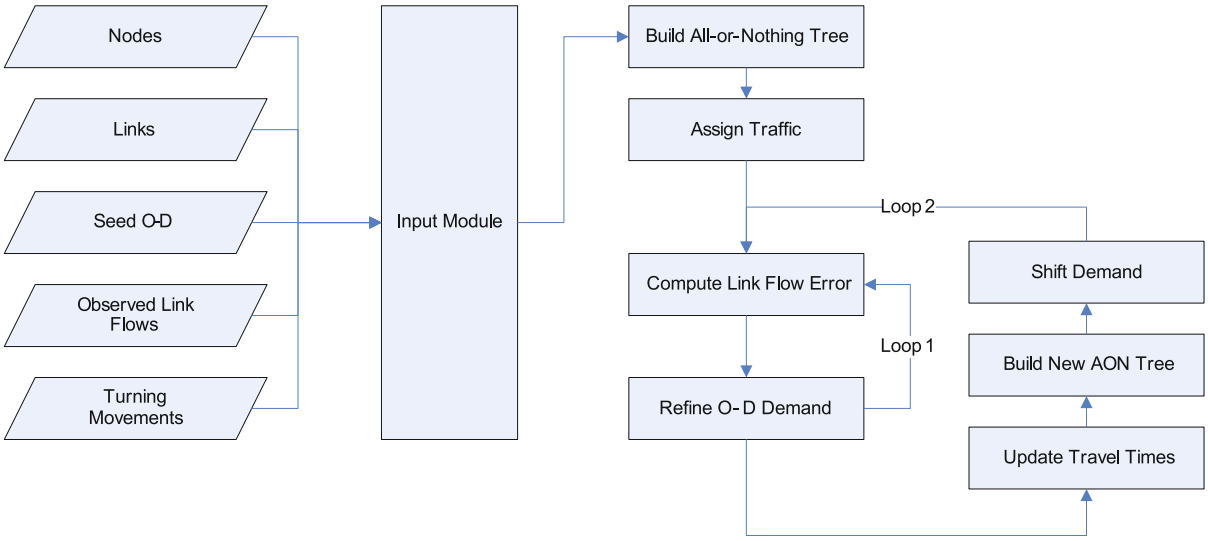

Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment, Figure 7
QueensOD Heuristic O-D Estimation Approach (Synthetic O-D Estimator)

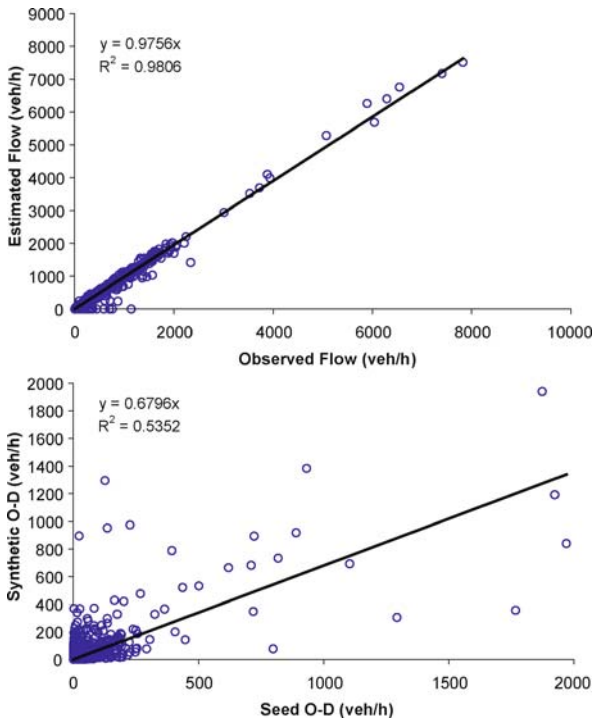

Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment, Figure 8
Example Application of QUEENSOD to the Bellevue Network in Seattle

Brescia-Verona-Vicenza-Padua motorway in Italy. They found that the accuracy of their model depended heavily on the number of links with observed traffic counts. Van Aerde et al. [238] introduced the QUEENSOD method and demonstrated its applicability on a 35-km section of Highway 401 in Toronto, Canada. Ashok [16] evaluated the use of a Kalman filtering-based method, which was first presented by Okutani [158] and estimates unobserved link traffic counts from observed link traffic counts. The method used was formulated by Ashok and Ben Akiva [17] and Ashok [16] and was evaluated using actual data from the Massachusetts Turnpike, Massachusetts, a stretch of I-880 near Hayward, California and a freeway encircling the city of Amsterdam, Netherlands. Later, Hellinga and Van Aerde [93] compared a least square error model and a least relative error model on a 35-km section of Highway 401 in Toronto, Canada. Zhou and Sachse [260] compared the use of three different O-D estimators and on a motorway network in Europe. They concluded that the models, although characterized by different computational loads, produced satisfactory results. They also commented on the need to decide on locations of detectors and aggregation time intervals. Van Der Zijpp and Romph [239] experimented their model on the Amsterdam Beltway. They tested their model using two different days worth of data and compared their model results with real and historical average data. While their model performed better in cases of accidents, the historical average data did, at least as good, in normal traffic. They stressed on the importance of correct modeling of the network and traffic flow characteristics for the production of good results. Kim et al. [112] introduced a genetic algorithm based method to overcome the shortcoming of the bi-level programming method when there is a significant difference between target and true O-D matrices. They tested their model on a small network of 9 nodes. Bierlaire and Crittin [27] formulated a least-square based method to overcome some of the shortcomings of the Kalman filter approach. They tested their method on a simple network as well as two real networks: a medium scale network, Central Artery Network, Boston, MA, and a large scale network, Irvine Network, Irvine, CA. Yun and Park developed a genetic algorithm based method with the purpose of solving dynamic O-D matrices for large networks. They compared their model's results with the results of QUEENSOD, and they tested their method on the City of Hampton network using the PARAMICS microscopic traffic simulation software. Nie et al. [156] developed a formulation that incorporates a decoupled path flow estimator in a generalized least squares framework with the objective of developing an efficient, simplified solution algorithm for realistic size networks. They tested their method on a small (9-node) and mid-size (100 nodes) network. Zhou and Mahmassani [259] developed a multi-objective optimization framework for the estimation of the O-D matrices using automatic vehicle identification data. They tested their method on a simplified Irvine testbed network (31 nodes). Finally, Castillo et al. [48] developed a method for the reconstruction and estimation of the trip matrix and path flows based on plate scanning and link observations. They tested their method on the Nguyen–Dupius Network, and concluded the superiority of plate scanning on link counts.

It should be noted at this point that the O-D estimation formulations and techniques that were presented and described in this section are heuristics and thus there is no mathematical proof that the algorithms converge to the unique optimum solution either in the static or dynamic context. While we have demonstrated that the solution matches the observed link flows for complex networks (Fig. 8, unfortunately the actual O-D demand is typically not available for real-life applications and thus it is not possible to measure how good the solution compares to the unique optimum O-D matrix.

Dynamic Estimation of Measures of Effectiveness

Dynamic assessment of traffic network performance requires the estimation of various measures of effectiveness in a dynamic context. This section provides a brief overview of the procedures for estimating delay, vehicle stops, and vehicle energy consumption and emissions.

Estimation of Delay

A key parameter in the dynamic assessment of traffic networks is the estimation of vehicle delay. The computation of delay requires the computation of travel times. Significant research has been conducted to develop analytical models for estimating delay especially at signalized intersections. Examples of such research efforts are provided for the interested reader [10,32,33,44,46,49,58,60,61,63,65,66,74,76,77,79,86,114,116,124,154,212,213,230].

Roadway travel times can be computed for any given vehicle by providing that vehicle with a *time card* upon its entry to any roadway or link. Subsequently, this *time card* is retrieved when the vehicle leaves the roadway. The difference between these entry and exit times provides a direct measure of the roadway travel time experienced by each vehicle. The delay can then be computed as the difference between the actual and free-flow travel time.

Alternatively, vehicle delay can be computed microscopically every deci-second as the difference in travel time between travel at the vehicle's instantaneous speed and travel at free-flow speed, as

$$d(t_i) = \Delta t \left(1 - \frac{u(t_i)}{u_f}\right) \quad \forall\ i\,. \tag{36}$$

The summation of these instantaneous delay estimates over the entire trip provides an estimate of the total delay. This model has been validated against analytical time-dependent queuing models, shockwave analysis, the Canadian Capacity Guide, the Highway Capacity Manual (HCM), and the Australian Capacity Guide procedures [71]. The procedure has also been incorporated in the INTEGRATION traffic simulation software [233,234] and utilized with second-by-second Global Positioning System (GPS) data [72,190].

Estimation of Vehicle Stops

Numerous researchers have dealt with the problem of estimating vehicle stops especially at signalized intersections. An important early contribution is attributed to Webster [247], who generated stop and delay relationships by simulating uniform traffic flows on a single-lane approach to an isolated intersection. In particular, the equations that Webster [247] derived have been fundamental to traffic signal setting procedures since their development. Later, Webster and Cobbe [248] developed a formula for estimating vehicle stops at under-saturated intersections assuming random vehicle arrivals. Other models were developed by Newell [153] and Catling [49]. Catling adapted equations of classical queuing theory to over-saturated traffic conditions and developed a comprehensive queue length estimation procedure that captured the time-dependent nature of queues to be applied to both under-saturated and over-saturated conditions. In addition, Cronje [60,61,62,63] developed stop and delay equations by treating traffic flow through a fixed-time signal as a Markov process. The approach assumed that the number of queued vehicles at the beginning of a cycle could be expressed by a geometric distribution. These models, however, were not designed to account for the partial stops that vehicles may incur. Furthermore, the models that account for partial stops do not estimate vehicle partial and full stops for over-saturated conditions. A study by Rakha et al. [187] developed a procedure for estimating vehicle stops while accounting for partial stops, as

$$S(t_i) = \frac{u(t_i) - u(t_{i-1})}{u_f} \qquad \forall\ i \ni u(t_i) < u(t_{i-1})\,. \tag{37}$$

The sum of these partial stops is also recorded. This sum, in turn, provides a very accurate explicit estimate of the total number of stops that are encountered along a roadway. Again the model can be implemented within a microscopic traffic simulation software or applied to second-by-second speed measurements using a GPS system.

Estimation of Vehicle Energy Consumption and Emissions

Estimating accurate mobile source emissions has gained interest among transportation professionals as a result of increasing environmental problems in large metropolitan urban areas. While current emission inventory models in the US, such as MOBILE and EMPAC, are capable of estimating large scale inventories, they are unable to estimate accurate vehicle emissions that result from operational-level projects. Alternatively, microscopic emission models are capable of assessing the impact of transportation projects on the environment and performing project-level analyzes. Consequently, the focus of this discussion will be on these microscopic and also mesoscopic models. Two models that are emerging include the Comprehensive Modal Emissions Model (CMEM) and the Virginia Tech Microscopic (VT-Micro) model. These models

are briefly described in terms of their structure, logic, and validity.

Comprehensive Modal Emission Model The Comprehensive Modal Emissions Model (CMEM), which is one of the newest power demand-based emission models, was developed by researchers at the University of California, Riverside [20]. The CMEM model estimates LDV and LDT emissions as a function of the vehicle's operating mode. The term "comprehensive" is utilized to reflect the ability of the model to predict emissions for a wide variety of LDVs and LDTs in various operating states (e. g., properly functioning, deteriorated, malfunctioning).

The development of the CMEM model involved extensive data collection for both engine-out and tailpipe emissions of over 300 vehicles, including more than 30 high emitters. These data were measured at a second-by-second level of resolution on three driving cycles, namely: the Federal Test Procedure (FTP), US06, and the Modal Emission Cycle (MEC). The MEC cycle was developed by the UC Riverside researchers in order to determine the load at which a specific vehicle enters into fuel enrichment mode. CMEM predicts second-by-second tailpipe emissions and fuel consumption rates for a wide range of vehicle/technology categories. The model is based on a simple parametrized physical approach that decomposes the entire emission process into components corresponding to the physical phenomena associated with vehicle operation and emission production. The model consists of six modules that predict engine power, engine speed, air-to-fuel ratio, fuel use, engine-out emissions, and catalyst pass fraction. Vehicle and operation variables (such as speed, acceleration, and road grade) and model calibrated parameters (such as cold start coefficients, engine friction factor) are utilized as input data to the model.

Vehicles were categorized in the CMEM model based on a vehicle's total emission contribution. Twenty-eight vehicle categories were constructed based on a number of vehicle variables. These vehicle variables included the vehicle's fuel and emission control technology (e. g. catalyst and fuel injection), accumulated mileage, power-to-weight ratio, emission certification level (tier0 and tier1), and emitter level category (high and normal emitter). In total 24 normal vehicle and 4 high emitter categories were considered [20].

The Virginia Tech Microscopic Energy and Emission Model (VT-Micro Model) The VT-Micro emission models were developed from experimentation with numerous polynomial combinations of speed and acceleration levels. Specifically, linear, quadratic, cubic, and fourth degree combinations of speed and acceleration levels were tested using chassis dynamometer data collected at the Oak Ridge National Laboratory (ORNL). The final regression model included a combination of linear, quadratic, and cubic speed and acceleration terms because it provided the least number of terms with a relatively good fit to the original data (R^2 in excess of 0.92 for all measures of effectiveness [MOE]). The ORNL data consisted of nine normal-emitting vehicles including six light-duty automobiles and three light-duty trucks. These vehicles were selected in order to produce an average vehicle that was consistent with average vehicle sales in terms of engine displacement, vehicle curb weight, and vehicle type. The data collected at ORNL contained between 1,300 to 1,600 individual measurements for each vehicle and MOE combination depending on the vehicle's envelope of operation [8].

This method has a significant advantage over emission data collected from a few driving cycles because it is difficult to cover the entire vehicle operational regime with only a few driving cycles. Typically, vehicle acceleration values ranged from –1.5 to 3.7 m/s^2 at increments of 0.3 m/s^2 (–5 to 12 ft/s^2 at 1-ft/s^2 increments). Vehicle speeds varied from 0 to 33.5 m/s (0 to 121 km/h or 0 to 110 ft/s) at in increments of 0.3 m/s [8].

The model had the problem of overestimating HC and CO emissions especially for high acceleration levels. Since this problem arose from the fact that the sensitivity of the dependent variables to the positive acceleration levels is significantly different from that for the negative acceleration levels, a two-regime model for positive and negative acceleration regimes was developed as [8,183]

$$\ln(MOE_e) = \begin{cases} \sum_{i=0}^{3}\sum_{j=0}^{3}(L_{i,j}^{e} \times u^i \times a^j) & \text{for } a \geq 0 \\ \sum_{i=0}^{3}\sum_{j=0}^{3}(M_{i,j}^{e} \times u^i \times a^j) & \text{for } a < 0\,. \end{cases} \tag{38}$$

Where MOE_e is the instantaneous fuel consumption or emission rate (ml/s or mg/s); $K_{i,j}^{e}$ is the model regression coefficient for MOE "e" at speed power "i" and acceleration power "j"; $L_{i,j}^{e}$ is the model regression coefficient for MOE "e" at speed power "i" and acceleration power "j" for positive accelerations; $M_{i,j}^{e}$ is the model regression coefficient for MOE "e" at speed power "i" and acceleration power "j" for negative accelerations; u is the instantaneous speed (km/h); and a is the instantaneous acceleration rate (km/h/s).

Additionally, the VT-Micro model was expanded by including data from 60 light-duty vehicles (LDVs) and

trucks (LDTs). Statistical clustering techniques were applied to group vehicles into homogeneous categories using classification and regression tree (CART) algorithms. The 60 vehicles were classified into five LDV and two LDT categories [182]. In addition, HE vehicle emission models were constructed using second-by-second emission data. In constructing the models, HEVs are classified into four categories for modeling purposes. The employed HEV categorization was based on the comprehensive modal emission model (CMEM) categorization. The first type of HEVs has a chronically lean fuel-to-air ratio at moderate power or transient operation, which results in high emissions in NO. The second type has a chronically rich fuel-to-air ratio at moderate power, which results in high emissions in CO. The third type is high in HC and CO. The fourth type has a chronically or transiently poor catalyst performance, which results in high emissions in HC, CO, and NO. Each model for each category was constructed within the VT-Micro modeling framework. The HE vehicle model was found to estimate vehicle emissions with a margin of error of 10% when compared to in-laboratory bag measurements [9]. Furthermore, all the models were incorporated into the INTEGRATION software, and made it possible to evaluate the environmental impacts of operational level transportation projects [164].

Use of Technology to Enhance System Performance

Due to the recent extensive developments within the fields of artificial intelligence, communications, and computation algorithms, transportation and traffic engineers' goals have evolved. As mentioned earlier in the paper, current spatio-temporal distribution of trips is far from being optimum, either with respect to driver satisfaction and/or network performance. A part of the contemporary DTA research is directed towards influencing, as opposed to modeling, dynamic spatio-temporal trip distributions. Advanced Traveler Information Systems (ATISs) are definitely the main tool for such influence, and understanding driver behavior is critical to the design and implementation of such systems. Research which is directly related to the possibility of enhancing system performance through the use of technology may be categorized in the following main areas of research:

- Validation of models, lab experiments and real world behavior, which is the area concerned with verifying the different theories and their implicit assumptions with regards to real-life situations. Due to the extreme complexity and questionable possibility of this task, several attempts have been made to verify the models with respect to lab experiments rather than the real world behavior. Moreover, comparison and verification of spatial and temporal transferability of the models might as well fall within this area. Examples of current literature include the works of Chang and Mahmassani [50] and Mahmassani and Jou [129].
- Calibration of algorithms and models, which as the name suggests, is the area related to the calibration of the algorithms and model parameters. This also entails spatial and temporal calibration, for certain models and/or parameters might only be valid for certain locations and time periods rather than others. Examples of current literature include the works of Chang and Mahmassani [50] and Rakha and Arafeh [184].
- Real time deployment, which focuses on the possibility of deploying DTA models into the real world. This area of research is concerned with developing deployable DTA algorithms. Current literature states that although "a mathematically tractable analytical model that is adequately sensitive to traffic realism vis-à-vis real-time operation is still elusive", yet even with currently available models there is a tradeoff between solution accuracy and computational efficiency. Other real-time deployment issues include computational tractability; consistency checking; model robustness, stability, and error and fault tolerance; and demand estimation and prediction [180]. Examples of current literature include the works of Mahmassani et al. [131,132,133,134], Ben-Akiva et al. [25,26], Mahmassani and Peeta [130,134,135], Peeta and Mahmassani [168], Hawas [88], Hawas et al. [91], Hawas and Mahmassani [90,92], Cantarell and Cascetta [38], Anastassopoulos [12], and Jha et al. [104].
- Issues of uncertainty, which is, as mentioned earlier, a fundamental feature in most transportation phenomena. Uncertainty can be represented in trip makers' knowledge of different route travel times, in the compliance rates of drivers to information, in the accuracy of the disseminated control information, in the driver's perception of disseminated information reliability, in the controller's predicted and/or refined dynamic travel times and/or O-D matrices, among others. Uncertainty-related research issues have been addressed through several approaches, like stochastic modeling, fuzzy control, and reliability indices. Examples of current literature include the works of: Birge and Ho [28], Peeta and Zhou [109,167], Cantarell and Cascetta [38], Ziliaskopoulos and Waller [262], Waller and Ziliaskopoulos [245], Waller [244], Peeata and Jeong [177], Jha et al. [104], Peeta and Paz [171].
- DTA control, which is the area of research concerned

with modifying how trips are distributed on the network. Research within this area focuses on capturing current network performance, and works on modifying the system elements, such as drivers route, and/or departure time selection, as well as mode choice (possibly through pricing and information dissemination); and traffic management (primarily through signal operation), in order to optimize system performance. Examples of current literature include the works of Peeta and Paz [171].

- Realism of other system characteristics, which is the research area concerned with capturing other system realities that are not considered in current available literature. Examples of such realities may include [180],
  - Person rather than driver assignment. It is an undeniable fact that many people tend to make their mode choices based on daily, real-time decisions, i. e. this is a dynamic and not a static process. It is further anticipated that with the current (and predicted) maturity of information technology within the transportation arena, would require explicit modeling within DTA models.
  - The effect of interaction between the different vehicle classes and road infrastructure. It is beyond doubt that certain vehicle classes (such as trucks and busses for example) will not be able to comply with certain diversion-requesting disseminated information, due to road infrastructure constraints. However, in other occasions, these vehicle classes might be able to divert routes, yet with travel time penalties (example if the turning radius was inadequate) that might not only affect these vehicle classes, but all other diverting vehicles as well.
  - Capturing latest traffic control technology and strategies. Traffic control technology and strategies have been rapidly developing during the past couple of decades. Examples of this include transit signal preemption, real-time adaptive signal traffic control, electronic toll collection, etc. For efficient DTA control, DTA algorithms should be able to sufficiently capture and consider them.

Examples of current literature include the works of Ran and Boyce [200], Peeta et al. [174], Ziliaskopoulos and Waller [262], Dion and Rakha [72], Sivananden et al. [223], Rakha et al. [190,191], Rakha and Zhang [197].

Road Pricing

A number of approaches have been and continue to be considered in an attempt to reduce urban congestion. These approaches include supply and demand management strategies. Supply management strategies are addressed through the building of new infrastructure, expanding the existing infrastructure, or by increasing the utilization of existing infrastructure through the introduction of traffic management systems. Alternatively, demand strategies focus on either distributing the existing demand more evenly in space over the network or distributing the existing demand more evenly by spreading the peak period over a longer time. One of the emerging demand management strategies for enhancing traffic system performance is the use of electronic road pricing.

Road pricing entails the charging of the user of a road for the use of the facility. The road charges include fuel taxes, license fees, parking taxes, tolls, and congestion changes, including those which may vary by time-of-day, by the specific vehicle type being used. Road pricing has two distinct objectives: revenue generation, usually for road infrastructure financing, and congestion pricing for demand management purposes. Toll roads are the typical example of revenue generation. Changes for using high-occupancy toll (HOT) lanes or urban tolls for entering a restricted area of a city are typical examples of using road pricing for congestion management purposes.

In modeling route choice within the context of road pricing one has to consider at least two main types of route choices, one which is tolled and one which is not. An example of such a situation is a multi-lane freeway that splits at several locations to provide 3 lanes that are free of toll and 1 lane that involves a toll. While these are some unique traffic engineering issues associated with the weaving that can take place, when there are multiple breaks in the structure that divides the toll lanes from the free ones, many important equilibrium issues can already be analyzed by looking at single diverge sections. If traffic demand is low, and the distance and speed limits are similar on both types of lanes, there exists little incentive for drivers to utilize the toll lane alternative. However, as the traffic demand on the toll alternative increases, its speed will decrease. This will make its travel time become longer, making the free lanes less attractive. This sets up a condition in which it may become attractive to some, but not necessarily all drivers, to start utilizing the toll lane. The timing of when drivers start to use the toll lanes and the number of drivers who elect to use the toll lane depends primarily on two factors, namely the "value of the toll" and the "value of time" of the drivers involved.

The "value of the toll" and "value of time" are closely related quantities, as a higher toll value accompanied with a higher value of time will result in the same perceived disutility and therefore path impedance. Consequently, it is ratio of the "value of the toll" to the "value of time" that is

most important. In addition, one should also note that for the same toll value, those with the highest "value of time" will be least influenced by the toll. They will therefore also be the first ones to start using the toll lanes when the free lanes become busy and slower. In contrast, those with the lowest value of time perceive the toll as having disutility or travel time equivalency. They will therefore be the last ones to switch to the toll lane, if at all.

The decisions, of those drivers who have a lower value of time, are also influenced by the decisions of the drivers with a higher value of time. Specifically, any switch over of the latter drivers from the free lanes to the toll lanes will leave the free lanes less congested for those with a lower value of time who stay in the free lanes. This may leave it undesirable for those with a low value of time to also switch to the toll lane. However, if the percent of drivers with a high value of time is relatively small, or if the total flow A to B is very high, even some of those with a low value of time will perceive themselves as being better off if they use the toll lanes. The issue of toll lane usage therefore requires one to answer two interrelated questions, namely: (a) what group or sub-population of drivers will utilize the toll road, and (b) what fraction of this group or sub-population will utilize the toll road. Consequently, extensive research is currently underway to model driver route selection under various tolling conditions. Examples of some research efforts on this topic include [23, 139,140,141,146]. This list is by no means comprehensive but does provide some applications on this emerging topic.

Related Transportation Areas

Research within the following two transportation areas definitely precedes DTA research. However, their significance to the DTA field is based on the fact that DTA theories are mostly dependent on older theories stemming from these two areas. Hence, advances within these two areas could probably significantly affect the advances within the DTA arena.

- Traffic flow models encompass the mathematical representation, or perhaps simulation of the traffic flow characteristics, such as modeling traffic flow propagation, queue spillbacks, lane-changing, signal operation, travel time computation, etc. are crucial in determining driver expectations and behavior. In addition, these are also fundamental in the calculations of travel times, which are vital in the combined problem of departure time and route choice. The quantity of research available in this area is probably as big as the quantity of research done in the area of DTA all together, if not even more. However, as mentioned earlier, all of the research done within this area has direct influence on the realization of the traffic flow models, which are also used within the DTA models.
- Planning applications, which in spite of being a quite under-researched area at the moment, is a vitally important one. There is no doubt DTA models are superior to static models, hence, it is probably only a matter of time before the industry abandons static models for dynamic models. "Dynamic models are simply the natural evolution in the transportation field that like any other new effort suffers from early development shortcomings" [180]. Examples of current literature within this area of research include the works of Li [125], Friesz et al., Waller [244], Waller and Ziliaskopoulos [245], Ziliaskopoulos and Waller [263], Ziliaskopoulos and Wardell [261].

Future Directions

Future research challenges and directions include:

- Enhance traffic flow modeling and driver behavior modeling. These include the modeling of person as opposed vehicle route choices, the separation of driver and vehicle within the traffic modeling framework, the explicit modeling of vehicle dynamics, enhancing car-following, lane-changing, and routing behavior.
- Develop more efficient algorithms that would be suitable for real-time deployment, without making any compromises in the computational accuracy, i. e. without trading-off the solution accuracy for the computational efficiency. In precise, without compromising any dimension of the traffic flow theory, nor driver behavior assumptions. As a matter of fact, further research should be done to capture more of the traffic, as well as the driver behavior theory. Hence, this should help in improving the realism of the available DTA models.
- Conduct more research on the driver behavior theory. Especially, since human factors cognitive research has significantly improved in the previous couple of decades, then modeling driver behavior from this perspective might lead to valuable outcomes.
- Critical examination of the validity of network equilibrium as a framework for network flow analysis [152]. Many of the current algorithms are based on the assumption that drivers become rational and homogeneous with learning. Hence, resulting in network equilibrium. A number of recent research efforts suggest that some drivers remain less rational, and heterogeneous drivers make up the system; drivers' attitudes

toward uncertainty become bipolar; and some drivers are sometimes deluded. Further research is required to characterize and model such behavior.

- Validate current models by comparing current model outputs with real world experiments, and possibly with controlled lab experiments (as mid-way experiments before conducting real world evaluations).
- Enhance traffic modeling tools within DTA models to capture the effect of diversion compliance of different vehicle modes (especially heavy vehicles) to more geometrically restrictive highways.
- Possibly calibrating hybrid fuzzy-stochastic models and comparing results to traditional models. According to the work done by Chen [52], probabilistic methods are better than possibility-based methods if sufficient information is available, on the other hand, possibility-based methods can be better if little information is available. However, when there is little information available about uncertainties, a hybrid method may be optimum.
- Conduct further research on the dynamic synthetic O-D estimation from link flow measurements and vehicle probe data. Further research is required to quantify the impact of erroneous or missing data on the accuracy of O-D estimates.
- Conduct further research on the temporal distribution of demand, analyzing and modeling it. Then, including the estimation and forecast of time-dependent demand within the planning process, in addition to the dynamic traffic management and control processes. This should, hopefully, help to fill-in the gap between the three mentioned processes.
- Incorporating person assignment, rather than mode assignment in the DTA and planning models, for as mentioned earlier, mode split is currently more of a daily real-time dynamic, rather than a static decision.
- Research is needed to develop models for driver behavior to different ATIS systems: (types and/or scenarios). Current literature is mainly based on stated preference surveys, which are known for their lack of accuracy. Before the deployment of ATIS systems, stated preference surveys were the best approach for prediction and modeling drivers' reactions. However now, after the deployment of many ATIS systems, more research is needed to capture the actual (possibly revealed) drivers' behavior, rather than the stated behavior.
- Develop approaches that are capable of realistically capturing traffic flow, traffic control, and their interactions; and simultaneously optimizing traffic flow routing and control. In other words, developing algorithms that are actually capable of capturing real-time driver behavior, and are able to control it, in order to improve network performance.

Appendix

Term Abbreviations

ANN	Artificial Neural Networks
ATIS	Advanced Traveler Information System
AVI	Automatic Vehicle Identification
AVL	Automatic Vehicle Location
DTA	Dynamic Traffic Assignment
FHWA	Federal Highway Administration
GA	Genetic Algorithm
GPS	Global Positioning System
HCM	Highway Capacity Manual
HOV	High Occupancy Vehicle
ITS	Intelligent Transportation Systems
LDV	Light Duty Vehicle
LMC	Link Marginal Cost
LP	Linear Programming
MOE	Measure of Effectiveness
NLP	Non-Linear Programming
O-D	Origin – Destination
PMC	Path Marginal Cost
SO	System Optimum
SOV	Single Occupancy Vehicle
TT	Travel Time
UE	User Equilibrium
VMS	Variable Message Sign

Variable Definitions

v_i	Traffic volume on route i
N	Set of network nodes
A	Set of network arcs (links)
R	Set of origin centroids
S	Set of destination centroids
k_{rs}	Set of paths connecting O-D pair $(r-s)$; $r \in R$, $s \in S$
x_a	Flow on arc (a)
x_b	Flow on arc (b)
t_a	Travel time on arc (a)
t_b	Travel time on arc (b)
f_k^{rs}	Flow on path (k) connecting O-D pair $(r\text{-}s)$
f_l^{mn}	Flow on path (l) connecting O-D pair $(m\text{-}n)$
c_k^{rs}	Travel time on path (k) connecting O-D pair $(r\text{-}s)$
q_{rs}	Trip rate between origin (r) and destination (s)
$\delta_{a,k}^{rs}$	Indicator variable, = 1 if arc (a) is on path (k) between O-D pair $(r\text{-}s)$, and 0 otherwise
x	Vector of flows on all arcs, $= (\ldots, x_a, \ldots)$
t	Vector of travel times on all arcs, $= (\ldots, t_a, \ldots)$

$\boldsymbol{f^{rs}}$ Vector of flows on all paths connecting O-D pair r-s, $= (\ldots, f_k^{rs}, \ldots)$
$\boldsymbol{f}$ Matrix of flows on all paths connecting all O-D pairs, $= (\ldots, \boldsymbol{f^{rs}}, \ldots)$
$\boldsymbol{c^{rs}}$ Vector of travel times on all paths connecting O-D pair *r-s*, $= (\ldots, c_k^{rs}, \ldots)$
$\boldsymbol{c}$ Matrix of travel times on all paths connecting all O-D pairs, $= (\ldots, \boldsymbol{c^{rs}}, \ldots)$
$\boldsymbol{q}$ Origin-destination matrix (with elements $= q_{rs}$)
$\boldsymbol{\Delta^{rs}}$ Link-path incidence matrix (with $\delta_{a,k}^{rs}$ elements) for O-D pair *r-s*, as discussed below
$\boldsymbol{\Delta}$ Matrix of link-path incidence matrices (for all O-D pairs), $= (\ldots, \Delta^{rs}, \ldots)$
$\boldsymbol{z}$ Objective function
$\boldsymbol{L}$ Lagrange (transformation of the) objective function
$\boldsymbol{u_{rs}}$ Dual variable associated with the flow conservation constraint for O-D pair (*r-s*)
$\boldsymbol{t_{i,k}}$ Observed average travel time along link i within the *k*th sampling interval
$\boldsymbol{\tilde{t}_{i,k}}$ Smoothed average travel time along link *i* in the *k*th sampling interval
$\boldsymbol{s_{i,k}^2}$ Variance of the observed travel times relative to the observed average travel time in the *k*th sampling interval
$\boldsymbol{\tilde{s}_{i,k}^2}$ Variance of the observed travel times relative to the smoothed travel time in the *k*th sampling interval
$\boldsymbol{n_{i,k}}$ Number of valid travel time readings on link *i* in the *k*th sampling interval
$\boldsymbol{\alpha}$ Exponential smoothing factor that varies as a function of the number of observations $n_{i,k}$ within the sampling interval
$\boldsymbol{\beta}$ Constant that varies between 0 and 1
$\boldsymbol{T_{ij}}$ Number of trips between production zone *i* and attraction zone *j*
$\boldsymbol{P_i}$ Number of trip productions from the origin zone
$\boldsymbol{A_j}$ Number of trip attractions to the destination zone
$\boldsymbol{F_{ij}}$ Impedance factor between production zone *i* and attraction zone *j*
$\boldsymbol{K_{ij}}$ Socio-economic adjustment factor for trips between production zone *i* and attraction zone *j*
$\boldsymbol{c_{ij}}$ Generalized cost of inter-zonal travel between production zone *i* and attraction zone *j*
$\boldsymbol{t_{ij}}$ Prior information on the number of trips between production zone *i* and attraction zone *j*
$\boldsymbol{V_a}$ Traffic flow on link (*a*)
$\boldsymbol{V_a'}$ Complementary traffic flow on link (*a*)
$\boldsymbol{p_{ij}^a}$ Probability of traffic flow between origin (*i*) and destination (*j*) to use link (*a*)
$\boldsymbol{T_r}$ Total demand departing during time-slice (*r*)
$\boldsymbol{t_r}$ Total seed matrix demand departing during time-slice (*r*)
$\boldsymbol{T_{rij}}$ Traffic demand departing during time-slice (*r*) traveling between origin (*i*) and destination (*j*)
$\boldsymbol{t_{rij}}$ Seed traffic demand departing during time-slice (*r*) traveling between origin (*i*) and destination (*j*)
$\boldsymbol{l_{rij}}$ Lagrange multiplier for departure time-slice, origin, and destination combination (*rij*)
$\boldsymbol{V_{sa}}$ Observed volume on link (*a*) during time-slice (*s*)
$\boldsymbol{p_{rij}^{sa}}$ Probability of (*a*) demand between origin (*i*) and destination (*j*) during time-slice (r) is observed on link (a) during time-slice (*s*)
$\boldsymbol{d(t_i)}$ Vehicle delay at time (t_i)
$\boldsymbol{u(t_i)}$ Vehicle instantaneous speed at time (t_i)
$\boldsymbol{u_f}$ Free-flow speed
$\boldsymbol{S(t_i)}$ Vehicle full and partial stops at time (t_i)
$\boldsymbol{MOE_e}$ Instantaneous fuel consumption or emission rate
$\boldsymbol{K_{i,j}^e}$ Model regression coefficient for MOE (*e*) at speed power (*i*) and acceleration power (*j*)
$\boldsymbol{L_{i,j}^e}$ Model regression coefficient for MOE (*e*) at speed power (*i*) and acceleration power (*j*) for positive accelerations
$\boldsymbol{M_{i,j}^e}$ Model regression coefficient for MOE (*e*) at speed power (*i*) and acceleration power (*j*) for negative accelerations
$\boldsymbol{u}$ Vehicle instantaneous speed
$\boldsymbol{a}$ Vehicle instantaneous acceleration rate
$\boldsymbol{q}$ Traffic stream flow (veh/h)
$\boldsymbol{k}$ Traffic stream density (veh/km)
$\boldsymbol{u}$ Traffic stream space-mean speed (km/h)
$\boldsymbol{u_f}$ Expected traffic stream free-flow speed (km/h)
$\boldsymbol{u_c}$ Expected traffic stream speed-at-capacity (km/h)
$\boldsymbol{k_j}$ Expected traffic stream jam density (veh/km)
$\boldsymbol{q_c}$ Expected traffic stream capacity (veh/km)
$\boldsymbol{c_1}$ Model coefficient (km/veh)
$\boldsymbol{c_3}$ Model constant (h/km^2 -veh)
$\boldsymbol{c_2}$ Model constant (h^{-1})

Bibliography

1. Abdel-Aty MA, Kitamura R, Jovanis PP (1997) Using Stated Preference Data for Studying the Effect of Advanced Traffic Information on Drivers' Route Choice. Transp Res Part C:(Emerg Technol) 5(1):39–50
2. Abdelfatah AS, Mahmassani HS (2001) A simulation-based signal optimization algorithm within a dynamic traffic assignment framework. IEEE Intelligent Transportation Systems Proceedings, IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC 2001, Oakland
3. Abdelghany AF, Abdelghany KF et al (2000) Dynamic traffic assignment in design and evaluation of high-occupancy toll lanes. Transp Res Rec 1733:39–48

4. Abdelghany KF, Mahmassani HS (2001) Dynamic trip assignment-simulation model for intermodal transportation networks. Transp Res Rec 1771:52–60
5. Abdelghany KF, Valdes DM et al (1999) Real-time dynamic traffic assignment and path-based signal coordination: Application to network traffic management. Transp Res Rec 1667:67–76
6. Abdulhai B, Porwal H, Recker W (2002) Short-term Freeway Traffic Flow Prediction Using Genetically Optimized Time-Delay-Based Neural Networks. ITS J: Intell Transp Syst J 7(1): 3–41
7. Ahmed M, Cook (1982) Analysis of Freeway Traffic Time Series Data by Using Box-Jenkins Techniques. Transp Res Rec 722:1–9
8. Ahn K, Rakha H et al (2002) Estimating vehicle fuel consumption and emissions based on instantaneous speed and acceleration levels. J Transp Eng 128(2):182–190
9. Ahn K, Rakha H et al (2004) Microframework for modeling of high-emitting vehicles. Transp Res Rec 1880:39–49
10. Akcelik R, Rouphail NM (1994) Overflow queues and delays with random and platooned arrivals at signalized intersections. J Adv Transp 28(3):227–251
11. Allen RW, Stein AC, Rosenthal TJ, Ziedman D, Torres JF, Halati A (1991) Human factors simulation investigation of driver route diversion and alternate route selection using in-vehicle navigation systems. In: Vehicle Navigation & Information Systems Conference, Dearborn, 20–23 Oct 1991. Proceedings Part 1 (of 2) Society of Automotive Engineers. SAE, Warrendale, pp 9–26
12. Anastassopoulos I (2000) Fault-Tolerance and Incident Detection using Fourier Transforms. Purdue University, Westlafayette
13. Arafeh M, Rakha H (2005) Genetic Algorithm Approach for Locating Automatic Vehicle Identification Readers. In: IEEE Intelligent Transportation System Conference, Vienna, 2005. Proceedings ITSV'05 IEEE Intelligent Conference on Transportations Systems, p 1153–1158
14. Arnott R, de Palma A, Lindsey R (1991) Does providing information to drivers reduce traffic congestion? Transp Res Part A (General) 25A(5):309
15. Arrow KJ (1951) Alternative Approaches to the Theory of Choice in Risk-Taking Situations. Econometrica 19(4):404–437
16. Ashok K (1996) Estimation and Prediction of Time-Dependent Origin-Destination Flows. Boston, Massachusetts Institute of Technology. Ph D Thesis, Massachusetts Institute of Technology
17. Ashok K, Ben-Akiva ME (1993) Dynamic Origin-Destination Matrix Estimation and Prediction for Real-Time Traffic Management Systems. In: Daganzo CF (ed) 12th International Symposium on Transportation and Traffic Theory. Elsevier, New York, pp 465–484
18. Ashok K, Ben-Akiva ME (2000) Alternative approaches for real-time estimation and prediction of time-dependent Origin-Destination flows. Transp Sci 34(1):21–36
19. Balakrishna R, Koutsopoulos HN et al (2005) Simulation-Based Evaluation of Advanced Traveler Information Systems. Transp Res Rec 1910:90–98
20. Barth M, An F et al (2000) Comprehensive Modal Emission Model (CMEM): Version 2.0 User's Guide. University of California, Riverside
21. Bell M, Iida Y (1997) Transportation network analysis. Iida Y translator. Wiley, Chichester / New York
22. Ben-Akiva M, Bierlaire M et al (1998) DynaMIT: a simulation-based system for traffic prediction. DACCORD Short Term Forecasting Workshop, Delft, February 1998
23. Ben-Akiva M, Bolduc D et al (1993) Estimation of travel choise models with randomly distributed values of time. Transp Res Rec 1413:88–97
24. Ben-Akiva M, Kroes E et al (1992) Real-Time Prediction of Traffic Congestion. Vehicle Navigation and Information Systems, IEEE, New York
25. Ben-Akiva M, Bierlaire M, Bottom J, Koutsopoulos H et al (1997) Development Of A Route Guidance Generation System For Real-Time Application. In: 8th International Federation of Automatic Control Symposium on Transportation Systems, Chania, 16–18 June 1997
26. Ben-Akiva MMB, Koutsopoulos H, Mishalani R (1998) DynaMIT: a simulation-based system for traffic prediction. DACCORD Short Term Forecasting Workshop, Delft
27. Bierlaire M, Crittin F (2004) An efficient algorithm for real-time estimation and prediction of dynamic OD tables. Oper Res 52(1):116–27
28. Birge JR, Ho JK (1993) Optimal flows in stochastic dynamic networks with congestion. Oper Res 41(1):203–216
29. Bolland JD, Hall MD et al (1979) SATURN: Simulation and Assignment of Traffic in Urban Road Networks. In: International Conference on Traffic Control Systems, Berkeley
30. Boyce DE, Ran B, Leblanc LJ (1995) Solving an instantaneous dynamic user-optimal route choice model. Transp Sci 29(2):128–142
31. Braess D (1968) Über ein Paradoxon der Verkehrsplanung. Unternehmensforschung 12:258–268
32. Brilon W (1995) Delays at oversaturated unsignalized intersections based on reserve capacities. Transp Res Rec 1484:1–8
33. Brilon W, Wu N (1990) Delays at fixed-time traffic signals under time-dependent traffic conditions. Traffic Eng & Control 31(12):8
34. Burell JE (1968) Multipath Route Assignment and its Application to Capacity-Restraint. Fourth International Symposium on the Theory of Traffic Flow, Karlsruhe
35. Burell JE (1976) Multipath Route Assignment: A Comparison of Two Methods. In: Florian M (ed) Traffic Equilibrium Methods. Lectue Notes in Economics and Mathematical Systems, vol 118. Springer, New York, pp 210–239
36. Busemeyer JR, Townsend JT (1993) Decision Field Theory: A Dynamic-Cognitive Approach to Decision Making in an Uncertain Environment. Psychol Rev 100(3):432
37. Byung-Wook Wie TRL, Friesz TL, Bernstein D (1995) A discrete time, nested cost operator approach to the dynamic network user equilibrium problem. Transp sci 29(1):79–92
38. Cantarella GE, Cascetta ES (1995) Dynamic processes and equilibrium in transportation networks: towards a unifying theory. Transp Sci 29(4):305–329
39. Carey M (1986) Constraint qualification for a dynamic traffic assignment model. Transp Sci 20(1):55–58
40. Carey M (1987) Optimal time-varying flows on congested networks. Oper Res 35(1):58–69
41. Carey M (1992) Nonconvexity of the dynamic traffic assignment problem. Transp Res Methodol 26B(2):127
42. Carey M, Subrahmanian E (2000) An approach to mod-

elling time-varying flows on congested networks. Transp Res Methodol 34B(3)
43. Cascetta E, Marquis G (1993) Dynamic estimators of origin-destination matrices using traffic counts. Transp Sci 27(4):363–373
44. Cassidy MJ, Han LD (1993) Proposed model for predicting motorist delays at two-lane highway work zones. J Transp Eng 119(1):27–42
45. Cassidy MJ, Rudjanakanoknad J (2005) Increasing the capacity of an isolated merge by metering its on-ramp. Transp Res Part B Methodol 39(10):896–913
46. Cassidy MJ, Son Y et al (1994) Estimating motorist delay at two-lane highway work zones. Transp Res Part A, Policy Pract 28(5):433–444
47. Cassidy MJ, Windover JR (1995) Methodology for assessing dynamics of freeway traffic flow. Transp Res Rec 1484:73–79
48. Castillo E, Menendez JM, Jimenez P (2008) Trip matrix and path flow reconstruction and estimation based on plate scanning and link observations. Transp Res Part B: Methodol 42(5):455–481
49. Catling I (1977) A Time-Dependent Approach to Junction Delays. Traffic Eng Control 18(11):520–523:526
50. Chang GL, Mahmassani HS (1988) Travel Time Prediction And Departure Time Adjustment Behavior Dynamics In A Congested Traffic System. Transp Res, Part B: Methodol 22B(3):217–232
51. Chang GL, Tao X (1999) Integrated model for estimating time-varying network origin-destination distributions. Transp Res, Part A: Policy Pract 33(5):381–399
52. Chen SQ (2000) Comparing Probabilistic and Fuzzy Set Approaches for Design in the Presence of Uncertainty. In: Aerospace and Ocean Engineering. Ph D, Polytechnic Institute and State University, Blacksburg
53. Chiu YC, Mahmassani HS (2001) Toward hybrid dynamic traffic assignment-models and solution procedures. In: IEEE Intelligent Transportation Systems Proceedings, IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC 2001, Oakland
54. Coifman B (1998) New algorithm for vehicle reidentification and travel time measurement on freeways. In: Proceedings of the 1998 5th International Conference on Applications of Advanced Technologies in Transportation, Newport Beach, Proceedings of the International Conference on Applications of Advanced Technologies in Transportation Engineering, ASCE, Reston
55. Coifman B, Banerjee B (2002) Vehicle reidentification and travel time measurement on freeways using single loop detectors-from free flow through the onset of congestion. In: Proceedings of the seventh International Conference on: Applications of Advanced Technology in Transportation, Cambridge, 5-7 Aug 2002. Proceedings of the International Conference on Applications of Advanced Technologies in Transportation Engineering. American Civil Engineers
56. Coifman B, Cassidy M (2001) Vehicle reidentification and travel time measurement, Part I: Congested freeways. In: IEEE Intelligent Transportation Systems Proceedings, Conference IEEE on Intelligent Transportation Systems, Proceedings, ITSC 2001, Oakland
57. Coifman B, Ergueta E (2003) Improved vehicle reidentification and travel time measurement on congested freeways. J Transp Eng 129(5):475–483
58. Colyar JD, Rouphail NM (2003) Measured Distributions of Control Delay on Signalized Arterials. Transp Res Rec 1852:1–9
59. Cremer M, Keller H (1987) New Class Of Dynamic Methods For The Identification Of Origin-Destination Flows. Transp Res, Part B: Methodol 21(2):117–132
60. Cronje WB (1983) Analysis of Existing Formulas for Delay, Overflow, and Stops. Transp Res Rec 905:89–93
61. Cronje WB (1983) Derivation of Equations for Queue Length, Stops, and Delay for Fixed-Time Traffic Signals. Transp Res Rec 905:93–95
62. Cronje WB (1983) Optimization Model for Isolated Signalized Traffic Intersections. Transp Res Rec 905:80–83
63. Cronje WB (1986) Comparative Analysis of Models for Estimating Delay for Oversaturated Conditions at Fixed-Time Traffic Signals. Transp Res Record 1091:48–59
64. Dafermos S (1980) Traffic equilibrium and variational inequalities. Transp Sci 14(1):42–54
65. Daganzo CF, Laval JA (2005) On the numerical treatment of moving bottlenecks. Transp Res Part B Methodol 39(1):31–46
66. Daniel J, Fambro DB et al (1996) Accounting for nonrandom arrivals in estimate of delay at signalized intersections. Transp Res Rec 1555:9–16
67. Dantzig GB (1957) The Shortest Route Problem. Oper Res 5:270–273
68. Dial R (1971) A Probabilistic Multipath Traffic Assignment Model which Obviates Path Enumeration. Transp Res 5: 83–111
69. Dijkstra EW (1959) A Note on Two Problems in Connection with Graphics. Numeriche Math 1:209–271
70. Dion F, Rakha H (2006) Estimating dynamic roadway travel times using automatic vehicle identification data for low sampling rates. Transp Res Part B 40:745–766
71. Dion F, Rakha H et al (2004) Comparison of delay estimates at under-saturated and over-saturated pre-timed signalized intersections. Transp Res Part B Methodol 38(2):99–122
72. Dion F, Rakha H et al (2004) Evaluation of potential transit signal priority benefits along a fixed-time signalized arterial. J Transp Eng 130(3):294–303
73. Elefteriadou L, Fang C et al (2005) Methodology for evaluating the operational performance of interchange ramp terminals. Transp Res Rec 1920:13–24
74. Engelbrecht RJ, Fambro DB et al (1996) Validation of generalized delay model for oversaturated conditions. Transp Res Rec 1572:122–130
75. Evans JL, Elefteriadou L et al (2001) Probability of breakdown at freeway merges using Markov chains. Transp Res Part B Methodol 35(3):237–254
76. Fambro DB, Rouphail NM (1996) Generalized delay model for signalized intersections and arterial streets. Transp Res Rec 1572:112–121
77. Fang FC, Elefteriadou L et al (2003) Using fuzzy clustering of user perception to define levels of service at signalized intersections. J Transp Eng 129(6):657–663
78. Fisk C (1979) More Paradoxes in the Equilibrium Assignment Problem. Transp Res 13B:305–309
79. Flannery A, Kharoufeh JP et al (2005) Queuing delay models for single-lane roundabouts. Civ Eng & Environ Syst 22(3):133–150
80. Frank M (1981) The Braess Paradox. Math Program 20: 283–302

81. Frank M, Wolfe P (1956) An Algorithm of Quatdratic Programming. Nav Res Logist 3:95–110
82. Friesz TL, DB, Smith TE, Tobin RL, Wie BW (1993) A Variational Inequality Formulation of the Dynamic Network User Equilibrium Problem. Oper Res 41(1):179–191
83. Friesz TL, JL, Tobin RL, Wie B-W (1989) Dynamic network traffic assignment considered as a continuous time optimal control problem. Oper Res 37(6):893–901
84. Ghali MO, Smith MJ (1995) A model for the Dynamic System Optimum Traffic Assignment Problem. Transp Res 29B(3):155–170
85. Greenshields BD (1934) A study of traffic capacity. In: Proc Highway Research Board 14:448–477
86. Hagring O, Rouphail NM et al (2003) Comparison of Capacity Models for Two-Lane Roundabouts. Transp Res Rec 1852:114–123
87. Hall MD, Van Vliet D et al (1980) SATURN - A Simulation-Assignment Model fo the Evalaution of Traffic Management Schemes. Traffic Eng Control 4:167–176
88. Hawas YE (1995) A Decentralized Architecture And Local Search Procedures For Real-Time Route Guidance In Congested Vehicular Traffic Networks. University of Texas, Austin
89. Hawas YE (2004) Development and calibration of route choice utility models: Neuro-fuzzy approach. J transp eng 130(2):171–182
90. Hawas YE, Mahmassani HS (1995) A Decentralized Scheme For Real-Time Route Guidance In Vehicular Traffic Networks. In: Second World Congress on Intelligent Transport Systems, Yokohama, 1995, p 1965–1963
91. Hawas YE, Mahmassani HS (1997) Comparative analysis of robustness of centralized and distributed network route control systems in incident situations. Transp Res Rec 1537:83–90
92. Hawas YE, Mahmassani HS, Chang GL, Taylor R, Peeta S, Ziliaskopoulos A (1997) Development of Dynasmart-X Software for Real-Time Dynamic Traffic Assignment. Center for Transportation Research, The University of Texas, Austin
93. Hellinga BR, Van Aerde M (1998) Estimating dynamic O-D demands for a freeway corridor using loop detector data. Canadian Society for Civil Engineering, Halifax, Montreal
94. Ho JK (1980) A successive linear optimization approach to the dynamic traffic assignment problem. Transp Sci 14(4): 295–305
95. Hu SR, Madanat SM, Krogmeier JV, Peeta S (2001) Estimation of dynamic. assignment matrices and OD demands using adaptive Kalman Filtering. Intell Transp Syst J 6:281–300
96. Huey-Kuo C, Che-Fu H (1998) A model and an algorithm for the dynamic user-optimal route choice problem. Transp Res, Part B Methodol 32B(3):219–234
97. Ishak S, Al-Deek H (2003) Performance Evaluation of a Short-Term Freeway Traffic Prediction Model. Transportation Research Board 82nd Annual Meeting, Washington DC
98. Janson BN (1991) Convergent Algorithm for Dynamic Traffic Assignment. Transp Res Rec 1328:69–80
99. Janson BN (1991) Dynamic Traffic Assignment For Urban Road Networks. Transp Res, Part B Methodol 25B:2–3
100. Jayakrishnan R, Mahmassani HS (1990) Dynamic simulation-assignment methodology to evaluate in-vehicle information strategies in urban traffic networks. Winter Simulation Conference Proceedings, New Orleans 1990. 90 Winter Simulation Conf Winter Simulation Conference Proceedings. IEEE, Piscataway (IEEE cat n 90CH2926–4)
101. Jayakrishnan R, Mahmassani HS (1991) Dynamic modelling framework of real-time guidance systems in general urban traffic networks. In: Proceedings of the 2nd International Conference on Applications of Advanced Technologies in Transportation Engineering, Minneapolis. ASCE, New York
102. Jayakrishnan R, Mahmassani HS et al (1993) User-friendly simulation model for traffic networks with ATIS/ATMS. Proceedings of the 5th International Conference on Computing in Civil and Building Engineering - V ICCCBE, Anaheim 1993. ASCE, New York
103. Jeffery DJ (1981) The Potential Benefits of Route Guidance. TRRL, Department of Transportation, Crowthorne
104. Jha M, Madanat S, Peeta S (1998) Perception updating and day-to-day travel choice dynamics in traffic networks with information provision. Transp Res Part C Emerg Technol 6C(3):189–212
105. Katsikopoulos KV, Duse-Anthony Y et al (2000) The Framing of Drivers' Route Choices when Travel Time Information Is Provided under Varying Degrees of Cognitive Load. J Hum Factors Ergonomics Soc 42(3):470–481
106. Kerner BS (2004) The physics of traffic. Springer, Berlin
107. Kerner BS (2004) Three-phase traffic theory and highway capacity. Physica A 333(1-4):379–440
108. Kerner BS (2005) Control of spatiotemporal congested traffic patterns at highway bottlenecks. Physica A 355(2-4):565–601
109. Kerner BS, Klenov SL (2006) Probabilistic breakdown phenomenon at on-ramp bottlenecks in three-phase traffic theory: Congestion nucleation in spatially non-homogeneous traffic. Physica A 364:473–492
110. Kerner BS, Rehborn H et al (2004) Recognition and tracking of spatial-temporal congested traffic patterns on freeways. Transp Res Part C Emerging Technologies 12(5):369–400
111. Khattak AJ, Schofer JL, Koppelman FS (1993) Commuters' en-route diversion and return decisions: analysis and implications for advanced traveller information systems. Transp Res (Policy and Practice) 27A(2):101
112. Kim H, Baek S et al (2001) Origin-destination matrices estimated with a genetic algorithm from link traffic counts. Transp Res Rec 1771:156–163
113. Koutsopoulos HN, Polydoropoulou A et al (1995) Travel simulators for data collection on driver behavior in the presence of information. Transp Res Part C Emerging Technologies 3(3):143
114. Krishnamurthy S, Coifman B (2004) Measuring freeway travel times using existing detector infrastructure. In: Proceedings - 7th International IEEE Conference on Intelligent Transportation Systems, ITSC, Washington, 2004
115. Laval JA, Daganzo CF (2006) Lane-changing in traffic streams. Transp Res Part B Methodol 40(3):251–264
116. Lawson TW, Lovell DJ et al (1996) Using input-output diagram to determine spatial and temporal extents of a queue upstream of a bottleneck. Transp Res Rec 1572:140–147
117. LeBlanc LJ (1975) An Algorithm for Discrete Network Design Problem. Tranp Sci 9:183–199
118. LeBlanc LJ, Abdulaal M (1970) A Comparison of User-Optimum versus System-Optimum Traffic Assignment in Transportation Network Design. Transp Res 18B:115–121
119. LeBlanc LJ, Morlok EK et al (1974) An Accurate and Efficient Approach to Equilibrium Traffic Assignment on Congested Networks. Transp Res Rec 491:12–23
120. Lee S, Fambro D (1999) Application of the Subset ARIMA

Model for Short-Term Freeway Traffic vol Forecasting. Transp Res Rec 1678:179–188
121. Leonard DR, Tough JB et al (1978) CONTRAM - A Traffic Assignment Model for Predicting Flows and Queues During Peak Periods. TRRL SR 568. Transport Research Laboratory, Crowthome
122. Lertworawanich P, Elefteriadou L (2001) Capacity estimations for type B weaving areas based on gap acceptance. Transp Res Rec 1776:24–34
123. Lertworawanich P, Elefteriadou L (2003) A methodology for estimating capacity at ramp weaves based on gap acceptance and linear optimization. Transp Res Part B Methodol 37(5):459–483
124. Li J, Rouphail NM et al (1994) Overflow delay estimation for a simple intersection with fully actuated signal control. Transp Res Rec 1457:73–81
125. Li Y (2001) Development of Dynamic Traffic Assignment Models for Planning Applications. Northwestern University, Evanston
126. Lighthill MJ, Witham GB (1955) On Kinematic Waves. I: Flood Movement in Long Rivers, II. A Theory of Traffic Flow on Long Crowded Roads. In: Proceedings of the Royal Society of London A 229, pp 281–345
127. Lorenz MR, Elefteriadou L (2001) Defining freeway capacity as function of breakdown probability. Transp Res Rec 1776: 43–51
128. Lotan T (1997) Effects of familiarity on route choice behavior in the presence of information. Transp Res Part C Emerg Technol 5(3-4):225–243
129. Mahmassani H, Jou R-C (2000) Transferring insights into commuter behavior dynamics from laboratory experiments to ®eld surveys. Transp Res Part A Policy Pract 34A(4):243–260
130. Mahmassani H, Peeta S (1992) System optimal dynamic assignment for electronic route guidance in a congested traffic network. In: Gartner NH, Improta G (eds) Urban Traffic Networks. Dynamic Flow Modelling and Control. Springer, Berlin, pp 3–37
131. Mahmassani HS, Chiu Y-C, Chang GL, Peeta S, Ziliaskopoulos A (1998) Off-line Laboratory Test Results for the DYNASMART-X Real-Time Dynamic Traffic Assignment System. Center for Transportation Research, The University of Texas, Austin
132. Mahmassani HS, Hawas Y, Abdelghany K, Abdelfatah A, Chiu Y-C, Kang Y, Chang GL, Peeta S, Taylor R, Ziliaskopoulos A (1998) DYNASMART-X, vol II: Analytical and Algorithmic Aspects. Center for Transportation Research, The University of Texas, Austin
133. Mahmassani HS, Hawas Y, Hu T-Y, Ziliaskopoulos A, Chang G-L, Peeta S, Taylor R (1998) Development of Dynasmart-X Software for Real-Time Dynamic Traffic Assignment. Technical Report ST067-85-Tast E (revised) submitted to Oak Ridge National Laboratory under subcontract 85X-SU565C,
134. Mahmassani HS, Peeta S (1993) Network Performance under System Optimal and User Equilibrium Dynamic Assignments: Implications for ATIS. Transp Res Rec 1408:83–93
135. Mahmassani HS, Peeta S (1995) System Optimal Dynamic Assignment for Electronic Route Guidance in a Congested Traffic Network. In: Gartner NH, Improta G (eds) URBAN TRAFFIC NETWORKS: Dynamic Flow Modeling and Control. Springer, Berlin, pp 3–37
136. Mahmassani HS, Peeta S, Hu T, Ziliaskopoulos A (1993) Algorithm for Dynamic Route Guidance in Congested Networks with Multiple User Information Availability Groups. In: 26th International Symposium on Automotive Technology and Automation, Aachen
137. Matsoukis EC (1986) Road Traffic Assignment - A Review Part I: Non-Equilibrium Methods. Transp Plan Technol 11:69–79
138. Matsoukis EC, Michalopolos PC (1986) Road Traffic Assignment - A Review Part II: Equilibrium Methods. Transp Plan Technol 11:117–135
139. Mekky A (1995) Toll revenue and traffic study of highway 407 in Toronto. Transp Res Rec 1498:5–15
140. Mekky A (1996) Modeling toll pricing strategies in greater Toronto areas. Transp Res Rec 1558:46–54
141. Mekky A (1998) Evaluation of two tolling strategies for Highway 407 in Toronto. Transp Res Rec 1649:17–25
142. Merchant DK, Nemhauser GL (1978) A Model And An Algorithm For The Dynamic Traffic Assignment Problems. Transp Sci 12(3):183–199
143. Merchant DK, Nemhauser GL (1978) Optimality Conditions For A Dynamic Traffic Assignment Model. Transp Sci 12(3):200–207
144. Minderhoud MM, Elefteriadou L (2003) Freeway Weaving: Comparison of Highway Capacity Manual 2000 and Dutch Guidelines. Transp Res Rec 1852:10–18
145. Moskowitz K (1956) California Method for Assigning Directed Traffic to Proposed Freeways. Bull Highw Res Board 130:1–26
146. Munnich LW Jr, Hubert HH et al (2007) L-394 MnPASS high-occupancy toll lanes planning and operational issues and outcomes (lessons learning in year 1). Transp Res Rec 1996:49–57
147. Murchland JD (1970) Braess's Paradox of Traffic Flow. Transp Res 4:391–394
148. Nagel K (1996) Particle Hopping Model and Traffic Flow Theory. Phys Rev E 53 (5):4655–4672
149. Nagel K, Schrekenberg M (1992) Cellular Automaton Model for Freeway Traffic. J Phys 2(20):2212–2229
150. Nagel K, Schrekenberg M (1995) Traffic Jam Dynamics in Stochastic Cellular Automata. US D Energy, Los Alamos National Laboratory, LA-UR-95-2132, Los Alamos
151. Nakayama S, Kitamura R (2000) Route choice model with inductive learning. Transp Res Rec 1725:63–70
152. Nakayama S, Kitamura R et al (2001) Drivers' route choice rules and network behavior: Do drivers become rational and homogeneous through learning? Transp Res Rec 1752:62–68
153. Newell GF (1965) Approximation Methods for Queues with Application to the Fixed-Cycle Traffic Light. SIAM Rev 7: 223–240
154. Newell GF (1999) Delays caused by a queue at a freeway exit ramp. Transp Res Part B Methodol 33(5):337–350
155. Nguyen S (1969) An Algorithm for the Assignment Problem. Transp Sci 8:203–216
156. Nie Y, Zhang HM et al (2005) Inferring origin-destination trip matrices with a decoupled GLS path flow estimator. Transp Res Part B Methodol 39(6):497–518
157. Noonan J, Shearer O (1998) Intelligent Transportation Systems Field Operational Test: Cross-Cutting Study Advance Traveler Information Systems. US Department of Transportation, Federal Highways Administration, Intelligent Transportation System, Washington, DC
158. Okutani I (1987) The Kalman Filtering Approaches in Some Transportation and Traffic Problems. In: Proceedings of the

Tenth International Symposium on Transportation and Traffic Theory. Elsevier, New York
159. Park B (2002) Hybrid neuro-fuzzy application in short-term freeway traffic vol forecasting. Transp Res Rec 1802:190–196
160. Park D, Rilett LR (1998) Forecasting multiple-period freeway link travel times using modular neural networks. Transp Res Rec 1617:163–170
161. Park D, Rilett LR (1999) Forecasting freeway link travel times with a multilayer feedforward neural network. Comput-Aided Civ & Infrastruct Eng 14(5):357–367
162. Park D, Rilett LR et al (1998) Forecasting multiple-period freeway link travel times using neural networks with expanded input nodes. In: Proceedings of the 1998 5th International Conference on Applications of Advanced Technologies in Transportation, Newport Beach and Proceedings of the International Conference on Applications of Advanced Technologies in Transportation Engineering 1998, ASCE, Reston
163. Park D, Rilett LR et al (1999) Spectral basis neural networks for real-time travel time forecasting. J Transp Eng 125(6): 515–523
164. Park S, Rakha H (2006) Energy and Environmental Impacts of Roadway Grades. Transp Res Rec 1987:148–160
165. Pavlis Y, Papageorgiou M (1999) Simple decentralized feedback strategies for route guidance in traffic networks. Transp Sci 33(3):264–278
166. Peeta S (1994) System Optimal Dynamic Traffic Assignment in Congested Networks with Advanced Information Systems. University of Texas, Austin
167. Peeta S, Bulusu S (1999) Generalized singular value decomposition approach for consistent on-line dynamic traffic assignment. Transp Res Rec 1667
168. Peeta S, Mahmassani HS (1995) Multiple user classes real-time traffic assignment for online operations: a rolling horizon solution framework. Transp Res Part C Emerg Technol 3C(2):83
169. Peeta S, Mahmassani HS (1995) System optimal and user equilibrium time-dependent traffic assignment in congested networks. Ann Oper Res 60:81–113
170. Peeta S, Mahmassani HS et al (1991) Effectiveness of real-time information strategies in situations of non-recurrent congestion. In: Proceedings of the 2nd International Conference on Applications of Advanced Technologies in Transportation Engineering, Minneapolis. ASCE, New York
171. Peeta S, Paz A (2006) Behavior-consistent within-day traffic routing under information provision. In: IEEE Intelligent Transportation Systems Conference, Toronto, pp 212–217
172. Peeta S, Ramos JL (2006) Driver response to variable message signs-based traffic information. Intell Transp Syst 153(1):2–10
173. Peeta S, Ramos JL, Pasupathy R (2000) Content of Variable Message Signs and On-line Driver Behavior. Transp Res Rec 1725:102–108
174. Peeta S, Yang T-H (2000) Stability of Large-scale Dynamic Traffic Networks under On-line Control Strategies. In: 6th International Conference on Applications of Advanced Technologies in Transportation Engineering, Singapore,paper no. 11 (eProceedings on CD), p 9
175. Peeta S, Yang T-H (2003) Stability Issues for Dynamic Traffic Assignment. Automatica 39(1):21–34
176. Peeta S, Yu JW (2004) Adaptability of a Hybrid Route Choice Model to Incorporating Driver Behavior Dynamics Under Information Provision. In: IEEE Transactions On Systems, Man, And Cybernetics Part A: Systems And Humans 34(2):243–256
177. Peeta S, Yu JW (2006) Behavior-based consistency-seeking models as deployment alternatives to dynamic traffic assignment models. Transp Res Part C Emerg Technol 14(2): 114–138
178. Peeta S, Zhou C (1999) On-Line Dynamic Update Heuristics for Robust Guidance. In: International Conference Modeling and Management in Transportation, Cracow, October 1999
179. Peeta S, Zhou C (1999) Robustness of the Off-line A Priori Stochastic Dynamic Traffic Assignment Solution for On-Line Operations. Transp Res Part C: Emerg Technol 7C(5):281–303
180. Peeta S, Ziliaskopoulos AK (2001) Foundations of Dynamic Traffic Assignment: The Past, the Present and the Future. Netw Spat Econ 1(3-4):233
181. Rakha H (1990) An Evaluation of the Benefits of User and System Optimised Route Guidance Strategies. In: Civil Engineering. Queen's University, Kingston
182. Rakha H, Ahn K (2004) Integration modeling framework for estimating mobile source emissions. J transp eng 130(2): 183–193
183. Rakha H, Ahn K et al (2004) Development of VT-Micro model for estimating hot stabilized light duty vehicle and truck emissions. Transp Res Part D Transport and Environment 9(1):49–74
184. Rakha H, Arafeh M (2007) Tool for calibrating steady-state traffic stream and car-following models. In: Transportation Research Board Annual Meeting, Washington, 22–25 Jan 2008
185. Rakha H, Crowther B (2002) Comparison of Greenshields, Pipes, and Van Aerde car-following and traffic stream models. Transp Res Rec 1802:248–262
186. Rakha H, Flintsch AM et al (2005) Evaluating alternative truck management strategies along interstate 81. Transp Res Rec 1925:76–86
187. Rakha H, Kang Y-S et al (2001) Estimating vehicle stops at undersaturated and oversaturated fixed-time signalized intersections. Transp Res Rec 1776:128–137
188. Rakha H, Lucic I (2002) Variable power vehicle dynamics model for estimating maximum truck acceleration levels. J Transp Eng 128(5):412–419
189. Rakha H, Lucic I et al (2001) Vehicle dynamics model for predicting maximum truck acceleration levels. J Transp Eng 127(5):418–425
190. Rakha H, Medina A et al (2000) Traffic signal coordination across jurisdictional boundaries: Field evaluation of efficiency, energy, environmental, and safety impacts. Transp Res Rec 1727:42–51
191. Rakha H, Paramahamsan H et al (2005) Comparison of Static Maximum Likelihood Origin-Destination Formulations. Transportation and Traffic Theory: Flow, Dynamics and Human Interaction. In: Proceedings of the 16th International Symposium on Transportation and Traffic Theory (ISTTT16), pp 693–716
192. Rakha H, Pasumarthy P et al (2004) Modeling longitudinal vehicle motion: issues and proposed solutions. In: Transport Science and Technology Congress, Athens, Sep 2004
193. Rakha H, Pasumarthy P et al (2004) The INTEGRATION framework for modeling longitudinal vehicle motion. TRANSTEC, Athens
194. Rakha H, Snare M et al (2004) Vehicle dynamics model for estimating maximum light-duty vehicle acceleration levels. Transp Res Rec 1883:40–49

195. Rakha H, Van Aerde M et al (1989) Evaluating the benefits and interactions of route guidance and traffic control strategies using simulation. In: First Vehicle Navigation and Information Systems Conference - VNIS '89, Toronto. IEEE, Piscataway
196. Rakha H, Van Aerde M et al (1998) Construction and calibration of a large-scale microsimulation model of the Salt Lake area. Transp Res Rec 1644:93–102
197. Rakha H, Zhang Y (2004) INTEGRATION 2.30 framework for modeling lane-changing behavior in weaving sections. Transp Res Rec 1883:140–149
198. Rakha H, Zhang Y (2004) Sensitivity analysis of transit signal priority impacts on operation of a signalized intersection. J Transp Eng 130(6):796–804
199. Rakha HA, Van Aerde MW (1996) Comparison of simulation modules of TRANSYT and integration models. Transp Res Rec 1566:1–7
200. Ran B, Boyce DE (1996) A link-based variational inequality formulation of ideal dynamic user-optimal route choice problem. Research Part C Emerg Technol 4C(1):1–12
201. Ran B, Boyce DE (1996) A link-based variational inequality formulation of ideal dynamic user-optimal route choice problem. Research Part C (Emerging Technologies) 4C(1):1–11
202. Ran B, Boyce DE, LeBlanc LJ (1993) A new class of instantaneous dynamic user-optimal traffic assignment models. Oper Res 41(1):192–202
203. Ran B, Hall RW, Boyce DE (1996) A link-based variational inequality model for dynamic departure time/route choice. Transp Res Methodol 30B(1):31–46
204. Ran B, Shimazaki T (1989) A general model and algorithm for the dynamic traffic assignment problems. In: Fifth World Conference on Transport Research, Transport Policy, Management and Technology Towards, Yokohama, 2001
205. Ran B, Shimazaki T (1989) Dynamic user equilibrium traffic assignment for congested transportation networks. In: Fifth World Conference on Transport Research, Yokohama, 1989
206. Randle J (1979) A Convergence Probabilistic Road Assignment Model. Traffic Eng Control 11:519–521
207. Richards PI (1956) Shock waves on the highway. Oper Res 4:42–51
208. Rilett L, Aerde V (1993) Modeling Route Guidance Using the Integration Model. In: Proceedings of the Pacific Rim Trans Tech Conference, Seattle, 1993 and Proceedings of the ASCE International Conference on Applications of Advanced Technologies in Transportation Engineering. ASCE, New York
209. Rilett L, Van Aerde M (1991) Routing based on anticipated travel times. In: Proceedings of the 2nd International Conference on Applications of Advanced Technologies in Transportation Engineering, Minneapolis. ASCE, New York
210. Rilett LR, Van Aerde M et al (1991) Simulating the TravTek route guidance logic using the integration traffic model. In: Vehicle Navigation & Information Systems Conference Proceedings Part 2 (of 2). Dearborn, 1991. In: Proceedings - Society of Automotive Engineers n P-253, SAE. Warrendale
211. Rilett LR, van Aerde MW (1991) Modelling distributed real-time route guidance strategies in a traffic network that exhibits the Braess paradox. In: Vehicle Navigation & Information Systems Conference Proceedings Part 2 (of 2). Dearborn, 1991. Proceedings - Society of Automotive Engineers n P-253. SAE, Warrendale
212. Rouphail NM (1988) Delay Models for Mixed Platoon and Secondary Flows. J Transp Eng 114(2):131–152
213. Rouphail NM, Akcelik R (1992) Preliminary model of queue interaction at signalised paired intersections. In: Proceedings of the 16th ARRB Conference, Perth, 9–12 November 1992. Congestion Management Proceedings - Conference of the Australian Road Research Board, Australian Road Research Board, Nunawading
214. Schofer AJKFSKJL (1993) Stated preferences for investigating commuters' diversion propensity. Transportation 20(2): 107–127
215. Sheffi Y (1985) Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods, Prentice Hall, Englewood Cliffs
216. Sheffi Y, Powell W (1981) A Comparison of Stochastic and Deterministic Traffic Assignment over Congested Networks. Transp Res 15B:65–88
217. Shen W, Nie Y et al (2006) Path-based System Optimal Dynamic Traffic Assignment Models: Formulations and Solution Methods. In: IEEE Intelligent Transportation Systems Conference. IEEE, Toronto, pp 1298–1303
218. Sherali HD, Arora N, Hobeika AG (1997) Parameter optimization methods for estimating dynamic origin-destination triptables. Transp Res Part B Methodol 31B(2):141–157
219. Sherali HD, Desai J et al (2006) A discrete optimization approach for locating Automatic Vehicle Identification readers for the provision of roadway travel times. Transp Res Part B 40:857–871
220. Simon H (1957) Models of Man, Social and Rational. Adm Sci Q 2(2)
221. Simon HA (1947) Administrative Behavior. Am Political Sci Rev 41(6)
222. Simon HA (1955) A Behavioral Model of Rational Choice. Q J Econ 69(1):99–118
223. Sivanandan R, Dion F et al (2003) Effect of Variable-Message Signs in Reducing Railroad Crossing Impacts. Transp Res Rec 1844:85–93
224. Smock R (1962) An Iterative Assignment Approach to Capacity-Restraint on Arterial Networks. Bulleton Highw Res Board 347:226–257
225. Srinivasan KK, Mahmassani HS (2000) Modeling inertia and compliance mechanisms in route choice behavior under real-time information. Transp Res Rec 1725:45–53
226. Srinivasan KK, Mahmassani HS (2000) Modeling inertia and compliance mechanisms in route choice behavior under real-time information. Transp Res Rec 1725:45–53
227. Steinberg R, Zangwill WI (1983) The Prevalence of Braess' Paradox. Transp Sci 17:301–318
228. Stewart N (1980) Equilibrium versus System-Optimal Flow: Some Examples. Transp Res 14A:81–84
229. Talaat H, Abdulhai B (2006) Modeling Driver Psychological Deliberation During Dynamic Route Selection Processes. In: 2006 IEEE Intelligent Transportation Systems Conference, Toronto, pp 695–700
230. Tarko A, Rouphail N et al (1993) Overflow delay at a signalized intersection approach influenced by an upstream signal. An analytical investigation. Transp Res Rec 1398:82–89
231. Van Aerde M (1985) Modelling of Traffic Flows, Assignment and Queueing in Integrated Freeway/Traffic Signal Networks. In: Civil Engineering. Ph D, University of Waterloo, Waterloo
232. Van Aerde M, Rakha H (1989) Development and Potential of System Optimized Route Guidance Strategies. In: IEEE Ve-

hicle Navigation and Information Systems Conference. IEEE, Toronto, pp 304–309
233. Van Aerde M, Rakha H (2007) INTEGRATION © Release 2.30 for Windows: User's Guide - vol I: Fundamental Model Features. M Van Aerde & Assoc, Ltd, Blacksburg
234. Van Aerde M, Rakha H (2007) INTEGRATION © Release 2.30 for Windows: User's Guide - vol II: Advanced Model Features. M Van Aerde & Assoc, Ltd, Blacksburg
235. Van Aerde M, Rakha H et al (2003) Estimation of Origin-Destination Matrices: Relationship between Practical and Theoretical Considerations. Transp Res Rec 1831:122–130
236. Van Aerde M, Yagar S (1988) Dynamic Integrated Freeway/Traffic Signal Networks: A Routeing-Based Modelling Approach. Transp Res 22A(6):445–453
237. Van Aerde M, Yagar S (1988) Dynamic Integrated Freeway/Traffic Signal Networks: Problems and Proposed Solutions. Transp Res 22A(6):435–443
238. Van Aerde, M, Hellinga BR et al (1993) QUEENSOD: A Method for Estimating Time Varying Origin-Destination Demands For Freeway Corridors/Networks. In: 72nd Annual Meeting of the Transportation Research Board, Washington DC, 1993
239. Van Der Zijpp NJ, De Romph E (1997) A dynamic traffic forecasting application on the Amsterdam beltway. Int J Forecast 13:87–103
240. Van Vliet D (1976) Road Assignment. Transp Res 10:137–157
241. Van Vliet D (1982) SATURN - A Modern Assignment Model. Traffic Eng Control 12:578–581
242. Van Zuylen JH, Willumsen LG (1980) The most likely trip matrix estimated from traffic counts. Transp Res 14B:281–293
243. Walker N, Fain WB et al (1997) Aging and Decision Making: Driving-Related Problem Solving. J Hum Factors Ergon Soc 39(3):438–444(7)
244. Waller ST (2000) Optimization and Control of Stochastic Dynamic Transportation Systems: Formulations, Solution Methodologies, and Computational Experience. Ph D, Northwestern University, Evanston
245. Waller ST, Ziliaskopoulos AK (2006) A chance-constrained based stochastic dynamic traffic assignment model: Analysis, formulation and solution algorithms. Transp Res Part C Emerg Technol 14(6):418–427
246. Wardrop J (1952) Some Theoretical Aspects of Road Traffic Research. Institute of Civil Engineers, pp 325–362
247. Webster F (1958) Traffic Signal Settings. HMsSO Road Research Laboratory, London
248. Webster FV, Cobbe BM (1966) Traffic Signals. HMsSO Road Research Laboratory, London
249. Wie BW (1991) Dynamic Analysis Of User-Optimized Network Flows With Elastic Travel Demand. Transp Res Rec 1328: 81–87
250. Willumsen LG (1978) Estimation of an O-D matrix from traffic counts: A review. Institute for Transport Studies, Working paper no 99, Leeds University, Leeds
251. Wilson AG (1970) Entropy in Urban and Regional Modelling. Pion, London
252. Wu J, Chang G-L (1996) Estimation of time-varying origin-destination distributions with dynamic screenline flows. Transp Res Part B Methodol 30B(4):277–290
253. Yagar S (1971) Dynamic Traffic Assignment by Individual Path Minimization and Queueing. Transp Res 5:179–196
254. Yagar S (1974) Dynamic Traffic Assignment by Individual Path Minimization and Queueing. Transp Res 5:179–196
255. Yagar S (1975) CORQ - A Model for Predicting Flows and Queues in a Road Corridor. Transp Res 553:77–87
256. Yagar S (1976) Measures of the Sensitivity and Effectiveness of the CORQ Traffic Model. Transp Res Rec 562:38–48
257. Yang Q, Ben-Akiva ME (2000) Simulation laboratory for evaluating dynamic traffic management systems. Transp Res Rec 1710:122–130
258. Yang T-H (2001) Deployable Stable Traffic Assignment Models for Control in Dynamic Traffic Networks: A Dynamical Systems Approach. Ph D, Purdue University, West Lafayette
259. Zhou X, Mahmassani HS (2006) Dynamic origin-destination demand estimation using automatic vehicle identification data. In: IEEE Transactions on Intell Transp Syst 7(1):105–114
260. Zhou Y, Sachse T (1997) A few practical problems on the application of OD-estimation in motorway networks. TOP 5(1):61–80
261. Ziliaskopoulos A, Wardell W (2000) Intermodal optimum path algorithm for multimodal networks with dynamic arc travel times and switching delays. Eur J Oper Res 125(3):486–502
262. Ziliaskopoulos AK (2000) A linear programming model for the single destination system optimum dynamic traffic assignment problem. Transp Sci 34(1):37–49
263. Ziliaskopoulos AK, Waller ST (2000) An Internet-based geographic information system that integrates data, models and users for transportation applications. Transp Res Part C EmergTechnol 8C:1–6

Traffic Networks, Optimization and Control of Urban

NATHAN H. GARTNER, CHRONIS STAMATIADIS
University of Massachusetts, Lowell, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Traffic Signal Control
The Basic Model – Single Intersection
Arterial Systems – Progression Models
Delay-Based Models – Cyclic Flow Profiles
Demand-Responsive Models
Multi-Level Traffic Control Strategies
Control System Performance – Analysis of Complexity
Future Directions
Bibliography

Glossary

Assignment (traffic) The allocation of traffic volumes to routes.

Arterial A road with two or more intersections.

Bandwidth Time interval in an arterial progressive system during which vehicles can travel unimpeded.

Controller A device which controls the sequence and duration of indications displayed by traffic signals.

Coordination The establishment of a definite timing relationship between adjacent traffic signals.

Cycle time The time required for one complete sequence of signal indications.

Delay (traffic) The time lost by vehicle(s) due to traffic interference or control devices.

Detector (traffic) A device to detect the presence or passage of a vehicle in the roadway.

Flow (traffic) The rate at which vehicles cross a given line on the road; the number of vehicles per unit of time.

Headway The time interval between successive vehicle crossings of a given line on the road.

Intelligent transportation system The addition of information and communications technology to enhance performance of transport infrastructure and of vehicles.

Offset The time difference between the start of green at one intersection and the start of green at a subsequent intersection, or with respect to a system time base.

Phase The time interval of a cycle time allocated to any combination of movements receiving the right-of-way simultaneously.

Performance index/disutility index A numerical quantity used to measure the performance of a traffic in a network in an optimization or a simulation model; also objective function.

Platoon A tight group of vehicles traveling along an arterial.

Progression A timing plan enabling a platoon to travel unimpeded along an arterial.

Saturation flow The rate of flow at which vehicles are discharged from a queue stopped at a signal; the maximum rate of flow on a street.

Synchronization A condition under which traffic signals operate with the same cycle time.

Volume (traffic) The number of vehicles crossing a line on the road per hour; traffic flow.

Definition of the Subject

Much of our economic and social life is dependent on communication and transportation systems. Travel in urban networks represents an interaction between the demand for transportation and the supply of transportation means and facilities which includes the vehicles, the road networks, as well as the control systems that govern them. The critical elements within these systems are junctions, or intersections, which are controlled by means of traffic signals. Public authorities, which are responsible for the operation of those systems, develop control policies that have a dominant impact on the quality of travel and the level of service provided by the networks. An understanding of the basic models involved in urban traffic control is essential for the development of optimal operating policies that would lead to the effective performance of urban traffic networks.

Introduction

The critical elements within an urban traffic network are the junctions, or the intersections, many of which are controlled by traffic signals. Therefore, traffic control signals are a key determinant for the efficient operation of urban street networks and an essential element of Intelligent Transportation Systems.

In this chapter we focus on the complex array of urban street networks and their control by means of traffic signals. We start by delineating the basic models involved in traffic flow at single intersections. We then present control models on arterial streets and networks, both progression models and delay-based models. Next we introduce demand-responsive systems and we conclude by outlining a strategy for multi-level control and an analysis of the complexity involved in the performance of the different systems.

Traffic Signal Control

Traffic control signals are defined as "power-operated traffic devices which alternately direct traffic to stop and to proceed." More specifically, traffic signals are used to control the assignment of right-of-way at locations where conflicts exist or where passive devices, such as signs and markings, do not provide the necessary flexibility of control to properly move traffic in a safe and efficient manner.

Traffic control signals are usually described as either *pre-timed* or *traffic actuated*. Each type may be used in either an independent (isolated) or interconnected (system) application. *Pre-timed control* assigns the right of way at an intersection according to a predetermined schedule. The length of the time interval for each signal indication in the cycle is fixed, based on historic traffic patterns. *Actuated control* differs from pre-timed in that signal phases are not of fixed length. Through the use of vehicle detectors, this type of control assigns the right of way on the basis of actual traffic conditions (demands) within given limitations. The full range of actuated control capabilities depends on the type of equipment employed and the operational requirements.

Traffic signal timings are divided into time phases. A signal phase may be defined as that part of the cycle

length allocated to a traffic movement receiving the right of way or to any combination of traffic movements receiving the right of way simultaneously. A traffic movement is a single vehicular movement, a single pedestrian movement, or a combination of vehicular and pedestrian movements. The sum of all traffic phases at an intersection is equal to the cycle length.

The Basic Model – Single Intersection

The material in this section is based on models developed by F.V. Webster at the British Road Research Laboratory [1] for estimating delays to vehicles at fixed-time traffic signals and for computing optimum control settings for such signals. The methods developed in this study can be applied both to fixed-time signals and to vehicle-actuation.

Webster's classical results provide a more definite basis for setting fixed-time signals than any which have so far been published. They are else applicable to those vehicle-actuated signals where the green periods owing to heavy traffic demands are frequently running to maximum, giving in effect fixed-time control. In addition, linked systems of traffic signals often work on a fixed cycle which has been set to meet the requirements of the main intersection of the system and the desired speed of traffic along the road.

Flow Model

The first step in the analysis of traffic control signals is to determine performance, most often measured by the delay caused to traffic. Traffic may be considered as arriving at random provided that the point at which it is observed is some distance from a disturbing factor such as a controlled intersection. Total delay to traffic is computed by the area under the queue-length curve. The average delay is obtained by dividing the total delay by the number of arrivals. This is illustrated in Fig. 4 below.

When the signal is red, arriving traffic queues at the signal stop line. After a green period commences, a certain time elapses while vehicles are accelerating to normal running speed, but after a few seconds the queue discharges at a constant rate, called the saturation flow. This is illustrated in the time-space (or distance) diagram of Fig. 1. If there is still a queue at the end of the green period, some vehicles will make use of the yellow interval to cross the intersection. In these circumstances traffic moves on both green and yellow signals but the discharge rate is less than the saturation flow both at the beginning and at the end of the right-of-way period, as shown in Fig. 2. The green and yellow together $(k + a)$ may be replaced by an 'effective' green (g) and a 'lost' time (l), such that the product of the

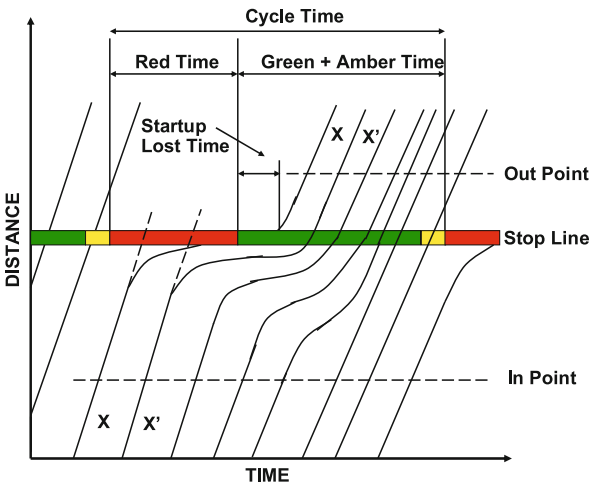

Traffic Networks, Optimization and Control of Urban, Figure 1
Distance/time diagram for flow through intersection

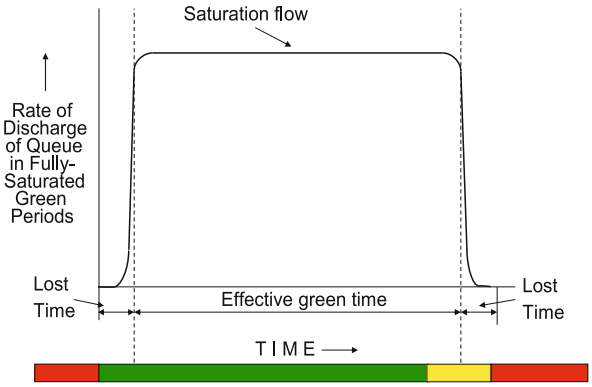

Traffic Networks, Optimization and Control of Urban, Figure 2
Variation of discharge rate of queue with time in a fully-saturated green period

effective green and the saturation flow is equal to the correct number of vehicles (say, b) discharged from the queue on the average in a saturated green period (i. e. a green period during which the queue never clears).

Thus, $k + a = g + 1$ and $b = g \cdot s$, where s is the saturation flow rate.

In the basic model presented in this section the saturation flow is assumed to be constant. The signal sequence on any approach is thus reduced to an 'effective' green period and a 'red' period which comprises all the times when traffic cannot run, i. e. red periods plus 'lost' time. The delay to traffic on a single approach to an intersection controlled by fixed-time traffic signals is computed over a range of values of green times, cycle times, traffic flows and saturation flows covering most practical conditions.

Signalized Intersections – Fluid Traffic Model

There are several models that may be used in the investigation of queues and delays at signalized intersections. In this section a continuum (i. e., deterministic) or fluid model is first considered for which basic measures of queue length and delay are developed [2]. In the next section the stochastic nature of traffic is taken into consideration. In estimating delay at intersections, traffic flow is considered to consist of identical passenger car units (pcu's). A truck, for example, may be considered as 1.5 or 2 equivalent pcu's and a turning vehicle may also be assigned some value depending on the type of maneuver that is made.

The following notation is used:

$c =$ the cycle time (sec);
$g =$ the effective green time (sec);
$r =$ the effective red time (sec);
$q =$ the average arrival rate of traffic on the approach (pcu's/sec);
$s =$ the saturation flow on the approach (pcu's/sec);
$d =$ the average delay to a pcu's on the approach (sec);
$Q_0 =$ the overflow (pcu's);
$\lambda = g/c$ (the proportion of the cycle that is effectively green);
$y = q/s$ (the ratio of average arrival rate to saturation flow); and
$x = q \cdot c/g \cdot s$ (the ratio of average number of arrivals/cycle to the maximum number of departures/cycle).

Thus, $r + g = c$ and $\lambda \cdot x = y$. The ratio x is called the degree of saturation of the approach and y is called the flow ratio of the approach.

The effective green time is the portion of the cycle time during which pcu's are assumed to pass the signal at a constant rate s, provided vehicles are waiting on the approach. Greenshields [3] observed that the times to cross the signal stop line are 3.8, 3.1, 2.7, 2.4, 2.2, 2.1, 2.1 seconds for the consecutive queued vehicles. The cumulative time for a queue of n stopped vehicles to pass a signal can then be given by:

$$\textit{Cumulative time} = 14.2 + 2.1(n-5)\,\text{sec} \quad \text{for } n \geq 5\,.$$

Had all the vehicles departed at the saturation rate $s = 1/2.1$ veh/sec $= 1714$ veh/hr, the first five vehicles would have required 10.5 sec; that is, the effective green is the signal green time less 3.7 sec. In most studies it is also assumed that a waiting queue of vehicles will take advantage of the yellow clearance interval, although the effective green time may be adjusted to reflect particular operating conditions.

Delay can be calculated by illustrating the meaning of arrival time and departure time for a pcu on an approach. They are demonstrated by reference to Fig. 3, in which distance-time curves are plotted for each of four vehicles. AB represents the passage of an un-delayed vehicle, where the line PQ represents the stopline at which the first vehicle waits when there is a queue. CDEF represents the trajectory of the first vehicle that is delayed by a signal. The straight portions of CD and EF are parallel to AB and projected to meet PQ and X and Y so that the length XY is the delay to the first vehicle. Similarly, $X'Y'$ and $X''Y''$ represent the delays for the next tow vehicles. XX' and $X'X''$ represent the arrival headways.

Continuum Model for a Pre-timed Signal

Representation of a continuum model of traffic flow at a signalized intersection is given in Fig. 4. The vertical axis represents the number of vehicle arrivals of the stop line and the horizontal axis the time t. The figure illustrates the behavior when the capacity of the green interval exceeds the arrival during the green + red time i. e., an under-saturated period. The top figure illustrates the queue length as a function of time (t is the queue clearance time, measured from the start of green). The bottom figure illustrates the cumulative arrival and departure functions. In Fig. 4 the vertical distance ca represents the number of vehicles that have accumulated since the signal entered the red phase. The horizontal distance ab represents the total time from arrival to departure for any given vehicle. The shaded areas, in each figure, represent the total vehicle delay.

The following measures of queue behavior are developed:

1. Time after start of green that queue is dissipated (t_0);
2. Proportion of cycle with queue (P_q);
3. Proportion of vehicles stopped (P_s) ;
4. Maximum number of vehicles in queue (Q_m);
5. Average number of vehicles in queue ($\overline{Q}$);
6. Total vehicle-time of delay per cycle (D);
7. Average individual vehicle delay (d);
8. Maximum individual vehicular delay (d_m);

The formulas for these measures are developed from simple geometric relationships:

1. For any given cycle it is evident that at time t_0 after the start of green, the total number of arrivals must equal the number discharged: $q \cdot (r + t_0) = s \cdot t_0$. Letting $y = q/s$ we obtain:

$$t_0 = y \cdot r/(1-y)\,.$$

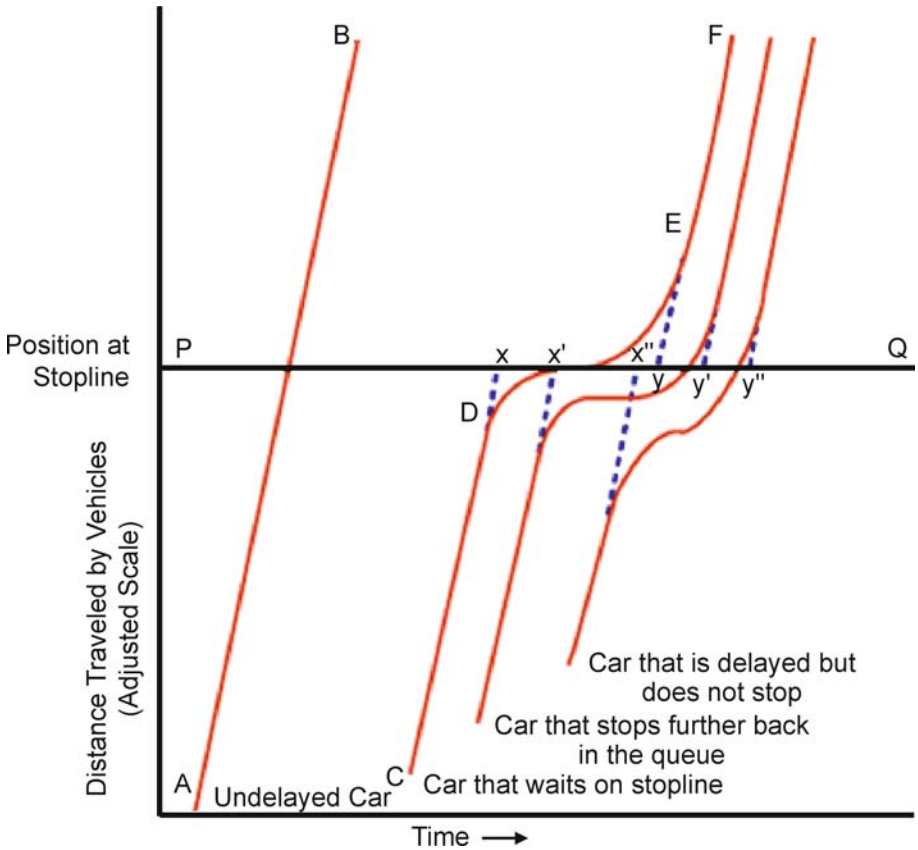

Traffic Networks, Optimization and Control of Urban, Figure 3
Vehicle arrival-departure times and delay of the stop-line

2. Proportion of vehicles stopped is equal to queue time/cycle length

$$P_q = (r + t_0)/c\,.$$

3. Proportion of vehicles stopped is equal to vehicles stopped/total vehicles per cycle

$$P_s \cdot q \cdot (r + t_0)/q \cdot (r + g) = t_0/(y \cdot c)\,.$$

4. The maximum number of vehicles in queue can be seen by inspection to be the height of the triangle at $t = r$, i. e., at the end of the red period.

$$Q_m = q \cdot r\,.$$

5. The average number of vehicles in the queue, over the total length of cycle c is

$$\overline{Q} = \frac{(qr/2)r + (qr/2)t_0 + o(g - t_0)}{r + t_0 + g - t_0}\,.$$

which yields: $\overline{Q} = (r + t_0) \cdot q \cdot r/(2 \cdot c)$.

6. The total vehicle-time of delay during the cycle is given by the area of the triangle

$$D = (q \cdot r/2)(r + t_0) = (q \cdot r/2)(r/(1 - y)) = q \cdot r^2/(2(1 - y))\,.$$

7. The average individual delay is given by dividing the total delay by the number of vehicles, or

$$d = \left[\frac{qr^2}{2(1 - y)}\right]\frac{1}{qc} = \frac{r^2}{2c(1 - y)}\,.$$

8. The maximum individual vehicular delay will be seen from Fig. 4 to be

$$d_m = r\,.$$

If the possible departures during the cycle sg (i. e., the capacity flow) are less than the arrivals qc, the queue grows with each successive cycle and the foregoing formulas are no longer applicable. We have, then, over-saturation.

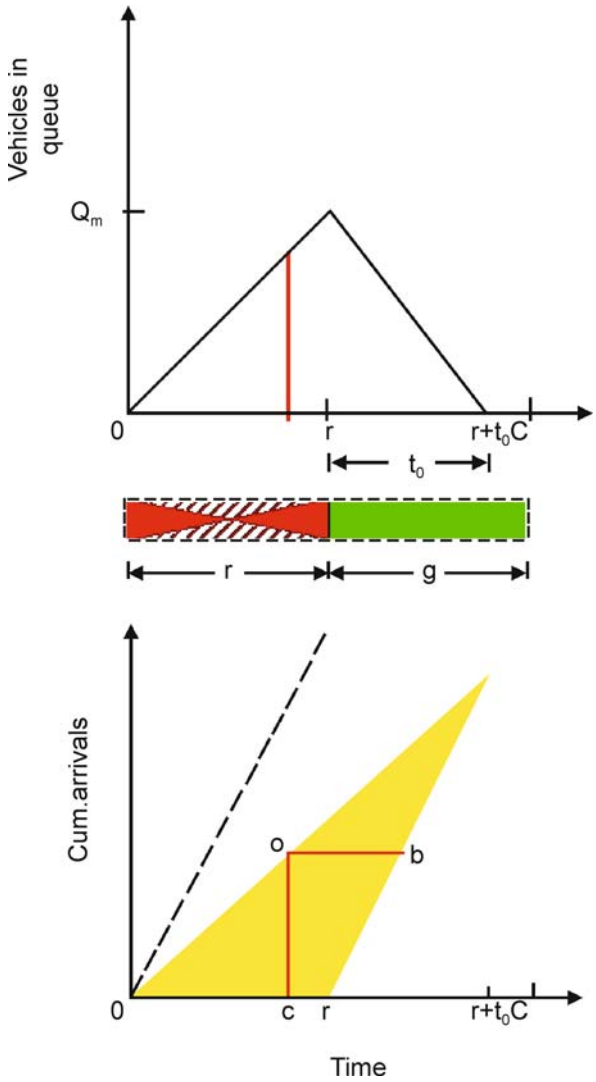

Traffic Networks, Optimization and Control of Urban, Figure 4
Cumulative arrival-departure diagrams and queuing at a signal (continuum model)

Average Delay per Vehicle

Webster found that the results of his computations using a stochastic simulation model could be expressed to a close approximation by the equation

$$d = \frac{c(1-\lambda)^2}{2(1-\lambda x)} + \frac{x^2}{2q(1-x)} - 0.65\left(\frac{c}{q^2}\right)^{\frac{1}{3}} x^{(2+5\lambda)} \quad (1)$$

where:

$d =$ average delay per vehicle on the particular arm of the intersection

$c =$ cycle time

$\lambda =$ proportion of the cycle which is effectively green for the phase under consideration (i. e. g/c)

$q =$ flow

$s =$ saturation flow

$x =$ the degree of saturation – this is the ratio of the actual flow to the maximum flow which can be passed through the intersection from this arm, and is given by: $x = q/\lambda s$.

An estimate of the average delay per vehicle is required to evaluate the performance of traffic control signals. To enable the expected delay to be estimated more easily, Eq. (1) can be rewritten as

$$d = cA + \frac{B}{q} - C \quad (2)$$

where $A = (1-\lambda)^2/2(1-\lambda x)$, $B = x^2/2(1-x)$ and C is a correction term.

The first two terms in Eq. (1) have a theoretical underpinning but the last term is purely empirical. The first term is the expression for the delay when traffic can be considered to be arriving at a uniform rate. It is identical to the formula in Item No. 7 above. Although the agreement between computed delays and those derived from this term is fairly good at low flows, as shown by Fig. 5, it is not so at higher values where the computed delays, owing to the random nature of the arrivals, are far in excess of values calculated from this term only. The second term makes some allowance for the random nature of the arrivals. It is an expression for the delay experienced by vehicles arriving randomly in time at a 'bottleneck', queuing up, and leaving at constant intervals. In queuing terminology it is the average delay of an M/D/1 queue. The addition of the second term improves the correspondence with Eq. (1), but overestimates it. To avoid the need to calculate the correction factor C, a suitable approximation can be obtained by one of the following:

$$d = \left[cA + \frac{B}{q}\right]\frac{100-P}{100} \quad (3)$$

where (a) $P = 10\%$, or (b) $P = 15x$, since the value of the correction term C is generally in the range 5 to 15 percent of d.

The 'lost time' illustrated in Fig. 2 plays an important role in determining the efficiency of the traffic signal operation. Information available on lost time suggests a value of about 4.5 seconds per phase plus any all-red periods. This is composed of 2.5 sec start-up lost time plus half the typical 4 sec yellow time. Since lost time depends on gradients, type of traffic, etc. its value will vary from site to site and even at the same site will probably vary with time of day.

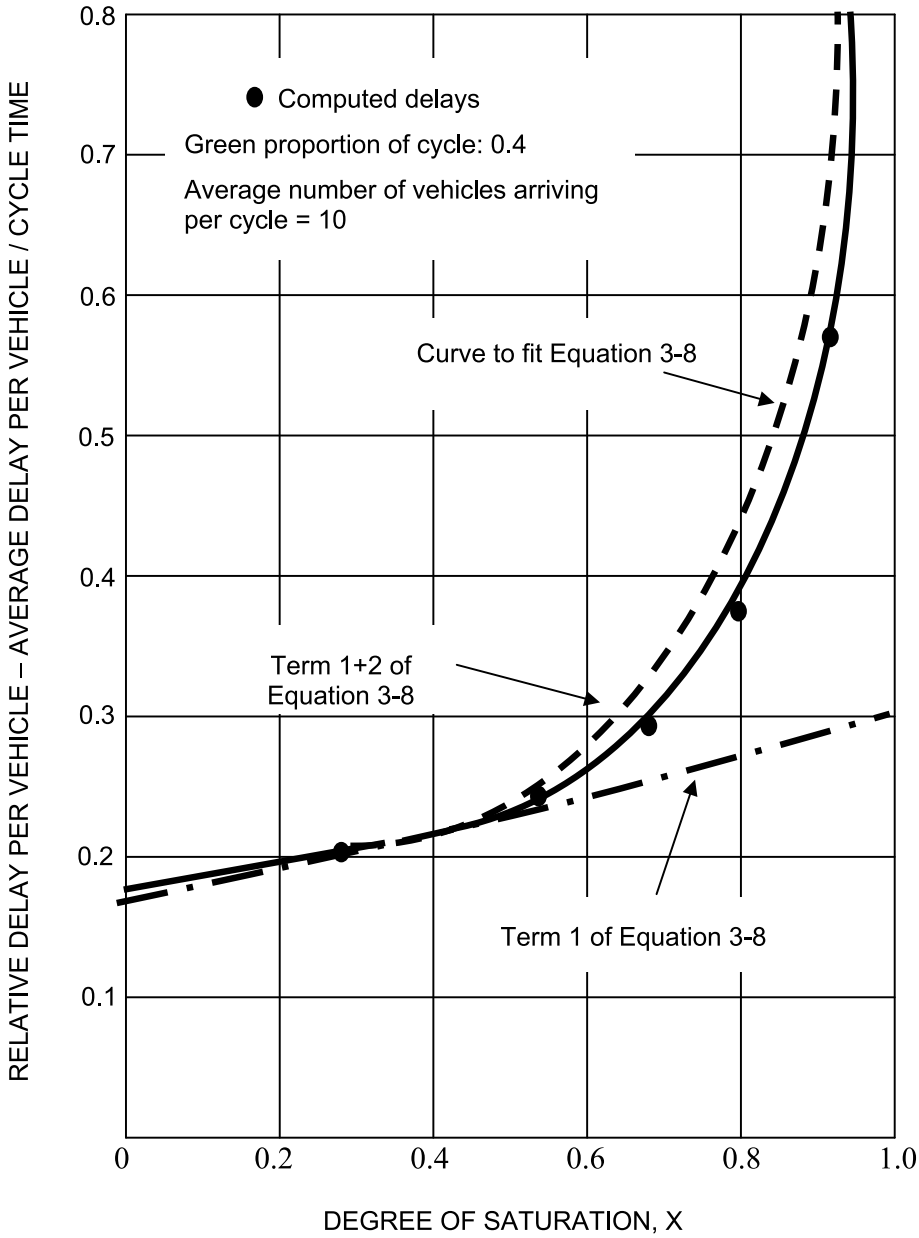

Traffic Networks, Optimization and Control of Urban, Figure 5
Typical fixed-time delay curve

Optimum Settings of Fixed-Time Signals

In general, all approaches belonging to the same phase will have the same green period even though the traffic requirements of the approaches may be different. To simplify calculations, and with negligible loss of accuracy, each phase can be represented by one approach only – the one with the highest ratio of flow to saturation flow. Let this ratio be denoted by the symbol y. In deriving the optimum green times and cycle time the empirical correction term of the delay equation is neglected since calculations show that its variation with respect to these quantities is slight.

Green Times Least delay to traffic is obtained when the green periods of the phases are in proportion to the corresponding ratios of flow to saturation flow, assuming this

ratio to be the same for all arms of the same phase. This division of the cycle time makes the capacity of the phases proportional to the average flows of the phases. That this is approximately the best division of the cycle time is shown by the examples given below.

The total delay experienced by vehicles at an intersection was calculated using the approximate form of the delay Eq. (2). The lost time was assumed to be 10 seconds per cycle and delays were calculated for a variety of cycle times and ratios of the effective green times. A typical result is shown in Fig. 6, where the ratio of the y values is 2.0. It can be seen that the best ratios of the effective green times is between 1.88 and 2.17 over a range of cycle times of 35 to 80 seconds.

Cycle Time In deriving an expression for the optimum cycle time it is assumed that the effective green times of the phases are in the ratio of their respective y values. Equation (3) representing the delay to traffic on one arm of an intersection is modified to cover the general case of an intersection with n phases, and the modified equation is differentiated with respect to cycle time to determine the value of cycle time which gives the least delay of all traffic using the intersection.

The derivation of the optimum cycle-time formula is given in [1]. This formula is too complicated for most purposes and a simple approximation is therefore derived as follows:

$$c_0 = \frac{1.5L + 5}{1 - Y} \text{ sec} \tag{4}$$

where Y is the sum of the y values and refers to the intersection as a whole and L is the total lost time per cycle in seconds. The lost time can be expressed by: $L = nl + R$.

Where:

n is the number of phases

l is the average lost time per phase (excluding any all-red periods)

R is the time during each cycle when all signals display red (including red-with-amber) simultaneously.

The possible aspect scheduling for a 2-phase signal is shown in Fig. 7.

Several examples of hypothetical intersections (including some fairly extreme cases) have been studied to show the effect of changes in cycle time on the delay. For a symmetrical intersection, values of delay have been computed by Eq. (1) and are shown in Fig. 8 plotted against the cycle length for several values of the total flow entering the intersection. The cycle time giving the least delay in each case is approximately twice the least possible cycle which will just allow the traffic to pass through (although with long delays). The latter is called the minimum cycle, and is the vertical asymptote to the delay/cycle-time curve. When traffic is of a truly random character the minimum cycle is associated with infinite delay. For uniform flow it is the cycle such that all traffic arriving during the cycle can just be discharged during the green. We can approximate the optimum cycle as follows:

$$c_{\text{opt}} \cong 2c_{\text{min}} = \frac{2L}{1 - Y}$$

From graphs such as Fig. 8, it was found that in most practical cases the delay for cycle times within the range ¾ to 1½ times the optimum value is never more than 10 to 20 per cent greater than that given by the optimum cycle. This fact can be used in deducing a compromise cycle time when the level of flow varies considerably throughout the day. It would be better either to change the cycle time to take account of this, or, as is more common, to use vehicle-actuated signals. However, for a single setting of fixed-time signals the simple approximate method outlined below may be used.

1. Calculate the optimum cycle for each hour of the day when the traffic flow is medium of heavy, e. g. between the hours of 8 a.m. and 7 p.m. and average over the day.
2. Evaluate three-quarters of the optimum cycle calculated for the heaviest peak hour.
3. Select whichever is greater for the cycle time.

It is suggested as a reasonable procedure that the division of the available green time $(c_0 - L)$ should be in proportion to the average y values for peak periods only, i. e.:

$$\frac{g_1}{g_2} = \frac{(y_1)_{\text{PEAK}}}{(y_2)_{\text{PEAK}}}$$

where $(y_1)_{\text{PEAK}}$ is the average y value during the peak periods for phase 1 and $(y_2)_{\text{PEAK}}$ that for phase 2.

Arterial Systems – Progression Models

An arterial is a road consisting of two or more intersections. A *signal system* is defined as having two or more individual signal installations which are linked together for coordination purposes. To obtain coordination all signals must operate with the same (common) cycle length, i. e., the signals must be synchronized. In rare instances, some intersections within the system may operate at double or one-half the cycle length of the system in which case there is half- (or double-) synchronization. It is desirable that the arterial for which coordination is being provided have a green plus yellow interval equivalent to at least 50% of

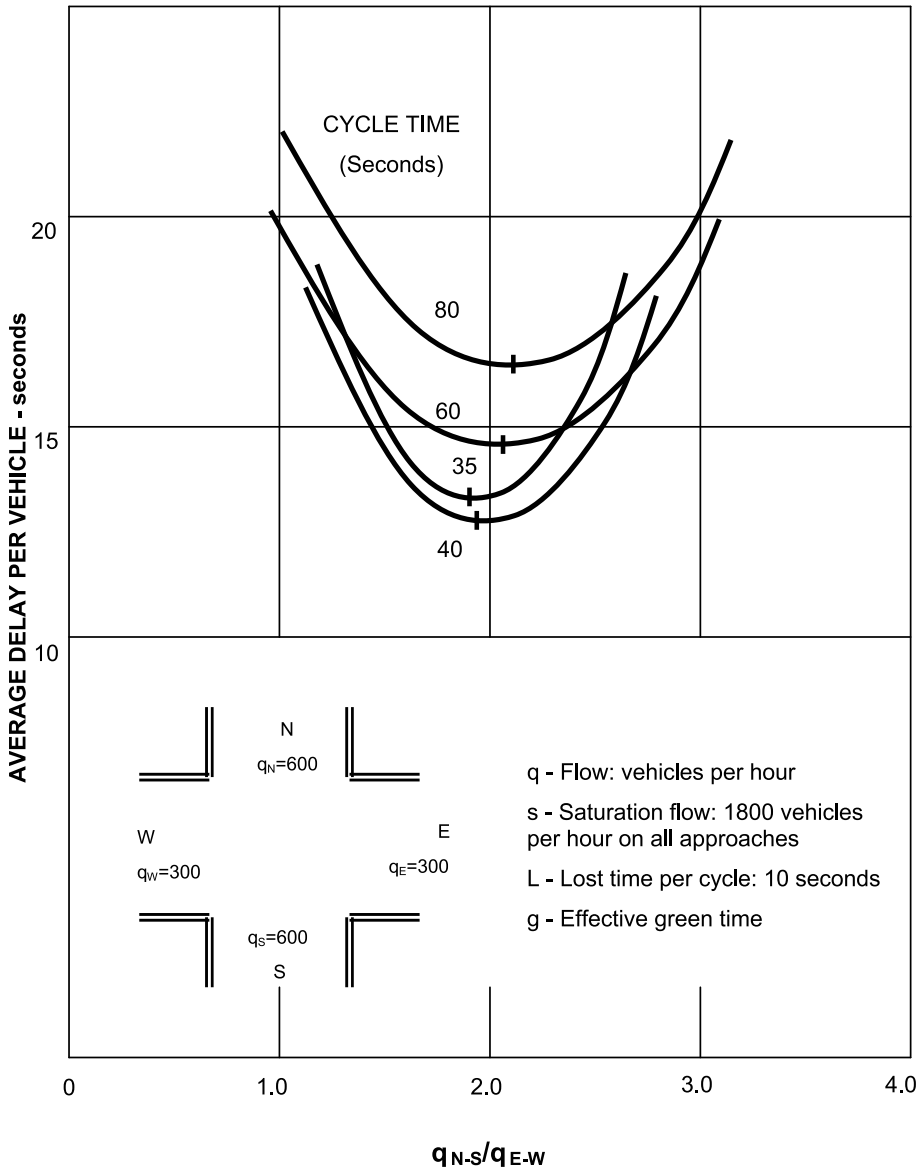

Traffic Networks, Optimization and Control of Urban, Figure 6
Effect on delay of the variation of the ratio of green periods

the cycle length. Two intersecting systems form an *open network* whenever they have only one intersection in common. A *closed network* is formed whenever there are three or more common intersections. A grid network of $m \times n$ arterials is a closed network with $m \times n$ common intersections. Closed networks are characterized by having *closed loops*.

The predominant flow of traffic in urban street networks is along arterial streets. Optimal control of the traffic signals along these arteries is essential for the effective operation of the grid network. Coordination of traffic lights along arterial streets provides numerous advantages:

1. A higher level of traffic service is provided in terms of higher overall speed and reduced number of stops.
2. Traffic flows more smoothly, often with an improvement in capacity due to reduced headways.
3. Vehicle speeds are more uniform because there is no incentive to travel at excessively high speeds to reach a signalized intersection within a green interval that is

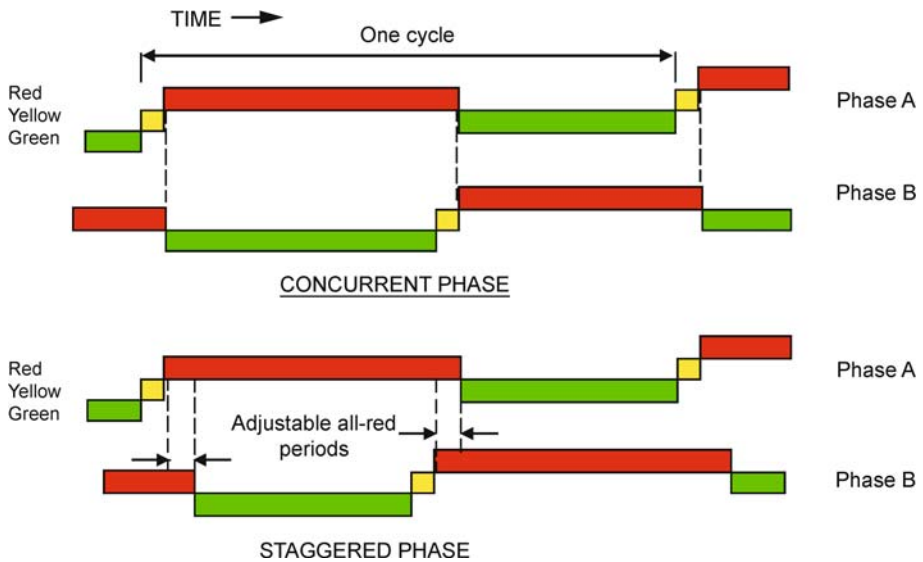

Traffic Networks, Optimization and Control of Urban, Figure 7
Possible phase aspect scheduling at a two-phase intersection

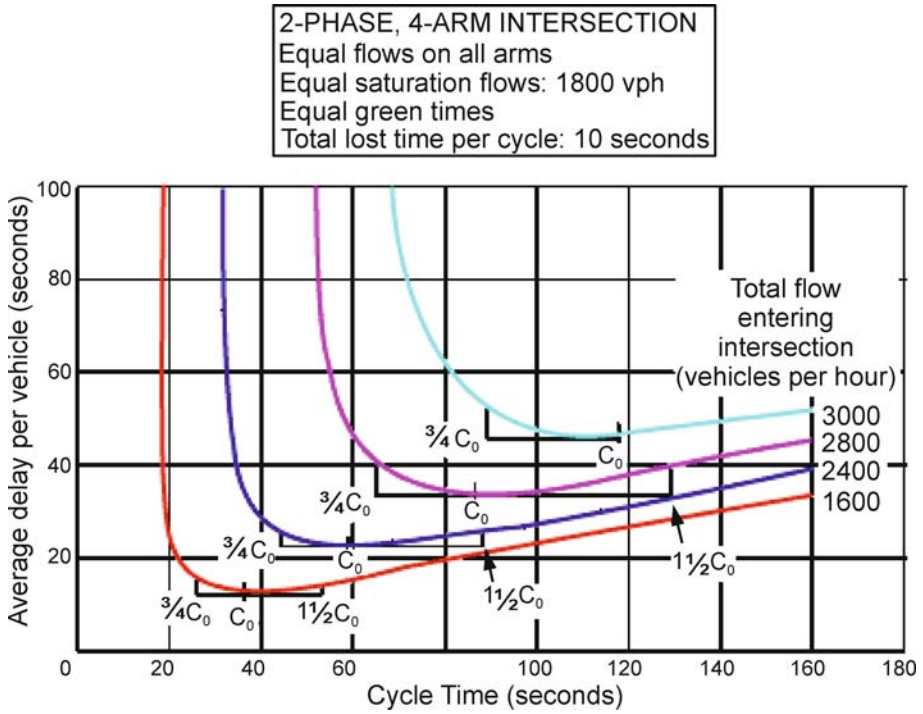

Traffic Networks, Optimization and Control of Urban, Figure 8
Effect on delay of variation of the cycle length

not in step. Also, the slow driver is encouraged to speed up in order to avoid having to stop for a red light.

4. There are fewer accidents because the platoons of vehicles arrive at each signal when it is green, thereby reducing the possibility of red signal violations and rear-end collisions.
5. Greater obedience to the signal commands is obtained from both motorists and pedestrians. The motorist tries

to keep within the green interval and the pedestrian stays at the curb because the vehicles are more tightly spaced.

6. Through traffic tends to stay on the arterial streets instead of diverting onto parallel minor streets in search of alternative routes.

There are two basic approaches to computing arterial signal timings: (a) maximize the bandwidth of the progression, or (b) minimize overall delays and stops.

Arterial progression schemes are widely used for signal coordination. The conceptual basis for progression design is that traffic signals tend to group vehicles into a "platoon" with more uniform headways than would otherwise occur. The platooning effect is most prevalent on major streets with signalized intersections at frequent intervals. It is desirable, in these circumstances, to encourage platooning so that continuous movement (or progression) of vehicle platoons through successive traffic lights can be maintained. Signal timings, in this case, are designed to maximize the width of continuous green bands in both directions along the artery at the expected or recommended speed of travel. Although maximizing the bandwidth is only a surrogate performance measure, it provides significant benefits in increased travel speeds and smoothness of traffic flow and is the preferred approach in many countries. Such systems operate best when main-street flow is predominantly through traffic and when the number of vehicles turning onto the main street is small compared with the through volume. The second approach uses direct performance models to minimize delay, stops or other measures of disutility. Due to the complexity of these models and the high degree of non-linearity involved, they are generally not amenable to solution by mathematical programming methods and heuristic approaches have to be used. As a consequence, the resulting solutions are non-optimal. In this section we concentrate on models of the first kind.

The first optimization model used for arterial bandwidth maximization was developed by Morgan and Little using a combinatorial optimization scheme [4]. A more advanced model, using mixed-integer linear programming, was later formulated by Little [5]. Advances in optimization techniques and computational capabilities have steadily increased the sophistication and versatility of the traffic signal optimization models. We introduce, first, single arterial bandwidth optimization models using both uniform and variable-width bandwidth progressions. These models are then extended to the optimization of arterial grid networks. Heuristic decomposition approaches are developed to speed-up the computation and to make it suitable for on-line implementation as well as for combination with traffic assignment and routing models.

The models presented below use mixed-integer linear programming (MILP) for optimization. They are presented with progressively increasing complexity. MILP-1 is the basic uniform-width bandwidth maximization model for a single arterial. MILP-2 introduces a multi-band, multi-weight approach that generates variable-width two-way progression bands along the arterial. MILP-3 extends the previous models to a grid network of intersecting arterials.

The Basic Bandwidth Maximization Problem

The geometric relations for the uniform bandwidth model are shown in Fig. 9. Consider an arterial having n signalized intersections. Let S_i denote the ith signal on the arterial and L_i denote the ith link (between signals i and $i+1$) of the arterial, with $i = 1, \ldots, n_j$. All time variables are defined in units of the cycle time. The following variables are defined:

$C =$ cycle time (sec);

$b(\bar{b}) =$ outbound (inbound) bandwidth;

$r_i(\bar{r}_i) =$ outbound (inbound) red time at S_i;

$w_i(\bar{w}_i) =$ interference variables, time from right (left) side of red at S_i to left (right) side of outbound (inbound) green band;

$t_{hi}(\bar{t}_{hi}) =$ travel time from S_i to S_h in the outbound (inbound) direction;

$d_{hi}(\bar{d}_{hi}) =$ distance from S_i to S_h in the outbound (inbound) direction;

$\phi_{hi}(\bar{\phi}_{hi}) =$ internode offsets, time from the center of the outbound (inbound) red at S_h to the center of the outbound (inbound) red at S_i;

$\Delta_i =$ directional node phase shift, time from center of $\bar{r}_i$ to nearest center of r_i;

$\tau_i(\bar{\tau}_i) =$ queue clearance time for advancement of outbound (inbound) bandwidth at S_i to clear turning-in traffic before arrival of main-street platoon;

$V_i(\bar{V}_i) =$ outbound (inbound) progression speed on link L_i (m/sec).

In the case of uniform bandwidths the objective function has the following form:

$$\text{Maximize } b + k \cdot \bar{b} \tag{5}$$

where k is the target ratio of inbound to outbound bandwidth taken as the ratio of total inbound to total outbound volumes along the arterial.

The first set of constraints that we introduce are the *directional interference constraints*. These constraints make sure that the progression bands use only the available

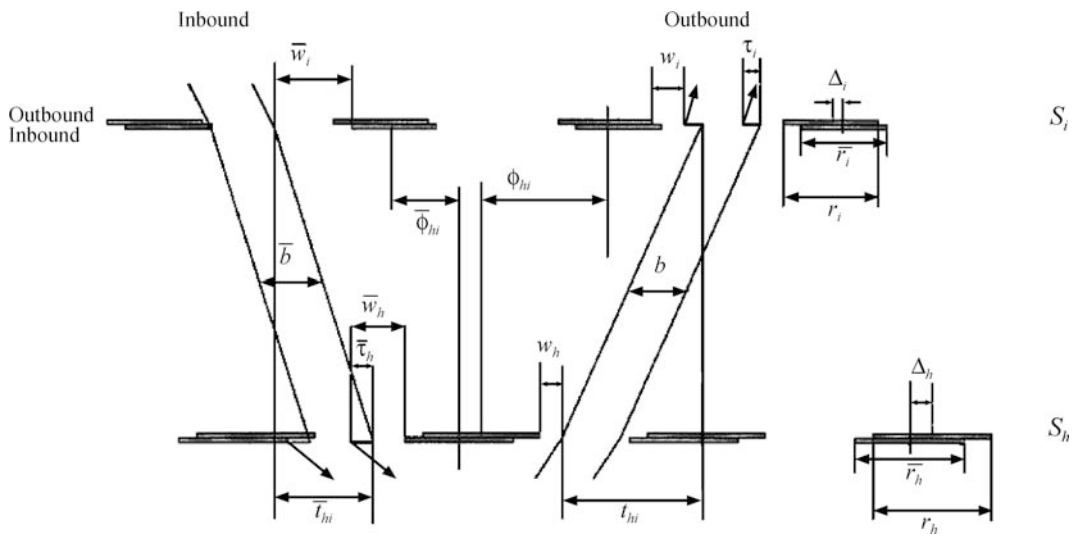

Traffic Networks, Optimization and Control of Urban, Figure 9
Time space diagram for MILP-1

green time and do not infringe upon any of the red times. From Fig. 9 we see that:

$$w_i + b \leq 1 - r_i \tag{6a}$$

$$\bar{w}_i + \bar{b} \leq 1 - \bar{r}_i \,. \tag{6b}$$

Next, we calculate the *arterial-loop integer constraint*. This constraint is due to the fact that the signals of the arterial are synchronized, i. e., they operate in a common cycle time. If we start at the center of the outbound red at S_j and proceed along a loop consisting of the following points:

- Center of outbound red at S_i-
- Center of inbound red at S_i-
- Center of inbound red at S_h-
- Center of outbound red at S_h,

we must end up at a point that is removed an integral number of cycle times from the point of departure. Summing algebraically the appropriate internode offsets and the directional node phase shift along the loop, we obtain

$$\phi_{hi} + \bar{\phi}_{hi} + \Delta_h - \Delta_i = \nu_{hi} \tag{7}$$

where ν_{ij} is the corresponding loop integer variable.

From Fig. 9, we observe the following identities:

$$\phi_{hi} + \frac{r_i}{2} + w_i + \tau_i = \frac{r_h}{2} + w_h + t_{hi} \tag{8a}$$

and:

$$\bar{\phi}_{hi} + \frac{\bar{r}_i}{2} + \bar{w}_i + \bar{\tau}_i = \frac{\bar{r}_h}{2} + \bar{w}_h + \bar{t}_{hi} \,. \tag{8b}$$

Substituting (8b) into (8a) to eliminate ϕ and $\bar{\phi}$ gives:

$$t_{hi} + \bar{t}_{hi} + \frac{r_h + \bar{r}_h}{2} + (w_h + \bar{w}_h) - \frac{r_i + \bar{r}_i}{2} - (w_i + \bar{w}_i) - (\tau_i + \bar{\tau}_h) + \Delta_h - \Delta_i = \nu_{hi} \,. \tag{9}$$

So far we have assumed that S_i follows S_h in the outbound direction, but this restriction is not necessary. Let x represent any of the variables t, $\bar{t}$, ν, ϕ, $\bar{\phi}$ then the following relationship is satisfied:

$$x_{hj} = x_{hi} + x_{ij}$$

whence, by setting $h = j$, we obtain $x_{hi} = -x_{ih}$. Thus Eqs. (8a) and (8b) hold for arbitrary S_h and S_i.

To simplify notation we number the signals sequentially from 1 to n in the outbound direction, and we define $x_i = x_{i,i+1}$. We can now rewrite Eq. (9) as follows:

$$t_i + \bar{t}_i + \frac{r_i + \bar{r}_i}{2} + (w_i + \bar{w}_i) - \frac{r_{i+1} + \bar{r}_{i+1}}{2} - (w_{i+1} + \bar{w}_{i+1}) - (\tau_{i+1} + \bar{\tau}_i) + \Delta_i - \Delta_{i+1} = \nu_i \,. \tag{10}$$

The common signal cycle time C (seconds) and the link specific progression speed $V_i(\bar{V}_i)$ (m/sec) are optimizable variables as well. This introduces considerable flexibility in the calculation of the best arterial progression. Each of these variables is constrained by upper and lower limits. In addition, changes in speed from one link to the next can also be limited. Let the limits be as follows:

$C_1, C_2 =$ lower and upper bounds on cycle length;
$(e_i, f_i), (\bar{e}_i, \bar{f}_i) =$ lower and upper bounds on outbound (inbound) speed $V_i(\bar{V}_i)$ (m/sec);
$(g_i, h_i), (\bar{g}_i, \bar{h}_i) =$ lower and upper bounds on change in outbound (inbound) speed $V_i(\bar{V}_i)$ (m/sec).

To obtain constraints that are linear in the decision variables we define the inverse of the cycle time, i. e., the signal frequency $z = 1/C$ (cycles/second), such that:

$$1/C_2 \le z \le 1/C_1$$

The progression speeds and the changes in speeds are also expressed in terms of their reciprocals:

$$\frac{1}{f_i} \le \frac{1}{V_i} \le \frac{1}{e_i} \quad \text{and} \quad \frac{1}{\bar{f}_i} \le \frac{1}{\bar{V}_i} \le \frac{1}{\bar{e}_i}$$
$$\frac{1}{h_i} \le \frac{1}{V_{i+1}} - \frac{1}{V_i} \le \frac{1}{g_i} \quad \text{and} \quad \frac{1}{\bar{h}_i} \le \frac{1}{\bar{V}_{i+1}} - \frac{1}{\bar{V}_i} \le \frac{1}{\bar{g}_i}$$

Using the notation $d_i = d_{i,i+1}$, we obtain

$$t_i = \frac{d_i}{V_i} z \quad \text{and} \quad \bar{t}_i = \frac{\bar{d}_i}{\bar{V}_i} z$$

Substituting these expressions above and manipulating the variables, we obtain

$$\frac{d_i}{f_i} z \le t_i \le \frac{d_i}{e_i} z \quad \text{and} \quad \frac{\bar{d}_i}{\bar{f}_i} z \le \bar{t}_i \le \frac{\bar{d}_i}{\bar{e}_i} z$$
$$\frac{d_i}{h_i} z \le \frac{d_i}{d_{i+1}} t_{i+1} - t_i \le \frac{d_i}{g_i} z \quad \text{and}$$
$$\frac{\bar{d}_i}{\bar{h}_i} z \le \frac{\bar{d}_i}{\bar{d}_{i+1}} \bar{t}_{i+1} - \bar{t}_i \le \frac{\bar{d}_i}{\bar{g}_i} z\,.$$

Thus, t_i, $\bar{t}_i$ and z are decision variables which, once known, determine the progression speeds. The complete MILP-1 is given below:

MILP-1. Find $b, \bar{b}, z, w_i, \bar{w}_i, t_i, \bar{t}_i, \nu_i$ **to**

Maximize $b_+ k.\bar{b}$ **subject to**

$$1/C_2 \le z \le 1/C_1$$
$$w_i + b \le 1 - r_i \quad \text{and} \quad \bar{w}_i + \bar{b} \le 1 - \bar{r}_i$$
$$i = 1, \ldots, n$$
$$t_i + \bar{t}_i + \frac{r_i + \bar{r}_i}{2} + (w_i + \bar{w}_i) - \frac{r_{i+1} + \bar{r}_{i+1}}{2}$$
$$- (w_{i+1} + \bar{w}_{i+1}) - (\tau_{i+1} + \bar{\tau}_i) + \Delta_i - \Delta_{i+1} = \nu_i$$
$$i = 1, \ldots, n-1$$
$$\frac{d_i}{f_i} z \le t_i \le \frac{d_i}{e_i} z \quad \text{and} \quad \frac{\bar{d}_i}{\bar{f}_i} z \le \bar{t}_i \le \frac{\bar{d}_i}{\bar{e}_i} z$$
$$i = 1, \ldots, n-1$$

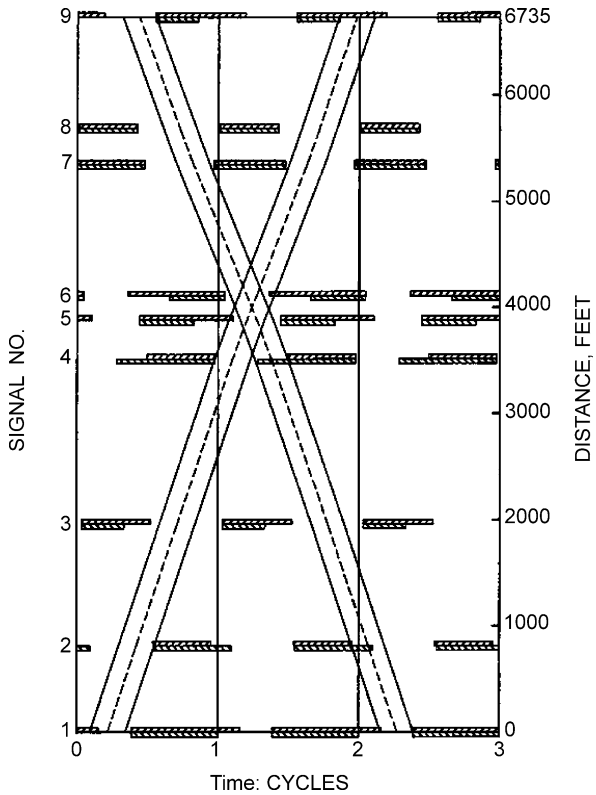

Traffic Networks, Optimization and Control of Urban, Figure 10
Time space diagram for a uniform bandwidth progression scheme generated by MILP-1

$$\frac{d_i}{h_i} z \le \frac{d_i}{d_{i+1}} t_{i+1} - t_i \le \frac{d_i}{g_i} z \quad \text{and}$$
$$\frac{\bar{d}_i}{\bar{h}_i} z \le \frac{\bar{d}_i}{\bar{d}_{i+1}} \bar{t}_{i+1} - \bar{t}_i \le \frac{\bar{d}_i}{\bar{g}_i} z \quad i = 1, \ldots, n-1$$
$$b, \bar{b}, z, w_i, \bar{w}_i, t_i, \bar{t}_i \ge 0 \quad \text{and} \quad \nu_i \text{ integer}\,.$$

In addition to bandwidths and interference variables, this program also calculates optimal cycle length ($1/z$) and progression speeds. To calculate the offsets, green splits need to be given or calculated based on flow/capacity considerations. Phase sequencing can also be optimized. MILP-1 is the model used by the MAXBAND program developed by Little et al. [6]. An example of a uniform bandwidth solution obtained by MILP-1 is shown in Fig. 10. The scheme contains signals with multi-phase sequences.

The Variable Bandwidth Problem

In this section we describe a more versatile multi-band/multi-weight approach for the bandwidth maximization problem. A different bandwidth is defined for each directional road section of the arterial and is individually

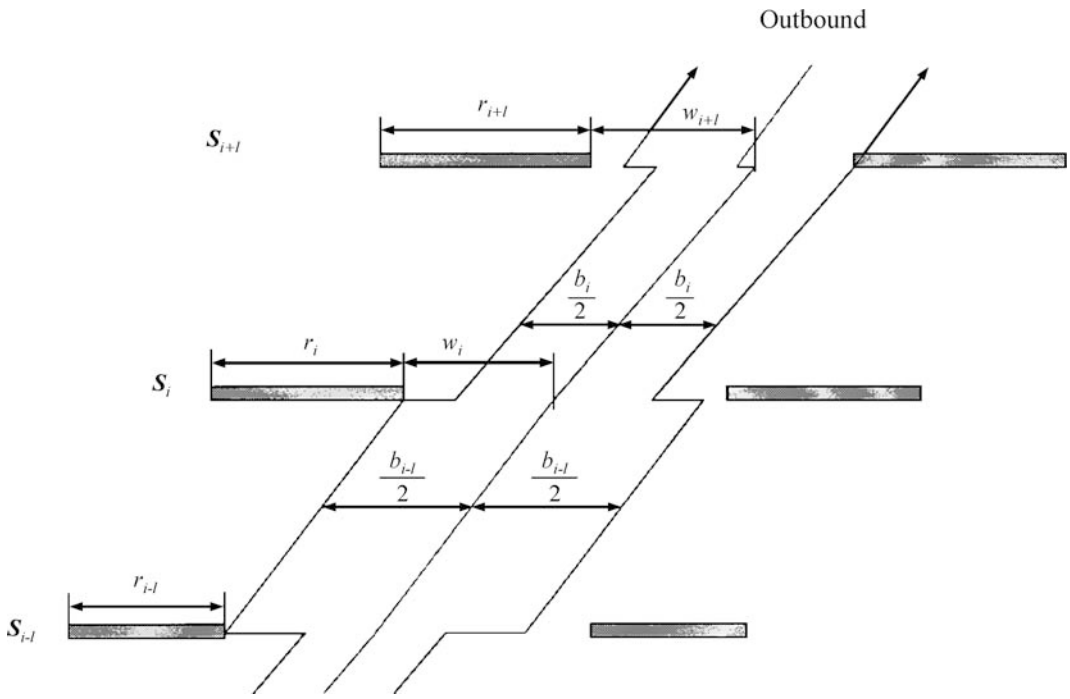

Traffic Networks, Optimization and Control of Urban, Figure 11
Geometric relations for MILP-2

weighted with respect to its contribution to the overall objective function. While the band is continuous, its width can vary and adapt to the prevailing traffic flow on each link. In this way the formulation is sensitive to varying traffic conditions and can tailor the progression scheme to the different possible traffic flow patterns. The user can still choose a uniform bandwidth progression if desired, but this is only one of the many possible options. Referring to the geometry in Fig. 11, we define the following variables:

$b_i(\bar{b}_i) =$ outbound (inbound) bandwidth of link i; there is now a specific band for each link L_i;

$w_i(\bar{w}_i) =$ the time from right (left) side of red at S_i to the centerline of the outbound (inbound) green band; the reference point at each signal has been moved from the edges to the centerline of the band.

The following constraints apply in the outbound directions at signal S_i:

$$w_i + \frac{b_i}{2} \leq 1 - r_i \quad \text{and} \quad w_i + \frac{b_i}{2} \geq 0 \,.$$

The pair of constraints can be combined as follows:

$$\frac{b_i}{2} \leq w_i \leq (1 - r_i) - \frac{b_i}{2} \,.$$

A similar relation must be observed at S_{i+1}, since band b_i must be constrained at both ends:

$$\frac{b_i}{2} \leq w_{i+1} \leq (1 - r_{i+1}) - \frac{b_i}{2} \,.$$

Corresponding relationships exist in the inbound direction (marked by a bar on all variables):

$$\frac{\bar{b}_i}{2} \leq \bar{w}_i \leq (1 - \bar{r}_i) - \frac{\bar{b}_i}{2}$$
$$\frac{\bar{b}_i}{2} \leq \bar{w}_{i+1} \leq (1 - \bar{r}_{i+1}) - \frac{\bar{b}_i}{2} \,.$$

These constraints are referred to as the *directional interference constraints*. The time reference points are redefined to the *centerline* of the bands (or, the *progression* line), rather than the edges. The *loop integer constraint* given by Eq. (10) remains unchanged and so do the travel time and speed-change constraints.

The most important change occurs in the objective function. Since the bands are link-specific, they can be weighted disaggregately to reflect user-defined traffic performance objectives which can be individually weighted for each link of the arterial. The objective function now

has the following form:

$$\text{Maximize} \quad \frac{1}{n-1}\sum_{i=1}^{n} a_i \cdot b_i + \bar{a}_i \cdot \bar{b}_i \tag{11}$$

where a_i and $\bar{a}_i$ are the link specific weighting coefficients for the outbound and inbound directions respectively. There is a multitude of possible options for choosing the weighting coefficients in Eq. (11). Some of the more common expressions are as follows:

$$a_i = \left(\frac{q_i}{s_i}\right)^p \quad \text{and} \quad \bar{a}_i = \left(\frac{\bar{q}_i}{\bar{s}_i}\right)^p$$

where

$q_i(\bar{q}_i) =$ outbound (inbound) directional flow rate on link L_i; either the total or the through volume can be used;

$s_i(\bar{s}_i) =$ saturation flow rate outbound (inbound) directional volume on link V_{ij}; either the total flow rate or the through flow rate can be used;

$p =$ an integer exponent; the values 0, 1, 2 and 4 were used.

To obtain an objective function value that is consistent with those used previously, we normalize the weighting coefficients to obtain

$$\sum_{i=1}^{n-1} a_i = n-1 \quad \text{and} \quad \sum_{i=1}^{n-1} \bar{a}_i = n-1$$

The multi-band/multi-weight optimization program is summarized in MILP-2.

MILP-2. Find $b_i, \bar{b}_i, z, w_i, \bar{w}_i, t_i, \bar{t}_i, \nu_i$ **to**

$$\textbf{Maximize} \ \frac{1}{n-1}\sum_{i=1}^{n} a_i \cdot b_i + \bar{a}_i \cdot \bar{b}_i \ \textbf{subject to}$$

$$1/C_2 \le z \le 1/C_1$$

$$\frac{b_i}{2} \le w_i \le (1-r_i) - \frac{b_i}{2}, \quad \frac{b_i}{2} \le w_{i+1} \le (1-r_{i+1}) - \frac{b_i}{2},$$

$$\frac{\bar{b}_i}{2} \le \bar{w}_i \le (1-\bar{r}_i) - \frac{\bar{b}_i}{2} \quad \text{and}$$

$$\frac{\bar{b}_i}{2} \le \bar{w}_{i+1} \le (1-\bar{r}_{i+1}) - \frac{\bar{b}_i}{2} \quad i = 1, \ldots, n-1$$

$$t_i + \bar{t}_i + \frac{r_i + \bar{r}_i}{2} + (w_i + \bar{w}_i) - \frac{r_{i+1} + \bar{r}_{i+1}}{2} - (w_{i+1} + \bar{w}_{i+1}) - (\tau_{i+1} + \bar{\tau}_i) + \Delta_i - \Delta_{i+1} = \nu_i$$

$$i = 1, \ldots, n-1$$

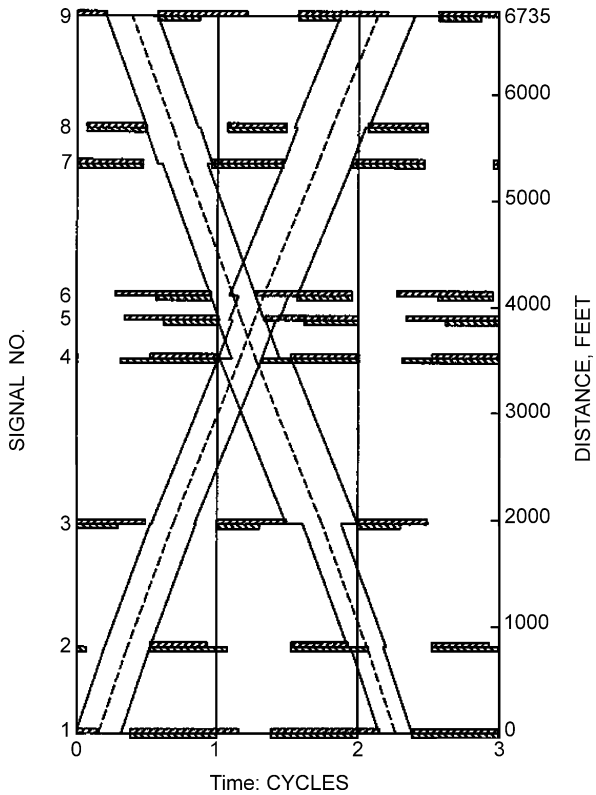

Traffic Networks, Optimization and Control of Urban, Figure 12
Time space diagram for a variable bandwidth progression scheme calculated by MILP-2

$$\frac{d_i}{f_i} z \le t_i \le \frac{d_i}{e_i} z \quad \text{and} \quad \frac{\bar{d}_i}{\bar{f}_i} z \le \bar{t}_i \le \frac{\bar{d}_i}{\bar{e}_i} z$$

$$i = 1, \ldots, n-1$$

$$\frac{d_i}{h_i} z \le \frac{d_i}{d_{i+1}} t_{i+1} - t_i \le \frac{d_i}{g_i} z \quad \text{and}$$

$$\frac{\bar{d}_i}{\bar{h}_i} z \le \frac{\bar{d}_i}{\bar{d}_{i+1}} \bar{t}_{i+1} - \bar{t}_i \le \frac{\bar{d}_i}{\bar{g}_i} z \quad i = 1, \ldots, n-1$$

$$b_i, \bar{b}_i, z, w_i, \bar{w}_i, t_i, \bar{t}_i \ge 0 \quad \text{and} \quad \nu_i \text{ integer}.$$

MILP-2 involves $(18n - 20)$ constraints and $(6n - 3)$ continuous variables, and $n - 1$ unrestricted integer variables, not counting slack variables. Green or red splits are determined the same as in MILP-1. The increased decision capabilities of MILP-2 require a corresponding increase in the size of the mathematical program. The formulation given in MILP-2 corresponds to the MULTIBAND optimization program developed by Gartner et al. [7]. The variable bandwidth solution for the same arterial as for MILP-1 (Fig. 9) is shown in Fig. 12.

The Network Problem

The MILP-2 problem can be extended for controlling traffic flow in a network of intersecting arterials. This represents a significantly more challenging problem: progressions must be provided on all the arterials of the network, simultaneously. Consider a network of m arterials with each arterial having n_j signalized intersections. Let S_{ij} denote the ith signal on the jth arterial of the network and L_{ij} denote the ith link (between signals i and $i+1$) of the jth arterial, with $j = 1, \ldots, m$ and $i = 1, \ldots, n_j$. The variables presented in MILP-1 and MILP-2 are redefined as follows:

$b_{ij}(\bar{b}_{ij}) =$ outbound (inbound) bandwidth of link i on arterial j;

$w_{ij}(\bar{w}_{ij}) =$ interference variables, the time from right (left) side of red at S_{ij} to the centerline of the outbound (inbound) green band;

$r_{ij}(\bar{r}_{ij}) =$ outbound (inbound) red time at S_{ij};

$t_{ij}(\bar{t}_{ij}) =$ travel time on link i of arterial j in the outbound (inbound) direction;

$d_{ij}(\bar{d}_{ij}) =$ length of link i of arterial j in the outbound (inbound) direction;

$\phi_{(ij),(jl)}(\bar{\phi}_{(ij),(kj)}) =$ internode offsets, time from the center of the outbound (inbound) red at S_{ij} to the center of the outbound (inbound) red at S_{kj};

$\Delta_{ij} =$ directional node phase shift, time from center of $\bar{r}_{ij}$ to nearest center of r_{ij};

$\tau_{ij}(\bar{\tau}_{ij}) =$ queue clearance time for advancement of outbound (inbound) bandwidth at S_{ij} to clear turning-in traffic before arrival of main-street platoon;

$V_{ij}(\bar{V}_{ij}) =$ outbound (inbound) progression speed on link L_{ij} (m/sec).

$e_{ij}, f_{ij}, (\bar{e}_{ij}, \bar{f}_{ij}) =$ lower and upper bounds on outbound (inbound) speed $V_{ij}(\bar{V}_{ij})$ (m/sec);

$g_{ij}, h_{ij}, (\bar{g}_{ij}, \bar{h}_{ij}) =$ lower and upper bounds on change in outbound (inbound) speed $V_{ij}(\bar{V}_{ij})$ (m/sec).

In addition to the above variables, the *intranode offset* is defined as:

$\omega_{(ij),(kl)}(\bar{\omega}_{(ij),(kl)}) =$ intranode offsets, the time from the center of the inbound (outbound) red at ith node of the jth arterial (S_{ij}), to the center of the inbound (outbound) red in the crossing direction at the same node which is also the kth node of the lth arterial (S_{kl}).

The directional interference constraints, the arterial loop integer constraint and the travel time and speed-change constraints remain unchanged and are applied, in turn, for each arterial. A new set of constraints, though, is required for the synchronization of all signals. Similar to the arterial loop-integer constraints Eq. (7), for any closed loops of the network consisting of more than 2 links, the summation of *internode* and *intranode offsets* around the loop of intersecting arterials must be an integer multiple of the cycle time. This results in the formulation of the *network loop-integer constraints*. The number of network loop constraints and the choice of a fundamental set of loops is given by Gartner [8].

The network optimization problem is summarized in MILP-3:

MILP-3. Find $b_{ij}, \bar{b}_{ij}, z, w_{ij}, \bar{w}_{ij}, t_{ij}, \bar{t}_{ij}, \nu_{ij}, \mu_i$ **to**

$$\textbf{Maximize} \sum_{j=1}^{m} \frac{1}{n-1} \sum_{i=1}^{n} a_{ij} \cdot b_{ij} + \bar{a}_{ij} \cdot \bar{b}_{ij} \textbf{ subject to}$$

$$1/C_2 \le z \le 1/C_1$$

$$\frac{b_{ij}}{2} \le w_{ij} \le (1 - r_{ij}) - \frac{b_{ij}}{2},$$

$$\frac{b_{ij}}{2} \le w_{i+1,j} \le (1 - r_{i+1,j}) - \frac{b_{ij}}{2},$$

$$\frac{\bar{b}_{ij}}{2} \le \bar{w}_{ij} \le (1 - \bar{r}_{ij}) - \frac{\bar{b}_{ij}}{2} \quad \text{and}$$

$$\frac{\bar{b}_{i,j}}{2} \le \bar{w}_{i+1,j} \le (1 - \bar{r}_{i+1,j}) - \frac{\bar{b}_{ij}}{2}$$

$$i = 1, \ldots, n_j - 1; j = 1, \ldots, m$$

$$t_{ij} + \bar{t}_{ij} + \frac{r_{ij} + \bar{r}_{ij}}{2} + (w_{ij} + \bar{w}_{ij}) - \frac{r_{i+1,j} + \bar{r}_{i+1,j}}{2} - (w_{i+1,j} + \bar{w}_{i+1,j}) - (\tau_{i+1,j} + \bar{\tau}_{i,j}) + \Delta_{\mathrm{i,j}} - \Delta_{i+1,j} = \nu_{\mathrm{i,j}}$$

$$i = 1, \ldots, n_j - 1; \quad j = 1, \ldots, m$$

$$\phi_{(ia),(i+1,a)} + \omega_{(i+1,a),(jb)} + \phi_{(jb),(j+1,b)} + \omega_{(j+1,b),(kc)} + \phi_{(kc),(k+1,c)} + \omega_{(k+1,c),(ld)} + \phi_{(ld),(l+1,d)} + \omega_{(ld),(ia)} = \mu_N$$

$$i = 1, \ldots, n_a - 1; \quad j = 1, \ldots, n_b - 1;$$

$$k = 1, \ldots, n_c - 1; \quad l = 1, \ldots, n_d - 1;$$

a, b, c, d arterials forming a fundamental network loop;

$$\frac{d_{ij}}{f_{ij}} z \le t_{ij} \le \frac{d_{ij}}{e_{ij}} z \quad \text{and} \quad \frac{\bar{d}_{ij}}{\bar{f}_{ij}} z \le \bar{t}_{ij} \le \frac{\bar{d}_{ij}}{\bar{e}_{ij}} z$$

$$i = 1, \ldots, n_j - 1; j = 1, \ldots, m$$

$$\frac{d_{ij}}{h_{ij}} z \le \frac{d_{ij}}{d_{i+1,j}} t_{i+1,j} - t_{ij} \le \frac{d_{ij}}{g_{ij}} z \quad \text{and}$$

$$\frac{\bar{d}_{ij}}{h_{ij}} z \le \frac{\bar{d}_{ij}}{d_{i+1,j}} \bar{t}_{i+1,j} - \bar{t}_{ij} \le \frac{\bar{d}_{ij}}{g_{ij}} z$$

$$i = 1, \ldots, n_j - 1; j = 1, \ldots, m$$

$$b_{ij}, \bar{b}_{ij}, z, w_{ij}, \bar{w}_{ij}, t_{ij}, \bar{t}_{ij} \ge 0 \quad \text{and}$$

ν_{ij}, μ_i integer.

For an $m \times n$ closed grid network ($m \times n$ intersections and $m+n$ arterials), MILP-3 may involve up to $(23mn-14m-14n+1)$ constraints and $(12mn-6m-6n)$ continuous variables, and $(3mn-2m-2n+1)$ unrestricted integer variables, not counting slack variables.

The principal difficulty in solving mixed-integer problems, such as the ones described above, is the number of integer variables and their range, i. e., the size of the integer feasible set. Virtually all MILP codes use branch-and-bound strategies for calculating the optimal values. Solving the primary multi-band problem by general purpose branch-and-bound is still a rather formidable task. For this reason, heuristic methods which quickly lead to a good solution have been developed [9].

Delay-Based Models – Cyclic Flow Profiles

TRANSYT

TRANSYT, the traffic network study tool, is a computer model to optimize traffic signal timings and perform traffic signal simulation. TRANSYT has two main elements – the traffic model which is used to calculate the performance index of the traffic network for a given set of signal timings and an optimizing process that makes changes to the settings and determines whether they improve the performance index or not.

Simulation Model

The traffic simulation model used in TRANSYT is a macroscopic traffic simulation model. In a macroscopic model platoons of vehicles are considered rather than individual vehicles. TRANSYT simulates traffic flow macroscopically, but in a step-wise manner. The cycle length is divided into small, equal time increments, called steps. All TRANSYT calculations are made on the basis of flow rates (such as vehicles per hour). The average flow pattern past a point in the road network is represented by a histogram also known as Cyclic Flow Profile. There are three traffic flow patterns considered while simulating behavior of a signalized intersection: the "IN", "GO", and "OUT" patterns (see figure below). [10].

The IN-Pattern or Arrival Flow Pattern

The IN-Pattern is the pattern of traffic arriving at the stop line. The arrival flow pattern is expressed mathematically as follows:

$$\mathrm{IN}_{it} = \sum_{j}^{n} F_{ij}(P_{ij} \cdot \mathrm{OUT}_{jt'})$$

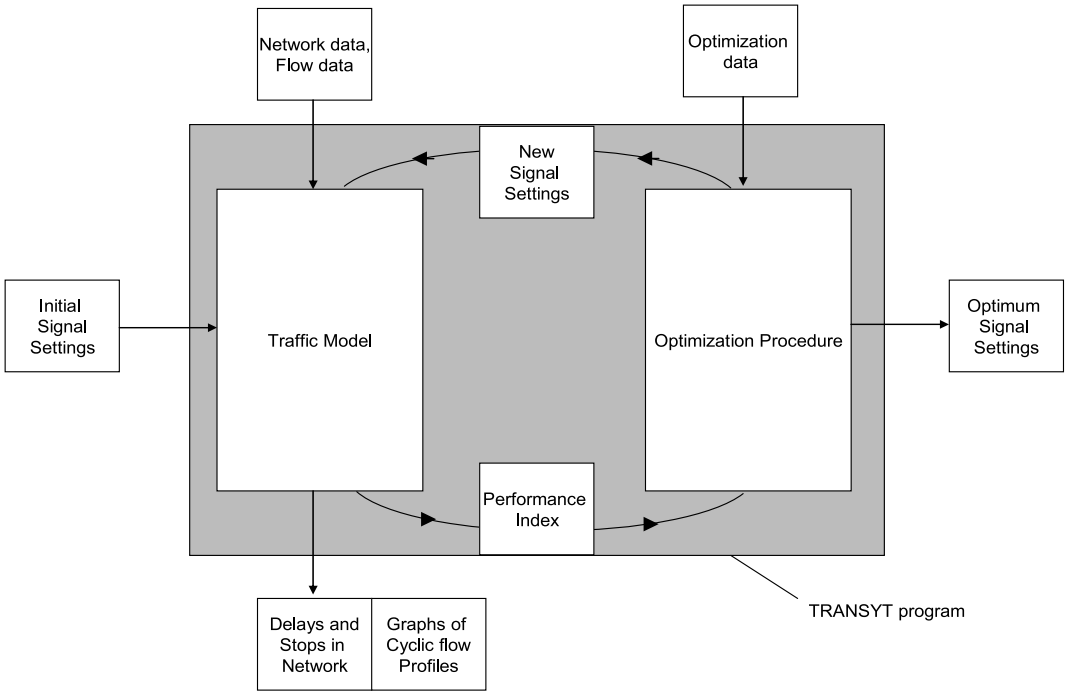

Traffic Networks, Optimization and Control of Urban, Figure 13
Structure of TRANSYT program

where,
IN_{it} = the IN-pattern on link i for time step t
F_{ij} = smoothing process related to platoon dispersion for flow to link i from link j
P_{ij} = the proportion of the feeding link OUT-pattern that feeds the subject link
$\mathrm{OUT}_{jt'}$ = the OUT-pattern of link j for step t'
t' = the time step t minus the travel time for flow to link i from link j
n = the number of links (j) that feed link i

The IN-pattern is computed for each step, t, in the cycle, thus forming a pattern, or flow profile.

The GO-Pattern or Saturation Flow Pattern

The GO-pattern is the flow rate at each step that would leave the stop line when the traffic signal turns green if there were enough traffic to saturate the green. There is a certain Startup Lost Time (SLT) involved due to reaction of driver etc.

The OUT-Pattern or Departure Flow Pattern

The OUT-pattern is the profile of traffic actually leaving the stop line. It is usually equal to the GO-pattern as long as there is a queue. After the queue dissipates, it is equal to the IN-pattern for the remainder of the effective green.

The queue (or the number of vehicles held at the stop line during any time interval, t) is first determined to determine the OUT-pattern

$$m_t = \max\{(m_{t-1} + q_t - s_t), 0\}$$

where,
m_t = number of vehicles in the queue in time interval t on a given link (and similarly for m_{t-1})
q_t = number of vehicles arriving in interval t, given by the IN-pattern
s_t = number of vehicles allowed to leave in interval t, given by the GO-pattern

The OUT-pattern is given for link i during time interval t by the expression:

$$\mathrm{OUT}_{it} = m_{i,t-1} + q_{it} - m_{it}\,.$$

The traffic model further utilizes a platoon dispersion algorithm that simulates the normal dispersion (i. e., the "spreading out") of platoons as they travel downstream. The start-and-stop operation of signals tends to create platoons of vehicles that travel along a link. TRANSYT models the dispersion of these platoons as they progress along a link.

On internal links, for each time interval (step), t, the downstream arrival flow is determined by the following recurrence equation:

$$v'_{(t+\beta T)} = F.v_t + [(1-F)\,.v'_{(t+\beta T-1)}$$

where,
$v'_{(t+\beta T)}$ = predicted flow rate in time interval $t + \beta T$ of the predicted platoon
v_t = flow rate of the initial platoon during step t
β = factor to take into account vehicles traveling faster than speed limit
T = the cruise travel time on the link, in steps
F = a smoothing factor, where:

$$F = (1 + \alpha\,\beta\,T)^{-1}\,,$$

where α = an empirically derived constant, called the platoon dispersion factor (PDF) takes into account site-specific factors such as grade, curvature, parking, opposing flow interference, and other sources of impedance.

Optimization Model

TRANSYT uses a Performance Index (PI) (also the objective function) allowing the user to define their preferences regarding performance of the traffic network. TRANSYT develops a signal timing plan that produces an optimal value of the PI. TRANSYT-7F (the US-adopted version) offers numerous PI's to choose from, which can reflect anything that the user desires including delay, progression, stops, fuel consumption, queuing, and throughput. These PI options are listed below:

$$\mathrm{PI} = \begin{bmatrix} \text{DI only} \\ \text{PROS only} \\ \text{PROS \& DI} \\ \text{PROS/DI} \\ \text{DI} \sum \frac{\text{Average back of queue on link } i}{\text{Queuing capacity on link } i} \\ \text{Throughput only} \\ \text{Throughput \& DI} \\ \text{Throughput/DI} \end{bmatrix}$$

In general, TRANSYT-7F always attempts to maximize the PI, unless the disutility index (DI) has been selected as the PI. Since the DI can be a combination of delay, stops, queuing, and fuel consumption, this value must be minimized. The other objective functions that involve progression or throughput must be maximized.

The DI is a combination of vehicle delay, stops, and fuel consumption. Weighting factors are available in order

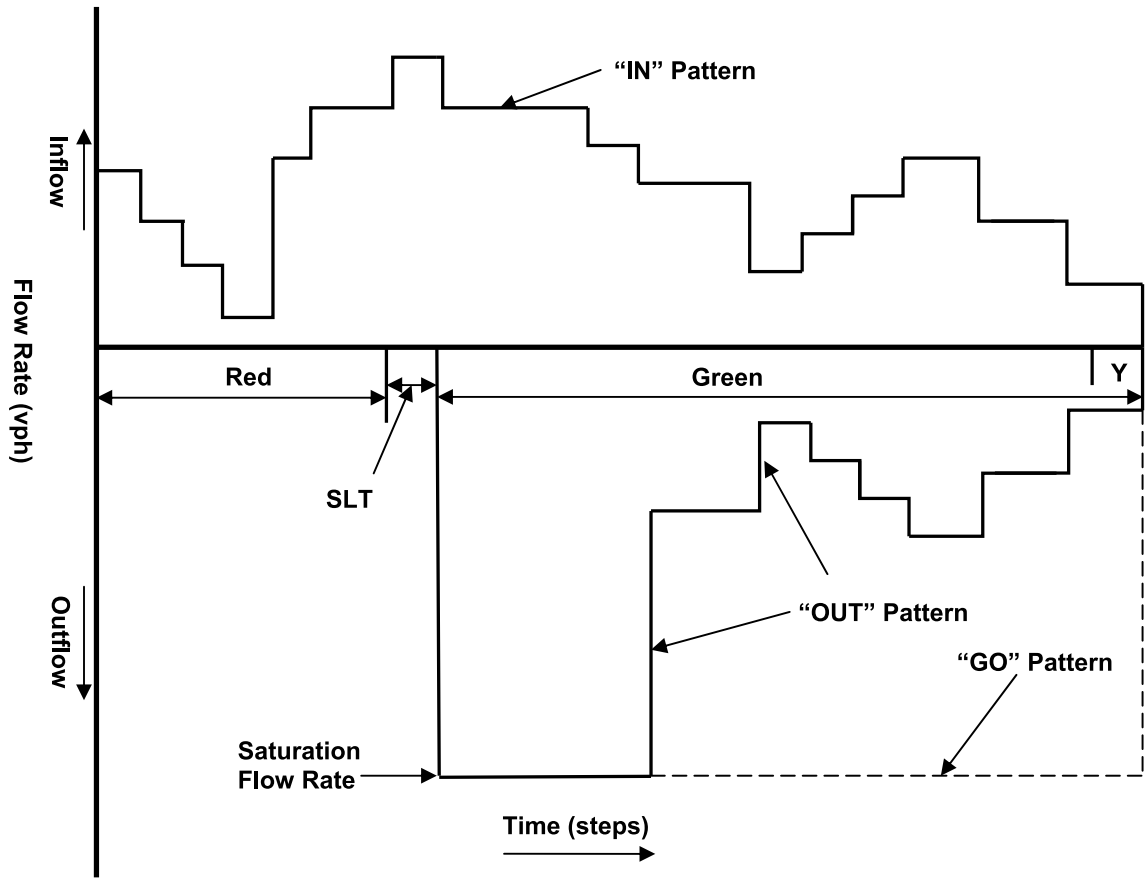

Traffic Networks, Optimization and Control of Urban, Figure 14
TRANSYT Flow Patterns

to place more emphasis on any of these three DI components if desired. Delay-only optimization results in excessive stops and fuel consumption. Excess fuel consumption (minimization) is considered to be a good compromise between bandwidth and delay-based optimization.

The "standard" delay and stops DI optimization objective function is defined as follows:

$$\mathrm{DI} = \sum_{i=1}^{n} \{(w_{d_i} d_i + K w_{S_i} S_i) + QP\}$$

where,

DI = disutility index, analogous to the original TRANSYT performance index

d_i = delay on link i (of n links) and on an optional user-specified upstream input link $i - 1$

K = user-coded stop penalty factor to express the importance of stops relative to delay

S_i = stops on link i per second

w_{xi} = link-specific weighting factors for delay (d) and stops (s) on link i

U_i = binary variable that is '1' if link-to-link weighting has been established, zero otherwise

QP = queuing penalty

A queuing penalty is used to minimize the possibility of spillback.

$$\mathrm{QP} = \mathrm{QB}_i W_{\mathrm{q}} (q_i - \mathrm{qc}_i)^2$$

where,

Q = a binary variable, '1' if the queue penalty is included in the DI '0' otherwise

B_i = a binary variable, '1' if the maximum back of queue (qi) exceeds the queuing capacity '0' otherwise

W_{q} = a network-wide penalty applied to the excess queue spillback

q_i = computed maximum back of queue on link i

qc_i = queuing capacity for link i

Progression opportunities (PROS) represent the ability of vehicles to progress through multiple intersections without stopping. Points are scored every time a vehicle is able to traverse additional intersections. Therefore it is possible to achieve PROS without having a wide bandwidth stretching from the beginning of an arterial street to the end.

It is also possible to select a combination of PROS and DI in order to achieve a compromise between the two objectives. For example, if the PI is defined as PROS/DI, the program attempts to simultaneously maximize the numerator (PROS) and minimize the denominator (DI). For optimizing bandwidth and progression speeds, PROS/DI is recommended. PROS-only is rarely acceptable, because although it often produces a good bandwidth, the minor movements are allowed to fail.

Demand-Responsive Models

A number of demand-responsive, or adaptive signal control strategies have been developed. Two such strategies are described herein: SCOOT and OPAC.

SCOOT

Adaptive traffic signal systems adjust signal timings based on measured flow. Reacting to these flow variations results in reduced delay, shorter queues and decreased travel time. SCOOT (Split Cycle Offset Optimization Technique) is one of the earliest real time adaptive traffic control systems [11]. It coordinates the operation of all the traffic signals in an area to give good progression to vehicles through the network.

The operation of the SCOOT model is illustrated in Fig. 15. SCOOT obtains information on traffic flows from detectors. As an adaptive system, SCOOT depends on good traffic data so that it can respond to changes in flow. Detectors are normally required on every link. Their location is important and they are usually positioned at the upstream end of the approach link.

When vehicles pass the detector, SCOOT receives the information and converts the data into its internal units and uses them to construct "Cyclic flow profiles" for each link. The sample profile shown in the diagram is color coded green and red according to the state of the traffic signals when the vehicles will arrive at the stop line at normal cruise speed. Vehicles are modeled down the link at cruise speed and join the back of the queue (if present). During the green, vehicles discharge from the stop line at the validated saturation flow rate.

The data from the model is then used by SCOOT in three optimizers which are continuously adapting three key traffic control parameters – the amount of green for each approach (Split), the time between adjacent signals (Offset) and the time allowed for all approaches to a signaled intersection (Cycle time). These three optimizers are used to continuously adapt these parameters for all intersections in the SCOOT controlled area, minimizing wasted green time at intersections and reducing stops and delays by synchronizing adjacent sets of signals. This means that signal timings evolve as the traffic situation changes without any of the harmful disruption caused by changing fixed time plans on more traditional urban traffic control systems.

OPAC – Optimization Policies for Adaptive Control

This section presents an adaptive control strategy for co-coordinating and synchronizing signals in a network using the virtual-fixed-cycle concept and is labeled *VFC-OPAC*. The strategy is based on the single intersection dynamic-programming-based *OPAC* (optimization policies for adaptive control) adaptive controller developed by Gartner [12]. *VFC-OPAC* consists of a distributed control strategy featuring a dynamic optimization algorithm that calculates signal timings to minimize a performance function of total intersection delays and stops. The algorithm uses a combination of measured and modeled demand to determine, in a distributed manner, phase durations at each signal that are constrained by minimum and maximum green times and, when running in a coordinated mode, by coordination and synchronization parameters that can be updated based on real-time data.

OPAC was developed from the outset as a distributed strategy featuring a dynamic optimization algorithm for traffic signal control without requiring a rigid, fixed cycle time. Signal timings are calculated to directly minimize performance measures, such as vehicle delays and stops, and are only constrained by minimum and maximum phase lengths and, if running in a coordinated mode, by a virtual cycle length and by a virtual offset. Development of the strategy has progressed through several versions, each one serving as a base for a subsequent version. The principal features of each version are outlined below.

OPAC I: Dynamic Programming

The first version, designated *OPAC I*, was designed to serve as a basis for subsequent *opac* strategy development. *OPAC I* utilizes a dynamic programming (DP) model for the determination of the traffic control parameters. Since DP is a global optimization strategy for multistage decision processes, it provides a standard against which all other strategies can be compared [13,14].

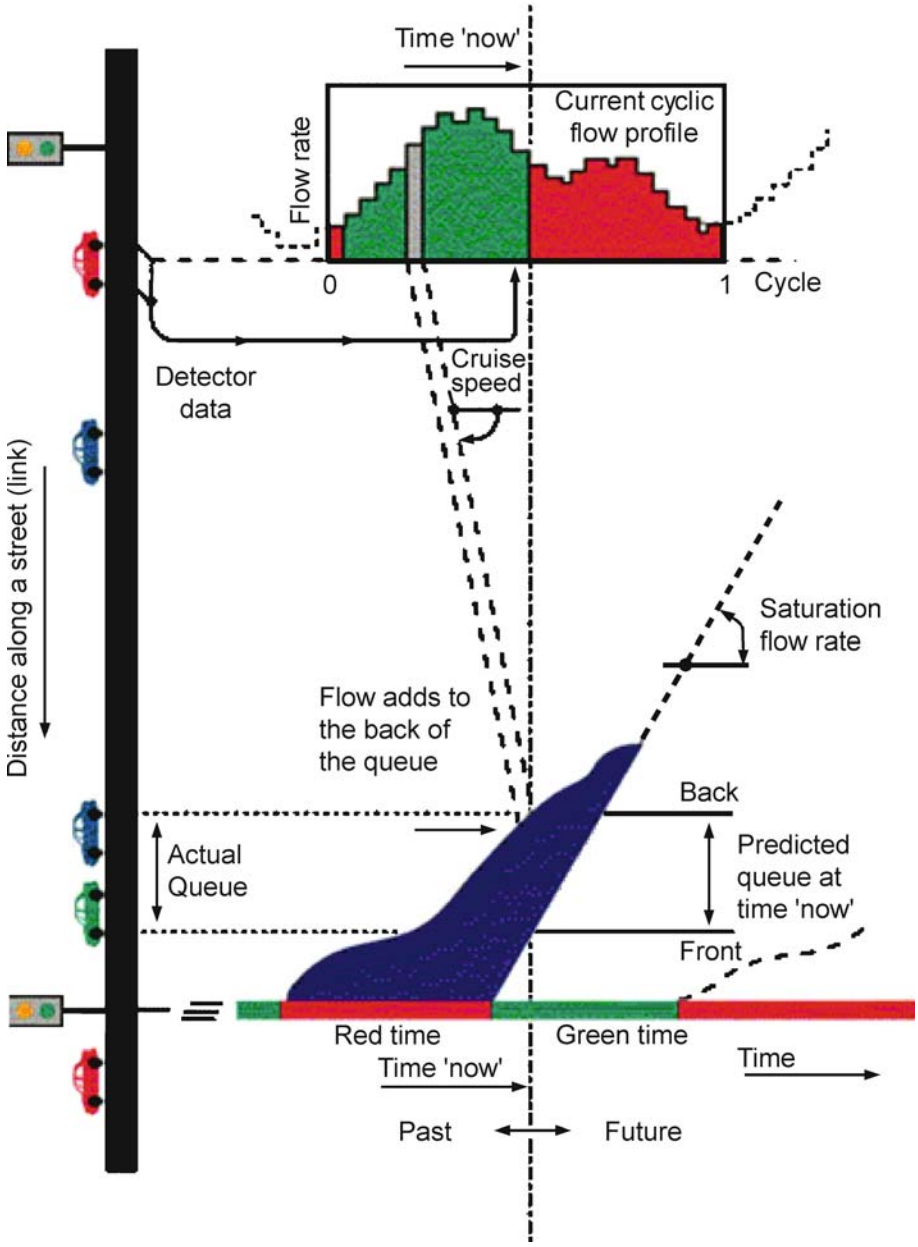

Traffic Networks, Optimization and Control of Urban, Figure 15
Schematic diagram showing operation of SCOOT model

The optimization process is decomposed into N stages, where each stage corresponds to a discrete time interval in which the arrivals are measured, (2 to 5 sec intervals). The total number of stages N corresponds to the *horizon length* (HL) of the input predictions. For exploratory purposes the horizon length was assumed to be several minutes long, e. g., 5, 10, or 15 mins. A typical stage i is illustrated in Fig. 16.

At stage i we have an input state vector I_i, an arrival vector A_i, output state vector O_i, input variable x_i, economic return (cost) output r_i, and a set of transformations:

$$O_i = T_i(I_i, A_i, x_i)$$
$$r_i = R_i(I_i, A_i, x_i)\,.$$

The state of the intersection is characterized by the state of the signal (green or red) and by the queue-length

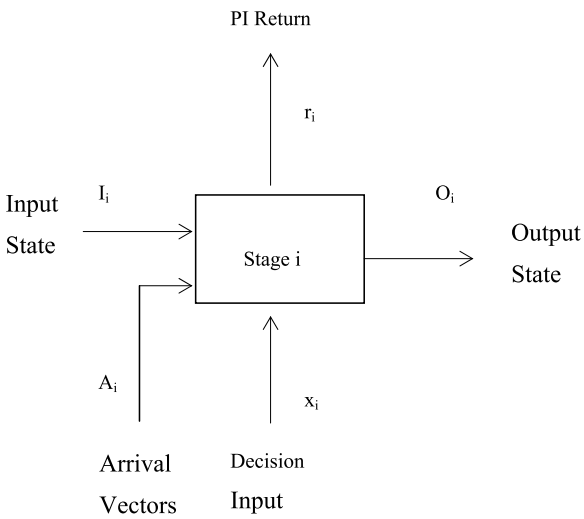

Traffic Networks, Optimization and Control of Urban, Figure 16
Dynamic Programming stage in OPAC I

on each of the approaches. The input decision variable indicates whether the signal is to be switched at this stage or to remain in its present state. The return cost output is the intersection's index of performance (e. g., total delay time and /or number of stops), which has to be minimized. The functional relationships between the input and output variables are based on the queuing-discharge process occurring at the intersection, i. e., the vehicle inflow and outflow as a function of the signal settings.

We define the following variables (all corresponding to stage i):

$a =$ approach designation, by direction, $a = \mathrm{N, S, E, W}$
$A^a =$ number of arrivals during the stage (for a four-leg intersection)
$D^a =$ number of departures (discharges) during the stage
$Q^a =$ queue-length on approach at beginning of stage
$S^a =$ status of signal at start of stage (green/red)
$x =$ stage input decision variable (change/no-change)

The input state vector (transposed) is: $I^{\mathrm{t}} = [I^{\mathrm{N}}, I^{\mathrm{S}}, I^{\mathrm{E}}, I^{\mathrm{W}}]$, where for each approach the state I

$$I^a = [S^a, \mathrm{Q}^a]^{\mathrm{t}} .$$

The output state vector contains the same elements as the input state vector and equals to the input state of the succeeding stage, i. e., $O_i = I_{i+1}$. The signal status for each approach a, is a binary variable:

$$S^a = \{0 \text{ for Green; } 1 \text{ for Red}\}$$

The input decision variable is also binary

$$x = \begin{cases} 0 & \text{no change in signal status} \\ 1 & \text{change current signal status} \end{cases}$$

The transformation of input to output, at stage i, is as follows:

$$S^a_{i+1} = (S^a_i + x_i)_{\text{ mod } 2}$$

$$Q^a_{i+1} = Q^a_i + A^a_i - D^a_i$$

The mod 2 operator ensures that $S^a_{\mathrm{i+1}}$ will always be 0 or 1. The arrivals at each stage i, $\mathrm{A}^a_{\mathrm{i}}$ are an observed input (e. g., from detectors); the departures are a function of the state and decision variables:

$$D = \begin{cases} 0 & \text{if } S = 1 \\ \min(Q + A, d_{\max}) & \text{if } S = 0 \end{cases}$$

where $d_{\max}$ is the saturation discharge rate (in veh/int). The DP algorithm goes backward in time, i. e., starting from the last interval and back-tracking to the first, at which time an optimal switching policy for the entire horizon is determined. The switching policy consists of the sequence of phase switch-ons and switch-offs throughout the horizon. The recursive optimization functional is:

$$f^*_i(I_i) = \min_{x_i} \left\{ R_i(I_i, A_i, x_i) + f^*_{i+1}(I_i, A_i, x_i) \right\} .$$

The Performance Index (return) at stage i is the queuing delay and number of stops N_{s} incurred at this stage:

$$r_i = R_i(I_i, A_i, x_i) = \sum_a (Q^a_i + A^a_i - D^a_i) + \sum_a N^a_s .$$

When the optimization is terminated at stage $i = 1$ we have,

$$f^*_1(I_i) = \min_{x_i} \left\{ \sum_{i=1}^{N} R_i(I_i, A_i, x_i) \right\} = \sum_{i=1}^{N} R_i(I_i, A_i, x^*_i) \tag{12}$$

which is the minimized Performance Index over the horizon period for a given initial input state I_1. Since the initial conditions at stage 1 are specified (i. e., the queue-lengths on all approaches are given as well as the initial signal status), we can retrace the optimal policy by taking a forward pass through the arrays of $X^*_i(I_1)$. The pol-

icy consists of the optimal sequence of switching decisions $\{x_i^*,\ i = 1, \ldots, N\}$ at all stages of the optimization process.

While this procedure assures globally optimal controls for the given horizon length, it requires complete information of arrivals over the entire control period. It cannot be used for real-time implementation due to both the (excessive) amount of processing involved and due to the lack of a practical method to gather real-time information for such a length of time. Much of the output generated by the procedure is never implemented because optimized policies are calculated for all possible combinations of initial conditions at each stage of the control period. In practice, only one 'optimum policy' is implemented. Nonetheless, *OPAC-I* serves an important function as a standard for the evaluation of the relative effectiveness of other, more practical strategies.

OPAC II: Sequential Optimization

OPAC II breaks up the horizon into sequential optimization stages to speed up the optimization process. The model is a reformulation of the *OPAC I* algorithm. The purpose is to create a building block for a distributed online strategy. *OPAC II* has the following features:

- The control period is divided into successive (back-to-back) horizon lengths of T seconds each (T may encompass one or more cycle lengths).
- Each horizon is divided into an integral number of intervals 't' seconds long; typically, $t = 2 - 5$ sec.
- During each horizon there must be a sufficient number of phase changes to guarantee that no optimal solution is missed. The phase change (switching) times are measured from the start of the horizon in time units of t.
- For any given switching sequence in a horizon, the performance function for each approach computes the total delay and/or stops.

The optimization problem in OPAC II can be stated as follows: For each horizon length, given the initial queues on each approach and the arrivals for each interval of the horizon, determine the sequence of switching times, in terms of intervals, which yield the least delay and/or stops to vehicles over the entire horizon.

The procedure used for solving the problem consists of an intelligent search over the set of all possible combinations of feasible switching times within the horizon to determine the optimum sequence. Valid switching times are constrained by minimum and maximum phase durations. The problem can be re-formulated as an alternative dynamic programming problem by re-defining the control variable x_j to denote the amount of green plus yellow time allocated to stage s_i. The stages, in this case, correspond to the phase lengths during the horizon. The recursive optimization functional (forward DP), is:

$$f_j(s_j) = \underset{x_j}{\text{Min}} \left\{ R_j(s_j, x_j) + f_{j-1}^*(s_{j-1}) \right\} . \tag{13}$$

The return (or Performance Index) is now

$$R = \sum_a \int_{s_{j-1}}^{s_j} A^a(\text{t})\text{dt} + \sum_a N_s^a . \tag{14}$$

This calculation results in the optimal sequence of signal phase (stage) lengths during the horizon. By reformulating the DP model, computation is considerably more economical than in *OPAC I*. *OPAC II* lends itself more readily to operation in real-time than does *OPAC I*; however, it still requires information on arrivals (flows) over the entire horizon length.

OPAC III: A Rolling Horizon Approach

A typical horizon length is 1–2 minutes long. Obtaining accurate arrival predictions for this length of time is not feasible with current technology. To use only readily available flow data without degrading the performance of the optimization procedure, a 'rolling horizon' strategy is applied to the *OPAC II* algorithm. In this version, the horizon length, or the *Projection Horizon* is the period for which traffic patterns are projected and optimum phase change information is calculated. The key feature is that real-time data are required for only a small portion of the horizon.

Figure 17 is an illustration of the rolling horizon procedure. From detectors placed upstream of each approach actual arrival data for k intervals can be obtained for the beginning, or head, portion of the horizon. For the remaining $n - k$ intervals, the tail of the horizon, flow data may be obtained from a model. A simple model consists of a moving average of all previous arrivals on the approach. An optimal switching policy is calculated for the entire horizon, but only those changes which occur within the head portion are actually being implemented. In this way, *OPAC III* can dynamically revise the switching decisions as more recent (i. e., more accurate) real-time data continuously become available.

By placing the detectors well upstream of the intersection (10 to 15 sec. travel time) one can obtain actual arrival information for the head period. This allows for a more reliable estimation of delay for any given phase change decision. At the conclusion of the current head pe-

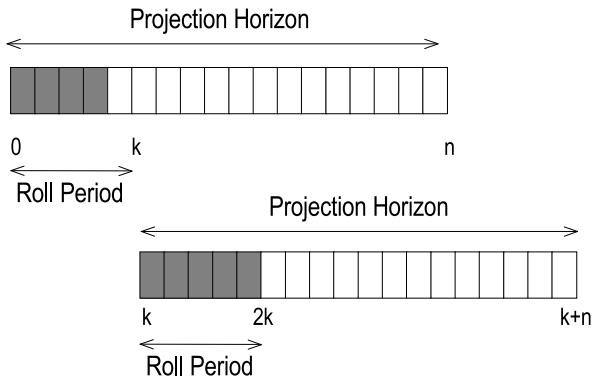

Traffic Networks, Optimization and Control of Urban, Figure 17
Implementation of the rolling horizon approach in OPAC

riod, a new projection horizon containing new head and tail periods is defined with the new horizon beginning at (rolled to) the termination of the old head period. Calculations are then repeated for the new projection horizon. The roll period can be any multiple number of steps, including one. A shorter roll period implies more frequent calculations and, generally, closer to optimum (i. e., ideal) results. Extensive simulation and field tests have demonstrated the effectiveness of this strategy in achieving performance that approaches the theoretical optimum established by *OPAC I*. The placing of the detectors and the information flow among the intersection controllers is illustrated in Fig. 18.

OPAC IV: The Virtual Fixed Cycle Strategy

Rhythmic, or cyclic operation is essential for coordination of signals. Whereas an independent *opac* controller is cycle-free, linking of adaptive signals in a network configuration requires coordinated operation to facilitate opportunities for unimpeded progression. This is achieved in the *vfc-opac* model by re-introducing the concepts of cycle time and offsets in a benign manner that allows for increased flexibility of signal timing selection and control strategy implementation. *Vfc-opac* controlled intersections interact with neighboring intersections (fixed-time, or other *vfc-opac* controlled) in response to projected arrival flows from upstream feeder signals. The model offers, at the option of the user, a coordination-synchronization strategy that is suitable for implementation in arterials and in networks. The strategy is referred to as *virtual-fixed-cycle* because from cycle to cycle the yield point, or local cycle reference point, is allowed to range about the fixed yield points dictated by the virtual cycle length and the virtual offset. This allows the synchronization phases to terminate early or extend later to better manage dy-

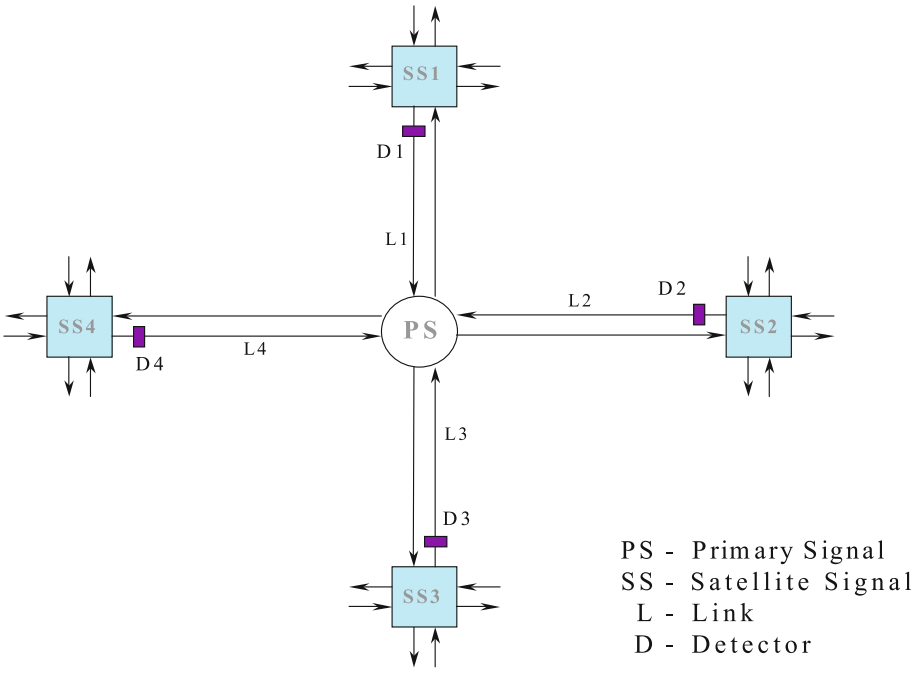

Traffic Networks, Optimization and Control of Urban, Figure 18
Information processing at an OPAC controlled intersection

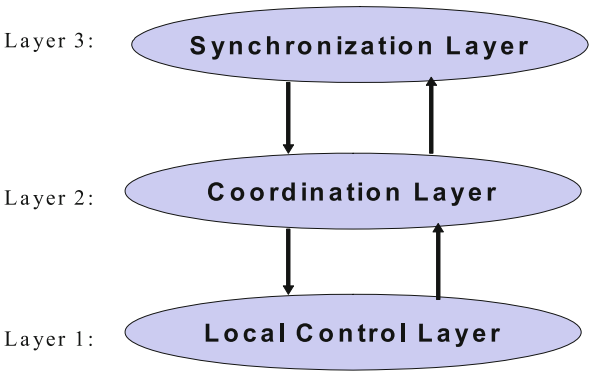

Traffic Networks, Optimization and Control of Urban, Figure 19
Control architecture in VFC-OPAC

namic traffic conditions. *Vfc-opac* consists of a three-layer control architecture as shown in Fig. 19.

Layer 1: The *Local Control Layer* implements the *OPAC III* rolling horizon procedure using the dynamic programming model of *OPAC II*. It continuously calculates optimal switching sequences for the Projection Horizon, subject to the *VFC* constraint communicated from Layer 3.

Layer 2: *The Coordination Layer* optimizes the offsets at each intersection (once per cycle). This is done by searching for the best offset of the *PS* (primary signal) within the mini-network shown in Fig. 4. A choice of three offset increments is being considered: 0 (no change), $+2-$ sec (move right one interval), $-2-$ sec (move left one interval). The cyclic flow profile associated with each incoming link is being discharged by the model through the intersection and projected to the downstream intersections, *SS* (satellite signals). All other parameters are being kept at their latest values in a relaxation mode. Since the coordination process is carried out in a distributed fashion at each intersection, each *SS*, in its turn, is also considered a *PS* of its own mini-network once during each cycle.

Layer 3: *The Synchronization Layer* calculates the network-wide *virtual-fixed-cycle* (once every few minutes, as specified by the user) in order to maintain a rhythmic operation of the signals in the network. The objective is to provide maximum leeway of phase switching timings as dictated by local conditions, yet maintain a capability for coordination with neighboring intersections by maintaining synchronicity of the signals which are linked in the network. The *virtual-fixed-cycle* (VFC) is calculated in a way that provides sufficient capacity at the most heavily loaded intersection while, at the same time, maintaining suitable progression opportunities among adjacent intersections. The VFC is calculated as follows:

- Check if pre-set time period (3–5 min), or number of cycles counted, has elapsed
- Identify the dominant intersection (based on flow/saturation flow ratios)
- Establish bounds on VFC to satisfy the following requirements:
  a. Provide sufficient capacity

$$C_a \geq \sum_j (G_j)_{\min} + L$$

  b. Keep the degree of saturation under a preset maximum k_m (e. g., $k_m = 0.90$)

$$C_b \geq \frac{k_m L}{k_m - \sum_j y_j}$$

  c. Do not exceed upper phase length limits

$$C_c \leq \sum_j (G_j) \max + L$$

  where, G_j = length of (green) phase j
  - $y_j = (q/s)_j$ = ratio of flow/sat-flow on phase j
  - L = total lost time at the intersection

The virtual-fixed-cycle is chosen as the lowest value satisfying these requirements;

$$\mathrm{VFC} = \min(C_a, C_b, C_c)\,.$$

In addition, there may be exogenously determined limits on the cycle time

$$C_{\min} \leq \mathrm{VFC} \leq C_{\max}\,.$$

The three-layer architecture is an effective means for implementing the distributed dynamic programming (DDB) algorithm. The VFC can be calculated separately for groups of intersections, as desired. The position of the marker line also determines the "virtual offset" of the signal. Over time the flexible cycle length and offsets are updated as the system adapts to changing traffic conditions. Due to the embedded flexibility this approach can provide improved local adaptability while maintaining network coordinability. Further information on the virtual-fixed-cycle model is given in [15].

Multi-Level Traffic Control Strategies

The performance of the transportation system is the result of a complex interaction between the physical supply of transportation facilities, and the individual trip-makers' decisions. The supply includes the management policies and operational controls that are applied in these facilities. The flows in a transportation network are the result of an "equilibration" process between the demand for transport services and the supply of transport capacity. This was the basis for the development of "static" models for the design and planning of TSM actions [16,17].

The extension of the static framework to the quasi-dynamic and dynamic cases, where on-line control and guidance is being provided in response to real-time traffic information. The application of advanced technologies in sensing, communications and computation in intelligent transportation systems, coupled with advanced modeling concepts, provides great opportunities to improve the performance of traffic networks under both recurrent and non-recurrent conditions. Current research and development efforts are directed toward development of a Dynamic Traffic Assignment (DTA) capability to predict future traffic conditions and a Real Time Traffic Adaptive Control System (RT-TRACS) for the generation of signal control strategies. Although these models are intimately connected, their development has proceeded independently of each other so far. In the framework described herein the two models are integrated into a combined system.

The Traffic Management Problem

The traffic management problem can be viewed as the interaction between the urban traffic manager and the individual trip-maker and their (sometimes differing) perceptions of the performance of, or the supply provided by, the transportation system. A schematic illustration of the interactions in a transportation network is shown in Fig. 20. The physical transportation system and the socioeconomic activity system interact, via the equilibration process, to produce a set of flows on the links of the network. Two major feedback loops affect this process. In Loop I, the traffic manager assesses the system's performance according to his measures-of-effectiveness and intervenes in the physical transportation system to achieve his desired objectives (or, more accurately, the objectives of the community which he represents), such as reducing total travel times, minimizing pollution, increasing safety, etc. The trip-makers, on the other hand, assess the flows according to their own perceptions, which may be different from those of the traffic manager, and propagate an adjustment in the travel demand pattern via Loop II. Performance of the system is a result of the combined interaction of the demand and the supply feedback loops.

The problem can be set in the following context: given (fixed) demands for travel in an urban area and a (fixed) supply of transportation facilities, the traffic manager must consider a variety of management strategies to induce a traffic flow pattern that will meet, in an optimal way, the overall objectives of the community. The manager's criteria for evaluating his actions may include public interest measures of performance (MOP) such as travel time, energy consumption, noise and pollutant emissions, etc. On the other hand, individual trip-makers are assumed to minimize only their own travel costs (usually the travel time), subject to the constraints imposed by the physical system supply, the manager's actions, and the interactions resulting from the other trip-makers' decisions. The problem can be cast as *system-optimization* (the manager's objective) for flows that result from *user-optimization* (the trip-makers' objective). This leads to a compound mathematical optimization formulation, as follows:

$$\min_{M} \left\{ Z_1 \left[M, \arg\min_{F} Z_2(F) \right] \right\} \tag{15}$$

subject to flow conservation constraints:

$$\sum_{P \in P_{(i,k)}} h_p = r_{(i,k)} \quad \text{with } (i,k) \in I \tag{16a}$$

and non-negativity constraints:

$$h_p; r_{(i,k)} \geq 0 \tag{16b}$$

where:

$M =$ the set of management variables under the control of the traffic manager

$I =$ set of all OD pairs (i, k)

$P_{(i,k)} =$ set of all simple paths between OD pair (i, k)

$f =$ set of all link flows

$f_j =$ flow on link j

$r_{(i,k)} =$ rate of trip interchange (demand in vehicles) between origin node i and destination node k, with $(i, k) \in I$

$h_p =$ flow on path p, with $p \in P_{(i,k)}$.

The compound objective function in Eq. (15) consists of the following components:

$$Z_1 = \sum_j f_j \cdot w_j(f_j, M) \tag{17a}$$

$$Z_2 = \sum_j \int_0^{f_j} c_j(x) \cdot \mathrm{d}x \tag{17b}$$

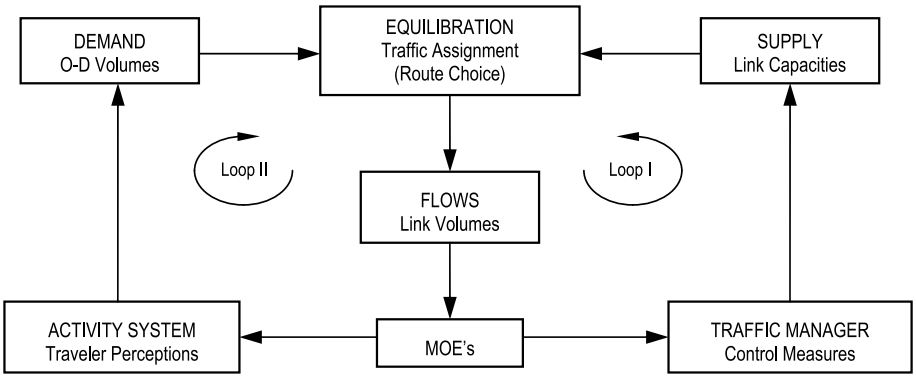

Traffic Networks, Optimization and Control of Urban, Figure 20
Static interactions in a transportation system

where $c_j(f_j)$ is the average user-perceived travel cost function on link j and $w_j(f_j, M)$ is a more general performance function reflecting the multiplicity of objectives pursued by the traffic manager in the public's interest. The traffic management problem stated by Eq. (15) consists of a primary optimization program (the *system-optimization*), and a secondary optimization program (the *user-optimization*). The "argmin" of a mathematical program is the optimal solution of the program; therefore, $Z_1(\cdot)$ has to be evaluated at the optimal solution of the secondary optimization program and represents aggregate traffic performance in the network with user-optimal flows.

An alternative mathematical representation of the traffic management problem is given by Tan et al. [18] and is labeled "hybrid optimization". In this case the objective function of the traffic management problem coincides with the system optimization objective described in Eq. (17a), while the user optimization objective acts as a constraint. Tan et al. show that Wardrop's First Principle describing the user optimization problem can be fully expressed mathematically as follows:

1. Every path p must carry non-negative flows:

$$h_p \geq 0$$

2. Path flows of the same O-D pair sum to the required flow $\sum_{P \in P_{(i,k)}} h_p = r_{(i,k)}$ with $(i, k) \in I$
3. Utilized paths have the same cost c_q which is equal to the minimum cost and paths with a greater cost than the minimum carry no flow. This is expressed by the following:

$$c_q(h_q, w_q) \geq \sum_{p \in P_{(i,k)}} h_p \cdot c_p(h_p, w_p) \cdot c_p(h_p, w_p)/r_{(i,k)}$$

$$\text{with } (i, k) \in I \text{ and } p \in P_{(i,k)} .$$

Thus the hybrid optimization formulation of the traffic management problem is:

$$\text{Minimize} \quad Z_1 = \sum_j f_j \cdot w_j(f_j, M)$$

subject to conditions (1.), (2.) and (3.) above.

Decision variables in the primary program are the set of management variables M under the control of the traffic manager. Decision variables in the secondary program are the link flows F, given the OD demands $r_{(i,k)}$. Major categories of actions that can be considered are: (a) actions to ensure the efficient use of existing road space; (b) actions to reduce vehicle use in congested areas; and (c) actions to improve transit service. For example, representative variables that are particularly amenable to formal optimization include traffic signalization variables (cycle time, green splits, offsets, signal phasing and phase sequencing) traffic operations variables (e. g., channelization, one-way streets, metering, variable message signs, reversible lanes) and preferential treatment variables (e. g., reserved or preferential lanes, exclusive transit lanes, transit vehicle preemption of traffic signals, special turning lanes, truck routes, parking control). The compound formulation and the hybrid-optimization are similar models; however, the latter has been more thoroughly investigated and, therefore, is more amenable to computation.

The framework given above for the static traffic management problem can be extended to the dynamic case. In principle, the same model applies; however, both the OD demands and the control actions M are now time-dependent. Further information is given in [19].

Signal Control and Traffic Assignment

Although signal control and route choice are intimately intertwined on the street, there is little practical experience with the integration of signal control and traffic as-

signment from a modeling standpoint. As indicated previously, most modeling efforts were done independently, i. e., assignment separate from control.

Smith [20] proposed a combined assignment-control model to take into consideration the effect of signal settings on route choice. The model is based on the P_0 control policy which is designed to maximize the "travel capacity" of the road network, under the assumption that all drivers seek their own best route. From computations conducted by Ghali and Smith [21] it was concluded that the P_0 control policy performs overall better than two standard control policies, (a) the local delay minimization for current traffic patterns at each intersection and (b) Webster's equisaturation method, where green times are selected to equalize the degree of saturation on competing approaches. Gartner and Al-Malik [22] presented a combined network model which simultaneously accounts for both the route choices made by motorists and the desired signal controls to match these choices. An important feature of this approach is that offset optimization can also be incorporated.

Real Time Traffic-Adaptive Control System (RT-TRACS)

RT-TRACS consists of a multi-level hierarchy of traffic control strategies and is illustrated in Fig. 21. This system provides a modern approach to adaptive traffic signal control and has been specifically designed to facilitate integration with traffic prediction [23]. The section below describes an RT-TRACS version that can also be incorporated within the framework of a dynamic traffic management system which is a dynamic version of the system illustrated in Fig. 20.

The degree of adaptability corresponds to the generation of the control strategy development, where n-GC is the nth generation control. Each level encompasses the capabilities of the lower levels in a nested fashion. The different levels and their interaction with predicted traffic data are described below.

Level 0 is the most basic type of signal control. The traffic engineer determines timing plans manually, or by some PC-based optimization program, using historical volume data. This level is suitable for recurrent traffic patterns, i. e., when there is little need to predict traffic flows, and there is infrequent updating of plans.

Level 1 corresponds to the 1-GC system: centralized control with limited availability of traffic surveillance and communications. Timing plans are calculated off-line using a multitude of scenarios to create a library of plans that can be activated in response to changing traffic and network conditions. Time of day (TOD) and traffic responsive (TRSP) options are available. Update interval is 15 minutes or higher. This level is suitable for application in the quasi-dynamic case. Integration with DTA is achieved in the following way: given a predicted O-D matrix, the system selects the most suitable timing plan in the library for the upcoming period. The plan should be created by a combined traffic assignment/signal optimization procedure and

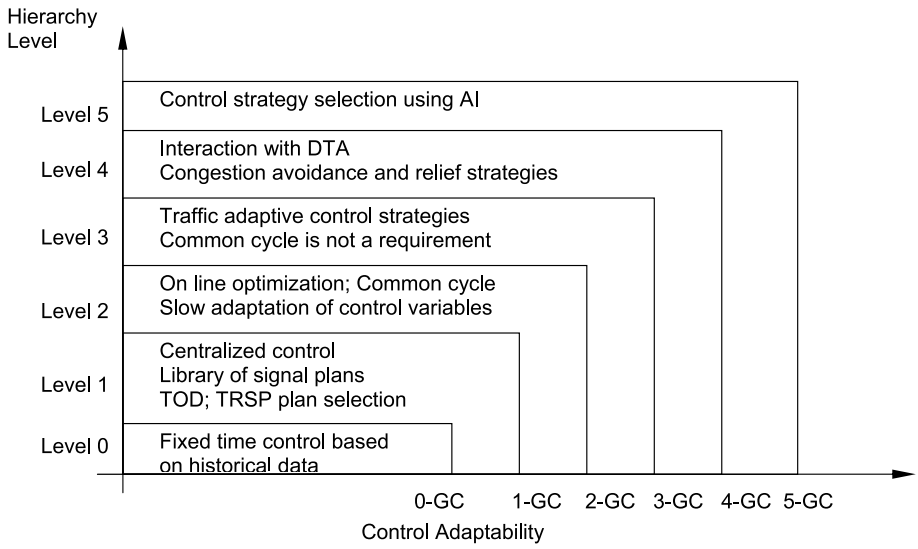

Traffic Networks, Optimization and Control of Urban, Figure 21
Nested hierarchy of RT-TRACS control levels

should match the current demand and supply conditions.

Level 2 (2-GC) is centralized control with on-line optimization using a fixed, common cycle. Similar to Level 1, time frames for optimization are shorter (5 to 10 minutes), plans are being generated on-line. This level of control can respond more rapidly to changes in travel demands and/or capacities. For example, progression schemes for priority routes can be generated on-line in conjunction with the traffic prediction module.

Level 3 (3-GC) is a fully-adaptive traffic signal control system that may also incorporate the capabilities of the previous levels if warranted. A common, fixed cycle time is not required for coordination. Capability of phase sequence optimization is available at selected locations. Example of a 3-GC strategy: the OPAC network version. This level of control is capable of reacting to rapidly changing traffic conditions and can also be used in conjunction with assignment capabilities to provide optimal control and rerouting of traffic. The latter capability is incorporated in the next level.

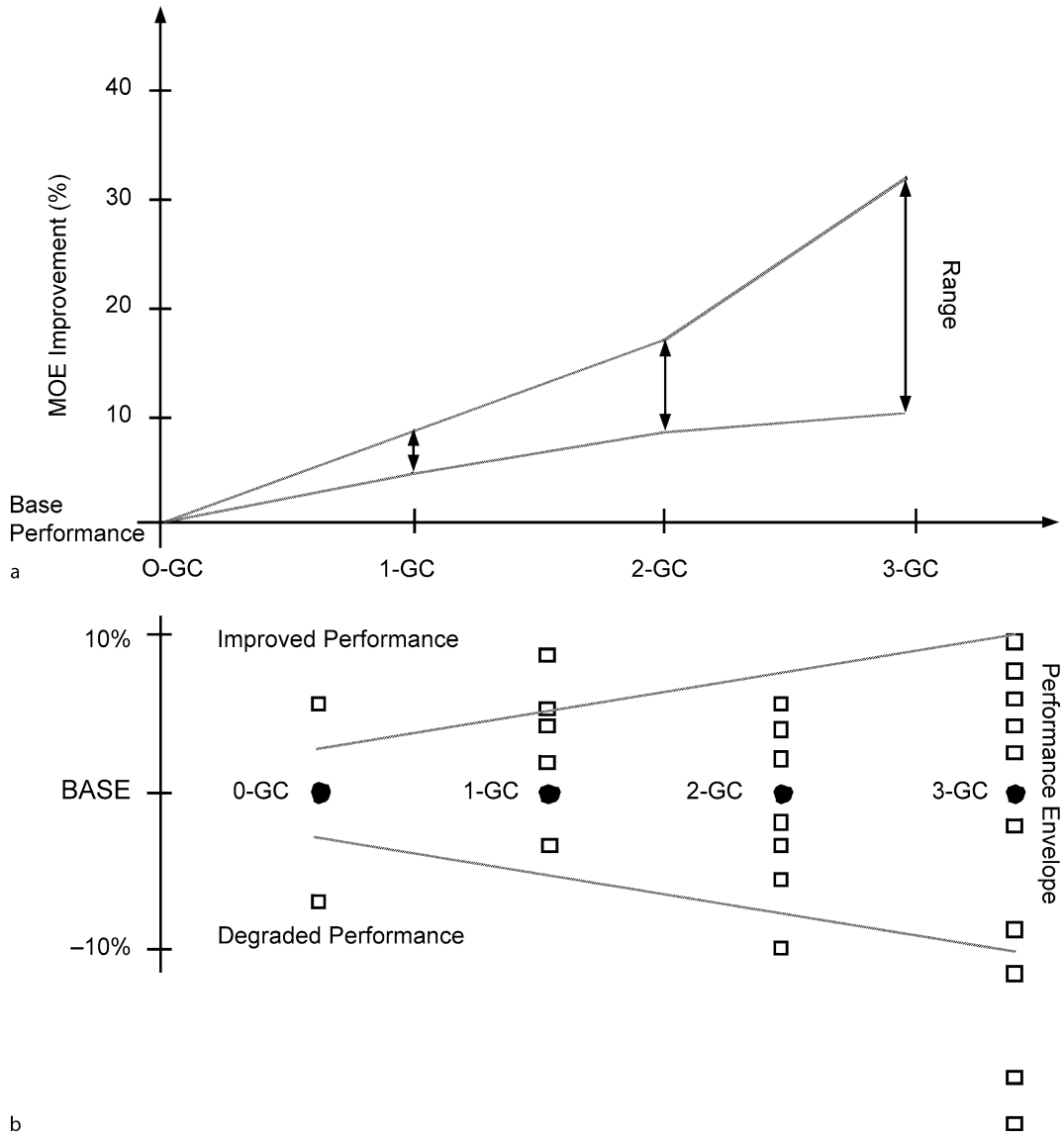

Traffic Networks, Optimization and Control of Urban, Figure 22
a Expected relative performance of control generation. b Reported relative performance of control generation

Level 4 includes all the capabilities of 3-GC with additional intelligence so that the system can interact with the DTA module. It can provide dynamic priority control on selected routes and implement congestion avoidance and relief strategies. This level is most suitable for handling incident and accident conditions.

Level 5 is a super level that makes the most efficient use of the available array of control strategies based on accumulated experience under local conditions. Selection of the appropriate control strategy for a particular condition can be invoked by Artificial Intelligence technology. Practical experience has shown that no single strategy is optimal for the entire spectrum of demand and supply variations. Therefore, it is beneficial to have available an arsenal of strategies and to be able to choose the most appropriate one when needed.

RT-TRACS provides a new design that is different from the traditional signal control generation strategies that are still state-of-the-art. In the traditional strategies no consideration is given to re-assignment of traffic, whereas here assignment is an integral part of the signal control generation process. The mutual interdependence between route choice and signal control is embodied at all levels of the hierarchy.

Control System Performance – Analysis of Complexity

The hierarchical framework described above offers the possibility to tailor the system's capabilities to particular needs and means. Advanced strategies can be deployed gradually and can coexist with lower level strategies. The full spectrum of capabilities can be built-up gradually in terms of deployment of sensors, surveillance equipment, controllers and communications, as well as central control hardware and software. It has been common wisdom that increasing responsiveness will contribute to improved traffic performance. Experience indicates that this has not always been the case. The more complex a system is the wider is the dispersion of the performance results [23]. The argument can be best illustrated with reference to Fig. 22. It shows the spreads in performance that are being perceived to exist among the different control generations (Fig. 22a), as well as the spreads in performance that were actually measured in the field (Fig. 22b). The points shown in the figure were collected from published reports in the literature. These figures illustrate an interesting phenomenon: the more "advanced" (i. e., responsive) strategies do not lead to improved performance all the time, notwithstanding the common perception that they do. More likely, they lead to a wider spread in performance results compared with a common basis. In some cases, due to reasons that are not completely explained, traditional off-line methods perform better than responsive methods. Therefore, one of the principal objectives of any intelligent control system would be to invoke a particular control strategy that will be most suitable for existing conditions so that overall performance of the system is optimized.

An expert system can be used to make sure that we implement a particular generation of control strategies when it is likely to yield the best performance. Development of the system requires careful characterization of signal networks and identification of the particular traffic flow patterns that are most amenable to benefit from a particular control strategy. The underlying proposition is that lower level strategies may often be as good, or better than higher level, more advanced strategies. By recognizing the conditions under which the particular strategies perform best, we can optimize overall system performance. This is the task of the 5th generation control level.

Future Directions

With the continuing developments in technology, one can expect to have gradual improvements in the technology of detection, communication and optimization which would lead to better real-time information on the status of the transportation system and, thus, to improved online control and performance. Specific areas in which advances can be sought and are likely to be achieved are: traffic-adaptive control of signal systems and the combined optimization of controls and routing. In particular, the latter would benefit from the proliferation of in-vehicular route guidance devices that are tied-in to advanced traveler information systems.

Bibliography

Primary Literature

1. Webster FV (1958) Traffic signal settings. HMSO, London
2. Gerlough DL, Huber MJ (1975) Traffic flow theory. Special report 165. Transportation Research Board, Washington DC
3. Greenshields BD, Schapiro D, Ericksen EL (1947) Traffic performance at urban street intersections. Technical Report, No 1. Yale Bureau of Highway Traffic
4. Morgan JT, Little JDC (1964) Synchronizing traffic signals for maximal bandwidth. Oper Res 12:896–912
5. Little JDC (1966) The synchronizing of traffic signals by mixed-integer linear programming. Oper Res 14:568–594
6. Little JD, Kelson MD, Gartner NH (1982) MAXBAND: A program for setting signals on arteries and triangular networks. Transp Res Rec 795:40–46

7. Gartner NH, Assmann SF, Lasaga F, Hou DL (1991) A multi-band approach to arterial traffic signal optimization. Transp Res 25B(1):55–74
8. Gartner NH (1972) Constraining relations among offsets in synchronized signal networks. Transp Sci 6:88–93
9. Gartner NH, Stamatiadis C (2002) Arterial-based control of traffic flow in urban grid networks. Math Comput Model 35: 657–671
10. Robertson DI (1969) TRANSYT: A traffic network study tool. TRRL Report No LR 253. Transportation and Road Research Laboratory, Crowthorne
11. Hunt PB, Robertson DI, Bretherton RD, Winton RI (1981) SCOOT – A traffic responsive method of coordinating signals. TRRL Report No LR 1014. Transportation and Road Research Laboratory, Crowthorne
12. Gartner NH (1983) OPAC: A demand-responsive strategy for traffic signal control. Transp Res Rec 906:75–81
13. Bellman R, Dreyfus S (1962) Applied dynamic programming. Princeton University Press, Princeton
14. Grafton RB, Newell GF (1967) Optimal policies for the control of an undersaturated intersection. In: Edie LC, Herman R, Rothery R (eds) Vehicular traffic science. Elsevier, New York, pp 239–257
15. Gartner NH, Pooran FJ, Andrews CM (2002) Optimized policy for adaptive control strategy in real-time adaptive control systems: Implementation and field testing. Transp Res Rec 1811. Transportation Research Board, Washington DC, pp 148–156
16. Gartner NH, Gershwin SB, Little JDC, Ross P (1980) Pilot study of computer-based urban traffic management. Transp Res 14(1/2):203–217
17. Gershwin SB, Ross P, Gartner NH, Little JDC (1978) Optimization of large traffic systems. Transp Res Rec 682:8–15
18. Tan HN, Gershwin SB, Athans M (1979) Hybrid optimization in urban traffic networks. Report No DOT-TSC-RSPA-79-7. Transportation Systems Center, US Dept of Transportation
19. Gartner NH, Stamatiadis C (1998) Integration of dynamic traffic assignment with real-time traffic adaptive signal control. Transp Res Rec 1644:150–156
20. Smith MJ (1980) A local traffic control policy which automatically maximizes the overall travel capacity of an urban network. Traffic Eng Control June 1980:298–302
21. Ghali MO, Smith MJ (1994) Comparisons of the performances of three responsive traffic control policies, taking drivers' day-to-day route choices into account. Traffic Eng Control October 1994:555–560
22. Gartner NH, Al-Malik M (1996) A combined model for signal control and route choice in urban networks. Transp Res Rec 1554:27–35
23. Gartner NH, Stamatiadis C, Tarnoff PJ (1995) Development of advanced traffic signal control strategies for ITS: A multi-level design. Transp Res Rec 1494:98–105

Books and Reviews

Gartner NH, Improta G (eds) (1995) Urban traffic networks: Dynamic flow modeling and control. Springer, Berlin
Gordon RL, et al (1996) Traffic control systems handbook. Federal highway administration. Report FHWA-SA-95-032. US Dept. of Transportation, Washington DC
Homburger WS (ed) (1982) Transportation and traffic engineering handbook, 2nd edn. Prentice-Hall, Englewood Cliffs
Kerner BS (2004) The physics of traffic. Springer, Berlin
Organization for economic cooperation and development (1987) Dynamic traffic management in urban and suburban road systems. OECD, Paris
Papageorgiou M (ed) (1991) Concise encyclopedia of traffic & transportation systems. Pergamon Press, Oxford
RiLSA (2002) Richtlinien für Lichtsignalanlagen für den Strassenverkehr. Forschungsgesellschaft für Strassen-und Verkehrswesen, Köln

Traffic Prediction of Congested Patterns

Hubert Rehborn[1], Sergey L. Klenov[2]
[1] Group Research GR/ETF, HPC: G021, Daimler AG, Sindelfingen, Germany
[2] Department of Physics, Moscow Institute of Physics and Technology, Dolgoprudny, Russia

Article Outline

Glossary
Definition of the Subject
Introduction
Congested Traffic Pattern Features on Freeways Relevant for Prediction
Congested Patterns in Urban Areas and Their Prediction
Reconstruction of Freeway Congested Traffic Patterns Based on Measured Data
Methods for Traffic Prediction: Spatiotemporal Pattern Analysis Based on Kerner's Three-Phase Traffic Theory Versus Other Approaches
Long-Term Traffic Prediction
Applications of Traffic Prediction
Future Directions
Acknowledgment
Bibliography

Glossary

Measurements of traffic flow variables Traffic variables like flow rate (normally given as number of vehicles per hour), vehicle speed and occupancy (percentage of time in which a road location (i. e., normally a detector) is occupied by a vehicle in a certain time interval) are measured with local detectors (single or double induction loop detectors) in the road network. The locally measured traffic variables are averaged and transferred from the roadside detectors to online traffic management centers in usually fixed time intervals (commonly 30 s in the USA, 60 s in Germany). Individual vehicles can measure travel times on road segments, which can be aggregated to travel times.

Spatiotemporal congested traffic pattern Spatiotemporal congested traffic patterns are traffic patterns on freeways that emerge and develop in space and time as a result of the onset of congestion in an initially free traffic flow. Spatiotemporal congested traffic patterns occur mostly at freeway bottlenecks associated with on- and off-ramps, roadwork, decreasing freeway lane numbers, etc.

Kerner's three-phase traffic theory In Kerner's three-phase traffic theory, empirical (measured) features of spatiotemporal congested traffic patterns are explained. Three traffic phases are identified and the features of spatiotemporal dynamics within these traffic phases, as well as the phase transitions between them are examined. These three traffic phases are (i) free traffic, (ii) synchronized flow and (iii) wide moving jam. The synchronized flow and wide moving jam traffic phases are associated with a congested traffic state. These two traffic phases are defined through the following empirical criteria. The definition of the wide moving jam phase [J]: a wide moving jam is a moving jam that propagates through any bottleneck while maintaining the mean velocity of the downstream jam front. The definition of the synchronized flow phase [S]: the downstream front of a synchronized flow region does not exhibit the above characteristic jam feature; in particular, the downstream front of a synchronized flow is often fixed at the location of the bottleneck.

Onset of congestion Onset of congestion (traffic breakdown) is defined as congested traffic pattern emergence in an initial free flow. The onset of congestion is observed mostly at freeway bottlenecks. In Kerner's three-phase traffic theory, the onset of congestion is explained by a phase transition from free flow to synchronized flow at a bottleneck ($F \rightarrow S$ transition for short).

Recurring and non-recurring congestion The congestion emerging by the routine presence of large numbers of vehicles on roads is called recurring congestion. Unexpected disruptions of traffic caused by like lane blockings, incidents from work zones, weather events etc. are called non-recurring congestion. Spatiotemporal congested patterns are caused by both recurring and non-recurring congestion. After recognition of a congested pattern the further prediction is mainly independent from its original reason. The methods described in this contribution are valid for both kinds of congestion reasons.

Wide moving jam emergence In empirical observations, wide moving jams emerge spontaneously in synchronized flow. As explained by Kerner's three-phase traffic theory, this wide moving jam emergence is associated with a first-order phase transition from synchronized flow to wide moving jam ($S \rightarrow J$ transition for short). Thus, wide moving jams emerge in free flow as a result of a sequence of two phase transitions: Firstly, a phase transition at a certain freeway location from free flow to synchronized flow occurs. Later and usually at another road location this synchronized flow can transform into a spatiotemporal congested pattern with propagating wide moving jams.

Prediction of spatiotemporal congested traffic pattern The function of the prediction approach is to forecast the whole congestion picture in space and time. Parameters of the congested patterns, such as size and form, at specific bottlenecks can provide a "fingerprint" of the road network, because they are typical, recurring and specific to the related bottleneck. These features facilitate the predictability of traffic congestion.

Time series of traffic variables Traffic variables can be archived and combined as daily or hourly time series. For example, a flow rate time series could be used as the traffic demand estimation for vehicles driving into congestion. A time series of travel time for a road network could be used for dynamic route guidance applications. Important parameters for the choice of the most probable time series used for prediction are at least the characteristics of the specific day (e. g., Monday morning rush-hour) and the probability of occurrence of each of the time series of traffic variables for days with the same characteristics.

Short-term prediction with time series A short-term prediction with time series can be performed via the comparison of a historical database of traffic variables and the current traffic data measurements. The choice of the appropriate traffic data curve that is then used for traffic prediction is called "matching". This choice is performed from the historical database finding the closest match to the current traffic data. The best matched time series is not necessarily the one with the highest probability among the different time series in the database. Therefore, matching can increase the prediction quality considerably.

Long-term prediction with time series Long-term prediction is defined to be independent from the current situation, i. e., traffic variables solely from a historic database are used. This is the established method for traffic prediction with larger time horizons, e. g., planning tomorrow's roadwork. In contrast, short term traffic prediction is influenced by the current traffic

situation. Without this influence the most probable time series from the database can be chosen directly to perform a specific traffic variable prediction.

Data mining or physics of traffic – two different approaches to traffic prediction In "data mining" approaches reproducible features of measured traffic data are identified and analyzed through the use of various learning systems like neural networks, regression techniques, time series, wavelet analysis, Bayesian networks, etc. The nature of the traffic phenomena does not play a key role for this reproducible traffic feature learning. Instead, reproducible phenomena of the input data are learned by those methods. In contrast, "physics of traffic" approaches understand, explain and model these reproducible features of measured traffic data. Then features of traffic phenomena are used as the basis for traffic prediction methods.

Definition of the Subject

Empirical congested traffic patterns and their predictions are an interesting field of complex systems because of their complex spatiotemporal behavior. The physics of traffic with its non-linear phenomena, self-organization processes based on individual driver behavior and phase transitions from free to congested traffic states are a great challenge for traffic prediction approaches. The road infrastructure approaches its capacity with severe congestion problems because of the high and increasing mobility demand in many countries of the world: a better understanding of traffic congestion patterns, their emergence and their dissolution as well as their prediction may improve both the complex individual and collective management of traffic networks. Due to advancing traffic data measurement techniques, the variety and complexity of available traffic data is steadily increasing. The amount of available traffic data can be used as archive information for the time series of traffic variables. Traffic predictions in the long term need historical knowledge about the traffic variables. The available historical traffic data varies from single vehicle trajectory data and large amounts of local induction loop detector data up to traffic message databases from service broadcasters. Specifically, the traffic flow rate and/or the travel times are needed for an estimation of future traffic situation. In the long-term, traffic patterns are recurring and regular (like a morning peak traffic hour with commuters). This regularity of traffic makes long-term time series based traffic predictions possible and feasible. On the other hand, a comparison with a current situation is a basis of short-term traffic prediction.

Relevant applications of traffic predictions are, from the collective point of view, the network predictions of traffic states as basic information for traffic management and control as well as the planning of the probable congestion consequences of roadwork. From an individual driver's point of view, traffic prediction is important for efficient route choice and the estimation of arrival times. Additionally, precise congestion information could be used for safety and comfort applications in vehicles like the adaptation of cruise control systems.

Introduction

Every morning, the commuters in urban areas of the world suffer from time delays caused by traffic congestion. The congestion is regular, recurring and permanent at certain locations of the freeway road network. Therefore, from a better understanding of the underlying physical processes in traffic, the wish for precise and reliable traffic prediction arises for multiple traffic management purposes, both collective (e. g., traffic control) or individual (e. g., dynamic route guidance). To reach these goals, empirical congestion patterns must first be analyzed and understood correctly.

Throughout the modern world, traffic congestion has enormous associated costs in terms of delay, fuel consumption and emissions, etc. One way to alleviate this problem is to build new transportation facilities and expand the transportation networks which is very cost intensive and not always possible. Another cost-effective method would be the implementation of real-time and predicted traffic generated and disseminated to the traveling public for both pre-trip planning and en-route usage. Having accurate and up-to-date traffic information, travelers would be able to have the opportunity to plan their trip and adjust their routes accordingly.

Infrastructure planning needs information on the traffic demand for its long-term construction processes. Additionally, the planning of roadwork must ensure efficiency, minimizing the effect on the current traffic demand while also considering requirements regarding costs and working necessities. The individual driver is interested in long-term traffic information for his pre-trip planning which can range from tomorrow's trip to work to his next holiday journey further in the future. Long-term time-series of traffic based on historical data are an opportunity to fulfil such individual or collective wishes for forecasted traffic information. The goal of the prediction approach is the precise prediction of a traffic variable like flow rate or travel time. Next, the probability of traffic breakdown, i. e., the possible onset of congestion at a certain location in the

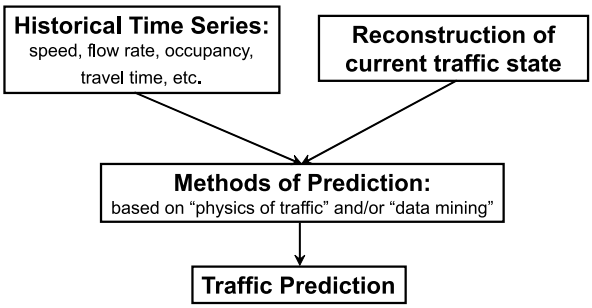

Traffic Prediction of Congested Patterns, Figure 1
Possible qualitative scheme of traffic prediction

road network, should be predicted. After this congestion emergence, the spatiotemporal behavior of the congested region must be forecasted.

Figure 1 illustrates a possible qualitative scheme to perform a traffic prediction: historical time series of traffic variables and the reconstruction of the current traffic state are the "input data" for traffic prediction methods including "matching", i. e., a choice of the most probable traffic pattern, which are used then for the traffic prediction of traffic variables.

There are a huge number of various approaches to reconstruction of current traffic state and short-term prediction (Fig. 1). Related approaches can be subdivided into two main categories, distinguished by their principal concept: driven by "physics of traffic" or "data mining" (Fig. 2). Some approaches, as presented in this article, combine both "physics of traffic" and "data mining".

In the article, the data mining approaches are illustrated with regards to their prediction capability; in particular, we discuss the following methods for traffic prediction: "statistical regression", "neural networks", "wavelet techniques" and "cluster algorithms". It should be noted, that data mining technologies may be used in addition for the generation of databases of historical time series.

The approaches associated with "physics of traffic" are devoted to the understanding of empirical (measured) spatiotemporal features of traffic as well as to modeling of these features. Then these features are used for traffic prediction. Although there were a huge number of studies of measured data on freeways (see classic works, e. g., by Edie and Foote [36], Treiterer [143] and Koshi et al. [86] and references in the book [69]), only recently Kerner solved the puzzle of empirical spatiotemporal features of freeway traffic congestion [64,67,69,72,75,76,77], ► Traffic Congestion, Modeling Approaches to. Kerner's three-phase traffic theory has been developed to explain the puzzle of freeway measured data.

It turns out that earlier known modeling approaches to freeway traffic congestion (e. g., [24,25,26,51,98,102,132,142]), which are based on the fundamental diagram hypothesis (Fig. 2), cannot explain and predict the empirical probabilistic features of the onset of congestion and resulting congested patterns occurring on freeways that are required for traffic prediction (see [70,72], ► Traffic Congestion, Modeling Approaches to for more detail). These traffic flow models are standard ones for validation of freeway traffic prediction. Thus the related simulations or other applications of these models cannot be used for a reliable freeway traffic prediction. In other words, at time Kerner's three-phase traffic theory is the only traffic theory, which can explain the observed spatiotemporal phenomena of congested freeway traffic. For this reason, we will use in this article often Kerner's three-phase traffic theory [69] as the basic approach associated with "physics of traffic" application for traffic prediction on freeways.

Empirical features of freeway traffic that make the prediction difficult and complex are the probabilistic nature of the traffic breakdown and the probabilistic features of the spatiotemporal patterns of freeway traffic [64,67,69,72,75,76,77], ► Traffic Congestion, Modeling Approaches to. As one result to be explained later, the travel times could be predicted based on measured and archived travel times in principle precisely only in free traffic situations which are the less interesting from the driver perspective because his personal trip would not be influenced by congestion. Possible applications of freeway traffic predictions are the use in navigation systems, for safety and comfort functions in vehicles and on the infrastructure side for the traffic assignment in networks and the choice of appropriate traffic control strategies.

The article starts with a description of the variety of congested traffic pattern both on freeways and in urban areas with roads interrupted by intersections (Sects. "Congested Traffic Pattern Features on Freeways Relevant for Prediction" and "Congested Patterns in Urban Areas and Their Prediction", respectively). FOTO and ASDA models for reconstruction, tracking, and prediction of spatiotemporal congested pattern dynamics are the subject of Sect. "Reconstruction of Freeway Congested Traffic Patterns Based on Measured Data". These methods are also used in Sect. "Methods for Traffic Prediction: Spatiotemporal Pattern Analysis Based on Kerner's Three-Phase Traffic Theory Versus Other Approaches" in which various approaches to reconstruction of current traffic state and short-term prediction (Fig. 2) are discussed. A long-time traffic prediction based on historical time serious derived from measured traffic data are considered in Sect. "Long-Term Traffic Prediction". In Sect. "Applica-

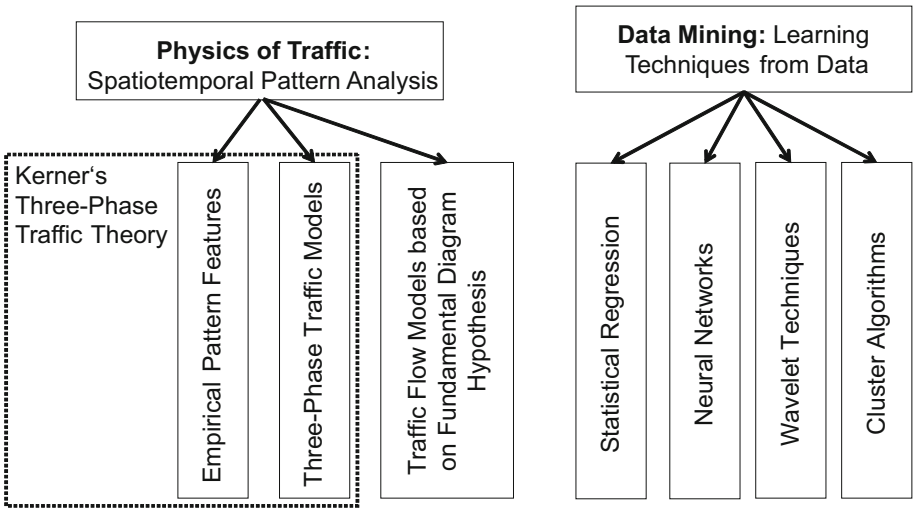

Traffic Prediction of Congested Patterns, Figure 2
General overview on approaches to reconstruction of current traffic state and short-term prediction

tions of Traffic Prediction", currently known applications of traffic prediction are explained. Finally, we discuss some future directions and remaining tasks in the field followed by an extended bibliography.

Congested Traffic Pattern Features on Freeways Relevant for Prediction

After extensive traffic data analyses of available stationary measurements spanning several years Kerner discovered that in congested freeway traffic two different traffic phases must be differentiated: "synchronized flow" and "wide moving jam" [64,67,68,69,75,76,77]. Experimental features of traffic on freeways were investigated and interpreted intensively in the late 1990ies ([67,75,76,77] and references in [69]). Empirical macroscopic spatiotemporal objective criteria for the phases as important elements of Kerner's three-phase traffic theory (see the book [69] for comprehensive explanations) are as follows:

(i) A wide moving jam is a moving jam that maintains the mean velocity of the downstream jam front, even when the jam propagates through any other traffic states or freeway bottlenecks.
(ii) In contrast, the downstream front of the "synchronized flow" phase is often fixed at a freeway bottleneck. Within the front, vehicles accelerate from lower speeds in synchronized flow to higher speeds in free flow.

It must be noted that the observation of speed synchronization in congested traffic is no criterion for the phase differentiation as well as some other relationships and features of congested traffic measured at a freeway location (e. g., in the flow-density plane). The clear differentiation between the synchronized flow and wide moving jam phases can be made on these above objective criteria (i) and (ii) *only*. These objective criteria define the features of the two congested traffic phases [69]. Traffic phase definition in congested traffic cannot be made definitely and consistently via differences in vehicle speeds, densities or vehicle flow rates or flow-density criteria in measured traffic variables.

As a consequence, the identification of the location and of the parameters of both traffic phases on the freeway is the basis for successful traffic prediction: the objective criteria have to be used to identify the traffic phases. After this identification, both traffic phases have predictable features and interdependences among each other which will be used in the prediction method. Some elements and issues from Kerner's three-phase traffic theory will introduce those predictable features. The aspects described here focus on the relevant features for traffic prediction.

Figure 3a illustrates a vehicle speed profile over time and location with measured traffic data and shows the three traffic phases. A wide moving jam propagates as a "low speed valley" through the freeway stretch. In contrast, a second speed valley is almost fixed at the bottleneck location in Fig. 3a: this congested traffic phase belongs to the synchronized flow phase.

The fundamental hypothesis of Kerner's three-phase traffic theory [69] is as follows: steady states of synchronized flow cover a two-dimensional (2D) region in

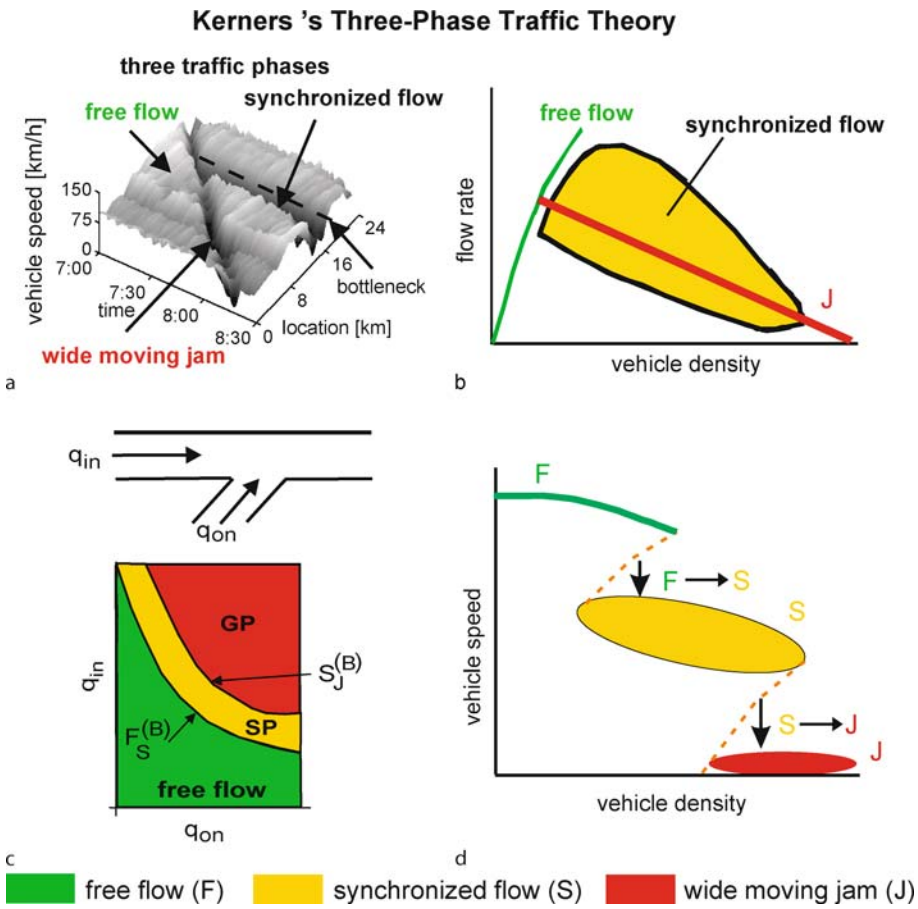

Traffic Prediction of Congested Patterns, Figure 3

Explanations to Kerner's three-phase traffic theory: a phase definition in measured data, b two-dimensional region of hypothetical steady states of synchronized flow, c simplified traffic phase diagram at an on-ramp bottleneck and d qualitative representation of the double Z-characteristic for phase transitions between free flow, synchronized flow and wide moving jams (with low speed traffic states within a wide moving jam [69])

the flow-density plane (so-called "Kerner's 2D-states") (Fig. 3b). This means that in contrast with the fundamental hypothesis of all earlier traffic flow theories there is no fundamental diagram for steady states of traffic flow in this theory. The red line J (so-called "Kerner's line J", [96]) represents the velocity of the downstream front of a wide moving jam (Fig. 3b).

Some other hypotheses and results of the theory: at any density of free flow at which both an $F \to S$ transition and an $F \to J$ transition are possible, a critical speed (density) disturbance needed for the $F \to S$ transition is considerably smaller than that needed for the $F \to J$ transition. This should explain why empirical traffic breakdown is associated with a $F \to S$ transition. This explains also why despite of free flow metastability with respect to a $F \to J$ transition empirically observed, this $F \to J$ transition does not occur spontaneously in empirical observations.

One of the effects of the fundamental hypothesis of the three-phase traffic theory is as follows. When a vehicle approaches the preceding vehicle that moves with a slower time-independent speed and cannot pass it, the vehicle adapts its own speed to the speed of the preceding vehicle at a space gap, which is only one possible space gap from the infinite amount of space gaps within the above 2D-region at this speed (Fig. 3b): there is no desired (or optimal) space gap at a time independent, i. e., constant speed in the car-following behavior in this theory. This feature is called the speed adaptation effect in synchronized flow.

Traffic breakdown ($F \to S$ transition) is explained in Kerner's three-phase traffic theory as a competition be-

tween the speed adaptation effect and a well-known over-acceleration effect, i. e., a tendency of a vehicle to accelerate to a free flow speed when it seems to be possible.

All steady states on the line J (an infinite number) are threshold states for wide moving jam existence and emergence. The line J separates all steady states of synchronized flow in the flow-density plane into two different classes: All states below the line J are stable with respect to wide moving jam existence and emergence. All states on and above the line J are metastable states with respect to wide moving jam existence and emergence (Fig. 3b).

The simplified phase diagram at an on-ramp bottleneck illustrates the phase transitions. While increasing q_{on} (the flow rate to the on-ramp) at a same level of q_{in} (the flow rate on the main road) in free flow (Fig. 3c) the following occurs: the onset of congestion in an initial free flow at a freeway bottleneck is associated with a phase transition from free flow to synchronized flow. Firstly, synchronized flow emerges ($F \to S$ transition) at the bottleneck. The boundary $F_S^{(B)}$ is associated with this $F \to S$ transition. Wide moving jams can emerge spontaneously only in synchronized flow ($F \to S$ transition). The boundary $F_J^{(B)}$ is associated with the $S \to J$ transition (Fig. 3c). Thus, wide moving jams emerge due to a sequence of two phase transitions called $F \to S \to J$ transitions.

A double Z-characteristic in the speed-density plane illustrates the sequence of the $F \to S \to J$ transitions and explains these phase transitions [69]. The first Z-characteristic related to higher speeds explains the transition (arrow labeled $F \to S$ in Fig. 3d) and includes the traffic states F and S. The second Z-characteristic, which is related to lower speeds, explains the $S \to J$ transition (arrow labeled $S \to J$ in Fig. 3d) and includes the traffic states S and J. Dashed curves in the double Z-characteristic are associated with the so-called "critical speeds" within local disturbances in traffic flow required for a phase transition (see for more detailed the book [69]). It must be noted that the double Z-characteristic of the three-phase traffic theory (Fig. 3d) can usually be correctly analyzed only after spatiotemporal features of the phase transitions and congested patterns have been studied. This is because $F \to S$ and $S \to J$ transitions occur at different locations.

Considering the case of an isolated effectual bottleneck (a bottleneck, which is far enough from other bottlenecks), the following two main congested pattern types can emerge [69]:

1. "Synchronized Pattern" (SP): this consists exclusively of an area of synchronized flow upstream of the effectual bottleneck, i. e., no wide moving jams develop

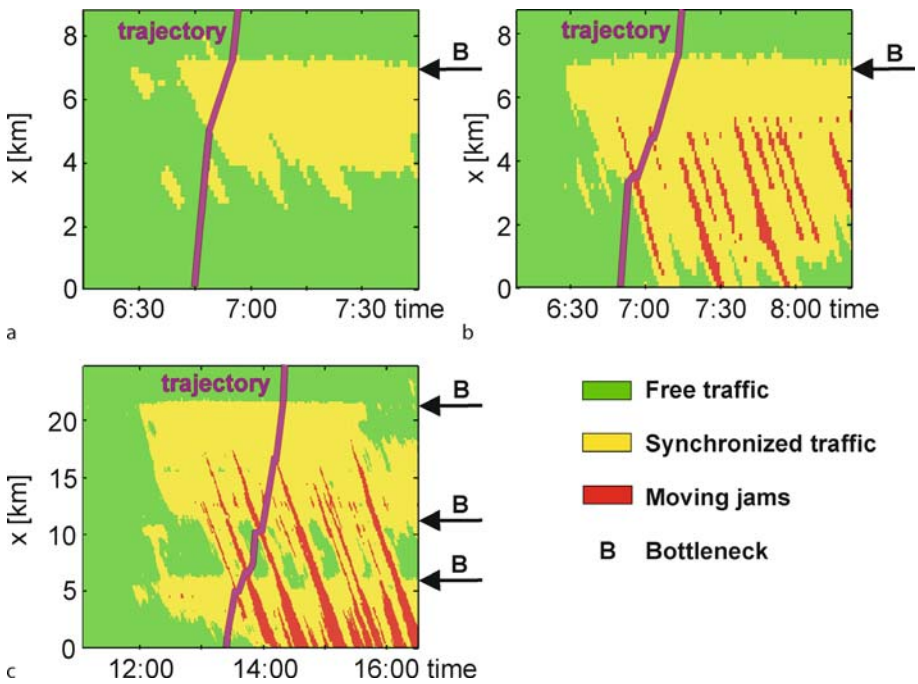

Traffic Prediction of Congested Patterns, Figure 4

Spatiotemporal traffic pattern classification: a empirical SP (Synchronized Pattern), b empirical GP (General Pattern), c empirical EP (Expanded Pattern) [69]. Solid lines labeled "trajectory" illustrate some vehicle trajectories through these congested patterns

in this pattern (Fig. 4a). In its variants, SP could be localized at the bottleneck ("LSP"), moving upstream ("MSP") or increasing over time ("WSP").

2. "General Pattern" (GP): this consists of an area of synchronized flow upstream from the effectual bottleneck and wide moving jams, which develop in this synchronized flow spontaneously. GP covers thus both phases of congested traffic. It is the most frequently arising pattern at isolated bottlenecks on freeways (Fig. 4b).

If several effectual bottlenecks are closely neighboring, an expanded traffic pattern can develop (EP: Expanded Pattern) with areas of synchronized flow affecting several adjacent effectual bottlenecks (Fig. 4c). For each effectual bottleneck or for each set of several neighboring effectual bottlenecks the spatiotemporal traffic patterns possess predictable, i. e., characteristic, and recurring characteristics, e. g., the most frequently arising traffic patterns or the average expansion of synchronized flow in a GP or EP. These characteristics can be almost alike on different days and years. The existence of traffic patterns also remain preserved within a large range of the flow rate. The prediction approach based on three-phase traffic theory uses this pattern definition and pattern features for the traffic pattern database concept. The pattern classification scheme is the key element of the related prediction approach. One of the main differences to other known approaches is the definition and use of different traffic phases including their specific features and not based on aggregated information from freeways (e. g., "traffic states") or drivers (e. g., "travel times").

Congested Patterns in Urban Areas and Their Prediction

In contrast to freeway sections, in urban areas a large part of the network is built up by short road sections, which are separated by each other through traffic regulations at intersections (signalized or non-signalized). The emergence and propagation of spatial congested traffic patterns is, therefore, hindered and limited by these intersections rather than self-organized spatiotemporal congested traffic patterns associated with driver interaction effects can emerge. In other words, the urban traffic prediction is mostly determined by traffic light signals at road intersections: traffic flow is interrupted during the red phase.

The recognition, tracking and prediction of those queues built at road intersections and queue parameters as number of vehicles in the queue, waiting time of the vehicles, etc. are the key elements of an urban prediction method for short road sections. Based on parameters like red and green cycle times, vehicle speeds, inflows into the network and the section length the prediction approaches calculate a traffic state on the urban road sections. Because driver interaction effects are not important for urban traffic with *short* road sections, most models and tools of the fundamental diagram approach, which cannot be applied for an adequate prediction of freeway traffic congestion, can nevertheless be used for such an urban traffic (e. g., [41,52,101,105,123,157]); examples for tools useful for *short* road sections: VISSIM (e. g., [39,153]), PARAMICS (e. g., [13]), AIMSUN (e. g., [5,44]), INTEGRATION (e. g., [122], DynaMIT e. g., [8]), OPAC [42], TRANSYT [130], SCOOT [57]. A review of optimization and control of urban traffic networks, in particular based on OPAC, TRANSYT and SCOOT appears in this Encyclopedia ► Traffic Networks, Optimization and Control of Urban.

It must be noted that for traffic management purposes the results of the prediction should be much faster than real-time, i. e., a 60 min prediction should be calculated several hundred times in less than a system cycle of, e. g., 5 min: then the predictions can be calculated fast enough to support the choice of the optimal traffic management strategy by predicting the effects of different possible control strategies. For that purpose, Kerner introduced a model UTA (Urban Traffic Analysis) [65] (see section 22.4 in [69]) that has been further developed and implemented for large networks by Kerner, Klenov and Aleksic [74]. Goal of the model UTA is a traffic management oriented macroscopic fast network prediction in a time interval which is much smaller than the prediction horizon and calculates thousands times faster than microscopic models like the above mentioned ones.

In the model UTA, for dynamical traffic prediction in urban areas [65,72], some effective macroscopic traffic variables (the flow rates, the waiting queue's lengths, the number of vehicles in the net's links, etc.) are considered. For example, the effective number of vehicles in a queue $N_q(t)$ is connected with the actual number of vehicles in a queue $N_q^{(\mathrm{actual})}$ corresponding to the formula:

$$N_q(t) = \frac{1}{T} \int_{t-T/2}^{t+T/2} N_q^{(\mathrm{actual})}(t')\mathrm{d}t' \tag{1}$$

All other traffic variables are averaged over time corresponding to the above formula. Thus, the effective traffic variables are related to those changes in traffic where the time scale is considerably longer than the duration of the period of light signal T. This approximation allows us fore-

casting of traffic variables in large urban traffic nets much easier and quicker than a calculation of the actual traffic variables, which can be very complex functions of time due to the interruption of traffic during the red phases at intersections of the net. In particular, in contrast to the actual flow rates out from the links of the net $q_{\text{out}}^{(\text{actual})}$, the effective flow rates q_{out} are continuous functions of time because they are not interrupted during the red phase (Fig. 5). The durations of the red and green phases T_{R}, T_{G} influences the related effective flow rates.

Dashed curves in Fig. 6 illustrate an urban definition of congestion at an intersection with signalization: if during the cycle time T within the green phase T_{G} the waiting queue of the number of vehicles $N_q^{(\text{actual})}$ dissolves, the state of the related section will be undersaturated. If after the green phase T_{G} there are still vehicles in the queue, the traffic state is oversaturated (i. e., it could be called "congested"). Solid curves in Fig. 6 illustrate the effective number of vehicles in the queue $N_q(t)$ used in the model UTA.

The model UTA [65] calculates the future values for the lengths of a queue L_q on a link and the related number of vehicles in the queue N_q, the number of vehicles inside the link N, the travel time across the link, the delay time in the queue, the flow rate to the queue $q_{\text{in},q}$ and all other macroscopic traffic variables in the net as functions of time.

As mentioned, the calculation time of the model UTA can be about several thousand times shorter than the one of microscopic models. It is linked to two reasons: (i) instead of a calculation of the behavior of each vehicle only the numbers of vehicles on the links are considered in UTA; (ii) the time scale considered in UTA is considerably higher than the one in microscopic models because

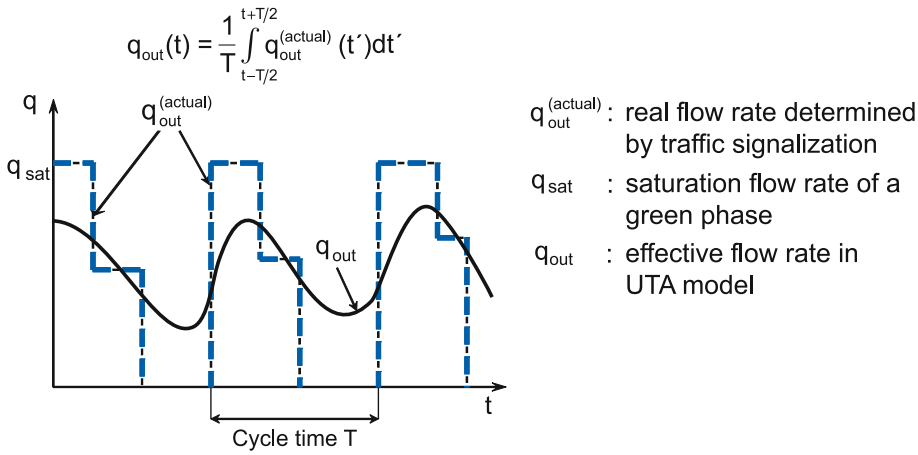

Traffic Prediction of Congested Patterns, Figure 5
Qualitative explanation of effective flow rate q_{out} in UTA model [65]

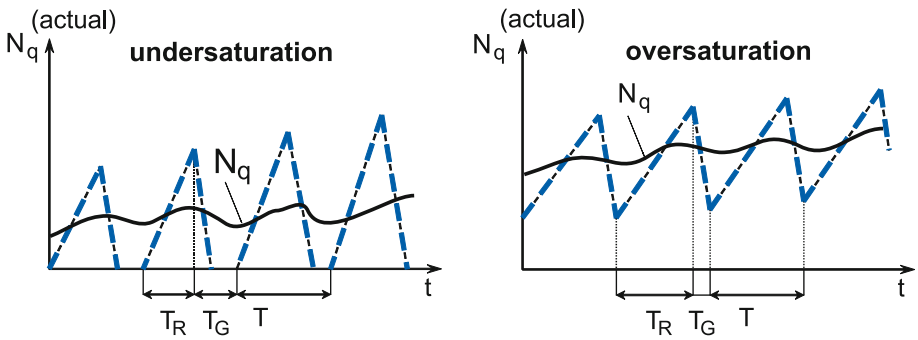

Traffic Prediction of Congested Patterns, Figure 6
Definition of under- and over-saturation at a traffic light signal on an urban road section ("urban congestion") [65]. Dashed curve – real number of vehicles in a queue. Solid curve – effective number of vehicles in the queue used in the model UTA

UTA goes well beyond the duration of the period of light signals.

The required input information of the UTA model contains [65,69]:

- Description of the network (lanes, intersections, location of light signals),
- Traffic light signal programs (cycle time, red and green phases),
- Historical time series of saturation flow rates (per section),
- Percentages of long vehicles,
- Turning ratios at each intersection,
- Incoming flows at the boundary of the network.

Then model UTA produces for future time intervals of the prediction horizon:

- Predicted length of queues (number of vehicles in the queue),
- Travel times associated with the links of the network and with any route through the network,
- Waiting times within queues.

In contrast to other models, UTA require only parameters which can be measured directly (no "origin-destination (OD)-matrices" are necessary) The macroscopic UTA model is computational efficient because of the concept of effective continuous in space and time variables.

In a laboratory simulation environment developed, large road networks (e. g., Tokyo) have been calculated some thousand times faster than any microscopic simulation tools known for us [74]. This feature allows proving various control strategies for vehicle route guidance, traffic management and control and traffic assignment before their real execution. An empirical example with measured and historical traffic data as well as real light signal control has been developed in the city of Stuttgart area (Fig. 7 shows a small part of this area).

Reconstruction of Freeway Congested Traffic Patterns Based on Measured Data

FOTO and ASDA Models

One of the engineering applications of this three-phase traffic theory are the FOTO and ASDA models proposed by Kerner and further developed by Rehborn, Aleksic and Haug for recognition and tracking of spatiotemporal congested traffic patterns occurring at freeway bottlenecks [66,69,78,79,81], ► Traffic Congestion, Modeling Approaches to. The FOTO and ASDA models recognize and track the synchronized flow and wide moving jam traffic phases in congested traffic based on traffic measurements.

Within the FOTO model, traffic phase identification is performed first. Secondly, FOTO performs the recognition of the locations of the upstream and downstream fronts

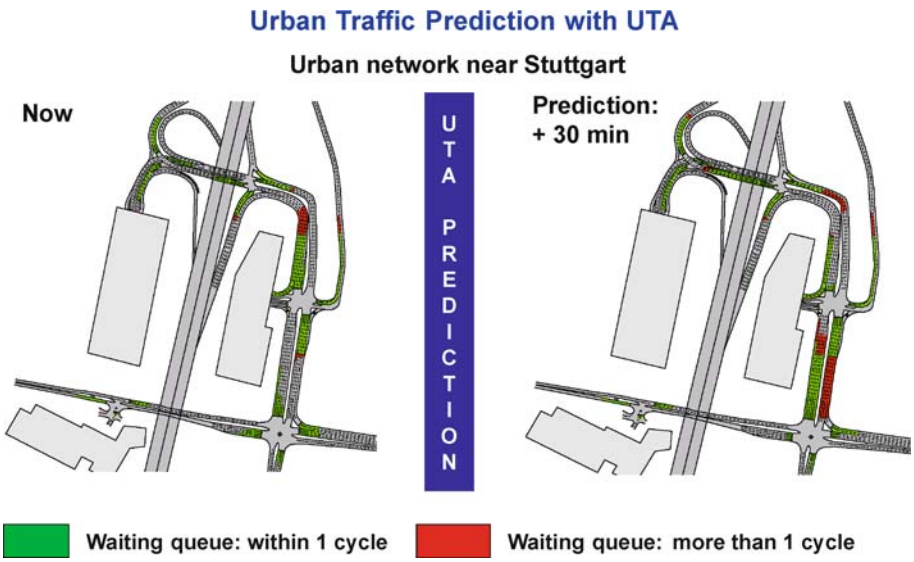

Traffic Prediction of Congested Patterns, Figure 7
Empirical urban prediction example with UTA [74] for a small network part near Stuttgart: current state (*left*) and predicted in 30 min (*right*)

$x_{\text{up}}^{(\text{syn})}$, $x_{\text{down}}^{(\text{syn})}$ of synchronized flow. While the downstream $x_{\text{down}}^{(\text{syn})}$ front of synchronized flow is fixed at the bottleneck, the position of the upstream front $x_{\text{up}}^{(\text{syn})}$ with respect to the bottleneck is calculated by a "cumulative flow" approach with the inflowing vehicles $q_0(t)$[vehicles/h] and the vehicles escaping downstream from the synchronized flow $q_n(t)$[vehicles/h]

$$x_{\text{up}}^{(\text{syn})}(t) = \mu \frac{1}{n} \int_{t_{\text{syn}}}^{t} (q_n(t) - q_0(t))\mathrm{d}t, \quad t \geq t_{\text{syn}} \tag{2}$$

with $20 < \mu < 40$ [m/vehicles] as parameter n as the number of freeway lanes and t_{syn} as time at which the syncronized flow is first registered.

Then, the ASDA model recognizes the upstream and downstream fronts $x_{\text{up}}^{(\text{jam})}$, $x_{\text{down}}^{(\text{jam})}$ of wide moving jams. The positions of these fronts are calculated by:

$$x_{\text{up}}^{(\text{jam})}(t) = \int_{t_0}^{t} v_{\text{gl}}(t)\mathrm{d}t \approx -\int_{t_0}^{t} \frac{q_0(t) - q_{\min}}{\rho_{\max} - (q_0(t)/v_0(t))}\mathrm{d}t, \quad t \geq t_0 \tag{3}$$

$$x_{\text{down}}^{(\text{jam})}(t) = \int_{t_1}^{t} v_{\text{gr}}(t)\mathrm{d}t \approx -\int_{t_1}^{t} \frac{q_n(t) - q_{\min}}{\rho_{\max} - (q_n(t)/v_n(t))}\mathrm{d}t, \quad t \geq t_1 \tag{4}$$

with $q_0(t)$, $q_n(t)$, as the upstream (downstream) flow rates, $v_0(t)$, $v_n(t)$ as the upstream (downstream) vehicle speeds, $q_{\min}$ as the flow rate in the wide moving jam (normally set to zero) and the parameter $\rho_{\max}$ as the density inside the wide moving jam, t_0 and t_1 are time moments at which upstream and downstream fronts are first registered at the downstream position, respectively. The densities $\rho_{\max}$ and $\rho_{\min} = q_n(t)/v_n(t)$ in case when free flow is formed downstream of the jam are the two distinct characteristic densities that determine the velocity of the downstream front of a wide moving jam ► Traffic Congestion, Modeling Approaches to.

The formulas (3), (4) are associated with the classic Stokes formula for the shockwave velocity v_s as follows:

$$v_s = \frac{q_2 - q_1}{\rho_2 - \rho_1} \tag{5}$$

with q_1, ρ_1 as the related flow rate and density downstream of the shockwave and q_2, ρ_2 as the related flow rate and density upstream of the shockwave. It must be stressed that this Stokes formula Eq. (5) and the related ASDA formulas Eq. (3) and (4) are fundamentally different from the well-known formula for shockwaves in the classic Lighthill–Whitham–Richards (LWR)-theory [152]:

$$v_s^{(\text{LWR})} = \frac{Q(\rho_2) - Q(\rho_1)}{\rho_2 - \rho_1} \tag{6}$$

where $Q(\rho_2)$ and $Q(\rho_1)$ are flow rates on the fundamental diagram that gives a single correspondence between the density and the flow rate. In contrast, in Eq. (5) and ASDA formulas Eq. (3) and (4) there is no given correspondence between the density ρ_2 and the flow rate q_2, as well as between the density ρ_1 and the flow rate q_1. ASDA solves the problem to find $q_2 = q_0$ in Eq. (3) through the use of measured data from an upstream detector. In turn, $q_1 = q_n$ in Eq. (4) is found through the use of measured data at the downstream detector. For each time interval (e. g., 1 min) based on measured data, ASDA formulas Eqs. (3) and (4) find the density as a ratio of the flow rate and the speed. As shown in empirical observations in synchronized flow there is no single relationship between flow rate and density. Instead there are an infinite number of such relations covering a two-dimensional region in the flow-density plane ("2D states" are associated with Kerner's hypothesis). In other words, in contrast to the LWR-theory, there is no fundamental diagram of congested traffic that gives a single flow rate for a given density or vice versa a single density for a given flow rate. For this reason, as we mentioned in the fundamental hypotheses of Kerner's three-phase traffic theory, there is a 2D-region of steady states in synchronized flow. Both the "cumu-

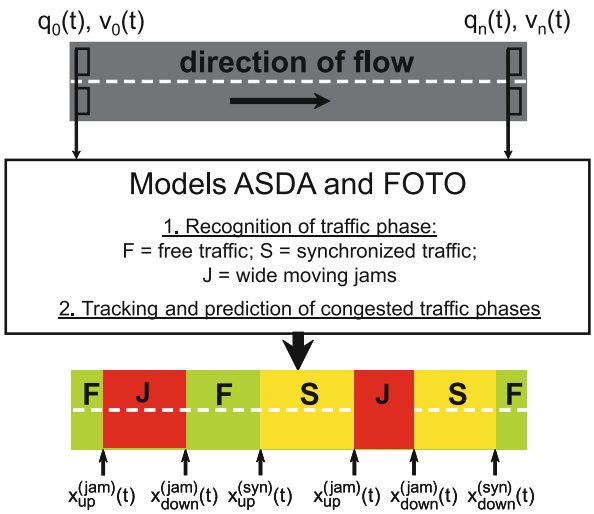

Traffic Prediction of Congested Patterns, Figure 8
Illustration of the FOTO and ASDA model approach [66,69,78, 79,81], ► Traffic Congestion, Modeling Approaches to

lative flow" approach and Stokes shockwave formula are well-known in traffic science. The main feature of FOTO and ASDA models is that the models identify firstly two different traffic phases in congested traffic (synchronized flow and wide moving jam), and then synchronized flow is tracked with the "cumulative flow" approach, while in contrast wide moving jams are tracked with the Stokes shockwave formula.

The front locations $x_{\mathrm{up}}^{(\mathrm{syn})}$, $x_{\mathrm{down}}^{(\mathrm{syn})}$ and $x_{\mathrm{up}}^{(\mathrm{jam})}$, $x_{\mathrm{down}}^{(\mathrm{jam})}$ define the spatial size and location of the related "synchronized flow" and "wide moving jam" objects, respectively. Finally, the FOTO and ASDA models track these object fronts $x_{\mathrm{up}}^{(\mathrm{jam})}(t)$, $x_{\mathrm{down}}^{(\mathrm{jam})}(t)$, $x_{\mathrm{up}}^{(\mathrm{syn})}(t)$, $x_{\mathrm{down}}^{(\mathrm{syn})}(t)$ in time and space (see Fig. 8). Note, that through the use of the FOTO and ASDA models the tracking of congested traffic objects is also carried out between detectors, i. e., when the object fronts cannot be measured at all. Additionally, the FOTO and ASDA models perform without any validation of model parameters in different environmental and traffic conditions. Model applications are not limited to station-

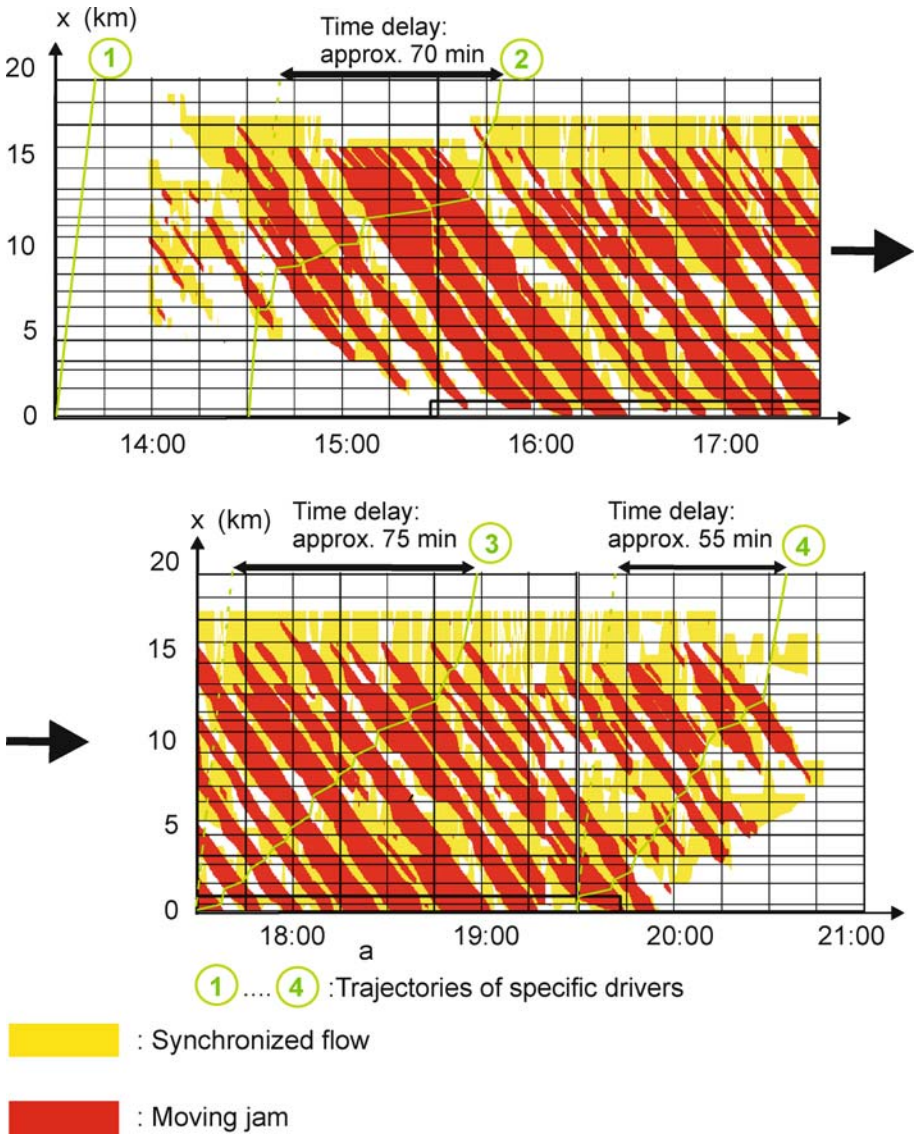

Traffic Prediction of Congested Patterns, Figure 9

Congested traffic pattern reconstructed by FOTO and ASDA models: space-time diagram with vehicle trajectories 1–4 and related delay times from freeway A5-North in Hessen, 14th June 2006

ary detector measurements which could measure the necessary flow rates and vehicle speed directly; the use of more advanced measurement technologies like floating car data (vehicles acting as moving traffic sensors) or phone probes (phones acting as moving traffic sensors) is also possible.

Tracking of spatiotemporal congested patterns at freeway bottlenecks is done for 1200 km of freeway network in the state Hessen (Germany) through the use of the FOTO and ASDA models in an online application in the traffic control center and evaluated comprehensively in a three-month investigation of all measured congestion on the freeway A5 [85]. Since April 2004 measured data of nearly 2500 detectors are automatically analyzed by FOTO and ASDA models [69,79]. The resulting spatiotemporal traffic patterns are illustrated in a space-time diagram showing congested pattern features.

Figure 9 shows the space-time diagram for very large traffic congestion at the A5-North in Hessen reconstructed from detector data by FOTO and ASDA models. On the 14th June, 2006, for about 18 km and nearly seven hours an "Expanded Pattern" exists on the freeway. The numbers 1 to 4 illustrated in Fig. 9 represent the trajectory for individual drivers at different times of the measurement period with different travel times for each driver on the 20 km section: the first one starts at 13:30 h in free traffic and passes in 10 min, the second starts one hour later and needs approx. 80 min (in heavy congested traffic), the third starts at 17:30 h and it takes approx. 85 min to pass this section and the fourth driver starts at 19:30 h and needs approx. 65 min to drive through the congested pattern in the decreasing congestion. The delay times of more than one hour are caused by the reduced vehicle speeds in both the "synchronized flow" and "wide moving jam" traffic phases, respectively.

The traffic control center in Hessen with its model application of FOTO and ASDA models is the laboratory environment for first realization of freeway traffic prediction based on three-phase traffic theory, because of online and archived detector measurements, historical time series and the recognition of the current situation by FOTO and ASDA models.

Detection of Current Traffic States with Floating Cars

Floating cars can send information on their individual drive based on pre-defined protocols. A specific onboard-algorithm interprets the vehicle data and decides if an event criterion like a slow speed for a certain time interval of a freeway is recognized [6,38]. Then, a part of the vehicle trajectory is communicated to the traffic control center. The vehicle can act as a moving sensor in traffic in addition or even as a replacement to stationary roadside sensors of the infrastructure (e. g., [9]).

The empirical Figs. 10 to 12 show the vehicle data solely and in combination with FOTO and ASDA results from the traffic control center based on stationary detectors (horizontal lines in Figs. 10 to 12) (see [125] for further examples). Several vehicles have sent their messages on this 20 km freeway stretch. The colors of the vehicle trajectories illustrate the different speeds. Solely based on this vehicle information, the blue dotted locations of congested patterns, i. e., a region of synchronized flow and two propagating wide moving jams, would have been assumed and registered in the traffic control center. In Fig. 10 the existence of spatiotemporal traffic patterns in congested traffic has been detected solely based on moving detectors, i. e., moving vehicles. The vehicles detect in this congested situation one traffic phase with a fixed downstream location (blue region between 2–7 km in Fig. 10) and two examples of a propagating traffic phase in congested traffic (blue bars diagonal on the whole freeway stretch in Fig. 10). As a consequence, the vehicles have measured empirically two different traffic phases in congested traffic as it is stated in Kerner's three-phase traffic theory [69].

The number of vehicles, which have sent messages in this example, has been about 20 per hour which correlates to a penetration rate of about 0,5% (estimated 4000 vehicles/h on the two lane freeway). These quantitative results

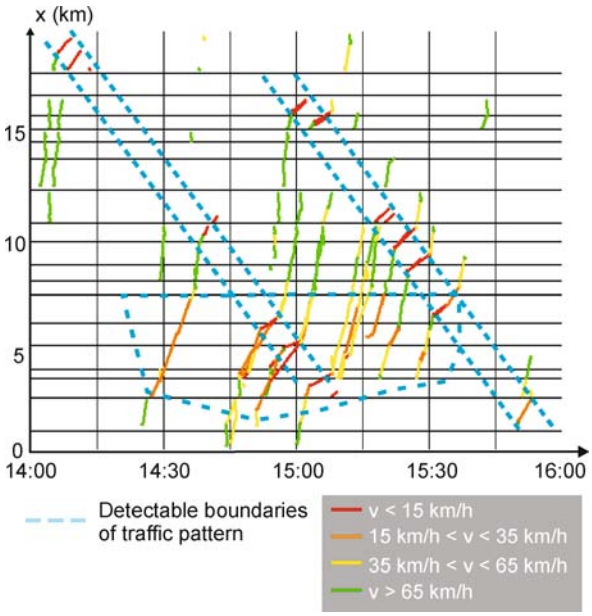

Traffic Prediction of Congested Patterns, Figure 10
Vehicle trajectories (approx. 20 per hour) from equipped vehicle (event messages per vehicle have been sent to the center) (see [125])

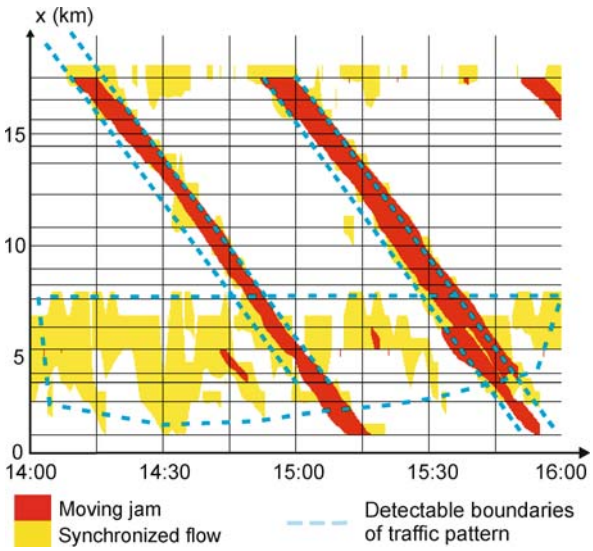

Traffic Prediction of Congested Patterns, Figure 11
Congested pattern reconstruction with FOTO and ASDA models (see [125])

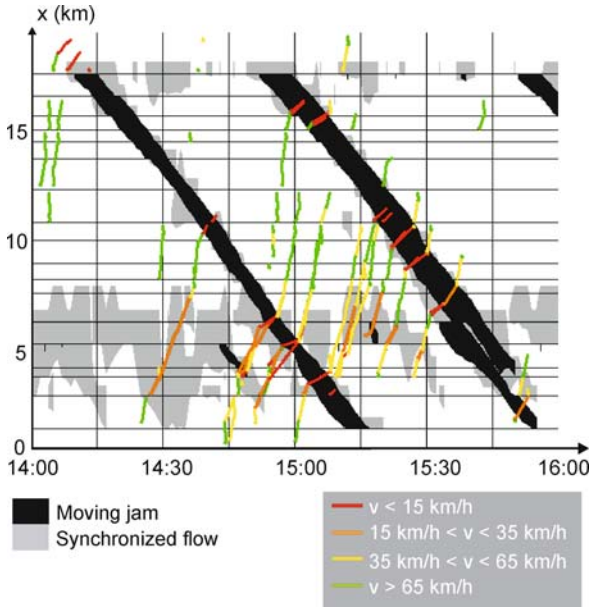

Traffic Prediction of Congested Patterns, Figure 12
Overview of FOTO and ASDA traffic reconstruction and vehicle trajectories (see [125])

give the impression that such a small penetration rate of equipped vehicles could be sufficient for congested traffic state recognition on freeways without any stationary sensors.

The same traffic situation has been reconstructed by FOTO and ASDA models solely based on local detectors (Fig. 11). Qualitatively the same regions of the two propagating wide moving jams and the synchronized flow have been recognized (i. e., the blue dotted regions).

The overview of all available data sources, i. e., local detectors with FOTO and ASDA reconstruction and vehicle trajectories prove the technical feasibility of detecting congested traffic patterns with different measurement techniques, both local and moving like a vehicle. It has to be taken into account that based on only one vehicle it is impossible to reconstruct the whole picture of the congested traffic. To perform such a congested traffic pattern reconstruction inside a vehicle, the vehicle has to have additional information from other drivers, beacons or traffic centers.

Methods for Traffic Prediction: Spatiotemporal Pattern Analysis Based on Kerner's Three-Phase Traffic Theory Versus Other Approaches

As explained in Introduction, from a methodological point of view two main traffic prediction concepts can be distinguished (see Fig. 2):

(i) "data mining": statistical methods based on regression (e. g., [29,139,155,156,160]), wavelet (e. g., [61,140]), or filtering models as well as neural networks (or in hybrid combinations) (e. g., [28,116])
(ii) "physics of traffic": approaches with the use of empirical (measured) spatiotemporal features of traffic congestion as well as traffic flow modeling (e. g., [2,7,22,24,25,26,102,142,150,151]).

We have also stated in Introduction that the first kind of approaches more or less take techniques from other disciplines like mathematics, information technology or artificial intelligence and try to transfer successful concepts into traffic science. The nature of empirical traffic process does not play a key role in the data mining approaches. The "data mining" techniques have the principal problem of determining which data features they should focus on and which are less or not relevant for the accurate prediction.

The "physics of traffic" approaches are based on the nature of traffic phenomena. They begin with the understanding of measurements of traffic variables made in space and time as well as with simulations of driver's behavior, roads with possible bottlenecks, etc. The key element for a successful prediction concept here lies in the correct and reliable understanding of the empirical features of the traffic process and their reproducible characteristics. Furthermore, this understanding of real traffic should be incorporated in a traffic flow model, which should explain and predict the empirical traffic features.

It is well-known that traffic flow is subject to occasional abrupt disturbances (e. g., incidents, congestion) that can

potentially change the underlying dynamics and the stability of the data [34]. For example, when a traffic accident occurs, the average speed may suddenly drop to a new level. Not only discrete events like accidents can alter the underlying dynamics but also the level of congestion. Nevertheless, reliable predictions are especially valuable and needed on occasions when unexpected events occur [92]. In fact, researchers also acknowledge the occurrence of larger prediction errors in their models whenever traffic congestion sets in and dissolves [145,146] and when accidents occur [165].

Many researchers have investigated the field of traffic science (see [23,27,51,52,69,91,98,115,132,142]). However, due to a lack of understanding of the underlying physical processes and the availability of the relevant high-quality traffic data as well as its interpretation, applications of traffic prediction are today still in their infancy. In Kerner's three-phase traffic theory [69] empirical phenomena on freeways have been explained well enough to develop reliable methods for traffic prediction. In this paper, traffic prediction methods are based on recurring and predictable features of traffic congestion as they are empirically proven in the three-phase traffic theory. The combination of empirical features of freeway traffic and its correct interpretation with a pattern database is the key for successful and precise traffic prediction.

Successful traffic prediction should be focused more or less on a congestion prediction. In the industrialized world with heavy vehicle traffic on roads, the congestion is still of much less duration than the free traffic: a very simple traffic prediction citing "free traffic" at all times on all roads would be correct in large parts of the networks. The time basis for traffic prediction evaluation must not cover the whole day, but the duration of the relatively seldom congestion events. The traffic prediction quality should be evaluated and focused on for the congested time periods of the day. Prediction errors of about 30% are reported for those heavily congested situations [59]. The deviations of the "real" congested situation on the road and the predicted situation have to be compared later on in the laboratory to evaluate the prediction approach. It has to be taken into account, that in today's measurement systems the detection of the reality in freeway networks is still imperfect, but those results of a detection system have to be used as a "gold standard" for traffic prediction evaluation.

Definition of "Short-Term" and "Long-Term" Traffic Predictions

Freeway traffic prediction could additionally be differentiated based on the time horizon: the short-term prediction in a horizon of hours can be influenced by the current situation on the road. If larger time horizons have to be forecasted in traffic, the traffic prediction should be based more on historical assumptions and less on the current situation.

In literature, there is no single definition which allows distinguishing between short-term and long-term predictions. Sometimes, traffic prediction on a current day is considered as a short-term and traffic prediction on the next days as long-term. In other publications, short term is a prediction during a chosen time interval, e. g., that is equal or shorter than 60 min [137]. Overviews on short-term prediction approaches can be found in, e. g., [4,90,93,99,135,136,148,163,165]. Short and long term prediction can also be distinguished through the use of current measured data: if in the traffic prediction the current measured data are taken into account, then the prediction is considered as a short-term, because this prediction has a sense for the current day only. Traffic prediction in which only historical time series are used is then long-term prediction. The prediction for the next days is automatically a long-term prediction, because current measured data have usually no influence on prediction for the next days. In general distinguishing between long and short term predictions made in the article is used for clarity of consideration rather than for analysis of the related prediction methods. Indeed, many of the methods discussed below can be used both for short time and long time predictions.

A possible definition of "short-term" could be based on the physical features of the congested patterns found in Kerner's three phase traffic theory and the methodology of the prediction approach. If the prediction of the congested pattern is influenced by the current situation on the freeway, i. e., there is already a congested traffic pattern emerging, this situation is called "short-term prediction". If there is no use of the current traffic situation in the traffic pattern prediction method, the traffic prediction will be called "long-term prediction", because it is performed on any kind of historical assumption of breakdown probabilities or traffic pattern correlations and not the current situation. With this definition, a short-term prediction in contradiction to a solely time-based definition as in [137] could last several hours (e. g., for more than five hours in Fig. 9 after the pattern emergence).

Traffic Flow and Travel Time Prediction

Many of the proposed forecasting models are focused on the traffic flow (e. g., [1,17,18,33,46,48,61,89,95,106,113,120,134,135,136,137,138,154,155,156]). Furthermore, most of these approaches predict the traffic flow condi-

tions at a location given previous observations at that point only. This neglects the spatiotemporal behavior of traffic with upstream moving structures. As proposed by [155], an efficient forecasting approach should involve the use of traffic flow data from upstream detectors to improve the forecasting at downstream locations. The traffic is flowing downstream, but the congestion is propagating upstream, as will be explained later in more detail; therefore the flow rate upstream is influenced by the congestion emergence downstream which is very relevant for successful predictions. Although travel times have been identified as an important parameter for the real-time performance of transportation management systems, most elder studies have been focused on predicting traffic flows. This is mainly because most traffic surveillance systems, especially in the USA, use single inductive loop detectors. These have point-measured information regarding traffic flow and occupancy. For many applications, the travel time is the most useful prediction variable and therefore much research effort has been dedicated to predicting travel times [18,21,28,30,31,37,56,63,83,87,88,90,93,97,103,104,106,108,110,112,118,119,127,129,133,145,146,166]. Those travel times are very relevant as static input data for trip planning, i. e., the reliable dynamic adaptation of navigation systems [19].

Statistical Approaches and Time Series

Other known methods [108] use approaches from mathematical statistics to "learn" coefficients of traffic prediction in the form of correlations in time and space. These ideas try to learn implicitly spatiotemporal features of the traffic process and try to find correlations between heterogeneous traffic data for time series prediction. For example, the velocity of the propagating wide moving jam of 15 km/h against the direction of the flow as characteristic parameter of this traffic phase [69] can be "learned" statistically: such a wide moving jam registered downstream could be predicted with high probability to be 3 km upstream in about 12 min.

Another widely used approach is the use of time series of traffic variables (e. g., [33,120,144,149]). Average flow rates and speeds vary over time, but are regular and recurring for the same network location. Therefore, clustering methods based on time series similarities are used for traffic prediction. Such average flow rates can be useful for an estimation of the expected congestion due to tomorrow's roadwork on a freeway section. Here the average flow rate is taken as traffic demand estimation. One of the principal problems of these clustering methods is the fact that the congested situations are part of the measurements which are used for the traffic time series. Therefore, if congestion is part of a flow rate time series at a specific location or section, the demand is underestimated because no vehicles have passed the detector but have had to wait in the congested area further upstream: for prediction the expected flow rate is systematically underestimated.

Several authors have used statistical approaches (e. g., [20,53,100,106,107,108,116,120,133,156]) and time series for traffic prediction. The spikes of travel time curves caused by (sudden) traffic breakdowns are a problem for these kinds of linear approaches, because they are underrepresented in the data set and the regression filters and smoothes the traffic breakdowns. Finding a proper prediction model defining the prediction variable and the number of necessary data cycles is a question of finding adequate parameters. Therefore, the success of regression techniques for short-term congestion predictions seems questionable.

Neural Networks

One widely used "data driven" approach was the use of Artificial Neural Networks (ANN) for traffic prediction. As a relatively new mathematical model, ANN were first introduced in the 1940s (e. g., [47,49,131]). The basic idea of the neural network is to emulate the human brain, which consists of a very large number of inter-connected neurons. Each of the neurons ("cells") have similar behavior and there is the possibility of having several layers of neurons in an ANN. ANNs try to adopt features of a given traffic data set in a learning phase and then perform a traffic prediction based on a given data set (i. e., the current traffic situation). Many researchers have used ANN for data completion or short-term forecasting of traffic variables which very often focus on travel times [1,16,31,32,35,37,55,58,60,82,83,94,97,117,118,129,134,144,145,158,159,161,162,163,164]. ANNs try to learn and predict spatiotemporal relationships in traffic data. They are capable of dealing with the spatiotemporal relationships implicitly without understanding the underlying complex processes. The ANN predictions are a "black-box"-process and, however, depend strongly on the completeness of the input data set. Therefore, the success of the ANN approach rises and falls with the choice of the training data set, which makes the approach difficult in prediction realization.

Pattern Matching: Fitting Traffic Pattern from Database

As it is said before, the congested traffic patterns SP, GP, and EP with its variants on freeways are composed of two basic elements ("traffic phases"), i. e., the regions of

wide moving jams and of synchronized flow. The prediction will be done by forecasting the current spatiotemporal traffic objects, i. e., the regions of the wide moving jams and synchronized flow [69]. Position and width have to be predicted: this is performed based on the most similar traffic pattern for the related effectual bottleneck in the pattern database. The traffic prediction is calculated as follows: firstly, the current situation is reconstructed by FOTO and ASDA models. Next, a filtered current ("generalized") traffic pattern is used as a search pattern for the traffic pattern database. The clustering is carried out to generalize the traffic pattern into some basic "puzzle" pieces of wide moving jam and synchronized flow objects. The reason for this is the dependence of the traffic pattern reconstruction quality on the detector system, which often has limitations in precision and reliability. Therefore, a congested traffic pattern measured by inductive loop detectors and interpreted by FOTO and ASDA models like in Fig. 9 is clustered and combined into larger regions of synchronized flow and wide moving jams, respectively. The similarity between the current clustered pattern and the pattern database is determined based on a distance measure which allows deviations in space and time (e. g., some hundreds of meters and some minutes as parameters). Then, the traffic pattern from the database is linked to the current pattern. The aggregated predicted delay time of the congested pattern consists of the delay time contributions of (i) the propagation of the current wide moving jams already detected, (ii) the predicted regions of the synchronized flow from the traffic pattern database, and (iii) the predicted new wide moving jams emerging in the future as information from the traffic pattern database.

If traffic congestion has been detected, how can the prediction pattern be identified from the database? Criteria for this "matching" process, i. e., finding the most probable traffic pattern for the future, are freeway, location of the effectual bottleneck, time, size of the current synchronized flow region, location of the first wide moving jam (if GP or EP) and the number and frequency of the wide moving jams (if GP or EP).

For applications it should be noted, that the matching process could use the current traffic situation from a traffic control center (e. g., based on online traffic reconstruction with FOTO and ASDA models) or from autonomous vehicles based on the vehicle internal measurements (Fig. 13).

Prediction of Propagation of the Fronts of Wide Moving Jam

An approach for wide moving jam short-term prediction is as follows: from three-phase traffic theory [69] the upstream stable propagation of the downstream front of a wide moving jam is a characteristic parameter of traffic. The upstream front is influenced by the incoming traffic flow into the wide moving jam. A short-term prediction of a wide moving jam can be based on this stable propagation feature: if a wide moving jam is detected and registered by ASDA model, the mean velocities of the fronts can be used for short-term predictions of the wide moving jam. After calculation of the mean velocities of the fronts, those values are used to predict the wide moving jam's position in the future. Therefore, for the wide moving jams stable movement in space and time can be calculated based on the registration times at the local detectors.

A short-term prediction of wide moving jams is suggested in [81]. If both fronts of a wide moving jam have been registered, the clearing up of the jam can be obtained. This could be useful to predict the time, when traffic control methods regarding the wide moving jam could be eliminated. The time t_{clear} is obtained as [81]:

$$t_{\text{clear}} = t_{\text{reg}} - \frac{\int_{t_0}^{t_{\text{reg}}} v_{\text{gl}}(t)\mathrm{d}t}{v_{\text{gl}}(t_{\text{reg}}) - v_{\text{gr}}(t_{\text{reg}})} \quad \text{at } \left|v_{\text{gl}}\right| < \left|v_{\text{gr}}\right| \quad (7)$$

with v_{gl} and v_{gr} as the related velocities of the up- and downstream front of the wide moving jam, respectively. t_0, t_{reg} are the registration times of the upstream and downstream front of the wide moving jam at the downstream position. Figure 14 illustrates scheme and empirical example for a wide moving jam clearing up.

A short-time travel time prediction can be established in the case of a propagating wide moving jam, if the downstream front of the wide moving jam has been registered at a downstream location. As illustrated in Fig. 15, during the first drive section until the upstream front reached at t_{up}, the drive line is based on an average free speed of v_0 the vehicle. For the subsequent drive section within the wide moving jam, the average vehicle speed of v_{jam} is used to generate the drive line. The last drive section after the downstream front of the wide moving jam is then calculated with the speed v_n. Therefore, the individual vehicle travel time for this section can be obtained as [81]:

$$t_{\text{R}} = \frac{L + v_n\, t_{\text{down}} - v_0\, t_{\text{up}} - v_{\text{jam}}(t_{\text{down}} - t_{\text{up}})}{v_n} \quad \text{at } t_0 = 0\,. \quad (8)$$

In many cases, the remaining speed within the wide moving jam v_{jam} can be neglected, so that for the travel time the following approximation formula is obtained [81]:

$$t_{\text{R}} = \frac{L + v_n t_{\text{down}} - v_0 t_{\text{up}}}{v_n} \quad (9)$$

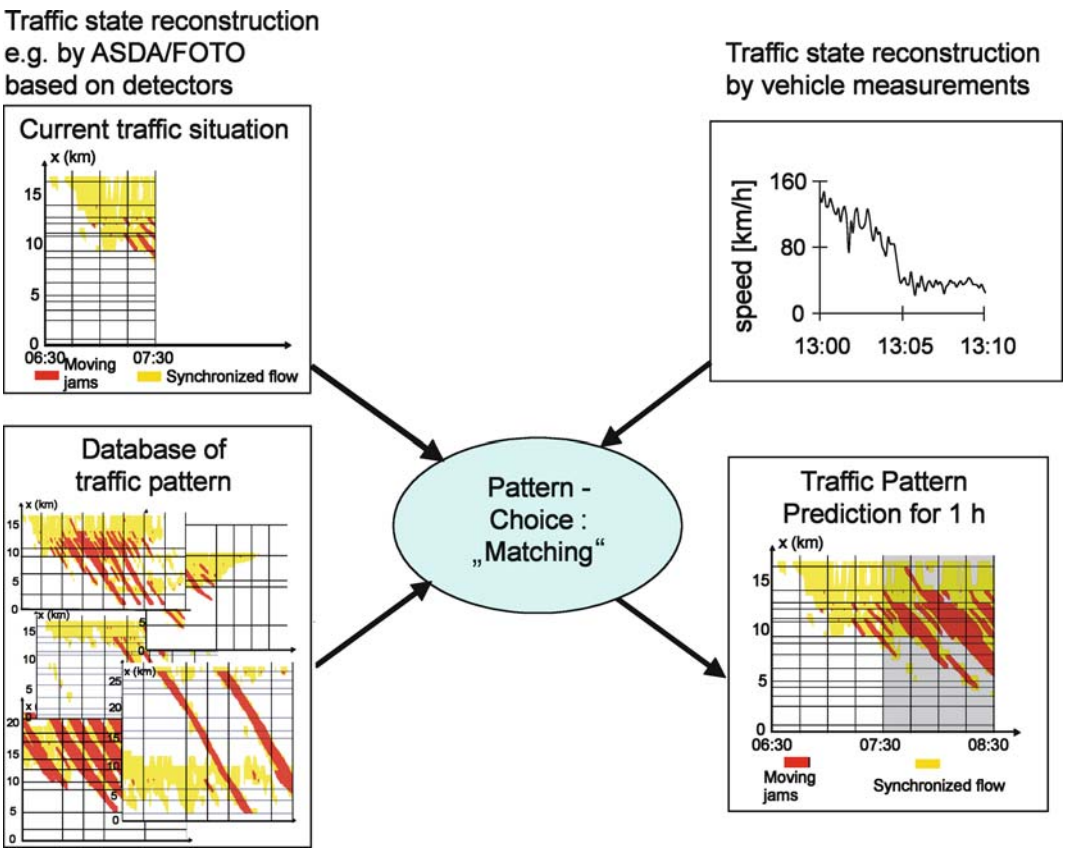

Traffic Prediction of Congested Patterns, Figure 13
Use of spatiotemporal traffic pattern database for traffic prediction with "matching" [69]

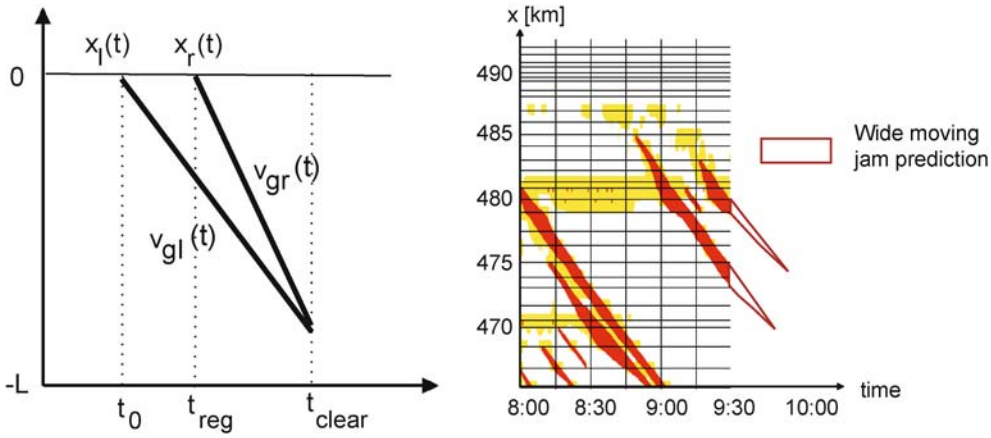

Traffic Prediction of Congested Patterns, Figure 14
Scheme for a wide moving jam clearing-up (*left*: scheme, *right*: empirical example) [81]

This method can also be used when several wide moving jams occur between two measuring points. In this case, assuming that no entry or exit roads exist in the monitoring area, the plausible assumption used is that the flow and average speed of vehicles corresponds both upstream and downstream of the wide moving jam at the time after the downstream front has passed the downstream measuring point [81].

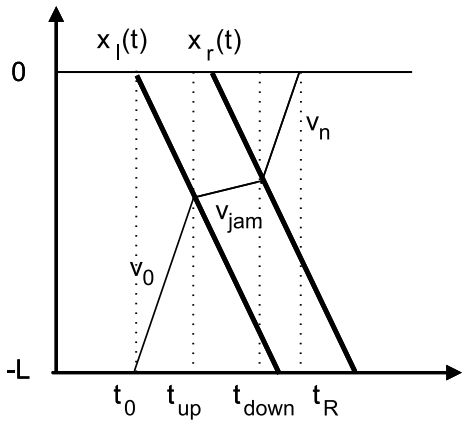

Traffic Prediction of Congested Patterns, Figure 15
Scheme for a travel time prediction for a vehicle [81]

Traffic Pattern Prediction with Traffic Flow Models

Traffic flow models offer a possibility to perform traffic predictions in different scenarios (e. g., [22,24,25,26,96,102,151]): through the use of a model, different (possibly future) influences on traffic like e. g., incidents, roadwork, accidents, etc. could be simulated. Therefore, the consequences of such future events could be estimated and used for current decisions, e. g., the choice of the best traffic management strategy. In this context, a traffic prediction based on traffic modeling could be a valuable instrument.

Many authors have used traffic flow models for traffic prediction, e. g., [22,24,25,26,102,151]. One of the key requirements for successful modeling is the correspondence with the empirical features of freeway congestion. In [73] is shown that all earlier traffic flow model approaches like the above mentioned ones based on the so-called "fundamental-diagram" approach are not able to explain those empirical features. As a consequence, the related traffic prediction results are doubtful.

On the other hand, traffic prediction and scenarios modeled based on KKW-model [69] could be a reliable instrument for decision support with traffic simulation: the consequences of different influences on traffic flow could be simulated in advance. More detail information on traffic simulation can be found in ► Traffic Congestion, Modeling Approaches to.

Long-Term Traffic Prediction

Many empirical studies of traffic have shown that the time-dependence of traffic flow possesses some predictable features (e. g., [50,52,69,80,98,109,120]) because they are recurring and regular. These features depend on characteristics of the day (for example working day or weekend) or whether there is an event like a soccer match. The regularity and recurrence is even stronger if a certain time shift of the congestion emergence is allowed, the morning traffic peak with the same traffic congestion 10 min later is not an "anomaly", but still a regular and similar traffic pattern.

If the characteristics of a day are known, there is a certain set of dependences of the flow rates over time for a specific freeway section. Each of these dependences is one of the historical time series for flow rates. A historical archive of traffic flow rates is therefore the basis for a time series database structured by the characteristics of the different days. At least two parameters of the databases are the day characteristics and an event which influences the traffic. The historical database contains a related probability of the occurrence for each of the time series for each of the different parameters. A cluster analysis of the historical data is a common approach to produce the time series database.

For urban areas time series based predictions have been proposed (e. g., [54,69,162]). Those methods should predict the traffic state for different time horizons. The database should use data from the infrastructure (road net, lanes, traffic lights), basic traffic engineering data (e. g., capacities, speed limits), historical traffic data as expectation values of, for example, the flow rates and current measurements.

Two types of prediction based on long-term time series are possible. Firstly, the choice of the most probable time series for a similar event or day in the future (e. g., soccer event, weekday or weekend) could be applied. Based on archived traffic data the database contains time series with additional parameters like day, time of day, event, weather, etc. The prediction for the next day is then performed based on a fulfilment of such criteria. Secondly, the prediction can include "matching of time series", i. e., the prediction is not necessarily the most probable curve for the future from the historical database, but based on time series with minimal deviation from the current traffic variables.

Long term time series of traffic variables have two aspects: (i) processing of the raw data to produce a time series database and (ii) their later usage in a predictive application. For the processing of raw traffic data, approaches from signal processing have been proposed [45,121,140,141,147,158] to extract and detect the relevant events in the data. Others take different and multiple methodologies from statistics (e. g., [10,11,15,33,40,43]) like cluster analysis [50,120], ARIMA (Auto-Regressive Integrated Moving Average) approaches for stationary linear processes ([14,144,156]), stochastic methods [3,107,108] or Bayes'

Theorem (e. g., [20,53]). Additionally, neural network approaches from artificial intelligence have been used for unsupervised learning transforming raw traffic data into time series [144,161].

In general, the methods for long-term time series prediction try to reproduce the traffic phenomena by their different aggregation methods. The resulting time series from the raw data should model a realistic curve of the traffic variable to be useful for a predictive application. The mentioned approaches have a commonality in their usage of a large variety of data analysis techniques. Some of them do not take into account the known physical and predictable features of the traffic process [69] while others even neglect the principle feature of non-linearity and non-stationarity of traffic (e. g., ARIMA [144,156]). Without proper understanding of the relevant features of traffic congestion it is doubtful that those methods could be optimal and successful.

Also microscopic traffic simulation has been used to estimate long-term effects in traffic in combinations of time series and modeling of traffic process [22,62,84,102,142,151]. The goal is a network wide traffic prediction in a combination of measurements and modeling techniques.

Predictable Features of Traffic Time Series

Long-term time series of flow rates are used for infrastructure planning and modeling of the traffic demand (e.g [50,52,120]). For the long-term the typical minute values of traffic data measurements are aggregated in different time intervals. For usage as traffic demand time series the influence of traffic congestion has to be minimized or filtered. Data aggregation is one key technique to extract predictable features like the flow rate curve from raw measurement data [111].

An example of typical detector measurements for double induction loop detectors is shown in Fig. 16. Three loops count the number of vehicles ("flow rate") and their average speed in one minute intervals on each lane. Here, the aggregated values for the freeway direction A5-North are illustrated for Friday, 23rd March, 2001. The minute values show several local minima of flow rates in the afternoon, both clearly visible in the flow and speed time series (Fig. 16, top). The emergence of the congestion has occurred at a downstream location, and the local detector on the freeway has registered the local minima. These local minima maybe associated with a part of the whole congested traffic pattern, e. g., some wide moving jams propagating upstream as part of a General Pattern (see [69] for explanation): in the dynamic spatiotemporal traffic on freeways, the local detector only registers a part of the whole congested situation). A maximum flow rate of approximately 6000 veh/h on this day is a good "rule of thumb" for a three lane freeway, i. e., 2000 vehicles/h per lane is a good estimation for the maximum possible throughput (including long vehicles).

The moving average of 5 min (Fig. 16, middle) smoothes the fluctuations of the minute interval and shows clearly similar local minima, because they usually last longer than 5 min. If the moving averaging interval is set to one hour (Fig. 16, bottom), the resulting time series is smoothed considerably. For long-term applications, e. g., a traffic prediction of the next day, those aggregated hourly traffic data is sufficient. The form of the 60 min-time series is typical for a freeway: an increasing flow rate in the morning, higher flow rates during the day, an afternoon peak and decreasing flows during the evening hours.

Predictable Features of Effectual Bottlenecks

In addition to the temporal aggregation of the previous chapter, a spatial aggregation would enhance the long-term time series prediction to a long-term congested traffic pattern prediction with use of time and location as key predictive parameters. According to three-phase traffic theory with the defined two phases in congested traffic, i. e., synchronized flow and wide moving jam, spatiotemporal congested traffic patterns emerge after a first phase transition from free flow to synchronized flow that explains traffic breakdown [68,69]. A bottleneck at which traffic breakdown is observed is called "effectual bottleneck". It has two empirical features: (i) the phase transition from free to synchronized flow, i. e., traffic breakdown, occurs more frequently at this location in comparison to other freeway locations and (ii) the downstream front of the synchronized flow region where the vehicles accelerate from the congested region to free flow is almost fixed at this effectual bottleneck (see B_1-B_3 in Fig. 17). The reason for the existence of an effectual bottleneck is a non-homogeneity of the freeway infrastructure caused by, e. g., on- and off-ramps, a decrease in the number of lanes, a higher slope and so on which influences the undisturbed homogeneous traffic flow. From a study of congested pattern formation from 1995–2003, it has been shown that the phase transitions from free to synchronized flow at the freeway A5-South have occurred mainly in the vicinity of the effectual bottlenecks B_1 (intersection "Nordwestkreuz"), B_2 (intersection "Bad Homburg") and B_3 (intersection "Friedberg") (see Fig. 17).

The resulting congested patterns emerging after the phase transition at the effectual bottlenecks have pre-

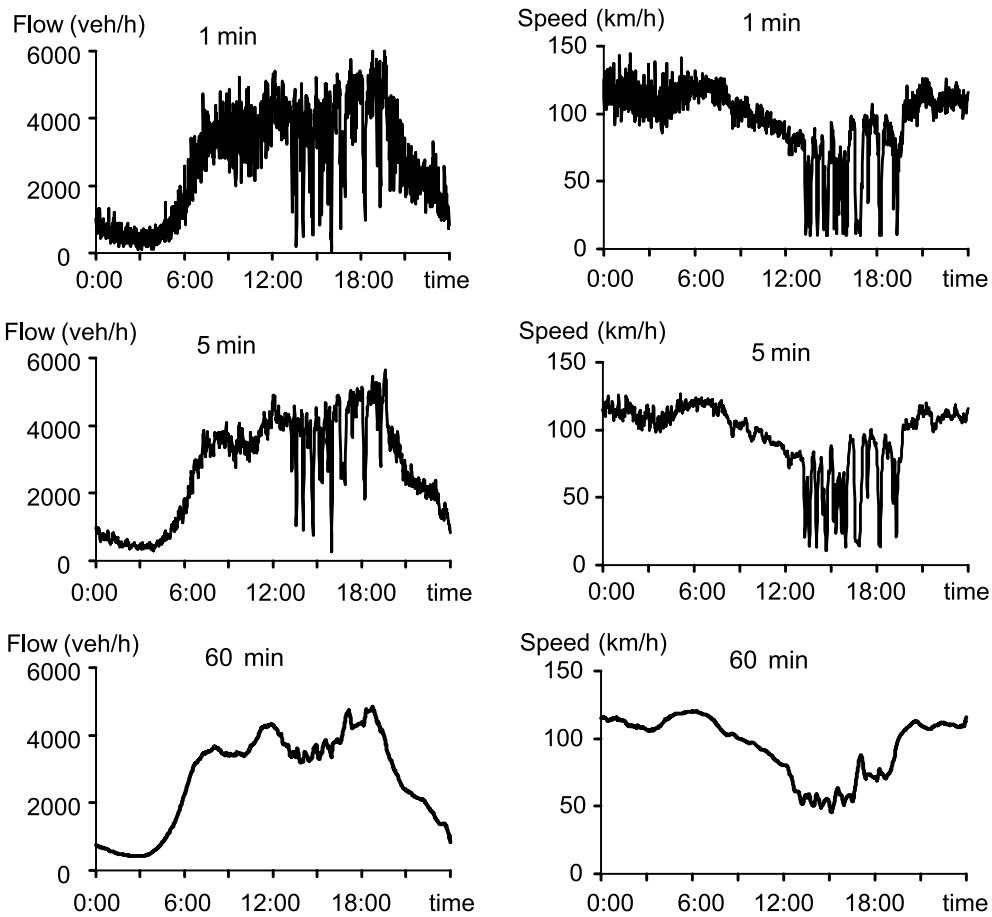

Traffic Prediction of Congested Patterns, Figure 16

Minute values (*top*), moving averages of this data for 5 min (*middle*) and 60 min (*bottom*) of flow rates (*left*) and vehicle speed (*right*) from a typical detector from three-lane freeway A5-North between intersections "Nordwestkreuz" and "Bad Homburger Kreuz" on Friday, 23rd March 2001

dictable and characteristic features: the pattern features are the same for different days and years (see Fig. 17 with space over time diagrams). From a global analysis many similarities among the patterns can be identified. The size, form and duration of the synchronized flow regions, wide moving jam emergence and their propagation through different bottlenecks, as well as the time dependences of phase transitions at different bottlenecks (i. e., phase transition from free to synchronized flow always earlier at B_3 than at B_2) are very similar on different days in different years. This supports predictability in a long-term congested traffic pattern database: prediction at a certain time on a certain day could be performed based on the current traffic conditions (current congestion, flow rate, weather, time of day, etc.) and a correlation to the pattern parameters of the database. Furthermore, the data analysis shows that the congested patterns are like "fingerprints" of the road network: each effectual bottleneck may have its specific characteristic parameters of the congested patterns, i. e., type, form and duration of the congested pattern are linked to the specific bottleneck at which the congestion emerges [69].

For long-term time series of congested traffic patterns, the possible degree of predictability has to be taken into account. According to probabilistic features of the congested patterns, even under the same traffic parameters (e. g., flow rates, weather, road conditions, etc.) the pattern features could be different for different situations (days) with different degrees of predictability. Some deterministic features with a high degree of predictability are as follows [69]:

(i) Location of the effectual bottleneck,
(ii) Emergence of wide moving jams in synchronized flow regions upstream of bottlenecks in a distance of about 3 km or more,

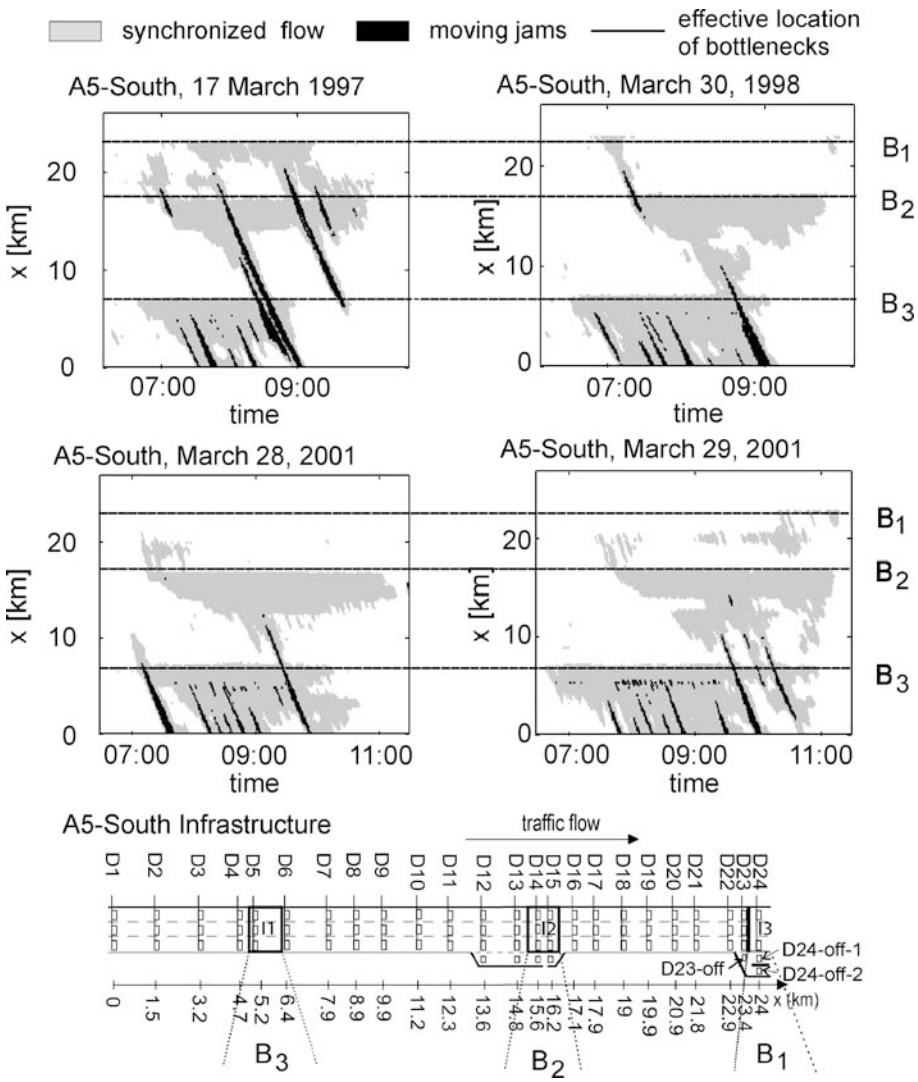

Traffic Prediction of Congested Patterns, Figure 17

Empirical examples of predictable congested patterns in different years at the effectual bottlenecks B_1–B_3 on the freeway A5-South (*bottom*: infrastructure). Overview on regions of free flow (*white*), synchronized flow (*light grey*) and wide moving jams (*black*) in space over time diagrams [69]

(iii) Stable propagation of wide moving jams through any other traffic states,
(iv) Considerably lower flow rate in the outflow of the wide moving jam if free flow emerges after the jam (about two thirds of the maximum flow rate in free traffic).

Congested pattern features with a middle degree of predictability are:

(i) The instant of a phase transition from free to synchronized flow at the effectual bottleneck at the same traffic demand ("probabilistic nature of speed breakdown"),
(ii) The type of congested pattern occurring at an effectual bottleneck. Some of the pattern types [69] may not occur at a specific set of adjacent effectual bottlenecks.

Thirdly, pattern features with a low degree of predictability are:

(i) The kind of phase transition from free to synchronized flow (induced, i. e., influenced by another adjacent effectual bottleneck or spontaneous, i. e., originally emerging at the related bottleneck),
(ii) The instant of wide moving jam emergence in synchronized flow.

A long-term pattern database therefore should contain the empirical congested pattern features and store the predictable information with related probabilities. An approach for development of such a congested traffic pattern database is described later on.

Approaches for Producing Long-Term Traffic Time Series

Cluster Analysis for Flow Rates One kind of detector on freeways and major roads in Germany allows hourly measurements of the flow rates at certain locations in the freeway network ("Dauerzählstellen"). Performing a cluster analysis of this hourly traffic data, typical recurring and regular time series of the flow rates are extracted from the traffic data (Fig. 18). The typical bimodal structure for weekdays (e. g., Monday in Fig. 18) and the unimodal structure for weekend days has been identified clearly [50,52,120]. In [120] seven types of time series for weekdays based on a statictical cluster analysis have been identified which can be characterized by their features:

- Type A: strong morning peak
- Type B: morning peak and lower afternoon peak
- Type C: relatively balanced distribution during the daytime
- Type D: double peak
- Type E: afternoon peak, low morning peak
- Type F: strong afternoon peak
- Type G: more than average morning, then decreasing steadily.

According to [120] Type A and B are dominant on freeways with commuting traffic with the opposite direction of Type E and F.

These time series represent an estimation of the traffic demand for the freeway and major road network, they do not take traffic congestion into account. In a comparison of the averaged detector measurements of Fig. 16 (bottom, left) and the hour values and the clustered results of Fig. 18, the freeway stretch of the A5-North between intersections "Nordwestkreuz" and "Bad Homburger Kreuz" on Friday, 23rd March, 2001 shows a "Type C" form with a relatively balanced distribution during the daytime. In a database, only this type classification and some parameters like the maximum flow rate must be stored for a long-term flow rate time series.

As an additional source of information for long-term time series the radio traffic messages via RDS/TMC (Radio Data System/Traffic Message Channel) protocol can be used as archived traffic data [124]. The main events of the broadcasted information are the "stationary" and "queuing" traffic which are coded in the broadcast data stream. In the year 2000, approximately one million messages were broadcasted in Germany. A daily distribution of those messages with relative percentages per hour shows in Fig. 19 a double peak in the morning and afternoon hours similar to Type D in Fig. 18. These time series illustrate the time distribution of the traffic congestion; for the whole network, the evening peak is wider than the morning one and approximately 10–12% of all congestion occur in the freeway network during the two peak hours of the day (from 7:30-8:30 h and from 15:30-16:30 h, respectively).

The comparison of Figs. 18 and 19 illustrate the correlation between the flow rate time series and the congestion message distribution: from a global perspective high traffic flows correlate with highly congested traffic states.

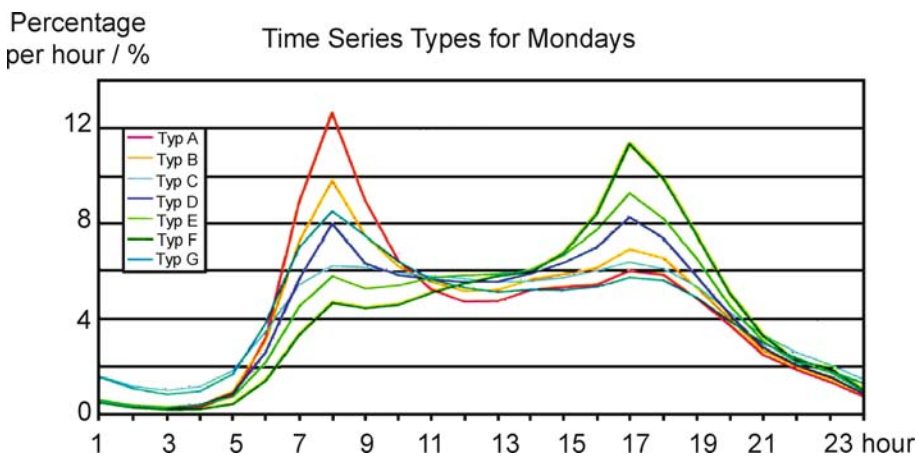

Traffic Prediction of Congested Patterns, Figure 18
Daily time series (Weekday Mondays, Type A – Type G) [120]

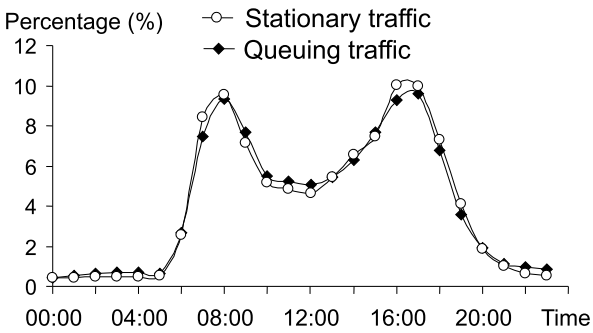

Traffic Prediction of Congested Patterns, Figure 19
Relative percentage of RDS/TMC radio messages ("stationary traffic" and "queuing traffic") per hour for all days of year 2000 in Germany [124]

The interpretation of the RDS/TMC messages of "stationary" and "queuing" traffic is related to speed and flow parameter values as definitions for the TMC protocol, e. g., all measured speeds below 60 km/h are defined as being "queuing" traffic and speeds below 30 km/h are defined as "stationary" traffic. This is in contradiction with empirical features of congested patterns like in Fig. 17 and therefore those time series cannot be used for successful traffic prediction. The TMC protocol definitions of today make it impossible to have information on the real congested traffic situation, but this information is broadcasted by radio stations and all radio listeners in most European (and recently USA) regions are experienced with this common form of traffic information. For dynamic route guidance applications the "stationary" and "queuing" traffic could be transformed into travel delay times and then be used in navigation systems. A specific high degree of predictability based on the real empirical traffic phases has been lost on this aggregated level of information.

Wavelet Analyses The extracting and producing of traffic data time series is cumbersome and time-consuming, as well as computer storage intensive. The frequency domain analyses, using Fourier series [11], offer a possible alternative, especially regarding storage capacity. The Fourier analysis is appropriate for stationary time series models without spikes. The local minima in traffic variables are on the contrary very important for traffic data analyses because of traffic congestion: they can occur suddenly and are relatively seldom events in comparison to all traffic data measurements. The wavelet analyses try to solve these problems [12,45]. Wavelets have been applied to traffic pattern analysis [61], traffic prediction [140,158], incident detection [141], and determining data aggregation intervals [121] in transportation technology. Wavelet analysis consists of breaking a given signal into other waves of different frequencies, thus forming both a high pass and low pass filter at the same time. The ability to analyze the coefficients of the wavelets significantly reduces the dimension of the data vector without losing too much information. This is an efficient approach considering the computational storage costs of the database as well as the performance of the pattern database.

Traffic patterns have some regularity or repetition of some basic underlying structure [45]. An example of this regularity is the typical structure of the bimodal daily pattern of weekday traffic with peaks in the morning and in the afternoon [147] (see Figs. 16 and 18).

For long-term traffic prediction the traffic data time series are averaged over longer time intervals (see Fig. 16). Therefore, it seems questionable that the mentioned dominant advantages of wavelet transformation are relevant for long-term time series.

Cluster Analysis for a Database of Congested Traffic Patterns In an approach supporting Kerner's three-phase traffic theory concepts [69], a database of congested traffic patterns can be used for long-term predictions. In a first step the empirical congested pattern has to be clustered. The clustering builds the congested regions into larger spatiotemporal regions and suppresses smaller congested regions. Therefore, a minimum duration of 30 min of the congestion and a width of at least 1 km distance between upstream and downstream front of a congested region are chosen for the clustering approach. Additionally, smaller spatiotemporal gaps have to be filled. Those occur typically in an online detection system: if at a specific location the detector malfunctions or in one cycle the average speed increases a little, there will be one cycle of free traffic within a congested pattern. As a chosen parameter, time gaps up to a 20 min interval will be filled in during the clustering process. The method for both traffic phases is single linkage clustering based on the minimum distance from two clusters A and B

$$\min_{a \in A,\, b \in B} \{d\,(a, b)\} < X\,, \tag{10}$$

where X is a parameter for cluster merging. The distance between a and b for two traffic flow cluster is calculated by Euclidian distance:

$$d(a, b) = \sqrt{\sum_{k=1}^{p} (a_k - b_k)^2}\,, \tag{11}$$

where k and p are the index and the length of arrays a and b, respectively. The given data points a and b have

to be compared for the calculation of the spatiotemporal distance. Figure 20 illustrates the result: from the empirical pattern based on minute values from a detection system and the reconstruction with FOTO and ASDA models [69] on the left side the number of wide moving jams is reduced in the clustered traffic pattern on the right. Similarly the regions of the synchronized traffic flow are clustered. The resulting clustered congested pattern could be stored in the database with parameters for both synchronized flow and wide moving jam. In comparison, the clustered pattern can reduce the amount of storage capacity for a congested traffic pattern by about 90% which is very computationally efficient for larger road networks.

In the database of congested traffic patterns the patterns are stored with some basic parameters which are relevant for each of the traffic phases ("puzzle pieces"). From three-phase traffic theory it is known that the downstream boundary of the synchronized flow is almost fixed at the bottleneck location. For a synchronized flow region, the starting location and starting time of the synchronized flow region is stored. Secondly, the duration and the expanse are relevant for each of those regions. Two speed parameters describe the form of emergence and dissolution of the synchronized flow regions. The puzzle pieces for synchronized flow have, therefore, four edges in the database.

For the wide moving jams in the traffic pattern database relevant parameters are the length of the wide moving jams and the maximum time of its propagation life cycle. Additionally, for the frequency of wide moving jams following one after the other, the distance between two following upstream jam fronts is given. To define the angle in the space-time diagram, the speed of the front is stored. The latest wide moving jam of a pattern has to take into account the dissolution of the synchronized flow region, i. e., there is no wide moving jam emerging without a synchronized flow region further downstream. This simplified database approach is the first step into online traffic prediction application. Furthermore, detailed information of the congested traffic patterns could be stored in prediction applications [69]:

- the average speed in synchronized flow region,
- the mean velocity and time dependence of the upstream front of the synchronized flow,
- transformations between different types of pattern (SP, GP, EP) after pattern emergence,
- additional information about EP characteristics depending on bottleneck locations,
- probabilities of pattern occurrence dependent on bottleneck location and traffic demand.

The probability of the first pattern occurrence (e. g., from Fig. 21 at a typical bottleneck) is in itself distributed over time. After the traffic breakdown the further development of the congested region is more deterministic than the first phase transition (occurring from free to synchronized flow in Fig. 21 at approx. 15 km and 14:20 h).

Matching with Current Traffic Situation for Prediction

If a structured database with sets of time series of traffic variables exists, the current traffic situation can be used as criterion for choice of the "best fit" prediction (Fig. 22). Measurements in the vehicle or at the traffic center support the search in the archive: a spatiotemporal criterion determines the prediction output. The choice of the "best fit" from the historical archive should classify a "reference" pattern from an anomaly. A time-shift component has to be established to ensure that for example, a morning peak 10 min later is not considered as an anomaly, but

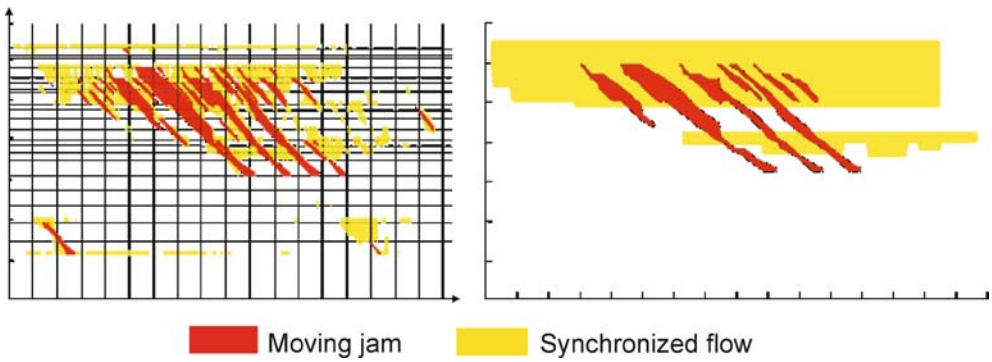

Traffic Prediction of Congested Patterns, Figure 20
Empirical congested traffic pattern at a bottleneck reconstructed by FOTO and ASDA models (*left*) and clustered traffic pattern for the database (*right*)

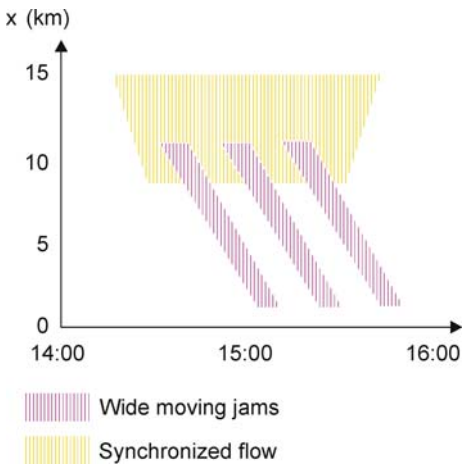

Traffic Prediction of Congested Patterns, Figure 21
Time series of congested traffic pattern database (General Pattern)

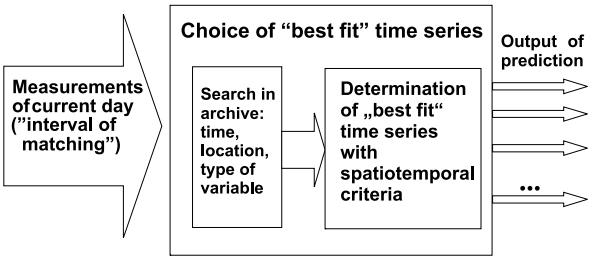

Traffic Prediction of Congested Patterns, Figure 22
Matching of time series with current measurements

still a regular pattern [69,147]. Because of the features of the traffic process, the acceptable tolerance defining similarity among patterns has to integrate spatiotemporal distances. "Best fit" time series must therefore be found in the environment of the specific location, i. e., neighboring detectors or road sections, as well as with time shifts of the current prediction time.

Applications of Traffic Prediction

Currently known applications of traffic prediction are split into several areas. They could be implemented on the infrastructure or the vehicle side. The Internet offers a broadcast medium for distributing traffic prediction to the user without individualizing the prediction information which is dependent on current position and time. One other use of prediction on the infrastructure side is the planning of roadwork: the prediction system estimates the optimal time period of a planned roadwork or estimates the consequences of such a roadwork. The online application of FOTO and ASDA models offers traffic prediction information which can be used in the traffic control center for traffic management purposes. An application concept for prediction usage in vehicles closes this overview.

Online Internet Applications of Traffic Prediction

From the currently known online prediction applications with use of long-term time series a model-driven approach (in Germany: North-Rhine Westphalia) and a data-driven approach (in the USA) are illustrated. They are available via the Internet as collective network short-term predictions and provide dynamic section oriented prediction information. The applications give a global overview similar to a weather prediction approach. Predictions are not individualized for a driver at a certain location at a certain time, for that, the individual route travel time has to be calculated which is dependent on position and time of the traffic prediction.

In addition, both are principally lacking in interpretation of the underlying process in traffic: the traffic state as well as the travel time aggregates the traffic phases on the road, i. e., the same traffic state or travel time value could be calculated possibly from different regions of different traffic phases and its different features. Therefore the precise predictability has been lost in principle on this aggregated information level. In contrast, traffic prediction must be done based on the basic elements of traffic, which are the regions of the traffic phases, because the traffic phases have different deterministic and predictable behavior. Again, traffic prediction methods based on aggregated information levels will not reach a higher accuracy in principle, because deterministic predictable features of traffic phases have been neglected. One example of this problem is the fact, that one traffic phase propagates through the freeway network independent of intersections, on- and off-ramps, etc., while another traffic phase remains almost fixed at one location. However, both traffic phases might be interpreted as the same traffic state or the same travel time value for a road section could be given.

Based on a cellular automata model [102] and the detector measurements, a network traffic state prediction for 30 min and 60 min is performed in a federal state of Germany (http://www.autobahn.nrw.de/nrw.html). It has been developed for fast computation performance in larger road networks. The basic element of this approach is long-term time series of the detector network of North-Rhine Westphalia. The prediction for the next 30 or 60 min is a combination of the model calculations and the estimated speed and flow values from the time series database. It is stated [22], that the real-world mea-

surements adapt in each online cycle of simulation results: therefore, the measurements and the time-series "overrule" the cellular automaton simulation. Strictly speaking, the system shows interpolated detector measurements and long-term assumptions of traffic states generated from archived speed and flow data. The traffic state prediction of this approach makes it impossible to evaluate the prediction quality of the traffic model itself because of the strong combination with the real measurements. The overall system prediction quality could be evaluated with comparisons of predicted and real travel times.

In the USA, nationwide traffic prediction applications have only recently been offered (e. g., [53], http://www.inrix.com). They collect all available detector measurements (flow and occupancy sensor from public detector networks) and dynamic traffic data (incident data, GPS based (Global Positioning System) vehicle data, etc.). Their methodological approach uses Bayesian networks [53]: all factors that combine to result in a congestion event occurring are attempted to be determined. Obvious factors for the approach are the day of the week, the prevailing weather, the occurrence of an accident or some other incident, sporting or other types of event and road construction. The approach views predictive and causal relationships in the context of the effect that one variable has on the probability distribution of outcomes of the other: those probabilistic relationships are in the framework of Bayesian statistics. Historical time series are integrated in these Bayesian networks as traffic congestion are recurring and regular. This approach neglects physical features of the traffic process and Bayes' theorem in itself is not a prediction model, but a probabilistic formula for dependences among variables. The prediction quality in comparison to the reality is currently unknown. The representation of the prediction results is similar to the German approach at http://www.autobahn.nrw.de/nrw.html: the current as well as estimated predicted traffic states are colored for each road segment in up to four different states.

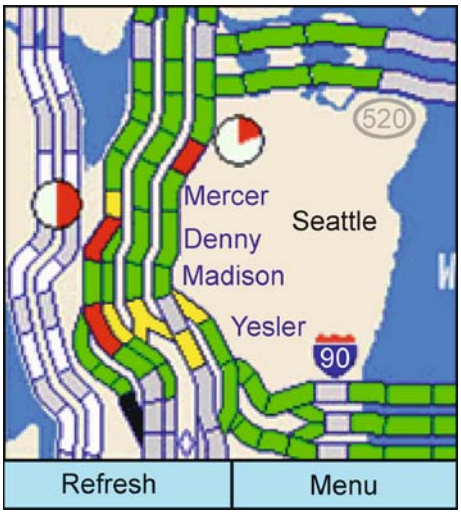

Traffic Prediction of Congested Patterns, Figure 23
Seattle traffic map illustrating a sample user interface, highlighting currently congested road segments in red and smooth flowing traffic in green and predictive information (clocks for duration of the events. Taken from: http://www.inrix.com/predictive.asp)

Roadwork Planning

Long-term time series of flow rates are used in the planning of road construction work and estimate how much higher than the remaining reduced capacity of the network the traffic demand will be [109,128]. Free flow capacities of different kinds of roadwork (depending of the number of (reduced) lanes, the reduced lane width, the form of way in the roadwork) have been measured [52,126] and could be used as parameters for the planning process. If the estimated traffic demand from the long-term flow rate is higher than the expected remaining free flow capacity of the planned roadwork, the number of vehicles in the "queue" for the predicted congestion is counted (dark grey region in Fig. 24). If the estimated flow decreases below the estimated capacity later, a time delay for the dissolution of the predicted "queue" is modeled. This simple model neglects dynamic processes of traffic congestion, probabilistic features and uses a simple queuing of vehicles upstream of the roadwork. Nevertheless, the traffic control center can move the time interval of the planned roadwork to a period where no congestion is predicted, because the traffic demand is assumed to be below the reduced capacity due to the roadwork. For the management of roadwork in a freeway network, this approach is used in the German federal state of Hessen [128]. Because the dynamic processes in real traffic could not be modeled by this approach, i. e., the flow rate is not the only sufficient parameter for congestion emergence [69], the results should be taken only as a guide for public roadwork planning and not as a precise traffic congestion prediction with correct time intervals.

Traffic Control Centers

For the control of freeway traffic, traffic control centers have some instrumentation like speed advisory, lane blockings, collective route choice, etc. To decide the best possible control strategy, a traffic prediction is relevant in-

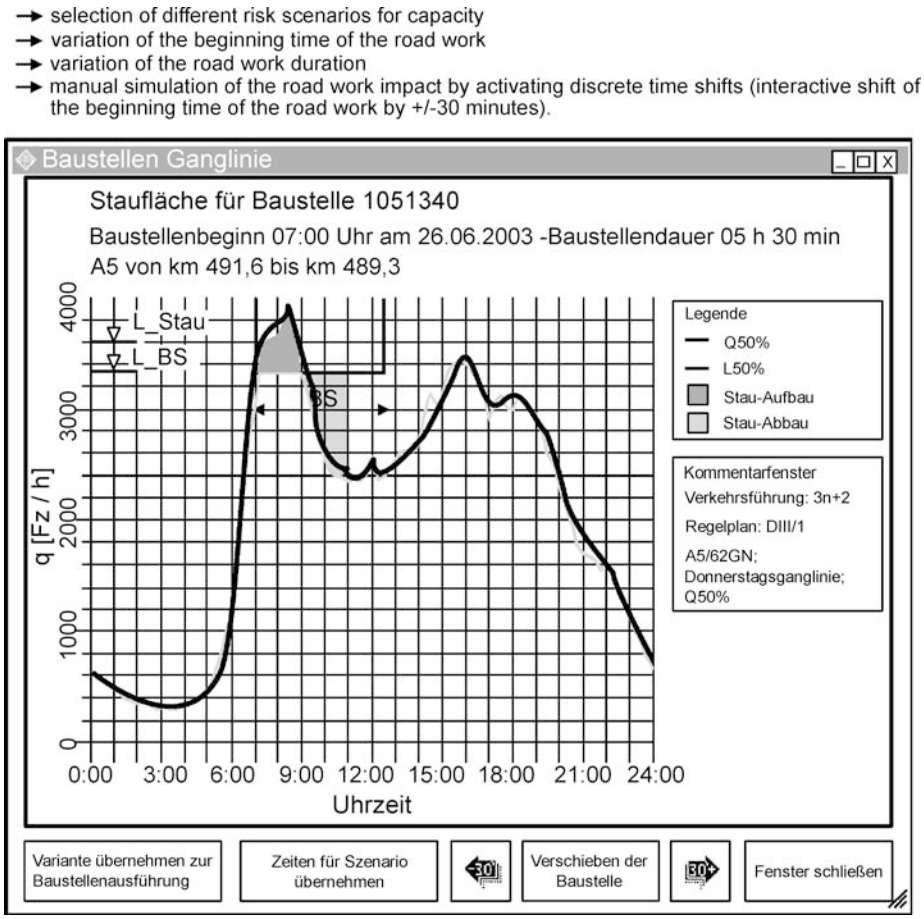

Traffic Prediction of Congested Patterns, Figure 24
Congestion prediction for planned roadwork with long-term flow time series from the German online system in Hessen [128]

formation. In the control center of the German federal state of Hessen, online traffic prediction based on FOTO and ASDA models has been installed and used.

The online application – called FOTOwin integrating both FOTO and ASDA models – offers the possibility to predict the movement of a wide moving jam (see above Subsect. "Pattern Matching: Fitting Traffic Pattern from Database"). In case of wide moving jam recognition, the online system predicts the subsequent propagation of the related wide moving jam. This information could be very useful for vehicle applications of local danger warnings: the sudden traffic breakdown of vehicle speed could be predicted, if such a wide moving jam approaches the individual vehicle.

An empirical example is illustrated in Fig. 25. At 15:00 h the future positions of a wide moving jam have been predicted. The test vehicle has met the upstream front at 14 km at about 15:08 h. A local danger warning of a speed breakdown could have been given about 8 min and 9 km in advance. Inside the vehicle this could be used for driver information, safety and comfort functions. If this information could be given with enough quality in larger road networks, the use for vehicle assistance systems like adaptive cruise control becomes technically feasible.

The online FOTOwin system in Hessen is able to "match" the current traffic pattern with a pattern database. Therefore, the infrastructure gets information on the subsequent development of the congested traffic pattern. This is relevant for decision support for the current traffic management strategy.

One empirical result of a short-term pattern prediction with matching is represented in the following space-time diagram (Fig. 26). The horizontal lines represent the positions of the detectors on the freeway. The light grey area denotes the prediction period with red and yellow vertical lines signifying wide moving jams and synchronized flow,

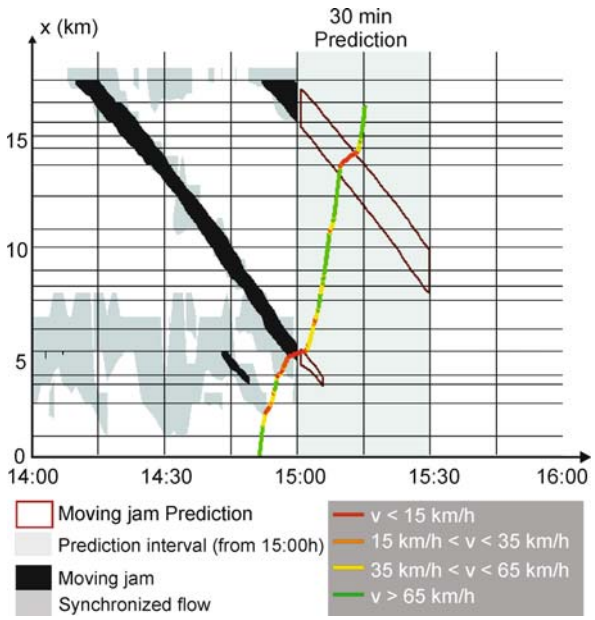

Traffic Prediction of Congested Patterns, Figure 25
Empirical wide moving jam short term prediction with FOTO and ASDA models for a vehicle (colored trajectory)

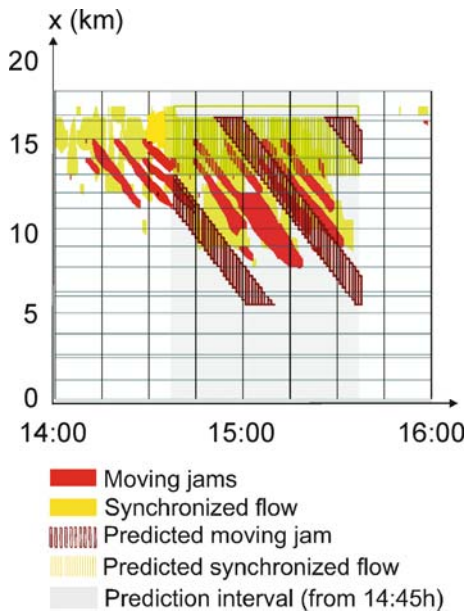

Traffic Prediction of Congested Patterns, Figure 26
Traffic pattern prediction, freeway A3-North, 06th June 2006, prediction at 14:38h

respectively. The related matched traffic database elements from the pattern database are also shown. Together with the prediction from 14:38 h the reality is shown, which has been reconstructed later.

The online FOTOwin system has been installed recently for North-Rhine Westphalia freeways [114]. In principle, the traffic control center located there has similar detector loop data available as in Hessen. These raw traffic data are transferred to WDR, the major public radio broadcasting station from North-Rhine Westphalia in Cologne, who offers traffic messages to the end customer (e. g., radio listener or driver) via broadcast channel RDS (Radio Data System). The application FOTOwin covers a part of the whole freeway network with 1900 km of freeway and more than 1000 double loop detectors.

As an empirical example, a large congested pattern on freeway A45-North is illustrated in Fig. 28. It is clearly visible, that such a pattern is very similar to congested traffic patterns in Hessen (see Fig. 9): emergence of wide moving jams and their stable propagation through many other bottlenecks does eventually occur on every freeway with bottlenecks.

In this example, for almost 7 h on the 17 km stretch a large number of propagating wide moving jams have emerged. At the locations approximately at 10 and 16 km phase transitions from free to synchronized flow occur at

Traffic Prediction of Congested Patterns, Figure 27
Freeway road network observed by FOTOwin in North-Rhine Westphalia [114]

different time moments. Later on, wide moving jams propagate from the downstream bottleneck at approximately 16 km to the upstream bottleneck in the vicinity of 10 km:

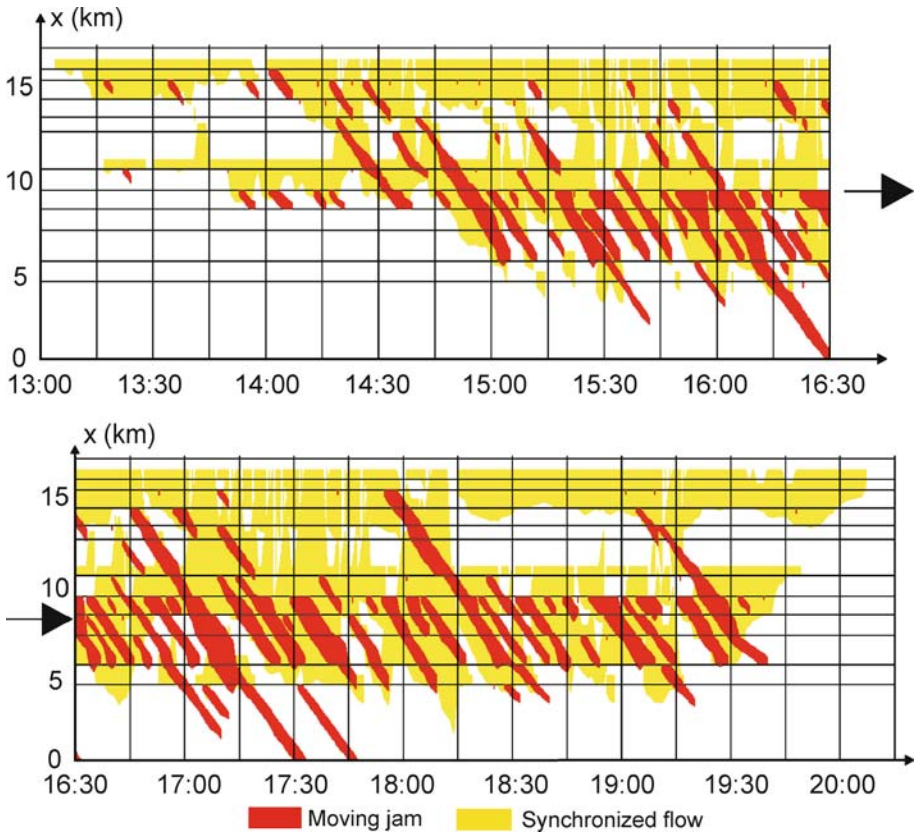

Traffic Prediction of Congested Patterns, Figure 28
Congested traffic pattern reconstructed by FOTO and ASDA models: space-time diagram from freeway A45-North in North-Rhine Westphalia, 31st August 2007 [114]

the occurring empirical Expanded Pattern illustrates the complex interactions of congested traffic patterns which can be explained based on Kerner three-phase traffic theory. Because of the dense freeway network in North-Rhine Westphalia, there are many possible bottlenecks on the freeways due to a large variety of on- and off-ramps with the occurrence of EPs.

In-Vehicle Traffic Prediction

The concept of autonomous traffic prediction in a vehicle is presented in [80]. Figure 29 illustrates the approach: a vehicle has a position unit, a digital map, and the current time. Additionally, the vehicle has an on board memory unit for historical travel time series (e. g., stored on CD/DVD). The linkage of the map information with stored traffic congestion information leads to an adequate choice of travel time prediction associated with the current time and the specific vehicle location. Subsequently, the driver can see this traffic prediction indicated in a vehicle monitor. Furthermore, the traffic prediction is useful for dynamic route guidance in navigation systems (e. g., [19]). Results of traffic prediction can also be presented in different visualization forms [80].

A possible extension of the above concept is the integration of historical spatiotemporal patterns [69] to also make short range traffic prediction available (Fig. 30). To reach this goal, a historical spatiotemporal pattern database should be stored in the vehicle, in addition to historical travel times (or flow rates). In this case, based on the current vehicle location and speed measured by the vehicle, an appropriate choice of a spatiotemporal pattern can be made in the vehicle, which the vehicle should meet in the near future (see vehicle trajectory in Fig. 30, bottom, right).

As it is shown (e. g., [33]), traffic is highly regular with recurring and typical characteristics. Therefore the historical patterns are relatively stable within proposed update

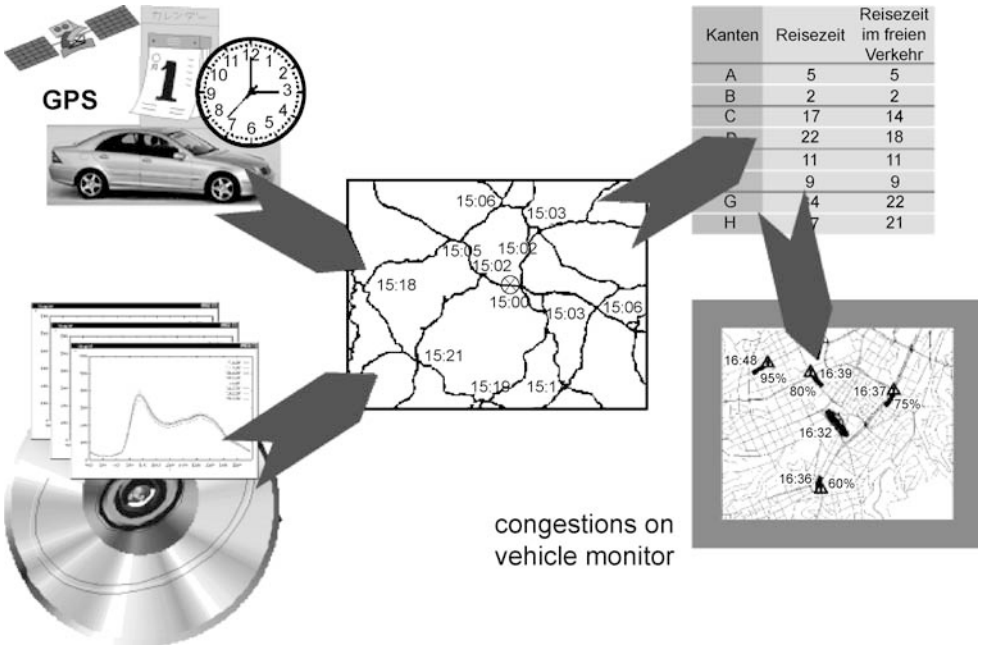

Traffic Prediction of Congested Patterns, Figure 29
Concept of autonomous traffic prediction [80]

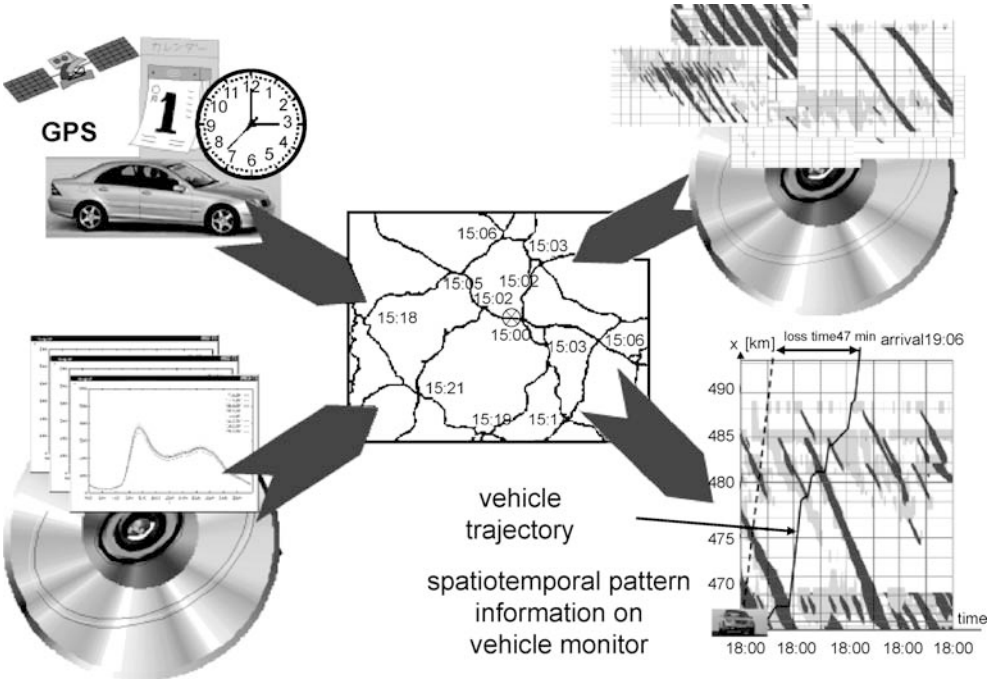

Traffic Prediction of Congested Patterns, Figure 30
Concept of autonomous traffic prediction with use of spatiotemporal pattern database in vehicle [80]

cycles of 1–2 years because of long-term changes in traffic demand and infrastructure: the pattern database in the vehicles could be relatively steady. Communication channels into the vehicle could update the spatiotemporal pattern database.

Future Directions

It must be stated that in the field of traffic prediction a large number of different concepts are coexisting. The proof of reliability especially in times of congested traffic is a still open question to the traffic prediction approaches. Traffic predictions in online information systems and used for collective traffic management or individual navigation system in established applications are in its infancies because of the heterogeneous and often unreliable prediction methods. Therefore, quality definitions and measurement techniques are a key issue for the future evaluation of traffic prediction approaches.

One of the missing elements is a quality definition for precise traffic predictions, because the step beforehand – an area covering precise measuring of the current traffic situation – is still not realized. The measurements of the current situation could be enhanced by using the vehicle as a moving sensor (FCD: Floating Car Data) or using any mobile device in the vehicle (FPD: Flow Phone data). Both technological ideas have been understood conceptually, but are still not introduced in the market because of the lack of successful business models. A definition of prediction quality tested in a laboratory environment should be based on deviation of the predicted values from real measured and archived travel times of the congested time periods.

Microscopic simulation is a very helpful instrument to predict consequences of both traffic control as well as individual traffic assistance systems (e. g., automatic cruise control) before introducing them into traffic control center or into the vehicles. Especially microscopic modeling of vehicle behavior in the case of automatic vehicle assistance systems and their influence of the global traffic system is a complex field of vehicle dynamics. A quality assured network wide traffic prediction with microscopic simulation analogous to weather prediction is still a task of the future.

The traffic pattern database could be enhanced for planned roadwork and accidents which both show characteristic predictable features. The traffic pattern database then has the characteristic content of long-term time series information.

A short-term precise prediction could be used as input for driver assistance systems in vehicles, e. g., for an adapted automatic speed control of the vehicle in the case of approaching congestion.

Traffic simulation in good accordance with the reality on the roads will be a valuable instrument for network traffic predictions. Benchmarking criteria for evaluation of model results should be defined to have quality measures for their application.

One relevant further issue is the enhancement and combination of long-term time series with time series of breakdown probabilities. Anticipation effects will arise with further distribution of dynamic traffic information: people will react on the information in different ways adapting their routing decisions. As consequence, the problem of dynamic traffic assignment under more and more dynamic current traffic information increases: how should the long-term traffic prediction information be given to have an efficient use of the available network resources and not the problem of emerging new congestion in regions where people use roads trying to avoid the already congested regions.

Looking on the network perspective, the effects of prediction on the driver behavior must be taken into account if larger information penetration rates are realized. In case of more available information, the dynamic traffic assignment problem arises, i. e., which part of the driver collective should be advised on which route, finding a system optimum for the road network.

Acknowledgment

We would like to thank Paul Mathias for his critical remarks and fruitful discussions as well as Andreas Haug, Mario Aleksic, Boris Kerner, Jörg Weber and Andrea Köhler for their valuable support. In addition, we thank Gerd Riegelhuth, Ulrich Steiger, Katrin Beckroth and Hendrik Zurlinden from the Hessen Landesamt für Straßen- und Verkehrswesen in Frankfurt for the possibility of research projects and the help with the preparation of traffic data and Susanne Breitenberger from BMW in Munich for the support with vehicle data.

Bibliography

Primary Literature[1]

1. Abdulhai B, Porwal H, Recker W (1999) Short term freeway traffic flow prediction using genetically-optimized time-delay-based neural networks. In: Proceedings 78th Annual Meeting Transportation Research Board, National. Academies Press, Washington DC

[1] Patents information at www.depatisnet.de

2. Acha-Daza JA, Hall FL (1993) A graphical comparison of the predictions for speed given by catastrophe theory and some classic models. Transp Res Rec 1398:119–124
3. Ahmed MS, Cook AR (1979) Analysis of Freeway Traffic Time-Series Data by Using Box-Jenkins Techniques. Transp Res Rec 722:1–9
4. Arem BV, Kirby HR, Van Der Vlist MJM, Whittaker JC (1997) Recent Advances and Applications in the Field of Short-Term Traffic Forecasting. Int J Forecast 13:1–12
5. Barceló J, Ferrer JL, García D, Florian M, Le Saux E (1998) Parallelization of Microscopic Traffic simulation for ATT Systems Analysis. In: Marcotte P, Nguyen S (eds) Equilibrium and Advanced Transportation Modeling. Kluwer, Dordrecht
6. Becker M, Fastenrath U (1998) Method for transmitting local data and measurement data from a terminal, including a telematic terminal, to a central traffic control unit. German Patent Publication DE19755875A1, USA: US6426709B1
7. Ben-Akiva M, Cuneo D, Hasan M, Jha M, Yang Q (2003) Evaluation of freeway control using a microscopic simulation laboratory. Transportation Research, Part C. Emerg Technol 11(1):29–50
8. Ben-Akiva M, Bierlaire M, Koutsopoulos H, Mishalani R (1998) DynaMIT: a simulation-based system for traffic prediction. In: Proceedings of the DACCORD Short-Term forecasting workshop. Delft University, Delft
9. Boker G, Lunze J (2001) State estimation in freeway traffic with floating car data. Automatisierungstechnik 49(11): 497–504
10. Box GEP, Jenkins GM (1976) Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco
11. Brockwell PJ, Davis RA (2002) Introduction to Time Series and Forecasting, 2nd edn. Springer, New York
12. Burrus CS, Gopinath RA, Guo HT (1998) Introduction to Wavelets and Wavelet Transforms: A Primer. Prentice Hall, Upper Saddle River
13. Cameron GDB, Duncan GID (1996) PARAMICS: Parallel Microscopic Simulation of Road Traffic. J Supercomput 10(1):25–53
14. Cetin M, Comert G (2006) Short-Term Traffic Flow Prediction with Regime Switching models. Transp Res Rec 1965:23–31
15. Chatfield C (2001) Time-Series Forecasting. Chapman and Hall/CRC, London
16. Chen H, Grant-Muller S, Mussone L, Montgomery F (2001) A study of hybrid neural network approaches and the effects of missing data on traffic forecasting. Neural Comput Appl 10:277–286
17. Chen H, Grant-Muller S (2001) Use of sequential learning for short-term traffic flow forecasting. Transp Res C 9:319–336
18. Chen M, Chien SIJ (2001) Dynamic freeway travel-time prediction with probe vehicle data – link based versus path based. Transp Res Rec 1768:157–161
19. Chen Y, Bell MGH, Bogenberger K (2007) Reliable Multipath Planning and Dynamic Adaptation for a Centralized Road Navigation System. IEEE Trans ITS 8(1):14–20
20. Chickering DM, Heckerman D, Meek C (1997) A Bayesian approach to learning Bayesian networks with local structure. In: Proceedings 13th Conference on Uncertainty in Artificial Intelligence. Rhode Island, USA, pp 80–89
21. Chien SIJ, Kuchipudi CM (2003) Dynamic travel time prediction with real-time and historic data. ASCEJ, Transp Eng 129(6):608–616
22. Chrobok R, Wahle J, Schreckenberg M (2001) Traffic Forecast Using Simulations of Large Scale Networks, In: Stone B, Conroy P (eds) Broggi A4th International IEEE Conference on Intelligent Transportation Systems. IEEE, Oakland, pp 434–439
23. Cremer M (1979) Traffic Flow on Freeways (in German). Springer, Berlin
24. Daganzo CF (1994) The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. Transp Res B 28(4):269–287
25. Daganzo CF (1995) The cell transmission model, Part II: Network Traffic. Transp Res B 29(2):79–93
26. Daganzo CF (1999) The Lagged Cell-Transmission Model. In: Ceder A (ed) Proceedings of the 14th International Symposium on Transportation and Traffic Theory. Jerusalem, Israel, pp 81–104
27. Daganzo CF (1997) Fundamentals of Transportation and Traffic Operations. Elsevier Science, Oxford
28. D'Angelo MP, Al-Deek HM, Wang MC (1999) Travel-time prediction for freeway corridors. Transp Res Rec 1676:184–191
29. Davis GA, Nihan NL (1991) Nonparametric regression and short-term freeway traffic forecasting. J Transp Eng 117(2):178–188
30. de Rham C, Lange R (2000) Short Term Forecast and Evaluation for Intelligent VMS Settings. In: Proceedings of the 7th World Congress on ITS. ERTICO ITS Congress Association, Torino, Italy
31. Dharia A, Adeli H (2003) Neural network model for rapid forecasting of freeway link travel time. Eng Appl Artif Intell 16(7–8):617–613
32. Dia H (2001) An object oriented neural network approach to short term traffic forecasting. European J Oper Res 131: 253–261
33. Ding A, Zhao X, Jiao L (2002) Traffic flow time series prediction based on statistics learning theory. In: IEEE 5th International Conference on Intelligent Transportation Systems. Singapore, pp 727–730
34. Disbro JE, Frame M (1989) Traffic Flow Theory and Chaotic Behaviour. Transp Res Rec 1225:109–125
35. Dougherty M (1995) A review of neural networks applied to transport. Transp Res C 3(4):247–260
36. Edie LC, Foote RS (1960) Effect of Shock Waves on Tunnel Traffic Flow. In: Highway Research Board Proceedings 39 National Research Council. Washington DC, pp 492–505
37. Fallah-Tafti M (2001) The application of artificial neural networks to anticipate the average journey time of traffic in the vicinity of merges. Knowledge-Based Syst 14:203–211
38. Fastenrath U (1998) Method for determining traffic data and traffic information exchange. German Patent Publication DE19737440A1, USA: US6329932B1
39. Fellendorf M, Vortisch P (2001) Validation of the microscopic traffic model VISSIM in different real-world situations. In: 80th Annual Meeting Transportation Research Board. National Academies Press, Washington DC
40. Fuller WA (1996) Introduction to Statistical Time Series, 2nd edn. Wiley, New York
41. Gartner NH (1973) Highway Research Record 445:12–23
42. Gartner NH (1983) OPAC – A Demand-Responsive Strategy for Traffic Signal Control. TRB, Transp Res Rec 906:75–81
43. Gazis D, Knapp C (1971) Online Estimation of Traffic Densities From Time Series of Traffic and Speed Data. Transp Sci 5: 283–301

44. Gipps PGA (1981) A Behavioural Car-Following Model for Computer Simulation. Transp Res B 15:105–111
45. Grenander U (1996) Elements of Pattern Theory. Johns Hopkins University Press, Baltimore
46. Hamed MM, Al-Masaeid HR, Bani Said ZM (1995) Short-Term Prediction of Traffic Volume in Urban Arterials. J Transp Eng 121(3):249–254
47. Haykin S (1999) Neural Networks: A Comprehensive Foundation. Prentice Hall, Upper Saddle River
48. Head LK (1995) Event-based Short-term Traffic Flow Prediction Model. Transp Res Rec 1510:45–52
49. Hecht-Nielsen R (1990) Neurocomputing. Addison-Wesley, Reading
50. Heidemann D, Wimber P (1982) Types of traffic flow rate time series based on clustering methods, vol 26. BASt, Straßenverkehrszählungen (in German)
51. Helbing D (1997) Traffic Dynamics: New Modeling Concepts in Physics (in German). Springer, Berlin
52. Highway Capacity Manual 2000 (2000) Transportation Research Board. National Research Council, Washington DC
53. Horvitz E, Apacible J, Sarin R, Liao L (2005) Prediction, Expectation, and Surprise: Methods, Designs, and Study of a Deployed Traffic Forecasting Service. In: Proceedings of the Conference on Uncertainty and Artificial Intelligence. AUAI Press, Edinburgh, pp 275–283
54. Hoyer R, Chrobok R, Feldges M, Folkerts G, Friedrich B, Huber W, Kates R, Kemper C, Kirschfink H, Lange R, Listl G, Mathias P, Offermann F, Pinkofsky L, Rehborn H, Schlichting B, Stieler P, Thiemann O, Vortisch P (2003) Advice for data completion and data aggregation in traffic management applications (in German). Hinweispapier der Forschungsgesellschaft für Straßen- und Verkehrswesen, FGSV-Papier, vol 382
55. Huang SH, Ran B (2003) An application of neural network on traffic speed prediction under adverse weather condition. In: 82nd TRB Annual Meeting. National Academies Press, Washington DC
56. Huisken G, Van Berkum EC (2003) A comparative analysis of short-range travel time prediction methods. In: 82nd TRB Annual Meeting Transportation Research Board. National Academies Press, Washington DC
57. Hunt PB, Robertson DI, Bretherton RD, Winton RI (1981) SCOOT – A traffic responsive method of coordinating signals. TRRL report No. LR1014. Transport and Road Research Laboratory, Crowthorne
58. Innama S (2001) Short term prediction of highway travel time using MLP neural networks. In: 8th World Congress on Intelligent Transportation Systems. Sydney, Australia, pp 1–12
59. Ishak S, Al-Deek H (2002) Performance evaluation of short term time series traffic prediction model. ASCEJ, Transp Eng 128(6):490–498
60. Ishak S, Alecsandru C (2004) Optimizing traffic prediction performance of neural networks under various topological, input, and traffic condition settings. J Transp Eng 130:452–465
61. Jiang X, Adeli H (2004) Wavelet Packet-Autocorrelation Function Method for Traffic Flow Pattern Analysis. Comput Aided Civ Infrastruct Eng 19:324–337
62. Kaumann O, Froese K, Chrobok R, Wahle J, Neubert L, Schreckenberg M (2000) Online Simulation of the Freeway Network of NRW. In: Helbing D, Hermann HJ, Schreckenberg M, Wolf DE (eds) Traffic and Granular Flow '99. Springer, Berlin, pp 351–356
63. Kaysi I, Ben-Akiva M, Koutsopoulos H (1993) An integrated approach to vehicle routing and congestion prediction for real-time driver guidance. Transp Research Rec 1408:66–74
64. Kerner BS (1998) Experimental features of self-organization in traffic flow. Phys Rev Lett 81:3797
65. Kerner BS (1999) Traffic prediction method for road network with traffic controlled network nodes (in German). German Patent DE19940957C2
66. Kerner BS (1999) Method for monitoring the condition of traffic for a traffic network comprising effective narrow points. German Patent DE19944075C2, USA Patent: US6813555B1, Japan Patent: JP2002117481
67. Kerner BS (1999) Congested traffic flow: Observations and theory. Transp Res Rec 1678:160–167
68. Kerner BS (2002) Empirical macroscopic features of spatial-temporal traffic patterns at highway bottlenecks. Phys Rev E 65:046138
69. Kerner BS (2004) The Physics of Traffic. Springer, Berlin, New York
70. Kerner BS (2007) On-ramp metering based on three-phase traffic theory. Traffic Eng Control 48(1):28–35
71. Kerner BS, Aleksic M, Denneler U (1999) Traffic condition supervision in traffic network, undertaking inquiry of current position and/or prognosis of future position of flank between area of free traffic and area of synchronized traffic continuously. German Patent DE19944077C1
72. Kerner BS, Herrtwich RGH (2001) Traffic Forecasting. Automatisierungstechnik 49:505–511
73. Kerner BS, Klenov SL (2006) Deterministic microscopic three-phase traffic flow models. J Phys A Math Gen 39:1775–1809
74. Kerner BS, Klenov SL, Aleksic M, Rehborn H: Development and Implementation of UTA model for urban traffic prediction. (Unpublished)
75. Kerner BS, Rehborn H (1996) Experimental properties of complexity in traffic flow. Phys Rev E 53:R4257
76. Kerner BS, Rehborn H (1996) Experimental features and characteristics of traffic jams. Phys Rev E 53:1297
77. Kerner BS, Rehborn H (1997) Experimental properties of phase transitions in traffic flow. Phys Rev Lett 79:4030
78. Kerner BS, Rehborn H (1998) Traffic surveillance method and vehicle flow control in a road network. German Patent Publication DE19835979A1, USA Patent: US6587779B1
79. Kerner BS, Rehborn H, Aleksic M, Haug A (2004) Recognition and Tracing of Spatial-Temporal Congested Traffic Patterns on Freeways. Transp Res C 12:369–400
80. Kerner BS, Rehborn H, Haug A, Aleksic M (2005) Traffic Prediction in Vehicles. In: Proceedings 8th IEEE Conference on Intelligent Transportation Systems. Vienna, pp 251–256
81. Kerner BS, Rehborn H, Kirschfink H (1998) Method for the automatic monitoring of traffic including the analysis of back-up dynamics. German Patent DE19647127C2, Dutch Patent: NL1007521C, USA Patent US5861820
82. Kirby HR, Watson SM, Dougherty MS (1997) Should we use neural networks or statistical models for short-term motorway traffic forecasting? I J Forecast 13:43–50
83. Kisgyorgy L, Rilett LR (2002) Travel time prediction by advanced neural network. Period Polytech Ser Civ Engin 46(1):15–32
84. Kitamura K, Kuwahara M (eds) (2005) Simulation Approaches in Transportation Analysis: Recent Advances and Challenges.

Operations Research/Computer Science Interfaces Series, vol 31. Springer, New York
85. Kniss HC (2000) Evaluation of ASDA/FOTO in traffic control centre Hessen (internal report, in German)
86. Koshi M, Iwasaki M, Ohkura I (1983) Some Findings and an Overview on Vehicular Flow Characteristics. In: Proceedings 8th International Symposium on Transportation and Traffic Theory, p 403
87. Kuchipudi CM, Chien SIJ (2003) Development of a hybrid model for dynamic travel time prediction. In: 82nd Annual Meeting Transportation Res Board. Transportation Res. Board, Washington DC
88. Kwon J, Coifman B, Bickel P (2000) Day-to-day travel-time trends and travel-time prediction from loop-detector data. Transp Res Rec 1717:120–129
89. Lan CJ, Miaou SP (1999) Real-time prediction of traffic flows using dynamic generalized linear models. Transp Res Rec 1678:168–178
90. Lee S, Kim D, Kim J, Cho B (1998) Comparison of models for predicting short-term travel speeds. In: 5th World Congress on Intelligent Transportation Systems. ERTICO ITS Congress Association, Seoul
91. Leutzbach W (1988) Introduction to the theory of traffic flow. Springer, Berlin
92. Lieu HC (2000) Traffic estimation and prediction system. Transp Res News 208:3–6
93. Lindveld CDR, Thijs R, Bovy PHL, Van der Zijpp NJ (2000) Evaluation of online travel time estimators and predictors. Transp Res Rec 1719:45–53
94. Lingras P, Sharma S, Zhong M (2002) Prediction of recreational travel using genetically designed regression and time-delay neural network models. Transp Res Rec 1805:16–24
95. Lu J (1990) Prediction of Traffic Flow by an Adaptive Prediction System. Transp Res Rec 1287:13–20
96. Maerivoet S, De Moor B (2005) Cellular Automata models of road traffic. Phys Rep 419:1–64
97. Matsui H, Fujita M (1998) Travel time prediction for freeway traffic information by neural network driven fuzzy reasoning. In: Himanen V, Nijkamp P, Reggiani A, Raito J (eds) Neural networks in transportation applications. Ashgate Publishers, Burlington, Vermont, pp 355–364
98. May AD (1990) Traffic Flow Fundamentals. Prentice Hall, Upper Saddle River
99. Middelham F (2001) Predictability: Some thoughts on modeling. Future Gener Comput Syst 17(5):627–636
100. Moorthy CK, Ratcliffe BG (1998) Short Term Traffic Forecasting Using Time Series Methods. Transp Plan Technol 12(1):45–56
101. Miyata S, Noda M, Usami T (1995) STREA. Proceedings of the 2nd World Congress on Intelligent Transport Systems. Yokohama 1:289–297
102. Nagel K, Schreckenberg M (1992) A cellular automaton model for freeway traffic. J Phys I France 2:2221–2229
103. Nam DH, Drew DR (1996) Traffic dynamics: Method for estimating freeway travel times in real time from flow measurements. J Transp Eng 122(3):186–191
104. Nanthawichit C, Nakatsuji T, Suzuki H (2003) Application of probe vehicle data for real-time traffic state estimation and short term travel time prediction on a freeway. In: Proceedings 82nd Annual Meeting Transportation Research Board. Transportation Res. Board, Washington DC
105. Newell GF (1982) Applications of Queuing Theory. Chapman Hall, London
106. Nicholson H, Swann CD (1974) The Prediction of Traffic Flow Volumes Based on Spectral Analysis. Transp Res 8:533–538
107. Nihan NL, Holmesland KO (1980) Use of the Box and Jenkins Time Series. Technique in Traffic Forecasting. Transportation 9:125–14372
108. Nikovski D, Nishiuma N, Goto Y, Kumazawa H (2005) Univariate Short-Term Prediction of Road Travel Times. In: International IEEE Conference on Intelligent Transportation Systems. IEEE 2005, Vienna, pp 1074–1079
109. Ober-Sundermeier A, Zackor H (2001) Prediction of Congestion due to Road Works on Freeways. In: Proceedings IEEE Intelligent Transportation Systems. Oakland, pp 240–244
110. Oda T (1990) An algorithm for prediction of travel time using vehicle sensor data. In: IEEE 3rd International Conference on Road Traffic Control, pp 40–44
111. Oh C, Ritchie SG, Oh JS (2005) Exploring the relationship between data aggregation and predictability toward providing better predictive traffic information. Transp Res Rec 1935: 28–36
112. Ohba Y, Koyama T, Shimada S (1997) Online learning type of traveling time prediction model in expressway. In: IEEE Conference on Intelligent Transportation Systems. Boston, Massachusetts, pp 350–355
113. Okutani I, Stephanedes YI (1984) Dynamic prediction of traffic volume through Kalman Filtering theory. Transp Res B 18B(1):1–11
114. Palmer J, Rehborn H (2008) ASDA/FOTO based on Kerner's Three-Phase Traffic Theory in North-Rhine Westfalia (in German). Straßenverkehrstechnik 8:463–470
115. Papageorgiou M (1983) Application of Automatic Control Concepts in Traffic Flow Modeling and Control. Springer, Berlin, New York
116. Pancratz A (1991) Forecasting with dynamic regression models. Wiley, New York
117. Park B, Messer CJ, Urbanik TII (1998) Short term traffic volume forecasting using radial basis function neural network. Transp Res Rec 1651:39–47
118. Park DJ, Rilett LR, Han G (1999) Spectral basis neural networks for real-time travel time forecasting. J Transp Eng 125(6): 515–523
119. Petty KF, Bickel P, Ostland M, Rice J, Schoenberg F, Jiang J, Ritov Y (1998) Accurate estimation of travel times from single loop detectors. Transp Res A 32(1):1–17
120. Pinkofsky L (2002) Types of Time Series (in German). In: Verkehrsentwicklung auf Bundesfernstraßen, Bericht der Bundesanstalt für Straßenwesen, Reihe Verkehrstechnik, vol V99. Bergisch Gladbach
121. Qiao F, Wang X, Yu L (2003) Optimizing Aggregation Level for ITS data based on Wavelet Decomposition. In: Proceedings 82nd Annual Meeting Transportation Research Board, National. Academies Press/Transportation Res. Board, Washington DC
122. Rakha H, Crowther B (2003) Comparison and Calibration of FRESIM and INTEGRATION Steady-State Car-Following Behaviour. Transp Res A 37:1–27
123. Ran R, Boyce D (1996) Modeling Dynamic Transportation Networks. Springer, Berlin
124. Rehborn H, Haug A, Aleksic M, Kerner BS, Fastenrath U (2002) Statistical analysis of traffic message archives as decision sup-

port for road construction up to traffic management (in German). Straßenverkehrstechnik 9:478–485
125. Rehborn H, Haug A, Kerner BS, Aleksic M, Fastenrath U (2003) Floating Car Data and methods for recognition and tracking of spatiotemporal traffic patterns (in German). Straßenverkehrstechnik 9:461–468
126. Ressel W (1994) Investigation of traffic at roadworks in the region of maximum capacity (in German). Informationen Verkehrsplanung und Straßenwesen, Universität der Bundeswehr München, vol 7. Munich
127. Rice J, Van Zwet E (2001) A simple and effective method for predicting travel times on freeways. In: Proceedings of the IEEE Conference on Intelligent Transportation Systems. Oakland, pp 227–232
128. Riegelhuth G, Kirschfink H (2003) Management with decision support of road works for traffic flow optimization on freeways. In: Proceedings of ITS World Congress, paper No. 2255T. ERTICO Congress Association, Madrid
129. Rilett LR, Park D (2001) Direct forecasting of freeway corridor travel times using spectral basis neural networks. Transp Res Rec 1752:140–147
130. Robertson DI (1969) TRANSYT: A traffic network study tool. TRRL Report No LR 253, Transportation and Road Research Laboratory. Crowthorne
131. Rumelhart DE, McClelland JL (1986) Parallel Distributed Processing: Exploration in the Microstructure of Cognition. MIT press, Cambridge
132. Schönhof M, Helbing D (2007) Empirical features of Congested Traffic States and Their Implications for Traffic Modeling. Transp Sci 41(2):135–166
133. Schrader CC, Kornhauser AL, Friese LM (2004) Using historical information in forecasting travel times. In: 82nd Annual Meeting Transportation Research Board, National. Academies Press, Washington DC
134. Smith BL, Demetsky MJ (1994) Short term traffic flow prediction: Neural network approach. Transp Res Rec 1453:98–104
135. Smith BL, Demetsky MJ (1997) Traffic flow forecasting: Comparison of modeling approaches. J Transp Eng 123(4): 261–266
136. Smith BL, Williams BM, Oswald KR (2002) Comparison of parametric and non-parametric models for traffic flow forecasting. Transp Res C 10(4):303–321
137. Smith BL, Oswald KR (2003) Meeting real-time traffic flow forecasting requirements with imprecise computations. Comput-Aided Civ Infrastruct Eng 18:201–213
138. Stathopoulos A, Karlaftis MG (2003) A multivariate state-space approach for urban traffic flow modeling and prediction. Transp Res C 11:121–135
139. Sun H, Liu HX, Xiao H, He RR, Ran B (2003) Short-term traffic forecasting using the local linear regression model. J Transp Res Board 1836:143–150
140. Sun H, Xiao HX, Yang F, Ran B, Tao Y, Oh Y (2004) Wavelet Preprocessing For Local Linear Traffic Prediction. In: 83rd Transportation Research Board Annual Meeting. Transportation Res. Board, Washington DC
141. Teng H, Qi Y (2003) Application of wavelet technique to freeway incident detection. Transp Res Part C 11(3–4):289–308
142. Traffic Flow Theory 2006 (2006) Monograph with 22 papers on the subject of traffic flow theory. Transp Res Rec 1965
143. Treiterer J (1975) Investigations of Traffic Dynamics by Aerial Photogrammetry. Ohio State University Technical, Report PB 246 094. Columbus, Ohio
144. Van der Voort M, Dougherty M, Watson S (1996) Combining KOHONEN maps with ARIMA time series models to forecast traffic flow. Transp Res Part C 4:307–318
145. Van Lint JWC, Hoogendoorn P, Van Zuylen HJ (2002) Freeway travel time prediction with state-space neural networks-modeling state-space dynamics with recurrent neural networks. Transp Res Rec 1811:30–39
146. Van Lint JWC, Van der Zijpp NJ (2003) Improving a travel time estimation algorithm by using dual loop detectors. Transp Res Rec 1855:41–48
147. Venkatanarayana R, Smith BL, Demetsky MJ (2005) Traffic Pattern Identification using Wavelets Transforms. In: 84th Transportation Research Board Annual Meeting. Washington DC
148. Vlahogianni EI, Golias JC, Karlaftis MG (2004) Short-term Traffic Forecasting: Overview of Objectives and Methods. Transp Rev 24(5):533–557
149. Vlahogianni EI, Karlaftis MG, Golias JC (2006) Statistical Methods for Detecting Non-linearity and Non-stationarity in Univariate Short-term Time-series of Traffic Volume. Transp Res C 14(5):351–367
150. Wang Y, Papageorgiou M (2005) Real-Time Freeway Traffic State Estimation based on Extended Kalman Filter: A General Approach. Transp Res B 39:141–167
151. Wahle J, Bazzan A, Klügl F, Schreckenberg M (2000) Anticipatory Traffic Forecast Using Multi-Agent Techniques. In: Helbing D, Hermann HJ, Schreckenberg M, Wolf DE (eds) Traffic and Granular Flow '99. Springer, Berlin, pp 87–92
152. Whitham G (1974) Linear and Nonlinear Waves. Wiley, New York
153. Wiedemann R (1974) Simulation of Traffic Flow (in German). Schriftenreihe des Instituts für Verkehrswesen der Universität Karlsruhe, No 8
154. Wild D (1997) Short-Term Forecasting Based on A Transformation and Classification of Traffic Volume Time Series. Int J Forecast 13:63–72
155. Williams BM (2001) Multivariate Vehicular Traffic Flow Prediction: An Evaluation of ARIMAX Modeling. Transp Res Rec 1776:194–200
156. Williams BM, Hoel LA (2003) Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical results. J Transp Eng 129(6):664–672
157. Williams JC, Mahmassani HS, Herman R (1987) Urban Network Flow Models. Transp Res Rec 1112:78–88
158. Xiao H, Sun H, Ran B, Oh Y (2003) Fuzzy-Neural Network Traffic Prediction with Wavelet Decomposition. Transp Res Rec 1836:16–20
159. Yang F, Sun H, Tao Y, Ran B (2004) Temporal difference learning with recurrent neural network in multi-step ahead freeway speed prediction. In: 83rd Transportation Research Board Annual Meeting. Washington DC
160. Yang F, Lin Z, Liu H X, Ran B (2004) Online Recursive Algorithm for Short-Term Traffic Prediction. Transp Res Rec 1879:1–9
161. Yasdi R (1999) Prediction of Road Traffic using a Neural Network. Neural Computing and Applications, vol 8. Springer, Berlin, pp 135–142
162. Yin H, Wong SC, Xu J (2002) Urban Traffic Prediction Using a Fuzzy-Neural Approach. Transp Res C 10:85–98
163. Zhang G, Patuwo E, Hu MY (1998) Forecasting with artificial neural networks: The state of the art. Int J Forecast 14:35–62

164. Zhang HM (2000) Recursive Prediction of Traffic Conditions with Neural Networks. J Transp Eng 126(6):472–481
165. Zhang X, Rice J (2003) Short term travel time prediction. Transp Res C 11:187–210
166. Zwahlen HT, Russ A (2002) Evaluation of the accuracy of a real-time travel time prediction system in a freeway construction work zone. Transp Res Rec 1803:87–93

Books and Reviews

Kalman R (1960) A new approach to linear filtering and prediction problems. ASME Basic Eng J
Kants H, Schreiber T (2004) Nonlinear Time Series Analysis. Cambridge University Press, Cambridge

Travel Behavior and Demand Analysis and Prediction

Konstadinos G. Goulias
University of California Santa Barbara,
Santa Barbara, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Dynamic Planning Practice
Sustainable and Green Visions
New Research and Technology
The Evolving Modeling Paradigm
Examples of Mathematical Models
Summary
Future Directions
Bibliography

Glossary

Activity-based approach A modeling method that accounts for the interdependent relationships among activities and persons to derive travel demand equations.

Dynamic planning The incorporation of trends, cycles, and feedback mechanisms into a process of actively shaping our future. Desired futures are first defined in terms of performance measures and a combination of forecasting and backcasting methods are used to identify the right paths to follow in achieving these futures.

Microsimulation A method to represent the movement in space and time of the most elementary units of a phenomenon. When applied in traffic engineering the units are vehicles. When applied in travel behavior the units are persons and households. Multi-agent microsimulation allows to also represent human interaction with each person modeled as an agent.

Travel demand The amount of travel within a time interval such as number of trips in a day, total amount of distance and total amount of travel time, the locations (destinations) visited, the means used to reach these locations, departure time and arrival time of trips, routes followed in reaching these locations, the sequencing and assembly of trips in groups, and the purpose or activity engaged in at the end of each trip.

Definition of the Subject

Transportation modeling and simulation aims at the design of an efficient infrastructure and service to meet our needs for accessibility and mobility. At its heart is good understanding of human behavior that includes the identification of the determinants of behavior and the change in human behavior when circumstances change either due to control (e. g., policy actions), trends (e. g., demographic change), or unexpectedly (e. g., disasters). This is the key ingredient that drives most decisions in transportation planning and traffic operations. Since transportation systems are the backbone connecting the vital parts of a city (a region, a state or an entire country), in-depth understanding of transportation-related human behavior is essential to the planning, design, and operational analysis of all the systems that make a city function.

Understanding human nature requires us to analyze and develop synthetic models of human agency in its most important dimensions and the most elemental constituent parts. This includes, and it is not limited to, understanding of individual evolution along a life cycle path (from birth to entry in the labor force to retirement to death) and the complex interaction between an individual and the anthropogenic environment, natural environment, and the social environment. Travel behavior research is one aspect of analyzing human nature and aims at understanding how traveler values, norms, attitudes, perception and constraints lead to observed behavior. Traveler values and attitudes refer to motivational, cognitive, situational, and disposition factors determining human behavior. Travel behavior refers primarily to the modeling and analysis of travel demand, based on theories and analytical methods from a variety of scientific fields. These include, but are not limited to, the use of time and its allocation to travel and activities, methods to study this in a variety of time contexts and stages in the life of people, and the arrangement or artifacts and use of space at any level of social organization such as the individual, the household, the community, and other formal or informal

groups This includes the movement of goods and the provision of services having strong interfaces and relationships with the engagement in activities and the movement of persons.

Travel behavior analysis and synthesis can be examined from both objective (observed by an analyst) and subjective (perceived by the human) perspectives in an integrated manner among four dimensions of time, geographic space, social space, and institutional context. In a few occasions the models reviewed here include and integrate time and space as conceived in science with perceptions of time and space by humans in their everyday life. For this reason research includes theory formation, data collection, modeling, inference, and simulation methods to produce decision support systems for policy assessment and evaluation that combine different views of time and space. Another objective of understanding human behavior is conceptual integration. Explanation of facts from different perspectives can be considered jointly to form a comprehensive understanding of people and their groups and their interactions with the natural and built environment. In this way, we may see explanations of human behavior fusing into the same universal principles. These principles eventually will lead to testable hypotheses from different perspectives offering Wilson's, 1998, famous consilience among, for example, psychology, anthropology, economics, the natural sciences, geography, and engineering. Unavoidably this is a daunting task with many model propositions in the research domain and very few ideas finding fertile ground in applications. The analysis-synthesis path in travel behavior gave us methods that help us understand and predict human (travel) behavior only partially leaving many gaps [163]. However, policy questions are becoming increasingly impossible to address with old tools, a large pool of researchers is actively working on new methods, and many public agencies commenced a variety of tool development projects to fill the travel behavior analysis gaps. To capture these trends, we see modeling examples with ideas from a transdisciplinary viewpoint and contributors to modeling and simulation from a variety of merged backgrounds (e. g., see the evolution of ideas in a sequence of the International Association for Travel Behavior Research conferences – www.public.asu.edu/~rpendyal/iatbr/iatbr_index.htm).

In the next sections the evolving paradigm of modeling and simulation is reviewed in detail and three of its fundamental sources are presented. Through the lens of contemporary planning practice the analytical requirements for modeling and simulation are discussed. Then, these same requirements are refined by examining contemporary visions about the world surrounding us and the theories and technologies we can use to build policy analysis models. This article ends with a section describing the emerging modeling and simulation paradigm, a brief section of mathematical models and closes with a summary.

Introduction

The impressive movement forward of transportation modeling and simulation emerges from three related but distinct sources. The first source is a fundamental change in planning practice that one could name *dynamic planning practice* to indicate the existence of bi-directional time (from the past to the future and from the future to today), as well as, assessment cycles and adjustments taking place within the short term, medium term, and long term horizons. These cycles are also bidirectional in time. This source contains three fundamental directions of practice that are *inventory creation and maintenance*, *strategy measurement and evaluation*, *forecasting and backcasting*. The second source is a vision that generates the substantive problems that we need to solve and the specific policies we need to examine. It is named *sustainable and green visions*. Problems and solutions in this general area motivate and inspire contemporary substance and content of policies throughout the world. One can identify three complementary and mutually strengthening directions in the *economy*, *environment*, and *society* that are the three fundamental pillars of sustainability. The third source is the never ending research for improved understanding of the world surrounding us. This source is named *new research and technology* to capture the most important elements of new discovery and new techniques enabling new discovery but also modeling and simulation. Key directions of inquiry within research and technology are *theory building*, *modeling and simulation*, and *enabling technologies*.

Dynamic Planning Practice

Dynamic thinking means that time and change are intrinsic in the thought processes underlying planning activities. In the past, assumptions about the existence of a tenable and general equilibrium and our ability to build the infrastructure needed to meet demand did not require careful orchestration of actions. This was radically changed in the industrialized world to meet specific goals using available finite resources to maximize benefits. Together with our inability to build at will and a tendency to the preservation of non-renewable resources (e. g., land and open space, fossil fuels, time) we are much more motivated to think strategically and to consider in a more careful way

the performance of the overall anthropogenic system as we plan, design, operate, and manage transportation systems. Any action of this type, however, requires that we have a detailed and accurate picture of our facilities, their interconnectedness, their status within the hierarchy of movements, their conditions, and their evolving role. An accurate and more complete picture like this is called an *inventory* herein.

Many planning activities at all geographical levels are preceded by data gathering steps of identifying all the sources of data and information about the specific study area's transportation system and its relationship with the rest of the world. These inventories include the typical information about the resident population – demographics and employment, land available and land uses, economic development and growth, and so forth. It is worth pointing out the inventory contains data and relationships within the geographic area of interest (region) but also the region's relationship with other areas with which substantial flow of people, goods, and communication takes place. Inventories may also include data and information about cultural and historical factors. For example, statewide plans identify a variety of corridors as buffers of land and communities around major routes of the movement of people and goods. Some of these routes were created centuries ago when pioneers were still exploring uncharted lands. These routes experienced a major change when waterways were the main links among economic and military centers, and they are still evolving. Today these same routes contain as backbones railways, freeways, rivers, and often they surround major distribution locations such as ports and airports. Their nature is heavily influenced by their historical and cultural context.

Travel behavior analysts are familiar with inventories created for the regional long range plans, which subdivide the study area in traffic analysis zones with data from the Decennial Census suitably reformatted and packaged for use in a specific application (i. e., the long range regional plan). Then, additional data are assigned to these same subdivisions to build a richer context for modeling and simulation. Thus, the inventory for a typical long range plan is an electronic map of where people live and work, the network(s) that connect different locations, availability of different modes on each segment of the network, as well as information about travel network performance (e. g., link capacities, speeds on links, congestion, and connectivity). Today the tool of choice for data storage and visualization is a Geographic Information System (GIS).

One of the thorniest problems within this context is maintaining an up to date inventory (e. g., characteristics of the population in each zone, presence of certain types of businesses, location and characteristics of intermodal facilities). This is a particularly important issue for periods in between decennial censuses. Year to year updates are very often required to provide "fresh" data. Many of these updates are becoming widely available and much less expensive than in the past. For example, the inventory of the highway network, with suitable additions and improvements, is available from the same private providers of in-vehicle navigation systems. In a similar way, inventories of businesses and residences can also be purchased from vendors. Census data, however, are required even when one uses data from private providers because they contain complementary data (e. g., the age distribution of the resident population) and they tend to provide wider coverage of a country. Although the need for inventories is undoubtedly extremely important many important issues are yet to be resolved. This is the core issue of two Transportation Research Board (TRB) conference proceedings on the National Household Travel Survey http://www.trb.org/Conferences/NHTS/Program.pdf and the US Census and the Census American Community Survey http://www.trb.org/conferences/censusdata/). Examples of unresolved issues include levels of detail we should use in updating the data we have, treatment of errors in the data and model sensitivity to these errors, frequency of data updates and treatment of missing data, and questions about merging different databases. Obviously, the answers to these questions are in the form of "it depends". It depends on the budget (time and money) available, consequences of errors in the data, and the use of models in decision making. In fact, one particular type of data collection is strategy measurement where some of these questions become even more important. We turn now to the second dimension in the dynamic planning practice which is about strategy and performance.

Strategic planning and performance-based planning changed the way we plan for the future. This has been a 20 year long process in the United States as its transportation policy at the Federal, State, and Metropolitan levels is shaped by three consecutive legislative initiatives (ISTEA, TEA-21, and SAFETEA-LU). Under all three legislative frameworks and independently of role, location and perceived need for investment, the overall goal of funding allocation has been to maximize the performance of the transportation system in its entirety and avoid major new infrastructure building initiatives. As a result, planning practice at the Federal, State, and local levels is becoming heavily performance based and designed in a way that motivates the measurement of policy and program outcomes and judging these outcomes for funding allo-

cation. Two examples of performance-based planning are the Program Assessment Rating Tool (PART) at the federal level and performance-based transportation planning at the state level. PART is used to assess the management and performance of individual programs from homeland security to education, employment, and training. This is a tool that offers assessments about programs based on 25 questions divided into sections. For each program a tailored analysis yields summaries that receive a rating from 0 to 100 ranging from ineffective to effective [172]). In a different way but in the same spirit many states have created long range plans that are strategic and they measure transportation performance. Yearly evaluative updates are also used for a state's strategic transportation plan. After a comprehensive public involvement campaign a few themes capturing the desires of the resident population are first identified. To these themes technical requirements based on planners and agency inputs are added, a large number of objectives are created and then a variety of measures of performance are developed. These measures are given target levels that evolve over time to a desired future performance for the entire state and for a finite number of corridors of statewide significance. Yearly evaluations contain measures of target achievement and they should be used to guide an agency in its investments. The interface with regions is also included in this performance-based framework. Many infrastructure improvement projects in the US are selected from lists of projects that regions (called Metropolitan Planning Organizations) submit to their state to be included in a list of projects in the Transportation Improvement Program (TIP) and become candidates for funding. Under statewide performance-based planning, these projects are evaluated with respect to their contribution in meeting the statewide performance measures and in some states the performance measures of the relevant corridor [122]. Although these examples are far ranging in time and space, they contain operations components and yearly evaluations that: a) require data collection, modeling, and simulation at finer spatial and temporal scales than their counterpart planning feedbacks used in the long range transportation planning practice, and b) need a method that is able to coordinate the short, medium, and long term impacts. Emerging from these considerations are questions about the types of consistency we need among geographic scales for planning and operations actions to perform evaluations, policy requirements for coordination among planning activities to ensure consistency, need for suitable methods to coordinate smaller projects in broader contexts (either of policy assessment or geographical area), development of tools required to perform measurement of impacts and program evaluation at the newly defined assessment cycles, and optimal planning activity with evaluation methods. Only a few solutions to the issues above are offered by contemporary projects such as the TRANSLAND project [70]. Within the context of integration between land use and transportation planning and the context of the European Union some of the conclusions include a call to strengthen regional plans, a stronger emphasis on public transport, strategic planning involving all actors, and the packaging of policies aiming at the same objectives. These themes are very similar to statewide and US Federal and European Union levels of planning. Very little, however, is said about the assessment methods and the choices we make in impact estimation. Performance assessment and evaluation of program effectiveness require the use of the inventory discussed before and a battery of models to forecast future expectations as well as to identify the actions required today to achieve desired futures.

As illustrated later in this article a new approach emerges in which models of discrete choice are applied to individual decision makers that are then used to (micro)simulate most of the possible combinations of choices in a day. The result is in essence a synthetic generation of traveling for the entire population. When the microsimulation also includes activities and duration at activity locations it becomes a synthetic schedule. In parallel, for forecasting purposes a synthetic population is first created for each land subdivision with all the relevant characteristics and then models are applied to the residents of each subdivision to represent areawide behavior. Changes are then imposed on each individual as a response to policies and predictive scenarios of policy impacts are thus developed. The evolution of individuals, their groups, and the entire study area can be used for trend analysis that includes details at the level of decision makers (either for passenger travel and/or for freight). In addition, progression in time happens from the present to the future and one could identify paths of change by individuals and groups if the application has been designed in the proper way (e. g., keeping detailed accounting of individuals as they move in time, using models that are designed for transitions over time and so forth). In a forecasting setting progression in time follows calendar time, temporal resolution is most often a year, and the treatment of dynamics is an one-way causal stream to the future.

Within the broader study of futures, forecasting is the method we use to develop *projective scenarios*. Performance-based planning, however, requires tools that can extrapolate from future performance targets the actions required today to reach them. In essence we also need *prospective studies* that start from a desirable future and

Travel Behavior and Demand Analysis and Prediction, Table 1
Backcasting schema

Content	Method
Determine objectives, purpose of the analysis, temporal, spatial and substantive scope of the analysis, decide the number and type of scenarios. Identify endogenous and exogenous variables	Problem orientation with technical representatives and stakeholders
Specify goals, constraints and targets for each scenario analysis and exogenous variables	Stakeholder creativity workshop and brainstorming sessions
Describe present system (building and updating of inventories), patterns and trends. Define processes, their actors, and determinants of outcomes. Identify exogenous variables and inputs to scenario analysis.	Scenario development by technical experts
Scenario analysis. Select suitable approach, analyze system evolution at end time points and intermediate time points, develop scenarios, iterate to make sure all components are consistent/coherent	Scenario assessment by technical experts and stakeholders
Undertake impact analysis. Consolidate scenario results. Analyze social, economic and environmental impacts. Compare results of the last with targets, iterate analysis with any other step as required to ensure consistency between goals and results	Backcasting workshops and stakeholder consultation (repeat to follow the iterations)
Implement Policy Actions	

move backwards to identify specific actions that will lead us to that prospect. *Backcasting* was invented in a study of future energy options by [141], to do exactly this through a participatory process. Scenarios in backcasting are the "images" of the future and the possible paths that will take us to that future. A typical application includes the stages shown in Table 1. An open question, however, remains with respect to scenario construction and assessment. This is particularly important when one considers the serious issues we face with inadequate design of experiments/trials in the forecasting setting. Forecasting and backcasting have some important differences in their objectives. On one hand forecasting is employed to identify likely futures and to develop methods to help us identify small changes in our policies. It is also a method to extrapolate past trends into the future and possibly identify paths of changes that are heavily influenced by habit and inertia. Backcasting, on the other hand is designed to discover new ways to build desirable futures. It is perfectly aligned with strategic planning and it is a better suited method for developing a program of conditions to meet targets. Many of the models developed to date are designed for forecasting applications (either to inform the design of forecasting model systems or to create necessary components in the model systems). Yet, planning practice is moving towards strategy development and therefore needs model components that fit within a backcasting scenario building (see the reversed four-step model in Miller and Demetsky [120], and its neural network implementation in Sadek et al. [143] and the participatory tools in California (http://www.sacregionblueprint.org/ – accessed May 2007).

Sustainable and Green Visions

Policy actions also view the world surrounding us as an integral ecosystem placing more emphasis on its overall survival by examining direct and indirect effects of individual policy actions and entire policy packages or programs (see the examples in [116]). This trend is not limited to transportation. Lomborg [104], shows that a sustainable and green vision encompasses the entire range of human activity and the entirety of the ecosystem we live in. Although these are good news, because the approach enables analyzes and policies that are consistent in their vision about futures, comprehensive views also reveal that the pace of economic growth and development is in clear conflict with the biological pace of evolution with unknown consequences [162] strengthening the view that more comprehensive analytical frameworks are required.

In fact, one of the most recent studies on research needs, which addresses the transportation and environment relationship by the Transportation Research Board of the National Academies [167,168], expands the envelope to incorporate ecology and natural systems and addresses human health in a more comprehensive way than in the past reiterating the urgency to address unresolved issues about environmental damage. As a result, we also experience a clear shift to policy analysis approaches that have an expanded scope and domain and they are characterized by explicit recognition of transportation system complexity and uncertainty.

Reflecting all this, *sustainable transportation* is now often used to indicate a shift in the mentality of the community of transportation analysts to represent a vision

of a transportation system that attempts to provide services that minimize harm to the environment. In fact, in one of the most comprehensive reviews of policies in North America, Meyer and Miller [116], contrast the non-sustainable to the sustainable approaches. They provide a compelling argument about the change in these policies and pathways toward a more sustainable path. In the US during the past twenty years, the need, to examine these new and more complex policy initiatives, has also become increasingly pressing due to the passage of a series of legislative initiatives (Acts) and associated Federal and State regulations on transportation policy, planning, and programming. The multi-modal character of the new legislation, its congestion management systems and the taxing air quality requirements for selected US regions have motivated many new forecasting applications that in the early years were predominantly based on the Urban Transportation Planning System and related processes but during the last five years motivated a shift to richer conceptual frameworks. In point of fact, air quality mandates motivated impact assessments of the so called transportation control measures and the creation of statewide mobile source air pollution inventories [65,107,154] that require different analytical forecasting tools than in any pre-1990 legislative initiatives [124]. An added motivation is also lack of substantial funding for transportation improvement projects and a shift to charge the firms that benefit the most from transportation system improvements creating a need for impact fee-assessment for individual private developers. These assessments create the need for higher resolution in the three dimensions of geography (space), time (time of day), and social space (groups of people with common interests and missions, households, individuals) used in typical regional forecasting models but also the domain of jurisdictions where major decisions are made. They also create a pressing need for interfaces with traffic engineering simulation tools that are approved and/or endorsed in legislation (for examples see Paaswell et al. [126]). Another push for new tools is the assessment of technologies under the general name of Intelligent Transportation Systems (i. e., bundles of technological solutions in the form of user services attempting to solve chronic problems such as congestion, safety, and air pollution). Natural and anthropogenic tragic recent events are adding requirements for modeling and simulation and urgency in their development and implementation as well as more detail in time and space [75].

As Garrett and Wachs [46], discuss in the context of a lawsuit against a regional planning agency in the Bay Area, traditional four-step regional simulation models [30, 80,125] are outpaced by the same legislative stream of the past 20 years that defined many of the policies described above. Unlike the "energy crisis" of the 1970s, the urgency and timeliness of modeling and simulation is becoming more urgent, more complex, and requires an "integrated" approach. Under these initiatives, forecasting models, in addition to long-term land use trends and air quality impacts, need to also address issues related to technology use and information provision to travelers in the short and medium terms. Similarly, the European Union focuses on issues such as: increasing citizen participation, intra-European integration, decentralization, deregulation, privatization, environmental concerns, mobility costs, congestion management by population segments, and private infrastructure finance (see van der Hoorn [174]). Table 2 provide an overview of policy tools that are loosely ordered from the longer term of land use and governance to medium and shorter term operational improvements depending on the lag time required for their impacts to be realized.

These policy initiatives place more complex issues in the domain of regional policy analysis and forecasting and amplify the need for methods that produce forecasts at the individual traveler and her/his household levels instead of the traffic analysis zone level. In addition to the long range planning activities and the typical traffic operations management activities, analysts and researchers in planning need to also evaluate the following: a) traveler and transportation system manager information provision and use (e. g., location based services, smart environments providing real time information to travelers, vehicles, and operators); b) combinations of transportation management actions and their impacts (e. g., parking fee structures and city center restrictions, congestion pricing), and c) assessment of combinations of environmental policy actions (e. g., carbon taxes and information campaigns about health effects of ozone).

To perform all this we need tool that also have forecasting and backcasting capabilities that are more accurate and detailed in space and time. In fact, planning initiatives are moving toward parcel by parcel analysis and yearly assessments. It is also conceivable that we need separate analyzes for different seasons of a year and days of the week to capture seasonal and within a week variations of travel. Echoing all this and in the context of the Dutch reality Borgers, Hofman, and Timmermans [21] have identified five information need domains that the new envisioned policy analysis models will need to address and they are (in a modified format from the original list):

1. social and demographic trends that may produce a structural shift in the relationship between places

Travel Behavior and Demand Analysis and Prediction, Table 2
Examples of Policy Tools

Type of policy tool	Brief description	Source of information*
Land use growth and management programs	Legislation that controls for the growth of cities in sustainable paths	www.smartgrowth.org, www.awcnet.org www.fhwa.dot.gov/planning/ppasg.htm www.compassblueprint.org
Land use design and attention to neighborhood design for non-motorized travel	Similar to the previous but with attention paid to individual neighborhoods	www.sustainable.doe.gov/landuse/luothtoc.shtml www.planning.dot.gov/Documents/DomesticScan/domscan2.htm
City annexations and spheres of influence	City boundaries are divided into incorporated, within the sphree of influence, and external to manage growth	http://countypolicy.co.la.ca.us/BOSPolicyFrame.htm www.ite.org/activeliving/files/Jeff_Summary.pdf
Accelerated retirement of vehicles programs	Programs to eliminate high emitting and older technology vehicles	ntl.bts.gov/DOCS/SCRAP.html
Public involvement and education programs	Programs aiming at defining goals based on the public's desires	www.fhwa.dot.gov/reports/pittd/contents.htm
Health promoting programs	Programs that promote physical activity in travel to benefit health	www.activelivingbydesign.org
Safety measures	A process to incorporate safety considerations in transportation planning	tmip.fhwa.dot.gov/clearinghouse/docs/safety/ www.fhwa.dot.gov/planning/scp/ www.safetyanalyst.org/
Emission control, vehicle miles traveled, and other fee programs (including carbon taxes and trading)	Programs that shift taxation from traditional sources towards pollutant emissions and natural- resource depletion agents	www.fresh-energy.org/ www.fhwa.dot.gov/environment/ www.fightglobalwarming.com/
Congestion pricing and toll collection programs	A premium is charged to travelers that wish to travel during the most congested periods	www.vtpi.org/london.pdf
Parking fee management	Parking pricing used as a tool to restrict access by space and time	www.gmu.edu/depts/spp/programs/parkingTaxes.pdf
Non-motorized systems	Programs to support walking and biking	www.vtpi.org/tdm/tdm25.htm www.psrc.org/projects/nonmotorized
Telecommuting and Teleshopping	The employment of telecommunications to substitute-complement-enhance travel	www.telework-mirti.org www.vtpi.org/tdm/tdm43.htm
Flexible and staggered work programs	Programs that change the workweek of individuals and firms	www.its.dot.gov/JPODOCS/REPTS_PR/13669/section05.htm
Goods movements (freight) programs to improve operations	A variety of programs to facilitate and minimize the damage for freight movement	ntl.bts.gov/DOCS/harvey.html
Highway system improvements in traffic operations and flow	Improved data collection, monitoring, and traffic management	www.transportation.org ite.org/mega/default.asp
Intelligent Transportation Systems (ITS)	Use of telecommunications and information technology to manage and control travel	www.itsa.org/ www.ertico.com/ www.its.dot.gov/index.htm/
Special event planning and associated traffic management	Enhanced procedures to handle the demands of a special event	tmcpfs.ops.fhwa.dot.gov/cfprojects/new_detail.cfm?id=32xxxnew=0
Security preparedness through metropolitan planning processes	A process to incorporate safety considerations in transportation planning	www.planning.dot.gov/Documents/Securitypaper.htm
Individualized marketing techniques with improved information and communication with the "customer"	Public programs to provide personal help in changing travel behavior in favor of environmentally friendly modes	www.local-transport.dft.gov.uk/travelplans/index.htm http://www.travelsmart.gov.au/

*accessed May 2007

and time allocation by individuals invalidating existing travel behavior model systems;

2. increasing scheduling and location flexibility and degrees of freedom for individuals in conducting their every day business leading to the need to consider additional choices (e. g., departure time from home, work at home, shopping by the internet, shifting activities to the weekend) in modeling travel behavior;

3. changing quality and price of transport modes based on market dynamics and not on external to the travel behavior policies (e. g., the effect of deregulation in public transport);
4. shifting of attitudes and potential cycles in the population outlook about travel options; and
5. changing scales/jurisdictions (scale is the original term used to signify the different jurisdictions) – different policy actions in different sectors have direct and indirect effects on transportation and different policy actions in transportation have direct and indirect effects in the other sectors (typical example in the US is the welfare to work program).

The first substantive implication of all these considerations is an expanded envelope of modeling and simulation. Many processes that were left outside the realm of transportation modeling and simulation need to be included as stages of the travel model system. One notable example is the inclusion of *residential location choice, work location choice, and school location choice* to capture the spatial distribution and relative location of important anchor points on travel behavior and to also capture the impact of transportation system availability and level of service on these choices. In this way when implemented policies lead to improved level of service and the relative attractiveness of locations change, shifts in residential location, work location, and possibly school location can be incorporated as impacts of transportation. A similar treatment is needed for *car ownership and car type choices* of households or *fleet sizes and composition* for firms. These car-related choices are expressed as functions of parking availability, energy and other costs and level of service offered by the transportation system (highway and transit). To account for other resources and facilities available for household travel we also need to consider processes for *driver's licensing*, acquiring of *public transportation subscription* (*passes*), and participation in *car sharing programs*. In this way, variables of car availability and public transportation availability in households can be used as determinants of travel behavior. Similar treatment is required for policies that change attitudes, perceptions and knowledge about travel options.

To address some of the policies of Table 2, we need to transition to a domain that contains a variety of outputs that include shares of program participation, sensitivity to accessibility and prices, and the usual indicators of travel on networks using input variables from the processes and behaviors discussed up to this point. Although the number of vehicles per hour per lane is the typical input of traffic operations software, a variety of other variables such as speeds on network links and types of vehicles are also needed for other models such as emissions estimation.

Ideally longer term social, economic, demographic, resource/facilities, and circumstances of people should be converted into yearly schedules identifying periods of vacation, workdays, special occasions, and so forth. These in turn should lead to weekly schedules separating days during which people stay at home from days during which people go to work and days during which they run errands and/or engage in other non-work and non-school related activities. In this way patterns of working days versus not working days can be derived in a natural (con)sequence. As we will see in a later section, a fundamental leap of faith intervenes in practice and converts all this background information into a representative day that is used to create a more or less complete sequence of activities and trips with their destinations and modes used.

In this way decisions and choices people make are organized along the time scale in terms of the time it takes for these events to occur and their implications. For example, decisions about education, careers and occupation, and residential and job location are considered first and they condition everything that happens next. These should be formulated in terms of one or more life course long projects and not represented by a cross-sectional choice model. Similarly, decisions about yearly school and work schedules that determine work days and vacation days in a year should also be modeled as a stream of interrelated choices. Conditional on all this are the daily schedules of individuals and the myriad of decisions determining a daily schedule, which are modeled in much more detail and paying closer attention to the mutual dependency among the different facets of a within a day schedule. The next section explores this further in the context of research and enabling technology. A section on mathematical models later in this article shows the beginning of a new way in modeling a simulation that emphasizes human interaction.

New Research and Technology

The planning and policy analysis discussion identified many requirements for modeling and simulation. Planning and policy expanded the context of travel behavior models to entire life paths of individuals and for this reason a more general modeling framework is emerging. In fact, modeling made tremendous progress toward a comprehensive approach to, in essence, build simulated worlds on computer enabling the study of complex policy scenarios. Although, passenger travel received the bulk of the attention, similar contributions to new research

and technology are found in modeling the movement of goods [151,153]. The emerging framework, although incomplete, is rich in the directions taken and potential for scientific discovery, policy analysis, and more comprehensive approaches in dealing with sustainability issues.

There are four dimensions that one can identify in building taxonomies of simulation models. The first is the *geographic space* and its conditional continuity, the second is the *temporal scale* and calendar continuity, the third is interconnectedness of *jurisdictions*, and the fourth and most important is the set of relationships in *social space* for individuals and their communities. The first dimension, *geographic space* here is intended as the physical space in which human action occurs. This dimension has played important roles in transportation planning and modeling because the first preoccupation of the transportation system designers has been to move persons from one location to another (i. e., overcoming spatial separation). Initial applications considered the territory divided into large areas (traffic analysis zones), represented by a virtual center (centroid), and connected by facilities (higher level highways). The centroids were connected to the higher level facilities using a virtual connector summarizing the characteristics of all the local roads within the zone. As computational power increased and the types of policies/strategies required increased resolution, the zone became smaller and smaller. Today, is not unreasonable to expect software to handle zones that are as small as a parcel of land and transportation facilities that are as low in the hierarchy as a local road (the centroid becomes the building on a parcel and the centroid connector is the driveway of the unit and they are no longer virtual).

In modeling and simulation we are interested in understanding human action. For this reason in some applications geographic space needs to consider more than just physical features (p. 387 in [49]) moving us into the notion of place and social space (see also below). The second dimension is *time* that is intended here as continuity of time, irreversibility of the temporal path, and the associated artificiality of the time period considered in many models. For example, models used in long range planning applications use typical days (e. g., a summer day for air pollution). In many regional long-range models the unspoken assumption is that we target a typical work weekday in developing models to assess policies. Households and their members, however, may not always (if at all) obey this strict definition of a typical weekday to schedule their activities and they may follow very different decision making horizons in allocating time to activities within a day, spreading activities among many days including weekends, substituting out of home with in home activities in some days but doing exactly the opposite on others, and using telecommunications only selectively (e. g., on Fridays and Mondays more often than on other days). Obviously, taking into account these scheduling activities is by far more complex than what is allowed in existing transportation planning models. The third dimension is *jurisdictions* and their interconnectedness. The actions of each person are "regulated" by jurisdictions with different and overlapping domains such as federal agencies, state agencies, regional authorities, municipal governments, neighborhood associations, trade associations and societies, religious groups, and formal and informal networks of families and friends. In fact, the federal government defines many rules and regulations on environmental protection. These may end up being enforced by a local jurisdiction (e. g., a regional office of an agency within a city). On the one hand, we have an organized way of governance that clearly defines jurisdictions and policy domains (e. g., tax collection in the US). On the other hand, however, the relationships among jurisdictions and decision making about allocation of resources does not follow always this orderly governance principle of hierarchy. A somewhat different and more "bottom up" relationship is found in the social network and for this reason requires a different dimension that is the fourth and final dimension named *social space* and the relationships among persons within this space. For example, individuals from the same household living in a neighborhood may change their daily time allocation patterns and location visits to accommodate and/or take advantage of changes in the neighborhood such as elimination of traffic and the creation of pedestrian zones. Depending on the effects of these changes on the pedestrian network we may also see a shift in the within the neighborhood social behavior. In contrast, increase in traffic to surrounding places may create an outcry by other surrounding neighborhoods, thus, complicating the relationships among the residents.

One important domain and entity within this social space is the household. This has been a very popular unit of analysis in transportation planning recognizing that strong relationships within a household can be used to capture behavioral variation (e. g., the simplest method is to use a household's characteristics as explanatory variables in a regression model of travel behavior). In this way any changes in the household's characteristics (e. g., change in the composition due to birth, death, or children leaving the nest or adults moving into the household) can be used to predict changes in travel behavior. New model systems are created to study this interaction within a household looking at the patterns of using time in a day and the changes across days and years. It is therefore very important in modeling and simulation to incorporate

in the models used for policy analysis interactions among these four fundamental dimensions, which bring us to the next major issue that of scale.

The typical long range planning analysis is usually defined for larger geographical areas (region, states, and countries) and addresses issues with horizons from 10 to 50 years. In many instances we may find that large geographic scale means also longer time frames applied to wider mosaics of social entities and including more diverse jurisdictions. On the other side of the spectrum issues that are relevant to smaller geographic scales are most likely to be accompanied by shorter term time frames applied to a few social entities that are relatively homogeneous and subject to the rule of very few jurisdictions. This is one important organizing principle but also an indicator of the complex relationships we attempt to recreate in our computerized models for decision support. In developing the blueprints of these models one can choose from a variety of theories (e. g., neoclassical microeconomics) and conceptual representations of the real world that help us develop these models. At the heart of our understanding of how the world (as an organization, a household, a formal or informal group, or an individual human being) works are models of decision making and conceptual representations of relationships among entities making up this world.

Transportation planning applications are about judgment and decision making of individuals and their organizations. There are different settings of decision making that we want to understand. Three of these settings are the travelers and their social units from which motivations for and constraints to their behavior emerge; the transportation managers and their organizations that serve the travelers and their social units, and the decision makers surrounding goods movement and service provision that contain a few additional actors, Southworth [151]. These include land use markets (see www.urbansim.org). Travelers received considerable attention in transportation planning and the majority of the models in practice aim at capturing their decision making process. The remaining settings received much less attention and they are poorly understood and modeled.

Conceptual models of this process are transformed into computerized models of a city, a region, or even a state in which we utilize components that are in turn models of human judgment and decision making (e. g., travelers moving around the transportation network and visiting locations where they can participate in activities). Models of this behavior are simplified versions of strategies used by travelers when they select among options that are directly related to their desired activities. In some of these models we also make assumptions about hierarchies of motivations, actions, and consequences. Some of these assumptions are explicit (e. g., when deriving the functional forms of models as in the typical disaggregate choice models or the rules in a production system) and in other models these assumptions are implicit.

When designing transportation planning model interfaces for transportation planners and managers we also implicitly make assumptions about the managers' ability to understand the input, agent representation, internal functioning, and output of these computerized models. Our objective is therefore not only to understand travel behavior and build models that describe and predict human behavior but also to devise tools that allow transportation managers to understand the assumed behavior in the models, study scenarios of policy actions, and define and explain policy implications to others. This, in essence, implies that we, the model system designers, create a platform for a relationship between planners and travelers. A similar but more direct relationship also exists between travelers and transportation managers when we design the observation methods that provide the data for modeling but also the data used to measure attitudes and opinions such as travel surveys. In fact, this relationship is studied in much more detail in the survey design context and linked directly to the image of the agency conducting the survey and the positive or negative impression of the travelers about the sponsoring agency [33]. Most transportation research for modeling and simulation, however, has emphasized traveler behavior when building surveys and their models neglecting the interface with the planners. The summary of theories below, however, applies to individuals traveling in a network but also to organizations and planners in the sense used by H.A. Simon in his Administrative Behavior [150].

Rational decision making is a label associated with human behavior that follows a strategy in identifying the best course of action. In summary, a decision maker solves an optimization problem and identifies the best existing solution to this problem. Within this more general strategy when an operational model is needed and this operational model provides quantitative predictions about human behavior some kind of mathematical apparatus is needed to produce the predictions. One such machinery is the subjective expected utility [146] formulation of human behavior. In developing alternative models to SEU Simon [149] defines four theoretical components:

- A person's decision is based on a utility function assigning a numerical value to each option – *existence and consideration of a cardinal utility function*;

- The person defines an exhaustive set of alternative strategies among which just one will be selected – *ability to enumerate all strategies and their consequences*;
- The person can build a probability distribution of all possible events and outcome for each alternate option – *infinite computational ability*; and
- The person selects the alternative that has the maximum utility – *maximizing utility behavior*.

This behavioral paradigm served as the basis for a rich production of models in transportation that include the mode of travel, destinations to visit as well as the household residence (see the examples in the seminal textbook by Ben-Akiva and Lerman [9]. It served also as the theoretical framework for consumer choice models and for attempts to develop models for hypothetical situations (see the comprehensive book by Louviere, Hensher, and Swait [108]. It has also replaced the aggregate modeling approaches to travel demand analysis as the orthodoxy against which many old and new theories and applications are compared and compete with. SEU can be considered to be a model from within a somewhat larger family of models under the label of weighted additive rule (WADD) models [127]. Real humans, however, may never behave according to SEU or related maximizing and infinitely computational capability models (Simon labels this the Olympian model, [149]). Based on exactly this argument different researchers in psychology have proposed a variety of decision making strategies (or heuristics). For example, Simon created alternate model paradigms under the label of *bounded rationality – the limited extent to which rational calculation can direct human behavior* [149,150] to depict a sequence of a person's actions when searching for a suitable alternative. The modeled human is allowed to make mistakes in this search giving a more realistic description of observed behavior (see also Rubinstein [142]). Tversky is credited with another stream of decision making models starting with the *lexicographic approach* [86, in which *a person first identifies the most important attribute, compares all alternatives on the value of this attribute, and chooses the alternative with the best value on this most important attribute*. Ties are resolved in a hierarchical system of attributes. Another Tversky model [169] assumes *a person selects an attribute in a probabilistic way and influenced by the importance of the attribute, all alternatives that do not meet a minimum criterion value (cutoff point) are eliminated*. The process proceeds with all other attributes until just one alternative is left and that one is the chosen. This has been named the *elimination by aspects strategies* (EBA) model. Later, Kahneman and Tversky [170] developed *prospect theory* and its subsequent version of *cumulative prospect theory* in Tversky and Kahneman [171] in which a simplification step is first undertaken by the decision maker editing the alternatives. Then, a value is assigned to each outcome and *a decision is made based on the sum of values multiplying each by a decision weight*. Losses and gains are treated differently. All these alternatives to SEU paradigms did not go unnoticed in transportation research with early significant applications appearing in the late 1980s. In fact, a conference was organized attracting a few of the most notable research contributors to summarize the state of the art in behavior paradigms and documented in Garling, Laitila, and Westin [45]. One of the earlier examples using another of Simon's inventions, the *satisficing behavior – acceptance of viable choices that may not be optimal* – is a series of transportation-specific applications described in Mahmassani and Herman [110]. Subsequent contributions continue along the path of more realistic models and the most recent example, discussing a few models, by Avineri and Prashker [7], uses cumulative prospect theory giving a preview of a movement toward more realistic travel behavior models. As Garling et al. [45] and Avineri and Prashker [7] point out, these paradigms are not ready for practical applications, contrary to the Mahmassani and colleagues efforts that have been applied, and additional work is required to use them in a simulation framework for applications. Another aspect is contextual *adaptation*. Payne, Bettman, and Johnson [127] provide an excellent review of decision making models and their differentiating aspects. They also provide evidence that decision makers *adapt* by switching between decision making paradigms *to the task and the context of their choices*. They also make mistakes and they may also fail to switch strategies. As Vause [175] discusses to some length transportation applications are possible using multiple decision making heuristics within the same general framework and employing a production system approach [123]. A key consideration, however, that has received little attention in transportation is the definition of context within which decision making takes place. Recent production systems [5] are significant improvements over past simulation techniques. However, travelers are still assumed to be passive in shaping the environment within which they decide to act (action space). This action space is viewed as largely made by constraints and not by their active shaping of their context. Goulias [58,60] reviews another framework from human development that is designed to treat decision makers in their active and passive roles and explicitly accounts for mutual influence between an agent (active autonomous decision maker) and her environment.

Transportation modeling and simulation experienced a few tremendously innovative and progressive steps forward. Interestingly these key innovations are from non-engineering fields but very often transferred and applied to transportation systems analysis and simulation by engineers. These are listed here in a somewhat sequential chronological order merging technological innovations and theoretical innovations. At exactly the time that the Bay Area Rapid Transit system was studied and evaluated in the 1960s, Dan McFadden (the Year 2000 Nobel Laureate in Economics) and a team of researchers produced practical mode choice regression models at the level of an individual decision maker (see http://emlab.berkeley.edu/users/mcfadden/ – accessed June 2007). The models are based on random utility maximization (of the SEU family) and their work opened up the possibility to predict mode choice rates more accurately than ever before. These models were initially named *behavioral travel-demand models* [155] and later the more appropriate term of *discrete choice models* [9] prevailed. Although restrictive in their assumptions, these models are still under continuous improvement and they have become the standard tool in evaluating discrete choices. Some of the most notable and recent developments advancing the state of the art and practice are:

- Better understanding of the theoretical and particularly behavioral limitations of these models [45,50,115];
- more flexible functional forms that resolve some of the problems raised in Williams and Ortuzar [184] allowing for different choices to be correlated when using the most popular discrete choice regression models [14,16,95];
- combination of revealed preference, stated choices by travelers, with stated preferences and intentions, answers to hypothetical questions by travelers, availability of data in the same choice framework to extract in a more informative way travelers willingness to use a mode and willingness to pay for a mode option [10,108]. This latter "improvement" enables us to assess situations that are impossible to build in the real world;
- computer-based interviewing and laboratory experimentation to study more complex choice situations and the transfer of the findings to the real world [111]. This direction, however, is also accompanied by a wide variety of research studies aiming at more realistic behavioral models that go beyond mode choice and travel behavior [50]; and
- expansion of the discrete choice framework using ideas from *latent class models* with covariates that were first developed by Lazarsfeld in the 1950s and their estimation finalized by Goodman in the 1970s (see the review in [56], and discrete choice applications in [20]). This family of models was used in Goulias [57] to study the dynamics of activity and travel behavior and in the study of choice in travel behavior [12].

As mentioned earlier the rational economic assumption of the maximum utility model framework (that underlies many but not all of the disaggregate models) is very restrictive and does not appear to be a descriptive behavioral model except for a few special circumstances when the framing of decisions is carefully designed (something we cannot expect to happen every time a person travels on the network). Its replacement, however, requires conceptual models that can provide the types of outputs needed in regional planning applications. A few additional research paths, labeled as *studies of constraints*, are also functioning as gateways into alternate approaches to replace or complement the more restrictive utility-based models. A few of these models also consider knowledge and information provision to travelers. The first aspect we consider is about the choice set in discrete choice models. Choice set is the set of alternatives from which the decision maker selects one. These alternatives need to be mutually exclusive, exhaustive, and finite in number [166]. Identification, counting, and issues related to the alternatives considered have motivated considerable research in choice set formation [77,78,140,158,159]. Key threat to misspecification of the choice set is the potential for incorrect predictions [161]. When this is an issue of considerable threat as in destination choice models where the alternatives are numerous, a model of choice set formation appears to be the additional burden [71]. Other methods, however, also exist and they may provide additional information about the decision making processes. Models of the processes can be designed to match the study of specific policies in specific contexts. One such example and a more comprehensive approach defining the choice sets is the situational approach [25]. The method uses in depth information from survey respondents to derive sets of reasons for which alternatives are not considered for specific choice settings (individual trips). This allows separation of analyst observed system availability from user perceived system availability (e. g., due to misinformation and willingness to consider information). This brings us to the duality between "objective choice attributes" and "subjective choice attributes". Most transportation applications, independently of the decision making paradigm adopted, assume the analysts (modelers) and the travelers (modeled) measured attributes to be the same. Modeling the process of perceived constraints may be far more complex when one

considers the influence of the context within which decisions are made. Golledge and Stimpson (pp. 33–34 in [49]) describe this within a conceptual model of decision making that has a cognitive feel to it. They also link the situational approach to the activity-based framework of travel extending the framework further (pp. 315–328 in [49]).

Chapin's research [28], providing one of the first comprehensive studies about time allocated to activity in space and time, is also credited for motivating the foundations of activity-based approaches to travel demand analysis. His focus has been on the propensity of individuals to participate in activities and travel linking their patterns to urban planning. In about the same period Becker also developed his theory of time allocation from a household production viewpoint [8] applying economic theory in a non-market sector and demonstrating the possibility of formulating time allocation models using economics reasoning (i. e., activity choice). In parallel another approach was developing in geography and Hagerstrand's seminal publication on time geography [72] presents the foundations of the approach. The idea of constraints in the movement of persons was taken a step further by this time-geography school in Lund. In that framework, the movement of persons among locations can be viewed as their movement in space and time under external constraints. Movement in time is viewed as the one way (irreversible) movement in the path while space is viewed as a three dimensional domain. It provides the third base about *constraints* in human paths in time and space for a variety of planning horizons. These are *capability constraints* (e. g., physical limitations such as speed); *coupling constraints* (e. g., requirements to be with other persons at the same time and place); and *authority constraints* (e. g., restrictions due to institutional and regulatory contexts such as the opening and closing hours of stores). Figure 1 provides a pictorial representation in space and time of a typical activity-travel pattern of two persons (P1 and P2) and the three types of constraints. H indicates home, W indicates work, L indicates leisure, and S indicates shopping.

Cullen and Godson [31] also reviewed by Arentze and Timmermans [5] and Golledge and Stimpson [49] appear to be the first researchers attempting to bridge the gap between the motivational (Chapin) approach to activity participation and the constraints (Hagerstrand) approach by creating a model that depicts a routine and deliberated approach to activity analysis. The Cullen and Dobson study also defined many terms often used today in activity-based approaches. For example, each activity (stay-home, work, leisure, and shopping) is an episode characterized by start time, duration, and end time. Activities are also classified into fixed and flexible and they can be engaged alone or with others. Moreover, they also analyzed sequencing of activities as well as pre-planned, routine, and on the spur of the moment activities. Within this overall theoret-

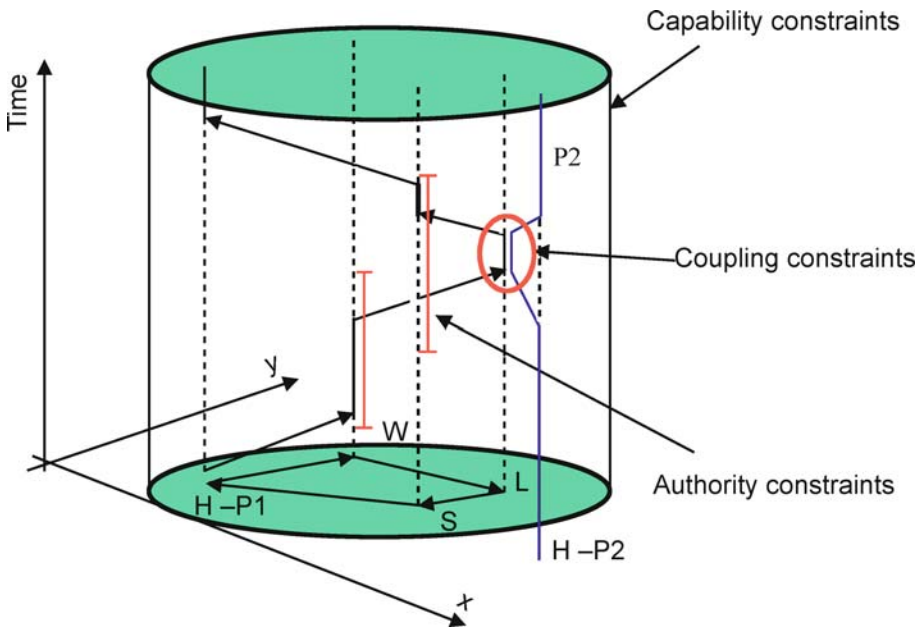

Travel Behavior and Demand Analysis and Prediction, Figure 1
A two-person (P1 and P2) activity-travel pattern and the time and space limits imposed by constraints [132]

ical framework is the idea of a project which according to Golledge and Stimpson [49], *is a set of linked tasks that are undertaken somewhere at some time within a constraining environment* (pp. 268–269). This idea of the project underlies one of the most exciting developments in activity-based approaches to travel demand analysis and forecasting because seemingly unrelated activity and trip episodes can be viewed as parts of a "big-picture" and given meaning and purpose completing in this way models of human agency and explaining resistance to change behavior.

Most subsequent contributions to the activity-based approach emerge in one way or another from these initial frameworks with important operational improvements (for reviews see [5,17,89,114]). The basic ingredients of an activity based approach for travel demand analysis [5,84] are:

a) explicit treatment of travel as derived demand [112], i. e., participation in activities such as work, shop, and leisure motivate travel but travel could also be an activity as well (e. g., taking a drive). These activities are viewed as episodes (i. e., they are characterized by starting time, duration, and ending time) and they are arranged in a sequence forming a pattern of behavior that can be distinguished from other patterns (i. e., a sequence of activities in a chain of episodes). In addition, these events are not independent and their interdependency is accounted for in the theoretical framework;
b) the household is considered to be the fundamental social unit (i. e., decision making unit) and the interactions among household members are explicitly modeled to capture task allocation and roles within the household, relationships at one time point and change in these relationships as households move along their life cycle stages and the individual's commitments and constraints change and these are depicted in the activity-based model; and
c) explicit consideration of constraints by the spatial, temporal, and social dimensions of the environment is given. These constraints can be explicit models of time-space prisms [130] or reflections of these constraints in the form of model parameters and/or rules in a production system format [5].

Input to these models are the typical regional model data of social, economic, and demographic information of potential travelers and land use information to create schedules followed by people in their everyday life. The output are detailed lists of activities pursued, times spent in each activity, and travel information from activity to activity (including travel time, mode used, and so forth). This output is very much like a "day-timer" for each person in a given region. Figure 2 provides an example of time allocation to different activities from an application that collected activity participation data [2,3]. It displays time allocation by one segment of the population showing the proportion of persons engaging in each activity by each hour of a day. Figure 3 shows the output from a model that predicts the presence of persons in each building during each hour of a day engaging in each activity type. Combining an activity model with a typical travel demand model produces "volumes" of individuals at specific locations and on the network of a city as shown in Figure 4 (a more detailed description of this study can be found in [67,97,99]).

Many planning and modeling applications, however, aim at forecasting. Inherent in forecasting are the time changes in the behavior of individuals and their households and their response to policy actions. At the heart of behavioral change are questions about the process followed in shifting from a given pattern of behavior to another. In addition to measuring change and the relationships among behavioral indicators that change in their values over time, we are also interested in the timing, sequencing, and staging of these changes. Moreover, we are interested in the triggers that may accelerate desirable or delay undesirable changes and the identification of social and demographic segments that may follow one time path versus another in systematic patterns. Knowledge about all this is required to design policies but it is also required to design better forecasting tools. Developments in exploring behavioral dynamics and advancing models for them have progressed in a few arenas. First, in the data collection arena with panel surveys, repeated observation of the same persons over time that are now giving us a considerable history in developing new ideas about data collection but also about data analysis [55,61] and interactive and laboratory data collection techniques [34] that allow a more in-depth examination of behavioral processes. The second arena is in the development of microeconomic dynamic formulations for travel behavior that challenge conventional assumptions and offer alternative formulations [91]. The third arena, is in the behavior from a developmental viewpoint as a single stochastic process, a staged development process [57], or as the outcome from multiple processes operating at different levels [59]. Experimentation with new theories from psychology emphasizing development dynamics is a potential fourth area that is just beginning to emerge [60]. Behavioral dynamics are also examined using more comprehensive analyzes [68] and models [136].

These models focus more on the paths of persons in space and time within a somewhat short time horizon such a day, a week, or maybe a month. The consideration of

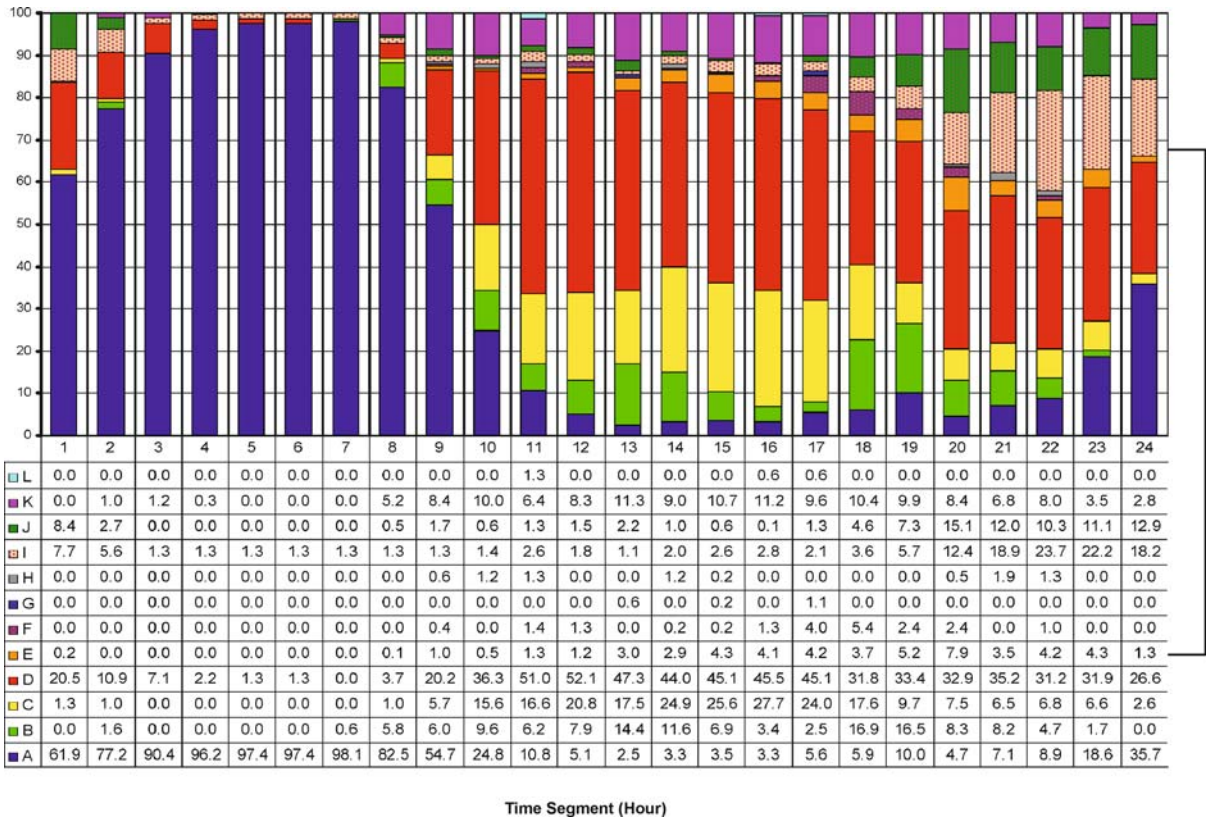

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
L	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.3	0.0	0.0	0.0	0.0	0.6	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0
K	0.0	1.0	1.2	0.3	0.0	0.0	0.0	5.2	8.4	10.0	6.4	8.3	11.3	9.0	10.7	11.2	9.6	10.4	9.9	8.4	6.8	8.0	3.5	2.8
J	8.4	2.7	0.0	0.0	0.0	0.0	0.0	0.5	1.7	0.6	1.3	1.5	2.2	1.0	0.6	0.1	1.3	4.6	7.3	15.1	12.0	10.3	11.1	12.9
I	7.7	5.6	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.4	2.6	1.8	1.1	2.0	2.6	2.8	2.1	3.6	5.7	12.4	18.9	23.7	22.2	18.2
H	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	1.2	1.3	0.0	0.0	1.2	0.2	0.0	0.0	0.0	0.0	0.5	1.9	1.3	0.0	0.0
G	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.2	0.0	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
F	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	1.4	1.3	0.0	0.2	0.2	1.3	4.0	5.4	2.4	2.4	0.0	1.0	0.0	0.0
E	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.1	1.0	0.5	1.3	1.2	3.0	2.9	4.3	4.1	4.2	3.7	5.2	7.9	3.5	4.2	4.3	1.3
D	20.5	10.9	7.1	2.2	1.3	1.3	0.0	3.7	20.2	36.3	51.0	52.1	47.3	44.0	45.1	45.5	45.1	31.8	33.4	32.9	35.2	31.2	31.9	26.6
C	1.3	1.0	0.0	0.0	0.0	0.0	0.0	1.0	5.7	15.6	16.6	20.8	17.5	24.9	25.6	27.7	24.0	17.6	9.7	7.5	6.5	6.8	6.6	2.6
B	0.0	1.6	0.0	0.0	0.0	0.0	0.6	5.8	6.0	9.6	6.2	7.9	14.4	11.6	6.9	3.4	2.5	16.9	16.5	8.3	8.2	4.7	1.7	0.0
A	61.9	77.2	90.4	96.2	97.4	97.4	98.1	82.5	54.7	24.8	10.8	5.1	2.5	3.3	3.5	3.3	5.6	5.9	10.0	4.7	7.1	8.9	18.6	35.7

Travel Behavior and Demand Analysis and Prediction, Figure 2

Time allocation to different activities in a day [2]. A: Personal Needs (includes sleep), B: Eat meal, C: Paid work, D: Education, E: Household and family care, F: shopping, G: medical, H: Volunteering/Community, I: Socializing, J: Sports and Hobbies, K: Travel, L: All other

behavioral dynamics has expanded the temporal horizons to a few years. However, regional simulation models are very often designed for long range plans spanning 25 years or even longer time horizons. Within these longer horizons, changes in the spatial distribution of activity locations and residences (land use) are substantial, changes in the demographic composition and spatial distribution of demographic segments are also substantial, and changes in travel patterns, transport facilities, and quality of service offered can be extreme. Past approaches in modeling and simulating the relationship among land use, demographics, and travel in a region attempted to disengage travel from the other two treating them as mutually exogenous. As interactions among them became more interesting and pressing, due to urban sprawl and suburban congestion, increasing attention was paid to their complex interdependencies. This led to a variety of attempts to develop "integrated model systems" that enable the study of scenarios of change and mutual influence between land use and travel. An earlier review of these models with heavy emphasis on discrete choice models can be found in Anas [4]. Miller [117] and Waddell and Ulfarsson [180] twenty years later provide two comprehensive reviews of models that have integrated many aspects in the interdependent triad of demographics-travel-land use models. Both reviews trace the history of some of the most notable developments and both link these models to the activity-based approach above. Both reviews also agree that a microeconomic and/or macroeconomic approach to modeling land and transportation interactions are not sufficient and more detailed simulation of the individuals and their organizations "acting" in an time-space domain need to be simulated in order to obtain the required output for informed decision making. They also introduce the idea of simulating interactive agents in a dynamic environment of other agents (multi-agent simulation). The vast literature is reviewed by Timmermans [163] and Miller [118], from different viewpoints about progress made until now. However, they both agree that progress is rapidly made and that integration of land use and transportation models needs

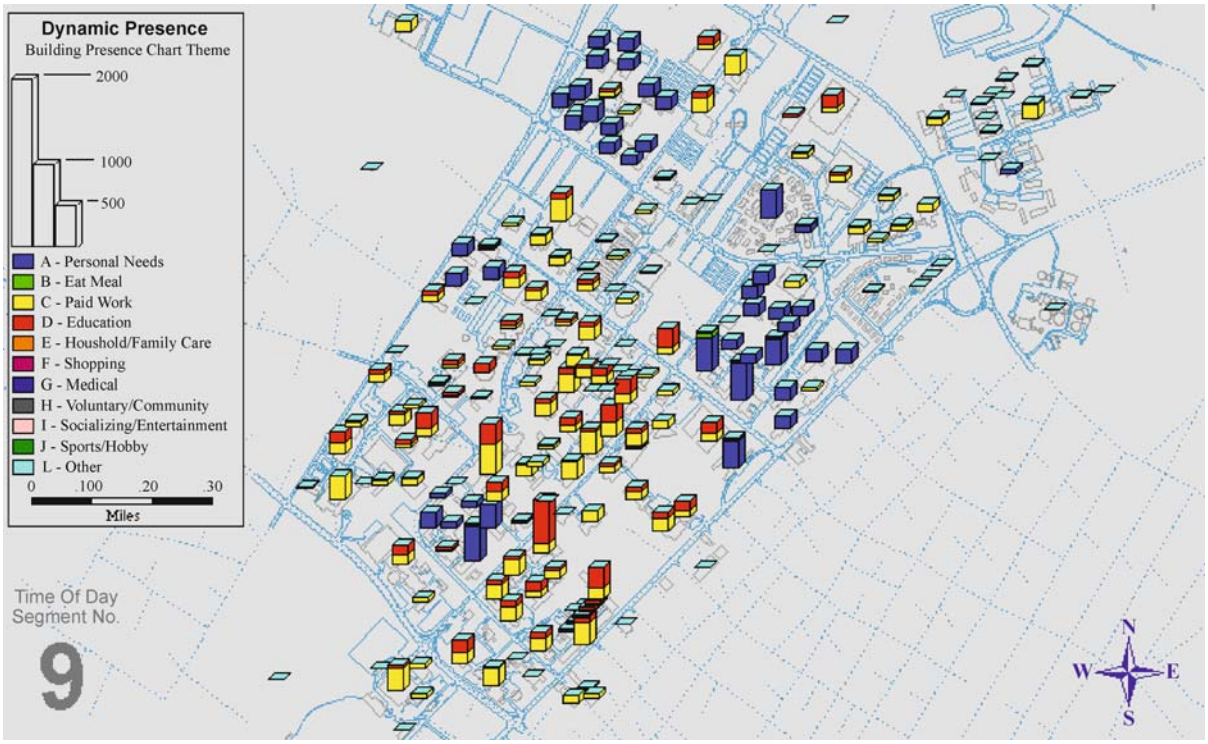

Travel Behavior and Demand Analysis and Prediction, Figure 3
Persons and activities assigned to buildings [2]

to move forward. Creation of integrated systems is further complicated by the emergence of an entire infrastructural system as another layer of human activity – telecommunication. Today telecommunication and transportation relationships are mostly absent from regional simulation planning and modeling as well from the most advanced land use and transportation integrated models. Considerable research findings, however, have been accumulating since the 1970s [53,66,81,96,111,113,121,128,129,144,182]. Another type of technologies (named enabling herein) helped us move modeling and simulation further.

A few of the most important technologies are *stochastic simulation*, *production systems*, *geographic information systems*, *interactive and technology-aided data collection approaches*, and more *flexible data analysis techniques*.

Stochastic microsimulation, as intended here, is an evolutionary engine software that is used to replicate the relationships among social, economic, and demographic factors with land use, time use, and travel by people. As discussed above the causal links among these groups of entities are extremely complex, non-linear, and in many instances unknown or incompletely specified. This is the reason that no closed form solution can be created for such a forecasting model system. An evolutionary engine, then, provides a realistic representation of person and household life histories (e. g., birth, death, marriages, divorces, birth of children, etc.), spatio-temporal activity opportunity evolution, and a variety of models that account for uncertainties in data, models, and behavioral variation (see [59,117], for overviews and [157] for an application).

Production systems were first developed by Newell and Simon [123] to explicitly depict the way humans go about solving problems. These are a series of condition-action (note the parallel with stimulus-response) statements in a sequence. From this viewpoint they are search processes that may never reach an absolute optimum and they replicate (or at least attempt to) human thought and action. Models of this kind are called *computational process models* (CPM) and through the use of IF ... THEN... rules have made possible the creation of a variety of new models.

Geographic information systems are software systems that can be used to collect, store, analyze, modify, and display large amounts of geographic data. They include layers of data that are able to incorporate relations among the variables in each layer and allow to build relationships in data across layers. One can visualize a GIS as a live map

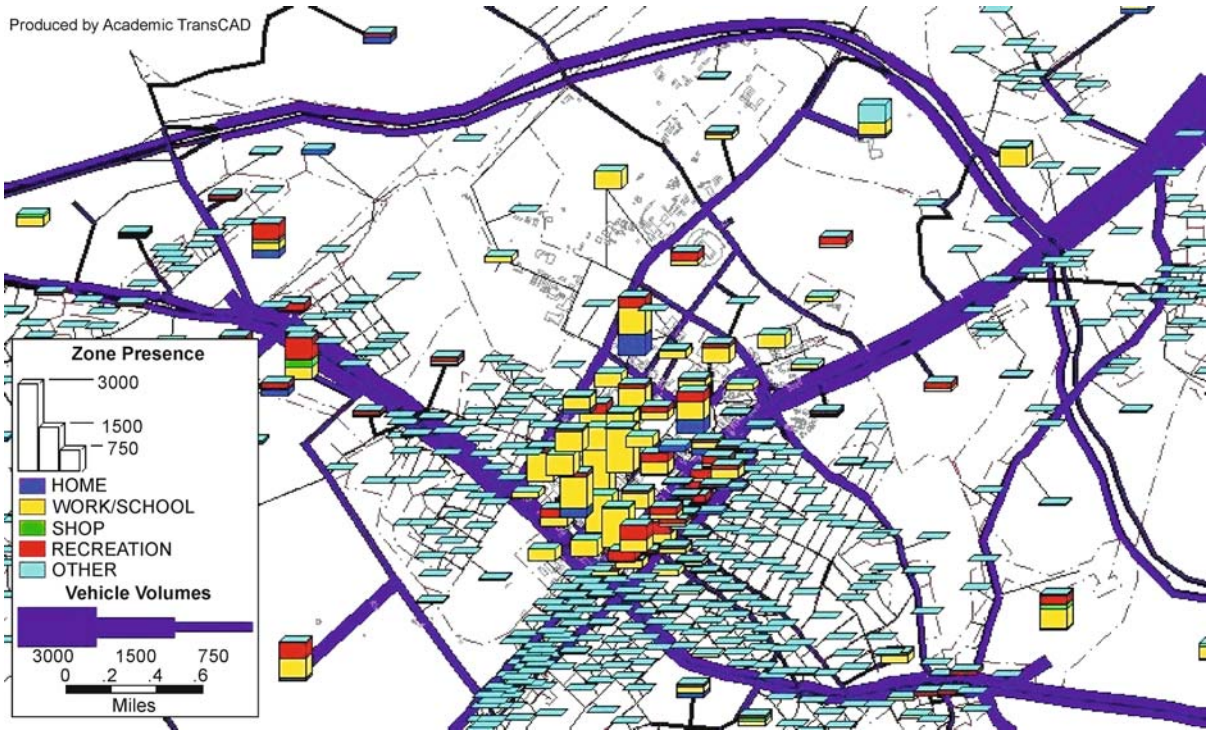

Travel Behavior and Demand Analysis and Prediction, Figure 4
Persons and activities assigned to buildings and travel to the network [67]

that can display almost any kind of spatio-temporal information. Maps have been used by transportation planners and engineers for long time and they are a natural interface to use in modeling and simulation. GIS today is moving beyond this relational database definition and is transforming the entire field into GI Science, which is beyond the scope of this article.

Advanced data collection methods and devices that are technologies that merit a note, although, not strictly developed for modeling. The first is about data collection and particularly data collection using internet technologies to build complex interviews that are interactive and dynamic [34]. In the same line of development we also see the use of geographic positioning systems (GPS) that allow one to develop a trace of individual paths in time and space [35,186]. Very important development is also the emergence of devices that can record the bulk of environmental data surrounding a person's movement, classify the environment in which the individual moves, and then ask simplified questions [74].

Soft computing and non-parametric data analysis is the last innovation mentioned here. In the data analysis we see greater strides in using data mining and artificial intelligence-born techniques to extract travel behavior patterns [134,160] and advanced and less restrictive statistical methods to discover relationships in travel behavior data (e. g., [88]). Soft computing is increasingly finding many applications in activity-based models (see www.imob.uhasselt.be). For a more recent and accessible review see Pribyl [133].

The Evolving Modeling Paradigm

Policies are dictating to create and test increasingly more sophisticated policy assessment instruments that account for direct and indirect effects of behavior, procedures for behavioral change, and to provide finer resolution in the four dimensions of geographic space, time, social space, and jurisdictions. Dynamic planning is also stressing the need to examine trends, cycles, and the inversion of time progression to develop paths from the future visions to today's actions. New model developments are also becoming increasingly urgent. Although, tremendous progress has been observed in the past 20 years, development requires a faster pace to create new policy tools. These policy tools need to disentangle the actions of persons under different policy actions and the impact of policy actions on aggregates to identify conflicts and resolutions. Supporting all

this is a rich collection of decision paradigms that are already used and a few new ideas are starting to migrate to practice as illustrated below.

Early models incorporating activity-based behavioral processes into applications were published in the late 1970s and early 1980s as proof-of-concept and experimental applications. Following Hagerstrand's time-geography approach, PESASP [103] is one of the first models to operationally show the use of a time-space prism and to account for the relationship among activities. The Cullen and Godson [31] study was also the first comprehensive treatment of activities that brought different research findings together. In parallel, models were developed that were utility-maximizing models such as Adler and Ben-Akiva model [1] and much later the Kawakami and Isobe model [87]. Following these studies, BSP [79] and Computational Algorithms for Rescheduling Lists of Activities – CARLA [83] also use the activities within a time-space prism paradigm and define the foundations of data collection for activity-based approaches.

After this period of experimentation three streams in model development emerged. The first is in deriving representative activity patterns (RAPs) and then using regression techniques to correlate RAPs to person and household social and demographic data and then forecasting. The second development refines the methods used to simulate persons and adds to the forecasting repertoire other forecasting tasks via *microsimulation*. The third is a movement that expands the envelope to include cognition and explicit representation of mental processes through CPMs.

The Simulation of Travel/Activity Responses to Complex Household Interactive Logistic Decisions (STARCHILD – Recker and McNally [138,139] derived RAPs, employed a utility-based model and incorporated constraints. It is considered a fundamental transition development from research to practical application of an activity-based approach and it is still the foundation of models that first derive representative patterns and then forecast travel behavior. The more recent SIMAP [100] is a direct derivation of STARCHILD. In this line of development, Ma [109] created a model system that combined long term activity patterns (Long-term activity and travel planning – LATP) with a within-a-day activity scheduling and simulation (Daily Activity and Travel Scheduling – DATS) incorporating day-to-day variation and history dependence. Her model system produced very accurate forecasts. However it required panel survey data (the repeated observation of the same persons and households over time) that are rarely collected. In the LATP/DATS system longitudinal statistical models are extracted from longitudinal records and they capture important aspects of behavioral dynamics such as habit persistence, day-to-day switching behaviors, and account for observed and unobserved heterogeneity contributed by the person, the household, the area of residence, and the area of workplace.

One of the first models to include a microsimulation in its paradigm is ORIENT [152]. This methodology suitably refined was demonstrated in a countrywide model for the Netherlands developed between 1989 and 1991 and named the Microanalytic Integrated Demographic Accounting System (MIDAS – Goulias and Kitamura [63,64]). MIDAS integrates demographic microsimulation, with dynamic car ownership models and a comprehensive suite of travel behavior equations. A cross-sectional version of MIDAS using data from the United States was also developed by Chung and Goulias [29]. MIDAS-USA simulates the evolution of households along with car ownership and travel behavior for Centre County, PA, and it is linked to a model to assign fees for development using GIS. A more ambitious development is the Activity Mobility Simulator – AMOS – by Kitamura et al. [93], which defines a few RAPs as templates. Then, uses a neural network to identify choices and a satisficing rule to simulate schedule changes due to policies. While MIDAS is a strictly longitudinal process econometric model progressing one year at a time, AMOS is constraint-based model designed for much finer temporal resolution. DEMOS, developed by Sundararajan and Goulias [157], is another MIDAS derivative. DEMOS is an object-oriented environment designed to simulate the evolution of people and their households using a variety of external data with the core models based on the Puget Sound Transportation Panel. It also simulates activity participation, travel, and telecommunication market penetration using a few representative patterns that were derived in Ma's LATP/DATS supplemented by telecommunications and travel behavior models.

SCHEDULER (Gärling et al. [43] is the first CPM that adds a psychometric cognitive implementation based on the Hayes-Roth and Hayes-Roth [73] planning model. In SCHEDULER, activities, selected from the long term calendar that represents a person's long term memory, comprise a schedule that is "mentally executed". Models start to combine CPM, microsimulation, and data derived behavioral patterns with random utility models to fill different modeling needs. The Simulation Model of Activity Scheduling Heuristics (SMASH – Ettema et al. [38]) is a CPM and econometric utility-based hybrid model that focuses on the pre-trip planning process predicting sequences of activities. In parallel, COMRADE [37], uses competing risk hazard models for activity scheduling and incorporates duration models in the system. The Model of

Action Space in Time Intervals and Clusters (MASTIC – Dijst and Vidakovic [32]), identifies clusters in the action space to perform and schedule activities. Time-space prisms are also the foundation of the Prism-Constrained Activity-Travel Simulator (PCATS – Kitamura [90], Kitamura and Fujii [92]), which is also a utility-based model. A direct operational derivative of SCHEDULER [44] was developed by Kwan, in her 1994 dissertation [101,102], and named GIS-Interfaced Computational-process modeling for Activity Scheduling (GISICAS). It is a simplified CPM, that uses time-space constraints and GIS to incorporate spatial information into a behavioral model to create individual schedules, starting with activities at higher levels of priority. Other models also attempt to recreate personal schedules such as Vause's model [175], a CPM that creates a restricted choice set for creating activity patterns, a model by Ettema [39], and VISEM [41], a data-driven model that is a part of PTV Vision, an urban and regional transportation planning system, that creates daily activity patterns for behaviorally homogeneous groups within the population. Stopher et al. [156] also proposed the Simulation Model for Activity Resources and Travel (SMART) using a time geography framework and a taxonomy of activities in a GIS environment. All these use observed patterns to derive behavioral models. In contrast, Recker [137], developed the Household Activity Pattern Problem (HAPP) as a normative model based on the pick up and delivery time window problem to be used as a yardstick model testing optimal behavioral hypotheses.

The model framework that impacted practice the most in the United States is the Daily Activity Schedule model by Ben-Akiva et al. in [11]. This model, was used to create the Portland Daily Activity Schedule Model [23], advocating modeling lifestyle and mobility decisions on a scale of years. These influence daily activity schedules, which are comprised of primary and secondary tours constrained in time and space. It contains two key elements that simplify activity-based model development and takes advantage of the research surge in developing more general discrete choice models. A similar simplification using conditional probabilities was also developed for Los Angeles by Kitamura et al. [94].

Figure 5 shows this hierarchy of decisions and the scheme used to convert the daily pattern into a system of discrete choices. This framework was used to design new models for the regions around San Francisco, New York, Columbus, Denver, Atlanta, and Sacramento [24].

Arentze and Timmermans [5] designed the most complete CPM named ALBATROSS, which is a multi-agent simulation and predicts the time, location, duration, activity companionship, and travel modes subjecting everything to spatio-temporal, institutional, and household constraints. The theoretical underpinnings of this model are by far wider and all encompassing than any other activity-based model. However, it does not simulate route choice and does not produce data suitable for traffic assignment algorithms. Development of the third version of ALBATROSS is currently underway [76]. This model is also representative of raising the ambitions of travel modelers. The Alam Penn State Emergency Management model (Alam-PSEM, Alam and Goulias [3]) is a building-by-building simulation of activity participation and presence at specific locations of a university campus for each hour of a typical day. In parallel Bhat and his co-workers [15,18] developed the Comprehensive Activity-Travel Generation System for Workers (CATGW), which is a se-

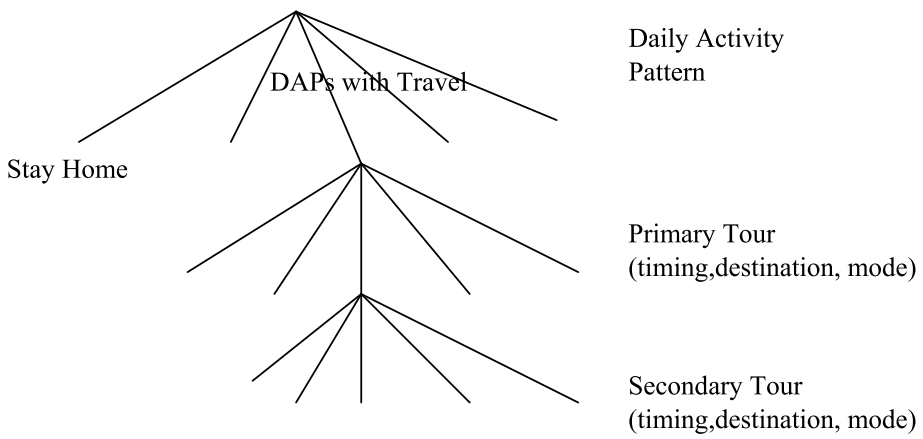

Travel Behavior and Demand Analysis and Prediction, Figure 5
The Bowman and Ben-Akiva daily activity model formulation [23]

ries of econometric models that replicate a commuter's evening mode choices, number of evening commute stops, and the number of stops after arriving home. The models developed by Bhat and colleagues are characterized by the use of hazard/duration regression models that were specifically developed for activity-based approaches and are by far more flexible that other regression methods. Another econometric model, the Conjoint-Based Model to Predict Regional Activity Patterns (COBRA), developed by Wang and Timmermans in [181], generates general patterns of stops for specific activities using a conjoint-based model with stated preference data instead of travel or activity diary data. The Wen and Koppelman model [183] utilizes three layers of decisions that are influenced by exogenous variables to generate activity patterns.

All these models point to new directions such as spatial choice needs to be dealt in more detail [3], activity choice and duration need to be dealt in a way the recognizes satiation in activity participation (e. g., in the duration models of Bhat [15]), sooner of later we will need to account for unobserved patterns and lack of experimental data (e. g., using conjoint experiments Wang and Timmermans [181]), and relations within the household need to also receive attention and inserted in the model hierarchy [183].

Spatial aspects of model development were considered in the CentreSIM regional model [67,98,99] that uses time-of-day activity and travel data for different market segments to predict hour-by-hour presence at locations and travel among zones. In 2004, as a part of the Longitudinal Integrated Forecasting Environment (LIFE) framework [58], Pribyl and Goulias [135] developed CentreSIM (medoid simulation) to derive a few representative patterns and simulate daily schedules accounting explicitly for within-household interactions for entire daily patterns. In the Netherlands, PATRICIA (Predicting Activity-Travel Interdependencies with a Suite of Choice-Based, Interlinked Analyzes), was developed by Borgers et al. [22] to help assess the performance of ALBATROSS. PATRICIA is a suite of linked models that incorporates an expanded set of activity choices, based on 63 distinct patterns, and activity destinations and describes activity transport modes and sequences. AURORA [82,165], which is a complementary model to ALBATROSS, is a utility-based system that models the dynamics of activity scheduling and rescheduling decisions as a function of many choice facets. AURORA is for short-term adaptation and rescheduling using just a few critical parameters. The model has since been expanded to include many new facets [76]. A much simpler model is PETRA [42] that allows the model to work with a small number of daily travel patterns with some statistical advantages (see also Henson et al. [76]). Microsimulation software experienced another push forward by the development of a multi-million investment in TRansportation ANalysis SIMulation System. This model system was developed in the decade 1995–2005 and one of its versions is now available via a NASA open source license from TMIP at http://tmip.fhwa.dot.gov/transims/. TRANSIMS is a survey data-driven cellular automata microsimulation and was developed by a team at Los Alamos National Laboratory [106]. It was one of the first simulation packages to contain models that create a synthetic population, generate activity plans for individuals using directly observed data in travel surveys, formulate routes on a network based on these, and execute activity plans. Microsimulation models also evolved in the interface between land use and travel behavior. The Integrated Land Use, Transportation and Environment (ILUTE) model [145] model is designed to simulate the evolution of people and their activity patterns, transportation networks, houses, commercial buildings, the economy, and the job market over time. Within this vision, Miller and Roorda [119], developed the Toronto Area Scheduling model for Household Agents (TASHA) that uses *projects* to organize activity episodes into schedules of persons. Schedules for members in a household are simultaneously generated to allow for joint activities. Both ILUTE and TASHA utilize CPMs and econometric utility-based paradigms.

Another microsimulation that uses econometric models to simulate daily activity travel patterns for an individual, is the Comprehensive Econometric Microsimulator for Daily Activity-travel Patterns (CEMDAP) model [19] that is based on land use, socio-demographic, activity system, and level-of-service (LOS) attributes. Key distinctive element of CEMDAP is its reliance on hazard-based regression models to account for the continuous nature time of activity duration. It includes population synthesis as well as the activity-pattern generation and scheduling of children, which is missing form many other simulators. Another model that utilizes constraints is the Florida Activity Mobility Simulator (FAMOS) [131]. FAMOS encompasses two modules, the Household Attributes Generation System (HAGS) and PCATS. Together, they comprise a system for modeling the activity patterns of individuals in Florida. The output is a series of activity-travel records. FAMOS is currently being further enhanced to include intra-household interactions and capture task allocation behavior among household members. Most recently, Ettema et al. [40] developed PUMA (Predicting Urbanization with Multi-Agents), a full-fledged multi-agent system of urban processes that represents land use changes

in a behaviorally realistic way. These processes include the evolution of population, businesses, and land use as well as daily activity and travel patterns of people. To simulate activity-travel patterns, an updated version of AURORA by Arentze et al. [6] will be created and also in the model FEATHERS (Forecasting Evolutionary Activity-Travel of Household and their Environmental Repercussions) to simulate activity-level scheduling decisions, within-a-day rescheduling, and learning processes in high resolutions of time and space. Developed as a complement to ALBATROSS, FEATHERS is econometric utility-based microsimulation that utilizes constraints that focuses on the short-term dynamics of activity-travel patterns. Members from this same Dutch team also developed MERLIN [173] and RAMBLAS [176].

Microsimulations have continued to gain in popularity in the activity-based modeling universe as they move from research applications to practice. Besides the Portland Daily Activity Schedule Model mentioned previously, New York's "Best Practice" Model (2002) and the Mid-Ohio Regional Planning Commission (MORPC) Model [179], both developed by Vovsha et al., and the San Francisco model [85] are currently being utilized by their respective MPO. The San Francisco model is currently being updated to implement enhanced destination choice models and being recalibrated using more recent household and census data. Four other models for Atlanta, Sacramento, the San Francisco Bay Area, and Denver are currently in various stages of implementation [24].

Although many past activity-based models have undefined or large time resolutions, STARCHILD already in mid-1980s used 15-min temporal resolution. The most recent models, however, go even further to simulate activities at small time intervals such as 5 min (TASHA) and 10 min intervals (SIMAP), minute by minute (MASTIC, CentreSIM, MASTIC, GISICAS, and RAMBLAS), and second-by-second (TRANSIMS-LANL, ALBATROSS, AURORA, CATGW, CEMDAP, FAMOS, and FEATHERS). Many applications, however, operate with large resolutions of one hour and they are implemented with a target of 30 min to one hour [24]. Spatial resolution of the models is still dominated by the zonal level. ALBATROSS and MORPC both can operate at the subzone level. Alam-PSEM, AURORA, CEMDAP, FEATHERS, GISICAS, ILUTE, PUMA, SIMAP, SMASH, and TRANSIMS-LANL utilize data at essentially the building or point level. Only two applications have spatial resolutions below the zonal level (Denver model that contains a two-stage destination locator to predict the address within a zone and the Sacramento model that operates at the parcel level). Cognitive theories (models of knowledge and memory as well as behavioral process for planning activities) were used only in SCHEDULER and based on that in ALBATROSS and FEATHERS. Behavior is most often incorporated as intra-household interaction in ALBATROSS, CEMDAP, FAMOS, FEATHERS, ILUTE/TASHA, and CentreSIM as well as some of the applications in regions such as MORPC.

Examples of Mathematical Models

In this section additional details of two examples of mathematical models for activity and travel behavior analysis are offered. Both examples aim at incorporating human interaction in time allocation models and they are multilevel regression models (based on Goulias [59]) and group decision making utility maximization models (based on Zhang et al. [187]).

Multilevel Regression Models

These regression models are known by different names in different fields of research such as random coefficient models ([69] and p. 669 in [105]), multilevel models [48], mixed models [147], and hierarchical linear models [26]. They describe the contextual nature of the data and/or the way of accounting for dependent variable variation from multiple sources. Key advantages of these models are: explicit recognition in model formulation of the hierarchical, multiple level and nested structure of the data we analyze, and model specification using three groups of regression components in the same regression model. The first group assumes constant sensitivity to explanatory variables among the units of analysis representing the mean effect of an explanatory variable on the dependent variable. The second group assumes a random deviation around this mean and the third group is the usual random error term(s) of the regression equation. When compared to traditional regression models, which contain only one level, multilevel models do not underestimate the standard errors of coefficient estimates avoiding overstatements about the statistical significance of policy variables (e. g., we do not exaggerate the effect of taxation on car ownership or the effect of time and cost on route choice). A system of multilevel regression models can be written as follows.

$$Y^q_{tij} = \alpha^q_{tij} + \beta^q_k X_{tij} + \gamma^q_m Z_{tij} \tag{1}$$

$$a^q_{tij} = \gamma^q_0 + \nu^q_j + u^q_{ij} + \varepsilon^q_{tij}\,, \quad \text{where } q = 1, \ldots, Q\,, \tag{2}$$

$$\beta^q_{k1} = \gamma^q_{k1} + u^q_{k1ij}\,, \quad \beta^q_{k2} = \gamma^q_{k2} + \nu^q_{k2j}\,. \tag{3}$$

Equation (1), represents Q equations that are one for each Y^q_{tij} variable that we want to explain and use in travel

demand forecasting. They can be the amount of time dedicated to activities and travel or distances to specific destinations or even attributes of routes considered by trip makers. The index t represents the time at which an observation was made for a person i from within a household j (with $t = 1, 2, 3, \ldots, T$, $i = 1, 2, \ldots$, number of people in household j, $j = 1, 2, \ldots$, number of households in sample). In this way we can identify change from one time point to another by an individual and study the relationships among individuals within social units (e. g., households, associations, neighborhoods and so forth).

The time points can be the same for all individuals or they may vary depending on the data collection procedures and willingness of respondents to provide information. Equation 1 is called the level 1 model because it is written at the level of the time point (observation occasion). The first term in the right hand side of Eq. (1) is a random intercept, α, given by Eq. (2). This component has specific meaning. For example, α^q_{tij} is the mean value of person i in household j at time t for variable q. The term ε^q_{tij} is a random temporal variation (also called within person variation) and it is the deviation of time expenditure around γ^q_0. The term u^q_{ij} is a random person to person variation (also called within household variation) and it is also a deviation of around γ^q_0. The term v^q_j is a random household to household variation and it is also a deviation around γ^q_0. These are also called random error components and they are usually assumed normally distributed with $E(\varepsilon) = E(u) = E(v) = 0$, with $\mathrm{Var}(\varepsilon) = \sigma^2_\varepsilon$, $\mathrm{Var}(u) = \sigma^2_u$, and $\mathrm{Var}(v) = \sigma^2_v$ to be estimated. It is worth noting that the system of equations represented by Eq. (1) contain a set of gamma coefficients (associated with a matrix Z representing explanatory variables) that are defined in a similar way as in typical regression models. The βs, however, that multiply the matrix X are defined as random with a mean and a variation around the means γs. This variation can be due to the temporal, personal, and/or household levels. In this way, we can define a variety of equations at each of these levels to represent heterogeneous behavior that is either due to temporal fluctuations, personal variation, or household variation. Equation (3) differentiates between βs that vary within individuals and those that vary within households. In this way, at each level we have a level-specific variance-covariance matrix of all the random terms (εs, us, vs). The significance of the elements in each of these three matrices can be tested using goodness-of-fit measures based on the deviance, which is the difference in the −2Log(likelihood) at convergence between two nested (in terms of specification) models. In addition, the γs can also be tested if they are significantly different than zero using a t-test. The γs in Eq. (1) are called the *fixed effects* and the remaining terms are called the *random effects* at each of the three levels in the hierarchy. Estimation of all the fixed and random parameters can be accomplished either by Full Information Maximum Likelihood, FIML, applied to Y directly or applied to the least-squares residuals, called Restricted Maximum Likelihood-REML that can be used in tandem with a generalized least squares approach. Longford [105], Bryk and Raudensbush [26] and [48] provide a comprehensive review of estimation techniques, their performance assessment, and detailed algorithms.

Household Utility Models

The second example is also representative of a movement toward more detailed consideration of within household decision making dynamics. Although the model was specified by Zhang et al. [187] for time allocation to shared (j) and non-shared activities (s), it is a potentially useful model for other trip making decisions. Each person in a household is assumed to form two utility functions. One utility is for the shared activities (i. e., engagement in activities with other household members) and non-shared activities. These utility functions are given by Eq. (4) (shared activity) and 5 (non-shared activity).

$$u_{is} = \exp\left(\left(\alpha_s + \sum_k \beta_{sk} x_{isk}\right) \ln\left(\sum_m \kappa_{sm} \tau_{ism}\right) + \varepsilon_{is}\right) \ln(t_{is}) \quad (4)$$

$$u_{ij} = \exp\left(\left(\alpha_j + \sum_k \beta_{jk} x_{ijk}\right) \ln\left(\sum_m \kappa_{jm} \tau_{ijm}\right) + \varepsilon_{ij}\right) \ln(t_{ij}) \,, \quad (5)$$

where:

- α_s is the constant term for each shared activity s.
- x_{isk} is the kth explanatory variable (and/or attribute) of household member i for shared activity s.
- β_{sk} is the parameter associated with the kth attribute of the shared activity.
- τ_{ism} is the travel time by mode m for each activity s by person i.
- κ_{sm} is the parameter associated with travel time by mode m.
- ε_{is} is a random error term of the shared activity s by person i.
- t_{is} is the amount of time dedicated to activity s by person i.

α_j is the constant term for each non-shared activity.
x_{ijk} is the kth explanatory variable (attribute) of household member i for non-shared activity j.
β_{jk} is the parameter associated with the kth attribute of non-shared activity.
τ_{ijm} is the travel time by mode m for each activity s by person i.
κ_{jm} is the parameter associated with travel time by mode m.
ε_{ij} is a random error term of the non-shared activity j by person i.
t_{ij} is the amount of time dedicated to activity j by person i.

The overall utility of activity participation and travel for each person i under the assumption of a multi-linear utility is given by Eq. (6).

$$u_i = \sum_{j=1}^{J+S} r_{ij} u_{ij} + \sum_{j=1}^{J+S} \sum_{j'>j} \delta_i r_{ij} r_{ij'} u_{ij} u_{ij'} \tag{6}$$

where,
u_{ij} is the utility of activity j for person i.
r_{ij} is the relative interest of person i for activity j.
δ_i is parameter of activity dependency for member i.
$J+S$ is the number of non-shared and shared activities for a person within the unit of time under consideration.

In a similar way the household utility function is a multi-linear combination of the individual utilities in Eq. (7).

$$\text{HUF} = \sum_{i=1}^{n} w_i u_i + \lambda \sum_{i=1}^{n} \sum_{i'>i} (w_i w_{i'} u_i u_{i'}) \tag{7}$$

where,
HUF is the household utility combining the utilities of all household members n.
u_i is the utility of household member i.
w_i is the relative influence of each household member i.
$\boldsymbol{\lambda}$ is a parameter of within household interaction.

Under the assumption of maximizing HUF it is possible to create a Lagrangian function that accounts for constraints (i. e., total amount of time available, signs of parameters and so forth) and through a maximization solution derive equations that can be used to estimate the unknown parameters in Eqs. (4–7) (details are provided in Zhang et al. [187] for time allocation). It is worth noting that Zhang et al. [187], derived two alternate model systems by changing the utility functions to represent different intra-household bargaining models (for a detailed review see [13]). Then through a linearization process they developed a system of linear equations and estimated the parameters using a multiple equations econometric approach (the Seemingly Unrelated Regression Estimation [69]) that is a simplifying alternative to the multilevel models described earlier in this section. A more general review of this type of model formulation is also provided by Timmermans [164].

Summary

Similarities and differences among the implemented modeling ideas are:

- A hierarchy of decisions by households is assumed that identifies longer term choices determining the shorter term choices. In this way different blocks of variables can be identified and their mutual correlation used to derive equations that are used in forecasting.
- Anchor points (Home location – work location – school location) are inserted in the first choice level and they define the overall spatial structure of activity scheduling.
- Out-of-home activity purposes include work, school, shopping, meals, personal business, recreation, and escort. These expand the original home-based and non-home based purposes in travel behavior and the three activity types in home economics (labor for pay, labor at home, and leisure).
- In-home activities are explicitly modeled or allowed to enter the model structure as a "stay-at-home" choice with some models allowing for activity choice at home (work, maintenance and discretionary). In this way limited substitution between at home and outside home can be reflected in the models.
- Stop frequencies and activities at stops are modeled at the day pattern and tour levels to distinguish between activities and trips that can be rescheduled with little additional efforts versus the activities and trips that cannot be rescheduled (e. g., school trips).
- Modes and destinations are modeled together. In this way the mutual influence – sequential and/or simultaneous relationships can be reflected in the model structure.
- Time is included in a few instances in activity-based models. For example departure time for trips and tour time of day choice are modeled explicitly. Model time periods are anywhere between 30 min and second-by-second and time windows are used to account for scheduling. This modeling component allows to incorporate time-of-day in the modeling suites. It also allows to identify windows of activity and travel opportunities. The presence of departure time also enables models to

trip matrices for any desired periods in a day. In fact, output of time periods depends on route choice and traffic assignment needs and can be adjusted almost at will.

- Human interaction, although limited for now to the within-household interaction, is incorporated by relating the day pattern of one person to the day patterns of other persons within a household, their joint activities and trip making are explicitly modeled (joint recreation, escort trips), and allocation of activity-roles are also modeled.
- Spatial aspects of a region are accounted for using methods that produce spatially distributed synthetic populations using as external control totals averages and relative frequencies of population characteristics.
- Accessibility measures are used to capture spatial interaction among activity locations and the level of service offered by the transportation systems. These are also the indicators used to account for feedback among the lower level in the hierarchy decisions (e. g., activity location choices, routes followed, congestion) and the higher level such as residence location choice.
- Spatial resolution is heavily dependent on data availability and it reached already the level of a parcel and/or building at its most disaggregate level. Outputs of models are then aggregated to whatever level is required by traffic assignment, mode specific studies (nonmotorized and/or transit) and reporting needs and requirements.

Overall, the plethora of advances includes: a) models and experiments to create computerized virtual worlds and synthetic schedules at the most elementary level of decision making using microsimulation and computational process models; b) data collection methods and new methods to collect extreme details about behavior and to estimate, validate, and verify models using advanced hardware, software, and data analysis techniques; and c) integration of models from different domains to reflect additional interdependencies such as land use and telecommunications.

Future Directions

Much more work remains to be done in order to develop models that can answer more complex questions in policy analysis and for this reason a few steps are outlined here. In policy and program evaluation, transportation analysis appears to be narrowly applied to only one method of assessment that does not follow the ideal of a randomized controlled trial and does not explicitly define what experimental setting we are using for our assessments. Unfortunately this weakens our findings about policy analysis and planning activities. Although we have many laboratory experiments that were done for intelligent transportation systems we lack studies and guidelines to develop experimental and quasi-experimental procedures to guide us in policy development and large scale data collection.

In addition, many issues remain unresolved in the areas of coordination among scale in time and space and related issues. In addition very little is known about model sensitivity and data error tolerance and their mapping to strategy evaluations. This is partially due to the lack of tools that are able to make these assessments but also due to lack of scrutiny of these issues and their implications on impact assessment.

Regarding strategic planning and evaluation, we also lack models designed to be used in scenario building exercises such as backcasting and related assessments. The models about change are usually defined for forecasting and simple time inversion may not work to make them usable in backcasting. This area does not have the long tradition of modeling and simulation to help us develop suitable models. Should more attention be paid to this aspect? Is there room for a combination of techniques including qualitative research methods? What is the interface between this aspect and the experimental methods questions in program evaluation?

In the new research and technology area, since we are dealing with the behavior of persons, it is unavoidable to consider perceptions of time and space. What role should perceptions of time and space [51] play in behavioral models and what is the most appropriate use of these perceptions? The multiple dimensions of time such as tempo, duration, and clock time are neglected in behavioral models – is there a role for them in behavioral models?

Human interaction is considered important and is receiving attention in more recent research Golob and McNally [54], Chandrasekharan and Goulias [27], Simma and Axhausen [148], Gliebe and Koppelman [47], Goulias and Kim [62], Zhang et al. [187], but only partially accounted for in applications as illustrated by Vovsha and Petersen [177]. Future applications will increasingly pay attention to motivations for human interactions and the nature of these interactions within households and in a wider social network context.

Bibliography

1. Adler T, Ben-Akiva M (1979) A theoretical and empirical model of trip chaining behavior. Transp Res B 13:243–257
2. Alam BS (1998) Dynamic emergency evacuation management system using GIS and spatio-temporal models of behavior. MS Thesis. Department of Civil and Environmnetal

Engineering, The Pennsylvania State University, University Park
3. Alam BS, KG Goulias (1999) Dynamic emergency evacuation management system using GIS and spatio-temporal models of behavior. Transp Res Record 1660:92–99
4. Anas A (1982) Residential location markets and urban transportation: Economic theory, econometrics and policy analysis with discrete choice models. Academic Press, New York
5. Arentze T, Timmermans H (2000) ALBATROSS – A learning based transportation oriented simulation system. European Institute of Retailing and Services Studies (EIRASS), Technical University of Eindhoven, Eindhoven
6. Arentze T, Timmermans H, Janssens D, Wets G (2006) Modeling short-term dynamics in activity-travel patterns: From aurora to feathers. Presented at the Innovations in Travel Modeling Conference, Austin 21–23 May 2006
7. Avineri E, Prashker Y (2003) Sensitivity to uncertainty: The need for a paradigm shift. CD-TRB ROM Proceedings, Paper presented at the 82nd Annual Transportation Research Board Meeting, 12–16 January 2003, Washington DC
8. Becker GS (1976) The economic approach to human behavior. The University of Chicago Press, Chicago
9. Ben-Akiva ME, Lerman SR (1985) Discrete choice analysis: Theory and application to travel demand. MIT Press, Cambridge
10. Ben-Akiva ME, Morikawa T (1989) Estimation of mode switching models from revealed preferences and stated intentions. Paper presented at the International Conference on Dynamic Travel Behavior at Kyoto University Hall, Kyoto
11. Ben-Akiva M, Bowman JL, Gopinath D (1996) Travel demand model system for the information era. Transportation 23: 241–266
12. Ben-Akiva ME, Walker J, Bernardino AT, Gopinath DA, Morikawa T, Polydoropoulou A (2002) Integration of choice and latent variable models. In: Mahmassani HS (ed) In perceptual motion: Travel behavior research opportunities and application challenges. Pergamon, Amsterdam
13. Bergstrom TC (1995) A survey of theories of the family. Department of Economics, University of California Santa Barbara, Paper 1995D. http://repositories.cdlib.org/ucsbecon/bergstrom/1995D/
14. Bhat CR (2000) Flexible model structures for discrete choice analysis. In: Hensher DA, Button KJ (eds) Handbook of transport modelling. Pergamon, Amsterdam, pp 71–89
15. Bhat C (2001) A comprehensive and operational analysis framework for generating the daily activity-travel pattern of workers. Paper presented at the 78th Annual Meeting of the Transportation Research Board, Washington DC, 10–14 January 2001
16. Bhat CR (2003) Random utility-based discrete choice models for travel demand analysis. In: Goulias KG (ed) Transportation systems planning: Methods and applications. CRC Press, Boca Raton, pp 10-1–10-30
17. Bhat CR, Koppelman F (1999) A retrospective and prospective survey of time-use research. Transportation 26(2):119–139
18. Bhat CR, Singh SK (2000) A comprehensive daily activity-travel generation model system for workers. Transp Res A 34(1):1–22
19. Bhat CR, Guo J, Srinivasan S, Sivakumar A (2003) Activity-based travel demand modeling for metropolitan areas in Texas: Software-related processes and mechanisms for the activity-travel pattern generation microsimulator. Research Report 4080-5, Center for Transportation Research, Austin
20. Bockenholt U (2002) Comparison and choice: Analyzing discrete preference data by latent class scaling models. In: Hagenaars JA, McCutcheon AL (eds) Applied latent class analysis. Cambridhe University Press, Cambridge, pp 163–182
21. Borgers AWJ, Hofman F, Timmermans HJP (1997) Activity-based modelling: Prospects. In: Ettema DF, Timmermans HJP (eds) Activity-based approaches to travel analysis. Pergamon, Oxford, pp 339–351
22. Borgers AWJ, Timmermans AH, van der Waerden P (2002) Patricia: Predicting activity-travel interdependencies with a suite of choice-based, interlinked analysis. Transp Res Rec 1807:145–153
23. Bowman JL, Bradley M, Shiftan Y, Lawton TK, Ben-Akiva M (1998) Demonstration of an activity-based model system for Portland. Paper presented at the 8th World Conference on Transport Research, Antwerp, June 1998
24. Bradley M, Bowman J (2006) A summary of design features of activity-based microsimulation models for US MPOs. Conference on Innovations in Travel Demand Modeling, Austin 21–23 May 2006
25. Brög W, Erl E (1980) Interactive measurement methods – Theoretical bases and practical applications. Transp Res Rec 765:1–6
26. Bryk AS, Raudenbush SW (1992) Hierarchical linear models. Sage, Newberry Park
27. Chandrasekharan B, Goulias KG (1999) Exploratory longitudinal analysis of solo and joint trip making in the Puget Sound transportation panel. Transp Res Rec 1676:77–85
28. Chapin Jr FS (1974) Human activity patterns in the city: Things people do in time and space. Wiley, New York
29. Chung J, Goulias KG (1997) Travel demand forecasting using microsimulation: Initial results from a case study in Pennsylvania. Transp Res Rec 1607:24–30
30. Creighton RL (1970) Urban transportation planning. University of Illiniois Press, Urbana
31. Cullen I, Godson V (1975) Urban networks: The structure of activity patterns. Progr Plan 4(1):1–96
32. Dijst M, Vidakovic V (1997) Individual action space in the city. In: Ettema DF, Timmermans HJP (eds) Activity-based approaches to travel analysis. Elsevier Science Inc, New York, pp 117–134
33. Dillman DA (2000) Mail and internet surveys: The tailored design method, 2nd edn. Wiley, New York
34. Doherty S (2003) Interactive methods for activity scheduling processes. In: Goulias KG (ed) Transportation systems planning: Methods and applications. CRC Press, Boca Raton, pp 7-1 to 7-25
35. Doherty ST, Noel N, Lee M-G, Sirois C, Ueno M (2001) Moving beyond observed outcomes: Global positioning systems and interactive computer-based travel behavior surveys. Transportation Research Circular, E-C026, March 2001, Transportation Research Board, Washington DC
36. Ettema DF, Timmermans HJP (1997) Activity-based approaches to travel analysis. Elsevier Science Inc, New York, p xiii
37. Ettema DF, Borgers AWJ, Timmermans HJP (1995) Competing risk hazard model of activity choice, timing, sequencing and duration. Transp Res Rec 1439:101–109

38. Ettema DF, Borgers AWJ, Timmermans H (1996) SMASH (Simulation Model of Activity Scheduling Heuristics): Some simulations. Transp Res Rec 1551:88–94
39. Ettema DF, Daly A, de Jong G, Kroes E (1997) Towards an applied activity-based travel demand model. Paper presented at the IATBR Conference, Austin 21–25 September 1997
40. Ettema DF, de Jong K, Timmermans H, Bakema A (2006) PUMA: Multi-agent modeling of urban systems. 2006 Transportation Research Board CD-ROM
41. Fellendorf M, Haupt T, Heidl U, Scherr W (1997) PTV vision: Activity based demand forecasting in daily practice. In: Ettema DF, Timmermans HJP (eds) Activity-based approaches to travel analysis. Elsevier Science Inc, New York, pp 55–72
42. Fosgerau M (2002) PETRA – an activity-based approach to travel demand analysis. In: Lundquist L, Mattsson L-G (eds) National transport models: Recent developments and prospects. Royal Institute of Technology, Stockholm, Sweden. Springer, Berlin
43. Gärling T, Brannas K, Garvill J, Golledge RG, Gopal S, Holm E, Lindberg E (1989) Household activity scheduling. In: Transport policy, management and technology towards 2001. Selected Proceedings of the Fifth World Conference on Transport Research, vol 4. Western Periodicals, Ventura, pp 235–248
44. Gärling T, Kwan M, Golledge R (1994) Computational-process modeling of household travel activity scheduling. Transp Res Part B 25:355–364
45. Gärling T, Laitila T, Westin K (1998) Theoretical foundations of travel choice modeling: An introduction. In: Garling T, Laitila T, Westin K (eds) Theoretical foundations of travel choice modeling. Pergamon, Elsevier, Amsterdam, pp 1–30
46. Garrett M, Wachs M (1996) Transportation planning on Trial. The clean air act and travel forecasting. Sage Publications, Thousand Oaks
47. Gliebe JP, Koppelman FS (2002) A model of joint activity participation. Transportation 29:49–72
48. Goldstein H (1995) Multilevel statistical models. Edward Arnold, London, New York
49. Golledge RG, Stimpson RJ (1997) Spatial behavior: A geographic perspective. The Guilford Press, New York
50. Golledge RG, Gärling T (2003) Spatial behavior in transportation modeling and planning. In: Goulias KG (ed) Transportation systems planning: Methods and applications. CRC Press, Boca Raton, pp 1–27
51. Golledge RG, Gärling T (2004) Cognitive maps and urban travel. In: Hensher D, Button K, Haynes K, Stopher P (eds) Handbook of transport geography and spatial systems, vol 5. Elsevier, Amsterdam, pp 501–512
52. Golledge RG, Smith TR, Pellegrino JW, Doherty S, Marshall SP (1985) A conceptual model and empirical analysis of children's acquisition of spatial knowledge. J Environ Psychol 5(2):125–152
53. Golob TF (2001) Travelbehaviour.com: Activity approaches to modeling the effects of information technology on personal travel behaviour, in travel behavior research. In: Hensher D (ed) The leading edge. Elsevier Science/Pergamon, Kidlington, Oxford, pp 145–184
54. Golob TF, McNally M (1997) A model of household interactions in activity participation and the derived demand for travel. Transp Res B 31:177–194
55. Golob TF, Kitamura R, Long L (eds) (1997) Panels for transportation planning: Methods and applications. Kluwer, Academic Publishers, Massachusetts
56. Goodman LA (2002) Latent class analysis: The empirical study of latent types, latent variables, and latent structures. In: Hagenaars JA, McCutcheon AL (eds) Applied latent class analysis. Cambridhe University Press, Cambridge, pp 3–55
57. Goulias KG (1999) Longitudinal analysis of activity and travel pattern dynamics using generalized mixed Markov latent class models. Transp Res B 33:535–557
58. Goulias KG (2001) A Longitudinal integrated forecasting environment (LIFE) for activity and travel forecasting. In: Villacampa Y, Brebbia CA, Uso J-L (eds) Ecosystems and sustainable development III. WIT Press, Southampton, pp 811–820
59. Goulias KG (2002) Multilevel analysis of daily time use and time allocation to activity types accounting for complex covariance structures using correlated random effects. Transportation 29(1):31–48
60. Goulias KG (2003) Transportation systems planning. In: Goulias KG (ed) Transportation systems planning: Methods and applications. CRC Press, Boca Raton, pp 1–1 to 1–45
61. Goulias KG, Kim T (2003) A longitudinal analysis of the relationship between environmentally friendly modes, weather conditions, and information-telecommunications technology market penetration. In: Tiezzi E, Brebbia CA, Uso JL (eds) Ecosystems and sustainable development, vol 2. WIT Press, pp 949–958
62. Goulias KG, Kim T (2005) An analysis of activity type classification and issues related to the with whom and for whom questions of an activity diary. In: Timmermans H (ed) Progress in activity-based analysis. Elsevier, pp 309–334
63. Goulias KG, Kitamura R (1992) Travel demand analysis with dynamic microsimulation. Transp Res Rec 1607:8–18
64. Goulias KG, Kitamura R (1997) Regional travel demand forecasting with dynamic microsimulation models. In: Golob T, Kitamura R, Long L (eds) Panels for transportation planning: Methods and applications. Kluwer, Academic Publishers, Massachusetts, pp 321–348
65. Goulias KG, Litzinger T, Nelson J, Chalamgari V (1993) A study of emission control strategies for Pennsylvania: Emission reductions from mobile Sources, cost effectiveness, and economic impacts. Final report to the Low Emissions Vehicle Commission. PTI 9403. The Pennsylvania Transportation Institute, University Park
66. Goulias KG, Kim T, Pribyl O (2003) A longitudinal analysis of awareness and use for advanced traveler information systems. Paper to be presented at the European Commission Workshop on Behavioural Responses to ITS – 1–3 April 2003, Eindhoven
67. Goulias KG, Zekkos M, Eom J (2004) CentreSIM3 Scenarios for the South Central Centre County Transportation Study. CentreSIM3 Report submitted to McCormick Taylor Associates and the Mid-Atlantic Universities Transportation Center, April 2004, University Park
68. Goulias KG, Blain L, Kilgren N, Michalowski T, Murakami E (2007) Catching the next big wave: Are the observed behavioral dynamics of the baby boomers forcing us to rethink regional travel demand models? Paper presented at the 86th Transportation Research Board Annual Meeting, 21–25 January 2007, Washington DC and included in the CD ROM proceedings

69. Greene WH (1997) Econometric analysis, 3rd edn. Prentice Hall, New Jersey
70. Grieving S, Kemper R (1999) Integration of transport and land use policies: State of the art. Deliverable 2b of the Project TRANSLAND, 4th RTD Framework Programme of the European Commission
71. Haab TC, Hicks RL (1997) Accounting for choice set endogeneity in random utility models of recreation demand. J Environ Econ Manag 34:127–147
72. Hagerstrand T (1970) What about people in regional science? Pap Reg Sci Assoc 10:7–21
73. Hayes-Roth B, Hayes-Roth F (1979) A cognitive model of planning. Cogn Sci 3:275–310
74. Hato E (2006) Development of behavioral context addressable loggers in the shell for travel activity analysis. Paper presented at the IATBR conference, Kyoto
75. Henson K, Goulias KG (2006) Preliminary assessment of activity and modeling for homeland security applications. Transportation Research Record: J Transportation Research Board, No. 1942, Transportation Research Board of the national Academies, Washington DC, pp 23–30
76. Henson K, Goulias KG, Golledge R (2006) An assessment of activity-based modeling and simulation for applications in operational studies, disaster preparedness, and homeland security. Paper presented at the IATBR conference, Kyoto
77. Horowitz JL (1991) Modeling the choice of choice set in discrete-choice random-utility models. Environ Plan A 23: 1237–1246
78. Horowitz JL, Louviere JJ (1995) What is the role of consideration sets in choice modeling? Int J Res Marketing 12:39–54
79. Huigen PPP (1986) Binnen of buiten bereik?: Een sociaal-geografisch onderzoek in Zuidwest-Friesland, Nederlandse Geografische Studies 7, University of Utrecht, Utrecht
80. Hutchinson BG (1974) Principles of urban transport systems planning. Scripta, Washington DC
81. JHK & Associates, Clough, Harbour & Associates, Pennsylvania Transportation Institute, Bogart Engineering (1996) Scranton/Wilkes-barre area strategic deployment plan. Final Report. Prepared for Pennsylvania Department of Transportation District 4-0. August 1996, Berlin
82. Joh C-H, Arentze T, Timmermans H (2004) Activity-travel scheduling and rescheduling decision processes: Empirical estimation of aurora model. Transp Res Rec 1898:10–18
83. Jones PM, Dix MC, Clarke MI, Heggie IG (1983) Understanding travel behaviour. Gower, Aldershot
84. Jones P, Koppelman F, Orfeuil J (1990) Activity analysis: State-of-the-art and future directions. In: Jones P (ed) Developments in dynamic and activity-based approaches to travel analysis. A compendium of papers from the 1989 Oxford Conference. Avebury, Gower-Aldershot, pp 34–55
85. Jonnalagadda N, Freedman J, Davidson WA, Hunt JD (2001) Development of microsimulation activity-based model for San Francisco. Transp Res Rec 1777:25–35
86. Kahneman D, Tversky A (1979) Prospect theory: An analysis of decisions under risk. Econometrica 47(2):263–291
87. Kawakami S, Isobe T (1989) Development of a travel-activity scheduling model considering time constraint and temporal transferability test of the model. In: Transport policy, management and technology towards 2001: Selected Proceedings of the Fifth World Conference on Transport Research, vol 4. Western Periodicals, Ventura, pp 221–233
88. Kharoufeh JP, Goulias KG (2002) Nonparametric identification of daily activity durations using Kernel density estimators. Transp Res B Methodological 36:59–82
89. Kitamura R (1988) An evaluation of activity-based travel analysis. Transportation 15:9–34
90. Kitamura R (1997) Applications of models of activity behavior for activity based demand forecasting. In: Engelke LJ (ed) Activity-based travel forecasting conference: Summary, recommendations and compendium of papers. Report of the Travel Model Improvement Program. Texas Transportation Institute, Arlington, pp 119–150
91. Kitamura R (2000) Longitudinal methods. In: Hensher DA, Button KJ (eds) Handbook of transport modelling. Pergamon, Ansterdam, pp 113–128
92. Kitamura R, Fujii S (1998) Two computational process models of activity-travel choice. In: Garling T, Laitila T, Westin K (eds) Theoretical foundations of travel choice modeling. Pergamon, Elsevier, Amsterdam, pp 251–279
93. Kitamura R, Pas EI, Lula CV, Lawton TK, Benson PE (1996) The sequenced activity simulator (SAMS): an integrated approach to modeling transportation, land use and air quality. Transportation 23:267–291
94. Kitamura R, Chen C, Pendyala RM (1997) Generation of synthetic daily activity-travel patterns. Transp Res Rec 1607: 154–162
95. Koppelman FS, Sethi V (2000) Closed-form discrete-choice models. In: Hensher DA, Button KJ (eds) Handbook of transport modelling. Pergamon, Amsterdam, pp 211–225
96. Krizek KJ, Johnson A (2003) Mapping of the terrain of information and communications technology (ICT) and household travel, Transportation Research Board annual meeting CD-ROM, Washington DC, January 2003
97. Kuhnau JL (2001) Activity-based travel demand modeling using spatial and temporal models in the urban transportation planning system. MS Thesis. Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park
98. Kuhnau JL, Goulias KG (2002) Centre SIM: Hour-by-hour travel demand forecasting for mobile source emission estimation. In: Brebbia CA, Zannetti P (eds) Development and application of computer techniques to environmental studies IX. WIT Press, Southampton, pp 257–266
99. Kuhnau JL, Goulias KG (2003) Centre SIM: First-generation model design, pragmatic implementation, and scenarios. In: Goulias KG (ed) Transportation systems planning: Methods and applications. CRC Press, Boca Raton, pp 16-1–16-14
100. Kulkarni A, McNally MG (2001) A microsimulation of daily activity patterns. Paper presented at the 80th Annual Meeting of the Transportation Research Board, Washington, 7–11 January 2001
101. Kwan M-P (1994) A GIS-based model for activity scheduling in intelligent vehicle highway systems (IVHS). Unpublished Ph D, Department of Geography, University of California Santa Barbara, Santa Barbara
102. Kwan M-P (1997) GISICAS: An activity-based travel decision support system using a GIS-interfaced computational-process model. In: Ettema DF, Timmermans HJP (eds) Activity-based approaches to travel analysis. Elsevier Science Inc, New York, pp 263–282
103. Lenntorp B (1976) Paths in space-time environment: A time geographic study of possibilities of individuals. The Royal

University of Lund, Department of Geography. Lund Studies in Geography, Series B Human Geography 44
104. Lomborg B (2001) The skeptical environmnetalist: Measuring the real state of the world. Cambridge University Press, Cambridge
105. Longford NT (1993) Random coefficient models. Clarendon Press, Oxford
106. Los Alamos National Laboratory (2003) TRANSIMS: Transportation analysis system (Version 3.1). LA-UR-00-1725
107. Loudon WR, Dagang DA (1994) Evaluating the effects of transportation control measures. In: Wholley TF (ed) Transportation planning and air quality II. American Society of Civil Engineers, New York
108. Louviere JJ, Hensher DA, Swait JD (2000) Stated choice methods: Analysis and application. Cambridge University Press, Cambridge
109. Ma J (1997) An activity-based and micro-simulated travel forecasting system: A pragmatic synthetic scheduling approach. Unpublished Ph D Dissertation, Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park
110. Mahmassani HS, Herman R (1990) Interactive experiments for the study of tripmaker behaviour dynamics in congested commuting systems, In: Developments in dynamic and activity-based approaches to travel analysis. A compendium of papers from the 1989 Oxford Conference. Avebury
111. Mahmassani HS, Jou R-C (1998) Bounded rationality in commuter decision dynamics: Incorporating trip chaing in departure time and route switching decisions. In: Garling T, Laitila T, Westin K (eds) Theoretical foundations of travel choice modeling. Pergamon, Elsevier, Amsterdam
112. Manheim ML (1979) Fundamentals of transportation systems analysis, vol 1: Basic Concepts. MIT Press. Cambridge
113. Marker JT, Goulias KG (2000) Framework for the analysis of grocery teleshopping. Transp Res Rec 1725:1–8
114. McNally MG (2000) The activity-based approach. In: Hensher DA, Button KJ (eds) Handbook of transport modelling. Pergamon, Amsterdam, pp 113–128
115. McFadden D (1998) Measuring willingness-to-pay for transportation improvements. In: Garling T, Laitila T, Westin K (eds) Theoretical foundations of travel choice modeling. Pergamon, Elsevier, Amsterdam, pp 339–364
116. Meyer MD, Miller EJ (2001) Urban transportation planning, 2nd edn. McGrawHill, Boston
117. Miller EJ (2003) Land use: Transportation modeling. In: Goulias KG (ed) Transportation systems planning: Methods and applications. CRC Press, Boca Raton, pp 5–1 to 5–24
118. Miller EJ (2006) Resource paper on integrated land use-transportation models. IATBR, Kyoto, 2006
119. Miller EJ, Roorda MJ (2003) A prototype model of household activity/travel scheduling. Transp Res Rec 1831:114–121
120. Miller JS, Demetsky MJ (1999) Reversing the direction of transportation planning process. ASCE J Transp Eng 125(3):231–237
121. Mokhtarian PL (1990) A typology of relationships between telecommunications and transportation. Transp Res A 24(3):231–242
122. National Cooperative Highway Research Program (2000) Report 446. Transp Res Board, Washington DC
123. Newell A, Simon HA (1972) Human problem solving. Prentice Hall, Englewood Cliffs
124. Niemeier DA (2003) Mobile source emissions: An overview of the regulatory and modeling framework. In: Goulias KG (ed) in Transportation systems planning: Methods and applications. CRC Press, Boca Raton, pp 13–1 to 13–28
125. Ortuzar JdD, Willumsen LG (2001) Modelling transport, 3rd edn. Wiley, Chicester
126. Paaswell RE, Rouphail N, Sutaria TC (eds) (1992) Site impact traffic assessment. Problems and solutions. ASCE, New York
127. Payne JW, Bettman JR, Johnson EJ (1993) The adaptive decision maker. Cambridge University Press, Cambridge
128. Patten ML, Goulias KG (2001) Test plan: motorist survey – Evaluation of the Pennsylvania turnpike advanced travelers information system (ATIS) project, Phase III PTI-2001-23-I. April 2001. University Park
129. Patten ML, Hallinan MP, Pribyl O, Goulias KG (2003) Evaluation of the Smartraveler advanced traveler information system in the Philadelphia metropolitan area. Technical memorandum. PTI 2003–33. March 2003. University Park
130. Pendyala R (2003) Time use and travel behavior in space and time. In: Goulias KG (ed) Transportation systems planning: Methods and applications. CRC Press, Boca Raton, pp 2-1–2-37
131. Pendyala RM, Kitamura R, Kikuchi A, Yamamoto T, Fujii S (2005) The florida activity mobility simulator (FAMOS): An overview and preliminary validation results. Presented at the 84th Annual Transportation Research Board Conference and CD-ROM
132. Pribyl O (2004) A microsimulation model of activity patterns and within household interactions. Ph D Dissertation, Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park
133. Pribyl O (2007) Computational intelligence in transportation: Short user-oriented guide. In: Goulias KG (ed) Transport science and technology. Elsevier, Amsterdam, pp 37–54
134. Pribyl O, Goulias KG (2003) On the application of adaptive neuro-fuzzy inference system (ANFIS) to analyze travel behavior. Paper presented at the 82nd Transportation Research Board Meeting and included in the CDROM proceedings and accepted for publication in the Transportation Research Record, Washington DC, January 2003
135. Pribyl O, Goulias KG (2005) Simulation of daily activity patterns. In: Timmermans H (ed) Progress in activity-based analysis. Elsevier Science, Amsterdam, pp 43–65
136. Ramadurai G, Srinivasan KK (2006) Dynamics and variability in within-day mode choice decisions. Role of state dependence, habit persistence, and unobserved heterogeneity. Transportation Research Record, J Transportation Research Board, No. 1977, Transportation Research Board of the National Academies, Washington DC, pp 43–52
137. Recker WW (1995) The household activity pattern problem: General formulation and solution. Transp Res B 29:61–77
138. Recker WW, McNally MG, Root GS (1986) A model of complex travel behavior: Part I – Theoretical development. Transp Res A 20(4):307–318
139. Recker WW, McNally MG, Root GS (1986) A model of complex travel behavior: Part II – An operational model. Transp Res A 20(4):319–330
140. Richardson A (1982) Search models and choice set generation. Transp Res Part A 16(5–6):403–416
141. Robinson J (1982) Energy backcasting: a proposed method of policy analysis. Energ Policy 10(4):337–344

142. Rubinstein A (1998) Modeling bounded rationality. The MIT Press, Cambridge
143. Sadek AW, El Dessouki WM, Ivan JI (2002) Deriving land use limits as a function of infrastructure capacity. Final Report, Project UVMR13-7, New England Region One University Transportation Center. MIT, Cambridge
144. Salomon I (1986) Telecommunications and travel relationships: A review. Transp Res A 20(3):223–238
145. Salvini P, Miller EJ (2003) ILUTE: An operational prototype of a comprehensive microsimulation model of urban systems. Paper presented at the 10th International Conference on Travel Behaviour Research, Lucerne, August 2003
146. Savage LJ (1954) The foundations of statistics. Reprinted version in 1972 by Dover Publications, New York
147. Searle SR, Casella G, McCulloch CE (1992) Variance components. Wiley, New York
148. Simma A, Axhausen KW (2001) Within-household allocation of travel-The case of Upper Austria. Transportation Research Record: J Transportation Research Board, No. 1752, TRB, National Research Council, Washington DC, pp 69–75
149. Simon HA (1983) Alternate visions of rationality. In: Simon HA (ed) Reason in human affairs. Stanford University Press, Stanford, pp 3–35
150. Simon HA (1997) Administrative behavior, 4th edn. The Free Press, New York
151. Southworth F (2003) Freight transportation planning: Models and methods. In: Goulias KG (ed) Transportation systems planning: Methods and applications. CRC Press, Boca Raton, pp 4.1–4.29
152. Sparmann U (1980) Ein verhaltensorientiertes Simulationsmodell zur Verkehrsprognose. Schriftenreihe des Instituts für Verkehrswesen 20. Universität (TH) Karlsruhe, Karlsruhe
153. Stefan KJ, McMillan JDP, Hunt JD (2005) An urban commercial vehicle movement model for calgary. Paper presented at the 84th Transportation Research Board Meeting, Washington DC
154. Stopher PR (1994) Predicting TCM responses with urban travel demand models. In: Wholley TF (ed) Transportation planning and air quality II. American Society of Civil Engineers, New York
155. Stopher PR, Meyburg AH (eds) (1976) Behavioral travel-demand models. Lexington Books, Lexington
156. Stopher PR, Hartgen DT, Li Y (1996) SMART: simulation model for activities, resources and travel. Transportation 23:293–312
157. Sundararajan A, Goulias KG (2002) Demographic microsimulation with DEMOS 2000: Design, validation, and forecasting, In: Goulias KG (ed) Transportation systems planning: Methods and applications. CRC Press, Boca Raton, pp 14-1–14-23
158. Swait J, Ben-Akiva M (1987) Incorporating random constraints in discrete models of choice set generation. Transp Res Part B 21(2):91–102
159. Swait J, Ben-Akiva M (1987) Empirical test of a constrained choice discrete model: Mode choice in Sao Paolo, Brazil. Transp Res Part B 21(2):103–115
160. Teodorovic D, Vukadinovic K (1998) Traffic control and transport planning: A fuzzy sets and neural networks approach. Kluwer, Boston
161. Thill J (1992) Choice set formation for destination choice modeling. Progr Human Geogr 16(3):361–382
162. Tiezzi E (2003) The end of time. WIT Press, Southampton
163. Timmermans H (2003) The saga of integrated land use-transport modeling: How many more dreams before we wake up? Conference keynote paper at the Moving through net: The physical and social dimensions of travel. 10th International Conference on Travel Behaviour Research, Lucerne, 10–15, August 2003. In: Proceedings of the meeting of the International Association for Travel Behevaior Research (IATBR). Lucerne, Switzerland
164. Timmermans H (2006) Analyses and models of household decision making processes. Resource paper in the CDROM proceedings of the 11th IATBR International Conference on Travel Behaviour Research, Kyoto, Japan
165. Timmermans H, Arentze T, Joh C-H (2001) Modeling effects of anticipated time pressure on execution of activity programs. Transp Res Rec 1752:8–15
166. Train KE (2003) Discrete choice methods with simulation. Cambridge University Press, Cambridge
167. Transportation Research Board (1999) Transportation, energy, and environment. Policies to promote sustainability. Transportation Research Circular 492. TRB Washington DC
168. Transportation Research Board (2002) Surface transportation environmental research: A long-term strategy. Transportation Research Board, Washington DC
169. Tversky (1969) Intransitivity of preferences. Psychol Rev 76:31–48
170. Tversky (1972) Elimination by aspects: A theory of choice. Psychol Rev 79:281–299
171. Tversky A, Kahneman D (1992) Advances in prospect theory: Cumulative representation of uncertainty. J Risk Uncertain 9:195–230
172. US Government (2006) Analytical perspectives. Budget of the United States Government, Fiscal year 2007. US Government printing Office, Washington DC
173. Van Middelkoop M, Borgers AWJ, Timmermans H (2004) Merlin. Transp Res Rec 1894:20–27
174. Van der Hoorn T (1997) Practitioner's future needs. Paper presented at the Conference on Transport Surveys, Raising the Standard. Grainau, Germany, May 24–30
175. Vause M (1997) A rule-based model of activity scheduling behavior. In: Ettema DF, Timmermans HJP (eds) Activity-based approaches to travel analysis. Elsevier Science Inc, New York, pp 73–88
176. Veldhuisen J, Timmermans H, Kapoen L (2000) RAMBLAS: a regional planning model based on the microsimulation of daily activity travel patterns. Transp Res A 32:427–443
177. Vovsha P, Petersen E (2005) Escorting children to school: Statistical analysis and applied modeling approach. Transp Res Rec: J Transp Res Board 1921, Transportation Research Board of the National Academies, Washington DC, pp 131–140
178. Vovsha P, Peterson EJ, Donnelly R (2002) Microsimulation in travel demand modeling: Lessons learned from the New York best practice mode. Transp Res Rec 1805:68–77
179. Vovsha P, Peterson EJ, Donnelly R (2003) Explicit modeling of joint travel by household members: Statistical evidence and applied approach. Transp Res Rec: J Transp Res Board 1831: 1–10
180. Waddell P, Ulfarsson GF (2003) Dynamic simulation of real estate development and land prices within an integrated land use and transportation model system. Presented at the 82nd Annual Meeting of the Transportation Research Board, 12–16 January 2003, Washington DC. Also available in http://www.urbansim.org/papers/ – accessed April 2003

181. Wang D, Timmermans H (2000) Conjoint-based model of activity engagement, timing, scheduling, and stop pattern formation. Transp Res Rec 1718:10–17
182. Weiland RJ, Purser LB (2000) Intelligent transportation systems. In: Transportation in the New Millennium. State of the art and future directions. Perspectives from transportation research board standing committees. Transportation Research Board. National Research Council. The National Academies, Washington DC, p 6. Also in http://nationalacademies.org/trb/
183. Wen C-H, Koppelman FS (2000) A conceptual and methodological framework for the generation of activity-travel patterns. Transportation 27:5–23
184. Williams HCWL, Ortuzar JD (1982) Behavioral theories of dispersion and the mis-specification of travel demand models. Transp Res B 16(3):167–219
185. Wilson EO (1998) Consilience, the unity of knowledge. Vintage Books, New York
186. Wolf J, Guensler R, Washington S, Frank L (2001) Use of electronic travel diaries and vehicle instrumentation packages in the year 2000. Atlanta Regional Household Travel Survey. Transportation Research Circular, E-C026, March 2001, Transportation Research Board, Washington DC
187. Zhang J, Timmermans HJP, Borgers AWJ (2005) A model of household task allocation and time use. Transp Res B 39:81–95

Treasury Market, Microstructure of the U.S.

Bruce Mizrach[1], Christopher J. Neely[2]
[1] Department of Economics, Rutgers University, New Brunswick, USA
[2] Research Department, Federal Reserve Bank of St. Louis, St. Louis, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Types of Treasury Issues
Treasury Market Participants
Stages of the Treasury Bond Market
The Treasury Futures Market
Seasonality and Announcement Effects
Discontinuities in the US Treasury Market
Order Flow in the US Treasury Market
Modeling the Limit Order Book
Price Discovery
Future Directions
Bibliography

Glossary

Algorithmic trading Algorithmic trading is the practice of automatically transacting based on a quantitative model.

Broker A broker is a firm that matches buyers and sellers in financial transactions. An *interdealer broker (IDB)* is an intermediary providing trading services to hedge funds, institutions, and other dealers. IDB's handle the majority of Treasury securities transactions in the secondary market.

Coupons Owners of Treasury notes and bonds receive periodic payments called coupons. They are fixed by the Treasury at auction and are typically paid semi-annually.

Depth Depth is the quantity the dealer is willing to sell at the bid or offer.

Electronic communications networks (ECN) The Securities and Exchange Commission defines electronic communications networks (ECNs) as "electronic trading systems that automatically match buy and sell orders at specified prices".

Market microstructure Market microstructure is a field of economics that studies the price formation process and trading procedures in security markets.

On-the-run On-the-run refers to the most recently auctioned Treasury security of a particular maturity. After the next auction, the security goes *off-the-run*.

Price discovery The process by which prices adapt to new information.

Primary dealers Primary dealers are large brokerage firms and investment banks that are permitted to trade directly with the Federal Reserve in exchange for making markets in Treasuries. They provide the majority of liquidity in the Treasury market, participate in Treasury auctions, and provide information to assist the Fed in implementing open market operations.

Secondary market After the initial auction of Treasury instruments, trading in on-the-run and off-the-run securities makes up the *secondary* Treasury market.

When issued When-issued bonds are those Treasuries whose auctions have been announced but have not yet settled.

Definition of the Subject

This article discusses the microstructure of the *US Treasury securities market*.

US Treasury securities are default risk free debt instruments issued by the US government. These securities play an important, even unique, role in international financial markets because of their safety, liquidity, and low transac-

tions costs. Treasury instruments are often the preferred safe haven during financial crises, a process often referred to as a "flight to quality".

According to the US Treasury, there was more than $9 trillion in US government debt outstanding as of August 31, 2007. Of this quantity, the public holds more than $5 trillion and $4.5 trillion is tradable on financial markets. Foreigners hold approximately $2.4 trillion of the marketable supply, with Japan and China together holding more than $1 trillion. According to the Securities Industry and Financial Markets Association (SIFMA), average daily trading volume in the US Treasury market in 2007 was $524.7 billion.

Microstructure is the study of the institutional details of markets and trading behavior. Microstructural analysis takes three ideas seriously that are often overlooked: the institutional features of the trading process influence how private information is impounded into prices; agents are heterogeneous; and information is asymmetric. Empirical microstructure research studies topics such as the causes and effects of market structure, how market structure influences price discovery, how trading and order flow reveal private information, how quickly public information is impounded into prices, the volatility-volume relation, and the determinants of transactions costs (i. e., the components of bid-ask spreads). The relatively recent availability of tick-by-tick financial data and limit order book data, as well as the computer resources to manipulate them, have been a great boon to financial market microstructure research.

Introduction

We begin by describing the types of Treasury issues and the major Treasury market participants, including the Federal Reserve, primary dealers and the major electronic brokers. We then outline the stages of the Treasury market, from auction announcements to the secondary market. Next, we examine several closely related areas of the literature: Seasonality in the Treasury market and the reactions of the Treasury market to macro and monetary announcements; discontinuities in Treasury prices; and the effect of order flow in Treasury markets. We then discuss modeling and other academic questions about the Treasury market.

Types of Treasury Issues

As of October 2007, the US Treasury issued four types of debt instruments. The shortest-maturity instruments are known as Treasury *bills*. 22.6% of the marketable US debt is in bills, securities with maturities of 1 year or less. Bills are sold at a discount and redeemed at their face value at maturity. They do not pay any coupons prior to maturity and currently have maturities up to 26 weeks. Treasury bill prices are usually quoted in "discount rate" terms, which are calculated with an actual/360 day count convention,

$$\text{T-bill discount rate} = [\text{face value} - \text{bill price}] \times (360/\text{number of days until maturity}) .$$

Thus, a bill with a face value of $100,000, a cash price of $97,500 and 90 days to maturity will have a discount rate of $10\% = [100 - 97.5] \times (360/90)$ in a newspaper. Treasury bill yields are often quoted as "bond equivalent yields", which are defined as,

$$\text{T-bill yield} = \left[\frac{\text{face value} - \text{bill price}}{\text{bill price}}\right] \times (365/\text{number of days until maturity}) .$$

Treasury instruments with intermediate maturities (2-, 5- and 10-year) are known as *Treasury notes*. *Notes* pay semi-annual coupons, and make up 54.7% of the debt. In February 2006, the US Treasury also resumed issuing 30-year instruments, known as *Treasury bonds*. *Bonds* also pay semi-annual coupons, and make up 12.5% of the US debt.

The price of both notes and bonds are quoted as a percentage of their face value in thirty-seconds of a point. A quoted price of 98-08 means that the quoted price of the note (or bond) is $(98 + 8/32 =)$ $98.25 for each $100 of face value. The cash price of bonds and notes is equal to the quoted price plus accrued interest since the last coupon payment, calculated with an actual/actual day count convention. Quoted prices are sometimes called "clean" prices, while cash prices are said to be "dirty".

The US Treasury also issues 5-, 10-, and 20-year Treasury Inflation–Protected Securities ("TIPS"), whose payoff is linked to changes in the US Consumer Price Index (CPI). These make up about 10.2% of the total value of Treasuries outstanding. The principal value of TIPS is adjusted daily and the semi-annual coupon payments and principal payment are then based on the adjusted principal amount. Economists extract inflation forecasts by comparing the TIPS yields to those on similar nominal instruments. The Federal Reserve Bank of Saint Louis provides "TIPS spreads" through its publication, *Monetary Trends*.

There is also an active market in STRIPS (Separate Trading of Registered Interest and Principal of Securities) which are popularly known as "zero coupon" bonds. These instruments are created by the Treasury through an accounting system which separates coupon interest payments and principal. Finally, the US Treasury also is-

sues savings bonds, low denomination securities for retail investors.

Treasury Market Participants

The Federal Reserve in the Treasury Market

The Federal Reserve Bank of New York, under the guidance of the Federal Open Market Committee (FOMC), is a uniquely important player in the Treasury market. The FOMC meets approximately every six weeks to review economic conditions and determine a target for the federal funds rate, the rate at which US banks borrow/lend reserve balances from/to each other. The manager of the Open Market Desk (a.k.a., "the Desk") at the Federal Reserve Bank of New York is responsible for ensuring that the average federal funds transaction is close to the target by buying and selling Treasury instruments (primarily short-term). In practice, the Desk accomplishes this in two ways. First the Desk buys sufficient Treasuries to satisfy most but not all the markets' demand for deposits at the Fed. Secondly, the Desk buys Treasuries via repurchase (repos) agreements (overnight and for terms of several days) to achieve a desired repo rate that influences the federal funds rate and other short-term interest rates through arbitrage.

To determine day-to-day actions, every morning, staff at both the Division of Monetary Affairs of the Board of Governors of the Federal Reserve System and the Desk forecast that day's demand for reserve balances. The Desk staff also consults market participants to get their views on financial conditions. The relevant Desk and Board staffs then exchange views in a 9 am conference call. Finally, the relevant Desk staff, the Board staff, and at least one of the voting Reserve Bank Presidents then confer during a second conference call at about 9:20 am. The Desk staff summarizes market conditions, projects actions for the day and asks the voting Reserve Bank President(s) for comments. Open market operations commence shortly after the conclusion of this call.

When the Desk buys Treasuries, it increases available liquidity (reserves) in debt markets and tends to lower interest rates. Selling Treasuries has the opposite effect, lowering reserves and raising interest rates. If the intention is to make a permanent change in reserves, then outright purchases or sales are undertaken. In contrast, if the Desk anticipates that only temporary changes in reserves are necessary, it uses repos (for purchases) or reverse repos (for sales). Bernanke [7] notes that actual open market sales of debt instruments are rare; it is more common for the Federal Reserve to allow such securities to expire without replacing them. Both open market sales and allowing the Fed's securities to expire have the same balance sheet effects: The Fed holds fewer bonds and more cash, while the public will hold more bonds and less cash.

The Federal Reserve provides several valuable references on its operating procedures. The Annual Report of the Markets Group of the Federal Reserve Bank of New York describes open market operations and current procedures (Federal Reserve Bank of New York, Markets Group [26]). Meulendyke [57] provides a comprehensive view of Federal Reserve monetary policy operations with a historical perspective. Akhtar [1] explains how monetary policy is decided and how such policies affect the economy. Finally, Harvey and Huang [43] gives some historical perspective on operating procedures in the 1980s.

Primary Dealers

Among the most important private sector players in the Treasury markets are the 21 *primary dealers*. The Federal Reserve Bank of New York explains that primary dealers must "participate meaningfully in both the Fed's open market operations and Treasury auctions and … provide the Fed's trading desk with market information and analysis that are helpful in the formulation and implementation of monetary policy". The Federal Reserve does not regulate primary dealers, but does subject them to capital requirements. The Federal Reserve can withdraw a firm's primary dealer designation if it fails to participate in auctions or open market operations or if its capital reserves fall below desired levels.

Interdealer Brokers

Prior to 2000, voice-assisted brokers dominated secondary market trading in Treasuries. Except for Cantor–Fitzgerald, all these brokers reported their trading activity to GovPX, a consortium. In the face of demands by the Securities and Exchange Commission and bond market dealers for greater transparency, five IDBs formed GovPX as a joint venture in 1991. In March 1999, Cantor–Fitzgerald opened up its internal electronic trading platform, eSpeed, to clients. The eSpeed system quickly grabbed a dominant market share, and Cantor Fitzgerald spun off eSpeed as a public company in December 1999. In 2000, a competing electronic brokerage, BrokerTec, joined the market. As in foreign exchange and equity markets, most interdealer and institutional trading in Treasuries quickly migrated from voice networks to these electronic communications networks (ECNs), which have dominated trading in Treasury instruments since 2001. Mizrach and Neely [58] describe the transition from voice assisted trading, largely through the primary dealers, to electronic trading in the Treasury market.

As of November 2007, the two dominant ECNs are eSpeed and BrokerTec. London-based ICAP, PLC, owns BrokerTec while eSpeed merged in the summer of 2007 with BGC, another London based interdealer brokerage. eSpeed and ICAP compete for both on- and off-the-run liquidity. Hilliard Farber and Tullett–Prebon hold the largest brokerage share outside of the dominant two platforms.

Stages of the Treasury Bond Market

The sale of Treasuries undergoes four distinct phases: when issued, primary, on-the-run and off-the-run. Each of these stages has a distinct market structure.

The Primary Market

In the *primary* market, the US Treasury sells debt to the public via auction. The US Treasury usually publishes a calendar of upcoming tentative auction dates on the first Wednesday of February, May, August, and November and bids may be submitted up to 30 days in advance of the auction. In practice, however, the Treasury only announces firm auction information several days in advance and most bids are submitted at that time. Since August 8, 2002, the Treasury has made auction announcements (for all new securities) at 11:00 am Eastern Time (ET). 13- and 26-week bills are auctioned weekly; 2- and 5-year notes are auctioned monthly; 10-year notes are auctioned eight times a year. 30-year bonds, which were reintroduced on February 9, 2006 after a five year hiatus, are auctioned four times a year.

The US Treasury has used a single price auction exclusively since November 1998. Garbade and Ingber [35] discuss the transition from multiple price auctions to the current format single price auctions. All securities are allocated to bidders at the price that, in the aggregate, will result in the sale of the entire issue. This mitigates the risk of a "buyer's curse" – the highest bidder paying more than other auction participants. To prevent a single large buyer from manipulating the auction, the Treasury restricts anyone from buying more than 35% of any single issue. Bids may be submitted up to thirty days prior to the auction, and large institutions make use of the Treasury Automated Auction Processing System (TAAPS). Retail investors can participate through the Treasury Direct program. The Treasury allocates a portion of nearly every auction to small investors at the same price as the large institutions. These are called *non-competitive bids*, and they are quantity only orders that are filled at the market clearing price.

Primary dealers dominate the auction process. In 2003, they submitted 86% of auction bids, totalling more than $6 trillion. They were awarded $2.4 trillion, or 78% of the total auction supply.

The Secondary Market

The secondary market is composed of the when-issued, on-the-run and off-the-run issues.

When-Issued Even prior to the primary auction, there is an active forward market in Treasury securities (apart from TIPS) that are about to be issued. Trading in the *when-issued* security market typically begins several days prior to an auction and continues until settlement of auction purchases. Nyborg and Sundaresan [61] document that when-issued trading provides important information about auction prices prior to the auction and also permits market participants to reduce the risk they take in bidding. Fabozzi and Fleming [25] estimate that 6% of total interdealer trading is in the when-issued market. Just prior to auctions though, these markets become substantially more active. In the bill market, when-issued trading volume exceeds the volume for the bills from the previous auction.

On-the-Run Upon completion of the auction, the most recently issued bill, note or bond becomes *on-the-run* and the previous on-the-run issue goes *off-the-run*. Overall Treasury trading volume is concentrated in a small number of on-the-run issues. Trading in these benchmark on-the-run issues, which Fabozzi and Fleming [25] say constitutes approximately 70% of total trading volume, has migrated almost completely to the electronic networks. Mizrach and Neely [58] estimate a 61% market share for the BrokerTec platform and a 39% share for eSpeed in 2005, which is consistent with industry estimates.

Off-the-Run With more than 200 off-the-run issues trading in October 2007 – 44 bills, 116 notes, and 45 bonds – most off-the-run volume takes place in voice and electronic interdealer networks. Barclay, Hendershott and Kotz [5] document the fall in ECN market share when issues go off the run. They also report that transaction volume falls by more than 90%, on average, once a bond goes off-the-run. The ECN market share falls from 75.2% to 9.9% for the 2-year notes, from 83.5% to 8.5% for the 5-year notes, and from 84.5% to 8.9% for the 10-year notes. Several IDBs handle most off-the-run securities trading.

On- Versus Off-the-Run Liquidity and Prices Off-the-run securities trade at a higher yield (lower price) than on-

the-run securities of similar maturity. Many researchers have attempted to explain the yield differential with relative liquidity. Vayanos and Weill [68] utilize a search theoretic model that is motivated by the fact that bonds may be difficult to locate once they go off-the-run. Goldreich, Hanke, and Nath [36] compare on-the-run and off-the-run Treasuries and show that the liquidity premium depends primarily on the amount of remaining future liquidity, which is highly predictable. The study exploits the fact that the liquidity of a Treasury is predictable. Duffie [18] argues that legal or institutional restrictions on supplying collateral induces "special" repo rates that are much less than market riskless interest rates. The price of the underlying instrument is increased by the present value of the savings in borrowing costs.

Supply Variation and Prices Although it is generally accepted that the on-the-run premium is due to greater liquidity, the theoretical relation between the supply of a given bond issue and prices is not clear. Do issue sizes produce lower yields (higher prices) through their liquidity effects or does downward-sloping demand for individual securities produce higher prices (lower yields) for larger issues? Empirically, the evidence is mixed. Simon [65,66], Duffie [18], Seligman [64] and Fleming [29] find that the larger issues lead to lower prices (higher yields), while Amihud and Mendelson [2], Kamara [51], Warga [69], and Elton and Green [23] find the opposite: The liquidity effect predominates, resulting in higher prices (lower yields) for larger issues. There might be a nonlinear relationship. Liquidity may increase prices up to a certain point, but then finite demand for any individual security reduces the attractiveness of additional supply.

The Treasury Futures Market

Spot markets are not the only markets for US Treasuries. The Chicago Board of Trade (CBOT) has active futures markets for 2-, 5-, 10- and 30-year US Treasuries. Table 1 briefly describes the CBOT contracts and pricing conventions.

Like other exchange-traded derivatives, Treasury futures have two advantages: trading is highly liquid and marking-to-market minimizes counterparty risk. The CBOT open auction trading hours are 7:20 am to 2:00 pm, Central Time, Monday through Friday; the CBOT electronic market functions from 6:00 pm to 4:00 pm, Central Time, Sunday through Friday. All Treasury contracts have a March–June-September–December cycle.

A variety of Treasury instruments meet the criteria to be deliverable issues. Table 1 describes the pricing conventions and the characteristics of the assets that may be delivered to satisfy the contracts. The CBOT defines "conversion factors" that adjust the quoted futures prices for the asset that is actually delivered. Despite these conversion factors, one issue will be the "cheapest to deliver". Cash prices at delivery depend on both the conversion factor for a particular bond and the interest accrued on that bond since the last coupon payment.

Treasury Market, Microstructure of the U.S., Table 1
Contract Details from the CBOT Treasury Market

Contract	Quote convention	Pricing example	Deliverable asset characteristics
2-year	1/32 and quarters of 32nds	$95-060 = 95+6/32$ $95-062 = 95+6.25/32$ $95-065 = 95+6.5/32$ $95-067 = 95+6.75/32$	US Treasury notes with a face value $\geq$ \$200,000 and original maturity $\leq$ 5 years and 3 months and remaining maturity $\geq$ 1 year and 9 months from the first day of the delivery month and and remaining maturity $\leq$ than 2 years from the last day of the delivery month.
5-year	1/32 and halves of 32nds	$90-170 = 90+17/32$ $90-175 = 90+17.5/32$	US Treasury notes with a face value $\geq$ \$100,000 and original maturity $\leq$ 5 years and 3 months and remaining maturity $\geq$ 4 year and 2 months from the first day of the delivery month
10-year	1/32 and halves of 32nds	$90-170 = 90+17/32$ $90-175 = 90+17.5/32$	US Treasury notes with a face value $\geq$ \$100,000 and remaining maturity $\leq$ 10 years remaining maturity $\geq$ 6 year and 6 months from the first day of the delivery month
30-year	1/32nds	$85-12 = 85+12/32$	US Treasury bonds with a face value $\geq$ \$100,000 and if callable: Not callable for at least 15 years from the first day of the delivery month; if not callable: Remaining maturity $\geq$ 15 years from the first day of the delivery month.

Although agents frequently use the futures markets for hedging or taking positions on future price movements, only a modest amount of microstructure research has focused on futures markets. Brandt, Kavajecz, and Underwood [11] show that futures and spot market order flow are useful in predicting daily returns in each market and that the type of trader influences the effect of order flow. Mizrach and Neely [59] show that futures markets contribute a substantial amount of price discovery to US Treasury markets. Campbell and Hendry [12] compare price discovery in the 10-year bond and futures contracts in both the United States and Canada.

Seasonality and Announcement Effects

Seasonality and announcement effects are intimately related to the microstructure literature in that the latter seeks to explain how markets with heterogeneous agents react to the release of information.

Seasonality and Macroeconomic Announcements

The earliest studies considered the issue of daily seasonality in Treasuries. Flannery and Protopapadakis [27] document differing day-of-the-week patterns in Treasuries and stock indices. The patterns in the prices of Treasuries securities vary by maturity and differ from those found in stock indices. They conclude that no single factor explains seasonal patterns across asset classes. In contrast to this day-of-the-week effect in spot T-bills, Johnston et al. [50] find day-of-the-week effects in government national mortgage association (GNMA) securities, T-note, and T-bond futures, but not in T-bill futures. The fact that day-of-the-week effects exist in spot T-bills but not in T-bill futures points up the importance of futures settlement rules.

Later studies began to consider the effects of macro announcements on price changes, volatility, volume and spreads. Macroeconomic announcements have been an especially popular subject of study because they occur at regular intervals that can be anticipated by market participants. The existence of survey expectations about upcoming macro announcements permits researchers to identify the "shock" component of the announcement, which allows them to investigate the differential effects of anticipated and unanticipated news releases of different magnitudes.

Ederington and Lee [20,21] did the seminal modern work with intraday data on macro announcement effects in bond markets. They found that volatility increases before the announcement and remains elevated for some time afterwards. The employment, PPI, CPI and durable goods orders releases produce the greatest impact of the 9 significant announcements, out of 16 studied. Ederington and Lee [22] follow up on their earlier studies by linking the literatures on seasonality and announcements in the bond market. Comparing the contributions of past volatility, seasonality and announcements in predicting intraday volatility bond futures data and exchange rates, these authors argue that announcements account for much of the apparent seasonality in interest rate volatility.

One of the earliest important results was that bond market prices react more strongly to macro announcements than do equity markets. Fleming and Remolona [32, 34] examined the 25 largest price changes in the GovPX data and related them all to macroeconomic announcements. Fleming and Remolona [34] note: "In contrast to stock prices, US Treasury security prices largely react to the arrival of public information on the economy". Fleming and Remolona [32,33] attribute the relative sensitivity of bond markets to the fact that bond prices depend only on expected discount rates while stock prices are also determined by future expected dividends. Macro announcements can have little or no effect on stock prices if their effects on expected dividends and discount rates offset each other.

Several studies used more sophisticated econometric procedures to evaluate the impact of announcements on persistence in volatility in a full model. Jones, Lamont and Lumsdaine [49] examine volatility patterns in the 5-year Treasury market around US announcements. Daily volatility from an ARCH-M does not persist for days after announcements and the authors interpret this as indicating that agents rapidly incorporate announcement information into prices. Weekly volatility displays a U-shaped pattern; the largest price changes occur on Mondays and Fridays. Further, Jones, Lamont and Lumsdaine [49] find a risk premium in returns on days of announcements. Bollerslev, Cai, and Song [8] also consider the interaction of announcements and persistence in volatility with 5-minute US Treasury bond data. Modeling the intraday volatility patterns and accounting for announcements reveals long-memory in bond market volatility.

An important issue in microstructure is the determination of bid-ask spreads. Balduzzi, Elton, and Green [4] use intraday GovPX data to look at the effects of macro announcements on volume, prices and spreads. Confirming previous findings, prices adjust to news within one minute while increases in volatility and volume persist for up to 60 minutes. Spreads initially widen but then return to normal after 5 to 15 minutes. News releases explain a substantial amount of bond market volatility. Importantly, Balduzzi, Elton, and Green [4] argue that the dif-

ferential impact of news on long and short bond prices indicates that at least two factors will be needed for models of the yield curve. They also present evidence that discontinuities (jumps) will be important in modeling bond prices.

Some recent papers have relaxed the restrictive assumption that announcements influence Treasury market variables in a linear, symmetric fashion. For example, Christie–David, Chaudhry, and Lindley [15] allow the effects of announcement shocks to depend on the size and sign of the shock. They measure these nonlinear effects on the intraday 10- and 30-year Treasury futures from 1992 to 1996.

Most studies of the effects of volatility have measured such variation with some function of squared returns. One can use the volatility implied by options prices, however, to measure expected volatility over longer horizons. Heuson and Su [45], for example, show that implied volatilities from options on Treasuries rise prior to macro announcements and that volatilities quickly return to normal levels after announcements. Beber and Brandt [6] use intraday, tick data from 1995 to 1999 to determine that macro announcements reduce the variance of the option-implied distribution of US Treasury bond prices. The content of the news and economic conditions explain these changes in higher-order moments. The study attributes the results to time-varying risk premia rather than relative mispricing or changing beliefs.

In a comprehensive study of the impact of US macroeconomic announcements across asset markets, Andersen, Bollerslev, Diebold and Vega [3] study the reaction of international equity, bond and foreign exchange markets. They confirm that US macroeconomic news drives bond prices, as well as those of the other assets.

Monetary Policy Announcements

Researchers have carefully investigated the effects of the Federal Reserve's actions on the Treasury market. While the literature has examined the effect of a wide variety of monetary policy behavior and communications – e. g., open market operations, FOMC news releases, speeches, etc. – on many aspects of Treasury market behavior, a large subset of these papers deal with one specific topic: The effect of federal funds target changes on the Treasury yield curve.

Federal Funds Target Changes and the Treasury Yield Curve The "expectations hypothesis of the term structure" motivates research on how the short- and long-end of the Treasury yield curve react to unexpected changes in the federal funds target rate. That is, if the FOMC increases overnight interest rates, how does this change short- and long-term rates?

Using data on 75 changes in the federal funds target from September 1974 through September 1979, Cook and Hahn [16] find that these target changes caused larger movements in short-term rates than in intermediate- and long-term Treasury rates. A difficulty with interpreting the Cook and Hahn [16] results is that efficient markets presumably can often anticipate most or all of a target change and such expectations are already incorporated into the yield curve. To confront this problem, Kuttner [53] decomposes target changes into anticipated and unanticipated components, finding –unsurprisingly – that Treasury rates respond much more strongly to unanticipated changes and that the results are consistent with the expectations hypothesis of the term structure. That is, the anticipated component of an interest rate change does not affect expectations. Hamilton [41] carefully reexamines the work of Kuttner [53], showing that it is robust to uncertainty about the dates of target changes and the effect of learning by market participants.

Poole and Rasche [62] also decompose federal funds target changes into expected and unexpected components – but use a later contract month than Kuttner [53] to avoid problems associated with computation of the contract payoff. They find that interest rates across the maturity spectrum fail to respond to the anticipated components of the changes in the intended funds rate.

Poole, Rasche and Thornton [63] consider how changes in FOMC procedures affect the impact of target changes on interest rates. This study first succinctly describes the changes in FOMC procedures in the 1990s. The FOMC began to contemporaneously announce policy actions in 1994 and adopted this as formal policy in 1995. Starting in August 1997, each policy directive has included the quantitative value of the "intended federal funds rate". And since 1999, the FOMC has issued a press release after each meeting with the value for the "intended federal funds rate" and, in most cases, an assessment of the balance of risks. After describing such procedural changes, Poole, Rasche and Thornton [63] go on to consider the response of the Treasury yield curve to funds rate target changes both before and after the FOMC began contemporaneously announcing target changes in 1994. In doing so, these authors account for measurement error in expectations and uncertainty about the dates of target changes and even whether market participants understood that the Federal Reserve was targeting the funds rate prior to 1994. They assess the market's knowledge of targeting by examining news reports. While short-rates respond similarly in

both subperiods, long rates do not respond as strongly to funds rate target changes after 1994. The authors interpret their results as being consistent with the Fed's greater transparency about long-run policy in the second subsample. With long-run expectations more firmly anchored, unexpected changes in the funds target have smaller effects on long rates.

One puzzle that has emerged from this literature is that the average effect of changes in the federal funds target on the yield curve is modest, despite the facts that such changes should be an important determinant of the yield curve and that yields are highly volatile around FOMC announcements. Fleming and Piazzesi [31] claim to partially resolve this puzzle by illustrating that such yield changes depend on the shape of the yield curve.

This literature on the reaction of the Treasury market to monetary policy has become progressively more sophisticated in assessing market expectations of Fed policy and modeling institutional features of the futures market and Fed operations. Nevertheless, the underlying conclusion that unanticipated target changes lead to large price increases on short-term Treasuries and smaller changes on the prices of long-term Treasuries has been remarkably robust.

Other Federal Reserve Behavior and the Treasury Market There has been a substantial literature analyzing how other types of Federal Reserve behavior have influenced the Treasury market. The literature has considered open market operations, FOMC statements, Congressional testimonies, and FOMC member speeches.

Open market operations are similar to macroeconomic announcements in that they are potentially important bond market events, occurring at regularly scheduled times. Harvey and Huang [43] used intraday data from 1982 to 1988 to examine how Federal Reserve open market operations influenced foreign exchange and bond markets. The paper finds that Treasury market volatility increases during open market operations, irrespective of whether they add or drain reserves. Oddly, volatility increases even more during the usual time for open market operations if there are no such transactions. The authors interpret this finding as indicating that open market operations actually smooth volatility.

Early studies made the simplifying assumption that the effect of macro announcements on the Treasury market was constant over time. This is not necessarily the case, of course. For example, the effect of macro announcements on the Treasury market might depend on monetary policy priorities. Kearney [52] characterizes the changing response of daily 3-month Treasury futures to the employment report over 1977 to 1997 and relates it to the changing importance of employment in the Fed's reaction function.

de Goeij and Marquering [17] also considers how both macro announcements and monetary policy events affect the US Treasury market. Using daily data from 1982 to 2004 de Goeij and Marquering [17] find that macro news announcements strongly affect the daily volatility of longer-term Treasury instruments while FOMC events affect the volatility of shorter-term instruments.

Some studies have explored more esoteric components of information about monetary policy. Boukus and Rosenberg [9], for example, use Latent Semantic Analysis to decompose the information content of FOMC minutes from 1987 to 2005. They then relate the information content to current and future economic conditions. Chirinko and Curran [13] argue that Federal Reserve speeches, testimonies, and meetings increase price and trading volatility on the 30-year bond market. FOMC meetings are the most important of the events considered. They go on to consider whether these Federal Reserve events merely create noise or transmit information about the future policy decisions or the state of the economy. They conclude that such events may reduce welfare by "overwhelming private information", creating herding behavior.

Announcements and Liquidity Variation

The literature on variation in liquidity and price effects overlaps with the literature on macroeconomic announcements. The seminal work of Amihud and Mendelson [2] showed that yields on short-time-to-maturity Treasuries vary inversely with liquidity. That is, more liquid assets have lower yields/higher prices. Harvey and Huang [43] discovered elevated volatility in interest rate (and foreign exchange) futures markets, in the first 60–70 minutes of trading on Thursdays and Fridays. Ederington and Lee [20] confirmed Harvey and Huang [43]'s speculation that major macroeconomic announcements – especially the employment report, the PPI, the CPI, and durable goods orders – create the intraday and intraweek patterns in the volatility of Treasury bond futures. Volatility is very high after announcements and remains elevated for hours. Fleming and Remolona [32] extend this work to show that the 25 greatest surges in activity in the 5-year on-the-run bond market came on macroeconomic announcement days, within 70 minutes of the announcement. The most important announcements for trading surges were employment reports, fed funds targets, 30-year auctions, 10-year auctions, the CPI, NAPM surveys, GDP, retail sales, and 3-year auctions. Releases that affect prices also

matter for trading activity. Fleming and Remolona [32] observe that timeliness, the degree of surprise in the announcement and market uncertainty also increase announcements' impact on trading.

Researchers continued to explore the impact of variation in liquidity caused by other events. For example, Fleming [28] exploits exogenous variation in Treasury issuance to show that securities that are "reopened" – the Treasury sells additional quantities of existing securities – have greater liquidity, lower spreads, than comparable assets. Paradoxically, this higher liquidity does not produce lower yields for the reopened securities.

More recent papers have explored variation in liquidity and volatility across markets. Chordia, Sarkar and Subrahmanyam [14] estimate a vector autoregression (VAR) in liquidity and volatility variables in stock and bond markets. They find that common factors make the variables' innovations highly correlated. Volatility shocks predict liquidity variables.

End-of-the-Year Patterns in One-Month Treasury Bills

The previous sets of papers studied daily and intraday seasonality, often as caused by macroeconomic or Federal Reserve announcements. Short-term Treasury bills also exhibit year-end seasonality, however. Market participants consider Treasury market instruments of 30 days or less to be highly liquid, close – but not perfect – substitutes for cash. The fact that short-term Treasuries are not perfect substitutes for cash is presumably what allows the New York Desk to use open market operations to manipulate short-term interest rates through a liquidity effect. A peculiar year-end pattern in one-month Treasury yields reinforces this evidence that such Treasuries are not perfect substitutes for cash.

Following on related work of Griffiths and Winters [40] in repos, Griffiths and Winters [39] find that yields on one month T-Bills (and other one-month securities) increase significantly at the beginning of December, remain high during December, and return to normal a few days before the year-end. This pattern does not exist in three-month T-bills. Neely and Winters [60] find similar patterns in the one-month LIBOR futures market.

Griffiths and Winters [38,39,40] explain this December effect by asserting that a year-end preference for liquidity drives the year-end surge in short-term interest rates. Debt holder (lenders in the money markets) start to liquidate their one-month securities in the last few days of November to meet cash obligations at the end of December. This preference for liquidity drives up one-month interest rates for most of December. Liquidity demand returns to normal at the end of December as investors repurchase short-term instruments, and interest rates return to normal levels.

Discontinuities in the US Treasury Market

The literature on discontinuities (or jumps) in Treasury prices is closely related to the literature on announcements, as announcements are obvious candidates to explain jumps. Three recent papers have looked at discontinuities in US Treasury prices. Huang [47] estimates daily jumps with bi-power variation on 10 years of 5-minute data on S&P 500 and US T-bond futures to measure the response of volatility and jumps to macro news. He identifies a major role for payroll news in bond market jumps by analyzing their conditional distributions and regressing continuous and jump components on measures of disagreement and uncertainty concerning future macroeconomic states. Huang [47] also finds that the bond market is relatively more responsive than the equity market.

Dungey, McKenzie, and Smith [19] estimate jumps and cojumps (simultaneous discontinuities in multiple markets) in the term structure of US Treasury rates. They find that the middle of the yield curve often cojumps with one of the ends, while the ends of the curve exhibit a greater tendency for idiosyncratic jumps. Macro news is strongly associated with cojumps in the term structure. Using BrokerTec data from 2003–2005, Jiang, Lo, and Verdelhan [48] extend this work by focusing on the role of liquidity shocks – estimated from the limit order book – in jumps and the relation of jumps to order flow and price discovery.

Lahaye, Laurent and Neely [54] examine jumps and cojumps across foreign exchange, stock, gold and 30-year Treasury futures. Discontinuities in bond futures prices were larger but less frequent than those in foreign exchange rates and smaller and about as frequent as those in equity markets. News announcements appear to cause many cojumps of bond prices with prices of other types of assets.

Order Flow in the US Treasury Market

The effect of order flow on prices has been a popular recent topic in microstructure. Several papers have explored the impact of order flow on prices and the ways in which macro/monetary announcements influence these impacts.

Huang, Cai, and Wang [46] use intraday 1998 GovPX spot data on the 5-year Treasury note to characterize trading patterns of primary dealers, announcement effects and

volatility-volume relations. The paper finds that both public information (i. e., announcements) and dealer inventory/order flow affect trading frequency.

Green [37] uses the Madhavan, Richardson, and Roomans [55] model to study the impact of GovPX trading in 5-year around announcements. Order flow has its largest price impact after large macro surprises, times of greater uncertainty about the announcement, and times of high liquidity. Green [37] concludes that order flow does reveal information about riskless rates.

Brandt and Kavajecz [10] find that order flow imbalances can explain up to 26% of the day-to-day variation in yields on non-announcement days. In contrast to Green [37], they find that order flow has its strongest impact at times of low liquidity. Brandt, Kavacejz, and Underwood [11] extend the work of Brandt and Kavajecz [10] to control for trader type and macroeconomic announcements in explaining the impact of bond market order flow on futures prices.

Menkveld, Sarkar, and Van der Wel [56] confirm earlier conclusions that announcements have significant effects on 30-year Treasury yields and they also find that customer order flow is much more informative on announcement days than on non-announcement days. They go on to investigate the profits that different types of traders make on announcement and non-announcement days.

At high frequencies, order flow is highly autocorrelated. A dynamic analysis of the market resilience requires modeling this formally. We turn to empirical modeling of the Treasury market order book in the next section.

Modeling the Limit Order Book

A purchase or a sale of a Treasury bond influences prices directly as trades work their way up the supply or demand curves. We would like to know whether these effects are large and long-lasting. To address this question, we must introduce a dynamic model of the limit order book.

Hasbrouck [44] proposed to study intra-day price formation with a standard bivariate vector autoregressive (VAR) model. Time t here is measured in 1-minute intervals. Let r_t be the percentage change in the transaction price and x_t^0 be the sum of signed trade indicators (+1 for buyer initiated, −1 for seller initiated) over minute t. Treasury market data sets typically indicate trade initiation as a "hit" −1 or a "take" +1.

The bivariate vector autoregression assumes that causality flows from trade initiation to returns by permitting r_t to depend on the contemporaneous value for x_t^0, but not allowing x_t^0 to depend on contemporaneous r_t. The model for returns is specified as follows

$$\begin{bmatrix} r_t \\ x_t^0 \end{bmatrix} = \sum_{i=1}^{5} \begin{bmatrix} a_{r,i} \\ a_{x,i} \end{bmatrix} r_{t-i} + \begin{bmatrix} \sum_{i=0}^{15} b_{r,i} \\ \sum_{i=1}^{15} b_{x,i} \end{bmatrix} x_{t-i}^0 + \begin{bmatrix} u_{r,t} \\ u_{x,t} \end{bmatrix} . \quad (1)$$

Mizrach and Neely [58] use 5 lags of the return series and 15 lags of the signed trades. The market impact is then defined as the dynamic effect of a buy shock to the return series,

$$\frac{\partial r_{t+n}}{\partial x_t} . \quad (2)$$

Mizrach and Neely [58] provide 15 minute market impact estimates from the GovPX market in 1999. The 2-year note is most resilient with prices only 0.0042% higher following a buyer initiated trade. The 30-year bond is the least liquid, with prices rising 0.0229% following a buy order. Mizrach and Neely also report 2004 estimates for the Cantor electronic limit order book. Market impacts range from 45 to 88% lower in the more liquid eSpeed ECN market. Fleming and Mizrach [30] find further reductions in market impacts on the BrokerTec ECN for 2005 and 2006.

Price Discovery

A crucial issue in the market microstructure literature is *price discovery*. This is the process by which prices embed new information. In the Treasury market, price discovery occurs in both the secondary spot market and in the futures markets at the Chicago Board of Trade (CBOT). The degree to which each market contributes to price discovery is a natural issue to address.

To investigate relative price discovery in these two Treasury markets, Mizrach and Neely [59] follow Hasbrouck [44] and assume that the price series have a unit root, are cointegrated, and have an rth order VAR representation,

$$p_t = \Phi_1 p_{t-1} + \Phi_2 p_{t-2} + \cdots + \Phi_r p_{t-r} + u_t .$$

It follows that the N returns,

$$r_t = \begin{bmatrix} p_{1,t} - p_{1,t-1} \\ \vdots \\ p_{N,t} - p_{N,t-1} \end{bmatrix} = \Delta p_t , \quad (3)$$

have the convenient Engle–Granger [24] error-correction

representation,

$$\Delta p_t = \alpha z_{t-1} + A_1 \Delta p_{t-1} + \cdots + A_r \Delta p_{t-r-1} + u_t\,, \quad (4)$$

where z_t is an error-correction term of rank $N-1$.

We analyze price discovery using the moving average representation of our return process (3),

$$\Delta p_t = \Theta(L)\varepsilon_t\,. \quad (5)$$

The disturbances are mean zero and serially uncorrelated, $E[\varepsilon_{i,t}] = 0$ and $\mathrm{cov}[\varepsilon_{i,t}, \varepsilon_{i,t-r}] = 0$, but they may be contemporaneously correlated, $\mathrm{cov}[\varepsilon_{i,t}, \varepsilon_{j,t}] \neq 0$.

The information share is related to the long run impulse responses, $\Theta(1) = \sum_{j=0}^{\infty} \Theta(L^j)$, the permanent effect of the shock vector on the Treasury prices. Cointegration makes the long run multipliers common across all markets,

$$\Theta(1) = \begin{bmatrix} \theta_1 & \cdots & \theta_N \\ \vdots & & \vdots \\ \theta_1 & \cdots & \theta_N \end{bmatrix}. \quad (6)$$

To eliminate contemporaneous correlation among the error terms in (5), we decompose $\Omega = E\left[\varepsilon_t \varepsilon_t'\right]$, the $N \times N$ covariance matrix, to find a lower triangular matrix M, whose i, jth element we denote m_{ij}, such that $MM' = \Omega$. The Hasbrouck [44] information share for market j is defined as

$$H_j = \frac{\left[\sum_{i=j}^{n} \theta_i m_{ij}\right]^2}{\left[\sum_{i=1}^{n} \theta_i m_{i1}\right]^2 + \left[\sum_{i=2}^{n} \theta_i m_{i2}\right]^2 + \cdots + (\theta_n m_{nn})^2}, \quad (7)$$

where the θ_is are the elements of row i of the long-run multipliers in (6). Because the Choleski decomposition is not unique, the information share will vary with the order of the equations in the VAR.

Mizrach and Neely [59] pair spot and maturity matched futures for the 2-year, 5-year and 10-year on-the-run spot notes. This calculation requires us to adjust futures prices according to the on-the-run spot instruments with which we compare them. The CBOT provides adjustment factors for each instrument. These adjustments typically make a single bond the cheapest to deliver (CTD), but the CTD is typically off-the-run. Nevertheless, the CTD off-the-run bonds and the most liquid on-the-run bonds are very close substitutes – their daily returns are highly correlated – so it is reasonable to examine price discovery between futures prices and on-the-run bonds, despite the fact that they are not identical.

Mizrach and Neely [59] find that information shares rise with the growth of the GovPX market, but fall as the ECNs take market share from GovPX voice markets. The spot market share is highest for the 2-year note, reaching 86%, while the 10-year spot market share never exceeds 50%. In addition, relative market liquidity measures like spreads, trades and volatility each strongly explain daily relative price discovery shares. Mizrach and Neely [59] compute both upper and lower bound estimates of the information shares. They also report estimates based on the Harris, McInish and Wood [42] methodology.

Campbell and Hendry [12] find similar results for the Canadian government bond market. They find that the information share in the 10-year spot note is below 50% in nearly all their sample of several months between 2002 and 2004. Upper and Werner [67] find that price discovery in the German Bund is dominated by the futures market, and in times of stress, like the 1998 Long Term Capital Management Crisis, the spot market information share falls to essentially zero. Upper and Werner [67], however, compare the futures market to the relatively illiquid, CTD bonds. This might explain their finding that the spot market does very little price discovery.

Future Directions

This article has reviewed the microstructure of the US Treasury market. The Open Market Desk at the Federal Reserve Bank of New York plays a uniquely important role in the Treasury market by using transactions in those securities to adjust the level of bank reserves. Primary dealers are key players in both Treasury auctions and the Fed's open market operations. The Treasury market consists of several phases: when-issued, primary, on-the-run and off-the-run. Two ECNs, eSpeed and BrokerTec, intermediate the most active trading, during the on-the-run phase. The Treasury futures market at the CBOT complements trading in the spot market.

Treasury markets exhibit end-of-year, daily and intraday seasonality. Macro and Federal Reserve announcements are responsible for a substantial part of the daily and intraday seasonality. The literature studying the impact of order flows on Treasury prices has also considered how macro news and Federal Reserve actions influence such impact.

The futures markets in Chicago play an important role in price discovery, and a discussion of Treasury microstructure needs to take this into account. Both spot and futures markets are quite resilient and recent research on the Treasury ECNs suggest that the market continues to become more liquid. Fleming and Mizrach [30] report that volume has increased almost 5 times since 2001. This increase in trading volume accompanies a decline

in the importance of the primary dealers. The Financial Times reported in March 2007 that hedge funds accounted for 80% of trading activity in the Treasury market with only a 20% share for the primary dealers. One large fund alone, Citadel, accounts for 10% of the trading volume on eSpeed and BrokerTec. It was perhaps inevitable that trading by the millisecond would come to the Treasury market as it did to equities and foreign exchange. Perhaps we should only be surprised that it took so long.

The Treasury market plays a central role in the credit market. Times of financial crisis highlight the Treasury market's role as a safe haven for investors both in the US and overseas. Treasury securities also serve as benchmarks for complex derivatives like mortgage backed securities and structured loans like collateralized debt obligations. The microstructure of the US Treasury market is fundamental to our understanding of the global financial markets.

Bibliography

1. Akhtar MA (1997) Understanding open market operations. Public Information Department, Federal Reserve Bank of New York, New York
2. Amihud Y, Mendelson H (1991) Liquidity, maturity, and the yields on US treasury securities. J Finance 46:1411–25
3. Andersen TG, Bollerslev T, Diebold FX, Vega C (2007) Real-time price discovery in stock, bond and foreign exchange markets. J Int Econ 73:251–77
4. Balduzzi P, Elton EJ, Green TC (2001) Economic news and bond prices: Evidence from the US treasury market. J Financial Quant Anal 36:523–43
5. Barclay MJ, Hendershott T, Kotz K (2006) Automation versus intermediation: Evidence from treasuries going off the run. J Finance 61:2395–2414
6. Beber A, Brandt MW (2006) The effect of macroeconomic news on beliefs and preferences: Evidence from the options market. J Monetary Econ 53:1997–2039
7. Bernanke BS (2005) Implementing monetary policy, remarks at the redefining investment strategy education symposium, Dayton. This is a public speech by Bernanke http://www.federalreserve.gov/boarddocs/speeches/2005/20050330/default.htm
8. Bollerslev T, Cai J, Song FM (2000) Intraday periodicity, long memory volatility, and macroeconomic announcement effects in the us treasury bond market. J Empir Finance 7:37–55
9. Boukus E, Rosenberg JV (2006) The information content of FOMC minutes. Working paper, Federal Reserve Bank of New York, New York
10. Brandt MW, Kavajecz KA (2004) Price discovery in the us treasury market: The impact of orderflow and liquidity on the yield curve. J Finance 59:2623–2654
11. Brandt MW, Kavajecz KA, Underwood SE (2007) Price discovery in the treasury futures market. J Futur Mark 27:1021–1051
12. Campbell B, Hendry S (2007) Price discovery in canadian and US 10-year government bond markets. Working Paper 07–43, Bank of Canada, Ottawa
13. Chirinko RS, Curran C (2006) Greenspan shrugs: Formal pronouncements, bond market volatility, and central bank communication. Presented at: The American Economic Association Meetings
14. Chordia T, Sarkar A, Subrahmanyam A (2005) An empirical analysis of stock and bond market liquidity. Rev Financial Stud 18:85–129
15. Christie-David R, Chaudhry M, Lindley JT (2003) The effects of unanticipated macroeconomic news on debt markets. J Financial Res 26:319–39
16. Cook T, Hahn T (1989) The effect of changes in the federal funds rate target on market interest rates in the 1970s. J Monetary Econ 24:331–351
17. de Goeij P, Marquering W (2006) Macroeconomic announcements and asymmetric volatility in bond returns. J Bank Finance 30:2659–2680
18. Duffie D (1996) Special repo rates. J Finance 51:493–526
19. Dungey M, McKenzie M, Smith V (2007) News, no-news and jumps in the us treasury market. Cambridge University (unpubl)
20. Ederington LH, Lee JH (1993) How markets process information: News releases and volatility. J Finance 48:1161–91
21. Ederington LH, Lee JH (1995) The short-run dynamics of the price adjustment to new information. J Financial Quant Anal 30:117–34
22. Ederington LH, Lee JH (2001) Intraday volatility in interest-rate and foreign-exchange markets: ARCH, announcement, and seasonality effects. J Futur Mark 21:517–52
23. Elton EJ, Green TC (1998) Tax and liquidity effects in pricing government bonds. J Finance 53:1533–1562
24. Engle R, Granger C (1987) Co-integration and error correction representation, estimation and testing. Econometrica 55: 251–276
25. Fabozzi FJ, Fleming MJ (2005) US treasury and agency securities. In: Fabozzi FJ (ed) The Handbook of Fixed Income Securities, 7th edn. McGraw Hill, New York, pp 229–250
26. Federal Reserve Bank of New York, Markets Group (2007) Domestic open market operations during 2006. Federal Reserve Bank of New York. http://www.ny.frb.org/markets/omo/omo2006.pdf
27. Flannery M, Protopapadakis A (1988) From T-bills to common stocks: Investigating the generality of intra-week return seasonalities. J Finance 43:431–450
28. Fleming M (2002) Are larger treasury issues more liquid? Evidence from bill reopenings. J Money Credit Bank 34: 707–735
29. Fleming M (2003) Measuring treasury market liquidity. Fed Reserv Bank New York Econ Policy Rev 9:83–108
30. Fleming M, Mizrach B (2008) The microstructure of a US treasury ECN: The BrokerTec platform. Working paper. Federal Reserve Bank of New York, New York
31. Fleming M, Piazzesi M (2005) Monetary policy tick-by-tick. Working paper No ID. Federal Reserve Bank of New York, New York
32. Fleming M, Remolona EM (1997) What moves the bond market? Fed Reserv Bank New York Econ Policy Rev 3:31–50
33. Fleming M, Remolona EM (1999) Price formation and liquidity in the US treasury market: The response to public information. J Finance 54:1901–1915
34. Fleming M, Remolona EM (1999) What moves bond prices? J Portf Manag 25:28–38

35. Garbade KD, Ingber JF (2005) The treasury auction process: Objectives, structure, and recent adaptations. Fed Reserv Bank New York Curr Issues Econ Finance 11:1–11
36. Goldreich D, Hanke B, Nath P (2005) The price of future liquidity: Time-varying liquidity in the US treasury market. Rev Finance 9:1–32
37. Green TC (2004) Economic news and the impact of trading on bond prices. J Finance 59:1201–1233
38. Griffiths M, Winters D (1997) On a preferred habitat for liquidity at the turn-of-the-year: Evidence from the term-repo market. J Financial Serv Res 12:21–38
39. Griffiths M, Winters D (2005) The turn-of-the-year in money markets: Tests of risk-shifting window dressing and preferred habitat hypotheses. J Bus 78:1337–1364
40. Griffiths M, Winters D (2005) The year-end price of risk in a market for liquidity. J Invest Manag 3:99–109
41. Hamilton JD (2008) Assessing monetary policy effects using daily fed funds futures contracts. Fed Reserv Bank St Louis Rev 90:377–393
42. Harris F, McInish T, Wood R (2002) Security price adjustment across exchanges: An investigation of common factor components for dow stocks. J Financial Mark 5:277–308
43. Harvey CR, Huang RD (2002) The impact of the federal reserve bank's open market operations. J Financial Mark 5:223–57
44. Hasbrouck J (1991) Measuring the information content of stock trades. J Finance 46:179–207
45. Heuson AJ, Su T (2003) Intra-day behavior of treasury sector index option implied volatilities around macroeconomic announcements. Financial Rev 38:161–77
46. Huang RD, Cai J, Wang X (2002) Information-based trading in the treasury note interdealer broker market. J Financial Intermed 11:269–296
47. Huang X (2006) Macroeconomic news announcements, financial market volatility and jumps. Duke University (unpubl)
48. Jiang GJ, Lo I, Verdelhan A (2007) Why do bond prices jump? A study of the US treasury market. Eller College of Management, University of Arizona (unpubl)
49. Jones CM, Lamont O, Lumsdaine RL (1998) Macroeconomic news and bond market volatility. J Financial Econ 47:315–337
50. Johnston E, Kracaw W, McConnell J (1991) Day-of-the-week effects in financial futures: An analysis of GNMA, T-bond, T-note, and T-bill contracts. J Financial Quant Anal 26:23–44
51. Kamara A (1994) Liquidity, taxes, and short-term treasury yields. J Financial Quant Anal 29:403–417
52. Kearney AA (2004) The changing impact of employment announcements on interest rates. J Econ Bus 54:415–429
53. Kuttner KN (2001) Monetary policy surprises and interest rates: Evidence from the fed funds futures market. J Monetary Econ 47:523–544
54. Lahaye J, Laurent S, Neely CJ (2007) Jumps, cojumps and macro announcements. Working Paper 2007–032A, Federal Reserve Bank of St. Louis, St. Louis
55. Madhavan A, Richardson M, Roomans M (1997) Why do securities prices change? A transaction-level analysis of NYSE stocks. Rev Financial Stud 10:1035–1064
56. Menkveld AJ, Sarkar A, Van der Wel M (2006) Customer flow, intermediaries, and the discovery of the equilibrium riskfree rate. Working paper. Federal Reserve Bank of New York, New York
57. Meulendyke AM (1998) US monetary policy & financial markets. Federal Reserve Bank of New York, New York
58. Mizrach B, Neely CJ (2006) The transition to electronic communication networks in the secondary treasury market. Fed Reserv Bank St. Louis Rev 88:527–541
59. Mizrach B, Neely CJ (2008) Information shares in the US treasury market. J Bank Finance 32:1221–1233
60. Neely CJ, Winters DB (2006) Year-end seasonality in one-month LIBOR derivatives. J Derivatives 13:47–65
61. Nyborg KG, Sundaresan S (1986) Discriminatory versus uniform treasury auctions: Evidence from when-issued transactions. J Financial Econ 42:63–104
62. Poole W, Rasche RH (2000) Perfecting the market's knowledge of monetary policy. J Financial Serv Res 18:255–298
63. Poole W, Rasche RH, Thornton DL (2002) Market anticipations of monetary policy actions. Fed Reserv Bank St. Louis Rev 84:65–94
64. Seligman J (2006) Does urgency affect price at market? An analysis of US treasury short term finance. J Money Credit Bank 38:989–1012
65. Simon DP (1991) Segmentation in the treasury bill market: Evidence from cash management bills. J Financial Quant Anal 26:97–108
66. Simon DP (1994) Further evidence on segmentation in the treasury bill market. J Bank Finance 18:139–151
67. Upper C, Werner T (2002) Tail wags dog? Time-varying information shares in the bund market. Working Paper 24/02, Bundesbank, Frankfurt
68. Vayanos D, Weill P (2008) A search-based theory of the on-the-run phenomenon. J Finance 63:1361–1398
69. Warga A (1992) Bond returns, liquidity, and missing data. J Financial Quant Anal 27:605–17

Tsunami Earthquakes

Jascha Polet[1], H. Kanamori[2]
[1] Geological Sciences Department, California State Polytechnic University, Pomona, USA
[2] Seismological Laboratory, Caltech, Pasadena, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Characteristics of Tsunami Earthquakes
Factors Involved in the Seismogenesis and Tsunamigenesis of Tsunami Earthquakes
A Model for Tsunami Earthquakes
Future Directions
Bibliography

Glossary

m_b body wave magnitude, based on the amplitude of the direct P wave, period of the measurement: 1.0–5.0 s. Also see: Seismic Magnitude.

M_S surface wave magnitude, based on the amplitude of surface waves, period of the measurement: 20 s. Also see: Seismic Magnitude.

M_w moment magnitude, determined from the seismic moment of an earthquake, typical period of the measurement: > 200 s. Also see: Seismic Magnitude.

Magnitude saturation due to the shape of the seismic source spectrum, relatively short period measurements of seismic magnitude will produce similar magnitudes for all earthquakes above a certain size. The value of this threshold earthquake size depends on the period of the measurement: magnitude measurements using shorter period waves will saturate at lower values than magnitude measurements using longer period waves. M_w will not saturate.

Run-up height difference between the elevation of maximum tsunami penetration (inundation line) and the sea level at the time of the tsunami.

Tsunami earthquake an earthquake that directly causes a regional and/or teleseismic tsunami that is greater in amplitude than would be expected from its seismic moment magnitude.

Tsunami magnitude a scale for the relative size of tsunamis generated by different earthquakes, M_t in particular is calculated from the logarithm of the maximum amplitude of the tsunami wave measured by a tide gauge distant from the tsunami source, corrected for the distance to the source (also see: Satake, this volume).

Seismic magnitude a scale for the relative size of earthquakes. Many different scales have been developed, almost all based on the logarithmic amplitude of a particular seismic wave on a particular type of seismometer, with corrections for the distance between source and receiver. These measurements are made for different wave types at different frequencies, and thus may lead to different values for magnitude for any one earthquake.

Seismic moment the product of the fault surface area of the earthquake, the rigidity of the rock surrounding the fault and the average slip on the fault.

Definition of the Subject

The original definition of "tsunami earthquake" was given by Kanamori [37] as "an earthquake that produces a large size tsunami relative to the value of its surface wave magnitude (M_S)". Therefore, the true damage potential that a tsunami earthquake represents may not be recognized by conventional near real-time seismic analysis methods and may only become apparent upon the arrival of the tsunami waves on the local coastline. Although tsunami earthquakes occur relatively infrequently, the effect on the local population can be devastating, as was most recently illustrated by the July 2006 Java tsunami earthquake. This event (moment magnitude $M_w = 7.8$) was quickly followed by tsunami waves two to seven meters high, traveling as far as two kilometers inland and killing at least 668 people (http://www.searo.who.int/en/Section23/Section1108/Section2077_11956.htm).

It is important to note that the definition of "tsunami earthquake" is distinct from that of "tsunamigenic earthquake". A tsunamigenic earthquake is any earthquake that excites a tsunami. Tsunami earthquakes are a specific subset of tsunamigenic earthquakes, which we will later in this article more precisely define as earthquakes that directly cause a regional and/or teleseismic tsunami that is greater in amplitude than would be expected from their seismic moment magnitude.

Introduction

Shallow oceanic earthquakes may excite destructive tsunamis. Truly devastating tsunamis occur only infrequently, but as the natural disaster of the tsunami following the 2004 Sumatra–Andaman Islands earthquake has shown, may cause widespread damage (in this case in the region of the Indian Ocean) and lead to hundreds of thousands of casualties. In general, tsunamis are caused by shallow earthquakes beneath the ocean floor displacing large volumes of water. Thus, the magnitude of the earthquake plays an important role in determining its tsunamigenic potential. However, a particular subclass of shallow subduction zone earthquakes: "tsunami earthquakes", poses a special problem.

For the purpose of this article, we will define the term "tsunami earthquake" as follows: an earthquake that directly causes a regional and/or teleseismic tsunami that is greater in amplitude than would be expected from its seismic moment magnitude. With this definition we specifically exclude seismic events that were followed by tsunamis directly caused by slides or slumps resulting from the original earthquake (as was the case for the 1992 Flores [28,32] and the 1998 Papua New Guinea earthquakes [22,73] for example). We further exclude events that only very locally caused large tsunamis as a result of, for example, focusing effects due to features of the ocean floor bathymetry (e. g. [63]) or directivity effects combined with the shape of the coastline, as was the case for the tsunamis that hit the harbor of Crescent City after the 1964 Aleutian Islands earthquake [12] and the November 15, 2006 Kurile Island event (http://www.usc.

edu/dept/tsunamis/california/Kuril_2006/). Furthermore, this definition compares the size of the tsunami with the moment magnitude of the earthquake and not its surface wave magnitude, slightly modifying the definition given by Kanamori [37], in order to exclude great events for which the surface wave magnitude saturates.

Our primary objective in this article is to describe the characteristics of tsunami earthquakes and the possible factors involved in the anomalously strong excitation of tsunamis by these events. We will also discuss a possible model for these infrequent, but potentially very hazardous, events. The earthquakes listed in Table 1 and plotted in Fig. 1 are considered tsunami earthquakes, according to our modified definition presented in the previous paragraph, by the majority of the community of earthquake and tsunami researchers. However, we note that the interpretation of the 1994 Java and the 1946 Aleutian Islands earthquakes varies with investigators. The 1994 Java earthquake occurred off the southeastern coast of this island, near the east end of the Java Trench in the Indian Ocean, at 1:18 am local time. It generated a devastating tsunami that took the lives of more than 200 East Java coastal residents. Run-up measured along the southeastern Java coast ranged from 1 to 14 m, while run-up measured along the southwestern coast of Bali ranged from 1–5 m [74,82]. Although the anomalously high tsunami excitation of the 1994 event is not in doubt, its earthquake source characteristics have been debated [3,59]. The 1946 Aleutian Islands earthquake off Unimak Island produced one of the largest trans-Pacific tsunamis and had a tsunami magnitude of 9.3 [1], but its moment magnitude is only $M_w = 8.2$, making it a tsunami earthquake [36]. Some of the great tsunami heights measured (exceeding 30 m in height on Unimak Island and 16 m in run-up at the Hawaiian islands; [78]) can be attributed to slumping [19]. However, its anomalously high tsunamis are probably primarily due to the seismic source directly [47].

The other, less controversial, tsunami earthquakes listed in our table are: the 1896 Sanriku event near the coast of Japan, two events near the Kurile Islands: one in 1963 and the other in 1975, the 1992 Nicaragua earthquake, the Peru earthquake 4 years later, as well as an earlier event in this region in 1960 and most recently the 2006 Java earthquake. The June 15, 1896 Sanriku earthquake generated devastating tsunamis with a maximum run-up of 25 m and caused the worst tsunami disaster in the history of Japan with over 20,000 deaths, despite its moderate surface wave magnitude ($M_S = 7.2$) and weak seismic intensity [2,26,76]. The November 20, 1960 Peru earthquake excited a tsunami that was anomalously large for an earthquake of moderate magnitude [57], resulting in 66 fatalities (from the tsunami event database of the National Geophysical Data Center, http://www.ngdc.noaa.gov/nndc/struts/form?t=101650&s=70&d=7). The October 20, 1963 Kurile earthquake was an aftershock to the great Kurile Islands underthrusting earthquake ($M_w = 8.5$) of October 13, 1963 and produced a maximum run-up height of 10–15 m at Urup Island, much larger than the height of the main shock tsunami of 5 m [2]. The 1975 earthquake occurred south of the Kurile Islands and was weakly felt along the entire southern part of the Kurile Islands. Like the 1963 tsunami earthquake, this event can be considered an aftershock [21] of a larger event ($M_S = 7.7$) that occurred essentially at the same location on June 17, 1973. The maximum run-up height was 5 m on Shikotan Island, while the main shock had a run-up height measured at 4.5 m. A fairly strong tsunami was also recorded on tide gauges in Alaska and Hawaii (from the tsunami event database of the National Geophysical Data Center http://www.ngdc.noaa.gov/nndc/struts/form?t=101650&s=70&d=7). After a time period of almost two decades without a significant tsunami earthquake, the 1992 Nicaragua earthquake was the first tsunami earthquake to be captured by modern broadband seismic networks. This tsunami caused 179 deaths (from the Emergency Disasters Database, http://www.em-dat.net/disasters/list.php) and significant damage to the coastal areas of Nicaragua and Costa Rica, reaching heights of up to 8 m. The 1996 Peru earthquake struck at 7:51 am local time, approximately 130 km off the northern coastal region of Peru. Approximately one hour after the main shock, a damaging tsunami reached the Peruvian coast, with run-up heights of 1 to 5 meters along a coastline of 400 km [27], resulting in twelve deaths [11]. Finally,

Tsunami Earthquakes, Table 1

Tsunami Earthquakes (references for most entries are listed in main text, others are from the National Geophysical Data Center Global Tsunami Database (http://www.ngdc.noaa.gov/seg/hazard/tsu_db.shtml) and the Centennial Earthquake Catalog [18])

Date	Geographical Region	M_w	m_b	M_S	M_t	Deaths
1896/06/15	Japan			7.2	8.0	26360
1946/04/01	Aleutian Islands	8.2		7.3	9.3	165
1960/11/20	Peru	7.6	7.0	7.0		66
1963/10/20	Kurile Islands	7.8	7.1	7.2		
1975/06/10	Kurile Islands	7.5	5.6	7.0		
1992/09/02	Nicaragua	7.7	5.4	7.2		179
1994/06/02	Java	7.8	5.7	7.1		250
1996/02/21	Peru	7.5	5.8	6.6		12
2006/07/17	Java	7.7	6.2	7.2		668

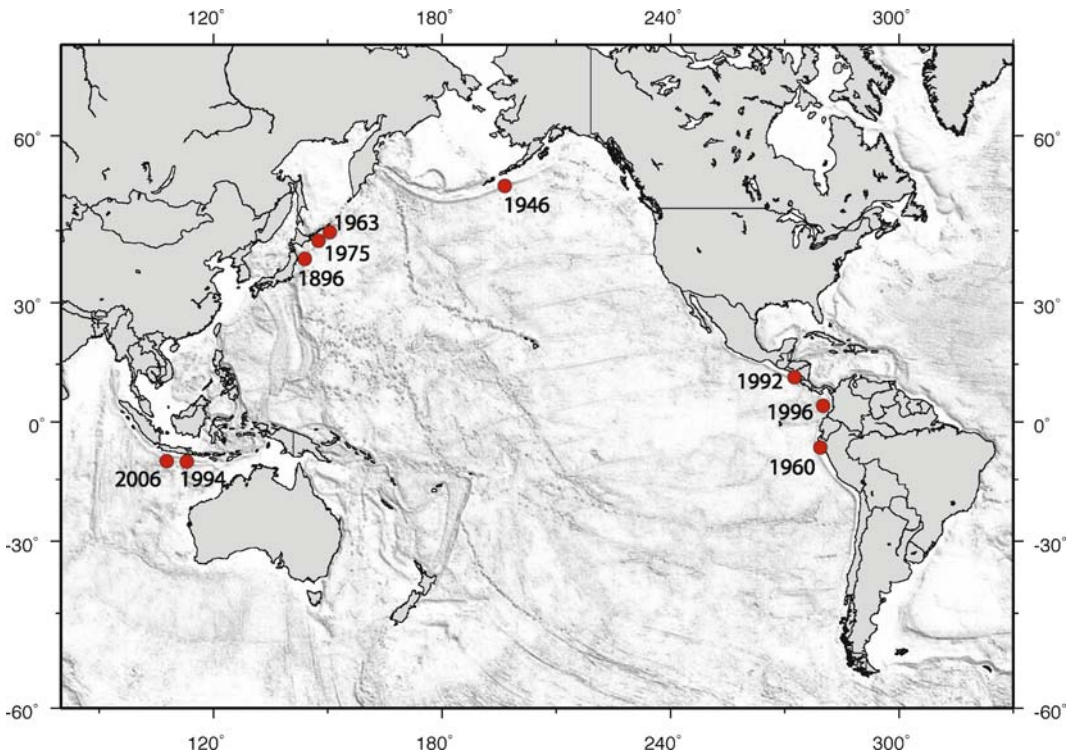

Tsunami Earthquakes, Figure 1
Map of tsunami earthquakes (listed in Table 1). Location for 1896 earthquake from [2] and for 2006 earthquake from the Global CMT catalog. All other earthquake locations from the Centennial Earthquake Catalog [18]

the 2006 Java earthquake was located only about 600 km west-northwest of the tsunami earthquake that occurred 12 years earlier in the same subduction zone. The Ministry of Health reported that approximately 668 people died and 65 are missing (http://www.searo.who.int/en/Section23/Section1108/Section2077_11956.htm) due to a tsunami that had a maximum run-up height of 15.7 m along the coast of central Java [45].

Several other earthquakes in the last few years have produced damaging tsunamis and have been mentioned as possible tsunami earthquakes. Seno and Hirata [69] suggest that the great 2004 Sumatra–Andaman earthquake also likely involved a component of tsunami earthquakes, because tsunamis larger than expected from seismic slip occurred, possibly due to slow slip in the shallow subduction boundary. It has also been proposed that the Kurile Islands earthquake of November 15, 2006 [35] may have exhibited some characteristics of tsunami earthquakes and, even more recently, the Solomon Island earthquake of April 1, 2007 excited large tsunamis, at least locally (http://soundwaves.usgs.gov/2007/04/). However, the disparity between seismic and tsunami excitation by these events is not nearly as large as for the events in Table 1, and we do not list these events as tsunami earthquakes.

Characteristics of Tsunami Earthquakes

Fortunately, tsunami earthquakes occur only infrequently. Fewer than ten of these events have occurred in the last three decades since the global installation of seismic broadband instruments and tide gauges and easy availability of their data were established. However, from the detailed investigations of the most recent events and comparisons with the limited data available for the older earthquakes in Table 1, several characteristics of these earthquakes clearly emerge.

Slow Character

The slow character of tsunami earthquakes manifests itself in several different, yet related, ways. One well-established characteristic of tsunami earthquakes is the discrepancy between the determined values of the different seismic magnitude types, calculated from various kinds of seismic waves or waves of different frequency ranges.

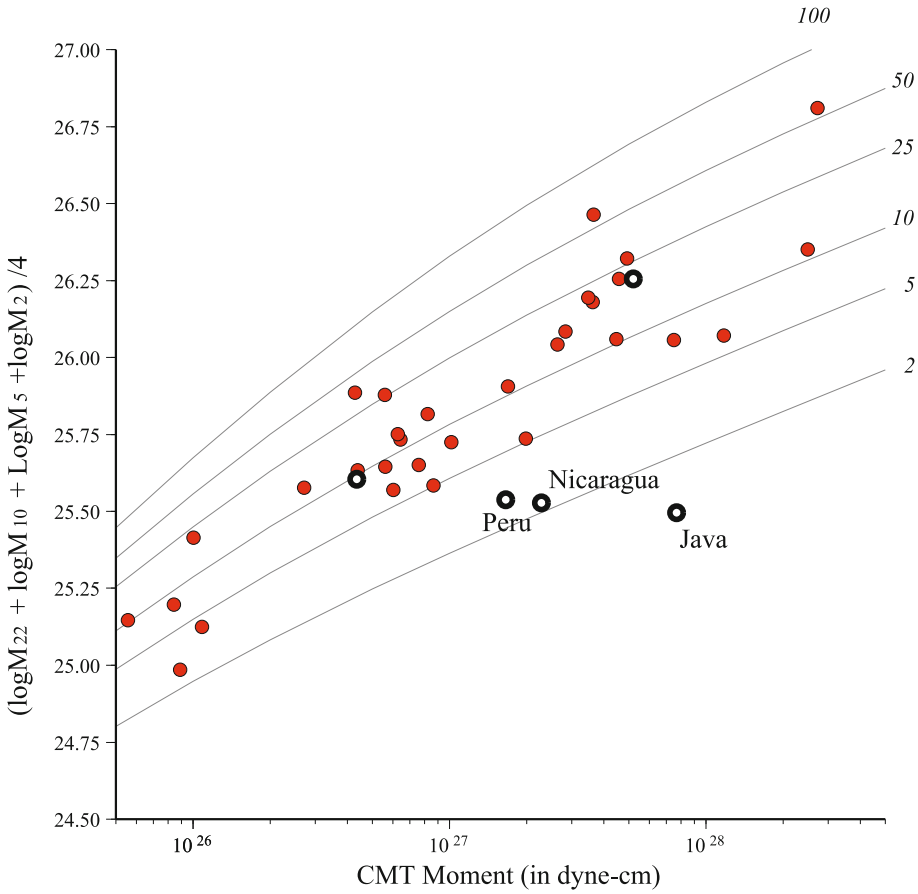

Tsunami Earthquakes, Figure 2

Average of log of moment rate spectrum at four periods (2, 5, 10, 22 s) as a function of the seismic moment as determined by CMT inversion of long-period surface waves. Reference curves were calculated for an ω^2 model [13]. Values next to the *grey curves* indicate the stress drop used to calculate the reference curve. Events shown are all shallow subduction zone earthquakes from 1992 to 1996 with moment magnitude 7.0 or greater. Earthquakes followed by anomalously large tsunamis are indicated with *open circles*. Of these events, only the 1992 Nicaragua, 1994 Java and 1996 Peru earthquakes are slow tsunami earthquakes, as is shown in this figure by their relatively low moment rate spectrum at shorter periods. Adapted from [59]

A typical comparison is that of the body wave magnitude, m_b, or the surface wave magnitude, M_S, with the moment magnitude of the earthquake, M_w. For tsunami earthquakes m_b is typically much smaller than the other two magnitudes, M_S and M_w, and M_w typically exceeds M_S. The discrepancy between these different magnitudes is more pronounced than for regular subduction zone earthquakes with similar moment magnitudes. For example: for the 1992 Nicaragua earthquake: $m_b = 5.4$; $M_S = 7.2$, $M_w = 7.7$ [18]; for the 1994 Java tsunami earthquake: $m_b = 5.7$; $M_S = 7.1$, $M_w = 7.8$ [59]; for the 2006 Java earthquake: $m_b = 6.2$; $M_S = 7.2$, $M_w = 7.7$ [5]. Since the body wave magnitude is calculated from short period P-waves, the surface wave magnitude is determined by the amplitude of surface waves with a period of 20 seconds and the moment magnitude is generally based on longer periods for big events, this consistent discrepancy is an indication of the relatively greater seismic energy release at longer periods (or the "slow" character) of these tsunami events. Similarly, investigations of teleseismic P-waves [5,59] have shown that their source spectra are depleted in high frequency energy at periods shorter than 20 seconds as compared to other shallow subduction zone earthquakes (see Fig. 2, from [59]), as well as their own aftershocks (see Fig. 3, adapted from [60]). Modeling of the rupture processes shows that the rupture velocities for tsunami earthquakes are slower than for most other subduction zone earthquakes (for several events: [58] and [33]; for Aleutians 1946: [47]; for Nicaragua 1992: [40] and [30]; for Peru 1996: [31]; for

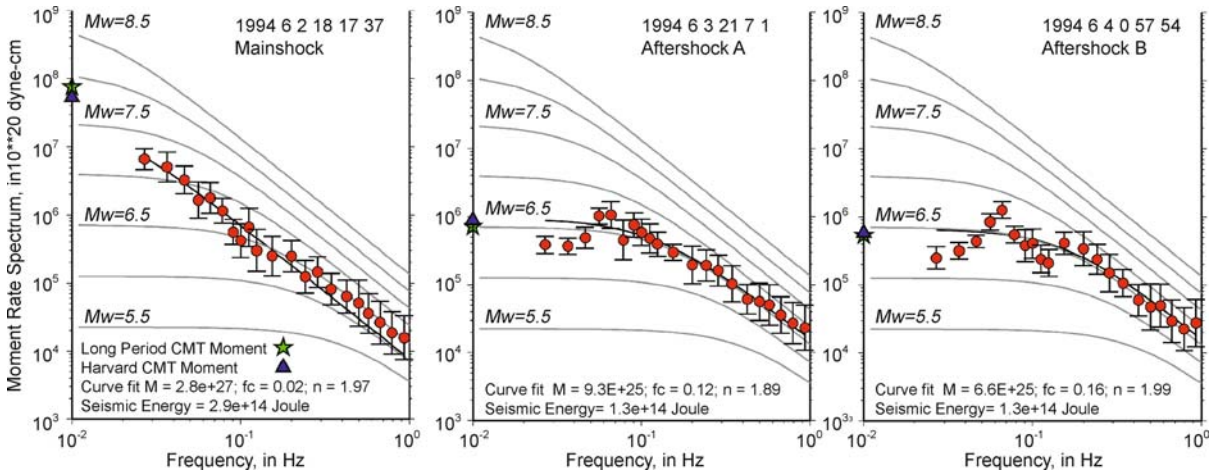

Tsunami Earthquakes, Figure 3

Moment rate spectra for the Java thrust mainshock (*left*) and two of its largest aftershocks, tensional events in the outer rise (*right panels*). The stars indicate the Harvard CMT moment, the *triangles* the moment determined using very long period surface waves. *Grey reference curves* were calculated for an ω^2 model [13] with a stress drop of 30 bars and an S-wave velocity of 3.75 km/s. The moment rate for the main shock is similar to that of its much smaller (in terms of M_w) aftershocks for periods shorter than 10 seconds. Adapted from [60]

Java 2006: [5]). Correspondingly, the centroid times and source durations or rise times determined for these events are also relatively large with respect to other large subduction zone earthquakes ([59]; for Kurile Islands 1975: [71]; for Peru 1960: [57]; for Java 2006: [5] and [25]), although they may not be anomalous relative to other, smaller, subduction zone earthquakes at very shallow depth [9]. The energy that is radiated by these slow rupture processes is also anomalously low, as is shown by analysies of the radiated energy to moment ratio (for several recent tsunami earthquakes: [51]; for the 1946 Aleutian Islands earthquake: [47]) and radiation efficiency [83].

Unfortunately, the slow character of tsunami earthquakes also means that local residents are not warned by strong ground shaking of the possibility of an impending tsunami. Field surveys and first-person accounts describe the motion of tsunami earthquakes more as a weak "rolling motion" than the usual impulsive character of local events. In the case of the Nicaragua earthquake, some felt a very feeble shock before the tsunami, but most did not feel the earthquake at all [30]. For the 1994 Java event earthquake-induced ground shaking was not noticed by the coastal residents interviewed in Bali and Java [74]. Interviews with local residents carried out for the 2006 Java earthquake [50] also indicate that they felt little or no shaking.

Most designs for tsunami earthquake discriminators and early warning systems make use of a number of the manifestations of the unusually slow character of tsunami earthquakes listed above and many incorporate the use of long period seismic waves for robust estimation of the size of the event. For example, the pulse width of the P-wave, used to calculate moment magnitude M_{wp}, can give an accurate estimate of source duration time. The combination of M_{wp} and the source duration can provide an effective tool to issue early tsunami warnings [81]. A slightly later arrival on a seismogram, the W-phase, is a distinct ramp-like long-period (up to 1000 s) phase that begins between P and S waves on displacement seismograms and is particularly pronounced for slow earthquakes; thus it can be used for identification of these types of events [38]. Another method for fast regional tsunami warning uses the ratio of the total seismic energy to the high-frequency energy (between 1 and 5 Hz), computed from the seismograms [70]. Similarly, the detection of deficient values of seismic energy-to-seismic moment ratio can be accomplished in automated, real-time mode [51].

Location: Close to Trench

The hypocenters of the recent tsunami earthquakes are located relatively close to the trench, as compared to regular subduction zone earthquakes. The Global CMT and other [59] centroid locations for several of these events are located even on the seawards side of the trench (also see Fig. 4). It may be possible that the inversion process mislocates the centroid of the event due to the unusually long duration of the seismic source for its moment

and thus its unusually late centroid time. Inversions using seismic and/or tsunami waveforms and other waveform investigations for the 1896 Sanriku event [76], the 1946 Aleutian Islands earthquake [36,78], the 1960 Peru earthquake [58], the 1963 Kurile earthquake [7,58,86], the 1975 Kurile earthquake [58,86], the 2006 Java earthquake [5,20], the Nicaragua earthquake [40,41,62] and the 1996 Peru event [31,64] indicate the presence of concentrated slip in a narrow region near the trench (see Fig. 4). Although in many of these inversions only 1-D Green's functions were used, preliminary research results using a more realistic velocity model for the shallow subduction zone [54] shows a similar picture for the 2006 Java tsunami earthquake.

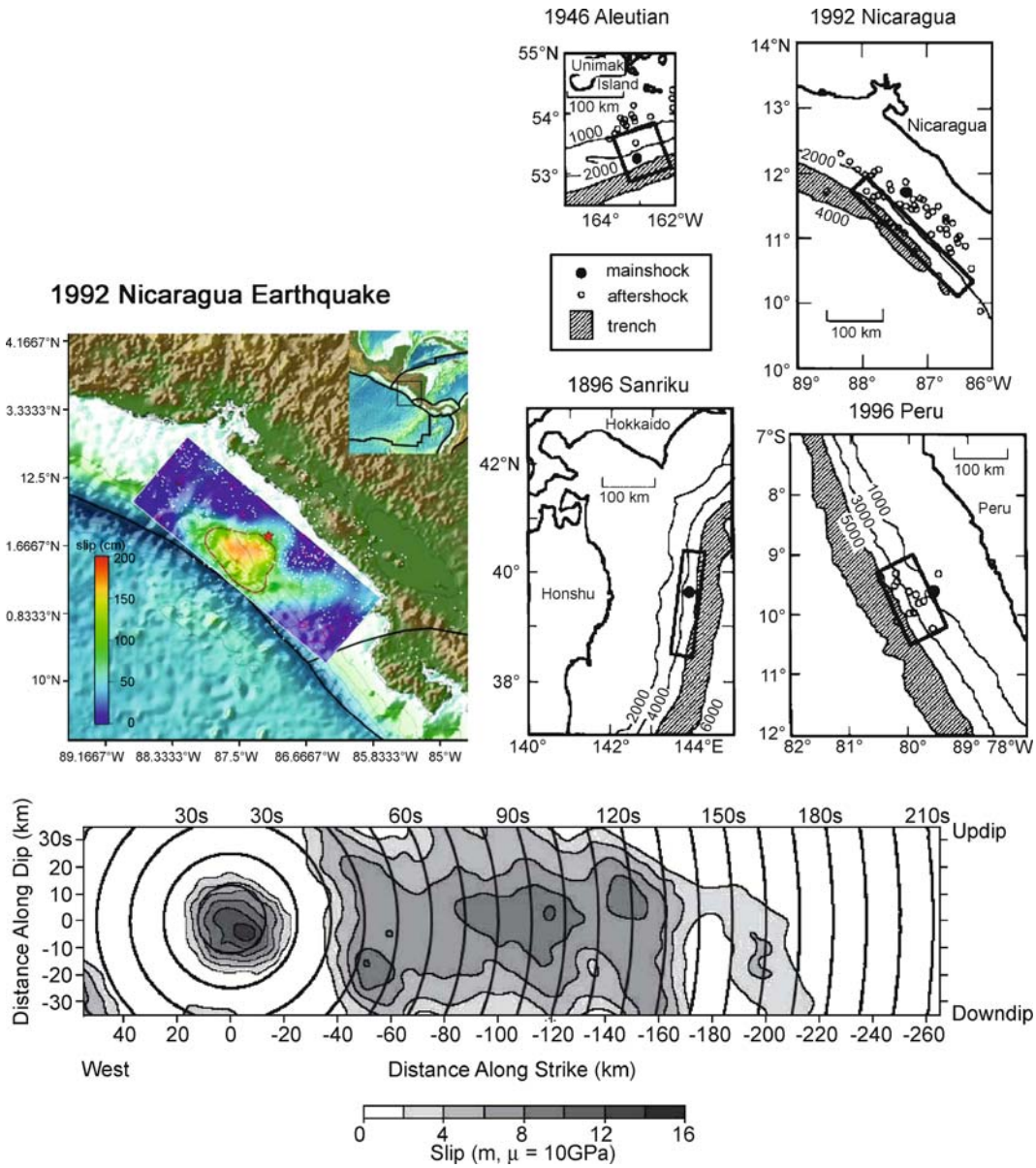

Tsunami Earthquakes, Figure 4

Slip or fault models determined for various tsunami earthquakes. The models shown are for: Nicaragua 1992 (*top left*, from [33]) Java 2006 (*bottom*, from [5]) and four tsunami earthquakes (*top right*, from [64]; the main shock and aftershock epicenters are shown and the *hatched area* indicates the trench). These models all show the presence of substantial slip close to the trench

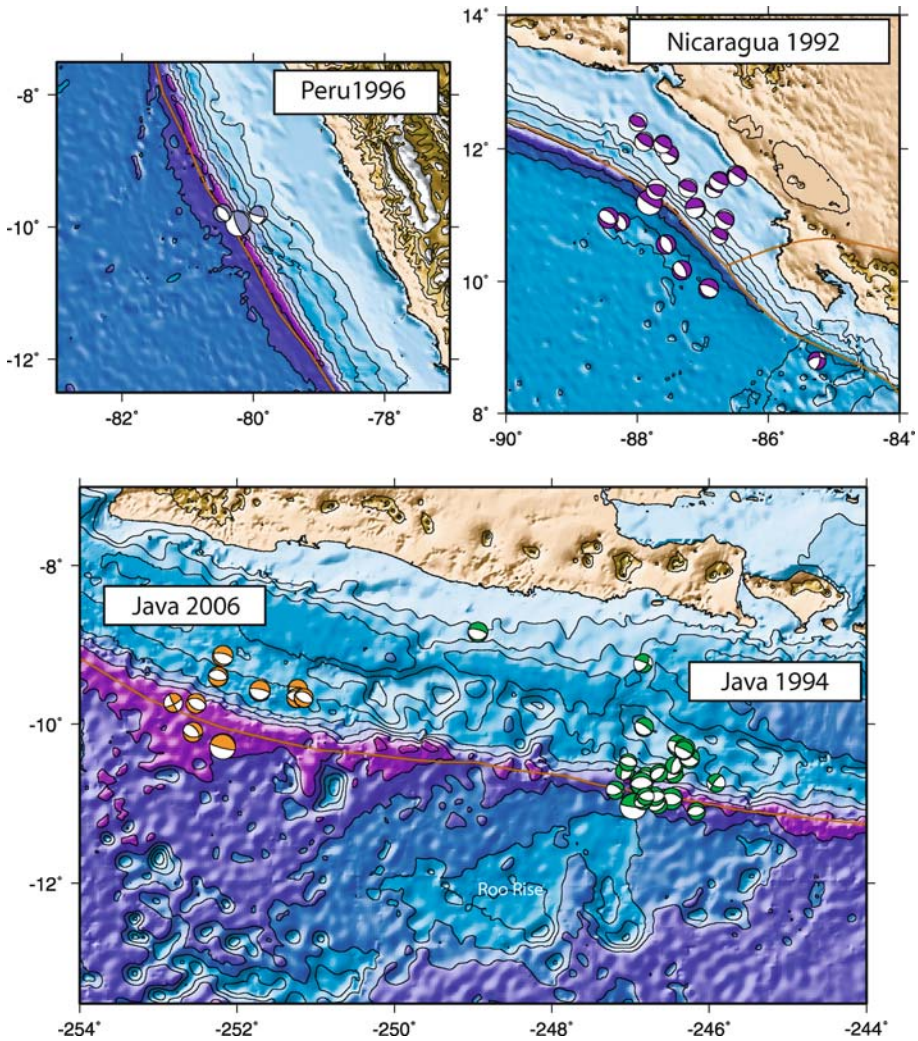

Tsunami Earthquakes, Figure 5

Aftershock sequences (seismicity in the region of the main shock within a period of 4 years after its occurrence) for several tsunami earthquakes, from the Global CMT catalog. Relatively few interplate aftershocks occur on the fault plane after a tsunami earthquake occurs, but a preponderance of normal-fault, probably intraplate, aftershocks is apparent

Aftershocks

The aftershock sequences of (recent) tsunami earthquakes are unusual in the preponderance of events not located on the interface between overriding and subducting plates [59]. Some of these aftershocks are located in the outer rise according to their Global CMT centroid locations and, in the case of the Java 2006 aftershocks, relocations using a 2.5-D model of the subduction zone [54] confirm this location. Others are located in the overriding plate ([8] for the 2006 Java earthquake), with some deeper within the accretionary prism (for the Java 2006 earthquake: [54]). The low number, or non-existence, of large (greater than magnitude 5.5) interplate earthquakes suggests that the main shock almost completely relieved the stress on the interface or may be related to the frictional properties of the fault. Several explanations have been proposed for the anomalously high number of intraplate earthquakes following tsunami earthquakes. Because of the proximity of the areas of high slip to the trench, the stress-change in the outer rise and trench area due to a tsunami earthquake are greater than for the "standard" subduction zone earthquake. Several modeling studies of stress alterations caused by large subduction earthquakes

suggest that the subduction slip will act to increase the tensional stress and favor normal events in zones towards the ocean from the upper limit of the rupture [16,80]. The subducting plate also seems to have a highly-broken up or rough character in many areas in which tsunami earthquakes have occurred ([49,75]; for Java: [48]; for Peru: [29] and [44]), which suggests the presence of more pervasive pre-existing weak zones, due to, for example, seafloor spreading related fabric. These weak zones may be re-activated in outer rise, or deeper intraplate, earthquakes following a tsunami earthquake.

Factors Involved in the Seismogenesis and Tsunamigenesis of Tsunami Earthquakes

Based on the consistent characteristics of tsunami earthquakes, as described in the previous section, and observations of their tectonic environments, hypotheses have been developed as to the cause of their extraordinary tsunami excitation and unusual seismic source process. In this section, we will document the factors that have been proposed to affect the seismo- and tsunami-genesis of tsunami earthquakes. Some are associated with the numerical prediction of tsunami wave-heights based on observed seismic waveforms, and others with possible unusual conditions of the tectonic environment in which these events occur; many of these factors are closely or at least somewhat related.

Slow Character May Lead to Underestimation of Earthquake Size

Prior to the installation of the broadband Global Seismic Network, the magnitudes of earthquakes were often determined from the amplitude of their teleseismic P-waves only. In the case of tsunami earthquakes, using this technique to determine their magnitude would lead to an initial underestimation of their true size, since their source spectra are depleted in the relatively high frequency energy that usually dominates the direct P-wave signals of regular earthquakes [59]. A similar issue would occur, although to a lesser degree, when using surface waves of periods of 20 seconds to determine the surface wave magnitude of these slow events [58]. With the advent of broadband sensors in the past several decades, it has now however become possible to investigate seismic waves to very long periods, hundreds or even thousands of seconds. Using more sophisticated techniques and the waveforms from technologically advanced sensors, the source spectrum of the recent tsunami earthquakes can now be modeled to these very long periods. Thus, no long period seismic energy that would excite tsunami waves should be "hidden" from the view of seismologists in the computation of moment magnitude or full rupture models using long period surface and body waves. However, for the earthquakes discussed in this paper, the observed tsunami are still larger than would be expected, even for their moment magnitude.

Effect of the Presence of Weak Materials with Low Shear Modulus

Most earthquake source inversions (either for full rupture or Centroid Moment Tensor parameters) implement a simple one-dimensional velocity and rigidity model to compute synthetic seismograms and model the recorded waveforms. If tsunami earthquakes are unusually shallow and/or involve sediments of low seismic wave speed, this is probably a poor approximation of the actual structure near the source region [23,64]. Some authors have attempted to rectify this error by using moment or slip distributions determined by seismic inversions in a structural model with significantly reduced shear modulus, more appropriate for the shallow trench region, and forward modeling the tsunami waves (e. g. [23]). This approach, however, is not satisfactory because a seismic inversion for moment or slip using this more appropriate rigidity model would produce a different distribution of moment or slip. Thus, a simple "correction" for the use of an inappropriate value of rigidity cannot be carried out after the rupture model has already been computed. If available, a correct rigidity model should be part of the seismic inversion process itself. To do this correctly, Green's functions should be computed for a three-dimensional (or possibly two dimensional, if the velocity structure is relatively uniform in the trench-parallel direction) velocity structure of the shallow subduction zone and incorporated in the modeling of the seismic waveforms. However, such sophisticated models are currently only available for very few subduction zones and the computational power required for these calculations (for body wave frequencies) is substantial. Unfortunately, whichever approach is chosen to go from recordings of seismic waveforms of tsunami earthquakes to the prediction of tsunami wave heights, a good model of the velocity and elastic properties of the shallow subduction zone is an unavoidable requirement.

Even when the moment distribution or moment magnitude of a tsunami earthquake has been determined using an appropriate model for the velocity and elastic parameters, there exists also the issue of enhanced tsunami excitation in material with weaker elastic properties, such as sedimentary layers. Modeling suggests that an event for which 10% of the moment is in sediments generates

a tsunami 10 times larger than its seismic moment would suggest [52] mainly because the slip in this material would be much greater than that for the same seismic moment in a stronger material (moment is the product of slip, area and rigidity after all). Therefore, the moment of a tsunami earthquake, even if determined correctly, may not directly reflect its tsunamigenic potential when low velocity sediments are present in the rupture zone. Since tsunami wave heights are mainly determined by the vertical displacement of the ocean floor, which in turn is primarily controlled by the slip on the fault plane, the slip (distribution) is more directly indicative of tsunamigenic potential. Since variations in shear modulus of a factor of five are not uncommon in shallow subduction zones [23], earthquakes with similar moments can result in substantially different slip models and tsunami excitation.

Shallow Depth of Slip Causes Relatively Great Displacement of Ocean Floor

Shallower earthquakes produce greater and shorter wavelength vertical displacement of the ocean floor, and thus greater and shorter wavelength tsunami waves right above the source region. However, higher frequency waves travel more slowly than longer period waves and, after a few hundred or thousand kilometers of travel, they drift to the back of the wave train and do not contribute to the maximum amplitude. Beyond about 2000 km distance, any earthquake at a depth less than 30 km appears to be equally efficient in tsunamigenesis ([85] and Fig. 6). Therefore, the exact depth of the slip in a shallow earthquake is not a key factor in determining its teleseismic tsunami wave-heights. Although the teleseismic tsunami wave-heights for a shallow slip event may not be significantly greater in amplitude than those for a somewhat deeper slip event [85], at local and regional distances the depth of the slip is an important factor. Therefore, for tsunami earthquakes, which have anomalously great tsunami height at mainly local and regional distances close to the rupture, the depth of the rupture should be a significant factor in their tsunamigenesis (see Fig. 6). Furthermore, modeling of shallow subduction zone earthquakes using a specific crack model [24] indicate that a rupture intersecting the free surface results in approximately twice the average slip. However, under the assumption of other specific frictional and crack models, the modeling of subduction zone earthquakes by Wang and He [84] produces slip models that have less vertical displacement of the ocean floor when the slip reaches the surface, due to a different slip distribution and the curvature of the subduction interface.

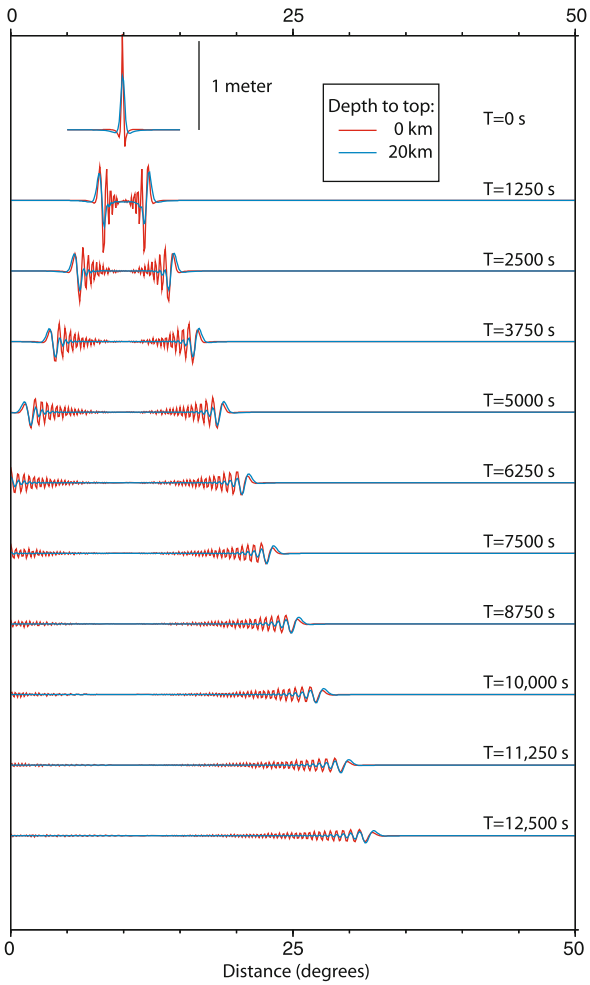

Tsunami Earthquakes, Figure 6
Cross sections of an expanding tsunami from a M7.5 thrust earthquake. The fault strikes north south (into the page) and the sections are taken east west. Elapsed time in seconds is given at the *left* and *right sides*. *Red lines* are for a fault that breaks the surface and *blue lines* for a fault with its top at a depth of 20 km. Deeper earthquakes make smaller and longer wavelength tsunamis at relatively short distances

Shallow Fault Dip May Lead to Underestimation of Slip from Seismic Waves

It is notoriously difficult to resolve moment, M_0, and dip, δ, independently for shallow thrust earthquakes. The excitation functions of Rayleigh waves for shallow thrust earthquakes show that only the product of dip and moment (more precisely $M_0 \sin(2\delta)$) can be resolved [39] when using these surface waves in inversions for a source mechanism. If Love wave data are also included in the inversion, it may be possible to add constraints by concentrating on fitting the amplitudes of those Love waves

recorded at azimuths corresponding to the along strike direction [34], however, this approach may be complicated by directivity effects. The polarity of body waves can only be used to constrain the focal mechanism of an earthquake at a limited range of incidence angles and thus also cannot provide any additional constraints on the dip of the shallowly dipping plane, unless assumptions are made about the rake angle. It is thus possible to (severely) underestimate the amount of slip in the earthquake, if the dip of the mechanism is poorly constrained and the inversion leads to a value for dip that is too high. This could be particularly important for very shallow subduction earthquakes, since the dip is expected to be small for these events. Thus, a difference between a dip of 3 or 6 degrees in a CMT solution may not seem significant, but it could lead to a difference in moment (and thus slip) of a factor of two. To illustrate the importance of this issue for very shallowly dipping thrust events, we show in Fig. 7 three different homogeneous slip models (only slip was changed, fault surface area and rigidity were kept constant) that will produce similar surface waves because the product of their slip and dip is identical. However, the vertical deformation of the ocean floor and the tsunami waveforms resulting from these three different earthquakes would be significantly different in amplitude.

Tsunami Earthquakes, Figure 7
Tsunami wave heights as a function of time for three different slip models for a shallow thrust earthquake, which will produce similar surface wave recordings (because the product of dip and moment is held constant). Thus, it is difficult to resolve between these different models using an inversion of surface waves, yet they produce very different vertical displacement of the ocean floor and thus very different tsunamis

Horizontal Deformation of the Ocean Floor May Lead to Great Displacement of Water, Yet Is Neglected in Tsunami Modeling

Most tsunami modelers only consider a water surface displacement identical to the vertical deformation of the ocean bottom due to faulting when computing the tsunami height resulting from an earthquake and neglect the effect of horizontal deformation. However, when the tsunami source is located close to a steep slope and the horizontal displacement is large relative to the vertical displacement, which is generally the case for tsunami earthquakes due to their mechanism and shallow depth, the effect of horizontal deformation may become significant [77].

Furthermore, it has been suggested that the lateral collision force of a continental slope into the ocean due to faulting could also play an important part in the tsunamigenesis of these events [72]. This type of dynamic excitation of tsunami waves would be particularly important for very shallowly dipping, shallow, thrust events, which would have a large component of horizontal motion. However, this concept has still been debated. This problem will be resolved if tsunami excitation is computed using the three-dimensional bathymetry in the source region with the displacement of the sea-floor as the initial condition.

Subduction of Bathymetric Features May Enable Seismic Slip in Usually Aseismic Region and Be Related to Unusual Aftershock Sequences

Sandbox experiments show the pervasive influence on the geomorphology of the shallow subduction zone margin when a seamount on the subducting plate is being subducted [17]. Subduction and underplating of relatively undeformed and water-laden sediments beneath the rear

part of the margin could, together with the dense fracture network generated by seamount subduction, modify the fluid pressure and introduce significant variations of the effective basal friction and thus the local mechanical plate coupling. More directly, subduction of a seamount may increase the normal stress across the subduction interface and hence enhance seismic coupling [15,67]. Unusual earthquakes have been documented in regions where ridges, seamounts or other bathymetric features are being subducted (e. g. [14,43]) and investigations of rupture characteristics of large underthrusting earthquakes provide evidence that seamounts can be subducted to seismogenic depths and that variations in seafloor bathymetry of the subducting plate may strongly influence the earthquake rupture process [10,61]. In the case of the Java tsunami earthquake of 1994, the bathymetry of the area landwards of the trench suggests that a local high is in the process of being subducted close to the area of maximum slip ([3], Fig. 4). A bathymetric high in the form of the Roo Rise can also be found just seaward of the trench region. In the bathymetry of the area around the 2006 Java event no such pronounced local feature can be found (Fig. 4), but the regional bathymetry south of the Java trench region is distinguished by an overall rough character. Similarly, the Nicaragua subduction zone is characterized by a highly developed horst and graben structure in the subducting plate, but no large-scale features, like a subducting seamount, are obvious. However, in case of the Peru earthquakes, both the 1996 and 1960 tsunami earthquakes occur at the intersection of the trench with major topographic features on the Nazca plate: the Mendaña fracture zone and the Trujillo trough, respectively [53].

Thus, subduction of either pronounced local bathymetric features or more regional seafloor roughness or horst-and-graben structures, which may modify the local coupling between subducting and overriding plates, has been documented in or near the rupture zone of many tsunami earthquakes.

Accretionary Prism: Uplift, Slides or Splay Faulting May Displace a Relatively Great Volume of Water

Tsunami earthquakes may involve seismic slip along the normally aseismic basal decollement of the accretionary prism [58,76]. Sediments near the toe of an inner trench slope may be scraped off by a large horizontal movement over the decollement due to an earthquake and thus cause an additional inelastic uplift, which could have a large effect on tsunami generation (for the 1896 Sanriku earthquake: [79]; for the 1946 Aleutian Islands earthquake: [78]).

The existence of splay faulting, which would be more effective in exciting tsunamis due to their steeper dip, within the accretionary prism itself has been suggested to be a cause of the large tsunami excitation for the 1994 Java earthquake [3], the 2004 Sumatra mega-thrust event [46] and the 1963/1975 Kurile earthquakes [21]. Splay faulting can further promote extensive vertical deformation of the ocean floor, and hence large tsunamis, through partitioning or branching of a rupture upwards from the interface along multiple splay faults leading up to the surface [21,56].

Presence of Fluids Influences Seismic Behavior

In subduction zones, fluids expelled from the subducting plate play an important role in many different subduction related phenomena such as volcanism, metamorphism and seismogenesis. Zones of high pore fluid pressure in the shallow subduction zone would change the effective normal stress significantly, possibly extending the region in which seismic slip is possible to shallower depths and generate slip of a slow nature. The presence of such zones has been suggested to be related to the occurrence of silent slip in the Nankai trough region [42]. In the case of tsunami earthquakes, the presence of large zones of elevated fluid pressure has been proposed to cause fairly rapid seismic slip close to the trench axis, following the breakage of asperities [68].

A Model for Tsunami Earthquakes

Most explanations for the slowness of tsunami earthquakes involve the presence of low velocity, low strength and low rigidity sediments in the accretionary prism and between the overriding and subducting plate in the shallow subduction zone. Rupture through these slow sediments is thought to promote a slow rupture velocity. In Scholz's [65,66] model of the typical subduction zone (see Fig. 8 for an interpretation of this model from [9]), three possible stability regimes exist. In the stable zone, seismic slip cannot be supported and aseismic creep releases all strain. In the unstable zone episodic seismic slip occurs. In the conditionally stable zone, in between these two zones, slip can be abrupt if it experiences loading from a nearby slip patch. This conditionally stable zone may be very heterogeneous due to roughness of the thrust fault [75] creating isolated asperities or due to permeability changes [55] from the subduction of low permeability materials or the presence of fluids [68]. Tsunami earthquakes may represent a rupture of one or several large "unstable" asperities, which then propagates for a significant distance in the conditionally stable sedimentary materials [9,59].

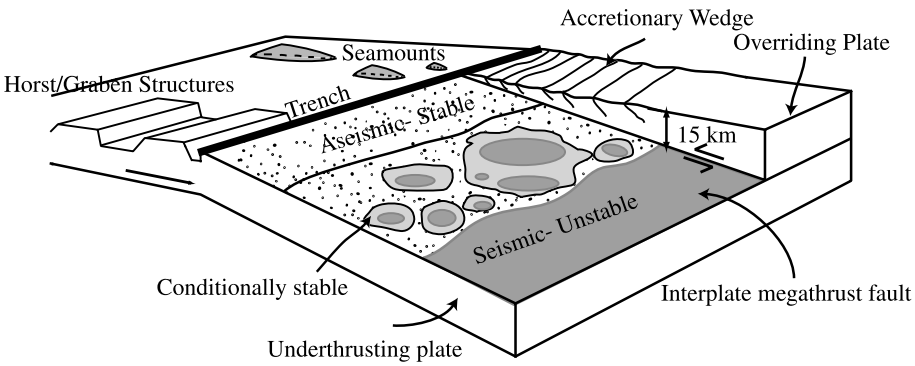

Tsunami Earthquakes, Figure 8

Cartoon illustrating frictional conditions of the subduction interface between subducting and overriding plate. Individual unstable sliding contact areas (*dark gray*) can provide the nucleation sites for rupture in the shallow subduction zone environment, which is typically a stable (*stippled*) or conditionally stable (*light gray*) frictional region. From [9]

Alternatively, but consistently, we could interpret tsunami earthquakes in the context of fracture mechanics as displaying a lack of radiated energy and a low rupture speed in a high G_C (fracture energy) environment [5,83]. As stated above, if these ruptures also involve localized patches of relatively strong unstable friction (that would be associated with high rupture speed and low G_C) it would allow the rupture to propagate seismically, instead of as a continuous creep process. From this point of view, tsunami earthquakes dissipate a large amount of energy during the fracture process and are left with little energy to radiate. It is possible that the highly faulted trench and deformed sediments result in larger energy dissipation during failure due to an excessive amount of branching and bifurcation of cracks which gives rise to inelastic behavior and hence a large dissipation of energy [6,83], possibly involving the branching of the rupture into multiple splay faults in the accretionary prism.

Tsunami earthquakes therefore would represent slip at unusually shallow depths that would typically be dominated by creep-processes. Their nucleation would be made possible through the existence of localized asperities or patches of unstable friction in a typically stable or conditionally stable region. The presence of compartments of elevated fluid pressure may aid in the propagation of the seismic slip by creating zones of nearly zero friction surrounding the asperities [68]. These asperities may be created by the subduction of bathymetric features like seamounts or ridges or by the broken up nature of the subducting plate itself, creating a horst-and-graben system, which would act as buckets for sediment subduction [59,75]. The stress-release on these asperities would be near complete, and any additional unloading of stress on the plate interface due to the rupture may occur mostly through creep. This would result in a relatively low number of aftershocks occurring on the interface between overriding and subducting plate. However, the static stress change in the outer rise area would be significant due to the shallow nature of most of the slip and thus normal faulting outer rise earthquakes would be more likely to be triggered [16,29,80]. The subduction of a bathymetric feature would also likely result in fracturization of the margin [17] in the overriding plate, thus further promoting the occurrence of intraplate aftershocks in this area of the shallow subduction zone. If the subducting plate itself is highly broken-up and thus contains pre-existing weak zones, this may facilitate further faulting within the subducting plate, in particular close to the lower edge of the rupture where the stress change due to the main shock is relatively large.

In this model for tsunami earthquakes discussed above, the unusually high effectiveness in the excitation of tsunami waves can be attributed to several factors, with the shallowness of the slip as the main underlying cause. Other important concerns coming into play are the possible failure of seismological techniques to provide an accurate estimate of the slip due to complexities associated with a very shallowly propagating rupture in a subduction zone, the possible failure of tsunami modeling to determine accurate wave heights due to similar complexities and the possible involvement of splay faulting or uplift of sediments near the trench in the accretionary prism.

Future Directions

The Sumatra–Andaman earthquake and tsunami of 2004 renewed interest in the development of near real-time methods to estimate the true size of large earthquakes and

the tsunamis that might follow them, and the installation of instrumentation that will facilitate these measurements. Because the time between the earthquake and the arrival of the first tsunami waves at the local coastline is short, it is still unclear how effective these types of early warning systems are for saving lives at short distances from the tsunami source, but they will be useful at large distances.

New technologies and surveys will enhance our knowledge of the geomorphology and velocity structure of the shallow subduction zone. Ocean bottom seismometers, tide gauges, buoys and other seafloor monitoring devices will provide high quality data, which will enable us to place better constraints on where exactly the slip in shallow earthquakes occurs and in what tectonic environment. Although tsunami earthquakes occur relatively infrequently and thus may be difficult to capture, comprehensive characterizations of their rupture processes placed in the context of detailed three-dimensional models of the shallow subduction zones they occurred in will be an important next step in understanding their unusual seismo- and tsunami-genic processes.

Bibliography

Primary Literature

1. Abe K (1979) Size of great earthquakes of 1873–1974 inferred from tsunami data. J Geophys Res 84:1561–1568
2. Abe K (1989) Quantification of tsunamigenic earthquakes by the M_t scale. Tectonophysics 166:21–34
3. Abercrombie RE, Antolik M, Felzer K, Ekstrom G (2001) The 1994 Java tsunami earthquake- Slip over a subducting seamount. J Geophys Res 106:6595–6608
4. Ammon CJ, Ji C, Thio HK, Robinson D, Ni S, Hjorleifsdottir V, Kanamori H, Lay T, Das S, Helmberg D, Ichinose G, Polet J, Wald D (2005) Rupture Process of the 2004 Sumatra–Andaman Earthquake. Science 308:1133. doi:10.1126/science.1112260
5. Ammon CJ, Kanamori H, Lay T, Velasco AA (2006) The 17 July 2006 Java tsunami earthquake. Geophys Res Lett 33:24. doi:10.1029/2006GL028005
6. Barragan BE, Giaccio GM, Zerbino RL (2001) Fracture and failure of thermally damaged concrete under tensile loading. Mater Struct 34:312–319
7. Beck SL, Ruff LJ (1987) Rupture process of the great 1963 Kuril Islands earthquake sequence: asperity interaction and multiple event rupture. J Geophys Res 92:14123–14138
8. Bilek SL, Engdahl ER (2007) Rupture characterization and relocation aftershocks of for the 1994 and 2006 tsunami earthquakes in the Java subduction zone. Geophys Res Lett. 34:L20311. doi:10.1029/2007GL031357
9. Bilek SL, Lay T (2002) Tsunami earthquakes possibly widespread manifestations of frictional conditional stability. Geophys Res Lett 29:18-1. doi:10.1029/2002GL01521
10. Bilek SL, Schwartz SY, Deshon HR (2003) Control of seafloor roughness on earthquake rupture behavior. Geology 31:455–458. doi:10.1130/0091-7613(2003)031
11. Bourgeois J, Petroff C, Yeh H, Titov V, Synolakis CE, Benson B, Kuroiwa J, Lander J, Norabuena E (1999) Geologic Setting, Field Survey and Modeling of the Chimbote, Northern Peru, Tsunami of 21 February 1996. Pure Appl Geophys 154: 513–540
12. Brown DL (1964) Tsunami activity accompanying the alaskan earthquake of 27 March 1964. US Army Engr Dist, Alaska, 20 pp
13. Brune JN (1970) Tectonic stress and spectra of seismic shear waves from earthquakes. J Geophys Res 75:4997–5009
14. Chung WY, Kanamori H (1978) Subduction process of a fracture zone and aseismic ridges – the focal mechanism and source characteristics of the New Hebrides earthquake of 1969 January 19 and some related events. Geophys J Int 54(1):221–240. doi:10.1111/j.1365-246X.1978.tb06764.x
15. Cloos M (1992) Thrust-type subduction-zone earthquakes and seamount asperities; a physical model for seismic rupture. Geology 20:601–604
16. Dmowska R, Zheng G, Rice JR (1996) Seismicity and deformation at convergent margins due to heterogeneous coupling. J Geophys Res 101:3015–3029
17. Dominguez S, Malavieille J, Lallemand SE (2000) Deformation of accretionary wedges in response to seamount subduction: insight from sandbox experiments. Tectonics 19:182–196
18. Engdahl ER, Villaseñor A (2002) Global seismicity: 1900–1999. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology. Academic Press, Amsterdam, Part A, chapt 41, pp 665–690
19. Fryer GJ, Watts P, Pratson LF (2004) Source of the great tsunami of 1 April 1946: a landslide in the upper Aleutian forearc. Mar Geol 203:201–218
20. Fujii Y, Satake K (2006) Source of the July 2006 West Java tsunami estimated from tide gauge records. Geophys Res Lett 33:L24317.1–L24317.5. doi:10.1029/2006GL028049
21. Fukao Y (1979) Tsunami earthquakes and subduction processes near deep-sea trenches. J Geophys Res 84:2303–2314
22. Geist EL (2000) Origin of the 17 July 1998 Papua New Guinea tsunami: Earthquake or landslide? Seism Res Lett 71:344–351
23. Geist EL, Bilek SL (2001) Effect of depth-dependent shear modulus on tsunami generation along subduction zones. Geophys Res Lett 28:1315–1318
24. Geist EL, Dmowska R (1999) Local tsunamis and distributed slip at the source. Pure Appl Geophys 154:485-512
25. Hara T (2006) Determination of earthquake magnitudes using duration of high-frequency energy radiation and maximum displacement amplitudes: application to the July 17, 2006 Java earthquake and other tsunami earthquakes. Eos Trans AGU 87(52):Fall Meet Suppl, Abstract S21A-0132
26. Hatori T (1967) The generating area of the Sanriku earthquake of 1896 and its comparison with the tsunami of 1933. J Seismol Soc Jap Ser 2, 20:164–170
27. Heinrich P, Schindele F, Guibourg S, Ihmlé PF (1998) Modeling of the February 1996 Peruvian tsunami. Geophys Res Lett 25:2687–2690
28. Hidayat D, Barker JS, Satake K (1995) Modeling the seismic source and tsunami generation of the December 12, 1992 Flores island, Indonesia, earthquake. Pure Appl Geophys 144: 537–554
29. Hilde TWC (1983) Sediment subduction versus accretion around the Pacific. Tectonophys 99:381–397
30. Ide S, Imamura F, Yoshida Y, Abe K (1993) Source character-

istics of the Nicaraguan tsunami earthquake of September 2, 1992. Geophys Res Lett 20:863–866
31. Ihmlé PF, Gomez JM, Heinrich P, Guibourg S (1998) The 1996 Peru tsunamigenic earthquake: Broadband source process. Geophys Res Lett 25:2691–2694
32. Imamura F, Gica E, Takahashi T, Shuto N (1995) Numerical simulation of the 1992 Flores tsunami: Interpretation of tsunami phenomena in northeastern Flores Island and damage at Babi Island. Pure Appl Geophys 144:555–568
33. Ji C (2006) A comparison study of 2006 Java earthquake and other Tsunami earthquakes. Eos Trans AGU 87(52):Fall Meet Suppl, Abstract
34. Ji C (2006) Resolving the trade-off between the seismic moment and fault dip of large subduction earthquakes and its impact on tsunami excitation. Tsunami Sources Workshop. Menlo Park
35. Ji C. Zeng Y, Song AT (2007) Rupture process of the 2006 Mw 8.3 Kuril Island Earthquake inferred from joint inversion of teleseismic body and surface waves. SSA meeting. Kona
36. Johnson JM, Satake K (1997) Estimation of seismic moment and slip distribution of the April 1, 1946, Aleutian tsunami earthquake. J Geophys Res 102:11765–11774
37. Kanamori H (1972) Mechanism of tsunami earthquakes. Phys Earth Planet Inter 6:346–359
38. Kanamori H (1993) W phase. Geophys Res Lett 20:1691–1694
39. Kanamori H, Given JW (1981) Use of long-period surface waves for rapid determination of earthquake-source parameters. Phys Earth Planet Int 27:8–31
40. Kanamori H, Kikuchi M (1993) The 1992 Nicaragua earthquake – A slow tsunami earthquake associated with subducted sediments. Nature 361:714–716
41. Kikuchi M, Kanamori H (1995) Source characteristics of the 1992 Nicaragua tsunami earthquake inferred from teleseismic body waves. Pure Appl Geophys 144:441–453
42. Kodaira S, Iidaka T, Kato A, Park JO, Iwasaki T, Kaneda Y (2004) High pore fluid pressure may cause silent slip in the Nankai trough. Science 304:1295–1298. doi:10.1126/science.1096535
43. Kodaira S, Takahashi N, Nakanishi A, Miura S, Kaneda Y (2000) Subducted seamount imaged in the rupture zone of the 1946 Nankaido earthquake. Science 289:104–106. doi:10.1126/science.289.5476.104
44. Kulm LD, Prince RA, French W, Johnson S, Masias A (1981) Crustal structure and tectonics of the central Peru continental margin and trench. In: Kulm LD, Dymond J, Dasch EJ, Hussong DM (eds) Nazca Plate: Crustal formation and Andean Convergence. Geol Soc Am Mem 154:445–468
45. Lavigne F, Gomes C, Giffo M, Wassmer P, Hoebreck C, Mardiatno D, Prioyono J, Paris R (2007) Field observations of the 17 July 2006 Tsunami in Java. Nat Hazards Earth Syst Sci 7: 177–183
46. Lay T, Kanamori H, Ammon CJ, Nettles M, Ward SN, Aster RA, Beck SL, Bilek BL, Brudzinski MR, Butler R, DeShon HR, Ekström G, Satake K, Sipkin S (2005) The great Sumatra–Andaman earthquake of 26 December 2004. Science 308:1127–1133. doi:10.1126/science.1112250
47. Lopez AM, Okal EA (2006) A seismological reassessment of the source of the 1946 Aleutian 'tsunami' earthquake. Geophys J Int 165(3):835–849. doi:10.1111/j.1365-246X.2006.02899.x
48. Masson DG, Parson LM, Milsom J, Nichols G, Sikumbang N, Dwiyanto B, Kallagher H (1990) Subduction of seamounts at the Java trench – a view with long-range sidescan sonar. Tectonophys 185:51–65
49. McAdoo BG, Capone MK, Minder J (2004) Seafloor geomorphology of convergent margins: Implications for Cascadia seismic hazard. Tectonics 23:TC6008. doi:10.1029/2003TC001570
50. Mori J, Mooney WD, Afnimar Kurniawan S, Anaya AI, Widiyantoro S (2007) The 17 July 2006 tsunami earthquake in west Java, Indonesia. Seismol Res Lett 78:291
51. Newman AV, Okal EA (1998) Teleseismic estimates of radiated seismic energy: The E/M_0 discriminant for tsunami earthquakes. J Geophys Res 103:26885–26898
52. Okal EA (1988) Seismic parameters controlling far-field tsunami amplitudes: A review. Nat Haz 1:67–96
53. Okal EA, Newman AV (2001) Tsunami earthquakes: The quest for a regional signal. Phys Earth Planet Inter 124:45–70
54. Okamoto T, Takenaka H (2006) Source process of the July 17, 2006 off Java island earthquake by using a fine crustal structure model of the Java trench and a 2.5D FDM computations. Eos Trans AGU 87(52):Fall Meet Suppl, Abstract
55. Pacheco JF, Sykes LR, Scholz CH (1993) Nature of seismic coupling along simple plate boundaries of the subduction type. J Geophys Res 98:14133–14159
56. Park J-O, Tsuru T, Kodaira S, Cummins PR, Kaneda Y (2002) Splay Fault branching along the Nankai subduction zone. Science 297:1157–1160
57. Pelayo AM, Wiens DA (1990) The November 20, 1960 Peru tsunami earthquake: Source mechanism of a slow event. Geophys Res Lett 17:661–664
58. Pelayo AM, Wiens DA (1992) Tsunami earthquakes – Slow thrust-faulting events in the accretionary wedge. J Geophys Res 97:15321–15337
59. Polet J, Kanamori H (2000) Shallow subduction zone earthquakes and their tsunamigenic potential. Geophys J Int 142:684–702. doi:10.1046/j.1365-246x.2000.00205.x
60. Polet J, Thio HK (2003) The 1994 Java Tsunami earthquake and its "Normal" Aftershocks. Geophys Res Lett 30:27–1. doi:10.1029/2002GL016806
61. Robinson DP, Das S, Watts AB (2006) Earthquake rupture stalled by a subducting fracture zone. Science 312:1203–1205. doi:10.1126/science.1125771
62. Satake K (1994) Mechanics of the 1992 Nicaragua tsunami earthquake. Geophys Res Lett 21:2519–2522
63. Satake K, Kanamori H (1991) Abnormal tsunamis caused by the June 13, 1984, Torishima, Japan, earthquake. J Geophys Res 96:19933–19939
64. Satake K, Tanioka Y (1999) Sources of tsunami and tsunamigenic earthquakes in subduction zones. Pure Appl Geophys 154:467–483. doi:10.1007/s000240050240
65. Scholz CH (1990) The mechanics of earthquakes and faulting. Cambridge Univ Press, New York
66. Scholz CH (1998) Earthquakes and friction laws. Nature 391: 37–42
67. Scholz CH, Small C (1997) The effect of seamount subduction on seismic coupling. Geol 25:487–490
68. Seno T (2002) Tsunami earthquakes as transient phenomena. Geophys Res Lett 29(10):58.1–58.4. doi:10.1029/2002GL014868
69. Seno T, Hirata K (2007) Did the 2004 Sumatra–Andaman earthquake involve a component of tsunami earthquakes? Bull Seismol Soc Am 97:S296–S306. doi:10.1785/0120050615
70. Shapiro NM, Singh SK, Pacheco J (1998) A fast and simple diag-

nostic method for identifying tsunamigenic earthquakes. Geophys Res Lett 25:3911–3914
71. Shimazaki K, Geller RJ (1977) Source process of the Kurile Islands tsunami earthquake of June 10, 1975. Eos Trans Am Geophys Union 58:446
72. Song Y, Fu L, Zlotnicki V, Ji C, Hjorleifsdottir V, Shum C, Yi Y (2006) Horizontal motions of faulting dictate the 26 December 2004 tsunami genesis. Eos Trans AGU 87(52):Fall Meet Suppl, Abstract U53C-02
73. Synolakis CE, Bardet JP, Borrero JC, Davies HL, Okal EA, Silver EA, Sweet S, Tappin DR (2002) The slump origin of the 1998 Papua New Guinea Tsunami. Proc Royal Soc A Math Phys Eng Sci 458(2020):763–789. doi:10.1098/rspa.2001.0915
74. Synolakis CE, Imamura F, Tsuji Y, Matsutomi H, Tinti S, Cook B, Chandra YP, Usman M (1995) Damage, conditions of East Java tsunamis of 1994 analyzed. EOS 76:26
75. Tanioka Y, Ruff L, Satake K (1997) What controls the lateral variation of large earthquake occurrence along the Japan trench. Isl Arc 6:261–266
76. Tanioka Y, Satake K (1996) Fault parameters of the 1896 Sanriku tsunami earthquake estimated from tsunami numerical modeling. Geophys Res Lett 23:1549–1552
77. Tanioka Y, Satake K (1996) Tsunami generation by horizontal displacement of ocean bottom. Geophys Res Lett 23: 861–864
78. Tanioka Y, Seno T (2001) Detailed analysis of tsunami waveforms generated by the 1946 Aleutian tsunami earthquake. Nat Haz Earth Syst Sci 1:171–175
79. Tanioka Y, Seno T (2001) Sediment effect on tsunami generation of the 1896 Sanriku tsunami earthquake. Geophys Res Lett 28:3389–3392
80. Taylor MAJ, Zheng G, Rice JR, Stuart WD, Dmowska R (1996) Cyclic stressing and seismicity at strong coupled subduction zones. J Geophys Res 101:8363–8381
81. Tsuboi S (2000) Application of M_{wp} to tsunami earthquake. Geophys Res Lett 27:3105–3108
82. Tsuji Y, Imamura F, Matsutomi H, Synolakis CE, Nanang PT, Jumadi, Harada S, Han SS, Arai K, Cook B (1995) Field survey of the east Java earthquake and tsunami of June 3, 1994. Pure Appl Geophys 144(3–4):839–854
83. Venkataraman A, Kanamori H (2004) Observational constraints on the fracture energy of subduction zone earthquakes. J Geophys Res 109:B05302.1–05302.20. doi:10.1029/2003JB002549
84. Wang K, He J (2008) Effects of frictional behavior and geometry of subduction fault on coseismic seafloor deformation. Bull Seismol Soc Am 98(2):571–579
85. Ward SN (2002) Tsunamis. In: Meyers RA (ed) The Encyclopedia of Physical Science and Technology, vol 17. Academic Press, San Diego, pp 175–191
86. Wiens D (1989) Bathymetric effects on body waveforms from shallow subduction zone earthquakes and application to seismic processes in the Kurile Trench. J Geophys Res 94: 2955–2972

Books and Reviews

Bebout G, Kirby S, Scholl D, Platt J (eds) (1996) Subduction from Top to Bottom. American Geophysical Union Monograph, no 96. American Geophysical Union, Washington DC
Satake K, Imamura F (1995) Tsunamis 1992–1994. Special Issue of Pure Appl Geophys 144(3–4):373–890

Tsunami Forecasting and Warning

Osamu Kamigaichi
Japan Meteorological Agency, Tokyo, Japan

Article Outline

Glossary
Definition of the Subject
Introduction
Complexity Problem in Tsunami Forecasting
Components of a Tsunami Early Warning System (TEWS)
Tsunami Early Warning System in Japan
Future Outlook
Acknowledgments
Bibliography

Glossary

Tsunami early warning system It consists of four components, namely, 1) Seismic network, 2) Seismic data processing system, 3) Tsunami forecast system, and 4) Sea level data monitoring system. In a broader sense, warning transmission system (downlink to disaster management organizations and public) is also included.

Tsunami amplitude Amplitude is measured from undisturbed sea level to peak or trough of the wave. By definition, it can be positive or negative. Tsunami amplitude can be measured in real-time by instruments like tide gauge, pressure sensor, etc., and can be reproduced by numerical tsunami propagation simulation. One needs not to be confused with the term 'run-up height'. It is the maximum height of inundation on land, and measured in post-tsunami field surveys from traces of tsunami (i. e. damage of constructions, vegetative markers, etc.).

Simulation point In this article, it is defined as the surface projection location of the hypothetical earthquake fault center. The vertical component of ocean bottom deformation due to earthquake fault dislocation calculated by elastic theory gives the initial tsunami waveform for the numerical tsunami propagation simulation.

Forecast point In this article, it is defined as the location of the offshore point where tsunami amplitude is evaluated by using numerical tsunami propagation simulation.

Intergovernmental coordination group The group established under UNESCO/IOC to facilitate interna-

tional cooperation for the tsunami disaster mitigation. There are four ICGs as of now (Pacific Ocean, Indian Ocean, Caribbean Sea, North Eastern Atlantic Ocean and Mediterranean Sea). One of the most important characteristics of tsunami is that it can cause huge disaster even after long distance propagation due to amplification near the coast. Therefore, international cooperation, especially the prompt data and information exchange, is essential for the disaster mitigation.

Centroid moment tensor solution One of the representation of seismic source process. It is represented by the moment tensor, which is a combination of six independent equivalent force couples, and is a weighted average of the source process in time and space. Since it represents an overall image of the source process, it is suitable to evaluate tsunamigenic potential of the earthquake.

Earthquake early warning Earthquake Early Warning is to enable countermeasures in advance for strong motion disaster by detecting seismic *P* wave at stations near the epicenter, quickly estimate seismic intensity and arrival time of *S* wave, and transmit these estimation before the S wave arrival. The Japan Meteorological Agency (JMA) started to provide EEW to the general public in October, 2007. This technique is applicable to quicken tsunami warning dissemination.

Definition of the Subject

A tsunami is, along with strong motion, one of the two major disasters caused by earthquake. To mitigate tsunami disaster, it is important to integrate software countermeasures like tsunami forecast to enable timely evacuation from area at risk before tsunami strikes the coast, as well as to intensify hardware countermeasures particularly in vulnerable coastal areas like building banks and water-gates. Tsunami disaster mitigation can be achieved effectively by the appropriate combination of the software and hardware countermeasures. Also, improving people's awareness on the tsunami disaster, necessity of spontaneous evacuation when they notice an imminent threat of tsunami on their own (feeling strong shaking near the coast, seeing abnormal sea level change, etc.) and how to respond to a tsunami forecast, and conducting tsunami evacuation drill are very important issues for disaster mitigation.

In this article, a tsunami forecast, as the most typical software countermeasure that a national organization can provide, is mainly described. Recent progress in science has deepened our understanding of the tsunami-generating source mechanisms, tsunami propagation and inundation process and enabled the development of sophisticated numerical simulation programs. But at the same time, complexities of focal process and tsunami behavior, especially near the coast, have also been revealed. Therefore, careful consideration of the complicated nature of tsunami phenomena is essential in order to make tsunami forecast contents and the dissemination of warnings effective for disaster mitigation.

Introduction

A tsunami is an oceanic gravity wave generated by submarine fault dislocation or other origins such as mud or rock slumps on steep continental margin slopes, marine volcanic eruptions and others. A large tsunami may cause disasters along densely populated or built-up coasts and sometimes also by the inundation of low land areas, up to several km inland. Most tsunamis are generated by large earthquakes occurring in oceanic areas, and it is possible to estimate tsunami generation to a certain extent from seismic wave analysis. By taking advantage of the propagation velocity difference between the much faster seismic and the slower tsunami waves, it is possible to mitigate tsunami disasters by issuing tsunami forecast before the tsunami arrives at the coast, thus enabling evacuation and other countermeasures. For other origins of tsunami, it is still difficult to quantitatively forecast tsunami generation until it is observed actually by the sea level change sensors.

In earlier years, an empirical method had been used to estimate tsunami amplitude via a regression formula that relates tsunami amplitude to the magnitude of the earthquake and its epicentral distance from the coast of interest. Recently, significant progress has been made in the understanding of the tsunami characteristics. This permits numerical simulation of the tsunami propagation once the initial tsunami wave distribution in the source area is correctly known, and bathymetry data are available with sufficient spatial resolution. In Japan, numerical simulation technique has been used in the operational tsunami forecasting procedure in the Japan Meteorological Agency (JMA) since 1999. Efforts have been made also in other countries to incorporate numerical simulation technique in tsunami forecasting. In this article, mainly based on JMA experience, procedures are introduced how to conduct tsunami forecast service by using numerical simulation techniques and giving due consideration to the complexity of this problem.

Complexity Problem in Tsunami Forecasting

The complexity of tsunami forecasting with numerical methods is twofold and mainly due to the following:

- The uncertainty of the initial tsunami wave distribution in the source area
- The complexity of bathymetry and coastal topography

Uncertainty of the Initial Tsunami Wave Distribution

For a local tsunami, only limited time is left from the generation of tsunami to its arrival at the coast. In order to assure maximum lead time for evacuation, this necessitates the tsunami forecast to be based on the earliest available data from seismic wave analysis. On the other hand, in the case of a distant tsunami sources, a relatively long lead time is left until the tsunami strikes the coast, and data of sea level change can be recorded in near real-time by tidal stations on the way from the source to the coast. This allows the tsunami forecast to be based on the actual observation of generated tsunami waves and thus to reduce the uncertainty in the initial spatial distribution of the tsunami wave in the source area. Also for local events it is possible to improve the accuracy and reliability of the tsunami forecast in a step-by-step manner, if more detailed data and reliable analysis results become available after the first forecast has been issued. Therefore, the most practical approach in view of the disaster mitigation is to assure the rapid issuance of the first forecast based on the seismic wave analysis, and then to update it with improved data.

There are several uncertainties when the initial tsunami wave distribution is inferred from seismological data analysis alone.

Uncertainty of the Relative Location of the Hypocenter in the Rupture Area The hypocenter of an earthquake is merely the location from where the rupture starts. It is calculated from first arriving seismic waves. Depending on the direction of rupture propagation (e. g., unilateral, bilateral, radial) and the complexity/irregularity of the actual fault rupture, the seismologically determined hypocenter may neither lie in the center of the rupture area nor coincide with the area of maximum fault displacement. The bigger the earthquake and thus the larger the spatial extent of the rupture area, the less representative the seismologically derived hypocenter as a parameter for characterizing the earthquake rupture is. To reduce this uncertainty, some methods have been developed, like a rupture process inversion by using teleseismic broadband seismogram [20], or near-field strong motion data [11]. A fault model inversion by using a co-seismic step estimated by real-time GPS data analysis is under development [5,22,25]. And the spatial extent of the tsunami source area can also be constrained by using inverse refraction diagrams from tidal stations [1] or deep ocean water pressure sensor (DART (Deep-ocean Assessment and Reporting of Tsunamis))-type stations [6] that recorded tsunami arrivals as well as by satellite sea surface altimetry data taken over the source area [7,10].

In general, the rupture process is complicated, and the slip distribution on the fault is not uniform. Non-uniformity of the slip distribution can make the initial tsunami wave distribution much more complicated, especially in its short wavelength component, and may strongly affect the tsunami behavior on the coast. To estimate the slip distribution, in addition to the seismological and geodetic methods mentioned above, fault slip distribution inversion techniques have been developed which compare actually observed tsunami waveforms with pre-calculated theoretical tsunami waveforms from unit fault dislocation segments as Green functions [26] ► Tsunamis, Inverse Problem of.

Uncertainty of the Magnitude The maximum amplitude of a seismic body or surface wave group is usually used for estimating the earthquake magnitude (see this volume) as a measure of earthquake size in a certain period range, which is often limited by the bandwidth of the seismometer response. Therefore, the tsunamigenic potential may be underestimated, especially for gigantic earthquakes (such as the Sumatra-Andaman Mw9.3 earthquake of 26 December 2004) and "tsunami earthquakes" that generate bigger tsunami than expected from its shaking potential and magnitude values in the short-period range. To overcome this difficulty, some fast body-wave methods have been developed that consider besides single maximum amplitudes also multiple rupture amplitudes over the rupture duration [2], ► Earthquake Magnitude, integrated seismic energy and/or total rupture duration time [8,21], while others use successfully long-period spectral mantle surface-wave amplitudes in real-time applications in the context of tsunami warning [38], and *W*-phase for quick estimation of CMT solution for a practical tsunami warning service [17]. Besides these seismological methods for estimating earthquake magnitude, geodetic methods that use the co-seismic step or ultra-long period component observed by crustal deformation sensors [5,13,18] play an increasing role for estimating the size of great earthquakes.

Uncertainty of the Fault Parameters It is difficult to estimate fault geometry promptly after the earthquake occurrence. Sea floor deformation due to fault motion is usually assumed to be identical to initial tsunami wave distribution. Accordingly, the latter is significantly affected by the specifics of fault geometry, fault and slip orienta-

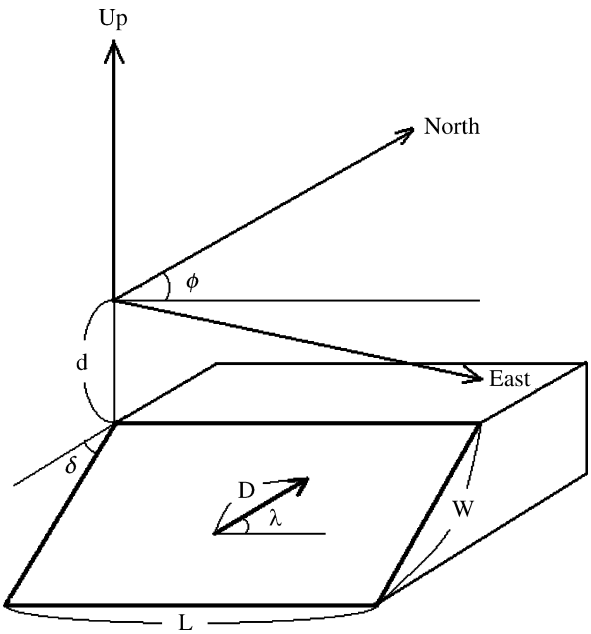

Tsunami Forecasting and Warning, Figure 1
Definitions of fault parameters. *L*, *W*, *D* and *d* denote the fault length, width, average slip amount and depth of the top margin of the fault respectively. The fault dip δ is measured down from horizontal, strike Φ clockwise round from North and slip (*rake*) λ anti-clockwise round from strike direction along the fault plane (motion of the hanging wall relative to the footwall)

tion in space as well as rupture complexity. These complexities are usually neglected and the real fault rupture is roughly approximated by a dislocation model that is represent by a rectangular fault that is described by 6 parameters, namely the fault length L, width W, average slip amount D, and the three angles of fault dip δ, strike Φ and slip (rake) λ (Fig. 1).

Fault Length, Width and Average Slip Amount Different scaling laws exist between L, W and D with earthquake magnitude (e. g., [3,4,16,37,39]). They allow us to roughly estimate these values from determined magnitudes. Such scaling relations are based on the assumption of more or less constant stress drop (i. e. High stress drop earthquake has larger D for the same L and W of low stress drop earthquake). But the stress drop is different between inter-plate thrust events and intra-plate events. In general, intra-plate event has higher stress drop than inter-plate event. Such differences may cause significant errors when estimating these fault parameters via scaling relations with magnitudes. Best estimates of these parameters, that characterize the earthquake size, are possible via the determination of the seismic moment $\mathrm{Mo} = \mu DA$ with μ being the rigidity of the crustal/lithosphere material in the rupture area $A = L \times W$, and L and W estimation by rupture process analysis described in Sect. "Uncertainty of the Relative Location of the Hypocenter in the Rupture Area".

Fault Dip, Strike and Slip Angle Representative values for dip, strike and slip angles can be set based on the analysis of past great earthquakes having occurred in each region. The fault plane of a newly occurred earthquake is likely to be close to either one of the two nodal planes derived from centroid moment tensor (CMT) solutions. But to select from these two candidates the real acting fault plane, precise aftershock location or rupture process analysis as described in Subsect. "Uncertainty of the Relative Location of the Hypocenter in the Rupture Area" are necessary.

Complexity due to Complicated Bathymetry

If the spatial distribution of the initial tsunami wave is given correctly, the tsunami propagation can be forecast in a deterministic manner by using numerical simulation technique. In a long-wave approximation, which considers only wavelengths of the tsunami that are much larger than the sea depth, the propagation velocity v of the tsunami is represented by $v = \sqrt{gd}$, where $g = 9.81\ldots$ m/s^2 is the Earth's gravity acceleration and d the sea depth in m. This approximation is valid for most tsunami cases. As wavelength is the product of velocity times period, tsunami wavelength is proportional to the square root of sea depth. Accordingly, the wavelength of the tsunami become shorter as it approaches the coast because the sea depth becomes shallower, and the wave becomes much more sensitive to finer bathymetry changes and coastal feature. Therefore, a finer mesh of bathymetry and coastal features is necessary near the coast in order to represent the tsunami behavior correctly. But, even if very fine mesh bathymetry data are available, it is not appropriate to use it in an early stage of tsunami propagation simulation when the initial tsunami wave distribution is still uncertain as mentioned in Subsect. "Uncertainty of the Initial Tsunami Wave Distribution".

Further, very fine-mesh simulations require a substantially long time for the conduct of simulation. Time constraints are crucial when incorporating numerical simulation techniques in an operational tsunami forecast procedure, except for distant events. Therefore, especially for local events, it is most practical to conduct tsunami simulations for a variety of scenarios in advance, to store these results in a database, and to conduct tsunami forecast by retrieving the most appropriate case for the determined hypocenter. As described later, the JMA has adopted this

way. Even in that case, the mesh size to be used in the simulations must be carefully examined taking into account the required accuracy and spatial resolution in the tsunami forecast for guiding disaster mitigation efforts and the need to complete the simulations for all scenarios in a realistic time span.

When very fine mesh is used, depending on the complexity of the bathymetry and topography near the coast, tsunami amplitude distribution along the coast shows significant scatter, and sometimes, extremely large tsunami amplitude will be estimated very locally. Therefore, it must be clearly defined in advance as to what kind of statistical average value, using a finite number of estimated tsunami amplitudes for the area of interest, should be adopted in the tsunami forecast.

Components of a Tsunami Early Warning System (TEWS)

In general, a tsunami early warning system consists of the following constituents:

1. Seismic network (seismometers and real-time data transmission link)
2. Real-time seismic data processing system for hypocenter and magnitude determination
3. Tsunami forecast system (including warning criteria, assembling of the text, and dissemination)
4. Sea level data monitoring system (tide gauge/tsunamimeter and real-time/near real-time data transmission link)

Such a composition is adopted in Japan, USA, Russia, Chile, Australia, French Polynesia, New Caledonia and in other countries around the Pacific Ocean. The up-to-date status of the TEWS of these countries can be consulted on the UNESCO/Intergovernmental Oceanographic Commission's (IOC) website as based on the national reports submitted to the latest meeting of the Intergovernmental Coordination Group/Pacific Tsunami Warning System (ICG/PTWS). Also in Indian Ocean countries, after the tremendous tsunami disaster brought by the Great Sumatra Earthquake in 2004, efforts have been made to establish tsunami early warning systems of the similar composition in different countries (e. g. GITEWS: German Indonesian Tsunami Early Warning System, see http://www.gitews.org/). Similar efforts are under way in the Caribbean, North Eastern Atlantic and Mediterranean regions too, inspired by the Sumatra event and taking into account potential tsunami risk assessments for these areas based on historical records.

To accomplish tsunami warning for local events, local seismic networks (see Subsect. "Seismic Network") and real-time seismic data processing systems (see Subsect. "Real-Time Seismic Data Processing System") are indispensable. For a distant event, one can utilize the hypocenter parameters contained in the international tsunami watch information provided by the Pacific Tsunami Warning Center (PTWC), West Coast/Alaska Tsunami Warning Center (WC/ATWC) of US and North West Pacific Tsunami Advisory Center (NWPTAC) of the JMA.

As a typical example, technical details of Japan's Tsunami Early Warning System, in which all four constituents are fully deployed, is explained in the following section.

As for the item 4 (sea level data monitoring system), the US and also GITEWS are now deploying DART type of buoys [6] system. This is to measure a sea surface vertical displacement by observing the related pressure change at the ocean bottom. DART buoys are placed in far offshore regions where bathymetry is simple, and very simple tsunami waveforms can be observed without the influence of complicated bathymetry near a coast. Such measured data are preferable for a comparison with simulated waveforms. And by placing such buoys with proper spacing at a certain distance from the tsunamigenic zone over the deep ocean basin, tsunami waves, that are sufficiently separated in time from seismic waves, can be observed early enough after the earthquake occurrence to be useful for tsunami EW from distant sources.

The Pacific Marine Environmental Laboratory (PMEL) of the National Oceanic and Atmospheric Administration (NOAA) is developing a tsunami forecasting system named SIFT (Short-term Inundation Forecasting) [33] based on the DART buoy data. Like in the JMA's tsunami simulation database, described in detail below, tsunami simulations are conducted for many different unit faults located along the tsunamigenic zone, and the calculated tsunami waveforms at DART buoy locations are stored in a database as Green functions, together with simulated waveforms at coastal points. In the case that a tsunami wave is observed by DART buoys, the observed waveforms are represented by a linear combination of several Green functions. Then, tsunami amplitudes at coastal points can be estimated from the linear combination coefficients and the simulated waveforms at coastal points (see web-page of NOAA/PMEL (http://nctr.pmel.noaa.gov/) for applications of their method to recent actual tsunami events). This system is planned to be introduced in the actual tsunami forecasting procedures of the US tsunami warning centers.

In Japan, a similar study is in progress at Tohoku-University [35]. They conduct tsunami simulations originat-

ing from unit vertical sea surface displacements in a unit sea area, and the resulting simulated waveforms at offshore tsunami meters are stored as Green functions in a database. They do not adhere to a specific fault model. The procedure for the estimation of tsunami waveform at coastal points is similar to that of SIFT.

Tsunami Early Warning System in Japan

Before describing the present status of the JMA's TEWS, a historical review is briefly given.

The JMA started tsunami warning service in 1952. Since then, the JMA has been making an effort to integrate its TEWS, and this is exactly a history of 'fight against time'.

At the commencement of the tsunami warning service, the JMA had 46 seismometers at meteorological observatories (basically near populated area on a sedimental layer). In case of earthquake occurrence, seismograms were read at observatories, and their results (*P* and *S* arrival times and maximum amplitudes) were transmitted to the headquarters in a telegram format. At the headquar-

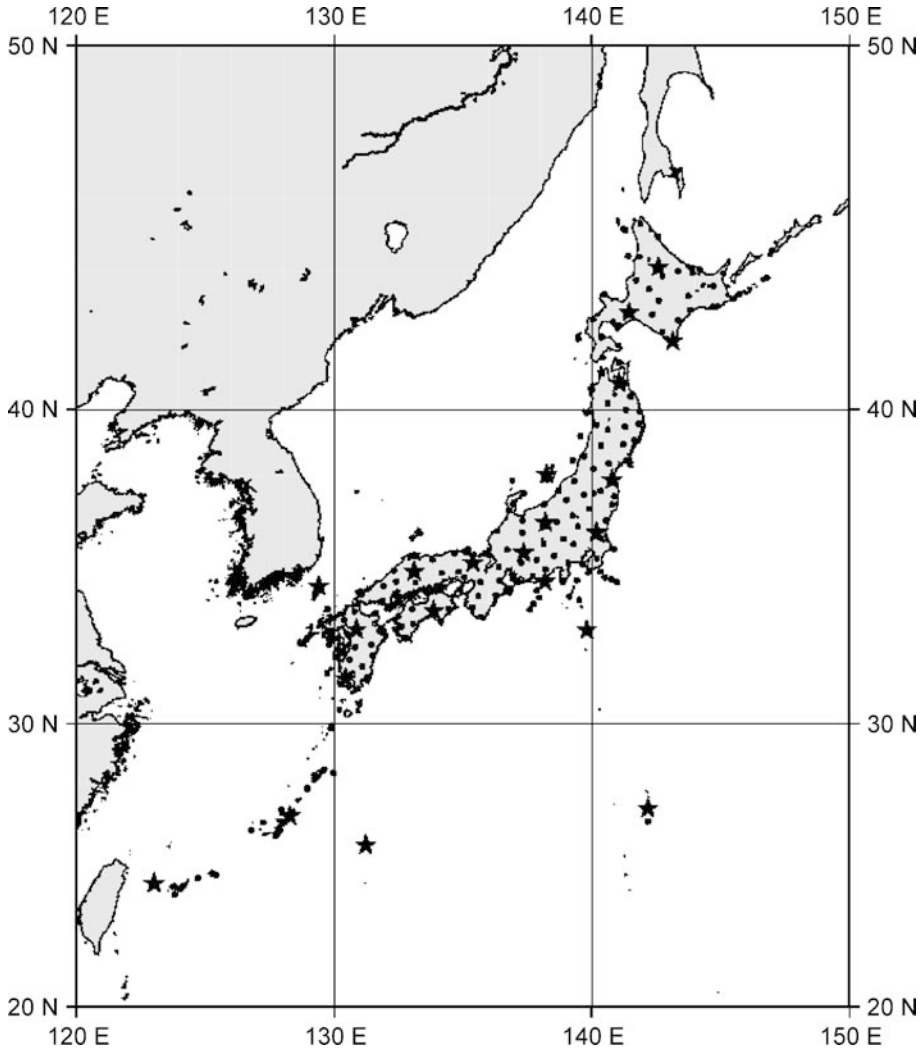

Tsunami Forecasting and Warning, Figure 2
Seismic network of the JMA used for tsunami forecasting. *Solid circles* and *stars* denote the locations of seismic stations operated by the JMA. The average spacing between the about 180 stations is 50 to 60 km. *Stars* denote the location of stations where STS-2 velocity broadband seismometers have been installed in addition to the Japanese short-period velocity-type seismometers and accelerometers

ters, hypocenter and magnitude were determined manually, and the tsunami warning grade was determined by using an empirical chart based on the relation between tsunami amplitude, earthquake magnitude and epicentral distance to a coast of the past events. It took about 15 to 20 minutes to disseminate a tsunami warning.

From the late 1960s to early 1980s, telemetry technology to transmit seismic waveform data from observatories to the headquarters and a processing computer were introduced. Detection capability of the earthquake was improved by monitoring collected seismic waveforms at one place, and the *P* and *S* picking precision was also improved by introducing a digitizer. By these, dissemination time was reduced to 12 to 13 minutes.

In 1983, the Mid-Japan-Sea earthquake (Mjma7.7) occurred. The JMA disseminated a tsunami warning in 14 minutes, but in about 7 minutes, the tsunami struck the nearest coast, killing 104 people. After this event, the JMA deployed a more sophisticated computer system for the seismic waveform processing. A graphical man–machine interface was introduced for more accurate and quicker phase reading and hypocenter and magnitude calculation. Still, the empirical method was used for the tsunami amplitude estimation. Dissemination time was reduced to between 7 and 8 minutes.

In 1993, the Southwest off Hokkaido earthquake (Mjma7.8) occurred. The JMA disseminated a tsunami warning in 5 minutes, but the tsunami struck Okushiri-Island in a few minutes, and 230 people were killed or lost by tsunami. After this event, the JMA totally replaced the seismic network. All seismometers were installed in remote un-manned sites on hard rock, and the total number of the site increased to about 180. Dissemination time was reduced to 3 to 5 minutes.

The present status of the JMA's TEWS is as follows.

Seismic Network

The JMA operates about 180 seismometers installed in Japan (Fig. 2). The seismic waveform data are sent to JMA continuously on a real-time basis through dedicated telephone lines.

There are two kinds of sensors deployed at each station: A short-period velocity sensor and an accelerometer. The records of the short-period sensor are mainly used for precise picking of the onset times of *P* and *S* phases which are required for an accurate hypocenter determination. Accelerometer records are used for the calculation of magnitude of large earthquakes in the case that the ground motion amplitude exceeds the dynamic range of the short-period sensor.

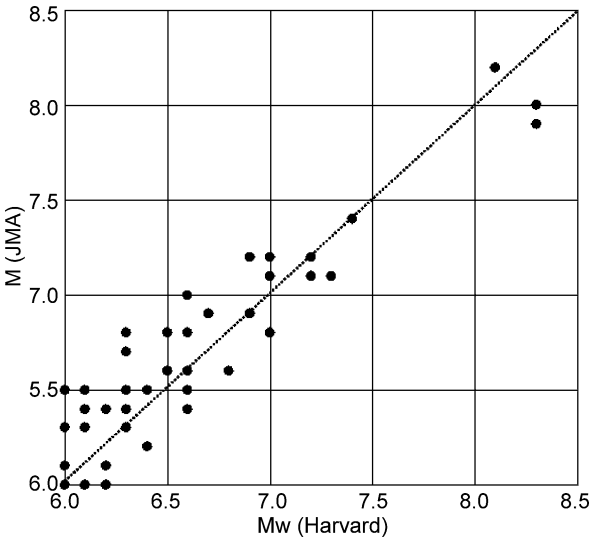

Tsunami Forecasting and Warning, Figure 3
Relation between Mw (Global CMT) and Mjma, based on events that occurred after September 2003 in and around Japan. Both magnitudes agree on average well up to around magnitude 8.0

STS-2 broad-band seismometers have additionally been installed at 20 stations. These broadband velocity seismic records are mainly used for obtaining CMT solutions and moment magnitudes Mw.

Real-Time Seismic Data Processing System

The JMA applies the short term average/long term average (STA/LTA) method as trigger criteria. It is most widely adopted in the seismological community and used to limit seismic recordings to relevant seismic event waveforms. When the ratio of STA/LTA exceeds the trigger threshold, it is supposed that a significant signal, possibly an earthquake signal, has been detected by the considered station.

STA/LTA trigger may occur due to some troubles or local noise at one station. To avoid "false trigger", the JMA also applies other criteria including the group trigger criterion. A group is set to include several neighboring stations. When the STA/LTA trigger turns on at a certain number of stations in one group, the system considers it to be caused by an earthquake.

For the precise picking of the *P* onset, an auto regressive (AR) model is applied to represent seismogram time series. Akaike's information criterion (AIC) is used to search the optimal time to separate the seismogram into two sub-stationary time series (noise part and signal part), which are represented by AR models, respectively, in a certain time window around trigger time [32,40].

The usual least square method is used in the hypocenter calculation. As for the velocity structure, the Japan local standard model JMA2001 [36] is applied. To better locate the hypocenter, *S* arrival times are used in addition to *P* arrival times. First, a preliminary hypocenter is calculated by using only *P* arrival times picked by an automatic picker. Then, *S* arrival times are picked automatically by applying the AR model and AIC in a time window centered at the theoretical *S* arrival times estimated from the preliminary hypocenter location and origin time. The final hypocenter is then calculated by using both the *P* and *S* arrival times.

For magnitude calculation, one of the local magnitude definitions in Japan, the JMA magnitude (Mjma) according to Katsumata [19] is used, in which average amplitude decay characteristic in Japan is taken as a decay correction term of the formula. Maximum amplitude of the horizontal ground displacement is the measured input value in the formula. The displacement waveform is calculated by applying real-time recursive filter that integrates twice the accelerometer data and applies a high-pass filter to avoid computational instability. Due to historical reasons, the long-period cut-off of the high-pass filter is set at 5 seconds. Since unsaturated strong-motion data are used, magnitude calculation can be started promptly after the earthquake occurrence by using data from stations very close to the epicenter. Therefore, such magnitude data are suitable for the tsunami warning purposes of local events.

The comparison between Mjma and the moment magnitude Mw of the Global CMT Project's solution is shown in Fig. 3. Mjma values are on average comparable with Mw for earthquakes with magnitudes between M6.0 and about 8.0.

All procedures mentioned above are run automatically by the computer. Considering the possible impact of a tsunami warning on the potentially affected communities, a human check of the adequacy of the calculated results is indispensable for minimizing false alarms. The fol-

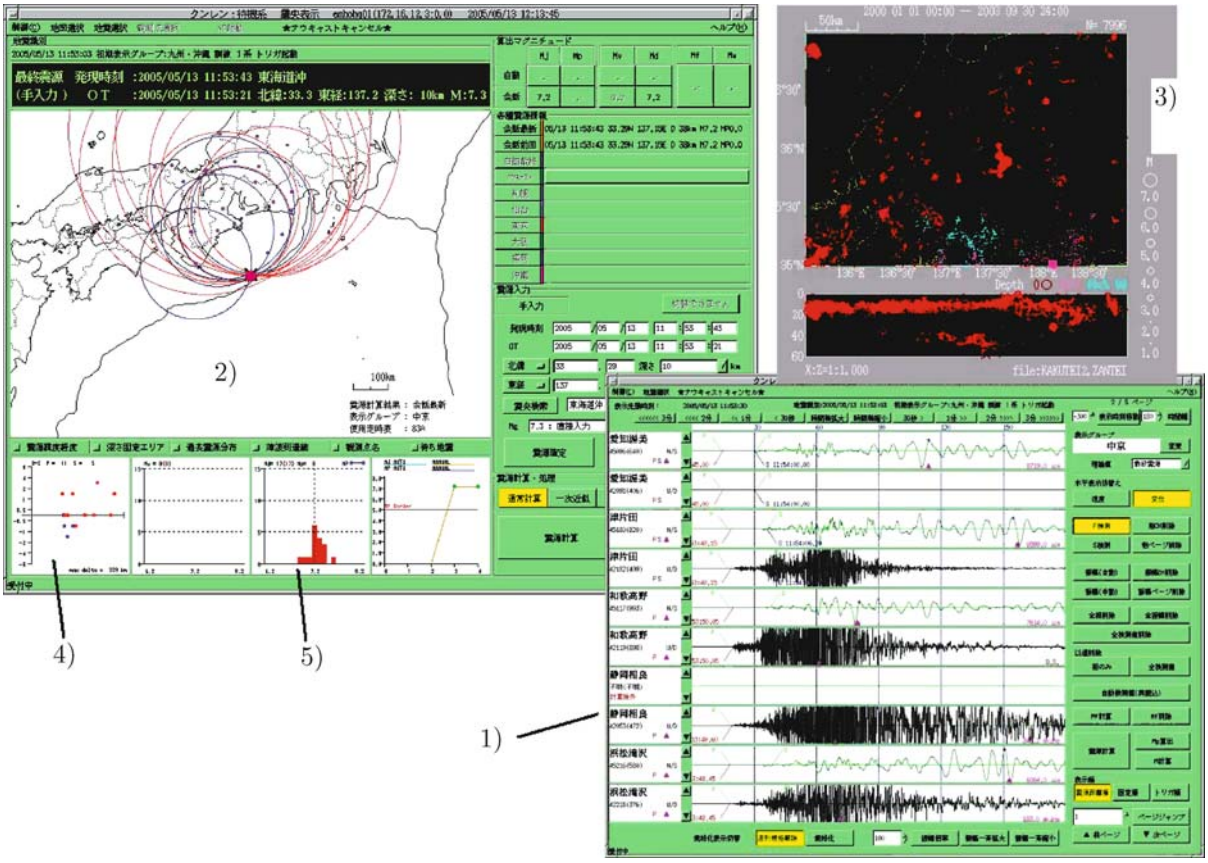

Tsunami Forecasting and Warning, Figure 4
Examples of the JMA's man–machine interface screen image in the seismic data processing system. *Numbers* in the figure are related to the respective numbers of items explained in the main text

lowing items are checked by the operators, using a man–machine interface:

1. Adequacy of the phase pickings and maximum amplitude readings: The seismic waveforms are plotted together with picked and theoretical arrival times for *P* and *S*, and the marked position at which the maximum amplitude was measured. This plot also allows them to discriminate noise from seismic signals.
2. Adequacy (mathematical) of the hypocenter location: A hypocenter plot map is used in which the locations of used seismic stations and the circles denoting the calcu-

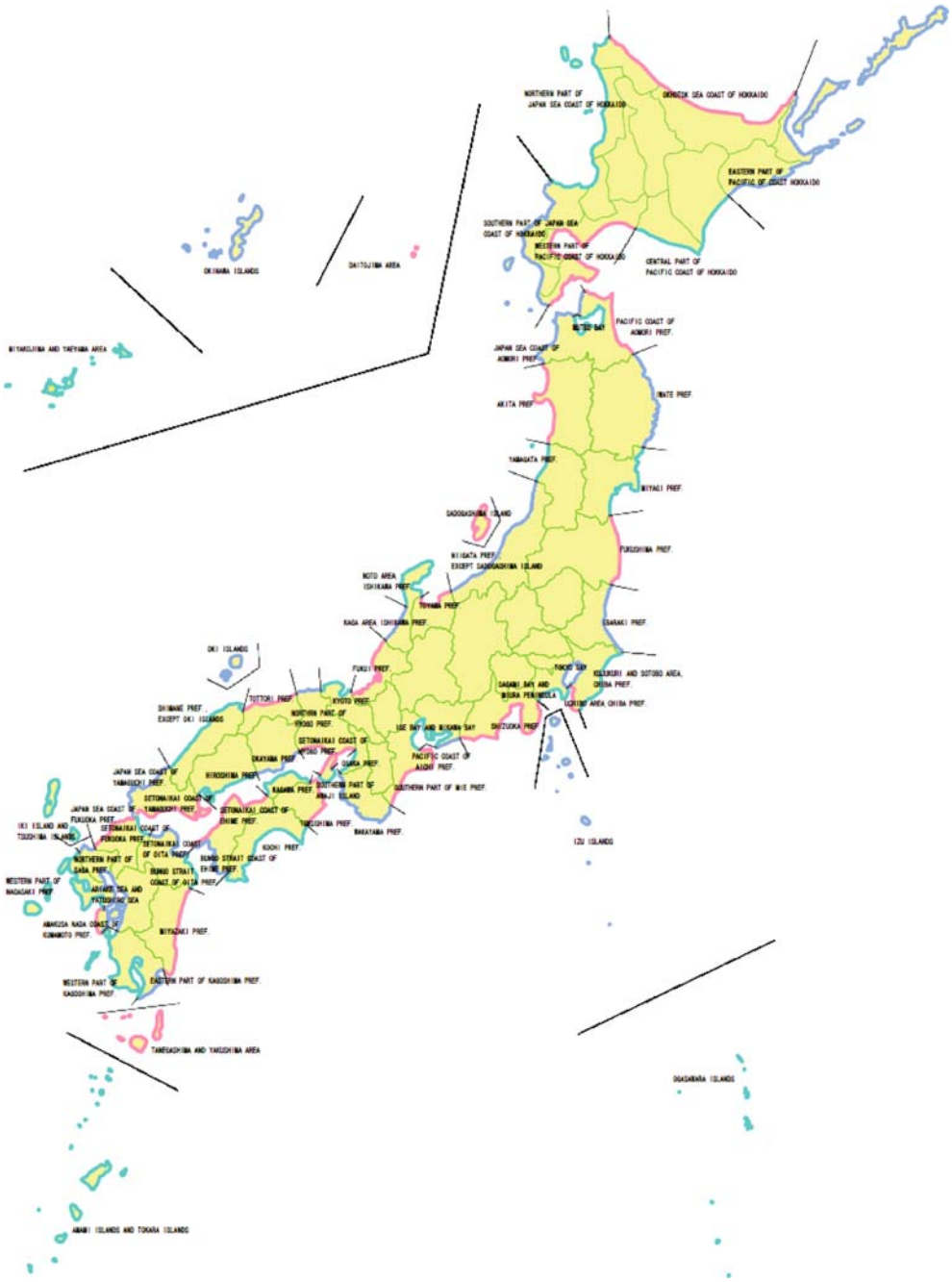

Tsunami Forecasting and Warning, Figure 5
Coastal block partitioning for tsunami forecast in Japan. *Straight lines* denote the borders between the 66 coastal blocks

Tsunami Forecasting and Warning, Table 1
Tsunami forecast grades and corresponding levels of expected maximum tsunami amplitudes as used in tsunami information issued by the JMA for each coastal block after a tsunami forecast has been made

Forecast Grade		Levels of Estimated Tsunami Amplitude
Warning	Major Tsunami	3 m, 4 m, 6 m, 8 m, 10 m or greater
	Tsunami	1 m, 2 m
Advisory		0.5 m

lated epicentral distance from each station are shown. If the circles intersect densely at one point, and if the used stations assure a wide azimuthal coverage around the source, then the calculation is judged as fine.

3. Adequacy (seismological) of the hypocenter location: The hypocenter plot on the background seismicity map is used together with the vertical cross section. If the hypocenter is located in a region where no background seismicity is found, it should be re-examined carefully.
4. Adequacy of the depth estimation: Travel time residuals are plotted as a function of epicentral distance. If the residuals depend on epicentral distance, the depth estimation should be revised.
5. Adequacy of the magnitude calculation: A histogram of station magnitudes is used. Outliers are checked and excluded in the re-calculation.

Sample images of the man–machine interface are shown in Fig. 4.

The JMA's man–machine interface is designed so that re-picking and recalculation of hypocenter and magnitude can be done within about 30 seconds.

Calculated hypocenter parameters (latitude, longitude, depth and magnitude) are transferred to the following tsunami forecast system as its input.

Tsunami Forecast System

Tsunami Forecast in Japan For a tsunami forecast to be useful, a mere general description of the tsunamigenic potential estimated from calculated hypocentral parameters is not sufficient. From the issued message, the recipient disaster management organization should be able to clearly grasp the severity of the expected disaster in its own jurisdiction. In Japan, a tsunami forecast is disseminated to each coastal block (the coast line of Japan is separated into certain number of blocks), giving three kinds of grade corresponding to the anticipated severity of the disaster.

Tsunami Forecast to Coastal Block The coastal blocks divide the Japanese coasts into 66 regions (Fig. 5). These blocks are defined by taking into account:

a) The administrative districts of local governments, which are the disaster management units to take actions in emergency situations
b) The sea areas to which each coast is facing
c) The uniformity in the behavior of tsunami
d) The precision of tsunami forecast technique

Basically, one coastal block corresponds to one prefecture, or finer.

Before the introduction of numerical tsunami simulation technique, the total number of coastal blocks was 18, due to a low precision of empirical tsunami amplitude estimation method.

Tsunami Forecast Criterion and Category The JMA categorizes tsunami forecast into "Tsunami Warning" and "Tsunami Advisory". Further, "Tsunami Warning" is divided into two grades, "Tsunami" and "Major Tsunami".

In Japan, there exist ample number of materials on the relation between tsunami disaster and observed tsunami amplitude along the coast. From the materials, high correlation between these two can be seen. Therefore, estimated maximum tsunami amplitude at the coast is used as a criterion of the tsunami forecast in Japan. The speed of water current in the sea near the coast could be another good index for assessing the tsunami disaster potential. But, up to now, only a few examples of the current observation are available. Therefore, the estimated tsunami amplitude at the coast is still the most appropriate criterion for tsunami forecast.

In Japan, tsunami disaster in the land area occurs when the tsunami amplitude exceeds 1 meter [30], so the warning criterion is set at 1 meter. When tsunami amplitude exceeds 3 meters, the proportion of damaged wooden houses and fishing boats increases remarkably [9,28,29]. Therefore, the JMA issues the highest warning grade "Major Tsunami" when the estimated maximum tsunami amplitude is 3 meters or higher. When the estimated maximum tsunami amplitudes range between 1 meter to 3 meters, the grade name is "Tsunami". If a "Tsunami Warning" is issued, the head of municipality must order evacuation to the residents in the "Tsunami Evacuation Zone" assigned by each municipality in advance.

Even when the tsunami amplitude is less than 1 meter, it is occasionally the case that bathing people are affected and aquaculture rafts for marine products industry are damaged. Considering safety criteria in sea bathing places (beaches, bays), and the relation between the tsunami amplitude near the coast and damage on aquaculture rafts in the past, the JMA's criterion for "Tsunami Advisory" is set at 20 centimeters. In the case that a "Tsunami Advisory" is issued, there is no need of evacuation in the land areas, except vulnerable very low-land areas, but people on beaches or swimming should go to higher places.

Table 1 shows the tsunami forecast grades and levels of the estimated maximum tsunami amplitude that are used by the JMA when issuing tsunami information for each coastal block after a tsunami forecast has been made.

When the estimated tsunami amplitude is less than 20 cm, no warnings or advisories are issued, but a message "No threat of tsunami disaster" is promptly provided to the public in order to prevent unnecessary spontaneous evacuations in accordance with the widely spread recommendation in Japan: "When you feel a strong shaking near a coast, run to a high place without waiting for tsunami forecasts from the JMA".

Tsunami Database Creation The JMA introduced in 1999 a numerical simulation technique to implement accurate tsunami warning. However, even most-advanced computers require substantial time for the completion of the calculations. Therefore it is impossible to issue prompt tsunami warning based on numerical simulation tech-

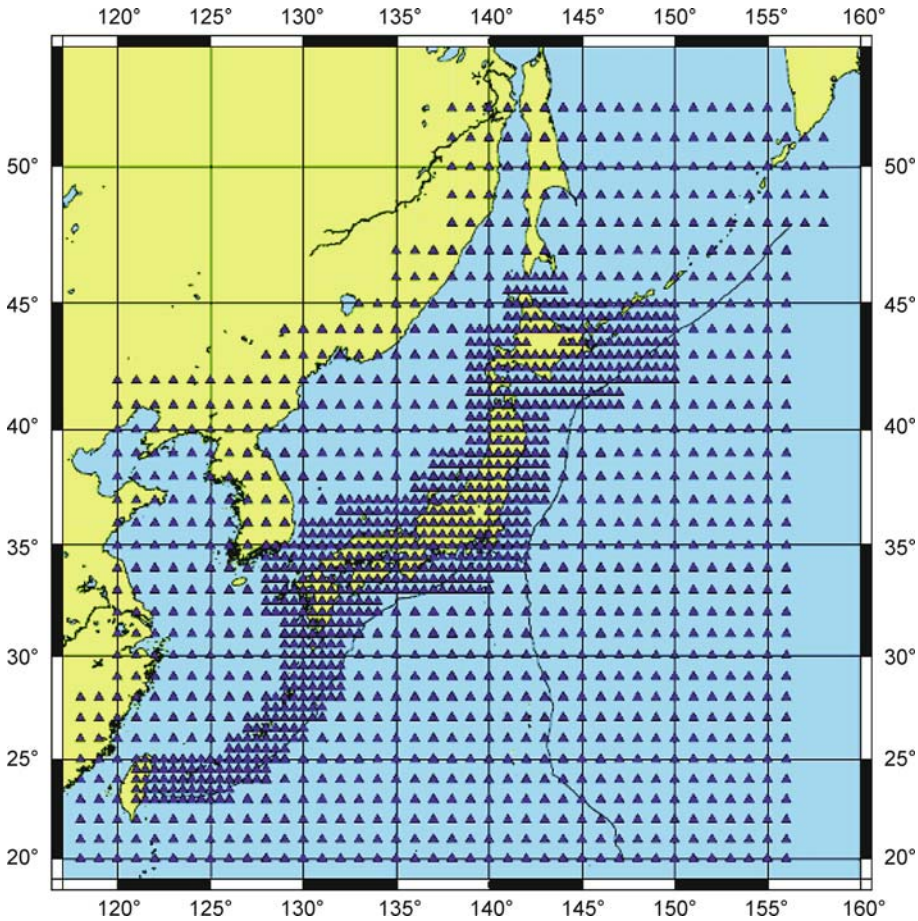

Tsunami Forecasting and Warning, Figure 6

Distribution of simulation points in and around Japan. *Solid triangles* denote locations of simulation points to which likely hypothetical fault models are attributed. *Spacing* between the simulation points is 0.5 degrees in near coastal areas, and 1.0 degree in off-shore areas

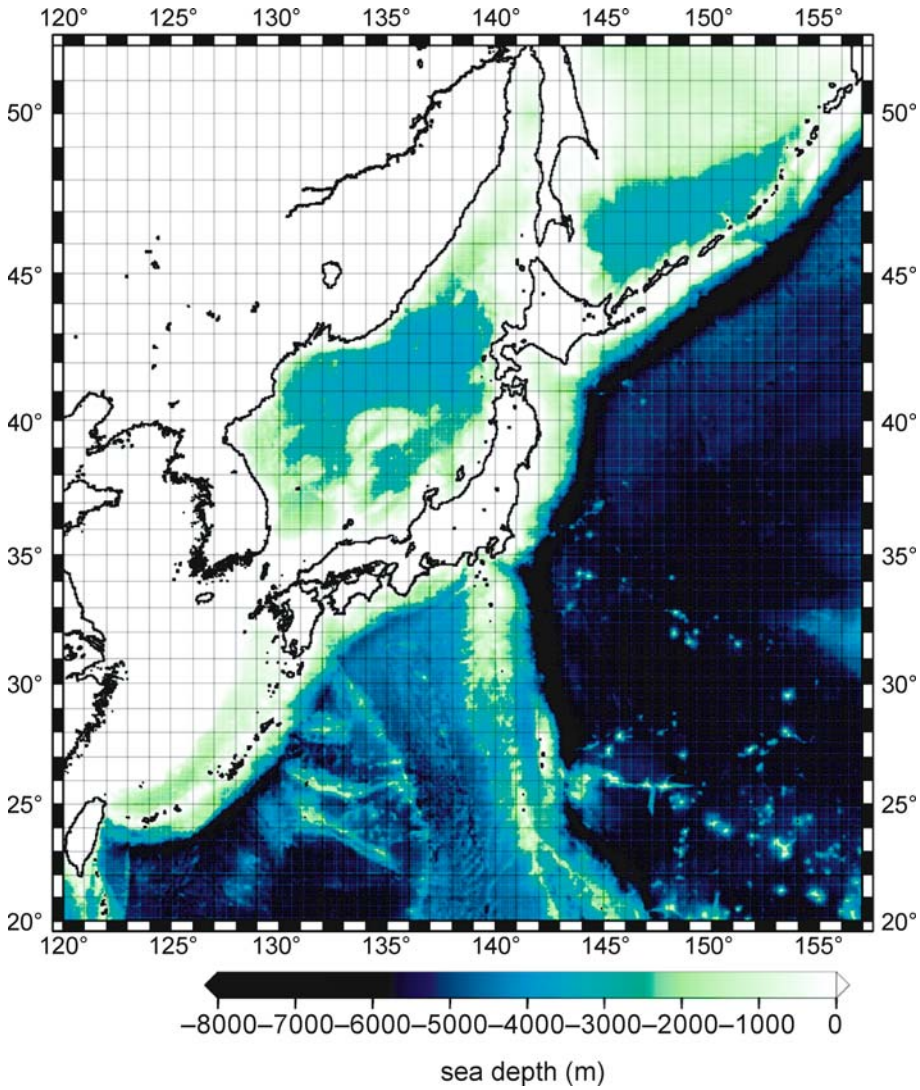

Tsunami Forecasting and Warning, Figure 7

Computational area for creating the Japanese tsunami simulation database. Bathymetry is given by a *color grade scale*. One arc minute mesh bathymetry data are used throughout this area for the tsunami simulation

nique even if the calculations start simultaneously with the occurrence of an earthquake.

Alternatively, the JMA employs a database method to achieve a breakthrough on this issue. Tsunami propagation originating from various locations, fault sizes and likely rupture mechanisms are calculated in advance and the results, namely estimates of maximum tsunami amplitudes and arrival times, are stored in a database together with the associated hypocenter location and magnitude.

If an earthquake occurs, the most appropriate case for the actual hypocenter location and magnitude is retrieved from the database and the tsunami forecast is issued accordingly.

Hypocentral Location The first step to create a tsunami database is to set fault models. 'Simulation points' are defined as surface projection locations of the center of possible causing faults, and are placed in the areas where tsunamigenic earthquakes are likely to occur. These points are selected on the basis of background seismicity and the records of the past earthquakes that generated tsunami.

Figure 6 shows the locations of simulation points in Japan and surrounding areas.

The total number of scenarios calculated for these points is about 64,000 (lat. × lon. × dep. × mag.). Earthquakes with hypocenter depth of more than 100 km or magnitude less than 6.2 are not considered, because they do not cause disastrous tsunami [23].

Fault Parameters Setting Figure 1 shows fault parameters used in the numerical simulation. Length L, width W, and slip amount D of the fault are represented in terms of empirical formulas as functions of magnitude. The following are common relations used in Japan [37] based on M = Mjma:

$$\log L = 0.5M - 1.8 \quad \text{with } L \text{ in km},$$
$$W = L/2 \quad \text{or} \quad \log W = 0.5M - 2.1 \quad \text{with } W \text{ in km},$$
$$\log D = 0.5M - 3.3 \quad \text{with } D \text{ in m}.$$

These formulas are consistent with the next two formulas, assuming a common rigidity $\mu = 2.0 \times 10^{10}$ Newton/m^2 of the rock-material in the source area

$$\mathrm{Mo} = \mu \times L \times W \times D$$
$$\log \mathrm{Mo} = 1.5\,\mathrm{Mw} + 9.1$$

(in international standard units, i. e., Mo in Nm = Newton meter).

The dip (δ), strike (ϕ) and slip (λ) angles of the fault at each simulation point are based on average values of past earthquakes at or near these locations. But if they are uncertain, they are assumed to be those of a pure reverse fault ($\lambda = 90$ degrees) whose strike is parallel to the trench or nearby coast, and with a dip angle of 45 degrees, corresponding to an efficient tsunami generation in view of disaster management.

Initial Value for Numerical Simulation of Tsunami Propagation The vertical deformation of sea floor due to fault motion is calculated by the elastic theory [24], using fault parameters as set above. The initial deformation of the sea surface is assumed to be identical to the vertical deformation of the sea floor and used as input value for the numerical simulation, because in most cases, rupture propagation velocity is much higher than the tsunami propagation velocity, and the ocean bottom deformation can be treated to occur instantaneously [14].

Numerical Simulation of Tsunami Propagation Non-linear long-wave approximation equation with advection term and ocean-bottom friction term is adopted as the equation of motion, and solved together with the equation of continuity as shown below by finite difference method

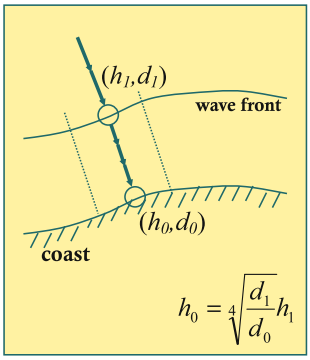

Tsunami Forecasting and Warning, Figure 8
Schematic sketch to illustrate the application of Green's Law for tsunami amplitude calculation in near coastal areas. Ray convergence and divergence are ignored

with staggered leap frog scheme [27].

$$\frac{\partial Vx}{\partial t} + Vx\frac{\partial Vx}{\partial x} + Vy\frac{\partial Vx}{\partial y} = -fVy - g\frac{\partial h}{\partial x} - C_f\frac{Vx\sqrt{Vx^2 + Vy^2}}{d + h}$$
$$\frac{\partial Vy}{\partial t} + Vx\frac{\partial Vy}{\partial x} + Vy\frac{\partial Vy}{\partial y} = fVx - g\frac{\partial h}{\partial y} - C_f\frac{Vy\sqrt{Vx^2 + Vy^2}}{d + h}$$
$$\frac{\partial h}{\partial t} + \frac{\partial}{\partial x}\{Vx(h + d)\} + \frac{\partial}{\partial y}\{Vy(h + d)\} = 0,$$

where V_x and V_y are the x (east) and y (south) components of the average water particle motion velocity in the depth direction, and h and d are sea surface displacement and sea depth, respectively. f is the Coriolis parameter ($= 2\Omega \cos\theta$, Ω is the angular frequency of Earth's self rotation, θ is the co-latitude), and C_f is the sea bottom friction coefficient.

► **Tsunami Forecasting and Warning, Figure 9**
The map shows the distribution of forecast points (*red dots*) along the southern and eastern shore of Hokkaido and part of NE Honshu together with the surface projections of two hypothetical faults representing earthquakes with magnitude 8.0 (*red rectangle*) and 6.8 (*violet rectangle*), respectively. The depth of the fault top is 1 km from the sea bottom for both cases. Pure reverse fault with dip angle 45 degrees is assumed. In the *upper half* of the figure the coastal blocks are enlarged, showing the locations of forecast points denoted by numerals and the *blue separation lines* between sub-sections of the coast. The numbers correspond to the 'forecast point number' given on the abscissa of Fig. 10. Also depicted are bathymetry contour lines near the coast at 20 m depth intervals

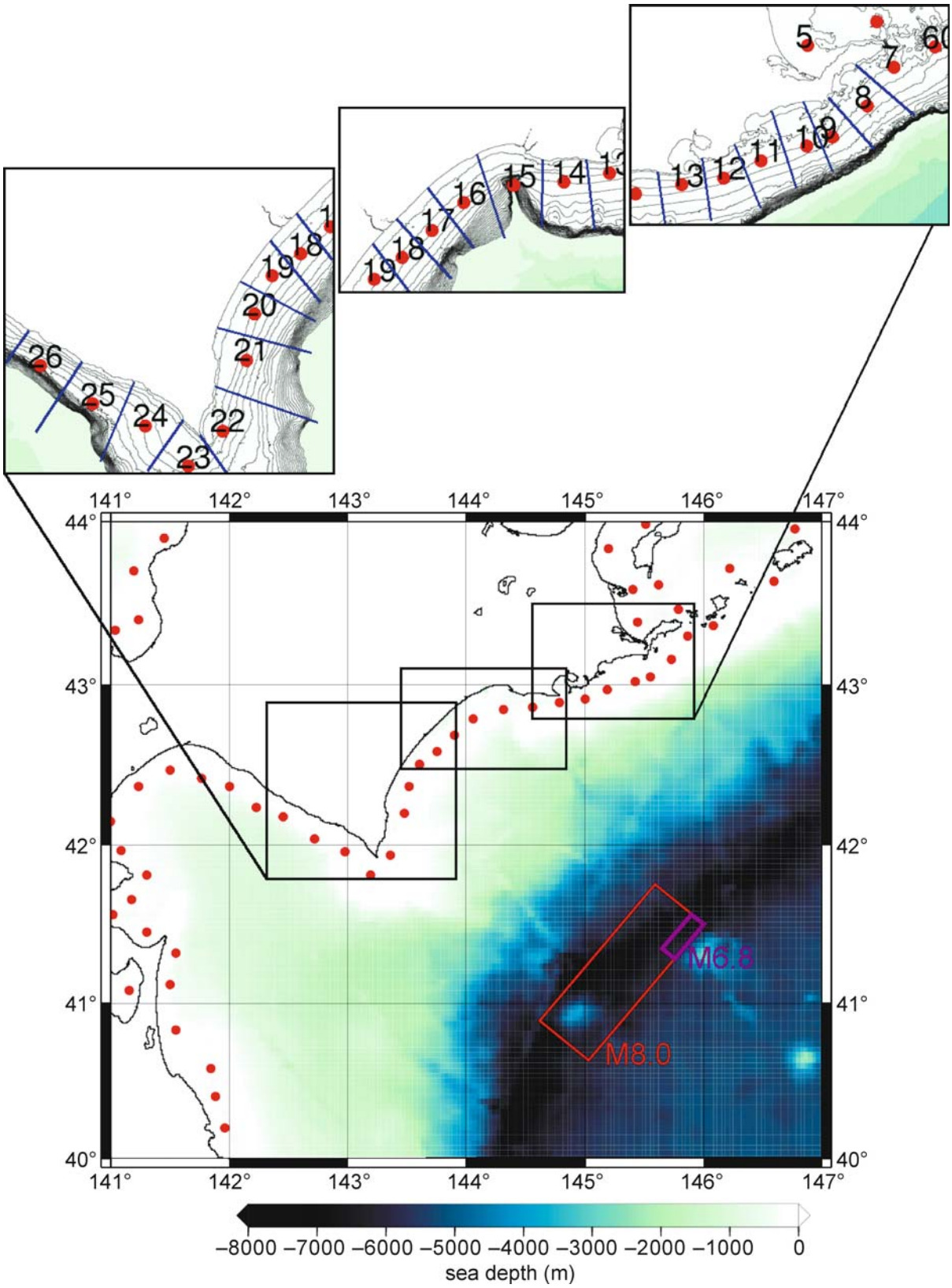
141°
142°
143°
144°
145°
146°
147°
44°
43°
42°
41°
40°
M6.8
M8.0
−8000 −7000 −6000 −5000 −4000 −3000 −2000 −1000 0
sea depth (m)

Advection and ocean bottom friction are considered only in sea areas with sea depth less than 100 m. Coriolis force is considered only in the distant tsunami case (see Sect. "Tsunami Forecast for Distant Event and Northwest Pacific Tsunami Advisory (NWPTA)").

Figure 7 shows the computational area. Mesh size is 1 arc minute throughout the computational area. Eight hours of propagation simulation is conducted for all scenarios with a fixed time interval three seconds for the integration. The Coriolis force is not considered for local and regional tsunami simulations. Total reflection boundary condition is used at the land-ocean boundary, and open boundary condition is used outside of the computational area.

To represent the tsunami waveform correctly in a shallow sea area, very fine bathymetry data mesh is necessary (in a strict sense, 20 or more grid points are necessary within one wave-length [31]), and a vast time is required for the completion of such detailed calculations. To overcome this difficulty, the numerical simulation with the long-wave approximation is applied only to points which are a few to a few ten kilometers seaward from the coast ("forecast points") where sea depth is about 50 m. Then, tsunami amplitude at the coast is calculated by using Green's law described in the next section.

Derivation of Tsunami Amplitude at Coast To estimate tsunami amplitude at the coast, Green's law (= energy conservation law) is applied to the estimated amplitude at the forecast points just seaward of the coast. As shown in Fig. 8, Green's law says that the tsunami amplitude at the coast is represented by fourth root of the ratio of sea depth at the forecast point and at the coast:

$$h_0 = \sqrt[4]{\frac{d_1}{d_0}}\, h_1 \,.$$

h and d denote tsunami amplitude and sea depth, and suffix '0' and '1' denotes the value at the coast and forecast point.

As the forecast points are set close enough to the coast, horizontal convergence and divergence of the rays can be ignored, i. e., the wave front is regarded as almost parallel to the coast due to a refraction towards the direction of the largest sea-depth gradient. As for the "d_0" value, the sea depth at the coast, 1 m is assumed, but actually measured value would be more appropriate, if available.

Adequacy of the Green's Law Application The adequacy of Green's Law application to estimate tsunami amplitude at the coast is examined by using the TSUNAMI-N2 [12] program. It is equivalent to the program used at the JMA [27] to create tsunami simulation database. Up to 150 m mesh at the finest was used by adopting nesting technique, and the tsunami amplitudes estimated at the coastal mesh are considered "true" ones. We compared the case using 1,350 m mesh throughout the computational area (which is equivalent to 1 arc minute mesh) and then estimating the tsunami amplitude at the coast by applying Green's Law to the value at the offshore forecast point, with the expected "true" amplitude at the coastal point using finer mesh. Forecast points actually used for the database creation are shown in Fig. 9. Hypothetical fault models corresponding to two different magnitudes (8.0 and 6.8) have been depicted in this Figure as red and violet rectangles, respectively, south off the Pacific coast of Hokkaido Island. The results are shown in Fig. 10a,b. For both events the estimated tsunami amplitudes at the coast by using a 1,350 m mesh and applying Green's Law to the forecast points (solid squares) agree well with the means (open squares) and medians (open circles) of tsunami amplitudes calculated by using a finer mesh in each coastal subsection. The coastal subsections are separated by middle points between the forecast points (see upper part of Fig. 9). Thus, it was confirmed that the application of Green's Law to tsunami amplitudes at the offshore forecast points yields amplitude estimates at the coast that are comparable with those calculated with much more time consuming fine mesh simulations. Therefore, we consider the former as representative estimates of the tsunami amplitude in the coastal subsection centered at the coastal projection of the forecast point.

It could also be shown that direct estimates of tsunami amplitudes at the coast based on coarse mesh simulations will underestimate values. This trend is even clearer for smaller event with shorter typical tsunami wavelengths (Fig. 10b). Therefore, it is not proper to estimate the tsunami amplitude directly at the coastal mesh when the mesh size is coarse relative to the typical tsunami wavelength.

However, there will be some issues to be considered in future in this method as follows.

(a) Strictly speaking, the Green's Law should only be applied to a direct wave. Therefore, in case the maximum amplitude is created by the superposition of multiply reflected waves, then this method can give overestimated tsunami amplitude.
(b) Scatter of estimated tsunami amplitude in one coastal subsection differs from section to section depending on the complexity of the coast line feature. It would be very difficult to precisely represent tsunami ampli-

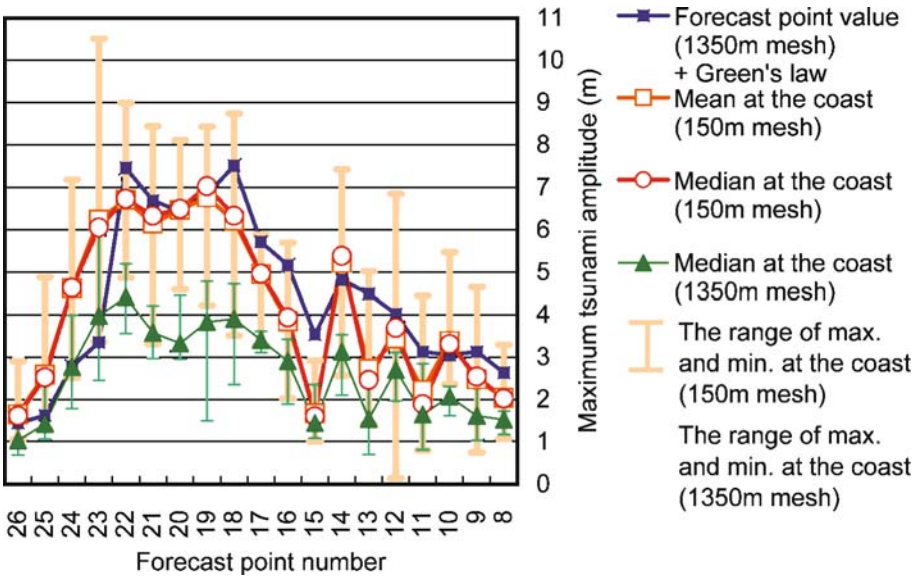

Tsunami Forecasting and Warning, Figure 10a

The presented diagram illustrates the adequacy of the Green's law application. Shown are the tsunami forecast results for the hypothetical earthquakes of magnitude 8.0 (*above*) and 6.8 (*below*). On the abscissa the numbers of the 'forecast points' depicted in Fig. 9 are given. On the ordinate the maximum tsunami amplitudes in meter are given, as they result from the application of different modeling procedures (fine mesh as well as course mesh with and without applying Green's law; see legend). *Vertical thick bars* denote the range between maximum and minimum values of estimated tsunami amplitudes in each sub-section using an up to 150 m mesh. *Vertical thin bars* denote the range of directly estimated maximum and minimum tsunami amplitudes at the coast in each sub-section using a 1,350 m mesh

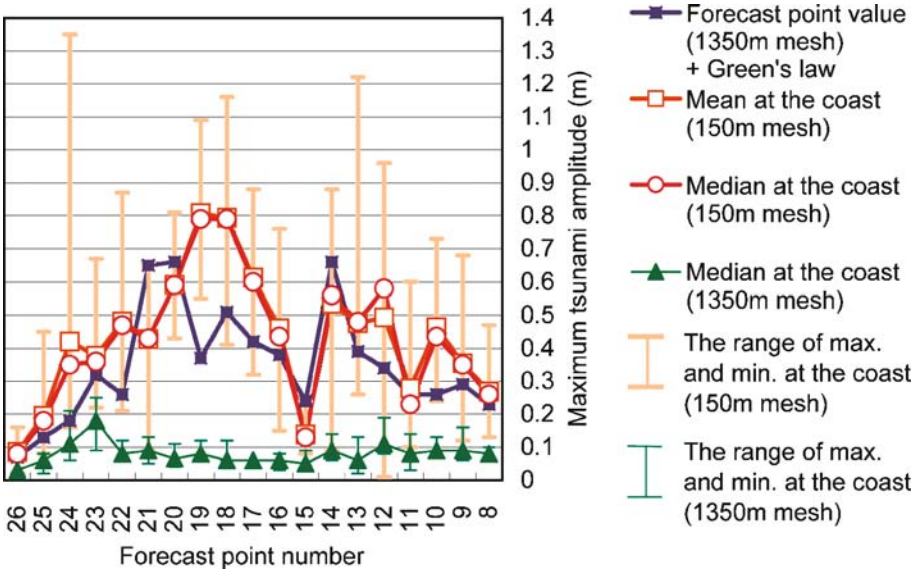

Tsunami Forecasting and Warning, Figure 10b
(continued)

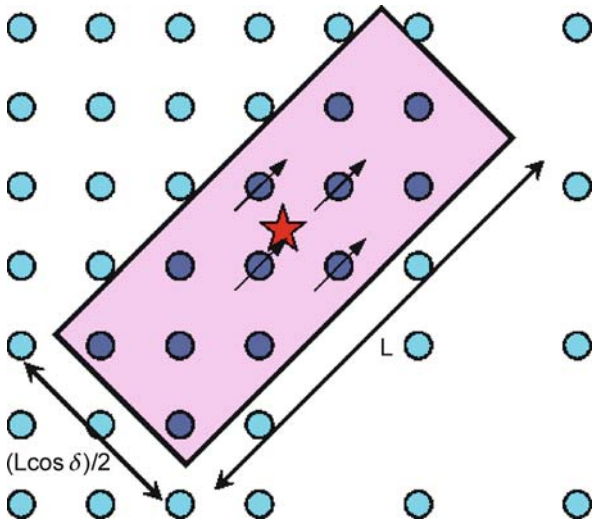

Tsunami Forecasting and Warning, Figure 11
Maximum Risk Method: The *star* denotes the epicenter and the *circles* the locations of simulation points. The *rectangle* delimitates the searching area from which the simulation points have to be retrieved. This rectangle is centered at the epicenter, and its length and width are L and $0.5L\cos\delta$, respectively, where L is given by the scaling law (see Subsect. "Fault Parameters Setting"). The strike of the length is the same as that given to the nearest simulation point to the epicenter. All simulation points inside of the rectangle (*solid circles*) are selected. *Small arrows* point into the strike direction assigned to simulation points

Tsunami Forecasting and Warning, Table 2a
Tsunami forecast/information examples of the JMA. Original texts are in Japanese and these are English translations: a Tsunami forecast of the first issuance, b Tsunami information (estimated tsunami arrival time and amplitude) that follows a, c Tsunami information (estimated high tide and tsunami arrival times) that follows a, d Tsunami information (Observed results), e Tsunami forecast revision, f Tsunami forecast cancellation

Tsunami Warning Issued by the Japan Meteorological Agency (JMA) Issued at 2029JST 15 Nov 2006
"Tsunami Warning"
Eastern Part of Pacific Coast of Hokkaido
Okhotsk Sea Coast of Hokkaido
"Tsunami Advisory"
Central Part of Pacific Coast of Hokkaido
Western Part of Pacific Coast of Hokkaido
Northern Part of Japan Sea Coast of Hokkaido
Pacific Coast of Aomori Pref.
Iwate Pref.
Miyagi Pref.
Fukushima Pref.
Ibaraki Pref.
Kujukuri and Sotobo Area of Chiba Pref.
Uchibo Area of Chiba Pref.
Izu Islands
Sagami Bay and Miura Peninsula
Shizuoka Pref.

tudes for individual coastal points even if very fine bathymetry data are used in a simulation, because a short wavelength tsunami is significantly affected by small scale bathymetry and coastal feature or by a small change in an initial tsunami waveform. Therefore, incorporating a standard deviation of estimated tsunami amplitude to represent a degree of scatter in one coastal subsection, as well as a median or mean, in the tsunami warning/advisory grade determination would be effective for disaster management.

Retrieval of the most Appropriate Case from the Database In general, no identical case for actually determined latitude, longitude, hypocenter depth, magnitude exists in the database. Thus, some kind of selection rule is necessary to choose the most appropriate one.

Hypocenter Depth and Magnitude The logarithm of the estimated tsunami amplitude changes approximately positively linear with magnitude and negatively linear with hypocenter depth. Therefore, first, one has to calculate the logarithm of the estimated tsunami amplitudes for four cases (i. e., for two different depths and magnitudes each) which include the calculated depth and magnitude in between, then apply a two-dimensional linear interpolation and finally convert the log values back to a tsunami amplitudes.

Latitude and Longitude As for the location of the epicenter, either one of the two methods is used depending on the hypocentral area.

Interpolation method One selects the nearest four locations of simulation points in the data base which enclose the epicenter determined for the real event and performs a two-dimensional linear interpolation. This method is used far off the coast areas where horizontal shifts of the source location cause relatively smooth changes in the estimated tsunami amplitude at the coast. This is applicable in areas with a simulation point spacing of 1 degree (see Fig. 6).

Maximum risk method In the case that the assumed fault locations in the data base differ from the real location, tsunami amplitude estimates near the epicenter may differ substantially from observed ones for events oc-

Tsunami Forecasting and Warning, Table 2b
(continued)

Tsunami Information (Estimated Tsunami Arrival Time and Amplitude) Issued by the Japan Meteorological Agency (JMA) Issued at 2030JST 15 Nov 2006		
— Estimated Tsunami Arrival Time and Amplitude —		
Coastal Block	Arrival Time	Amplitude
"Tsunami Warning"		
Eastern Part of Pacific Coast of Hokkaido	2110 15 Nov	1 m
Okhotsk Sea Coast of Hokkaido	2120 15 Nov	2 m
"Tsunami Advisory"		
Central Part of Pacific Coast of Hokkaido	2130 15 Nov	0.5 m
Western Part of Pacific Coast of Hokkaido	2150 15 Nov	0.5 m
Northern Part of Japan Sea Coast of Hokkaido	2250 15 Nov	0.5 m
Pacific Coast of Aomori Pref.	2140 15 Nov	0.5 m
Iwate Pref.	2140 15 Nov	0.5 m
Miyagi Pref.	2140 15 Nov	0.5 m
Fukushima Pref.	2210 15 Nov	0.5 m
Ibaraki Pref.	2210 15 Nov	0.5 m
Kujukuri and Sotobo Area of Chiba Pref.	2210 15 Nov	0.5 m
Uchibo Area of Chiba Pref.	2210 15 Nov	0.5 m
Izu Islands	2210 15 Nov	0.5 m
Sagami Bay and Miura Peninsula	2220 15 Nov	0.5 m
Shizuoka Pref.	2220 15 Nov	0.5 m
Tsunami may be higher than the estimation in some places.		
Sea level may slightly fluctuate in other coastal areas but no danger.		
— Earthquake Information —		
Origin Time: 2015JST 15 Nov 2006		
Epicenter: 46.6 North, 153.6 East		
Depth: 30 km		
Magnitude: 8.1		

curring near to the coast. The uncertainty in the relative location of the epicenter on the surface projection of the seismic fault has to be considered at an early stage of the seismic data analysis, and all possible cases should be taken into account. Four cases with the epicenter located on one of the four corners of the fault are the most extreme cases, and the truth lies in between. Therefore, in the case of near coastal sources and with a view to disaster management, cases corresponding to every simulation point inside the rectangular area shown in Fig. 11 (solid circles) have to be considered. This rectangular area is centered at the epicenter (star mark). Its length and width are determined via the empirical formulas given in Subsect. "Fault Parameters Setting". Note, however, that the projected width in Fig. 11 is only $0.5L\cos\delta$ and depends on the source dip. Also one has to remember that each simulation point is located in the center of the respective scenario fault in the data base. Therefore, the length and width of the rectangle to select all possible 'simulation points' are, L (not $2L$) and $0.5L\cos\delta$ respectively, and the strike of the length is that given to the simulation point (denoted by an arrow), which is closest to the determined epicenter. And the worst case is chosen amongst them for each of the forecast points.

When readings of the tsunami arrivals at tidal gauges or tsunami meter become available, improbable simulation points can be excluded by applying inverse refraction diagrams. Thus the uncertainty in the spatial extent of the tsunami source area can be reduced.

If the number of the simulation points inside of the rectangular area is less than 4, the nearest 4 simulation points from the epicenter are retrieved.

Tsunami Forecasting and Warning, Table 2c
(continued)

Tsunami Information (Estimated High Tide and Tsunami Arrival Time) Issued by the Japan Meteorological Agency (JMA) Issued at 2030JST 15 Nov 2006		
— Estimated High Tide and Tsunami Arrival Time —		
If tsunami arrives at coasts at around high tide time, tsunami becomes higher.		
Coastal Block	High Tide Time	Tsunami Arrival
"Tsunami Warning"		
Eastern Part of Pacific Coast of Hokkaido		2110 15 Nov
Kushiro	2335 15 Nov	2130 15 Nov
Hanasaki	2338 15 Nov	2120 15 Nov
Okhotsk Sea Coast of Hokkaido		2120 15 Nov
Abashiri	2226 15 Nov	2140 15 Nov
Monbetsu	2206 15 Nov	2200 15 Nov
Esashi-ko	2228 15 Nov	2220 15 Nov
"Tsunami Advisory"		
Central Part of Pacific Coast of Hokkaido		2130 15 Nov
Urakawa	2353 15 Nov	2140 15 Nov
Tokachi-ko	2359 15 Nov	2140 15 Nov
Western Part of Pacific Coast of Hokkaido		2150 15 Nov
Muroran	2337 15 Nov	2210 15 Nov
Hakodate	0008 16 Nov	2220 15 Nov
Tomakomai-nishi-ko	0007 16 Nov	2210 15 Nov
Yoshioka	0026 16 Nov	2230 15 Nov
Northern Part of Japan Sea Coast of Hokkaido		2250 15 Nov
Wakkanai	0229 16 Nov	2310 15 Nov
Rumoi	0114 16 Nov	2250 15 Nov
Otaru	0105 16 Nov	2250 15 Nov
Pacific Coast of Aomori Pref.		2140 15 Nov
Hachinohe	0000 16 Nov	2200 15 Nov
Sekinehama	0005 16 Nov	2200 15 Nov
Iwate Pref.		2140 15 Nov
Miyako	0006 16 Nov	2150 15 Nov
Ofunato	0017 16 Nov	2150 15 Nov
Kamaishi	0011 16 Nov	2150 15 Nov
(abbreviated)	(abbreviated)	(abbreviated)
Sea level may slightly fluctuate in other coastal areas but no danger.		
— Earthquake Information —		
Origin Time: 2015JST 15 Nov 2006		
Epicenter: 46.6 North, 153.6 East		
Depth: 30 km		
Magnitude: 8.1		

Tsunami Arrival Time Estimation Forecast points are located not along the coast but offshore. Travel time estimation by adding the travel time from the source area to the forecast point to that from the forecast point to the coastal point may lead to substantial errors, especially when the tsunami source area is close to the coast. Therefore, for the tsunami arrival time estimation at the coastal point, an inverse refraction diagram from the coastal point to the source area is used to reduce such errors. Finite spatial extent of the source area, depending on the calculated mag-

Tsunami Forecasting and Warning, Table 2d
(continued)

Tsunami Information(Observation Results)				
Issued by the Japan Meteorological Agency (JMA)				
Issued at 2207JST 15 Nov 2006				
— Tsunami Observation at Sea Level Stations as of 2205 15 Nov —				
Higher tsunami may have arrived at some places other than the sea level stations.				
Tsunami may glow higher later.				
Kushiro	First Wave	2143 15 Nov	(+)	0.2 m
	Maximum Wave	2155 15 Nov		0.2 m
Hanasaki	First Wave	2129 15 Nov	(+)	0.4 m
	Maximum Wave	2143 15 Nov		0.4 m
Tokachi-ko	First Wave	2149 15 Nov		unclear
	Maximum Wave	(arriving)		
"Tsunami Warning" has been in effect for the following coastal blocks.				
Eastern Part of Pacific Coast of Hokkaido				
Okhotsk Sea Coast of Hokkaido				
"Tsunami Advisory" has been in effect for the following coastal blocks.				
Central Part of Pacific Coast of Hokkaido				
Western Part of Pacific Coast of Hokkaido				
Northern Part of Japan Sea Coast of Hokkaido				
Pacific Coast of Aomori Pref.				
Iwate Pref.				
Miyagi Pref.				
Fukushima Pref.				
Ibaraki Pref.				
Kujukuri and Sotobo Area of Chiba Pref.				
Uchibo Area of Chiba Pref.				
Izu Islands				
Sagami Bay and Miura Peninsula				
Shizuoka Pref.				
Sea level may slightly fluctuate in other coastal areas but no danger.				
— Earthquake Information —				
Origin Time: 2015JST 15 Nov 2006				
Epicenter: 46.6 North, 153.6 East				
Depth: 30 km				
Magnitude: 8.1				

nitude value, is considered, and the earliest arrival time corresponding to the crossing point of the inverse refraction diagram and the outer rim of the source area is used with a view to disaster management.

Tsunami Forecast Assembling The tsunami forecast grade for a considered coastal block depends on the maximum of the expected tsunami amplitudes at the forecast points located in that block which are converted to amplitudes at the coast by applying Green's law. Forecast points are placed off-shore parallel to the coast with spacing of about 20 km. The average number of forecast points in one coastal block is 9. As mentioned above, expected tsunami amplitude at a forecast point, including Green's law application when using 1 arc minute bathymetry mesh, agrees reasonably well with expected mean or median tsunami amplitudes on the coast that have been calculated by using very fine bathymetry mesh in a coastal subsection separated in the middle between forecast points. This means that the expected tsunami amplitude at the forecast point

Tsunami Forecasting and Warning, Table 2e (continued)

Tsunami Warning (Revision)
Issued by the Japan Meteorological Agency (JMA)
Issued at 2330JST 15 Nov 2006
"Tsunami Warning" has been revised into "Tsunami Advisory"
for the following coastal blocks.
Eastern Part of Pacific Coast of Hokkaido
Okhotsk Sea Coast of Hokkaido
"Tsunami Advisory" has been in effect for the following coastal blocks.
Eastern Part of Pacific Coast of Hokkaido
Okhotsk Sea Coast of Hokkaido
Central Part of Pacific Coast of Hokkaido
Western Part of Pacific Coast of Hokkaido
Northern Part of Japan Sea Coast of Hokkaido
Pacific Coast of Aomori Pref.
Iwate Pref.
Miyagi Pref.
Fukushima Pref.
Ibaraki Pref.
Kujukuri and Sotobo Area of Chiba Pref.
Uchibo Area of Chiba Pref.
Izu Islands
Sagami Bay and Miura Peninsula
Shizuoka Pref.
Ogasawara Islands

Tsunami Forecasting and Warning, Table 2f (continued)

Cancellation of Tsunami Warning
Issued by the Japan Meteorological Agency (JMA)
Issued at 0130JST 16 Nov 2006
Tsunami warning has been all canceled.
"Tsunami Advisory" has been canceled for the following coastal blocks.
Eastern Part of Pacific Coast of Hokkaido
Central Part of Pacific Coast of Hokkaido
Western Part of Pacific Coast of Hokkaido
Pacific Coast of Aomori Pref.
Iwate Pref.
Miyagi Pref.
Fukushima Pref.
Ibaraki Pref.
Kujukuri and Sotobo Area of Chiba Pref.
Uchibo Area of Chiba Pref.
Izu Islands
Sagami Bay and Miura Peninsula
Ogasawara Islands
However, sea level is expected to fluctuate in above coastal areas.
Be careful in sea bathing and fishing.

can be regarded as being representative for the respective coastal subsection. We think this approach is reasonable with a view to disaster management requirements because the forecast grades for a coastal block are based on the maximum of these representative values in each coastal subsection. This prevents that the forecast grade is affected by abnormally large local maxima as they may result from very fine-mesh simulations (see Fig. 10). However, the JMA informs in its public information releases that tsunami amplitude might very locally be much higher than in the forecast.

The content of tsunami warning/advisory is very simple, namely, tsunami grades and corresponding coastal block names. This is to enable recipient organizations to understand the most important information easily, namely the necessity of evacuation. Tsunami information that follows just after the warning/advisory provides then more detailed estimates of tsunami amplitudes and arrival times for each coastal block, as well as hypocentral parameters. Examples of tsunami forecast/information are given in Table 2a–f.

Tsunami Forecast Dissemination Like other forecast/information disseminated from the JMA, tsunami forecast/information is transmitted to relevant organizations by fully utilizing existing online facilities. In case of land line failure, forecast/information is also transmitted through satellite link. Loud speaker/sires operated by municipalities/prefectures are used in 'the last mile' to reach the public. Police and fire departments have their own robust transmission routes, which are also used. Additionally, the public broadcasting company plays a very important role because of its wide outreach and timeliness.

Tsunami Monitoring System

Data from more than 100 tidal observation stations (including those operated by other institutes, like the Japan Coast Guard) are collected on-line at the JMA (Fig. 12). Data sampling and transmission rate are every one second.

Arrival time, initial amplitude, polarity, maximum tsunami amplitude and corresponding time are measured on the man–machine interface depicted in Fig. 13, after prior removal of the ocean tidal component. Tsunami observation results are issued in tsunami information.

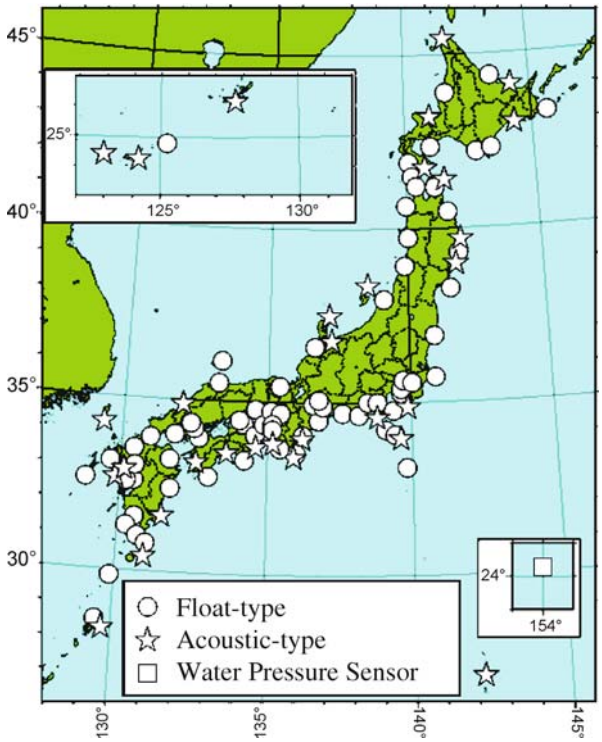

Tsunami Forecasting and Warning, Figure 12
Tidal stations network used for tsunami monitoring by the JMA. More than 100 tidal stations are monitored in real time. Symbols denote instrumentation type. *Circle*, *star* and *square* denote float-type, acoustic-type and pressure sensor type, respectively

Sea level data are used for the following: 1) Deciding about the time of tsunami warning/advisory cancellation; 2) Revision of the grade of warning/advisory in the case that estimated and observed tsunami amplitudes differ significantly and 3) Informing the public on the up-to-date status of tsunami observation.

Tsunami Forecast for Distant Event and Northwest Pacific Tsunami Advisory (NWPTA)

For a distant event, a similar method to a local event as mentioned above is used, but the seismic waveform data from global network which are available in real-time through the Internet is used for hypocenter and magnitude calculation, instead of those from domestic seismic network. The PTWC's bulletins are also incorporated in the framework of ICG/PTWS. For magnitude estimation, Mwp [34] is used commonly with PTWC and WC/ATWC. Tsunami simulation database has been created also for distant events, taking the Coriolis force effect into account considering long distance propagation. In case enough time is left until the earliest estimated tsunami arrival time at Japan coast, the JMA monitors sea level data collected via the data collection platform function of geostationary meteorological satellites, and circulated through GTS (global telecommunication network operated by the World Meteorological Organization of UN) circuit in near-real-time. After confirming the generation of tsunami at sea level observation sites near the source, tsunami warnings/advisories are disseminated for the same coastal blocks as for local tsunami if necessary. Estimated tsunami amplitudes in the database can be calibrated by the actually observed tsunami amplitudes before the dissemination. Furthermore, if a reliable causing fault model is available, the JMA conducts tsunami propagation simulation for a specific setting of fault location, depth, magnitude and mechanism so that tsunami warnings/advisories can be based on more reliable tsunami estimation.

By using the same method as described above in this section, the JMA, as a regional center of the ICG/PTWS, is providing international tsunami information (NWPTA: Northwest Pacific Tsunami Advisory) to relevant countries when an earthquake with magnitude 6.5 or larger occurs in the northwest Pacific area since March 2005. NWPTA contains estimated tsunami amplitudes and arrival times, as well as estimated hypocenter parameters.

Some Lessons Learnt from a Recent Event

At 11:14 (UTC) on Nov. 15, 2006, a large earthquake with Mw8.3 occurred in the Kuril Island Region which generated a Pacific-wide tsunami. Tsunami records at tidal gauges in Japan are shown in Fig. 14. The JMA disseminated tsunami forecasts at 11:29 (UTC) (corresponding to 20:29 JST in the related case example given in Table 2) based on the estimated tsunami amplitudes in the database. About 4 hours after the first detection of the tsunami on the Japan coast, the JMA cancelled all the forecasts at 16:30 (UTC) judging from the amplitude decay in the sea level record of Chichijima station (cf. Fig. 14) that no higher tsunami will be observed. But, after the cancellation, maximum amplitudes were recorded at many of the tidal gauges along the Pacific coast of Japan up to 5 hours later. The largest tsunami amplitude recorded in Japan for this event was 84 cm at Miyake Is. (Tsubota).

The JMA conducted a close examination of this case and found that these large amplitudes were tsunami waves reflected from the Emperor Sea Mount Chain in the middle of the Pacific (Fig. 15). Since the computational area for the tsunami simulation database of the JMA was limited around Japan (cf. Fig. 7), such reflected wave could

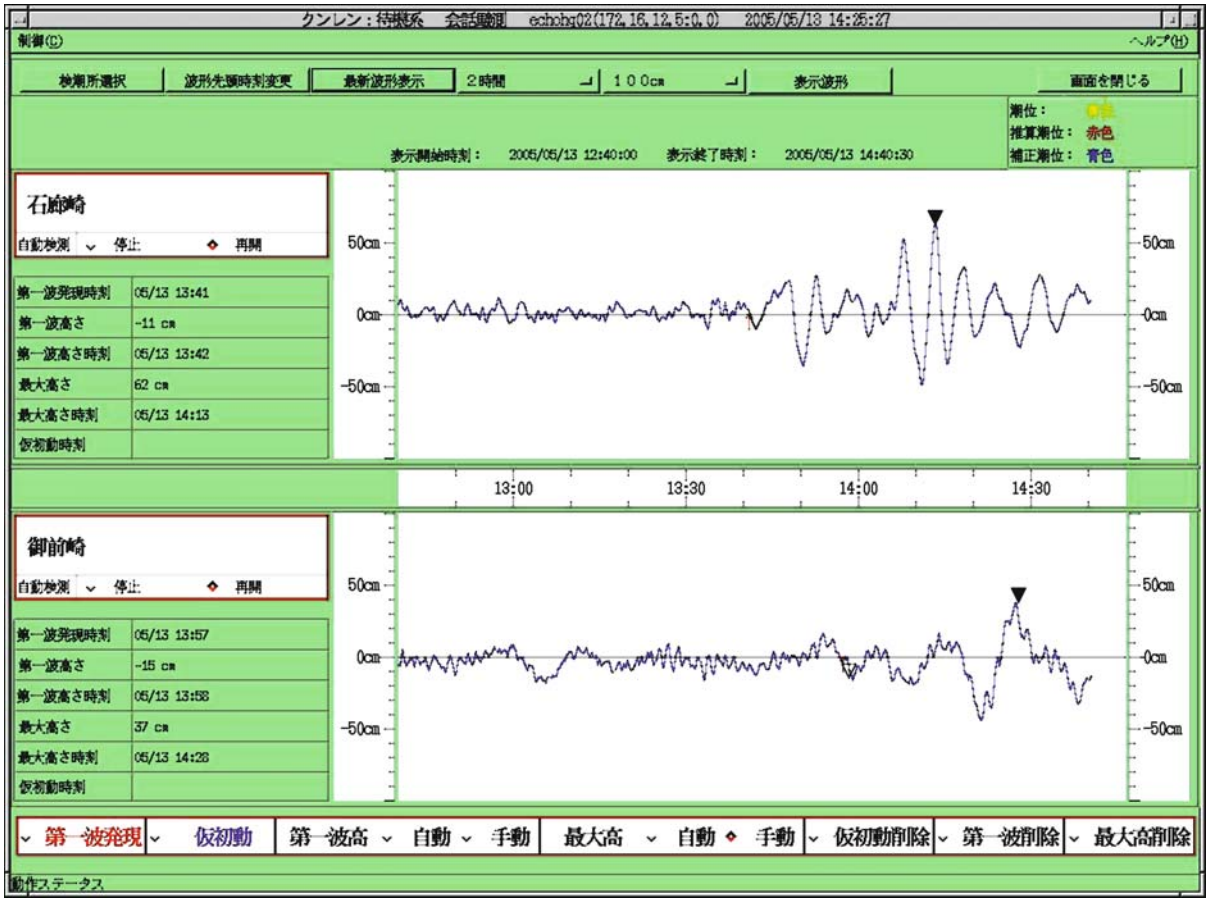

Tsunami Forecasting and Warning, Figure 13

Example of the JMA's man–machine interface screen image in the Sea Level Data Monitoring System. Shown are two sea level recordings by different tidal gauge stations. The ocean tide component has been removed for more precise reading of the relevant tsunami wave parameters. Arrival time, initial wave amplitude, polarity, maximum amplitude and corresponding time are picked. The reversed *solid triangles* denote the maximum amplitude within the analyzed time window

not be represented by the simulation. Therefore, the JMA examines now the scenarios for which the computational area should be expanded.

Recent Improvements

Application of Earthquake Early Warning (EEW) Technique to Quicken Tsunami Forecast In the near-coastal area (i. e., roughly within 100 km from the coast) EEW hypocenter and magnitude estimates [15] are equivalent to those resulting from the common procedures described above. Therefore, EEW results can be used as input data for tsunami forecast. The JMA started this incorporation in October 2006. In the case of the Mjma6.8 Niigataken Chuetu-oki earthquake of 16 July 2007, the JMA disseminated tsunami advisory one minute after the first detection of the seismic wave.

Quicker Revision/Cancellation of Tsunami Forecasts by Utilizing CMT Solutions In JMA, automatic CMT solutions are available in 10 to 20 minutes after the earthquake occurrence. They yield more realistic magnitudes than Mjma for really great earthquakes (Mw > 8) and reliable fault plane solutions. On the other hand, it is not proper to wait for the CMT solution for issuing the first tsunami warning, and it is not realistic neither to prepare the tsunami simulation database for so many different dip, strike and slip angle settings with fine increments and choose the best fitting one, because it requires huge storage and computational time for the conduct of simulations for all parameter settings. Therefore, we use the CMT solution for the revision/cancellation of the first tsunami warning, and we limit the number of different fault parameter settings to 4, namely pure reverse fault with dip angles 10, 45 and 80 degrees and pure strike fault. Normal fault with

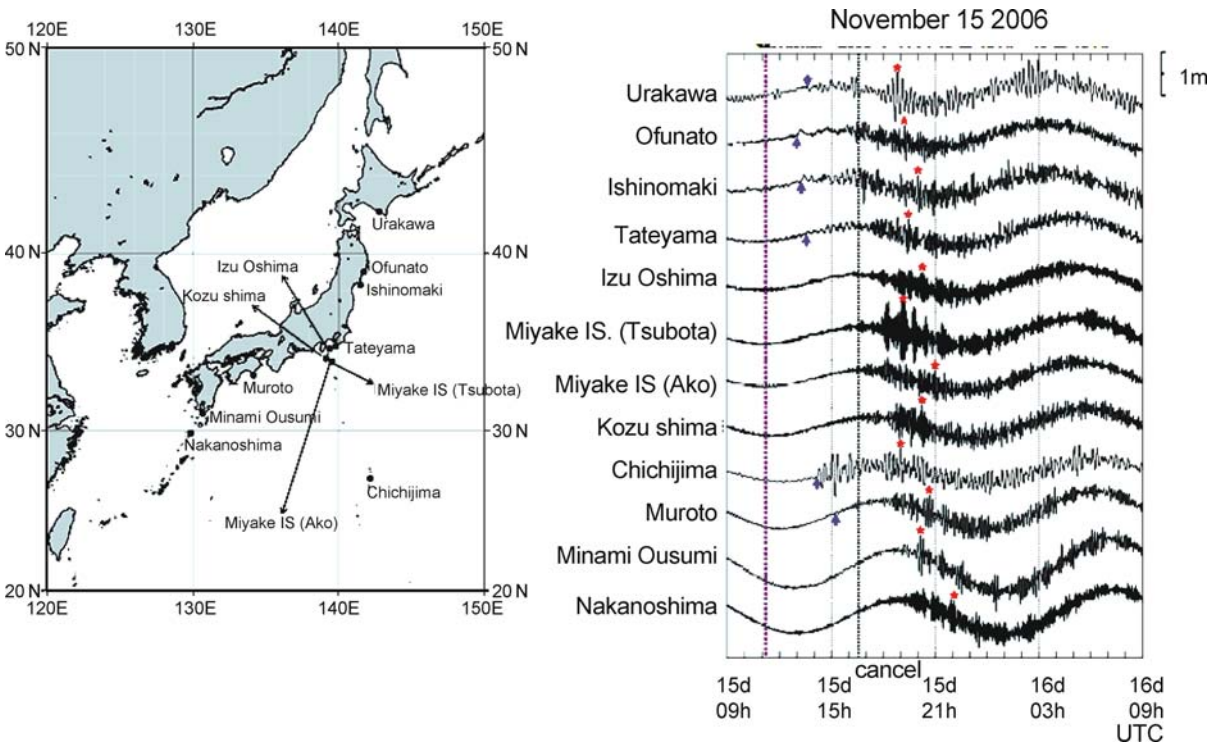

Tsunami Forecasting and Warning, Figure 14

Tidal station records of regular ocean tides superimposed by tsunami waves generated by the 15 Nov. 2006 Kuril earthquake (Mw = 8.3). *Left*: Locations of tidal gauges whose data are shown in the right figure. *Right*: Tidal records of the stations shown in the left figure. Time is in UTC. The *thick blue dotted line* denotes the origin time of the earthquake and the blue arrows denote the arrival time of tsunami at each site. The *thin dotted line* denotes the time of tsunami forecast cancellation. *Red stars* mark the time of maximum amplitude at each site. At all sites maximum amplitudes were recorded about 2 to 5 hours after the forecast had been canceled. The largest tsunami amplitude recorded in Japan for this event was 84 cm at Miyake Is. (Tsubota)

the opposite slip angle (λ) direction to the reverse fault with the same values for other fault parameters has almost identical tsunamigenic potential, and the difference is just a polarity of initial tsunami wave. Therefore, normal fault is treated as reverse fault with the same dip angle for this purpose. Strike angles have been fixed to the representative values of the considered regions, or set parallel to the nearby trench axis or coast line. The case with the highest resemblance to the actually calculated CMT solution with respect to its tsunamigenic potential is then retrieved.

The JMA incorporates CMT solutions in the revision/cancellation of tsunami forecast as follows:

(a) In the case of reverse or normal faults, and if the centroid depth is less than the depth determined by P and S arrival times $+30$ km, and Mw 0.5 or more larger than Mjma, the tsunami forecast is then upgraded in accordance with respective database cases.
(b) In the case of strike faults, taking the database into account, and after confirmation by low or not observed tsunami amplitudes at tide gauges, the tsunami forecast is then downgraded or canceled.

The JMA started this incorporation in July, 2007 for the sea area where additional database creation has been completed.

Future Outlook

In tsunami forecasting, trade-off exists between promptness and accuracy/reliance. The JMA's tsunami forecasting strategy, especially for local events, is to satisfy both by using state-of-the-art technologies. They assure promptness of the first issuance of tsunami forecast, based on preliminary seismological results including that of EEW, and its subsequent revision, if required, as soon as more accurate and reliable data such as sea level change, and the results of a more detailed complex data analysis have become available. And as for accuracy/reliability there are two ways to be taken. One is quicken the process of reduc-

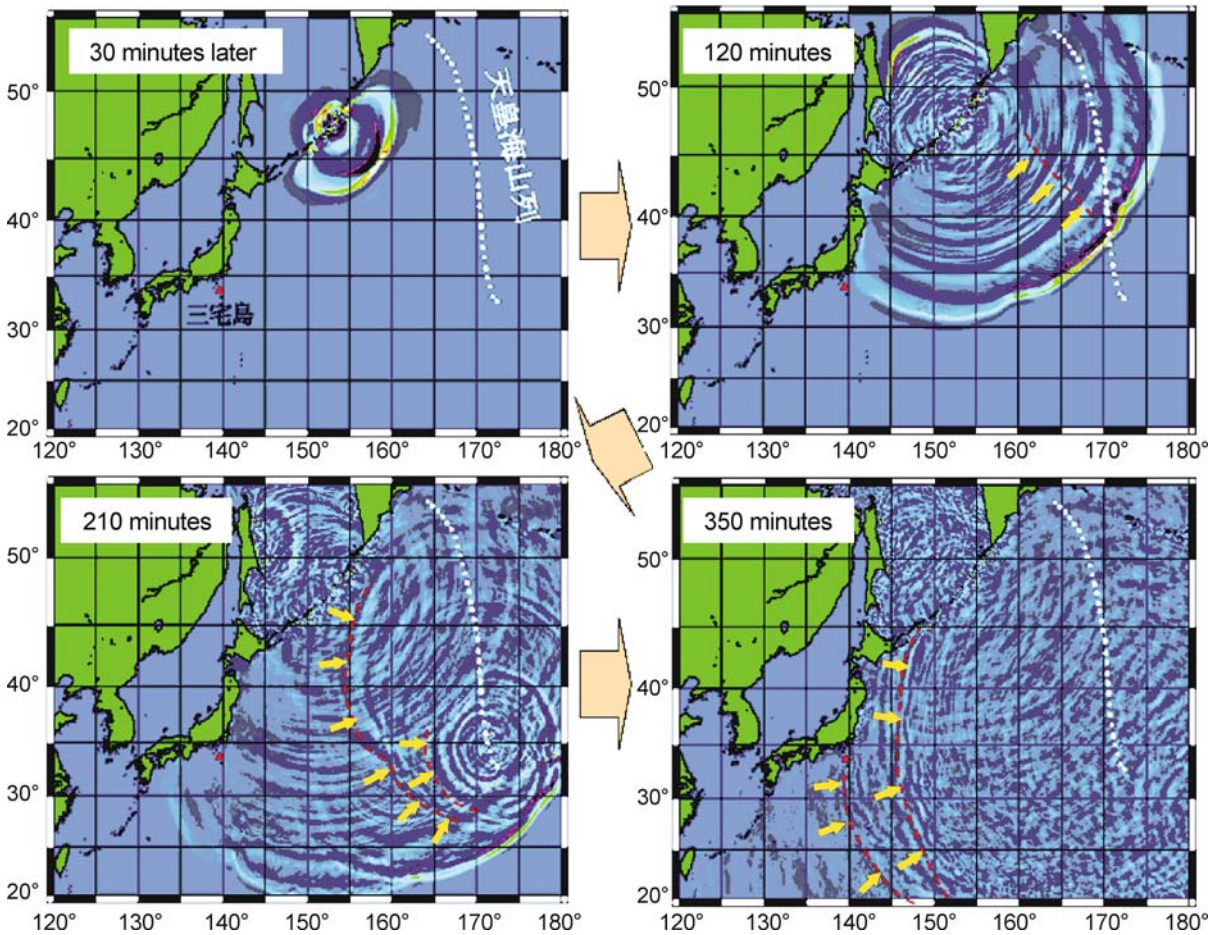

Tsunami Forecasting and Warning, Figure 15

Snap shots of numerical tsunami simulation for the 2006 Nov. 15 Kuril event (Mw8.3). Coriolis force has been considered for this case. Four snap shots (30, 120, 210 and 350 minutes after the earthquake occurrence) are shown. The *white dotted line* marks the position of the Emperor Sea Mount Chain. *Broken read curves with arrows* denote reflected tsunami waves from the Emperor Sea Mount Chain. The estimated arrival times of the reflected waves at the coast of Japan are consistent with the actual tidal records

ing the uncertainty of initial tsunami wave distribution assessments. Quick focal process inversion technique, new magnitude definitions applicable for gigantic or tsunami earthquakes and tsunami source inversion techniques using sea level data (also from satellite altimetry) will become effective for this purpose soon. The other way is very detailed numerical simulation, after reduction of uncertainties in the initial tsunami wave distribution, using fine bathymetry mesh data. Along with the development of integrated calculation algorithms, the improvement of the computer performance might solve this problem in future. But even then, due to the stochastic nature of tsunami behavior near the coast, one has carefully to examine on what statistical quantity of simulated results the tsunami forecast criterion has to be based.

Acknowledgments

We thank Dr. Peter Bormann and Dr. Kenji Satake for reviewing the manuscript, their comments and suggestions greatly improved it.

Bibliography

Primary Literature

1. Abe K (1973) Tsunami and mechanism of great earthquakes. Phys Earth Planet Inter 7:143–153
2. Bormann P, Wylegalla K (2005) Quick estimator of the size of great earthquakes. Eos 86(46):464
3. Bormann P, Baumbach M, Bock G, Grosser H, Choy GL, Boatwright J (2002) Seismic sources and source parameters. In: Bormann P (ed) IASPEI new manual seismological observatory

practice, vol 1, Chap 3. GeoForschungsZentrum Potsdam, Potsdam, pp 1–94
4. Geller RJ (1976) Scaling relations for earthquake source parameters and magnitudes. Bull Seism Soc Am 66:1501–1523
5. Geographical Survey Institute of Japan (2006) Real-time collection and analysis of crustal deformation data, Report on technical development and promotion plan concerning prompt disaster mitigation countermeasures based on disaster information, Chap 2. Ministry of Land, Infrastructure and Transport, Tokyo
6. González FI, Bernard EN, Meinig C, Eble M, Mofjeld HO, Stalin S (2005) The NTHMP tsunameter network. Nat Hazards (Special Issue, US National Tsunami Hazard Mitigation Program) 35(1):25–39
7. Jim Gower J (2005) Jason 1 detects the 26 December 2004 tsunami. EOS Trans Am Geophys Union 86(4):37–38
8. Hara T (2007) Measurement of duration of high-frequency energy radiation and its application to determination of magnitudes of large shallow earthquakes. Earth Planets Space 59:227–231
9. Hatori T (1984) On the damage to houses due to tsunamis. Bull Earthq Res Inst 59:433–439 (in Japanese)
10. Hayashi Y (2008) Extracting the 2004 Indian Ocean tsunami signals from sea surface height data observed by satellite altimetry. J Geophys Res 113:C01001
11. Ide S, Takeo M, Yoshida Y (1996) Source process of the 1995 Kobe earthquake: Determination of spatio-temporal slip distribution by Baysian modeling. Bull Seism Soc Am 87: 547–566
12. Imamura F (1997) IUGG/IOC TIME PROJECT Numerical method of tsunami simulation with the leap-frog scheme, part 3 (Programme lists for near field tsunami), vol 35. IOC Manuals and Guides, Paris
13. Japan Meteorological Agency (2005) A magnitude estimation using borehole volume strainmeters for earthquake events near the coast of Sumatra, Indonesia. Rep Coord Comm Earthq Predict 74:575–577 (in Japanese)
14. Kajiura K(1970) Tsunami source, energy and the directivity of wave radiation. Bull Earthq Res Inst (Univ. of Tokyo) 48: 835–869
15. Kamigaichi O (2004) JMA earthquake early warning. J Japan Assoc Earthq Eng (Special Issue) 4:134–137
16. Kanamori H, Anderson DL (1975) Theoretical basis of some empirical relations in seismology. Bull Seism Soc Am 65: 1073–1095
17. Kanamori H, Rivera L (2007) Speeding up seismic tsunami warning using *W* phase. In: Abstracts of AGU Fall Meeting 2007, S43C-06
18. Kasahara M, Sasatani T (1986) Body wave analyses of strain seismograms observed at Erimo, Hokkaido, Japan. J Fac Sc Hokkaido Univ Ser. VII (Geophysics) 8:83–108
19. Katsumata A (2004) Revision of the JMA displacement magnitude. Q J Seismol 67:1–10 (in Japanese)
20. Kikuchi M, Kanamori H (1991) Inversion of complex body waves, III. Bull Seism Soc Am 81:2335–2350
21. Lomax A, Michelini A, Piatanesi A (2007) An energy-duration procedure for rapid determination of earthquake magnitude and tsunamigenic potential. Geophys J Int 170:1195–1209
22. Matsu'ura M, Hasegawa Y (1987) A maximum likelihood approach to nonlinear inversion under constraints. Phys Earth Planet Inter 47:179–187
23. Okada M, Tanioka Y (1998) Relation of tsunami generation ratio with earthquake magnitude and hypocentral depth. Mon Kaiyo (Special issue) 15:18–22 (in Japanese)
24. Okada Y (1985) Surface deformation due to shear and tensile faults in a half-space. Bull Seism Soc Am 75:1135–1154
25. Ozawa S (1996) Geodetic inversion for the fault model of the 1994 Shikotan Earthquake. Geophys Res Lett 23(16): 2009–2012
26. Satake K (1989) Inversion of tsunami waveforms for the estimation of heterogeneous fault motion of large submarine earthquakes – the 1968 Tokachi-Oki and 1983 Japan Sea earthquakes. J Geophys Res 94:5627–5636
27. Satake K (1995) Linear and nonlinear computations of the 1992 Nicaragua earthquake tsunami. PAGEOPH 144:455–470
28. Shuto N (1991) Historical changes in characteristics of tsunami disasters. In: Proc of international symposium on natural disaster reduction and civil engineering. Japan Society of Civil Engineering, Tokyo, pp 77–86
29. Shuto N (1992) Tsunami Intensity and damage, Tsunami engineering technical report. Tohoku Univ 9:101–136 (in Japanese)
30. Shuto N (1998) Present state of tsunami research and defense works. Bull Coastal Oceanogr 35(2):147–157 (in Japanese)
31. Shuto N et al (1986) A study of numerical techniques on the tsunami propagation and run-up. Sci Tsunami Hazard 4: 111–124
32. Takanami T, Kitagawa G (eds) (2002) Methods and application of signal processing in seismic network operations. Lecture Notes in Earth Science vol 98. Springer, Berlin
33. Titov VV, González FI, Bernard EN, Eble MC, Mofjeld HO, Newman JC, Venturato AJ (2005) Real-time tsunami forecasting: Challenges and solutions. Nat Hazards (Special Issue, US National Tsunami Hazard Mitigation Program) 35(1):41–58
34. Tsuboi S, Abe K, Takano K, Yamanaka Y (1995) Rapid determination of Mw from broadband *P* waveforms. Bull Seism Soc Am 83:606–613
35. Tsushima H, Hino R, Fujimoto H, Tanioka Y (2007) Application of cabled offshore ocean bottom tsunami gauge data for real-time tsunami forecasting. In: Proc symposium on underwater technology 2007/Workshop on scientific use of submarine cables and related technologies 2007. The University of Tokyo, Tokyo, pp 631–639
36. Ueno H, Hatakeyama S, Aketagawa T, Funasaki J, Hamada N (2002) Improvement of hypocenter determination procedures in the Japan Meteorological Agency. Q J Seism 65:123–134 (in Japanese)
37. Utsu T, Shima E, Yoshii T, Yamashina K (2001) Encyclopedia of Earthquakes, 2nd edn. Asakura, Tokyo, pp 657
38. Weinstein S, Okal E (2005) The mantle magnitude Mm and the slowness parameter theta: Five years of real-time use in the context of tsunami warning. Bull Seosm Soc Am 85: 779–799
39. Wells DL, Coppersmith KJ (1994) New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement. Bull Seism Soc Am 84(4): 974–1002
40. Yokota T, Zhou S, Mizoue M, Nakamura I (1981) An automatic measurement of arrival time of seismic waves and its application to an on-line processing system. Bull Earthq Res Inst 56:449–484 (in Japanese)

Books and Reviews

Bormann P (ed) (2002) IASPEI new manual of seismological observatory practice, vol 1 and 2. GeoForschungsZentrum Potsdam, Potsdam

Satake K (2007) Tsunamis, chap 4, 17. Treatise on geophysics, vol 4. Elsevier, Amsterdam, pp 483–511

Tsunami Inundation, Modeling of

PATRICK J. LYNETT
Texas A&M University, College Station, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Brief Review of Tsunami Generation and Open Ocean Propagation
Physics of Nearshore Tsunami Evolution
Effects of Bathymetric and Topographical Features on Inundation
Hydrodynamic Modeling of Tsunami Evolution
Moving Shoreline Algorithms
Future Directions
Bibliography

Glossary

Beach profile A cross-shore, or normal to the beach, survey of the seafloor and dry ground elevation (bathymetry and topography); a series of spatial location and bottom elevation data pairs.

Bore A steep hydraulic front which transitions between areas of different water level. Tsunamis can approach land as a turbulent, breaking bore if the incident tsunami is of sufficiently large height.

Boussinesq equations An approximate equation model, used for waves with wave length of at least two times the local water depth; a long-wave-based model, but includes some frequency dispersion

Dispersion, amplitude The separation of wave components due to a wave-height related difference in wave speed; all else being equal, a wave with a large height will travel faster than one with a small height.

Dispersion, frequency The separation of wave components due to a frequency related difference in wave speed; all else being equal, a wave with a longer period will travel faster than one with a short period.

Navier–Stokes equations The full equations of fluid motion, including dissipation through the fluid molecular viscosity only. Other models discussed here, namely the Shallow Water Wave and Boussinesq equations, are approximations to these equations.

Runup, or runup height The ground elevation (a vertical measure) at the furthest point of landward inundation.

Shallow water wave equations An approximate equation model, used for waves with wave length many times larger than the water depth; a non-dispersive, long-wave model; there is no frequency dispersion in this model.

Tsunami inundation The spatial area flooded as a tsunami rushes inland.

Definition of the Subject

Tsunami inundation is the one of the final stages of tsunami evolution, when the wave encroaches upon and floods dry land. It is during this stage that a tsunami takes the vast majority of its victims. Depending on the properties of the tsunami (e. g. wave height and period) and the beach profile (e. g. beach slope, roughness), the tsunami may approach as a relatively calm, gradual rise of the ocean surface or as an extremely turbulent and powerful bore – a wall of white water. The characteristics of this approach determine the magnitude and type of damage to coastal infrastructure and, more importantly, the actions required of coastal residents to find a safe retreat or shelter.

To gage the nearshore impact of tsunami inundation, engineers and scientists rely primarily on three different methods: 1) Field survey of past events, 2) Physical experimentation in a laboratory, and 3) Numerical modeling. It is the last of these methods – numerical simulation of tsunami inundation – that will be the focus of this article. With numerical simulation, it is possible to predict the consequence of future tsunamis on existing coastal towns and cities. This information allows for the establishment of optimum evacuation routes, identification of high-risk and/or unprepared areas, re-assessment of building codes to withstand wave impact, and placement of tsunami shelters, for example. It is the hope that, through accurate prediction of tsunami effects, in conjunction with policy makers willing to implement recommended changes and a strong public education program, communities will show resiliency to tsunami impact, with minimal loss of life and damage to critical infrastructure.

Introduction

On December 26, 2004, the boundary between the Indo–Australian and Eurasian plates off the coast of north-

ern Sumatra ruptured in a great (Mw 9.3) earthquake at 00:58:53 universal time (U.T.). Up to 15m of thrust on the plate interface [31] displaced tens of cubic kilometers of seawater and propagated a tsunami across the Indian Ocean. The earthquake was widely felt throughout South Asia and was locally destructive in Sumatra and the Andaman and Nicobar islands, but it was the tsunami that caused widespread damage to densely populated coastal communities both nearby and thousands of kilometers away.

Due to the extensive damage left behind by large tsunamis such as the Indian Ocean tsunami, it is difficult if not impossible to put together a complete picture of the event with field observations alone. Additionally, for some parts of the world that have not seen a tsunami in recent times, there are no field observations on which to develop safety procedures and protect residences from future tsunamis. It is for these purposes – understanding the detail of tsunami inundation and to estimate tsunami hazard – that we must rely on modeling of tsunamis. There are two primary modeling approaches - physical and numerical. The physical, or experimental, approach uses scaled-down models to look at a particular aspect of a phenomenon. While this approach is integral to the fundamental understanding of waves, because of the huge wavelengths of tsunamis, experiments are limited. For example, a tsunami might have a wavelength of 100 km in a deep ocean depth of 1 km, with a wave height of 1 m. Note that the above values represent approximate order of magnitudes for a large subduction zone tsunami. Now, to scale this down for the laboratory with a wave tank depth of 1 m – the tank would have to be 100 m long and the created lab-tsunami would have a hardly measurable height of 1 mm. Numerical modeling, while not "real" in the sense that modeling is done on a computer chip with approximated equations of motion rather than in the laboratory, does not suffer from this scaling problem, and can generally accommodate any type of arbitrary wave and ocean depth profile.

Numerical simulations of tsunami propagation have been greatly improved in the last 30 years. In the United States, several computational models are being used in the National Tsunami Hazard Mitigation Program, sponsored by the National Oceanic and Atmospheric Administration (NOAA), to produce tsunami inundation maps and predict tsunami runup in real-time for the warning system. In addition, there are numerous other models used by researchers and engineering companies in an attempt to better understand tsunami impact. In this article, an overview of these models, as well as how they are validated and utilized, is provided.

Brief Review of Tsunami Generation and Open Ocean Propagation

Before introducing the physics behind propagating a tsunami across oceans and overland, we must first discuss how a tsunami is created. For earthquake generated mega-tsunamis, such as the Indian Ocean event, a huge undersea earthquake along a great fault length of a subduction zone must occur. These earthquakes create large vertical motions of the seafloor. This vertical motion of the seafloor pushes the water above it, essentially creating a small displacement of water above the earthquake. This displacement of water will immediately try to spread out and reach a gravitational equilibrium, and it does so as waves propagating away from the earthquake zone – this is the tsunami.

To represent the tsunami in numerical models, we use an initial condition. Simply put, there is placed some irregular ocean surface profile at the instant after the earthquake, when the numerical simulations will start. Then, based on physics – Newton's Laws written for fluid – the initial condition evolves and transits oceans. As a tsunami travels unhindered across ocean basins, it does so quickly and with little noticeable change. In the deep ocean, even the largest tsunamis have heights only near 1 m and currents of 10 cm/s, and are not likely to be identified by ships or surface buoys in the presence of wind waves.

Physics of Nearshore Tsunami Evolution

A tsunami in the deep ocean is long and travels extremely fast. As the wave reaches shallow water, near the coastline, the tsunami begins the shoaling process. The speed at which long wave such as a tsunami moves, or celerity, is a function of the local water depth. The less the depth, the slower the wave moves. A tsunami, with its very long length, experiences different water depths at any given instant as it travels up a slope; the depth at the front of the wave, the portion of the tsunami closest to the shoreline, will generally be in the shallowest water and thus is moving the slowest. The back of the tsunami, on the other hand, will be in deeper water and will be moving faster than the front. This leads to a situation where the back part of the wave is moving faster than the front, causing the wavelength to shorten. With a shortening tsunami, the wave energy is in essence squeezed into a smaller region, forcing the height to grow. It is for this reason that, despite having a height of only a meter in the deep ocean, the tsunami elevation over land can easily exceed 10 m. With this great increase in wave height comes a more dynamic and complex phenomenon.

Presented in a more technical manner, a tsunami in the open ocean is generally a linear, non-dispersive wave. First, what is meant by a non-dispersive tsunami will be discussed. Also, the discussion here will be in terms of a large earthquake generation tsunami, such as the 2004 Indian Ocean event. Other impulsive waves, such as landslide or asteroid impact generated waves, are more difficult to generalize and will be introduced separately at the end of this section.

Any wave condition, whether it is a tsunami or a typical wind wave in the ocean, can be mathematically described as a superposition, or summation, of a series of separate sine (or cosine) waves, each with independent amplitude and speed. For example, with the right choice of individual sine waves, it is possible to construct even the idealized tsunami: a single soliton. If a wave is considered a dispersive wave, then the various sine wave components will have different wave speeds, and the wave will disperse as the faster moving components move away from the slower ones. If a wave is non-dispersive, then all the components move at the same speed, and there is no lengthwise dispersal, or spreading, of the tsunami wave energy. It is for this reason that tsunamis can be devastating across such a large spatial region; the tsunami wave energy will not disperse but will remain in a focused pulse.

The dispersion described above is generally what scientists are referring to during a discussion of dispersive vs non-dispersive waves. However, it is more precisely called "frequency dispersion" as it is dominantly dependent on the period of the component. There is another type of dispersion, called "amplitude dispersion." This second type of dispersion is a function of the nonlinearity of the wave, and is usually discussed under the framework of linear vs nonlinear waves. For tsunamis, the nonlinearity of the waves is given by the ratio of the tsunami height to the water depth. When this ratio is small, such as in the open ocean, the wave is linear; on the other hand, in shallow water the ratio is order unity and thus the wave is no longer linear. The linear/nonlinear nomenclature is not an intuitive physical description of the waves, but comes from the equations describing the tsunami motion, described later in this section. When this nonlinear effect is taken into account, it is found that the wave speed is no longer just a function of the local depth, but of the wave height as well. More specifically, looking at two components of the same period but with different amplitudes, the component with the larger amplitude will have a slightly larger wave speed. Except for the interesting cases of wave fission, discussed later in this section, the nonlinear effect of amplitude dispersion does not spread tsunami energy with an end result of lessening nearshore impact; in fact it will act to focus wave energy at the front, often leading to a powerful breaking bore.

Thus, open ocean propagation of a conventional tsunami is a relatively uncomplicated process which translates wave energy across basins, subject to wave speed changes that are a function of the local depth. As a tsunami enters the nearshore region, roughly characterized by water depths of 100m and less, the wave can undergo a major physical transformation. The properties of this transformation depend heavily on the characteristics of the beach profile and the wave itself. In the simplest inundation case, the beach profile is relatively steep (footnote: here "steep" should be thought of in terms of the tsunami wavelength. If the horizontal distance along the slope connecting deep water to the shoreline is small compared to the tsunami wavelength, the beach would be considered steep) and the tsunami wave height is small, then the runup process closely resembles that of a wave hitting a vertical wall, and the runup height will be approximately twice the offshore tsunami height. In these special cases, a breaking bore front would not be expected; in fact horizontal fluid velocities near the shoreline would be very small. Here, the tsunami inundation would closely resemble that of a quickly rising tide with only very minor turbulent, dynamic impacts. However, even in these cases, overland flow constrictions and other features can create localized energetic inundation.

If the beach profile slope is mild, typical of continental margins, and/or the tsunami wave height is large, then the shallow water evolution process becomes highly nonlinear. However, while the nonlinear effect becomes very important, in the large majority if cases, frequency dispersion is still very small and can be neglected. Nearshore nonlinear evolution is characterized by a strong steepening, and possible breaking, of the wave front with associated large horizontal velocities. In these cases, turbulent dissipation can play a major role.

While it may be intuitive to postulate that wave breaking dissipation at the tsunami front plays a significant role in the tsunami inundation, this may not be altogether correct. This breaking dissipation, while extraordinarily intense, is fairly localized at the front which, both spatial and temporal, often represents only a small fraction of the tsunami. So, for tsunamis such as the 2004 Indian Ocean event, the related dissipation likely had only minor impact on leading-importance quantities such as the maximum runup and inland (off-beachfront) flow velocities. However, the properties of breaking are of great importance to other aspects of tsunami inundation. The maximum forces on beachfront infrastructure, such as ports, terminals, piers, boardwalks, and houses, should include

the bore impact force as well as the drag force associated with the following quasi-steady flow [54,68]. If one was interested in understanding how bottom sediments are suspended, transported, and deposited by a tsunami, the bore turbulence again may play an important role. Thus, understanding the dynamics of a breaking tsunami front is not of particular importance for near real-time or operational tsunami forecast models. This information is of great use for engineers and planners, who can utilize it to design tsunami-resistant structures, for example.

A second energy dissipation mechanism, one that does play a major role in determining maximum runup, is bottom friction. On a fundamental level, this dissipation is caused by the flow interaction with the bottom, where bottom irregularities lead to flow separations and the resulting turbulence. All natural bottoms result in some bottom friction; a smooth, sandy beach may generate only minor dissipation, while a coral reef or a mangrove forest can play a huge role in reducing tsunami energy [12]. Such features will be discussed in additional detail in the next section. Other means of energy dissipation will be largely local, and may include enhanced mixing due to sediment or debris entrainment, large shallow-flow vortex generation by headlands or other natural or artificial obstacles and the resulting dissipation, and flow through/around buildings and other infrastructure, sometimes termed macro-roughness and grouped with bottom friction.

Up to this point, we have only discussed the "typical" nearshore tsunami evolution which is portrayed as a wave without frequency dispersion, and may be called a linear or nonlinear tsunami, depending on a number of physical properties. The rest of this section will be devoted to those situations where the above characterization may no longer be adequate. Looking first to the tsunami source, waves that are generated by underwater landslides, underwater explosions, or asteroid impacts will often not behave as non-dispersive waves in the open ocean [44,72]. These source regions tend to be at least an order of magnitude smaller in spatial extent compared to their subduction zone counterparts. Physically, this implies that the generated waves will be of shorter wavelength. As a rule of thumb, if these generated waves have length scales of less than 10 times the local depth, then it should be anticipated that frequency dispersion will play a role [42]. Under this constraint, individual component wave speeds near the dominant period become frequency dependent.

Understanding that an impulsively generated wave can be dispersive has serious implications. Take, for example, a hypothetical landslide located in the Atlantic Ocean which generates a dispersive tsunami (e. g. Ward and Day, 2001). As this tsunami travels across the Atlantic, to either the USA east coast or the European west coast, frequency dispersion effects will spread the wave energy in the direction of propagation. This will convert the initial short-period pulse into a long train of waves. By spreading this energy out, the inundation impact will be greatly reduced. First, by taking a high-density energy pulse and stretching it into a longer, lower-density train, the maximum energy flux, and thus intensity, hitting the shoreline will decrease. Second, by increasing the duration of the time series, and creating many individual crests, energy dissipation can play a bigger role. Using simple energy arguments, it can be shown that, comparing a high-density, short-period pulse to a low-density, long-period train, more energy will be removed through bottom friction and breaking. This increase will be related to the ratio of the period of the entire dispersive wave train to the period of the pulse. Numerically studies have shown that for such cases, the individual wave crests are largely dissipated, and runup is dominated by the carrier wave, or in other terms it becomes a time-dependent, wave setup problem [26].

While a topic of current research in the tsunami community, frequency dispersion may occasionally play a non-negligible role in even the long wavelength, subduction zone tsunamis. To date, there have been two categories of argument that dispersion is important for these tsunamis: 1) short-period energy generated at the source is significant and leads to different patterns of runup if included (e. g. [21,28]), and 2) shallow-water nonlinear interactions can generate short-period components which can become decoupled (or un-locked) from the primary wave, and will change the incident tsunami properties [47].

Thinking of an arbitrary and complex initial free surface displacement generated by an undersea earthquake, there does exist the possibility that dispersive wave energy can be initially generated here. This irregular wave condition can be constructed as a continuous wave energy spectrum, and by definition there will be finite (albeit small) energy at all frequencies. The obvious question in this case is: what length scale characteristics of the initial free surface displacement, or the preceding earthquake, will lead to a significant measure of dispersive wave energy? To provide an answer to this question, a simple order-of-magnitude scaling argument is presented here; see Hammack and Segur [19], for example, for a mathematically rigorous attempt at insight. Let us define a characteristic change in vertical free surface elevation, $\Delta\eta$, and a horizontal length scale across which this vertical change occurs, ΔL. Reducing to the simplest case, a regular wave with wave height equal to $\Delta\eta$, then the wavelength would be $2\Delta L$. Follow-

ing this analogy, which will hold in a proportional sense for a Fourier series of wave components, for any $\Delta\eta$ measured along a tsunami initial condition, there exists a wave component with wave length equal to $2\Delta L$. For that component to be significant to the tsunami evolution, the local vertical change, $\Delta\eta$, must be some non-negligible fraction of the maximum tsunami height, H. Note that the H discussed here is a global property of the entire initial tsunami wave condition. For the individual wave component, with length $2\Delta L$, to be dispersive, its length should be less than roughly 10 times the local water depth, h_0. Additionally, the difference between dispersive wave propagation and a non-dispersive propagation is cumulative. For example, if the full linear wave theory predicts, or a specific wave component, a wave speed that is 10% less than the long wave speed, then the predicted arrival time difference will grow by 0.1 times the period for each wavelength of propagation. Thus, the impact of dispersion is related to the distance of propagation, and is proportional to D/λ, where D is the total distance traveled by the wave, and λ is the average wavelength of the wave across D which can be expected to be proportional to $2\Delta L$. Assuming that λ is approximately equal to $2\Delta L$, it can be said that in order for frequency dispersion effects to play a role in tsunami evolution,

$$\max\left|\frac{h_0\Delta\eta}{\Delta L^2}\right|\frac{D}{H} > \delta\,, \tag{1}$$

where δ is some minimum threshold for importance. What this value should be is an open question, although it is likely to be near 0.1. From this term, it is clear the impact of dispersion is a function of a number of the properties of the initial tsunami condition, and should be taken into consideration when creating tsunami initial conditions. For example, use of discontinuous block-type segments (e. g. [30]), with sharp edges (very small ΔL) may lead to the conclusion that frequency dispersion is important, while it could be a direct result of a coarsely approximated initial tsunami condition. Also note that this exercise does not include the effects of radial spreading, which could very likely be important for small-scale irregularities in the initial condition. Wave height decrease by radial spreading is proportional the horizontal curvature of the initial condition and to $(\lambda/D)^n$, and decreases faster for dispersive waves ($n \sim 1$) as compared to non-dispersive waves ($n = 0.5$) (e. g. [72]). Thus, for this case of significant radial spreading, it would be very difficult for source-based dispersion effects to play a meaningful role in the far field.

Under certain conditions, namely a nonlinear tsunami propagating across a wide shallow shelf, a process called fission may occur. Wave fission is a separation process where wave energy, initially part of a primary wave or pulse, attains certain properties, such as higher or lower phase speed, that allow it to disconnect from the primary wave and propagate as an independent wave. In the context of nearshore tsunami evolution, there is a standard mechanism which is the cause of this fission. First, it is necessary to describe what a nonlinear, phase-locked wave is. To do this, we will examine the acceleration terms of the 1D conservation of momentum equation for the velocity component u:

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = -\frac{\partial p}{\partial x} + \mu\frac{\partial^2 u}{\partial x^2}\,. \tag{2}$$

Now assume that there is a single wave component, under which the velocity oscillates as

$$\cos(kx - \omega t) = \cos\theta\,, \tag{3}$$

where k is the wavenumber, ω the frequency, and the speed of the wave is given by ω/k. If the wave is nonlinear, which is to say that the convective acceleration term in the above momentum equation is not negligible, the convective term will include the product of

$$\cos\theta * \cos\theta = \cos 2\theta = \cos(2kx - 2\omega t)\,. \tag{4}$$

Thus, through this nonlinear term, a new wave component, with twice the wavenumber and frequency (or half the wavelength and period) has been generated. From linear wave theory, it is expected that this new wave, with a shorter period, will have a different wave speed than the original, primary wave. However, from the phase function of this new wave, there is the speed $2\omega/2k = \omega/k$, which is the identical speed of the primary wave. Thus this new wave is *locked* to the *phase* of the original wave. This connection can be rather delicate, and any disruptions to the primary wave, such as a varying seafloor, dissipation, or interactions with other free waves in the train or wave pulse, can cause the new waves to become unlocked. When this occurs, the now free waves retain their frequency 2ω, but take a wavenumber as given by linear wave theory. Since these freed waves will be of a shorter period than the primary wave, they will travel at a slower speed and generally trail the main wave front.

Long wave fission is most commonly discussed in the literature via a solitary wave propagating over an abrupt change in depth, such as a step (e. g. [17,24,33,37,45,55]). In these cases, there is a deep water segment of the seafloor profile, where a solitary wave initially exists. In this depth, the solitary wave is of permanent form. As the solitary wave passes over the change in depth, into shallower water,

the leading wave energy will try to re-discover a balance between nonlinearity and dispersion; the solitary wave. Since this new solitary wave will be a different shape and contain a lower level of mass, by conservation there must be some trailing disturbance to account for the deficient. This trailing disturbance will take the form of a rank-ordered train of solitons. Figure 1 depicts this process. The solitons in the trailing train, while smaller in height than the leading solitary wave, tend to have a similar wavelength; this has been shown both analytically, numerically, and experimentally. Note, however, that discussion of fission in this sense is not particularly relevant to "real" tsunami modeling, where the offshore wave approaching the shelf break rarely resembles a solitary wave solution [62]. However, the offshore wave does not need to specifically be a solitary wave for this process to occur.

In numerous eyewitness accounts and videos recorded of the 2004 Indian Ocean tsunami, there is evidence of the tsunami approaching the coastline as a series of short period (on the order of 1 min and less) breaking fronts, or bores (e. g. [30]). These short period waves may be the result of fission processes of a steep tsunami front propagating across a wide shelf of shallow depth. Along the steep front of a very long period wave, nonlinearity will be very important. There will be a large amount of energy in high-frequency components with wavelengths similar the horizontal length of the tsunami front (on the order of 1 km). As the wave continues to shoal, the high-frequency locked waves may eventually become free waves, and will take the form of very short waves "riding" the main wave pulse. This situation is akin to an undular bore in a moving reference frame. This process is, in fact, identical to that described in the above paragraph, it simply takes place over a much longer distance. The newly freed waves, in the nonlinear and shallow environment, will attempt to reach an equilibrium state, where frequency dispersion and nonlinearity are balanced. Thus, the fission waves will appear as solitary waves, or more generally, cnoidal waves. This fact provides some guidance as to the wavelength of these fission waves; they can be approximately calculated via solitary wave theory using the tsunami height and depth of the shelf. For example, on a shelf with depth of 30 m and an incident tsunami height of 5 m, fission waves with a wavelength of approximately 240 m and period of 13 s would be generated. In recent work looking at tsunamis along the eastern USA coast, where there exists a wide shallow shelf, this fission process has been investigated [14]. Figure 2 gives a few numerical simulation snapshots, and shows where the fission occurs, and the eventually impact on the waveform. This simulation, run with the dispersive equations, generated fission waves with lengths in the range of 100–200 m, and required a grid size of 5 m to attain numerically convergent results. In this example, the steep fission waves break offshore, and have little impact on the maximum runup. A conclusion of this fission issue is that, if one attempts to simulate tsunami propagation with dispersive equations, and if the grid is not chosen to be fine enough to resolve the short fission waves,

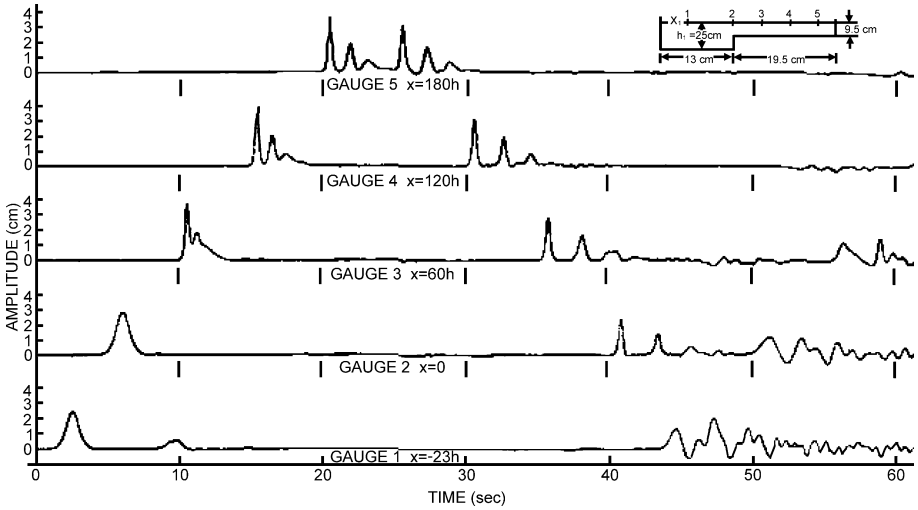

Tsunami Inundation, Modeling of, Figure 1

Example of experimental data looking at solitary wave fission by propagation onto a shelf from Goring and Raichlen [17]. Note that the flume layout and measurement location is given up in the *upper right*. The initial solitary wave undergoes the fission process and results in three distinct solitary waves

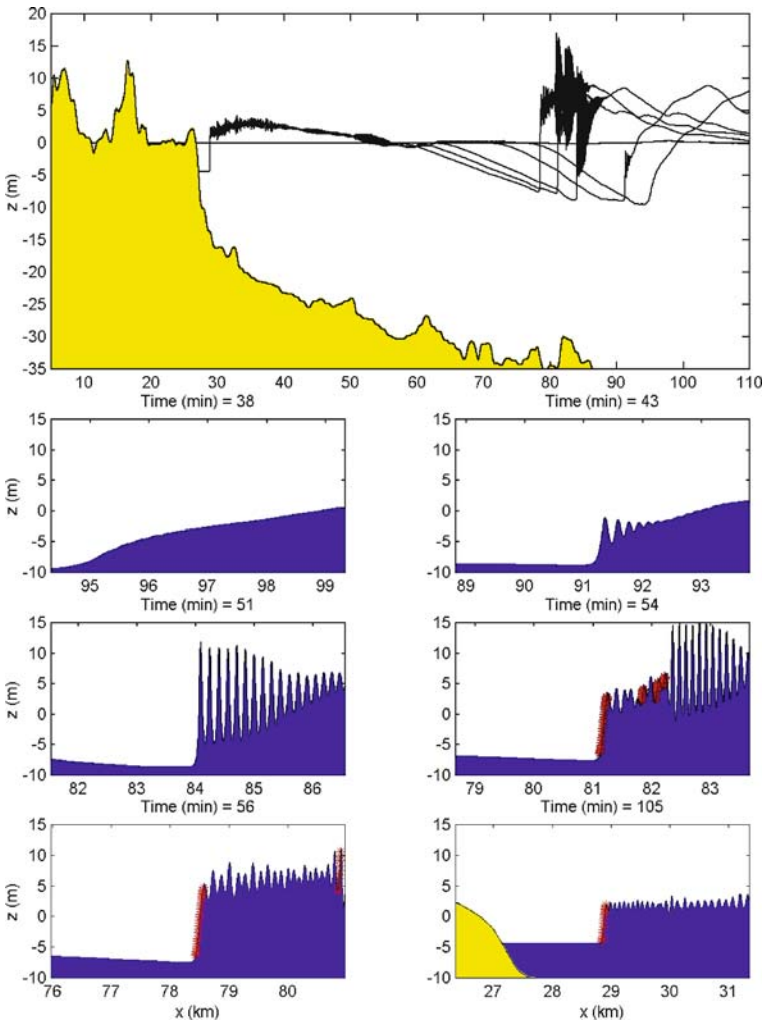

Tsunami Inundation, Modeling of, Figure 2

Example of tsunami fission. Simulation results are from Geist et al. [14] for a landslide-generated tsunami off the east coast of the USA. The *top plot* shows the beach profile and six free surface profiles at different times. The *lower subplots* are zoom-in's of those six profiles, with the times given in the individual plot titles. The *red marks* visible in the *lowest plots* indicate regions where the wave is breaking

the justification to use the dispersive model is greatly degraded.

Effects of Bathymetric and Topographical Features on Inundation

It is well established that large-scale coastal features, such as small islands, large shoals, canyons, and shelves, can play am important role in tsunami inundation due to conventional shallow water effects such as shoaling and refraction (e. g. [4,5,16,36,70]). On the other hand, understanding of the impact of smaller scale features is just now being developed. This work was largely initiated by field observations. Synolakis et al. [60], surveying the coast of Nicaragua for information about the 1992 tsunami in the region, noted that the highest levels of damage along a particular stretch of beach were located directly landward of a reef opening used for boat traffic. It was postulated that the reef gap acted as a lower resistance conduit for tsunami energy, behaving like a funnel and focusing the tsunami. Along neighboring beaches with intact reefs, the tsunami did not have the intensity to remove even beach umbrellas. Investigating impacts from the same tsunami, Borrero et al. [3], discussed how small scale bathymetry variations affected coastal inundation. One of the conclusions of this work was that bathymetry features with

Tsunami Inundation, Modeling of, Figure 3
Example of foundation scour. This image was taken by the International Tsunami Survey Team to Sri Lanka in January 2005

Tsunami Inundation, Modeling of, Figure 4
Example of damage to the backside of a coastal residence. These images were taken by the International Tsunami Survey Team to Sri Lanka in January 2005. The *top photograph* shows the front side of the structure, facing the ocean; there is damage but the main structure is intact. The *lower photo* shows the backside of the same building, showing the walls blown out, away from the center of the structure

length scales 50 m and less had leading order impact on the runup. Looking to the recent Indian Ocean tsunami, a survey team in Sri Lanka inferred from observations that reef and dune breaks lead to locally increased tsunami impact [38,39]. Also in Sri Lanka, Fernando et al. [12] performed a more thorough survey along the southeastern coastline, and concluded that there was a compelling correlation between coral mining and locally severe tsunami damage. While additional research is needed to quantify the effects of small scale features, the observations hint that defense measures such as seawalls, once thought to be inconsequential to tsunami inundation, may provide some protection.

Onshore, tsunami propagation is effected by the general topography (ground slope), ground roughness, and obstacles (e. g. [41,59,61,67]). The composition of the ground, be it sand, grass, mangroves, or pavement, controls the roughness and the subsequent bottom friction damping. To predict tsunami inundation with high confidence, the ground type must be well mapped and the hydrodynamic interaction with that type must be well understood. If the tsunami approaches the shoreline as a bore, the process of "bore collapse", or the conversion of potential to kinetic energy, will cause the fluid to rapidly accelerate [56,69]. This fast flow equates to high fluid forces on obstacles such as buildings. Tsunami interaction with these obstacles can lead to a highly variable local flow pattern (e. g. [10,67]). As the flow accelerates around the corners of a building, for example, the scour potential of that flow increases greatly, and foundation undermining is a concern (e. g. see Fig. 3). As with any fluid flow past an obstacle, the backface of the obstacle is characterized by a low-pressure wake. Combined with the interior flooding of a building, this low pressure wake may lead to an outward "pull" force on the back wall, causing it to fail by falling away from the center of the building. Such failures were observed during field surveys of the 2004 event, as shown in Fig. 4. Increasing the topographical complexity, in built coastal environments, structures are located within

close enough proximity to each other such that their disturbances to the flow may interact. This can lead to irregular and unexpected loadings, where for example a 2nd row building experiences a larger force than beach front buildings due to a funneling effect. These types of interactions are very poorly understood, and require additional research.

Hydrodynamic Modeling of Tsunami Evolution

Numerical simulations of tsunami propagation have made great progress in the last thirty years. Several tsunami computational models are currently used in the National Tsunami Hazard Mitigation Program, sponsored by the National Oceanic and Atmospheric Administration, to produce tsunami inundation and evacuation maps for the states of Alaska, California, Hawaii, Oregon, and Washington. The computational models include MOST (Method Of Splitting Tsunami), developed originally by researchers at the University of Southern California [66]; COMCOT (Cornell Multi-grid Coupled Tsunami Model), developed at Cornell University [35]; and TUNAMI-N2, developed at Tohoku University in Japan [29]. All three models solve the same depth-integrated and 2D horizontal (2DH) nonlinear shallow-water (NSW) equations with different finite-difference algorithms. There are a number of other tsunami models as well, including the finite element model ADCIRC (ADvanced CIRCulation Model For Oceanic, Coastal And Estuarine Waters; e. g., [53]). For a given source region condition, existing models can simulate propagation of a tsunami over a long distance with sufficient accuracy, provided that accurate bathymetry data exist.

The shallow-water equation models commonly lack the capability of simulating dispersive waves, which, however, could well be the dominating features in landslide-generated tsunamis and for the fission processes described previously. Several high-order depth-integrated wave hydrodynamics models (Boussinesq models) are now available for simulating nonlinear and weakly dispersive waves, such as COULWAVE (Cornell University Long and Intermediate Wave Modeling Package; [42]) and FUNWAVE [25]. The major difference between the two is their treatment of moving shoreline boundaries. Lynett et al. [44] applied COULWAVE to the 1998 PNG tsunami with the landslide source; the results agreed with field survey data well. Recently, several finite element models have also been developed based on Boussinesq-type equations (e. g., [73]). Boussinesq models require higher spatial and temporal resolutions, and therefore are more computationally intensive. Moreover, most of model validation work was performed for open-ocean or open-coast problems. In other words, the models have not been carefully tested for wave propagation and oscillations in semi-enclosed regions – such as a harbor or bay – especially under resonant conditions.

Being depth-integrated and horizontally 2D, NSW and Boussinesq models lack the capability of simulating the details of many coastal effects, such as wave overturning and the interaction between tsunamis and coastal structures, which could be either stationary or movable. At present, stationary coastal structures are parametrized as bottom roughness and contribute to frictional effects in these 2DH models. Although by adjusting the roughness and friction parameter satisfactory results can be achieved for maximum runup and delineation of the inundation zone (e. g., [35]), these models cannot provide adequate information for wave forces acting on coastal structures.

As a tsunami propagates into the nearshore region, the wave front undergoes a nonlinear transformation while it steepens through shoaling. It is in this nearshore region that dissipative effects can be important. Bottom friction can play a major role in the maximum runup and area of inundation (e. g. [67]). In depth-integrated models, bottom friction is typically approximated through a quadratic (drag) friction term, where the friction factor is calculated often through a Manning's coefficient or a Darcy–Wiesbach type friction factor (e. g. [25,35]). The validity of these steady-flow based coefficients has yet to be rigorously validated for use with tsunamis. If the tsunami is large enough, it can break at some offshore depth and approach land as a bore – the white wall of water commonly referenced by survivors of the Indian Ocean tsunami. Wave breaking in traditional NSW tsunami models has not been handled in a satisfactory manner. Numerical dissipation is commonly used to mimic breaking (e. g. [36]), and thus results become grid dependant. In Boussinesq models, this breaking is still handled in an approximate manner due to the fact that the depth-integrated derivation does not allow for an overturning wave; however these breaking schemes have been validated for a wide range of nearshore conditions (e. g. [40]).

Being depth-integrated, NSW and Boussinesq models lack the capability of simulating the vertical details of many coastal effects, such as strong wave breaking/overturning and the interaction between tsunamis and irregularly shaped coastal structures. To address this deficiency, several 2D and 3D computational models based on Navier–Stokes equations have been developed, with varying degrees of success. An example is COBRAS (Cornell Breaking waves and Structures model Lin and Liu 1998a,b, Lin et al. 1999), which is capable of describing the

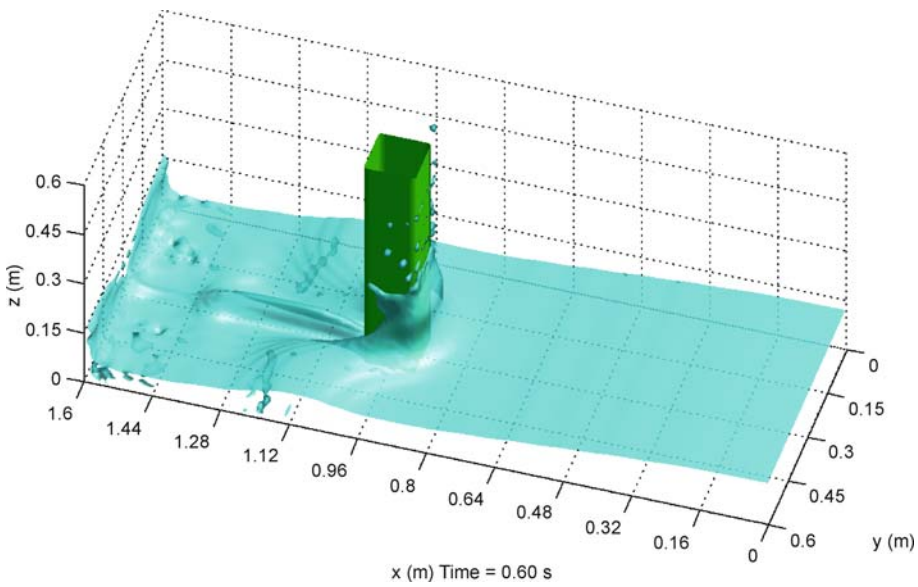

Tsunami Inundation, Modeling of, Figure 5
An simulation snapshot taken from a 3D Navier–Stokes solver with an LES turbulence closure. This setup is looking at a bore impacting a column; the wake and vertical splash are clearly visible (Image provided by Philip L.-F. Liu, Cornell University)

interactions between breaking waves and structures that are either surface piercing or submerged [6,22]. COBRAS adopted the Volume of Fluid (VOF) method to track free surface movement along with a Large Eddy Simulation (LES) turbulence closure model; several other computational models using different free surface tracking methods are also in use, such as the micro surface cell technique developed by Johnson et al. [23]. This 3D Navier–Stokes equation model has been tested by two tsunami related experiments. The first is 3D landslide experiments [38,39], while the second involves measurements of solitary wave forces on vertical cylinders. Both experiments were conducted in the NEES tsunami basin at Oregon State. An example of a LES numerical solution of a solitary wave impinging on a circular is shown in Fig. 5.

Due to their high computational costs, full 3D models would best be used in conjunction with a depth-integrated 2DH model (i. e., NSW or Boussinesq). While the 2DH model provides incident far-field tsunami information, the 3D model computes local wave-structure interactions. The results from 3D models could also provide a better parametrization of small-scale features (3D), which could then be embedded in a large-scale 2DH model. One-way coupling (e. g. using a NSW-generated time series to drive a 3D model, but not permitting feedback from the 3D model back into the NSW) is fairly straightforward to construct (e. g. [18]). Two-way coupling, however, is difficult and requires consistent matching of physics and numerical schemes across model interfaces. Previous work in this area of two-way coupling of hydrodynamic models is limited. Fujima et al. [13] two-way coupled a NLSW model with a fully 3D model. While the results appear promising, the approach used by Fujima et al. requires ad-hoc and unphysical boundary conditions at the model matching locations, in the form of spatial gradients forced to zero, to ensure numerical stability. Even with these ad-hoc treatments, their hybrid model compares very well with the completely-3D-domain simulation, requiring roughly 1/5 of the total 3D CPU time to achieve similar levels of accuracy. Sittanggang et al. [58] presented work on two-way coupling of a Boussinesq model and 2D Navier–Stokes model. These results indicate that there is large potential for hybrid modeling, in terms of more rapid simulation as well as the ability to approach a new class of problems.

Moving Shoreline Algorithms

In order to simulate the flooding of dry land by a tsunami, a numerical model must be capable of allowing the shoreline to move in time. Here, the shoreline is defined as the spatial location where the solid bottom transitions from submerged to dry, and is a function of the two horizontal spatial coordinates and time. Numerical models generally require some type of special consideration and treatment to accurately include these moving boundaries; the logic

and implementation behind this treatment is called a moving shoreline, or runup, algorithm.

For typical tsunami propagation models, it is possible to divide runup algorithms into two main approaches: those on a fixed grid and those on a Lagrangian or transformed domain. Both approaches have their advantages and disadvantages; currently fixed grid methods are found more commonly in operational-level models (e. g. [66]), likely due in large part to their conceptual simplicity. A review of these two classes of models will be given in this section, followed by a review of the standard analytical, experimental, and field benchmarks used to validate the runup models. For additional information, the reader is directed to the comprehensive review given in Pedersen [49].

With a fixed grid method, the spatial locations of the numerical grid points or control volumes are determined at the start of a simulation, and do not change shape or location throughout the simulation duration. These methods can be classified into extrapolation, stairstep, auxiliary shoreline point, and permeable beach techniques. The extrapolation method has its roots in Sielecki and Wurtele [57], with extensions by Hibberd and Peregrine [20], Kowalik and Murty [27], and Lynett et al. [43]. The basic idea behind this method is that the shoreline location can be extrapolated using the nearest wet points, such that its position is not required to be locked onto a fixed grid point; it can move freely to any location. Theoretically, the extrapolation can be of any order; however, from stability constraints a linear extrapolation is generally found. Hidden in the extrapolation, the method is roughly equivalent to the use of low-order, diffusive directional differences taken from the last wet point into the fluid domain [43]. Additionally, there are no explicit conservation constraints or physical boundary conditions prescribed at the shoreline, indicating that large local errors may result if the flow in the extrapolated region cannot be approximately as linear in slope. The extrapolation approach can be found in both NLSW and Boussinesq models with finite difference, finite volume, and finite element solution schemes, and has shown to be accurate for a wide range of non-breaking, breaking, two horizontal dimension, and irregular topography problems (e. g. [8,9,26,44,51]).

Stairstep moving shoreline methods, one of the more common approaches found in tsunami models (e. g. [35]), reconstruct the naturally continuous beach profile into a series of constant elevation segments connected through vertical transitions. In essence, across a single cell width, the bottom elevation is taken as the average value. A cell transitions from a dry cell to a wet cell when the water elevation in a neighboring cell exceeds the bottom elevation, and transitions from wet to dry when the local total water depth falls below some small threshold value. These methods are particularly useful in finite volume and *C*-grid [1] type approaches (e. g. [32,36]), but can be difficult to implement in centered difference models, particularly high-order models or those sensitive to fluid discontinuities, where the "shock" of opening and closing entire cells can lead to numerical noise.

Auxiliary shoreline point methods require dynamic regridding very near the shoreline, such that the last wet point is always located immediately at the shoreline. Obviously, this method requires a numerical scheme that can readily accommodate non-uniform and changing node locations. There is some relation to the extrapolation methods discussed above; the moving shoreline point must be assigned some velocity, and it is extrapolated from the neighboring wet points. However, it is fundamentally different in that the shoreline point is explicitly included in the fluid domain. Thus, it would be expected that the governing conservation equations near the shoreline are more precisely satisfied here, although still dependent on the appropriateness of the extrapolation. One such method can be found in Titov and Synolakis [65], and has been successfully applied in NSLW equation models.

Another fixed grid treatment of moving boundary problems is employing a slot or permeable-seabed technique [63,64]. Conceptually, this method creates porous slots, or conduits, through the dry beach, such that there is always some fluid in a "dry" beach cell, although it may exist below the dry beach surface. These porous, "dry" nodes use a modified form of the NLSW; it is noted here that although in concept this approach is modeling a porous beach, it is not attempting to simulate the groundwater flow under a real, sandy beach, for example. The equations governing the "dry" domain contain a number of empirical parameters that are tuned to provide reasonable runup agreement with benchmark datasets. The advantage of this approach is that is allows the entire domain; including the fluid and "dry" nodes, to be a determined via a somewhat consistent set of governing equations, without requiring a direct search routine to determine the shoreline location. The method has gained some popularity in wind wave models (e. g. [25,46]) when a highly accurate estimate of the shoreline location is not the highest priority. However, the approach has been used with some success in tsunami studies (e. g. [30]) despite the fact that the empirical coefficients that govern the model accuracy cannot be universally determined for a wide range of problems [7].

Alternative to fixed grid methods is the Lagrangian approach. Here, the fluid domain is descritized into particles,

or columns of fluid in depth-integrated models, that are transported following the total fluid derivative. There are no fixed spatial grid locations; the columns move freely in space and time and thus these techniques require numerical flexibility, in terms of utilizing constantly changing space and time steps. The Lagrangian approach can be described as both the more physically consistent and mathematical elegant method of describing shoreline motion. The shoreline "particle" is included in the physical formulation just as any other point in the domain (i. e. no extrapolations are necessary), and thus the shoreline position accuracy will be compromised only by the overarching physical approximation (e. g. long wave approximation) and the numerical solution scheme (e. g. second-order time integration). The cost for this accuracy is a mathematical system that can be more difficult and tedious to solve numerically, typically requiring domain transformations, mappings, and/or re-griddings. Lagrangian methods have been used successfully in finite difference and finite element nonlinear shallow water (NLSW) and Boussinesq equation models (e. g., [2,15,48,50,52,74]).

Future Directions

Towards a more robust simulation of tsunami inundation, there are two major issues which require additional fundamental investigations: dissipation mechanisms and interaction with infrastructure. Bottom friction, known to play an important role in inundation, needs to be re-examined starting from its basic formulation. Can a steady-flow based Mannings-type expression for bottom friction be used for tsunami? Does the unsteady nature of the tsunami flow make use of these existing formulations invalid? The answer to these questions may be different depending on what part of the wave is investigated (e. g. front). In addition, the hydrodynamic effect of common coastal vegetation, such as mangroves, needs to be quantified. There is current discussion of the use of such natural roughness as a tsunami defense (e. g. [11]); confidence cannot be put in such measures until it is understood how they behave. In addition to bottom friction, which exists at all locations and times under an inundating tsunami, wave breaking can increase the total energy dissipation. While breaking is generally confined to the leading front of a tsunami, the characteristics of this front are important for hydrodynamic loadings on beachfront structures, and may be significant to the net sediment and debris transport of a tsunami. Three-dimensional tsunami breaking is poorly understood and has received little attention.

Wave loadings and interactions with infrastructure are not well understood. To tackle this problem, tsunami hydrodynamic models need to be coupled with structural and geotechnical models. Ideally, these models should all be two-way coupled, such that the displacement of a structure, be it a single collapsed wall, will change the flow pattern, and scour underneath the foundation will change the structure stability. Additionally, impacts of flow-transported debris (e. g. cars) should be included in this framework. If such a modeling capacity existed, engineering design of coastal structures could be undertaken in a very efficient manner.

Bibliography

1. Arakawa A, Lamb VR (1977) Computational design of the basic dynamical processes of the UCLA general circulation model. In: Chang J (ed) Methods in computational physics. Academic Press, New York, pp 174–267
2. Birknes J, Pedersen G (2006) A particle finite element method applied to long wave run-up. Int J Numer Methods Fluids 52(3):237–261
3. Borrero JC, Bourgeois J, Harkins G, Synolakis CE (1997) How small-scale bathymetry affected coastal inundation in the 1992 Nicaraguan tsunami. Fall AGU Meeting, San Francisco
4. Briggs MJ, Synolakis CE, Harkins GS, Green D (1994) Laboratory experiments of tsunami runup on a circular island. PAGEOPH 144(3/4):569–593
5. Carrier GF (1966) Gravity waves on water of variable depth. Fluid J Mech 24:641–659
6. Chang K-A, Hsu T-J, Liu PL-F (2001) Vortex generation and evolution in water waves propagating over a submerged rectangular obstacle. Part I solitary waves. Coast Eng 44:13–36
7. Chen Q, Kirby JT, Dalrymple RA, Kennedy AB, Chawla A (2000) Boussinesq modeling of wave transformation, breaking, and runup: Part I 2D. J Waterw Port Coast Ocean Eng 126(1): 57–62
8. Cheung K, Phadke A, Wei Y, Rojas R, Douyere Y, Martino C, Houston S, Liu PL-F, Lynett P, Dodd N, Liao S, Nakazaki E (2003) Modeling of storm-induced coastal flooding for emergency management. Ocean Eng 30:1353–1386
9. Cienfuegos R, Barthelemy E, Bonneton P (2007) A fourth-order compact finite volume scheme for fully nonlinear and weakly dispersive Boussinesq-type equations. Part II: Boundary conditions and model validation. Int J Numer Meth Fluids 53(9):1423–1455
10. Cross RH (1967) Tsunami surge forces, ASCE JI Waterways & Harbors Division WW4:201–231
11. Danielsen F, Sørensen MK, Olwig MF, Selvam V et al (2005) The Asian tsunami: A protective role for coastal vegetation. Science 310:643
12. Fernando HJS, McCulley JL, Mendis SG, Perera K (2005) Coral poaching worsens tsunami destruction in Sri Lanka. Eos Trans AGU 86(33):301
13. Fujima K, Masamura K, Goto C (2002) Development of the 2d/3d hybrid model for tsunami numerical simulation. Coastal Eng J 44(4):373–397
14. Geist E, Lynett P, Chaytor J (2008) Hydrodynamic modeling of tsunamis from the currituck landslide, Marine Geology (in press)

15. Gopalakrishnan TC, Tung CC (1983) Numerical analysis of a moving boundary problem in coastal hydrodynamics. Intl J Numer Meth Fluids 3:179–200
16. González FI, Satake K, Boss EF, Mofjeld HO (1995) Edge wave and non-trapped modes of the 25 April 1992 Cape Mendocino tsunami. Pure Appl Geophys PAGEOPH 144(3–4): 409–426
17. Goring DG, Raichlen F (1992) Propagation of long waves onto shelf. J Waterw Port Coast Ocean Eng 118(1):43–61
18. Guignard S, Grilli ST, Marcer R, Rey V (1999) Computation of shoaling and breaking waves in nearshore areas by the coupling of BEM and VOF methods. Proc. 9th Offshore and Polar Engng. Conf., vol III, ISOPE99, Brest, France, pp 304–309
19. Hammack J, Segur H (1978) Modelling criteria for long water waves. J Fluid Mech 84(2):359–373
20. Hibberd S, Peregrine DH (1979) Surf and run-up on a beach. J Fluid Mech 95:323–345
21. Horrillo, Juan, Kowalik, Zygmunt, Shigihara, Yoshinori (2006) Wave dispersion study in the Indian Ocean–Tsunami of December 26, 2004. Marine Geodesy 29(3):149–166(18)
22. Hsu T-J, Sakakiyama T, Liu PL-F (2002) A numerical model for waves and turbulence flow in front of a composite breakwater. Coast Eng 46:25–50
23. Johnson DB, Raad PE, Chen S (1994) Simulation of impacts of fluid free surfaces with solid boundaries. Int J Num Methods Fluids 19:153–176
24. Johson RS (1972) Some numerical solutions of a variable-coefficient Korteweg–de Vries equation (with application to solitary wave development on a shelf). J Fluid Mech 54:81
25. Kennedy AB, Chen Q, Kirby JT, Dalrymple RA (2000) Boussinesq modeling of wave transformation, breaking, and runup. 1: 1D. J Waterway Port Coastal Ocean Eng ASCE 126:39–47
26. Korycansky DG, Lynett P (2007) Runup from impact tsunami. Geophys J Int 170:1076–1088
27. Kowalik Z, Murty TS (1993) Numerical simulation of two-dimensional tsunami runup. Marine Geodesy 16:87–100
28. Kulikov E (2005) Dispersion of the Sumatra tsunami waves in the Indian Ocean detected by satellite altimetry. Report from P.P. Shirshov Institute of Oceanology, Russian Academy of Sciences, Moscow
29. Imamura F (1995) Review of tsunami simulation with a finite difference method, long-wave runup models. World Scientific, Singapore, pp 25–42, http://en.wikipedia.org/wiki/2004_Indian_Ocean_earthquake
30. Ioualalen M, Asavanant JA, Kaewbanjak N, Grilli ST, Kirby JT, Watts P (2007) Modeling of the 26th December 2004 Indian Ocean tsunami: Case study of impact in Thailand. J Geophys Res 112:C07024, doi:10.1029/2006JC003850
31. Lay T, Kanamori H, Ammon C, Nettles M, Ward S, Aster R, Beck S, Bilek S, Brudzinski M, Butler R, DeShon H, Ekström G, Satake K, Sipkin S (2005) The great Sumatra–Andaman earthquake of December 26, 2004. Science 308:1127–1133, doi:10.1126/science.1112250
32. LeVeque RJ, George DL (2004) High-resolution finite volume methods for the shallow water equations with bathymetry and dry states. In: Liu PL, Yeh H, Synolakis C (eds) Advanced numerical models for simulating tsunami waves and runup. vol 10 of Advances in Coastal and Ocean Engineering. World Scientific, Singapore
33. Losada MA, Vidal V, Medina R (1989) Experimental study of the evolution of a solitary wave at an abrupt junction. J Geophys Res 94:14557
34. Liu PL-F, Synolakis CE, Yeh H (1991) Report on the international workshop on long-wave run-up. J Fluid Mech 229:675–688
35. Liu PL-F, Cho Y-S, Yoon SB, Seo SN (1994) Numerical simulations of the 1960 Chilean tsunami propagation and inundation at Hilo, Hawaii. in: El-Sabh MI (ed) Recent development in tsunami research. Kluwer, Dordrecht, pp 99–115
36. Liu PL-F, Cho Y-S, Briggs MJ, Kanoglu U, Synolakis CE (1995) Runup of solitary waves on a circular island. J Fluid Mech 320:259–285
37. Liu PL-F, Cheng Y (2001) A numerical study of the evolution of a solitary wave over a shelf. Phys Fluids 13(6):1660–166
38. Liu PL-F, Lynett P, Fernando H, Jaffe BE, Fritz H, Higman B, Morton R, Goff J, Synolakis C (2005) Observations by the International Tsunami Survey Team in Sri Lanka. Science 308: 1595
39. Liu PL-F, Wu T-R, Raichlen F, Synolakis CE, Borrero JC (2005) Runup and rundown generated by three-dimensional sliding masses. J Fluid Mech 536:107–144
40. Lynett P (2006) Nearshore modeling using high-order Boussinesq equations. J Waterway Port Coastal Ocean Eng (ASCE) 132(5):348–357
41. Lynett P (2007) The effect of a shallow water obstruction on long wave runup and overland flow velocity. J Waterway Port Coastal Ocean Eng (ASCE) 133(6):455–462
42. Lynett P, Liu PL-F (2002) A numerical study of submarine landslide generated waves and runup. Proc R Soc London A 458:2885–2910
43. Lynett P, Wu T-R, Liu PL-F (2002) Modeling wave runup with depth-integrated equations. Coast Eng 46(2):89–107
44. Lynett P, Borrero J, Liu PL-F, Synolakis CE (2003) Field survey and numerical simulations: A review of the 1998 Papua New Guinea tsunami. Pure Appl Geophys 160:2119–2146
45. Madsen OS, Mei CC (1969) The transformation of a solitary wave over an uneven bottom. J Fluid Mech 39:781
46. Madsen PA, Sorensen OR, Schaffer HA (1997) Surf zone dynamics simulated by a Boussinesq-type model: Part I. Model description and cross-shore motion of regular waves. Coast Eng 32:255–287
47. Matsuyama M, Ikeno M, Sakakiyama T, Takeda T (2007) A study on tsunami wave fission in an undistorted experiment. Pure Appl Geophys 164:617–631
48. Özkan-Haller HT, Kirby JTA (1997) Fourier-Chebyshev collocation method for the shallow water equations including shoreline run-up. Appl Ocean Res 19:21–34
49. Pedersen G (2006) On long wave runup models. In: Proceedings of the 3rd International Workshop on Long-Wave Runup Models, in Catalina Island, California
50. Pedersen G, Gjevik B (1983) Runup of solitary waves. J Fluid Mech 142:283–299
51. Pedrozo-Acuna A, Simmonds DJ, Otta AK, Chadwick AJ (2006) On the cross-shore profile change of gravel beaches. Coast Eng 53(4):335–347
52. Petera J, Nassehi V (1996) A new two-dimensional finite element model for the shallow water equations using a Lagrangian framework constructed along fluid particle trajectories. Int J Num Methods Eng 39:4159–4182
53. Priest GR et al (1997) Cascadia subduction zone tsunamis: Haz-

ard mapping at Yaquina Bay, Oregon. Final Technical Report to the National Earthquake Hazard Reduction Program DOGAMI Open File Report 0-97-34, p 143
54. Ramsden JD (1996) Forces on a vertical wall due to long waves, bores, and dry-bed surges. J Waterway Port Coast Ocean Eng 122(3):134–141
55. Seabra-Santos FJ, Renouard DP, Temperville AM (1987) Numerical and experimental study of the transformation of a solitary wave over a shelf or isolated obstacles. J Fluid Mech 176:117
56. Shen M, Meyer R (1963) Climb of a bore on a beach Part 3. Runup. J Fluid Mech 16(1):113–125
57. Sielecki A, Wurtele MG (1970) The numerical integration of the nonlinear shallow-water equations with sloping boundaries. J Comput Phys 6:219–236
58. Sittanggang K, Lynett P, Liu P (2006) Development of a Boussinesq-RANSVOF Hybrid Wave Model. in Proceedings of 30th ICCE, San Diego, pp 24–25
59. Synolakis CE (1987) The runup of solitary waves. J Fluid Mech 185:523–545
60. Synolakis CE, Imamura F, Tsuji Y, Matsutomi S, Tinti B, Cook B, Ushman M (1995) Damage, Conditions of East Java tsunami of 1994 analyzed. EOS, Transactions, American Geophysical Union 76(26):257 and 261–262
61. Tadepalli S, Synolakis CE (1994) The runup of N-Waves on sloping beaches, Proc R Soc London A 445:99–112
62. Tadepalli S, Synolakis CE (1996) Model for the leading waves of tsunamis. Phys Rev Lett 77:2141–2144
63. Tao J (1983) Computation of wave runup and wave breaking. Internal Report. Danish Hydraulics Institute, Denmark
64. Tao J (1984) Numerical modeling of wave runup and breaking on the beach. Acta Oceanol Sin 6(5):692–700 (in chinese)
65. Titov VV, Synolakis CE (1995) Modeling of breaking and nonbreaking long wave evolution and runup using VTCS-2. J Harbors Waterways Port Coast Ocean Eng 121(6): 308–316
66. Titov VV, Synolakis CE (1998) Numerical modeling of tidal wave runup. J Waterway Port Coast Ocean Eng ASCE 124(4): 157–171
67. Tomita T, Honda K (2007) Tsunami estimation including effect of coastal structures and buildings by 3D model. Coastal Structures '07, Venice
68. Yeh H (2006) Maximum fluid forces in the tsunami runup zone. J Waterway Port Coast Ocean Eng 132(6):496–500
69. Yeh H, Ghazali A, Marton I (1989) Experimental study of bore runup. J Fluid Mech 206:563–578
70. Yeh H, Liu P, Briggs M, Synolakis C (1994) Propagation and amplification of tsunamis at coastal boundaries. Nature 372: 353–355
71. Ward SN, Day S (2001) Cumbre Vieja Volcano – Potential collapse and tsunami at La Palma, Canary Islands. Geophys Res Lett 28(17):3397–3400
72. Weiss R, Wunnemann K, Bahlburg H (2006) Numerical modelling of generation, propagation, and run-up of tsunamis caused by ocean impacts: model strategy and technical solutions. Geophys J Int 67:77–88
73. Woo S-B, Liu PL-F (2004) A finite element model for modified Boussinesq equations. Part I: Model development. J Waterway Port Coast Ocean Eng 130(1):1–16
74. Zelt JA (1991) The run-up of nonbreaking and breaking solitary waves. Coast Eng 15(3):205–246

Tsunamis, Inverse Problem of

Kenji Satake
Earthquake Research Institute, University of Tokyo, Tokyo, Japan

Article Outline

Glossary
Definition of the Subject
Introduction
Tsunami Generation by Earthquakes
Tsunami Propagation
Tsunami Observations
Estimation of Tsunami Source
Estimation of Earthquake Fault Parameters
Future Directions
Bibliography

Glossary

Inverse problem Unlike a forward problem which starts from a tsunami source then computes propagation in the ocean and predicts travel times and/or water heights on coasts, an inverse problem starts from tsunami observations to study the generation process. While forward modeling is useful for tsunami warning or hazard assessments, inverse modeling is a typical approach for geophysical problems.

Shallow water (long) waves In hydrodynamics, water waves can be treated as shallow water, or long, waves when the wavelength is much larger than the water depth. In such a case, the entire water mass from water bottom to surface moves horizontally and the wave propagation speed is given as a square root of the product of the gravitational acceleration and the water depth.

The 2004 Indian Ocean tsunami On December 26, 2004, a gigantic earthquake, the largest in the last half century in the world, occurred off the west coast of Sumatra Island, Indonesia. With the source extending more than 1,000 km through Nicobar and Andaman Islands, the earthquake generated tsunami which attacked the coasts of Indian Ocean and caused the worst tsunami disaster in history. The total casualties were about 230,000 in many countries as far away as Africa.

Fault parameters Earthquake source is modeled as a fault motion, which can be described by nine static parameters. Once these fault parameters are specified, the

seafloor deformation due to faulting, or initial condition of tsunamis, can be calculated by using the elastic dislocation theory.

Refraction and inverse refraction diagrams (travel time map) Refraction diagram is a map showing isochrons or lines of equal tsunami travel times calculated from the source toward coasts. Inverse refraction diagram is a map showing arcs calculated backwards from observation points. The tsunami source can be estimated from the arcs corresponding to tsunami travel times.

Definition of the Subject

Forward modeling of tsunami starts from given initial condition, computes its propagation in the ocean, and calculates tsunami arrival times and/or water heights on coasts. Once the initial condition is provided, the propagation and coastal behavior can be numerically computed on actual bathymetry (Fig. 1).

Recent technological developments make it possible to carry out tsunami forward modeling with speed and accuracy usable for the early tsunami warning or detailed hazard assessments. However, the initial condition, or the tsunami generation process, is still poorly known, because large tsunamis are rare and the tsunami generation in the ocean is not directly observable. Indirect estimation of tsunami source, mostly on the basis of seismological analyzes, is used as the initial condition of tsunami forward modeling. More direct estimation of tsunami source is essential to better understand the tsunami generation process and to more accurately forecast the tsunami on coasts.

Inverse modeling of tsunami starts from observed tsunami data, to study the tsunami source. The propagation process can be evaluated by using numerical simulation, as in the forward modeling. As the observed tsunami data, tsunami arrival times, heights or waveforms recorded on instruments are used. For historical tsunamis, tsunami heights can be estimated from description of damage on historical documents. For prehistoric tsunamis, geological studies of tsunami deposits can be used to estimate the coastal tsunami heights or flooding areas.

Introduction

Tsunamis are oceanic gravity waves generated by seafloor deformation due to submarine earthquakes or other submarine geological processes such as volcanic eruptions, landslides, or asteroid impacts. While earthquake tsunamis, such as the 2004 Indian Ocean tsunami caused by the Sumatra–Andaman earthquake, are most frequent, large volcanic eruptions such as the 1883 Krakatau eruption off Sumatra Island also cause ocean-wide tsunamis. Landslides, often triggered by earthquakes, cause locally large tsunamis, but the effects are usually limited to the area around the source.

Most tsunamigenic geological processes produce seafloor deformation. When horizontal scale, or wavelength, of the seafloor deformation is much larger than the water depth, a similar disturbance appears on the water surface and becomes the source of tsunami. This is called shallow water, or long-wave, approximation. For large earthquakes, wavelength of seafloor deformation is

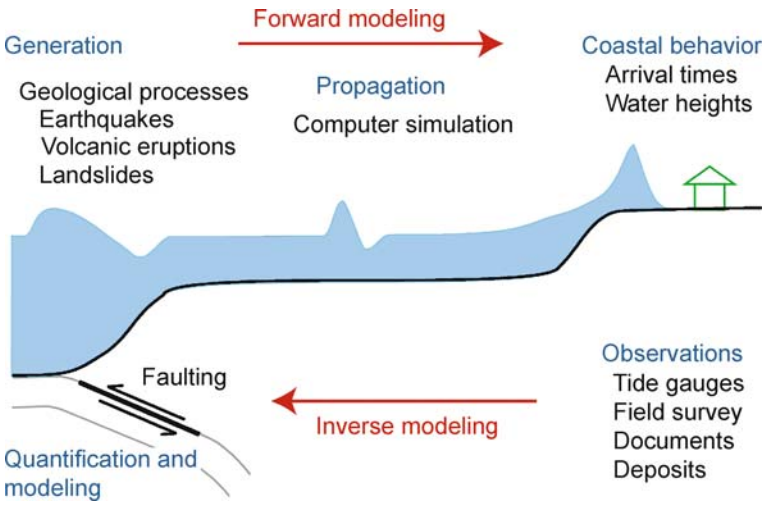

Tsunamis, Inverse Problem of, Figure 1
Schematic diagram showing tsunami generation, propagation and coastal behavior. Forward modeling starts from tsunami source and forecasts the coastal behavior, while inverse modeling starts from observed data to estimate the tsunami source

an order of several tens to hundreds of km, while the ocean depth is up to several km, hence the long-wave approximation is valid. For small scale disturbance relative to water depth, such as submarine landslides or volcanic eruptions in deep seas, the shallow water approximation may not be valid.

This paper reviews inverse methods to study tsunami sources from the observations. Section "Tsunami Generation by Earthquakes" describes the tsunami generation by earthquakes, with emphasis on the fault parameters and their effects on tsunamis. Section "Tsunami Propagation" describes tsunami propagation: shallow water theory and numerical computation. Section "Tsunami Observations" summarizes the tsunami observation: instrumental sea-level data, runup height estimates for modern, historical and prehistoric tsunamis. Section "Estimation of Tsunami Source" describes methods of modeling and quantifying tsunami source, and of analyzing tsunami travel times, amplitudes and waveforms, including some historical developments. Section "Estimation of Earthquake Fault Parameters" focuses on the estimation of earthquake fault parameters, including the waveform inversion of tsunami data to estimate heterogeneous fault motion and its application for tsunami warning.

Tsunami Generation by Earthquakes

Fault Parameters and Seafloor Deformation

The seafloor deformation due to earthquake faulting can be calculated by using the elastic theory of dislocation. The displacement, u_k, in an infinite homogeneous medium due to dislocation Δu_i across surface Σ is given by the Voltera's theorem as [1]

$$u_k = \frac{1}{8\pi\mu} \int_{\Sigma} \Delta u_i \left\{ \lambda\, \delta_{ij}\, u_k^{n,n} + \mu \left(u_k^{i,j} + u_k^{j,i} \right) \right\} \nu_j \, \mathrm{d}S \quad (1)$$

where λ and μ are Lame constants, δ_{ij} is Kronecker's delta, ν is the unit normal to the surface. The expression u_i^j denotes the ith component of the displacement due to the jth component of a point force at the source whose magnitude is $8\pi\mu$, and $u_i^{j,k}$ indicates its spatial derivative with respect to the kth coordinate. For a half-space with free surface, a mirror image can be used to cancel the stress components on the free surface. The explicit formulas are given by Manshinha and Smyle [2] or Okada [3].

The fault parameters needed to compute surface deformation are summarized in Fig. 2. They are: fault length (L), width (W), strike (ϕ), dip (δ), rake (λ), slip amount (u) and location (x, y, z). The slip u can be decomposed into strike-slip component u_s and dip-slip component u_d. The strike ϕ is measured clockwise from North, dip angle δ is downward from horizontal, and rake angle λ is a movement of hanging wall measured counter-clockwise from horizontal (see Fig. 2). Therefore, the fault motion is reverse if $\lambda > 0°$ and normal if $\lambda < 0°$. The fault motion has left-lateral component if $|\lambda| < 90°$ and right-lateral component if $|\lambda| > 90°$.

The physical parameter to quantify the fault motion is the seismic moment, M_0, defined as

$$M_0 = \mu u S = \mu u L W \,. \quad (2)$$

More conventional parameter of earthquake size is a magnitude scale, which has been determined from amplitudes of seismograms. To relate the seismic moment and magnitude scales, the moment magnitude scale, M_w, is defined as [4,5]

$$M_w = \frac{2}{3} \log M_0 - 10.7 \quad (3)$$

where M_0 is given in dyne.cm (10^{-7} Nm).

Most of the above fault parameters can be estimated from seismic wave analysis. The location and depth of fault (x, y, z) correspond to hypocenter, which is estimated from arrival times of seismic waves. The fault geometry (ϕ, δ, λ) is estimated from the polarity distribution of body wave first motions or azimuthal distribution of surface wave amplitudes. The seismic moment is estimated from waveform modeling of seismic waves. The fault size, L and W, are more difficult to estimate; they are usually estimated from aftershock distribution or detailed waveform modeling of seismic body waves. The slip amount, u, is indirectly estimated, from seismic moment M_0 by assuming μ and fault size (L and W). All such estimates assume

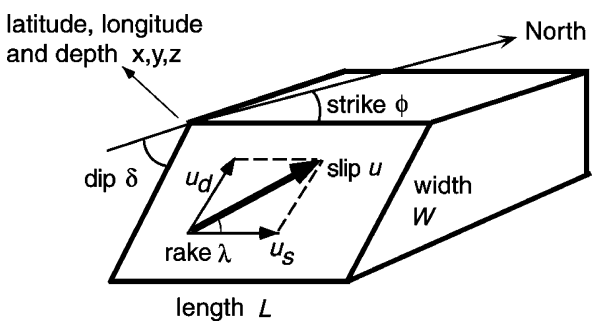

Tsunamis, Inverse Problem of, Figure 2
Fault parameters. Seafloor deformation can be computed from these static parameters

that the faulting is planar and continuous, which most often is a simplification of real, more complex faulting.

The 2004 Sumatra–Andaman earthquake was the largest earthquake since the 1960 Chilean earthquake (M_w 9.5) or 1964 Alaskan earthquake (M_w 9.2). The seismic moment estimates range $4 - 12 \times 10^{22}$ Nm, and the corresponding moment magnitude M_w ranges 9.0–9.3 from the seismological analyzes [6,7,8]. The aftershock area extended from off Sumatra through the Nicobar to Andaman Islands with the total fault length of 1,200 to 1,300 km [6]. For such a gigantic earthquake, multiple fault planes with different strike and slip amounts are needed to represent the fault motion, as shown later (Sect. "Estimation of Earthquake Fault Parameters").

Effect of Fault Parameters on Tsunami Generation

Among the above static fault parameters, the slip amount has the largest effect on the vertical seafloor deformation and the tsunami amplitude. The dip angle and fault depth are also important parameters to control tsunami amplitude [9,10]. Dynamic parameters such as rupture velocity are found to be insignificant for tsunami generation. However, for a gigantic earthquake such as the 2004 Sumatra–Andaman earthquake with the source length over 1,000 km, rupture propagation effect on tsunamis is not negligible [11].

Amplitude of far-field seismic waves, either body waves or surface waves, is controlled by seismic moment, while amplitude of tsunami is controlled by fault slip. Satake and Tanioka [12] found for the 1998 Papua New Guinea tsunami that the far-field tsunami amplitudes are proportional to the volume of displaced water, while the near-field tsunami amplitudes are controlled by the potential energy of the displacement.

Traditionally, only vertical component of seafloor deformation has been considered for tsunami generation (Fig. 3a). If an earthquake occurs on a steep ocean slope such as trench slope, horizontal displacement due to faulting moves the slope and contributes the tsunami generation [13]. The effective vertical movement (positive upward) due to faulting can be written as follows (Fig. 3b),

$$u_z + u_x \frac{\partial H}{\partial x} + u_y \frac{\partial H}{\partial y} \tag{4}$$

where u_z is vertical component and u_x and u_y are horizontal components of seafloor deformation, and H is water depth (measured positive downward).

Recently, Song et al. [14] proposed that horizontal motion of slope also transfers kinetic energy from seafloor to water. They claimed that the kinetic energy thus transferred was about five times larger than the potential energy for the 2004 Indian Ocean tsunami, from comparisons of the observed tsunami (sea surface height) data with the computed ones by using an ocean-general-circulation-model.

Tsunami Propagation

Shallow Water Theory

The equation of motion, or conservation of momentum, for shallow water, or long-wave, theory is given as follows.

$$\frac{\partial \boldsymbol{V}}{\partial t} + (\boldsymbol{V} \cdot \nabla)\, \boldsymbol{V} = -g \nabla h + C_f \frac{\boldsymbol{V}\,|\boldsymbol{V}|}{d + h} \tag{5}$$

where $\boldsymbol{V}$ is the depth-averaged horizontal velocity vector, h is the water height or tsunami amplitude, d is the water depth and g is the gravitational acceleration. On the left-hand side, the first term represents local acceleration and the second term represents nonlinear advection. On the right-hand side, the first term represents pressure gradient, or restoring force due to gravity, and the second term represents nonlinear bottom friction where C_f is the non-dimensional frictional coefficient.

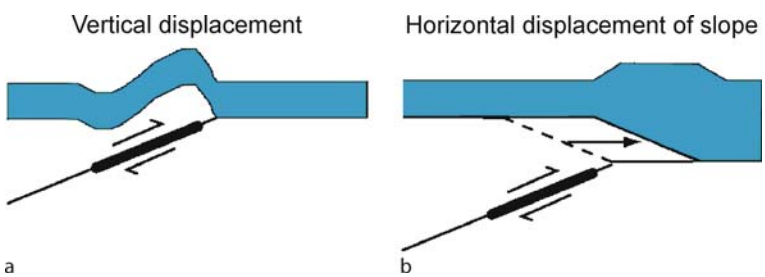

Tsunamis, Inverse Problem of, Figure 3
Seafloor deformation and tsunami source. a Vertical seafloor deformation becomes the tsunami source. b When the seafloor is not flat, horizontal displacement of the slope also affects the tsunami generation [13]

The equation of continuity, or conservation of mass, can be written as

$$\frac{\partial(d+h)}{\partial t} = -\nabla \cdot \{(d+h)\,\boldsymbol{V}\}\,. \tag{6}$$

These equations can be linearized as

$$\frac{\partial \boldsymbol{V}}{\partial t} = -g\nabla h \tag{7}$$

$$\frac{\partial h}{\partial t} = -\nabla \cdot (d\boldsymbol{V}) \tag{8}$$

when the tsunami amplitude h is small compared to water depth d, and the bottom friction can be neglected. Such an assumption is valid for deep ocean, or most of the tsunami propagation path. Near the coasts, nonlinear terms play important roles, hence linearization may not be valid. The major advantage of the linear theory is the superposition principle; the computational results can be easily scaled to estimate with different initial water heights.

From Eqs. (7) and (8), the wave equation with wave velocity (celerity) $\sqrt{gd}$ can be derived. This indicates that the tsunami speed is controlled by water depth. Once the water depth distribution, or ocean bottom bathymetry, is known, then the tsunami propagation can be computed numerically.

Numerical Computations

Equations (5) and (6), or (7) and (8) for linearized case, can be directly solved by numerical methods, once the initial condition is given. The tsunami wave velocity distribution, which is given by the bathymetry, is much better known than the velocity distribution of seismic waves, hence actual values can be used. Finite-difference method with staggered grids is popularly used [15,16], while use of other methods such as finite-element methods have been also proposed [17]. Grid size for finite-difference computations is typically a few km for deep ocean, but grids as fine as several tens to hundreds of meters are used near coasts. The temporal changes in water height at grid points corresponding to the observation points are used as computed tsunami waveforms.

The database of global water depth or bathymetry data such as ETOPO2 (NOAA/NGDC) or GEBCO (British Oceanographic Data Centre) are popularly used. The ETOPO2 database is based on predicted bathymetry from satellite altimetry data [18] with interval of 2 minutes (about 3.5 km), while GEBCO data are digitized from nautical charts with grid interval of 1 minute. Higher resolution bathymetry data near coasts are open to public in some countries such as US (NOAA/NGDC) or Japan (JODC).

Tsunami Observations

Instrumental Data

Traditional instrumental data for tsunami observation are tide gauge records. Tide gauges are typically installed on ports or harbors to define datum or to monitor ocean tides. The temporal resolution is usually low with a sampling interval of several minutes or longer. For the tsunami monitoring, higher sampling rate, at least 1 min or shorter interval, is required. While the recorded tsunami waveforms contains coastal effects such as coastal reflections or resonance particularly for the later phase, the initial part of tsunami signals is more dominant by the tsunami source effect hence the source information can be retrieved. Currently, sea level measurement data from many tide gauge stations are transmitted through weather satellite and available in real time. Figure 4 (left) shows some of tide gauge records of the 2004 Indian Ocean tsunami.

Tsunami waveforms are simpler offshore or in deep ocean, free from nonlinear coastal effects, though the signal is smaller. Offshore and deep ocean tsunami observation facilities have been significantly developed recently. Offshore tsunami gauges such as GPS tsunami gauge have been developed and deployed around Japan [19]. In 2003, cabled bottom pressure gauges have recorded the tsunami generation process in the source area [20]. The US NOAA deployed dozens of bottom pressure gauges, called Deep-ocean Assessment and Reporting of Tsunamis (DART) or simply Tsunameters [21]. The bottom pressure signals are sent to surface buoys via acoustic telemetry in the ocean, then to land station via satellite. As described later, the DART data are used for real-time data assimilation and tsunami warning purposes [22]. After the 2004 Indian Ocean tsunami, many DART-type bottom pressure gauges have been deployed in the Pacific and Indian Oceans.

Satellite altimeters captured the propagation of the 2004 Indian Ocean tsunami (Fig. 4). Three satellites flew over the Indian Ocean at a few hours after the earthquake and measured the sea surface height (SSH) of about 0.8 m in the middle of Indian Ocean. The tsunami amplitudes in deep ocean are much smaller than the maximum coastal heights of more than 10 m. The SSH data are used to study the tsunami source [11,23].

Modern, Historical and Prehistoric Tsunami Heights

After damaging tsunamis, tsunami height distribution is often measured by survey teams [24]. Measurements are usually made for flow depth above ground, on the basis of

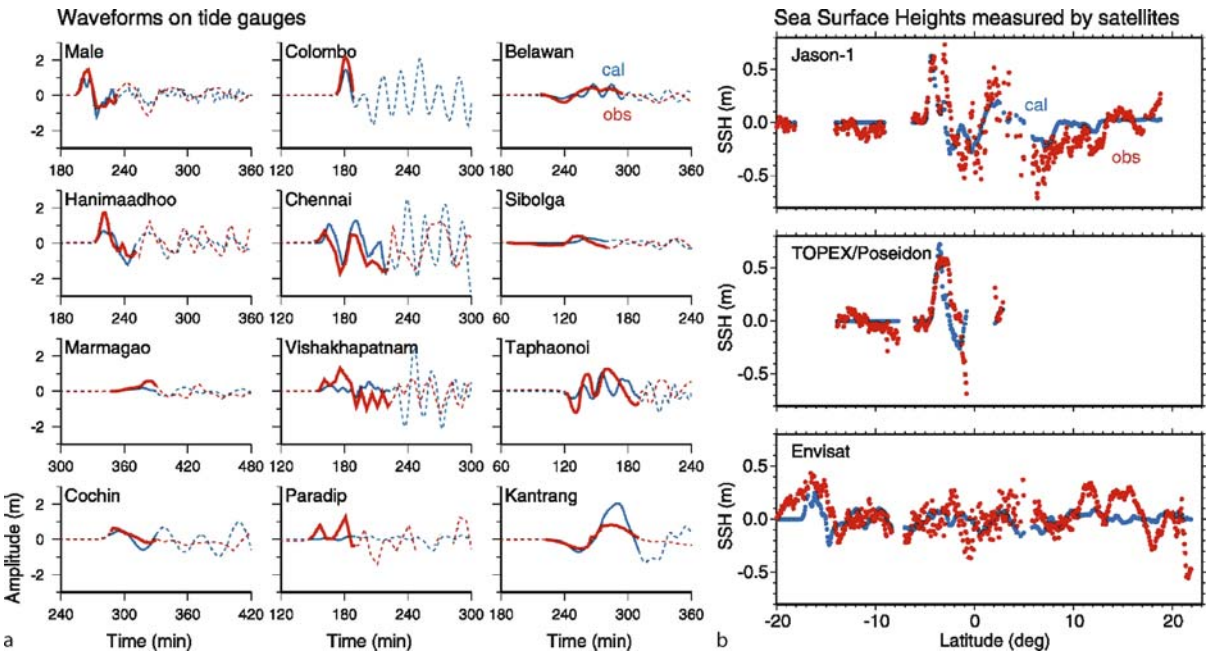

Tsunamis, Inverse Problem of, Figure 4

The sea level data from the 2004 Indian Ocean tsunami [11]. a Tsunami waveforms on tide gauges. *Red curves* indicate observed waveforms and *blue ones* are computed. Data shown in *solid lines* are used for the waveform inversion. b Sea surface heights measured by three different satellites (see Fig. 5 for the tracks). *Red* shows the observed data and *blue* is for computed surface heights

various watermarks, then converted to inundation height above sea level [25]. The tsunami inundation heights are usually not constant along a profile from beach, and the height at most inland point is called runup height.

For historical tsunamis, coastal tsunami heights can be estimated from descriptions of tsunami or its damage recorded in historical documents. Such estimates include various assumptions on sea levels and a relationship between tsunami damage and flow depth, but provide important tsunami data for historical tsunamis. For example, date and size of the last gigantic earthquake in the Cascadia subduction zone off North America were estimated as January 26, 1700 and $M_w \sim 9.0$ from the Japanese tsunami records [26].

Geological traces such as tsunami deposits can also be used to estimate tsunami heights for prehistoric tsunamis. In the last few decades, many studies of tsunami deposits, combined with numerical computations, have been made to analyze prehistoric tsunamis [27,28]. For example, in Hokkaido, prehistoric tsunami deposits indicate past tsunamis with larger inundation area and longer recurrence interval than those from the recent plate-boundary earthquakes along the southern Kuril trench, which were attributed to multi-segment earthquakes with $M_w \sim 8.4$ [29].

Estimation of Tsunami Source

Refraction Diagram

Tsunami propagation can be computed and described as a refraction diagram. When the tsunami wavelength is smaller than the scale length of velocity heterogeneity, or the water depth variation is smooth, then the geometrical ray theory of optics can be applied. The wavefronts of propagating tsunami can be drawn on the basis of Huygens' principle. Alternatively, propagation of rays, which is orthogonal to wavefronts, can be traced from an assumed source. While refraction diagram do not provide information on water height, the relative amplitudes can be estimated from density of rays [30].

Refraction diagrams can be prepared for major tsunami sources and used for tsunami warning; as soon as the epicenter is known, the tsunami arrival times can be readily calculated. The refraction diagrams are usually drawn from a point source, but it is possible to draw it from an extended source for a great or giant earthquake. Figure 5a shows the refraction diagram from the 2004 Sumatra–Andaman earthquake with wavefronts at each hour. To the east of the assumed source, the tsunami is expected to arrive at the Thai coast in about two hours through Andaman Sea. To the west, through deeper Bay

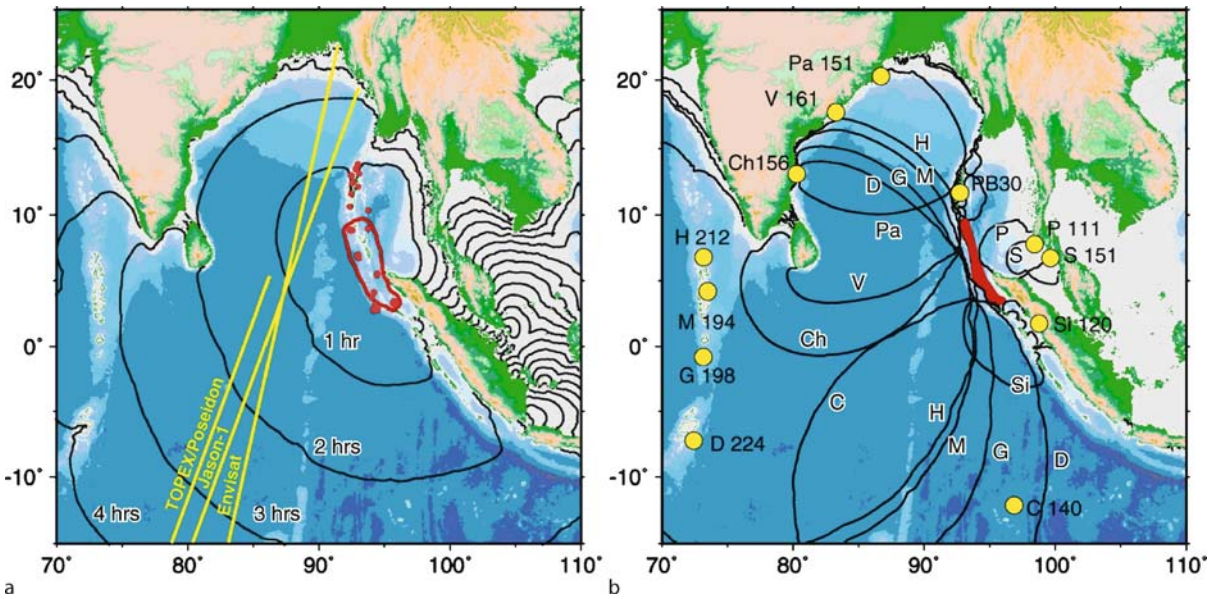

Tsunamis, Inverse Problem of, Figure 5

a Tsunami refraction diagram for the 2004 Sumatra–Andaman earthquake. *Red dots* indicate aftershocks within 1 day according to USGS. The *red curve* shows the assumed tsunami source. Tracks of three satellites with altimeters are shown by *yellow lines*. *Black curves* indicate tsunami wavefronts at each hour after the earthquake. b Tsunami inverse refraction diagram for the same event. Station code and tsunami arrival times (in min) are attached to tide gauge stations (*yellow circles*) where the tsunami was instrumentally recorded. *Black curves* are the travel-time arcs computed for each station. *Red area* indicates inferred tsunami source

of Bengal, the tsunami is expected to arrive at Sri Lanka also in two hours. The predicted tsunami arrival times are similar to the actually observed values [31].

Inverse Refraction Diagram

Refraction diagram can be drawn backwards from coasts. Such a diagram is called inverse refraction diagram and is used to estimate the tsunami source area. When the tsunami travel time, that is tsunami arrival time minus earthquake origin time, is known, the corresponding wavefront, or travel-time arc, drawn from the tsunami observation point (typically tide gauge stations) would indicate the initial wavefront at the tsunami source. The tsunami inverse refraction diagram was first drawn for the 1933 Sanriku tsunami [32], although the estimated tsunami source was much larger than modern estimates, because both tsunami travel times and the bathymetry were poorly known.

The 2004 Indian Ocean tsunami was observed at many tide gauge stations in the Indian Ocean [33,34]. The tsunami arrival times were read from the tide gauge records and tsunami travel times were calculated from the earthquake origin time. The tsunami propagation was then computed from each tide gauge station, and wavefronts corresponding to the travel time were drawn as travel-time arcs (Fig. 5b). These travel-time arcs surround the tsunami source, and the source area was estimated as about 900 km long [6,35].

Estimation of Tsunami Source

Tsunami data can be used to study earthquake source processes in a similar way that seismic waves are used. This was first demonstrated for the 1968 Tokachi-oki earthquake ($M_0 = 2.8 \times 10^{21}$ Nm or $M_w = 8.3$) [36]. The tsunami source area estimated from an inverse refraction diagram agrees well with the aftershock area (Fig. 6a, b). In addition, the initial water surface disturbance was estimated as uplift at the southeastern edge and subsidence at the northwestern edge, from the first motion of recorded tsunami waveforms on tide gauges. This pattern is very similar to the vertical bottom deformation due to the faulting, which was independently estimated from seismological analysis (Fig. 6c).

Green's Law and Tsunami Heights

The water height in the tsunami source area can be estimated from the observed tsunami heights along the coasts, by using the Green's law. The Green's law is derived from

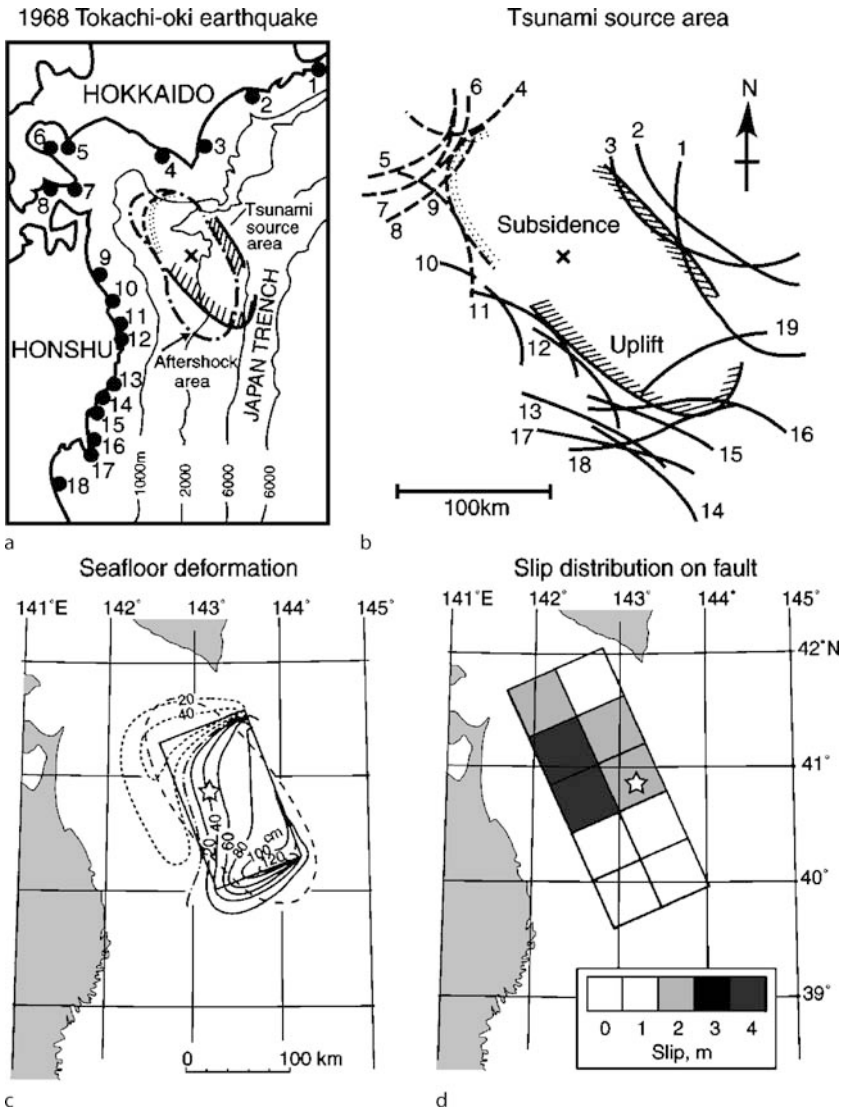

Tsunamis, Inverse Problem of, Figure 6

The tsunami source area and seafloor deformation of the 1968 Tokachi-oki earthquake [36,37]. a Estimated tsunami source, aftershock area and distribution of tide gauge stations. b *Travel-time arcs* drawn by inverse refraction diagram. The *numbers* correspond to tide gauge stations in (a). *Solid* and *dashed curves* show uplift and subsidence, respectively. c Seafloor deformation computed from a seismological fault model. d Slip distribution on fault estimated by an inversion of tsunami waveforms

the conservation of potential energy along rays [38],

$$b_0\, d_0^{1/2}\, h_0^2 = b_1\, d_1^{1/2}\, h_1^2 \tag{9}$$

where d is the water depth, b is the distance between the neighboring rays, h is the tsunami amplitude, and the subscripts 0 and 1 indicate two different locations on the same ray. If the tsunami amplitude at location 0 (e. g., on the coast) is known, the tsunami amplitude at location 1 (e. g., at the source) can be estimated as

$$h_1 = \left(\frac{b_0}{b_1}\right)^{1/2} \left(\frac{d_0}{d_1}\right)^{1/4} h_0\,. \tag{10}$$

The ratio b_0/b_1 represents the spreading of rays, which can be graphically obtained from refraction diagrams. For the Tokachi-oki earthquake, the average tsunami height at the source was estimated as 1.8 m using the Green's law, which is very similar to 1.6 m, the average vertical seafloor displacement computed from the fault model [36].

The Green's law is also used to estimate the shoaling effects. For plane waves approaching the coast, the spreading ratio is unity, hence the amplitude is proportional to a 1/4 power of water depth change. For example, when the water depth becomes a half, the amplitude becomes 1.18 times larger.

Tsunami Magnitude

Tsunami magnitude scale, M_t, was introduced to quantify earthquake source that generated tsunamis [39]. The formulas were calibrated with the moment magnitude scale, M_w, of earthquakes. It is different from other tsunami magnitude or intensity scales that simply quantify the observed tsunamis. The definition of M_t for a trans-Pacific tsunami is [39]:

$$M_t = \log H + C + 9.1 \tag{11}$$

and for a regional (100 km < Δ < 3500 km) tsunami is [40]:

$$M_t = \log H + \log \Delta + 5.8 \tag{12}$$

where H is a maximum amplitude on tide gauges in meters, C is a distance factor depending on a combination of the source and the observation points, and Δ is the nautical distance in km. The tsunami magnitude M_t was assigned as $M_t = 8.2$ for the 1968 Tokachi-oki earthquake and $M_t = 9.0$ for the 2004 Sumatra–Andaman earthquake.

Because the tsunami magnitude scale M_t is defined from tsunami amplitudes, it can be used to characterize "tsunami earthquakes" that produce much larger tsunamis than expected from seismic waves (see Polet and Kanamori: ► Tsunami Earthquakes). Abe [41] defined "tsunami earthquakes" for such events with tsunami magnitude M_t larger than surface wave magnitude M_s by more than 0.5. It should not be confused with "tsunamigenic earthquake" which refers to any earthquake that generates tsunami.

Estimation of Earthquake Fault Parameters

For earthquake tsunamis, the fault parameters can be estimated by inverse modeling of tsunamis. Such attempts were first made by a trial and error approach. In order to estimate the heterogeneous fault parameters, inversion of tsunami waveforms or runup heights has been introduced.

Trial and Error Approach

Numerical simulation of tsunami has been carried out for many tsunamigenic earthquakes around Japan [42]. For the 1968 Tokachi-oki earthquake, tsunami waveforms were computed from two models, one based on seismological analysis (Fig. 6c) and another horizontally shifted by 28 km, and were compared with the observed tsunami waveforms recorded on tide gauges. Comparison of waveforms indicates that the latter model, shifted from that based on seismological analysis, shows better match in terms of tsunami arrival times. The slip amount on the fault was estimated as 4 m.

The best fault models are judged by comparison of the observed and computed tsunami waveforms. A few statistical parameters used to quantify the comparison are geometric mean, logarithmic standard deviation and correlation coefficient. The geometric mean K of the amplitude ratio O_i/C_i, where O_i and C_i are the observed and computed tsunami amplitudes at station i, is given as

$$\log K = \frac{1}{n}\sum_i \log \frac{O_i}{C_i} . \tag{13}$$

The logarithmic standard deviation κ is defined as

$$\log \kappa = \left[\frac{1}{n}\sum_i \left(\log \frac{O_i}{C_i}\right)^2 - (\log K)^2\right]^{1/2} . \tag{14}$$

If the logarithmic amplitude ratios obey the normal distribution $N(\log K, \log \kappa)$, then parameter κ can be considered as an error factor, because its logarithm shows the standard deviation. The geometric mean K indicates the relative size of the observed and computed tsunami models. The logarithmic standard deviation κ indicates the goodness of the model; the smaller κ means the better model. The arrival times of observed and computed waveforms are also compared, as indicated in the above example. Another parameter is correlation coefficient of the observed and computed waveforms, which are also used for the comparison of models.

While the above parameters (K and κ) were originally defined for maximum amplitudes of waveforms, they are also used for comparison of observed and computed runup heights. For tsunami hazard evaluation of nuclear power plants in Japan, tsunami source models need to satisfy $0.95 < K < 1.05$ and $\kappa < 1.45$ for the observed and computed coastal heights [43].

Heterogeneous Fault Motion

Seismological studies of large earthquakes have indicated that the slip amount is not uniform but heterogeneous on faults. For the 1968 Tokachi-oki earthquake, inversion of far-field body waves or regional Rayleigh waves showed

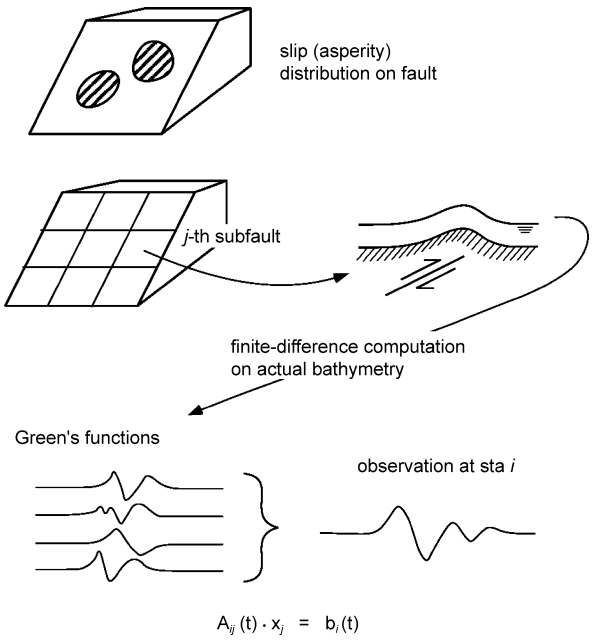

Tsunamis, Inverse Problem of, Figure 7
Schematic illustration of tsunami waveform inversion method

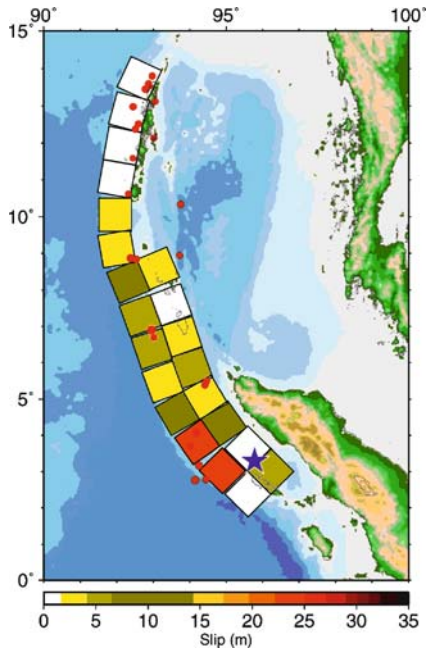

Tsunamis, Inverse Problem of, Figure 8
Slip distribution on 22 subfaults of the 2004 Sumatra–Andaman earthquake estimated from a joint inversion of tsunami waveforms on tide gauges and sea surface heights measured by satellite altimeters [11]

the slip distributions similar to that estimated from tsunami waves [44,45]. The large slip area, called asperity, produces high-frequency seismic waves thus important for strong-motion prediction for earthquake hazard assessments. The asperity produces large seafloor deformation, hence it is also important for tsunami generation and its hazard estimation. Lay and Kanamori [46] suggested that characteristic size of asperities differs from one subduction zone to another and is controlled by geological setting. Yamanaka and Kikuchi [47], from studies of recurrent earthquakes off northern Honshu, showed that the same asperity ruptures in repeated earthquakes. Their asperity map can be used for earthquake and tsunami hazard assessment.

Waveform Inversion

The asperity distribution can be estimated by an inversion of tsunami waveforms. In this method (Fig. 7), the fault plane is first divided into several subfaults, and the seafloor deformation is computed for each subfault with a unit amount of slip. Using these as the initial conditions, tsunami propagation is numerically computed for actual bathymetry and the waveforms at tide gauge stations, called Green's functions, are computed. Assuming that the tsunami generation and propagation are linear process, the observed tsunami waveforms are expressed as a linear superposition of Green's functions as follows,

$$A_{ij}(t) \cdot x_j = b_i(t) \tag{15}$$

where $A_{ij}(t)$ is the computed waveform as a function of time t, or Green's function, at the ith station from the jth subfault; x_j is the amount of slip on the jth subfault; and $b_i(t)$ is the observed tsunami waveform at the ith station. The slip x_j on each subfault can be estimated by a least-square inversion of the above set of equations, by minimizing the l_2 norm of the residuals J

$$J = \|\boldsymbol{A} \cdot \boldsymbol{x} - \boldsymbol{b}\| \rightarrow \min \tag{16}$$

where $\boldsymbol{A}$, $\boldsymbol{x}$, and $\boldsymbol{b}$ indicate matrix representations of elements in Eq. (15). Figure 6d shows the slip distribution on the fault for the 1968 Tokachi-oki earthquake. The source fault was divided into 10 subfaults, and slip distribution on the subfaults were estimated. The largest slip, about 3.7 m, was estimated on the subfaults to the west of epicenter, but the average slip is 1.2 m, much smaller than that estimated by the trial and error method (4 m) which compared the maximum tsunami amplitudes [37].

The 2004 Indian Ocean tsunami, caused by the Sumatra–Andaman earthquake, was recorded by satellite altimeters, as well as tide gauges. A joint inversion of

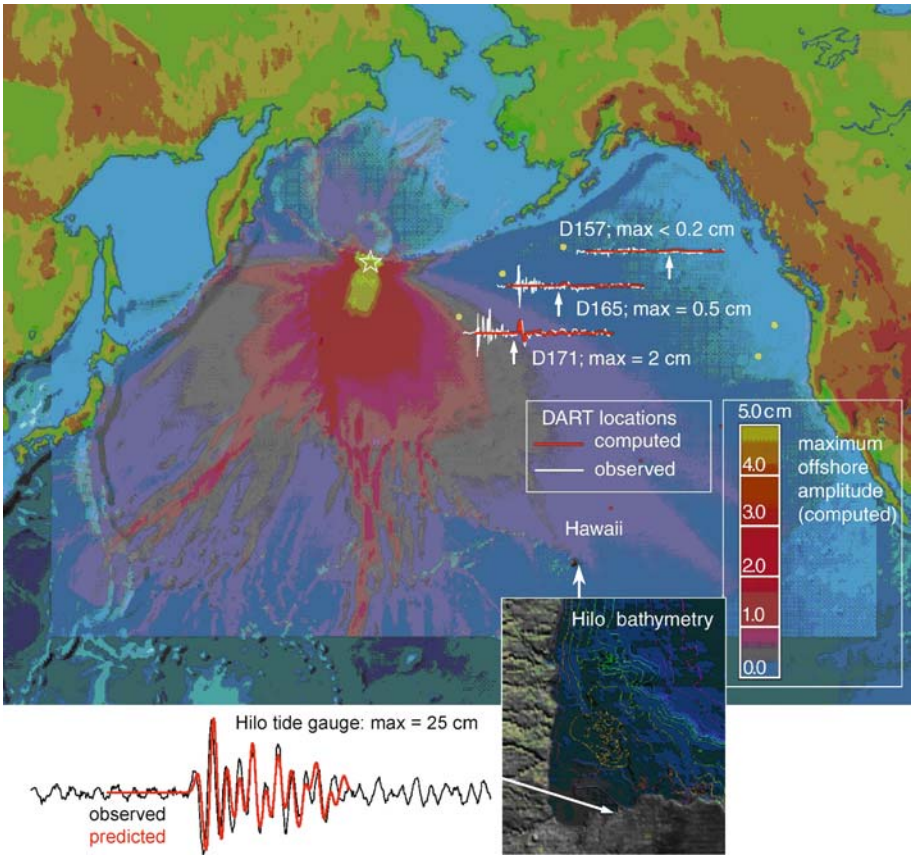

Tsunamis, Inverse Problem of, Figure 9

Real-time data assimilation for the November 17, 2003 Rat Island Tsunami [22]. *Star* indicates the epicenter. *Yellow dots* are locations of DART systems. The tsunami waveforms recorded on DART (*white curves*) are compared with computed waveforms (*red*). *Filled colors* on ocean show computed maximum tsunami amplitudes of the updated source. The *bottom plot* shows location of Hilo tide gauge (*right*) and comparison of the predicted tsunami waveforms (*red*) with the recorded tide gauge signal (*black*)

tsunami waveforms recorded at 12 tide gauge stations and the sea surface heights measured by three satellites indicates that the tsunami source was about 900 km long [11]. The estimated slip distribution (Fig. 8) indicates that the largest slip, about 13 to 25 m, was located off Sumatra Island and the second largest slip, up to 7 m, near the Nicobar Islands. Inversion of satellite altimeter data alone supports a longer, about 1,400 km long, tsunami source [23], but such a model produces much larger tsunami waveforms than observed at Indian tide gauge stations. Inversion of tide gauge data alone does not support tsunami source beneath Andaman Islands [48]. The slip distribution estimated by the joint inversion is similar to those estimated from seismological analyses. The fault slip was the largest near off the northern Sumatra, followed by off Nicobar Islands [49]. Fault slip around Andaman Islands was estimated to be small from seismological analysis [50].

Nonlinear Inversion Methods

In the above inversion method, both tsunami generation and propagation process are assumed to be linear. Because slip amount u, among the nine static fault parameters, is linearly related with the seafloor deformation and tsunami amplitudes, and it has the largest effect on tsunamis, other parameters are fixed in the above method. However, the tsunami propagation, particularly near coasts, might be affected by some nonlinear process such as advection or bottom friction.

Nonlinear inversion methods to overcome these limitations have been proposed. Pires and Miranda [51] proposed an adjoint method, which consists of four steps: source area delimitation by backward ray-tracing, the optimization of the initial sea state, nonlinear adjustments of the fault model, and final optimization of fault parameters.

The minimum of residual J is obtained iteratively through gradient descent method using the partial derivative or gradient of J with respect to parameters to be inverted.

Inversion of Tsunami Heights

Maximum tsunami heights on coasts, rather than tsunami waveforms, have also been used for tsunami inversion. Distribution of maximum tsunami heights along the coasts is available from field surveys, historical or geological studies, and is valuable to study tsunami source. Piatanesi et al. [52] used coastal tsunami heights of the 1992 Nicaragua "tsunami earthquake" and estimated the slip distribution on the fault, as well as the mean amplification factor of computed coastal heights and measured runup heights.

Annaka et al. [53] proposed a method of joint inversion of tsunami waveforms and runup heights. As a residual to be minimized, they used a weighted sum of difference in waveforms (similar to Eq. 16) and logarithm of runup heights. They first tried linear inversion to estimate the initial value, then estimated the perturbation by the nonlinear inversion. Similar nonlinear inversion methods, based on the initial solution estimated by either linear inversion or other data such as geodetic data, have been proposed [54,55].

Real-Time Data Assimilation

The tsunami waveform inversion can be done in real-time for the purpose of tsunami warning. The real-time data assimilation using the DART records and tsunami forecast have been made by NOAA [22]. They first use seismological information to determine the source location and parameters, then using the database of pre-computed simulation results, invert the DART data to estimate the tsunami source size (slip amounts). The tsunami forecast is made for farther locations where the tsunami has not arrived. For the 2003 Rat Island (Aleutians) earthquake, they successfully forecasted tsunami waveforms at Hilo, Hawaii, before the tsunami arrivals (Fig. 9).

Future Directions

Inverse modeling methods of tsunami need to be further developed to better understand the tsunami generation process. Future developments are expected in each field of observation, propagation modeling and application to seismic and nonseismic tsunamis.

The tsunami observation system has been improved recently, particularly after the 2004 Indian Ocean tsunami. Many instrumental data, both coastal and offshore, become available for the studies of tsunami generation process [56]. Maintenance of the systems, particularly for offshore systems, is sometimes costly, but essential to record infrequent tsunami. Open and real-time availability of such data is also important for tsunami studies as well as for tsunami warning purposes.

For the past tsunamis, more studies are needed to estimate tsunami heights from historical documents, as well as geological data such as distribution of tsunami deposits. Such historical tsunami database has been developed, e. g., at NOAA/NGDC (http://www.ngdc.noaa.gov/seg/hazard/tsu_db.shtml).

For modeling tsunamis recorded on coastal tide gauges or runup heights, nonlinear computations with very fine bathymetry data are essential. While computational methods have been developed, fine bathymetry data are not always available. Developments of nonlinear inversion methods are also important.

Finally, inversion of tsunami data can be applied to tsunamis generated from submarine processes other than earthquakes, such as volcanic eruptions or landslides. For such nonseismic tsunamis, parametrization is essential to quantify the geological process and to solve inverse problems.

Bibliography

Primary Literature

1. Steketee JA (1958) On Volterra's dislocations in a semi-infinite elastic medium. Can J Phys 36:192–205
2. Mansinha L, Smylie DE (1971) The displacement fields of inclined faults. Bull Seism Soc Am 61:1433–1440
3. Okada Y (1985) Surface deformation due to shear and tensile faults in a half-space. Bull Seism Soc Am 75:1135–1154
4. Hanks T, Kanamori H (1979) A moment magnitude scale. J Geophys Res 84:2348–2350
5. Kanamori H (1977) The energy release in great earthquakes. J Geophys Res 82:2981–2987
6. Lay T, Kanamori H, Ammon CJ, Nettles M, Ward SN, Aster RC, Beck SL, Bilek SL, Brudzinski MR, Butler R, DeShon HR, Ekstrom G, Satake K, Sipkin S (2005) The great Sumatra–Andaman earthquake of 26 December 2004. Science 308:1127–1133
7. Stein S, Okal EA (2005) Speed and size of the Sumatra earthquake. Nature 434:581–582
8. Tsai VC, Nettles M, Ekstrom G, Dziewonski AM (2005) Multiple CMT source analysis of the 2004 Sumatra earthquake. Geophys Res Lett 32. doi:10.1029/2005GL023813
9. Geist E (1998) Local tsunamis and earthquake source parameters. Adv Geophys 39:117–209
10. Yamashita T, Sato R (1974) Generation of tsunami by a fault model. J Phys Earth 22:415–440
11. Fujii Y, Satake K (2007) Tsunami source model of the 2004 Sumatra–Andaman earthquake inferred from tide gauge and satellite data. Bull Seism Soc Am 97:S192–S207

12. Satake K, Tanioka Y (2003) The July 1998 Papua New Guinea earthquake: Mechanism and quantification of unusual tsunami generation. Pure Appl Geophys 160: 2087–2118
13. Tanioka Y, Satake K (1996) Tsunami generation by horizontal displacement of ocean bottom. Geophys Res Lett 23:861–864
14. Song YT, Fu L-L, Zlotnicki V, Ji C, Hjorleifsdottir V, Shum CK, Yi Y (2008) The role of horizontal impulses of the faulting continental slope in generating the 26 December 2004 tsunami. Ocean Modell 20:362–379
15. Intergovernmental Oceanographic Commission (1997) IUGG/IOC TIME Project Numerical Method of Tsunami Simulation with the Leap-frog Scheme. UNESCO, Paris
16. Mader CL (1988) Numerical modeling of water waves. University of California Press, Berkeley
17. Yeh H, Liu P, Synolakis C (1996) Long-wave runup models. World Scientific, Singapore
18. Smith WHF, Scharroo R, Titov VV, Arcas D, Arbic BK (2005) Satellite altimeters measure tsunami, early model estimates confirmed. Oceanography 18:10–12
19. Kato T, Terada Y, Kinoshita M, Kakimoto H, Isshiki H, Matsuishi M, Yokoyama A, Tanno T (2000) Real-time observation of tsunami by RTK-GPS. Earth Planet Space 52:841–845
20. Mikada H, Mitsuzawa K, Matsumoto H, Watanabe T, Morita S, Otsuka R, Sugioka H, Baba T, Araki E, Suyehiro K (2006) New discoveries in dynamics of an M8 earthquake-phenomena and their implications from the 2003 Tokachi-oki earthquake using a long term monitoring cabled observatory. Tectonophysics 426:95–105
21. Gonzalez FI, Bernard EN, Meinig C, Eble MC, Mofjeld HO, Stalin S (2005) The NTHMP tsunameter network. Nat Hazard 35: 25–39
22. Titov VV, Gonzalez FI, Bernard EN, Eble MC, Mofjeld HO, Newman JC, Venturato AJ (2005) Real-time tsunami forecasting: Challenges and solutions. Nat Hazard 35:41–58
23. Hirata K, Satake K, Tanioka Y, Kuragano T, Hasegawa Y, Hayashi Y, Hamada N (2006) The 2004 Indian Ocean tsunami: Tsunami source model from satellite altimetry. Earth Planet Space 58:195–201
24. Synolakis CE, Okal EA (2005) 1992–2002: Perspective on a decade of post-tsunami surveys. In: Satake K (ed) Tsunamis: Case studies and recent developments. Springer, Dordrecht, pp 1–29
25. Intergovernmental Oceanographic Commission (1998) Post-tsunami survey field guide. UNESCO, Paris
26. Satake K, Wang KL, Atwater BF (2003) Fault slip and seismic moment of the 1700 Cascadia earthquake inferred from Japanese tsunami descriptions. J Geophys Res 108. doi:10.1029/2003JB002521
27. Atwater BF, Musumi-Rokkaku S, Satake K, Tsuji Y, Ueda K, Yamaguchi DK (2005) The orphan tsunami of 1700. USGS Prof Paper 1707:133
28. Dawson AG, Shi SZ (2000) Tsunami deposits. Pure Appl Geophys 157:875–897
29. Nanayama F, Satake K, Furukawa R, Shimokawa K, Atwater BF, Shigeno K, Yamaki S (2003) Unusually large earthquakes inferred from tsunami deposits along the Kuril trench. Nature 424:660–663
30. Satake K (1988) Effects of bathymetry on tsunami propagation – Application of ray tracing to tsunamis. Pure Appl Geophys 126:27–36
31. Rabinovich AB, Thomson RE (2007) The 26 December 2004 Sumatra tsunami: Analysis of tide gauge data from the world ocean Part 1, Indian Ocean and South Africa. Pure Appl Geophys 164:261–308
32. Miyabe N (1934) An investigation of the Sanriku tunami based on mareogram data. Bull Earthq Res Inst Univ Tokyo Suppl 1:112–126
33. Merrifield MA, Firing YL, Aarup T, Agricole W, Brundrit G, Chang-Seng D, Farre R, Kilonsky B, Knight W, Kong L, Magori C, Manurung P, McCreery C, Mitchell W, Pillay S, Schindele F, Shillington F, Testut L, Wijeratne EMS, Caldwell P, Jardin J, Nakahara S, Porter FY, Turetsky N (2005) Tide gauge observations of the Indian Ocean tsunami, December 26, 2004. Geophys Res Lett 32. doi:10.1029/2005GL022610
34. Nagarajan B, Suresh I, Sundar D, Sharma R, Lal AK, Neetu S, Shenoi SSC, Shetye SR, Shankar D (2006) The great tsunami of 26 December 2004: A description based on tide-gauge data from the Indian subcontinent and surrounding areas. Earth Planet Space 58:211–215
35. Neetu S, Suresh I, Shankar R, Shankar D, Shenoi SSC, Shetye SR, Sundar D, Nagarajan B (2005) Comment on "The great Sumatra–Andaman earthquake of 26 December 2004". Science 310:1431a
36. Abe K (1973) Tsunami and mechanism of great earthquakes. Phys Earth Planet Inter 7:143–153
37. Satake K (1989) Inversion of tsunami waveforms for the estimation of heterogeneous fault motion of large submarine earthquakes – The 1968 Tokachi-Oki and 1983 Japan Sea earthquakes. J Geophys Res 94:5627–5636
38. Mei CC (1989) The applied dynamics of ocean surface waves. World Scientific, Singapore
39. Abe K (1979) Size of great earthquakes of 1873–1974 inferred from tsunami data. J Geophys Res 84:1561–1568
40. Abe K (1981) Physical size of tsunamigenic earthquakes of the northwestern Pacific. Phys Earth Planet Inter 27:194–205
41. Abe K (1989) Quantification of tsunamigenic earthquakes by the Mt scale. Tectonophysics 166:27–34
42. Aida I (1978) Reliability of a tsunami source model derived from fault parameters. J Phys Earth 26:57–73
43. Yanagisawa K, Imamura F, Sakakiyama T, Annaka T, Takeda T, Shuto N (2007) Tsunami assessment for risk management at nuclear power facilities in Japan. Pure Appl Geophys 164: 565–576
44. Kikuchi M, Fukao Y (1985) Iterative deconvolution of complex body waves from great earthquakes – The Tokachi-oki earthquake of 1968. Phys Earth Planet Inter 37:235–248
45. Mori J, Shimazaki K (1985) Inversion of intermediate-period Rayleigh waves for source characteristics of the 1968 Tokachi-oki earthquake. J Geophys Res 90:11374–11382
46. Lay T, Kanamori H (1981) An asperity model of large earthquake sequences. In: Simpson DW, Richards PG (eds) Earthquake prediction – An international review. American Geophysical Union, Washington DC, pp 579–592
47. Yamanaka Y, Kikuchi M (2004) Asperity map along the subduction zone in northeastern Japan inferred from regional seismic data. J Geophys Res 109:B07307, doi:10.1029/2003JB002683
48. Tanioka Y, Yudhicara, Kusunose T, Kathiroli S, Nishimura Y, Iwasaki S-I, Satake K (2006) Rupture process of the 2004 great Sumatra–Andaman earthquake estimated from tsunami waveforms. Earth Planet Space 58:203–209
49. Ammon CJ, Ji C, Thio HK, Robinson D, Ni SD, Hjorleifsdottir V,

Kanamori H, Lay T, Das S, Helmberger D, Ichinose G, Polet J, Wald D (2005) Rupture process of the 2004 Sumatra–Andaman earthquake. Science 308:1133–1139
50. Velasco AA, Ammon CJ, Lay T (2006) Search for seismic radiation from late slip for the December 26, 2004 Sumatra–Andaman (Mw=9.15) earthquake. Geophys Res Lett 33:L18305, doi:10.1029/2006GL027286
51. Pires C, Miranda PMA (2001) Tsunami waveform inversion by adjoint methods. J Geophys Res 106:19773–19796
52. Piatanesi A, Tinti S, Gavagni I (1996) The slip distribution of the 1992 Nicaragua earthquake from tsunami run-up data. Geophys Res Lett 23:37–40
53. Annaka T, Ohta K, Motegi H, Yoshida I, Takao M, Soraoka H (1999) A study on the tsunami inversion method based on shallow water theory. Proc Coastal Engin JSCE 46:341–345
54. Yokota T, Nemoro M, Masuda T (2004) Estimate of slip distribution by tsunami height data inversion. Abst Jpn Earth Planet Sci Joint Meeting S043-P0005
55. Namegaya Y, Tsuji Y (2007) Distribution of asperities of the 1854 Ansei Nankai earthquake. Abst Jpn Earth Planet Sci Joint Meeting S142-009
56. Satake K, Baba T, Hirata K, Iwasaki S, Kato T, Koshimura S, Takenaka J, Terada Y (2005) Tsunami source of the 2004 off the Kii peninsula earthquakes inferred from offshore tsunami and coastal tide gauges. Earth Planet Space 57:173–178

Books and Reviews

Lawson CL, Hanson RJ (1974) Solving least squares problems. Prentice-Hall, Englewood Cliffs. (Republished by Society for Industrial and Applied Mathematics, 1995)
Lay T, Wallace TC (1995) Modern global seismology. Academic Press, San Diego
Menke W (1989) Geophysical data analysis: Discrete inverse theory (revised edition). Academic Press, San Diego
Satake K (2007) Tsunamis. In: Kanamori H (ed) Treatise on Geophysics, vol 4. Elsevier, Amsterdam

Tunneling Through Quantum Dots with Discrete Symmetries

Yshai Avishai[1,2], Konstantin Kikoin[3]
[1] Physics Department and Ilse Katz Institute for Nanotechnology, Ben-Gurion University of the Negev, Beer-Sheva, Israel
[2] RTRA-Triangle de la Physique, LPS (Orsay) and CEA-SPHT (Saclay), Gif sur Yvette, France
[3] Department of Physics, Tel-Aviv University, Tel-Aviv, Israel

Article Outline

Glossary
Definition of the Subject
Introduction
Complex Quantum Dots
Dynamical Symmetries and Kondo Effect in Tunneling Through Complex Quantum Dots
Discrete Degrees of Freedom in Kondo Tunneling
Tunneling in External Fields
Connections with other Nano-objects and Future Prospects
Bibliography

Glossary

Quantum dot Quantum dot (QD) is an element of artificial nanostructures. It arises in a situation when a finite number of electrons is confined in a puddle of 10^0–10^2 nanometer size either by means of external electrical potential imposed on semiconductor heterostructure or in a process of formation of non-equilibrium self-assembled structures. The electron wavelength in a QD is comparable with its size. As a result all energy levels are spatially quantized like in atoms or molecules. Quantum systems with fully quantized energy levels are defined as zero-dimensional nano-objects. Complex quantum dots consist of several QDs organized in linear or ring structures.

Discrete symmetry Group theory classifies all physical objects in accordance with their symmetry properties relative to symmetry transformations (rotations and translations) in space and time. These operations may be continuous or discrete. In the latter case one speaks about discrete rotational or translational symmetry.

Dynamical symmetry Dynamical symmetry characterizes not only the eigenstates of a Hamiltonian, but also the symmetry of transition between the states belonging to different irreducible representation of the symmetry group of the Hamiltonian under external dynamical perturbation. Dynamical symmetry may be continuous and discrete or combine both types of symmetry operations.

Kondo effect Kondo effect is the many particle phenomenon which arises in magnetically doped metals due to exchange scattering of metallic electrons on localized magnetic moments of impurities. Multiple electron scattering processes result in dynamical screening of impurity magnetic moment. As a result, impurity-related electrical resistivity has a minimum at some temperature, and reaches the unitarity limit at $T \to 0$. This limit corresponds to maximum backward electron scattering. Kondo effect exists also in tunneling through small QDs. In this case the unitarity limit corresponds to maximum tunnel transparency of QD.

Quantum tunneling Quantum tunneling of electrons through potential barriers arises at low temperatures,

where all overbarrier activation processes are frozen. In nanostructures this type of transport is realized on the interfaces between different materials forming nanostructure or in especially designed tunneling channels.

Definition of the Subject

Electrons may be confined within a nano-size quantum box by various methods. The first example of such confinement was demonstrated in the studies of optical properties of semiconductor precipitates in glasses [10]. Later on such confinement was realized in planar quantum dots. These dots are fabricated in semiconductor heterostructures, where the electrons already confined in a two-dimensional layer between two semiconductors (usually GaAs/GaAlAs) are locked in a nano-size puddle by electrostatic potential created by electrodes superimposed on the heterostructure (see [25,51] for a description of the early stage of the physics of quantum dots). QDs may be also prepared by means of colloidal synthesis [4], grown as self-assembled structures of semiconductor droplets on a strained surface of another semiconductor [36], etc. In particular, quantum dots may be fabricated in a form of vertical structures possessing cylindrical symmetry [32].

Discrete symmetries become relevant in tunneling through QD when this nanoobject consists of several individual dots. In such complex quantum dots (CQDs) electrons may occupy any of the individual quantum wells, having the same or nearly the same energy provided all wells are identical or nearly identical. Thus, additional degeneracy of the electron spectrum arises, which is characterized by the discrete permutation symmetry group. If CQD has the form of an "artificial molecule" or has a form of a regular polygon, this permutation symmetry transforms into point symmetry. Interplay between the discrete symmetries of CQD, $SU(2)$ spin symmetry of the electron confined in the dot and $U(1)$ symmetry of its charge results in new exciting properties of these artificial systems which are unimaginable in "natural" objects such as atoms, molecules or impurities in metals and semiconductors.

Introduction

To study electron tunneling through QDs, one should build it into an electric circuit. In the case of the planar quantum dot the metallic electrodes (source and drain leads) are formed in the two-dimensional layer filled by electrons. The tunnel current emerges when the bias voltage V_b, is applied to the leads, and the number of electrons in QD is regulated by the gate voltage v_g applied to the puddle. The number of electrons in such QD may be varied experimentally from 1–2 to several tens. The most remarkable property of QDs is a single electron tunneling, which is realized in small enough dots, provided the charging energy Q spent for injecting an electron from the dot exceeds both the inter-level spacing $\delta\varepsilon$ and the tunneling rate Γ. In this regime the current-voltage characteristics $I(V_b)$ acquires the step-wise form of the "Coulomb ladder" instead of the conventional linear Ohm's law. Tunnel conductance $G = \delta I/\delta V_b$ is nothing but a sequence of delta-function-like peaks. Each peak corresponds to injection of the next electron in the dot. This possibility of "counting electrons by number" opens the way to construction of single electron transistors and other nano-devices.

Semiconductor QDs may be used as single photon emitters ("single photon on demand"). Besides, an ensemble of QDs is proposed as a reliable carrier of spin qubits for quantum computers [37,53].

On the other hand, a quantum dot with strong Coulomb-exchange correlation in a tunnel contact with two electron reservoirs (Fermi-seas) is an excellent model system for studying fundamental quantum mechanical and quantum statistical phenomena, such as the Kondo effect [15,42], edge singularity phenomenon [39], Fano-resonances [24,29], etc. Quantum dots may be organized in complex self-assembled arrays, and these arrays possess features, which combine the properties of solid state and atomic physics [6,55].

Complex Quantum Dots

In the case of a strong Coulomb blockade, the number of electrons in a QD is fixed, and one should discriminate between the dots with even and odd electron occupation. Electrons in a QD occupy discrete levels in accordance with the Pauli principle. Then the dot with odd occupation is characterized by spin 1/2, whereas the dot with odd occupation has zero spin. Disk-like planar QDs and vertical QDs possess cylindrical symmetry, so these nano-objects may be considered as few electron systems with shell structures ("artificial cylindrical atoms" [32]). The electron shells are occupied in accordance with Hund rules modified for cylindrical symmetry [23]. As a result the ground and low-lying excited states in these objects are manifolds consisting of states with spin $\boldsymbol{S} = 1/2, 3/2$ (odd occupation) and $\boldsymbol{S} = 1, 0$ (even occupation).

Complex quantum dots may be treated as "artificial molecules". The simplest of such complex objects is the double quantum dot (DQD), first realized experimentally in planar geometry [18,41]. It consists of two islands with confined electrons. Both capacitive (electrostatic) and tunneling coupling may exist between these islands. If two is-

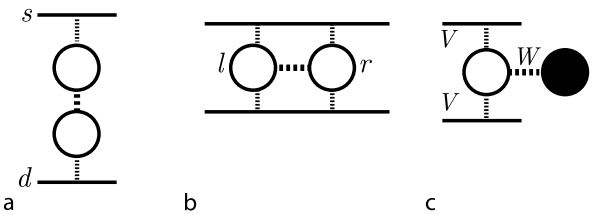

Tunneling Through Quantum Dots with Discrete Symmetries, Figure 1
DQD in sequential (a), parallel (b) and T-shape (c) geometry. The *filled black circle* marks the dot detached from the leads

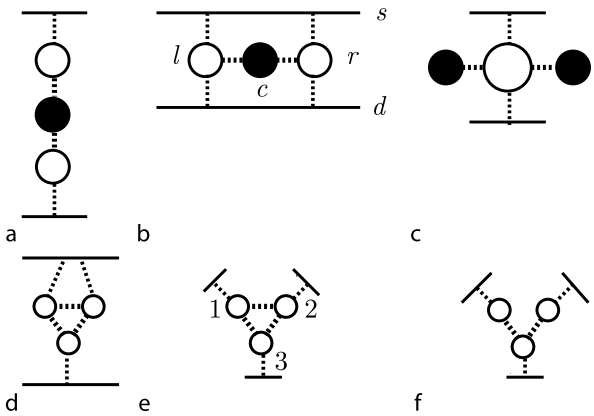

Tunneling Through Quantum Dots with Discrete Symmetries, Figure 2
TQD in numerous configurations: sequential (a), parallel (b), cross (c), two-terminal triangular (d), three-terminal triangular (e) and fork (f). The *filled black circles* mark the dots which are detached from the leads

lands are equivalent, then the structure of the electronic spectrum mimics that of elementary molecules: DQD with two electrons simulates H_2, the dots with one or three electrons look like positively and negatively charged H molecules, etc. If the two dots are not equivalent (e. g., two potential wells have different depth and/or width) then the electronic structure of DQD mimics that of a polar molecule, e. g. DQD with four electrons reminds us of the LiH molecule. One may tune the balance between ionic and covalent components of the inter-dot coupling by varying the gate voltage applied to two dots and the width of the tunnel channel between them.

DQD may be coupled to source and drain electrodes in several ways (Fig. 1). In cases (a) and (b) there is a one and two tunneling channel, respectively. In case (c) tunneling through the single channel is controlled by charge and spin states of the side dot.

In the case of even occupation and strong left-right (l-r) tunnel coupling, all electrons are shared between two islands, the dot is characterized only by its spin state (singlet or triplet), and the electron motion from source (s) to drain (d) is possible only in the co-tunneling regime (where the electron from the source tunnels into the DQD only when another electron leaves it on its way into the drain). In case of odd occupation one "unpaired" electron oscillates between the two islands. As a result, the DQD acquires an orbital degree of freedom, which is characterized by a discrete l-r symmetry.

Like in other two-level systems, this symmetry may be characterized by a pseudo-spin operator $\boldsymbol{\tau}$ acting in (l-r) space.

More diverse discrete symmetries are realized in triple quantum dots (TQD). These dots may have linear structure [8,16] or form closed loops (triangles) [56]. Some interesting and experimentally accessible geometric structures of tunneling devices with TQD are shown in Fig. 2.

The linear TQDs shown in the upper panel are characterized by l-r reflection symmetry, whereas the triangular TQD with equivalent constituent smaller dots (configurations (d, e) in the lower panel), have a discrete symmetry of a triangle, that is represented by the group C_{3v} which is equivalent to a permutation symmetry P_3. This symmetry is violated in case (d) by the tunnel contacts with source and drain. The QD with cross symmetry (c) is a generalization of the T-shape symmetric QD displayed in Fig. 1c.

Dynamical Symmetries and Kondo Effect in Tunneling Through Complex Quantum Dots

Any characteristic (internal) symmetry of a QD may be (and generically is) violated due to an interaction with the environment (usually electrons in the leads forming a Fermi sea). All systems which are schematically depicted in Figs. 1, 2 can be described by the Hamiltonian

$$H = H_{\mathrm{d}} + H_{\mathrm{r}} + H_{\mathrm{c}} \tag{1}$$

Here the indices d, r, c denote dot, electron reservoir and dot-reservoir coupling, respectively. In a simple special case the variables, which are conserved in isolated QD are charge and spin, so that

$$H_{\mathrm{d}} = H_{\mathrm{d0}} + QN^2 + \lambda \boldsymbol{S}^2 \,. \tag{2}$$

Here H_{d0} describes the discrete energy spectrum of the QD, Q is the electrostatic (capacitive) energy fixing the number of electrons N in the dot, and $\lambda < 0$ is the exchange energy responsible for stabilization of the dot spin (described by the operator $\boldsymbol{S}$). The Hamiltonian (1) is a generic Hamiltonian for a class of strongly correlated

electron systems (SCES) [13,27]: it describes coupling between two subsystems, one of which H_r contains nearly free particles with weak interaction, whereas the low-energy excitations in the dot Hamiltonian H_d are fully determined by the strong correlation, which acquires the form of the constraint as in (2). As a rule, the coupling term H_c between the two subsystems results in a complete reconstruction of the low-energy part of the energy spectrum of SCES within some interval E_K due to various many-body effects characterized by infrared singularities. Since the symmetry of the Hamiltonian (2) is violated by H_c, this coupling initiates transitions between the states belonging to different irreducible representations (irreps) of the symmetry group G_H of H_d within the energy interval E_K. This means that the dynamical symmetry of CQD influences the physics of tunneling through this object.

In the context of a quantum tunneling problem, the dynamical symmetry group G_D is determined as the group generated by the collection of operators $\boldsymbol{R}$ inducing transitions both within a given irrep and between different irreps of the symmetry group G_H [28]. This set of operators form a closed algebra with commutation relations

$$[R_i, R_j] = f_{ijk} R_k \,, \tag{3}$$

where f_{ijk} are the structure factors. The simplest example of such a group is the group $SO(4)$ describing the dynamical symmetry of triplet-singlet manifold $\boldsymbol{S} = 1, 0$ of DQD with even occupation [26]. This group is formed by six generators. Three of them are the components of the operator $\boldsymbol{S}$, and the other three are the components of another vector $\boldsymbol{P}$ with the components

$$\begin{aligned} P^+ &= \sqrt{2}(|T_\uparrow\rangle\langle S| - |S\rangle\,|\langle T_\downarrow|\rangle \,, \\ P_z &= -(|T_0\rangle\langle S| + |S\rangle\langle T_0|) \end{aligned} \tag{4}$$

describing transitions between singlet (S) and three components of spin triplet (T_ν). In the case of DQD with $N = 2$, where each dot in Fig. 1 is occupied by one electron with spin $\mathbf{s}$, the two vectors may be represented as

$$\boldsymbol{S} = \boldsymbol{s}_1 + \boldsymbol{s}_2 \,, \quad \boldsymbol{P} = \boldsymbol{s}_1 - \boldsymbol{s}_2 \,. \tag{5}$$

Kinematics imposes two constraints on these vectors, namely

$$\boldsymbol{s}_1^2 + \boldsymbol{s}_2^2 = 3/2 \,; \quad \boldsymbol{s}_1^2 - \boldsymbol{s}_2^2 = 0 \,.$$

These two constraints are in fact the two Casimir operators $\hat{\mathbf{C}}_1, \hat{\mathbf{C}}_2$ supplementing the $o(4)$ algebra of the $SO(4)$ group. In other words, these two constraints are $\hat{\mathbf{C}}_1 = \boldsymbol{S}^2 + \boldsymbol{P}^2 = 3$, $\hat{\mathbf{C}}_2 = \boldsymbol{S}\cdot\boldsymbol{P} = 0$. The first of them arises in H_d instead of the usual Casimir operator $\boldsymbol{S}^2 = S(S+1)$. This Hamiltonian then acquires the form

$$H_d = H_{d0} + QN^2 + (E_T \boldsymbol{S}^2 + E_S \boldsymbol{P}^2)/2 \,, \tag{6}$$

where E_T, E_S are the energies of triplet and singlet states respectively, and $E_T - E_S = -\lambda$.

Among the many-body features, which accompany tunneling through QDs the most salient and universal is the Kondo effect. This effect, originally found in many-particle scattering of electrons by magnetic impurities in metals [31] is responsible also for the well pronounced temperature and magnetic field dependent zero bias anomalies in conductance through quantum dots [15,42]. The physical explanation of this mapping of the scattering problem in bulk metals on the tunneling problem in low-dimensional systems is that the strong Coulomb blockade allows electron propagation through the QD only in the co-tunneling regime (see above), which does not preserve the spin of the dot. The co-tunneling Hamiltonian which arises in second order in H_c, projects out charge excitations and takes into account spin reversal processes. It has the form of the effective exchange term,

$$H_{\text{cot}} = J\boldsymbol{S}\cdot\boldsymbol{\sigma} \,, \tag{7}$$

where the operator $\boldsymbol{\sigma}$ describes spin excitations of the itinerant electrons in a metallic reservoir and $J \sim V^2$ is the coupling constant quadratic in the tunneling amplitude V between the reservoir and the QD.

In complex quantum dots the Kondo effect should take into account dynamical symmetries of CQD, provided the characteristic energy scale $E_K \sim \exp(-1/J)$ (Kondo energy) is comparable with the scale of low-lying states in the manifold of eigenstates of H_d. In DQD where this manifold includes the states E_T, E_S, dynamical symmetry $SO(4)$ is involved in Kondo tunneling provided the singlet-triplet gap $|\lambda| \sim E_K$ [11,26]. Then both vectors $\boldsymbol{S}$ and $\boldsymbol{P}$ are involved in Kondo cotunneling, and the effective Hamiltonian has the form

$$H_{\text{cot}} = J_1 \boldsymbol{S}\boldsymbol{\sigma} + J_2 \boldsymbol{P}\cdot\boldsymbol{\sigma} \,. \tag{8}$$

In this situation the Kondo energy scale is a function of the exchange gap $E_K(\lambda)$, and the zero bias anomaly in the tunnel conductance becomes sensitive to its value.

Discrete Degrees of Freedom in Kondo Tunneling

The short chains of QDs represented in Fig. 1a–c and Fig. 2a–c, are among the most simple objects, where the interplay between discrete symmetries and the original Kondo physics may be demonstrated. An electron in

a DQD in the charge sector $N = 1$ is represented not only by its spin $\boldsymbol{\sigma}$ with projections $\pm$, but also by the pseudospin vector operator $\boldsymbol{\tau}$ describing its position in l-r space within the double well:

$$\boldsymbol{\tau}_z = \sum_\sigma \left(d^\dagger_{l\sigma} d_{l\sigma} - d^\dagger_{r\sigma} d_{r\sigma} \right) , \quad \boldsymbol{\tau}^+ = \sum_\sigma d^\dagger_{l\sigma} d_{r\sigma} ,$$
$$\boldsymbol{\tau}^- = \sum_\sigma d^\dagger_{r\sigma} d_{l\sigma} . \tag{9}$$

This vector, together with the four spin vectors $\boldsymbol{S}_{ij}$ representing spin in a double well $(i, j = l, r)$ form a set of 15 generators of the $SU(4)$ group:

$$\{(\boldsymbol{\tau}^+, \boldsymbol{\tau}^-, \boldsymbol{\tau}_z, I) \otimes (\sigma^+, \sigma^-, \sigma_z, I)\} - \{I \otimes I\} . \tag{10}$$

The effective co-tunneling Hamiltonian in this case acquires the form

$$H_{\text{cot}} = \sum_{ij} J_{ij} \boldsymbol{S}_{ij} \cdot \boldsymbol{s}_{ji} + K \boldsymbol{\tau} \cdot \boldsymbol{t} . \tag{11}$$

Here I is the unit 2×2 matrix, $\boldsymbol{s}_{ij}$ is the spin operator for electrons in the leads (s, d), $\mathbf{t}$ is the pseudospin operator for lead electrons similar to (9). In sequential tunneling geometry of Fig. 1a, only the coupling (ls) and (rd) is taken into account.

Electron levels in the double well are split due to the inter-well tunneling $\boldsymbol{\tau}^\pm$. Both split levels are involved in Kondo cotunneling provided the splitting energy Δ is comparable with the Kondo energy.

The Kondo energy itself is a function of this splitting, $E_{\text{K}}(\Delta)$. This energy splitting plays the same role in the discrete dynamical $SU(4)$ symmetry as the exchange splitting λ does in the dynamical symmetry $SO(4)$. In both cases, involvement of excited states in the Kondo tunneling results in the increase of E_{K}, and thereby it causes the enhancement of the zero bias anomaly of the tunnel conductance [11,12,26,49]. This increment is encoded within the following asymptotic power law,

$$E_{\text{K}}(\delta) = E_{\text{K}}(0)[E_{\text{K}}(0)/\delta]^\gamma \qquad (\delta = \lambda, \Delta) . \tag{12}$$

Here the exponent γ is specific for a given dynamical symmetry.

In many cases, the exchange splitting energy may change its sign following tuning of the dot parameters, as well as gate voltages. In this case, one speaks about triplet singlet crossover. When the triplet state is involved in Kondo tunneling at finite energy but the zero bias anomaly disappears at low temperatures $kT \ll |\lambda|$ [19,49], one even speaks of a quantum phase transition.

An even more multifarious situation arises in linear TQD [33]. For example, in a charge sector $N = 4$ the electron distribution in a TQD with strong Coulomb blockade in the central dot and weak blockade in the side dots (Fig. 2a, b) has a shell-like structure: the first electron occupies the deepest level in the central dot, the two additional electrons which are shared between the two side dots form a sort of closed shell and the last electron dwells in a double (l-r) well. The low-lying part of the spin manifold consists of two triplets and two singlets. The relative positions of these levels may be changed by varying the gate voltages applied to the side dots as well as by tuning other experimentally controllable parameters of the device.

In the most symmetric situation where the l-r reflection symmetry is preserved, the symmetry of the TQD is $P_2 \otimes SO(4) \otimes SO(4)$. This is the symmetry of two singlet/triplet pairs degenerate under left-right permutation. It can be reduced to the $SO(7)$ dynamical symmetry, which is realized for the manifold consisting of two triplets and one singlet. The set of 21 operators, which describe transitions within this manifold form a closed algebra $o(7)$. These operators are grouped in six vectors and three scalars. The pattern of commutation relations has a more complicated form than (3). One may also construct five Casimir operators describing kinematical constraints on spin variables in the TQD. Another dynamical symmetry is the $SO(5)$ symmetry of a manifold containing two singlets and one triplet. The corresponding algebra $o(5)$ is formed by ten generators grouped in three vectors and one scalar (the latter describes transition between two singlets). Three Casimir operators constrain this dynamical symmetry. Other symmetries realizable in TQD are $SO(4)$, $P_2 \otimes SO(3) \otimes SO(3)$ and $SO(3)$.

Thus, because of interplay between discrete l-r reflection symmetry and dynamical $SO(n)$ symmetries of spin multiplets, the linear TQD possesses a quite complicated phase diagram, where the index n may be changed. In accordance with the general law (12), the Kondo energy depends on the value of the gaps δ separating the lowest excitation of the ground state of the system. Figure 3 illustrates this variation for several parts of the phase diagram.

Thus, the parameters of the dynamical group straightforwardly influence the tunnel conductivity of the TQD.

The problem of Kondo tunneling through a TQD in cross and fork geometries (Fig. 2c, f) was analyzed in [35]. It was found that these TQDs with discrete mirror symmetry at odd occupation $N = 1, 3$ possess properties, which were observed earlier only in QD with even occupation $N = 2$. In particular, the Kondo tunneling may be absent in the ground spin-doublet state due to the special symmetry of the electron wave function in the TQD, but, on the

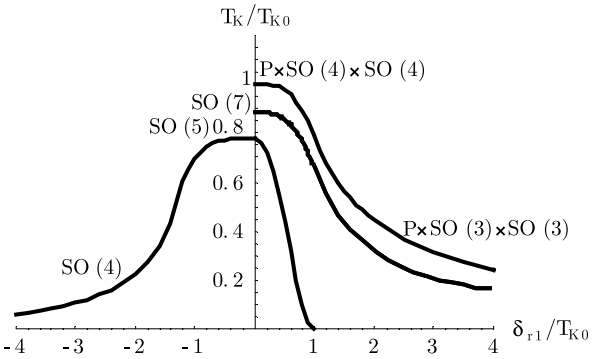

Tunneling Through Quantum Dots with Discrete Symmetries, Figure 3

Variation of Kondo temperature with the difference between gate voltages $\delta_{r1} = v_{gr} - v_{gl}$ applied to the left and right dot

other hand, Kondo-active states are involved in tunneling at higher excitation energies. This behavior is similar to that encountered in DQD with $N = 2$ in a singlet ground state with low-lying triplet excitation [19].

A system consisting of TQD in the fork geometry looks promising also for studying the physics of spin entanglement [52]. In a charge sector $N = 4$, the spins of the electrons located in the side dots are entangled in a singlet state via the doubly occupied central dot.

Consider now the discrete point symmetry C_{3v} of an equilateral triangular TQD shown in Fig. 2d, e. It is superimposed on the spin symmetry of electrons localized in the three wells [9,34]. In the charge sector $N = 1$ an electron may occupy three equivalent positions, so that the total symmetry of the TQD is $SU(6)$ (spin + three colors) but this degeneracy is lifted by the inter-dot tunneling W, so that the orbital triplet is split into a singlet (a) and a doublet (b). The splitting energy is $\Delta_{\rm tr} = E_a - E_b = 3W$. Since $W < 0$, the ground state is the orbital doublet, and the discrete degrees of freedom are quenched, provided $E_{\rm K} \ll \Delta_{\rm tr}$, which is usually the case. However, the orbital degeneracy may be induced by an external magnetic field (see next section).

Two-electron states in the charge sector $N = 2$ are formed by electrons occupying adjacent wells in triplet and singlet states, and this classification is supplemented by the same orbital classification as in the case of $N = 1$. The manifold consists of four levels ordered as $E_{sa} < E_{tb} < E_{ta} < E_{sb}$. As a result, the possibility opens for dynamical mixing of singlet and triplet states with different quantum numbers.

In the three-electron sector $N = 3$, the discrete symmetry of the triangle opens a new possibility of ordering spins. Since the indirect exchange interaction induced by virtual inter-dot tunneling is always antiferromagnetic, there is no possibility of energy minimization by means of simultaneous anti-parallel orientation of all three spins. Because of these frustrations the low-energy spin manifold is formed by three configurations of spin 1/2 localized in one of the three dots, while two other spins are coupled into a spin singlet (see Fig. 4). Only two of these states are linearly independent, so that the ground state is doubly degenerate. This degeneracy (stemming from the discrete point symmetry of a triangle) is lifted by the contact with source and drain (in the geometries of Fig. 2d, f).

The effective spin Hamiltonian describing these configurations has non-Heisenberg form [53]

$$H_{\rm d} = \sum\nolimits_{ij} J_{ij} \boldsymbol{S}_i \cdot \boldsymbol{S}_j + \sum_{ijk} D_{jkl} \boldsymbol{S}_i \cdot [\boldsymbol{S}_j \times \boldsymbol{S}_k] \,. \quad (13)$$

Here J_{ij} is the indirect inter-dot exchange integral, D_{jkl} is triple co-tunneling constant describing cyclic electron exchange between three sites. The quantity $\chi_{ijk} = \boldsymbol{S}_i \cdot [\boldsymbol{S}_j \times \boldsymbol{S}_k]$ is characterized by a quantum number referred to as scalar spin chirality. Lead-dot co-tunneling processes involve also the excited high spin state with $\boldsymbol{S} = 3/2$. As a result of interplay between discrete point symmetry and con-

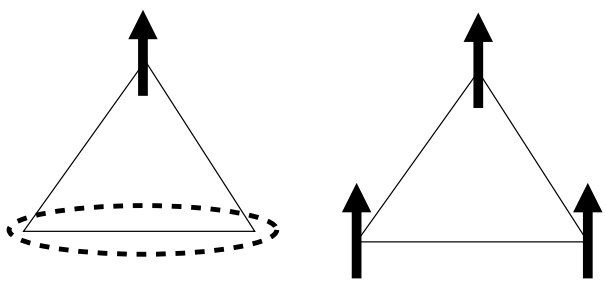

Tunneling Through Quantum Dots with Discrete Symmetries, Figure 4

Possible configurations of three spins in TQD: two spin-polarized states with $S = 1/2$ and high spin state with $S = 3/2$

tinuous spin symmetry of the TQD, dynamical symmetry includes spin variables, spin chirality number and pseudo-spin describing the position of an un-compensated spin in the low-spin state of the TQD.

Tunneling in External Fields

External electric field $\boldsymbol{E}$ and magnetic field $\boldsymbol{B}$ influence tunneling through CQD in two different ways. First, these fields introduce additional energy scales into the dynamics of co-tunneling processes. These are the electrostatic potential associated with the electric field and the Zeeman splitting energy $g\mu_{\mathrm{B}}B$ (g is the gyromagnetic ratio for the electron and μ_{B} is the Bohr magneton). Second, these fields lower the symmetry of the system. Application of an electric field for achieving confinement in two-dimensional systems introduces an additional vector $\boldsymbol{n}$ normal to the confinement plane (Rashba vector [7]). The magnetic field breaks spin rotational symmetry of CQD and introduces chirality of orbital states in case of closed configurations (by breaking time reversal invariance).

At finite bias eV applied to the source and drain, electron tunneling becomes a non-equilibrium process, but in weak enough fields, one may neglect non-equilibrium repopulation of dot states and consider the problem within a quasi equilibrium approximation, where the two chemical potentials μ_{s} and μ_{d} are introduced for the source and drain reservoirs which otherwise remain in an equilibrium state, so that $eV = \mu_{\mathrm{s}} - \mu_{\mathrm{d}}$. The energy acquired by an electron in the source electrode and then accelerated by the potential eV may compensate the exchange energy gap $\lambda < 0$ and thus activate a Kondo tunneling processes which is otherwise quenched in the singlet ground state of the CQD at $kT \ll |\lambda|$. As a result a finite bias anomaly at $eV = \lambda$ shows up in the tunneling conductance instead of zero bias anomaly [30]. Such behavior was observed experimentally in tunneling through single-wall carbon nanotubes [44].

The exchange energy gap may also be compensated by the Zeeman splitting energy under an external magnetic field. Indeed, in a "resonance" field defined as $g\mu_{\mathrm{B}}B_{\mathrm{r}} - |\lambda| \sim E_{\mathrm{K}}$ transitions between the singlet and "up" projection of the triplet are involved in the dynamical symmetry, so that one may introduce operators $P^{+} = \sqrt{2}\left|T_{\uparrow}\right\rangle\langle S|\,, P_z = \left|T_{\uparrow}\right\rangle\left\langle T_{\uparrow}\right| - |S\rangle\langle S|$ in analogy with (4). In these resonance conditions, the effective co-tunneling Hamiltonian has the form $H_{\mathrm{cot}} = J\boldsymbol{P}\cdot\boldsymbol{s}$. It describes Kondo tunneling due to triplet-singlet transitions induced by an external magnetic field [48]. This type of Kondo effect was also observed in single-wall carbon nano-tubes [43].

In cases where the tunneling channels and/or the CQD configuration form closed loops [Figs. 1b, 2b, d], the "which pass" situation arises in the trajectories leading from the source to the drain. In the absence of a magnetic field, the currents through the tunneling channels simply add, but in the presence of a perpendicular magnetic field $B^{\perp}$ a more complicated superposition pattern arises due to the Aharonov–Bohm effect. Besides, the field $B^{\perp}$ introduces charge chirality into the scheme of classification of electron states of the CQD [9,34]. This type of chirality arises because an electron acquires a $U(1)$ gauge phase $\phi = \Phi/3$ at each tunneling hopping event. In an anticlockwise direction between the vertices of the equilateral triangle the hopping amplitude W is then modified as,

$$W \to W\exp(i\phi)\,, \tag{14}$$

where Φ is the magnetic flux through the TQD. As a result, the electron spectrum displays a rich pattern of degenerate levels. Instead of the levels $E_{a,b}$ there is now a magnetic field dependent spectrum $E_{\mathrm{K}}(\phi)$. For example, in a charge sector $N = 1$,

$$E_{\mathrm{K}}(\phi) = \varepsilon - 2W\cos(K - \phi)\,, \quad K = 0, \pm 1\,. \tag{15}$$

As a result, accidental orbital degeneracy arises at all $\Phi_r = (2n+1)/2$ (half-integer magnetic flux quanta) as can be seen in Fig. 5 (upper panel). This degeneracy transforms the spin symmetry $SU(2)$ of the TQD into a spin+orbital symmetry $SU(4)$. In accordance with the general law (11), the Kondo temperature, and hence the tunnel conductance $G \sim G_0 \ln^{-2}(T/T_{\mathrm{K}})$ at $T \gg T_{\mathrm{K}}$ have peaks around Φ_r (see Fig. 5, lower panel).

A similar level crossing scenario takes place at $N = 2$, but in this case one of the orbital degeneracies is accompanied by the singlet/triplet degeneracy [9]. At $N = 3$, the amplitude of energy oscillations in the magnetic field, $E_{N=3}(\Phi)$ are weak because all the orbital states are occupied and oscillations of different states nearly cancel each other. However, the level crossing between the states with $S = 3/2$ and $S = 1/2$ split due to the Zeeman effect [9,17] results in magnetic-field-induced Kondo tunneling similar to that discussed in [43].

Aharonov–Bohm interference arises, when an electron injected from the source electrode passes through two channels in the presence of a magnetic flux Φ and the two trajectories interfere in the drain electrode. The interplay between the Kondo and the Aharonov–Bohm effects may be directly seen in the tunneling conductance. Such a regime may be realized both in DQD [21] (Fig. 1b without l-r tunneling channel) and TQD [34] (Fig. 2d). In the former case the states in the DQD possessing a discrete l-r

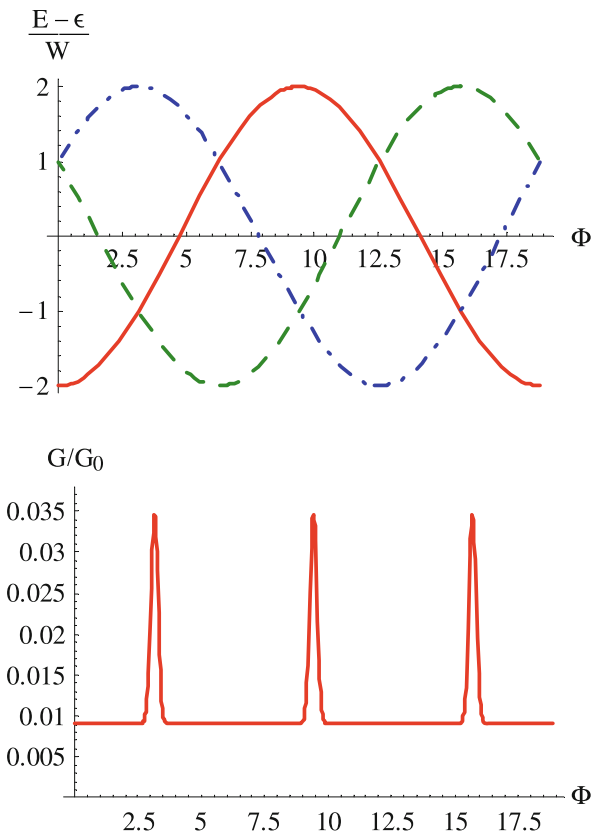

Tunneling Through Quantum Dots with Discrete Symmetries, Figure 5

***Upper panel*: Evolution of the energy levels E_0 (*solid line*) and $E_\pm$ (*dashed* and *dot-dashed lines*, respectively). *Lower panel*: The corresponding evolution of the tunnel conductance (here $G_0 = \pi e^2/\hbar$ is a quantum of conductance)**

reflection symmetry are classified as even (e) and odd (o) relative to this reflection plane, so that the zero-temperature value (unitarity limit of the Kondo effect) of the tunnel conductance is determined by the difference between transparencies of the two channels,

$$G(T=0)/G_0 = \sin^2\left\{\frac{\pi}{2}\left[\langle n_e\rangle - \langle n_o\rangle\right]\right\}\,, \tag{16}$$

where $\langle n_{e,o}\rangle$ are the ground state occupation numbers in the even and odd channels, respectively. Since different $U(1)$ gauge phases (14) are acquired by the electron hopping parameters W in the two channels, interference of these channels results in a magnetic field dependence conductance $G_\phi(T=0)$. Besides, the Kondo energy E_K itself becomes a function of the magnetic flux, because the effective exchange parameter depends on the magnetic field ($J \sim W^2$) in H_{cot} (7). The resulting Aharonov–Bohm dependence is $J \sim \cos\Phi$, and this destructive interference suppresses Kondo tunneling at magnetic fields corresponding to a half-integer magnetic flux quantum. In the TQD shown in Fig. 2d, the ground state is an orbital singlet, and the most striking manifestation of an Aharonov–Bohm destructive interference pattern is due to the dependence $J(\Phi)$, which has a slightly more complicated form than that in DQD. In any case, electron tunneling through the "Kondo–Aharonov–Bohm interferometer" is completely suppressed at some values of the magnetic flux Φ.

Another mechanism exhibiting the influence of external fields on tunneling properties of CQD is the so-called Thomas–Rashba precession of the magnetic quantization axis initiated by an external electric field that confines the electrons in two dimensions [7,50]. In a system with relativistic spin-orbit interaction, the generic form of the Thomas–Rashba interaction is

$$H_{TR} = \alpha \boldsymbol{n}\cdot[\boldsymbol{S}\times\boldsymbol{p}] \tag{17}$$

where α is a spin-orbit coupling constant, $\boldsymbol{n}$ is a unit vector characterizing the direction of confining electric field and $\boldsymbol{p}$ is the electron momentum operator. If this relativistic interaction exists in the leads (reservoirs), it results in the appearance of a $\boldsymbol{p}$-dependent spin quantization axis and splitting of the electron dispersion law $\varepsilon_s(\boldsymbol{p})$ into two branches characterized by "spirality" quantum number s. This may cause an indirect Ruderman–Kittel–Katsuya–Yoshida (RKKY) interaction exchange between two electrons in DQD with $N=2$ due to lead-dot tunneling which results in a complicated spin-coupling Hamiltonian of the form [20],

$$H_{RKKY} = F_{12}\boldsymbol{S}_1\cdot\boldsymbol{S}_2(\theta_{12})\,, \tag{18}$$

where F_{12} is the RKKY coupling constant and θ_{12} is the angle between two twisted quantization axes. The origin of this twisting is the Thomas–Rashba precession in the leads. The mismatch between two quantization axes may be represented as an effective Dzyaloshinskii–Moriya interaction [cf. (13)] plus an additional Ising-type term

$$\begin{aligned}\boldsymbol{S}_1\cdot\boldsymbol{S}_2(\theta_{12}) = \boldsymbol{S}_1\cdot\boldsymbol{S}_2 + \sin\theta_{12}[\boldsymbol{S}_1\times\boldsymbol{S}_2]\\ + (1-\cos\theta_{12})S_1^y S_2^y\,.\end{aligned} \tag{19}$$

If a perpendicular magnetic field $B^\perp$ is then switched on, two types of quantum interference arise in the Kondo tunneling through DQD due to Aharonov–Bohm charge gauge phase ϕ induced by the magnetic field and Aharonov–Casher effect due to spin gauge phase χ induced by the electric field. Both phases control the conductance $G(\phi,\chi)$ through the DQD in this regime [1].

Connections with Other Nano-objects and Future Prospects

The physics of tunnel spectroscopy of CQD was developed in parallel with that of tunnel spectroscopy of single molecular devices. Generic similarities between strong correlation effects in bulk materials, molecular complexes and quantum dots were noticed by many authors (see, e. g., [13,26,40,47]). Experimentally, tunnel devices may be prepared in many ways: tunnel spectra of molecules adsorbed on metallic layers may be measured by means of scanning tunneling microscopy. In this case, the nano-tip of the microscope and metallic substrate play the role of source and drain electrodes. Otherwise, the molecule may be suspended between two metallic electrodes by electrochemical methods. All characteristic features of tunneling through quantum dots (Coulomb blockade, Kondo effect, Fano effect etc) are observed in molecular tunneling spectroscopy as well.

Single-wall carbon nanotubes (SWCN) may serve as part of a "bridge" between artificial and natural molecular complexes. These long cylindrical macromolecules usually behave as quasi one-dimensional metallic or semiconductor wires, but if one confines the electrons in a segment of SWCN by imposing a pair of electrodes on it, the electron wave functions within this segment are quantized in all directions and the system as a whole behaves like a quantum dot with characteristic single-electron tunneling behavior due to the Coulomb blockade [54] and a Kondo-like zero bias anomaly due to spin screening in the metallic electrodes [43]. Quantum dots formed in SWCN possess the discrete symmetry inherited from two sub-lattice crystal structures of prototype graphene sheets. The spatially quantized states in these QDs are doubly degenerate [2] and this degeneracy manifests itself as a $SU(2) \to SU(4)$ crossover in Kondo tunneling similar to that in DQD [38]. A pair of nanotubes may form a two-channel device in the form of a ring, where both the Fano effect and Aharonov–Bohm effect are involved in electron tunneling [3].

Other representatives of the carbon family, like various modifications of the fullerene molecules also demonstrate those features characteristic for nano-size quantum dots. Coulomb blockade features [45] and Kondo anomalies [59] were observed in conventional C_{60} molecules deposited onto pair of connected gold electrodes. There are some proposals to form double quantum dots from C_{140} molecule having the form of a dumb-bell [46] or pair of endofullerene molecules, like GdC_{82} [5], where a behavior similar to that of DQD is expected. It should be noticed, however that in real molecules (unlike artificial ones), phonon emission/absorption processes assist single electron tunneling.

Molecular trimers, whose behavior should remind that of TQD may be absorbed on metallic surfaces. The first example of such system is chromium trimers on gold surfaces [22]. These trimers may be both linear and triangular, so from the point of view of Kondo tunneling their behavior should be similar to that of magnetic TQD with $N = 3$.

Another family of real molecules possessing discrete point symmetries was studied both experimentally and theoretically within the same context. These are molecular complexes containing transition and rare-earth metal ions secluded in cages formed by CH, CN and other organic radicals. These molecular cages are in direct contact with the electrodes, whereas the magnetic ions form spin multiplets. Endofullerenes and lanthanocene families are the simplest examples of such complexes. It was noticed [26,27] that their behavior in quantum tunneling should be similar to that of DQD in T-shape geometry (Fig. 1c) There are also molecular complexes containing several cages, which form 2×2, 3×3 and more complicated grids, which also may be incorporated into tunnel spectroscopy [57]. In these complexes the complicated structure of the spin multiplet is influenced by magnetic anisotropy due to the discrete symmetry of a grid.

Finally, one should mention the family of single-molecule magnets, namely, large molecular complexes containing about ten transition-metal atoms (Fe_8, Mn_{12}, etc). In these molecules, the magnetic moments of individual spins are added, forming a nano-object with high spin and strong magnetic anisotropy. The discrete molecular symmetry predetermines the complicated quantum dynamics of this high spin object. Complicated spin selection rules for the Kondo effect in some cases suppress two-electron co-tunneling, and reversal $M_S \to -M_S$ of spin projection appears only as a fourth order process [58]. Kondo tunneling through magnetic molecules in an external magnetic field displays very complicated structure. Moreover, spin reversal is also characterized by a Berry phase due to multiple excited level crossings [14]. Tunnel transport measurements can directly probe the Berry phase of individual single-molecule magnets.

Future prospects of quantum tunneling through nano-objects with discrete symmetries are closely connected with modern advance in quantum electronics and the quest for manipulating spin systems (spintronics). The most important challenge for theory, experiment and material science engineering concerns the fabrication and study of macroscopic networks formed from quantum dots and molecular complexes. In these composite struc-

tures, the dynamical and tunneling properties of each individual nano-object will serve as building blocks for complex systems, in which the information may be stored and read out on the quantum level in a controllable way.

Bibliography

Primary Literature

1. Aono T (2007) Two-impurity Kondo problem under Aharonov–Bohm and Aharonov–Casher effects. Phys Rev B 76:073304
2. Babic B, Kontos T, Schonenberger C (2004) Kondo effect in carbon nanotubes at half filling. Phys Rev B 70:235419
3. Babic B, Schonenberger C (2004) Observation of Fano resonances in single-wall carbon nanotubes. Phys Rev B 70:195408
4. Bawendi MG et al (1990) Electronic structure and photoexcited-carrier dynamics in nanometer-size CdSe clusters. Phys Rev Lett 65:1623–1626
5. Bergeret FS et al (2006) Interplay between Josephson effect and magnetic interactions in double quantum dots. Phys Rev B 74:132505
6. Bimberg D, Grundman M, Ledentsov NN (1999) Quantum Dot Heterostructures. Wiley, New York
7. Bychkov YA, Rashba EI (1984) Oscillatory effects and the magnetic susceptibility of carriers in inversiojn layers. J Phys C: Solid State Phys 17:6039–6045
8. Craig NJ et al (2004) Tunable nonlocal spin control in coupled-quantum-dot system. Science 304:565–567
9. Delgado F et al (2007) Theory of spin, electronic and transport properties of the lateral triplet quantum dot molecule in a magnetic field. Phys Rev B 76:115332
10. Ekimov AI, Onushchenko AA (1984) Size quantization of the electron energy spectrum in a microscopic semiconductor crystal. JETP Lett 40:1136–1139
11. Eto M (2001) Mean-field theory of the Kondo effect in quantum dots with an even number of electrons. Phys Rev B 64:085322
12. Eto M (2005) Enhancement of Kondo effect in Multilevel quantum dots. J Phys Soc Jpn 74:95–102
13. Fulde P (1995) Electron Correlations in Molecules and Solids, Chap 12, 3rd edn. Springer, Heidelberg
14. Garg A (1993) Topologically quenched tunnel splitting in spin systems without Kramers degeneracy. Europhys Lett 22: 205–210
15. Glazman LI, Raikh ME (1988) Resonant Kondo transparency of a barrier with quasilocal impurity states. JETP Lett 47: 452–455
16. Guadreau L et al (2006) Stability diagram of a few-electron triple dot. Phys Rev Lett 97:036807
17. Nojiri H, Ishikawa E, Yamase T (2005) Effect of spin chirality on magnetization in triangular rings. Progr Theor Phys Suppl No 159:292–296
18. Hofmann F et al (1995) Single electron switching in a parallel quantum dot. Phys Rev B 51:13872–13875
19. Hofstatter W, Schoeller H (2002) Quantum phase transitions in a multilevel dot. Phys Rev Lett 88:016803
20. Imamura H, Bruno P, Utsumi Y (2004) Twisted exchange interaction between localized spins embedded in a one- and two-dimensional electron gas with Rashba spin-orbit coupling. Phys Rev B 69:121303
21. Izumida W, Sakai O, Shimizu Y (1997) Many body effects on electron tunneling through quantum dots in and Aharonov–Bohm circuit. J Phys Soc Jpn 66:717–726
22. Jamneala T, Madhavan V, Crommie MF (2001) Kondo response of a single antiferromagnetic chromium trimer. Phys Rev Lett 87:256804
23. Joault B, Santoro G, Tagliacozzo A (2000) Sequential magnetotunneling in a vertical quantum dot tuned at the crossing to higher spin states. Phys Rev B 61:10242–10246
24. Kang K et al (2001) Anti-Kondo resonance in transport through a quantum wire with a side coupled quantum dot. Phys Rev B 63:113304
25. Kastner MA (1993) Artificial atoms. Phys Today 46:24–31
26. Kikoin K, Avishai Y (2001) Kondo tunneling through real and artificial molecules. Phys Rev Lett 86:2090–2093
27. Kikoin K, Avishai Y (2002) Kondo singlet versus Zhang-Rice and Heitler-London singlets. Physica B 312–313:165–166
28. Kikoin K, Avishai Y, Kiselev MN (2006) Dynamical symmetries in nanophysics. In: Nanophysics, Nanoclusters and Nanodevices. Nova Publishers, New York, pp 39–86
29. Kim TS, Hershfield S (2001) Suppression of current in transport through parallel double quantum dots. Phys Rev B 63:245326
30. Kiselev MN, Kikoin K, Molenkamp LW (2003) Resonance Kondo tunneling through a double quantum dot at finite bias. Phys Rev B 68:155323
31. Kondo J (1964) Resistance minimum in dilute magnetic alloys. Progr Theor Phys 32:37–49
32. Kowenhoven LP, Austing DG, Tarucha S (2001) Few-electron quantum dots. Rep Progr Phys 64:701–736
33. Kuzmenko T, Kikoin K, Avishai Y (2004) Kondo effect in systems with dynamical symmetries. Phys Rev B 69:195109
34. Kuzmenko T, Kikoin K, Avishai Y (2006) Magnetically tunable Kondo–Aharonov–Bohm effect in triangular quantum dot. Phys Rev Lett 96:046601
35. Kuzmenko T, Kikoin K, Avishai Y (2006) Tunneling through triple quantum dots with mirror symmetry. Phys Rev B 73:235310
36. Leonhard D et al (1993) Direct formation of quantum-size dots from uniform coherent islands of InGaAs on GaAs surfaces. Appl Phys Lett 63:3203–3205
37. Loss D, DiVincenco DP (1998) Quantum computation with quantum dots. Phys Rev A 57:120–126
38. Makarovski A et al (2007) *SU*(4) and *SU*(2) Kondo effects in carbon nanotube quantum dots. Phys Rev B 75:241407
39. Matveev KA, Larkin AI (1992) Interaction-induced threshold singularities in tunneling via localized levels. Phys Rev B 46:15337–15347
40. Mitra A, Aleiner I, Millis AJ (2004) Phonon effects in molecular transistors: quantum and classical treatment. Phys Rev B 69:245302
41. Molenkamp LW et al (1995) Scaling of the Coulomb energy due to quantum fluctuations in the charge on a quantum dot. Phys Rev Lett 75:4282–4285
42. Ng TK, Lee PA (1988) On-site Coulomb repulsion and resonant tunneling. Phys Rev Lett 61:1768–1771
43. Nugard J et al (2000) Kondo physics in carbon nanotubes. Nature 408:342–345
44. Paaske J et al (2006) Nonequilibrium singlet-triplet Kondo effect in carbon nanotubres. Nat Phys 2:460–464
45. Park H et al (2000) Nanomechanical oscillations in a single- C_{60} transistor. Nature 407:57–60

46. Pasupathy AN et al (2005) Vibration assisted electron tunneling in C_{140} single electron transistors. Nano Lett 5:203–207
47. Plihal M, Gadzuk JW (2001) Nonequilibrium theory of scanning tunneling spectroscopy via adsorbate resonances. Phys Rev B 63:085404
48. Pustilnik M, Avishai Y, Kikoin (2000) Quantum dots with even number of electrons: Kondo effect in a finite magnetic field. Phys Rev Lett 84:1756–1759
49. Pustilnik M, Glazman LI (2001) Kondo effect induced by magnetic field. Phys Rev B 64:45328
50. Rashba EI (2006) Modern trends in semiconductor Spintronics. In: Ivanov AL, Tikhodeev SG (eds) Problems of Condensed Matter Physics. Clarendon Press, Oxford, pp 188–198
51. Read MA (1993) Quantum dots. Sci Amer 268:118–123
52. Saraga DS, Loss D (2003) Spin-entangled currents created by a triple quantum dot. Phys Rev Lett 90:166803
53. Scarola VW, Das Sarma S (2005) Exchange gate in solid-state spin quantum computation: the applicability of the Heisenberg model. Phys Rev A 71:032340
54. Tans S et al (1998) Electron-electron correlations in carbon nanotubes. Nature 394:761–764
55. Teichert C (2002) Self-organization of nanostructures in semiconductor heteroepitaxy. Phys Repts 365:335–432
56. Vidan A et al (2004) Triple quantum dot charging rectifier. Appl Phys Lett 85:3602–3604
57. Waldmann O, Zhao L, Thompson LK (2002) Field-dependent anisotropy change in a supramolecular Mn(II)-[3 × 3] grid. Phys Rev Lett 88:066401
58. Wegewijs M et al (2007) Magnetotransport through single-molecule magnets: Kondo peaks, zero-bias dips, molecular symmetry and Berry's phase. New J Phys 9:344
59. Yu LH, Natelson D (2004) The Kondo effect in C_{60} single-molecule transistors. Nano Lett 4:79–83

Books and Reviews

Cuniberti G, Fagas G, Richter K (eds) (2005) Molecular electronics. Lecture Notes in Physics, vol 680. Springer, New York
Englefield MJ (1972) Group Theory and the Coulomb Problem. Wiley, New York
Kouwenhoven LP, Glazman LI (2001) Revival of the Kondo effect. Phys World 14:33–38
Natelson D (2006) Handbook of Organic Electronics and Photonics. American Scientific Publishers, Valencia
Sachrajda PH, Ciorga M (2003) Nano-spintronics with lateral quantum dots. Kluwer, Boston

Two-Sided Matching Models

Marilda Sotomayor[1,2], Ömer Özak[2]
[1] Department of Economics, University of São Paulo/SP, São Paulo, Brazil
[2] Department of Economics, Brown University, Providence, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Discrete Two-Sided Matching Models
Continuous Two-Sided Matching Model with Additively Separable Utility Functions
Hybrid One-to-One Matching Model
Incentives
Future Directions
Bibliography

Glossary

Two-sided matching model is a game theoretical model whose elements are (i) two disjoint and finite sets of agents: F with m elements, and W with n elements, referred to as the sides of the matching model; (ii) the structure of agents' preferences and (iii) the agents' quotas.
The rules of the game determine the feasible outcomes. The main activity of the agents from one set is to form partnerships with the agents on the other set. Players derive their payoffs from the set of partnerships they form. The agents belonging to F and W are called F-agents and W-agents, respectively.

Quota of an agent in a two-sided matching model is the maximum number of partnerships an agent is allowed to form. When every participant can form one partnership at most the matching model is called one-to-one. If only the players of one of the sides can form more than one partnership the matching model is said to be many-to-one. Otherwise the matching model is many-to-many.

Allowable set of partners for $f \in F$ with quota $r(f)$ is a family of elements of $F \cup W$ with k distinct W-agents, $0 \leqslant k \leqslant r(f)$, and $r(f) - k$ repetitions of f.

Discrete two-sided matching model In the discrete two-sided matching models agents have preferences over allowable sets of partners. The allowable sets of partners for f of the type $\{w, f, \ldots, f\}$ are identified with the individual agent $w \in W$ and the allowable set of partners $\{f \ldots f\}$ is identified with f. Under this identification, agent w is *acceptable* to agent f if and only if f likes w as well as him/her/it self. Similar definitions and identifications apply to an agent $w \in W$. These preferences are transitive and complete, so they can be represented by ordered lists of preferences. The model can then be described by (F, W, P, r, s), where P is the profile of preferences and r and s are the arrays of quotas for the F-agents and W-agents, respectively.

Continuous two-sided matching model In this model the structure of preferences is given by utility functions

which are continuous in some money variable which varies continuously in the set of real numbers. A particular case is obtained when agents place a monetary value on each possible partner or on each possible set of partners.

Hybrid two-sided matching model is a unification of the discrete and the continuous models. It is obtained by allowing the agents of both markets to trade with each other in the same market.

Matching μ in a two-sided matching model with sides F and W is a function that maps every agent into an allowable set of partners for him/her/it, such that f is in $\mu(w)$ if and only if w is in $\mu(f)$, for every $(f, w) \in F \times W$. If we relax this condition the function is called a *pre-matching*.

A matching describes the set of partnerships of the type (f, w), (f, f) or (w, w), with $f \in F$ and $w \in W$, formed by the agents. We say that a player that does not enter any partnership is unmatched. Agents compare two matchings by comparing the two allowable sets of partners they obtain.

Feasible assignment for a two-sided matching model with sides F and W is an $m \times n$ matrix $x = (x_{fw})$ whose entries are zeros or ones such that $\sum_f x_{fw} \leqslant s(w)$ for all $w \in W$ and $\sum_w x_{fw} \leqslant r(f)$ for all $f \in F$. We say that $x_{fw} = 1$ if f and w form a partnership and $x_{fw} = 0$ otherwise. A feasible assignment x corresponds to a matching μ which matches f to w if and only if $x_{fw} = 1$. Thus, if $\sum_f x_{fw} = 0$ then w is unassigned at x or, equivalently, unmatched at μ, and if $\sum_w x_{fw} = 0$, then f is likewise unassigned at x or, equivalently, unmatched at μ.

Responsive preference in a discrete two-sided matching model with sides F and W. Agent $f \in F$ has a responsive preference relation over allowable sets of partners if whenever (i) A and B are two allowable sets of partners for player f; (ii) j and k are two elements of $W \cup \{f\}$ and (iii) $A = B \cup \{w\} \setminus \{w'\}$ with $w \notin B$ and $w' \in B$, then f prefers A to B if and only if f prefers w to w'. Similarly we define the responsive preference for $w \in W$.

$r(f)$-Separable preference in a discrete two-sided matching model with sides F and W. Agent $f \in F$ with a quota of $r(f)$ has a $r(f)$-separable preference relation over allowable sets of partners if whenever $A = B \cup \{w\} \setminus \{f\}$ with $w \notin B$ and $f \in B$, then f prefers A to B if and only if f prefers w to f. Similarly we define $s(w)$-separable preference for $w \in W$.

Maximin preferences in a discrete two-sided matching model with sides F and W. Agent $f \in F$ with a quota of $r(f)$ has a maximin preference relation over allowable sets of partners if whenever two allowable sets C and C' contained in W, such that f prefers C' to C and no w in C is unacceptable to f, then a) all of C' are acceptable to f and b) if $|C| = r(f)$, then the least preferred worker in $C' - C$ is preferred by f to the least preferred worker in $C - C'$. Similarly we define maximin preference for $w \in W$.

Choice set of $f \in F$ from $A \subseteq W(Ch_f(A))$ in a discrete two-sided matching model with sides F and W. Let $B = \{A' | A'$ is an allowable set of partners for f and $A' \cap W$ is contained in $A\}$. Then, $A' \in Ch_f(A)$ if and only if $A' \in B$ and f likes A' at least as well as A'', for all $A'' \in B$. Similarly we define $Ch_w(A)$ for $w \in W$ and $A \subseteq F$.

Substitutable preferences in a discrete two-sided matching model with sides F and W. Agent $f \in F$ has a substitutable preference relation over allowable sets of partners if whenever $A \subseteq W$ and $B \subseteq W$ are such that $A \cap B = \phi$ then (i) for all $S' \in Ch_f(A \cup B)$ there is some $S \in Ch_f(A)$ such that $S' \cap A \subseteq S$ and (ii) for all $S \in Ch_f(A)$ there is some $S' \in Ch_f(A \cup B)$ such that $S' \cap A \subseteq S$. If an agent's preference is responsive then it is substitutable.

When preferences are strict, conditions (i) and (ii) are equivalent to requiring that if $Ch_f(A \cup B) = S'$ then $S' \cap A \subseteq Ch_f(A)$. This concept is similarly defined for $w \in W$.

Strongly substitutable preferences in a discrete two-sided matching model with sides F and W. Agent $f \in F$ has a strongly substitutable preference relation over allowable sets of partners if for every pair of allowable sets of partners A and B such that $A >_f B$, if $w \in Ch_f(W \cap A \cup \{w\})$, then $w \in Ch_f(W \cap B \cup \{w\})$. This is a stronger condition than substitutability and responsiveness.

Additively separable preferences in a discrete two-sided matching model with sides F and W. Agent $f \in F$ has additively separable preferences if he/she/it assigns a nonnegative number a_{fw} to each $w \in W$ and assigns the value $v(A) = \sum_{w \in A} a_{fw}$ to each allowable set A of partners for f. Agent f compares two allowable sets by comparing the values of these sets. This concept is similarly defined for $w \in W$. If the agents have additively separable preferences we can think that if a partnership $(f, w) \in F \times W$ is formed then the partners participate in some joint activity that generates a payoff a_{fw} for player f and b_{fw} for player w. These numbers are fixed, i. e., they are not negotiable. If the preferences of the agents are additively separable then they are responsive. The converse is not true (see Kraft, Pratt and Seidenberg [47]).

***T*-map** in a discrete two-sided matching model with sides F and W is defined as follows. For every pre-matching μ, let $T(\mu(f)) = Ch_f(U(f,\mu))$ for all $f \in F$, where $U(f,\mu) = \{w \in W | f \in Ch_w(\mu(w) \cup \{f\})\}$. Similarly, $T(\mu(w)) = Ch_w(U(w,\mu))$ for all $f \in F$, where $U(w,\mu) = \{f \in F | w \in Ch_f(\mu(f) \cup \{w\})\}$.

Lattice property A set L endowed with a partial order relation $\geqslant$ has the lattice property if $\sup\{x,y\} \equiv x \vee y$ and $\inf\{x,y\} \equiv x \wedge y$ are in L, for all $x, x' \in L$. The lattice is complete if all its subsets have a supremum and an infimum. (See Birkhoff [11]).

Pareto-optimal matching A feasible matching μ is Pareto-optimal if there is no feasible matching which is weakly preferred to μ by all players and it is strictly preferred by at least one of them.

Outcome For the discrete two-sided matching models the outcome is a matching or at least corresponds to a matching; for the continuous two-sided matching models the outcome specifies a payoff for each agent and a matching.

Stable outcome It is the natural solution concept for a two-sided matching model. It is also referred as set-wise-stable outcome. See the definition bellow.

***F*-optimal stable matching (respectively, payoff)** for a discrete (respectively, continuous) two-sided matching model is the stable matching (respectively, payoff) which is weakly preferred by every agent in F. Similarly we define the W-optimal stable matching (respectively, payoff).

Achievable mate for agent y in a discrete two-sided matching model is any y's of partner under some stable matching.

Matching mechanism For the discrete two-sided matching models, a matching mechanism is a function h whose range is the set of all possible inputs $X = (F, W, P, r, s)$, and whose output $h(X)$ is a matching for X.

Stable matching mechanism It is a matching mechanism h such that $h(X)$ is always stable for the market X. If $h(X)$ always produces the F-optimal stable matching for X then it is called an *F-optimal stable matching mechanism,* and so on.

Revelation mechanism Given the discrete two-sided matching model (F, W, P, r, s), a revelation mechanism is the restriction of a matching mechanism h to the set of discrete two-sided matching markets (F, W, Q, r, s) where the sets of agents and quotas are fixed.

Revelation game It is the strategic game induced by a revelation mechanism for the discrete two-sided matching (F, W, P, r, s): the set of players is given by the union of F and W; a strategy of player j is any possible list of preferences $Q(j)$ that player j can state; the outcome function is given by the mechanism h and the preferences of the players over the set of outcomes are determined by P.

Sincere strategy for a player j in a revelation game is the true list of preferences $P(j)$.

Manipulable mechanism A mechanism h is *manipulable* or it is *not strategy-proof* if in some revelation game induced by h, stating the true preferences is not a dominant strategy for at least one player. A mechanism h is *collectively manipulable* if in some revelation game induced by h, there is a coalition whose members can be better off by misrepresenting its preferences.

Rematching proof equilibrium is a Nash equilibrium profile from which no pair of players $(f, w) \in F \times W$ can profitably deviate given that the other players do not change their strategies.

Truncation Let $P(a)$ be the a's preference list over individuals for a discrete two-sided matching model. For $a \in F \cup W$, the list of preferences $Q(a)$ over individuals is a truncation of $P(a)$ at some agent b if $Q(a)$ keeps the same ordering as $P(a)$ but ranks as unacceptable all agents which are ranked below b.

Definition of the Subject

This article describes the basic elements of the cooperative and non-cooperative approaches for two-sided matching models and analyzes the fundamental differences and similarities between some of these models.

Basic Definitions

Feasible outcome is an outcome that is specified by the rules of the game. In the discrete case, a feasible outcome is a feasible matching μ or at least corresponds to a feasible matching. The usual definition is the following. *The matching μ is feasible if it matches every agent to an allowable set of partners and $\mu(f) \in Ch_f(\mu(f) \cap W)$ and $\mu(w) \in Ch_w(\mu(w) \cap F)$ for every $(f, w) \in F \times W$.* Then, if preferences are responsive, every matched pair is mutually acceptable. An implication of this definition is that a feasible outcome is always individually rational.

In the continuous case, the rules of the game may specify, for example, whether the agents negotiate their payoffs individually within each partnership or if they negotiate in blocks. In the former case a feasible outcome specifies an array of individual payoffs for each player, indexed according to the partnerships formed under the corresponding matching. In the latter case

the feasible outcome only specifies a total payoff for each agent.

Corewise-stability is a solution concept that assigns to each two-sided matching model the set of feasible outcomes which are not *dominated* by any feasible outcome via a coalition.

The feasible outcome x *dominates* the feasible outcome y via the coalition $S \neq \phi$, if (i) every player in S prefers x to y and (ii) if $j \in S$ then all of j's partners under x belong to S. Coalition S is said to *block* the outcome y.

It is a natural solution concept for the one-to-one matching models and for the continuous many-to-one matching models.

Strong-corewise-stability is a solution concept that assigns to each two-sided matching model the set of feasible outcomes which are not *weakly dominated* by any feasible outcome via a coalition.

The feasible outcome x *weakly dominates* the feasible outcome y via the coalition $S \neq \phi$, if (i) all players in S weakly prefer x to y and at least one of them strictly prefers x to y; (ii) if $j \in S$ then all of j's partners under x belong to S. Coalition S is said to *weakly block* the outcome y.

Strong-corewise-stability is a natural solution concept for the discrete many-to-one matching models.

Pairwise-stability is a solution concept that assigns to each two-sided matching model the set of feasible outcomes which are not *quasi-dominated* by any feasible outcome via a pair of agents $(f, w) \in F \times W$.

For the discrete case an outcome can be identified with a matching. We say that the feasible matching x *quasi-dominates* the feasible matching y via the coalition $S \neq \phi$, if (i) every player in S prefers the matching x to the matching y and (ii) if $j \in S$ and $k \in x(j)$ then $k \in S \cup y(j)$.

For the continuous case in which agents negotiate their payoffs individually with each partner, the feasible outcome x *quasi-dominates* the feasible outcome y via the coalition $S \neq \phi$, if (i) every player in S gets a higher total payoff under x than under y and (ii) if $j \in S$ and $k \in x(j)$ then $k \in S$ or $k \in y(j)$, in which case, k keeps the same individual payment obtained with j under y.

For the continuous case in which agents cannot negotiate their individual payments, the definition of *quasi-dominance* is equivalent to that of dominance. In any case coalition S is said to *destabilize* the outcome y. Thus, for the discrete models, for example, a matching μ is pairwise-stable if it is feasible and it is not destabilized by any pair $(f, w) \in F \times W$. The pair of agents (f, w) destabilizes the matching μ if these agents prefer each other to some of their current mates. Pairwise-stability is a natural solution concept for the marriage and the college admission models as well as for continuous matching models in which the agents negotiate their individual payoffs.

Setwise-stability is the solution concept that assigns to each two-sided matching model the set of feasible outcomes which are not *quasi-dominated* by any feasible outcome via a coalition.

It is a generalization of the group stability concept defined by Roth [65] for the college admissions model. An attempt to extend the concept of group stability to a discrete many-to-many matching market with substitutable and strict preferences was not successful in Roth [62]. The concept introduced in that paper is equivalent to pairwise stability. It turns out that pairwise-stable matchings may be blocked by coalitions of size greater than two in this model. In fact, an example of a pairwise-stable matching that is corewise-unstable is presented in Blair [12] and another one in Sotomayor [80]. Thus, the pairwise-stability concept cannot be regarded as the natural cooperative solution concept for this model.

Setwise-stability regarded as a new cooperative equilibrium concept, different from the core concept, was obtained for the first time in Sotomayor [75] in a many-to-many matching model with additively separable utilities. In this model the set of setwise-stable outcomes equals the set of pairwise-stable outcomes and may be smaller than the core.

Setwise-stability concept was introduced in Sotomayor [80] as a refinement of the core concept and as the natural cooperative equilibrium concept for a two-sided matching model. In the discrete many-to-many matching models it is stronger than pairwise-stability plus corewise-stability (see Example 3 below). Since the essential coalitions in the marriage and in the college admission models are pairs of players made up of one agent of each side of the market, a setwise-stable matching is a pairwise-stable matching in these models. If the preferences are strict, the pairwise-stable matchings are the corewise-stable matchings in the marriage model and are the strong corewise-stable matchings in the college admissions model.

A Brief Historical Account

The theory of the two-sided matching markets was born in 1962, year of the publication of the seminal paper by Gale and Shapley, who formulated the stable matching problem for the marriage and the college admissions markets. The

problem of "college admissions" as described in that paper involves a set of colleges and a set of students. Each student lists in order of strict preference those colleges he/she wishes to attend and each college lists in order of strict preference those students it is willing to admit. Furthermore, each college has a quota, representing the maximum number of students it can admit. The problem is then to devise some allocation procedure of students to colleges in a way that takes account of their respective preferences. More specifically, given any set of colleges and students, together with their preferences and quotas, can one find a stable matching?

The answer to this question is affirmative.

For the existence proof Gale and Shapley constructed a simple deferred-acceptance algorithm, which, starting from the given data, leads to a stable matching in a finite number of steps. The matching obtained in this way is the unique stable matching (there may be many) which is preferred by all students to any other such matching. For this reason, it is called optimal stable matching for the students.

In the 1962s paper they remark that: "…even though we no longer have the symmetry of the marriage problem, we can still invert our admissions procedure to obtain the unique "ccollege optimal" assignment. The inverted method bears some resemblance to a fraternity "crush week" it starts with each college making bids to those applicants it considers most desirable, up to its quota limit, and then the bid-for students reject all but the most attractive offer, and so on."

In the paper mentioned above the authors express some reservation on the possibilities of application of their algorithm. It turns out that since 1951, a mathematically equivalent algorithm was being used by the National Resident Matching Program (NRMP), located in Evanston, Illinois. The NRMP has the task each year of assigning graduates of all medical schools in the United States to hospitals where they are required to serve a year's residency. The algorithm used by the NRMP was mathematically the equivalent to the one described in Gale and Shapley [30] to produce the optimal stable matching for the colleges. Thus, in this algorithm, each hospital applies to its quota of students. The confirmation of this fact was obtained by David Gale in 1975. In a letter from December 8, 1975, in response to a letter from Gale to the NRMP, Elliott Peranson, consultant to NRMP, responsible for the technical operation of the matching program, says the following:

"However, I might point out that the NRMP algorithm in fact uses the inverse procedure and produces the unique "ccollege optimal" assignment rather than this "student optimal" assignment. This procedure more closely parallels the actual admissions process where a matching algorithm is not used. In this case students apply to all hospitals they would consider (not just their first choice), each hospital then selects the most desirable students, up to its quota limit, from amongst all applicants, then the "bid-for" students reject all but the most desirable offer, and so on".

Hence, the proof that the NRMP was yielding a stable matching, which was the optimal stable matching for the colleges, is that such an outcome is always obtained by the Gale and Shapley's algorithm with the colleges proposing.

The discovery that the two algorithms were mathematically equivalent was first spread orally and later reported in Gale and Sotomayor [32] with an equivalent description of the NRMP algorithm. (Roth [63] also presents the NRMP algorithm). This was the first application of matching theory of which we have knowledge. The algorithms proposed by Gale and Shapley are described below. Their description is quoted from Gale and Sotomayor [32].

Gale–Shapley-Algorithm with the Colleges Proposing to the Applicants

Each hospital H tentatively admits its quota q_H consisting of the top q_H applicants on its list. Applicants who are tentatively admitted to more than one hospital *tentatively accept* the one they prefer. Their names are then removed from the lists of all other hospitals which have tentatively admitted them. This gives the *first tentative matching.* Hospitals which now fall short of their quota again admit tentatively until either their quota is again filled or they have exhausted their list. Admitted applicants again reject all but their favorite hospital, giving the *second tentative matching,* etc. The algorithm terminates when, after some tentative matching, no hospitals can admit any more applicants either because their quota is full or they have exhausted their list. The tentative matching then becomes permanent.

Gale–Shapley-Algorithm with the Applicants Proposing to the Colleges

Each applicant *petitions* for admission to his/her favorite hospital. In general, some hospitals will have more petitioners than allowed by their quota. Such oversubscribed hospitals now reject the lowest petitioners on their preference list so as to come within their quota. This is the *first tentative matching.* Next, rejected applicants petition for admission to their second favorite hospital and again oversubscribed hospitals reject the overflow, etc. The algorithm terminates when every applicant is tentatively admitted or has been rejected by every hospital on his list.

The fact that the matching produced by the NRMP algorithm is stable stands for one of the most important applications of game theory to Economics. During about fifty years the allocation procedures used to assign interns to hospitals in the United States produced unstable matchings. Unsuccessful procedures were often proposed by the Association of American Medical Colleges. This sort of events culminated with a centralized mechanism that employed the NRMP algorithm. Such a centralized mechanism lasted for almost fifty years, suggesting that interns and hospitals had reached an equilibrium. And the paper written by Gale and Shapely corroborated that the game-theoretical predictions were, once more, correct.

Before the publication of Gale and Sotomayor [32], a few, but important, contributions were made to the theory of two-sided matching markets. The famous paper of 1972 by Shapley and Shubik establishes the assignment game via the introduction of money, as a continuous variable, into the marriage model. The book *Marriages Estables* by Knuth was published in 1976. In this volume, the proof, attributed to Conway, that the set of stable matchings for the marriage model is a lattice is presented. The assignment game was generalized by Kelso and Crawford [42] to a model where the utilities satisfy some gross substitute condition. Another generalization, which considers continuous utility functions, non-necessarily linear, was presented in Demange and Gale [18].

Nevertheless, among the contributions of this period one of them caused considerable impact. This was the non-manipulability theorem by Dubins and Freedman [21]. In a stable revelation mechanism, for every profile of preferences that can be selected by the agents, some algorithm that yields a stable matching is used. These authors prove that the revelation mechanism that produces the optimal stable matching for a given side of the marriage market is not collectively manipulable by the agents of that side. Also, this non-manipulability result holds for the college admission market when the mechanism yields the student-optimal stable matching. An analogue to this result was proved in Demange and Gale [18] for the continuous model through a key lemma that became known in the literature as the Blocking Lemma. The main challenge that motivated Gale and Sotomayor [32] was to prove the discrete version of the blocking lemma, which allowed to prove the non-manipulability theorem in just three lines. Two simple and short proofs (one with the use of the algorithm and the other one without the use of the algorithm) of Dubins and Freedman's theorem were presented as an alternative to the original proof by the authors which was about twenty pages long. An example in Dubins and Freedman [21], where some woman can be better off by falsifying her preference list when the man-optimal stable matching is to be employed, motivated Gale and Sotomayor [31]. This paper proves that such a mechanism can almost always be manipulated by the women and then treats the strategic possibilities for these agents in the corresponding strategic game. Another paper of this period was Roth [60], which proves, via an example, that any rule for selecting a stable matching is manipulable (either by some man or some woman).

The existence theorem of manipulability by the women, the Impossibility Theorem and the non-manipulability theorem attracted the authors' interest toward a fruitful line of investigation concerning the incentives facing the agents when an allocation mechanism is employed. Algorithms have been developed for this purpose for several matching models. The games induced by such mechanisms are played non-cooperatively by the agents, and in general, their self-enforcing agreements lead to a stable outcome. In these cases a non-cooperative implementation of the set of stable outcomes is provided. Analyzing the strategic behavior of the agents in such games has been an important subject of research of several authors in an attempt to get precise answers to the strategic questions raised. In this direction we can cite Roth [60,61], Gale and Sotomayor [31], Perez-Castrillo and Sotomayor [56], Sotomayor [87,88,94], Kamecke [38], Kara and Sönmez [40,41], Alcalde [5], Alcalde, Perez–Castrillo and Romero–Medina [6], among others.

Over all these years the popularity of the matching theory has spread among mathematicians and economists, mainly due to the publication in 1990 of the first edition of the book *Two-sided matchings. A study in game theoretical modeling and analysis*, by Roth and Sotomayor, which attempts a comprehensive survey of the main results on the two-sided matching theory until that date. (An extensive bibliography can also be found in http://kuznets.fas.harvard.edu/~aroth/bib.html#matchbib).

The stable matching problem has been generalized to several two-sided matching models, which have been widely modeled and analyzed under both cooperative and non-cooperative game theoretic approaches. Through these models a variety of markets has become better understood, which has considerably contributed to their organization. The deferred acceptance algorithm of Gale and Shapley and adaptations of it have been applied in the reorganization of admission processes of many two-sided matching markets. And the design of these mechanisms has also raised new theoretical questions. (In this connection see e. g. Balinski and Sönmez [8], Ergin and Sönmez [28], and Pathak and Sönmez [55], Abdulkadiroglu and Sönmez [1] and Bardella and Sotomayor [9]).

Introduction

The two-sided matching theory is the study of game theoretical models in which the set of players is partitioned into the disjoint union of two finite sets, and the main activity of the agents from one set is to form partnerships with the agents on the other set. In addition, a structure of preferences is available for the players, as well as an array of quotas, one quota for each participant, representing the maximum number of partnerships that he/she/it is allowed to form.

The rules of the game determine the feasible outcomes. These models are called two-sided matching models. They are suitable for modeling a great variety of labor markets of firms and workers, markets of buyers and sellers, markets of students and schools, etc. They are grouped into three categories: the discrete, the continuous and the hybrid matching models. The distinctions between them are based on the kind of preference structure they are endowed with. Within each category the models vary as to the possibilities of the agents to form multiple partnerships, the rules of the game and some variation of the structure of the preferences.

Given the sets of agents, the structure of preferences and quotas and the rules of the game, the question that emerges is:

A. Which partnerships should be formed?
B. If a partnership is formed, what payoff should be awarded to each agent?

The prediction should be outcomes that cannot be upset by any coalition. The idea is that, in games where players form partnerships, it should be allowed for coalitions to be formed in which its members keep some of their current partners if they wish, replace some others with new partners belonging to the coalition. Furthermore, by doing this they get a preferable outcome. This intuitive idea is captured by a refinement of the core concept introduced in Sotomayor [80], called setwise-stability (stability, for short). For the marriage model and the college admission model with responsive preferences this concept coincides with the one introduced by Gale and Shapley.

This review is an attempt to understand some of the differences and similarities between some matching models. We do that by analyzing both the cooperative and the non-cooperative structure of these models. Some future directions will then be presented.

The rest of the article is organized as follows. Section "Discrete Two-Sided Matching Models" is devoted to the cooperative approach of the discrete matching models with responsive preferences. It introduces the basic cooperative one-to-one, many-to-one and many-to-many matching models and discusses the main properties that characterize the set of stable outcomes for these models. This kind of analysis is also provided in Sect. "Continuous Two-Sided Matching Model with Additively Separable Utility Functions" for the continuous matching models where the utilities are additively separable. Section "Hybrid One-to-One Matching Model" discusses some fundamental similarities and differences between a one-to-one hybrid model and its corresponding non-hybrid matching models. Strategic questions are treated in Sect. "Incentives". Section "Future Directions" presents some future directions and open problems.

Discrete Two-Sided Matching Models

There are two finite and disjoint sets of agents, F and W, which we may think of as being sets of men and women, colleges and students, firms and workers, etc. To fix ideas let us describe the model in terms of firms and workers. Then, F is a set of m firms and W is a set of n workers. Salaries cannot be negotiated and are part of the job description. Each worker w has a quota $s(w)$ representing the maximum number of jobs in different firms she can take. Each firm f has a quota $r(f)$ representing the maximum number of workers it can hire. Set r and s the array of quotas of the firms and the workers, respectively.

Since the agents form partnerships, they always have preferences over potential individual partners; that is, over allowable sets of partners with only one agent belonging to the opposite side of the market. These preferences are assumed to be strict. They are transitive and complete and so, they can be represented by ordered lists of preferences. Thus, the individual preference relation of firm f can be represented by an ordered list of preferences $P(f)$ on the set $W \cup \{f\}$; the individual preference relation of worker w can be represented by an ordered list of preferences $P(w)$, on the set $F \cup \{w\}$. The array of these preferences will be denoted by P. Then, an agent w is acceptable to an agent f if $w \geqslant_f f$. Similarly, an agent f is acceptable to an agent w if $f \geqslant_w w$.

If an agent may form more than one partnership then his/her/its preferences are not restricted to the individual potential partners. That is, agents have preferences over any allowable set of partners. Given two allowable sets of partners, A and B, for agent $y \in F \cup W$, we write $A >_y B$ to mean y prefers A to B, and $A \geqslant_y B$ to mean y likes A at least as well as B. In order to fix ideas we will assume that these preferences are *responsive* to the agents' individual preferences and are not necessarily strict.

The rules of the market are that any firm and worker pair may sign an employment contract with each other if they both agree; any firm may choose to keep one or more of its positions unfilled, and any worker may choose not to fill his or her quota of jobs if he or she wishes to do so.

When the quota of every agent is one, we have the *marriage model*. In this case an allowable set of partners for any agent is a singleton. Therefore, every agent only has preferences over individuals.

If only the agents of one of the sides are allowed to have a quota greater than one, then we have the so called *college admission model* with responsive preferences.

In both models it is a matter of verification that the sets of setwise-stable matchings, pairwise- stable matchings and strong corewise-stable matchings coincide. For the many-to-many case this is not always true. The strong corewise-stability concept is not a natural solution concept for this model as Example 1 shows.

Example 1 (Sotomayor [80]) (The corewise-stability concept is not a natural solution concept for the many-to-many case)

Consider two firms f_1 and f_2 and two workers w_1 and w_2. Each firm may employ and wants to employ both workers; worker w_1 may take, at most, one job and prefers f_1 to f_2; worker w_2 may work and wants to work for both firms. If the agents can communicate with each other, the outcome that we expect to observe in this market is obvious: f_1 hires both workers and f_2 hires only worker w_2. Of course this outcome is in the strong core. Since f_1 has a quota of two and w_1 prefers f_1 to f_2 we cannot expect to observe the strong corewise-stable outcome where f_1 hires only w_2 and f_2 hires both workers. That is, both outcomes are in the strong core, but only the first one is expected to occur. Our explanation for this is that only the first outcome is setwise-stable.

On the other hand, the pairwise-stability concept is not a refinement of the core for the discrete many-to-many matching models. See Example 2.

Example 2 (Sotomayor [80]) (A pairwise-stable matching which is not in the core)

Consider the following labor market with a set of firms $F = \{f_1, f_2, f_3, f_4\}$ and a set of workers $W = \{w_1, w_2, w_3, w_4\}$, where each firm can hire two workers, and each worker can work for two firms. If firm f_i hires worker w_j then f_i gets the profit a_{ij} and w_j gets the salary b_{ij}. The pairs of numbers (a_{ij}, b_{ij}) are given in Table 1.

Consider the matching μ where f_1 and f_2 are matched to $\{w_3, w_4\}$ and f_3 and f_4 are matched to $\{w_1, w_2\}$. (The payoffs of each matched pair are presented in boldface). This matching is pairwise stable. In fact, f_3 and f_4 do not belong to any pair which causes instability, because they are matched to their two best choices: w_1 and w_2; (f_1, w_1) and (f_1, w_2) do not cause instabilities since f_1 is the worst choice for w_1 and w_2 is the worst choice for f_1; (f_2, w_1) and (f_2, w_2) do not cause instabilities since w_1 is the worst choice for f_2 and f_2 is the worst choice for w_2. Nevertheless, f_1 and f_2 prefer $\{w_1, w_2\}$ to $\{w_3, w_4\}$ and w_1 and w_2 prefer $\{f_1, f_2\}$ to $\{f_3, f_4\}$. Hence this matching is not in the core, since it is blocked by the coalition $\{f_1, f_2, w_1, w_2\}$.

Two-Sided Matching Models, Table 1

Payoff Matrix of Example 2. Each row represents a firm and each column a worker. The values in the cell represent the payoffs to the corresponding firm (*first value*) and worker (*second value*)

10,1	1,10	**4,10**	**2,10**
1,10	10,1	**4,4**	**2,4**
10,4	**4,4**	2,2	1,2
10,2	**4,2**	2,1	1,1

Two-Sided Matching Models, Table 2

Payoff Matrix of Examples 3 and 4. Each row represents a firm and each column a worker. The values in the cell represent the payoffs to the corresponding firm (*first value*) and worker (*second value*)

13,1	**14,10**	**4,10**	1,10	0,0	0,0	3,10
1,10	0,0	0,0	10,1	**4,10**	**2,10**	0,0
10,4	0,0	0,0	0,0	0,0	0,0	0,0
10,2	0,0	0,0	0,0	0,0	0,0	0,0
0,0	**9,9**	0,0	**10,4**	0,0	0,0	0,0
0,0	0,0	0,0	**10,2**	0,0	0,0	0,0

Example 3 shows that setwise stability is a strictly stronger requirement than pairwise-stability plus strong corewise-stability. It presents a situation in which the set of setwise-stable matchings is a proper subset of the intersection of the strong core with the set of pairwise-stable matchings.

Example 3 (Sotomayor [80]) (A strong corewise-stable matching which is pairwise-stable and is not setwise-stable)

Consider the following labor market with a set of firms $F = \{f_1, f_2, f_3, f_4, f_5, f_6\}$ and a set of workers $W = \{w_1, w_2, w_3, w_4, w_5, w_6, w_7\}$, where $r_1 = 3$, $r_2 = r_5 = 2$, $r_3 = r_4 = r_6 = 1$, $s_1 = s_2 = s_4 = 2$ and $s_3 = s_5 = s_6 = s_7 = 1$. If firm f_i hires worker w_j then f_i gets profit a_{ij} and the worker gets a salary b_{ij}. The pairs of numbers (a_{ij}, b_{ij}) are given in Table 2.

Let μ be the matching given by

$$\mu(f_1) = \{w_2, w_3, w_7\}\,, \quad \mu(f_2) = \{w_5, w_6\}\,,$$
$$\mu(f_3) = \mu(f_4) = \{w_1\}\,, \quad \mu(f_5) = \{w_2, w_4\}\,, \quad \text{and}$$
$$\mu(f_6) = \{w_4\}\,.$$

The associated payoffs are shown in bold in Table 2. This matching is in the strong core. In fact, if there is a matching μ' which weakly dominates μ via some coalition A, then, under μ', no player in A is worse off and at least one player in A is better off. Furthermore, matching μ' must match all members of A among themselves. By inspection we can see that the only players that can be better off are f_1, f_2, w_1 and w_4, for all remaining players are matched to their best choices. However, if A contains one player of the set $\{f_1, f_2, w_1, w_4\}$, then A must contain all four players. In fact, If $f_1 \in A$, then f_1 must form a new partnership with w_1, so $w_1 \in A$. If $w_1 \in A$, then w_1 must form a new partnership with f_2, so $f_2 \in A$. If $f_2 \in A$, then f_2 must form a new partnership with w_4, so $w_4 \in A$. Finally, if $w_4 \in A$ then w_4 must form a new partnership with f_1, so $f_1 \in A$. Thus, if μ' weakly dominates μ via A, then f_1, f_2, w_1 and w_4 are in A and f_1 and f_2 form new partnerships with w_1 and w_4. Nevertheless, f_1 must keep his partnership with w_2, his best choice. Then w_2 must be in A, so she cannot be worse off and so f_5 must also be in A. But f_5 requires the partnership with w_4, who has quota of 2 and has already filled her quota with f_1 and f_2. Hence f_5 is worse off at μ' than at μ and then μ' cannot weakly dominate μ via A.

The matching μ is clearly pairwise stable. Nevertheless, the coalition $\{f_1, f_2, w_1, w_4\}$ causes an instability in μ. In fact, if f_1 is matched to $\{w_1, w_2, w_4\}$ and f_2 is matched to $\{w_1, w_4\}$ then f_1 gets 28 and the rest of the players in the coalition get 11 instead of 6. Hence, the matching μ is not setwise-stable.

The question is then to know if, given any two sets of agents with their respective preferences and quotas, one can always find a setwise-stable matching. The answer is affirmative for the one-to-one matching model and for the many-to-one matching models with substitutable preferences. The existence of setwise-stable matchings for the marriage model was first proved by Gale and Shapley [30]. Sotomayor [76] also provides a simple proof of the existence of stable matchings for the marriage model that connects stability with a broader notion of stability with respect to unmatched agents. Gale and Shapley construct a deferred acceptance algorithm described below and prove that it yields a stable matching in a finite number of steps.

The deferred acceptance algorithm with the men making the proposals. Each man begins by proposing to his favorite woman (the first woman on his preference list). Each woman rejects the proposal of any man unacceptable to her, and in case she gets several proposals, she keeps only her most preferred one. If a man is not rejected at this step he is kept engaged. At any step, any man who was rejected at the previous step proposes to his next choice (his most preferred woman among those who have not rejected him), as long as there remains an acceptable woman to whom he has not yet proposed (if at one step a man has been rejected by all of his acceptable women, he issues no further proposals). The algorithm terminates after any step in which no man is rejected, because then every man is either engaged to some woman or has been rejected by every woman on his list of acceptable women. Women who did not receive any acceptable proposals, and men rejected by all women acceptable to them remain single.

To see that the matching yielded by this algorithm is stable first observe that no agent has an unacceptable partner. In addition, if there is some man f and woman w not matched to each other and such that f prefers w to his current partner, then woman w is acceptable to man f and so he must have proposed to her at some step of the algorithm. But then he must have been rejected by w in favor of someone she prefers to f. Therefore, w is matched to a man whom she prefers to f, by the transitiveness of the preferences and so f and w do not destabilize the matching.

For the college admission model with responsive preferences Gale and Shapley defined a related marriage model in which each college is replicated a number of times equal to its quota, so that in the related model, every agent has a quota of one. If $f_1, \ldots, f_{r(f)}$ are the $r(f)$ copies of college f then each of these f_i's has preferences over individuals that are identical with those of f. Each student's preference list is changed by replacing f, wherever it appears on his/her list, by the string $f_1, \ldots, f_{r(f)}$ in that order of preference. Therefore, the stable matchings of the related marriage market are in natural one-to-one correspondence with the stable matchings of the college admission market. By using the existence theorem for the marriage model we obtain the corresponding result for the college admission model.

The existence proof for the many-to-one case with strict and substitutable preferences was first given by Kelso and Crawford [42] through a variant of the deferred-acceptance algorithm.

If the market does not have two-sides or the many-to-one matching model does not have substitutable preferences then the set of setwise-stable matchings may be empty. Gale and Shapley [30] present an example of a one-sided matching model that does not have any stable matchings. This model was called by these authors the "roommate problem". In this example there are four agents $\{a, b, c, d\}$, such that a's first choice is b, b's first choice is c, and c's first choice is a, and d is the last choice of all the other agents. Of course if some agent is un-

matched then there will be two unmatched agents and they will destabilize the matching. If every one is matched, the agent who is matched to d will form a blocking pair with the agent who lists him at the head of his list. Therefore, there is no stable matching in this example.

A small amount of literature has grown around the issues of finding conditions in which the set of stable matchings is non-empty for the roommate problem, and the performance of algorithms that can produce them when they exist. (See Abeledo and Isaak [3], Irving [37], Tan [95], Chung [13] and Sotomayor [89].)

If preferences are not substitutable Example 2.7 of Roth and Sotomayor [69] shows that setwise-stable matchings may not exist in the many-to-one case.

Even in the simplest case of preferences representable by additively separable utility functions, setwise-stable matchings may not exist for the many-to-many case. See the example below.

Example 4 (Sotomayor [80]) (Nonexistence of stable matchings) Consider again the matching model of Example 2. We are going to show that the set of setwise-stable matchings is empty. First, observe that f_3 prefers $\{w_1, w_2\}$ to any other set of players and f_3 is the second choice for w_1 and w_2; w_3 prefers $\{f_1, f_2\}$ to any other set of players and w_3 is the second choice for f_1 and f_2. Then, in any stable matching μ, f_3 must be matched to w_1 and w_2, while w_3 must be matched to f_1 and f_2. Separate the cases by considering the possibilities for the second partner of w_1, under a supposed stable matching μ:

Case 1: (w_1 is matched to $\{f_2, f_3\}$). Then f_2 is not matched to w_4 and we have that $\{f_2, w_4\}$ causes an instability in the matching, since f_2 prefers w_4 to w_1 and f_2 is the second choice for q_4.

Case 2: (w_1 is matched to $\{f_3, f_4\}$). Then the following possibilities occur:

(i) (w_2 is matched to $\{f_3, f_4\}$.) Then $\{f_1, f_2, w_1, w_2\}$ causes an instability in the matching. This matching is pairwise-stable, but it is not in the core.

(ii) (w_2 is matched to $\{f_3, f_1\}$.) Then $\{f_1, w_4\}$ causes an instability in the matching, since f_1 is the first choice for w_4 and f_1 prefers w_4 to w_2.

(iii) (w_2 is matched to $\{f_3, f_2\}$ or $\{f_3\}$.) Then $\{f_4, w_2\}$ causes an instability in both cases, since w_2 is the second choice for f_4 and w_2 prefers f_4 to f_2 and prefers f_4 to have an unfilled position.

Case 3: (w_1 is matched to $\{f_1, f_3\}$ or $\{f_3\}$.)} Then $\{f_4, w_1\}$ causes an instability in both cases, since w_1 is the first choice for f_4 and w_1 prefers f_4 to f_1 and prefers f_4 to have an unfilled position.

Hence, there are no stable matchings in this example.

Pairwise-stable matchings always exist when the preferences are substitutable. When preferences are strict, Roth [62] presents an algorithm that finds a pairwise-stable matching for a many-to-many matching model with substitutable preferences. Sotomayor [80] provides a simple and non-constructive proof of the existence of pairwise-stable matchings for the general discrete many-to-many matching model with substitutable and not-necessarily strict preferences. Martínez et al. [51] construct an algorithm, which allows finding the whole set of pairwise-stable matchings, when they exist, for the many-to-one matching model.

Authors have looked for sufficient conditions on the preferences of the agents for the existence of setwise-stable matchings in the many-to-many cases. Sotomayor [88] proves that if the preferences of the firms satisfy the *maximin* property, then the set of pairwise-stable matchings coincides with the set of setwise-stable matchings. An example in that paper shows that the set of setwise stable matchings may be empty if this condition is not satisfied. It is assumed there that the preferences are responsive and it is conjectured that the result above extends to the case of substitutable preferences.

Echenique and Oviedo [24] also address this problem with a different condition. They show that if agents on one side of the market have strongly substitutable preferences, while the other side has substitutable preferences, then the set of setwise stable matchings coincides with the set of pairwise-stable matchings.

Konishi and Ünver [46] give conditions on the preferences of the agents in a many-to-many matching market under which a pairwise-stable matching cannot be quasi-dominated by a pairwise-unstable matching via a collation.

Eeckhout [26], under the assumption of strict preferences and that every man (woman) is acceptable to every woman (man), presents a sufficient condition for uniqueness of the stable matchings in the marriage market. The condition on preferences is simple: for every $f_i \in F = \{f_1, f_2, \ldots, f_m\}$, $w_i >_{f_i} w_k$ for all $k > i$, and for every $w_j \in W = \{w_1, w_2, \ldots, w_n\}$, $f_j >_{w_j} f_k$ for all $k > i$.

One line of investigation that has been developed in the theory of two-sided matchings concerns the mathematical structure of the set of stable matchings, because it captures fundamental differences and similarities between the several kinds of models. For the marriage model and the college admission model with responsive preferences,

assuming that the preferences over individuals are strict, the set of setwise-stable (stable for short) matchings have the following characteristic properties:

A1. *Let μ and μ' be stable matchings. Then $\mu \geqslant_F \mu'$ if and only if $\mu' \geqslant_W \mu$.}*
That is, there exists an opposition of interests between the two sides of the market along the whole set of stable matchings.

A2. *Every agent is matched to the same number of mates under every stable matching.*
Consequently, *if an agent is unmatched under some stable matching then he/she/it is unmatched under any other stable matching.*
When preferences are strict, there are two natural partial orders on the set of all stable matchings. The partial order $\geqslant_F$ is defined as follows: $\mu \geqslant_F \mu'$ if $\mu(f) \geqslant_f \mu'(f)$ for all $f \in F$. The partial order $\geqslant_W$ is analogously defined. The fact that these partial orders are well defined follows from A1. Then,

A3. *The set of stable matchings has the algebraic structure of a complete lattice under the partial orders $\geqslant_F$ and $\geqslant_W$.*

The lattice property means the following: if μ and μ' are two stable matchings, then some workers (respectively firms) will get a preferable set of mates under μ than under μ' and others will be better off under μ' than under μ. The lattice property implies that there is then a stable matching which gives each agent the most preferable of the two sets of partners and also one which gives each of them the least preferred set of partners. That is, if μ and μ' are stable matchings the lattice property implies that the functions λ, υ, η and τ are stable matchings, where $\lambda = \mu \vee_F \mu'$ is defined by $\lambda(f) = \max\{\mu(f), \mu'(f)\}$ and $\lambda(w) = \min\{\mu(w), \mu'(w)\}$; the function $\eta = \mu \vee_W \mu'$ is analogously defined; the function $\upsilon = \mu \wedge_F \mu'$ is defined by $\upsilon(f) = \min\{\mu(f), \mu'(f)\}$ and $\upsilon(w) = \max\{\mu(w), \mu'(w)\}$. Analogously we define $\tau = \mu \wedge_W \mu'$ (notice that $\mu \vee_F \mu'$ is the same as $\mu \wedge_W \mu'$ and $\mu \vee_W \mu'$ is the same as $\mu \wedge_F \mu'$).

The fact that the lattice is complete implies the existence and uniqueness of a maximal element and a minimal element in the set of stable payoffs, with respect to the partial order that is being considered. Thus, there exists one and only one stable matching μ_F and one and only one stable matching μ_W such that $\mu_F \geqslant_F \mu$ and $\mu_W \geqslant_W \mu$ for all stable matchings μ. Property A1 then implies that $\mu \geqslant_W \mu_F$ and $\mu \geqslant_F \mu_W$. That is,

A4. *There is an F-optimal stable matching μ_F with the property that for any stable matching μ, $\mu_F \geqslant_F \mu$ and $\mu \geqslant_W \mu_F$; there is a W-optimal stable matching μ_W with symmetrical properties.*

Property A1 was first proved by Knuth [45] for the marriage model. The result for the college admission model with responsive preferences was proved in Roth and Sotomayor [69] by making use of the following proposition of Roth and Sotomayor [68]: *Suppose colleges and students have strict individual preferences, and let μ_1 and μ_2 be stable matchings for the college admission model such that $\mu_1(f) \neq \mu_2(f)$. Let μ_1* and μ_2* be the stable matchings corresponding to μ_1 and μ_2 in the related marriage model. If $\mu_1 * (f_i) >_f \mu_2 * (f_i)$ for some position f_i of f then $\mu_1 * (f_j) \geqslant_f \mu_2 * (f_j)$ for all positions f_j of f.*

Property A2 was proved by Gale and Sotomayor [32] for both models. For the college admission model with responsive preferences Roth [66] added that *if a college does not fill its quota at some stable matching then it has the same set of mates at every stable matching.* The restriction of property A2 to the marriage model where every pair of partners is mutually acceptable was proved by McVitie and Wilson [52].

For the many-to-one case with substitutable preferences, Martínez et al. [50] presents an example in which there are agents who are unmatched under some stable matching and are matched under another one. By introducing quotas in the model with substitutable preferences of Roth and Sotomayor [69], these authors prove that if the preferences of the colleges are strict, substitutable and $r(f)$-separable for every college f, then property A2 holds. Furthermore Roth's result mentioned above also applies.

The lattice property of the set of stable matchings for the marriage model is attributed by Knuth [45] to Conway. The existence of the optimal stable matchings for each side of the marriage market and the college admission market with responsive preferences was first proved in Gale and Shapley [30] by using the deferred acceptance procedure. The idea of their elegant proof is to show that a proposer is never rejected by an achievable mate, so he/she/it ends up with his/hers/its best achievable mate.

The lattice property for the college admission model with responsive preferences was obtained in Roth and Sotomayor [69]. To show that the functions λ, υ, η and τ above are well defined these authors used the following theorem from Roth and Sotomayor [68]: *If colleges and students have strict preferences over individuals, then colleges have strict preferences over those groups of students that they may be assigned at stable matchings. That is, if μ_1 and μ_2 are stable matchings, then a college f is indifferent between $\mu_1(f)$ and $\mu_2(f)$ only if $\mu_1(f) = \mu_2(f)$.*

This result is an immediate consequence of the proposition mentioned above, due to the responsiveness of the preferences. Therefore, if μ_1 and μ_2 are two stable matchings then f prefers $\mu_1(f)$ to $\mu_2(f)$ if and only if the $r(f)$ most preferred students by f in the set formed by the union of $\mu_1(f)$ and $\mu_2(f)$ are those ones in $\mu_1(f)$.

For the many-to-many matching market with strict and substitutable preferences, Blair [12] proved that the set of pairwise-stable matchings (not necessarily setwise-stable) has the lattice structure under some partial order relation that is not defined by the preferences of the agents. The definition of the partial order $\geqslant_F$ uses that *if μ_1 and μ_2 are pairwise-stable matchings then $\mu_1 \geqslant_f \mu_2$ if and only if $Ch_f(\mu_1(f) \cup \mu_2(f)) = \mu_1(f)$, for every agent $f \in F$*. Similarly the partial order $\geqslant_W$ is defined. As remarked above the partial order defined by Blair coincides with the one defined by the preferences of the players in the college admission model with responsive preferences.

Adachi [4] introduces a map, which is called a T-map, defined over the set of pre-matchings, in order to show that *the set of stable matchings is a non-empty lattice in the marriage market under strict preferences.* Adachi defines the T-map as follows: given a pre-matching μ, $T(\mu(f))$ is f's most preferred worker in $\{w \in W | f \geqslant_w \mu(w)\} \cup \{f\}$ for all $f \in F$, and similarly, $T(\mu(w))$ is w's most preferred firm in $\{f \in F | w \geqslant_f \mu(f)\} \cup \{w\}$ for all $w \in W$. Clearly, any fixed point of the T-map is a matching, and Adachi showed it has to be stable. Using the partial order $\geqslant_F$ defined by the agents' preferences, he showed that the set of pre-matchings endowed with this partial order is a complete lattice and the T-map is an isotone function (order preserving). Thus, Tarski's fixed-point theorem implies that the set of fixed points of the T-map, which is the set of stable matchings, is a non-empty complete lattice.

Theorem 1 (Tarski's Theorem [96]) *Let E be a complete lattice with respect to some partial order $\geqslant$, and let f be an isotone function from E to E. Then the set of fixed points of f is non-empty and is itself a complete lattice with respect to the partial order $\geqslant$.*

In this same vein, Echenique and Oviedo [23,24] extend Adachi's [4] approach and the T-map in order to analyze the many-to-one and many-to-many models, respectively. Again, any fixed point of the T-map is a matching. Echenique and Oviedo [23] show that for the many-to-one model the set of fixed points of the T-map is equal to the set of stable matchings. By making successive iterations of the T-map, starting from some specific pre-matching, until a fixed point is reached, this map can be used to find all the stable matchings, as long as they exist. This procedure is called the T algorithm. These authors show that *as long as the strong core is non-empty, the T algorithm always converges, and if the strong core is empty, it cycles.* They present an example of a situation in which the preferences are not substitutable and the T algorithm finds strong core allocations, but the algorithm with firms proposing according to their preference lists over allowable sets of workers does not do so. Finally, they give a bound on the computational complexity of the T algorithm and show how it can be used to calculate both the supremum and infimum under Blair's partial order, which under non-substitutable preferences might not be easily computed. Furthermore, under substitutability, the set of pre-matchings endowed with the partial order defined by Blair is again a complete lattice and the T-map is isotone, so Tarski's theorem implies that *the strong core is a non-empty lattice under the partial order introduced in Blair* [12].

For the many-to-many model, the set of fixed-points of the T-map studied by Echenique and Oviedo [24] is shown to be, under substitutability, equal to the set of pairwise-stable matchings, and a superset of the set of setwise-stable matchings. Furthermore, the set of pre-matchings endowed with Blair's partial order is again a complete lattice. Then Tarski's theorem applies and the set of pairwise-stable matchings is a non-empty complete lattice. *If both sides of the market satisfy the strong substitutability property then the set of setwise stable matchings is a complete lattice, both for Blair's partial order and for the partial order defined by the agents' preferences.*

Martínez et al. [51] propose an algorithm which allows them to calculate the whole set of pairwise-stable matchings under substitutability.

Echenique and Yenmez [25] study the college admission problem when students have preferences over colleagues. Using the T-map they construct an algorithm, which finds all the core allocations, as long as the core is non-empty. In a similar setup, Pycia [57] finds necessary and sufficient conditions for the existence of stable matchings. Dutta and Massó [22] studied a many-to-one version of the model of Echenique and Yenmez [25]. They showed that under certain conditions on preferences the core is non-empty.

Hatfield and Milgrom [36] present a general many-to-one model in which, the central concept is that of a contract, which allows a different formalization of a matching. In their model there are a finite set of firms, a finite set of workers and a finite set of wages offered by firms. Each contract c specifies the pair (f, w) involved and the wage the worker w gets from firm f, so that the set of contracts is $C = F \times W \times WP$ (where WP is the set of wages offered by the firms). Clearly, if each firm offers a unique wage level to all workers, their model is a college admis-

sion model. Agents have preferences over the contracts in which they could be involved. In this model, a feasible allocation is a set of contracts $C' \subseteq C$ in which each worker w appears at most in one contract $c \in C'$ and each firm f appears in at most $r(f)$ contracts $c_1, \ldots, c_{r(f)} \in C'$ and such that for each agent $a \in F \cup W$ we have that $Ch_a(C_a) = C_a$, where C_a is the set of contracts in C' in which agent a appears. According to Hatfield and Milgrom a feasible allocation C' is stable if there does not exist an alternative feasible allocation which is strictly preferred by some firm f and weakly preferred by all of the workers it hires. Making use of Tarski's fixed point theorem, they prove that if preferences are strict and satisfy substitutability over the set of contracts, then the set of stable allocations is a non-empty lattice. They introduce the condition of the law of aggregate demand on preferences, which requires that for all allowable sets X, Y if $X \subseteq Y$ then $|Ch_f(Y)| \geqslant |Ch_f(X)|$. By assuming that firms' preferences satisfy substitutability and the law of aggregate demand, they prove some characteristic results on the structure of the set of stable allocations and analyze the incentives facing the agents when a mechanism which produces the W-optimal stable allocation is adopted. They show that under this mechanism, it is a dominant strategy for the workers to state their true preferences.

Ostrovsky [54] generalizes the model presented in Hatfield et al. [36] to a K-sided discrete many-to-many matching model. A set of contracts, which allows the production and consumption of some goods, is called a network. This author considers supply networks where goods are sold and transformed through many stages, starting from the suppliers of initial inputs, going through different intermediaries until they reach the final consumer. He generalizes the concept of pairwise stability to this setting and calls it *chain stability*, which requires that there does not exist a chain of contracts such that all members of this chain are better off. A chain of contracts specifies a sequence of agents, each of whom is the seller in one contract and the buyer in the next one. Under certain conditions on the preferences of the agents he proves the existence of chain stable networks and, by using fixed point methods, he shows that the set of chain stable networks is a non-empty lattice. Furthermore, he proves that there exists a consumer optimal network and an initial supplier optimal network, similar to the F-optimal and W-optimal matchings in other models. Finally, for the case in which each agent can be the seller (buyer) in at most one contract he shows that the set of chain stable networks is equal to the core.

Crawford [15] proposes to allow offers in the NRMP mechanism to include salaries and demonstrates how the resulting market can generate stable outcomes, which might Pareto dominate the ones in the current form of the NRMP. This model can be seen as an application of the Hatfield and Milgrom [36] paper.

Another line of investigation that has grown in the last decade concerns the special case of the college admission model in which colleges have fixed preferences, known nowadays as the school choice model. The seminal paper is Sotomayor [77] which was motivated by the admission market of economists to graduate centers of economics in Brazil. The students take some tests and each institution places weights on each of these tests in order to rank the students according to the weighted average of the tests. See Ergin and Sönmez [28], Pathak and Sönmez [55] and Balinski and Sönmez [8].

Continuous Two-Sided Matching Model with Additively Separable Utility Functions

The two-sided matching model with *additively separable utility functions* involves two finite and disjoint sets of players which will be denoted by B, with m elements, and Q, with n elements. Each $b \in B$ has a quota $r(b)$ and each $q \in Q$ has a quota $s(q)$ representing the maximum number of partnerships they can form. The main characteristic of this model is that agents are able to negotiate their individual payoffs: If a partnership (b, q) is formed, the partners undertake an activity together that produces a payoff v_{bq} which is divided between them into the payoffs u_{bq} for b and w_{bq} for q respectively, as a result of a negotiation process. Therefore, an outcome for this game is a matching, along with individual payoffs u_{bq}'s and w_{bq}'s. Dummy players, denoted by 0, are included for technical convenience in both sides of the market. We have that $v_{b0} = v_{0q} = 0$ for all $b \in B$ and $q \in Q$. As for the quotas, a dummy player may form as many partnerships as needed to fill the quotas of the non-dummy players. Then, an allowable set of partners for agent b, with only $r(b) - k$ elements of Q, has k copies of the dummy Q-agent introduced in the model.

A matching μ is feasible if each player is matched to an allowable set of partners. A feasible outcome, denoted by $(u, w; \mu)$, is a feasible matching μ and a pair of payoffs (u, w), where the individual payoffs of each $b \in B$ and $q \in Q$ are given by the arrays of numbers u_{bq}, with $q \in \mu(b)$, and w_{bq}, with $b \in \mu(q)$, respectively, such that $u_{bq} + w_{bq} = v_{bq}, u_{bq} \geqslant 0$ and $w_{bq} \geqslant 0$. Consequently, $u_{b0} = u_{0q} = w_{b0} = w_{0q} = 0$ in case these payoffs are defined. We say that the matching μ is compatible with the payoff (u, w).

The value of μ is $\sum_{q \in Q, b \in \mu(q)} v_{bq}$. The matching μ is *optimal* if it attains the maximum value among all feasible matchings.

This model generates the following game in coalitional function form with side payments. The set of players is $N = B \cup Q$, and the characteristic function v satisfies the following:

(a) $v(\phi) = 0$,
(b) $v(S) = 0$ if $S \subseteq B$ or $S \subseteq Q$,
(c) $v(S) \leqslant v(T)$ if $S \subseteq T$,
(d) $v(b, q) = v_{bq}$ for all $(b, q) \in B \times Q$.

For every $b \in B$ and for all sets $S \subseteq Q$ with $|S| \geqslant r(b)$,

(e) $v(b \cup S) = \max\{v(b \cup S'); S' \subseteq S \text{ and } |S'| = r(b)\}$.
(c) and (d) imply that for all sets $S \subseteq W$ with $|S| \geqslant r(b)$, $v(b \cup S) = v(b \cup S')$, for some $S' \subseteq S$ with $|S'| = r(b)$

Analogously we define $v(q \cup S)$ for every $q \in Q$ and all sets $S \subseteq B$ with $|S| \geqslant s(q)$.

The condition that the game has *additively separable utilities* means that for every coalition $S = R \cup T, R \subseteq B$ and $T \subseteq Q$,

(f) $v(R \cup T) = \max\{\sum_{(b,q) \in RxT} x_{bq} v_{bq}$, for every feasible assignment $x\}$. Consequently, for all $T \subseteq Q$ with $|T| \leqslant r(b)$ and for all $R \subseteq B$ with $|R| \leqslant s(q)$, $v(b \cup T) = \sum_{q \in T} v_{bq}$ and $v(q \cup R) = \sum_{b \in R} v_{bq}$.

When the quota of any agent is one, the model is the well known assignment game introduced in Shapley and Shubik [72]. In this case the set of individual payoffs of an agent is a singleton, so these payoffs need not be indexed according to the partnerships formed under the matching. Furthermore, the concepts of setwise-stability, pairwise-stability and corewise-stability are equivalent. *The outcome $(u, w; \mu)$ is stable if it is feasible and $u_b + w_q \geqslant v_{bq}$ for all pairs (b, q).*

The existence of stable payoffs for the assignment game was proved in Shapley and Shubik [72] with the use of linear programming. A different, but also simple proof, is obtained in Sotomayor [81] using only combinatorial arguments. Crawford and Knoer [16] consider a discrete version (as well as the continuous version) of the assignment game and develop a version of the deferred acceptance algorithm of Gale and Shapley to prove the non-emptiness of the set of pairwise-stable payoffs.

The main properties that characterize the set of stable outcomes of the assignment game of Shapley and Shubik are the following:

B1. *Let (u, w) be some stable payoff. Then μ is an optimal matching if and only if it is compatible with (u, w).*

This result means that the set of stable payoffs is the same under every optimal matching. Then we can concentrate on the payoffs of the agents rather than on the underlying matching.

B2. *Let (u, w) and (u', w') be stable payoffs. Then $u \geqslant u'$ if and only if $w' \geqslant w$.*

That is, there exists an opposition of interests between the two sides of the market along the whole set of stable payoffs.

B3. *If an agent is unmatched under some stable outcome then he/she gets a zero payoff under any other stable outcome.*

This means, for example, that if a worker is unemployed under some stable outcome then this worker will get a zero salary under any other stable outcome.

B4. *The set of stable payoffs forms a convex and compact lattice under the partial orders $\geqslant_B$ and $\geqslant_Q$.*

The partial order $\geqslant_B$ on the set of stable payoffs is defined as follows: $(u, w) \geqslant_B (u', w')$ if $u_b \geqslant u'_b$ for all $b \in B$. Property B2 implies that $w_q \leqslant w'_q$ for all $q \in Q$, so this partial order is well defined. The partial order $\geqslant_Q$ is symmetrically defined. Then, $(u, w) \vee_B (u', w') = (\max\{u, u'\}, \min\{w, w'\})$ and $(u, w) \wedge_B (u', w') = (\min\{u, u'\}, \max\{w, w'\})$.

The lattice property implies that there exist a maximal element and a minimal element in the set of stable payoffs. The fact that the lattice is complete implies the uniqueness of these extreme points. Thus, there exists one and only one stable payoff $(u*, w_*)$ and one and only one stable payoff $(u_*, w*)$ such that $(u*, w_*) \geqslant_B (u, w)$ and $(u_*, w*) \geqslant_Q (u, w)$ for all stable payoffs (u, w). That is,

B5. *There is a B-optimal stable payoff $(u*, w_*)$ with the property that for any stable payoff (u, w), $u* \geqslant u$ and $w_* \leqslant w$; there is a Q-optimal stable payoff $(u_*, w*)$ with symmetrical properties.*

B6. *The set of stable payoffs equals the core and the set of competitive equilibrium payoffs.*

Excluding property B3 which follows from property 1 of Demange and Gale [18], all the other properties were first proved in Shapley and Shubik [72].

The general quota case is a version of the model studied in Crawford and Knoer [16] and was first presented in Sotomayor [75] in the context of a labor market of firms and workers. Under this approach, the number $r(b)$ is the maximum number of workers firm b can hire; the number $s(q)$ is the maximum number of jobs worker q can

take and the number v_{bq} is the productivity of worker q in firm b. The natural cooperative solution concept is that of setwise-stability which is shown to be equivalent to the concept of pairwise-stability. Then, *the feasible outcome* $(u, w; \mu)$ *is setwise-stable if* $u_b(\min) + w_q(\min) \geqslant v_{bq}$ *for all pairs* (b, q) *with* $q \notin \mu(b)$, *where* $u_b(\min)$ *is the smallest individual payoff of firm b;* $w_q(\min)$ *is the smallest individual payoff of worker q.*

The existence of setwise-stable payoffs for this model was proved in Sotomayor [75,79] through the use of linear programming.

Another interpretation of this model considers a buyer-seller market: B is a set of buyers and Q is a set of sellers. Buyers are interested in sets of objects owned by different sellers and each seller owns a set of identical objects. The number $r(b)$ is the number of objects buyer b is allowed to acquire; the number $s(q)$ is the number of identical objects seller q owns and the number v_{bq} is the amount of money buyer b considers to pay for an object of seller q. We say that v_{bq} is the value of object q (object owned by seller q) to buyer b. An artificial *null-object*, 0, owned by the dummy seller, whose value is zero to all buyers and whose price is always zero is introduced for technical convenience. Under this approach a buyer will be assigned to an allowable set of objects at a feasible allocation, meaning that she is matched to the set of sellers who own the objects in the given set.

Given a price vector $p \in R^{\sigma}_{+}$, with $\sigma \equiv \sum_{q \in Q} s(q)$, the preferences of buyers over objects are completely described by the numbers v_{bq}'s: For any two allowable sets of objects S and S', buyer b prefers S to S' at prices p if her total payoff when she buys S is greater than her total payoff when she buys S'. She is indifferent between these two sets if she gets the same total payoff with both sets. Usually, given the prices of the objects, buyers demand their favorite allowable sets of objects at those prices. The set of such allowable sets is called the demand set of buyer b at prices p. *An equilibrium is reached if every buyer is assigned to an allowable set of objects of her demand set, every seller with a positive price sells all of his items and the number of objects in the market is enough to meet the demand of all buyers.* The solution concept that captures this intuitive idea of equilibrium is that of *competitive equilibrium payoff* defined in Sotomayor [91] as an extension of the concept of competitive equilibrium price for the Assignment game given in Demange, Gale and Sotomayor [19]. *Formally,* $(u, p; \mu*)$ *is a* ***competitive equilibrium outcome*** *if (i) it is feasible, (ii)* $\mu*$ *is a feasible allocation such that, if* $\mu * (b) = S$ *then S is in the demand set of b at prices p for all* $b \in B$ *and (iii)* $p_q = 0$ *if object q is left unsold.*

If $(u, p; \mu*)$ is a competitive equilibrium outcome we say that (u, p) is a competitive equilibrium payoff, $(p, \mu*)$ is a competitive equilibrium and p is a *competitive equilibrium price* or an *equilibrium price* for short.

One characteristic of the additively separable utility function is that if a buyer demands a set A of objects at prices p and some of these objects have their prices raised, then the buyer will continue to want to buy the objects in A whose prices were not changed. That is, the function $v(\{b\} \cup A)$ over all allowable sets A of partners for b satisfies the *gross substitute condition.* Kelso and Crawford [42] formulated a discrete and a continuous many-to-one matching model where the functions $v(\{b\} \cup A)$ satisfy the gross substitute condition and are not necessarily additively separable. In this model, the core, the set of setwise-stable payoffs and the set of competitive equilibrium payoffs coincide and are non-empty. These authors prove, through an example, that without this condition the core may be empty.

A consequence of the competitive equilibrium concept for the many-to-many case with additively separable utility functions is that sellers do not discriminate buyers under a competitive equilibrium payoff, as they might do under a stable outcome. The competitive equilibrium payoffs for this model are characterized *as the setwise-stable payoffs where every seller has identical individual payoffs.* It is interesting to point out that if the identical objects are owned by different sellers, they need not be sold at the same price unless the two sellers have the same number of objects and the selling price is the minimum competitive equilibrium price. (Sotomayor [91]).

A stable (respectively, competitive equilibrium) payoff is called a B-optimal stable (respectively, competitive equilibrium) payoff if every agent in B weakly prefers it to any other stable (respectively, competitive equilibrium) payoff. That is, the B-optimal stable (respectively, competitive equilibrium) payoff gives to each agent in B the maximum total payoff among all stable (respectively, competitive equilibrium) payoffs. Similarly we define a Q-optimal stable (respectively, competitive equilibrium) payoff.

The existence and uniqueness of the B-optimal and of the Q-optimal stable payoffs are proved in Sotomayor [79] by showing that the set of stable payoffs is a lattice under two conveniently defined partial orders. This result runs into the difficulties of defining a partial order relation in the set of stable payoffs, due to the fact that, on the one hand the arrays of individual payoffs are unordered sets of numbers indexed according to the current matching and on the other hand, the agents' preferences do not define a partial order relation, since they violate the anti-symmetric property.

To solve this problem Sotomayor [79] defines a partnership (b, q) to be nonessential if it occurs in some but not all optimal matchings and essential if it occurs in all optimal matchings. Then, two matchings differ only by their nonessential partnerships. According to Theorem 1 of that paper, *(i) in every stable outcome a player gets the same payoff in any nonessential partnership; furthermore this payoff is less than or equal to any other payoff the player gets under the same outcome; (ii) given a stable outcome* $(u, w; \mu)$ *and a different optimal matching* μ'*, we can reindex the* u_{bq}*'s and* w_{bq}*'s according to* μ' *and still get a stable outcome.*

Therefore, the array of individual payoffs of a player can be represented by a vector in a Euclidean space whose dimension is the quota of the given player. The first coordinates are the payoffs that the player gets from his essential partners (if any), following some given ordering. The remaining coordinates (if any) are equal to a number which represents the payoff the player gets from all his nonessential partners. This representation is clearly independent of the matching, so any optimal matching is compatible with a stable payoff. Hence, by ordering the players in B (respectively, Q), we can immerse the stable payoffs of these players in a Euclidean space, whose dimension is the sum of the quotas of all players in B (respectively, Q). Then, the natural partial order relation of this Euclidean space induces the partial order relation $\geqslant_B$ (respectively $\geqslant_Q$) in the set of stable payoffs. We say that $(u, w) \geqslant_B (u', w')$ if the vector of individual payoffs of any buyer, under (u, w), is greater than or equal to her vector of individual payoffs under (u', w'). Similarly we define $(u, w) \geqslant_Q (u', w')$.

The main results of Sotomayor [79] are that, under the vectorial representation of the stable payoffs, properties B1, B2, B3, B4 and B5 hold for the general many-to-many case.

An implication of property B2 is the conflict of interests that exists between the two sides of the market with respect to two comparable stable payoffs. That is, if payoffs (u, w) and (u', w') are stable and comparable, then for all $(b, q) \in B \times Q$ we have that b's total payoff under the first outcome is greater than b's total payoff under the second outcome if and only if q's total payoff under the second outcome is greater than q's total payoff under the first outcome. From property B3, if a seller has some unsold object under a stable outcome, then one of his individual payoffs will be zero under any other such outcome.

Even though the preferences of the players do not define the partial orders $\geqslant_B$ and $\geqslant_Q$, the property stated in B4 is of interest because the two extreme points of the lattice have an important meaning for the model. The extreme points of the lattice are precisely the B-optimal and the Q-optimal stable payoffs. Also every buyer weakly prefers any stable payoff to the Q-optimal stable payoff and any seller weakly prefers any stable payoff to the B-optimal stable payoff.

Indeed, the set of competitive equilibrium payoffs is a sub lattice of the set of stable payoffs. This connection is given by the following theorem of Sotomayor [91] that states that *the set of competitive equilibrium payoffs is contained in the set of stable payoffs and is a non-empty and complete lattice under the partial order* $\geqslant_B$ *(respectively* $\geqslant_Q$*) whose supremum (respectively infimum) is B-optimal and whose infimum (respectively supremum) is Q-optimal.*

The idea of the proof is that the set of competitive equilibrium payoffs can be obtained by "shrinking" the set of stable payoffs through the application of a convenient isotone (order preserving) map f whose fixed points are exactly the competitive equilibrium payoffs. The desired result is concluded via the *algebraic fixed point theorem* due to Alfred Tarski [96].

It is also proved in Sotomayor [91] that *the B-optimal stable payoff is a fixed point of f, so the B-optimal stable payoff is the B-optimal competitive equilibrium payoff.*

As for property B6, Sotomayor [84] shows that the core coincides with the set of pairwise-stable payoffs in the many-to-one case where sellers have a quota of one. Since a seller only owns one object then he cannot discriminate the buyers, so the core coincides with the set of competitive equilibrium payoffs in this model. Thus, the set of competitive equilibrium payoffs is a lattice in this model. The same result is reached in Gül and Stacchetti [33] for the many-to-one case in which the utilities satisfy the gross substitute condition.

However, in the general quota case under additively separable utilities, the core may be bigger than the set of stable payoffs, which in its turn contains and may contain properly the set of competitive equilibrium payoffs, as it is illustrated in the example below from Sotomayor [91]. This example also shows that the core may not be a lattice and, the polarization of interests, observed in the sets of stable payoffs and of competitive equilibrium payoffs, does not always carry over to the core payoffs: The best core payoff for the buyers is not necessarily the worst core payoff for the sellers.

Example 5 (Sotomayor, [91]*)* Consider the following situation. The B-players will be called firms and the Q-players will be called workers. There are two firms, b and b', and two workers q and q'. Each firm may employ and wants to employ both workers; worker q may take, at most, one job and worker q' may work and wants to work for both

firms. The first row of matrix v is $(3, 2)$ and the second one is $(3, 3)$.

There are two optimal matchings: μ and μ', where $\mu(b) = \{q, q'\}, \mu(b') = \{q', 0\}$ and $\mu'(b') = \{q, q'\}$, $\mu'(b) = \{q', 0\}$. The core is described by the set of individual payoffs (u, w) whose total payoffs (U, W), satisfy the following system of inequalities: $0 \leqslant U_b \leqslant 2$, $0 \leqslant U_{b'} \leqslant 3$; $W_q + W_{q'} \geqslant 3$, $W_{q'} - W_q \leqslant 2$, $1 \leqslant W_q \leqslant 3$. It is not hard to see that the outcome (u, w) is stable if and only if seller q always gets payoff $w_q = 3$, seller q' gets individual payoffs $w_{bq'} \in [0, 2]$ and $w_{b'q'} \in [0, 3]$; the individual payoffs of buyers b and b' are given by $(u_{bq} = 0, u_{bq'} = 2 - w_{bq'})$ and $(u_{b'q'} = 3 - w_{b'q'}, u_{b0} = 0)$, respectively.

To see that corewise-stability is not adequate to define the cooperative equilibrium for this market, let $(u, w; \mu)$ be such that $u_{bq} = 1$, $u_{bq'} = 1$, $u_{b'q'} = 1$, $u_{b'0} = 0$; $w_{bq} = 2$, $w_{bq'} = 1$, $w_{b'q'} = 2$. That is, firm b hires workers q and q', obtains from each one of them a profit of one and pays two to q and one to q'; firm b' hires worker q' at a salary of two and obtains a profit of one. Observe that b' has quota of two, so it has one unfilled position. It happens that b' can pay more than two to q. Thus, if agents can communicate with each other and behave cooperatively, this outcome will not occur, because worker q will not accept to receive only two from firm b, since she knows that she can get more than two by working with firm b'. Hence, this outcome cannot be a cooperative equilibrium. Observe that $2 = u_{b'0} + w_{bq} < v_{b'q} = 3$, so this outcome is not stable. On the other hand, it is in the core. In fact, if there is a blocking coalition then it must contain $\{b', q\}$. These agents cannot increase their total payoffs by themselves; b' needs to hire both workers. However $\{b', q, q'\}$ does not block the outcome, because q' is worse off by taking only one job. Nevertheless, the coalition of all agents does not block the outcome, since b loses worker q, so it will be worse off.

Now, consider the outcome $(u', w'; \mu)$, where $u'_{bq} = 0$, $u'_{bq'} = 1$, $u'_{b'q'} = 1$, $u'_{b'0} = 0$; $w'_{bq} = 3$, $w'_{bq'} = 1$, $w'_{b'q'} = 2$. Firm b' cannot offer more than three to worker q, so the structure of the outcome cannot be ruptured. Then, although both outcomes $(u, w; \mu)$ and $(u', w'; \mu)$ are corewise-stable, only the second one can be expected to occur, so only this outcome is a cooperative equilibrium. Our explanation for this is that only $(u', w'; \mu)$ is stable.

The connection between the core, the set of stable payoffs and the set of competitive payoffs, exhibited in this example, can be better understood via Fig. 1. Now, if the reader prefers, sets B and Q are better interpreted as being the set of buyers and the set of sellers, respectively.

In Fig. 1, $C(W)$ is the set of seller's total payoffs, which

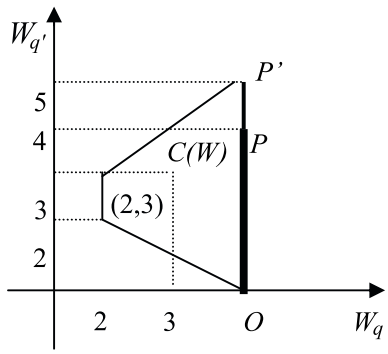

Two-Sided Matching Models, Figure 1
Core, stable and competitive equilibrium payoffs in Example 5

can be derived from any core payoff. The segment OP' is the set of sellers total payoffs which can be derived from any stable payoff. That is, $(W_q, W_{q'}) \in OP'$ if and only if there is a stable outcome $(u, w; \mu)$ such that $W_q = w_{bq}$ and $W_{q'} = w_{bq'} + w_{b'q'}$. The segment OP is the set of seller's total payoffs, which can be derived from any competitive equilibrium price. That is, $(W_q, W_{q'}) \in OP$ if and only if there is a competitive equilibrium price p such that $W_q = p_q$ and $W_{q'} = p_{q'} + p_{q'}$.

We can see that $C(W)$ is bigger than OP' which, in its turn, is bigger than OP. The point $(2, 3) \in C(W) - OP'$. It corresponds to the outcome (u, w, μ) described above that is in the core but is not stable.

It is clear in Fig. 1 that $C(W)$ is not a lattice, so the core is not a lattice under neither $\geqslant_B$ nor $\geqslant_Q$. In fact, the outcome which corresponds to the point $(3, 0)$ gives the individual payoff of 3 to q and two individual payoffs of zero to q'. On the other hand, the core outcome that corresponds to $(2, 3)$ gives the individual payoff of two to q . Then, the infimum (respectively supremum) of these two core payoffs under $\geqslant_Q$ (respectively $\geqslant_B$) gives payoff two to seller q and two individual payoffs of zero to seller q'. This payoff corresponds to the vector of total payoffs $(2, 0)$, which is not in $C(W)$.

It is evident that the stable payoff corresponding to point $P' = (3, 5)$ is the Q-optimal stable payoff. Seller q receives three and seller q' receives two from b and three from b'. Buyers get zero from the sellers. By applying the function f we obtain the Q-optimal competitive equilibrium payoff, corresponding to point $P = (3, 4)$, where q receives three and q' receives two from both buyers.

Point $O = (3, 0)$ corresponds to the outcome $(u'', w''; \mu)$, where $u''_{bq} = 0$, $u''_{bq'} = 2$, $u''_{b'q'} = 3$, $u''_{b'0} = 0$; $w''_{bq} = 3$, $w''_{bq'} = 0$, $w''_{b'q'} = 0$. Payoff (u'', w'') is the B-optimal stable payoff, the B-optimal competitive equilibrium

payoff and the B-optimal core payoff. It is the worst stable payoff and the worst competitive equilibrium payoff for the sellers. However, it is not the worst core payoff for the sellers since it is the best core payoff for q. Indeed, as it can be observed, there is no minimum core payoff for the sellers.

A fruitful line of investigation has been the design of mechanisms to produce a competitive equilibrium price. Demange, Gale and Sotomayor [19] propose a generalization of the English auction for the assignment game, which yields the minimum competitive equilibrium price in a finite number of steps. In the same spirit, Sotomayor [83] presents a descending bid auction mechanism which leads the maximum competitive equilibrium price as a generalization of the Dutch auction. Gül and Stacchetti [34] obtain the minimum competitive equilibrium price through a generalization of the auction of Demange et al. [19] to the many-to-one case in which the utility functions satisfy the gross substitute condition. Sotomayor [90] obtains the same result by considering a dynamic mechanism for the many-to-many case with additively separable utility functions. This mechanism leads to the B-optimal stable payoff. Using the symmetry of the model, the Q-optimal stable payoff is obtained by reverting the roles between buyers and sellers in the mechanism.

A number of works related to the assignment game can be found in the literature. Demange and Gale [18] generalize the assignment game by allowing agents' preferences to be represented by any continuous utility function in the money variable. For that model, Roth and Sotomayor [67] generalize a previous result of Rochford [58], using Tarski's fixed point theorem, and show that the set of fixed points of a "rebargaining" function is a subset of the core, which maintains the lattice structure of the core. Another approach of the many-to-many assignment game with additively separable utilities is treated in Sotomayor [75]. There, agents are not allowed to negotiate their individual payments and act in blocks. An outcome only specifies their total payoffs. In this model, the core coincides with the set of stable payoffs and is not a lattice. Also, pairwise-stability is not equivalent to corewise-stability. Sotomayor [84] considers an extension of the assignment game to a many-to-many matching model in the context of firms and workers in which the quotas of the agents are not the number of partnerships they are allowed to form. Instead, they are given by the units of labor time they can supply or employ. Bikhchandani and Mamer [10] analyze the existence of market clearing prices in an exchange economy in which agents have interdependent values over several indivisible objects. Although an agent can be both a buyer and a seller, such an exchange economy can be transformed into a many-to-one matching market where each seller owns only one object and buyers want to buy a bundle of objects, and can be viewed as an extension of the assignment game. See also Demange [17], Leonard [48], Perez–Castrillo and Sotomayor [56], Sotomayor [86], Thompson [97] and Kaneko [39].

Hybrid One-to-One Matching Model

The hybrid one-to-one matching model is the name given in the literature to a unified model due to Eriksson and Karlander [29] and inspired in the unification proposed in Roth and Sotomayor [70]. Agents from the marriage market and the assignment market are put together so that they can trade with each other in the same market. We can interpret the hybrid model as being a labor market of firms and workers: P is the set of firms and Q is the set of workers. There are two classes of agents in each set: rigid agents and flexible agents. For each pair $(p, q) \in P \times Q$ there is a number c_{pq} representing the productivity of the pair. If a firm p hires worker q and both agents are flexible then the number c_{pq} is allocated into salary v_q for the worker and profit u_p for the firm as a result of a negotiation process. If one of the agents is rigid, then the payoffs of the agents are pre-set and fixed and are part of the job description. In this case the profit of p and the salary of q will be a_{pq} and b_{pq}, respectively. The definitions of feasible outcome, corewise-stable outcomes and setwise-stable outcomes are straightforward extensions from the respective concepts for the non-hybrid models. Therefore, the concepts of setwise-stability and corewise-stability are equivalent.

This model is motivated by the fact that, in practice, a wide range of real-world matching markets are neither completely discrete nor completely continuous. In the United States, for example, new law school graduates may enter the market for associate positions in private law firms, which negotiate salaries, or they may seek employment as law clerks to federal circuit court judges, which are civil service positions with predetermined fixed salaries. In the market for academic positions and professors, for example, the American universities compete with each other in terms of salaries, while the French public universities offer a preset and fixed salary. In Brazil, new professors may enter the market for permanent positions (with preset and fixed salaries) in federal universities, or they may seek employment in private universities, which do not offer such positions, but compensate the entrants with better and negotiable salaries.

Eriksson and Karlander [29] present an algorithm to find a stable outcome under the assumption that the num-

bers a_{pq}, b_{pq} and c_{pq} are integer numbers of some unit. A non-constructive proof of the existence result without imposing any restriction is provided in Sotomayor [81]. In this paper it is proved that, under the assumption that the core, C, is equal to the strong core, $C*$, the main properties that characterize the stable payoffs of the marriage and of the assignment models carry over to the hybrid model. That is, for the hybrid model,

C1. *Let (u, w) and (u', w') be stable payoffs for the hybrid model. If $C = C*$ then $u \geqslant u'$ if and only if $w' \geqslant w$.*
C2. *If $C = C*$ and an agent is unmatched under some stable outcome then he/she gets a zero payoff under any other stable outcome.*
C3. *If $C = C*$ then the set of stable payoffs forms a complete lattice under the partial orders $\geqslant_P$ and $\geqslant_Q$.*
C4. *If $C = C*$ then there is a P-optimal stable payoff $(u*, w_*)$ with the property that for any stable payoff (u, w), $u* \geqslant u$ and $w_* \leqslant w$; there is a Q-optimal stable payoff $(u_*, w*)$ with symmetrical properties.*

Sotomayor [92] studies the hybrid model without imposing the assumption that the core is equal to the strong core. Instead she assumes that the preferences of the rigid agents, as well as the preferences of the flexible agents over rigid agents, are strict. It is shown there that the core of the hybrid model has a non-standard algebraic structure given by the disjoint union of complete lattices endowed with the properties above. The extreme points of the lattices of the core partition are called *quasi-optimal stable payoffs for firms* and *quasi-optimal stable payoffs for workers*. When the workers are always flexible, then the marriage market is obtained when the flexible firms leave the hybrid market and the assignment game is obtained when the rigid firms leave the hybrid market.

Each subset of the core partition of the hybrid model is obtained as follows. For any matching μ which is compatible with some stable payoff, decompose the market participants into two disjoint subsets. One subset contains all rigid firms and their mates at μ and the other one contains all flexible firms, their mates at μ and the unmatched workers. Now, fix such a partition of the agents. The desired subset $C(\mu)$ of the core partition is formed with the core payoffs $(u, w; \mu')$ such that all agents in the first set are matched among themselves under μ' and all agents in the second set are matched among themselves under μ'.

Clearly, as the rigid firms exit the hybrid market, the core partition for the corresponding assignment market is reduced to only one set, since any stable matching is compatible with any core payoff by property B1. An analogous result holds as the flexible firms leave the hybrid market, due to the fact that the matched agents in the marriage market are the same at every stable matching, which is implied by property A2. Therefore, as all flexible firms leave the hybrid market, or as all rigid firms leave the hybrid market, the restriction of the algebraic structure to the core of the resulting non-hybrid market is that of a complete lattice. Then, the extreme points of the resulting lattice are exactly the firm-optimal and the worker-optimal stable payoffs.

This algebraic structure was used in Sotomayor [92] to investigate the comparative effects on the quasi-optimal stable payoffs for firms and on the quasi-optimal stable payoffs for workers caused by the entrance of rigid firms into the assignment market or by the entrance of flexible firms into the marriage model. The results of that paper can be summarized as follows: *Whether agents are allocated according to a quasi-optimal stable payoff for firms or according to a quasi-optimal stable payoff for workers, it will always be the case that if flexible firms enter the rigid market, no rigid firm will be made better off and no worker will be made worse off; if rigid firms enter the flexible market, no flexible firm will be made better off and no worker will be made worse off.*

Comparative static results of adding agents from the same side to the marriage market or to the assignment market have been obtained in the literature under the assumption that the agents are allocated according to the optimal-stable outcome for firms or according to the optimal-stable outcome for workers. However, in the approach considered in Sotomayor [92],

1. *The firms that are added are different from the firms, which are already in the market.* For example, in the marriage market, where utility is non-transferable, the comparative static adds firms with flexible wages who can transfer utility.
2. *The points which are compared belong to cores with quite distinct algebraic structures.*
3. *There may exist several quasi-optimal stable outcomes for firms and several quasi-optimal stable outcomes for workers in the hybrid model. Despite the multiplicity of these outcomes, all of them reveal the same kind of comparative static effects.*

Therefore, the result above has no parallel in the non-hybrid models.

It is argued in Sotomayor [92] that if the resulting core partition is not reduced to only one set when, say, the flexible firms leave the hybrid model, the comparative statics may be meaningless. This happens, for example, if we define a set of the core partition as the set of all stable payoffs compatible with some given matching. Then each lattice of the core partition for the marriage market has only one

stable matching, which is both the supremum and the infimum of the lattice. Of course the distinctions between, say, the best stable payoff for workers of some lattice of the core partition of the hybrid market and an arbitrary core point of the marriage model, cannot be attributed to the entrance of the flexible firms into the marriage market.

Results of comparative statics were originally obtained by Gale and Sotomayor [32] for the marriage model and the college admission model: *If agents from the same side of the market enter the market then no agent from this side is better off and no agent of the opposite side is worse off, if any of the two optimal stable matchings prevails.* A similar result was proved by Demange and Gale [18] for a continuous one-to-one matching model that includes the assignment game.

For the assignment game, Shapley [71] showed that the optimal stable payoff for an agent weakly decreases when another agent is added to the same side and weakly increases when another agent is added to the other side. Still with regard to the assignment game, Mo [53] showed that if the incoming worker is allocated to some firm in some stable outcome for the new market, there is a set of agents such that every firm is better off and every worker is worse off in the new market than in the previous one. A symmetric result holds when the incoming agent is a firm. An analogous result is demonstrated by Roth and Sotomayor [69] for the marriage market.

For the many-to-one matching markets with substitutable preferences, Kelso and Crawford [42] showed that, within the context of (flexible) firms and workers, *the addition of one or more firms to the market weakly improves the workers' payoffs, and the addition of one or more workers weakly improves the firms' payoffs, under the firm-optimal stable allocation*. Similar conclusions were obtained by Crawford [14] for a many-to-many matching model with strict and substitutable preferences, by comparing pairwise-stable outcomes instead of setwise-stable outcomes.

Incentives

The strategic questions that emerge when a stable revelation mechanism is adopted concerns its non-manipulability and, for the games induced by the mechanism, the existence of strategic equilibria and the implementability of the set of stable matchings via such equilibria. For the marriage model with strict preferences the equilibrium analysis of a game induced by a stable matching mechanism leads to the following results:

1. (Impossibility Theorem) (Roth and Sotomayor, [69]) *When any stable mechanism is applied to a marriage market in which preferences are strict and there is more than one stable matching, then at least one agent can profitably misrepresent his or her preferences, assuming the others tell the truth. (This agent can misrepresent in such a way as to be matched to his or her most preferred achievable mate under the true preferences at every stable matching under the mis-stated preferences).*
2. (Limits on successful manipulation.) (Demange, Gale, and Sotomayor [20]). *Let P be the true preferences (not necessarily strict) of the agents, and let P' differ from P in that some coalition C of men and women misstate their preferences. Then there is no matching μ, stable for P', which is preferred to every stable matching under the true preferences P by all members of C.*

A corollary of this result is due to Dubins and Freedman [21] which states that *the man-optimal stable matching mechanism is non-manipulable, individually and collectively, by the men.*

3. (Gale and Sotomayor [31]) *When all preferences are strict, let μ be any stable matching for (F, W, P). Suppose each woman w in $\mu(F)$ chooses the strategy of listing only $\mu(w)$ on her stated preference list of acceptable men (and each man states his true preferences). This is a Nash equilibrium in the game induced by the man-optimal stable matching mechanism (and μ is the matching that results).*
4. (Roth, [61]) *Suppose each man chooses his dominant strategy and states his true preferences, and the women choose any set of strategies (preference lists) $P'(w)$ that form a Nash equilibrium for the revelation game induced by the man-optimal stable mechanism. Then the corresponding man-optimal stable matching for (F, W, P') is one of the stable matchings for (F, W, P).*
5. (Gale and Sotomayor, [31]) *Suppose each man chooses his dominant strategy and states his true preferences, and the women truncate their true preferences at the mate they get under the woman-optimal stable mechanism. This profile of preferences is a strong equilibrium for the women in the game induced by the man-optimal stable mechanism (and the woman-optimal stable matching under the true preferences is the matching that results).*

Results (3) and (4) imply that the man-optimal stable mechanism implements the core correspondence via Nash equilibria.

For the college admission model with responsive and strict preferences, the theorem of Dubins and Freedman implies *that the student-optimal stable mechanism is non-manipulable individually and collectively by the students.*

Roth [65] shows through an example that *the college-optimal stable mechanism is manipulable by the colleges* due to the fact that the colleges may have a quota greater than one.

Sotomayor [78,82,94] analyzes the strategic behavior of the students in a school choice model where participants have strict preferences over individuals. This paper proves that *the college-optimal stable mechanism implements the set of stable matchings via the Nash equilibrium concept.* When some other stable mechanism is used, an example shows that the strategic behavior of the students may lead to unstable matchings under the true preferences. A sufficient condition for the stability of the Nash equilibrium outcome is then proved to be that the set of stable matchings for the Nash equilibrium profile is a singleton. A random stable matching mechanism is proposed and the Nash equilibrium concept ex-ante is shown to be equivalent to the Nash equilibrium concept ex-post of the game induced by such a mechanism. This refinement of the Nash equilibrium concept is called *Nash equilibrium in the strong sense. Under this equilibrium concept any stable matching mechanism (and in particular the random stable matching mechanism) implements the set of stable matchings.* Also, *if the students only play truncations of the true preferences, any stable matching mechanism implements the student-optimal stable matching via strong equilibrium in the strong sense and Nash equilibrium in the strong sense.*

Ergin and Sönmez [28] and Pathak and Sönmez [55] analyze the Boston mechanism, which is used to assign students to schools in many cities in the US and show that students' parents do not have incentives to report preferences truthfully.

Ma [49] analyzes the strategic behavior of both students and colleges in the college admission model with responsive preferences. This author proves that *the set of stable matchings is implemented by any stable mechanism via rematching proof equilibrium and strong equilibrium in truncation strategies at the match point.*

The implementability of the set of stable matchings through stable and non-necessarily stable mechanisms has also been investigated by several authors. Alcalde [5] presents a mechanism for the marriage market closely related to the algorithm of Gale and Shapley, which implements the core correspondence in undominated equilibria. Kara and Sönmez [40] analyze the problem of implementation in the college admission market. They show that the set of stable matchings is implementable in Nash equilibrium. Nevertheless, no subset of the core is Nash implementable. Romero–Medina [59] studies the mechanism employed by the Spanish universities to distribute the students to colleges, which can produce unstable matchings for the stated preferences. However, when students play in equilibrium, only stable allocations are reached. Sotomayor [85] investigates a mechanism for the marriage model which is not designed for producing stable matchings. Here also the equilibrium outcomes are stable matchings under the true preferences.

For the discrete many-to-one matching model with responsive preferences, Alcalde and Romero–Medina [7] analyze the following mechanism: firms announce a set of workers they want to hire. Then each worker selects the firm she wants to work for. This paper proves that such a mechanism implements the set of stable matchings in subgame perfect Nash equilibrium. For the many-to-many case, Sotomayor [88] shows that this result does not carry over. This paper proves that subgame perfect Nash equilibria always exist, while strong equilibria may not exist. The subgame perfect Nash equilibrium outcomes are precisely the pairwise-stable matchings, which may be out of the core when the preferences of the agents in one of the sides are not maximin. Under this condition, the equilibrium outcomes are the setwise-stable matchings and every subgame perfect Nash equilibrium is a strong equilibrium.

By assuming non-strict preferences, Abdulkadiroğlu, Pathak, Roth and Sönmez [2] show that no mechanism (stable or not, and Pareto optimal or not), which is better for the students than the student proposing deferred acceptance algorithm with tie breaking, can be strategy proof.

Sönmez [73] analyzes a model which he calls the *generalized indivisible allocation problem* that includes the roommate and the marriage markets. He looks for conditions which explain the differences on strategy-proofness results that have been generated in the literature. He shows how some of the results in the literature can be seen as corollaries of his results.

Ehlers and Massó [27] study Bayesian Nash equilibria of stable mechanisms (such as the NRMP) in matching markets under incomplete information. They show that truth-telling is an equilibrium of the Bayesian revelation game induced by a common belief and a stable mechanism if and only if all the profiles in the support of the common belief have singleton cores.

For the continuous matching models the idea is to use competitive equilibrium as an allocation mechanism to produce outcomes with the desirable properties of fairness and efficiency. It involves having agents specify their supply and demand functions. The competitive equilibria are then calculated and allocations are made accordingly. Demange [17] and Leonard [48] considered the assignment game of Shapley and Shubik and, independently, proved that the allocation mechanism that yields the minimum

competitive equilibrium prices is individually non-manipulable by the buyers. Demange and Gale [18] consider a one-to-one matching model in which the utilities are continuous functions in the money variable and not necessarily additively separable. These authors prove a sort of non-manipulability theorem which states that *if the mechanism which produces the buyer-optimal stable payoff is adopted then no coalition of buyers by falsifying demands can achieve, only through the mechanism, higher payoffs to all of its members.* We added *only through the mechanism* because this model allows monetary transfers within any coalition. As in the marriage model, this result is an immediate consequence of a more general theorem due to Sotomayor [74] which states the following: *Let* $(u', w'; \mu)$ *be any stable outcome for the market* M' *where* $B' \cup Q'$ *is the set of agents who misrepresent their utility functions. Let* $(u*, w*)$ *be the true payoff under* $(u', w'; \mu)$. *Then, there exists a stable payoff* (u, w) *for the original market such that* $u_b \geqslant u*_b$ *for at least one* b *in* B' *or* $w_q \geqslant w'_q$ *for at least one* q *in* Q'.

Demange and Gale [18] also addresses the strategic behavior by the sellers when the mechanism produces the buyer-optimal stable payoff. These authors show that by specifying their supply functions appropriately the sellers can force, by strong Nash equilibrium strategies, the payoff to be given by the maximum rather than the minimum equilibrium price. Under the assumption that the sellers only manipulate their reservation prices then, if a profile of strategies does not give the maximum equilibrium price allocation, then either some seller is using a dominated strategy or the strategy profile is not a Nash equilibrium. For this model Sotomayor [74] proves that the outcome produced by a Nash equilibrium strategy is stable for the original market.

Sotomayor [87] considers, for the assignment game of Shapley and Shubik, the strategic games induced by a class of market clearing price mechanisms. In these procedures, buyers and sellers, in different stages reveal their demand and supply functions and a competitive equilibrium is produced by the mechanism. For each vector of reservation prices selected by the sellers, the buyers play the subgame that starts and can force the buyer-optimal stable payoff through Nash equilibrium strategies. However sellers can reverse this outcome by forcing the subgame perfect equilibrium allocation to be the seller-optimal stable payoff for the original market.

Kamecke [38] and Perez–Castrillo and Sotomayor [56] consider the assignment game of Shapley and Shubik. The former paper presents two mechanisms for this market. In the first one, agents act simultaneously. In the second game, the strategies are chosen sequentially. These mechanisms implement the social choice correspondences that yield the core and the optimal stable payoff for the sellers, respectively. The second paper analyzes a sequential mechanism, which implements the social choice correspondence that yields the optimal stable payoff for the sellers.

Future Directions

In this section we present some directions for future investigations and some open problems which have intrigued matching theorists.

1. The discrete two-sided matching models with non-necessarily strict preferences have been explored very little in the literature. In the discrete models under strict preferences and in the continuous models, due to the fact that there is no weak blocking pairs, the set of Pareto-stable outcomes coincides with the set of setwise-stable outcomes. However, under weak preferences, setwise-stable matchings may not be Pareto-optimal. Sotomayor [93] proposes that in this case the Pareto-stability concept, which requires that the matching is stable and Pareto optimal, should be considered the natural solution concept. The justification for the Pareto-stability concept relies in the argument that in a decentralized setting, where agents freely get together in groups, recontracts between pairs of agents already allocated according to a stable matching leading to a (weak) Pareto improvement of the original matching should be allowed. Thus, weak blockings can upset a matching once they come from the grand coalition.
   We think that the study of the discrete two-sided matching models with non-necessarily strict preferences and the search for algorithms to produce the Pareto-stable matchings is a new and interesting line of investigation.
2. Consider the hybrid model where no worker is rigid. If rigid firms enter the flexible market or flexible firms enter the rigid market then no firm gains and no worker loses if a quasi-optimal stable outcome for one of the sides always prevails. Suppose now that some rigid firm becomes flexible or some flexible firm becomes rigid. What kind of comparative static effect is caused by this change in the market?
3. One line of investigation not yet explored in the literature concerns the incentives faced by the agents in the hybrid model when some stable allocation mechanism is used.
4. Consider the discrete many-to-many matching market with substitutable preferences where a matching μ is

feasible if $Ch_y(\mu(y)) = \mu(y)$ for every agent y. Is the core always non-empty for this model?

5. Consider the assignment game of Shapley and Shubik in the context of buyers and sellers. Consider a sealed bid auction in which the buyers select a monetary value for each of the items. The auctioneer then chooses a competitive equilibrium price vector for the profile of selected values, according to some pre-set probability distribution. It is of theoretical interest the investigation of the buyers' strategic behavior.
6. We know that the core of the many-to-many assignment model of Sotomayor [75] in which the agents negotiate in blocks is not a lattice. However, a problem that is still open is to know if the optimal stable payoffs for each side of the market always exist.

Bibliography

1. Abdulkadiroglu A, Sönmez T (2003) School Choice: A Mechanism Design Approach. Am Econ Rev 93(3):729–747
2. Abdulkadiroglu A, Pathak P, Roth A, Sönmez T (2006) Changing the Boston School Choice Mechanism: Strategy-proofness as Equal Access, working paper. Boston College and Harvard University, Boston
3. Abeledo H, Isaak G (1991) A characterization of graphs which assure the existence of stable matchings. Math Soc Sci 22(1):93–96
4. Adachi H (2000) On a characterization of stable matchings. Econ Lett 68(1):43–49
5. Alcalde J (1996) Implementation of Stable Solutions to Marriage Problems. J Econ Theory 69(1):240–254
6. Alcalde J, Pérez-Castrillo D, Romero-Medina A (1998) Hiring Procedures to Implement Stable Allocations. J Econ Theory 82(2):469–480
7. Alcalde J, Romero-Medina A (2000) Simple Mechanisms to Implement the Core of College Admissions Problems. Games Econ Behav 31(2):294–302
8. Balinski M, Sönmez T (1999) A Tale of Two Mechanisms: Student Placement. J Econ Theory 84(1):73–94
9. Bardella F, Sotomayor M (2006) Redesign and analysis of an admission market to the graduate centers of economics in Brazil: a natural experiment in market organization, working paper. Universidade de São Paulo, São Paulo
10. Bikhchandani S, Mamer J (1997) Competitive Equilibrium in an Exchange Economy with Indivisibilities. J Econ Theory 74(2):385–413
11. Birkhoff G (1973) Lattice theory, vol v. 25. Colloquium publications. American Mathematical Society, Providence
12. Blair C (1988) The Lattice Structure of the Set of Stable Matchings with Multiple Partners. Math Oper Res 13(4):619–628
13. Chung K (2000) On the Existence of Stable Roommate Matchings. Games Econ Behav 33(2):206–230
14. Crawford V (1991) Comparative statics in matching markets. J Econ Theory 54(2):389–400
15. Crawford V (2008) The Flexible-Salary Match: A Proposal to Increase the Salary Flexibility of the National Resident Matching Program. J Econ Behav Organ 66(2):149–160
16. Crawford V, Knoer E (1981) Job Matching with Heterogeneous Firms and Workers. Econometrica 49(2):437–450
17. Demange G (1982) Strategyproofness in the assignment market game, working paper. Ecole Polytechnique, Laboratoire D'Econometrie, Paris
18. Demange G, Gale D (1985) The Strategy Structure of Two-Sided Matching Markets. Econometrica 53(4):873–888
19. Demange G, Gale D, Sotomayor M (1986) Multi-Item Auctions. J Political Econ 94(4):863–872
20. Demange G, Gale D, Sotomayor M (1987) A further note on the stable matching problem. Discret Appl Math 16(3):217–222
21. Dubins L, Freedman D (1981) Machiavelli and the Gale-Shapley Algorithm. Am Math Mon 88(7):485–494
22. Dutta B, Massó J (1997) Stability of Matchings When Individuals Have Preferences over Colleagues. J Econ Theory 75(2): 464–475
23. Echenique F, Oviedo J (2004) Core many-to-one matchings by fixed-point methods. J Econ Theory 115(2):358–376
24. Echenique F, Oviedo J (2006) A theory of stability in many-to-many matching markets. Theor Econ 1(2):233–273
25. Echenique F, Yenmez M (2007) A solution to matching with preferences over colleagues. Games Econ Behav 59(1):46–71
26. Eeckhout J (2000) On the uniqueness of stable marriage matchings. Econ Lett 69(1):1–8
27. Ehlers L, Massó J (2004) Incomplete Information and Small Cores in Matching Markets, working paper. CREA, Barcelona
28. Ergin H, Sönmez T (2006) Games of school choice under the Boston mechanism. J Public Econ 90(1–2):215–237
29. Eriksson K, Karlander J (2000) Stable matching in a common generalization of the marriage and assignment models. Discret Math 217(1):135–156
30. Gale D, Shapley L (1962) College Admissions and the Stability of Marriage. Am Math Mon 69(1):9–15
31. Gale D, Sotomayor M (1985) Ms. Machiavelli and the Stable Matching Problem. Am Math Mon 92(4):261–268
32. Gale D, Sotomayor M (1983,1985) Some remarks on the stable matching problem. Discret Appl Math 11:223–232
33. Gül F, Stacchetti E (1999) Walrasian Equilibrium with Gross Substitutes. J Econ Theory 87(1):95–124
34. Gül F, Stacchetti E (2000) The English Auction with Differentiated Commodities. J Econ Theory 92(1):66–95
35. Gusfield D (1988) The Structure of the Stable Roommate Problem: Efficient Representation and Enumeration of All Stable Assignments. SIAM J Comput 17:742–769
36. Hatfield J, Milgrom P (2005) Matching with Contracts. Am Econ Rev 95(4):913–935
37. Irving R (1985) An efficient algorithm for the stable roommates problem. J Algorithms 6:577–595
38. Kamecke U (1989) Non-cooperative matching games. Int J Game Theory 18(4):423–431
39. Kaneko M (1982) The central assignment game and the assignment markets. J Math Econ 10(2–3):205–232
40. Kara T, Sönmez T (1996) Nash Implementation of Matching Rules. J Econ Theory 68(2):425–439
41. Kara T, Sönmez T (1997) Implementation of college admission rules. J Econ Theory 9(2):197–218
42. Kelso Jr A, Crawford V (1982) Job Matching, Coalition Formation, and Gross Substitutes. Econometrica 50(6):1483–1504
43. Kesten O (2004) Student placement to public schools in the US: Two new solutions, working paper. University of Rochester, Rochester

44. Kesten O (2006) On two competing mechanisms for priority-based allocation problems. J Econ Theory 127(1):155–171
45. Knuth D (1976) Marriage Stables. Les Presses de l'Université de Montréal, Montréal
46. Konishi H, Ünver MU (2006) Credible group stability in many-to-many matching problems. J Econ Theory 127(1):57–80
47. Kraft C, Pratt J, Seidenberg A (1959) Intuitive Probability on Finite Sets. Ann Math Stat 30(2):408–419
48. Leonard H (1983) Elicitation of Honest Preferences for the Assignment of Individuals to Positions. J Political Econ 91(3): 461–479
49. Ma J (2002) Stable matchings and the small core in Nash equilibrium in the college admissions problem. Rev Econ Des 7(2):117–134
50. Martínez R, Massó J, Neme A, Oviedo J (2001) On the lattice structure of the set of stable matchings for a many-to-one model. Optimization 50(5):439–457
51. Martínez R, Massó J, Neme A, Oviedo J (2004) An algorithm to compute the full set of many-to-many stable matchings. Math Soc Sci 47(2):187–210
52. McVitie D, Wilson L (1970) Stable marriage assignment for unequal sets. BIT Numer Math 10(3):295–309
53. Mo J (1988) Entry and structures of interest groups in assignment games. J Econ Theory 46(1):66–96
54. Ostrovsky M (2008) Stability in Supply Chain Networks. Am Econ Rev 98(3):897–923
55. Pathak P, Sönmez T (2006) Leveling the Playing Field: Sincere and Strategic Players in the Boston Mechanism, working paper. Boston College and Havard University, Boston
56. Pérez-Castrillo D, Sotomayor M (2002) A Simple Selling and Buying Procedure. J Econ Theory 103(2):461–474
57. Pycia M (2007) Many-to-One Matching with Complementarities and Peer Effects, working paper. Penn State Working Paper, Pennsylvania
58. Rochford S (1984) Symmetrically Pairwise-Bargained Allocations in an Assignment Market. J Econ Theory 34(2):262–281
59. Romero-Medina A (1998) Implementation of stable solutions in a restricted matching market. Rev Econ Des 3(2):137–147
60. Roth A (1982) The Economics of Matching: Stability and Incentives. Math Oper Res 7(4):617–628
61. Roth A (1984) Misrepresentation and Stability in the Marriage Problem. J Econ Theory 34(2):383–387
62. Roth A (1984) Stability and Polarization of Interests in Job Matching. Econometrica 52(1):47–58
63. Roth A (1984) The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory. J Political Econ 92(6):991–1016
64. Roth A (1985) Conflict and Coincidence of Interest in Job Matching: Some New Results and Open Questions. Math Oper Res 10(3):379–389
65. Roth A (1985) The College Admissions Problem is not Equivalent to the Marriage Problem. J Econ Theory 36(2):277–288
66. Roth A (1986) On the Allocation of Residents to Rural Hospitals: A General Property of Two-Sided Matching Markets. Econometrica 54(2):425–427
67. Roth A, Sotomayor M (1988) Interior points in the core of two-sided matching markets. J Econ Theory 45(1):85–101
68. Roth A, Sotomayor M (1989) The College Admissions Problem Revisited. Econometrica 57(3):559–570
69. Roth A, Sotomayor M (1990) Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis, vol 18. In: Econometric Society Monographs. Cambridge University Press, New York
70. Roth A, Sotomayor M (1996) Stable outcomes in discrete and continuous models of two-sided matching: A unified treatment. Braz Rev Econom 16:1–4
71. Shapley L (1962) Complements and substitutes in the Optimal Assignment Problem. Navals Res Logist Q 9:45–48
72. Shapley L, Shubik M (1972) The Assignment Game I: The core. Int J Game Theory 1(1):111–130
73. Sönmez T (1999) Strategy-Proofness and Essentially Single-Valued Cores. Econometrica 67(3):677–689
74. Sotomayor M (1986) On incentives in a two-sided matching market, working paper. Depart of Mathematics, PUC/RJ, Rio de Janeiro
75. Sotomayor M (1992) The multiple partners game in Equilibrium and dynamics. In: Majumdar M (ed) Essays in honour of David Gale. MacMillan Press Ltd, New York, pp 322–336
76. Sotomayor M (1996) A Non-constructive Elementary Proof of the Existence of Stable Marriages. Games Econ Behav 13(1):135–137
77. Sotomayor M (1996) Admission mechanisms of students to colleges. A game-theoretic modeling and analysis. Braz Rev Econom 16(1):25–63
78. Sotomayor M (1998) The strategy structure of the college admissions stable mechanisms. In: Annals of Jornadas Latino Americanas de Teoria Econômica, San Luis, Argentina, 2000; First World Congress of the Game Theory Society, Bilbao, 2000; 4th Spanish Meeting, Valencia, 2000; International Symposium of Mathematical Programming, Atlanta, 2000; World Congress of the Econometric Society, Seattle, 2000, http://www.econ.fea.usp.br/marilda/artigos/roommates_1.doc
79. Sotomayor M (1999) The lattice structure of the set of stable outcomes of the multiple partners assignment game. Int J Game Theory 28(4):567–583
80. Sotomayor M (1999) Three remarks on the many-to-many stable matching problem. Math Soc Sci 38(1):55–70
81. Sotomayor M (2000) Existence of stable outcomes and the lattice property for a unified matching market. Math Soc Sci 39(2):119–132
82. Sotomayor M (2000) Reaching the core through college admissions stable mechanisms. In: Annals of abstracts of the following congresses: International Conference on Game Theory, 2001, Stony Brook; Annals of Brazilian Meeting of Econometrics, Salvador, Brazil, 2001; Latin American Meeting of the Econometric Society, Buenos Aires, Argentina, 2001, http://www.econ.fea.usp.br/marilda/artigos/reaching_%20core_random_stable_allocation_mechanisms.pdf
83. Sotomayor M (2002) A simultaneous descending bid auction for multiple items and unitary demand. Revista Brasileira Economia 56:497–510
84. Sotomayor M (2003) A labor market with heterogeneous firms and workers. Int J Game Theory 31(2):269–283
85. Sotomayor M (2003) Reaching the core of the marriage market through a non-revelation matching mechanism. Int J Game Theory 32(2):241–251
86. Sotomayor M (2003) Some further remark on the core structure of the assignment game. Math Soc Sci 46:261–265
87. Sotomayor M (2004) Buying and selling strategies in the assignment game, working paper. Universidade São Paulo, São Paulo

88. Sotomayor M (2004) Implementation in the many-to-many matching market. Games Econ Behav 46(1):199–212
89. Sotomayor M (2005) The roommate problem revisited, working paper. Universidade São Paulo, São Paulo. http://www.econ.fea.usp.br/marilda/artigos/THE_ROOMMATE_PROBLEM_REVISITED_2007.pdf
90. Sotomayor M (2006) Adjusting prices in the many-to-many assignment game to yield the smallest competitive equilibrium price vector, working paper. Universidade de São Paulo, São Paulo. http://www.econ.fea.usp.br/marilda/artigos/A_dynam_stable-mechan_proof_lemma_1.pdf
91. Sotomayor M (2007) Connecting the cooperative and competitive structures of the multiple partners assignment game. J Econ Theory 134(1):155–174
92. Sotomayor M (2007) Core structure and comparative statics in a hybrid matching market. Games Econ Behav 60(2):357–380
93. Sotomayor M (2008) The Pareto-Stability concept is a natural solution concept for the Discrete Matching Markets with indifferences, working paper. Universidade São Paulo, São Paulo. http://www.econ.fea.usp.br/marilda/artigos/ROLE_PLAYED_SIMPLE_OUTCOMES_STABLE_COALITI_3.pdf
94. Sotomayor M (2007) The stability of the equilibrium outcomes in the admission games induced by stable matching rules. Inter Jour Game Theory 36(3–4):621–640
95. Tan J (1991) A Necessary and Sufficient Condition for the Existence of a Complete Stable Matching. J Algorithms 12(1): 154–178
96. Tarski A (1955) A lattice-theoretical fixpoint theorem and its applications. Pac J Math 5(2):285–309
97. Thompson G (1980) Computing the core of a market game. In: Fiacco AA, Kortane K (eds) Extremal Methods and Systems Analysis, vol 174. Springer, New York, pp 312–324

U

Ultracold Atomic Gases: Novel States of Matter

Ludwig Mathey[1], Shan-Wen Tsai[2], Antonio H. Castro Neto[3]
[1] Harvard University, Cambridge, USA
[2] University of California, Riverside, USA
[3] Boston University, Boston, USA

Article Outline

Glossary
Definition of the Subject
Introduction
One-Dimensional Lattices
Phase-Locking Transition of Coupled Low-Dimensional Superfluids
Bose–Fermi Mixtures in Two-Dimensional Optical Lattices
Conclusions
Bibliography

Glossary

Bose–Einstein condensation (BEC) Low temperature phase of systems of identical bosons, characterized by superfluidity.

Boson Particles with integer spin $S = 0, 1, 2, \ldots$ Mediators of interactions, such as photons and gluons are bosons. Objects made of an even number of fermions are bosons: positronium (electron + positron), meson (two quarks), ^{87}Rb (37 protons, 48 neutrons and 37 electrons), ^{7}Li (3 protons, 4 neutrons, 3 electrons).

Cooper pairs At low temperatures and for attractive interactions fermions form a superconducting state, in which fermions form pairs which condense.

Fermi surface Since fermions obey Pauli's exclusion principle, the ground state of N non-interacting fermions in d-dimensions is the state with the N lowest energy states occupied. In momentum space the last occupied state and the first unoccupied state define a surface of dimensions $d - 1$, called the Fermi surface.

Fermion Particles with half-odd integer spin $S = 1/2, 3/2, 5/2, \ldots$ Examples include elementary particles such as electrons and quarks. Objects made of an odd number of fermions are also fermionic, such as protons, ^{40}K (19 protons, 21 neutrons, and 19 electrons), and ^{6}Li (3 protons, 3 neutrons, and 3 electrons).

Laser cooling In a typical experimental setup, the atoms are cooled to the regime of $10^2 \mu K$, by using pairs of counterpropagating laser beams that are slightly red-detuned below an atomic transition. Due to the Doppler effect the atoms can only absorb a photon if they travel towards the beam with a high velocity. From that process the atoms experience a recoil, which slows them down.

Evaporative cooling To slow the atoms down further, to the μK regime, one applies radio frequency radiation that flips the internal state to a high-field seeking, i. e. non-trapped, state in such a way, that only atoms of high kinetic energy can escape. Due to thermalization, this leads to cooling of the remaining atomic ensemble.

Magnetic trap The atoms are trapped by applying a spatially inhomogeneous magnetic field. This field leads to an energy shift due to the Zeeman effect, which the atoms experience as an external potential, for large energy splittings of the magnetic levels. Different geometric designs are in use, such as the TOP trap, or the Ioffe–Pritchard trap.

Optical lattice Counterpropagating laser beams create a standing wave field, which the atoms experience as a periodic potential, due to the ac Stark shift. If the temperature and all energy scales are small compared to the energy splitting due the spatial confinement in each well, this system is well approximated by a Hubbard model, i. e. by taking into account nearest-neighbor hopping and on-site interaction.

Nesting Fermi surface with portions that are parallel. The vector that connects different parallel portions is called the nesting vector $\vec{Q}$.

Definition of the Subject

The work presented in this article belongs to the recently emerging interface of atomic physics and condensed matter theory. One of the crucial connections between these fields is the fact that ultra-cold atom ensembles in optical lattices, i. e. periodic potentials provided by standing waves of laser light, are well described by Hubbard models, the quintessential model of many-body theory. Therefore, these experiments allow for the study of many-body effects in a well-defined and tunable environment.

The subject of this article is the study of quantum phases of ultra-cold atoms in optical lattices. The objective is to propose experimental configurations, such as what lattice geometry or which types of atoms to use, for which unusual many-body effects can be found. Besides the applicability to ultra-cold atom systems, and given the generic nature of the underlying models, the resulting phases are also of interest in solid state systems.

Using techniques such as a numerical implementation of functional renormalization group equations and Luttinger liquid theory, we find the phase diagrams of various low-dimensional systems of different geometry, and discuss how the various phases could be detected.

Introduction

The technology of cooling and trapping atomic ensembles has been one of the most important developments in physics over the last decades. It has been a critical ingredient in creating Bose–Einstein condensates [4,14], improving atomic clocks [79], and studying atomic properties [43,55]. A new direction in this development was the realization of the Mott insulator transition [27] with ultra-cold atoms, which demonstrated that these systems can be used to create various types of quantum phases in a tunable and well-defined environment. The subsequent progress that has been made in controlling and manipulating ensembles of ultra-cold atoms [52,58,59,92], was followed by a number of experiments to create and study more and more sophisticated many-body effects, such as fermionic superfluids [28,42,109], one-dimensional strongly correlated Fermi and Bose systems [47,70,77], or noise correlations in interacting atomic systems [3,18,29,61] . These developments established the notion of 'engineering' many-body states in a tunable environment, i. e. manipulating ensembles of ultra-cold atoms in optical lattices.

This article further explores this development. The first step of creating novel states of matter is to determine the phase diagram of the system under consideration. For this purpose we use Luttinger liquid theory for studying one-dimensional quantum systems and two-dimensional thermal systems, and functional renormalization group equations to study two-dimensional quantum systems, which are both sophisticated methods that generate a lot of insight into the physics of these systems.

This article contains three main sections, which can be read independently of each other, organized as follows: In Sect. "One-Dimensional Lattices" we first study the phase diagram of an incommensurate Bose–Fermi mixture in one dimension, which can be understood as a Luttinger liquid of polarons (see [60,64]). We then broaden the scope of this study to include the effects of commensurate densities (see [63]). In Sect. "Phase-Locking Transition of Coupled Low-Dimensional Superfluids", we study the phases of two coupled two-dimensional superfluids, and we propose how the phase-locking transition of such systems can be used to realize the Kibble–Zurek mechanism, i. e. to create topological defects by ramping across a phase transition (see [66]). In Sect. "Bose–Fermi Mixtures in Two-Dimensional Optical Lattices", we use a numerical implementation of functional renormalization group equations to study the phase diagrams of Bose–Fermi mixtures in optical lattices in two dimensions. For both, a square and a triangular lattice, we find a rich structure of competing phases (see [50,62,65]).

One-Dimensional Lattices

The theory of one-dimensional many-body systems has been a highly active and fascinating field of physics for many decades, the centerpiece of which is the notion of the Luttinger liquid [22,24,90]. In this section, we propose several systems that display various features of Luttinger liquids, such as quasi-long range order, competing orders, and Kosterlitz–Thouless transitions due to commensurate densities, as will be explained.

Recent advances in controlling ultra-cold atoms lead to the realization of truly one-dimensional systems, and the study of many-body effects therein. Important benchmarks, such as the Tonks–Girardeau gas [47,77] and the Mott transition in one dimension [92], have been achieved by trapping bosonic atoms in tight tubes formed by an optical lattice potential. Novel transport properties of one-dimensional lattice bosons have been studied using these techniques [17]. More recently, a strongly interacting one-dimensional Fermi gas was realized using similar trapping methods [70]. Interactions between the fermion atoms were controlled by tuning a Feshbach resonance in these experiments. On the theory side, numerous proposals were given for realizing a variety of different phases in ultra-cold Fermi systems [10,19,83], Bose–

Fermi mixtures [9,60,64,88], as well as Bose–Bose mixtures [37,38].

In the first part of this section, we describe the phase diagram of an incommensurate Bose–Fermi mixture, in the second part we consider the effect of commensurate fillings.

Luttinger Liquid of Polarons in One-Dimensional Bose–Fermi Mixtures

In this section we investigate one-dimensional (1D) Bose–Fermi mixtures (BFM) using bosonization [8,32]. The resulting quantum phases can be understood by introducing polarons, i. e. atoms of one species surrounded by screening clouds of the other species. In our analysis the polarons emerge as the most well-defined quasi-particles in the interacting system while quantum phases of the system arise from a competition of various ordering instabilities of such polarons. The phase diagrams we obtain show a remarkable similarity to the Luttinger liquid phase diagrams of 1D interacting electron systems [90,102], suggesting that 1D BFM may be understood as Luttinger liquids of polarons.

To illustrate the results of this section, we show a typical phase diagram for a BFM in an optical lattice in Fig. 1, as a function of experimentally controlled parameters. We consider two types of atoms, one fermionic and one bosonic, moving in a lattice potential with the amplitude $V_{b,\parallel}$ (see [64]), and interacting via a short-ranged interaction characterized by the scattering length a_{bf} between bosons and fermions. We use these parameters, the scattering length a_{bf} and the strength of the longitudinal optical lattice for bosonic atoms ($V_{b,\parallel}$)[1], as tuning parameters in Fig. 1. For relatively weak interactions and slow bosons (i. e. large $V_{b,\parallel}$) the system is in the charge-density-wave (CDW) phase, in which the densities of fermions and bosons have a periodic modulation[2]. For very strong interactions the system is unstable to phase separation (PS) [1,7,9]. The two regimes are separated by a p-wave pairing phase of fermionic polarons (f-PP). Our analysis is carried out for the most promising system of atoms in an optical lattice. However, qualitative results should also apply to atoms in a tight 1D cigar-shaped magnetic trap [23]. A sketch of the two phases is shown in Fig. 2.

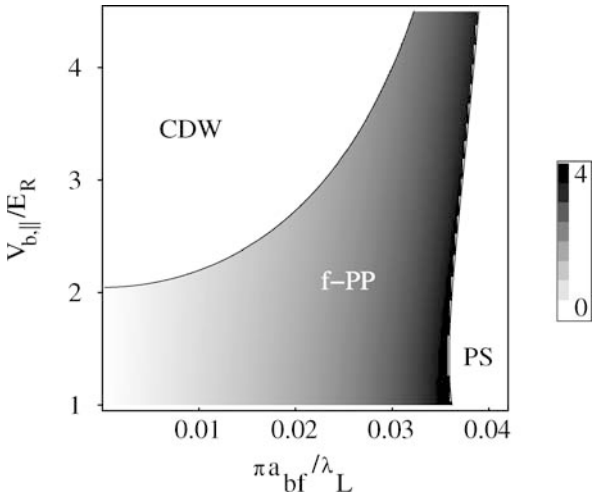

Ultracold Atomic Gases: Novel States of Matter, Figure 1
Phase diagram for a mixture of bosonic and spinless fermionic atoms in a 1D optical lattice. *Shading* in the f-PP phase describes the strength of the bosonic screening cloud (2λ, see Eq. (4)) around a pair of fermions. λ_L and E_R are respectively the lattice period and recoil energy. Other parameters used for this figure are (see text for notations, [64] for details): $\nu_b = 4$, $\nu_f = 0.5$, $V_{b,\perp} = V_{f,\perp} = 20E_R$, $V_{f,\parallel} = 2E_R$, boson-boson scattering length $a_{bb} = 0.01\lambda_L$

The essence of the bosonization procedure is to diagonalize the effective low-energy Hamiltonian, which allows for the exact calculation of all relevant correlation functions. The phase diagrams are determined by finding the order parameter which has the most divergent susceptibility [90,102]. Bosonization approach has been applied to BFM in [9]. However, that work did not consider the formation of polarons and, as a result, did not describe most of the quantum phases discussed here. The present system also has a close analogy to 1D electron–phonon systems discussed previously (see e. g. [101]). A qualitative difference of the electron–phonon system is that the sound velocity is usually much smaller than the Fermi velocity, whereas for a BFM the velocity of the phonon modes (of the bosonic condensate) can be larger than the Fermi velocity. We also note that the 1D p-wave superfluid we obtain here may be of relevance to a recent proposal for quantum computation [48].

We now give an overview over the, somewhat technical, derivation of this phase diagram, before we discuss issues concerning the experimental realization and detection of these phases, and conclude. We consider a mixture of spinless fermionic (f) and bosonic (b) atoms. For a sufficiently strong optical potential the microscopic

[1] In this paper, we use $V_{f/b,\parallel(\perp)}$ to denote the optical lattice potential experienced by fermionic/bosonic atoms in the longitudinal (perpendicular) directions. Independent tuning of the optical lattices for two species of atoms can be achieved even with a single pair of lasers by controlling the laser detuning and intensity separately.

[2] We note that in a homogeneous 1D system only quasi-long-range order for CDW phase is possible. However, the inhomogeneous trap boundary in a realistic experiment can pin the CDW phase to generate a true density modulation.

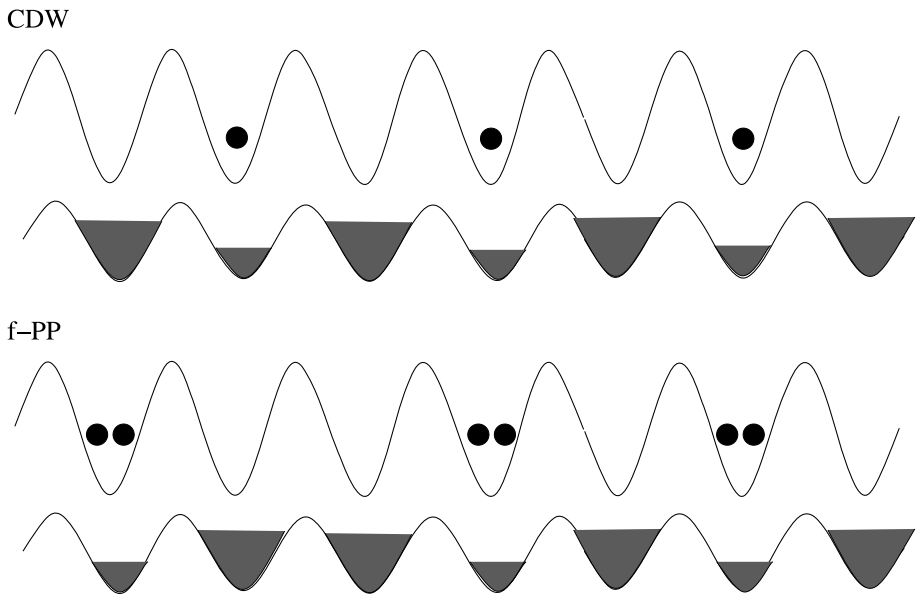

Ultracold Atomic Gases: Novel States of Matter, Figure 2
Illustration of the two phases that occur in a BFM with spinless fermions, CDW and *f*-PP. In the CDW phase the system develops a $2k_f$ density modulation in both the fermionic and the bosonic liquid. In the *f*-PP phase, the fermions form polarons, indicated by the reduced bosonic density in their vicinity, that is, their polarization cloud. This polarization leads to an effective attractive interaction, which causes these fermionic polarons to pair up and form a superfluid state

Hamiltonian is given by a single band Hubbard model

$$H = -\sum_{\langle ij\rangle}\left(t_b b_i^\dagger b_j + t_f f_i^\dagger f_j\right) - \sum_i \left(\mu_f n_{f,i} + \mu_b n_{b,i}\right) + \frac{U_b}{2}\sum_i n_{b,i}(n_{b,i}-1) + U_{bf}\sum_i n_{b,i} n_{f,i}\,, \quad (1)$$

where $n_{b/f,i}$ are the boson/fermion density operators with $\mu_{b/f}$ being their chemical potentials. The tunneling amplitudes $t_{f/b}$, and the particle interactions U_b and U_{bf} can be expressed explicitly in terms of the *s*-wave scattering lengths, the laser beam intensities and atomic masses [39]. For simplicity we assume that the filling fraction of fermions $\nu_f \equiv \langle n_{f,i}\rangle$ is not commensurate with the lattice or with the filling fraction of bosons ν_b. Therefore, we can neglect lattice-assisted backward/Umklapp scattering. The Fermi momentum and velocity are given by $k_f = \pi\nu_f$ and $v_f = 2t_f \sin(k_f)$, respectively.

In Haldane's bosonization approach [8,32] 1D fermion and boson operators can be represented by $f(x) = [\nu_f + \Pi_f]^{1/2}\sum_{m=-\infty}^{\infty} e^{(2m+1)i\Theta_f} e^{i\Phi_f}$ and $b(x) = [\nu_b + \Pi_b]^{1/2}\sum_{m=-\infty}^{\infty} e^{2mi\Theta_b} e^{i\Phi_b}$, where x is a continuous coordinate that replaces the site index i. The operators $\Pi_{f/b}(x)$ and $\Phi_{f/b}(x)$ are the bosonized density and phase fluctuation operators. The $\Theta_{f/b}(x)$ fields are given by $\Theta_{f/b} \equiv \pi\nu_{f/b}x + \pi\int^x dy\,\Pi_{f/b}(y)$. The low-energy effective Hamiltonian thus can be written as:

$$H_{eff} = \sum_{\alpha=b,f}\frac{v_\alpha}{2}\int dx\left[\frac{K_\alpha}{\pi}(\partial_x\Phi_\alpha)^2 + \frac{\pi}{K_\alpha}\Pi_\alpha^2\right] + U_{bf}\int dx\,\Pi_b\Pi_f + \frac{2G}{2\pi}\int dx\left[\pi^2\Pi_f^2 - (\partial_x\Phi_f)^2\right]. \quad (2)$$

where v_b and K_b are the phonon velocity and Luttinger exponent of the bosons and $K_f = 1$ for noninteracting fermion atoms.

To obtain the last term of H_{eff} we have integrated out the high energy ($2k_f$) phonons within the instaneous approximation (i. e. assuming $v_b \gg v_f$). $G \equiv g_{2k_f}^2/\omega_{2k_f}$, where ω_k is the (Bogoliubov) phonon energy dispersion [98] and $g_k = U_{bf}\sqrt{\nu_b\,\varepsilon_{b,k}/2\pi\omega_k}$ is the fermion–phonon (FP) coupling vertex with $\varepsilon_{b,k}$ being the noninteracting boson band energy. In the long wavelength limit we have a conventional FP coupling $g_k = g|k|^{1/2}$ with $g \equiv U_{bf}\sqrt{K_b}/2\pi$. The effective Hamiltonian, Eq. (2), is quadratic and can be diagonalized [16]. The resulting two eigenmode velocities are given by [9]

$$v_{a,A}^2 = \frac{1}{2}(v_b^2 + \tilde{v}_f^2) \pm \frac{1}{2}\sqrt{(v_b^2 - \tilde{v}_f^2)^2 + 16\tilde{g}^2 v_b \tilde{v}_f}\,, \quad (3)$$

where $\tilde{v}_f \equiv (v_f^2 - 4G^2)^{1/2}$ and $\tilde{g} \equiv g\,e^\theta$ with $e^\theta = ((v_f - 2G)/(v_f + 2G))^{1/4}$. When the FP coupling g becomes suf-

ficiently strong the eigenmode velocity v_A becomes imaginary, indicating an instability of the system. This instability corresponds to phase separation (global collapse) for positive (negative) $U_{\rm bf}$ [9].

To understand the nature of the many-body state of BFM outside of the instability region we analyze the long distance behavior of the correlation functions. For the bare bosonic and fermionic particles we find $\langle b(x)b^\dagger(0)\rangle \sim |x|^{-\frac{1}{2}K_\epsilon^{-1}}$ and $\langle f(x)f^\dagger(0)\rangle \sim \cos(k_{\rm f}x)|x|^{-\frac{1}{2}(K_\beta+K_\gamma^{-1})}$ [3]. To describe particles dressed by the other species we introduce the composite operators

$$\tilde{f}(x) \equiv {\rm e}^{-i\lambda\Phi_{\rm b}(x)}f(x)\,, \quad \tilde{b}(x) \equiv {\rm e}^{-i\eta\Phi_{\rm f}(x)}b(x)\,, \qquad (4)$$

with λ and η being some real numbers. The correlation functions of these operators are given by $\langle\tilde{f}(x)\tilde{f}^\dagger(0)\rangle \sim \cos(k_{\rm f}x)|x|^{-\frac{1}{2}(K_\beta+\lambda^2K_\epsilon^{-1}+K_\gamma^{-1}-2\lambda K_{\gamma\epsilon}^{-1})}$ and $\langle\tilde{b}(x)\tilde{b}^\dagger(0)\rangle \sim |x|^{-\frac{1}{2}(K_\epsilon^{-1}+\eta^2K_\gamma^{-1}-2\eta K_{\gamma\epsilon}^{-1})}$ [3]. We observe that the exponents of the correlation functions are maximized for $\lambda_{\rm c} = K_\epsilon/K_{\gamma\epsilon}$ and $\eta_{\rm c} = K_\gamma/K_{\gamma\epsilon}$. From now on we will use Eq. (4) with $\lambda_{\rm c}$ and $\eta_{\rm c}$ to construct polaronic particles. In the limit of weak interactions we have $\lambda_{\rm c} \to U_{\rm bf}/U_{\rm b}$ and $\eta_{\rm c} \to 2U_{\rm bf}/\pi v_{\rm b}$. This result can be understood by a simple density counting argument that a fermionic polaron (f-polaron) locally suppresses (enhances) a bosonic cloud by $\lambda_{\rm c}$ particles, whereas a bosonic polaron (b-polaron) depletes (enhances) the fermionic system by $\eta_{\rm c}$ atoms for positive (negative) g.

The polaronic operators defined in Eq. (4) can also be introduced via the canonical polaron transformation (CPT), which is often used in polaron theory [2,57]. The CPT operator is given by $U = {\rm e}^{-i\lambda\sum_{k\neq0}(F_k\beta_k\rho_k^\dagger+{\rm h.c.})}$, where β_k is the phonon annihilation operator, ρ_k is the fermion density operator, F_k is some function of wavevector k, and λ specifies the strength of the phonon dressing. When applied to a fermion operator, the CPT transforms it to a polaron operator, $U^{-1}f(x)U = f(x)\exp[-i\lambda\sum_{k\neq0}(F_k\beta_k{\rm e}^{-ik\cdot x}+{\rm h.c.})]$ [2,57], which is the same as Eq. (4), provided that one takes $F_k = \sqrt{2\pi/(K_{\rm b}|k|L)}\,{\rm sgn}(k)$. (Note that in 1D fermionic systems density operators correspond to Luttinger bosons.) We note, however, unlike in ordinary polaron theory, where further approximations after the CPT have to be made [2,57], in the 1D BFM system we consider here, the full low energy quantum fluctuations have been included via bosonization method and exact diagonalization of the resulting Hamiltonian Eq. (2). This allows for an essentially exact determination of the polarization parameter λ.

Now we study the many-body ground state phase diagram of a 1D BFM, which is characterized by specifying the order parameters that have the slowest long distance decay of the correlation functions [90,102]. Two types of ordering were found to occur: $2k_{\rm f}$-ordering due to a Peierls-type instability and f-polaron pairing due to their effective attractive interactions induced by the screening clouds, see Fig. 2. For the $2k_{\rm f}$ CDW order parameter, $O_{\rm CDW} = f_{\rm L}^\dagger f_{\rm R}$, we find $\alpha_{\rm CDW} = 2 - 2K_\beta$, and for the f-polaron pairing field, $O_{f-{\rm PP}} = \tilde{f}_{\rm L}\tilde{f}_{\rm R}$, we obtain $\alpha_{f-{\rm PP}} = 2 - 2[\lambda_{\rm c}^2K_\epsilon^{-1} + K_\gamma^{-1} - 2\lambda_{\rm c}K_{\gamma\epsilon}^{-1}]$. We did not include polaron dressing in $O_{\rm CDW}$, since this operator has no net fermionic charge and the exponent of $O_{\rm CDW}$ does not change if we replace f by $\tilde{f}$. Scaling exponents shown in Fig. 3a demonstrate that divergencies of the CDW and f-PP susceptibilities (corresponding to positive α) are mutually exclusive and cover the entire phase diagram outside the PS regime. In the same figure, we also show the scaling exponents calculated for bare fermion pairing ($O_{BFP} = f_{\rm L}f_{\rm R}$), bare boson condensate ($O_{BB} = b$), and b-polaron condensate ($O_{{\rm b}-P} = \tilde{b}$). It is easy to see that the polaronic order parameters always have larger exponents than their counterparts constructed with bare atoms, showing the stability of polaronic quasi-particle in a 1D BFM system. Moreover, the necessity to consider f-polaron pairing instead of bare fermion pairing is further supported by considering the stability of superfluidity: we introduce a single weak impurity potential in the 1D BFM and determine its relevance by a renormalization group (RG) calculation [44]. We find that the impurity potential is relevant within the CDW phase and irrelevant outside of it. This indicates that there should be a superfluid phase outside of the insulating CDW phase, which supports the existence of f-polaron pairing instead of bare fermion pairing according to Fig. 3a.

In Fig. 3b we show a global phase diagram of a BFM considering the FP coupling (g) and effective fermion–fermion interaction (G) as independent variables. One can see that the polaronic effects and the associated pairing phase are important when FP coupling (g) is large, while the CDW phase dominates when the effective fermion interaction (G) is increased. This phase diagram is very similar to what one finds for spinless electrons in Luttinger liquid theory [90,102], where CDW and pairing phase compete with each other in the whole phase diagram. Therefore one can introduce a Luttinger liquid of

[3] Here we defined $K_\beta \equiv {\rm e}^{2\theta}\tilde{v}_{\rm f}(\cos^2\psi/v_A + \sin^2\psi/v_{\rm a})$, $K_\delta \equiv K_{\rm b}v_{\rm b}(\sin^2\psi/v_A + \cos^2\psi/v_{\rm a})$, $K_\gamma^{-1} \equiv {\rm e}^{-2\theta}/\tilde{v}_{\rm f}(v_A\cos^2\psi + v_{\rm a}\sin^2\psi)$, $K_\epsilon^{-1} \equiv K_{\rm b}^{-1}/v_{\rm b}(v_A\sin^2\psi + v_{\rm a}\sin^2\psi)$, $K_{\beta\delta} = {\rm e}^\theta\sqrt{K_{\rm b}\tilde{v}_{\rm f}v_{\rm b}}\sin(2\psi)/2(1/v_{\rm a} - 1/v_A)$, and $K_{\gamma\epsilon}^{-1} = {\rm e}^{-\theta}/\sqrt{K_{\rm b}\tilde{v}_{\rm f}v_{\rm b}}\sin(2\psi)/2(v_{\rm a} - v_A)$. ψ is given by $\tan 2\psi = 4\tilde{g}(v_{\rm b}\tilde{v}_{\rm f})^{1/2}/(v_{\rm b}^2 - \tilde{v}_{\rm f}^2)$. These expressions are obtained from diagonalizing $H_{\rm eff}$. For details see [64].

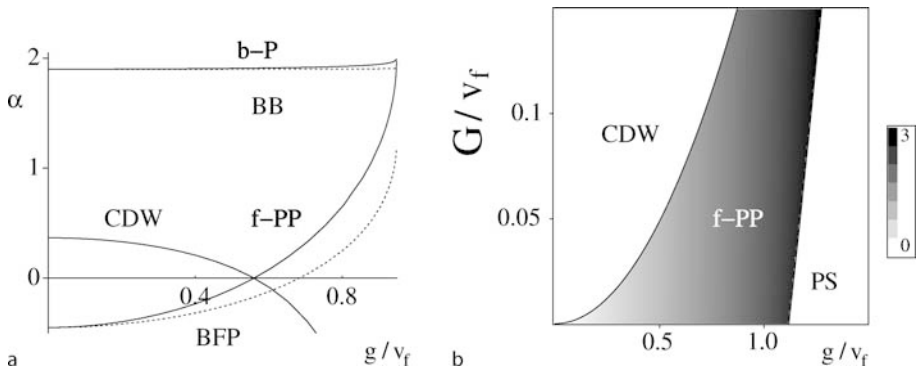

Ultracold Atomic Gases: Novel States of Matter, Figure 3

Ground state of a BFM with spinless fermions. a Scaling exponents of different order parameters (see the text). Parameters are chosen to be $v_b/v_f = 3$, $K_b = 5$ and $G/v_f = 0.1$. b Global phase diagram for $v_b/v_f = 5$ and $K_b = 10$. *Shading* density indicates the strength of the screening clouds of a polaron pair, $2\lambda_c$

polarons to describe BFM in 1D systems. The phase diagram in terms of experimentally controlled parameters was shown in Fig. 1. When considering finite temperature effects in a realistic experiment, we note that the correlation function is cut-off by thermal correlation lengths, which are approximately given by $\xi \sim v_f/k_B T$. Therefore the zero temperature ground states should appear when $\xi > L$ with L being the system size. This corresponds to a temperature regime of 1% of the Fermi temperature for systems of approximately 100 sites in the longitudinal direction.

Several approaches can be used to detect the quantum phases discussed above. First, in the CDW phase the fermion density modulation will induce a $2k_f$ density wave in the boson field in addition to the zero momentum condensation so that the CDW phase can be observed as interference peaks at momentum $k = 2k_f$ in a standard time-of-flight (TOF) measurement for bosons[4]. Secondly, the polaron pairing phase can be observed by measuring the noise correlation of fermions in a TOF experiment as proposed in [3]. Thirdly, a laser stirring experiment [72,82] can be used to probe the phase transition between the insulating (pinned by trap potential) CDW and the superfluid f-PP phase: one can use a laser beam focused at the center of the cloud and stir such local potential to measure the response of the BFM. If the system is in the pairing phase, the laser beam can be moved through the system without dissipation if only its velocity is slower than some critical value [72,82]. At the f-PP/CDW phase boundary this critical velocity goes to zero, reflecting a transition to the insulating (CDW) state. This scenario follows from the above described RG analysis of a single impurity potential [44]. Finally, a way to probe the PS boundary could be to measure the dipolar collective oscillations of the system, generated by a sudden displacement of the harmonic trap potential with respect to the lattice potential [21,56,100]. When the system is near the PS boundary, fermion–boson interaction will strongly reduce the frequency of the dipolar mode.

In summary, we used bosonization to investigate the quantum phases of 1D mixtures of bosonic and fermionic atoms involving spinless fermions. The phase diagram that we found can be understood in terms of a Luttinger liquid of polarons. We also described several experimental techniques for probing these quantum phases.

Commensurate Mixtures of Ultra-Cold Atoms in One Dimension

In this section we explore the behavior of ultra-cold atomic mixtures, confined to one-dimensional (1D) motion in an optical lattice, that exhibit different types of commensurability, by which we mean that the atomic densities and/or the inverse lattice spacing have an integer ratio. Commensurable fillings arise naturally in many ultra-cold atom systems, because the external trap potential approximately corresponds to a sweep of the chemical potential through the phase diagram, and therefore passes through points of commensurability. At these points the system can develop an energy gap, which fixes the density commensurability over a spatially extended volume. This was demonstrated in the celebrated Mott insulator experiment by Greiner et al. [27], where Mott phases with integer filling occurred in shell-shaped regions in the atom trap. These gaped

[4]We note that in a homogeneous 1D system only quasi-long-range order for CDW phase is possible. However, the inhomogeneous trap boundary in a realistic experiment can pin the CDW phase to generate a true density modulation.

phases gave rise to the well-known signature in the time-of-flight images[5], and triggered the endeavor of 'engineering' many-body states in optical lattices. Further examples include the recently created density-imbalanced fermion mixtures [78,110] in which the development of a balanced, i. e. commensurate, mixture at the center of the trap is observed.

In 1D, this phenomenon is of particular importance, because it is the only effect that can lead to the opening of a gap, for a system with short-range interactions. In contrast to higher-dimensional systems, where, for instance, pairing can lead to a state with an energy gap, in 1D only discrete symmetries can be broken, due to the importance of fluctuations. Orders that correspond to a continuous symmetry can, at most, develop quasi long range order (QLRO), which refers to a state in which an order parameter $O(x)$ has a correlation function with algebraic scaling, $\langle O(x)O(0)\rangle \sim |x|^{-(2-\alpha)}$, with a positive scaling exponent α.

Due to its importance in solid state physics, the most thoroughly studied commensurate 1D system is the SU(2) symmetric system of spin-1/2 fermions. This system develops a spin gap for attractive interaction and remains gapless for repulsive interaction, as can be seen from a second order RG calculation. However, the assumed symmetry between the two internal spin states, which is natural in solid state systems, does not generically occur in Fermi–Fermi mixtures (FFMs) of ultra-cold atoms, where the 'spin' states are in fact different hyperfine states of the atoms. An analysis of the generic system is therefore highly called for. Furthermore, we will extend this analysis to both Bose–Fermi (BFMs) and Bose–Bose mixtures (BBMs), as well as to the dual commensurability, in which the charge field, and not the spin field, exhibits commensurate filling, as will be explained below.

The main results of this section are the phase diagrams shown in Figs. 4 and 5. We find that both attractive and repulsive interactions can open an energy gap. For FFMs the entire phase diagram is gaped, except for the repulsive SU(2) symmetric regime (cf. [10]), for BFMs or BBMs the bosonic liquid(s) need(s) to be close to the hard-core limit, otherwise the system remains gapless. Furthermore, we find a rich structure of quasi-phases, including charge and spin density wave order (CDW, SDW), singlet and triplet pairing (SS, TS), polaron pairing [60,64], and a supersolid phase, which is the first example of a supersolid phase in 1D. These results are derived within a Luttinger liquid (LL) description, which treats bosonic and fermionic liquids on equal footing.

We will now classify the types of commensurability that can occur in a system with short-ranged density–density interaction. We consider Haldane's representation [8,32] of the densities for the two species:

$$n_{1/2} = [\nu_{1/2} + \Pi_{1/2}] \sum_m \mathrm{e}^{2mi\Theta_{1/2}} \,. \tag{5}$$

ν_1 and ν_2 are the densities of the two liquids, $\Pi_{1/2}(x)$ are the low-k parts (i. e. $k \ll 1/\nu$) of the density fluctuations; the fields $\Theta_{1/2}(x)$ are given by $\Theta_{1/2}(x) = \pi\nu_{1/2}x + \theta_{1/2}(x)$, with $\theta_{1/2}(x) = \pi\int^x \mathrm{d}y\Pi_{1/2}(y)$. These expressions hold for both bosons and fermions. If we use this representation in a density–density interaction term $U_{12}\int \mathrm{d}x n_1(x)n_2(x)$, we generate to lowest order a term of the shape $U_{12}\int \mathrm{d}x\Pi_1(x)\Pi_2(x)$, but in addition an infinite number of nonlinear terms, corresponding to all harmonics in the representation. However, only the terms for which the linear terms $(2\pi m_{1/2}\nu_{1/2}x)$ cancel, can drive a phase transition. For a continuous system this happens for $m_1\nu_1 - m_2\nu_2 = 0$, whereas for a system on a lattice we have the condition $m_1\nu_1 - m_2\nu_2 = m_3$, where m_1, m_2 and m_3 are integer numbers. In general, higher integer numbers correspond to terms that are less relevant, because the scaling dimension of the non-linear term scales quadratically with these integers. We are therefore lead to consider small integer ratios between the fillings and/or the lattice if present. In [64], we considered two cases of commensurabilities: A Mott insulator transition coupled to an incommensurate liquid, and a fermionic liquid at half-filling coupled to an incommensurate bosonic liquid. In both cases the commensurability occurs between one species and the lattice, but does not involve the second species. Here, we consider the two most relevant, i. e. lowest order, cases which exhibit a commensurability that involves both species. The first case is the case of equal filling $\nu_1 = \nu_2$, the second is the case of the total density being unity, i. e. $\nu_1 + \nu_2 = 1$, where the densities ν_1 and ν_2 themselves are incommensurate. The first case can drive the system to a spin-gaped state, the second to a charge gaped state. We will determine in which parameter regime these transitions occur, and what type of QRLO the system exhibits in the vicinity of the transition. These two cases can be mapped onto each other via a dual mapping, which enables us to study only one case and then infer the results for the second by using this mapping. We will write out our discussion for the case of equal filling and merely state the corresponding results for complementary filling.

[5] Due to the confining trap the transition is, while visible, 'blurred' into a gradual cross-over, due to finite size effects, and, more importantly, due to the coexistence of several phases in the trap. This can also be expected for the phase transitions predicted in this paper.

The action of a two-species mixture with equal filling in bosonized form is given by:

$$S = S_{0,1} + S_{0,2} + S_{12} + S_{\text{int}} \,. \tag{6}$$

The terms $S_{0,j}$, with $j = 1, 2$, are given by

$$S_{0,j} = \frac{1}{2\pi K_j} \int \mathrm{d}^2 r \Big(\frac{1}{v_j} (\partial_\tau \theta_j)^2 + v_j (\partial_x \theta_j)^2 \Big) \,. \tag{7}$$

Each of the two types of atoms, regardless of being bosonic or fermionic, are characterized by a Luttinger parameter $K_{1/2}$ and a velocity $v_{1/2}$. Here we integrate over $\mathbf{r} = (v_0 \tau, x)$, where we defined the energy scale $v_0 = (v_1 + v_2)/2$. The term S_{12} describes the acoustic coupling between the two species, and is bilinear:

$$S_{12} = \frac{U_{12}}{\pi^2} \int \mathrm{d}^2 r \partial_x \theta_1 \partial_x \theta_2 + \frac{V_{12}}{\pi^2} \int \mathrm{d}^2 r \partial_\tau \theta_1 \partial_\tau \theta_2 \,. \tag{8}$$

The second term is created during the RG flow; its prefactor therefore has the initial value $V_{12}(0) = 0$. We define $S_0 = S_{0,1} + S_{0,2} + S_{12}$, which is the diagonalizable part of the action. S_{int} corresponds to the non-linear coupling between the two liquids, which we study within an RG approach:

$$S_{\text{int}} = \frac{2 g_{12}}{(2\pi\alpha)^2} \int \mathrm{d}^2 r \cos(2\theta_1 - 2\theta_2) \,. \tag{9}$$

This bosonized description applies to a BBM, a BFM, and a FFM. Depending on which of these mixtures we want to describe we either construct bosonic or fermionic operators according to Haldane's construction [8,32]:

$$f/b = [v_0 + \Pi]^{1/2} \sum_{m \text{ odd/even}} \mathrm{e}^{mi\Theta} \mathrm{e}^{i\Phi} \,. \tag{10}$$

v_0 is the zero-mode of the density, $\Phi(x)$ is the phase field, which is the conjugate field of the density fluctuations $\Pi(x)$. The action for a mixture with complementary filling, $\nu_1 + \nu_2 = 1$, is of the form $S_0 + S'_{\text{int}}$, where the interaction S'_{int} is given by:

$$S'_{\text{int}} = \frac{2 g_{12}}{(2\pi\alpha)^2} \int \mathrm{d}^2 r \cos(2\theta_1 + 2\theta_2) \,. \tag{11}$$

To map the action in Eq. (6) onto this system we use the mapping: $\theta_2 \to -\theta_2$, $\phi_2 \to -\phi_2$, and $g_{12} \to -g_{12}$, which evidently maps a mixture with complementary filling and attractive (repulsive) interaction and onto a mixture with equal filling with repulsive (attractive) interaction.

To study the action given in Eq. (6), we perform an RG calculation along the lines of the treatment of the sine-Gordon model in [24,51]. In our model, a crucial modification arises: The linear combination $\theta_1 - \theta_2$, that appears in the non-linear term, is not proportional to an eigenmode of S_0, and therefore the RG flow does not affect only one separate sector of the system, as in an SU(2)-symmetric system. The RG scheme that we use here proceeds as follows: First, we diagonalize S_0 through the transformation (see [63]) $\theta_1 = B_1 \tilde{\theta}_1 + B_2 \tilde{\theta}_2$, and $\theta_2 = D_1 \tilde{\theta}_1 + D_2 \tilde{\theta}_2$, where $B_{1/2}$ and $D_{1/2}$ are some coefficients, and $\tilde{\theta}_{1/2}$ are the eigenmode fields with velocities $\tilde{v}_{1/2}$. Now we introduce an energy cut-off Λ on $\tilde{\theta}_{1/2}$ according to $\omega^2/\tilde{v}_{1/2} + \tilde{v}_{1/2} k^2 < \Lambda^2$. We shift this cut-off by an amount $\mathrm{d}\Lambda$, and correct for this shift up to second order in g_{12}. At first order, only g_{12} is affected, its flow equation is given by:

$$\frac{\mathrm{d} g_{12}}{\mathrm{d} l} = \Big(2 - K_1 - K_2 - \frac{2}{\pi} \frac{U_{12} + V_{12} v_1 v_2}{v_1 + v_2} \Big) g_{12} \,, \tag{12}$$

with $\mathrm{d}l = \mathrm{d}\Lambda/\Lambda$. At second order several terms are created that are quadratic in the original fields θ_1 and θ_2. We undo the diagonalization, and absorb these terms into the parameters of the action, which concludes the RG step. By iterating this procedure we obtain these flow equations at second order in g_{12}:

$$\frac{\mathrm{d} K_{1/2}}{\mathrm{d} l} = -\frac{g_{12}^2}{16\pi^2} \Big(2 + \Big(\frac{v_2}{v_1} + \frac{v_1}{v_2} \Big) \Big) \,, \tag{13}$$

$$\frac{\mathrm{d} v_1}{\mathrm{d} l} = v_1 \frac{g_{12}^2}{16\pi^2} \Big(\frac{v_2}{v_1} - \frac{v_1}{v_2} \Big) \,, \tag{14}$$

$$\frac{\mathrm{d} v_2}{\mathrm{d} l} = v_2 \frac{g_{12}^2}{16\pi^2} \Big(\frac{v_1}{v_2} - \frac{v_2}{v_1} \Big) \,, \tag{15}$$

$$\frac{\mathrm{d} U_{12}}{\mathrm{d} l} = -\frac{g_{12}^2}{8\pi} (v_1 + v_2) \,, \tag{16}$$

$$\frac{\mathrm{d} V_{12}}{\mathrm{d} l} = -\frac{g_{12}^2}{8\pi} (1/v_1 + 1/v_2) \,. \tag{17}$$

The system of differential equations, Eqs. (12) to (17), can show two types of qualitative behavior: The coefficient g_{12} of the non-linear term (9) can either flow to zero, i. e. S_{int} is irrelevant, or it diverges, leading to the formation of an energy gap. In the first case, the system flows to a fixed point that is described by a renormalized diagonalizable action of the type S_0, from which the quasi-phases can be determined.

When S_{int} is relevant, we introduce the fields [102] $\theta_{\rho/\sigma} = \frac{1}{\sqrt{2}} (\theta_1 \pm \theta_2)$, which define the charge and the spin sector of the system. In this regime, these sectors decouple. Each of the two sectors is characterized by a Luttinger parameter and a velocity, $K_{\rho/\sigma}$ and $v_{\rho/\sigma}$, which are related to

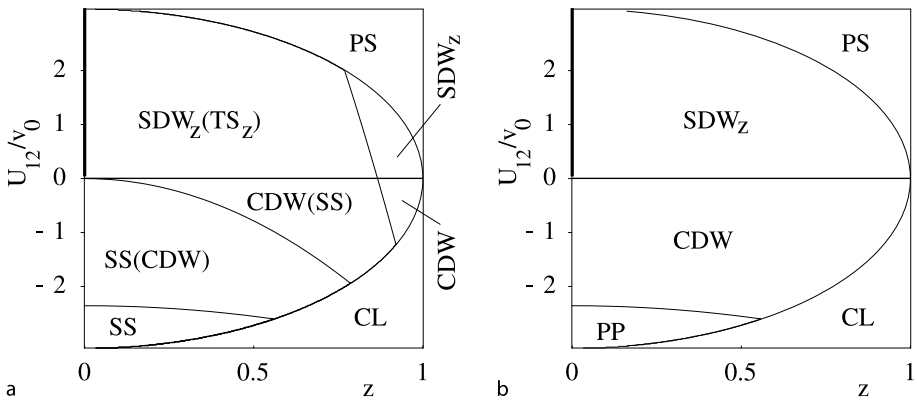

Ultracold Atomic Gases: Novel States of Matter, Figure 4

a Phase diagram of a commensurate FFM or a BBM of hardcore bosons (with the replacement $TS_z \rightarrow SS$), b phase diagram of a BFM with hardcore bosons, in terms of the interaction U_{12} and the parameter $z = |v_1 - v_2|/(v_1 + v_2)$. For both attractive and repulsive interactions a spin gap opens, except for $z = 0$ and positive interaction. In the attractive regime, a FFM or a BBM shows either singlet pairing or CDW order, or a coexistence of these phases, a BFM shows either CDW order or polaron pairing. For repulsive interaction all mixtures show SDW ordering, with FFMs and BBMs showing subdominant triplet or singlet pairing, respectively, for a large range of z. In the gapless regime, a FFM shows degenerate SDW and CDW order, a BFM shows CDW order for the fermions and SF for the bosons, and a BBM shows SF with subdominant CDW, i. e. supersolid behavior. For very large positive values of U_{12} the system undergoes phase separation (PS); for very large negative values it collapses (CL)

the original parameters in S_0 in a straightforward way. Using the numerical solution of the flow equations, we find that $K_\sigma \rightarrow 0$, as can be expected for an ordering of the nature of a spin gap, leaving K_ρ the only parameter characterizing the QLRO in this phase.

In order to determine the QLRO in the system we will determine the scaling exponents of various order parameters. The order parameter with the largest positive scaling exponent shows the dominant order, whereas other orders with positive exponent are subdominant.

We will now apply this procedure to the different types of mixtures. For a FFM we find that the system always develops a gap, with the exception of the repulsive SU(2) symmetric regime (cf. [10]). To determine the QLRO we introduce the following operators [22,102]:

$$O_{\mathrm{SS}} = \sum_{\sigma,\sigma'} \tilde{\sigma} f_{R,\sigma} \delta_{\sigma,\sigma'} f_{L,3-\sigma'} ,$$

$$O^a_{TS} = \sum_{\sigma,\sigma'} \tilde{\sigma} f_{R,\sigma} \sigma^a_{\sigma,\sigma'} f_{L,3-\sigma'} ,$$

$$O_{\mathrm{CDW}} = \sum_{\sigma,\sigma'} f^\dagger_{R,\sigma} \delta_{\sigma,\sigma'} f_{L,\sigma'} , \text{ and}$$

$$O^a_{SDW} = \sum_{\sigma,\sigma'} \tilde{\sigma} f^\dagger_{R,\sigma} \sigma^a_{\sigma,\sigma'} f_{L,\sigma'} ,$$

with $\sigma, \sigma' = 1, 2$, $\tilde{\sigma} = 3 - 2\sigma$, and $a = x, y, z$. In the gapless SU(2) symmetric regime, both CDW and SDW show QLRO, with both scaling exponents of the form $\alpha_{\mathrm{SDW/CDW}} = 1 - K_\rho$ [102], which shows that these orders are algebraically degenerate. Within the gaped regime the scaling exponents of these operators are given by $\alpha_{SS,TS_z} = 2 - K_\rho^{-1}$ and $\alpha_{\mathrm{CDW,SDW}_z} = 2 - K_\rho$. As discussed in [22], the sign of g_{12} determines whether CDW or SDW_z, and SS or TS_z appears. In Fig. 4a, we show the phase diagram based on these results. In addition to these phases we indicate the appearance of the Wentzel–Bardeen instability, shown as phase separation for repulsive interaction and collapse for attractive interaction.

We will now use the dual mapping to obtain the phase diagram of a FFM with complementary filling from Fig. 4a. Under this mapping, the attractive and repulsive regimes are exchanged with the following replacements: $\mathrm{CDW} \rightarrow \mathrm{SDW}_z$, $\mathrm{SDW}_z \rightarrow \mathrm{CDW}$, $\mathrm{SS}, \mathrm{TS}_z \rightarrow \mathrm{SDW}$, and $\mathrm{SDW} \rightarrow \mathrm{SS}$. Note that the gapless regime is now on the attractive side, with degenerate CDW and SS pairing.

For BBMs we proceed in the same way as for FFMs. We introduce the following set of order parameters: $O_{\mathrm{CDW}} = b_1^\dagger b_1 + b_2^\dagger b_2$, $O_{\mathrm{SS}} = b_1 b_2$, $O_{\mathrm{SDW}_z} = b_1^\dagger b_1 - b_2^\dagger b_2$, $O_{\mathrm{SDW}_x} = b_1^\dagger b_2 + b_2^\dagger b_1$, $O_{\mathrm{SDW}_y} = -i(b_1^\dagger b_2 - b_2^\dagger b_1)$, and in addition the superfluid (SF) order parameters b_1 and b_2. In Fig. 4a we show the phase diagram of a mixture of a BBM of hardcore bosons, which is almost identical to the one of a FFM. The phase diagram of the mixture with complementary filling, as obtained from the dual mapping, is

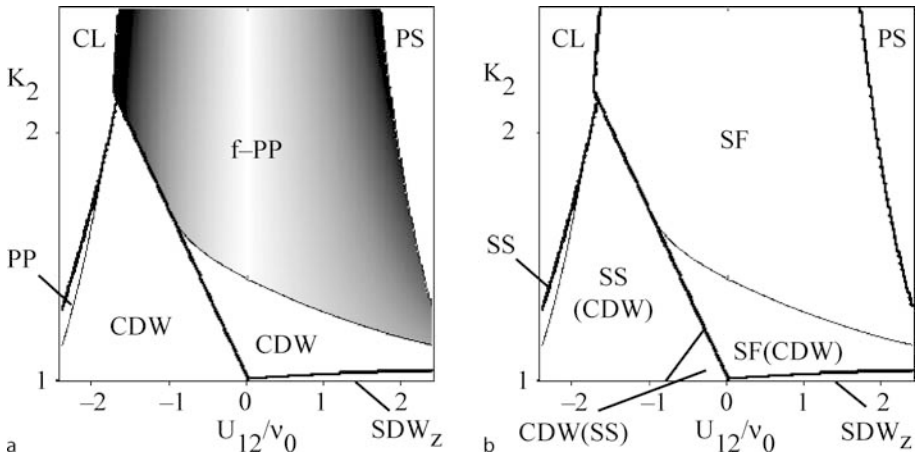

Ultracold Atomic Gases: Novel States of Matter, Figure 5

a Phase diagram of a BFM, b phase diagram of a BBM with the first species being in the hardcore limit, in terms of U_{12}, and the Luttinger parameter of the second species (K_2), at the fixed velocity ratio $|v_1 - v_2|/(v_1 + v_2) = 0.5$. For large repulsive interaction the system undergoes phase separation (PS), for large attractive interaction the system collapses (CL). In the regime below the *thick line* the system opens a gap, i. e. if species 2 is close to the hardcore limit. However, for larger values of K_2, the gapless phase is restored. Close to the transition, the properties of the fermions, respectively hardcore bosons, are still affected by the RG flow, leading to CDW order for the fermions and to supersolid behavior for the bosons

also of the same form as its fermionic equivalent, with the exception of the gapless regime, in which BBMs show supersolid behavior (coexistence of SF and CDW order), and with the replacement $TS_z \to SS$.

In Fig. 5b, we show the phase diagram of a mixture of hardcore bosons (species 1) and bosons in the intermediate to hardcore regime (species 2). If species 2 is sufficiently far away from the hardcore limit, the system remains gapless. However, in the vicinity of the transition the scaling exponents of the liquids are affected by the RG flow. As indicated, the effective scaling exponent of the hardcore bosons is renormalized to a value that is smaller than 1, and therefore we find both SF and CDW order, i. e. supersolid behavior. The phase diagram of the dual mixture is of the following form: the attractive and the repulsive regime are exchanged, and in the gaped phase we again have the mapping: $CDW \to SDW_z$, $SDW_z \to CDW$, $SS \to SDW$, and $SDW \to SS$. The gapless regime is unaffected.

For a BFM we find that the order parameters O_{CDW}, O_{SDW_z}, $O_{f-PP} = f_R f_L e^{-2i\lambda\Phi_b}$ [60,64], and b can develop QLRO in the gapless regime. In the gaped regime, the order parameters $O_{PP} \equiv f_R b f_L b$ and $O_{PP'} \equiv f_R b^\dagger f_L b^\dagger$, in addition to O_{CDW}, show QLRO. ($O_{PP/PP'}$ are special cases of the polaron pairing operators discussed in [60,64].) In Fig. 4b we show the phase diagram of a BFM with hardcore bosons, and in Fig. 5a, we vary the Luttinger parameter of the bosons. In both the gapless phase and the gaped phase, we find that CDW and f-PP or PP, respectively, are mutually exclusive and cover the entire phase diagram, cf. [60,64]. The dual mapping again maps attractive and repulsive regimes onto each other. Within the gaped phase we find the mapping $CDW \to SDW_z$, $SDW_z \to CDW$, and $PP \to PP'$, the gapless regime is unaffected.

Before we conclude, we discuss how these predictions could be measured experimentally. CDW order will create additional peaks in TOF images, corresponding to a wavevector $Q = 2k_f$. As demonstrated and pointed out in [3,18,29,61], the noise in TOF images allows to identify the different regimes of both gaped and gapless phases. As discussed in [60,64], a laser stirring experiment could determine the onset of CDW order for fermions, or the supersolid regime for bosons. RF spectroscopy [12] can be used to determine the presence and the size of an energy gap.

In conclusion, we have studied mixtures of ultra-cold atoms in 1D with commensurate filling. We used a Luttinger liquid description which enables us to study FFMs, BFMs, and BBMs in a single approach. We find that FFMs are generically gaped for both attractive and repulsive interactions, whereas for BFMs and BBMs the bosons need to be close to the hardcore limit. We find a rich structure of quasi-phases in the vicinity of these transitions, in particular a supersolid phase for BBMs, that occurs close to the hardcore limit. Experimental methods to detect the predictions were also discussed.

Phase-Locking Transition of Coupled Low-Dimensional Superfluids

Most phase transitions that have been realized in ultracold atom systems are generic first or second order transitions. However, the paradigm of phase transitions in two dimensions at finite temperature is of a more intricate type, a Kosterlitz–Thouless transition, which is characterized by a change of the functional form of the correlation function of the order parameter, from algebraic decay to exponential decay. In an intriguing new development in studying low-dimensional strongly correlated systems, such a Kosterlitz–Thouless (KT) transition [11] was indeed realized and observed [30]. In this experiment the interference amplitude between two independent two-dimensional (2D) Bose systems was studied as a function of temperature. This analysis revealed the jump in the superfluid stiffness (see also [80]) and the emergence of unpaired isolated vortices as they crossed the phase transition.

The other focus of this section, the physics of ramping across a phase transition, is also triggered by a recent experiment: Sadler et al. observed spontaneous generation of topological defects in the spinor condensate after a sudden quench (i. e. a rapid, non-adiabatic ramp) through a quantum phase transition [84]. A similar experiment in a double-layer system was reported in [85]. The topological defects are generated [45] at a density which is related to the rate at which the transition is crossed [108]. Later it was argued that the dependence of the number of such defects on the swipe rate across a quantum critical point can be used as a probe of the critical exponents characterizing the phase transition [81]. This Kibble–Zurek (KZ) mechanism was originally considered as an early universe scenario creating cosmic strings, which would serve as an ingredient for the formation of galaxies[6]. Cold atom systems appear to be a very suitable laboratory for performing such "cosmological experiments", since these systems are highly tunable and well isolated from the environment. So far the experiments and the theoretical proposals addressed the KZ scenario across a quantum phase transition. The main reason is that it is generally hard to cool such systems sufficiently fast to observe non-equilibrium effects. In this work we provide an example of a particular system where this difficulty can be easily overcome by quenching the transition temperature T_c instead of T. Thus the relevant ratio T/T_c can be tuned with an arbitrary rate and the KZ mechanism can be observed. Specifically, we examine a system of two superfluids (SF): As we show below, by turning on tunneling between the two systems the transition temperature increases rapidly, and the system attempts to create long-range order (LRO). However, in this process, defects in the SF phase are created, which develop into long-lived vortex-anti-vortex pairs or in finite system unbalanced population between vortices and anti-vortices. We note that because the systems are isolated and there is no external heat bath, the temperature itself also changes due to the quench. However, the long-wavelength fluctuations relevant for the KT transition are only a small subset of all degrees of freedom, majority of which are only weakly affected by small inter-layer tunneling. So we believe that the change of the T_c is the main effect of the quench.

In this section we consider two SFs coupled via tunneling and/or interactions. In the experiments the hopping or tunneling rate between two systems can be tuned to a high precision [30,31,77,87]. Interactions between the atoms in different systems can either be realized in ensembles of polar molecules or by using mixtures of two hyperfine states, where the tunneling rate is controlled by an infrared light source [41], which induces spin-flipping between the hyperfine states. In this case the atoms in different states naturally interact with each other since they are not physically separated in space. The main results of our analysis are the phase diagrams of coupled SFs in Figs. 6 and 7, the behavior of T_c and the energy gap shown in Fig. 8, as well as the

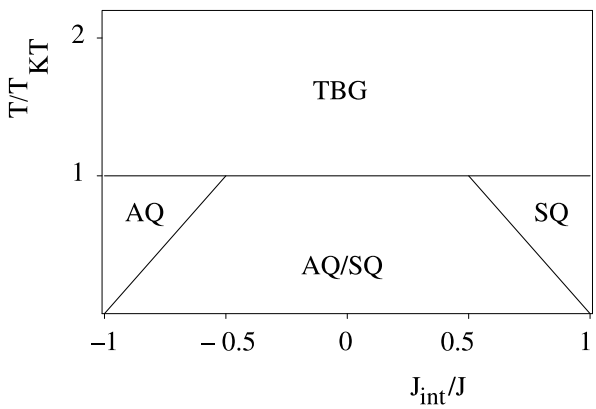

Ultracold Atomic Gases: Novel States of Matter, Figure 6
Phase diagrams of two 2D SFs, coupled through a term of the form S_{12}, Eq. (22), in terms of J_{int}/J and T/T_{KT}. For low temperatures we find antisymmetric quasi-order (AQ) and/or symmetric quasi-order (SQ), which either simultaneously undergo a KT transition due to single vortices (AQ/SQ to thermal Bose gas (TBG) phase), or individually due to correlated vortex pairs: symmetric (anti-symmetric) vortex pairs drive the AQ/SQ to AQ (SQ) transition

[6] Since then, it has been established that cosmic strings can only contribute a small fraction of the initial density perturbations. We thank T. Kibble for this comment.

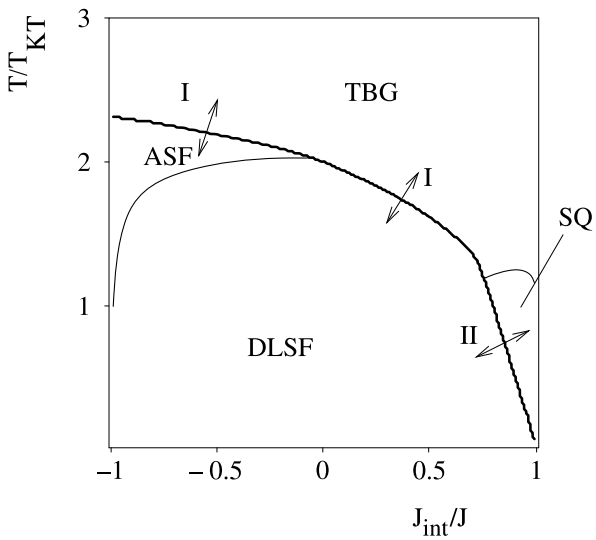

Ultracold Atomic Gases: Novel States of Matter, Figure 7
Phase diagram, temperature (in units of T_{KT}) versus interaction (in units of J). We assume $J_\perp/J \sim 10^{-3}$ and $A_1/J \sim 10^{-3}$. DLSF: double layer superfluid; TBG: thermal Bose gas; ASF: anti-symmetric superfluid; SQ: symmetric quasi-order. The order of the transition lines are either first (I), second order (II), or KT (*thin lines*)

proposal of realizing the KZ mechanism by switching on the tunneling between two SFs.

2D Superfluids

In this section we consider two 2D SFs, each characterized by a KT temperature T_{KT}. We write the bosonic operators $b_{1/2}$ in the two layers in a phase-density representation [11,24], $b_{1/2} \sim \sqrt{\rho_{1/2}}\exp(i\phi_{1/2})$, where $\rho_{1/2}$ are the density operators of the two systems, and $\phi_{1/2}$ the phases. The low-momentum fluctuations of the phase fields are described by Gaussian contributions to the Hamiltonian $\mathcal{H}_0$. Because of the formal analogy between the quantum 1D and thermal 2D systems [22] we adopt the quantum terminology throughout the paper and refer to the ratio of the Hamiltonian and the temperature as the action. Then

$$S_0 \equiv \frac{\mathcal{H}_0}{T} = \frac{J}{2T}\int \mathrm{d}^2 r\left[(\nabla\phi_1)^2 + (\nabla\phi_2)^2\right] . \qquad (18)$$

The energy scale J here is related to T_{KT} by $J = 2T_{KT}/\pi$. Besides these long-wavelength fluctuations, the system also contains additional degrees of freedom, vortex-anti-vortex pairs [11]. The corresponding term in the action is expressed through the dual fields $\theta_{1,2}$ [24]:

$$S_1 = \frac{2A_1}{T}\int \frac{\mathrm{d}^2 r}{(2\pi\alpha)^2}\left[\cos(2\theta_1) + \cos(2\theta_2)\right] , \qquad (19)$$

where α is a short-distance cut-off of the size of the vortex core, and A_1 is proportional to the single-vortex fugacity: $A_1 \sim J\exp(-J/T)$, where we assume both SFs to have the same effective parameters J and A_1. Operators of the type $\exp(2i\theta)$ create kinks in the field ϕ: $\exp(-2i\theta(x))\phi(x')\exp(2i\theta(x)) \sim \phi(x') + 2\pi\Theta(x - x')$, $\Theta(x)$ being the step function, which corresponds to the effect of vortices in the original 2D problem ([22], p 92).

In addition, the two systems are coupled by a hopping term $\sim t_\perp b_1^\dagger b_2 + \text{h.c.}$, which results in the following con-

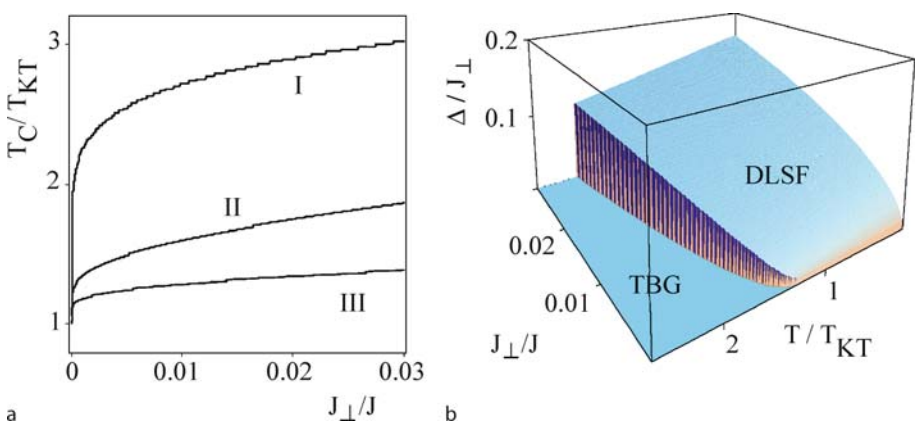

Ultracold Atomic Gases: Novel States of Matter, Figure 8
a Critical temperature T_c of the DLSF-TBG transition (in units of T_{KT}) for different values of A_1/J: 10^{-3}, 0.1, 0.4 (I–III), and for $J_{int} = 0$. b Energy gap in the anti-symmetric sector (in units of $J_\perp$) as a function of $J_\perp/J$ and temperature (in units of T_{KT}). We have set $A_1/J = 0.1$ and $J_{int} = 0$

tribution to the action:

$$S_\perp = \frac{2J_\perp}{T}\int \frac{\mathrm{d}^2 r}{(2\pi\alpha)^2}\cos(\phi_1-\phi_2)\,, \tag{20}$$

where the bare value of $J_\perp$ corresponds approximately to $t_\perp\rho_0$. In principle, the hopping term is modified by the vortex contributions, however, these corrections are always irrelevant under renormalization group (RG).

For most of the discussion in this paper we use the symmetric and anti-symmetric combinations of $\phi_{1/2}$ and $\theta_{1/2}$:

$$\phi_{\mathrm{s/a}} = (\phi_1 \pm \phi_2)/\sqrt{2}\,, \quad \theta_{\mathrm{s/a}} = (\theta_1 \pm \theta_2)/\sqrt{2}\,. \tag{21}$$

Written in these fields, the term S_0 in Eq. (18) is again a sum of Gaussian models, now in the fields ϕ_{s} and ϕ_{a}, with the same energy scale J. However, we will consider a broader class of actions, in which the energy scales of the symmetric and anti-symmetric sector differ. We include the following term in the action:

$$S_{12} = \frac{J_{\mathrm{int}}}{T}\int \mathrm{d}^2 r \nabla\phi_1\nabla\phi_2\,. \tag{22}$$

With this, the quadratic part of the action is given by:

$$S_0 + S_{12} = \frac{J_{\mathrm{s}}}{2T}\int \mathrm{d}^2 r(\nabla\phi_{\mathrm{s}})^2 + \frac{J_{\mathrm{a}}}{2T}\int \mathrm{d}^2 r(\nabla\phi_{\mathrm{a}})^2\,, \tag{23}$$

where J_{s} and J_{a} are given by $J_{\mathrm{s/a}} = J \pm J_{\mathrm{int}}$.

We now motivate the existence of such a term S_{12} in ultra-cold atom systems, by considering two BECs coupled by a short-range density–density interaction. Starting from a Hamiltonian of the form $H = \sum_k[\epsilon_{\mathbf{k}} b^\dagger_{\mathbf{k}} b_{\mathbf{k}} + (g/2V)\rho^\dagger_{\mathbf{k}}\rho_{\mathbf{k}}]$, where $b_{\mathbf{k}}$ is the boson operator, $\epsilon_{\mathbf{k}}$ the free dispersion $\epsilon_{\mathbf{k}} = \mathbf{k}^2/2m$, g is the interaction strength of the contact interaction, V the volume, and $\rho_{\mathbf{k}}$ is the density operator of momentum $\mathbf{k}$, given by $\rho_{\mathbf{k}} = \sum_{\mathbf{p}} b^\dagger_{\mathbf{p}} b_{\mathbf{p}+\mathbf{k}}$, we assume that the zero momentum mode is macroscopically occupied, and formally replace the operator b_0 by a number, $b_0 \to \sqrt{N_0}$, where N_0 is the number of condensed atoms which is comparable to the total atom number N, i. e. $N_0 \sim N$. Next we keep all terms that are quadratic in $b_{\mathbf{k}}$ (with $\mathbf{k} \neq 0$), and perform a Bogoliubov transformation, given by: $b_{\mathbf{k}} = u_{\mathbf{k}}\beta_{\mathbf{k}} + v_{\mathbf{k}}\beta^\dagger_{-\mathbf{k}}$, to diagonalize the Hamiltonian. The eigenmodes $\beta_{\mathbf{k}}$ have a dispersion relation $\omega_{\mathbf{k}} = \sqrt{\epsilon_{\mathbf{k}}(\epsilon_{\mathbf{k}} + 2gn)}$, with n being the density N/V. The low-$\mathbf{k}$ limit is given by $\omega^2_{\mathbf{k}} \sim v^2|\mathbf{k}|^2$, with $v = \sqrt{gn/m}$, which corresponds to the contribution in Eq. (18) of the action. Next, we consider the sum of two copies of the previous Hamiltonian with boson operators $b_{1/2}$. In addition we consider an interaction $H_{12} = g_{12}/V\sum_{\mathbf{k}}\rho^\dagger_{1,\mathbf{k}}\rho_{2,\mathbf{k}}$, where the density operators $\rho_{1/2,\mathbf{k}}$ are given by $\rho_{1/2,\mathbf{k}} = \sum_{\mathbf{p}} b^\dagger_{1/2,\mathbf{p}} b_{1/2,\mathbf{p}+\mathbf{k}}$. Following the same procedure as before, we find two eigenmode branches, corresponding to in-phase and out-of-phase superpositions of the modes of each condensate, with the dispersions $\omega^2_{\mathrm{s/a},\mathbf{k}} \sim v^2_{\mathrm{s/a}}|\mathbf{k}|^2$, with the velocities $v_{\mathrm{s/a}} = \sqrt{(g \pm g_{12})n/m}$. Therefore, for this example, the energy scale J_{int} is related to $g_{12}n/m$, which would be of similar order as J for a system interacting via contact interaction, for small temperatures. This discussion only applies to the weakly interacting limit of a true condensate. However, it demonstrates that a density–density contact interaction term can lead to a substantial energy splitting of the in-phase and out-of phase modes.

Finally, in addition to single vortices in each SF, we have to consider the possibility of correlated vortex pairs, i. e. one vortex in each layer at the same location of either the same or of opposite vorticity. We will refer to these vortex configurations as *symmetric* or *anti-symmetric vortex pairs*, respectively. These excitations appear as the following terms in the action:

$$S_{\mathrm{s,a}} = \frac{2A_{\mathrm{s,a}}}{T}\int \frac{\mathrm{d}^2 r}{(2\pi\alpha)^2}\cos(2\sqrt{2}\,\theta_{\mathrm{s,a}})\,. \tag{24}$$

These correlated vortex terms, which describe new degrees of freedom, can be the most relevant non-linear terms in the action, which derives from the possibility that the vortices in different layers interact with each other, through the terms (20) and (22). The effect of these terms is the following: At low temperatures the energy between two single vortices of opposite vorticity due to tunneling grows as the square of the distance D between them, i. e. as $J_\perp(D/\alpha)^2$. As a result, the tunneling term attempts to confine vortices of opposite vorticity, leading to phase-locking between the layers, which we describe further later on. The interaction S_{12} changes the energy of correlated vortex pairs as follows: The energy of a single vortex is given by $2\pi J \log L/\alpha$, where L is the system size, whereas a symmetric/anti-symmetric vortex pair has an energy of $4\pi(J \pm J_{\mathrm{int}})\log L/\alpha$. Therefore, symmetric vortex pairs are the lowest energy vortex excitations for $J_{\mathrm{int}} < -J/2$, whereas for $J_{\mathrm{int}} > J/2$ anti-symmetric vortex pairs are the lowest energy excitations. As we will see below, in these regimes correlated vortex pairs drive transitions to phases, in which one sector is (quasi-)SF whereas the other is disordered. We will also see that these terms are generated under the RG flow, even if not present at the onset.

We note that a similar system has been studied in [8]. Here, we consider a larger class of systems by including the interaction term (22), which in turn requires us to include

the correlated vortex excitations (24). These terms give rise to additional phases as we will see in the following.

Next we analyze our system within the RG approach. This RG flow is perturbative in the vortex fugacities A_1, A_s, and A_a, and the tunneling energy $J_\perp$, and therefore applies to the weak-coupling limit (in particular $J_\perp \to +0$). At second order the flow equations are given by [6]:

$$\frac{\mathrm{d}J_\perp}{\mathrm{d}l} = \left(2 - \frac{T}{2\pi J_a}\right) J_\perp \,, \tag{25}$$

$$\frac{\mathrm{d}A_s}{\mathrm{d}l} = \left(2 - 2\pi\frac{J_s}{T}\right) A_s + \alpha_3 \frac{A_1^2(J_a - J_s)}{2T^2} \,, \tag{26}$$

$$\frac{\mathrm{d}A_a}{\mathrm{d}l} = \left(2 - 2\pi\frac{J_a}{T}\right) A_a + \alpha_3 \frac{A_1^2(J_s - J_a)}{2T^2} \,, \tag{27}$$

$$\frac{\mathrm{d}A_1}{\mathrm{d}l} = \left(2 - \frac{\pi(J_s + J_a)}{2T} + \alpha_3 \frac{A_s J_s + A_a J_a}{T^2}\right) A_1 \,, \tag{28}$$

$$\frac{\mathrm{d}J_a}{\mathrm{d}l} = \alpha_2 \left(\frac{J_\perp^2}{4\pi^4 J_a} - 4\frac{A_a^2}{T^4} J_a^3 - \frac{A_1^2}{2T^4}(J_s + J_a) J_a^2\right) \,, \tag{29}$$

$$\frac{\mathrm{d}J_s}{\mathrm{d}l} = -\alpha_2 \left(2\frac{A_s^2}{T^4} J_s^2 + \frac{A_1^2}{4T^4}(J_s + J_a) J_s\right) 2J_s \,. \tag{30}$$

The coefficients $\alpha_{2/3}$ are non-universal parameters that appear in the RG procedure [51], and which do not affect the results qualitatively. For consistency, we have to expand the right-hand site of the above equations up to second order, around the resulting Gaussian fixed point: $J_{s/a} = J \pm J_{\mathrm{int}} + j_{s/a}$. We emphasize again that J_{int} near the fixed point can be generated by RG and be nonzero even if it is not present at the onset.

Before we consider the full RG flow, we consider the simpler case of no tunneling, i. e. we solve the RG equations while setting $J_\perp = 0$. In Fig. 6 we show the phase diagram of two 2D SFs coupled by S_{12}, Eq. (22). Such a system would be realized by a 2D mixture of bosonic atoms in two different hyperfine states, interacting via some short-range potential. The order parameters we consider are $O_s(x) = b_1(x)b_2(x)$ and $O_a(x) = b_1^\dagger(x)b_2(x)$. To obtain the phase diagram we consider the correlation functions of each of these order parameters, which can either scale algebraically or exponentially. In Fig. 6 we refer to algebraic scaling of $O_s(x)$ as symmetric quasi-order (SQ), and of $O_a(x)$ as anti-symmetric quasi-order (AQ). In each of the sectors a KT transition marks the transition from the algebraic to the exponential regime, which occur either simultaneously and are driven by single-vortex excitations, or at different temperatures and are driven by correlated vortex pairs. As a result we find four regimes: At temperatures above T_{KT}, both sectors are disordered, giving rise to a thermal Bose gas (TBG) phase. For temperatures below T_{KT}, and for a wide range of J_{int}, we find that both sectors are quasi SF (AQ/SQ), which is the only phase in which the correlation function of the single boson operators show algebraic scaling. We also find regimes in which only one sector shows algebraic scaling, whereas the other is disordered (AQ and SQ). From the perspective of vortices, the TBG phase is a gas of free single vortices in each layer, whereas the AQ (SQ) phase is a gas of symmetric (anti-symmetric) vortex pairs.

We now consider the full RG system, including $J_\perp$. We numerically integrate the RG equations, and find the phase diagram shown in Fig. 7 in terms of the temperature T and the interaction J_{int}. We again find four different phases that are different combinations of LRO, QLRO, and disorder in the symmetric and anti-symmetric sector. At high temperatures we find that both sectors are disordered in a TBG phase, as before. For lower temperatures, and for a wide range of J_{int}, the system is in a double-layer SF phase (DLSF): The symmetric sector shows algebraic scaling, whereas the exponent of the anti-symmetric sector is renormalized to zero, i. e. we find two SFs that are phase-locked due to $J_\perp$. Note that the transition temperature T_c between DLSF and TBG has been noticeably increased relative to the decoupled value T_{KT}, as we will discuss further later on. We also find two additional phases, which are partially (quasi-)SF and partially disordered. One of them is the SQ phase, as before, whereas the other one (ASF), now shows true LRO in the anti-symmetric sector due to $J_\perp$, whereas the symmetric sector remains disordered. We note that the generic double-layer action that we discuss in this paper does not show a sliding phase [71], for any non-zero $J_\perp$. Either S_1 or S_a, which is generated by RG, drives the anti-symmetric sector to a disordered state, or $S_\perp$ creates true LRO in the field ϕ_a.

We also use the RG flow to find the order of the phase transitions in the weak-coupling limit that the anti-symmetric sector undergoes, by determining the energy gap using a 'poor-man's scaling' argument: when the coupling amplitude $J_\perp(l^*)$ is of order unity the corresponding gap is given by the expression $\Delta \sim J_\perp \exp(-l^*)$. From the behavior of Δ at the phase transition we can read off whether it is of first or second order, as indicated in Fig. 7.

Given the nature of an effective theory, only approximate statements can be made about how the different regimes of the phase diagram relate to the microscopic interactions. To create the ASF or AQ phase an attraction between the two atom species is needed that is of order J, whereas to create the SQ phase, a repulsion of that order would be needed. To detect the different phases, one could use the interference method used in [30] to distinguish the phase-locked phases (DLSF and ASF), which would show a well-defined interference pattern, from the uncorrelated

phases. Another approach would be time-of-flight images: The DLSF phase would display a quasi-condensate signature, whereas the other phases would appear disordered. However, at the transition from ASF or SQ to TBG, the width of the distribution would abruptly increase.

Kibble–Zurek Mechanism

In this section we discuss how the phase-locking transition found in the previous section could be used to realize the KZ mechanism. The defining property of this mechanism is the generation of topological defects by ramping across a phase transition, coming from the disordered phase. The disordered phase that we propose to use is the TBG phase of the *decoupled* 2D systems, that is, we consider the experimental setup reported in ref. [30] for a temperature T above the KT temperature T_{KT}. The ordered phase we consider is the DLSF phase, i. e. the phase-locked phase of two *coupled* SFs. The ramping is achieved by turning on the tunneling between the two layers, which can be done by lowering the potential barrier between them. For this procedure the critical temperature T_{c} of the DLSF-TBG transition needs to be above the KT temperature of the uncoupled systems. We now show that the RG flow indeed predicts such a scenario. In the experiments in [30], the atoms in different layers do not interact with each other. Therefore, it can be expected that J_{int} is small, of order $J_{\perp}$, which motivates us to discuss the case $J_{\mathrm{int}} = 0$ here. We note however, that the desired scenario of an increased critical temperature, is found for a wide range of J_{int}, as can be seen in Fig. 7. In Fig. 8a we show how the critical temperature of the DLSF-TBG transition behaves, predicted by the RG flow, for different values of A_1. The critical temperature shows a sizable increase, due to the phase-locking transition. Due to the perturbative nature of the RG scheme, the RG flow underestimates the effects of the term S_1, and predicts a finite jump of the critical temperature when $J_{\perp}$ is turned on. However, to lock the SFs together in the regime slightly above T_{KT}, $J_{\perp}$ needs to be at least of the order of the vortex core energy, giving rise to a finite slope of T_{c} instead of a jump. The energy gap of this transition is shown in Fig. 8 from which we can see that the transition is of first order, in contrast to the second order transition described in [45,108], which is advantageous because the onset of order is instantaneous rather than continuous. We note that the phase diagram was obtained using the assumption that the bare parameters of the model, in particular J, do not depend on temperature. This is true only if temperature is close to T_{KT}. Here we find that the ratio $T_{\mathrm{c}}/T_{\mathrm{KT}}$ can be relatively large. In fact $T_{\mathrm{c}}/T_{\mathrm{KT}}$ will be always smaller than that shown in Fig 8a, however, qualitatively the behavior of $T_{\mathrm{c}}/T_{\mathrm{KT}}$ as a function of $J_{\perp}$ should remain intact. We point out that our results can be generalized to a system of $N > 2$ coupled SFs. One finds that the SFs still show a strong tendency to phase-lock together. As a result the critical temperature should approximately satisfy the equality $\pi J(T_{\mathrm{c}})N = 2T_{\mathrm{c}}$. Thus as N increases T_{c} approaches the mean-field critical temperature at which the stiffness J vanishes and we recover the usual 3D result.

In finite size systems there is another constraint on the minimum value of $J_{\perp}$: We consider the free energy of a single vortex in the anti-symmetric field: $\phi_{\mathrm{a}} \sim \arctan(x/y)$. For the decoupled system we get for the free energy [51]: $F \sim 2(\pi J - 2T)\log L/\alpha$, where L is the system size. The coupling term gives a free energy contribution $F_{\perp} \sim J_{\perp}(L/\alpha)^2$. In the thermodynamic limit, $L \to \infty$, this term diverges faster than the others, which is consistent with our finding of LRO in the antisymmetric sector. For a finite system, comparing these terms gives the estimate $J_{\perp} \sim J\log(L/\alpha)/(L/\alpha)^2$, that is required for this order to develop. With a system size $L/\alpha \sim 10^2$, that would require $J_{\perp} \sim 10^{-3}J$, which, for the setup in [30], would be around $10^2 s^{-1}$.

As an estimate of the number of domains that would be created, we follow the argument in [45]: The coherence scale of the DLSF phase is given by $(J/\Delta)^{1/2}\alpha$, which is the scale of a Klein–Gordon model with a kinetic energy scale J and a 'mass-term' with a prefactor Δ/α^2. The domain size is then given by $(J/\Delta)\alpha^2$, and the number of domains by $\sim (\Delta/J)L^2/\alpha^2$. As we show in Fig. 8b for $J_{\perp}/J \approx 10^{-2}$, we find $\Delta/J_{\perp} \sim 10^{-1}$, and therefore $J/\Delta \sim 10^3$. With $L/\alpha \sim 10^2$, we would get $N_{\mathrm{dom}} \sim 10^1 - 10^2$, which would generate a similar number of vortices. We estimate the vortex-antivortex imbalance by considering the number of domains around the periphery of the system, which scales as L/ξ. If we imagine that the phase behaves like a random walk, the total phase mismatch, corresponding to the vortex-antivortex imbalance, will scale as $\sqrt{L/\xi} \sim N_{\mathrm{dom}}^{1/4}$, which, for $L/\alpha \sim 10^2$, is of the order 10^0–10^1.

In summary, we propose the following procedure: i) Prepare two uncoupled SFs at a temperature T slightly above T_{KT}. ii) Switch on the tunneling between the two layers, which creates a DLSF phase with a critical temperature T_{c} higher than T. As a result, one should find a number of long-lived vortex-antivortex pairs in the anti-symmetric phase field ϕ_{a}, which would be visible in an interference measurement, at a temperature where there would be none in thermal equilibrium.

In conclusion of the section, we studied the phase-locking transition of 2D superfluids, within an renormalization group approach. We find that this transition is accompanied by an increase of the transition temperature.

We suggest that this effect can be used to probe the Kibble–Zurek mechanism in cold atom systems by rapidly changing the ratio T/T_c. When we include interactions between the layers we find additional phases, in which either the symmetric or the anti-symmetric sector is disordered, and the other sector stays superfluid or quasi-superfluid.

Bose–Fermi Mixtures in Two-Dimensional Optical Lattices

In the spirit of engineering many-body systems that are relevant in other fields, we now turn to atomic mixtures that resemble qualitatively, i. e. in terms of degrees of freedom of the system, electron–phonon systems. In two dimensions, these systems are actively studied and prove to be of considerable complexity. In order to study their atomic counterparts, Bose–Fermi mixtures in optical lattices, we use the powerful method of functional renormalization group equations, with which we can determine their phase diagrams in the weak-coupling limit in a systematic fashion. We find a rich competition of phases for both the square lattice and triangular lattice geometry that we consider.

In this section we consider mixtures of one bosonic type of atom and either two fermionic types that are SU(2) symmetric or spinless fermions. The Hamiltonian for a mixture on a square lattice is given by:

$$\begin{aligned} H = &-t_\mathrm{f} \sum_{\langle ij\rangle,s} f_{i,s}^\dagger f_{j,s} - t_\mathrm{b} \sum_{\langle ij\rangle} b_i^\dagger b_j \\ &- \sum_i (\mu_\mathrm{f} n_{\mathrm{f},i} + \mu_\mathrm{b} n_{\mathrm{b},i}) \\ &+ \sum_i \left[U_\mathrm{ff} n_{\mathrm{f},i,\uparrow} n_{\mathrm{f},i,\downarrow} + \frac{U_\mathrm{bb}}{2} n_{\mathrm{b},i} n_{\mathrm{b},i} + U_\mathrm{bf} n_{\mathrm{b},i} n_{\mathrm{f},i} \right], \end{aligned} \tag{31}$$

where $f_{i,s}^\dagger$ ($f_{i,s}$) creates (annihilates) a fermion at site i with pseudo-spin s ($s = \uparrow, \downarrow$), $b_i^\dagger$ (b_i) creates (annihilates) a boson at site i, $n_{\mathrm{f},i} = \sum_s f_{i,s}^\dagger f_{i,s}$ ($n_{\mathrm{b},i} = b_i^\dagger b_i$) is the fermion (boson) number operator, t_f and t_b are the fermionic and bosonic tunneling energies between neighboring sites, μ_f (μ_b) is the chemical potential for fermions (bosons), U_bb is the repulsion energy between bosons on the same site, U_ff is the repulsion energy between the two species of fermions, and U_bf is the repulsion energy between bosons and fermions. The two fermion species have been treated as a pseudo-spin-1/2 index ($\uparrow$ and $\downarrow$). The case of spinless fermions can be immediately obtained from (31) by ignoring one of the spin states. In momentum space, the Hamiltonian (31) is written as:

$$\begin{aligned} H = \sum_\mathbf{k} \Bigg\{ &(\epsilon_{\mathrm{f},\mathbf{k}} - \mu_\mathrm{f}) \sum_s f_{\mathbf{k},s}^\dagger f_{\mathbf{k},s} + (\epsilon_{\mathrm{b},\mathbf{k}} - \mu_\mathrm{b}) b_\mathbf{k}^\dagger b_\mathbf{k} \\ &+ \frac{U_\mathrm{ff}}{V} \rho_{\mathrm{f},\mathbf{k},\uparrow} \rho_{\mathrm{f},-\mathbf{k},\downarrow} + \frac{U_\mathrm{bb}}{2V} \rho_{\mathrm{b},\mathbf{k}} \rho_{\mathrm{b},-\mathbf{k}} + \frac{U_\mathrm{bf}}{V} \rho_{\mathrm{b},\mathbf{k}} \rho_{\mathrm{f},-\mathbf{k}} \Bigg\}, \end{aligned} \tag{32}$$

where $\rho_{\mathrm{f},\mathbf{k}} = \sum_{\mathbf{q},s} f_{\mathbf{k}+\mathbf{q},s}^\dagger f_{\mathbf{k},s}$ ($\rho_{\mathrm{b},\mathbf{k}} = \sum_\mathbf{q} b_{\mathbf{k}+\mathbf{q}}^\dagger b_\mathbf{k}$) is the fermion (boson) density operator, $\epsilon_{\mathrm{b/f},\mathbf{k}} = -2t_{\mathrm{b/f}}(\cos k_x + \cos k_y)$, is the bosonic/fermionic dispersion relation.

We consider the limit of weakly interacting bosons that form a BEC [62,104], where we assume that the zero momentum bosonic mode is macroscopically occupied, and the corresponding operator b_0 can be formally replaced by a real number $b_0 \to \sqrt{N_0}$, where N_0 is the number of condensed atoms. After this replacement we keep all terms that are quadratic in $b_\mathbf{k}$ (with $\mathbf{k} \neq 0$), and perform a Bogoliubov transformation, given by: $b_\mathbf{k} = u_\mathbf{k} \beta_\mathbf{k} + v_\mathbf{k} \beta_{-\mathbf{k}}^\dagger$, to diagonalize the bosonic Hamiltonian. The resulting eigenmodes $\beta_\mathbf{k}$ have a dispersion relation given by $\omega_\mathbf{k} = \sqrt{\epsilon_{\mathrm{b},\mathbf{k}}(\epsilon_{\mathrm{b},\mathbf{k}} + 2U_\mathrm{bb} n_\mathrm{b})}$, with the low-$\mathbf{k}$ limit $\omega_\mathbf{k} \sim v_\mathrm{b}|\mathbf{k}|$, with $v_\mathrm{b} = \sqrt{2t_\mathrm{b} U_\mathrm{bb} n_\mathrm{b}}$. The parameters $u_\mathbf{k}$ and $v_\mathbf{k}$ are given by: $u_\mathbf{k}^2 = (\omega_\mathbf{k} + \epsilon_{\mathrm{b},\mathbf{k}} + U_\mathrm{bb} n_\mathrm{b})/(2\omega_\mathbf{k})$ and $v_\mathbf{k}^2 = (-\omega_\mathbf{k} + \epsilon_{\mathrm{b},\mathbf{k}} + U_\mathrm{bb} n_\mathrm{b})/(2\omega_\mathbf{k})$.

The density fluctuations of the bosons are approximated by: $\rho_{\mathrm{b},\mathbf{k}} \approx \sqrt{N_0}(u_\mathbf{k} + v_\mathbf{k})(\beta_\mathbf{k} + \beta_{-\mathbf{k}}^\dagger)$, with $\mathbf{k} \neq 0$. The interaction between bosons and fermions is then given by $U_\mathrm{bf}\sqrt{N_0}/V \sum_\mathbf{k} (u_\mathbf{k} + v_\mathbf{k})(\beta_\mathbf{k} + \beta_{-\mathbf{k}}^\dagger)\rho_{\mathrm{f},-\mathbf{k}}$. As a next step we integrate out the bosonic modes and use an instantaneous approximation, leading to the following effective Hamiltonian:

$$\begin{aligned} H_\mathrm{eff.} = \sum_\mathbf{k} \Bigg\{ &(\epsilon_\mathbf{k} - \mu_\mathrm{f}) \sum_s f_{\mathbf{k},s}^\dagger f_{\mathbf{k},s} + \frac{U_\mathrm{ff}}{V} \rho_{\mathrm{f},\mathbf{k},\uparrow} \rho_{\mathrm{f},-\mathbf{k},\downarrow} \\ &+ \frac{1}{2V} V_{\mathrm{ind.},\mathbf{k}} \rho_{\mathrm{f},\mathbf{k}} \rho_{\mathrm{f},-\mathbf{k}} \Bigg\}, \end{aligned} \tag{33}$$

where the induced potential $V_{\mathrm{ind.},\mathbf{k}}$ is given by:

$$V_{\mathrm{ind.},\mathbf{k}} = -\tilde{V}/(1 + \xi^2(4 - 2\cos k_x - 2\cos k_y)), \tag{34}$$

with $\tilde{V}$ given by $\tilde{V} = U_\mathrm{bf}^2/U_\mathrm{bb}$, and ξ is the healing length of the BEC and is given by $\xi = \sqrt{t_\mathrm{b}/2n_\mathrm{b}U_\mathrm{bb}}$. This approach is only valid when $v_\mathrm{b} \gg v_\mathrm{f}$, so that the fermion–fermion interaction mediated by the bosons can be considered as instantaneous. Away from this limit, retardation effects are present. In this case, one has to consider the frequency dependence of the interaction explicitly [50,96,97]. The full effective interaction, including re-

tardation, is given by:

$$V_{\text{ind.}}(\omega, \mathbf{k}) = -\left[\frac{\tilde{V}}{1+\xi^2(4-2\cos k_x - 2\cos k_y)}\right]\frac{\omega_{\mathbf{k}}^2}{\omega^2+\omega_{\mathbf{k}}^2}\,, \quad (35)$$

and the static limit (34) is recovered when $\omega_{\mathbf{k}} \gg \omega$. Equation (33) describes the scattering of two fermions from momenta $\mathbf{k}_1$ and $\mathbf{k}_2$, that are scattered into momenta $\mathbf{k}_3$ and $\mathbf{k}_4$. Momentum conservation at the interaction vertex requires that $\mathbf{k}_4 = \mathbf{k}_1 + \mathbf{k}_2 - \mathbf{k}_3$, and hence the interaction vertex, $U(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3)$, depends on three momenta. Its bare value from (33) can be written as:

$$U(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3) = U_{\text{ff}} + V_{\text{ind.},\mathbf{k}_1-\mathbf{k}_3}\,. \quad (36)$$

For the case with retardation, there is dependence on both the momenta and frequencies of the electrons so we have $U(k_1, k_2, k_3)$, with $k_4 = k_1 + k_2 - k_3$, where $k_i = (\omega, \mathbf{k})$.

Starting from non-interacting fermions, we ask the general question of what new many-body phases can emerge when the system is subjected to a given interaction $U(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3)$. Our approach to address this question is the renormalization-group method, described in the next section.

Renormalization-Group Method

Starting with a microscopic model of interacting electrons on a lattice, the renormalization-group (RG) method provides the effective model at a given temperature or energy scale [89]. The RG is implemented by systematically tracing out high energy degrees of freedom in a region between Λ and $\Lambda + \mathrm{d}\Lambda$, where Λ is the energy cut-off of the problem. In this process, the vertex U is renormalized. At the initial value of the cut-off $\Lambda = \Lambda_0$, the value of U is given by its bare value. For the BFM system we describe here, it is given by (36). At one loop, the RG flow is obtained from a series of coupled integral-differential equations [106] for all the different interaction vertices $U(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3)$. The RG equations read:

$$\begin{aligned}\partial_\ell U_\ell(\mathbf{k_1}, \mathbf{k_2}, \mathbf{k_3}) = &-\int_{p,\omega} \partial_\ell[G_{p\ell}G_{k\ell}]U_\ell(\mathbf{k_1}, \mathbf{k_2}, \mathbf{k})U_\ell(\mathbf{p}, \mathbf{k}, \mathbf{k_3})\\ &-\int_{p,\omega} \partial_\ell[G_{p\ell}G_{q_1\ell}]U_\ell(\mathbf{p}, \mathbf{k_2}, \mathbf{q_1})U_\ell(\mathbf{k_1}, \mathbf{q_1}, \mathbf{k_3})\\ &-\int_{p,\omega} \partial_\ell[G_{p\ell}G_{q_2\ell}]\Big\{-2U_\ell(\mathbf{k_1}, \mathbf{p}, \mathbf{q_2})U_\ell(\mathbf{q_2}, \mathbf{k_2}, \mathbf{k_3})\\ &\qquad + U_\ell(\mathbf{p}, \mathbf{k_1}, \mathbf{q_2})U_\ell(\mathbf{q_2}, \mathbf{k_2}, \mathbf{k_3})\\ &\qquad + U_\ell(\mathbf{k_1}, \mathbf{p}, \mathbf{q_2})U_\ell(\mathbf{k_2}, \mathbf{q_2}, \mathbf{k_3})\Big\} \quad (37)\end{aligned}$$

where $\ell = \ln(\Lambda_0/\Lambda)$, $\mathbf{k} = \mathbf{k}_1 + \mathbf{k}_2 - \mathbf{p}$, $\mathbf{q}_1 = \mathbf{p} + \mathbf{k}_2 - \mathbf{k}_3$, $\mathbf{q_2} = \mathbf{p} + \mathbf{k}_1 - \mathbf{k}_3$, and $G_{k\ell} = \Theta(|\xi_{\mathbf{k}}| - \Lambda)/(i\omega - \xi_{\mathbf{k}})$ with $\xi_{\mathbf{k}} = \epsilon_{\text{f},\mathbf{k}} - \mu_{\text{f}}$ and $k = (\omega, \mathbf{k})$.

From the general interaction vertices $U(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3)$, the specific interaction channels, such as charge-density wave (CDW), antiferromagnetic (AF), and superconducting (BCS), can be obtained:

$$V^{\text{CDW}} = 4\,U_{\text{c}}(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_1 + \mathbf{Q})\,, \quad (38)$$

$$V^{\text{AF}} = 4\,U_\sigma(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_1 + \mathbf{Q})\,, \quad (39)$$

$$V^{\text{BCS}} = U(\mathbf{k}_1, -\mathbf{k}_1, \mathbf{k}_2)\,, \quad (40)$$

where we have used the notation: $U_{\text{c}} = (2 - \hat{X})U/4$, $U_\sigma = -\hat{X}U/4$ with $\hat{X}U(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3) = U(\mathbf{k}_2, \mathbf{k}_1, \mathbf{k}_3)$, and $\mathbf{Q}$ is the nesting vector, $\mathbf{Q} = (\pi, \pi)$.

In a numerical implementation, one discretizes the Fermi surface into M patches, and hence each of the interaction channels (38), (39), (40) is represented by an $M \times M$ matrix. At each RG step, we diagonalize each of these matrices. The channel with the largest eigenvalue (with the caveat that a BCS-channel needs to be attractive to drive a transition) corresponds to the dominant order. The elements of the eigenvector are labeled by the discrete patch indices around the Fermi surface and the symmetry of the order parameter is given by this angular dependence.

The RG method for interacting fermions has been extended to also include retardation effects, as for the case of interacting electrons which are also coupled to phonons in a crystal [96,97]. In this case: (i) the interaction vertices also depend on frequencies of the incoming and outgoing fermions, so the RG equations are written for given external frequencies and the integral over intermediate frequencies can not be done analytically, and (ii) there are important self-energy corrections (in particular the imaginary part of the self-energy is non-zero). Eliashberg equations for strong-coupling superconductivity has been derived with this method [96,97] for the case of electrons, with a circular Fermi surface, coupled to phonons. This method has also been applied to other electron–phonon problems [49,94], and to mixtures of cold atoms in an optical lattice [50].

Phase Diagram and Sub-Dominant Orders

The microscopic parameters in the Hamiltonian (31) determine the initial conditions for the RG flow, and the shape of the Fermi surface. With these, we write the RG flow equations and solve them numerically. We first discuss the case without retardation. For some parameters,

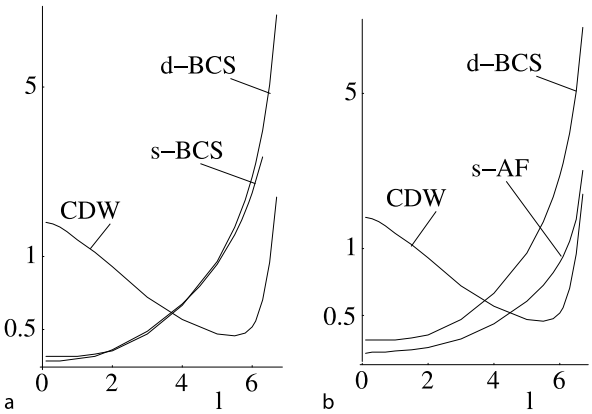

Ultracold Atomic Gases: Novel States of Matter, Figure 9
RG flow for the different effective interactions (in units of t_f) as a function of the RG parameter l ($\tilde{V}/t_f = 3$ and $\xi = 1$). a $U_{ff}/t_F = 0.5$; b $U_{ff}/t_f = 1.2$

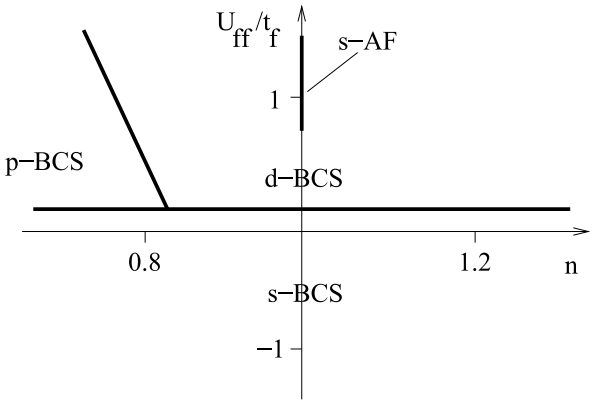

Ultracold Atomic Gases: Novel States of Matter, Figure 10
Phase diagram, interaction strength, U_{ff}/t_f, versus number of fermions per site, n, for a Fermi–Bose mixture in a square lattice in 2D ($\tilde{V}/t_f = 2$, $\xi = 1$)

we encounter a divergence in the RG flow, indicating the onset of ordering with a gap that is in the detectable regime, i. e. larger than $10^{-3} t_f$. In other regimes, where such a divergence is not reached, one can read off the dominant tendency of the RG flow. In Fig. 9 we show examples of RG flows as a function of ℓ. In Fig. 9a, we show the competition between d-wave and s-wave pairing, with d-wave being dominant and s-wave being subdominant. In Fig. 9b we show an example with dominant d-wave channel and subdominant AF channel. In both cases we find that for short distances (or high energies) CDW fluctuations are dominant, giving rise to a state that resembles the findings for high-T_c superconductors [33,34,99]. In some situations the many-body states are almost degenerate and small changes in the initial conditions (that is, changes in the form of the interactions) can be used to select one particular ground state.

With this procedure we determine the phase diagram of the system, which is shown in Fig. 10. We now discuss the general features of the phase diagram. In the absence of any coupling to the bosons, i. e. for $\tilde{V} = 0$, the system shows s-wave pairing for attractive interaction, $U_{ff} < 0$, and no ordering for $U_{ff} > 0$, i. e. Fermi liquid behavior, except for the special case of half-filling where Fermi surface nesting drives the system to AF order for repulsive interactions, and to s-wave pairing (degenerate with CDW) for attractive interaction. If we now turn on the interaction to the bosons, this picture is modified in the following way: The boundary of the s-wave regime is moved into the regime of positive U_{ff}, approximately to a value of U_{ff} where the effective interaction at the nesting vector $\mathbf{Q}$ between the fermions, $U_{ff} + V_{\text{ind.},\mathbf{Q}}$, is positive, i. e. for $U_{ff} \approx \tilde{V}/(1 + 8\xi^2)$. On the repulsive side, and away from half-filling, we find the tendency to form a paired state, either d-wave or p-wave. This tendency becomes weaker the further the system is away from half-filling. We typically find a gap in the vicinity of half-filling and further away from $\mu = 0$ we find only an increasing strength of the corresponding interaction channel. For the half-filled system, we find that for attractive interactions the degeneracy between s-wave pairing and CDW ordering is lifted, with s-wave pairing being the remaining type of order. For repulsive interactions, we find an intermediate regime of d-wave pairing, and for larger values of U_{ff} we obtain AF order.

The RG approach also allows the extraction of the many-body gaps in the system through a "poor man's scaling" analysis of the divergent flow: at the point where the coupling becomes of order of t_f the scaling parameter ℓ reaches the maximum value $\ell_c = \ln(t_f/\Delta)$, where Δ is the value of the gap. Hence, $\Delta/t_f \approx \exp\{-\ell_c\}$ can be obtained from the RG flows such as the ones in Fig. 9. In Fig. 11 we show the gaps of the problem as a function of U_{ff}/t_f in the half-filled case. One can see that as U_{ff} increases, from negative to positive values, the s-wave gap is replaced by a d-wave gap, and finally for an antiferromagnetic gap. As is apparent from this figure, the gap in the d-wave phase is much smaller than the gaps of the AF order and the s-wave pairing, and, furthermore, almost independent of the value of U_{ff}. The latter is the case because the U_{ff} term is a pure s-wave contribution to the interaction and therefore does not contribute to the d-wave channel. The d-wave channel has an initial contribution which is entirely due to the anisotropy of the induced interaction,

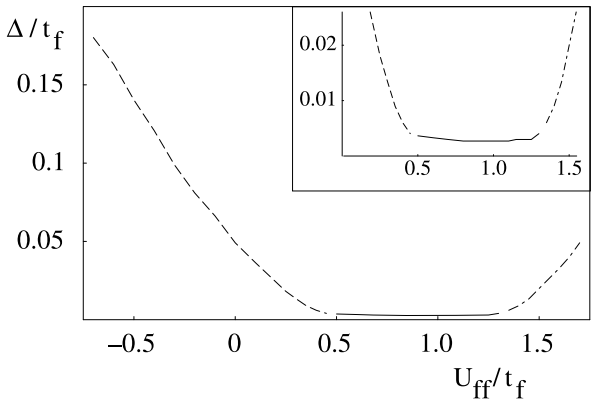

Ultracold Atomic Gases: Novel States of Matter, Figure 11
Many-body energy gaps at half-filling ($\mu = 0$), as a function of U_{ff}/t_f, for $\tilde{V}/t_f = 3$ and at a fixed value of the coherence length $\xi = 1$. *Dashed line*: *s*-wave gap; *continuous line*: *d*-wave gap; *dotted-dashed line*: antiferromagnetic gap. The *inset* shows a magnified plot of the *d*-wave regime

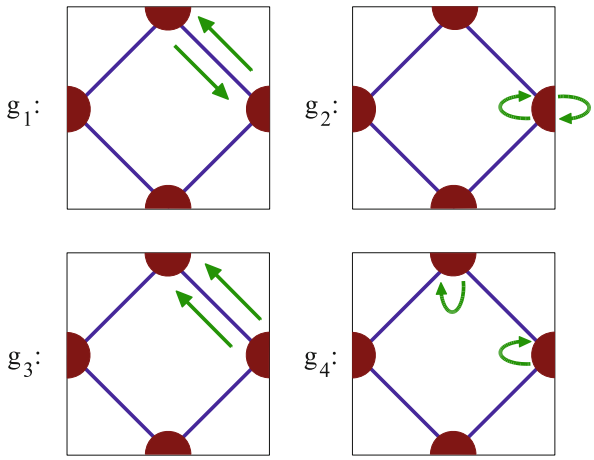

Ultracold Atomic Gases: Novel States of Matter, Figure 12
Relevant processes in the two-patch approximation

which gives only a small value, and as a consequence only a small value for the gap. The value of the gap (in units of t_f) can be numerically fitted with a BCS expression of the form $a\exp(-b/\tilde{V})$, with the parameters a and b given by $a = 0.31$ and $b = 14.2$.

For a system of spinless fermions, one can simply suppress one of the spin indices in (31) and (32). In this case there is a major simplification in the problem since U_{ff} is absent: in a spinless problem there can be only one fermion per site, as per Pauli's principle. Hence, in the absence of bosons, the spinless gas is non-interacting. The bosons, however, mediate the interaction between the fermions. Since the fermions are in different lattice sites the pair wavefunction has necessarily a node and hence, no *s*-wave pairing is allowed. In other words, in the spinless case the anti-symmetry of the wavefunction requires pairing in an odd angular momentum channel. In fact, we find that throughout the entire phase diagram the fermions develop *p*-wave pairing. At half-filling we find a similar behavior of CDW fluctuations on short scales, analogous to the flow shown in Fig. 9. One should point out that in real solids the conditions of "spinlessness" behavior is hard to achieve since it usually requires complete polarization of the electron gas, that is, magnetic energies of the order of the Fermi energy (a situation experimentally difficult to achieve in good metals). However, in cold atom lattices this situation can be easily accomplished with the correct choice of atoms.

Finally, when retardation is important, the numerical task of solving the RG flow equations become much more demanding. In addition to the discretization of the Fermi surface, one has to also discretize the frequency (for $T = 0$), or consider a certain number of Matsubara frequencies ($T \neq 0$). This has been done for this Bose–Fermi system only for a fixed density of fermions corresponding to one half [50]. In this case the Fermi surface has a diamond shape and scattering processes are dominated by the van Hove points (corners of the diamond) where the density of states is singular. The Fermi surface can therefore be approximated by the van Hove points only [86] so that he types of relevant processes are reduced, as shown in Fig. 12. Each of the processes still depend on frequencies: $g_i(\omega_1, \omega_2, \omega_3)$, for $i = 1, 2, 3, 4$. The phase diagram is shown in Fig. 13. Retardation leads to additional phases at half-filling and by tuning the lattice parameters, the system undergoes AF (or spin-density-wave SDW), *d*-wave SC pairing, *s*-wave-pairing, and CDW. In the limit of $v_b \gg v_f$ discussed previously, when retardation is not important, CDW does not become dominant (Fig. 10). It is at most degenerate with *s*-wave pairing for $U_{ff} < 0$. As the bosons become slower, there is stronger tendency for CDW formation (Fig. 13).

Quantum Frustration in Triangular Lattices

It is known that the geometric shape of the lattice is a crucial factor in determining the properties of interacting many-body systems. For instance, localized spins interacting antiferromagnetically on a triangular lattice suffer from the phenomenon of *frustration*, when antiferromagnetic order cannot be achieved because of the particular lattice structure. For itinerant fermionic systems, the lattice structure, together with the dispersion relation and the filling fraction, determine the shape of the

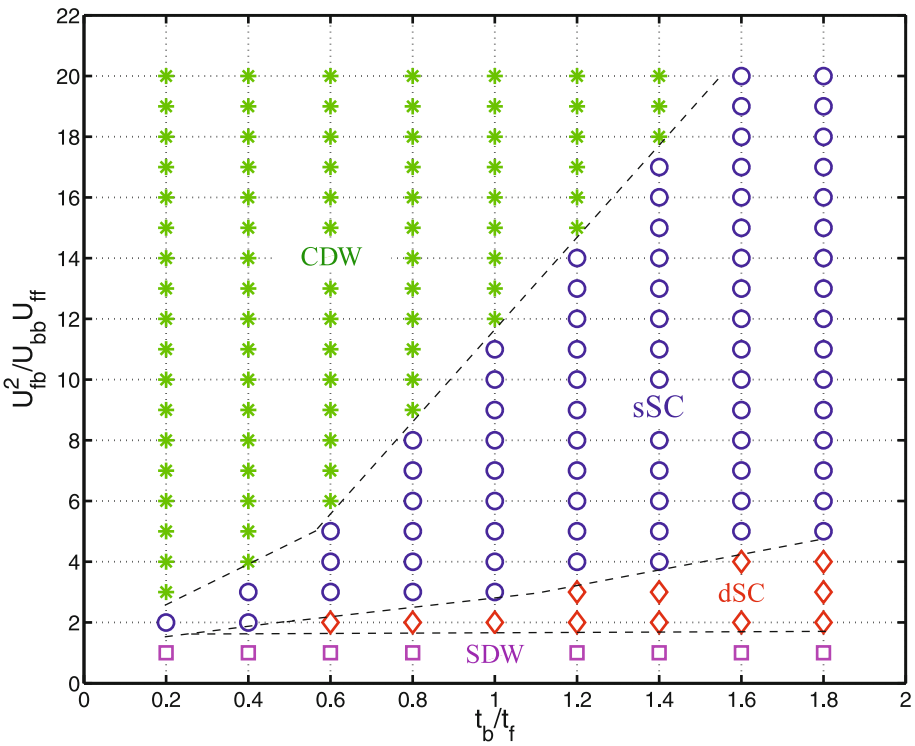

Ultracold Atomic Gases: Novel States of Matter, Figure 13
Phase diagram for $U_{ff} = 0.4t_f$, $U_{bb} = 0.8t_f$, and $n_b = 2.5$. *Blue circles* indicate *s*-wave SC, *red rhombuses* indicate *d*-wave SC, *magenta squares* AF (also called spin-density-wave SDW), and *green stars* CDW type of ordering. *Dashed lines* are guides to the eye

Fermi surface. The Fermi surface, by its turn, is a crucial factor in determining what type of orders the system can develop. Indeed, for the triangular lattice we consider in this section, which shows a rich and subtle competition between superconducting phases with different symmetries, small changes in the shape of the FS determine which pairing symmetry is dominant. This is a reflection of the "lattice frustration" on the superconducting phases. In solids, this intriguing lattice geometry is realized in materials such as cobaltates [93], transition metal dichalcogenides [105] and κ-(ET)$_2$X layered organic crystals [40] (if each lattice site is represented by one ET dimer [46]), and has been the subject of several theoretical studies [5,13,20,35,54,95,103,107].

In this section we consider a BFM on a triangular lattice. The geometry of the lattice under consideration here is shown in Fig. 15a. This system is described by a similar Hubbard model as for the square lattice. But now, besides the triangular geometry, we allow for two different values for the hopping amplitudes, for two types of lattice bonds, as indicated in Fig. 15a by dashed and continuous lines. $t_{f,a}$ and $t_{b,a}$ with $a = 1, 2$ are the fermionic and bosonic tunneling amplitudes between neighboring sites, where the index $a = 1$ ($a = 2$) refers to the continuous (dashed) bonds. For the description of the isotropic case we equate $t_{b/f,1}$ and $t_{b/f,2}$, and define $t_f \equiv t_{f,1} = t_{f,2}$ and $t_b \equiv t_{b,1} = t_{b,2}$. μ_f (μ_b) is the chemical potential for fermions (bosons), U_{bb}, U_{ff}, and U_{bf} are the on-site boson-boson, fermion–fermion and boson–fermion repulsion energy, respectively.

Just as for the case of a square lattice, we consider the limit of weakly interacting bosons, in which the bosons form a BEC, for which we use the same description. The resulting dispersion relation is now given by $\omega_{\mathbf{k}} = \sqrt{(\epsilon_{b,\mathbf{k}} - \epsilon_{b,0})(\epsilon_{b,\mathbf{k}} - \epsilon_{b,0} + 2U_{bb}n_b)}$, where the bare lattice dispersion is given by:

$$\begin{aligned}\epsilon_{b,k} = &- t_{b,1} 2\cos k_x \\ &- t_{b,2}(2\cos(k_x/2 + \sqrt{3}k_y/2) \\ &+ 2\cos(k_x/2 - \sqrt{3}k_y/2))\,.\end{aligned} \tag{41}$$

For small values of k_x and k_y, $\omega_{\mathbf{k}}$ can be expanded as: $\omega_{\mathbf{k}} \sim \sqrt{((2t_{b,1} + t_{b,2})k_x^2 + 3t_{b,2}k_y^2)U_{bb}n_b}$, which gives us

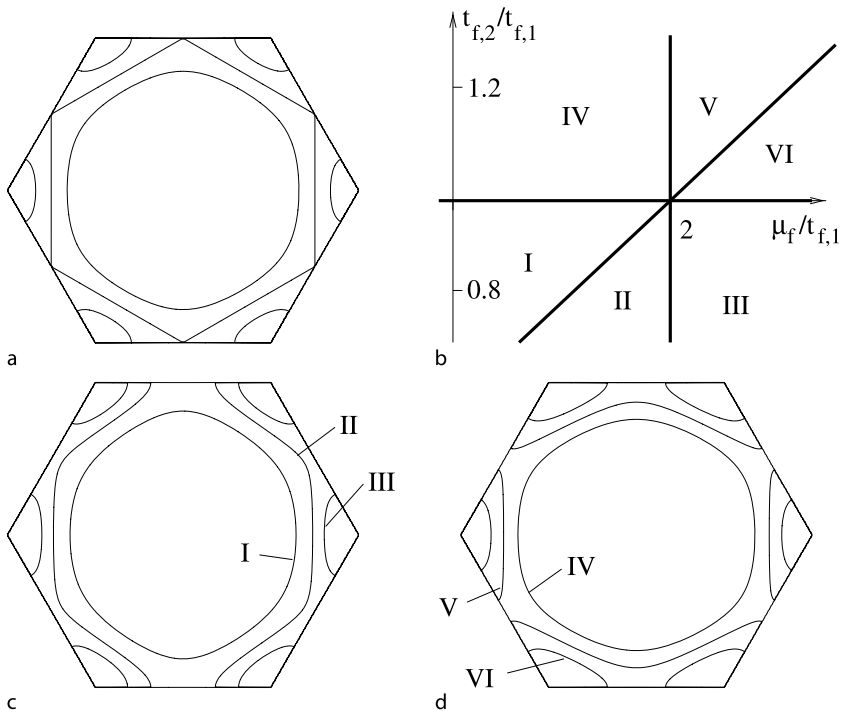

Ultracold Atomic Gases: Novel States of Matter, Figure 14

a Fermi surfaces for an isotropic lattice, for $\mu_f < 2t_{f,1/2}$, $\mu_f = 2t_{f,1/2}$ (*hexagonal shape*), and $\mu_f > 2t_{f,1/2}$ (*six disjoint arcs*). b Diagram of the different types of Fermi surfaces that can be created on an anisotropic lattice, by varying the ratio $t_{f,2}/t_{f,1}$ and μ_f. c Fermi surfaces for $t_{f,2} < t_{f,1}$, for $\mu_f < 4t_{f,2} - 2t_{f,1}$, $4t_{f,2} - 2t_{f,1} < \mu_f < 2t_{f,1}$, and $\mu_f > 2t_{f,1}$, corresponding to the regimes I–III, respectively. d Fermi surfaces for $t_{f,2} > t_{f,1}$, for $\mu_f < 2t_{f,1}$, $2t_{f,1} < \mu_f < 4t_{f,2} - 2t_{f,1}$, and $\mu_f > 4t_{f,2} - 2t_{f,1}$, corresponding to the regimes IV–VI, respectively

the two velocities $v_{b,x} = \sqrt{(2t_{b,1} + t_{b,2})U_{bb}n_b}$ and $v_{b,y} = \sqrt{3t_{b,2}U_{bb}n_b}$.

We again assume that these velocities of the condensate fluctuations are much larger than the Fermi velocity, which corresponds to the conditions $v_{b,x/y} > t_{f,1/2}$. Therefore, large bosonic hopping amplitudes, a bosonic density of $\approx 1-3$, and some intermediate value for $t_{f,1/2}$ will satisfy this requirement. As before, the bosonic modes can be integrated out, and we obtain an approximately non-retarded fermion–fermion interaction. The induced potential $V_{\text{ind.},\mathbf{k}}$ is given by: $V_{\text{ind.},\mathbf{k}} = -\tilde{V}/(1+\xi_1^2(2-2\cos k_x)+\xi_2^2(4-4\cos(k_x/2)\cos(\sqrt{3}k_y/2)))$ with $\tilde{V} = U_{bf}^2/U_{bb}$, and ξ_a are the healing lengths of the Bose–Einstein condensate (BEC) and are given by $\xi_a = \sqrt{t_{b,a}/2n_bU_{bb}}$ with $a = 1, 2$. We again arrive at a purely fermionic, non-retarded description of the same form as before. This is the effective model that we study with a numerical implementation of the functional renormalization group. For the isotropic case, perfect nesting occurs at 3/4-filling, with three possible nesting vectors: $\mathbf{Q}_1 = (0, 2\pi)$, $\mathbf{Q}_2 = (\pi, \sqrt{3}\pi)$, and $\mathbf{Q}_3 = (-\pi, \sqrt{3}\pi)$, leading to three different possible types of instabilities per density wave channel. For the anisotropic case, only $\mathbf{Q}_1$ can be a nesting vector, for the condition $\mu_f = 2t_{f,1}$. To determine the scale of the gaps, Δ, associated with each of these order parameters, we again use a 'poor man's' scaling estimate, specifically: $\Delta \approx \Lambda_0 e^{-\ell_c}$, where ℓ_c is the point at which the RG flow diverges and the instability occurs.

The RG is implemented numerically by discretizing the FS into M patches. For the results shown in this section, $M = 24$ or 36 was used. The CDW, AF and BCS channels are diagonalized at each RG step. The dominant instability is the channel that has an eigenvalue (divided by the dimension of the matrix) with the largest magnitude (for BCS one has to ensure that such eigenvalue is negative so that the channel is attractive). Each element of the corresponding eigenvector represents a given FS patch, and hence, the symmetry of the dominant order parameter is reflected on the patch (i. e., angular) dependence of each element around the FS. Using this method, we determine the phase diagram of the system in various limits.

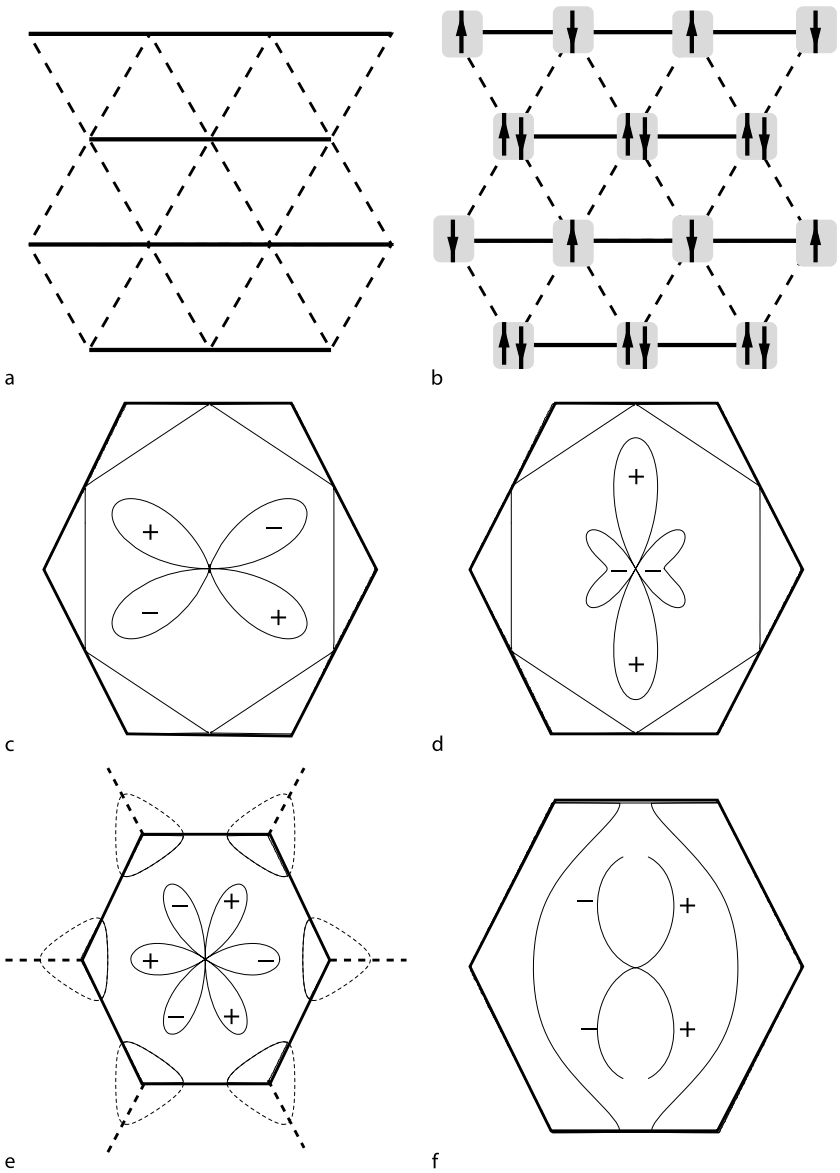

Ultracold Atomic Gases: Novel States of Matter, Figure 15

a Lattice geometry of the system. The continuous (*dashed*) bonds correspond to the hopping amplitudes $t_{b/f,1(2)}$. For $t_{b/f,1} = t_{b/f,2}$, the lattice is an isotropic triangular lattice. b schematic representation of the AF order corresponding to nesting vector Q_1. c, d Order parameters of the extended *d*-wave orders D_1 and D_2. e Order parameter of the *f*-wave phase. This order can also be interpreted as two *s*-wave paired hole states whose order parameters are out of phase by π. f Order parameter of the extended *p*-wave phase, that appears in anisotropic lattices

We first consider spin-1/2 fermions on an isotropic triangular lattice, i. e. with $t_{f,1} = t_{f,2} \equiv t_f$. The FS for such a lattice behaves as follows: For small filling the FS consists of one near-circular piece, which then approaches the shape of a hexagon as μ_f approaches the special value $\mu_f = 2t_f$. At this special chemical potential, which corresponds to 3/4-filling, the FS is nested with the three distinct nesting vectors $\mathbf{Q}_i$. For filling fractions larger than 3/4 the FS breaks into six disjoined arcs. Examples for these different regimes are shown in Fig. 14a. Without coupling to the BEC, the fermions form an *s*-wave pairing phase for attractive interactions, and a Fermi liquid phase for repulsive interactions (ignoring high angular momentum pairing phases predicted by the Kohn–Luttinger the-

orem [53] which would occur at energy scales much lower than the experimentally accessible regime), except for the specific case $\mu_f = 2t_f$, where the system shows AF order for repulsive interactions. A schematic picture of this order is shown in Fig. 15b for the nesting vector $\mathbf{Q}_1$. This behavior is similar to the one found for isotropic square lattice in the previous section [62]: *s*-wave pairing for attractive interaction, and Fermi liquid behavior for repulsive interaction, except at an special filling, for which we find AF order due to nesting. An interesting difference for the triangular lattice is the three-fold degeneracy of the AF phase, an indication of frustration.

When the coupling to the BEC is turned on, the isotropic triangular lattice shows a phase diagram of the type shown in Fig. 8. The *s*-wave pairing phase slightly extends into the regime of positive U_{ff}, because of the induced attractive interaction mediated by the bosonic fluctuations. The regime that showed Fermi liquid behavior in the absence of the induced interaction now shows a rich competition of various types of pairing. In the regime where the density is below half-filling, when the FS is approximately circular, the system shows *p*-wave pairing. For fillings larger than 3/4, when the FS consists of six disjoined parts, the fermions Cooper pair in a superconducting state with *f* symmetry. As shown in Fig. 15e, the FS in this regime can also be interpreted as two distinct near-circular Fermi surfaces of holes. In this interpretation each of these two fermionic systems is in an *s*-wave pairing phase, but the relative phase between the two order parameters is π. At 3/4-filling and large values of U_{ff}, the system still shows AF order. However, for smaller values of U_{ff}, and also for smaller fillings, two phases with degenerate extended *d* symmetry develop. These superconducting orders have a sizable *g*-wave component and are approximately given by:

$$\psi_{D_1} = \sin 2\theta + 0.5 \sin 4\theta \,, \tag{42}$$

$$\psi_{D_2} = \cos 2\theta - 0.5 \cos 4\theta \,. \tag{43}$$

These order parameters are shown in Fig. 15c and d. The shapes of the order parameters are energetically advantageous because, on the one hand, the order parameter maxima are located at points at which the system has a high density of states (the 'corners' of the FS). Hence, when the superconducting gap opens, there is a large gain of condensation energy coming from these regions on the FS. On the other hand, the *d*-wave state has lower kinetic energy than the *f*-wave, and hence is selected.

The phase diagram Fig. 8 has a number of similarities to the phase diagram for a BFM on a square lattice, such as

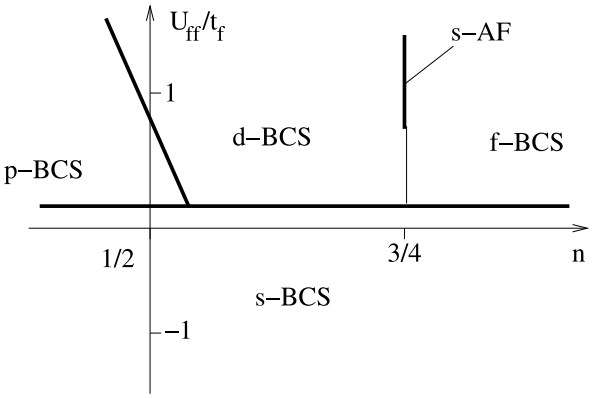

Ultracold Atomic Gases: Novel States of Matter, Figure 16
Phase diagram of a Bose–Fermi mixture on a 2D isotropic triangular lattice. The *vertical axis* corresponds to the interaction strength, U_{ff}/t_f, whereas the *horizontal axis* corresponds to the filling fraction of the fermions per site, *n*. The other parameters are given by $\tilde{V}/t_f = 3$, and $\xi_a = 1$ with $a = 1, 2$

the *s*- and the *p*-wave pairing phase, and the existence of AF order for a nested Fermi surface for large U_{ff}.

However, the competition of pairing orders for positive U_{ff} and intermediate and large filling is much richer, due to the more complex shape of the Fermi surface.

The energy gaps associated with these order parameters can be determined as we did in the previous section [62], by using a 'poor man's' scaling argument. We find for the *s*-wave pairing and the AF order, that they are around $0.1T_F$, where T_F is the Fermi temperature of the system. For most of the exotic phases, we energy gaps of the order of $0.01 - 0.001T_F$.

We now consider a BFM with spin-1/2 fermions on an anisotropic triangular lattice, i. e. with unequal hopping amplitudes, $t_{f(b),1} \neq t_{f(b),2}$. The shape of the FS behaves as follows: For $t_{f,2} > t_{f,1}$, as one increases the chemical potential, the FS first breaks into four arcs at $\mu_f = 2t_{f,1}$, and then breaks into six arcs at $\mu_f = 4t_{f,2} - 2t_{f,1}$, corresponding to the regimes IV–VI, in Fig. 14b and d. For $t_{f,2} < t_{f,1}$ the FS first breaks into two arcs at $\mu_f = 4t_{f,2} - 2t_{f,1}$, and then breaks into six arcs at $\mu_f = 2t_{f,1}$ corresponding to the regimes I–III, in Fig. 14b and c. At the special chemical potential $\mu_f = 2t_{f,1}$ the FS is still nested, but there is only one nesting vector along the direction of the bonds with hopping amplitude $t_{f,1}$. In the absence of the coupling to the BEC the phase diagram has a similar structure as for the isotropic case: *s*-wave pairing for attractive interaction, Fermi liquid behavior for repulsive interaction, with the exception of the nested FS at $\mu_f = 2t_{f,1}$ where one finds AF order (notice that in this case the filling is not 3/4).

When the coupling to the bosons is turned on, one generates an even more complicated competition of

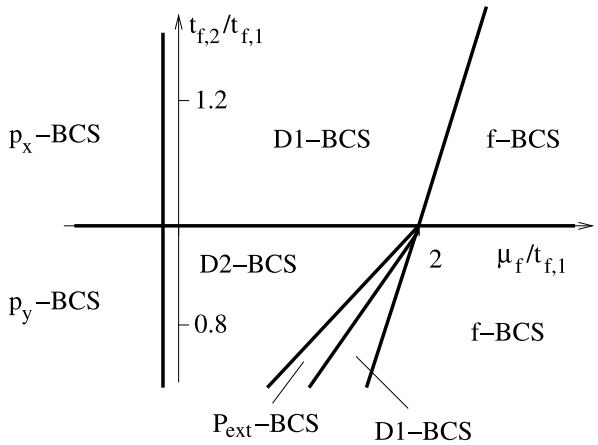

Ultracold Atomic Gases: Novel States of Matter, Figure 17
Phase diagram of a Bose–Fermi mixture on an anisotropic triangular lattice. The *vertical axis* corresponds to the ratio $t_{f,2}/t_{f,1}$, the *horizontal axis* corresponds to the chemical potential μ_f. The other parameters are given by $\tilde{V}/t_f = 3$, $U_{ff}/t_{f,1} = 2$, and $\xi_a = 1$ with $a = 1, 2$)

pairing phases for repulsive U_{ff} in the vicinity of the point $\mu_f = 2t_{f,1}$, as is shown in Fig. 17. Generally, for unequal hopping the degeneracy between D_1 and D_2 in (43), as well as p_x and p_y is lifted: In the regime with $t_{f,2} > t_{f,1}$ ($t_{f,2} < t_{f,1}$), D_1 (D_2) and p_x (p_y) dominate. For $t_{f,2} > t_{f,1}$, in the intermediate regime, in which the FS consists of four disjoined arcs, corresponding to the regime V in Fig. 14, the type of ordering changes from D_1 to f. For $t_{f,2} < t_{f,1}$, the type of pairing also eventually becomes f-wave, but first develops two other types of pairing, in the regime II in Fig. 14. Firstly, one finds an unusual extended p-wave symmetry, which is schematically shown in Fig. 15f. Its wavefunction is of the form:

$$\psi_{P_{ext}} = \begin{cases} \sin^2\theta & -\pi/2 < \theta < \pi/2 \\ -\sin^2\theta & \pi/2 < \theta < 3\pi/2\,. \end{cases} \tag{44}$$

The second type of pairing that appears before the system develops f-wave pairing is D_1. These unusual pairing states are energetically favorable because of the anisotropic shape of the FS. For the regime in which the FS has just barely broken up into two arcs, the order parameter assumes p-wave symmetry and the maxima are located along the y-axis, where the density of states is highest. As the region of open FS widens (see Fig. 15f), this pairing becomes energetically unfavorable, and the system develops D_1-pairing, so that the maxima of the order parameter can again be located near the point of highest density of states. The energy gaps associated with these order parameters are of the same order of magnitude as for the isotropic lattice.

Finally, we consider a BFM with spinless fermions on an isotropic lattice. Due to the absence of s-wave scattering between fermions of the same spin state, there is no direct interaction, that is, $U_{ff} = 0$. Hence, in the absence of bosons, the spinless gas is non-interacting. The boson fluctuations, however, mediate an induced interaction between the fermions. Due to the anti-symmetry of the Cooper pair wavefunction, pairing occurs in an odd angular momentum channel. We find a competition between p and f-wave pairing symmetry. For small to intermediate filling ($n < 0.65$), p-wave pairing dominates. For larger fillings, for which the FS first approaches the shape of a hexagon and then breaks up into six arcs, the system shows f-wave pairing. Since these larger fillings of fermions are typically realized in the center of an atomic trap, this result would suggest a comparatively easy way to create an exotic pairing state experimentally. In contrast to this, a spinless BFM on a square lattice only shows p-wave pairing, since for the quadrangular shape of its FS, channels of higher angular momentum are of no advantage energetically.

The many-body states discussed in this section can be observed through various methods: AF order could be revealed in time-of-flight images and Bragg scattering [91], noise correlations [3,18,29,61] can be used to detect the various pairing phases, laser stirring experiments [72,82] can be used to detect the phase boundary between AF order and pairing. The short-scale CDW fluctuations should give a signature in a photo-association measurement. RF spectroscopy [12] can be used to quantify the gaps of the various phases.

Conclusions

In this article we studied the phase diagrams of various low-dimensional ultra-cold atom systems. In Sect. "One-Dimensional Lattices" we studied atomic mixtures in one dimension, using Luttinger liquid theory. We argued that a Bose–Fermi mixture can be naturally looked at as a Luttinger liquid of polarons, and we discussed the rich phase diagrams of commensurate mixtures. In Sect. "Phase-Locking Transition of Coupled Low-Dimensional Superfluids", we studied the phases of two coupled two-dimensional superfluids at finite temperature. We found that the critical temperature of the phase-locked phase is significantly increased over its bare value, which we propose to use for realizing the Kibble–Zurek mechanism. When interactions between the two superfluids are present, we find additional phases which are partially superfluid, and partially disordered. In Sect. "Bose–Fermi Mixtures in Two-Dimensional Optical Lattices", we used

the powerful method of functional renormalization group equations to determine the phases of Bose–Fermi mixtures in two-dimensional optical lattices. We found an intricate competition of orders, including new types of exotic pairing for triangular lattices. In all these sections, ideas how to probe our predictions were also given.

Bibliography

Primary Literature

1. Albus AP et al (2003) Phys Rev A67:063606
2. Alexandrov AS, Mott SN (1995) Polarons and Bipolarons. World Scientific, Singapore
3. Altman E, Demler E, Lukin M (2004) Probing many-body states of ultracold atoms via noise correlations. Phys Rev A 70:013603
4. Anderson MH, Ensher JR, Matthews MR, Wieman CE, Cornell EA (1995) Observation of Bose–Einstein condensation in a dilute atomic vapor below 200 nanokelvin. Science 269:198
5. Baskaran G (2003) Electronic model for CoO_2 layer based systems: chiral resonating valence bond metal and superconductivity. Phys Rev Lett 91:097003
6. Benfatto L et al (2007) Kosterlitz-Thouless behavior in layered superconductors: the role of the vortex-core energy. arXiv:cond-mat/0609287
7. Bijlsma MJ et al (2000) Phys Rev A 61:053601
8. Cazalilla MA (2004) Phys J B 37:1–47
9. Cazalilla MA, Ho AF (2003) Phys Rev Lett 91:150403
10. Cazalilla MA, Ho AF, Giamarchi T (2005) Phys Rev Lett 95:226402
11. Chaikin PM, Lubensky TC (1995) Principles of condensed matter physics. Cambridge University Press, Cambridge
12. Chin C, Bartenstein M, Altmeyer A, Riendl S, Jochim S, Hecker Denschlag J, Grimm R (2004) Observation of the pairing gap in a strongly interacting Fermi gas. Science 305:1128
13. Choy TP, Galanakis D, Phillips P (2007) Proposed insulating ground state of $Na_{0.5}CoO_2$: Squaring the triangle. Phys Rev B 75:073103
14. Davis KB, Mewes M-O, Andrews MR, NJ van Druten, Durfee DS, Kurn DM, Ketterle W (1995) Bose–Einstein condensation in a gas of sodium atoms. Phys Rev Lett 75:3969
15. de Pue MT, McCormick C, Winoto SL, Oliver S, Weiss DS (1999) Unity occupation of sites in a 3D optical lattice. Phys Rev Lett 82:2262
16. Engelsberg S, Varga BB (1964) Phys Rev 136:A1582
17. Fertig CD et al (2004) Strongly inhibited transport of a 1D Bose gas in a lattice. arXiv:cond-mat/0410491
18. Fölling S, Gerbier F, Widera A, Mandel O, Gericke T, Bloch I (2005) Spatial quantum noise interferometry in expanding ultracold atom clouds. Nature 434:481
19. Fuchs JN, Recati A, Zwerger W (2004) Phys Rev Lett 93:090408
20. Gan JY, Chen Y, Zhang FC (2006) Superconducting pairing symmetries in anisotropic triangular quantum antiferromagnets. Phys Rev B 74:094515
21. Gensemer SD, Jin DS (2001) Phys Rev Lett 87:173201
22. Giamarchi T (2004) Quantum Physics in one dimension. Clarendon Press, Oxford
23. Görlitz A et al (2001) Phys Rev Lett 87:130402
24. Gogolin AO, Nerseyan AA, Tsvelik AM (1998) Bosonization and Strongly Correlated Systems. Cambridge University Press, Cambridge
25. Goldwin J, Inouye S, Olsen ML, Newman B, DePaola BD, Jin DS (2004) Measurement of the interaction strength in a Bose–Fermi mixture with ^{87}Rb and ^{40}K. Phys Rev A 70:021601
26. Greiner M, Bloch I, Mandel O, Hänsch TW, Esslinger T (2001) Exploring phase coherence in a 2D lattice of Bose–Einstein condensates. Phys Rev Lett 87:160405
27. Greiner M, Mandel O, Esslinger T, Hänsch TW, Bloch I (2002) Quantum phase transition from a superfluid to a Mott insulator in a gas of ultracold atoms. Nature 415:39
28. Greiner M, Regal CA, Jin DS (2003) Emergence of a molecular Bose–Einstein condensate from a Fermi gas. Nature 426:537
29. Greiner M, Regal CA, Stewart JT, Jin DS (2005) Probing pair-correlated fermionic atoms through correlations in atom shot noise. Phys Rev Lett 94:110401
30. Hadzibabic Z, Krüger P, Cheneau M, Battelier B, Dalibard JB (2006) Nature 441:1118
31. Hadzibabic Z, Stock S, Battelier B, Bretin V, Dalibard J (2004) Phys Rev Lett 93:180403
32. Haldane FDM (1981) Phys Rev Lett 47:1840
33. Hanaguri T, Lupien C, Kohsaka Y, Lee D-H, Azuma M, Takano M, Takagi H, Davis JC (2004) A 'checkerboard' electronic crystal state in lightly hole-doped $Ca_{2-x}Na_xCuO_2Cl_2$. Nature 430:1001
34. Hoffman JE, Hudson EW, Lang KM, Madhavan V, Eisaki H, Uchida S, Davis SC (2002) A four unit cell periodic pattern of quasi-particle states surrounding vortex cores in $Bi_2Sr_2CaCu_2O_{8+\delta}$. Science 295:466
35. Honerkamp C (2003) Instabilities of interacting electrons on the triangular lattice. Phys Rev B 68:104510
36. Hubbard J (1963) Electron correlations in narrow energy band. Proc Soc R Lond A276:238
37. Isacsson A, Girvin SM (2005) Phys Rev A 72:053604
38. Isacsson A, Cha M-C, Sengupta K, Girvin SM (2005) Phys. Rev. B 72:184507
39. Jaksch D, Bruder C, Cirac JI, Gardiner CW, Zoller P (1998) Cold bosonic atoms in optical lattices. Phys Rev Lett 81:3108
40. Jérome D (1991) The physics of organic superconductors. Science 252:1509
41. Jo G-B, Shin Y, Will S, Pasquini TA, Saba M, Ketterle W, Pritchard DE, Vengalattore M, Prentiss M (2007) Long phase coherence time and number squeezing of two Bose–Einstein condensates on an atom chip. arXiv:cond-mat/0608585
42. Jochim S, Bartenstein M, Altmeyer A, Hendl G, Riedl S, Chin C, Denschlag JH, Grimm R (2003) Science 302:2101
43. Jones K, Julienne P, Lett P, Phillips WD, Tiesinga E, Williams C (1996) Measurement of the atomis Na(3P) lifetime and of retardation in the interaction between two atoms bound in a molecule. Europhys Lett 35:85
44. Kane C, Fisher MPA (1992) Phys Rev Lett 68:1220
45. Kibble TWB (1976) Phys J A 9:1387
46. Kino H, Fukuyama H (1995) Electronic states of conducting organic κ-(BEDT-TTF)$_2$. Phys XJ Soc Jpn 64:2726
47. Kinoshita T, Wenger T, Weiss DS (2004) Science 305:1125
48. Kitaev AY (2000) Unpaired Majorana fermions in quantum wires. arXiv:cond-mat/0010440
49. Klironomos FD, Tsai S-W (2006) Phonon-mediated tuning of instabilities in the Hubbard model at half-filling. Phys Rev B 74:205109
50. Klironomos FD, Tsai S-W (2007) Pairing and density-wave

phases in Boson–Fermion mixtures at fixed filling. arXiv:cond-mat/0702660
51. Kogut JB (1979) Rev Mod Phys 51:659
52. Köhl M, Moritz H, Stöferle T, Günter, Esslinger T (2005) Fermionic atoms in a three dimensional optical lattice: observing Fermi surfaces, dynamics, and interactions. Phys Rev Lett 94:080403
53. Kohn W, Luttinger JM (1965) New mechanism for superconductivity. Phys Rev Lett 15:524
54. Lee S-S, Lee PA (2005) $U(1)$ Gauge theory of the Hubbard model: spin liquid states and possible application to κ-(BEDT-TTF)$_2$Cu$_2$(CN)$_3$. Phys Rev Lett 95:036403
55. Lett PD, Julienne PS, Phillips WD (1995) Photoassociative spectroscopy of laser-cooled atoms. Annu Rev Phys Chem 46:423
56. Maddaloni P et al (2000) Phys Rev Lett 85:2413
57. Mahan GD (1990) Many-particle physics. Plenum Press, New York
58. Mandel O, Greiner M, Widera A, Rom T, Hänsch TW, Bloch I (2003) Phys Rev Lett 91:010407
59. Mandel O, Greiner M, Widera A, Rom T, Hänsch TW, Bloch I (2003) Nature 425:937
60. Mathey L, Wang D-W, Hofstetter W, Lukin MD, Demler E (2004) Luttinger liquid of polarons in one-dimensional boson–fermion mixtures. Phys Rev Lett 93:120404
61. Mathey L, Altman E, Vishwanath A (2008) Noise correlations in one-dimensional systems of ultra-cold fermions. Phys Rev Lett 100:240401
62. Mathey L, Tsai S-W, Castro Neto AH (2006) Competing types of order in two-dimensional Bose–Fermi mixtures. Phys Rev Lett 97:030601
63. Mathey L (2007) Commensurate mixtures of ultra-cold atoms in one dimension. Phys Rev B 75:144510
64. Mathey L, Wang D-W (2007) Phase diagrams of one-dimensional Bose–Fermi mixtures of ultra-cold atoms. Phys Rev A 75:013612
65. Mathey L, Tsai S-W, Castro Neto AH (2007) Exotic superconducting phases of ultracold atom mixtures on triangular lattices. Phys Rev B 75:144510
66. Mathey L, Polkovnikov A, Castro Neto AH (2008) Phase-locking transition of coupled low-dimensional superfluids. Europhys Lett 81:10008
67. Mckenzie RH (1997) Similarities between organic and cuprate superconductors. Science 278:820
68. Modugno G, Roati G, Riboli F, Ferlaino F, Brech RJ, Inguscio M (2002) Collapse of a degenerate Fermi gas. Science 297:2240
69. Modugno G, Ferlaino F, Heidemann R, Roati G, Inguscio M (2003) Production of a Fermi gas of atoms in an optical lattice. Phys Rev A 68:011601
70. Moritz H, Stöferle T, Günter K, Köhl M, Esslinger T (2005) Phys Rev Lett 94:210401
71. Mukhopadhyay R, Kane CL, Lubensky TC (2001) Phys Rev B 64:045120. CSO'Hern, Lubensky TC, Toner J (1999) Phys Rev Lett 83:2745
72. Onofrio R, Raman C, Vogels JM, Abo-Shaeer JR, Chikkatur AP, Ketterle W (1999) Observation of superfluid flow in a Bose–Einstein condensed gas. Phys Rev Lett 85:2228
73. Ospelkaus C, Ospelkaus S, Sengstock K, Bongs K (2006) Interaction-driven dynamics of ^{40}K-^{87}Rb fermion–boson gas mixtures in the large-particle-number limit. Phys Rev Lett 96:020401
74. Ospelkaus C, Ospelkaus S, Humbert L, Ernst P, Sengstock K, Bongs K (2006) Ultracold heteronuclear molecules in a 3optical D lattice. Phys Rev Lett 97:120402
75. Ospelkaus S, Ospelkaus C, Wille O, Succo M, Ernst P, Sengstock K, Bongs K (2006) Localization of bosonic atoms by fermionic impurities in a three-dimensional optical lattice. Phys Rev Lett 96:180403
76. Ospelkaus S, Ospelkaus C, Humbert L, Sengstock K, Bongs K (2006) Tuning of heteronuclear interactions in a degenerate Fermi–Bose mixture. Phys Rev Lett 97:120403
77. Paredes B, Widera A, Murg V, Mandel O, Foelling S, Cirac I, Shlyapnikov GV, Hänsch TW, Bloch I (2004) Nature 429:277
78. Partridge GB, et al (2005) Pairing and phase separation in a polarized fermi gas. arXiv:cond-mat/0511752
79. Phillips WD, Ekstrom C, Goldin W, Rolston SL (1996) Laser cooling of neutral atoms for frequency standards. In: Bergquist J (ed) Symposium on Frequency Standards and Metrology. World Scientific, Singapore, pp 5–10
80. Polkovnikov A, Altman E, Demler E (2006) Proc Natl Acad Sci USA 103:6125
81. Polkovnikov A (2005) Phys Rev B 72:161201(R). Zurek WH, Dorner U, Zoller P (2005) Phys Rev Lett 95:105701. Dziarmaga J (2005) Phys Rev Lett 95:245701
82. Raman C, Köhl M, Onofrio R, Durfee DS, Kuklewicz CE, Hadzibabic Z, Ketterle W (1999) Evidence for a critical velocity in a Bose–Einstein condensed gas. Phys Rev Lett 83:2502
83. Recati A, Fedichev PO, Zwerger W, Zoller P (2003) Phys Rev Lett 90:020401
84. Sadler LE et al (2006) Nature 443:312
85. Scherer DR et al (2006) Vortex formation by interference of multiple trapped Bose–Einstein condensates. arXiv:cond-mat/0610187
86. Schulz HJ (1987) Superconductivity and antiferromagnetism in the two-dimensional Hubbard model – scaling theory. Europhys Lett 4:609
87. Schumm T, Hofferberth S, Andersson LM, Wildermuth S, Groth S, Bar-Joseph I, Schmiedmayer J, Krüger P (2005) Nature Phys 1:57
88. Sengupta P, Pryadko LP (2005) Quantum degenerate Bose-Fermi mixtures on 1D optical lattices. arXiv:cond-mat/0512241
89. Shankar R (1994) Renormalization group approach to interacting fermions. Rev Mod Phys 68:129
90. Solyom J (1979) Adv. Phys 28:201
91. Stenger J, Inouye S, Chikkatur AP, Stamper-Kurn DM, Pritchard DE, Ketterle W (1999) Bragg spectroscopy of a Bose–Einstein condensate. Phys Rev Lett 82:4569
92. Stöferle T, Moritz H, Schori C, Köhl M, Esslinger T (2004) Transition from a strongly interacting 1D superfluid to a Mott insulator. Phys Rev Lett 92:130403
93. Takada K, Sakurai H, Takayama-Muromachi E, Izumi F, Dilanian RA, Sasaki T (2003) Nature 422:53
94. Tam K-M, Tsai S-W, Campbell DK, Castro Neto AH (2007) Retardation effects in the Holstein–Hubbard chain at half filling. Phys Rev B 75:161103
95. Tsai S-W, Marston JB (2001) κ-(BEDTTTF)$_2$X organic crystals: Superconducting versus anti-ferromagnetic instabilities in the Hubbard model on an anisotropic triangular lattice. Can. Phys J 79:1463
96. Tsai S-W, Castro Neto AH, Shankar R, Campbell DK (2005)

Renormalization group approach to strong-coupled superconductors. Phys Rev B 72:054531
97. Tsai S-W, Castro Neto AH, Shankar R, Campbell DK (2005) Renormalization group approach to superconductivity: from weak to strong electron–phonon coupling. Philos Mag 86:2631
98. Tsuchiya S, Griffin A (2003) Damping of Bogoliubov Excitations in Optical Lattices. arXiv:cond-mat/0311321
99. Vershinin M, Misra S, Ono S, Abe Y, Ando Y, Yazdani A (2004) Local ordering in the pseudogap state of the high-T_c superconductor $Bi_2Sr_2CaCu_2O_{8+\delta}$. Science 303:1995
100. Vichi L et al (1999) Phys Rev A 60:4734
101. Voit J, Schulz HJ (1987) Phys Rev B 36:968. ibid 37:10068 (1988)
102. Voit J (1995) Rep Prog Phys 58:977
103. Vojta M, Dagotto E (1999) Indications of unconventional superconductivity in doped and undoped triangular antiferromagnets. Phys Rev B 59:713
104. Wang D-W, Lukin MD, Demler E (2005) Engineering superfluidity in Bose–Fermi mixtures of ultracold atoms. Phys Rev A 72:(R)051604
105. Withers RL, Wilson JA (1986) An examination of the formation and characteristics of charge-density waves in inorganic materials with special reference to the two- and one-dimensional transition-metal chalcogenides. Phys J C 19:4809
106. Zanchi D, Schulz HJ (2000) Weakly correlated electrons on a square lattice: Renormalization-group theory. Phys Rev B 61:13609
107. Zhou A, Wang Z (2006) Charge and spin order on the triangular lattice – Na_xCoO_2 at $x = 0.5$. arXiv:cond-mat/0608068
108. Zurek WH (1985) Nature 317:505
109. Zwierlein MW, Stan CA, Schunck CH, Raupach SMF, Gupta S, Hadzibabic Z, Ketterle W (2003) Phys Rev Lett 91:250401
110. Zwierlein M et al (2005) Fermionic Superfluidity with Imbalanced Spin Populations and the Quantum Phase Transition to the Normal State. arXiv:cond-mat/0511197

Books and Reviews

Bloch I, Dalibard J, Zwerger W (2007) Many-body physics with ultracold gases. arXiv:0704.3011v1
Leggett AJ (2001) Bose–Einstein condensation in the alkali gases: Some fundamental concepts. Rev Mod Phys 73:307

Unconventional Computing, Introduction to

Andrew Adamatzky
University of the West of England, Bristol, UK

Unconventional Computing deals with computing and information processing derived from or implemented in physical, chemical and biological systems. As Toffoli said once "a computing scheme that today viewed as unconventional may well be so because its time hasn't come yet – or is already gone" [1].

Mechanical and analogue computers are characteristics representatives of such computing schemes. When mechanical computation concerned the unconventional computers can be traced back to many centuries old harmonic analyzers and synthesizers, Napier's bones, and Pascal's wheeled calculator, and they are revived recently in self-assembling devices and molecular machines (see ► Mechanical Computing: The Computational Complexity of Physical Devices).

In 1876 Lord Kelvin envisaged that a computation by directly exploiting law of Nature [2] and invented differential analyzer. Ideas of analog computation emerged, flourished, almost deceased by 1980s and then were resurrected in 1990s in ever growing field of computing based on continuous representations by means of continuous processes (see ► Analog Computation). This also branched into theory of computing with light and related continuous state machines (see ► Optical Computing).

Chemical and molecular media and substrates are now widely explored as prototypes of future non-standard computing devices. They can be classified into the following groups. First, models based interpretations of chemical reactions as computing processes, including autocatalytic polymer chemistries, chemistries inspired by Turing machines, lattice molecular systems (see ► Artificial Chemistry). Second, computers which utilize replication of macro-molecules. Here DNA and other polymer molecules are basic elements of computing devices, they solve Hamiltonian path problem, simulate implementation of chess problem, and binary arithmetic (see ► DNA Computing) and also molecular finite state machines, therapeutic and diagnostic automata (see ► Molecular Automata). Chemical computers of the third type compute by interactions between propagating diffusive and phase wave fronts in chemical media.

When waves, e. g. in excitable chemical medium, are confined to geometrical restrictions of channels and junctions one implement diodes, frequency transformers, logical gates and chemical sensors (see ► Computing in Geometrical Constrained Excitable Chemical Systems). Waves propagating in unconstrained, or free-space, chemical systems are also proved to be capable for implementation of logical circuits, and robot control and computational geometry (see ► Reaction-Diffusion Computing).

While going into domains of physics we must highlight two more types of computing devices: soliton-based computers and quantum-computers. The soliton-based devices compute by colliding self-preserved solitary waves, which change their state or trajectories in result of the collision and thus are capable of implementing basic logical gates (see ► Computing with Solitons). The quantum

computers are based on the laws of quantum mechanics, such as superposition and entanglement, interpreted in terms of logical operations (see ► Quantum Computing).

Consequently not only the laws of physics can be exploited in computation but reversely computing and computers themselves can be analyzed in terms of physical laws, e. g. computer equivalents of the first and second laws of thermodynamics, thermodynamics of digital and analog computers (see ► Thermodynamics of Computation). Fundamental microscopic physical properties of natural systems are reflected in the models of reversible computing (see ► Reversible Computing) which also forms the underlying paradigm of thermodynamically efficient computation.

The next range of unconventional computers in our – very conditional, and somewhat illusory – classification of non-standard computing devices would be those relying on biological substrates (in addition to DNA computing). They are bacterial and cellular computers (see ► Bacterial Computing, ► Cellular Computing) employing biochemistry, genetic circuitry and inter- and intra-cellular communication for transformation of information, immune-computers, which mimic behavior of immune system (see ► Immunecomputing), and computing based on comparmentalization of a computing medium with membranes (see ► Membrane Computing).

There is also a family of computers classified not by implementation substrate but some other characteristics. For instance, nano-computers are based on technology employing devices and wires with feature sizes in the order of a few nano-meters (see ► Nanocomputers). Amorphous computers are built of a collection of computational particles, with no a priori knowledge of their positions or orientations, dispersed irregularly on a surface or throughout a volume (see ► Amorphous Computing).

Evolving unconventional computers can learn, and programmed, to reconfigure their physical structure to meet certain computational demands, so far we have examples of evolving computing abilities of liquid crystals, conducting and electro-activated polymers, voltage controlled colloids (see ► Evolution in Materio).

Biological, chemical and other 'wet' substrates are indeed exciting prototypes for future computing devices. However, there is still a great potential for further enhancement and miniaturization of hardware systems, e. g. hardware analogs of reaction-diffusion systems, digital CMOS quasi-chemical chips, devices based on minority-carrier transport in semiconductors, networks of single-electron oscillators (see ► Unconventional Computing, Novel Hardware for).

Bibliography

1. Toffoli T (1998) Programmable matter methods. Future Gener Comput Syst 16:187–201
2. Thomson W (1876) On an instrument for calculating integral of the product of two given functions (Lord Kelvin). Proc R Soc 24:266–275

Unconventional Computing, Novel Hardware for

Tetsuya Asai
Hokkaido University, Sapporo, Japan

Article Outline

Glossary
Definition of the Subject
Introduction
Constructing Electrical Analog of Reaction-Diffusion Systems
Digital CMOS Reaction-Diffusion Chips
Analog CMOS Reaction-Diffusion Chip
Reaction-Diffusion Computing Devices Based on Minority-Carrier Transport in Semiconductors
Single-Electron Reaction-Diffusion System
Collision-Based RD Computers
Future Directions
Bibliography

Glossary

Analog circuit An electronic circuit that operates with currents and voltages that vary continuously with time and have no abrupt transitions between levels. Since most physical quantities, e. g., velocity and temperature, vary continuously, as does audio, an analog circuit provides the best means of representing them.

Current mirror A circuit that copies single input current to single (or multiple) output nodes. Two types of current mirrors exist; nMOS for current sinks and pMOS for current sources. Combining both types of current mirrors, one can invert a direction of currents; e. g., sink to source or source to sink.

Digital circuit An electronic circuit that can take on only a finite number of states. Binary (two-state) digital circuits are the most common. The two possible states of a binary circuit are represented by the binary digits, or bits, 0 and 1. The simplest forms of digital circuits are built from logic gates, the building blocks of the digital computer.

Diode A device that allows current flow only in one direction. Chemical diode allows for propagation of chemical waves only in one direction.

Flip-flop circuit A synchronous bistable device where the output changes state only when the clock input is triggered. That is, changes in the output occur in synchronization with the clock.

Floating-gate transistor A device consisting of a control gate, floating gate and the thin oxide layer; when floating gate is given an electrical charge, the charge is trapped in the insulating think oxide layer. The transistors are used as non-volatile storage devices because they store electrical charge for a long time without powering.

LSI, large-scale integration circuit An electronic circuit built on a semiconductor substrate, usually one of single-crystal silicon. It contains from 100 to 1000 transistors. Some LSI circuits are analog devices; an operational amplifier is an example. Other LSI circuits, such as the microprocessors used in computers, are digital devices.

Minority-carrier transport A physical phenomenon in forwardly-biased semiconductor *p*-*n* junctions. Minority carriers are generated in both area of *p*- and *n*-type semiconductors. For *p*-type semiconductors, the minority carriers are electrons, while they are holes in *n*-type semiconductors. Once minority carriers are generated, they diffuse among the semiconductor and finally disappears by the recombination of electrons and holes.

nMOS FET Abbreviation of *n*-type metal-oxide-semiconductor field effect transistor, where semiconductor is negatively charged so the transistors are controlled by movement of electrons; these transistors have three modes of operation: cut-off, triode, and saturation (active).

pMOS FET A device which works by analogy to nMOS FET but the transistors are moved on and off by movement of electron vacancies.

Single-electron circuit An electrical circuit that is functionally constructed by controlling movements of single electrons. Single-electron circuit consists of tunneling junctions and electrons are controlled by using physical phenomena called the Coulomb blockade.

Definition of the Subject

Natural systems give us examples of amorphous, unstructured devices, capable of fault-tolerant information processing, particularly with regard to the massive parallel spatial problems that digital processors have difficulty with. For example, reaction-diffusion (RD) chemical systems have the unique ability to efficiently solve combinatorial problems with natural parallelism [5]. In liquid-phase parallel RD processors (RD chemical computers), both the data and the results of the computation are encoded as concentration profiles of the reagents. The computation is performed via the spreading and interaction of the wave fronts. In experimental chemical processors, data are represented by local disturbances in the concentrations, and computation is accomplished via the interaction of waves caused by the local disturbances.

The RD chemical computers operate in parallel since the chemical medium's micro-volumes update their states simultaneously, and the molecules diffuse and react in parallel. We see a similar parallelism in cellular automata (CA). Various RD systems can be modeled in terms of CA, including the Belousov–Zhabotinsky (BZ) reaction [5,30], the Turing system [62], a precipitating BZ system for computation of Voronoi diagram [1,2], and so on. A two-dimensional CA is particularly well suited for the coming generation of massively parallel machines, in which a very large number of separate processors act in parallel. If an elemental processor in the CA is constructed from a smart processor and photosensor, various CA algorithms can easily be used to develop intelligent image sensors.

Implementing RD systems in hardware has several advantages. Hardware RD systems are very useful in simulating RD phenomena, even if the phenomena never occur in nature. This implies that a hardware system is a possible candidate for developing an artificial RD system that is superior to a natural system. For instance, hardware RD systems can operate at much faster speeds than actual RD systems. The velocity of chemical waves in a BZ reaction is $O(10^{-2})$ m/s [56], while that of a hardware RD system will be over a million times faster than that of the BZ reaction, independent of system size. This property is useful for developers of RD applications because every RD application benefits from high speed operations. These properties encouraged us to develop novel hardware for unconventional (RD-based) computing.

Introduction

This chapter presents an overview of the semiconductor implementation of reaction-diffusion (RD) computers in large-scale integrated (LSI) circuits for unconventional computing. There, we see how to model RD processes in LSI circuits and discuss several designs of RD digital chips, based on cellular-automaton models of RD and excitable systems. Feasibility of a RD digital chip is demonstrated in the construction of a Voronoi diagram and decomposition

of images. The chapter concludes with analogue RD chips, where closer to physical reality nonlinear characteristics of chemical systems are employed. We see designs of RD chips based on Oregonator, Turing and so on. Moreover, functionality of analogue RD chips in feature extraction and fingerprint reconstruction tasks are exemplified.

A RD chip consists of i) reaction circuits that emulate elementary interactions between neurons (or chemical substances) and ii) diffusion devices that imitate synapses (or chemical diffusion of the substances). RD chips were mostly designed by digital, analog, or mixed-signal complementally metal-oxide-semiconductor (CMOS) circuits of cellular neural networks (CNNs) or cellular automata (CA). Electrical cell circuits were designed to implement several CA and CNN models of RD systems [8,17,20,38], as well as fundamental RD equations [18,21,24,33,50]. Each cell is arranged on a 2D square or hexagonal grid and is connected with adjacent cells through coupling devices that transmit a cell's state to its neighboring cells, as in conventional CAs. For instance, an analog-digital hybrid RD chip [17] was designed for emulating a conventional CA model for BZ reactions [30]. A precipitating BZ system for computation of Voronoi diagram [1,2] was also implemented on an analog-digital hybrid RD chip [20]. A full-digital RD processor [38] was also designed on the basis of a multiple-valued CA model, called *excitable lattices* [5]. Furthermore, a RD CA processor for complex image processing has been proposed [19]. It performs quadrilateral-object extraction based on serial and parallel CA algorithms. An analog cell circuit was also designed to be equivalent to spatial-discrete Turing RD systems [24]. A full-analog RD chip that emulates BZ reactions has also been designed and fabricated [21]. Furthermore, blueprints of non-CMOS RD chips have been designed; i. e., a RD device based on minority-carrier transport in semiconductor devices [18]. In the following sections, we see how to construct an artificial RD system on *solid-state media*, and to develop some applications using the solid-state RD system that could cope with conventional digital computers.

Constructing Electrical Analog of Reaction-Diffusion Systems

The behavior of RD systems, or the spatiotemporal patterns of chemical concentration, can be expressed by the reaction-diffusion equation, a partial differential equation with chemical concentrations as variables:

$$\frac{\partial \mathbf{u}}{\partial t} = \mathbf{f}(\mathbf{u}) + D\Delta\mathbf{u}\,, \quad [\mathbf{u} = (u_1, u_2, u_3, \ldots)]\,, \qquad (1)$$

where t is time, $\mathbf{u}$ is the vector of chemical concentrations,

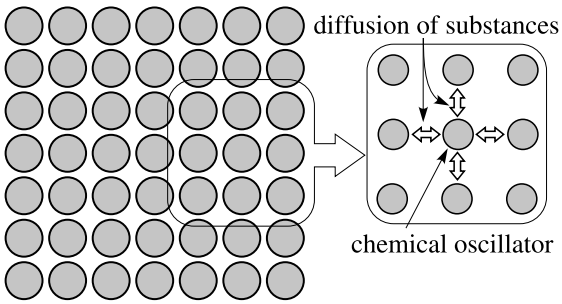

Unconventional Computing, Novel Hardware for, Figure 1
Simplified model of RD systems, consisting of many chemical oscillators. Each oscillator has variables corresponding to chemical concentrations $u_1, u_2, u_3, \ldots$ in Eq. (1) and interacts with its neighbors through diffusion of substances

u_i is the concentration of the ith substance, and D is the diagonal matrix of diffusion coefficients. Nonlinear function $\mathbf{f}(\mathbf{u})$ is the reaction term that represents the reaction kinetics of the system. Spatial derivative $D\Delta\mathbf{u}$ is the diffusion term that represents the change of $\mathbf{u}$ due to the diffusion of the substance. A greater number of variables results in more complex dynamics and a more complicated dissipative structure. A simple reaction-diffusion system with few variables, however, will still exhibit dynamics similar to biological activity.

An RD system can be considered an aggregate of coupled chemical oscillators, or a chemical cellular automaton, as described in Fig. 1. Each oscillator represents the local reaction of chemical substances and generates nonlinear dynamics $\mathrm{d}\mathbf{u}/\mathrm{d}t = \mathbf{f}(\mathbf{u})$ that corresponds to reaction kinetics in Eq. (1). The oscillator interacts with its neighbors through nonlocal diffusion of substances; this corresponds to the diffusion term in Eq. (1) and produces dynamics $\mathrm{d}\mathbf{u}/\mathrm{d}t = D\Delta\mathbf{u}$. Because of diffusion, all oscillators correlate with one another to generate synchronization and entrainment. Consequently, the system as a whole produces orderly dissipative structures on a macroscopic level. The size of each oscillator, or the size of the local space in which chemical concentrations are roughly uniform, depends on the diffusion coefficients and reaction velocities in the system. It is several micrometers in diameter in many liquid RD systems; therefore, even a tiny RD system in a test tube contains millions of oscillators.

An electrical analog of RD systems can be created by using electrical oscillation circuits instead of chemical oscillators and coupling these circuits with one another in a way that imitates diffusion. Variables are the electrical potential of nodes in the oscillation circuits in this electrical RD system. The system will produce electrical dissipative structures, i. e., orderly spatiotemporal patterns of node potentials, under appropriate conditions.

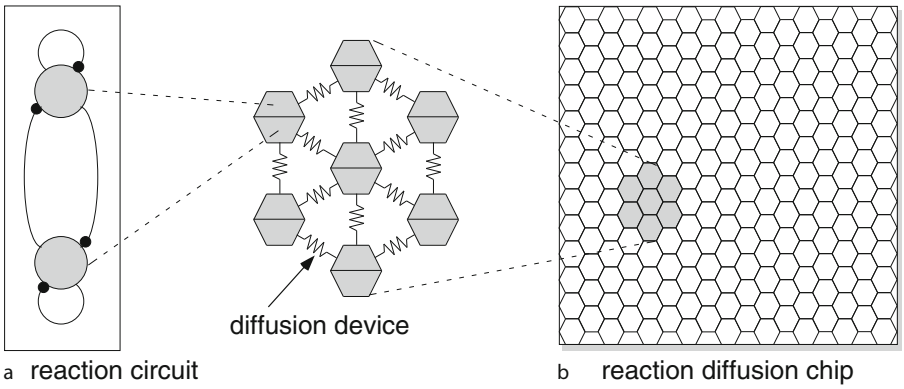

Unconventional Computing, Novel Hardware for, Figure 2
Basic construction of RD chip

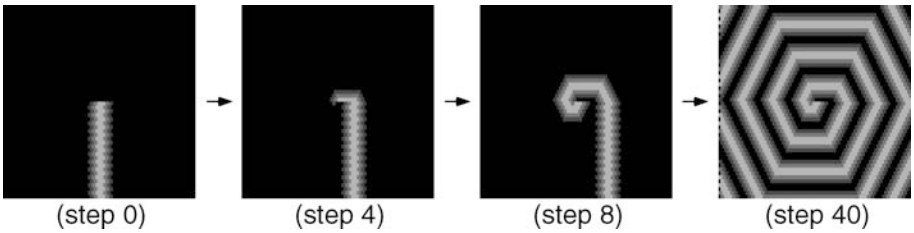

Unconventional Computing, Novel Hardware for, Figure 3
Excitatory modelock operations of RD circuit [17]

The key to building an electrical RD system is to integrate a large number of oscillation circuits on a chip with coupling subcircuits. A large arrangement of oscillators (e. g., 1000 × 1000 or more) is needed to generate complex, varied dissipative structures as observed in chemical RD systems. Oscillators constructed by micro- or nanoscale circuits are thus useful to achieve such large scale integration. Such circuits can generate nonlinear oscillation through a simple circuit structure, so it can effectively be used in producing small oscillators for electrical RD systems.

Digital CMOS Reaction-Diffusion Chips

A Reaction-Diffusion Circuit Based on Cellular-Automaton Processing Emulating the Belousov–Zhabotinsky Reaction

The Belousov–Zhabotinsky (BZ) reaction provides us important clues to control 2D phase-lagged stable synchronous patterns in excitable medium. Because of the difficulty in computing RD systems in large systems using conventional digital processors a cellular-automaton (CA) circuit that emulates the BZ reaction was proposed [17]. In the circuit, a two-dimensional array of parallel processing cells, shown in Fig. 2, is responsible for the emulation, and its operation rate is independent of the system size. The operations of the CA circuit were demonstrated by using a simulation program with integrated circuit emphasis (SPICE). In the circuit's initial state, cells adjacent to inactive cells were in a refractory period (step 0 in Fig. 3). The inactive cells adjacent to the white bar in Fig. 3 were suppressed by adjacent cells in the refractory period (cells in the white bar). The inactive cells then entered an active, inactive, or refractory period, depending on the degree of the refractory condition. When the inactive cells were in an active or inactive period, the tip of the bar rotated inward (step 4 to 8 in Fig. 3), resulting in the generation of the modelock (spiral patterns) typically observed in the BZ reaction (step 40 in Fig. 3). A hexagonal distortion of the propagating waves was generated by interactions between adjacent cells. These results indicated that the RD chip could be easily integrated into existing digital systems and can be used to clarify RD systems, aiming at developing further novel applications.

Reaction-Diffusion Chip Implementing Excitable Lattices with Multiple-Valued Cellular Automata

A RD chip was fabricated based on a multiple-valued CA model of excitable lattices [38]. The experiments

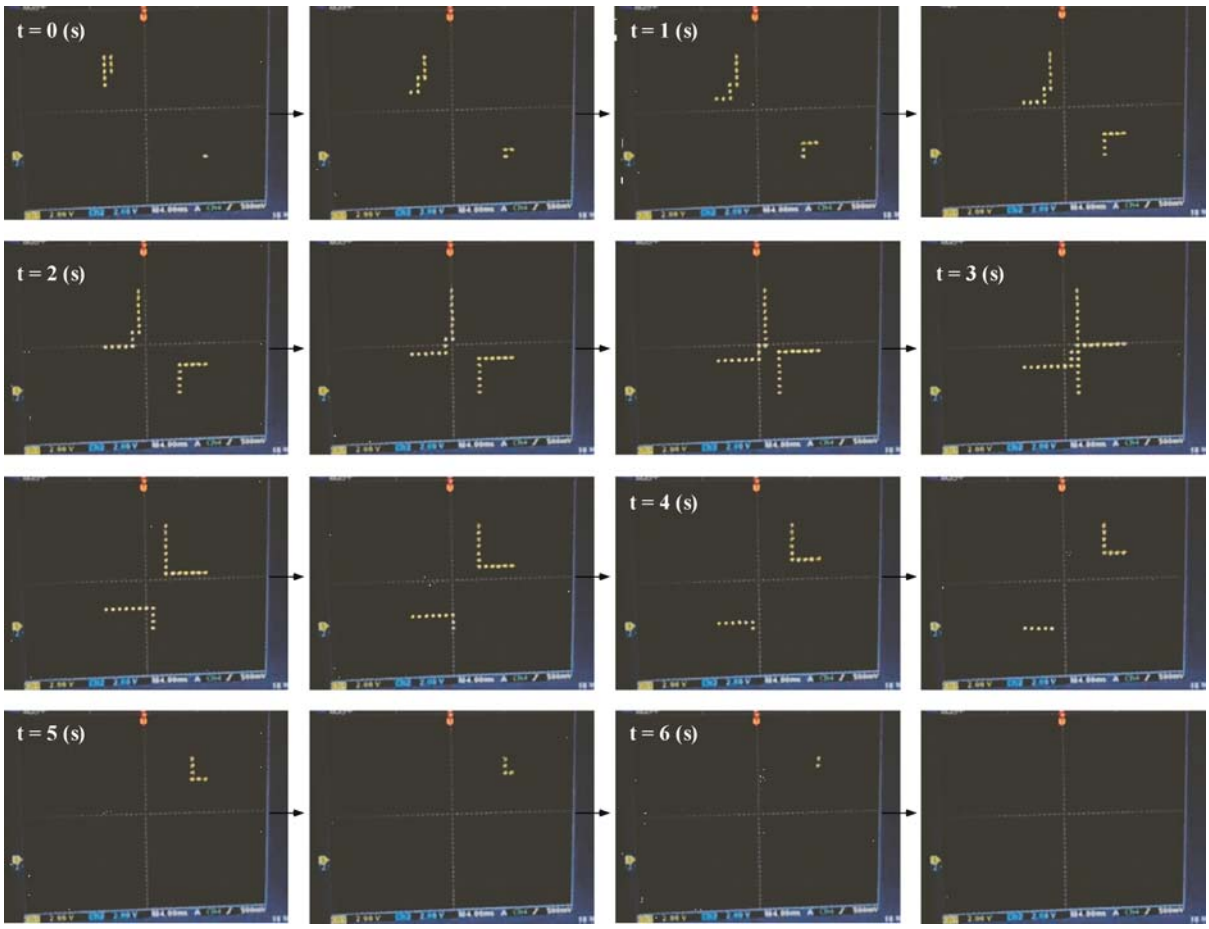

Unconventional Computing, Novel Hardware for, Figure 4
Snapshots of recorded movie obtained from fabricated RD chip [38]

confirmed the expected operations, i. e., excitable wave propagation and annihilation. One could obtain a binary stream from the common output wire by selecting each cell sequentially. Using a conventional displaying technique, the binary stream was reconstructed on a 2-D display. Figure 4 shows snapshots taken from the recorded movie. Each dot represents an excitatory cell where EXC is logical "1". In the experiment, the supply voltage were set at 5 V, and the system clock was set at low frequency (2.5 Hz) so that "very-slow" spatiotemporal activities could be observed visually (the low frequency was used only for the visualization, and was not the upper limit of the circuit operation). Pin-spot lights were applied to several cells at top-left and bottom right corners of the chip. The circuit exhibited the expected results; i. e., two excitable waves of excited cells triggered by the corner cells propagated toward the center and disappeared when they collided. This result suggests that if we use a more microscopic process and a large number of cells were implemented, we would observe the same complex (BZ-like) patterns, as observed in the original excitable lattices [5].

This chip can operate much faster than real chemical RD systems, even when the system clock frequency is $O(1)$ Hz, and is much easier to use in various experimental environments. Therefore, the chip should encourage RD application developers who use such properties of excitable waves to develop unconventional computing schemes, e. g., chemical image processing, pattern recognition, path planing, and robot navigation.

Silicon Implementation of a Chemical Reaction-Diffusion Processor for Computation of Voronoi Diagram

RD chemical systems are known to realize sensible computation when both data and results of the computation

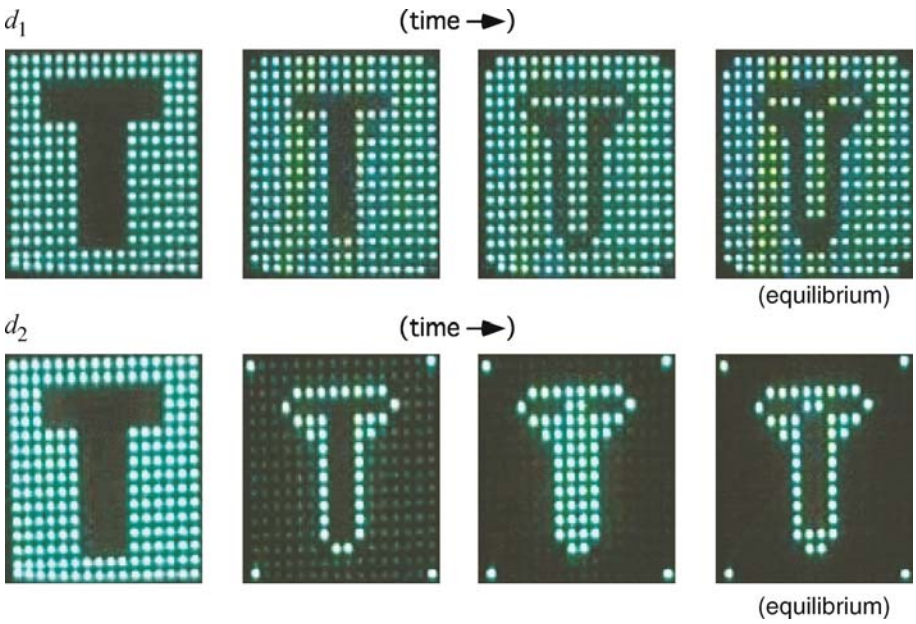

Unconventional Computing, Novel Hardware for, Figure 5
Skeleton operation with "T" shape [20]

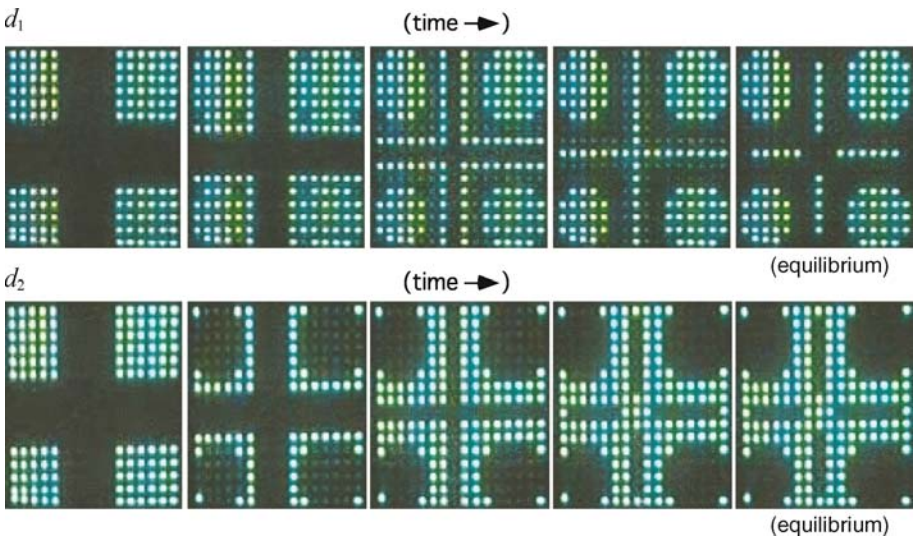

Unconventional Computing, Novel Hardware for, Figure 6
Skeleton operation with "+" shape [20]

are encoded in concentration profiles of chemical species; the computation is implemented via spreading and interaction of either diffusive or phase waves, while a silicon RD chip is an electronic analog of the chemical RD systems.

A prototype RD chip implementing a chemical RD processor for a well-known NP-complete problem of computational geometry – computation of a Voronoi diagram was fabricated. Here we see experimental results for fabricated RD chips and compare the accuracy of information processing in silicon analogs of RD processors and their experimental 'wetware' prototypes [20]. Figures 5 and 6 show examples of skeleton operation of a T and '+' shaped images. As initial images, a glass mask was prepared where 'T' and '+' areas were exactly masked. Therefore, cells under the 'T' and '+' areas are initially resting and the rest are initially excited. At its equilibrium, skeletons of 'T' and '+' were successfully obtained.

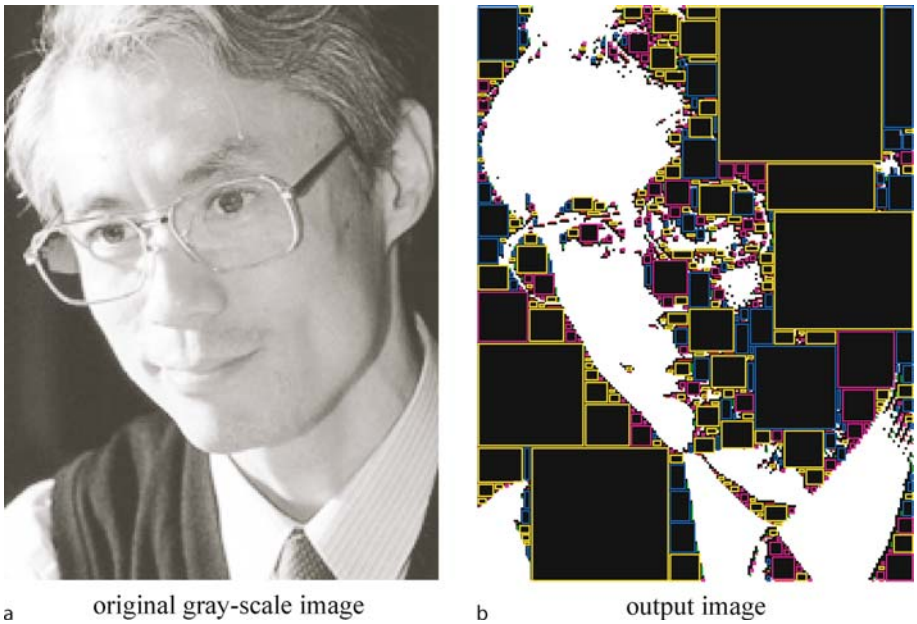

Unconventional Computing, Novel Hardware for, Figure 7
Simulation results for 181 × 238 image; a original image, b mixed image of quantized image given to CA LSI and detected quadrilateral objects [19]

A Quadrilateral-Object Composer for Binary Images with Reaction-Diffusion Cellular Automata

A CA LSI architecture that extracted quadrilateral objects from binary images was proposed in [19] with a serial combination of parallel CA algorithms, based on RD chemical systems model. Each cell in the CA was implemented by a simple digital circuit called an elemental processor. The CA LSI can be constructed by a large number of elemental processors and their controllers operating in serial and parallel.

Figure 7a demonstrates object extraction for a natural image. The image was quantized, and given to the CA LSI. Figure 7b show the results. The maximum boxes were correctly detected in order, as predicted. The input bitmap image that consisted of 181 × 238 pixels was decomposed of 1020 quadrilateral objects. The bitmap image occupied 43,078 bits (5385 bytes) on memory, while the objects used 4080 bytes (8-bit address of 4 corners × 1020 boxes). The important thing is not discussing the compression rate between bitmap images and extracted objects, but that bitmap images were represented by small number of vector objects, which facilitates picture drawing in terms of the drawing speed if we have variable box window.

One of the most important application targets for the proposed chip is a computer-aided design (CAD) system for LSIs. Conventional LSI CAD tools use polygons to represent device structures. However, recent LSIs include not only polygon patterns but also graphical patterns, consisting of large number of dots, usually imported from image files such as JPEGs, to implement complex analog structures. In the mask manufacturing process, exposing a large number of dot patterns is quite a time-consuming task. Recently, electron beam (EB) lithography systems that can expose wide areas through a quadrilateral window have been produced on a commercial basis. The proposed LSI can produce efficient stream files from binary image files that can easily be handled by the new EB systems, by developing simple software that converts the box format, produced by the proposed LSI, to a conventional stream format.

Analog CMOS Reaction-Diffusion Chip

Analog Reaction-Diffusion Chip with Hardware Oregonator Model

Silicon devices that imitate the autocatalytic and dissipative phenomena of RD systems were developed [21]. Numerical simulations and experimental results revealed that an RD device could successfully produce concentric and spiral waves in the same way as natural RD systems. These results encouraged us to develop new applications based on natural RD phenomena using hardware RD devices.

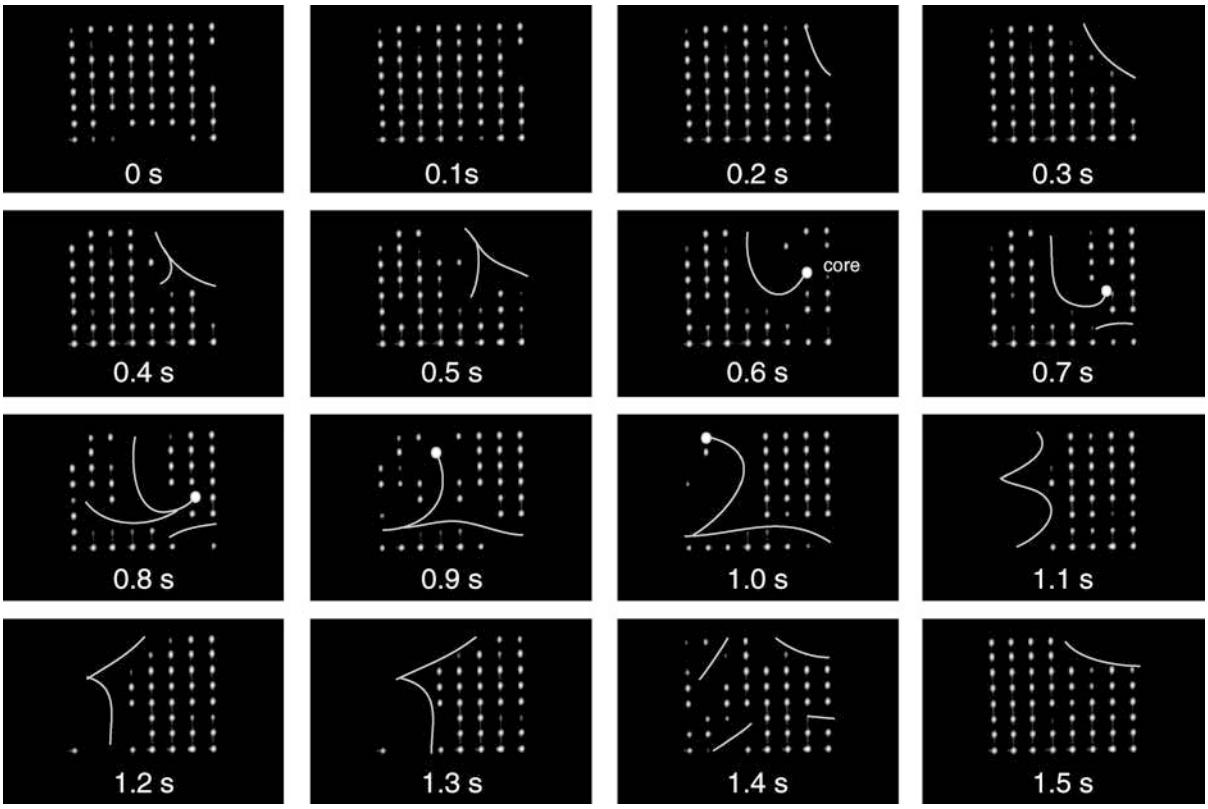

Unconventional Computing, Novel Hardware for, Figure 8
Spiral patterns on RD chip (weak connections between cells) [21]

Figure 8 presents an example where some spiral (modelock) patterns of cell clusters were observed. Snapshots were taken with at intervals of 100 ms. The active-cell clusters and cores of the spirals are superimposed with the figure by white curves and white circles, respectively. Although observing "beautiful" spirals as in chemical RD systems is difficult because of the small number of cells, the appearance and disappearance of small sections of spiral waves were successfully observed.

RD devices and circuits are useful not only for hardware RD systems but also for constructing modern neurochips. The excitatory and oscillatory behaviors of an RD device and circuit are very similar to actual neurons that produce sequences of identically shaped pulses in time, called *spikes*. Recently, Fukai demonstrated that an inhibitory network of spiking neurons achieves robust and efficient neural competition on the basis of a novel timing mechanism of neural activity [28]. A network with such a timing mechanism may provide an appropriate platform to develop analog LSI circuits and could overcome problems with analog devices, namely their lack of precision and reproducibility.

Striped and Spotted Pattern Generation on RD Cellular Automata

A novel RD model that is suitable for LSI implementation and its basic LSI architecture were proposed in [55]. The model employs linear diffusion fields of activators and inhibitors and a discrete transition rule after diffusion. Image-processing LSI circuits based on pattern formation in RD systems were developed. Continuous diffusion fields and an analog state variable were introduced to improve the Young's local activator-inhibitor model [62]. A model pattern diagram was produced on a 2D parameter space through extensive numerical simulations. The spatial frequency and form (striped or spotted) could be controlled with only two parameters. Theoretical analysis of the one dimensional model proved that i) spatial distribution given by a periodic square function is stable at the equilibrium and ii) the spatial frequency is inversely proportional to the square root of a diffusion coefficient of the inhibitors.

A basic circuit for the proposed model was designed, i. e., an RD LSI based on the analog computing method

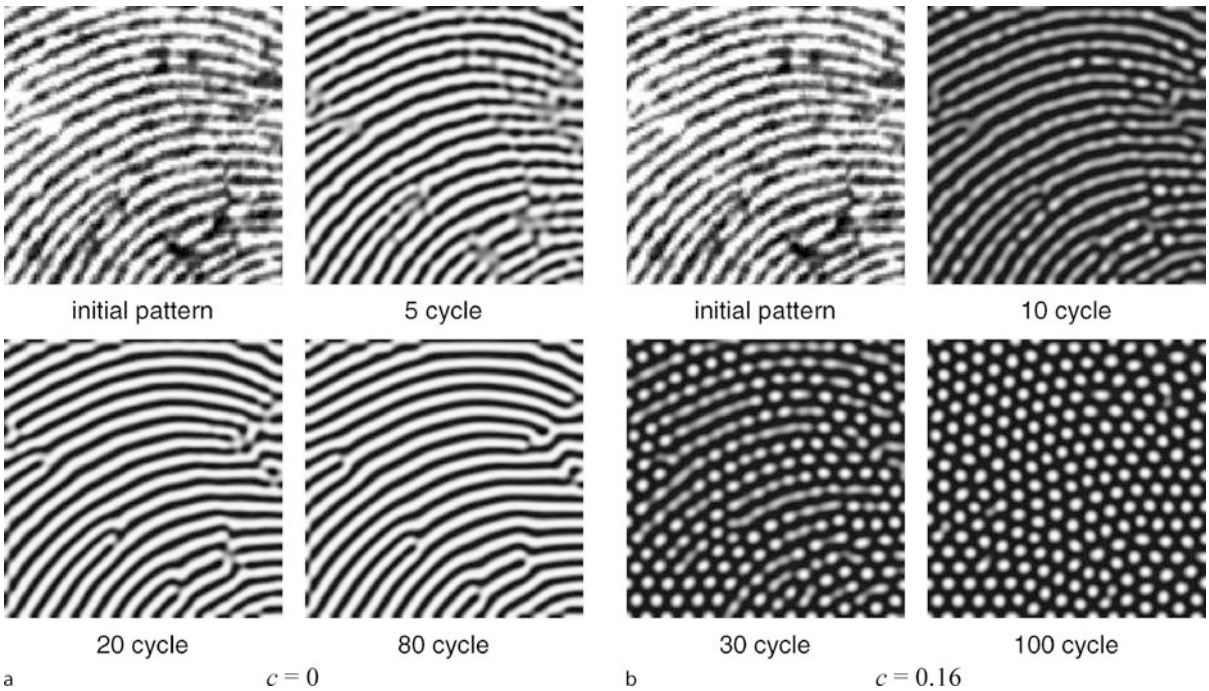

Unconventional Computing, Novel Hardware for, Figure 9
Snapshots of pattern formation from initial fingerprint image [55]

where the concentration of chemicals was represented by a two-dimensional voltage distribution and the cell voltage was diffused step by step. By mimicking two diffusion fields with the proposed model in one diffusion circuit on the LSI, one can reduce the area of the unit cell circuit.

Figure 9a has snapshots of pattern formation in the circuit. One can estimate that the system produces striped patterns from the theory [55]. Therefore, a fingerprint pattern was used as an initial input. Noisy local patterns were repaired by their surrounding striped patterns, as time increased. The circuit required 50 cycles (8000 clocks) to reach equilibrium. Figure 9b shows the results for spot pattern generation. The same initial input as in Fig. 9a was given to the circuit. As expected from the theory, spotted patterns were obtained. The pattern formation process was the same as in Fig. 9a where noisy local spots were restored by surrounding global spotted patterns. Therefore, this circuit would be suitable for restoring regularly-arranged spotted patterns such as polka-dot patterns. The system took 100 cycles (16,000 clocks) until it reached equilibrium to restore the spotted patterns.

Reaction-Diffusion Computing Devices Based on Minority-Carrier Transport in Semiconductors

A massive parallel computing device was designed [18] based on principles of information processing in RD chemical media [4,5] (Figs. 10 and 11). This device imitates auto-catalytic and dissipative phenomena of the chemical RD systems, however comparing to real chemical medium the semi-conductor analog of RD computers, functions much faster. Operational characteristics of the RD silicon devices and feasibility of the approach on several computational tasks are shown in [18]. The results indicate that the proposed RD device will be a useful tool for developing novel hardware architectures based on RD principles of information processing.

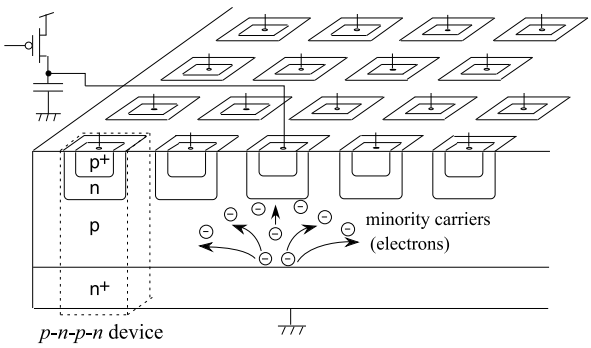

Unconventional Computing, Novel Hardware for, Figure 10
Construction of two-dimensional RD device with vertical *p-n-p-n* device [18]

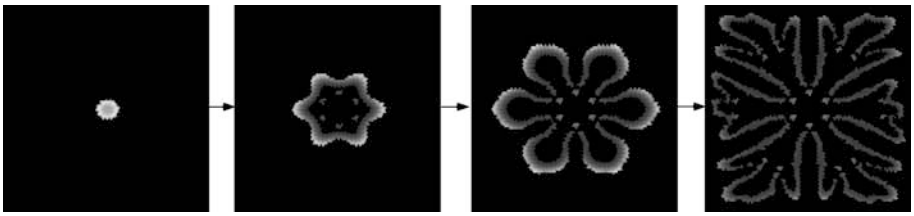

Unconventional Computing, Novel Hardware for, Figure 11
Simulation results of RD device producing multiplicating patterns [18]

Practical value of RD chemical systems are significantly reduced by low speed of traveling waves which makes real-time computation senseless. One of the cost-efficient options to overcome the speed-limitations of RD computers while preserving unique features of wave-based computing is to implement RD chemical computers in silicon. The velocity of traveling wavefronts in typical reaction diffusion systems, e. g., BZ reaction, is 10^{-2} m/s [56], while that of a hardware RD system will be over a million times faster than that of the BZ reaction, independent of system size [7]. The increase in speed will be indispensable for developers of RD computers. Moreover, if a RD system is implemented in integrated circuits, then we would be able to artificially design various types RD spatio-temporal dynamics and thus develop parallel computing processors for novel applications. Basing on experimental evidences of RD-like behavior, namely traveling current density filaments [41], in *p-n-p-n* devices a novel type of semiconductor RD computing device, where minority carriers diffuse as chemical species and reaction elements are represented by *p-n-p-n* diodes, was proposed.

Single-Electron Reaction-Diffusion System

This section introduces a single-electron device that is analogous to the RD system. This electrical RD device consists of a two-dimensional array of single-electron nonlinear oscillators that are combined with one another through diffusive coupling. The device produces animated spatiotemporal patterns of node voltages, e. g., a rotating spiral pattern similar to that of a colony of cellular slime molds and a dividing-and-multiplying pattern that reminds us of cell division. A method of fabricating actual devices by using self-organized crystal growth technology is also described. The following is an excerpt from [43]. For details, see the reference.

Figure 12 shows an electrical oscillator constructed by a single-electron circuit. It consists of tunneling junction C_j and high resistance R connected in series at node 1 and biased by positive voltage V_{dd}. This circuit is an elementary component of single-electron circuits known as the single-electron transistor (SET) cell (see [31] for detailed explanation). A SET cell only has a single variable, voltage V_1 of node 1, but it can be oscillatory or excitatory in operation – which is indispensable in creating RD systems – because the node voltage can produce a discontinuous change because of electron tunneling. In continuous-variable systems such as chemical reaction systems, two or more variables are needed for oscillatory and excitatory operations.

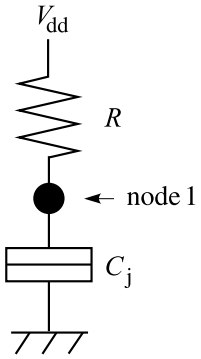

Unconventional Computing, Novel Hardware for, Figure 12
Single-electron oscillator (a SET cell) consisting of tunneling junction C_j, high resistance R connected at node 1, and positive bias voltage V_{dd}

The SET cell operates as a nonlinear oscillator at the low temperatures at which the Coulomb-blockade effect occurs. It is oscillatory (astable) if $V_{dd} > e/(2C_j)$ (e is elementary charge) and produces nonlinear oscillation in voltage at node 1 (Fig. 13a). The node voltage gradually increases as junction capacitance C_j is charged through resistance R, then drops discontinuously because of electron tunneling through the junction, again gradually increasing to repeat the same cycles. In contrast, the oscillator is excitatory (monostable) if $V_{dd} < e/(2C_j)$ and produces single-pulse operation excited by an external trigger (Fig. 13b). A modified Monte Carlo method is used for the simulation. Kuwamura and his colleagues [35] have given details

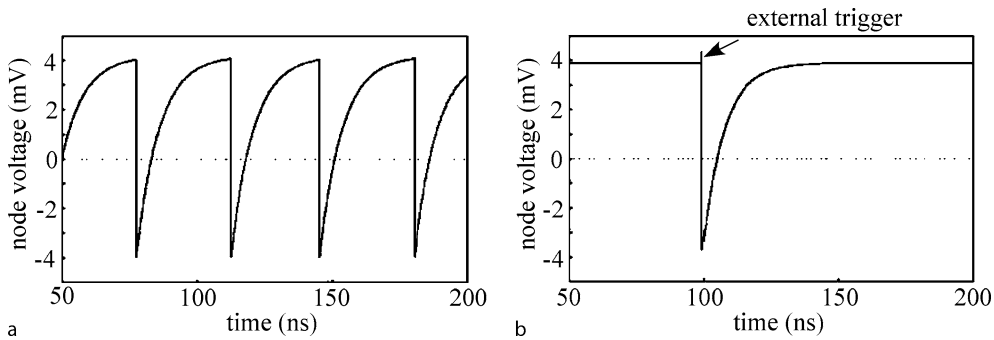

Unconventional Computing, Novel Hardware for, Figure 13
Operation of the oscillator. Waveforms of node voltage are shown for a self-induced oscillation and b monostable oscillation, simulated with following set of parameters: tunneling junction capacitance $C_j = 20$ aF, tunneling junction conductance = 1 μS, high resistance R = 400 MΩ, and zero temperature. Bias voltage $V_{dd} = 4.2$ mV for self-induced oscillation, and $V_{dd} = 3.8$ mV for monostable oscillation

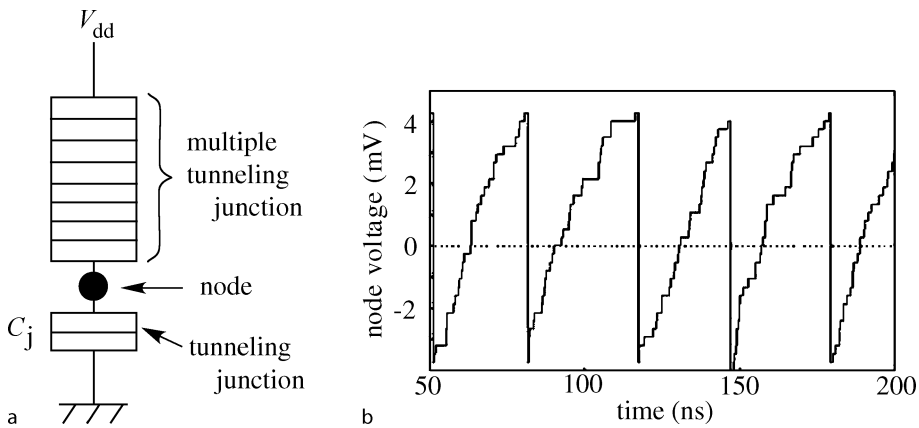

Unconventional Computing, Novel Hardware for, Figure 14
Single-electron oscillator with a multiple tunneling junction: a circuit configuration, and b simulated self-induced oscillation. Parameters are: capacitance of single tunneling junction $C_j = 10$ aF, conductance of the single tunneling junction = 1 μS, 30 tunnelling junctions in the multiple tunneling junction, capacitance and conductance of a tunneling junction in the multiple tunneling junction are 300 aF and 50 nS, bias voltage $V_{dd} = 8.6$ mV, and zero temperature

of this method. Also see Appendix in [59]. For constructing electrical RD systems, oscillatory oscillators and excitatory ones, or both, can be used.

The oscillator exhibits discontinuous, probabilistic kinetics resulting from electron tunneling. The kinetics is given in the form of

$$\frac{dV_1}{dt} = \frac{V_{dd} - V_1}{RC_j} - \frac{e}{C_j}\delta\left(V_1 - \frac{e}{2C_j} - \Delta V\right),$$

where $\delta(\cdot)$ represents a discontinuous change in node voltage caused by electron tunneling. Probabilistic operation arises from the stochastic nature of tunneling; i. e., a time lag (a waiting time) exists between when junction voltage exceeds tunneling threshold $e/(2C_j)$ and when tunneling actually occurs. This effect is represented by delay term ΔV in the equation. Because the value of ΔV has probabilistic fluctuations in every tunneling event and cannot be expressed in analytical form, so Monte Carlo simulation is necessary for studying the behavior of the oscillator.

In fabricating actual oscillators, a high resistance of hundreds of mega-ohms or more is not easy to implement on an LSI chip. A better way is to use a multiple tunneling junction, i. e., a series of many tunneling junctions, instead of high resistance (Fig. 14a). This structure also enables oscillatory and excitatory operations to be obtained because sequential electron tunneling through a multiple tunneling junction has a similar effect to current flowing at high

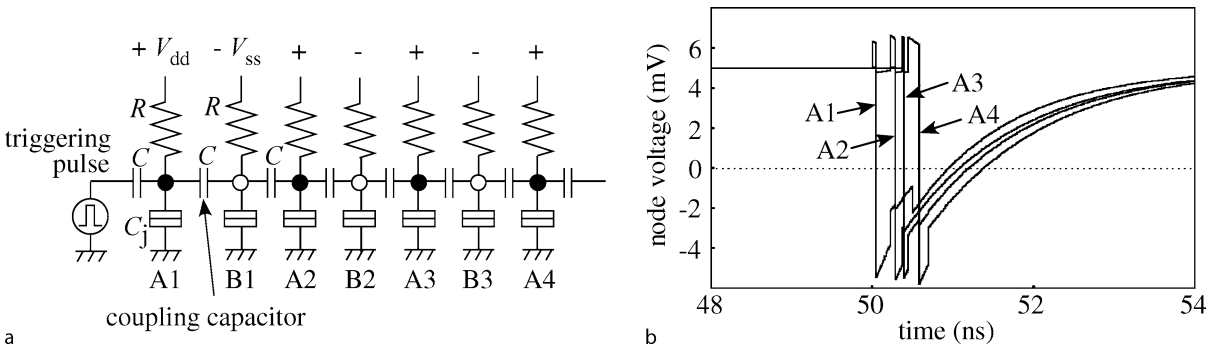

Unconventional Computing, Novel Hardware for, Figure 15

Diffusive connection of oscillators. a One-dimensional chain of oscillators (A1, A2, ...) with intermediary cells (B1, B2, ...) and coupling capacitors *C*. For study of the transmission of tunneling, a triggering pulse generator is connected to the left end. b Transmission of tunneling through the chain of excitatory oscillators. The waveform of node voltage is plotted for each oscillator. A triggering pulse was applied to the leftmost oscillator, and tunneling started at the oscillator to transmit along the chain with delay. Jumps in curves A1–A4 result from electron tunneling in oscillators A1–A4. Simulated with a set of parameters: $C_j = 10$ aF, $C = 2$ aF, $R = 77$ MΩ, tunneling junction conductance = 5 μS, $V_{dd} = 5$ mV, $-V_{ss} = -5$ mV, and zero temperature

resistance (Fig. 14b). In the following sections, however, the high-resistance SET cell (Fig. 12) is used to construct electrical RD systems because less computing time is required in simulating RD operation. One can expect that the knowledge obtained from high-resistance RD systems will be able to be applied to RD systems consisting of multiple-junction oscillators.

To construct RD systems, oscillators have to be connected with one another so that they will interact through "diffusive" coupling to generate synchronization and entrainment. To do this, the oscillators are connected by means of intermediary oscillation cells and coupling capacitors. Figure 15a illustrates the method of connection with a one-dimensional chain of oscillators. The oscillators (SET cells denoted by A1, A2, ..., with their nodes represented by closed circles) are connected with their neighboring oscillators through intermediary oscillation cells (SET cells denoted by B1, B2, ..., with their nodes represented by open circles) and coupling capacitors *C*. One can use an excitatory SET cell biased with a negative voltage $-V_{ss}$ as the intermediary oscillation cell. An excitatory SET cell biased with a negative voltage $-V_{ss}$ is used here as the intermediary oscillation cell.

When electron tunneling occurs in an oscillator in this structure, the node voltage of the oscillator changes from positive to negative, and this induces, through coupling capacitor *C*, electron tunneling in an adjacent intermediary cell. The induced tunneling changes the node voltage of the intermediary cell from negative to positive, and this induces electron tunneling in an adjacent oscillator. In this way, electron tunneling is transmitted from one oscillator to another along the oscillator chain. There is a time lag between two tunneling events in two neighboring oscillators as if these oscillators interacted through diffusion. This phenomenon is not diffusion itself and cannot be expressed in the form $D\Delta\mathbf{u}$ in Eq. (1) but can be used as a substitute for diffusion.

The transmission of tunneling with delay is illustrated in Fig. 15b with simulated results for a chain of excitatory oscillators with intermediary cells. Electron tunneling was induced in the leftmost oscillator by a triggering pulse, and it was transmitted to the right along the chain with delay. In other words, an excitation wave of tunneling traveled to the right along the chain. Its delay in traveling from one oscillator to a neighbor has probabilistic fluctuations because of the stochastic nature of tunneling, but this is not a problem for applications to RD systems.

An electrical RD system can be constructed by connecting oscillators into a network by means of intermediary cells and coupling capacitors (Fig. 16). Each oscillator is connected to its neighboring 4 oscillators by means of 4 intermediary cells and coupling capacitors. This is a two-dimensional RD system. A three-dimensional RD system can also be constructed in a similar way by arranging oscillators into a cubic structure and connecting each oscillator with its 6 neighboring oscillators by means of 6 intermediary cells and coupling capacitors.

In the single-electron RD system, the node voltage of each oscillator changes temporally as the oscillators operate through mutual interactions. Consequently, a two-dimensional spatiotemporal pattern of the node voltages is produced on the RD system. Since this voltage pattern corresponds to the dissipative structure in chemical RD systems, it can be called an "electrical dissipative structure".

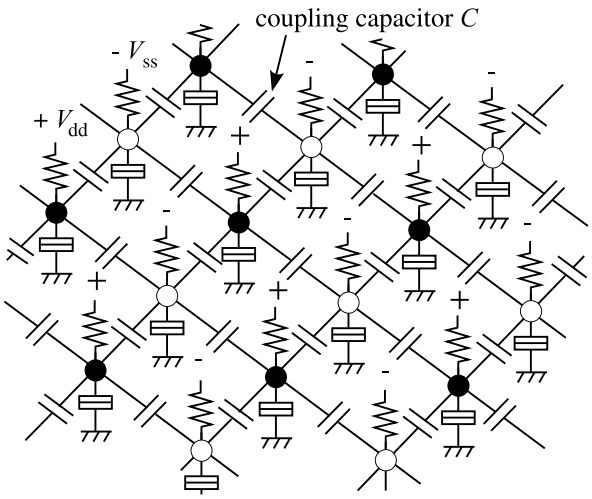

Unconventional Computing, Novel Hardware for, Figure 16
Two-dimensional RD system consisting of the network of single-electron oscillators. Each oscillator (closed-circle node) is connected with 4 neighboring oscillators by means of 4 intermediary cells (open-circle nodes) and coupling capacitors

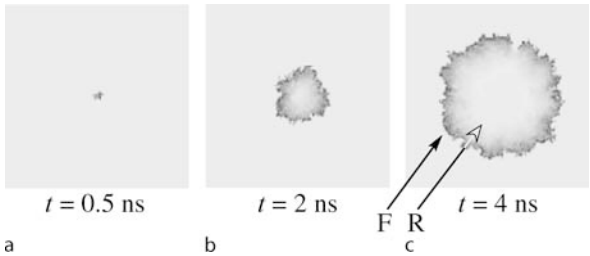

Unconventional Computing, Novel Hardware for, Figure 17
Expanding circular pattern in the single-electron RD system. Snapshots for three time steps. Simulated with parameters: tunneling junction capacitance $C_j = 1$ aF, tunneling junction conductance = 1 μS, high resistance $R = 137.5$ MΩ, coupling capacitance $C = 1$ aF, bias voltage $V_{dd} = 16.5$ mV, $-V_{ss} = -16.5$ mV, and zero temperature

A variety of electrical dissipative structures are produced from different sets of system parameters. To understand the behavior of an electrical RD system entirely, a phase diagram for the system must be drawn, i. e., a diagram that depicts – in the multidimensional space of system parameters – what kind of dissipative structure will appear for each set of parameter values. However, a phase diagram for the RD system cannot be drawn without a long numerical computer simulation because its reaction-diffusion kinetics cannot be expressed in analytical form. Instead, few examples of electrical dissipative structures simulated with a few sample sets of parameter values are demonstrated.

Although a single-electron RD system differs greatly from chemical RD systems in terms of reaction-diffusion kinetics, it can produce dissipative structures similar to those of chemical RD systems. Here, three examples are exhibited, i. e., an expanding circular pattern, a rotating spiral pattern, and a dividing-and-multiplying pattern. The following will have the results simulated for a RD system consisting of 201 × 201 excitatory oscillators and 200 × 200 intermediary cells.

Expanding Circular Pattern

A single-electron RD system consisting of excitatory oscillators is in a stable uniform state as it stands. Once a triggering signal is applied to an oscillator in the system, an excitation wave of tunneling starts at the oscillator and propagates in all directions to form an expanding circular pattern. This can be seen in Fig. 17; the node voltage of each oscillator is represented by a gray scale: the light shading means high voltage, and the dark means low voltage. The front F of the wave is the region where tunneling just occurred and, therefore, the node voltage of the oscillators is at the lowest negative value. The front line is uneven or irregular because the velocity of the traveling wave fluctuated in each direction throughout the process because of the stochastic waiting time of tunneling.

After the excitation wave passed through, the node voltage of each oscillator gradually increased to return to its initial value, the positive bias voltage. This is indicated in the figure by the light shading on the rear R of the wave. If a triggering signal is applied repeatedly to one oscillator, a concentric circular wave – called a target pattern in chemical RD systems – will be generated.

Rotating Spiral Pattern

This pattern appears when an expanding circular wave is chipped by external disturbance, thereby making an endpoint to appear in the wave front. With this endpoint acting as a center, the wave begins to curl itself to form a rotating spiral pattern (Fig. 18). The principle of curling is similar to that in chemical RD systems.

In this example, a triggering signal was applied to the middle oscillator on the left of the RD system. When an excitation wave started and expanded a little, the lower half of the wave was chipped by resetting the node voltage of oscillators to zero (Fig. 18a). After that, the RD system was left to operate freely, and a rotating spiral pattern of node voltages automatically generated as can be seen in Figs. 18b–f.

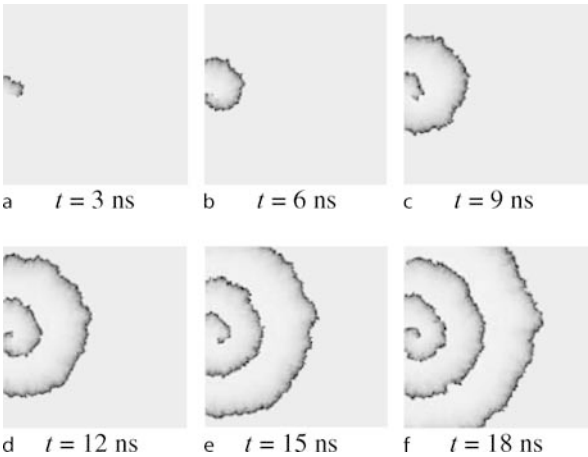

Unconventional Computing, Novel Hardware for, Figure 18
Rotating spiral pattern. Snapshots for six time steps. Simulated with the same parameters as used for Fig. 17

Dividing-and-Multiplying Pattern

This pattern appears when the coupling between oscillators is weak (i. e., small coupling capacitance or low bias voltage). When this happens, electron tunneling in an oscillator cannot be transmitted to all four adjacent intermediary cells; e. g., tunneling can be transmitted to the right and left cells but not to the upper and lower cells. As a result, an expanding node-voltage pattern splits into pieces, and each piece again expands to split again. This produces dividing-and-multiplying patterns (Fig. 19). The principle of division is different from that in chemical RD systems,

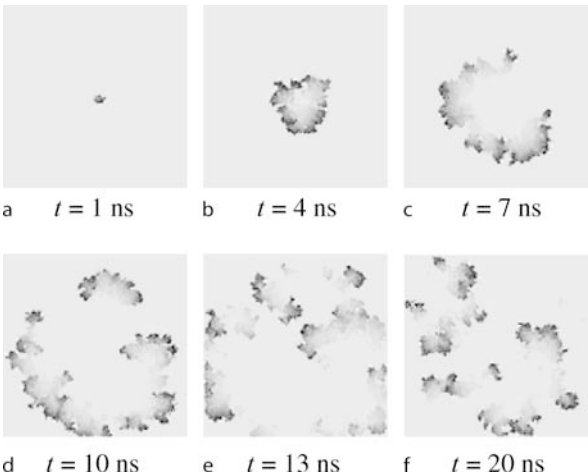

Unconventional Computing, Novel Hardware for, Figure 19
Dividing-and-multiplying pattern. Snapshots for six time steps. Simulated with the same parameters as used for Fig. 17 except that $R = 150.5\,\mathrm{M\Omega}$, $V_{dd} = 15.8\,\mathrm{mV}$, and $-V_{ss} = -15.8\,\mathrm{mV}$

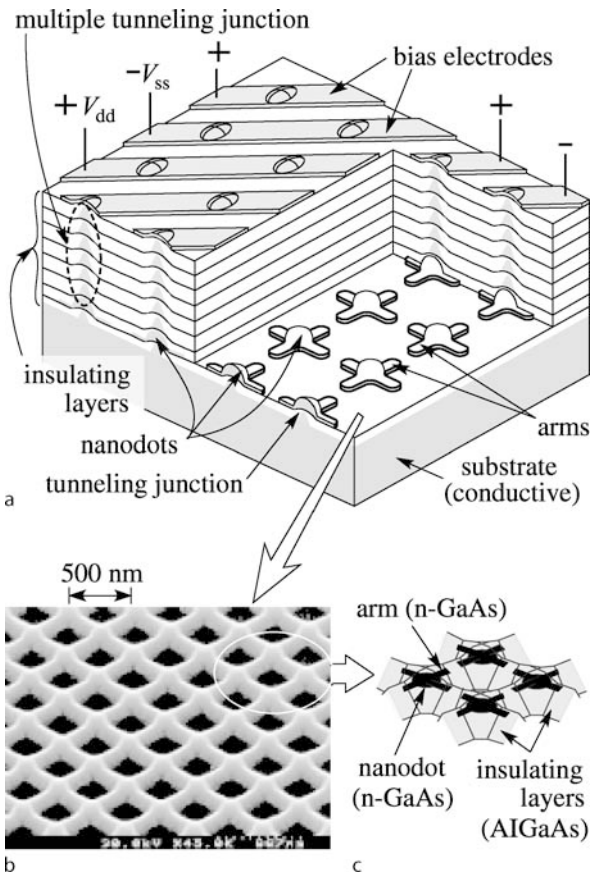

Unconventional Computing, Novel Hardware for, Figure 20
Device structure for the single-electron RD system. a Three-dimensional and cross-sectional schematics; b SEM photograph of a two-dimensional array of GaAs nanodots with coupling arms and tunneling junctions; c schematic diagram of the nanodots with coupling arms

but the behavior of created patterns is somewhat similar. In a way, we may consider that there are electrical microbes consisting of negative voltages living on the RD system, eating positive charges on nodes as food, and propagating so that they can will spread all over the system.

The unit element in this RD system is a single-electron oscillator coupled with four neighbors. The multiple-tunneling-junction oscillator (Fig. 14) is preferable for this element because it can be made without high resistance, which is difficult to implement on an LSI chip. Arranging such oscillators into a two-dimensional array produces an RD system, so the next task is to fabricate many identical oscillators on a substrate. Figure 20a shows the three-dimensional and cross-sectional schematics for the structure of the device. Each oscillator consists of a conductive nanodot (minute dot) with four coupling arms, and there is

a tunneling junction between the nanodot and the conductive substrate beneath it. Many series-connected junctions run between the nanodot and a positive-bias or a negative-bias electrode. Capacitive coupling between neighboring oscillators can be achieved by laying their coupling arms close to each other.

The key in this construction is to prepare a large arrangement of nanodots with coupling arms and tunneling junctions. A process technology that could be used to fabricate the RD-system structure was previously proposed and demonstrated [42]. This technology uses self-organized crystal growth achieved by selective-area metalorganic vapor-phase epitaxy (SA-MOVPE), and it can be used to fabricate GaAs nanodots with arms and tunneling junctions on a GaAs substrate by making use of the dependence of the crystal-growth rate on crystal orientation (for detailed explanation, see [36] and [37]). With this technology, a nanodot with four coupling arms can be formed automatically in a self-organizing manner. This technology can also be used to automatically create the structure for multiple tunneling junctions on nanodots simply by repeating the growth of an *n*-type GaAs layer and an insulating AlGaAs layer. Using such a process, the formation of GaAs nanodots with their arms and tunneling junctions beneath them was succeeded, in the form of a two-dimensional array on a substrate (Figs. 20b and c), though the technology is not yet perfect and a complete device has not been fabricated yet. An improved process technology to form GaAs nanodots with arms and multiple tunneling junctions is now under development, where the arms and multiple tunneling junctions are arranged regularly with a smaller pitch of 100 nm or less (corresponding to 1010 oscillators/cm^2). With the improved process technology, we will be able to integrate coupled single-electron oscillators on a chip and proceed from there to develop reaction-diffusion LSIs.

Collision-Based RD Computers

Present digital LSI systems consist of a number of combinational and sequential logic circuits as well as related peripheral circuits. A well-known basic logic circuit is a two-input NAND circuit that consists of four metal-oxide semiconductor field-effect transistors (MOS FETs) where three transistors are on the current path between the power supply and the ground. Many complex logic circuits can be constructed by not only populations of a large number of NAND circuits but also special logic circuits with a small number of transistors (there are more than three transistors on the current path) compared with NAND-based circuits.

A straight-forward way to construct low-power digital LSIs is to decrease the power-supply voltage because the power consumption of digital circuits is proportional to the square of the supply voltage. In complex logic circuits, where many transistors are on the current paths, the supply voltage cannot be decreased due to stacking effects of transistors' threshold voltages, even though the threshold voltage is decreasing as LSI fabrication technology advances year by year. On the other hand, if two-input basic gates that have the minimum number of transistors (three or less) on the current path are used to decrease the supply voltage, a large number of the gates will be required for constructing complex logic circuits.

The Reed–Muller expansion [40,48], which expands logical functions into combinations of AND and XOR logic, enables us to design 'specific' arithmetic functions with a small number of gates, but it is not suitable for arbitrary arithmetic computation. Pass-transistor logic (PTL) circuits use a small number of transistors for basic logic functions but additional level-restoring circuits are required for every unit [52]. Moreover, the acceptance of PTL circuits into mainstream digital design critically depends on the availability of tools for logic, physical synthesis, and optimization. Current-mode logic circuits also use a small number of transistors for basic logic, but their power consumption is very high due to the continuous current flow in turn-on states [15]. Subthreshold logic circuits where all the transistors operate under their threshold voltage are expected to exhibit ultra-low power consumption, but the operation speed is extremely slow [53]. Binary decision diagram logic circuits are suitable for next-generation semiconductor devices such as single-electron transistors [16,51], but not for present digital LSIs because of the use of PTL circuits.

To address the problems above concerning low-power and high-speed operation in digital LSIs, a method of designing logic circuits with collision-based fusion gates, which is inspired by collision-based RD computing (RDC) [6,7], is described. In the following sections, we see a new interpretation of collision-based RDC, especially concerning directions and speeds of propagating information quanta. We also see basic logical functions constructed by collision-based fusion gates, and discuss the number of transistors in classical and fusion-gate logic circuits [60].

Collision-Based Reaction-Diffusion Computing for Digital LSIs

Adamatzky proposed how to realize arithmetical scheme using wave fragments traveling in a RD medium where ex-

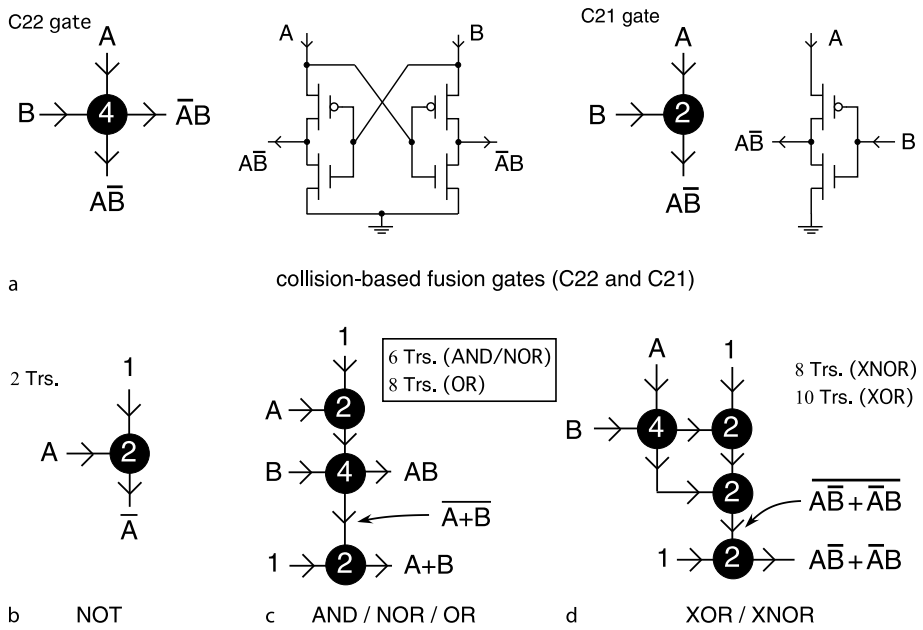

Unconventional Computing, Novel Hardware for, Figure 21
Construction of collision-based fusion gates [60]. a Definition of 2-in 2-out (C22) and 2-in 1-out (C21) gates and their corresponding circuits; b NOT; c AND, NOR, and OR; d XOR and XNOR functions

citable chemical waves disappear when they collide each other [6,7]. His cellular-automaton model mimicked localized excitable waves (wave fragments) traveling along columns and rows of the lattice and along diagonals. The wave fragments represented values of logical variables where logical operations were implemented when wave fragments collided and were annihilated or reflected as a result of the collision. One can achieve basic logical gates in the cellular-automaton model, and build an arithmetic circuit using the gates [6].

The cellular-automaton model for basic logic gates has been implemented on digital LSIs [7]. Each cell consisted of several tens of transistors and was regularly arranged on a 2D chip surface. To implement a one-bit adder, for example, by collision-based cellular automata, at least several tens of cells are required to allocate sufficient space for the collision of wave fragments [6]. This implies several hundreds of transistors are required for constructing just a one-bit adder. Direct implementation of the cellular automaton model is therefore a waste of chip space, as long as the single cell space is decreased to the same degree of chemical compounds in spatially-continuous RD processors.

What happens if wave fragments travel in 'limited directions instantaneously'? Our possible answers to this question are depicted in Fig. 21. Figure 21a shows 2-in 2-out (C22) and 2-in 1-out (C21) units representing two perpendicular 'limited' directions of wave fragments, i. e., North-South and West-East fragments. The number of MOS transistors in each unit is written inside the black circle in the figure. The input fragments are represented by values A and B where A (or B) = '1' represents the existence of a wave fragment traveling North-South (or West-East), and A (or B) = '0' represents the absence of wave fragments. When A = B = '1' wave fragments collide at the center position (black circle) and then disappear. Thus, East and South outputs are '0' because of the disappearance. If A = B = '0', the outputs will be '0' as well because of the absence of the fragments. When A = '1' and B = '0', a wave fragment can travel to the South because it does not collide with a fragment traveling West-East. The East and South outputs are thus '0' and '1', respectively, whereas they are '1' and '0', respectively, when A = '0' and B = '1'. Consequently, logical functions of this simple 'operator' are represented by $\bar{A}B$ and $A\bar{B}$, as shown in Fig. 21a left. We call this operator a collision-based 'fusion gate', where two inputs correspond to perpendicular wave fragments, and one (or two) output represents the results of the collision (transparent or disappear) along the perpendicular axes. Figures 21b to d represent basic logic circuits constructed by combining several fusion gates. The simplest example is shown in Fig. 21b where the NOT function

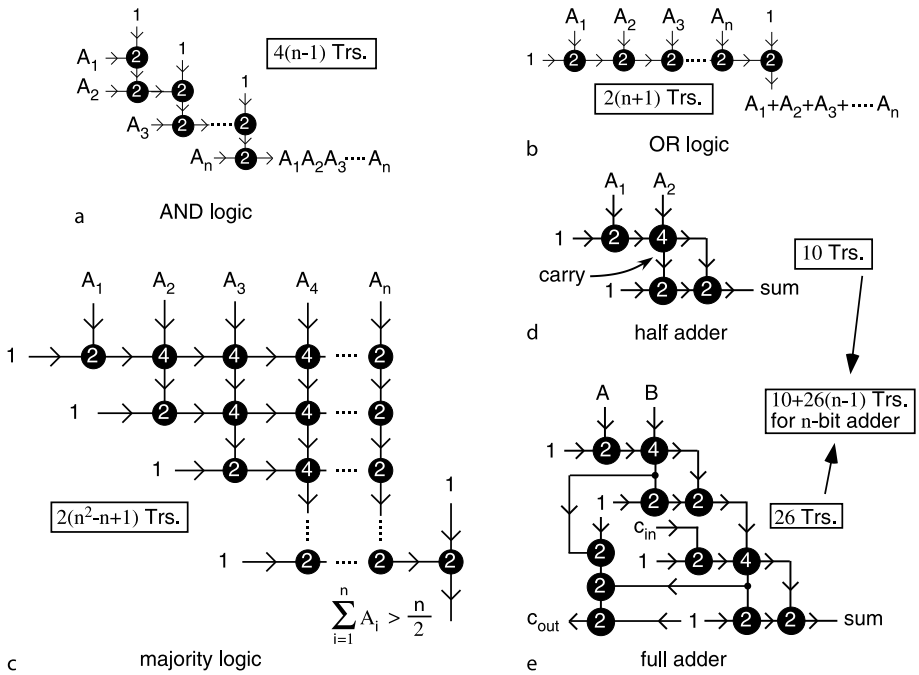

Unconventional Computing, Novel Hardware for, Figure 22
Fusion gate architectures of multiple-input functions [60]; a AND, b OR, c majority logic gates. Half and full adders are shown in d and e, respectively

is implemented by a C21 gate (2 transistors). The North input is always '1', whereas the West is the input (A) of the NOT function. The output appears on South node ($\overline{A}$). Figure 21c represents a combinational circuit of three fusion gates (two C21 gates and one C22 gate) that produces AND, NOR, and OR functions. Exclusive logic functions are produced by three (for XNOR) or four (for XOR) fusion gates as shown in Fig. 21d. The number of transistors for each function is depicted in the figure (inside the white boxes).

A collision-based fusion gate receives two logical inputs (A and B) and produces one (C21) or two (C22) logical outputs; i. e., $A\overline{B}$ for C21, $\overline{A}B$ and $A\overline{B}$ for C22. Unit circuits for C22 and C21 gates receive logical (voltage) inputs (A and B) and produce these logic functions. The minimum circuit structure is based on PTL circuits where a single-transistor AND logic is fully utilized. For instance, in Fig. 21a right, a pMOS pass transistor is responsible for the $\overline{A}B$ function, and an additional nMOS transistor is used for discharging operations. When the pMOS transistor receives voltages A and B at its gate and drain, respectively, the source voltage approaches $\overline{A}B$ at equilibrium. If a pMOS transistor is turned off, an nMOS transistor connected between the pMOS transistor and the ground discharges the output node, which significantly increases the upper bound of the operation frequency. When A = B = '0', the output voltage is not completely zero because of the threshold voltage of the MOS transistors, however, this small voltage shift is restored to logical '0' at the next input stage. Therefore additional level-restoring circuits are unnecessary for this circuit.

Figure 22 shows constructions of multiple-input logic functions with fusion gates. In classical circuits, two-input AND and OR gates consist of six transistors. To decrease the power supply voltage for low-power operation, a small number of transistors (three or less) should be on each unit's current path. Since each unit circuit has six transistors, n-input AND and OR gates consist of $6(n-1)$ transistors ($n \geq 2$). On the other hand, in fusion gate logic [(a) and (b)], a n-input AND gate consisted of $4(n-1)$ transistors, whereas $2(n+1)$ transistors were used in an n-input OR gate. Therefore, in case of AND logic, the number of transistors in fusion gate circuits is smaller than that of classical circuits. The difference will be significantly expanded as n increases. Figure 22c shows fusion gate implementation of majority logic circuits with multiple inputs. Again, in classical circuits, the number of transistors on each unit's current path is fixed to three. For n-bit inputs (n must be an odd number larger than 3), the number of transistors in the classical circuit was $30 + 36(n-3)^2$,

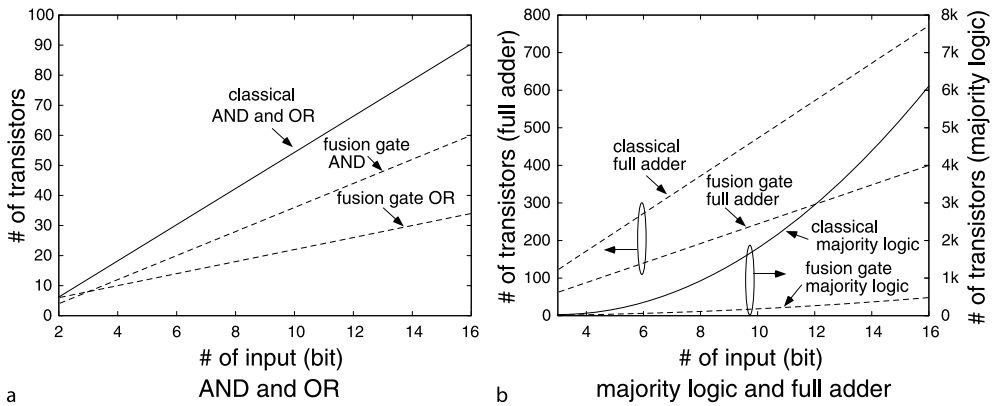

Unconventional Computing, Novel Hardware for, Figure 23
Total number of transistors in classical and collision-based (CB) multiple-input logic gates [60]

while in the fusion gate circuit, it was $2(n^2 - n + 1)$, which indicates that the collision-based circuit has a great advantage in the number of transistors. Half- and full adders constructed by fusion gate logic are illustrated in Figs. 22d and e. The number of transistors in a classical half adder was 22, while it was 10 in a fusion gate half adder (Fig. 22d). For n-bit full adders ($n \geq 1$), the number of transistors in a classical circuit was $50n - 28$, while it was $26(n - 1) + 10$ in a fusion gate circuit (Fig. 22e). Again, the fusion gate circuit has a significantly smaller number of transistors, and the difference will be increased as n increases.

Figure 23 summarizes the comparison of the number of transistors between classical and fusion gate logic. The number of transistors in fusion gate logic was always smaller than that of transistors in classical logic circuits, especially in majority logic gates.

Future Directions

Before starting any computation one should input data-information in the RD medium. A parallel input is an essential feature of an edge-cutting parallel computing architecture. Serial inputs, so common for vast majority of massively-parallel processors, dramatically decrease performance of the computing devices, particularly those operating in transducer mode, where information is constantly fed into the processor (e. g. in tasks of image processing). Experimental RD chemical computers, at least in certain case, may well have analogs of parallel inputs. It has been demonstrated widely that applying light of varying intensity we can control excitation dynamic in BZ-medium [22,27,32,44], wave velocity [49], patter formation [58]. Of particular interest are experimental evidences of light-induced back propagating waves, wavefront splitting and phase shifting [61]; we can also manipulate medium's excitability by varying intensity of the medium's illumination [23]. In fact, optical input of data-information has been already used at the beginning of RD research [34]. This was proved to particularly important in experiments in image processing in BZ-medium-based computing devices [34,45,46,47]. We are not aware of any rigorous experimental results demonstrating a possibility of optical inputs of RD semi-conductor devices there is however a simulation-related evidence of a possibility of optical parallel inputs. Paper [39] discusses particulars of a photo-response, photon-induced generation of electron-hole pairs in *p-n-p-n* devices, depending on primary color components in the stimulating input because elementary processors can also act as color sensors.

There exists a possibility of parallel optical outputs on semiconductor devices. Technologies for integrating optoelectronic devices and electronic circuitry are fully developed, but limited hybrid integration is available commercially at present. An important problem of such integration is that pursuing it involves simultaneous development of sophisticated technologies for optoelectronic devices (III-V semiconductors) and silicon integrated circuits. Indeed, recent development in optoelectronic integrated circuits (OEICs) enables us to implement light-emitting devices (LEDs) on silicon substrate by controlling defects at III-V/silicon interface. For example, Furukawa et al. demonstrated that lattice-matched and defect-free GaPN epilayers can be grown on silicon with a thin GaP buffer layer [29]. The task is enormous and as a practical matter only small scale OEICs have been demonstrated. While the integration levels of III-V OEICs have remained low, the degree of integration in commercial GaAs inte-

grated circuits has reached LSI levels in recent years. These advances offer a route to achieving much higher levels of optoelectronic integration through epitaxial growth of III-V heterostructures on GaAs-based LSI electronics.

Most experimental prototypes of chemical excitable computing devices suffer from difficulties with representation of results of wave interaction. This is because being excited a medium's micro-volume becomes refractory and then recovers back to the resting state. Excitation wave-fronts usually annihilate in the result of collision. Therefore, experimental excitable RD computers require some external devices, like digital camera, to record their spatio-temporal dynamics. Thus, e. g. to compute a collision-free path around obstacles in thin-layer BZ-medium one must record snapshots of BZ-medium's activity and then analyze this series of snapshots to extract results of the computation [11,14]. This is the cost we pay for reusable (because the excitable medium eventually returns to resting state) RD processors. Another option of preserving results of computation may be to employ non-excitable RD media, where a precipitate is formed (or do not formed) in the result of diffusion wave interaction with a substrate (or competition of several wave fronts for the substrate) [9,10,13,25]. Precipitation is an analog of infinite memory. The feature is priceless however makes experimental prototypes simply disposable, in contrast to excitable media the precipitate-forming media can be used just once.

RD semiconductor computers, because they are essentially man-made devices, may allow us to combine reusability and rich space-time dynamics of excitable RD media with low post-processing costs of precipitate-forming RD media. This can be done by embedding a lattice of oscillatory semiconductor elements into a lattice of excitatory semiconductor elements. The excitatory elements (EEs) will form a substrate to support traveling excitation waves while oscillatory elements (OEs) will play a role of rewritable memory. For example, to represent sites of wave-front collision we must adjust an activation threshold of OEs in such manner that front of a single wave will not trigger OEs, however when two or more wave-fronts collide the OEs at the sites of collision are triggered and continue oscillate even when the wave-fronts annihilate.

What computational tasks can be realized in RD semiconductor devices? Most primitive operations of image processing (see overview in [5]), e. g. detection of contour and enhancement, are straightly mapped onto the silicon architecture. Silicon implementation of RD (precipitate formation based) algorithms for computational geometry – Voronoi diagram [13,25,26] and skeletonization [12] – requires embedding of oscillatory elements in the lattice of excitatory *p-n-p-n* devices; the oscillating elements will represent bisectors of a Voronoi diagram and segments of a skeleton.

Universal computation can be realized in RD semiconductor devices using two approaches. Firstly, by employing collision-based mode of computation in excitable media [3], where functionally complete sets of Boolean gates are implemented by colliding waves. In collision-based mode, however we must force the medium to be in a sub-excitable regime which may pose some problems from fabrication point of view. Secondly, we can 'physically' embed logical circuits, namely their diagram-based representations, into the excitable medium [54,57]. In this we employ particulars of photo-response of *p-n-p-n* elements and project the logical circuit onto the medium as pattern of heterogeneous illumination.

Bibliography

1. Adamatzky A (1994) Reaction–diffusion algorithm for constructing discrete generalized Voronoi diagram. Neural Netw World 6:635–643
2. Adamatzky A (1996) Voronoi–like partition of lattice in cellular automata. Mathl Comput Modeling 23:51–66
3. Adamatzky A (1998) Universal dynamical computation in multi-dimensional excitable lattices. Int J Theor Phys 37:3069–3108
4. Adamatzky A (2000) Reaction-diffusion and excitable processors: A sense of the unconventional. Parallel Distrib Comput Theor Pract 3:113–132
5. Adamatzky A (2001) Computing in nonlinear media and automata collectives. Institute of Physics Publishing, Bristol
6. Adamatzky A (ed) (2002) Collision-based computing. Springer, London
7. Adamatzky A, De Lacy Costello B, Asai T (2005) Reaction-diffusion computers. Elsevier, Amsterdam
8. Adamatzky A, Arena P, Basile A, Carmona-Galán R, De Lacy Costello B, Fortuna L, Frasca M, Rodríguez-Vázquez A (2004) Reaction-diffusion navigation robot control: From chemical to VLSI analogic processors. IEEE Trans Circuit Syst I 51:926–938
9. Adamatzky A, De Lacy Costello BPJ (2002) Experimental logical gates in a reaction-diffusion medium: The XOR gate and beyond. Phys Rev E 66:046112
10. Adamatzky A, De Lacy Costello BPJ (2003) On some limitations of reaction-diffusion computers in relation to Voronoi diagram and its inversion. Phys Lett A 309:397–406
11. Adamatzky A, De Lacy Costello BPJ (2002) Collision-free path planning in the Belousov-Zhabotinsky medium assisted by a cellular automaton. Naturwissenschaften 89:474–478
12. Adamatzky A, De Lacy Costello B, Ratcliffe NM (2002) Experimental reaction-diffusion pre-processor for shape recognition. Phys Lett A 297:344–352
13. Adamatzky A, Tolmachiev D (1997) Chemical processor for computation of skeleton of planar shape. Adv Mater Optics Electron 7:135–139

14. Agladze K, Magome N, Aliev R, Yamaguchi T, Yoshikawa K (1997) Finding the optimal path with the aid of chemical wave. Physica D 106:247–254
15. Alioto M, Palumbo G (2005) Model and design of bipolar and MOS current-mode logic: CML, ECL and SCL digital circuits. Springer, Berlin
16. Asahi N, Akazawa M, Amemiya Y (1998) Single-electron logic systems based on the binary decision diagram. IEICE Trans Electron E81-C:49–56
17. Asai T, Nishimiya Y, Amemiya Y (2002) A CMOS reaction-diffusion circuit based on cellular-automaton processing emulating the Belousov–Zhabotinsky reaction. IEICE Trans Fundam E85-A:2093–2096
18. Asai T, Adamatzky A, Amemiya Y (2004) Towards reaction-diffusion computing devices based on minority-carrier transport in semiconductors. Chaos Solit Fractals 20:863–876
19. Asai T, Ikebe M, Hirose T, Amemiya Y (2005) A quadrilateral-object composer for binary images with reaction-diffusion cellular automata. Int J Parallel Emergent Distrib Syst 20:57–68
20. Asai T, De Lacy Costello B, Adamatzky A (2005) Silicon implementation of a chemical reaction-diffusion processor for computation of Voronoi diagram. Int J Bifurc Chaos 15:3307–3320
21. Asai T, Kanazawa Y, Hirose T, Amemiya Y (2005) Analog reaction-diffusion chip imitating the Belousov–Zhabotinsky reaction with hardware Oregonator model. Int J Unconv Comput 1:123–147
22. Beato V, Engel H (2003) Pulse propagation in a model for the photosensitive Belousov–Zhabotinsky reaction with external noise. In: Schimansky-Geier L et al (eds) Noise in complex systems and stochastic dynamics. Proc SPIE 5114:353–362
23. Brandtstädter H, Braune M, Schebesch I, Engel H (2000) Experimental study of the dynamics of spiral pairs in light-sensitive Belousov–Zhabotinskii media using an open-gel reactor. Chem Phys Lett 323:145–154
24. Daikoku T, Asai T, Amemiya Y (2002) An analog CMOS circuit implementing Turing's reaction-diffusion model. Proc Int Symp Nonlinear Theor Appl 809–812
25. De Lacy Costello B, Adamatzky A (2003) On multitasking in parallel chemical processors: experimental findings. Int J Bifurc Chaos 13:521–533
26. De Lacy Costello B, Adamatzky A, Ratcliffe N, Zanin AL, Liehr AW, Purwins HG (2004) The formation of Voronoi diagrams in chemical and physical systems: Experimental findings and theoretical models. Int J Bifurc Chaos 14:2187–2210
27. Flessels JM, Belmonte A, and Gáspár V (1998) Dispersion relation for waves in the Belousov–Zhabotinsky reaction. J Chem Soc Faraday Trans 94:851–855
28. Fukai T (1996) Competition in the temporal domain among neural activities phase-locked to subthreshold oscillations. Biol Cybern 75:453–461
29. Furukawa Y, Yonezu H, Ojima K, Samonji K, Fujimoto Y, Momose K, Aiki K (2001) Control of N content of GaPN growun by molecular beam epitaxy and growth of GaPN lattice-matched to Si(100) substrate. Jpn J Appl Phys 41:528–532
30. Gerhardt M, Schuster H, Tyson JJ (1990) A cellular automaton model of excitable media. Physica D 46:392–415
31. Gravert H, Devoret MH (1992) Single charge tunneling – Coulomb blockade phenomena in nanostructures. Plenum, New York
32. Grill S, Zykov VS, Müller SC (1996) Spiral wave dynamics under pulsatory modulation of excitability. J Phys Chem 100:19082–19088
33. Karahaliloglu K, Balkir S (2005) Bio-inspired compact cell circuit for reaction-diffusion systems. IEEE Trans Circuits Syst II Express Briefs 52:558–562
34. Kuhnert L, Agladze KL, Krinsky VI (1989) Image processing using light-sensitive chemical waves. Nature 337:244–247
35. Kuwamura N, Taniguchi K, Hamakawa C (1994) Simulation of single-electron logic circuits. IEICE Trans Electron J77-C-II:221–228
36. Kumakura K, Nakakoshi K, Motohisa J, Fukui T, Hasegawa H (1995) Novel formation method of quantum dot structures by self-limited selective area metalorganic vapor phase epitaxy. Jpn J Appl Phys 34:4387–4389
37. Kumakura K, Motohisa J, Fukui T (1997) Formation and characterization of coupled quantum dots (CQDs) by selective area metalorganic vapor phase epitaxy. J Crystal Growth 170:700–704
38. Matsubara Y, Asai T, Hirose T, Amemiya Y (2004) Reaction-diffusion chip implementing excitable lattices with multiple-valued cellular automata. IEICE Electron Express 1:248–252
39. Mohajerzadeh S, Nathan A, Selvakumar CR (1994) Numerical simulation of a p-n-p-n color sensor for simultaneous color detection. Sens Actuators A 44:119–124
40. Muller DE (1954) Application of boolean algebra to switching circuit design and to error detection. IRE Trans Electr Comp EC-3:6–12
41. Niedernostheide FJ, Kreimer M, Kukuk B, Schulze HJ, Purwins HG (1994) Travelling current density filaments in multilayered silicon devices. Phys Lett A 191:285–290
42. Oya T, Asai T, Fukui T, Amemiya Y (2002) A majority-logic nanodevice using a balanced pair of single-electron boxes. J Nanosci and Nanotech 2:333–342
43. Oya T, Asai T, Fukui T, Amemiya Y (2005) Reaction-diffusion systems consisting of single-electron circuits. Int J Unconv Comput 1:123–147
44. Petrov V, Ouyang Q, Swinney HL (1997) Resonant formation in a chemical system. Nature 388:655–657
45. Rambidi NG (1998) Neural network devices based on reaction-duffusion media: An approach to artificial retina. Supramol Sci 5:765–767
46. Rambidi NG, Yakovenchuk D (2001) Chemical reaction-diffusion implementation of finding the shortest paths in a labyrinth. Phys Rev E63:0266071–0266076
47. Rambidi NG, Shamayaev KE, Peshkov GY (2002) Image processing using light-sensitive chemical waves. Phys Lett A 298:375–382
48. Reed IS (1954) A class of multiple-error-correcting codes and their decoding scheme. IRE Trans Inform Th PGIT-4:38–49
49. Schebesch I, Engel H (1998) Wave propagation in heterogeneous excitable media. Phys Rev E 57:3905–3910
50. Serrano-Gotarredona T, Linares-Barranco B (2003) Log-domain implementation of complex dynamics reaction-diffusion neural networks. IEEE Trans Neural Networks 14:1337–1355
51. Shelar RS, Sapatnekar SS (2001) BDD decomposition for the synthesis of high performance PTL circuits. Workshop Notes IEEE IWLS 2001 298–303
52. Song M, Asada K (1998) Design of low power digital VLSI circuits based on a novel pass-transistor logic. IEICE Trans Electron E81-C:1740–1749

53. Soeleman H, Roy K, Paul BC (2001) Robust subthreshold logic for ultra-low power operation. IEEE Trans VLSI Syst 9:90–99
54. Steinbock O, Toth A, Showalter K (1995) Navigating complex labyrinths: optimal paths from chemical waves. Science 267:868–871
55. Suzuki Y, Takayama T, Motoike IN, Asai T (2007) Striped and spotted pattern generation on reaction-diffusion cellular automata: Theory and LSI implementation. Int J Unconv Comput 3:1–13
56. Tóth Á, Gáspár V, Showalter K (1994) Propagation of chemical waves through capillary tubes. J Phys Chem 98:522–531
57. Tóth Á, Showalter K (1995) Logic gates in excitable media. J Chem Phys 103:2058–2066
58. Wang J (2001) Light-induced pattern formation in the excitable Belousov–Zhabotinsky medium. Chem Phys Lett 339:357–361
59. Yamada T, Akazawa M, Asai T, Amemiya Y (2001) Boltzmann machine neural network devices using single-electron tunneling. Nanotechnology 12:60–67
60. Yamada K, Asai T, Hirose T, Amemiya Y (2008) On digital LSI circuits exploiting collision-based fusion gates. Int J Unconv Comput 4:45–59
61. Yoneyama M (1996) Optical modification of wave dynamics in a surface layer of the Mn-catalyzed Belousov–Zhabotinsky reaction. Chem Phys Lett 254:191–196
62. Young DA (1984) A local activator-inhibitor model of vertebrate skin patterns. Math Biosci 72:51–58

V

Vehicular Traffic: A Review of Continuum Mathematical Models

Benedetto Piccoli, Andrea Tosin
Istituto per le Applicazioni del Calcolo "Mauro Picone",
Consiglio Nazionale delle Ricerche,
Rome, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Macroscopic Modeling
Kinetic Modeling
Road Networks
Future Directions
Bibliography

Glossary

Mathematical model Simplified description of a real world system in mathematical terms, e. g., by means of differential equations or other suitable mathematical structures.

Scale of observation/representation Level of detail at which a system is observed and mathematically modeled.

Macroscopic scale Scale of observation focusing on the evolution in time and space of some *macroscopic quantities* of the system at hand; that is, quantities that can be experimentally observed and measured.

Kinetic scale Scale of representation focusing on a statistical description of the evolution of the microscopic states of the system at hand.

Vehicular traffic The overall dynamics produced by cars or other transports in motion along a road.

Road network A set of roads connected to one another by junctions.

Definition of the Subject

Vehicular traffic is attracting a growing scientific interest because of its connections with other important problems, like, e. g., environmental pollution and congestion of cities. Rational planning and management of vehicle fluxes are key topics in modern societies under both economical and social points of view, as the increasing number of projects aimed at monitoring the quality of the road traffic demonstrates. In spite of their importance, however, these issues cannot be effectively handled by simple experimental approaches. On the one hand, observation and data recording may provide useful information on the physics of traffic, highlighting some typical features like, e. g., clustering of the vehicles, the appearance of stop-and-go waves, the phase transition between the regimes of free and congested flow, the trend of the traffic in uniform flow conditions. The books by Kerner [48] and Leutzbach [54] extensively report about traffic phenomena, real traffic data, and their phenomenological interpretation. On the other hand, processing and organization of (usually huge amounts of) experimental measurements hardly allow one to catch the real unsteady traffic dynamics, which definitely makes this approach scarcely predictive. Therefore vehicular traffic is not only an engineering matter but also a challenging mathematical problem.

This paper reviews some of the major macroscopic and kinetic mathematical models of vehicular traffic available in the specialized literature. Application to road networks is also discussed, and an extensive list of references is provided to both the original works presented here and other sources for further details. Microscopic models are only briefly recalled in the introduction for the sake of thoroughness, as they are actually beyond the scope of the review. However, suitable references are included for the interested readers.

Introduction

The mathematical modeling of vehicular traffic requires, first of all, the choice of the scale of representation. The

relevant literature offers many examples of models at any scale, from the microscopic to the macroscopic through the kinetic one. Each of them implies some technical approximations, and suffers therefore from related drawbacks, either analytical or computational.

The microscopic scale is like a magnifying glass focused on each single vehicle, whose dynamics is described by a system of ordinary second-order differential equations of the form

$$\ddot{x}_i = a_i\left[t, \{x_k\}_{k=1}^N, \{\dot{x}_k\}_{k=1}^N\right], \quad i = 1, \ldots, N, \tag{1}$$

where t is time, $x_i = x_i(t)$ the scalar position of the ith vehicle along the road, $\dot{x}_i = \dot{x}_i(t)$ its velocity, and $N \in \mathbb{N}$ the total number of vehicles. The function a_i describes the acceleration of the ith vehicle, which in principle might be influenced by the positions and velocities of all other vehicles simultaneously present on the road. As a matter of fact, each vehicle is commonly assumed to be influenced by its heading vehicle only (*follow-the-leader models*), so that the acceleration a_i depends at most on x_i, x_{i+1} and on $\dot{x}_i$, $\dot{x}_{i+1}$. It is immediately seen that the size of system (1) rapidly increases with the number N of vehicles considered, which frequently makes the microscopic approach not competitive for computational purposes. Furthermore, from the analytical point of view it is often difficult to investigate the relevant global features of the system, also in connection with control and optimization problems. In this paper we will not be concerned specifically with microscopic modeling of vehicular traffic. The interested reader can find further details on the subject in Aw et al. [4], Chakroborty et al. [20], Gazis et al. [36], Helbing [41,42], Kerner and Klenov [49], Nagel et al. [56] Treiber et al. [66,67,68] and in the main references listed therein, as well as in the review by Hoogendoorn and Bovy [46].

Macroscopic and kinetic models aim at describing the big picture without looking specifically at each single subject of the system, hence they are computationally more efficient: Few partial differential equations, that can be solved numerically in a feasible time, are normally involved, and the global characteristics of the system are readily accessible. Nevertheless, now the modeling is, in a sense, less accurate than in the microscopic case, due to the fact that both approaches rely on the *continuum hypothesis*, clearly not physically satisfied by cars along a road: The number of vehicles should be large enough so that it makes sense to introduce concepts like macroscopic density, average speed, or kinetic distribution function as continuous functions of space and, in the latter case, also velocity. However, the continuum hypothesis can be profitably accepted as a technical approximation of the physical reality, regarding macroscopic quantities globally as measures of traffic features and as tools to depict the spatial and temporal evolution of traffic waves.

Macroscopic modeling is based on the idea, originally due in the fifties to Lighthill and Whitham, and independently, to Richards, that the classical Euler and Navier-Stokes equations of fluid dynamics describing the flow of fluids could also describe the motion of cars along a road, provided a large-scale point of view is adopted so as to consider cars as small particles and their density as the main quantity at which one looks. This analogy remains nowadays in all macroscopic models of vehicular traffic, as terms like traffic pressure, traffic flow, traffic waves demonstrate.

Kinetic modeling, first used by Prigogine in the sixties, relies instead on the principles of statistical mechanics introduced by Boltzmann to describe the unsteady evolution of a gas. The kinetic approach as well requires in principle a large number of particles in the system but, unlike the macroscopic approach, it is based on a microscopic modeling of their mutual interactions. The evolution of the system is described by means of a distribution function over the microscopic state of the particles, whence the interesting macroscopic quantities are obtained via suitable microscopic average procedures.

In this review we present and discuss some of the main continuum mathematical models of vehicular traffic available in the literature (see also Bellomo et al. [10]). Specifically, the paper is organized in four more sections that follow this Introduction. Section "Macroscopic Modeling" is devoted to macroscopic models of both first and second order, which describe the gross evolution of traffic via either the sole mass conservation equation supplemented by suitable closure relations or a coupled system of mass conservation and momentum balance equations. Section "Kinetic Modeling" reports about traffic modeling by methods of the kinetic theory, from the very first attempts recalled above to the more recent contributions in the field based on the discrete kinetic theory. These two sections account mainly for traffic models on single one-way roads, whereas Sect. "Road Networks" finally deals with traffic problems on road networks. Such topic provides, in the present context, a strong link between mathematical modeling and real world applications, as most prediction, control, and optimization problems concern precisely the management of car fluxes over complex systems of interconnected roads, e. g., urban networks or some specific sub-networks. In addition, it poses nontrivial theoretical questions, which motivate the de-

velopment of innovative mathematical methods and tools susceptible of application also to different fields. These sections of the paper are conceived so as to give first the general mathematical framework, which is then detailed to obtain particular models. Section "Future Directions" finally sketches some research perspectives in the field.

Macroscopic Modeling

In the macroscopic approach to the modeling of vehicular traffic the flow of cars along a road is assimilated to the flow of fluid particles, for which suitable balance or conservation laws can be written. For this reason, macroscopic models are often called in the present context *hydrodynamic models*.

In the macroscopic approach, cars are not followed individually. The point of view is rather that of the classical continuum mechanics, i. e., one looks at the evolution in time and space of some average quantities of interest such as the mass density or the mean velocity, referred to an infinitesimal reference volume individuated by a point in the geometrical space. In more detail, since the motion of cars along a road is usually modeled as one-dimensional, the spatial coordinate x is a scalar independent variable of the problem which ranges in an appropriate subset of $\mathbb{R}$, possibly in the whole real axis.

The main dependent variables introduced to describe the problem mathematically are the density of cars $\rho = \rho(t, x)$ and their average velocity $u = u(t, x)$ at time t in the point x. From these quantities another important variable is derived, namely the flux $q = q(t, x)$ given by $q = \rho u$, which is of great interest for both theoretical and experimental purposes. Indeed, many real data, used to construct and to validate macroscopic models, refer precisely to flux measurements, because these can be performed with higher precision than measurements of other variables.

Since, as said before, macroscopic models of vehicular traffic are inspired to the continuum equations of fluid dynamics, the Eulerian point of view is normally adopted, meaning that the above dependent variables evaluated at (t, x) yield the evolution at time t of the vehicles flowing through the fixed in space position x. In other words, the spatial coordinate is not linked to any reference configuration but rather to the actual geometrical space, hence at different times the values of the state variables computed in the point x refer in general to different continuum particles flowing through x. Of course, all models can in principle be converted in Lagrangian form, up to a proper change of variable.

The basic evolution equation translates the principle of conservation of the vehicles:

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho u) = 0 \,. \tag{2}$$

According to this equation, the time variation in the amount of cars within any stretch of road comprised between two locations $x_1 < x_2$ is only due to the difference between the incoming flux at x_1 and the outgoing flux at x_2. Integration of Eq. (2) over the interval $[x_1, x_2]$ produces indeed

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{x_1}^{x_2} \rho(t, x) \mathrm{d}x = q(t, x_1) - q(t, x_2) \,,$$

where the integral at the left-hand side defines precisely the mass of cars contained in $[x_1, x_2]$ at time t.

It can be questioned that Eq. (2) does not give rise by itself to a self-consistent mathematical model, as it involves simultaneously two variables, ρ and u. To overcome this difficulty, one possibility is to devise suitable closures which allow one to express the velocity u, or equivalently the flux q, as a function of the density ρ. This way Eq. (2) becomes an evolution equation for the sole density ρ:

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x} f(\rho) = 0 \,, \tag{3}$$

where $q = f(\rho)$ is the closure relation. Macroscopic models based on Eq. (3) are usually called *first order models*.

We briefly recall here that a diffusion term is sometimes added to the right-hand side of Eq. (3) in order to smooth the resulting density field ρ, which, from the theory of nonlinear conservation laws (see e. g., Bressan [18], Dafermos [27], Serre [63,64]), is known to possibly develop discontinuities in finite time even for smooth initial data. The resulting equation reads

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x} f(\rho) = \epsilon \frac{\partial}{\partial x} \left(\kappa(\rho) \frac{\partial \rho}{\partial x} \right) \,, \tag{4}$$

where $\epsilon > 0$ is a parameter accounting for the strength of the diffusion with respect to the other terms of the equation, and $\kappa(\rho)$ is the (possibly nonlinear) diffusion coefficient. Notice that if κ is a positive constant then Eq. (4) reduces formally to the linear heat equation with nonlinear advection. However, unlike Eq. (3), models based on Eq. (4) do not preserve in general the total amount of vehicles.

A second possibility consists instead in joining Eq. (2) to an evolution equation for the flux q inspired by the momentum balance of a continuum:

$$\frac{\partial q}{\partial t} + \frac{\partial}{\partial x}(qu) = A[\rho, u, D\rho, Du] \,, \tag{5}$$

where A is some material model for the generalized forces acting on the system and responsible for momentum variations. These forces express the macroscopic outcome of the microscopic interactions among the vehicles; as we will see later, in most models they are assumed to be determined by either the local density ρ or the local velocity u of the cars, as well as by their respective variations in time and/or space which in the above equation are denoted by the operator D. Using $q = \rho u$ and taking Eq. (2) into account, Eq. (5) can be further manipulated to obtain

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = a[\rho, u, D\rho, Du] \,, \tag{6}$$

where $a := A/\rho$ is a material model for the acceleration of the vehicles to be specified. Finally, the set of equations resulting from this second approach is

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \dfrac{\partial}{\partial x}(\rho u) = 0 \\[2ex] \dfrac{\partial u}{\partial t} + u\dfrac{\partial u}{\partial x} = a[\rho, u, D\rho, Du] \end{cases} \tag{7}$$

which is self-consistent in the unknowns ρ, u once the acceleration a has been conveniently designed. Mathematical models based on the system (7) are commonly termed in the literature *second order models* or, less frequently, *two-equations models*.

Lighthill–Whitham–Richards and Other First Order Models

First order models are obtained from the conservation law (3) by resorting to a suitable closure relation which expresses the flow of cars q as a function of the density ρ. A relation of the form $q = f(\rho)$ is named in this context *fundamental diagram*.

If $\rho_{\max} > 0$ denotes the maximum vehicle density allowed along the road according, for instance, to the capacity, i. e., the maximum sustainable occupancy of the latter, the function f is often required

1. to be monotonically increasing from $\rho = 0$ up to a certain density value $\sigma \in (0, \rho_{\max})$;
2. to be decreasing for $\sigma \le \rho \le \rho_{\max}$;
3. to have $\rho = \sigma$ as unique maximum point in $[0, \rho_{\max}]$;
4. to be concave in the interval $[0, \rho_{\max}]$.

These assumptions aim mainly at reproducing the typical behavior of experimentally computed fundamental diagrams (see Fig. 1 and, e. g., Bonzani and Mussone [16,17], Kerner [47,48,49]). In addition, some of them, for instance

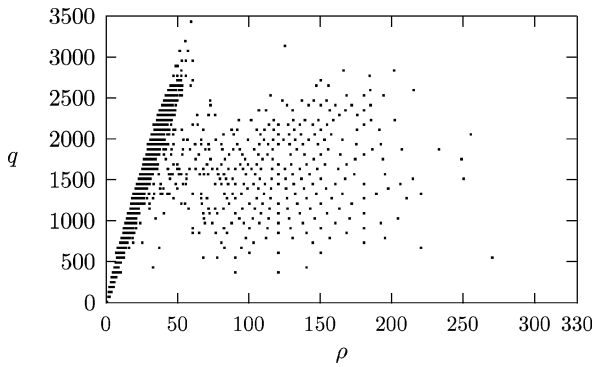

Vehicular Traffic: A Review of Continuum Mathematical Models, Figure 1
Fundamental diagram based on experimental data referring to one-week traffic flow in viale del Muro Torto, Rome, Italy. Measures kindly provided by ATAC S.p.A.

assumptions (3) and (4), are stated for theoretical purposes linked to the qualitative properties of the conservation law (3). We observe that sometimes assumption (4) is relaxed, allowing the function f to possibly change concavity in the decreasing branch.

By prescribing the flux f one implicitly assigns also the relation between the average velocity u and the density ρ (*velocity diagram*), indeed using $q = \rho u$ one has

$$u(\rho) = \frac{f(\rho)}{\rho} \,.$$

Since the function f often vanishes for $\rho = 0$, this formula commonly determines a finite value for u in the limit $\rho \to 0^+$.

The prototype of all fluxes complying with the above assumptions is the parabolic profile firstly proposed by Lighthill and Whitham [55,71] and then, independently, by Richards [62], which gives rise to the so-called LWR model:

$$f(\rho) = u_{\max}\rho\left(1 - \frac{\rho}{\rho_{\max}}\right) \,, \tag{8}$$

from which the following velocity diagram is obtained:

$$u(\rho) = u_{\max}\left(1 - \frac{\rho}{\rho_{\max}}\right) \tag{9}$$

(see Fig. 2a). The parameter $u_{\max} > 0$ identifies the maximum velocity of the cars in a situation of completely free road ($\rho = 0$). Notice that Eq. (8) assumes zero flux of vehicles when both $\rho = 0$ and $\rho = \rho_{\max}$; correspondingly, the average velocity given by Eq. (9) is maximum in the first case and vanishes in the second case.

A generalization of Eqs. (8), (9) is provided by the Greenshield model (Fig. 2b):

$$f(\rho) = u_{\max}\rho\left[1-\left(\frac{\rho}{\rho_{\max}}\right)^n\right],$$
$$u(\rho) = u_{\max}\left[1-\left(\frac{\rho}{\rho_{\max}}\right)^n\right] \quad (n \in \mathbb{N}),$$

while an example of fundamental diagram satisfying all the previous assumptions but which generates an unbounded velocity diagram for $\rho \to 0^+$ is due to Greenberg (Fig. 2c):

$$f(\rho) = u_0\rho\log\left(\frac{\rho_{\max}}{\rho}\right),$$
$$u(\rho) = u_0\log\left(\frac{\rho_{\max}}{\rho}\right),$$

where $u_0 > 0$ is a tuning parameter. These formulas were specifically conceived to fit some experimental data on the traffic flow in the Lincoln tunnel in New York.

Additional fundamental and velocity diagrams, for which some of the above discussed assumptions fail, are reported in the book by Garavello and Piccoli [34]. Moreover, other closure relations of the mass conservation equation for first order models are discussed in Bellomo and Coscia [8]. Here we simply mention a further diagram due to Bonzani and Mussone [17] (Fig. 2d):

$$f(\rho) = u_{\max}\rho e^{-\alpha\frac{\rho}{\rho_{\max}-\rho}},$$
$$u(\rho) = u_{\max}e^{-\alpha\frac{\rho}{\rho_{\max}-\rho}} \tag{10}$$

which depends on a phenomenological parameter $\alpha > 0$ linked to road and environmental conditions. In particular, the lower the α, the more the flow of vehicles is favored; according to the authors, comparisons with suit-

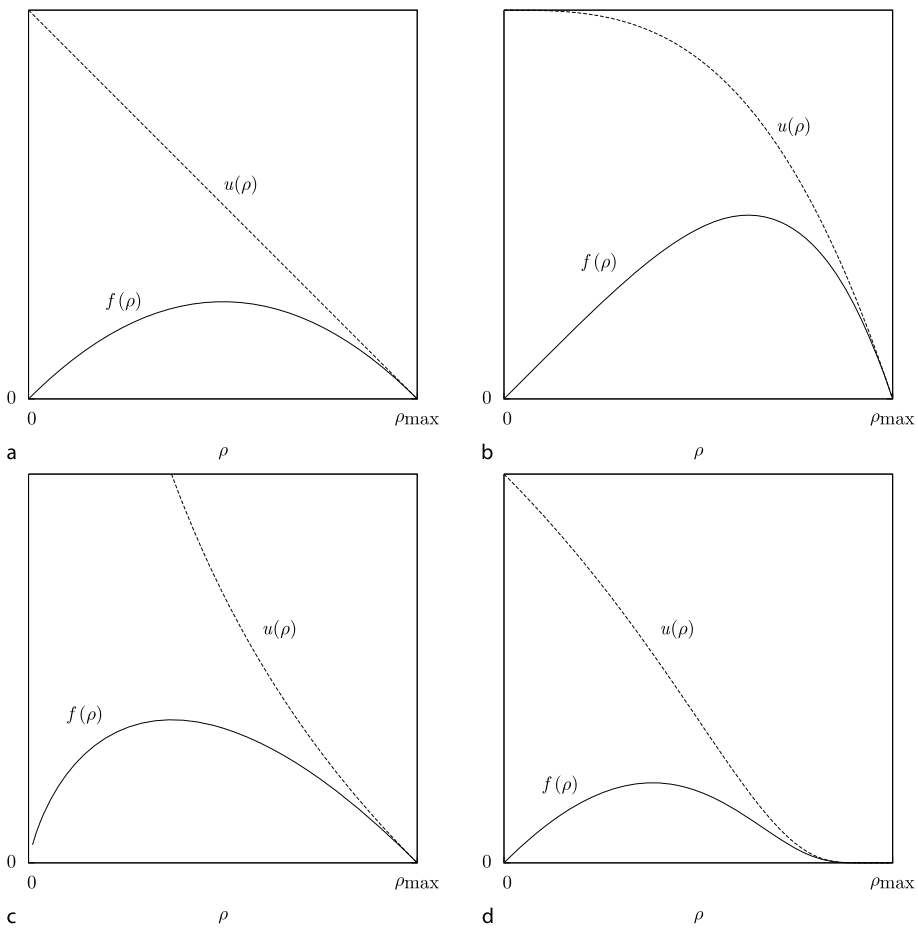

Vehicular Traffic: A Review of Continuum Mathematical Models, Figure 2
Fundamental diagrams (*solid lines*) and velocity diagrams (*dashed lines*) for a LWR model, b Greenshield's model, c Greenberg's model, and d Bonzani–Mussone's model

able sets of experimental data suggest the interval $[1, 2.5]$ as a standard range for α. Notice that in this case the function f may have an inflection point in the interval $(\sigma,\ \rho_{\max})$, hence the hypothesis (4) is potentially violated.

Fictitious Density Model

In view of the analogy with the flow of fluid particles, the macroscopic modeling framework accounts for a fully mechanical evolution of the traffic. On the other hand, it can be argued that cars are actually not completely mechanical subjects, for the presence of the drivers is likely to affect their behavior in an essentially unconventional 'human-like' fashion.

Still in the context of first order hydrodynamic models, De Angelis [32] tries to explicitly account for the action of the drivers on the evolution of the traffic by introducing the concept of *fictitious density*. The fictitious density ρ^* represents the density of cars along the road as perceived by the drivers, hence it expresses the personal feelings of the latter about the local conditions of traffic. This new variable, which is in general different from the actual density ρ, is used to model the instantaneous reactions of the drivers.

Specifically, the fictitious density is linked to the actual density and to its spatial gradient in such a way that if ρ is increasing (respectively, decreasing) then ρ^* increases (respectively, decreases) in turn but more quickly than ρ, while for ρ approaching the limit threshold $\rho_{\max}$ also ρ^* tends to saturate to $\rho_{\max}$. In other words, the sensitivity of the drivers should consist in feeling a fictitious congestion state of the road higher or lower than the real one, and by consequence in anticipating certain behaviors that a purely mechanical agent would exhibit only in presence of the appropriate external conditions.

The dependence of ρ^* on ρ and $\partial_x\rho$ proposed by De Angelis is

$$\rho^*[\rho,\ \partial_x\rho] = \rho + \eta\left(1 - \frac{\rho}{\rho_{\max}}\right)\frac{\partial\rho}{\partial x}\,, \tag{11}$$

where $\eta > 0$ is a parameter related to the sensitivity of the drivers, which weights the effect of the spatial variations of the density ρ. It can be observed that for increasing actual density, that is $\partial_x\rho > 0$, the fictitious density is larger than ρ, thus suggesting a more careful attitude of the system car-driver with respect to the purely mechanical case. Conversely, for decreasing actual density, namely $\partial_x\rho < 0$, the fictitious density is lower than ρ, which implies a more aggressive behavior induced by the presence of the driver.

Specific models are obtained by plugging Eq. (11) into the expression of the velocity diagram. This amounts essentially to devising a closure relation of the mass conservation equation (2) of the form $u = u(\rho^*)$, so that the evolution equation finally reads

$$\frac{\partial\rho}{\partial t} + \frac{\partial}{\partial x}(\rho u(\rho^*)) = 0\,.$$

In view of the dependence of ρ^* on $\partial_x\rho$, this class of models often specializes in nonlinear, possibly degenerate, parabolic equations in the unknown ρ. For instance, technical calculations with the Lighthill–Whitham–Richards diagram (9) give rise to

$$\frac{\partial\rho}{\partial t} + h(\rho)\frac{\partial\rho}{\partial x} = \frac{\eta}{\rho_{\max}}\left[f(\rho)\frac{\partial^2\rho}{\partial x^2} + h(\rho)\left(\frac{\partial\rho}{\partial x}\right)^2\right]$$

where the function h is defined by

$$h(\rho) = u_{\max}\left(1 - \frac{2\rho}{\rho_{\max}}\right)$$

while the flux f is given by Eq. (8).

Different models of fictitious density, as well as their correlations with diffusion models based on Eq. (4), are considered by Bonzani [15].

The Payne–Whitham Model

Payne [58] and Whitham's [71] model is a second order model in which the acceleration a (cf. Eq. (6)) is constructed as the sum of two terms:

$$a[\rho, u, D\rho, Du] = a_r[\rho, u] + a_a[\rho, \partial_x\rho]\,. \tag{12}$$

The former, a_r, is called *relaxation* and models the tendency of each driver to travel at a certain desired speed depending on the level of congestion of the road:

$$a_r[\rho, u] = \frac{V_e(\rho) - u}{\tau}\,, \tag{13}$$

where u is the actual speed and $V_e(\rho)$ the desired one, while $\tau > 0$ is a parameter representing the relaxation time or, in other words, the reaction time of the drivers in adjusting their actual speed to $V_e(\rho)$. The desired speed in Eq. (13) may be thought of as the speed possibly observed in stationary homogeneous traffic conditions for a specific value of the density ρ. As a matter of fact, the function $V_e(\rho)$ can be one of the velocity diagrams used in first order models (see Subsect. "Lighthill–Whitham–Richards and Other First Order Models").

The second term forming the acceleration, a_a, is named *anticipation* and accounts for the reaction of the drivers to the variations in the traffic conditions ahead of

them. In more detail, it models the tendency to accelerate or to decelerate in response to a decrease or an increase, respectively, of the density of cars, therefore it is linked to the density ρ and to its spatial gradient $\partial_x \rho$ as

$$a_a[\rho,\ \partial_x \rho] = -\frac{1}{\rho}\frac{\partial}{\partial x} p(\rho)\ , \tag{14}$$

where the function $p = p(\rho)$ is the analogous to the *pressure* in the fluid dynamics equations. A common hypothesis on this traffic pressure is that it is a nondecreasing function of the density; assuming moreover differentiability of p with respect to ρ, the anticipation term can be formally rewritten in the equivalent form

$$a_a[\rho,\ \partial_x \rho] = -\frac{p'(\rho)}{\rho}\frac{\partial \rho}{\partial x}\ , \tag{15}$$

whence one sees that the ratio $p'(\rho)/\rho$ is a kind of weight on the gradient $\partial_x \rho$, which amplifies or reduces the effect of the latter on the basis of the local density of cars along the road.

In view of Eqs. (13)–(15), the final form of Payne–Whitham's model is as follows:

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \dfrac{\partial}{\partial x}(\rho u) = 0 \\[2ex] \dfrac{\partial u}{\partial t} + u\dfrac{\partial u}{\partial x} = \dfrac{V_e(\rho) - u}{\tau} - \dfrac{p'(\rho)}{\rho}\dfrac{\partial \rho}{\partial x}\ . \end{cases} \tag{16}$$

According to the derivation of these equations proposed by Whitham himself in the book [71] as continuous approximation of a microscopic car-following model, one has the following representation of the traffic pressure:

$$p'(\rho) = \frac{1}{2\tau}\left|V_e'(\rho)\right|\ .$$

More in general, the relation between p and ρ can be expressed in the form (see e. g., Günther et al. [39], Klar and Wegener [50])

$$p(\rho) = \rho\Theta_e(\rho)\ ,$$

where Θ_e denotes the equilibrium value in homogeneous traffic conditions of the variance of the microscopic velocity of cars (cf. Eq. (25)).

For the sake of thoroughness, we recall that sometimes a diffusive acceleration term of the form

$$a_d\left[\partial_x^2 u\right] = \frac{\mu}{\rho}\frac{\partial^2 u}{\partial x^2}$$

is introduced in the decomposition (12), the parameter $\mu > 0$ representing a sort of viscosity by analogy with the theory of Newtonian fluids.

The Aw–Rascle Model

The hydrodynamic model by Payne and Whitham suffers from some severe drawbacks which led Daganzo [28] to write a celebrated 'requiem' for this kind of second order approximation of traffic flow. Flaws of the model include the possible onset of negative speeds of the vehicles as well as the violation of the so-called *anisotropy principle*, i. e., the fact that a car should be influenced only by the traffic dynamics ahead of it, being practically insensitive to what happens behind. By computing the characteristic speeds λ_1, λ_2 of the hyperbolic system (16) in the variables ρ, u one finds indeed $\lambda_{1,2} = u \pm \sqrt{p'(\rho)}$, meaning that some information generated by the motion of the vehicles travels faster than the vehicles themselves, hence affects the forward traffic dynamics.

Aw and Rascle [3] suggest that all of these inconsistencies are caused by a too direct application to cars of the analogy with fluid particles, to which, on the other hand, the principles of nonnegativity of the velocity and of anisotropy do not apply. In particular, they observe that the anticipation term a_a should not be regarded as an Eulerian quantity, as it is specifically linked to the perception of the traffic conditions by the drivers while in motion within the flow of vehicles. An external (Eulerian) observer might see an increasing vehicle density in a certain fixed spatial position x while, at the same time, a driver, i. e., an internal (Lagragian) observer, flowing through x sees a decreasing density in front of her/him if, for example, cars ahead move faster. In such a situation, the behavior predicted by the Payne–Whitham's anticipation term, that is a deceleration, is obviously inconsistent with the presumable reaction of the driver, who is instead expected to maintain the current speed if not to accelerate.

The problem is solved by Aw and Rascle by making the pressure p a material (i. e., Lagrangian) variable of the model, which technically amounts to expressing the anticipation term via the material derivative $D_t := \partial_t + u\partial_x$ of p:

$$\begin{aligned} &a_a[\rho, u, \partial_t \rho,\ \partial_x \rho] \\ &\qquad = -D_t p(\rho) = -\partial_t p(\rho) - u\partial_x p(\rho)\ . \end{aligned} \tag{17}$$

The original Aw–Rascle's (AR) model assumes in addition no relaxation term in the acceleration, thus it writes

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \dfrac{\partial}{\partial x}(\rho u) = 0 \\[2ex] \dfrac{\partial}{\partial t}(u + p(\rho)) + u\dfrac{\partial}{\partial x}(u + p(\rho)) = 0 \end{cases} \tag{18}$$

for a smooth and increasing function p of the density ρ, e. g.

$$p(\rho) = \rho^\gamma, \quad \gamma > 0 .$$

Notice that, owing to the conservation of the vehicle mass, the second equation in (18) can be regarded as the non-conservative form of a conservation equation for the generalized flux ρw, with $w := u + p(\rho)$. For this reason, this equation is frequently encountered in the literature as

$$\frac{\partial}{\partial t}(\rho w) + \frac{\partial}{\partial x}(\rho w u) = 0 . \tag{19}$$

Computing the characteristic speeds of the hyperbolic system (18) in the variables ρ, u gives $\lambda_1 = u$, $\lambda_2 = u - \rho p'(\rho)$, whence, in view of the assumption of nondecreasing pressure, one sees that no information can now be advected faster than cars.

Further extensions of the AR model involve the introduction of a relaxation term (Greenberg [38], Rascle [61]) to let drivers travel at some maximum desired speed when the traffic conditions allow. In particular, Greenberg's relaxation term takes the form

$$a_r[\rho, u] = \frac{V_m - (u + p(\rho))}{\tau} ,$$

where $V_m > 0$ is a constant value representing the maximum possible speed along the road and $\tau > 0$ is the relaxation time, while Rascle's relaxation term is more similar to the corresponding Payne–Whitham's term (cf. Eq. (13)):

$$a_r[\rho, u] = \frac{V(\rho) - u}{\tau} ,$$

though here the function $V(\rho)$ need not represent an equilibrium velocity.

The Phase Transition Model

Numerous experimental observations (see e. g., Kerner [47,48,49]) report about strongly different qualitative and quantitative features of the flow of vehicles in *free* and *congested* regimes, corresponding respectively to low and high density of cars, far from or close to the road capacity. In particular, from experimentally computed fundamental diagrams it can be inferred that the assumption that the average speed be a one-to-one function of the density applies well to free traffic, while being completely inadequate to describe congested flows. Indeed, in the latter case the density ρ no longer determines univocally the velocity u, or equivalently the flux q, of the vehicles, for relevant fluctuations are recorded so that the admissible pairs (ρ, q) are scattered over a two-dimensional region of the state space. In the specialized literature, this phenomenon is called a *phase transition*.

Starting from these experimental facts, Colombo [23, 24] proposed a mathematical model in which the existence of the phase transition is postulated and accounted for by splitting the state space (ρ, q) in two regions, Ω_f and Ω_c, corresponding to the regimes of free and congested flow, respectively. In Ω_f a classical relation of the form $u = u(\rho)$ is imposed, and the model simply consists in the conservation law (3) with $f(\rho) = \rho u(\rho)$. In particular, the LWR velocity diagram (9) is chosen. Conversely, in Ω_c the density ρ and the flux q are regarded as two independent state variables, and the model is similar to a second order model in which the mass conservation equation (2) is joined to an evolution equation for the flux of the form

$$\frac{\partial q}{\partial t} + \frac{\partial}{\partial x}((q - q_*)u) = 0 . \tag{20}$$

Furthermore, the average velocity is specified as a function of both ρ and q which, in a way, generalizes the LWR velocity diagram:

$$u(\rho, q) = \frac{q}{\rho}\left(1 - \frac{\rho}{\rho_{\max}}\right) . \tag{21}$$

Notice that Eq. (20) closely recalls the momentum balance (5) of second order models with zero external force A. The parameter $q_* > 0$ is introduced to obtain that at high car density and high car speed breaking produces shocks and accelerating produces rarefaction waves, while at low car density and speed the opposite occurs (see [23] for further details).

Finally, Colombo's phase transition model writes:

Free flow: $(\rho, q) \in \Omega_f$

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \dfrac{\partial}{\partial x}(\rho u) = 0 \\ u = u(\rho) , \end{cases}$$

Congested flow: $(\rho, q) \in \Omega_c$

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \dfrac{\partial}{\partial x}(\rho u) = 0 \\ \dfrac{\partial q}{\partial t} + \dfrac{\partial}{\partial x}((q - q_*)u) = 0 \\ u = u(\rho, q) . \end{cases}$$

The regions Ω_f, Ω_c are defined so as to be positively invariant with respect to their corresponding systems of equations: If the initial datum is entirely contained in

one of the two regions, then the solution will remain in that region at all successive times of existence. In other words, if the traffic is initially free (respectively, congested), then it will remain free (respectively, congested) during the whole subsequent evolution. Technically, letting $\Omega = [0, \rho_{\max}] \times [0, +\infty) \ni (\rho, q)$ denote the state space:

$$\Omega_f = \{(\rho, q) \in \Omega \,:\, u(\rho) \geq \hat{V}_f,\ q = u_{\max}\rho\},$$
$$\Omega_c = \left\{(\rho, q) \in \Omega \,:\, u(\rho, q) \leq \hat{V}_c,\ \frac{Q_1 - q_*}{\rho_{\max}} \leq \frac{q - q_*}{\rho} \leq \frac{Q_2 - q_*}{\rho_{\max}}\right\},$$

where $\hat{V}_f$, $\hat{V}_c$ are two fixed speed thresholds such that above the former the flow is free while below the latter the flow is congested, with furthermore $\hat{V}_c < \hat{V}_f \leq u_{\max}$ in order for the two phases not to overlap, and $Q_1 \in (0, q_*)$, $Q_2 \in (q_*, +\infty)$ are tuning parameters related to various environmental conditions.

Kinetic Modeling

The kinetic approach to the modeling of vehicular traffic is based on the choice of an intermediate representation scale between the macroscopic and the microscopic ones, technically called *mesoscopic*. Rather than looking at each single car of the system, like in the microscopic approach, a *distribution function*, usually denoted by f, is introduced over the microscopic state (x, v) of the vehicles, x standing for position and v for velocity assumed one-dimensional with moreover $v \geq 0$. The distribution function is such that the quantity

$$f(t, x, v)\mathrm{d}x\mathrm{d}v$$

represents a measure of the number of vehicles whose position at time t is comprised between x and $x + dx$, with a velocity comprised between v and $v + dv$. Consequently, under suitable integrability assumptions on f, if $D_x \subseteq \mathbb{R}$, $D_v \subseteq \mathbb{R}_+$ denote the spatial domain and the velocity domain, respectively, the quantity

$$N(t) = \int_{D_v}\int_{D_x} f(t, x, v)\mathrm{d}x\mathrm{d}v \tag{22}$$

gives the total amount of vehicles present in the system at time t. It is worth pointing out that f is in general not a probability distribution function over (x, v), as the integral in Eq. (22) may not equal 1 at all times. However, if the total mass of cars is preserved then $N(t)$ is constant in t, hence it is possible to rescale the distribution function so that it have a unit integral on $D_x \times D_v$ for all times t of existence.

The knowledge of the distribution function enables one to recover the usual macroscopic variables of interest as local averages over the microscopic states of the vehicles. For instance, the density ρ of cars at time t in the point x is given by

$$\rho(t, x) = \int_{D_v} f(t, x, v)\mathrm{d}v\,, \tag{23}$$

while the flux q and the average velocity u are obtained from the first momentum of f with respect to v as

$$q(t, x) = \int_{D_v} v f(t, x, v)\mathrm{d}v, \quad u(t, x) = \frac{q(t, x)}{\rho(t, x)}\,. \tag{24}$$

Higher order momenta are related to other macroscopic variables, such as the average kinetic energy and the variance of the velocity. In particular, the latter is given by

$$\Theta(t, x) = \frac{1}{\rho(t, x)}\int_{D_v} (v - u(t, x))^2 f(t, x, v)\mathrm{d}v \tag{25}$$

and is often linked, at a macroscopic level, to the traffic pressure responsible for the anticipation terms in the hydrodynamic equations of traffic (see e. g., the Payne–Whitham's model, Subsect. "The Payne–Whitham Model", and Günther et al. [39] or Klar and Wegener [50] for further details).

Modeling of the system is obtained by stating an evolution equation in time and space for the distribution function f. Unlike the macroscopic approach, such an equation is written considering the influence of microscopic interactions among the vehicles on the microscopic states. Concerning this, it is useful to introduce a specific terminology to distinguish in an interaction the vehicle which is likely to change its state from the one which potentially causes such a change. The former is technically called the *candidate vehicle*, while the latter is called the *field vehicle*. Practically all kinetic models of vehicular traffic present in the literature describe the interactions by appealing to the following general guidelines:

1. Cars are regarded as points, their dimensions being negligible.
2. Interactions are binary, in the sense that those involving simultaneously more than two vehicles are disregarded.
3. Interactions modify by themselves only the velocity of the vehicles, not their positions. More specifically, in an interaction the sole velocity of the candidate vehicle may vary, but that of the field vehicle remains instead unchanged.

4. Vehicles are anisotropic particles which react mainly to frontal than to rear stimuli (see Daganzo [28]), therefore candidate vehicles only interact with field vehicles located ahead of them.
5. There exists a *probability of passing* P such that when a candidate vehicle encounters a slower field vehicle it may overtake it instantaneously with probability P without modifying its own velocity.
6. Interactions are conservative, in the sense that they preserve the total number of vehicles of the system.

The evolution equation for the distribution function is obtained from a balance principle which states that the time variation in the number of vehicles belonging to any subset of the state space $D_x \times D_v$ is determined by the state transitions due to the interactions of the vehicles with each other. Therefore, in the absence of external actions on the system it writes

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} = J[f] \,, \tag{26}$$

where $J[f] = J[f](t, x, v)$ is an operator acting on the distribution function f, charged to describe the interactions and their effects on the states of the vehicles. Following the classical terminology coming from the collisional kinetic theory of the gas dynamics (see e. g., Villani [69]), J is frequently termed the *collisional operator*, though in the present context this is a slight abuse of speech because interactions among vehicles are not like collisions among mechanical particles.

In view of the assumption (6) above, the operator J is required to satisfy

$$\int_{D_v} J[f](t, x, v)\mathrm{d}v = 0, \quad \forall x \in D_x, \quad \forall t \geq 0$$

so that integrating Eq. (26) with respect to v and recalling Eqs. (23), (24) yields the macroscopic mass conservation equation (2).

Specific models are obtained from Eq. (26) by detailing the form of the collisional operator.

The Prigogine Model

In his pioneering work on the mathematical modeling of vehicular traffic by methods of the kinetic theory, Prigogine [59,60] constructs the collisional operator through the contribution of two terms:

$$J[f] = J_r[f] + J_i[f_2] \,. \tag{27}$$

The first one, $J_r[f]$, is called *relaxation term* and models the tendency of each driver to adapt the state of her/his vehicle to a desired standard state. The latter is described by the *desired distribution function* $f_0 = f_0(t, x, v)$, which in essence corresponds to the driving program each driver aims at. In Prigogine's model, the function f_0 is expressed as

$$f_0(t, x, v) = \rho(t, x)\tilde{f}_0(v) \,, \tag{28}$$

where $\tilde{f}_0$ is a prescribed probability distribution over the variable v, independent of x and t. The factorization (28) translates the hypothesis that the desired distribution of the velocity is, for each driver, unaffected by the local vehicle density. The relaxation term is then taken to be

$$J_r[f] = -\frac{f - f_0}{T} \,,$$

where T is the relaxation time which depends on the probability of passing P as

$$T = \tau\frac{1-P}{P} \quad \text{for} \quad P = 1 - \frac{\rho}{\rho_{\max}} \,. \tag{29}$$

In these formulas, τ is a positive parameter of the model and $\rho_{\max}$ denotes the maximum vehicle density locally allowed along the road according to the road capacity.

The second term appearing in the decomposition (27), $J_i[f_2]$, is called *interaction term* and is represented by an operator which models the interactions among candidate and field vehicles. As usual in the kinetic framework, it is further split into two more operators:

$$J_i[f_2] = G[f_2] - L[f_2] \,.$$

The *gain operator* $G[f_2]$ accounts for the interactions of candidate vehicles having velocity $v_* > v$ with field vehicles with velocity v, which force the former to slow down to v if they cannot overtake the latter, causing this way a gain of cars with velocity state v:

$$G[f_2] = (1-P)\int_v^{+\infty} (v_* - v) f_2(t, x, v_*, x, v)\mathrm{d}v_* \,. \tag{30}$$

The *loss operator* $L[f_2]$ accounts instead for the interactions of candidate vehicles having velocity v with field vehicles with velocity $v_* < v$, which force the former to slow down to v_* if they cannot overtake the latter, thus giving rise to a loss of cars with velocity state v:

$$L[f_2] = (1-P)\int_0^{v} (v - v_*) f_2(t, x, v, x, v_*)\mathrm{d}v_* \,. \tag{31}$$

In Eq. (30) it is assumed $D_v = \mathbb{R}_+$, that is no upper limitation is imposed on the velocity of the cars.

The function f_2 appearing in Eqs. (30), (31) is the joint distribution function of candidate and field vehicles, such that $f_2(t, x, v, x_*, v_*)$ gives a measure of the joint probability to find a candidate vehicle in the state (x, v) and simultaneously a field vehicle in the state (x_*, v_*). Introducing the hypothesis of *vehicular chaos* (see e. g., Hoogendoorn and Bovy [46]), which states that vehicles are actually uncorrelated due to the mixing caused by overtaking, one can express the function f_2 in terms of f as

$$f_2(t, x, v, x_*, v_*) = f(t, x, v) f(t, x_*, v_*) ,$$

so that J_i takes the form of a bilinear operator acting on f:

$$J_i[f, f] = (1 - P) f(t, x, v) \int_0^{+\infty} (v_* - v) f(t, x, v_*) \mathrm{d}v_*$$

and Prigogine's model finally reads

$$\frac{\partial f}{\partial t} + v \frac{\partial f}{\partial x} = -\frac{f(t, x, v) - f_0(t, x, v)}{T} + (1 - P) f(t, x, v) \int_0^{+\infty} (v_* - v) f(t, x, v_*) \mathrm{d}v_* . \quad (32)$$

Notice that the forms of G and L in Eqs. (30), (31) implicitly assume localized interactions ($x = x_*$) like in the classical Boltzmann collisional kinetic theory.

The Paveri Fontana Model

One of the main criticisms to Prigogine's model is that the desired speed distribution function $\tilde{f}_0$ in Eq. (28) is prescribed *a priori*, and is thus independent of the evolution of the system. In order to correct this drawback, Paveri Fontana [57] conceived a model in which the desired speed is taken into account as a further state variable w ranging in the same domain D_v of the true speed v. At the same time, a *generalized distribution function* $g = g(t, x, v, w)$ is introduced such that

$$g(t, x, v, w) \mathrm{d}x \mathrm{d}v \mathrm{d}w$$

gives a measure of the number of vehicles that at time t have a position comprised between x and $x + \mathrm{d}x$, a true speed comprised between v and $v + \mathrm{d}v$, and a desired speed comprised between w and $w + \mathrm{d}w$. Notice that the distribution function f and the desired distribution function f_0 are readily recovered by integrating out w and v, respectively, from g:

$$f(t, x, v) = \int_{D_v} g(t, x, v, w) \mathrm{d}w, \qquad f_0(t, x, w) = \int_{D_v} g(t, x, v, w) \mathrm{d}v. \quad (33)$$

Using these relations, an evolution equation for g is straightforwardly derived from the corresponding equation (26) for f:

$$\frac{\partial g}{\partial t} + v \frac{\partial g}{\partial x} = I[g] , \quad (34)$$

where I is a new collisional operator acting on the generalized distribution function.

Analogously to Eq. (27), Paveri Fontana's choice consists in splitting I in a twofold contribution:

$$I[g] = I_r[g] + I_i[g, g] ,$$

where the first term is again a relaxation toward the desired speed having now the form

$$I_r[g] = -\frac{\partial}{\partial v} \left(\frac{w - v}{T} g \right) , \quad (35)$$

while the second one describes the interactions among the vehicles. Specifically, in the same spirit as Prigogine's model (cf. Eqs. (30), (31)), gain and loss terms can be identified such that

$$I_i[g, g] = G[g, g] - L[g, g]$$

with

$$G[g, g] = (1 - P) \int_v^{+\infty} (v_* - v) f(t, x, v) g(t, x, v_*, w) \mathrm{d}v_* , \quad (36)$$

$$L[g, g] = (1 - P) \int_0^v (v - v_*) f(t, x, v_*) g(t, x, v, w) \mathrm{d}v_* , \quad (37)$$

the probability of passing P being defined as

$$P = \left(1 - \frac{\rho}{\rho_{\max}}\right) H\left(1 - \frac{\rho}{\rho_c}\right)$$

where $H(\cdot)$ is the Heaviside function and $\rho_c \in (0, \rho_{\max})$ a critical density threshold above which overtaking is inhibited. Notice in particular that candidate vehicles are described via the generalized distribution function g in order to take their desired speed into account, while field

vehicles are described by means of the distribution function f, in which the dependence on w has been integrated out, to focus rather on their actual speed. Moreover, Eqs. (36), (37) still assume localized interactions as confirmed by the fact that both f and g are evaluated at the same spatial position x.

Putting Eqs. (34)–(37) together we deduce the final form of Paveri Fontana's model:

$$\frac{\partial g}{\partial t} + v\frac{\partial g}{\partial x} = -\frac{\partial}{\partial v}\left(\frac{w-v}{T}g\right) + (1-P)\left[f(t,x,v)\int_v^{+\infty}(v_*-v)g(t,x,v_*,w)\mathrm{d}v_* - g(t,x,v,w)\int_0^v (v-v_*)f(t,x,v_*)\mathrm{d}v_*\right]. \quad (38)$$

Integrating Eq. (38) over w and using the first of Eqs. (33) gives an evolution equation for the distribution function f:

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} = -\frac{\partial}{\partial v}\left[\frac{1}{T}\int_0^{+\infty} wg(t,x,v,w)\mathrm{d}w - \frac{vf(t,x,v)}{T}\right] + (1-P)f(t,x,v)\int_0^{+\infty}(v_*-v)f(t,x,v_*)\mathrm{d}v_*\,.$$

This equation differs from the corresponding Eq. (32) of Prigogine's model only for the relaxation term at the right-hand side, which here maintains the explicit dependence on the generalized distribution function g.

Similarly, integration of Eq. (38) with respect to v, along with the second of Eqs. (33), yields the following evolution equation for the desired distribution function f_0:

$$\frac{\partial f_0}{\partial t} + \frac{\partial}{\partial x}\left[\int_0^{+\infty} vg(t,x,v,w)\mathrm{d}v\right] = 0$$

which definitely confirms that f_0 depends now on the overall evolution of the system.

Enskog-Like Models

As noticed by some authors (see e. g., Klar and Wegener [50]), the localized interactions framework used in classical kinetic models of traffic (like Prigogine's and Paveri Fontana's ones) prevents backward propagation of the perturbations in the negative x direction. This is essentially due to the fact that the flow of vehicles is unidirectional, given the positivity constraint on the velocity v. On the other hand, the localized interactions assumption is a heritage of the Boltzmann collisional kinetic theory, where however the aforementioned drawback is not present because the flow of gas particles need not be unidirectional.

To obviate this difficulty of the theory, Klar and coworkers suggest, in a series of papers [39,40,50,51,70] focusing among other things on this topic, see also the review by Klar and Wegener [52], to describe the collisional operator J of the kinetic equation (26) as:

$$J[f] = G[f,f] - fL[f]$$

where the gain operator $G[f,f]$ is given by

$$G[f,f] = \sum_{j=1}^{M}\iint_{\Omega_j}|v_1-v_2|\sigma_j(v;v_1,v_2,\rho) \cdot f(t,x,v_1)f(t,x+h_j,v_2)k(h_j,\rho)\mathrm{d}v_1\mathrm{d}v_2 \quad (39)$$

and the loss operator by

$$L[f] = \sum_{j=1}^{M}\int_{\omega_j}|v-v_2|f(t,x+h_j,v_2)k(h_j,\rho)\mathrm{d}v_2\,. \quad (40)$$

Notice that the interacting pairs are not supposed to occupy the same spatial position; in particular, the field vehicle is assumed to be located at $x+h_j$, x being the position of the candidate vehicle. The thresholds $h_j > 0$, $j = 1,\ldots,M$, are introduced to delocalize the interactions, similar to an Enskog-like kinetic setting, and to trigger different behaviors of the candidate vehicle (acceleration, deceleration) on the basis of the distance separating it from its leading field cars. Integration is performed over subsets of the velocity domain, $\Omega_j \subseteq D_v^2, \omega_j \subseteq D_v$, associated with each of the thresholds h_j. Finally, the function k weights the interactions according to the distance of the interacting pairs and the local congestion of the road. Its presence in the collisional operator can be formally justified by referring to the Enskog kinetic equations of a dense gas. The interested reader might want to consult the book by Bellomo et al. [9] and the paper by Cercignani and Lampis [19] for further details on this part of the theory.

In the simplest case, two constant thresholds h_1, h_2 are considered to model, respectively, a deceleration of the candidate vehicle when the distance from the field vehicle falls below h_1, and an acceleration when it grows instead

above h_2. Correspondingly, the deceleration domains in G and L are given by

$$\Omega_1 = \left\{(v_1, v_2) \in D_v^2 : v_1 > v_2\right\},$$
$$\omega_1 = \{v_2 \in D_v : v_2 > v\},$$

while the acceleration domains are

$$\Omega_2 = \left\{(v_1, v_2) \in D_v^2 : v_1 < v_2\right\},$$
$$\omega_2 = \{v_2 \in D_v : v_2 < v\}.$$

Specifically, in the case of a deceleration it is assumed that the post-interaction velocity v of the candidate vehicle either remains unchanged, thus equal to the initial velocity v_1, if an overtaking of the field vehicle is possible, or stochastically reduces to a certain fraction of the velocity v_2 of the field vehicle, if the candidate is forced to queue. This is expressed by the following form of the *transition probability* σ_1:

$$\sigma_1(v; v_1, v_2, \rho) = P\delta(v-v_1)+(1-P)\frac{1}{(1-\beta)v_2}\chi_{[\beta v_2, v_2]}(v),$$

where P is the probability of passing defined like in Eq. (29), which carries the dependence of σ_1 on ρ, $\beta \in [0, 1]$ is a parameter, and χ_I is the indicator function of the set I.

Conversely, in the case of an acceleration the post-interaction velocity v of the candidate vehicle is assumed to be uniformly distributed between the values v_1 and $v_1 + \alpha(v_{\max} - v_1)$, where α depends on ρ through P as $\alpha = \alpha_0 P$ for a suitable parameter $\alpha_0 > 0$, and $v_{\max}$ is the maximum allowed velocity on the road. Therefore, the transition probability associated to this second threshold takes the form

$$\sigma_2(v; v_1, v_2, \rho) = \frac{1}{\alpha(v_{\max} - v)}\chi_{[v_1, v_1+\alpha(v_{\max}-v)]}(v).$$

The final form of Klar and coworkers' model with two equal interaction thresholds ($h_1 = h_2 \equiv h$) writes:

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} = \sum_{j=1}^{2}\iint_{\Omega_j} |v_1 - v_2|\sigma_j(v; v_1, v_2, \rho)$$
$$\cdot f(t, x, v_1)f(t, x + h, v_2)k(h, \rho)\mathrm{d}v_1\mathrm{d}v_2$$
$$-f(t, x, v)\sum_{j=1}^{2}\int_{\omega_j} |v-v_2|f(t, x+h, v_2)k(h, \rho)\mathrm{d}v_2.$$

We observe that the spirit in which the collisional operator is constructed in model (42) somehow differs from Prigogine's and Paveri Fontana's models in that interactions are now defined by detailing the short-range reactions of each driver to the dynamics of the neighboring vehicles rather than by interpreting her/his overall behavior.

Discrete Velocity Models

Recently new mathematical models of vehicular traffic, based on the discrete kinetic theory, have been proposed in the literature, with the aim of taking into account, still at a mesoscopic level of description, some aspects of the strong granular nature of the flow of vehicles. Indeed, as reported by several authors (see e. g., Ben-Naim and Krapivsky [11,12], Berthelin et al. [13]), cars along a road tend to cluster, which gives granularity in space, with a nearly constant speed within each cluster, which makes also the velocity a discretely distributed variable.

Referring to Delitala and Tosin [33], the main idea is to make the velocity variable v discrete by introducing in the domain D_v a grid $I_v = \{v_i\}_{i=1}^n$ of the form

$$0 = v_1 < v_2 < \cdots < v_n = 1$$

and letting then $v \in I_v$, while time and space are left continuous. Each v_i is interpreted as a *velocity class*, encompassing a certain range of velocities v which are not individually distinguished. Correspondingly, the distribution function f is expressed as a linear combination of Dirac functions in the variable v, with coefficients depending on time and space:

$$f(t, x, v) = \sum_{i=1}^{n} f_i(t, x)\delta(v - v_i), \tag{41}$$

where $f_i(t, x)$ gives the distribution of cars in the point x having at time t a velocity comprised in the ith velocity class. Using this representation, the following expressions for the macroscopic variables of interest are easily derived from Eqs. (23)–(24):

$$\rho(t, x) = \sum_{i=1}^{n} f_i(t, x),$$
$$q(t, x) = \sum_{i=1}^{n} v_i f_i(t, x),$$
$$u(t, x) = \frac{\sum_{i=1}^{n} v_i f_i(t, x)}{\sum_{i=1}^{n} f_i(t, x)}.$$

Moreover, plugging Eq. (41) into Eq. (26) yields the following system of evolution equations for the distribution functions f_i:

$$\frac{\partial f_i}{\partial t} + v_i\frac{\partial f_i}{\partial x} = J_i[\mathbf{f}], \quad i = 1, \ldots, n, \tag{42}$$

where $\mathbf{f} = (f_1, \ldots, f_n)$ and $J_i[\mathbf{f}]$ is the ith collisional operator. Modeling of $J_i[\mathbf{f}]$ is performed by appealing to

a discrete stochastic game theory (see Bertotti and Delitala [14]), after identifying gain and loss contributions:

$$J_i[\mathbf{f}] = G_i[\mathbf{f},\ \mathbf{f}] - f_i L_i[\mathbf{f}]$$

with

$$G_i[\mathbf{f},\ \mathbf{f}](t,x) = \sum_{h,k=1}^{n} \int_{x}^{x+\xi} \eta[\rho](t,x_*) A^i_{hk}[\rho](t,x_*) \cdot f_h(t,x) f_k(t,x_*) w(x_*-x)\mathrm{d}x_* \quad (43)$$

and

$$L_i[\mathbf{f}](t,x) = \sum_{k=1}^{n} \int_{x}^{x+\xi} \eta[\rho](t,x_*) f_k(t,x_*) w(x_*-x)\mathrm{d}x_* \,. \quad (44)$$

One of the main features of Delitala–Tosin's model is that interactions are not local, but are distributed over a characteristic length $\xi > 0$, which can be interpreted as the *visibility length* of the drivers: A candidate vehicle located in x is supposed to be affected, in average, by all field vehicles comprised in the *visibility zone* $[x, x+\xi]$ ahead of it. One can therefore speak of *averaged binary interactions*. Averaging is carried out by the weight function w, which is nonnegative and supported in the above visibility zone, with unit integral there.

Acceleration and deceleration of the vehicles are described, like in Klar and coworkers' model, as short-range reactions of the drivers to the traffic dynamics in terms of velocity class transitions. Specifically, a *table of games* A^i_{hk} is introduced, which gives the probability that a candidate vehicle in the hth velocity class adjusts its speed to the ith class after an interaction with a field vehicle in the kth class. These probabilities are parameterized by the density of cars ρ, so as to account for a dependence of the interactions on the local traffic congestion, and are subject to the additional requirements

$$A^i_{hk}[\rho] \geq 0, \quad \sum_{i=1}^{n} A^i_{hk}[\rho] = 1, \quad \forall i,h,k \in \{1,\dots,n\}, \quad \forall \rho \in [0,\rho_{\max}]\,,$$

where $\rho_{\max}$ is the road capacity.

Finally, interactions are assumed to happen with more or less frequency according to the local degree of occupancy of the road, which is expressed in Eqs. (43), (44) by the *interaction rate* $\eta[\rho] > 0$. In particular, the latter can be thought of as a measure of the reactiveness of the drivers to the evolution of the traffic in their visibility zone.

The framework depicted by Eqs. (42)–(44) can be formally derived from very general kinetic structures (see Arlotti et al. [1], Bellomo [7]) as shown by Tosin [65]. Specific models are then obtained by detailing the table of games and the interaction rate as functions of the macroscopic density of cars ρ. In [33] the following form of $\eta[\rho]$ is proposed:

$$\eta[\rho] = \frac{1}{1-\rho/\rho_{\max}}$$

which models an increasing reactiveness of the drivers as the road becomes more and more congested, while the table of games is constructed by distinguishing three main cases:

(i) If the candidate vehicle interacts with a faster field vehicle then it either maintains its current speed or possibly accelerates of one velocity class (follow-the-leader strategy), depending on the available surrounding free space.
(ii) If the candidate vehicle interacts with a slower field vehicle then it either maintains its current speed, provided it can overtake the leader, or slows down to the velocity class of the latter and queues.
(iii) If the candidate vehicle interacts with an equally fast field vehicle then a mix of the two previous situations arises, according to the current evolution of the system.

Technically, the expressions of the coefficients A^i_{hk}, where h, k denote the pre-interaction velocity classes of the candidate and the field vehicle, respectively, are as follows:

(i) for $h < k$

$$A^i_{hk}[\rho] = \begin{cases} 1-\alpha(1-\rho/\rho_{\max}) & \text{if } i = h \\ \alpha(1-\rho/\rho_{\max}) & \text{if } i = h+1 \\ 0 & \text{otherwise;} \end{cases}$$

(ii) for $h > k$

$$A^i_{hk}[\rho] = \begin{cases} 1-\alpha(1-\rho/\rho_{\max}) & \text{if } i = k \\ \alpha(1-\rho/\rho_{\max}) & \text{if } i = h \\ 0 & \text{otherwise;} \end{cases}$$

(iii) for $h = k$

$$A^i_{hk}[\rho] = \begin{cases} \alpha\rho/\rho_{\max} & \text{if } i = h-1 \\ 1-\alpha & \text{if } i = h \\ \alpha(1-\rho/\rho_{\max}) & \text{if } i = h+1 \\ 0 & \text{otherwise,} \end{cases}$$

where $\alpha \in [0,1]$ is a parameter related to the quality of the road ($\alpha = 0$ stands for bad road, $\alpha = 1$ for good road). Notice that in case (ii) the coefficient $A^h_{hk}[\rho]$, linked to the possibility for the candidate vehicle to maintain its speed

because it can overtake the heading field vehicle, defines in practice a probability of passing

$$P = \alpha \left(1 - \frac{\rho}{\rho_{\max}}\right)$$

which closely recalls that of Prigogine's model (cf. Eq. (29)). Finally, in case (iii) a modification of the table of games is needed for $h = 1$ or $h = n$, for in these cases the candidate vehicle cannot brake or accelerate, respectively, due to the lack of further lower or higher velocity classes. The deceleration or the acceleration are then simply merged into the tendency of the drivers to preserve the current speed.

Figure 3 shows the dimensionless equilibrium diagrams for the flux q (fundamental diagram) and the average speed u vs. the vehicle density ρ, obtained as stationary solutions to Delitala–Tosin's discrete velocity model for two different values of the parameter α appearing in the table of games. Notice the almost linear growth of q, with a corresponding nearly constant u close to the maximum dimensionless allowed value $u_{\max} = 1$, for low density, and their subsequent strongly nonlinear decrease to zero for high density. This behavior, in good agreement with the experimental results, suggests that the model is able to capture in average the phase transition between free and congested flow occurring in traffic (see Subsect. "The Phase Transition Model" and also Kerner [47,48,49]). Further, we observe that when the quality of the road decreases the phase transition is correspondingly anticipated at lower values of ρ.

In the discrete velocity model by Coscia et al. [26] the velocity grid I_v is conceived so as to have a variable step, which tends to zero for high vehicle concentrations. Specifically, the discretization consists of $2n - 1$ points given by

$$v_i(\rho) = \frac{i-1}{n-1} V_e(\rho), \quad i = 1, \ldots, 2n-1 ,$$

where $V_e(\rho)$ is an average equilibrium velocity, so that $v_1 = 0$, $v_{2n-1} = v_{\max} = 2V_e(\rho)$, and the grid is symmetric with respect to its central point $v_n = V_e(\rho)$. Since $V_e(\rho) \to 0$ for $\rho \to \rho_{\max}$, we note that $v_i(\rho) \to 0$ each $i = 1, \ldots, 2n-1$ as the density locally approaches its maximum admissible value. The authors recommend for $V_e(\rho)$ the use of the Bonzani–Mussone's velocity diagram (10). Alternatively, one can use the classical diagram (9) of the LWR model.

The mathematical structure of the discrete kinetic equations to generate particular models relies on the localized Boltzmann-like collisional framework:

$$\frac{\partial f_i}{\partial t} + \frac{\partial}{\partial x}(v_i(\rho) f_i) = G_i[\mathbf{f}, \mathbf{f}] - f_i L_i[\mathbf{f}], \quad i = 1, \ldots, 2n-1$$

for gain and loss operators given respectively by

$$G_i[\mathbf{f}, \mathbf{f}] = \sum_{h,k=1}^{2n-1} \gamma |v_h - v_k| A^i_{hk} f_h f_k ,$$

$$L_i[\mathbf{f}] = \sum_{k=1}^{2n-1} \gamma |v_i - v_k| f_k ,$$

where $\gamma > 0$ is a constant.

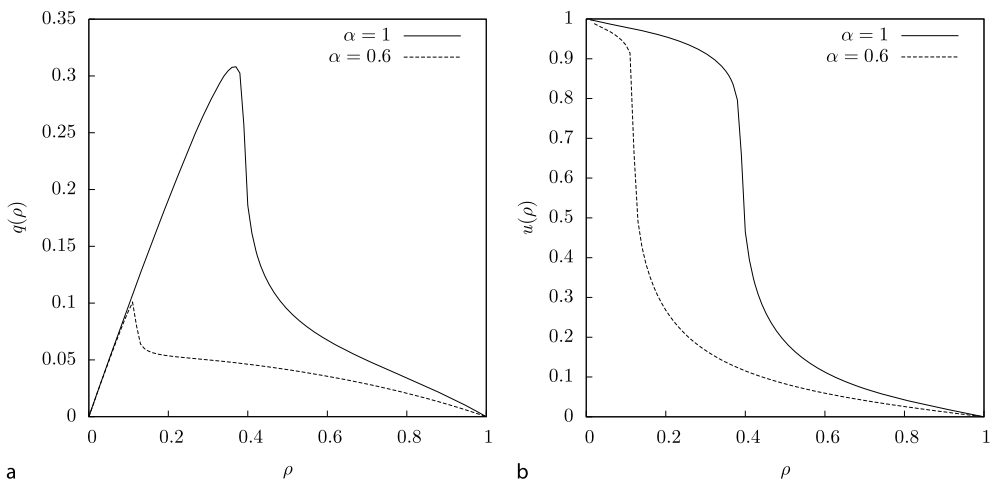

Vehicular Traffic: A Review of Continuum Mathematical Models, Figure 3

Fundamental diagram (a) and velocity diagram (b) computed as stationary dimensionless solutions to Delitala–Tosin's model Eqs. (42)–(44) for two different values of the parameter α of the table of games, corresponding to mediocre ($\alpha = 0.6$) and good ($\alpha = 1$) road conditions

The table of games A^i_{hk} is assumed to be constant in time and space, and is constructed so as to allow interactions only among vehicles belonging to sufficiently close velocity classes ($|h-k| \le 1$ in [26]). After an interaction, the candidate vehicle can only fall in a velocity class adjacent to its current one, thus the sole potentially non-zero coefficients A^i_{hk} are those for which $i = h-1$, $i = h$, and $i = h+1$. In addition, slow and fast cars, which are distinguished by comparing their velocity class to the central class v_n, are supposed to interact only with faster and slower cars, respectively, with a possible tendency to accelerate in the former case and to decelerate in the latter case. If $\epsilon_a, \epsilon_d \in [0,1]$ denote constant acceleration and deceleration probabilities, the table of games proposed by Coscia and coworkers reads as follows:

(i) for $h < n$

$$A^h_{h,h+1} = 1-\epsilon_a, \quad A^{h+1}_{h,h+1} = \epsilon_a, \quad A^i_{hk} = 0 \quad \text{otherwise;}$$

(ii) for $h > n$

$$A^{h-1}_{h,h-1} = \epsilon_d, \quad A^h_{h,h-1} = 1-\epsilon_d, \quad A^i_{hk} = 0 \quad \text{otherwise.}$$

A further particularization of these expressions is obtained by setting $\epsilon_a = \epsilon, \epsilon_d = \nu\epsilon$ for $0 \le \nu \le 1$, in order to identify a basic reaction probability $\epsilon \in [0,1]$ and a measure ν of the relative strength of ϵ_a and ϵ_d.

Road Networks

A *road network* is a finite set of $N \in \mathbb{N}$ interconnected roads, which communicate with each other through exchanges of vehicle fluxes at junctions. Therefore, a network can be modeled as an oriented graph in which roads are assimilated to the arcs and junctions to the vertices. Each arcs is represented by an interval $I_i = [a_i, b_i] \subset \mathbb{R}$, $i = 1, \dots, N$, $a_i < b_i$ with possibly either $a_i = -\infty$ or $b_i = +\infty$, whereas each vertex J is univocally identified by the sets of its incoming and outgoing arcs as $J = ((i_1, \dots, i_n), (j_1, \dots, j_m))$, where the first n-uple and the second m-uple contain the indexes of the incoming and the outgoing arcs, respectively ($m, n \le N$). Hence, the complete model is given by a couple $(\mathcal{I}, \mathcal{J})$, where $\mathcal{I} = \{I_i\}_{i=1}^N$ is the collection of arcs and $\mathcal{J}$ the collection of vertices.

The evolution of traffic on road networks is usually described at a macroscopic level via car densities and other conserved quantities, see Chitour and Piccoli [21], Coclite et al. [22], Garavello and Piccoli [34,35], Herty et al. [43,44], Holden and Risebro [45], Lebacque and Khoshyaran [53]. Similar ideas were also used for telecommunication networks, see D'Apice et al. [31], gas pipelines networks, see Banda et al. [5], Colombo and Garavello [25], and supply chains, see Armbruster et al. [2], D'Apice and Manzo [29], Göttlich et al. [37]. In all these cases one is faced with a system of conservation laws:

$$\frac{\partial \mathbf{u}_i}{\partial t} + \frac{\partial}{\partial x} f(\mathbf{u}_i) = 0, \quad i = 1, \dots, N, \tag{45}$$

where $\mathbf{u}_i = \mathbf{u}_i(t,x) : \mathbb{R}_+ \times I_i \to \mathbb{R}^d$ is the vector of the conserved quantities on the ith arc of the graph and $f : \mathbb{R}^d \to \mathbb{R}^d$ is the flux function.

In the case of vehicular traffic, Eq. (45) may represent one of the first or second order models (if $d = 1, 2$, respectively) reviewed in Sect. "Macroscopic Modeling". For instance, choosing the Lighthill–Whitham–Richards model one can identify $\mathbf{u}_i$ with the density of cars $\rho_i = \rho_i(t,x) : \mathbb{R}_+ \times I_i \to \mathbb{R}$ along the ith road, so that on each arc of the graph the following scalar conservation law is set:

$$\frac{\partial \rho_i}{\partial t} + \frac{\partial}{\partial x} f(\rho_i) = 0, \quad i = 1, \dots, N, \tag{46}$$

where the flux $f : \mathbb{R} \to \mathbb{R}$ is given by Eq. (8). Conversely, the Aw–Rascle model yields the 2×2 system

$$\frac{\partial}{\partial t}\begin{pmatrix} \rho_i \\ q_i \end{pmatrix} + \frac{\partial}{\partial x}\begin{pmatrix} q_i - \rho_i p(\rho_i) \\ \dfrac{q_i^2}{\rho_i} - q_i p(\rho_i) \end{pmatrix} = 0, \quad i = 1, \dots, N, \tag{47}$$

where now $\mathbf{u}_i = (\rho_i, q_i)$ with $q_i = \rho_i w_i$ the ith generalized flux (cf. Eq. (19)).

To complete the model, system (45) has to be supplemented by the dynamics at the junctions of the network. A natural assumption is the conservation of the quantities $\mathbf{u}_i$, $i = 1, \dots, N$ at vertices, which is for instance obtained by requiring that $\mathbf{u}_i$ be a solution at each given vertex: This means that it satisfies Eq. (45), possibly in weak form, for test functions defined on the whole graph and smooth through that vertex. However, in some cases the conservation is not strictly necessary from the modeling point of view. This may either be a modeling choice, like in Garavello and Piccoli [35], or due to some absorption at vertices, e. g. for the presence of queues (Göttlich et al. [37]).

The conservation of the quantities $\mathbf{u}_i$ at vertices is in general not sufficient to describe univocally the dynamics on the network. This was first shown by Holden and Risebro [45], who introduced an additional maximization of some functional at vertices to get uniqueness of the solution to the problem. Then various other choices were proposed, along with techniques to solve Cauchy problems on complete networks. Usually one assigns some traffic distribution rules from incoming to outgoing arcs, joined to

traffic flux (or possibly other functionals) maximization. In some cases, extra rules must be used: For instance, for a junction of a road network with two incoming and one outgoing roads, one needs to describe the right of way of the two incoming roads.

Definition of Solution and Dynamics at Vertices. Riemann Solvers

In this subsection we briefly report about mathematical details needed to state the concept of solution for a system of conservation laws on a network, including the proper formal definition of the dynamics at the junctions.

We will assume in the sequel that the network load be described on each arc I_i by a vector-valued variable $\mathbf{u}_i = \mathbf{u}_i(t,x) : \mathbb{R}_+ \times I_i \to \mathbb{R}^d$ satisfying the conservation law (45) for a given flux $f : \mathbb{R}^d \to \mathbb{R}^d$. The function $\mathbf{u}_i$ is required to be a weak entropic solution to Eq. (45) on $I_i = [a_i, b_i]$, thus for every smooth positive test function $\varphi = \varphi(t,x) : \mathbb{R}_+ \times I_i \to \mathbb{R}_+$ compactly supported in $(0,+\infty) \times (a_i, b_i)$ the following relation must hold:

$$\int_0^{+\infty} \int_{a_i}^{b_i} \left(\mathbf{u}_i \frac{\partial \varphi}{\partial t} + f(\mathbf{u}_i) \frac{\partial \varphi}{\partial x} \right) \mathrm{d}x\mathrm{d}t = 0$$

with, in addition, entropy conditions fulfilled, see Bressan [18], Dafermos [27], Serre [63,64]. It is well known that for every initial datum with bounded total variation (possibly sufficiently small, especially in the case of systems of conservation laws) Eq. (45) on $\mathbb{R}$ admits a unique entropic weak solution depending continuously on the initial datum in L^1_{loc}. Moreover, if the initial datum is taken in $L^1 \cap L^\infty$ then Lipschitz continuous dependence of the mapping $t \mapsto \mathbf{u}_i(t,\cdot)$ in L^1 is achieved.

To define solutions at vertices, fix a vertex J with n incoming arcs, say $I_1, \ldots, I_n$, and m outgoing arcs, say $I_{n+1}, \ldots, I_{n+m}$. A weak solution at J is a collection of functions $\mathbf{u}_l : \mathbb{R}_+ \times I_l \to \mathbb{R}^d$, $l = 1, \ldots, n+m$, such that

$$\sum_{l=1}^{n+m} \left[\int_0^{+\infty} \int_{a_l}^{b_l} \left(\mathbf{u}_l \frac{\partial \varphi_l}{\partial t} + f(\mathbf{u}_l) \frac{\partial \varphi_l}{\partial x} \right) \mathrm{d}x\mathrm{d}t \right] = 0 \quad (48)$$

for every smooth test function $\varphi_l : \mathbb{R}_+ \times I_j \to \mathbb{R}$ compactly supported either in $(0,+\infty) \times (a_l, b_l]$ if $l = 1, \ldots, n$ (incoming arcs) or in $(0,+\infty) \times [a_l, b_l)$ if $l = n+1, \ldots, n+m$ (outgoing arcs). Test functions are also required to stick smoothly across the vertex, i. e.,

$$\varphi_i(\cdot, b_i) = \varphi_j(\cdot, a_j), \quad \frac{\partial \varphi_i}{\partial x}(\cdot, b_i) = \frac{\partial \varphi_j}{\partial x}(\cdot, a_j)$$

for all $i = 1, \ldots, n$ and all $j = n+1, \ldots, n+m$. As a consequence, the solution satisfies the Rankine–Hugoniot condition at the vertex J, namely

$$\sum_{i=1}^{n} f(\mathbf{u}_i(t, b_i^-)) = \sum_{j=n+1}^{n+m} f(\mathbf{u}_j(t, a_j^+))$$

for almost every $t > 0$. This Kirchhoff-like condition also guarantees the conservation of the solution at vertices.

For a system of conservation laws on the real line, a *Riemann problem* consists in a Cauchy problem for an initial datum of Heaviside type, that is a piecewise constant function with only one discontinuity. Then one looks for self-similar solutions constituted by simple waves, which are the building blocks to deal with more general Cauchy problems via wave-front tracking algorithms (see Bressan [18], Garavello and Piccoli [34]). These solutions are formed by (a combination of) continuous waves, called *rarefactions*, and traveling discontinuities, called *shocks*, whose speeds are related to the eigenvalues of the Jacobian matrix of the flux f (namely, to the derivative of f in the scalar case, see Bressan [18]). A mapping yielding the solution to a Riemann problem as a function of the initial datum is said to be a *Riemann solver*.

Analogously, a Riemann problem for a vertex in a network is the Cauchy problem corresponding to an initial datum which is constant on each arc entering or issuing from that vertex. Correspondingly, the following definition of Riemann solver at a vertex is given:

Definition 1 For an initial datum $\underline{u}^0 = (\mathbf{u}_1^0, \ldots, \mathbf{u}_{n+m}^0) \in \mathbb{R}^{d\times(n+m)}$ prescribed at a vertex J, a *Riemann solver* at J is a mapping $RS : \mathbb{R}^{d\times(n+m)} \to \mathbb{R}^{d\times(n+m)}$ that associates to $\underline{u}^0$ a vector $\underline{\hat{u}} = (\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_{n+m}) \in \mathbb{R}^{d\times(n+m)}$ such that:

(i) On every incoming arc I_i, $i = 1, \ldots, n$, the solution is given by the waves produced by the Riemann problem $(\mathbf{u}_i^0, \hat{\mathbf{u}}_i)$;
(ii) On every outgoing arc I_j, $j = n+1, \ldots, n+m$, the solution is given by the waves produced by the Riemann problem $(\hat{\mathbf{u}}_j, \mathbf{u}_j^0)$.

Any Riemann solver at J must fulfil the following consistency condition:

$$RS(RS(\underline{u}^0)) = RS(\underline{u}^0)\,.$$

Additional hypotheses are usually stated for a Riemann solver at a vertex, namely:

(H1) The waves generated from the vertex must have negative speeds on the incoming arcs and positive speeds on the outgoing ones.

(H2) The solution to a Riemann problem at a vertex must satisfy Eq. (48).
(H3) The mapping $\mathbf{u}_l^0 \mapsto f(\hat{\mathbf{u}}_l)$ has to be continuous each $l = 1, \ldots, n+m$.

Hypothesis (H1) is a consistency condition on the description of the dynamics at the vertex. In fact, if (H1) does not hold then some waves generated by the Riemann solver may disappear through the vertex. Condition (H2) is necessary to have a weak solution at the vertex. However, in some cases (H2) may be violated if only some components of the solution are conserved across the junctions, see for instance Garavello and Piccoli [35]. Finally, (H3) is a regularity condition needed to have a well-posed theory. We simply remark here that continuity of the mapping $\underline{u}^0 \mapsto \underline{\hat{u}}$ might not hold in case (H1) holds true.

An important consequence of hypothesis (H1) is that some restrictions on the admissible values of the solutions of the problem, and of the corresponding fluxes, arise. For instance, consider a single conservation law for a bounded quantity u, say $u \in [0, u_{\max}]$ for a certain fixed threshold $u_{\max} > 0$, and assume the following:

(F) The flux $f : [0, u_{\max}] \to \mathbb{R}$ is strictly concave, with in addition $f(0) = f(u_{\max}) = 0$. Thus f has a unique maximum point $\sigma \in (0, u_{\max})$.

If $\tau : [0, u_{\max}] \to [0, u_{\max}]$ is defined to be the mapping such that $f(\tau(\varrho)) = f(\varrho)$ for all $\varrho \in [0, u_{\max}]$, with however $\tau(\varrho) \neq \varrho$ for all $\varrho \in [0, u_{\max}] \setminus \{\sigma\}$, then the following results are obtained:

Proposition 2 *Consider a single conservation law for a bounded variable $u \in [0, u_{\max}]$ between an incoming arc I_i and an outgoing arc I_j at a vertex J, and assume that the flux function satisfies assumption* (F) *above. Let $RS : \mathbb{R}^2 \to \mathbb{R}^2$ be a Riemann solver for the vertex J, where the initial datum $\underline{u}^0 = (u_i^0, u_j^0)$ is prescribed, and set $\underline{\hat{u}} := RS(\underline{u}^0) = (\hat{u}_i, \hat{u}_j)$. Then*

$$\hat{u}_i \in \begin{cases} \{u_i^0\} \cup \left(\tau\left(u_i^0\right), u_{\max}\right] & \text{if } 0 \le u_i^0 \le \sigma \\ [\sigma, u_{\max}] & \text{if } \sigma \le u_i^0 \le u_{\max}, \end{cases}$$

$$\hat{u}_j \in \begin{cases} [0, \sigma] & \text{if } 0 \le u_j^0 \le \sigma \\ \left\{u_j^0\right\} \cup \left(0, \tau\left(u_j^0\right)\right] & \text{if } \sigma \le u_j^0 \le u_{\max}. \end{cases}$$

Proposition 3 *Consider a single conservation law like in the previous Proposition 2 and assume again the hypothesis (F) on the flux f. For an initial datum $\underline{u}^0 = (u_i^0, u_j^0)$, a Riemann solver at a vertex J complying with the hypothesis (H1) above is univocally defined by prescribing the values $f(\underline{\hat{u}})$ of the flux. Moreover, there exist maximal possible fluxes in the incoming and outgoing arcs I_i, I_j given respectively by*

$$f_i^{\max}(\underline{u}^0) := \begin{cases} f(u_i^0) & \text{if } 0 \le u_i^0 \le \sigma \\ f(\sigma) & \text{if } \sigma \le u_i^0 \le u_{\max}, \end{cases}$$

$$f_j^{\max}(\underline{u}^0) := \begin{cases} f(\sigma) & \text{if } 0 \le u_j^0 \le \sigma \\ f(u_j^0) & \text{if } \sigma \le u_j^0 \le u_{\max}. \end{cases}$$

After assigning a Riemann solver RS at a vertex J, the admissible solutions $\mathbf{u}_l$ across J are defined as those functions of bounded total variation:

$$TV(\mathbf{u}_l(t, \cdot); I_l) < +\infty, \; l = 1, \ldots, n+m$$

for almost every time t of existence, with the further property:

$$RS(\underline{u}_J(t)) = \underline{u}_J(t)$$

where $\underline{u}_J(t) := (\mathbf{u}_1(t, b_1^-), \ldots, \mathbf{u}_n(t, b_n^-), \mathbf{u}_{n+1}(t, a_{n+1}^+), \ldots, \mathbf{u}_{n+m}(t, a_{n+m}^+)) \in \mathbb{R}^{d \times (n+m)}$.

Finally, for an arc $I_i = [a_i, b_i]$ such that either $a_i > -\infty$ and I_i is not an outgoing arc of any vertex of the graph or $b_i < +\infty$ and I_i is not an incoming arc of any vertex of the graph, a boundary condition $\boldsymbol{\psi}_i = \boldsymbol{\psi}_i(t) : \mathbb{R}_+ \to \mathbb{R}^d$ on the unknown $\mathbf{u}_i$ is prescribed. In particular, the latter is required to satisfy

$$\mathbf{u}_i(t, a_i) = \boldsymbol{\psi}_i(t) \quad \text{or} \quad \mathbf{u}_i(t, b_i) = \boldsymbol{\psi}_i(t), \forall\, t > 0$$

in the sense specified by Bardos et al. [6].

Constructing Solutions on a Network

To solve the Cauchy problem on a network for a prescribed initial datum $\underline{u}^0 = (\mathbf{u}_1^0, \ldots, \mathbf{u}_N^0)$, where each $\mathbf{u}_i^0 = \mathbf{u}_i^0(x) : I_i \to \mathbb{R}^d$ is a measurable function, and possibly a set of suitable boundary conditions amounts to finding N vector-valued functions $\mathbf{u}_i = \mathbf{u}_i(t, x) : \mathbb{R}_+ \times I_i \to \mathbb{R}^d$ such that:

(i) The mapping $t \mapsto \|\mathbf{u}_i(t, \cdot)\|_{L^1_{\text{loc}}}$ is continuous each $i = 1, \ldots, N$;
(ii) Each $\mathbf{u}_i$ is a weak entropic solution to Eq. (45) on the corresponding arc I_i;
(iii) At each vertex of the graph the proper collection of incoming and outgoing $\mathbf{u}_i$ defines an admissible solution in the sense discussed in the previous section;
(iv) $\mathbf{u}_i(0, x) = \mathbf{u}_i^0(x)$ for almost every $x \in I_i$.

There is a general strategy to prove existence of admissible solutions on a network, which basically relies on the following steps:

1. Construct approximate solutions via wave-front tracking algorithms, using suitable Riemann solvers at vertices to deal with the interactions among different arcs;
2. Estimate the total variation of the flux at vertices, then on the whole network;
3. Use appropriate compactness properties to pass to the limit and recover the solution to the problem.

The wave-front tracking algorithm invoked by step 1 can be roughly described as follows. First one approximates the initial datum $\mathbf{u}_i^0$ by a piecewise constant function, then uses classical self-similar solutions to Riemann problems on arcs combined with solutions generated by Riemann solvers at vertices. Rarefaction waves are split in *rarefaction shocks fans*, i. e., collections of small non-entropic shocks. When waves interact, either along arcs or at junctions, new Riemann problems are generated to be sequentially solved. In order to successfully perform this construction, one needs to estimate the number of waves and of interactions, which is achieved via suitable functionals (see Garavello and Piccoli [34,35] for details).

Estimation of the total variation of the flux on the network (step 2) can be reduced to the case of a single junction, thanks to the following result:

Theorem 4 *Let $K \subset \mathbb{R}^{d\times N}$ be a fixed compact set and assume $\underline{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_N) \in K$. For each vertex J of the network, consider the new network which has J as unique junction, whose incoming and outgoing arcs are prolonged to infinity. If for this network there exists a constant $C > 0$ such that for every initial datum $\underline{u}^0 = (\mathbf{u}_1^0, \ldots, \mathbf{u}_N^0)$ with bounded total variation it results*

$$TV[f(\underline{u}(t))] \le C \cdot TV[f(\underline{u}^0)] \ ,$$

then the same estimate holds for the flux on the entire network with a possibly time-dependent constant $C_t > 0$.

Finally, passing to the limit to obtain the true entropic solution to the original problem as required by step 3 is usually the most delicate task. Indeed, even if some compactness of the sequence $\{f(\underline{u}_\nu)\}_{\nu\in\mathbb{N}}$ holds (where $\{\underline{u}_\nu\}_{\nu\in\mathbb{N}}$ is the sequence generated by the wave-front tracking algorithm), still u_ν may fail to converge. This issue was solved for the scalar case $u_i \in \mathbb{R}$ with concave flux f by proving that the number of waves crossing the maximum of the flux can be controlled. To be precise, assume that condition (F) holds true on the function $f : [0, u_{\max}] \to \mathbb{R}$, then define

Definition 5 A wave (u_i^-, u_i^+) in the ith arc I_i is said to be a *big shock* if $u_i^- < u_i^+$ and

$$\operatorname{sgn}(u_i^- - \sigma) \cdot \operatorname{sgn}(u_i^+ - \sigma) < 0.$$

Definition 6 Fix a vertex J. An incoming arc I_i is said to have a *good datum* at J at time $t > 0$ if

$$u_i(t, b_i^-) \in [\sigma, u_{\max}]$$

and a *bad datum* otherwise. Conversely, an outgoing arc I_j is said to have a good datum at J at time $t > 0$ if

$$u_j(t, a_j^+) \in [0, \sigma]$$

and a bad datum otherwise.

Then one can prove (see Garavello and Piccoli [34]):

Lemma 7 *If an arc I_i of a vertex J has a good datum, then the datum remains good after all interactions with waves coming in J from other arcs. Furthermore, no big shocks are produced. Conversely, if I_i has a bad datum then after any interaction with waves coming in J from other arcs either the datum of I_i remains unchanged or a big shock is produced with a resulting good new datum.*

Once the number of big shocks is bounded via Lemma 7, one can invert the flux $u \mapsto f(u)$, which allows one to pass to the limit also in the sequence $\{\underline{u}_\nu\}_{\nu\in\mathbb{N}}$. The interested reader is referred to the book by Garavello and Piccoli [34] for further details also on this part of the theory.

Riemann Solvers at Vertexes

In this section we focus specifically on vehicular traffic networks, and describe several Riemann solvers at vertices recently proposed in the literature for the Lighthill–Whitham–Richards model (46). In particular, we consider on each arc I_i the density of cars as a bounded variable $\rho_i \in [0, \rho_{\max}^i]$, where $\rho_{\max}^i$ stands for the capacity of the ith road, and we look for Riemann solvers at vertices complying with the hypothesis (H1). Notice that if the flux f is given on each arc by the parabolic profile (8) (with $\rho_{\max}$ replaced by $\rho_{\max}^i$) then condition (F) is straightforwardly satisfied with $\sigma = \rho_{\max}^i/2$.

According to Proposition 3, a Riemann solver at a vertex J is defined by simply determining the proper values of the flux f. In the papers by Coclite et al. [22] and by D'Apice et al. [31] two different Riemann solvers at vertices, hereafter denoted by R1 and R2, are considered, based on the following algorithms:

- Solver R1:
  (a) The traffic from incoming roads is distributed on outgoing roads according to fixed rates;
  (b) Under rule (a), the flow of cars through the junction is maximized.

- Solver R2:
  The flow of cars through the junction is maximized over both incoming and outgoing roads.

Concerning the solver R1, for a certain junction J let $I_1, \dots, I_n$ be the n incoming roads and, analogously, $I_{n+1}, \dots, I_{n+m}$ the m outgoing ones. Rule (a) corresponds to defining a stochastic matrix

$$A = (\alpha_{ji})_{\substack{i=1,\dots,n \\ j=n+1,\dots,n+m}} \tag{49}$$

whose entries α_{ji} represent the rate of traffic coming from I_i and directed to I_j. For this, the coefficients α_{ji} are required to satisfy

$$0 < \alpha_{ji} < 1, \qquad \sum_{j=n+1}^{n+m} \alpha_{ji} = 1 \,.$$

Notice that the second condition entails in particular that no cars are lost at the junction, indeed if

$$\rho_j = \sum_{i=1}^{n} \alpha_{ji} \rho_i$$

is the total density of cars flowing through the jth outgoing road after passing the junction then it results

$$\sum_{j=n+1}^{n+m} \rho_j = \sum_{i=1}^{n} \rho_i \,.$$

Furthermore, it can be shown that rule (a) implies the fulfillment of hypothesis (H2) by the Riemann solver R1.

In view of Proposition 3, the incoming and outgoing fluxes at J must take values in the sets

$$\begin{aligned} \Omega_{\text{in}} &= \bigotimes_{i=1}^{n} [0, f_i^{\max}] \subset \mathbb{R}^n, \\ \Omega_{\text{out}} &= \bigotimes_{j=n+1}^{n+m} [0, f_j^{\max}] \subset \mathbb{R}^m \,. \end{aligned} \tag{50}$$

Moreover, owing to rule (a) the incoming fluxes must specifically belong to the region

$$\tilde{\Omega}_{\text{in}} = \{\gamma \in \Omega_{\text{in}} : A\gamma \in \Omega_{\text{out}}\} \,,$$

A being the matrix defined in Eq. (49), which is a convex subset of Ω_{in} determined by linear constraints. By consequence, rule (b) is equivalent to perform a maximization over the incoming fluxes only, the outgoing ones being univocally recoverable using rule (a). Finally, rules (a) and (b) give rise to a linear programming problem at vertex J, whose solution always exists and is in addition unique provided the gradient of the cost function (here, the vector with all unit components) is not orthogonal to the linear constraints describing the set $\tilde{\Omega}_{\text{in}}$.

The orthogonality condition cannot hold if $n > m$. To be specific, we consider in the sequel the simple case study $(n, m) = (2, 1)$ corresponding to a junction with two incoming roads I_1, I_2 and one outgoing road I_3. If not all traffic coming from I_1 and I_2 can flow to I_3, then one should assign an additional yielding or priority rule:

(c) There exists a priority vector $p \in \mathbb{R}^n$ such that the vector of incoming fluxes must be parallel to p.

In the case under consideration, the matrix A reduces to the vector $(1, 1)$, which gives only the trivial information that all vehicles flow toward the road I_3. Rule (b) amounts instead to determining the through flux as

$$\Gamma = \min\{f_1^{\max} + f_2^{\max}, f_3^{\max}\} \,.$$

If $\Gamma = f_1^{\max} + f_2^{\max}$ then one simply takes the maximal flux over both incoming arcs. Conversely, if the opposite happens, consider the set of all possible incoming fluxes (γ_1, γ_2), which must belong to the region $\Omega_{\text{in}} = [0, f_1^{\max}] \times [0, f_2^{\max}]$, and define the following lines:

$$r_p = \{tp : t \in \mathbb{R}\}, \quad r_\Gamma = \{(\gamma_1, \gamma_2) : \gamma_1 + \gamma_2 = \Gamma\} \,,$$

where p is the vector invoked by rule (c). These lines intersect at a point P. As shown by Fig. 4, two situations may arise, namely either P belongs to Ω_{in} or P lies outside Ω_{in}. In the first case the incoming fluxes are individuated by P, whereas in the second case they are determined by the projection Q of P onto the convex set $\Omega_{\text{in}} \cap r_\Gamma$.

The reasoning can be repeated also in the case of n incoming arcs. In $\mathbb{R}^n$, the general definition of the set Ω_{in} given by Eq. (50) applies; furthermore, the line r_p is defined as above (for $p \in \mathbb{R}^n$), while r_Γ is replaced by the hyperplane

$$H_\Gamma = \left\{ (\gamma_1, \dots, \gamma_n) : \sum_{i=1}^{n} \gamma_i = \Gamma \right\} \,.$$

The point $P = r_p \cap H_\Gamma$ exists and is unique. If $P \in \Omega_{\text{in}}$ then it is used again to determine the incoming fluxes, otherwise one chooses its projection Q over $\Omega_{\text{in}} \cap H_\Gamma$. Notice that the point Q is unique since $\Omega_{\text{in}} \cap H_\Gamma$ is a closed convex subset of H_Γ.

It is to check that the hypothesis (H3) is verified for this Riemann solver.

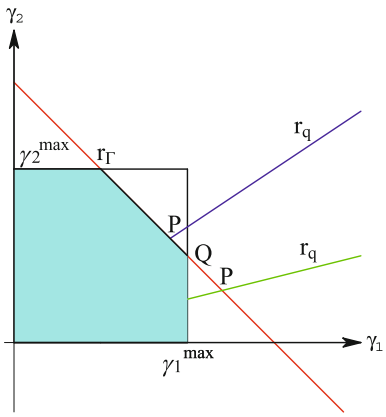

Vehicular Traffic: A Review of Continuum Mathematical Models, Figure 4
The two cases $P \in \Omega_{\text{in}}$ and $P \notin \Omega_{\text{in}}$

Concerning the solver (R2), one defines the maximal incoming and outgoing fluxes using again Proposition 3:

$$\Gamma_{\text{in}} = \sum_{i=1}^{n} f_i^{\max}, \quad \Gamma_{\text{out}} = \sum_{j=n+1}^{n+m} f_j^{\max},$$

so that the through flux is simply determined by

$$\Gamma = \min\{\Gamma_{\text{in}}, \Gamma_{\text{out}}\}.$$

Then one uses rule (c) over both incoming and outgoing arcs to find the values of the flux over the arcs. Once again hypothesis (H1) is assumed, thus densities are uniquely determined. Property (H2) holds by definition, and (H3) is easily proved. Different Riemann solvers at vertices, along with related qualitative analyzes, are studied by D'Apice and Piccoli [30].

Future Directions

Various research perspectives may be envisaged in the field of vehicular traffic, aimed on the one hand at an improvement of the mathematical framework, in order to include in the theory further specific features of this nonstandard system, and on the other hand at a coupling with other topics intimately correlated to traffic, so as to enlarge the scenario of applications to real world problems. Without claiming to be thorough, we simply outline here some possible basic ideas that require to be appropriately developed.

In both macroscopic and kinetic models presented in this review vehicles are described at an essentially mechanical level. Indeed, the macroscopic approach relies on the analogy with fluid particles, hence on principles proper of continuum mechanics, whereas the kinetic approach accounts for a microscopic state of the cars fully characterized by mechanical variables only, such as position and velocity. However, it can be argued that cars are actually not completely mechanical subjects, for the presence of the drivers is likely to affect their behavior in a strongly unconventional way. The concept of fictitious density proposed by De Angelis (see Subsect. "Fictitious Density Model" and [32]) is an attempt to introduce this aspect of the problem in the macroscopic equations of traffic, but in a sense it is too deterministic to catch the real essence of the human behavior, which is instead quite unpredictable. On the other hand, at the kinetic level of representation no model has been so far introduced in the literature taking the action of the driver explicitly into account. Promising ideas toward this issue are expressed in the recent book by Bellomo [7] dealing with generalized kinetic methods for systems of *active particles*, namely particles whose microscopic state includes, besides the standard mechanical variables, also an additional internal variable named *activity*. In the case of vehicular traffic, the latter may represent the driving ability, or alternatively the personality, of the drivers, and can be conceived so as to affect the velocity transition probabilities in an essentially stochastic way.

Further research perspectives concern, primarily at a macroscopic level of description, the development of new hydrodynamic models and algorithms for real-time reconstruction of traffic flows in a road network by few experimental information. Using few data, as well as few model parameters, is indeed necessary for compatibility reasons with actual traffic measurements, which are usually pointwise scattered on the network. In addition, the development of fluid dynamic and gas dynamic models to simulate the emission and propagation of pollutants produced by cars in an urban context might provide powerful technological tools to rationalize the design and use of public and private transportation resources, and to reduce unpleasant effects of urban traffic on the environment.

Bibliography

1. Arlotti L, Bellomo N, De Angelis E (2002) Generalized kinetic (Boltzmann) models: mathematical structures and applications. Math Models Methods Appl Sci 12(4):567–591
2. Armbruster D, Degond P, Ringhofer C (2006) A model for the dynamics of large queuing networks and supply chains. SIAM J Appl Math (electronic) 66(3):896–920
3. Aw A, Rascle M (2000) Resurrection of "second order" models of traffic flow. SIAM J Appl Math 60(3):916–938
4. Aw A, Klar A, Materne T, Rascle M (2002) Derivation of continuum traffic flow models from microscopic follow-the-leader models. SIAM J Appl Math (electronic) 63(1):259–278
5. Banda MK, Herty M, Klar A (2006) Gas flow in pipeline networks. Netw Heterog Media (electronic) 1(1):41–56

6. Bardos C, le Roux AY, Nédélec JC (1979) First order quasilinear equations with boundary conditions. Comm Partial Differential Equations 4(9):1017–1034
7. Bellomo N (2007) Modelling complex living systems. A kinetic theory and stochastic game approach. Modeling and Simulation in Science, Engineering and Technology. Birkhäuser, Boston
8. Bellomo N, Coscia V (2005) First order models and closure of the mass conservation equation in the mathematical theory of vehicular traffic flow. C R Mec 333:843–851
9. Bellomo N, Lachowicz M, Polewczak J, Toscani G (1991) Mathematical topics in nonlinear kinetic theory II. The Enskog equation. Series on Advances in Mathematics for Applied Sciences, vol 1. World Scientific Publishing, Teaneck
10. Bellomo N, Delitala M, Coscia V (2002) On the mathematical theory of vehicular traffic flow I. Fluid dynamic and kinetic modelling. Math Models Methods Appl Sci 12(12):1801–1843
11. Ben-Naim E, Krapivsky PL (1998) Steady-state properties of traffic flows. J Phys A 31(40):8073–8080
12. Ben-Naim E, Krapivsky PL (2003) Kinetic theory of traffic flows. Traffic Granul Flow 1:155
13. Berthelin F, Degond P, Delitala M, Rascle M (2008) A model for the formation and evolution of traffic jams. Arch Ration Mech Anal 187(2):185–220
14. Bertotti ML, Delitala M (2004) From discrete kinetic and stochastic game theory to modelling complex systems in applied sciences. Math Models Methods Appl Sci 14(7):1061–1084
15. Bonzani I (2000) Hydrodynamic models of traffic flow: drivers' behaviour and nonlinear diffusion. Math Comput Modelling 31(6–7):1–8
16. Bonzani I, Mussone L (2002) Stochastic modelling of traffic flow. Math Comput Model 36(1–2):109–119
17. Bonzani I, Mussone L (2003) From experiments to hydrodynamic traffic flow models I. Modelling and parameter identification. Math Comput Model 37(12–13):1435–1442
18. Bressan A (2000) Hyperbolic systems of conservation laws. The one-dimensional Cauchy problem. In: Oxford Lecture Series in Mathematics and its Applications, vol 20. Oxford University Press, Oxford
19. Cercignani C, Lampis M (1988) On the kinetic theory of a dense gas of rough spheres. J Stat Phys 53(3–4):655–672
20. Chakroborty P, Agrawal S, Vasishtha K (2004) Microscopic modeling of driver behavior in uninterrupted traffic flow. J Transp Eng 130(4):438–451
21. Chitour Y, Piccoli B (2005) Traffic circles and timing of traffic lights for cars flow. Discret Contin Dyn Syst Ser B 5(3):599–630
22. Coclite GM, Garavello M, Piccoli B (2005) Traffic flow on a road network. SIAM J Math Anal (electronic) 36(6):1862–1886
23. Colombo RM (2002) A 2×2 hyperbolic traffic flow model, traffic flow – modelling and simulation. Math Comput Model 35(5–6):683–688
24. Colombo RM (2002) Hyperbolic phase transitions in traffic flow. SIAM J Appl Math 63(2):708–721
25. Colombo RM, Garavello M (2006) A well posed Riemann problem for the *p*-system at a junction. Netw Heterog Media (electronic) 1(3):495–511
26. Coscia V, Delitala M, Frasca P (2007) On the mathematical theory of vehicular traffic flow, II. Discrete velocity kinetic models. Int J Non-Linear Mech 42(3):411–421
27. Dafermos CM (2005) Hyperbolic conservation laws in continuum physics. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol 325, 2nd edn. Springer, Berlin
28. Daganzo CF (1995) Requiem for second-order fluid approximation of traffic flow. Transp Res 29B(4):277–286
29. D'Apice C, Manzo R (2006) A fluid dynamic model for supply chains. Netw Heterog Media (electronic) 1(3):379–398
30. D'Apice C, Piccoli B (2008) Vertex flow models for network traffic. Math Models Methods Appl Sci (submitted)
31. D'Apice C, Manzo R, Piccoli B (2006) Packet flow on telecommunication networks. SIAM J Math Anal (electronic) 38(3): 717–740
32. De Angelis E (1999) Nonlinear hydrodynamic models of traffic flow modelling and mathematical problems. Math Comput Model 29(7):83–95
33. Delitala M, Tosin A (2007) Mathematical modeling of vehicular traffic: a discrete kinetic theory approach. Math Models Methods Appl Sci 17(6):901–932
34. Garavello M, Piccoli B (2006) Traffic flow on networks. In: AIMS Series on Applied Mathematics, vol 1. American Institute of Mathematical Sciences (AIMS), Springfield
35. Garavello M, Piccoli B (2006) Traffic flow on a road network using the Aw–Rascle model. Comm Partial Differ Equ 31(1–3):243–275
36. Gazis DC, Herman R, Rothery RW (1961) Nonlinear follow-the-leader models of traffic flow. Oper Res 9:545–567
37. Göttlich S, Herty M, Klar A (2006) Modelling and optimization of supply chains on complex networks. Commun Math Sci 4(2):315–330
38. Greenberg JM (2001/02) Extensions and amplifications of a traffic model of Aw and Rascle. SIAM J Appl Math (electronic) 62(3):729–745
39. Günther M, Klar A, Materne T, Wegener R (2002) An explicitly solvable kinetic model for vehicular traffic and associated macroscopic equations. Math Comput Model 35(5–6):591–606
40. Günther M, Klar A, Materne T, Wegener R (2003) Multivalued fundamental diagrams and stop and go waves for continuum traffic flow equations. SIAM J Appl Math 64(2):468–483
41. Helbing D (1998) From microscopic to macroscopic traffic models. In: A perspective look at nonlinear media. Lecture Notes in Phys, vol 503. Springer, Berlin, pp 122–139
42. Helbing D (2001) Traffic and related self-driven many-particle systems. Rev Mod Phys 73(4):1067–1141, doi:10.1103/RevModPhys.73.1067
43. Herty M, Kirchner C, Moutari S (2006) Multi-class traffic models on road networks. Commun Math Sci 4(3):591–608
44. Herty M, Moutari S, Rascle M (2006) Optimization criteria for modelling intersections of vehicular traffic flow. Netw Heterog Media (electronic) 1(2):275–294
45. Holden H, Risebro NH (1995) A mathematical model of traffic flow on a network of unidirectional roads. SIAM J Math Anal 26(4):999–1017
46. Hoogendoorn SP, Bovy PHL (2001) State-of-the-art of vehicular traffic flow modelling. J Syst Cont Eng 215(4):283–303
47. Kerner BS (2000) Phase transitions in traffic flow. In: Helbing D, Hermann H, Schreckenberg M, Wolf DE (eds) Traffic and Granular Flow '99. Springer, New York, pp 253–283
48. Kerner BS (2004) The physics of traffic. Springer, Berlin
49. Kerner BS, Klenov SL (2002) A microscopic model for phase transitions in traffic flow. J Phys A 35(3):L31–L43
50. Klar A, Wegener R (1997) Enskog-like kinetic models for vehicular traffic. J Stat Phys 87(1–2):91–114

51. Klar A, Wegener R (2000) Kinetic derivation of macroscopic anticipation models for vehicular traffic. SIAM J Appl Math 60(5):1749–1766
52. Klar A, Wegener R (2004) Traffic flow: models and numerics. In: Modeling and computational methods for kinetic equations. Model Simul Sci Eng Technol. Birkhäuser, Boston, pp 219–258
53. Lebacque JP, Khoshyaran MM (1999) Modelling vehicular traffic flow on networks using macroscopic models. In: Finite volumes for complex applications II. Hermes Sci Publ, Paris, pp 551–558
54. Leutzbach W (1988) Introduction to the Theory of Traffic Flow. Springer, New York
55. Lighthill MJ, Whitham GB (1955) On kinematic waves, II. A theory of traffic flow on long crowded roads. Proc Roy Soc Lond Ser A 229:317–345
56. Nagel K, Wagner P, Woesler R (2003) Still flowing: approaches to traffic flow and traffic jam modeling. Oper Res 51(5): 681–710
57. Paveri Fontana SL (1975) On Boltzmann-like treatments for traffic flow. Transp Res 9:225–235
58. Payne HJ (1971) Models of freeway traffic and control. Math Models Publ Syst Simul Council Proc 28:51–61
59. Prigogine I (1961) A Boltzmann-like approach to the statistical theory of traffic flow. In: Theory of traffic flow. Elsevier, Amsterdam, pp 158–164
60. Prigogine I, Herman R (1971) Kinetic theory of vehicular traffic. American Elsevier Publishing, New York
61. Rascle M (2002) An improved macroscopic model of traffic flow: Derivation and links with the Lighthill–Whitham model. Math Comput Model, Traffic Flow Model Simul 35(5–6): 581–590
62. Richards PI (1956) Shock waves on the highway. Oper Res 4: 42–51
63. Serre D (1996) Hyperbolicité, entropies, ondes de choc. [Hyperbolicity, entropies, shock waves.] In: Systèmes de lois de conservation, I. Fondations. [Foundations.] Diderot Editeur, Paris
64. Serre D (1996) Structures géométriques, oscillation et problémes mixtes. [Geometric structures, oscillation and mixed problems.] In: Systèmes de lois de conservation, II. Fondations. [Foundations.] Diderot Editeur, Paris
65. Tosin A (2008) Discrete kinetic and stochastic game theory for vehicular traffic: Modeling and mathematical problems. Ph D thesis, Department of Mathematics, Politecnico di Torino
66. Treiber M, Helbing D (2003) Memory effects in microscopic traffic models and wide scattering in flow-density data. Phys Rev E 68(4):046–119, doi:10.1103/PhysRevE.68.046119
67. Treiber M, Hennecke A, Helbing D (2000) Congested traffic states in empirical observations and microscopic simulations. Phys Rev E 62(2):1805–1824, doi:10.1103/PhysRevE.62.1805
68. Treiber M, Kesting A, Helbing D (2006) Delays, inaccuracies and anticipation in microscopic traffic models. Physica A 360(1): 71–88
69. Villani C (2002) A review of mathematical topics in collisional kinetic theory. In: Handbook of mathematical fluid dynamics, vol I. North-Holland, Amsterdam, pp 71–305
70. Wegener R, Klar A (1996) A kinetic model for vehicular traffic derived from a stochastic microscopic model. Transp Theory Stat Phys 25(7):785–798
71. Whitham GB (1974) Linear and nonlinear waves. Wiley-Interscience, New York

Viral Protein Nano-Actuators, Computational Studies of Bio-nanomachines

Atul Dubey, M. Silvina Tomassone
Department of Chemical and Biochemical Engineering, Rutgers, The State University of New Jersey, Piscataway, USA

Article Outline

Glossary
Definition of the Subject
Introduction
The VPL Nanoactuator
Methods
Simulation Results and Discussion
Conclusions
Future Directions
Acknowledgments
Bibliography

Glossary

Nanodevice A Device with size dimensions of the order of the nanometer (10^{-9} m)

Actuator A device used to generate motion by converting some form of energy (such as chemical or light energy) into mechanical energy.

ATP Adenosine Triphosphate molecule used by living organisms to store and transport energy.

End-effector The last link in a robotic assembly that performs the desired tasks by the robot by interacting with or acting upon the elements outside the robot.

Retrovirus A type of virus that has a protein envelope around it, stores its genetic information in an RNA molecule instead of double stranded DNA and transcripts its genome into DNA which integrates into the chromosome of the infected cell thereby replicating the virus when the cell replicates.

Glycoprotein A protein molecule which has a carbohydrate attached to it.

Membrane fusion A process when viral and hose cellular membranes fuse together during infection.

Hemagglutinin A type of protein on the surface of the influenza virus.

Fusogenic The unit that facilitates or takes part in membrane fusion.

Disulfide bond A sulfur–sulfur atomic bond which typically joins two biological segments such as protein chains.

Proteolytic cleavage A process of breaking the peptide bond between amino acids in a protein chain.

Endosome Intracellular vesicle formed by the outer cell membrane used to transport material inside the animal cell.

Endocytosis The process of encapsulation of an external element by the cell membrane and subsequent formation of an endosome for transport within the cell.

Nucleocapsid The genome of a virus enclosed in a protein coat.

α-Helix A secondary structure of protein molecules characterized by a helical shape and a specific pitch and stabilized by hydrogen bonds along the length of the helix.

β-Sheet A secondary structure of protein molecules where two or more chains are interconnected by hydrogen bonds forming a sheet-like structure.

Titin A big filamentous protein used by muscles.

Holonomic constraint A constraint that puts a restriction on the motion of a system by the use of integrable system of differential equations.

Chymotrypsin An enzyme in the pancreas which is involved in the catalysis of hydrolysis of proteins in the small intestine.

PDB The protein data bank which contains structural and other information about a large number of protein molecules. http://www.rcsb.org/pdb/home/home.do

Implicit solvation A technique used in molecular simulation which allows for a potential function to be used in place of explicit solvent molecules surrounding the molecule of interest.

Protonation The process of accepting a positively charged hydrogen ion or a proton by an amino acid.

Helicity The measure of the helical content of a given peptide molecule.

Denaturation Process of loss of secondary structure of a peptide molecule, which can happen for example due to heating beyond a certain level.

Definition of the Subject

In this paper, we describe computational aspects of an initiative to create a nanodevice for chemo-mechanical energy conversion based on a biological system. We provide insights into the available methods and their applications to gain knowledge about whether a given molecule is suitable to function as a nanoactuator. The field of bio-nanotechnology is new and it is developing fast. Noticeably, its roots go back decades of studies in the fields of biophysics, biochemistry and engineering. These disciplines can now come together to create useful and controllable devices at the nano-scale thereby adding another dimension to human scientific capabilities. Their applications in the fields of medicine, space exploration and colonization, and military are very promising.

Introduction

There is a persistent need of miniaturization of machines and energy conversion devices for various engineering applications. A great part of modern research is shifting from macro to nano-scale as consumer and scientific devices become smaller, faster and more energy efficient. In the field of robotics, there is a need to develop nano-scale actuators, joints, motors and other machine components. Since these nanodevices can be either single-molecule or macromolecular devices, the laws governing their behavior are different from the ones that govern macro-scale machines. Hence in order to study and develop them it is crucial to fully understand the material properties and the work environment and be able to create appropriate models that account for them.

Miniature molecular machines have been operating for centuries in living organisms. Nature has created and optimized these complex machines to carry out specific tasks, such as moving cellular cargo, and oxidizing high-energy molecules etc. With modern microscopic abilities, finer structural and functional details of these molecules are being discovered. Mechanical motions, structural features, operating principles, limitations, and design features of many natural molecular machines are being investigated. It is possible to use these machines or their individual parts under artificial stimuli in order to perform tasks not necessarily assigned to them by nature. Attempts at creating artificial molecular machines are also being made. Some of the principles and applications of molecular machines have been reviewed [4,5,7,26,32,38,47].

Molecular machines can be broadly classified into protein-based, DNA-based or chemical machines. From each of the three classes, completely different types of machines are obtained. Each of these machines has a different operating environment, force and displacement capabilities, and fuel requirements. Hence all of them figure in a molecular part-list and can be utilized when requirements are specific to them. For example protein-based ATP motors require protons or high energy molecules, whereas DNA-based machines require DNA strands as fuel. Both these machines can be used to produce rotary motion, although in different environmental conditions.

Molecular devices for use in robotics have not yet been thoroughly investigated. There is a need to find nanoscale equivalents of robot parts such as actuators, links, end-ef-

fectors, sensors etc. This study aims to develop one such protein based linear actuator, called the Viral Protein Linear (VPL) actuator. The inspiration for this type of actuator comes from nature – from the infection mechanism of some viruses. There is a particular segment of the surface protein of retroviruses such as influenza virus that undergoes a conformational change which can be used as a linear actuator producing about 10 nm of linear motion. In this work, we characterize the VPL protein actuator by computationally studying its behavior and predicting its performance.

The VPL Nanoactuator

The role of envelope glycoproteins (surface proteins) of various retroviruses in the process of membrane fusion has been investigated and understood over the years. Such viruses infect their target cells by the process of membrane fusion. Membrane fusion is necessary for a large number of diverse processes in biology such as protein trafficking, protein secretion, fertilization, viral invasion and neurotransmission. The mechanism is best understood among enveloped viruses such as the influenza virus [31,35,43,44]. Specialized viral proteins are required to promote membrane fusion – a process which is otherwise very slow. In many cases, these membrane-fusion proteins also serve as agents that promote the binding of the virus onto the cell surface receptors. In the influenza virus, a protein segment called Hemagglutinin (HA) mediates both the binding of the virus to the cell surface and the subsequent fusion of viral and cellular membranes. The receptor binding subunit of HA is termed HA1, while the fusogenic subunit is denoted as HA2. Figure 1 shows a schematic of the influenza virus.

HA1/HA2 complex consists of the disulfide bonded HA1 and HA2 peptides. Each HA monomer is synthesized as a fusion-incompetent precursor polypeptide known as HA0 which undergoes proteolytic cleavage to give rise to the two chains [25,34]. The two subunits have different functions. For example, in the case of the Human Immunodeficiency Virus (HIV), HIV 1, the precursor glycoprotein is gp160, which is proteolytically cleaved into gp120 and gp41 subunits. The gp120 is the surface subunit and the gp41 is the transmembrane (TM) subunit. The surface subunit serves to recognize the cell to be infected when it comes in the vicinity of the virus with the help of receptors located on the cell surface. The gp41 mediates membrane fusion between the viral and cellular membranes. It has been found that gp41, and corresponding TM subunits in other retroviruses acquire an alpha-helical conformation when the virus is in its active or fusogenic state. The structure is like a hairpin composed of three coils, having one C terminal (carboxy- end) and the other N terminal (amino-end). The carboxy regions pack in an anti-parallel manner around the three hydrophobic amino ends as shown in Fig. 3.

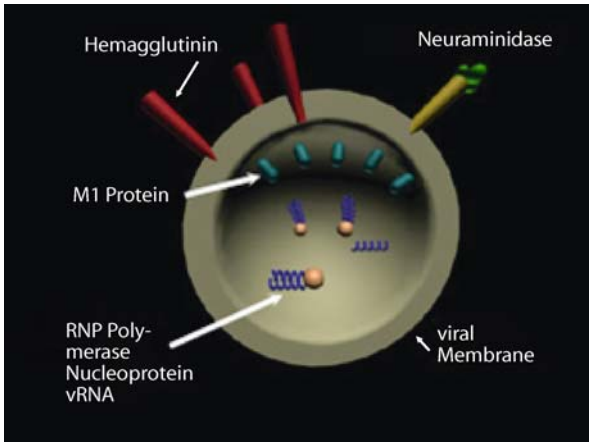

Viral Protein Nano-Actuators, Computational Studies of Bio-nanomachines, Figure 1
A schematic of the influenza virus. Hemagglutinin (HA) polypeptides lie on the surface of the virus while viral RNA resides inside. HA has subunits that attach to cells and then facilitate membrane fusion in order to infect the cell

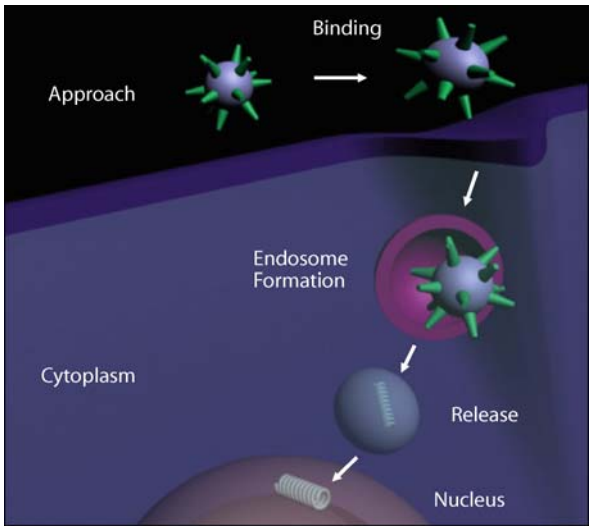

Viral Protein Nano-Actuators, Computational Studies of Bio-nanomachines, Figure 2
Various stages in influenza virus infection. The virus in endocytosed by the cell and an endosome is formed, wherein a conformational change of interest occurs

The native HA1/HA2 complex in the viral envelope is fusion-inactive. The cell-surface receptor needed for the

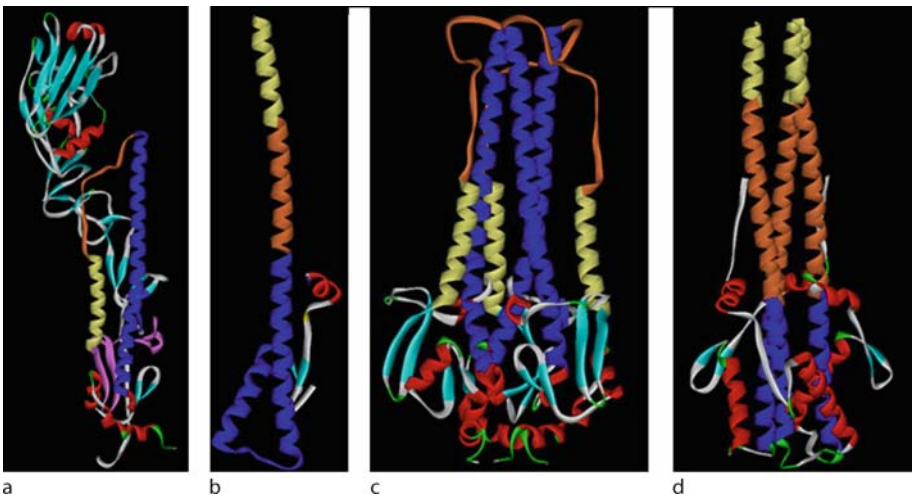

Viral Protein Nano-Actuators, Computational Studies of Bio-nanomachines, Figure 3

a Hairpin like structure of central HA2 native state monomer. The monomer consists of two segments A and B. At the N-terminal is the fusion peptide contained in segment A (residues 1–25, shown in *pink*) followed by short anti-parallel β-sheet (residues 26–37, shown in *cyan*). The outer arm of the hairpin is a small α-helix (residues 38–53, shown *yellow*) which is connected to a long α-helix (residues 82–125, shown *blue*) by a loop region (residues 54–81, shown in *orange*); b HA2 chains A and B in fusogenic states. The loop region (*orange*) converts into a relatively more rigid α-helix and forms a continuation of the long α-helix (*blue*). The short α-helix is thus translated upwards along with the fusion peptide (not shown). c The native state HA2 trimer. The remaining structure other than the helices (*bottom*) is the fusion peptide from the N-terminal and the other smaller secondary structures following the long helices. d Fusogenic state of the HA2 trimer. The helical hairpins open up to form an extended three stranded coiled coil structure with the previously loop regions now helical. Such coiled coil motif is found in many other proteins such as the leucine zipper domain of some transcription factors. The α-helices are wrapped around each other with a left handed superhelical twist

virus to bind onto the cell is known as sialic acid. HA1 binds to this receptor and hence serves the purpose of bringing the virus and cell together. Upon binding with sialic acid, the virus is endocytosed by the cell (Fig. 2). HA remains dormant until the endosome begins to mature, and the pH in the HA surroundings drops to a value of about 5.

At this pH, there is a conformational change in HA2 domain of HA that induces the viral membrane to fuse with the cellular, endosomal membrane, thereby permitting the nucleocapsid of the virus to be deposited into the cytoplasm of the cell. Hence it can be inferred that the acidic pH acts as the physiological trigger for the HA conformational change. Since the low pH also activates membrane fusion, the low pH conformation of HA is also known as fusogenic conformation.

The crystal structures of HA in both the native and the fusogenic conformations are known [8,40,41,45]. As stated above, the HA1 only serves the purpose of cellular recognition and binding to the cell surface receptor and is not of much interest as a nanoactuator. The HA2 subunit, however, is then responsible for the membrane fusion activity. In the native state, the central region of HA2 folds as a helical hairpin-like monomeric structure (Fig. 3a). Two more such subunits (consisting of two segments each) are present to form a trimer together. The long α-helices from each subunit in the native state form a well known structure called the three stranded coiled coil system. Coiled coils can be visualized as an intertwined rope of three interacting helices. As is visible in Fig. 3c and 3d, in addition to the helical domains, there are other regions – small helices, β-sheets, turns and random loop regions. The blue regions in the native stage in Fig. 3c are triple stranded coiled coils wound against each other. They are connected to smaller helices (yellow) in the HA2 domain by loop regions shown in orange. The smaller helices connect to the fusion peptide (residue numbers 1–25). After attachment and pH drop, the orange loop regions convert into extended coiled coils effectively extending the blue coils and moving the fusion peptide by about 10 nm [11].

Coiled coil motifs contain hydrophobic and hydrophilic amino acids in a repeating heptad pattern (positions *a* through *g*) as shown in Figs. 4 and 5. Looking down the axis of the helix, hydrophobic residues tend to occur at the positions *a* and *d* of the heptad repeat and these residues form the interface between the helices. It is believed that HA folds into its thermodynamically most stable state at neutral pH, known as its native state. But

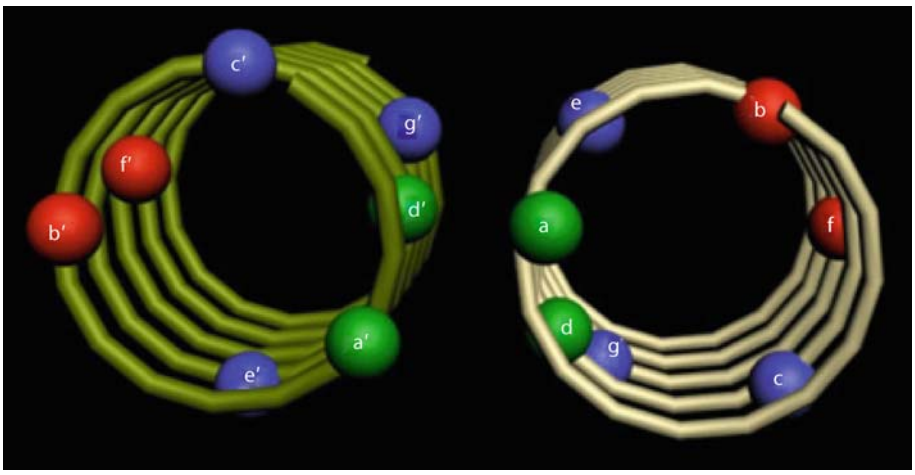

Viral Protein Nano-Actuators, Computational Studies of Bio-nanomachines, Figure 4
Two helices of a coiled coil system. The seven positions (*a*–*e*) denote locations of hydrophobic/hydrophilic residues. Hydrophobic residues (green balls) tend to occur on the inside positions *a* and *d* whereas the hydrophilic residues (red balls) reside in other positions

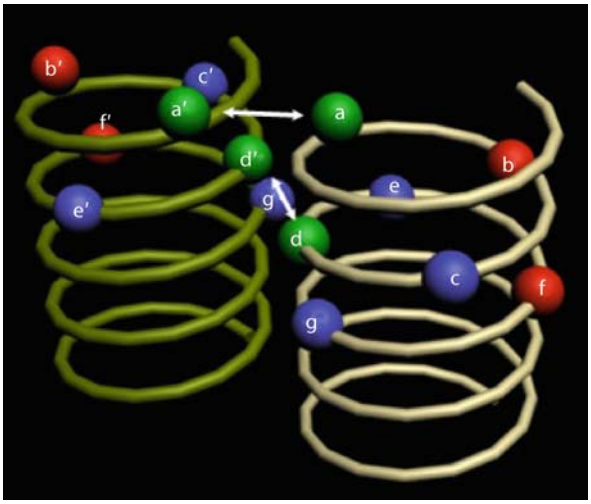

Viral Protein Nano-Actuators, Computational Studies of Bio-nanomachines, Figure 5
Packing of two helices of a coiled coil system. The hydrophobic residues in positions *a* — *a′* and *d* — *d′* of each chain interact with each other to stabilize the coiled coil system

at reduced pH, the fusogenic state becomes thermodynamically more feasible and hence the protein changes conformation in order to achieve it [10]. Another model that is widely accepted suggests that the native state is a metastable state that can be disturbed by any agent, not only pH. It has been observed that heat can also produce this conformational change at neutral pH and so does a denaturant such as urea [10].

In addition to Influenza, the following other viruses show similar infection mechanism and can be candidates for a VPL-like nanoactuator:

(i) The Human Immunodeficiency Virus type 1 (HIV 1) peptide gp41 [12]
(ii) The Human Respiratory Syncytial Virus (HRSV) protein subunit F1 [33]
(iii) The Simian Immunodeficiency Virus (SIV) protein gp41 [9]
(iv) The Human T cell Leukemia virus type 1, protein gp21 [23]
(v) The Simian Parainfluenza Virus peptide unit SV5 [3]
(vi) Ebola virus protein gp2 [42]

Each one of these peptides can result in a different actuator that can have different properties such as weight, volume, range of motion, force and speed capabilities. However, their principle of actuation is the same. Studies have shown [13,39,43] that the common characteristic in these viruses is the structure of a portion of the envelope glycoprotein and the mode of infection.

Projected Robotic Applications of VPL Nanoactuators

Figure 6 shows a schematic of the VPL motor supporting a moving platform. The motor is shown in its initial, "contracted" phase in the first image that corresponds to the virus' native state and at its extended, fusogenic state in the second image of the figure. The total outward protrusion is measured to be 10 nanometers. The estimated time for filling up of the hydrophobic membrane defect by lipids

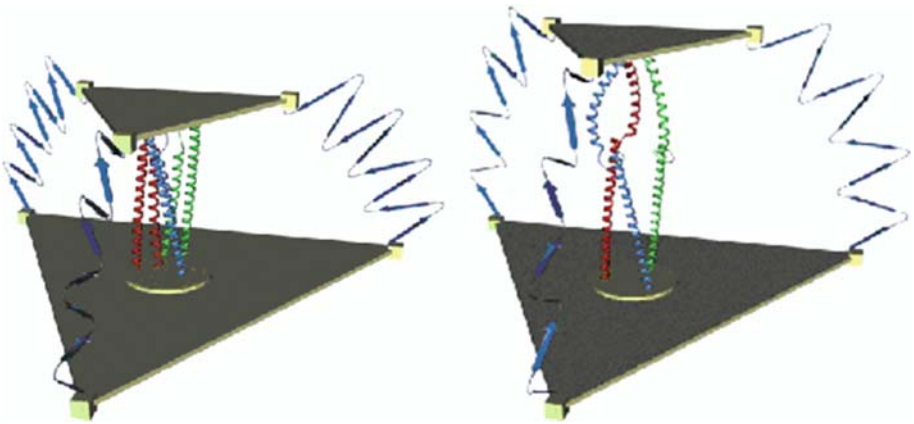

Viral Protein Nano-Actuators, Computational Studies of Bio-nanomachines, Figure 6
A single degree of freedom VPL actuated parallel platform VPL actuator

is 2–4 ns, which gives an idea that the peptide unfolding process takes about the same time. To augment the force capabilities of the VPL motors, several VPL actuating elements could be attached in parallel as it is shown in Fig. 7. Such parallel attachment of multiple VPL motors could result extremely powerful, micro, meso or even macro actuators that will be able to apply ultra large forces while their dimensions are extremely small.

In a similar way, to increase the displacement capability of the VPL motors, several VPL elements could be connected in series (see Fig. 7). For attaching the titin fibers to the platforms DNA joints could be manufactured. By biochemical matching, the platforms could be designed to have proteins that bind to DNA on their corners. DNA strands could then bind onto the corners from one end and to the fibers on the other end to create a flexible joint.

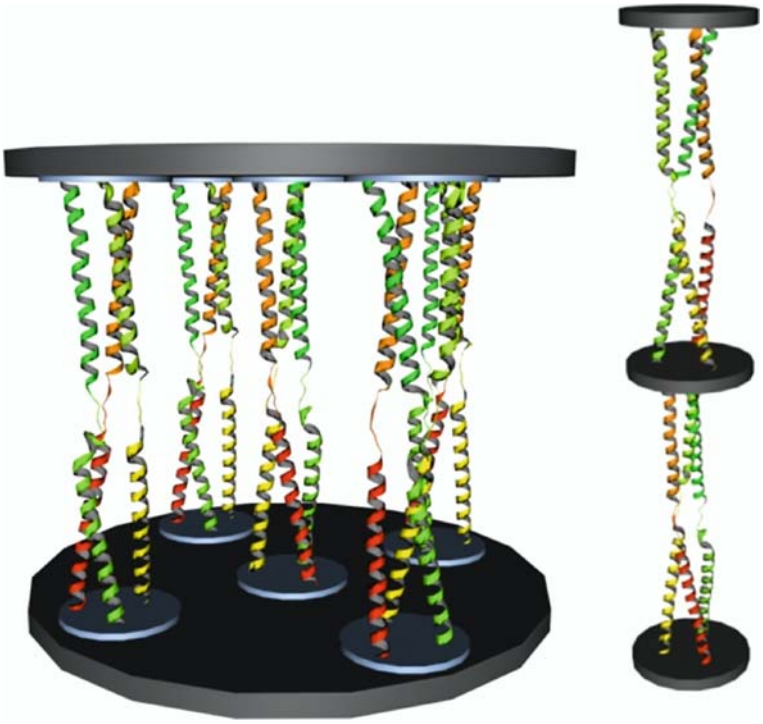

Viral Protein Nano-Actuators, Computational Studies of Bio-nanomachines, Figure 7
Several VPL motors placed in parallel and series

The titin fibers will have to be functionalized in such a way that their ends bind to the DNA as well.

Methods

Targeted Molecular Dynamics (TMD) is a technique used to simulate conformational pathways between two known states by resorting to Newtonian molecular dynamics calculations with a physical potential and a holonomic constraint [29,30]. It provides an approximate modeling of processes that would otherwise span long time-scales (up to milliseconds) in realistic time scales for simulations. Protein folding is a microsecond scale phenomenon [46] and cannot be captured by unbiased MD simulations in realistic time frames. TMD is an example of a biasing technique that drives the trajectory from one known state to another. A number of such techniques have been investigated, for example, internal coordinate constraints [36], minimum biasing methods [20], biased molecular dynamics [27], essential dynamics sampling [14], steered molecular dynamics [22], modified targeted molecular dynamics [2], accelerated molecular dynamics [19] and more recently, the restricted perturbation targeted molecular dynamics (RP-TMD) [37] method.

TMD applies a holonomic constraint that reduces the root mean square deviation (RMSD) of the protein from a predefined target structure by a preset value at each dynamics step [29]. It has been shown that TMD can produce results equivalent to those from traditional MD simulations techniques in a much shorter time by using and lower temperatures. Ferrara et al. [17] have shown using the example of *chymotrypsin inhibitor 2* that the energies of conformation sampled using TMD are in agreement with those obtained from higher temperature unrestrained MD simulations and also with simulations performed using different force fields.

Because the distance to be traveled by the N-terminal of the VPL peptide is relatively very large, unbiased MD techniques cannot be applied to simulate the entire conformational change. High resolution crystal structures of initial and final states of VPL peptides are available from PDB. The closed state was termed as the initial state I and the open state as the final state F. TMD applies the holonomic constraint to reduce the root mean square deviation (RMSD) from I to F by a finite value at each MD step. The 3N position coordinates corresponding to N atoms in the molecule can be defined as:

$$\mathbf{X} = (X_1, X_2, \ldots, X_{3N})^{\mathrm{T}} , \tag{1}$$

where 3N are the Cartesian coordinates of the position vectors $\mathbf{r_1}, \mathbf{r_2}, \ldots, \mathbf{r_N}$ of each individual atom.

The position of an individual atom is time dependent, and is described at time t as $X_i(t)$. Let X_{Fi} be the final (target) position of an atom i. The holonomic constraint is of the form:

$$\Phi[\underline{X}(t)] = \sum \left[\underline{X}_i(t) - \underline{X}_{Fi}\right]^2 - \underline{\rho}^2(t) = 0 . \tag{2}$$

The underlined terms indicate that they have been scaled by a factor of $\sqrt{m_i/\langle m\rangle}$ where m_i is the mass of atom i and $\langle m\rangle$ is the average atomic mass. This mass scaling helps avoid the net translation of the system as a whole [30]. The holonomic constraint in Eq. (2) gives rise to a constraint force of the form:

$$F_i^c = \lambda \nabla_i \Phi = \frac{2\lambda m_i}{\langle m\rangle}(X_i - X_{Fi}) , \tag{3}$$

where λ is a Lagrange multiplier. It is determined by the following method. The Leap-frog [1] propagator is a common method used to update variables along a trajectory of positions. According to this algorithm, the position of an atom i at time t is given by:

$$\begin{aligned} X_i(t) &= \left[X_i(t-\Delta t) + v_i\left(t - \frac{3}{2}\Delta t\right)\Delta t \right. \\ &\quad \left. + \frac{F_i(t-\Delta t)}{m_i}(\Delta t)^2\right] + \frac{F_i^c(t-\Delta t)}{m_i}(\Delta t)^2 \\ &= x_i(t) + p_i(t) \end{aligned} \tag{4}$$

In the above equation, Δt is the time increment, F_i is the force on atom i due to unbiased Newtonian molecular dynamics, and v_i is the velocity of the atom. The variable $x_i(t)$ is the position of the atom in the absence of the constraint and $p_i(t)$ is the perturbation due to the holonomic constraint [37]. Noting that:

$$p_i(t) = \frac{F_i^c(t-\Delta t)}{m_i}(\Delta t)^2$$

from Eq. (4), and

$$F_i^c(t-\Delta t) = \frac{2\lambda m_i}{\langle m\rangle}\{X_i(t-\Delta t) - X_{Fi}\}$$

from Eq. (3), and defining

$$\gamma = \frac{2(\Delta t)^2\lambda}{\langle m\rangle}$$

we get:

$$p_i(t) = \gamma[X_i(t-\Delta t) - X_{Fi}] . \tag{5}$$

Hence the perturbation on each atom is now simply the product of a Δt dependent scaling factor γ and the difference in position between the current and previous time

step. In order to obtain γ (and hence λ), substitute Eq. (4) into the original constraint Eq. (2) to get:

$$\Phi[\underline{X}(t)] = \sum \left[\underline{X}_i(t) - \underline{X}_{Fi}\right]^2 - \underline{\rho}^2(t) = 0\,.$$

Which can be written as:

$$\sum \left[\left\{\underline{x}_i(t) + \underline{p}_i(t)\right\} - \underline{X}_{Fi}\right]^2 - \underline{\rho}^2(t) = 0\,,$$

and further simplified as:

$$\begin{aligned}
&\sum \Big[\left\{\underline{x}_i(t) + \underline{p}_i(t)\right\}^2 \\
&\qquad - 2\left\{\underline{x}_i(t) + \underline{p}_i(t)\right\}\underline{X}_{Fi} - \underline{X}_{Fi}^2\Big] \\
&\qquad - \underline{\rho}^2(t) = 0 \\
&\sum \Big[\left\{\underline{x}_i^2(t) + \underline{p}_i^2(t) + 2\underline{x}_i(t)\underline{p}_i(t)\right\} \\
&\qquad - 2\underline{x}_i(t)\underline{X}_{Fi} - 2\underline{p}_i(t)\underline{X}_{Fi} - \underline{X}_{Fi}^2\Big] \\
&\qquad - \underline{\rho}^2(t) = 0 \\
&\sum \Big[\left\{\underline{p}_i^2(t) + 2\underline{x}_i(t)\underline{p}_i(t) - 2\underline{p}_i(t)\underline{X}_{Fi}\right\} \\
&\qquad + \underline{x}_i^2(t) - 2\underline{x}_i(t)\underline{X}_{Fi} + \underline{X}_{Fi}^2\Big] \\
&\qquad - \underline{\rho}^2(t) = 0 \\
&\sum \Big[\underline{p}_i^2(t) + 2\underline{p}_i(t)\{\underline{x}_i(t) - \underline{X}_{Fi}\}\Big] \\
&\qquad + \sum \left[\underline{x}_i(t) - \underline{X}_{Fi}\right]^2 - \underline{\rho}^2(t) = 0\,.
\end{aligned} \tag{6}$$

Using Eq. (5) into (6) yields:

$$\begin{aligned}
&\sum \big[\gamma^2\{\underline{X}_i(t-\Delta t) - \underline{X}_{Fi}\}^2 + 2\gamma\{\underline{X}_i(t-\Delta t) - \underline{X}_{Fi}\} \\
&\cdot \{\underline{x}_i(t) - \underline{X}_{Fi}\}\big] + \sum [\underline{x}_i(t) - \underline{X}_{Fi}]^2 - \underline{\rho}^2(t) = 0\,,
\end{aligned}$$

which can be rearranged as –

$$\begin{aligned}
&\gamma^2 \sum \left[\{\underline{X}_i(t-\Delta t) - \underline{X}_{Fi}\}^2\right] \\
&+ \gamma \sum 2[\{\underline{X}_i(t-\Delta t) - \underline{X}_{Fi}\}\{\underline{x}_i(t) - \underline{X}_{Fi}\}] \\
&\qquad + \sum [\underline{x}_i(t) - \underline{X}_{Fi}]^2 - \underline{\rho}^2(t) = 0\,.
\end{aligned} \tag{7}$$

Defining,

$$\begin{aligned}
a &= \sum \left[\{\underline{X}_i(t-\Delta t) - \underline{X}_{Fi}\}^2\right], \\
b &= 2\sum [\{\underline{X}_i(t-\Delta t) - \underline{X}_{Fi}\}\{\underline{x}_i(t) - \underline{X}_{Fi}\}]\,.
\end{aligned}$$

And noting that:

$$\Phi\left[\underline{x}(t)\right] = \sum \left[\underline{x}_i(t) - \underline{X}_{Fi}\right]^2 - \underline{\rho}^2(t)\,.$$

Equation (7) can be simply written as –

$$a\gamma^2 + b\gamma + \Phi[\underline{x}(t)] = 0\,. \tag{8}$$

The Eq. (8) yields two roots for γ. In order to minimize the total perturbation $\sum |p_i|$ (from Eq. (5)), the root with lower absolute value is chosen.

The RMSD is decreased by a predetermined (and user defined) value at each time step by a value of $\Delta\rho$ as:

$$\rho(t) = \rho(t - \Delta t) - \Delta\rho\,. \tag{9}$$

The TMD algorithm steps are:

(i) Set $\rho = \rho_0 = |\mathbf{x}_\mathrm{I} - \mathbf{x}_\mathrm{F}|$ where I is the initial and F is the final conformation.
(ii) Numerically solve the equations of motion with the additional constraint force F^c.
(iii) After each time step Δt, diminish ρ as in Eq. (9).

At the end of the simulation, the final distance ρ_f is reached. In this way, a monotonous reduction of ρ forces the system to find a pathway from $\mathbf{x}_\mathrm{I}$ to a final configuration $\mathbf{x}_\mathrm{F}$.

Simulation Results and Discussion

Molecular Dynamics Simulations were performed on VPL peptides in order to study the performance of the nanoactuator. Specifically, a fragment from the VPL segment, the loop36 was chosen for these studies due to its established role in the conformational change [11]. The large conformational change was simulated using the Targeted Molecular Dynamics (TMD) technique. The TMD analyzes were performed using implicit solvent conditions. In order to simulate the effect of environmental pH change on the VPL motor protein, acidic amino acids were protonated. In the TMD simulations the protein was allowed to undergo a transition from one known state to another. This is a useful approach to ensure the feasibility of the conformational change and to get an idea of the magnitude of the energy change involved. The VPL nanoactuator works in a solution where there are protons bombarding the protein due to a reduction in the solution pH. This allows acidic amino acids in the protein to accept the protons on their negatively charged sites, and hence get protonated. Protonation changes the energy balance of the solvent-protein system and can create conditions that promote a conformational change. Low pH simulations on the small fragment loop36 were performed to study the effect on the protein conformation and the way it interacts energetically with its surroundings.

In addition to the TMD analyzes, classical molecular dynamics approaches were performed to quantify some

other aspects such as helicity and free energy of the peptides. In order to be able to perform classical unbiased MD simulations in a realistic time scale, it is important to energize the system by using high temperatures. The open and closed structures were taken from the protein data bank (PDB) with the open state as the target state and the closed state as the starting point [16]. The simulations were performed using an implicit solvation model known as Effective Energy Function for proteins EEF1 [24].

Structural Details of Loop36

Loop36 is a 36 residue long peptide of the VPL protein that forms a hinge region of the viral protein joining the two α-helical regions of each monomer [8,11]. An initial state as obtained from PDB [6] file 1HGF at pH of about 7.0, loop36 consists of a 15 residue long α-helical part with the remaining fragment in a random and flabby form (Fig. 8). It is located in the segment B of the Influenza Hemagglutinin protein sequence from residues 54 to 89. It has been recognized as being critical for a pH dependent conformational change [10,11]. The loop36 wild type (naturally occurring) sequence is as follows:

ARG VAL ILE GLU LYS THR ASN GLU LYS PHE HIS GLN ILE GLU LYS GLU PHE SER GLU VAL GLU GLY ARG ILE GLN ASP LEU GLU LYS TYR VAL GLU ASP THR LYS ILE

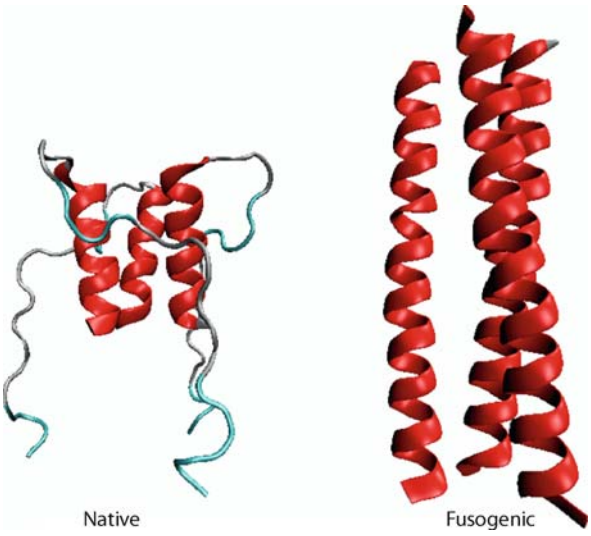

Viral Protein Nano-Actuators, Computational Studies of Bio-nanomachines, Figure 8
The peptide loop36 in the initial and final states of the conformational change. Coiled coils are shown in red and the random loop regions in white and cyan. The random loop regions rise and convert into α-helical coiled coils upon activation

The initial and final states of *loop36* as obtained by the crystal structures 1HGF (pH $\sim$ 7.0) and 1HTM (pH $\sim$ 5.0) from PDB are shown in Fig. 8.

In order to simulate low pH, the biochemical response of certain amino acids to the acidic conditions was considered. Some amino acids are acidic in nature, some basic and some neutral; and they get protonated at different pH values. Out of the 20 constituent types of amino acids, Glutamic Acid (GLU), Aspartic Acid (ASP), and Histidine (HIS) are in their unprotonated states at neutral pH (7.0). There is only one histidine residue in loop36, whereas there are eight glutamic acid residues and two aspartic acid residues which make a large part of loop36. Hence, overall, there were eleven residues out of the 36 that were protonated.

Figure 9 shows the Visual Molecular Dynamics (VMD) [21] rendered transition of a loop36 monomer from initial (closed) state to an open (fusogenic) state. The entire transition is achieved in 66 picoseconds when using TMD, while it is not achievable in 30 nanoseconds using traditional MD techniques. As seen in the figure, after 20 ps the loop regions of the peptide are halfway open and after 50 ps they are completely opened but have not attained an α-helical conformation.

The deviation of the final state from the initial state of the monomer in Fig. 9 can be seen in the root mean square deviation (RMSD) from the open state. Results are depicted in Fig. 10 and values of RMSD indicate that the open and close conformation differ by approximately 19 Å. The rate of opening of the monomer can be estimated by the slope of this curve which in this case is 0.2 Å/ps. It must be kept in mind that to have an estimation of the real rate of opening, this value should be rescaled to account for the increase in speed of the process given by TMD.

Mutational Analysis

Mutations in a protein can be made by replacing one or more constituent amino acids with different amino acids. The structure and properties of a protein change when they undergo a mutation. It is therefore possible to design a protein according to specific requirements; for example, if a high hydrophobicity is desired, hydrophobic amino acids should replace some of the residues.

The loop36 peptide contains one Glycine residue which is known as a helix-breaker amino acid because its side chain is simply one hydrogen atom. In experimental studies, it is observed that replacing this residue by an Alanine results in better α-helix formation [11]. In line with the experiments, in our simulations the Glycine is replaced

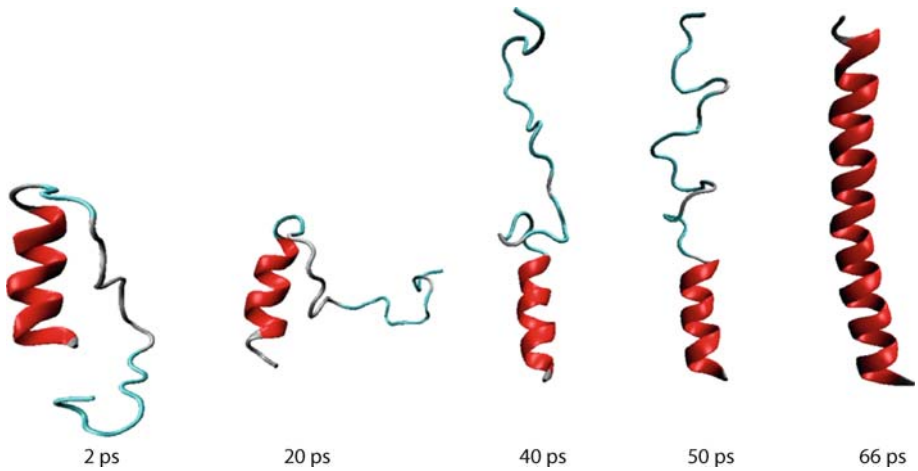

Viral Protein Nano-Actuators, Computational Studies of Bio-nanomachines, Figure 9
Transition of a loop36 monomer from native to fusogenic state during a TMD simulation

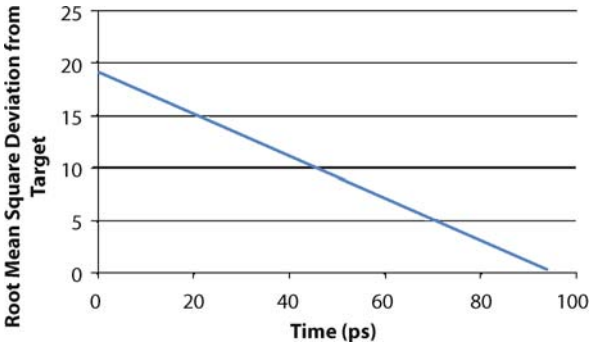

Viral Protein Nano-Actuators, Computational Studies of Bio-nanomachines, Figure 10
In the beginning the RMSD from the final state has a maximum value of about 19 Å and it decreases linearly with time as the open state is achieved due to the holonomic constraint applied by TMD

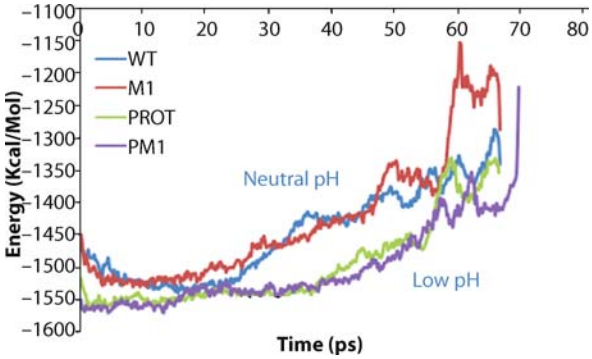

Viral Protein Nano-Actuators, Computational Studies of Bio-nanomachines, Figure 11
Comparison of potential energy paths of four types of loop36 monomers. (i) wild type WT (the protein in this state does not have any mutation or protonation), (ii) mutated type M1 (in this state, GLY (glycine) is replaced by ALA (alanine), (iii) protonated type PROT (GLU (glutamic acid)) is replaced by protonated GLU, ASP (aspartic acid) by protonated ASP, HIS (histidine) replaced by doubly protonated histidine and (iv) protonated and mutated type PM1 (in this state, the protein has both protonation and mutation). The low pH structures follow lower energy pathways than the neutral pH structures although all structures eventually achieve similar endstate energy values

with the Alanine. The glycine (GLY) residue is the 22nd residue in loop36, and a mutated version of the loop36 with GLY replaced by alanine (ALA) is called mutation M1. The unmutated version of the peptide is simply called the wild type (WT). GLY contains no side chain and it has no atom attached to the α-carbon (CA). On the other hand, ALA contains a simple side chain (CH3) attached to its CA atom.

The four different cases compared in this study are as follows:

(i) the wild type (no mutation or protonation) guided to an open state
(ii) one mutation M1
(iii) protonation of amino acids (with no mutations), and
(iv) mutation and protonation (PM1). Conformational energies of these simulations are compared in Fig. 11.

H-Bond Analysis

The intra-helical H-bonds stabilize α-helical structures. The hydrogen bonds for the loop36 conformational transition (Fig. 12) were found to be oscillatory in nature and hence a moving average with a period of 15 time units was

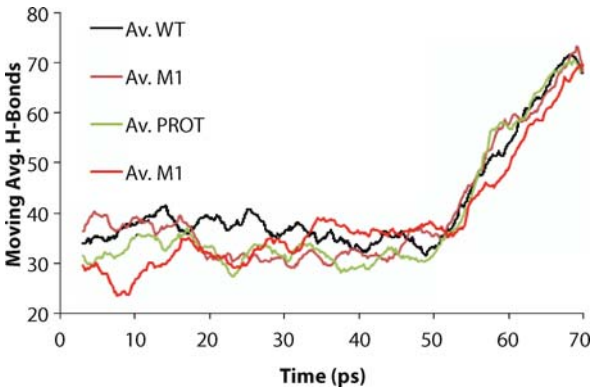

Viral Protein Nano-Actuators, Computational Studies of Bio-nanomachines, Figure 12
The number of H-bonds in each of the monomer WT, M1, PROT and PM1 modified structures varies as the open state is achieved. A moving average of 15 points (3 ps) is shown in each curve. The H-bonds oscillate around a constant value up to about 50 ps in each of the cases. A steep rise in the number is observed beyond this point. This region corresponds to the helix formation which results in intra-helical H-bond formation

chosen. Each observation was made every 0.2 ps hence 15 points correspond to 3 ps time interval. In the initial stages of the simulation (time < 50 ps) the partial helix of loop36 is the only region with a significant number of H-bonds. The random loop region is converted into α-helix beyond 50 ps and hence a steep increase in the numbers is observed. The variation in the open state contacts, namely the H-bonds quantitatively describe the achievement of the open state.

Classical Molecular Dynamics Studies

Even though classical methods cannot capture a large conformational change such as that exhibited in the VPL nanoactuator they provide insight into important structural details and properties. Typically higher temperatures are used to accelerate structural events in macromolecule simulations. When the protein is exposed to high temperatures, not only can it fold/unfold but it may also become denatured. Heating a protein corresponds to increasing the kinetic energy of its atoms, and this, in turn causes the molecules to vibrate more, possibly disrupting many of the hydrogen bonds and non-polar interactions and hence the secondary structure of the protein. The classical simulations on VPL protein were performed to understand the behavior of the VPL trimer at elevated temperatures. Knowledge of this will also help to ascertain whether there is a range of temperatures in which changes of structure such as folding or α-helix formation were initiated, before the protein denatures [15]. Figure 13 shows temperature effects on the VPL trimer loop50, which is an extended loop36 trimer. In contrast to the TMD simulations in which the open state was achievable in about 60 ps, the classical simulations need an approximate time of the order of few microseconds to achieve the same results [18].

It has been shown experimentally [28] that viral membrane fusion in the hemagglutinin of influenza virus oc-

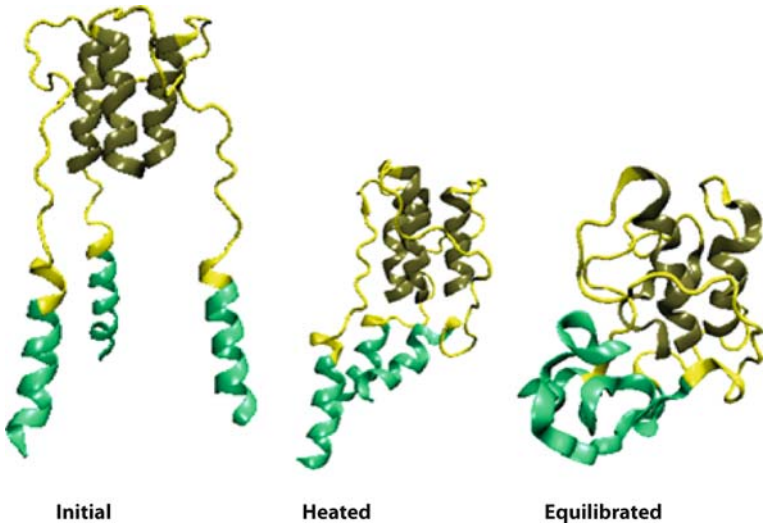

Viral Protein Nano-Actuators, Computational Studies of Bio-nanomachines, Figure 13
Effect of temperature on VPL peptide. The heated state corresponds to 500 °C temperature achieved in 150 ps and the equilibrated state corresponds to a constant temperature equilibration for up to 6 ns. The peptide gets denatured at this point

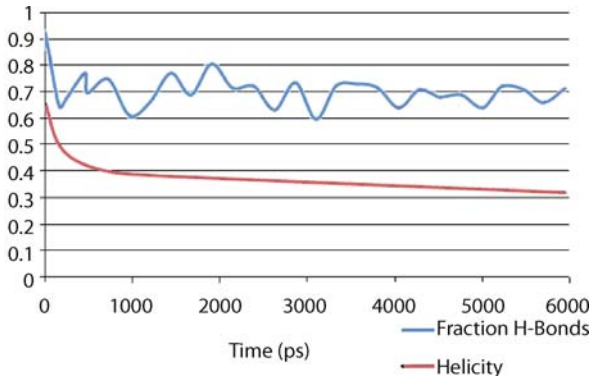

Viral Protein Nano-Actuators, Computational Studies of Bio-nanomachines, Figure 14

Helicity and fraction of open state contacts for a 6 ns classical high temperature molecular dynamics simulation

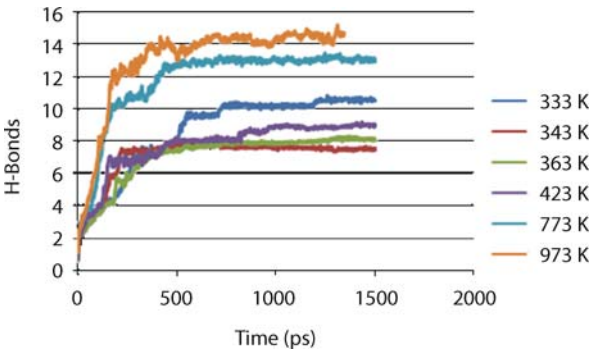

Viral Protein Nano-Actuators, Computational Studies of Bio-nanomachines, Figure 15

RMSD of the loop50 trimer from its initial state for each of the simulations ranging from 333 K to 973 K. Higher temperatures take the peptide further away from the initial state which indicates a larger conformational change. However, there is significant denaturation of the peptide at 773 K and 973 K

curs at 335 K. The loop50 trimer was subjected to a range of simulated temperatures ranging from 333 K to 973 K with an attempt to capture the conformational changes required for fusion and hence the performance of the VPL nanoactuator. The RMSD from the initial equilibrated state at various temperatures is shown in Fig. 15. The expected trend would be that increasing temperature increases the RMSD from the initial state as it allows the peptide to jump energy barriers and denatures it. This was observed in simulations at 773 K and 973 K. However, the 333 K simulation indicates that the peptide has a larger difference with the initial state (larger RMSD values) in the same time frame than those at 343 K, 363 K and 423 K. This result is in agreement with the experimental observation and suggests that larger conformational changes may

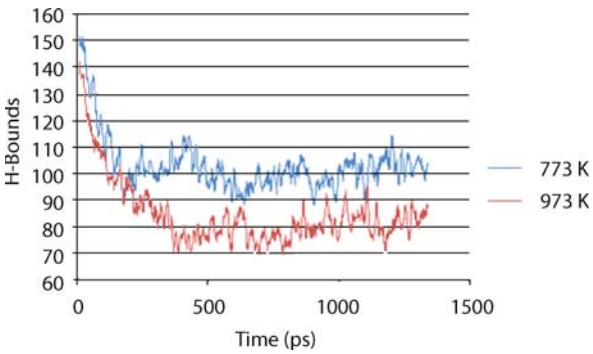

Viral Protein Nano-Actuators, Computational Studies of Bio-nanomachines, Figure 16

H-bonding data for the conformational changes achieved at 773 K and 973 K. The higher temperature results in loss of H-bonds

take place in the temperature range of 333–343 K [11]. The extremely high temperatures 773 K and 973 K result in denaturation of the peptides as seen by the H-bond data in Fig. 16.

Conclusions

The recent explosion in nanotechnology has given place to the creation of biomolecular machines and motors such as viral protein linear actuators (VPL). The idea of the VPL motors, whose constituent proteins belong to a family of retroviruses such as HIV and influenza virus, originates from the mechanism they utilize to infect a cell. When these types of viruses are endocytosed by the cell, they experience a drop in pH and change their structure from a coil strand to an alpha helix conformation. In order to exploit the functional mechanism of viral proteins, it is important to characterize and understand the sensing capabilities of the peptides under different conditions of the environment. In this study, some aspects of the response of the VPL motor to different stimuli, using molecular dynamics have been illustrated. For the study of nanoactuators, in particular, molecular simulations provide a very valuable tool, especially since sometimes; experiments at the nanoscale may not be simple to perform. Alteration to the protein sequence (mutations) is a good example of a process in which computational studies are much easier to perform than experimental ones due to the complexities involved. One of the challenges is to learn how to control structural mechanisms, behavior and properties of the basic nanocomponents involved in molecular motors. In this study we learn about changes in the structure experienced by the viral protein actuator monomer as it interacts with other monomers in con-

ditions of trimerization, mutation of its sequence, and changes in temperature.

The VPL undergoes a remarkably large conformational change upon a drop in pH. In order to simulate the motion of the protein, targeted and traditional MD techniques are employed. Structural changes induced by pH variations are studied by protonating a number of amino acids in the peptide sequence. The TMD results showed a clear preference to conformational transition paths at low pH values. Our findings confirm that the protein forms a helical coil at acidic pH values but it shows a coiled structure at neutral pH. More studies are needed to establish conclusive results on the effect of pH and mutations.

The experimental observation that influenza hemagglutinin peptides showed a conformational change with a significant formation of alpha helix occurs in the wild type (unmutated) at 65°C at neutral pH [10]. In our simulations of a reduced loop50 trimer, this conformational change was not captured within 6 ns. Given the remarkably short time frames of our simulations and the size of the peptides, we are unable to observe regeneration of alpha helix. Even more, in order to be able to observe visible conformational changes, we had to work at very high temperatures. Even at that value of the temperature we do not see complete denaturation, again due to the short time frames under study. In an experimental situation the protein denatures at a much lower value of the temperature (at approximately 71°C). At these relatively low temperatures, a much longer time scale (micro- or even millisecond is required to fully capture alpha helix regeneration.

It is important to mention that high temperature approaches remain a valuable tool for conformational exploration; one must understand, however, that torsional transition kinetics are not correct at higher temperatures. Furthermore, rescaling to lower temperatures yields only approximate results. In spite of this, our goal is to present a methodology for studying large conformational changes in nanoscale molecular systems using both targeted and traditional molecular dynamics techniques that can ultimately be extended to much longer time frames using parallel computing.

We are at the dawn of a new era in the development of molecular machinery, and we are just starting to elucidate some of the challenges that these nanoactuators represent. Future studies will involve investigating interfacing with other molecular components such as carbon nanotubes, biological membranes, inorganic substrates, ions etc. As mentioned earlier, the end-effector of VPL can be designed according to the requirement. Moreover, the composition of the peptide can be varied according to the environment, as long as the rules of having hydrophobic and polar residues to ensure stability of the coiled-coil system are adhered to. The length of the VPL peptides, the role of the solvent (and solvent composition), binding energies to various target objects, are other variables that need to be investigated in order to assess conditions for optimum performance, (i. e. for example maximum force and stability, a given velocity etc.)

The VPL peptides are known to work better with specific mutations [11]. This opens up doors for further mutational analysis of the peptide, again directed towards achieving optimal performance parameters. The goal will be to determine the optimal sequence of the VPL peptide that results in the most stable, quickest, powerful and robust nanoactuator.

Future Directions

The future of this study involves the use of larger peptides, employment of better computational algorithms and interfacing of the VPL nanoactuator to other nanodevices to produce complex nanosystems. This work has shown great potential for further research. Better algorithms need to be developed for the study of the forces exerted by the nanoactuator. These studies have concentrated on developing a suitable approach to characterize and study conformational changes of loop36 and loop50. Preliminary studies have shown that longer peptides, such as loop114 (an extended version of loop50) are promising nanoactuator candidates as they provide a larger displacement ($\sim$ 9 nm) and they have a larger helical structure. The added secondary structure may provide more stability and larger forces to the actuator.

Mutations M1 to M7 were applied to loop36. Future studies will involve mutations to loop50 and larger peptides, which provide more mutation points and can significantly alter the performance of the actuator. The VPL nanoactuator can also be interfaced with other protein/lipid based molecular platform/devices such as lipid bilayer membranes. A possible scenario is to attach a cysteine tag to the end of VPL peptide and a lipid bilayer via a disulfide bond. TMD and Steered Molecular Dynamics (SMD) [22] can be used to study the performance of the nanoactuator while bound to a platform.

The TMD algorithm that was used to simulate conformational changes by applying a bias in the force field in this study has recently been succeeded by an updated restricted perturbation TMD (RP-TMD) [37] algorithm. The RP-TMD method depends not only on reduction of RMSD at each time step, but also on the minimal energy path. The two-way check yields lower energy pathways

for the trajectories. The TMD algorithm has a few disadvantages:

(i) Transition is based solely on RMSD, which is a geometric parameter. No attempt is made to ensure that the conformational transition follows the minimum possible energy pathway.
(ii) Due to the above, large energy barriers are crossed which may not be realistic.
(iii) Some trajectories obtained are not reversible.

In the RP-TMD method, the above deficiencies are considerably reduced.

Acknowledgments

The authors gratefully acknowledge support of the National Science Foundation (Grant NSF NIRT #0303950).

Bibliography

1. Allen M, Tildesley D (1987) Computer Simulation of Liquids. Oxford University Press, New York
2. Apostolakis J, Ferrara P, Caflisch A (1999) Calculation of conformational transitions and barriers in solvated systems: application to the alanine dipeptide in water. J Chem Phys 110(4):2099–2108
3. Baker KA et al (1999) Structural basis for paramyxovirus-mediated membrane fusion. Mol Cell 3(3):309–19
4. Balzani VV et al (2000) Artificial Molecular Machines. Angew Chem Int Edn Engl 39(19):3348–3391
5. Berg HC (2003) The Rotary Motor of Bacterial Flagella. Annu Rev Biochem 72:19–54
6. Berman HM et al (2000) The Protein Data Bank. Nucl Acids Res 28(1):235–42
7. Boyer PD (1998) Energy, life, and ATP. Biosci Rep 18(3):97–117
8. Bullough PA et al (1994) Structure of influenza haemagglutinin at the pH of membrane fusion. Nature 371(6492):37–43
9. Caffrey M et al (1998) Three-dimensional solution structure of the 44 kDa ectodomain of SIV gp41. EMBO J 17(16):4572–84
10. Carr CM, Chaudhry C, Kim PS (1997) Influenza hemagglutinin is spring-loaded by a metastable native conformation. Proc Natl Acad Sci USA 94(26):14306–13
11. Carr CM, Kim PS (1993) A spring-loaded mechanism for the conformational change of influenza hemagglutinin. Cell 73(4):823–32
12. Chan DC et al (1997) Core structure of gp41 from the HIV envelope glycoprotein. Cell 89(2):263–73
13. Colman PM, Lawrence MC (2003) The structural biology of type I viral membrane fusion. Nat Rev Mol Cell Biol 4(4):309–19
14. Daidone I et al (2003) Molecular Dynamics Simulation of Protein Folding by Essential Dynamics Sampling: Folding Landscape of Horse Heart Cytochrome c. Biophys J 85(5):2865–2871
15. Dubey A, Mavroidis C, Tomassone MS (2006) Molecular Dynamic Studies of Viral-Protein Based Nano-Actuators. J Comput Theor Nanosci 3:885
16. Dubey A, Sharma G, Mavroidis C, Tomassone MS, Nikitczuk K, Yarmush ML (2004) Computational Studies of Viral Protein Nano-Actuator. J Comput Theor Nanosci 1(1):1–11
17. Ferrara P, Apostolakis J, Caflisch A (2000) Targeted Molecular Dynamics Simulations of Protein Folding. J Phys Chem B 104:4511–4518
18. Ferrara P, Apostolakis J, Caflisch A (2000) Thermodynamics and Kinetics of Folding of Two Model Peptides Investigated by Molecular Dynamics Simulations. J Phys Chem B 104(20): 5000–5010
19. Hamelberg D, Mongan J, McCammon JA (2004) Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. J Chem Phys 120(24):11919–29
20. Harvey SC, Gabb HA (1993) Conformational transitions using molecular dynamics with minimum biasing. Biopolymers 33(8):1167–72
21. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. J Mol Graph 14(1):27–8, 33–8
22. Isralewitz B, Gao M, Schulten K (2001) Steered molecular dynamics and mechanical functions of proteins. Curr Opin Struct Biol 11(2):224–230
23. Kobe B et al (1999) Crystal structure of human T cell leukemia virus type 1 gp21 ectodomain crystallized as a maltose-binding protein chimera reveals structural evolution of retroviral transmembrane proteins. Proc Natl Acad Sci USA 96(8): 4319–24
24. Lazaridis T, Karplus M (1999) Effective energy function for proteins in solution. Proteins 35(2):133–52
25. Lazarowitz SG, Compans RW, Choppin PW (1971) Influenza virus structural and nonstructural proteins in infected cells and their plasma membranes. Virology 46(3):830–43
26. Oster G, Wang H (2003) Rotary protein motors. Trends Cell Biol 13(3):114–21
27. Paci E, Karplus M (1999) Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations. J Mol Biol 288(3):441–59
28. Ruigrok RW et al (1986) Conformational changes in the hemagglutinin of influenza virus which accompany heat-induced fusion of virus with liposomes. Virology 155(2):484–97
29. Schlitter J (1994) Targeted Molecular Dynamics: A New Approach for Searching Pathways of Conformational Transitions. J Mol Graph 12:84–89
30. Schlitter JEM, Krüger P, Jacoby E, Wollmer A (1993) Targeted molecular dynamics simulation of conformational change: application to the T↔R transition in insulin. Mol Simul 10(2–6):291–308
31. Schoch C, Blumenthal R, Clague MJ (1992) A long-lived state for influenza virus-erythrocyte complexes committed to fusion at neutral pH. FEBS Lett 311(3):221–5
32. Seeman NC (2003) DNA in a material world. Nature 421(6921):427–31
33. Singh M, Berger B, Kim PS (1999) LearnCoil-VMF: computational evidence for coiled-coil-like motifs in many viral membrane-fusion proteins. J Mol Biol 290(5):1031–41
34. Skehel JJ, Waterfield MD (1975) Studies on the primary structure of the influenza virus hemagglutinin. Proc Natl Acad Sci USA 72(1):93–7
35. Skehel JJ, Wiley DC (2000) Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. Annu Rev Biochem 69:531–69
36. Tobias DJ, Brooks CL III (1988) Molecular dynamics with internal coordinate constraints. J Chem Phys 89:5115–5127
37. van der Vaart A, Karplus M (2005) Simulation of conformational

transitions by the restricted perturbation–targeted molecular dynamics method. J Chem Phys 122:114903
38. Vale RD, Milligan RA (2000) The way things move: looking under the hood of molecular motor proteins. Science 288(5463):88–95
39. Walker JE (1998) ATP Synthesis by Rotary Catalysis (Nobel Lecture). Angew Chem Int Edn 37:2308–2319
40. Weis WI et al (1988) Structure of the influenza virus haemagglutinin complexed with its receptor, sialic acid. Nature 333(6172):426–31
41. Weis WI et al (1990) Refinement of the influenza virus hemagglutinin by simulated annealing. J Mol Biol 212(4):737–61
42. Weissenhorn W et al (1998) The central structural feature of the membrane fusion protein subunit from the Ebola virus glycoprotein is a long triple-stranded coiled coil. Proc Natl Acad Sci USA 95(11):6032–6
43. Weissenhorn W et al (1999) Structural basis for membrane fusion by enveloped viruses. Mol Membr Biol 16(1):3–9
44. Wiley DC, Skehel JJ (1987) The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. Annu Rev Biochem 56:365–94
45. Wilson IA, Skehel JJ, Wiley DC (1981) Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 A resolution. Nature 289(5796):366–73
46. Yang W et al (2003) The missing link between thermodynamics and structure in F1-ATPase. Proc Natl Acad Sci USA 100(3): 874–9
47. Yurke B, Turberfield AJ, Mills AP, Simmel FC, Neumann JL (2000) A DNA-Fuelled Molecular Machine Made of DNA. Nature 415:62–65

Volcanic Eruptions: Cyclicity During Lava Dome Growth

Oleg Melnik[1,2], R. Stephen J. Sparks[2], Antonio Costa[2,3], Alexei A. Barmin[1]
[1] Institute of Mechanics, Moscow State University, Moscow, Russia
[2] Earth Science Department, University of Bristol, Bristol, UK
[3] Istituto Nazionale di Geofisica e Vulcanologia, Naples, Italy

Article Outline

Glossary
Definition of the Subject
Introduction
Dynamics of Magma Ascent During Extrusive Eruptions
Short-Term Cycles
Long-Term Cycles
Future Directions
Acknowledgments
Bibliography

Glossary

Andesite Magma or volcanic rock is characterized by intermediate SiO_2 concentration. Andesite magmas have rheological properties that are intermediate between basalt and rhyolite magmas. Silica content in andesites ranges from approximately 52 to 66 weight percent. Common minerals in andesite include plagioclase, amphibole and pyroxene. Andesite is typically erupted at temperatures between 800 to 1000°C Andesite is particularly common in subduction zones, where tectonic plates converge and water is introduced into the mantle.

Basalt Magma or volcanic rock contains not more than about 52% SiO_2 by weight. Basaltic magmas have a low viscosity. Volcanic gases can escape easily without generating high eruption columns. Basalt is typically erupted at temperatures between 1100 to 1250°C. Basalt flows cover about 70% of the Earth's surface and huge areas of the terrestrial planets and so are the most important of all crustal igneous rocks.

Bingham liquid is a fluid that does not flow in response to an applied stress until a critical yield stress is reached. Above the critical yield stress, strain rate is proportional to the applied stress, as in a Newtonian fluid.

Bubbly flow A multi-phase flow regime, in which the gas phase appears as bubbles suspended in a continuous liquid phase.

Conduit A channel, through which magma flows towards the Earth's surface. Volcanic conduits can commonly be approximately cylindrical and typically a few 10's meters across or bounded by near parallel sides in a magma-filled fracture. Conduits can be vertical or inclined.

Crystallization Conversion, partial or total, of a silicate melt into crystals during solidification of magma.

Degassing n. (degas v.) The process by which volatiles that are dissolved in silicate melts come out of solution in the form of bubbles. Open- and closed-system degassing can be distinguished. In the former, volatiles can be lost or gained by the system. In the latter, the total amount of volatiles in the bubbles and in solution in the magma is conserved.

Differentiation The process of changing the chemical composition of magma by processes of crystallization accompanied by separation melts from crystals.

Dome A steep-sided, commonly bulbous extrusion of lava or shallow intrusion (cryptodome). Domes are commonly, but not exclusively, composed of SiO_2-rich magmas. In dome-forming eruptions the erupted

magma is so viscous, or the discharge rate so slow, that lava accumulates very close to the vent region, rather than flowing away. Pyroclastic flows can be generated by collapse of lava domes. Recent eruptions producing lava domes include the 1995–2006 eruption of the Soufrière Hills volcano, Montserrat, and the 2004–2006 eruption of Mount St. Helens, USA.

Dyke A sheet-like igneous intrusion, commonly vertical or near vertical, that cuts across pre-existing, older, geological structures. During magmatism, dykes transport magma toward the surface or laterally in fracture-like conduits. In the geologic record, dykes are preserved as sheet-like bodies of igneous rocks.

Explosive eruption A volcanic eruption in which gas expansion tears the magma into numerous fragments with a wide range of sizes. The mixture of gas and entrained fragments flows upward and outward from volcanic vents at high speed into the atmosphere. Depending on the volume of erupted material, eruption intensity and sustainability, explosive eruptions are classified as Strombolian, Vulcanian, sub-Plinian, Plinian or Mega–Plinian; this order is approximately in the order of increasing intensity. Strombolian and Vulcanian eruptions involve very short-lived explosions.

Extrusive flow or eruption A non-explosive (non-pyroclastic) magma flow from a volcanic conduit during a lava dome-building eruption or lava flow.

Mafic Magma, lava, or tephra with silica concentrations of approximately SiO_2 <55%.

Magma Molten rock that consists of up three components: liquid silicate melt, suspended crystalline solids, and gas bubbles. It is the raw material of all volcanic processes. Silicate magmas are the most common magma type and consist of long, polymeric chains and rings of Si–O tetrahedra, between which are located cations (e. g. Ca^{2+}, Mg^{2+}, Fe^{2+}, and Na^{+}). Anions (e. g. OH^{-}, F^{-}, Cl^{-}, and S^{-}) can substitute for the oxygen in the silicate framework. The greater the silica (SiO_2) content of the magma, the more chains and rings of silicate tetrahedra there are to impede each other and hence the viscosity of the magma increases. The pressure regime and composition of the magma control the minerals that nucleate and crystallize from a magma when it cools or degasses.

Magma chamber A subsurface volume within which magma accumulates, differentiates and crystallizes. Igneous intrusions can constrain the form and size of some magma chambers, but in general the shape and volume of magma chambers beneath active volcanoes are poorly known. Magma reservoir is an equivalent term.

Melt Liquid part of magma. Melts (usually silicate) contain variable amounts of dissolved volatiles. The primary volatiles are usually water and carbon dioxide.

Newtonian liquid A liquid for which the strain rate is proportional to the applied stress. The proportionality coefficient is called the viscosity.

Microlite Crystal with dimensions less than 100 μm. Usually microlites crystallize at shallow levels of magmatic system.

Phenocryst Crystal with dimensions larger than 100 μm. Usually phenocrysts grow in magmatic reservoirs prior to an eruption and or are entrained by magma in the chamber.

Pyroclastic flow or surge A gas-particle flow of pyroclasts suspended in a mixture of hot air, magmatic gas, and fine ash. The flow originates by the gravitational collapse of a dense, turbulent explosive eruption column at the source vent, or by dome collapse, and moves down-slope as a coherent flow. Pyroclastic flows and surges are distinguished by particle concentration in the flow, surges being more dilute. Variations in particle concentration result in differences in the deposits left by flows and surges.

Silicic Magma, lava, or tephra with silica concentrations of approximately $SiO_2 > 55\%$. The magmas are commonly rich in Al, Na– and K– bearing minerals. Silicic magmas are typically very viscous and can have high volatile contents. Rhyolite is an example of a silicic magma.

Volatile A component in a magmatic melt which can be partitioned in the gas phase in significant amounts during some stage of magma history. The most common volatile in magmas is water vapor H_2O, but there are commonly also significant quantities of CO_2, SO_2 and halogens.

Definition of the Subject

We consider the process of slow extrusion of very viscous magma that forms lava domes. Dome-building eruptions are commonly associated with hazardous phenomena, including pyroclastic flows generated by dome collapses, explosive eruptions and volcanic blasts. These eruptions commonly display fairly regular alternations between periods of high and low or no activity with time scales from hours to years. Usually hazardous phenomena are associated with periods of high magma discharge rate, thus, understanding the causes of pulsatory activity during extrusive eruptions is an important step towards forecasting volcanic behavior, especially the transition to explosive activity when magma discharge rate increases by a few orders

of magnitude. In recent years the risks have increased because the population density in the vicinity of many active volcanoes has increased.

Introduction

Many volcanic eruptions involve the formation of lava domes, which are extrusions of very viscous, degassed magmas. The magma is so viscous that it accumulates close to the vent. Extrusion of lava domes is a slow and long-lived process, and can continue for many years or even decades [71,83,85]. Typical horizontal dimensions of lava domes are several hundred meters, heights are of an order of tens to several hundred meters, and volumes several million to hundreds of million cubic meters. Typical magma discharge rates (measured as the increase of dome volume with time in dense rock equivalent (DRE)) can reach up to 20–40 m^3/s, but are usually below 10 m^3/s [83].

Dome-building eruptions are commonly associated with hazardous phenomena, including pyroclastic flows and tsunamis generated by dome collapses, explosive eruptions and volcanic blasts. Dome-building eruptions can also contribute to edifice instability and sector collapse, as occurred on Montserrat on 26 December 1997 [87]. Lava dome activity can sometimes precede or follow major explosive eruptions; the eruption of Pinatubo, Philippines (1991) is an example of the former [37], and the eruption of Mount St. Helens, USA (1980–1986) is an example of the latter [89].

Several lava dome eruptions have been documented in detail and show quite complex behaviors. Substantial fluctuations in magma discharge rate have been documented. In some cases these fluctuations can be quite regular (nearly periodic), as in the extrusion of lava in 1980–1982 on Mount St. Helens [89] and in the 1922–2002 activity of the Santiaguito lava dome, Guatemala [35]. In these cases, periods of high magma discharge rate alternate with longer periods of low magma discharge rate or no extrusion. In some volcanoes, such as Shiveluch, Kamchatka, the intervals of no extrusion are so long compared with the periods of dome growth that the episodes of dome growth have been described as separate eruptions of the volcano rather than episodes of the same eruption. Other dome-building activity can be nearly continuous and relatively steady, as observed at Mount St. Helens in 1983 [89] and at the Soufrière Hills Volcano, Montserrat between November 1999 and July 2003. In yet other cases the behavior can be more complex with quite sudden changes in magma discharge rate, which cannot be related to any well-defined regularity or pattern (e. g. Lascar volcano, Chile, [57]).

Pauses during lava dome-building eruptions are quite common. For example, at Mount St. Helens there were 9 pulses of dome growth with a period of ~74 days, a duration of 1–7 days and no growth in between [89]. The Soufrière Hills Volcano Montserrat experienced a long (20 months) pause in extrusion after the first episode of growth [72]. On Shiveluch volcano in Kamchatka episodes of dome growth occurred in 1980, 1993 and 2000, following a major explosion in 1964 [28]. Each episode of dome growth began with magma discharge rate increasing over the first few weeks to a peak of 8–15 m^3/s, with a gradual decline in magma discharge rate over the following year. In between the episodes very minimal activity was recorded.

Fluctuations in magma discharge rate have been documented on a variety of time-scales from both qualitative and quantitative observations. Several lava dome eruptions are characterized by extrusion of multiple lobes and flow units [68,94]. In the case of the Soufrière Hills Volcano, extrusion of shear lobes can be related to spurts in discharge rate and is associated with other geophysical changes, such as onset of seismic swarms and marked changes in temporal patterns of ground tilt [90,91,94]. These spurts in discharge rate have been fairly regular for substantial periods, occurring every 6 to 7 weeks over a 7 month period in 1997 [21,87,91]. These spurts are commonly associated with large dome collapses and pyroclastic flows and, in some cases, with the onset of periods of repetitive Vulcanian explosions [14,26]. Consequently the recognition of this pattern has become significant for forecasting activity for hazard assessment purposes. In the Soufrière Hills Volcano and Mount Pinatubo much shorter fluctuations in magma discharge rate have been recognized from cyclic variations in seismicity, ground tilt, gas fluxes and rock-fall activity [23,91,93]. This cyclic activity has typical periods in the range of 4 to 36 hours. Cyclic activity has been attributed to cycles of gas pressurization and depressurization with surges in dome growth related to degassing, rheological stiffening and stick-slip behavior [23,49,61,91,98].

Dome eruptions can show transitions to explosive activity, which sometimes can be linked to spurts in magma discharge rate. For example, in 1980, periodic episodes of lava dome extrusion on Mount St. Helens were initiated by explosive eruptions, which partly destroyed the dome that had been extruded in each previous extrusion episode [89]. At Unzen Volcano, Japan a single Vulcanian explosive eruption occurred in June 1991 when the magma discharge rate was at its highest [68]. At the Soufrière Hills Volcano, repetitive series of Vulcanian explosions have occurred following large dome collapses in

periods when magma discharge rates were the highest of the eruption [26,88]. In the case of Lascar Volcano, Chile, an intense Plinian explosive eruption occurred on 18 and 19 April, 1993, after nine years of dome extrusion and occasional short-lived Vulcanian explosions [57].

Lava dome eruptions require magma with special physical properties. In order to produce a lava dome rather than a lava flow, the viscosity of the magma must be extremely high so that the lava cannot flow easily from the vent. High viscosity is a consequence of factors such as relatively low temperature (typically 750–900°C), melt compositions rich in network-forming components (principally Si and Al) efficient gas loss during magma decompression, and crystallization as a response to cooling and degassing. Viscosities of silica-rich magmas, such as rhyolites and some andesites, are increased by several orders of magnitude by the loss of dissolved water during decompression. Many, but not all, domes also have high crystal content (up to 60 to 95 vol%), with crystallization being triggered mostly by degassing [10,86]. In order to avoid fragmentation that leads to an explosive eruption, magma must have lost gas during ascent. Consider, for example, a magma at 150 MPa containing 5wt% of dissolved water decompressed to atmospheric pressure. Without gas loss the volume fraction of bubbles, will be more than 99%. Typical dome rock contains less than 20 vol% of bubbles, although there is evidence that magma at depth can be more bubble-rich (e. g., [13,74]). On the other hand, very commonly there is no change in temperature or bulk magma composition in the products of explosive and extrusive eruptions for a particular volcano. This suggests that the properties of magma that are conducive to the formation of lava domes are controlled by physico-chemical transformations that occurred during magma ascent to the surface.

Two other important factors that influence whether lava domes or flows form are topography and discharge rate. The same magma can form a dome if the discharge rate is low, and a lava flow if the rate is high [29,92]. The discharge rate is controlled by overall conduit resistance that is a function of viscosity, conduit size and shape, and driving pressure (the difference between chamber pressure and atmospheric pressure). Additionally the same magma can form a dome on low slopes, such as a flat crater (e. g. the mafic andesite dome of the Soufrière Volcano, St. Vincent; [40]) and a lava flow on steep slopes.

Prior to an eruption, magma is usually stored in a shallow crustal reservoir called a magma chamber. For several volcanoes magma chambers can be detected and characterized by earthquake locations, seismic tomography, petrology or interpretation of ground deformation data [55]. Typical depths of magma chambers range from a few kilometers to tens of kilometers. Volumes range from less than one to several thousand km^3 [55], but are usually less than a hundred km^3. Magma chambers are connected to the surface by magma pathways called conduits. There is evidence that the conduits that feed lava dome eruptions can be both dykes or cylindrical. Dykes of a few meters width are commonly observed in the interior of eroded andesite volcanoes. Dyke feeders to lava domes have been intersected by drilling at Inyo crater, California, USA [56] and at Mount Unzen [67]. Geophysical studies point to dyke feeders; for example fault-plane solutions of shallow volcano-tectonic earthquakes indicate pressure fluctuations in dykes [75,76]. Deformation data at Unzen, combined with structural analysis, indicate that the 1991–1995 dome was fed by a dyke [68]. Dykes are also the only viable mechanism of developing a pathway through brittle crust from a deep magma chamber to the surface in the initial stages of an eruption [50,77].

Cylindrical conduits commonly develop during lava dome eruptions. The early stages of lava dome eruptions frequently involve phreatic and phreatomagmatic explosions that create near surface craters and cylindrical conduits [12,73,87,89,96,99]. These explosions are usually attributed to interaction of magma rising along a dyke with ground water. Cylindrical conduits formed by explosions are confined to relatively shallow parts of the crust, probably of order hundreds of meters depth and < 1 km, as indicated by mineralogical studies [73]. Examples of such initial conduit forming activity include Mount Usu (Japan) Mt. St. Helens, and Soufrière Hills Volcano [12,87,99]. Many lava dome eruptions are also characterized by Vulcanian, sub-Plinian and even Plinian explosive eruptions. Examples include Mount Unzen, Mount St.Helens, Santiaguito and Soufrière Hills Volcano [68,87,89,96]. Here the fragmentation front may reach to depths of several kilometers [54] with the possibility of cylindrical conduit development due to severe underpressurization and mechanical disruption of conduit wallrocks. Subsequently domes can be preferentially fed along the cylindrical conduits created by earlier explosive activity. On the Soufrière Hills Volcano, early dome growth was characterized by extrusion of spines with nearly cylindrical shape [87].

Observations of magma discharge rate variations on a variety of time-scales highlight the need to understand the underlying dynamic controls. Research has increasingly focused on modeling studies of conduit flow dynamics during lava-dome eruptions. We will restrict our discussions here to mechanisms that lead to cyclic and quasi-periodic fluctuations in magma discharge rate on various timescales, mainly focusing on long-term cycles. Issues

concerning the transition between explosive and extrusive activity are discussed in detail in [45,80,81,97] and in the special volume of Journal of Volcanology and Geothermal Research dedicated to modeling of explosive eruptions [79]. Several papers consider the processes that occur at the surface and relate dome morphology and dimensions with controlling parameters.

Combined theoretical, experimental and geological studies identify four main types of dome: spiny, lobate, platy, and axisymmetric [5,29,30]. These types of dome reflect different regimes which are controlled by discharge rates, cooling rates and yield strength, and the viscosity of the dome-forming material. In recent years mathematical modeling has been used to semi-quantitatively describe spreading of lava domes, including models based on the thin layer approximation [1,2] and fully 2D simulations of lava dome growth which account for visco-elastic and -plastic rheologies [32,33].

Dynamics of Magma Ascent During Extrusive Eruptions

In order to understand the causes of cyclic behavior during extrusive eruptions, first we need to consider the underling dynamics of volcanic systems, and discuss physical and chemical transitions during magma ascent.

The physical framework for the model of a volcanic system is shown in Fig. 1. Magma is stored in a chamber at depth L, with a chamber pressure P_{ch} that is higher than hydrostatic pressure of magma column and drives magma ascent. Magma contains silicate melt, crystals and dissolved and possibly exsolved volatiles. During ascent of the magma up the conduit the pressure decreases and volatiles exsolve forming bubbles. As the bubble concentration becomes substantial, bubble coalescence take place and permeability develops [27,45], allowing gas to escape from ascending magma, both in vertical (through the magma) and horizontal (to conduit wallrocks) directions. If magma ascends slowly, gas escape results in a significant reduction of the volume fraction of bubbles; such a process is termed open system degassing. In comparison closed system degassing is characterized by a negligible gas escape. Reduction in bubble content, combined with relatively low gas pressures and efficient decoupling of the gas and melt phases, prevents magma fragmentation and, thus, development of explosive eruption [59,60,82].

Due to typical low ascent velocities (from millimeters to a few centimeters per second) magma ascent times to the Earth's surface from the magma chamber range from a several hours to many weeks. These ascent times are often comparable with the times that are required for crystals to grow significantly and for heat exchange between magma and wallrocks. The main driving force for crystallization is related to exsolution of volatiles from the magma, leading to increase in the liquidus temperature T_L, and development of magma undercooling $\Delta T = T_L - T$ [10]. Crystallization leads to release of latent heat, and magma temperature can increase with respect to the initial temperature [6]. As a consequence of increasing crystal content, magma viscosity increases by several orders of magnitude [16,20,21] and magma becomes a non-Newtonian fluid [78]. As will be shown later, crystallization induced by degassing can become a key process in causing variable flow rates.

Due to the long duration of extrusive eruptions, the magma chamber can be replenished with significant amounts of magma from underlying sources [41,65]. Replenishment can lead to pressure build-up in the magma chamber, volatile and heat exchange between host and new magmas. The composition of the magma can also change over time, due to differentiation, crustal rock assimilation, or magma mixing. Thus, any model that explains magma ascent dynamics needs to deal with many complexities. Of course there is no single model that can take into account all physical processes in a volcanic system. Additional complications arise from the fact that the physical properties of magma at high crystal contents, such as rheology or crystal growth kinetics and geometry of volcanic systems, are typically poorly constrained. Several issues regarding the dynamics of multiphase systems have not been resolved theoretically, especially for cases where the volume fractions of components of the multiphase system are comparable.

Below we will present a review of existing models that treat cyclic behavior during extrusive eruptions on different timescales.

Short-Term Cycles

Cyclic patterns of seismicity, ground deformation and volcanic activity (Fig. 2 from [23]) have been documented at Mount Pinatubo, Philippines, in 1991 [37] and Soufrière Hills volcano, Montserrat, British West Indies, in 1996–1997 [90,91]. At Soufrière Hills, periodicity in seismicity and tilt ranged from ~4 to 30 h, and the oscillations in both records continued for weeks. Cyclic behavior was first observed in the seismicity (RSAM) records beginning in July 1996, when the record of dome growth constrained the average supply rate to between 2 and 3 m^3/s [88]. The oscillations in the RSAM records initially had low amplitudes, and no tilt-measurement station was close enough to the vent to detect any pressure oscillations in the con-

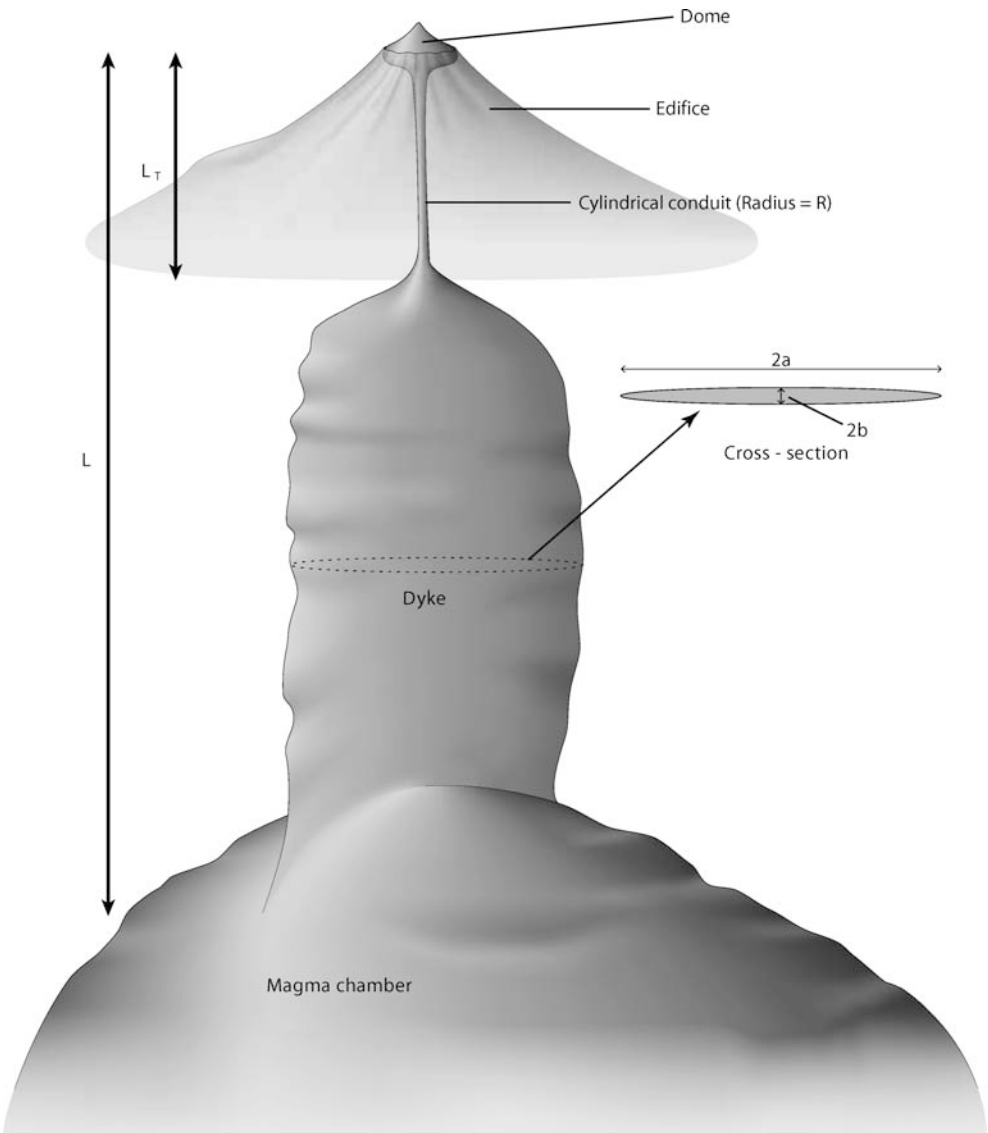

Volcanic Eruptions: Cyclicity During Lava Dome Growth, Figure 1

Schematic view of the volcanic system. In the upper part the conduit is cylindrical with a radius R. A transition from the cylinder to a dyke occurs at depth L_T. The length scale for the transition from cylinder to dyke is w_T. The dyke has an elliptical cross-section with semi-axis lengths a_0 and b_0. The chamber is located a depth L. In the text we used also the following auxiliary variables: $D = 2R$ for conduit diameter, $L_d = L - L_T$ for the dyke vertical length, $W_d = 2a_0$ for the dyke width, and $H_d = 2b_0$ for the dyke thickness. After [20]

duit. By August 1996, RSAM records showed strong oscillatory seismicity at dome-growth rates between 3 and 4 m^3/s. Tilt data, taken close enough to the vent (i. e., Chances Peak [90]) to be sensitive to conduit pressure oscillations, are only available for February 1997 and May–August 1997. In the latter period dome growth rate increase from $\sim$5 m^3/s in May to between 6 and 10 m^3/s in August. Both near-vent tilt and RSAM displayed oscillatory behavior during this period and were strongly correlated in time. Similar RSAM oscillations having periods of 7 to 10 h were observed at Mount Pinatubo following the climactic eruption in 1991 [37]. At both volcanoes, oscillation periods were observed that do not fit any multiple of Earth or ocean tides.

The cyclic activity at both Pinatubo and Soufrière Hills Volcano are strongly correlated with eruptive behavior and other geophysical phenomena. In the Pinatubo case and on Soufrière Hills Volcano in August, September and

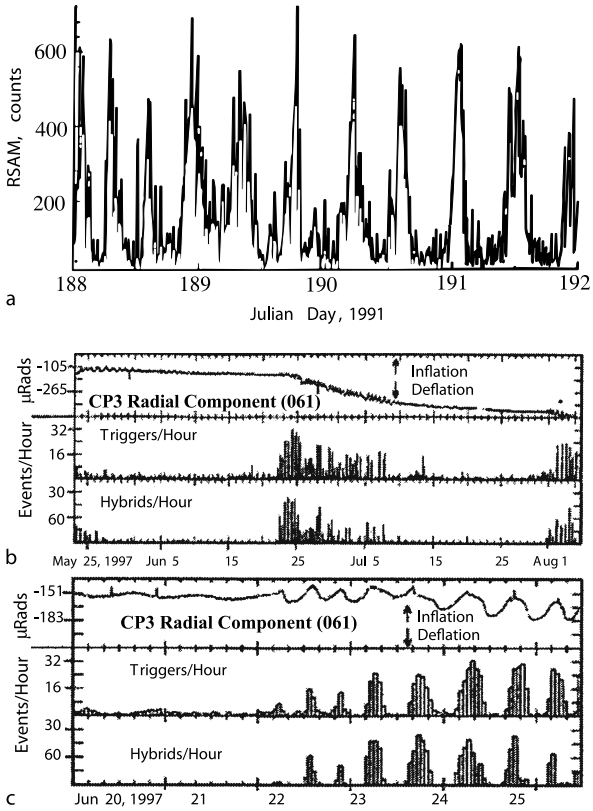

Volcanic Eruptions: Cyclicity During Lava Dome Growth, Figure 2
Examples of cyclic behavior in a lava dome eruption. RSAM records from Mt. Pinatubo, Philippines following its climatic eruption in June 1991 (a, after [23]). Radial tilt, triggered earthquakes, and hybrid earthquakes for 17 May to 6 August 1997, at Soufrière Hills Volcano (b, after [21,90]). Parts of three 5–7 week cycles are shown, with each cycle showing high amplitude tilt and seismicity pulsations with short time scale that last for several weeks after the start of the cycle. The onset of a cycle is rapid, as detailed in c for the cycle initiating on 22 June 1997

October 1997, the cycles were linked to short-lived volcanic explosions. In the case of the Soufrière Hills Volcano, explosions in August 1997 occurred at the peak in the tilt. The peak in tilt also marked the onset of episodes of increased rock falls [7,8] and [91] is attributed to increased magma discharge rates. SO_2 flux data show that the cycles are linked to surges in gas release, which reach a peak about an hour after the tilt peak [93]. Green et al. [31] have shown that several families of near identical long period earthquakes occur during the tilt cycle, starting at the inflexion point on the up-cycle and finishing before the inflexion point on the down-cycle.

Several models [23,49,70,98,98] have been proposed to explain the observed cyclicity. In these models [23,98] the conduit is divided into two parts. Magma is assumed to be forced into the lower part of volcanic conduit at a constant rate. In Denlinger and Hoblitt [23] magma in the lower part of the conduit is assumed to be compressible. In Wylie et al. [98] the magma is incompressible but the cylindrical conduit is allowed to expand elastically. In both models the lower part of the conduit, therefore, acts like a capacitor that allows magma to be stored temporally in order to release it during the intense phase of the eruption. In the upper part of the conduit friction is dependent on magma discharge rate, with a decrease in friction resulting in an increase in discharge rate, over a certain range of discharge rates. In Denlinger and Hoblitt [23], when magma discharge rate reaches a critical value, magma detaches from the conduit walls and a stick-slip transition occurs. Rapid motion of magma leads to depressurization of the conduit and a consequent decrease in discharge rate, until at another critical value the magma again sticks to the walls. Pressure starts to increase again due to influx of new magma into the conduit. On a pressure-discharge diagram the path of eruption is represented by a hysteresis loop. In Wylie et al. [98] the friction is controlled by volatile-dependent viscosity. Volatile exsolution delay is controlled by diffusion. When magma ascends rapidly, volatiles have no time to exsolve and viscosity remains low. Depressurization of the upper part of the conduit leads to a decrease in magma discharge rate and an increase in viscosity due to more intense volatile exsolution.

In Neuberg et al. [70] a steady 2D conduit flow model was developed. The full set of Navier–Stokes equations for a compressible fluid with variable viscosity was solved by means of a finite element code. Below some critical depth the flow was considered to be viscous and Newtonian, with a no-slip boundary condition at the wall. Above this depth a plug develops, with a wall boundary condition of frictional slip. The slip criterion was based on the assumption that the shear stress inside the magma overcomes some critical value. Simulations reveal that slip occurs in the shallow part of the conduit, in good agreement with locations of long-period volcanic earthquakes for the Soufrière Hills Volcano, [31]. However, some parameters used in the simulations (low crystallinity, less than 30% and high discharge rate, more than 100 m^3/s) are inconsistent with observations.

Lensky et al. [49] developed the stick-slip model by incorporating degassing from supersaturated magma together with a sticking plug. Gas diffuses into the magma, which cannot expand due to the presence of a sticking plug, resulting in a build up of pressure. Eventually the pressure exceeds the strength of the plug, which fails in stick-slip motion and the pressure is relieved. The magma sticks again when the pressure falls below the dynamic

friction value. In this model the time scale of the cycles is controlled by gas diffusion. The influence of permeable gas loss, crystallization and elastic expansion of the conduit on the period of pulsations was studied.

A shorter timescale of order of minutes was investigated in Iverson et al. [43] in relation to repetitive seismic events during the 2004–2006 eruption of Mount St. Helens. The flow dynamics is controlled by the presence of a solid plug that is pushed by a Newtonian liquid, with the possibility of a stick-slip transition. Inertia of the plug becomes important on such short timescales.

Models to explain the occurrence of Vulcanian explosions have also been developed by Connor et al. [15], Jaquet et al. [44] and Clarke et al. [13]. A statistical model of repose periods between explosions by Connor et al. [15] shows that data fit a log-logistic distribution, consistent with the interaction of two competing processes that decrease and increase gas pressure respectively. Jaquet et al. [44] show the explosion repose period data have a memory. The petrological observations of Clarke et al.. [13] on clasts from Vulcanian explosions associated with short-term cycles support a model where pressure builds up beneath a plug by gas diffusion, but is opposed by gas leakage through a permeable magma foam. Although these models include some of the key processes and have promising explanatory power, they do not consider the development of the magma plug explicitly. This process is considered in [24], where a model of magma ascent with gas escape is proposed.

Long-Term Cycles

Figure 3 shows views of three lava domes (Mount St. Helens, USA, Santiaguito, Guatemala and Shiveluch, Russia). Measurements of magma discharge rate variations with time are presented below [25,35,89]. Behavior at the first two volcanoes is rather regular, whereas at Shiveluch, long repose periods are followed by an initial rapid increase in eruptive activity with subsequent decrease and complete stop of the eruption. Growth of the lava dome at Unzen volcano, Japan, 1991–1995 was similar to Shiveluch [25,68].

There are three types of conceptual models that attempt to explain long term variation in magma discharge rate. Maeda [53], after [42], considers a simple system that contains a spherical magma chamber located in elastic rocks with a cylindrical conduit located in visco-elastic rocks. Magma viscosity is assumed to be constant. The magma chamber is replenished with a time dependent influx rate. The model reproduces discharge rate variation at Unzen volcano by assuming a bell-shaped form of influx rate dependence on time. There are two controversial assumptions in the model. First, the assumption that the conduit wallrocks are visco-elastic, while the magma chamber wallrocks are purely elastic cannot be justified because near the chamber rock temperature is the highest and it is more reasonable to expect viscous properties for chamber wallrocks rather than for the conduit. The equation that links conduit diameter with

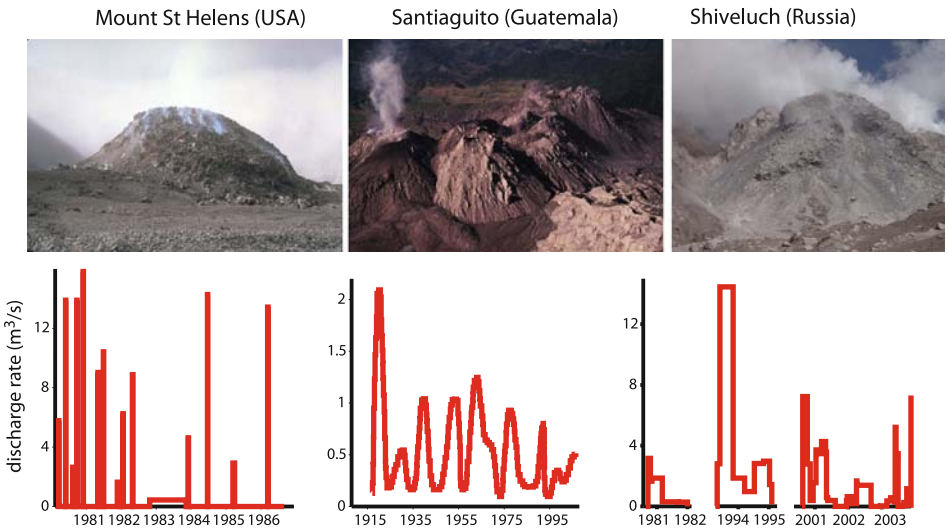

Volcanic Eruptions: Cyclicity During Lava Dome Growth, Figure 3
Observed discharge rate versus time for (a) Mount St.Helens dome growth and (b) Santiaguito volcano and Shiveluch (c). Photos: Mount St.Helens by Lyn Topinka (1985), Santiaguito by Gregg Bluth (2002), Shiveluch by Pavel Plechov (2001)

magmatic overpressure assumes viscous rock properties up to infinity. If the chamber is located in visco-elastic rocks, oscillations in discharge rate are not possible. Second, in order to obtain reasonable timescales, the rock viscosity must be rather small, of order of 10^{13} Pa s, which is only slightly higher than the typical viscosities of the magma.

Another set of models attribute cyclic behavior to heat exchange between ascending magma and wallrocks, which accounts for temperature dependent viscosity [17,95]. The idea of both models is that magma cools down as it ascends, and heat flux is proportional to the difference between the average temperature of the magma and the temperature of the wallrocks. If magma ascends quickly than heat loss is small in comparison with heat advection. Magma viscosity remains low as a consequence and allows high magma discharge rates. In contrast, when magma ascends slowly it can cool substantially and viscosity increases significantly. Both models suggest that, for a fixed chamber pressure, there can be up to three steady state solutions with markedly different discharge rates. Transition between these steady-state solutions leads to cyclic variations in discharge rate. Whitehead and Helfrich [95] demonstrated the existence of cyclic regimes in experiments using corn syrup. In application to magma ascent in a volcanic conduit, these models have strong limitations, because a constant wall-rock temperature is assumed. However, as an eruption progresses the wallrocks heat up and heat flux decreases, a condition that makes periodic behavior impossible for long-lived eruptions. For such a long-lived eruption like Santiaguito (started in 1922) wallrocks are expected to be nearly equilibrated in temperature with the magma, and heat losses from magma became small. It is possible that this decrease in heat flux contributes to a slow progressive increase in temperature that is observed on timescales longer than the period of pulsations. For example, magma at Santiaguito becomes progressively less viscous, resulting in a transition from mainly lava dome to lava flow activity.

Models, developed by authors of this manuscript, consider that degassing-induced crystallization is a major controlling process for the long-term cyclicity during lava dome building eruptions. There is increasing evidence that there is a good correlation between magma discharge rate and crystallinity of the magma [10,68]. An increase in crystal content leads to an increase in magma viscosity [16,20,21] and, thus, influences magma ascent dynamics. First, we consider a simplified model of magma ascent in a volcanic conduit that accounts for crystallization and rheological stiffening.

A Simplified Model

In Barmin et al. [3] the following simplifying assumptions have been made in order to develop a semi-analytical approach to magma ascent dynamics.

1. Magma is incompressible. The density change due to bubble formation and melt crystallization is neglected.
2. Magma is a viscous Newtonian fluid. Viscosity is a step function of crystal content. When the concentration of crystals β reaches a critical value β_* the viscosity of magma increases from value μ_1 to a higher value μ_2. Later on we will consider magma rheology in more detail (see Sect."Rheology of Crystal-Bearing Magma and Conduit Resistance"), but a sharp increase in viscosity over a narrow range of crystal content has been confirmed experimentally (e. g., [9,48]).
3. Crystal growth rate is constant and no nucleation occurs in the conduit. The model neglects the fact that magma is a complicated multi-component system and its crystallization is controlled by the degree of undercooling (defined as the difference between actual temperature of the magma and its liquidus temperature). Later in the paper a more elaborate model for magma crystallization will be considered.
4. The conduit is a vertical cylindrical pipe. Elastic deformation of the wallrocks is not included in the model. This assumption is valid for a cylindrical shape of the conduit at typical magmatic overpressures, but is violated when the conduit has a fracture shape. Real geometries of volcanic conduits and their inclination can vary significantly with depth.
5. The magma chamber is located in elastic rocks and is fed from below, with a constant influx rate. For some volcanoes, like Santiaguito or Mount St.Helens, average magma discharge rate remained approximately constant during several periods of pulsation. Thus the assumption of constant influx rate is valid. For volcanoes like Mount Unzen or Shiveluch, there is an evidence of pulse-like magma recharge [25,53].

With above simplification the system of equations for unsteady 1D flow is as follows:

$$\frac{\partial}{\partial t}\rho + \frac{\partial}{\partial x}\rho u = 0\,; \quad \frac{\partial}{\partial t}n + \frac{\partial}{\partial x}nu = 0 \tag{1a}$$

$$\frac{\partial p}{\partial x} = -\rho g - \frac{32\mu u}{\delta^2}\,; \quad \mu = \begin{cases} \mu_1, & \beta < \beta_* \\ \mu_2, & \beta \geqslant \beta_* \end{cases} \tag{1b}$$

$$\frac{\partial}{\partial t}\beta + u\frac{\partial \beta}{\partial x} = 4\pi n r^2 \chi = (36\pi n)^{\frac{1}{3}}\,\beta^{\frac{2}{3}}\,\chi \tag{1c}$$

Here ρ is the density of magma, u is the vertical cross-section averaged ascent velocity, n is the number density of

crystals per unit volume, p is the pressure, g is the acceleration due to gravity, δ is the conduit diameter, β is the volume concentration of crystals, β_* is a critical concentration of crystals above which the viscosity changes from μ_1 to μ_2, r is the crystal radii, χ is the linear crystal growth rate, and x is the vertical coordinate. The first two Eqs. (1a) represent the conservation of mass and the number density of crystals, the second (1b) is the momentum equation with negligible inertia, and the third (1c) is the crystal growth equations with χ = constant. We assume the following boundary conditions for the system (1):

$$x = 0 : \frac{\mathrm{d}p_{\mathrm{ch}}}{\mathrm{d}t} = \frac{\gamma}{V_{\mathrm{ch}}}\left(Q_{\mathrm{in}} - Q_{\mathrm{out}}\right); \ \beta = \beta_{\mathrm{ch}}; \ \ n = n_{\mathrm{ch}}$$
$$x = l : p = 0$$

Here γ is the rigidity of the wall-rock of the magma chamber, V_{ch} is the chamber volume, β_{ch} and n_{ch} are the crystal concentration and number density of crystals per unit volume in the chamber, p_{ch} is the pressure in the chamber, l is the conduit length, Q_{in} is the flux into the chamber and $Q_{\mathrm{out}} = \pi\delta^2 u/4$ is the flux out of the chamber into the conduit. Both β_{ch} and n_{ch} are assumed constant. We neglect the influence of variations of the height of the lava dome on the pressure at the top of the conduit and assume that the pressure there is constant. As the magma is assumed to be incompressible, the pressure at the top of the conduit can be set to zero because the atmospheric pressure is much smaller than magma chamber pressure.

From the mass conservation equation for the case of constant magma density, $u = u(t)$ and $n = n_{\mathrm{ch}}$ everywhere. Equations (1) can be integrated and transformed from partial differential equations to a set of ordinary differential equations with state-dependent delay representing a "memory" effect on crystal concentration (see [3] for details).

Results and Applications

The general steady-state solution for magma ascent velocity variations with chamber pressure is shown in Fig. 4a. Solutions at high magma ascent velocities, when the critical concentration of crystals is not reached inside the conduit, result in a straight line (CB), which is the same as for the classical Poiseuille solution for a fluid with constant viscosity. At low ascent velocities there is a quadratic relationship (OAC) between chamber pressure and ascent velocity (see [3] for derivation of the equation). A key feature of the steady-state solution is that, for a fixed chamber pressure, it is possible to have three different magma ascent velocities. We note that for $\mu_2/\mu_1 = 1$ only the branch CB exists and for $1 < \mu_2/\mu_1 \leqslant 2$ there is a smooth transition between the lower branch (ODA) and the upper branch (CB) and multiple steady-state regimes do not exist.

We first consider the case where chamber pressure changes quasi-statically and the value of Q_{in} is between Q_{A} and Q_{C}. Starting at point O the chamber pressure increases, because the influx into the chamber is higher than the outflux. At point A, a further increase in pressure is not possible along the same branch of the steady-state solution and the system must change to point B, where the outflux of magma is larger than the influx. The chamber pressure and ascent velocity decrease along BC until the point C is

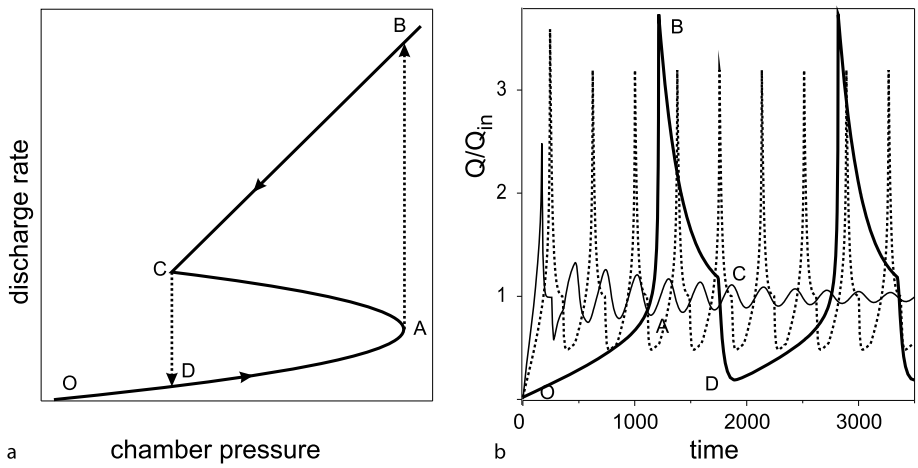

Volcanic Eruptions: Cyclicity During Lava Dome Growth, Figure 4
The general steady-state solution and possible quasi-static evolution of an eruption. After [3]. Different curves correspond to different values of dimensionless parameter $\kappa = (\pi\delta^2\gamma)/(4V_{\mathrm{ch}}\rho g)$ (0.12 – *thin line*, 0.05 – *dotted line*, and 0.005 – *bold line*)

reached and the system must change to point D. The cycle DABC then repeats itself. Provided the chamber continues to be supplied at the same constant rate repetition of this cycle results in periodic behavior. The transitions AB and CD in a cycle must involve unsteady flow.

Oscillations in magma discharge rate involve large variations in magma crystal content. This relation is observed on many volcanoes. For example, pumice and samples of the Soufrière Hills dome that were erupted during periods of high discharge [88] have high glass contents (25 to 35%) and few microlites [65], whereas samples derived from parts of the dome that were extruded more slowly (days to weeks typically) have much lower glass contents (5 to 15%) and high contents of groundmass microlites. These and other observations [34,68] suggest that microlite crystallization can take place on similar time scales to the ascent time of the magma.

Of more general interest is to consider unsteady flow behaviour. We assume that the initial distribution of parameters in the conduit corresponds to the steady-state solution of system (1) with the initial magma discharge rate, Q_0, being in the lowest regime. The behaviour of an eruption with time depends strongly on the value of Q_{in}. If Q_{in} corresponds to the upper or the lower branch of the steady state solution the eruption stabilizes with time with $Q = Q_{in}$ and $dp_{ch}/dt = 0$. However, if Q_{in} corresponds to the intermediate branch of the steady-state solution, with Q_{in} between Q_A and Q_C, periodic behaviour is possible. Figure 4b shows three eruption scenarios for different values of the magma chamber volume V_{ch}. When V_{ch} is small the eruption stabilizes with time. In contrast, undamped periodic oscillations occur for values of V_{ch} larger than some critical value. For very large magma chamber volumes the transient solution almost exactly follows the steady-state solution, with unsteady transitions between the regimes. The time that the system spends in unsteady transitions in this case is much shorter than the period of pulsations. The maximum discharge rate during the cycle is close to Q_B, the minimum is close to Q_D and the average is equal to Q_{in}. The period of pulsations increases as the volume of magma chamber increases.

Now we apply the model to two well-documented eruptions: the growth of lava domes on Mount St. Helens (1980–1986) and on Santiaguito (1922-present). Our objective here is to establish that the model can reproduce the periodic behaviors observed at these two volcanoes. Estimates can be obtained for most of the system parameters. However, magma chamber size is not well-constrained and so the model can be used to make qualitative inferences on relative chamber size. Given the uncertainties in the parameter values and the simplifications in the model development, the approach can be characterized as mimicry. Adjustments in some parameters were made to achieve best fits with observations, but the particular best-fits are not unique.

For Mount St. Helens our model is based on data presented by [89]. Three periods of activity can be distinguished during the period of dome growth. The first period consists of 9 pulses of activity with average peak magma discharge rates $\sim$15 m^3/s during 1981–1982. Each pulse lasted from 2 to 7 days (with a mean value of 4 days) with the average period between the pulses being 74 days and the average discharge rate during the period being $Q_1 \sim 0.67$ m^3/s. The second period is represented by continuous dome growth and lasted more than a year (368 days) with a mean magma discharge rate of $Q_2 \sim 0.48$ m^3/s. During the last period there were 5 episodes of dome growth with peak magma discharge rates up to ~ 15 m^3/s, an average period of pulsation of ~ 230 days and a mean discharge rate of $Q_3 \sim 0.23$ m^3/s. We assume that the intensity of influx into the magma chamber, Q_{in}, is equal to the average magma discharge rate over the corresponding periods. There might have been a progressive decrease in the intensity of influx during the eruption, but, due to limitations of the model, we assume that Q_{in} changes as a step function between the periods.

The best-fit model for the eruption is presented in Fig. 5a, with the parameters used for the simulation summarized in Table 2 in [3]. During the first period of the eruption $Q_{in} = Q_1$. This corresponds to the intermediate branch of the steady-state solution and cyclic behaviour occurs. In the second period $Q_{in} = Q_2$ the system moves to the lower regime and the eruption stabilizes with time. For the third period of the eruption the parameters of the system have been changed, so that $Q_{in} = Q_3 < Q_2$, corresponding to the intermediate regime once again and periodic behaviour occurs. This condition can be satisfied by a decrease in the diameter of the conduit, or a decrease in crystal growth rate, or the number density of crystals. All these mechanisms are possible: decrease in the diameter could be a consequence of magma crystallization on the conduit walls, while a decrease in either crystal growth rate or number density of crystals could be reflect observed changes in magma composition [89]. The influence of conduit diameter is the strongest because ascent velocity, for the same discharge rate, depends on the square of the diameter. The required change in diameter is from 18 to 12 m, but this change can be smaller if we assume a simultaneous decrease in crystal growth rate.

Since 1922, lava extrusion at Santiaguito has been cyclic [35]. Each cycle begins with a 3–6 year long high (0.5–2.1 m^3/s) magma discharge rate phase, followed by

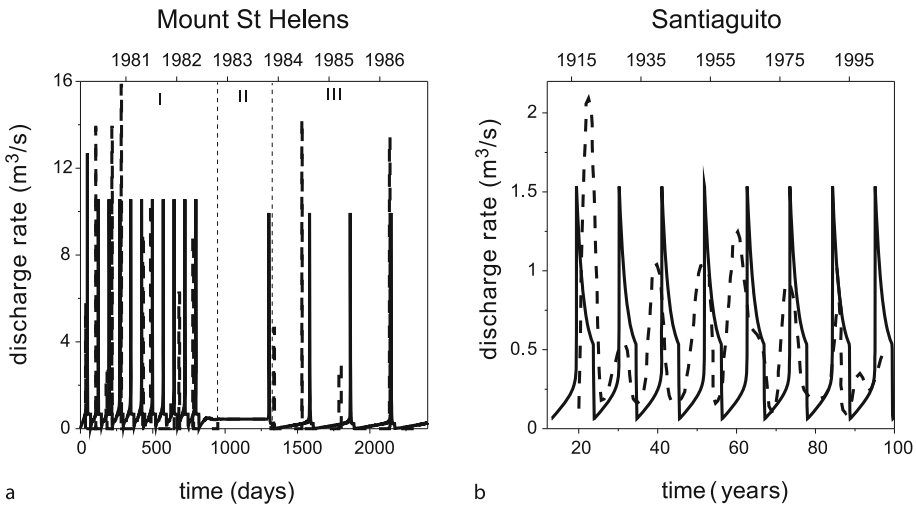

Volcanic Eruptions: Cyclicity During Lava Dome Growth, Figure 5
Discharge rate versus time for a Mount St.Helens dome growth and b Santiaguito volcano. *Doted lines* represent are the observed values of discharge rates and the *solid lines* are the best fit simulations. After [3]

a longer (3–11 years) low ($\sim 0.2\,m^3/s$) discharge rate phase. The time-averaged magma discharge rate was almost constant at $\sim 0.44\,m^3/s$ between 1922 and 2000. The first peak in discharge rate had a value $>2\,m^3/s$, whereas the second peak had a much smaller value. The value for the second peak is underestimated as it is calculated based on the dome volume only, but does not include the volume of dome collapse pyroclastic flows. Later peaks show an increase in magma discharge rate until 1960 (Fig. 5b, dashed line). Post-1960, the duration of the low discharge rate phase increased, the peak discharge and the time-averaged discharge rates for each cycle decreased, and the difference between discharge rates during the high and low discharge rate phases of each cycle decreased. Our best-fit model is shown in Fig. 5b and the parameters estimates are listed in Table 2 in [3]. The model reproduces the main features of the eruption, including the period of pulsations, the ratio between low and high magma discharge rates, and the range of observed discharge rates. We cannot, however, reproduce the decrease in the amplitude of pulsations within the framework of the model using fixed parameter values.

The theory provides a potential method to estimate magma chamber volumes. For Mount St.Helens our estimate of the chamber size ($\sim 0.6\,km^3$) is comparable with the total erupted volume in the entire 1980–1986 eruption and is consistent with the fact that geophysical imaging did not identify a large magma body. Santiaguito volcano erupted more than $10\,km^3$ in the 1902 explosive eruption [96] and more than $1\,km^3$ of lava domes since 1922. The best-fit model estimate of a large ($64\,km^3$) chamber is consistent with much larger eruption volumes, long periods, and longevity of the eruption in comparison to Mount St. Helens. One limitation of the model is that the supply of deep magma from depth to the chamber is assumed to be constant.

Model Development

In this section we further develop models to examine new effects and relax some of the simplifications of earlier models. We investigate a number of effects that were not fully explained or considered in previous studies [3,61,62]. The new model incorporates a more advanced treatment of crystallization kinetics based on the theoretical concepts developed in [38,46], and is calibrated by experimental studies in andesitic systems [22,34]. In particular, we distinguish growth of phenocrysts formed in the magma chamber from crystallization of microlites during magma ascent. Previous models have assumed that magma is always Newtonian, so we study models of conduit flow assuming non-Newtonian rheology, with rheological properties being related to crystal content. Latent heat is released during the crystallization of ascending magma due to degassing and we show that this can have an important influence on the dynamics. Elastic deformation of conduit walls leads to coupling between magma ascend and volcano deformations.

System of Equations We model magma ascent in a dyke-shaped conduit with elliptical cross-section using

a set of 1D transient equations written for horizontally averaged variables [20,21]:

$$\frac{1}{S}\frac{\partial}{\partial t}(S\rho_{\mathrm{m}}) + \frac{1}{S}\frac{\partial}{\partial x}(S\rho_{\mathrm{m}}V) = -G_{\mathrm{mc}} - G_{\mathrm{ph}} \tag{2}$$

$$\frac{1}{S}\frac{\partial}{\partial t}(S\rho_{\mathrm{mc}}) + \frac{1}{S}\frac{\partial}{\partial x}(S\rho_{\mathrm{mc}}V) = G_{\mathrm{mc}} \tag{3a}$$

$$\frac{1}{S}\frac{\partial}{\partial t}(S\rho_{\mathrm{ph}}) + \frac{1}{S}\frac{\partial}{\partial x}(S\rho_{\mathrm{ph}}V) = G_{\mathrm{ph}} \tag{3b}$$

$$\frac{1}{S}\frac{\partial}{\partial t}(S\rho_{\mathrm{d}}) + \frac{1}{S}\frac{\partial}{\partial x}(S\rho_{\mathrm{d}}V) = -J \tag{4a}$$

$$\frac{1}{S}\frac{\partial}{\partial t}(S\rho_{\mathrm{g}}) + \frac{1}{S}\frac{\partial}{\partial x}(S\rho_{\mathrm{g}}V_{\mathrm{g}}) = J \tag{4b}$$

Here t denotes time, x the vertical coordinate, ρ_{m}, ρ_{ph}, ρ_{mc}, ρ_{d} and ρ_{g} are the densities of melt, phenocrysts, microlites, dissolved gas and exsolved gas respectively, and V and V_g are the velocities of magma and gas, respectively. G_{ph}, G_{mc} represent the mass transfer rate due to crystallization of phenocrysts and microlites, respectively, and J the mass transfer rate due to gas exsolution, S is the cross-section area of the conduit. Equation (2) represents the mass conservation for the melt phase, Eqs. (3a) and (3b) are the conservation equations for microlites and phenocrysts respectively, Eqs. (4a) and (4b) represent the conservation of the dissolved gas and of the exsolved gas respectively.

$$\frac{\partial p}{\partial x} = -\rho g - F_c \tag{5}$$

$$V_{\mathrm{g}} - V = -\frac{k}{\mu_g}\frac{\partial p}{\partial x} \tag{6}$$

Here p is the pressure, ρ the bulk density of magma, g the acceleration due to gravity, μ is the magma viscosity, k is the magma permeability and μ_{g} is the gas viscosity. Eq. (5) represents the equation of momentum for the mixture as a whole, in which the pressure drops due to gravity and conduit resistance are calculated for laminar flow in an elliptic pipe. Equation (6) is the Darcy law for the exsolved gas flux through the magma.

$$\frac{1}{S}\frac{\partial}{\partial t}(S\rho C_{\mathrm{m}}T) + \frac{1}{S}\frac{\partial}{\partial x}(S\rho C_{\mathrm{m}}VT) = L_*(G_{\mathrm{mc}} + G_{\mathrm{ph}}) - C_{\mathrm{m}}TJ - Q_{\mathrm{cl}} + Q_{\mathrm{vh}} \tag{7}$$

Here C_m is the bulk specific heat of magma, T is the bulk flow-averaged temperature, L_* is latent heat of crystallization, Q_{cl} denotes the total heat loss by conduction to the conduit walls, and Q_{vh} denotes the total heat generation due to viscous dissipation. Here we consider the case of the latent heat release only. This assumption is valid when both $Q_{\mathrm{cl}} \approx 0$ and $Q_{\mathrm{vh}} \approx 0$ or when $Q_{\mathrm{cl}} + Q_{\mathrm{vh}} \approx 0$. The study of the effects of both heat loss and viscous heating, which are intrinsically two-dimensional [18,19], and their parametrization is the subject of ongoing research.

$$\rho_{\mathrm{m}} = \rho_{\mathrm{m}}^0(1-\alpha)(1-\beta)(1-c)\,; \quad \rho_{\mathrm{c}} = \rho_{\mathrm{c}}^0(1-\alpha)\beta \tag{8a}$$

$$\rho_{\mathrm{d}} = \rho_{\mathrm{m}}^0(1-\alpha)(1-\beta)c\,; \quad \rho_{\mathrm{g}} = \rho_{\mathrm{g}}^0\alpha \tag{8b}$$

$$\rho = \rho_{\mathrm{m}} + \rho_{\mathrm{c}} + \rho_{\mathrm{d}} + \rho_{\mathrm{g}} \tag{8c}$$

$$\alpha = \frac{4}{3}\pi r_{\mathrm{b}}^3 n; \quad \frac{\partial}{\partial t}(Sn) + \frac{\partial}{\partial x}(SnV) = 0\,; \quad p = \rho_{\mathrm{g}}^0 RT \tag{9}$$

Here α is the volume concentration of bubble, β is the volume concentration of crystals in the condensed phase (melt plus crystals), and c is mass concentration of dissolved gas (equal to volume concentration as we assume that the density of dissolved volatiles is the same as the density of the melt), ρ_0^{m} denotes the mean density of the pure melt phase, ρ_0^{c} is density of the pure crystal phase (with $\rho_{\mathrm{c}} = \rho_{\mathrm{ph}} + \rho_{\mathrm{mc}}$, $\beta = \beta_{\mathrm{ph}} + \beta_{\mathrm{mc}}$), r_b is the bubble radius, and n the number density of bubble per unit volume. Concerning the parametrization of mass transfer rate functions, we use:

$$J = 4\pi r_{\mathrm{b}} n D \rho_{\mathrm{m}}^0 \left(c - C_f\sqrt{p}\right) \tag{10}$$

$$G_{\mathrm{mc}} = 4\pi\rho_{\mathrm{c}}^0(1-\beta)(1-\alpha) \times U(t)\int_0^t I(\omega)\left(\int_\omega^t U(\eta)\,\mathrm{d}\eta\right)^2 \mathrm{d}\omega \tag{11a}$$

$$G_{\mathrm{ph}} = 3\gamma_s\left(\frac{4\pi N_{\mathrm{ph}}\beta_{\mathrm{ph}}^2}{3}\right)^{\frac{1}{3}}\rho_{\mathrm{c}}^0(1-\beta)(1-\alpha)\,U(t) \tag{11b}$$

Here J is parametrized using the analytical solution described in [69], U is the linear crystal growth rate ($\mathrm{m\,s^{-1}}$), I is the nucleation rate ($\mathrm{m^{-3}s^{-1}}$), which defines the number of newly nucleated crystal per cubic meter, and γ_{s} is a shape factor of the order of unity, D and C_f are the diffusion and the solubility coefficients, respectively. Concerning the mass transfer due to crystallization G_{mc}, we adapt a model similar to that described in [38]. Assuming spherical crystals, the Avrami–Johnson-Mehl–Kolmogorov equation in the form adopted by [46], for the crystal volume increase rate, is:

$$\frac{\mathrm{d}\beta}{\mathrm{d}t} = 4\pi\Upsilon_t U(t)\int_0^t I(\omega)\left(\int_\omega^t U(\eta)\,\mathrm{d}\eta\right)^2 \mathrm{d}\omega$$

where $\Upsilon_t = (1-\beta)(1-\alpha)$ is the volume fraction of melt remaining uncrystallized at the time t. Therefore, we have $G_{\mathrm{mc}} = \rho_{\mathrm{mc}}\mathrm{d}\beta/\mathrm{d}t$. For the phenocryst growth rate $G_{\mathrm{ph}}()$

we assume that it is proportional to the phenocryst volume increase rate $\mathrm{d}\beta_{\mathrm{ph}}/\mathrm{d}t = 4\pi R_{\mathrm{ph}}^2 N_{\mathrm{ph}} U(t)$ times the crystal density ρ_{c}^0 times the volume fraction of melt remaining uncrystallized at the time t. A detailed description of the parametrization used for the different terms is reported in [63].

For parametrizations of magma permeability k and magma viscosity μ we use:

$$k = k(\alpha) = k_0 \alpha^j \tag{12}$$

$$\mu = \mu_{\mathrm{m}}(c, T)\, \theta(\beta)\, \eta(\alpha, Ca) \tag{13}$$

where k is assumed to depend only on bubble volume fraction α. Magma viscosity μ depends on water content, temperature, crystal content, bubble fraction and capillary number as described in detail in the next section.

Regarding equations for semi-axes, a and b, we assume that the elliptical shape is maintained and that pressure change gradually in respect with vertical coordinate and time so that the plain strain analytical solution for an ellipse subjected to a constant internal overpressure [64,66], remains valid:

$$a = a_0 + \frac{\Delta P}{2G}\left[-(1-2\nu)a_0 + 2(1-\nu)b_0\right] \tag{14a}$$

$$b = b_0 + \frac{\Delta P}{2G}\left[2(1-\nu)a_0 - (1-2\nu)b_0\right] \tag{14b}$$

where ΔP is the overpressure, i. e. the difference between conduit pressure and far field pressure (here assumed to be lithostatic for a sake of simplicity), a_0 and b_0 are the initial values of the semi-axes, ν is the host rock Poisson ratio, and G is the host rock rigidity.

Equations (2)–(14) are solved between the top of the magma chamber and the bottom of the lava dome that provides some constant load by using the numerical method described in [63]. The effects of dome height and morphology changes are not considered in this paper. We consider three different kinds of boundary conditions at the inlet of the dyke: constant pressure, constant influx rate and the presence of a magma chamber located in elastic rocks. The case of constant pressure is applicable when a dyke starts from either a large magma chamber or unspecified source, so that pressure variations in the source region remain small. An estimate of the volume of magma stored in the source region that allows pressure to be approximated as constant depends on wall-rock elasticity, magma compressibility (volatile content), and the total volume of the erupted material. If the magma flow at depth is controlled by regional tectonics, the case of constant influx rate into the dyke may be applicable if total variations in supply rate are relatively small on the timescale of the eruption.

For the case where magma is stored in a shallow magma chamber prior to eruption, and significant chamber replenishment occurs, the flow inside the conduit must be coupled with the model for the magma chamber. In this case, as explained in detail in [63], we assume that the relationship between the pressure at the top of the magma chamber p_{ch} and the intensity of influx Q_{in} and outflux Q_{out} of magma to and from the chamber is given by:

$$\frac{\mathrm{d}p_{\mathrm{ch}}}{\mathrm{d}t} = \frac{4G\langle K\rangle}{\langle\rho\rangle V_{\mathrm{ch}}(3\langle K\rangle + 4G)}(Q_{\mathrm{in}} - Q_{\mathrm{out}}) \tag{15}$$

where V_{ch} is the magma chamber volume, $\langle\rho\rangle$ and $\langle K\rangle$ are the average magma density and magma bulk modulus, respectively, and G is the rigidity of rocks surrounding the chamber.

Cases of constant influx and of constant source pressure are the limit cases of Eq. (15) in the case of infinitely small and infinitely large magma chamber volume. We assume that the volume concentration of bubbles and phenocrysts are determined by equilibrium conditions and that the temperature of the magma is constant. The effect of temperature change on eruption dynamics, due to interaction between silicic and basaltic magma, was studied in [63].

We use a steady-state distribution of parameters along the conduit as an initial condition for the transient simulation. The values are calculated for a low magma discharge rate, but the particular value of this parameter is not important because the system deviates from initial conditions to a cyclic or stabilized state, which does not depends on the initial conditions.

Rheology of Crystal-Bearing Magma and Conduit Resistance Magma viscosity is modeled as a product of melt viscosity $\mu_{\mathrm{m}}(c, T)$, the relative viscosity due to crystal content $\theta(\beta) = \Theta(\beta)\varphi(\beta)$, and the relative viscosity due to the presence of bubbles $\eta(\alpha, Ca)$. The viscosity of the pure melt $\mu_{\mathrm{m}}(c, T)$ is calculated according to [36]. Viscosity increase due to the presence of the crystals is described through the function $\Theta(\beta)$ [16,20,21]. As crystallization proceeds, the remaining melt becomes enriched in silica- and melt viscosity increases. The parametrization of this effect is described by the function $\varphi(\beta)$ in [21,25]. Effects of the solid fraction are parametrized as described in [21].

Effects due to the presence of bubbles are accounted for by adopting a generalization of [51] for an elliptical conduit [20,21].

In the case of Newtonian magma rheology, the friction force in an elliptical conduit can be obtained from

a classical Poiseuille solution for low Reynolds number flow $F_c = 4\mu(a^2 + b^2)/(a^2 b^2)V$ [47]. High crystal or bubble content magmas may show non-Newtonian rheology. One possible non-Newtonian rheology is that of a Bingham material characterized by a yield strength τ_b [4]. The stress-strain relation for this material is given by:

$$\tau_{ij} = \left(\mu + \frac{\tau_b}{\gamma_b}\right)\gamma_{ij} \Leftrightarrow \tau > \tau_b \qquad \gamma_{ij} = 0 \Leftrightarrow \tau \leqslant \tau_b \tag{16}$$

Here τ_{ij} and γ_{ij} are the stress and strain rate tensors, τ and γ_b are second invariants of corresponding tensors. According to this rheological law, the material behaves linearly when the applied stress is higher than a yield strength. No motion occurs if the stress is lower than a yield strength. In the case of a cylindrical conduit the average velocity can be calculated in terms of the stress on the conduit wall τ_w [52]:

$$V = \frac{1}{12}\frac{r}{\tau_w^3 \mu}\left(\tau_b^4 + 3\tau_w^4 - 4\tau_b \tau_w^3\right) \tag{17}$$

Here r is the conduit radii. This form of equation gives an implicit relation between ascent velocity and pressure drop, and is not convenient to use. By introducing dimensionless variables $\Pi = \mu V/\tau_b r$ and $\Omega = \tau_w/\tau_b \geqslant 1$ relation (17) can be transformed into:

$$\Omega^4 - \frac{1}{6}(8 + 3\Pi)\,\Omega^3 + \frac{1}{3} = 0 \tag{18}$$

Following [63] a semi-analytical solution can be used for (18) and the conduit friction force can be expressed finally as:

$$F_c = \frac{2\tau_w}{r} = \frac{2\tau_b \Omega\,(\Pi)}{r}\,.$$

We note that a finite pressure gradient is necessary to initiate the flow in the case of Bingham liquid, in contrast to a Newtonian liquid.

Results and Applications

Influence of non-Newtonian Properties on Eruption Behaviour Now we compare the dynamics of magma extrusion in the cases of Newtonian and Bingham rheology. We will assume that yield strength is reached when the concentration of crystals reaches a critical value:

$$\tau = \begin{cases} \tau_b & \text{for } \beta > \beta_{\text{cr}} \\ 0 & \text{for } \beta \leqslant \beta_{\text{cr}} \end{cases} \tag{19}$$

Figure 6a shows a set of steady-state solutions for different values of τ_b. Values of τ_b and β_{cr} depend on crystal shape, crystal size distribution, magma temperature and other properties, but here are assumed to be constant. To illustrate the influence of Bingham rheology, the value of $\beta_{\text{cr}} = 0.65$ was chosen so that, for discharge rate larger then $\sim 5\,\text{m}^3/\text{s}$, the magma has Newtonian rheology (see Fig. 6a). A more detailed study would require measurements of the rheological properties of magma for a wide range of crystal content and crystal size distributions. As the value of τ_b the chamber pressure that is necessary to start the eruption increases.

Figure 6b shows the influence of these two rheological models on the dynamics of magma extrusion. In the case of Bingham rheology, magma discharge rate between the two pulses is zero until a critical chamber overpressure is reached. Then the magma discharge rate increases rapidly with decrease in crystal content, leading to a significant reduction of both magma viscosity and the length of the part of the conduit that is occupied by the Bingham liquid where $\beta_c > \beta_{\text{cr}}$. There is a transition in the system to the uppermost flow regime and the pressure then decreases quickly. Because the pressure at the onset of the pulse was significantly larger than in the case of a Newtonian liquid, the resulting discharge rate in the case of Bingham rheology is also significantly higher.

Modeling of Conduit Flow during Dome Extrusion on Shiveluch Volcano The maximum intensity of extrusion was reached at an early stage in all three eruptions (Fig. 3). We therefore suggest that dome extrusion was initiated by high overpressure in the magma chamber with respect to the lithostatic pressure. Depressurization of the magma chamber occurred as a result of extrusion. Without magma chamber replenishment, depressurization results in a decrease in magma discharge rate. In open system chambers replenishment of the chamber during eruption can lead to pulsatory behaviour [3]. The following account is derived from [25]. For the 1980–1981 eruption the monotonic decrease in discharge rate indicates that there was little or no replenishment of the magma chamber. During the 1993–1995 and 2001–2004 episodes, however, the magma discharge rate fluctuated markedly, suggesting that replenishment was occurring. The influx of new magma causes an increase in magma chamber pressure, and a subsequent increase in magma discharge rate. During the 2001–2004 eruption there were at least three peaks in discharge rate. Replenishment of the magma chamber with new hot magma can explain the transition from lava dome extrusion to viscous lava flow that occurred on Shiveluch after 10 May 2004, and which continues at the time of writing (2007).

We simulated dome growth during the 2001–2002, because this dataset is the most complete and is supported

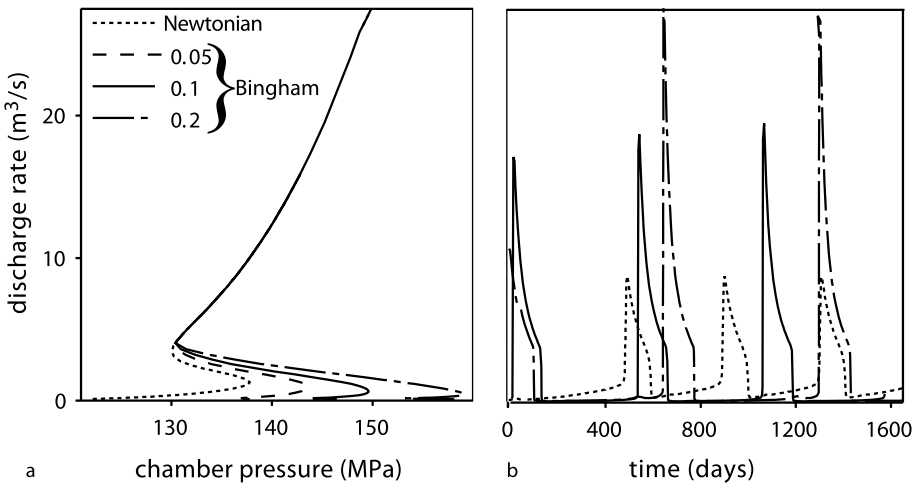

Volcanic Eruptions: Cyclicity During Lava Dome Growth, Figure 6

a Steady-state solutions and dependence of discharge rate on time for Newtonian and Bingham rheology of the magma. Yield strength is a parameter marked on the curves (values in MPa). For Bingham rheology discharge rate remains zero between the pulses of activity. Bingham rheology results in much higher chamber pressures prior to the onset of activity and, therefore, much higher discharge rates in comparison with Newtonian rheology. b Comparison of the period of pulsation in discharge rate for Newtonian and Bingham rheologies. After [63]

by petrological investigations [39]. We assume the shape of the influx curve:

$$Q_{\text{in}} = \begin{cases} 0, & t < t_s \\ Q_0, & t_s \leqslant t \leqslant t_f \\ 0, & t > t_f \end{cases} \tag{20}$$

Influx occurs with constant intensity Q_0 between times t_s and t_f. We have examined many combinations of values of these parameters within the constraints provided by observations. The best simulation results use the following values of parameters: $Q_0 = 3.8\ \text{m}^3/\text{s}$, $t_s = 77$ and $t_f = 240$ days. A more continuous influx, dependent on time, is plausible, but there is no geophysical evidence that allows us to constrain the intensity of the influx, because ground deformation data are absent for Shiveluch volcano. The output of the model gives a magma chamber volume of 12 km^3, assuming a spherical chamber. Figure 7a shows the time-dependence of magma discharge rate, and Fig. 7b shows the increase in the volume of erupted material with time after 6th June 2001. The timing of magma influx is in good agreement with the residence time of basaltic magma in the system, as calculated from the olivine reaction rims. For further details see [25].

5 to 7 Weeks Cycles on the Soufriere Hills Volcano: Evidence for a Dyke? An approximately 5 to 7 week cyclic pattern of activity was recognized at the Soufrière Hills Volcano (SHV) [87,91] between April 1997 and March 1998 from peaks in the intensity of eruptive activity and geophysical data, including tilt and seismicity (Fig. 2).

In models discussed above, the time-scale of pulsations depends principally on the volume of the magma chamber, magma rheology and the cross-sectional area of the conduit. These models might provide an explanation for the 2–3 year cycles of dome extrusion observed at SHV, where deformation data indicate that the magma chamber regulates the cycles. However, the models cannot simultaneously explain the 5–7 week cycles. Thus another mechanism is needed.

The evidence for a dyke feeder at SHV includes GPS data [58], distribution of active vents, and seismic data [76]. We have assumed that, at dept, the conduit has an elliptical shape that transforms to a cylinder at shallow level. In order to get a smooth transition from the dyke at depth to a cylindrical conduit (see Fig. 1) the value of a_0 in Eqs. (14) is parametrized as:

$$a_0(x) = A_1 \arctan\left(\frac{x - L_T}{w_T}\right) + A_2 \tag{21}$$

Here L_T and w_T are the position and the vertical extent of the transition zone between the ellipse and the cylinder and constants A_1 and A_2 are calculated to satisfy conditions $a_0(L) = R$ and $a_0(0) = a_0$, where R is the radius of the cylindrical part of the conduit and a_0 is the length of semi-major axis at the inlet of the dyke. The value of b_0 is calculated in order to conserve the cross-section area of

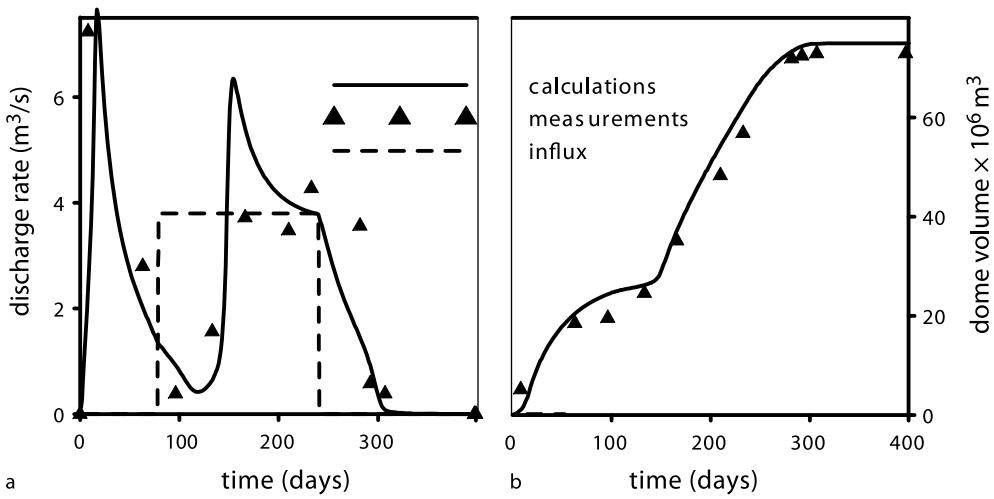

Volcanic Eruptions: Cyclicity During Lava Dome Growth, Figure 7
a Comparison of calculated and measured discharge rates (a) and volumes of the dome (b) for the episode of the dome growth in 2001–2002. Influx into the magma chamber is shown by a *dashed line* in a. Time in days begins on 6th June, 2001. After [25]

the unpressurized dyke, although it can also be specified independently.

In order to de-couple the influence of the dyke geometry from the oscillations caused by magma chamber pressure variations, we have assumed a fixed chamber pressure as a boundary condition for the entrance to the conduit. This assumption is valid because the timescale of chamber pressure variations are much longer than the period of the cycle (2–3 years in comparison with 5–7 weeks).

Results presented in Figure 8 show that, even with a fixed chamber pressure, there are magma discharge rate oscillations. At the beginning of a cycle the magma discharge rate is at a minimum, while the overpressure (here presented for 1 km depth by a dashed line) and dyke width are at a maximum. At point A in Fig. 8a the crystal content and viscosity have reached their maximum values. Beyond this threshold condition, an increase in magma discharge rate results in decreasing pressure and dyke width. However, crystal content and viscosity also decrease and this effect decreases friction, resulting in flow rate increase and pressure decrease. At C the system reaches minimum viscosity and crystal content, which cannot decline further. Thereafter the magma discharge rate decreases, while the pressure and dyke width increase. The dyke acts like a capacitor, storing volume during this part of the cycle.

The period of oscillation depends on several parameters such as influx rate and dyke aspect ratio a/R. Typically the period decreases with increasing aspect ratio. The range of calculated periods varies between 38 and 51 days for semi major-axis lengths, a, from 175 to 250 m and semi minor-axes, b, from 2 to 4 m. These results match observed cyclicity at SHV. The start of a cycle is quite sharp (Fig. 2), with the onset of shallow hybrid-type (impulsive, low-frequency coda) earthquakes. The change to shorter period and higher amplitude tilt pulsations indicate a marked increase in average magma discharge rate [91,98]. The model cycles also have rapid onsets. The high-amplitude tilt pulsations lasted for several weeks [91], consistent with the duration of higher magma discharge rates early in each 5–7 week cycle. Tilt data (Fig. 2) are consistent with the model in that the episode of high magma discharge is associated with a marked deflation that lasts several weeks (see dashed curve at Fig. 8, representing magmatic overpressure at 1 km depth). The magma pressure builds up in the swelling dyke and then reaches a threshold, whereupon a surge of partly crystallized magma occurs, accompanied by elevated seismicity.

The models presented above have certain general features that are necessary to show cyclic behaviour. First of all, the resistance of the conduit must depend on magma discharge rate in a way that resistance decreases when discharge rate increases in some range of discharge rate. This dependence is reproduced by a sigmoidal curve. Resistance is a product of viscosity and velocity, and is linearly proportional to discharge rate. This means that magma viscosity must decrease as discharge rate strongly increases. There may be many reasons for this behaviour, including crystallization, temperature variation or gas diffusion. The second condition is that there must be some capacitor in the system that can store magma in a period of low discharge rate and release it in a period of high discharge rate. The role of this capacitor can be played

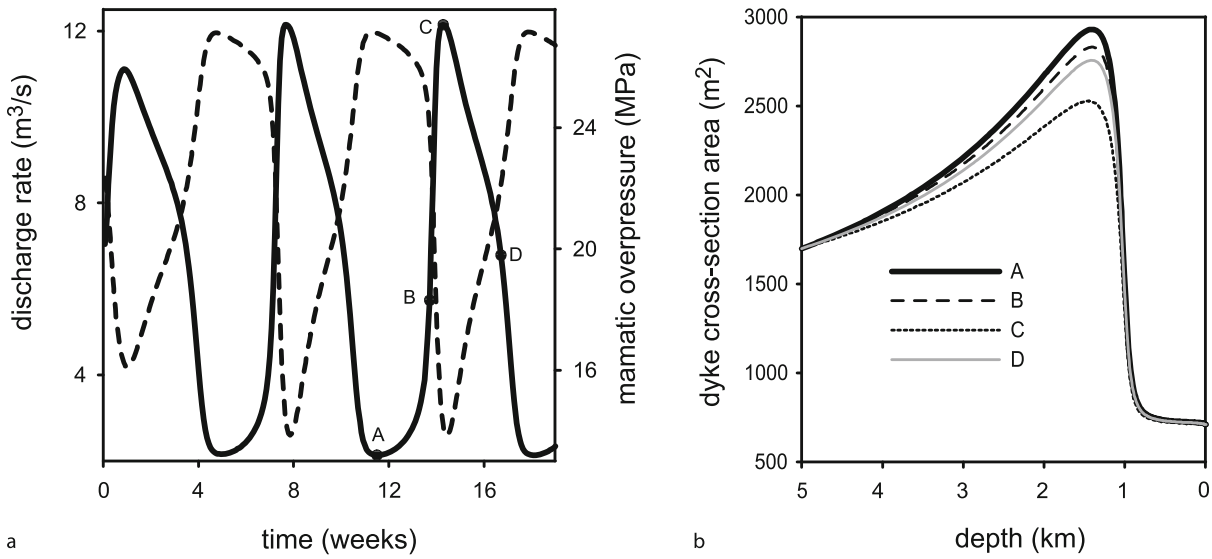

Volcanic Eruptions: Cyclicity During Lava Dome Growth, Figure 8
a Dependence of magma discharge rate (*solid line*) and magmatic overpressure at depth of 1 km (*dashed line*) on time, for $a = 240$ m and $b = 2.25$ m at the inlet of the dyke. The period of cycle is 46 days, average discharge rate is 6.2 m^3/s, with peak rate about 12 m^3/s. b Profiles of cross-section areas of the conduit during one cycle. Curve A corresponds to the beginning of the cycle, B – to a point on the curve of ascending discharge rate, C – to maximum discharge, and D to the middle of descending discharge curve. At the beginning of the cycle, due to large viscosity of magma (at low discharge rate crystal content is high) large magmatic overpressure develops, reaching a maximum near the transition between the dyke and cylindrical conduit; the dyke inflates providing temporary magma storage. Minimum dyke volume corresponds to maximum discharge rate (curve C). After [21]

by a magma chamber or dyke-shaped conduit located in elastic rocks, or by compressibility of the magma itself. The volumes of these capacitors are different and thus will cause pulsations with different periods. Currently there is no single model that can account for pulsations with multiple timescales.

Future Directions

Our models indicate that magmatic systems in lava dome eruptions can be very sensitive to small changes in parameters. This sensitivity is most marked when the system is close to the cusps of steady-state solutions. If magma discharge rate becomes so high that gas cannot escape efficiently during magma ascent, then conditions for magma fragmentation and explosive eruption can arise. Empirical evidence suggests that conditions for explosive eruption arise when magma discharge rates reach approximately 10 m^3/s or more in dome eruptions [45,83]. Calculations show the possibility of such high discharge rates for the system parameters typical of lava dome-building eruptions.

We have illustrated model sensitivity of results by varying only one parameter at a time on plots of chamber pressure and discharge rate. However, magmatic systems have many controlling parameters that may vary simultaneously. Furthermore, some controlling parameters are likely to be interdependent (such as temperature, volatile content and phenocryst content, for example) and others may be independent (such as magma temperature and conduit dimensions). An eruption can be expected to move through n-parameter space, making simulation and its parameter depiction difficult. Our results are simplified, so system sensitivity and behaviour in the real world may be yet more complex. A volcanic system may be quite predictable when it is within a stable regime, but may become inherently unpredictable [84,85] when variations in the parameters move the system towards transition points and flow regime boundaries.

As in all complex systems there are many controlling parameters. Our models capture some of the key dynamics, but are still simplified in many respects, so do not fully capture the real variations. Our models do not, for example, consider variations in dome height, gas escape to surrounding rocks, strain-rate dependent rheological effects or time dependent changes in conduit diameter. The model for porosity is based on interpretation of measurements of porosity of erupted magma. The role of post-eruptive alterations of pore structure, for example, formation of cooling cracks, cannot be easily estimated. The

model of bubble coalescence and permeability formation is important for understanding gas escape mechanisms and will provide constraints on transitions between extrusive and explosive activity. Because the model remains 1D, lateral distribution of parameters cannot be studied. These includes: lateral pressure gradients, magma crystallization on the conduit walls, wallrocks melting or erosion, formation of shear zones and shear heating, heat flux to surrounding rocks. The models also make the simplifying assumption that influx into the chamber from a deep source is a constant, or given as a function of time. The dynamics of the magma chamber itself are oversimplified in all existing conduit flow models. Changes in magma properties in magma chamber can affect the long-term evolution of eruptions. We have considered water as the only volatile and the addition of other gas species (e. g. CO_2 and SO_2) would add further variability.

There are large uncertainties in some parameters, which are likely to be very strong controls, such as the rheological properties of high crystalline magmas and crystal growth kinetic parameters, notably at low pressures (< 30 MPa) where experiments are very difficult to do (e. g. [22]). More experiments are necessary to understand the rheology of multiphase systems containing melt, crystals and bubbles. The effects of crystal shape, crystal size distribution and strain rate remain largely unclear. Some parameters, such as conduit geometry variation with depth, are highly uncertain. With so many parameters, good fits can be achieved by selecting plausible values for real systems. Barmin et al. [3], for example, were able to reproduce the patterns of discharge rate at Mount St.Helens and Santiaguito quite accurately. However, such models are not unique, partly because the actual values of some parameters may be quite different to the assumed values and partly because of the model simplifications.

Results obtained from waveform inversions of very-long-period seismic data over the past few years point to the predominance of a crack-like geometry for volcanic conduits [11], and it is becoming increasingly evident that the details of this geometry, such as conduit inclination, a sudden change in conduit direction, a conduit bifurcation, or a sudden increase in cross section, are all critically important in controlling magma flow dynamics. Future modeling attempts will need to be closely tied to information on conduit geometry derived from seismology to provide a more realistic view of volcanic processes.

The full simulation of any particular volcanic eruption in such a non-linear and sensitive system may appear a hopeless task. However, some reduction in uncertainties will certainly help to make the models more realistic. Further experimental studies of crystallization kinetics and the rheological properties of magma at high crystallinities are among the most obvious topics for future research. Advances in understanding the controls on magma input into an open-system chamber would be beneficial, since the delicate balance between input and output is a prime control on periodic behaviour.

Further model development includes 2D effects, elastic deformation effects in dyke-fed domes and coupling between magma chamber and conduit flow dynamics. Even with such improvements, large parameter uncertainties and modeling difficulties will remain. In such circumstances the logical approach is to start quantifying the uncertainties and sampling from them to produce probabilistic outputs based on ensemble models where numerical models of the kind discussed here can be run many times. A future challenge for numerical models will also be to produce simulated outputs which compare in detail with observations, in particular time series of magma discharge rates.

Acknowledgments

This work was supported by NERC research grant reference NE/C509958/1. OM and AB acknowledge Russian Foundation for Basic Research (08-01-00016) and President of Russian Federation program (NCH-4710.2006.1). RSJS acknowledges a Royal Society Wolfson Merit Award. The Royal Scoiety exchange grants and NERC grants had supported the Bristol/Moscow work over the last 10 years.

Bibliography

Primary Literature

1. Balmforth NJ, Burbidge AS, Craster RV (2001) Shallow Lava Theory. In: Balmforth NJ, Provenzale A (eds) Geomorphological Fluid Mechanics. Lecture Notes in Physics, vol 582. Springer, Berlin, pp 164–187
2. Balmforth NJ, Burbidge AS, Craster RV, Rust AC, Sassi R (2006) Viscoplastic flow over an inclined surface. J Non-Newtonian Fluid Mech 139:103–127
3. Barmin A, Melnik O, Sparks RSJ (2002) Periodic behaviour in lava dome eruptions. Earth Planet Sc Lett 199:173–184
4. Bingham EC (1922) Fluidity and Plasticity. McGraw–Hill, New York, p 215
5. Blake S (1990) Viscoplastic models of lava domes. In: Fink JH (ed) Lava flows and domes, Vol 2. In: Fink JH (ed) Lava flows and domes; emplacement mechanisms and hazard implications. Springer, Berlin, pp 88–126
6. Blundy JD, Cashman KV, Humphreys MCS (2006) Magma heating by decompression-driven crystallisation beneath andesite volcanoes. Nature 443:76–80. doi:10.1038/nature05100
7. Calder ES, Luckett R, Sparks RSJ, Voight B (2002) Mechanisms of lava dome instability and generation of rockfalls and pyroclastic flows at Soufrière Hills Volcano, Montserrat. In: Druitt TH, Kokelaar BP (eds) The Eruption of Soufrière Hills Volcano,

Montserrat, from 1995 to 1999, Geological Society, London, Memoirs No 21, pp 173–190
8. Calder ES, Cortés JA, Palma JL, Luckett R (2005) Probabilistic analysis of rockfall frequencies during an andesite lava dome eruption: The Soufrière Hills Volcano, Montserrat. Geophys Res Lett 32:L16309. doi:10.1029/2005GL023594
9. Caricchi L, Burlini L, Ulmer P, Gerya T, Vassalli M, Papale P (2007) Non-Newtonian rheology of crystal-bearing magmas and implications for magma ascent dynamics. Earth Planet Sci Lett 10.1016/j.epsl.2007.09.032
10. Cashman KV, Blundy JD (2000) Degassing and crystallization of ascending andesite and dacite. In: Francis P, Neuberg J, Sparks RSJ (eds) Causes and consequences of eruptions of andesite volcanoes. Phil Trans Royal Soc London A, pp 1487–1513
11. Chouet B, Dawson P, Arciniega–Ceballos A (2005) Source mechanism of Vulcanian degassing at Popocate'petl Volcano, Mexico, determined from waveform inversions of very long period signals. J Geophys Res 110:B07301. doi:10.1029/2004JB003524
12. Christiansen RL, Peterson DW (1981) Chronology of the 1980 Eruptive Activity. In: Lipman PW, Mullineaux DR (eds) The 1980 Eruptions of Mount St. Helens. Washington, US Geological Survey Professional Paper 1250, p 844
13. Clarke AB, Stephens S, Teasdale R, Sparks RSJ, Diller K (2007) Petrological constraints on the decompression history of magma prior to Vulcanian explosions at the Soufrière Hills volcano, Montserrat. J Volcanol Geotherm Res 161:261–274. doi:10.1016/j.jvolgeores.2006.11.007
14. Cole P, Calder ES, Sparks RSJ, Clarke AB, Druitt TH, Young SR, Herd R, Harford CL, Norton GE (2002) Deposits from dome-collapse and fountain-collapse pyroclastic flows at Soufrière Hills Volcano, Montserrat. In: Druitt TH, Kokelaar BP (eds) The eruption of the Soufrière Hills Volcano, Montserrat from 1995 to 1999. Geological Society, London, Memoir No 21, pp 231–262
15. Connor CB, Sparks RSJ, Mason RM, Bonadonna C, Young SR (2003) Exploring links between physical and probabilistic models of volcanic eruptions: the Soufriere Hills Volcano, Montserrat. Geophys Res Lett 30. doi:10.1029/2003GLO17384
16. Costa A (2005) Viscosity of high crystal content melts: dependence on solid fraction. Geophys Res Lett 32:L22308. doi:10.1029/2005GL02430
17. Costa A, Macedonio G (2002) Nonlinear phenomena in fluids with temperature-dependent viscosity: an hysteresis model for magma flow in conduits. Geophys Res Lett 29(10). doi:1029/2001GL014493
18. Costa A, Macedonio G (2003) Viscous heating in fluids with temperature-dependent viscosity: implications for magma flows. Nonlinear Proc Geophys 10:545–555
19. Costa A, Macedonio G (2005) Viscous heating in fluids with temperature-dependent viscosity: triggering of secondary flows. J Fluid Mech 540:21–38
20. Costa A, Melnik O, Sparks RSJ (2007) Controls of conduit geometry and wallrock elasticity on lava dome eruptions. Earth Planet Sci Lett 260:137–151. doi:10.1016/j.epsl.2007.05.024
21. Costa A, Melnik O, Sparks RSJ, Voight B (2007) The control of magma flow in dykes on cyclic lava dome extrusion. Geophys Res Lett 34:L02303. doi:1029/2006GL027466
22. Couch S, Sparks RSJ, Carroll MR (2001) Mineral disequilibrium in lavas explained by convective self-mixing in open magma chambers. Nature 411:1037–1039
23. Denlinger RP, Hoblitt RP (1999) Cyclic eruptive behaviour of silicic volcanoes. Geology 27(5):459–462
24. Diller K, Clarke AB, Voight B, Neri A (2006) Mechanisms of conduit plug formation: Implications for vulcanian explosions. Geophys Res Lett 33:L20302. doi:10.1029/2006GL027391
25. Dirksen O, Humphreys MCS, Pletchov P, Melnik O, Demyanchuk Y, Sparks RSJ, Mahony S (2006) The 2001–2004 dome-forming eruption of Shiveluch Volcano, Kamchatka: Observation, petrological investigation and numerical modelling. J Volcanol Geotherm Res 155:201–226. doi:10.1016/j.jvolgeores.2006.03.029
26. Druitt TH, Young S, Baptie B, Calder E, Clarke AB, Cole P, Harford C, Herd R, Luckett R, Ryan G, Sparks RSJ, Voight B (2002) Episodes of cyclic Vulcanian explosive activity with fountain collapse at Soufrière Hills volcano, Montserrat. In: Druitt TH, Kokelaar BP (eds) The eruption of the Soufrière Hills Volcano, Montserrat from 1995 to 1999. Geological Society, London, Memoir No 21, pp 231–262
27. Eichelberger JC, Carrigan CR, Westrich HR, Price RH (1986) Non-explosive silicic volcanism. Nature 323:598–602
28. Fedotov SA, Dvigalo VN, Zharinov NA, Ivanov VV, Seliverstov NI, Khubunaya SA, Demyanchuk YV, Markov LG, Osipenko LG, Smelov NP (2001) The eruption of Shiveluch volcano on May–July 2001. Volcanol Seis 6:3–15
29. Fink JH, Griffiths RW (1990) Radial spreading of viscous gravity currents with solidifying crust. J Fluid Mech 221:485–509
30. Fink JH, Griffiths RW (1998) Morphology, eruption rates, and rheology of lava domes: Insights from laboratory models. J Geophys Res 103:527–545
31. Green, DN, Neuberg J (2006) Waveform classification of volcanic low-frequency earthquake swarms and its implication at Soufrire Hills Volcano, Montserrat. J Volcanol Geotherm Res 153:51–63. doi:10.1016/j.jvolgeores.2005.08.003
32. Hale AJ, Bourgouin L, Mühlhaus HB (2007) Using the level set method to model endogenous lava dome growth. J Geophys Res 112:B03213. doi:10.1029/2006JB004445
33. Hale AJ, Wadge G (2003) Numerical modeling of the growth dynamics of a simple silicic lava dome. Geophys Res Lett 30(19). doi:10.1029/2003GL018182
34. Hammer JE, Rutherford MJ (2002) An experimental study of the kinetics of decompression-induced crystallization in silicic melt. J Geophys Res 107:(B1). doi:10.1029/2001JB000281
35. Harris AL, Rose WI, Flynn LP (2002) Temporal trends in Lava Dome extrusion at Santiaguito 1922–2000. Bull Volcanol 65:77–89
36. Hess KU, Dingwell DB (1996) Viscosities of hydrous leucogranite melts: A non-Arrhenian model. Am Mineral 81:1297–1300
37. Hoblitt RP, Wolfe EW, Scott WE, Couchman MR, Pallister JS, Javier D (1996) The preclimactic eruptions of Mount Pinatubo, June 1991. In: Newhall CG, Punongbayan RS (eds) Fire and Mud: Eruptions and Lahars of Mount Pinatubo, Philippines. Philippine Institute of Volcanology and Seismology, Quezon City, and University of Washington Press, Seattle, pp 457–511
38. Hort M (1998) Abrupt change in magma liquidus temperature because of volatile loss or magma mixing: effects of Nucleation, crystal growth and thermal history of the magma. J Petrol 39:1063–1076
39. Humphreys M, Blundy, JD, Sparks RSJ (2006) Magma Evolution and Open-system processes at Shiveluch Volcano: insights from phenocryst zoning. J Petrol 47:(12) 2303–2334. doi:10.1093/petrology/eg1045

40. Huppert HE, Shepherd JB, Sigurdsson H, Sparks RSJ (1982) On lava dome growth, with application to the 1979 lava extrusion of the Soufriere, St Vincent. J Volcanol Geotherm Res 14: 199–222
41. Huppert HE, Woods AW (2002) The role of volatiles in magma chamber dynamics. Nature 420:493–495
42. Ida Y (1996) Cyclic fluid effusion accompanied by pressure change: Implication for volcanic eruptions and tremor. Geophys Res Lett 23:1457–1460
43. Iverson RM et al (2006) Dynamics of seismogenic volcanic extrusion at Mount St. Helens in 2004–05. Nature 444:439–443
44. Jaquet O, Sparks RSJ, Carniel R (2006) Magma Memory recorded by statistics of volcanic explosions at the Soufriere Hills Volcano, Montserrat. In: Mader HM, Coles SG, Connor CB, Connor LJ (eds) Statistics in Volcanology. Geological Society, London, Special Publication of IAVCEI, vol 1. pp 175–184
45. Jaupart C, Allegre CJ (1991) Gas content, eruption rate and instabilities of eruption regime in silicic volcanoes. Earth Planet Sci Lett 102:413–429
46. Kirkpatrick R (1976) Towards a Kinetic Model for the Crystallization of Magma Bodies. J Geophys Res 81:2565–2571
47. Landau L, Lifshitz E (1987) Fluid Mechanics, 2nd edn. Butterworth–Heinmann, Oxford
48. Lejeune A, Richet P (1995) Rheology of crystal-bearing silicate melts: An experimental study at high viscosità. J Geophys Res 100:4215–4229
49. Lensky NG, Sparks RSJ, Navon O, Lyakhovsky V (2007) Cylic activity at Soufriere Hills volcano, Montserrat: degassing-induced pressurization and stick-slip extrusion. In: Lane SJ, Gilbert JS (eds) Fluid motions in volcanic conduits: a source of seismic and acoustic signals. Geolocical Society, London, Special Publications, vol 307, pp 169–188. doi:10.1144/SP307.100305-8719/08/$15.00 The Geological Society of London 2008
50. Lister JR, Kerr RC (1991) Fluid mechanical models of crack propagation and their application to magma transport in dykes. J Geophys Res 96:10049–10077
51. Llewellin EW, Manga M (2005) Bubble suspension rheology and implications for conduit flow. J Geotherm Res 143: 205–217
52. Loitsyansky LG (1978) Fluid and gas mechanics. Nauka, Moscow, pp 847 (in Russian)
53. Maeda I (2000) Nonlinear visco-elastic volcanic model and its application to the recent eruption of Mt. Unzen. J Volcanol Geotherm Res 95:35–47
54. Mason RM, Starostin AB, Melnik O, Sparks RSJ (2006) From Vulcanian explosions to sustained explosive eruptions: The role of diffusive mass transfer in conduit flow dynamics. J Volcanol Geotherm Res 153:148–165. doi:10.1016/j.jvolgeores.2005.08.011
55. Marsh BD (2000) Reservoirs of Magma and Magma chambers. In: Sigurdsson H (ed) Encyclopedia of volcanoes. Academic Press, New York, pp 191–206
56. Mastin GL, Pollard DD (1988) Surface Deformation and Shallow Dike Intrusion Processes at Inyo Craters, Long Valley, California. J Geophys Res 93(B11):13221–13235
57. Matthews SJ, Gardeweg MC, Sparks RSJ (1997) The 1984 to 1996 cyclic activity of Lascar Volcano, northern Chile: Cycles of dome growth, dome subsidence, degassing and explosive eruptions. Bull Volcanol 59:72–82
58. Mattioli G, Dixon TH, Farina F, Howell ES, Jansma PE, Smith AL (1998) GPS measurement of surface deformation around Soufriere Hills volcano, Montserrat from October 1995 to July 1996. Geophys Res Lett 25(18):3417–3420
59. Melnik O (2000) Dynamics of two-phase conduit flow of high-viscosity gas-saturated magma: large variations of sustained explosive eruption intensity. Bull Volcanol 62:153–170
60. Melnik O, Barmin A, Sparks RSJ (2005) Dynamics of magma flow inside volcanic conduits with bubble overpressure buildup and gas loss through permeable magma. J Volcanol Geotherm Res 143:53–68
61. Melnik O, Sparks RSJ (1999) Non-linear dynamics of lava dome extrusion. Nature 402:37–41
62. Melnik O, Sparks RSJ (2002) Dynamics of magma ascent and lava extrusion at Soufrière Hills Volcano, Montserrat. In: Druitt TH, Kokelaar BP (eds) The eruption of the Soufrière Hills Volcano, Montserrat from 1995 to 1999. Geological Society, London, Memoir No 21, pp 223–240
63. Melnik O, Sparks RSJ (2005) Controls on conduit magma flow dynamics during lava dome building eruptions. J Geophys Res 110(B02209). doi:10.1029/2004JB003183
64. Mériaux C, Jaupart C (1995) Simple fluid dynamic models of volcanic rift zones. Earth Planet Sci Lett 136:223–240
65. Murphy MD, Sparks SJ, Barclay J, Carroll MR, Brewer TS (2000) Remobilization origin for andesite magma by intrusion of mafic magma at the Soufrière Hills Volcano. In: Montserrat WI (ed) A trigger for renewed eruption. J Petrol 41:21–42
66. Muskhelishvili N (1963) Some Basic Problems in the Mathematical Theory of Elasticity. Noordhof, Leiden, The Netherlands
67. Nakada S, Eichelberger JC (2004) Looking into a volcano: drilling Unzen. Geotimes 49:14–17
68. Nakada S, Shimizu H, Ohta K (1999) Overview of the 1990–1995 eruption at Unzen Volcano. J Volcanol Geoth Res 89:1–22
69. Navon O, Lyakhovsky V (1998) Vesiculation processes in silicic magmas. In: Gilbert J, Sparks RSJ (eds) The Physics of explosive volcanic eruption. Geological Society London, Special Pubblication, vol 145. pp 27–50
70. Neuberg JW, Tuffen H, Collier L, Green D, Powell T, Dingwell D (2006) The trigger mechanism of low-frequency earthquakes on Montserrat. J Volcanol Geotherm Res 153:37–50
71. Newhall CG, Melson WG (1983) Explosive activity associated with the growth of volcanic domes. J Volcanol Geoth Res 17:111–131
72. Norton GE, Watts RB, Voight B, Mattioli GS, Herd RA, Young SR, Devine JD, Aspinall WP, Bonadonna C, Baptie BJ, Edmonds M, Harford CL, Jolly AD, Loughlin SC, Luckett R, Sparks RSJ (2002) Pyroclastic flow and explosive activity of the lava dome of Soufrière Hills volcano, Montserrat, during a period of no magma extrusion (March 1998 to November 1999). In: Druitt TH, Kokelaar BP (eds) The eruption of the Soufrière Hills Volcano, Montserrat from 1995 to 1999. Geological Society, London, Memoir No 21, pp 467–482
73. Ohba T, Kitade Y (2005) Subvolcanic hydrothermal systems: Implications from hydrothermal minerals in hydrovolcanic ash. J Volcanol Geotherm Res 145:249–262
74. Robertson R, Cole P, Sparks RSJ, Harford C, Lejeune AM, McGuire WJ, Miller AD, Murphy MD, Norton G, Stevens NF, Young SR (1998) The explosive eruption of Soufriere Hills Volcano, Montserrat 17 September, 1996. Geophys Res Lett 25:3429–3432
75. Roman DC (2005) Numerical models of volcanotectonic earthquake triggering on non-ideally oriented faults. Geophys Res Lett 32, doi:10.1029/2004GL021549

76. Roman DC, Neuberg J, Luckett RR (2006) Assessing the likelihood of volcanic eruption through analysis of volcanotectonic earthquake fault-plane solutions. Earth Planet Sci Lett 248:244–252
77. Rubin AM (1995) Propagation of magma-filled cracks. Annu Rev Planet Sci 23:287–336
78. Saar MO, Manga M, Katharine VC, Fremouw S (2001) Numerical models of the onset of yield strength in crystal–melt suspensions. Earth Planet Sci Lett 187:367–379
79. Sahagian D (2005) Volcanic eruption mechanisms: Insights from intercomparison of models of conduit processes. J Volcanol Geotherm Res 143(1–3): 1–15
80. Slezin YB (1984) Dispersion regime dynamics in volcanic eruptions, 2. Flow rate instability conditions and nature of catastrophic explosive eruptions. Vulkanol Seism 1:23–35
81. Slezin YB (2003) The mechanism of volcanic eruptions (a steady state approach). J Volcanol Geotherm Res 122:7–50
82. Sparks RSJ (1978) The dynamics of bubble formation and growth in magmas – a review and analysis. J Volcanol Geotherm Res 3:1–37
83. Sparks RSJ (1997) Causes and consequences of pressurization in lava dome eruptions. Earth Planet Sci Lett 150:177–189
84. Sparks RSJ (2003) Forecasting Volcanic Eruptions. Earth and Planetary Science Letters Frontiers in Earth Science Series 210:1–15
85. Sparks RSJ, Aspinall WP (2004) Volcanic Activity: Frontiers and Challenges. In: Forecasting, Prediction, and Risk Assessment. AGU Geophysical Monograph "State of the Planet" 150, IUGG Monograph 19, pp 359–374
86. Sparks RSJ, Murphy MD, Lejeune AM, Watts RB, Barclay J, Young SR (2000) Control on the emplacement of the andesite lava dome of the Soufriere Hills Volcano by degassing-induced crystallization. Terra Nova 12:14–20
87. Sparks RSJ, Young SR (2002) The eruption of Soufrière Hills volcano, Montserrat (1995–1999): overview of scientific results. In: Druitt TH, Kokelaar BP (eds) The eruption of the Soufrière Hills Volcano, Montserrat from 1995 to 1999. Geological Society, London, Memoir No 21, pp 45–69
88. Sparks RSJ, Young SR, Barclay J, Calder ES, Cole PD, Darroux B, Davies MA, Druitt TH, Harford CL, Herd R, James M, Lejeune AM, Loughlin S, Norton G, Skerrit G, Stevens NF, Toothill J, Wadge G, Watts R (1998) Magma production and growth of the lava dome of the Soufrière Hills Volcano, Montserrat, West Indies: November 1995 to December 1997. Geophys Res Lett 25:3421–3424
89. Swanson DA, Holcomb RT (1990) Regularities in growth of the Mount St. Helens dacite dome 1980–1986. In: Fink JH (ed) Lava flows and domes; emplacement mechanisms and hazard implications. Springer, Berlin, pp 3–24
90. Voight B, Hoblitt RP, Clarke AB, Lockhart AB, Miller AD, Lynch L, McMahon J (1998) Remarkable cyclic ground deformation monitored in real-time on Montserrat, and its use in eruption forecasting. Geophys Res Lett 25:3405–3408
91. Voight B, Sparks RSJ, Miller AD, Stewart RC, Hoblitt RP, Clarke A, Ewart J, Aspinall W, Baptie B, Druitt TH, Herd R, Jackson P, Lockhart AB, Loughlin SC, Lynch L, McMahon J, Norton GE, Robertson R, Watson IM, Young SR (1999) Magma flow instability and cyclic activity at Soufrière Hills Volcano, Montserrat. Science 283:1138–1142
92. Walker GPL (1973) Lengths of lava flows. Philos Trans Royal Soc A 274:107–118
93. Watson IM et al (2000) The relationship between degassing and ground deformation at Soufriere Hills Volcano, Montserrat. J Volcanol Geotherm Res 98(1–4):117–126
94. Watts RB, Sparks RSJ, Herd RA, Young SR (2002) Growth patterns and emplacement of the andesitic lava dome at Soufrière Hills Volcano, Montserrat. In: Druitt TH, Kokelaar BP (eds) The eruption of the Soufrière Hills Volcano, Montserrat from 1995 to 1999. Geological Society, London, Memoir No 21, pp 115–152
95. Whitehead JA, Helfrich KR (1991) Instability of flow with temperature-dependent viscosity: a model of magma dynamics. J Geophys Res 96:4145–4155
96. Williams SN, Self S (1983) The October 1902 Plinian eruption of Santa Maria volcano, Guatemala. J Volcanol Geotherm Res 16:33–56
97. Woods AW, Koyaguchi T (1994) Transitions between explosive and effusive eruption of silicic magmas. Nature 370:641–645
98. Wylie JJ, Voight B, Whitehead JA (1999) Instability of magma flow from volatile-dependent viscosity. Science 285: 1883–1885
99. Yokoyama I, Yamashita H, Watanabe H, Okada H (1981) Geophysical characteristics of dacite volcanism – 1977–1978 eruption of Usu volcano. J Volcanol Geotherm Res 9:335–358

Books and Reviews

Dobran F (2001) Volcanic Processes: Mechanisms In Material Transport. Kluwer, New York, pp 620
Gilbert JS, Sparks RSJ (eds) (1998) The Physics of Explosive Volcanism. Special Publication of the Geological Society of London, vol 145, pp 186
Gonnermann H, Manga M (2007) The fluid mechanics inside a volcano. Ann Rev Fluid Mech 39:321–356
Mader HM, Coles SG, Connor CB, Connor LJ (2006) Statistics in Volcanology. IAVCEI Publications, Geological Society Publishing House, p 296

Volcanic Eruptions, Explosive: Experimental Insights

Stephen J. Lane, Michael R. James
Lancaster Environment Centre, Lancaster University, Lancaster, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Volcanic Materials
Volcanic Processes
Analogue Approach
Future Directions
Acknowledgments
Bibliography

Glossary

Conduit Subterranean pathways to the surface created and sustained by the flow of volcanic materials are known as volcanic conduits. In the laboratory, tubes are used as scale equivalents. The two terminologies ('conduits' and 'tubes') should be kept distinct to avoid inadvertent acceptance that a laboratory experiment is a direct replica of a volcanic process.

Magma When rock melts underground it becomes magma. When magma flows on the surface it is known as lava and when it is projected into the atmosphere it forms pyroclasts. Most magma is silicate-based and comprises a three-phase mixture of solid (crystals, called microlites when small), liquid (melt) and gas (bubbles or vesicles).

Rheology When a force is applied to matter it deforms. The rheology of a material describes the relationship between applied force and the material response. Volcanic materials have a wide range of rheological behaviors dependent on the proportions of gas, liquid and solid, temperature, chemical composition, deformation history and rate of deformation.

Scaling Explosive volcanic eruptions involve temperatures of order 1000°C, pressures of many MPa and physical sizes from μm to km. Laboratory study of these processes generally involves observation and measurement of processes operating at lower temperature or pressure, or at smaller physical size. Scaling is the means of making laboratory-scale analogues representative of volcano-scale natural processes.

Volatiles Explosive volcanic eruptions are driven by the exsolution of a dissolved gas phase as magma pressure falls. These gaseous species, dominated by water vapor on Earth, are known as volatiles.

Definition of the Subject

The magnitude of volcanic eruptions may be classified using the Volcanic Explosivity Index (VEI) [95,99]. The lowest VEI index of zero covers non-explosive activity (such as the production of lava flows) and is generally associated with low magma effusion rates or low viscosity magmas, for example, carbonate-based carbonatites [127]. For silicate-based magmas, which are generally more viscous than carbonatites, the boundary between effusive and explosive volcanism is commonly marked by Hawaiian fire-fountaining activity [96]. Magmas with a high silica content, or a lower temperature (and thus higher viscosity), can erupt either explosively or non-explosively as lava flows or dome-building eruptions [128]. Some eruptions switch between effusive and explosive behavior dependent on prevailing physicochemical conditions (e. g., [54,62]), with magmas tending to become more explosive at higher effusion rates. High VEI eruptions are, therefore, associated with high flow rates, large erupted volumes and substantial eruption plumes.

Such explosive activity comprises a wide range of phenomena, which result from complex non-linear interactions and feedback mechanisms. Processes that initiate in the subsurface plumbing or conduit system determine the nature of following events and control eruption styles. On emerging from the vent, volcanic material enters the atmosphere and the ensuing interactions are key in determining the consequent transport of volcanic debris that defines the impact on the environment and on human lives and infrastructure. On Earth, our capability to mitigate such volcanic hazard relies in large part on forecasting explosive events, a process which requires a high degree of understanding about the physicochemical factors operating during explosive volcanism.

In common with many science disciplines, the approaches taken to gain an understanding of explosive volcanism have relied on a combination of field observations, theoretical models and laboratory models of materials and mechanisms. Field observations form the 'ground truth' for numerical and experimental modeling (and provide vital historical data), but are often very difficult to unambiguously interpret due to their scarcity in both space and time. Numerical modeling complements field observations by providing much greater information density and, hence, the potential to examine processes in great detail. However, the accuracy of the results depends on correctly including all relevant materials properties and system processes in the model and the results can be difficult to verify against often-unobservable field phenomena. Laboratory investigation of the properties of volcanic materials, their modes of interaction and the behaviors of analogous systems provides the third complementary approach; this is the area we focus on here.

Introduction

In the context of explosive volcanic activity, laboratory experiments cover a wide range of research, the most fundamental of which is to gain an understanding of the properties of the volcanic material involved. Without this to underpin further work, neither accurate computer nor laboratory simulations could be carried out. The materials properties that are relevant to explosive volcanism include the solubility in silicate melts of volatiles (most commonly water and carbon dioxide), melt viscosity as a function of chemical composition, magma rheology as a function of

the size, shape and proportion of crystals and bubbles, gas permeability of magma, magma deformation mode (e. g. brittle or ductile) as a function of strain-rate, and volatile diffusivity. Experimental determination of these parameters and their interdependencies results in empirically formulated relationships that can be used in numerical models and compared with the properties of analogue fluids. We first review aspects of the volcanic materials literature in Sect. "Volcanic Materials", with the aim of illustrating the nature of molten rock, the complexity of which underpins most explosive volcanic processes.

Experimental modeling of these processes can then build on the materials understanding. Such experiments involve investigation of the behavior of natural volcanic products at laboratory time and length scales, including the response of magma samples to rapid changes in pressure and temperature, the fall behavior of silicate particles in the atmosphere, and the generation and separation of electrostatic charge during explosive eruptions. Here, we encounter the concept of scaling [11,20], which links the applicability of experimental results using small samples of natural materials to natural scales that are often several orders of magnitude larger. For experiments that replace natural volcanic products with analogue materials (for example, in order to relax the experimental pressure and temperature requirements), scaling arguments are crucial. A scaling approach permits adjustment of experimental durations and sizes, whilst maintaining similitude with much larger and longer events. Consequently, analogue experimentation has allowed investigation of explosive phenomena in the broader context of fluid-dynamic behavior, and is a powerful means of giving 'insight' into volcanic processes.

Burgisser et al. [20] review and discuss scaling criteria for experiments and natural phenomena involving dilute multiphase mixtures in magmatic systems, and conclude "… that, despite numerous experimental studies on processes relevant to magmatic systems, some and perhaps most geologically important parameter ranges still need to be addressed at the laboratory scale". This emphasizes that small-scale experiments aiming to mimic larger-scale natural processes need to be considered in two contexts. Firstly, laboratory experiments yield valid results in their own right and provide a means of validating numerical models. Secondly, laboratory experiments are not facsimiles of the natural process, but can give perception into how aspects of natural systems behave provided those aspects are scaled; a process normally considered at design stage. Consequently, comparing experimental results with the natural process must always be undertaken with due caution. The analogue experiment literature is reviewed in Sects. "Volcanic Processes" and "Analogue Approach".

In Sect. "Definition of the Subject", the combined tactics of field observation, numerical model and experimental model was discussed. In practice, complete overlap of these techniques can be difficult to achieve because the parameters measurable in the field are often incompatible with experimentally accessible variables, which in turn may not coincide with the output from numerical models. Explosive volcanism is also a difficult phenomenon to observe in action, and there are three reasons for this. Firstly, explosions do not happen very often, and the larger the explosion the more infrequent it is, with, for example, repeat timescales of order 10^3 years for VEI 6 events like the Krakatau, 1883 eruption. This makes detailed syn-event measurements of large explosive episodes infrequent on the timescale of contemporary volcanology. Secondly, high VEI events tend to destroy near-field observation equipment, making measurements very difficult to obtain even when an event does occur. Thirdly, processes that comprise an explosive event can be difficult to access directly; perhaps a prime example of this is how to obtain data about the nature of flow in the volcanic conduit system during explosive activity. Direct measurements are not possible and our main source of information is the deformation of the conduit wall created by unsteadiness in the flow, which in turn is calculated from the interpretation of ground motion signals detected by seismometers (e. g., [31]).

These considerations mean that there are many field observations, numerical and experimental models that await comparison with their companion approaches, however, the detailed study of volcanoes with low VEIs is one area where combined tactics are possible. As an example, in Sect. "Analogue Approach" we review the contribution that experimental modeling has made to the understanding of fluid flow in the conduit feeding VEI 1–2 explosions at Stromboli Volcano, Italy. Here it has been possible to measure signals from many thousand events, over the period of a few years (e. g., [7]), with a network of seismometers and other instruments, as well as detailed sampling of eruption products (e. g., [72]). Such field data are key to testing both numerical and experimental models of explosive volcanic activity wherever it occurs.

Volcanic Materials

Magma Rheology

Rheology describes the way in which materials deform and flow under the influence of an applied stress. Explosive eruptions only occur because considerable stress is applied

to magma. The rheology of natural silicate materials was being measured before World War II (e. g., [14]), with empirical models (derived from experimental data) becoming established in the 1960s and 1970s (e. g., [18,92,107] and references therein).

When stress is applied to a solid it deforms or strains elastically, the strain being the macroscopic result of atoms in the solid moving closer together or further apart. These two parameters are related by a constant called a modulus, for example, in solids under tension $\sigma = E\gamma$, where σ is tensile stress, γ is tensile strain and E is Young's modulus. Similar relationships exist for deformation under shear or by pressure change. These moduli are likely to be a constant for a range of values of applied stress, but could become dependent on strain, or indeed strain rate.

When stress (σ) is applied to a fluid (the collective term for gas and liquid), viscosity (η) is the parameter that relates the rate of flow (strain rate, $\mathrm{d}\gamma/\mathrm{d}t$) to the stress that is being applied to it, giving the relationship $\sigma = \eta \mathrm{d}\gamma/\mathrm{d}t$. The flow of a fluid is the macroscopic result of fluid molecules moving past each other; therefore, changes that influence molecular motion also affect viscosity. The viscosity can be independent of the rate of flow, a case known as Newtonian behavior, and is best exhibited by gases and liquids of low molecular weight. Fluids whose viscosity is a function of strain-rate are described as non-Newtonian, and many foodstuffs, for instance, exhibit this behavior (e. g. tomato ketchup). Viscosity is also a function of temperature, generally decreasing as temperature increases and molecules move further apart (thermal expansion). An everyday example would be Golden or Maple syrup that becomes easier to pour on heating. The concept of viscosity fails when molecular interactions are rare, for example in gases at low pressure.

Newtonian Viscosity Natural silicates in the liquid state have been assumed to be Newtonian in behavior and demonstrate a strong viscosity dependence on temperature. The relationship was initially represented by an Arrhenius-type equation, such as $\eta = A \exp[E/RT]$, where the constants A and E are dependent on the chemical composition of the silicate melt (e. g., Murase and McBirney [92]), R is the gas constant and T the absolute temperature. With considerable further laboratory effort in measuring the viscosities of a wide compositional range of natural silicate liquids, it became clear that viscosity was not strictly an Arrhenian function of temperature. Hess and Dingwell [49] developed an empirical model to describe laboratory values of viscosity for leucogranitic melts as a function of temperature and water content. The compilation of data used by Hess and Dingwell [49] is shown in Fig. 1, and the empirical equation generated to fit the data is:

$$\log \eta = \frac{[-3.545 + 0.833\ln(w)] + [9601 - 2368\ln(w)]}{\{T - [195.7 + 32.25\ln(w)]\}},$$

where w is the water % w/w. The constants in this equation will be different for silicate melts of different composition; see Zhang et al. [133] for an updated and modified version of this model. Water is important here because it is a ubiquitous component of explosive volcanism on Earth. The concentration of water has a large and non-linear impact on viscosity (Fig. 1a) because it acts as a network modifier by breaking Si-Si bonds in a silicate melt, thus decreasing the average molecular weight.

The use of laboratory measurement to provide quantitative understanding of the dependence of viscosity on temperature, water content and other parameters has provided a key component in the numerical modeling of the explosive flow behavior of natural silicates, as well as pointing the way to identifying fluids for analogue experiments.

Non-Newtonian Viscosity There is laboratory evidence that the viscosity of liquid magmas is not a constant as once assumed, but may be subtly non-Newtonian [110]. The contents of a volcanic conduit are rarely 100% liquid, with crystals present in many silicate melts below 1250°C, considerably higher than most eruption temperatures. The effect of crystallization is to significantly change the relationship between viscosity and temperature [91], especially as crystal content is both temperature and time dependent. The presence of bubbles is also ubiquitous in explosive volcanic materials, and these too alter magma behavior in response to an applied stress. The presence of both yields a rheologically complex 3-phase material.

Combined laboratory and field-measurement of the viscosity of lavas from Mount Etna [98] showed that above 1120°C Newtonian behavior was exhibited. Eruption temperatures ranged from 1084–1125°C; therefore, for much of this range the magma behaved in a non-Newtonian fashion as a consequence of the interaction between crystals suspended within the liquid silicate melt. The Etnean basalt was found to be thixotropic, meaning that the viscosity decreased the longer it was deformed, perhaps indicating that crystals were becoming aligned and could consequently move past each other more readily. The basalt also showed pseudoplastic or shear-thinning behavior, meaning that viscosity decreased as the rate of flow increased, perhaps again, because of the dynamics of crystal alignment. Finally, the basalt showed a small but measurable yield stress of < 78 Pa, below which it behaved more

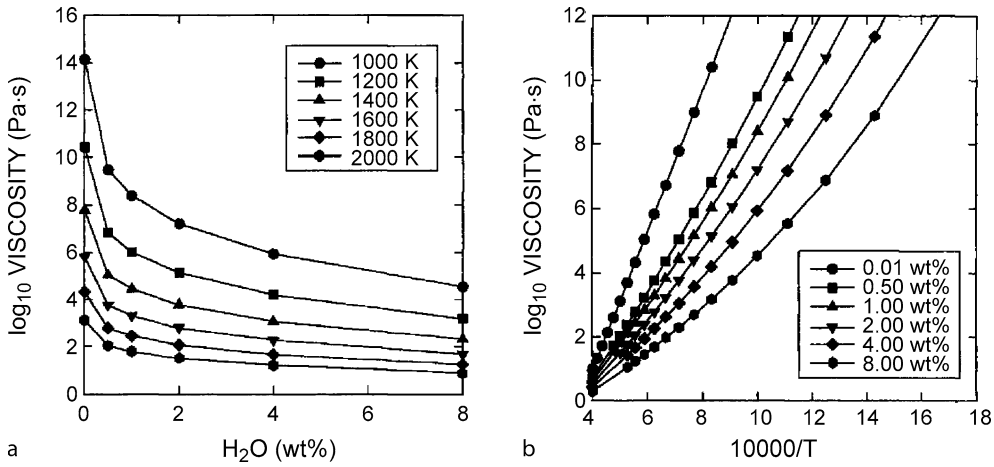

Volcanic Eruptions, Explosive: Experimental Insights, Figure 1

Model curves, based on experimental measurement, of the dependence of viscosity of metaluminous leucogranitic melts on dissolved water content (a) and temperature (b) from a compilation of data (Hess and Dingwell [49]). Note the logarithmic viscosity scale and non-linear nature of the relationships. Explosive volcanic activity often results in water loss from the melt, and cooling results from expansion and interaction with the atmosphere; melt viscosity increases result. The role this plays on the dynamics of expanding vesiculated flows may be studied by analogue experimental means. Figure reproduced with kind permission of the Mineralogical Society of America

like a solid material. The yield stress is likely to be a consequence of crystals weakly binding together into a mesh-like structure. The small nature of the yield stress means that the viscosity of these magmas could be described by a power law equation (i. e., the yield stress was ignored) where $\sigma = \eta\,(\mathrm{d}\gamma\mathrm{d}t)^n$. Above 1120°C, $n = 1$ and Newtonian behavior was observed, but as temperature falls to 1084°C, n decreases to 0.46, meaning that viscosity decreases as strain rate increases. Pinkerton and Norton [98] also noticed that physical motion of the silicate melt could induce crystallization as seen by Kouchi et al. [68]. Similar behavior has been found more recently for laboratory measurements of basalts from Japan [55].

Hoover et al. [51] showed that crystals induce yield strength in magmas at volume fractions as low as 0.2, although this is strongly dependent on crystal shape. These experiments were carried out on natural basaltic samples, as well as using an analogous system of neutrally buoyant particles in corn syrup (Fig. 2). Natural samples do, by definition, provide data directly applicable to the natural system, but analogue experiments are complementary because particles have controlled size, shape and concentration. These experiments provide a good example of the kind of complex interplay of time-dependent processes that can occur in magmatic systems that tend only to be quantified by detailed experimentation.

At eruption temperatures, the non-Newtonian behavior of magmas is due in large part to the presence of crystal networks; however, during flow, bubbles must also move past each other and contribute to fluid rheology. Explosive eruptions only occur because bubbles form and expand; therefore, understanding bubble behavior is central to understanding explosive processes. When both bubble size and the fluid flow rate are small, then the surface tension between the silicate liquid and the water vapor in the bubble resists viscous forces and is able to minimize the bubble surface area maintaining spherical bubbles; the bubbles act as if they are solid spheres suspended in the silicate melt. However, if bubbles and flow rate are larger, then surface tension is no longer able to maintain a spherical shape against the viscous drag, and bubbles elongate in the flow field. Under conditions of steady flow, when bubble shape is stable with time, these two behaviors may be distinguished by the capillary number, $Ca = \lambda\gamma'$ where $\lambda = a\eta/\sigma$, a being bubble diameter, η viscosity, γ' strain rate and σ surface tension (e. g., Llewellin and Manga [74]). When Ca is significantly less than unity, surface tension forces dominate and bubbles remain near spherical. However, when Ca is significantly greater than unity surface tension is overcome by the combination of liquid shear and viscosity, forcing bubbles to elongate. Ca provides an example of a parameter that may be used to scale between volcanic behavior and a laboratory experiment. If bubbles preserved in pumice from an eruption were highly elongated, then any experiment simulating this process would need to also produce highly elongated

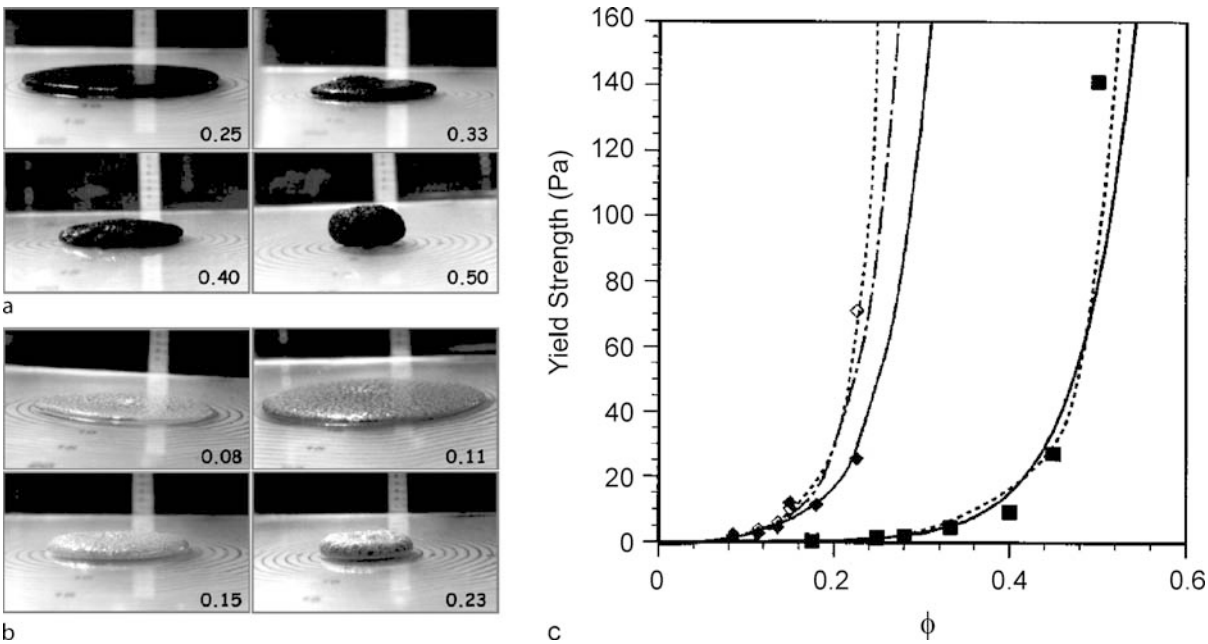

Volcanic Eruptions, Explosive: Experimental Insights, Figure 2

Magma has complex rheology partly because of its crystal content. Hoover et al. [51] found that basalt with crystals showed solid-like behavior at low stress. Analogue experiments using near-spherical poppy seeds (a) and highly irregular shavings of kaolin/wax (b) in corn syrup demonstrate the role of particle concentration and shape in giving yield strength to solid-liquid mixtures, the number in each panel being the volume fraction of particles. Yield strength increased rapidly (c) for volume fractions of round poppy seed (filled squares) above about 0.4, allowing extruded material to retain considerable vertical topography (a). The same effect occurs at volume fractions around 0.2 for irregular particles (b) because they interact more strongly with each other than spherical particles. Crystals in magma not only act as nucleation sites for bubbles, but also have a very strong rheological influence in their own right. Reprinted with kind permission from Elsevier

bubbles and not spherical bubbles. The shape-change undergone by bubbles provides an added complexity over the effects of crystals, whose shapes are generally constant.

Bagdassarov and Dingwell [8,9] experimentally measured the rheological behavior of vesicular rhyolite, using oscillatory rheometry, with a liquid viscosity between 10^9 and 10^{12} Pa s and gas bubble volume fractions in the region between zero and 0.3. Under these conditions, where $Ca \gg 1$, it was shown that vesicular rhyolite decreased in viscosity with increasing gas volume fraction. Such rheological experiments on hot, high-viscosity natural silicate melts are difficult, and the scaling of the small-strain results to larger scales is not straightforward. Using an oscillating viscometer, Llewellin et al. [75] measured the viscosity of bubbly Golden Syrup, a fluid analogous to molten silicate with water vapor bubbles, but with liquid viscosity between 10^1 and 10^2 Pa s. The analogous fluid has the advantage of being liquid at room temperature, thereby placing much less stringent demands on any measurement apparatus. Llewellin et al. [75] showed the effect of increasing gas volume fraction from zero to 0.5 with small deformations ($Ca \ll 1$) caused viscosity to change by over an order of magnitude. They also discriminated between steady flow (bubbles not changing shape) and unsteady flow (bubbles in the process of changing shape) by proposing the dynamic capillary number $Cd = \lambda\gamma''/\gamma'$, where λ represents the timescale over which a deformed bubble can relax. Under steady flow conditions bubbles tended to increase viscosity, but under conditions of rapidly changing strain rate increasing gas volume fraction reduced viscosity; a result consistent with the oscillatory rheometry results of Bagdassarov and Dingwell [9]. This experimental result largely unified the seemingly contradictory literature that preceded it, and the consequences of these two regimes of behavior are discussed lucidly by Llewellin and Manga [74], where the influence of bubbles is shown to change some model outcomes by factors of two or more. This is an excellent example of the role that analogue experimental measurement can play in helping to understand complex and inaccessible natural phenomena.

Viscosity or Modulus? The concept of a relaxation time was encountered above in relation to how long a deformed (non-spherical) bubble takes to approach the relaxed spherical shape. This same approach can also be applied at the molecular level as liquids flow. During viscous flow, molecules move past each other, and this requires them to be able to reorient or deform. If the timescale of molecular deformation is much shorter than the timescale required by the strain rate during flow, then liquid or viscous behavior ensues. However, if the strain rate is such that molecules cannot reorient or deform, then the material behaves, at least in part, like an elastic solid, with strain being accommodated by bond stretching, or fracture, if yield stress is exceeded.

Materials that demonstrate aspects of liquid and solid behavior are known as viscoelastic. Viscoelasticity was first investigated in the nineteenth century using materials like rubber and glass. Viscoelastic behavior in volcanic materials begins with the onset of non-Newtonian liquid behavior at deformation timescales about 100 times longer than the relaxation time of silicate molecules in the melt [39]. The onset of non-Newtonian behavior suggests that there is a range of relaxation timescales for different molecules, and this may be a consequence of silicate melts having a range of molecular weights rather than a single value.

The *timescale* of molecular relaxation is given by the Maxwell relation, *viscosity*/*modulus*, i. e., the relative weighting of liquid to solid response when stress is applied. In volcanic materials the modulus is effectively constant as temperature and composition change, but viscosity changes by a factor of about 10^9 or more. At low magma viscosities, the relaxation timescale is short and viscous liquid behavior occurs at volcanically plausible strain rates. However, magmas of high viscosity have long relaxation timescales. Here, the onset of non-Newtonian behavior, and ultimately the glass transition to brittle solid behavior, occurs at much lower strain rates. Dingwell [38] discusses the relevance of this to explosive volcanic eruptions, and proposes the onset of brittle failure in high-viscosity magma under significant strain rates as a criterion for magma fragmentation. Zhang [130] showed that magma fragmentation also depends on the number density and size of bubbles within the melt.

Volatile Solubility

Volatiles are defined as those compounds that are dissolved within the matrix of a silicate melt at high pressures, but become supersaturated and, consequently, then exsolve in the supercritical or gaseous state as pressure drops towards atmospheric. On Earth, H_2O is the most common magmatic volatile, with CO_2, SO_2, HCl, and the halogens also present. For water, exsolution occurs into the supercritical state, or gas phase at depths less than about 1 km; therefore, understanding the physicochemical behavior of these compounds at volcanic temperatures and pressures is a key component in the modeling of explosive volcanic eruptions. Water itself has been the subject of extensive laboratory and numerical experiments (with Kalinichev [65] providing a review) because of its anthropogenic usefulness as well as geological importance. The densities of volatile species are less than that of the silicate melt, for instance at the depths of exolution the density of water is about 200–400 kg m^{-3}, a factor of $\sim$ 10 less than magma. The bulk modulus of water (order 10^7 Pa) is much less than that of silicate melt (about 10^{10} Pa, Alidibirov et al. [5]). Consequently, as pressure falls, any exsolved volatile phase expands by about three orders of magnitude more than that of the surrounding silicate melt; this is the fundamental driving mechanism of explosive volcanism.

The solubility of volatiles in silicate melts has, and continues to be, the subject of extensive experimentation. Carroll and Holloway [28] review volatiles in magmas, and Liu et al. [73] provide a well-referenced introduction to water and carbon dioxide in rhyolite, and give new experimental data and empirical relationships so important to numerical models of explosive volcanic activity. Empirical H_2O and CO_2 solubility in rhyolitic melt (700–1200°C and 0–500 MPa) are expressed by Liu et al. [73] as

$$\begin{aligned} H_2O_t = {} & \frac{\left(354.94P_w^{0.5} + 9.623P_w - 1.5223P_w^{1.5}\right)}{T} \\ & + 0.0012439P_w^{1.5} \\ & + P_{CO_2}\left(-1.084\times10^{-4}P_w^{0.5} - 1.362\times10^{-5}P_w\right) , \end{aligned}$$

$$\begin{aligned} CO_2 = {} & \frac{P_{CO_2}(5668 - 55.99P_w)}{T} \\ & + P_{CO_2}\left(0.4133P_w^{0.5} + 2.041\times10^{-3}P_w^{1.5}\right) , \end{aligned}$$

where H_2O_t is the total dissolved H_2O% w/w, CO_2 content is in mass-ppm, T is absolute temperature and P the partial pressure of water or carbon dioxide.

Figure 3 shows how the solubility of H_2O in rhyolite varies with pressure from a range of experimental work. Experiments focused on explosive volcanism have concentrated on the solubility of water at a temperature of 850°C, because this is a common eruption temperature for rhyolitic melts. Although water solubility increases as pressure increases, this relationship is not linear, especially at lower pressures, and the nature of the non-linearity is dependent on magma composition (see [86] for review). The root of the complexity of water solubility is that water

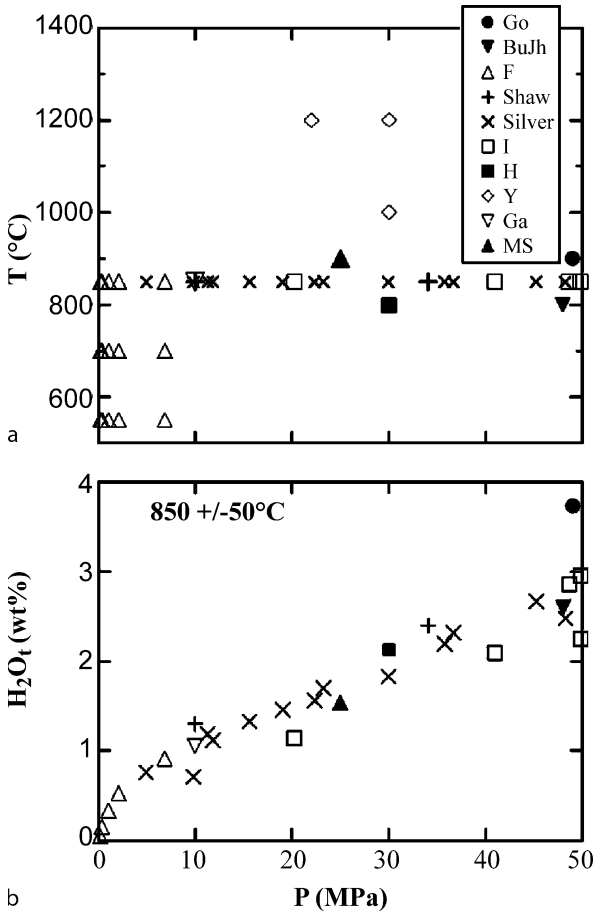

Volcanic Eruptions, Explosive: Experimental Insights, Figure 3
a shows pressures and temperatures of experiments carried out on rhyolite melts to investigate water solubility as a function of pressure, see Liu et al. [73] for sources. The bulk of the experiments are carried out at 850°C because this is a common temperature for rhyolitic eruptions. b The solubility of water decreases with pressure, becoming increasingly non-linear below 10 MPa (shallower than about 50 m depth for lithostatic pressure). The transition, and consequent expansion, of water going from the dissolved to the vapor (or supercritical) states as pressure decreases is the driving mechanism of explosive volcanic eruptions. Reprinted with kind permission from Elsevier

chemically reacts with the silica glass matrix as a network modifier. This reduction in the average molecular weight is responsible for decreasing the viscosity (Fig. 1), as well as having a range of complex geochemical effects (see [12] for popular review).

Nucleation and Diffusion

Decreasing the pressure applied to a volatile-saturated magma causes supersaturation of the volatile species. For explosive volcanism to occur, the volatile must make the transition from the dissolved state to a lower-density (and therefore higher volume) exsolved state, i. e., the process of bubble nucleation. Determining the degree of volatile supersaturation required for nucleation to take place has been the subject of significant laboratory effort. Magmas are complex chemical compounds that are rarely free of interfaces. Hurwitz and Navon [52] found experimentally that heterogeneous bubble nucleation dominated, with only small supersaturations of water vapor (< 1 MPa, or ~ 50 m head of magma) being required to nucleate bubbles at the surfaces of Fe-Ti oxide microlites. Other crystal-melt interfaces required supersaturations > 10 MPa for nucleation of water bubbles. Even at supersaturations of > 130 MPa, nucleation was heterogeneous even in apparently crystal-free melts.

More recently, Gardner [43] found experimentally that certain crystal interfaces or features acted as nucleation sites, in particular blocky shaped magnetite or the ends of needle-shaped hematite (Fig. 4). Gardner's [43] study, and others cited therein, indicate that bubble nucleation is a complex function of the composition and temperature of the melt, the composition and morphology of crystals, and the degree of supersaturation prevailing. Accurate modeling of bubble nucleation requires detailed knowledge of the melt as it becomes supersaturated. Nucleation could occur at low supersaturations, with conditions never far from equilibrium. Conversely, in the absence of suitable

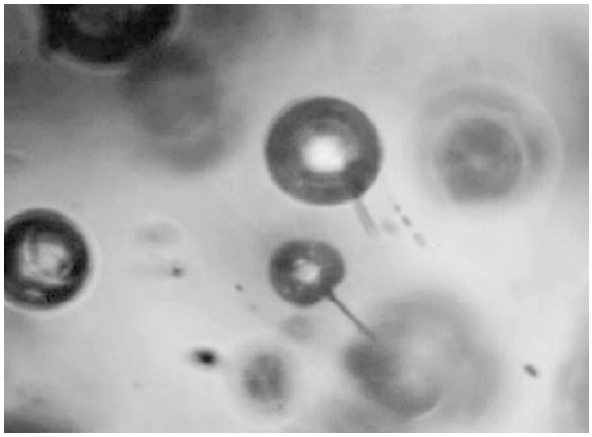

Volcanic Eruptions, Explosive: Experimental Insights, Figure 4
As pressure decreases, the melt becomes supersaturated in water. During the decompression of small samples of hydrated rhyolite, bubbles of water vapor nucleated then grew at the ends of needle-shaped hematite crystals, but not along their lengths (scale bar = 20 µm, Gardner [43]). Such 'imperfections' allow heterogeneous bubble nucleation and prevent large supersaturations developing. Reprinted with kind permission from Elsevier

nucleation sites, volatile supersaturations of the order of 100 MPa could develop as pressure in the melt decreases, and when nucleation does occur, rapid bubble growth can be expected.

Once bubbles of volatile have nucleated, growth is promoted by diffusion of volatiles from the supersaturated melt, and by expansion during pressure decrease. Any decrease in pressure also acts to increase volatile supersaturation (Fig. 3), encouraging either new bubble nucleation or, if bubbles are nearby, the diffusion of volatile into existing bubbles [17]. The diffusivity of volatiles is, therefore, a key parameter in determining the nature of the degassing process and has been reviewed by Watson [124]. Volatile diffusion into a bubble from the local melt surrounding it implies that the bubble will be surrounded by a shell of melt with relatively low volatile content and, thus, with elevated viscosity (Fig. 1a). This viscous melt shell may play an important role in the dynamics of bubble expansion during eruption, further adding to the importance of understanding diffusivity. Zhang and Behrens [131] experimentally investigated the diffusivity of water in rhyolitic melts, which, in combination with previous studies, covers the parameter range 400–1200°C, 0.1–810 MPa, and 0.1–7.7% w/w total H_2O content. The resulting empirical relationship between diffusivity of molecular water (D) and pressure (P), temperature (T) and water concentration (X) is given by

$$D_{H_2O_m} = \exp\left[\left(14.08 - \frac{13128}{T} - \frac{2.796P}{T}\right) + \left(-27.21 + \frac{36892}{T} + \frac{57.23P}{T}\right)X\right].$$

Such experimental determination of water diffusivity then forms key aspects of models of the behavior of silicate melts that undergo decompression, for instance, in how rapidly bubble pressure recovers after a perturbation [32], or in the nature of the resulting explosive activity [84].

Permeability

Erupting magma usually includes gas bubbles, which can approach and interact with each other due to differences in their rise velocities within the melt and due to expansion. If the residence timescale of magma in the plumbing system is long compared to the time required for bubbles to move through it, then the bubbles can be considered to be separating from the melt. This is often the situation in low VEI eruption styles such as Strombolian activity. Here, gas bubbles can ascend through the low viscosity magma relatively rapidly and are able to collide and concentrate, possibly forming bubble rafts or foam layers. Bubble coalescence may then create fewer, but larger bubbles with significantly increased ascent velocity that, on reaching the surface, result in Strombolian activity.

Conversely, if magma residence timescale is short in comparison to bubble rise velocity, then bubble interactions cannot take place so readily. Under these conditions, the rapid depressurization and expansion of the trapped gas bubbles often results in high VEI events, but not universally. The ability of some such magmas to erupt relatively passively indicates that, as they ascend, they are able to loose gas by some other mechanism. One way for this to happen is for the vesiculated magma to become increasingly able to accommodate gas flow and separation as the volume fraction of bubbles (porosity) increases, allowing gas to move from bubble to bubble to the surface or into the surrounding rock. Such gas escape can be quantified using permeability, which is proportional to the effective pore diameter. Establishing how bubbly magmas become permeable, and the nature of that permeability is, therefore, of key importance to understanding eruption mechanisms.

The permeability of eruption products has only been measured experimentally in recent decades, and by few workers (e. g., [42,67,87,104]). Eichelberger et al. [42] determined the room-temperature gas permeability of rhyolite with between 37% v/v and 74% v/v porosity. Over this range, permeability increased by over nearly four orders of magnitude from 10^{-16} to 10^{-12} m^2, precipitating the hypothesis that high permeability can suppress explosive behavior in silicic volcanism, provided gas can escape. Klug and Cashman [67] made laboratory measurements on 73 pumice samples with porosities ranging from 30% v/v to 92% v/v. Permeability ranged from 10^{-14} to 10^{-12} m^2, with the relationship that $k \propto \phi^{3.5}$, where k is permeability and ϕ is porosity. Permeabilities at high porosity values agree between these two investigations, but disagreement at low porosities suggests greater complexity. Saar and Manga [104] demonstrated that there is not a straightforward relationship between permeability and the volume fraction of bubbles. Samples of basaltic scoria, which expand isotropically then cool rapidly, and therefore contain near-spherical bubbles, are well behaved in their permeability-porosity relationship and may be modeled by established percolation theory. However, many multiphase volcanic products do not have spherical bubbles because anisotropic flow causes high values of the capillary number. The elongate nature of the resulting bubbles causes high permeabilities at low porosities [104]. However, the permeability, like the porosity, is no longer isotropic, and gas escapes in directions determined by the local direction

of flow as well as the prevailing pressure gradient. Laboratory measurements of volcanic products have been used to scale and validate theoretical models of magma permeability as a function of porosity. These models have, amongst others, been based on analogy with networks of resistors in electric circuits [16], and combination of classical approaches with fractal pore-space geometry [35]. Such models can then be incorporated into numerical simulations of flow in volcanic conduits that test the importance of permeability on explosive-effusive transitions.

Consequences

The consequences of the materials properties of magmas responsible for explosive volcanic flows might now be explored. At pressures higher than the total volatile vapor-pressure, magma remains bubble-free. When pressure is reduced sufficiently below that of the volatile vapor pressure, then bubbles of volatile nucleate in the magma. These bubbles grow, and as bubble density is less than that of the magma, the bulk magma density decreases. This may act to increase magma ascent rate, creating a mechanism of positive feedback. Simultaneously, the reducing volatile content of the magma rapidly increases the liquid viscosity, reducing the strain rate required for viscoelastic and brittle behavior. This process occurs at the length-scale of diffusion into bubbles as well as within the liquid as a whole and, depending on the timescales, crystal growth may also be changing, or indeed driving, the rheological evolution of the system.

These non-linear interactions underpin the complexity of understanding how magma flows, and illustrate the subsurface control on the nature of eruptions. In order to investigate how the properties of volcanic materials interact during the eruption process, experiments of a different nature are required.

Volcanic Processes

Experiments investigating materials properties are mainly designed to measure parameters under static or steady conditions. However, explosive eruptions are not static and only pseudo-steady at best; therefore, different experiments are needed to explore bulk flow behaviors. Experiments investigating the explosive response of volcanic materials to possible eruption triggers are best conducted using natural silicates that have been well characterized by previous study. These experiments can be designed to minimize the inherent complexity of the overall process by concentrating on specific aspects. Models verified by experiments then contribute to larger models of the whole process.

Explosive Processes

Magma is a hydrated silicate liquid (normally also containing some solid crystals) that, during the explosive eruption process, is converted into silicate fragments ranging in size from 10^1 to 10^{-6} m. The mechanism responsible is fragmentation and understanding fragmentation processes is a major goal of physical volcanology.

Explosive volcanic activity predominantly requires the formation and growth of gas bubbles within magma. Exsolution processes are initiated at pressures generally > 10 MPa, and may extend to pressures of several hundred MPa for water. Such pressures result from overburdens of several kilometers and flow experiments carried out at volcanic temperature and pressure (1100–1500K, up to at least 100 MPa) place severe constraints on containment vessels, measurement transducers and the volume of experimental samples. Another explosive mechanism occurs when volatile species mingle with hot magma at lower pressures resulting in hydrovolcanic activity. This is driven by expansion as liquid turns to gas on heating.

Direct observation of explosive processes is unlikely, but experiments mimicking volcanic conditions give valuable insight. Experimental exploration of explosive processes using natural samples requires that appropriate proportions of gas, liquid and solid are present together with suitable pressures and temperatures.

Hydrovolcanism Experiments investigating the response of hot natural silicate melts to explosive 'molten fuel-coolant interactions' (MFCIs) with liquid water have been carried out by Wohletz [126] and Zimanowski et al. [134]. MFCIs are not confined to geological processes and are a dangerous industrial hazard, for example, on 4 November 1975, at the Queen Victoria Blast Furnace, Appleby–Frodingham Steelworks, Scunthorpe, UK, up to 90 tons of molten metal were thrown over a wide area after an MFCI explosion involving about 2 tons of water and 180 tons of molten metal. Berthoud [13] reviews MFCIs from the perspective of the nuclear power and other industries, but the physics described also applies to hydrovolcanism.

Volcano-specific research into MFCIs has studied the size and morphology of silicate particles from volcanic eruptions and compared them to those generated experimentally by MFCI processes (Wohletz [126]). The ratio of water to magma was found to be a key parameter in determining the nature of the particles resulting from the explosive process, with particle size reducing as explosivity increases. Such experiments provide insight into hydrovolcanic eruptions where magma may interact explosively

with surface water (Surtseyan eruptions) or ground water (phreatomagmatic eruptions). Two processes dominantly produce the explosivity: (1) the rapid 'flashing' of liquid water to vapor; and (2) rapid cooling of the silicate melt causing thermal shock. Zimanowski et al. [135] studied how explosive fragmentation of a remelted volcanic rock could be interpreted from the size and shape of expelled silicate fragments. Two methods of fragmentation were used: (1) injection of air at 5 or 10 MPa into the base of the crucible containing the melt (at 1653 K); and (2) initiation of a MFCI using liquid water and a triggering perturbation. With an air injection at 5 MPa pressure, fine particles were uncommon and the particles generally rounded; however, the particle size distribution suggested that two unspecified fragmentation processes had operated. Three unspecified fragmentation modes were implied for the 10 MPa air injections, with finer, but still rounded, particles produced. Interaction with water (MFCI) also had three fragmentation mechanisms, but produced the finest particles, many of which were blocky. The MFCI experiments resulted in a much more complex fragmentation process than the air injection experiments (Zimanowski et al. [135]), with a key difference being the rapid cooling of the silicate melt by interaction with water. The combination of rapid cooling and high strain rates generated by large pressures is considered to cause brittle failure of the melt and generate the fine, blocky fragments during the MFCI experiment.

Although not hydrovolcanism, the air-injection method of Zimanowski et al. [135] has been used by Taddeucci et al. [117] to experimentally investigate the role of crystal volume fraction during the explosive fragmentation of magma. This is of interest because ascending magmas are thought to have both axial and radial heterogeneity in their crystal contents, both before and during explosive activity. Preliminary results showed a strong correlation between the fragment morphology and the crystal content in both experimental and natural samples, with increasing crystallinity promoting brittle fracture. This suggests a number of interpretations: (1) a higher viscosity liquid phase naturally exists in more crystalline samples; and/or (2) strain-rates in the smaller liquid volume of high crystallinity samples are higher, i. e., there is less liquid to accommodate a similar motion; and/or (3) the macroscopic rheological behavior of the two-phase mixture promotes brittle processes [51].

Comparisons between experimental and natural particles generated by MFCI experiments [22] show that quite subtle surface features can be used to distinguish between events involving either partial or complete water vaporization (Fig. 5). Where vaporization is complete, both experimental and natural particles are blocky and equant in shape, and a brittle fragmentation process is implied from observations of stepped surfaces on the particles, created by mechanical etching. Particles resulting from incomplete vaporization of water show similar features, except that their surfaces host a network of cracks that are considered to be the result of rapid surface cooling (thermal shock), created by interaction with excess liquid water. Büttner et al. [22] comment on the experimental relevance of their research; "In general, the artificial pyroclasts produced in the experiments cover all grain sizes and shape properties of the natural ones, thus illustrating the quality of the experimental scaling in terms of geometry and energy release." Such phenomenological similarities are a powerful indicator that experimental processes are

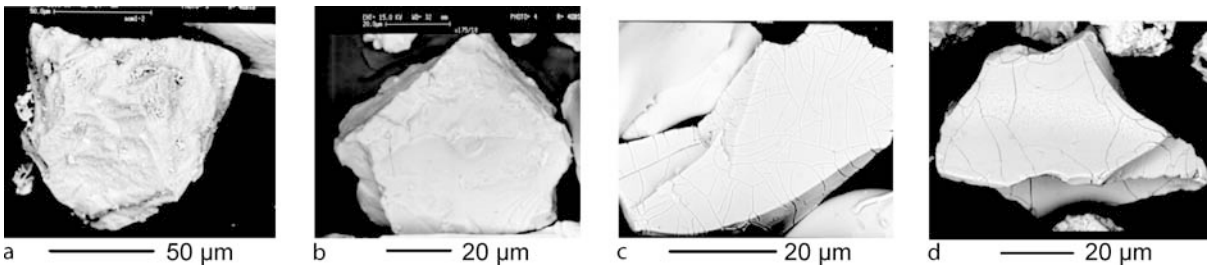

Volcanic Eruptions, Explosive: Experimental Insights, Figure 5

Magma and water may interact explosively during phreatomagmatic eruptions. Molten fuel-coolant interactions (MFCI) or magma-water interactions (MWI) can be carried out experimentally (Fig. 6). a shows a fragment generated by experimental MWI (Büttner et al. [22]) where all water was vaporized. b shows a fragment from a natural phreatomagmatic eruption that has not interacted with excess water. c shows an experimental fragment that 'erupted' through excess water, and d the product of phreatomagmatic activity thought to have interacted with excess water. The phenomenological similarities between a and b, and c and d provide powerful evidence that experiments are mimicking the processes that occurred during natural phreatomagmatic eruptions. Reprinted by kind permission from Macmillan Publishers Ltd: Nature (Büttner et al. [22]), copyright (1999)

mimicking the intended natural processes, and this is additional evidence of similarity. However, the experiments did not mimic all the post-explosion fragmentation processes that lead to the final natural product, demonstrating that care must be taken in comparing the outcomes of controlled experiments with the final product of the volcanic process.

Grunewald et al. [48] introduced a chemical aspect to MFCI experiments by using near-saturated NaCl solution as well as pure water, with the intention that the addition of other species would add complexity. Experimental methods were similar to that of Zimanowski et al. [135] (Fig. 6), with the triggering of a rapid thermal interaction by the injection of hot molten silicate with cool liquid brine. It was found that NaOH and HCl were formed during the heating of the NaCl solution, and that explosion energies were lower than with pure water. The lower explosion energies resulted in more rounded and larger silicate fragments as found by Zimanowski et al. [135]. The reason for the reduced physical energy of fragmentation was attributed to energy coupling into the endothermic chemical process that produced NaOH and HCl, but the heating of brine may well also have introduced a liquid NaCl phase (1074 K–1738 K) that altered the physical MFCI process. These experiments raise the interesting prospect of being able to use the chemistry of volcanic gases and particle surfaces to make inferences about physical processes occurring during hydrovolcanism.

A different experimental approach to studying magma-water interactions (MWI) was used by Trigila et al. [121] in which experiments could be carried out at pressures up to 200 MPa, and temperatures up to 1473 K. In contrast to the method of Zimanowski et al. [135], where water is injected into hot (1650 K), liquid-dominated silicate melt at 0.1 MPa, Trigila et al. [121] 'infused' water into permeable sintered samples of solid-dominated basalt at lower temperature (1100 K), but high pressure (8 MPa). The samples were manufactured by grinding basalt to grain sizes of < 10 μm (powder) or about 0.7 mm (granular). These were heated to experimental temperatures and allowed to sinter for 30 minutes to form a permeable cylinder (about 2 cm in diameter and 6 cm in length) into which water was introduced. Only the more permeable granulated samples underwent spontaneous explosive behavior on mixing with water, generating particles similar in size to the original grains. These exploratory experiments indicate the importance of permeability to MFCI processes where groundwater could interact with hot, porous country rock as well as magma. We have already explored the importance of permeability in allowing exsolving volatiles to escape and reduce explosivity; under different circumstances we now we see that permeability can also increase explosivity.

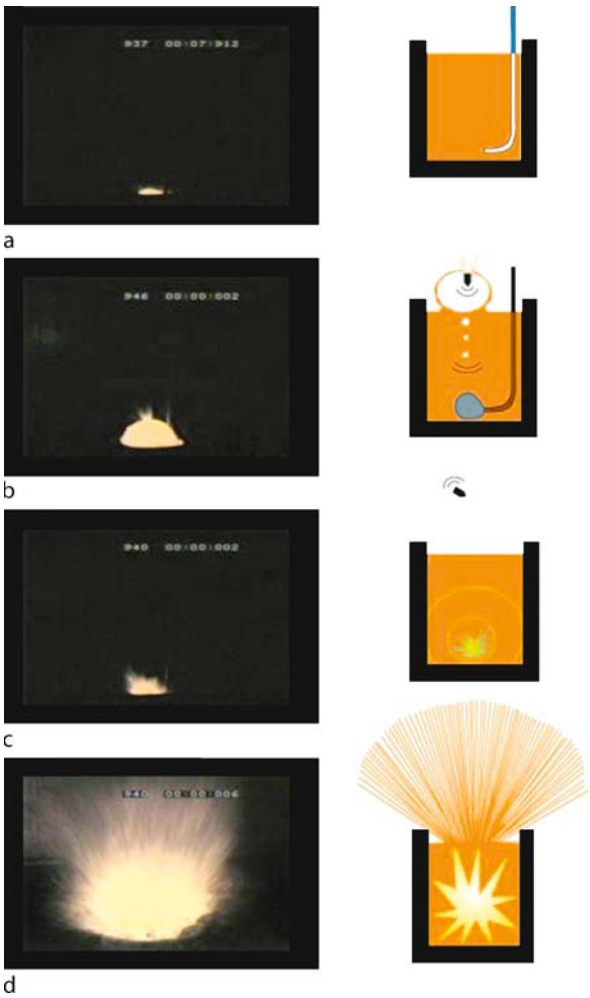

Volcanic Eruptions, Explosive: Experimental Insights, Figure 6
Schematic diagrams and frames from high-speed video illustrate one experimental method of simulating explosive phreatomagmatic eruptions (Grunewald et al. [48]). a Pre-experiment conditions with water injection tube in crucible of hot (1300°C) basalt melt. b 15 ml of water is injected, with the air expelled from the tube forming a bubble. The bubble was impacted by an air gun pellet to trigger (c) the explosive mixing of water and melt. d rapid transition of water from liquid to vapor results in pressure rise. The melt is fragmented by the subsequent rapid expansion. Reprinted with kind permission from Elsevier

Exsolution The nature of geological materials at pressures and temperatures existing within the crust and mantle has been the subject of considerable research by petrologists. Experimental petrology generally studies samples of material in small (millimeter-scale) capsules that are

sealed, then pressurized up to tens of GPa and heated up to 2000 K; easily covering volcanically relevant conditions. Adaptation of this experimental technology to volcanic processes by Hurwitz and Navon [52] was used to study the response of hydrated rhyolite to decompressions up to 135 MPa (equivalent to about 5 km silicate overburden) at eruption temperatures (780°C to 850°C). Extended flow processes cannot be simulated with such small samples at these experimentally challenging conditions, but the pseudo-static response of hydrated silicate melt to pressure decreases of varying rate and magnitude provides important dynamic information about the processes of volatile exsolution (bubble nucleation and growth) and diffusion, as well as crystal growth. Hurwitz and Navon [52] revealed that nucleation of water bubbles (as a supercritical fluid or gas) in natural magmas is likely to be dominated by various heterogeneous processes operating over a range of supersaturations.

Lyakhovsky et al. [76] carried out further analysis of the experimental results of Hurwitz and Navon [52] to reveal the controls on bubble growth following nucleation. Numerical modeling of the experimental data showed that diffusion of water from the melt into the bubbles controlled bubble growth at modest supersaturations in hydrated melts (3–6% w/w water, viscosity 10^4–10^6 Pa s). The viscosity of the melt played a controlling role only when bubble surface area was small in the initial stages (< 1 s) of growth. At high supersaturations, the rapidly increasing area of bubble wall physically moves larger volumes of melt closer to the bubble and volatile advection is a significant contributor to bubble growth under certain conditions. As bubbles grow, the water content of the melt declines and the melt viscosity rapidly increases (Fig. 1a). Bubble growth then again becomes influenced by viscosity. These experimental results indicated that at water concentrations $> 3\%$ w/w in rhyolite, equilibrium degassing was a good approximation. However, as water content declines then equilibrium is not maintained and melt some distance away from a bubble wall, and therefore out of range of significant diffusion, becomes supersaturated in water. Lyakhovsky et al. [76] postulated that these supersaturated melt pockets then encourage additional episodes of bubble nucleation; a process supported by Blower et al. [17] using analogue experimentation and computer modeling. Constrained rapid-decompression experiments provided key insights into the role of continuous nucleation in near-equilibrium degassing of high-viscosity, low-diffusivity melts, with supersaturation limited to that required by heterogeneous nucleation sites.

The role of high viscosities (10^7–10^9 Pa s) on bubble nucleation was studied by Gardner [43], using methods similar to those of Hurwitz and Navon [52]. Heterogeneous and continuous nucleation of water bubbles in a rhyolite with less than 1% v/v microlites of Fe–Ti oxides and plagioclase were again exhibited (Fig. 4). Increasing viscosity and decreasing supersaturation were found to increase bubble nucleation times from a few seconds to greater than 100 s. The controlling parameter for nucleation rate was found not to be temperature, diffusivity or viscosity, but the surface tensions of interfaces between melt, crystal and bubble.

All these experiments were fully sealed, with no opportunity for the exolving volatile phase to escape. This condition was relaxed by Burgisser and Gardner [19] who decompressed 3 mm diameter, 7 mm long cylinders of rhyolite hydrated within gold capsules at 150 MPa and 825°C, with some capsules modified to incorporate a sink for escaping water (pseudo-open conditions). Three bubble growth regimes were identified that depended on balances between diffusivity and viscosity in relation to the rate of sample decompression. The open system experiments showed that, given sufficient time (about 180 s here) water bubbles begin to coalesce once bubble volume fraction exceeds about 43%; note that the permeability of natural samples tends to decline rapidly below about 40% porosity. This is important because bubble coalescence promotes an increase in permeability as preferential degassing pathways form in the melt. The limited volume of the sample, and its confining capsule, prevents the establishment of the range of flow behaviors that exist in volcanic conduits during explosive eruptions. However, the experiments of Burgisser and Gardner [19] provide useful insight into the role of flow on bubble coalescence for this very reason; features seen in the experimental samples are likely to have been difficult to distinguish in any experiments aimed at flow behavior. This demonstrates the usefulness of experiments for investigating spatially and temporally small (but tractable) aspects of a complex bigger picture. The results of such experiments can be used to verify models of the partial-process, with the facility to eventually combine these into a description of the full process.

The link between porosity and permeability was further explored by Takeuchi et al. [119] who isothermally decompressed 5 mm diameter, 15 mm long samples of Usu dacite in gold tubes from 150 MPa at 900°C, to pressures ranging from 50–10 MPa. Samples were then measured for permeability along the tube axis and, with their bubble fraction ranging from 22 to 83% v/v, covered the porosity range over which permeability has been measured in natural samples. However, the experimental samples showed significantly lower permeabilities than natural samples, except for the low porosity data of Eichelberger et al. [42],

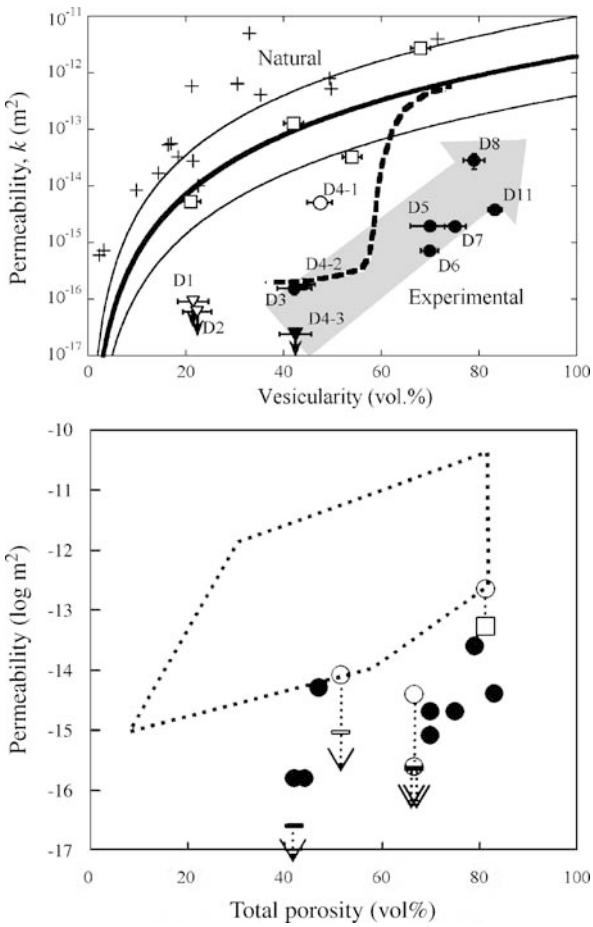

◄ **Volcanic Eruptions, Explosive: Experimental Insights, Figure 7**
As bubble volume fraction (vesicularity) increases, more bubbles coalesce to allow easy gas exchange between them. The result of this is increasing sample permeability with increasing vesicularity. Permeability allows gas to escape from the magma and reduces explosive potential, making permeability an important parameter in understanding the transition from effusive to explosive volcanism. Experiments carried out by Takeuchi et al. [119] (*circles* and *squares* with *error bars*, the *solid* and *open symbols* indicate isotropic and deformed vesicular textures of the products, respectively) are compared with measurement of natural samples in the *top panel* (*dashed curve* indicates the relationship of Eichelberger et al. [42], the *bold solid curve* with two fine satellite lines indicates the relationship of Klug and Cashman [67] with the *upper* and *lower* limit of the scattered data, the *crosses* and *squares* represent permeabilities of Melnik and Sparks [87], and Takeuchi et al. [119] respectively). Permeabilities have been corrected at a later date because of measurement artefact (*bottom panel*, Takeuchi et al. [118]). The trend of increasing permeability with vesicularity is demonstrated, but experiments generally show lower permeability than natural samples. One explanation for this could be that most of the natural samples have been subject to significant flow, whereas the experimental samples were generally not (*open symbols* show some evidence of flow). The role that flow plays in the development of permeability remains unclear. Reprinted with kind permission of AGU

and for experiments where bubbles were elongated along the permeability measurement direction (Fig. 7). This suggests that flow processes act to increase permeability by 1–2 orders of magnitude, and that most of the natural samples measured have undergone significant flow; or as Takeuchi et al. [119] state "The quenched experimental products can be considered as snapshots in vesiculating magma during decompression without deformation of vesicular structure in the late stages of eruption". Describing experiments as snapshots of a larger, more-complex process is an apt analogy. More recently, Takeuchi et al. [118] have revisited their experimental samples and, after removing the gold tubes, found that the sample permeabilities were further reduced (Fig. 7), especially at low vesicularity. The original measurements had included some gas leakage between the silicate sample and the gold tube and, when operating at the limits of permeability detection, the leakage added a significant systematic error to the measurements. The new results further emphasize the importance of the flow history of vesicular materials in the development of permeability, and illustrate the complexity of interactions in the magmatic degassing process that determine the eruptive behavior.

Decompression-Driven Flows At Mt. St. Helens in 1980, a landslide acted to reduce the pressure applied to a body of vesiculated magma. Because the vesicles contained gas, the reduced pressure drove extensive magma expansion. This created the explosive lateral blast that preceded sustained explosive activity on May 18th 1980. Similar behavior may also be exhibited when portions of lava domes collapse. The response of vesiculated magma samples to relatively rapid decompression (10^{-3}–10^{1} GPa s^{-1} (Alidibirov and Dingwell [3])) forms a major area of experimental volcanology.

Laboratory apparatus known as a 'fragmentation bomb' (Fig. 8) was conceived by Alidibirov and Dingwell [1,2] and has been developed by others since. This shock-tube apparatus rapidly decompresses experimental samples at 850°C from up to 35 MPa to 0.1 MPa (Spieler et al. [112]). The samples, cylinders of approximate dimensions of 2 cm diameter by 20 cm length, are natural volcanic rocks heated to the required temperature in an atmosphere of Ar within a pressure vessel. On reaching the temperature of interest, the sample is further pressurized with Argon until a system of calibrated rupture discs fail.

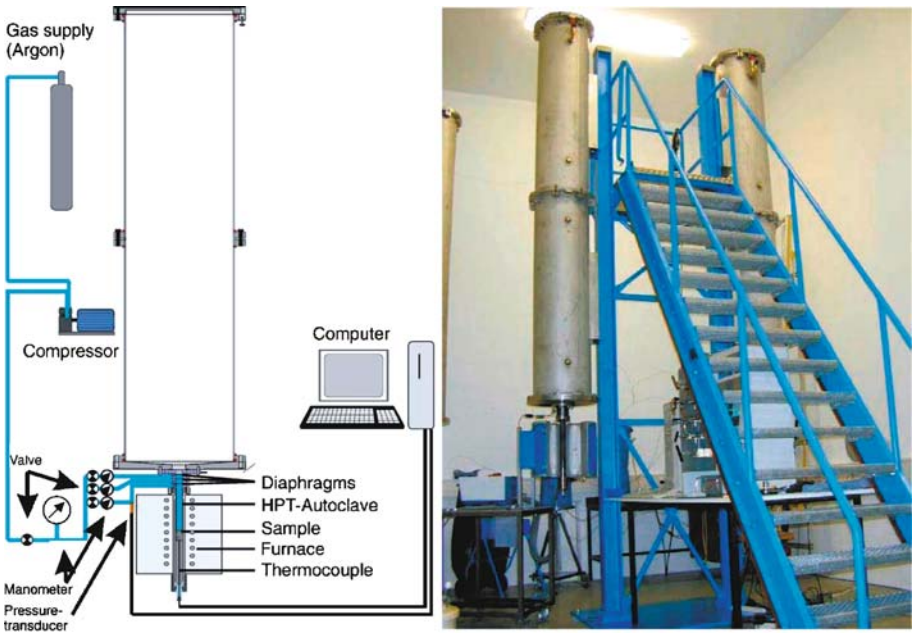

Volcanic Eruptions, Explosive: Experimental Insights, Figure 8
Schematic diagram and photograph of 'fragmentation bomb' shock-tube apparatus (Kueppers et al. [69]). Natural samples are pressurized up to 50 MPa with gas and heated up to 850°C in the autoclave. Samples can be used at room temperature with a transparent experimental section for high-speed imaging (Spieler et al. [111] (Fig. 9); Taddeucci et al. [116] (Fig. 22)). The diaphragms are ruptured to decompress the sample to atmosphere (0.1 MPa) and the resulting flow is collected in the large tank. Pressures, images and fragment morphology then provide information on flow dynamics. Reprinted with kind permission from Elsevier

On failure, the sample rapidly decompresses (1–100 GPa s^{-1}) and any explosive products freely expand into the collecting tank for further analysis. The use of Ar for these experiments removes any additional complexities that may arise from the nucleation, diffusion and growth of water bubbles allowing the detailed study of the mechanical response of vesiculated silicates to a rapid pressure drop.

Alidibirov and Dingwell [3] provide a review of the findings of experiments to that date. Samples of Mt. St. Helens Dacite from the May 18th 1980 eruption were decompressed by up to 18.5 MPa at temperatures up to 950°C. Observation of the resulting fragments showed angular material characteristic of brittle failure of the samples. Calculations of the strain rates imposed on the samples indicated that, even at the highest temperatures used, the dacite would behave as a brittle solid. Alidibirov and Dingwell [3] analyze three mechanisms that could explain explosive fragmentation of the sample. The dacite samples had interconnected pores and were, therefore, permeable. Rapid removal of the Ar gas above the sample creates a pressure gradient within the sample. If permeability if very high then the Ar gas can readily escape and pressure gradients will be relatively low. At the other extreme, if bubbles are not interconnected then pressure gradients within the sample will be maximized (note, the experimental samples must have some permeability in order to pressurize them with Ar, but this does not need to be the case volcanically). In reality, the samples have a degree of permeability that allows pressure gradients to develop within the dacite. The pressure gradient was found to be at its highest at the fragmentation front, where the yield strength of the sample was exceeded. The fragmentation front was found to intrude the sample at tens of m s^{-1}, much slower than the sound speed and known as the fragmentation wave. In this instance, fragmentation can be described as a pressure-gradient driven change in flow regime from permeable filtration flow where only the gas moves, to two-phase flow where gas and solid are both in motion.

Martel et al. [82] synthesised 2 cm long by 2 cm diameter samples of hydrated haplogranitic glass with water contents ranging from 1.4 to 5.7% w/w. Water bubbles were nucleated and grown within the samples under a range of temperatures and pressures within the fragmentation bomb apparatus (Fig. 8) on timescales of about 1 hour, giving vesicularities ranging from zero to 0.91 v/v. Samples that developed large bubbles and high vesicularities underwent significant expansion and tended to develop

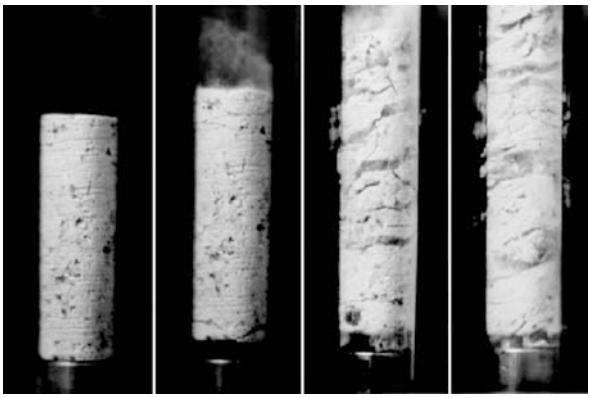

Volcanic Eruptions, Explosive: Experimental Insights, Figure 9
The rapid decompression of a sample of Santorini pumice, at room temperature, by at least 9 MPa (3 times higher than at 850°C) in the 'fragmentation bomb' (Fig. 8) results in spalling fragmentation (Spieler et al. [111]). The *first frame* shows the 50-mm length by 20-mm diameter sample just prior to decompression. 0.25 ms after diaphragm rupture (*second frame*) the ejection of fine dust from the sample surface indicates that pressure is falling in the *upper reaches* of the sample and gas is starting to escape from the pores. The combined elastic response of the sample and apparatus has detached the sample from the holder, fracturing the sample base. 0.5 ms after decompression (*third frame*), spalling fragmentation propagates through the sample perpendicular to the axis of decompression. Fragmentation results from the build up of a pressure gradient within gas in the upper sample sufficient to exceed the tensile strength of the bubbly solid. After 0.75 ms (*fourth frame*) primary fragmentation is complete and fragments are ejected into the large tank (Fig. 8). The timescale of this transient flow is of order 1 ms, but such visualizations give great insight into possible volcanological fragmentation mechanisms. Reproduced with kind permission from Springer Science and Business Media

elongated, or tube, bubbles during this process. Subsequent rapid depressurizations of between 5 and 18 MPa resulted in explosive fragmentation of the pre-foamed samples purely from expansion of the gas phase within the bubbles; timescales were too short for any significant further diffusion and exsolution of water from the glass. The resulting fragments were collected and analyzed for size, shape and vesicularity. The primary fragmentation process was found to be a sequential brittle spalling of the upper sample surface exposed to the decompression and was imaged by Spieler et al. [111] using Santorini pumice at room temperature in a transparent pressure vessel (reproduced here in Fig. 9). Low vesicularities and small decompressions were found to yield larger fragments than high vesicularities and large decompressions, a result consistent with the amount of energy released during the decompression process. The fragment size distributions may be dominated by the primary fragmentation mechanism (surface spalling), but impacts between fragments, and with the apparatus wall, may increase the proportion of the smallest fragments at the expense of the largest, especially with large decompressions (giving high fragment velocities). Interestingly, the presence of tube bubbles acted to decrease fragment size, demonstrating again that flow history is important in determining the products of explosive events.

In a refinement of these experiments, Martel et al. [83] investigated the effect of adding relatively inert alumina particles to the hydrated haplogranitic samples as an analogue to crystals in a silicate melt. Solid particles influenced the final particle size distribution because they tended to remain intact during the experimental fragmentation process. An increasing proportion of alumina particles acted to increase the fragment size for a given decompression, suggesting that energy release was reducing. This would be consistent with the reduced proportion of gas phase that results from an increasing proportion of solids. This effect may also account for the observed increase in the decompression needed to fragment samples as the proportion of crystals increased, with the limit that a zero-vesicularity sample (100% crystals) would not fragment under decompression. Interestingly, tube bubbles were not formed in these experiments due to enhanced heterogeneous bubble nucleation acting to reduce bubble diameter.

Spieler et al. [112] carried out a systematic investigation of the rapid pressure drop required to cause the brittle spalling fragmentation of natural samples with a wide range of chemistry, porosity, permeability and crystallinity. The major control on fragmentation threshold pressure (ΔP_{frag}) was found to be the sample porosity (ϕ). Figure 10 shows the relationship to be approximated by $\Delta P_{\text{frag}} = 1/\phi$, with pressure given in MPa. Such a relationship makes qualitative sense because, as vesicularity tends to zero, ΔP_{frag} will tend to a large value because of little or no gas phase to drive the fragmentation. At large vesicularities, there will be little melt to resist the expansion of the volatile and fragmentation will take place with small decompressions. Figure 10 suggests that the rapid removal of order 100 m of overburden from a vesiculated magma with $\phi > 0.2$ will result in explosive fragmentation caused by pressure-retention in the pores. Such an event could happen by slope failure, for example, at Mt. St. Helens in 1980, or by a dome collapse. At lower vesicularities, the rapid removal of an overburden thickness of order 1000 m is required to initiate explosive fragmentation. However, although smaller overburden removals may not result in immediate spalling fragmentation, slower diffusively driven bubble growth will act to expand the magma and could thus trigger more sustained explosive activity.

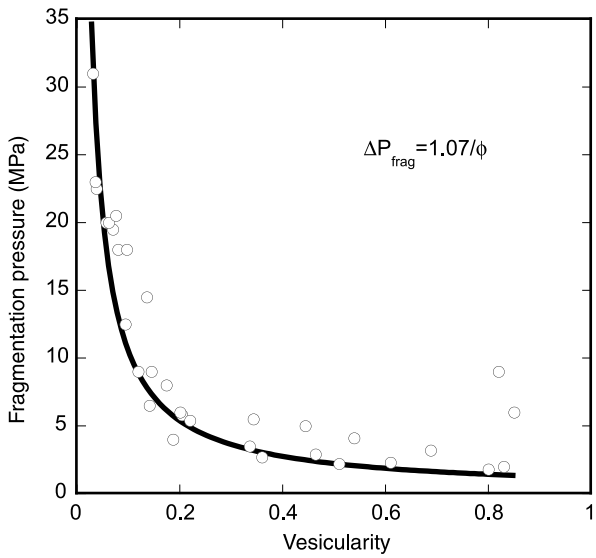

Volcanic Eruptions, Explosive: Experimental Insights, Figure 10
Samples of natural volcanic material with a range of vesicularities were heated to 850°C and decompressed in the 'fragmentation bomb' (Fig. 8) by Spieler et al. [112]. The pressure drop required to cause explosive brittle fragmentation was inversely related to the sample vesicularity, with the constant of proportionality (here, about 1.07 MPa) being the tensile strength of the solid phase. Intriguingly, this suggests that heterogeneity in vesicularity will result in selective fragmentation of more vesicular regions of magma, particularly at vesicularities below 0.3. Data replotted from Spieler et al. [112]

Another mechanism of increasing pore pressure in magma is by crystallization of the silicate melt (e. g., Sparks [108]). This occurs because water is not incorporated into the structure of growing crystals and concentrates in the declining proportion of melt. Taddeucci et al. [115] used the 'fragmentation bomb' to ascertain the fragmentation threshold of crystal-rich magma from the 2001 basalt erupted at Mt. Etna, as a function of porosity. This was combined with models of pressures generated by crystal growth, and calculations of conduit pressures required to eject magma blocks observed during the explosions. This approach supported an explosion model sourced in the heterogeneous fragmentation of crystallizing magma plugs. The particularly interesting aspect of this model is the introduction of porosity heterogeneity within the magma. Figure 10 demonstrates that as porosity increases, the pressure change required for fragmentation decreases. Therefore, the competence of a body of magma will depend on its weakest link, namely any regions of high porosity. Once these begin to fragment, then the less vesicular regions will be removed as large blocks.

Although volcanically accurate materials were used in fragmentation bomb experiments, the effects of reductions in length and timescale by three orders of magnitude remain unknown. Nevertheless, these important experiments give specific insight into key mechanisms that could operate during the rapid decompression (or internal pressurization) of a vesiculated body of magma, and general insight into the behavior of volcanic materials undergoing explosive eruption.

Explosion Products

Explosive expansion of magma, driven by the low bulk modulus of gas-phase volatile species, generates a range of consequent phenomena. The primary magma fragmentation (one mechanism is illustrated in Fig. 6, and another in Fig. 9) involves a large increase in surface area within the volcanic system. The mechanism of this increase is the breaking of molecular bonds.

On exiting the volcanic vent, hot silicate fragments and volatile gas (mainly water vapor on Earth) interact with any atmosphere and cool. In the absence of a significant atmosphere, for example on Io, the volcanic ejecta expands, then the silicate fragments follow ballistic trajectories once they have decoupled from any volcanic gases and cool by radiation. Interaction between silicate fragments is now unlikely and their trajectories end when they impact Io's surface; some may escape Io altogether. On Earth, the presence of an appreciable atmosphere results in different behavior. Explosion products emerge from the vent and interact with the atmosphere. If the silicate fragments are large and the explosive power low then fragment trajectories tend towards ballistic control. Conversely, if eruption power is high and fragments small, then atmospheric interaction may dominate. The mixing of air with hot eruption material and its consequent expansion may result in a thermally buoyant plume that ascends many kilometers into the atmosphere and may circumnavigate the globe by long-range atmospheric transport. If insufficient atmospheric air is entrained in order to develop and maintain buoyancy then some ejecta may fall back to ground to form a pyroclastic flow. On a smaller scale, the presence of electrostatic fields and liquid phases encourages sub-millimeter silicate fragments to aggregate. Many experimental studies have been carried out on the processes that follow explosive volcanic events and we review a selection here.

Fragment Size Distributions The fragment size distribution resulting from explosive volcanic eruption has been found to follow a power law distribution, $N = \lambda r^{-D}$, where N is the number of fragments greater or equal in

size to r, and λ is a constant of proportionality. For a single or primary fragmentation event the exponent (D) is predicted to be less than 3.0. Experiments carried out by Kaminski and Jaupart [66] impacted a sample of Minoan pumice with a solid piston to give an exponent of 2.7. James et al. [57] impacted two pumices together and found an exponent of 3.5. Pyroclastic flow or fall deposits from explosive volcanic eruptions exhibit an exponent range between 2.9 and 3.9.

Martel et al. [82] rapidly decompressed hot vesiculated rhyolite in a 'fragmentation bomb'. Decompressions of less than 15 MPa produced exponents of less than 3.0 for the resulting fragment size distribution, regardless of initial sample vesicularity. However, decompressions of > 15 MPa, for samples of vesicularity > 0.5, resulted in exponents > 3.0 and ranging up to 4.1 (Fig. 11). These results suggest that secondary fragmentation mechanisms, such as those induced by fragment-wall and fragment-fragment collision, could be occurring within the apparatus after the primary spalling-fragmentation event (Fig. 9). Alternatively, as high exponents were observed at high vesicularities, the highly heterogeneous nature of a foamed silicate could invalidate the assumptions made in suggesting 3.0 as the exponent boundary for primary fragmentation, at least for fragment size ranges smaller than the order of bubble size. Interpretation of fragment size distributions between experimental and natural systems in terms of fragmentation processes is further complicated by difficulties in obtaining complete natural samples from explosive events. For example, winnowing in the atmosphere separates fragments as a function of fall velocity, which is a function of size and shape; therefore, the products of any individual fragmentation process are subsequently separated.

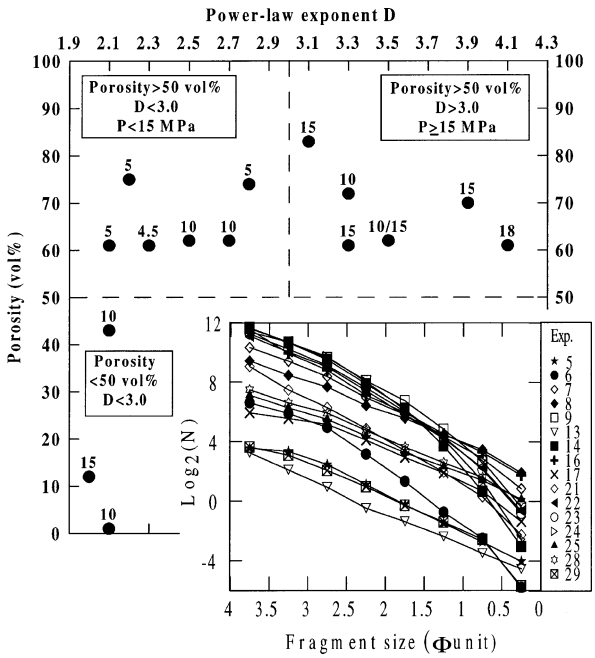

Volcanic Eruptions, Explosive: Experimental Insights, Figure 11
Hot samples of bubbly rhyolite glass with a range of vesicularities were decompressed through various pressures in the 'fragmentation bomb' (Fig. 8), and the fragment size distribution (FSD) measured (Martel et al. [82]). The FSD was compared with a power-law distribution given by $N = \lambda r^{-D}$, where N is the number of fragments with radii larger than r. The power-law exponent, D, was found to increase with the magnitude of decompression (numbers next to *data points* in MPa), as well as with higher vesicularity. Compare this to an exponent of 3.5 for colliding pumices in Fig. 12. Theoretically, $D > 3$ suggests more than one fragmentation process, but these are complex, heterogeneous, multiphase materials and established fragmentation theory based on dense homogeneous materials may not give the full picture. Reprinted with kind permission from Elsevier

Fall Velocity of Silicate Fragments Understanding the behavior of silicate fragments in the atmosphere requires knowledge of the fall velocities of real volcanic pyroclasts, which rarely conform to an idealized spherical shape. Walker et al. [123] dropped samples of pumice from heights up to 30 m and found that terminal fall velocities were best approximated by the fall equation for cylindrical bodies. Wilson and Huang [125] carried out extensive experiments on silicate crystals and glass fragments in the size range 30 to 500 μm. A fragment shape factor ($F = (b + c)/2a$) was defined in terms of the longest (a), intermediate (b) and shortest (c) principal axes of the fragment, allowing definition of the empirical formula $Cd = (24/\mathrm{Re})F^{-0.828} + 2(1.07 - F)^{0.5}$, where the fragment drag coefficient is given by Cd, as a function of Reynolds number, Re, and the shape factor. Suzuki [114] revisited the experimental data and proposed a modified equation, $Cd = (24/\mathrm{Re})F^{-0.32} + 2(1.07 - F)^{0.5}$. This empirical, experimentally informed approach allowed calculation of the fall velocity of a fragment in the atmosphere and provided input into numerical models of the dispersal of products from explosive volcanism. Further information can be found in [101,109,120].

Fragment Electrification James et al. [56] investigated possible mechanisms by which silicate fragments could become electrically charged during explosive activity. Two natural pumices were collided together resulting in the generation of particles generally smaller than 70 μm in diameter as the silicate foam fragmented in brittle fashion. Fracto-emission generated ions and charged silicate frag-

ments. Ions were found to have one net charge and silicate fragments the opposite net charge. Experiments carried out at atmospheric pressure produced charge densities similar to those found on ash particles resulting from explosive terrestrial volcanism, whilst those carried out at 0.1 Pa sustained at least an order of magnitude more charge. The absolute amount of charge was dependent on the impact energy, (a result comparable to [23] findings of increasing electric field with greater fragmentation energy), with the net charge representing a slight imbalance between positively and negatively charged fragments. The mechanism of producing charged silicate fragments from pumice-pumice collision adequately accounts for the fragment charges and electric fields measured in proximity to explosive volcanic eruptions. This suggests that pumice-pumice collision, i. e., secondary fragmentation, could be a major source mechanism for sub-100 μm silicate fragments.

Experimental hydrovolcanic explosions were found to produce an electrically charged fragment and ion cloud [24]. Interestingly, fragmentation using high-pressure Argon gas produced electrical effects of smaller magnitude than magma-water interaction, even though expansion rates of the ejecta from the experimental crucible were similar. This suggests that the cooling effect of water does produce much more brittle failure resulting in enhanced fracto-emission [37,40], consistent with the production of smaller blocky fragments from MFCI experiments [135]. It is also possible that polar water molecules dissociated into reactive ions during the MFCI process, a mechanism not so accessible to Argon atoms. Büttner et al. [24] found striking similarities between experimental electric fields and those generated during explosive events at Stromboli volcano.

Silicate Fragments and Their Aggregation Accretionary lapilli are widespread in the deposits from explosive volcanic eruptions, and anomalous variation in the deposit thickness and grain size distribution as a function of distance from the explosive vent is commonly observed. Both of these processes are explainable if silicate fragments smaller that about 100 μm clump together, or aggregate, to form a larger entity which has different aerodynamic properties than its constituent particles. Sparks et al. (see Chapter 16 in [109]) provides a review of particle aggregation in volcanic plumes prior to 1997.

Accretionary lapilli are obvious in the geologic record as millimeter to centimeter sized spheroidal aggregates (density range 1200 to 1600 kg m^{-3}) of silicate fragments and secondary minerals such as gypsum, and they formed the early focus of study into aggregation in volcanic plumes. Two complementary experimental studies provided insight into the formation mechanisms of accretionary lapilli. Gilbert and Lane [45] focused on formation in atmospheric plumes. A wind tunnel was used to demonstrate experimentally that particles of volcanic ash collide and adhere to the surfaces of objects falling through the atmosphere that are covered with a thin layer of liquid. The liquid layer was found to act as a control on the sizes of silicate fragments that adhere, with high sticking efficiencies at 10 μm diameter, reducing by two orders of magnitude for fragments 100 μm in diameter. A key experimental finding was the importance of hygroscopic chemical species in maintaining thin liquid films. Experimentally this was achieved using NaCl; volcanically, sulphuric acid is the most likely hygroscopic compound that could maintain liquid films at low relative humidity and down to temperatures of −70°C. These experiments strongly suggest that accretionary lapilli are diagnostic of a three-phase environment within the volcanic plume in which they formed, supporting earlier hypotheses (e. g., [88]).

Schumacher and Schmincke [105] focused on accretionary lapilli formation in pyroclastic flow processes. Experiments were carried out providing insight into interpreting the structure of accretionary lapilli in terms of their formation mechanisms. Accretionary lapilli were synthesized by a method similar to the industrial production of fertilizer pellets, with mixtures of volcanic ash and variable proportions of water being rotated in a steel pan. The resulting spheroidal agglomerates varied with the proportion of water present, formation being optimal at between 15 and 25% w/w water, and grain size sorting was found to occur at low water mass fractions The mechanism of formation, namely capillary binding, is identical to that of Gilbert and Lane [45], but under conditions of much higher fragment number-concentration.

The occurrence of secondary thickness maxima and bimodality of grain size distribution, exemplified by Carey and Sigurdsson [27], implies the mechanism of silicate fragments being transported within aggregates. These aggregates could be accretionary lapilli, but these are often absent from the deposits suggesting that the aggregates are fragile and lose their identities once incorporated into a volcanic deposit. Fragile aggregates have been observed falling from volcanic plumes and one explanation for their formation is the presence of electrostatic charge. During fragmentation, silicate particles become electrostatically charged (see Sect. "Fragment Electrification"), and whilst in the atmosphere these charges have no path to escape to earth unless electric fields exceed the level required for electrostatic discharge in air. This leads to long-lived attractive electrostatic forces between silicate

fragments [81], and consequently the formation of aggregates. Once on the ground an electrically conductive path to earth can be established, especially in the presence of moisture, and the aggregates disintegrate.

Schumacher and Schmincke [105] carried out experiments using natural volcanic ash and a stream of ions from an electrostatic paint gun. Silicate fragments smaller that about 180 μm diameter were found to rapidly aggregate into loose fragile clusters demonstrating the viability of the electrostatic aggregation process for fine volcanic particles.

James et al. [57] studied the generation of aggregates resulting from the interaction of ions and charged silicate fragments using the fragment generation and charging mechanism of James et al. [56] with pumice from the May 1980, Mount St. Helens fall deposit. Silicate fragments were allowed to fall and interact over a distance of 0.5 to 1.1 m onto a variety of analysis platforms. Fallout was collected and analyzed for fragment size (Fig. 12) using an agitated dispersion of the particles in a conducting electrolyte to maintain disaggregation. The particle size distribution showed power law behavior with exponent 3.5 (Fig. 12), somewhat above the theoretical maximum of 3.0 for a primary fragmentation event. These experiments indicated that the aggregation process happens shortly after the fragmentation event has generated charged silicate fragments and ions (Fig. 12). The aggregation process significantly changes the size distribution relevant to aerodynamic behavior (Fig. 12) from the single-fragment form $N = \lambda r^{-3.5}$ to aggregate form $N = 15\,089 \exp[-0.011 d']$, where d' is the aggregate diameter. Experimental measurement of aggregate fall velocities enabled estimation of aggregate densities between 80 and 200 kg m^{-3}, at least an order of magnitude less than both the fragments that comprise aggregates and accretionary lapilli. The main consequence of electrostatic aggregation for sub-100 μm silicate fragments is to reduce atmospheric residence time because aggregation acts to increase fall velocity above the single fragment value. The size distribution of individual fragments within an aggregate was experimentally examined by James et al. [58], who found empirically that for aggregate diameters less than 140 μm, $n = 0.000272 d_a^3 [\exp(-0.22 d_p) + 0.008 \exp y(-0.083 d_p)$, where n is the number of fragments larger than diameter d_p within an aggregate of diameter d_a. As electrostatic aggregates exceed about 140 μm in diameter the size distribution of their constituent fragments was found to change. This was attributed to the aggregate growth mechanism becoming dominated by aggregate-aggregate interaction rather than aggregate-fragment interaction (Fig. 12), and stabilization of the size distribution of material being incorporated into the aggregate. Electrostatic aggregates can be represented as spheres of density about 200 kg m^{-3} and an experimentally determined empirical relationship between drag coefficient (Cd) and Reynolds number (Re) for aggregate diameter range 50 to 500 μm is $Cd_{\mathrm{agg}} = 23\mathrm{Re}^{-0.637}$ (James et al. [58]). Such experimental data can then be used within numerical models of the transport of sub-100 μm volcanic ejecta in the atmosphere, although at the time of writing this has yet to be undertaken.

Pyroclastic Flows Pyroclastic flows comprise a hot mixture of silicate particles and gas formed by the collapse of negatively buoyant regions of a volcanic plume or avalanching of material from a growing lava dome. By nature, pyroclastic flows are destructive to infrastructure and almost invariably fatal to plants and animals. Insight into the behavior of pyroclastic flows may be gained by studying their deposits. However, it is difficult to definitively identify fluid dynamic processes operating during a pyroclastic flow from the complex structures within deposits that are revealed once motion has ceased. Here, experiments can help in isolating flow processes and their subsequent effect on deposits. However, scaling between experiment and natural process requires care to ensure behavioral similarity as particles of a range of sizes interact with the gas phase to different degrees.

Dellino et al. [36] designed experiments to investigate the mechanics of pyroclastic flows on a scale comparable to the natural process. In order to approach this, the physical scale of the experiments was such that they required an outdoor laboratory. The experimental material was sourced from a natural pyroclastic flow deposit in order to mimic the interaction between fragments and gas as closely as possible. Varying masses of this material were packed into an experimental tube 60 cm in diameter and 2.2 m high. 14 liters of gas at pressures over 10^7 Pa was rapidly injected into the base of the tube to eject the experimental material, with experiments triggered and logged under computer control. The experiments were carried out at ambient temperature and were, therefore, not scaled for any thermal effects; this has the benefit of reducing both equipment and process complexity. Ejected material could be sampled for analysis and compared to natural pyroclastic deposits. Preliminary experiments showed two main types of behavior as the specific mechanical energy (SME, given by *gas pressure* × *gas volume* ÷ *mass of pyroclastic material*) of the system was changed (Fig. 13). At high SME values a dilute particle plume develops with the particles settling out individually. At low SME values the ejected material forms a collapsing column that generates a particle flow similar to a pyroclastic flow. Intermediate

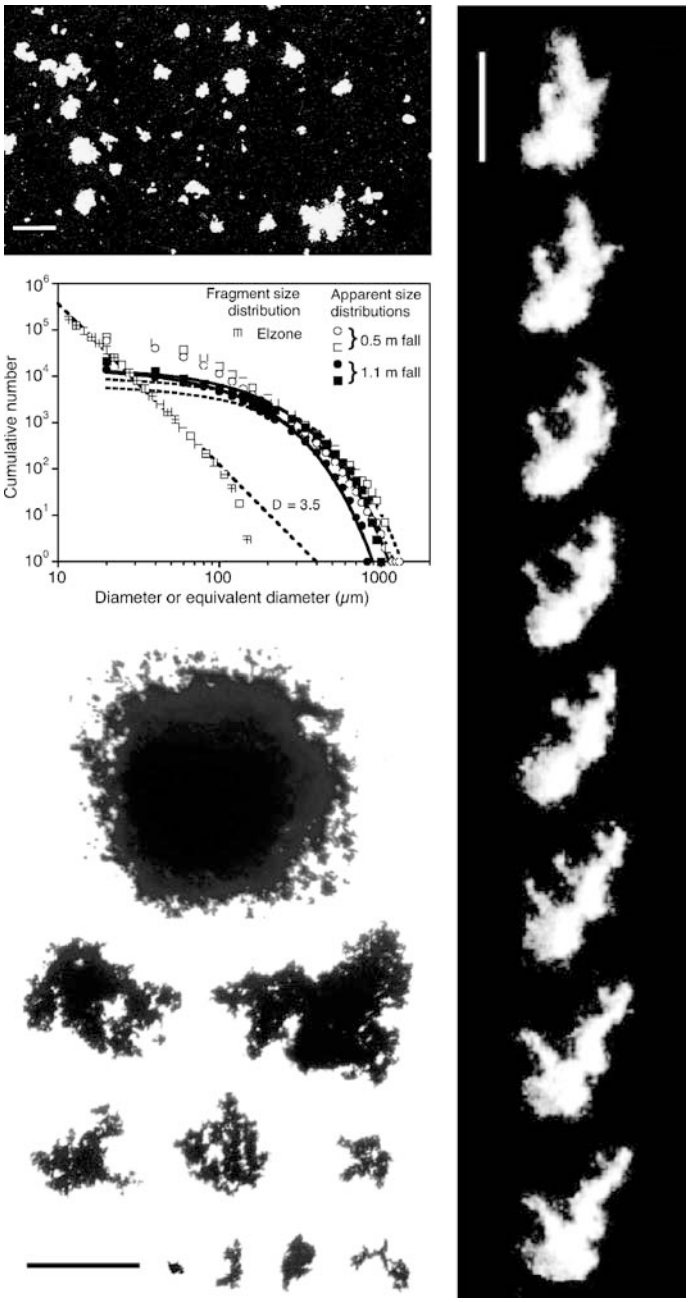

Volcanic Eruptions, Explosive: Experimental Insights, Figure 12

Fragmentation during explosive eruptions produces electrically charged fragments and ions by fracto-emission processes (James et al. [57,58]). Electrostatic forces are significant in comparison to gravity for fragments less than about 100 μm in size, resulting in fragment aggregation. Aggregates have different aerodynamic properties than their constituent fragments, thus altering the fallout behavior of sub-100 μm volcanic ash and consequently its atmospheric transport. The top image shows experimental aggregates collected on a plate (scale 1 mm), with the graph showing the size distribution of the aggregates (*curved lines*), as well as that of the constituent fragments (*straight line*). The *bottom panel* (scale 0.5 mm) shows that experimental aggregates of different sizes have different morphologies, and the *right-hand panel* (scale 0.5 mm) shows strobe images of a falling and rotating experimental aggregate. Reprinted with kind permission of AGU

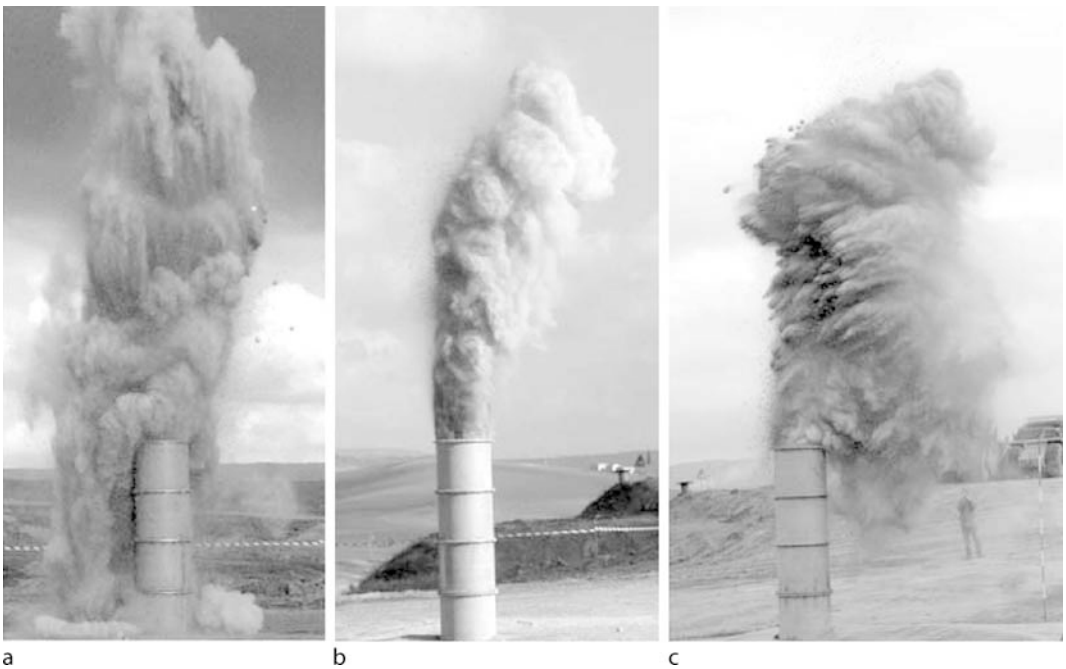

Volcanic Eruptions, Explosive: Experimental Insights, Figure 13

Ejection of natural pyroclastic material by a burst of gas can be used to study both volcanic plumes and pyroclastic flows (Dellino et al. [36]), this also looks enormous fun! a shows an experiment with low specific mechanical energy (SME) generating a collapsing column and (pyro)clastic flow. c shows a high SME experiment generates a plume that, whilst not buoyant, disperses on the prevailing wind, and b shows the transitional case. Such large-scale investigation helps to overcome some of the scaling difficulties imposed by smaller-scale laboratory experiments. Reprinted with kind permission of AGU

SME values give transitional behavior with aspects of both end-members. These experimental observations have considerable phenomenological similarity to natural explosive eruptions where particle-dilute regions ascend in the atmosphere and more particle-concentrated regions fall under gravity to give pyroclastic flows. Such similarity gives added confidence that mechanisms operating in the natural process also operate in the simulation.

Experiments producing a collapsing-column deposited material at the point of impact of the column with the ground. The resulting deposit was found to be structureless, just as many natural proximal pyroclastic deposits, because grain-sorting processes did not operate in this concentrated particle flow. Other similarities included the formation of a concentrated undercurrent, bed load at the base of the turbulent flow, and continuous atmospheric suspension of small silicate particles; this is very interesting in the absence of thermal processes. These preliminary experiments demonstrate great potential for linking processes in scaled experimental flows with the resulting experimental deposits and, therefore, providing a powerful interpretation tool for natural pyroclastic flows.

Experiments carried out at smaller scale, but which included heating of the volcanic ash [41], have also investigated gas retention in pyroclastic flows. A bed of pyroclastic flow material with a wide fragment size range was fluidized by the drag force of rising hot gas. Both the bed expansion and collapse (degassing) after the gas supply was cut were studied. During expansion, smaller fragment size, lower density and higher temperature all promoted more uniform and easier expansion of the fluidizing bed. Figure 14 illustrates the nature of flow patterns that develop. Collapse experiments were imaged using X-rays and carried out between 20°C and 550°C giving temperatures comparable to those in natural flows. The generic collapse process is illustrated in Fig. 14. Experiments and numerical models identified that an aerated bed loses its gas by permeable flow on a diffusive timescale. A fluidized and uniformly expanded bed was controlled by timescales of hindered settling and diffusive degassing of the resulting sediment layer. The relative magnitudes of these timescales depended on the thickness of the degassing bed, with settling dominating in thin beds and diffusion in thick ones. These experiments indicate that hot pyroclas-

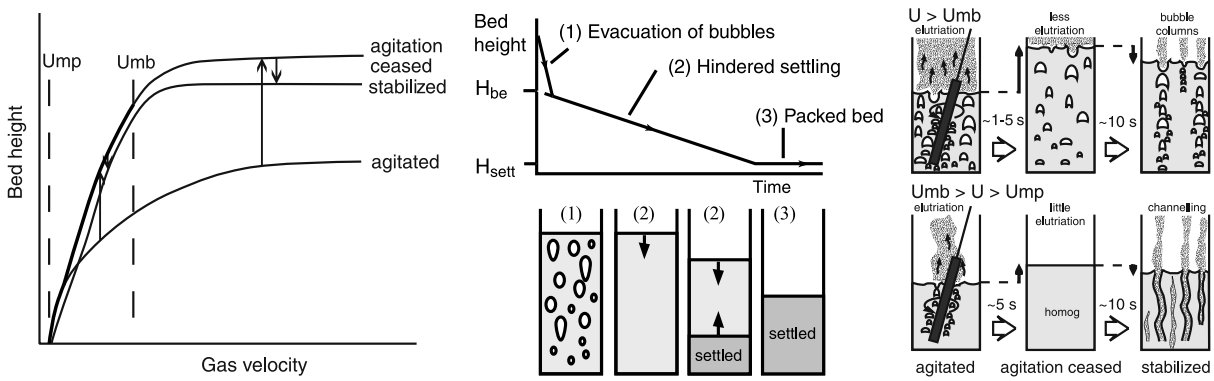

Volcanic Eruptions, Explosive: Experimental Insights, Figure 14

The inflation and deflation of a bed of natural pyroclastic material using hot gas shows a range of flow structures (Druitt et al. [41]). During inflation, the bed can be homogeneous or develop gas bubbles or channels, and fine particles can be removed by elutriation. When fluidizing gas velocity exceeds Ump (*left panel*), the bed inflates and particles start to segregate on the basis of size and density. Bubbling occurs when gas velocity exceeds Umb (*upper right panel*), and the bed expands only weakly with increasing gas velocity (*left panel*). Agitation of the flow acts to reduce the gas volume fraction. Once the gas supply is closed off then bubble collapse results in rapid deflation ((1) in *middle panel*). An extended period of hindered settling then generates a layer of settled material at the base of the bed ((2–3) in *middle panel*). These experiments were used to investigate the gas retention properties of pyroclastic materials as a function of fragment size and temperature. Gas retention was favored by finer fragment sizes, with high temperature having a secondary retention effect. Hot, fine-grained pyroclastic flows are, therefore, predicted to flow further. Reproduced with kind permission from Springer Science and Business Media

tic flows with small fragment sizes retain gas longer than cool, large fragment flows, with temperature having a secondary effect. The gas retention translates into the distance a flow can travel and these experimental results are consistent with observations of natural pyroclastic flows.

Analogue Approach

Experiments with volcanic materials can suffer constraints placed by the high temperatures and pressures at which the processes of interest occur. If similar processes can be studied under more amenable environments, then the practical limitations are relaxed and, generally, fluid volumes can be increased and a greater range of transducers are available for measurement. This can be achieved by substituting natural fluids with analogues that can be studied at near room temperatures and pressures. However, one difficulty with this approach is in ensuring the relevance of analogue results to volcanic systems and, perversely, natural volcanic materials are now often characterized more fully than many of the analogue fluids used.

Explosive Processes

Direct observation of the motion of magma within volcanic conduits would reveal the how and why of volcanic explosive activity. However, direct observation is considered near-impossible and indirect observation open to interpretation. Experiments with natural materials provide constraints on physico-chemical processes within conduits, but extending the experimental parameter space by using analogue materials provides greater insight and enables testing of numerical models.

High VEI Events Volcanic eruptions of high VEI involve the transition of a low-velocity, high-density, high-viscosity bubbly magma, for which the liquid phase is considered continuous, to a high-velocity, low-density, low-viscosity 'dusty gas' where the gas phase is considered continuous. This fragmentation process represents the point at which the liquid and gas phases start to separate. Studies designed to investigate this process in general, but with volcanic application in mind, form a major part of experimental research using the analogue approach.

As pressure declines, the exsolution and expansion of volatile (specifically water) from magma provides the driving mechanism for most explosive events. In order to study similar behavior in analogue systems the shock tube has been almost universally adopted, as it has for the investigation of fragmentation in natural materials (e. g., [1,2]). Such experiments became of great importance around WW2 when understanding of shock waves and flame fronts from explosions, and supersonic flight, became of interest to the fluid dynamic community; volcanological application of shock tubes emerged from this background. Mader [77] provides a review of early shock tube experimentation relating to flow within volcanic con-

duits, we briefly summarize these here, but focus on more recent work.

Post-Fragmentation Flows The decompression of beds of small, incompressible particles with compressible, gas-filled pore space can give insight into the development of two-phase flows within and above the conduit subsequent to primary fragmentation [6]. The resulting rapidly expanding, but transient flows show development of significant heterogeneity, which declines as expansion reduces particle concentration. The development of heterogeneity within expanding particle-laden gas suggests mechanisms for the formation of pyroclastic flows from high-density sections of eruption columns, whist lower-density sections become buoyant and form an eruption plume. Using smaller particles, Cagnoli et al. [25] carried out similar experiments to explore the dynamics of short-lived Vulcanian explosions. In similarity (we think, but see Fig. 9) with post-fragmentation volcanic flows, the experimental tube was much larger than the particle size. As in Anilkumar et al. [6], these transient flows demonstrated heterogeneous distribution of particles as the mixture expanded and the gas separated from the particles (Fig. 15). These features include sub-horizontal gas-rich regions that invite comparison with those found by Spieler et al. [111] in decompressed pumice samples; the implication being that the primary fragmentation process may be incidental to the nature of the evolving flow. Heterogeneities also comprise gas-rich bubbles, sometimes approaching tube diameter in size, and more elongated towards the flow front. Streamers of particles also emerge from the flow front indicating inhomogeneity in the distribution of gas escaping the particle flow (Fig. 15). As the flow ages and velocities increase, turbulence acts to smear out heterogeneities. The source of the heterogeneities in such transient flows is unclear. They may represent the expansion of gas rich regions in the original packed bed, suggesting that the transient flow reflects heterogeneity in the source material. Variability in permeability, and the dynamics of initial particle acceleration may also act to amplify heterogeneity as the flow expands. These experiments demonstrate the complex nature of even 'simple' two-phase flows that are undergoing rapid changes in variables such as pressure and velocity. The volcanic implication is that flow within a volcanic conduit can be highly heterogeneous and that modeling such flows as having smoothly changing density, as a function of time or space, will be far from representing the real richness of the flow physics. Such modeling may give accurate prediction of time-averaged parameters, but will not reflect variability. However, it could be the variability that ultimately results in hazards like pyroclastic flows.

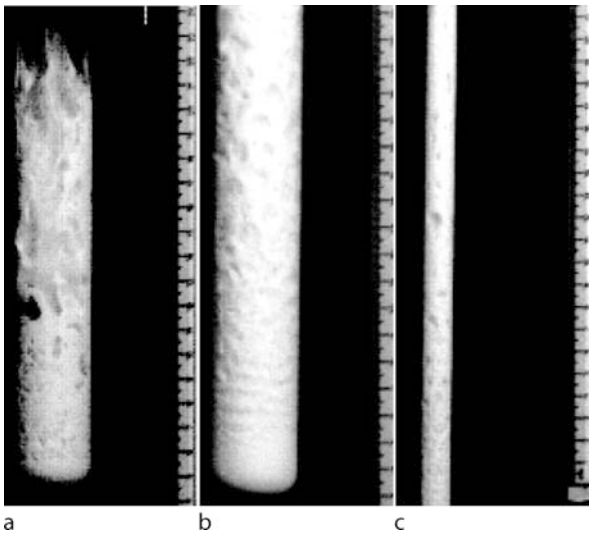

Volcanic Eruptions, Explosive: Experimental Insights, Figure 15
Rapid 20 to 90 kPa decompression of beds of sub-100 µm glass beads generates a particle flow as the gas phase expands in response to the pressure drop (Cagnoli et al. [25]). a shows the decompression of 100 g of beads (with average diameter of 38 µm through 51 kPa) after 28 ms. Note the wispy *flow front*, indicating escape of gas, and the development of large gas bubbles within the flow (scale shows 1 cm and 0.5 cm gradations). b (150 g of 95 µm beads, 69 kPa decompression, after 23 ms) shows the development of sub-horizontal particle-poor regions, giving some phenomenological similarity to spalling fragmentation (Fig. 9). c shows flow in a narrower tube (100 g of 38 µm beads, 68 kPa decompression, after 20 ms) where large particle-poor regions develop at the flow margin. Reprinted with kind permission from Elsevier

Chojnicki et al. [29] demonstrated that models of explosive volcanic flows based on pseudogas approximations underestimated the initial shock wave strength and velocity. Pseudogas models, and others based on steady-state experiments, then overestimated the subsequent particle bed expansion rate when applied to laboratory scale experiments. Such discrepancies suggest that one or both of the approaches, numerical and experimental, requires modification to reconcile the differences in order to understand the natural phenomena. In this case, Chojnicki et al. [29] identify processes not accounted for in the numerical approach, but which became apparent from experimental observation. Existing theory was then empirically modified to fit the unsteady experimental flow. Volcanically, these experiments suggest that the pressures present in conduits prior to Vulcanian explosive events were underestimated by a factor of approximately five when calculated using established pseudogas and inviscid shock theories. The pressures calculated from equations based

on volcanically relevant experiments are consistent with the overpressures associated with Vulcanian explosions, as well as the rupture strength of volcanic rock. This demonstrates the importance of experimental design to progressing understanding of complex volcanic phenomena, and using the results of these experiments to test, modify and develop the numerical approach.

Decompression-Driven Flows The rapid decompression of natural vesicular solids causes fragmentation if the pressure drop is large enough for pressure gradients within the sample to exceed the tensile strength (Fig. 9). Such transient experiments have also been carried out on analogue solids in order to further characterize the behavior of this system of flows. Decompression of synthetic organic resin, with about 90% porosity and significant permeability (Alidibirov and Panov [4]), resulted in fragmentation phenomenologically similar to that of natural samples (Figs. 9, 16) undergoing brittle failure. A fragmentation front or wave was observed to propagate through the sample creating spalling-type fragmentation. Fragment size was observed to decrease, and exit velocity increase as the magnitude of decompression increases from 0.1 to 0.8 MPa. The propagation velocity of the observed fragmentation wave was less than sound speed in either the solid resin or the gas in the pores. Combining this observation with the dependence of fragment size on pressure drop suggests that fragmentation occurs as gas escapes from the decompressed foam surface, creating a pressure gradient within the foam. When this pressure gradient is sufficiently steep, the foam fails revealing the next gas-escape surface. The expanding gas then accelerates the foam fragments. This implies that fragment size will also depend on sample permeability; an impermeable sample fragmenting on a scale comparable with bubble size and at small decompressions, whilst a permeable sample requires large decompressions and fragments on a scale much larger than bubble size. Very permeable materials will not fragment because the gas can escape without applying sufficient tensile force to the foam. Variability in permeability and porosity may well be reflected in fragment size for any decompression event.

Ichihara et al. [53] depressurized a vesiculated silicone compound and studied the expansion of vesicular materials across the viscoelastic transition by varying the decompression rate (Fig. 17). Fragmentation occurred at decompression rates $> 2.7\,\mathrm{MPa\ s^{-1}}$, with only sample expansion occurring at $< 2.7\,\mathrm{MPa\ s^{-1}}$. At high decompression rates, fragmentation occurred before expansion of the fragments. These experiments showed that both the fragmentation threshold and the nature of the fragmentation

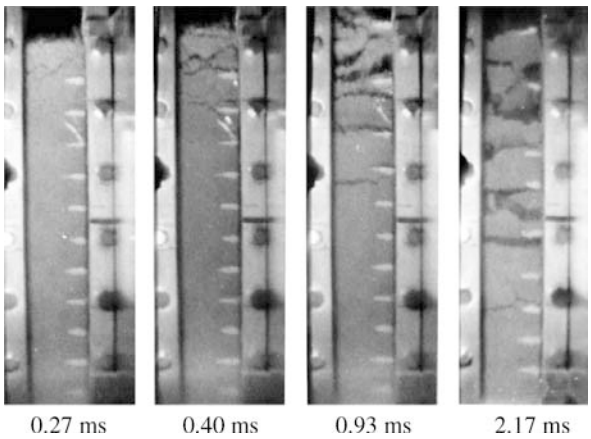

Volcanic Eruptions, Explosive: Experimental Insights, Figure 16
An 0.8-MPa rapid decompression of vesiculated solid foam (Plastiprin) shows very similar spalling fragmentation behavior to natural vesiculated magma (compare with Fig. 9). This demonstrates the usefulness of analogue experimentation in identifying the universality (or not) of flow behaviors (Alidibirov and Panov [4]). Reproduced with kind permission from Springer Science and Business Media

process depend not on the magnitude of decompression (above a minimum), but on the rate of pressure change. Furthermore, the rate of decompression needed to exceed the critical value for more than about 0.1 s [53], indicating that a pressure change approaching 300 kPa represents the minimum pressure drop for fragmentation and also relates to the tensile strength of the experimental material. This fragmentation timescale was similar to the structural relaxation time for the glass transition (about 0.3 s), suggesting that crossing the relaxation timescale results in brittle failure of the sample. Coincidentally, the timescale for viscous control of bubble growth in the experimental material is also in the region of 0.1 s, suggesting that bubble expansion may determine the fragmentation timescale. This coincidence occurs because the rigidity of the silicone fluid is within an order of magnitude of the experimental pressure; in magmas the difference is in the region of three orders of magnitude. The nature of the fracture surfaces (Fig. 17, compare with Figs. 9 and 16), and the observation of minimal expansion before fragmentation suggests, however, that fragmentation occurred because instantaneous strain rate exceeded that required for brittle failure.

Using viscoelastic and shear-thinning solutions of 0.1%, 0.3% and 0.5% w/w polysaccharide (xanthan gum) in water, transient flow experiments exploring a wider parameter range, including variable porosity, were undertaken by Namiki and Manga [93]. Air bubbles were introduced using a hand-mixer, giving gas volume fractions

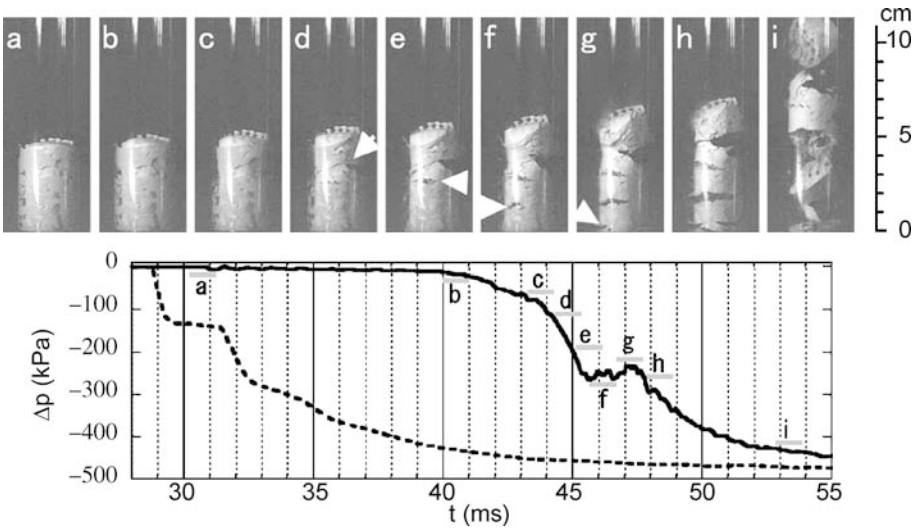

Volcanic Eruptions, Explosive: Experimental Insights, Figure 17
Rapid (50 MPa s^{-1}) 0.5-MPa decompression of a vesiculated dilatant (*shear-thickening*) silicone compound that acts as a brittle solid at high strain rates ($> 3s^{-1}$), but a viscous liquid at low strain rates. Pressure above the sample is given by the *dashed line*, whilst that below by the *solid line*. The surface of the sample is decompressing in (a), but pressure beneath the sample remains steady until (b), decreasing more rapidly from (c). Brittle failure of the sample occurs over a < 10 ms time span (d–g) in a spalling-fragmentation manner (Ichihara et al. [53]), about 40 ms after initial decompression. Fractures appear to initiate at the tube wall and then open to expand the flow (h, i). There is little or no expansion of the vesiculated sample or fragments. Reprinted with kind permission of AGU

up to 0.79. Samples were rapidly decompressed from atmospheric pressure to pressures ranging from 0.8 to 0.06 of an atmosphere; pressures much lower than this would encounter the saturated vapor pressure of water, causing the flows to be influenced by the production of water vapor. The shock tube was 0.05 m diameter by 0.25 m length, with fluid added to a depth of about 0.05 m. A range of flow patterns was observed (Fig. 18) as system parameters were varied. These include: fragmentation, partial rupture, detachment, deformation, and nothing; only 'fragmentation' results in ejection of fluid from the shock tube. Large decompression of highly vesicular samples leads to fragmenting flows (Fig. 18), which initially expand rapidly, then fragment layer by layer (Fig. 19) showing some similarity to spalling fragmentation (Figs. 9, 16, 17). In 'partial rupture', large-scale net-like structures develop that allow gas to escape from the fluid, but the liquid phase does not fragment *en masse*. At lower flow energies, the liquid phase detaches and contracts from the tube wall to yield an inverse annular, or 'detached' flow. Transient experiments allow the measurement of flow-front position, thus enabling estimation of strain rates within the flow. In these experiments the flow expansion was found to be independent of viscosity, which ranged from about 30 Pa s to 0.01 Pa s and was strain-rate dependent. For all experiments, it was found that shear rates were sufficient for elastic behavior, consistent with the independence from viscosity.

The changes in flow pattern are attributable to the behavior of the bubble walls. Detachment of the fluid from the tube walls can be explained by the Poisson's ratio of the axially expanding foam, and represents the development of radial heterogeneity within an axial flow. However, if bubble walls rupture, elastic energy is released resulting in radial expansion and, thus, suppression of detachment. This suggests that fragmentation cannot take place from the detached flow pattern. Failure of bubble walls also increases permeability and allows gas to escape. Failure of the thin film, or plateau, between two bubbles, but not of the thicker plateau borders between three or more bubbles, is considered to result in the partial-rupture flow pattern. The plateau borders form the net-like structure. Wholesale failure of plateau and plateau borders results in fragmentation. Application of experimental processes to volcanic scales (Fig. 18) by utilizing results from studies of transient flows of natural materials, gives indication of the decompressions needed to fragment vesicular magma for a range of vesicularities, but defining other flow patterns at volcanic scale requires further research.

Using a similar technique to Ichihara et al. [53], Namiki and Manga [94] investigated the role of transient decompression rate on aqueous xanthan gum so-

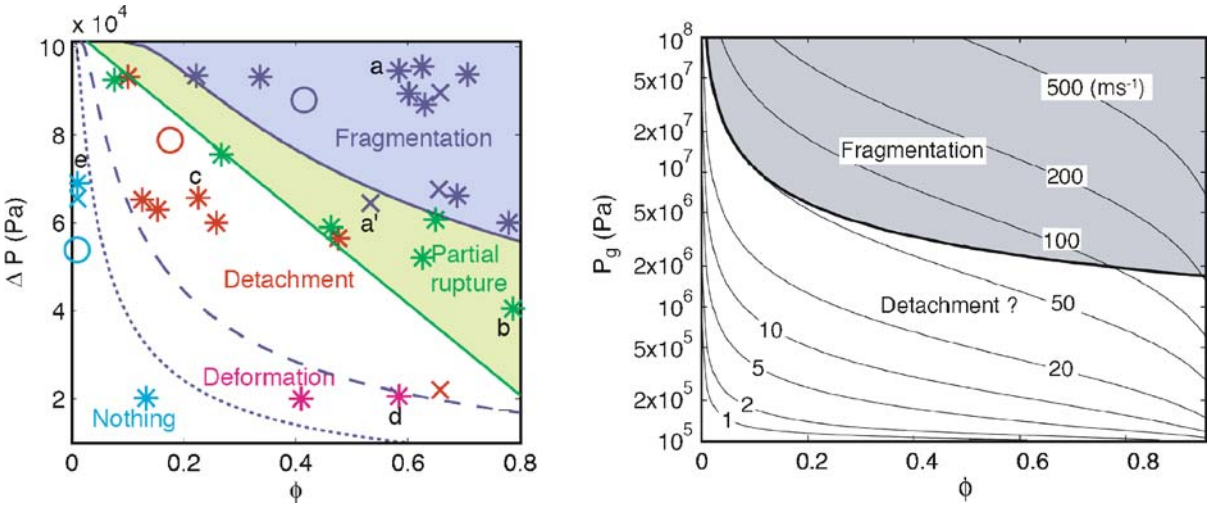

Volcanic Eruptions, Explosive: Experimental Insights, Figure 18

Aqueous xanthan gum solutions are shear thinning and develop brittle solid behavior at strain rates above about 0.1 to 1 s^{-1}, depending on concentration. The *left panel* shows that rapid decompressions (< 0.1 MPa) of pre-vesiculated xanthan gum solution results in a range of flow patterns (Namiki and Manga [93]). Combining data established from the decompression of natural samples with the analogue xanthan gum experiments allows the calculated fragmentation field for explosive volcanic eruptions (*right panel*) in terms of gas pressure and vesicularity. It was not possible to define the parameter space for other experimental flow patterns applied to magmatic systems. Reprinted with kind permission from Elsevier

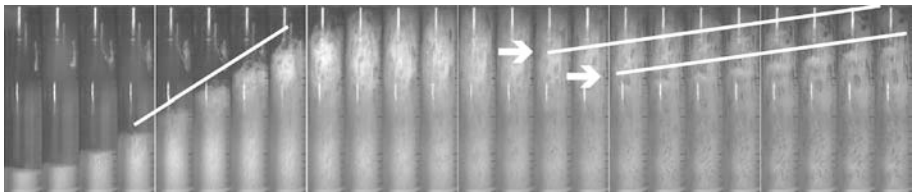

Volcanic Eruptions, Explosive: Experimental Insights, Figure 19

Fragmentation of vesiculated xanthan gum solutions occurs if decompression and vesicularity are high enough (Fig. 18). The *first frame* (*zero time*) shows the sample of vesiculated solution. On decompression, bubbles expand to give an approximately constant-velocity flow front (Namiki and Manga [93]), shown by the *line in frames 4–8* (3 ms between frames). During expansion, bubbles coalesce through film breakage to increase significantly in size, especially in the *upper parts* of the flow. The foam then fragments layer by layer, as shown by the *two lines* following liquid-poor regions. Most of the xanthan gum is ejected from the shock tube during the experiment. Reprinted with kind permission from Elsevier

lution. The variables here also included porosity (0.19–0.82 v/v) and absolute decompression (0.41–0.94 atmosphere). High decompression rates result in a variety of flow patterns, including fragmentation, depending on porosity and absolute decompression. At low decompression rates only sample deformation or expansion takes place (Fig. 20). Rough upper surfaces develop on flows during high decompression rates, indicating that bubble walls are rupturing; conversely, lower rates of decompression leave flows with smooth upper surfaces. These experiments suggest that decompression rate is an important parameter in determining eruption behavior, in particular the transition between effusive and explosive eruption in low-viscosity vesiculated magmas subjected to decompression. Explosive response (fragmentation) to decompression results from disequilibrium expansion, where pressure in the bubbles is significantly higher than in the surrounding fluid and bubble expansion is limited by the enthalpy change resulting from the decompression. Bubbles cannot physically expand fast enough to maintain the same pressure as the surrounding fluid. Effusive response results from bubble pressure being close to fluid pressure.

Models of equilibrium and disequilibrium expansion were tested against experimental results. At high decompressions, equilibrium expansion would be faster than dis-

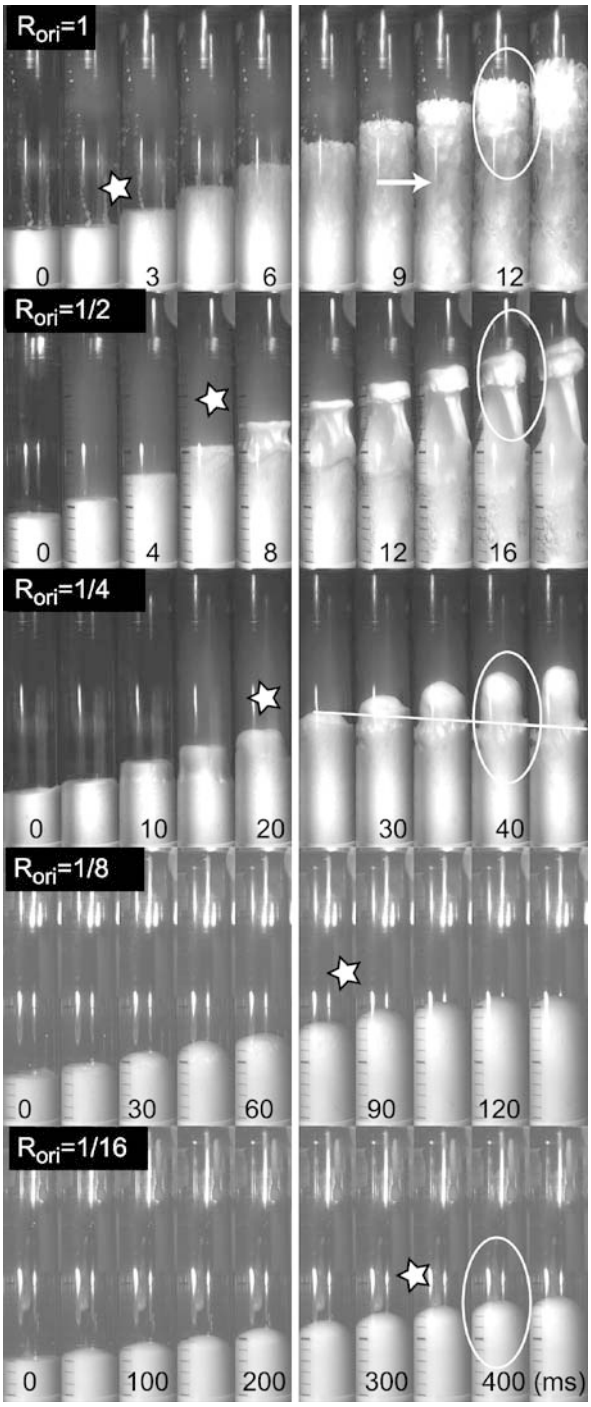

◀ **Volcanic Eruptions, Explosive: Experimental Insights, Figure 20**

Images of a 60-kPa decompression of xanthan gum solution of initial vesicularity about 0.6 (Namiki and Manga [94]). In order to study the effect of decompression rate, the top of the shock tube has been choked with an orifice whose radius is a fraction of the tube diameter (R_{ori}). Experiments with rapid decompressions have short durations (time in ms on individual frames), and slow decompressions significantly longer durations, with the star indicating the time decompression ended. For $R_{ori} = 1$, the flow develops a rough surface suggesting rupturing of bubble walls (*ellipse*), and the *arrow* indicates the onset of fragmentation. When $R_{ori} = 1/2$, the flow top starts to separate (*ellipse*), but surfaces are *smooth* indicating that bubble walls are staying intact. Fragmentation does not occur when $R_{ori} = 1/4$, but the flow detaches from the tube wall as it expands. Smaller orifices result in flow expansion that retains wall contact. Different values of vesicularity and absolute decompression yield different patterns of behavior. Reprinted with kind permission of AGU

equilibrium expansion; experiments follow the slower disequilibrium expansion model. At lower decompressions the disequilibrium model predicts more rapid expansion than the equilibrium model; experiments again follow the slower expansion, this time in equilibrium. Namiki and Manga [94] gave the critical decompression rate for inviscid explosive behavior as $-P'_{Ot} > (2\gamma/\rho_L\phi_i P_{Gi}(1-\phi_i)(\gamma-1))^{0.5}(P^2_{Ot}/h_{Fi})$, where P_{Ot} is pressure outside bubble during decompression, γ is isentropic exponent, ρ_L is liquid density, ϕ_i is initial gas volume fraction, P_{Gi} is internal bubble pressure and h_{Fi} the initial height of the bubbly fluid column. This experimentally based threshold decompression rate is consistent with the estimated decompression rate for the explosive/effusive transition in natural basaltic eruptions (Fig. 21), with Hawaiian eruptions suggested as mainly driven by the sufficiently rapid decompression of bubbly magma.

Degassing-Driven Flows The decompression and expansion of materials with pre-existing vesicularity provides valuable insight into volcanic behavior. However, magmas also degas on decompression, potentially adding to the explosivity of eruptions. In contrast to experiments with natural materials, where high temperatures and pressures make spatially and temporally extensive flows difficult to achieve, analogue experiments offer the possibility of studying the evolution of flow processes from single phase liquid to multi-phase fragmented flow using high-speed imaging and other data collection. Decompression of an unvesiculated hydrated magma results in, amongst other effects, the nucleation and growth of bubbles of gas phase or supercritical water, complex changes in the rheology of the liquid phase, increasing permeability and decreasing diffusivity. Incorporating all these interdependent variables into an experimental system adds complexity and makes understanding the first order effects more difficult, even if scaling between experiment and volcano is possible. Designing experiments to investigate the domi-

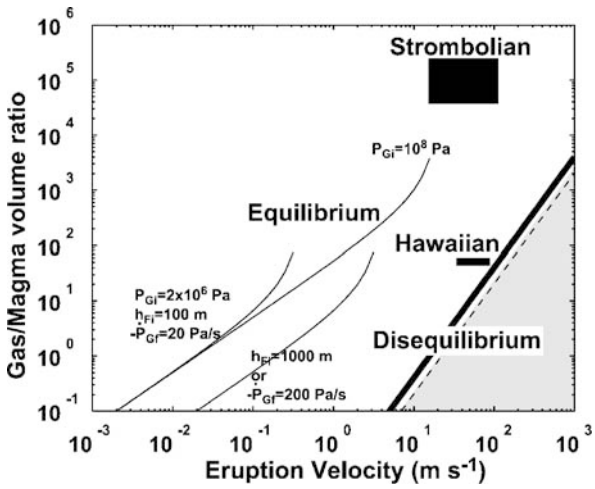

Volcanic Eruptions, Explosive: Experimental Insights, Figure 21
Application of experiments from Fig. 20 to the volcanic case give an estimate of the magma velocity required to cause sufficient decompression rate to initiate explosive eruption (*gray region*) as a function of magma vesicularity (Namiki and Manga [94]), where P_{Gi} is initial bubble gas pressure, h_{Fi} initial height of bubble column and $-P'_{Gf}$ the bubble decompression rate. Eruptions that result from the rapid expansion of bubbly magma are likely to plot on the thick line separating equilibrium and disequilibrium fields. Hawaiian eruptions approach the transition and, on this basis, result from the rapid expansion of vesiculated basalt magma. Strombolian eruptions, however, are indicated to result from another mechanism. Reprinted with kind permission of AGU

nant physics using a canonical approach reduces complexity and eases scaling requirements. The first order physical process occurring during these flows is the phase change of one system component from liquid (or more strictly, dissolved) to gas, with the associated reduction in bulk modulus and response to pressure change. Therefore, the aim of analogue experiments here is primarily to study the effect of degassing on flow.

The decompression of a liquid below its saturated vapor pressure results in boiling of that liquid as it becomes superheated. Experiments carried out by Hill and Sturtevant [50] used a shock tube to rapidly decompress a refrigerant. The refrigerant was Newtonian in rheology and of low viscosity, which, whilst dissimilar to the physical properties of magma, reduces the experiments to the first order process. Boiling produced a flow of vapor with entrained liquid drops. The superheating of the experimental liquid can be likened to the supersaturation of water in magma, and the liquid boiling to the exsolution of water from magma. The emerging gas phase expands, entrains liquid droplets and rapidly accelerates the flow. These experiments showed the development of a sharply defined plane at which explosive boiling occurred.

The addition of inert particles to a volatile liquid [113] reduces the proportion of volatile phase and maintains particles in the post-fragmentation flow. Decompression, combined with the imposition of a steep pressure gradient within a centrifuge, results in a well-defined fragmentation front when decompressions are large. However, the mass of inert material acts to reduce post-fragmentation acceleration thus maintaining pressure and suppressing heterogeneous bubble growth below the fragmentation front. Small decompressions result in bubble nucleation on the inert particles throughout the mixture, illustrating that system behavior depends on the rates at which processes respond to imposed changes. Although these experiments probably represent end-member behavior, they provided a tantalizing glimpse of possible explosive fragmentation mechanisms that may operate in volcanic conduits given suitable conditions.

Volcanically, only a few weight percent of the magma degasses to drive the flow, whilst boiling experiments represent an extreme expression of degassing flows where all the liquid can change phase to gas. Experiments that retain low liquid viscosity and Newtonian rheology, but degas only a fraction of their mass, extend our knowledge of degassing flows and are likely to behave in a different, possibly more volcanic fashion. Mader et al. [78,79] rapidly generated bubbles of CO_2 in water, and throughout the fluid rather than just at the top interface. In volcanic systems, if bubbles nucleate and grow on timescales over which they can rise and escape then significant fluid expansion does not occur. However, if the CO_2 can be experimentally generated with sufficient supersaturation to drive bubble growth on timescales much shorter than those needed for bubble rise, then the two-phase mixture rapidly expands and fragments in a ductile fashion, producing liquid and foam drops suspended in gas, as the gas volume fraction increases. Two methods were used to achieve CO_2 supersaturation up to 455 times ambient. Water was saturated in CO_2 and decompressed in a shock tube developing high supersaturations. An alternative approach used rapid chemical reaction between K_2CO_3 and HCl to generate a supersaturated CO_2 solution. An intriguing positive feedback mechanism was observed whereby agitation of the liquid by the growing gas bubbles enhanced the release of gas thus increasing the violence of the flow. It was found that considerable expansion and acceleration took place before the major change in flow pattern induced by fragmentation. This contrasts with brittle systems that tend to fragment then expand [53]. Accelerations approached 180 g and were proportional to volatile

supersaturation; velocities peaked at 15 m s^{-1}, and strain rates were in the region of 30 s^{-1}. These are considered to be in the same range to those experienced volcanically and indicate a degree of similarity between laboratory and natural systems, with strain rates suggesting that high-viscosity magmas will be fragmenting in brittle fashion.

Injection apparatus in the chemical mixing experiments introduces discontinuities in the tube geometry and this appears to play a role in the behavior of the flow. Mader et al. [80] carried out relatively sustained explosive experiments, lasting about 1.5 s, where a considerable volume of carbonated water was decompressed. The vessel used was a round-bottomed flask. The discharge rate of water was found to fluctuate because of flow instabilities at the base of the flask neck, probably accompanied by pressure fluctuations. The flask neck was also found to pin the point of fragmentation as the flow evolved, with similar results being found by Zhang [129]. These experiments demonstrated the importance of sustained flows in allowing quasi-steady behavior to emerge, as well as demonstrating that conduit geometry influences flow patterns.

Although water-CO_2 experiments illuminate the behavior of explosively degassing systems, water is many orders of magnitude less viscous than magma. Zhang et al. [132] investigated the role of viscosity by adding polymers to the water, increasing liquid viscosity by between 1 and 3 orders of magnitude. Degassing styles ranged between passive degassing with little or no expansion for low supersaturation decompressions, through to rapidly expanding and fragmenting flows when pressure reduces by a factor of 50 or more on decompression. Bubbles were found to appear uniformly throughout the liquid on decompression, and were considered to be the result of heterogeneous bubble nucleation on minute particles in the liquid. This differs from boiling liquid flows and is more consistent with heterogeneous bubble nucleation on crystals in magmas. Initial acceleration of the liquid was found to be constant if ambient pressure was constant, with bubbles growing as $time^{2/3}$. Viscosity was found to be less important than the dissolved gas content in determining explosive capability, indicating that the low solubility of CO_2 in magmas made explosive eruptions driven solely by CO_2 unlikely. Experimental bubble growth was found to match well with models formulated for bubbles in magma, adding to confidence that the experimental system was, at least in part, similar to the natural phenomena.

Taddeucci et al. [116] increased complexity by carrying out degassing experiments with a viscoelastic analogue material, i. e., one that can change from viscous liquid to brittle solid behavior as strain rate increases. The silicate analogue was polydimethyl siloxane, with various additives, sold commercially as 'Silly Putty'. Silly Putty is a dilatant material with yield strength, and is viscoelastic [26]. The volatile analogue was the inert gas Argon, which had minimal effect on the rheology of the Silly Putty. Pressures up to 10 MPa were also found to have minor effect on rheology. The apparatus used was similar to the 'Fragmentation Bomb' (Fig. 8) used for decompression of non-degassing samples; however, because the Silly Putty experiments were carried out at room temperature, a visibly transparent shock tube could be used to facilitate high-speed imaging. Samples were loaded into the base of the shock tube, then exposed to Ar gas at pressures varying between 4 and 13 MPa, for times between 5 and 340 hours, then rapidly decompressed to 0.1 MPa.

Figure 22 shows the flow response on a depressurization of 11 MPa after 5.7 hours of exposure to Ar gas. On decompression, a small upward bulging of the sample surface was followed by detachment of the upper part of the sample from the tube walls. This is indicated by the change in wall refraction between Figs. 22c and 22d as the Silly Putty 'unsticks' from 'wetting' the tube wall and allows a thin film of gas to change the refractive index structure of the interface. The detachment possibly represents a solid-like radial contraction to axial extension (Poisson's ratio) caused by elastic expansion of the bubble-free sample on decompression. Alternatively, the momentary appearance of doming of the sample surface suggests a significant degree of friction with the tube wall and the generation of high strain rate; the Silly Putty wetting of the tube wall breaks down, but wall friction remains significant. Bubbles nucleate and start to expand. The sample surface domes up, suggesting that despite detachment, wall friction is still appreciable (compare with Fig. 20). The sample then expands and develops fractures at the tube wall. Sharp fractures near the flow head were defined as 'rim fractures', with other more numerous and widespread fractures called 'pervasive fractures'. These fractures represent a solid-like response to the high strain rate imposed at the flow margin and give experimental support for a theoretical process proposed by Gonnermann and Manga [47] who postulated that fragmentation was occurring due to viscous shear along flow edges. The fracturing process may also create seismic signals [122] and act as a trigger for seismically detectable resonant phenomena. The fractures only propagate as far as strain rates maintain brittle behavior, which acts to confine fractures within the outer section of the flow. Fragmentation will occur if the strain rate is sufficient for brittle behavior to persist right to the center of the flow. The fractures were also observed to close and heal because, once formed, lower strain rates caused reversion to viscous behavior. The development of such frac-

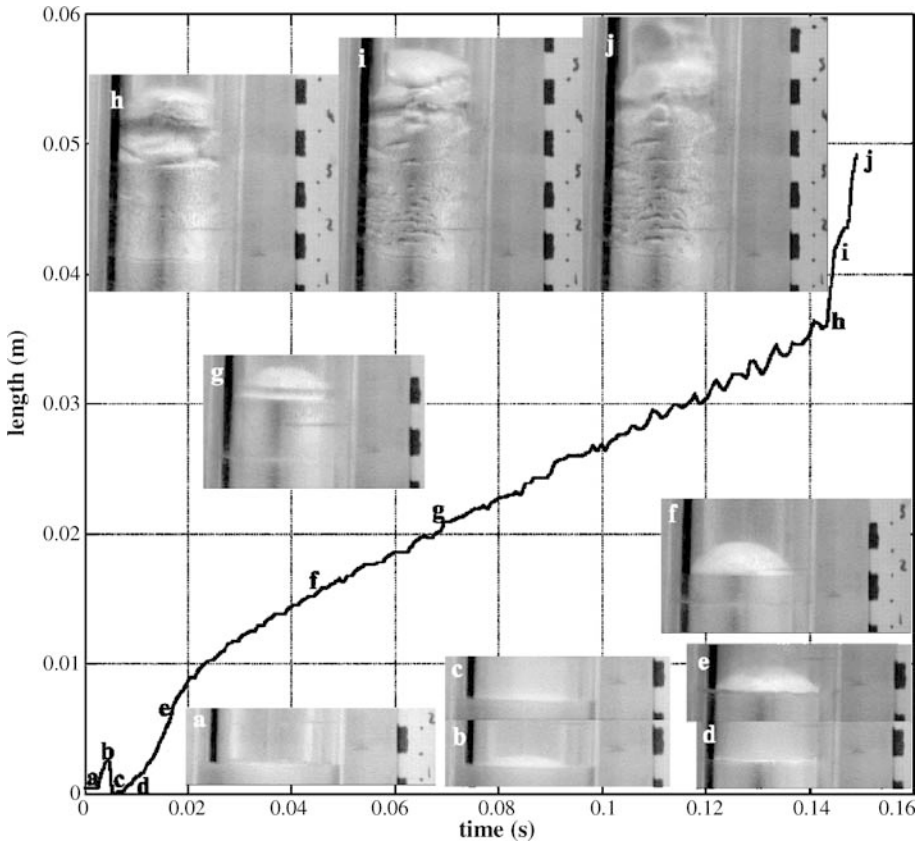

Volcanic Eruptions, Explosive: Experimental Insights, Figure 22

An 11-MPa rapid decompression of viscoelastic 'Silly Putty', saturated with Ar gas (Taddeucci et al. [116]), resulted in an expanding degassing flow. The domed flow top indicates the presence of significant wall friction, despite optical indication that the "Silly Putty" did not wet the tube wall during expansion (compare images c and d). Constant velocity expansion then occurred (e to h) with a circumferential fracture forming near the flow front at g. Flow front oscillation is apparent between g and h, but no concurrent pressure data are reported. From h, the flow front velocity increases dramatically as pervasive brittle fractures appear in the flow margin (i and j) where strain rates are highest. Fragmentation did not occur because the fractures did not propagate through the flow center where strain rates were low and viscous behavior persisted. The fractures could then close and heal due to viscous flow once strain rate declined. These experiments give valuable insight into the process of margin fracture, which could provide an efficient means for gas escape, thus suppressing explosive behavior. They also demonstrate a mechanism of generating fluctuations in magma effusion rate. At volcanic scale, margin fracture may be expected to generate a seismic response, and field evidence of such processes exists (Tuffen and Dingwell [122]). Reprinted with kind permission from Elsevier

ture networks has significant implication for mechanisms of magma degassing. The continual formation and healing of margin fractures gives rapid access to the conduit margin for gas deeper in the flow. This has consequences for transitions between effusive and explosive behavior.

Only a few Silly Putty experiments underwent explosive fragmentation. All these were triggered at a sudden tube widening, emphasizing the importance of conduit geometry on flow behavior in volcanic systems. In comparison, similar decompressions of pre-vesiculated samples [53] resulted in much more extensive fragmentation. This demonstrates that, although there may be more gas dissolved in an unvesiculated sample than contained in the pores of another non-degassing sample, the pre-vesiculated material is more explosive. The process of bubble nucleation and diffusive growth provides a regulating mechanism on the effective maximum rate of decompression and reduces initial explosive activity. However, the continued supply of volatile into bubbles means that degassing flows are likely to expand more given sufficient time, but in the absence of geometric structures, will they fragment in a brittle solid manner?

The viscosity of the liquid phase in magma increases by several orders of magnitude (Fig. 1) as water degasses,

also decreasing the strain-rates over which viscous flow turns to brittle solid behavior. Water and melt interact to change the lengths of silicate molecules and 'lubricate' differential molecular motion, hence changing viscosity. As an analogue material, gum rosin derives from removing volatile compounds from tree sap, leaving 3-ring aromatic hydrocarbons typified by abietic acid ($C_{19}H_{29}COOH$). It is a glassy organic solid at room temperature, but melts in the region of 80°C, and becomes increasingly less viscous as temperature rises. Warm gum rosin behaves in a viscoelastic fashion (Bagdassarov and Pinkerton [10]), being brittle when deformed rapidly but viscous under more gentle treatment. Such behavior is phenomenologically similar to a silicate melt. Phillips et al. [97] found that certain oxygen-bearing organic solvents, such as acetone, dissolved in gum rosin to yield a liquid at room temperature. At concentrations below about 20% w/w solvent these appear to be true solutions [71], with a partial pressure of acetone below that of the pure liquid suggesting some chemical interaction; another phenomenological similarity to hydrated magma. The viscosity of this liquid was found to be a non-linear function of the acetone content of the solution, again similar to a hydrated silicate melt (Fig. 23). Decompression of gum rosin-acetone solution below the saturated vapor pressure of acetone (about 20 kPa at room temperature) was found to result in volatile degassing and consequent increase in liquid viscosity. Slow decompression resulted in the establishment of a degassing interface that consumed the liquid solution. No measures were taken to remove nucleation sites from the liquid, suggesting that nucleation was sensitive to pressure and was confined to a surface layer in the liquid. This is similar behavior to boiling experiments [50,113], where the reaction force from expansion of degassing volatile and accompanying liquid was considered to elevate pressure in the underlying liquid and suppress bubble nucleation and growth, therefore acting to confine rapid degassing to a thin interface. This process may not be as prevalent in H_2O-CO_2 and Silly Putty systems because absolute pressures were larger, in relation to hydrostatic, than in gum rosin-acetone or boiling refrigerant experiments.

Rapid decompressions of gum rosin-acetone resulted in expanding foam growing from the liquid surface [97]. Flow-front acceleration was generally constant, except for large supersaturations with high acetone content where accelerations increased with time. These were reinterpreted by Mourtada-Bonnefoi and Mader [89] in the light of additional experiments that found bubble growth prior to fragmentation is diffusion-driven, but limited by the high-viscosity skins that develop around growing bubbles because of the reducing volatile concentration. This infers

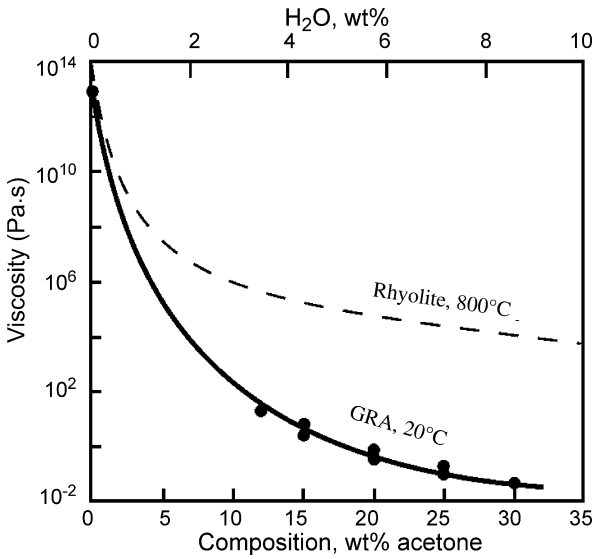

Volcanic Eruptions, Explosive: Experimental Insights, Figure 23
The viscosity of hydrated silicate melts changes non-linearly with water content (Fig. 1). One experimental system uses gum rosin as an analogy for silicate melt, and organic solvents such as acetone and diethyl ether instead of water. The degassing of gum rosin acetone (GRA) results in increasing liquid viscosity (*lower scale*), in similarity with the degassing of hydrated magma (*upper scale*) (Mourtada-Bonnefoi and Mader [90]). The volatile content between natural and analogue systems can be compared by quantifying in terms of moles of volatile per unit volume of liquid; 4% water w/w in magma equates to about 20% w/w acetone in gum rosin. The scales here are not matched on this basis. Reprinted with kind permission from Elsevier

that diffusion-limited degassing is unlikely before fragmentation removes viscous control as a physical mechanism. Accelerations and velocities using gum rosin-acetone were an order of magnitude less than those produced by CO_2-H_2O experiments, but this is consistent with the order of magnitude lower driving pressures. Experiments carried out with high supersaturations allowed extensive acetone degassing to take place. This resulted in the production of solid gum rosin foam with elongated bubbles, phenomenologically similar to woody or tube pumice.

The experiments of Phillips et al. [97] were carried out using small volumes of solution in un-instrumented and relatively short tubes that did not allow sufficient time or length for flow patterns to develop and stabilize. Lane et al. [70] carried out similar experiments using longer tubes, larger volumes of gum rosin-acetone and gum rosin-diethyl ether solutions, and high-speed pressure measurement. Instantaneous decompression from 10^5 to 10^2 Pa resulted in rapid degassing from the liquid surface. Flow became pseudo-steady after about 0.2 s, with foam expanding from the liquid interface, and continued

explosive behavior for about 1.5 s. Pockets of gas developed at the tube wall, eventually merging to allow the expanding foam to detach from the wall. Removal of wall drag allowed the foam core to accelerate as an inverse annular, or detached flow, and break up into foam fragments that were ejected from the top of the shock tube (Fig. 24). The fragmentation of the foam core was precluded during expansion of a non-degassing system [93], suggesting that degassing systems have different flow patterns to those driven by gas expansion alone. The reason for this difference may lie with the smaller overpressure in bubbles of pseudo-steady degassing systems compared with those in transient expanding flows. The lower bubble pressures prevent bubble wall breakage and stabilize detached flow until fragmentation commences [93]. This suggests that flow behavior preceding fragmentation may differ between explosive volcanism triggered by rapid decompression of vesiculated magma, for example the May 18, 1980 Mt. St. Helens lateral blast triggered by a landslide, and sustained explosive volcanism such as the Plinian eruption which followed the blast. The fragmentation mechanism may also be different, with brittle spalling fragmentation before expansion (Figs. 9, 16 and 17) suggested for the lateral blast and fluid expansion before 'plastic' fragmentation (Figs. 19, 20 and 24) for the Plinian eruption.

The size of fragments of gum rosin foam was initially significantly smaller than tube diameter, but increased to near tube diameter and elongated with time, suggesting a declining flow energy. Interestingly, some of the flow features found by Cagnoli et al. [25] in expanding particle beds appear similar to those in expanding gum rosin foams, with bubbles appearing at the margins of both flows and being stretched out (Fig. 15). The sequence of flow pattern changes in the gum rosin flows was found to result in the generation of characteristic pressure fluctuations. Pressure fluctuations may be inherent in the unsteady nature of flows undergoing changes in behavior and such phenomena could be detectable by seismic means during volcanic flow.

Magma can fragment in the presence of pre-existing bubbles alone without any degassing (e. g., Fig. 9), and the presence of crystals increases the number density of nucleation sites (e. g., Fig. 4), and hence the surface area over which degassing can occur. Magmas are decompressed with crystals and bubbles often present, an aspect investigated analogously by Mourtada-Bonnefoi and Mader [90] using gum rosin-acetone mixtures and a variety of particles and existing bubbles. It was found that fragmentation occurred at lower volatile contents and lower decompressions in the presence of particles and bubbles. It was also found that flow behavior depends on the distribu-

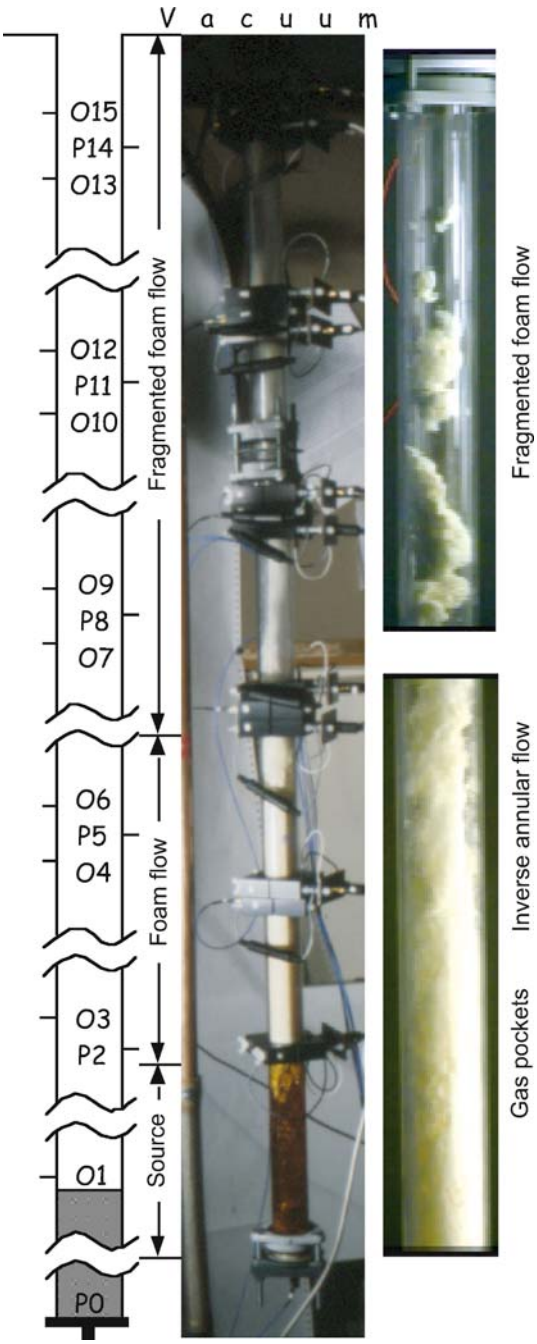

Volcanic Eruptions, Explosive: Experimental Insights, Figure 24
An 0.1-MPa rapid decompression of a solution of gum rosin 19% w/w diethyl ether generates a degassing flow (Lane et al. [70]). Bubble growth occurs predominantly at the upper liquid surface. The expanding flow develops gas pockets at the wall, which merge and allow a central foam core to detach from the tube wall giving inverse annular flow. The detached foam core is no longer constrained by wall drag and expands until it breaks up to give a fragmented flow. The fragments of gum rosin foam are ejected from the shock tube. Reprinted with kind permission of AGU

tion of particles in the liquid, with more rapid degassing where particle concentrations were high. This suggests that heterogeneity in the distribution of crystals and pre-existing bubbles will develop into sizable features within a degassing flow, at least before fragmentation. This raises the possibility that flow heterogeneity over a range of size scales may then play a role in the fragmentation process.

Low VEI Events High VEI events do not occur very frequently compared to anthropogenic timescales, and do not lend themselves readily to detailed field investigation of subterranean flow processes. Conversely, low VEI events, typified by Strombolian activity, occur on anthropogenically short timescales and form relatively safe natural laboratories to study magma flow in the volcanic conduit. Low VEI eruptions commonly occur at volcanoes where magma viscosity is low and degassing water is able to separate relatively easily from the magma; indeed, this separation is required to produce low VEI events. Analogue experiments of separating flows may, therefore, be compared with relatively well-studied natural processes in ways that high-VEI events cannot. Experimental conditions are also simplified in relation to those applied to high-VEI processes, with the assumptions that liquid rheology and temperature remain constant being easier to justify. Low-VEI processes may also be studied without shock tube or high-pressure apparatus, making observation and measurement more straightforward.

Strombolian eruptions are characterized by short-lived explosive events that erupt significantly higher water to silicate mass ratio than that of the bulk magma [15,30]. These events are separated by longer periods of relative quiescence, suggesting the presence of a gas separation process that produces gas-rich regions that rise up the conduit between intervals of relatively gas-depleted magma. In two-phase flow terminology, one such regime is that of slug flow, where large ascending gas slugs can form from the coalescence of many much smaller bubbles. However, evidence from extensive research on water-gas systems indicates that for such a flow pattern to be stable in the conduit the overall gas volume fraction would probably need to exceed about 0.25 [34], and the time intervals between explosions would be similar to, or less than explosion duration. It is also very questionable that slug flow would form from numerous small bubbles in a relatively high viscosity liquid like basalt magma. This suggests that other mechanisms operate to promote formation of gas slugs. Jaupart and Vergniolle [63,64] were among the first to recognized that volcanic conduits are not straight vertical tubes, but exhibit significant geometrical heterogeneity that may play a major role in eruptive behavior. Experiments were carried out to examine the effect of a flat roof in trapping the numerous small bubbles rising through the magma (Fig. 25). Small air bubbles were injected at variable rate into the base of a tank containing liquids of various viscosities. The bubbles rose to the roof of the tank, which had a vertical small diameter outlet tube. Three modes of system behavior were observed. At lower liquid viscosity, a raft of bubbles formed at the roof. At some critical thickness the bubble raft collapsed on the scale of the tank roof and flowed up the outlet tube as a large gas bubble. This be-

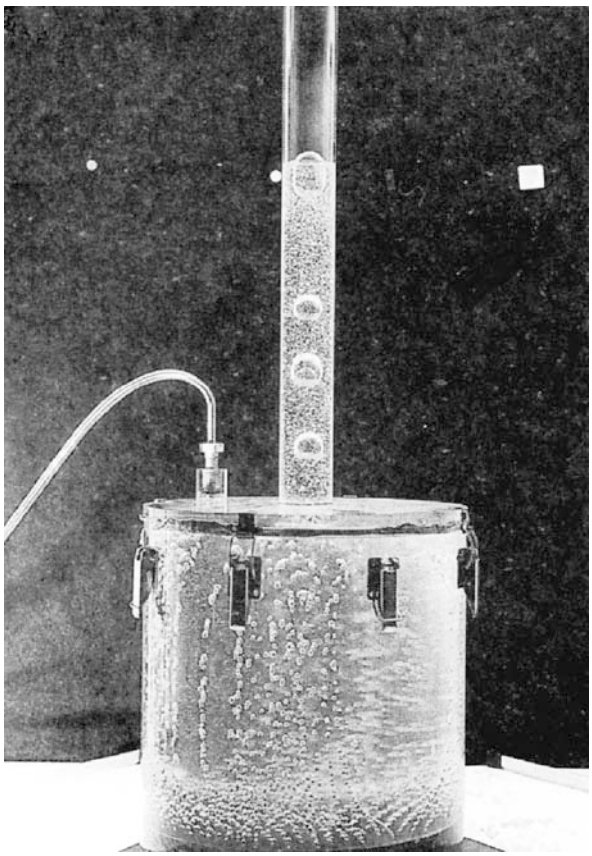

Volcanic Eruptions, Explosive: Experimental Insights, Figure 25
Numerous small nitrogen bubbles rising in a silicone oil of viscosity 1 Pa s (Jaupart and Vergniolle [63]) encounter a structural discontinuity in the form of the horizontal roof of the experimental tank. Bubbles trapped under the roof collide and coalesce to emerge up the narrow exit pipe as much larger bubbles. Such a mechanism can be envisaged for the production of large bubbles some 1–3 km beneath the vent at Stromboli volcano (Burton et al. [21]); however, the experimental geometry canonically demonstrates the process rather than simulating conduit structure. Large bubbles emerging from the trap then expand and become overpressured on approach to the surface (James et al. [61]). Reproduced with kind permission of Cambridge University Press

havior was considered analogous to Hawaiian eruptions. At higher liquid viscosities, the bubble raft collapsed on a scale smaller than the tank roof and a series of bubbles that are much larger than those injected, but smaller than the low viscosity case, emerge up the outlet pipe. This gives insight into Strombolian eruptive activity. The final mode of behavior was the development of a relatively stable bubble raft and the escape of small bubbles up the outlet pipe to give bubbly flow at the surface.

The ascent of large gas bubbles, or slugs, up a tube can be controlled by surface tension, liquid viscosity or liquid inertia. In basaltic systems surface tension has negligible control, and inertia dominates with some component of viscous influence [106]. Conveniently, this allows the use of water and air at laboratory scale to give inertia-controlled slug flow. Seyfried and Freundt [106] carried out experiments applying the flow patterns that develop with increasing air flux through a column of water to basaltic volcanism. Single gas slug ascent was found to only occur through sudden incidents of limited volume gas release, with one mechanism being that of foam collapse from a trap [63,64]. It was observed that slug burst in the tube was a gentle process, but that if the liquid level was in the reservoir atop the tube (akin to a lava lake) then burst was more vigorous and could be likened to Strombolian activity; another example of the impact of conduit geometry on flow. Experiments with gas traps in the water column found that the coalescence of trapped gas pockets resulted in significant pressure oscillations that were a potential source mechanism for volcanic tremor; linking volcano-seismic signals to fluid flow could be a powerful method of experimental scaling inaccessible in high VEI events.

Seismic signals may result from the collapse of a foam layer trapped at a roof. Measurements made within apparatus similar to that of Jaupart and Vergniolle [63,64] reveal pressure fluctuations both within the fluid and in the air above the fluid surface in the outlet tube [102]. These could be linked to seismic and acoustic signals measured at Stromboli volcano and give insight into the role of fluid-flow on the generation of seismo-acoustic waves during explosive volcanism. During the build up of the foam layer, no pressure changes were detected. Collapse of the foam layer and motion of the gas slug up the outlet pipe produced pressure changes in the liquid tank of a few-hundred Pa, and of a few-tenths Pa in the air above the liquid level. Collapse of the foam layer generates upward gas flow in the outlet pipe thereby producing downward motion of the dense liquid phase in the outlet pipe, reducing pressure. Oscillation of liquid on the gas spring of the slug produces pressure oscillations within and above the liquid. Slug burst results in a sudden increase in oscillation frequency above the liquid, and a small pressure pulse in the liquid due to decay of liquid flow around the slug nose [59]. Ripepe et al. [102] postulate that seismic signals measured at Stromboli are generated during rapid gas expansion during foam collapse, and estimated that the slug be about 4.5 m long and induce a pressure drop of 0.1 MPa in the magma.

Inversion of highly repeatable seismic waveforms measured at Stromboli Volcano, Italy [31,33] has enabled imaging of the conduit as well as revealing the forces and pressures being created by magma flow during Strombolian eruptions. In summary, Stromboli's plumbing system appears to comprise a dyke dipping about 72° that runs from the vent system to an intersection with a dyke dipping about 45° and 900 m below the vent system. Stable and repeatable seismic sources are inferred at a depth of about 220 m below the vent system, and have the characteristics of a dyke segment intersecting the main conduit. Another, weaker, seismic source is located at the intersection of the two main dykes about 900 m below the vents. The dominant seismic source at 220 m below vent generates a downward vertical force of about 10^8 N followed by an upward force of similar magnitude (Fig. 26a). The source volume increases during the downward force and decreases with the upward force. These measurements provided a means of testing laboratory models of Strombolian eruptions.

James et al. [59] studied pressure oscillations resulting from the ascent of single gas slugs in a vertical tube. The oscillations observed at the top of the liquid column (Fig. 27) showed an increasing amplitude and decreasing period as the slug ascended, with oscillations ceasing as the slug reached the surface. Burst of the meniscus produced further pressure fluctuation. Modeling of the oscillations revealed the source mechanism as bouncing of the liquid mass above the slug on the spring of gas within the slug. The observed pressure oscillations were not similar to the change in volume of the VLP seismic source at Stromboli (Fig. 26a), nor were oscillation amplitudes sufficient to account for forces of 10^8 N at volcanic scale. Bubble bursting was also quiescent rather than explosive.

Such experiments on Strombolian processes were not scaled for expansion of the gas slug as it rises, and gas chemistry suggests that slug source depths could be in the range 900–2700 m (Burton et al. [21]). A slug ascending from 1000 m in magma will expand about 200 times to reach atmospheric pressure, but in the experiments expansions were in the region of 1.2. In order to investigate more appropriate values, James et al. [61] used low vapor-pressure vacuum pump oil as an analogue liquid and varied

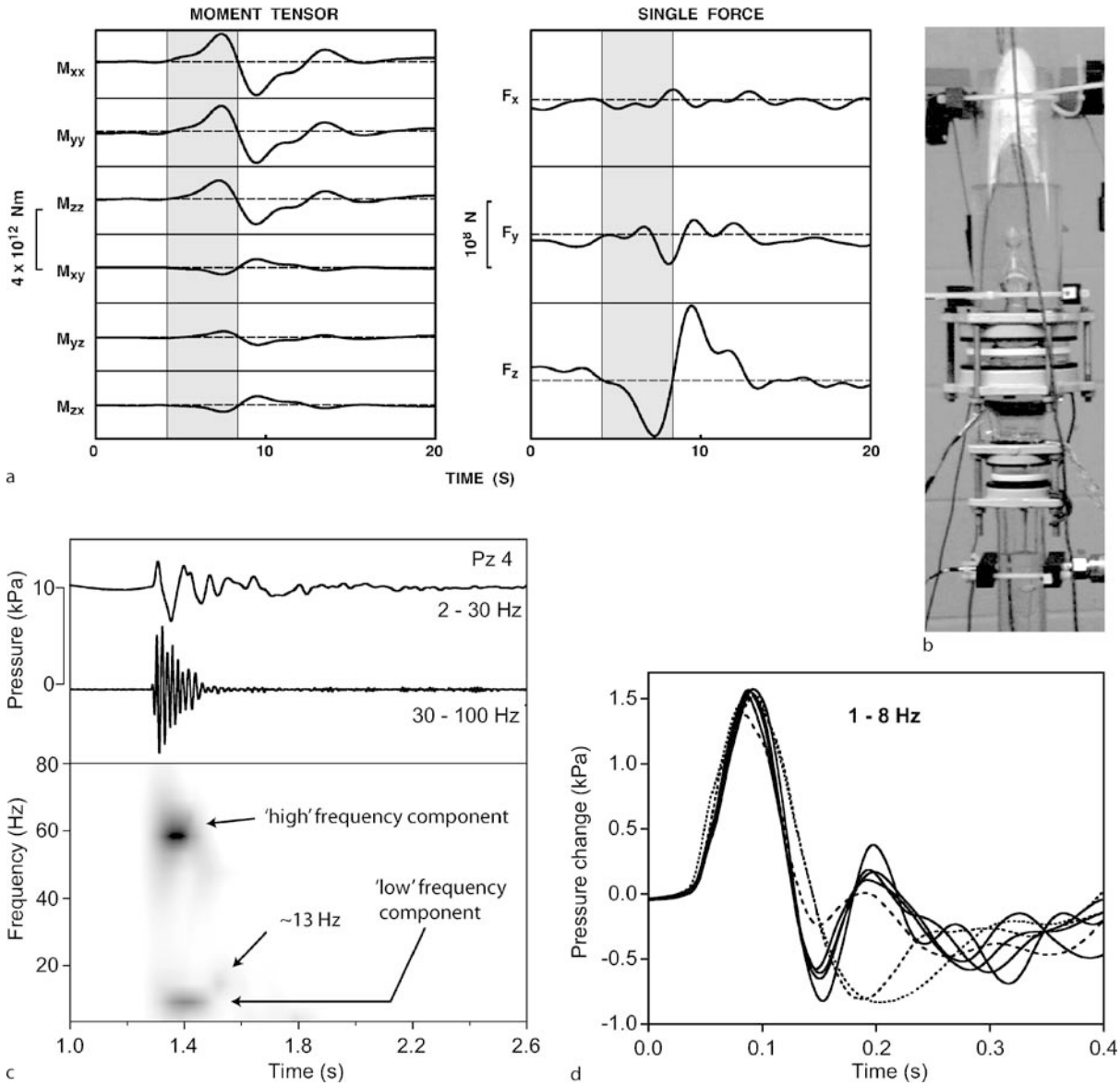

Volcanic Eruptions, Explosive: Experimental Insights, Figure 26

The geometry of experimental tubes has been observed to change flow behavior in a number of instances. James et al. [60] investigated the dynamics of a gas slug ascending through a tube widening. Inversion of seismic data measured at Stromboli volcano (a, Chouet et al. [31]) indicated an initial downward force of about 10^8 N, accompanied by pressure increase expanding the conduit. This was followed by upward force and conduit contraction with a period of about 10 s. The experimental ascent of a gas slug through a tube widening (b) caused breakup of the slug and production of a transient gas jet. Spectral analysis of pressure (c) showed three resonant components. The low frequency component (d, about 6 Hz) showed one to two cycles of pressure increase followed by decrease, qualitatively consistent with pressure change in the conduit at Stromboli during explosive eruption. Measurement of apparatus displacement indicated a peak downward force of 30 N, which when naïvely scaled by 100^3 for physical dimension and 2.5 for density difference gives a force of 7.5×10^7 N, encouragingly close to that estimated from seismic inversion. Reprinted with kind permission of AGU

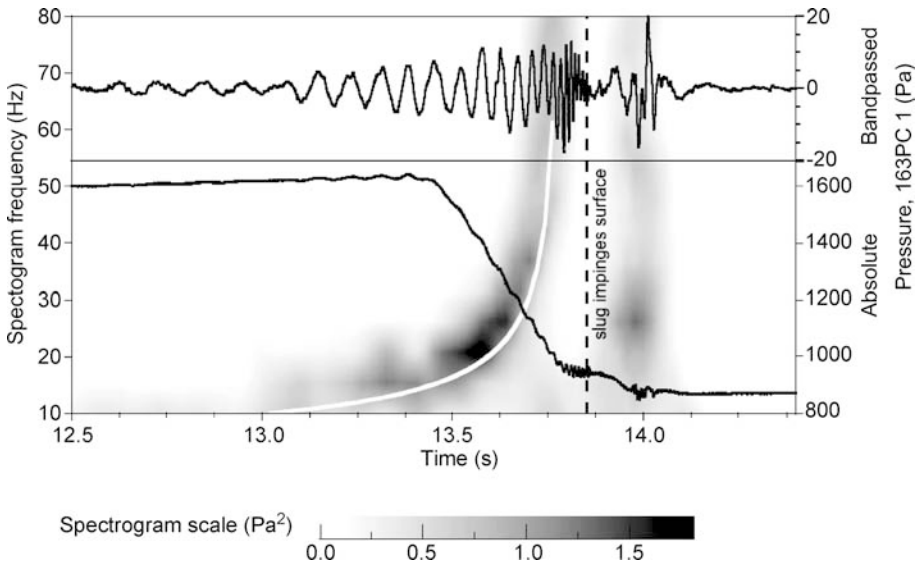

Volcanic Eruptions, Explosive: Experimental Insights, Figure 27

Strombolian eruptions are short-lived, periodic and have high volatile content. For these reasons they are associated with the formation, ascent and bursting of large bubbles of water vapor (often known as slugs), rather than the expansion of a vesicular magma (Fig. 21). The ascent of air slugs in water during small-scale experiments caused pressure oscillations (James et al. [59]). A spectrogram (*gray shading*) of band-passed (*upper trace*) pressure data (*lower trace*) was found to match well with calculations of the oscillation frequency (*white trace*) of the water above the gas slug bouncing resonantly on the gas slug. This oscillatory mechanism could not be used to account for the seismic signals used to image the conduit at Stromboli (Chouet et al. [31]). Reprinted with kind permission from Elsevier

the ambient pressure above the liquid surface between 10^5 and 10 Pa. The reduction in ambient pressure provided a proxy for slug ascent from depth and allowed potential gas expansions of factors up to 10^4. The positions of the slug base, slug nose and liquid surface for ascending slugs of various gas masses and ambient pressures are shown in Fig. 28. Regardless of ambient pressure and slug mass, the ascent velocity of the slug base remained essentially constant. However, the slug nose accelerated rapidly on approach to the surface when ambient pressure was low, and burst with a pressure higher than ambient. Gas slugs that underwent little expansion burst at near ambient pressure. Scaling to the volcanic case was undertaken using experimentally verified CFD modeling. This showed that slug burst pressure is a function of conduit radius and slug mass, with large slugs in narrow conduits producing gas overpressures in the region of 1 MPa at burst. This overpressure will fragment the upward-moving (tens of m s^{-1}) meniscus of liquid existing just prior to burst. The effect of such decompression on the vesiculated basalt magma flowing down the conduit walls and slug base could be explosive fragmentation (Namiki and Manga [94]). This would eject pyroclasts, adding greatly to the pyroclast volume from the meniscus and explaining how explosive slug burst results in the ejection of significant masses of pyroclasts. There is also the possibility that the gas slug is not a single bubble, but more accurately described as a foam raft, although such a structure may not survive the degrees of expansion undergone in ascent from 1000 m depth with its bubble walls intact. The rapid 1 MPa decompression of a foam raft would also produce pyroclasts as the foam fragments.

Rapid slug expansion, which generates the explosivity of Strombolian eruptions, is a process operating within the top few tens of meters of the magma column (Fig. 28c). This, and the fact that the forces generated are of order 10^6 N upward, means that slug expansion is not the fluid dynamic source process responsible for the VLP seismic signals inverted by Chouet et al. [31]. Slug expansion probably plays little role in this particular seismic signal because the magma head is probably too large for the slug to be expanding as any appreciable rate. The stable position of the seismic source suggests geometric control of a flow process, a phenomena we have seen a number of times before. James et al. [60] measured pressure and force as a gas slug ascended from a narrow tube through a flare into a wider tube. The gas slug segmented on passing through the flare and liquid flow during this process generated

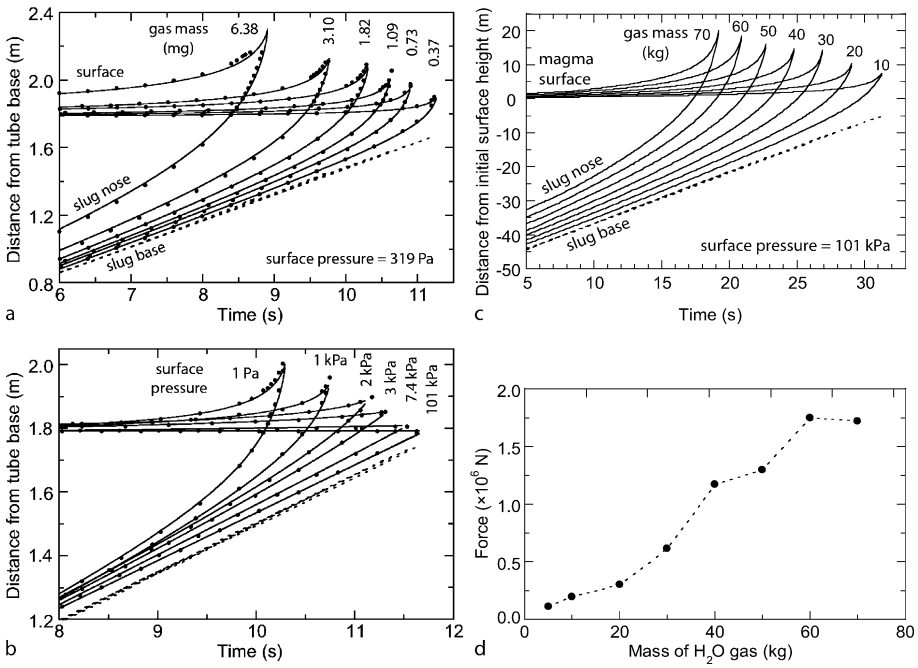

Volcanic Eruptions, Explosive: Experimental Insights, Figure 28

Experiments on slug flow are numerous in the engineering literature, but volcanic flows operate under conditions not investigated for anthropogenic purposes. One of these conditions is the large pressure change a gas bubble will experience in rising from some depth. James et al. [61] investigated the dynamics of rapid gas slug expansion as a result of this decompression using an air bubble and vacuum pump oil as analogue fluids. Using surface pressures below atmospheric allows the air bubble to undergo considerable decompression under laboratory conditions. The degree of slug expansion, and hence explosivity on burst, was found to depend on (a) the mass of gas in the slug as well as (b) the applied surface pressure. Computational fluid dynamic modeling was used to apply the experimental findings at volcanic scale (c). The process of slug expansion generated significant pressure change and vertical force (d), but the amplitudes are two orders of magnitude less than those calculated at Stromboli (Chouet et al. [31,33]). Reproduced with kind permission of the Geological Society of London

pressure oscillations (Fig. 29). Analysis of these oscillations (Fig. 26) revealed three components centered on 60, 13 and 6 Hz, with the low frequency component showing increasing pressure followed by a decrease, resembling the expansion-contraction waveform from inversion of VLP seismic data (Fig. 26a [31]). The vertical force generated during the experimental pressure increase was 30 N downward. Applying a simplistic scaling based on density and volume differences between laboratory and volcano gives a force in the region of 10^8 N, of similar magnitude to that emerging from inversion of VLP seismic data. It was, therefore, suggested that the source mechanism for VLP seismic signals measured at Stromboli by Chouet et al. [31] was the ascent of slugs sourced at depth [21] through a geometric widening of the feeder dyke inclined at 72°. The upward widening disrupts the downward flow of liquid around the ascending slug, allowing significant thickening of the falling film. As this thickening falls through the downward narrowing, it pinches closed and segments the slug. The resulting liquid piston continues to descend because of its inertia, and increases pressure below it. The piston then oscillates for a cycle or two before being disrupted by the buoyancy of the gas beneath. Experimental insight suggests that the source mechanism for the VLP seismic signal is the deceleration of the liquid piston.

Explosion Consequences

Various consequences of explosive degassing of magma have been studied by analogue means, and many of these are reviewed in Sparks et al. [109]. Here, we review a sample of more recent literature.

Cinder Cones Explosive volcanic eruptions produce pyroclastic constructs, the most common form being known as cinder or scoria cones. The basis for understanding cinder cone growth was given by McGetchin et al. [85]. These structures, which are piles of generally uncon-

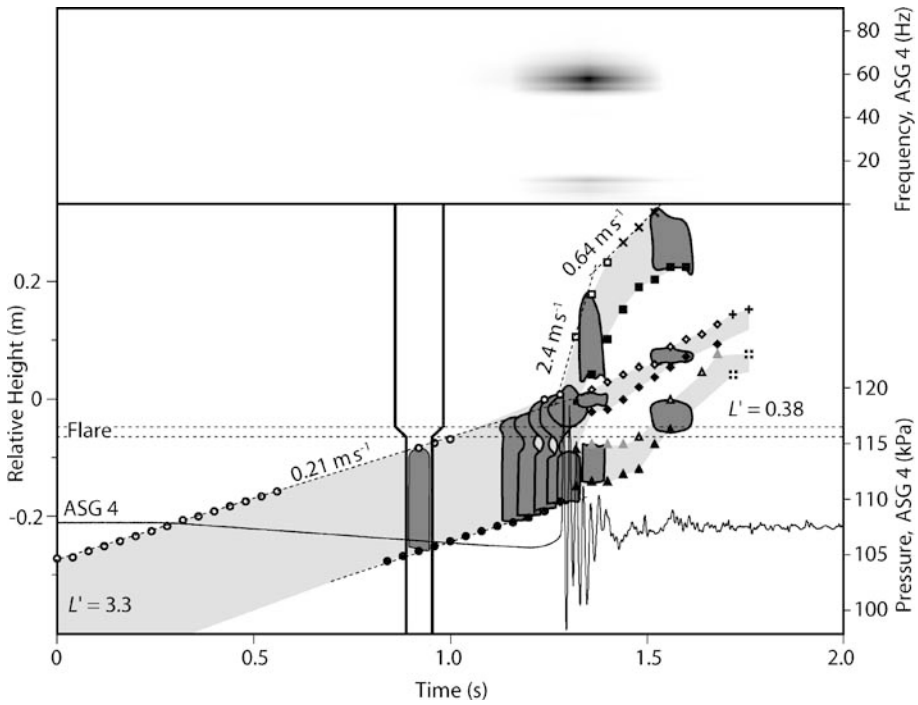

Volcanic Eruptions, Explosive: Experimental Insights, Figure 29

The source mechanism suggested by laboratory experiment (Fig. 26) for seismic signals measured by Chouet et al. [31] at Stromboli Volcano, Italy is the creation and rapid deceleration of a downward-moving liquid piston, formed as a gas slug breaks up in a conduit widening (James et al. [60]). The liquid piston 'bounces' on the lower portion of the slug for a cycle or two before being consumed by the ascending remains of the original slug. The cartoon superimposed on the pressure trace correlates the position of gas bubbles with pressure measurement and spectral analysis of pressure fluctuations. The dramatic velocity increase of the slug nose does not appear to play an obvious role as a source mechanism and may be associated with the highly unstable bubble in the tube widening. Reprinted with kind permission of AGU

solidated magma fragments, are prone to gravitational collapse and rapid erosion. To test the hypothesis that cinder cone geometry is controlled by the behavior of unconsolidated fragments, Riedel et al. [100] carried out small-scale analogue experiments using grains of sugar and sand. These were built into cones, drained in the center to represent the crater. These grain piles showed similar geometry to natural cinder cones (Fig. 30). Quantification of the geometry showed that analogue and natural cones were identical in shape to within 5%. This demonstrates that the angles of repose exert fundamental control on cinder cone geometry, with surface grain flow maintaining angle of repose during cone growth. The angle of repose nature of cinder cones emphasizes their vulnerability to erosion and collapse.

Pyroclastic Flows Pyroclastic flows 'hug' the ground because their density is greater than that of the atmosphere. Gravity then drives the pyroclastic material in a lateral fashion; hence pyroclastic flows are part of the family of gravity currents, and the sub-family of particle-driven flows that includes turbidity currents. The flow behavior of pyroclastic flows, which result from explosive volcanism or dome collapse, has been investigated at laboratory scales using analogue materials for a number of years [109]. Gladstone et al. [46] experimentally examined the dynamics of gravity currents that were initially stratified in density using a 3 m long flume. A range of brine solutions of different density were layered into a lock at one end of the flume and released into tap water. The solutions were used singly, in two or three layers to investigate how stratification can develop or be homogenized in pyroclastic flows and turbidity currents. Flows were visualized by coloring the solutions allowing detection of mixing and stratification.

Flow behavior was found to depend on two dimensionless parameters. Firstly, the relative values of buoyancy between brine solutions, B^*, where most salt is in the lower layer for $0 < B^* < 0.5$, and in the upper layer for $0.5 < B^* < 1$. Secondly, the density ratio of the solutions,

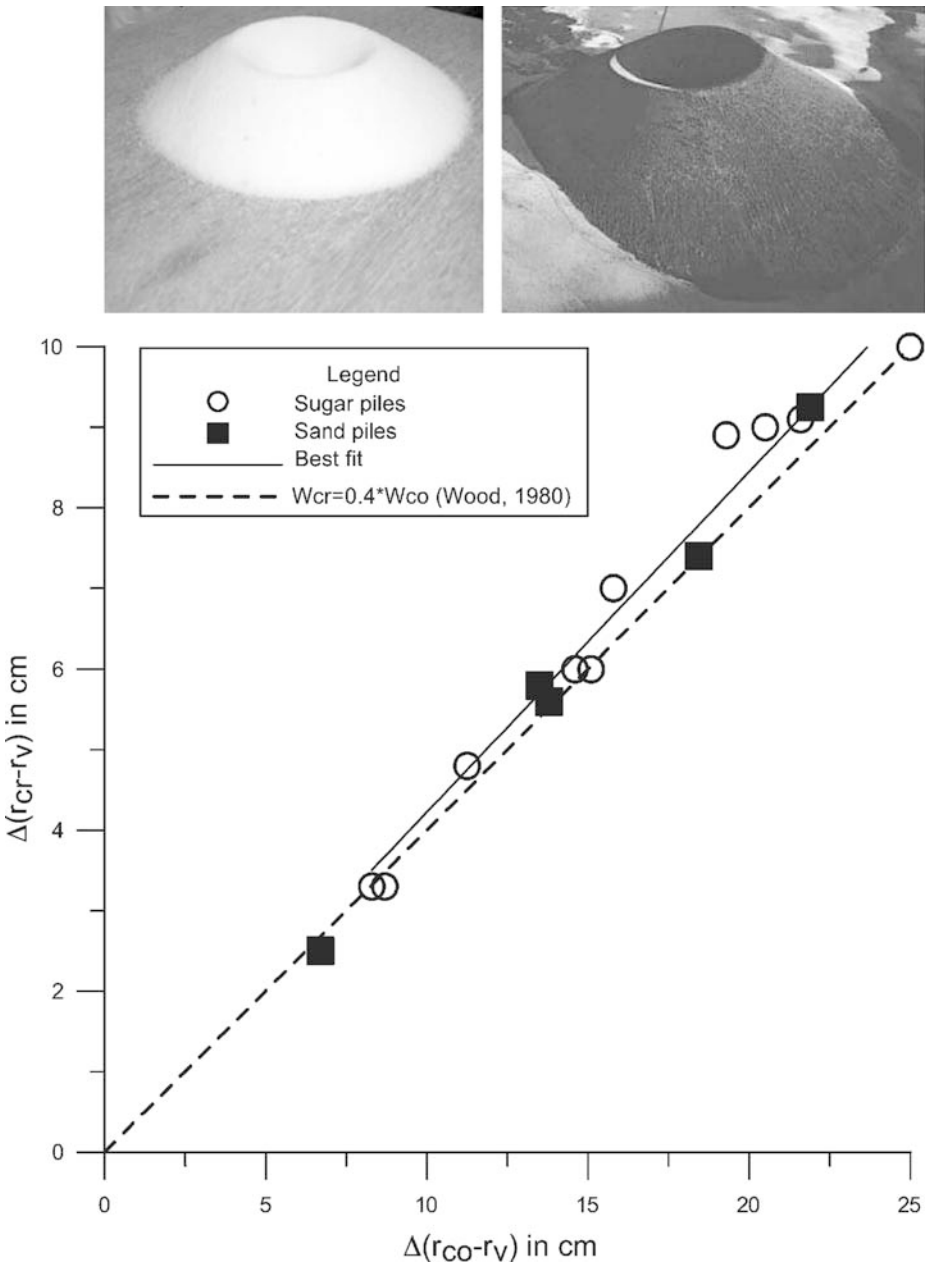

Volcanic Eruptions, Explosive: Experimental Insights, Figure 30

Cinder cones are the proximal expression of fallout from explosive volcanic eruptions. Riedel et al. [100] found that small piles of sugar (*left image*) or sand look similar in shape to much large piles of cinder (*right image*). Comparison of characteristic dimensions of laboratory cones (r_{cr} is distance from central hole to crater rim, r_{co} is distance from central hole to pile extremity and r_v is vent radius) with established measurements of natural cinder cones confirms the similarity. These experiments showed that cinder cones grow by the surface transport of grains under gravity to leave the slopes at angle of repose. Reprinted with kind permission from Elsevier

ρ^*, which is small for strong stratification, approaches 1 for weak stratification and equals 1 for unstratified conditions. When the lower layer has a greater proportion of the driving buoyancy ($B^* < 0.5$), it can run ahead leading to flow stratification (ρ^* approaches 0, Fig. 31), or the layers can mix to produce a homogeneous current (ρ^* approaches 1, Fig. 31). If the upper layer is more buoyant ($B^* > 0.5$), it travels to the nose of the current by mix-

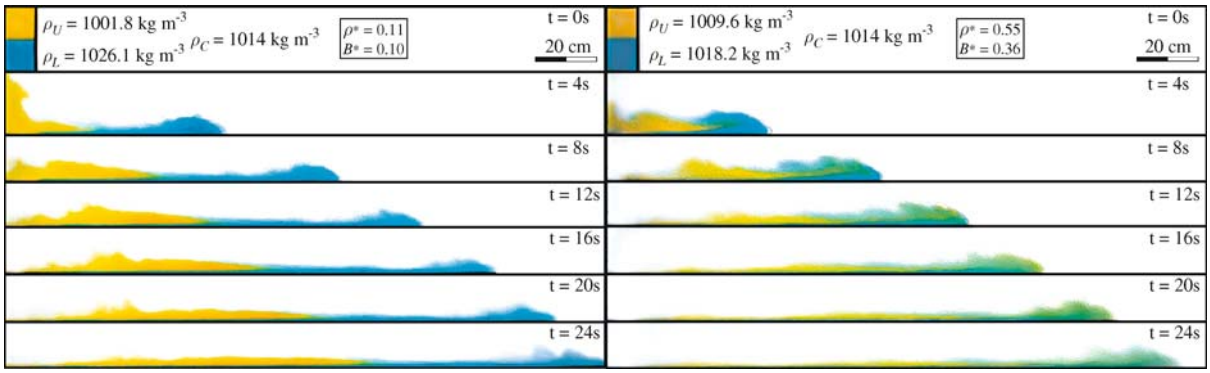

Volcanic Eruptions, Explosive: Experimental Insights, Figure 31

The first order behavior of gravity currents, which include pyroclastic flows, can be studied by analogue experiment using dyed brine solutions of different density propagating in a water-filled flume tank (Gladstone et al. [46]). Density stratification in the flow source material may be preserved under certain conditions of buoyancy and density (*left panel*). Pyroclastic deposits from such flows are likely to show vertical and horizontal stratification. Under other buoyancy and density conditions (*right panel*) the initial stratification is removed by turbulent mixing and resultant pyroclastic flow deposits are expected to be unstratified. Experiments that reduce complexity, as here by not including suspended particles, often give greater insight into primary controlling parameters. Complexity can then be added to understand the smaller scale processes that lead to diagnostic features identifiable in volcanic deposits. Reprinted with kind permission from Blackwell Publishing Ltd

ing to produce a homogeneous flow (ρ^* approaches 1) or overtaking, leading to flow stratification (ρ^* approaches 0). This suggests that buoyancy-contrast controls separation, whilst density-contrast controls mixing. These observations give insights into the origin of stratification and grain size contrast in pyroclastic flow deposits and allow constraint of flow conditions from field observation.

Caldera Collapse The highest VEI eruptions result from the ejection of large volumes of magma. As magma emerges, country rocks overlying the magma storage area subside, resulting in the formation of a caldera. The caldera roof may fail in a number of modes. Roche and Druitt [103] carried out a force balance calculation for a caldera roof to fail as a coherent single piston, predicting that $\Delta P/\tau_c \geq 4R$, where ΔP is the pressure below lithostatic at which failure occurs, τ_c is the rupture shear stress and R is the ratio of roof thickness to diameter. Experiments were carried out using sand or a slightly cohesive sand-flour mixture as an analogue to country rock, and a draining reservoir of silicone fluid to simulate removal of magma. Comparison of experiment and theory (Fig. 32) shows that coherent roof collapse did occur in line with theoretical prediction. The experiments also showed two other more complex modes of collapse that occur at roof aspect ratios larger than the coherent collapse field, and that these collapse mechanisms may take considerable time to complete. Comparison of the experimental data with known parameters for seven calderas suggested that four had collapsed as coherent pistons, with the other collapses being non-coherent.

Experiments have helped to constrain possible volumes of caldera forming eruptions from pre-caldera conditions [44], showing a potentially powerful route of predicting the time and size of caldera-forming eruptions. Experiments were carried out using a bed of quartz sand as a rock analogue and a deflating water-filled latex balloon to represent the magma body. Extensive arguments are presented to justify application of the experimental results at volcanic scales. Coherent piston collapse was observed at low roof aspect ratios (roof thickness/diameter), becoming more non-coherent as aspect ratio increased. These experiments showed different fault structures and demonstrated the complex 3D nature of the faults that result from caldera collapse (Fig. 33). Fault development was monitored as the balloon was drained to establish the sequence of events prior to, and during, caldera collapse. The relationship between the proportion of water drained from the balloon to make collapse inevitable (f_{crit}) and the aspect ratio of the roof (r) is empirically given as $f_{crit} = 22.4ln(r) + 43.5$ (Fig. 33). Data from caldera-forming eruptions were compared to the experimental trend and found to be in reasonable agreement. This suggests that the small-scale experiments mimic the first-order behavior of much larger-scale caldera collapse, but that there are issues of field measurement and second order process that require further investigation. Nevertheless, this work demonstrates the potential of be-

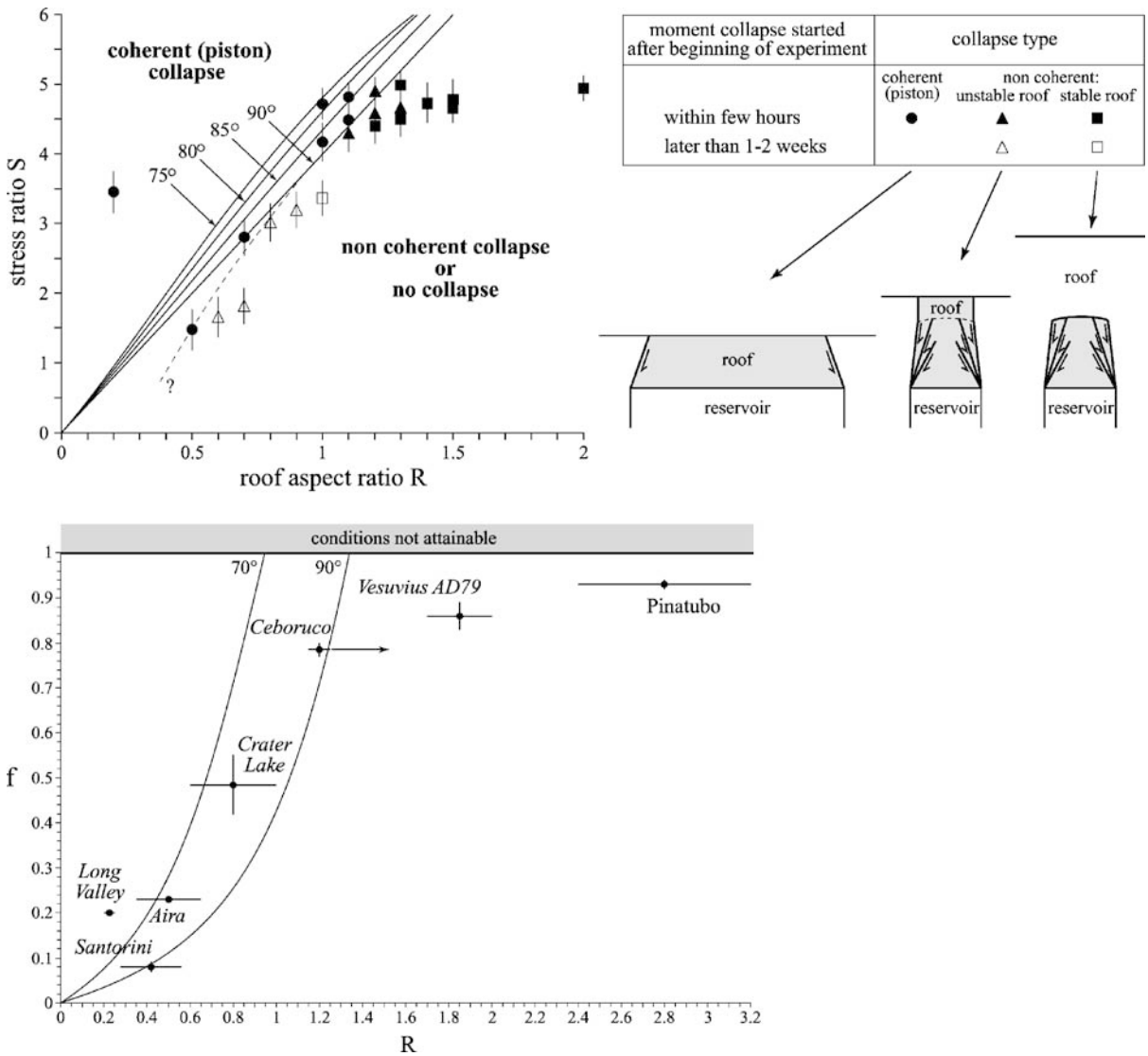

Volcanic Eruptions, Explosive: Experimental Insights, Figure 32

Large VEI eruptions may develop into caldera collapse as the country rock overlying the magma body subsides with pressure drop in the magma. Roche and Druitt [103] experimentally verified a model of coherent, piston-like formation of a caldera (*upper left panel*). Different modes of collapse were also found experimentally, some of which required significant time to elapse (*upper right panel*). This suggests that the roof of a deep, narrow magma body may not collapse simultaneously with eruption, but some time afterwards. Comparison of experimentally verified theory and estimates for proportion of magma erupted (f) and roof aspect ratio (R) in natural caldera collapses (*bottom panel*) indicate that coherent collapse does occur. Reprinted with kind permission from Elsevier

ing able to predict the onset and volume of caldera collapse from combining field observation and laboratory experiment.

Future Directions

Experiments investigating processes that could occur both during, and as a consequence of, explosive volcanic eruptions are many and varied. All the experiments provide insight into volcanic processes because they are all analogous in certain aspects, whether in size in the case of experiments with natural materials, or with analogue materials as well. Curiosity into the fluid dynamics of degassing and expanding multi-phase flows with non-Newtonian and viscoelastic materials will continue to expand knowledge over a wide parameter range of which volcanic processes

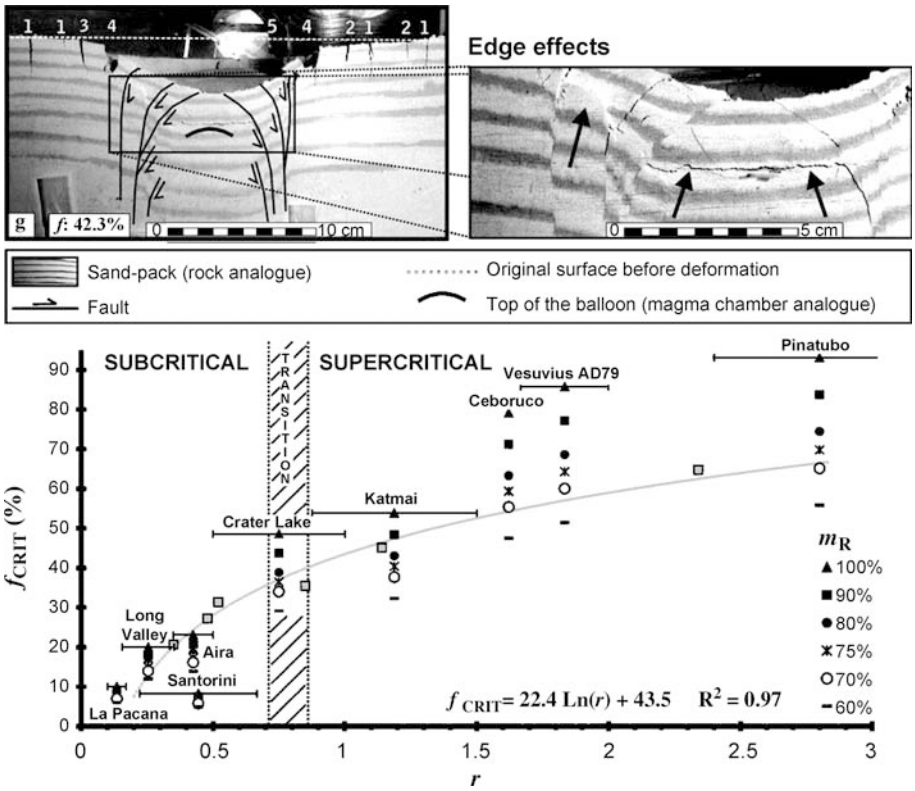

Volcanic Eruptions, Explosive: Experimental Insights, Figure 33

Geyer et al. [44] experimentally ascertained (*top panel*) the critical proportion of a magma chamber that must be erupted for caldera collapse to occur (f_{CRIT}). This was related to the aspect ratio of the roof (r) by an empirical equation. Experimental data (*gray squares*) were compared with values estimated for caldera-forming eruptions (*bottom panel*) for both subcritical (coherent or piston) collapse and supercritical (non-coherent) collapse. The estimated f_{CRIT} values are a function of the percentage of magma actually removable (m_R). This work raises the possibility of being able to forecast the timing and volume of caldera collapse on the basis of previous eruption history and estimates of chamber depth and volume. Reprinted with kind permission from Elsevier

form a small but motivating part. Experiments will form a key part of such research and will continue to inspire numerical models that can provide a data density unavailable experimentally. One future direction could be the development of experiments that use much greater volumes of fluid and run for longer time periods. This would be especially attractive using natural materials, but the apparatus and instrumentation challenge is considerable.

The richness of phenomena that emerge from this complex system suggests that volcanoes are individuals in their eruptive behavior, an observation common to field measurement of volcanoes. Eruption models will increasingly be tested by their ability to account for observed eruption phenomena, and linking of detailed field measurement, volcano-specific experimentation and a unified modeling approach is one of the major current drivers of experimental volcanism. Highly instrumented volcanoes where measurement, in time and space, of such parameters as wide-spectrum ground and atmosphere motion, thermal outputs, chemistry and physics of magma components, electrical effects, flow rates and flow features are required to drive the understanding and forecasting of volcanic events; a task that, in general, becomes more challenging as VEI increases.

The role of experimental models is in providing a real system that can be repeated and measured in some detail. Computer modeling, verified by many experiments, may then be applied at volcanic scale. Quantitative comparison to field observation then tests the model system. Such combined tactics are likely to play a major role in the advancement of volcanology, and it is possible to envisage near real-time interpretation of field measurement in terms of fluid-dynamic, elasto-dynamic, atmospheric and chemical models built on the results of field, numerical

and experimental volcanism. Such an autonomous system would be a major tool in quantifying fluid behavior in volcanic conduits and provide rapid assessment of changes taking place.

Acknowledgments

We thank Bernard Chouet for a highly constructive review of this chapter, William H. K. Lee for editing this section and the staff at Springer including Julia Koerting and Kerstin Kindler.

Bibliography

Primary Literature

1. Alidibirov M, Dingwell DB (1996) Magma fragmentation by rapid decompression. Nature 380:146–149
2. Alidibirov M, Dingwell DB (1996) An experimental facility for the investigation of magma fragmentation by rapid decompression. Bull Volcanol 58:411–416
3. Alidibirov M, Dingwell DB (2000) Three fragmentation mechanisms for highly viscous magma under rapid decompression. J Volcanol Geotherm Res 100:413–421
4. Alidibirov M, Panov V (1998) Magma fragmentation dynamics: experiments with analogue porous low-strength material. Bull Volcanol 59(7):481–489
5. Alidibirov M, Dingwell DB, Stevenson RJ, Hess K-U, Webb SL, Zinke J (1997) Physical properties of the 1980 Mount St. Helens cryptodome magma. Bull Volcanol 59:103–111
6. Anilkumar AV, Sparks RSJ, Sturtevant B (1993) Geological implications and applications of high-velocity two-phase flow experiments. J Volcanol Geotherm Res 56(1–2): 145–160
7. Auger E, D'Auria L, Martini M, Chouet B, Dawson P (2006) Real-time monitoring and massive inversion of source parameters of very long period seismic signals: an application to Stromboli Volcano, Italy. Geophys Res Lett 33(4):L04301
8. Bagdassarov NS, Dingwell DB (1992) A rheological investigation of vesicular rhyolite. J Volcanol Geophys Res 50: 307–322
9. Bagdassarov NS, Dingwell DB (1993) Frequency dependent rheology of vesicular rhyolite. J Geophys Res 98:6477–6487
10. Bagdassarov N, Pinkerton H (2004) Transient phenomena in vesicular lava flows based on laboratory experiments with analogue materials. J Volcanol Geotherm Res 132:115–136
11. Barenblatt GI (2003) Scaling. Cambridge University Press, Cambridge. ISBN 0-521-53394-5
12. Behrens H, Gaillard F (2006) Geochemical aspects of melts: volatiles and redox behavior. Elements 2(5):275–280. ISSN 1811-5209
13. Berthoud G (2000) Vapor explosions. Ann Rev Fluid Mech 32:573–611
14. Birch F, Dane EB (1942) Viscosity. In: Birch F, Schairer JF, Spicer HC (eds) Handbook of physical constants, vol 97. Geological Society of America Special Paper, vol 36, pp 97–173
15. Blackburn EA, Wilson L, Sparks RSJ (1976) Mechanisms and dynamics of Strombolian activity. J Geol Soc Lond 132: 429–440
16. Blower JD (2001) Factors controlling permeability-porosity relationships in magma. Bull Volcanol 63:497–504
17. Blower JD, Keating JP, Mader HM, Phillips JC (2001) Inferring volcanic degassing processes from vesicle size distributions. Geophys Res Lett 28(2):347–350
18. Bottinga Y, Weill DF (1972) The viscosity of magmatic silicate liquids: a model for calculation. Am J Sci 272:438–475
19. Burgisser A, Gardner JE (2005) Experimental constraints on degassing and permeability in volcanic conduit flow. Bull Volcanol 67:42–56
20. Burgisser A, Bergantz GW, Breidenthal RE (2005) Addressing complexity in laboratory experiments: the scaling of dilute multiphase flows in magmatic systems. J Volcanol Geotherm Res 141:245–265
21. Burton M, Allard P, Muré F, La Spina A (2007) Magmatic gas composition reveals the source depth of slug-driven Strombolian explosive activity. Science 317:227–230
22. Büttner R, Dellino P, Zimanowski B (1999) Identifying magma-water interaction from the surface features of ash particles. Nature 401(6754):688–690
23. Büttner R, Roder H, Zimanowski B (1997) Electrical effects generated by experimental volcanic explosions. Appl Phys Lett 70(14):1903–1905
24. Büttner R, Zimanowski B, Roder H (2000) Short-time electrical effects during volcanic eruption: Experiments and field measurements. J Geophys Res 105(B2):2819–2827
25. Cagnoli B, Barmin A, Melnik O, Sparks RSJ (2002) Depressurization of fine powders in a shock tube and dynamics of fragmented magma in volcanic conduits. Earth Planet Sci Lett 204(1–2):101–113
26. Cambridge Polymer Group (2002) The Cambridge Polymer Group Silly Putty™ "Egg". http://www.campoly.com/documents/appnotes/sillyputty.pdf, accessed 10/10/08
27. Carey SN, Sigurdsson H (1982) Influence of particle aggregation on deposition of distal tephra from the may 18, 1980, eruption of Mount St-Helens volcano. J Geophys Res 87(NB8):7061–7072
28. Carroll MR, Holloway JR (eds) (1994) Volatiles in magmas. Rev Miner 30:517. ISBN 0-939950-36-7
29. Chojnicki K, Clarke AB, Phillips JC (2006) A shock-tube investigation of the dynamics of gas-particle mixtures: Implications for explosive volcanic eruptions. Geophys Res Lett 33:L15309. doi:10.1029/2006GL026414
30. Chouet BA, Hamisevicz B, McGetchin TR (1974) Photoballistics of volcanic jet activity at Stromboli, Italy. J Geophys Res 79:4961–4976
31. Chouet B, Dawson P, Ohminato T, Martini M, Saccorotti G, Giudicepietro F, De Luca G, Milana G, Scarpa R (2003) Source mechanisms of explosions at Stromboli Volcano, Italy, determined from moment-tensor inversions of very-long-period data. J Geophys Res 108(B1):2019. doi:10.1029/2002JB001919
32. Chouet B, Dawson P, Nakano M (2006) Dynamics of diffusive bubble growth and pressure recovery in a bubbly rhyolitic melt embedded in an elastic solid. J Geophys Res 111: B07310
33. Chouet B, Dawson P, Martini M (2008) Shallow-conduit dynamics at Stromboli Volcano, Italy, imaged from waveform inversions. In: Lane SJ, Gilbert JS (eds) Fluid motion in volcanic conduits: a source of seismic and acoustic signals. Geological

Society, London. Special Publication 307:57–84, doi:10.1144/SP307.5
34. Clift R, Grace JR, Weber ME (1978) Bubbles, drops and particles. Academic Press, London, p 380. ISBN 012176950X
35. Costa A (2006) Permeability-porosity relationship: a reexamination of the Kozeny–Carman equation based on a fractal pore-space geometry assumption. Geophys Res Lett 33:L02318
36. Dellino P, Zimanowski B, Büttner R, La Volpe L, Mele D, Sulpizio R (2007) Large-scale experiments on the mechanics of pyroclastic flows: design, engineering, and first results. J Geophys Res 112:B04202
37. Dickinson JT, Langford SC, Jensen LC, McVay GL, Kelso JF, Pantano CG (1988) Fractoemission from fused-silica and sodium-silicate glasses. J Vac Sci Technol A 6(3): 1084–1089
38. Dingwell DB (1996) Volcanic dilemma: flow or blow? Science 273(5278):1054–1055
39. Dingwell DB, Webb SL (1989) Structural relaxation in silicate melts and non-Newtonian melt rheology in geologic processes. Phys Chem Miner 16:508–516
40. Donaldson EE, Dickinson JT, Bhattacharya SK (1988) Production and properties of ejecta released by fracture of materials. J Adhes 25(4):281–302
41. Druitt TH, Avard G, Bruni G, Lettieri P, Maez F (2007) Gas retention in fine-grained pyroclastic flow materials at high temperatures. Bull Volcanol 69(8):881–901
42. Eichelberger JC, Carrigan CR, Westrich HR, Price RH (1986) Non-explosive silicic volcanism. Nature 323:598–602
43. Gardner JE (2007) Heterogeneous bubble nucleation in highly viscous silicate melts during instantaneous decompression from high pressure. Chem Geol 236:1–12
44. Geyer A, Folch A, Marti J (2006) Relationship between caldera collapse and magma chamber withdrawal: an experimental approach. J Volcanol Geotherm Res 157(4):375–386
45. Gilbert JS, Lane SJ (1994) The origin of accretionary lapilli. Bull Volcanol 56(5):398–411
46. Gladstone C, Ritchie LJ, Sparks RSJ, Woods AW (2004) An experimental investigation of density-stratified inertial gravity currents. Sedimentology 51(4):767–789
47. Gonnermann HM, Manga M (2003) Explosive volcanism may not be the inevitable consequence of magma fragmentation. Nature 42:432–435
48. Grunewald U, Zimanowski B, Büttner R, Phillips LF, Heide K, Büchel G (2007) MFCI experiments on the influence of NaCl-saturated water on phreatomagmatic explosions. J Volcanol Geotherm Res 159:126–137
49. Hess K-U, Dingwell DB (1996) Viscosities of hydrous leucogranitic melts: a non-Arrhenian model. Am Miner 81:1297–1300
50. Hill LG, Sturtevant B (1990) An experimental study of evaporation waves in a superheated liquid. In: Meier GEA, Thompson PA (eds) Adiabatic waves in liquid-vapor systems. IUTAM Symposium Göttingen, Germany. Springer, Berlin, pp 25–37. ISBN 3-540-50203-3
51. Hoover SR, Cashman KV, Manga M (2001) The yield strength of subliquidus basalts – experimental results. J Volcanol Geotherm Res 107(1–3):1–18
52. Hurwitz S, Navon O (1994) Bubble nucleation in rhyolitic melts: experiments at high pressure, temperature, and water content. Earth Planet Sci Lett 122(3–4):267–280
53. Ichihara M, Rittel D, Sturtevant B (2002) Fragmentation of a porous viscoelastic material: implications to magma fragmentation. J Geophys Res 107(B10):2229. doi:10.1029/2001JB000591
54. Ida Y (2007) Driving force of lateral permeable gas flow in magma and the criterion of explosive and effusive eruptions. J Volcanol Geotherm Res 162(3–4):172–184
55. Ishibashi H, Sato H (2007) Viscosity measurements of subliquidus magmas: alkali olivine basalt from the Higashi–Matsuura district, Southwest Japan. J Volcanol Geotherm Res 160(3–4):223–238
56. James MR, Lane SJ, Gilbert JS (2000) Volcanic plume electrification: experimental investigation of a fracture-charging mechanism. J Geophys Res 105(B7):16641–16649
57. James MR, Gilbert JS, Lane SJ (2002) Experimental investigation of volcanic particle aggregation in the absence of a liquid phase. J Geophys Res 107(B9):2191
58. James MR, Lane SJ, Gilbert JS (2003) Density, construction, and drag coefficient of electrostatic volcanic ash aggregates. J Geophys Res 108(B9):2435
59. James MR, Lane SJ, Chouet B, Gilbert JS (2004) Pressure changes associated with the ascent and bursting of gas slugs in liquid-filled vertical and inclined conduits. J Volcanol Geotherm Res 129(1–3):61–82
60. James MR, Lane SJ, Chouet BA (2006) Gas slug ascent through changes in conduit diameter: laboratory insights into a volcano-seismic source process in low-viscosity magmas. J Geophys Res 111:B05201. doi:10.1029/2005JB003718
61. James MR, Lane SJ, Corder SB (2008) Modelling the rapid near-surface expansion of gas slugs in low-viscosity magmas. In: Lane SJ, Gilbert JS (eds) Fluid motion in volcanic conduits: a source of seismic and acoustic signals. Geological Society, London. Special Publication 307:147–167, doi:10.1144/SP307.9
62. Jaupart C, Allegre CJ (1991) Gas content, eruption rate and instabilities of eruption regime in silicic volcanoes. Earth Planet Sci Lett 102:413–429
63. Jaupart C, Vergniolle S (1988) Laboratory models of Hawaiian and Strombolian eruptions. Nature 331(6151):58–60
64. Jaupart C, Vergniolle S (1989) The generation and collapse of a foam layer at the roof of a basaltic magma chamber. J Fluid Mech 203:347–380
65. Kalinichev AG (2001) Molecular simulations of liquid and supercritical water: thermodynamics, structure, and hydrogen bonding. Rev Miner Geochem 42:83–129
66. Kaminski E, Jaupart C (1998) The size distribution of pyroclasts and the fragmentation sequence in explosive volcanic eruptions. J Geophys Res 103(B12):29759–29779
67. Klug C, Cashman KV (1996) Permeability development in vesiculating magmas: implications for fragmentation. Bull Volcanol 58:87–100
68. Kouchi A, Tsuchiyama A, Sunagawa I (1986) Effect of stirring on crystallisation kinetics of basalt: texture and element partitioning. Contrib Miner Petrol 93:429–438
69. Kueppers U, Perugini D, Dingwell DB (2006) "Explosive energy" during volcanic eruptions from fractal analysis of pyroclasts. Earth Planet Sci Lett 248(3–4):800–807
70. Lane SJ, Chouet BA, Phillips JC, Dawson P, Ryan GA, Hurst E (2001) Experimental observations of pressure oscillations and flow regimes in an analogue volcanic system. J Geophys Res 106(B4):6461–6476

71. Lane SJ, Phillips JC, Ryan GA (2008) Dome-building eruptions: insights from analogue experiments. In: Lane SJ, Gilbert JS (eds) Fluid motion in volcanic conduits: a source of seismic and acoustic signals. Geological Society, London
72. Lautze NC, Houghton BF (2007) Linking variable explosion style and magma textures during 2002 at Stromboli volcano, Italy. Bull Volcanol 69(4):445–460
73. Liu Y, Zhang Y, Behrens H (2005) Solubility of H_2O in rhyolitic melts at low pressures and a new empirical model for mixed H_2O–CO_2 solubility in rhyolitic melts. J Volcanol Geotherm Res 143:219–235
74. Llewellin EW, Manga M (2005) Bubble suspension rheology and implications for conduit flow. J Volcanol Geotherm Res 143:205–217
75. Llewellin EW, Mader HM, Wilson SDR (2002) The rheology of a bubbly liquid. Proc R Soc Lond A 458:987–1016
76. Lyakhovsky V, Hurwitz S, Navon O (1996) Bubble growth in rhyolitic melts: experimental and numerical investigation. Bull Volcanol 58:19–32
77. Mader HM (1998) Conduit flow and fragmentation. Geol Soc London 145:51–71
78. Mader HM, Zhang Y, Phillips JC, Sparks RSJ, Sturtevant B, Stolper E (1994) Experimental simulations of explosive degassing of magma. Nature 372(6501):85–88
79. Mader HM, Phillips JC, Sparks RSJ, Sturtevant B (1996) Dynamics of explosive degassing of magma: Observations of fragmenting two-phase flows. J Geophys Res Solid Earth 101(B3):5547–5560
80. Mader HM, Brodsky EE, Howard D, Sturtevant B (1997) Laboratory simulations of sustained volcanic eruptions. Nature 388(6641):462–464
81. Marshall JR, Sauke TB, Cuzzi JN (2005) Microgravity studies of aggregation in particulate clouds. Geophys Res Lett 32(11):L11202
82. Martel C, Dingwell DB, Spieler O, Pichavant M, Wilke M (2000) Fragmentation of foamed silicic melts: an experimental study. Earth Planet Sci Lett 178:47–58
83. Martel C, Dingwell DB, Spieler O, Pichavant M, Wilke M (2001) Experimental fragmentation of crystal- and vesicle-bearing silicic melts. Bull Volcanol 63:398–405
84. Mason RM, Starostin AB, Melnik OE, Sparks RSJ (2006) From Vulcanian explosions to sustained explosive eruptions: the role of diffusive mass transfer in conduit flow dynamics. J Volcanol Geotherm Res 153(1–2):148–165
85. McGetchin TR, Settle M, Chouet B (1974) Cinder cone growth modeled after Northeast Crater, Mount Etna, Sicily. J Geophys Res 79:3257–3272
86. McMillan PF (1994) Water solubility and speciation models. In: Carroll MR, Holloway JR (eds) Volatiles in magmas.
87. Melnik O, Sparks RSJ (2002) Dynamics of magma ascent and lava extrusion at Soufriére Hills Volcano, Montserrat. In: Druitt T, Kokelaar P (eds) The eruption of Soufriére Hills Volcano, Montserrat, from 1995 to 1999. Geological Society, London, pp 153–171
88. Moore JG, Peck DL (1962) Accretionary lapilli in volcanic rocks of the western continental United-States. J Geol 70(2): 182–193
89. Mourtada-Bonnefoi CC, Mader HM (2001) On the development of highly-viscous skins of liquid around bubbles during magmatic degassing. Geophys Res Lett 28(8):1647–1650
90. Mourtada-Bonnefoi CC, Mader HM (2004) Experimental observations of the effects of crystals and pre-existing bubbles on the dynamics and fragmentation of vesiculating flows. J Volcanol Geotherm Res 129:83–97. doi:10.1016/S0377-0273(03)00233-6
91. Murase T (1962) Viscosity and related properties of volcanic rocks at 800° to 1400°C. Hokkaido Univ Fac Sci Jour Ser VII 1:487–584
92. Murase T, McBirney AR (1973) Properties of some common igneous rocks and their melts at high-temperatures. Geol Soc Am Bull 84(11):3563–3592
93. Namiki A, Manga M (2005) Response of a bubble bearing viscoelastic fluid to rapid decompression: implications for explosive volcanic eruptions. Earth Planet Sci Lett 236:269–284
94. Namiki A, Manga M (2006) Influence of decompression rate on the expansion velocity and expansion style of bubbly fluids. J Geophys Res 111:B11208
95. Newhall CG, Self S (1982) The explosivity index (VEI) – an estimate of explosive magnitude for historical volcanism. J Geophys Res 87(NC2):1231–1238
96. Parfitt EA (2004) A discussion of the mechanisms of explosive basaltic eruptions. J Volcanol Geotherm Res 134:77–107
97. Phillips JC, Lane SJ, Lejeune A-M, Hilton M (1995) Gum rosin – acetone system as an analogue to the degassing behaviour of hydrated magmas. Bull Volcanol 57:263–268
98. Pinkerton H, Norton G (1995) Rheological properties of basaltic lavas at sub-liquidus temperatures – laboratory and field-measurements on lavas from Mount Etna. J Volcanol Geotherm Res 68(4):307–323
99. Pyle DM (2000) Sizes of volcanic eruptions. In: Sigurdsson H, Houghton B, McNutt SR, Rymer H, Stix J (eds) Encyclopaedia of volcanoes. Academic Press, pp 263–269. ISBN 0-12-643140-X
100. Riedel C, Ernst GGJ, Riley M (2003) Controls on the growth and geometry of pyroclastic constructs. J Volcanol Geotherm Res 127(1–2):121–152
101. Riley CM, Rose WI, Bluth GJS (2003) Quantitative shape measurements of distal volcanic ash. J Geophys Res Solid Earth 108(B10):2504
102. Ripepe M, Ciliberto S, Della Schiava M (2001) Time constraints for modeling source dynamics of volcanic explosions at Stromboli. J Geophys Res Solid Earth 106(B5):8713–8727
103. Roche O, Druitt TH (2001) Onset of caldera collapse during ignimbrite eruptions. Earth Planet Sci Lett 191(3–4):191–202
104. Saar MO, Manga M (1999) Permeability-porosity relationship in vesicular basalts. Geophys Res Lett 26:111–114
105. Schumacher R, Schmincke HU (1995) Models for the origin of accretionary lapilli. Bull Volcanol 56(8):626–639
106. Seyfried R, Freundt A (2000) Experiments on conduit flow and eruption behavior of basaltic volcanic eruptions. J Geophys Res Solid Earth 105:23727–23740
107. Shaw HR (1972) Viscosities of magmatic silicate liquids: an empirical method of prediction. Am Jour Sci 272:870–893
108. Sparks RSJ (1997) Causes and consequences of pressurisation in lava dome eruptions. Earth Planet Sci Lett 150:177–189
109. Sparks RSJ, Bursik MI, Carey SN, Gilbert JS, Glaze LS, Sigurdsson H, Woods AW (1997) Volcanic plumes. Wiley, Chichester. ISBN 0-471-93901-3
110. Spera FJ, Borgia A, Strimple J, Feigenson M (1988) Rheology of melts and magmatic suspensions I. Design and calibration of a concentric cylinder viscometer with application to rhyolitic magma. J Geophys Res 93:10273–10294

111. Spieler O, Alidibirov M, Dingwell DB (2003) Grain-size characteristics of experimental pyroclasts of 1980 Mount St. Helens cryptodome dacite: effects of pressure drop and temperature. Bull Volcanol 65:90–104
112. Spieler O, Kennedy B, Kueppers U, Dingwell DB, Scheu B, Taddeucci J (2004) The fragmentation threshold of pyroclastic rocks. Earth Planet Sci Lett 226:139–148
113. Sugioka I, Bursik M (1995) Explosive fragmentation of erupting magma. Nature 373(6516):689–692
114. Suzuki T (1983) A theoretical model for dispersion of tephra. In: Shimozuru D, Yokoyama I (eds) Arc volcanism, physics and tectonics. Terra Scientific publishing company Terrapub, Tokyo, pp 95–113
115. Taddeucci J, Spieler O, Kennedy B, Pompilio M, Dingwell DB, Scarlato P (2004) Experimental and analytical modeling of basaltic ash explosions at Mount Etna, Italy, 2001. J Geophys Res 109:B08203. doi:10.1029/2003JB002952
116. Taddeucci J, Spieler O, Ichihara M, Dingwell DB, Scarlato P (2006) Flow and fracturing of viscoelastic media under diffusion-driven bubble growth: an analogue experiment for eruptive volcanic conduits. Earth Planet Sci Lett 243:771–785
117. Taddeucci J, Scarlato P, Andronico D, Cristaldi A, Büttner R, Zimanowski B, Küppers U (2007) Advances in the study of volcanic ash. Eos Trans AGU 88(24):253
118. Takeuchi S, Nakashima S, Tomiya A (2008) Permeability measurements of natural and experimental volcanic materials with a simple permeameter: Toward an understanding of magmatic degassing process. J Volcanol Geotherm Res doi:10.1016/j.jvolgeores.2008.05.010
119. Takeuchi S, Nakashima S, Tomiya A, Shinohara H (2005) Experimental constraints on the low gas permeability of vesicular magma during decompression. Geophys Res Lett 32: L10312
120. Textor C, Graf HF, Longo A, Neri A, Ongaro TE, Papale P, Timmreck C, Ernst GG (2005) Numerical simulation of explosive volcanic eruptions from the conduit flow to global atmospheric scales. Ann Geophys 48(4–5):817–842
121. Trigila R, Battaglia M, Manga M (2007) An experimental facility for investigating hydromagmatic eruptions at high-pressure and high-temperature with application to the importance of magma porosity for magma-water interaction. Bull Volcano 69:365–372
122. Tuffen H, Dingwell D (2005) Fault textures in volcanic conduits: evidence for seismic trigger mechanisms during silicic eruptions. Bull Volcanol 67:370–387
123. Walker GPL, Wilson L, Bowell ELG (1971) Explosive volcanic eruptions. 1. Rate of fall of pyroclasts. Geophys J Royal Astron Soc 22(4):377–383
124. Watson EB (1994) Diffusion in volatile-bearing magmas. In: Carroll MR, Holloway JR (eds) Volatiles in magmas.
125. Wilson L, Huang TC (1979) Influence of shape on the atmospheric settling velocity of volcanic ash particles. Earth Planet Sci Lett 44(2):311–324
126. Wohletz KH (1983) Mechanisms of hydrovolcanic pyroclast formation: grain-size, scanning electron microscopy, and experimental studies. J Volcanol Geotherm Res 17(1–4): 31–63
127. Woolley AR, Church AA (2005) Extrusive carbonatites: a brief review. Lithos 85(1–4):1–14
128. Yokoyama I (2005) Growth rates of lava domes with respect to viscosity of magmas. Ann Geophys 48(6):957–971
129. Zhang YX (1998) Experimental simulations of gas-driven eruptions: kinetics of bubble growth and effect of geometry. Bull Volcanol 59(4):281–290
130. Zhang Y (1999) A criterion for the fragmentation of bubbly magma based on brittle failure theory. Nature 402:648–650
131. Zhang Y, Behrens H (2000) H_2O diffusion in rhyolitic melts and glasses. Chem Geolog 169:243–26
132. Zhang YX, Sturtevant B, Stolper EM (1997) Dynamics of gas-driven eruptions: experimental simulations using CO_2-H_2O-polymer system. J Geophys Res Solid Earth 102(B2): 3077–3096
133. Zhang Y, Xu Z, Liu Y (2003) Viscosity of hydrous rhyolitic melts inferred from kinetic experiments: a new viscosity model. Am Miner 88:1741–1752
134. Zimanowski B, Lorenz V, Frohlich G (1986) Experiments on phreatomagmatic explosions with silicate and carbonatitic melts. J Volcanol Geotherm Res 30(1–2):149–153
135. Zimanowski B, Büttner R, Lorenz V, Häfele H-G (1997) Fragmentation of basaltic melt in the course of explosive volcanism. J Geophys Res 102(B1):803–814

Books and Reviews

Barnes HA (1997) Thixotropy – a review. J Non-Newtonian Fluid Mech 70:1–33

Barnes HA (1999) The yield strength – a review or 'παντα ρει' everything flows? J Non-Newtonian Fluid Mech 81:133–178

Chhabra RP (2006) Bubbles, drops, and particles in non-Newtonian fluids. CRC Press, London, p 771. ISBN 0824723295

Chouet B (2003) Volcano seismology. Pure and Applied Geophysics 160:739–788

De Kee D, Chhabra RP (2002) Transport processes in bubbles, drops, and particles. Taylor and Francis, p 352. ISBN 1560329068

Fan L-S, Zhu C (1998) Principles of gas-solid flows. Cambridge University Press, Cambridge, p 575. ISBN 0521581486

Gilbert JS, Sparks RSJ (1998) The physics of explosive volcanic eruptions. The Geological Society, London. ISBN 1-86239-020-7

Gonnermann HM, Manga M (2007) The fluid mechanics inside a volcano. Ann Rev Fluid Mech 39:321–356

Henderson G, Calas G, Stebbins J (eds) (2006) Glasses and melts: linking geochemistry and materials science. Elements 2(5):257–320. ISSN 1811–5209

Kaminski E, Tait S, Carazzo G (2005) Turbulent entrainment in jets with arbitrary buoyancy. J Fluid Mech 526:361–376

Mather TA, Harrison RG (2006) Electrification of volcanic plumes surveys. Geophysics 27(4):387–432

Moretti R, Richet P, Stebbins JF (2006) Physics, chemistry and rheology of silicate melts and glasses. Chem Geol 229(1–3): 1–226

Parfitt L, Wilson L (2008) Fundamentals of physical volcanology. Blackwell, p 256. ISBN 0632054433

Sigurdsson H, Houghton B, McNutt SR, Rymer H, Stix J (eds) (2000) Encyclopaedia of volcanoes. Academic Press. ISBN 0-12-643140-X

Tong L-S, Tang YS (1997) Boiling heat transfer and two-phase flow. Taylor and Francis, p 542. ISBN 1560324856

Webb S (1997) Silicate melts: relaxation, rheology, and the glass transition. Rev Geophys 35(2):191–218

Wood CA (1980) Morphometric evolution of cinder cones. J Volcanol Geotherm Res 7:387–413

Volcanic Eruptions: Stochastic Models of Occurrence Patterns

MARK S. BEBBINGTON
Massey University, Palmerston North, New Zealand

Article Outline

Glossary
Definition of the Subject
Introduction
Data
Temporal Models
Volcanic Regimes
Spatial Aspects
Yucca Mountain
Interactions with Earthquakes
Future Directions
Bibliography

Glossary

Onset The beginning of an eruption. The term **event** will be used interchangeably.

Repose Periods during which an eruption is not in progress.

Onset time Time at which an eruption begins. The ith onset time will be denoted t_i, for $i = 1, \ldots, n$.

Inter-onset time Time between successive onsets, denoted by $r_i = t_{i+1} - t_i$, for $i = 1, \ldots, n-1$. In many papers this is termed the **repose time**, which we shall use interchangeably, although the latter is strictly the time between the *end* of one eruption and the *start* of the next.

Polygenetic vent Site of multiple events, in contrast to **monogenetic vents**, at which only one eruption occurs. The latter occur predominately in **volcanic fields**.

Absolute time is denoted by t or s, with the latter usually being the time of the last known event. The **elapsed time** since the last known event is denoted by u, or, equivalently, $t - s$.

A parameter estimate of the parameter λ, say, is denoted by $\hat{\lambda}$.

Definition of the Subject

Depending on the definition, there are approximately 1300 Holocene (last 10,000 years) active volcanoes [83], some 55–70 of which are typically active in any given year. Being tectonic phenomena, they are spatially over-represented along littorals and in island arcs where populations congregate, and millions of lives are potentially at risk. Hence the quantitative forecasting of hazard has long been a key aim of volcanology. This requires models to quantify the likelihood of future outcomes, which may involve the size, location and/or style of eruption, in addition to temporal occurrence. While a wide range of complex deterministic models exist to model various volcanic processes, these provide little in the way of information about future activity. Being the (partially) observed realization of a complex system, volcanological data are inherently stochastic in nature, and need to be modeled using statistical models. To paraphrase Decker [25] concerning potentially active volcanoes, geologists can find out what did happen, geophysicists can determine what is happening, but to find out what might happen requires statistics. Reminding ourselves that "all models are wrong, but some models are useful", the most useful models for volcanic eruptions are ones that provide new insight into the physical processes, and can in turn be informed by our present understanding of those physical processes.

Introduction

The hazard is a product of many elements, including the likelihood of an eruption, the distribution of the eruption size, and the style of the eruption. Depending on the nature of the volcanic products, climatological or geographical factors may also play a part.

The style of an eruption strongly determines the relative risks to life and property. Effusive eruptions of molten lava pose little threat to life with, for example, only one life being lost due to a volcanic eruption on Hawaii in the 20th century. However, in the same time period 5% of the island has been covered by new lava flows [25]. Explosive eruptions, in particular pyroclastic flows, travel at high speeds and can cause extensive fatalities such as the 29,000 killed in the Mont Pelee (West Indies) eruption of 1902. Lahars, or mudflows, such as those that killed 25,000 people in the Nevado del Ruiz (Columbia) eruption of 1985, are also a major danger.

Forecasting volcanic eruptions is easier than forecasting earthquakes on several grounds. First, at least in the case of polygenetic volcanoes, there need be no spatial element involved. Second, it is possible using geological techniques to look backward (often a long way, depending on geography and finances) to see what happened in the past. Thirdly, and of most importance for short-term warning, magma intrusions can be detected by seismic networks, although not all seismic signals indicate magma of course, and other monitoring techniques [25].

In this article we will limit our consideration to models of eruption occurrence, omitting techniques for fore-

casting the nature and effect of the eruption. As the track record of a potentially active volcano provides the best method of assessing its future volcanic hazards on a long-term basis [21,25], we will first briefly review the provenance and characteristics of the data available. Various taxonomies for stochastic models are possible. As the majority of models applied to volcanoes are purely temporal, and this also includes the simplest models, we will look at these first. Such models can be further broken down into (nonhomogeneous) Poisson processes, which consider absolute time relative to some fixed origin, and renewal processes, which consider only the time since the previous event. Markov processes, where the state of the process moves between various eruptive and repose states, and models based on eruptive volume, are the simplest ways of incorporating some element of eruption size into the model. Fractals have been used as a descriptive technique for volcanoes, and are reviewed next. The above models, with the exception of the nonhomogeneous Poisson process, all assume that the volcano is in steady state, i. e., that the activity pattern does not change over time. For volcanoes with lengthy records, this appears to be an untenable assumption over longer time scales [99]. This raises the issues of, first, how to detect a departure from stationary stochastic behavior, or change in volcanic regime and, secondly, how to model it, which are discussed in Sect. "Volcanic Regimes", illustrated with results from Mount Etna. While purely temporal models are suitable for polygenetic volcanoes, monogenetic volcanic fields require a spatial element, to model the location, as well as the time, of future events. This is usually estimated using kernel smoothing, but in cases where the location can be treated as a discrete variable, such as summit versus flank eruptions, Markov chain techniques can be used. These are reviewed, along with methods for detecting clustering and linear alignments, in Sect. "Spatial Aspects". A very high profile, particularly in the USA, exercise in volcanic hazard estimation concerns the risk to a proposed high-level radioactive waste repository planned for Yucca Mountain. The history of published hazard estimates is used in Sect. "Yucca Mountain" to illustrate the application of both temporal and spatio-temporal models. The final substantive section looks at methods for detecting links between volcanic eruptions and large earthquakes. Determining whether correlation in the occurrence patterns exists is a necessary precursor to formulating models to take advantage of any potential new information from the additional process.

The focus of this article is squarely on the techniques. Hence the applications jump around a bit, as they are those presented in the original paper(s) dealing with the technique. The examples of Mount Etna and Yucca Mountain were selected for more detailed examination partly on the basis that many, somewhat contradictory, results exist. Different models make different assumptions, and vary in how much information they can extract from data. In addition, the data used often varies from study to study, and the sensitivity of models to data is important, but too often ignored. Subjective judgments on the merits of various techniques is minimized, except for comments on details necessary to their application.

Data

Information about past eruptions forms the basis of any stochastic model for future behavior. Using such data poses a number of challenges because of its inhomogeneity, particularly in the longer records. This inhomogeneity is due to incomplete observation, which becomes more incomplete the further back one goes, and by variations in dating and measurement precision, again becoming less precise the further into the past one looks. We will now outline the nature and limitations of the available information.

There are several functional distinctions in the types of data used in volcanic hazard models. The most fundamental is at the question of scale. Certain volcanoes, notably Stromboli (which lends its name to the behavior) and Soufriere Hills, erupt in distinct events separated by minutes or hours. Hence a catalog of hundreds of events may only cover a few days of activity. More commonly, eruptions are events with durations ranging from days to years. Eruptions are usually modeled simply as a point process of onsets, but a few studies consider duration, volume, or spatial extent.

A global catalog of Holocene volcanism is maintained by Seibert and Simkin [82]. This gives onset dates, of various precision, and Volcanic Explosivity Index (VEI) for all known eruptions. The VEI, as a logarithmic measuring scale of eruption size analogous to the magnitude of an earthquake, was proposed by Newhall and Self [68]. It is assigned to historical eruptions on the basis of (in decreasing order of reliability): explosion size; volume of ejecta; column height, classification ('Hawaiian', 'Strombolian', 'Vulcanian', '(Ultra-)Plinian'); duration; most explosive activity type; tropospheric and stratospheric injection. Many historical (i. e., observed and recorded at the time) observations also have a duration, and some an estimate of eruptive volume. Estimates of volume, where they exist, typically contain uncertainties of 10–50% [101] or more [54] and, provided the eruption styles are similar, the duration of an eruption can be used as a proxy [3,67].

The frequency of observed volcanic eruptions has increased over the last few hundred years, although this is largely an artifact of the global population increase [84], which indicates that many of the historical records are incomplete. A number of published catalogs for individual volcanoes are of greater detail, and probably more complete. Some of the better known are Etna [65,67,74] and Vesuvius [11,77] (Italy), Ruapehu (New Zealand) [5], Colima (Mexico) [27,63] and Kilaeua/Mauna Loa (Hawaii) [46].

While it is reasonably easy, although sometimes controversial, to define events in an historical catalog, it is not so easy in pre-historical cases. Ideally events correspond to eruptions, but subsequent geologic processes can obscure, or even obliterate, the evidence of previous eruptions. The identification of events is dependent on looking in the right place. Pre-historical eruptions can be identified from a combination of stratigraphy (which provides an ordering for the eruptions) and dating techniques. Radiocarbon dating is feasible for material within the last 50,000 years, and provides an estimated age and (Gaussian) error given in C^{14} years. See [28] for an example of such a catalog. Because of the variable rate of C^{14} deposition, this has to be converted to calendar years [8], in the process losing the Gaussian distribution (and possibly even becoming disjoint). One way [94] of dealing with the resulting untidiness is to treat the data in Monte Carlo fashion, repeatedly drawing random realizations of the eruption sequence from the distributions (ensuring that the stratigraphic ordering is not violated), and fitting the model to each realization to provide a population of hazard models. Beyond 50,000 years, dating is via a number of other techniques, such as K-Ar radiometric age determinations which, with resolutions of the order of 10^6–10^7 years, are much less precise. Conway et al. [20] take the approach of simulating event dates from a uniform distribution, bounded by the $\pm 2\sigma$ error limits on the dates, in order to determine the uncertainty in the estimated recurrence rate.

Because of the expense, and due to the difficulty of finding suitable material in the sample to date, it is common to date only a portion of the samples. Geomorphology is then used to interpolate or bound the unknown dates. For example, in a single core, a spline can be fitted to the depth and known dates, representing the sediment deposition rate, and the unknown dates calculated from their depths [94], as shown in Fig. 1. This gets much more complicated when there is a spatial element involved [14].

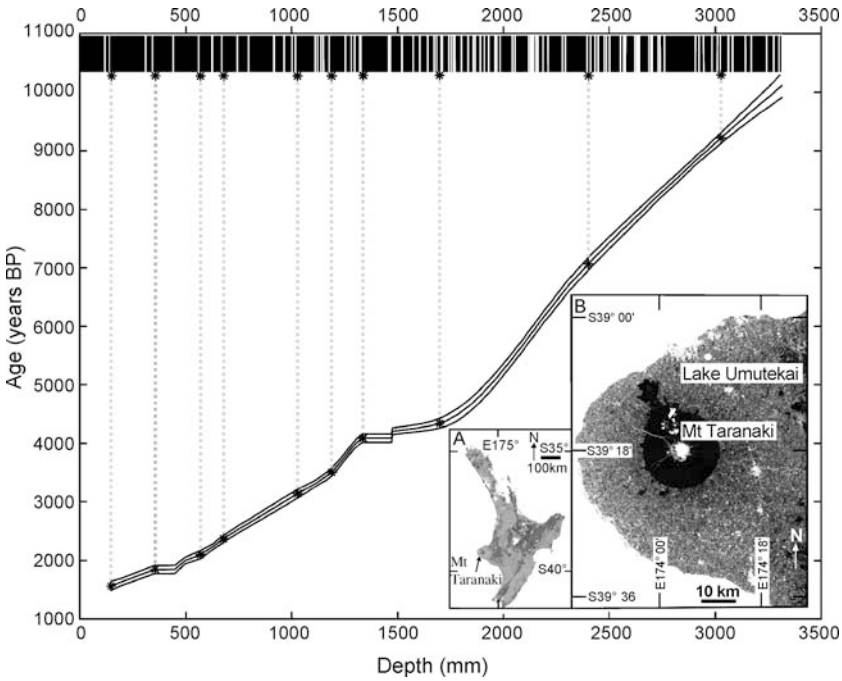

Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 1

Radiocarbon age versus depth model (mm) for tephra layers in the sediment core from Lake Umutekai (*inset A:* North Island of New Zealand and *inset B:* Taranaki – the lake is directly above the "k– in "Lake"). The core stratigraphy is displayed at the top of the plot with white representing tephra deposits. Radiocarbon dated layers are indicated by *dotted lines*

While deposition cores can provide a complete (above a certain minimum size), and ordered tephra record for a given site, the record extracted from another site may only partially overlap. Various criteria of tephra composition and geochemistry can be used to identify common events. Composite records from many sources will almost certainly be incomplete due to missing events, and may well unknowingly contain duplicates of other events.

Temporal Models

Models for the occurrence of volcanic eruptions are based on the idea that the past behavior of a volcano is the best predictor of its future activity. This was dramatically illustrated by Crandell et al. [21] in their somewhat non-quantitative forecast of an eruption of Mount St Helens:

> "The repetitive nature of the eruptive activity at Mount St. Helens during the last 4000 years, with dormant intervals typically of a few centuries or less, suggests that the current quiet period will not last a thousand years. Instead, an eruption is likely within the next hundred years, possibly before the end of this century."

Published in 1975, as a warning that dormant intervals can be a recurring element in the life of a volcano, this foreshadowed the 1980 eruption and subsequent activity.

Let the number of events in a time interval (s, t) be denoted $N(s, t)$, where we use the shorthand $N(t)$ if $s = 0$. The *intensity* of a process is

$$\lambda(t) = \lim_{\Delta t \downarrow 0} \frac{\Pr(N(t, t + \Delta t) = 1)}{\Delta t}, \tag{1}$$

i. e., in the short period of time $(t, t + \Delta t)$, the probability of an event is approximately $\lambda(t)\Delta t$.

Guttorp and Thompson [35] provided a non-parametric method of estimating the intensity, based on the observed onset counts Y_k in intervals of length Δ, indexed by $k = 1, \ldots, K$. They first estimate the reporting probability $p(t)$ by smoothing the observed onset counts (see [86] for an alternative approach assuming a non-decreasing reporting probability), and calculate $Z_k = Y_k / \int_{(k-1)\Delta}^{k\Delta} p(t)\mathrm{d}t$. This is a reconstruction of the underlying process obtained by re-inflating the observed counts by the corresponding observance probability. Assuming that the intensity is locally fairly smooth, an estimate, based on the hanning operation from time-series analysis, is

$$\hat{\lambda}(j\Delta) = \frac{\sum_{i=1}^{K-j-1} Z_i Z_{i+j-1} + 2\sum_{i=1}^{K-j} Z_i Z_{i+j} + \sum_{i=1}^{K-j+1} Z_i Z_{i+j-1}}{4\Delta(K-j)\sum_{k=1}^{K} Z_k / K}. \tag{2}$$

Application to the global catalog, assuming a constant rate of underlying events, produced the estimated observance probability in Fig. 2 that increases slowly from 1500 until the second half of the 18th century, and then more sharply at a constant rate to an assumed probability $p(1980) = 1$.

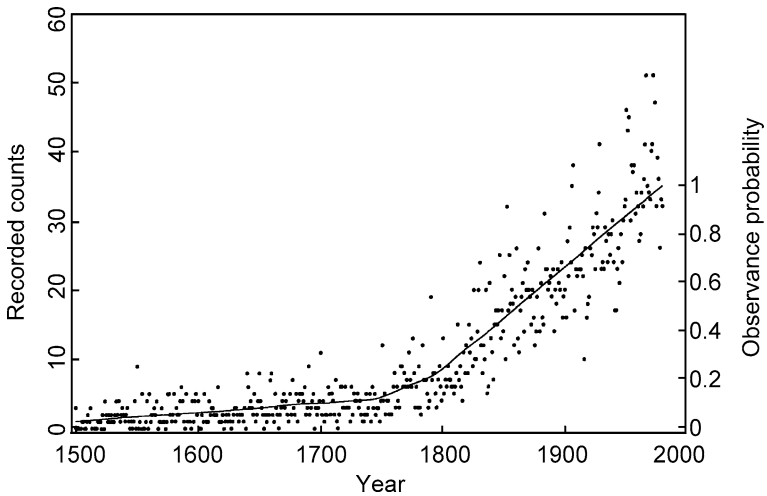

Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 2
Recorded counts (*left-hand scale*) and estimated observance probability (*right-hand scale*) for yearly eruption starts in the Simkin catalog from A.D. 1500 on. Reprinted with permission from the *Journal of the American Statistical Association*. Copyright 1991 by the American Statistical Association

The resulting estimated intensity for global eruption starts did not differ from a Poisson (constant intensity) process. However, in order to see if more structure could be observed in a series of eruptions from a restricted area with homogeneous tectonic structure, the methodology was applied to a catalog of Icelandic eruptions. As expected, the estimated observance probability for the last 200 years was approximately constant, which resulted in a periodicity of about 40 years in the estimated intensity, as shown in Fig. 3.

There have been a few other attempts to discern periodic fluctuations in volcanic activity. Mauk and Johnston [62] examined the correlation of the onset dates of 680 eruptions 1900–1971 with the fortnightly solid earth tide. There was a statistically significant tendency for both andesitic and basaltic eruptions to occur at the maxima of the earth tide. In addition, basaltic eruptions had a significant peak at the minima of the earth tide. A more detailed examination of 18 Japanese volcanoes indicated that this may be related to Bouguer anomalies. Similarly, Martin and Rose [54], in Fig. 4, found that eruption onsets of Fuego Volcano, Guatemala were correlated with the fortnightly lunar tidal maxima, with 23 out of 48 eruptions occurring within ± 2 days of the peak gravity acceleration. Casetti et al. [12] binned eruptions at Mt Etna and examined the resulting statistics, in particular the month of an eruption. Eruptions were shown to be significantly more likely during March–April and November, when there is an increase in the earth-velocity rate. Although Stothers [89] repeated this analysis for a global catalog of eruptions with VEI ≥ 3, 1500–1980, and detected no significant monthly or seasonal variation, Mason et al. [61] found that seasonal fluctuations account for 18% of the historical average monthly eruption rate. The idea of a deterministic cycle is incorporated in a renewal model by Jupp et al. [45], but the result was not fitted to data.

Poisson Processes

The simplest model is a *simple (homogeneous) Poisson process*, which is characterized by $\lambda(t) = \lambda_0$, a constant independent of t. In this case $N(t)$ has a Poisson distribution with mean $\mu = \lambda_0 t$, and the inter-event times have an exponential distribution with mean $1/\lambda_0$. This is sometimes referred to as *Poissonian behavior*. The parameter is estimated as $\hat{\lambda}_0 = N(T)/T$, for an observed history spanning the time interval $(0, T)$.

In order to weaken the requirement that the mean and variance of $N(t)$ be equal, Ho [36] suggested a Bayesian approach, in which the parameter λ_0 has a gamma prior distribution. This leads to $N(t)$ having a negative binomial distribution. Solow [87] augmented this to an empirical Bayes formulation by suggesting that an informative prior could be constructed from the eruption records of a group of similar volcanoes. Bebbington and Lai [6] introduced a generalized negative binomial distribution incorporating

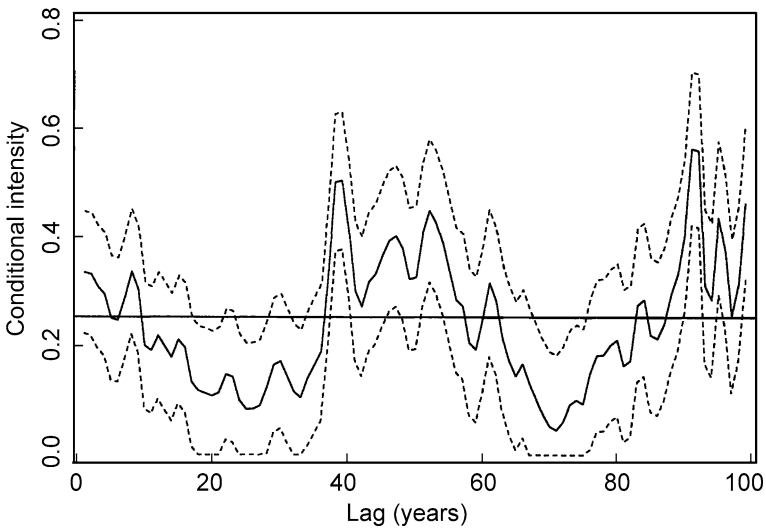

Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 3

Estimated conditional intensity function for Icelandic eruption starts using data from the last 200 years. The *dashed lines* are approximate 95% pointwise confidence bands under the assumption of a homogeneous Poisson process. Reprinted with permission from the *Journal of the American Statistical Association*. Copyright 1991 by the American Statistical Association

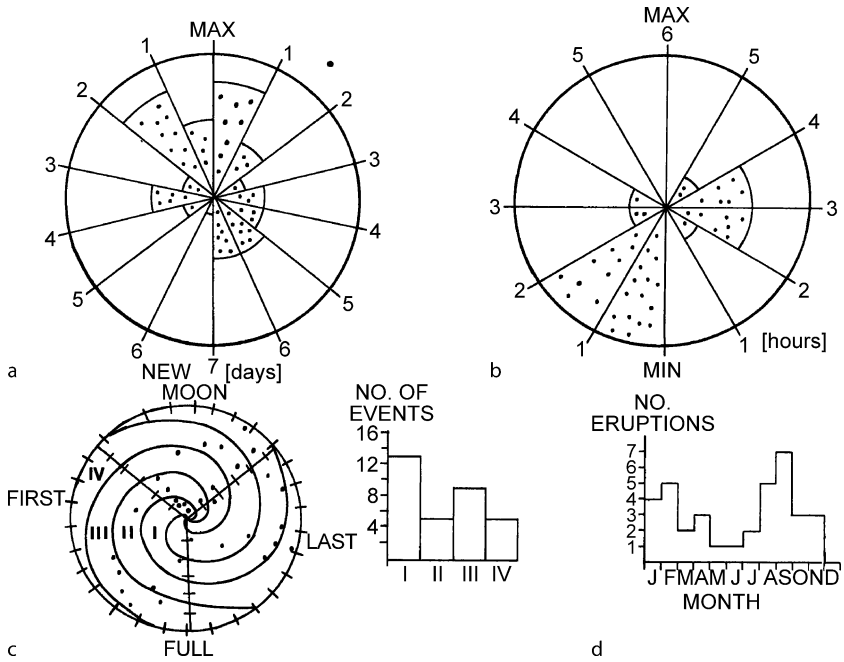

Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 4

a Lambert equal area polar histogram showing the percentage of eruptions of Fuego since 1800 in relationship to the fortnightly tidal maximum of the vertical gravity acceleration. b Lambert equal-area polar histogram showing the percentage of all the known times of the beginning of eruptions since 1957 compared to the 12.4-hour (semi-diurnal) minimum of the vertical gravity acceleration. c A spiral diagram indicating the position of an eruption in relation to the synodic month (days plotted around the circumference) and the reduced anomalistic age (days plotted as the radius, with perigee at the center), in which it occurred. There are 27.5 days in the anomalistic cycle, and 29.6 days in the synodic cycle. Right of the spiral diagram, is a histogram of the four spirals of the above diagram, which are drawn around ±3.5 days of the maxima and minima of the 2 cycles. d Monthly histogram of Fuego's eruptions which have occurred since 1800. Reprinted from [54] © 1981 Elsevier B.V., with permission from Elsevier

serial dependence and applied it to eruptions from Mount Sangay (Ecuador).

The Poisson process is stationary in time, in that the distribution of the number of events in an interval depends only on the length of the interval, not its location. In other words, events occur at the same average rate at all times, i. e., there is no trend in the occurrence rate with time. Marzocchi [55] analyzed possible trends by calculating the (biased) autocorrelation function

$$a(s) = \frac{\sum_{i=1}^{n-1-|s|}(r_i - \bar{r})(r_{i+|s|} - \bar{r})}{\sum_{i=1}^{n-1}(r_i - \bar{r})^2} , \qquad (3)$$

for $s = 0, \pm 1, \ldots \pm m$, where $r_i, i = 1, \ldots, n-1$ is the time between the ith and $(i+1)$th onsets, $\bar{r} = (1/(n-1)) \sum_{i=1}^{n-1} r_i$ and $m < n-1$. Under the null hypothesis that the process is purely random, the distribution of $a(s)$ is asymptotically normal with mean zero and variance $1/(n-1)$ for $s \neq 0$.

More generally, a process can be a *nonhomogeneous Poisson process*, where $N(t)$ has a Poisson distribution with mean $\mu(t) = \int_0^t \lambda(s)\mathrm{d}s$. Ho [37] used an example of a nonhomogeneous Poisson process known as the *Weibull Process* [2], with

$$\lambda(t) = \frac{\beta}{\theta}\left(\frac{t}{\theta}\right)^{\beta-1} , \qquad (4)$$

where t is absolute time with respect to some fixed origin, not the elapsed time since the previous eruption. This includes the homogeneous Poisson process as a special ($\beta = 1$) case, while if $\beta \neq 1$ the process is non-stationary. The intensity (4) is monotonic, and hence can model either an increase or decrease in volcanic activity, but not both. Note that the parameter estimates,

$$\hat{\beta} = N(T)/\sum_{i=1}^{N(T)} \ln(T/t_i) \quad \text{and} \quad \hat{\theta} = T/N(T)^{1/\hat{\beta}} , \qquad (5)$$

for the intensity (4) are sensitive to the position of the time origin [4]. Furthermore, estimates of $\hat{\beta} > 2$, which are quite feasible, indicate a constantly accelerating, or

convex, intensity. Neither of these properties is particularly desirable from a physical viewpoint, and hence the Weibull process is not suitable to model entire volcanic histories. Salvi et al. [73] used a statistical technique to identify that the most recent part of the Mount Etna sequence was non-stationary, and fitted the Weibull process to it. The Weibull process can be tested for goodness of fit to a series of data [4] either via a standard χ^2 test, or by using the fact that $x_1, \ldots, x_{n-1}$, where $x_i = \ln(t_n/t_{n-i})$, should be a random sample from an exponential distribution of unknown mean.

Jaquet et al. [44] proposed a temporal occurrence model using the Cox process. This is a doubly stochastic process with

$$\Pr(N(s, s+t) = k) = \mathbb{E}\left(\frac{Z(s, s+t)^k}{k!} \mathrm{e}^{-Z(s,s+t)}\right), \quad (6)$$

where

$$Z(s, s+t) = \int_s^{s+t} \omega(t)\mathrm{d}t\,, \quad (7)$$

and $\omega(t)$ is the conditional random frequency of events. The process of determining $\omega(t)$ is not spelt out in detail, involving geostatistical methods, simulation, variograms and a conditioning technique. The advantage of (6) over the Poisson process is that it introduces the possibility of correlation in time.

Renewal Processes

The homogeneous Poisson process without trend is a special case of a *renewal process*. In a renewal process, the intervals between events are independent and identically distributed, and so

$$\lambda(t) = \frac{f(t-s;\boldsymbol{\theta})}{1 - F(t-s;\boldsymbol{\theta})}\,, \quad t > s\,, \quad (8)$$

where the most recent event occurred at time $s < t$, and $f = F'$ is a density with parameter vector $\boldsymbol{\theta}$. This was first suggested for eruption onsets by Wickman [103], who coined the term *age-specific eruption rate* for (8). The character of the renewal process is that only the elapsed time since the last eruption controls the time to the next eruption. Previous eruptions exert an influence only through their contribution to the parameter estimates $\hat{\boldsymbol{\theta}}$. A number of tests to check whether a series of eruptions is consistent with a renewal process can be found in the paper by Reyment [72]: One first checks for the existence of a trend using the standard normal statistic

$$U = \frac{\sum_{i=1}^n t_i/n - T/2}{T\sqrt{n/12}}\,, \quad (9)$$

where onsets are observed at times $0 < t_1 < \cdots < t_n < T$. If no trend is observed, one can then examine the serial correlation coefficients

$$\rho_j = \frac{\mathrm{Cov}(\mathbf{r}, \mathbf{r}^{(j)})}{\mathrm{Var}(\mathbf{r})}\,, \quad (10)$$

where $\mathbf{r} = (r_1 \cdots r_{n-1-j})$, $\mathbf{r}^{(j)} = (r_{1+j} \cdots r_{n-1})$, and r_i is the time between the ith and $(i+1)$th onsets. For a renewal process, there should be no correlation, and $\mathrm{Var}(\rho_j) \sim 1/(n-j)$. Hence a test of $\rho_j = 0, j = 1, 2, \ldots$ can be performed using the standard normal statistic $\rho_j\sqrt{(n-j)}$. One can also examine the periodogram [72]. The special case of a Poisson process can be tested in many ways, the easiest of which is to test the inter-onset intervals for exponentiality using, for example, the Kolmogorov–Smirnov statistic of the maximum distance between the empirical and hypothesized distribution functions. Other tests are mentioned in [4,5].

Given a density $f(u;\boldsymbol{\theta})$ and observed inter-onset times r_i, $i = 1, \ldots, n-1$, the parameters $\boldsymbol{\theta}$ can be estimated by maximum likelihood. That is, the values are chosen, either algebraically or numerically, to maximize the likelihood

$$L(r_1, \ldots, r_{n-1}, t-s;\boldsymbol{\theta}) = [1 - F(t-s;\boldsymbol{\theta})]\prod_{i=1}^{n-1} f(r_i;\boldsymbol{\theta})\,, \quad (11)$$

where the first term accounts for the current unfinished interval, and is often omitted. Occasionally other methods are used. The method of moments involves setting the parameter values to equalize the observed and theoretical moments. As many moments are used as one has parameters. For example, two parameters can be fitted using the mean and standard deviation. An alternative is based on minimizing the sum of squared differences between an observed curve and that produced by the model. This is usually done using linear regression, but direct numerical optimization is possible.

If f is the exponential density $f(u) = \nu\mathrm{e}^{-\nu u}$, then

$$\lambda(t) = \frac{\nu\mathrm{e}^{-\nu(t-s)}}{1 - (1 - \mathrm{e}^{-\nu(t-s)})} = \nu\,, \quad (12)$$

and we recover the homogeneous Poisson process. Klein [46] tested the eruptive patterns of Hawaiian volcanoes against a Poisson process (random model) using this formulation, and showed that reposes following large eruptions differed significantly from the whole. The Eq. (12) is sometimes referred to as the memoryless property, as it says that the time elapsed since the last eruption provides no information about the time of the next eruption.

Thorlaksson [93] suggested that an over-dispersed (standard deviation σ greater than the mean μ) sequence of inter-onset intervals should be modeled by the Pareto density $f(u) = a(1+bu)^{-a/b-1}$, $u > 0$, with intensity

$$\lambda(t) = \frac{a}{1+b(t-s)}\,, \quad t > s\,, \tag{13}$$

where $\hat{a} = 2\sigma^2/(\mu^3+\mu\sigma^2)$ and $\hat{b} = (\sigma^2-\mu^2)/(\mu^3+\mu\sigma^2)$ are estimated by the method of moments. This models a decrease in eruption probability with time, and was applied to Colima by Medina Martinez [63]. For under-dispersed inter-onset intervals, Thorlakssson [93] suggested the density

$$f(u) = \begin{cases} (a+bu)\exp(-au-(b/2)u^2) & \sqrt{4/\pi-1} < \sigma/\mu < 1 \\ b(u-a)\exp(-(b/2)(u-a)^2) & \sigma/\mu < \sqrt{4/\pi-1} \end{cases}, \quad u > 0\,, \tag{14}$$

leading to the intensity

$$\lambda(t) = \begin{cases} a+b(t-s) & \sqrt{4/\pi-1} < \sigma/\mu < 1 \\ b(t-s-a)\,H(t-s-a) & \sigma/\mu < \sqrt{4/\pi-1} \end{cases}, \quad t > s\,, \tag{15}$$

where

$$H(x) = \begin{cases} 1 & x > 0 \\ 0 & x \le 0 \end{cases} \tag{16}$$

is the Heaviside function. These imply an increase in eruption probability with time, while the case $\sigma/\mu < \sqrt{4/\pi-1}$ adds a loading time a, and the parameters can be estimated as $\hat{a} = \mu - \sigma/\sqrt{4/\pi-1}$ and $\hat{b} = (2-\pi/2)/\sigma^2$. This latter case was applied to the volcanoes Hekla and Katla, which had ratios σ/μ of 0.44 and 0.43 respectively. Of the other 28 volcanoes examined, none had a ratio less than 0.76, and 23 had a ratio of greater than unity.

Settle and McGetchin [78] fitted a Gaussian density to the inter-onset times of eruptions at Stromboli over four days. Note that the formulation (8) is always strictly positive, even for distributions such as the Gaussian whose support includes the negative axis. However, fitting and simulating such processes presents certain practical difficulties.

Bebbington and Lai [4] fitted renewal models to a number of volcanoes, using the Weibull density with scale parameter β and shape parameter α,

$$f(u) = \alpha\beta(\beta u)^{\alpha-1}\exp(-(\beta u)^{\alpha})\,, \quad u > 0\,. \tag{17}$$

The intensity is then

$$\lambda(t) = \alpha\beta(\beta(t-s))^{\alpha-1}\,, \quad t > s\,. \tag{18}$$

We see that while $\alpha = 1$ is the exponential distribution, $\alpha < 1$ corresponds to an 'over-dispersed', or clustering, distribution (cf. [34]), and $\alpha > 1$ to a more periodic distribution with a mode at $u = \beta^{-1}(1-1/\alpha)^{1/\alpha}$. The density and intensity are shown in Fig. 5.

De la Cruz-Reyna and Carrasco-Nunez [28] suggested using the gamma density, which also allows for clustering and periodicity, but the computational details are more complex. Also, the Weibull is the consequence of a material failure model [98], and in any case, little qualitative difference is usually observed between Weibull and gamma distributions for small samples, such as eruptive records. Note that the intensity (18) is monotonic, and can model either increasing probability of eruption as the repose time increases, such as at Hekla [103], or decreasing probability with increasing repose time, such as at Colima [63]. Bebbington and Lai [4,5] provided a number of tests for assessing the applicability of the Weibull distribution, based on the fact that r_i^{β}, where $r_i = t_{i+1} - t_i$, should be exponentially distributed.

The lognormal density,

$$f(u) = \frac{1}{u\sigma\sqrt{2\pi}}\exp\left[-\frac{(\ln u-\mu)^2}{2\sigma^2}\right]\,, \quad u > 0\,, \tag{19}$$

with intensity

$$\lambda(t) = \frac{\dfrac{\sqrt{2}}{(t-s)\sigma\sqrt{\pi}}\exp\left[-\dfrac{(\ln(t-s)-\mu)^2}{2\sigma^2}\right]}{\operatorname{erfc}\left(\dfrac{\ln(t-s)-\mu}{\sqrt{2}\sigma}\right)}\,, \quad t > s\,, \tag{20}$$

was also considered by Bebbington and Lai [4], and used by Marzocchi and Zacarelli [59] and Eliasson et al. [30]. As the density (19) is unimodal at $u = \exp(\mu-\sigma^2)$, this works better for non-clustering events, such as the very large eruptions considered in [59]. Eliasson et al. [30] noted significant departure from the distribution in one of their models. The density and intensity are shown in Fig. 6. Note that (20) is 'upside-down bathtub shaped', i. e., that as t increases it rises to a maximum and then declines.

The log-logistic density,

$$f(u) = \frac{\eta\gamma u^{\gamma-1}}{(1+\eta u^{\gamma})^2}\,, \quad u > 0\,, \tag{21}$$

with a mode at $u = \max\{0, (\gamma-1)/((\gamma+1)\eta)^{1/\gamma}\}$ was used by Connor et al. [18] to model the intervals between

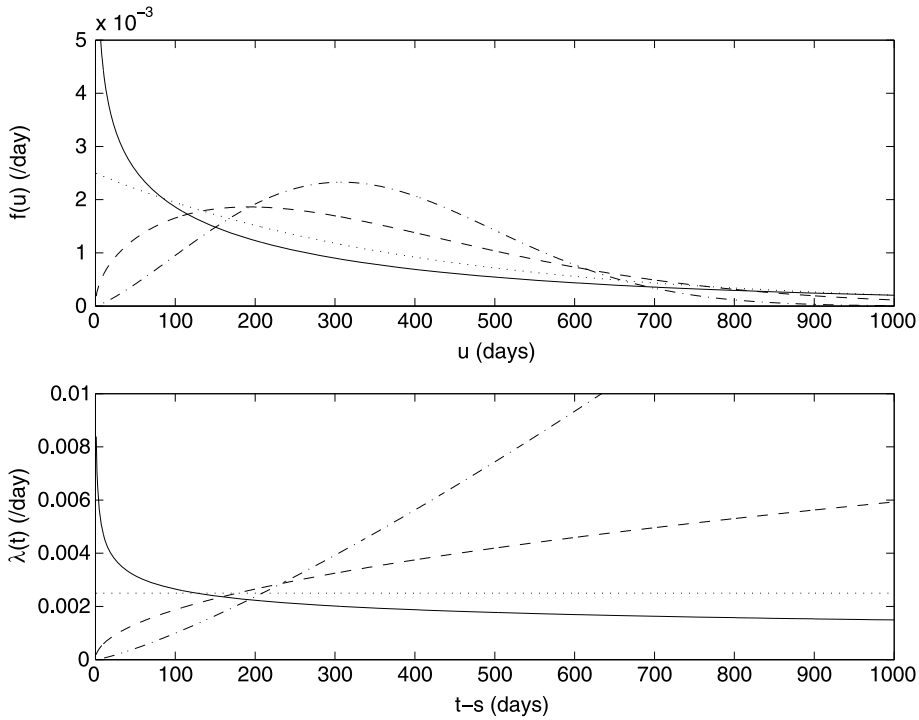

Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 5
Density f and conditional intensity λ for the Weibull distribution with $\beta = 0.0025$ and $\alpha = 0.75$ (*solid line*), $\alpha = 1$ (*dotted line*), $\alpha = 1.5$ (*dashed line*) and $\alpha = 2.25$ (*dot-dash line*)

Vulcanian explosions of Soufriere Hills volcano, Montserrat. The density and the intensity

$$\lambda(t) = \frac{\eta\gamma(t-s)^{\gamma-1}}{1+\eta(t-s)^{\gamma}}\,, \quad t > s\,, \tag{22}$$

are shown in Fig. 7. The intensity (22) has a maximum at $t = s + \max\{0, ((\gamma-1)/\eta)^{1/\gamma}\}$, prior to which it is increasing, subsequently decreasing. Thus we can see that the log-logistic distribution possesses some of the character of the Weibull and of the lognormal distributions.

Pyle [71] used the power-law density

$$f(u \mid u > c) = \frac{u^{-1-1/b}}{bc^{-1/b}}\,, \quad u > c > 0 \tag{23}$$

to model the tail of the inter-onset distribution. Note that the density is not defined for $c = 0$. The intensity is

$$\lambda(t) = \frac{1}{b(t-s)}\,, \quad t > s\,, \tag{24}$$

from which we see that the parameter b controls the rate of decay of the intensity, or equivalently, the 'thickness of the tail'.

Turner et al. [94] fitted a renewal density consisting of a mixture of Weibull densities, as exemplified in Fig. 8, to model a multimodal repose distribution. Cronin et al. [22] proposed a variant on this, where onsets occurred in 'episodes', each of variable numbers of eruptions, with different repose distributions corresponding to inter- and intra-episode reposes.

The drawback of renewal models is that they commonly fail to explain variations in eruption rate, corresponding to changes in activity level, although Wickman [103] did suggest using a small number of different levels. This idea was taken up in [3], where a renewal process with parameters controlled by a hidden Markov model was fitted to flank eruptions from Mt Etna. Using the Weibull distribution (17) for the inter-onset times, the preferred model had long sequences of approximately Poisson behavior, interspersed with shorter sequences of more frequent and regular eruptions.

Markov Processes

A *Markov process* (continuous time Markov chain) $Z(t), t \geq 0$ has transition intensity matrix $Q = (q_{ij})$,

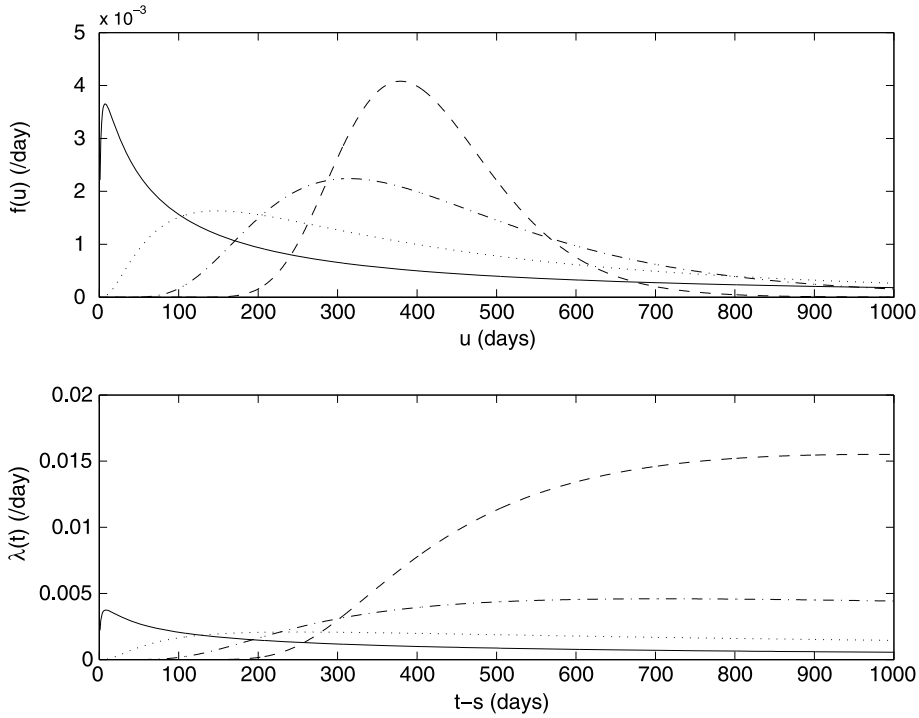

Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 6
Density *f* and conditional intensity λ for the lognormal distribution with $\mu = 6$ and $\sigma = 2$ (*solid line*), $\sigma = 1$ (*dotted line*), $\sigma = 0.5$ (*dashed line*) and $\sigma = 0.25$ (*dot-dash line*)

where

$$q_{ij} = \lim_{t \downarrow 0} \frac{\Pr(Z(s+t) = j \mid Z(s) = i)}{t} \geq 0 \,, \qquad (25)$$

for all $i \neq j$, and $-q_i = q_{ii} = -\sum_{j \neq i} q_{ij}$. Thus the sojurn time in state i is exponential with mean $1/q_i$, and when it ends, the process moves to state j with probability $p_{ij} = q_{ij} / \sum_{k \neq i} q_{ik}$.

Wickman [104] proposed the use of Markov processes to model eruption-repose patterns. This entailed defining a number of eruptive and repose states, with prescribed permitted transitions between them. The initial models [104] allowed for one or two magma chambers, with eruptions and 'solidification' of the vent occurring in Poisson processes of differing rate. In the two-state case this solidification can only be removed by an eruption from the second, larger and deeper, magma chamber. If the two magma chambers erupt at rates ν and μ ($\nu > \mu$), respectively, and the vent solidifies at rate η, then the eruption rate can be derived as

$$\lambda(t) = \frac{\mu\eta + \nu(\mu + \nu + \eta)\exp[-(\nu + \eta)(t - s)]}{\eta + \nu \exp[-(\eta + \nu)(t - s)]} \,, \qquad t > s \,, \qquad (26)$$

where the previous eruption occurred at time s. The parameters can be estimated by numerically fitting this equation to the empirical renewal intensity.

Wickman [105] extended this to six possibilities modeled on various well-known volcano types, and simulated them to produce reasonable appearing synthetic data. The process of fitting these models to observed data has required manual identification of the various states in the data [11,54], but the use of hidden Markov model techniques to automate the process has been suggested [3]. In a Markov process, the time between onsets is the passage time through the eruptive and repose states, which will have a compound exponential distribution, i.e, be the sum of a number of exponential random variables of differing means.

Time and Size Predictable Models

Bacon [1] noted that the intervals between rhyolitic and basaltic eruptions in the Coso Range (California) appeared to be proportional to the volume of the preceding eruption. This *time-predictable* behavior (cf. [81]) is most simply explained by assuming that the rate of magma input is constant, and an eruption occurs when a certain magma level is reached. However, Bacon [1] suggested the covariate of accumulated extensional strain as the controlling

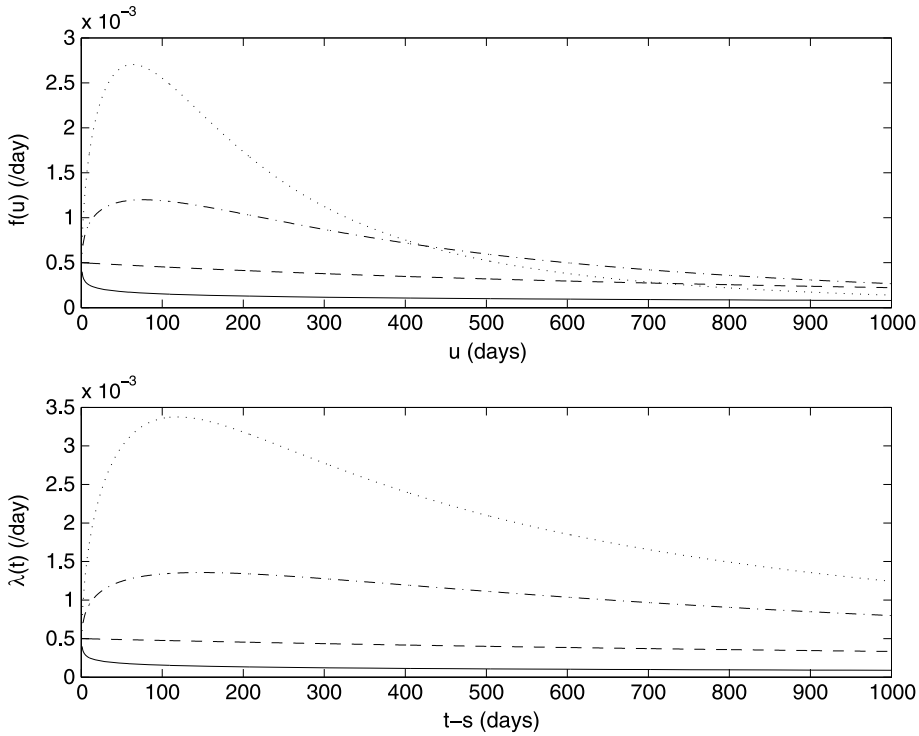

Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 7
Density f and conditional intensity λ for the log-logistic distribution with $\eta = 0.0005$ and $\gamma = 0.8$ (*solid line*), $\gamma = 1$ (*dashed line*), $\gamma = 1.2$ (*dot-dash line*) and $\gamma = 1.4$ (*dashed line*)

variable, and considered eruptions a passive response of the magmatic system to the tectonic stress field.

Similar behavior was observed for Fuego Volcano, Guatemala [54], together with a weaker proportionality between volume and the length of repose before an eruption, although the authors did note that "the volume estimates probably have an error of less than ± 1 order of magnitude". Martin and Rose [54] also examined a possible link between the length of repose before an eruption and its geochemical composition. Santacroce [75] found a similar phenomenon at the Somma-Vesuvius complex, and suggested that eruptions there occur in cycles, of approximately 3000 years, each initiated by a large-scale Plinian eruption, and terminated by a long repose of several centuries. Klein [46] found evidence of time-predictable behavior at both Kilauea and Mauna Loa (Hawaii).

Wadge and Guest [101] suggested that the time-predictable model has a limited predictive ability, based on drawing an empirical envelope on the recent eruptive volume curve. This has a slope determined as the mean eruptive volume rate, and limits defined by the extreme points of the curve in the considered history. This enables a forecast of both the longest the current repose is likely to last, and the maximum volume of the next eruption, provided that steady-state is maintained. Figure 9 shows an example using flank eruptions from Mt Etna 1970–2006. Wadge [99] estimated, however, that volcanoes may spend only 1/4–1/2 of the time in steady-state, being dormant much of the remainder. Changes in eruptive rate have also been found in a number of historical records [9,26,65].

De la Cruz-Reyna [26] proposed a general *load-and-discharge* model, where the 'energy' of the volcano increases at a constant rate σ between eruptions. The ith eruption occurs when this exceeds the threshold H_i, during which the stored energy drops to the threshold L_i, releasing an energy $\xi_i = H_i - L_i$. The interval between eruptions is thus

$$r_i = t_{i+1} - t_i = \frac{H_{i+1} - H_i + \xi_i}{\sigma} \,. \tag{27}$$

Hence successive repose intervals are not independent, as they include a common threshold term, and should be negatively correlated. In general, the thresholds can be random, but this formulation includes the time-predictable model. In this, the thresholds H_i are constant, and

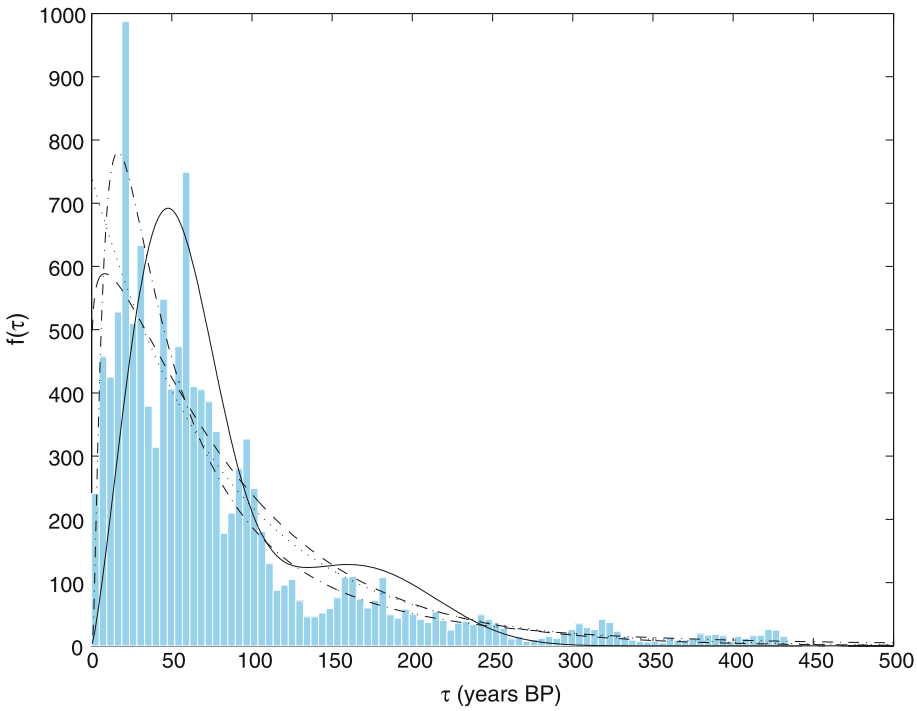

Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 8
Histogram of 122,000 sampled inter-event times for Mt Taranaki based on 1000 Monte Carlo runs. Curves show the fitted densities for this data set: *Dotted line* = exponential distribution, *dot-dash* = lognormal, *dashed* = Weibull, *solid* = mixture of Weibulls

thus the interval

$$r_i = \xi_i/\sigma \tag{28}$$

is proportional to the energy of the preceding eruption, and successive intervals are independent. Alternatively, if the thresholds L_i are held constant ($= L$, say), $H_i = \xi_i - L$ and from (27)

$$r_i = (H_{i+1} - L)/\sigma \ . \tag{29}$$

Thus successive intervals are again uncorrelated, and the energy drop of the eruption is proportional to the length of the interval preceding it. This is termed the *size-predictable* model.

In a global analysis of eruptions, De la Cruz-Reyna [26] noted the existence of a regression relationship $\log \lambda_0 = 3.494 - 0.789\text{VEI}$ with correlation -0.999, where λ_0 is the Poisson intensity for eruptions of a given VEI. This led De la Cruz-Reyna [27] (see also [28]) to model the onsets of different VEI events as independent Poisson processes, in a special case of the size-predictable model. This approach was also used in the quantification of volcanic risk due to Vesuvius by Scandone et al. [76], based on the sequence of equilibrium states (repose, persistent activity, intermediate eruption, final eruption) identified by Carta et al. [11].

Burt et al. [9], in an examination of the basaltic volcano Nyamuragia (Zaire), formulated the 'pressure-cooker' and 'water-butt' models. The former, where an eruption occurs when the magma volume reaches a threshold determined by the rock strength and reservoir shape, is the time-predictable model, while the latter, where eruptions always drain the magma reservoir to the same level, is the size-predictable model. Burt et al. [9] suggested physical explanations for stochastic perturbations observed in these models, and proposed that a volcano could be tested for time-predictability by performing a regression analysis of $(v_i^{(1)})$ on (r_i), where $r_i = t_{i+1} - t_i$, and $t_i, v_i^{(1)}$ are the onset time and eruptive volume of the ith eruption. Similarly the size-predictable model entails a regression analysis of $(v_i^{(2)})$ on (r_i), where $v_i^{(2)} = v_{i+1}^{(1)}$. Sandri et al. [74] generalized the test for time-predictability to a regression analysis of $(\log v_i^{(1)})$ on $(\log r_i)$, so that an estimated slope of b significantly greater than zero implies a time-predictable relation $v_i^{(1)} \propto r_i^b$. Marzocchi and Zaccarelli [59] performed a similar regression analysis of $(\log v_i^{(2)})$ on $(\log r_i)$ for the size predictable model, resulting in a relation $v_i^{(2)} \propto r_i^b$. Be-

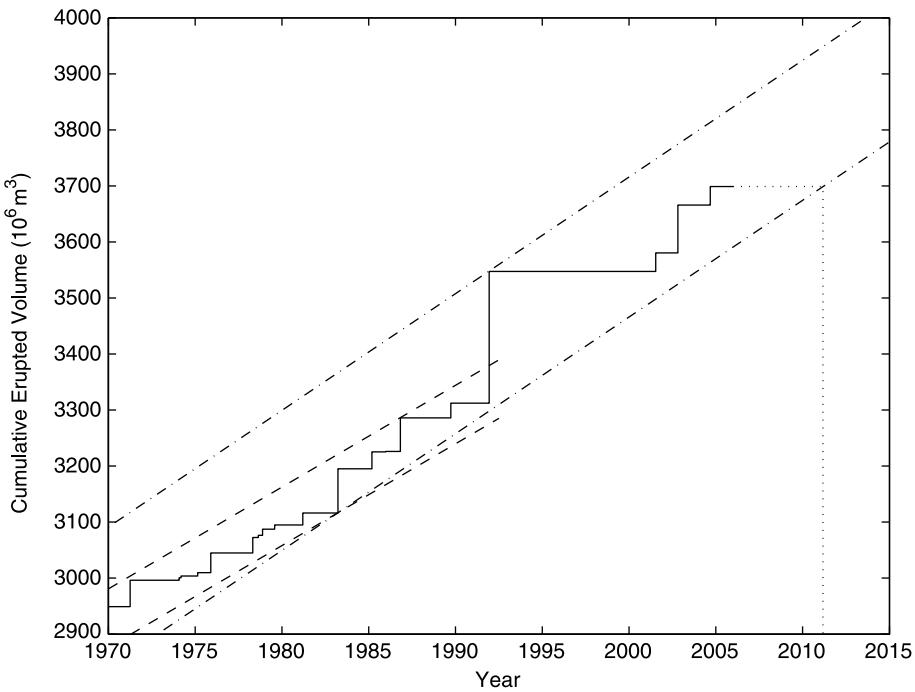

Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 9

Volume-time curve for Mt Etna flank eruptions (*solid line*). *Dashed limits* based on 1970–1990 data were exceeded, or equivalently the volcano departed from steady-state, by the 1991 eruption. *Dot-dash limits* based on 1970–2006 data have a steeper slope and wider limits to accommodate the 1991 eruption. The *dotted lines* show that the method predicts the next eruption to occur by the end of February 2011, and the eruptive volume to be less than $142.1 + 20.8t$ (10^6 m^3), where t is the number of years to the date of the eruption from January 2007

cause both the repose and volume distributions are highly skewed, the logarithmic transformation is advisable to reduce the high leverage of the tail points.

Marzocchi et al. [57] estimated the VEI of the next eruption of Mt Vesuvius by considering a) the global catalog, b) a catalog of 'analog' volcanoes, and c) the catalog of Vesuvius itself. In each case, the catalogs are trimmed to those eruptions following a repose interval of at least the presently observed 60 years. The estimated probability of an eruption of VEI ≥ 5 ranged from 1% to 20%, and could not be considered negligible. This has implications for the Emergency Plan of Mount Vesuvius, which assumes a maximal VEI $= 4$.

Marzocchi and Zaccarelli [59] proposed a renewal model based on the time-predictable model. In their *open conduit system* model, the probability density of the time u to the next eruption is conditional on the previous eruptive volume v. Using a log-normal distribution, this is

$$f(u \mid v) = \frac{1}{u\sigma\sqrt{2\pi}} \exp\left[-\frac{(\log u - (a + bv))^2}{2\sigma^2}\right] , \quad (30)$$

where a, b are the intercept and slope regression coefficients from the regression of $(\log v_i^{(1)})$ on $(\log r_i)$.

Chaos and Fractals

A sufficiently complicated deterministic model is indistinguishable by finite observation from a random process. However, it is possible to construct systems with few degrees of freedom, but very complex dynamics, which are known as *chaotic systems*, or *strange attractors*. This possibility can be tested by examining the correlation properties of the observation sequence. The underlying principle of fractals is one of *self-similarity*, i. e., that the statistical properties of the process are invariant at differing time (or space) scales.

Sornette et al. [88] computed the dimension D of the attractor as a function of the dimension d of the embedding phase space. Given onset times t_i, let $X_i^{(k)} = r_{i+k-1}$ for $k = 1, \ldots, d-1$, where $r_i = t_{i+1} - t_i$, for $i = 1, \ldots, n-1$. Then the correlation function is determined as

$$C_d(w) = \frac{1}{n^2} \#\{i, j\colon \|\mathbf{X}_i - \mathbf{X}_j\| < w\} , \quad (31)$$

where $\mathbf{X}_i$ is the vector $(X_i^{(1)} \cdots X_i^{(d)})$. The correlation dimension is then defined by the power law relation

$$C(w) \sim w^{D(d)} . \quad (32)$$

If this saturates at some value $D_{\max}$ as d increases, this is taken as evidence that the underlying dynamics is deterministic and has dimension of the order of $D_{\max}$. Applications to the volcanoes Piton de la Fournaise (La Reunion Island) and Mauna Loa/Kilauea (Hawaii) produced estimates of $D_{\max}$ of 1.7 and 4.6, respectively. The latter value was interpreted as evidence of two independent dynamical systems for the two Hawaiian volcanoes, contradicting the conclusion of Klein [46]. Sornette et al. [88] also constructed a series of 'return maps' (r_{i+1} as a function of r_i) by truncating the data in various ways.

Dubois and Cheminee [29] used the Cantor dust model for volcanic eruption onsets. In this method, the eruptive record is divided into segments of length T, and the fraction of these intervals containing at least one onset is denoted X. If fractal clustering with fractal dimension D is present, then $X \sim T^{1-D}$. This implies a renewal model with intensity $\lambda(t) = D/(t-s)$, $t > s$, where the last eruption occurred at time s (cf. Eq. (24)). An obvious method for estimating D is a regression analysis of $(\log X_i)$ against $(\log T_i)$ for various interval lengths T_i. Applied to the Piton de la Fournaise (La Reunion Island), two slopes were observed, one with $D \approx 0.5$ for $T < 10$ months, and one with $D \approx 0.7$ for $T > 10$ months. Kilauea had two regimes with $D = 0.58$ for $T < 24$ months and $D = 0.81$ for $T > 24$ months, while Mauna Loa had $D \approx 0.35$ for $T < 6$ years and $D \approx 0.85$ for $T > 6$ years. The two regimes all indicated clustering at short repose lengths, and more regular activity for eruptions separated by longer reposes. This was interpreted as evidence of multiple magma chambers, with different sizes and recharge rates, at different depths. Applied to a combined catalog of flank and summit eruptions of Mt Etna 1950–1987, the result was a single regime with $D = 0.88$. This high value of D was interpreted as being characteristic of the regular eruptive activity of the volcano.

Marzocchi [55] calculated the statistic

$$a_{\max} = \max_{1 \le s \le 5} a(s) \,, \tag{33}$$

where $a(s)$ is defined in (3), although a process might first be de-trended by subtracting an auto-regressive model. A value $a_{\max} > a_c^*$ indicates a chaotic system, while a value $a_{\max} < a_s^*$ indicates a stochastic system. Intermediate values provide insufficient evidence for either. The critical values a_c^* and a_s^* are determined by simulation. Both Vesuvius and Etna evidenced significantly stochastic behavior.

A multifractal analysis using the Grassberger–Procaccia algorithm was performed by Godano and Civetta [33]. The correlation interval is calculated as

$$C(t,q) = \frac{1}{n-1} \sum_{i=1}^{n-1} \left[\frac{1}{n-2} \sum_{j \ne i} H\left(t - \mid r_i - r_j \mid\right) \right]^q \,, \tag{34}$$

where $r_1, \ldots, r_{n-1}$ are the inter-onset times, and $H(\cdot)$ is the Heaviside function. The *generalized dimensions* are then evaluated as

$$D_{q+1} = \phi(q)/q \,, \tag{35}$$

where $\phi(q)$ is the estimated slope from a regression analysis of $\log C(t,q)$ versus $\log t$. The fractal, information, and correlation dimensions are D_0, D_1 and D_2, respectively. More generally, D_{q+1} is constant as q varies for a homogeneous fractal distribution, but varies with q for a multifractal distribution. The latter implies that local clustering properties are different from global clustering properties. Applied to eruptions of Vesuvius, the estimated dimensions were $D_0 = 1.05 \pm 0.07$, $D_2 = 0.75 \pm 0.04 \approx D_\infty = 0.73 \pm 0.02$ and $D_{-\infty} = 1.09 \pm 0.13$. This indicates that at a global scale, the catalog exhibits Poissonian behavior ($D = 1$), and the most intense clustering is not a particularly strong effect. The catalog exhibits only weak multifractal behavior, and this only for $q < 1$. Thus the cyclic behavior outlined by Carta et al. [11] cannot be identified statistically, as the intervals between events in a cycle are not distinguishable from those between cycles.

Nishi et al. [69] examined the long term memory of the time series of 3796 eruptions of Sakurajima 1981–1999 using the Hurst exponent. Let $r_i = t_{i+1} - t_i$, $i = 1, \ldots, n-1$, be the inter-onset times, and set

$$X(k,t,m) = \sum_{i=1}^{k} \left(r_{i+t-1} - \frac{1}{m} \sum_{j=1}^{m} r_{j+t-1} \right) \,. \tag{36}$$

Then the *self-adjust range* is

$$R(t,m) = \max_{1 \le k \le t} X(k,t,m) - \min_{1 \le k \le t} X(k,t,m) \,, \tag{37}$$

which is re-scaled by the standard deviation

$$S(t,m) = \sqrt{\frac{1}{t} \sum_{i=1}^{t} \left(r_{i+m-1} - \frac{1}{m} \sum_{j=1}^{m} r_{j+t-1} \right)^2} \,, \tag{38}$$

and averaging over the length of the time series, the *Hurst exponent* H is then defined by

$$\frac{\sum_{m=1}^{n-1} R(t,m)}{\sum_{m=1}^{n-1} S(t,m)} = \left(\frac{t}{2} \right)^H \,. \tag{39}$$

The value of H can be estimated by a regression analysis of (39), after taking logarithms. The Poisson process produces a value of $H = 0.5$. A value of H greater than this implies persistent behavior, with each value depending on the previous one. Anti-persistence behavior is characterized by $H < 0.5$. The Hurst exponents for Sakurajima were $H = 0.72$ for the whole series, and $H = 0.74$ for a 'high frequency' period 1983–1985. A Monte Carlo test consisting of randomly reordering the observed inter-onset times and recalculating H confirms that these values are significantly greater than 0.5. Telesca and Lapenna [92] performed the same analysis for eruptions of VEI ≥ 0 of 14 volcanoes worldwide. The requirements were a minimum of 40 events during 1800–2000. Seven of the volcanoes, with Hurst exponents ranging from 0.79 to 1.3, were characterized by persistent behavior, while the remainder, with Hurst exponents ranging from 0.60 to 0.73, were not significantly different from a Poisson process.

Telesca et al. [91] examined 35 sequences of VEI ≥ 0 eruptions (minimum 25 eruptions 1800–2000) using the Fano factor method. The Fano factor is defined as the variance of the number of events in a specified interval width Δ divided by the mean,

$$\mathrm{FF}(\Delta) = \frac{(\Delta/T)\sum_{i=1}^{T/\Delta} N_i^2 - \left((\Delta/T)\sum_{i=1}^{T/\Delta} N_i\right)^2}{(\Delta/T)\sum_{i=1}^{T/\Delta} N_i}, \quad (40)$$

where $N_i = N((i-1)\Delta, i\Delta)$. Telesca et al. [91] represented the FF of a fractal point process as a monotonic power law

$$\mathrm{FF}(\Delta) = 1 + \left(\frac{\Delta}{\Delta_0}\right)^{\alpha}, \quad (41)$$

for $\Delta > \Delta_0$, where Δ_0 is the fractal onset time, marking the lower limit for significant scaling behavior, with negligible clustering below this. Again the fractal exponent α is estimated by a regression analysis of (41), after taking logarithms. For Poisson processes, the FF is approximately one for all interval lengths, and so $\alpha \approx 0$. Of the 35 sequences examined, 30 had fractal exponents ranging from 0.2 to 0.9 (mean of 0.5). The remaining five exhibited no value Δ_0 above which (41) held.

Gusev et al. [34] examined clustering of eruptions on Kamchatka during the last 10,000 years using the Weibull density (17) for the inter-onset times, the correlation dimension (32), and the Hurst exponent (39). No self-similar behavior was found for intervals greater than 800 years, but for shorter delays there was a significant degree of clustering. Gusev et al. [34] also generalized the correlation dimension by replacing (31) by a weighted version, where the contribution of a pair of events is proportional to the product of their eruptive volumes. This identified self-similar behavior over the entire range of inter-onset times (100–10,000 years).

Fractals have basically been used in volcanology as a descriptive technique. They can provide, at the level of a single volcano, some indication of the sort of stochastic model that might fit the observed behavior. Applied to multiple volcanoes, they can provide a similarity index, useful for sorting volcanoes into homogeneous groups.

Volcanic Regimes

Apart from the nonhomogeneous model (4), the parametric models examined above have all been stationary, in the sense that the statistical properties of the intensity function do not vary over time. However, Wadge [99] postulated that, while a volcano could be in steady state on a time scale of years to decades, in the longer term the activity can wax and wane. Thus, particularly for long-term hazard forecasts, it is desirable to formulate methods of detecting (and modeling) such non-stationary activity.

Wickman [103] suggested that the activity of a volcano could change over time, perhaps in discrete steps, which we will term *regimes*. Mulargia et al. [65] addressed the importance of objectively identifying regimes of a volcano, noting however that a) the number of regimes is unknown, b) they may follow different distributions, and c) sample sizes are generally small. Different patterns of eruption and magma output are supposedly features of most volcanoes, and fundamental to the understanding and modeling of eruptions. These regimes may represent changes in the eruption mechanism, the mechanism for transport of magma to the surface, or the eruptive style (see, for example, the interpretation by Wadge et al. [102]). The estimated properties of such regimes can exclude certain models or mechanisms of volcanic eruption [46].

Various methods have been used to identify changes in regime, one based on cumulative eruptive volume [99,102] already having been illustrated in Fig. 9. Another uses a running mean of time between onsets. Suppose that the individual reposes are denoted by $r_i = t_{i+1} - t_i$, $i = 1, \ldots, n-1$, and that $\sum_{i=1}^{n-1} r_i = T$. If the reposes are divided into disjoint groups of m consecutive reposes, then $(2(n-1)/T)\sum_{i=(k-1)m+1}^{km} r_i$, for $k = 1, \ldots, [\![(n-1)/m]\!]$, where $[\![\cdot]\!]$ is the greatest integer function, are independent and have a chi-square distribution with $2m$ degrees of freedom [46].

Mulargia et al. [65] presented a method based on the theory of change-point problems. Suppose that we have

a time series of data $X_1, \ldots, X_n$, which might be, for example, inter-onset intervals, eruptive volumes, or effusion rates. The algorithm begins by setting a significance level α, and examining the time series to determine the most significant change point. If a change point is found, the time series is divided into two parts at this point, and each examined to determine the most significant change point. The algorithm proceeds recursively until no further significant change points are found. The most significant change point is identified by a search process: For $m = 3, \ldots, n-3$, divide the time series into two segments $X^{(1)} = X_1, \ldots, X_m$, and $X^{(2)} = X_{m+1}, \ldots, X_n$, and calculate the Kolmogorov–Smirnov (two-tail) statistic as

$$J = \left(\frac{m(n-m)}{n}\right)^{1/2} \max_x \mid F_1(x;m) - F_2(x;n-m) \mid, \quad (42)$$

where $F_j(x;k) = \#\{i\colon X_i^{(j)} \le x\}/k$ is the empirical distribution function of the first segment, et cetera. For large $m, n-m (> 30)$, the approximate critical value J_α of the statistic can be determined from

$$\Pr(J < J_\alpha) = \sum_{k=-\infty}^{\infty} (-1)^k \exp(-2k^2 J_\alpha^2) = 1 - \alpha\,. \quad (43)$$

Exact critical values for small $m, n-m$ are also available.

Later approaches used the cumulative count of eruptions in a statistical control chart [39], applied by Burt et al. [9] to Nyamuragia, and rank order statistics for the size of event [71]. Bebbington [3] noted that a hidden Markov model for the activity, with the unobserved state representing the regime, provides for regime identification via the Viterbi algorithm, which finds the most likely path through the hidden states.

Mount Etna

Mount Etna possesses one of the most complete records of volcanism known. A variety of eruption styles have been displayed, but since 1500 the activity has been predominately effusive eruptions from lateral vents on the volcano's flanks, or from the central craters. The record of summit eruptions is incomplete prior to 1970 [101], but data for flank eruptions is considered complete since 1600 or so [65]. The volumes of all such eruptions have been estimated [74], and these have been shown to be well correlated with the observed duration [67]. As the summit eruptions appear to be of different style [102], and to have different mechanisms to the flank eruptions [66] the longer record of the latter can be analyzed in isolation. In addition, only flank eruptions pose a major threat to inhabited areas [73].

Using the cumulative erupted volume curve, Wadge and Guest [101] found that the volume output was in steady state 1971–1981. Mulargia et al. [67] found that the events 1600 to 1980 were satisfactorily fit by a stationary Poisson process, but that the distribution of eruptive durations changed at the end of the 17th century, confirming the observation of Wadge et al. [102]. However, Mulargia et al. [65] identified a change point in the inter-onset times at 1865, both parts being consistent with a Poisson process, and, when augmented by volume data, change points at 1670 and 1750. Using the effusion rate (= volume/duration), Mulargia et al. [65] found a change point at 1950 in the 1600–1980 data, with the later sequence being of significantly lower rate. A repeat of the analysis by Gasperini et al. [31] for the data 1978–1987 identified a change point in the inter-onset times at November 1978, although no change point occurred in the volume series. Ho [39] fitted a statistical control chart based on the Weibull process, and found departures from the trend (4) in 1702 and 1759. Marzocchi [55] found some evidence of a trend in inter-onset times in the 1600–1994 data, while Sandri et al. [74] found no change points in the inter-onset times or eruptive volumes in the 1971–2002 data. Salvi et al. [73] determined that the activity 1536–2001 was not Poissonian, and although there was no trend 1536–1980, there was a trend in the data 1536–2001. They therefore concluded that the last 20 years of data is from a nonhomogeneous Poisson process, and fitted a Weibull process model with increasing trend.

The regime change points identified in the Mount Etna flank eruption data are summarized in Table 1. The lack of consistency is due to differences in both models and, via the fitting of them, in the coverage of the data. These models are 'first order' in the data being analyzed. Using hidden Markov models, Bebbington [3] was able to examine the 'second order' quantity of the correlation between durations and subsequent reposes. The conclusion was that changes in volcanic regime may be more frequent and/or fleeting, and thus undetectable by change-point methods, than has been thought. Notably, the identified change points corresponded closely to those identified manually by Behncke and Neri [7], showing that these are statistically identifiable. The complex relation between the duration of an eruption and the subsequent repose appears to be at least partially explained by an open/closed conduit system model with transitions between them. On the other hand, the correlation between reposes and subsequent durations may be sensitive to the longer of the two cycles identified by Behncke and Neri [7]. This would indi-

Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Table 1
Regime change points for Etna flank eruptions

Data type	Source	Date Range	Change Points
Onset counts	Mulargia et al. [67]	1600–1980	None
Inter-onset times	Mulargia et al. [65]	1600–1978	1865
	Gasperini et al. [31]	1978–1987	1987
	Ho [39]	1600–1978	1702, 1759
	Marzocchi [55]	1600–1994	(Trend)
	Sandri et al. [74]	1971–2002	None
	Salvi et al. [73]	1536–2001	1980
Volume -repose	Wadge et al. [102]	1535–1974	1610, 1669, 1759/1763
	Wadge and Guest [101]	1971–1981	None
	Mulargia et al. [65]	1600–1978	1670, 1750, 1950
	Gasperini et al. [31]	1978–1987	None
	Sandri et al. [74]	1971–2002	None

cate that flank eruptions are being fed from more than one magma chamber, of widely different sizes and recharge rates.

Spatial Aspects

In the case of polygenetic cones, the major aspect of hazard modeling is temporal. Spatial aspects are limited to the direction of flank eruptions (cf. [73]), and the likely paths of lava flows and lahars. In monogenetic volcanic fields, where cones correspond to single eruptions, there is a true spatio-temporal element to modeling occurrence data.

Spatio-temporal Intensities

The models used to describe spatio-temporal eruption occurrence have been largely nonparametric, or kernel-type. Connor and Hill [17] gave three methods:

1. Spatio-temporal nearest-neighbor estimate: Suppose we have n volcanoes, that the formation of the ith volcano occurred at time t_i, and let u_i be the area of a circle with radius equal to the distance between the point x of interest and the ith volcano. Further, let $j = 1, \ldots, m$ index the mth nearest neighbors to the point x using the distance metric $u_i(t - t_i)$. Then the estimated intensity at the point x is

$$\lambda(x, t; m) = \frac{m}{\sum_{j=1}^{m} u_j(t - t_j)} , \tag{44}$$

where m is the number of nearest neighbors used in the estimation, which was set equal to the number of volcanoes, although a smaller choice is possible. The temporal hazard for an area A can then be obtained by integrating the intensity (44) over the area. Note that in practice, to avoid singularities, $u_i > c$, a constant, and the intensity is summed over a grid. The choice of c may be an interesting question.
Condit and Connor [14] proposed that the optimal m can be determined by matching the average recurrence rate $\int_{x \in A} \lambda(x, t; m) \mathrm{d}x$ to the observed average recurrence rate $\bar{\lambda}(t) = N(t - s, t)/As$, in a window of length s, for the area A. Further, Condit and Connor [14] also suggested that overestimation during rapidly waning stages of activity can be addressed by defining a threshold time beyond which prior eruptions are not included in the calculation.
2. Kernel estimate: If we have n volcanoes, then using the Epanechnikov kernel

$$\kappa_i = \max\left\{0, \frac{2}{\pi}\left[1 - \left(\frac{d_i}{h}\right)^2\right]\right\} , \tag{45}$$

where d_i is the distance from x to the ith volcano, we have the spatial intensity

$$\lambda(x) = \frac{1}{e_h} \sum_{i=1}^{n} \frac{\kappa_i}{h^2} , \tag{46}$$

where e_h is an edge correction. Conway et al. [20] use the Gaussian kernel instead. The parameter h is a smoothing constant. A small value of h concentrates the probability close to existing volcanoes, while a large value distributes it more uniformly. Estimating the best value of h is a difficult task – see Vere-Jones [96] for an earth sciences based discussion.
3. Nearest-neighbor kernel estimate: Here we replace the smoothing constant h in the Kernel estimate with $d_{(m)}$, the distance to the mth nearest neighbor using the d_i metric.

The two (spatial) kernel estimates above can then be made into spatio-temporal estimates by multiplying them by a spatially independent temporal rate $\hat{\lambda}_0 = N/T$, resulting in

$$\lambda(x, t) = \lambda(x)\hat{\lambda}_0 \,. \tag{47}$$

Ho and Smith [42] generalized (47) by replacing $\hat{\lambda}_0$ by the Weibull process intensity (4).

Martin et al. [53] extended the spatio-temporal idea with a Bayesian formulation to incorporate information from other geophysical data such as P-wave velocity perturbations or geothermal gradients. The spatio-temporal

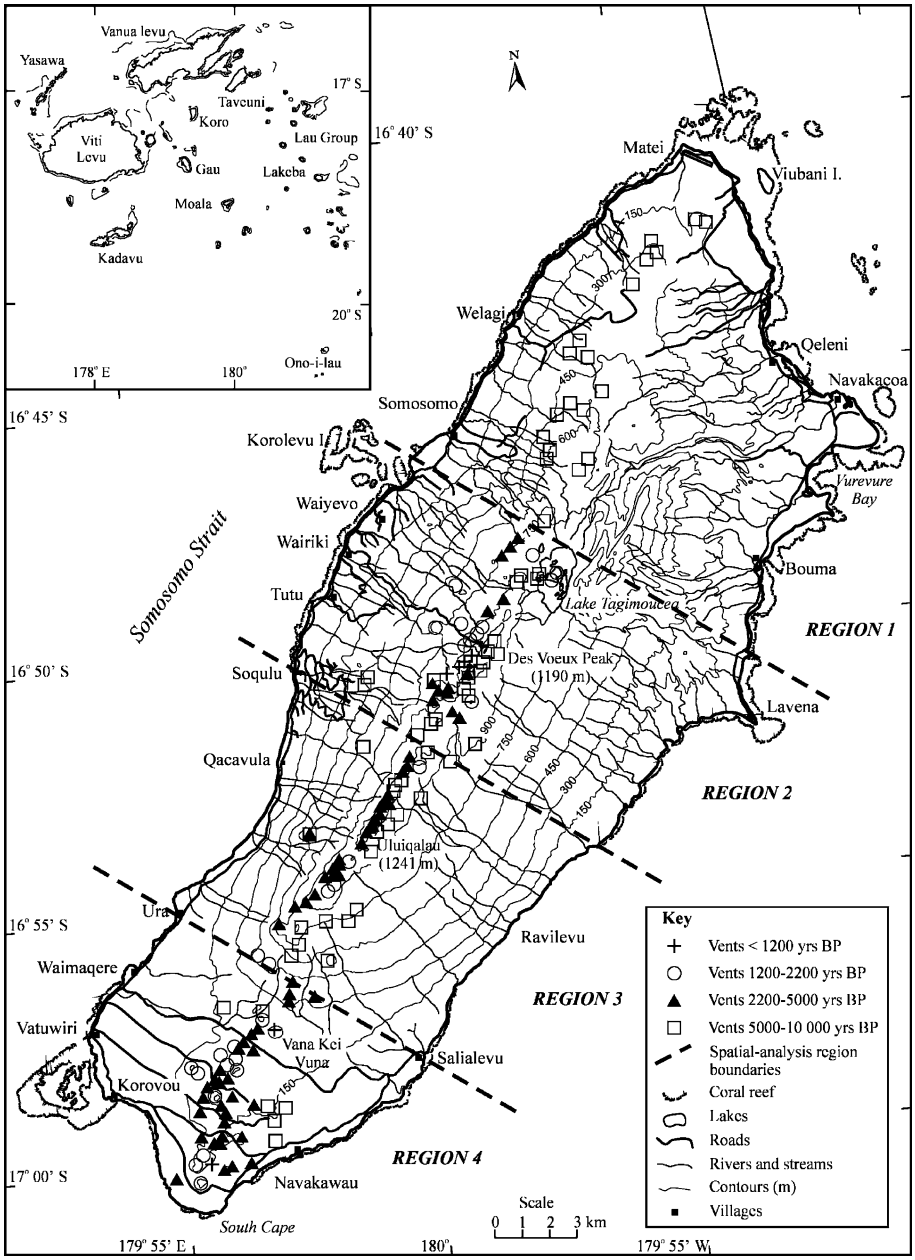

Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 10
Location map of Taveuni (*inset:* Fiji group), with main villages, positions of Holocene mapped vents, and regions used for the spatial analysis. From [22], Figure 1, © Springer-Verlag 2001. With kind permission of Springer Science and Business Media

intensity (46) is used as an a priori intensity $P(x)$, and a likelihood function $L(\theta \mid x)$ is generated by conditioning the geophysical data θ on the locations of volcanic events. The a posterior intensity can then be obtained from Bayes' theorem as

$$P(x \mid \theta) = \frac{P(x)L(\theta \mid x)}{\int_{y \in A} P(y)L(\theta \mid y)\mathrm{d}A}\,, \tag{48}$$

where A is the volcanic field.

Markov Chains

In the case where there is a geologically useful spatial classification, and sufficient data, a simple Markov chain approach can be used. If the ith eruption occurs in spatial region X_i, then the transition matrix is $P = (p_{jk})$, where $p_{jk} = \Pr(X_{i+1} = k \mid X_i = j)$. The maximum likelihood estimates are $\hat{p}_{jk} = N_{jk}/\sum_l N_{jl}$, where N_{jk} is the number of transitions from j to k observed in the data. The Markov chain can be tested against a null hypothesis of independence using the statistic

$$\chi^2 = 2\sum_j \sum_k N_{jk} \log\left[\frac{N_{jk}/\sum_l N_{jl}}{\sum_l N_{lk}/\sum_l \sum_m N_{lm}}\right]\,, \tag{49}$$

which asymptotically has a χ^2 distribution with K^2 degrees of freedom, where K is the number of states. Cronin et al. [22] used this approach, made feasible by the essentially linear rift formation shown in Fig. 10, to estimate the hazard at the heavily populated south-west end of Taveuni, Fiji. This was combined (cf. Eq. (47)) with a hierarchical renewal model for the temporal aspect to provide an alternative to the spatio-temporal hazard estimate from (44) shown in Fig. 11. Eliasson et al. [30] likewise divided the Katla caldera, Iceland, into three sectors to examine the spatial progression of volcanogenic flood events.

In the same vein, a few studies have considered possible correlations between multiple vents on the same volcano. Klein [46] showed using a runs test that eruptions and intrusions at Kilauea tend to cluster in time, and that intrusions occur in place of eruptions during long reposes. By dividing Kilauea reposes into summit-summit, summit-flank, flank-summit and flank-flank pairs, and finding that their distributions did not differ statistically, Klein [46] showed that the location of the previous event provided no information about the time of the next eruption. At Mauna Loa, however, reposes following flank eruptions were significantly longer than those following summit eruptions, consistent with the fact that the former are generally more voluminous than the latter. Examination of the sequences of summit and flank eruptions disclosed that Kilauea summit eruptions tend to occur in runs, due solely to the long summit sequence of 1924–1954, while Mauna Loa displays no tendency for clustering or alternation of summit and flank eruptions. However, Lockwood [49], in an analysis of 170 well-dated prehistoric lava flows from Mauna Loa, identifies cycling between summit overflows and flank eruptions with a periodicity of about 2000 years.

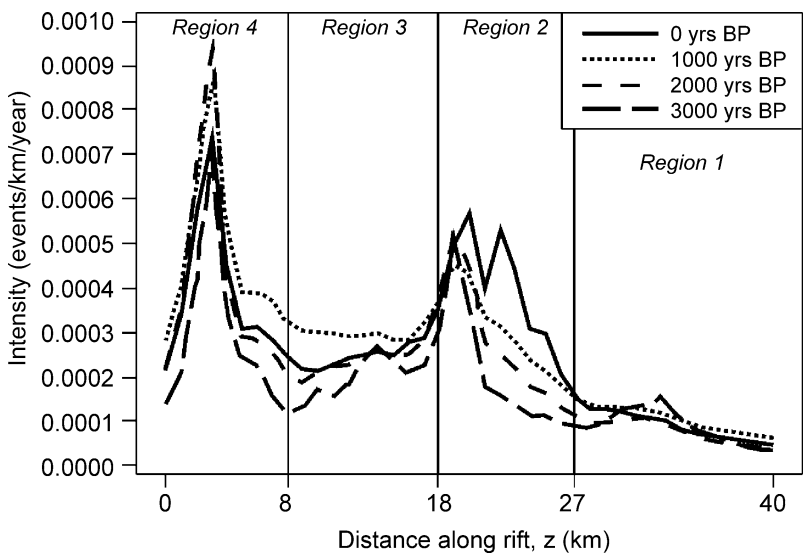

Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 11
Plot of spatio-temporal intensities (events/km year^{-1}) over the length of the Taveuni rift zone and at different times between 3000 B.P. and the present. From [22], Figure 8, © Springer-Verlag 2001. With kind permission of Springer Science and Business Media

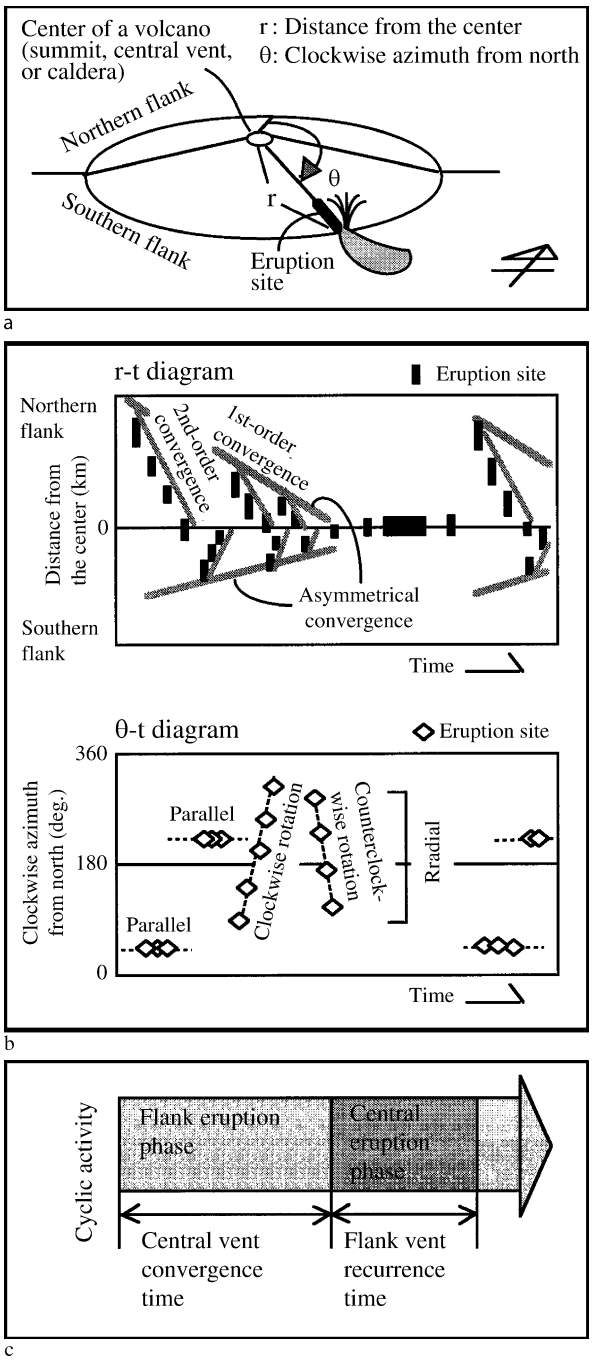

◀ **Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 12**

a Parameters representing the position of an eruption site. The flank of a volcano is divided into two regions, e. g., the northern flank and the southern one. b Two types of time-series diagram for eruption sites, representing the distance of eruption sites from the center (r-t diagram) and the clockwise azimuth from the north (θ-t diagram), respectively. c Schematic cyclic activity consisting of flank eruption and central eruption phases From [90], Figure 1, © Springer-Verlag 1997. With kind permission of Springer Science and Business Media

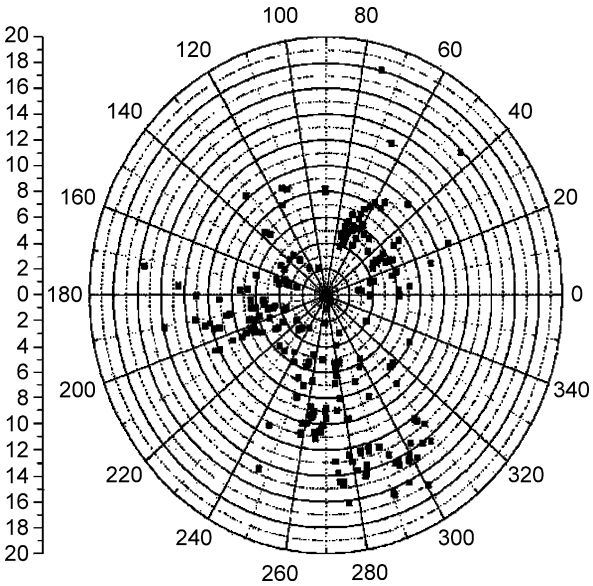

Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 13

Azimuthal distribution of cones of lateral eruption on Etna. The center is located on "La Voragine" central crater. The scale on the *left* is in km from the center. Reprinted from [73], © 2006 Elsevier B.V., with permission from Elsevier

Takada [90] proposed using time-series plots, as shown in Fig. 12, to visually express the temporal relationship between flank and summit eruptions, in particular plots of flank-vent distance and direction from the central vent. Central eruption phases, defined as periods where at least 70% of eruptions are central vent eruptions without accompanying flank eruptions, alternate with flank eruption phases. Salvi et al. [73] considered the azimuth distribution of flank eruptions from Mount Etna, as in Fig. 13, determining that there is no significant difference in the pre- and post-1536 distributions.

Klein [46] also examined the relation between activity at the nearby volcanoes of Kilauea and Mauna Loa. Separating the record into three parts at the longest repose of each volcano disclosed that the activities were inversely related. The rank correlation of the number of events in five year periods 1915–1980 was –0.646, significant at a level of 0.01, which indicates that they share the same magma source. Bebbington and Lai [5] repeated this analysis for the New Zealand volcanoes Ruapehu and Ngauruhoe, finding that the combined eruption sequence was indistinguishable from a superposition of indepen-

dent Poisson processes. This independence of the eruption sequences argues for separate magma sources.

Alignments and Clusters

There has also been considerable interest in detecting alignments or clusters of volcanic vents, in the hope that these can be related to geological features.

Lutz [50] proposed the 'two-point azimuth method', where the azimuth between every two vents is measured, generating $n(n-1)/2$ measurements for n vents. These are then binned in 10° intervals, although the bin size could vary, and the result tested for departure from randomness. As non-circular fields and/or a non-homogeneous density of points can produce a preferred orientation, a correction must be made in these cases. One way of doing this is to use Monte Carlo simulation of random points to produce a reference distribution from which departures can be detected. Lutz and Gutmann [51] noted that using kernel smoothing methods for the vent location provides an improved distribution from which to perform the Monte Carlo simulation.

Wadge and Cross [100] suggested the use of the Hough transform. In this technique, a point (x, y) is converted into the normal parameterization (ρ, θ), where $\rho = x\cos\theta + y\sin\theta$. Thus each point generates a curve in the (ρ, θ) plane as θ varies, and multiple (more than two curves) intersections of these define the alignment of co-linear points. In practice the scales for ρ and θ are quantized into discrete bins, which allows for a small degree of departure from strict co-linearity. Significance levels can be obtained via Monte Carlo techniques.

Connor [15] suggested the use of uniform kernel density fusion cluster analysis, using the distance metric

$$d(i,j) = \begin{cases} \frac{1}{2}\left(\frac{1}{m_i(w)} + \frac{1}{m_j(w)}\right), & \|x_i - x_j\| < w \\ \infty, & \text{otherwise}, \end{cases} \tag{50}$$

where $\|\cdot\|$ denotes map distance, and $m_i(w)$ is the number of cones within a circle of radius w centered on the ith vent. Cones are linked into clusters starting with the smallest values of d. Clusters are only combined if the 'density fusion' d between the clusters is greater than the maximum d between any two events within either cluster. This makes it possible to recognize overlapping clusters. The search radius w is determined experimentally.

Connor [15] and Connor et al. [16] used the two-point azimuth method, Hough transform and cluster analysis to map clusters of multiple cones with common alignments in the TransMexican volcanic belt and Springerville volcanic field (Arizona).

Magill et al. [52] measured vent clustering in the Auckland volcanic field using the statistic

$$L(d) = \sqrt{\frac{K(d)}{\pi}} - d\,, \tag{51}$$

where

$$K(d) = \frac{A}{n}\sum_{j=1}^{n}\sum_{i\neq j}\frac{H(d-d_{ij})}{w_{ij}} \tag{52}$$

is Ripley's K-function, A is the area of the field, d_{ij} is the distance between the ith and jth vents, $H(\cdot)$ is the Heaviside function, and w_{ij} is an edge correction. Peaks of positive $L(d)$ indicate clustering, and troughs of negative values regularity, at that distance. The Auckland field exhibits clustering at between 900 and 1600 m, and spatial regularity at about 4600 m. This clustering was used to collapse

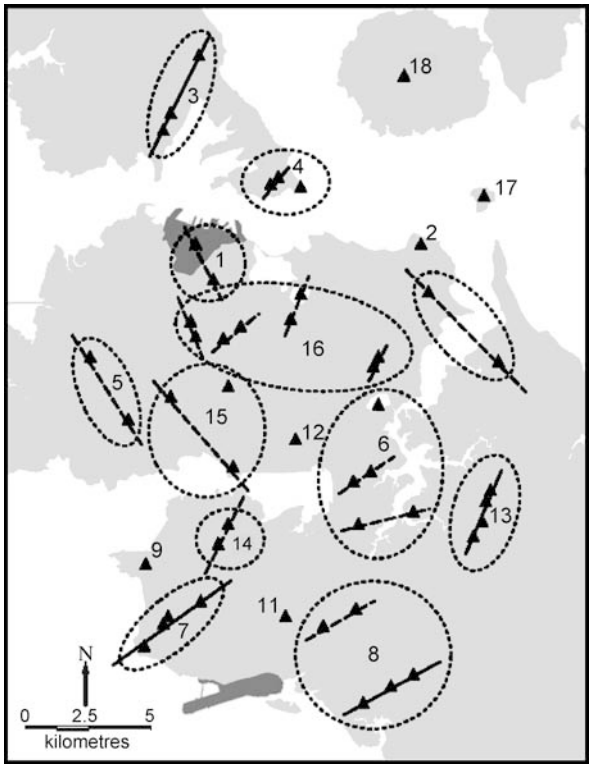

Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 14

Groups and alignments of vents within the Auckland volcanic field, identified by position and age, and their relative order of eruption. *Solid lines* indicate alignments for which there are three or more vents. *Dashed lines* are alignments inferred from two vents. From [52], Figure 5, © 2005 International Association for Mathematical Geology. With kind permission of Springer Science and Business Media

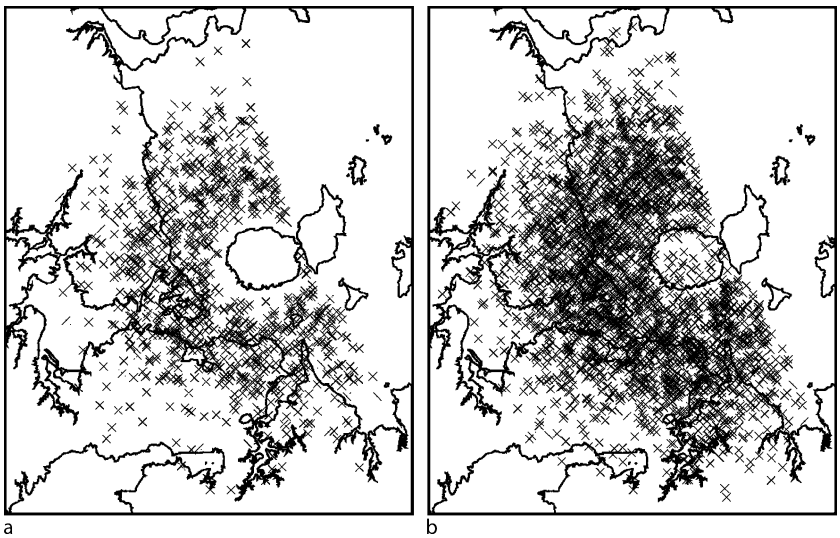

Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 15
a event centroid locations, and b eruption locations for 1000 simulated events. From [52], Figure 8, © 2005 International Association for Mathematical Geology. With kind permission of Springer Science and Business Media

49 cones into 18 events. Magill et al. [52] then examined the distance separating consecutively erupted vents, and the distance between clustered eruptions, concluding that eruptions do not occur randomly, and occur preferentially closer to the previous eruption. Vent alignments in NE–SW and NW–SE directions were observed within clusters, as shown in Fig. 14. Using these observations, Magill et al. [52] constructed a Monte Carlo simulation model for the location of the next vent. Based on the last vent location (Rangitoto), the distance to the center of the next cluster was simulated from a distribution fitted to the observed distances. The number of vents was described by a negative binomial distribution, and vent locations were then simulated from a fitted distribution for the distance from the center of the cluster, modified by the observed alignment distribution, with the results shown in Fig. 15. Some 57% of the simulated eruption locations occur either on land, or within 1 km of the coast, rising to 69% if one considers the probability of at least one eruption from the next cluster falling within these limits.

Yucca Mountain

After Mount Etna, the next most studied volcano is probably the volcanic field at Yucca Mountain, Nevada (cf. Fig. 16), due to the high-level radioactive waste repository planned there. The available data [13] consists of a large number of basaltic volcanoes, of ages 80,000 years to 10,500,000 years. These of course have large dating errors associated with them, but most analysis do not seem to have included this. The object is to estimate the probability of disruption of the repository by a volcanic intrusion. This is dependent on the probability of an intrusion, the distribution of the intrusion's dimensions, and on the dimensions of the repository itself.

The first analysis [23] used a Poisson process for the probability of an event, coupled with a Bernoulli probability of an event disrupting the repository:

$$\Pr(\text{no disruption before time } t) = \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!}(1-p)^n = \mathrm{e}^{-\lambda t p} . \tag{53}$$

The probability of an event being disruptive was estimated as $p = a/A$, where a is the area of the repository, and A the area, including the repository, used to define λ. The latter was estimated as N/T, where N is the number of scoria cones, and T the corresponding time period. Ho et al. [43] elaborated on means of estimating λ through event counts, repose intervals and magma volumes, and Ho [38] suggested a Bayesian framework for estimating p.

The next approach [38] attempted to incorporate a trend in the rate of volcanism using the Weibull process (4). As the data consisted of a number of clusters each containing several cones, but only the former had been dated, the event dates input to the Weibull process

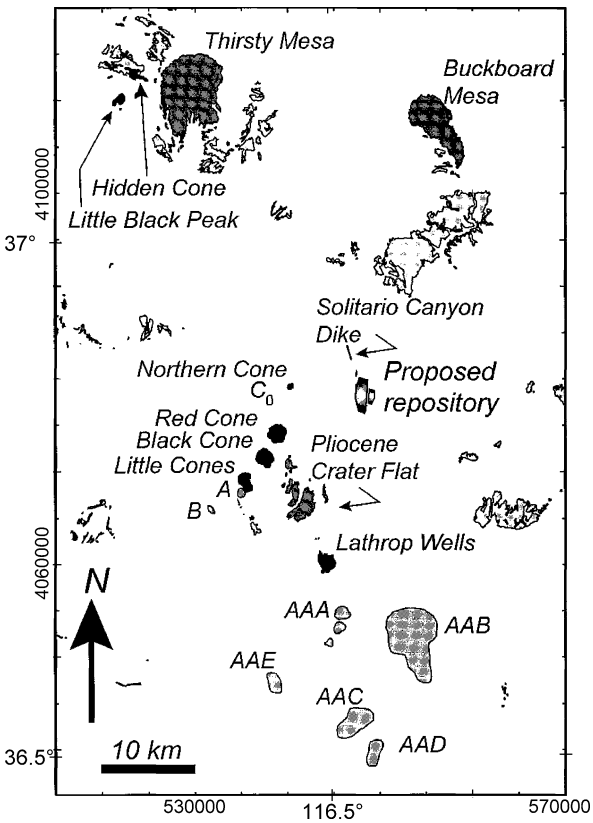

Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 16

Distribution of basaltic rocks in the Yucca Mountain Region: *light gray shading*, Miocene; *dark gray shading*, Pliocene; *black shading*, Quaternary. Magnetic anomalies are shown in *light gray shading*, but only AAB is dated (Pliocene). Location is given in latitude/longitude, and in Universal Transverse Mercator (UTM) coordinate pairs (in meters). From [19], Figure 1b, © 2000 by the American Geophysical Union

consisted of multiples of each of a few dates. Depending on the details of the implementation, there was either a strong ($\beta > 2$) increasing trend, or no significant trend at all. Notably, Ho [38] used a homogeneous Poisson process, with rate equal to the fitted Weibull process intensity at the present, for future prediction, although a continuing Weibull process was one possibility considered in the sensitivity analysis [40]. Ho and Smith [41] combined the Weibull process and Bayesian estimation of p in a formulation allowing the incorporation of 'expert knowledge'.

Sheridan [80] included a spatial element by fitting a bivariate Gaussian distribution with five parameters (the (x, y) coordinates of the center, the length of the major and minor axes, and their orientation) to the volcanic field as a whole. Connor and Hill [17] used the spatio-temporal estimate (44), and integrated the resulting intensity over the area of the proposed repository, using a Gamma distribution for the area of a new volcano. The effects on the intensity of future events beyond the vicinity of the repository were ignored. Similar results were obtained using kernel-type estimators. Ho and Smith [42] elaborated on the latter by making the independent temporal component a Weibull process, while Connor et al. [19], while retaining the temporal Poisson process component, imposed a variety of geological constraints from crustal density, vent and fault alignments, and vent positions, on the spatial Gaussian kernel intensity in Fig. 17. A sensitivity analysis restricting the vent locations to various date-subsets was also conducted.

Crowe et al. [24], in their 'final report' of hazard studies at Yucca Mountain, summarized previous work, in particular that of Geomatrix Consultants [32]. They also fitted uniform, triangular (two varieties) and normal distributions to event count data, and simulated the results to obtain hazard estimates.

Smith et al. [85] suggested that the volcanism in the Yucca Mountain area could be episodic volcanism, and linked to an area of hot mantle. This would make possible another future peak of volcanism, with consequent underestimation of the recurrence rate. It was suggested that recurrence rates up to four times those previously used might be feasible. On the other hand, Coleman et al. [13] considered the lack of any detected dikes in or above the potential repository block, treating intrusions as an isolated Poisson process with zero observations in 13,000,000 years (although there was one 'near-miss'). The conclusion was that high probabilities (10^{-6} per year) of disruption are 'unrealistic'. Table 2 summarizes the published hazard estimates.

Interactions with Earthquakes

A common question is whether an eruption is related to a preceding large regional earthquake (or vice-versa). Stress changes from earthquakes resolved at dike locations tend to be small, in fact smaller than those associated with solid earth tides. The exception are those earthquakes within a few fault-lengths. However, small increments of stress may of course advance the occurrence of an already imminent eruption. On a longer scale, as both eruptions and earthquakes are tectonic in nature, they must be related. Any such correlation may provide additional information about the likelihood of an imminent eruption which may be caused, or at least 'detected', by seismic activity. Alternatively, one might be interested in the likeli-

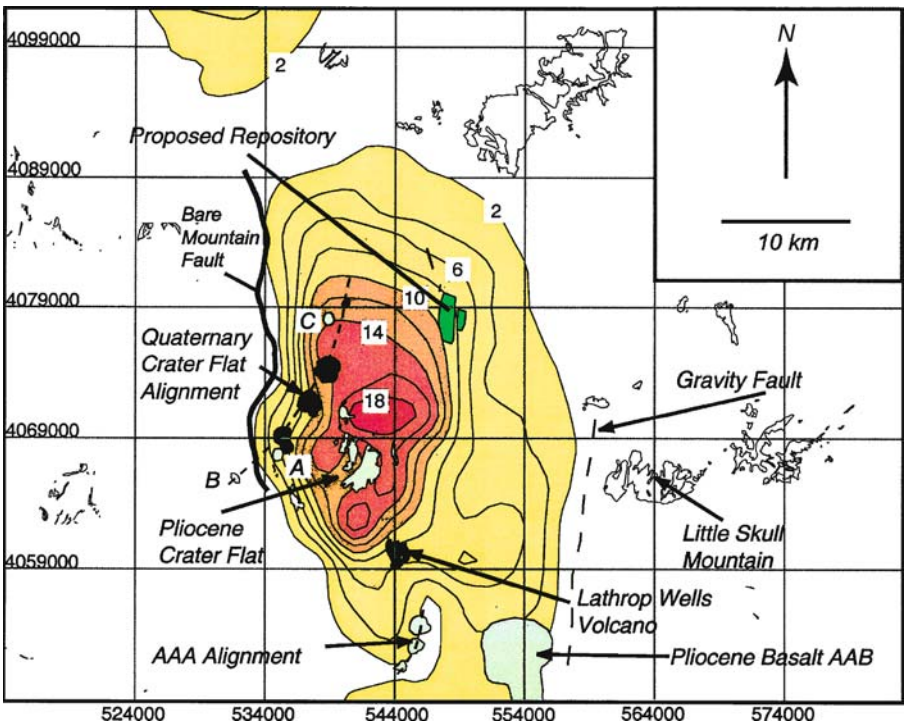

Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 17
The spatial recurrence rate (volcanic events/km^2) contoured for the Yucca Mountain Region, based on the distribution of Quaternary volcanism and its relationship to the Bare Mountain Fault. The contour interval is 2×10^{-4} volcanic events/km^2. Location is given in Universal Transverse Mercator (UTM) coordinate pairs (in meters). From [19], Plate 2, © 2000 by the American Geophysical Union

Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Table 2
Summary of hazard assessment for Yucca Mountain

Source	Estimation Method		Probability of disruption
	Temporal	Spatial	in 10^5 years
Crowe et al. [23]	Poisson process	$p = a/A$	10^{-5} to 10^{-3}
Ho et al. [43]	Poisson process	(p)	$0.5p$
Sheridan [80]	(λ)	bivariate Gaussian	$6\lambda \times 10^{-3}$ to $1.7\lambda \times 10^{-2}$
Ho [38]	Weibull (past)/ Poisson (future) process	Bayesian	10^{-3} to 7×10^{-3} (10^4 years)
Ho [40]	Weibull process	Bayesian	2×10^{-5} to 7×10^{-3}
Connor and Hill [17]	Spatio -temporal		10^{-4} to 5×10^{-4}(10^4 years)
Connor and Hill [17]	Poisson process	Kernel	10^{-4} to 5×10^{-4}(10^4 years)
Geomatrix [32]	Expert Panel: Various		5×10^{-5} to 5×10^{-3}
Ho and Smith [41]	Weibull process	Bayesian	10^{-3} to 3×10^2
Crowe et al. [24]	Summary: Various		7×10^{-5} to 4×10^{-3}
Ho and Smith [42]	Weibull process	Kernel	10^{-4} to 6×10^{-4}
Connor et al. [19]	Weighted spatio -temporal		10^{-3} to 10^{-2}
Coleman et al. [13]	Poisson process		10^{-3} to 10^{-2}

hood of potentially dangerous earthquakes during a period of volcanic activity.

Sharp et al. [79] considered the relationship between eruptions of Mount Etna and earthquakes from various structural zones of southern Italy. Both the earthquakes and eruption onsets were treated as Poisson processes, an assumption that had to be tested, with intensities λ_a and λ_b, respectively. A variable $\xi(t)$ is formed by summing

the number of A-series events in the intervals $(B_i, B_i + t)$, where (B_i) are the observed times of the B-series events. If the two series are independent, then $\xi(t)$ should have a Poisson distribution with mean $\hat{\lambda} = \hat{\lambda}_a \hat{\lambda}_b T$, where T is the length of the observation period, and $\hat{\lambda}_i = N_i(T)/T$, for $i = a, b$. For example, if B-series events consistently lead A-series events by some interval less than t, $\hat{\xi}$ will be significantly larger than $\hat{\lambda}$. The significance of an observed $\hat{\xi}(t)$, or confidence limits, can be obtained from the null distribution, which is Poisson with mean $\hat{\lambda}$. Repeating the procedure for varying t identifies significant correlations. Sharp et al. [79] found that the earthquakes were not Poissonian, but could be represented as a non-stationary Poisson process. A significant relationship was found between local earthquakes and subsequent flank eruptions within a few hundred days. Also, summit eruptions were found to occur before flank eruptions not preceded by earthquakes, but the summit eruptions themselves were not correlated with earthquakes. This suggests two mechanisms for flank eruptions of Etna: Those preceded by summit eruptions may be caused by magma pressure in the reservoir, while those preceded by earthquakes may be due to fracturing of the flank by earthquakes caused by tensile forces associated with the east-west extension of eastern Sicily.

Klein [47] examined various precursors of volcanic eruptions at Kilauea by examining the ratio of these precursors preceding an eruption to their long-term average value. The candidates included summit tilt, daily counts of small caldera earthquakes, repose since the last eruption, earth tides, and rainfall, as shown in Fig. 18. The number of caldera earthquakes was highly significant as a short-term (10 days or less) predictor, but tilt rate was a better predictor beyond 20 days. The fortnightly earth tidal modulation was also significant (cf. [54]), as it appears to trigger a volcano that is nearly ready to erupt anyway. Similarly, Mulargia et al. [66] used a statistical non-parametric pattern-recognition algorithm to examine possible precursors of eruptions at Mount Etna 1974–1990, including the number of earthquakes in five surrounding tectonic regions, and their maximum magnitude. An eighty day interval straddling the eruption onset was considered, and the pattern of 12 factors in each interval noted. The only statistically significant result was that flank eruptions were linked to seismicity with more than six earthquakes in the surrounding regions during the 80 day period. All 11 of the flank eruptions considered satisfied this condition, but only 11 out of 24 (in a total of 62) windows with this level of seismicity corresponded to earthquakes. Summit eruptions possessed no discernable pattern.

Mulargia [64] used the intensity cross product for two series of events at times $0 < s_1 < \cdots < s_{M(T)} < T$ and

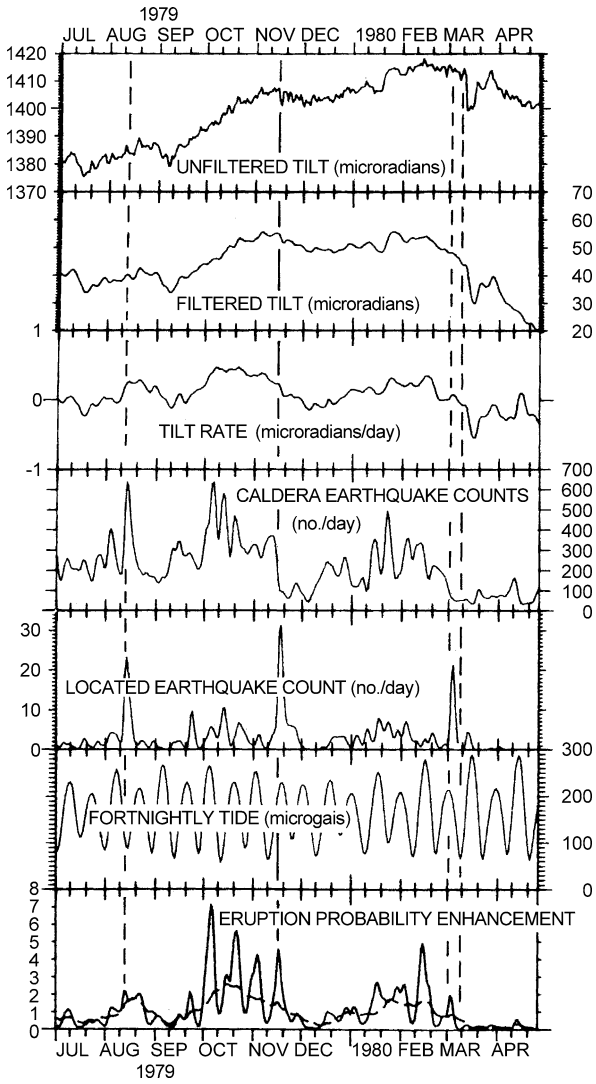

Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 18

Kilauea volcano: Several parameters and eruption probability enhancement at an expanded time scale. The figure spans 10 months during 1979–1980 and includes the November 16, 1979, eruption (*solid vertical line*), and intrusions on August 12, 1979, and March 2 and 10, 1980 (*dashed vertical lines*). Intrusions, like eruptions, are generally accompanied by summit deflation, an earthquake swarm and harmonic tremor, but magma fails to reach the surface. The eruption probability enhancement is plotted as a 1-day outlook (*solid line*) and 30-day outlook (*dashed line*). The 1-day probability depends on the fortnightly tide and oscillates with it. The 30-day probability is a function of smoothed tilt and earthquake parameters and does not depend on the tide. Note that eruption probability enhancement was higher than average (greater than 1) for about 2 months prior to the November 1979 eruption and March 1980 intrusions. From [47], Figure 4 © 1984 by the American Geophysical Union

$0 < t_1 < \cdots < t_{N(T)} < T$. For a lag u, this is estimated by

$$\hat{p}_{MN}(u) = \frac{n(u,h)}{2hT}\,, \tag{54}$$

where $n(u,h) = \#\{i\colon t_j + u - h \le s_i \le t_j + u + h\,,\ \text{for some}\ j\}$ is the number of events from the process M that occur within a window of length $2h$ of the lag of a point in the process N. Under a null hypothesis of two independent Poisson processes, the significance level can be calculated as

$$P = \sum_{k=n(u,h)}^{N(T)} \binom{k}{N(T)} \left(\frac{2hM(T)}{T}\right)^k . \tag{55}$$

This was applied to the flank eruptive activity and local seismicity at Mt Etna 1974–1990. The data consisted of 11 eruptions and 12 seismic sequences [31]. Using a window of length $2h = 10$ days, Mulargia [64] found a highly significant association peak at a lag u between –7 and 3 days. This indicates that local earthquakes and flank eruptions are largely concomitant phenomena.

Marzocchi et al. [58] examined the correlation between eruptions of Vesuvius and earthquakes of $M > 5.4$ in the southern Apennines, Calabrian Arc and Sicily, 1631–1944. The eruptions and earthquakes were binned into intervals of 1, 2, 3 and 4 years, and the Spearman rank correlation

$$\rho_{XY}(k) = \frac{\sum_{i=1}^{m} R(X_{i+k})R(Y_i) - m(m+1)^2/4}{\left(\sum_{i=1}^{m} R(X_{i+k})^2 - m(m+1)^2/4\right)^2 \cdot \left(\sum_{i=1}^{m} R(Y_i)^2 - m(m+1)^2/4\right)^2} \tag{56}$$

calculated, where X_i, Y_i are the total seismic moment and number of eruptions in the ith time bin, $R(\cdot)$ is the rank, from smallest to largest, of the quantities, and k is the lag. Varying this lag produces the Spearman correlogram. Marzocchi et al. [58] observed changes in eruptive activity following increases in seismicity in the southern Apennines after 6–13 years, and with delays of 27–30 years and 36–39 years for earthquakes in Sicily and the Calabrian arc, respectively. Nostro et al. [70] investigated this coupling (cf. Fig. 19) in terms of the Coulomb stress change on a magma body beneath Vesuvius caused by a slip on Apennine normal faults (promoting eruptions by earthquakes) and on fault planes caused by voiding a buried magma body (promoting earthquakes by eruptions), as shown in Fig. 20. Using these results to select earthquakes and eruptions improved the significance of the Marzocchi et al. [58] analysis.

Cardaci et al. [10] used a cross-correlation technique to investigate the relation between seismic and volcanic data at Mount Etna 1975–1987. The seismic and eruptive data were averaged over seven day intervals, with the cross-correlation being

$$\rho_{XY}(k) = \frac{\sigma_{XY}(k)}{\sqrt{\sigma_X^2 \sigma_Y^2}} \tag{57}$$

for $k = 0, \pm 1, \ldots, \pm K$, where

$$\sigma_{XY}(k) = \begin{cases} (1/n)\sum_{i=1}^{n-k}(X_i - \mu_X)(Y_{i+k} - \mu_Y) & k = 0, 1, \ldots, K \\ (1/n)\sum_{i=1-k}^{n}(X_i - \mu_X)(Y_{i+k} - \mu_Y) & k = -1, \ldots, -K \end{cases} \tag{58}$$

is the cross covariance function and μ_X, μ_Y and σ_Y^2, σ_Y^2 are the means and variances, respectively, of the two time series $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. The significance of the cross-correlation can be assessed from the fact that $\rho_{XY}(k)\sqrt{(m-2)/(1-\rho_{XY}(k))}$, is asymptotically distributed as a t-distribution with $m-2$ degrees of freedom, where m is the number of overlapped positions between the two series. Cardaci et al. [10] arrived at a similar conclusion to Mulargia et al. [66], identifying low frequency events, whose rate of occurrence increases from 17 to 108 days prior to flank eruptions, as the best precursor. Summit events had a less clear relation with volcanic tremor rather than seismic activity.

Linde and Sacks [48] examined the historical record of eruptions (VEI ≥ 2) and large earthquakes, to see whether there are more of the former within two days of the latter. For earthquakes with (M ≥ 8), there are a significant excess of eruptions within 750 km of the epicenter, while for earthquakes with $7.0 \le M \le 7.9$, the effect is only observed within a radius of 200 km. In all, 20 such earthquake-volcano pairs are observed 1587–1974. In addition, Linde and Sacks [48] hypothesized that eruptions in unison of volcanoes separated by hundreds of kilometers could be due to the second eruption being triggered by earthquakes associated with the first. Marzocchi [56] elaborated on this analysis by defining a perturbation function

$$\phi_i^{(k)}(\Delta) = \sum_j M_j w(d_{jk}) \mathbf{I}_{((i-1)\Delta, i\Delta)}(t_k - s_j)\,, \tag{59}$$

where

$$\mathbf{I}_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases} \tag{60}$$

is the indicator function, s_j and M_j are the time and seismic moment of the jth earthquake, t_k is the onset time of

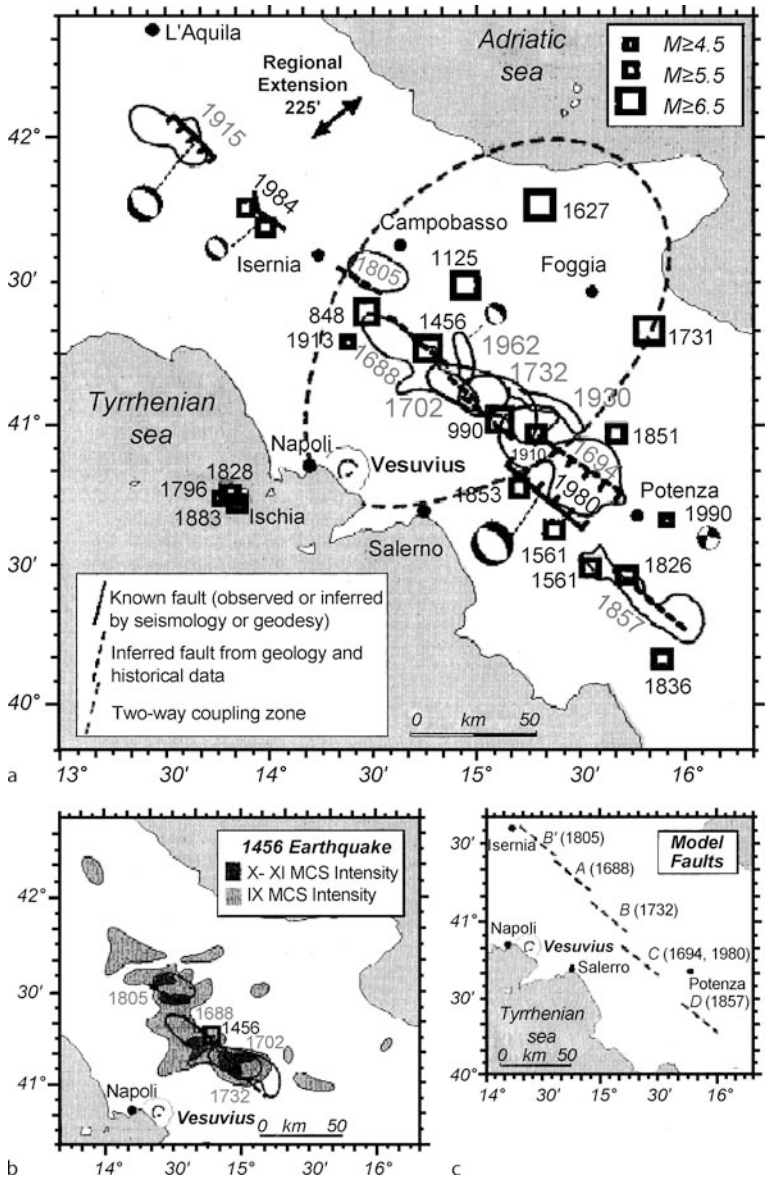

Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 19

a Sites of historical earthquakes in south central Italy, associated faults and Vesuvius volcano. Epicentres or, if unknown, isoseismals of modified Mercalli intensity, are shown. The coupling zone is the region in which normal-faulting earthquakes may promote Vesuvius eruptions and vice versa. b Isoseismals for the 1456 earthquake with 1688–1805 isoseismal zones. c Simplified model faults A-D used for calculations, with the dates of earthquakes associated with each fault segment. From [70], Figure 1 © 1998 by the American Geophysical Union

the kth eruption, and i is the lag. The distance between the epicenter of the jth earthquake and the kth eruption is d_{jk}, which is weighted by a (sharply) decreasing function $w(\cdot)$. The statistical significance of the calculated values is assessed by a randomization test. The idea is that the perturbation on the volcano by earthquakes is proportional to the seismic moment and decays with distance. Of the eight eruptions with VEI ≥ 5 examined, using $\Delta = 3$ and 5 years, five have a significant peak at 0–5 years in the perturbation function, calculated from earthquakes with M ≥ 7 (792 events, 1900–1999). Using an earthquake catalog with M ≥ 6 (3186 events, 1950–1997), two of four events have a significant peak at 30–35 years before the eruption. Marzocchi et al. [60] expanded this analysis to

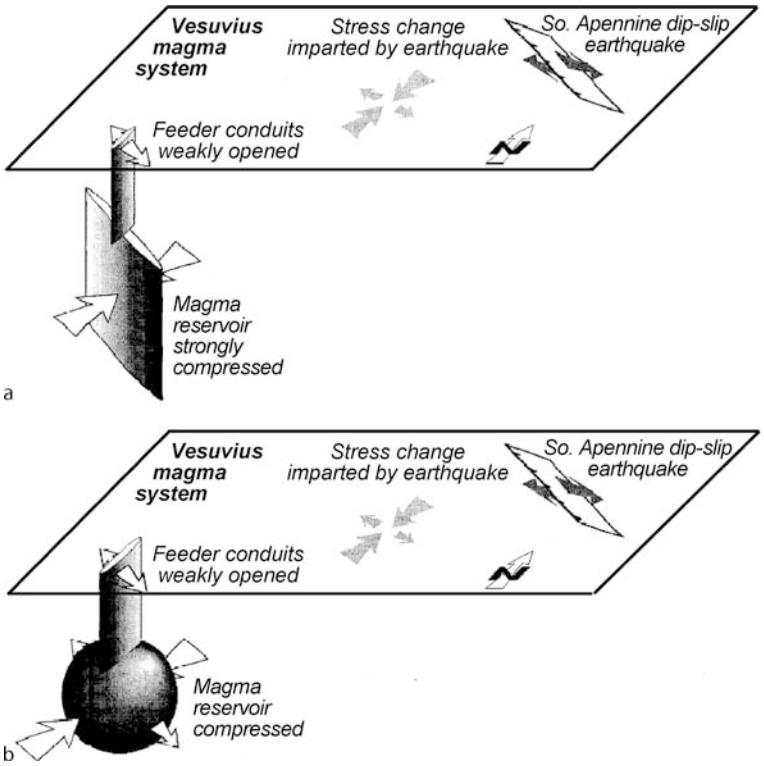

Volcanic Eruptions: Stochastic Models of Occurrence Patterns, Figure 20
Schematic illustration of the response of a hypothetical Vesuvius magmatic system to a southern Apennine normal-faulting earthquake for (a) a buried dike in which at least one fissure or feeder conduit strikes NE and (b) a buried spherical magma chamber with a NE striking fissure. From [70], Figure 3 © 1998 by the American Geophysical Union

the 62 20th Century eruptions with VEI ≥ 4, identifying peaks at 0–5 and 20–35 years. The perturbations for VEI ≥ 4 eruptions were shown to be significantly larger than those for VEI $= 1$ and VEI $= 2$ eruptions. This suggests that the likelihood of large explosive eruptions is affected by post-seismic stress variations induced by large earthquakes up to decades earlier and hundreds of km distant. No spatio-temporal clustering of the earthquakes and eruptions was detected.

Future Directions

There are two places in the above review where the effect of a second process on the likelihood of eruptions has been identified, but an appropriate model incorporating this is lacking. These are deterministic (fortnightly and/or annual) cycles in the intensity, and correlation between eruptions and large earthquakes. A candidate to model the first is a cyclic Poisson process [97], and the second could be examined via a bivariate Point process, or perhaps a non-homogeneous hidden Markov model.

More generally, much of the relevant literature on stochastic processes has not been applied to problems in volcanology. In particular, (marked) point processes, alternating renewal processes, and semi-Markov processes would seem to be natural candidates for modeling volcanic activity. A number of techniques from time series analysis (see [106] for a brief review) may prove useful for the investigation of volcanic data at various time scales.

In the application of stochastic models, and the understanding to be gleaned from them, volcanology lags well behind seismology. This is despite the inherent advantages of a phenomenon in which spatial degrees of freedom can often be ignored. An obvious future direction is the borrowing of more techniques from seismology. Unlike seismology, a certain amount of effort (and money) can often produce additional data through coring and dating. This is a great advantage in statistical terms, especially as the data can be 'targeted' to some degree as required by the model. This additional data makes it possible in volcanology to detect, and forecast, non-stationary behavior, and to relate this to physical changes in the volcano [95].

Although this article has concentrated on models for occurrence patterns, actual volcanic hazard is dependent on the nature and behavior of volcanic products such as lava, ash fall, lahars and pyroclastic flows. Incorporating existing sophisticated numerical models of such behavior with stochastic models of occurrence, location/direction and magnitude will allow more accurate estimation of the hazard to life and property.

Bibliography

Primary Literature

1. Bacon CR (1982) Time-predictable bimodal volcanism in the Coso range, California. Geology 10:65–69
2. Bain LJ (1978) Statistical Analysis of Reliability and Life Testing Models. Marcel Dekker, New York
3. Bebbington MS (2007) Identifying volcanic regimes using hidden Markov models. Geophys J Int, 171:921–942
4. Bebbington MS, Lai CD (1996) On nonhomogeneous models for volcanic eruptions. Math Geol 28:585–600
5. Bebbington MS, Lai CD (1996) Statistical analysis of New Zealand volcanic occurrence data. J Volcanol Geotherm Res 74:101–110
6. Bebbington MS, Lai CD (1998) A generalised negative binomial and applications. Commun Statist Theor Methods 27:2515–2533
7. Behncke B, Neri M (2003) Cycles and trends in the recent eruptive behaviour of Mount Etna (Italy). Canadian J Earth Sci 40:1405–1411
8. Bronk Ramsey C (1995) Radiocarbon calibration and analysis of stratigraphy: The OxCal program. Radiocarbon 37:425–430
9. Burt ML, Wadge G, Scott WA (1994) Simple stochastic modelling of the eruption history of a basaltic volcano: Nyamuragira, Zaire. Bull Volcanol 56:87–97
10. Cardaci C, Falsaperla S, Gasperini P, Lombardo G, Marzocchi W, Mulargia F (1993) Cross-correlation analysis of seismic and volcanic data at Mt Etna volcano, Italy. Bull Volcanol 55: 596–603
11. Carta S, Figari R, Sartoris G, Sassi R, Scandone R (1981) A statistical model for Vesuvius and its volcanological implications. Bull Volcanol 44:129–151
12. Casetti G, Frazzetta G, Romano R (1981) A statistical analysis in time of the eruptive events on Mount Etna (Italy) from 1323 to 1980. J Volcanol Geotherm Res 44:283–294
13. Coleman NM, Abramson LR, Marsh BD (2004) Testing claims about volcanic disruption of a potential geologic repository at Yucca Mountain, Nevada. Geophys Res Lett 31:L24601
14. Condit CD, Connor CB (1996) Recurrence rates of volcanism in basaltic volcanic fields: an example from the Springerville volcanic field, Arizona. Geol Soc Am Bull 108:1225–1241
15. Connor CB (1990) Cinder cone clustering in the TransMexican volcanic belt: Implications for structural and petrologic models. J Geophys Res 95:19395–19405
16. Connor CB, Condit CD, Crumpler LS, Aubele JC (1992) Evidence of regional structural controls on vent distribution: Springerville volcanic field, Arizona. J Geophys Res 97: 349–359
17. Connor CB, Hill BE (1995) Three nonhomogeneous Poisson models for the probability of basaltic volcanism: application to the Yucca Mountain region, Nevada. J Geophys Res 100:10107–10125
18. Connor CB, Sparks RSJ, Mason RM, Bonadonna C, Young SR (2003) Exploring links between physical and probabilistic models of volcanic eruptions: The Soufriere Hills volcano, Montserrat. Geophys Res Lett 30:1701
19. Connor CB, Stamatakos JA, Ferrill DA, Hill BE, Ofoegbu GI, Conway FM, Sagar B, Trapp J (2000) Geologic factors controlling patterns of small-volume basaltic volcanism: Application to a volcanic hazards assessment at Yucca Mountain, Nevada. J Geophys Res 105:417–432
20. Conway FM, Connor CB, Hill BE, Condit CD, Mullaney K, Hall CM (1998) Recurrence rates of basaltic volcanism in SP cluster, San Francisco volcanic field, Arizona. Geology 26:655–658
21. Crandell DR, Mullineaux DR, Rubin M (1975) Mount St. Helens volcano: Recent and future behaviour. Science 187:438–441
22. Cronin S, Bebbington M, Lai CD (2001) A probabilistic assessment of eruption recurrence on Taveuni volcano, Fiji. Bull Volcanol 63:274–288
23. Crowe BM, Johnson ME, Beckman RJ (1982) Calculation of the probability of volcanic disruption of a high-level radioactive waste repository within southern Nevada, USA. Radioact Waste Manag 3:167–190
24. Crowe BM, Wallmann P, Bowker LM (1998) Probabilistic modeling of volcanism data: Final volcanic hazard studies for the Yucca Mountain site. In: Perry FV et al. (eds) Volcanism Studies: Final Report for the Yucca Mountain project. Los Alamos National Laboratory Report LA-13478, Los Alamos National Laboratory, Los Alamos, 415 pp
25. Decker RW (1986) Forecasting volcanic eruptions. Ann Rev Earth Planet Sci 14:267–291
26. De la Cruz-Reyna S (1991) Poisson-distributed patterns of explosive eruptive activity. Bull Volcanol 54:57–67
27. De la Cruz-Reyna S (1993) Random patterns of occurrence of explosive eruptions at Colima vulcano, Mexico. J Volcanol Geotherm Res 55:51–68
28. De la Cruz-Reyna S, Carrasco-Nunez G (2002) Probabilistic hazard analysis of Citlaltepetl (Pico de Orizaba) volcano, eastern Mexican volcanic belt. J Volcanol Geotherm Res 113: 307–318
29. Dubois J, Cheminee JL (1991) Fractal analysis of eruptive activity of some basaltic volcanoes. J Volcanol Geotherm Res 45:197–208
30. Eliasson J, Larsen G, Gudmundsson MT, Sigmundsson F (2006) Probabilistic model for eruptions and associated flood events in the Katla caldera, Iceland. Comput Geosci 10: 179–200
31. Gasperini P, Gresta S, Mulargia F (1990) Statistical analysis of seismic and eruptive activities at Mt. Etna during 1978–1987. J Volcanol Geotherm Res 40:317–325
32. Geomatrix Consultants (1996) Probabilistic volcanic hazard analysis for Yucca Mountain, Nevada. Report BA0000000-1717-220-00082. Geomatrix Consultants, San Francisco
33. Godano C, Civetta L (1996) Multifractal analysis of Vesuvius volcano eruptions. Geophys Res Lett 23:1167–1170
34. Gusev AA, Ponomareva VV, Braitseva OA, Melekestsev IV, Sulerzhitsky LD (2003) Great explosive eruptions on Kamchatka during the last 10,000 years: Self-similar irregularity of the output of volcanic products. J Geophys Res 108(B2):2126

35. Guttorp P, Thompson ML (1991) Estimating second-order parameters of volcanicity from historical data. J Amer Statist Assoc 86:578–583
36. Ho CH (1990) Bayesian analysis of volcanic eruptions. J Volcanol Geotherm Res 43:91–98
37. Ho CH (1991) Nonhomogeneous Poisson model for volcanic eruptions. Math Geol 23:167–173
38. Ho CH (1992) Risk assessment for the Yucca Mountain high-level nuclear waste repository site: Estimation of volcanic disruption. Math Geol 24:347–364
39. Ho CH (1992) Statistical control chart for regime identification in volcanic time-series. Math Geol 24:775–787
40. Ho CH (1995) Sensitivity in volcanic hazard assessment for the Yucca Mountain high-level nuclear waste repository site: The model and the data. Math Geol 27:239–258
41. Ho CH, Smith EI (1997) Volcanic hazard assessment incorporating expert knowledge: application to the Yucca Mountain region, Nevada, USA. Math Geol 29:615–627
42. Ho CH, Smith EI (1998) A spatial-temporal/3-D model for volcanic hazard assessment: application to the Yucca Mountain region, Nevada. Math Geol 30:497–510
43. Ho CH, Smith EI, Feuerbach DL, Naumann TR (1991) Eruptive probability calculation for the Yucca Mountain site, USA: Statistical estimation of recurrence rates. Bull Volcanol 54:50–56
44. Jaquet O, Low S, Martinelli B, Dietrich V, Gilby D (2000) Estimation of volcanic hazards based on Cox stochastic processes. Phys Chem Earth (A) 25:571–579
45. Jupp TE, Pyle DM, Mason BG, Dade WB (2004) A statistical model for the timing of earthquakes and volcanic eruptions influenced by periodic processes. J Geophys Res 109:B02206
46. Klein FW (1982) Patterns of historical eruptions at Hawaiian volcanoes. J Volcanol Geotherm Res 12:1–35
47. Klein FW (1984) Eruption forecasting at Kilauea Volcano, Hawaii. J Geophys Res 89:3059–3073
48. Linde AT, Sacks IS (1998) Triggering of volcanic eruptions. Nature 395:888–890
49. Lockwood JP (1995) Mauna Loa eruptive history – the preliminary radiocarbon record, Hawai'i. In: Rhodes JM, Lockwood JP (eds) Mauna Loa Revealed: Structure, Composition, History, and Hazards. American Geophysical Union Monograph, vol 92. American Geophysical Union, Washington DC, pp 81–94
50. Lutz TM (1986) An analysis of the orientation of large-scale crustal structures: A statistical approach based on areal distributions of pointlike features. J Geophys Res 91:421–434
51. Lutz TM, Gutmann JT (1995) An improved method for determining and characterizing alignments of pointlike features and its immplications for the Pinacate volcanic field, Sonora, Mexico. J Geophys Res 100:17659–17670
52. Magill CR, McAneney KJ, Smith IEM (2005) Probabilistic assessment of vent locations for the next Auckland volcanic field event. Math Geol 37:227–242
53. Martin AJ, Umeda K, Connor CB, Weller JN, Zhao D, Takahashi M (2004) Modeling long-term volcanic hazards through Bayesian inference: An example from the Tohoku volcanic arc, Japan. J Geophys Res 109:B10208
54. Martin DP, Rose WI (1981) Behavioral patterns of Fuego volcano, Guatemala. J Volcanol Geotherm Res 10:67–81
55. Marzocchi W (1996) Chaos and stochasticity in volcanic eruptions the case of Mount Etna and Vesuvius. J Volcanol Geotherm Res 70:205–212
56. Marzocchi W (2002) Remote seimic influence on large explosive eruptions. J Geophys Res 107:2018. doi:10.1029/2001JB000307
57. Marzocchi W, Sandri L, Gasparini P, Newhall C, Boschi E (2004) Quantifying probabilities of volcanic events: The example of volcanic hazard at Mount Vesuvius. J Geophys Res 109:B11201
58. Marzocchi W, Scandone R, Mulargia F (1993) The tectonic setting of Mount Vesuvius and the correlation between its eruptions and the earthquakes of the southern Apennines. J Volcanol Geotherm Res 58:27–41
59. Marzocchi W, Zaccarelli L (2006) A quantitative model for the time-size distribution of eruptions. J Geophys Res 111: B04204
60. Marzocchi W, Zaccarelli L, Boschi E (2004) Phenomenological evidence in favor of a remote seimic coupling for large volcanic eruptions. Geophys Res Lett 31:L04601
61. Mason BG, Pyle DM, Dade WB, Jupp T (2004) Seasonality of volcanic eruptions. J Geophys Res 109:B04206
62. Mauk FJ, Johnston MJS (1973) On the triggering of volcanic erutpions by earth tides. J Geophys Res 78:3356–3362
63. Medina Martinez F (1983) Analysis of the eruptive history of the Volcan de Colima, Mexico (1560–1980). Geof Int 22: 157–178
64. Mulargia F (1992) Time association between series of geophysical events. Phys Earth Plan Int 71:147–153
65. Mulargia F, Gasperini P, Tinti S (1987) Identifying regimes in eruptive activity: An application to Etna volcano. J Volcanol Geotherm Res 34:89–106
66. Mulargia F, Marzocchi W, Gasperini P (1992) Statistical identification of physical patterns which accompany eruptive activity on Mount Etna, Sicily. J Volcanol Geotherm Res 53: 289–296
67. Mulargia F, Tinti S, Boschi E (1985) A statistical analysis of flank eruptions on Etna volcano. J Volcanol Geotherm Res 23: 263–272
68. Newhall CG, Self S (1982) The volcanic explosivity index (VEI): An estimate of explosive magnitude for historical volcanism. J Geophys Res 87:1231–1238
69. Nishi Y, Inoue M, Tnaka T, Murai M (2001) Analysis of time sequences of explosive volcanic eruptions of Sakurajima. J Phys Soc Japan 70:1422–1428
70. Nostro C, Stein RS, Cocco M, Belardinelli ME, Marzocchi W (1998) Two-way coupling between Vesuvius eruptions and southern Apennine earthquakes, Italy, by elastic stress transfer. J Geophys Res 103:24487–24504
71. Pyle DM (1998) Forecasting sizes and repose times of future extreme volcanic events. Geology 26:367–370
72. Reyment RA (1969) Statistical analysis of some volcanologic data regarded as series of point events. Pure Appl Geophys 74:57–77
73. Salvi F, Scandone R, Palma C (2006) Statistical analysis of the historical activity of Mount Etna, aimed at the evaluation of volcanic hazard. J Volcanol Geotherm Res 154:159–168
74. Sandri L, Marzocchi W, Gasperini P (2005) Some insights on the occurrence of recent volcanic eruptions of Mount Etna volcano (Sicily, Italy). Geophys J Int 163:1203–1218
75. Santacroce R (1983) A general model for the behaviour of the Somma-Vesuvius volcanic complex. J Volcanol Geotherm Res 17:237–248
76. Scandone R, Arganese G, Galdi F (1993) The evaluation of

volcanic risk in the Vesuvian area. J Volcanol Geotherm Res 58:263–271
77. Scandone R, Giacomelli L, Gasparini P (1993) Mount Vesuvius: 2000 years of volcanological observations. J Volcanol Geotherm Res 58:5–25
78. Settle M, McGetchin TR (1980) Statistical analysis of persistent explosive activity at Stromboli, 1971: Implications for eruption prediction. J Volcanol Geotherm Res 8:45–58
79. Sharp ADL, Lombardo G, David PM (1981) Correlation between eruptions of Mount Etna, Sicily, and regional earthquakes as seen in historical records from AD 1582. Geophys J R astr Soc 65:507–523
80. Sheridan MF (1992) A Monte Carlo technique to estimate the probability of volcanic dikes. In: High-Level Radioactive Waste Management: Proceedings of the Third Annual International Conference, Las Vegas, April 12–16, 1992. American Nuclear Society, La Grange Park, pp 2033–2038
81. Shimazaki K, Nakata T (1980) Time-predictable recurrence model for large earthquakes. Geophys Res Lett 7:279–282
82. Siebert L, Simkin T (2002) Volcanoes of the World: an Illustrated Catalog of Holocene Volcanoes and their Eruptions, Smithsonian Institution, Global Volcanism Program Digital Information Series, GVP-3. http://www.volcano.si.edu/world/. Accessed 20 Jun 2008
83. Simkin T (1993) Terrestrial volcanism in space and time. Ann Rev Earth Panet Sci 21:427–452
84. Simkin T (1994) Distant effects of volcanism – how big and how often? Science 264:913–914
85. Smith EI, Keenan DL, Plank T (2002) Episodic volcanism and hot mantle: Implications for volcanic hazard studies at the proposed nuclear waste repository at Yucca Mountain, Nevada. GSA Today April 2002:4–9
86. Solow AR (1993) Estimating record inclusion probability. The Amer Statist 47:206–208
87. Solow AR (2001) An empirical Bayes analysis of volcanic eruptions. Math Geol 33:95–102
88. Sornette A, Dubois J, Cheminee JL, Sornette D (1991) Are sequences of volcanic eruptions deterministically chaotic? J Geophys Res 96:11931–11945
89. Stothers RB (1989) Seasonal variations of volcanic eruption frequencies. Geophys Res Lett 16:453–455
90. Takada A (1997) Cyclic flank-vent and central-vent eruption patterns. Bull Volcanol 58:539–556
91. Telesca L, Cuomo V, Lapenna V, Macchiato M (2002) Time-clustering analysis of volcanic occurrence sequences. Phys Earth Planet Int 131:47–62
92. Telesca L, Lapenna V (2005) Identifying features in time-occurrence sequences of volcanic eruptions. Environmentrics 16:181–190 58
93. Thorlaksson JE (1967) A probability model of volcanoes and the probability of eruptions of Hekla and Katla. Bull Volcanol 31:97–106
94. Turner M, Cronin S, Bebbington M, Platz T (2008) Developing a probabilistic eruption forecast for dormant volcanos; a case study from Mt Taranaki, New Zealand. Bull Volcanol 70: 507–515
95. Turner M, Cronin S, Smith I, Bebbington M, Stewart RB (2008) Using titanomagnetite textures to elucidate volcanic eruption histories. Geology 36:31–34
96. Vere-Jones D (1992) Statistical methods for the description and display of earthquake catalogs. In: Walden AT, Guttorp P (eds) Statistics in the Environmental and Earth Sciences, Edward Arnold, London, pp 220–246
97. Vere-Jones D, Ozaki T (1982) Some examples of statistical estimation applied to earthquake data. Ann Inst Statist Math 34:189–207
98. Voight B (1988) A method for prediction of volcanic eruptions. Nature 332:125–130
99. Wadge G (1982) Steady state volcanism: Evidence from eruption histories of polygenetic volcanoes. J Geophys Res 87:4035–4049
100. Wadge G, Cross A (1988) Quantitative methods for detecting aligned points: An application to the volcanic vents of the Michoacan–Guanajuato volcanic field, Mexico. Geology 16: 815–818
101. Wadge G, Guest JE (1981) Steady-state magma discharge at Etna 1971–1981. Nature 294:548–550
102. Wadge G, Walker WPL, Guest JE (1975) The output of Etna volcano. Nature 255:385–387
103. Wickman FE (1966) Repose-period patterns of volcanoes. I. Volcanic eruptions regarded as random phenomena. Arch Mineral Geol 4:291–367
104. Wickman FE (1966) Repose-period patterns of volcanoes. V. General discussion and a tentative stochastic model. Arch Mineral Geol 4:351–367
105. Wickman FE (1976) Markov models of repose-period patterns of volcanoes. In: Merriam DF (ed) Random Processes in Geology. Springer, New York, pp 135–161
106. Young PC (2006) New approaches to volcanic time-series analsyis. In Mader HM, Coles SG, Connor CB, Connor LJ (eds) Statistics in Volcanology. Geological Society of London, London, pp 143–160

Books and Reviews

Guttorp P (1995) Stochastic Modeling of Scientific Data, Chapman and Hall, London
Hill DP, Pollitz FP, Newhall C (2002) Earthquake-volcano interactions. Phys Today 55:41–47
Lindsay JK (2004) Statistical Analysis of Stochastic Processes in Time. Cambridge University Press, Cambridge
Karr AF (1991) Point Processes and Their Statistical Inference, 2nd edn. Marcel Dekker, New York
Mader HM, Coles SG, Connor CB, Connor LJ (eds) (2006) Statistics in Volcanology. Geological Society of London, London

Volcanic Hazards and Early Warning

Robert I. Tilling
Volcano Hazards Team, US Geological Survey,
Menlo Park, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Scope of Problem:

Challenge for Emergency–Management Authorities
Variability in Possible Outcomes of Volcano Unrest
Some Recent Examples of Actual Outcomes of Volcano Unrest
Challenges in Achieving Refined Predictive Capability
Future Directions
Acknowledgments
Bibliography

Glossary

Volcano hazards Potentially damaging volcano-related processes and products that occur during or following eruptions [see [31,37] for overviews]. In quantitative hazards assessments, the probability of a given area being affected by potentially destructive phenomena within a given period of time.

Volcano risk Probability of harmful consequences – individual or societal – or expected losses (deaths, injuries, property, livelihoods, economic activity disrupted or environment damaged) resulting from interactions between volcano hazards, human development, and vulnerable conditions. Though definitions vary in detail, risk is conventionally expressed by the general relation: *risk = hazards x vulnerability* [13].

Vulnerability The conditions determined by physical, social, economic, and environmental factors or processes, which increase the susceptibility of a community to the impact of hazards [13].

Volcano status The current activity or mode of behavior of a volcano. Status is commonly described as follows: *active* (having one or more recorded historical eruptions); *dormant* (currently inactive but having potential for renewed eruptive activity); and *extinct* (dormant for long time and not expected to erupt again). These terms, while commonly used in the scientific literature, are imperfect and are undergoing serious reexamination within the global volcanological community.

Factual statement Following the recommended definition of Swanson et al. [35], is the description of the current status and conditions of a volcano but does not anticipate future events.

Forecast As defined by Swanson et al. [35], is a comparatively imprecise statement of the time, place, and nature of expected activity; forecasts of eruptions and earthquakes sometimes are probabilistic (e. g., [43,47]). A forecast usually covers a longer time period than that for a prediction.

Prediction As defined by Swanson et al. [35], is a comparatively precise statement of the time, place, and, ideally, the nature and size of impending activity. Forecasts and predictions can be either *long term* (typically years to decades or longer) or *short term* (typically hours to months).

Volcano unrest Visual and (or) measurable physical and (or) chemical changes, surface or subsurface, in the status of the volcano, relative to its "normal" historical behavior; duration of unrest can vary from hours to decades. The initiation or escalation of volcano unrest, regardless of duration, may or may not culminate in eruption.

Magma intrusion The subsurface movement or injection of *magma* (molten rock containing associated crystals and gases) from one part of a volcanic system into another. Typically, an intrusion involves transport of magma from a central zone of storage (i. e., magma "reservoir") into peripheral, structurally weaker zones (e. g., faults or rifts). Some intrusions culminate in surface eruptive outbreaks, others do not.

Volcano monitoring The systematic collection, analysis, and interpretation of visual observations and instrumental measurements of changes at volcanoes before, during, and after the onset of volcano unrest and (or) eruptive activity.

"Baseline" monitoring data Volcano-monitoring data acquired for a volcanic system documenting its range of variation during its "normal" behavior prior to volcano unrest and (or) eruptions. The longer the time span covered by "baseline" monitoring, the more diagnostic the dataset for identifying any significant departures from normal behavior in anticipating the possible outcomes of escalating volcano unrest.

Volcanic crisis An unstable situation or time of heightened public concern when the level of volcano unrest exceeds its "baseline" level, thereby increasing the prospects of possible eruption at some future, but indeterminate time. In general, during a crisis, emergency managers face a relatively narrow "decision window" in which to take timely mitigative actions to ensure public safety.

"Decision window" The period of time – typically weeks to months, but can be longer – during a developing volcanic crisis between the onset or escalation of volcano unrest. During this time, emergency managers are under high-stress political and public pressure to make decisions regarding mitigative actions, including possible evacuation of populations at risk.

Warning An official message issued by government authorities, usually given to a specific community or communities when a direct response to a volcanic threat is required. To be useful, warnings must be cred-

ible and effectively communicated – in clear, easily understandable language – with sufficient lead-time, ideally, well before the volcano unrest escalates into a volcanic crisis.

Definition of the Subject

The hazards and risks posed by volcanic eruptions are increasing inexorably with time. This trend is the direct result of continuing exponential growth in global population and progressive encroachment of human settlement and economic development into hazardous volcanic areas. One obvious strategy in reducing volcano risk is the total abandonment of hazardous volcanic regions for human habitation. Clearly, this is utterly unrealistic; many hazardous volcanoes are located in densely populated areas, for most of which land-use patterns have been fixed by history, culture, and tradition for centuries or millennia. Moreover, people also are exposed to potential volcano hazards by simply being passengers aboard commercial airliners flying over volcanic regions and possibly encountering a drifting volcanic ash cloud from a powerful explosive eruption [5]. Thus, the only viable option in reducing volcano risk is the timely issuance of early warning of possible impending eruptions, allowing emergency-management officials to take mitigative actions. It is then imperative for volcanologists to effectively communicate the best-possible hazards information to emergency-management authorities, and to convince them to implement timely mitigative countermeasures, including evacuation, if necessary, of people at risk. However, volcanologists face daunting challenges in providing reliable, precise early warnings that government officials demand and need to make informed decisions to ensure public safety.

Introduction

About 1500 volcanoes have erupted one or more times during the Holocene (i. e., past 10,000 years); since A.D. 1600, volcanic disasters have killed about than 300,000 people and resulted in property damage and economic loss exceeding hundreds of millions of dollars [39,40]. During recent centuries, on average 50 to 70 volcanoes worldwide are in eruption each year – about half representing activity continuing from the previous year and the other half new eruptive outbreaks [32]. Most of these erupting volcanoes are located in developing countries, which often lack sufficient scientific and economic resources to conduct adequate volcano-monitoring studies. Moreover, in any given year, many more volcanoes exhibit observable and (or) measurable anomalous behavior (*volcano unrest*) that does not lead to eruption.

For scientists and emergency-management authorities alike, the critical issues in providing early warning of possible eruption, or of possible escalation of an ongoing but weak eruption, involve basic questions such as (1) will the initiation of unrest at a currently dormant volcano culminate in eruption?; (2) if so, what would be its time of onset, eruptive mode (explosive, non-explosive, etc.) and duration?; (3) at an already restless or weakly erupting volcano, what are the prospects of escalation to major activity?; and (4) what would be the nature and severity of the associated hazards and which sectors of the volcano are the most vulnerable?

Systematic geoscience studies of a volcano's geologic history and past behavior can provide answers to some of the above-listed and related questions relevant to *long-term* forecasts of a volcano's present and possible future activity. In general, the fundamental premise applied in making long-term forecasts is that the volcano's past and present behavior provides the best clues to its future behavior. However, data from volcano monitoring – showing increased activity relative to pre-crisis or "background" levels – constitute the only scientific basis for *short-term* forecasts or predictions of a possible future eruption or a significant change during an ongoing eruption. To respond effectively to a developing volcanic crisis, timely early warnings are absolutely essential, and these warnings must be reliable and precise. Improvement in the reliability and precision of warnings can be achieved only by a greatly improved capability for eruption prediction, which in turn depends on the quantity and quality of volcano-monitoring data and the diagnostic interpretation of such information [36,38,40].

At the outset, it must be emphasized that, in this article, the term *prediction* is used in the strict sense (see Glossary) proposed by Swanson et al. [35]; thus, for a forecast or prediction to be genuine, it must be included in an official warning statement made publicly in advance of the predicted event. While retrospective forecasts and predictions serve valid scientific purposes in testing or verifying theories or models, they are of little or no use to emergency-management authorities. It is beyond the scope of this brief paper to consider in detail volcano-monitoring studies and their applications to eruption prediction. Suffice it to say that instrumental volcano monitoring is a multi-faceted, multidisciplinary endeavor that involves the ground- and (or) satellite-based measurement of seismicity, ground deformation, gas and fluid chemistry, temperature variations, secular changes in micro-gravity and geomagnetism, etc. For more information about volcano-monitoring approaches, techniques, and instrumentation, interested readers are referred to nu-

merous recent comprehensive reviews and references cited therein (e. g., [6,10,21,22,23,36] ► Volcanoes, Non-linear Processes in). Instead, this paper highlights the range in possible outcomes of volcano unrest and reviews some recent examples of the actual outcomes documented for several well-monitored volcanoes. Some implications from the observations discussed in this paper are considered in the final two parts of this article (Sect. "Challenges in Achieving Refined Predictive Capability" and "Future Directions").

Scope of Problem: Challenge for Emergency–Management Authorities

Because earthquake and eruption dynamics intrinsically involve non-linear processes (the focus of this volume), the attainment of a reliable capability to predict eruptions (especially explosive events) still eludes volcanologists. This dilemma persists despite notable advances in volcanology and the availability in recent decades of increasingly sophisticated models, stochastic and deterministic. Such models, which commonly involve pattern-recognition methodologies, are developed from theoretical or statistical analysis of volcano-monitoring data, eruption-occurrence patterns, or other time-space attributes for a volcanic system (e. g., [14,15,18,25,29,33,45,46] ► Volcanic Eruptions: Stochastic Models of Occurrence Patterns). The non-attainment of predictive capability to date reflects the reality that most volcanic systems are highly complex and that the relevant datasets for the vast majority of volcanoes are too inadequate and (or) incomplete to enable reliable and precise predictions.

Given the fact that the emerging science of eruption prediction is still nascent, scientists and emergency-management officials are at a substantial disadvantage when a long-dormant or weakly active volcano begins to exhibit unrest above its historical (i. e., "normal") levels. A principal concern is that the heightened unrest, should it persist or escalate, could culminate in renewed or greatly increased, possibly hazardous activity. Because of public-safety and socio-economic issues and consequences, emergency managers in particular must have a definitive answer to the question: What is the most likely outcome of the escalating unrest? The answer(s) to this critical question then shape(s) possible decisions they must make – within a narrow "decision window" (typically weeks to months), generally under confusing, stressful conditions – in managing the growing volcanic crisis.

Figure 1 depicts a common scenario during a volcanic crisis and lists some major factors that complicate the determination of the most likely outcome of the unrest. Unfortunately, despite advances in volcanology and volcano-monitoring techniques, scientists are unable to give emergency-management authorities the specific information they require. The greatest frustration for the authorities is that scientists cannot guarantee that an eruption will ***not*** occur. In the section below, I review some common possible outcomes of volcano unrest, and, in Sect. "Some Recent Examples of Actual Outcomes of Volcano Unrest", some actual outcomes of recent episodes of volcano unrest.

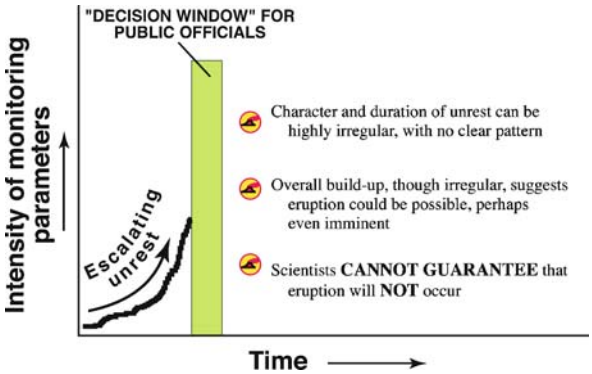

Volcanic Hazards and Early Warning, Figure 1
A common scenario confronting emergency-management authorities when unrest initiates or escalates at a volcano, as documented by increasing intensity of monitoring parameters (seismicity, ground deformation, gas emission, thermal, microgravity, etc.) during a volcanic crisis. The "decision window", though shown schematically, represents a very real and usually short timeframe during which emergency managers must decide and implement – by legal requirement, political and (or) citizen pressure, or moral conscience – strategies and actions to mitigate the potential risk from volcano hazards. (Figures 1 and 2 modified from unpublished conceptual sketches of C. Dan Miller, US Geological Survey)

Variability in Possible Outcomes of Volcano Unrest

Experience worldwide for historical eruptions demonstrates a great range in possible outcomes of volcano unrest. The three types of outcomes discussed below – for which adequate eruptive-history and volcano-monitoring data exist – probably represent the commonest possibilities. Given the non-linear nature of eruptive phenomena, however, other scenarios can be envisaged.

Culmination in Major Eruption or Return to Dormancy Following Short Duration of Unrest

Perhaps the simplest scenario – and certainly the most easily manageable for public officials – is when a volcano reawakens from a long dormancy (i. e., greater than a century) and begins to exhibit significant unrest. Such unrest

then persists at a high level for a relatively short duration (several weeks or months) before culminating in a major eruption (Fig. 2a). With adequate volcano-monitoring studies, preferably in real time or near-real time, conducted during the precursory build-up, scientists may be able to make a short-term forecast or prediction of the impending eruption. The emergency-management authorities then could have the needed guidance during the "decision window" in issuing early warnings and implementing mitigative countermeasures. Alternatively, diagnostic analysis of the volcano-monitoring data, combined with detailed knowledge of eruptive history, could lead the scientists to conclude that an eruption is not likely. In this situation, emergency managers also receive the required scientific input they need to make informed decisions to manage the volcanic crisis, including the declaration of the end of the crisis and the cancellation of any warnings in effect. In either case, scientists need sufficient information to try to determine whether an imminent eruption or a return of the volcano to dormancy is the most likely outcome. However, it must be emphasized that scientists in assessing possible outcomes can reach incorrect conclusions, even using abundant scientific data and constraints from quantitative models.

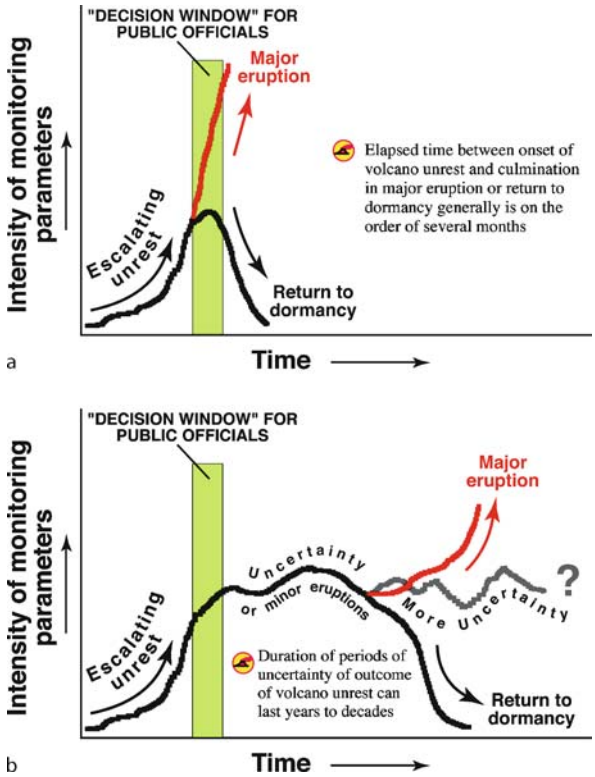

Volcanic Hazards and Early Warning, Figure 2
a Two common possible outcomes of volcano unrest that are arguably the most tractable for public officials managing a volcanic crisis (see text). b Some other possible outcomes of volcano unrest that pose much greater challenges for scientists and public officials alike during a volcanic crisis (see text)

Culmination in Major Eruption or Return to Dormancy Following a Long Lull in Unrest

A scenario fraught with greater uncertainty is one that, after an initial short period (weeks to months) of sustained and heightened volcano unrest, the volcano neither erupts nor quickly returns to quiescence. Instead, the volcanic system exhibits a decreased level of activity characterized by fluctuating but low-level unrest and (or) weak, nonhazardous eruptions that may persist for years or decades, before possibly exhibiting another episode of heightened unrest. Under such circumstances, scientists and public officials confront anew the crisis conditions and challenges inherent in attempting to anticipate the outcome of the renewed unrest.

This scenario is similar to that previously discussed above (Sect. "Culmination in Major Eruption or Return to Dormancy Following Short Duration of Unrest") but differs by the complication posed by another possible outcome in addition to those of an imminent eruption or a return to dormancy: namely, the possibility that the volcano could revert to a mode of continuing uncertainty involving only low-level activity or cessation of unrest for an indeterminate period of time (Fig. 2b). The prolongation of the periods of uncertainty has some unfortunate consequences, including (1) the potential for scientists and public officials to lose credibility by incorrectly anticipating the outcomes during the pulses of heightened unrest and by issuing warnings ("false alarms") of events that fail to materialize (see Sect. "The Dilemma of "False Alarms""); and (2) increased difficulty in maintaining public interest in, and awareness of, potential hazards posed by some eventual volcanic crisis that actually culminates in eruption.

Ongoing Unrest over Long Periods with no Clear Indication of Possible Future Activity

Some long-dormant volcanic systems undergo unrest that can persist for many years to decades, fluctuating in intensity but not showing any long-term definitive trend suggestive of possible renewed eruptive activity (Fig. 2b). It is extremely difficult, perhaps impossible, to anticipate possible outcomes of such long-duration, ongoing unrest. The reasons for this problem are simply that the eruptive his-

tories and dynamics of long-dormant systems are poorly understood and (or) their eruption frequencies are longer than the time span of the available pre-eruption *baseline* monitoring data. Some examples of such modes of volcano unrest include: Three Sisters volcanic center, Oregon, USA [48]; Long Valley caldera, California, USA [2,12]; Yellowstone caldera, Wyoming–Montana, USA [48]; and Campi Flegrei Caldera, Italy [3,42].

Some Recent Examples of Actual Outcomes of Volcano Unrest

The previous discussion of possible hypothetical outcomes of volcano unrest (Sect. "Variability in Possible Outcomes of Volcano Unrest") has its basis in several recent examples of actual volcano unrest, as documented by detailed investigations. These examples are reviewed herein (Fig. 3).

Mount St. Helens (USA) and Mount Pinatubo (Philippines)

In 1975, on the basis of many years of detailed stratigraphic and dating studies, together with an analysis of repose intervals, scientists of the US Geological Survey (USGS) made a long-term forecast of possible renewed eruptive activity at Mount St. Helens Volcano, Washington State, which had been dormant since 1857. This forecast, albeit qualitative, was remarkably prescient, stating: "…an eruption is likely within the next hundred years, possibly before the end of this century" (see p. 441 in [7]).

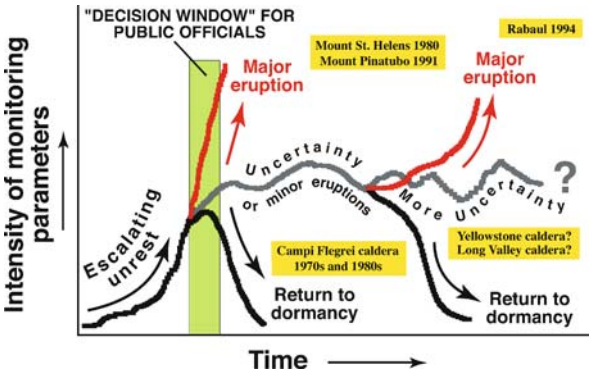

Volcanic Hazards and Early Warning, Figure 3
Some actual outcomes of recent episodes of volcano unrest at selected volcanoes. Even though these well-studied volcanoes are closely monitored in real time or near-real time by instrumental arrays, it was or is not possible to determine the outcome of the unrest at most of them (see text for discussion)

Volcanic Hazards and Early Warning, Figure 4
19 km-high eruption column of the climactic magmatic eruption of Mount St. Helens Volcano, Washington State, USA, on 18 May 1980. This eruption occurred following about two months of volcano unrest that began after a 123-year quiet period. (Photograph by Austin Post taken about noon on 18 May 1980)

Five years after the forecast, phreatic eruptions began at Mount St. Helens in late March 1980 following a week of precursory seismicity. Sustained seismicity, accompanied by large-scale ground deformation of the volcano's north flank, ultimately produced a cataclysmic magmatic eruption on 18 May 1980 (Fig. 4). This catastrophic event caused the worst volcanic disaster in the history of the United States (see [16] and chapters therein for detailed summary).

For Mount St. Helens, the volcano unrest, which was well monitored visually and instrumentally, lasted for only about three months before culminating in a major eruption (Fig. 3). Scientists monitoring Mount St. Helens were not able to make a precise prediction of the 18 May events (flank collapse, laterally directed blast, and vertical magmatic eruption). From analysis of the volcano-monitoring data and hazards assessments, however, they were able to give early warning and to convince emergency-management authorities to order evacuation and take other risk-reduction actions well within the "decision window" prior to the paroxysmal eruption.

Another illustrative example of a major eruptive activity following a brief duration of volcano unrest was the 1991 eruption of Mount Pinatubo, Luzon, Philippines (see [26], and chapters therein). Prior to 1991, Pinatubo had been quiescent for about 600 years with no recorded

historical eruptions; it had been little studied and was not monitored. In March 1991, Pinatubo began to stir from its long sleep, as indicated by increased fumarolic activity and precursory earthquakes, some large enough to be felt by local inhabitants. Phreatic explosions began in early April, and the first magmatic activity (dome intrusion) was observed on 7 June. On the basis of the analysis and interpretation of data from volcano monitoring (initiated in late April), combined with observation of increased frequency of high (> 15 km) eruption columns, scientists of the Philippine Institute of Volcanology and Seismology (PHIVOLCS) and the USGS Volcano Disaster Assistance Program (VDAP) issued warnings and made a forecast on 12 June that a major eruption was imminent. The PHIVOLCS-USGS scientific team recommended to the Commandant of the US Clark Air Base and the Philippine emergency-management authorities to order major evacuations of the air base and other population centers around the volcano. In all, more than 300,000 people were evacuated from hazardous zones. Three days later on 15 June, Pinatubo's climactic eruption occurred. Even though this eruption was the largest in the world since the 1912 eruption of Novarupta (Alaska, USA), the timely early warnings and evacuations saved many thousands of lives and minimized economic loss.

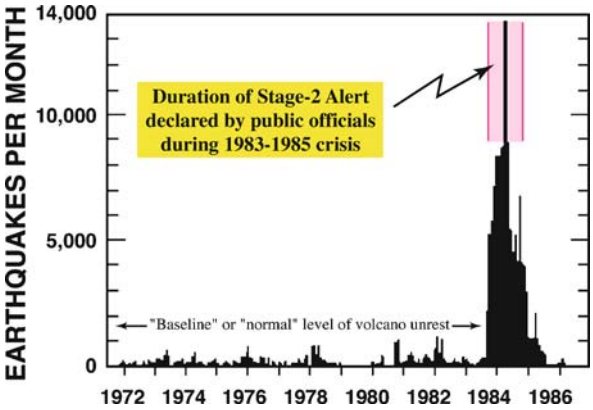

Volcanic Hazards and Early Warning, Figure 5
Occurrences of earthquakes at Rabaul Caldera, Papua New Guinea, during the period 1972–1986 (modified from McKee et al. [20]). The dramatically increased seismicity during 1983–1985, together with significant ground deformation, constituted a serious volcanic crisis and prompted public officials to declare a "Stage-2 Alert" (see text). The time span bracketed by this alert (*pink shading*) is somewhat comparable to the "decision window" schematically shown in Figs. 1–3. However, the 1983–1985 unrest did not immediately culminate in eruption, which did occur ten years later

Rabaul Caldera (Papua New Guinea)

The outcomes for the volcano unrest at Mount St. Helens (1980) and Pinatubo (1991) represent examples of simple (linear?) and relatively short progression (several months) from onset of precursory activity to major eruption. However, most other recent episodes of volcano unrest involve complex or uncertain outcomes; below I discuss one well-documented example of a complex outcome.

Rabaul caldera, Papua New Guinea, after being quiescent for several decades began to exhibit volcano unrest in the mid-1980s, as manifested by dramatically increased seismicity (Fig. 5) and accompanying caldera uplift of more than 1 m [20]. This intense and rapidly escalating unrest prompted the scientists of the Rabaul Volcanological Observatory (RVO) to warn of possible imminent eruption. In response, emergency-management authorities declared in October 1983 a "Stage-2 Alert" (Fig. 5), stating that the volcano could erupt within weeks. Accordingly, the civil authorities implemented a number of mitigative measures – including widening of roads to serve as evacuation routes, preparation of contingency plans, conduct of evacuation drills, etc. – to prepare for the expected soon-approaching eruption. However, soon after the issuance of the alert, the volcano unrest, after peaking in early 1985, began to diminish sharply and returned to normal levels within 6 months (Fig. 5). No eruption occurred, and the "Stage-2 Alert" was lifted in November 1984.

The abrupt cessation of volcano unrest and the non-occurrence of the forecasted eruption were generally viewed as a needless, disruptive "false alarm". As fallout from the unanticipated turn of events, the scientists of the RVO lost credibility with the emergency-response authorities, who in turn lost the confidence of the general population.

Ultimately however, the initially unsatisfactory outcome of the 1983–1985 unrest positively influenced the successful response to a later major eruption. After remaining at the normally low level of activity after for about ten years the mid-1980s crisis, Rabaul erupted suddenly on 19 September 1994 following only 27 hours of precursory seismicity [4]. This major eruption (Fig. 6) devastated nearly all of Rabaul City, but fortunately caused only 5 deaths. The lower-than-expected fatalities can in part be attributable to the fact that people largely "self-evacuated" without waiting for official advisories to do so. They apparently well remembered the lessons (e. g., enhanced awareness of volcano hazards, measures to take in case of eruption, evacuation drills) learned from the volcanic crisis ten years earlier.

Volcanic Hazards and Early Warning, Figure 6
View from space of the powerful explosive eruption of 19 September 1994 from Rabaul Caldera. (Photograph courtesy of astronauts aboard the NASA Shuttle)

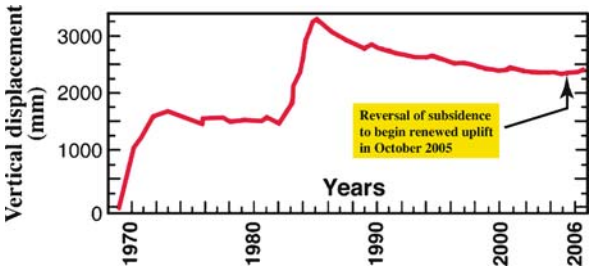

Volcanic Hazards and Early Warning, Figure 7
Uplift and subsidence at Campi Flegrei Caldera, Italy, during the period 1969–2006 (modified from Fig. 1b in [42]). Vertical displacements measured by precise leveling, together with associated horizontal displacements and increased seismicity, indicate two episodes of caldera uplift, one during 1969–1973 and another during 1982–1985. Neither episode culminated in eruption (see text). Note that the current "baseline" level of activity is higher than that prior to 1982

Ongoing Irregular Long-Duration Volcano Unrest but no Clear Indication of Possible Eruption

Recent volcano unrest at several calderas, which are well studied and systematically monitored in real time or near-real time by geophysical and geochemical networks, provide good examples of the difficulty in determining the most likely outcomes of unrest. Because of the abundance of historical observations and volcano-monitoring data, Campi Flegrei Caldera, Italy, located adjacent to Vesuvius Volcano in the Naples metropolitan area, is a particularly instructive example. During the early 1970s and again during 1983–1985, the caldera exhibited pronounced episodes of volcano unrest (Fig. 7) that generated scientific and public concern. Neither of these unrest episodes culminated in eruption, and each was followed by longer intervals of relatively low, fluctuating activity or net subsidence, hinting of an outcome whereby the restless volcano would return to dormancy.

From a recent study by Troise et al. [42], however, the post-1985 subsidence trend reversed in late 2005, and the caldera since has risen by a few centimeters (Fig. 7). Historical records, extending back in time for more than two millennia (see Fig. 1a in [42]), demonstrate a pattern of relatively short-duration uplifts (lasting years to decades) interspersed with long-lived gradual subsidence (lasting centuries). Of these several ups and downs of Campi Flegrei caldera over the past two millennia, only one episode of volcano unrest culminated in eruptive activity – the Monte Nuovo eruption in 1538 [9]. The historical behavior of Campi Flegrei emphasizes the challenging task of attempting to determine the outcome of volcano unrest with only a few years or decades of volcano-monitoring data.

Similarly, in recent years, ground-based and space-based geodetic data [e. g., **I**nterferometric **S**ynthetic **A**perture **R**adar (InSAR) and GPS], combined with seismic monitoring and other geophysical and geochemical investigations, have documented uplift and (or) subsidence at several other well-monitored long-dormant volcanic systems (e. g., Three Sisters volcanic center, Oregon; Long Valley Caldera, California; and Yellowstone Caldera, Wyoming–Montana). For more information concerning these systems as well as other volcanic systems, the interested reader is referred to the following works: Wicks et al. [48,49]; Battaglia et al. [2]; Hill [12]; and Dzurisin [10 and chapters therein]. It is important to stress that, while the level and intensity at these restless volcanic systems have varied within the periods of systematic instrumental monitoring (only a few decades at most), the available datasets and interpretations are inadequate to determine whether or not the ongoing unrest will culminate in eruption.

Kilauea Volcano, Hawaii (USA)

Brief mention should be made of frequently active volcanic systems – typically basaltic in composition – that are characterized by conspicuous and measurable unrest occurring nearly continuously over long time periods. Perhaps the best-studied volcanic systems exhibiting such behavior are those of Kilauea and Mauna Loa volcanoes on the Island of Hawaii (see [8] and chapters therein). These two volcanoes are among the most active in the world and have been well monitored since the establishment in 1912

of the Hawaiian Volcano Observatory, operated continuously by the USGS since 1948. Abundant "baseline" monitoring data, particularly for Kilauea, demonstrate well-defined patterns and outcomes of unrest (e. g., Fig. 5 in [41]; Fig. 8 in [37]). However, it must be emphasized that the behavior and outcomes of unrest observed for Kilauea – culminating either in subsurface intrusion only or in eruption – occur over much shorter time scales than those for the examples of explosive volcanoes discussed in this paper (e. g., Figs. 2, 3, and 7).

For Kilauea, because of its high eruption frequency and, at times, nearly continuous activity (e. g., [11]), the outcomes of its volcano unrest and the need to issue early warnings are manageable for the civil authorities with little or no anxiety. This fortunate situation reflects a positive circumstance: because of the volcano's frequent and, at times, continuous eruptive activity, the dissemination of hazards information and early warnings is easily and routinely accomplished among the scientists, emergency-management officials, news media, and the affected public. Moreover, analyzes of data for well-monitored, frequently active volcanoes, such as Kilauea, provide many case histories that can serve to test and refine volcano-monitoring techniques as well as eruption-forecasting methodologies. Nonetheless, it should be noted that eruptions at Kilauea and other basaltic volcanoes differ fundamentally from those of more explosive – hence, potentially more dangerous – stratovolcanoes that erupt more viscous materials of intermediate composition (e. g., Mount St. Helens, Mount Pinatubo).

Challenges in Achieving Refined Predictive Capability

The discussion in the preceding sections paints a somewhat discouraging picture: namely, possibly with very rare exceptions (e. g., the Mount St. Helens and Pinatubo case histories), to date scientists still lack a reliable capability for precisely *and* accurately predicting the outcomes of unrest at explosive volcanoes with low eruption frequency. This unsatisfactory state of affairs does not well serve societal needs to reduce the risks of volcanic hazards and demands improvement. Considered below are some elements that I deem to be essential to achieve refined predictive capability and early warnings of possible eruptive activity.

More Geologic Mapping and Dating Studies of Volcanoes

Of the world's active and potentially active volcanoes, only a handful (mostly in the developed countries) have been studied sufficiently for the detailed and complete reconstruction of their past behavior, eruption frequencies, and associated hazards. Specifically, geologic mapping (at a scale of 1:25,000 or better) and dating studies of stratigraphically well-controlled eruptive products have been done for only a few volcanoes. Recorded historical eruptions for most volcanoes – especially those in the New World – rarely extend back in time for more than five centuries. Their eruptive histories generally are too short to serve as a comprehensive time-series dataset for quantitative statistical analysis of eruption frequency or repose intervals. Thus, it is essential to expand the time-series dataset by obtaining radiometric ages of prehistoric eruptive products. Moreover, in studies of eruption frequency, no matter how abundant the data used, the calculated frequencies obtained are necessarily *minima*, because the products of small eruptive events are lost to erosion and not preserved in the geologic record.

More Volcano Monitoring at More Volcanoes: Importance of Establishing Long-Term "Baseline" Behavior

As noted by Tilling (see p. 9 in [40]), most volcanologists believe that "…if a volcano is monitored extensively in real- or near-real time by well-deployed instrumental networks, it should be possible to make much better forecasts of the outcomes of volcano unrest." Unfortunately, many of the world's active and potentially active volcanoes are poorly monitored, and some are not monitored at all. Thus, to advance our understanding of eruptive dynamics in order to make more precise early warnings, initiating or expanding volcano monitoring at many more volcanoes is imperative. Ideally, volcano monitoring should be conducted in real- or near-real time, and it should involve a combination of monitoring techniques rather than reliance on any single one [36]. Moreover, the longer the time span of volcano monitoring, the more diagnostic is the relevant database needed for detecting and interpreting any significant departure from the overall variation in the long-term "baseline" behavior of the volcano.

The Dilemma of "False Alarms"

With the current state-of-the-art in volcanology and monitoring techniques, it is difficult, and generally impossible, for volcanologists to determine with any certainty: (1) the most likely outcome of onset of unrest at a long-dormant volcano; and (2) the eventual possible occurrence of a major eruption in the course of long-duration, but relatively low-level ongoing unrest. Nonetheless, scientists still have an obligation to emergency managers and the threatened populations to provide the best-possible hazards in-

formation, forecasts or predictions, and warnings within the limitations of available knowledge, technology, and volcano-monitoring techniques. In so doing, the scientists run the unavoidable risk of raising a "false alarm" and the associated criticism and loss of credibility when the forecasted event fails to materialize. On the other hand, emergency-management officials as well as the general public must be prepared to accept the disruptive socio-economic consequences and costs of occasional "false alarms". The Rabaul case history (previously discussed in Sect. "Rabaul Caldera (Papua New Guinea)") underscores the sentiment of Banks et al. (see p. 78 in[1]): "…False alarms themselves can provide, through objective assessment of the scientific and public response to a volcanic crisis that ended without eruption, valuable lessons useful … for the next crisis, which could culminate in an eruption." Even so, every attempt should be made to minimize "false alarms" to the extent possible, in order to maintain scientific credibility and the confidence of the emergency-management authorities and the populace at risk.

Future Directions

An obvious critical need is the comprehensive geoscience studies of more volcanoes, to reconstruct their prehistoric eruptive histories, thereby extending the time-series database needed for quantitative analysis and interpretation. In addition, volcano monitoring, even if rudimentary – deploying only 1 to 3 seismometers and making simple field observations and measurements [34] – is needed for many more volcanoes, both to establish "baselines" and to provide early warning. Obviously, future research necessarily should also be focused on methodologies to make more robust the acquisition and processing of real- or near-real-time volcano-monitoring data, using or refining ***existing*** instrumentation and techniques, volcano databases, analytical tools, and empirical and theoretical models. An especially promising avenue of future research would require the wider utilization of broadband instruments in seismic-monitoring networks at active and potentially volcanoes. This would enable systematic and quantitative analysis and interpretation of Long-Period (LP) and Very-Long-Period (VLP) seismicity, which from worldwide experience precedes and accompany nearly all eruptions ► Volcanoes, Non-linear Processes in.

A long-standing and vexing problem for volcanologists is that, at present, no geophysical or geochemical criteria exist to identify the distinguishing characteristics between volcano unrest that merely ends in a subsurface intrusion of magma and unrest that culminates with magma breaking the surface to produce an eruption. From the vantage point of the emergency-management officials, subsurface intrusion or movement of magma, while of great academic interest to volcanologists in deciphering volcano dynamics, poses no threat to public safety. In contrast, the threat of an imminent eruption requires timely decisions and mitigative actions to ensure public safety. Thus, the future development, if possible, of quantitative criteria to distinguish the precursory pattern of a subsurface intrusion from that for an eruption would mark a quantum leap in the young science of volcanology. In this regard, an improved understanding of magma intrusion *vs.* eruptive dynamics might provide scientists more powerful tools in choosing between the various paths shown in Fig. 3 as the most likely outcome of volcano unrest.

In attempting to quantify estimates of the probabilities of specified hazardous volcanic events within specified timeframes (i. e., eruption predictions), scientists have made increasing use of "materials-failure" models (e. g., [25,45,46]), "event trees" (e. g., [27]), and more statistically rigorous variants of the "event-tree", "decision-tree", "pattern-recognition", and "occurrence-pattern" approaches (e. g., [17,18,19,24,29] ► Volcanic Eruptions: Stochastic Models of Occurrence Patterns). Depending on the particular circumstances of the volcano(es) and eruptions under study, these and other statistical tools have their individual strengths and weaknesses. Taken together, however, the utility and potential successful application of such approaches are inherently hampered by the incompleteness and (or) too-short time span of the eruptive-history or volcano-monitoring datasets analyzed. Thus, to become more robust, the probabilistic methodologies must employ larger and more complete datasets. In any case, while useful, the existing techniques or models cannot yield results with the precision and accuracy that emergency-managers demand and need as they confront a "decision window" during a volcanic crisis. Simply stated, even if the calculated results prove to be accurate, their error bars generally exceed the time span bracketed by the "decision window".

In recent years, the World Organization of Volcano Observatories (WOVO) – a Commission of the International Association of Volcanology and Chemistry of the Earth's Interior – has launched an ambitious but poorly funded effort, called *WOVOdat*, to construct a global database of empirical data on volcanic unrest. This database is being developed as a user-friendly Internet-based system with query and analysis tools [30]. Even though a basic design and schema now exist for the construction of *WOVOdat* [44], the actual compilation and integration of data are just barely beginning. The *WOVOdat* project recently received major funding from

the Singapore Government and Nanyang Technological University [28]. This favorable development brightens the prospects for it ultimately to become a powerful tool in evaluating patterns and possible outcomes of volcanic unrest. Nonetheless, until *WOVOdat* becomes fully operational, we must continue to rely on the available datasets – however imperfect – and to employ and refine the existing analytical methodologies and statistical models.

Acknowledgments

This article has benefited from constructive reviews and helpful suggestions by L. J. Patrick Muffler and Fred Klein (both of the US Geological Survey, Menlo Park) on an earlier draft. To them, I offer them my sincere thanks. The views expressed in this article have been shaped by my personal involvement in responses to several of the volcanic crises in recent decades, and by enlightening and instructive interactions and discussions with many colleagues in the global volcanological community.

Bibliography

Primary Literature

1. Banks NG, Tilling RI, Harlow DH, Ewert JW (1989) Volcano monitoring and short-term forecasts. In: Tilling RI (ed) Short Courses in Geology, vol 1. Volcanic Hazards, American Geophysical Union, Washingnton, DC, chap 4, pp 51–80
2. Battaglia M, Roberts C, Segall P (2003) The mechanics of unrest at Long Valley caldera, California: 2. Constraining the nature of the source using geodetic and micro-gravity data. J Volcanol Geotherm Res 127:219–245
3. Berrino G, Corrado G, Luongo G, Toro B (1984) Ground deformation and gravity changes accompanying the 1982 Pozzuoli uplift. Bull Volcanol 47:187–200
4. Blong R, McKee C (1995) The Rabaul eruption 1994: Destruction of a Town. Natural Hazards Research Centre, Macquarie University, Sydney, 52 pp
5. Casadevall TJ (ed) (1994) Volcanic ash and aviation safety: Proceedings of the First International Symposium on Volcanic Ash and Aviation Safety. US Geological Survey Bulletin 2047. Government Printing Office, Washington, 450 pp
6. Chouet B (2004) Volcano seismology. Pure Appl Geophys 160:739–788
7. Crandell DR, Mullineaux DR, Rubin M (1975) Mount St. Helens volcano: Recent and future behaviour. Science 187:438–441
8. Decker RW, Wright TL, Stauffer PH (eds) (1987) Volcanism in Hawaii. US Geological Survey Professional Paper 1350. vol 1 and 2. U.S. Government Printing Office, Washington, 1667 pp
9. De Vito M, Lirer L, Mastrolorenzo G, Rolandi G (1987) The 1538 Monte Nuovo eruption (Campi Flegrei, Italy). Bull Volcanol 49:608–615
10. Dzurisin D (ed) (2006) Volcano deformation: Geodetic Monitoring Techniques. Springer-Praxis, Berlin, 441 pp
11. Heliker C, Swanson DA, Takahashi TJ (eds) (2003) The Pu'u O'o-Kupaianaha Eruption of Kilauea Volcano, Hawaii: The first 20 years. US Geological Survey Professional Paper 1676. U.S. Geological Survey, Reston, 206 pp
12. Hill DP (2006) Unrest in Long Valley Caldera, California, 1978–2004. Geol Soc London 269:1–24
13. ISDR (2004) Living with Risk: International Strategy for Disaster Reduction (ISDR). United Nations, New York and Geneva vol 1, 431 pp and vol 2, 126 pp
14. Klein F (1982) Patterns of historical eruptions at Hawaiian volcanoes. J Volcanol Geotherm Res 12:1–35
15. Klein F (1984) Eruption forecasting at Kilauea Volcano, Hawaii. J Geophys Res 89(B5):3059–3073
16. Lipman PW, Mullineaux DR (eds) (1981) The 1980 eruptions of Mount St. Helens, Washington. US Geological Survey Professional Paper 1250. U.S. Government Printing Office, Washington, 844 pp
17. Marzocchi W (1996) Chaos and stochasticity in volcanic eruptions the case of Mount Etna and Vesuvius. J Volcanol Geotherm Res 70:205–212
18. Marzocchi W, Sandri L, Gasparini P, Newhall C, Boschi E (2004) Quantifying probabilities of volcanic events: The example of volcanic hazards at Mt. Vesuvius. J Geophys Res 109:B11201. doi:10.1029/2004JB003155
19. Marzocchi W, Sandri L, Selva J (2008) BET_EF: A probabilistic tool for long- and short-term eruption forecasting. Bull Volcanol 70(5):623–632. http://dx.doi.org/
20. McKee CO, Johnson RW, Lowenstein PL, Riley SJ, Blong RJ, de St. Ours P, Talai B (1985) Rabaul caldera, Papua New Guinea: Volcanic hazards, surveillance, and eruption contingency planning. J Volcanol Geotherm Res 23:195–237
21. McNutt SR (1996) Seismic monitoring and eruption forecasting of volcanoes: A review of the state of the art and case histories. In: Scarpa R, Tilling RI (eds) Monitoring and Mitigation of Volcano Hazards. Springer, Heidelberg, pp 99–146
22. McNutt SR (2000) Seismic monitoring. In: Sigurdsson H, Houghton B, McNutt SR, Rymer H, Stix J (eds) Encyclopedia of Volcanoes. Academic Press, San Diego, chap 68, pp 1095–1119
23. McNutt SR (2000) Synthesis of volcano monitoring. In: Sigurdsson H, Houghton B, McNutt SR, Rymer H, Stix J (eds) Encyclopedia of Volcanoes. Academic Press, San Diego, chap 71, pp 1167–1185
24. Mulargia F, Gasparini P, Marzochhi W (1991) Pattern recognition applied to volcanic activity: Identification of the precursory patterns to Etna recent flank eruptions and periods of rest. J Volcanol Geotherm Res 45:187–196
25. Murray JB, Ramirez Ruiz JJ (2002) Long-term predictions of the time of eruptions using remote distance measurement at Volcán de Colima, México. J Volcanol Geotherm Res 117(1–2): 79–89
26. Newhall CG, Punongbayan RS (eds) (1996) Fire and Mud: Eruptions and Lahars of the Mount Pinatubo, Philippines. Philippine Institute of Volcanology and Seismology, Quezon City, and University of Washington Press, Seattle, 1126 pp
27. Newhall CG, Hoblitt RP (2002) Constructing event trees for volcanic crises. Bull Volcanol 64:3–20
28. Newhall CG (2008) written communication, 11 February 2008
29. Sandri L, Marzocchi W, Zaccarelli L (2004) A new perspective in identifying the precursory patters of eruptions. Bull Volcanol 66:263–275
30. Schwandner FM, Newhall CG (2005) *WOVOdat*: The World Organization of Volcano Observatories database of volcanic un-

rest. European Geosciences Union, Geophysical Research Abstracts, vol 7, abstract # 05-J-09267 (extended abstract), 2 pp
31. Sigurdsson H, Houghton B, McNutt SR, Rymer H, Stix J (eds) (2000) Encyclopedia of Volcanoes (and chapters therein). Academic Press, San Diego, 1417 pp
32. Simkin T, Siebert L (1994) Volcanoes of the World: A Regional Directory, Gazetteer, and Chronology of Volcanism During the Last 10,000 Years, 2nd edn. Smithsonian Institution, Washington and Geoscience Press, Inc., Tucson, Arizona, 349 pp
33. Sparks RSJ (2003) Frontiers: Forecasting volcanic eruptions. Earth Planet Sci Lett 210(1–2):1–15
34. Swanson DA (1992) The importance of field observations for monitoring volcanoes, and the approach of "Keeping Monitoring as Simple as Practical". In: Ewert JW, Swanson DA (eds) Monitoring volcanoes: Techniques and Strategies used by the staff of the Cascades Volcano Observatory, 1980–90. US Geological Survey Bulletin, vol 1966. U.S. Government Printing Office, Washington, 219–223
35. Swanson DA, Casadevall TJ, Dzurisin D, Holcomb RT, Newhall CG, Malone SD, Weaver CS (1985) Forecasts and predictions of eruptive activity at Mount St. Helens, USA: 1975–1984. J Geodyn 3:397–423
36. Tilling RI (1995) The role of monitoring in forecasting volcanic events. In: McGuire WJ, Kilburn CRJ, Murray JB (eds) Monitoring Active Volcanoes: Strategies, Procedures, and Techniques. UCL Press, London, pp 369–402
37. Tilling RI (2002) Volcanic Hazards. In: Meyer RA (ed) Encyclopedia of Physical Science and Technology, vol 17, 3rd edn. Academic Press, San Diego, pp 559–577
38. Tilling RI (2003) Volcano monitoring and eruption warnings. In: Zschau J, Küppers AN (eds) Early Warning Systems for Natural Disaster Reduction. Springer, Berlin, chap 5, pp 505–510
39. Tilling RI (2005) Volcano Hazards. In: Martí J, Ernst G (eds) Volcanoes and the Environment. Cambridge University Press, Cambridge, chap 2, pp 56–89
40. Tilling RI (2008) The critical role of volcano monitoring in risk reduction. Adv Geosci 14:3–11
41. Tilling RI, Dvorak JJ (1993) The anatomy of a basaltic volcano. Nature 363:125–133
42. Troise C, De Natale G, Pingue F, Obrizzo F, De Martino P, Tammaro U, Boschi E (2007) Renewed ground uplift at Campi Flegrei caldera (Italy): New insight on magmatic processes and forecast. Geophys Res Lett 34:L03301. doi:10.1029/2006GL28545
43. Turner MB, Cronin SJ, Bebbington MS, Platz T (2008) Developing probabilistic eruption forecasts for dormant volcanoes: A case study from Mt. Taranaki, New Zealand. Bull Volcanol 70:507–515
44. Venezky DY, Newhall CG (2007) WOVOdat design document: The schema, table descriptions, and create table statements for the database of worldwide volcanic unrest (WOVOdat version 1.0). US Geological Survey Open-File Report 2007-1117, 184 pp
45. Voight B (1988) A method for prediction of volcanic eruptions. Nature 332:125–130
46. Voight B, Cornelius RR (1991) Prospects for eruption prediction in near real-time. Nature 350:695–698
47. WG99, Working Group on California Earthquake Probabilities (1999) Earthquake probabilities in the San Francisco Bay region: 2000 to 2030 – A summary of findings. US Geological Survey Open-File Report 99-517, 60 pp
48. Wicks C, Dzurisin D, Ingebritsen S, Thatcher W, Lu Z, Iverson RM (2002) Magmatic activity beneath the quiescent Three Sisters volcanic center, central Oregon Cascade Range, USA. Geophys Res Lett 29(7):26–751,26–754. doi:10.1029/2001GL014205
49. Wicks C, Thatcher W, Dzurisin D, Svarc J (2006) Uplift, thermal unrest, and magmatic intrusion at Yellowstone Caldera. Nature 440:72–75

Books and Reviews

Chester D (1993) Volcanoes and Society. Edward Arnold (a Division of Hodder & Stoughton), London, 351 pp
Martí J, Ernst G (eds) (2005) Volcanoes and the Environment. Cambridge University Press, Cambridge, 471 pp
Scarpa R, Tilling RI (eds) (1996) Monitoring and Mitigation of Volcano Hazards. Springer, Heidelberg, 841 pp

Volcanoes, Non-linear Processes in

Bernard Chouet
US Geological Survey, Menlo Park, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Description of Seismic Sources in Volcanoes
Sources of Long-Period Seismicity
Source Processes of Very-Long-Period Signals
Future Directions
Acknowledgment
Bibliography

Glossary

Bubbly liquid Liquid-gas mixture in which the gas phase is distributed as discrete bubbles dispersed in a continuous liquid phase.

Choked flow Flow condition where the transonic flow of a compressible fluid through a constriction becomes choked. As the fluid flows through a narrowing channel, it accelerates at a rate that depends on the ratio of channel inlet pressure to channel outlet pressure. At a critical pressure ratio, which is a function of the fractional change of cross-sectional area, the speed of the fluid accelerating through the nozzle-like constriction reaches a maximum value equal to its sound speed. This flow condition is termed choked flow.

Crack wave A dispersive wave generated by fluid-solid interaction in a fluid-filled crack embedded in an elastic solid. The crack-wave speed is always smaller than the acoustic speed of the fluid.

Diffuse interface Thin mixing layer constituting the interface between two immiscible fluids. Interface dynamics is governed by molecular forces and is described by the mixing energy of the two fluid components.

Gas slug A gas bubble whose diameter is approximately the diameter of the pipe in which the slug is flowing. In a slug ascending a vertical pipe, the nose of the bubble has a characteristic spherical cap and the gas in the rising bubble is separated from the pipe wall by a falling film of liquid.

Long-period (LP) event A seismic event originating under volcanoes with emergent onset of *P* waves, no distinct *S* waves, and a harmonic signature with periods in the range 0.2–2 s as compared to a tectonic earthquake of the same magnitude. LP events are attributed to the involvement of fluid such as magma and/or water in the source process (see also VLP event).

Magma Molten rock containing variable amounts of dissolved gases (principally water vapor, carbon dioxide, and sulfur dioxide), crystals (principally silicates and oxides), and, occasionally, preexisting solid rock fragments.

Moment tensor A symmetric second-order tensor that completely characterizes an internal seismic point source. For an extended source, it represents a point-source approximation and can be quantified from an analysis of seismic waves whose wavelengths are much longer than the source dimensions.

Nonlinear process Process involving physical variables governed by nonlinear equations that reflect the fundamental micro-scale dynamics of the system. Nonlinear phenomena can arise from a number of different sources, including convective acceleration in fluid dynamics, constitutive relations, boundary conditions representing gas-liquid interfaces, nonlinear body forces, and geometric nonlinearities arising from large-scale deformation.

Phase-field method Method treating the interface between two immiscible fluids as a diffuse thin layer; also known as the diffuse interface method. In this approach, the physical and chemical properties of the interfacial layer are defined by a phase-field variable ϕ, whose dynamics are expressed by the Cahn–Hilliard convection-diffusion equation. Use of the phase-field variable avoids the necessity of tracking the interface and yields the correct interfacial tension from the free energy stored in the mixing layer.

Very-long-period (VLP) event A seismic event originating under volcanoes with typical periods in the range 2–100 s. VLP events are attributed to the involvement of fluid such as magma and/or water in the source process. Together with LP events, VLP events represent a continuum of fluid oscillations, including acoustic and inertial effects resulting from perturbations in the flow of fluid through conduits.

Volatile A chemical compound that is dissolved in magma at high pressure and exsolves from the melt and appears as a low-density gas at low pressure. The most common volatiles in magma are water, carbon dioxide, and sulfur dioxide.

Waveform inversion Given an assumed model of the wave-propagation medium, a procedure for determining the source mechanism and source location of a seismic event based on matching observed waveforms with synthetics calculated with the model.

Definition of the Subject

Magma transport is fundamentally episodic in character as a result of the inherent instability of magmatic systems at all time scales. This episodicity is reflected in seismic activity, which originates in dynamic interactions between gas, liquid and solid along magma transport paths involving complex geometries. The geometrical complexity plays a central role in controlling flow disturbances and also providing specific sites where pressure and momentum changes in the fluid are effectively coupled to the Earth.

Recent technological developments and improvements in the seismological instrumentation of volcanoes now allow the surface effects of subterranean volcanic processes to be imaged in unprecedented detail. Through detailed analyzes of the seismic wavefields radiated by volcanic activity, it has become possible for volcano seismologists to make direct measurements of the volcanic conduit response to flow processes, thus opening the way for detailed modeling of such processes.

However, unlike the description of seismic waves, which is based on the linear equations of elastodynamics, the description of the flow processes underlying the observed seismic source mechanisms is governed by the nonlinear equations of fluid dynamics. In the classic Navier–Stokes equations of fluid mechanics, the nonlinearity resides in the convective acceleration terms in the equations of conservation of momentum. In volcanic fluids, further complexity arises from the strong nonlinear dependence of magma rheology on temperature, pressure, and water and crystal content, and nonlinear characteristics of associated processes underlying the physico-chemical evolution of liquid-gas mixtures constituting magma. An example of nonlinearity in two-phase fluid mixtures

are the changing boundary conditions associated with the internal surfaces separating the phases, a situation commonly encountered during the spinodal decomposition of binary fluids [61]. Spinodal decomposition occurs when a mixture of two species A and B of fluid forming a homogeneous mixture at high temperature undergoes spontaneous demixing following a temperature drop (quench) below some critical temperature. As the temperature drops below the critical temperature, the initially homogeneous mixture becomes locally unstable, and a period of interdiffusion is initiated, which leads to the formation of patches of A-rich and B-rich fluids separated by sharp interfaces. In the late stages of demixing following a deep quench, the interfacial thickness reaches a molecular scale and the interfacial tension approaches an equilibrium surface tension. As it approaches local equilibrium, the mixture continues to evolve toward a minimization of energy by reducing its interfacial area, and stresses arising from changes in the local interfacial curvature drive fluid motion. A smooth evolution of the interfacial structure ensues, occasionally interrupted by local coalescence, breakup and reconnection, until final equilibrium is reached between the two bulk domains. Similar processes underlie the physics that governs the vesiculation, fragmentation, and collapse of bubble-rich suspensions to form separate melt and vapor in response to decreasing pressure along the ascent path of magma.

All of the above factors contribute to a diversity of oscillatory processes that intensifies as magma rheology becomes more complex; increasingly diverse behavior also characterizes hydrothermal fluids. Heat transfer to the hydrothermal system from magmatic gases streaming through networks of fractures pervading the rock matrix may induce boiling in ground water, resulting in the formation of steam bubbles and attendant bubble oscillations, thereby generating sustained tremor of hydrothermal origin.

Refined understanding of magma and hydrothermal transport dynamics therefore requires multidisciplinary research involving detailed field measurements, laboratory experiments, and numerical modeling. The comprehensive breadth of seismology must be used to understand and interpret the wide variety of seismic signals encountered in various volcanic processes. This effort must then be complemented by laboratory experiments and numerical modeling to better understand the complex flow dynamics at the origin of the source mechanisms imaged from seismic data. Such research is fundamental to monitoring and interpreting the subsurface migration of magma that often leads to eruptions, and thus enhances our ability to forecast hazardous volcanic activity.

Introduction

Degassing is the main driving force behind most volcanic phenomena. The separation of vapor and melt phases leads to the formation of bubbles, the presence of which decreases magma density, enhances magma buoyancy and propels magma ascent [140]. Bubbles in magma also naturally increase magma compressibility, resulting in a low acoustic speed of the bubbly mixture compared to that of the liquid phase [28,36,37,56]. At shallow depths, the high compressibility of the bubbles may further contribute to the generation of pressure oscillations in bubble clouds [28,145]. Bubble coalescence and fragmentation can also contribute to pressure disturbances [85]. In hydrothermal fluids, the collapse of small vapor bubbles can be the source of pressure pulses and sustained tremor [70,72].

Bubble dynamics play a key role in the transport of magmatic and hydrothermal fluids, not only as sources of acoustic energy, but also in providing a sharp contrast in velocity between the fluid and encasing solid, which facilitates the entrapment of acoustic energy at the source. This aspect of the source often manifests itself in the form of long-lasting, long-period harmonic oscillations produced by sustained resonance at the source.

Other types of gaseous fluid mixtures also favor source resonance. For example, gases laden with solid particles, or gases mixed with liquid droplets, may produce velocity contrasts that are similar to, or stronger, than those associated with bubbly liquids. Indeed, dusty gases made of micron-sized particles, or misty gases made of micron-sized droplets, can sustain resonance at the source over durations that far exceed those achieved with bubbly fluids [77].

Large gas slugs bursting at the liquid surface can act as active sources of acoustic and seismic radiation [59,118,136]. Gas slugs traversing discontinuities in conduit geometry can also cause transient liquid motions inducing pressure and momentum changes [60], which radiate elastic waves via their coupling to the conduit walls [32,35,109]. In magmas where a high fluid viscosity impedes flow, the diffusion of volatiles from a supersaturated melt may result in gradual pressurization of the melt and attendant deformation of the encasing rock matrix [33,34,104,124]. The fluid motions resulting from unsteady slug dynamics, large degassing bursts, and diffusion-dominated processes typically produce signals with characteristic periods longer than those commonly associated with acoustic resonance.

Oscillatory processes are thus ubiquitous during magma flow and are a natural expression of the release of

thermo-chemical and gravitational energy from volcanic fluids. The radiated elastic energy reflects both active and passive processes at the source, and features a large variety of signals over a wide range of periods.

Seismic signals originating in the dynamics of magmatic and hydrothermal fluids typically include Long-Period (LP) events and tremor [27,28]. LP events resemble small tectonic earthquakes in duration but differ in their characteristic frequency range and harmonic signature. Tremor is characterized by a harmonic signal of sustained amplitude lasting from minutes to days, and sometimes for months or even longer. In many instances, LP events and tremor are found to have essentially the same temporal and spectral components, suggesting that a common source process, differing only in duration, underlies these two types of events. Accordingly, LP events and tremor are often grouped under the common appellation LP seismicity. The periods in which LP seismicity is observed typically range from 0.2 to 2 seconds, and the characteristic oscillations of LP signals are commonly viewed as a result of acoustic resonance in a fluid-filled cavity. Slower processes involving inertial forces associated with unsteady mass transport may radiate seismic signals extending to periods much longer than 2 seconds. These types of signals commonly fall under the appellation of Very-Long-Period (VLP) seismicity. Collectively, LP and VLP seismicities provide a comprehensive view of mass transport dynamics under a volcano.

The present chapter offers a brief review of the state of the art in volcano seismology and addresses basic issues in the quantitative interpretation of processes operative in active volcanic systems. The chapter starts with an introduction of the seismic methodology used to quantify the source of volcano seismicity, with emphasis on sources originating in the dynamics of volcanic fluids. A review of some of the representative source mechanisms of LP and VLP signals and of their implications for volcanic processes follows in Sects. "Sources of Long-Period Seismicity" and "Source Processes of Very-Long-Period Signals". A final section describes a mesoscale computational approach for simulating two-phase flows of complex magmatic fluids. This method, called the phase-field method, grew out of the original work of [19], and relies on an energy-based variational formalism to treat interfacial dynamics and complex rheology in a unified framework. The phase-field method has been successfully applied to the modeling of complex micro-structured two-phase fluids in the field of engineering materials and appears well adapted to the numerical simulation of nonlinear volcanic processes.

Description of Seismic Sources in Volcanoes

A general kinematic description of seismic sources in volcanoes is commonly based on a moment-tensor and single-force representation of the source [3] (Fig. 1a). The seismic-moment tensor consists of nine force couples, with each corresponding to one set of opposing forces (dipoles or shear couples). This symmetric second-order tensor allows a description of any generally oriented dis-

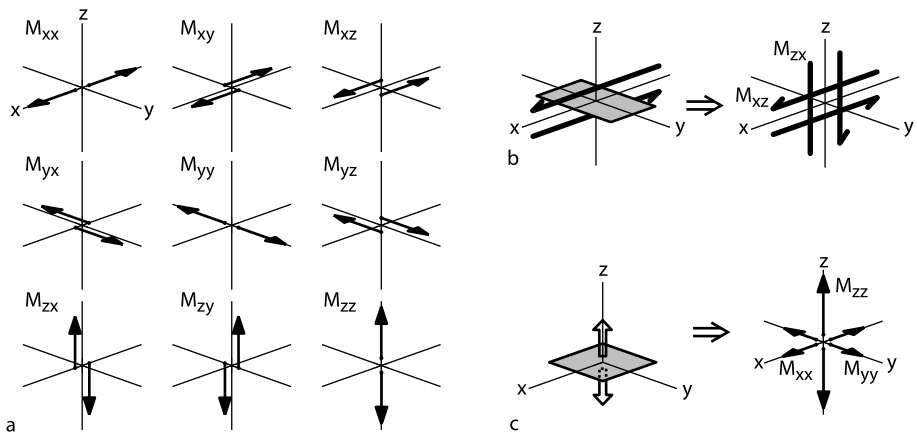

Volcanoes, Non-linear Processes in, Figure 1

a The nine possible couples corresponding to the moment-tensor components describing the equivalent force system for a seismic source in the Earth. **b** Slip on a fault can be described by a superposition of two force couples, in which each force couple is represented by a pair of forces offset in the direction normal to the force. Sources involving slip on a fault thus have an equivalent force system in the form of a double couple composed of four forces. **c** A tensile crack has a representation in the form of three vector dipoles, in which each dipole consists of a pair of forces offset in the direction of the force

continuity in the Earth in terms of equivalent body forces. For example, slip on a fault can be represented by an equivalent force system involving a superposition of two force couples of equal magnitudes – a double couple (Fig. 1b). Similarly, a tensile crack has an equivalent force system made of three vector dipoles, with dipole magnitudes with ratios 1:1:$(\lambda + 2\mu)/\lambda$, in which λ and μ are the Lamé coefficients of the rock matrix [28], and where the dominant dipole is oriented normal to the crack plane (Fig. 1c). Injection of fluid into the crack will cause the crack to expand and act as a seismic source. In general, magma movement between adjacent segments of conduit can be represented through a combination of volumetric sources of this type. Because mass-advection processes can also generate forces on the Earth, a complete description of volcanic sources commonly requires the consideration of single forces in addition to the volumetric source components expressed in the moment tensor. For example, a volcanic eruption can induce a force system that includes a contraction of the conduit/reservoir system in response to the ejection of fluid, and a reaction force from the eruption jet [64] (Fig. 2). Some volcanic processes can be described by a single-force mechanism only. An example is the single-force mechanism attributed to the massive landslide observed at the start of the 1980 eruption of Mount St. Helens [63,66]. A single-force source model was also proposed by Uhira et al. [134] to explain the mechanism of dome collapses at Unzen Volcano, Japan. In general terms, a single force can be generated by an exchange of linear momentum between the source and the rest of the Earth [129]. Thus, a single force on the Earth may result from an acceleration of the center of mass of the source in the direction opposite to the force, or deceleration of the center of mass of the source in the same direction as the force [129].

Waveform Inversion

Waveform inversions solving for the amplitudes and time histories of the moment-tensor and single-force components of the source have become the primary tool to

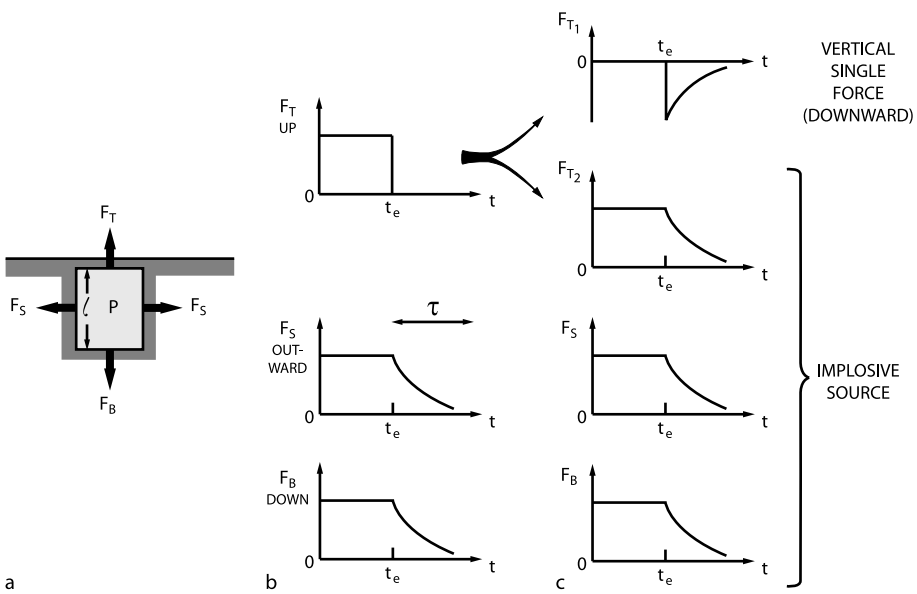

Volcanoes, Non-linear Processes in, Figure 2

Force system equivalent to a volcanic eruption (after [64]). a Pressurized cavity model for a volcanic eruption. A shallow vertically oriented cylindrical cavity initially sealed at the top by a lid contains a pressurized inviscid fluid which exerts an upward vertical force F_T on the lid, a horizontal outward force F_S on the sidewall, and a vertical downward force F_B on the bottom of the cylinder. b Time histories of forces acting on the top, side, and bottom cavity walls of the cavity. The eruption is simulated by the sudden removal of the lid at time $t = t_e$, at which point the force F_T vanishes instantaneously and the fluid pressure in the cylinder starts to decrease with a characteristic time constant τ fixed by the mass flux of the eruption, i. e., $\tau \sim \ell/v$, where ℓ is the length of the cylinder and v is the mean fluid velocity inside the cylinder. Since the forces F_S and F_B are both proportional to pressure, they decrease with the same time constant τ. c Decomposition of the force system to a downward single force and implosive source. The force F_T in b is decomposed into a vertical downward component F_{T_1} and vertical upward component F_{T_2} in such a way that F_{T_2} has the same time history as F_S and F_B. As a result, the three forces F_{T_2}, F_S, and F_B form an implosive source so that the eruption mechanism is represented by the superposition of a downward vertical force (the reaction force of the volcanic jet) with this volumetric implosion

identify and understand the mechanisms of generation of LP and VLP signals recorded by broadband seismometers. When the wavelengths of observed seismic waves are much longer than the spatial extent of the source, the source may be approximated by a point source, and the force system represented by the moment-tensor and single-force components is localized at this point. The waveform inversion procedure requires a calculation of the impulse responses (Green's functions) of the medium to this force system for each receiver component. The displacement field generated by a seismic source is described by the representation theorem which, for a point source, may be written as e. g., Chouet, [28]

$$u_n(t) = F_p(t) * G_{np}(t) + M_{pq}(t) * G_{np,q}(t)\,, \qquad p, q = x, y, z\,, \quad (1)$$

where $u_n(t)$ is the n-component of seismic displacement at a receiver at time t, $F_p(t)$ is the time history of the force applied in the p-direction, $M_{pq}(t)$ is the time history of the pq-component of the moment tensor, and $G_{np}(t)$ is the Green tensor which relates the n-component of displacement at the receiver position with the p-component of impulsive force at the source position. The notation, q indicates spatial differentiation with respect to the q-coordinate and the symbol $*$ denotes convolution. Summation over repeated indices is implied.

A common approach to the quantification of the seismic source mechanism relies on a discretized representation of the volcanic edifice based on a digital elevation model. Using this model, Green's functions are then calculated by the finite difference method [107]. Once the Green's functions are known, the nine independent mechanisms in Eq. (1) are retrieved through least-squares inversion expressing the standard linear problem $\mathbf{d} = \mathbf{Gm}$ in this equation [32].

Sources of Long-Period Seismicity

Long-Period (LP) seismicity, including individual LP events and tremor, is commonly observed in relation to magmatic and hydrothermal activities in volcanic areas and is recognized as a precursory phenomenon for eruptive activity (e. g., Chouet, [27]). The waveform of the LP event is characterized by simple decaying harmonic oscillations except for a brief interval at the event onset (Fig. 3). This signature is commonly interpreted as oscillations of a fluid-filled resonator in response to a time-localized excitation. By the same token, tremor may be viewed as oscillations of the same resonator in response to a sustained excitation. LP events are particularly useful in the quantification of magmatic and hydrothermal processes because the properties of the resonator system at the source of this event can be inferred from the properties of the decaying oscillations in the tail of the seismogram. The damped oscillations in the LP coda are quantified by two parameters, T, and Q, where T is the period of the dominant mode of oscillation, and Q is the quality factor of the oscillatory system representing the combined effects of intrinsic and radiation losses.

Interpretations of the oscillating characteristics of LP sources have mostly relied on a model of fluid-driven crack [23,24,25]. This model, which has the most natural geometry that satisfies mass-transport conditions at depth beneath a volcano, is supported by results from inversions of LP waveforms recorded at several volcanoes [80,82,97,137]. A simplified two-dimensional model of fluid-driven crack was originally proposed by Aki et al. [2]. Although this model considered both the driving excitation and geometry appropriate for transport, the fluid inside the crack was only treated as a passive cushion that did not support the acoustic propagation of the pressure disturbance caused by the motion of the crack wall. An extension of this model including active fluid participation was proposed by Chouet and Julian [22], who considered a simultaneous solution of the elastodynamics and fluid dynamics for a two-dimensional crack. This model was further extended to three dimensions by Chouet [23] and was extensively studied by Chouet [24,25].

Chouet's three-dimensional model consists of a single isolated rectangular crack embedded in an infinite elastic body and assumes zero mass transfer in and out of the crack (Fig. 4). Crack resonance is excited by a pressure transient applied symmetrically on both walls over a small patch of crack wall. In this model, the crack aperture is assumed to be much smaller than the seismic wavelengths of interest and the motion of the fluid inside the crack is treated as two-dimensional in-plane motion. The fluid dynamics are represented by the conservation of mass and momentum for the two in-plane components of fluid velocity averaged over the crack aperture, and convective terms in the momentum equations are neglected assuming their contributions are negligibly small compared to the time derivatives of pressure and flow velocity. A solution of the coupled equations of fluid dynamics and elastodynamics is then obtained by using a finite-difference approach. As formulated, the model accounts for the radiation loss only; intrinsic losses due to dissipation mechanisms within the fluid are treated separately [77].

Figure 5 shows snapshots of the normal component of particle velocity on the crack wall obtained at four different times following the onset of crack resonance. In this particular example, the crack excitation is triggered by a step

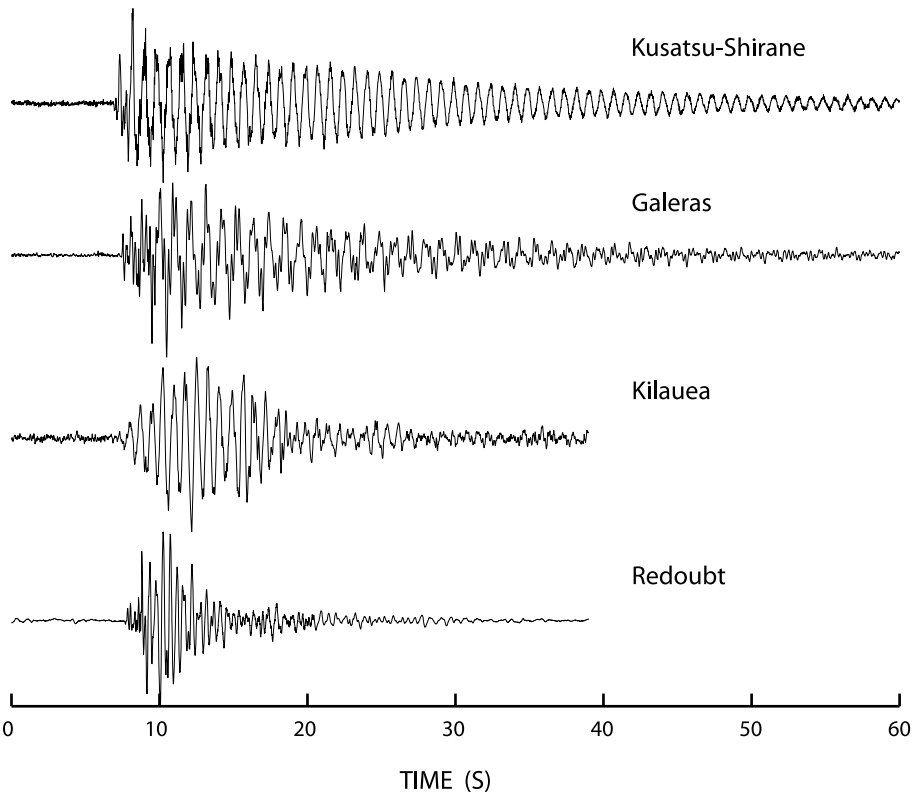

Volcanoes, Non-linear Processes in, Figure 3
Typical signatures of long-period events observed at Kusatsu-Shirane, Galeras, Kilauea, and Redoubt Volcanoes. The signatures are all characterized by a harmonic coda following a signal onset enriched in higher-frequencies

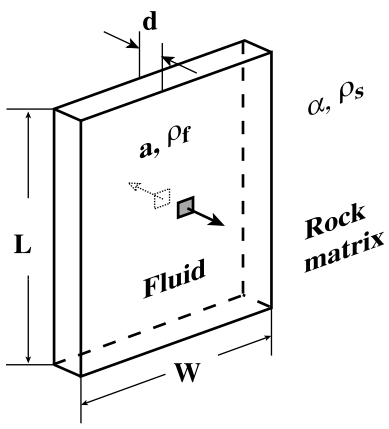

Volcanoes, Non-linear Processes in, Figure 4
Geometry of the fluid-filled crack model of Chouet [23]. The crack has length *L*, width *W*, and aperture *d*, and contains a fluid with sound speed *a* and density ρ_f. The crack is embedded in an elastic solid with compressional wave velocity α and density ρ_s. Excitation of the crack is provided by a pressure transient applied symmetrically on both walls over the small areas such as that indicated by the gray patch and dotted small square

in pressure applied at a point located a quarter of the crack width from the left edge of the crack and a quarter of the crack length from the front edge of the crack. The first snapshot at the upper left shows the propagation of the elastic disturbance on the crack surface shortly after the onset of the pressure transient. The second snapshot at the lower left shows the motion of the crack wall following the first reflections from the front, left and right edges of the crack. The third snapshot at the upper right represents the wall motion immediately preceding the reflection from the back edge of the crack, and the fourth snapshot at the lower right depicts the shape of the lateral and longitudinal modes of crack resonance, which by then are well established.

The excitation of modes in the crack model depends on the position of the pressure transient, the extent of crack surface affected by the transient, the temporal characteristics of the transient, and crack stiffness $C = (b/\mu)(L/d)$, where b is the bulk modulus of the fluid, μ is the rigidity of the rock, L is the crack length, and d is the crack aperture [2,23]. An example of synthetic seismo-

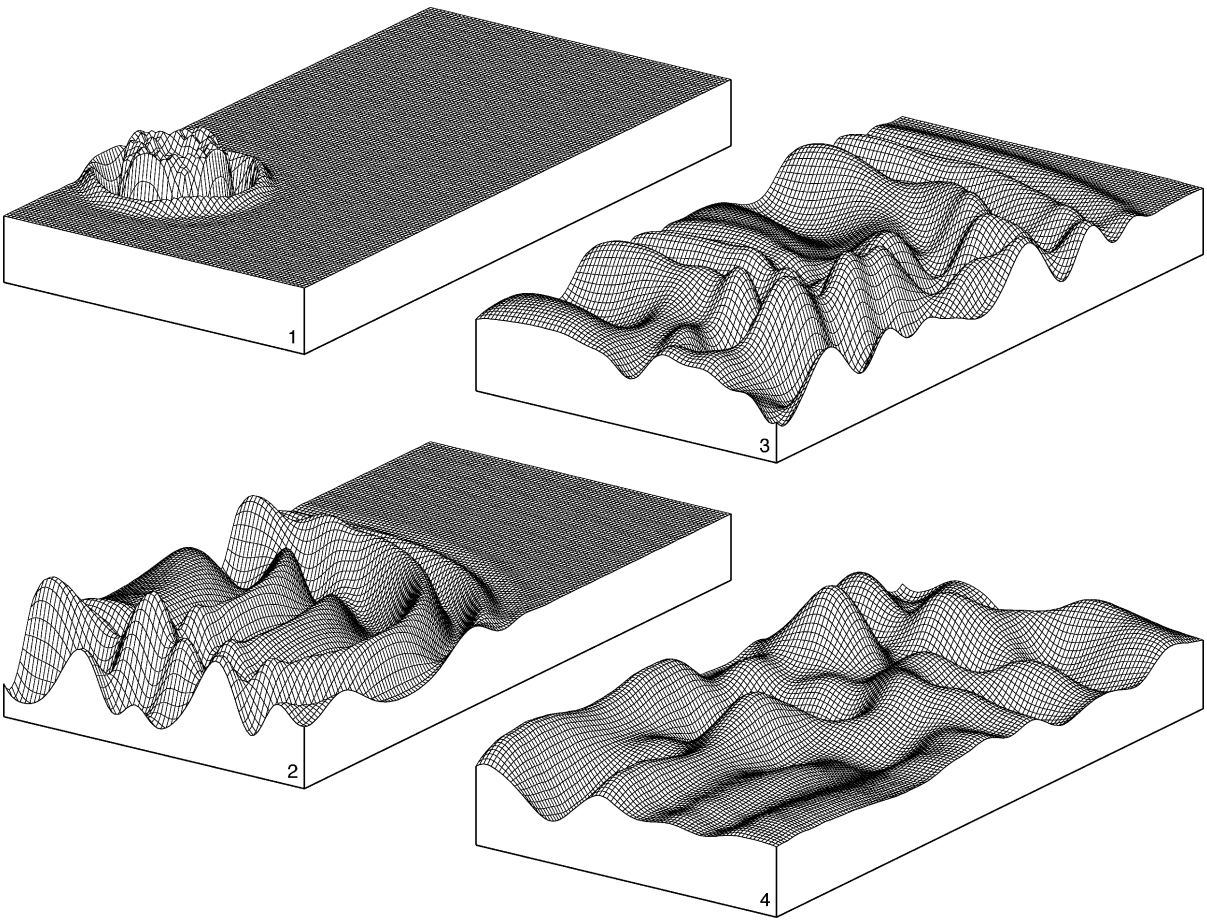

Volcanoes, Non-linear Processes in, Figure 5
Snapshots of the normal component of particle velocity on the surface of a fluid-filled crack excited by a step in pressure applied over a small patch of the crack wall. The snapshots (1–4) show the propagation of the disturbance on the crack wall at four different times in the history of crack oscillation specified by the number of time steps, N, in the finite-difference solution obtained by Chouet [23]. The time step has size $\Delta t = 0.003125L/\alpha$ (see Fig. 4 for explanation of symbols). Depicted are the shape of the wall motion at $N = 72$ (snapshot 1), $N = 312$ (snapshot 2), $N = 560$ (snapshot 3), and $N = 1200$ (snapshot 4). (See text for discussion)

gram representing the vertical ground velocity response in the near field of a fluid-filled crack excited into resonance by a step transient in pressure is shown in Fig. 6. The synthetics display sustained oscillations that bear strong resemblance to observed LP waveforms (compare with Fig. 3), convincingly demonstrating the critical involvement of active fluid participation in the source process of volcanic signals.

Chouet's studies demonstrated the existence of a slow wave propagating along the crack wall, which he named the "crack wave". The asymptotic behavior of this wave was investigated by Ferrazini and Aki [42] in an analytical study of normal modes trapped in a liquid layer sandwiched between two elastic half spaces. The crack wave speed is slower than the sound velocity of the liquid and

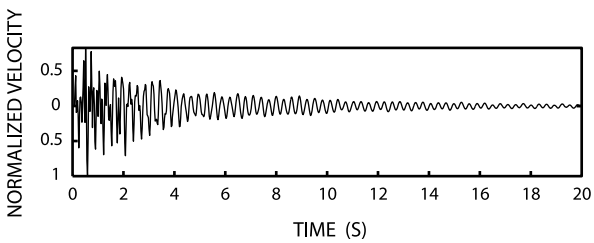

Volcanoes, Non-linear Processes in, Figure 6
Synthetic velocity waveform observed at an epicentral distance of 500 m and azimuth of 45° from the crack trace for a vertical crack (vertical extent $L = 150$ m and horizontal extent $W = 75$ m) buried at a depth of 500 m in a homogeneous half space. The aspect ratio is $L/d = 10^4$, and the parameters of the fluid and solid are $a = 300$ m/s, $\rho_f = 2120$ kg/m^3, and $\alpha = 4500$ m/s, $\rho_s = 2650$ kg/m^3. (Reproduced from [77])

decreases with increasing crack stiffness and increasing wavelength; this result reflects the elasticity of the crack wall, whose effect is to decrease the effective bulk modulus of the fluid. These slow characteristics of the crack wave lead to more realistic estimates of crack dimensions compared to estimates based on the sound speed of a fluid embedded in a resonator with perfectly rigid walls. Source dimensions estimated from LP data based on this model typically range from tens to several hundred meters [79, 82,120]. Detailed analyzes of the dependence of crack resonance on fluid composition carried out by Kumagai and Chouet [77], and systematic investigations of LP signatures based on their results, suggest that dusty gases or bubbly basalt are common fluids in LP events of magmatic origin [27,45,76], and that wet gases, steam and bubbly water are typically representative of the source of LP events of hydrothermal origin [79,120].

The model of Chouet [23] does not address the excitation mechanism of LP events or tremor. Rather, the spatio-temporal properties of the pressure transient triggering the crack resonance are preset as kinematic conditions in the model. The usefulness of this model is thus restricted to a quantification of the crack resonance and properties of the fluids at the source of LP events. For a better perspective of the excitation mechanism of LP seismicity, observations of the source over a wider band of frequencies are necessary, along with an adequate sampling of the evolutionary characteristics of volcanic seismicity. Observations carried out in different volcanic settings point to a wide variety of LP excitation mechanisms originating in magmatic-hydrothermal interactions, as well as magmatic instabilities.

Magmatic-Hydrothermal Interactions

In hydrothermal systems, LP seismicity can be induced by surges in the heat transfer from an underlying magma body. Kusatsu-Shirane Volcano in central Japan provides a prime example of the character of such interactions. Three crater lakes (Yugama, Mizugama, and Karagama) occupy the summit of the volcano along with numerous hot springs, pointing to an active magmatic system and enhanced circulation of hydrothermal fluids. Geochemical studies provide further support for hydrothermal activity resulting from the interaction of hot volcanic gases with groundwater [55,106]. LP events marked by nearly monochromatic oscillatory signatures have been frequently observed at this volcano (see Fig. 3) [43,46], and temporal variations in the decay characteristics of these events have also been documented [96]. Kumagai et al. [79] performed a systematic study of the temporal characteristics of LP events observed at Kustasu-Shirane over a period from August 1992 through January 1993, and Kumagai et al. [80] carried out waveform inversions of a typical event during this time interval. Results from these studies point to the repeated excitation of a fixed sub-horizontal crack at depths near 200 m under the Yugama crater lake, and consistently explain temporal variations in LP signatures by the dynamic response of this crack to a magmatic heat pulse. In this interpretation, the crack contains an initially wet gas, which became gradually drier with time, suggesting a "drying out" of the hydrothermal system in response to the heat pulse. Further analyzes by Nakano et al. [97] based on waveform inversions of the effective excitation functions (the signal components remaining after removal of the resonance characteristics) of individual LP events confirm the crack mechanism imaged by Kumagai et al. [80] and shed further light on the source process. This process involves a collapse and recovery of the sub-horizontal crack accompanied by an upward-directed force, and is consistent with a gradual buildup of pressure in the crack causing repeated discharges of steam from the crack. The vertical force imaged with the volumetric component of the source reflects the release of gravitational energy that occurs as a slug of steam ejected from the crack ascends toward the surface and is replaced by a downward flow of cooler water within a conduit linking the crack to the base of the Yugama crater lake. In this scenario, crack resonance is triggered by the sudden collapse of the crack induced by the venting of steam.

Similar mechanisms of interaction between magmatic and hydrothermal systems have been inferred for Kilauea, Mount St. Helens, and Redoubt Volcanoes. At Kilauea Volcano, Hawaii, Almendros et al. [4] imaged a shallow hydrothermal system from frequency-slowness analyses of LP seismicity recorded on three small-aperture seismic arrays deployed in the summit caldera. Located within the top 500 m below the caldera floor, this reservoir is perched immediately above the shallowest segment of magma conduit identified by [109] from inversions of VLP signals produced during a transient in the magma flow feeding the long-lived and ongoing east rift eruption of this volcano [48]. Episodic deflations of the Kilauea summit during brief pauses in the flow of magma are commonly followed by rapid reinflation of the summit during the resumption of flow, and subsequent slower deflation back to an original state preceding the pause [21]. Larger summit reinflation episodes associated with more energetic surges of magma are accompanied by the emission of VLP signals, production of shallow brittle failure earthquakes, and enhanced LP seismicity in the hydrothermal reservoir [39,109]. Detailed analyses of seismic array data by

Almendros et al. [5] further document the persistence of low-amplitude tremor generated by shallow hydrothermal activity in response to the flux of heat from the underlying magma conduit. Additional support for this scenario also comes from analyses of the source process of a hydrothermal LP event [82], which point to the resonance of a horizontal crack at depth of ~ 150 m immediately above the conduit imaged by Ohminato et al. [109]. The observed frequencies and attenuation characteristics of crack resonance are consistent with a crack filled with either bubbly water or steam [82].

The inferred scenario of magmatic-hydrothermal interactions at Kilauea, which is in harmony with all of the above observations, involves a chain of causes and effects that starts with a volumetric deformation of the underlying shallow segment of magma conduit induced by the surging magma. This conduit response causes brittle fractures in the surrounding rock matrix, and thus enhances the upward transport of magmatic gases through the fractured medium to shallow depths, where the hot gases heat and activate the hydrothermal system. The boiling of groundwater and attendant production of steam then raise the overall pressure of hydrothermal fluids in fractures permeating the medium, thereby leading to repeated impulsive discharges of fluid and resonant excitation of individual fractures.

The current dome-building eruption of Mount St. Helens Volcano, Washington, is marked by the extrusion of a stiff plug of dacitic magma [40,91] accompanied by shallow, repetitive LP events that have been dubbed "drumbeats" by scientists monitoring Mount St. Helens [94]. VLP signals commonly accompany the LP events, but are generally detected only in the immediate vicinity of the crater. Using high-quality data from a temporary broadband seismic network deployed in 2004–2005, Waite et al. [137] carried out systematic waveform inversions for both LP and VLP signals. Results from these inversions point to the perturbation of a common crack system linking the magma conduit and shallow water-saturated region of the volcano. A scenario consistent with the imaged source mechanisms involves the repeated pressurization of a sub-horizontal crack that lies beneath the growing lava dome in the southern part of the crater. Steam produced by the heating of ground water, as well as steam exsolved from the magma, feed into the crack and pressurize it. Upon reaching a critical pressure threshold, steam is vented out of the crack through fractures leading to the surface. The drop in pressure triggers a partial collapse of the crack, which resonates for 5 to 10s of seconds. The collapse of the crack in turn induces sagging of the overlying dome and triggers a passive response in the magma conduit that is observed in the VLP band. The observed force and volumetric components of this VLP response are both consistently explained as the result of a perturbation in an otherwise smooth steady flow of magma upward through a dike and into a sill underlying the old dome. In this model, the dike represents the top of the old conduit that fed the 1980–1986 dome-building eruptions, and the sill represents a bypass below the old dome. According to this interpretation, at the start of renewed activity in September 2004, magma moving up the conduit found an easier pathway to the surface by breaking out of the conduit near the surface and pushing sub-horizontally to the south rather than intruding the older dome complex. The imaged sill represents this segment of conduit, and the hydrothermal crack represents an extension of this fracture into the shallow water-soaked region beneath the crater floor. This picture of Mount St. Helens as a steam engine is quite distinct from an earlier model in which the seismic drumbeats were assumed to represent stick-slip motion between the extruding lava and conduit walls (see below). A sustained supply of heat and fluid from the magmatic system is necessary to keep the crack pressurized and keep the drumbeats beating in the model elaborated by Waite et al. [137]. As for Kilauea, a proper interpretation of this activity has important implications for the assessment of the potential for phreatic eruptions at Mount St. Helens.

Another perspective on a different type of instability presumed to be associated with dome growth activity at Mount St. Helens was proposed by Iverson et al. [57]. Their model, originally developed before results from the detailed seismic analyses by Waite et al. [137] became available, hypothesizes that repetitive stick-slip motion along the perimeter of the extruding solid plug may cause observable earthquakes. Although not supported by the character of LP seismicity observed at Mount St. Helens, this model nevertheless offers an intriguing view into a possible oscillatory behavior of plug flow. Oscillations in extrusion velocity originate in the interaction of plug inertia, a variable upward force due to magma pressure, and downward force due to plug mass. Assuming that a steady ascent of compressible magma drives the upward extrusion of a solidified plug exhibiting nonlinear rate-weakening friction along its margins, Iverson et al. [57] infer that stick-slip oscillations might be an inevitable component of such an extrusion process; however, the predicted amplitude of individual slip events, estimated at a few millimeters, suggests this mechanism may be difficult to detect. The seismic source associated with a stick-slip extrusion event may manifest itself as a combination of a near-vertical force (assuming vertical spine growth) representing the

reaction force on the Earth due to the upward acceleration of mass, and double couple resulting from shear near the plug margins where viscosity is very high and magma is in the glass transition. This particular mechanism does not involve LP source resonance and is apparently not part of the dominant expression of seismicity at Mount St. Helens, but could possibly be operating at a level below the threshold of seismic detection. Such mechanism may be observable in dome-building eruptions where the extrusion rate is sufficiently large and unsteady to produce oscillations of this type.

The reawakening of Redoubt Volcano, Alaska on 13 December, 1989, after 23 years of quiescence, was heralded by a rapidly accelerating swarm of repetitive LP events that merged into sustained tremor a few hours before the eruption onset [26,126]. These LP events were interpreted by Chouet et al. [26] as the resonant excitation of a crack linking a shallow, low-pressure hydrothermal system to a deeper supercharged magma-dominated reservoir. The initiation of the swarm was attributed by these authors to the failure of a barrier separating the two reservoirs. The actual mechanism triggering the LP excitation in this model is the unsteady choking of a supersonic flow of magmatic steam driven by the pressure gradient existing between the two reservoirs. Unsteady choking itself is the result of fluctuations in crack outlet pressure associated with the reaction of the hydrothermal system to the injection of mass and heat from below. Similarly, the emergence of sustained tremor late in the swarm is interpreted as a change in choked-flow regime related to a gradual weakening of the pressure gradient driving the flow [92]. Numerical simulations of the transonic flow through a crack featuring a nozzle-like constriction were performed by Morrissey and Chouet [95] using the Navier–Stokes equations representing a mixture of gas and suspended solid particles. In these simulations, shock waves develop immediately downstream from the nozzle-like constriction, and the flow acts as an energy transducer that converts smooth low-amplitude fluctuations of outlet flow pressure into strongly-amplified and repetitive step-like pressure transients applied to the crack wall immediately downstream from the nozzle. The magnitude of the pressure transient at the walls is fixed by the pressure of the supplying reservoir and the geometry of the nozzle aperture, presumably remaining constant as long as these conditions prevail. The temporal pattern of outlet pressure fluctuations controls the areal extent of the crack wall impacted by shock waves and, given a roughly constant shock magnitude, fixes the amplitude of the force applied to the wall by the fluid. This variable force applied at a fixed location along the crack wall is viewed as the force responsible for the scaling of amplitudes and the similarity of waveforms observed in the LP swarm at Redoubt [26].

Magmatic LP Events

In low-viscosity basaltic systems, LP seismicity may be part of a broadband source process associated with the unsteady transport of magma and gases through conduits. This is well illustrated in a broadband seismic record obtained at Kilauea (Fig. 7); most interesting is the displacement record shown at bottom of figure. This record, obtained after filtering the raw broadband signal (top trace in Fig. 7) with a 0.125 Hz low-pass filter, features a repetitive sawtooth signal with rise time of 2–3 min and drop time of 5–10 s. The observed repetitiveness of the VLP signal is a clear indication of the repeated action of a non-destructive source. Noteworthy also are the characteristic LP signatures obtained by filtering the broadband signal with a 0.33-Hz high-pass filter (third trace from the top). These LP events display a fixed dominant period of 2.5 s, whose onsets coincide with the start of the downdrop in the VLP sawtooth displacement signals. Such coincidence is strongly suggestive of a causative relationship between the VLP and LP signals.

The data in Fig. 7 are part of a seismic sequence observed during a surge in the magma flow feeding the ongoing east rift eruption of Kilauea, and the source mechanism of the VLP signals obtained by Ohmiato et al. [109] from waveform inversions of these data points to a repeated cycle of deformation of a sub-horizontal crack located roughly 1 km below the floor of the summit caldera. Each sawtooth in the recorded displacement signal represents a slow dilation of the crack, followed by a rapid collapse of the crack. The volume change associated with individual sawtooths ranges from 1000 to 4000 m^3 [109].

A conceptual model consistent with the seismic source mechanism imaged by Ohmiato et al. [109] is shown in Fig. 8. This model involves the injection of a large slug of gas into a sub-horizontal crack with a narrow constricted outlet; both liquid magma and gas flow through this outlet constriction. Initially, liquid magma occupies the entire outlet cross section and gas accumulates in the crack, forming a layer of gas on top of the liquid and deforming the crack walls as the back pressure of gas increases in the crack. Eventually, when the back pressure buildup in the crack reaches a critical threshold, the gas slug deforms and rapidly squeezes past the constriction along with the liquid, triggering a rapid deflation (collapse) of the crack in the process. This sequence is then repeated with the next gas slug. In this scenario, the restraining force of the liquid on the gas acts like a self-activated viscosity-controlled

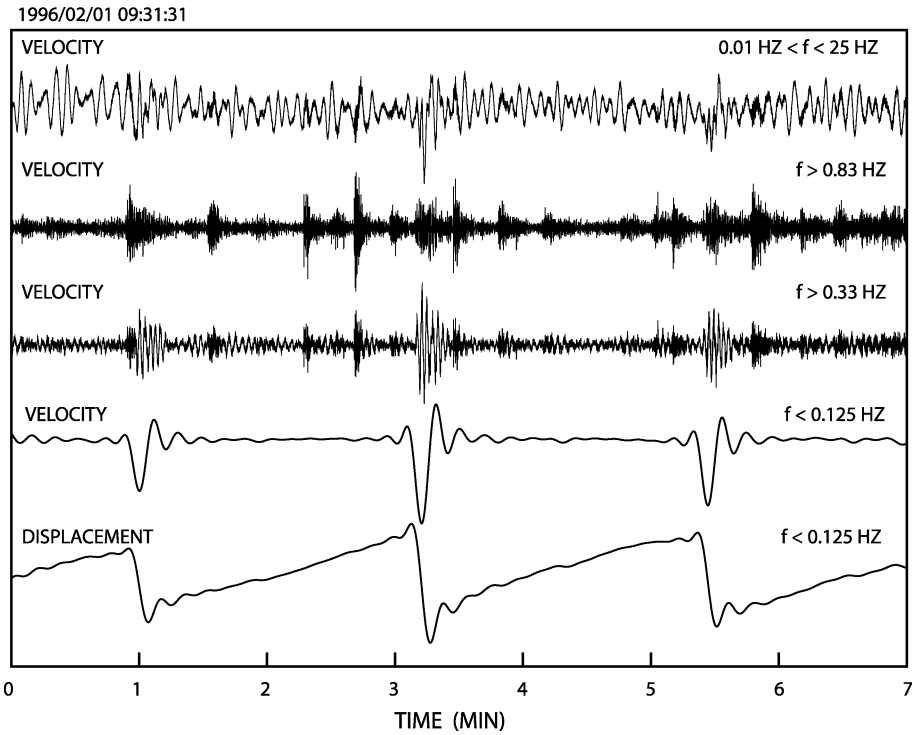

Volcanoes, Non-linear Processes in, Figure 7

Broadband record and associated filtered signals obtained at Kilauea Volcano during a volcanic event on February 1, 1996. The broadband signal represents ground velocity and is filtered in various frequency bands to produce five records for the same 7-min time interval. The top trace shows the broadband signal ($0.01 < f < 25$ Hz), which is dominated by the oceanic wave-action microseism with dominant periods in the range 3 to 7 s. The second trace shows the signal obtained after application of a high-pass filter ($f > 0.83$ Hz). The result is equivalent to a typical short-period record and shows a series of events superimposed on a background of tremor. The third trace also has a high-pass filter applied but with a lower corner frequency ($f > 0.33$ Hz); LP signals with a dominant period of about 2.5 s are enhanced in this record. The fourth trace shows the signal when a low-pass filter is applied ($f < 0.125$ Hz); a repetitive VLP signal consisting of pulses with period of about 20 s is observed. The fifth trace is the corresponding displacement record obtained with the same low-pass filter ($f < 0.125$ Hz), showing a repetitive sawtooth pattern with rise time of 2–3 min and drop time of 5–10 s. Notice that the onset of the LP signal seen above coincides with the onset of the downdrop in the VLP displacement. (Reproduced from [38])

valve. In the stratified flow through the nozzle, the liquid moves at a steady slow pace on the order of m/s but the gas flow itself is choked in the narrow opening between the liquid/gas interface and the upper wall [138]. The formation of a shock associated with compound choking of the flow is viewed here as a trigger of acoustic oscillations of the liquid/gas mixture, which is at the origin of the LP signature with characteristic period of 2.5 s observed in conjunction with the rapidly downgoing part of the sawtooth displacement signal shown in Fig. 7.

At Stromboli Volcano, Italy, broadband seismic signals including strong VLP and LP components are seen to accompany explosions. Waveform inversions of the VLP signals performed by Chouet et al. [32] show that the processes associated with eruptions involve volumetric changes marking a cycle of pressurization-depressurization-repressurization of the conduit, coupled with a single force component, both of which may be viewed as the result of a piston-like action of the magma associated with the disruption of a gas slug transiting through a sudden enlargement in conduit aperture in the steeply-inclined upper segment of conduit [32,60] (see Sect. "Slug Distribution at Stromboli" for details). The LP signal in that case is attributed to the oscillatory response of the shallowest segment of the fluid-filled conduit associated with the rapid expansion and ejection of the slug [29].

LP seismicity is also intimately linked with Vulcanian activity. The complex magma rheology, marked by a strong nonlinear dependence of viscosity on water content and temperature [49,147], finite yield strength, strain-rate dependent viscosity, and transition to brittle solid behavior at high strain rates [133] contributes to

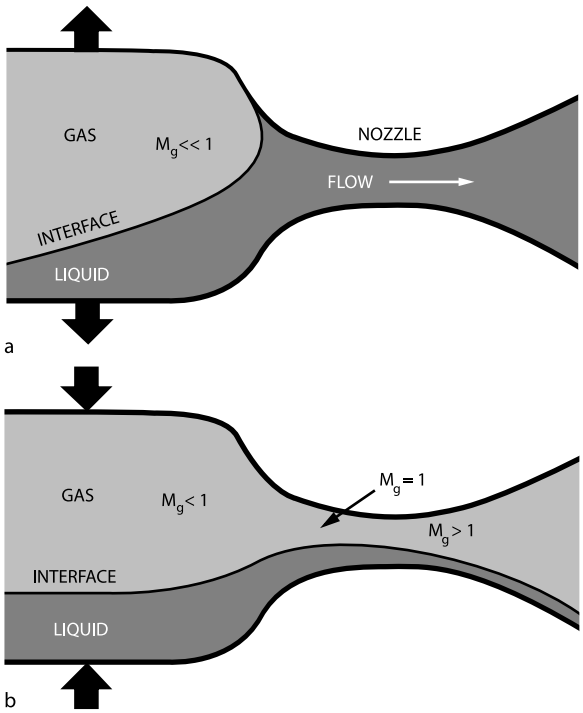

Volcanoes, Non-linear Processes in, Figure 8

Conceptual model of separated gas-liquid flow through a converging-diverging nozzle under choked conditions (see pp. 71–74 in [138]). a Inflation phase showing gas accumulation upstream of the nozzle, leading to pressure buildup and deformation of the crack. The gas slug is essentially stationary and the Mach number of the gas is $M_g \ll 1$. This stage coincides with the upgoing ramp in the sawtooth displacement signal at the bottom of Fig. 7. b Separated gas-liquid flow through the nozzle under compound choked conditions. The gas flow is choked ($M_g = 1$) in the nozzle defined by the gas-liquid interface and upper solid wall, and is supersonic ($M_g > 1$) immediately downstream of the nozzle. There exists a maximum possible gas flow rate that is fixed by the rate of liquid discharge for given upstream conditions and fixed throat geometry. In the limit, liquid fills the duct entirely and there is no gas flow. (Reproduced from [109])

widely varying oscillatory behaviors seen in Vulcanian systems. At Galeras Volcano, Colombia, LP events displaying extended quasi-monochromatic coda were found to be linked to the release of ash-laden gases from a shallow magma body through a preexisting system of cracks bisecting the dome capping the vent [45]. Similar to Galeras, eruptive activity at Popocatepetl Volcano, Mexico, is dominated by emissions of steam, gas and ash, and by repeated formation and destruction of lava domes, all of which generate a wide variety of signals [10,11]. LP events accompanying degassing bursts are commonly associated with VLP signals and display waveform features that remain stationary from event to event. The source of these events coincides with, or is very close to, the source of VLP signals, 1.5 km below the crater floor [11,33]. This LP activity represents an integral part of a degassing process similar to that inferred for Galeras, where the sudden expulsion of a pressurized pocket of gases induces elastic deformation of the conduit coupled with acoustic resonance of the conduit (see further discussion in Sect. "Coupled Diffusive-Elastic Pressurization at Popocatepetl Volcano").

LP events at Soufrière Hills, Montserrat, have been linked to the action of a repetitive, non-destructive, stationary source 1.5 km below the crater floor [100]. The temporal relationship between LP events and tilt patterns at Soufrière Hills clearly shows that LP seismicity starts in synchronicity with the passage of tilt through a turning point marked by a maximum in time derivative of the tilt signal. Seismicity subsequently ceases as tilt goes through a second turning point [100]. This correlation between LP events and inflation-deflation cycles is evidence of a strong relationship linking LP seismicity to magma movement and pressure conditions at depth. Based on these observations, [100] proposed a mechanism of LP excitation originating in the brittle failure of a highly viscous magma under high strain rate. In their model, loss of gas and heat

from the magma at the conduit wall leads to the generation of a sharp viscosity gradient at the wall. This in turn leads to a buildup of shear stress near the wall, where the flowing magma undergoing glass transition can fail in a brittle manner. The resulting brecciated zone along the conduit boundary provides a natural pathway for escaping gases. The LP events may then be viewed as an expression of acoustic resonance occurring in well-defined channels in this zone, or possibly as an expression of acoustic energy trapped in an adjacent region of conduit filled with bubbly magma [36,56].

The LP sources described so far all represent shallow sources occurring in the top 2 km below the surface. Although thoroughly documented, such sources do not represent the total expression of LP seismicity. An activity of deep (> 5 km depth) LP events has also been noted in several areas of the world. In particular, episodes of deep (30–60 km) and intermediate-depth (5–15 km) LP seismicity are often observed under Kilauea [74,75]. Analyzes of the intermittency of LP activity were carried out by Shaw and Chouet [122] and Shaw and Chouet [123], who viewed such tremor episodes as the relaxation oscillations of a percolation network of coupled magma-filled fractures. In their interpretation, each individual fracture in the network acts as a source of LP events, and sets of fractures give rise to multiple LP events and episodes of tremor. The aggregate nature of fracture propagation and vibration then lead to the existence of universal scaling laws that give rise to the long-recognized frequency invariance of LP events [122,123].

Other episodic bursts have been noted at depths of 10–20 km under Mammoth Mountain, California, where LP and VLP events have been seen to occur synchronously with spasmodic bursts of small brittle failure earthquakes [54]. A process involving the transport of a slug of CO_2-rich hydrous magmatic fluid derived from a plexus of basaltic dikes and sills has been invoked to explain such occurrences [54]. LP events at depths near 30 km under Mount Pinatubo, Philippines, have been attributed to a basaltic intrusion thought to have triggered processes leading to the cataclysmic eruption of June 15 [139]. Small LP events attributed to magmatic activity have also been detected at depths of 25–40 km beneath active volcanoes in northeastern Japan [47], and at depths of 10–45 km below volcanoes of the Aleutian Arc [114]. In all these instances, the occurrence of deep LP events appears to be more directly related to deep magma supply dynamics than near-surface or surface activity. Because of our lack of knowledge concerning the character and dynamics of deep-seated fluid transport under volcanoes, the actual origins of these events remain enigmatic.

Summary of Inferred Excitation Mechanisms for LP Seismicity

Shallow hydrothermal LP events at Kusatsu-Shirane, Kilauea, and Mount St. Helens all share a common excitation mechanism involving the repeated pressurization of a steam-filled fracture, causing the venting of steam, collapse of the fracture and recharge, in response to heat transfer from an underlying magma body. At Mount St. Helens, where the LP source is located very close to the magma conduit, LP activity is also seen to trigger a passive response of the conduit itself. LP seismicity observed at Redoubt represents a more energetic form of magmatic-hydrothermal interaction where an unsteady choked flow of magmatic gases provides a natural source of pressure perturbation at the origin of LP events and tremor. Magmatic LP events produced during bursts of ash-laden gases associated with Vulcanian activity at Galeras and Popocatepetl involve a pumping mechanism in shallow fractures similar to that inferred for hydrothermal LP events at Kusatsu-Shirane, Kilauea, and Mount St. Helens. LP events accompanying endogeneous dome growth at Soufrière Hills appear to be related to a more complex process involving the production of shear fractures in highly viscous magma at the glass transition, injection of dusty gases into these fractures, and resonance of the fracture network and/or possibly resonance of a bubble-rich magma excited by the energy release from the brittle failure. Magmatic LP events in the basaltic systems at Kilauea and Stromboli are the result of pressure disturbances generated during the transit of large slugs of gas through conduit discontinuities.

A nonlinear excitation mechanism of LP seismicity by fluid flow has also been proposed by Julian [62]. Using a simple lumped-parameter model, Julian investigated the elastic coupling of the fluid and solid as a means to produce self-excited oscillations in a viscous incompressible liquid flowing through a channel with compliant walls. In his model, an increase in flow velocity leads to decrease of fluid pressure via the Bernoulli effect. As a result, the channel walls move inward and constrict the flow, causing an increase in fluid pressure and forcing the channel open again. The cyclic repetition of this process is the source of sustained oscillations in Julian's model. Julian demonstrated that with increasing driving pressure the model can exhibit various oscillatory behaviors resembling tremor.

As illustrated in the above examples, magmatic-hydrothermal interactions and magmatic oscillations can take on a variety of forms. Studies of the mechanisms of tremor excitation are still very much in their infancy, owing to the paucity of accurate and detailed seismic ob-

servations essential in documenting and analyzing such processes.

Source Processes of Very-Long-Period Signals

Unlike the LP signals, which are mainly interpreted as manifestations of acoustic resonance, VLP signals are viewed as the results of inertial forces associated with perturbations in the flow of magma and gases through conduits. An increasingly widespread use of broadband seismometers on volcanoes has led to an increasing number of observations, and VLP signals have by now been documented in volcanic areas all over the world, including Sakurajima [67,68], Unzen [134], Aso [65,69,86,143], Satsuma-Iwojima [108,110], Iwate [101], Miyakejima [44,78], Usu [144], Bandai [103], Hachijo [81,83], and Asama [111] in Japan; Stromboli [14,30,32,35,99] in Italy; Merapi [50,51,52] in Indonesia; Kilauea [6,7,38,39,109] in Hawaii; Long Valley [53,54] and Mount St. Helens [137] in the United States; Erebus [13,119] in Antarctica; Popocatepetl [9,10,33] in Mexico; and Cotopaxi and Tungurahua [84] in Ecuador.

The wavelengths of VLP signals are in the range of tens to hundreds of kilometers, which greatly facilitates their analysis. Inversions of VLP waveforms have imaged crack geometries at the source in the forms of dikes or sills [32,81,109,143], as well as more complicated geometrical configurations involving a composite of a dike intersecting a sill [33], or composites of interesecting dikes [7,35], or two chambers connected to each other by a narrow channel [101]. Based on the characteristic period of VLP signals recorded at Hachijo Island, Japan, Kumagai [83] estimated a source with dimensions of a few km. The decaying harmonic oscillations with periods near 10 s and duration up to 300 s of the VLP signals seen at Hachijo are similar to the features of LP events with periods near 1 s (see Fig. 3) and appear consistent with the resonance of a dike containing a bubbly basalt. In that sense, these VLP signals probably represent a unique end member of the family of sources associated with resonant conduit excitation. Interestingly, the dominant periods and associated decay characteristics of the VLP signals at Hachijo were both found to vary with time, and these temporal variations were attributed by Kumagai [83] to a change in source dimensions, assuming fixed acoustic properties of the fluid at the source. A spread of about 4 km was also estimated by Chouet et al. [101] for two chambers at Iwate Volcano. The modeling of VLP waveforms at Stromboli by Nishimura et al. [35] points to two point sources, each involving two intersecting dikes, whose relative positions and orientations represent individual conduit segments with dimensions of a few hundred meters. Systematic modeling of sustained VLP tremor accompanying episodes of deflation-inflation-deflation at Kilauea has suggested a complex magma pathway composed of a plexus of intersecting dikes extending over the depth range 0.6–1.6 km below the summit caldera [7]. A crack-like conduit extending over the depth range 0.3–2.5 km has also been inferred under Aso Volcano, Japan [143].

Contrasting the above-discussed findings are results obtained by Ohminato [111] from waveform inversions of VLP signals produced by Vulcanian explosions at Asama Volcano, Japan. At Asama, the contribution from a vertical single force with magnitude $10^{10} - 10^{11}$ N was found to dominate the observed waveforms, and no obvious volumetric component was identified at the source. The depth of the VLP source is ~ 200 m beneath the summit crater, and the source-time history features two downward force components separated by an upward force component. The initial downward component was attributed by Ohminato [111] to the sudden removal of the lid capping the pressurized conduit (this force is analogous to the force F_{T_1} in Fig. 2), and the subsequent upward force was interpreted by these authors as a drag force induced by viscous magma moving up the conduit. Ohminato [111] also attributed the final downward force component to an explosive fragmentation of the magma, whose effect is to effectively cancel viscous drag so that the downward force due to jet recoil again dominates the VLP signal.

Available studies demonstrate that detailed investigations of the VLP oscillation characteristics of the source are critically important to our understanding of volcanic fluid dynamics. Below, I review two in-depth investigations of VLP signals representative of Vulcanian degassing bursts at Popocatepetl and gas-slug transport dynamics at Stromboli.

Coupled Diffusive-Elastic Pressurization at Popocatepetl Volcano

Vulcanian degassing bursts at Popocatepetl Volcano, Mexico, have been observed to be closely linked to the degassing of a sill-shaped volume of magma at a depth of 1.5 km below its summit crater [33]. The variations in volume of this sill were interpreted by Chouet et al. [33] as reflecting cyclic pressure oscillations originating in the magma filling the sill. Based on these observations, these authors proposed a model in which static magma in the sill becomes supersaturated because of groundmass crystallization. In this conceptual view, volatile exsolution and diffusion of gas from the melt into bubbles increase the internal pressure of the bubbles, because bubble expansion

is impeded by the viscous resistance of the surrounding liquid and by the confining effects due to the finite yield strength of the overlying column of magma and surrounding solid rock. Elastic inflation of the sill occurs as a result of bubble pressurization, and this inflation proceeds until the critical yield strength of the magma column is exceeded and magma starts flowing out of the sill. Magma fragmentation induced by viscous shear near the conduit wall then causes coalescence and collapse of bubbles intersected by fractures, allowing gas escape through a transient network of fractures. This, in turn, induces a pressure decrease in the sill, which results in the collapse and welding of the fracture network that shuts down the gas-escape pathway. Repeated cycles of shear-induced fracture and welding of magma provide a ratchet mechanism by which the separated gas phase in the magma can be recharged and evacuated.

The expansion dynamics of bubbles in supersaturated magma have been addressed in many previous studies under conditions of constant melt pressure [87,88,90,115,117,121,125], or constant decompression rate [116,131,132]. A canonical model of bubble growth in magma that includes the effect of finite spacing of bubbles in the melt was first developed by Proussevitch et al. [115]. This model considers a suspension of gas bubbles in an incompressible volatile-bearing liquid. The suspension is modeled as a three-dimensional lattice of closely-packed spherical cells, where each identical elementary cell is composed of a gas bubble surrounded by a shell of liquid. A step drop in pressure in the melt induces volatile exsolution and bubble expansion is driven by the diffusion of gas into the bubble. In this model, only the volatiles contained in the shell of liquid surrounding the bubble contribute to the bubble expansion so that bubble growth is limited by how closely packed the bubbles are in the liquid. Using the model of Proussevitch et al. [115], Chouet et al. [34] and Shimomura et al. [124] obtained a dynamic solution that includes the effect of melt compressibility. In their model, pressure recovery is driven by bubble growth in a supersaturated magma that is stressed by the surrounding crust as the magma volume increases with bubble expansion. Both models consider the pressure recovery following an instantaneous pressure drop in the melt. The magma is embedded in an infinite, homogeneous, elastic solid and consists of a melt containing numerous small spherical gas bubbles of identical size. No new bubbles are created and no bubbles are lost during pressure recovery, and the gas in the bubbles is assumed to be a perfect gas. Gravity and other body forces are not considered.

The conduit geometry considered by Chouet et al. [34] applies specifically to the observation made at Popocatepetl [33] and consists of a penny-shaped crack containing a melt with an isotropic distribution of bubbles (Fig. 9a). The initial oversaturation of the melt resulting from a pressure drop ΔP_0 is distributed uniformly in the melt, and each bubble grows by diffusion of volatiles from a surrounding shell of melt with finite radius (Fig. 9b). As diffusion is a slow process, with time scale much longer than that of heat transfer into the bubble [56], bubble growth is assumed to be isothermal. The mathematical model describing an elementary cell therefore consists of three equations: (1) a diffusion equation governing the transfer of volatiles in the melt shell; (2) an equation expressing the radial motion of the bubble; and (3) a relation representing the ideal gas approximation. These equations are then subjected to three boundary conditions expressing the phase equilibrium at the bubble wall, mass flux at the bubble wall, and zero mass flux through the outer shell wall. As the bubble grows by mass transfer from the melt, melt is compressed and the surrounding rock ma-

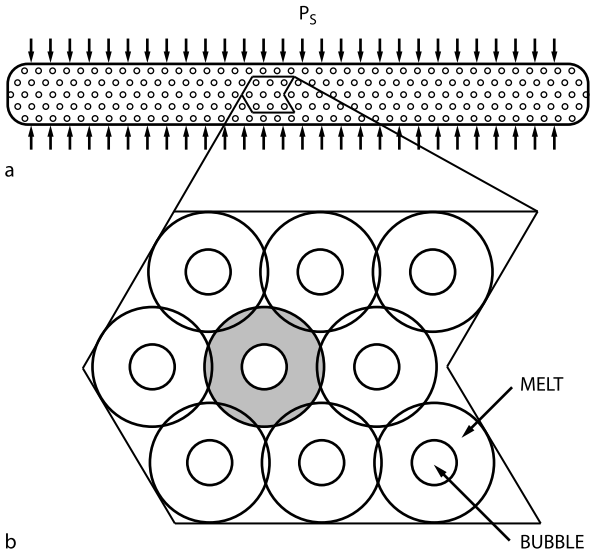

Volcanoes, Non-linear Processes in, Figure 9
Model used by Chouet et al. [34] to interpret the source mechanism of Vulcanian degassing bursts at Popocatepetl Volcano, Mexico. a Schematic illustration (cross section) of penny-shaped crack embedded in an elastic solid. The crack is under confining pressure P_s, and contains a melt with an isotropic distribution of gas bubbles of identical sizes. b Detailed view of the distribution of elementary cells composing the bubbly liquid (after [115]). Each cell (as in gray-shaded example) consists of a gas bubble surrounded by a finite volume of melt. The elementary cells are organized in a three-dimensional lattice with slight overlap of the cells, where the volumes of intersecting melt and gaps are equal, so that the gas-volume fraction is $(R/S)^3$, where R is the bubble radius and S is the outer shell radius

trix is deformed as a result. The pressure in the melt is balanced by the stress applied by the surrounding elastic medium, and in both Chouet et al. [34] and Shimomura et al. [124], this is formalized by assuming a quasi-static deformation of the magma conduit. Melt shrinkage from dehydration [105] is assumed to be negligible compared to melt compression due to bubble growth and is not considered.

In their treatment of the initial condition in the bubble, Chouet et al. [34] follow Proussevitch et al. [115] and assume an instantaneous drop in gas pressure inside the bubble. This induces an instantaneous drop in volatile concentration at the bubble wall, thereby producing a step-like gradient of concentration that kick-starts the diffusion process. This condition is different from that assumed by Shimomura et al. [124], who follow the model of Navon and Lyakhovsky [98], in which the bubble initially retains the original gas pressure and immediately expands in response to the difference between internal and external pressures. The relaxation expansion of the bubble then induces a decrease in gas pressure inside the bubble, which leads to the establishment of a gradient in volatile concentration at the bubble wall that starts the diffusion. Although this leads to distinct features in the instantaneous bubble response, this has essentially no effect on the volumetric response of the bubble at longer time scales.

Figure 10 shows the pressure recovery calculated by Chouet et al. [34] for input step drops $\Delta P_0 = 0.01$, 0.1, and 1 MPa applied to a bubbly rhyolitic melt encased in a penny-shaped crack with radius 100 m and thickness 5 m under ambient pressure $P_s = 40$ MPa appropriate for the depth of the source imaged for Popocatepetl Volcano [33]. The bubbles have a fixed initial radius 10^{-6} m, and the bubble number density is 10^{12} m^{-3}. Melt density, viscosity, and diffusivity are 2300 kg m^{-3}, 10^6 Pa s, and 10^{-11} m^2 s^{-1}, respectively; surface tension is 0.2 Nm^{-1}, and the bulk modulus of melt and elastic rigidity of the rock are both 10^{10} Pa. The amplitude ratio of pressure recovery to initial pressure drop displays a strong non-linear sensitivity to the magnitude of the input pressure drop, resulting in magnifications ranging from about 1 to nearly 30 over the range of input transients considered (Fig. 10). This overpressurization upon recovery was previously noted in a static solution obtained earlier by Nishimura [104] and becomes largest for tiny bubbles and a stiff elastic medium [34,124]. Figure 10 also indicates that the speed of recovery is a function of the input transient, with the crack response becoming markedly slower for small pressure transients compared to larger ones.

The variation of time and amplitude scales seen in Fig. 10 suggests that a wide range of responses may

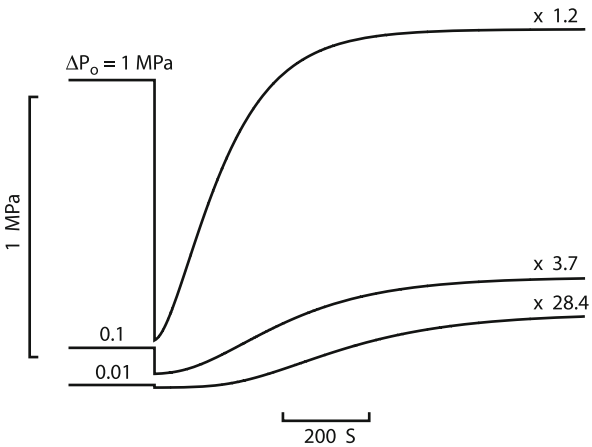

Volcanoes, Non-linear Processes in, Figure 10
Dependence of pressure recovery on the size of the input transient, ΔP_0, in the model of Chouet et al. [34]. The magnification factor for each response is indicated at the upper right of each pressure trace

be possible for cracks containing melts with a wider population of bubble sizes than the single-size bubble populations considered by Chouet et al. [34]. The net drop in volatile concentration associated with diffusion-driven bubble growth is quite small. Indeed, the results obtained by Chouet et al. [34] point to concentration drops $\sim 0.003 - 0.035$ wt% for initial bubble radii of $10^{-5} - 10^{-7}$ m in penny-shaped cracks with aperture to radius ratios $0.01 - 0.25$. The net concentration drop is small because pressure recovery in the melt acts to raise the saturation level of volatiles in the melt. This has important implications for the overall history of degassing in a Vulcanian system, as it suggests that this process may be repeated many times without significantly depleting the gases in the melt body.

The source process imaged under Popocatepetl involves a sequence of inflation, deflation, and reinflation, reflecting a cycle of pressurization, depressurization, and repressurization within a time interval of 3–5 min [33]. The volumetric component of the source processes in the sill associated with an eruption on 23 May 2000 is shown in Fig. 11. Other eruptions were found to produce similar waveform characteristics, hence the event in Fig. 11 may be viewed as an appropriate representation of overall source dynamics associated with these Vulcanian eruptions. For comparison, Fig. 11 also shows the volume change of a magma-filled sill calculated with the model of Chouet et al. [34]. The sill response has been band-pass filtered in the same band as the observed signal, and is illustrated for three distinct pressure drops. The model pa-

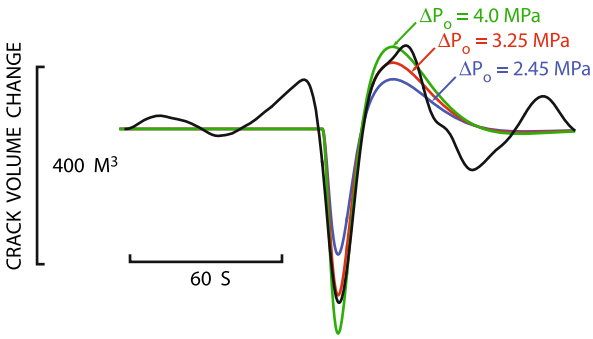

Volcanoes, Non-linear Processes in, Figure 11
Source-time function of volume change obtained by Chouet et al. [33] for a sill under Popocatepetl Volcano during a degassing burst on 23 May 2000 (black line), and corresponding volume changes in a sill filled with a bubbly melt in response to a step drop ΔP_0 in pressure (colored lines). The data from Popocatepetl and solutions from the crack model have both been band-pass filtered in the 15–70 s band

rameters in Fig. 11 are identical to those used in Fig. 10. A realistic fit of the peak-to-trough amplitude of the volume change imaged at Popocatepetl is obtained for an input pressure drop $\Delta P_0 = 3.25$ MPa.

Unfortunately, the longer-period components and other specific features of the sill response in the model are lost in the band-limited version of the signal compatible with observations made at Popocatepetl, so that the responses obtained in this band for different model parameters are all very similar [34]. Although it is not possible to infer specific values of melt viscosity, volatile diffusivity, initial bubble radius, bubble number density, or sill aperture to radius ratio based on available seismic data, these results strongly support the idea that diffusion pumping of bubbles in the magma under Popocatepetl provides a viable mechanism for pressure recovery following a pressure drop induced by a degassing event.

Slug Disruption at Stromboli

Eruptive behavior at Stromboli is characterized by mild, intermittent explosive activity, during which well-collimated jets of gases laden with molten lava fragments burst in short eruptions typically lasting 5–15 s. Modeling of VLP seismic data recorded during explosive activity in 1997 has imaged two distinct dike structures representative of explosive eruptions from two different vents located near the northern and southern perimeters of the summit crater [32,35].

Figures 12 and 13 show two representations of the conduit geometry underlying the northern vent area. Figure 12 shows the seismic source mechanisms obtained from inversion of VLP waveforms; the mechanisms include two point sources, each of which marks a flow disruption site along the upper conduit. The upper source (Fig. 12a) is located 300 m below the crater floor and represents a bifurcation in the conduit; this is the main flow disruption site for gas slugs ascending toward the northern vent. The lower source (Fig. 12b), located roughly 500 m below the upper source, involves a sharp corner in the conduit and represents a secondary flow disruption site. Both upper and lower sources feature a dominant crack sustaining the largest volume change (colored red), and a subdominant crack undergoing a smaller volume change (shaded gray). Both cracks in Fig. 12a display a similar sequence of inflation-deflation-inflation. The dominant crack in the lower source displays a volumetric response similar to that seen in the dominant crack at the upper source, but delayed by about 1 s with respect to the upper source. This implies a propagation speed of roughly 500 m/s between the two sources, consistent with the slow speed expected for the crack wave.

Figure 13 shows a picture of the upper conduit geometry consistent with the seismic source mechanisms imaged in Fig. 12. The closely matching dips of the two dominant cracks in Fig. 12 point to a conduit that extends essentially straight from 80 m below sea level to the crater floor, 760 m above sea level. At a depth of 80 m below sea level the conduit features a sharp corner leading into a dike segment dipping 40° to the southeast. The upper dominant dike segment, and deep segment below the abrupt corner both strike northeast-southwest along a direction parallel to the elongation of the volcanic edifice and a prominent zone of structural weakness, as expressed by lineaments, dikes, and brittle structures. The surface trace of the main dike segment trends through the northern vent area, while that of the upper subsidiary segment extends northwest-southeast in rough alignment with several vents active in the northwest quadrant of Stromboli in 2002–2003 [1]; the subsidiary dike trace intersects the main dike trace $\sim$ 170 m north of the northern vent area.

A striking aspect of the mechanisms imaged for the two sources in Fig. 12 is the presence of dominantly vertical single-force components with common-looking time histories, except for a polarity reversal in one source compared to the other. The upper source (Fig. 12a) displays an initially downward force followed by an upward force, while an upward force followed by a downward force is manifest in the lower source (Fig. 12b). These force components compensate each other so that the total momentum in the overall source volume is conserved. In contrast to the delay of $\sim$ 1 s in the volumetric components noted above, no significant delay is noted in the onsets of the ver-

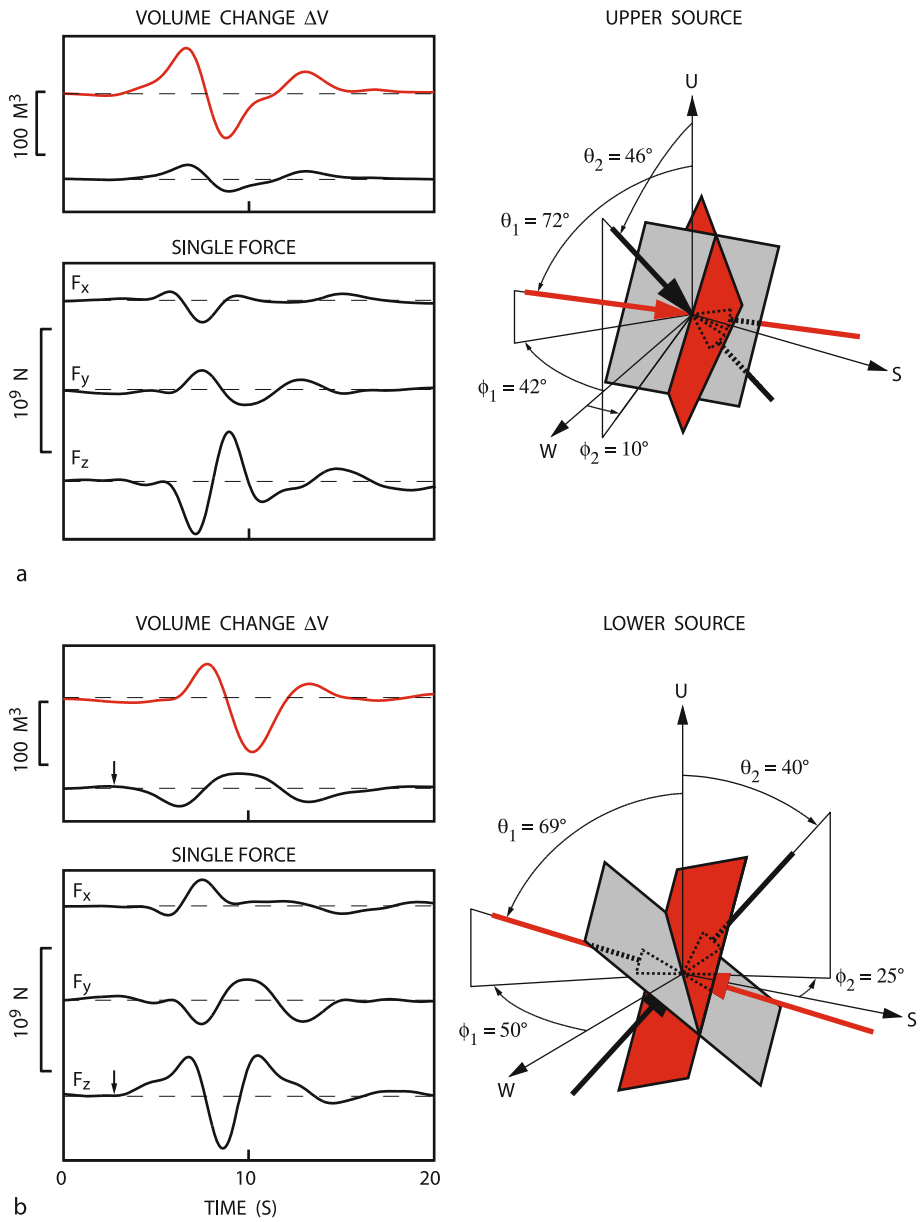

Volcanoes, Non-linear Processes in, Figure 12

Seismic source mechanisms imaged for the two flow disruption sites in the shallow conduit structure underlying the northern vent area at Stromboli. The two point-source mechanisms were obtained by Chouet et al. [35] from inversions of VLP waveforms recorded during explosions at Stromboli, and are positioned at different depths in the volcanic edifice (see text for details). Each source consists of two intersecting cracks and a single force with components F_x (east), F_y (north), and F_z (up). Volume changes are color-coded with the color of the cracks they represent in each source. Crack orientations are provided by the azimuth ϕ and polar angle θ of the dominant dipole normal to the crack plane, with arrow directions marking crack deflation (see Fig. 1c). The reference coordinates are W (west), S (south), and U (up). a Upper source. b Lower source. Small arrows in left panels mark the onset of deflation of the lower dike (gray dike in right panel) and synchronous start of the upward force. (Reproduced from [35])

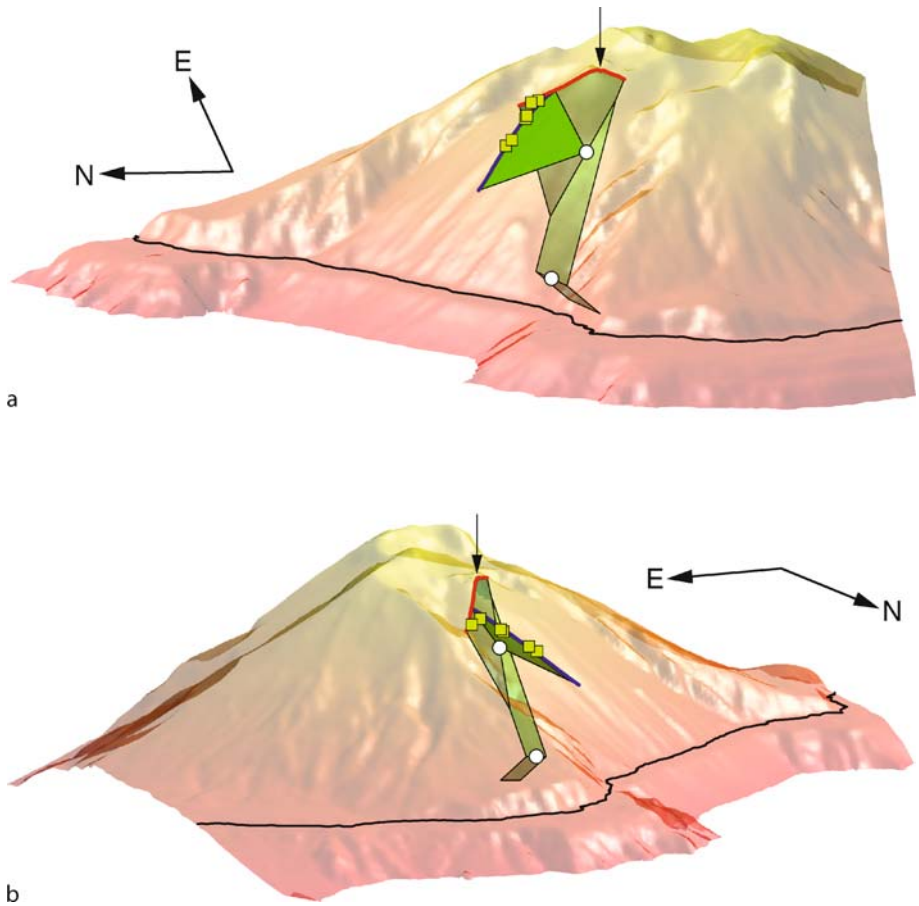

Volcanoes, Non-linear Processes in, Figure 13

Geometry of the upper 1 km of conduit underlying the northern vent area of Stromboli. A semi-transparent view of the northwest quadrant of the volcanic edifice provides the reference for the location and geometry of the conduit, which is derived from the seismic source mechanisms in Fig. 12. Thin black line indicates sea level. The summit of the volcano is 924 m above sea level (no vertical exaggeration). The two flow disruption sites that are sources of VLP elastic radiation are indicated by small circles. The irregular red and blue lines, respectively, represent the surface traces of the dominant and subdominant dike segments constituting the shallowest portions of the conduit system. The eruptive vent is marked by an arrow, and vents temporarily active during the flank eruption in 2002–2003 are marked by green squares. The lateral extents of individual dike segments are unknown and are shown for illustrative purpose only. a East-looking view. b South-looking view

tical forces at the two sources, suggesting that transmission of the force between the two sources occurs via the faster speed (3.5 km/s [32]) of the compressional wave in the rock matrix.

Laboratory simulations carried out by James et al. [60] provide insights into the origin of the initial pressurization and downward force observed at the upper source in Fig. 12a. These experiments investigate the ascent of a slug of gas in a vertical liquid-filed tube featuring a flare that abruptly doubles the cross sectional area. The tube is instrumented with pressure transducers mounted flush with the inner tube wall, and one accelerometer mounted on the exterior of the tube, and the whole assembly is free to move in the vertical direction. Detailed measurements of the flow transients obtained by James et al. [60] show that the transit of a gas slug through the tube flare involves complex changes in flow pattern. A characteristic pinching of the slug tail is observed to occur synchronously with strong pressure and acceleration transients at the time the slug clears the flare, a picture consistent with the downward and inward motion of a liquid piston formed by the thickening film of liquid falling past the slug expanding in the wider tube. The sudden deceleration of the liquid annulus as it impinges the narrower inlet to the lower tube segment generates a pressure pulse in the liquid below the flare and also induces a downward force on the apparatus.

These observations are consistent with the pressurization phase and initial downward force imaged for the upper source at Stromboli and a similar funneling mechanism was inferred by Chouet et al. [35] to be operative there. The repeatability of recorded pressure data and dependence of the magnitude of the pressure transient on slug size seen in the experiments of James et al. [60] are also in harmony with the observed spatio-temporal properties of VLP signals at Stromboli [32,35].

At the lower source, the start of the vertical force signal is synchronous with the onset of deflation of the lower dike segment (see arrows in the left panels in Fig. 12b). As the amplitude of the upward force increases, the lower dike segment continuously deflates. During the same interval, the dominant dike (red-colored volumetric trace in Fig. 12b) remains in a slightly deflated state. The lower dike reaches maximum deflation at the time the upper dike segment goes through a transition from weak contraction to expansion, and the upward force reaches its peak amplitude ~ 0.5 s later. This picture is consistent with a compression of the lower dike synchronous with a downward acceleration of the liquid mass, both of which are suggestive of increasing external pressure on the conduit wall resulting from the downward vertical force acting at the upper source. Compression of the lower dike segment proceeds unimpeded until this process is overprinted by the arrival of the much slower volumetric expansion signal from the upper source.

Although not illustrated here, a similar slug disruption mechanism was imaged in the dike system underlying the southern vent [35]. In both conduit systems, the early response of the lower source relative to the volumetric disturbance arriving from the upper source may be interpreted as the passive response of the liquid to the movement of the conduit wall induced by elastic radiation from the force acting at the upper source. The scenario emerging from these dynamics is that of an upper source representing an active fluid phase and passive solid phase, and a lower source representing an active solid phase and passive fluid phase. The overall seismic source process associated with eruptions at Stromboli may then be summarized as follows. A slug of gas formed in the deeper reaches of conduit [18] rises through the lower conduit corner (the lower seismic source). At this point, the slug is most likely a few meters long and traverses this corner aseismically; past this corner, the slug expands on its way to the upper conduit bifurcation (the upper seismic source). The transit time from the lower to the upper seismic source is probably in the range of 5–15 minutes [60], hence related changes in magmatic head are well beyond the capability of broadband seismometers to detect and are not apparent within the VLP band imaged in Fig. 12. As it traverses the upper seismic source, the slug length has expanded to tens of meters and the slug is by then seismically noisy. Gravitational slumping of the liquid occurs as the slug expands through a flare in the conduit at this location. The slumping liquid rapidly decelerates in the narrowing dike neck, increasing the liquid pressure and inducing a volume expansion in the main conduit and its subsidiary branch. The rapid deceleration and associated pressurization of the liquid couples to the conduit wall via the flare shoulders and induces a downward vertical force on the Earth. The volumetric signal propagates along the conduit at the slow speed of the crack wave, while the force signal itself propagates in the solid at the much higher speed of the compressional wave in the rock matrix, arriving at the lower conduit corner well before the crack wave. At this corner, the downward displacement of the rock induced by the force acting at the upper source impinges the bottom dike and squeezes this segment of conduit; this segment essentially acts like a spring that absorbs the downward motion of the rock. This scenario is consistent with both the small volume change of the lower dike segment, as well as its early response. The subsequent conduit response then reflects the combined effects of volumetric and mass oscillations of this liquid/gas/solid system, which are damped and eventually terminated by changing flow conditions.

The analyzes carried out by Chouet et al. [35] illuminate the subsurface processes driving eruptions at Stromboli and clearly point to the key role played by the conduit geometry in controlling fluid motion and resultant processes. Each irregularity or discontinuity in the conduit provides a site where pressure and momentum changes resulting from flow processes associated with the transit of a gas slug through the discontinuity are coupled to the Earth, or where the elastic response of the conduit can couple back into pressure and momentum changes in the fluid. The resulting processes are naturally oscillatory and involve diverse dynamics that reflect the complex physicochemical behavior of the volcanic system from the surface downward.

Future Directions

Seismology alone cannot directly see into the conduit and resolve details of the actual fluid dynamics at the origin of the seismic source mechanisms revealed by analyzes of LP or VLP signals. To develop a better understanding of fluid behavior responsible for these signals, laboratory experiments are required to explore the links between known flow processes and the resulting pressure and momentum changes. The pressure and mo-

mentum changes generated under laboratory conditions may then be compared with the pressure and momentum changes estimated from the time-varying moment-tensor and single-force components imaged from seismic data, yielding clues about the physical flow processes linked to the seismic source mechanism. Recent laboratory studies [59,60,85] have elucidated self-excitation mechanisms inherent to the fluid nonlinearity that are providing new insights into the mechanisms imaged from seismic data. In particular, the results obtained by James et al. [60] demonstrate that direct links between the moment-tensor and single-force seismic-source mechanism and fluid-flow processes are possible and could potentially provide a wealth of information not available from seismic data alone. Together with these laboratory advances, numerical studies of multiphase flows are required to shed light on both the micro- and macrophysics of such flows (e. g., Badalassi [15]), along with models exploring the coupled dynamics of fluid and solid [102]. The key to a better understanding of volcanic processes lies in a sustained effort aimed at cross-fertilization between increasingly realistic numerical and experimental models of the fluid dynamics and elastodynamics, spatially and temporally dense field measurements of diverse geophysical signals at all frequencies, and chemical and physical evidence recorded in the eruptive products. In the next section, I present a brief summary of a modeling approach to two-phase fluids that holds great promise for the quantification of volcanic and hydrothermal source processes.

Phase-Field Method

Liquid-gas mixtures are an ubiquitous feature of magmatic and hydrothermal systems. A major challenge in addressing these types of fluids is the description of the moving and deforming interface between the two components of fluid. Traditional fluid dynamics treats these as sharp interfaces on which matching boundary conditions must be imposed, which generally leads to intractable problems. Another difficulty is posed by the complex rheology of each component, whose internal microstructure is coupled with the flow field. The *phase-field method*, also known as the *diffuse-interface model*, circumvents these difficulties by considering the interface between components as a thin diffuse layer within which the two components are mixed and store a mixing energy. This idea follows from the original work of van der Waals [135] and expresses the properties of the interface by molecular forces and a mixing energy. In the limit of an interface width approaching zero, the diffuse-interface model reduces to the classical sharp-interface model and results in the proper expression for interfacial tension. Lowengrub and Truskinovsky [89] provide a detailed formulation of the conservative dynamics of the diffuse-interface model based on the classical procedure of Lagrangian mechanics. A review of diffuse-interface methods and related applications can also be found in Anderson et al. [8]. Noteworthy are the more recent applications by Kendon et al. [71], Xu et al. [141], Xu et al. [142], and Yue et al. [146]. The following discussion is restricted to a brief description of the basic principles of the method based on the work of Kendon et al. [71], who consider the evolution of a symmetric binary fluid mixture after a deep quench.

Equilibrium State A two-phase mixture is described by two scalar fields n_A and n_B expressing the local molar densities of the fluid components A and B. The corresponding local density, ρ, is given by:

$$\rho = m_A n_A + m_B n_B \,, \tag{2}$$

where m_A and m_B are the molecular weights of phases A and B, respectively. The local composition of the fluid is quantified by a dimensionless parameter, ϕ,

$$\phi = \frac{n_A - n_B}{n_A + n_B} \,, \tag{3}$$

called the order parameter; the values of ϕ span the range $[-1, +1]$ with the value $+1$ denoting the pure phase A and the value -1 denoting the pure phase B.

The chemical equilibrium between the two phases is described through a minimization of the free energy $\mathcal{F}$, which represents the integral over the body of the local free-energy density (with units of Jm^{-3} or, equivalently, Nm^{-2}), a function of the fluid composition and its gradient [19]:

$$\mathcal{F} = \int \left\{ f_0(\phi, P) + \frac{1}{2}\kappa(\nabla\phi)^2 \right\} \mathrm{d}\mathbf{r} \,. \tag{4}$$

The term $f_0(\phi, P)$ in the above equation represents the bulk free-energy density, which is assumed to take the form of a double-well with respect to ϕ. The detailed functional form of the double-well potential is not important [17,71,127,128,141,142,146] and may be set as [71,141]:

$$f_0(\phi, P) = \frac{a}{2}\phi^2 + \frac{b}{4}\phi^4 + P \ln \frac{P}{P_0} \,, \tag{5}$$

where the polynomial terms represent the bulk properties of the fluid. The parameter $a < 0$ and parameter b is always positive [141]. The term in P yields a positive

background pressure and does not affect the phase behavior (see discussion of pressure below). The form of this latter term is selected here for convenience, with P_0 representing an arbitrary reference pressure. As defined above, the function f_0 features two minima for $\phi = \pm\sqrt{-a/b}$ that correspond to the coexisting pure bulk phases. Assuming $b = -a$, one obtains the equilibrium values $\phi = \pm 1$ [141].

The squared gradient term in Eq. (4), with κ a positive constant, represents weakly non-local interactions between the components. The effect of these interactions favors mixing of the components in contrast to the tendency of the bulk free energy f_0, which is toward total separation of the phases into domains of pure components [146].

According to the model represented by Eq. (5), minimization of the free energy is achieved through the creation of two bulk domains at compositions $\phi = \pm 1$ separated by interfaces across which the composition varies smoothly from one phase to the other. The change in $\mathcal{F}$ induced by a small local change in composition is described by the chemical potential, μ,

$$\mu = \frac{\delta \mathcal{F}}{\delta \phi} , \tag{6}$$

and equilibrium between the coexisting bulk phases is reached when μ is everywhere zero [19,71]. Using Eqs. (4) and (5), one obtains the following expression for μ:

$$\mu = a\phi + b\phi^3 - \kappa \nabla^2 \phi . \tag{7}$$

The equilibrium interfacial profile is given by $\phi(x) = \tanh(2x/\xi)$, where x is the coordinate normal to the interface [141]. With the choice $b = -a$, this yields an interfacial width, $\xi = 2\sqrt{-2\kappa/a}$, and surface tension, $\sigma = (2/3)\sqrt{-2a\kappa}$ [141]. The interface thickness may then be defined as the width over which 90% of the variation of ϕ occurs; this yields an equilibrium interface thickness of 1.47222ξ.

As interfaces in the fluid can exert non-isotropic forces, pressure is a tensor. The thermodynamic pressure tensor P_{ij}^{th} is obtained from the free energy as [41,71,127]

$$\begin{aligned} P_{ij}^{\text{th}} &= \left(P\frac{\delta \mathcal{F}}{\delta P} + \phi \frac{\delta \mathcal{F}}{\delta \phi} - \left[f_0(\phi, P) + \frac{1}{2}\kappa(\nabla\phi)^2 \right] \right) \delta_{ij} \\ &\quad + \kappa \frac{\partial \phi}{\partial x_i}\frac{\partial \phi}{\partial x_j} \\ &= P + P_{ij}^{\text{chem}} , \end{aligned} \tag{8}$$

where P is the bulk pressure, and P_{ij}^{chem} is the chemical pressure component given by

$$\begin{aligned} P_{ij}^{\text{chem}} &= \left[\frac{1}{2}a\phi^2 + \frac{3}{4}b\phi^4 - \kappa\phi\nabla^2\phi - \frac{1}{2}\kappa(\nabla\phi)^2 \right] \delta_{ij} \\ &\quad + \kappa \frac{\partial \phi}{\partial x_i}\frac{\partial \phi}{\partial x_j} , \end{aligned} \tag{9}$$

in which δ_{ij} is the Kronecker symbol ($\delta_{ij} = 0$ for $i \neq j$, and $\delta_{ij} = 1$ for $i = j$). Note that the expression between brackets contributes to the isotropic fluid pressure and that only the last term is anisotropic.

Conservation Equations The temporal evolution of the fluid composition ϕ is described by the Cahn–Hilliard diffusion-advection equation [20]

$$\frac{\partial \phi}{\partial t} + \mathbf{v} \cdot \nabla \phi = \nabla \cdot (\gamma \nabla \mu) , \tag{10}$$

where γ is the order-parameter mobility and $\mathbf{v}(\mathbf{r})$ is the fluid velocity.

The dynamics of the fluid are described by the Navier–Stokes equations, whose exact forms depend on the system considered. The equations describing a viscous, compressible, isothermal flow are the conservation of mass:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0 , \tag{11}$$

and conservation of momentum:

$$\rho \left[\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} \right] = -\nabla \cdot \mathbf{P}^{\text{th}} + \nabla \cdot \tau + \rho \mathbf{g} , \tag{12}$$

where $\mathbf{g}$ is the acceleration of gravity vector, $\mathbf{P}^{\text{th}}$ is the thermodynamic pressure tensor with components given by Eqs. (8) and (9), and τ is the viscous stress tensor, which for a newtonian fluid is given by

$$\tau_{ij} = \lambda e_{kk}\delta_{ij} + 2\eta e_{ij} , \tag{13}$$

where η is the coefficient of viscosity (dynamic viscosity) and λ is the second coefficient of viscosity generally defined as $\lambda = -(2/3)\eta$ [130]. Summation over repeated indices is implied. In this equation e_{ij} is the strain rate tensor given by:

$$e_{ij} = \frac{1}{2}\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right) . \tag{14}$$

The field variables are the order parameter ϕ, the three components of fluid velocity $\mathbf{v}$, density ρ, and bulk pressure P. Under isothermal conditions P may be specified

through a van der Waals equation of state. The diffusion-advection Eq. (10), continuity Eq. (11), and the three momentum Eqs (12), along with an equation of state for P, therefore provide a complete description of the physicochemical evolution of the binary fluid mixture. All other quantities can be obtained from the relations (2), (3), (7), (8), (9), (13), and (14). For non-isothermal flow, a complete description is obtained from Eqs. (10–12), along with two additional equations describing the conservation of energy and entropy production [8].

The thermodynamic properties of the two-phase system enter the above equations through the chemical potential $\mu = \delta\mathcal{F}/\delta\phi$, and pressure tensor $\mathbf{P}^{\mathrm{th}}$, which are both obtained from the free energy Eq. (4). As a result, interfaces between the different fluid phases appear naturally within the model and do not need to be put in as boundary conditions. The procedure is quite general and may conceivably be applied to any complex fluid with a properly defined free energy.

The coupled Cahn–Hilliard/Navier–Stokes system of equations representing binary fluid systems is analytically intractable, and efforts have mainly been directed toward the development of robust, stable numerical schemes. The description of the fluid in terms of a smooth variation of the composition $\phi(\mathbf{r})$ provides a coarse-grained representation of the fluid where the smallest length scale is larger than the average distance between molecules. Under dynamical conditions, care must be taken to ensure that this coarse-grained description adequately represents the underlying microscopic physics. For the nominal interfacial thickness representing 90% of the variation of ϕ this typically requires 7–10 grids to resolve [146]. Numerical applications include the Lattice Boltzmann Method (LBM) [16,71,112,113,127,128,141,142], and various finite-difference approaches based on an Implicit-Explicit (IMEX) discretization [12], semi-implicit discretization [15,146], fully implicit discretization [73], or fully explicit central-differenced staggered-grid discretization [58].

An example of application of this method that may provide a useful metaphor for the process of Vulcanian degassing discussed in Sect. "Coupled Diffusive-Elastic Pressurization at Popocatepetl Volcano" is the 3D spinodal decomposition of a density-matched binary fluid mixture in a channel under shear [15]. A remarkable feature of this type of flow is the formation of string-like structures. Similar structures have been observed in immiscible viscoelastic systems subjected to complex flow fields [93], and such process may also play a role in the formation of degassing channels in a rhyolitic bubbly magma driven by slow pressurization in response to crystallization and degassing. A 2D application of the method to the dynamics of a gas slug ascending in a liquid-filled conduit featuring a flare is described in the chapter by *D'Auria and Martini*. The numerical results obtained by these authors reproduce the basic behavior observed in the experimental simulations of James et al. [59] and James et al. [60] and hold great potential as a key link between the seismic source mechanisms imaged at Stromboli (see Sect. "Slug Disruption at Stromboli ") and the underlying volcanic fluid dynamics.

Acknowledgment

I am grateful to Phil Dawson for his assistance in drafting figures. I am indebted to Robert Tilling and David Hill for careful reviews and helpful suggestions.

Bibliography

1. Acocella V, Neri M, Scarlato P (2006) Understanding shallow magma emplacement at volcanoes: Orthogonal feeder dykes during the 2002–2003 Stromboli (Italy) eruption. Geophys Res Lett 33:L17310. doi:10.1029/2006GL026862
2. Aki K, Fehler M, Das S (1977) Source mechanism of volcanic tremor: Fluid-driven crack models and their application to the 1963 Kilauea eruption. J Volcanol Geotherm Res 2:259–287
3. Aki K, Richards PG (1980) Quantitative Seismology. Freeman, New York, p 932
4. Almendros J, Chouet B, Dawson P (2001) Spatial extent of a hydrothermal system at Kilauea Volcano, Hawaii, determined form array analyses of shallow long-period seismicity. J Geophys Res 106:13581–13597
5. Almendros J, Chouet B, Dawson P, Huber C (2002) Mapping the sources of the seismic wave field at Kilauea Volcano, Hawaii, using data recorded on multiple seismic antennas. Bull Seismol Soc Am 92:2333–2351
6. Almendros J, Chouet B, Dawson P, Bond T (2002) Identifying elements of the plumbing system beneath Kilauea Volcano, Hawaii, from the source locations of very-long-period signals. Geophys J Int 148:303–312
7. Almendros J, Chouet B, Dawson P (2008) Shallow magma transport pathway under Kilauea Volcano, Hawaii, imaged from waveform inversions of very-long-period seismic data. 2-Source modeling J Geophys Res, submitted
8. Anderson DM, McFadden GB, Wheeler AA (1998) Diffuse-interface methods in fluid mechanics. Annu Rev Fluid Mech 30:139–165
9. Arciniega-Ceballos A, Chouet BA, Dawson P (1999) Very-long-period signals associated with Vulcanian explosions at Popocatepetl Volcano, Mexico. Geophys Res Lett 26: 3013–3016
10. Arciniega-Ceballos A, Chouet B, Dawson P (2003) Long-period events and tremor at Popocatepetl Volcano (1994–2000) and their broadband characteristics. Bull Volcanol 65: 124–135
11. Arciniega-Ceballos A, Chouet B, Dawson P, Asch G (2008) Broadband seismic measurements of degassing activity as-

sociated with lava effusion at Popocatepetl Volcano, Mexico. J Volcanol Geotherm Res 170:12–23
12. Ascher UM, Ruuth SJ, Wetton BTR (1995) Implicit-explicit methods for time-dependent partial differential equations. SIAM J Numer Anal 3:797–823
13. Aster R, Mah S, Kyle P, McIntosh W, Dunbar N, Johnson J, Ruiz M, McNamara S (2003) Very long period oscillations of Mount Erebus Volcano. J Geophys Res 108(B11). doi:10.1029/2002JB002101
14. Auger E, D'Auria L, Martini M, Chouet B, Dawson P (2006) Real-time monitoring and massive inversion of source parameters of very long period seismic signals: An application to Stromboli Volcano, Italy. Geophys Res Lett 33:L04301. doi:10.1029/2005GL024703
15. Badalassi VE, Ceniceros HD, Banerjee S (2003) Computation of multiphase systems with phase field models. J Comput Phys 190:371–397
16. Balazs AC, Verberg R, Pooley CM, Kuksenok O (2005) Modeling the flow of complex fluids through heterogeneous channels. Soft Matter 1:44–54
17. Bray AJ (1994) Theory of phase-ordering kinetics. Adv Phys 43:357–459
18. Burton M, Allard P, Muré F, La Spina A (2007) Magmatic gas composition reveals the source depth of slug-driven Strombolian explosive activity. Science 317(5835):227–230
19. Cahn JW, Hilliard JE (1958) Free energy of a nonuniform system. I. Interfacial free energy. J Chem Phys 28:258–267
20. Cahn JW, Hilliard JE (1959) Free energy of a nonuniform system. III, Nucleation in a two-component incompressible fluid. J Chem Phys 31:688–689
21. Cervelli PF, Miklius A (2003) The shallow magmatic system of Kilauea Volcano. US Geol Surv Prof Paper 1676:149–163
22. Chouet B, Julian BR (1985) Dynamics of an expanding fluid-filled crack. J Geophys Res 90:11187–11198
23. Chouet B (1986) Dynamics of a fluid-driven crack in three dimensions by the finite difference method. J Geophys Res 91:13967–13992
24. Chouet B (1988) Resonance of a fluid-driven crack: Radiation properties and implications for the source of long-period events and harmonic tremor. J Geophys Res 93:4375–4400
25. Chouet B (1992) A seismic source model for the source of long-period events and harmonic tremor. In: Gasparini P, Scarpa R, Aki K (eds) IAVCEI Proceedings in Volcanology, vol 3. Springer, New York, pp 133–156
26. Chouet BA, Page RA, Stephens CD, Lahr JC, Power JA (1994) Precursory swarms of long-period events at Redoubt Volcano (1989–1990), Alaska: Their origin and use as a forecasting tool. J Volcanol Geotherm Res 62:95–135
27. Chouet B (1996) Long-period volcano seismicity: its source and use in eruption forecasting. Nature 380:309–316
28. Chouet B (1996) New methods and future trends in seismological volcano monitoring. In: Scarpa R, Tilling RI (eds) Monitoring and Mitigation of Volcano Hazards. Springer, New York, pp 23–97
29. Chouet B, Saccorotti G, Martini M, Dawson P, De Luca G, Milana G, Scarpa R (1997) Source and path effects in the wave fields of tremor and explosions at Stromboli Volcano, Italy. J Geophys Res 102:15129–15150
30. Chouet B, Saccorotti G, Dawson P, Martini M, Scarpa R, De Luca G, Milana G, Cattaneo M (1999) Broadband measurements of the sources of explosions at Stromboli Volcano, Italy. Geophys Res Lett 26:1937–1940
31. Chouet B (2003) Volcano seismology. PAGEOPH 160(3–4): 739–788
32. Chouet B, Dawson P, Ohminato T, Martini M, Saccorotti G, Giudicepietro F, De Luca G, Milana G, Scarpa R (2003) Source mechanisms of explosions at Stromboli Volcano, Italy, determined from moment-tensor inversions of very-long-period data. J Geophys Res 108(B1):2019. doi:10.1029/2002JB001919
33. Chouet B, Dawson P, Arciniega-Ceballos A (2005) Source mechanism of Vulcanian degassing at Popocatepetl Volcano, Mexico, determined from waveform inversions of very long period signals. J Geophys Res 110:B07301. doi:10.1029/2004JB003524
34. Chouet B, Dawson P, Nakano M (2006) Dynamics of bubble growth and pressure recovery in a bubbly rhyolitic melt embedded in an elastic solid. J Geophys Res 111:B07310. doi:10.1029/2005JB004174
35. Chouet B, Dawson P, Martini M (2008) Shallow-conduit dynamics at Stromboli Volcano, Italy, imaged from waveform inversion. In: Lane SJ, Gilbert JS (eds) Fluid Motions in Volcanic Conduits A Source of Seismic and Accoustic Signals. Geol Soc Lond Special Pub 307:57–84
36. Collier L, Neuberg JW, Lensky N, Lyakhovsky V, Navon O (2006) Attenuation in gas-charged magma. J Volcanol Geotherm Res 153:21–36
37. Commander KW, Prosperetti A (1989) Linear pressure waves in bubbly liquids: Comparison between theory and experiments. J Acoust Soc Am 85:732–746
38. Dawson PB, Dietel C, Chouet BA, Honma K, Ohminato T, Okubo P (1998) A digitally telemetered broadband seismic network at Kilauea Volcano, Hawaii. US Geol Surv Open-File Rep 98–108:1–121
39. Dawson P, Whilldin D, Chouet B (2004) Application of near real-time radial semblance to locate the shallow magmatic conduit at Kilauea Volcano, Hawaii. Geophys Res Lett 31:L21606. doi:10.1029/2004GL021163
40. Dzurisin D, Vallance JW, Gerlach TM, Moran SC, Malone SD (2005) Mount St. Helens reawakens. Eos Trans AGU 86:25,29 doi:10.1029/2005EO030001
41. Evans R (1979) The nature of the liquid-vapour interface and other topics in the statistical mechanics of non-uniform, classical fluids. Adv Phys 28:148–200
42. Ferrazini V, Aki K (1987) Slow waves trapped in a fluid-filled infinite crack: Implications for volcanic tremor. J Geophys Res 92:9215–9233
43. Fujita E, Ida Y, Oikawa J (1995) Eigen oscillation of a fluid sphere and source mechanism of harmonic volcanic tremor. J Volcanol Geotherm Res 69:365–378
44. Fujita E, Ukawa M, Yamamoto E (2004) Subsurface cyclic magma sill expansions in the 2000 Miyakejima Volcano eruption: Possibility of two-phase flow oscillation. J Geophys Res 109:B04205. doi:10.1029/2003JB002556
45. Gil Cruz F, Chouet BA (1997) Long-period events, the most characteristic seismicity accompanying the emplacement and extrusion of a lava dome in Galeras Volcano, Colombia, in 1991. J Volcanol Geotherm Res 77:121–158
46. Hamada N, Jingu H, Ikumoto K (1976) On the volcanic earthquake with slowly decaying coda wave (in Japanese with English abstract). Bull Volcanol Soc Jpn 21:167–183
47. Hasegawa A, Zhao D, Hori S, Yamamoto A, Horiuchi S (1991)

Deep structure of the northeastern Japan arc and its relationship to seismic and volcanic activity. Nature 352:683–689
48. Heliker C, Mattox T (2003) The first two decades of the Pu'u O'o-Kupaianaha eruption: Chronology and selected bibliography. US Geol Surv Prof Paper 1676:1–27
49. Hess K, Dingwell DB (1996) Viscosities of hydrous leucogranitic melts: A non-Arrhenian model. Am Mineral 81:1297–1300
50. Hidayat D, Voight B, Langston C, Ratdomopurbo A, Ebeling C (2000) Broadband seismic experiment at Merapi Volcano, Java, Indonesia: Very-long-period pulses embedded in multiphase earthquakes. J Volcanol Geotherm Res 100:215–231
51. Hidayat D, Chouet B, Voight B, Dawson P, Ratdomopurbo A (2002) Source mechanism of very-long-period signals accompanying dome growth activity at Merapi Volcano, Indonesia. Geophys Res Lett 29(23):2118. doi:10.1029/2002GL015013
52. Hidayat D, Chouet B, Voight B, Dawson P, Ratdomopurbo A (2003) Correction to Source mechanism of very-long-period signals accompanying dome growth activity at Merapi Volcano, Indonesia. Geophys Res Lett 30(10):2118. doi:10.1029/2003GL017211
53. Hill DP, Dawson P, Johnston MJS, Pitt AM, Biasi G, Smith K (2002) Very-long-period volcanic earthquakes beneath Mammoth Mountain, California. Geophys Res Lett 29(10):1370. doi:10.1029/2002GL014833
54. Hill DP, Prejean S (2005) Magmatic unrest beneath Mammoth Mountain, California. J Volcanol Geotherm Res 146:257–283
55. Hirabayashi J (1999) Formation of volcanic fluid reservoir and volcanic activity. J Balneol Soc Jpn 49:99–105
56. Ichihara M, Kameda M (2004) Propagation of acoustic waves in a visco-elastic two-phase system: Influences of the liquid viscosity and the internal diffusion. J Volcanol Geotherm Res 137:73–91
57. Iverson RM, Dzurisin D, Gardner CA, Gerlach TM, LaHusen RG, Lisowski M, Major JJ, Malone SD, Messerich JA, Moran SC, Pallister JS, Qamar AI, Schilling SP, Vallance JW (2006) Dynamics of seismogenic volcanic extrusion at Mount St. Helens in 2004–05. Nature 444:439–443. doi:10.1038/nature05322
58. Jacqmin D (1999) Calculation of two-phase Navier–Stokes Flows using phase-field modeling. J Comput Phys 155:96–127
59. James MR, Lane SJ, Chouet B, Gilbert JS (2004) Pressure changes associated with the ascent and bursting of gas slugs in liquid-filled vertical and inclined conduits. J Volcanol Geotherm Res 129:61–82
60. James MR, Lane SJ, Chouet BA (2006) Gas slug ascent through changes in conduit diameter: Laboratory insights into a volcano-seismic source process in low-viscosity magmas. J Geophys Res 111:B05201. doi:10.1029/2005JB003718
61. Jones RAL (2002) Soft Condensed Matter. Oxford University Press, Oxford, p 195
62. Julian BR (1994) Volcanic tremor, Nonlinear excitation by fluid flow. J Geophys Res 99:11859–11877
63. Kanamori H, Given JW (1982) Analysis of long-period waves excited by the May 18, 1980, eruption of Mount St. Helens – A terrestrial monopole? J Geophys Res 87:5422–5432
64. Kanamori H, Given JW, Lay T (1984) Analysis of seismic body waves excited by the Mount St. Helens eruption of May 18:1980. J Geophys Res 89:1856–1866
65. Kaneshima S, Kawakatsu H, Matsubayashi H, Sudo Y, Tsutsui T, Ohminato T, Ito H, Uhira K, Yamasato H, Oikawa J, Takeo M, Iidaka T (1996) Mechanism of phreatic eruptions at Aso Volcano inferred from near-field broadband seismic observations. Science 273:642–645
66. Kawakatsu H (1989) Centroid single force inversion of seismic waves generated by landslides. J Geophys Res 94:12363–12374
67. Kawakatsu H, Ohminato T, Ito H, Kuwahara Y (1992) Broadband seismic observation at Sakurajima Volcano, Japan. Geophys Res Lett 19:1959–1962
68. Kawakatsu H, Ohminato T, Ito H (1994) 10-s-period volcanic tremors observed over a wide area in southwestern Japan. Geophys Res Lett 21:1963–1966
69. Kawakatsu H, Kaneshima S, Matsubayashi H, Ohminato T, Sudo Y, Tsutsui T, Uhira K, Yamasato H, Ito H, Legrand D (2000) Aso94: Aso seismic observation with broadband instruments. J Volcanol Geotherm Res 101:129–154
70. Kedar S, Sturtevant B, Kanamori H (1996) The origin of harmonic tremor at Old Faithful Geyser. Nature 379:708–711
71. Kendon VM, Cates ME, Pagonabarraga I, Desplat J-C, Bladon P (2001) Inertial effects in three-dimensional spinodal decomposition of a symmetric binary fluid mixture: A lattice Boltzmann study. J Fluid Mech 440:147–203
72. Kieffer SW (1977) Sound speed in liquid-gas mixtures: Water-air and water-steam. J Geophys Res 82:2895–2904
73. Kim J, Kang K, Lowengrub J (2004) Conservative multigrid methods for Cahn–Hilliard fluids. J Comput Phys 193:511–543
74. Klein FW, Koyanagi RY, Nakata JS, Tanigawa WR (1987) The seismicity of Kilauea's magma system. In: Decker RW, Wright TL, Stauffer RW (eds) Volcanism in Hawaii. US Geol. Surv. Prof. Pap., 1350. US government printing office, Washington, pp 1019–1185
75. Koyanagi RY, Chouet B, Aki K (1987) Origin of volcanic tremor in Hawaii. Part I: Compilation of seismic data from the Hawaiian Volcano Observatory, 1972 to 1985. In: Decker RW, Wright TL, Stauffer RW (eds) Volcanism in Hawaii, US Geol. Surv. Prof. Pap., 1350. US government printing office, Washington, pp 1221–1257
76. Kumagai H, Chouet BA (1999) The complex frequencies of long-period seismic events as probes of fluid composition beneath volcanoes. Geophys Int 138:F7–F12
77. Kumagai H, Chouet BA (2000) Acoustic properties of a crack containing magmatic or hydrothermal fluids. J Geophys Res 105:25493–25512
78. Kumagai H, Ohminato T, Nakano M, Ooi M, Kubo A, Inoue H, Oikawa J (2001) Very-long-period seismic signals and caldera formation at Miyake Island, Japan. Science 293:687–690
79. Kumagai H, Chouet BA, Nakano M (2002) Temporal evolution of a hydrothermal system in Kusatsu-Shirane Volcano, Japan, inferred from the complex frequencies of long-period events. J Geophys Res 107(B10):2236. doi:10.1029/2001JB000653
80. Kumagai H, Chouet BA, Nakano M (2002) Waveform inversion of oscillatory signatures in long-period events beneath volcanoes. J Geophys Res 107(B11)2301. doi:10.1029/2001JB001704
81. Kumagai H, Miyakawa K, Negishi H, Inoue H, Obara K, Suetsugu D (2003) Magmatic dyke resonances inferred from very-long-period seismic signals. Science 299:2058–2061
82. Kumagai H, Chouet BA, Dawson PB (2005) Source process of a long-period event at Kilauea Volcano, Hawaii. Geophys J Int 161:243–254
83. Kumagai H (2006) Temporal evolution of a magmatic dike

system inferred from the complex frequencies of very long period seismic signals. J Geophys Res 111:B06201. doi:10.1029/2005JB003881
84. Kumagai H, Yepes H, Vaca M, Caceres V, Nagai T, Yokoe K, Imai T, Miyakawa K, Yamashina T, Arrais S, Vasconez F, Pinajota E, Cisneros C, Ramos C, Paredes M, Gomezjurado L, Garcia-Aristizabal A, Molina I, Ramon P, Segovia M, Palacios P, Troncoso L, Alvarado A, Aguilar J, Pozo J, Enriquez W, Mothes P, Hall M, Inoue I, Nakano M, Inoue H (2007) Enhancing volcano-monitoring capabilities in Ecuador. Eos Trans Am Geophys Union 88:245–246
85. Lane SJ, Chouet BA, Phillips JC, Dawson P, Ryan GA, Hurst E (2001) Experimental observations of pressure oscillations and flow regimes in an analogue volcanic system. J Geophys Res 106:6461–6476
86. Legrand D, Kaneshima S, Kawakatsu H (2000) Moment tensor analysis of near-field broadband waveforms observed at Aso Volcano, Japan. J Volcanol Geotherm Res 101:155–169
87. Lensky NG, Lyakhovsky V, Navon O (2002) Expansion dynamics of volatile-saturated liquids and bulk viscosity of bubbly magmas. J Fluid Mech 460:39–56
88. Lensky NG, Navon O, Lyakhovsky V (2004) Bubble growth during decompression of magma: experimental and theoretical investigation. J Volcanol Geotherm Res 129:7–22
89. Lowengrub J, Truskinovsky L (1998) Quasi-incompressible Cahn-Hilliard fluids and topological transitions. Proc Royal Soc Lond A 454:2617–2654
90. Lyakhovsky V, Hurwitz S, Navon O (1996) Bubble growth in rhyolitic melts: experimental and numerical investigation. Bull Volcanol 58:19–32
91. Major JJ, Scott WE, Driedger C, Dzurisin D (2005) Mount St. Helens erupts again – Activity from September 2004 through March 2005. US Geological Survey Fact Sheet FS2005-3036. US goverment printing office, Washington, p 4
92. Meier GEA, Grabitz G, Jungowsky WM, Witczak KJ, Anderson JS (1978) Oscillations of the supersonic flow downstream of an abrupt inrease in duct crosssection. Mitteilungen aus dem Max-Plank Institut für Strömungsforschung und der Aerodynamischen Versuchsanstalt 65:1–172
93. Migler KB (2001) String formation in sheared polymer blends: coalescence, breakup, and finite size effects. Phys Rev Lett 86(6):1023–1026
94. Moran SC, Malone SD, Qamar AI, Thelen W, Wright AK, Caplan-Auerbach J (2007) 2004–2005 seismicity associated with the renewed dome-building eruption of Mount St. Helens. In: Sherrod DR, Scott WE, Stauffer PH (eds) A Volcano rekindled: the first year of renewed eruptions at Mount St. Helens, 2004–2006. US Geological Survey Prof. Pap. 2007-XXXX
95. Morrissey MM, Chouet BA (1997) A numerical investigation of choked flow dynamics and its application to the triggering mechanism of long-period events at Redoubt Volcano, Alaska. J Geophys Res 102:7965–7983
96. Nakano M, Kumagai H, Kumazawa M, Yamaoka K, Chouet BA (1998) The excitation and characteristic frequency of the long-period volcanic event: An approach based on an inhomogeneous autoregressive model of a linear dynamic system. J Geophys Res 103:10031–10046
97. Nakano M, Kumagai H, Chouet BA (2003) Source mechanism of long-period events at Kusatsu-Shirane Volcano, Japan, inferred from waveform inversion of the effective excitation functions. J Volcanol Geotherm Res 122:149–164
98. Navon O, Lyakhovsky V (1998) Vesiculation processes in silicic magmas. Geol Soc Lond Special Pub 145:27–50
99. Neuberg J, Luckett R, Ripepe M, Braun T (1994) Highlights from a seismic broadband array on Stromboli Volcano. Geophys Res Lett 21:749–752
100. Neuberg JW, Tuffen H, Collier L, Green D, Powell T, Dingwell D (2006) The trigger mechanism of low-frequency earthquakes on Montserrat. J Volcanol Geotherm Res 153:37–50
101. Nishimura T, Nakamichi H, Tanaka S, Sato M, Kobayashi T, Ueki S, Hamaguchi H, Ohtake M, Sato H (2000) Source process of very long period seismic events associated with the 1998 activity of Iwate Volcano, northeastern Japan. J Geophys Res 105:19135–19147
102. Nishimura T, Chouet B (2003) A numerical simulation of magma motion, crustal deformation, and seismic radiation associated with volcanic eruptions. Geophys. J Int 153: 699–718
103. Nishimura T, Ueki S, Yamawaki T, Tanaka S, Hasino H, Sato M, Nakamichi H, Hamaguchi H (2003) Broadband seismic signals associated with the 2000 volcanic unrest of Mount Bandai, northeastern Japan. J Volcanol Geotherm Res 119:51–59
104. Nishimura T (2004) Pressure recovery in magma due to bubble growth. Geophys Res Lett 31:L12613. doi:10.1029/2004GL019810
105. Ochs FA, Lange RA (1999) The density of hydrous magmatic liquids. Science 283:1314–1317
106. Ohba T, Hirabayashi J, Nogami K (2000) D/H and $^{18}O/^{16}O$ ratios of water in the crater lake of Kusatsu-Shirane Volcano, Japan. J Volcanol Geotherm Res 97:329–346
107. Ohminato T, Chouet BA (1997) A free-surface boundary condition for including 3D topography in the finite-difference method. Bull Seismol Soc Am 87:494–515
108. Ohminato T, Ereditato D (1997) Broadband seismic observations at Satsuma-Iwojima Volcano, Japan. Geophys Res Lett 24:2845–2848
109. Ohminato T, Chouet BA, Dawson PB, Kedar S (1998) Waveform inversion of very-long-period impulsive signals associated with magmatic injection beneath Kilauea volcano, Hawaii. J Geophys Res 103:23839–23862
110. Ohminato T (2006) Characteristics and source modeling of broadband seismic signals associated with the hydrothermal system at Satsuma–IwoJima volcano, Japan. J Volcanol Geotherm Res 158:467–490
111. Ohminato T, Takeo M, Kumagai H, Yamashina T, Oikawa J, Koyama E, Tsuji H, Urabe T (2006) Vulcanian eruptions with dominant single force components observed during the Asama 2004 volcanic activity in Japan. Earth Planets Space 58:583–593
112. Orlandini E, Swift MR, Yeomans JM (1995) A Lattice Boltzmann Model of binary fluid mixtures. Europhys Lett 32:463
113. Pooley CM, Kuksenok O, Balasz AC (2005) Convection-driven pattern formation in phase-separating binary fluids. Phys Rev E 71:030501(R)
114. Power JA, Stihler SD, White RA, Moran SC (2004) Observations of deep long-period (DLP) seismic events beneath Aleutian Arc Volcanoes, 1989–2002. J Volcanol Geotherm Res 138:243–266
115. Proussevitch AA, Sahagian DL, Anderson AT (1993) Dynamics of diffusive bubble growth in magmas: Isothermal case. J Geophys Res 98:22283–22307
116. Proussevitch AA, Sahagian DL (1996) Dynamics of coupled

diffusive and decompressive bubble growth in magmatic systems. J Geophys Res 101:17447–17455
117. Proussevitch AA, Sahagian DL (1998) Dynamics and energetics of bubble growth in magmas: Analytical formulation and numerical modeling. J Geophys Res 103:18223–18251
118. Ripepe M, Poggi P, Braun T, Gordeev E (1996) Infrasonic waves and volcanic tremor at Stromboli. Geophys Res Lett 23: 181–184
119. Rowe CA, Aster RC, Kyle PR, Schlue JW, Dibble RR (1998) Broadband recording of Strombolian explosions and associated very-long-period seismic signals on Mount Erebus Volcano, Ross Island, Antarctica. Geophys Res Lett 25:2297–2300
120. Saccorotti G, Chouet B, Dawson P (2001) Wavefield properties of a shallow long-period event and tremor at Kilauea Volcano, Hawaii. J Volcanol Geotherm Res 109:163–189
121. Scriven LE (1959) On the dynamics of phase growth. Chem Eng Sci 10:1–13
122. Shaw HR, Chouet B (1989) Singularity spectrum of intermittent seismic tremor at Kilauea Volcano, Hawaii. Geophys Res Lett 16:195–198
123. Shaw HR, Chouet B (1991) Fractal hierarchies of magma transport in Hawaii and critical self-organization of tremor. J Geophys Res 96:10191–10207
124. Shimomura Y, Nishimura T, Sato H (2006) Bubble growth processes in magma surrounded by an elastic medium. J Volcanol Geotherm Res 155:307–322
125. Sparks R (1978) The dynamics of bubble formation and growth in magmas: A review and analysis. J Volcanol Geotherm Res 3:1–38
126. Stephens CD, Chouet BA (2001) Evolution of the December 14, 1989 precursory long-period event swarm at Redoubt Volcano, Alaska. J Volcanol Geotherm Res 109:133–148
127. Swift MR, Osborn WR, Yeomans JM (1995) Lattice Boltzmann simulation of nonideal fluids. Phys Rev Lett 75(5):0–833
128. Swift MR, Orlandini E, Osborn WR, Yeomans JM (1996) Lattice Boltzmann simulations of liquid-gas and binary fluid systems. Phys Rev E 54(5):5041–5052
129. Takei Y, Kumazawa M (1994) Why have the single force and torque been excluded from seismic source models? Geophys J Int 118:20–30
130. Tannehill JC, Anderson DA, Pletcher RH (1997) Computational Fluid Mechanics and Heat Transfer, 2nd edn. Taylor and Francis, Philadelphia, p 792
131. Toramaru A (1989) Vesiculation process and bubble size distributions in ascending magmas with constant velocities. J Geophys Res 94:17523–17542
132. Toramaru A (1995) Numerical study of nucleation and growth of bubbles in viscous magmas. J Geophys Res 100:1913–1931
133. Tuffen H, Dingwell DB, Pinkerton H (2003) Repeated fracture and healing of silicic magma generates flow banding and earthquakes? Geology 31:1089–1092
134. Uhira K, Yamasato H, Takeo M (1994) Source mechanism of seismic waves excited by pyroclastic flows observed at Unzen Volcano, Japan. J Geophys Res 99:17757–17773
135. van der Waals JD (1979) The thermodynamic theory of capillarity under the hypothesis of a continuous variation of density. Verhandel Konink. Acad. Weten. Amsterdam, (Sect. 1), 1, pp 1–56. Translation by Rowlingson JS. J Statist Phys 20: 197–244
136. Vergniolle S, Brandeis G, Mareschal JC (1996) Strombolian explosions, 2. Eruption dynamics determined from acoustic measurements. J Geophys Res 101:20449–20466
137. Waite GP, Chouet BA, Dawson PB (2008) Eruption dynamics at Mount St. Helens imaged from broadband seismic waveforms: Interaction of the shallow magmatic and hydrothermal systems. J Geophys Res 113, B02305. doi:10.1029/2007JB005259
138. Wallis GB (1969) One-dimensional Two-phase Flow. MacGraw-Hill, New York, pp 408
139. White RA (1996) Precursory deep long-period earthquakes at Mount Pinatubo: Spatio-temporal link to a basalt trigger. In: Newhall CG, Punongbayan RS (eds) Fire and Mud: Eruptions and Lahars of Mount Pinatubo, Philippines. University of Washington Press, Seattle, pp 307–326
140. Wilson L, Head JW (1981) Ascent and eruption of basaltic magma on the earth and moon. J Geophys Res 86:2971–3001
141. Xu A, Gonnella G, Lamura A (2003) Phase-separating binary fluids under oscillatory shear. Phys Rev E 67:056105
142. Xu A, Gonnella G, Lamura A (2004) Phase separation of incompressible binary fluids with lattice Boltzmann methods. Physica A 331:10–22
143. Yamamoto M, Kawakatsu H, Kaneshima S, Mori T, Tsutsui T, Sudo Y, Morita Y (1999) Detection of a crack-like conduit beneath the active crater at Aso Volcano, Japan. Geophys Res Lett 26:3677–3680
144. Yamamoto M, Kawakatsu H, Yomogida K, Koyama J (2002) Long-period (12 sec) volcanic tremor observed at Usu 2000 eruption: seismological detection of a deep magma plumbing system. Geophys Res Lett 29(9):1329. doi:10.1029/2001GL013996
145. Yoon SW, Crum LA, Prosperetti A, Lu NQ (1991) An investigation of the collective oscillations of a bubble cloud. J Acoust Soc Am 89:700–706
146. Yue P, Feng JJ, Liu C, Shen J (2004) A diffuse-interface method for simulating two-phase flows of complex fluids. J Fluid Mech 515:293–317
147. Zhang Y, Xu Z, Liu Y (2003) Viscosity of hydrous rhyolitic melts inferred from kinetic experiments: A new viscosity model. Am Mineral 88:1741–1752

Volcano Seismic Signals, Source Quantification of

Hiroyuki Kumagai[1,2]
[1] National Research Institute for Earth Science and Disaster Prevention, Tsukuba, Japan
[2] IAVCEI/IASPEI Joint Commission on Volcano Seismology, Tsukuba, Japan

Article Outline

Glossary
Definition of the Subject
Introduction
Phenomenological Representation of Seismic Sources
Waveform Inversion

Spectral Analysis
Fluid-Solid Interactions
Future Directions
Acknowledgment
Appendix A: Green's Functions
Appendix B: Moment Tensor for a Spherical Source
Appendix C: Moment Tensor for a Cylindrical Source
Bibliography

Glossary

Moment tensor A point seismic source representation defined by the first-order moment of the equivalent body force or the stress glut. Slip on a fault as well as volumetric changes such as an isotropic expansion and tensile crack can be represented by the moment tensor.

Waveform inversion An approach to estimate source mechanisms and locations of seismic events by finding the best fits between observed and synthesized seismograms.

Autoregressive equation A difference form of the equation of motion of a linear dynamic system, which is a basic equation to determine the complex frequencies (frequencies and Q factors) of decaying harmonic oscillations in observed signals.

Crack wave A dispersive wave generated by fluid-solid interactions in a crack. The phase velocity of the crack wave is smaller than the acoustic velocity of the fluid in the crack.

Definition of the Subject

Volcano seismicity produces a wide variety of seismic signals that provide glimpses of the internal dynamics of volcanic systems. Quantitative approaches to analyze and interpret volcano-seismic signals have been developed since the late 1970s. The availability of seismic equipments with wide frequency and dynamic ranges since the early 1990s revealed a further variety of volcano-seismic signals in oscillation periods longer than a few seconds. Quantification of the sources of volcano-seismic signals is crucial to achieving a better understanding of the physical states and dynamics of magmatic and hydrothermal systems.

Introduction

Volcanoes occur in tectonically active regions of the earth where magmatic and hydrothermal fluids display complex interactions with volcanic rocks and the atmosphere. Volcano seismicity is the manifestation of such complex interactions occurring inside of volcanic edifices. Volcano seismology aims at achieving a better understanding of the physical states and dynamics of magmatic and hydrothermal systems beneath active volcanoes through observation, analysis, and interpretation of volcano-seismic signals. It is well known that volcano-seismic signals have a wide variety of signatures (Fig. 1), which have been classified by various schemes e. g., [25,61,80,84]. The volcano-tectonic (VT) earthquake, low-frequency (LF) or long-period (LP) event, tremor, very-long-period (VLP) event, hybrid event, and explosion earthquake have been traditionally used to specify volcano-seismic signals. Here, LF denotes frequencies between 0.5 and 5 Hz [61], and LP and VLP denote periods between 0.2 and 2 s and longer than a few seconds, respectively [25].

In contrast to the various classification schemes, source processes of volcano-seismic signals may be classified into three main families: (i) brittle failure in volcanic rock triggered by fluid movement, (ii) transient pressure disturbance caused by mass transport and/or volumetric change, and (iii) resonance of a fluid-filled resonator excited by a pressure disturbance. These three processes may be linked to the signals as follows: (i) VT earthquake, (ii) VLP event and explosion earthquake, (iii) LF or LP event, tremor, and VLP event, and (i)+(iii) hybrid event. VT earthquakes, sometimes called high-frequency (HF) earthquakes, are indistinguishable from ordinary tectonic earthquakes, but occur in volcanic regions due to process (i). LF or LP events have signals characterized by decaying oscillations, whereas the signature of tremor shows continuous oscillations. These oscillatory signatures can be interpreted as process (iii), in which the excitations are impulsive for LF or LP events and successive for tremor. Most VLP events are characterized by impulsive signatures, which may be generated by process (ii). However, some VLP events show sustained oscillations or decaying harmonic oscillations, which may reflect process (iii). Hybrid events show mixed characters of both VT earthquakes and LP events, suggesting that processes (i) and (iii) both take place at the source. Explosion earthquakes occur in association with eruptions in which process (ii) is dominant.

Recently, there have been remarkable advances in quantitative understanding of the sources of volcano-seismic signals. These advances were made possible by the development of source models and analysis techniques coupled with increased computer capacity and the availability of seismic observation equipment with wide frequency and dynamic ranges. Major studies that have contributed to these advances are summarized as follows: Detailed magma intrusion processes imaged by VT earthquakes e. g., [19,43,46,86,111,112,123]; Quantitative interpretations of oscillations in LP and VLP events and tremor

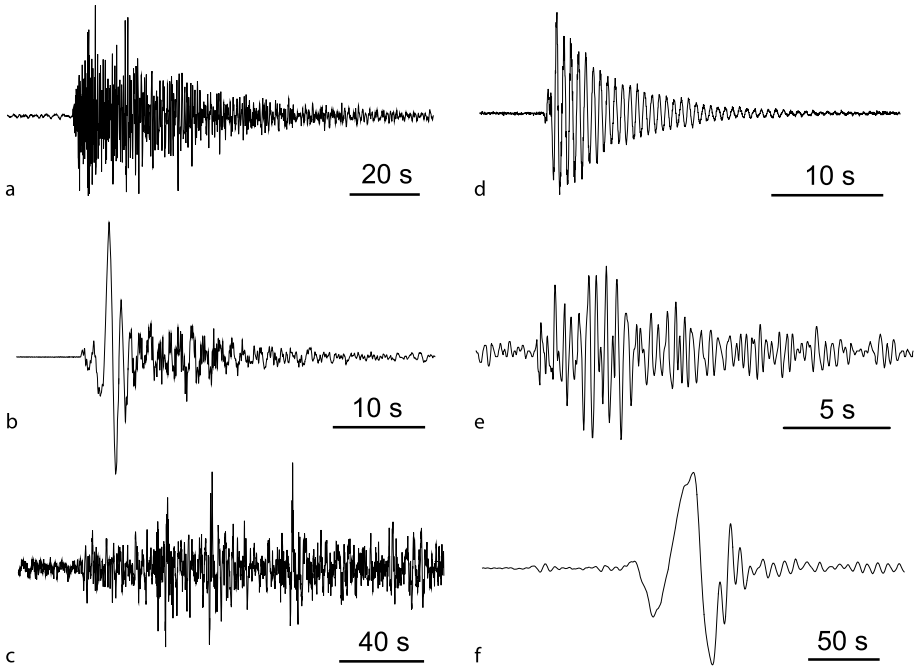

Volcano Seismic Signals, Source Quantification of, Figure 1
Various volcano-seismic signals: a volcano-tectonic earthquake that occurred between Miyakejima and Kozujima, Japan; b explosion earthquake observed at Asama volcano, Japan; c tremor that occurred beneath Mt. Fuji, Japan; d long-period event observed at Kusatsu–Shirane volcano, Japan; e long-period event observed at Guagua Pichincha volcano, Ecuador; f very-long-period event that occurred beneath Miyakejima, Japan (low-pass filtered at 20 s)

to diagnose the states of fluids at the sources of these signals e. g., [2,7,21,22,23,24,28,29,38,39,40,41,42,44,52,53,65, 66, 67, 70, 71, 73, 77, 78, 85, 87, 88, 90, 91, 92, 93, 94, 95, 98, 134]; Tracking the sources of LP events and tremor e. g., [3,4,5,13,30,45,82,113,114]; Waveform analysis of impulsive and oscillatory signatures in VLP events and explosion earthquakes to investigate source dynamics e. g., [6,8,9,10,11,31,32,33,34,36,47,48,49,51,54,55,56,57, 58,59,60,64,69,72,76,97,102,103,105,107,108,110,120,122, 126,132,133]. These volcano seismological studies have been reviewed by various authors [25,26,27,61,80,81,96].

The purpose of this article is to provide a systematic presentation of the theoretical basis for quantification of the sources of volcano-seismic signals suitable for readers, including graduate students and young researchers, who are new to the study of volcano seismology. I focus on four subjects: (1) phenomenological representation of seismic sources, (2) waveform inversion to estimate source mechanisms, (3) spectral analysis based on an autoregressive model, and (4) physical properties of fluid-solid coupled waves. Some additional notes are provided in Appendixes. These constitute the basic elements needed to quantify the sources of volcano-seismic signals. Applications of quantitative approaches are not fully reviewed in this manuscript. Readers are encouraged to consult both the original research papers and the review papers cited above.

Phenomenological Representation of Seismic Sources

We first derive the phenomenological representation of seismic sources using single force and moment tensor. This representation provides the basis to study source processes of volcano seismicity. I follow the approach presented by Backus and Mulcahy [12] (hereafter referred to as BM) and Aki and Richards [1] for the moment tensor representation. The theory elaborated by Takei and Kumazawa [118,119] is used to introduce single force in seismic sources.

Stress Glut

We consider indigenous seismic sources occurring within the closed earth system. This means that external bodies and detached mass are excluded. Let us assume an isotropic medium for the earth, which is represented by

density ρ and Lame's constants λ and μ. The exact equation of motion of the earth is given by

$$\rho \frac{\partial^2 u_i}{\partial t^2} = \frac{\partial \sigma_{ni}^{\text{true}}}{\partial x_n} , \tag{1}$$

where u_i and $\sigma_{ni}^{\text{true}}$ are the true displacement field and true stress tensor, respectively, actually produced in the earth. These are functions of space $\boldsymbol{x} = (x_1, x_2, x_3)$ and time t. The summation convention for repeated subscripts is used throughout this paper unless otherwise stated. In BM's theory, the failure of the elastic stress-strain relationship in source processes, such as slip across a plane that can not be described by the linear elastodynamics, is represented by an equivalent body force. We use a model stress $\sigma_{ni}^{\text{model}}$ given by the elastic stress-strain relationship as

$$\sigma_{ni}^{\text{model}} = \lambda \frac{\partial u_k}{\partial x_k} \delta_{ni} + \mu \left(\frac{\partial u_n}{\partial x_i} + \frac{\partial u_i}{\partial x_n} \right) , \tag{2}$$

where δ_{ni} is the Kronecker symbol ($\delta_{ni} = 0$ for $n \neq i$ and $\delta_{ni} = 1$ for $n = i$). By replacing the true stress by the model stress, Eq. (1) can be rewritten as follows:

$$\rho \frac{\partial^2 u_i}{\partial t^2} = \frac{\partial \sigma_{ni}^{\text{model}}}{\partial x_n} + f_i^S , \tag{3}$$

where f_i^S is the equivalent body force defined as

$$f_i^S = \frac{\partial}{\partial x_n} (\sigma_{ni}^{\text{true}} - \sigma_{ni}^{\text{model}}) . \tag{4}$$

The difference between $\sigma_{ni}^{\text{true}}$ and $\sigma_{ni}^{\text{model}}$ is called the stress glut. The equivalent body force, which is the spatial derivative of the stress glut, describes the source and vanishes outside the source region for motion in a completely elastic medium.

We introduce Green's function $G_{ij}(\boldsymbol{x}, t; \boldsymbol{\eta}, \tau)$, which is the ith component of displacement at $(\boldsymbol{x}, t)$ excited by the unit impulse applied at $\boldsymbol{x} = \boldsymbol{\eta}$ and $t = \tau$ in the j-direction. $G_{ij}(\boldsymbol{x}, t; \boldsymbol{\eta}, \tau)$ satisfies the following equation:

$$\rho \frac{\partial^2 G_{ij}}{\partial t^2} = \frac{\partial}{\partial x_n} \left\{ \lambda \frac{\partial G_{kj}}{\partial x_k} \delta_{ni} + \mu \left(\frac{\partial G_{nj}}{\partial x_i} + \frac{\partial G_{ij}}{\partial x_n} \right) \right\} + \delta(\boldsymbol{x} - \boldsymbol{\eta}) \delta(t - \tau) \delta_{ij} . \tag{5}$$

Note that $G_{ij}(\boldsymbol{x}, t; \boldsymbol{\eta}, \tau) = G_{ij}(\boldsymbol{x}, t - \tau; \boldsymbol{\eta}, 0)$. The displacement u_i can then be described by using the equivalent body force and Green's functions as follows ("Appendix A: Green's Functions"):

$$u_i(\boldsymbol{x}, t) = \int_{-\infty}^{\infty} \iiint_V f_j^S(\boldsymbol{\eta}, \tau) G_{ij}(\boldsymbol{x}, t - \tau; \boldsymbol{\eta}, 0) \mathrm{d}V(\boldsymbol{\eta}) \mathrm{d}\tau . \tag{6}$$

BM expanded f_j^S in terms of polynomial moments at a particular point $\boldsymbol{\xi_0}$ as

$$\begin{aligned} f_j^S(\boldsymbol{\eta}, \tau) &= F_j^{S0}(\boldsymbol{\xi_0}, \tau) \delta(\boldsymbol{\eta} - \boldsymbol{\xi_0}) \\ &\quad + F_{jk}^{S1}(\boldsymbol{\xi_0}, \tau) \frac{\partial \delta(\boldsymbol{\eta} - \boldsymbol{\xi_0})}{\partial \eta_k} + \cdots \\ &= f_j^{S0}(\boldsymbol{\eta}, \tau) + f_j^{S1}(\boldsymbol{\eta}, \tau) + \cdots . \end{aligned} \tag{7}$$

This expansion is schematically shown in Fig. 2, which may help the reader to understand that F_j^{S0} and F_{jk}^{S1} have the dimensions of force and moment, respectively.

Single Force

F_j^{S0} is given as

$$F_j^{S0}(\boldsymbol{\xi_0}, \tau) = \iiint_V f_j^S(\boldsymbol{\eta}, \tau) \mathrm{d}V(\boldsymbol{\eta}) , \tag{8}$$

which is the total force. This must be zero because of the conservation of linear momentum (no force appears or

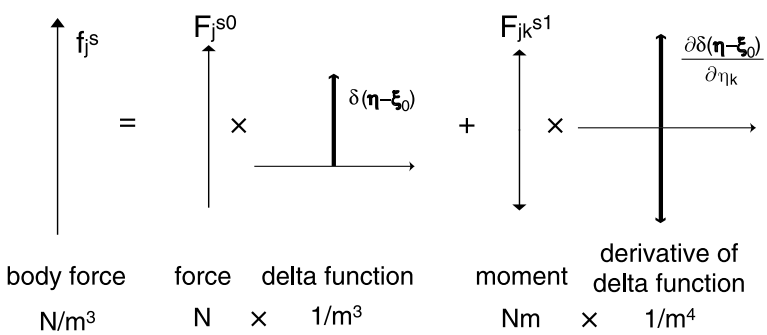

Volcano Seismic Signals, Source Quantification of, Figure 2
Schematic diagram of the expansion of the equivalent body force f_j^S in terms of polynomial moments

disappears in the earth). Therefore,

$$\iiint_V f_j^S(\boldsymbol{\eta},\tau)\mathrm{d}V(\boldsymbol{\eta}) = 0\,. \tag{9}$$

This states that no single force exists in any indigenous seismic sources in the scheme presented by BM. However, we will see later that a single force exists in indigenous sources if the non-linear effect of mass advection is taken into account.

Moment Tensor

We next examine the equivalent body force arising from the first-order moment,

$$f_j^{S1}(\boldsymbol{\eta},\tau) = F_{jk}^{S1}(\boldsymbol{\xi_0},\tau)\frac{\partial\delta(\boldsymbol{\eta}-\boldsymbol{\xi_0})}{\partial\eta_k}\,. \tag{10}$$

If we consider a displacement discontinuity across a surface Σ, f_j^{S1} may be expressed as [1]

$$f_j^{S1}(\boldsymbol{\eta},\tau) = -\iint_\Sigma m_{jk}(\boldsymbol{\xi},\tau)\frac{\partial}{\partial\eta_k}\delta(\boldsymbol{\eta}-\boldsymbol{\xi})\,\mathrm{d}\Sigma(\boldsymbol{\xi})\,, \tag{11}$$

where m_{jk} is the moment density tensor given as

$$m_{jk} = \lambda\nu_l[u_l]\delta_{jk} + \mu(\nu_j[u_k] + \nu_k[u_j])\,. \tag{12}$$

Here, $[\boldsymbol{u}]$ is the displacement discontinuity defined by $[\boldsymbol{u}] = \boldsymbol{u}|_{\Sigma_+} - \boldsymbol{u}|_{\Sigma_-}$ and $\boldsymbol{\nu} = (\nu_1,\nu_2,\nu_3)$ is the unit vector normal to Σ, where $\boldsymbol{u}|_{\Sigma_+}$ and $\boldsymbol{u}|_{\Sigma_-}$ are displacements on the Σ_+ and the Σ_- sides of Σ, respectively (see Fig. 3). An explicit expression of m_{jk} is

$$\boldsymbol{m} = \begin{pmatrix} \lambda\nu_k[u_k] + 2\mu\nu_1[u_1] & \mu(\nu_1[u_2]+\nu_2[u_1]) & \mu(\nu_1[u_3]+\nu_3[u_1]) \\ \mu(\nu_2[u_1]+\nu_1[u_2]) & \lambda\nu_k[u_k] + 2\mu\nu_2[u_2] & \mu(\nu_2[u_3]+\nu_3[u_2]) \\ \mu(\nu_3[u_1]+\nu_1[u_3]) & \mu(\nu_3[u_2]+\nu_2[u_3]) & \lambda\nu_k[u_k] + 2\mu\nu_3[u_3] \end{pmatrix}. \tag{13}$$

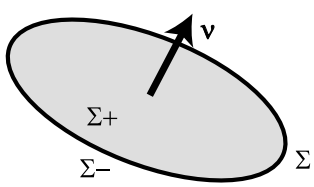

Volcano Seismic Signals, Source Quantification of, Figure 3
An internal surface Σ across which a displacement discontinuity occurs. The displacement discontinuity is denoted by $[u] = u|_{\Sigma_+} - u|_{\Sigma_-}$, where $u|_{\Sigma_+}$ and $u|_{\Sigma_-}$ are displacements on the Σ_+ and Σ_- sides of Σ, respectively [1]

Therefore $m_{jk} = m_{kj}$. Using Green's functions, the displacement field due to f_j^{S1} is expressed as

$$\begin{aligned} u_i^{S1}(\boldsymbol{x},t) &= \int_{-\infty}^{\infty}\iiint_V f_j^{S1}(\boldsymbol{\eta},\tau)G_{ij}(\boldsymbol{x},t-\tau;\boldsymbol{\eta},0)\mathrm{d}V(\boldsymbol{\eta})\,\mathrm{d}\tau \\ &= \int_{-\infty}^{\infty}\iiint_V\left[-\iint_\Sigma m_{jk}(\boldsymbol{\xi},\tau)\frac{\partial}{\partial\eta_k}\delta(\boldsymbol{\eta}-\boldsymbol{\xi})\,\mathrm{d}\Sigma(\boldsymbol{\xi})\right] \\ &\quad\cdot G_{ij}(\boldsymbol{x},t-\tau;\boldsymbol{\eta},0)\mathrm{d}V(\boldsymbol{\eta})\,\mathrm{d}\tau \\ &= \int_{-\infty}^{\infty}\iint_\Sigma\left[m_{jk}(\boldsymbol{\xi},\tau)\iiint_V -\frac{\partial}{\partial\eta_k}\delta(\boldsymbol{\eta}-\boldsymbol{\xi})G_{ij}(\boldsymbol{x},t\right. \\ &\quad\left.-\tau;\boldsymbol{\eta},0)\mathrm{d}V(\boldsymbol{\eta})\right]\mathrm{d}\Sigma(\boldsymbol{\xi})\,\mathrm{d}\tau \\ &= \int_{-\infty}^{\infty}\iint_\Sigma m_{jk}(\boldsymbol{\xi},\tau)\frac{\partial}{\partial\xi_k}G_{ij}(\boldsymbol{x},t-\tau;\boldsymbol{\xi},0)\,\mathrm{d}\Sigma(\boldsymbol{\xi})\,\mathrm{d}\tau\,, \end{aligned} \tag{14}$$

where I used the following property:

$$\iiint_V \frac{\partial}{\partial\eta_k}\delta(\boldsymbol{\eta}-\boldsymbol{\xi})G_{ij}(\boldsymbol{x},t-\tau;\boldsymbol{\eta},0)\mathrm{d}V(\boldsymbol{\eta}) = -\frac{\partial}{\partial\xi_k}G_{ij}(\boldsymbol{x},t-\tau;\boldsymbol{\xi},0)\,. \tag{15}$$

We assume that the source is small compared to the wavelengths of seismic waves and approximate the source as a point at $\boldsymbol{\xi_0}$. Then, Eq. (14) can be written as

$$u_i^{S1}(\boldsymbol{x},t) = \int_{-\infty}^{\infty} M_{jk}(\boldsymbol{\xi_0},\tau)\frac{\partial}{\partial\xi_k}G_{ij}(\boldsymbol{x},t-\tau;\boldsymbol{\xi_0},0)\,\mathrm{d}\tau\,, \tag{16}$$

where M_{jk} is the moment tensor defined as

$$M_{jk}(\boldsymbol{\xi_0},\tau) = \iint_\Sigma m_{jk}(\boldsymbol{\xi},\tau)\,\mathrm{d}\Sigma(\boldsymbol{\xi})\,. \tag{17}$$

Slip on a Fault We consider slip $[u_1]$ in the plane $\xi_3 = 0$ (see Fig. 4). In this case, $\boldsymbol{\nu} = (0,0,1)$ and $[\boldsymbol{u}] = ([u_1],0,0)$. Substituting $\boldsymbol{\nu}$ and $[\boldsymbol{u}]$ into Eq. (13), we obtain the moment density tensor m_{jk} as

$$\boldsymbol{m} = \begin{pmatrix} 0 & 0 & \mu[u_1(\boldsymbol{\xi},\tau)] \\ 0 & 0 & 0 \\ \mu[u_1(\boldsymbol{\xi},\tau)] & 0 & 0 \end{pmatrix}, \tag{18}$$

and the moment tensor M_{jk} as

$$\boldsymbol{M} = M(\boldsymbol{\xi_0},\tau)\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \tag{19}$$

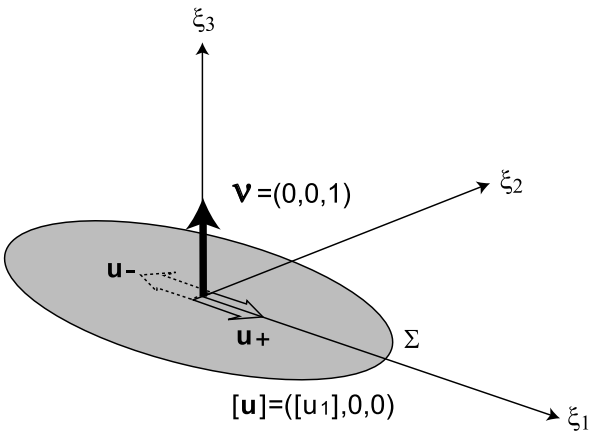

Volcano Seismic Signals, Source Quantification of, Figure 4
Slip $[u_1]$ in the plane $\xi_3 = 0$

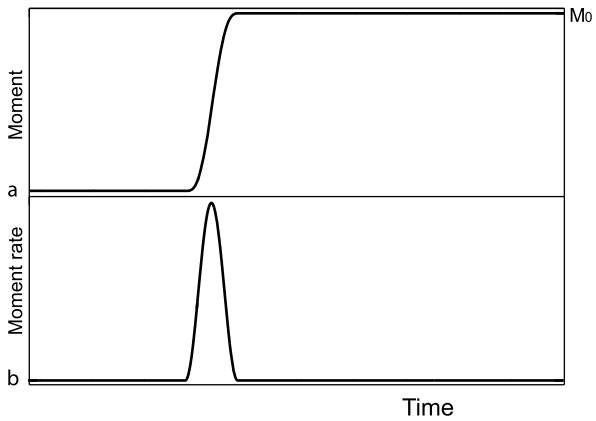

Volcano Seismic Signals, Source Quantification of, Figure 5
Moment and moment-rate functions for slip across a plane in the form of a step-like function

where

$$M(\boldsymbol{\xi}_0, \tau) = \mu \iint_{\Sigma} [u_1(\boldsymbol{\xi}, \tau)]\, \mathrm{d}\Sigma(\boldsymbol{\xi})\,. \tag{20}$$

This is called the moment function or source-time function. The time history of slip across a plane is typically represented by a step-like function, and therefore $M(\boldsymbol{\xi}_0, \tau)$ displays a step-like function (Fig. 5a). We can also define the time derivative of M as

$$\dot{M}(\boldsymbol{\xi}_0, \tau) = \mu \frac{\partial}{\partial \tau} \iint_{\Sigma} [u_1(\boldsymbol{\xi}, \tau)]\, \mathrm{d}\Sigma(\boldsymbol{\xi})\,. \tag{21}$$

This is called the moment-rate function, which is an impulsive function (Fig. 5b). After slip stops, M has a finite value M_0 (Fig. 5a), which is given as

$$M_0 = \mu A \bar{u_1}\,. \tag{22}$$

A and $\bar{u_1}$ are the slip area and average slip, respectively, which are defined as

$$A = \iint_{\Sigma} \mathrm{d}\Sigma\,, \tag{23}$$

$$\bar{u_1} = \frac{\iint_{\Sigma} [u_1(\boldsymbol{\xi}, \tau = \infty)]\, \mathrm{d}\Sigma}{A}\,, \tag{24}$$

where $\tau = \infty$ represents time after the termination of slip. M_0 is called the seismic moment. When $\tau = \infty$, M_{jk} becomes

$$\boldsymbol{M} = M_0 \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}\,. \tag{25}$$

This matrix can be diagonalized to

$$\boldsymbol{M}' = M_0 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}\,, \tag{26}$$

where three eigenvalues (1, 0, and -1) correspond to the T (maximum compression), B (null), and P (minimum compression) axes, respectively. As shown in Fig. 6, this system constitutes a double couple, which is equivalent to fault slip.

Tensile Crack We examine a crack such that slip $[u_3]$ is nonzero in the plane $\xi_3 = 0$ (a horizontal crack; see Fig. 7). In this case, $\boldsymbol{\nu} = (0, 0, 1)$ and $[\boldsymbol{u}] = (0, 0, [u_3])$. Therefore, we find from Eq. (13) that

$$\boldsymbol{m} = \begin{pmatrix} \lambda[u_3(\boldsymbol{\xi}, \tau)] & 0 & 0 \\ 0 & \lambda[u_3(\boldsymbol{\xi}, \tau)] & 0 \\ 0 & 0 & (\lambda + 2\mu)[u_3(\boldsymbol{\xi}, \tau)] \end{pmatrix}\,. \tag{27}$$

The moment tensor for a horizontal crack is given as

$$\boldsymbol{M} = M(\boldsymbol{\xi}_0, \tau) \begin{pmatrix} \lambda/\mu & 0 & 0 \\ 0 & \lambda/\mu & 0 \\ 0 & 0 & \lambda/\mu + 2 \end{pmatrix}\,, \tag{28}$$

and

$$\begin{aligned} M(\boldsymbol{\xi}_0, \tau) &= \mu \iint_{\Sigma} [u_3(\boldsymbol{\xi}, \tau)]\, \mathrm{d}\Sigma(\boldsymbol{\xi}) \\ &= \mu A \bar{u_3}(\tau) \\ &= \mu \Delta V(\tau)\,, \end{aligned} \tag{29}$$

where ΔV represents the incremental volume change of the crack.

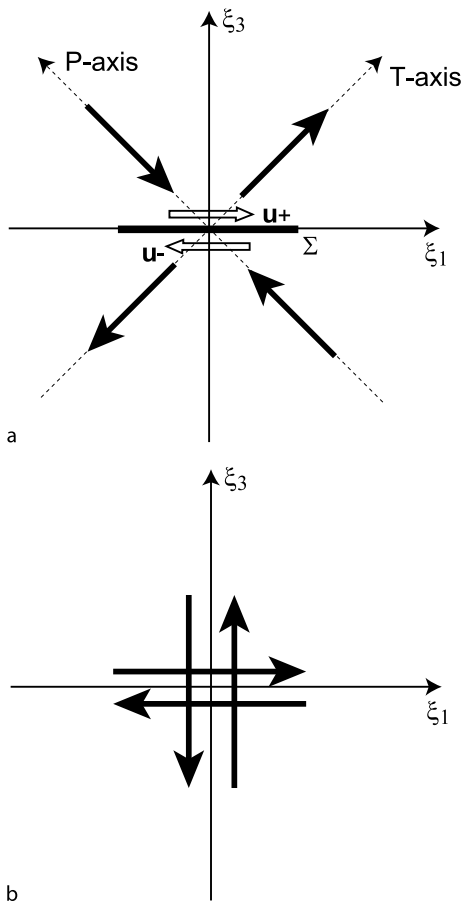

Volcano Seismic Signals, Source Quantification of, Figure 6
a Force system for slip $[u_1]$ in the plane $\xi_3 = 0$. b A double couple equivalent to the force system in Fig. 6a

Figure 8a is a tensile crack with normal direction (θ, ϕ), and we find that $\boldsymbol{\nu} = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$ and $[\boldsymbol{u}] = [u](\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$. Using Eqs. (13) and (17), we obtain the moment tensor for the crack:

$$\boldsymbol{M} = \mu\Delta V \begin{pmatrix} \lambda/\mu + 2\sin^2\theta\cos^2\phi & 2\sin^2\theta\sin\phi\cos\phi & 2\sin\theta\cos\theta\cos\phi \\ 2\sin^2\theta\sin\phi\cos\phi & \lambda/\mu + 2\sin^2\theta\sin^2\phi & 2\sin\theta\cos\theta\sin\phi \\ 2\sin\theta\cos\theta\cos\phi & 2\sin\theta\cos\theta\sin\phi & \lambda/\mu + 2\cos^2\theta \end{pmatrix}, \quad (30)$$

where $\Delta V = \iint_\Sigma [u]\,\mathrm{d}\Sigma$.

For a vertical crack ($\theta = \pi/2$),

$$\boldsymbol{M} = \mu\Delta V \begin{pmatrix} \lambda/\mu + 2\cos^2\phi & 2\sin\phi\cos\phi & 0 \\ 2\sin\phi\cos\phi & \lambda/\mu + 2\sin^2\phi & 0 \\ 0 & 0 & \lambda/\mu \end{pmatrix}. \quad (31)$$

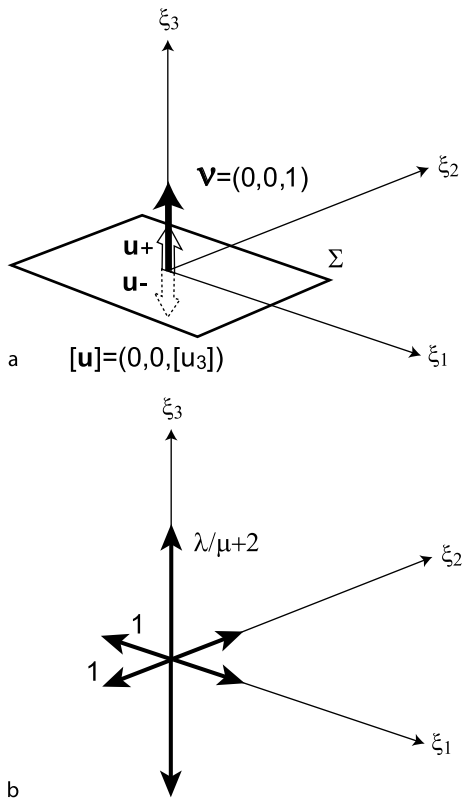

Volcano Seismic Signals, Source Quantification of, Figure 7
a A tensile horizontal crack in which slip $[u_3]$ is nonzero in the plane ξ_3. b Force system equivalent to a tensile horizontal crack

A tensile crack is equivalent to a superposition of three vector dipoles with amplitude ratios $1{:}1{:}(\lambda + 2\mu)/\lambda$ (Fig. 7). If $\lambda = \mu$, the amplitude ratios are represented by 1:1:3. If $\lambda = 2\mu$, which may be appropriate either for volcanic rock near liquidus temperature or for a highly heterogeneous medium, the ratios become 1:1:2.

Spherical Source The moment tensor for a spherical source is given as

$$\boldsymbol{M} = (\lambda + 2\mu)\Delta V_s \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad (32)$$

where $\Delta V_s = 4\pi R^2 \Delta_s$ is the volume change caused by the displacement Δ_s of a spherical surface of radius R [89] (see "Appendix B: Moment Tensor for a Spherical Source"). The moment tensor for a spherical source may be useful to describe the source of explosion earthquakes e. g., [56].

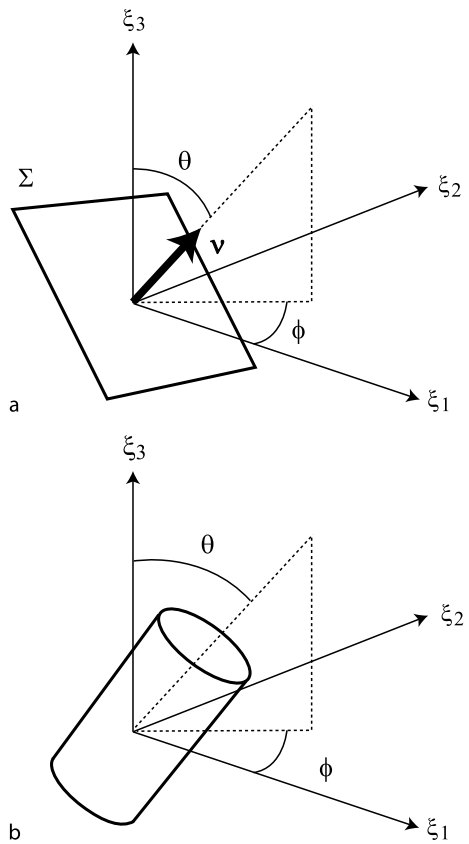

Volcano Seismic Signals, Source Quantification of, Figure 8
Source coordinates and geometries for a a crack and b a cylinder

Cylindrical Source The moment tensor for a vertical cylinder is

$$\boldsymbol{M} = \frac{\lambda + 2\mu}{\lambda + \mu}\mu\Delta V_c \begin{pmatrix} \lambda/\mu + 1 & 0 & 0 \\ 0 & \lambda/\mu + 1 & 0 \\ 0 & 0 & \lambda/\mu \end{pmatrix}, \tag{33}$$

where $\Delta V_c = 2\pi RL\Delta_c$ is the volume change caused by the radial displacement Δ_c of the surface of a cylinder of radius R and length L [89,126] (see "Appendix C: Moment Tensor for a Cylindrical Source").

The moment tensor for a cylinder with axis orientation angles (θ, ϕ) (Fig. 8b) is given as

$$\boldsymbol{M} = \frac{\lambda + 2\mu}{\lambda + \mu}\mu\Delta V_c \times \begin{pmatrix} \lambda/\mu + (\cos^2\theta\cos^2\phi + \sin^2\phi) & \sin^2\theta\sin\phi\cos\phi & \sin\theta\cos\theta\cos\phi \\ \sin^2\theta\sin\phi\cos\phi & \lambda/\mu + (\cos^2\theta\sin^2\phi + \cos^2\phi) & \sin\theta\cos\theta\sin\phi \\ \sin\theta\cos\theta\cos\phi & \sin\theta\cos\theta\sin\phi & \lambda/\mu + \sin^2\theta \end{pmatrix} \tag{34}$$

(see "Appendix C: Moment Tensor for a Cylindrical Source"). Radial volumetric changes of a pipe may occur in association with mass transport in a cylindrical conduit e. g., [126].

Inertial Glut

In the scheme presented by BM, no single force exists in indigenous seismic sources because of the requirement for conservation of linear momentum. However, Takei and Kumazawa [118] presented a theoretical justification for the existence of a single force in indigenous sources, if the non-linear effect of mass advection is taken into account.

Following the theory of Takei and Kumazawa [118], we consider the exact equation of motion of the earth, given as

$$\rho^{\text{true}}\frac{D^2u_i}{Dt^2} = \frac{\partial\sigma_{ni}^{\text{true}}}{\partial x_n}, \tag{35}$$

where ρ^{true} is true density as a function of both space and time. $\frac{D}{Dt}$ is the particle derivative:

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + v_k\frac{\partial}{\partial x_k}, \tag{36}$$

where $v_k = \frac{Du_k}{Dt}$ is particle velocity. Using model density ρ^{model} and model stress $\sigma_{ni}^{\text{model}}$ from Eq. (2), Eq. (35) can be rewritten as follows:

$$\rho^{\text{model}}\frac{\partial^2u_i}{\partial t^2} = \frac{\partial\sigma_{ni}^{\text{model}}}{\partial x_n} + f_i^V. \tag{37}$$

This equation is similar to Eq. (3), but the equivalent body force f_i^V is defined in this case as

$$\begin{aligned} f_i^V &= \left(\frac{\partial\sigma_{ni}^{\text{true}}}{\partial x_n} - \frac{\partial\sigma_{ni}^{\text{model}}}{\partial x_n}\right) \\ &\quad + \left(\rho^{\text{model}}\frac{\partial^2u_i}{\partial t^2} - \rho^{\text{true}}\frac{D^2u_i}{Dt^2}\right) \\ &= f_i^S + f_i^I. \end{aligned} \tag{38}$$

Here, f_i^V is f_i^S, arising from the stress glut, plus f_i^I, representing the difference between the inertia terms, which is called the inertial glut [118].

In the same way as presented by BM, we expand f_i^I in terms of polynomial moments at a particular point $\boldsymbol{\zeta_0}$ as

$$f_i^I(\boldsymbol{\eta}, \tau) = F_i^{I0}(\boldsymbol{\zeta_0}, \tau)\delta(\boldsymbol{\eta} - \boldsymbol{\zeta_0}) + \text{higher-order terms}, \tag{39}$$

where

$$\begin{aligned} F_i^{I0} &= \iiint_V f_i^I(\boldsymbol{\eta}, \tau)\mathrm{d}V(\boldsymbol{\eta}) \\ &= -\iiint_V \left(\rho^{\text{true}}\frac{D^2u_i}{Dt^2} - \rho^{\text{model}}\frac{\partial^2u_i}{\partial t^2}\right)\mathrm{d}V(\boldsymbol{\eta}). \end{aligned} \tag{40}$$

By using mass conservation

$$\frac{\partial \rho^{\text{true}}}{\partial t} + \frac{\partial}{\partial x_k}(\rho^{\text{true}} v_k) = 0 \tag{41}$$

and Gauss' theorem

$$\iiint_V \frac{\partial v_k}{\partial x_k} \mathrm{d}V = \iint_{S_e} v_k v_k \mathrm{d}S_e \,, \tag{42}$$

the first integral of Eq. (40) is expressed as

$$\iiint_V \rho^{\text{true}} \frac{D^2 u_i}{Dt^2} \mathrm{d}V = \frac{\partial}{\partial t} \iiint_V \rho^{\text{true}} v_i \mathrm{d}V + \iint_{S_e} (\rho^{\text{true}} v_i v_k) v_k \mathrm{d}S_e \,, \tag{43}$$

where S_e is the surface of the earth. The second term in the right-hand side of Eq. (43) vanishes because it is a surface integral of a non-linear term, which is infinitesimal. Therefore, we find that

$$\iiint_V \rho^{\text{true}} \frac{D^2 u_i}{Dt^2} \mathrm{d}V = \frac{\partial}{\partial t} \iiint_V \rho^{\text{true}} v_i \mathrm{d}V = \dot{L}_i^t \,. \tag{44}$$

Here, $\dot{L}_i^t$ is the time derivative of the total linear momentum, which is zero at any given time because of conservation of linear momentum of the earth. The second integral in Eq. (40) is given as

$$\iiint_V \rho^{\text{model}} \frac{\partial^2 u_i}{\partial t^2} \mathrm{d}V = \frac{\partial}{\partial t} \iiint_V \rho^{\text{model}} \frac{\partial u_i}{\partial t} \mathrm{d}V = \dot{L}_i^m \,, \tag{45}$$

which is the time derivative of linear momentum defined by the model density. Therefore, Eq. (40) is written as

$$F_i^{I0} = -(\dot{L}_i^t - \dot{L}_i^m) = \dot{L}_i^m \,, \tag{46}$$

which shows that force F_i^{I0} exists as an apparent change of the total linear momentum of the earth referred to a biased model density structure.

Takei and Kumazawa [118] stated that F_i^{I0} originates from (a) the difference of density structure of the prescribed model from the actual value in the source region before the event, and from (b) a temporal change of the density structure in the source region caused by finite displacement of mass during the event. The latter is a non-linear effect caused by mass advection, which may occur by fluid flow, especially in volcanic regions.

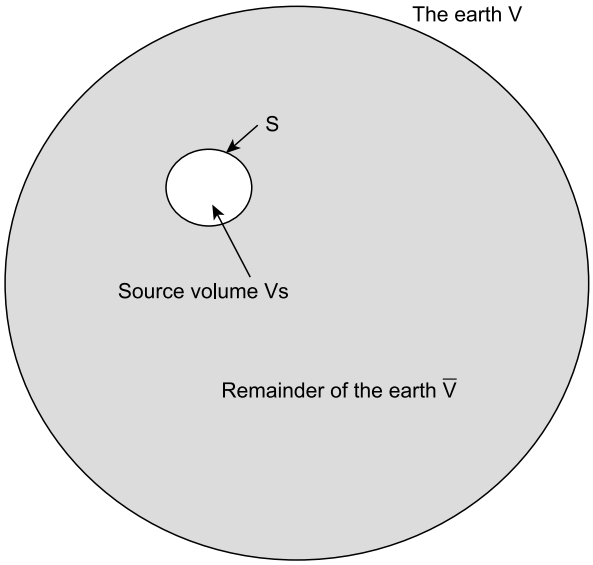

Volcano Seismic Signals, Source Quantification of, Figure 9
The earth V divided into a source volume V_s enclosed by a surface S and the remainder of the earth $\bar{V}$

Single Force Defined by Action and Reaction

We further consider the single force from the viewpoint of the interaction force between the source volume and the rest of the earth. Following Takei and Kumazawa [119], we define the source region V_s enclosed by a surface S and the remaining part of the earth $\bar{V}$ (Fig. 9). Using Eqs. (38), (9), (44), and (45), we obtain

$$\begin{aligned}
\iiint_V f_i^V \mathrm{d}V &= \iiint_V (f_i^S + f_i^I) \mathrm{d}V \\
&= \iiint_V f_i^I \mathrm{d}V \\
&= -\frac{\partial}{\partial t} \iiint_V \left(\rho^{\text{true}} v_i - \rho^{\text{model}} \frac{\partial u_i}{\partial t} \right) \mathrm{d}V \\
&= -\frac{\partial}{\partial t} \left[\iiint_{V_s} \left(\rho^{\text{true}} v_i - \rho^{\text{model}} \frac{\partial u_i}{\partial t} \right) \mathrm{d}V \right. \\
&\quad \left. - \iiint_{\bar{V}} \left(\rho^{\text{true}} v_i - \rho^{\text{model}} \frac{\partial u_i}{\partial t} \right) \mathrm{d}V \right] ,
\end{aligned} \tag{47}$$

where infinitesimal surface integral terms are omitted. The second integral in the right-hand side of Eq. (47) vanishes because $\rho^{\text{true}} = \rho^{\text{model}}$ and $v_i = \partial u_i / \partial t$ in $\bar{V}$. f_i^V exists only in the source region, and thus the volume integral of f_i^V in the left hand side of Eq. (47) can be limited to V_s.

Therefore, Eq. (47) is written as

$$\iiint_{V_s} f_i^V \mathrm{d}V = -\frac{\partial}{\partial t}\iiint_{V_s}\left(\rho^{\mathrm{true}} v_i - \rho^{\mathrm{model}}\frac{\partial u_i}{\partial t}\right)\mathrm{d}V\,. \tag{48}$$

This equation is then rewritten as

$$\begin{aligned}\frac{\partial}{\partial t}\iiint_{V_s}\rho^{\mathrm{true}} v_i \mathrm{d}V &= \frac{\partial}{\partial t}\iiint_{V_s}\rho^{\mathrm{model}}\frac{\partial u_i}{\partial t}\mathrm{d}V \\ &\quad - \iiint_{V_s} f_i^V \mathrm{d}V \\ &= \iiint_{V_s}\left(\rho^{\mathrm{model}}\frac{\partial^2 u_i}{\partial t^2} - f_i^V\right)\mathrm{d}V\,.\end{aligned} \tag{49}$$

From Eq. (37), we find that the integrand of the right-hand side of Eq. (49) equals $\partial\sigma_{ni}^{\mathrm{model}}/\partial x_n$. Therefore, we obtain

$$\begin{aligned}\frac{\partial}{\partial t}\iiint_{V_s}\rho^{\mathrm{true}} v_i \mathrm{d}V &= \iiint_{V_s}\frac{\partial\sigma_{ni}^{\mathrm{model}}}{\partial x_n}\mathrm{d}V \\ &= \iint_S \sigma_{ni}^{\mathrm{model}} \nu_n \mathrm{d}S\,.\end{aligned} \tag{50}$$

This indicates that a temporal change of linear momentum in the source region is equivalent to the total force exerted on the boundary S. Therefore, $\iint_S \sigma_{ni}^{\mathrm{model}}\nu_n \mathrm{d}S$ can be regarded as the action force originating from non-linear processes in the source volume V_s. In $\bar{V}$, the reaction force opposite to $\iint_S \sigma_{ni}^{\mathrm{model}}\nu_n \mathrm{d}S$ acts through the boundary S. We define traction P_i on S, which satisfies the following condition:

$$\sigma_{ni}^{\mathrm{model}}\nu_n = -P_i \quad \text{on } S\,. \tag{51}$$

Then, we find that

$$\begin{aligned}\iint_S P_i \mathrm{d}S &= -\iint_S \sigma_{ni}^{\mathrm{model}}\nu_n \mathrm{d}S \\ &= -\frac{\partial}{\partial t}\iiint_{V_s}\rho^{\mathrm{true}} v_i \mathrm{d}V\,,\end{aligned} \tag{52}$$

which is the reaction force opposite to the action force.

If the source is small compared to the wavelengths of seismic waves, the action force originating in a closed small source volume V_s may not be observable outside the source volume $\bar{V}$. Instead, the traction P_j exerted on S may be viewed in $\bar{V}$ as a traction discontinuity on S. As shown by Aki and Richards [1], a traction discontinuity is the source of seismic waves represented by an equivalent body force. We consider seismic waves in $\bar{V}$, and the equation of motion in $\bar{V}$ may be expressed as

$$\rho^{\mathrm{model}}\frac{\partial^2 u_i}{\partial t^2} = \frac{\partial\sigma_{ni}^{\mathrm{model}}}{\partial x_n} + f_i \quad \text{in } \bar{V}\,, \tag{53}$$

where f_i is the equivalent body force due to the traction discontinuity on S, given as

$$f_i(\boldsymbol{\eta},\tau) = \iint_S P_i(\boldsymbol{\zeta},\tau)\delta(\boldsymbol{\eta}-\boldsymbol{\zeta})\mathrm{d}S(\boldsymbol{\zeta})\,. \tag{54}$$

Note that

$$\iiint_{\bar{V}} f_i(\boldsymbol{\eta},\tau)\mathrm{d}V(\boldsymbol{\eta}) = \iint_S P_i(\boldsymbol{\zeta},\tau)\mathrm{d}S(\boldsymbol{\zeta})\,, \tag{55}$$

which is the total force acting through S in $\bar{V}$.

We may use Green's functions estimated from the model structure because the effect of the true structure in the source volume is negligible in the point source approximation. Therefore, by using Eq. (54), the displacement field u_i^f due to the single force may be described as

$$\begin{aligned}u_i^f(\boldsymbol{x},t) &= \int_{-\infty}^{\infty}\iiint_{\bar{V}} f_j(\boldsymbol{\eta},\tau)G_{ij}(\boldsymbol{x},t-\tau;\boldsymbol{\eta},0)\mathrm{d}V(\boldsymbol{\eta})\,\mathrm{d}\tau \\ &= \int_{-\infty}^{\infty}\iiint_{\bar{V}}\left[\iint_S P_j(\boldsymbol{\zeta},\tau)\delta(\boldsymbol{\eta}-\boldsymbol{\zeta})\mathrm{d}S(\boldsymbol{\zeta})\right] \\ &\qquad \cdot G_{ij}(\boldsymbol{x},t-\tau;\boldsymbol{\eta},0)\mathrm{d}V(\boldsymbol{\eta})\,\mathrm{d}\tau \\ &= \int_{-\infty}^{\infty}\iint_S P_j(\boldsymbol{\zeta},\tau)G_{ij}(\boldsymbol{x},t-\tau;\boldsymbol{\zeta},0)\mathrm{d}S(\boldsymbol{\zeta})\,\mathrm{d}\tau\,,\end{aligned} \tag{56}$$

or, equivalently,

$$u_i^f(\boldsymbol{x},t) = \int_{-\infty}^{\infty} F_j(\boldsymbol{\zeta_0},\tau)G_{ij}(\boldsymbol{x},t-\tau;\boldsymbol{\zeta_0},0)\,\mathrm{d}\tau\,, \tag{57}$$

where

$$\begin{aligned}F_j(\boldsymbol{\zeta_0},\tau) &= \iint_S P_j(\boldsymbol{\zeta},\tau)\mathrm{d}S(\boldsymbol{\zeta}) \\ &= -\frac{\partial}{\partial t}\iiint_{V_s}\rho^{\mathrm{true}} v_i \mathrm{d}V \\ &= -\dot{L}_j^s\,.\end{aligned} \tag{58}$$

Here, L_j^s is the total linear momentum inside the source volume. This single force representation in terms of the traction P_j is consistent with the relationship between the traction modes and equivalent body force given by Takei and Kumazawa [119] in small-source and long-wavelength approximations.

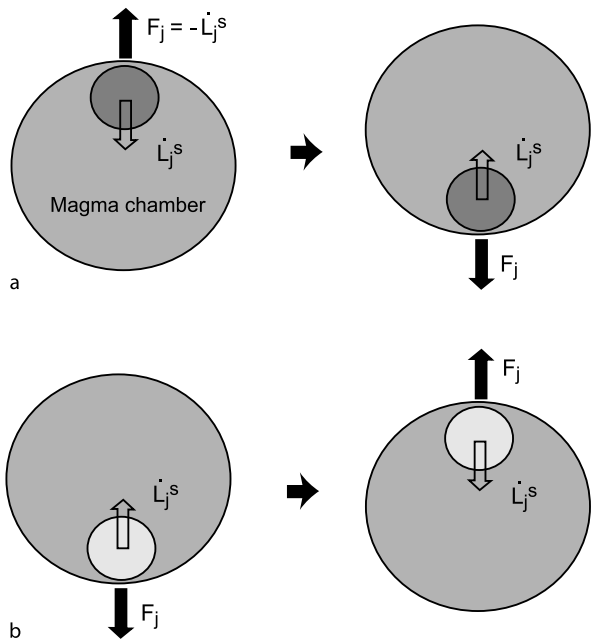

Volcano Seismic Signals, Source Quantification of, Figure 10
Schematic illustration of the reaction force (F_j) opposite to the action force ($\dot{L}_j^s$) generated by nonlinear processes in a source volume: a dense materials falling in a magma chamber and b light materials ascending in a magma chamber

F_j is the reaction force opposite to the action force generated by non-linear processes in the source volume. A simple example is dense materials falling in a magma chamber (Fig. 10a). During the acceleration stage of falling dense materials, $\dot{L}_j^s$ is a downward force and we therefore observe an upward single force as the reaction force; during the deceleration stage, we observe a downward single force (Fig. 10a). In the case of lighter materials ascending in a magma chamber, we observe a downward single force during the acceleration stage of ascending lighter materials, and an upward single force is observed during the deceleration stage of ascending lighter materials (Fig. 10b).

Summary

Here, we summarize the phenomenological representation of indigenous seismic sources. For point sources, the displacement field u_i can be expressed by the superposition of the displacement fields excited by single force u_i^f of Eq. (57) and moment tensor u_i^{S1} of Eq. (16):

$$u_i(\boldsymbol{x}, t) = \int_{-\infty}^{\infty} F_j(\boldsymbol{\zeta_0}, \tau) G_{ij}(\boldsymbol{x}, t-\tau; \boldsymbol{\zeta_0}, 0)\,\mathrm{d}\tau + \int_{-\infty}^{\infty} M_{jk}(\boldsymbol{\xi_0}, \tau) \frac{\partial}{\partial \xi_k} G_{ij}(\boldsymbol{x}, t-\tau; \boldsymbol{\xi_0}, 0)\,\mathrm{d}\tau\,, \quad (59)$$

where

$$F_j(\boldsymbol{\zeta_0}, \tau) = \iint_S P_j(\boldsymbol{\zeta}, \tau)\mathrm{d}S(\boldsymbol{\zeta}) \quad (60)$$

and

$$M_{jk}(\boldsymbol{\xi_0}, \tau) = \iint_\Sigma m_{jk}(\boldsymbol{\xi}, \tau)\,\mathrm{d}\Sigma(\boldsymbol{\xi})\,. \quad (61)$$

Here, P_j is the traction exerted on the boundary S between the source volume and the rest of the earth and m_{jk} is the moment density tensor defined by the displacement discontinuity across the surface Σ. F_j is related to the total linear momentum within the source volume L_j^s as

$$F_j = -\dot{L}_j^s \quad (j = 1, 2, 3)\,. \quad (62)$$

M_{jk} is the moment tensor consisting of six independent components:

$$\boldsymbol{M} = \begin{pmatrix} M_{11} & M_{12} & M_{13} \\ M_{12} & M_{22} & M_{23} \\ M_{13} & M_{23} & M_{33} \end{pmatrix}. \quad (63)$$

Waveform Inversion

Waveform inversion is a quantitative tool used to estimate the source mechanisms and locations of volcano-seismic signals. The source parameters are determined by finding the best fits between observed and synthetic seismic waveforms. This can be performed by solving an inverse problem by using the concept of linear inverse theory e. g., [83].

We may reasonably assume that the source of a tectonic earthquake is represented by the moment tensor corresponding to a double couple. However, volcano-seismic signals may be generated by complex source processes, requiring a general source representation consisting of single force and moment tensor. Furthermore, while we can assume a step-like function for the source-time function of a tectonic earthquake, we must determine the complex time history at the source of volcano-seismic signals, which involves more unknowns. Waveform inversion of volcano-seismic signals is thus a more complicated task than the inversion of a tectonic earthquake.

We assume a single force and moment tensor at a particular source location, and obtain from Eq. (59) the following relationship:

$$u_i(t) = \int_{-\infty}^{\infty} [F_j(\tau) G_{ij}(t-\tau) + M_{jk}(\tau) G_{ij,k}(t-\tau)]\,\mathrm{d}\tau\,, \quad (64)$$

where $G_{ij,k} = \partial G_{ij}/\partial \xi_k$. We rewrite Eq. (64) in the following form:

$$u_i(t) = \sum_{j=1}^{9} \int_{-\infty}^{\infty} m_j(\tau)\Gamma_{ij}(t-\tau)\,\mathrm{d}\tau\,, \tag{65}$$

where

$$m_j = [F_1, F_2, F_3, M_{11}, M_{22}, M_{33}, M_{12}, M_{23}, M_{13}] \tag{66}$$

and

$$\Gamma_{ij} = [G_{i1}, G_{i2}, G_{i3}, G_{i1,1}, G_{i2,2}, G_{i3,3}, G_{i1,2} + G_{i2,1}, G_{i2,3} + G_{i3,2}, G_{i1,3} + G_{i3,1}]\,. \tag{67}$$

We discretize t and τ as $t = k\Delta t$ $(k = 0, \ldots, N_t - 1)$ and $\tau = l\Delta t$ $(l = 0, \ldots, N_t - 1)$, respectively, where Δt is the sampling interval. This yields the discrete form of Eq. (65) given as

$$u_i(k\Delta t) = \sum_{j=1}^{9} \sum_{l=0}^{N_t-1} m_j(l\Delta t)\Gamma_{ij}(k\Delta t - l\Delta t)\Delta t\,. \tag{68}$$

If we have N_c traces of u_i, Eq. (68) is written in matrix form

$$\boldsymbol{u} = \boldsymbol{\Gamma}\,\boldsymbol{m}\,, \tag{69}$$

where

$$\boldsymbol{u} = [u_1(0), \ldots, u_1((N_t-1)\Delta t), \ldots, u_{N_c}(0), \ldots, u_{N_c}((N_t-1)\Delta t)]^{\mathrm{T}}\,, \tag{70}$$

$$\boldsymbol{m} = [m_1(0), \ldots, m_1((N_t-1)\Delta t), \ldots, m_9(0), \ldots, m_9((N_t-1)\Delta t)]^{\mathrm{T}}\,, \tag{71}$$

$$\boldsymbol{\Gamma} = \begin{bmatrix} \Gamma_{11}(0) & \cdots & \Gamma_{11}((1-N_t)\Delta t) & \cdots & \cdots & \Gamma_{19}(0) & \cdots & \Gamma_{19}((1-N_t)\Delta t) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \Gamma_{11}((N_t-1)\Delta t) & \cdots & \Gamma_{11}(0) & \cdots & \cdots & \Gamma_{19}((N_t-1)\Delta t) & \cdots & \Gamma_{19}(0) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \Gamma_{N_c1}(0) & \cdots & \Gamma_{N_c1}((1-N_t)\Delta t) & \cdots & \cdots & \Gamma_{N_c9}(0) & \cdots & \Gamma_{N_c9}((1-N_t)\Delta t) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \Gamma_{N_c1}((N_t-1)\Delta t) & \cdots & \Gamma_{N_c1}(0) & \cdots & \cdots & \Gamma_{N_c9}((N_t-1)\Delta t) & \cdots & \Gamma_{N_c9}(0) \end{bmatrix} \Delta t\,. \tag{72}$$

Here T denotes transpose. $\boldsymbol{\Gamma}$ is called the data kernel, which is an $N \times M$ matrix, and $\boldsymbol{u}$ and $\boldsymbol{m}$ are vectors of lengths N and M, respectively, where $N = N_c N_t$ and $M = 9N_t$.

We denote observed waveforms as $\boldsymbol{u}^{\mathrm{obs}}$, which satisfy the following equation:

$$\boldsymbol{u}^{\mathrm{obs}} - \boldsymbol{\Gamma}\,\boldsymbol{m} = \boldsymbol{e}\,, \tag{73}$$

where $\boldsymbol{e}$ is the prediction error or misfit. In the case of $N > M$, we estimate $\boldsymbol{m}$ by minimizing the residual R, defined as

$$R = \frac{\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}}{\boldsymbol{u}^{\mathrm{obs}\,\mathrm{T}}\boldsymbol{u}^{\mathrm{obs}}} = \frac{\sum_{i=1}^{N} e_i^2}{\sum_{i=1}^{N} u_i^{\mathrm{obs}}}\,, \tag{74}$$

where

$$\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} = (\boldsymbol{u}^{\mathrm{obs}} - \boldsymbol{\Gamma}\,\boldsymbol{m})^{\mathrm{T}}(\boldsymbol{u}^{\mathrm{obs}} - \boldsymbol{\Gamma}\,\boldsymbol{m})\,. \tag{75}$$

To minimize R is to find values of $\boldsymbol{m}$ that satisfy the following equations:

$$\frac{\partial R}{\partial m_q} = 0 \quad (q = 1, \ldots, M)\,. \tag{76}$$

This leads to the following relationship (the normal equation)

$$\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Gamma}\,\boldsymbol{m} - \boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{u}^{\mathrm{obs}} = 0\,. \tag{77}$$

Therefore, we have the following solution:

$$\boldsymbol{m}^{\mathrm{est}} = [\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Gamma}]^{-1}\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{u}^{\mathrm{obs}}, \tag{78}$$

which is the least-squares solution. This least-squares approach in the time domain has been used in many volcano seismological studies e. g., [32,33,48,69,71,72,73,92,94,102,105,107,108,120,122].

As can be seen from Eq. (68), u_i is calculated by the convolution of m_j with Γ_{ij}. This means that the displacement $u_i(k\Delta t)$ includes the entire effect of the time history of the single force and moment tensor as well as Green's functions before time $k\Delta t$. This may be easily understood if we consider $u_i((N_t-1)\Delta t)$ in Eq. (68):

$$\begin{aligned} u_i((N_t-1)\Delta t) = \Delta t \sum_{j=1}^{9} [& m_j(0)\Gamma_{ij}((N_t-1)\Delta t) \\ & + m_j(\Delta t)\Gamma_{ij}((N_t-2)\Delta t) + \cdots \\ & + m_j((N_t-2)\Delta t)\Gamma_{ij}(\Delta t) \\ & + m_j((N_t-1)\Delta t)\Gamma_{ij}(0)]\,. \end{aligned} \tag{79}$$

The convolution relationship leads to the large matrix of $\boldsymbol{\Gamma}$ of Eq. (69), which requires long computation time to obtain the least-squares solution. A better alternative is to solve the inverse problem in the frequency domain [11,95, 116]. The Fourier transform of Eq. (65) is given by

$$\bar{u}_i(\omega) = \sum_{j=1}^{9} \bar{m}_j(\omega)\bar{\Gamma}_{ij}(\omega)\,, \tag{80}$$

where ω is the angular frequency and $\bar{u}_i$, $\bar{m}_j$, and $\bar{\Gamma}_{ij}$ are the Fourier transforms of u_i, m_j, and Γ_{ij}. This frequency domain equation is discretized as follows:

$$\bar{u}_i(k\Delta\omega) = \sum_{j=1}^{9} \bar{m}_j(k\Delta\omega)\bar{\Gamma}_{ij}(k\Delta\omega)\,, \tag{81}$$

where $\Delta\omega = 2\pi/(N_t\Delta t)$ and $k = 0, \ldots, N_t/2$. Note that the Fourier transformed components are complex and we have a set of the equations corresponding to real and imaginary parts; therefore, Eq. (81) consists of N_t equations. We further note that $\bar{u}_i$, $\bar{m}_j$, and $\bar{\Gamma}_{ij}$ in Eq. (81) depend only on $k\Delta\omega$. This indicates that individual frequency components of $\bar{u}_i$, $\bar{m}_j$, and $\bar{\Gamma}_{ij}$ are independent of each other and Eq. (81) can be solved separately. This is different from the time-domain Eq. (65): u_i is a function of $k\Delta t$, whereas m_j is a function of $l\Delta t$, and Γ_{ij} is a function of $k\Delta t$ and $l\Delta t$, so we cannot separate the time-domain equation in individual values of $k\Delta t$. The use of the frequency-domain equation results in a smaller size of the data kernel, as shown below.

If we have N_c traces of $\bar{u}_i$, Eq. (81) is written in matrix form

$$\bar{\boldsymbol{u}} = \bar{\boldsymbol{\Gamma}}\,\bar{\boldsymbol{m}}\,. \tag{82}$$

Here, $\bar{\boldsymbol{u}}$ is a vector consisting of $\bar{u}_i$, which is arranged in individual frequency components as follows:

$$\begin{aligned}\bar{\boldsymbol{u}} = [&\bar{u}_1(0), \bar{u}_2(0), \ldots, \bar{u}_{Nc}(0), \bar{u}_1(\Delta\omega),\\ &\bar{u}_2(\Delta\omega), \ldots, \bar{u}_{Nc}(\Delta\omega), \ldots, \bar{u}_1((N_t/2)\Delta\omega),\\ &\bar{u}_2((N_t/2)\Delta\omega), \ldots, \bar{u}_{N_c}((N_t/2)\Delta\omega)]^{\mathrm{T}}\end{aligned} \tag{83}$$

or, equivalently,

$$\bar{\boldsymbol{u}} = [\bar{\boldsymbol{u}}(0), \bar{\boldsymbol{u}}(\Delta\omega), \ldots, \bar{\boldsymbol{u}}(k\Delta\omega), \ldots, \bar{\boldsymbol{u}}(N_t/2\Delta\omega)]^{\mathrm{T}}, \tag{84}$$

where

$$\bar{\boldsymbol{u}}(k\Delta\omega) = [\bar{u}_1(k\Delta\omega), \bar{u}_2(k\Delta\omega), \ldots, \bar{u}_{Nc}(k\Delta\omega)]^{\mathrm{T}}. \tag{85}$$

Similarly, $\bar{\boldsymbol{m}}$ is a vector consisting of $\bar{m}_j$, which is arranged in individual frequency components as

$$\begin{aligned}\bar{\boldsymbol{m}} = [&\bar{m}_1(0), \bar{m}_2(0), \ldots, \bar{m}_9(0), \bar{m}_1(\Delta\omega),\\ &\bar{m}_2(\Delta\omega), \ldots, \bar{m}_9(\Delta\omega), \ldots, \bar{m}_1((N_t/2)\Delta\omega),\\ &\bar{m}_2((N_t/2)\Delta\omega), \ldots, \bar{m}_9((N_t/2)\Delta\omega)]^{\mathrm{T}}\end{aligned} \tag{86}$$

or

$$\bar{\boldsymbol{m}} = [\bar{\boldsymbol{m}}(0), \bar{\boldsymbol{m}}(\Delta\omega), \ldots, \bar{\boldsymbol{m}}(k\Delta\omega), \ldots, \bar{\boldsymbol{m}}((N_t/2)\Delta\omega)]^{\mathrm{T}}, \tag{87}$$

where

$$\bar{\boldsymbol{m}}(k\Delta\omega) = [\bar{m}_1(k\Delta\omega), \bar{m}_2(k\Delta\omega), \ldots, \bar{m}_9(k\Delta\omega)]^{\mathrm{T}}. \tag{88}$$

Accordingly, $\bar{\boldsymbol{\Gamma}}$ is given as

$$\bar{\boldsymbol{\Gamma}} = \begin{bmatrix} \bar{\boldsymbol{\Gamma}}(0) & & & & & \\ & \bar{\boldsymbol{\Gamma}}(\Delta\omega) & & & 0 & \\ & & \ddots & & & \\ & & & \bar{\boldsymbol{\Gamma}}(k\Delta\omega) & & \\ & 0 & & & \ddots & \\ & & & & & \bar{\boldsymbol{\Gamma}}((N_t/2)\Delta\omega) \end{bmatrix}, \tag{89}$$

where $\bar{\boldsymbol{\Gamma}}(k\Delta\omega)$ is

$$\bar{\boldsymbol{\Gamma}}(k\Delta\omega) = \begin{bmatrix} \bar{\Gamma}_{11}(k\Delta\omega) & \cdots & \bar{\Gamma}_{19}(k\Delta\omega) \\ \vdots & \vdots & \vdots \\ \bar{\Gamma}_{N_c1}(k\Delta\omega) & \cdots & \bar{\Gamma}_{N_c9}(k\Delta\omega) \end{bmatrix}. \tag{90}$$

Since $\bar{\boldsymbol{\Gamma}}$ is the block diagonal matrix, Eq. (82) can be reduced to

$$\bar{\boldsymbol{u}}(k\Delta\omega) = \bar{\boldsymbol{\Gamma}}(k\Delta\omega)\bar{\boldsymbol{m}}(k\Delta\omega) \quad (k = 0, \ldots, N_t/2). \tag{91}$$

The dimension of matrix $\bar{\boldsymbol{\Gamma}}(k\Delta\omega)$ is $N_c \times 9$, which is N_t^2 times smaller than that of $\boldsymbol{\Gamma}$ in the time domain (Eq. (72)). This is because convolution in the time domain is equivalent to the multiplication in the frequency domain. The smaller size of $\bar{\boldsymbol{\Gamma}}(k\Delta\omega)$ contributes to faster computation of the least-squares solution. We denote the $k\Delta\omega$ component of the Fourier transform of observed seismograms as

$$\bar{\boldsymbol{u}}^{\mathrm{obs}}(k\Delta\omega) = \left[\bar{u}_1^{\mathrm{obs}}(k\Delta\omega), \bar{u}_2^{\mathrm{obs}}(k\Delta\omega), \ldots, \bar{u}_{N_c}^{\mathrm{obs}}(k\Delta\omega)\right]^{\mathrm{T}}. \tag{92}$$

The least-squares solution in the frequency domain for the $k\Delta\omega$ component is then given by

$$\bar{\boldsymbol{m}}^{\mathrm{est}}(k\Delta\omega) = \left[\bar{\boldsymbol{\Gamma}}^H(k\Delta\omega)\bar{\boldsymbol{\Gamma}}(k\Delta\omega)\right]^{-1} \cdot \bar{\boldsymbol{\Gamma}}^H(k\Delta\omega)\bar{\boldsymbol{u}}^{\mathrm{obs}}(k\Delta\omega)\,, \tag{93}$$

where H denotes the adjoint (Hermitian conjugate). We perform the inverse Fourier transform of $\bar{\boldsymbol{m}}^{\mathrm{est}}(k\Delta\omega)$ ($k = 0, \ldots, N_t/2$) to obtain the solution in the time domain.

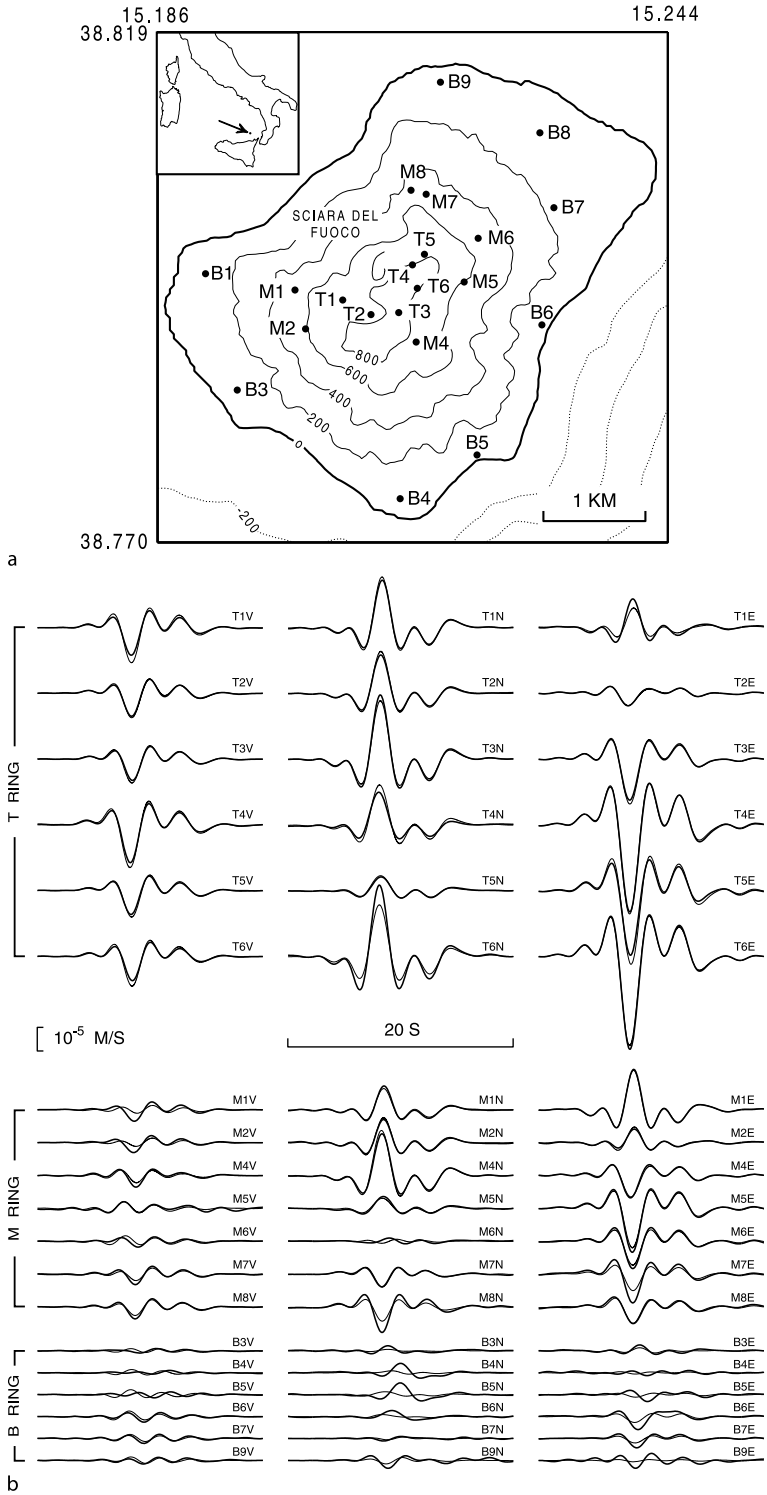

Volcano Seismic Signals, Source Quantification of, Figure 11

a Locations of broadband seismic stations at Stromboli Volcano, Italy. b Observed VLP velocity waveforms (*thick lines*) bandpassed between 2 and 20 s and the best-fit synthetic seismograms (*thin lines*) obtained from waveform inversion assuming six moment tensor and three single force components for the source mechanism [32]

The use of appropriate Green's functions is critically important to obtain the correct solution $\boldsymbol{m}^{\text{est}}$ in either the time or frequency domains. If the source is located in a medium that can be approximated by a homogeneous or layered structure with a flat surface, Green's functions can be calculated by using propagator matrices e. g., [63] and the discrete wavenumber method e. g., [16,17,20]. However, this situation is not common in volcanic regions, and we need to take into account the steep topographies and heterogeneous structures of volcanoes. The boundary integral equation method e. g., [18] and boundary element method e. g., [62] can be used to quantify the effect of three-dimensional topography in a homogeneous or layered structure. To treat both topography and structural heterogeneity simultaneously, the finite-difference method e. g., [106,109] and the discrete lattice method [104] are suitable. Once Green's functions (and their spatial derivatives) are calculated, we can obtain the least-squares solution $\boldsymbol{m}^{\text{est}}$ through Eq. (78) or (93). This solution is, however, obtained for a particular source point and a grid search in space is required to find the best-fit source location.

An example of waveform inversion is provided in Figs. 11 and 12. The source mechanism of VLP signals observed by a dense broadband seismic network at Stromboli volcano, Italy, was estimated by using the waveform inversion method in the time domain [32]. In the inversion, VLP waveforms from 18 stations (Fig. 11) were used, and six moment tensor and three single force components for a point source were assumed. Green's functions were calculated by the finite-difference method [106] using the topography of the volcano. The inversion result (Fig. 12) points to volumetric changes accompanying a dominantly vertical single force at the source of the VLP signals. The mechanism is interpreted as the movement of a magma column in response to the piston-like rise of a slug of gas in an inclined crack-like conduit of Stromboli volcano [32].

The waveform inversion approach discussed above might be successful if we have many (more than 10) stations covering the entire edifice of a volcano. However, such dense seismic networks exist at only a limited number of volcanoes. Many volcanoes are monitored by networks of just a few seismic stations. It is also possible that signals would be detected by only a few stations close to the source even in a dense seismic network. Nakano and Kumagai [92] proposed a practical approach for quantifying the source of volcano-seismic signals observed by a small number of seismic stations. In this approach, possible source geometries such as a crack and pipe, are assumed as a priori information in waveform inversion. This assumption reduces the number of free parameters and enables us to estimate the source mechanism and location with waveform data from 2 to 3 three-component seismic stations. Their approach is summarized as follows. The moment tensors for a crack and pipe are given in Eqs. (30) and (34), respectively. If we assume a value for λ/μ, free parameters in the moment tensors are the source-time function, θ, and ϕ. We can determine the source-time

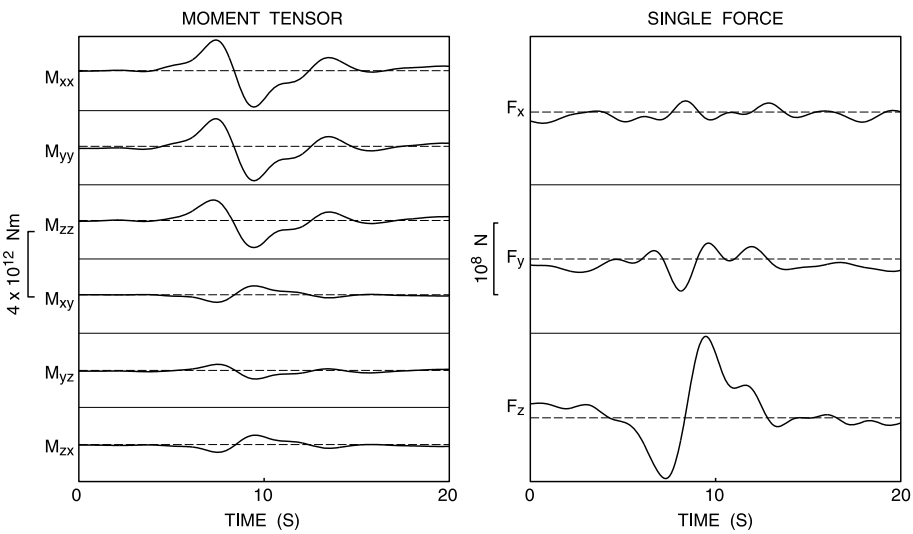

Volcano Seismic Signals, Source Quantification of, Figure 12
Source-time functions of six moment tensor and three single force components estimated by waveform inversion for the VLP signals shown in Fig. 11b [32]

function by the waveform inversion discussed above, in which a grid search with respect to θ and ϕ is conducted to find the best-fit angles at a fixed location. A grid search in space is also conducted to find the combination of the best-fit angles and source location, and their residuals are compared to adopt the best model. The applicability of this simple approach has been demonstrated by synthetic tests and observed seismograms [92].

Spectral Analysis

Volcano-seismic signals often show harmonic oscillatory signatures, which are interpreted as oscillations of fluid-filled resonators at the source of these signals. To determine the frequency characteristics of observed oscillations is essential to quantify the characteristic properties of the resonator systems. The discrete Fourier transform (DFT) has traditionally been used for this purpose. In DFT, the discrete time series $u(k\Delta t)$ $(k = 0, \ldots, N_t - 1)$ is transformed into the discrete frequency series

$$\bar{u}(l\Delta\omega) = \Delta t \sum_{k=0}^{N_t-1} u(k\Delta t) e^{2\pi i l k/N_t} \quad (l = 0, \ldots, N_t - 1), \tag{94}$$

where $\Delta\omega = 2\pi/(N_t\Delta t)$ and $i = \sqrt{-1}$. DFT is the decomposition of time series into a linear combination of the orthogonal basis functions $e^{2\pi i l k/N_t}$ in the frequency domain. Because of the stationary nature of the basis functions, DFT is especially useful to analyze stationary periodic signals. However, volcano-seismic signals often show decaying harmonic oscillations. These may be viewed as the impulsive response of a resonator system, which inherently contains dissipation mechanisms. A decaying harmonic oscillation can be represented by the complex frequency, which cannot be directly estimated by DFT based on the real frequency. The complex frequencies can be determined by solving an inverse problem for parameters in an appropriate model of a given time series. The Sompi method [75] is one such approach that is based on an autoregressive (AR) model of a linear dynamic system. This method has been successfully used to determine the complex frequencies of decaying harmonic oscillations in LP and VLP events. We follow the theory of Sompi presented by Kumazawa et al. [75] and Hori et al. [50].

Let us consider the equation of motion for a linear dynamic system consisting of a block and spring:

$$b\frac{d^2}{dt^2}x(t) + \nu\frac{d}{dt}x(t) + \kappa x(t) = 0, \tag{95}$$

where b is the mass of a block, κ is the spring constant, and $\nu(dx(t)/dt)$ represents a disspation term caused by friction, which is proportional to velocity. We set $x(t) = e^{i\omega t}$ in Eq. (95), which is an eigenfunction of this system. We then obtain

$$b(i\omega)^2 + \nu(i\omega) + \kappa = 0\,. \tag{96}$$

This is the characteristic equation of the system. If we consider oscillation solutions ($\nu^2 - 4bk < 0$), Eq. (96) has two characteristic solutions:

$$i\omega_1 = \frac{-\nu + i\sqrt{4b\kappa - \nu^2}}{2b}\,, \quad i\omega_2 = i\omega_1^*\,, \tag{97}$$

where $*$ denotes the complex conjugate. If we set $\omega = 2\pi(f - ig)$, we obtain

$$f_1 = -f_2 = \frac{1}{2\pi}\sqrt{\frac{\kappa}{b} - \frac{\nu^2}{4b^2}}\,, \tag{98}$$

$$g_1 = g_2 = -\frac{1}{2\pi}\frac{\nu}{2b}\,. \tag{99}$$

Let us consider the following difference equation:

$$\sum_{j=0}^{2} a(j)x((k-j)\Delta t) = 0 \tag{100}$$

or

$$a(0)x(k\Delta t) + a(1)x(k\Delta t - \Delta t) + a(2)x(k\Delta t - 2\Delta t) = 0\,. \tag{101}$$

If we set $x(k\Delta t) = z^k$, where $z = e^{i\omega\Delta t}$, Eq. (100) is given as

$$a(0)z^k + a(1)z^{k-1} + a(2)z^{k-2} = 0 \tag{102}$$

or, equivalently,

$$a(0)z^2 + a(1)z + a(2) = 0\,. \tag{103}$$

Equation (103) is the characteristic equation of the difference equation. If we consider oscillation solutions ($a^2(1) - 4a(0)a(2) < 0$), Eq. (103) has the following two characteristic solutions:

$$z_1 = \frac{-a(1) + i\sqrt{4a(0)a(2) - a^2(1)}}{2a(0)}\,, \quad z_2 = z_1^*\,. \tag{104}$$

These two solutions constitute complex conjugate pairs. By setting $\omega = 2\pi(f - ig)$, we obtain

$$f_1 = -f_2 = \frac{1}{2\pi\Delta t}\tan^{-1}\left(-\sqrt{\frac{4a(0)a(2)}{a^2(1)} - 1}\right)\,, \tag{105}$$

$$g_1 = g_2 = \frac{1}{2\pi\Delta t}\ln\sqrt{\frac{a(2)}{a(0)}}\,. \tag{106}$$

These solutions are equivalent to the solutions of the linear dynamic system given in Eqs. (98) and (99). Therefore, Eq. (100) is a difference form of the equation of motion (95).

We further consider a dynamic system consisting of two block-spring subsystems, which are coupled to each other. The equations of motion for this system are written as

$$M_1\frac{\mathrm{d}^2}{\mathrm{d}t^2}x_1(t)+L_1\frac{\mathrm{d}}{\mathrm{d}t}x_1(t)+K_1x_1(t) = C(x_2-x_1)\,, \tag{107}$$

$$M_2\frac{\mathrm{d}^2}{\mathrm{d}t^2}x_2(t)+L_2\frac{\mathrm{d}}{\mathrm{d}t}x_2(t)+K_2x_2(t) = C(x_1-x_2)\,, \tag{108}$$

where M_i, L_i, and K_i are the mass, dissipation coefficient, and spring constant of the ith block and C is the spring constant between the blocks. The set of the equations of motion (107) and (108) can be equivalently given in the following form:

$$\begin{pmatrix} M_1\frac{\mathrm{d}^2}{\mathrm{d}t^2}+L_1\frac{\mathrm{d}}{\mathrm{d}t}+K_1+C & -C \\ -C & M_2\frac{\mathrm{d}^2}{\mathrm{d}t^2}+L_2\frac{\mathrm{d}}{\mathrm{d}t}+K_2+C \end{pmatrix} \cdot \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = 0\,. \tag{109}$$

These equations have non-trivial solutions of $x_i(t)$ if, and only if,

$$\det\begin{vmatrix} M_1\frac{\mathrm{d}^2}{\mathrm{d}t^2}+L_1\frac{\mathrm{d}}{\mathrm{d}t}+K_1+C & -C \\ -C & M_2\frac{\mathrm{d}^2}{\mathrm{d}t^2}+L_2\frac{\mathrm{d}}{\mathrm{d}t}+K_2+C \end{vmatrix} = 0\,. \tag{110}$$

This leads to the following 4th-order differential equation:

$$\left[\left(M_1\frac{\mathrm{d}^2}{\mathrm{d}t^2}+L_1\frac{\mathrm{d}}{\mathrm{d}t}+K_1+C\right) \cdot\left(M_2\frac{\mathrm{d}^2}{\mathrm{d}t^2}+L_2\frac{\mathrm{d}}{\mathrm{d}t}+K_2+C\right)-C^2\right]x_i(t) = 0\,, \tag{111}$$

where $i = 1, 2$. Each 4th-order differential equation has four complex frequencies, of which two are complex conjugates. Therefore, each block has two independent complex frequencies in this dynamic system. Similarly, a dynamic system consisting of m block-spring subsystems, which are all connected to each other, can be described by the following $2m$-order differential equation:

$$\sum_{j=0}^{2m} b_j\frac{\mathrm{d}^j}{\mathrm{d}t^j}x_i(t) = 0 \quad (i = 1, 2, \ldots, m)\,, \tag{112}$$

where b_j denotes the coefficients determined by the characteristic properties of the system. Each block has m independent complex frequencies in this case. A difference form of Eq. (112) is

$$\sum_{j=0}^{2m} a(j)x_i(k\Delta t - j\Delta t) = 0\,, \tag{113}$$

which is a homogeneous $2m$-order AR equation. Although we have considered block-spring systems as an example, Eq. (112) is a general equation to describe the free oscillations of any linear systems. We use its difference form, Eq. (113), as the basic equation to determine the characteristic complex frequencies of linear systems from time series records.

Hereafter, I set $\Delta t = 1$ to simplify the following notations. We assume that observed discrete time series $u(k)$ consists of signal $x(k)$ and white noise $e(k)$:

$$u(k) = x(k) + e(k) \quad (k = 0, 1, \ldots, N_t - 1)\,, \tag{114}$$

and $x(k)$ satisfies the following homogeneous AR equation:

$$\sum_{j=0}^{2m} a(j)x(k-j) = 0\,. \tag{115}$$

To determine the $2m + 1$ AR coefficients, we minimize the residual R:

$$R = \frac{1}{N_t - 2m}\sum_{k=2m}^{N_t-1}\left[\sum_{j=0}^{2m} a(j)u(k-j)\right]^2\,. \tag{116}$$

Since Eq. (115) is homogeneous, we cannot determine absolute values of the AR coefficients. To avoid the trivial solution $a(j) = 0$ $(j = 0, 1, \ldots, 2m)$, we impose the condition

$$\sum_{j=0}^{2m} a^2(j) = 1\,. \tag{117}$$

To minimize the residual R under the condition of Eq. (117) is a conditional minimization problem. This problem can be solved by minimizing the residual R':

$$R' = R - \beta\left(\sum_{j=0}^{2m} a^2(j) - 1\right)\,, \tag{118}$$

where β is a constant. Consequently, we solve the following equation:

$$\frac{\partial}{\partial a(l)} R' = 0 \quad (l = 0, 1, \ldots, 2m) . \tag{119}$$

This equation leads to an eigenvalue problem

$$\boldsymbol{Pa} = \beta \boldsymbol{a} , \tag{120}$$

where $\boldsymbol{a} = (a(0), a(1), \ldots, a(2m))^{\mathrm{T}}$ and $\boldsymbol{P}$ is a non-Toeplitz matrix given by

$$\boldsymbol{P} = \frac{1}{N_t - 2m} \cdot \begin{bmatrix} \sum_{n=2m}^{N_t-1} u(n)u(n) & \cdots & \sum_{n=2m}^{N_t-1} u(n-2m)u(n) \\ \vdots & \ddots & \vdots \\ \sum_{n=2m}^{N_t-1} u(n)u(n-2m) & \cdots & \sum_{n=2m}^{N_t-1} u(n-2m)u(n-2m) \end{bmatrix} . \tag{121}$$

By solving the eigenvalue problem of Eq. (120) we obtain $2m + 1$ eigenvalues $\beta^{(r)} (r = 1, 2, \ldots, 2m + 1)$ and corresponding eigenvectors $\boldsymbol{a}^{(r)} (r = 1, 2, \ldots, 2m + 1)$. Note that the residual R in Eq. (116) is given as

$$R = \boldsymbol{a}^{\mathrm{T}} \boldsymbol{Pa} = \boldsymbol{a}^{\mathrm{T}} \beta \boldsymbol{a} = \beta . \tag{122}$$

Therefore, each eigenvector gives a local minimum of the residual R. We therefore adopt the eigenvector corresponding to the minimum eigenvalue, which minimizes the residual R. Hereafter, $\boldsymbol{a}$ denotes the eigenvector corresponding to the minimum eigenvalue.

The next step is to determine the complex frequencies from the eigenvector $\boldsymbol{a}$. The characteristic equation of the 2m-order homogeneous AR equation is given as

$$\sum_{j=0}^{2m} a(j) z^{-j} = 0 . \tag{123}$$

This equation is the $2m$-degree algebraic equation for z^{-1}, and has $2m$ complex solutions, which can be solved by a numerical procedure. As mentioned before, a complex solution is always accompanied by its complex conjugate, and therefore there are m independent solutions $z_l (l = 1, 2, \ldots, m)$ in Eq. (123). These solutions are related to the complex frequencies $(f_l - ig_l)$ as

$$z_l = \mathrm{e}^{2\pi i (f_l - i g_l)} \tag{124}$$

or

$$f_l - i g_l = \ln z_l / (2\pi i) . \tag{125}$$

The Q factor is defined through the complex frequencies as

$$Q_l = f_l / (-2 g_l) . \tag{126}$$

The final step is to determine the amplitudes of the solutions z_l. If we set a conjugate pair as $z_l = z^*_{m+l}$, the general solution of Eq. (123) is given as

$$\begin{aligned} x(k) &= \sum_{l=1}^{m} \left(\alpha_l z_l^k + \alpha_{m+l} z_{m+l}^k \right) \\ &= \sum_{l=1}^{m} \left(\alpha_l z_l^k + \left(\alpha_l z_l^k \right)^* \right) . \end{aligned} \tag{127}$$

Here α_l is the complex amplitude given as

$$\alpha_l = r_l \mathrm{e}^{i\phi_l} , \tag{128}$$

where r_l and ϕ_l are the real amplitude and phase, respectively. This is equivalently written as

$$\alpha_l = \alpha_l^r + i \alpha_l^i , \tag{129}$$

where α_l^r and α_l^i are the real and imaginary parts of α_l, respectively. Then, the general solution of Eq. (127) can be rewritten

$$x(k) = \sum_{l=1}^{m} \left(2\alpha_l^r C_l(k) - 2\alpha_l^i S_l(k) \right) , \tag{130}$$

where $C_l(k) = \mathrm{e}^{2\pi g_l k} \cos(2\pi f_l k)$ and $S_l(k) = \mathrm{e}^{2\pi g_l k} \sin(2\pi f_l k)$. The observed time series thus satisfies the following equation

$$u(k) - \sum_{l=1}^{m} \left(2\alpha_l^r C_l(k) - 2\alpha_l^i S_l(k) \right) = e(k) \quad (k = 0, 1, \ldots, N_t - 1) \tag{131}$$

or, in matrix form

$$\boldsymbol{u} - \boldsymbol{A\alpha} = \boldsymbol{e} , \tag{132}$$

where

$$\boldsymbol{u} = (u(0), u(1), \ldots, u(N_t - 1))^{\mathrm{T}} , \tag{133}$$

$$\boldsymbol{\alpha} = \left(2\alpha_1^r, \ldots, 2\alpha_m^r, 2\alpha_1^i, \ldots, 2\alpha_m^i \right)^{\mathrm{T}} , \tag{134}$$

$$\boldsymbol{e} = (e(0), e(1), \ldots, e(N_t - 1))^{\mathrm{T}} , \tag{135}$$

and

$$\boldsymbol{A} = \begin{bmatrix} C_1(0) & \cdots & C_m(0) & -S_1(0) & \cdots & -S_m(0) \\ \vdots & & \vdots & \vdots & & \vdots \\ C_1(N_t-1) & \cdots & C_m(N_t-1) & -S_1(N_t-1) & \cdots & -S_m(N_t-1) \end{bmatrix} . \tag{136}$$

We estimate $\boldsymbol{\alpha}$ by minimizing the residual R_a as follows:

$$R_a = (\boldsymbol{u} - \boldsymbol{A\alpha})^{\mathrm{T}}(\boldsymbol{u} - \boldsymbol{A\alpha}) . \tag{137}$$

This is an ordinary least-squares problem, and the solution is given as

$$\boldsymbol{\alpha} = [\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}]^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{u} . \tag{138}$$

The real amplitudes r_l and phases ϕ_l at $k = 0$ are determined from estimated $\boldsymbol{\alpha}$ as

$$r_l = \sqrt{\left(\alpha_l^r\right)^2 + \left(\alpha_l^i\right)^2} \quad (l = 1, \ldots, m) , \tag{139}$$

$$\phi_l = \tan^{-1}(\alpha_l^i/\alpha_l^r) \quad (l = 1, \ldots, m) . \tag{140}$$

An example of the Sompi analysis of a VLP waveform (Fig. 13a) is shown in Fig. 14. The VLP waveform in Fig. 13a shows decaying harmonic oscillations, which are interpreted as the resonances of a magmatic dike triggered by a time-localized excitation [65,72]. The amplitude spectrum of the VLP waveform based on DFT is shown in Fig. 13b, which indicates two dominant peaks. As such, DFT provides frequency and amplitude information simultaneously. Q values may be determined by the bandwidths of these spectral peaks at halves their peak amplitudes. The estimated Q values, however, can be biased, because the whole waveform including the tail of decaying oscillations as well as the onset of growing oscillations is used in this approach. A box-car window can be applied to extract the decaying oscillations, but this distorts the spectrum and may lead to biased estimations of Q values.

In Sompi, on the other hand, we delete the onset portion of the signal and use the decaying harmonic oscillations only, which satisfy the homogeneous assumption of the AR equation. Since the number of signals (oscilla-

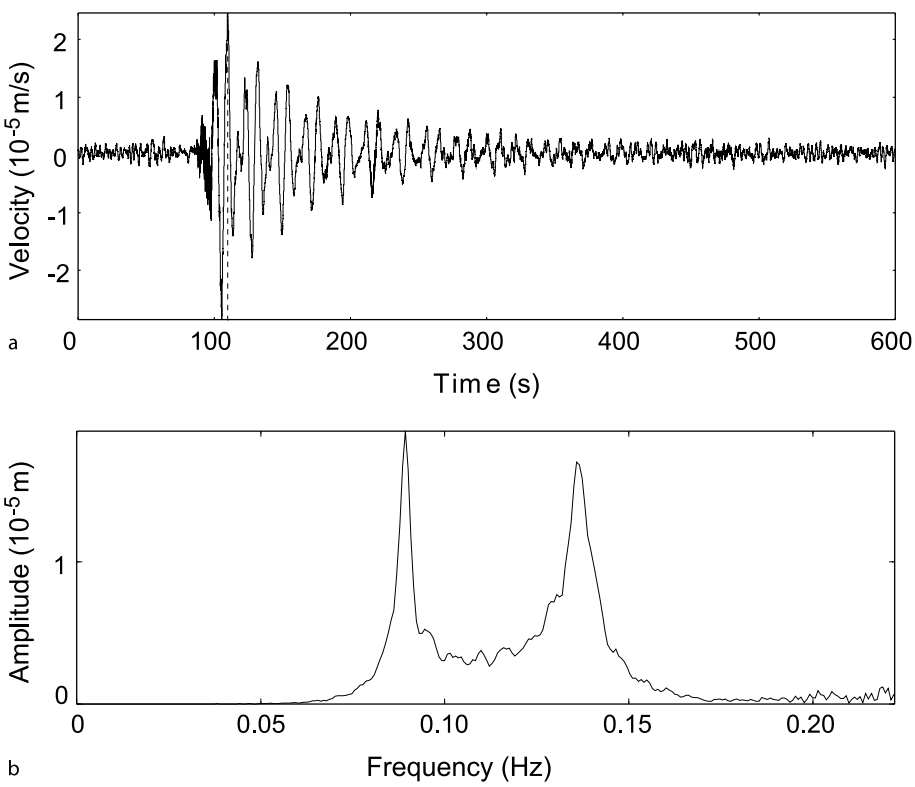

Volcano Seismic Signals, Source Quantification of, Figure 13
a VLP waveform observed at Hachijo Island, Japan, associated with a VT earthquake swarm [65,72]. b Amplitude spectrum of the waveform displayed in Fig. 13a

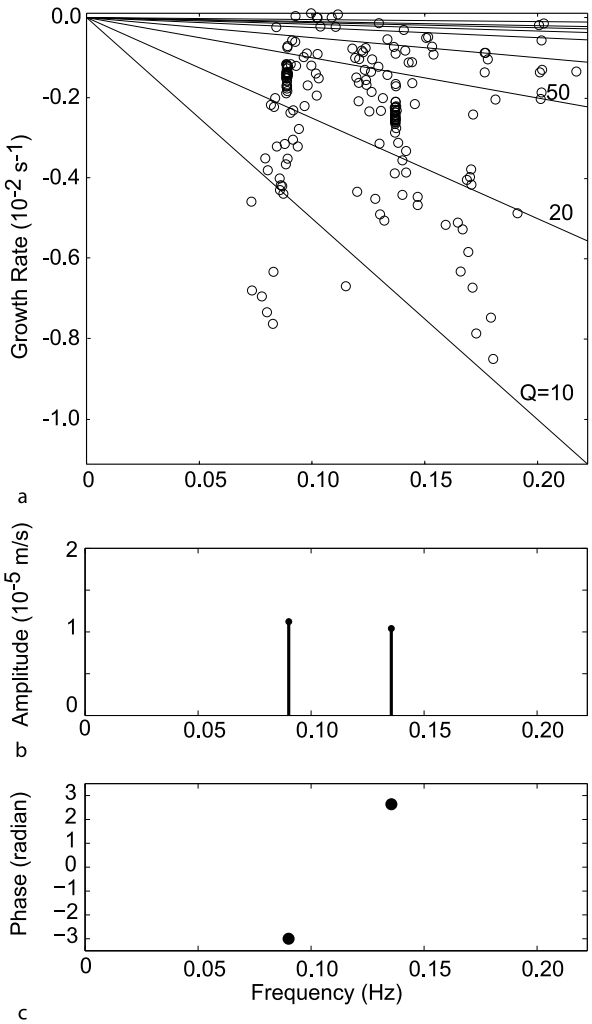

Volcano Seismic Signals, Source Quantification of, Figure 14
An example of the Sompi spectral analysis of the VLP waveform of Fig. 13a, in which Sompi was applied to the waveform after the time of the maximum amplitude indicated by the *dashed line* in Fig. 13a. a A frequency-growth rate diagram for estimated complex frequencies using trial AR orders between 4 and 60. *Solid lines* represent lines along which the *Q* factor is constant. We can identify two densely populated regions, which represent the signal, while other *scattered points* represent incoherent noise. b Amplitudes and c phases at the time of the maximum amplitude using the two complex frequencies determined at an AR order of 40

tions) is not known a priori, we use trial AR orders and solve the eigenvalue problems for individual trial AR orders to determine the complex frequencies, which are plotted in a frequency-growth rate (f-g) diagram. Figure 14a is the f-g diagram of the complex frequencies determined from the VLP waveform by using trial AR orders between 4 and 60. Densely populated regions on the f-g diagram represent the signal for which the complex frequencies are stably determined for different AR orders, while scattered points represent incoherent noise. Sompi then determines the complex amplitudes $re^{i\phi}$ by using the estimated complex frequencies. The optimum AR order and the number of signals may be determined by the two-parameter Akaike information criterion (AIC) [79] or the extended information criterion (EIC) [128]. Figure 14b and c plot the amplitudes r and phases ϕ, respectively, at the beginning of the decaying harmonic oscillations and using the two complex frequencies determined at an AR order of 40.

Some remarks on the theory of the Sompi method may be required. Although Hori et al. [50] and Kumazawa et al. [75] stated that Sompi is a new spectral method. This may be incorrect, as suggested by Ulrych and Sacchi [128]. Let us consider again the AR Eq. (115), which is equivalently rewritten as

$$x(k) = \sum_{j=1}^{2m} a'(j)x(k-j)\,, \tag{141}$$

where $a'(j) = -a(j)/a(0)$. From Eq. (114), we have

$$\begin{aligned} u(k) &= x(k) + e(k) \\ &= \sum_{j=1}^{2m} a'(j)x(k-j) + e(k)\,. \end{aligned} \tag{142}$$

Substituting $u(k-j) = x(k-j) + e(k-j)$ into Eq. (142) we obtain

$$u(k) = \sum_{j=1}^{2m} a'(j)u(k-j) + e(k) - \sum_{j=1}^{2m} a'(j)e(k-j)\,. \tag{143}$$

This is the special autoregressive-moving average (ARMA) model discussed by Ulrych and Clayton [127], who showed that the AR coefficients in Eq. (143) can be determined by an eigenvalue problem that is essentially the same as that given in Eq. (120). In this sense, Sompi is not a new method. However, the Sompi theory, which is based not only on mathematics or information theory but also on physical concepts, yielded a systematic approach to estimate the complex frequencies and complex amplitudes, which can be directly related to the characteristic properties of a linear dynamic system and the system excitation. This approach has been favorably used in analysis of LP and VLP events to estimate the complex frequencies and their temporal variations e. g., [38,65,66,70,72,78,85,91,93].

Fluid-Solid Interactions

Understanding the physics of the fluid-solid interactions is crucial to interpreting source mechanisms and spectral characters of volcano-seismic signals. The pioneering study on the fluid-solid interaction was by Biot [14], who studied seismic waves in a fluid-filled borehole and showed the existence of a slow wave propagating along the borehole boundary. This is called the tube wave. Aki et al. [2] first studied a fluid-filled crack, although the motion of the fluid in the crack was not adequately treated in their study. Chouet [22,23,24] fully studied the fluid-filled crack and demonstrated the existence of a slow wave in the fluid-filled crack, which was called the crack wave. We first consider a fluid layer sandwiched between two homogeneous elastic half-spaces following Ferrazzini and Aki [39] to examine the physical properties of the crack wave. Then, we examine the resonance of a crack of finite size containing a fluid by using the approach proposed by Chouet [22]. For simplicity, we assume an inviscid fluid.

Crack Wave

Let us consider a compressible fluid layer of thickness d, which is sandwiched between two elastic half-spaces bounded at $x_3 = -d/2$ and $x_3 = d/2$ (Fig. 15). The equation of motion for the fluid is given by

$$\rho_f \frac{\partial^2 u_1^f}{\partial t^2} = -\frac{\partial p}{\partial x_1} , \tag{144}$$

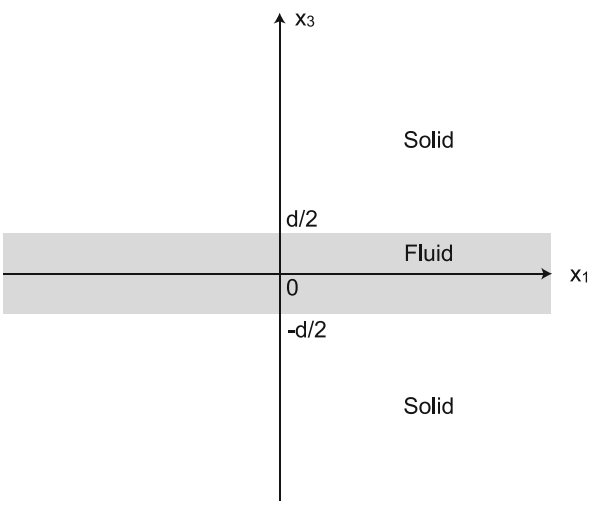

Volcano Seismic Signals, Source Quantification of, Figure 15
A fluid layer of thickness d sandwiched between two elastic half-spaces bounded at $x_3 = -d/2$ and $x_3 = d/2$

$$\rho_f \frac{\partial^2 u_3^f}{\partial t^2} = -\frac{\partial p}{\partial x_3} . \tag{145}$$

Here, ρ_f is the fluid density, u_1^f and u_3^f are the x_1 and x_3 components of the displacement in the fluid, respectively, and p is the fluid pressure satisfying the following continuity equation:

$$p = -b \left(\frac{\partial u_1^f}{\partial x_1} + \frac{\partial u_3^f}{\partial x_3} \right) , \tag{146}$$

where b is the bulk modulus of the fluid. The equation of motion for the solid is

$$\rho_s \frac{\partial^2 u_1}{\partial t^2} = \frac{\partial \sigma_{11}}{\partial x_1} + \frac{\partial \sigma_{13}}{\partial x_3} , \tag{147}$$

$$\rho_s \frac{\partial^2 u_3}{\partial t^2} = \frac{\partial \sigma_{13}}{\partial x_1} + \frac{\partial \sigma_{33}}{\partial x_3} . \tag{148}$$

Here, ρ_s is the solid density, u_1 and u_3 are the x_1 and x_3 components of the displacement in the solid, respectively, and σ_{11}, σ_{33}, and σ_{13} are the components of stress:

$$\sigma_{11} = \lambda \left(\frac{\partial u_1}{\partial x_1} + \frac{\partial u_3}{\partial x_3} \right) + 2\mu \left(\frac{\partial u_1}{\partial x_1} \right) , \tag{149}$$

$$\sigma_{33} = \lambda \left(\frac{\partial u_1}{\partial x_1} + \frac{\partial u_3}{\partial x_3} \right) + 2\mu \left(\frac{\partial u_3}{\partial x_3} \right) , \tag{150}$$

$$\sigma_{13} = \mu \left(\frac{\partial u_1}{\partial x_3} + \frac{\partial u_3}{\partial x_1} \right) . \tag{151}$$

The boundary conditions at the fluid-solid interfaces ($x_3 = \pm d/2$) are

$$u_3 = u_3^f , \tag{152}$$

$$\sigma_{33} = -p , \tag{153}$$

$$\sigma_{13} = 0 . \tag{154}$$

The first and second conditions mean that the normal displacement and stress across the interface are both continuous. The third condition states that the shear stress vanishes at the interface.

For waves propagating in the x_1 direction, the displacements in the fluid may be written in the forms

$$u_1^f = r_1(x_3; k, \omega) e^{i(kx_1 - \omega t)} , \tag{155}$$

$$u_3^f = i r_2(x_3; k, \omega) e^{i(kx_1 - \omega t)} , \tag{156}$$

Substituting these equations into Eq. (146) we obtain

$$p = -i r_3 e^{i(kx_1 - \omega t)} , \tag{157}$$

where k and ω are the wavenumber and angular frequency, respectively

$$r_3 = b\left(kr_1 + \frac{\mathrm{d}r_2}{\mathrm{d}x_3}\right) . \tag{158}$$

Substituting Eqs. (155) and (156) into Eqs. (144) and (145), respectively, and using Eq. (158) we have

$$\rho_f\omega^2 r_1 = kr_3 , \tag{159}$$

$$-\rho_f\omega^2 r_2 = \frac{\mathrm{d}r_3}{\mathrm{d}x_3} . \tag{160}$$

Equations (158), (159), and (160) can be arranged in the following matrix form:

$$\begin{bmatrix} \frac{\mathrm{d}r_2}{\mathrm{d}x_3} \\ \frac{\mathrm{d}r_3}{\mathrm{d}x_3} \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{b} - \frac{k^2}{\rho_f\omega^2} \\ -\rho_f\omega^2 & 0 \end{bmatrix} \begin{bmatrix} r_2 \\ r_3 \end{bmatrix} . \tag{161}$$

For the solid, the displacement components of waves propagating in the x_1 direction may be written as

$$u_1 = y_1(x_3; k, \omega)\mathrm{e}^{i(kx_1-\omega t)} , \tag{162}$$

$$u_3 = iy_2(x_3; k, \omega)\mathrm{e}^{i(kx_1-\omega t)} . \tag{163}$$

Substituting these equations into Eqs. (147), (148), (149), (150), and (151) we obtain the relationship:

$$\begin{bmatrix} \frac{\mathrm{d}y_1}{\mathrm{d}x_3} \\ \frac{\mathrm{d}y_2}{\mathrm{d}x_3} \\ \frac{\mathrm{d}y_3}{\mathrm{d}x_3} \\ \frac{\mathrm{d}y_4}{\mathrm{d}x_3} \end{bmatrix} = \begin{bmatrix} 0 & k & \frac{1}{\mu} & 0 \\ \frac{-k\lambda}{\lambda+2\mu} & 0 & 0 & \frac{1}{\lambda+2\mu} \\ k^2\zeta - \rho_s\omega^2 & 0 & 0 & \frac{k\lambda}{\lambda+2\mu} \\ 0 & -\rho_s\omega^2 & -k & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} , \tag{164}$$

where $\zeta = \frac{4\mu(\lambda+\mu)}{\lambda+2\mu}$ and

$$y_3 = \mu\left(\frac{\mathrm{d}y_1}{\mathrm{d}x_3} - ky_2\right) , \tag{165}$$

$$y_4 = (\lambda + 2\mu)\frac{\mathrm{d}y_2}{\mathrm{d}x_3} + k\lambda y_1 . \tag{166}$$

The boundary condition Eqs.(152), (153), and (154) can then be written as

$$y_2 = r_2 , \tag{167}$$

$$y_4 = r_3 , \tag{168}$$

$$y_3 = 0 , \tag{169}$$

respectively, at $x_3 = \pm d/2$. The systems of Eqs. (161) and (164) are to be solved under the boundary conditions of Eqs. (167), (168), and (169), in which we seek solutions satisfying the following two additional conditions

$$r_2(-x_3) = -r_2(x_3) \quad -d/2 \le x_3 \le d/2 , \tag{170}$$

$$y_1, y_2 \to 0 \quad x_3 \to \pm\infty . \tag{171}$$

The first condition means that motion in the fluid layer is symmetric with respect to x_3. There is the antisymmetric motion in the fluid layer with respect to x_3. This motion is the same as that observed in a fluid layer overlying a solid half-space, such as ocean, and has been investigated by Biot [15] and Tolstoy [124]. This motion does not produce the slow wave, and I do not discuss it here.

The solution for the systems of Eqs. (161) and (164) under the conditions mentioned above can be determined by the following numerical procedure, which was proposed by Takeuchi and Saito [121]. Since we assume symmetric motion in the fluid layer, it is enough to consider the region $x_3 \ge 0$. We choose $x_3 = h$, where h is appropriately far from the fluid-solid interface to satisfy $y_1 = y_2 = 0$. At $x_3 = h$, we assume two independent initial values $y_1 = y_2 = y_4 = 0, y_3 = 1$ and $y_1 = y_2 = y_3 = 0, y_4 = 1$ and integrate the equation for the solid (164) with values of k and ω by the Runge–Kutta method. Then, we have two independent solutions $y_i^A(x_3; k, \omega)$ and $y_i^B(x_3; k, \omega)(i = 1, 2, 3, 4)$. Any solution of Eq. (164) may be expressed as a linear combination of the two solutions:

$$y_i(x_3; k, \omega) = Ay_i^A(x_3; k, \omega) + By_i^B(x_3; k, \omega) \quad (i = 1, 2, 3, 4) , \tag{172}$$

where A and B are arbitrary constants of integration for the two solutions. At $x_3 = 0$, we assume initial values $r_2 = 0$ and $r_3 = 1$, and integrate the equation for the fluid (161) with values of k and ω. Then, we have a solution $r_i^C(x_3; k, \omega)$ $(i = 1, 2)$, and any solution of Eq. (161) may be given as

$$r_i(x_3; k, \omega) = Cr_i^C(x_3; k, \omega) \quad (i = 2, 3) , \tag{173}$$

where C is an arbitrary constant of integration. From the boundary conditions of Eqs. (167), (168), and (169), we have the following relationships at $x_3 = d/2$:

$$\begin{aligned} Ay_2^A(d/2; k, \omega) + By_2^B(d/2; k, \omega) - Cr_2^C(d/2; k, \omega) &= 0 , \\ Ay_4^A(d/2; k, \omega) + By_4^B(d/2; k, \omega) - Cr_3^C(d/2; k, \omega) &= 0 , \\ Ay_3^A(d/2; k, \omega) + By_3^B(d/2; k, \omega) &= 0 . \end{aligned} \tag{174}$$

These equations have a nontrivial solution if, and only if,

$$\det \begin{vmatrix} y_2^A(d/2;k,\omega) & y_2^B(d/2;k,\omega) & -r_2^C(d/2;k,\omega) \\ y_4^A(d/2;k,\omega) & y_4^B(d/2;k,\omega) & -r_3^C(d/2;k,\omega) \\ y_3^A(d/2;k,\omega) & y_3^B(d/2;k,\omega) & 0 \end{vmatrix} = 0 . \tag{175}$$

The dispersion relation of the fluid-solid coupled waves can be determined by searching for k and ω that satisfy Eq. (175). From Eq. (174) we have

$$\frac{B}{A} = -\frac{y_3^A(d/2;k,\omega)}{y_3^B(d/2;k,\omega)} , \tag{176}$$

$$\frac{C}{A} = \frac{y_2^A(d/2;k,\omega)}{r_2^C(d/2;k,\omega)} - \frac{y_3^A(d/2;k,\omega)}{y_3^B(d/2;k,\omega)} \frac{y_2^B(d/2;k,\omega)}{r_2^C(d/2;k,\omega)} , \tag{177}$$

and the eigenfunctions are given as

$$y_i(x_3;k,\omega)/A = y_i^A(x_3;k,\omega) + (B/A) y_i^B(x_3;k,\omega) \quad (i = 1,2,3,4) , \tag{178}$$

$$r_i(x_3;k,\omega)/A = (C/A) r_i^C(x_3;k,\omega) \quad (i = 2,3) . \tag{179}$$

The dispersion relationship and eigenfunctions of the crack wave calculated by the above procedure are shown in Figs. 16 and 17, respectively. In this example, I used the following parameter values: $\rho_s = 1.10\rho_f$, $b = 0.06\mu$, and $\lambda = 2\mu$, which mimic a dike containing basaltic magma with a gas-volume fraction of 15%. The dispersion relationship in Fig. 16 clearly indicates that the phase velocity of the fluid-solid coupled wave depends on the wavelength, in which the velocity decreases as the wavelength increases. This wave is called the crack wave. Figure 17 displays dimensionless normal displacements (y_2 and r_2) and normal stress and pressure (y_4 and r_3) as a function of dimensionless x_3. Figure 17 indicates that the normal displacement is symmetric with respect to x_3. This motion causes the deformation of the fluid-solid boundary. The deformation has the effect of lowering the phase velocity, which is examined below.

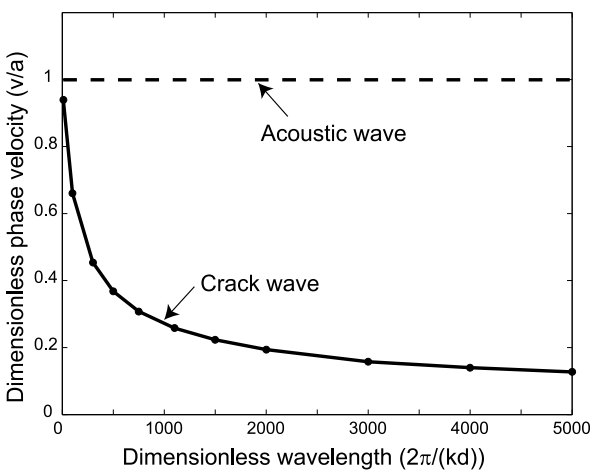

Volcano Seismic Signals, Source Quantification of, Figure 16
The phase velocities of the acoustic wave (*dashed line*) and crack wave (*solid line*). These are normalized by the fluid acoustic velocity, and plotted as a function of the wavelength normalized by the thickness of the fluid layer

Fluid-Filled Crack Model

Chouet [22,23,24] studied the motion of a finite crack containing a fluid in an infinite elastic medium. Following Chouet [22], we consider a crack set in the plane $x_3 = 0$, which extends from $-W/2$ to $W/2$ along the x_1 axis and from 0 to L in the x_2 direction (Fig. 18). The thickness of the crack is d, which is much smaller than L. Chouet [22] used the equations of motion for the solid given as

$$\rho_s \frac{\partial v_i}{\partial t} = \frac{\partial \sigma_{ij}}{\partial x_i} , \tag{180}$$

with the stress-strain relationship

$$\frac{\partial \sigma_{ij}}{\partial t} = \lambda \frac{\partial v_k}{\partial x_k} \delta_{ij} + \mu \left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right) , \tag{181}$$

where v_j is particle velocity of the solid and $i = 1,2,3$. The equations of motion for the fluid are given as

$$\rho_f \frac{\partial^2 u_i^f}{\partial t^2} = -\frac{\partial p}{\partial x_i} , \tag{182}$$

with the equation of continuity

$$p = -b \frac{\partial u_k^f}{\partial x_k} . \tag{183}$$

The boundary conditions on the crack surface S_1 are

$$u_3^f = u_3 , \quad \sigma_{33} = -p , \quad \sigma_{13} = \sigma_{23} = 0 , \tag{184}$$

which are the continuity of normal displacement and stress and the vanishment of shear stress on S_1. Let us introduce u_d, which represents the normal displacement of the crack wall and is identical to $u_3^f = u_3$ on S_1.

We integrate Eq. (183) with respect to x_3 within the crack and divide it by d:

$$\frac{1}{d} \int_{-d/2}^{d/2} p \, dx_3 = -b \frac{1}{d} \int_{-d/2}^{d/2} \left[\frac{\partial u_1^f}{\partial x_1} + \frac{\partial u_2^f}{\partial x_2} + \frac{\partial u_3^f}{\partial x_3} \right] dx_3 . \tag{185}$$

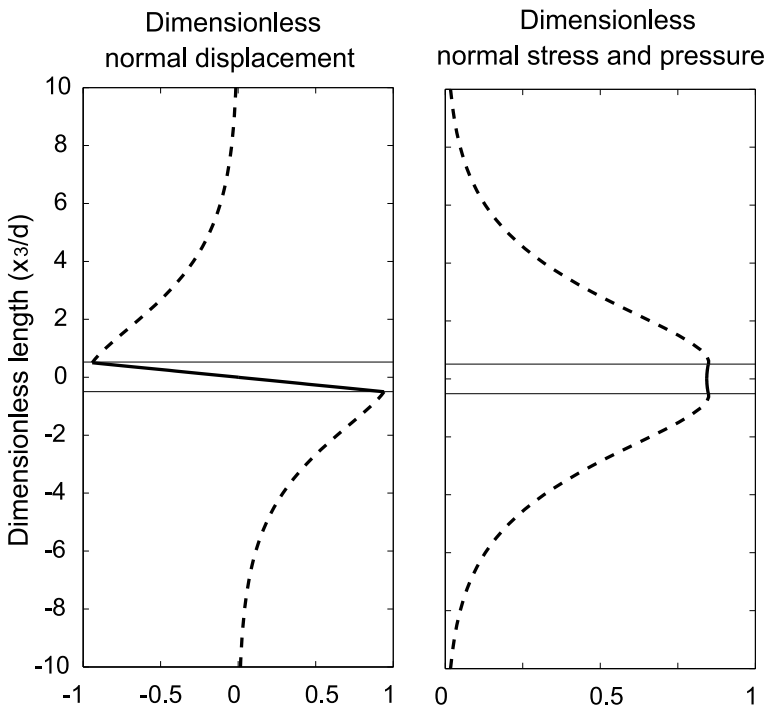

Volcano Seismic Signals, Source Quantification of, Figure 17

Dimensionless normal displacements, normal stress, and pressure as a function of x_3 normalized by the thickness of the fluid layer. *Thin solid lines* represent the boundaries between the fluid and solid, and *thick dashed* and *solid lines* represent the functions in the fluid and solid regions, respectively

If we define

$$P = \frac{1}{d}\int_{-d/2}^{d/2} p\,\mathrm{d}x_3 \tag{186}$$

and

$$U_i^f = \frac{1}{d}\int_{-d/2}^{d/2} u_i^f\,\mathrm{d}x_3 \quad (i = 1, 2)\,, \tag{187}$$

Eq. (185) is written as

$$P = -b\left(\frac{\partial U_1^f}{\partial x_1} + \frac{\partial U_2^f}{\partial x_2} + \frac{1}{d}\int_{-d/2}^{d/2}\frac{\partial u_3^f}{\partial x_3}\mathrm{d}x_3\right)\,. \tag{188}$$

Since u_3^f is symmetric with respect to $x_3(u_3^f(-x_3) = -u_3^f(x_3))$, we find that

$$\int_{-d/2}^{d/2}\frac{\partial u_3^f}{\partial x_3}\mathrm{d}x_3 = \int_{-d/2}^{d/2}\mathrm{d}u_3^f = [u_3^f]_{-d/2}^{d/2} = 2u_3^f(d/2)\,. \tag{189}$$

Note that $u_3^f(d/2) = u_3(d/2) = u_d$ from the boundary condition of Eq. (184) and, therefore, we have

$$P = -b\left(\frac{\partial U_1^f}{\partial x_1} + \frac{\partial U_2^f}{\partial x_2} + \frac{2}{d}u_d\right)\,. \tag{190}$$

From Eq. (182) we obtain

$$\rho_f\frac{\partial^2 U_1^f}{\partial t^2} = -\frac{\partial P}{\partial x_1}\,, \tag{191}$$

$$\rho_f\frac{\partial^2 U_2^f}{\partial t^2} = -\frac{\partial P}{\partial x_2}\,. \tag{192}$$

Using the fluid velocity, Eqs. (191), (192), and (190) can be equivalently written as

$$\rho_f\frac{\partial V_1^f}{\partial t} = -\frac{\partial P}{\partial x_1}\,, \tag{193}$$

$$\rho_f\frac{\partial V_2^f}{\partial t} = -\frac{\partial P}{\partial x_2}\,, \tag{194}$$

$$\frac{\partial P}{\partial t} = -b\left(\frac{\partial V_1^f}{\partial x_1} + \frac{\partial V_2^f}{\partial x_1} + \frac{2}{d}v_d\right)\,, \tag{195}$$

where $V_i^f = \partial U_i^f/\partial t$ $(i = 1, 2)$ and v_d is the normal velocity of the crack wall.

Thus, the three-dimensional equations of motion and continuity are reduced to two-dimensional equations with

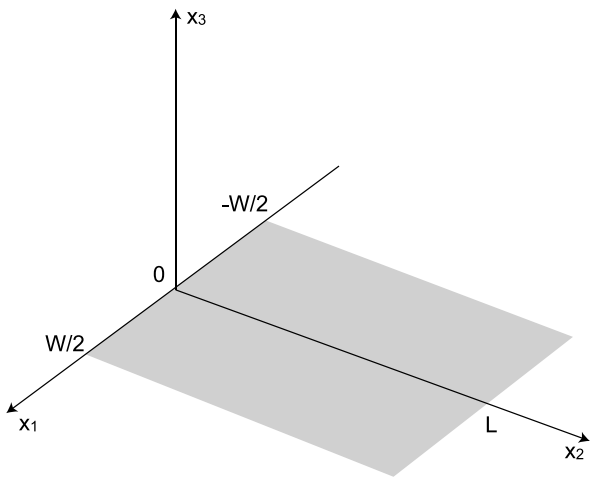

Volcano Seismic Signals, Source Quantification of, Figure 18
Coordinates and geometry of a crack set in the plane $x_3 = 0$, which extends from $-W/2$ to $+W/2$ on the x_1 axis and from 0 to L on the x_2 axis

respect to V_1^f, V_2^f, P, and v_d. In this problem, the boundary conditions are

$$\sigma_{33} = -p\,, \quad \sigma_{13} = \sigma_{23} = 0\,, \quad \text{on} \quad S_1\,, \tag{196}$$

$$u_3 = 0\,, \quad \sigma_{13} = \sigma_{23} = 0\,, \quad \text{on} \quad S_2\,, \tag{197}$$

where S_1 denotes the crack surface and S_2 denotes the crack plane outside the crack. The latter condition expresses that the displacement in the solid is symmetric with respect to x_3. In addition to these boundary conditions, Chouet [22] assumed zero mass transfer in and out of the crack, which are as follows:

$$V_1^f = v_1 \quad \text{along} \quad |x_1| = W/2\,, 0 \le x_2 \le L\,, \tag{198}$$

$$V_2^f = v_2 \quad \text{along} \quad -W/2 \le x_1 \le W/2\,, x_2 = 0 \;\; x_2 = L\,. \tag{199}$$

The equations of motion and stress-strain relationship for the solid in the form of Eqs. (180) and (181) and the equations of motion and continuity for the fluid in the forms of Eqs. (193), (194) and (195) are solved under the boundary conditions of Eqs. (196), (197), (198), and (199). Chouet [22] solved this problem by using a finite-difference method, in which the following scaling was used:

$$\bar{x}_i = x_i/L\,, \tag{200}$$

$$\bar{t} = \alpha t/L\,, \tag{201}$$

$$\bar{\sigma}_{ij} = \sigma_{ij}/\sigma_0\,, \tag{202}$$

$$\bar{u}_i = u_i \mu/(L\sigma_0)\,, \tag{203}$$

$$\bar{v}_i = v_i \mu/(\alpha\sigma_0)\,, \tag{204}$$

where nondimensional variables are indicated by a bar, σ_0 is an effective stress, and α is the compressional wave velocity $\alpha = \sqrt{(\lambda + 2\mu)/\rho_s}$. Let us define the first-order difference operators in time and space as:

$$\Delta_t f(l; i, j, k) = f(l+1; i, j, k) - f(l; i, j, k)\,, \tag{205}$$

$$\Delta_{x1} f(l; i, j, k) = f(l; i+1, j, k) - f(l; i, j, k)\,, \tag{206}$$

$$\Delta_{x2} f(l; i, j, k) = f(l; i, j+1, k) - f(l; i, j, k)\,, \tag{207}$$

$$\Delta_{x3} f(l; i, j, k) = f(l; i, j, k+1) - f(l; i, j, k)\,, \tag{208}$$

where $f(l; i, j, k)$ represents any function at time $l\Delta t$ and position $[i\Delta, j\Delta, k\Delta]$ or $[i\Delta_a, j\Delta_a, k\Delta_a]$. Here, Δt is the time increment and Δ and Δ_a are the grid spacings for the solid and fluid, respectively. A difference form of Eq. (195) is then obtained as

$$\frac{\Delta_t P}{\Delta t} = -b\left(\frac{\Delta_{x1} V_1^f}{\Delta_a} + \frac{\Delta_{x2} V_2^f}{\Delta_a} + \frac{2}{d} v_d\right) \tag{209}$$

or

$$\Delta_t P = -bH_a\left(\Delta_{x1} V_1^f + \Delta_{x2} V_2^f\right) - 2\frac{b}{d}\Delta t v_d\,, \tag{210}$$

where $H_a = \Delta t/\Delta_a$. Using the scaling given by Eqs. (200) to (204), Eq. (210) can be further modified to a nondimensional form:

$$\Delta_t \bar{P} = -\left(\frac{b}{\mu}\right)\alpha H_a\left(\Delta_{x1} \bar{V}_1^f + \Delta_{x2} \bar{V}_2^f\right) - 2C\Delta\bar{t}\bar{v}_d\,, \tag{211}$$

where $C = (b/\mu)(L/d)$ is called the crack stiffness [2]. In a similar manner, we obtain nondimensional difference forms of equations of motion for the fluid (193) and (194) as

$$\Delta_t \bar{V}_1^f = -\left(\frac{a}{\alpha}\right)^2\left(\frac{\mu}{b}\right)\alpha H_a \Delta_{x1}\bar{P}\,, \tag{212}$$

$$\Delta_t \bar{V}_2^f = -\left(\frac{a}{\alpha}\right)^2\left(\frac{\mu}{b}\right)\alpha H_a \Delta_{x2}\bar{P}\,, \tag{213}$$

where $a = \sqrt{b/\rho_f}$ is the acoustic velocity of the fluid and the ratios α/a and b/μ are related to ρ_f and ρ_s through the relationship

$$\frac{\rho_f}{\rho_s} = \left(\frac{\mu}{\lambda + 2\mu}\right)\left(\frac{\alpha}{a}\right)^2\left(\frac{b}{\mu}\right)\,. \tag{214}$$

Nondimensional difference forms of the equations of motion and the stress-strain relationship for the solid

(Eqs. (180) and (181)) are

$$\Delta_t \bar{v}_i = \frac{1}{2 + (\lambda/\mu)} \alpha H \Delta_{xi} \bar{\sigma}_{ij} \,, \tag{215}$$

$$\Delta_t \bar{\sigma}_{ij} = \alpha H \left[\frac{\lambda}{\mu} \Delta_{xk} \bar{v}_k \delta_{ij} + \left(\Delta_{xi} \bar{v}_i + \Delta_{xj} \bar{v}_j \right) \right] \,, \tag{216}$$

where $H = \Delta t / \Delta$. To satisfy the conditions for numerical stability in the solid and fluid, Chouet [22] used the following relationships:

$$H = \frac{\Delta t}{\Delta} = \frac{1}{4\alpha} \,, \tag{217}$$

$$H_a = \frac{\Delta t}{\Delta_a} = \frac{1}{2a} \,. \tag{218}$$

Applying a step increase in the fluid pressure over an area of the crack, the nondimensional difference equations for the solid and fluid are solved by a finite-difference method with a staggered grid scheme [129] to determine the normal velocity of the crack wall v_d. This source space-time function is used to synthesize the ground response to the fluid-filled crack. Note that Chouet [22] added an artificial dissipation term to the equations for particle velocities in the solid (Eq. (215)) and for fluid velocities (Eqs. (212) and (213)) to suppress spurious numerical oscillations associated with discretization in time and space in the finite difference calculations.

Snapshots of the normal velocity of the crack wall v_d for a fluid-filled crack excited by a step increase in the pressure applied over a strip extending over the entire width of the crack are shown in Fig. 19a. The results show that the crack oscillation is dominated by the longitudinal modes with wavelengths L and $2L/5$. Figure 19b shows snapshots of v_d for the crack with the excitation applied over a small area at the center of the crack. The crack oscillation shown in Fig. 19b displays more complex spatial patters than those of Fig. 19a, in which both the longitudinal and transverse modes with wavelengths $2L/3$ and $2W/3$, respectively, are dominantly excited. Chouet [22] showed that the dimensionless phase velocity v/a of the crack wave sustaining the crack resonance depends on the crack stiffness C. The Q factor of the crack resonance almost monotonically increases with the increasing α/a [67,68].

The slow wave and its dependence on the crack stiffness may be understood by the following simple theoretical consideration. Using Eqs. (190), (191), and (192), we obtain the following relationship:

$$\frac{\partial^2}{\partial t^2} \left(P + \frac{2b}{d} u_d \right) = \left(\frac{b}{\rho_f} \right) \left(\frac{\partial^2 P}{\partial x_1^2} + \frac{\partial^2 P}{\partial x_2^2} \right) \,. \tag{219}$$

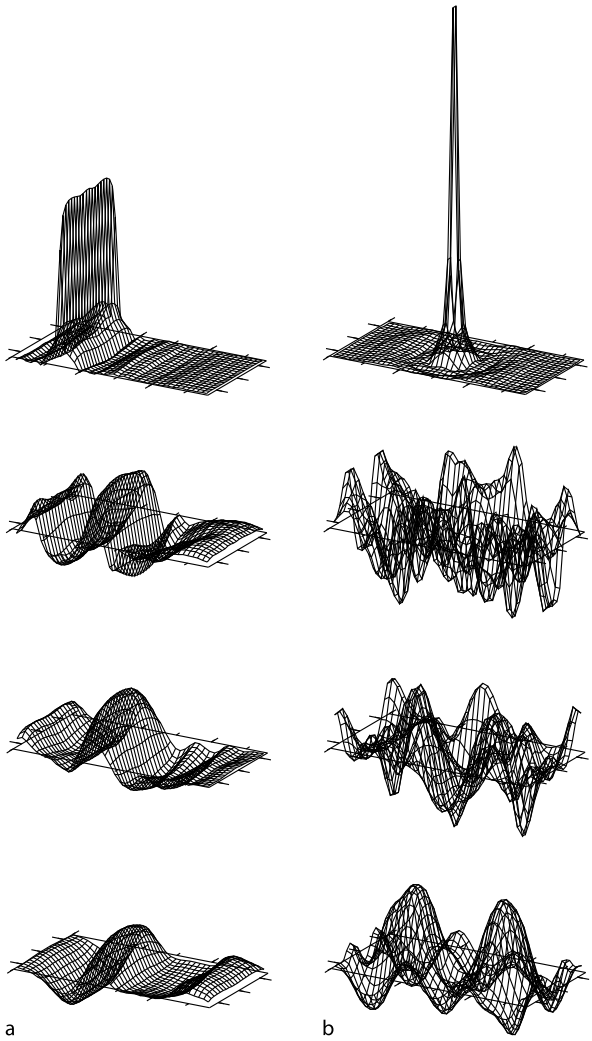

Volcano Seismic Signals, Source Quantification of, Figure 19
a Snapshots of the normal velocity of the crack wall v_d for a fluid-filled crack excited by a step increase in the pressure applied over a strip extending over the entire width of the crack. b Snapshots of v_d for the crack with the excitation applied over a small area at the center of the crack (modified from Nakano et al. [95])

Note that $a = \sqrt{b/\rho_f}$ is the acoustic velocity of the fluid. Let us consider a simple case for which u_d is proportional to P:

$$u_d = \epsilon P \,, \tag{220}$$

where ϵ is a proportionality constant. Using the scaling given in Eqs. (202) and (203), the constant ϵ may be written in a nondimensional form as

$$\epsilon = \bar{\epsilon} L / \mu \,, \tag{221}$$

where $\bar{\epsilon}$ is a nondimensional proportionality constant. Using Eqs. (220) and (221), Eq. (219) can be written as

$$\left(1 + 2\bar{\epsilon}\frac{b}{\mu}\frac{L}{d}\right)\frac{\partial^2 P}{\partial t^2} = \left(\frac{b}{\rho_f}\right)\left(\frac{\partial^2 P}{\partial x_1^2} + \frac{\partial^2 P}{\partial x_2^2}\right) \tag{222}$$

or, equivalently,

$$\frac{\partial^2 P}{\partial t^2} = \left(\frac{b_e}{\rho_f}\right)\left(\frac{\partial^2 P}{\partial x_1^2} + \frac{\partial^2 P}{\partial x_2^2}\right), \tag{223}$$

where b_e is an effective bulk modulus defined as

$$b_e = b/(1 + 2\bar{\epsilon}C). \tag{224}$$

Here, $C = (b/\mu)(L/d)$ is the crack stiffness. Equation (223) is the two-dimensional wave equation with phase velocity $a_e = \sqrt{b_e/\rho_f}$. If no deformation of the crack wall occurs ($\bar{\epsilon} = 0$), the phase velocity a_e is equal to the fluid acoustic velocity a. If the deformation of the crack wall occurs such that the wall moves inward to increase fluid pressure ($\bar{\epsilon} < 0$), a_e becomes smaller than a through a reduction of the effective bulk modulus, which depends on the crack stiffness as given in Eq. (224).

The slow crack wave leads to more realistic estimates of the size and volume of a fluid-filled resonator as compared to a resonator with spherical geometry e. g. [35,40,41,115]. The slow wave has been used to interpret spectral characteristics of LP and VLP events observed at various active volcanoes [29,44,52,65,66,67,70,72,73,85,88,91,98,117,131].

Future Directions

Recent volcano seismological studies have contributed greatly to achieving a better understanding of magmatic and hydrothermal systems. It is clear that the theoretical development of source models and analysis techniques has played an essential role in recent advances. The current studies in volcano seismology may be summarized as follows: (a) physical understanding and modeling of source dynamics of LP and VLP events [34,99,100,125], (b) development of a waveform inversion method of volcano-seismic signals for an extended source to achieve a better understanding of the resonance characteristics and triggering mechanisms [95], (c) extension of the crack model to incorporate more realistic fluid acoustic properties [65,131], and (d) dense broadband seismic observation and the use of waveform inversion techniques to monitor active volcanoes e. g., [11,37,74]. Future studies building on these will contribute further to our ability to quantify the sources of volcano-seismic signals, which will help us in our efforts to predict eruptive behavior and thus mitigate volcanic hazards through improved seismic monitoring of magmatic and hydrothermal activity.

Acknowledgment

I am deeply grateful to Masaru Nakano for numerous discussions on all the subjects presented in this manuscript. I thank Yasuko Takei for constructive comments on the phenomenological source representation. Comments from Pablo Palacios, Luca D'Auria, Takeshi Nishimura, and an anonymous reviewer helped improve the manuscript. I used the Generic Mapping Tools (GMT) [130] in the preparation of figures.

Appendix A: Green's Functions

Here, I briefly explain the relationship between the displacement and Green's functions. To simplify the explanation, I use two-dimensional equations in an infinite medium. The extension of the equations into three-dimension is straightforward. Green's functions defined by Eq. (5) are explicitly written as

$$\rho\frac{\partial^2 G_{11}}{\partial t^2} = (\lambda + 2\mu)\frac{\partial^2 G_{11}}{\partial x_1^2} + (\lambda + \mu)\frac{\partial^2 G_{21}}{\partial x_2 \partial x_1} + \mu\frac{\partial^2 G_{11}}{\partial x_2^2} + \delta(\boldsymbol{x} - \boldsymbol{\eta})\delta(t - \tau), \tag{A1}$$

$$\rho\frac{\partial^2 G_{21}}{\partial t^2} = (\lambda + 2\mu)\frac{\partial^2 G_{21}}{\partial x_2^2} + (\lambda + \mu)\frac{\partial^2 G_{11}}{\partial x_2 \partial x_1} + \mu\frac{\partial^2 G_{21}}{\partial x_1^2}, \tag{A2}$$

$$\rho\frac{\partial^2 G_{12}}{\partial t^2} = (\lambda + 2\mu)\frac{\partial^2 G_{12}}{\partial x_1^2} + (\lambda + \mu)\frac{\partial^2 G_{22}}{\partial x_2 \partial x_1} + \mu\frac{\partial^2 G_{12}}{\partial x_2^2}, \tag{A3}$$

$$\rho\frac{\partial^2 G_{22}}{\partial t^2} = (\lambda + 2\mu)\frac{\partial^2 G_{22}}{\partial x_2^2} + (\lambda + \mu)\frac{\partial^2 G_{12}}{\partial x_2 \partial x_1} + \mu\frac{\partial^2 G_{22}}{\partial x_1^2} + \delta(\boldsymbol{x} - \boldsymbol{\eta})\delta(t - \tau). \tag{A4}$$

Note that (G_{11}, G_{21}) represents the x_1 and x_2 components of the wavefield at $(\boldsymbol{x}, t)$, which is excited by the impulse in the x_1 direction applied at $\boldsymbol{x} = \boldsymbol{\eta}$ and $t = \tau$. Similarly, (G_{12}, G_{22}) represents the x_1 and x_2 components of the wavefield excited by the impulse in the x_2 direction at $\boldsymbol{x} = \boldsymbol{\eta}$ and at $t = \tau$. Therefore, i and j in G_{ij} represent the component and direction of the impulse, respectively. To specify the relationship between the receiver and source, we use the notation $G_{ij}(\boldsymbol{x}, t; \boldsymbol{\eta}, \tau)$. We

obtain $G_{ij}(\boldsymbol{x}, t; \boldsymbol{\eta}, \tau) = G_{ij}(\boldsymbol{x}, t-\tau; \boldsymbol{\eta}, 0)$, since Green's functions are independent of the time of origin.

The displacement $u_i(\boldsymbol{x}, t)$ satisfies Eq. (3), which is explicitly written as

$$\rho \frac{\partial^2 u_1}{\partial t^2} = (\lambda + 2\mu)\frac{\partial^2 u_1}{\partial x_1^2} + (\lambda + \mu)\frac{\partial^2 u_2}{\partial x_1 \partial x_2} + \mu \frac{\partial^2 u_1}{\partial x_2^2} + f_1^S(\boldsymbol{x}, t)\,, \quad \text{(A5)}$$

$$\rho \frac{\partial^2 u_2}{\partial t^2} = (\lambda + 2\mu)\frac{\partial^2 u_2}{\partial x_2^2} + (\lambda + \mu)\frac{\partial^2 u_1}{\partial x_1 \partial x_2} + \mu \frac{\partial^2 u_2}{\partial x_1^2} + f_2^S(\boldsymbol{x}, t)\,. \quad \text{(A6)}$$

Equation (6) indicates that u_i is described by the following relationships:

$$u_1(\boldsymbol{x}, t) = \int_{-\infty}^{\infty}\iint_{-\infty}^{\infty} \left[f_1^S(\boldsymbol{\eta}, \tau) G_{11}(\boldsymbol{x}, t-\tau; \boldsymbol{\eta}, 0) + f_2^S(\boldsymbol{\eta}, \tau) G_{12}(\boldsymbol{x}, t-\tau; \boldsymbol{\eta}, 0)\right] \mathrm{d}\eta_1\, \mathrm{d}\eta_2\, \mathrm{d}\tau\,, \quad \text{(A7)}$$

$$u_2(\boldsymbol{x}, t) = \int_{-\infty}^{\infty}\iint_{-\infty}^{\infty} \left[f_1^S(\boldsymbol{\eta}, \tau) G_{21}(\boldsymbol{x}, t-\tau; \boldsymbol{\eta}, 0) + f_2^S(\boldsymbol{\eta}, \tau) G_{22}(\boldsymbol{x}, t-\tau; \boldsymbol{\eta}, 0)\right] \mathrm{d}\eta_1\, \mathrm{d}\eta_2\, \mathrm{d}\tau\,. \quad \text{(A8)}$$

We can verify that the displacement given by Eqs. (A7) and (A8) satisfies Eqs. (A5) and (A6) in the following way. We can rewrite Eqs. (A5) and (A6) as

$$f_1^S(\boldsymbol{x}, t) = \rho \frac{\partial^2 u_1}{\partial t^2} - (\lambda + 2\mu)\frac{\partial^2 u_1}{\partial x_1^2} - (\lambda + \mu)\frac{\partial^2 u_2}{\partial x_1 \partial x_2} - \mu \frac{\partial^2 u_1}{\partial x_2^2}\,, \quad \text{(A9)}$$

$$f_2^S(\boldsymbol{x}, t) = \rho \frac{\partial^2 u_2}{\partial t^2} - (\lambda + 2\mu)\frac{\partial^2 u_2}{\partial x_2^2} - (\lambda + \mu)\frac{\partial^2 u_1}{\partial x_1 \partial x_2} - \mu \frac{\partial^2 u_2}{\partial x_1^2}\,. \quad \text{(A10)}$$

We denote the right-hand side of Eq. (A9) as L_1, which is derived from Eqs. (A7) and (A8) as

$$L_1 = \int_{-\infty}^{\infty}\iint_{-\infty}^{\infty} \left[\rho\left(f_1^S \frac{\partial^2 G_{11}}{\partial t^2} + f_2^S \frac{\partial^2 G_{12}}{\partial t^2}\right) - (\lambda + 2\mu)\left(f_1^S \frac{\partial^2 G_{11}}{\partial x_1^2} + f_2^S \frac{\partial^2 G_{12}}{\partial x_1^2}\right) - (\lambda + \mu)\left(f_1^S \frac{\partial^2 G_{21}}{\partial x_1 \partial x_2} + f_2^S \frac{\partial^2 G_{22}}{\partial x_1 \partial x_2}\right) - \mu\left(f_1^S \frac{\partial^2 G_{11}}{\partial x_2^2} + f_2^S \frac{\partial^2 G_{12}}{\partial x_2^2}\right)\right] \mathrm{d}\eta_1\, \mathrm{d}\eta_2\, \mathrm{d}\tau\,. \quad \text{(A11)}$$

This equation can be modified as follows:

$$L_1 = \int_{-\infty}^{\infty}\iint_{-\infty}^{\infty} \left[f_1^S\left\{\rho \frac{\partial^2 G_{11}}{\partial t^2} - (\lambda + 2\mu)\frac{\partial^2 G_{11}}{\partial x_1^2} - (\lambda + \mu)\frac{\partial^2 G_{21}}{\partial x_1 \partial x_2} - \mu \frac{\partial^2 G_{11}}{\partial x_2^2}\right\} + f_2^S\left\{\rho \frac{\partial^2 G_{12}}{\partial t^2} - (\lambda + 2\mu)\frac{\partial^2 G_{12}}{\partial x_1^2} - (\lambda + \mu)\frac{\partial^2 G_{22}}{\partial x_1 \partial x_2} - \mu \frac{\partial^2 G_{12}}{\partial x_2^2}\right\}\right] \mathrm{d}\eta_1\, \mathrm{d}\eta_2\, \mathrm{d}\tau\,. \quad \text{(A12)}$$

From Eq. (A3), we find that the second integral of Eq. (A12) is zero. Therefore, Eq. (A12) from Eq. (A1) becomes

$$L_1 = \int_{-\infty}^{\infty}\iint_{-\infty}^{\infty} f_1^S(\boldsymbol{\eta}, \tau)\delta(\boldsymbol{x} - \boldsymbol{\eta})\delta(t - \tau)\, \mathrm{d}\eta_1\, \mathrm{d}\eta_2\, \mathrm{d}\tau = f_1^S(\boldsymbol{x}, t)\,. \quad \text{(A13)}$$

Similarly, we can show that the right-hand side of Eq. (A10) is equivalent to $f_2^S(\boldsymbol{x}, t)$, and thus the displacement in the forms of Eqs. (A7) and (A8) satisfies the equation of motion (A5) and (A6).

Appendix B: Moment Tensor for a Spherical Source

Let us consider the moment density tensors for three tensile cracks in the planes $\xi_1 = 0, \xi_2 = 0$, and $\xi_3 = 0$. If we sum and average these three tensors and assume that $[u_1] = [u_2] = [u_3] = D_s$, we obtain the following moment density tensor:

$$\boldsymbol{m} = D_s \begin{pmatrix} \lambda + 2\mu/3 & 0 & 0 \\ 0 & \lambda + 2\mu/3 & 0 \\ 0 & 0 & \lambda + 2\mu/3 \end{pmatrix}. \quad \text{(B1)}$$

This represents the moment density tensor for the isotropic expansion of a cubic element. Following Müller [89], we consider a spherical crack surface of radius R where a constant radial expansion $D_s = (d_s + \Delta_s)$ occurs (Fig. A1). Here, the inner wall of the crack moves inward by d_s and the outer wall moves outward by Δ_s. The moment tensor of the spherical expansion may be obtained by integration of the moment density tensor (B1) over the surface R:

$$\boldsymbol{M} = \Delta V \begin{pmatrix} \lambda + 2\mu/3 & 0 & 0 \\ 0 & \lambda + 2\mu/3 & 0 \\ 0 & 0 & \lambda + 2\mu/3 \end{pmatrix}, \quad \text{(B2)}$$

where ΔV is the volume given as

$$\Delta V = 4\pi R^2 D_s = 4\pi R^2 (d_s + \Delta_s)\,. \quad \text{(B3)}$$

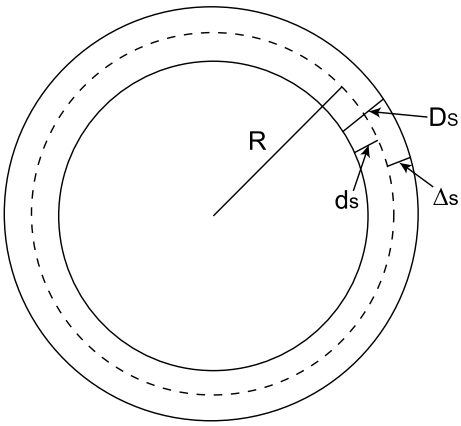

Volcano Seismic Signals, Source Quantification of, Figure A1
A spherical crack surface of radius R where a constant radial expansion $D_s = (d_s + \Delta_s)$ occurs. Here, the inner wall of the crack moves inward by d_s and the outer wall moves outward by Δ_s [89]

The volume $4\pi R^2 d_s$ is caused by the inward motion, which compresses the sphere. The volume $\Delta V_s = 4\pi R^2 \Delta_s$, on the other hand, is caused by the outward motion, which excites seismic waves in the region outside the sphere. ΔV_s can be determined by solving an elastostatic boundary-value problem in the following way. We assume an isotropic medium, and denote the regions inside and outside the sphere as regions 1 and 2, respectively. Since the motion is radial only, the equation of motion is given as e. g., [121]

$$\rho \frac{\partial^2 u}{\partial t^2} = \frac{1}{r^2}\frac{\partial}{\partial r}(r^2 \sigma_{rr}) - \frac{1}{r}(\sigma_{\theta\theta} + \sigma_{\phi\phi}) \tag{B4}$$

and

$$\sigma_{rr} = (\lambda + 2\mu)\frac{\partial u}{\partial r} + \lambda \frac{2}{r} u \,, \tag{B5}$$

$$\sigma_{\theta\theta} = \sigma_{\phi\phi} = \lambda \frac{\partial u}{\partial r} + (\lambda + \mu)\frac{2}{r} u \,, \tag{B6}$$

where u is the radial displacement and σ_{rr}, $\sigma_{\theta\theta}$, and $\sigma_{\phi\phi}$ are the stress components in the spherical coordinate. Equations (B4) and (B5) apply to both regions 1 and 2. Substituting Eq. (B5) into Eq. (B4) and setting $\rho(\partial^2 u_r/\partial t^2) = 0$, we obtain the following static equilibrium equation:

$$\frac{\partial^2 u}{\partial r^2} + \frac{2}{r}\frac{\partial u}{\partial r} - \frac{2}{r^2} u = 0 \,. \tag{B7}$$

This equation has two solutions: $u = ar$ and $u = b/r^2$, where a and b are constants. The former is the interior solution for region 1 ($r \le R$), and the latter is the exterior solution for region 2 ($r \ge R$). I denote the interior and exterior solutions as u_i and u_e, respectively. The constants a and b are determined by the boundary conditions for the radial displacement and the continuity of the radial stress at $r = R$:

$$u_e(R) - u_i(R) = \frac{b}{R^2} - aR = D_s \,, \tag{B8}$$

$$\sigma^i_{rr}(R) - \sigma^e_{rr}(R) = -\frac{4\mu}{R^3} b - (3\lambda + 2\mu) a = 0 \,, \tag{B9}$$

where σ^i_{rr} and σ^e_{rr} are the radial stresses in regions 1 and 2, respectively. Accordingly, we obtain

$$u_i(r) = -\frac{4\mu D_s}{3(\lambda + 2\mu)} \frac{r}{R} \quad (r \le R) \,, \tag{B10}$$

$$u_e(r) = \frac{(\lambda + 2\mu/3) D_s}{\lambda + 2\mu} \frac{R^2}{r^2} \quad (r \ge R) \,. \tag{B11}$$

Then, we obtain

$$d_s = -u_i(R) = \frac{4\mu D_s}{3(\lambda + 2\mu)} \,, \tag{B12}$$

$$\Delta_s = u_e(R) = \frac{(\lambda + 2\mu/3) D_s}{\lambda + 2\mu} \,. \tag{B13}$$

Equation (B3) can be modified as

$$\Delta V = 4\pi R^2 \Delta_s (D_s/\Delta_s) = \frac{\lambda + 2\mu}{\lambda + 2\mu/3} \Delta V_s \,. \tag{B14}$$

Finally, we obtain the moment tensor for the spherical expansion as

$$\boldsymbol{M} = (\lambda + 2\mu)\Delta V_s \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \,. \tag{B15}$$

Appendix C: Moment Tensor for a Cylindrical Source

We consider the moment density tensors for two tensile cracks in the planes $\xi_1 = 0$ and $\xi_2 = 0$, where ξ_1 and ξ_2 are two horizontal axes. If we sum and average these two tensors and assume that $[u_1] = [u_2] = D_c$, we obtain

$$\boldsymbol{m} = D_c \begin{pmatrix} \lambda + \mu & 0 & 0 \\ 0 & \lambda + \mu & 0 \\ 0 & 0 & \lambda \end{pmatrix} \,. \tag{C1}$$

This represents the moment density tensor for the expansion of a cubic element in the two horizontal directions. Let us consider a vertical cylinder of length L and radius R. The cylinder surface at radius R can be regarded as a cylindrical crack, where the radial expansion $D_c = (d_c + \Delta_c)$

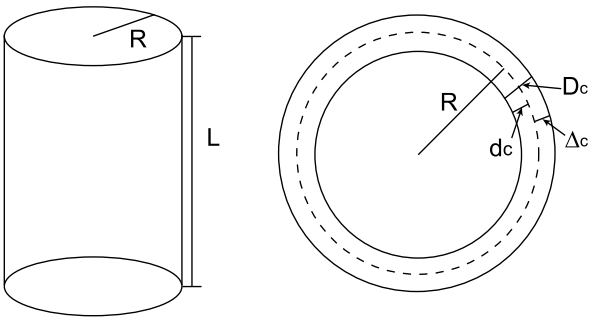

Volcano Seismic Signals, Source Quantification of, Figure A2
A vertical cylinder of length L and radius R. The cylinder surface can be regarded as a cylindrical crack, where the radial expansion $D_c = (d_c + \Delta_c)$ occurs. Here, the inner wall of the crack moves inward by d_c and the outer wall moves outward by Δ_c [89]

occurs (Fig. A2). The moment tensor for the radial expansion of the cylinder may be obtained by integration of the moment density tensor (C1) over the surface R:

$$\boldsymbol{M} = \Delta V \begin{pmatrix} \lambda+\mu & 0 & 0 \\ 0 & \lambda+\mu & 0 \\ 0 & 0 & \lambda \end{pmatrix}, \tag{C2}$$

where ΔV is the volume given as

$$\Delta V = 2\pi RLD_c = 2\pi RL(d_c + \Delta_c)\,. \tag{C3}$$

We obtain the static equilibrium equation for the radial expansion of the cylinder u as

$$\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} - \frac{1}{r^2}u = 0\,. \tag{C4}$$

This equation has internal and external solutions, which are given as $u = ar$ and $u = b/r$, respectively. The constants a and b are determined by the boundary conditions

$$u_e(R) - u_i(R) = \frac{b}{R} - aR = D_c\,, \tag{C5}$$

$$\sigma^i_{rr}(R) - \sigma^e_{rr}(R) = -\frac{2\mu}{R^2}b - 2(\lambda+\mu)a = 0\,, \tag{C6}$$

where superscripts i and e denote the interior and exterior solutions. We then obtain

$$d_c = -u_i(R) = \frac{\mu D_s}{(\lambda+2\mu)}\,, \tag{C7}$$

$$\Delta_c = u_e(R) = \frac{(\lambda+\mu)D_c}{\lambda+2\mu}\,, \tag{C8}$$

and

$$\Delta V = \frac{\lambda+2\mu}{\lambda+\mu}\Delta V_c\,. \tag{C9}$$

The moment tensor for the vertical cylinder is therefore given as

$$\boldsymbol{M} = \frac{\lambda+2\mu}{\lambda+\mu}\Delta V_c \begin{pmatrix} \lambda+\mu & 0 & 0 \\ 0 & \lambda+\mu & 0 \\ 0 & 0 & \lambda \end{pmatrix}. \tag{C10}$$

Let us rotate the vertical cylinder with axis orientation angles ϕ and θ (Fig. 8b). This can be done in the following steps: (1) fix the cylinder and (2) rotate the ξ_1 and ξ_2 axes around the ξ_3 axis through an angle $-\phi$, and (3) further rotate the ξ_1 and ξ_3 axes around the ξ_2 axis through an angle $-\theta$. The rotation matrix $\boldsymbol{R}$ is given as

$$\boldsymbol{R} = \begin{pmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{pmatrix} \begin{pmatrix} \cos\phi & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{C11}$$

$$= \begin{pmatrix} \cos\theta\cos\phi & -\cos\theta\sin\phi & -\sin\theta \\ \sin\phi & \cos\phi & 0 \\ \sin\theta & -\sin\theta\sin\phi & \cos\theta \end{pmatrix}. \tag{C12}$$

Using the matrix $\boldsymbol{R}$, we obtain the moment tensor for a cylinder with the axis orientation angles (ϕ, θ) as

$$\boldsymbol{M}' = \boldsymbol{R}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{R}\,, \tag{C13}$$

which leads to Eq. (34).

Bibliography

Primary Literature

1. Aki K, Richards PG (2002) Quantitative seismology, 2nd ed. University Science Books, Sausalito
2. Aki K, Fehler M, Das S (1977) Source mechanism of volcanic tremor: Fluid-driven crack models and their application to the 1963 Kilauea eruption. J Volcanol Geotherm Res 2:259–287
3. Almendros J, Chouet BA, Dawson PB (2001) Spatial extent of a hydrothermal system at Kilauea Volcano, Hawaii, determined from array analyses of shallow long-period seismicity 1. Method. J Geophys Res 106:13565–13580
4. Almendros J, Chouet BA, Dawson PB (2001) Spatial extent of a hydrothermal system at Kilauea Volcano, Hawaii, determined from array analyses of shallow long-period seismicity 2. Results. J Geophys Res 106:13581–13597
5. Almendros J, Chouet BA, Dawson PB, Huber C (2002) Mapping the sources of the seismic wave field at Kilauea Volcano, Hawaii, using data recorded on multiple seismic antennas. Bull Seism Soc Am 92:2333–2351
6. Almendros J, Chouet BA, Dawson PB, Bond T (2002) Identifying elements of the plumbing system beneath Kilauea Volcano, Hawaii, from the source locations of very-long-period signals Geophys J Int 148:303–312

7. Aoyama H, Takeo M (2001) Wave properties and focal mechanisms of N-type earthquakes at Asama volcano. J Volcanol Geotherm Res 105:163–182
8. Arciniega-Ceballos A, Chouet BA, Dawson PB (1999) Very long-period signals associated with vulcanian explosions at Popocatepetl Volcano, Mexico. Geophys Res Lett 26: 3013–3016
9. Arciniega-Ceballos A, Chouet BA, Dawson PB (2003) Long-period events and tremor at Popocatepetl volcano (1994–2000) and their broadband characteristics. Bull Volcanol 65: 124–135
10. Aster R, Mah S, Kyle P, McIntosh W, Dunbar N, Johnson J, Ruiz M, McNamara S (2003) Very long period oscillations of Mount Erebus Volcano. J Geophys Res 108:2522. doi:10.129/2002JB002101
11. Auger E, D'Auria L, Martini M, Chouet BA, Dawson PB (2006) Real-time monitoring and massive inversion of source parameters of Very-Long-Period (VLP) seismic signals: An application to Stromboli Volcano, Italy. Geophys Res Lett 33:L04301. doi:10.1029/2005GL024703
12. Backus G, Mulcahy M (1976) Moment tensors and other phenomenological descriptions of seismic sources–I. Continuous displacements. Geophys J R Astr Soc 46:341–361
13. Battaglia J, Aki K (2003) Location of seismic events and eruptive fissures on the Piton de la Fournaise volcano using seismic amplitudes. J Geophys Res 108:2364. doi:10.1029/2002JB002193
14. Biot MA (1952) Propagation of elastic waves in a cylindrical bore containing a fluid. J Appl Phys 23:997–1005
15. Biot MA (1952) The interaction of Rayleigh and Stoneley waves in the ocean bottom. Bull Seism Soc Am 42:82–92
16. Bouchon M (1979) Discrete wave number representation of elastic wave fields in three-space dimensions. J Geophys Res 84:3609–3614
17. Bouchon M (1981) A simple method to calculate Green's functions for elastic layered media. Bull Seism Soc Am 71:959–971
18. Bouchon M, Schultz CA, Toksöz MN (1996) Effect of 3D topography on seismic motion. J Geophys Res 101:5835–5846
19. Brandsdottir B, Einarsson P (1979) Seismic activity associated with the September 1977 deflation of the Krafla central volcano in northeastern Iceland. J Volcanol Geotherm Res 6: 197–212
20. Chouet BA (1982) Free surface displacements in the near field of a tensile crack expanding in three dimensions. J Geophys Res 87:3868–3872
21. Chouet BA (1985) Excitation of a buried magmatic pipe: A seismic source model for volcanic tremor. J Geophys Res 90:1881–1893
22. Chouet BA (1986) Dynamics of a fluid-driven crack in three dimensions by the finite difference method. J Geophys Res 91:13967–13992
23. Chouet BA (1988) Resonance of a fluid-driven crack: Radiation properties and implications for the source of long-period events and harmonic tremor. J Geophys Res 93:4375–4400
24. Chouet BA (1992) A seismic model for the source of long-period events and harmonic tremor. In: Gasparini P, Scarpa R, Aki K (eds) Volcanic Seismology. Springer, Berlin, pp 133–156
25. Chouet BA (1996) New methods and future trends in seismological volcano monitoring. In: Scarpa R, Tilling RI (eds) Monitoring and Mitigation of Volcano Hazards. Springer, New York, pp 23–97
26. Chouet BA (1996) Long-period volcano seismicity: its source and use in eruption forecasting. Nature 380:309–316
27. Chouet BA (2003) Volcano Seismology. Pure Appl Geophys 160:739–788
28. Chouet BA, Julian B (1985) Dynamics of an expanding fluid-filled crack. J Geophys Res 90:11187–11198
29. Chouet BA, Page RA, Stephens CD, Lahr JC, Power JA (1994) Precursory swarms of long-period events at Redoubt Volcano (1989–1990), Alaska: Their origin and use as a forecasting tool. J Volcanol Geotherm Res 62:95–135
30. Chouet BA, Saccorotti G, Martini M, Dawson PB, De Luca G, Milana G, Scarpa R (1997) Source and path effects in the wavefields of tremor and explosions at Stromboli volcano, Italy. J Geophys Res 102:15129–15150
31. Chouet BA, Saccorotti G, Dawson PB, Martini M, Scarpa R, De Luca G, Milana G, Cattaneo M (1999) Broadband measurements of the sources of explosions at Stromboli Volcano, Italy. Geophys Res Lett 26:1937–1940
32. Chouet BA, Dawson PB, Ohminato T, Martini M, Saccorotti G, Guidicepierto F, De Luca G, Milana G, Scarpa R (2003) Source mechanisms of explosions at Stromboli Volcano, Italy, determined from moment-tensor inversions of very-long-period data. J Geophys Res 108:2331. doi:10.1029/2003JB002535
33. Chouet BA, Dawson PB, Arciniega-Ceballos A (2005) Source mechanism of Vulcanian degassing at Popocatepetl Volcano, Mexico, determined from waveform inversions of very long period signals. J Geophys Res 110:B07301. doi:10.1029/2004JB003524
34. Chouet BA, Dawson PB, Nakano M (2006) Dynamics of diffusive bubble growth and pressure recovery in a bubbly rhyolitic melt embedded in an elastic solid. J Geophys Res 111:B07310. doi:10.1029/2005JB004174
35. Crosson RS, Bame DA (1985) A spherical source model for low frequency volcanic earthquakes. J Geophys Res 90:10237–10247
36. Dawson PB, Dietel C, Chouet BA, Honma K, Ohminato T, Okubo P (1998) A digitally telemetered broadband seismic network at Kilauea Volcano, Hawaii. US Geol Surv Open File Report 98–108:1–121
37. Dawson PB, Whilldin D, Chouet BA (2004) Application of near real-time semblance to locate the shallow magmatic conduit at Kilauea Volcano, Hawaii. Geophys Res Lett 31:L21606. doi:10.1029/2004GL021163
38. De Angelis S, McNutt SR (2005) Degassing and hydrothermal activity at Mt. Spurr, Alaska during the summer of 2004 inferred from the complex frequencies of long-period events. Geophys Res Lett 32:L12312. doi:10.1029/2005GL022618
39. Ferrazzini V, Aki K (1987) Slow waves trapped in a fluid-filled infinite crack: Implication for volcanic tremor. J Geophys Res 92:9215–9223
40. Fujita E, Ida Y (2003) Geometrical effects and low-attenuation resonance of volcanic fluid inclusions for the source mechanism of long-period earthquakes. J Geophys Res 108:2118. doi:10.1029/2002JB001806
41. Fujita E, Ida Y, Oikawa J (1995) Eigen oscillation of a fluid sphere and source mechanism of harmonic volcanic tremor. J Volcanol Geotherm Res 69:365–378
42. Fujita E, Ukawa M, Yamamoto E (2004) Subsurface cyclic magma sill expansions in the 2000 Miyakejima volcano eruption: Possibility of two-phase flow oscillation. J Geophys Res 109:B04205. doi:10.1029/2003JB002556

43. Fukuyama E, Kubo A, Kawai H, Nonomura K (2001) Seismic remote monitoring of stress field. Earth Planets Space 53: 1021–1026
44. Gil Cruz F, Chouet BA (1997) Long-period events, the most characteristic seismicity accompanying the emplacement and extrusion of a lava dome in Galeras Volcano, Colombia, in 1991. J Volcanol Geotherm Res 77:121–158
45. Goldstein P, Chouet BA Array measurements and modeling of sources of shallow volcanic tremor at Kilauea volcano, Hawaii. J Geophys Res 99:2637–2652
46. Hayashi Y, Morita Y (2003) An image of a magma intrusion process inferred from precise hypocenter migrations of the earthquake swarm eat of the Izu Peninsula. Geophys J Int 153:159–174
47. Hidayat D, Voight B, Langston C, Ratdomopurbo A, Ebeling C (2000) Broadband seismic experiment at Merapi volcano, Java, Indonesia: very-long-period pulses embedded in multiphase earthquakes. J Volcanol Geotherm Res 100: 215–231
48. Hidayat D, Chouet BA, Voght B, Dawson P, Ratdomopurbo A (2002) Source mechanism of very-long-period signals accompanying dome growth activity at Merapi volcano, Indonesia. Geophys Res Lett 23:2118. doi:10.1029/2002GL015013
49. Hill DP, Dawson PB, Johnston MJS, Pitt AM (2002) Very-long-period volcanic earthquakes beneath Mammoth Mountain, California. Geophys Res Lett 29:1370 doi:10.1029/2002GL014833
50. Hori S, Fukao Y, Kumazawa M, Furumoto M, Yamamoto A (1989) A new method of spectral analysis and its application to the earth's free oscillations: the "Sompi" method. J Geophys Res 94:7535–7553
51. Iguchi M (1994) A vertical expansion source model for the mechanisms of earthquakes originated in the magma conduit of an andesitic volcano: Sakurajima, Japan. Bull Volcanol Soc Jpn 39:49–67
52. Jousset P, Neuberg JW, Sturton S (2003) Modelling the time-dependent frequency content of low-frequency volcanic earthquakes. J Volcanol Geotherm Res 128:201–223
53. Julian BR (1994) Volcanic tremor: Nonlinear excitation by fluid flow. J Geophys Res 99:11859–11877
54. Julian BR, Miller AD, Foulger GR (1998) Non-double-couple earthquakes 1. Theory. Rev Geophys 36:525–549
55. Kanamori H, Given J (1982) Analysis of long-period seismic waves excited by the May 18, 1980, eruption of Mount St. Helens – A terrestrial monopole. J Geophys Res 87: 5422–5432
56. Kanamori H, Given J, Lay T (1984) Analysis of seismic body waves excited by the Mount St. Helens eruption of May 18, 1980. J Geophys Res 89:1856–1866
57. Kaneshima S et al (1996) Mechanism of phreatic eruptions at Aso volcano inferred from near-field broadband observations. Science 273:642–645
58. Kawakatsu H, Ohminato T, Ito H, Kuwahara Y, Kato T, Tsuruga K, Honda S, Yomogida K (1992) Broadband seismic observation at Sakurajima Volcano, Japan. Geophys Res Lett 19: 1959–1962
59. Kawakatsu H, Ohminato T, Ito H (1994) 10s-period volcanic tremors observed over a wide area in southwestern Japan. Geophys Res Lett 21:1963–1966
60. Kawakatsu H et al (2000) Aso-94: Aso seismic observation with broadband instruments. J Volcanol Geotherm Res 101:129–154
61. Kawakatsu H, Yamamoto M (2007) Volcano Seismology. In: Kamanori H (ed) Treatise on Geophysics. Earthquake Seismology, vol 4. Elsevier, Amsterdam, pp 389–420
62. Kawase H (1988) Time domain response of a semi-circular canyon for incident SV, P and Rayleigh waves calculated by the discrete wavenumber boundary element method. Bull Seism Soc Am 78:1415–1437
63. Kennet LN, Kerry NJ (1979) Seismic waves in a stratified half-space. Geophys J R Astr Soc 57:557–583
64. Kobayashi T, Ohminato T, Ida Y (2003) Earthquakes series preceding very long period seismic signals observed during the 2000 Miyakejima volcanic activity. Geophys Res Lett 30:1423. doi:10.1029/2002GL016631
65. Kumagai H (2006) Temporal evolution of a magmatic dike system inferred from the complex frequencies of very long period seismic events. J Geophys Res 111:B06201. doi:10.1029/2005JB003881
66. Kumagai H, Chouet BA (1999) The complex frequencies of long-period seismic events as probes of fluid composition beneath volcanoes. Geophys J Int 138:F7–F12
67. Kumagai H, Chouet BA (2000) Acoustic properties of a crack containing magmatic or hydrothermal fluids. J Geophys Res 105:25493–25512
68. Kumagai H, Chouet BA (2001) The dependence of acoustic properties of a crack on the mode and geometry. Geophys Res Lett 28:3325–3328
69. Kumagai H, Ohminato T, Nakano M, Ooi M, Kubo A, Inoue H, Oikawa J (2001) Very-long-period seismic signals and caldera formation at Miyake Island, Japan. Science 293:687–690
70. Kumagai H, Chouet BA, Nakano M (2002) Temporal evolution of a hydrothermal system in Kusatsu-Shirane Volcano, Japan, inferred from the complex frequencies of long-period events. J Geophys Res 107:2236. doi:10.1029/2001JB000653
71. Kumagai H, Chouet BA, Nakano M (2002) Waveform inversion of oscillatory signatures in long-period events beneath volcanoes. J Geophys Res 107:2301. doi:10.1029/2001JB001704
72. Kumagai H, Miyakawa K, Negishi H, Inoue H, Obara K, Suetsugu D (2003) Magmatic dike resonances inferred from very-long-period seismic signals. Science 299:2058–2061 (10.1126/science.1081195)
73. Kumagai H, Chouet BA, Dawson PB (2005) Source process of a long-period event at Kilauea Volcano, Hawaii. Geophys J Int 161:243–254
74. Kumagai H et al (2007) Enhancing volcano-monitoring capablities in Ecuador. Eos trans AGU 88:245–246
75. Kumazawa M, Imanishi Y, Fukao Y, Furumoto M, Yamamoto A (1990) A theory of spectral analysis based on the characteristic property of a linear dynamic system. Geophys J Int 101:613–630
76. Legrand D, Kaneshima S, Kawakatsu H (2000) Moment tensor analysis of near field broadband waveforms observed at Aso volcano, Japan. J Volcanol Geotherm Res 101:155–169
77. Lesage P, Glangeaud F, Mars J (2002) Applications of autoregressive models and time-frequency analysis to the study of volcanic tremor and long-period events. J Volcanol Geotherm Res 114:391–417
78. Lin CH, Konstantinou KI, Liang WT, Pu HC, Lin YM, You SH, Huang YP (2005) Preliminary analysis of volcanoseismic sig-

nals recorded at the Tatun Volcano Group, northern Taiwan. Geophys Res Lett 32:L10313. doi:10.1029/2005GL022861
79. Matsuura T, Imanishi Y, Imanari M, Kumazawa M (1990) Application of a new method of high-resolution spectral analysis, "Sompi", for free induction decay of nuclear magnetic resonance. Appl Spectrosc 44:618–626
80. McNutt SR (1996) Seismic monitoring and eruption forecasting of volcanoes: A review of the state-of-the art and cased histories. In: Scarpa R, Tilling RI (eds) Monitoring and Mitigation of Volcano Hazards. Springer, New York, pp 99–146
81. McNutt SR (2005) Volcanic Seismology. Annu Rev Earth Planet Sci 33:461–491
82. Métaxian JP, Lesage P, Dorel J (1997) Permanent tremor of Masaya volcano, Nicaragua: wavefield analysis and source location. J Geophys Res 102:22529–22545
83. Menke W (1989) Geophysical data analysis: Discrete inverse theory, Revised edition. Academic Press, San Diego
84. Minakami T (1974) Seismology of volcanoes in Japan. In: Civetta L, Gasparini P, Luongo G, Rapolla A (eds) Physical Volcanology. Elsevier, Amsterdam, pp 1–27
85. Molina I, Kumagai H, Yepes H (2004) Resonances of a volcanic conduit triggered by repetitive injections of an ash-laden gas. Geophys Res Lett 31:L03603. doi:10.1029/2003GL018934
86. Morita Y, Nakao S, Hayashi Y (2006) A quantitative approach to the dike intrusion process inferred from a joint analysis of geodetic and seismological data for the 1998 earthquake swarm off the east coast of Izu Peninsula, central Japan. J Geophys Res 111:B06208. doi:10.1029/2005JB003860
87. Morrissey M, Chouet BA (1997) A numerical investigation of choked flow dynamics and its application to the triggering mechanism of long-period events at Redoubt Volcano, Alaska. J Geophys Res 102:7965–7983
88. Morrissey M, Chouet BA (2001) Trends in long-period seismicity related to magmatic fluid compositions. J Volcanol Geotherm Res 108:265–281
89. Müller G (2001) Volume change of seismic sources from moment tensors. Bull Seism Soc Am 91:880–884
90. Nakamichi H, Hamaguchi H, Tanaka S, Ueki S, Nishimura T, Hasegawa A (2003) Source mechanisms of deep and intermediate-depth low-frequency earthquakes beneath Iwate volcano, northeastern Japan. Geophys J Int 154:811–828
91. Nakano M, Kumagai H (2005) Response of a hydrothermal system to magmatic heat inferred from temporal variations in the complex frequencies of long-period events at Kusatsu-Shirane Volcano, Japan. J Volcanol Geotherm Res 147: 233–244
92. Nakano M, Kumagai H (2005) Waveform inversion of volcano-seismic signals assuming possible source geometries. Geophys Res Lett 32:L12302. doi:10.1029/2005GL022666
93. Nakano M, Kumagai H, Kumazawa M, Yamaoka K, Chouet BA (1998) The excitation and characteristic frequency of the long-period volcanic event: An approach based on an inhomogeneous autoregressive model of a linear dynamic system. J Geophys Res 103:10031–10046
94. Nakano M, Kumagai H, Chouet BA (2003) Source mechanism of long-period events at Kusatsu-Shirane Volcano, Japan, inferred from waveform inversion of the effective excitation functions. J Volcanol Geotherm Res 122:149–164
95. Nakano M, Kumagai H, Chouet BA, Dawson PB (2007) Waveform inversion of volcano-seismic signals for an extended source. J Geophys Res 112:B02306. doi:10.1029/2006JB004490
96. Neuberg JW (2000) Characteristics and causes of shallow seismicity in andesite volcanoes. Phils Trans R Soc Lond A 358:1533–1546
97. Neuberg JW, Luckett R, Ripepe M, Braun T (1994) Highlights from a seismic broadband array on Stromboli Volcano. Geophys Res Lett 21:749–752
98. Neuberg JW, Luckett R, Baptie B, Olsen K (2000) Models of tremor and low-frequency earthquake swarms on Montserrat. J Volcanol Geotherm Res 101:83–104
99. Neuberg JW, Tuffen H, Collier L, Green D, Powell T, Dingwell D (2006) The trigger mechanism of low frequency earthquakes on Montserrat. J Volcanol Geotherm Res 153:37–50
100. Nishimura T (2004) Pressure recovery in magma due to bubble growth. Geophys Res Lett 31:L12613. doi:10.1029/2004GL019810
101. Nishimura T, Chouet BA (2003) A numerical simulation of magma motion, crustal deformation, and seismic radiation asscoiated with volcanic eruptions. Geophys J Int 153: 699–718
102. Nishimura T, Nakamichi H, Tanaka S, Sato M, Kobayashi T, Ueki S, Hamaguchi H, Ohtake M, Sato H (2000) Source process of very long period seismic events associated with the 1998 activity of Iwate Volcano, northeastern Japan. J Geophys Res 105:19135–19417
103. Nishimura T, Ueki S, Yamawaki T, Tanaka S, Hashino H, Sato M, Nakamichi H, Hamaguchi H (2002) Broadband seismic signals associated with the 2000 volcanic unrest of Mount Bandai, northeastern Japan. J Volcanol Geotherm Res 119:51–59
104. O'Brien GS, Bean CJ (2004) A 3D discrete numerical elastic lattice method for seismic wave propagation in heterogeneous media with topography. Geophys Res Lett 31:L14608. doi:10.1029/2004GL020069
105. Ohminato T (2006) Characteristics and source modeling of broadband seismic signals associated with the hydrothermal system at Satsuma-Iwojima volcano, Japan. J Volcanol Geotherm Res 158:467–490
106. Ohminato T, Chouet BA (1997) A free-surface boundary condition for including 3D topography in the finite difference method. Bull Seism Soc Am 87:494–515
107. Ohminato T, Chouet BA, Dawson PB, Kedar S (1998) Waveform inversion of very long period impulsive signals associated with magmatic injection beneath Kilauea Volcano, Hawaii. J Geophys Res 103:23839–23862
108. Ohminato T, Takeo M, Kumagai H, Yamashina T, Oikawa J, Koyama E, Tsuji H, Urabe T (2006) Vulcanian eruptions with dominant single force components observed during the Asama 2004 volcanic activity in Japan. Earth Planets Space 58:583–593
109. Ripperger J, Igel H, Wasserman J (2003) Seismic wave simulation in the presence of real volcano topography. J Volcanol Geotherm Res 128:31–44
110. Rowe CA, Aster RC, Kyle PR, Schlue JW, Dibble RR (1998) Broadband recording of Strombolian explosions and associated very-long-period seismic signals on Mount Erebus volcano, Ross Island, Antarctica. Geophys Res Lett 25: 2297–2300
111. Rubin AM, Gillard D (1998) Dike-induced seismicity: theoretical considerations. J Geophys Res 103:10017–10030
112. Rubin AM, Gillard D, Got JL (1998) A reinterpretation of seis-

micity associated with the January 1983 dike intrusion at Kilauea Volcano, Hawaii. J Geophys Res 103:10003–10015
113. Saccorotti G, Del Pezzo E (2000) A probabilistic approach to the inversion of data from a seismic array and its application to volcanic signals. Geophys J Int 143:249–261
114. Saccorotti G, Chouet BA, Dawson PB (2001) Wavefield properties of a shallow long-period event and tremor at Kilauea Volcano, Hawaii. J Volcanol Geotherm Res 109:163–189
115. Sakuraba A, Oikawa J, Imanishi Y (2002) Free oscillations of a fluid sphere in an infinite elastic medium and long-period volcanic earthquakes. Earth Planets Space 54:91–106
116. Stump BW, Johnson LR (1977) The determination of source properties by the linear inversion of seismograms. Bull Seism Soc Am 67:1489–1502
117. Sturton S, Neuberg JW (2006) The effects of conduit length and acoustic velocity on conduit resonance: Implications for low-frequency events. J Volcanol Geotherm Res 151: 319–339
118. Takei Y, Kumazawa M (1994) Why have the single force and torque been excluded from seismic source models? Geophys J Int 118:20–30
119. Takei Y, Kumazawa M (1995) Phenomenological representation and kinematics of general seismic sources including the seismic vector modes. Geophys J Int 121:641–662
120. Takeo M, Yamasato H, Furuya I, Seino M (1990) Analysis of long-period seismic waves excited by the November 1987 eruption of Izu-Oshima Volcano. J Geophys Res 95:19377–19393
121. Takeuchi H, Saito M (1972) Seismic surface waves. In: Bolt BA (ed) Methods in Computational Physics, vol 11. Academic Press, New York, pp 217–295
122. Tameguri T, Iguchi M, Ishihara K (2002) Mechanism of explosive eruptions from moment tensor analyses of explosion earthquakes at Sakurajima volcano, Japan. Bull Volcanol Soc Jpn 47:197–216
123. Toda S, Stein R, Sagiya T (2002) Evidence from the AD 2000 Izu islands earthquake swarm that stressing rate governs seismicity. Nature 419:58–61
124. Tolstoy I (1954) Dispersive properties of a fluid layer overlying a semi-infinite elastic solid. Bull Seism Soc Am 44:493–512
125. Tuffen H, Dingwell D (2005) Fault textures in volcanic conduits: evidence for seismic trigger mechanisms during silicic eruptions. Bull Volcanol 67:370–387
126. Uhira K, Takeo M (1994) The source of explosive eruptions of Sakurajima volcano, Japan. J Geophys Res 99:17775–17789
127. Ulrych TJ, Clayton RW (1976) Time series modeling and maximum entropy. Phys Earth Planet Inter 12:188–200
128. Ulrych TJ, Sacchi MD (1995) Sompi, Pisarenko and the extended information criterion. Geophys J Int 122:719–724
129. Virieux J (1986) P-SV wave propagation in heterogeneous media: Velocity-stress finite-difference method. Geophysics 51:889–901
130. Wessel P, Smith WHF (1998) New improved version of Generic Mapping Tools released. Eos trans AGU 122:149–164
131. Yamamoto M (2005) Volcanic fluid system inferred from broadband seismic signals. Ph D Thesis, University of Tokyo
132. Yamamoto M, Kawakatsu H, Kaneshima S, Mori T, Tsutsui T, Sudo Y, Morita Y (1999) Detection of a crack-like conduit beneath active crater at Aso volcano, Japan. Geophys Res Lett 26:3677–3680
133. Yamamoto M, Kawakatsu H, Yomogida K, Koyama J (2002) Long-period (12 sec) volcanic tremor observed at Usu 2000 eruption: Seismological detection of a deep magma plumbing system. Geophys Res Lett 29:1329. doi:10.1029/2001GL013996
134. Yamamura K, Kawakatsu H (1998) Normal-mode solutions for radiation boundary conditions with an impedance contrast. Geophys J Int 134:849–855

Books and Reviews

Dahlen FA, Tromp J (1998) Theoretical Global Seismology. Princeton University Press, Princeton
Kay SM, Marple SL (1981) Spectrum analysis – A modern perspective. Proc IEEE 69:1380–1419

Voting

Alvaro Sandroni, Jonathan Pogach, Michela Tincani, Antonio Penta, Deniz Selman
University of Pennsylvania, Philadelphia, USA

Article Outline

Glossary
Definition of the Subject
Introduction
The Collective Choice Problem
Voting Rules
Welfare Economics
Arrow's Impossibility Theorem
Political Ignorance and the Condorcet Jury Theorem
Gibbard–Satterthwaite Theorem
Political Competition and Strategic Voting
The Common Value Setting with Strategic Agents
Future Directions
Bibliography

Glossary

Arrow's impossibility theorem Arrow's Impossibility Theorem states that there does not exist a complete social ranking over alternatives that meets minimum impositions of egalitarianism and efficiency, No Dictatorship and Pareto Optimality, respectively. Consequently, there is no voting mechanism that can simultaneously satisfy basic notions of egalitarianism and efficiency.

Cost of voting Any sacrifice in utility that voting entails, such as cognitive costs, time cost of going to the polls, etc.

Collective or social choice problem A collective or social choice problem is a setting in which a group of individuals must jointly decide on a single alternative from

a set. The outcome of such a problem potentially affects the welfare of all individuals.

Common values A common values problem is one in which agents share preferences over the alternatives, but may differ on their information regarding the attributes of alternatives.

Condorcet jury theorem In a common values setting, the result that under majority voting the correct candidate will (almost always) be elected so long as the following assumptions are satisfied: (1) each voter's belief is correct with a probability higher than half, (2) each voter votes according to her belief and (3) there is a large number of voters. In political science this result is sometimes referred to as *wisdom of crowds*.

Condorcet paradox See voting cycle.

Downsian model of political competition A model in which candidate strategically position themselves on a unidimensional policy space in order to win an election.

Efficiency Efficiency is a broad criterion that may be used to evaluate the social value of alternative outcomes by demanding that society make use of all valuable resources. Examples of efficiency measures are utilitarianism and Pareto Optimality.

Egalitarianism Egalitarianism is a broad criterion used to evaluate the social value of alternative outcomes by demanding that welfare or resources are evenly distributed across the population.

Gibbard–Satterthwaite theorem The Gibbard–Satterthwaite Theorem states that, under some assumptions, every non-dictatorial social choice function is manipulable.

Majority voting Majority voting is a voting rule that stipulates that each agent vote for a single alternative and an alternative that receives more than half of all votes is the collective choice.

Manipulability of a social choice function A social choice function is said to be manipulable if there is some agent who, given the social choice function, prefers to misreport his true preferences. In a voting context, this translates to voting for an alternative different from that which is most preferred.

Median voter theorem The median voter theorem states that if individuals' preferences are single peaked, then there exists an alternative that beats all others in a pairwise majority vote. Single peaked requires that each individual has a bliss point (most preferred alternative) and alternatives are less preferred the further they are from the bliss point. The selected alternative is then the median bliss point and the voter who has this bliss point is the median voter.

No Dictatorship The No Dictatorship criterion of Arrow's desiderata demands that there is no single individual whose preferences always determine those of society. Any egalitarian arrangement must satisfy the no dictatorship criterion, though arrangements that satisfy no dictatorship need not be egalitarian.

Pairwise majority voting rule A pairwise majority voting rule compares each pair of alternatives in a majority vote. Depending on agents preferences and votes, this rule may lead to a voting cycle.

Paradox of voting the puzzle of why there is high voter turnout in large elections, when the probability of any single vote to be determinant for the outcome is extremely small.

Pareto optimality or pareto efficiency An outcome is Pareto Optimal or Pareto Efficient if it is not possible to increase the welfare of one individual without lessening the welfare of another.

Political ignorance A state in which voters are not well informed about the issues and/or candidates they must vote on.

Social choice function A social choice function is a mapping of all individuals' preferences into an alternative, the social choice.

Strategic abstention The tactical decision of an uninformed citizen to abstain in order to allow informed citizens with the same preferences determine the outcome.

Strategic voting Voting by agents who aim to maximize their utility and might do so by misreporting their true preferences over electoral outcomes.

Utilitarianism Utilitarianism is a conception of efficiency that evaluates outcomes on aggregate utility.

Utility of voting Benefit to citizens from voting, usually divided into two components: a **non instrumental** component which includes utility derived from the mere act of voting and not related to the actual outcome of the election, and an **instrumental** component given by the utility of the outcome a voter would induce if determining the outcome, weighted by the probability that his vote determines the outcome.

Voting cycle or Condorcet's paradox A voting cycle or Condorcet's Paradox results when every feasible alternative is beaten by another in a pairwise majority vote. As such, any collective choice is less preferred to some other alternative by more than half of the population.

Voting rule A voting rule is a mapping of votes into a collective choice. Examples of different voting rules include majority voting and plurality voting. The collective choice may vary under alternative voting rules.

Definition of the Subject

Voting is a fundamental mechanism that individuals use to reach an agreement on which one of many alternatives to implement. The individuals might all be affected by the outcome of such a process and might have conflicting preferences and/or information over alternatives. In a voting mechanism, preferences and information are aggregated as individuals submit votes and a voting rule maps the compilation of votes into the alternative that is to be selected. The use of voting as a means of making a group decision dates back at least to ancient Greece, though French Revolutionary contemporaries Condorcet's and Borda's works are among the pioneers of voting theory. Meanwhile, welfare economists such as Bentham suggested formal definitions of socially desirable outcomes. As voting theory and welfare economics evolved, Arrow's result in the middle twentieth century showed that no mechanism, voting or otherwise, can produce outcomes consistent with some of welfare economists' definitions of socially desirable states. Moreover, results in the early 1970s suggested that voters have incentives to misrepresent their true preferences in elections. Contemporary voting theory has developed new models of strategic behavior to address questions on how political agents behave and which outcomes voting might produce.

Introduction

Voting is one of the most commonly used ways of making collective decisions. Two issues arise when a group of people must find an agreement on a choice that will potentially affect the welfare of all. First, individuals might have conflicting preferences over the set of alternatives they are choosing from. An interesting question is what is the best way to aggregate individual's preferences in order to reach a common decision when preferences conflict. Second, even if agents share the same preferences, they might possess different information about the alternatives. In this situation, an interesting question is what is the best way to aggregate people's conflicting information so as to make the right choice.

The optimal ways to aggregate preferences and information are among the most important questions that the literature on voting has tried to answer. Since voting concerns societies, as opposed to single individuals, this literature relies on works in the field of welfare economics. This branch of economics is primarily concerned with the analysis and the definition of the welfare of a society. There is no agreement among social scientists on a single definition of welfare of a society. However, two concepts are prominent: efficiency, which concerns the minimal waste scarce resources, and egalitarianism, which concerns the equal distribution of those resources. Using these concepts to define the welfare of a society, social scientists have attempted to answer questions on preference and information aggregation in making a collective choice.

Early works on different voting techniques date back to at least the late eighteenth century, with the works of French mathematicians such as Borda [6] and Condorcet [9]. Their works were among the first to analyze voting procedures with formal mathematical tools, which have since been prominently used in the literature on voting. Continuing in this tradition, a fundamental result in voting theory is Arrow's Impossibility Theorem. This is a formal treatment of the problem of aggregation of preferences that reached a striking result: there is no way, under some assumptions, to aggregate individuals' preferences that is minimally egalitarian and minimally efficient.

Unlike the problem of aggregation of conflicting preferences, studies on the aggregation of conflicting information obtain positive results. In particular, the Condorcet Jury Theorem finds a mathematical justification for the phenomenon that is known in Political Science as the *wisdom of the crowds*, that is, the observation that democracies seem to be better at making decisions than single individuals. However Condorcet's result relies on the assumption that people vote sincerely and subsequent works on voters' behavior suggest that this is not always the case. A fundamental theoretical result, known as Gibbard–Satterthwaite theorem, shows formally that, under some assumptions, in every election at least one individual has an incentive to vote non sincerely, i. e. to vote strategically.

Gibbard and Satterthwaite's result is a starting point for a branch of the voting literature that deals with strategic voting. Numerous works analyze strategic voting though the use of mathematical tools such as Game Theory. The latter has been used not only to explain voters' behavior, but also to describe competing candidates' behaviors in an election. The research in this game theoretic literature focuses on the outcome of political competition and on the development of a theory of turnout. The former is analyzed through the use of a model of political competition. A main result is that in a two party election, both parties choose the same political platform in order to maximize the probability of winning an election. The latter is motivated by the paradox of voting, which refers to the empirical observation that citizens vote even when the probability that their vote determines the outcome of the election is negligible, such as in large elections. At present, there is no widely accepted theory to explain this phenomenon.

As with the paradox of voting, the voting literature has still many unanswered interesting questions. As will be mentioned in the section on future directions, the voter's behavior is still only partially understood and much needs to be explained. On a broader level, the study of the historical evolution of democracies needs further developments. Much has been written on the subject of voting, and some interesting results have been obtained, but much still needs to be explained.

The Collective Choice Problem

The basic framework for understanding voting is a *collective* or *social choice problem*: a group of agents must reach an agreement on which alternative to select, and this decision potentially affects the welfare of everyone. Such problems could range from "what should be taught in public schools?" to "who should be president?" Consider two possible alternatives, A and B, and a group of individuals who must jointly choose one of the two. It might be the case that some individuals in society prefer A and some prefer B. These conflicting preferences pose a challenge in determining the appropriate social choice. Alternatively, consider a scenario where A and B are different characteristics that two candidates running for a public office may possess. Suppose that all individuals agree that A is more desirable than B, i. e. this a *common values* setting. However, individuals differ in their information; some think that candidate 1 possesses trait A, while others think candidate 2 does. In this case, the challenge is finding the best way to balance the conflicting information in arriving to a collective choice.

There are several ways to resolve the problems of conflicting preferences and information. For example, people can bargain to reach an agreement, or they can fight. A collective choice problem might also be solved through a dictatorship, or even through the toss of a coin. A fundamental mechanism used to resolve these conflicts is an election in which people submit votes on the feasible alternatives and a *voting rule* maps the collection of votes into a social choice. As can be seen in the next section, there exists a multitude of voting rules.

Before discussing voting rules, one should note the following distinction between elections: those in which the alternatives are policy and those in which the alternatives are candidates. The former is known as *direct democracy*, a common example of which is a *referendum*. In a referendum, citizens vote on a particular proposal, such as the adoption of a new law. The outcome of such an election is then to implement or not implement the proposal. This is in contrast to the case, known as *representative democracy*, where the election is over candidates. The collective choice in such a system is an agent or group of agents with the responsibility of choosing policy. The following voting rules apply to both situations.

Voting Rules

Majority voting is one of the most commonly used voting rules. The rule prescribes that each citizen vote for a single alternative and an alternative becomes the social choice if it receives more than half of all votes. Clearly, when there are more than two alternatives, majority voting does not necessarily produce a social choice.

To ensure a comparison between alternatives, one can resort to *pairwise majority voting rule*, in which alternatives are voted over pair by pair with a majority vote. That is, a majority vote is held between every pair of feasible alternatives and for each majority vote, the winner is deemed socially preferable to the loser. However, this voting rule might generate an intransitive social preference in which society chooses x to y, y to z, but z to x. Consider a three-individual committee comprised of voters 1, 2 and 3 who must choose one of three alternatives, x, y and z. Individual preferences are such that voter 1 prefers x to y to z, voter 2 prefers z to x to y, and voter 3 preferences y to z to x. By pairwise majority voting x beats y, which in turn beats z, which in turn beats x. This intransitivity over alternatives is known as a *voting cycle* or *Condorcet's Paradox*.

In the presence of a voting cycle, pairwise majority voting might not produce an overall winner, as each alternative might be beaten by another. So, an agenda setter could end a cycle by specifying the order in which the alternatives are to be compared in a pairwise vote. However, pairwise majority voting then places all the decision power in the hands of the agenda setter. Returning to the previous example, suppose the agenda prescribes that voting be carried out between alternatives x and y first and the winner is then to be compared with z. In the first round of voting x beats y and z then beats x, so z becomes collective choice. However, the agenda setter could instead choose an initial comparison between y and z. Since y beats z in the first round and x beats y in the second, x would then be the collective choice. Hence, the agenda setter decides the outcome of the election by choosing the order of the pairwise voting.

A *plurality voting rule* is an alternative way to make a collective choice: agents each vote for one alternative and the alternative with the most votes is chosen.

Other voting rules require agents to submit scores or rankings of all available alternatives, rather than just vot-

ing for a single one. In a *Borda Count*, agents rank all alternatives assigning the larger numbers to those that are more preferred. The voting rule sums the scores for each alternative across individuals and the alternative with the highest sum is the social choice.

A *supramajority* voting rule stipulates that the 'status quo' alternative is chosen unless another alternative receives at least some specified percentage of the vote larger than fifty percent. In the limit, there might be a *unanimity rule* that mandates one hundred percent of the electorate vote for an alternative for it to be chosen against the status quo. Examples of supramajority rules include the passing of constitutional amendments in the United States, where the current constitution is the status quo. Unanimity rules are commonly found in the judicial system in which all jurors must agree on the guilt of the defendant to override the status quo, the presumption of innocence.

Voting rules might also grant veto power to one or more agents. For example, each of the five permanent members of the fifteen member United Nations Security Council has the power to veto resolutions on particular matters. Any collective choice must therefore have the approval of all five permanent members.

The different voting rules are not simply different methods of arriving at the same social choice. Rather, the result of an election depends critically on the voting rule that is used. In fact, an alternative that is the social choice according to one voting rule might be the least preferred under another. For instance, consider alternatives x, y, and z and seven voters, three who prefer x to y to z, two who prefer y to z to x, and two who prefer z to y to x. By pairwise majority voting, x loses to both y and z. In contrast, x beats both y and z in a plurality vote. This suggests that in order to determine which voting rule to use, one must carefully analyze the various resulting outcomes.

To this end, it is useful to identify criteria that allow one to discriminate among the different outcomes produced under various voting rules. There are two main methods of doing this: efficiency and egalitarianism.

Welfare Economics

The analysis of efficiency and egalitarianism of a social state is among the objectives of a discipline called welfare economics. The first criterion for efficiency used by welfare economists dates back at least to Jeremy Bentham [2] and is known as utilitarianism. According to utilitarianism, the social interest is judged in terms of the total utility of a community. For example, if by moving from arrangement A to arrangement B Mr 1 benefits more than Ms 2 suffers, then the movement from A to B is judged as a social welfare improvement. Notice that in order to implement this criterion, the satisfaction intensities of different individuals must be comparable. In the 1930s this criterion was criticized by Lionel Robbins [35] and other welfare economists who claimed that the comparison of utilities across individuals has no scientific basis. In the 1940s a new criterion was developed which required no comparison of individual utilities: the Pareto criterion. A social outcome is said to be Pareto Optimal (Pareto Efficient) if there is no other outcome that would benefit at least one individual without hurting anyone else. Consider a scenario where ten dollars must be split among two individuals who value money and there are two alternatives: either person 1 receives five dollars, person 2 four and the remaining dollar is thrown away, or both receive five dollars. Clearly, the first alternative is not Pareto Optimal because person 2 can be made better off and person 1 would remain as well off if 2 is given the dollar that is being thrown away. Notice that the first alternative is also non utilitarian; in fact, the sum total of utilities can not be maximized when valuable resources are thrown away. However, a drawback of this efficiency criterion is that there exist multiple non comparable Pareto Optimal outcomes: any division of the ten dollars among the two individuals is Pareto Optimal as long as no money is thrown away, since to make one person better off one would have to take resources away from the other person. Hence, the Pareto Optimality criterion does not allow one to distinguish among multiple outcomes. Finally, notice that Pareto Optimal outcomes can be extremely non egalitarian: person 1 receiving ten dollars and person 2 zero is an unequal but Pareto Efficient division.

An alternative criterion often used to discriminate among social outcomes is egalitarianism, which focuses on the distribution of welfare across members of a society. One of the abstract principles behind egalitarianism is *the veil of ignorance* [18,19,33]. Consider a situation in which two persons must share a cake. Pareto Optimality does not help in selecting a division: as in the ten dollars example, any division of the cake is Pareto Optimal. Suppose that one of the two persons sharing the cake is asked to cut it in two without knowing a priori which piece she will receive. Her ignorance about which piece she will receive makes her cut the cake in two equal shares, an egalitarian division. Notice that there are a number of ways to define egalitarian outcomes. For example, Rawls's [33] *maximin rule* suggests that the social objective should be to maximize the welfare of the worst-off individual. Finally, notice that an egalitarian outcome might be extremely inefficient. Returning to the ten dollar example, an arrangement where person 1 is given nine dollars and person 2 one dollar is

not as egalitarian as one where both are given two dollars, though the latter does not make use of more than half of the available resources.

Arrow's Impossibility Theorem

Ideally, one would like to use the normative criteria of social efficiency and egalitarianism to discriminate between different voting rules and select the best one. However a general result known as *Arrow's Impossibility Theorem* [1] shows mathematically that this is impossible. In his seminal work, Arrow shows that there is no voting mechanism that generates a social consensus on the ordering of the different alternatives while satisfying a number of axioms, among which are the weakest forms of efficiency and egalitarianism: Pareto Optimality and No Dictatorship. The latter, which states that no individual always determines preferences of society, is a weak form of egalitarianism. While a non-dictatorial society can be quite unequal, any egalitarian society must be non-dictatorial.

A number of possibility results have been obtained by the relaxation of some of Arrow's axioms. For example, pairwise majority voting with a particular restriction on individual tastes, which violates what Arrow called Unrestricted Domain, satisfies Pareto Optimality and No Dictatorship, while generating an ordering of the social alternatives.

Black [4] noticed that pairwise majority voting produces an outcome that is not subject to Condorcet's paradox when individual preferences are single-peaked: every individual must have a most preferred alternative (*bliss point*) and between any two alternatives he prefers the one that is closer to his bliss point. An important result in voting theory, called the median voter theorem, shows that when individuals' preferences satisfy this condition, the bliss point of the *median voter* beats any other alternative by pairwise majority voting. The median voter is found by ordering voters according to their bliss points. The importance of this theorem derives from its ability to describe how democracies work in practice. It is commonly observed that candidates try to appeal to voters who are politically moderate, or "in the middle": this is consistent with the theory, which suggests that these are the preferences that will eventually prevail in a democratic system.

Political Ignorance and the Condorcet Jury Theorem

As mentioned earlier, voting is not only a way to aggregate conflicting preferences but it is also a way to aggregate individual, possibly conflicting, information when preferences are partially or totally aligned. This is the case in *common value* settings. When people would agree on the best choice if given the same information on alternatives, but differ in the information they actually receive, a natural question is which is the voting mechanism that aggregates information in a way that maximizes the probability of the right decision being made. A result called *Condorcet Jury Theorem* [9] shows that among all the possible voting rules, simple majority rule guarantees that the right decision is made under three crucial assumptions: that each voter has a correct belief with a probability higher than 50%, the voter votes according to his belief, and that the electorate is very large. Before going into the details of the theorem, it must be mentioned that this result has a practical importance. A number of works document that voters are ignorant over both policy and candidate alternatives over which they are to vote. [7] claim that "many people know the existence of few if any of the major issues of policy", while [12] discusses "mass political ignorance" and "mass political apathy" as playing key roles throughout the history of American politics. More recently, the *2004 American National Election Study* found that Americans performed extremely poorly when asked simple questions about the political system and the leaders in charge of it. This evidence is in favor of what is called *political ignorance*. In a setting where voters are politically ignorant, the Condorcet Jury Theorem provides a valuable insight as to how much political information matters in determining the outcome.

Consider a committee who has to elect an administrator out of two candidates, one "good" and one "bad." Assume that all the members of the committee share the same preferences: they all prefer the good administrator to be selected. Individuals differ, however, in the information they have about which candidate is the good one. Suppose that each voter has a belief about which is the good candidate and votes according to his belief. If each voter has a correct belief with probability higher than 50%, then by the Law of Large Numbers as the number of voters becomes very large the probability that more than half of the the electorate votes for the right candidate goes to one. Hence, under simple majority the probability that the right choice is made goes to one. Condorcet's conclusion is that in a common value setting a democratic decision is superior to an individual decision, because each voter makes the wrong decision with a non-negligible positive probability, whereas the population as a whole makes the right decision almost always. As far as political ignorance is concerned, this result shows that it is not necessary for an electorate to be well informed for it to make the right decision. So long as each voter has a correct belief with a probability higher than a half, the electoral outcome will almost always be identical to one in which the electorate

was perfectly informed. Therefore, Condorcet's result implies that the ignorance of individual voters is overcome by the aggregation of information in an election.

Gibbard–Satterthwaite Theorem

Thus far, citizens have been treated as if they disregard any tactical considerations when faced with a voting decision. However, a citizen might find it worthwhile to misrepresent his true preferences in order to achieve a social outcome more preferred than the one that would result if he voted naively. Consider an election in which a status quo will be replaced if a simple majority agrees on one of three candidates. Suppose the status quo is a conservative government, and the three alternative candidates to be voted for are a conservative, a moderate and a liberal. Imagine that there are only three voters, and two votes have already been cast: one is for the moderate candidate, one is for the conservative one. Suppose that the last individual who is called to vote is politically liberal. He knows that if he votes for his truly most preferred candidate, the liberal one, there would be a tie and the status quo conservative government would not be replaced. However, by voting for his second most preferred alternative, the moderate candidate, he would break a tie and the status quo government would be replaced with a moderate one. A liberal voter prefers this outcome to the one where conservatives win. Therefore he has an incentive to misrepresent his true preferences and vote tactically for his second-best alternative.

A powerful result in voting theory called Gibbard–Satthertwaite theorem [17,36] shows formally that in most electoral settings at least one citizen has an incentive to vote tactically. Define a *social choice function* as a mapping of all individuals' preferences into a collective choice. The theorem states that there is no social choice function that is Non Dictatorial and Non Manipulable, i. e. such that no agent has an incentive to vote tactically. Consequently, for every voting rule there is at least one agent with an incentive to misrepresent her preferences.

It should be mentioned that the Gibbard–Satterthwaite theorem places also other technical restrictions. Furthermore, in a setting such as a majority vote, a misrepresentation of one's true preferences coincides with voting for an alternative which the voter does not rank top. In the original formulation of their theorem, however, Gibbard and Satterthwaite dealt with mechanisms where agents are required to submit a ranking over all alternatives, and a misrepresentation of tastes in their framework does not coincide necessarily with a misrepresentation of only the most preferred alternative (see [26]).

This result suggests that for a deep understanding of voting one should not focus only on the assumption that citizens vote sincerely. Tactical voting is not only an abstract possibility, it is also an actual behavior that must be considered in any voting analysis. The following sections explore how the literature on voting has dealt with strategic voting.

Political Competition and Strategic Voting

The early works of Downs [10] and Tullock [39] initiated the analysis of political issues within a strategic framework, where voters and/or candidates are assumed to be rational decision makers. *Political competition* describes a situation in which candidates strategically position themselves in order to win an election, whereas *strategic voting* refers to individuals' decision to vote so to maximize utility by sometimes misreporting true preferences. This game theoretical framework developed due to positive arguments such as Gibbard–Satterthwaite theorem, which suggests that voters have an incentive to behave strategically, and is also due to spread of game theory as a dominant tool in economic analysis.

Political Competition

The classic model analyzing the candidates' choice of positioning on the political spectrum is that of Downs [10], who adapted the classical *Hotelling model* [21] to the analysis of the choice of political platforms by candidates. In the *Downsian model of political competition*, there is a unidimensional policy space, representing the political spectrum. There are two candidates who position themselves on this policy space. Each voter has a most preferred point on this space and prefers points closer to this point than those further away, i. e. each voter has single-peaked preferences. Downs argues that strategic candidates concerned only with winning position themselves at the point that is most preferred by the median voter. If candidate A were positioned anywhere else, say to the left of the median voter, candidate B could get the majority of votes by positioning himself between candidate A and the median voter; all the agents to the right of B, constituting more than half of the voters, would prefer B to A. Given this scenario, candidate A (for the same reason) would then have an incentive to position himself between B and the median voter, and so on. Hence, the result is that both candidates position themselves at the policy platform most preferred by the median voter. An interesting implication of this result is that under a democracy with two parties, both parties act identically, and therefore there are only as many positions (just one) taken by political parties as there

would be in a dictatorship. However, it is crucial that two parties exist so that the competition between them can allow the chosen policy point to represent the preferences of the voters. This is in contrast to a dictatorship in which the ruling party can implement its own preferred policy without voter approval.

Notice that the example above assumes a two-party system (for models that allow for more than two parties see [3,29]). [11] posits that in a representative democracy with a plurality voting rule, only two parties compete in the elections. This theory, known as *Duverger's Law*, states that a *proportional representation system*, in which parties gain seats proportional to the number of votes received, fosters elections with numerous parties. In contrast, a plurality system marginalizes smaller parties and results in only two parties entering into political competition.

The Decision to Vote: the Paradox of Voting

Another issue raised by Downs, one which focuses on voters' behavior rather than candidates', is known as *the paradox of voting*. It refers to the fact that in a large election, the probability that any single vote determines the outcome is vanishingly small. If every person only votes for the purpose of influencing the outcome of the election, even a small cost of voting would be sufficient to dissuade anyone from voting. Yet, it is commonly observed that turnout is very high, even in large elections. From the large empirical literature on turnout in elections, some facts seem to be acquired knowledge: (1) turnout is higher in more important elections (e. g. Presidential election in the US have a significantly higher turnout than Gubernatorial elections), (2) turnout is generally higher in close elections (i. e. with smaller margins of victory), and (3) turnout rates are different among groups with different demographic characteristics. For instance, from the thorough work by [41], it emerges that education has a substantial effect on the probability that one will vote. Income has less of an effect once it has been controlled for the impact of other variables. After education, the second most important variable is age, which appears to have a strong positive relationship with turnout. Other socio-economic variables are also important; in particular, racial minorities appear to be less likely to vote. Finally, turnout seems to be significantly influenced by factors such as the weather conditions on the day of the election and voters' distance from the polls (see [8]).

Such comparative statics suggest that it is appropriate to model voters' behavior as a rational choice problem within a standard utility maximization framework. The modern theory of voting applies the classic utilitarian framework to the voting problem, positing that agents decide whether or not to vote by comparing the *cost of voting* with the *utility of voting*. The traditional starting point for the modern theory of voting is [34], who formalize the insights of [10] and [39] in a simple utilitarian model of voting. The *cost of voting* comprises any sacrifice in utility that voting entails. The *utility of voting* is usually divided into two components: a *non-instrumental* component and an *instrumental* component. The *non-instrumental* component includes utility derived from the mere act of voting and not related to the actual outcome of the election. It may include, for instance, the sense of civic duty. There is considerable evidence that voters are motivated by a sense of civic duty (see, for example, [5]). The *instrumental* component is the utility of the outcome a voter induces if her vote determines the outcome, weighted by the probability that her vote actually determines the outcome.

The instrumental component of the utility of voting has attracted most of the attention in the literature. It is typically analyzed through the rational theory of voting, which is motivated by an empirical observation: there exists a strong positive correlation between turnout rate and closeness of the election. This fact suggests that, ceteris paribus, voters are more likely to vote if their vote is more likely to make a difference. The main theoretical problem is to endogenize the probability that each voter is pivotal, i. e. that his vote is determinant for the outcome of the election.

Ledyard [23,24] are among the early works in the literature on game theoretical models of the pivotal-voter. In these models, voters infer the probability of being pivotal from the equilibrium strategies of other voters. Subsequently, they decide whether or not to vote, trading off the cost of voting with the expected (instrumental) utility of voting. Although Ledyard did not focus on the magnitude of turnout in a strategic model, this question is addressed by Palfrey and Rosenthal ([30,31]) who model elections with uncertainty about the total number of voters. Voters strategically choose whether or not to vote for their favorite alternative amongst two candidates. However, Palfrey and Rosenthal's theories do not explain high turnout in large elections, when the cost of voting is not very (and unrealistically) low.

Ultimately, the game-theoretic approach to costly voting could not escape the *paradox of voting*. Since the probability of being pivotal is very small in large elections, the individual incentives to vote cannot justify high turnouts unless the cost of voting is sufficiently small. Conversely, regardless of how small the cost of voting is, the theory posits that there should be low turnout as the election becomes arbitrarily large, which is in contrast to empirical

evidence. The puzzle that remains open is how to reconcile the evidence of high turnout in large elections with the responsiveness of turnout levels to the closeness of the election.

Mobilization and Group-Based Notion of Welfare

Two strands of the literature try to overcome the paradox of voting by focusing on groups of like-minded people rather than on individual agents. These are models of *mobilization* and models incorporating a *group based* notion of welfare.

In models of mobilization, the population of voters is assumed to be divided into groups, each of which has a leader who has the same preferences as all agents in the group and coordinates their behavior. The turnout decision within each group is determined by how the leaders allocate costly resources to voters. It is as if leaders buy the votes of the agents in their group, compensating for the agents' costs of voting. Since leaders influence a large number of voters, their decisions have a non-negligible impact on the probability of affecting the electoral outcome and consequently, on the individual instrumental benefit from voting. [37,38] test group based models and provide some empirical support for the mobilization thesis. (see also [27,28,40]). The problem for models of mobilization is that it is not clear how leaders affect the individual behavior of voters.

Models of *group based welfare* consider groups of like-minded individuals whose actions are intended not to maximize their individual utilities, but rather that of the group (see [22,25]). In this case, there is no leader who prescribes behavior as in mobilization models, but instead there is an implicit understanding among agents in the group on appropriate behavior. This idea is developed by [16], who appeal to [20] *Group Rule-Utilitarian Theory* to endogenize the *non-instrumental component* of the utility from voting in a way that preserves the positive relation between closeness of the election and incentive to vote typical of the classic *pivotal-voter models*. In this model, agents derive utility from "doing their part": in the spirit of [20], this is understood to mean following the rule that, when followed by all the agents in a given group, would maximize some measure of the group's utility. The outcome is a set of rules for each group, which are mutually optimal (from the point of view of the group) given that individuals follow the rules within their group. Abstention still occurs because for some agents (those with higher costs of voting) the rule prescribes not to vote, since their contribution to increase the group's utility from the election's outcome does not compensate the increase in the group's total cost of voting. [8] provide some empirical support to the group rule.

The Common Value Setting with Strategic Agents

Feddersen and Pesendorfer [13,14] consider the Condorcet Jury Theorem in a strategic setting and reach a different conclusion than Condorcet. They model a voting problem in an almost common value setting as a game, that is, as a situation where agents interact strategically. The following simple example provides the basic insights of their model: suppose that there are three voters, 1, 2 and 3, and two candidates, A and B. Suppose that voter 1 is an A-*partisan*, meaning he prefers candidate A in all states of the world. Agents 2 and 3 instead prefer candidate A in state s_A and candidate B in state s_B. Let p be the probability of state s_A and $1-p$ the probability of s_B. Now, suppose that agent 2 is informed, i. e. he knows the state of the world before voting, while agent 3 is not. Finally, suppose that there is no cost of voting and that the election is decided by simple majority rule. In this situation, agent 1 votes for A, agent 3 for B, and agent 2 for A if he observes s_A and for B if he observes s_B. To understand why the uninformed agent 3 votes for B, notice that by doing so the outcome is that A is always selected in state s_A, while B is always selected in state s_B. This is clearly the best outcome for agent 3, and in all states this is also the best outcome for the majority of the population. In fact, if the true state is s_A, A is selected, which is preferred to B by all voters; if the true state is s_B, B is selected, which is preferred to B by two voters out of three.

The uninformed agent in the example votes for B no matter what the prior probability p is, even if p is close to one, that is, even if he is almost sure that A is the right candidate. By voting for B individual 3 counterbalances the A-partisan's vote, thereby allowing the informed voter (2) to induce the "right" outcome with probability one. In Condorcet's argument, if voters vote according to their belief about the state of the world, information is aggregated in a way that induces the right social choice to be made. In this setting, where preferences are only partially aligned, information is aggregated so as to always generate the decision that is preferred by the majority of the population only if uninformed voters vote strategically. In this setting, if voters vote sincerely, as is assumed by Condorcet, the result that the aggregation of information taking place during an election delivers the "right" social choice does not hold. In Feddersen and Pesendorfer's framework it is strategic voting that induces the "correct" social choice. This is the major difference with Condorcet's result, which was driven by the assumption of sincere voting.

Recognizing the strategic incentives that voters may have in an election, Feddersen and Pesendorfer [15] apply a similar analysis to the unanimity rule in juries. They find that in the context of a common value setting, unanimity voting might result in convicting the innocent more often than other rules because strategic jurors consider the probability of being pivotal (like voter 3 in the example) and make their decisions conditional on being pivotal.

Now suppose there is another voter, 4, who is uninformed and shares the same preferences as 2 and 3. In this setup, even with a zero cost of voting agent 4 would abstain, so that the informed voter is pivotal with probability one. Voter 4's behavior is known as *strategic abstention*: the act of abstaining by uninformed voters not because voting is costly, but because by doing so they allow the informed voters to be pivotal. By abstaining, uninformed voters in effect delegate the decision to the informed voters. In the example, voter 4's strategic abstention allows for information equivalence to arise: making voter 4 informed would not change the outcome of the election, as long as 4 strategically abstains so that the informed voter determines the outcome. Notice that this result is in the same spirit as Condorcet's Jury Theorem's result: there, having each voter's belief accurate with a probability slightly higher than 50% or substantially higher than 50% does not make a difference. As long as voters vote according to their belief, the outcome is the same no matter what the underlying belief accuracy is (as long as it is greater than a half). Although for different reasons, in both Condorcet's setting and Feddersen and Pesendorfer's model (under some circumstances) the aggregation of information that takes place during an election ensures that the outcome of the election does not vary if the electorate is made more informed.

Future Directions

There are a number of open questions in the literature on voting. As mentioned in earlier sections, at the time of writing there is no universally accepted theory of turnout. In particular, no theory delivers all of the relevant comparative statics observed empirically, that is, the variability of turnout across elections of difference size, importance, and closeness. Broadly speaking, one may say that the behavior of voters is still only partially understood. Whether voters vote based on strategic considerations or vote without regard to how other citizens might be voting is an unsettled issue. Furthermore, individuals' choices in hypothetical and real situations might differ. The act of voting in large elections is almost a hypothetical choice, in that the likelihood that a vote determines the outcome is negligible. An open question then is whether voters choose candidates as they would in a real situation or as they would in a hypothetical situation. As far as empirical work is concerned, little seems to be known of the empirical impact of political ignorance on the outcome of elections. Also, an issue that is both empirical and theoretical and has not been satisfactorily addressed is what types of election rules are best suited for different decisions. Finally, although much has been written on the origins and evolution of democracies, there is no general consensus as to what are the possible reasons of democracies historical evolution. As is clear from this section, many interesting questions in politics have not yet been satisfactorily answered, leaving space for future research.

Bibliography

Primary Literature

1. Arrow KJ (1950) A difficulty in the concept of social welfare. J Political Econ 58(4):328–46
2. Bentham J (1789) An introduction to the principles of morals and legislation. Payne, London
3. Besley T, Coate S (1997) An economic model of representative democracy. Quart J Econ 112(1):85–114
4. Black D (1948) The decisions of a committee using a special majority. Econometrica 16(3):245–61
5. Blais A (2000) To vote or not to vote: The merits and limits of rational choice theory. University of Pittsburgh Press, Pittsburgh
6. Borda J (1781) Mathematical derivation of an election system. Isis 44(1–2):42–51
7. Campbell A, Converse PE, Miller WE, Stokes DE (1960) The American voter. Wiley, New York
8. Coate S, Conlin M (2004) A group-rule utilitarian approach to voter turnout: Theory and evidence. Am Econ Rev 95(5):1476–1504
9. Condorcet M (1785) Essai sur l'application de l'analyse a la probabilite' des decisions rendues a la pluralite des voix. L'Imprimerie Royale, Paris. English translation by Baker (1976) Chelsea, New York
10. Downs A (1957) An economic theory of democracy. Harper, New York
11. Duverger M (1972) Factors in a two-party and multiparty system. In: Party Politics and Pressure Groups. Crowell Company, New York, pp 23–32
12. Dye TR, Zeigler TH (1970) The irony of democracy, an uncommon introduction to american politics. Duxbury, Belmont
13. Feddersen T, Pesendorfer W (1996) The swing voter's curse. Am Econ Rev 86(3):408–424
14. Feddersen T, Pesendorfer W (1997) Voting behavior and information aggregation in elections with private information. Econometrica 86(3):408–424
15. Feddersen T, Pesendorfer W (1998) Convicting the innocent: The inferiority of unanimous jury verdicts. Am Political Sci Rev 92:23–35
16. Feddersen T, Sandroni A (2006) Theory A of participation in elections. Am Econ Rev 96(4):1271–1282

17. Gibbard A (1973) Manipulation of voting schemes. Econometrica 41:587–601
18. Harsanyi JC (1953) Cardinal utility in welfare economics and the theory of risk-taking. J Political Econ 61:434–435
19. Harsanyi JC (1977) Rational behavior and bargaining equilibrium in games and social situations. Cambridge University Press, New York
20. Harsanyi JC (1980) Rule utilitarianism, rights, obligations and the theory of rational behavior. Theory Decis 12(1):115–33
21. Hotelling H (1929) Stability and competition. Econ J 39(1): 41–57
22. Kinder DR, Kiewiet DR (1979) Economic discontent and political behavior: The role of personal grievances and collective economic judgements in congressional voting. Am J Political Sci 23(3):495–527
23. Ledyard J (1981) The paradox of voting and candidate competition: A general equilibrium analysis. In: Hoorwich G, Quick JP (eds) Essays in Contemporary Fields of Economics. Purdue University Press, Lafayette, pp 54–80
24. Ledyard J (1984) The pure theory of two candidate elections. Public Choice 44(1):7–41
25. Markus G (1988) The impact of personal and national economic conditions on the presidential vote: A pooled cross-sectional analysis. Am J Political Sci 32(1):137–54
26. Mas-Colell A, Whinston M, Green J (1995) Microeconomic theory. Oxford University Press, Oxford
27. Morton R (1987) Group majority a model of voting. Social Choice and Welfare 4(2):117–31
28. Morton R (1991) Groups in rational turnout models. Am J Political Sci 35:758–76
29. Osborne MJ, Slivinski A (1996) Model a of political competition with citizen-candidates. Quart J Econ 111(1):65–96
30. Palfrey T, Rosenthal H (1983) A strategic calculus of voting. Public Choice 41(1):7–53
31. Palfrey T, Rosenthal H (1985) Voter participation and strategic uncertainty. Am Political Sci Rev 79(1):62–78
32. Persson T, Tabellini G (2001) Political economics. Press MIT, Boston
33. Rawls J (1971) A theory of justice. Harvard University Press, Cambridge
34. Riker W, Ordeshook P (1968) A theory of the calculus of voting. Am Political Sci Rev 62:25–42
35. Robbins L (1938) Interpersonal comparisons of utility: A comment. Econ J 48(192):635–41
36. Satterthwaite M (1975) Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. J Econ Theory 10: 187–217
37. Schram A (1991) Voter behavior in economic perspective. Springer, Heidelberg
38. Shachar R, Nalebuff B (1999) Follow the leader: Theory and evidence on political participation. Am Econ Rev 89(3):525–47
39. Tullock G (1967) Towards a mathematics of politics. The University of Michigan Press, Ann Arbor
40. Uhlaner C (1989) Rational turnout: The neglected role of groups. Am J Political Sci 33(2):390–422
41. Wolfinger, Rosenstone (1980) Who votes? Yale University Press, New Haven

Books and Reviews

Arrow KJ (1951) Social choice and individual values. Wiley, New York
Arrow KJ, Sen AK, Suzumura K (eds) (2002) Handbook of social choice and welfare, vol 1. Elsevier, Amsterdam
Austen-Smith D, Banks JS (1999) Positive political theory I: collective preference. The University of Michigan Press, Ann Arbor
Cox GW (1997) Making votes count: strategic coordination in the world's electoral systems. Cambridge University Press, Cambridge
Dummett M (1984) Voting procedures. Clarendon Press, Oxford
Duverger M (1959) Political parties: their organization and activity in the modern state. Methuen and Co. Ltd, London
Dworkin R (1981) What is equality? Part 2: Equality of resources. Phil and Public Affairs 10:283–345
Feddersen T (2004) Rational choice theory and the paradox of not voting. J Econ Perspect 18(1):99–112
Katz RS (1997) Democracy and elections. Oxford University Press, Oxford
McLean I (1987) Public choice: An introduction. Basil Blackwell Inc., New York
Milnor AJ (1969) Elections and political stability. Little, Brown, Boston
Myerson R (2000) Large poisson games. J Econ Theory 94(1):7–45
Mueller DC: Public choice II (1989) A revised edition of Public Choice. Cambridge University Press, Cambridge
Niemi RG, Weisberg HF (eds) (1972) Probability models of collective decision making. Charles Merrill Publishing E company, Columbus
Ordeshook PC (1986) Game theory and political theory: An introduction. Cambridge University Press, Cambridge
Ordeshook PC (ed) (1989) Models of strategic choice in politics. The University of Michigan Press, Ann Arbor
Rawls J (1958) Justice as fairness. Phil Rev 67(2):164–194
Riker WH (ed) (1993) Agenda formation. The University of Michigan Press, Ann Arbor
Sen AK (1973) On economic inequality. Oxford University Press, Oxford
Sen AK (2002) Rationality and freedom. The Belknap Press of Harvard University Press, Cambridge
Tidean N (2006) Collective decisions and voting: The potential for public choice. Ashgate, Burlington
Tullock G (1998) On voting. Edward Elgar Publishing, Northampton

Voting Procedures, Complexity of

Olivier Hudry
École Nationale Supérieure des Télécommunications, Paris, France

Article Outline

Glossary
Definition of the Subject
Introduction
Common Voting Procedures
Complexity Results

Further Directions
Acknowledgments
Bibliography

Glossary

Condorcet winner A candidate is a Condorcet winner if he or she defeats any other candidate in a one-to-one matchup. Such a candidate may not exist; at most, there is only one. Though it could seem reasonable to adopt a Condorcet winner (if any) as the winner of an election, many common voting procedures bypass the Condorcet winner in favor of a winner chosen by other criteria.

Majority relation, strict majority relation In a pairwise comparison method, each candidate is compared to all others, one at a time. If a candidate x is preferred to a candidate y by at least $m/2$ voters (a majority), where m denotes the number of voters, x is said to be preferred to y according to the majority relation. The strict majority relation is defined in a similar way, but with $(m+1)/2$ instead of $m/2$. If there is no tie, the strict majority relation is a tournament, i. e., a complete asymmetric binary relation, called the majority tournament.

Preference, preference aggregation A voter's preference is some relational structure defined over the set of candidates. Such a structure depends on the chosen voting procedure, and usually ranges between a binary relation on one extreme and a linear order on the other. Given a collection, called a profile, of individual preferences defined on a set of candidates, the aggregation problem consists in computing a collective preference summarizing the profile as well as possible (for a given criterion).

Profile A profile $\Pi = (R_1, R_2, \ldots, R_m)$ is an ordered collection (or a multiset) of m relations R_i $(1 \leqslant i \leqslant m)$ for a given integer m. As the relations R_i can be the same, another representation of a profile Π consists in specifying only the q relations R_i which are different, for an appropriate integer q, and the number m_i of occurrences of each relation R_i $(1 \leqslant i \leqslant q)$: $\Pi = (R_1, m_1; R_2, m_2; \ldots; R_q, m_q)$.

Social choice function, social choice correspondence A social choice function maps a collection of individual preferences specified on a set of candidates onto a unique candidate, while a social choice correspondence maps it onto a nonempty set of candidates. This provides a way to formalize what constitutes the most preferred choice for a group of agents.

Voting procedure, voting theory A voting procedure is a rule defining how to elect a winner (single-winner election) or several winners (multiple-winner election) or to rank the candidates from the individual preferences of the voters. Voting theory studies the (axiomatic, algorithmic, combinatorial, and so on) properties of the voting procedures designed in order to reach collective decisions.

Definition of the Subject

One main concern of voting theory is to determine a procedure (also called, according to the context or the authors, *rule*, *method*, *social choice function*, *social choice correspondence*, *system*, *scheme*, *count*, *rank aggregation*, *principle*, *solution* and so on), for choosing a winner from among a set of candidates, based on the preferences of the voters. Each voter's preference may be expressed as the choice of a single individual candidate or, more ambitiously, a ranked list including all or some of the candidates. Such a situation occurs, obviously, in the field of social choice and welfare (for a broader presentation of the field of social choice and welfare, see for instance [4,7,8,9,14,17,67,75,97,101,136,143,154]) and especially of elections (for more about voting theory, see [27,63,107,109,110,125,141,156,167,168,169]), but also in many other fields: games, sports, artificial intelligence, spam detection, Web search engines, Internet applications, statistics, and so on.

For a long time, much attention has been paid to the axiomatic properties fulfilled by different procedures that have been proposed. These properties are important in choosing a procedure, since there is no "ideal" procedure (see next section). More recently, in the late 1980s and the early 1990s, the question has arisen regarding the relative difficulty of computing winners according to a given procedure. (The first to study the question may have been J. Orlin in 1981 [142]; for a recent introduction to computational social choice, see [38].) From a practical point of view, it is crucial to be able to announce the winner in a "reasonable" time. This raises the question of the complexity of voting procedures, which should be taken into account to the same extent as their axiomatic characteristics.

Below, we will detail the complexity results of several procedures: plurality rule (one-round procedure), plurality rule with run-off (two-round procedure), preferential voting procedure (STV), Borda's procedure, Nanson's procedure, Baldwin's procedure, Condorcet's procedure, Condorcet–Kemeny problem, Slater problem, prudent orders (G. Köhler, K.J. Arrow and H. Raynaud), maximin procedure (P.B. Simpson), minimax procedure

(K.J. Arrow and H. Raynaud), ranked pairs procedure (T.N. Tideman), Copeland's procedure, the top cycle solution (J.H. Smith), the uncovered set solution (P.C. Fishburn, N. Miller), the minimal covering set solution (B. Dutta), Banks's solution, the tournament equilibrium set solution (T. Schwartz), Dodgson's procedure, Young's procedure, approval voting procedure, majority-choice approval procedure (F. Simmons), and Bucklin's procedure. Sect. "Introduction" is devoted to a historic overview and to basic definitions and notation. The common voting procedures are depicted in Sect. "Common Voting Procedures". Sect. "Complexity Results" specifies the complexity of these procedures. Other considerations linked to complexity in the field of voting theory can be found in Sect. "Further Directions".

Introduction

The Search for a "Good" Voting Procedure, from Borda to Arrow

It is customarily agreed that the search for a "good" voting procedure goes back at least to the end of the eighteenth century, to the works of the chevalier Jean-Charles de Borda (1733–1799) [24], and of Marie Jean Antoine Nicolas Caritat, marquis de Condorcet (1743–1794) [31] (for references upon the historical context, see [7,17,23,80,81,119,120,121,122,123,124] and references below).

In the 1770s in the 1780s, J.-C. de Borda [24], a member of the French Academy of Sciences, showed that the plurality rule used at that time by the academy was not satisfactory. Indeed, with such a voting procedure, the winner could be contested by a majority of voters who would all agree to choose another candidate over the elected winner. (We may notice that the plurality rule with run-off used in many countries has the same defect.) Borda then suggested another procedure (see below). But, as pointed out by Condorcet in 1784 [31], the procedure advocated by Borda had the same defect as the one depicted by Borda himself for the plurality rule: a majority of dissatisfied voters could agree to constitute a majority coalition against the candidate elected by Borda's procedure in favor of another candidate. Condorcet then designed a method based on pairwise comparisons. By nature, this method cannot elect a winner who will give rise to a majority coalition against him or her. But, unfortunately, Condorcet's method does not always succeed in finding a winner. If such a winner does exist, i. e., if there exists a candidate who defeats any other candidate in such a pairwise comparison, then this candidate is said to be a *Condorcet winner*; if he or she exists, a Condorcet winner is unique. Actually, the Academy of Sciences decided to adopt Borda's method (until 1803). By the way, notice that, according to I. McLean, H. Lorrey and J. Colomer [124], "both Ramon Llull (ca 1232–1316) and Nicolaus of Cusa (also known as Cusanus, 1401–1464) made contributions which have been believed to be centuries more recent. Llull promotes the method of pairwise comparison, and proposes the Copeland rule to select a winner. Cusanus proposes the Borda rule, which should properly be renamed the Cusanus rule". Despite these historical discoveries, we shall keep the usual names.

After these seminal works, in spite of some works by the Swiss Simon Lhuilier (1750–1840) [112] (see also [128]), the Spanish Joseph Isidoro Morales [139], the French Pierre Claude François Daunou (1761–1840) [54] and Pierre Simon, marquis de Laplace (1749–1827) [105], who, through a slightly different approach, rediscovered Borda's procedure, it seems that the history of the theory of social choice slowed down and almost disappeared until the 1870s or even the 1950s (see [23]). In the 1870s, the English reverend Charles Lutwidge Dodgson (1832–1898), also (or maybe better) known as Lewis Carroll, proposed a voting system in which the winner is the candidate who becomes a Condorcet winner with the fewest appropriate changes in voters' preferences [58,59,60]. Some years later, another English mathematician, Edward J. Nanson (1850–1936) [140] on the one hand, and the Australian Joseph M. Baldwin (1878–1945) [11] in 1926 on the other hand, slightly modified Borda's method by iteratively eliminating some candidates until only one remains.

None of these methods is utterly satisfactory when we consider the properties which are usually considered desirable. In 1951 and 1963, Kenneth J. Arrow [7] shows that there does not exist a "good" voting method, with respect to some "reasonable" axiomatic properties; this is known as the famous "impossibility theorem". More precisely, assuming that the preferences of the voters are complete preorders (see below; notice that the set of complete preorders includes the one of linear orders, quite often considered to model the preferences of the voters) and that the result of the voting procedure should also be a complete preorder, K.J. Arrow considered the following properties:

- Unrestricted domain or universality: the voting procedure must be able to provide a result whatever the preferences of the voters are.
- Independence of irrelevant candidates: the collective preference between candidates x and y must depend only on the individual preferences between x and y; in other words, the collective preference between x and y must remain the same as long as the individual preferences between x and y do not change.

- Unanimity (or Pareto property): if a candidate x is preferred to another candidate y by all the voters, then x must be preferred to y in the collective preference, too.

K.J. Arrow showed that, if there are at least three candidates (things are much more comfortable with only two candidates!) and at least two voters, the only procedure which satisfies all these conditions at once is the dictatorship, in which one voter (the dictator) imposes his or her preference. Though this impossibility theorem ruins the hope of designing a voting procedure fulfilling the usual desirable properties, several procedures have been suggested since this date; we shall describe some of them below. Among ways to escape Arrow's impossibility theorem, we find the following:

- The definition of other axiomatic systems that would lead to voting procedures which would not be dictatorship.
- The restriction of individual preferences to more constrained domains.
- Adapting the result, when this one is not satisfactory with respect to the required axiomatic properties, into a result fulfilling these properties and fitting the genuine result as well as possible, for some criterion which must be defined.

The main questions associated with the first possibility are "given some axiomatic properties, what are the voting procedures satisfying these properties?" or, conversely, "given a voting procedure, what is the proper axiomatic system characterizing this procedure?"; we will not consider this direction here. The second possibility will be illustrated below, by the restriction of individual preferences to *single-peaked preferences*. The third direction was followed by J.G. Kemeny in 1959 [102], when he studied the aggregation of orders into a *median order* (see below). Notice that this approach is also attributed to Condorcet; in the sequel, we will refer the search for a median linear order as the Condorcet–Kemeny problem (as other people rediscovered this problem or some of its variants [130], the problem has also been known under other names, see [36]). Another related problem, dealing with the *majority tournament* (see below) is one, stated explicitly by P. Slater in 1961 [160], of fitting a tournament into a linear order at minimum distance. This is one of the so-called *tournament solutions* (see [106] and [138]; for references on tournaments, see also [118,134,149,150]), of which the aim is to determine a winner from a tournament. Besides the Slater solution, we shall describe other tournament solutions, such as the top cycle, the uncovered set, the minimal covering set, Banks's solution, the tournament equilibrium set. The other common tournament solutions are polynomial or their complexities are not completely known (see [93] for more details).

Definitions, Notation and Partially Ordered Sets Used to Model Preferences

Let us assume that we are dealing with m voters who must choose between n candidates denoted $x_1, x_2, \ldots, x_n$ or $x, y, z \ldots$; X denotes this set of candidates; in the following, we suppose that n is large enough. A binary relation R defined on X is a subset of $X \times X = \{(x, y) : x \in X \text{ and } y \in X\}$. We use the notation xRy instead of $(x, y) \in R$ and $x\bar{R}y$ instead of $(x, y) \notin R$.

It is customary to represent the preferences R_i $(1 \leqslant i \leqslant m)$ of m voters as an ordered collection (or a multiset) $\Pi = (R_1, R_2, \ldots, R_m)$, called the *profile* of the m binary relations. Another representation of the preferences of the voters exists. Since these preferences may be the same for two different voters, we may consider only the different relations $R_1, R_2, \ldots, R_q$ arising from the opinions of the voters, where q denotes this number of different opinions. In this way, we combine all the voters sharing the same opinion. Last, if m_i $(1 \leqslant i \leqslant q)$ denotes the number of voters sharing R_i $(1 \leqslant i \leqslant q)$ as their preference (notice the equality $\sum_{i=1}^{q} m_i = m$), we associate this number m_i of occurrences with each type R_i of relations to describe Π. Then Π can be described as the set of such pairs: $\Pi = \{(R_1, m_1); (R_2, m_2); \ldots; (R_q, m_q)\}$. Such a representation does represent the data more compactly when m is large with respect to n. Usually, the complexity results are the same for the two representations, because their proofs stand even if m is bounded by a polynomial in n, and in this case there is no qualitative difference between the two representations. With respect to the results stated in Sect. "Complexity Results", there would be a difference only for the L^{NP}-completeness results of Theorems 5, 14 and 15 (which then should be replaced only by NP-hardness results). So, in the sequel, we shall consider only the first representation. Moreover, we shall assume that the preference of each voter i $(1 \leqslant i \leqslant m)$ is given by a binary relation R_i defined on X, and that R_i is described by its characteristic vector (see below), which requires n^2 bits. So the size of the data set is about mn^2.

We will consider two kinds of elections: we want to elect one candidate, or we want to rank all of them into a partially ordered set (or *poset*). One of the most common posets is the structure of linear order. Other posets can be defined from the following basic properties:

Reflexivity $\forall x \in X, xRx$;

Irreflexivity $\forall x \in X, x\bar{R}x$;

Antisymmetry $\forall(x, y) \in X^2$, (xRy and $x \neq y$) $\Rightarrow y\bar{R}x$;
Asymmetry $\forall(x, y) \in X^2, xRy \Rightarrow y\bar{R}x$;
Transitivity $\forall(x, y, z) \in X^3$, (xRy and yRz) $\Rightarrow xRz$;
Completeness $\forall(x, y) \in X^2$, with $x \neq y$, xRy or (inclusively) yRx.

By combining the above properties, we may define different types of binary relations (see for instance [17,32,74,77]). As a binary relation R defined on X is the same as the directed graph $G = (X, R) : (x, y)$ is an arc (i. e., a directed edge; similarly, a directed cycle will be called a circuit in the sequel) of G if and only if we have xRy, we illustrate these types with graph theoretic examples (for basic references on graph theory, see [12] or [22]). It is often possible to define reflexive or irreflexive versions of the following ordered structures. But, as reflexivity and irreflexivity do not matter for complexity results (see [91]), we give below only one version among these two possibilities.

- A *partial order* is an antisymmetric (if reflexive) or asymmetric (if irreflexive) and transitive binary relation (see Fig. 1); $\mathcal{O}$ will denote the set of the partial orders defined on X.
- A *linear order* is a complete partial order (see Fig. 2); $\mathcal{L}$ will denote the set of the linear orders defined on X. If L denotes a linear order defined on X, we will represent L as $x_{\sigma(1)} > x_{\sigma(2)} > \cdots > x_{\sigma(n)}$ for some appropriate permutation σ, with the agreement that the notation $x_{\sigma(i)} > x_{\sigma(i+1)}$ (for $1 \leqslant i < n$) means that $x_{\sigma(i)}$ is preferred to $x_{\sigma(i+1)}$ according to L, the relationship between the other elements of X being involved by transitivity. The element $x_{\sigma(1)}$ will be called *the winner of L*.
- A *tournament* is a complete and asymmetric binary relation (see Fig. 3); $\mathcal{T}$ will denote the set of the tournaments defined on X; notice that a transitive tournament is a linear order and conversely. As a tournament may contain circuits, they are usually not considered as an appropriate structure to represent the collective preference that we seek. Tournaments will be used below to summarize individual preferences in the so-called *majority tournament*.
- A *preorder* is a reflexive and transitive binary relation (see Fig. 4); $\mathcal{P}$ will denote the set of the preorders defined on X.
- A *complete preorder* is a reflexive, transitive and complete binary relation (see Fig. 5); C will denote the set of the complete preorders defined on X.
- An *acyclic relation* (we should rather say *without circuit*, but *acyclic* is the usual term) is a relation R of which the associated graph G is without any circuit (see Fig. 6); $\mathcal{A}$ will denote the set of the acyclic relations defined on X.

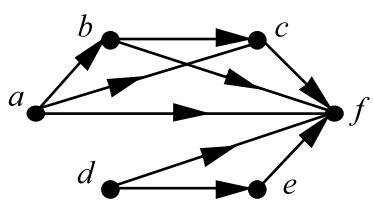

Voting Procedures, Complexity of, Figure 1
A partial order

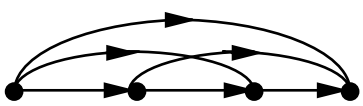

Voting Procedures, Complexity of, Figure 2
A linear order. The partial order of Fig. 1 is not a linear order, for instance because the vertices *a* and *d* are not compared

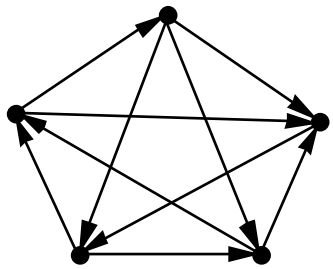

Voting Procedures, Complexity of, Figure 3
A tournament

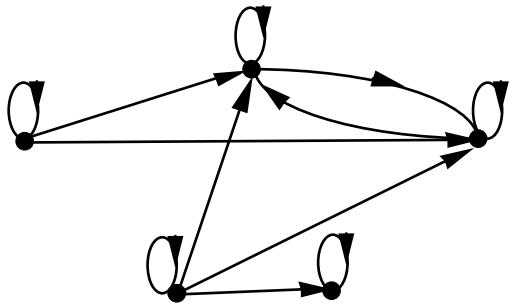

Voting Procedures, Complexity of, Figure 4
A preorder

As stated above, it is possible to obtain other structures by adding or removing reflexivity or irreflexivity from the above definitions. In fact, the distinction between reflexive and irreflexive relations is not relevant from the complexity point of view: the results remain the same. Thus, in the following, we do not take reflexivity or irreflexivity into account.

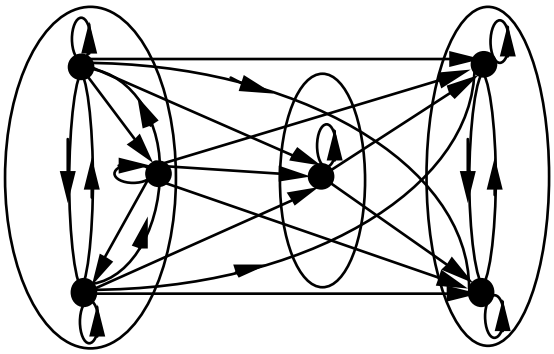

Voting Procedures, Complexity of, Figure 5
A complete preorder

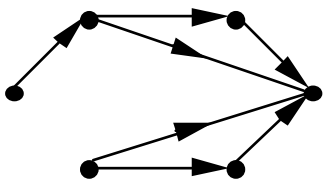

Voting Procedures, Complexity of, Figure 6
An acyclic relation. Its transitive closure is the partial order of Fig. 1

Other types of posets exist, obtained by considering the asymmetric parts of the previous relations (for instance, a *weak order* is sometimes defined as the asymmetric part of a complete preorder) or by adding extra structural properties (for instance to define interval orders or semiorders from partial orders). We will not consider these posets here; the interested reader is referred to [32] for their definitions and to [91] for some complexity results about them. On the contrary, we will pay attention to generic binary relations without any particular property; the set of the binary relations will be noted $\mathcal{R}$. We may notice several inclusions between these sets, especially the following ones: $\forall \mathcal{Z} \in \{\mathcal{A}, \mathcal{C}, \mathcal{L}, \mathcal{O}, \mathcal{P}, \mathcal{R}, \mathcal{T}\}$, $\mathcal{L} \subseteq \mathcal{Z} \subseteq \mathcal{R}$; in other words, a linear order can be considered as a special case of any one of the other types, and any type is a special case of binary relation.

To conclude this section, we summarize the notation for these sets in Table 1.

Common Voting Procedures

In this section, we describe the main voting procedures, i. e., plurality rule (one-round procedure), plurality rule with run-off (two-round procedure), preferential voting procedure (STV), Borda's procedure, Nanson's procedure, Baldwin's procedure, Condorcet's procedure, Condorcet–Kemeny problem, Slater problem, prudent orders (G. Köhler, K.J. Arrow and H. Raynaud), maximin procedure (P.B. Simpson), minimax procedure (K.J. Arrow and H. Raynaud), ranked pairs procedure (T.N. Tideman), Copeland's procedure, the top cycle solution (J.H. Smith), the uncovered set solution (P.C. Fishburn, N. Miller), the minimal covering set solution (B. Dutta), Banks's solution, the tournament equilibrium set solution (T. Schwartz), Dodgson's procedure, Young's procedure, approval voting procedure, majority-choice approval procedure (F. Simmons), and Bucklin's procedure (see also [27]. For other tournament solutions, see [106] and [138]).

Plurality Rule, Plurality Rule with Run-Off, Preferential Voting Procedure

One of the easiest voting procedures by which to elect one candidate as a winner is the *plurality rule* (also called *one-round procedure* or *relative majority*, or sometimes *first-past-the-post*, or *winner-takes-all*, or also *majoritarian voting*…; see [95]). In this procedure, each voter gives one point to his or her favorite candidate (so it is not necessary to know the preferences of the voters on the whole set of candidates). The candidate who gains the maximum number of points is the winner.

This procedure belongs to the family of *scoring procedures*. In such a procedure, a *score-vector* $(s_1, s_2, \ldots, s_n)$ is fixed independently of the voters, with $s_1 \geqslant s_2 \geqslant \cdots \geqslant s_n$. For each voter, a candidate x receives s_i points if x is ranked at the ith position by the considered voter. The *score* of x is the total number of points that x received. The winner is the candidate with the maximum score. If the aim is to rank the candidates, we may also sort them according to decreasing scores and then consider the linear extensions of the complete preorder provided by this sorting. (Another possibility would be to apply the procedure $n-1$ times, after having removed the winner of the current it-

Voting Procedures, Complexity of, Table 1
Meaning of the notation $\mathcal{A}, \mathcal{C}, \mathcal{L}, \mathcal{O}, \mathcal{P}, \mathcal{R}, \mathcal{T}$

$\mathcal{A}$: acyclic relation	$\mathcal{L}$: linear order	$\mathcal{P}$: preorder	$\mathcal{T}$: tournament
$\mathcal{C}$: complete preorder	$\mathcal{O}$: partial order	$\mathcal{R}$: binary relation	

eration.) For the plurality rule, the score vector is (1, 0, 0, ..., 0).

There are two rounds in the *plurality rule with run-off*, also called *two-round* (or *two-ballot*) *procedure*. The first round is like the plurality rule described above. At the end of this first step, if a candidate has gained at least $(m+1)/2$ points (the strict majority), he or she is the winner. Otherwise, the two candidates with the maximum numbers of points remain for a second round and the others are removed from the election. Then the plurality rule is applied again, but only to the remaining two candidates. This method is designed to elect only one winner. If we want to obtain k winners, it is sufficient to apply it k times. Notice that the repetition of a given procedure always makes it possible to elect several winners.

A generalization of these procedures consists in performing a given number of rounds. For each round, each voter gives one point for his or her favorite candidate. If, at the end of a round, there is a candidate who has gained at least $(m+1)/2$ points, he or she is the winner. Otherwise, the candidates with the lowest numbers of points are eliminated from the competition; the number of candidates eliminated at each round depends on the number of rounds, but is such that only two candidates will remain for the last round. The winner of this last round is the winner of the election. Special cases are the plurality rule with run-off, described above, and the one with at most $n-1$ rounds. For this latter case, exactly one candidate is removed at each round. This variant, in which the candidate who is the least often ranked at the first position is removed, is also known as *preferential voting* (or *preference voting*) or as *single transferable vote* (STV), or as *instant-runoff voting* (IRV).

Other variants of these voting procedures may be defined by the successive eliminations of the losers. For instance, the candidate who is most often ranked last is removed, and we iterate this process while there remain at least two candidates.

Borda's Procedure and Some Variants (Nanson's and Baldwin's Procedures)

As related above, Borda considered the plurality rule unsatisfactory because the winner can be contested by a majority of voters who would all agree to choose another candidate instead of the elected winner (the same holds for the plurality rule with run-off, see Example 1 below). Borda suggested another procedure in which the voters rank the n candidates according to their preferences. For each voter, the candidate who is ranked first is given $n-1$ points, a candidate ranked second is given $n-2$ points, and so on: more generally, a candidate ranked at the ith position is given $n-i$ points. Then, all these points are summed up for each candidate: this sum is the *Borda score* s_B of the candidate. The candidate with a maximum Borda score is the *Borda winner*. So, Borda's procedure is also a scoring procedure, of which the score-vector is $(n-1, n-2, \ldots, 1, 0)$.

Using Borda's procedure, one can easily obtain a ranking of all the candidates. A first possibility consists, as for the plurality rule or the plurality rule with run-off, in iterating Borda's procedure $n-1$ times, after having removed the Borda winner of the current iteration. But we may also apply Borda's procedure only once, and then rank the candidates according to the decreasing values of their Borda scores. This gives a complete preorder. Any linear extension of this complete preorder can be considered as the collective ranking according to Borda's procedure. Notice that these two possibilities do not necessarily provide the same rankings (see Example 1).

Several variants of Borda's procedure have been studied. For instance, we may apply other score-vectors: instead of $n-i$ points given to the candidate ranked at the ith position, we may choose to assign other values. For instance, given an integer k, we may credit k points to the candidate ranked at the first position, $k-1$ points to the second, and so on, through the candidate ranked at the kth position, who gains 1 point, after which the following candidates gain nothing (the shape of the score-vector is $(k, k-1, \ldots, 1, 0, \ldots, 0)$). For $k=1$, this system is the plurality rule. For $k \geqslant n-1$, this system gives the same results as Borda's procedure.

Other systems are based on Borda's procedure, but with the elimination of some candidates. *Nanson's procedure* [140] modifies Borda's procedure by eliminating the candidates whose Borda scores are below the average Borda score and by repeating the computations of the Borda scores with respect to the remaining candidates after these eliminations, until there remains only one candidate. Another variant of Borda's procedure is one suggested by Baldwin in 1926 [11]; as in Nanson's procedure, candidates are iteratively removed from the election. But, in *Baldwin's procedure*, only one candidate is removed at each iteration, the one whose Borda's score is the lowest.

Condorcet's Procedure

Condorcet designed a method based on pairwise comparisons. More precisely, for each candidate x and each candidate y with $x \neq y$, we compute the number m_{xy}, that we will call the *pairwise comparison coefficient* below, of voters who prefer x to y. Then x is considered as better

than y if a majority of voters prefers x to y, i. e., if we have $m_{xy} > m_{yx}$. This defines the (*strict*) *majority relation* T: $xTy \Leftrightarrow m_{xy} > m_{yx}$. In some cases, there exists a *Condorcet winner*, i. e., a candidate x defeating any other candidate: $\forall y \neq x, m_{xy} > m_{yx}$. If there exists a Condorcet winner, then he or she is unique. It may even happen that T is a linear order and allows us to rank all the candidates. Such is the case for the following example (due to B. Monjardet [133]), which illustrates the previous voting procedures.

Example 1 Assume that $m = 13$ voters must rank $n = 4$ candidates x, y, z, and t. Suppose the preferences of the voters are given by the following linear orders:

- the preferences of two voters are: $x > y > z > t$;
- the preference of one voter is: $y > z > x > t$;
- the preference of one voter is: $y > z > t > x$;
- the preferences of four voters are: $z > y > x > t$;
- the preferences of five voters are: $t > x > y > z$.

According to the plurality rule, t is the winner with 5 points (2 points for x and for y, 4 points for z). According to the plurality rule with run-off, z is the winner (the four voters who voted for x or y prefer z to t). The Borda scores s_B of the candidates are:

- $s_B(x) = 2 \times 3 + 5 \times 2 + 5 \times 1 + 1 \times 0 = 21$;
- $s_B(y) = 2 \times 3 + 6 \times 2 + 5 \times 1 + 0 \times 0 = 23$;
- $s_B(z) = 4 \times 3 + 2 \times 2 + 2 \times 1 + 5 \times 0 = 18$;
- $s_B(t) = 5 \times 3 + 0 \times 2 + 1 \times 1 + 7 \times 0 = 16$.

So, the winner according to Borda's procedure is y, and the ranking of the four candidates is $y > x > z > t$ (notice that if, while there are at least two candidates, we apply the variant consisting in removing the winner and reapplying Borda's procedure, then we obtain the orders $y > x > z > t$ and $y > z > x > t$ as the possible rankings; the distance to the ranking provided by one application of Borda's procedure may be much more important). In Nanson's procedure, as the Borda scores of z and t are below the average (which is equal to 19.5), z and t are removed and only x and y remain. Then the Borda scores of x and y for the second step become $s_B(x) = 7$ and $s_B(y) = 6$: hence x is the winner according to Nanson's procedure. In Baldwin's procedure, t is first removed. The Borda scores of x, y and z become $s_B(x) = 14$, $s_B(y) = 15$, $s_B(z) = 10$. Thus z is removed and the remaining computations are as in Nanson's procedure: here also x is the winner (but it may happen that Nanson's procedure and Baldwin's procedure do not provide the same winners).

Let us now compute the pairwise comparison coefficients m_{xy} necessary to apply Condorcet's procedure:

- $m_{xy} = 2 + 5 = \mathbf{7}$; $m_{yx} = 1 + 1 + 4 = 6$;
- $m_{xz} = 2 + 5 = \mathbf{7}$; $m_{zx} = 1 + 1 + 4 = 6$;
- $m_{xt} = 2 + 1 + 4 = \mathbf{7}$; $m_{tx} = 1 + 5 = 6$;
- $m_{yz} = 2 + 1 + 1 + 5 = \mathbf{9}$; $m_{zy} = 4$;
- $m_{yt} = 2 + 1 + 1 + 4 = \mathbf{8}$; $m_{ty} = 5$;
- $m_{zt} = 2 + 1 + 1 + 4 = \mathbf{8}$; $m_{tz} = 5$.

The bold values are the ones greater than the majority $(m + 1)/2$. These values show that the majority relation, here, is a linear order, namely $x > y > z > t$.

By the way, this example shows the importance of the voting procedure, as already noticed by Borda (see [113]): four procedures, four different winners; so all four of our candidates may claim to be the winners of the election... It is often the case that different procedures provide different winners.

Median Orders, Condorcet–Kemeny Problem, Slater Problem

As Condorcet himself discovered, his procedure may lead to a majority relation which is not transitive. The simplest example in this respect is one with $n = 3$ candidates x, y and z, and with $m = 3$ voters whose preferences are respectively $x > y > z$, $y > z > x$, $z > x > y$. It is easy to verify that the majority relation T is defined by xTy, yTz, zTx, hence a lack of transitivity. If there is no tie (which is necessarily the case if m is odd, since the individual preferences of the voters are assumed to be linear orders), then T is a tournament, called the *majority tournament*. We may or may not weight T if we want to take the intensity of the preferences into account. If T is not weighted, the search for a winner or for a ranking of the candidates leads to the definition of *tournament solutions* (see [106] for a comprehensive study of these, and [93] for a survey of their complexities, some of which are given below). In the Condorcet–Kemeny problem, T is weighted and the aim is to compute a linear order or, more generally, a poset fitting T "as well as possible".

To specify what "as well as possible" means, we use the *symmetric difference distance* δ defined, for two binary relations R and S defined on X, by:

$$\delta(R, S) = \left|\left\{(x, y) \in X^2 \colon \left(xRy \text{ and } x\bar{S}y\right) \text{ or } \left(x\bar{R}y \text{ and } xSy\right)\right\}\right| .$$

This quantity $\delta(R, S)$ measures the number of disagreements between R and S. Though it is possible to consider other distances, δ is widely used and is appropriate for many applications. J.-P. Barthélemy [15] shows that δ satisfies a number of naturally desirable properties. J.-P.

Barthélemy and B. Monjardet [17] recall that $\delta(R, S)$ is the Hamming distance between the characteristic vectors (see below) of R and S and point out the links between δ and the L_1-metric or the square of the Euclidean distance between these vectors (see also [129] and [130]).

So, for a profile $\Pi = (R_1, R_2, \ldots, R_m)$ of m relations, we can define the *remoteness* $\Delta(\Pi, R)$ between a relation R and the profile Π by:

$$\Delta(\Pi, R) = \sum_{i=1}^{m} \delta(R, R_i) .$$

The remoteness $\Delta(\Pi, R)$ measures the total number of disagreements between Π and R.

Our aggregation problem can now be seen as a combinatorial optimization problem: given a profile Π, determine a binary relation R^* minimizing Δ over one of the sets $\mathcal{A}, \mathcal{C}, \mathcal{L}, \mathcal{O}, \mathcal{P}, \mathcal{R}, \mathcal{T}$. Such a relation R^* will be called a *median relation* of Π [17]. According to the number m of relations of the profile, and according to the properties assumed for the relations belonging to Π or required from the median relation, we get many combinatorial problems. We note them as follows:

Problems $P_m(\mathcal{Y}, \mathcal{Z})$. For $\mathcal{Y}$ belonging to $\{\mathcal{A}, \mathcal{C}, \mathcal{L}, \mathcal{O}, \mathcal{P}, \mathcal{R}, \mathcal{T}\}$ and $\mathcal{Z}$ also belonging to $\{\mathcal{A}, \mathcal{C}, \mathcal{L}, \mathcal{O}, \mathcal{P}, \mathcal{R}, \mathcal{T}\}$, for a positive integer m, $P_m(\mathcal{Y}, \mathcal{Z})$ denotes the following problem: given a finite set X of n elements, given a profile Π of m binary relations all belonging to $\mathcal{Y}$, find a relation R^* belonging to $\mathcal{Z}$ minimizing Δ over $\mathcal{Z}$: $\Delta(\Pi, R^*) = \mathrm{Min}\{\Delta(\Pi, R) \text{ for } R \in \mathcal{Z}\}$.

With this notation, the initial problem, possibly considered by Condorcet and then surely by J.G. Kemeny [102], is $P_m(\mathcal{L}, \mathcal{L})$, consisting in aggregating m linear orders into a median linear order. A *Condorcet–Kemeny winner*, or here simply a *Kemeny winner*, is the winner of any median linear order (for references about the Condorcet–Kemeny problem, see for example [36,98,131,132,151] and references therein).

We may easily state the problems $P_m(\mathcal{Y}, \mathcal{Z})$ as 0-1 linear programming problems (see [17,36,89,91,173] for instance) for any profile $\Pi = (R_1, R_2, \ldots, R_m)$ of m binary relations R_i $(1 \leqslant i \leqslant m)$ all belonging to $\mathcal{Y}$. To this end, consider the characteristic vectors $r^i = (r^i_{xy})_{(x,y)\in X^2}$ of the relations R_i $(1 \leqslant i \leqslant m)$ defined by $r^i_{xy} = 1$ if xR_iy and $r^i_{xy} = 0$ otherwise, and similarly the characteristic vector $r = (r_{xy})_{(x,y)\in X^2}$ of any binary relation R. Then, after some computations, we obtain $\Delta(\Pi, R) = C - \sum_{(x,y)\in X^2} \alpha_{xy} r_{xy}$, where $C = \sum_{i=1}^{m} \sum_{(x,y)\in X^2} r^i_{xy}$ is a constant and with $\alpha_{xy} = 2\sum_{j=1}^{m} r^i_{xy} - m$. Notice that, for the problem $P_m(\mathcal{L}, \mathcal{L})$, the sum $\sum_{i=1}^{m} r^i_{xy}$ is equal for $x \neq y$ to the pairwise comparison coefficient m_{xy} introduced above. So, because of the equality $m = \sum_{i=1}^{m} r^i_{xy} + \sum_{i=1}^{m} r^i_{yx}$, we get $\alpha_{xy} = m_{xy} - m_{yx}$: α_{xy} measures the number of voters who prefer x to y minus the number of voters who prefer y to x. More generally, α_{xy} is equal to twice the gap between the number of voters who prefer x to y and the majority. It is a non-positive or non-negative integer with the same parity as m. To obtain a 0-1 linear programming statement of $P_m(\mathcal{Y}, \mathcal{Z})$, it is then sufficient to express the constraints defining the set $\mathcal{Z}$, which is easy for the posets described above. For instance, the transitivity of R can be expressed by the following inequalities:

$$\forall (x, y, z) \in X^3, 0 \leqslant r_{xy} + r_{yz} - r_{xz} \leqslant 1 .$$

As stated above, it is also common to represent a preference R defined on X by a graph. The properties of the graph are the properties of R: it can be antisymmetric, complete, transitive, and so on. Similarly, the profile $\Pi = (R_1, R_2, \ldots, R_m)$ can also be represented by a directed, weighted, complete, symmetric graph $G_\Pi = (X, U_X)$: its set of vertices is X, and G_Π contains all the possible arcs except the loops, i. e., $U_X = X \times X - \{(x, x) \text{ for } x \in X\}$. (The loops would be associated with reflexivity; this property, as well as irreflexivity, has no impact on the complexity status of the studied problems, hence this simplification.) The weights of the arcs (x, y) give the intensity of the preference for x over y. The computations above lead us to assign α_{xy} as the weight of any arc (x, y) of G_Π. With this choice, minimizing $\Delta(\Pi, R)$ is the same, from the graph theoretic point of view, as drawing from G_Π a subset of arcs with a maximum total weight and satisfying the structural properties required from R. Notice that characterizations of the graphs that we can associate with profiles Π have been provided by different authors (see [36,55,56,68,91,117,166]); the construction of the profiles can be done in polynomial time, which allows us to study the complexities of the problems $P_m(\mathcal{Y}, \mathcal{Z})$ through their graph theoretic representations G_Π.

Example 2 illustrates these considerations.

Example 2 Assume that $m = 9$ voters must rank $n = 4$ candidates x, y, z, and t. Assume, also, that the preferences of the voters are given by the following linear orders:

- the preferences of three voters are: $x > y > z > t$;
- the preferences of two voters are: $y > t > z > x$;
- the preference of one voter is: $t > z > x > y$;
- the preference of one voter is: $x > z > y > t$;
- the preference of one voter is: $t > y > x > z$;
- the preference of one voter is: $z > t > y > x$.

The quantities m_{xy} involved in Condorcet's procedure are the following, where the bold values, again, denote those greater than the majority $(m+1)/2$:

- $m_{xy} = \mathbf{5}$; $m_{yx} = 4$;
- $m_{xz} = \mathbf{5}$; $m_{zx} = 4$;
- $m_{xt} = 4$; $m_{tx} = \mathbf{5}$;
- $m_{yz} = \mathbf{6}$; $m_{zy} = 3$;
- $m_{yt} = \mathbf{6}$; $m_{ty} = 3$;
- $m_{zt} = \mathbf{5}$; $m_{tz} = 4$.

Here, the majority relation is not a linear order, but the tournament of Fig. 7. Figure 7 also displays the graph G_Π summarizing the data. We may observe that the majority tournament is given by the arcs of G_Π with a positive weight.

For these data, it is not too difficult to verify that there is only one median linear order, which is $x > y > z > t$, this order keeps all the positive weights except that of the arc (t, x), of which the weight is the minimum, while the arcs with positive weights do not define a linear order. Hence, x is the only Kemeny winner.

Attention has also been paid to $P_1(\mathcal{T}, \mathcal{L})$, i. e., the approximation of a tournament (which can be the majority tournament of the election) by a linear order at minimum distance. This problem is also known as *Slater problem*, since P. Slater explicitly stated it in 1961 ([160]; see also [36] for a survey of this problem). A linear order L^* at minimum distance (with respect to the symmetric difference distance) from the tournament T constituting the considered instance of $P_1(\mathcal{T}, \mathcal{L})$ is called a *Slater order* of T: $\delta(T, L^*) = \min_L \delta(T, L)$. This minimum distance is usually called the *Slater index* $i(T)$ of T: $\delta(T, L^*) = i(T)$. A *Slater winner of* T is the winner of any Slater order of T.

Prudent Orders, Maximin Procedure, Minimax Procedure, Ranked Pairs Procedure

Other procedures are based on the quantities m_{xy} defined above.

Such is the case for *prudent orders* proposed by G. Köhler [103] and by K.J. Arrow and H. Raynaud [8]. Given the pairwise comparison coefficients m_{xy}, let us define the *cut relation* $R_{>t}$ for any integer t with $-m \leqslant t \leqslant m$ by the following:

$$\text{for } x \in X \text{ and } y \in X \text{ with } x \neq y, \qquad x R_{>t} y \Leftrightarrow m_{xy} - m_{yx} > t\,.$$

Then define a threshold $t_{\min}$ by:

$$t_{\min} = \min\{t \text{ with } -m \leqslant t \leqslant m \text{ and such that } R_{>t} \text{ contains no circuit}\}\,.$$

A *prudent order* is any linear order which contains $R_{>t_{\min}}$. A winner according to the procedure of the prudent orders is the winner of any prudent order of the profile being considered. For Example 2, we have $t_{\min} = 1$; with respect to the graph G_Π of Fig. 7, $R_{>t_{\min}}$ contains only the arcs (y, z) and (y, t). There are eight prudent orders, including $x > y > z > t$ or $y > t > z > x$, but none of them has z or t as its winner: x and y are the winners according to the prudent order procedure.

The *maximin procedure*, proposed by P.B. Simpson in 1969 [159], is also based on the quantities m_{xy}. In this procedure, for each candidate x, we consider the worst performance $W(x)$ in the pairwise comparisons: $W(x) = \min_{y \neq x} m_{xy}$. The winner is any candidate x^* with a maximum worst performance: $W(x^*) = \max_{x \in X} W(x)$. We may also use these quantities to sort the candidates in decreasing order with respect to the quantities W to obtain

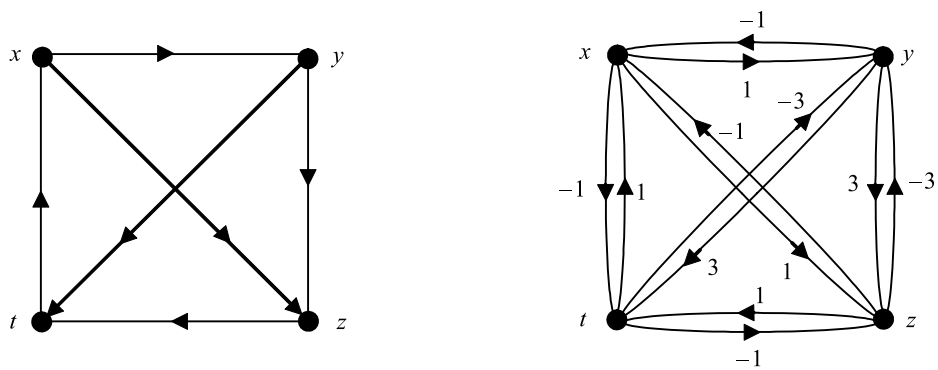

Voting Procedures, Complexity of, Figure 7
The majority tournament *(left)* and the graph G_Π *(right)* associated with the data of Example 2 and weighted by the quantities $\alpha_{xy} = m_{xy} - m_{yx}$

a collective preference over the whole set of candidates (as usual, another possibility is to remove the winner and to iterate the process). For instance, applied to Example 2, we obtain $W(x) = 4$, $W(y) = 4$, $W(z) = 3$, $W(t) = 3$: x and y are the Simpson winners.

The *minimax procedure*, in which the previous roles of "min" and "max" are switched and in which the candidates are sorted in increasing order with respect to the obtained maxima, has been proposed by K.J. Arrow and H. Raynaud [8]. For Example 2, the winners would be x, z and t.

In 1987, T.N. Tideman introduced the *ranked pairs procedure* [170]. Here, we first sort the quantities m_{xy} decreasingly. Then we scan them in this order to build a linear order L for the collective preference. If m_{xy} is the current quantity that we scan, then we definitely set xLy if such a decision is compatible with the previous ones. We do so until a linear order is completely defined. For instance, for Example 2, we first fix yLz and yLt since m_{yz} and m_{yt} are maxima and z and t cannot be winners. Then we must choose between adding xLy or xLz or tLx or zLt, since m_{xy}, m_{xz}, m_{tx} and m_{zt} are equal. We cannot add all of them simultaneously, because of incompatibilities with yLz or yLt, already fixed. If we choose to add xLy, xLz and zLt, we obtain the linear order $x > y > z > t$, and x is a winner. If we choose to add tLx and zLt, we obtain the linear order $y > z > t > x$, and y is a winner. Here again, x and y are the winners.

Tournament Solutions

The procedures described in this section apply to tournaments, which can function as the majority tournament of an election. They deal with unweighted tournaments, but some of them can be extended to weighted tournaments, as in the case of the solution designed by P. Slater (see above).

Number of Wins: Copeland's Procedure The procedure designed by A.H. Copeland in 1951 [48] is also based on the m_{xy}'s. For any two different candidates x and y, we set: $C(x, y) = 1$ if $m_{xy} > m_{yx}$, $C(x, y) = 0.5$ if $m_{xy} = m_{yx}$, $C(x, y) = 0$ if $m_{xy} < m_{yx}$. The Copeland score $C(x)$ of x is $C(x) = \sum_{y \neq x} C(x, y)$. The Copeland winner is any candidate with a maximum Copeland score (here also, we may rank the candidates according to decreasing Copeland scores).

The application of Copeland's procedure to Example 2 gives: $C(x) = 2$, $C(y) = 2$, $C(z) = 1$, $C(t) = 1$. Here, x and y are the Copeland winners.

The Copeland score has a meaning from a graph theoretic point of view. If we assume that there is no tie, the majority relation is a tournament (the majority tournament). Then the Copeland score of a candidate x is also the out-degree of the vertex associated with x. Copeland's procedure is one of the common tournament solutions (see [106]). It can be extended to weighted tournaments by sorting the vertices x according to the sum $\sum_{y \neq x} m_{xy}$ of the pairwise comparison coefficients m_{xy}; this variant leads to the same ranking as Borda's procedure.

Top Cycle: Smith's Solution Another tournament solution is the so-called *top cycle* (also called the *Smith set*) [161]. Any directed graph G can be decomposed into its strongly connected components. We may then define another graph H derived from G. The vertices of H are associated with the strongly connected components of G. Let x (respectively y) be a vertex of G associated with the strongly connected components C_x (respectively C_y) of G. There will exist an arc (x, y) from x to y if there is at least one arc in G from C_x to C_y (notice that, in this case, all the arcs between C_x and C_y are from C_x towards C_y). Then H does not contain any circuits. Moreover, if G is a tournament, H is a linear order and admits a winner (but H contains only one vertex if G is strongly connected). The top cycle of G is the strongly connected component of G associated with the winner of H.

The top cycle solution, when applied to a tournament T, consists of considering all vertices of the top cycle of T as the winners of T. For instance, for the tournament of Fig. 7, which is strongly connected, the top cycle contains the four vertices x, y, z, and t, which are the four winners according to this solution.

Uncovered Set: Fishburn's and Miller's Solution A refinement of the top cycle is provided by the set of uncovered vertices. Given a tournament T and two vertices x and y of T, we say that x covers y if (x, y) is an arc of T and if, for any arc (y, z), (x, z) is also an arc of T. (In other words, considering T as a majority tournament, x beats y and any vertex beaten by y is also beaten by x.) A vertex is said to be *uncovered* if no other vertex covers it. The uncovered set of T is noted $\mathrm{UC}(T)$. Adopting the elements of $\mathrm{UC}(T)$ as the winners of T has been independently suggested by P. Fishburn in 1977 [76] and by N. Miller in 1980 [126].

For the tournament of Fig. 7, it is easy to see that z is covered by y while x, y and t are uncovered: here, x, y, and t are the three winners according to this solution.

The definition of a covered vertex can easily be extended to weighted tournaments, though the possibility of weights equal to 0 makes the situation more difficult than expected (see [35] for details).

Minimal Covering Set: Dutta's Solution The uncovered set can also be refined, for example by the minimal covering set proposed by B. Dutta [64]. Consider a tournament T defined on X. We say that a subset Y of X is a *covering set* of T if we have the following property: $\forall x \in X - Y, x \notin \mathrm{UC}(Y \cup \{x\})$. For instance, $\mathrm{UC}(T)$ is a covering set of T. Let $\Xi(T)$ denote the set of the covering sets of T. Then there exists a minimal element of $\Xi(T)$ with respect to inclusion; this minimal element is called the *minimal covering set* $\mathrm{MC}(T)$ of T.

For the tournament of Fig. 7, the minimal covering set contains x, y, and t, which are the three winners according to this solution, here as for the uncovered set. (But there exist tournaments for which the general inclusion $\mathrm{MC}(T) \subseteq \mathrm{UC}(T)$ is strict.)

Maximal Transitive Subtournaments: Banks's Solution Among other tournament solutions, one designed by J. Banks in 1985 [13] is of interest. When the tournament being considered, T, is transitive (i. e., T is a linear order), there exists a unique winner who is selected as the winner of T by the usual tournament solutions. If that is not the case, we may consider the transitive subtournaments of T which are maximal with respect to inclusion, and then select the winner of each of them as the winners of T. This defines the *Banks's solution* [13]: a *Banks winner* of T is the winner of any maximal (with respect to inclusion) transitive subtournament of T.

If we consider the majority tournament of Example 2 (see Fig. 7), three Banks winners exist: x (because of the maximal transitive subtournament $x > y > z$), y (because $y > z > t$) and t (because $t > x$).

Tournament Equilibrium Set: T. Schwartz's Solution The solution that we deal with in this subsection was designed by T. Schwartz [158] and is called the *tournament equilibrium set*. To define it, we need extra definitions. Let G be a directed graph. The top set $\mathrm{TS}(G)$ of G is defined as the union of the strongly connected components of G with no in-coming arcs (if G is a tournament, then $\mathrm{TS}(G)$ is equal to $\mathrm{TC}(G)$).

Let *Sol* be a tournament solution and T be a tournament. When *Sol* is applied to T, we define the *contestation graph* $G(Sol, T)$ associated with *Sol* and T as follows: The vertex set of $G(Sol, T)$ is X; there is an arc (x, y) from x to y in $G(Sol, T)$ if and only if (x, y) is an arc of T and x is a winner according to *Sol* when *Sol* is applied to the subtournament of T induced by the predecessors of y in T. In other words, the arcs (x, y) of $G(Sol, T)$ describe the following situation: if y is considered as a possible winner, then x will contest the election of y because x beats y in T and because x is a winner among the candidates who beat y. We may notice that $G(Sol, T)$ is a subgraph of T. By considering the top set $\mathrm{TS}[G(Sol, T)]$ of $G(Sol, T)$, we get a new tournament solution. T. Schwartz proved in [158] that there exists a unique tournament solution, that he called the *tournament equilibrium set* TEQ, which is a fixed point with respect to this process: $\forall T, \mathrm{TEQ}(T) = \mathrm{TS}[G(\mathrm{TEQ}, T)]$.

For the majority tournament of Example 2 (see Fig. 7), the *tournament equilibrium set* contains x, y and t.

Dodgson's Procedure

In the procedure proposed in 1876 by C.L. Dodgson (or Lewis Carroll) [60], each voter ranks all the candidates into a linear order. If a Condorcet winner exists, he or she is also the Dodgson winner. Otherwise, Dodgson's procedure consists of choosing as winners all the candidates who are "closest" to being Condorcet winners: for each candidate $x \in X$, let $D(x)$ be the *Dodgson score of* x, defined as the minimum number of swaps between consecutive candidates in the preferences of the voters such that x becomes a Condorcet winner; a *Dodgson winner* is any candidate x^* minimizing D: $D(x^*) = \mathrm{Min}_{x \in X} D(x)$.

Specifically, consider a profile $\Pi = (L_1, L_2, \ldots, L_m)$ of m linear orders. Let i be an index between 1 and m and let y and z be two consecutive candidates in L_i: yL_iz and there is no candidate t with yL_it and tL_iz simultaneously. Then define a linear order L'_i obtained from L_i by swapping y and z in L_i: zL'_iy and, for any other ordered pair of candidates $(t, v) \neq (y, z)$, we have tL'_iv if and only if we had tL_iv. By substituting L'_i to L_i, we obtain a new profile. By repeating such swaps as many times as necessary, we may generate all the possible profiles with m linear orders. Among them, some admit x as a Condorcet winner. The Dodgson score $D(x)$ of a candidate x is the minimum number of swaps that must be applied from Π in order to obtain a profile with x as a Condorcet winner.

Let us illustrate these definitions on Example 2. First, let us compute the Dodgson scores of the four candidates. If we want x to become a Condorcet winner, it is necessary that x defeats t, which is not currently the case. Since x is never just after t in the preferences of the voters, one swap is not enough to reach this aim. On the other hand, two swaps are enough, for instance by swapping first y and x in the preference of the last voter, and then by swapping x and t in this new preference. Thus: $D(x) = 2$. One swap is sufficient (and necessary) to make y a Condorcet winner, for instance by swapping y and x in the preference of one of the first three voters: $D(y) = 1$. The computation of $D(z)$ is a little more difficult. To become a Con-

dorcet winner, z must defeat y. Since there are only 3 voters who prefer z to y against 6 who prefer y to z, it is necessary to perform at least 2 swaps so that a new majority of voters will prefer z to y. But, if such swaps exist, they do not help to alleviate the difficulty in helping z to defeat x. For z to defeat x necessitates at least one extra swap. Hence the inequality $D(z) \geqslant 3$. On the other hand, swapping x and z in the preference of the voter whose preference is $x > z > y > t$ and swapping y and z in the preferences of two of the first three voters makes z a Condorcet winner: $D(z) = 3$. The computation of $D(t)$ is similar to that of $D(z)$, and we also find $D(t) = 3$. These values show that there is only one Dodgson winner (who is not a Kemeny winner): y.

Young's Procedure

Like Dodgson's procedure, H.P. Young's procedure [177] is based on altered profiles. But, instead of performing the fewest possible number of swaps of consecutive candidates in the voters' preferences, we now find the minimum number of voters that must be removed in order for a Condorcet winner to emerge. More precisely, for each candidate x, we define the *Young score* $Y(x)$ of x as the minimum number of voters whose simultaneous removals allow x to become a Condorcet winner. Any candidate with a minimum Young score is a *Young winner*.

For Example 2, we may verify that the Young scores of the four candidates are: $Y(x) = 2$, $Y(y) = 2$, $Y(z) = 4$, $Y(t) = 4$. For instance, for z to defeat x and y requires at least four removals for y. But these removals must keep z defeating t. It is easy to see that removing two of the first three voters, one of the two voters sharing the same preferences and the voter for whom z is in the last position, suffices to make z a Condorcet winner of the new election.

Approval Voting Procedure, Majority-Choice Approval Procedure and Variants

The approval voting procedure was popularized by Steven J. Brams and Peter C. Fishburn in 1978 [25] and 1983 [26]. But, according to [40,52,111], it was already in use in the 13th century in Venice and in papal elections, and then in elections in England during the 19th century, among other places. Its name seems to comes from R.J. Weber in 1976 (see [175]; several other persons seem to have found this procedure independently during the late 1960s and early 1970s). The approval voting procedure consists of giving to the voters the possibility of voting for several candidates simultaneously. In other words, instead of appointing their preferred candidate, the voters are invited to answer the question: "who are the candidates for whom you would like to vote?". Each voter then gives one point to the candidates with whom he or she agrees. The candidate with the greatest number of points is the winner. Notice that, if each voter actually chooses only one candidate, this voting procedure is the same as the plurality rule. In this respect, the approval voting procedure can be seen as a generalization of the plurality rule.

More formally, this procedure assumes that the preferences of the voters are complete preorders with only two classes (one class for the approved candidates and one for the disapproved ones). These preorders must then be aggregated into a collective complete preorder also with two classes, one of them with only one element (the winner), the other class with all the other elements (the losers). Extension for electing several candidates simultaneously is immediate.

Variants (sometimes known as *range voting*, *ratings summation*, *average voting*, *cardinal ratings*, *0–99 voting*, the *score system* or the *point system*) can be based on the same idea. For instance, each voter has a maximum number of points, and he or she can share them out among the candidates as he or she pleases, with or without a constraint on the maximum number of points per candidate. We can also add several rounds, and thus define new procedures. The *majority-choice approval* procedure (MCA) designed by F. Simmons in 2002 can also be seen as a variant of approval voting. In this system, each voter has three possibilities for rating each candidate: "favored", "accepted" or "disapproved". If a candidate is ranked "favored" by an absolute majority of the voters, then any candidate marked "favored" by a maximum number of voters is a winner. Otherwise, the winner is any candidate with the largest number of "favored" or "accepted" marks. It is sometimes required this number to be at least $(m + 1)/2$, otherwise no one will be elected. Ties can be broken using the number of "favored" marks. We may, of course, obtain new variants by increasing the number of levels.

Bucklin's Procedure

The procedure proposed by James W. Bucklin in the early 20th century is also called the *Grand Junction system* (Grand Junction is a city in Colorado, where Bucklin's procedure was applied from 1909 to 1922). In this procedure, we first count, for each candidate x, the number of times that x is ranked first, as in the plurality rule. If a candidate has at least the absolute majority $(m + 1)/2$, he or she is the winner. Otherwise, second choices are added to the first choices. Once again, if a candidate has the absolute majority $(m + 1)/2$, he or she becomes a winner. Otherwise, we consider third choices and so on, until at least

one candidate obtains the absolute majority: then he or she becomes a winner. Since after the first round there are more votes than voters, Bucklin's procedure is sometimes considered as a variant of approval voting procedure, but a main difference is that Bucklin's procedure may require several rounds. It can also be considered as a variant of scoring procedures, since it is the same as iteratively applying a scoring procedure with successively (1, 0, 0, …, 0), (1, 1, 0, 0, …, 0), (1, 1, 1, 0, …, 0), … as the score-vectors, until a candidate obtains the majority.

Once again, consider Example 2. In the first round, x gets 4 points; y, 2 points; z, 1 point; t, 2 points. As there is no candidate with at least 5 points, we consider the second choices. Then we obtain: x keeps 4 points; y obtains 6 points; z, 3 points; t, 5 points. As y and t obtain at least 5 points, the process stops here and y and t are the Bucklin winners.

A variant of Bucklin's procedure would be to consider, at the last iteration, the candidates with a maximum number of points as the winners (only y for Example 2).

Complexity Results

After a reminder of the complexity classes that are useful for our purposes, this section examines the complexity of the voting procedures described in the previous section (see also [72]). Among those tournament solutions which can be applied to the majority tournament, we restrict ourselves to the solutions depicted above; the reader interested in the complexity of the other common tournament solutions will find some answers in [93].

Main Complexity Classes

Some complexity classesz are well-known: P, NP, co-NP, the sets of polynomial problems, of NP-complete problems, of NP-hard problems… Let us recall some other notations (for references upon the theory of complexity, see for instance [10,78,82,96]; see also [1] for a list of about 500 complexity classes). Like D.S. Johnson in [96], we will distinguish between decision problems and other types of problems. Let n denote the size of the data. The class L^{NP}, or $\mathrm{P}^{\mathrm{NP}[\log]}$, $\mathrm{P}^{\mathrm{NP}[\log n]}$, Θ_2^p, or also $\mathrm{P}_{\parallel}^{\mathrm{NP}}$, contains those decision problems which can be solved by a deterministic Turing machine with an oracle in the following manner: the oracle can solve an appropriate NP-complete problem in unit time, and the number of consultations of the oracle is upper-bounded by $\log(n)$. The class P^{NP} or Δ_2^p (or sometimes simply Δ_2) or also $\mathrm{P}^{\mathrm{NP}[n^{\mathrm{O}(1)}]} = \bigcup_{k\geqslant 0} \mathrm{P}^{\mathrm{NP}[n^k]}$ is defined similarly, but with a polynomial of n instead of $\log(n)$. Notice the inclusions: $\mathrm{NP} \cup \text{co-NP} \subseteq \mathrm{L}^{\mathrm{NP}} \subseteq \mathrm{P}^{\mathrm{NP}}$.

For problems which are not decision problems (i. e., optimization problems in which we look for the optimal value of a given function f, or search problems in which we look for an optimal solution of f, or enumeration problems in which we look for all the optimal solutions of f; these problems are sometimes called *function problems*), these classes are extended with an "F" in front of their names (see [96]); for instance we thus obtain the classes $\mathrm{FL}^{\mathrm{NP}}(= \mathrm{F}\Theta_2^p \ldots)$ or $\mathrm{FP}^{\mathrm{NP}}(= \mathrm{F}\Delta_2^p \ldots)$. The usual definition of "completeness" is extended to these classes in a natural way. Though the notation Δ_2^p or Θ_2^p are sometimes more common in complexity theory (especially when dealing with the polynomial hierarchy), we shall keep the pseudonyms P^{NP} and L^{NP} (as well as $\mathrm{FP}^{\mathrm{NP}}$ and $\mathrm{FL}^{\mathrm{NP}}$), as perhaps being more informative.

Similarly, there exist refinements of P. In particular, the class L denotes the subset of P containing the (decision) problems which can be solved by an algorithm using only logarithmic space (in a deterministic Turing machine), the input itself not being counted as part of memory. The classes AC^0 and TC^0 are more technical. The first one, AC^0, consists of all the problems solvable by uniform constant-depth Boolean circuits with unbounded fan-in and a polynomial number of gates. The second one, TC^0, consists of all the problems solvable by polynomial-size, bounded-depth, unbounded fan-in Boolean circuits augmented by so-called "threshold" gates, i. e., unbounded fan-in gates that output "1" if and only if more than half their inputs are non-zero (see [96] and the references therein for details). A problem of TC^0 is said to be TC^0-complete if it is complete under AC^0 Turing reductions. Notice the inclusions $\mathrm{AC}^0 \subset \mathrm{TC}^0 \subseteq \mathrm{L} \subseteq \mathrm{P}$.

Complexity Results for the Usual Voting Procedures

We can now examine complexity results of the voting procedures described above. When dealing with linear orders, we assume that we can have direct access to the ordered list of the candidates, especially to their winners (see above). Remember that n denotes the number of candidates and m the number of voters.

Theorem 1 *The following procedures are polynomial.*

- *The plurality rule (one-round procedure) is in $O(n + m)$.*
- *The plurality rule with run-off (two-round procedure) is in $O(n + m)$.*
- *The preferential voting procedure (STV) is in $O(nm + n^2)$.*
- *Borda's procedure is in $O(nm)$.*

- *If the preferences of the voters are known through a profile of complete preorders with two classes, the approval voting procedure (and its variant suggested by F. Simmons) is in O(nm).*
- *Bucklin's procedure is in O(nm).*

More generally, the previous polynomial results can usually be extended to procedures based on score-vectors, often with a complexity of $O(nm)$.

The previous results are rather easy to obtain. For instance, for the plurality rule, an efficient way to determine the winners consists of computing an n-vector V which provides, for each candidate x, the number of voters who prefer x. As the preferred candidate of each voter is assumed to be accessible directly, computing V requires $O(n+m)$ operations (or fewer if we assume that voters sharing similar preferences are gathered, see Subsect. "Definitions, Notation and Partially Ordered Sets Used to Model Preferences"). More precisely, we first initialize V to 0 in $O(n)$. Then, for each voter, we consider his or her preferred candidate x and we increment $V(x)$ by 1 in $O(1)$ for each voter, thus in $O(m)$ for all the voters. By scanning V in $O(n)$, we determine the maximum value contained in V and, once again by scanning V in $O(n)$, the winners: they are the candidates x whose values $V(x)$ are maximum. The whole process requires $O(n+m)$ operations.

When it succeeds in finding an order (or at least a Condorcet winner), Condorcet's procedure is also polynomial. This is also the case for Simpson's procedure, since computing the pairwise comparison coefficients m_{xy} can obviously be done in polynomial time. The same happens for Tideman's ranked pairs procedure, since the detection of a circuit in a graph can be done in a time proportional to the number of arcs of this graph, i. e., here in $O(n^2)$ (see [51]), and for the prudent order procedure or the minimax procedure.

Theorem 2 *When there exists a Condorcet winner, Condorcet's procedure is polynomial, in $O(n^2m)$. The prudent orders procedure, the minimax procedure, Simpson's procedure (maximin procedure), and Tideman's procedure (ranked pairs procedure) are also polynomial.*

When there is no Condorcet winner, the situation is more difficult to manage. In this case, we may pay attention to the problems called $P_m(\mathcal{Y},\mathcal{Z})$ above. The problems $P_m(\mathcal{Y},\mathcal{T})$ and $P_m(\mathcal{Y},\mathcal{R})$, i. e., the aggregation of m preferences into a tournament or into a binary relation in which no special property is required, are polynomial for any m and any set $\mathcal{Y}$ as specified by Theorem 3.

Theorem 3 *Let m be any integer with $m \geqslant 1$ and let $\mathcal{Y}$ be any subset of $\mathcal{R}$. Consider a profile $\Pi \in \mathcal{Y}^m$. Then $P_m(\mathcal{Y},\mathcal{R})$ and $P_m(\mathcal{Y},\mathcal{T})$ are polynomial.*

For the Condorcet–Kemeny problem, and more generally for the problems $P_m(\mathcal{Y},\mathcal{Z})$ with $\mathcal{Y}$ belonging to $\{\mathcal{A},\mathcal{C},\mathcal{L},\mathcal{O},\mathcal{P},\mathcal{R},\mathcal{T}\}$ and $\mathcal{Z}$ to $\{\mathcal{A},\mathcal{C},\mathcal{L},\mathcal{O},\mathcal{P}\}$ and where m is a positive integer, most are NP-hard when m is large enough (see [5,16,19,34,41,65,85,89,91,92,94,173,174]; see also [91] for the NP-hardness of similar problems $P_m(\mathcal{Y},\mathcal{Z})$ when extended to other posets, including interval orders, interval relations, semiorders, weak orders, quasi-orders).

Theorem 4 *For m large enough, we have:*

- *For any set $\mathcal{Y}$ containing $\mathcal{L}$ (this is the case in particular for $\mathcal{Y}$ belonging to $\{\mathcal{A},\mathcal{C},\mathcal{L},\mathcal{O},\mathcal{P},\mathcal{R},\mathcal{T}\}$ or to unions or intersections of such sets), $P_m(\mathcal{Y},\mathcal{Z})$ is NP-hard and the decision problem associated with $P_m(\mathcal{Y},\mathcal{Z})$ is NP-complete for $\mathcal{Z} \in \{\mathcal{A},\mathcal{C},\mathcal{L}\}$.*
- *$P_m(\mathcal{R},\mathcal{Z})$ is also NP-hard and the decision problem associated with $P_m(\mathcal{R},\mathcal{Z})$ is also NP-complete for $\mathcal{Z} \in \{\mathcal{O},\mathcal{P}\}$.*
- *The complexity of $P_m(\mathcal{Y},\mathcal{Z})$ is unknown for $\mathcal{Y} \in \{\mathcal{A},\mathcal{C},\mathcal{L},\mathcal{O},\mathcal{P},\mathcal{T}\}$ and for $\mathcal{Z} \in \{\mathcal{O},\mathcal{P}\}$, but $P_m(\mathcal{Y},\mathcal{O})$ and $P_m(\mathcal{Y},\mathcal{P})$ have the same complexity.*

The minimum value of m for which $P_m(\mathcal{Y},\mathcal{Z})$ is NP-hard in Theorem 4 is usually unknown and, moreover, the parity of m plays a role. Table 2 gives the ranges of lower bounds of m from which $P_m(\mathcal{Y},\mathcal{Z})$ is known to be NP-hard; for lower values of m, when not trivial, the complexity of $P_m(\mathcal{Y},\mathcal{Z})$ is usually unknown. (This is the case, for instance, for $P_1(\mathcal{R},\mathcal{C})$, $P_1(\mathcal{R},\mathcal{O})$ or $P_1(\mathcal{R},\mathcal{P})$; notice that $P_2(\mathcal{L},\mathcal{L})$ is polynomial.) A question mark (?) means that the complexity of the problem is still unknown.

So, the Condorcet–Kemeny problem $P_m(\mathcal{L},\mathcal{L})$ is NP-hard for m odd and large enough or for m even with $m \geqslant 4$ (and, as noted above, is polynomial for $m=2$). More specific results deal with this problem $P_m(\mathcal{L},\mathcal{L})$ (see [19,85,92]). To state them, let us define, for each element x of X and for any given profile Π, the *Kemeny score K(x) of x* (with respect to Π) as the minimum remoteness $\Delta(\Pi, L_x)$ between Π and the linear orders L_x with x as their winner. The *Kemeny index of Π*, $K(\Pi)$ is the minimum remoteness between Π and any linear order; thus it is the minimum that we look for in $P_m(\mathcal{L},\mathcal{L})$; it is also the minimum of the Kemeny scores over X.

Theorem 5 *Let Π be a profile of m linear orders with m large enough.*

Voting Procedures, Complexity of, Table 2
Lower bounds of m from which $P_m(\mathcal{Y}, \mathcal{Z})$ is known to be NP-hard

Median relation ($\mathcal{Z}$)	$\mathcal{L} \subseteq \mathcal{Y}$	$\mathcal{L} \subseteq \mathcal{Y}$	$\mathcal{Y} = \mathcal{T}$	$\mathcal{Y} = \mathcal{T}$	$\mathcal{Y} = \mathcal{R}$	$\mathcal{Y} = \mathcal{R}$
	m odd	m even	m odd	m even	m odd	m even
Acyclic relation ($\mathcal{A}$)	$\theta(n^2)$	4	1	2	1	2
Complete preorder ($\mathcal{C}$)	$\theta(n^4)$	$\theta(n^2)$	$\theta(n^3)$	$\theta(n)$	3	2
Linear order ($\mathcal{L}$)	$\theta(n^2)$	4	1	2	1	2
Partial order ($\mathcal{O}$)	?	?	?	?	3	2
Preorder ($\mathcal{P}$)	?	?	?	?	3	2

1. *The following problems are NP-complete:*
   - *Given Π, a candidate $x \in X$ and an integer k, is $K(x)$ lower than or equal to k?*
   - *Given an integer k, is $K(\Pi)$ lower than or equal to k?*
2. *The following problems are NP-hard:*
   - *Given Π and two candidates $x \in X$ and $y \in X$ with $x \neq y$, is $K(x)$ lower than or equal to $K(y)$ (in other words, is x "better" than y ?)? Moreover, if we assume that the profiles Π are given by the ordered list of the m preferences of the voters (and not by a set of different linear orders with their multiplicities, see Subsect. "Definitions, Notation and Partially Ordered Sets Used to Model Preferences"), this problem is L^{NP}-complete.*
   - *Given Π and a candidate $x \in X$, is x a Kemeny winner? Moreover, if we assume that the profiles Π are given by the ordered list of the m preferences of the voters, this problem is L^{NP}-complete.*
   - *Given Π and a candidate $x \in X$, is x the unique Kemeny winner? Moreover, if we assume that the profiles Π are given by the ordered list of the m preferences of the voters, this problem is L^{NP}-complete.*
   - *Given Π, determine a Kemeny winner of Π. Moreover, this problem belongs to FP^{NP}.*
   - *Given Π, determine all the Kemeny winners of Π. Moreover, this problem belongs to FP^{NP}.*
   - *Given Π, determine a Kemeny order of Π. Moreover, this problem belongs to FP^{NP}.*
3. *The following problem belongs to the class co-NP:*
   - *Given Π and a linear order L defined on X, is L a Kemeny order of Π?*

Attention has also been paid to $P_1(\mathcal{T}, \mathcal{L})$, i. e., the approximation of a tournament (which can be the majority tournament of the election) by a linear order at minimum distance (Slater problem [160]). The complexity of the Slater problem derives from a recent result dealing with the problem called the *feedback arc set problem*, which is known to be NP-complete from work by R. Karp [100] for general graphs. This problem consists of removing a minimum number of arcs in a directed graph in order to obtain a graph without any circuits. Recent results [5,34,41] show that this problem remains NP-complete even when restricted to tournaments. For a tournament T, removing a minimum number of arcs to obtain a graph without any circuits is the same as reversing a minimum number of arcs to make T transitive, i. e., a linear order (see [92]). From this, we may prove the following theorem (see [92] for details):

Theorem 6 *For any tournament T, we have the following results:*

- *The computation of the Slater index $i(T)$ of T is NP-hard; this problem belongs to the class FL^{NP}; the associated decision problem is NP-complete.*
- *The computation of a Slater winner of T is NP-hard; this problem belongs to the class FP^{NP}.*
- *Checking that a given vertex is a Slater winner of T is NP-hard; this problem belongs to the class L^{NP}.*
- *The computation of a Slater order of T is NP-hard; this problem belongs to the class FP^{NP}.*
- *The computation of all the Slater winners of T is NP-hard; this problem belongs to the class FP^{NP}.*
- *The computation of all the Slater orders of T is NP-hard.*
- *Checking that a given order is a Slater order is a problem which belongs to the class co-NP.*

Some of the previous results may be generalized to other definitions of remoteness, for instance, if the sum of the symmetric difference distances to the individual preferences is replaced by their maximum, or by the sum of their squares, or by the sum of any of their positive powers, as specified below (see [94]):

Theorem 7 *Let Φ denote any remoteness defined for any profile Π such that, when m is equal to 1 with $\Pi = (R_1)$, then the minimization of $\Phi(\Pi, R)$ over the relations R belonging to $\mathcal{A}$ yields to the same optimal solutions as the minimization of $\delta(R_1, R)$ over the same set. Then, for any fixed m with $m \geqslant 1$, the aggregation of m binary relations*

or m tournaments obtained by minimizing Φ (the problems similar to $P_m(\mathcal{R}, \mathcal{Z})$ and $P_m(\mathcal{T}, \mathcal{Z})$ for $Z \in \{\mathcal{A}, \mathcal{L}\}$ but with respect to Φ) are NP-hard, and the associated decision problems are NP-complete.

An important case exists for which $P_m(\mathcal{L}, \mathcal{L})$ becomes polynomial for any m: it is the one for which the voters' preferences are *single-peaked* linear orders. To define them, assume that we can order the candidates on a line, from left to right, and assume that this linear order Ω does not depend on the voters (from a practical point of view, Ω is not always easy to define, even for political elections). For any voter α, let x_α denote the candidate preferred by α. The preference of α is said to be *Ω-singled-peaked* if, for any candidates y and z with $x \neq y \neq z \neq x$ and located on the same side of Ω from x_α, y is preferred to z by α if and only if y is closer to x_α than z with respect to Ω. Let $\mathcal{U}_\Omega$ denote the set of Ω-single-peaked linear orders. D. Black [23] showed that, for any order Ω, the aggregation of Ω-single-peaked linear orders is an Ω-single-peaked linear order (see also [42] for another study dealing with single-peaked preferences). Hence the polynomiality of the aggregation of Ω-single-peaked linear orders into an Ω-single-peaked linear order:

Theorem 8 *For any linear order Ω defined on X and any positive integer m, for any set $\mathcal{Z}$ containing $\mathcal{U}_\Omega$ as a subset (this is the case for the sets $\mathcal{A}, \mathcal{C}, \mathcal{L}, \mathcal{O}, \mathcal{P}, \mathcal{R}, \mathcal{T}$), $P_m(\mathcal{U}_\Omega, \mathcal{Z})$ is polynomial. More precisely, $P_m(\mathcal{U}_\Omega, \mathcal{Z})$ can be solved in $O(n^2 m)$.*

The NP-hardness of Slater's problems shows that tournament solutions can be difficult to compute. This obviously depends on the solutions considered. For instance, Copeland's procedure is polynomial, as specified by the next theorem.

Theorem 9 *Copeland's procedure is polynomial, in $O(n^2 m)$.*

If we assume that the tournament related to Copeland's procedure is already computed, and if we do not take the memory-space necessary to code this tournament into account, it is easy to see that the memory-space necessary to compute the maximum of the Copeland scores and then to decide whether a given vertex is a Copeland winner can be bounded by a constant. So, deciding whether a given vertex is a Copeland winner is a problem belonging to class L. In fact, a stronger result is shown by F. Brandt, F. Fischer and P. Harrenstein in [29]:

Theorem 10 *Checking that a given vertex is a Copeland winner is a TC^0-complete problem.*

Similar results can be stated for Smith's tournament solution (top cycle):

Theorem 11 *Let T be a tournament.*

- *The computation of the Smith winners of T (the elements of the top cycle of T) can be done in $O(n^2)$ (or even in linear time with respect to the cardinality of the top cycle if the sorted scores are known).*
- *Checking that a given vertex is a Smith winner is a TC^0-complete problem [29].*

For the other tournament solutions depicted above, we have the results stated in Theorem 12 (see [93] for details and for results upon other tournament solutions). This theorem shows that checking whether a given vertex of a given tournament is a Banks winner is NP-complete [176] (and [30] for an alternative proof), while computing a Banks winner is polynomial [90]. Of course, when such a Banks winner is computed, we do not choose the winner that we compute among the set of Banks winners (and so there is no contradiction between the two results if P and NP are different). The polynomiality of the minimal covering set solution (through n resolutions of a linear programming problem, which can be done in polynomial time using L. Khachiyan's algorithm [99]), and the NP-hardness of the tournament equilibrium set solution, are new results respectively due to F. Brandt and F. Fischer [28] and to F. Brandt, F. Fischer, P. Harrenstein and M. Mair [30].

Theorem 12 *Let T be a tournament.*

- *Computing the uncovered elements of T can be done within the same complexity as the multiplication of two $(n \times n)$-matrices, and so can be done in $O(n^{2.38})$ operations.*
- *Computing the elements of the minimal covering set of T can be done in polynomial time with respect to n.*
- *The following problem is NP-complete: given a tournament T and a vertex x of T, is x a Banks winner of T?*
- *Computing a Banks winner is polynomial, and more precisely, can be done in $O(n^2)$ operations.*
- *Computing all the Banks winners of T is an NP-hard problem. More precisely, it is a problem belonging to the class FP^{NP}.*
- *The following decision problem is NP-hard (but is not known to be inside NP): given a vertex x of T, does x belong to TEQ(T)?*

J.J. Bartholdi III, C.A. Tovey and A. Trick stated in [19] that Dodgson's procedure is NP-hard. More precisely,

they proved the NP-completeness of the problem of Theorem 13 and the NP-hardness of the first two problems of Theorem 14. E. Hemaspaandra, L. Hemaspaandra and J. Rothe [84] sharpened their results (Theorem 14), assuming that Π is given by the ordered list of the m preferences of the voters (and not by a set of different linear orders with their multiplicities).

Theorem 13 *The following problem is NP-complete: given a profile Π of linear orders defined on X, a candidate $x \in X$ and an integer k, is the Dodgson score $D(x)$ of x less than or equal to k?*

Theorem 14 *The following problems are NP-hard, and more precisely L^{NP}-complete if the considered profiles Π are given by the ordered lists of the m preferences of the voters.*

- *Given a profile Π of linear orders defined on X and a candidate $x \in X$, is x a Dodgson winner?*
- *Given a profile Π of linear orders defined on X and a candidate $x \in X$, is x the unique Dodgson winner?*
- *Given a profile Π of linear orders defined on X and two candidates $x \in X$ and $y \in X$ with $x \neq y$, is $D(x)$ less than or equal to $D(y)$ (in other words, is x "better" than y)?*

Notice that a variant of Dodgson's procedure, called *homogeneous Dodgson's procedure* (see [76] and [177]) has been shown to be polynomial by J. Rothe, H. Spakowski and J. Vogel [153]. Consider that each voter is replicated p times, each copy having the same preference. A procedure is said to be *homogeneous* if such a replication does not change the winners. Dodgson's procedure and Young's procedure are not homogeneous [76]. Then, instead of the Dodgson score defined above, we may consider, for each candidate x, the limit, when p tends to infinity, of the ratio between, on the one hand, the Dodgson score of x after the replication of each voter p times, and, on the other hand, p. J. Rothe, H. Spakowski and J. Vogel provide in [153] a linear program for computing such a limit for any candidate. Hence the polynomiality of the homogeneous version of Dodgson's procedure.

They give, in the same paper [153], the complexity of Young's procedure.

Theorem 15 *The following problems are NP-hard, and more precisely L^{NP}-complete.*

- *Given a profile Π of linear orders defined on X and a candidate $x \in X$, is x a Young winner?*
- *Given a profile Π of linear orders defined on X and two candidates $x \in X$ and $y \in X$ with $x \neq y$, is $Y(x)$ less than or equal to $Y(y)$ (in other words, is x "better" than y)?*

Notice that all these problems become obviously polynomial if we assume that the number of candidates is upper-bounded by a constant.

Further Directions

Complexity is a prominent feature of voting procedures. NP-hardness may involve too much important CPU time to solve a given instance if the number of candidates is not small. In this respect, polynomial procedures can be preferable to NP-hard ones. In practice, the number of candidates is not always too large, and the instance considered can be tractable. This also depends on the efficiency of the applied algorithms (see [17,35,36,37,41,88,89,98,108,151,152,173] for references on algorithms designed to solve some of the previous problems). Moreover, complexity is stated here in the worst case. A study about the average complexity (for appropriate probability distributions) remains to be done.

Other directions can be investigated.

For instance, the parameterized complexity (see [62] for a global presentation of this field) of the feedback arc set problem applied to tournaments has also been studied. V. Raman and S. Saurabh [148] (see also [61], where the case of bipartite tournaments is also considered) show that the feedback arc set problem for weighted or unweighted tournaments is fixed-parameter tractable (FPT) by providing appropriate algorithms. (Remember that a problem is said to be FPT if there exist an arbitrary function f and a polynomial function Q such that, for any instance (I, k) where I is an instance of the non-parameterized version of the problem and k is the considered complexity parameter, (I, k) can be solved within a CPU time upper-bounded by $f(k)Q(|I|)$, where $|I|$ denotes the size of I. In particular, if k is upper-bounded by a constant, then the problem becomes polynomial.) V. Raman and S. Saurabh give several algorithms to solve the parameterized version of the feedback arc set problem applied to tournaments. The best complexity of their algorithms for this problem is $O(2.415^k n^{\omega})$, where ω denotes the exponent of the running time of the best matrix multiplication algorithm (for instance, in the method designed by D. Coppersmith and S. Winograd [49], ω is about 2.376). Parameterized complexity appears also in [39,115,116] for Kemeny's, Dodgson's and Young's procedures.

Another direction deals with algorithms with approximation guarantees (see [10] or [172] for a presentation of this field). For instance, while the feedback arc set problem is APX-hard in general [69] (remember that, from a practical point of view, this implies that we do not know polynomial-time approximation scheme – PTAS – to solve this

problem), N. Ailon, M. Charikar and A. Newman [3] designed randomized 3-approximation and 2.5-approximation algorithms for the feedback arc set problem when restricted to unweighted tournaments. The 3-approximation algorithm has been derandomized by A. van Zuylen [171], still for unweighted tournaments and with an approximation ratio equal to 3. These methods may be adapted to weighted tournaments with an approximation ratio equal to 5 (see [3] and [50]). N. Alon [5] (see also [2]) shows that, for any fixed $\varepsilon > 0$, it is NP-hard to approximate the minimum size of a feedback arc set for a tournament on n vertices up to an additive error of $n^{2-\varepsilon}$ (but approximating it up to an additive error of εn^2 can be done in polynomial time). Ranking the vertices of a tournament according to their out-degrees (Copeland's procedure) for unweighted tournaments, or according to the sum of the weights of the arcs leaving them minus the sum of the weights of the arcs entering them (Borda's procedure) for weighted tournaments, also provides a method with approximation ratio equal to 5, irrespective of how ties are broken (see [50] and [170]).

Probabilistic algorithms (see [6] for a global presentation of these methods) have also been applied to the feedback arc set problem (or rather to the search for a maximum subdigraph without circuits, which is the same for tournaments), for unweighted [53,57,144,145,163,164,165] or weighted [53] tournaments.

If NP-hardness is usually considered as a drawback (because the CPU time necessary to solve the instances becomes quickly prohibitive, because the algorithms may be difficult to explain to the voters, and so on), it can also be an asset with respect to manipulation, bribery or other attempts to control the election (by adding or deleting candidates or voters, etc.). This is an emerging but already flourishing topic, which becomes a subject on its own (see [18,20,21,33,43,44,45,46,47,66,71,72,73,79,83,86,87,108,114,127,135,137,146,147,155,157,162,169]). A voting procedure is said to be *manipulable* when a voter, who knows how the other voters vote, has the opportunity to benefit by *strategic* or *tactical voting*, i. e., when the voter supports a candidate other than his or her sincerely preferred candidate in order to prevent an undesirable outcome (the main difference between manipulation and bribery is that, for manipulability, the manipulators are known as a part of the instance; for bribery, the number of corrupted voters is bounded but these manipulators are not given). It is known from the theorem by A. Gibbard [79] and M. Satterthwaite [157] that, for at least three candidates, any voting procedure without a dictator is manipulable. Because of this result, manipulation cannot be precluded in any reasonable voting procedure on at least three candidates. As suggested in [18] and in [20] (see also [72]), the situation is less bad if we adopt a procedure for which manipulation is NP-hard. Some procedures are easy to manipulate: this is the case, for instance, for the plurality rule, Borda's procedure, maximin procedure or Copeland's procedure [20] (see also [72] for variants). Some others are NP-hard: this is the case for some variants of some procedures based on score-vectors [44,47,83], or for other procedures [71,87]. But we must remember that NP-hardness refers to the complexity in the worst case; V. Conitzer and T. Sandholm show in [46] that, under certain assumptions, no voting procedure is hard to manipulate, on average.

Among further related issues, we can still mention complexity issues in voting on combinatorial domains (see for instance [104]), or the complexity of determining if the output of a vote is determined even when some voters have not yet been elicited (see for instance [43]).

Computational complexity has become an important criterion in evaluating voting procedures, even if real elections are not necessary the most difficult instances (for example because the number of candidates can be limited). Voting methods are increasingly used in multiagent systems. When autonomous software agents are voting over all sorts of issues, then large numbers of candidates may be more likely, and also agents are possibly more likely to manipulate. Thus, complexity is now a property which must be taken into account along with axiomatic properties. In spite of Arrow's impossibility theorem, or maybe because of it, it is still necessary to design new voting procedures, to study them and to compare them to existing procedures. In the mathematical tool box available to do so, there is now a place for computational complexity.

Acknowledgments

I would like to thank Ulle Endriss, Jérôme Lang and Bernard Monjardet for their help. Their comments were very useful to improve the text.

Bibliography

Primary Literature

1. Aaronson S, Kuperberg G (2008) Complexity Zoo. http://qwiki.caltech.edu/wiki/Complexity_Zoo
2. Ailon N, Alon N (2007) Hardness of fully dense problems. Inf Comput 205:117–1129
3. Ailon N, Charikar M, Newman A (2005) Aggregating inconsistent information: ranking and clustering. Proceedings of the 37th annual ACM symposium on Theory of computing (STOC), pp 684–693
4. Aizerman MA, Aleskerov FT (1995) Theory of choice. North Holland, Elsevier, Amsterdam

5. Alon N (2006) Ranking tournaments. SIAM J Discret Math 20(1):137–142
6. Alon N, Spencer J (2000) The probabilistic method, 2nd edn. Wiley, New York
7. Arrow KJ (1963) Social choice and individual values, rev edn. Wiley, New York
8. Arrow KJ, Raynaud H (1986) Social choice and multicriterion decision-making. MIT Press, Cambridge
9. Arrow KJ, Sen AK, Suzumura K (eds) (2002) Handbook of social choice and welfare, vol 1. North-Holland, Amsterdam
10. Ausiello G, Crescenzi P, Gambosi G, Kann V, Marchetti-Spaccamela A, Protasi M (2003) Complexity and Approximation, 2nd edn. Springer, Berlin
11. Baldwin JM (1926) The technique of the Nanson preferential majority system of election. Proc Royal Soc Victoria 39:42–52
12. Bang-Jensen J, Gutin G (2001) Digraphs: theory, algorithms, and applications. Springer, Berlin
13. Banks J (1985) Sophisticated voting outcomes and agenda control. Soc Choice Welf 2:295–306
14. Barnett WA, Moulin H, Salles M, Schofield NJ (eds) (1995) Social choice, welfare and ethics. Cambridge University Press, New York
15. Barthélemy J-P (1979) Caractérisations axiomatiques de la distance de la différence symétrique entre des relations binaires. Math Sci Hum 67:85–113
16. Barthélemy J-P, Guénoche A, Hudry O (1989) Median linear orders: heuristics and a branch and bound algorithm. Eur J Oper Res 41:313–325
17. Barthélemy J-P, Monjardet B (1981) The median procedure in cluster analysis and social choice theory. Math Soc Sci 1: 235–267
18. Bartholdi III JJ, Orlin J (1991) Single transferable vote resists strategic voting. Soc Choice Welf 8(4):341–354
19. Bartholdi III JJ, Tovey CA, Trick MA (1989) Voting schemes for which it can be difficult to tell who won the election. Soc Choice Welf 6:157–165
20. Bartholdi III JJ, Tovey CA, Trick MA (1989) The Computational Difficulty of Manipulating an Election. Soc Choice Welf 6: 227–241
21. Bartholdi III JJ, Tovey CA, Trick MA (1992) How hard is it to control an election? Math Comput Model 16(8/9):27–40
22. Berge C (1985) Graphs. North-Holland, Amsterdam
23. Black D (1958) The theory of committees and elections. Cambridge University Press, Cambridge
24. Borda J-C (1784) Mémoire sur les élections au scrutin. Histoire de l'Académie Royale des Sciences pour 1781, Paris, pp 657–665. English translation: de Grazia A (1953) Mathematical Derivation of an Election System. Isis 44:42–51
25. Brams SJ, Fishburn PC (1978) Approval Voting. Am Political Sci Rev 72(3):831–857
26. Brams SJ, Fishburn PC (1983) Approval Voting. Birkhauser, Boston
27. Brams SJ, Fishburn PC (2002) Voting Procedures. In: Arrow K, Sen A, Suzumura K (eds) Handbook of Social Choice and Welfare, vol 1. Elsevier, Amsterdam, pp 175–236
28. Brandt F, Fischer F (2008) Computing the minimal covering set. Math Soc Sci 58(2):254–268
29. Brandt F, Fischer F, Harrenstein P (2006) The computational complexity of choice sets. In: Endriss U, Lang J (eds) Proceedings of the conference Computational Social Choice 2006. University of Amsterdam, The Netherlands, pp 63–76
30. Brandt F, Fischer F, Harrenstein P, Mair M (2008) A computational analysis of the tournament equilibrium set. In: Fox D, Gomes CP (eds) Proc. of AAAI, pp. 38–43
31. Caritat MJAN, marquis de Condorcet (1785) Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. Imprimerie Royale, Paris
32. Caspard N, Monjardet B, Leclerc B (2007) Ensembles ordonnés finis: concepts, résultats et usages. Springer, Berlin
33. Chamberlin JR (1985) An investigation into the effective manipulability of four voting systems. Behav Sci 30:195–203
34. Charbit P, Thomassé S, Yeo A (2007) The minimum feedback arc set problem is NP-hard for tournaments. Comb Probab Comput 16(1):1–4
35. Charon I, Hudry O (2006) A branch and bound algorithm to solve the linear ordering problem for weighted tournaments. Discret Appl Math 154:2097–2116
36. Charon I, Hudry O (2007) A survey on the linear ordering problem for weighted or unweighted tournaments. 4OR 5(1):5–60
37. Charon I, Guénoche A, Hudry O, Woirgard F (1997) New results on the computation of median orders. Discret Math 165–166:139–154
38. Chevaleyre Y, Endriss U, Lang J, Maudet N (2007) A short introduction to computational social choice. Proceedings of the 33rd Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM-2007). Lecture Notes in Computer Science, vol 4362. Springer, Berlin, pp 51–69
39. Christian R, Fellows M, Rosamond F, Slinko A (2006) On complexity of lobbying in multiple referenda. Proceedings of the First International Workshop on Computational Social Choice (COMSOC 2006). University of Amsterdam, pp 87–96
40. Colomer JM, McLean I (1998) Electing Popes: Approval Balloting and Qualified-Majority Rule. J Interdisciplinary History 29(1):1–22
41. Conitzer V (2006) Computing Slater Rankings Using Similarities Among Candidates. In: Proceedings of the 21st National Conference on Artificial Intelligence, AAAI-06, Boston, pp. 613–619
42. Conitzer V (2007) Eliciting single-peaked preferences using comparison queries. Proceedings of the 6th International Joint Conference on Autonomous Agents and Multi Agent Systems (AAMAS-07). Honolulu, USA, pp 408–415
43. Conitzer V, Sandholm T (2002) Vote elicitation: complexity and strategy-proofness. Proceedings of the National Conference on Artificial Intelligence (AAAI), pp 392–397
44. Conitzer V, Sandholm T (2002) Complexity of manipulating elections with few candidates. Proceedings of the 18th National Conference on Artificial Intelligence (AAAI), pp 314–319
45. Conitzer V, Sandholm T (2003) Universal voting protocal tweaks to make manipulation hard. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03), Acapulco, Mexico, pp 781–788
46. Conitzer V, Sandholm T (2006) Nonexistence of voting rules that are usually hard to manipulate. Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06), Boston, pp 627–634
47. Conitzer V, Lang J, Sandholm T (2003) How many candidates are needed to make elections hard to manipulate? Theoretical Aspects of Rationality and Knowledge (TARK), pp 201–214
48. Copeland AH (1951) A "reasonable" social welfare function.

Seminar on applications of mathematics to the social sciences. University of Michigan
49. Coppersmith T, Winograd S (1987) Matrix multiplication via arithmetic progression. In: Proc 19th Ann ACM Symp Theor Comput, pp 1–6
50. Coppersmith D, Fleischer L, Rudra A (2006) Ordering by weighted number of wins gives a good ranking for weighted tournaments. Proceedings of the 17th annual ACM-SIAM symposium on discrete algorithms (SODA'06), pp 776–782
51. Cormen T, Leiserson C, Rivest R (1990) Introduction to algorithms, 2nd edn 2001. MIT Press, Cambridge
52. Cox GW (1987) The Cabinet and the Development of Political Parties in Victorian England. Cambridge University Press, New York
53. Czygrinow A, Poljak S, Rödl V (1999) Constructive quasi-Ramsey numbers and tournament ranking. SIAM J Discret Math 12(1):48–63
54. Daunou PCF (1803) Mémoire sur les élections au scrutin. Baudoin, Paris, an XI
55. Debord B (1987) Caractérisation des matrices de préférences nettes et méthodes d'agrégation associées. Math Sci Hum 97:5–17
56. Debord B (1987) Axiomatisation de procédures d'agrégation de préférences. Ph D thesis, Université scientifique technologique et médicale de Grenoble
57. de la Vega WF (1983) On the maximal cardinality of a consistent set of arcs in a random tournament. J Comb Theor B 35:328–332
58. Dodgson CL (1873) A Discussion of the Various Methods of Procedure in Conducting Elections. Imprint by Gardner EB, Hall EP, Stacy JH. Printers to the University, Oxford. Reprinted In: Black D (1958) The Theory of Committees and Elections. Cambridge University Press, Cambridge, pp 214–222
59. Dodgson CL (1874) Suggestions as to the Best Method of Taking Votes, Where More than Two Issues are to be Voted on. Imprint by Hall EP, Stacy JH. Printers to the University, Oxford. Reprinted In: Black D (1958) The Theory of Committees and Elections. Cambridge University Press, Cambridge, pp 222–224
60. Dodgson CL (1876) A method of taking votes on more than two issues, Clarendon Press, Oxford. Reprint In: Black D (1958) The theory of committees and elections, Cambridge University Press, Cambridge, pp 224–234; and In: McLean I, Urken A (1995) Classics of social choice. University of Michigan Press, Ann Arbor
61. Dom M, Guo J, Hüffner F, Niedermeier R, Truß A (2006) Fixed-Parameter Tractability Results for Feedback Set Problems in Tournaments. Lecture Notes in Computer Science, vol 3998. Springer, Berlin, pp 320–331
62. Downey RG, Fellows MR (1999) Parameterized complexity. Springer, Berlin
63. Dummett M (1984) Voting Procedures. Clarendon Press, Oxford
64. Dutta B (1988) Covering sets and a new Condorcet choice correspondence. J Economic Theory 44:63–80
65. Dwork C, Kumar R, Naor M, Sivakumar D (2001) Rank aggregation methods for the Web. Proceedings of the 10th international conference on World Wide Web (WWW10), Hong Kong, pp 613–622
66. Elkin E, Lipmaa H (2006) Hybrid voting protocols and hardness of manipulation. In: Endriss U, Lang J (eds) Proceedings of the First International Workshop on Computational Social Choice (COMSOC 2006). University of Amsterdam, pp 178–191
67. Elster J, Hylland A (eds) (1986) Foundations of Social Choice Theory. Cambridge University Press, New York
68. Erdös P, Moser L (1964) On the representation of directed graphs as unions of orderings. Magyar Tud Akad Mat Kutato Int Közl 9:125–132
69. Even G, Naor JS, Sudan M, Schieber B (1998) Approximating minimum feedback sets and multicuts in directed graphs. Algorithmica 20(2):151–174
70. Fagin R, Kumar R, Mahdian M, Sivakumar D, Vee E (2005) Rank Aggregation: An Algorithmic Perspective. unpublished manuscript
71. Faliszewski P, Hemaspaandra E, Hemaspaandra L (2006) The complexity of bribery in elections. In: Endriss U, Lang J (eds) Proceedings of the First International Workshop on Computational Social Choice (COMSOC 2006). University of Amsterdam, pp 178–191
72. Faliszewski P, Hemaspaandra E, Hemaspaandra L, Rothe J (2006) A richer understanding of the complexity of election systems. Technical report TR-2006-903, University of Rochester. To appear in: Ravi S, Shukla S (eds) (2008) Fundamental Problems in Computing: Essays in honor of Professor Daniel J. Rosenkrantz. Springer, Berlin
73. Faliszewski P, Hemaspaandra E, Hemaspaandra L, Rothe J (2007) Llull and Copeland Voting Broadly Resist Bribery and Control. Technical report TR-2006-903. University of Rochester, NY
74. Fishburn PC (1973) Interval representations for interval orders and semiorders. J Math Psychol 10:91–105
75. Fishburn PC (1973) The theory of social choice. Princeton University Press, Princeton
76. Fishburn PC (1977) Condorcet social choice functions. SIAM J Appl Math 33:469–489
77. Fishburn PC (1985) Interval orders and interval graphs, a study of partially ordered sets. Wiley, New York
78. Garey MR, Johnson DS (1979) Computers and intractability, a guide to the theory of NP-completeness. Freeman, New York
79. Gibbard A (1973) Manipulation of voting schemes. Econometrica 41:587–602
80. Guilbaud GT (1952) Les théories de l'intérêt général et le problème logique de l'agrégation. Économie Appl 5(4):501–584; Éléments de la théorie des jeux, 1968. Dunod, Paris
81. Hägele G, Pukelsheim F (2001) Llull's writings on electoral systems. Studia Lulliana 3:3–38
82. Hemaspaandra L (2000) Complexity classes. In: Rosen KH (ed) Handbook of discrete and combinatorial mathematics. CRC Press, Boca Raton, pp 1085–1090
83. Hemaspaandra E, Hemaspaandra L (2007) Dichotomy for voting systems. J Comput Syst Sci 73(1):73–83
84. Hemaspaandra E, Hemaspaandra L, Rothe J (1997) Exact analysis of Dodgson elections: Lewis Carroll's 1876 voting system is complete for parallel access to NP. J ACM 44(6):806–825
85. Hemaspaandra E, Spakowski H, Vogel J (2005) The complexity of Kemeny elections. Theor Comput Sci 349:382–391
86. Hemaspaandra E, Hemaspaandra L, Rothe J (2006) Hybrid elections broaden complexity-theoretic resistance to control. Proceedings of the First International Workshop on Computational Social Choice (COMSOC 2006), University of Amster-

dam, pp 234–247; (2007) Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007). AAAI Press, pp 1308–1314
87. Hemaspaandra E, Hemaspaandra L, Rothe J (2007) Anyone but him: the complexity of precluding an alternative. Artif Intell 171(5–6):255–285
88. Homan C, Hemaspaandra L (2006) Guarantees for the success frequency of an algorithm for finding Dodgson-election winners. Proceedings of the 31st International Symposium on Mathematical Foundations of Computer Science. Lecture Notes in Computer Science, vol 4162. Springer, Berlin, pp 528–539
89. Hudry O (1989) Recherche d'ordres médians: complexité, algorithmique et problèmes combinatoires. Ph D thesis, ENST, Paris
90. Hudry O (2004) A note on Banks winners. In: Woeginger GJ (ed) tournaments are difficult to recognize. Soc Choice Welf 23:1–2
91. Hudry O (2008) NP-hardness results on the aggregation of linear orders into median orders. Ann Oper Res 163(1):63–88
92. Hudry O (2008) NP-hardness of Slater and Kemeny problems. (submitted)
93. Hudry O (2008) A survey on the complexity of tournament solutions. Math Soc Sci (to appear)
94. Hudry O (2008) Complexity of the aggregation of binary relations. (submitted)
95. Inada K (1969) The Simple Majority Decision Rule. Econometrica 37:490–506
96. Johnson DS (1990) A catalog of complexity classes. In: van Leeuwen J (ed) Handbook of Theoretical Computer Science Vol. A: Algorithms and Complexity. Elsevier, Amsterdam, pp 67–161
97. Johnson PE (1998) Social Choice Theory and Research, CA (Quantitative Applications in the Social Sciences Series, vol 123). Sage Publications, Thousand Oaks
98. Jünger M (1985) Polyhedral combinatorics and the acyclic subdigraph problem. Heldermann, Berlin
99. Khachiyan L (1979) A polynomial algorithm in linear programming. Sov Math Dokl 20:191–194
100. Karp RM (1972) Reducibility among combinatorial problems. In: Miller RE, Tatcher JW (eds) Complexity of computer computations. Plenum Press, New York, pp 85–103
101. Kelly JS (1987) Social Choice Theory: an Introduction. Springer, Berlin
102. Kemeny JG (1959) Mathematics without Numbers. Daedalus 88:571–591
103. Köhler G (1978) Choix multicritère et analyse algébrique de données ordinales. Ph D thesis, université scientifique et médicale de Grenoble
104. Lang J (2004) Logical preference representation and combinatorial vote. Ann Math Artif Intell 42:37–71
105. Laplace (marquis de) PS (1795) Journal de l'École Polytechnique, tome II vol. 7–8; Théorie analytique des probabilités. Essai philosophique sur les probabilités. Œuvres de Laplace, tome VII, Paris, 1847
106. Laslier J-F (1997) Tournament Solutions and Majority Voting. Springer, Berlin
107. Laslier J-F (2004) Le vote et la règle majoritaire. Analyse mathématique de la politique éditions du CNRS
108. LeGrand R, Markakis E, Mehta A (2006) Approval voting: local search heuristics and approximation algorithms for the minimax solution. Proceedings of the First International Workshop on Computational Social Choice (COMSOC 2006), University of Amsterdam, pp 234–247
109. Levenglick A (1975) Fair and reasonable election systems. Behav Sci 20:34–46
110. Levin J, Nalebuff B (1995) An introduction to vote-counting schemes. J Economic Perspectives 9(1):3–26
111. Lines M (1986) Approval Voting and Strategy Analysis: A Venetian. Ex Theor Decis 20:155–172
112. Lhuilier S (1794) Examen du mode d'élection proposé à la Convention nationale de France en février 1793 et adopté à Genève, Genève. Reprint In: (1976) Math Sci Hum 54:7–24
113. Mascart J (1919) La vie et les travaux du chevalier Jean-Charles de Borda (1733–1799): épisodes de la vie scientifique au XVIIIe siècle. Annales de l'université de Lyon vol. II (33). New edition, Presses de l'université de Paris-Sorbonne, 2000
114. Maus S, Peters H, Storcken T (2006) Anonymous voting and minimal manipulability. Proceedings of the First International Workshop on Computational Social Choice (COMSOC 2006), University of Amsterdam, pp 317–330
115. McCabe-Dansted J (2006) Feasibility and approximability of Dodgson's rule. Master's thesis, University of Auckland
116. McCabe-Dansted J, Pritchard G, Slinko A (2006) Approximability of Dodgson's rule. Proceedings of the First International Workshop on Computational Social Choice (COMSOC 2006), University of Amsterdam, pp 234–247
117. Mc Garvey D (1953) A theorem on the construction of voting paradoxes. Econometrica 21:608–610
118. McKey B (2006) http://cs.anu.edu.au/pp~bdm/data/digraphs.html
119. McLean I (1995) The first golden age of social choice, 1784–1803. In: Barnett WA, Moulin H, Salles M, Schofield NJ (eds) Social choice welfare, and ethics: proceedings of the eighth international symposium in economic theory and econometrics. Cambridge University Press, Cambridge, pp 13–33
120. McLean I, Hewitt F (1994) Condorcet: Foundations of Social choice and Political Theory. Edward Elgar, Hants
121. McLean I, Urken A (1995) Classics of social choice. University of Michigan Press, Ann Arbor
122. McLean I, Urken A (1997) La réception des œuvres de Condorcet sur le choix social (1794–1803): Lhuilier, Morales et Daunou, in Condorcet, Homme des Lumières et de la Révolution, Chouillet A-M, Pierre Crépel (eds) ENS éditions, Fontenay-aux-roses, pp 147–160
123. McLean I, McMillan A, Monroe BL (1995) Duncan Black and Lewis Carroll. J Theor Politics 7:107–124
124. McLean I, Lorrey H, Colomer JM (2007) Social Choice in Medieval Europe. workshop Histoire des Mathématiques Sociales, Paris
125. Merrill III S, Grofman B (1999) A Unified Theory of Voting. Cambridge University Press, Cambridge
126. Miller N (1980) A new solution set for tournaments and majority voting: Further graph-theoretical approaches to the theory of voting. Am J Political Sci 24(1):68–96
127. Mitlöhner J, Eckert D, Klamler C (2006) Simulating the effects of misperception on the manipulability of voting rules. Proceedings of the First International Workshop on Computational Social Choice (COMSOC 2006), University of Amsterdam, pp 234–247
128. Monjardet B (1976) Lhuilier contre Condorcet au pays des paradoxes. Math Sci Hum 54:33–43

129. Monjardet B (1979) Relations à éloignement minimum de relations binaires, note bibliographique. Math Sci Hum 67:115–122
130. Monjardet B (1990) Sur diverses formes de la "règle de Condorcet" d'agrégation des préférences. Math Inf Sci Hum 111:61–71
131. Monjardet B (2008) Acyclic domains of linear orders: a survey. (to appear)
132. Monjardet B (2008) Mathématique Sociale and Mathematics. A case study: Condorcet's effect and medians. Electron J Hist Probab Stat 4(1):1–26
133. Monjardet B Private communication
134. Moon JW (1968) Topics on tournaments. Holt, Rinehart and Winston, New York
135. Moulin H (1980) On strategy-proofness and single peakedness. Public Choice 35:437–455
136. Moulin H (1983) The Strategy of Social Choice. North Holland, Amsterdam
137. Moulin H (1985) Fairness and strategy in voting. In: Young HP (ed) Fair Allocation, American Mathematical Society. Proc Symp Appl Math 33:109–142
138. Moulin H (1986) Choosing from a tournament. Soc Choice Welf 3:272–291
139. Morales JI (1797) Memoria matemática sobre el cálculo de la opinión en las elecciones. Imprenta Real, Madrid. Translated in: McLean I, Urken A (1995) Classics of social choice. University of Michigan Press, Ann arbor
140. Nanson EJ (1882) Methods of Election. Trans Proc Royal Soc Victoria 18:197–240
141. Nurmi H (1987) Comparing Voting Systems. D. Reidel Publishing Company, Dordrecht
142. Orlin J (1981) unpublished
143. Pattanaik PK, Salles M (eds) (1983) Social Choice and Welfare. North-Holland, Amsterdam
144. Poljak S, Turzík D (1986) A polynomial time heuristic for certain subgraph optimization problems with guaranteed lower bound. Discret Math 58:99–104
145. Poljak S, Rödl V, Spencer J (1988) Tournament ranking with expected profit in polynomial time. SIAM J Discret Math 1(3):372–376
146. Procaccia A, Rosenschein J (2006) Junta distribution and the average-case complexity of manipulating elections. Proceedings of the 5th International Joint Autonomous Agents and Multiagent Systems, ACM Press, pp 497–504
147. Procaccia A, Rosenschein J, Zohar A (2006) Multi-winner elections: complexity of manipulation, control, and winner-determination. Proceedings of the 8th Trading Agent Design and Analysis and Agent Mediated Electronic Commerce Joint International Workshop (TADA/AMEC 2006), pp 15–28
148. Raman V, Saurabh S (2006) Parameterized algorithms for feedback set problems and their duals in tournaments. Theor Comput Sci 351:446–458
149. Reid KB (2004) Tournaments. In: Gross JL, Yellen J (eds) Handbook of Graph Theory. CRC Press, Boca Raton, pp 156–184
150. Reid KB, Beineke LW (1978) Tournaments. In: Beineke LW, Wilson RJ (eds) Selected topics in graph theory. Academic Press, London, pp 169–204
151. Reinelt G (1985) The linear ordering problem: algorithms and applications. Research and Exposition in Mathematics 8. Heldermann, Berlin
152. Rothe J, Spakowski H (2006) On determining Dodgson winners by frequently self-knowingly correct algorithms and in average-case polynomial time. Proceedings of the First International Workshop on Computational Social Choice (COMSOC 2006), University of Amsterdam, pp 234–247
153. Rothe J, Spakowski H, Vogel J (2003) Exact complexity of the winner problem for Young elections. Theor Comput Syst 36(4):375–386
154. Rowley CK (ed) (1993) Social Choice Theory, vol 1: The Aggregation of Preferences. Edward Elgar Publishing Company, London
155. Saari D (1990) Susceptibility to manipulation. Public Choice 64:21–41
156. Saari D (2001) Decisions and Elections, Explaining the Unexpected. Cambridge University Press, Cambridge
157. Satterthwaite M (1975) Strategy-proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions. J Econ Theor 10:187–217
158. Schwartz T (1990) Cyclic tournaments and cooperative majority voting: A solution. Soc Choice Welf 7:19–29
159. Simpson PB (1969) On Defining Areas of Voter Choice. Q J Econ 83(3):478–490
160. Slater P (1961) Inconsistencies in a schedule of paired comparisons. Biometrika 48:303–312
161. Smith JH (1973) Aggregation of preferences with variable electorate. Econometrica 41(6):1027–1041
162. Smith D (1999) Manipulability measures of common social choice functions. Soc Choice Welf 16:639–661
163. Spencer J (1971) Optimal ranking of tournaments. Networks 1:135–138
164. Spencer J (1978) Nonconstructive methods in discrete mathematics. In: Studies in Combinatorics, Rota GC (ed) Mathematical Association of America, Washington, pp 142–178
165. Spencer J (1987) Ten lectures on the probabilistic method. CBMS-NSF Regional Conference Series in Applied Mathematics N 52, SIAM, Philadelphia
166. Stearns R (1959) The voting problem. Am Math Monthly 66:761–763
167. Straffin PD Jr. (1980) Topics in the Theory of Voting. Birkhäuser, Boston
168. Taylor AD (1995) Mathematics and Politics Strategy, Voting, Power, and Proof. Springer, Berlin
169. Taylor AD (2005) Social Choice and the Mathematics of Manipulation. Cambridge University Press, Cambridge
170. Tideman TN (1987) Independence of clones as criterion for voting rules. Soc Choice Welf 4:185–206
171. van Zuylen A (2005) Deterministic approximation algorithms for ranking and clusterings. Cornell ORIE Tech Report No. 1431
172. Vazirani VV (2003) Approximation Algorithms. Springer, Berlin
173. Wakabayashi Y (1986) Aggregation of binary relations: algorithmic and polyhedral investigations. Ph D thesis, Augsburg
174. Wakabayashi Y (1998) The Complexity of Computing Medians of Relations. Resenhas 3(3):323–349
175. Weber RJ (1995) Approval voting. J Econ Perspectives 9(1):39–49
176. Woeginger GJ (2003) Banks winner in tournaments are difficult to recognize. Soc Choice Welf 20:523–528
177. Young HP (1977) Extending Condorcet's rule. J Economic Theor 16(2):335–353

Books and Reviews

Aleskerov FT (1999) Arrovian Aggregation Models, Mathematical and statistical methods. Theory and decision library, vol 39. Kluwer, Boston

Aleskerov FT, Monjardet B (2002) Utility Maximisation, Choice and preference. Springer, Berlin

Baker KM (1975) Condorcet from Natural Philosophy to Social Mathematics. The University of Chicago Press, Chicago. Reissued 1982

Balinski M, Young HP (1982) Fair Representation. Yale University Press, New Haven

Barthélemy J-P, Monjardet B (1988) The median procedure in data analysis: new results and open problems. In: Bock HH (ed) Classification and related methods of data analysis. North Holland, Amsterdam

Batteau P, Jacquet-Lagrèze É, Monjardet B (eds) (1981) Analyse et agrégation des préférences dans les sciences économiques et de gestion. Economica, Paris

Black D (1996) Formal contributions to the theory of public choice. In: Brady GL, Tullock G (eds) The unpublished works of Duncan Black. Kluwer, Boston

Bouyssou D, Marchant T, Pirlot M, Tsoukias A, Vincke P (2006) Evaluation and decision models with multiple criteria. Springer, Berlin

Campbell DE (1992) Equity, Efficiency, and Social Choice. Clarendon Press, Oxford

Coughlin P (1992) Probabilistic Voting Theory. Cambridge University Press, Cambridge

Danilov V, Sotskov A (2002) Social Choice Mechanisms. Springer, Berlin

Dubois D, Pirlot M, Bouyssou D, Prade H (eds) (2006) Concepts et méthodes pour l'aide à la décision. Hermès, Paris

Endriss U, Lang J (eds) (2006) Proceedings of the First International Workshop on Computational Social Choice, COMSOC 2006, University of Amsterdam

Enelow J, Hinich M (eds) (1990) Advances in the Spatial Theory of Voting. Cambridge University Press, Cambridge

Farquharson R (1969) Theory of Voting. Yale University Press, New Haven

Feldman AM (1980) Welfare Economics and Social Choice Theory. Martinus Nijhoff, Boston

Felsenthal DS, Machover M (1998) The Measurement of Voting Power: Theory and Practise, Problems and Paradoxes. Edward Elgar, Cheltenham

Gaertner W (2001) Domains Conditions in Social Choice Theory. Cambridge University Press, Cambridge

Greenberg J (1990) The Theory of Social Situations. Cambridge University Press, Cambridge

Grofman B (1981) When is the Condorcet Winner the Condorcet Winner? University of California, Irvine

Grofman B, Owen G (eds) (1986) Information Pooling and Group Decision Making. JAI Press, Greenwich

Heal G (ed) (1997) Topological Social Choice. Springer, Berlin

Hillinger C (2004) Voting and the cardinal aggregation cardinal of judgments. Discussion papers in economics 353, University of Munich

Holler MJ (ed) (1978) Power Voting and Voting Power. Physica, Wurtsburg-Wien

Holler MJ, Owen G (eds) (2001) Indices and Coalition Formation. Kluwer, Boston

Kemeny J, Snell L (1960) Mathematical models in the social sciences. Ginn, Boston

Laslier J-F (2006) Spatial Approval Voting. Political Analysis 14(2):160–185

Laslier J-F, Van Der Straeten K (2008) A live experiment on approval voting. Exp Econ 11:97–105

Lieberman B (ed) (1971) Social Choice. Gordon and Breach, New York

Mirkin BG (1979) Group Choice. Winston, Washington

Moulin H (2003) Fair Division and Collective Welfare. Institute of Technology Press, Boston

Nurmi H (1999) Voting Paradoxes and how to Deal with Them. Springer, Berlin

Nurmi H (2002) Voting Procedures under Uncertainty. Springer, Berlin

Pattanaik PK (1971) Voting and Collective Choice. Harvard University Press, Cambridge

Pattanaik PK (1978) Strategy and Group Choice. North Holland, Amsterdam

Peleg B (1984) Game Theoretic Analysis of Voting in Committes. Cambridge University Press, Cambridge

Rothschild E (2001) Economic Sentiments: Adam Smith, Condorcet, and the Enlightenment. Harvard University Press, Cambridge

Saari DG (1994) Geometry of Voting. Springer, Berlin

Saari DG (1995) Basic Geometry of Voting. Springer, Berlin

Saari DG (2000) Chaotic Elections! American Mathematical Society, Providence

Schofield N (1984) Social Choice and Democracy. Springer, Berlin

Schofield N (ed) (1996) Collective Decision Making: Social Choice and Political Economy. Kluwer, Boston

Schwartz T (1986) The Logic of Collective Choice. Columbia University Press, New York

Sen AK (1979) Collective Choice and Social Welfare. North Holland, Amsterdam

Sen AK (1982) Choice, Welfare and Measurement. Basil Blackwell, Oxford

Suzumura K (1984) Rational Choice, Collective Decisions and Social Welfare. Cambridge University Press, Cambridge

Tanguiane AS (1991) Aggregation and Representation of Preferences, Introduction to Mathematical Theory of Democracy. Springer, Berlin

Tideman N (2006) Collective Decisions and Voting: The Potential for Public Choice. Ashgate, Burlington

Woodall DR (1997) Monotonicity of single-seat preferential election rules. Discret Appl Math 77:81–98

Young HP (1974) An Axiomatization of Borda's Rule. J Econ Theor 9:43–52

Young HP (1986) Optimal ranking and choice from pairwise comparisons. In: Grofman B, Owen G (eds) Information pooling and group decision making. JAI Press, Greenwich, pp 113–122

Young HP (1988) Condorcet Theory of Voting. Am Political Sci Rev 82:1231–1244

Young HP (1995) Optimal voting rules. J Econ Perspect 9(1): 51–64

Young HP, Levenglick A (1978) A Consistent Extension of Condorcet's Election Principle. SIAM J Appl Math 35:285–300

W

Water Waves and the Korteweg–de Vries Equation

Lokenath Debnath
Department of Mathematics,
University of Texas – Pan American, Edinburg, USA

Article Outline

Glossary
Definition of the Subject
Introduction
The Euler Equation of Motion in Rectangular Cartesian and Cylindrical Polar Coordinates
Basic Equations of Water Waves with Effects of Surface Tension
The Stokes Waves and Nonlinear Dispersion Relation
Surface Gravity Waves on a Running Stream in Water
History of Russell's Solitary Waves and Their Interactions
The Korteweg–de Vries and Boussinesq Equations
Solutions of the KdV Equation: Solitons and Cnoidal Waves
Derivation of the KdV Equation from the Euler Equations
Two-Dimensional and Axisymmetric KdV Equations
The Nonlinear Schrödinger Equation and Solitary Waves
Whitham's Equations of Nonlinear Dispersive Waves
Whitham's Instability Analysis of Water Waves
The Benjamin–Feir Instability of the Stokes Water Waves
Future Directions
Bibliography

Glossary

Axisymmetric (concentric) KdV equation This is a partial differential equation for the free surface $\eta(R,t)$ in the form $(2\eta_R + \frac{1}{R}\eta + 3\eta\eta_\xi) + \frac{1}{3}\eta_{\xi\xi\xi} = 0$.

Benjamin–Feir instability of water waves This describes the instability of nonlinear water waves.

Bernoulli's equation This is a partial differential equation which determines the pressure in terms of the velocity potential in the form $\phi_t + \frac{1}{2}(\nabla\phi)^2 + \frac{P}{\rho} + gz = 0$, where ϕ is the velocity potential, P is the pressure, ρ is the density and g is the acceleration due to gravity.

Boussinesq equation This is a nonlinear partial differential equation in shallow water of depth h given by

$$u_{tt} - c^2 u_{xx} + \frac{1}{2}\left(u^2\right)_{xx} = \frac{1}{3}h^2 u_{xxtt}\,, \qquad \text{where } c^2 = \sqrt{gh}\,.$$

Cnoidal waves Waves are represented by the Jacobian elliptic function $cn(z,m)$.

Continuity equation This is an equation describing the conservation of mass of a fluid. More precisely, this equation for an incompressible fluid is $div\,\mathbf{u} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0$, where $\mathbf{u} = (u,v,w)$ is the velocity field, and $\mathbf{x} = (x,y,z)$.

Continuum hypothesis It requires that the velocity $\mathbf{u} = (u,v,w)$, pressure p and density ρ are continuous functions of position $\mathbf{x} = (x,y,z)$ and time t.

Crapper's nonlinear capillary waves Pure progressive capillary waves of arbitrary amplitude.

Dispersion relation A mathematical relation between the wavenumber, frequency and/or the amplitude of a wave.

Euler equations This is a nonlinear partial differential equation for an inviscid incompressible fluid flow governed by the velocity field $\mathbf{u} = (u,v,w)$ and pressure $P(\mathbf{x},t)$ under the external force $\mathbf{F}$. More precisely, $\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} = -\frac{1}{\rho}\nabla P + \mathbf{F}$, where ρ is the constant density of the fluid.

Group velocity The velocity defined by the derivative of the frequency with respect to the wavenumber ($c_g = \mathrm{d}\omega/\mathrm{d}k$).

Johnson's equation This is a nearly concentric KdV equation for $\eta(R,\xi,\theta)$ in cylindrical polar coordinates in the form

$$\left(2\eta_R + \frac{1}{R}\eta + 3\eta\eta_\xi + \frac{1}{3}\eta_{\xi\xi\xi} + \frac{1}{R^2}\eta_{\theta\theta}\right) = 0\,.$$

Kadomtsev–Petviashvili (KP) equation This is a two-dimensional KdV equation in the form

$$\left(2\eta_t + 3\eta\eta_\xi + \frac{1}{3}\,\eta_{\xi\xi\xi}\right)_\xi + \eta_{yy} = 0\,.$$

KdV–Burgers equation This is a nonlinear partial differential equation in the form

$$\eta_t + c_0\,\eta_x + d\,\eta\,\eta_x + \mu\,\eta_{xxx} - \nu\,\eta_{xx} = 0\,,$$
$$\text{where } \mu = \frac{1}{6}\,c_0 h_0^2\,.$$

Korteweg–de Vries (KdV) equation This is a nonlinear partial differential equation for a solitary wave (or soliton). This equation in a shallow water of depth h is governed by the free surface elevation $\eta(x,t)$ in the form $\frac{\partial\eta}{\partial t} + c(1 + \frac{3}{2h}\,\eta)\eta_x + \left(\frac{ch^2}{6}\right)\eta_{xxx} = 0$, where $c = \sqrt{gh}$ is the shallow water speed.

Laplace equation This is partial differential equation of the form $\nabla^2\phi = \phi_{xx} + \phi_{yy} + \phi_{zz}$ where the $\phi = \phi(x,y,z)$ is the potential.

Linear dispersion relation A mathematical relation between the wavenumber k and the frequency ω of waves ($\omega = \omega(k)$).

Linear dispersive waves Waves with the given dispersion relation between the wavenumber and the frequency.

Linear Schrödinger equation This can be written in the form $i\,a_t + \frac{1}{2}\,\omega''(k)\frac{\partial^2 a}{\partial x^2} = 0$, where $a = a(x,t)$ is the amplitude and $\omega = \omega(k)$.

Linear wave equation in Cartesian coordinates In three dimensions this can be written in the form $u_{tt} = c^2\,\nabla^2 u$, where c is a constant and $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial}{\partial z^2}$ is the three-dimensional Laplacian.

Linear wave equation in cylindrical polar coordinates This can be written in the form $u_{tt} = u_{rr} + \frac{1}{r}\,u_r + \frac{1}{r^2}\,u_{\theta\theta}$.

Navier–Stokes equations This is a nonlinear partial differential equation for an incompressible and viscous fluid flow governed by the velocity field $\mathbf{u} = (u,v,w)$ and pressure $P(\mathbf{x},t)$ under the action of external force $\mathbf{F}$. More precisely, $\frac{\partial\mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} = -\frac{1}{\rho}\,\nabla P + \nu^2\mathbf{u} + \mathbf{F}$, where ρ is the density and $\nu = (\mu/\rho)$ is the kinematic viscosity.

Nonlinear dispersion relation A mathematical relation between the wavenumber k, frequency ω and the amplitude a that is $D(\omega,k,a) = 0$.

Nonlinear dispersive waves Waves with the given dispersion relation between the wavenumber, frequency and the amplitude.

Non-linear Schrödinger (NLS) equation This is a nonlinear partial differential equation for the nonlinear modulation of a monochromatic wave. The amplitude, $a(x,t)$ of the modulation satisfies the equation $i\left(\frac{\partial a}{\partial t} + \omega_0'\,\frac{\partial a}{\partial x}\right) + \frac{1}{2}\,\omega_0''\,\frac{\partial^2 a}{\partial x^2} + \gamma|a|^2 a = 0$, where $\omega_0 = \omega_0(k)$, and γ is a constant.

Ocean waves Waves observed on the surface or inside the ocean.

Phase velocity The velocity defined by the ratio of the frequency ω and the wavenumber k, ($c_p = \frac{\omega}{k}$).

Resonant or critical phenomenon Waves with unbounded amplitude.

Sinusoidal (or exponential) wave A wave is of the form $u(x,t) = a\,\mathrm{Re}\exp[i(kx - \omega t)] = a\cos(kx - \omega t)$, where a is amplitude, k is the wavenumber ($k = \frac{2\pi}{\lambda}$), λ is the wavelength and ω is the frequency.

Solitary waves (or soliton) Waves describing a single hump of given height travel in a medium without change of shape.

Stokes expansion This is an expansion of the frequency in terms of the wavenumber k and the amplitude a, that is, $\omega(k) = \omega_0(k) + \omega_2(k)a^2 + \ldots$.

Stokes wave Water waves with dispersion relation involving the wavenumber, frequency and amplitude.

Surface-capillary gravity waves Waves under the joint action of the gravitational field and surface tension.

Surface gravity waves Water waves under the action of the gravitational field.

Variational principle For three-dimensional water waves, it is of the form $\delta I = \delta \iint_D L\,\mathrm{d}\mathbf{x}\,\mathrm{d}t = 0$, where L is called the Lagrangian.

Velocity potential A single valued function $\phi = \phi(\mathbf{x},t)$ defined by $\mathbf{u} = \nabla\phi$.

Water waves Waves observed on the surface or inside of a body of water.

Waves on a running stream Waves observed on the surface or inside of a body of fluid which is moving with a given velocity.

Whitham averaged variational principle This can be formulated in the form $\delta \iint \mathcal{L}\,\mathrm{d}\mathbf{x}\,\mathrm{d}t = 0$, where $\mathcal{L}$ is called the Whitham average Lagrangian over the phase of the integral of the Lagrangian L defined by $\mathcal{L}(\omega,\mathbf{k},a,\mathbf{x},t) = \frac{1}{2\pi}\int_0^{2\pi} L\,\mathrm{d}\theta$, where L is the Lagrangian.

Whitham's conservation equations These are first order nonlinear partial differential equations in the form $\frac{\partial k}{\partial t} + \frac{\partial\omega}{\partial x} = 0$, $\frac{\partial}{\partial t}\left\{f(k)A^2\right\} + \frac{\partial}{\partial x}\left\{f(k)C(k)A^2\right\} = 0$, where $k = k(x,t)$ is the density of waves, $\omega = \omega(x,t)$ is the flux of waves, $A = A(x,t)$ is the amplitude and $f(k)$ is an arbitrary function.

Whitham's equation This first order nonlinear partial differential equation represents the conservation of waves. Mathematically, $(\partial k/\partial t) + (\partial\omega/\partial x) = 0$ where

$k = k(x, t)$ is the density of waves and $\omega = \omega(x, t)$ is the flux of waves.

Whitham's equation for slowly varying wavetrain This is written in the form $\frac{\partial}{\partial t}\mathcal{L}_\omega - \frac{\partial}{\partial x_i}\mathcal{L}_{k_i} = 0$, where $\mathcal{L}$ is the Whitham averaged Lagrangian.

Whitham's nonlinear nonlocal equations It is in the form $u_t + duu_x + \int_{-\infty}^{\infty} K(x-s)\, u_s(s, t)\, \mathrm{d}s = 0$, where $K(x) = \mathcal{F}^{-1}\{c(k) = \omega/k\}$ and $\mathcal{F}^{-1}$ is the inverse Fourier transformation.

Definition of the Subject

A *wave* is usually defined as the propagation of a disturbance in a medium.

The simplest example is the exponential or sinusoidal wave which has the form

$$\begin{aligned} u(x, t) &= a\,\mathrm{Re}\exp\left[i(kx - \omega t)\right] \\ &= a\cos(kx - \omega t)\,, \end{aligned} \tag{1}$$

where a is called the *amplitude*, *Re* stands for the real part, $k\,(= \frac{2\pi}{\lambda})$ is called the *wavenumber* and λ is called the *wavelength* of the wave, and ω is called the *frequency* and it is a definite function of the wavenumber k and hence, $\omega = \omega(k)$ is determined by the particular equation of the problem. The quantity $\theta = kx - \omega t$ is called the *phase* of the wave so that a wave of a constant phase propagate with $kx - \omega t =$ constant.

The mathematical relation

$$\begin{aligned} &\omega = \omega(k) \quad \text{or} \\ &D(\omega, k) = 0\,, \end{aligned} \tag{2}$$

is called the *dispersion relation*.

The *phase* (or *wave*) *velocity* is defined by

$$c(k) = \frac{\omega(k)}{k}\,. \tag{3}$$

This shows that the phase velocity, in general, depends on the wavenumber k (or wavelength λ) so that waves with different wavelength propagate with different phase velocity. The waves are called *dispersive* if the phase velocity is not a constant, but depends on the wavenumber k. On the other hand, waves are called nondispersive if $c(k)$ is constant, that is, independent of k.

In general, the dispersion relation (2) can be written in the complex form

$$\begin{aligned} \omega &= \omega(k) \\ &= \sigma(k) + i\,\nu(k)\,, \end{aligned} \tag{4}$$

where $\nu\,(k) < 0$, $\nu\,(k) = 0$, or $\nu\,(k) > 0$.

In this case, the sinusoidal wave takes the form

$$u(x, t) = a\,\exp\left[\nu(k)t\right]\exp\left[i(kx - \sigma t)\right]\,. \tag{5}$$

If $\nu(k) < 0$, the amplitude of the wave decays to zero as $t \to \infty$ and the waves are called *dissipative*. When $\nu(k) = 0$, waves are dispersive. On the other hand, if $\nu(k) > 0$ for some or all k, the solution grows exponentially as $t \to \infty$. This case corresponds to instability.

The group velocity of the wave (1) is defined by

$$C(k) = \frac{\mathrm{d}\omega}{\mathrm{d}k}\,. \tag{6}$$

So, in general, $c(k) \neq C(k)$. It is convenient to modify the definition slightly. Waves are called dispersive if $\omega'(k)$ is not a constant, that is, $\omega''(k) \neq 0$.

In higher dimensions, all the above ideas can be generalized without any difficulty. The sinusoidal waves in three space dimensions are defined by

$$u(\mathbf{x}, \mathbf{t}) = a\ \mathrm{Re}\ \exp\left[i(\boldsymbol{\kappa}\cdot\mathbf{x} - \omega t)\right]\,, \tag{7}$$

where a is the amplitude, $\mathbf{x} = (x, y, z)$ is the displacement vector, $\boldsymbol{\kappa} = (k, l, m)$ is the wavenumber vector and ω is the frequency which is related to $\boldsymbol{\kappa}$ by the dispersion relation

$$\omega = \omega(\boldsymbol{\kappa}) \quad \text{or} \quad D(\omega, \boldsymbol{\kappa}) = 0\,. \tag{8}$$

This function D is also determined by the particular equation of the problem. In this case, $\theta = \boldsymbol{\kappa}\cdot\mathbf{x} - \omega t$ is the phase function.

Similarly, the phase velocity of the waves are defined as follows:

$$c(k) = \frac{\omega(\boldsymbol{\kappa})}{\boldsymbol{\kappa}}\,\widehat{\boldsymbol{\kappa}}\,, \tag{9}$$

where $\widehat{\boldsymbol{\kappa}} = (\widehat{k}, \widehat{l}, \widehat{m})$ is the unit vector in the $\boldsymbol{\kappa}$ direction and the group velocity is defined by

$$C(\boldsymbol{\kappa}) = \nabla_{\boldsymbol{\kappa}}\omega(\boldsymbol{\kappa}) = \widehat{\mathbf{k}}\,\frac{\partial\omega}{\partial k} + \widehat{\mathbf{l}}\,\frac{\partial\omega}{\partial l} + \widehat{\mathbf{m}}\,\frac{\partial\omega}{\partial m}\,. \tag{10}$$

If the dispersion relation (2) depends on the amplitude a so that the dispersion relation becomes

$$\omega = \omega(k, a) \quad \text{or} \quad D(\omega, k, a) = 0\,. \tag{11}$$

In general, the *linear plane waves* are recognized by the existence of periodic wavetrains in the form

$$u(x, t) = f(\theta) = f(kx - \omega t)\,, \tag{12}$$

where f is a periodic function of the phase θ.

For linear problems, solutions more general than (1) can be obtained by the principle of superposition to form Fourier integrals as

$$u(x,t) = \int_{-\infty}^{\infty} U(k) \exp\left[i\{kx - \omega(k)t\}\right] \mathrm{d}k \,, \tag{13}$$

where the arbitrary function $U(k)$ may be chosen to fit arbitrary initial or boundary conditions, provided the data are reasonable enough to admit Fourier transform $U(k) = \mathcal{F}\{u(x,0)\}$, and $\omega = \omega(k)$ is the dispersion relation (2) appropriate to the problem.

On the other hand, the *nonlinear dispersive* waves are also recognized by the existence of periodic wavetrains (12) and the solution must include the amplitude parameter a and it also requires a nonlinear dispersion relation of the form (11).

In shallow water of constant depth h, the free surface elevation $\eta(x,t)$ satisfies the Korteweg and de Vries (KdV) equation

$$\eta_t + c\left(1 + \frac{3}{2h}\eta\right)\eta_x + \left(\frac{ch^2}{6}\right)\eta_{xxx} = 0 \,, \tag{14}$$

where $c = \sqrt{gh}$ is the shallow water velocity, g is the acceleration due to gravity, and the total depth $H = h + \eta(x,t)$. The first two terms $(\eta_t + c\,\eta_x)$ describe the wave evolution at speed c, the third term with the coefficient $(3c/2h)$ represents the nonlinear wave steepening and the last term with the coefficient $(ch^2/6)$ describes linear dispersion.

The KdV equation admits an exact solution in the form

$$\eta(x,t) = a\,\mathrm{sech}^2\left[\left(\frac{3a}{4h^3}\right)^{\frac{1}{2}} X\right] \,, \tag{15}$$

where $X = x - Ut$, and U is the wave velocity given by

$$U = C\left(1 + \frac{a}{2h}\right) \,. \tag{16}$$

The solution (15) is called the *solitary wave* (or *soliton*) describing a single hump of height a above the undisturbed depth h and tending rapidly to zero away from $X = 0$. The solitary wave propagates to the right with velocity $U(> c)$ which is directly proportional to the amplitude a and has width $b^{-1} = (3a/4h^3)^{-\frac{1}{2}}$, that is, b^{-1} is inversely proportional to the square root of the amplitude a. Another significant feature of the solitary wave is that it travels in the medium *without* change of shape.

In general, the solution for $\eta(x,t)$ can be expressed in terms of the Jacobian elliptic function $cn(X,m)$

$$\eta(X) = a\,cn^2\left[\left(\frac{3b}{4h^3}\right)^{\frac{1}{2}} X, m\right] \,, \tag{17}$$

where $m = (a/b)^{\frac{1}{2}}$ is the modulus of the *cn* function. Consequently, the solution (17) is called the *cnoidal wave*.

The *nonlinear Schrödinger (NLS) equation* can be written in the standard form

$$i\,\psi_t + \psi_{xx} + \gamma|\psi|^2\psi = 0 \,, \qquad -\infty < x < \infty \,, \quad t > 0 \,, \tag{18}$$

and γ is a constant.

With $X = x - Ut$, we seek the solution in the form

$$\psi = \exp\left[i(mX - nt)\right] f(X) \,, \tag{19}$$

where $f(X)$ can be expressed in terms of the Jacobian elliptic function in the form

$$f(X) = (\alpha_1/\alpha_2)^{\frac{1}{2}}\, sn(\sigma X, \kappa) \,, \tag{20}$$

where α_1, α_2, $\sigma = (\alpha_2\beta_2/\beta_1\alpha_2)$ and $\kappa = (\alpha_1\beta_2)/(\beta_1\alpha_2)$ are constants.

The limiting case of the solitary wave is possible and has the form

$$f(X) = \left(\frac{2\alpha}{\gamma}\right)^{\frac{1}{2}} \mathrm{sech}\left[\sqrt{\alpha}(x - Ut)\right] \,, \tag{21}$$

where α and γ are positive constants. This solution represents a solitary wave solution which propagates without change of shape with constant velocity. However, unlike the KdV solitary waves, the amplitude and the velocity are independent parameters. It is important to note that the solution (21) is possible only in the unstable case $\gamma > 0$. This suggests that the end result of an unstable wavetrain subject to small modulation is a series of solitary waves.

Introduction

Water waves are the most common observable phenomena in Nature. The subject of water waves is most fascinating and highly mathematical, and varied of all areas in the study of wave motions in the physical world. The mathematical as well as physical problems deal with water waves and their breaking on beaches, with flood waves in rivers, with ocean waves from storms, with ship waves on water, with free oscillations of enclosed waters such as lakes and harbors. The study of water waves and their various ramifications remain central to fluid dynamics in general, and to the dynamics of oceans in particular. The mathematical theory of water waves is quite interesting in its own merit and intrinsically beautiful. It has provided the solid background and impetus for the development of the theory of nonlinear dispersive waves. Indeed, most of the fundamental ideas and results for nonlinear dispersive waves and solitons originated in the investigation of water waves.

Historically, the problems of water waves in oceans originated from the classic work of Leonhard Euler (1707–1783), A.G. Cauchy (1789–1857), S.D. Poisson (1781–1840), Joseph Boussinesq (1842–1929), George Airy (1801–1892), Lord Kelvin (1824–1907), Lord Rayleigh (1842–1919), George Stokes (1819–1903), Scott Russell (1808–1882) and many others. Indeed, Euler formulated the boundary value problem to understand water wave phenomena and provided successful mathematical and physical description of inviscid and incompressible fluid motion in general. Based on Newton's second law of motion, Euler first formulated his celebration equation of motion of an inviscid fluid about 250 years ago. Euler's equation is still considered the basis of all inviscid fluid flows. In 1821, Claude Navier (1785–1836) included the effect of viscosity to the Euler equation, and first developed the equations of motion of viscous fluid. In 1845, Sir George Stokes provided a sound mathematical foundation of viscous fluid flows and rederived the equations of motion for an incompressible viscous fluid that is universally known as the *Navier–Stokes equations* in the form

$$\frac{\mathrm{D}\,\mathbf{u}}{\mathrm{D}\,t} = \frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} = -\frac{1}{\rho}\nabla P + \mathbf{F} + \nu\nabla^2\mathbf{u}\,, \quad (22)$$

where $(\mathrm{D}\,/\,\mathrm{D}\,t)$ is the *total* (or *convective*) derivative given by

$$\frac{\mathrm{D}}{\mathrm{D}\,t} = \frac{\partial}{\partial t} + (\mathbf{u}\cdot\nabla)\,, \quad (23)$$

$\mathbf{u}(\mathbf{x}, t) = (u, v, w)$ is the fluid velocity, $P(\mathbf{x}, t)$ is the pressure field at a point $\mathbf{x} = (x, y, z)$, t is time, ρ is a constant density, $\mathbf{F}(\mathbf{x}, t)$ is a general body force per unit mass, $\nu = (\mu/\rho)$ is known as kinematic viscosity, μ is called the dynamic viscosity, $\nabla^2 = \nabla\cdot\nabla$ is the *Laplace operator* and $\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right)$ is the familiar gradient operator.

In particular, when $\nu = 0$ ($\mu = 0$), the Navier–Stokes Eq. (22) reduces to the celebrated *Euler equation* of motion for an inviscid fluid

$$\frac{\mathrm{D}\,\mathbf{u}}{\mathrm{D}\,t} = \frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} = -\frac{1}{\rho}\nabla P + \mathbf{F}\,. \quad (24)$$

Both the Euler Eq. (24) and the Navier–Stokes Eq. (22) form a closed set when the *equation of mass conservation* (or the *continuity equation*)

$$div\,\mathbf{u} = \nabla\cdot\mathbf{u} = 0 \quad (25)$$

is added to (22) or (24) so that there are four equations for four unknown quantities u, v, w, and P. The study of these equations is based on the *continuum hypothesis* which requires that $\mathbf{u}$, p and ρ are continuous function of $\mathbf{x} = (x, y, z)$ and t.

Remarkably, this 150-year old system of the Navier–Stokes Eqs. (22) and (25) provided the fundamental basis of modern fluid mechanics. However, there are certain major difficulties associated with the Navier–Stokes equations. First, there are no general results for the Navier–Stokes equations on existence of solutions, uniqueness, regularity, and continuous dependence on the initial conditions. Second, another difficulty arises from the strong nonlinear convective term, $(\mathbf{u}\cdot\nabla)\mathbf{u}$ in Eq. (22).

The Euler Equation of Motion in Rectangular Cartesian and Cylindrical Polar Coordinates

With $\mathbf{F} = (0, 0, -g)$, where g is the acceleration due to gravity and constant density ρ, the three components of the Euler's equation of motion in Cartesian coordinates and the continuity equation are

$$\frac{\mathrm{D}\,u}{\mathrm{D}\,t} = -\frac{1}{\rho}\frac{\partial P}{\partial x}\,, \quad \frac{\mathrm{D}\,v}{\mathrm{D}\,t} = -\frac{1}{\rho}\frac{\partial P}{\partial y}\,,$$
$$\frac{\mathrm{D}\,w}{\mathrm{D}\,t} = -\frac{1}{\rho}\frac{\partial P}{\partial z} - g\,, \quad (26)$$

where the total derivative is given by

$$\frac{\mathrm{D}}{\mathrm{D}\,t} = \frac{\partial}{\partial t} + u\frac{\partial}{\partial x} + v\frac{\partial}{\partial y} + w\frac{\partial}{\partial z}\,, \quad (27)$$

and

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0\,. \quad (28)$$

The basic assumption in continuum mechanics is that the motion of the fluid can be described mathematically as a topological deformation that depends continuously on the time t. Consequently, we assume the fluid under consideration to have a boundary surface S, fixed or moving, which separated it from other media. We consider the case of a body of water with air above it so that S is the interface between them. We represent that surface S by an equation $S(x, y, z, t) = 0$. The kinematic condition is derived from the fact that the normal fluid velocity of the surface $(-S_t/|\nabla S|)$ is equal to the normal velocity $(\mathbf{u}\cdot\mathbf{n} = \mathbf{u}\cdot\nabla S/|\nabla S|)$, that is,

$$\frac{\mathrm{D}\,S}{\mathrm{D}\,t} = S_t + (\mathbf{u}\cdot\nabla)S = 0\,. \quad (29)$$

This means that any fluid particle originally on the boundary surface, S will remain on it.

It is often convenient to represent the free surface by the equation $z = h(x, y, t)$ so that the equation for the boundary surface S is

$$S = z - h(x, y, t) = 0\,, \tag{30}$$

where z is independent of other variables.

Thus, the kinematic free surface condition follows from (29) in the form

$$w - (h_t + u\,h_x + v\,h_y) = 0 \quad \text{on } z = h(x, y, t)\,, \quad t > 0\,. \tag{31}$$

Since the gravitational force is only body force which acts in the negative z-direction, the continuity Eq. (28) and the Euler equation in the form

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} = -\frac{1}{\rho}\nabla P - g\,\widehat{\mathbf{m}}\,, \tag{32}$$

where $\widehat{\mathbf{m}}$ is the unit vector in the positive z-direction, represent the fundamental equations for water wave motion.

In general, the water wave motion is unsteady, and irrotational which physically means that the individual fluid particle do not rotate. Mathematically, this implies that vorticity $\boldsymbol{\omega} = curl\ \mathbf{u} = \mathbf{0}$. So, there exists a single-valued velocity potential ϕ so that $\mathbf{u} = \nabla\phi$, where $\phi = \phi(\mathbf{x}, t)$. Consequently, the continuity equation reduces to the Laplace equation

$$\nabla^2\phi = \frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} + \frac{\partial^2\phi}{\partial z^2} = 0\,. \tag{33}$$

So, ϕ is a harmonic function. This is, indeed, a great advantage because the velocity field $\mathbf{u}$ can be obtained from a single potential function ϕ which satisfies a linear partial differential equation. This equation with prescribed boundary conditions can readily be solved in many simple cases without difficulty.

In view of a vector identity

$$(\mathbf{u}\cdot\nabla)\mathbf{u} = \frac{1}{2}\nabla u^2 - \mathbf{u}\times\boldsymbol{\omega} \tag{34}$$

combined with $\boldsymbol{\omega} = curl\ \mathbf{u} = \mathbf{0}$ and $\mathbf{u} = \nabla\phi$, the Euler Eq. (32) may be written as

$$\nabla\left[\phi_t + \frac{1}{2}(\nabla\phi)^2 + \frac{P}{\rho} + gz\right] = 0\,. \tag{35}$$

This can be integrated with respect to the space variables to give the equation for pressure P at every point of the fluid

$$\phi_t + \frac{1}{2}(\nabla\phi)^2 + \frac{P}{\rho} + gz = C(t)\,, \quad t \geq 0\,, \tag{36}$$

where $C(t)$ is an arbitrary function of time only ($\nabla C = 0$). Since the pressure gradient affects the flow, a function of t alone added to the pressure field P has no effect on the motion. So, without loss of generality, we can set $C(t) = 0$ in Eq. (36). Thus, the pressure Eq. (36) becomes

$$\phi_t + \frac{1}{2}(\nabla\phi)^2 + \frac{P}{\rho} + gz = 0\,. \tag{37}$$

This is the so-called the *Bernoulli's equation* which determines the pressure in terms of the velocity potential ϕ. Thus, Eqs. (33) and (37) are used to determine the potential ϕ (hence, three velocity components u, v, and w) and the pressure field P.

We consider a body of inviscid, incompressible fluid occupying the region $b(x, y) \leq z \leq h(x, y, t) = h_0 + a\eta(x, y, t)$, where $z = b(x, y)$ is the bottom boundary surface and $z = h_0$ is the undisturbed (initial) typical constant depth, a is the typical amplitude and $\eta(x, y, t)$ is nondimensional the free surface elevation that tends to zero as $t \to 0$. We suppose that the bottom boundary $z = b(x, y)$ is a rigid solid surface.

Since the upper boundary is the surface exposed to a constant atmospheric pressure P_a, we have $P = P_a$ on this surface, S. Thus, Eq. (35) assumes the form

$$\phi_t + \frac{1}{2}(\nabla\phi)^2 + \frac{1}{\rho}P_a + gz = C(t) \quad \text{on } S\,, \quad t \geq 0\,. \tag{38}$$

Absorbing $\frac{1}{\rho}P_a$ and $C(t)$ into ϕ_t, this equation may be rewritten as

$$\phi_t + \frac{1}{2}(\nabla\phi)^2 + gz = 0\,, \qquad \text{on } S\,, \quad t \geq 0\,. \tag{39}$$

Since S is a upper boundary surface of the fluid, it contains the same fluid particles for all time t, that is, S is a material surface. Hence, it follows from (39) that

$$\frac{\mathrm{D}}{\mathrm{D}\,t}\left[\phi_t + \frac{1}{2}(\nabla\phi)^2 + gz\right] = 0\,, \quad \text{on } S\,, \quad t \geq 0\,. \tag{40}$$

Or, equivalently,

$$\begin{aligned}\left[\frac{\partial}{\partial t} + (\nabla\phi\cdot\nabla)\right]\left[\frac{\partial\phi}{\partial t} + \frac{1}{2}(\nabla\phi)^2 + gz\right] &= \phi_{tt} + 2\nabla\phi\cdot\nabla(\phi_t) + \frac{1}{2}\nabla\phi\cdot\nabla(\nabla\phi)^2 + g\phi_z \\ &= 0\,, \qquad \text{on } S\,, \quad t \geq 0\,.\end{aligned} \tag{41}$$

We next include the effects of *surface tension force* per unit length which does support a pressure difference across

a curved surface so that $P - P_a = -(T/R)$, where $R^{-1} = (R_1^{-1} + R_2^{-1})$ is called the *Gaussian curvature* expressed as the sum of two principal radii of curvatures R_1^{-1} and R_2^{-1} given by

$$R_1^{-1} = \frac{\partial}{\partial x}\left[h_x(1 + h_x^2 + h_y^2)^{-\frac{1}{2}}\right] ,$$
$$R_2^{-1} = \frac{\partial}{\partial y}\left[h_y(1 + h_x^2 + h_y^2)^{-\frac{1}{2}}\right] . \tag{42ab}$$

Explicitly, R^{-1} can be written in terms of $h(x, y, t)$ and its partial derivatives as

$$R^{-1} = (R_1^{-1} + R_2^{-1}) = \frac{h_{xx}(1 + h_y^2) - 2\, h_x\, h_y\, h_{xy} + h_{yy}(1 + h_x^2)}{(1 + h_x^2 + h_y^2)^{3/2}} . \tag{43}$$

For small deviation of the free surface $z = h(x, y, t)$, its first partial derivatives h_x and h_y are small so that (43) takes the linearized form

$$R^{-1} \approx (h_{xx} + h_{yy}) . \tag{44}$$

Consequently, the linearized dynamic condition at the free surface $z = h$ becomes

$$\phi_t + gz - \frac{T}{\rho}(h_{xx} + h_{yy}) = 0 . \tag{45}$$

For an inviscid fluid, like the free surface condition (29), there is a bottom boundary condition which follows from the fact that

$$\frac{\mathrm{D}}{\mathrm{D}\,t}\left[z - b(x, y, t)\right] = 0 , \tag{46}$$

where $z = b(x, y, t)$ is the equation of the fixed solid bottom boundary surface. Thus, Eq. (46) assumes the form

$$w = b_t + (\mathbf{u} \cdot \nabla)b = b_t + u\, b_x + v\, b_y \quad \text{on } z = b(x, y, t) , \tag{47}$$

where $z = b(x, y, t)$ is given. However, there is a class of problems including sediment movement, where b is not known. For stationary bottom, b is independent of time so that (47) becomes

$$w = u\, b_x + v\, b_y \quad \text{on } z = b(x, y) . \tag{48}$$

For one-dimensional problem, $h = b(x)$ with $\mathbf{u} = (u, 0)$ so that (48) reduces to the simple bottom condition

$$w = u\, b'(x) \quad \text{on } z = b(x) . \tag{49}$$

It is convenient to introduce a typical amplitude parameter a by writing the free surface $z = h(x, y, t)$ in the form

$$h = h_0 + a\eta(x, y, t) , \tag{50}$$

where h_0 is the undisturbed depth of water. The pressure field can be rewritten as

$$P = P_a + g\rho(h_0 - z) + (g\rho h_0)p , \tag{51}$$

where P_a is the constant atmospheric pressure, p is the pressure variable which measures the deviation from the hydrostatic pressure, $g\rho(h_0 - z)$, and $g\rho h_0$ is the typical pressure scale based on the pressure at depth $h = h_0$.

We next introduce two fundamental parameters ε and δ as

$$\varepsilon = \frac{a}{h_0} \quad \text{and} \quad \delta = \frac{h_0^2}{\lambda^2} , \tag{52}$$

where λ is the typical wavelength of the surface gravity wave, ε and δ are called the *amplitude* and *long wavelength* parameters respectively.

In terms of typical depth scale h_0, λ is the typical wavelength, $c_0 = \sqrt{g\, h_0}$ is the typical horizontal velocity scale, (λ/c_0) is the typical time scale, it is also convenient to introduce non-dimensional flow variables denoted by asterisks

$$(x^*, y^*) = \frac{1}{\lambda}(x, y) , \quad (z^*, b^*) = \frac{1}{h_0}(z, b) ,$$
$$t^* = \left(\frac{c_0}{\lambda}\right) t , \tag{53}$$

$$(u^*, v^*) = \frac{1}{c_0}(u, v) , \quad w^* = \left(\frac{\lambda c_0}{h_0}\right) w \quad \text{and}$$
$$p^* = \frac{p}{g\rho h_0} . \tag{54}$$

In terms of these nondimensional flow variables and parameters, the Euler Eqs. (26) and the continuity Eq. (28) can be rewritten, dropping the asterisks, in the form

$$\frac{\mathrm{D}}{\mathrm{D}\,t}(u, v) = -\left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right) p, \quad \delta\, \frac{\mathrm{D}\,w}{\mathrm{D}\,t} = -\frac{\partial p}{\partial z} , \tag{55}$$

where

$$\frac{\mathrm{D}}{\mathrm{D}\,t} = \frac{\partial}{\partial t} + u\frac{\partial}{\partial x} + v\frac{\partial}{\partial y} + w\frac{\partial}{\partial z} ,$$

and

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0 . \tag{56}$$

The most general dynamic condition is $P = P_a - \frac{T}{R}$ on the free surface $z = h = h_0 + a\eta$. Without surface tension ($T = 0$), it becomes $P = P_a$ on $z = h_0 + a\eta$. Using the pressure field (51), this free surface dynamic condition reduces to the nondimensional form $p = \varepsilon\eta$ on $z = 1 + \varepsilon\eta$. Thus, the nondimensional kinematic condition (31) and the dynamic condition at the free surface are given by

$$w - \varepsilon(\eta_t + u\,\eta_x + v\,\eta_y) = 0\,, \qquad p = \varepsilon\eta \quad \text{on } z = 1 + \varepsilon\eta\,. \tag{57ab}$$

The nondimensional form of the bottom boundary condition (48) remains the unchanged form.

Consistent with the governing equations and free surface boundary conditions, we introduce a set of scaled flow variables

$$(u, v, w, p) \to \varepsilon(u, v, w, p) \tag{58}$$

so that the Euler Eqs. (55) and the continuity Eq. (56) reduce to the form

$$\frac{\mathrm{D}}{\mathrm{D}\,t}(u, v) = -\left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)p, \quad \delta\,\frac{\mathrm{D}\,w}{\mathrm{D}\,t} = -\frac{\partial p}{\partial z}\,, \tag{59}$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0\,, \tag{60}$$

where

$$\frac{\mathrm{D}}{\mathrm{D}\,t} = \frac{\partial}{\partial t} + \varepsilon\left(u\,\frac{\partial}{\partial x} + v\,\frac{\partial}{\partial y} + w\,\frac{\partial}{\partial y}\right)\,. \tag{61}$$

The free surface boundary conditions (57ab) remain the same. The horizontal bottom ($b = 0$) boundary condition is

$$w = 0\,, \quad \text{on } z = 0\,, \tag{62}$$

In cylindrical polar coordinates (r, θ, z), the Euler equations and the continuity equation are given by

$$\frac{\mathrm{D}\,u}{\mathrm{D}\,t} - \frac{v^2}{r} = -\frac{1}{\rho}\,P_r, \quad \frac{\mathrm{D}\,v}{\mathrm{D}\,t} + \frac{uv}{r} = -\frac{1}{\rho}\,\frac{1}{r}\,P_\theta\,, \quad \frac{\mathrm{D}\,w}{\mathrm{D}\,t} = -\frac{1}{\rho}\,P_z - g\,, \tag{63}$$

where

$$\frac{\mathrm{D}}{\mathrm{D}\,t} = \frac{\partial}{\partial t} + u\,\frac{\partial}{\partial r} + \frac{v}{r}\,\frac{\partial}{\partial\theta} + w\,\frac{\partial}{\partial z}\,, \tag{64}$$

and

$$\frac{1}{r}\,\frac{\partial}{\partial r}(ru) + \frac{1}{r}\,\frac{\partial v}{\partial\theta} + \frac{\partial w}{\partial z} = 0\,. \tag{65}$$

In terms of the above nondimensional flow variables except for x and y that are replaced by $r^* = \frac{1}{\lambda}\,r$, the above Eqs. (63)–(65) assume the following nondimensional form, dropping the asterisks,

$$\frac{\mathrm{D}\,u}{\mathrm{D}\,t} - \frac{v^2}{r} = -\frac{\partial p}{\partial r}\,: \quad \frac{\mathrm{D}\,v}{\mathrm{D}\,t} + \frac{uv}{r} = -\frac{1}{r}\,\frac{\partial p}{\partial\theta}\,, \quad \delta\,\frac{\mathrm{D}\,w}{\mathrm{D}\,t} = -\frac{\partial p}{\partial z}\,, \tag{66}$$

where

$$\frac{\mathrm{D}}{\mathrm{D}\,t} = \frac{\partial}{\partial t} + u\,\frac{\partial}{\partial r} + \frac{v}{r}\,\frac{\partial}{\partial\theta} + w\,\frac{\partial}{\partial z}\,, \tag{67}$$

and

$$\frac{1}{r}\,\frac{\partial}{\partial r}(ru) + \frac{1}{r}\,\frac{\partial v}{\partial\theta} + \frac{\partial w}{\partial z} = 0\,. \tag{68}$$

Basic Equations of Water Waves with Effects of Surface Tension

We consider an irrotational unsteady motion of an inviscid and incompressible water occupying the region $b(x, y) \le z \le h_0 + a\,\eta(x, y, t)$ in a constant gravitational field g which is in the negative z-direction. The equation of uneven bottom boundary is $z = b(x, y)$. Including the effects of surface tension T, the basic equations, two free surface boundary conditions and the bottom boundary condition for the velocity potential $\phi = \phi(x, y, z, t)$ and the free surface elevation $\eta = \eta(x, y, t)$ are

$$\nabla^2\phi = \phi_{xx} + \phi_{yy} + \phi_{zz} = 0\,, \qquad b(x, y) \le z \le h_0 + a\,\eta\,, \quad t > 0\,, \tag{69}$$

$$\eta_t + (\phi_x\eta_x + \phi_y\eta_y) - \phi_z = 0\,, \qquad \text{on } z = h_0 + a\,\eta\,, \quad t > 0\,, \tag{70}$$

$$\phi_t + \frac{1}{2}(\nabla\phi)^2 + g\,\eta - \frac{T}{\rho}(R_1^{-1} + R_2^{-1}) = 0\,, \qquad \text{on } z = h_0 + a\,\eta\,, \quad t > 0\,, \tag{71}$$

$$\phi_z - (\phi_x b_x + \phi_y b_y) = 0\,, \quad \text{on } z = b(x, y)\,, \tag{72}$$

where R_1^{-1} and R_2^{-1} are given by (42ab), h_0 is the typical depth of water and a is the typical amplitude of the surface gravity wave.

In water of infinite depth with the origin at the free surface, the bottom boundary condition (72) is replaced by $(\nabla\phi) \to 0$ as $z \to -\infty$.

Because of the presence of nonlinear terms in the free surface boundary conditions (70)–(71), the determination of ϕ and η in the general case is a difficult task.

Similarly, we can write the basic equations for the water wave problems in cylindrical polar coordinates.

In terms of the nondimensional flow variables and the parameters ε and δ stated above, the basic water wave Eqs. (69)–(72) without surface tension can be written in the nondimensional form

$$\delta(\phi_{xx} + \phi_{yy}) + \phi_{zz} = 0\,, \quad \text{on } b \le z \le 1 + \varepsilon\eta\,, \quad t > 0\,, \tag{73}$$

$$\delta\left[\eta_t + \varepsilon(\phi_x\eta_x + \phi_y\eta_y)\right] - \phi_z = 0\,, \quad \text{on } z = 1 + \varepsilon\eta\,, \tag{74}$$

$$\phi_t + \eta + \frac{\varepsilon}{2}(\phi_x^2 + \phi_y^2) + \frac{\varepsilon}{2\delta}\phi_z^2 = 0\,, \quad \text{on } z = 1 + \varepsilon\eta\,, \tag{75}$$

$$\phi_z - \delta(\phi_x b_x + \phi_y b_y) = 0\,, \quad \text{on } z = b(x, y)\,, \tag{76}$$

where $(\varepsilon/\delta) = (a\lambda^2/h_0^3)$ is another fundamental parameter in water-wave theory.

Luke [47] first explicitly formulated a variational principle for two-dimensional water waves and proved that the basic Laplace equation, free surface and bottom boundary conditions can be derived from the Hamilton principle. We formulate the variational principle for three-dimensional water waves in the form

$$\delta I = \delta \iint_D L \, \mathrm{d}\mathbf{x} \, \mathrm{d}t = 0\,, \tag{77}$$

where the Lagrangian L is assumed to be equal to the pressure so that

$$L = -\rho \int_{-h(x,y)}^{\eta(\mathbf{x},t)} \left[\phi_t + \frac{1}{2}(\nabla\phi)^2 + gz\right] \mathrm{d}z\,, \tag{78}$$

where D is an arbitrary region in the $(\mathbf{x}, t)$ space, and $\phi(\mathbf{x}, z, t)$ is the velocity potential of an unbounded fluid lying between the rigid bottom $z = -h(x, y)$ and the free boundary surface $z = \eta(x, y, t)$. Using the standard procedure in the calculus of variations (see Debnath [19]), the following nonlinear system of equations for the classical water waves can be derived:

$$\nabla^2\phi = 0\,, \quad -h < z < \eta\,, \quad -\infty < (x, y) < \infty\,, \tag{79}$$

$$\eta_t + (\phi_x\eta_x + \phi_y\eta_y) - \phi_z = 0\,, \quad \text{on } z = \eta\,, \tag{80}$$

$$\phi_t + \frac{1}{2}(\nabla\phi)^2 + gz = 0\,, \quad \text{on } z = \eta\,, \tag{81}$$

$$\phi_z + \phi_x h_x + \phi_y h_y = 0\,, \quad \text{on } z = -h\,. \tag{82}$$

These results are in perfect agreement with those of Luke for the two-dimensional waves on water of arbitrary but uniform depth h. In his pioneering work, Whitham [58,59] first developed a general approach to linear and nonlinear dispersive waves using a Lagrangian. It is now well known that most of the general ideas about dispersive waves have originated from the classical problems of water waves.

Making reference to Debnath [19], we first state the solution of the linearized two-dimensional problem of the classical water waves on water of uniform depth h governed by the following equation, free surface and boundary conditions

$$\phi_{xx} + \phi_{zz} = 0\,, \quad -h \le z \le 0\,, \quad t > 0\,, \tag{83}$$

$$\eta_t = \phi_z\,, \quad \phi_t + g\eta = 0\,, \quad \text{on } z = 0\,,\ t > 0\,, \tag{84ab}$$

$$\phi_z = 0\,, \quad \text{on } z = -h\,, \quad t \ge 0\,. \tag{85}$$

The solutions for $\phi(x, z, t)$ and $\eta(x, t)$ representing a sinusoidal wave propagating in the x-direction are given by

$$\phi(x, z, t) = \mathrm{Re}\ a\left(\frac{ig}{\omega}\right)\frac{\cosh k(z + h)}{\cosh kh}\exp\left[i(\omega t - kx)\right]\,, \tag{86}$$

$$\eta(x, t) = \mathrm{Re}\ a\ \exp\left[i(\omega t - kx)\right]\,, \tag{87}$$

where $a = (C\omega/ig)\cosh kh = \max|\eta|$ is the amplitude and C is an arbitrary constant.

Using (84ab), we obtain the celebrated *dispersion relation* between the frequency ω and the wavenumber k in the form:

$$\omega^2 = gk\tanh kh\,. \tag{88}$$

Physically, this relation describes the interaction between inertial and gravitational forces. This can also be rewritten in terms of the wave (or phase) velocity, $c(k)$ as

$$c(k) = \frac{\omega}{k} = (gk^{-1}\tanh kh)^{\frac{1}{2}} = \left[\left(\frac{g\lambda}{2\pi}\right)\tanh\left(\frac{2\pi h}{\lambda}\right)\right]^{\frac{1}{2}}\,. \tag{89}$$

This formula shows that the phase velocity $c(k)$ depends on the gravity g and depth h as well as the wavenumber k or wavelength $\lambda = (2\pi/k)$. Thus, water waves of different wavelengths travel with different wave (or phase) velocities. Such waves are called *dispersive*, as time passes, these waves disperse (or spread out) into various group of waves such that each group of waves such that each group would consist of waves having approximately the

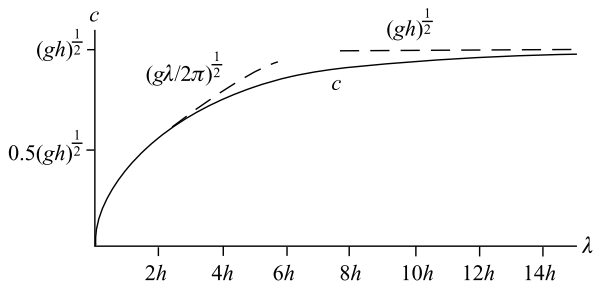

Water Waves and the Korteweg–de Vries Equation, Figure 1
The phase velocity $c(\lambda)$ against the wavelength λ (from [46], Figure 52, page 217)

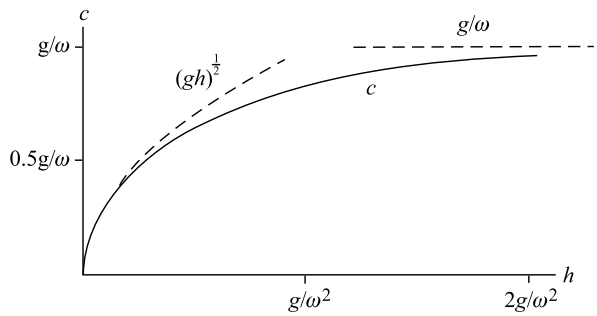

Water Waves and the Korteweg–de Vries Equation, Figure 2
The phase velocity c for waves of frequency ω on water of depth h (from [46], Figure 53, page 218)

same wavelength. Figure 1 showing a plot of the phase velocity c given by (89) against the wavelength λ reveals a transition between the deep water limit $c \sim (g\lambda/2\pi)^{\frac{1}{2}}$ (parabolic form) when $\lambda < (3.5)h$ and the shallow water limit $c \sim \sqrt{gh}$ for $\lambda > (14)h$.

The quantity $(\mathrm{d}\omega/\mathrm{d}k)$ represents the velocity of such a group in the direction of propagation and is called the *group velocity*, denoted by $C(k)$ so that

$$C(k) = \frac{\mathrm{d}\omega}{\mathrm{d}k} = \left(\frac{g}{2\omega}\right)(\tanh kh + kh\,\mathrm{sech}^2 kh)\,, \qquad (90)$$

which is, using (89)

$$= \tfrac{1}{2}\,c(k)\left[1 + 2kh\,\mathrm{cosech}(2kh)\right]\,. \qquad (91)$$

Evidently, the group velocity is, in general, different from the phase velocity.

Two limiting cases are of special interest: (i) Shallow water waves and (ii) Deep water waves.

For shallow water, the wavelength; $\lambda = (2\pi/k)$ is large compared with the depth h so that $kh \ll 1$, and hence, $\tanh kh \approx kh$ and $\sin 2kh \approx 2kh$. In such a case, (88)–(91) reduce to

$$\omega^2 = gk\,h^2\,, \quad c(k) = \sqrt{gh} = C(k)\,. \qquad (92)$$

Both phase and group velocities are independent of the wavenumber k. Thus, shallow water waves are nondispersive, and their phase velocity is equal to group velocity. Both vary as the square root of the depth h.

In the other limiting case dealing with deep water waves, the wavelength is very small compared with the depth so that $kh \gg 1$. In the limit as $kh \to \infty$, $\tanh kh \to 1$, $\left[\cosh k(z+h)/\cosh kh\right] \to \exp(kz)$, and the corresponding solutions for ϕ and η become

$$\phi = \mathrm{Re}\left(\frac{iag}{\omega}\right)\mathrm{e}^{kz}\exp\left[i(\omega t - kx)\right]$$
$$= \left(\frac{ag}{\omega}\right)\mathrm{e}^{kz}\sin(kx - \omega t)\,, \qquad (93)$$

$$\eta = \mathrm{Re}\;a\exp\left[i(\omega t - kx)\right] = a\,\cos(kx - \omega t)\,. \qquad (94)$$

Evidently, for deep water waves, the dispersion relation (88), the phase velocity (89) and the group velocity (90) become

$$\omega^2 = gk = \left(\frac{2\pi g}{\lambda}\right)\,, \quad c(k) = (g/k)^{\frac{1}{2}} = \left(\frac{g\lambda}{2\pi}\right)^{\frac{1}{2}}\,,$$
$$C(k) = \frac{1}{2}\,c(k)\,. \qquad (95abc)$$

Thus, deep water waves are dispersive and their phase velocity is proportional to the square root of their wavelengths. Also, the group velocity is equal to one-half of the phase velocity.

Another equivalent form of the phase velocity $c(k)$ follows from the dispersion relation (88) in the form

$$c(k) = \frac{\omega}{k} = \left(\frac{\omega^2}{\omega k}\right) = \frac{g}{\omega}\tanh\left(\frac{\omega h}{c}\right)\,. \qquad (96)$$

The phase velocity $c(k)$ for water waves of frequency ω against depth h is shown in Fig. 2. This figure shows a transition between the deep water limit $c \sim (g/\omega)$ when $(\omega h/c) \gg 2$ or $h \geq 2g/\omega^2$ and the shallow water limit $c \sim \sqrt{gh}$ when $(\omega h/c) \ll 1$ or $h \ll g/\omega^2$.

Similarly, an alternative form of the group velocity follows from the dispersion relation (88) as

$$C(k) = \frac{\mathrm{d}\omega}{\mathrm{d}k} = \frac{\mathrm{d}}{\mathrm{d}k}(ck) = c + k\,\frac{\mathrm{d}c}{\mathrm{d}k} = c - \lambda\,\frac{\mathrm{d}c}{\mathrm{d}\lambda}\,. \qquad (97)$$

Finally, we close the section by adding surface capillary-gravity waves on water of constant depth h with the free surface at $z = 0$. For such waves, the linearized free surface conditions (70)–(71) with the effect of surface tension are modified as follows:

$$\eta_t = \phi_z\,, \quad \phi_t + g\eta - \frac{T}{\rho}\,\eta_{xx} = 0 \quad \text{on } z = 0\,. \qquad (98)$$

These conditions can be combined to obtain

$$(\phi_{tt} + g\phi_z) - \frac{T}{\rho}\phi_{xxz} = 0 \quad \text{on } z = 0\,. \tag{99}$$

In this case, the solutions for the velocity potential $\phi(x, z, t)$ and the free surface elevation $\eta(x, t)$ are exactly the same as (86) and (87) on water of constant depth h or as (93)–(94) for deep water. Evidently, the dispersion relation for capillary-gravity waves on water of constant depth h is given by

$$\omega^2 = gk\left(1 + \frac{Tk^2}{\rho g}\right)\tanh kh\,. \tag{100}$$

The phase velocity is

$$c(k) = \left[\frac{g}{k}\left(1 + \frac{Tk^2}{\rho g}\right)\tanh kh\right]^{\frac{1}{2}}\,. \tag{101}$$

The corresponding dispersion relation for deep water ($kh \gg 1$) is

$$\omega^2 = gk(1 + T^*)\,, \tag{102}$$

where $T^* = (Tk^2/g\rho) = (4\pi^2 T/g\rho\lambda^2)$ is a parameter giving the relative importance of surface tension and gravity. The phase velocity is

$$c(k) = \left[\frac{g}{k}(1 + T^*)\right]^{\frac{1}{2}} = \left[(\frac{g\lambda}{2\pi})(1 + T^*)\right]^{\frac{1}{2}}\,. \tag{103}$$

The phase velocity $c(k)$ has a minimum value at $k = k_m = \sqrt{g\rho/T}$ (or $T^* = 1$). The corresponding minimum value for $c = c_m$ attained at $k = k_m = \sqrt{g\rho/T}$ (or $T^* = 1$) is

$$c = c_m = \left(\frac{4gT}{\rho}\right)^{\frac{1}{4}} \tag{104}$$

at wavelength $\lambda = \lambda_m = 2\pi(T/g\rho)^{\frac{1}{2}}$.

The inequality $k \ll k_m$ holds for waves to be effectively pure gravity waves, since the surface tension is negligible by comparison. This inequality is equivalent to the wavelength λ to be large compared with $\lambda_m = (2\pi/k_m) = 2\pi(T/g\rho)^{\frac{1}{2}}$. The phase velocity $c(\lambda)$ for capillary-gravity waves against λ in deep water is shown in Fig. 3. This figure shows the transition between the pure capillary-wave value $(2\pi T/\lambda\rho)^{\frac{1}{2}}$ and the pure gravity-wave value $(g\lambda/2\pi)^{\frac{1}{2}}$. Also shown in this figure are (i) the minimum phase velocity c_m attained at $\lambda = \lambda_m$, (ii) the gravity-wave curve ($T^* = 0$) for $\lambda > \lambda_m$, and (iii) the capillary-wave curve with no gravitational effect $g \to 0$ or $T^* \to \infty$. Figure 3 also shows that for $\lambda > 4\lambda_m$ the capillary-gravity wave curve is rapidly running closer with the gravity-wave curve. On the other hand, when $\lambda < \lambda_m$, the rapid tendency is for wave speeds to increase again so that the capillary-gravity wave curve approaches the capillary wave curve as $T^* \to \infty$, which corresponds to very short waves called *capillary waves* (or *ripples*). In fact, when $\lambda < \frac{1}{4}\lambda_m$, results (102) and (103) for $T^* \to \infty$ give

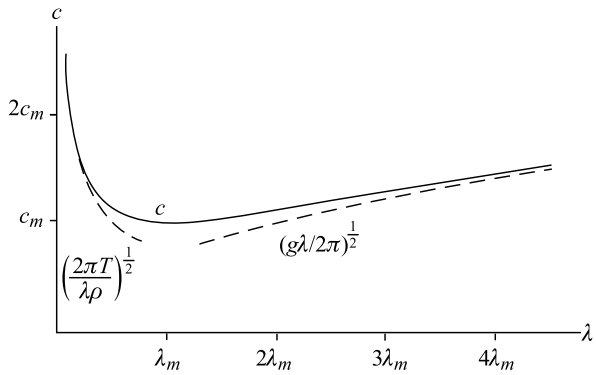

Water Waves and the Korteweg–de Vries Equation, Figure 3
The phase velocity $c(\lambda)$ given by (103) against λ (from [46], Figure 56, page 224)

$$\omega^2 = \rho^{-1}Tk^3 \quad \text{and} \quad c(k) = (\rho^{-1}Tk)^{\frac{1}{2}}\,. \tag{105ab}$$

For such capillary waves, surface tension is the only significant restoring force.

The Stokes Waves and Nonlinear Dispersion Relation

In his 1847 classic paper, Stokes [54] first established the nonlinear solutions for periodic plane waves on deep water. We consider here some of the nonlinear effects neglected in the linearized theory. A more simple approach is to recall the exact free surface dynamic condition (41) with constant atmospheric pressure and the negligible surface tension so that (41) becomes

$$(\phi_{tt} + g\phi_z) + 2\nabla\phi\cdot\nabla\phi_t + \frac{1}{2}\nabla\phi\cdot\nabla(\nabla\phi)^2 = 0\,, \quad \text{on } z = \eta\,, \quad \text{for } t \geq 0\,. \tag{106}$$

This condition is applied on the unknown free surface $z = \eta$ given by (39), that is,

$$\eta = -\frac{1}{g}\left[\phi_t + \frac{1}{2}(\nabla\phi)^2\right]\,. \tag{107}$$

In the absence of nonlinear terms, results (106) and (107) reduce to (84ab) and (84b), respectively.

We use Taylor series expansions of ϕ and its derivatives from $z = \eta$ to $z = 0$ in the form

$$\phi(x, y, \eta, t) = \phi(x, y, 0, t) + \eta \left(\frac{\partial \phi}{\partial z}\right)_{z=0} + \frac{1}{2}\eta^2 \left(\frac{\partial^2 \phi}{\partial z^2}\right)_{z=0} + \dots . \quad (108)$$

Using this expansion procedure for each of the derivatives in (106) and (107), we can generate a series of boundary conditions on the surface plane $z = 0$. Thus, we obtain first three conditions:

$$L\phi = \phi_{tt} + g\phi_z = 0 + O(\phi^2) , \quad (109)$$

$$L\phi + 2\nabla\phi \cdot \nabla\phi_t - \frac{1}{g}\phi_t \frac{\partial}{\partial z}(L\phi) = 0 + O(\phi^3) , \quad (110)$$

$$\begin{aligned} L\phi &+ 2\nabla\phi \cdot \nabla\phi_t \\ &+ \frac{1}{2}\nabla\phi \cdot \nabla(\nabla\phi)^2 \\ &- \frac{1}{g}\phi_t \frac{\partial}{\partial z}(L\phi + 2\nabla\phi \cdot \nabla\phi_t) \\ &- \frac{1}{g}\left[-\frac{1}{g}\phi_t\phi_{zt} + \frac{1}{2}(\nabla\phi)^2\right]\frac{\partial}{\partial z}(L\phi) \\ &+ \frac{1}{2g^2}\phi_t^2 \frac{\partial^2}{\partial z^2}(L\phi) \\ &= 0 + O(\phi^4) , \end{aligned} \quad (111)$$

where the symbol $O(\,)$ is used to indicate the magnitude of the neglected terms.

If we substitute the first-order velocity potential (93) for deep water in the second-order boundary condition (110), the second-order terms in (110) vanish. Thus, the first-order potential is a solution of the second-order boundary value problem, and we can write

$$\phi = \left(\frac{ga}{\omega}\right) e^{kz} \sin(kx - \omega t) + O(a^3) . \quad (112)$$

Similarly, using (108), we can incorporate the second-order effects in $\eta(x, z, t)$ so that

$$\begin{aligned} \eta(x, z, t) &= -\frac{1}{g}\left[\phi_t + \frac{1}{2}(\nabla\phi)^2\right]_{z=\eta} \\ &= -\frac{1}{g}\left[\phi_t + \frac{1}{2}(\nabla\phi)^2\right]_{z=0} \\ &\quad + \eta\frac{\partial}{\partial z}\left[-\frac{1}{g}\left\{\phi_t + \frac{1}{2}(\nabla\phi)^2\right\}\right]_{z=0} + \dots \\ &= -\frac{1}{g}\left[\phi_t + \frac{1}{2}(\nabla\phi)^2 - \frac{1}{g}\phi_t\phi_{zt}\right]_{z=0} + \dots . \end{aligned} \quad (113)$$

Direct substitution of (112) in (113) gives the following second-order result:

$$\begin{aligned} \eta &= a\cos(kx - \omega t) \\ &\quad + \frac{1}{2}ka^2\left\{2\cos^2(kx - \omega t) - 1\right\} + \dots \\ &= a\cos\theta + \frac{1}{2}ka^2\cos 2\theta + \dots , \end{aligned} \quad (114)$$

where the phase $\theta = kx - \omega t$. Clearly, the second term in (114) represents the nonlinear effects on the free-surface profile, and it is positive both at the crests $\theta = 0, 2\pi, 4\pi, \dots$ and at the troughs $\theta = 0, 3\pi, 5\pi, \dots$. The notable feature of this solution (114) is that the wave profile is no longer sinusoidal as shown in Fig. 2.7 of Debnath [19].

The actual shape of this wave profile is a curve known as a *trochoid*: The crests are steeper and the troughs are flatter. This feature becomes accentuated as the wave amplitude is increased.

For the third-order free-surface condition (111), direct substitution of the plane wave potential (112) in the nonlinear terms in (111) eliminates all but one term. Therefore, the free-surface boundary condition for the third-order plane wave solution is

$$L\phi + \frac{1}{2}\nabla\phi \cdot \nabla(\nabla\phi)^2 = 0 + O(\phi^4) . \quad (115)$$

The first-order solution (112) satisfies this third-order boundary condition so that the dispersion relation (95a) includes a second-order effect of the form

$$\omega^2 = gk(1 + a^2k^2) + O(a^3k^3) . \quad (116)$$

This remarkable dependence of frequency on wave amplitude is usually known as *amplitude* (or *nonlinear*) *dispersion*. The modified phase velocity expression is

$$\begin{aligned} c(k) &= \frac{\omega}{k} = \left(\frac{g}{k}\right)^{\frac{1}{2}}(1 + k^2a^2)^{\frac{1}{2}} \\ &\approx \left(\frac{g}{k}\right)^{\frac{1}{2}}(1 + \frac{1}{2}a^2k^2) . \end{aligned} \quad (117)$$

The significant change from the linearized theory is (116) or (117), which confirms that the phase velocity now depends on amplitude a as well as on wavelength; steep waves travel faster than less steep waves of the same wavelength. Surface gravity waves thus acquire amplitude dispersion as a second-order correction. The dependence of c on amplitude is known as the *amplitude dispersion* in contrast to the *frequency dispersion* as given by (95a). It may be noted that Stokes' results (114), (116), and (117) can easily be approximated further to obtain solutions for long waves (or shallow water) and for short waves (or deep water).

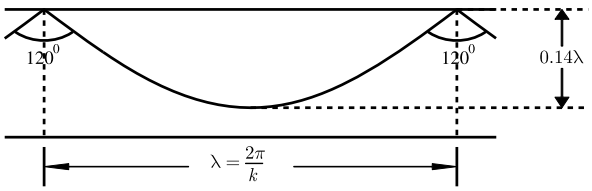

Water Waves and the Korteweg–de Vries Equation, Figure 4
The steepest wave profile

We next discuss the phenomenon of breaking of water waves which is one of the most common observable phenomena on an ocean beach. A wave coming from deep ocean changes shape as it moves across a shallow beach. Its amplitude and wavelength also are modified. The wavetrain is very smooth some distance offshore, but as it moves inshore, the front of the wave steepens noticeably until, finally, it breaks. After breaking, waves continue to move inshore as a series of bores or hydraulic jumps, whose energy is gradually dissipated by means of the water turbulence. Of the phenomena common to waves on beaches, breaking is the physically most significant and mathematically least known. In fact, it is one of the most intriguing longstanding problems of water waves.

For waves of small amplitude in deep water, the maximum particle velocity is $v = a\omega = ack$. But the basic assumption of small amplitude theory implies that $\frac{v}{c} = ak \ll 1$. Therefore, wave breaking can never be predicted by the small amplitude wave theory. That possibility arises only in the theory of finite amplitude waves. It is to be noted that the Stokes expansions are limited to relatively small amplitude and cannot predict the wavetrain of maximum height at which the crests are found to be very sharp. For a wave profile of constant shape moving at a uniform velocity, it can be shown that the maximum total crest angle as the wave begins to break is 120° as shown in Fig. 4.

The upshot of the Stokes analysis reveals that the inclusion of higher-order terms in the representation of the surface wave profile distorts its shape away from the linear sinusoidal curve. The effects of nonlinearity are likely to make crests narrower (sharper) and the troughs flatter as depicted in Figure 2.7 of Debnath [19]. The resulting wave profile more accurately portrays the water waves that are observed in nature. Finally, the sharp crest angle of 120° was first found by Stokes.

On the other hand, in 1865, Rankine conjectured that there exists a wave of extreme height. In a moving reference frame, the Euler equations are Galilean invariant, and the Bernoulli Eq. (36) on the free surface of water with $\rho = 1$ becomes

$$\tfrac{1}{2}|\nabla\phi|^2 + gz = E\,. \tag{118}$$

Thus, this equation represents the conservation of local energy, where the first term is the kinetic energy of the fluid and the second term is the potential energy due to gravity. For the wave of maximum height, $E = gz_{\max}$, where $z_{\max}$ is the maximum height of the fluid. Thus, the velocity is zero at the maximum height so that there will be a stagnation point in the fluid flow. Rankine conjectured that a cusp is developed at the peak of the free surface with a vertical slope so that the angle subtended at the peak is 120° as also conjectured by Stokes [54]. Toland [56] and Amick et al. [4] have proved rigorously the existence of a wave of greatest height and the Stokes conjecture for the wave of extreme form. However, Toland [56] also proved that if the singularity at the peak is *not* a cusp, that is, if there is no vertical slope at the peak of the free surface, then the Stokes remarkable conjecture of the crest angle of 120° is true. Subsequently, Amick et al. [4] confirmed that the singularity at the peak is *not* a cusp. Therefore, the full Euler equations exhibit singularities, and there is a limiting amplitude to the periodic waves.

The nonlinear solutions for plane waves based on systematic power series in the wave amplitude are known as *Stokes expansions*. Stokes [54] showed that the free surface elevation η of a plane wavetrain on deep water can be expanded in powers of the amplitude a as

$$\eta = a\cos\theta + \frac{1}{2}ka^2\cos 2\theta + \frac{3}{8}k^2a^3\cos 3\theta + \ldots\,, \tag{119}$$

where $\theta = kx - \omega t$ and the square of the frequency is

$$\omega^2 = gk\left(1 + a^2k^2 + \frac{5}{4}a^4k^4 + \ldots\right)\,. \tag{120}$$

A question was raised about the convergence of the Stokes expansion in order to prove the existence of solution representing periodic water waves of permanent form. Considerable attention had been given to this problem by several authors. The problem was eventually resolved by Levi–Civita [43], who proved formally that the Stokes expansion for deep water converges provided the wave steepness (ak) is very small. Struik [55] extended the proof of Levi–Civita to small-amplitude waves on water of arbitrary, but constant depth. Subsequently, Kraskovskii [40,41] finally established the existence of permanent periodic waves for all amplitudes less than the extreme value at which the waves assume a sharp-crested form. Although the success of the perturbation expansion as a means of representing waves of finite amplitude depends on the convergence of Stokes' expansion, convergence *does not* at all imply stability. In spite of the preceding attempts to establish the possibility of finite-am-

plitude water waves of permanent form, the question of their stability was altogether ignored until the 1960s. The independent work of Lighthill [44,45], Benjamin [5] and Whitham [60,61] finally established that Stokes waves on deep water are definitely *unstable*! This is one of the most remarkable discoveries in the history of the subject of the theory of water waves.

As a complement to the Stokes theory of pure gravity waves, Crapper [16] first discovered the *exact nonlinear* solution for pure progressive capillary waves of arbitrary amplitude by using a complex variable method. He obtained a remarkably new result for the phase velocity

$$c(k) = \left(\frac{kT}{\rho}\right)^{\frac{1}{2}} \left(1 + \frac{\pi^2 H^2}{4\lambda}\right)^{-1/4} , \tag{121}$$

where $H = 2a$ is the wave height representing the vertical distance between crest and trough. This result clearly shows that the phase velocity decreases as the wave amplitude increases for a fixed wavelength. Result (121) is now known as *Crapper's nonlinear capillary wave solution*. It has also been shown by Crapper that the capillary wave of greatest height occurs when $H = 0.73\lambda$, which has a striking contrast with the corresponding result due to Michell [48] for pure gravity waves, $H = 0.142\lambda = \frac{1}{7}\lambda$. Figure 5 exhibits Crapper's nonlinear capillary wave profiles.

One of the important features of Crapper's solution is that there is a maximum possible steepness for pure capillary waves, $H/\lambda = 0.73$, at which the waves hit each other. In the case of pure gravity waves, Stokes suggested that the wave steepness would possibly be greatest when the crest actually becomes a point with the maximum included crest angle of 120°. Miche [49] obtained the maximum wave height in water of finite depth h as

$$\left(\frac{H}{\lambda}\right)_{\max} = (0.14) \tanh kh , \tag{122}$$

which, in very shallow water ($kh \ll 1$), gives

$$\left(\frac{H}{h}\right)_{\max} = 0.88 . \tag{123}$$

This result is in excellent agreement with experimental observations. As the ratio of wave amplitude to local water depth increases, waves are gradually deformed and seem to behave more and more like a series of solitary waves or a train of cnoidal waves. As revealed by reliable observations, soon they become unstable and then break, forming a whitecap. However, little is known about the detailed nature of wave breaking except for the fact that it is a typical complicated nonlinear phenomenon. It has also been predicted that swell in shallow water tends to break when the crest-to-trough height is about 0.88 times the local water depth.

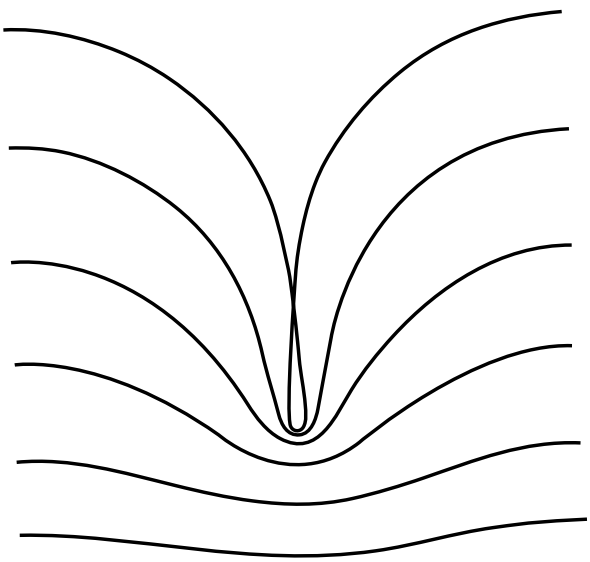

Water Waves and the Korteweg–de Vries Equation, Figure 5
Surface profile for nonlinear capillary waves up to maximum possible wave, $(H/\lambda) = 0.73$ (from [16])

Surface Gravity Waves on a Running Stream in Water

Debnath and Rosenblat [18] solved the *initial value problem* for the generation and propagation of two-dimensional surface waves at the free surface of a running stream on water of finite depth. With the aid of generalized functions, it is proved that the asymptotic solution of the initial value problem leads to the ultimate steady-state solution and the transient solution *without* the need to resort to the use of a radiation condition at infinity or equivalent device.

The fundamental two-dimensional water wave equations on a running stream with velocity U in the x-direction in water of depth h (see Debnath [19], page 115) are given by

$$\phi_{xx} + \phi_{zz} = 0 , \quad -h \le z \le 0 , \quad -\infty < x < \infty , \quad t > 0 , \tag{124}$$

$$\eta_t + U\eta_x - \phi_z = 0 , \quad \text{on } z = 0 , \quad t > 0 , \tag{125}$$

$$\phi_t + U\phi_x + g\eta = -\frac{P}{\rho} p(x) e^{i\omega t} , \quad z = 0 , \quad t > 0 , \tag{126}$$

$$\phi_z = 0 , \quad \text{on } z = -h , \tag{127}$$

where the term on the right hand side of (126) represents the arbitrary applied pressure with frequency ω, P is a constant and $p(x)$ is arbitrary function of x.

Making reference to Debnath [19] for a detailed method of solution and asymptotic analysis of the solution of the initial value problem, we discuss only the *inherent resonant* (*critical*) *phenomenon* involved in this problem. For certain limiting values of the three quantities U, ω, and h, there exists a critical speed U_c such that

$$\text{(a)} \quad U = U_c = \frac{1}{4}\left(\frac{g}{\omega}\right) \quad \text{for } h \to \infty\,, \quad \omega \text{ and } U \text{ are finite}\,. \tag{128}$$

and

$$\text{(b)} \quad U = U_c = \sqrt{gh} \quad \text{for } \omega = 0\,, \quad h \text{ and } U \text{ are finite.} \tag{129}$$

A special property of (a) and (b) is that in both situations there exists a critical value of the speed U, above and below which the steady-state wave system has respectively quite a different character; and that when U assumes its critical value, U_c the solution is singular due presumably to a breakdown of the linearized theory. More precisely, when $U > U_c$, these are two surface waves downstream of the origin traveling with speeds (ω/s_1) and (ω/s_2), respectively, in the position x-direction; there are none upstream, where $-s_1$ and $-s_2$ are two roots (see Debnath and Rosenblat [18]) of the equations

$$(\omega + kU) \mp (gk \tanh kh)^{\frac{1}{2}} = 0\,. \tag{130ab}$$

On the other hand, when $U < U_c$, these are two more surface waves which move with speeds (ω/σ_1) and (ω/σ_2), both in the negative x-direction, where σ_1 and σ_2 are arising from (130a) at positive values of $k = \sigma_1$ and $k = \sigma_2$. The former of these exists only on the upstream side of the origin, and the latter only on the downstream side. This latter wave thus appears to originate at infinity, but this is only when it is viewed from the moving coordinate system. It is easily verified from equation (130a) that the speed (ω/σ_1) is always less than U, the speed of the frame, so that relative to axes at rest this wave travels in a positive direction. Depending on U, ω and h, there is a delimiting case when roots σ_1 and σ_2 coalesce into a double root. This is the critical case which demand a certain relationship between physical quantities involved, which can be combined into velocities U, $\sqrt{gh}$ and (g/ω). Indeed, when $h \to \infty$, we have the case (a), and when $\omega = 0$, the case (b) occurs.

When $\omega = 0$, the wave system is degenerate. In this case, $s_1 = s_2 = \sigma_1 = 0$ and the only wave that occurs is the one associated with the root σ_2. As before, it exists on the downstream side, when $U > U_c$, only, but is now a steady motion relative to the moving frame. This is the case discussed by Stoker ([53], p. 214) who found a similar behavior.

Combining the results (128) and (129) together, we see that there are three quantities involved having the dimensions of velocity, namely U, $\left(\frac{g}{\omega}\right)$ and $\sqrt{gh}$. In general, all these three quantities are finite, and hence the critical situation can be expressed as a functional relationship between them of the form

$$F\left(U,\ g/\omega,\ \sqrt{gh}\right) = 0\,. \tag{131}$$

This function F is found and is seen to yield U as a single-valued function of (g/ω) and $\sqrt{gh}$, with formulas (128) and (129) emerging as limiting cases.

Debnath and Rosenblat [18] also showed that the free surface elevation $\eta(x, t)$ becomes singular when roots σ_1 and σ_2 coalesce into a double root, σ. Their asymptotic evaluation of $\eta(x, t)$ for large $|x|$ and t reveals that it represents two waves where the amplitude of one wave increases like x, while the amplitude of other wave is of $t^{\frac{1}{2}}$ as $t \to \infty$. This singular behavior is not unexpected and is in accord with the findings of Stoker [53] and Kaplan [37]. Mathematically, this critical situation corresponds to the coalescence of two roots $(\sigma_1 = \sigma_2 = \sigma)$, which is in turn is equivalent to the reinforcement by superposition of two like waves leading to a *resonance-type effect*. Physically, this situation would reveal itself through wave motions of large amplitude, which cannot come within the scope of linearized theory. Consequently, it would be necessary to include nonlinear terms in the original formulation of the problem in order to achieve a mathematically valid and physically reasonable solution.

Akylas [2] developed a nonlinear theory near the critical speed $U_c = \sqrt{gh}$ under the action of surface pressure $p(x)$ traveling at a constant speed U. With slow time scale $T = \varepsilon t$ and the slow space variable $X = \varepsilon^{\frac{1}{3}} x$, where $\varepsilon = \frac{a}{h}$, his asymptotic analysis reveals that the nonlinear response is of bounded amplitude $A(X, t)$ and is governed by a *forced Korteweg and de Vries* (*KdV*) equation in the form

$$A_T + \gamma A_X - 2AA_X - \tfrac{1}{6} A_{XXX} = \pi\, \overline{p}(0)\delta'(X)\,, \tag{132}$$

where $(U/\sqrt{gh}) = 1 + \gamma\, \varepsilon^{\frac{2}{3}}$, $\gamma = O(1)$, $\overline{p}(k)$ is the Fourier transform of $p(X)$ and $\delta(X)$ is the Dirac delta function.

The main conclusion of this asymptotic analysis is that the far field disturbance is of relatively large amplitude,

but it remains bounded. The main question whether the nonlinear response evolves to solitons, or disperses out or even gives a cnoidal wave, still remains open. However, it was shown by numerical study that Eq. (132) admits a series of solitons that are generated in front of the pressure field. For $\gamma < 0$, the linearized solution consists of a periodic sinusoidal wave in $X < 0$. The nonlinear evolution of the wave disturbance at the resonance condition ($\gamma = 0$) remains bounded. The major conclusion of this study is that a series of solitary waves (or solitons) is successively generated when $X < 0$ and propagates in front of the pressure field. This prediction is in theoretical agreement with the nonlinear study of Wu and Wu [64], and also in good agreement with experimental findings of Huang et al. [32] who observed solitons in their experiments involving a ship moving in shallow water.

Akylas [3] also developed a nonlinear theory to extend Debnath and Rosenblat's linearized study for the evolution of surface waves generated by a moving oscillatory pressure near the *critical* (or *resonant*) speed. Based on the slow time scale $T = \varepsilon t$ and the slow spatial variable $X = \sqrt{\varepsilon}\, x$, his nonlinear asymptotic analysis reveals that the evolution for the wave envelope $A(X, T)$ is governed by the *forced nonlinear Schrödinger equation*. For details, the reader is referred to Akylas [3] or Debnath [19].

History of Russell's Solitary Waves and Their Interactions

Historically, John Scott Russell first experimentally observed the solitary wave, a long water wave without change in shape, on the Edinburgh–Glasgow Canal in 1834. He called it the "*great wave of translation*" and then reported his observations at the British Association in his 1844 paper "Report on Waves." Thus, the solitary wave represents, not a periodic wave, but the propagation of a single isolated symmetrical hump of unchanged form. His discovery of this remarkable phenomenon inspired him further to conduct a series of extensive laboratory experiments on the generation and propagation of such waves. Based on his experimental findings, Russell discovered, empirically, one of the most important relations between the speed U of a solitary wave and its maximum amplitude a above the free surface of liquid of finite depth h in the form

$$U^2 = g(h + a)\,, \tag{133}$$

where g is the acceleration due to gravity. His experiments stimulated great interest in the subject of water waves and his findings received a strong criticism from G.B. Airy [1] and G.G. Stokes [54]. In spite of his remarkable work on the existence of periodic wavetrains representing a typical feature of nonlinear dispersive wave systems, Stokes' conclusion on the existence of the solitary wave was erroneous.

However, Stokes [54] proposed that the free surface elevation of the plane wavetrains on deep water can be expanded in powers of the wave amplitude. His original result for the dispersion relation in deep water is given by (120). Despite these serious attempts to prove the existence of finite-amplitude water waves of permanent form, the independent question of their stability remained unanswered until the 1960s except for an isolated study by Korteweg and de Vries in 1895 on long surface waves in water of finite depth. But one of the most remarkable discoveries made in the 1960s was that the periodic Stokes waves on sufficiently deep water are *definitely* unstable. This result seems revolutionary in view of the sustained attempts to prove the existence of Stokes waves of finite amplitude and permanent form. Scott Russell's discovery of solitary waves contradicted the theories of water waves due to Airy and Stokes; they raised questions on the existence of Russell's solitary waves and conjectured that such waves cannot propagate in a liquid medium without change of form. It was not until the 1870s that Russell's prediction was finally and independently confirmed by both J. Boussinesq [9,10,11,12] and Lord Rayleigh [51]. From the equations of motion for an inviscid incompressible liquid, they derived formula (133). In fact, they also showed that the Russell's solitary wave profile (see Fig. 6) $z = \eta(x, t)$ is given by

$$\eta(x, t) = a\ \mathrm{sech}^2\left[\beta(x - Ut)\right]\,, \tag{134}$$

where $\beta^2 = 3a \div \left\{4h^2(h + a)\right\}$ for any $a > 0$.

Although these authors found the sech^2 solution, which is valid only if $a \ll h$, they did not write any equation for η that admits (134) as a solution. However, Boussinesq made much more progress and discovered several

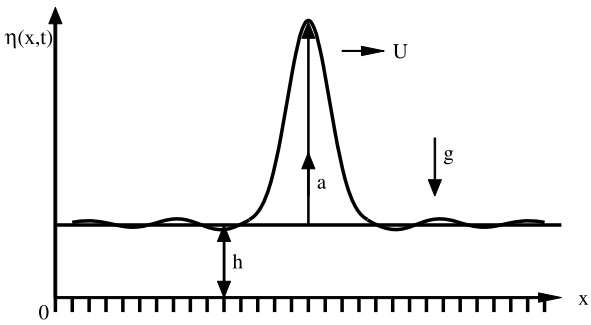

Water Waves and the Korteweg–de Vries Equation, Figure 6
A solitary wave

new ideas, including a nonlinear evolution equation for such long water waves in the form

$$\eta_{tt} = c^2 \left[\eta_{xx} + \frac{3}{2} \left(\frac{\eta^2}{h} \right)_{xx} + \frac{1}{3} h^2 \eta_{xxx} \right] , \tag{135}$$

where $c = \sqrt{gh}$ is the speed of the shallow water waves. This is known as the *Boussinesq (bidirectional) equation*, which admits the solution

$$\eta(x,t) = a \operatorname{sech}^2 \left[(3a/h^3)^{1/2} (x \pm Ut) \right] . \tag{136}$$

This represents solitary waves traveling in both, the positive and negative x-directions.

More than 60 years later, in 1895, two Dutchmen, D.J. Korteweg (1848–1941) and G. de Vries [39], formulated a mathematical model equation to provide an explanation of the phenomenon observed by Scott Russell. They derived the now-famous equation for the propagation of waves in one direction on the surface water of density ρ in the form

$$\eta_t = \frac{c}{h} \left[\left(\varepsilon + \frac{3}{2} \eta \right) \eta_X + \frac{1}{2} \sigma \, \eta_{XXX} \right] , \tag{137}$$

where X is a coordinate chosen to be moving (almost) with the wave, $c = \sqrt{gh}$, ε is a small parameter, and

$$\sigma = h \left(\frac{h^2}{3} - \frac{T}{g\rho} \right) \sim \frac{1}{3} h^3 , \tag{138}$$

when the surface tension $T \left(\ll \frac{1}{3} g \rho h^2 \right)$ is negligible. Equation (137) is known as the *Korteweg–de Vries (KdV) equation*. This is one of the simplest and most useful nonlinear model equations for solitary waves, and it represents the longtime evolution of wave phenomena in which the steepening effect of the nonlinear term is counterbalanced by the smoothening effect of the linear dispersion.

It is convenient to introduce the change of variables $\eta = \eta(X^*, t)$ and $X^* = X + (\varepsilon/h)ct$ which, dropping the asterisks, allows us to rewrite Eq. (137) in the form

$$\eta_t = \frac{c}{h} \left(\frac{3}{2} \eta \, \eta_X + \frac{1}{2} \sigma \, \eta_{XXX} \right) . \tag{139}$$

Modern developments in the theory and applications of the KdV solitary waves began with the seminal work published as a Los Alamos Scientific Laboratory Report in 1955 by Fermi, Pasta, and Ulam [23] on a numerical model of a discrete nonlinear mass-spring system. In 1914, Debye suggested that the finite thermal conductivity of an anharmonic lattice is due to the nonlinear forces in the springs. This suggestion led Fermi, Pasta, and Ulam to believe that a smooth initial state would eventually relax to an equipartition of energy among all modes because of nonlinearity. But their study led to the striking conclusion that there is no equipartition of energy among the modes. Although all the energy was initially in the lowest modes, after flowing back and forth among various low-order modes, it eventually returns to the lowest mode, and the end state is a series of recurring states. This remarkable fact has become known as the *Fermi–Pasta–Ulam* (FPU) *recurrence phenomenon*. Cercignani [13] and later on Palais [50] described the FPU experiment and its relationship to the KdV equation in some detail.

This curious result of the FPU experiment inspired Martin Kruskal and Norman Zabusky [65] to formulate a continuum model for the nonlinear mass-spring system to understand why recurrence occurred. In fact, they considered the initial-value problem for the KdV equation,

$$u_t + u u_x + \delta \, u_{xxx} = 0 , \tag{140}$$

where $\delta = \left(\frac{h}{\ell} \right)^2$, ℓ is a typical horizontal length scale, with the initial condition

$$u(x,0) = \cos \pi x , \quad 0 \le x \le 2 , \tag{141}$$

and the periodic boundary conditions with period 2, so that $u(x,t) = u(x+2,t)$ for all t. Their numerical study with $\sqrt{\delta} = 0.022$ produced remarkably new interesting results, which are shown in Fig. 7.

They observed that, initially, the wave steepened in regions where it had a negative slope, a consequence of the dominant effects of nonlinearity over the dispersive term, δu_{xxx}. As the wave steepens, the dispersive effect then becomes significant and balances the nonlinearity. At later times, the solution develops a series of *eight* well-defined waves, each like sech^2 functions with the taller (faster)

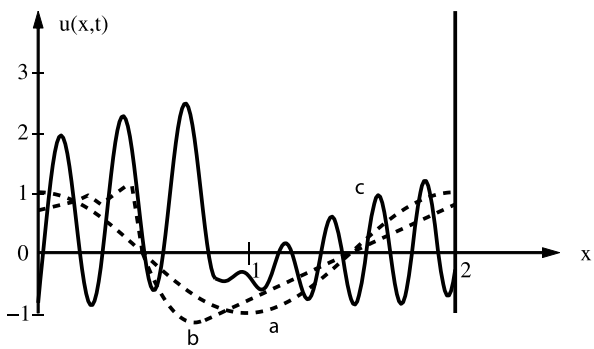

Water Waves and the Korteweg–de Vries Equation, Figure 7
Development of solitary waves: a initial profile at $t = 0$, b profile at $t = \pi^{-1}$, and c wave profile at $t = (3.6)\pi^{-1}$ (from [65])

waves ever catching up and overtaking the shorter (slower) waves. These waves undergo nonlinear interaction according to the KdV equation and then emerge from the interaction without change of form and amplitude, but with only a small change in their phases. So, the most remarkable feature is that these waves retain their identities after the nonlinear interaction. Another surprising fact is that the initial profile reappears, very similarly to the FPU recurrence phenomenon. In view of their preservation of shape and the resemblance to the particle-like character of these waves, Kruskal and Zabusky called these solitary waves, *solitons*, like photon, proton, electron, and other terms for elementary particles in physics.

Historically, the famous 1965 paper of Zabusky and Kruskal [65] marked the birth of the new concept of the *soliton*, a name intended to signify particle-like quantities. Subsequently, Zabusky [66] confirmed, numerically, the actual physical interaction of two solitons, and Lax [42] gave a rigorous analytical proof that the identities of two distinct solitons are preserved through the nonlinear interaction governed by the KdV equation. Physically, when two solitons of different amplitudes (and hence, of different speeds) are placed far apart on the real line, the taller (faster) wave to the left of the shorter (slower) wave, the taller one eventually catches up to the shorter one and then overtakes it. When this happens, they undergo a nonlinear interaction according to the KdV equation and emerge from the interaction completely preserved in form and speed with only a phase shift. Thus, these two remarkable features, (i) steady progressive pulse-like solutions and (ii) the preservation of their shapes and speeds, confirmed the particle-like property of the waves and, hence, the definition of the soliton. Subsequently, Gardner et al. [24,25] and Hirota [29,30,31] constructed analytical solutions of the KdV equation that provide the description of the interaction among N solitons for any positive integral N. After the discovery of the complete integrability of the KdV equation in 1967, the theory of the KdV equation and its relationship to the Euler equations of motion as an approximate model derived from the theory of asymptotic expansions became of major interest. From a physical point of view, the KdV equation is not only a standard nonlinear model for long water waves in a dispersive medium, it also arises as an approximate model in numerous other fields, including ion-acoustic plasma waves, magnetohydrodynamic waves, and anharmonic lattice vibrations. Experimental confirmation of solitons and their interactions has been provided successfully by Zabusky and Galvin [67], Hammack and Segur [26], and Weidman and Maxworthy [57]. Thus, these discoveries have led, in turn, to extensive theoretical, experimental, and computational studies over the last 40 years. Many nonlinear model equations have now been found that possess similar properties, and diverse branches of pure and applied mathematics have been required to explain many of the novel features that have appeared.

The Korteweg–de Vries and Boussinesq Equations

We consider an inviscid liquid of constant mean depth h and constant density ρ without surface tension. We assume that the (x, y)-plane is the undisturbed free surface with the z-axis positive upward. The free surface elevation above the undisturbed mean depth h is given by $z = \eta(x, y, t)$, so that the free surface is at $z = H = h + \eta$ and $z = 0$ is the horizontal rigid bottom (see Fig. 8).

It has already been recognized that the parameters $\varepsilon = (a/h)$ and $\kappa = ak$, where a is the surface wave amplitude and k is the wavenumber, must both be small for the linearized theory of surface waves to be valid. To develop the nonlinear shallow water theory, it is convenient to introduce the following nondimensional flow variables based on a different length scale h (which could be the fluid depth):

$$(x^*, y^*) = \frac{1}{l}(x, y)\,, \quad z^* = \frac{z}{h}\,, \quad t^* = \left(\frac{ct}{l}\right)\,,$$
$$\eta^* = \frac{\eta}{a}\,, \quad \phi^* = \left(\frac{h}{alc}\right)\phi\,, \qquad (142)$$

where $c = \sqrt{gh}$ is the typical horizontal velocity (or shallow water wave speed).

We next introduce two fundamental parameters to characterize the nonlinear shallow water waves:

$$\varepsilon = \frac{a}{h} \quad \text{and} \quad \delta = \frac{h^2}{l^2}\,, \qquad (143)$$

where ε is called the *amplitude* parameter and $\sqrt{\delta}$ is called the *long wavelength* or *shallowness* parameter.

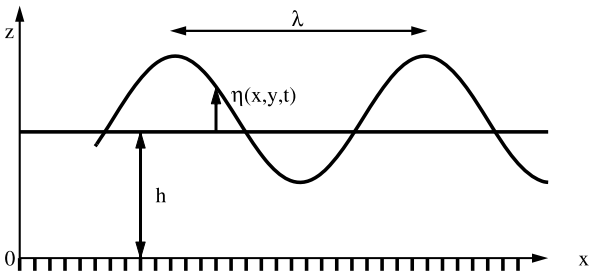

Water Waves and the Korteweg–de Vries Equation, Figure 8
A shallow water wave model

In terms of the preceding nondimensional flow variables and the parameters, the basic equations for water waves (73)–(76) can be written in the nondimensional form, dropping the asterisks:

$$\delta(\phi_{xx} + \phi_{yy}) + \phi_{zz} = 0\,, \quad 0 \le z \le 1 + \varepsilon\eta\,, \tag{144}$$

$$\frac{\partial\phi}{\partial t} + \frac{\varepsilon}{2}(\phi_x^2 + \phi_y^2) + \frac{\varepsilon}{2\delta}\phi_z^2 + \eta = 0 \quad \text{on } z = 1 + \varepsilon\eta\,, \tag{145}$$

$$\delta\left[\eta_t + \varepsilon(\phi_x\eta_x + \phi_y\eta_y)\right] - \phi_z = 0 \quad \text{on } z = 1 + \varepsilon\eta\,, \tag{146}$$

$$\phi_z = 0 \quad \text{on } z = 0\,. \tag{147}$$

It is noted that the parameter $\kappa = ak$ does not enter explicitly in Eqs. (144)–(147), but an equivalent parameter $\gamma = (a/l)$ is associated with ε and δ through $\gamma = (a/h)\cdot(h/l) = \varepsilon\sqrt{\delta}$.

If ε is small, the terms involving ε in (145) and (146) can be neglected to recover the linearized free surface conditions. However, the assumption that δ is small might be interpreted as the characteristic feature of the shallow water theory. So, we expand ϕ in terms of δ without any assumption about ε, and write

$$\phi = \phi_0 + \delta\phi_1 + \delta^2\phi_2 + \ldots\,, \tag{148}$$

and then substitute in (144)–(146). The lowest-order term in (144) is

$$\phi_{0zz} = 0\,, \tag{149}$$

which, with (147), yields $\phi_{0z} = 0$, for all z, or $\phi_0 = \phi_0(x, y, t)$, which indicates that the horizontal velocity components are independent of the vertical coordinate z in lowest order. Consequently, we use the notation

$$\phi_{0x} = u(x, y, t) \quad \text{and} \quad \phi_{0y} = v(x, y, t)\,. \tag{150ab}$$

The first- and second-order terms in (144) are given by

$$\phi_{0xx} + \phi_{0yy} + \phi_{1zz} = 0\,, \tag{151}$$

$$\phi_{1xx} + \phi_{1yy} + \phi_{2zz} = 0\,. \tag{152}$$

Integrating (151) with respect to z and using (150ab) gives

$$\phi_{1z} = -z(u_x + v_y) + C(x, y, t)\,, \tag{153}$$

where the arbitrary function $C(x, y, t)$ becomes zero because of the bottom boundary condition (147). Integrating the resulting Eq. (153), again with respect to z and omitting the arbitrary constant, we obtain

$$\phi_1 = -\frac{z^2}{2}(u_x + v_y)\,, \tag{154}$$

so that $\phi_1 = 0$ at $z = 0$ and u and v are then the horizontal velocity components at the bottom boundary.

We next substitute (154) in (151) and (152), and then integrate with condition (147) to determine the arbitrary function. Consequently,

$$\phi_{2z} = \frac{1}{6}z^3\left[(\nabla^2 u)_x + (\nabla^2 v)_y\right]\,,$$
$$\phi_2 = \frac{1}{24}z^4\left[(\nabla^2 u)_x + (\nabla^2 v)_y\right]\,, \tag{155ab}$$

where ∇^2 is the two-dimensional Laplacian.

We next consider the free surface boundary conditions retaining all terms up to order δ, ε in (145), and δ^2, ε^2, and $\delta\varepsilon$ in (146). It turns out that conditions (145) and (146) become

$$\phi_{0t} - \frac{\delta}{2}(u_{tx} + v_{ty}) + \eta + \tfrac{1}{2}\varepsilon(u^2 + v^2) = 0\,, \tag{156}$$

$$\delta\left[\left\{\eta_t + \varepsilon(u\eta_x + v\eta_y)\right\} + (1 + \varepsilon\eta)(u_x + v_y)\right] = \frac{\delta^2}{6}\left[(\nabla^2 u)_x + (\nabla^2 v)_y\right]\,. \tag{157}$$

Differentiating (156) first with respect to x and then with respect to y gives two equations:

$$u_t + \varepsilon(uu_x + vv_x) + \eta_x - \frac{1}{2}\delta(u_{txx} + v_{txy}) = 0\,, \tag{158}$$

$$v_t + \varepsilon(uu_y + vv_y) + \eta_y - \frac{1}{2}\delta(u_{txy} + v_{tyy}) = 0\,. \tag{159}$$

Simplifying (157) yields

$$\eta_t + \left[u(1 + \varepsilon\eta)\right]_x + \left[v(1 + \varepsilon\eta)\right]_y = \frac{\delta}{6}\left[(\nabla^2 u)_x + (\nabla^2 v)_y\right]\,. \tag{160}$$

Evidently, Eqs. (158)–(160) represent the nondimensional *shallow water equations*.

Using the fact that ϕ_0 is irrotational, that is, $u_y = v_x$ and neglecting terms $O(\delta)$ in (158)–(160), we obtain the fundamental shallow water equations

$$u_t + \varepsilon(uu_x + vu_x) + \eta_x = 0\,, \tag{161}$$

$$v_t + \varepsilon(uv_x + vv_y) + \eta_y = 0\,, \tag{162}$$

$$\eta_t + \left[u(1 + \varepsilon\eta)\right]_x + \left[v(1 + \varepsilon\eta)\right]_y = 0\,. \tag{163}$$

This system of three, coupled, nonlinear equations is closed and admits some interesting and useful solutions for u, v, and η. It is equivalent to the boundary-layer equations in fluid mechanics. Finally, it can be linearized when $\varepsilon = (a/h) \ll 1$ to obtain the following dimensional equations:

$$u_t + g\,\eta_x = 0\,, \quad v_t + g\,\eta_y = 0\,,$$
$$\eta_t + h(u_x + v_y) = 0\,. \tag{164abc}$$

Eliminating u and v from these equations gives

$$\eta_{tt} = c^2(\eta_{xx} + \eta_{yy})\,. \tag{165}$$

This is a well-known *two-dimensional wave equation*. It corresponds to the nondispersive shallow water waves that propagate with constant velocity $c = \sqrt{gh}$. This velocity is simply the linearized version of $\sqrt{g(h+\eta)}$. The wave equation has the simple *d'Alembert solution* representing plane progressive waves

$$\eta(x,y,t) = f(k_1 x + l_1 y - \kappa_1 ct) + g(k_2 x + l_2 y - \kappa_2 ct)\,, \tag{166}$$

where f and g are arbitrary functions and $\kappa_r^2 = (k_r^2 + l_r^2)$, $r = 1, 2$.

We consider the one-dimensional case retaining both ε and δ order terms in (158)–(160) so that these equations reduce to the *Boussinesq* [9,10,11,12] *equations*

$$u_t + \varepsilon u u_x + \eta_x - \tfrac{1}{2}\,\delta\, u_{txx} = 0\,, \tag{167}$$

$$\eta_t + \big[u(1+\varepsilon\eta)\big]_x - \tfrac{1}{6}\,\delta\, u_{xxx} = 0\,. \tag{168}$$

On the other hand, Eqs. (161)–(163), expressed in dimensional form, read

$$u_t + u u_x + v u_y + g H_x = 0\,, \tag{169}$$

$$v_t + u v_x + v v_y + g H_y = 0\,, \tag{170}$$

$$H_t + (uH)_x + (vH)_y = 0\,, \tag{171}$$

where $H = (h + \eta)$ is the total depth and $H_x = \eta_x$, since the depth h is constant.

In particular, the one-dimensional version of the shallow water equations follows from (169)–(171) and is given by

$$u_t + u u_x + g H_x = 0\,, \tag{172}$$

$$H_t + (uH)_x = 0\,. \tag{173}$$

This system of approximate shallow water equations is analogous to the exact governing equations of gas dynamics for the case of a compressible flow involving only one space variable (see Riabouchinsky [52]).

It is convenient to rewrite these equations in terms of the wave speed $c = \sqrt{gH}$ by using $\mathrm{d}c = \big(g/2c\big)\,\mathrm{d}H$, so that they become

$$u_t + u u_x + 2c c_x = 0\,, \tag{174}$$

$$2c_t + c u_x + 2u c_x = 0\,. \tag{175}$$

The standard method of characteristics can easily be used to solve (174) and (175). Adding and subtracting these equations allows us to rewrite them in the characteristic form

$$\left[\frac{\partial}{\partial t} + (c+u)\frac{\partial}{\partial x}\right](u + 2c) = 0\,, \tag{176}$$

$$\left[\frac{\partial}{\partial t} + (c-u)\frac{\partial}{\partial x}\right](u - 2c) = 0\,. \tag{177}$$

Equations (176) and (177) show that $u + 2c$ propagates in the positive x-direction with velocity $c + u$, and $u - 2c$ travels in the negative x-direction with velocity $c - u$, that is, both $u + 2c$ and $u - 2c$ propagate in their respective directions with velocity c relative to the water. In other words,

$$u + 2c = \text{constant on curves } C_+ \text{ on which } \frac{\mathrm{d}x}{\mathrm{d}t} = u + c$$
$$u - 2c = \text{constant on curves } C_- \text{ on which } \frac{\mathrm{d}x}{\mathrm{d}t} = u - c \tag{178ab}$$

where C_+ and C_- are *characteristic curves* of the system of partial differential Eqs. (172) and (173). A disturbance propagates along these characteristic curves at speed c relative to the flow speed. The quantities $(u \pm 2c)$ are called the *Riemann invariants* of the system, and a simple wave is propagating to the right into water of depth h, that is, $u - 2c = c_0 = \sqrt{gh}$. Then, the solution is given by

$$u = f(\xi)\,, \quad x = \xi + \left(c_0 + \frac{3}{2}u\right)t\,, \tag{179}$$

where $u(x,t) = f(x)$ at $t = 0$. However, we note that

$$u_x = \left(1 - \frac{3}{2}u_x t\right) f'(\xi)\,,$$

giving

$$u_x = \frac{2f'(\xi)}{2 + 3t\,f'(\xi)}\,. \tag{180}$$

Thus, if $f'(\xi)$ is anywhere less than zero, u_x tends to infinity as $t \to -2/(3f')$. In terms of the free surface elevation, solution (179) implies that the wave profile progressively distorts itself, and, in fact, any forward-facing portion of such a wave continually steepens, or the higher parts of the wave tend to catch up with lower parts in front of them. Thus, all of these waves, carrying an increase of elevation, invariably break. The breaking of water waves on beaches is perhaps the most common and the most striking phenomenon in nature.

An alternative system equivalent to the nonlinear evolution Eqs. (167) and (168) can be derived from the nonlinear shallow water theory, retaining both ε and δ order terms with $\delta < 1$. This system is also known as the *Boussinesq equations*, which, in dimensional variables, are given by

$$\eta_t + \left[(h+\eta)u\right]_x = 0\,, \tag{181}$$

$$u_t + uu_x + g\eta_x = \frac{1}{3}\,h^2\,u_{xxt}\,. \tag{182}$$

They describe the evolution of long water waves that move in both positive and negative x-directions. Eliminating η and neglecting terms smaller than $O(\varepsilon,\delta)$ gives a single *Boussinesq equation* for $u(x,t)$ in the form

$$u_{tt} - c^2 u_{xx} + \frac{1}{2}(u^2)_{xt} = \frac{1}{3}h^2 u_{xxtt}\,. \tag{183}$$

The linearized Boussinesq equation for u and η follows from (181) and (182) as

$$\left[\frac{\partial^2}{\partial t^2} - c^2\frac{\partial^2}{\partial x^2} - \frac{1}{3}h^2\frac{\partial^4}{\partial x^2\partial t^2}\right]\begin{pmatrix} u \\ \eta \end{pmatrix} = 0\,. \tag{184}$$

This is in perfect agreement with the infinitesimal wave theory result expanded for small kh. Thus, the third derivative term in (182) may be identified with the frequency dispersion.

Another equivalent version of the Boussinesq equation is given by

$$\eta_{tt} - c^2\eta_{xx} = \frac{3}{2}\left(\frac{\eta^2}{h}\right)_{xx} + \frac{1}{3}h^2\eta_{xxxx}\,. \tag{185}$$

There are several features of this equation. It is a nonlinear partial differential equation that incorporates the basic idea of nonlinearity and dispersion. Boussinesq obtained three invariant physical quantities, Q, E, and M, defined by

$$Q = \int_{-\infty}^{\infty}\eta\,\mathrm{d}x\,, \quad E = \int_{-\infty}^{\infty}\eta^2\,\mathrm{d}x\,,$$
$$M = \int_{-\infty}^{\infty}\left[\eta_x^2 - 3\left(\frac{\eta}{h}\right)^3\right]\mathrm{d}x\,, \tag{186}$$

provided that $\eta \to 0$ as $|x| \to \infty$. Evidently, Q and E represent the volume and the energy of the solitary wave. The third quantity M is called the *moment of instability*, and the variational problem, $\delta M = 0$ with E fixed, leads to the unique solitary-wave solution. Boussinesq also derived the results for the amplitude and volume of a solitary wave of given energy in the form

$$a = \frac{3}{4}\left(\frac{E^{3/2}}{h}\right)\,, \quad Q = 2\,h\,E^{1/3}\,. \tag{187ab}$$

The former result shows that the amplitude of a solitary wave in a channel varies inversely as the channel depth h.

The Boussinesq equation can then be written in the normalized form

$$u_{tt} - u_{xx} - \tfrac{3}{2}(u^2)_{xx} - u_{xxxx} = 0\,. \tag{188}$$

This particular form is of special interest because it admits inverse scattering formalism. Equation (188) has steady progressive wave solutions in the form

$$u(x,t) = 4k^2 f(X)\,, \quad X = kx - \omega t\,, \tag{189}$$

where the equation for $f(X)$ can be integrated to obtain

$$f_{XX} = 6A + (4-6B)f - 6f^2\,, \tag{190}$$

where A is a constant of integration and

$$\omega^2 = k^2 + k^4(4-6B)\,. \tag{191}$$

For the special case $A = B = 0$, a single solitary-wave solution is given by

$$f(X) = \mathrm{sech}^2(X - X_0)\,, \tag{192}$$

where X_0 is a constant of integration. This result can be used to construct a solution for a series of solitary waves, spaced 2σ apart, in the form

$$f(X) = \sum_{n=-\infty}^{\infty}\mathrm{sech}^2(X - 2n\sigma)\,. \tag{193}$$

This is a 2σ periodic function that satisfies (190) for certain values of A and B.

We next assume that ε and δ are comparable, so that all terms $O(\varepsilon,\delta)$ in (158)–(160) can be retained. For the case of the two-dimensional wave motion ($v = 0$ and $\partial/\partial y = 0$), these equations become

$$u_t + \eta_x + \varepsilon\,u\,u_x - \tfrac{1}{2}\delta\,u_{txx} = 0\,, \tag{194}$$

$$\eta_t + \left[u(1+\varepsilon\eta)\right]_x - \tfrac{1}{6}\delta\,u_{xxx} = 0\,. \tag{195}$$

We now seek steady progressive wave solutions traveling to the positive x-direction only, so that $u = u(x - Ut)$ and $\eta = \eta(x - Ut)$. With the terms of zero order in ε and δ and $U = 1$, we assume a solution of the form

$$u = \eta + \varepsilon P + \delta\, Q\,, \tag{196}$$

where P and Q are unknown functions to be determined. Consequently, Eqs. (194) and (195) become

$$(\eta + \varepsilon P + \delta Q)_t + \eta_x + \varepsilon\eta\eta_x - \tfrac{1}{2}\delta\, \eta_{txx} = 0\,, \tag{197}$$

$$\eta_t + \big[(1 + \varepsilon\eta)(\eta + \varepsilon P + \delta Q)\big]_x - \tfrac{1}{6}\delta\, \eta_{xxx} = 0\,. \tag{198}$$

These equations must be consistent so that we stipulate for the zero order

$$\eta_t = -\eta_x\,, \quad P = -\tfrac{1}{4}\eta^2\,, \quad Q = \tfrac{1}{3}\eta_{xx} = -\tfrac{1}{3}\eta_{xt}\,. \tag{199}$$

We use these results to rewrite both (197) and (198) with the assumption that ε and δ are of equal order and small enough for their products and squares to be ignored, so that the ratio $(\varepsilon/\delta) = (al^2/h^3)$ is of the order one. Consequently, we obtain a single equation for $\eta(x, t)$ in the form

$$\eta_t + \left(1 + \tfrac{3}{2}\varepsilon\,\eta\right)\eta_x + \tfrac{1}{6}\delta\,\eta_{xxx} = 0\,. \tag{200}$$

This is now universally known as the *Korteweg and de Vries equation* as they discovered it in their 1895 seminal work. We point out that $(\varepsilon/\delta) = al^2/h^3$ is one of the *fundamental* parameters in the theory of nonlinear shallow water waves. Recently, Infeld [33] considered three-dimensional generalizations of the Boussinesq and Korteweg–de Vries equations.

Solutions of the KdV Equation: Solitons and Cnoidal Waves

To find solutions of the KdV equation, it is convenient to rewrite it in terms of dimensional variables as

$$\eta_t + c\left(1 + \frac{3}{2h}\eta\right)\eta_x + \frac{ch^2}{6}\eta_{xxx} = 0\,, \tag{201}$$

where $c = \sqrt{gh}$, and the total depth $H = h + \eta$. The first two terms $(\eta_t + c\,\eta_x)$ describe wave evolution at the shallow water speed c, the third term with coefficient $(3c/2h)$ represents a nonlinear wave steepening, and the last term with coefficient $(ch^2/6)$ describes linear dispersion. Thus, the KdV equation is a balance between time evolution, nonlinearity, and linear dispersion. The dimensional velocity u is obtained from (196) with (199) in the form

$$u = \frac{c}{h}\left(\eta - \frac{1}{4h}\eta^2 + \frac{h}{3}\eta_{xx}\right)\,. \tag{202}$$

We seek a traveling wave solution of (201) in the frame X so that $\eta = \eta(X)$ and $X = x - Ut$ with $\eta \to 0$, as $|x| \to \infty$, where U is a constant speed. Substituting this solution in (201) gives

$$(c - U)\eta' + \frac{3c}{2h}\eta\eta' + \frac{ch^2}{6}\eta''' = 0\,, \tag{203}$$

where $\eta' = \mathrm{d}\eta/\mathrm{d}X$. Integrating this equation with respect to X yields

$$(c - U)\eta + \frac{3c}{4h}\eta^2 + \frac{ch^2}{6}\eta'' = A\,, \tag{204}$$

where A is an integrating constant.

We multiply this equation by $2\eta'$ and integrate again to obtain

$$(c - U)\eta^2 + \left(\frac{c}{2h}\right)\eta^3 + \left(\frac{ch^2}{6}\right)\left(\frac{\mathrm{d}\eta}{\mathrm{d}X}\right)^2 = 2A\eta + B\,, \tag{205}$$

where B is also a constant of integration.

We now consider a special case when η and its derivatives tend to zero at infinity and $A = B = 0$, so that (205) gives

$$\left(\frac{\mathrm{d}\eta}{\mathrm{d}X}\right)^2 = \frac{3}{h^3}\eta^2(a - \eta)\,, \tag{206}$$

where

$$a = 2h\left(\frac{U}{c} - 1\right)\,. \tag{207}$$

The right-hand side of (206) vanishes at $\eta = 0$ and $\eta = a$, and the exact solution of (206) represents Russull's solitary wave in the form

$$\eta(X) = a\,\mathrm{sech}^2(bX)\,, \quad b = \left(\frac{3a}{4h^3}\right)^{1/2}\,. \tag{208ab}$$

Thus, the explicit form of the solution is

$$\eta(x, t) = a\,\mathrm{sech}^2\left[\left(\frac{3a}{4h^3}\right)^{1/2}(x - Ut)\right]\,, \tag{209}$$

where the velocity of the wave is

$$U = c\left(1 + \frac{a}{2h}\right)\,. \tag{210}$$

This is an *exact* solution of the KdV equation for all (a/h); however, the equation is derived with the approximation

$(a/h) \ll 1$. The solution (209) is called a *soliton* (or *solitary wave*) describing a single hump of height a above the undisturbed depth h and tending rapidly to zero away from $X = 0$. The solitary wave propagates to the right with velocity $U(> c)$, which is directly proportional to the amplitude a and has width $b^{-1} = (3a/4h^3)^{-1/2}$, that is, b^{-1} is inversely proportional to the square root of the amplitude a. Another significant feature of the soliton solution is that it travels in the medium without change of shape, which is hardly possible without retaining δ-order terms in the governing equation. A solitary wave profile has already been shown in Fig. 6.

In the general case, when both A and B are nonzero, (205) can be rewritten as

$$\frac{h^3}{3}\left(\frac{\mathrm{d}\eta}{\mathrm{d}X}\right)^2 = -\eta^3 + 2h\left(\frac{U}{c} - 1\right)\eta^2 + \frac{2h}{c}(2A\eta + B) = F(\eta)\,, \quad (211)$$

where $F(\eta)$ is a cubic with simple zeros.

We seek a real bounded solution for $\eta(X)$, which has a minimum value zero and a maximum value a and oscillates between the two values. For bounded solutions, all three zeros η_1, η_2, η_3 must be real. Without loss of generality, we set $\eta_1 = 0$ and $\eta_2 = a$. Hence, the third zero must be negative so that $\eta_3 = -(b - a)$ with $b > a > 0$. With these choices, $F(\eta) = \eta(a - \eta)(\eta - a + b)$ and Eq. (211) assumes the form

$$\frac{h^3}{3}\left(\frac{\mathrm{d}\eta}{\mathrm{d}X}\right)^2 = \eta(a - \eta)(\eta - a + b)\,, \quad (212)$$

where

$$U = c\left(1 + \frac{2a - b}{2h}\right)\,, \quad (213)$$

which is obtained by comparing the coefficients of η^2 in (211) and (212).

Writing $a - \eta = p^2$, it follows from Eq. (212) that

$$\left(\frac{3}{4h^3}\right)^{1/2} \mathrm{d}X = \frac{\mathrm{d}p}{\left[(a - p^2)(b - p^2)\right]^{1/2}}\,. \quad (214)$$

Substituting $p = \sqrt{a}\, q$ in (214) gives the standard elliptic integral of the first kind (see Dutta and Debnath [22] and Helal and Molines [28])

$$\left(\frac{3b}{4h^3}\right)^{1/2} X = \int_0^q \frac{\mathrm{d}q}{\left[(1 - q^2)(1 - m^2 q^2)\right]^{1/2}}\,, \qquad m = \left(\frac{a}{b}\right)^{1/2}\,, \quad (215)$$

and then, function q can be expressed in terms of the Jacobian elliptic *function*, $sn(z, m)$

$$q(X, m) = sn\left[\left(\frac{3b}{4h^3}\right)^{1/2} X, m\right]\,, \quad (216)$$

where m is the modulus of $sn(z, m)$.

Finally,

$$\eta(X) = a\left[1 - sn^2\left\{\left(\frac{3b}{4h^3}\right)^{1/2} X\right\}\right] = a\, cn^2\left[\left(\frac{3b}{4h^3}\right)^{1/2} X\right]\,, \quad (217)$$

where $cn(z, m)$ is also the Jacobian elliptic function with a period $2K(m)$, where $K(m)$ is the complete elliptic integral of the first kind defined by

$$K(m) = \int_0^{\pi/2} (1 - m^2 \sin^2\theta)^{-1/2}\, \mathrm{d}\theta\,, \quad (218)$$

and $cn^2(z) + sn^2(z) = 1$.

It is important to note that $cn\ z$ is periodic, and hence, $\eta(X)$ represents a train of periodic waves in shallow water. Thus, these waves are called *cnoidal waves* with wavelength

$$\lambda = 2\left(\frac{4h^3}{3b}\right)^{1/2} K(m)\,. \quad (219)$$

The upshot of this analysis is that solution (217) represents a nonlinear wave whose shape and wavelength (or period) all depend on the amplitude of the wave. A typical cnoidal wave is shown in Fig. 9. Sometimes, the cnoidal waves with slowly varying amplitude are observed in rivers. More often, wavetrains behind a weak bore (called an *undular bore*) can be regarded as cnoidal waves. Two limiting cases are of special physical interest: (i) $m \to 0$ and (ii) $m \to 1$.

In the first case, $sn\ z \to \sin\ z$, $cn\ z \to \cos\ z$ as $m \to 0$ $(a \to 0)$. This corresponds to small-amplitude waves where the linearized KdV equation is appropriate. So, in this limiting case, the solution (217) becomes

$$\eta(x, t) = \frac{1}{2} a\left[1 + \cos(kx - \omega t)\right]\,, \qquad k = \left(\frac{3b}{h^3}\right)^{1/2}\,, \quad (220)$$

Water Waves and the Korteweg–de Vries Equation, Figure 9
A cnoidal wave

where the corresponding dispersion relation is

$$\omega = Uk = ck\left(1 - \frac{1}{6}k^2h^2\right) . \tag{221}$$

This corresponds to the first two terms of the series expansion of $(gk \tanh kh)^{1/2}$. Thus, these results are in perfect agreement with the linearized theory.

In the second limiting case, $m \to 1 (a \to b)$, $cn\ z \longrightarrow \operatorname{sech} z$. Thus, the cnoidal wave solution tends to the classical KdV solitary-wave solution where the wavelength λ, given by (219), tends to infinity because $K(a) = \infty$ and $K(0) = \pi/2$. The solution identically reduces to (209) with (210).

We next report the numerical computation of the KdV Eq. (201) due to Berezin and Karpman [8]. In terms of new variables defined by

$$x^* = x - ct\,, \quad t^* = t\,, \quad \eta^* = \left(\frac{3c}{2h}\right)\eta\,, \tag{222}$$

omitting the asterisks, Eq. (201) becomes

$$\eta_t + \eta\eta_x + \beta\eta_{xxx} = 0\,, \tag{223}$$

where $\beta = \left(\frac{1}{6}\right) ch^2$.

We examine the numerical solution of (223) with the initial condition

$$\eta(x,0) = \eta_0\, f\left(\frac{x}{\ell}\right)\,, \tag{224}$$

where η_0 is constant and $f(\xi)$ is a nondimensional function characterizing the initial wave profile. It is convenient to introduce the dimensionless variables

$$\xi = \frac{x}{\ell}\,, \quad \tau = \frac{\eta_0\, t}{\ell}\,, \quad u = \frac{\eta}{\eta_0} \tag{225}$$

so that Eqs. (223) and (224) reduce to

$$u_\tau + uu_\xi + \sigma^{-2}\, u_{\xi\xi\xi} = 0\,, \tag{226}$$

$$u(\xi,0) = f(\xi)\,, \tag{227}$$

where the dimensionless parameter σ is defined by $\sigma = \ell(\eta_0/\beta)^{\frac{1}{2}}$.

Berezin and Karpman [8] obtained the numerical solution of (226) with the Gaussian initial pulse of the form $u(\xi,0) = f(\xi) = \exp(-\xi^2)$ and values of the parameter $\sigma = 1.9$ and $\sigma = 16.5$. Their numerical solutions are shown in Fig. 10.

As shown in Fig. 10 for case (a), the perturbation splits into a soliton and a wavepacket. In case (b), there are *six* solitons. It is readily seen that the peaks of the solitons lie nearly on a straight line. This is due to the fact that the velocity of the soliton is proportional to its amplitude, so that the distances traversed by the solitons would also be proportional to their amplitudes.

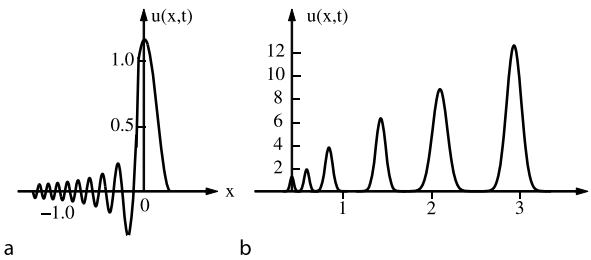

Water Waves and the Korteweg–de Vries Equation, Figure 10
The solutions of the KdV equation $u(x, t)$ for large values of t with the values of the similarity parameter σ: a $\sigma = 1.9$, b $\sigma = 16.5$ (from [8])

Zabusky's [66] numerical investigation of the interaction of two solitons reveals that the taller soliton, initially behind, catches up to the shorter one, they undergo a nonlinear interaction and, then, emerge from the interaction without any change in shape and amplitude. The end result is that the taller soliton reappears in front and the shorter one behind. This is essentially strong computational evidence of the stability of solitons.

Using the transformation

$$x^* = \varepsilon\beta(x - ct)\,, \quad t^* = \varepsilon^3 t\,, \quad \eta^* = \left(\alpha\,\varepsilon^2\right)^{-1}\eta$$

with $\alpha\beta(= 3c/2h) = 6$ and $\beta^3(= ch^2/6) = 1$, we write the KdV Eq. (201) in the normalized form, dropping the asterisks,

$$\eta_t + 6\eta\eta_x + \eta_{xxx} = 0\,. \tag{228}$$

We next seek a steady progressive wave solution of (228) in the form

$$\eta = 2k^2 f(X)\,, \quad X = kx - \omega t\,. \tag{229}$$

Then, the equation for $f(X)$ can be integrated once to obtain

$$f''(X) = 6A + (4 - 6B)f - 6f^2\,, \tag{230}$$

where A is a constant of integration and the frequency ω is given by

$$\omega = k^3(4 - 6B)\,. \tag{231}$$

A single-soliton solution corresponds to the special case $A = B = 0$ and is given by

$$f(X) = \operatorname{sech}^2(X - X_0)\,, \tag{232}$$

where X_0 is a constant.

Thus, for a series of solitons spaced 2σ apart, we write

$$f(X) = \sum_{n=-\infty}^{\infty} \operatorname{sech}^2(X - 2n\sigma) \,. \tag{233}$$

This is a 2σ periodic function that satisfies (230) for some A and B.

The general elliptic function solution of (228) can be obtained from the integral of (230), which can be written as

$$f_X^2 = -4C + 12Af + (4 - 6B)f^2 - 4f^3 \,, \tag{234}$$

where C is a constant of integration. Various asymptotic and numerical results lead to the relations (see Whitham [63])

$$C = -\frac{1}{2}\frac{\mathrm{d}A(\sigma)}{\mathrm{d}\sigma} = \frac{1}{4}\frac{\mathrm{d}^2B(\sigma)}{\mathrm{d}\sigma^2} \,, \tag{235ab}$$

and the cubic in (234) can be factorized as

$$-C + 3Af + \left(1 - \frac{3}{2}B\right) f^2 - f^3 = (f_1 - f)(f - f_2)(f - f_3) \,, \tag{236}$$

where $f_r(\sigma)(r = 1, 2, 3)$ are determined from $A(\sigma)$, $B(\sigma)$, and $C(\sigma)$. If we set $f_1 > f_2 > f_3$ and, then, the periodic solution oscillates between f_1 at $X = 0$ and f_2 at $X = \sigma$, one particular form of the solution is given by

$$f(X) = f_2 + (f_1 - f_2)cn^2(\sqrt{(f_1 - f_3)}\, X) \,, \tag{237}$$

where the modulus m of $cn(z, m)$ is given by

$$m^2 = \left(\frac{f_1 - f_2}{f_1 - f_3}\right) \,. \tag{238}$$

Thus, it follows from (233) and (237) that the following identity holds:

$$f_2 + (f_1 - f_2)cn^2(\sqrt{(f_1 - f_3)}\, X) = \sum_{n=-\infty}^{\infty} \operatorname{sech}^2(X - 2n\sigma) \,, \tag{239}$$

which can be verified by comparing the periods and poles of the two sides.

Finally, the higher-order modified KdV equation

$$v_t + (p + 1)v^p v_x + v_{xxxx} = 0 \,, \quad p > 2 \,, \tag{240}$$

admits single-soliton solutions in the form

$$v(x, t) = a \operatorname{sech}^{2/p}(kx - \omega t) \,. \tag{241}$$

However, in view of the fractional powers of the sech function, it seems, perhaps, unlikely that there will be any simple superposition formula.

Derivation of the KdV Equation from the Euler Equations

This problem was discussed in Sect. "The Korteweg–de Vries and Boussinesq Equations" by using the Laplace equation for the velocity potential under the assumption that $\delta = O(\varepsilon)$ as $\varepsilon \to 0$. Here we follow Johnson [35] to present another derivation of the KdV equation from the Euler equation in $(1 + 1)$ dimensions. This approach can be generalized to derive higher dimensional KdV equations.

We consider the problem of surface gravity waves which propagate in the positive x-direction over stationary water of constant depth. The associated Euler Eqs. (59) and the continuity Eq. (60) in $(1 + 1)$ dimensions are given by

$$u_t + \varepsilon(uu_x + wu_z) = -p_x \,, \tag{242}$$

$$\delta\left[w_t + \varepsilon(uw_x + ww_z)\right] = -p_z \,, \tag{243}$$

$$u_x + w_z = 0 \,. \tag{244}$$

The free surface and bottom boundary conditions are obtained from (57ab) and (62) in the form

$$w = \eta_t + \varepsilon u\eta_x \,, \quad p = \eta \,, \quad \text{on } z = 1 + \varepsilon\eta \,, \tag{245}$$

$$w = 0 \quad \text{on } z = 0 \,. \tag{246}$$

It can easily be shown that, for any $\sqrt{\delta}$ as $\varepsilon \to 0$, there exists a region in the (x, t)-space where there is a balance between nonlinearity and dispersion which leads to the KdV equation. The region of interest is defined by a scaling of the independent flow variables as

$$x \to \sqrt{\frac{\delta}{\varepsilon}}\, x \quad \text{and} \quad t \to \sqrt{\frac{\delta}{\varepsilon}}\, t \,, \tag{247}$$

for any ε and $\sqrt{\delta}$. In order to ensure consistency in the continuity equation, it is necessary to introduce a scaling of w by

$$w \to \sqrt{\frac{\varepsilon}{\delta}}\, w \,. \tag{248}$$

Consequently, the net effect of the scalings is to replace δ by ε in Eqs. (242)–(246) so that they become

$$u_t + \varepsilon(uu_x + wu_z) = -p_x \,, \tag{249}$$

$$\varepsilon\left[w_t + \varepsilon(uw_x + ww_z)\right] = -p_z \,, \tag{250}$$

$$u_x + w_z = 0 \,, \tag{251}$$

$$w = \eta_t + \varepsilon u\eta_x \,, \quad \text{and} \quad p = \eta \quad \text{on } z = 1 + \varepsilon\eta \,, \tag{252}$$

$$w = 0 \quad \text{on } z = 0 . \tag{253}$$

In the limit as $\varepsilon \to 0$, the first-order approximation of Eqs. (249) and (252) gives

$$\eta = p , \quad 0 \le z \le 1 , \quad \text{and} \quad u_t + \eta_x = 0 . \tag{254}$$

It then follows from (251) that $w = -z\, u_x$ which satisfies (253). The boundary condition (252) leads to $\eta_t + u_x = 0$ on $z = 1$, which can be combined with (254) to obtain the linear wave equation

$$\eta_{tt} - \eta_{xx} = 0 . \tag{255}$$

For waves propagating in the positive x-direction, we introduce the far-field variables

$$\xi = x - t \quad \text{and} \quad \tau = \varepsilon t , \tag{256}$$

so that $\xi = O(1)$ and $\tau = O(1)$ give the far-field region of the problem. This is the region where nonlinearity balances the dispersion to produce the KdV equation.

With the choice of the transformations (256), Eqs. (249)–(253) can be rewritten in the form

$$-u_\xi + \varepsilon(u_\tau + u u_\xi + w u_z) = -p_\xi , \tag{257}$$

$$\varepsilon\left[-w_\xi + \varepsilon(w_\tau + u w_\xi + w w_z)\right] = -p_z , \tag{258}$$

$$u_\xi + w_z = 0 , \tag{259}$$

$$w = -\eta_\xi + \varepsilon(\eta_\tau + u\eta_\xi) , \quad p = \eta , \quad \text{on } z = 1 + \varepsilon\eta , \tag{260}$$

$$w = 0 \quad \text{on } z = 0 . \tag{261}$$

We seek an asymptotic series expansion of the solutions of the system (257)–(261) in the form

$$\begin{aligned} \eta(\xi, \tau, \varepsilon) &= \sum_{n=0}^{\infty} \varepsilon^n \eta_n(\xi, \tau) , \\ q(\xi, \tau, z; \varepsilon) &= \sum_{n=0}^{\infty} \varepsilon^n q_n(\xi, \tau, z) , \end{aligned} \tag{262}$$

where q (and the corresponding q_n) denotes each of the variables u, w, and p.

Consequently, the leading-order problem is given by

$$u_{0\xi} = p_{0\xi} , \quad p_{0z} = 0 , \quad u_{0\xi} + w_{0z} = 0 , \tag{263}$$

$$p_0 = \eta_0 , \quad w + \eta_{0\xi} = 0 \quad \text{on } z = 1 , \tag{264}$$

$$w = 0 \quad \text{on } z = 0 . \tag{265}$$

These leading-order equations give

$$p_0 = \eta_0 , \quad u_0 = \eta_0 , \quad w_0 + z\,\eta_{0\xi} = 0 , \quad 0 \le z \le 1 , \tag{266}$$

with $u_0 = 0$ whenever $\eta_0 = 0$. The boundary condition on w_0 at $z = 1$ is automatically satisfied.

Using the Taylor series expansion of u, w, and p about $z = 1$, the two surface boundary conditions on $z = 1+\varepsilon\eta$ are rewritten on $z = 1$ and, hence, take the form

$$p_0 + \varepsilon\,\eta_0 p_{0z} + \varepsilon\, p_1 = \eta_0 + \varepsilon\eta_1 + O(\varepsilon^2) \quad \text{on } z = 1 . \tag{267}$$

$$w_0 + \varepsilon\,\eta_0\, w_{0z} + \varepsilon\, w_1 = -\eta_{0\xi} - \varepsilon\,\eta_{1\xi} + \varepsilon(\eta_{0\tau} + u_0\eta_{0\xi}) + O(\varepsilon^2) \quad \text{on } z = 1 . \tag{268}$$

These conditions are to be used together with (257), (258), and (261).

The equations in the next order are given by

$$-u_{1\xi} + u_{0\tau} + u_0 u_{0\xi} + w_0 u_{0z} = -p_{1\xi} , \quad p_{1z} = w_{0\xi} , \tag{269}$$

$$u_{1\xi} + w_{1z} = 0 , \tag{270}$$

$$\begin{aligned} &p_1 + \eta_0 p_0 = \eta_1 , \\ &w_1 + \eta_0 + w_{0z} = -\eta_{1\xi} + \eta_{0\tau} + u_0\eta_{0\xi} \quad \text{on } z = 1 , \end{aligned} \tag{271}$$

$$w_1 = 0 \quad \text{on } z = 0 . \tag{272}$$

Noting that

$$u_{0z} = 0 , \quad p_{0z} = 0 , \quad \text{and} \quad w_{0z} = -\eta_{0\xi} , \tag{273}$$

we obtain

$$p_1 = \tfrac{1}{2}(1 - z^2)\,\eta_{0\xi\xi} + \eta_1 , \tag{274}$$

and therefore,

$$\begin{aligned} w_{1z} &= -u_{1\xi} = -p_{1\xi} - u_{0\tau} - u_0 u_{0\xi} \\ &= -\left[\eta_{1\xi} + \tfrac{1}{2}(1 - z^2)\eta_{0\xi\xi\xi} + \eta_{0\tau} + \eta_0\eta_{0\xi}\right] . \end{aligned} \tag{275}$$

Finally, we find

$$w_1 = -\left[\eta_{1\xi} + \eta_{0\tau} + \eta_0 + \eta_{0\xi} + \tfrac{1}{2}\eta_{0\xi\xi\xi}\right] z + \tfrac{1}{6}\, z^3\, \eta_{0\xi\xi\xi} , \tag{276}$$

which satisfies the bottom boundary condition on $z = 0$.

The free surface boundary condition on $z = 1$ gives

$$(w_1)_{z=1} = -\left(\eta_{1\xi} + \eta_{0\tau} + \eta_0\eta_{0\xi} + \tfrac{1}{2}\,\eta_{0\xi\xi\xi}\right) + \tfrac{1}{6}\,\eta_{0\xi\xi\xi} = -\eta_{1\xi} + \eta_{0\tau} + 2\,\eta_0\eta_{0\xi}\,, \tag{277}$$

which yields the equation for $\eta_0(\xi, \tau)$ in the form

$$\eta_{0\tau} + \tfrac{3}{2}\,\eta_0\eta_{0\xi} + \tfrac{1}{6}\,\eta_{0\xi\xi\xi} = 0\,. \tag{278}$$

This is the *Korteweg–de Vries (KdV) equation*, which describes nonlinear plane gravity waves propagating in the x-direction. The exact solution of the general initial-value problem for the KdV equation can be obtained provided the initial data decay sufficiently rapidly as $|\xi| \to \infty$. We may raise the question of whether the asymptotic expansion for η (and hence, for the other flow variables) is uniformly valid as $|\xi| \to \infty$ and as $\tau \to \infty$. For the case of $\tau \to \infty$, this question is difficult to answer because the equations for $\eta_n (n \geq 1)$ are not easy to solve. However, all the available numerical evidence suggests that the asymptotic expansion of η is indeed uniformly valid as $\tau \to \infty$ (at least for solutions which satisfy $\eta \to 0$ as $|\xi| \to \infty$). From a physical point of view, if the waves are allowed to propagate indefinitely, then other physical effects cannot be neglected. In the case of real water waves, the most common physical effects include viscous damping. In practice, the viscous damping seems to be sufficiently weak to allow the dispersive and nonlinear effects to dominate before the waves eventually decay completely.

Two-Dimensional and Axisymmetric KdV Equations

It is well known that the KdV equation describes nonlinear plane waves which propagate only in the x-direction. However, there are many physical situations in which waves move on a two-dimensional surface. So it is natural to include both x- and y-directions with the appropriate balance of dispersion and nonlinearity. One of the simplest examples is the classical two-dimensional linear wave equation

$$u_{tt} = c^2(u_{xx} + u_{yy})\,, \tag{279}$$

which describes the propagation of long waves.

This equation has a solution in the form

$$u(\mathbf{x}, t) = a \exp\left[i(\omega t - \boldsymbol{\kappa} \cdot \mathbf{x})\right]\,, \tag{280}$$

where a is the wave amplitude, ω is the frequency, $\mathbf{x} = (x, y)$, and the wavenumber vector is $\boldsymbol{\kappa} = (k, \ell)$.

The dispersion relation is given by

$$\omega^2 = c^2(k^2 + \ell^2)\,. \tag{281}$$

For waves that propagate predominantly in the x-direction, the wavenumber component ℓ becomes small so that the approximate phase velocity is given by

$$c_p = c\left(1 + \frac{1}{2}\,\frac{\ell^2}{k^2}\right) \quad \text{as} \quad \ell \to 0\,. \tag{282}$$

It follows from (282) that the phase velocity suffers from a small correction provided by the wavenumber component ℓ in the y-direction. In order to ensure that this small correction is the same order as the dispersion and nonlinearity, it it necessary to require $\ell = O(\sqrt{\varepsilon})$ or $\ell^2 = O(\varepsilon)$. This requirement can be incorporated in the governing equations by a scaling of the flow variables as

$$y \to \sqrt{\varepsilon}\,y \qquad \text{and} \qquad v \to \sqrt{\varepsilon}\,v\,, \tag{283}$$

and using the same far-field transformations as (256).

Consequently, Eqs. (59)–(62) and the free surface condition (57ab) with the parameters δ replaced by ε reduce to the form

$$-u_\xi + \varepsilon(u_\tau + uu_\xi + \varepsilon v u_y + w u_z) = -p_\xi\,, \tag{284}$$

$$-v_\xi + \varepsilon(v_\tau + uv_\xi + \varepsilon v v_y + w v_z) = -p_y\,, \tag{285}$$

$$-\varepsilon\left[-w_\xi + \varepsilon(w_\tau + uw_\xi + \varepsilon v w_y + w w_z)\right] = -p_z\,, \tag{286}$$

$$-u_\xi + \varepsilon v_y + w_z = 0\,, \tag{287}$$

$$w = -\eta_\xi + \varepsilon(\eta_\tau + uu_\xi + \varepsilon v \eta_y) \quad \text{and} \quad p = \eta \quad \text{on} \quad z = 1 + \varepsilon\eta\,, \tag{288}$$

$$w = 0 \quad \text{on} \quad z = 0\,. \tag{289}$$

We seek the same asymptotic series solution (262) valid as $\varepsilon \to 0$ without any change of the leading order problem except that the flow variables involve y so that

$$p_0 = \eta_0\,, \quad u_0 = \eta_0, w_0 = -z\,\eta_{0\xi}\,, \quad 0 \leq z \leq 1\,, \tag{290}$$

$$v_{0\xi} = \eta_{0y}\,. \tag{291}$$

At the next order, the only difference from Sect. "Solutions of the KdV Equation: Solitons and Cnoidal Waves" is in the continuity equation which becomes

$$w_{1z} = -u_{1\xi} - v_{0y}\,. \tag{292}$$

This change leads to the following equation:

$$w_1 = -\left(\eta_{1\xi} + \eta_{0\tau} + \eta_0\eta_{0\xi} + \tfrac{1}{2}\,\eta_{0\xi\xi\xi} + v_{0y}\right) z + \tfrac{1}{6}\,z^3\eta_{0\xi\xi\xi}\,. \tag{293}$$

Using Eq. (291), the final result is the equation for the leading-order representation of the surface wave in the form

$$2\,\eta_{0\tau} + 3\,\eta_0\eta_{0\xi} + \tfrac{1}{3}\,\eta_{0\xi\xi} + v_{0y} = 0\,. \tag{294}$$

Consequently, differentiating (294) with respect to ξ and replacing $v_{0\xi}$ by η_{0y} give the evolution equation for $\eta_0(\xi,\tau,y)$ in the form

$$\left(2\,\eta_{0\tau} + 3\,\eta_0\eta_{0\xi} + \tfrac{1}{3}\,\eta_{0\xi\xi\xi}\right)_\xi + \eta_{0yy} = 0\,. \tag{295}$$

This is the *two-dimensional* or, more precisely, the $(1+2)$-*dimensional KdV equation*. Obviously, when there is no y-dependence, (295) reduces to the KdV Eq. (278). The two-dimensional KdV equation is also known as the *Kadomtsev–Petviashvili (KP) equation* (Kadomtsev and Petviashvili, [36]). This is another very special completely integrable equation, and it has an exact analytical solution that represents obliquely crossing nonlinear waves. Physically, any number of waves cross obliquely and interact nonlinearly. In particular, the nonlinear interaction of three waves leads to a *resonance phenomenon*. Such an interaction becomes more pronounced, leading to a strongly nonlinear interaction as the wave configuration is more nearly that of parallel waves. This situation can be interpreted as one in which the waves interact nonlinearly over a large distance so that a strong distortion is produced among these waves. The reader is referred to Johnson's [35] book for more detailed information on oblique interaction of waves.

We now consider the Euler Eqs. (63) and the continuity Eq. (65) in cylindrical polar coordinates (r,θ,z). It is convenient to use the nondimensional flow variables, parameters, and scaled variables similar to those defined by (53), (54), and (58) with $r = \ell r^*$ where r^* is a nondimensional variable so that the polar form of Euler Eqs. (66), continuity Eq. (65), the free surface boundary conditions (57ab) in nondimensional cylindrical polar form become

$$\frac{\mathrm{D}\,u}{\mathrm{D}\,t} - \frac{\varepsilon v^2}{r} = -\frac{\partial p}{\partial r}\,, \quad \frac{\mathrm{D}\,v}{\mathrm{D}\,t} + \frac{\varepsilon u v}{r} = -\frac{1}{r}\frac{\partial p}{\partial \theta}\,,$$
$$\delta\,\frac{\mathrm{D}\,w}{\mathrm{D}\,t} = -\frac{\partial p}{\partial z}\,, \tag{296}$$

$$\frac{1}{r}\frac{\partial}{\partial r}(ru) + \frac{1}{r}\frac{\partial v}{\partial \theta} + \frac{\partial w}{\partial z} = 0\,, \tag{297}$$

$$w = \eta_t + \varepsilon\left(u\,\eta_r + \frac{v}{r}\,\eta_\theta\right) \quad \text{and} \quad p = \eta \ \text{ on } \ z = 1 + \varepsilon\eta\,, \tag{298}$$

$$w = 0 \quad \text{on } \ z = 0\,, \tag{299}$$

where

$$\frac{\mathrm{D}}{\mathrm{D}\,t} = \frac{\partial}{\partial t} + \varepsilon\left(u\frac{\partial}{\partial r} + \frac{v}{r}\frac{\partial}{\partial \theta} + w\frac{\partial}{\partial z}\right)\,, \tag{300}$$

and ε and δ are defined by (52).

For the case of axisymmetric wave motions $\left(\frac{\partial}{\partial\theta} = 0\right)$, the governing equations become

$$u_t + \varepsilon(uu_r + wu_z) = -p_r\,, \tag{301}$$

$$\delta\left[w_t + \varepsilon(uw_r + ww_z)\right] = -p_z\,, \tag{302}$$

$$u_r + \frac{u}{r} + w_z = 0\,, \tag{303}$$

$$w = \eta_t + \varepsilon u\eta_r \quad \text{and} \quad p = \eta \ \text{ on } \ z = 1 + \varepsilon\eta\,, \tag{304}$$

$$w = 0 \quad \text{on } \ z = 0\,. \tag{305}$$

In the limit as $\varepsilon \to 0$, the linearized version of Eqs. (296)–(300) become

$$u_t = -p_r\,, \quad v_t = -\frac{1}{r}\,p_\theta\,, \quad \delta w_t = -p_z\,, \tag{306}$$

$$\frac{1}{r}\frac{\partial}{\partial r}(ru) + \frac{1}{r}\frac{\partial v}{\partial \theta} + \frac{\partial w}{\partial z} = 0\,, \tag{307}$$

$$w = \eta_t \quad \text{and} \quad p = \eta \ \text{ on } \ z = 1\,, \tag{308}$$

$$w = 0 \quad \text{on } \ z = 0\,, \tag{309}$$

For long waves ($\delta \to 0$), Eqs. (306)–(309) lead to the classical wave equation

$$\eta_{tt} = \eta_{rr} + \frac{1}{r}\,\eta_r + \frac{1}{r^2}\,\eta_{\theta\theta}\,. \tag{310}$$

For axisymmetric surface waves, the wave Eq. (310) becomes

$$\eta_{tt} = \eta_{rr} + \frac{1}{r}\,\eta_r\,. \tag{311}$$

This can be solved by using the Hankel transform (see Debnath and Bhatta [21]).

It is convenient to introduce the characteristic variable $\xi = r - t$ for outgoing waves and $R = \alpha r$ so that $\alpha \to 0$ which corresponds to large radius r. In other words, $R = O(1)$, $\alpha \to 0$ gives $r \to \infty$. The Eq. (311) reduces to the form

$$2\,\eta_{\xi R} + \frac{1}{R}\,\eta_\xi + \alpha\left(\eta_{RR} + \frac{1}{R}\,\eta_R\right) = 0\,. \tag{312}$$

In the limit as $\alpha \to 0$, it follows that

$$\sqrt{R}\,\eta_\xi = g(\xi)\,, \tag{313}$$

where $g(\xi)$ is an arbitrary function of ξ.

For outgoing waves, the correct solution takes the form for $\alpha \to 0 (r \to \infty)$,

$$\eta = \frac{1}{\sqrt{R}} \int g(\xi)\, d\xi = \frac{1}{\sqrt{R}} f(\xi)\,, \tag{314}$$

where $\eta = 0$ when $f = 0$.

This shows that the amplitude of axisymmetric waves decays as the radius $r \to \infty (R \to \infty)$. This behavior is totally different from the derivation of the KdV equation where the amplitude is uniformly $O(\varepsilon)$. In the present axisymmetric case, the amplitude decreases as the radius r increases so that there is no far-field region where the balance between nonlinearity and dispersion occurs. In other words, the amplitude is so small that nonlinear terms play no role at the leading order. However, there exists a scaling of the flow variables which leads to the *axisymmetric (concentric) KdV equation* as shown by Johnson [35].

We recall the axisymmetric Euler equations of motion and boundary conditions (301)–(305) and introduce scalings in terms of large radial variable R (see Johnson [35]),

$$\xi = \frac{\varepsilon^2}{\delta}(r - t) \quad \text{and} \quad R = \frac{\varepsilon^6}{\delta^2} r\,. \tag{315}$$

We next apply the transformations of the flow variables

$$(\eta, p, u, w) = \frac{\varepsilon^3}{\delta}\left(\eta^*, p^*, u^*, \frac{\varepsilon^2}{\delta} w^*\right)\,, \tag{316}$$

where large distance/time is measured by the scale (δ^2/ε^6) so that

$$\left(\frac{\delta^2}{\varepsilon^6}\right)^{-\frac{1}{2}} = \left(\frac{\varepsilon^3}{\delta}\right)\,,$$

which represents the scale of the amplitude of the waves. It is important to point out that the original wave amplitude parameter ε can now be interpreted based on the amplitude of the wave for $r = O(1)$ and $t = O(1)$. Consequently, the governing equations and the boundary conditions (301)–(305) become, dropping the asterisks in the variables,

$$-u_\xi + \alpha(u u_\xi + w u_z + \alpha\, u\, u_R) = -(p_\xi + \alpha\, p_R)\,, \tag{317}$$

$$\alpha\left[-w_\xi + \alpha(u w_\xi + w w_z + \alpha\, u\, w_R)\right] = -p_z\,, \tag{318}$$

$$u_\xi + w_z + \alpha\left(u_R + \frac{1}{R} u\right) = 0\,, \tag{319}$$

$$w = -\eta_\xi + \alpha(u \eta_\xi + \alpha\, u\, \eta_R)\,, \quad p = \eta \quad \text{on } z = 1 + \alpha\eta\,, \tag{320}$$

$$w = 0 \quad \text{on } z = 0\,, \tag{321}$$

where $\alpha = (\delta^{-1}\varepsilon^4)$ is a new parameter. These equations are similar in structure to those discussed above with parameter ε, which is now replaced by α in (317)–(321) so that the limit as $\alpha \to 0$ is required. This requirement is satisfied (for example $\varepsilon \to 0$ with δ fixed), and the scaling introduced by (315) describes the region where the appropriate balance occurs so that the wave amplitude in this region is $O(\alpha)$.

We now seek an asymptotic series solution in the form

$$q(\xi, R, z) = \sum_{n=0}^{\infty} \alpha^n\, q_n(\xi, R, z)\,, \quad \alpha \to 0\,, \tag{322}$$

where q represents each of η, u, w, and p.

In the leading order, we obtain the familiar equations

$$p_0 = \eta_0\,, \quad u_0 = \eta_0\,, \quad w_0 = -z\,\eta_{0\xi}\,, \quad 0 \le z \le 1\,. \tag{323}$$

It follows from the continuity Eq. (319) that

$$w_{1z} = -u_{1z} - \left(u_{0R} + \frac{1}{R} u_0\right)\,. \tag{324}$$

Without any more algebraic calculation, it turns out that $\eta_0(\xi, R)$ satisfies the nonlinear evolution equation

$$2\,\eta_{0R} + \frac{1}{R}\,\eta_0 + 3\,\eta_0 \eta_{0\xi} + \frac{1}{3}\,\eta_{0\xi\xi\xi} = 0\,. \tag{325}$$

This is usually referred to as the *axisymmetric (concentric) KdV equation* which includes a new term $R^{-1}\eta_0$. We may use the large time variable $\tau = (\delta^{-2}\varepsilon^6)t$ so that $R = \tau + \alpha\xi \approx \tau$ (see Johnson [35] or [34]).

Johnson [34] derived a new concentric KdV equation which incorporates weak dependence on the angular coordinate θ. In the derivation of KP Eq. (295), $\sqrt{\varepsilon}$ was chosen as the scaling parameter in the y-direction. Similarly, the appropriate scaling on the angular variable θ may be chosen as $\sqrt{\alpha}$. In the derivation of the concentric KdV Eq. (325), the parameter α plays the role of ε which is used as the small parameter in the asymptotic solution of the KdV equation.

Following the work of Johnson [34,35], we choose the variables ξ and R defined by (315) and the scaled θ variable as

$$\theta = \sqrt{\alpha}\,\theta^* = (\delta^{-1}\varepsilon^2)\,\theta^*\,, \tag{326}$$

which introduces a small angular deviation from purely concentric effects. We also use the scaled velocity component in the θ-direction as

$$v = (\delta^{-3/2}\varepsilon^5) v^*\,, \tag{327}$$

so that the scalings on $u = \phi_r$ and $v = \frac{1}{r}\,\phi_\theta$ are found to be consistent.

Using (315)–(316), Eqs. (296)–(300), and following Johnson [35], the equation in cylindrical polar coordinates can be obtained in the form:

$$\left(2\,\eta_{0R} + \frac{1}{R}\,\eta_0 + 3\,\eta_0\eta_{0\xi} + \frac{1}{3}\,\eta_{0\xi\xi\xi}\right)_\xi + \frac{1}{R^2}\,\eta_{0\theta\theta} = 0\,. \tag{328}$$

This is known as *Johnson's* (or *nearly concentric KdV*) *equation*, as it was first derived by Johnson [34]. In the absence of the θ-dependence, Eq. (328) reduces to the concentric KdV Eq. (325).

The Nonlinear Schrödinger Equation and Solitary Waves

We describe below that the nonlinear modulation of a quasi-monochromatic wave described by the nonlinear Schrödinger equation. To take into account the nonlinearity and the modulation in the far-field approximation, the wavenumber k and frequency ω in the linear dispersion relation are replaced by $k - i\frac{\partial}{\partial x}$ and $\omega + i\frac{\partial}{\partial t}$, respectively. It is convenient to use the nonlinear dispersion relation in the form

$$D\left(k - i\frac{\partial}{\partial x}\,,\quad \omega + i\frac{\partial}{\partial t}\,,\quad |A|^2\right)A = 0\,. \tag{329}$$

We consider the case of a weak nonlinearity and a slow variation of the amplitude, and hence, the amplitude A is assumed to be a slowly varying function of space and time. We next expand (329) with respect to $|A|^2$, $-i\frac{\partial}{\partial x}$, and $i\frac{\partial}{\partial t}$ to obtain

$$\begin{aligned} &D(k,\omega,0) - i\left(D_k\frac{\partial}{\partial x} - D_\omega\frac{\partial}{\partial t}\right)A \\ &- \frac{1}{2}\left(D_{kk}\frac{\partial^2}{\partial x^2} - 2D_{k\omega}\frac{\partial^2}{\partial x\partial t} + D_{\omega\omega}\frac{\partial^2}{\partial t^2}\right)A \\ &\qquad + \frac{\partial D}{\partial|A|^2}|A|^2A = 0\,, \end{aligned} \tag{330}$$

where the first term $D(k,\omega,0)$ corresponds to the linear dispersion so that $D(k,\omega,0) = 0$ represents the linear dispersion relation.

Introducing the transformation $x^* = x - C\,t$, $t^* = t$, assuming that $A = O(\varepsilon)$, $\frac{\partial}{\partial x^*} = O(\varepsilon)$, and $\frac{\partial}{\partial t^*} = O(\varepsilon^2)$, retaining all terms up to $O(\varepsilon^3)$, and dropping the asterisks, we find that

$$i\,\frac{\partial A}{\partial t} + p\,\frac{\partial^2 A}{\partial x^2} + q|A|^2A = 0\,, \tag{331}$$

where

$$p = \frac{1}{2}\left(\frac{\mathrm{d}C}{\mathrm{d}k}\right)\,,\quad q = \frac{1}{D_\omega}\left(\frac{\partial D}{\partial|A|^2}\right)\,, \tag{332}$$

and the following results

$$\begin{aligned} C &= -\frac{D\omega}{Dk}\,, \\ \frac{\mathrm{d}C}{\mathrm{d}k} &= (D_{kk} + 2\,C\,D_{\omega k} + C^2D_{\omega\omega})/D_\omega\,, \end{aligned} \tag{333}$$

have been used with D given by (329).

Equation (331) is known as the *cubic nonlinear Schrödinger (NLS) equation*. When the last cubic nonlinear term is neglected, Eq. (331) reduces to the corresponding *linear Schrödinger (NLS) equation* for $A(x,t)$ in the form

$$i\,\frac{\partial A}{\partial t} + \frac{1}{2}\,\omega''(k)\frac{\partial^2 A}{\partial x^2} = 0\,, \tag{334}$$

when $p = \frac{1}{2}\,C'(k) = \frac{1}{2}\,\omega''(k)$ is used.

More explicitly, if the nonlinear dispersion relation is given by

$$\omega = \omega(k, a^2)\,, \tag{335}$$

and if we expand ω in a Taylor series about $k = k_0$ and $|a|^2 = 0$, we obtain

$$\begin{aligned} \omega \approx\ &\omega_0 + (k-k_0)\omega_0' \\ &+ \frac{1}{2}(k-k_0)^2\omega_0'' + \left(\frac{\partial\omega}{\partial|a|^2}\right)_{|a|^2=0}|a|^2\,. \end{aligned} \tag{336}$$

Replacing $(\omega - \omega_0)$ by $i\left(\frac{\partial}{\partial t}\right)$, $k - k_0$ by $-i\left(\frac{\partial}{\partial x}\right)$, and assuming that the resulting operators act on the amplitude function $a(x,t)$, it turns out that $a(x,t)$ satisfies the equation

$$i(a_t + \omega_0'a_x) + \frac{1}{2}\,\omega_0''\,a_{xx} + \gamma|a|^2a = 0\,, \tag{337}$$

where

$$\gamma = -\left(\frac{\partial\omega}{\partial|a|^2}\right)_{|a|^2=0} = \text{constant}\,. \tag{338}$$

Equation (337) is known as the *cubic nonlinear Schrödinger equation*. If we choose a frame of reference moving with the linear group velocity ω_0', that is, $\xi = x - \omega_0't$ and $\tau = t$, the term involving a_x will drop out from (337), and therefore, the amplitude $a(x,t)$ satisfies the normalized *nonlinear Schrödinger equation*

$$i\,a_\tau + \tfrac{1}{2}\,\omega_0''\,a_{\xi\xi} + \gamma|a|^2a = 0\,. \tag{339}$$

The corresponding dispersion relation is given by

$$\omega = \tfrac{1}{2}\,\omega_0''\,k^2 - \gamma\,a^2\,. \tag{340}$$

According to the stability criterion established by Whitham [62], the wave modulation is stable if $\gamma\omega_0'' < 0$ or unstable if $\gamma\omega_0'' > 0$.

To study the solitary wave solution, it is convenient to use the NLS equation in the standard form

$$i\,\psi_t + \psi_{xx} + \gamma|\psi|^2\psi = 0\,, \qquad -\infty < x < \infty\,, \quad t > 0\,. \tag{341}$$

We then seek waves of permanent form by assuming the solution

$$\psi = f(X)\mathrm{e}^{i(mX-nt)}\,, \quad X = x - Ut\,, \tag{342}$$

for some functions f and constant wave speed U to be determined, and where m, n are constants.

Substitution of (342) in (341) gives

$$f'' + i(2m-U)f' + (n-m^2)f + \gamma|f|^2 f = 0\,. \tag{343}$$

We eliminate f' by setting $2m - U = 0$, and then write $n = m^2 - \alpha$, so that f can be assumed real. Thus, Eq. (343) becomes

$$f'' - \alpha f + \gamma f^3 = 0\,. \tag{344}$$

Multiplying this equation by $2f'$ and integrating, we find that

$$f'^2 = A + \alpha f^2 - \frac{\gamma}{2}f^4 = F(f)\,, \tag{345}$$

where $F(f) = (\alpha_1 - \alpha_2 f^2)(\beta_1 - \beta_2 f^2)$, so that $\alpha = -(\alpha_1\beta_2 + \alpha_2\beta_1)$, $A = \alpha_1\beta_1$, $\gamma = -2(\alpha_2\beta_2)$, and the α's and β's are assumed to be real and distinct.

Evidently, it follows from (345) that

$$X = \int_0^f \frac{\mathrm{d}f}{\sqrt{(\alpha_1 - \alpha_2 f^2)(\beta_1 - \beta_2 f^2)}}\,. \tag{346}$$

Setting $(\alpha_2/\alpha_1)^{1/2} f = u$ in this integral (346), we deduce the following elliptic integral of the first kind (see Dutta and Debnath [22] and Helal and Molines [28]):

$$\sigma X = \int_0^f \frac{\mathrm{d}f}{\sqrt{(1-u^2)(1-\kappa^2 u^2)}}\,, \tag{347}$$

where $\sigma = (\alpha_2\beta_1)^{1/2}$ and $\kappa = (\alpha_1\beta_2)/(\beta_1\alpha_2)$.

Thus, the final solution can be expressed in terms of the Jacobian elliptic function

$$u = sn(\sigma X, \kappa)\,. \tag{348a}$$

Thus, the solution for $f(X)$ is given by

$$f(X) = \left(\frac{\alpha_1}{\alpha_2}\right)^{1/2} sn(\sigma X, \kappa)\,. \tag{348b}$$

In particular, when $A = 0$, $\alpha > 0$, and $\gamma > 0$, we obtain a solitary wave solution. In this case, (345) can be rewritten

$$\sqrt{\alpha}X = \int_0^f \left\{f^2\left(1 - \frac{\gamma}{2\alpha}f^2\right)\right\}^{-\frac{1}{2}} \mathrm{d}f\,. \tag{349}$$

Substitution of $(\gamma/2\alpha)^{1/2} f = \operatorname{sech}\theta$ in this integral gives the exact solution

$$f(X) = \left(\frac{2\alpha}{\gamma}\right)^{1/2} \operatorname{sech}\left[\sqrt{\alpha}(x - Ut)\right]\,. \tag{350}$$

This represents a *solitary wave* solution that propagates without change of shape with constant velocity U. Unlike the solitary wave solution of the KdV equation, the amplitude and the velocity of the wave are independent parameters. It is noted that the solitary wave exists only for the unstable case ($\gamma > 0$). This means that small modulations of the unstable wavetrain lead to a series of solitary waves.

The well-known nonlinear dispersion relation (116) for deep water waves is

$$\omega = \sqrt{gk}\left(1 + \frac{1}{2}a^2k^2\right)\,. \tag{351}$$

Therefore,

$$\omega_0' = \frac{\omega_0}{2k_0}\,, \quad \omega_0'' = \frac{\omega_0}{4k_0^2}\,, \quad \text{and } \gamma = -\tfrac{1}{2}\omega_0 k_0^2\,, \tag{352}$$

and the NLS equation for deep water waves is obtained from (337) in the form

$$i\left(a_t + \frac{\omega_0}{2k_0}a_x\right) - \frac{\omega_0}{8k_0^2}a_{xx} - \frac{1}{2}\omega_0 k_0^2|a|^2 a = 0\,. \tag{353}$$

The normalized form of this equation in a frame of reference moving with the linear group velocity ω_0' is

$$i\,a_t - \left(\frac{\omega_0}{8k_0^2}\right)a_{xx} - \frac{1}{2}\omega_0 k_0^2|a|^2 a = 0\,. \tag{354}$$

Since $\gamma\omega_0'' = (\omega_0^2/8) > 0$, this equation confirms the instability of deep water waves. This is one of the most remarkable recent results in the theory of water waves.

We next discuss the uniform solution and the solitary wave solution of the NLS Eq. (354). We look for solutions in the form

$$a(x,t) = A(X)\exp(i\gamma^2 t)\,, \quad X = x - \omega_0' t\,, \tag{355}$$

and substitute this in Eq. (354) to obtain the following equation:

$$A_{XX} = \frac{8k_0^2}{\omega_0}\left(\gamma^2 A + \frac{1}{2}\omega_0 k_0^2 A^3\right)\,. \tag{356}$$

We multiply this equation by $2A_X$ and, then, integrate to find

$$A_X^2 = -\left(A_0^4 m'^2 + \frac{8}{\omega_0}\gamma^2 k_0^2 A^2 + 2k_0^4 A^4\right) = (A_0^2 - A^2)(A^2 - m'^2 A_0^3)\,, \quad (357)$$

where $A_0^4 m'^2$ is an integrating constant, $2k_0^4 = 1$, $m'^2 = 1 - m^2$, and $A_0^2 = 4\gamma^2/\omega_0 k_0^2(m^2 - 2)$, which relates A_0, γ, and m.

Finally, we rewrite Eq. (357) in the form

$$A_0^2\,\mathrm{d}X = \frac{\mathrm{d}A}{\left[\left(1 - \frac{A^2}{A_0^2}\right)\left(\frac{A^2}{A_0^2} - m'^2\right)\right]^{1/2}}\,, \quad (358a)$$

or, equivalently,

$$A_0(X - X_0) = \int^t \frac{\mathrm{d}s}{\left[(1 - s^2)(s^2 - m'^2)\right]^{1/2}}\,, \quad s = (A/A_0)\,. \quad (358b)$$

This can readily be expressed in terms of the Jacobi *dn* function (see Dutta and Debnath [22] and Helal and Molines [28]):

$$A = A_0\, dn\left[A_0(X - X_0),\ m\right]\,, \quad (359)$$

where m is the modulus of the *dn* function.

In the limit, $m \to 0$, $dn\, z \to 1$, and $\gamma^2 \to -\frac{1}{2}\omega_0 k_0^2 A_0^2$. Hence, the solution becomes

$$a(x,t) = A(t) = A_0 \exp\left(-\tfrac{1}{2} i\, \omega_0 k_0^2 A_0^2\, t\right)\,. \quad (360)$$

On the other hand, when $m \to 1$, $dn\, z \to \operatorname{sech} z$, and $\gamma^2 \to -\frac{1}{4}\omega_0 k_0^2 A_0^2$. Therefore, the solitary wave solution is

$$a(x,t) = A_0 \exp\left(-\frac{i}{4}\omega_0 k_0^2 A_0^2\, t\right) \cdot \operatorname{sech}\left[A_0(x - \omega_0' t - X_0)\right]\,. \quad (361)$$

An analysis of this section reveals several remarkable features of the nonlinear Schrödinger equation. Like the KdV equation, the nonlinear Schrödinger equation is the lowest-order approximation for weakly and strongly nonlinear dispersive waves system. This equation can also be used to investigate instability phenomena in many other physical systems. Like the various forms of the KdV equation, the NLS equation arises in many physical problems, including nonlinear water waves and ocean waves, waves in plasma, propagation of heat pulses in a solid, self-trapping phenomena in nonlinear optics, nonlinear waves in a fluid-filled viscoelastic tube, and various nonlinear instability phenomena in fluids and plasmas.

Whitham's Equations of Nonlinear Dispersive Waves

To describe a slowly varying nonlinear and nonuniform oscillatory wavetrain in a dispersive medium, we assume the existence of a one-dimensional solution in the form (see Whitham [62] or Debnath [20]), so that

$$\phi(x,t) = a(x,t)\exp\{i\theta(x,t)\} + c.c.\,, \quad (362)$$

where *c.c.* stands for the complex conjugate and $a(x,t)$ is the complex amplitude (see Debnath [20]). The phase function $\theta\,(x,t)$ is given by

$$\theta(x,t) = xk(x,t) - t\omega(x,t)\,, \quad (363)$$

and k, ω, and a are slowly varying functions of space variable x and time t.

Because of the slow variations of k and ω, it is reasonable to assume that these quantities still satisfy the linear Dispersion relation of the form

$$\omega = W(k)\,. \quad (364)$$

Differentiating (363) with respect to x and t, respectively, we obtain

$$\theta_x = k + \{x - t\,W'(k)\}\,k_x\,, \quad (365)$$

$$\theta_t = -W(k) + \{x - t\,W'(k)\}\,k_t\,. \quad (366)$$

In the neighborhood of stationary points defined by $W'(k) = (x/t) > 0$, these equations become

$$\theta_x = k(x,t) \quad \text{and} \quad \theta_t = -\omega(x,t)\,. \quad (367ab)$$

These results can be used as a definition of *local wavenumber* and *local frequency* of a slowly varying nonlinear wavetrain.

In view of (367), relation (364) gives a nonlinear partial differential equation for the phase $\theta(x,t)$ in the form

$$\frac{\partial\theta}{\partial t} + W\left(\frac{\partial\theta}{\partial x}\right) = 0\,. \quad (368)$$

The solution of this equation determines the geometry of the wave pattern.

We eliminate θ from (367ab) to obtain the equation

$$\frac{\partial k}{\partial t} + \frac{\partial\omega}{\partial x} = 0\,. \quad (369)$$

This is known as the *Whitham equation* for the conservation of waves, where k represents the *density* of waves and ω is the *flux* of waves.

Using the dispersion relation (364), Eq. (369) gives

$$\frac{\partial k}{\partial t} + C(k)\frac{\partial k}{\partial x} = 0\,, \tag{370}$$

where $C(k) = W'(k)$ is the group velocity. This represents the simplest nonlinear wave (hyperbolic) equation for the propagation of wavenumber k with group velocity $C(k)$.

Equations (370) and (364) reveal that ω also satisfies the first-order, nonlinear wave (hyperbolic) equation

$$\frac{\partial \omega}{\partial t} + W'(k)\frac{\partial \omega}{\partial x} = 0\,. \tag{371}$$

It follows from Eqs. (370) and (371) that both k and ω remain constant on the characteristic curves defined by

$$\frac{\mathrm{d}x}{\mathrm{d}t} = W'(k) = C(k) \tag{372}$$

in the (x, t)-plane. Since k or ω is constant on each characteristic, the characteristics are straight lines with slope $C(k)$.

Finally, it follows from the preceding analysis that any constant value of the phase θ propagates according to $\theta(x, t) = \text{constant}$, and hence,

$$\theta_t + \left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)\theta_x = 0\,, \tag{373}$$

which gives, by (367ab),

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\frac{\theta_t}{\theta_x} = \frac{\omega}{k} = c(k)\,. \tag{374}$$

Thus, the phase of the waves propagates with the phase speed $c(k)$. On the other hand, Eq. (370) ensures that the wavenumber k propagates with group velocity $C(k) = (\mathrm{d}\omega/\mathrm{d}k) = W'(k)$.

We next follow Whitham [62] or Debnath [20] to investigate how wave energy propagates in a dispersive medium. We consider the following integral involving the square of the wave amplitude (energy) between any two points $x = x_1$ and $x = x_2 (0 < x_1 < x_2)$:

$$Q(t) = \int_{x_1}^{x_2} |A|^2 \mathrm{d}x = \int_{x_1}^{x_2} AA^* \mathrm{d}x \tag{375}$$

$$= 2\pi \int_{x_1}^{x_2} \frac{F(k)F^*(k)}{t|W''(k)|}\mathrm{d}x\,, \tag{376}$$

which, due to a change of variable $x = tW'(k)$,

$$= 2\pi \int_{k_1}^{k_2} F(k)F^*(k)\mathrm{d}k\,, \tag{377}$$

where $x_r = t\,W'(k_r)$, $r = 1, 2$.

When k_r is kept fixed as t varies, $Q(t)$ remains constant so that

$$0 = \frac{\mathrm{d}Q}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\int_{x_1}^{x_2} |A|^2 \mathrm{d}x$$

$$= \int_{x_1}^{x_2} \frac{\partial}{\partial t}|A|^2 \mathrm{d}x + |A|_2^2 W'(k_2) - |A|_1^2 W'(k_1)\,. \tag{378}$$

In the limit, as $x_2 - x_1 \to 0$, this result reduces to the first-order, partial differential equation

$$\frac{\partial}{\partial t}|A|^2 + \frac{\partial}{\partial x}\left[W'(k)|A|^2\right] = 0\,. \tag{379}$$

This represents the equation for the conservation of wave energy where $|A|^2$ and $|A|^2 W'(k)$ are the energy density and energy flux, respectively. It also follows that the energy propagates with group velocity $W'(k)$. It has been shown that the wavenumber k also propagates with the group velocity. Thus, the group velocity plays a double role.

The preceding analysis reveals another important fact that (364), (369), and (379) constitute a closed set of equations for the three functions k, ω, and A. Indeed, these are the fundamental equations for nonlinear dispersive waves and are known as *Whitham's equations*.

In his pioneering work on nonlinear dispersive waves, Whitham [62] formulated a more general energy equation based on the amount of energy $Q(t)$ between two points x_1 and x_2

$$Q(t) = \int_{x_1}^{x_2} g(k)A^2\,\mathrm{d}x\,, \tag{380}$$

where $g(k)$ is an arbitrary proportionality factor associated with the square of the amplitude and energy.

In a new coordinate system moving with the group velocity, that is, along the rays $x = C(k)t$, Eq. (380) reduces to the form

$$Q(t) = 2\pi \int_{x_1}^{x_2} g(k)F(k)F^*(k)\mathrm{d}k\,, \tag{381}$$

where $\omega = \Omega(k)$ and $\Omega''(k) > 0$ and k_1 and k_2 are defined by $x_1 = C(k_1)t$ and $x_2 = C(k_2)t$, respectively.

Using the principle of conservation of energy, that is, stating that the energy between the points x_1 and x_2 traveling with the group velocities $C(k_1)$ and $C(k_2)$ remains invariant, it turns out from (380) that

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = \int_{x_1}^{x_2} \frac{\partial}{\partial t}\left\{g(k)A^2\right\}\mathrm{d}x + g(k_2)C(k_2)A^2(x_2, t) - g(k_1)C(k_1)A^2(x_1, t) = 0\,, \tag{382}$$

which is, in the limit as $x_2 - x_1 \to 0$,

$$\frac{\partial}{\partial t}\left\{g(k)A^2\right\} + \frac{\partial}{\partial x}\left\{g(k)C(k)A^2\right\} = 0\,. \tag{383}$$

This may be treated as the more general energy equation. Quantities $g(k)A^2$ and $g(k)C(k)A^2$ represent the *energy density* and the *energy flux*, so that they are proportional to $|A|^2$ and $C(k)|A|^2$, respectively. The flux of energy propagates with the group velocity $C(k)$. Hence, the group velocity has a double role: it is the propagation velocity for the wavenumber k and for the energy $g(k)|A|^2$.

Thus, (369) and (383) are known as *Whitham's conservation equations* for nonlinear dispersive waves. The former represents the conservation of wave-number k and the latter is the conservation of energy (or more generally, the conservation of wave action). Whitham also derived the conservation equations more rigorously from a general and extremely powerful approach that is now known as *Whitham's averaged variational principle*.

For slowly varying dispersive wavetrains, the solution maintains the elementary form $u = \Phi(\theta, a)$, but $\omega, \boldsymbol{\kappa}$, and a are no longer constants, so that θ is not a linear function of x_i and t. The local wavenumber and local frequency are defined by

$$k_i = \frac{\partial \theta}{\partial x_i}\,, \quad \omega = -\frac{\partial \theta}{\partial t}\,. \tag{384ab}$$

The *Whitham averaged Lagrangian* over the phase of the integral of the Lagrangian L is defined by

$$\mathcal{L}(\omega, \boldsymbol{\kappa}, a, \mathbf{x}, t) = \frac{1}{2\pi}\int_0^{2\pi} L\,\mathrm{d}\theta \tag{385}$$

and is calculated by assuming the uniform periodic solution $u = \Phi(\theta, a)$ in L. Whitham first formulated the *averaged variational principle* in the form

$$\delta \iint \mathcal{L}\,\mathrm{d}\mathbf{x}\,\mathrm{d}t = 0\,, \tag{386}$$

to derive the equations for $\omega, \boldsymbol{\kappa}$, and a.

It is noted that the dependence of $\mathcal{L}$ on $\mathbf{x}$ and t reflects possible nonuniformity of the medium supporting the wave motion. In a uniform medium, $\mathcal{L}$ is independent of $\mathbf{x}$ and t, so that the Whitham function $\mathcal{L} = \mathcal{L}(\omega, \boldsymbol{\kappa}, a)$. However, in a uniform medium, some additional variables also appear only through their derivatives. They represent potentials whose derivatives are important physical quantities.

The Euler equations resulting from the independent variations of δa and $\delta\theta$ in (386) with $\mathcal{L} = \mathcal{L}(\omega, \boldsymbol{\kappa}, a)$ are

$$\delta a : \mathcal{L}_a(\omega, \boldsymbol{\kappa}, a) = 0\,, \tag{387}$$

$$\delta\theta : \frac{\partial}{\partial t}\mathcal{L}_\omega - \frac{\partial}{\partial x_i}\mathcal{L}_{k_i} = 0\,. \tag{388}$$

The θ-eliminant of (384ab) gives the consistency equations

$$\frac{\partial k_i}{\partial t} + \frac{\partial \omega}{\partial x_i} = 0\,, \quad \frac{\partial k_i}{\partial x_j} - \frac{\partial k_j}{\partial x_i} = 0\,. \tag{389ab}$$

Thus, (387)–(389) represent the *Whitham equations* for describing the slowly varying wavetrain in a nonuniform medium and constitute a closed set from which the triad $\omega, \boldsymbol{\kappa}$, and a can be determined.

In linear problems, the Lagrangian L, in general, is a quadratic in u and its derivatives. Hence, if $\Phi(\theta) = a\cos\theta$ is substituted in (385), $\mathcal{L}$ must always take the form

$$\mathcal{L}(\omega, \boldsymbol{\kappa}, a) = D(\omega, \boldsymbol{\kappa})a^2\,, \tag{390}$$

so that the dispersion relation ($\mathcal{L}_a = 0$) must take the form

$$D(\omega, \boldsymbol{\kappa}) = 0\,. \tag{391}$$

We note that the stationary value of $\mathcal{L}$ is, in fact, zero for linear problems. In the simple case, L equals the difference between kinetic and potential energy. This proves the well-known principle of equipartition of energy, stating that average potential and kinetic energies are equal.

Whitham's Instability Analysis of Water Waves

Section "The Nonlinear Schrödinger Equation and Solitary Waves" deals with Whitham's new remarkable variational approach to the theory of slowly varying, nonlinear, dispersive waves. Based upon Whitham's ideas and, especially, Whitham's fundamental dispersion Eq. (388), Lighthill [44,45] developed an elegant and remarkably simple result determining whether very gradual – not necessarily small – variations in the properties of a wavetrain are governed by hyperbolic or elliptic partial differential equations. A general account of Lighthill's work with special reference to the instability of wavetrains on deep water was described by Debnath [19]. This section is devoted to the Whitham instability theory with applications to water waves.

According to Whitham's nonlinear dispersive wave theory $L_a = 0$ gives a dispersion relation that depends on wave amplitude a has the form

$$\omega = \omega(k, a)\,, \tag{392}$$

where equations for k and a are no longer uncoupled and constitute a system of partial differential equations. The first important question is whether these equations are hyperbolic or elliptic. This can be answered by a standard and simple method combined with Whitham's conservation Eqs. (389a) and (383). For moderately small amplitudes, we use the Stokes expansion of ω in terms of k and a^2 in the form

$$\omega(k) = \omega_0(k) + \omega_2(k)a^2 + \ldots\,. \tag{393}$$

We substitute this result in (389a) and (383), replace $\omega'(k)$ by its linear value $\omega_0'(k)$, and retain the terms of order a^2 to obtain the equations for k and a^2 in the form

$$\frac{\partial k}{\partial t} + \left[\omega'(k) + \omega_2'(k)a^2\right]\frac{\partial k}{\partial x} + \omega_2(k)\frac{\partial a^2}{\partial x} = 0\,, \quad (394)$$

$$\frac{\partial a^2}{\partial t} + \omega_0'(k)\frac{\partial a^2}{\partial x} + \omega_0''(k)a^2\frac{\partial k}{\partial x} = 0\,. \quad (395)$$

Neglecting the term $O(a^2)$, these equations can be rewritten as

$$\frac{\partial k}{\partial t} + \omega_0'\frac{\partial k}{\partial x} + \omega_2\frac{\partial a^2}{\partial x} = 0\,, \quad (396)$$

$$\frac{\partial a^2}{\partial t} + \omega_0'(k)\frac{\partial a^2}{\partial x} + \omega_0''a^2\frac{\partial k}{\partial x} = 0\,. \quad (397)$$

These describe the modulations of a linear dispersive wavetrain and represent a coupled system due to the nonlinear dispersion relation (393) exhibiting the dependence of ω on both k and a. In matrix form, these equations read

$$\begin{pmatrix} \omega_0' & \omega_2 \\ \omega_0''a^2 & \omega_0' \end{pmatrix}\begin{pmatrix} \frac{\partial k}{\partial x} \\ \frac{\partial a^2}{\partial x} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \frac{\partial k}{\partial t} \\ \frac{\partial a^2}{\partial t} \end{pmatrix} = 0\,. \quad (398)$$

Hence, the eigenvalues λ are the roots of the determinant equation

$$|a_{ij} - \lambda b_{ij}| = \begin{vmatrix} \omega_0' - \lambda & \omega_2 \\ \omega_0''a^2 & \omega_0' - \lambda \end{vmatrix} = 0\,, \quad (399)$$

where a_{ij} and b_{ij} are the coefficient matrices of (398). This determinant equation gives the characteristic velocities

$$\begin{aligned} \lambda &= \left(\frac{\mathrm{d}x}{\mathrm{d}t}\right) = C(k) \\ &= C_0(k) \pm a\left[\omega_2(k)\omega_0''(k)\right]^{1/2} + O(a^2)\,, \end{aligned} \quad (400ab)$$

where $C_0(k) = \omega_0'(k)$ is the linear group velocity, and, in general, $\omega_0''(k) \neq 0$ for dispersive waves. The equations are hyperbolic or elliptic depending on whether $\omega_2(k)\,\omega_0''(k) > 0$ or < 0.

In the hyperbolic case, the characteristics are real, and the double characteristic velocity splits into two separate velocities and provides a generalization of the group velocity of nonlinear dispersive waves. In fact, the characteristic velocities (400ab) are used to define the *nonlinear group velocities*. The splitting of the double characteristic velocity into two different velocities is one of the most significant results of the Whitham theory. This means that any initial disturbance of finite extent will eventually split into two separate disturbances. This prediction is radically different from that of the linearized theory, where an initial disturbance may suffer from a distortion due to dependence of the linear group velocity $C_0(k) = \omega_0'(k)$ on the wavenumber k, but would *never split* into two. Another significant consequence of nonlinearity in the hyperbolic case is that compressive modulations will suffer from gradual distortion and steepening in the typical hyperbolic manner discussed earlier. This leads to the multiple-valued solutions and hence, eventually, breaking of waves.

In the elliptic case ($\omega_2, \omega_0'' < 0$), the characteristics are imaginary. This leads to *ill-posed problems* in the theory of nonlinear wave propagation. Any small sinusoidal disturbances in a and k may be represented by solutions of the form $\exp\left[ia\{x - C(k)t\}\right]$, where $C(k)$ is given by (400ab) for the unperturbed values of a and k. Thus, when $C(k)$ is complex, these disturbances will grow exponentially in time, and hence, the periodic wavetrains become definitely *unstable*.

An application of this analysis to the Stokes waves on deep water reveals that the associated dispersion equation is elliptic in this case. For waves on deep water, the dispersion relation is

$$\omega = \sqrt{gk}\left(1 + \frac{1}{2}a^2k^2\right) + O(a^4)\,. \quad (401)$$

This result is compared with the Stokes expansion (393) to give $\omega_0(k) = \sqrt{gk}$ and $\omega_2(k) = \frac{1}{2}k^2\sqrt{gk}$. Hence, $\omega_0''\omega_2 < 0$, the velocities (400ab) are complex, and the Stokes waves on deep water are definitely *unstable*. The instability of deep water waves came as a real surprise to researchers in the field in view of the very long and controversial history of the subject. The question of instability went unrecognized for a long period of time, even though special efforts have been made to prove the existence of a permanent shape for periodic water waves for all amplitudes less than the critical value at which the waves assume a sharp-crested form. However, Lighthill's [45] theoretical investigation into the elliptic case and Benjamin and Feir's [6] theoretical and experimental findings have provided conclusive evidence of the instability of Stokes waves on deep water.

For more details of nonlinear dispersive wave phenomena, the reader is also referred to Debnath [19].

In his pioneering work on nonlinear water waves, Whitham [62] observed that the neglect of dispersion in the nonlinear shallow water equations leads to the development of multivalued solutions with a vertical slope, and hence, eventually breaking occurs. It seems clear that the third derivative dispersive term in the KdV equation produces the periodic and solitary waves which are not found in the shallow water theory. However, the KdV equation cannot describe the observed symmetrical peaking of the

crests with a finite angle. On the other hand, the Stokes waves include full effects of dispersion, but they are limited to small amplitude, and describe neither the solitary waves nor the peaking phenomenon.

Although both breaking and peaking are without doubt involved in the governing equations of the exact potential theory, Whitham [62] developed a mathematical equation that can include all these phenomena. It has been shown earlier that the breaking of shallow water waves is governed by the nonlinear equation

$$\eta_t + c_0\eta_x + d\,\eta\eta_x = 0\,, \quad d = 3c_0(2h_0)^{-1}\,. \tag{402}$$

On the other hand, the linear equation corresponding to a general linear dispersion relation

$$\frac{\omega}{k} = c(k) \tag{403}$$

is given by the integrodifferential equation in the form

$$\eta_t + \int_{-\infty}^{\infty} K(x-s)\eta_s(s,t)\mathrm{d}s = 0\,, \tag{404}$$

where the kernel K is given by the inverse Fourier transform of $c(k)$:

$$K(x) = \mathcal{F}^{-1}\{c(k)\} = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{ikx}\,c(k)\mathrm{d}k\,. \tag{405}$$

Whitham combined the above ideas to formulate a new class of *nonlinear nonlocal* equations

$$\eta_t + d\,\eta\eta_x + \int_{-\infty}^{\infty} K(x-s)\eta_s(s,t)\mathrm{d}s = 0\,. \tag{406}$$

This is well known as the *Whitham equation*, which can, indeed, describe symmetric waves that propagate without change of shape and peak at a critical height, as well as asymmetric waves that invariably break.

Once a wave breaks, it usually continues to travel in the form of a bore as observed in tidal waves. The weak bores have a smooth but oscillatory structure, whereas the strong bores have a structure similar to turbulence with no coherent oscillations. Since the region where waves break is a zone of high energy dissipation, it is natural to include a second derivative dissipative term in the KdV equation to obtain

$$\eta_t + c_0\,\eta_x + d\,\eta\eta_x + \mu\,\eta_{xxx} - \nu\,\eta_{xx} = 0\,, \tag{407}$$

where $\mu = \frac{1}{6}c_0h_0^2$. This is known as the *KdV–Burgers equation*, which also arises in nonlinear acoustics for fluids with gas bubbles (see [38]).

Whitham's [62] equations for the slow modulation of the wave amplitude a and the wavenumber k in the case of two-dimensional deep water waves are given by

$$\frac{\partial}{\partial t}\left(\tfrac{a^2}{\omega_0}\right) + \frac{\partial}{\partial x}\left(C\frac{a^2}{\omega_0}\right) = 0\,, \tag{408}$$

$$\frac{\partial k}{\partial t} + \frac{\partial \omega}{\partial x} = 0\,, \tag{409}$$

where $\omega_0 = \sqrt{gk}$ is the first-order approximation for the wave frequency $\omega(k)$ and $C = (g/2\omega_0)$ is the group velocity.

Chu and Mei [14,15] observed that certain terms of the dispersive type, neglected in Whitham's equations to the same order of approximation, must be included to extend the validity of these equations. Whitham's theory is based on the direct use of Stokes' dispersion relation for a uniform wavetrain,

$$\omega = \omega_0\left(1 + \tfrac{1}{2}\varepsilon^2a^2k^2\right)\,, \tag{410}$$

whereas Chu and Mei added terms of higher derivatives and dispersive type to the expression for ω, so that

$$\omega = \omega_0\left[1 + \varepsilon^2\left(\frac{1}{2}a^2k^2 + \left\{\left(\frac{a}{\omega_0}\right)_{tt} \div 2\omega_0 a\right\}\right)\right]\,. \tag{411}$$

They used the expression (411) to transform (408) and (409) in a frame of reference moving with the group velocity $C(k)$ and obtained the following nondimensional equations:

$$\frac{\partial a^2}{\partial t} + \frac{\partial}{\partial x}(a^2\phi_x) = 0\,, \tag{412}$$

$$-2\frac{\partial^2\phi}{\partial x\partial t} + \frac{\partial}{\partial x}\left[-\phi_x^2 + \frac{a^2}{4} + \frac{a_{xx}}{16a}\right] = 0\,, \tag{413}$$

where we have used Chu and Mei's result $W = -2\phi_x$ and ϕ is a small phase variation. Integrating (413) with respect to x and setting the constant of integration to be zero gives

$$\phi_t + \frac{1}{2}\phi_x^2 - \frac{1}{8}a^2 - \frac{a_{xx}}{32a} = 0\,. \tag{414}$$

A transformation $\Psi = a\exp(4i\phi)$ is used to simplify (412) and (414), which reduces to the *nonlinear Schrödinger equation*

$$i\,\Psi_t + \tfrac{1}{8}\Psi_{xx} + \tfrac{1}{2}\Psi|\Psi|^2 = 0\,. \tag{415}$$

This equation has also been derived and exploited by several authors, including Benney and Roskes [7], Hasimoto and Ono [27], and Davey and Stewartson [17] to examine the nonlinear evolution of Stokes' waves on water.

An alternative way to study nonlinear evolution of two- and three-dimensional wavepackets is to use the method of multiple scales in which the small parameter ε is explicitly built into the expansion procedure. The small parameter ε characterizes the wave steepness. This method has been employed by several authors in various fields and has also been used by Hasimoto and Ono [27] and Davey and Stewartson [17].

The Benjamin–Feir Instability of the Stokes Water Waves

One of the simplest solutions of the nonlinear Schrödinger Eq. (354) is given by (360), that is,

$$A(t) = A_0 \exp\left(-\tfrac{1}{2} i\omega_0 k_0^2 A_0^2\, t\right) , \tag{416}$$

where A_0 is a constant. This essentially represents the fundamental component of the Stokes wave. We consider a perturbation of (416) and express it in the form

$$a(x,t) = A(t)\left[1 + B(x,t)\right] , \tag{417}$$

where $B(x,t)$ is the perturbation function. Substituting this result in (354) gives

$$\begin{aligned} &i(1+B)A_t + iAB_t - \left(\frac{\omega_0}{8k_0^2}\right) AB_{xx} \\ &= \frac{1}{2}\omega_0 k_0^2 A_0^2 \left[(1+B) + BB^*(1+B) \right. \\ &\qquad \left. + (B+B^*)B + (B+B^*)\right] A , \end{aligned} \tag{418}$$

where $B^*(x,t)$ is the complex conjugate of the perturbed function $B(x,t)$. Neglecting squares of B, Eq. (418) reduces to

$$iB_t - \left(\frac{\omega_0}{8k_0^2}\right) B_{xx} = \frac{1}{2}\omega_0 k_0^2 A_0^2 (B + B^*) . \tag{419}$$

We look for a solution for perturbed quantity $B(x,t)$ in the form

$$\begin{aligned} B(x,t) &= B_1 \exp(\Omega t + i\ell x) \\ &\quad + B_2 \exp(\Omega^* t - i\ell x) , \end{aligned} \tag{420}$$

where B_1 and B_2 are complex constants, ℓ is a real wavenumber, and Ω is a growth rate (possibly a complex quantity) to be determined. Substituting the solution for B in (419) yields a pair of coupled equations:

$$\left(i\Omega + \frac{\omega_0 \ell^2}{8k_0^2}\right) B_1 - \frac{1}{2}\omega_0 k_0^2 A_0^2 (B_1 + B_2^*) = 0 , \tag{421}$$

$$\left(i\Omega^* + \frac{\omega_0 \ell^2}{8k_0^2}\right) B_2 - \frac{1}{2}\omega_0 k_0^2 A_0^2 (B_1^* + B_2) = 0 . \tag{422}$$

We take the complex conjugate of (422) to transform it into the form

$$\left(-i\Omega + \frac{\omega_0 \ell^2}{8k_0^2}\right) B_2^* - \frac{1}{2}\omega_0 k_0^2 A_0^2 (B_1 + B_2^*) = 0 . \tag{423}$$

The pair of linear homogeneous Eqs. (421) and (423) for B_1 and B_2^* admits a nontrivial eigenvalue for Ω provided

$$\begin{vmatrix} i\Omega + \frac{\omega_0 \ell^2}{8k_0^2} - \frac{1}{2}\omega_0 k_0^2 A_0^2 & -\frac{1}{2}\omega_0 k_0^2 A_0^2 \\ -\frac{1}{2}\omega_0 k_0^2 A_0^2 & i\Omega + \frac{\omega_0 \ell^2}{8k_0^2} - \frac{1}{2}\omega_0 k_0^2 A_0^2 \end{vmatrix} = 0 , \tag{424}$$

which is equivalent to

$$\Omega^2 = \frac{1}{2}\left(\frac{\omega_0 \ell}{2k_0}\right)^2 \left(k_0^2 A_0^2 - \frac{\ell^2}{8k_0^2}\right) . \tag{425}$$

The growth rate Ω is purely imaginary or real (and positive) depending on whether $\ell^2 > 8k_0^4 A_0^2$ or $\ell^2 < 8k_0^4 A_0^2$. The former case represents a wave solution for B, and the latter corresponds to the *Benjamin–Feir* (or *modulational*) *instability* with a criterion in terms of the nondimensional wavenumber $\tilde{\ell} = (\ell/k_0)$ as

$$\tilde{\ell}^2 < 8k_0^2 A_0^2 . \tag{426}$$

Thus, the range of instability is given by

$$0 < \tilde{\ell} < \tilde{\ell}_c = 2\sqrt{2}\, k_0 A_0 . \tag{427}$$

Since Ω is a function of $\tilde{\ell}$, the maximum instability occurs at $\tilde{\ell} = \tilde{\ell}_{\max} = 2k_0 A_0$, with a maximum growth rate given by

$$(\mathrm{Re}\,\Omega)_{\max} = \frac{1}{2}\omega_0 k_0^2 A_0^2 . \tag{428}$$

To establish the connection with the Benjamin–Feir instability [6], we have to find the velocity potential for the fundamental wave mode multiplied by $\exp(kz)$. It turns out that the term proportional to B_1 is the upper sideband, whereas that proportional to B_2 is the lower sideband. The main conclusion of this analysis is that Stokes water waves are definitely *unstable*.

Future Directions

Although this chapter has been devoted to water waves and the Korteweg and de Vries equations, there are some challenging problems dealing with solitary waves envelopes in the wake of a ship in oceans, and the KdV equation with variable coefficients in an ocean of variable depth. Despite some recent progress, there is no complete theory for ships in waves that takes into account of the effects of the finite ship volume, and possible interactions

between Kelvin wake and soliton envelopes. In order to respond to experimental results with a rough bottom, theories based on an empirical formula for the bottom stress have been developed, but they are not yet completely satisfactory when compared to experiments.

Of the systems that conserve energy, some are completely integrable in the sense of solitary wave theory, but many including those most important for applications are not. In some cases, useful analytical techniques are currently available, and others are waiting to be discovered. So there is a need for research in the area of nonlinear lattices. When we deal with real world problems, the conservation of energy does not hold because of dissipative effects and forcing terms. Consequently, the KdV equation and other relevant evolution equations need modification by including terms that describe such effects. So, there are challenging problems dealing with these modified evolution equations, their methods of solutions and the qualitative and quantitative understanding of the solutions. On the other hand, prospects for research in the transverse coupling between solitary waves and parallel systems are indeed bright.

Bibliography

Primary Literature

1. Airy GB (1845) Tides and Waves. In: Encyclopedia Metropolitana, Article 192
2. Akylas TR (1984) On the excitation of long nonlinear water waves by a moving pressure distribution. J Fluid Mech 141:455–466
3. Akylas TR (1984) On the excitation of nonlinear water waves by a moving pressure distribution oscillatory at resonant frequency. Phys Fluids 27:2803–2807
4. Amick CJ, Fraenkel LE, Toland JF (1982) On the Stokes Conjecture for the wave of extreme form. Acta Math Stockh 148:193–214
5. Benjamin TB (1967) Instability of periodic wave trains in nonlinear dispersive systems. Proc Roy Soc London A 299:59–75
6. Benjamin TB, Feir JE (1967) The disintegration of wavetrains on deep water. Part 1. Theory. J Fluid Mech 27:417–430
7. Benney DJ, Roskes G (1969) Wave instabilities. Stud Appl Math 48:377–385
8. Berezin YA, Karpman VI (1966) Nonlinear evolution of disturbances in plasmas and other dispersive media. Soviet Phys JETP 24:1049–1056
9. Boussinesq MJ (1871) Théorie de l'intumescence liquide appelée onde solitaire ou de translation se propageant dans un canal rectangulaire. Comptes Rendus 72:755–759
10. Boussinesq MJ (1871) Théorie generale des mouvenments qui sont propagés dans un canal rectangulaire horizontal. Comptes Rendus 72:755–759
11. Boussinesq MJ (1872) Théorie des ondes et des rumous qui se propagent le long d'un canal rectagulaire horizontal, en communiquant au liquide contenu dans ce canal des vitesses sensiblement pareilles de la surface au fond. J Math Pures Appl ser 2/17:55–108
12. Boussinesq MJ (1877) Essai sur la théorie des eaux courants. Mémoires Acad Sci Inst Fr ser 2 23:1–630
13. Cercignani C (1977) Solitons. Theory and application. Rev Nuovo Cimento 7:429–469
14. Chu VH, Mei CC (1970) On slowly-varying Stokes waves. J Fluid Mech 41:873–887
15. Chu VH, Mei CC (1971) The nonlinear evolution of Stokes waves in deep water. J Fluid Mech 47:337–351
16. Crapper GD (1957) An exact solution for progressive capillary waves of arbitrary amplitude. J Fluid Mech 2:532–540
17. Davey A, Stewartson K (1974) On three-dimensional packets of surface waves. Proc Roy Soc London A 338:101–110
18. Debnath L, Rosenblat S (1969) The ultimate approach to the steady state in the generation of waves on a running stream. Quart J Mech Appl Math XXII:221–233
19. Debnath L (1994) Nonlinear Water Waves. Academic Press, Boston
20. Debnath L (2005) Nonlinear Partial Differential Equations for Scientists and Engineers (Second Edition). Birkhauser, Boston
21. Debnath L, Bhatta D (2007) Integral Transforms and Their Applications. Second Edition. Chapman Hall/CRC Press, Boca Raton
22. Dutta M, Debnath L (1965) Elements of the Theory of Elliptic and Associated Functions with Applications. World Press Pub. Ltd., Calcutta
23. Fermi E, Pasta J, Ulam S (1955) Studies of nonlinear problems I. Los Alamos Report LA 1940. In: Newell AC (ed) Lectures in Applied Mathematics. American Mathematical Society, Providence 15:143–156
24. Gardner CS, Greene JM, Kruskal MD, Miura RM (1967) Method for solving the KdV equation. Phys Rev Lett 19:1095–1097
25. Gardner CS, Greene JM, Kruskal MD, Miura RM (1974) Korteweg–de Vries equation and generalizations, VI, Methods for exact solution. Comm Pure Appl Math 27:97–133
26. Hammack JL, Segur H (1974) The Korteweg–de Vries equation and water waves. Part 2. Comparison with experiments. J Fluid Mech 65:289–314
27. Hasimoto H, Ono H (1972) Nonlinear modulation of gravity waves. J Phys Soc Jpn 35:805–811
28. Helal MA, Molines JM (1981) Nonlinear internal waves in shallow water. A Theoretical and Experimental study. Tellus 33:488–504
29. Hirota R (1971) Exact solution of the Korteweg–de Vries equation for multiple collisions of solitons. Phys Rev Lett 27:1192–1194
30. Hirota R (1973) Exact envelope-soliton solutions of a nonlinear wave equation. J Math Phys 14:805–809
31. Hirota R (1973) Exact N-Solutions of the wave equation of long waves in shallow water and in nonlinear lattices. J Math Phys 14:810–814
32. Huang DB, Sibul OJ, Webster WC, Wehausen JV, Wu DM, Wu TY (1982) Ship moving in a transcritical range. Proc. Conf. on Behavior of Ships in Restricted Waters (Varna, Bulgaria) 2:26-1–26-10
33. Infeld E (1980) On three-dimensional generalizations of the Boussinesq and Korteweg–de Vries equations. Quart Appl Math XXXVIII:277–287
34. Johnson RS (1980) Water waves and Korteweg–de Vries equations. J Fluid Mech 97:701–719

35. Johnson RS (1997) A Modern Introduction to the Mathematical Theory of Water Waves. Cambridge University Press, Cambridge
36. Kadomtsev BB, Petviashvili VI (1970) On the stability of solitary waves in weakly dispersive media. Sov Phys Dokl 15:539–541
37. Kaplan P (1957) The waves generated by the forward motion of oscillatory pressure distribution. In: Proc. 5th Midwest Conf. Fluid Mech, pp 316–328
38. Karpman VI (1975) Nonlinear Waves in Dispersive Media. Pergamon Press, London
39. Korteweg DJ and de Vries G (1895) On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves. Phil Mag (5) 39:422–443
40. Kraskovskii YP (1960) On the theory of steady waves of not all small amplitude (in Russian). Dokl Akad Nauk SSSR 130:1237–1252
41. Kraskovskii YP (1961) On the theory of permanent waves of finite amplitude (in Russian). Zh Vychisl Mat Fiz 1:836–855
42. Lax PD (1968) Integrals of nonlinear equations of evolution and solitary waves. Comm. Pure Appl Math 21:467–490
43. Levi–Civita T (1925) Détermination rigoureuse des ondes permanentes d'ampleur finie. Math Ann 93:264–314
44. Lighthill MJ (1965) Group velocity. J Inst Math Appl 1:1–28
45. Lighthill MJ (1967) Some special cases treated by the Whitham theory. Proc Roy Soc London A 299:38–53
46. Lighthill MJ (1978) Waves in Fluids. Cambridge University Press, Cambridge
47. Luke JC (1967) A variational principle for a fluid with a free surface. J Fluid Mech 27:395–397
48. Michell JH (1893) On the highest waves in water. Phil Mag 36:430–435
49. Miche R (1944) Movements ondulatories de la mer en profondeur constants ou décreoissante. Forme limite de la houle lors de son déferlement. Application aux dignes maritimes. Ann Ponts Chaussées 114:25–78, 131–164, 270–292, 369–406
50. Palais RS (1997) The symmetries of solitons. Bull Amer Math Soc 34:339–403
51. Rayleigh L (1876) On waves. Phil Mag 1:257–279
52. Riabouchinsky D (1932) Sur l'analogie hydraulique des mouvements d'un fluide compressible, Institut de France, Académie des Sciences. Comptes Rendus 195:998
53. Stoker JJ (1957) Water Waves. Interscience, New York
54. Stokes G (1847) On the theory of oscillatory waves. Trans Camb Phil Soc 8:197–229
55. Struik DJ (1926) Détermination rigoureuse des ondes irrotationnelles périodiques dans un canal à profondeur finie. Math Ann 95:595–634
56. Toland JF (1978) On the existence of a wave of greatest height and Stokes' conjecture. Proc. Roy. Soc. Lond. A 363:469–485
57. Weidman PD, Maxworthy T (1978) Experiments on strong interactions between solitary waves. J Fluid Mech 85:417–431
58. Whitham GB (1965) A general approach to linear and nonlinear dispersive waves using a Lagrangian. J Fluid Mech 22:273–283
59. Whitham GB (1965) Nonlinear dispersive waves, Proc Roy Soc London A 283:238–261
60. Whitham GB (1967) Nonlinear dispersion of water waves. J Fluid. Mech 27:399–412
61. Whitham GB (1967) Variational methods and application to water waves. Proc Roy Soc London A 299:6–25
62. Whitham GB (1974) Linear and Nonlinear Waves. John Wiley, New York
63. Whitham GB (1984) Comments on periodic waves and solitons. IMA J Appl Math 32:353–366
64. Wu DM, Wu TY (1982) Three-dimensional nonlinear long waves due to a moving pressure. In: Proc 14th Symp. Naval Hydrodyn. National Academy of Sciences, Washington DC, pp 103–129
65. Zabusky NJ, Kruskal MD (1965) Interaction of solitons in a collisionless plasma and the recurrence of initial states. Phys Rev Lett 15:240–243
66. Zabusky NJ (1967) A synergetic approach to problems of nonlinear dispersive wave propagation and interaction. In: Ames WF (ed) Proc. Symp. on Nonlinear Partial Differential Equations. Academic Press, Boston. pp 223–258
67. Zabusky NJ, Galvin CJ (1971) Shallow water waves, the Korteweg–de Vries equation and solitons. J Fluid Mech 48:811–824

Books and Reviews

Ablowitz MJ, Clarkson PA (1991) Solitons, Nonlinear Evolution Equations and Inverse Scattering. Cambridge University Press, Cambridge

Bhatnagar PL (1979) Nonlinear Waves in One-Dimensional Dispersive Systems. Oxford University Press, Oxford

Crapper GD (1984) Introduction to Water Waves. Ellis Horwood Limited, Chichester

Debnath L (1983) Nonlinear Waves. Cambridge University Press, Cambridge

Dodd RK, Eilbeck JC, Gibbons JD, Morris HC (1982) Solitons and Nonlinear Wave Equations. Academic Press, London

Drazin PG, Johnson RS (1989) Solitons: An Introduction. Cambridge University Press, Cambridge

Infeld E, Rowlands G (1990) Nonlinear Waves, Solitons and Chaos. Cambridge University Press, Cambridge

Lamb H (1932) Hydrodynamics, 6th edition. Cambridge University Press, Cambridge

Leibovich S, Seebass AR (1972) Nonlinear Waves. Cornell University Press, Ithaca

Myint-U T, Debnath L (2007) Linear Partial Different Equations for Scientists and Engineers (fourth edition). Birkhauser, Boston

Russell JS (1844) Report on waves. In: Murray J (ed) Report on the 14th Meeting of the British Association for the Advancement of Science. London 311–390

Wavelets, Introduction to

EDWARD ABOUFADEL
Department of Mathematics,
Grand Valley State University, Allendale, USA

Wavelets emerged in the late 1980s as a valuable new tool in science and engineering, and as a topic for fruitful mathematical research. Wavelet transforms are now regularly applied in areas such as image processing, statistical analysis, and seismic research. Wavelets are at the heart of

the WSQ standard used by the United States' Federal Bureau of Investigation to compress fingerprint images [1]. They have been implemented in the Red professional digital video camera [2] and they are now being used in an attempt to analyze paintings by famous artists [3]. In the past twenty years, related "-lets" tools have also emerged, such as ridgelets and shearlets (see ► Curvelets and Ridgelets). The 1992 book by Daubechies [4] is a classic in the field.

Wavelet analysis shares with Fourier analysis the key idea of having a basis of functions with which to analyze signals and other functions. While Fourier analysis uses sine and cosine functions (or complex exponentials) – which are smooth and periodic – wavelet analysis utilizes functions that tend to be rough and localized (see ► Wavelets and the Lifting Scheme for examples of wavelet functions). To create the basis in Fourier analysis, the frequencies of the standard functions are varied. However, to create a system of wavelet functions, a scaling function is modified by translations and dilations, leading to various resolutions and hence the term *multiresolution* that is frequently used in the field. There are both discrete and continuous Fourier transforms, and we have the same for wavelets (see ► Comparison of Discrete and Continuous Wavelet Transforms). There are also analogues of the Fast Fourier Transform. Ironically, much of the derivation of results in wavelets uses the Fourier transform as a theoretical tool.

To be more specific, the creation of a wavelet basis starts with a *scaling function* ϕ and a *mother wavelet* ψ. (The scaling function is sometimes called the "father wavelet".) These functions tend to either have compact support (meaning they are equal to 0 outside of a bounded interval), or decay exponentially outside of a relatively small part of their domain. Translations and dilations of the mother wavelet are created in the form $\psi(2^n t - k)$, where t and k are integers. The resolution varies with n, with larger positive values of n corresponding to finer resolutions. By varying k, the support of the function can be translated, allowing localized analysis of other functions or signals.

Wavelet transformations are described and applied in multiple ways. As functions, wavelets can be studied with either the real line or the complex plane as a domain. High pass and low pass wavelet filters can be applied to tease out details in signals or images (see ► Popular Wavelet Families and Filters and Their Use). To think of wavelet transformations as a combination of "prediction" and "update" steps is the main idea behind the lifting scheme (see ► Wavelets and the Lifting Scheme). In addition, working with Fourier transforms of wavelet functions lead to various polyphase results (see ► Multiwavelets, for example).

As suggested above, the use of wavelets in various applications has contributed to their popularity. In statistics, wavelets are used to denoise signals and estimate properties of zero-mean stochastic processes (see ► Statistical Applications of Wavelets). Wavelet techniques have been combined with methods from partial differential equations to clean up images, or fill in damaged parts of an image (which is called *inpainting*) (see ► Wavelets and PDE Techniques in Image Processing, A Quick Tour of). Because images from astronomy tend to be isotropic (that is, having similar shapes or values in all directions, such as a sphere), wavelet methods have also been useful in that discipline (see ► Numerical Issues When Using Wavelets).

Wavelet analysis is preferred to Fourier analysis in many applications. One reason is that the relative roughness of the wavelet functions is a better match than trigonometric functions are for real data. This leads to a rather sparse representation of this data in the wavelet domain, which is helpful for compression and noise reduction. The Fourier transform breaks down a function into different frequencies; consequently this representation is based on averages of the function over its whole domain. Wavelets give you information about a function at different scales and different locations simultaneously. The localized nature of wavelet basis functions leads to better identification and elimination of local behavior in data such as spikes, edges, and other discontinuities. In addition, the nature of wavelets allows analysis at multiple resolutions, from coarse to fine.

A fertile area of research in wavelets is to consider extensions where the domain and/or range of the wavelet functions are multi-dimensional. For instance, since images are two-dimensional objects, a natural area to investigate are situations where the domain is $\mathbf{R}^2$ (see ► Bivariate (Two-dimensional) Wavelets). Basic wavelet families are orthogonal, but they don't need to be (see ► Numerical Issues When Using Wavelets). In order to work with radial images such as retinal scans [5], or anisotropic images (which are long and narrow, or tall and thin), other wavelet tools have been developed (see ► Curvelets and Ridgelets). Multiwavelets are functions on the real line whose range is two or more copies of the complex plane (see ► Multiwavelets). Finally, there is an intimate connection between wavelets and splines (see ► Multivariate Splines and Their Applications).

For the reader unfamiliar with wavelets, sections I, II, and III of the ► Wavelets and the Lifting Scheme article, along with the glossary of the ► Comparison of Discrete and Continuous Wavelet Transforms article, are good places to begin. A familiarity with the Fourier transform is critical to read many of the articles in this section.

Bibliography

1. Bradley J, Brislawn C, Hopper T (1993) The FBI wavelet/scalar quantization standard for gray-scale fingerprint image compression. In: Proc Conf Visual Info Process II Proc SPIE, vol 1961. pp 293–304
2. RED Digital Cinema Camera Company www.red.com
3. Lyu S, Rockmore D, Farid H (2004) A digital technique for art authentication. Proc Nat Acad Sci 101:17006–17010
4. Daubechies I (1992) Ten lectures on wavelets. SIAM
5. Feng P et al (2007) Enhancing retinal image by the Contourlet transform. Pattern Recognit Lett 28:516–522

Wavelets and the Lifting Scheme

Anders La Cour–Harbo[1], Arne Jensen[2]

[1] Department of Electronics Systems, Aalborg University, Aalborg East, Denmark

[2] Department of Mathematical Sciences, Aalborg University, Aalborg East, Denmark

Article Outline

Glossary
Definition of the Subject
Introduction
Lifting
Prediction and Update
Interpretation
Lifting and Filter Banks
Making Lifting Steps from Filters
Bibliography

Glossary

Lifting step An equation describing the transformation of an odd (even) sample of a signal by means of a linear combination of even (odd) samples, respectively. Changing an odd sample is sometimes called a prediction step, while changing an even sample is called an update step.

Discrete wavelet transform (DWT) Transformation of a discrete (sampled) signal into another discrete signal by means of a wavelet basis. The transform can be accomplished in a number of ways, typically either by a two channel filter bank or by lifting steps.

Filter bank A series of bandpass filters, which separates the input signal into a number of components, each with a distinct range of frequencies of the original signal.

z-Transform A mapping of a vector $\mathbf{x} = \{x[k]\}$ to a function in the complex plane by

$$X(z) = \sum_k x[k]z^{-k} .$$

For a vector with a finite number of non-zero entries $X(z)$ is a Laurent polynomial.

Filter A linear map, which maps a signal $\mathbf{x}$ with finite energy to another signal with finite energy. In the time domain it is given by convolution with a vector $\mathbf{h}$, which in the frequency domain is equal to $H(z)X(z)$. The vector $\mathbf{h}$ is called the impulse response (IR) of the filter (or sometimes the filter taps), and $H(e^{j\omega})$ the transfer function (or sometimes the frequency response). If $\mathbf{h}$ is a finite sequence, then $\mathbf{h}$ is called a FIR (finite impulse response) filter. An infinite $\mathbf{h}$ is then called an IIR filter. We only consider FIR filters in this article. We restrict ourselves to filters with real coefficients.

Filter taps The entries in the vector that by convolution with a signal gives a filtering of the signal. The filter taps are also called the impulse response of the filter.

Laurent polynomial A polynomial in the variables z and z^{-1}. Some examples: $3z^{-2} + 2z$, $z^{-3} - 2z^{-1}$, and $1 + z + z^3$. The z-transform of a FIR filter $\mathbf{h}$ is a Laurent polynomial of the form

$$H(z) = \sum_{k=k_b}^{k_e} h[k]z^{-k} , \quad k_b \le k_e .$$

This is in contrast to ordinary polynomials, where we only have non-negative powers of z. Assuming that $h[k_e] \neq 0$ and $h[k_b] \neq 0$, the degree of a Laurent polynomial is defined as $|h| = k_e - k_b$.

Monomial A Laurent polynomial of degree 0, for example $3z^7$, z^{-8}, or 12.

Definition of the Subject

The objective of this article is to give a concise introduction to the discrete wavelet transform (DWT) based on a technique called *lifting*. The lifting technique allows one to give an elementary, but rigorous, definition of the DWT, with modest requirements on the reader. A basic knowledge of linear algebra and signal processing will suffice. The lifting based definition is equivalent to the usual filter bank based definition of the DWT. The article does not discuss applications in any detail. The reader is referred to other articles in this collection.

The DWT was introduced in the second half of the eighties, through the work of Y. Meyer, I. Daubechies, S. Mallat, and many other researchers. The lifting technique was introduced by W. Sweldens in 1996, and the equivalence with the filter bank approach was established by I. Daubechies and W. Sweldens in an article published in 1998 [2], but available in preprint form from 1996.

This article is based on the book Ripples in Mathematics – The Discrete Wavelet Transform [3] with kind permission of Springer Science and Business Media. For more information on the background and history of wavelets, we again refer to other articles in this collection.

Introduction

We begin the exposition of the lifting technique with a simple, yet very descriptive example. It will shortly become apparent how this is related to lifting and to the discrete wavelet transform. We take a digital signal consisting of eight samples (this idea is due to Mulcahy [5], and we use his example, with a minor modification)

$$56, \quad 40, \quad 8, \quad 24, \quad 48, \quad 48, \quad 40, \quad 16\,.$$

We choose to believe that these numbers are not random, but have some structure that we want to extract. Perhaps there is some correlation between a number and its immediate successor. To reveal such a correlation we will take the numbers in pairs and compute the mean, and the difference between the first member of the pair and the computed mean. The result of this computation is shown as the second row in Table 1. This row contains the four means followed by the four differences, the latter being typeset in boldface. We then leave the four differences unchanged and apply the mean and difference computations to the first four entries of the second row. We repeat this procedure once more on the third row. The fourth row then contains a first entry, which is the mean of the original eight numbers, and the seven calculated differences. The boldface entries in the table are called the details of the signal.

It is important to observe that no information has been lost in this transformation of the first row into the fourth row. We can see this by reversing the calculation. Beginning with the last row, we compute the first two entries in the third row as $32 = 35 + (-3)$ and $38 = 35 - (-3)$. In the same way the first four entries in the second row are calculated as $48 = 32 + (16)$, $16 = 32 - (16)$, $48 = 38 + (10)$, and finally $28 = 38 - (10)$. Repeating this procedure we get the first row in the table.

Wavelets and the Lifting Scheme, Table 1
Mean and difference computation. Differences are in *boldface type*

56	40	8	24	48	48	40	16
48	16	48	28	**8**	**−8**	**0**	**12**
32	38	**16**	**10**	**8**	**−8**	**0**	**12**
35	**−3**	**16**	**10**	**8**	**−8**	**0**	**12**

So, what is the purpose of these computations? The four signals contain the same information, just in different ways, so how do we gain anything from this change of representation of the signal? If our assumption (pairwise equality of samples) is correct, the four means in the second row will equal the original numbers, and the four differences will be zero. If further, the original numbers are equal in groups of four, the two means in the third row will equal the original quadruples, and the differences will be zero. Finally, if all eight original samples are equal, the mean in row four will be equal to all eight samples, and the seven differences will be zero.

Apparently, our assumption is not correct, as the differences are not zero. However, the numbers in the fourth row are generally smaller than the original numbers. This indicates that while the numbers are not equal, they are 'similar', leaving us with small differences. We can say that we have achieved some kind of loss-free compression by reducing the dynamic range of the signal. By loss-free we mean that we can transform back to the original signal without any information being lost in the process. As a simple measure of compression we can count the number of digits used to represent the signal. The first row contains 15 digits. The last row contains 12 digits and two negative signs. So in this example the compression is not very large. But it is easy to give other examples, where the compression can be substantial.

We see in this example that the pair 48, 48 do fit our assumption of equality, and therefore the difference is zero. Suppose that after transformation we find that many 'difference entries' are zero. Then we can store the transformed signal more efficiently by only storing the nonzero entries (and their locations).

Processing the Transformed Signal

Let us now suppose that we are willing to accept a certain loss of quality in the signal, if we can get a higher rate of compression. One technique for lossy compression is called thresholding. We choose a threshold and decide to set equal to zero all entries with an absolute value less than this threshold value. Let us in our example choose 4 as the threshold. This means that we in Table 1 replace the entry **−3** by **0** and then perform the reconstruction. The result is in Table 2, left hand part.

The original and transformed signals are both shown on the left in Fig. 1. We have chosen to join the sample points by straight line segments to get a good visualization of the signals. Clearly the two graphs differ very little. Now let us perform a more drastic compression. This time we choose the threshold equal to 9. The computations are

Wavelets and the Lifting Scheme, Table 2
Reconstruction with threshold 4 (*top*) and threshold 9 (*bottom*)

59	43	11	27	45	45	37	13
51	19	45	25	**8**	**−8**	**0**	**12**
35	35	**16**	**10**	**8**	**−8**	**0**	**12**
35	**0**	**16**	**10**	**8**	**−8**	**0**	**12**

51	51	19	19	45	45	37	13
51	19	45	25	**0**	**0**	**0**	**12**
35	35	**16**	**10**	**0**	**0**	**0**	**12**
35	**0**	**16**	**10**	**0**	**0**	**0**	**12**

given on the right in Table 2, and the graphs are plotted on the right in Fig. 1. Notice that the peaks in the original signal have been flattened. We also note that the signal now is represented by only four non-zero entries.

There are several variations of the transformation. We could have stored differences instead of 'half-differences', or we could have used the difference between the second element of the pair and the computed average. The first choice will lead to boldface entries in the tables that can be obtained from the computed ones by multiplication by a factor -2. The second variant is obtained by multiplication by -1.

The 'mean-difference' transformation can be performed on any signal of even length, since all we need is the possibility of taking pairs. If we want to be able to keep transforming the signal until there is one signal mean left, we need a signal of length 2^N, and it will lead to a table with $N+1$ rows, where the first row is the original signal. If the given signal has a length different from a power of 2, then we will have to do some additional operations on the signal to compensate for that. One possibility is to add samples with value zero to one or both ends of the signal until a length of 2^N is achieved. This is called zero padding.

Lifting

The transformation performed by successively calculating means and differences of a signal is an example of the discrete wavelet transform. Specifically, the transform introduced in the example is called the Haar transform, and occasionally also the first of the Daubechies transforms. It can be undone by simply reversing the steps performed, and it provides a number of different representations of the same signal. Actually, all discrete wavelet transforms can be realized by a similar procedure, where the only modification from the Haar transform is the way, in which we compute the means part and differences part of the transformed signal. The procedure is known as 'lifting' in the literature, and was introduced by Wim Sweldens in 1996 in a series of papers [8,9].

To fully appreciate the concept of lifting we need to elaborate on the example in the previous section. We have a discrete signal of real (or complex) numbers

$$x[0], x[1], x[2], x[3], \ldots, x[N-1] .$$

Our convention is to start indexing at zero. In implementations one may need to adapt this to the programming language used. Here we assume the signal to have finite length N, but this is not a restriction. The theory also applies to infinite signals, provided they have finite energy. In the mathematical (and signal processing) literature finite energy means

$$\sum_{n=-\infty}^{\infty} |x[n]|^2 < \infty ,$$

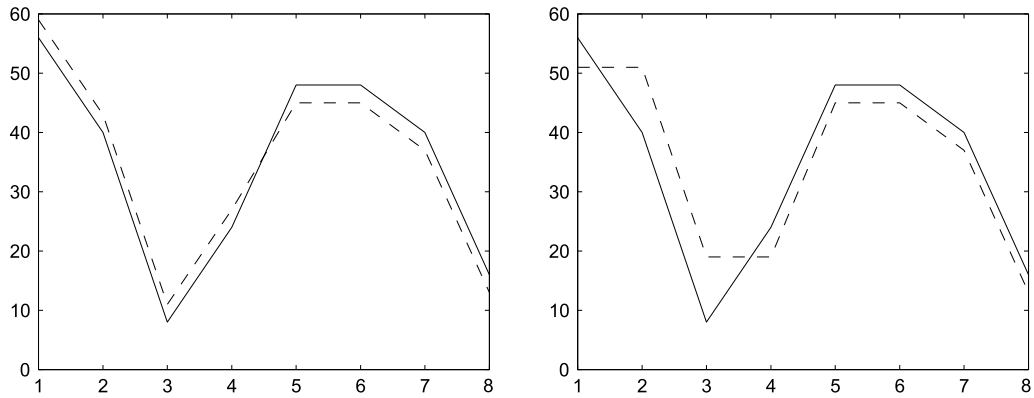

Wavelets and the Lifting Scheme, Figure 1
Original signal (*solid*) and modified signal (*dashed*) with threshold 4 (*left*) and 9 (*right*)

and the set of such signals is usually denoted by $\ell^2(\mathbf{Z})$. Sometimes we use the mathematical term sequence instead of signal, and we also use the term vector, in particular in connection with use of results from linear algebra.

Introducing Prediction and Update

Let us now return to the example in the Introduction. We took a pair of numbers a, b and computed the mean, and the difference between the first entry and the mean

$$s = \frac{a+b}{2}, \tag{1}$$

$$d = a - s\,. \tag{2}$$

Alternatively, almost the same result can be achieved in the following way

$$d = a - b\,, \tag{3}$$

$$s = \frac{a+b}{2} = \frac{d}{2} + b\,, \tag{4}$$

except we now have the difference rather than half the difference. This mathematical formulation of the Haar transform, albeit quite simple, brings us to the definition of lifting. The two operations, mean and difference, can be viewed as special cases of more general operations. Remember that we assumed that there is some correlation between two successive samples, and we therefore computed the difference. If two samples are almost equal, the difference is, of course, small. Thus one can use the first sample to predict the second sample, the prediction being that they are equal. If is is a good prediction, the difference between them is small. Thus we can call (3) a *prediction step*. We can use other and more sophisticated prediction steps than one based on just the previous sample. We will give a few examples of this later.

We also calculated the mean of the two samples. This step can be viewed in two ways. Either as an operation, which preserves some properties of the original signal (later we shall see how the mean value or the energy of a signal is preserved during transformation), or as an extraction of essential features of the signal. The latter viewpoint is based on the fact that the pair-wise mean values contain the overall structure of the signal, but with only half the number of samples. We use the term *update step* for the this operation, and (4) is the update step in the Haar transform. Just as the prediction operation, the update operation can be more sophisticated than just calculating the mean.

The prediction and update operations are shown in Fig. 2. The notation and the setup is a little different from the Introduction; we start with a finite sequence $\mathbf{s}_j$ of length 2^j and end with two sequences $\mathbf{s}_{j-1}$ and $\mathbf{d}_{j-1}$, each of length 2^{j-1}. Let us explain the steps in Fig. 2.

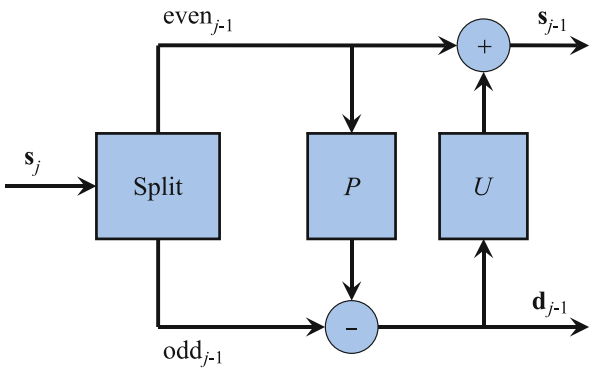

Wavelets and the Lifting Scheme, Figure 2
The three steps in a lifting building block. Note that the minus sign means 'the signal from the left minus the signal from the top'

split The entries are sorted into the even and the odd entries. It is important to note that we do this only to explain the algorithm. In (effective) implementations the entries are not moved or separated.

prediction If the signal contains some structure, then we can expect correlation between a sample and its nearest neighbors. In our first example the prediction is that the signal is constant. More elaborately, given the value at the sample number $2n$, we predict that the value at sample $2n+1$ is the same. We then compute the difference and in the figure let it replace the value at $2n+1$, since this value is not needed anymore. In our notation this is

$$d_{j-1}[n] = s_j[2n+1] - s_j[2n]\,.$$

In general, the idea is to have a prediction procedure P and then compute the difference signal as

$$\mathbf{d}_{j-1} = \mathrm{odd}_{j-1} - P(\mathrm{even}_{j-1})\,. \tag{5}$$

Thus in the $\mathbf{d}$ signal each entry is one odd sample minus some prediction based on a number of even samples.

update Given an even entry we predicted that the next odd entry has the same value, and stored the difference. We then update our even entry to reflect our knowledge of the signal. In the example above we replaced the even entry by the average. In our notation

$$s_{j-1}[n] = s_j[2n] + d_{j-1}[n]/2\,.$$

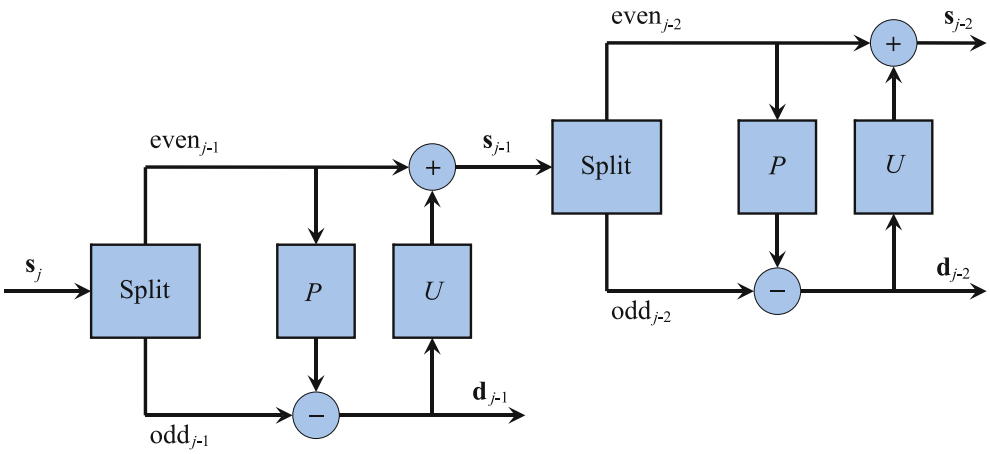

Wavelets and the Lifting Scheme, Figure 3
Two step discrete wavelet transform

In general we decide on an updating procedure, and then compute

$$\mathbf{s}_{j-1} = \text{even}_{j-1} + U(\mathbf{d}_{j-1})\,. \tag{6}$$

The algorithm described here is called one step lifting. It requires the choice of a prediction procedure P, and an update procedure U.

Wavelets and the Lifting Scheme, Table 3
This is Table 1 in the lifting step notation

$s_3[0]$	$s_3[1]$	$s_3[2]$	$s_3[3]$	$s_3[4]$	$s_3[5]$	$s_3[6]$	$s_3[7]$
$s_2[0]$	$s_2[1]$	$s_2[2]$	$s_2[3]$	$d_2[0]$	$d_2[1]$	$d_2[2]$	$d_2[3]$
$s_1[0]$	$s_1[1]$	$d_1[0]$	$d_1[1]$	$d_2[0]$	$d_2[1]$	$d_2[2]$	$d_2[3]$
$s_0[0]$	$d_0[0]$	$d_1[0]$	$d_1[1]$	$d_2[0]$	$d_2[1]$	$d_2[2]$	$d_2[3]$

Multiple Transforms

The discrete wavelet transform is obtained by combining a number of lifting steps. As in the example in Table 1 we keep the computed differences $\mathbf{d}_{j-1}$ and use the average sequence $\mathbf{s}_{j-1}$ as input for one more lifting step. This two step procedure is illustrated in Fig. 3. Obviously, we can apply more lifting steps if we want to transform the signal further. In general, we start with a signal $\mathbf{s}_j$ of length 2^j and we can repeat the transformation j times until we have a single number $s_0[0]$ and j difference sequences $\mathbf{d}_{j-1}$ to $\mathbf{d}_0$. If we let $j = 3$ and use the Haar transform $s_0[0]$ will be be the mean value of the eight entries in the original sequence, and $\mathbf{d}_0$, $\mathbf{d}_1$, and $\mathbf{d}_2$ will be the numbers in boldface in Table 1. In lifting step notation this table becomes as Table 3.

We have previously motivated the prediction operation with the reduction in dynamic range of the signal obtained in using differences rather than the original values, potentially leading to good compression of a signal. The update procedure has not yet been clearly motivated. The update performed in the first example was

$$s_{j-1}[n] = s_j[2n] + \tfrac{1}{2}d_{j-1}[n] = \tfrac{1}{2}(s_j[2n] + s_j[2n+1])\,.$$

It turns out that this update operation preserves the mean value. The consequence is that all the $\mathbf{s}$ sequences have the same mean value. It is easy to verify in the case of the example in Table 1, since

$$\frac{56+40+8+24+48+48+40+16}{8} = \frac{48+16+48+28}{4} = \frac{32+38}{2} = 35\,.$$

It is not difficult to see that this holds for any sequence $\mathbf{s}$ of length 2^j. In particular, $s_0[0]$ equals the mean value of the original samples $s_j[0], \ldots, s_j[2^j - 1]$.

Inverse Transform

A lifting step is easy to undo. In fact, looking at Fig. 2 it is clear that we can go backwards from $\mathbf{s}_{j-1}$ and $\mathbf{d}_{j-1}$ to $\mathbf{s}_j$ simply by first 'undoing' the update step and then 'undoing' the prediction step. This is accomplished by using the same update step as in the direct lifting step, but subtracting instead of adding, followed by the same prediction step, but using addition instead of subtraction. Finally, we need to merge the odd and even samples. The direct lifting step and the corresponding inverse lifting step are shown in Fig. 4. This merging step, where the sequences even$_{j-1}$ and odd$_{j-1}$ are merged to form the sequence $\mathbf{s}_j$, is shown to explain the algorithm. It is not performed in implementations, since the entries are not reordered in the first place.

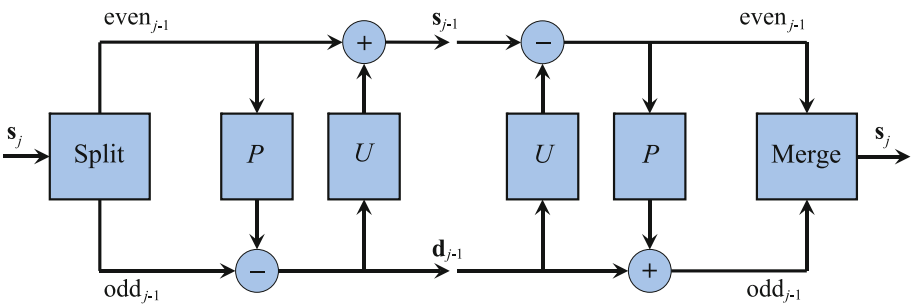

Wavelets and the Lifting Scheme, Figure 4
Direct and inverse lifting step

Mathematically, the inverse lifting step is accomplished by using the formulas in opposite order, and isolating the odd and even samples. In our example the inverse lifting step becomes

$$b = s - d/2\,, \tag{7}$$

$$a = d + b\,, \tag{8}$$

where (7) is actually (4) with b isolated, and (8) is (3) with a isolated. In the lifting step notation this is first (6) rewritten as

$$s_j[2n] = s_{j-1}[n] - d_{j-1}[n]/2$$

to undo the update step, and then (5) rewritten as

$$s_j[2n+1] = d_{j-1}[n] + s_j[2n]$$

to undo the prediction step. For the general lifting step

$$\mathbf{d}_{j-1} = \text{odd}_{j-1} - P(\text{even}_{j-1})$$
$$\mathbf{s}_{j-1} = \text{even}_{j-1} + U(\mathbf{d}_{j-1})$$

the inversion is accomplished by the steps

$$\text{even}_{j-1} = \mathbf{s}_{j-1} - U(\mathbf{d}_{j-1})$$
$$\text{odd}_{j-1} = \mathbf{d}_{j-1} + P(\text{even}_{j-1})\,.$$

That is all there is to the inverse transform.

Prediction and Update

Assuming that a signal is constant was useful in the first example to demonstrate the concept of lifting. However, this assumption is hardly the best for most signals. Fortunately, there are many other possible prediction and update procedures to accommodate various signal assumptions. We will present two examples of other lifting steps. The first example is the next logical step from the Haar transform when we assume the signal to be stepwise linear rather than just constant. The second example is the Daubechies 4 transform, which is not based on a specific assumption about the signal, but is taken from the classical wavelet theory and adapted to the lifting step method.

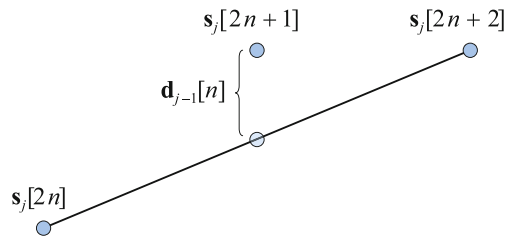

Wavelets and the Lifting Scheme, Figure 5
The linear prediction uses two even samples to predict one odd sample

Linear Prediction

Instead of basing the prediction on the assumption that the signal is constant, we will now base it on the assumption that the signal is linear. We really do mean an affine signal, but we stick to the commonly used term 'linear.' Thus a linear signal is in this context a signal of the form $s_j[n] = \alpha n + \beta$, that is, all the samples of the signal lie on a straight line. For a given odd entry $s_j[2n+1]$ we now need the two nearest even neighbors to predict the value (remember that for the constant signal we needed only one neighbor). The predicted value is $\frac{1}{2}(s_j[2n] + s_j[2n+2])$, the mean of the two even samples. This value is the circle on the middle of the line in Fig. 5.

The correction is then the difference between what we predict the middle sample to be and what it actually is

$$d_{j-1}[n] = s_j[2n+1] - \tfrac{1}{2}(s_j[2n] + s_j[2n+2])\,.$$

We decide to base the update procedure on the two most recently computed differences. We take it of the form

$$s_{j-1}[n] = s_j[2n] + A(d_{j-1}[n-1] + d_{j-1}[n])\,,$$

where A is a constant to be determined. In the first exam-

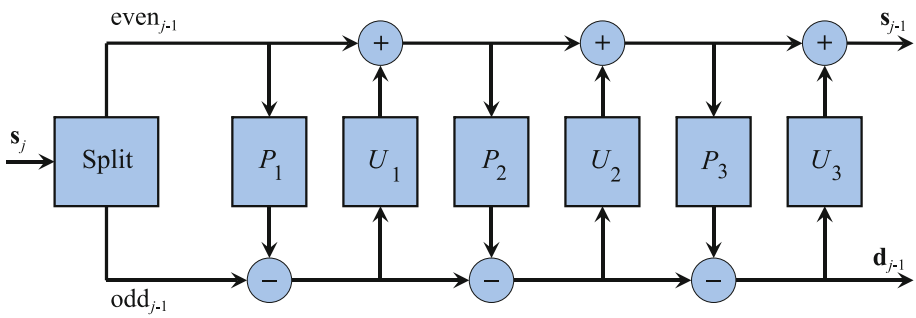

Wavelets and the Lifting Scheme, Figure 6
Multiple prediction and update steps

ple we had the property

$$\sum_n s_{j-1}[n] = \tfrac{1}{2}\sum_n s_j[n]\,, \tag{9}$$

that is, the mean was preserved. (Recall that $\mathbf{s}_j$ has length 2^j and $\mathbf{s}_{j-1}$ has length 2^{j-1}. This explains the factor $\frac{1}{2}$.) We would like to have the same property here. Let us first rewrite the expression for $s_{j-1}[n]$ above,

$$\begin{aligned} s_{j-1}[n] &= s_j[2n] + Ad_{j-1}[n-1] + Ad_{j-1}[n] \\ &= s_j[2n] + A\big(s_j[2n-1] - \tfrac{1}{2}s_j[2n-2] - \tfrac{1}{2}s_j[2n]\big) \\ &\quad + A\big(s_j[2n+1] - \tfrac{1}{2}s_j[2n] - \tfrac{1}{2}s_j[2n+2]\big)\,. \end{aligned}$$

Using this expression, and gathering even and odd terms, we get

$$\sum_n s_{j-1}[n] = (1-2A)\sum_n s_j[2n] + 2A\sum_n s_j[2n+1]\,.$$

To satisfy (9) we must choose $A = \frac{1}{4}$. Summarizing, we have the following two steps

$$d_{j-1}[n] = s_j[2n+1] - \tfrac{1}{2}(s_j[2n] + s_j[2n+2])\,, \tag{10}$$

$$s_{j-1}[n] = s_j[2n] + \tfrac{1}{4}(d_{j-1}[n-1] + d_{j-1}[n])\,. \tag{11}$$

The transform in this example also has the property.

$$\sum_n ns_{j-1}[n] = \tfrac{1}{4}\sum_n ns_j[n]\,. \tag{12}$$

Note that there is a misprint in this formula in [3]. We say that the transform preserves the *first moment* of the sequence. The mean is also called the *zeroth moment* of the sequence. Finally, we want to note that this transform is known in the literature as CDF(2,2). The origin of this name is explained later.

In the above presentation we have simplified the notation by not specifying where the finite sequences start and end, thereby for the moment avoiding keeping track of the ranges of the variables. In other words, we have considered our finite sequences as infinite, adding zeroes before and after the given entries. In implementations one has to keep track of these things, but doing so now would obscure the simplicity of the lifting procedure. For more details on transformation of finite signals, see Chap. 10 in [3].

Daubechies 4 Wavelet

We now turn to the second example, the Daubechies 4 wavelet. This transform is a bit different from what we have seen so far. So to adapt this transform to the lifting step method we need to first take a closer look at the construction of the lifting step. Looking at Fig. 4 once more we see that we can add another prediction step after the update step without changing the lifting concept; we still have even and odd samples, we still have $\mathbf{s}_{j-1}$ and $\mathbf{d}_{j-1}$ as output, and we can still invert the lifting step by reversing the order of the prediction and update steps. In fact, we can add as many prediction and update steps as we want, and still have the same lifting concept. As an illustration of this Fig. 6 shows a direct transform consisting of three pairs of prediction and update operations. If we have two prediction and only one update, but insist on having them in pairs (this is occasionally useful in the theory), we can always add an operation of either type, which does nothing.

It turns out that this generalization is crucial in applications. There are many important transforms, where the steps do not occur in pairs. The Daubechies 4 transform is an example, where there is a U operation followed by a P operation and another U operation. Furthermore, in the last two steps, in (16) and (17), we add a new type of operation which is called normalization, or sometimes rescaling. The resulting algorithm is applied to a signal $\mathbf{s}_j$

as follows

$$s_{j-1}^{(1)}[n] = s_j[2n] + \sqrt{3}s_j[2n+1] \tag{13}$$

$$d_{j-1}^{(1)}[n] = s_j[2n+1] - \tfrac{1}{4}\sqrt{3}s_{j-1}^{(1)}[n] - \tfrac{1}{4}(\sqrt{3}-2)s_{j-1}^{(1)}[n-1] \tag{14}$$

$$s_{j-1}^{(2)}[n] = s_{j-1}^{(1)}[n] - d_{j-1}^{(1)}[n+1] \tag{15}$$

$$s_{j-1}[n] = \frac{\sqrt{3}-1}{\sqrt{2}}s_{j-1}^{(2)}[n] \tag{16}$$

$$d_{j-1}[n] = \frac{\sqrt{3}+1}{\sqrt{2}}d_{j-1}^{(1)}[n]\,. \tag{17}$$

Since there is more than one U operation, we have used superscripts on the **s** and **d** signals in order to tell them apart. Note that in the normalization steps we have

$$\frac{\sqrt{3}-1}{\sqrt{2}} \cdot \frac{\sqrt{3}+1}{\sqrt{2}} = 1\,.$$

Normalization does not change the information on the signal, but is merely a means to have certain transform properties satisfied.

The Daubechies 4 equations are derived from the classical wavelet theory by a procedure called Euclidean factorization, see Sect. "Lifting and Filter Banks". While the Daubechies 4 transform does have a well-described origin in mathematics (see for instance Ten Lectures on Wavelets [1] by Daubechies), it becomes somewhat unclear in the lifting step version, what exactly this transform can provide. In particular, starting with an update step does not conform with our previous exposition of the lifting approach. However, this transform has one important property not possessed by the linear prediction transform CDF(2,2), or indeed any predicting lifting steps based on polynomial interpolation; in a frequency interpretation of the wavelet transforms the Daubechies 4 behaves somewhat nicer than CDF(2,2). This is because the former is an orthogonal transform, while the latter is a biorthogonal transform. We refer to other wavelet articles in this collection for more information on this subject.

While the Daubechies 4 is lacking a proper interpretation in terms of prediction and update, and thus may appear to be less appealing to include in the lifting concept, it's ability to be written as lifting step is in fact a consequence of the pleasant fact that all wavelet transforms (not matter their origin and design criteria) can be written as lifting steps. The two examples then show that some lifting steps comes from a sample-by-sample design method, while others may come from completely different design methods (the Daubechies transform was designed within the field of filter theory to have certain nice frequency related properties). This article touches upon the filter subject, see Sect. "Lifting and Filter Banks". For a more in-depth coverage of the subject, we refer the reader to other articles in this collection.

To find the inverse transform we have to use the prescription given above. We do the steps in reverse order and with the signs reversed. The normalization is undone by multiplication by the inverse constants. The result is

$$d_{j-1}^{(1)}[n] = \frac{\sqrt{3}-1}{\sqrt{2}}d_{j-1}[n] \tag{18}$$

$$s_{j-1}^{(2)}[n] = \frac{\sqrt{3}+1}{\sqrt{2}}s_{j-1}[n] \tag{19}$$

$$s_{j-1}^{(1)}[n] = s_{j-1}^{(2)}[n] + d_{j-1}^{(1)}[n+1] \tag{20}$$

$$s_j[2n+1] = d_{j-1}^{(1)}[n] + \frac{1}{4}\sqrt{3}s_{j-1}^{(1)}[n] + \frac{1}{4}(\sqrt{3}-2)s_{j-1}^{(1)}[n-1] \tag{21}$$

$$s_j[2n] = s_{j-1}^{(1)}[n] - \sqrt{3}s_j[2n+1]\,. \tag{22}$$

Incidentally, this transform illustrates one of the problems that has to be faced in computer implementations. For example, to compute $d_{j-1}^{(1)}[0]$ we need to know $s_{j-1}^{(1)}[0]$ and $s_{j-1}^{(1)}[-1]$. But to compute $s_{j-1}^{(1)}[-1]$ one needs the values $s_j[-2]$ and $s_j[-1]$, which are not defined. The easiest solution to this problem is to use zero padding to get a sample at this index value (zero padding means that all undefined samples are defined to be 0). There exists other more sophisticated methods, but that topic is beyond the scope of this article.

Finally, let us repeat our first example in the above notation. We also add a normalization step. In this form the transform is known as the Haar transform in the literature (we used this term previously, but in fact the scaling needs to be including for the transform to be the true Haar transform). The direct transform is

$$d_{j-1}^{(1)}[n] = s_j[2n+1] - s_j[2n] \tag{23}$$

$$s_{j-1}^{(1)}[n] = s_j[2n] + \tfrac{1}{2}d_{j-1}^{(1)}[n] \tag{24}$$

$$s_{j-1}[n] = \sqrt{2}s_{j-1}^{(1)}[n] \tag{25}$$

$$d_{j-1}[n] = \frac{1}{\sqrt{2}}d_{j-1}^{(1)}[n] \tag{26}$$

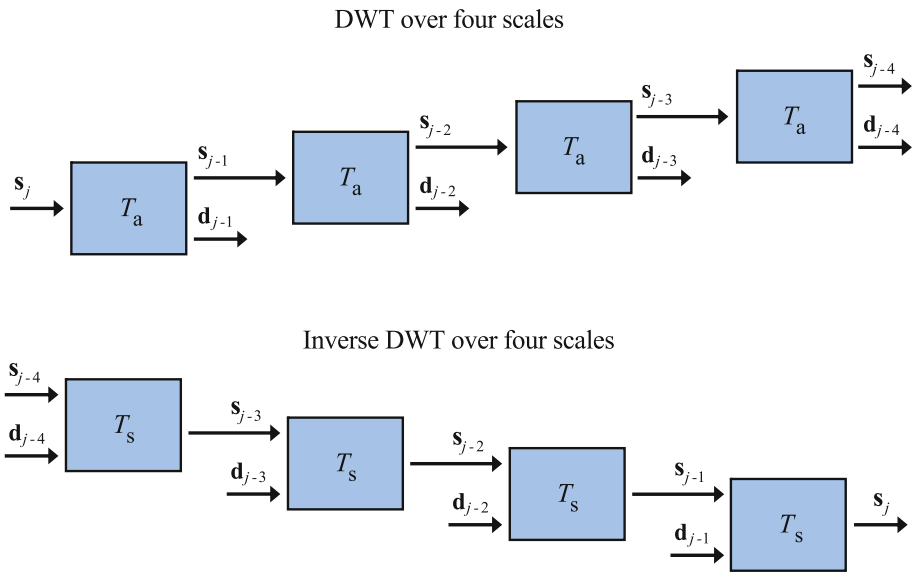

Wavelets and the Lifting Scheme, Figure 7
DWT and inverse DWT over four scales

and the inverse transform is given by

$$d_{j-1}^{(1)}[n] = \sqrt{2}d_{j-1}[n] \tag{27}$$

$$s_{j-1}^{(1)}[n] = \frac{1}{\sqrt{2}}s_{j-1}[n] \tag{28}$$

$$s_j[2n] = s_{j-1}^{(1)}[n] - \tfrac{1}{2}d_{j-1}^{(1)}[n] \tag{29}$$

$$s_j[2n+1] = s_j[2n] + d_{j-1}^{(1)}[n]\,. \tag{30}$$

We note that this transform can be applied to a signal of length 2^j without using zero padding. It turns out to be essentially the only transform with this property.

Discrete Wavelet Transform

We have now established a method for transforming one signal into another signal that has a natural division into two parts of equal length. This transform is the discrete wavelet transform (DWT). In many cases one is not interested in the actual implementation of the transform, though, and the DWT is in those cases often shown as a box with one input and two outputs. The direct transform is sometimes called 'analysis', and we will represent such a direct transform by the symbol T_a ('a' for analysis).

In our notation the input would be $\mathbf{s}_j$ and the outputs would be $\mathbf{s}_{j-1}$ and $\mathbf{d}_{j-1}$. The inverse transform is then naturally shown as another box with two inputs and one output, and is sometimes called 'synthesis'. Therefore, the inverse transform is represented by the symbol T_s.

The contents of the T_a box could be the direct Haar transform as given by (23)–(26), and the contents of the T_s box could be the inverse Haar transform as given by (27)–(30). Obviously, we must make sure to use the inverse transform corresponding to the applied direct transform. Otherwise, the results will be meaningless.

We can now combine these building blocks to get a multi-scale discrete wavelet transforms. We perform the transform over a certain number of scales k, meaning that we combine k of the building blocks as shown in Fig. 3 in the case of 2 scales, and with the upper-most diagram in Fig. 7 in the case of 4 scales. In the latter figure we use the building block representation of the individual steps.

It is possible to apply the DWT to the difference output of the transform as well, instead of just the average (or means) output. This is known as the wavelet packet transform, and is commonly found in the literature, and is often used in practical implementations. This concept introduces new possibilities and methods. We refer to other articles presenting this subject.

Interpretation

The main topic for this collection of articles is 'Wavelets'. This term refers to a function, often denoted by ψ in the wavelet literature, that takes many shapes and determines the properties of the wavelet transform. So what is the

relation between a series of lifting steps and this wavelet function, and is it useful to know this relation, if one just wants to use the wavelet transform? After all, the lifting step method works quite well without any apparent need for knowledge of the wavelet function.

It turns out that while the transform can be implemented easily with just a few equations, the understanding of what is actually happening with a signal, when it is transformed, relies heavily on a proper interpretation of the transform. Here we will present an interpretation based on linear algebra, which will provide some useful insights. However, this presentation is incomplete; we state some results from the general theory, and illustrate them with explicit computations, but we do not discuss the general theory in detail, since this is beyond the scope of this article.

In the following discussion we will continue to use the Haar transform example from the Introduction, partly because it allows for simple and yet convincing computations, partly because we believe that the reader has become familiar with this example by now. We will end with an interpretation of the Daubechies 4 transform.

Using Linear Algebra

We saw above that the first of the eight numbers in the transformed signal in the first example in the Introduction, that is, the number 35, had a special meaning; it was the mean of the entire signal. This property (though not the number itself) was actually independent of the signal; for any signal that first number $s_0[0]$ will always be the mean of the signal. In fact, it was a property derived from the transform, and it provided an interpretation of the number 35 in the transformed signal. Similar interpretations can be made of the other numbers in the transformed signal.

First, we examine the original example in Table 1 once more.

56	40	8	24	48	48	40	16
48	16	48	28	**8**	**−8**	**0**	**12**
32	38	**16**	**10**	**8**	**−8**	**0**	**12**
35	**−3**	**16**	**10**	**8**	**−8**	**0**	**12**

This table was constructed from top to bottom by multiple Haar transforms, and the 'signal mean' property was derived using equations, see for instance (9). However, we can reach the same conclusion by examining signal samples, and by starting at the bottom. To isolate the property of interest here, we start with a bottom row consisting of zeros at the entries we are not interested in interpreting, and a 1 at the entry we would like to interpret, that is, the first of the eight numbers. Then we do a three scale synthesis of the signal and get

1	1	1	1	1	1	1	1
1	1	1	1	**0**	**0**	**0**	**0**
1	1	**0**	**0**	**0**	**0**	**0**	**0**
1	**0**	**0**	**0**	**0**	**0**	**0**	**0**

Note that the normalization steps (25) and (26) has been omitted to make the table easier to read (this does not interfere with the interpretation, even though this fact may not be clear at this point). It is obvious that if we had started with any other number than 1, this number would then be the one repeated eight times in the top row after the three scale synthesis. This computation indicates that the entry at location 0 (recall that our indexing starts with 0) in the fourth row has an equal contribution from all eight original samples, and thus is the mean value.

Let us repeat this procedure with the remaining seven entries in the fourth row. Here are the first two, having a 1 at entry number 1 and 2.

1	1	1	1	−1	−1	−1	−1
1	1	−1	−1	**0**	**0**	**0**	**0**
1	−1	**0**	**0**	**0**	**0**	**0**	**0**
0	**1**	**0**	**0**	**0**	**0**	**0**	**0**

1	1	−1	−1	0	0	0	0
1	−1	0	0	**0**	**0**	**0**	**0**
0	0	**1**	**0**	**0**	**0**	**0**	**0**
0	**0**	**1**	**0**	**0**	**0**	**0**	**0**

Examining the first row in the left-most table reveals that the value at entry number 1, i. e. the value 1 in the fourth row, can be interpreted as the mean of the first four sample minus the mean of the last four sample. In filter terms this is roughly equal to a band pass filtering just above DC, i. e. DC does not contribute to entry number 1, and neither would any frequency higher than a single cycle.

The result of all eight computations are now represented using notation from linear algebra; they are inserted as columns in a matrix

$$\mathbf{W}_s^{(3)} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 & 1 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 & -1 \end{bmatrix}, \tag{31}$$

which is denoted by $\mathbf{W}$ with a subscript s for synthesis for the following reason. A signal of length 8 is a vector in the vector space $\mathbf{R}^8$. The process described above of reconstructing the signal, whose transform is one of the canonical basis vectors in $\mathbf{R}^8$, is the same as finding the columns in the matrix (with respect to the canonical basis) of the three scale synthesis transform. This is just the well known definition from linear algebra of the matrix of a linear transform. This in turn means that applying this matrix to any length 8 signal performs the inverse Haar transform. For example, multiplying it with the fourth row of Table 1 (regarded as a column vector) produces the first row of that same table.

The matrix of the direct three scale transform is obtained in a similar fashion; by computing the transforms of the eight canonical basis vectors in $\mathbf{R}^8$. In other words, we start with the signal $[1, 0, 0, 0, 0, 0, 0, 0]$ and carry out the three scale transform as shown here.

1	0	0	0	0	0	0	0
$\frac{1}{2}$	0	0	0	$\mathbf{\frac{1}{2}}$	**0**	**0**	**0**
$\frac{1}{4}$	0	$\mathbf{\frac{1}{4}}$	**0**	$\mathbf{\frac{1}{2}}$	**0**	**0**	**0**
$\frac{1}{8}$	$\mathbf{\frac{1}{8}}$	$\mathbf{\frac{1}{4}}$	**0**	$\mathbf{\frac{1}{2}}$	**0**	**0**	**0**

Computing the other seven transforms and inserting the resulting signals as column vectors gives

$$\mathbf{W}_{\mathrm{a}}^{(3)} = \begin{bmatrix} \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} \\ \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ \frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & -\frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & -\frac{1}{2} \end{bmatrix} . \tag{32}$$

Multiplying the matrices we find $\mathbf{W}_{\mathrm{s}}^{(3)} \cdot \mathbf{W}_{\mathrm{a}}^{(3)} = \mathbf{I}$ and $\mathbf{W}_{\mathrm{a}}^{(3)} \cdot \mathbf{W}_{\mathrm{s}}^{(3)} = \mathbf{I}$, where $\mathbf{I}$ denotes the 8×8 identity matrix. This is the linear algebra formulation of perfect reconstruction, or of the invertibility of the three scale transform.

Another interesting property to note here is that the structures of the two matrices are quite similar. The transpose of the one matrix bears a striking resembles the other matrix. In fact, if the above computations of the direct and inverse transforms of the canonical basis vectors had used the normalization steps, we would get an orthogonal matrix and its transpose, which by elementary linear algebra equals its inverse. Any wavelet transform with this property is called an orthogonal transform (i. e. all columns are orthogonal to each other and all have norm 1). All transforms in the Daubechies family have this property, and so do other well-known families like Coiflets and Symlets. The CDF families are not orthogonal (but rather biorthogonal), and consequently the structure of the CDF matrices is slightly more complicated.

It is clear that analogous constructions can be carried out for signal of length 2^j and transforms to all scales $k = 1, \ldots, j$. The linear algebra point of view is useful in understanding the theory. However, if one is trying to carry out numerical computations, then it is a bad idea to use the matrix formulation. The direct k scale wavelet transform using the lifting steps requires a number of operations (additions, multiplications, etc.) on the computer, which is proportional to the length L of the signal. If we perform the transform using its matrix, then in general a number of operations proportional to L^2 is needed. For longer signals this makes a huge difference for the computation time.

From Transform to Wavelets

Let us now consider the column vectors in (31) as a sampling of continuous, piecewise constant signals on the interval $[0; 1]$. These continuous signals are shown in Fig. 8. In this figure the signals have been group according to their origin. Recall that the first column came from a three scale synthesis of $[1, 0, 0, 0, 0, 0, 0, 0]$. This is equivalent to letting $\mathbf{s}_0 = 1$ and $\mathbf{d}_0 = 0$, $\mathbf{d}_1 = \mathbf{0}$, and $\mathbf{d}_2 = \mathbf{0}$ and using a three scale inverse DWT as shown in Fig. 7. The four right-most graphs in Fig. 8 can be obtained by inverse transforms where $\mathbf{d}_2$ equals $[1, 0, 0, 0]$, $[0, 1, 0, 0]$, $[0, 0, 1, 0]$, and $[0, 0, 0, 1]$, respectively, and $\mathbf{s}_0 = 0$, $\mathbf{d}_0 = 0$, and $\mathbf{d}_1 = \mathbf{0}$ in all four cases.

The $\mathbf{s}$ group consists of only one signal, which is constant. The $\mathbf{d}$ groups all consists of signals with the same shape (positive signal value followed by negative signal value). The three $\mathbf{d}$ differ by a factor 2 scaling in the x-axis direction, and the signals the $\mathbf{d}_1$ and $\mathbf{d}_2$ groups differ by translation along the x-axis.

This pattern can be contained in a single equation. First, define the function

$$\psi(t) = \begin{cases} 1, & t \in [0; \frac{1}{2}[\,, \\ -1, & t \in [\frac{1}{2}; 1]\,, \\ 0, & \text{otherwise}\,. \end{cases} \tag{33}$$

This is in fact the $\mathbf{d}_0$ signal. Now all the other $\mathbf{d}$ signal can be obtained by

$$\psi_{j,k}(t) = \psi(2^j t - k)\,,$$

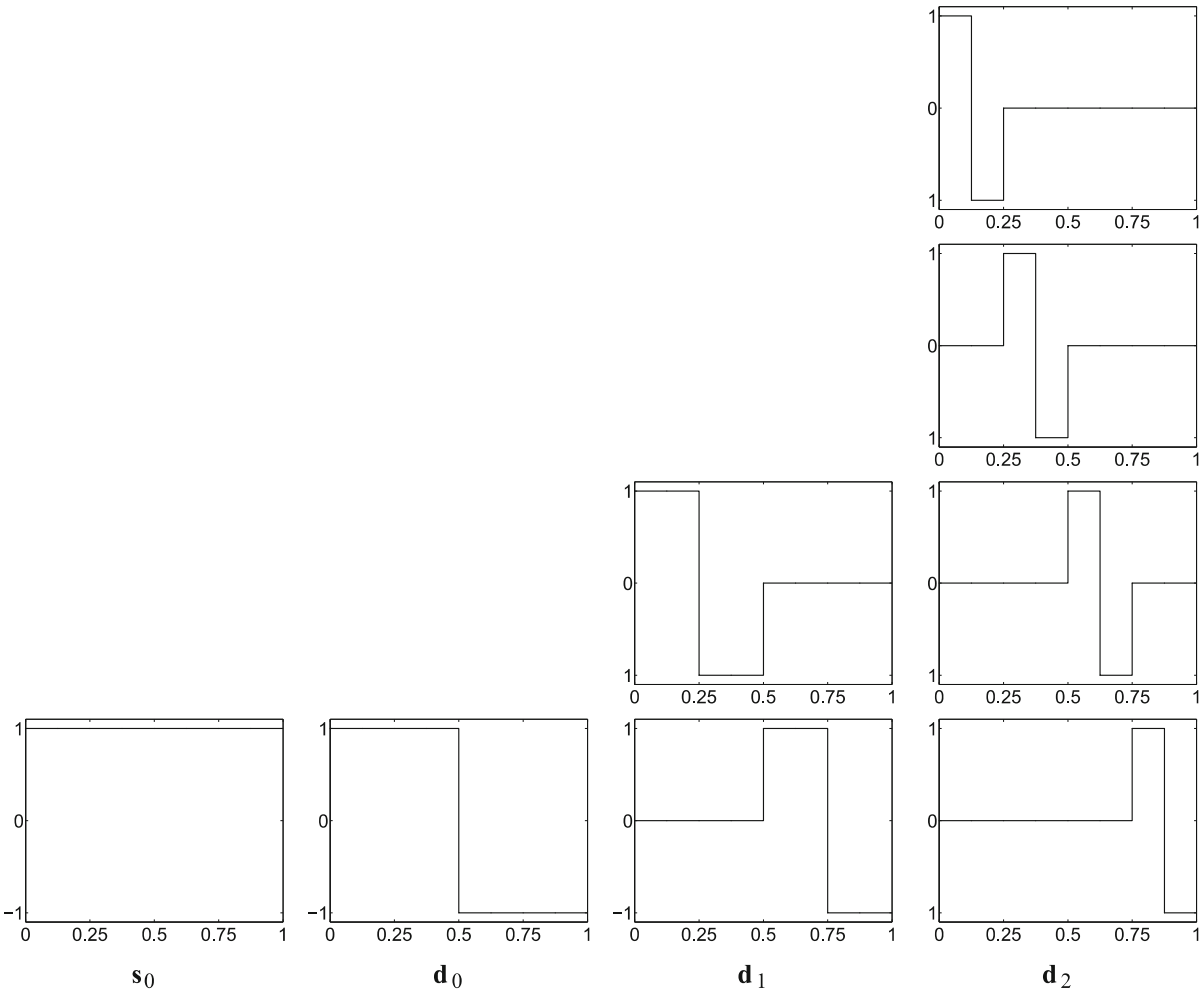

Wavelets and the Lifting Scheme, Figure 8

Graphs showing the 8 column vectors in (31) on the interval [0; 1]. The graphs are sorted according to the multiscale decomposition shown in Fig. 7

where j is equal to 0, 1, or 2 for the groups $\mathbf{d}_0$, $\mathbf{d}_1$, and $\mathbf{d}_2$, and k ranges from 0 to $2^j - 1$. The ψ function is the wavelet. Sometimes it is called the mother wavelet to signify that a wavelet transform consists of numerous similar functions originating from one function. If we include the normalization step the equation reaches its final form

$$\psi_{j,k}(t) = 2^{-j/2}\psi(2^j t - k) .$$

This framework does not include the $\mathbf{s}_0$. However, another equation combines $\mathbf{s}_0$ and $\mathbf{d}_0$. We noticed previously that while the vector $[1, 1, 1, 1, 1, 1, 1, 1]$, the first column in (31), represents the mean of the original signal, the vector $[1, 1, 1, 1, -1, -1, -1, -1]$ is mean of the first half of the original signal minus the mean of the other half of the signal. This link can be expressed with continuous functions. First, define

$$\phi(t) = \begin{cases} 1, & t \in [0; 1] , \\ 0, & \text{otherwise} , \end{cases}$$

that is, the function corresponding to the continuous $\mathbf{s}_0$ signal. This function is called the scaling function in wavelet literature. Now we can link $\mathbf{s}_0$ and $\mathbf{d}_0$ with

$$\psi(t) = \phi(2t) - \phi(2t - 1) . \tag{34}$$

This equation is called the two-scale equation, and is found in virtually any literature presenting the wavelet theory. Its general form is

$$\psi(t) = \sum_n g_n \phi(2t - n) , \tag{35}$$

and it is valid for all orthogonal wavelet transforms. For

the Haar transform we have $g_0 = 1$ and $g_1 = -1$ and $g_n = 0$ for all indices other than 0 and 1. The g_n for the Daubechies 4 transform are

$$g_0 = -\frac{1-\sqrt{3}}{4\sqrt{2}}\,, \quad g_1 = \frac{3-\sqrt{3}}{4\sqrt{2}}\,,$$
$$g_2 = -\frac{3+\sqrt{3}}{4\sqrt{2}}\,, \quad g_3 = \frac{1+\sqrt{3}}{4\sqrt{2}}\,.$$

There is also a two-scale equation that produces the scaling function.

$$\phi(t) = \sum_n h_n \phi(2t-n)\,. \tag{36}$$

For orthogonal transforms **h** is obtained from **g** by $h_n = (-1)^n g_{1-n}$, that is, reverse sequence order with alternating change of sign and index shift by 1.

Wavelet and Scaling Function for Daubechies 4

The appearance of the Haar wavelet function through transform synthesis in the previous section may appear to be a special property of the simple Haar transform. However, the principle applies to all wavelets, and the method for generating the wavelet function can be used to obtain any wavelet, at least as a graph. Actually, in many cases no closed form of the wavelet exist. This includes the Daubechies family.

Thus, to see the shape of the Daubechies 4 wavelet we start again with an 'all but one entry is zero'-signal. We saw that in fact when obtaining the wavelet graph it does not matter which entry is non-zero, as long as it is not the first entry. The only difference in the output is the dilation and translation of the resulting function. As an example we therefore choose [0, 0, 0, 0, 0, 1, 0, 0]. We perform a inverse three scale transform using the inverse Daubechies 4 Equations (18)–(22) and plot the result, see Fig. 8. Note that the problem with finite signals needs to be addressed in this inverse transform. For our purpose zero padding is sufficient.

This figure contains very little information. But let us now repeat the procedure for vectors of length 8, 32, 128, and 512, applied to a vector with a single 1 as its sixth entry. This requires inverse transforms over 3, 5, 7, and 9 scales, respectively. We fit each transform to the same interval, as if we have a finer and finer resolution. The result is shown in Fig. 9.

This figure shows that the graphs rapidly approach a limiting graph, as we increase the length of the vector. This is a result that can be established rigorously, but it is not easy to do so, and it is beyond the scope of this article.

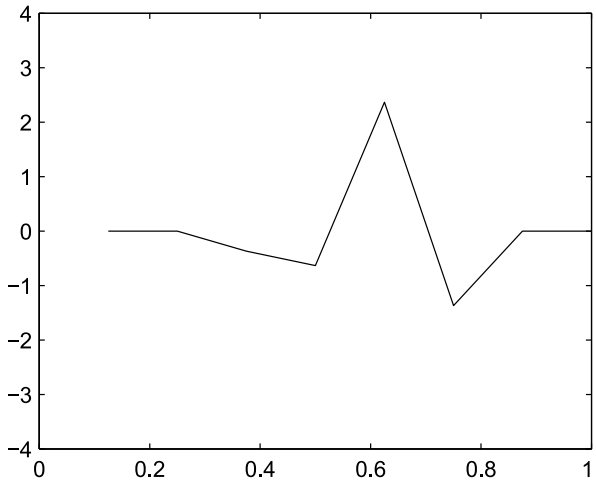

Wavelets and the Lifting Scheme, Figure 9
Inverse Daubechies 4 of [0, 0, 0, 0, 0, 1, 0, 0] over three scales

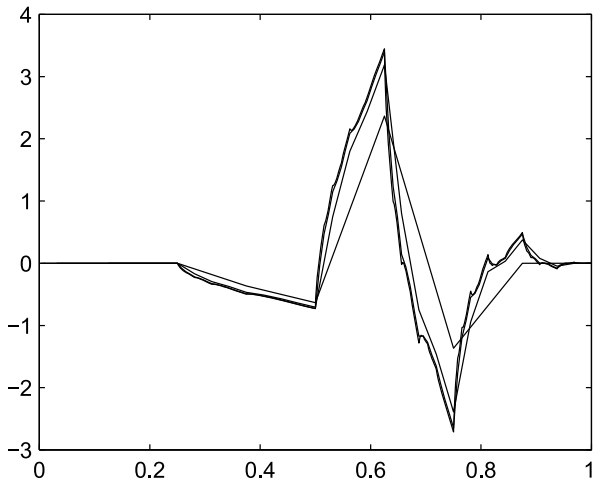

Wavelets and the Lifting Scheme, Figure 10
Inverse Daubechies 4 of sixth basis vector, length 8, 32, 128 and 512

One can interpret the limiting function in Fig. 9 as a function whose values, sampled at appropriate points, represent the entries in the inverse transform of a vector of length 2^N, with a single 1 as its sixth entry. For N just moderately large, say $N = 12$, this is a very good approximation to the actual value. See Fig. 10 for the result for $N = 12$, i. e. a vector of length 4096.

Wavelet and Scaling Function for CDF(2,2)

Finally, let us repeat the procedure for the CDF(2,2) transform in (10) and (11), and at the same time illustrate how the wavelet is translated depending on the placement of

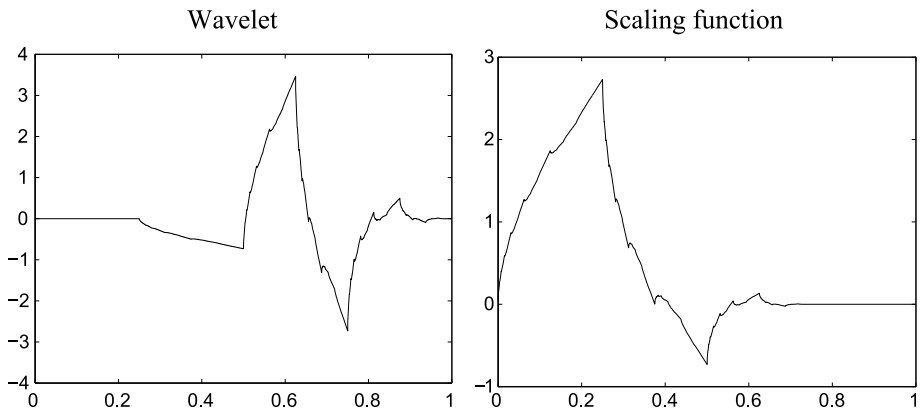

Wavelets and the Lifting Scheme, Figure 11
Inverse Daubechies 4 of sixth and first basis vectors, both length 4096. The result is the Daubechies 4 wavelet on the *left* and the scaling function on the *right*

the 1 in the otherwise 0 vector. An example is given in Fig. 11. The difference in the graphs in Fig. 11 and Fig. 10 is striking. It reflects the result that the Daubechies 4 wavelet has very little regularity (it is not differentiable), whereas the CDF(2,2) wavelet is a piecewise linear function.

Recall that the Haar and Daubechies 4 transform are orthogonal transforms, which means that in linear algebra terms the transpose of the synthesis matrix equals the analysis matrix. A consequence of this is that the same wavelet and scaling function are used for analysis and for synthesis. This is not the case for CDF(2,2), which is a biorthogonal transform. Therefore, it has one synthesis pair of scaling function and wavelets, and an analysis pair. If we use the forward transform instead of the inverse we can obtain the analysis functions. They are shown in Fig. 12. These functions are quite complicated, and it is interesting to see that while the analysis wavelet and scaling function are very simple functions (we have not shown the scaling function of CDF(2,2) though) the inverse of that same transform have some rather complex wavelet and scaling functions.

As an end remark notice that all CDF(2,2) functions are symmetrical. This is a special property that can only be achieved by biorthogonal wavelets. Orthogonal wavelets can come close to symmetry, but they can never become completely symmetrical. An almost symmetrical family has been constructed, called the Symlets, see for example [4].

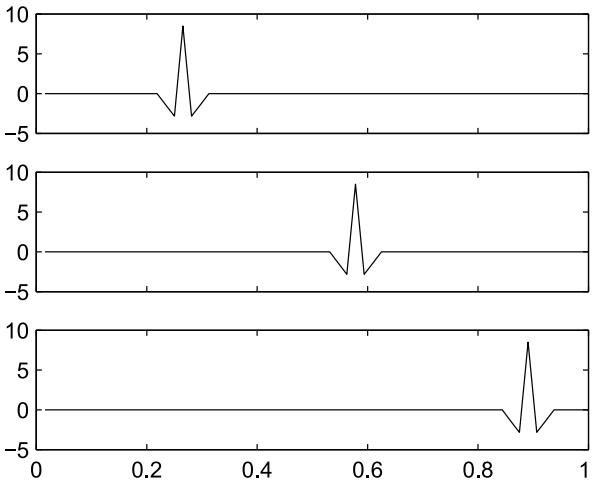

Wavelets and the Lifting Scheme, Figure 12
Inverse CDF(2,2) of three basis vectors of length 64, entry 40, or 50, or 60, equal to 1 and the remaining entries equal to zero. The result is the same function (the wavelet) with different translations

The General Case

The above computations lead us to the conclusion that there are just two functions underlying the direct transform, and another two functions underlying the inverse transform, in the sense that if we take sufficiently long vectors, say 2^N, and perform a k scale transform, with k large, then we get values that are sampled values of one of the underlying functions. More precisely, inverse transforms of unit vectors with a one in places from 1 to 2^{N-k} yield translated copies of the scaling function. Inverse transforms of unit vectors with a one in places from $2^{N-k}+1$ to 2^{N-k+1} yield translated copies of the wavelet. Finally, inverse transforms of unit vectors with a one at places from $2^{N-k+1}+1$ to 2^N yield scaled and translated copies of the wavelet.

These results are strictly correct only in a limiting sense, and they are not easy to establish. There is one fur-

ther complication which we have omitted to state clearly. If one performs the procedure above with a 1 close to the start or end of the vector, then there will in general be some strange effects, depending on how the transform has been implemented. We refer to these as boundary effects. They depend on how one makes up for missing samples in computations near the start or end of a finite vector, the so-called boundary corrections. We have already mentioned zero padding as one of the correction methods. This is what we have used indirectly in plotting for example Fig. 10, where we have taken a vector with a one at place 24, and zeroes everywhere else.

Readers interested in a rigorous treatment of the interpretation of the transforms given here, and with the required mathematical background, are referred to the literature, for example the books by I. Daubechies [1], S. Mallat [4], and M. Vetterli-J. Kovačević [10,11]. Note that these books base their treatment of the wavelet transforms on the concepts of multiresolution analysis and filter theory rather than lifting.

Lifting and Filter Banks

So far, the lifting technique has been carried out in the time domain; the correlation of samples was based on odd and even entries, interpretation was based on the signal in the time domain, and every reference to the signal was made by **s** and **d**, i. e.time representations of the signals. However, there is much to be learned about wavelets and about lifting when turning to the frequency domain.

The lifting method is a special case of a standard frequency-based signal processing method called filter banks. A filter bank is a series of bandpass filters which separates the input signal into a number of components, each with a distinct range of frequencies from the original signal. Often, in a filter bank, the filters are designed such that no information is lost, that is, it is possible to reconstruct the original signal from the components. The outputs of a filter bank are called subband signals, and it has as many subbands as there are filters in the filter bank.

The following introduction to lifting in the frequency domain requires some knowledge of the z-transform as well as basic Fourier theory. To ease the reading, we briefly state some important relations for the z-transform. The transform maps a signal $\mathbf{x} = \{x[n]\}$ to a function defined in the complex plane

$$X(z) = \sum_{n \in \mathbf{Z}} x[n] z^{-n} .$$

This is equivalent to the Fourier transform, when $z = e^{j\omega}$, and we use capital X to denote the z-transform of $\mathbf{x}$. In the following exposition we will needed up and down sampling by two of a signal. In the time domain this is accomplished by inserted zeros between all samples and by removing every other sample, respectively. Explicitly, if $\mathbf{x} = \{x[n]\}$, then the down sampled sequence $\mathbf{x}_{2\downarrow}$ is given by $x_{2\downarrow}[n] = x[2n]$. The up sampled sequence $\mathbf{x}_{2\uparrow}$ is given by $x_{2\uparrow}[n] = x[n/2]$, if n is even, and $x_{2\uparrow}[n] = 0$, if n is odd.

In the z-transform up sampling of $X(z)$ is accomplished by $X(z^2)$, and down sampling is accomplished by

$$\tfrac{1}{2}\big(X(z^{1/2}) + X(-z^{1/2})\big) . \tag{37}$$

Finally, we note that in this section all signals have finite length, and thus the z-transform has a finite number of non-zero terms. The z-transform is therefore a Laurent polynomial, which is a polynomial in the variables z and z^{-1}.

Lifting in the z-Transform Representation

The lifting method consists of a three basic operations; it is composed of splitting the signal in odd and even samples, predicting an odd sample, and updating an even sample. This is shown in Fig. 3. The splitting is given in the z-transform representation by

$$X(z) = X_0(z^2) + z^{-1} X_1(z^2) , \tag{38}$$

where

$$X_0(z) = \sum_n x[2n] z^{-n} = \tfrac{1}{2}\big(X(z^{1/2}) + X(-z^{1/2})\big) , \tag{39}$$

$$X_1(z) = \sum_n x[2n+1] z^{-n} = \tfrac{1}{2} z^{1/2}\big(X(z^{1/2}) - X(-z^{1/2})\big) , \tag{40}$$

are the z-transforms of the even and odd samples. Note that the second equality in both formulas comes from (37). We represent this decomposition by the left side of the diagram in Fig. 13. In the diagram the time shift is inserted prior to the down sampling, rather than after as in (40), since this gives multiplication with z instead of $z^{1/2}$. The decomposition in (38) is called a polyphase decomposition (with two components).

The inverse operation is obtained by reading Eq. (38) from right to left. The equation tells us that we can obtain $X(z)$ from $X_0(z)$ and $X_1(z)$ by first up sampling the two components by 2, then shifting $X_1(z^2)$ one time unit right (by multiplication by z^{-1}), and finally adding the two components. We represent this reconstruction by the

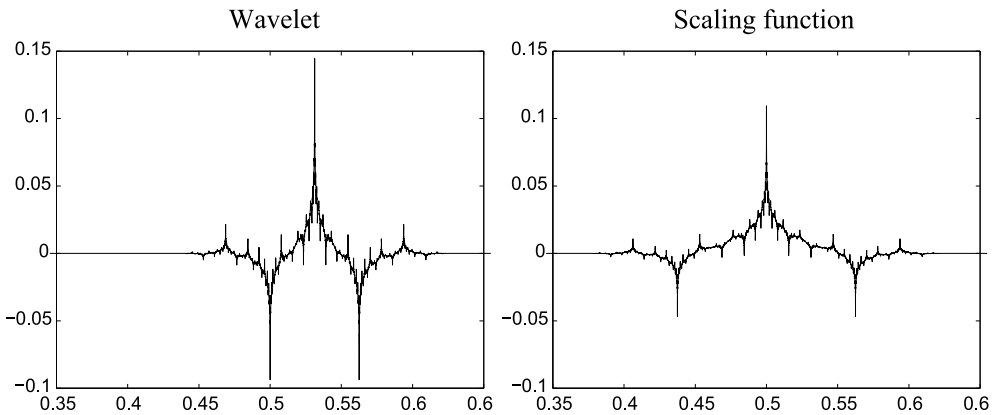

Wavelets and the Lifting Scheme, Figure 13
Wavelet and Scaling function for CDF(2,2)

right hand side of the diagram in Fig. 13. Notice that as with the lifting method this inversion simply 'undoes' the changes made to the signal. The transform pair represented in Fig. 13 is sometimes in the literature called the 'lazy' wavelet transform and its inverse.

Prediction and Update Steps Let us now see how we can implement the prediction step from Sect. "Lifting" in the z-transform representation. The prediction technique was to form a linear combination of the even entries and then subtract the result from the odd entry under consideration. The linear combination was formed independently of the index of the odd sample under consideration, and based only on the relative location of the even entries. For example, in the CDF(2,2) transform the first step in (10) can be implemented as $X_1(z) - T(z)X_0(z)$, where $T(z) = \frac{1}{2}(1+z)$. This is because $T(z)X_0(z)$ is the convolution $\mathbf{t} * \mathbf{x}_0$ in the time domain, which is exactly a linear combination of the even entries with weights t_n. Explicitly,

$$\begin{aligned} X_1(z) - T(z)X_0(z) &= X_1(z) - \tfrac{1}{2}(1+z)\sum_n x[2n]z^{-n} \\ &= X_1(z) - \tfrac{1}{2}\sum_n x[2n]z^{-n} + \tfrac{1}{2}\sum_n x[2n]z^{-n+1} \\ &= \sum_n x[2n+1] - \sum_n \tfrac{1}{2}(x[2n] + x[2n+2])z^{-n} \\ &= \sum_n \left(x[2n+1] - \tfrac{1}{2}(x[2n] + x[2n+2])\right) z^{-n}\,, \end{aligned}$$

which is the z-transform representation of the right hand side of (10). The transition is described by matrix multiplication as in

$$\begin{bmatrix} X_0(z) \\ X_1(z) - T(z)X_0(z) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -T(z) & 1 \end{bmatrix} \begin{bmatrix} X_0(z) \\ X_1(z) \end{bmatrix}.$$

Here we use 2×2 matrices whose entries are Laurent polynomials. An entirely analogous computation shows that if we define $S(z) = \frac{1}{4}(1 + z^{-1})$, then the update step in (11) is implemented in the z-transform representation as multiplication by the matrix

$$\begin{bmatrix} 1 & S(z) \\ 0 & 1 \end{bmatrix}.$$

The normalization step, as for example given in (26) and (25), can be implemented by multiplication by a matrix of the form

$$\begin{bmatrix} K & 0 \\ 0 & K^{-1} \end{bmatrix},$$

where $K > 0$ is a constant.

Entire Transform Since every prediction step is subtraction of a linear combination of odd samples, any prediction step can be implemented by means of multiplication by a matrix of the form

$$\mathbf{P}(z) = \begin{bmatrix} 1 & 0 \\ -T(z) & 1 \end{bmatrix}. \tag{41}$$

The same result applies to any update step, which can be implemented by multiplication by a matrix of the form

$$\mathbf{U}(z) = \begin{bmatrix} 1 & S(z) \\ 0 & 1 \end{bmatrix}. \tag{42}$$

Here $T(z)$ and $S(z)$ are both Laurent polynomials.

The general one scale DWT is a combination of multiple prediction and updates steps and one normalization step. This is shown in Fig. 6 (although the normalization

is not included in the figure). In the z-transform representation this becomes

$$\mathbf{H}(z) = \begin{bmatrix} K & 0 \\ 0 & K^{-1} \end{bmatrix} \begin{bmatrix} 1 & S_N(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -T_N(z) & 1 \end{bmatrix} \cdots \begin{bmatrix} 1 & S_1(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -T_1(z) & 1 \end{bmatrix} . \quad (43)$$

The order of the factors is determined by the order in which the steps are applied. First a prediction step, then an update step, repeated N times, and then finally the normalization step.

An important property of the DWT implemented via lifting steps was the invertibility of the transform, as illustrated for example in Fig. 4. It is easy to verify that we have

$$\mathbf{P}(z)^{-1} = \begin{bmatrix} 1 & 0 \\ T(z) & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{U}(z)^{-1} = \begin{bmatrix} 1 & -S(z) \\ 0 & 1 \end{bmatrix} . \quad (44)$$

We note that the inverse matrix is again a matrix with Laurent polynomials as entries. Thus, the inversion of the transform is accomplished simply by inverting each 2×2 matrix and reversing their order.

$$\mathbf{H}^{-1}(z) = \mathbf{G}(z) = \begin{bmatrix} 1 & 0 \\ T_1(z) & 1 \end{bmatrix} \begin{bmatrix} 1 & -S_1(z) \\ 0 & 1 \end{bmatrix} \cdots \begin{bmatrix} 1 & 0 \\ T_N(z) & 1 \end{bmatrix} \begin{bmatrix} 1 & -S_N(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} K^{-1} & 0 \\ 0 & K \end{bmatrix} . \quad (45)$$

Multiplying out the matrices in the product defining $\mathbf{H}(z)$ in (43), we get a 2×2 matrix with entries, which are Laurent polynomials. We use the notation

$$\mathbf{H}(z) = \begin{bmatrix} H_{00}(z) & H_{01}(z) \\ H_{10}(z) & H_{11}(z) \end{bmatrix} , \quad (46)$$

for such a general matrix. Written in matrix notation the DWT is given as

$$\begin{bmatrix} Y_0(z) \\ Y_1(z) \end{bmatrix} = \begin{bmatrix} H_{00}(z) & H_{01}(z) \\ H_{10}(z) & H_{11}(z) \end{bmatrix} \begin{bmatrix} X_0(z) \\ X_1(z) \end{bmatrix} , \quad (47)$$

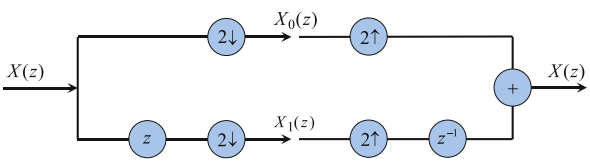

Wavelets and the Lifting Scheme, Figure 14
Splitting in even and odd components by means of up sampling followed by reconstruction of the signal from the odd and even components

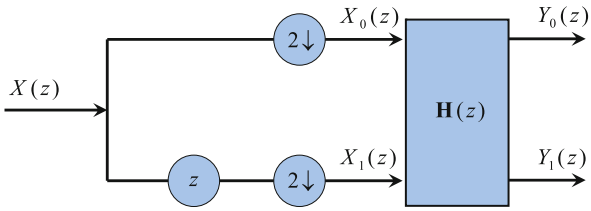

Wavelets and the Lifting Scheme, Figure 15
One scale DWT in the z-representation as given in (47). This is the polyphase representation and it comes directly from the lifting step method

and we can then represent the implementation of the complete one scale DWT in the z-transform representation by the diagram in Fig. 15.

This representation of a two channel filter bank, without any reference to lifting, is in the signal analysis literature called the polyphase representation, see for example [7,10,11].

Two Channel Filter Banks with Perfect Reconstruction

We have now seen how the time-based lifting method has an equivalent representation in the frequency domain. Lifting allows for perfect reconstruction of the signal after transformation, and we will now use this property (which is preserved in the frequency representation) to better understand the link between lifting and filter banks. We assume the reader is familiar with the concept of filtering and filter banks. Details on filtering can be found in any book on signal analysis, for example in [6,7]. The filter bank approach to wavelets is the one used in most introductions to the subject.

A two channel filter bank starts with two analysis filters, here denoted by the filter taps $\mathbf{h}_0$ and $\mathbf{h}_1$, and two synthesis filters, denoted by $\mathbf{g}_0$ and $\mathbf{g}_1$. Usually the filters with index 0 are chosen to be low pass filters, and the filters with index 1 to be high pass filters. The analysis and synthesis parts of the filter bank are shown in Fig. 15.

The analysis part transforms the input $X(z)$ to the output pair $Y_0(z)$ and $Y_1(z)$. The synthesis part then transforms this pair to the output $\widetilde{X}(z)$. The filtering scheme is said to have the perfect reconstruction property, if $X(z) = \widetilde{X}(z)$ for any $X(z)$.

The main point that we will now establish is that the two Figs. 15 and 16 do indeed show the same thing, and that it is possible to get the one from the other.

Conditions for Perfect Reconstruction To do this we first analyze which conditions are needed on the four filters in Fig. 15 in order to obtain the perfect reconstruction

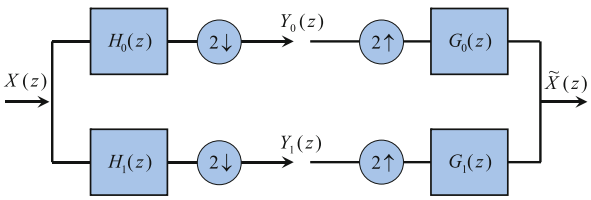

Wavelets and the Lifting Scheme, Figure 16
Two channel analysis and synthesis. This is the standard filter bank representation

property. Filtering by $\mathbf{h}_0$ transforms $X(z)$ to $H_0(z)X(z)$, and we then use (38) to down sample by two. Thus we have

$$Y_0(z) = \tfrac{1}{2}\Big(H_0(z^{1/2})X(z^{1/2}) + H_0(-z^{1/2})X(-z^{1/2})\Big)\,, \quad (48)$$

$$Y_1(z) = \tfrac{1}{2}\Big(H_1(z^{1/2})X(z^{1/2}) + H_1(-z^{1/2})X(-z^{1/2})\Big)\,. \quad (49)$$

Up sampling by two followed by filtering by the G-filters, and addition of the results leads to a reconstructed signal

$$\widetilde{X}(z) = G_0(z)Y_0(z^2) + G_1(z)Y_1(z^2)\,. \quad (50)$$

Perfect reconstruction means that $\widetilde{X}(z) = X(z)$. We combine the above expressions and then regroup terms to get

$$\begin{aligned}\widetilde{X}(z) = \;& \tfrac{1}{2}\big[G_0(z)H_0(z) + G_1(z)H_1(z)\big]X(z) \\ &+ \tfrac{1}{2}\big[G_0(z)H_0(-z) + G_1(z)H_1(-z)\big]X(-z)\,.\end{aligned}$$

To fulfill the condition $\widetilde{X}(z) = X(z)$ we need

$$G_0(z)H_0(z) + G_1(z)H_1(z) = 2\,, \quad (51)$$

$$G_0(z)H_0(-z) + G_1(z)H_1(-z) = 0\,. \quad (52)$$

These conditions mean that the four filters cannot be chosen independently, if we want to have perfect reconstruction. To determine what conditions these equations impose on H and G it is convenient to write them as a matrix equation,

$$\begin{bmatrix} H_0(z) & H_1(z) \\ H_0(-z) & H_1(-z) \end{bmatrix}\begin{bmatrix} G_0(z) \\ G_1(z) \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}. \quad (53)$$

We want to solve this equation, that is, determine $G_0(z)$ and $G_1(z)$ as functions of $H_0(z)$ and $H_1(z)$. Since the matrix is invertible (it performs a DWT that can be inverted), we can use Cramer's rule to get

$$G_0(z) = \frac{\begin{vmatrix} 2 & H_1(z) \\ 0 & H_1(-z) \end{vmatrix}}{d(z)} = 2d(z)^{-1}H_1(-z)\,.$$

The determinant $d(z)$ of an invertible 2×2 matrix with Laurent polynomials as entries is a monomial (see [3]), which in this case has the property that $d(-z) = -d(z)$. Therefore, the determinant can be assumed to be of the form $d(z) = \frac{1}{2}C^{-1}z^{-2k-1}$ for some integer k and some real constant C. Thus,

$$G_0(z) = \frac{\begin{vmatrix} 2 & H_1(z) \\ 0 & H_1(-z) \end{vmatrix}}{\frac{1}{2}C^{-1}z^{-2k-1}} = Cz^{2k+1}H_1(-z)\,, \quad (54)$$

$$G_1(z) = \frac{\begin{vmatrix} H_0(z) & 2 \\ H_0(-z) & 0 \end{vmatrix}}{\frac{1}{2}C^{-1}z^{-2k-1}} = -Cz^{2k+1}H_0(-z)\,. \quad (55)$$

These equations show that we can choose either the H-filter pair or the G-filter pair. We will assume that we have filters H_0 and H_1, subject to the condition that

$$\begin{aligned} d(z) &= H_0(z)H_1(-z) - H_0(-z)H_1(z) \\ &= \tfrac{1}{2}C^{-1}z^{-2k-1} \end{aligned} \quad (56)$$

for some integer k and nonzero constant C. Then G_0 and G_1 are determined by the Eqs. (54) and (55), which implies that they are unique up to a scaling factor and an odd shift in time. Note that this argument is valid for any two channel filter bank, subject to the condition (56).

Lifting and Two Channel Filter Banks with Perfect Reconstruction Are Equivalent

Many presentations of the wavelet transform start with a two channel filter bank with the perfect reconstruction property. The analysis part is then used to define the direct one scale DWT, and the synthesis part is used for reconstruction. It is an important result that the filtering approach, and the one based on lifting, actually are identical. Thus any set of lifting steps leads to a two channel filter bank with perfect reconstruction, and, quite remarkably, vice versa. This means that they are just two different ways of describing the same transformation from $X(z)$ to $Y_0(z)$ and $Y_1(z)$. This equivalence will be the subject of the remainder of the article.

The first step is to show that the analysis step in Fig. 15 coming from the two channel filter bank is equivalent to the analysis step summarized in Fig. 14 and in (47), i. e. coming from the lifting method. Thus, we want to find the equations relating the coefficients in the $\mathbf{H}$ matrix (46) and the filters H_0, H_1, G_0, and G_1. The analysis step by both methods should yield the same result. To avoid the square root terms we compare the results after up sampling by two. We start with the equality

$$\begin{bmatrix} Y_0(z^2) \\ Y_1(z^2) \end{bmatrix} = \mathbf{H}(z^2)\begin{bmatrix} \frac{1}{2}\left(X(z) + X(-z)\right) \\ \frac{1}{2}z\left(X(z) - X(-z)\right) \end{bmatrix},$$

where the left hand side is from the filter bank approach (48) and (49), and the right hand side from the lifting approach with polyphase matrix (46) and up sampled $X_0(z)$ and $X_1(z)$. The first equation can then be written as

$$\tfrac{1}{2}\big(H_0(z)X(z) + H_0(-z)X(-z)\big) = H_{00}(z^2)\tfrac{1}{2}\big(X(z)+X(-z)\big)+H_{01}(z^2)\tfrac{1}{2}z\big(X(z)-X(-z)\big).$$

This leads to the relation

$$H_0(z) = H_{00}(z^2) + zH_{01}(z^2)\,.$$

The relation for H_1 is found analogously, and then the relations for G_0 and G_1 can be found using the perfect reconstruction conditions (54) and (55) in the two cases. The results are summarized here.

$$H_0(z) = H_{00}(z^2) + zH_{01}(z^2)\,, \tag{57}$$

$$H_1(z) = H_{10}(z^2) + zH_{11}(z^2)\,, \tag{58}$$

$$G_0(z) = G_{00}(z^2) + z^{-1}G_{01}(z^2)\,, \tag{59}$$

$$G_1(z) = G_{10}(z^2) + z^{-1}G_{11}(z^2)\,. \tag{60}$$

Note the difference in the decomposition of the H-filters and the G-filters. Thus in the polyphase representation we use

$$\mathbf{H}(z) = \begin{bmatrix} H_{00}(z) & H_{01}(z) \\ H_{10}(z) & H_{11}(z) \end{bmatrix},$$

$$\mathbf{G}(z) = \mathbf{H}(z)^{-1} = \begin{bmatrix} G_{00}(z) & G_{10}(z) \\ G_{01}(z) & G_{11}(z) \end{bmatrix}.$$

Note the placement of entries in $\mathbf{G}(z)$, which differs from the usual notation for matrices. The requirement of perfect reconstruction in the polyphase formulation was the requirement that $\mathbf{G}(z)$ should be the inverse of $\mathbf{H}(z)$.

From Filters to Lifting Steps It is easy to start with a filter bank and derive $\mathbf{H}(z)$ using (57) and (58). However, going the other way is somewhat more tricky; given $\mathbf{H}(z)$ it is necessary to factorize it into a number of 2×2 matrices of a particular structure to obtain the lifting steps. The remarkable result is that this is always possible. This result was obtained by I. Daubechies and W. Sweldens in 1998 in the paper [2]. The factorization result for 2×2 matrices with Laurent entries was previously known in the mathematical literature. The importance of the paper by I. Daubechies and W. Sweldens lies in its impact on signal processing.

Theorem 1(Daubechies and Sweldens 1998) *Assume that* $\mathbf{H}(z)$ *is a* 2×2 *matrix of Laurent polynomials, normalized to* $\det \mathbf{H}(z) = 1$. *Then there exists a constant* $K \neq 0$ *and Laurent polynomials* $S_1(z), \ldots, S_N(z)$, $T_1(z), \ldots, T_N(z)$, *such that*

$$\mathbf{H}(z) = \begin{bmatrix} K & 0 \\ 0 & K^{-1} \end{bmatrix} \begin{bmatrix} 1 & S_N(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ T_N(z) & 1 \end{bmatrix} \cdots \begin{bmatrix} 1 & S_1(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ T_1(z) & 1 \end{bmatrix}. \tag{61}$$

The normalization $\det \mathbf{H}(z) = 1$ in the theorem can always be satisfied. As seen above, in the general case $\det \mathbf{H}(z) = Cz^{2k+1}$. One can always get the determinant equal to one by scaling and an odd shift in time. Therefore the result is stated with $\det \mathbf{H}(z) = 1$ without loss of generality.

The proof of this theorem is constructive. It gives an algorithm for finding the Laurent polynomials $S_1(z), \ldots, S_N(z)$, $T_1(z), \ldots, T_N(z)$ in the factorization. It is important to note that the factorization is not unique. Once we have a factorization, we can translate it into lifting steps.

The advantage of the lifting approach, compared to the filter approach, is that it is very easy to find perfect reconstruction filters H_0, H_1, G_0, and G_1. It is just a matter of multiplying the lifting steps as in (43), and then assemble the filters according to the Eqs. (57)–(60). There can also be algorithmic advantages in implementing a transform using lifting steps, since it may result in fewer arithmetic operations than a filter implementation.

This approach should be contrasted with the traditional signal analysis approach, where one tries to find (approximate numerical) solutions to the Eqs. (51) and (52), using for example spectral factorization. The weakness in constructing a transform based solely on the lifting technique is that it is based entirely on considerations in the time domain. Sometimes it is desirable to design filters with certain properties in the frequency domain, and once filters have been constructed in the frequency domain, we can use the constructive proof of the theorem to derive a lifting implementation.

We should mention that the numerical stability of transforms designed using lifting can be difficult to analyze.

Examples of Lifting Steps in the Frequency Domain

We will now see how the Daubechies 4 transform looks in the frequency domain, that is, what frequency response we would get when passing a signal through the Daubechies

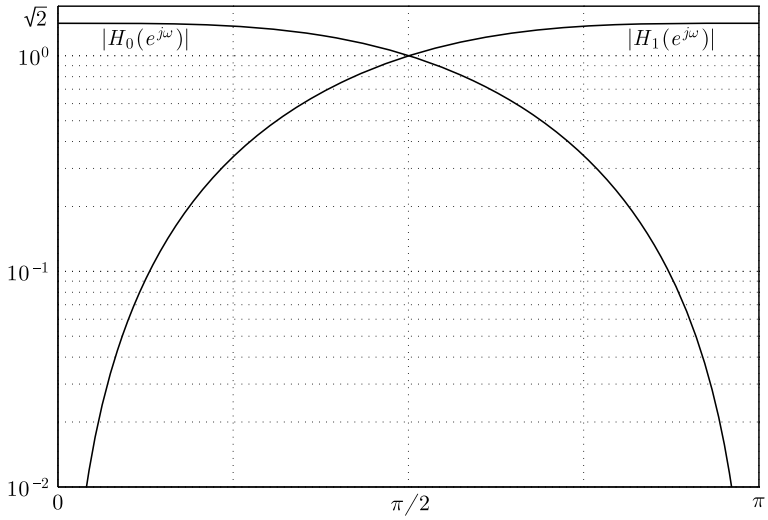

Wavelets and the Lifting Scheme, Figure 17
Frequency response of the Daubechies 4 lifting steps

4 transform. First, we need to find $\mathbf{H}(z)$ by multiplying the lifting steps originating in the Daubechies 4 Eqs. (13) through (17). Each equation has an equivalent lifting step matrix (41) or (42). The first of the Daubechies 4 equations

$$s_{j-1}^{(1)}[n] = s_j[2n] + \sqrt{3}s_j[2n+1]$$

is an update step, and it simply multiplies the odd sample with $\sqrt{3}$ and add the result to the even sample. In the z-transform this is achieved by multiplying with

$$\begin{bmatrix} 1 & \sqrt{3} \\ 0 & 1 \end{bmatrix},$$

which makes this matrix the first transform step. The second equation

$$d_{j-1}^{(1)}[n] = s_j[2n+1] - \tfrac{1}{4}\sqrt{3}s_{j-1}^{(1)}[n] - \tfrac{1}{4}(\sqrt{3}-2)s_{j-1}^{(1)}[n-1]$$

is a prediction step that uses the present value (index n) and the first previous value (index $n-1$) of the signal. This is achieved in the z-transform by

$$\begin{bmatrix} 1 & 0 \\ -\frac{\sqrt{3}}{4} - \frac{\sqrt{3}-2}{4}z^{-1} & 1 \end{bmatrix}.$$

The third equation converts to z-transform similarly. The two scaling Equation (16) and (17) are joined in one scaling matrix

$$\begin{bmatrix} \frac{\sqrt{3}-1}{\sqrt{2}} & 0 \\ 0 & \frac{\sqrt{3}+1}{\sqrt{2}} \end{bmatrix}.$$

Now, the polyphase matrix $\mathbf{H}(z)$ can be written

$$\begin{aligned}
\mathbf{H}(z) &= \begin{bmatrix} \frac{\sqrt{3}-1}{\sqrt{2}} & 0 \\ 0 & \frac{\sqrt{3}+1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & -z \\ 0 & 1 \end{bmatrix} \\
&\quad \begin{bmatrix} 1 & 0 \\ -\frac{\sqrt{3}}{4} - \frac{\sqrt{3}-2}{4}z^{-1} & 1 \end{bmatrix} \begin{bmatrix} 1 & \sqrt{3} \\ 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} \frac{3-\sqrt{3}}{4\sqrt{2}}z + \frac{1+\sqrt{3}}{4\sqrt{2}} & \frac{1-\sqrt{3}}{4\sqrt{2}}z + \frac{3+\sqrt{3}}{4\sqrt{2}} \\ -\frac{3+\sqrt{3}}{4\sqrt{2}} - \frac{1-\sqrt{3}}{4\sqrt{2}}z^{-1} & \frac{1+\sqrt{3}}{4\sqrt{2}} + \frac{3-\sqrt{3}}{4\sqrt{2}}z^{-1} \end{bmatrix} \\
&= \begin{bmatrix} H_{00}(z) & H_{01}(z) \\ H_{10}(z) & H_{11}(z) \end{bmatrix}
\end{aligned}$$

Thus, from (57) and (58) it follows that the low and high pass filter taps (in z-transform) are given by

$$\begin{aligned}
H_0(z) &= H_{00}(z^2) + zH_{01}(z^2) \\
&= \frac{1+\sqrt{3}}{4\sqrt{2}} + \frac{3+\sqrt{3}}{4\sqrt{2}}z + \frac{3-\sqrt{3}}{4\sqrt{2}}z^2 + \frac{1-\sqrt{3}}{4\sqrt{2}}z^3, \\
H_1(z) &= H_{10}(z^2) + zH_{11}(z^2) \\
&= -\frac{1-\sqrt{3}}{4\sqrt{2}}z^{-2} + \frac{3-\sqrt{3}}{4\sqrt{2}}z^{-1} - \frac{3+\sqrt{3}}{4\sqrt{2}} + \frac{1+\sqrt{3}}{4\sqrt{2}}z.
\end{aligned}$$

To determine the frequency response of these transfer functions, we simply replace z with $e^{j\omega}$, which gives the Fourier transform of the filter taps. We then plot $|H_0(e^{j\omega})|$ and $|H_1(e^{j\omega})|$ as function of ω to show the amplitude characteristics of the transfer functions $H_0(z)$ and $H_1(z)$ on the unit circle in the complex plane. This plot is shown in Fig. 16.

Making Lifting Steps from Filters

There are basically three ways of representing the building block in a DWT: (1) The transform can be represented by a pair of filters (usually low pass and high pass filters) satisfying the perfect reconstruction conditions, or it can be given as lifting steps, which are either given (2) in the time domain as a set of equations, or (3) in the frequency domain as a factored matrix of Laurent polynomials.

As an example the Daubechies 4 transform can be given in these three forms, as shown below.

Equation form:

$$s^{(1)}[n] = S[2n] + \sqrt{3}S[2n+1]\,, \tag{62}$$

$$d^{(1)}[n] = S[2n+1] - \frac{1}{4}\sqrt{3}s^{(1)}[n] - \frac{1}{4}(\sqrt{3}-2)s^{(1)}[n-1]\,, \tag{63}$$

$$s^{(2)}[n] = s^{(1)}[n] - d^{(1)}[n+1]\,, \tag{64}$$

$$\tilde{s}[n] = \frac{\sqrt{3}-1}{\sqrt{2}}s^{(2)}[n]\,, \tag{65}$$

$$\tilde{d}[n] = \frac{\sqrt{3}+1}{\sqrt{2}}d^{(1)}[n]\,. \tag{66}$$

Matrix form:

$$\mathbf{H}(z) = \begin{bmatrix} \frac{\sqrt{3}-1}{\sqrt{2}} & 0 \\ 0 & \frac{\sqrt{3}+1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & -z \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\frac{\sqrt{3}}{4} - \frac{\sqrt{3}-2}{4}z^{-1} & 1 \end{bmatrix} \begin{bmatrix} 1 & \sqrt{3} \\ 0 & 1 \end{bmatrix}. \tag{67}$$

Filter form:

$$\mathbf{h} = \frac{1}{4\sqrt{2}}\left[1+\sqrt{3},\ \ 3+\sqrt{3},\ \ 3-\sqrt{3},\ \ 1-\sqrt{3}\right], \tag{68}$$

$$\mathbf{g} = \frac{1}{4\sqrt{2}}\left[1-\sqrt{3},\ \ -3+\sqrt{3},\ \ 3+\sqrt{3},\ \ -1-\sqrt{3}\right]. \tag{69}$$

The equation form was presented in (13) through (17). Converting this to the matrix form has not been shown directly. However, using (41), (42), and (43) it is not difficult to derive the matrices from the equations. It is not difficult either to obtain the filter taps **h** and **g** from the matrix form. Simply multiply the matrix to get the polyphase matrix (46), and the odd entries in **h** will be the coefficient in $H_{00}(z)$, and the even entries the coefficients of $H_{01}(z)$. Likewise, the odd and even entries of **g** is the coefficients of $H_{10}(z)$ and $H_{11}(z)$.

Now, the hard part is to go in the other direction. Suppose we have the filter taps and want to find the lifting steps (in either time or frequency domain). That it is theoretical possible to obtain the lifting steps was made evident with Theorem 1. Now we want to show how to actually obtain the lifting steps.

The Euclidean Algorithm

The Euclidean algorithm is usually first presented as an algorithm for finding the greatest common divisor of two integers. But it can be applied to many other analogous problems. One application is to finding the greatest common divisor of two Laurent polynomials. This turns out to be the key step in factorizing a polyphase matrix.

Take two Laurent polynomials $a(z)$ and $b(z) \neq 0$ with $|a(z)| \geq |b(z)|$. Then there always exist a Laurent polynomial $q(z)$, the quotient, with $|q(z)| = |a(z)| - |b(z)|$, and a Laurent polynomial $r(z)$, the remainder, with $|r(z)| < |b(z)|$, such that

$$a(z) = b(z)q(z) + r(z)\,. \tag{70}$$

We use the notation $q(z) = a(z)/b(z)$ as 'division disregarding the remainder'. Now, let $a_0(z) = a(z)$ and $b_0(z) = b(z)$, and iterate the following steps starting from $n = 0$

$$a_{n+1}(z) = b_n(z)\,,$$
$$b_{n+1}(z) = a_n(z) - q_{n+1}(z)b_n(z)$$

which in matrix form is

$$\begin{bmatrix} a_{n+1}(z) \\ b_{n+1}(z) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & -q_{n+1}(z) \end{bmatrix} \begin{bmatrix} a_n(z) \\ b_n(z) \end{bmatrix},$$

where $q_{n+1}(z) = a_n(z)/b_n(z)$. Let N denote the smallest integer with $N \leq |b(z)| + 1$, for which $b_N(z) = 0$. Then $a_N(z)$ is a greatest common divisor for $a(z)$ and $b(z)$. In short, for two Laurent polynomials $a(z)$ and $b(z)$ with the given constraints we have

$$\begin{bmatrix} a_N(z) \\ 0 \end{bmatrix} = \prod_{n=N}^{1} \begin{bmatrix} 0 & 1 \\ 1 & -q_n(z) \end{bmatrix} \begin{bmatrix} a(z) \\ b(z) \end{bmatrix}.$$

We note that there is no unique greatest common divisor for $a(z)$ and $b(z)$, since if $d(z)$ divides both $a(z)$ and $b(z)$, and is of maximal degree, then $\alpha z^m d(z)$ is also a divisor of the same degree.

Note the order of the terms in this product, as given by the limits in the product. Inverting the matrices and moving them to the other side, we get (products in reverse order, as shown by the limits)

$$\begin{bmatrix} a(z) \\ b(z) \end{bmatrix} = \prod_{n=1}^{N} \begin{bmatrix} q_n(z) & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a_N(z) \\ 0 \end{bmatrix}. \tag{71}$$

Now, take $a(z) = H_{00}(z)$ and $b(z) = H_{01}(z)$ from the polyphase matrix (46). We then get

$$\begin{bmatrix} H_{00}(z) \\ H_{01}(z) \end{bmatrix} = \prod_{n=1}^{N} \begin{bmatrix} q_n(z) & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} K \\ 0 \end{bmatrix}. \tag{72}$$

The greatest common divisor $a_N(z)$ of $H_{00}(z)$ and $H_{01}(z)$ is a constant. This is because of the determinant assumption

$$H_{00}(z)H_{11}(z) - H_{01}(z)H_{10}(z) = 1\,,$$

since $a_N(z)$ divides the left hand side, it also divides the right hand side. Thus, $a_N(z) = Kz^{\ell}$. By shifting the indices in the z-transform of $H_0(z)$ with ℓ steps, the greatest common divisor becomes a constant.

Observing that

$$\begin{bmatrix} q(z) & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & q(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ q(z) & 1 \end{bmatrix},$$

and replacing

$$\begin{bmatrix} K \\ 0 \end{bmatrix} \quad \text{with} \quad \begin{bmatrix} K & 0 \\ 0 & K^{-1} \end{bmatrix},$$

and letting $M = \lceil N/2 \rceil$, we can rewrite (72) to

$$\begin{bmatrix} H_{00}(z) & H'_{10}(z) \\ H_{01}(z) & H'_{11}(z) \end{bmatrix} = \prod_{n=1}^{M} \begin{bmatrix} 1 & q_{2n-1}(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ q_{2n}(z) & 1 \end{bmatrix} \begin{bmatrix} K & 0 \\ 0 & K^{-1} \end{bmatrix},$$

where these equations *define* $H'_{10}(z)$ and $H'_{11}(z)$. If N is odd, let $q_{2M}(z) = 0$. Transposing both sides yields

$$\begin{bmatrix} H_{00}(z) & H_{01}(z) \\ H'_{10}(z) & H'_{11}(z) \end{bmatrix} = \begin{bmatrix} K & 0 \\ 0 & K^{-1} \end{bmatrix} \prod_{n=M}^{1} \begin{bmatrix} 1 & q_{2n}(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ q_{2n-1}(z) & 1 \end{bmatrix}. \tag{73}$$

which is clearly resembles the factorization in Theorem 1, which we are trying to achieve. All we need to do now is to find out how $H'_{10}(z)$ and $H'_{11}(z)$ are connected with $H_{10}(z)$ and $H_{11}(z)$. By using the fact that both polyphase matrices (46) and (73) must have determinant 1, it is possible to show that

$$\begin{bmatrix} 1 & 0 \\ -t(z) & 1 \end{bmatrix} \begin{bmatrix} H_{00}(z) & H_{01}(z) \\ H'_{10}(z) & H'_{11}(z) \end{bmatrix} = \begin{bmatrix} H_{00}(z) & H_{01}(z) \\ H_{10}(z) & H_{11}(z) \end{bmatrix}$$

for

$$t(z) = H'_{10}(z)H_{11}(z) - H'_{11}(z)H_{10}(z)\,. \tag{74}$$

Consequently, we have the relation

$$\begin{bmatrix} H_{00}(z) & H_{01}(z) \\ H_{10}(z) & H_{11}(z) \end{bmatrix} = \begin{bmatrix} K & 0 \\ 0 & K^{-1} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -K^2 t(z) & 1 \end{bmatrix} \prod_{n=M}^{1} \begin{bmatrix} 1 & q_{2n}(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ q_{2n-1}(z) & 1 \end{bmatrix}, \tag{75}$$

which by a suitable re-indexing of the q polynomials (and at the same time making $K^2 t(z)$ one of them), determines the $S_n(z)$ and $T_n(z)$ in Theorem 1.

Practical Implementation of the Euclidean Algorithm

We now have a step-by-step method for obtaining the lifting steps from any filter. It goes as follows.

1. Find the z-transform $H_0(z)$ of the low pass filter taps **h**.
2. Determine $H_{00}(z)$ and $H_{01}(z)$ by means of (57).
3. Assign $a_0(z) = H_{00}(z)$ and $b_0(z) = H_{01}(z)$, and let $n = 0$.
4. Determine the quotient $q_{n+1}(z) = a_n(z)/b_n(z)$. There may be multiple solutions to this step. Choose one.
5. Determine the remainder as $b_{n+1} = a_n(z) - q_{n+1}(z)\, b_n(z)$.
6. Assign $a_{n+1}(z) = b_n(z)$.
7. Increment n. If $b_n \neq 0$, go to step 3.
8. Let $K = a_n(z)$. Determine $H'_{10}(z)$ and $H'_{11}(z)$ using (73).
9. Find the z-transform $H_1(z)$ of **g** such that it is index-shifted an odd number compared to $H_0(z)$.
10. Determine $H_{10}(z)$ and $H_{11}(z)$ by means of (58).
11. Determine $t(z)$ from (74).
12. Let $T_n(z) = q_{2n-1}(z)$ for $n = 1, \ldots, M$.
13. Let $S_n(z) = q_{2n}(z)$ for $n = 1, \ldots, M$.
14. Let $T_{M+1}(z) = -K^2 t(z)$.
15. If K has non-zero power of z, you may discard this. Alternatively, choosing the right z-transform in step 2 will give a 'zero power' K.

This procedure will provide all the ingredients for building the lifting step matrices. Note that this will always give a prediction step as the first matrix, and as such this procedure seems unable to generate lifting steps for, say, Daubechies 4 which starts with an update step. However, in the polyphase representation prediction and update steps are merely linear combinations of even and odd samples applied to odd and even samples, and as such predictions and updates can be interchanged simply by an odd shift of the z-transform of the signal.

Factoring Daubechies 4 into Lifting Steps

We now give an example of creating lifting steps using the algorithm presented in the previous section. The example is factorization of Daubechies 4. This means that we show how to get the matrix form (67) from the filter tap form (68) and (69). We will follow the step-bystep method laid out in the previous section.

The Daubechies 4 filter taps are given by

$$\mathbf{h} = \begin{bmatrix} \frac{1+\sqrt{3}}{4\sqrt{2}} & \frac{3+\sqrt{3}}{4\sqrt{2}} & \frac{3-\sqrt{3}}{4\sqrt{2}} & \frac{1-\sqrt{3}}{4\sqrt{2}} \end{bmatrix}. \tag{76}$$

In the z-transform in step 1 we are free to choose a shift. For instance, we could choose

$$H_0(z) = h[0] + h[1]z + h[2]z^2 + h[3]z^3 . \tag{77}$$

We know that eventually this choice will provide a monomial K that may or may not have zero power. As we will see later, this choice does in fact lead to a real number K. $H_0(z)$ must be separated into even and odd indices. This happens in step 2. Combining this with assignment to $a_0(z)$ and $b_0(z)$ in step 3 we get

$$\begin{aligned} a_0(z) = H_{00}(z) &= h[0] + h[2]z \\ &= \frac{1+\sqrt{3}}{4\sqrt{2}} + \frac{3-\sqrt{3}}{4\sqrt{2}}z , \end{aligned} \tag{78}$$

$$\begin{aligned} b_0(z) = H_{01}(z) &= h[1] + h[3]z \\ &= \frac{3+\sqrt{3}}{4\sqrt{2}} + \frac{1-\sqrt{3}}{4\sqrt{2}}z . \end{aligned} \tag{79}$$

Note that the concept of even and odd applies to the chosen z-transform of the filters taps (77), not the filter tap vector itself (76). Then step 4 is to find $q_1(z)$. Since $a_0(z)$ and $b_0(z)$ have the same degree, the quotient is a monomial. Matching the z coefficients yields

$$\begin{aligned} q_1(z) &= \frac{z \text{ coefficient of } a_0(z)}{z \text{ coefficient of } b_0(z)} = \frac{\frac{3-\sqrt{3}}{4\sqrt{2}}}{\frac{1-\sqrt{3}}{4\sqrt{2}}} = \frac{3-\sqrt{3}}{1-\sqrt{3}} \\ &= \frac{(3-\sqrt{3})(1+\sqrt{3})}{(1-\sqrt{3})(1+\sqrt{3})} = \frac{2\sqrt{3}}{-2} = -\sqrt{3} . \end{aligned}$$

The remainder is then

$$\begin{aligned} b_1(z) &= a_0(z) - b_0(z)q_1(z) \\ &= \left(\frac{1+\sqrt{3}}{4\sqrt{2}} + \frac{3-\sqrt{3}}{4\sqrt{2}}z\right) \\ &\quad - \left(\frac{3+\sqrt{3}}{4\sqrt{2}} + \frac{1-\sqrt{3}}{4\sqrt{2}}z\right)\cdot(-\sqrt{3}) \\ &= \frac{1+\sqrt{3}+3\sqrt{3}+3}{4\sqrt{2}} = \frac{1+\sqrt{3}}{\sqrt{2}} , \end{aligned}$$

thus completing step 5. In step 6 we assign

$$a_1(z) = b_0(z) = \frac{3+\sqrt{3}}{4\sqrt{2}} + \frac{1-\sqrt{3}}{4\sqrt{2}}z .$$

In step 7 we increment $n = 0$ to $n = 1$, and since $b_1 \neq 0$ we repeat from step 4. This time the quotient has degree 1, since $b_1(z)$ is one degree less than $a_1(z)$. More specifically, $q_2(z)$ must be on the form $c + dz$. Further, since the degree of the remainder decreases, and $|b_1(z)| = 0$, the remainder must be zero this time. So

$$\begin{aligned} b_1(z)q_2(z) &= \frac{1+\sqrt{3}}{\sqrt{2}}(c+dz) = a_1(z) \\ &= \frac{3+\sqrt{3}}{4\sqrt{2}} + \frac{1-\sqrt{3}}{4\sqrt{2}}z . \end{aligned}$$

Thus,

$$\begin{aligned} c &= \frac{\frac{3+\sqrt{3}}{4\sqrt{2}}}{\frac{1+\sqrt{3}}{\sqrt{2}}} = \frac{3+\sqrt{3}}{4(1+\sqrt{3})} = \frac{\sqrt{3}}{4} \\ d &= \frac{\frac{1-\sqrt{3}}{4\sqrt{2}}}{\frac{1+\sqrt{3}}{\sqrt{2}}} = \frac{1-\sqrt{3}}{4(1+\sqrt{3})} = \frac{\sqrt{3}-2}{4} . \end{aligned}$$

Therefore,

$$\begin{aligned} q_2(z) &= \frac{\sqrt{3}}{4} + \frac{\sqrt{3}-2}{4}z , \qquad b_2(z) = 0 , \\ a_2(z) &= b_1(z) = \frac{1+\sqrt{3}}{\sqrt{2}} . \end{aligned}$$

Now, this time in step 7 we have $b_2(z) = 0$. Thus, we continue to step 8, assign

$$K = \frac{1+\sqrt{3}}{\sqrt{2}} ,$$

and use (73) to find H'_{10} and H'_{11}.

$$\begin{aligned} &\begin{bmatrix} H_{00}(z) & H_{01}(z) \\ H'_{10}(z) & H'_{11}(z) \end{bmatrix} \\ &= \begin{bmatrix} K & 0 \\ 0 & K^{-1} \end{bmatrix} \begin{bmatrix} 1 & q_2(z) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ q_1(z) & 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{\sqrt{3}+1}{\sqrt{2}} & 0 \\ 0 & \frac{\sqrt{3}-1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & \frac{\sqrt{3}-2}{4}z + \frac{\sqrt{3}}{4} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\sqrt{3} & 1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{1+\sqrt{3}}{4\sqrt{2}} + \frac{3-\sqrt{3}}{4\sqrt{2}}z & \frac{3+\sqrt{3}}{4\sqrt{2}} + \frac{1-\sqrt{3}}{4\sqrt{2}}z \\ \frac{3-\sqrt{3}}{\sqrt{2}} & \frac{\sqrt{3}-1}{\sqrt{2}} \end{bmatrix} . \end{aligned} \tag{80}$$

Now that we have H'_{10} and H'_{11}, we need H_{10} and H_{11} in order to determine $t(z)$. First, we need the z-transform of the high pass filter taps. The high pass filter $\mathbf{g}$ is the time reversed of $\mathbf{g}$ with alternating signs. From (76) we therefore get

$$\mathbf{g} = \mathbf{H}_1 = \begin{bmatrix} -\frac{1-\sqrt{3}}{4\sqrt{2}} & \frac{3-\sqrt{3}}{4\sqrt{2}} & -\frac{3+\sqrt{3}}{4\sqrt{2}} & \frac{1+\sqrt{3}}{4\sqrt{2}} \end{bmatrix}. \quad (81)$$

The z-transform of $\mathbf{g}$ must have the index shifted by 1 (or some other odd number) compared to $H_0(z)$. For instance,

$$H_1(z) = g[0]z^{-2} + g[1]z^{-1} + g[2] + g[3]z\,, \quad (82)$$

where $g[3] = h[0]$ now is coefficient for an odd power, whereas $h[0]$ is coefficient for an even power in $H_0(z)$. Splitting this in even and odd coefficients (where even and odd again relates to the z-transform), we get from (58) that

$$H_{10}(z) = g[0]z^{-1} + g[2] = -\frac{1-\sqrt{3}}{4\sqrt{2}}z^{-1} - \frac{3+\sqrt{3}}{4\sqrt{2}}\,,$$
$$H_{11}(z) = g[1]z^{-1} + g[3] = \frac{3-\sqrt{3}}{4\sqrt{2}}z^{-1} + \frac{1+\sqrt{3}}{4\sqrt{2}}\,,$$

thus completing step 11. We now insert these H_{10} and H_{11} together with H'_{10} and H'_{11} from (80) into (74).

$$\begin{aligned} t(z) &= H'_{10}(z)H_{11}(z) - H'_{11}(z)H_{10}(z) \\ &= \frac{\sqrt{3}-3}{\sqrt{2}}\Big(\frac{3-\sqrt{3}}{4\sqrt{2}}z^{-1} + \frac{1+\sqrt{3}}{4\sqrt{2}}\Big) \\ &\quad - \frac{\sqrt{3}-1}{\sqrt{2}}\Big(-\frac{1-\sqrt{3}}{4\sqrt{2}}z^{-1} - \frac{3+\sqrt{3}}{4\sqrt{2}}\Big) \\ &= \frac{1}{8}\Big((\sqrt{3}-3)(3-\sqrt{3})z^{-1} + (\sqrt{3}-3)(1+\sqrt{3}) \\ &\quad + (\sqrt{3}-1)(1-\sqrt{3})z^{-1} + (\sqrt{3}-1)(3+\sqrt{3})\Big) \\ &= \frac{z^{-1}}{8}\big((-9+6\sqrt{3}-3) + (-1+2\sqrt{3}-3)\big) \\ &= (\sqrt{3}-2)z^{-1}\,, \end{aligned}$$

which completes step 11. We can now determine the extra matrix we need, as shown in (75),

$$-K^2 t(z) = -\Big(\frac{1+\sqrt{3}}{\sqrt{2}}\Big)^2(\sqrt{3}-2)z^{-1} = z^{-1}\,.$$

It is now possible to determine all the matrix lifting steps that makes up the polyphase matrix, which is step 12, 13, and 14. We have $M = 1$, so we get one S matrix and two T matrices.

$$\begin{aligned} T_1(z) &= q_1(z) = -\sqrt{3} \\ S_1(z) &= q_2(z) = \frac{\sqrt{3}}{4} + \frac{\sqrt{3}-2}{4}z \\ T_2(z) &= -K^2 t(z) = z^{-1}\,. \end{aligned}$$

Since K is a real number, we do not need to shift the powers of z (as mentioned in step 15), and we can now write the complete factorization of the polyphase matrix $\mathbf{H}(z)$ as shown in (61) from Theorem 1.

$$\mathbf{H}(z) = \begin{bmatrix} \frac{\sqrt{3}+1}{\sqrt{2}} & 0 \\ 0 & \frac{\sqrt{3}-1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ z^{-1} & 1 \end{bmatrix} \begin{bmatrix} 1 & \frac{\sqrt{3}}{4} + \frac{\sqrt{3}-2}{4}z \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\sqrt{3} & 1 \end{bmatrix}.$$

But, this is fact not (67). The reason is that the presented algorithm for finding the lifting steps always will start with a prediction step, whereas the factorization in (67) starts with an update step. So while $\mathbf{H}(z)$ in (67) indeed is equal to $\mathbf{H}(z)$ above, the factorizations are not the same. This is a consequence of the fact that the distinction of the lifting matrices into update and prediction is useful for interpretation of the transform results, but not necessary for the implementation of the transform.

Actually, this is not the only 'non-uniqueness feature' of the wavelet transform. If instead of (77) and (82) we had chosen

$$H_0(z) = h[0]z^{-3} + h[1]z^{-2} + h[2]z^{-1} + h[3]\,,$$
$$H_1(z) = g[0]z^1 + g[1]z^2 + g[2]z^3 + g[3]z^4\,,$$

the even coefficients from before are now odd, and vice versa. That means that the relation between $\mathbf{h}$ and $\mathbf{g}$ changes signs, so for instance, $g[3] = -h[0]$ now, rather than $g[3] = h[0]$. The resulting factorization becomes

$$\mathbf{H}(z) = \begin{bmatrix} -\frac{\sqrt{3}-1}{\sqrt{2}} & 0 \\ 0 & -\frac{\sqrt{3}+1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ z^2 & 1 \end{bmatrix} \begin{bmatrix} 1 & -\frac{\sqrt{3}}{4}z^{-1} - \frac{\sqrt{3}+2}{4}z^{-2} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \sqrt{3}z & 1 \end{bmatrix}.$$

There are more examples of factorizations in Jensen and la Cour–Harbo [3] and in the original paper by Sweldens and Daubechies [2].

Bibliography

1. Daubechies I (1992) Ten Lectures on Wavelets. CBSM-NSF Regional Conference Series in Applied Mathematics, vol 60. SIAM, Philadelphia
2. Daubechies I, Sweldens W (1998) Factoring wavelet transforms into lifting steps. J Fourier Anal Appl 4(3):245–267
3. Jensen A, la Cour–Harbo A (2001) Ripples in mathematics – the discrete wavelet transform. Springer, Heidelberg
4. Mallat S, Wavelet A (1998) Tour of signal processing. Academic Press Inc, San Diego

5. Mulcahy C (1996) Plotting and scheming with wavelets. Math Mag 69(5):323–343
6. Oppenheim AV, Schafer R (1975) Digital signal processing. Prentice Hall, Upper Saddle River
7. Oppenheim AV, Schafer R, Buck JR (1999) Distrete–time signal processing. Prentice Hall, Upper Saddle River
8. Sweldens W (1996) The lifting scheme: A custom-design construction of biorthogonal wavelets. Appl Comput Harmon Anal 3(2):186–200
9. Sweldens W (1997) The lifting scheme: A construction of second generation wavelets. SIAM J Math Anal 29(2):511–546
10. Vetterli M, Kovačević J (1995) Wavelets and subband coding. Prentice–Hall, Englewood Cliffs
11. Vetterli M, Kovačević J (2007) Wavelets and subband coding, 2nd edn. http://waveletsandsubbandcoding.org, accessed on Sept. 5th, 2008

Wavelets and PDE Techniques in Image Processing, A Quick Tour of

Hao-Min Zhou[1], Tony F. Chan[2], Jianhong Shen[3]
[1] School of Mathematics, Georgia Institute of Technology, Atlanta, USA
[2] Department of Mathematics, University of California, Los Angeles, USA
[3] Barclays Capital, New York, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Wavelets in Image Processing
PDE Techniques
Wavelet Based Variational PDE Methods
Future Directions
Acknowledgments
Bibliography

Glossary

Wavelets Wavelets are selected functions that generate orthonormal bases of the square integrable function space L^2 (or more generally, frames of L^p spaces) by using dilations and translations. The basis functions have certain locality, such as compact support or fast decay property. And they are usually organized according to different scales or resolutions, which are called Multi-Resolution Analysis (MRA). Fast wavelet transforms are filtering procedures that compute the projection of any given function onto a wavelet basis.

Digital images Digital images usually refer to n dimensional data arrays recorded by optical or other imaging devices, such as digital cameras, Radar, Computed Tomography (CT), and Magnetic Resonance Imaging (MRI). They can also be generated by computer graphics software. Most digital images in the literature are 2- or 3-dimensional.

Image restoration Image restoration rebuilds high-quality images from given images that are corrupted or polluted during acquisition or transmission processes. The most commonly seen restoration tasks are denoising and deblurring. Denoising is to remove random perturbations to individual pixel values. Deblurring is to remove the unwanted correlation between nearby pixels and to recover the original clear images.

Image compression Compression converts images from n dimensional data arrays into "0" and "1" bit streams so that they can be stored or transmitted more efficiently. There are two types of compression, lossy and lossless, depending on whether information is permanently lost or recoverable, respectively. Many of the commonly used compression algorithms, such as the ones used by international image compression standards JPEG and JPEG2000, are transform-based compression, which consists of three basic steps: transform pixel values into frequency coefficients, quantization of the frequency coefficients, and coding to convert them into bit streams.

Image segmentation Segmentation partitions images into subregions (segments), on which images share similar features. Each region often corresponds to the image of an individual object in the 3-dimensional world.

Image inpainting Inpainting is an artistic word referring to filling in missing image information on damaged regions, e. g., scratches and damages in precious old photos, old Hollywood films, and ancient paintings. The objective of digital image inpainting is to fill in the missing information automatically and meaningfully.

Definition of the Subject

Explosive information has dominated nearly all aspects of modern society, science and technology. Visualization is one of the most direct and preferable ways to observe information carried by data, which are often massive in size and uncertain in data quality. To better reveal the information, especially when it is hidden, implicit, or corrupted, data must first be properly processed. In achieving this, image processing, which includes many differ-

ent tasks such as compression, restoration, inpainting, segmentation, pattern recognition and registration, has played a critical role. Historically, it has been viewed as a branch of signal processing, and many classical methods are adopted from traditional Fourier-based signal processing algorithms. In the past couple of decades, numerous new competing methods have emerged. Among them, wavelets, variational and PDE techniques, and stochastic methods have demonstrated outstanding performance due to their special properties. For instance, wavelets have become the dominant tool in image processing because of their multiresolution structure, energy concentration ability and fast transform algorithms. The popularity of variational PDE techniques is driven by their extraordinary properties in understanding and manipulating geometrical features. These new techniques have contributed to new observations, understandings, and discoveries in science and technology. Furthermore, many of these methods have been successfully applied to applications in the fields of medical, physical sciences, engineering, and even everyday life.

Introduction

Digital image processing analyzes or extracts certain information from digital images, which are often viewed as 2- or multi-dimensional data sets in mathematics. Each element in the data sets is called a pixel. Typical image processing tasks include segmentation, restoration, pattern recognition, analysis, compression, registration and motion detection [41,46]. Image processing has a wide range of applications including communication, computer vision, acoustics, satellite imaging, medical and industrial diagnosis and many more.

Image processing tasks often require large-scale computations, mainly due to the large amount of data to be processed. A typical gray scale still image with moderate resolution, such as 1024×1024, has over a million pixels. The size of a color image is three times as large given the same resolution. A video sequence usually consists of over 24 color frames per second with each frame being a still image. A multi-spectral image contains a collection of several (usually more than 3) monochrome images of the same scene, each of which is taken with a different wavelength by a different sensor. In addition, many applications, such as airport screening and unmanned vehicle navigation, require real time response. All of these demand efficient and reliable algorithms.

Traditional image processing methods are mainly based on Fourier/wavelets or statistical approaches. The best example is the current international image compression standards, JPEG and JPEG2000, which are based on discrete cosine transform (DCT) and wavelet transforms. For this reason, more images are stored using their wavelet coefficients. The tremendous success of wavelets in image processing is due to their positive properties, including multiresolution data structures, fast transform algorithms and superb energy concentration ability, which allows one to approximate functions (images) using only a relative small number of coefficients.

Thousands of researchers have devoted their efforts to the development of wavelet theory, analysis, and algorithms in different applications. Groundbreaking contributions include Meyer's wavelet theory [53], Daubechies' compact support orthogonal wavelets [33], Mallat's multiresolution analysis [49,50], Shapiro's progressive zero tree image coding algorithm [63], and many other works cited in books such as [28,34,44,51], and [64].

Roughly speaking, wavelet transforms can express any square integrable functions by superpositions of wavelet basis functions, which are generated by dilations and translations from a few (if not a single) wavelet functions. The summation coefficients are called wavelet coefficients, which are standard L^2 inner products between wavelets and the given functions. Wavelet transforms are realized by filtering procedures. Usually, wavelet coefficients are classified into two types: low or high frequencies. Low frequency coefficients correspond to certain kinds of weighted local averages of the data values. High frequency coefficients are related to certain order derivatives. Therefore, high frequency coefficients are small for smooth functions and large for functions containing discontinuities.

In applications, it is inevitable that some of the wavelet coefficients, especially the high frequencies, are unavailable due to intentional or involuntary reasons. For instance, in wavelet-based image compression, insignificant (smaller in magnitude) high frequency coefficients are discarded on purpose to save more storage space. In lossy channel communication, coefficients are lost or damaged during the transmission due to unwanted disturbances. Obviously, with incomplete wavelet coefficients, one cannot synthesize the exact original functions. Many problematic issues could arise as a result. One that has drawn the most attention is the assertion that oscillations are generated near discontinuities. This is the famous Gibbs' phenomenon in mathematics and edge artifacts in image processing.

Several directions have been taken to improve the performance of wavelet based image processing methods by reducing the Gibbs' oscillations, and by better preserving geometrical information in images. One strategy involves

the use of nonlinear thresholding procedures to allocate more storage resource to significant coefficients. Well-known examples include translation invariant denoising methods [31], wavelet hard thresholding, and wavelet shrinkage (also called soft thresholding) [37].

Another strategy is to build new geometry friendly wavelet-like multiresolution representations, such as ridgelets [8], curvelets [9], beamlets [36], bandelets [57] and many more recent developments. With geometry directly incorporated into the construction of multiresolution representations, it is expected that the decompositions have better performance near discontinuities.

The third direction is to modify the existing wavelet transforms so that fewer large high frequency coefficients are generated near discontinuities. Thus, less information is truncated in the thresholding process. Many methods have been proposed, such as Harten's remarkable general multiresolution framework [42] and its recent developments [2], the adaptive lifting scheme [30], and the adaptive Essential Non-Oscillatory (ENO) wavelet transforms [25,26]. Many recent contributions are collected in [65].

In a different direction, PDE techniques for image processing, pioneered by Mumford–Shah's segmentation functional [55], Rudin–Osher–Fatemi's Total Variation (TV) restoration [59], and Perona–Malik's anisotropic diffusion [58], have emerged more recently. Due to their outstanding properties in handling geometrical information, different variational PDE models and methods have been proposed and studied for a variety of image processing goals, such as affine scale space [62], fundamental equations for image processing [1], total variation image analysis [16], active contour for segmentation [12,23], blind deconvolution [24], image interpolation and inpainting [4,5,18,20,52], restoration [17,19], and compression [27,38]. The field is significantly enriched, and many books have been published in recent years (see [3,13,21,54,56,61,66] and references therein).

Given the developments in both wavelets and PDE techniques in image processing, it is natural to think of combining their advantages to gain more benefits in the applications, especially when geometrical features are important. Well designed wavelet PDE methods can retain the good properties of wavelets, such as multiresolution and fast algorithms. Meanwhile, they are able to use PDE concepts, such as gradients, curvatures to capture, control and manipulate the geometrical information to achieve image processing goals in more systematic manners. There are quite a few examples that have demonstrated the combined advantages in different applications [10,15,27,29,39,48].

In this paper, it is not our intention to give a complete survey on either wavelets or PDE techniques in image processing. Instead, we will focus on a recent trend that combines them together. To be self-contained, we start with a brief introduction to wavelets, and followed by PDE techniques in image processing. We hope to use selected topics based on our experience to help readers, especially beginners, to know some basic models and a few commonly used methodologies on the subject. The rest of the paper is arranged as follows. Section "Wavelets in Image Processing" is a brief introduction to wavelets and their applications in image processing. Section "PDE Techniques" presents some well known PDE models in image processing. In Sect. "Wavelet Based Variational PDE Methods" we give some new developments of combining wavelets and PDE techniques. A concise list of future directions is stated in the end.

Wavelets in Image Processing

Historically, Fourier decompositions, which express any given square integrable function by superpositions of sinusoidal functions, have been the major tool for image processing due to their efficient representations and fast Fourier transforms (FFT). This is particularly true for 1-D signals, such as audio sequences. However, all Fourier basis functions have global supports, which implies that any local change in the given function has to result in a global change in the representations. For this reason, Fourier bases are not efficient to represent local information, such as discontinuities. The well-known Gibbs' phenomena is an exhibition of this limitation. Unfortunately, most salient features, such as edges and corners in images, are local and discontinuous. Thus, all Fourier-based methods for image processing suffer from the ringing artifacts.

Facing this shortcoming, it is highly desirable to have efficient representations which can better handle local information, especially discontinuities. Or more precisely, the basis functions should have local support or fast decay properties so that any local perturbation will only cause changes in a small neighborhood but not to far away places. To a certain extent, wavelets are designed to fill up this expectation and have gained unsurpassed success in many applications of image processing.

After several decades of intensive studies, wavelets have been developed into a very rich mathematical theory. There are many different types of wavelets such as Meyer's wavelets, spline wavelets, and bi-orthogonal wavelets. Here, we present a very brief introduction based on Daubechies' compact supported wavelets and their connections to compression and denoising.

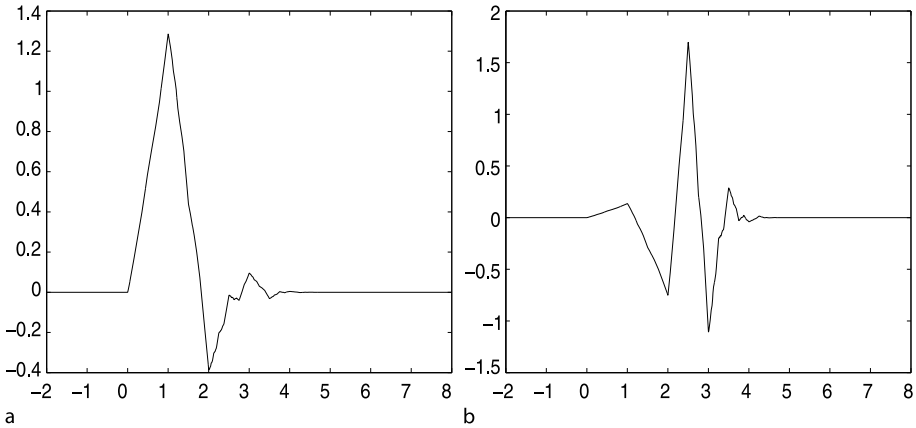

Wavelets and PDE Techniques in Image Processing, A Quick Tour of, Figure 1
a The scaling function for Daubechies-6 wavelet. b The corresponding wavelet

Wavelets

Wavelets can be viewed as orthonormal bases of the square integrable function space $L^2(\mathbf{R})$. It starts with carefully selected scaling function $\phi(x)$ and corresponding wavelet $\psi(x)$ defined on finite support $[0, l]$, where l is a positive integer. We refer to [34] for the detailed selection procedure for $\phi(x)$ and $\psi(x)$. Many commonly used software such as MATLAB have built-in routines for the scaling and wavelet functions already.

The functions $\phi(x)$ and $\psi(x)$ satisfy the dilation equations (also called two-scale relations or refinement equations in some literature):

$$\phi(x) = \sqrt{2}\sum_{s=0}^{l} c_s \phi(2x - s)\,, \tag{1}$$

and

$$\psi(x) = \sqrt{2}\sum_{s=0}^{l} h_s \phi(2x - s)\,, \tag{2}$$

where the c_s's and h_s's are constants, called low- and high-pass filters, respectively. To give examples, the famous Haar wavelet selects

$$\phi(x) = \begin{cases} 1 & x \in [0, 1) \\ 0 & \text{otherwise}\,, \end{cases}$$

and

$$\phi(x) = \begin{cases} 1 & x \in [0, \frac{1}{2}) \\ -1 & x \in [\frac{1}{2}, 1) \\ 0 & \text{otherwise}\,, \end{cases}$$

which are step functions. We also plot the scaling and wavelet functions of Daubechies-6 in Fig. 1.

Using dilation and translation, one can form families of functions from $\phi(x)$ and $\psi(x)$, as follows,

$$\phi_{j,k}(x) = 2^{\frac{j}{2}}\phi(2^j x - k)\,, \tag{3}$$

and

$$\psi_{j,k}(x) = 2^{\frac{j}{2}}\psi(2^j x - k)\,, \tag{4}$$

where (j, k) are integers. Then the collection of $\psi_{j,k}(x)$ form an orthonormal basis of $L^2(\mathbf{R})$. This means that for any given function $f(x) \in L^2(\mathbf{R})$, one has

$$f(x) = \sum_{j,k} < f(x), \psi_{j,k}(x) > \psi_{j,k}(x)\,, \tag{5}$$

where $< \cdot, \cdot >$ denotes the standard $L^2(\mathbf{R})$ inner product defined by

$$< f(x), g(x) >= \int_{\mathbf{R}} f(x)g(x)\mathrm{d}x\,.$$

There are many desirable properties for the scaling functions and wavelets. Among them, locality and oscillations are the most cited common features in all wavelets. Literally speaking, they make wavelets behave like localized small waves, which also explains the origination of the name.

The locality, which often refers to compact support or fast decay properties, enables wavelets to decompose or approximate functions locally. This satisfies the desire of many applications, particularly in image processing.

A good mathematical way to describe the oscillatory nature of wavelets is to use their vanishing moment property, which means

$$\int \psi(x)x^j\mathrm{d}x = 0\,, \quad j = 0, 1, \ldots, p-1\,, \tag{6}$$

where p is a positive integer. In this case, the wavelet $\psi(x)$ is said to have p vanishing moments. The more vanishing moments, the more oscillations in wavelets in general.

Locality and oscillation together have been the main driving engines for the success of wavelets in many applications.

Multi-Resolution Analysis

The success of wavelets also relies on their connection to Multi-Resolution Analysis, introduced by Mallat [49,50].

Consider the subspace of $L^2(\mathbf{R})$ defined by the scaling function $\phi_{j,k}(x)$,

$$V_j = \text{span}\{\phi_{j,k}(x), k \in \mathbf{Z}\} ,$$

for every fixed j. The dilation Eq. (1) implies that the subspaces form an ordered chain,

$$\cdots \subseteq V_{j-1} \subseteq V_j \subseteq V_{j+1} \subseteq V_{j+2} \ldots , j \in \mathbf{Z} ,$$

which also satisfies

$$\overline{\lim_{j\to\infty} V_j} = L^2(\mathbf{R}), \lim_{j\to-\infty} V_j = 0 .$$

Here, larger indexes j correspond to finer resolutions or scales.

Similarly, one can define the subspaces generated by wavelets $\psi_{j,k}(x)$,

$$W_j = \text{span}\{\psi_{j,k}(x), k \in \mathbf{Z}\} .$$

The dilation Eq. (2) implies the following connection between W_j and V_j,

$$V_j = V_{j-1} \oplus W_{j-1}, j \in \mathbf{Z} . \tag{7}$$

Therefore $L^2(\mathbf{R})$ can be decomposed into,

$$L^2(\mathbf{R}) = V_J \oplus \sum_{j>J} W_j = \sum_{j=-\infty}^{\infty} W_j ,$$

where J is an arbitrary reference resolution level. Consequently, $f(x) \in L^2(\mathbf{R})$ can be decomposed into a multi-resolution representation as,

$$f(x) = \sum_k \alpha_{J,k}\phi_{J,k}(x) + \sum_{j>J,k} \beta_{j,k}\psi_{j,k}(x) , \tag{8}$$

where $\alpha_{j,k} =< f(x), \phi_{j,k}(x) >$ is called a low frequency (or scaling) coefficient, and $\beta_{j,k} =< f(x), \psi_{j,k}(x) >$ is a high frequency (or wavelet) coefficient. Without causing confusion, we call them wavelet coefficients for simplicity in this paper.

The decomposition (7) and the dilation Eqs. (1), (2) lead to the following filtering and down-sampling procedures to compute the coarser scale wavelet coefficients from the finer scale coefficients,

$$\alpha_{j,k} = \sum_{s=0}^{l} c_s \alpha_{j+1,2k+s} , \tag{9}$$

and

$$\beta_{j,k} = \sum_{s=0}^{l} h_s \alpha_{j+1,2k+s} . \tag{10}$$

These are the famous fast wavelet transforms.

Apparently, fast wavelet transforms involve only the coefficients and can be started if one knows the low frequency coefficients $\{\alpha_{I,k}\}$ on a certain fine resolution I. Then, it is natural to ask how to obtain $\{\alpha_{I,k}\}$. Theoretically, $\{\alpha_{I,k}\}$ must be computed by $< f(x), \phi_{I,k}(x) >$ according to the definition. However, they are often replaced by the point-wise values $f(x_k)$ in practice, even though such an action is called a wavelet crime in [64]. The replacement makes sense when the function $f(x)$ is smooth and the resolution I is fine enough, because the low frequency coefficients $\alpha_{I,k}$, which are the weighted local averages of $f(x)$, are very close approximations to the point-wise values.

The above described wavelet transforms are for 1-D functions. Wavelet transforms for 2-D images are realized by simple tensor product in practice. More precisely, 2-D transforms are obtained by performing column-wise 1-D transforms followed by row-wise 1-D transforms.

Wavelet Thresholding and Image Processing

The wavelet representations (8) provide a mechanism to approximate functions in a multi-resolution fashion. For instance, the jth scale (resolution) approximation is simply defined as:

$$\begin{aligned} f_j(x) &= \sum_k \alpha_{j,k}\phi_{j,k}(x) \\ &= \sum_k \alpha_{J,k}\phi_{J,k}(x) + \sum_{J<i<j,k} \beta_{i,k}\psi_{i,k}(x) . \end{aligned} \tag{11}$$

This multi-resolution approximation satisfies a standard error bound,

$$\|f(x) - f_j(x)\| \le C2^{-jp}\|f^{(p)}(x)\| , \tag{12}$$

where C is a constant independent of j. It is obvious that the error is controlled by the vanishing moment p, the norm of the pth derivative of $f(x)$, and the resolution j. Better approximations with more detailed information can be easily obtained by adding more terms for the finer resolutions. Many have argued that this convenient zoom-in and zoom-out multi-resolution approximation is by far one of the best mathematical models that mimic human perceptions.

In addition to the multi-resolution structure, the success of wavelet decomposition in image processing also depends on the sparsity of the wavelet coefficients. Simple integration by parts can show that high frequency wavelet coefficients satisfy $|\beta_{j,k}| = |f^{(p)}(x)|O(2^{-jp})$, which suggests that the wavelet coefficients are small if the function $f(x)$ is smooth enough. For images containing many smooth regions, such as most of the natural scenery images, it is easy to observe that a large number of the high frequency coefficients are insignificant, and therefore can be ignored in applications. Thresholdings are mathematical procedures that realize this observation.

Loosely speaking, thresholding is setting selected wavelet coefficients to be zero. There are many different types of thresholdings. In fact, the jth scale approximation is one of them. It is constructed by ignoring all the scales higher than the given resolution j. This is often called linear thresholding, because the procedure is linear. Other nonlinear data dependent thresholdings, including commonly used hard and soft thresholdings, can achieve much better performance in image processing.

The hard thresholding simply sets any wavelet coefficients whose magnitudes are smaller than a given tolerance ϵ to be zero, i. e.

$$\bar{\beta}_{j,k} = \begin{cases} \beta_{j,k} & |\beta_{j,k}| > \epsilon \\ 0 & |\beta_{j,k}| \le \epsilon \,. \end{cases}$$

A similar formula holds for the low frequency coefficients too.

The soft wavelet thresholding is slightly different from the hard thresholding. It is a shrinkage procedure. In addition to setting the coefficients whose magnitudes are smaller than the tolerance to zero, it reduces the magnitudes of other coefficients by ϵ as well,

$$\bar{\beta}_{j,k} = \begin{cases} \mathrm{sign}(\beta_{j,k})(|\beta_{j,k}| - \epsilon) & |\beta_{j,k}| > \epsilon \\ 0 & |\beta_{j,k}| \le \epsilon \,, \end{cases}$$

where the sign$(\cdot)$ is the signum function.

The selection of the threshold ϵ has also been investigated by many groups. Among many proposed strategies, Donoho–Johnstone's *SQTWOLOG* [37] and Stein's unbiased risk estimate have been widely used.

Thresholding procedures have accomplished remarkable success in image processing, especially in compression. It is easy to understand that wavelet thresholdings are useful in this application because one does not have to store the coefficients that are zero. However, it is more subtle in practical compression schemes. The problem is that not only does one need to remember the non-zero wavelet coefficients, but also their locations. The location information may occupy more storage space than the coefficients if they are recorded in a naive way. Shapiro's zero tree scheme [63] introduces a tree structure for wavelet coefficients based on their multiresolution property. A branch of the tree can be represented by a single bit '0' if all coefficients in the branch are zero. This is used in conjunction with thresholdings to achieve very efficient compression. Many well known state-of-the-art compression methods, such as Set Partitioning in Hierarchical Trees (SPIHT) [60] and Group Test Wavelet (GTW) [45] compression algorithms, are based on the zero tree idea.

Simple thresholdings also provide fast and effective methods for noise removal. They have found many successful applications in communications, military and medical images. In Fig. 2, we display the denoising effects of wavelet hard (**b**) and soft (**c**) thresholdings of a test image with additive white noise (**a**).

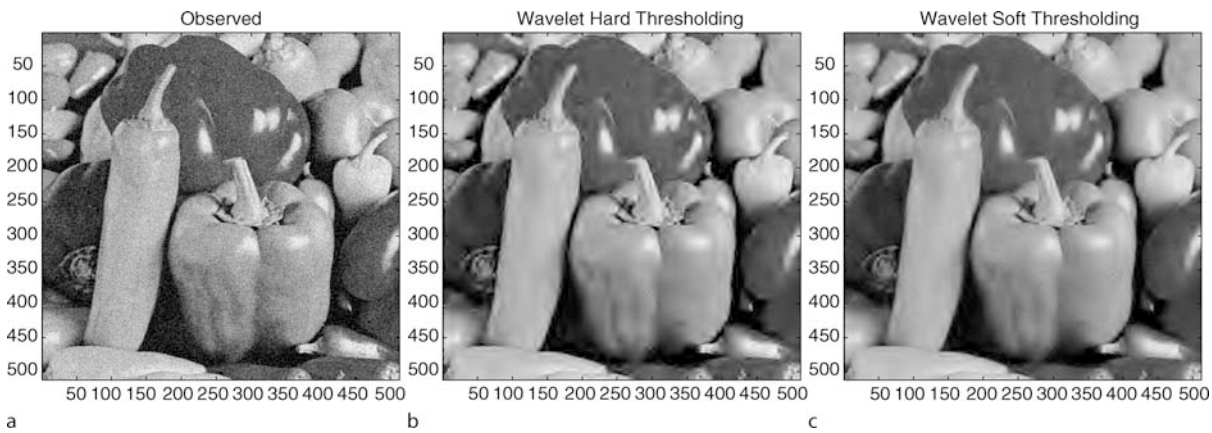

Wavelets and PDE Techniques in Image Processing, A Quick Tour of, Figure 2
a Test image corrupted by white noise. b Denoised image by hard thresholding. c Denoised image by soft thresholding

From a mathematical point of view, the success of thresholdings can be explained by their connections to optimizations. It has been shown that many thresholding results are optimal in a certain sense. In layman's terms, those thresholding results are best under certain criteria. For example, let us assume that the hard thresholding reconstruction

$$\bar{f}(x) = \sum_k \bar{\alpha}_{J,k}\phi_{J,k}(x) + \sum_{j,k} \bar{\beta}_{j,k}\psi_{j,k}(x)$$

has M nonzero wavelet coefficients. Then $\bar{f}(x)$ is the minimizer of the following optimization problem,

$$\min_g \|f-g\|_2 , \quad \text{subject to} \quad g \text{ has at most } M \text{ nonzero wavelet coefficients} .$$

This leads to the conclusion that the hard thresholding gives the best M-term approximation in $L^2(\mathbf{R})$ among all possible combinations.

In a more general setting as discussed in [15] and [67], it is proved that the soft thresholding gives the minimizer of the following optimization problem

$$\min_g \{\|f-g\|_2 + 2\epsilon\|g\|_{B_1^1(L^1)}\} ,$$

where $B_1^1(L^1)$ is a Besov space. And the linear thresholding gives an approximate minimizer of the following optimization problem,

$$\min_g \{\|f-g\|_2 + 2\epsilon\|g\|_{W^m(L^2)}\} ,$$

where $W^m(L^2)$ is a Sobolev space. We refer readers to [15] for a detailed discussion.

PDE Techniques

Compared to wavelets, modern PDE techniques in image processing have appeared more recently, even though some traditional image processing methods can be interpreted from PDE perspective. For instance, the classical Gaussian filter for image denoising is accomplished by convolving the noisy image u_0 with the Gaussian kernel (also called heat kernel in literature) $G(x,t) = \frac{1}{2\pi t}\exp(-\frac{x^2}{2t})$,

$$u = G * u_0 = \int u_0(y)G(x-y,t)\mathrm{d}y . \tag{13}$$

This denoised image u is actually the solution $u(x,t)$ of the following diffusion PDE,

$$u_t(x,t) = D\Delta u(x,t) , \quad u(x,0) = u_0(x) , \tag{14}$$

where Δ is the Laplace operator, and $D = 1/2$ is diffusive coefficient.

Modern PDE techniques have drawn great attention and reached remarkable success in the past two decades. This is due to their extraordinary ability to handle geometrical features, which are lacking in traditional statistical or Fourier/wavelet based approaches. Two different strategies are commonly used to design PDE techniques for different image processing goals.

1. Construct PDE-based evolution processes and incorporate geometry in the equations.
2. Pose image processing tasks in variational framework and derive corresponding Euler–Lagrange equations to compute the minimizers.

In both strategies, image processing goals are achieved by solving PDE's. Next, we use a few well-known examples to demonstrate these two strategies.

Anisotropic Diffusion for Denoising

Image denoising removes unwanted disturbances in images. Very often, those disturbances, such as white noise and pepper-and-salt noise, are highly localized and oscillatory. This makes it harder to separate noise from edges, which are also local and discontinuous. As an anti-oscillation procedure, diffusion is a natural selection for denoising. As mentioned earlier, the classical Gaussian filter for denoising is equivalent to the linear isotropic diffusion (14). However, it has been observed in both experimental and theoretical studies that isotropic diffusion unavoidably smears sharp edges, corners and other geometrical features embedded in u_0 while filtering out noise. This is because it treats all orientations identically and never recognizes the presence of spatially coherent discontinuities – edges. In addition, the larger the diffusive coefficient D, the quicker the smoothing out.

To remedy this drawback, Perona–Malik [58] proposed using anisotropic diffusion instead,

$$u_t = \nabla \cdot (D(x,u,\nabla u)\nabla u) . \tag{15}$$

The diffusivity coefficient D is data dependent and must sense the existence of edges, so that the PDE stops diffusion across the discontinuities. For this purpose, it is desirable to have D satisfying the following requirements,

$$D = \begin{cases} \text{large}, & \text{when } |\nabla u| \text{ is small on intra-regions}, \\ \text{small}, & \text{when } |\nabla u| \text{ is large near edges} . \end{cases} \tag{16}$$

Therefore, the evolution only smooths out the oscillations away from edges but not across them . In [58], D is selected as

$$D = g(|\nabla u|^2) ,$$

where g is a smooth positive concave function satisfying $g(+\infty) = 0$. For example, g can be taken as

$$g(|\nabla u|^2) = \mathrm{e}^{-\frac{|\nabla u|^2}{2\sigma^2}}\,, \quad \text{or} \quad g(|\nabla u|^2) = \frac{1}{1 + b|\nabla u|^2}\,,$$

where σ and $b > 0$ are constants.

In practice, the anisotropic diffusion (15) must face a challenge on how to compute the coefficient D robustly. This may be troublesome especially in the beginning of the diffusion process when u_0 contains highly oscillatory noise, because $|\nabla u|$ is large almost everywhere, so D is small everywhere. Thus, the diffusion is not effective in removing noise. To overcome this difficulty, the use of a mollified image in g has been proposed in [14], which takes the form as

$$u_t = \nabla \cdot (g(|\nabla(G_\sigma * u)|^2)\nabla u)\,, \quad u(x,0) = u_0(x)\,,$$

where G_σ is again the Gaussian kernel.

Along the lines of anisotropic diffusion, much more research has been done including the well known general axiomatic scale-space theory in [1]. We refer readers to [21,66] for more discussion.

Total Variation Image Denoising

A different viewpoint for denoising is to reduce the uncorrelated local oscillations in images. Mathematically speaking, total variation (TV) is a quantity that measures oscillations in functions. It is intuitive that oscillatory noise greatly increases the TV norm. Naturally, one can think of denoising as reducing the total variations of images. In fact, this observation leads to the famous TV model proposed by Rudin–Osher–Fatemi [59],

$$\min_u \int |\nabla u| \mathrm{d}x \quad \text{subject to} \quad \|u - u_0\|_2 \le \sigma\,, \tag{17}$$

where σ is related to the noise level. The objective functional is to reduce oscillations in the reconstruction, and the constraint term is a fitting requirement. This optimization problem can be read as to find the least oscillatory image within a small ball of radius σ centered at the noisy image u_0.

The model is often re-formulated as a non-constraint minimization problem as

$$\min_u \int |\nabla u| \mathrm{d}x + \frac{\lambda}{2}\|u - u_0\|_2\,, \tag{18}$$

where $\lambda \ge 0$ is a Lagrange multiplier, which is the factor that balances the competition between oscillations and fidelity. The smaller the λ, the fewer details in the denoised images. In extreme situations, the solution for (18) is a flat constant when λ is zero, or is the noisy image u_0 when λ is infinite.

The most outstanding advantage of TV denoising model (18) is that it allows sharp edges to be preserved in the reconstruction. This implies that TV model has the ability to reduce small oscillations (noise) without penalizing the edges. This feature has been well understood in the context of computational fluid dynamics (CFD), especially in shock capturing, where TV semi norm is intensively used. In fact, the authors of [59] are also experts in CFD, and it is no doubt that (17) is inspired by numerical shock capturing.

Another attraction of TV denoising is its geometrical properties. For functions with finite TV semi norms, this can be seen clearly through an equivalent coarea formula,

$$\int |\nabla u| \mathrm{d}x = \int_{-\infty}^{+\infty} \int_{\{u=\gamma\}} \mathrm{d}s\,\mathrm{d}\gamma\,.$$

Here the term $\int_{\{u=\gamma\}}$ is the length of the level set $\{u = \gamma\}$. The TV semi norm is obtained by integrating along all level contours of $\{u = \gamma\}$ for all values of γ. This suggests that TV semi norm controls both the size of the jumps and the geometry of the level sets.

The geometric connection of TV minimization is more visible if we analyze the optimization (18) by calculus of variation. The standard theory shows that the minimizer must satisfy the following Euler–Lagrange equation,

$$-\nabla \cdot \left(\frac{\nabla u}{|\nabla u|}\right) + \lambda(u - u_0) = 0\,. \tag{19}$$

The first term, the functional derivative of TV semi norm, is precisely the curvature of the image, which makes the method more geometric friendly. For noisy pixels, the jumps are isolated and their curvature is large. They will be wiped out much quicker than the edges that are coherent jumps with relatively smaller curvature.

The best known, but not necessarily the most efficient, algorithm to solve (18) is the gradient descent method, which introduces an artificial time to form an evolution PDE,

$$u_t = \nabla \cdot \left(\frac{\nabla u}{|\nabla u|}\right) - \lambda(u - u_0)\,. \tag{20}$$

Compared to (15), the gradient descent of TV minimization (20) is also an anisotropic diffusion with a degenerate diffusive coefficient $D = 1/|\nabla u|$. And it satisfies the anisotropic diffusion requirement (16). In particular, if an edge is sharp, D will be zero and no diffusion is performed across the edge. In practice, to prevent numerical blow-up caused by $|\nabla u| = 0$ in the denominator, it is often replaced by $\sqrt{|\nabla u|^2 + \epsilon}$, where ϵ is a small positive number. Actually, this replacement can be derived from variational framework too.

An interesting and natural question is why would one want to use TV semi-norm in (18) instead of $\int |\nabla u|^2 \mathrm{d}x$, which is the famous Sobolev H_1 semi-norm in PDE's. In fact, a very similar calculation can show that the H_1 minimization leads to exactly the isotropic diffusion (14), which loses the geometrical properties.

Variational Models for Image Segmentation

The purpose of image segmentation is to divide an image into regions within which the image has similar features, such as intensity values, texture pattern, or belonging to same objects. Segmentation is a crucial building block for many high level image processing and vision tasks such as object detection, recognition, and tracking. Obviously, one image may produce different partitions because of different segmentation criteria. This non-uniqueness nature, which is also true for many other image processing tasks, makes the segmentation problem very challenging. There is extensive literature on the subject and many methods have been proposed using different strategies. For example, the celebrated intensity-edge mixture model is statistic-based [40], while the widely-used active contour (also called snake) model [47] uses variational framework. We take the well-known Mumford–Shah segmentation model [55] and Chan–Vese region-based active contour model [23] as examples to demonstrate how mathematical formulations and computational strategies can contribute to segmentation.

The original Mumford–Shah segmentation model is stated in a variational format,

$$\min \lambda_1 \int_\Gamma \mathrm{d}s + \frac{\lambda_2}{2} \int_{\Omega\setminus\Gamma} |\nabla u|^2 \mathrm{d}x + \frac{\lambda_3}{2} \int_\Gamma (u - u_0)^2 \mathrm{d}x \,, \quad (21)$$

where λ_1, λ_2 and λ_3 are three constants, u is the partitions with different segments. Ω is the region where the image is defined and Γ is the interior boundary separating different segments. The first term is the length of the interior boundary curves. The second term is isotropic diffusion within each homogeneous region. Similar to the TV minimization model (18), the third term is the fitting term. From formulation (21), the segmentation is achieved by balancing the competitive three terms. Different ratios among λ_1, λ_2, and λ_3 give different partitions.

The Mumford–Shah model has many desirable properties and is very general. Many other known models can be viewed as special cases of it. However, it also faces serious computational challenges because the partition boundary Γ is unknown. And thus the first term involving line integral along Γ has no easy way to compute. To ease the challenges, many other models are proposed for better computation properties. Among them, Chan–Vese's *active contour without edge* model [23] has gained remarkable success due to its simplicity and robustness.

Assume that C is a closed curve partitioning the segments. The model is designed to move C so that the following energy is minimized,

$$\min \lambda_1 \cdot \mathrm{Length}(\Gamma) + \lambda_2 \cdot \mathrm{Area}(\mathrm{inside}(\Gamma)) + \lambda_3 \int_{\mathrm{inside}(\Gamma)} (u_0 - c_1)^2 \mathrm{d}x + \lambda_4 \int_{\mathrm{outside}(\Gamma)} (u_0 - c_2)^2 \mathrm{d}x \,, \quad (22)$$

where $\lambda_1, \lambda_2, \lambda_3$ and λ_4 are positive fixed parameters. This model uses piecewise constant approximations inside and outside the partition curves. If one picks $\lambda_2 = 0$, (22) becomes the minimal partition model which is a special case for (21).

The Chan–Vese model (22) can also be formulated in a level set framework, and it leads to a fast and robust computation method, which sparkles a large amount of follow-up researches in using level set-based active contour methods for segmentations in different applications.

PDE Method for Image Inpainting

Image inpainting, or its mathematical synonym image interpolation, fills in missing or damaged image regions based on known surrounding information. It is a very fundamental problem having numerous prior work in existence. It also shares common ground with many other image processing tasks, such as image replacement, error concealment, edge completion and image editing. Here we only use 1) a third order nonlinear inpainting PDE by Bertalmio et al. [6], 2) a variational inpainting model by Chan–Shen [20], as two examples to illustrate how modern mathematics is used for this traditional labor-intensive task, because image inpainting used to be done by hand.

Similar to segmentation, image inpainting is an inverse problem having possible multiple solutions. It is obvious that when information is missing, different people may have different ways to patch information to the regions, and all of them may look reasonable. However, it is commonly agreed that the inpainted regions must have consistent geometrical features and texture patterns with their surroundings. For this reason, many of the inpainting methods are based on geometrical interpolations or extrapolations. One example is the remarkable third order inpainting PDE introduced by Bertalmio et al. [6]. In fact, the term *image inpainting* was first used by them, and

the work has stimulated a wave of interest in inpainting related problems.

The PDE is given as

$$u_t = \nabla(\Delta u) \cdot \nabla^{\perp} u \,, \tag{23}$$

where $\nabla^{\perp}$ is the orthogonal gradient direction (isophote direction as called in the original paper). The idea behind (23) is a brilliant intuition of information transport along broken level lines (isophotes). The PDE is solved only inside the inpainting regions with proper boundary conditions based on the gray values and isophote directions. It is discovered later that (23) actually connects to the famous Navier–Stokes equations in CFD [5].

The Chan–Shen's inpainting model tackles the problem from a different angle. It starts with a variational principle inspired by TV restoration model [59],

$$\min_u \int_{\Omega} |\nabla u| \mathrm{d}x + \frac{\lambda}{2} \int_{\Omega \setminus D} (u - u_0)^2 \mathrm{d}x \,, \tag{24}$$

where D is the inpainting region. Similar to (17), a straightforward interpretation of this model is that the minimizer u is the least oscillatory image that is close enough to the given image u_0 outside the inpainting region. For the regions inside of D, the restored u has no restriction except matching its surroundings in a least oscillatory fashion. This model can lead to a nonlinear data dependent PDE similar to (19) and can be solved numerically. The results are impressive and much follow-up work has been performed to analyze the model and extend it to include more sophisticated measurements such as Euler's Elastica into consideration for better curve treatments [18].

Wavelet Based Variational PDE Methods

As discussed in the previous sections, both wavelets and PDE techniques have been used extensively in image processing and achieved tremendous success in numerous applications. Their success is based on different properties of both approaches. Wavelets have multi-resolution data structure, energy concentration (sparsity) and fast algorithms. PDE techniques are geometrically friendly and often tied to variational principles. A closer look at both approaches can easily reveal that those properties do not overlap, and one cannot be used to replace the other. In a certain sense, they are complementary to each other, and it seems natural to combine the advantages of both to gain benefits. In fact, many research efforts have been put forward in this direction.

There are two different strategies that have been explored to merge PDE techniques with wavelets,

(a) Use computational PDE skills to modify the standard wavelet transforms to form new transforms having better geometric properties.
(b) Design new wavelet-based variational models for different image processing tasks.

In this section, we will select a few examples to demonstrate both strategies.

ENO-Wavelet Transforms

As mentioned earlier in Sect. "Wavelets in Image Processing", it is a well-known fact that Fourier-based algorithms suffer Gibbs' oscillations. Wavelets can remarkably reduce the severity of oscillations due to their locality. But they still exist unless one retains all discontinuity related coefficients, which is not practical in many applications. To improve the image quality, one needs to reduce the oscillations by lowering the threshold ϵ in thresholding procedures. As a consequence, more coefficients, especially edge-related ones, must be retained, which is why a majority of the storage is allocated to edge-related coefficients in JPEG2000.

To further improve the performance and reduce the ringing artifacts, it is desirable to design wavelet-like transforms so that fewer significant high-frequency coefficients are generated. One way to achieve this goal is to incorporate geometrical features in the design of wavelet-like basis (or redundant frame) functions so that discontinuous functions can be more effectively represented. Many efforts have been proposed, such as curvelets [9], beamlets [36], and bandelets [57] to name a few. A different approach is to reduce the wavelet filter length. This is based on the fact that the larger the supports of wavelets, the longer the filters, and the worse the Gibbs' oscillations. Then, one can adaptively use shorter wavelet filters near the vicinities of edges. The adaptive lifting scheme proposed in [30] employs this idea.

ENO-wavelet transforms approach the problem from a different angle. It is inspired by the original Harten's multi-resolution framework [42], which has a profound impact on many new methods in the field. ENO-wavelet transforms borrow a key idea, the one-sided interpolation strategy, from ENO schemes for shock capturing [43]. Different from the fore-mentioned methods, which adapt the filters or basis functions to better fit the data, ENO-wavelet transforms change the data near edge areas and feed them into the same standard wavelet filters. The data is changed in a special way so that the filters do not see the discontinuities.

Let us imagine that we filter around a jump discontinuity. A high-frequency coefficient is large if the high

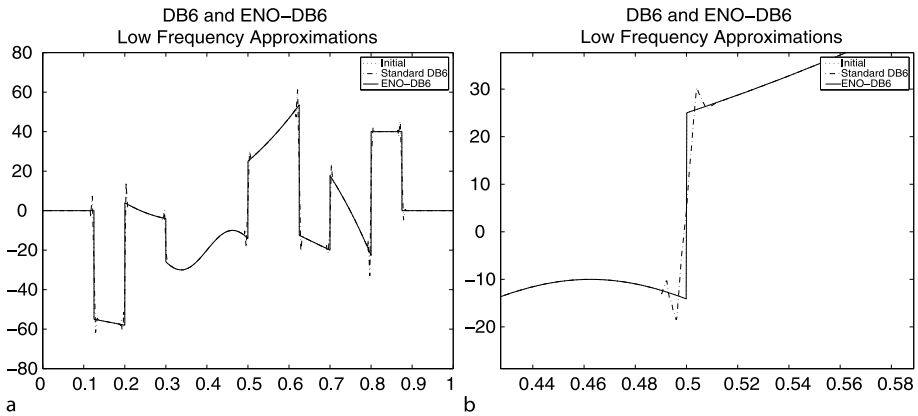

Wavelets and PDE Techniques in Image Processing, A Quick Tour of, Figure 3
a The comparison between 4-level ENO-Daubechies-6 (*solid line*) and standard Daubechies-6 (*dash-dotted line*) approximations. b A zoom-in of the picture on the left near a discontinuity. Standard Daubechies-6 generates oscillations near discontinuities, but the ENO-Daubechies-6 does not

pass filter is convolved with data across the jump. However, one can extend the data from both left and right sides in smooth ways and feed the extended data to the filters. Then the high-frequency coefficients are small as the high pass filter only sees the smooth data on both sides. Of course, one may immediately question the fact that near the jump region, we have actually two different pieces of data overlapping in the area. In fact, this is a serious issue, in that it causes a double storage problem, which means we have doubled the number of wavelet coefficients in the jump region. And this is directly against the goals of many image processing tasks, especially image compression. Fortunately, the problem can be avoided by a strategy called coarse-level extrapolation, which extends the data in such ways that some of the jump-related wavelet coefficients are predictable and do not need to be memorized. And the storage can be reduced to the same as that of standard wavelet transforms. We refer to [25] for detailed ENO-wavelet transform algorithms. Here we just point out the main ideas and some important results.

ENO-wavelet transforms can be used as functional replacements of standard wavelet transforms. Indeed, ENO-wavelet transforms perform standard wavelet transforms if no discontinuity is detected. ENO-wavelet transforms retain the essential properties and advantages of standard wavelet transforms, such as energy concentration, multiresolution framework and fast transform algorithms, all without any edge artifacts. They also achieve uniform approximation accuracy up to the discontinuities. If $\hat{f}_j(x)$ is the jth resolution approximation to $f(x)$ by using ENO-wavelet transforms, then

$$\|\hat{f}_j(x) - f(x)\| \le C2^{-jp}\|f^{(p)}(x)\|_{\Omega\setminus\Gamma}\,, \tag{25}$$

where Γ is the set of discontinuous points. It is worth noting that the error (12) for standard wavelet transforms depends on the pth derivative of $f(x)$ on the entire region Ω, which is unbounded if the discontinuous set Γ is not empty. In contrast, the error for ENO-wavelet transforms (25) depends on $f^{(p)}(x)$ only on the domain Ω excluding Γ. This ensures that ENO-wavelet transforms perform uniformly accurate, regardless of the presence of discontinuities, which is probably the best result one may expect. In Fig. 3, we show a comparison between ENO-wavelets and standard wavelets.

Wavelet Based Minimal Energy Methods for Denoising

As discussed in Sect. "PDE Techniques", anisotropic diffusion and total variation minimization for image denoising have a tremendous ability to extract image features, especially edges, for better image quality preservation. However, it is also commonly recognized that such PDE techniques often post higher computational demands, because numerical solutions for nonlinear PDE's need to be computed iteratively. To achieve reasonable solutions, many iterations must be performed. This has been a major criticism of PDE techniques, especially when one compares them with wavelets, which have ultra fast filtering algorithms.

To retain the capability in feature extraction while eliminating the need of iterations for anisotropic diffusion, there have been efforts to formulate geometric friendly energy minimizations in wavelet spaces so that the minimizers can be obtained directly from wavelet coefficients without iterations. In fact, as discussed in Subsect. "Wavelet Thresholding and Image Processing", classical wavelet

thresholdings, including linear, hard and soft thresholdings, have corresponding energy optimizations in certain functional spaces. But those minimization problems are not built to handle geometrical features. In [29], Chui–Wang suggested a new geometrical energy minimization in wavelet space given as

$$\min E_\lambda(\rho,\beta) = \lambda E_i(\rho,\beta) + \frac{1}{2}\|\beta-\alpha\|^2 , \tag{26}$$

where $E_i(\rho,\beta)$ is a selected internal energy which can be expressed by the wavelet coefficients β. The second term is the standard L^2 fitting requirement. In their original paper, a blended internal energy is chosen as

$$E_i(\rho,\beta) = \sum_{j,k}(\rho(m_{j,k}(p)) + \rho(\beta^d_{j,k})) , \tag{27}$$

where $m_{j,k}(p) = (|\beta^h_{j,k}|^p + |\beta^v_{j,k}|^p)^{1/p}$, and $\rho(s) = |s|$. In (27), the notations β^d, β^h and β^v are 2-D tensor product wavelet coefficients along diagonal, horizontal and vertical directions respectively. It is clear that the energy functional is resolution, orientation and spatial dependent. In this way, the energy functional can "see" the corners and edges in wavelet spaces because those geometrical structures create correlated wavelet coefficients along diagonal, horizontal and vertical directions respectively. When $p = 2$, the minimizers of (26) and (27) can be attained explicitly from the wavelet coefficients as

$$(\beta^*_\lambda)^{h,v}_{j,k} = (\beta^0)^{h,v}_{j,k}\left(1-\frac{\lambda}{m^0_{j,k}}\right)_+ , \tag{28}$$

and

$$(\beta^*_\lambda)^{d}_{j,k} = \mathrm{sign}(\beta)^{0,d}_{j,k}\left(|\beta^{0,d}_{j,k}-\lambda|\right)_+ , \tag{29}$$

where $(\cdot)_+$ denotes the nonnegative value function.

Along this line, there has been a recent trend in the computational harmonic analysis community to design data dependent nonlinear filters based on PDE techniques, for example, the adaptive digital TV filter presented in [19]. More recently, Chui and collaborators have proposed a new anisotropic filtering strategy based on ideas of finding approximate solutions of anisotropic diffusion equations discussed in Subsect. "Anisotropic Diffusion for Denoising". Their method realizes image denoising by one sweep of nonlinear filtering.

Diffusion Wavelets

Diffusion wavelets has been proposed by Coifman and collaborators in [32]. It is a different way to generalize classical wavelets using PDE and geometry concepts. The goal is to construct a multiresolution analysis framework on general geometric structures, such as manifolds, graphs or even discrete point sets, so that image processing tasks can be performed for functions defined on these structures.

As discussed in Subsect. "Wavelets", standard wavelet multi-resolutions are based on dilation and translations. However, this is often impossible for general data structures, especially when little geometric information is known. To overcome this difficulty, diffusion wavelets use dyadic powers of a diffusion operator T (with $\|T\| < 1$), such as the heat operator defined on the general data structure to create scales. The following two properties are crucial for constructing diffusion wavelets. One is that the spectral of high powers of T decay faster as the power gets higher. Consequently, one can use a few leading eigenfunctions of T^j (j large) to approximate the range spaces of T^j accurately.

The other property is that applying higher powers of T to local functions, such as Dirac delta functions defined on a point in a discrete data set produces smoother functions with wider supports, because T is a diffusion process. After a non-trivial process involving orthonormalization, which we refer to [32] for details, one can construct a multi-resolution analysis based on T^{2^j} ($j \in \mathbf{Z}^+$). Especially, once the multi-resolution analysis is formed, T^{2^j} can be expressed in a highly compressed format.

There are many potential applications, such as in data mining and learning theory. Here, we pick the following simple example to illustrate their usage. Let us consider computing the inverse of Laplacian $(I-T)^{-1}$ applied to an arbitrary vector f defined on a general data structure. The operator $(I-T)^{-1}$ is a deblurring process commonly seen in image restorations. It is well known that

$$(I-T)^{-1}f = \sum_{k=1}^{+\infty} T^k f .$$

Define

$$S_K = \sum_{k=1}^{2^K} T^k ,$$

then an approximation to $(I-T)^{-1}$ can be achieved by

$$S_{K+1}f = (S_K + T^{2^K}S_K)f = \prod_{k=0}^{K}(I+T^{2^k})f .$$

Since the powers T^{2^k} have been compressed in the multi-resolution analysis and can be efficiently applied to f, $(I-T)^{-1}f$ is computed efficiently.

TV Wavelet Inpainting

Wavelet inpainting, or more generally wavelet interpolation, refers to the problem of filling in missing or dam-

aged wavelet coefficients due to lossy image transmission or communication. Obviously, the task is closely related to classical inpainting problems, as discussed in Subsect. "PDE Method for Image Inpainting", but also differs remarkably in that the inpainting regions are in the wavelet domain.

Working in the wavelet domain, instead of the pixel domain, changes the nature of the inpainting problem, since damages to wavelet coefficients can create correlated damage patterns in the pixel domain. For instance, there usually exists no corresponding regular geometric inpainting regions, which are however necessary for many PDE-based inpainting models in pixel domains. Such lack of spatial geometric regularity of inpainting regions also prohibits many other existent inpainting techniques applied to pixel domains. On the other hand, direct interpolation in the wavelet domain is also problematic, because wavelet coefficients are constructed to be uncorrelated in the L^2 sense and neighboring coefficients provide minimum information to the missing ones. In addition, degradation in wavelet inpainting problems is often spatially inhomogeneous, which demands different treatments in different regions.

A closer examination may find that all these new challenges are actually caused by a simple fact: Damage happens in the wavelet domain while human perception prefers to see images with certain regularity in pixel domain. Therefore, it seems natural to create wavelet inpainting methods by filling in the coefficients in wavelet domain while controlling the regularity in the pixel domain. TV wavelet inpainting models presented in [22] exactly follow this strategy.

Two different models have been proposed based on the noise level in images. The first one is for noiseless images,

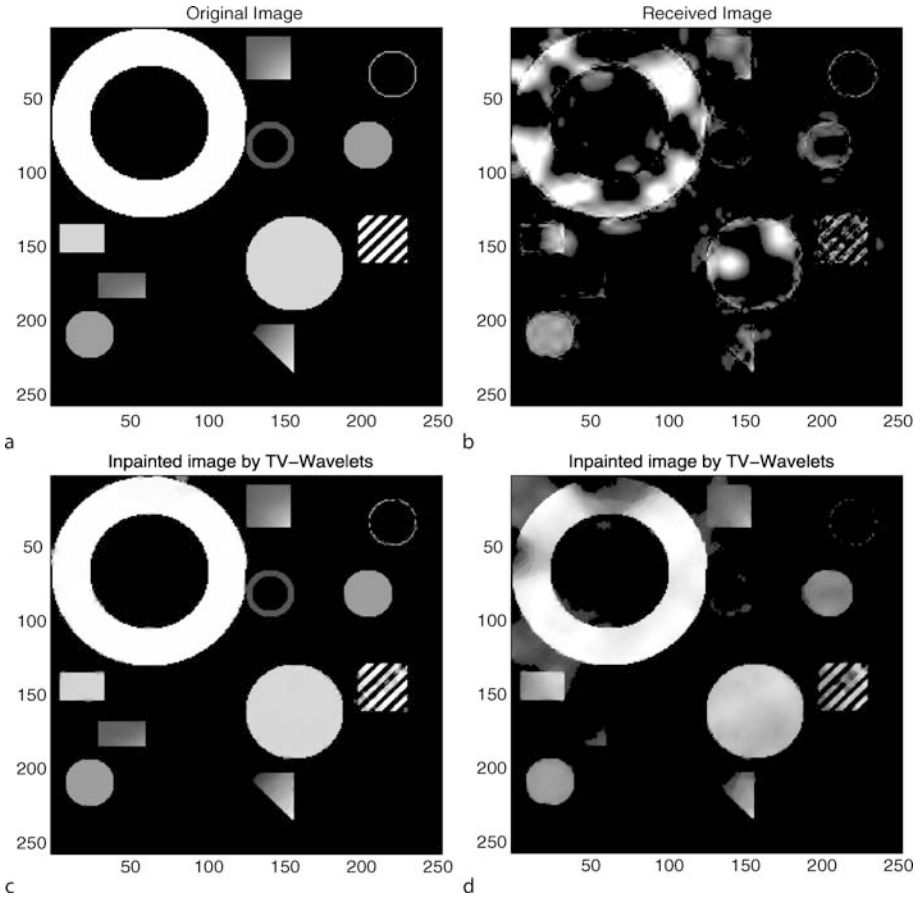

Wavelets and PDE Techniques in Image Processing, A Quick Tour of, Figure 4

a Original synthetic image. **b** 50 % of the wavelet coefficients are randomly lost, including some low-frequency coefficients, which results in large, damaged regions in the pixel domain. Notice that there are no well-defined inpainting regions in the pixel domain. **c** Restored image by Model I, **d** Restored image by Model II. They not only fill in missing regions properly, but also restore the sharp edges and geometrical shapes

in which the retained coefficients are considered to be correct and will not be alerted.

Model I:

$$\min_{\beta_{j,k}:(j,k)\in I} F(u,z) = \int_{\mathbf{R}^2} |\nabla_x u(\beta,x)|\mathrm{d}x = \mathrm{TV}(u(\beta,x)) \,, \quad (30)$$

where I is the inpainting index region in wavelet domain, and $u(\beta, x)$ has the wavelet transform:

$$u(\beta,x) = \sum_{j,k} \beta_{j,k}\psi_{j,k}(x) \,.$$

For noisy images, since every coefficient is also corrupted by noise. Then one has to modify (denoise) them too.

Model II:

$$\min_{\beta_{j,k}} F(u,z) = \int_{\mathbf{R}^2} |\nabla_x u(\beta,x)|\mathrm{d}x + \sum_{(j,k)} \lambda_{j,k}(\beta_{j,k} - \alpha_{j,k})^2 \,, \quad (31)$$

and the parameter $\lambda_{(j,k)}$ is zero if $(j, k) \in I$; otherwise, it equals a positive constant λ.

Clearly, these two models are inspired by TV denoising model for their exceptional ability of handling geometries. Both models recover the wavelet coefficients so that the restored images are least oscillatory while matching the known information. The key difference is that the arguments are restricted to the inpainting regions I only for Model I, so the dimension of unknowns is the number of coefficients in I. While in Model II, the parameter λ is taken to be zero in the inpainting regions I in the wavelet domain, in contrast to the standard denoising and compression models, where λ is usually taken to be a constant everywhere. This difference essentially puts no constraint on the missing wavelet coefficients so that they can change freely. In Fig. 4, we show an example of wavelet inpainting by the two models.

Compressive Sampling

Compressive sampling [7], also goes by the name compressed sensing [35], is an emerging theory addressing the sampling problem in image and signal processing. In information theory, the classical Shannon–Nyquist sampling theorem states that "Exact reconstruction of a continuous-time baseband signal from its samples is possible if the signal is bandlimited and the sampling frequency is greater than twice the signal bandwidth". More precisely, bandlimited signals refer to functions whose Fourier frequencies are in a bounded interval. And the sampling theorem says that a bandlimited signal can be fully reconstructed from its evenly spaced samples, provided that the sampling rate must exceed twice the maximum frequency in the bandlimited signal. This rate is often called Nyquist rate.

Compressive sampling considers a different scenario. In the simplest case, let us assume that a signal $f(t)$ is sparse in the frequency space or any other convenient spaces. For instance, the signal consists of only a few Fourier terms, i. e.

$$f(t) = \sum_{j=1}^{n} \beta_j \mathrm{e}^{-ik_j t} \,,$$

where n is an integer much smaller than the signal resolution N, $\{k_j\}$'s are frequencies, and i the imaginary unit. But the actual values of frequency k_j are not known. Obviously, $f(t)$ is a bandlimited signal. Without loss of generality, we assume k_n is the highest frequency. Then Shannon–Nyquist sampling theory requires at least $2k_n$ evenly spaced observations to exactly reconstruct $f(t)$. However, since the value k_n is not known, the actual number of samples (resolution N) needed may be much higher than $2k_n$.

Compressive sampling asks a different question: Can one exactly recover a sparse signal $f(t)$ using a small number of samples $\{f(t_j)\}_1^m$ observed at randomly selected time t_j $(j = 1, \ldots, m)$? Here m may be much smaller than N. Given the knowledge of sparsity, ideally one can convert this problem into the following minimization problem,

$$\min \|\beta\|_{l_0} \,, \quad \text{subject to} \quad (F^*\beta)(t_j) = f(t_j) \,, \quad (32)$$

where β is a vector containing the Fourier coefficients of the reconstructed signal, F is the Fourier matrix, and $\|\cdot\|_{l_0}$ is the l_0 norm of a discrete sequence, which is defined as the number of nonzero elements. Then F^* gives the inverse Fourier transform, and $(F^*\beta)$ is the reconstructed signal. l_0 minimization (32) finds the sparsest reconstruction $(F^*\beta)$ among all possible functions that agree with the observations $f(t_j)$.

However, l_0 minimization (32) is essentially a large non-convex integer optimization problem, which is computationally prohibitive. Then compressive sampling suggests that it is still possible to exactly recover $f(t)$ from the samples $\{f(t_j)\}_1^m$. The exact reconstruction is realized by the following l_1 optimization,

$$\min \|\beta\|_{l_1} \,, \quad \text{subject to} \quad (F^*\beta)(t_j) = f(t_j) \,. \quad (33)$$

In other words, the compressive sampling achieves exact recovery by finding the signal having the smallest l_1 norm

in frequency space among all signals matching the sample values $f(t_j)$ at t_j.

There are many reasons to select l_1 norm in the optimization, including the remarkable mathematical insights given in [11]. We do not intend to present their results here. Instead, we list the following two reasons that are more intuitive and may explain the essence of l_1 optimization in sparse recovery.

1. l_1 norm exhibits interesting sparsity in many applications. In other words, l_1 minimization in frequency space intends to drive more frequencies to zero.
2. l_1 norm has the least index p among all l_p norms that are convex.

The convex property ensures that (33) may be solved efficiently by the standard convex optimization algorithms.

It is believed that compressive sampling may have many implications. One of the most attractive potentials is that it suggests the possibility of new data acquisition protocols that translate analog information into digital form with fewer sensors than what was considered necessary. There are many interesting studies on how to advance the theory and applications, and even design new hardware to realize the implications. We refer to [7] for more information on the subject.

It is worth noting that even though TV wavelet inpainting and compressive sampling are developed independently, there is an interesting connection between them. For instance, the derivative of a piecewise constant image, ∇u, can be viewed as sparse in the pixel space. If one makes measurements in the wavelet space, then Model I (30) is the l_1 minimization of the derivative in the pixel space with constraints in the wavelet space, which fits well in the framework of compressive sampling. In this sense, they are complementary to each other, and can be viewed as dual formulations.

Future Directions

Driven by rapidly developing imaging sciences and technologies, the last couple of decades have witnessed the tremendous success of wavelets and PDE techniques in mathematical image processing. Many researchers have been working in the field, and exciting new developments are constantly reported. However, compared to the even faster growing demands, there still is a long way to go to meet the ever-increasing expectations. The following is just a very short list of directions that are being or shall be pursued in the near future.

(1) Developing more sophisticated models, methods to better preserve features for images, or general data sets in higher dimensions, such as video or hyper-spectral images. Merging traditional wavelets and PDE techniques seems to be a promising is direction. For example, developing wavelets and PDE models for segmentation is interesting. To our knowledge, it has not been attempted yet.

(2) New applications in high-level vision, such as pattern recognition, auto navigation and tracking, which demands better understanding of the problems and more accurate extraction of the connections among data sets.

(3) Robust and efficient implementation strategies to compute the solutions of mathematical image processing models, especially those involving solutions of nonlinear PDE's.

The list is based on the authors' experience and reflects our own perspective. Certainly it does not cover all aspects of this large field. Interested readers are encouraged to read up-to-date literature to follow the latest advancements on the subject.

Acknowledgments

Research supported in part by grants ONR-N00014-06-1-0345, NSF CCF-0430077, CCF-0528583, DMS-0610079, DMS-0410062 and CAREER Award DMS-0645266, NIH U54 RR021813, and STTR Program from TechFinity Inc.

Bibliography

1. Alvarez L, Guichard F, Lions PL, Morel JM (1993) Axioms and Fundamental Equations of Image Processing. Arch Ration Mech Anal 16:200–257
2. Arandiga F, Donat R (2000) Nonlinear Multiscale Decompositions: The approach of A Harten. Numer Algorithm 23:175–216
3. Aubert G, Kornprobst P (2001) Mathematical Problems in Image Processing: Partial Differential Equations and the Calculas of Variations, vol 147. Springer
4. Ballester C, Bertalmio M, Caselles V, Sapiro G, Verdera J (2001) Filling-in by Joint Interpolation of Vector Fields and Grey Levels. IEEE Trans Image Process 10(8):1200–1211
5. Bertalmio M, Bertozzi AL, Sapiro G (2001) Navier–Stokes, fluid dynamics, and image and video inpainting. IEEE Conference on Computer Vision and Pattern Recognition Dec Kauai, Hawaii
6. Bertalmio M, Sapiro G, Caselles V, Ballester C (1999) Image Inpainting, Tech Report. ECE-University of Minnesota
7. Candés E (2006) Compressive Sampling. Proceedings of the International Congress of Mathematicians, Madrid
8. Candés E, Donoho D (1999) Ridgelets: a Key to Higher-dimensional Intermittency? Phil Trans R Soc Lond A 357(1760):2495–2509
9. Candès E, Donoho D (1999) Curvelets: A Surprisingly Effective Nonadaptive Representation of Objects with Edges. Tech Report. Dept of Stat, Stanford University

10. Candès E, Romberg J, Tao T (2006) Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information. IEEE Trans Inform Theory 52:489–509
11. Candès E, Tao T (2006) Near Optimal Signal Recovery From Random Projections and Universal Encoding Strategies. IEEE Trans Inform Theory 52:5406–5425
12. Caselles V, Kimmel R, Sapiro G (1997) On Geodesic Active Contours. Int J Comput Vis 22(1):61–79
13. Caselles V, Morel JM, Sapiro G, Tannenbaum A (1998) Special Issue on Partial Differential Equations and Geometry-Driven Diffusion in Image Processing and Analysis. IEEE Tran Image Proc 7(3)
14. Catte F, Lions PL, Morel JM, Coll T (1992) Image Selective Smoothing and Edge Detection by Nonlinear Diffusion. SIAM J Numer Anal 29:182–193
15. Chambolle A, DeVore R, Lee N, Lucier B (1998) Nonlinear Wavelet Image Processing: Variational Problems, Compression, and Noise Removal Through Wavelet Shrinkage. IEEE Tran Image Proc 7(3):319–333
16. Chambolle A, Lions PL (1997) Image Recovery via Total Variational Minimization and Related Problems. Numer Math 76:167–188
17. Chan TF, Golub GH, Mulet P (1996) A Nonlinear Primal-Dual Method for Total Variation-Based Image Restoration In: ICAOS'96, 12th International Conference on Analysis and Optimization of Systems: Images, Wavelets and PDEs, Paris, 26–28 June, 1996. Lecture Notes in Control and Information Sciences, vol 219. Springer, pp 241–252
18. Chan TF, Kang SH, Shen J (2002) Euler's Elastica and Curvature Based inpainting. SIAM J Appl Math 63(2):564–592
19. Chan TF, Osher S, Shen J (2001) The Digital TV Filter and Nonlinear Denoising. Trans IEEE Image Process 10(2):231–241
20. Chan TF, Shen J (2002) Mathematical Models for Local Non-Texture Inpainting. SIAM J Appl Math 62(3) 1019–1043
21. Chan TF, Shen J (2005) Image Processing and Analysis - Variational, PDE, Wavelet and Stochastic Methods. SIAM Publisher, Philadelphia
22. Chan TF, Shen J, Zhou HM (2006) Total Variation Wavelet Inpainting. J Math Imaging Vis 25(1):107–125
23. Chan TF, Vese L (2001) Active Contour Without Edges. IEEE Trans Image Process 10(2):266–277
24. Chan TF, Wong CK (1998) Total Variation Blind Deconvolution. IEEE Trans Image Process 7:370–375
25. Chan TF, Zhou HM (2002) ENO-wavelet Transforms for Piecewise Smooth Functions. SIAM J Numer Anal 40(4):1369–1404
26. Tony Chan F, Zhou HM (2003) ENO-wavelet Transforms and Some Applications In: Stoeckler J, Welland GV (eds) Beyond Wavelets. Academic Press, New York
27. Chan TF, Zhou HM (2000) Optimal Constructions of Wavelet Coefficients Using Total Variation Regularization in Image Compression, CAM Report, No 00–27. Dept of Math, UCLA
28. Chui CK (1997) Wavelet: A Mathematical Tool for Signal Analysis. SIAM, Philadelphia
29. Chui CK, Wang J (2007) Wavelet-based Minimal-Energy Approach to Image Restoration. Appl Comp Harmon Anal 23(1):114–130
30. Claypoole P, Davis G, Sweldens W, Baraniuk R (1999) Nonlinear Wavelet Transforms for Image Coding. Correspond. Author: Baraniuk, Dept of Elec and Comp Sci; Submit to IEEE Transactions on Image Processing 12(12):1449 –1459
31. Coifman R, Donoho D (1995) Translation invariant de-noising In: Antoniadis A, Oppenheim G (eds) Wavelets and Statistics. Springer, New York, pp 125–150
32. Coifman R, Maggioni M (2006) Diffusion Wavelets. Appl Comp Harm Anal 21(11):53–94
33. Daubechies I (1988) Orthonormal Bases of Compactly Supported Wavelets. Comm Pure Appl Math 41:909–996
34. Daubechies I (1992) Ten Lectures on Wavelets. SIAM, Philadelphia
35. Donoho D (2004) Compressed Sensing. Tech Report, Stanford University
36. Donoho D, Huo X (2002) Beamlets and multiscale image analysis. In: Barth TJ, Chan T, Haimes R (eds) Multiscale and Multiresolution Methods. Lecture Notes in Computational Science and Engineering, vol 20. Springer, pp 149–196
37. Donoho D, Johnstone I (1995) Adapting to Unknown Smoothness via Wavelet Shrinkage. J Amer Stat Assoc 90:1200–1224
38. Dugatkin D, Zhou HM, Chan TF, Effros M (2002) Lagrangian Optimization of A Group Testing for ENO Wavelets Algorithm. Proceedings to the 2002 Conference on Information Sciences and Systems, Princeton University, New Jersey, 20–22 March, 2002
39. Durand S, Froment J (2001) Artifact Free Signal Denoising with Wavelets In: Proceedings of ICASSP'01, vol 6. pp 3685–3688
40. Geman S, Geman D (1984) Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Trans Patt Anal Mach Intell 6:721–741
41. Gonzalez RC, Woods RE (1993) Digital Image Processing. Addison Wesley, Reading
42. Harten A (1994) Multiresolution Representation of Data, II General Framework. Dept of Math, UCLA, CAM Report 94-10
43. Harten A, Engquist B, Osher S, Chakravarthy S (1987) Uniformly High Order Essentially Non-Oscillatory Schemes, III. J Comput Phys 71:231–303
44. Hernandez E, Weiss G (1996) A First Course on Wavelets. CRC Press, Boca Raton
45. Hong ES, Ladner RE (2000) Group testing for image compression. IEEE Proceedings of the Data Compression Conference, Snowbird, UT, March 2000, pp 2–12
46. Jain AK (1989) Fundamentals of Digital Image Processing. Prentice Hall, Englewood Cliffs
47. Kass M, Witkin A, Terzopoulos D (1987) Snakes: Active Contour Models. Int J Comput Vis 1:321–331
48. Malgouyres F, Guichard F (2001) Edge Direction Preserving Image Zooming: a Mathematical and Numerical Analysis. SIAM J Num Anal 39(1):1–37
49. Mallat S (1989) Multiresolution Approximation and Wavelet Orthonormal Bases of $L^2(R)$. Trans Amer Math Soc 315:69–87
50. Mallat S (1989) A Theory of Multiresolution Signal Decomposition: The Wavelet Representation. IEEE Trans PAMI 11:674–693
51. Mallat S (1998) A Wavelet Tour of Signal Processing. Academic Press, San Diego
52. Masnou M, Morel J (1998) Level-lines Based Disocclusion. IEEE ICIP 3:259–263
53. Meyer Y (1992) Wavelets and operators, Advanced mathematics. Cambridge University Press, Cambridge
54. Morel JM, Solimini S (1994) Variational Methods in Image Segmentation. Birkhauser, Boston
55. Mumford D, Shah J (1989) Optimal Approximation by Piecewise Smooth Functions and Associated Variational Problems. Comm, Pure Appl Math 42:577–685

56. Osher S, Fedkiw R (2002) Level Set Methods and Dynamic Implicit Surfaces. Springer, New York
57. Le Pennec E, Mallat S (2000) Image Compression with Geometrical Wavelets In: IEEE Conference on Image Processing (ICIP), Vancouver, September 2000
58. Perona P, Malik J (1990) Scale-space and edge detection using anisotropic diffusion. IEEE T Pattern Anal 12(7):629–639
59. Rudin L, Osher S, Fatemi E (1992) Nonlinear Total Variation Based Noise Removal Algorithms. Phys D 60:259–268
60. Said A, Pearlman W (1996) A new, fast, and efficient image codec based on set partitioning in hierarchical trees. IEEE Trans Circuits Syst Video Technol 6(3):243–250
61. Sapiro G (2001) Geometric Partial Differential Equations and Image Processing. Cambridge University Press,Cambridge
62. Sapiro G, Tannenbaum A (1993) Affine Invariant Scale-Space. Internet J Comput Vis 11:25–44
63. Shapiro J (1993) Embedded Image Coding Using Zerotrees of Wavelet Coefficients. IEEE Trans Signal Process 41(12): 3445–3462
64. Strang G, Nguyen T (1996) Wavelets and Filter Banks. Wellesley-Cambridge Press, Wellesley
65. Welland GV (ed) (2003) Beyond Wavelets. Academic Press, New York
66. Weickert J (1998) Anisotropic Diffusion in Image Processing. ECMI Series, Teubner, Stuttgart
67. Xu J, Osher S (2006) Iterative Regularization and Nonlinear Inverse Scale Space Applied to Wavelet Based Denoising. UCLA CAM Report 06–11, March 2006

Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis

Kelin Wang[1,2], Yan Hu[2], Jiangheng He[1]
[1] Pacific Geoscience Centre, Geological Survey of Canada, Sidney, Canada
[2] School of Earth and Ocean Sciences, University of Victoria, Victoria, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
Stable and Critical Coulomb Wedges
Dynamic Coulomb Wedge
Stress Drop and Increase in a Subduction Earthquake
Tsunamigenic Coseismic Seafloor Deformation
Future Directions
Acknowledgments
Bibliography

Glossary

Subduction zone earthquake cycle Megathrust fault, the interface between the two converging lithospheric plates at a subduction zone, moves in a stick-slip fashion. In the "stick" phase, the fault is locked or slips very slowly, allowing elastic strain energy to be accumulated in both plates around the fault. Every few decades or centuries, the fault breaks as high-rate "slip" to release the strain energy, causing a large or great earthquake, usually accompanied with a tsunami. An interseismic period and the ensuing earthquake together is called a subduction zone earthquake cycle. The word cycle by no means implies periodicity. Neighboring segments of the same subduction zone may exhibit different temporal patterns of earthquake cycles.

Accretionary wedge (prism) At some subduction zones, as one plate subducts beneath the other, some sediment is scraped off the incoming plate and accreted to the leading edge of the upper plate. Because of its wedge shape, the accreted sedimentary body is called the accretionary wedge (or accretionary prism). If all the sediment on the incoming plate is subducted, there is still a sedimentary wedge in the frontal part of the upper plate, but it is usually very small and consists of sediments derived from the upper plate by surface erosion.

Coulomb plasticity Coulomb plasticity is a macroscopic, continuum description of the most common type of permanent deformation of Earth materials such as sand, soil, and rock at relatively low temperature and pressure and is widely used in civil engineering and Earth science. In detail, the deformation mechanism is actually brittle shear failure, with or without emitting elastic wave energy. The macroscopic yield criterion is the Coulomb's law, in which shear strength increases linearly with confining pressure. If the strength does not change with permanent deformation, the material is said to be perfectly plastic. Note that in Earth science the word plasticity is also used to indicate thermally activated creep, but it is very different from the meaning used herein.

Velocity-weakening and strengthening These are macroscopic descriptions of dynamic frictional behavior of contact surfaces. Velocity-weakening, featuring a net decrease in frictional strength with increasing slip rate, is the necessary condition for a fault to produce earthquakes. It differs from slip-weakening in that a velocity-weakened fault will regain its strength when the slip slows down or stops. Velocity-strengthening is the opposite of velocity-weakening and is the necessary condition for a fault to resist earthquake rupture. Detailed physical processes on the contact surfaces or within the fault zones controlling their frictional behavior are still being investigated.

Definition of the Subject

Mechanics of wedge-shaped geological bodies such as accretionary prisms at subduction zones and fold-and-thrust belts at collision zones is of great scientific interest, mainly because it enables us to use the observed morphology and deformation of the wedge-shaped bodies to constrain properties of the thrust faults underlying them. Davis et al. [12] drew analogy between these wedge-shaped bodies and the sand wedge in front of a moving bulldozer and established a mathematical model. Their work triggered wide applications of wedge mechanics to geology. The fundamental process described in wedge mechanics is how gravitational force, in the presence of a sloping surface, is balanced by basal stress and internal stress. The internal state of stress depends on the rheology of the wedge. The most commonly assumed wedge rheology for geological problems is perfect Coulomb plasticity [12], and the model based on this rheology is referred to as the Coulomb wedge model.

The Coulomb wedge model was designed to explain geological processes of timescale of hundreds of thousands of years. The state of stress is understood to be an average over time, and the wedge is assumed to be in a critical state, that is, uniformly at the Coulomb yield stress. In the application of the model to the sedimentary wedges at subduction zones, attention is now being paid to the temporal changes in stress and pore fluid pressure associated with great subduction earthquakes which have recurrence intervals of decades to centuries. To account for the short-term stress changes, Wang and Hu [50] expanded the Coulomb wedge model by introducing the elastic – perfectly Coulomb plastic rheology. The expanded Coulomb wedge model links long-term geology with coseismic processes and provides a new perspective for the study of subduction zone earthquakes, tsunami generation, frontal wedge structure, and forearc hydrogeology.

Introduction

Sloping surfaces are commonly dealt with in engineering problems such as dam design and landslide hazard mitigation. In the presence of a sloping surface, materials under gravitational force tend to "flow" downhill and thus generate tensile stress, but whether collapse actually occurs depends on the strength of the material. Materials used in construction can easily sustain stresses caused by the presence of a vertical surface, but a material of no shear strength such as stationary water cannot support any surface slope. A geological wedge, such as the accretionary prism at a subduction zone (see Fig. 1a), overlies a dipping fault, and hence its internal stress is controlled by the strength of the fault as well. If the basal fault is a thrust fault and is strong relative to the strength of the wedge material, the wedge can undergo compressive failure. Wedge mechanics thus consists of two aspects: the frictional behavior of the basal fault and the deformation of the wedge itself. Given a wedge with density ρ, surface slope angle α, and basal dip β (see Fig. 1b), Elliot [16], Chapple [8], and later workers all considered the following equations of force balance (exact form varies between publications depending on the assumed coordinate systems)

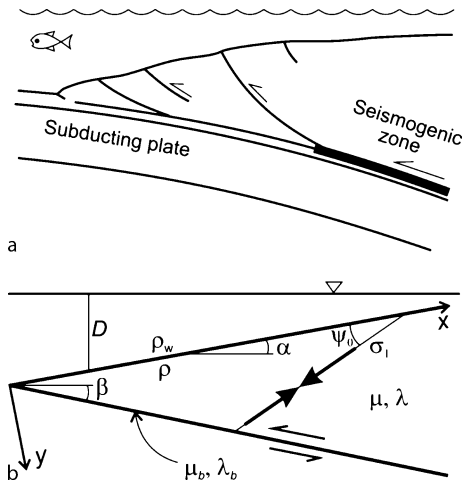

Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 1

a Schematic illustration of a subduction zone accretionary wedge. b Coordinate system used in this article (*x*, *y*). The surface of the wedge and the subduction fault are simplified to be planar, with slope angle α and dip β, respectively, and the *x* axis is aligned with the upper surface. ψ_o is the angle between the maximum compressive stress σ_1 and the upper surface. *D* is water depth. ρ and ρ_w are densities of the wedge material and overlying water, respectively. μ, λ, μ_b and λ_b are as defined in Eqs. (2) and (3)

$$\frac{\partial \sigma_x}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} - \rho g \sin\alpha = 0 \,, \tag{1a}$$

$$\frac{\partial \tau_{xy}}{\partial x} + \frac{\partial \sigma_y}{\partial y} + \rho g \cos\alpha = 0 \,, \tag{1b}$$

where g is gravitational acceleration, σ_x and σ_y are normal stresses, and τ_{xy} is shear stress.

It was Davis et al. [12] who introduced Coulomb plasticity into wedge mechanics. Coulomb plasticity was initially proposed by French engineer Coulomb in 1773 to describe the mechanical strength of soil and sand. The Coulomb failure criterion in its simplest form is

$$\tau = S - \mu(\sigma_n + P) = S - \mu\bar{\sigma}_n \,, \tag{2}$$

where τ is shear strength, σ_n is normal stress, S is cohesion,

and $\mu = \tan\varphi$ is the coefficient of internal friction with φ called the internal friction angle. P is the pressure of fluids present in the pore space between solid grains and in various small fractures and is loosely referred to as the pore fluid pressure. Note that the effective stress $\bar{\sigma}_n = \sigma_n + P$ is normal stress with P *subtracted*. The plus sign is due to the custom in mechanics (except rock mechanics) that compressive stress is defined to be negative, but pressure, although also compressive, is defined to be positive. Here the plane on which the stress is evaluated is oriented in any arbitrary direction, but failure will start on the set of planes that meets the above criterion. This is a generalization of the Coulomb friction criterion for a fixed fault plane

$$\tau_b = S_b - \mu_b(\sigma_{n_b} + P_b) = S_b - \mu_b\bar{\sigma}_{n_b}\,, \qquad (3)$$

where S_b is the cohesion of the fault, $\mu_b = \tan\varphi_b$ is its friction coefficient, and P_b is fluid pressure along the fault. S_b is usually negligibly small and almost always taken to be zero, and μ_b is normally significantly lower than μ. A well developed fault such as a plate boundary fault is a zone of finite thickness filled with gouge material, so that the "friction" described by (3) or other friction laws is actually the shear deformation of the gouge in the fault zone, and P_b is actually pore fluid pressure of the gouge. Fault gouge is often made very weak by the presence of hydrous minerals [5,42], such that the collective strength of the fault zone material is much less than the strength of the rocks on both sides. Another process that may weaken the fault is that the local hydrogeological regime may dynamically maintain P_b in the fault zone to stay higher than P on both sides [14]. For both Coulomb plasticity and Coulomb friction, strength increases with depth because of the increasing pressure thus normal stress.

It has long been recognized that Coulomb plasticity, featuring strong depth dependence, applies to the shallow part of Earth's lithosphere. The most common example is the use of Byerlee's law of rock friction [6] to describe brittle strength in "Christmas-tree"-like vertical strength-depth profiles of the lithosphere. The Byerlee's law is an empirical Coulomb friction law. By assuming that faults, i. e., potential failure planes, are oriented in all directions, we regard the brittle part of the lithosphere as being Coulomb plastic. This example also illustrates how a system of numerous discrete structures can be regarded as a continuum at a much larger scale. Similarly, although a geological wedge actually consists of numerous blocks divided by fractures, Coulomb plasticity can be used to describe its overall rheology. However, specific values of friction parameters for submarine wedges may be quite different from those in the Byerlee's law.

The Coulomb wedge model explains how the geometry (taper) of the wedge is controlled by the interplay between the gravitational force, the strength of the wedge material, and the strength of the basal fault. The wedge strength and fault strength are both strongly influenced by pore fluid pressure, and the most popular application of the model is to estimate pore fluid pressure from observed wedge geometry. Since the work of Davis et al. [12], more rigorous analytical solutions have been derived [9,10,18,50,55], and some extensions have been proposed, e. g., [2,3,17,54]. Lithospheric scale numerical models are often used to study the evolution of geological wedges including such effects as erosion and sedimentation in collision or subduction zone settings, e. g., [20,53]. Utilizing the bulldozer – sand wedge analogy, important physical insights have been obtained from sandbox experiments [7,11,29,30,31,35,41,52]. For a list of applications of the Coulomb wedge model to subduction-zone accretionary prisms, see [50].

Stable and Critical Coulomb Wedges

Depending on the state of stress, a Coulomb wedge can be at a critical state, that is, everywhere at Coulomb failure, or a stable (also referred to as supercritical) state, that is, everywhere not at failure (see Fig. 2). The taper of a critical wedge will not change if the stress does not change; note that a critical wedge of stable geometry should not be confused with a stable wedge. Stress solutions have been derived for both critical and stable states. Here we only summarize the simplest, exact solution for a cohesionless wedge ($S = 0$) derived in the coordinate system shown in Fig. 1b, because of the convenience of its application as compared to other solutions. The wedge is assumed to be elastic – perfectly Coulomb plastic. If it is at the critical

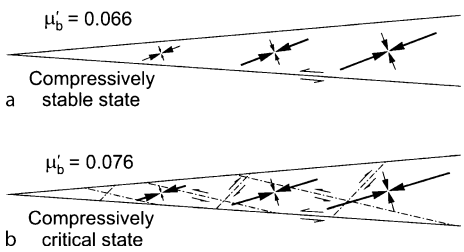

Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 2

An example to show how stresses in an elastic – perfectly Coulomb-plastic wedge, with $\alpha = 5°$, $\beta = 4°$, $\mu = 0.6$, and $\lambda = 0.86$, are affected by basal friction $\mu'_b = \mu_b(1 - \lambda_b)$. Converging arrows represent principal stresses, with the larger one being σ_1. a Compressively stable state. b Compressively critical state. *Dot-dashed* lines are plastic slip lines (potential failure planes)

state, it obeys (2); if it is at the stable state, it obeys the Hooke's law of elasticity. The basal thrust fault obeys (3). The wedge is assumed to be under water of density ρ_w and depth D (a function of x; see Fig. 1b). By defining a Hubbert–Rubey fluid pressure ratio within the wedge [9]

$$\lambda = \frac{P - \rho_w g D}{-\sigma_y - \rho_w g D}\,, \tag{4}$$

and a similar parameter along the basal fault [51]

$$\lambda_b = \frac{P_b - \rho_w g D + \lambda H}{-\sigma_y - \rho_w g D + H}\,, \tag{5a}$$

where

$$H = \frac{\sigma_y - \sigma_n}{1 - \lambda}\,, \tag{5b}$$

and assuming $S_b = 0$, we can rewrite (3) into

$$\tau_b = -\mu_b \bar{\sigma}_{n_b} = -\mu_b'' \bar{\sigma}_n\,, \tag{6a}$$

where

$$\mu_b'' = \frac{1 - \lambda_b}{1 - \lambda}\mu_b = \frac{\mu_b'}{1 - \lambda}\,. \tag{6b}$$

The strength of the basal fault is represented by $\mu_b' = \mu_b(1 - \lambda_b)$ which always appears as a single parameter and is commonly referred to as the effective friction coefficient. The hydrological process in the fault zone may differ from that in the wedge and thus cause a sharp gradient in fluid pressure across the wedge base. By allowing λ_b to be different from λ, we use a pressure discontinuity to approximate the sharp gradient. Because stress is continuous across the basal fault, the discontinuity in pore fluid pressure leads to a discontinuity in effective stress. The second equality of Eq. (6a) shows the relation between the effective stress along the fault ($\bar{\sigma}_{n_b}$) and the effective stress just above the fault ($\bar{\sigma}_n$). The establishment of (6) allows the following exact stress solution to be derived. In this expression, all stress components are normalized by $\rho g y$ (e. g., $\bar{\sigma}_x' = \bar{\sigma}_x/\rho g y$).

$$\bar{\sigma}_x' = -m(1 - \lambda)\cos\alpha\,, \tag{7a}$$

$$\bar{\sigma}_y' = -(1 - \lambda)\cos\alpha\,, \tag{7b}$$

$$\tau_{xy}' = (1 - \rho')\sin\alpha\,, \tag{7c}$$

where $\rho' = \rho_w/\rho$, and the effective stress ratio $m = \bar{\sigma}_x/\bar{\sigma}_y$ depends on whether the wedge is stable or critical. If the wedge is in a stable state (elastic) [50],

$$m = 1 + \frac{2\left(\tan\alpha' + \mu_b''\right)}{\sin 2\theta\left(1 - \mu_b''\tan\theta\right)} - \frac{2\tan\alpha'}{\tan\theta}\,, \tag{8}$$

where $\theta = \alpha + \beta$, and,

$$\tan\alpha' = \frac{1 - \rho'}{1 - \lambda}\tan\alpha\,. \tag{9}$$

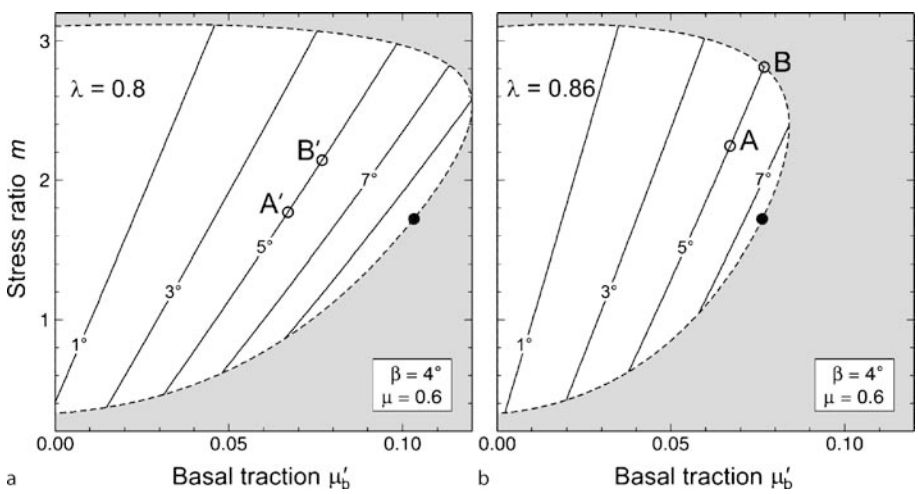

Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 3

Effective stress ratio *m* (Eq. (7)) as a function of basal friction $\mu_b' = \mu_b(1 - \lambda_b)$ for wedges of the same basal dip ($\beta = 4°$) but different surface slope angles (α) as labelled on the curves (*solid* lines). a and b are for two different pore fluid pressure ratio values within the wedge. Each curve is terminated at the extensionally critical state at a lower μ_b' and the compressively critical state at a higher μ_b'. The end points (connected by a *dashed* line) outline the stable region (*white*). No solution exists outside this region. The *solid circle* marks the state in which the surface slope is at the angle of repose. It divides the line of critical states (*dashed* line) into the compressive part (above) and extensional part (below). *Open circles* in b labelled A and B mark the states shown in Fig. 2a and b, respectively. State A′ in (a) is for the same wedge with the same basal friction as state A in (b) except for a lower pore fluid pressure ratio, and state B′ in (a) corresponds to state B in (b) in the same fashion. Comparison of B with B′ shows how an increase in pore fluid pressure weakens the wedge

The angle ψ_0 between the most compressive stress σ_1 and the upper surface is uniform (see Fig. 1b). A more general solution for a purely elastic wedge can be found in [23]. If the wedge is in a critical state (perfectly plastic) [9]

$$m = m^c = 1 + \frac{2\tan\alpha'}{\tan 2\psi_0^c}\,, \tag{10}$$

where ψ_0^c is the value of ψ_0 in the critical state and is given by the following relation

$$\frac{\sin\varphi \sin 2\psi_0^c}{1 - \sin\varphi \cos 2\psi_0^c} = \tan\alpha'\,. \tag{11}$$

In the above expressions, superscript c indicates critical state. If the wedge has a cohesion that is proportional to depth, (10) and (11) will be modified only slightly [50,55].

A wedge of fixed geometry has two m^c values. The lower one defines the extensionally critical state, in which the wedge is on the verge of gravitational collapse. This occurs if friction along the basal fault is very low relative to the strength of the wedge material. The higher one defines the compressively critical state, in which the wedge is everywhere on the verge of thrust failure. If m lies between these two critical values, the wedge is in a stable state and only experiences elastic deformation. A change of basal friction μ_b' will cause a change in m and thus may potentially cause the wedge to switch between the stable and critical states (see Fig. 3). An example of a wedge being in a stable or compressively critical state as controlled by basal friction is shown in Fig. 2. The plastic slip lines (potential failure planes) in the critical wedge (see Fig. 2b) are reminiscent of the out-of-sequence faults in a real accretionary prism (see Fig. 1a).

If we fix the values of all material properties, the critical-wedge solution ($m = m^c$) defines a relation between α and β representing all possible geometries of a critically tapered wedge (dashed line in Fig. 4). This is a very commonly used diagram, in which the lower branch of the $\alpha - \beta$ curve represents compressively critical states, and the upper branch represents extensionally critical states. Combinations of α and β outside of the stability region comprise unstable geometries and cannot exist in steady state. If sedimentary wedges of subduction zones are compressively critical, their observed $\alpha - \beta$ pairs should line up with the lower branch. However, it has been shown that most of them fall in the stable or even extensionally unstable region of this type of diagram [30,44]. To fit observations, we need to move the lower branch upward by a significant amount. In order to do this, we need to assume either a weaker wedge or a stronger basal fault, or both. This is more simply illustrated by Fig. 3. Given wedge geometry and internal friction, if the state of the wedge is to be changed from stable to compressively critical, we need to have a higher friction (μ_b') along the basal fault (i. e., greater stress coupling) and/or higher pore fluid pressure within the wedge (i. e., greater λ weakening the wedge material by reducing effective pressure). Wang and Hu [50] proposed that higher μ_b' and λ can occur at the time of a great earthquake and introduced the concept of dynamic Coulomb wedge.

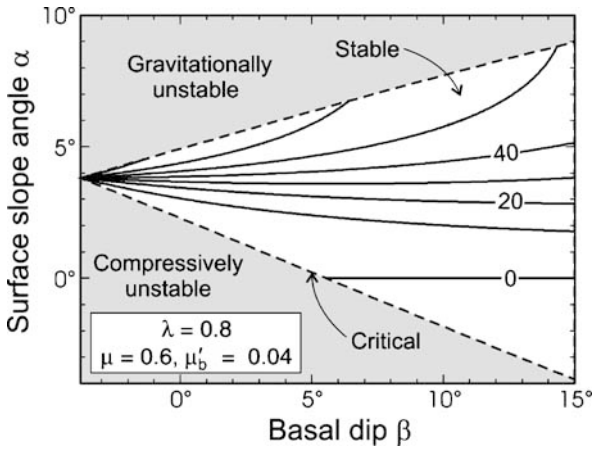

Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 4

Possible wedge geometry (α and β) given material strength and basal friction. The *dashed* line indicates critical wedge geometry, with the upper and lower branches representing extensionally and compressively critical states, respectively. Contours of ψ_0 (in degrees) are shown in the stable region (*white*)

Dynamic Coulomb Wedge

The concept of dynamic Coulomb wedge is based on the widely recognized frictional behavior of subduction faults. Ignoring the presence of along-strike variations of frictional properties, we can summarize the frictional behavior in a simplified cross-section view (see Fig. 5). The seismogenic zone exhibits velocity-weakening behavior: It weakens in response to high-rate slip, resulting in slip instability, that is, earthquakes. The segments updip and downdip of the seismogenic zone exhibits velocity-strengthening: They strengthen at the time of the earthquake to develop higher stress to resist rupture, but they may slip aseismically after the earthquake to relieve the high stress attained during the earthquake. We assume that the actively deforming sedimentary wedge overlies the updip segment (also see Fig. 1a).

Microscopic mechanisms for the velocity-weakening behavior of the seismogenic zone and the velocity-strengthening behavior of the aseismic zones are subjects of intense research. The aseismic behavior of the deeper

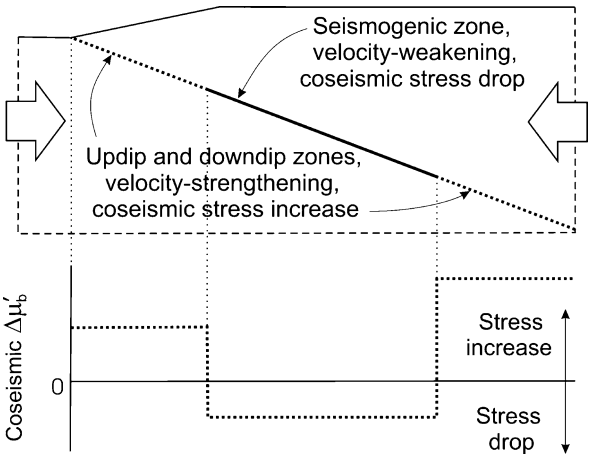

Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 5

Schematic illustration of the subduction zone model considered in this work. *Large arrows* represent interseismic strain accumulation. An earthquake is represented by a sudden decrease in the effective friction coefficient μ'_b of the seismogenic zone by $\Delta\mu'_b$. Coseismic strengthening of the updip and downdip zones is represented by a sudden increase in their μ'_b values

part of any fault is intuitively easy to comprehend; higher temperature at greater depths increasingly enhances viscous deformation and inhibits brittle faulting. There are different physical explanations for the velocity-weakening behavior of the seismogenic zone as summarized in [37]. The mechanism responsible for the velocity-strengthening behavior of the updip segment is yet to be identified, although it is widely accepted that the presence of clay minerals has something to do with it [24,33,34,45,46]. Laboratory experiments indicate that dilatancy of granular fault zone material during fast slip can lead to velocity-strengthening [32]. We think part of the reason for the velocity-strengthening behavior of the updip segment may be its inability to localize into a very thin slip zone. Seismic rupture occurs along very thin slip zones of a few millimeters thickness that are parts of a thicker fault zone [40]. Fast slip of the updip segment, if triggered by the rupture of the deeper seismogenic zone, may tend to drag along fault zone materials over a more distributed band and thus meet greater resistance. This view is different from the velocity-strengthening process described by laboratory-derived rate- and state-dependent friction laws [15,39] in which dynamic changes in the thickness of the slip zone plays no role.

Regardless of the microscopic mechanisms, the downdip variation of the frictional behavior is expected to bring direct consequences to wedge deformation. In an earthquake, thrust motion of the seismogenic zone causes the frontal wedge to be pushed from behind, and velocity-strengthening of the updip megathrust segment gives rise to higher stress at its base. If the wedge is originally in a stable state, the coseismic strengthening of the basal fault may increase the m value in (7a) from subcritical to critical. After the earthquake, when the seismogenic zone has returned to a locked state, the stress along the updip segment will relax. The decrease in μ'_b leads to a smaller m, and therefore the frontal wedge returns to a stable state.

If we ignore the change in fluid pressure, the above described process can be seen as the stress ratio m moving up-and-down along one of the solid lines in Fig. 3 in response to changes in μ'_b in earthquake cycles. During a big earthquake, it will hit the upper end of the line (m^c). The state of stress for a stable wedge before the earthquake ($m < m^c$) and that for a compressively critical wedge during the earthquake ($m = m^c$) are illustrated by examples in Fig. 2a and b, respectively. Fluid pressure variation should not be ignored, however. For example, the same basal friction as shown in Fig. 2b (also see point B in Fig. 3b) will not drive the wedge to failure if the pore fluid pressure is lower, as shown by point B′ in Fig. 3a.

During an earthquake, the sudden compression of the frontal wedge will cause its internal pore fluid pressure to increase, coseismically weakening the wedge material. By comparing Fig. 3a and b, we can see that if the pore fluid pressure ratio in the wedge is higher, the increase in basal friction required to push the wedge into a critical state can be smaller. We may envisage the following scenario. The pore fluid pressure in a frontal wedge may decrease to some degree over the interseismic period due to fluid drainage through fractures and stress relaxation, and the mechanical state of the wedge before an earthquake can be represented by point A′ in Fig. 3a as opposed to point A in Fig. 3b. An earthquake not only causes the basal friction to increase but also the pore fluid pressure within the wedge to rise, such that the wedge enters a critical state represented by point B in Fig. 3b. Therefore, the coseismic strengthening of the basal fault and coseismic weakening of the wedge both work toward bringing the wedge to failure.

All previous applications of the Coulomb wedge model to subduction zones assume $m = m^c$. The dynamic Coulomb wedge model of [50] explains the meaning of this long-term m^c: At least as an end-member scenario, it is the value of m briefly achieved in numerous large earthquakes. The "average" basal stress that determines the wedge geometry in long-term Coulomb wedge models is actually the peak stress achieved at the time of large earthquakes. Thus, the common illustration of the peaceful scene of a bulldozer pushing a sand wedge in classical

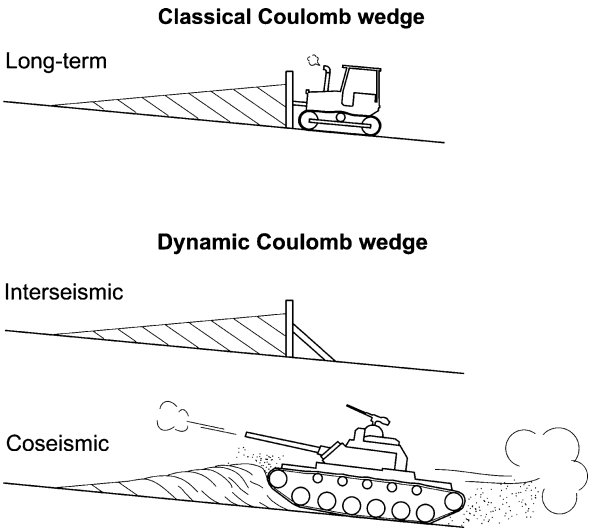

Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 6
Cartoon illustrating the difference between the classical Coulomb wedge model and the dynamic Cou lomb wedge model for subduction zone accretionary prisms. In the classical wedge model, $m = m^c$. In the dynamic wedge model, $m < m^c$ in the interseismic period but $m = m^c$ at the time of a large earthquake. See Eqs. (7) and (10) for definition of m and m^c

Coulomb wedge papers (see Fig. 6a) should be modified to reflect the unpleasant reality of the world (see Fig. 6b).

Stress Drop and Increase in a Subduction Earthquake

It is important to know the possible amount of stress increase in the frontal wedge for a given earthquake. The increase cannot be arbitrarily large; it is limited by the level of the "push" on the wedge from behind during an earthquake. For this purpose, a numerical model of a larger scale embracing the essential components of the subduction fault as shown in Fig. 5 must be considered, because the stress interaction between the frontal wedge and the material overlying the seismogenic zone cannot be handled by the analytical Coulomb wedge solutions. For an illustration, we consider the following model geometry, representative of most subduction zones. The subduction fault has a constant dip $\beta = 4°$. The frontal 50 km of the upper plate has a surface slope $\alpha = 5°$, representing the sedimentary wedge, and the rest of the upper plate has a flat surface. We wish to focus on the process of stress transfer from the seismogenic zone to the updip segment during an earthquake, and a static, uniform, and purely elastic model suffices. For simplicity, the effect of pore fluid pressure change on deformation is neglected.

We use a 2D plane-strain finite element model and simulate Coulomb friction along the fault using the method of Lagrange-multiplier Domain Decomposition [48]. The model boundaries are set sufficiently far away so that the model resembles a half-space. For numerical stability, we invoke gravity (assuming a rock density of 2800 kg/m^3 only when determining yield stress along the fault but exclude it from the deformation calculation. The effect of gravity on coseismic elastic deformation is very small and is neglected in most earthquake cycle deformation models, but gravity is important in the calculation of frictional slip of the fault.

We first generate a pre-stress field by moving the remote seaward and landward model boundaries toward each other against a locked fault. At this stage of "interseismic" strain accumulation, we use a μ'_b of 0.04 for the seismogenic and updip segments and 0.004 for the deeper segment. The low strength of the subduction fault is based on the weak fault argument as summarized in Wang and Hu (2006). The nearly zero strength of the deeper part represents a relaxed state after a long time of locking of the seismogenic zone. However, the absolute strength of the fault has no effect on our results. It is the incremental change in fault strength ($\Delta\mu'_b$) at the time of the earthquake that is relevant. A negative $\Delta\mu'_b$ represents the net effect of weakening, and a positive $\Delta\mu'_b$ represents the net effect of strengthening. The velocity-dependent evolution of $\Delta\mu'_b$ through time is not explicitly simulated.

Three examples are shown in Fig. 7. In all cases, $\Delta\mu'_b = -0.01$ is assigned to the 150-km wide seismogenic zone. This value is chosen to produce an average stress drop of a few MPa (see Fig. 7a), typical of values observed for great subduction earthquakes. The stress drop releases elastic strain energy initially stored in the system, leading to fault slip that represents an earthquake rupture. The deepest segment is assigned a sufficiently large positive $\Delta\mu'_b$ so that it cannot slip. The examples differ in the $\Delta\mu'_b$ values assumed for their 50 km wide updip segment, which is the coseismic increase in basal friction in the dynamic Coulomb wedge model.

Example 1 No trench-breaking rupture (solid line in Fig. 7b). In this case, the strengthening of the updip segment is $\Delta\mu'_b = 0.036$. This particular value of $\Delta\mu'_b$ creates a situation in which the entire updip segment is at failure but is just short of breaking the trench. This is the minimum value of the $\Delta\mu'_b$ of the updip segment required to prevent trench-breaking rupture and is denoted $\Delta\mu'_{b_t}$. The value of $\Delta\mu'_{b_t}$ depends on the product of the stress drop and the area of the seismogenic zone, a quantity we refer to as "force drop". That is, if the upper edge of the

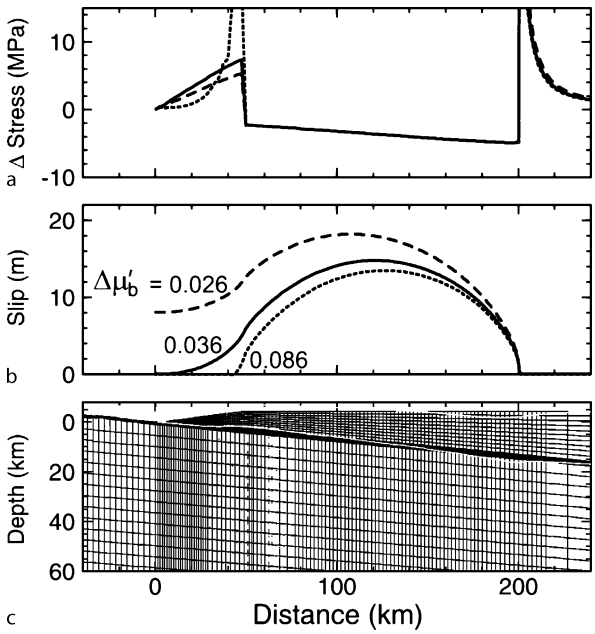

Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 7

Three examples of the stress transfer model. The examples differ in the coseismic strength increase of the updip fault segment, indicated in (b) as $\Delta\mu'_b$. a Shear stress drop (or increase) along the fault. b Slip distribution along the fault. c Central portion of the finite element mesh (*thin lines*). The "*thick line*" along the plate interface is actually a group of very densely spaced elements. *Thick gray lines* indicate deformed fault and surface after the earthquake (exaggerated by a factor of 2000)

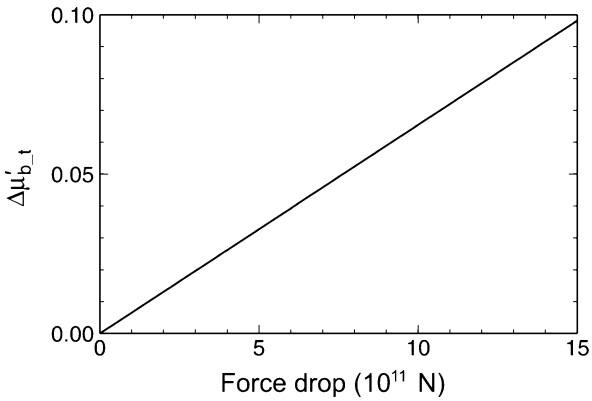

Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 8

Minimum increase in $\Delta\mu'_b$ of the updip megathrust segment (denoted $\Delta\mu'_{b_t}$) required to prevent trench-breaking rupture as a function of force drop along the seismogenic zone for the model shown in Fig. 7c, with the upper edge of the seismogenic zone fixed at 50 km from the trench

seismogenic zone is fixed, increasing its downdip width or stress drop gives the same result. For the same model geometry as used for this example, $\Delta\mu'_{b_t}$ as a function of force drop per unit strike length is shown in Fig. 8, assuming the upper edge of the seismogenic zone is fixed. Using a different model geometry or position of the seismogenic zone upper edge will change the slope of this function.

Example 2 Trench-breaking rupture (dashed line in Fig. 7b). Given the same force drop in the seismogenic zone, if $\Delta\mu'_b$ of the updip segment is less than $\Delta\mu'_{b_t}$, the rupture will break the trench. Knowing whether coseismic trench-breaking rupture exists or is common awaits future seafloor monitoring observations. This example shows that a trench-breaking rupture does not necessarily indicate the updip segment exhibits velocity-weakening. Conceivably, the slip of the velocity-strengthening updip segment may be slower than that of the seismogenic zone and may not generate much seismic waves.

Example 3 Fully buried rupture (dotted line in Fig. 7b). If $\Delta\mu'_b$ of the updip segment is greater than $\Delta\mu'_{b_t}$, rupture may only extend into its lower part. For a very high $\Delta\mu'_b$, most of the segment does not slip at all, because a tiny portion immediately updip of the seismogenic zone is sufficient to stop the rupture. This is just the buried-rupture scenario of the crack model commonly used in earthquake simulation [21]. Because most of the updip segment is "protected" and does not experience coseismic stress increase, this scenario is not applicable to the dynamic Coulomb wedge model. A very large stress increase just updip of the seismogenic zone is considered unrealistic.

These examples show the consequences of changes in the strength of the basal fault of the frontal wedge resulting from the rupture of the seismogenic zone. Whether the given strength increase $\Delta\mu'_b$ can drive the wedge from a stable state into a critical state depends on two factors. First, it depends on the value of μ'_b before the earthquake. The value of 0.04 used in the above examples is only one of the numerous possible values. If μ'_b is already near a critical value, that is, m in Fig. 3 for a given α is already near m^c, a small increase will do. Conceivably, μ'_b before an earthquake may be relatively high if the strengthened state of the fault caused by the previous earthquake has not fully relaxed. Second, given μ'_b, it depends on the strength of the wedge material. A weaker wedge becomes critical at a lower $\Delta\mu'_b$. The average internal friction value μ of an actively deforming frontal prism is lower than the rest of the lithosphere because of its low degree of consolidation and high degree of fracturing. The pore fluid pressure within it, represented by λ in the Coulomb wedge model, may increase due to coseismic compression of the prism,

further weakening the wedge, as discussed in Sect. "Dynamic Coulomb Wedge".

Tsunamigenic Coseismic Seafloor Deformation

To understand the process of tsunami generation by a great subduction zone earthquake, we must know how the seafloor deforms at the time of the earthquake. Despite the dramatic worldwide improvement of geodetic, seismological, and oceanographic monitoring networks over the past few decades, our knowledge of coseismic seafloor deformation (CSD) is surprisingly poor and is based almost entirely on theoretical models. The problem is the rarity of near-field observations. Except for seafloor pressure sensor records at the time of the M8.2 Tokachi-oki earthquake of 2003 [1] and continuous GPS measurements from islands very near the Sumatra trench at the time of the M8.7 Nias-Simeulue earthquake of 2005 [4,22], most observations are made at sites too remotely located to resolve reliably CSD within about 100 km of the trench. The lack of near-field CSD information causes severe nonuniqueness in the inversion of tsunami, seismic, and geodetic data to determine coseismic slip patterns of the shallow part of the subduction fault, for which the only cure is to introduce a priori constraints on the basis of theoretical models. Until the situation is improved by the establishment of seafloor monitoring systems, we must continue to resort to what we are able to deduce from these models.

The largest uncertainty in our knowledge of the processes that control CSD is how coseismic slip along the subduction fault varies in the downdip direction [21,49]. It may overshadow uncertainties in our knowledge of the timescale of the deformation and the spatial variation of mechanical properties of the rock medium. In comparison, along-strike variations of the coseismic slip, usually described in terms of "asperities", are much easier to determine using high-density terrestrial monitoring networks. Theoretical models discussed above can help us understand the downdip slip distribution. Using the same type of model as shown in Fig. 5 and Fig. 7 but with a realistic, curved fault, we illustrate how the frictional behavior of the updip segment affects the CSD (see Fig. 9). The simulated earthquakes in the two shown examples have the same "size", quantified by the seismic moment – the product of rigidity, slip area, and average slip, but they cause very different CSD patterns.

Example 4 No trench-breaking rupture (solid line in Fig. 9a). This model is similar to the first example in the preceding section in that the updip segment is assumed to strengthen by $\Delta\mu'_{b_t}$, and the rupture is on the verge of breaking the trench.

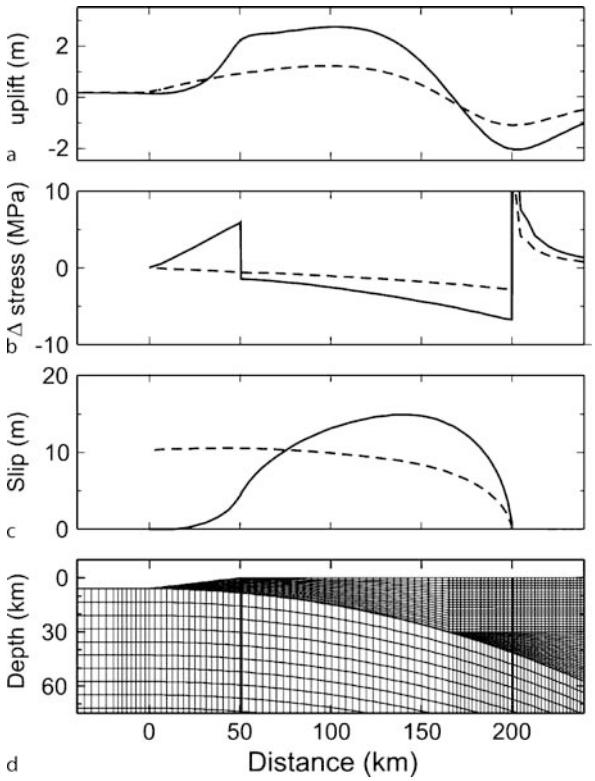

Wedge Mechanics: Relation with Subduction Zone Earthquakes and Tsunamis, Figure 9

Two examples showing how the frictional behavior of the updip segment (see Fig. 5) affects CSD. In one example (*solid line*), the segment strengthens by $\Delta\mu'_{b_t}$, and in the other example (*dashed line*), the segment weakens and thus becomes part of the seismogenic zone. The simulated earthquakes in both examples have the same seismic moment. a Surface uplift. b Stress drop (or increase) along the fault. c Slip distribution along the fault. d Central part of the finite element mesh (*thin lines*). The two *vertical "thick lines"* at distances 50 km and 200 km bracketing the seismogenic zone of the model of no trench-breaking rupture are actually groups of very densely spaced elements

Example 5 Full trench-breaking rupture (dashed line in Fig. 9a). In this model, there is no velocity-strengthening updip segment, and the seismogenic zone extends to the trench. The stress drop of the resultant much wider seismogenic zone features a monotonic increase from the trench.

The model of full trench-breaking rupture yields much smaller vertical CSD than does the model of no trench-breaking rupture. The reason is two-fold. First, without a velocity-strengthening updip segment to resist rupture, the maximum slip occurs in the most shallowly dipping near-trench part of the fault where seafloor displacement is predominantly horizontal. This effect would not be

obvious had a straight fault geometry been used in the model. Second, without the resistance of an updip segment, the upper plate does not experience horizontal compression and the resultant vertical expansion. If we rescale the two models so that they have the same maximum slip, the model of full trench-breaking rupture will have a greater seismic moment but still a much lower vertical CSD [49]. This result demonstrates the importance of the frictional behavior of the shallowest fault segment in affecting seafloor uplift. However, it addresses only one aspect of tsunamigenic CSD. Many other factors contribute to tsunami generation. For example, although the full trench-breaking model yields very low vertical CSD, its horizontal CSD may contribute to tsunami generation. If the seafloor slopes at angle α, its horizontal motion D in the slope direction should raise the seafloor by $D \tan \alpha$ relative to a fixed water column above, an effect addressed by Tanioka and Satake [43]. The speed of the coseismic slip is also an important factor in tsunami generation. In some rare cases, the rupture is too low to generate much seismic wave energy yet fast enough to generate rather large tsunamis, giving rise to a class of earthquake called tsunami earthquakes [27].

Elastic deformation of the ocean floor as discussed above is the primary cause of tsunami generation during subduction earthquakes, but inelastic deformation can be locally important. For example, although the lower continental slope of active margins is on average the expression of a critically tapered Coulomb wedge, seafloor topography at these margins is rugged at smaller scales due to sedimentation, erosion, and deformation processes, and where the local surface slope is sufficiently high earthquake shaking may trigger gravitational failure. Such submarine "landslides" may have a locally significant effect on tsunami generation. Another potentially important inelastic process is the coseismic activation of out-of-sequence thrust faults (splay faults) in the accretionary prism (Fig. 1a). Splay faults are much more steeply dipping, and their thrust motion will serve to "redirect" the low-angle slip of the megathrust to a higher angle and thus may greatly enhance local seafloor uplift and contribute to tsunami generation [19,36]. As mentioned in the Introduction, from the continuum perspective, such faulting is a manifestation of Coulomb plasticity. At the local scale, it is actually frictional sliding of a contact surface with elastically deforming rocks on both sides. By comparison of the splay faults schematically illustrated in Fig. 1a and the potential failure planes of the critical wedge in Fig. 2b, we can see that some of the splay faults are oriented optimally for thrust failure if the frontal wedge is compressed during a megathrust earthquake.

Future Directions

The connection between wedge mechanics and great earthquakes and tsunamis at subduction zones is an emerging new field of study. It leads to challenges in both theoretical development and experimental design and thus excellent research opportunities. We need better constraints on how stresses along different downdip segments of the subduction fault evolve with time throughout an earthquake cycle and how the evolution impacts wedge and seafloor deformation. A number of outstanding questions are to be addressed: Can we constrain the updip limit of the seismogenic zone using wedge morphology? What is the timescale of stress relaxation along the updip segment of the megathrust after an earthquake? Does the seismogenic zone stay locked in the interseismic period? How does pore fluid pressure evolve in an earthquake cycle? How is the transfer of material from the incoming plate to the upper plate (accretion), from the subducted plate to the upper plate (underplating), or from the upper plate to the subducted plate (tectonic erosion) accomplished? What determines the dominant mode of material transfer? What does the spatial change in wedge morphology tells us about changes in the mechanical state of the wedge and the megathrust fault? These questions should be put in the proper context of larger-scale processes such as the viscoelastic relaxation of the mantle following a megathrust earthquake and the deformation of the subducting plate in earthquake cycles [47].

Sandbox experiments designed to study wedge mechanics and dynamic friction experiments designed to study fault mechanics are traditionally separate research activities addressing processes of vastly different timescales. The linkage between subduction earthquakes and submarine wedge evolution suggests the need to combine these experiments. Rapid motion used to simulate earthquakes has begun to be introduced into sandbox experiments [38].

The most promising type of field observation is continuous monitoring of deformation, such as strain and tilt, and fluid pressure using submarine borehole and seafloor observatories. Seafloor elevation change in response to the 2003 Tokachi-oki, northeast Japan, earthquake (M8.2), continuously recorded by two seafloor pressure sensors, clearly indicated coseismic strengthening behavior of the shallowest segment of the subduction fault [1]. Formation fluid pressure changes detected at subsea borehole observatories at the Nankai Trough subduction zone, southwest Japan, have been interpreted to indicate transient aseismic motion of a part of the locked seismogenic zone and/or dynamics of the incoming plate [13]. A number of very-low-

frequency earthquakes have been remotely detected within the Nankai Trough accretionary prism using land-based seismic networks [25], revealing the need for near-field observation using seafloor systems. Submarine monitoring in conjunction with land-based monitoring at subduction zones that are currently in different phases of earthquake cycles will allow us to understand the evolution of fault and wedge stresses during the interseismic period. In this regard, cabled seafloor monitoring networks including borehole observatories, being designed or implemented at different subduction zones [26,28] will surely yield valuable data in the near future.

Acknowledgments

We thank EE Davis, N Kukowski, SE Lallemand, and K Satake for reviewing the article and providing valuable comments. This work is Geological Survey of Canada contribution 20070221.

Bibliography

Primary Literature

1. Baba T, Hirata K, Hori T, Sakaguchi H (2006) Offshore geodetic data conducive to the estimation of the afterslip distribution following the 2003 Tokachi-oki earthquake. Earth Planet Sci Lett 241:281–292
2. Barr TD, Dahlen FA (1990) Constraints on friction and stress in the Taiwan fold-and-thrust belt from heat flow and geochronology. Geology 18:111–115
3. Breen NA, Orange DL (1992) The effects of fluid escape on accretionary wedges 1. Variable porosity and wedge convexity. J Geophys Res 97:9265–9275
4. Briggs RW, Sieh K, Meltzner AJ, Natawidjaja D, Galetzka J, Suwargadi B, Hsu Y-J, Simons M, Hananto N, Suprihanto I, Prayudi D, Avouc J-P, Prawirodirdjo L, Bock Y (2006) Deformation and slip along the Sunda megathrust in the great 2005 Nias-Simeulue earthquake. Science 311:1897–1901
5. Brown K, Kopf A, Underwood MB, Weinberger JL (2003) Compositional and fluid pressure controls on the state of stress on the Nankai subduction thrust: A weak plate boundary. Earth Planet Sci Lett 241:589–603
6. Byerlee JD (1978) Friction of rocks. Pure Appl Geophys 116:615–626
7. Byrne DE, Davis DM, Sykes LR (1988) Local and maximum size of thrust earthquakes and the mechanics of the shallow region of subduction zones. Tectonics 7:833–857
8. Chapple WM (1978) Mechanics of thin-skinned fold-and-thrust belts. Geol Soc Am Bull 89:1189–1198
9. Dahlen FA (1984) Noncohesive critical Coulomb wedges: An exact solution. J Geophys Res 89:10125–10133
10. Dahlen FA, Suppe J, Davis DM (1984) Mechanics of fold-and-thrust belts and accretionary wedges: Cohesive Coulomb theory. J Geophys Res 89:10087–10101
11. Davis DM (1990) Accretionary mechanics with properties that vary in space and time. In: Debout GE et al. (eds) Subduction: Top to Bottom. AGU Monograph 96, Washington, DC, pp 39–48
12. Davis DM, Suppe J, Dahlen FA (1983) Mechanics of fold-and-thrust belts and accretionary wedges. J Geophys Res 88:1153–1172
13. Davis EE, Becker K, Wang K, Obara K, Ito Y (2006) A discrete episode of seismic and aseismic deformation of the Nankai subduction zone accretionary prism and incoming Philippine Sea plate. Earth Planet Sci Lett 242:73–84
14. Dewhurst DN, Clennell MB, Brown KM, Westbrook GK (1996) Fabric and hydraulic conductivity of sheared clays. Géotechnique 46:761–768
15. Dieterich JH (1979) Modeling of rock friction: 1. Experimental results and constitutive equations. J Geophys Res 84:2161–2168
16. Elliot D (1976) The motion of thrust sheets. J Geophys Res 81:949–963
17. Enlow RL, Koons PO (1998) Critical wedges in three dimensions: Analytical expressions from Mohr–Coulomb constrained perturbation analysis. J Geophys Res 103:4897–4914
18. Fletcher RC (1989) Approximate analytical solutions for a cohesive fold-and-thrust wedge: Some results for lateral variation in wedge properties and for finite wedge angle. J Geophys Res 94:10347–10354
19. Fukao Y (1979) Tsunami earthquakes and subduction processes near deep-sea trenches. J Geophys Res 84:2303–2314
20. Fuller CW, Willett SD, Brandon MT (2006) Formation of forearc basins and their influence on subduction zone earthquakes. Geology 34:65–68
21. Geist EL, Dmowska R (1999) Local tsunamis and distributed slip at the source. Pure Appl Geophys 154:485–512
22. Hsu Y-J, Simons M, Avouac J-P, Galetzka J, Sieh K, Chlieh M, Natawidjaja D, Prawirodirdjo L, Bock Y (2006) Frictional afterslip following the 2005 Nias-Simeulue earthquake, Sumatra. Science 312:1921–1926
23. Hu Y, Wang K (2006) Bending-like behavior of wedge-shaped thin elastic fault blocks. J Geophys Res 111; doi:10.1029/2005JB003987
24. Hyndman RD, Wang K (1993) Thermal constraints on the zone of major thrust earthquake failure: The Cascadia subduction zone. J Geophys Res 98:2039–2060
25. Ito Y, Obara K (2006) Dynamic deformation of the accretionary prism excites very low frequency earthquakes. Geophys Res Lett 33; doi:10.1029/2005GL025270
26. Juniper K, Bornhold B, Barnes C (2006) NEPTUNE Canada community science experiments. Eos Tranctions, American Geophysical Union 87(52), Fall Meeting Supplement: Abstract OS34F-04
27. Kanamori H (1972) Mechanism of tsunami earthquakes. Phys Earth Planet Interior 6:246–259
28. Kaneda Y (2006) The advanced dense ocean floor observatory network system for mega-thrust earthquakes and tsunamis in the Nankai Trough – precise real-time observatory and simulating phenomena of earthquakes and tsunamis. Eos Tranctions, American Geophysical Union 87(52), Fall Meeting Supplement: Abstract OS34F-01
29. Kukowski N, von Hune R, Malavieille J, Lallemand SE (1994) Sediment accretion against a buttress beneath the Peruvian continental margin at 12° S as simulated with sandbox modeling. Geol Rundsch 83:822–831
30. Lallemand SE, Schnürle P, Malavieille J (1994) Coulomb theory

applied to accretionary and nonaccretionary wedges: Possible causes for tectonic erosion and/or frontal accretion. J Geophys Res 99:12033–12055
31. Lohrmann J, Kukowski N, Adam J, Oncken O (2003) The impact of analogue material properties on the geometry, kinematics, and dynamics of convergent sand wedges. J Struct Geol 25:1691–1711
32. Morone C (1998) Laboratory-derived friction laws and their application to seismic faulting. Annu Rev Earth Planet Sci 26:649–696
33. Moore DE, Lockner DA (2007) Friction of the smectite clay montmorillonite: A review and interpretation of data. In: Dixon T, Moore JC (eds) The Seismogenic Zone of Subduction Thrust Faults, Columbia University Press, New York
34. Moore JC, Saffer D (2001) Updip limit of the seismogenic zone beneath the accretionary prism of southwest Japan: An effect of diagenetic to low grade metamorphic processes and increasing effective stress. Geology 29:183–186
35. Mourgues R, Cobbold PR (2006) Thrust wedges and fluid overpressures: Sandbox models involving pore fluids. J Geophys Research 111; doi:10.1029/2004JB003441
36. Park JO, Tsuru T, Kodaira S, Cummins PR, Kaneda Y (2002) Splay fault branching along the Nankai subduction zone. Science 297:1157–1160
37. Rice J (2006) Heating and weakening of faults during earthquake slip. J Geophys Res 111; doi:10.1029/2005JB004006
38. Rosenau M, Melnick D, Brookhagen B, Echtler HP, Oncken O, Strecker MR (2006) About the relationship between forearc anatomy and megathrust earthquakes. Eos Tranctions, American Geophysical Union 87(52), Fall Meeting Supplement: Abstract T12C-04
39. Ruina A (1983) Slip instability and state variable friction laws. J Geophys Res 88:10359–10370
40. Sibson RH (2003) Thickness of the seismic slip zone. Bull Seismol Soc Am 93:1169–1178
41. Smit JHW, Brun JP, Sokoutis D (2003) Deformation of brittle-ductile thrust wedges in experiments and nature. J Geophys Res 108; doi:10.1029/2002JB002190
42. Takahashi M, Mizoguchi K, Kitamura K, Masuda K (2007) Effects of clay content on the frictional strength and fluid transport property of faults. J Geophys Res 112; doi:10.1029/2006JB004678
43. Tanioka Y, Satake K (1996) Tsunami generation by horizontal displacement of ocean bottom. Geophys Res Lett 23:861–864
44. von Huene R, Ranero CR (2003) Subduction erosion and basal friction along the sediment-starved convergent margin off Antofagasta, Chile. J Geophys Res 108; doi:10.1029/2001JB001569
45. Vrolijk P (1990) On the mechanical role of smectite in subduction zones. Geology 18:703–707
46. Wang CY (1980) Sediment subduction and frictional sliding in a subduction zone. Geology 8:530–533
47. Wang K (2007) Elastic and viscoelastic models of subduction earthquake cycles. In: Dixon T, Moore JC (Eds) The Seismogenic Zone of Subduction Thrust Faults, Columbia University Press, New York
48. Wang K, He J (1999) Mechanics of low-stress forearcs: Nankai and Cascadia. J Geophys Res 104:15191–15205
49. Wang K, He J (2007) Effects of Frictional Behaviour and Geometry of Subduction Fault on Coseismic Seafloor Deformation. Bull Seismol Soc Am 98:571–579
50. Wang K, Hu Y (2006) Accretionary prisms in subduction earthquake cycles: The theory of dynamic Coulomb wedge. J Geophys Res 111; doi:10.1029/2005JB004094
51. Wang K, He J, Hu Y (2006) A note on pore fluid pressure ratios in the Coulomb wedge theory. Geophys Res Lett 33; doi:10.1029/2006GL027233
52. Wang WH, Davis DM (1996) Sandbox model simulation of forearc evolution and noncritical wedges. J Geophys Res 101:11329–11339
53. Willett S, Beaumont C, Fullsack P (1993) Mechanical model for the tectonics of doubly vergent compressional orogens. Geology 21:371–374
54. Xiao HB, Dahlen FA, Suppe J (1991) Mechanics of extensional wedges. J Geophys Res 96:10301–10318
55. Zhao W L, Davis DM, Dahlen FA, Suppe J (1986) Origin of convex accretionary wedges: Evidence from Barbados. J Geophys Res 91:10246–10258

Books and Reviews

Dahlen FA (1990) Critical taper model of fold-and-thrust belts and accretionary wedges. Annu Rev Earth Planet Sci 18:55–99
Dixon T, Moore JC (eds) The Seismogenic Zone of Subduction Thrust Faults, Columbia University Press, New York
Scholz CH (2003) The Mechanics of Earthquakes and Faulting, 2nd edn. Cambridge University Press, Cambridge, 471 p

World Wide Web, Graph Structure

Lada A. Adamic
School of Information and the Center for the Study of Complex Systems, University of Michigan, Ann Arbor, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Graph Structure
Algorithms
Subsets of the Web Graph
Future Directions
Bibliography

Glossary

Graph A set of nodes (vertices) connected by links (edges, arcs). In the Web graph, the nodes are webpages, and the edges are the hyperlinks between them.

Indegree The number incoming edges to a node; in the case of the Web, it is the number of webpages pointing to a page.

Outdegree The number of outgoing edges; in the case of the Web, it is the number of webpages a webpage points to.

Strongly connected component A set of webpages such that any page can be reached from any other page by following hyperlinks.

Weakly connected component A set of webpages such that any page can be reached from any other page by treating the hyperlinks as undirected.

URL A unique resource locator that corresponds to an online information source.

Hypertext transfer protocol (HTTP) The communications protocol that allows web clients to communicate with web servers.

Web server A program that responds to requests to for web pages.

Webpage An information resource, identified by a URL, that is usually but not necessarily in the HTML (Hypertext Markup Language) format. A webpage may be *static*, meaning that it is stored on the server as a document, or *dynamic*, meaning that it is generated dynamically at the point that it is requested by the browser, using scripts and/or back-end databases.

Domain A name that identifies one or more IP addresses, e. g. umich.edu.

Top level domain The suffix of the domain name, sometimes corresponding to the purpose of the website or the country of origin for the website, e. g. "edu", "com", "gov", "uk", "cn", etc.

Website A collection of webpages that is hosted on one or more web servers and that share a common root URL, e. g. "www.springer.com".

Randomized network A network that preserves the degrees of each node relative to the original, but the edges themselves are rewired.

Definition of the Subject

In the period of a few short years, the World Wide Web has become the primary information source. Billions of webpages, tied together by hyperlinks, constitute the world's largest publicly accessible store of data, and are accessed by hundreds of millions of people worldwide. Understanding how the webpages are connected to form the Web graph is important for building comprehensive and accurate information retrieval systems, as well as characterizing real-world phenomena that are reflected in its content and link structure.

Introduction

The Web is becoming the single most indispensable source of information. A massive graph of billions of pages, the Web is continuously growing and evolving. People use the Web to do everything from reading the news, socializing, making purchases, paying their bills, watching videos, playing games, and finding romance. Given how pervasive the Web is in our daily lives, it is hard to believe that it is a relatively recent invention.

The World Wide Web was invented by Tim Berners Lee in 1990 at CERN in Geneva, Switzerland. Berners' idea was to combine hypertext – the linking of documents via terms – with the internet, allowing users to navigate between documents that are distributed and accessible across the globe. Starting with the first site created at CERN in 1991, the Web grew to 130 sites by 1993 [1]. By 1997 there were around a million sites [1], and by 2006, they numbered 100 million, about half of them active [2].

Such massive growth is only possible through the distributed contribution by many individuals. Thus online information is not contributed by a few vetted experts, nor is access limited to a few individuals. Rather, almost anyone can contribute to online content, especially with the proliferation of easy-to-use web authoring tools. One might expect then that there is a great deal of randomness in how pages are linked to one another, and how users navigate those links. Contrary to this, there are a number of strong regularities both in the structure of the Web and the pattern of access by its users. These regularities can be discovered when one studies the graph structure of the Web.

Graph Structure

The Web is a graph. The webpages are the nodes or vertices, and the hyperlinks are the directed links or edges between them. Characterizing the Web graph has lead to improved crawling, sampling, and ranking algorithms. It has also lead to novel insights about the underlying human dynamics that are reflected in the link patterns of the Web.

Empirical Measurements

Given the enormity of the Web graph, only a handful of projects, WebBase [3] and WebGraph [4,5] among them, have sought to gather comprehensive crawls gathering hundreds of millions of the web's hundreds of billions of pages and have made them available to the research community. Early characterizations of the Web graph were made from Web crawls made by commercial search engines [6,7], and major search engine companies continue to host a variety of research on the Web graph [8,9,10,11].

Degree Distributions One of the most striking features of the Web graph that was discovered early on is the power

law distribution of incoming hyperlinks. While billions of webpages, comprising a substantial portion of the web, receive few or no hyperlinks, a few very popular pages have attracted millions. More precisely, the distribution of incoming links is a power-law (also known as a Zipf's law or Pareto distribution) [12,13,14]

$$p(k) \sim k^{-\alpha} \tag{1}$$

The indegree distribution has shown remarkable consistency across many different studies, having an exponent α of 2.1 [7,15]. The outdegree distribution has a steeper power law exponent measured at $\alpha = 2.5$ [15] and $\alpha = 2.7$ [7]. The steeper power law exponent corresponds to a smaller variance in the number of hyperlinks stemming from a web page as opposed to pointing to it; while a very useful or informative webpage such as a major search engine or news site can attract millions of hyperlinks, it is rather impractical, due to limitations in content length, for a single webpage to contain a very large number of hyperlinks.

Power law degree distributions have been shown to not be preserved for some subsamples of the Web graph, however. Pennock et al. [16] found that for sets of company, university, newspaper, and scientist homepages, distributions of inlinks deviated for power-laws, especially in the number of pages of low degree. Intuitively, one might expect a university homepage, or a newspaper homepage to attract at least a few links, producing a peak in the distribution away from the minimum. This is in contrast to power law distributions, which are scale-free, appearing the same on all scales, e. g. 10 to 100 links or 100 to 1000 links, and having no "typical" degree.

Much research has gone into deriving generative models that can produce the observed degree distributions. We will return to this topic in Subsect. "Generative Models".

Degree Correlations Degree correlation, also termed assortativity, characterizes to what extent high degree nodes link to other high degree nodes. In the case of the Web graph, a question may be whether pages that receive many links, link to other pages with high indegree. In an undirected version of the nd.edu webpage network, the webpages were found to be mildly disassortative with a correlation of -0.065 in the combined undirected degree of vertices [17]. This webpage network, crawled in 1999, consists of 325,729 documents and 1,469,680 hyperlinks. We can resolve this correlation further by considering pairwise the in and outdegrees of both the page that is giving the link, and the one that is being linked to. Using the nd.edu domain data set, we find that there is a slight negative correlation between both the indegree and outdegree of the webpage containing the hyperlink and the indegree of the page it is pointing to. This indicates that there is a slight tendency of link-rich pages to link to less link-rich pages and less popular pages, and link-poor pages to link to link-rich and popular pages. Interestingly, in this domain, there was a strong *positive* correlation between the indegree of the linking page and the outdegree of the linked page. This may be an indication that the most cited pages themselves link to link-rich pages and so serve a funneling function. The above is summarized in Table 1. The correlation between the pages' own indegree and outdegree was positive ($\rho = 0.244$), indicating that link rich pages themselves tended to receive more links.

World Wide Web, Graph Structure, Table 1
Assortativity of webpages in the nd.edu domain

		to	
		indegree	outdegree
from	indegree	-0.023	0.256
	outdegree	-0.062	-0.014

Connected Components and Bow-Tie Structure The bow-tie structure describes the connectivity of the Web graph as a whole [7]. The middle knot of the bow-tie is the largest strongly connected component (LSCC). The component is called strongly connected because within it, any page can be reached from any other by following directed links. The OUT component consists of those pages that can be reached from the LSCC, but do not have paths leading into the LSCC. The IN component consist of those pages that have paths leading into the LSCC, but are not reachable from the LSCC. The remainder of the graph is composed of tubes, tendrils, and islands. Tubes connect the IN component to the OUT component, bypassing the LSCC. Tendrils are those pages that can either be reached from the IN component, or lead to the OUT component, but are not in tubes or the LSCC itself. Islands are connected components of vertices that are not connected via any links to the other components.

Each of these four regions (LSCC, IN, OUT, and the rest) of the Web's bow-tie accounts for roughly a fourth of the entire graph. This implies that if two pages A and B are selected at random, there is only roughly a 1 in 4 chance that there is a directed chain of hyperlinks leading from A to B. The reasoning is the following: most often, if there is a path from A to B, A is either in the IN or LSCC components, while B is in the LSCC or OUT components. This is true in approximately $1/2 \times 1/2 = 1/4$ of the cases. By comparison, about 90% of the webpages are

in the largest connected component if the links are treated as undirected.

The bow-tie structure was also observed to exist for subsets of webpages corresponding to a given topic, so that the web appears self-similar on several scales [18]. This structure has important implications for a crawler attempting to traverse the web in its entirety. By starting at a single webpage, the crawler will likely not be able to reach significant portions of the Web graph by following directed hyperlinks. Instead, it may adopt a strategy of discovering web pages to start from by scanning IP addresses to see if they are hosting a web server.

Shortest Paths and Diameter In modeling user behavior and web crawler performance, it is useful to know not just whether B is reachable via a directed path of hyperlinks from A, but how many hyperlinks must be traversed. Albert et al. [19] found for the above mentioned nd.edu domain crawl, that the average shortest path over all pairs of vertices was 11.2. This closely matched the value of 11.6 predicted by their preferential attachment model, which will be described in Subsect. "Preferential Attachment and the BA Model". The model produces an average shortest path $\langle d \rangle$ of

$$\langle d \rangle = 0.35 + 2.06 \log(N) \tag{2}$$

The model likewise provides a close prediction of 17.5 for a graph of the size of the sample of 203 million nodes described by Broder et al. [7]. The measured shortest path between the 24% of the pairs for this sample that were *reachable* was 16. As discussed in Subsect. "Connected Components and Bow–Tie Structure", the other three quarters of the pairs of web pages have no directed path connecting them, and hence their distance is infinite. On the other hand, treating the graph as undirected produces an average shortest path of just 6 hops for the 90% of the webpages that are connected.

Another quantity used to describe the width of the Web is the diameter, the maximal shortest path between any two connected vertices. Broder et al. [7] measured a diameter of at least 28 in the central core and over 500 for the graph as a whole, indicating that while most reachable pairs can be connected in just a small number of hops, others are far removed.

Reciprocity, Clustering and Motifs Reciprocity is simply a measure of how often when one webpage links to another, that page links back. The measure should take into account the expected number of reciprocated links based simply on the link density $\bar{a} = m/N/(N-1)$, where N is the number of nodes and m is the number of edges. An unbiased form of the reciprocity measure is given by

$$\rho = \frac{m_{\mathrm{bd}}/m - \bar{a}}{1 - \bar{a}} \tag{3}$$

where m_{bd} is the number of bidirectional links [20]. For the crawl of pages in the nd.edu domain, $\rho = 0.52$ [19].

Clustering or transitivity quantifies how often webpages that are linked to a common page are linked to each other as well. The measure, usually applied to an undirected version of the network, is given by the following equation.

$$C = \frac{6 \times \text{ number of triangles in the network}}{\text{number of connected triples of vertices}} \tag{4}$$

A 1998 Alexa crawl of 260,000 websites yielded a clustering coefficient of 0.014 at the site level (0.11 using the Watts Strogatz clustering coefficient [35]), significantly higher than a randomized version of the network [6].

The clustering coefficient captures the degree of transitivity in the graph, but does not reveal the configuration of those triads. Triad motifs [21] shed further insight into the local structure of the Web graph. There are 13 triad motifs, and together their abundance compared to randomized versions of the web graph constitute a *triad significance profile* [22]. The significance profile SP is a normalized vector

$$\mathrm{SP}_i = \frac{Z_i}{\left(\sum_i Z_i^2\right)^{1/2}} \tag{5}$$

of z scores measuring the statistical significance of the number n_i^{web} of observed motifs relative to the randomized counterparts of the network

$$Z_i = \frac{n_i^{\mathrm{web}} - \langle N_i^{\mathrm{rand}} \rangle}{\sigma_i^{\mathrm{rand}}} \tag{6}$$

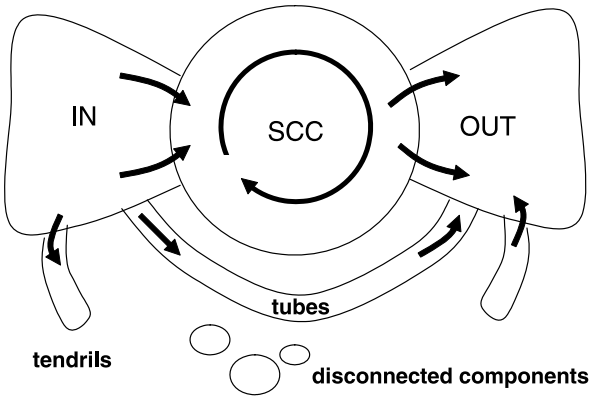

World Wide Web, Graph Structure, Figure 1
A schematic of the bow-tie structure of the web

where $\langle N_i^{\text{rand}} \rangle$ and σ_i^{rand} are the mean and standard deviation of the frequency of occurrence of motif i in randomized versions of the network.

The motif profile of Web graphs is like that of social networks, with closed triads and reciprocated links more frequent than motifs without closure and with unreciprocated links. The most underexpressed profile is that of a webpage that has reciprocated links with two other webpages, but those two webpages do not link to one another. The most highly expressed motif is that of the fully reciprocated closed triad. The probability of such a triad occurring in a random graph is very small, but it occurs rather frequently in the Web graph where clusters of related pages link to one another reciprocally.

Time Evolution

The web is growing exponentially, with pages being continuously added, modified, and deleted. Addition and deletion clearly affect which nodes are present in the graph, while modifications in content [23,24] possibly result in link addition and deletion. Cho and Garcia–Molina [25] found that 30% of a sample of 720,000 pages disappeared in an interval of a month. Douglis et al. [26] studied 950,000 web page requests from a server, and found that for pages that were accessed more than once, 60% had changed one or more of their hyperlinks. Making the task of discovering new pages easier is the fact that 85–95% of new pages appear on existing sites [27].

Web Surfing

An important aspect of the Web graph is how it is navigated. Huberman et al. [28] discovered the *Law of Web Surfing*, that describes the distribution in the number of links a user will access on a particular website. This distribution was found to hold for web surfing patterns of AOL users across a million web sites, users accessing the Xerox website, and students, faculty, and staff at Georgia Tech's College of Computing. The underlying process was argued to be the following. The value V_L of the Lth page in a click stream where L links were followed is assumed to be related to the previous page, but also to deviate from it in a random way.

$$V_L = V_{L-1} + \epsilon_L \tag{7}$$

where ϵ_L is a normally distributed random variable. The individual will continue to surf until the expected discounted value of information to be found on future pages is less than the cost (in time) of continuing. This maps to an option pricing model in finance, and the distribution of the number of links a user follows before the expected utility of continuing falls below a threshold is given by

$$P(L) = \sqrt{\frac{\lambda}{2\pi L^3}} \exp\left[\frac{-\lambda(L-\mu)^2}{2\mu^2 L}\right] \tag{8}$$

with a mean number of clicks $\mathrm{E}(L) = \mu$ and variance $\mathrm{Var}(L) = \mu^3\lambda$, λ being a scale parameter. This heavy tailed distribution accurately describes users' surfing behavior on the Web graph: many follow just a link or two, while a few explore dozens or even hundreds of links in a single session.

Crawling

The web is only partly navigated through following hyperlinks. Users jump from one part of the Web graph to another by querying search engines and selecting one or more pages that match their query. In order to deliver web pages matching particular user queries, search engines must first crawl the Web graph. Web crawlers have two tasks: revisiting previously crawled pages in order to refresh them and discovering new pages by following new links. The crawler needs to prioritize both the order in which it will crawl the new content and how often it will recrawl pages it has previously crawled, in part to discover links to new content [27].

Search Engine Coverage It is not possible for search engines to crawl the entire Web. There are infinite websites that will return valid and unique content for an infinite number of URLs. Still, it is possible to explore the overlap in search engine coverage of the content that should be covered by the search engines. In 1998 Lawrence and Giles set out to estimate the size of the "indexable" web by issuing queries to multiple search engines [29]. By measuring the overlap in the search results and assuming that each search engine samples the web independently, they estimated the total size of the Web at that point to be at least 320 million pages, with no single search engine achieving more than a third of the coverage.

Generative Models

Generative models help describe the underlying processes that are shaping the Web, and therefore can be used to make predictions about the future evolution of its properties.

Preferential Attachment and the BA Model The World Wide Web inspired the Barabasi–Albert preferential attachment model of network growth [15,30], that has since been applied to networks in many other domains, including biological, social, and technological. As discussed in

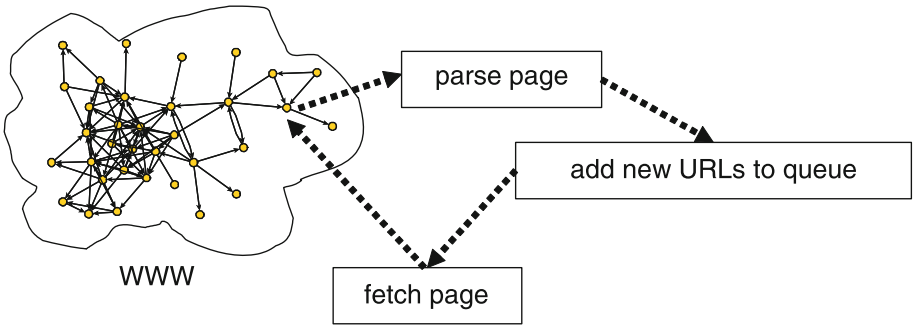

World Wide Web, Graph Structure, Figure 2
A basic schematic of a crawler traversing webpages. As new pages are discovered, they are parsed, and their links added to the queue of pages to be visited if they had not been seen before

Subsect. "Degree Distributions" Albert et al. observed that the distribution of incoming hyperlinks for webpages was highly skewed. They devised a very simple model of preferential attachment, which had previously been explored in other networked and non-networked contexts [31,32,33], wherein a new page is added at each timestep t, and attaches to m existing webpages, each with probability $\Pi(k_i)$ proportional to that page's current degree k_i.

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j} \tag{9}$$

An intuitive argument for why this is so is that the more popular a web resource is, the more likely an author of a new webpage is to learn about it in the course of browsing and searching the Web, and the more likely she is to cite it.

The links are treated as undirected, and the model yields a scale-free indegree distribution with a power law exponent $\alpha = 3$, independent of m. While reproducing some characteristics of the Web graph, such as a scale free distribution and short average path lengths, the model does not capture others. The power law exponent of 3 is much higher than the consistently observed $\alpha = 2.1$ for the Web. Its clustering coefficient depends on the size of the graph as $N^{-0.75}$, and is much smaller than that of the Web graph. It also predicts that the degree of each node depends on its time of introduction i as follows

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\beta}, \quad \beta = \frac{1}{2}. \tag{10}$$

However, there is only very weak correlation between the age of a webpage or website on the web and its indegree [34].

Extensions of the Barabasi–Albert model were proposed to overcome its limitations. Rewiring of existing edges with the endpoint being chosen preferentially produces power law disributions with cut-offs as well as exponents in the range $2 \leq \alpha < \infty$. Other models have introduced aging factors, where webpages lose attractiveness over time [36], which produced cutoffs in the power law distributions. In the fitness model, webpages have a different inherent attractiveness η_i [37], with preferential attachment combining this attractiveness with the degree k_i so that $\Pi_i \sim \eta_i * k_i$. Interestingly, it was found that nonlinear preferrential attachment [38] where

$$\Pi(k_i) \sim {k_i}^{\nu} \tag{11}$$

does not always produce power law degree distributions when $\nu \neq 1$ [39].

Copying Models The basic preferential attachment models do not explicitly explain why a new webpage would link to existing webpages in proportion to their indegree. Such models also do not capture features such as clustering, since the only factor influencing the likelihood of being linked to is the indegree of the page. In copying models, both clustering and preferential attachment are produced [40,41,42,43] as a result of links being copied, without content being explicitly taken into account. New pages are added over time, and each page selects an existing page, termed a *prototype* or *ambassador* page, and copies several of the links on that page. The process leads to preferential attachment, since the probability that any given page is linked to is proportional to the probability that it is the neighbor of the ambassador node, which is in turn proportional to its indegree. In the case where the new page links to the ambassador page in addition to copying some of its links, the process also produces clustering; each copied link results in a closed triad: the new page c links to the ambassador page a, a links to b, and c copies that link and also links to b. In the case where the new page does not link to the ambassador page, the process produces bi-cliques, sets of pages that link to a second set of pages.

Kumar et al. [42] constructed a stochastic copying model where new pages are created linearly or exponentially in time. Each page has a fixed outdegree. With probability p_r, the ith link will point to a webpage at random, otherwise it will copy the ith link of the ambassador node. In the case of linear addition of pages in time, the probability $p(k, t)$ that a node has degree k at time t is given by a power law:

$$p(k, t) \sim tk^{-(2-p_r)/(1-p_r)} \tag{12}$$

where the power-law exponent α of the degree distribution is bounded from below by 2 in the case of pure copying, and grows steeper as the proportion p_r of random links is increased. The model also produces a large number of bipartite cliques, commensurate with empirically observed proportions.

Related to the copying model is the forest fire model, where the copying continues recursively to all nodes that are encountered by following links [43]. The process is analogous to a forest fire which spreads from tree to tree. Two probabilities are assigned. With forward burning probability p, the outgoing links from page are copied. With probability q incoming links are copied. This last step models something akin to awebpage author searching for other pages that link to the current page and deciding to link to them as well.

1. As in the copying model, a new node chooses an ambassador node uniformly at random from the set of all nodes
2. Two geometrically distributed random numbers, x and y are generated with means $p/(1-p)$ and $q/(1-q)$ respectively. x out-links and y in-links are copied from the current node (if x or y are greater than the total number, all the links are copied).
3. Step 2 is applied recursively to each of the nodes the new node now has links to. Cycles are disallowed.

Although the forest fire model is so far analytically intractable, it has several properties that match the web. Heavy-tailed indegree distributions occur because in copying links from other nodes, new nodes are attaching preferentially to existing nodes in proportion to their indegree. Heavy-tailed outdegree distributions are produced by the cascading process of recursive link copying. Many cascades will terminate already with the first node, but some will continue for many steps. Community structure and clustering are produced, as in the basic stochastic copying model, because the link copying leads to triadic closure, as a new webpage links to a page and some of its neighbors, and those neighbors' neighbors, etc.

Content-Based Models The above generative models have not explicitly modeled the textual content of the webpages. The intuition behind content-based or similarity-driven models is that webpages will not only link to popular pages, but also to topically related webpages (or both). Menczer [44] used a cosine similarity measure

$$s(p_1, p_2) = \frac{\sum_k w_{kp_1} w_{kp_2}}{\sqrt{\left(\sum_k w_{kp_1}^2\right)\left(\sum_k w_{kp_2}^2\right)}} \tag{13}$$

where w_{kp} is some weight function, typically, a TFIDF (term frequency, inverse document frequency) weight for the term k in page p. In 110,000 webpages sampled from the Open Directory, Menczer found that lexical similarity closely correlates with webpages being clustered (either linking directly or overlapping in their neighbors). He proposed a content based generative network model to take into account that once two pages exceed a threshold overlap ($s > s^*$) in their content, their probability of being linked levels off. Below that threshold, the probability decreases as a power-law, so that while neighboring pages tend to be lexically similar, there is a long tail of lexically distant pages that are neighbors in the Web graph. As in the Barabasi–Albert model, at each time step t a new page is added and links to m existing pages. But rather than linking to any page in proportion to their degree, it will only do so below a threshold lexical similarity. Above that threshold the probability that a new page appearing at time t will link to an existing page i with degree $k(i)$ will be proportional to the lexical similarity:

$$Pr(p, t) = \begin{cases} \frac{k(i)}{mt} & \text{if } s(p_i, p_t) > s^* \\ c_1 \left([s(p_i, p_t)]^{-1} - 1\right)^{-\alpha} & \text{otherwise .} \end{cases} \tag{14}$$

The model produces a degree distribution that deviates from a pure powerlaw, but closely matches the sample webpages from the Open Directory.

Algorithms

Ranking

Most search engines utilize the structure of the Web graph to rank their search results. Some structural node properties, such as indegree, are purely local and utilize only the nearest neighbors of the webpage. Others, such as PageRank, utilize the entire graph, and yet converge within a short number of iterations.

Indegree The simplest link based ranking of webpages is based on their indegree,

$$s(k) \sim k \tag{15}$$

where each inlink serves as a popularity vote. This measure is susceptible to spamming. One can create webpages and websites with relative ease whose primary purpose is to boost the ranking of a particular target page.

PageRank

The PageRank algorithm was developed by Larry Page, one of the co-founders of the Google search engine [45]. PageRank is less susceptible to spamming because it takes into account not just how many pages link to a particular target page, but also how many pages link to those pages etc. It does so by simulating a random walk on the Web graph. This PageRank score represents the proportion of time a random web surfer would spend at each webpage, traversing the graph by following random links from a succession of web pages. Roughly speaking, PageRank corresponds to the probability distribution given by the principal eigenvector of the normalized webgraph adjacency matrix. To prevent the random walk from being trapped in a region of the graph (remember that only a fraction of the pages is in the LSCC), a damping factor d (also termed a teleportation probability) is introduced. With probability d the surfer follows a random link from the page she finds herself on. With probability $(1-d)$ the surfer jumps to a random node in the Web graph. Let A be the adjacency matrix for a Web graph with n pages. Then the matrix B corresponding to the Markov chain is given by

$$b_{ij} = d\,\frac{a_{ij}}{\sum_i a_{ij}} + \frac{1-d}{n}\,. \tag{16}$$

Rather than solving for the eigenvector exactly, one can obtain a close approximation by applying a power iteration technique, iteratively multiplying an initial vector x by B

$$x = Bx\,. \tag{17}$$

The convergence of the iterations depends on the initial vector and the damping factor, but is tractable even for very large web crawls. The damping factor is typically set to $d = 0.85$. For most applications, PageRank values are precomputed for all the webpages at once, and are combined with a textual match score to a user's query and other factors in determining the final ranking of webpages on any given search engine result page. Although PageRank is designed to take into account more than direct links, studies have found it to be highly correlated with indegree when applied to the Web graph [46].

Personalization of PageRank can be achieved by biasing the random jump toward pages a particular user has visited already, rather than all webpages uniformly at random [47]. Since pages that link to one another are likely to be topically similar, the personalized PageRank algorithm will boost the scores of pages in the areas of the web the user has previously explored, and hence bias the search results toward pages of topical interest to the user.

Application to Unbiased Sampling During a crawl, the probability of encountering a link to a page is roughly proportional to the PageRank of that page, since the PageRank represents the random walk probability of finding the page. Therefore crawling will produce a biased sample where popular pages are more likely to be included. In order to create an unbiased sample, one can simply select pages in inverse proportion to their estimated PageRank [48].

HITS

The HITS (Hypertext Inducted Topic Selection) algorithm was developed by Jon Kleinberg, and starts with a smaller subset of query results, rather than with the entire web graph [49]. Instead of than assigning a webpage a single score, as PageRank does, HITS assigns two scores, a hub score y and authority score x. A good authority is a page with authoritative information, and a good hub points to good authorities. Hub may constitute good starting points for a user to explore an information space, while authorities are likely to contain definitive information.

In the first step of the HITS algorithm, an initial root set of webpages (e. g. 200 pages) is retrieved based on a textual match. Some of these pages, containing text relevant, though not necessarily authoritative, on the subject of the query, can be expected to link to prominent authorities on the subject, and some will be linked to by hubs. This root set is therefore expanded to a base set that includes (up to a cutoff of e. g. 1000–3000) pages linked to by the root set and linking to the root set. At this point, links between pages located on the same website are removed, as they are likely to serve a navigational purpose and don't necessarily confer authority.

Much like the PageRank scores, hub scores x and authority scores y can be computed iteratively, after being initialized to uniform constants. For a page j, the value x_j is updated to be the sum of y_i over all the pages i that link to j.

$$x_j = \sum_{i \text{ such that } i \to j} y_i \tag{18}$$

Similarly, the hub score y_i of a page i is updated according to the authority scores x_j of the pages it points to.

$$y_i = \sum_{j \text{ such that } i \to j} x_j \tag{19}$$

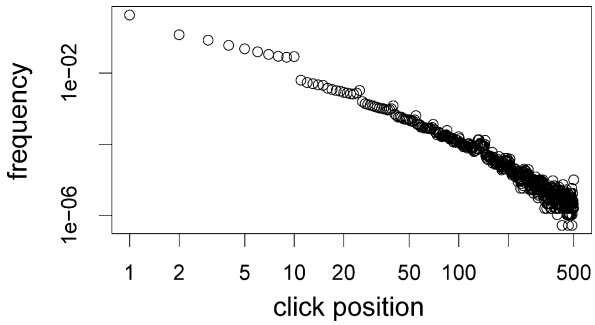

World Wide Web, Graph Structure, Figure 3
Frequency of clicks on different search result positions in 20 million queries issued by 600,000 users of the AOL search engine, from March to May 2006 [50]

The above can be also be computed using a power iteration technique, with the authority score x corresponding to the principal eigenvector of $A^{\mathrm{T}}A$, while y corresponds to the principal eigenvector of $A\,A^{\mathrm{T}}$, A being the adjacency matrix of the base set:

$$x \leftarrow A^{\mathrm{T}} y \leftarrow A^{\mathrm{T}} A\, x$$
$$y \leftarrow A\, x \leftarrow A A^{\mathrm{T}} y\,.$$

Unlike PageRank, where the scores for all pages are precomputed and subsequently used for all queries, HITS is computed at query-time.

Search Engine Bias

Since most search engines utilize a link-based ranking algorithm to order their search results, there is the possibility that they further bias web traffic toward already popular sites on a given topic. As shown in Fig. 3, 90% of the search results users click on are on the first page, with 43% being the very top search result. In part the higher rate of click-throughs for highly ranked results is due to users' tendency to scan content in order, but in part it is also due to search engines' ability to rank relevant content successfully. This may inhibit newer, but high quality pages, or pages presenting a different perspective, from receiving attention from search engine users.

While search engines drive traffic to the most heavily linked sites, the search engine users spread themselves out much more widely by searching for a wide variety of queries. In a 2005 study, Fortunato et al. [51] found that a majority of queries to Google return less than 30,000 hits (representing less than one millionth of the entire set of the web crawled by Google at that time). Hence, while some web pages have tens of thousands and even millions of in-links, they do not grab all of the search engine users' attention.

Finding Communities in the Web Graph

Webpages on similar topics naturally form densely connected regions of the Web graph, termed communities. Various algorithms have been developed to discover web communities based on the link structure alone, without considering the textual content or starting from a given seed set. Kleinberg et al. [52] *trawled* for emerging cyber-communities by identifying bipartite cores: sets of i pages that all cite j authoritative pages. Of thousands of such cores identified, the vast majority corresponded to groups of related pages. This approach has been extended to finding larger, but incomplete bipartite cores [53]. Communities can also be identified by performing the HITS expansion step from a root set that is comprised of a core [54].

Flake et al. [55] used maximum-flow minimum-cut algorithms to discover communities, the definition of a community being such that any page within the community has more links within the community than outside of it. The Web graph is augmented by creating an artificial source s with infinite capacity edges routed to all seed vertices S for which one would like to find a community. All preexisting edges are made bidirectional and rescaled to a heuristically chosen constant k. A virtual sink is connected to all vertices except the seed vertices and virtual source. A maximum flow procedure is used to produce a residual-flow graph (the graph of edges that have excess capacity). All vertices that are accessible from s through the residual graph satisfy the definition of community.

Many other network based clustering algorithms exist, some of which have been applied to subsets of the Web. Some, based on the concept of modularity, divide the network into communities so as to maximize the within-community edges compared to what one would expect for a random assignment of pages to communities. The modularity can be expressed as:

$$Q = \frac{1}{4m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \tag{20}$$

where A is the adjacency matrix, m is the total number of edges in the network, k_i is the degree of vertex i and c_i is the community i is assigned to. The modularity can be maximized through a greedy hierarchical algorithm that scales up to millions of nodes [56], a simulated annealing approach [57], or by directly computing the eigenvectors of the modularity matrix B

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m}\,. \tag{21}$$

Beyond identifying topically related sets of pages, community finding algorithms may be applied to identify spam

hosts and pages, since pages within so called link farms tend to link to one another much more often than to legitimate sites [9,58].

Subsets of the Web Graph

The Web contains various interesting subgraphs, reflecting different topics, and social and collaborative activity. Here we discuss the graph properties of just a few of those subsets.

Query Connection Graphs

Subraphs corresponding to search engine results for given queries can be informative both regarding the quality of the search results and the likelihood that a user will reformulate the query, making it more specific or more general [59]. The process of constructing query connection graphs is shown in Fig. 4. A *projection graph* is constructed, consisting of the search results and the hyperlinks between them. A *connection graph* is constructed by adding the minimum number of other webpages necessary to reconnect the projection graph. One can also construct the two graphs on the domain level, where a domain encompasses all of e. g. "springer.com" or "umich.edu", and two domains share an edge if a page in one domain links to a page in the other. As one might expect, domain query projection graphs are much more densely linked than projection graphs on individual webpages. A complete domain graph of the web from February 2006 contained 39 million domain names and 720 million directed edges, with an average shortest path of 4 and the largest weakly connected component containing 99% of the nodes. Whether at the page or domain level, when compared with human ratings of the quality of a set of pages for a given query, the sets that had more tightly connected projection graphs (and correspondingly smaller connection graphs) were good predictors of quality of the pages.

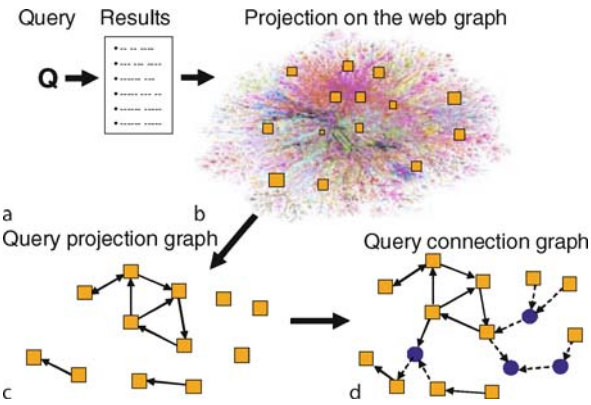

World Wide Web, Graph Structure, Figure 4
The process of constructing a query connection graph. A search engine is used to retrieve web pages relevant to the query. The pages are projected onto the Web graph. The projection subgraph contains only the search results and the hyperlinks between them. The connection subgraph also includes the nodes (webpages) that are minimally needed to reconnect the pages

The success of a user's search is only in part due to the accuracy of the search engine ranking algorithm. Another component is the skill of the user in formulating a query. By looking at the query subgraphs of issued queries, one can predict whether a user will reformulate a query. Queries with result sets that are more tightly knit and contain a few central high degree nodes are less likely to be reformulated. On the other hand, if a query projection graph lacks central, high degree nodes, the search engine user is more likely to reformulate the query. However, if the largest connected component constitutes e. g. 10 of the top 20 search results, the query may be too specialized, and the user is more likely to reformulate the query to be more general.

Weblogs

A rapidly growing part of the Web are weblogs (blogs), webpages containing time-stamped posts. Blogs are interconnected through direct links from one post to another, which may occur when bloggers discuss the same topic. Blogs also frequently include a sidebar of links to other blogs and websites. Millions are used simply as personal journals, but many specialize in particular topics, such as politics or technology. Some blogs today attract as much attention as mainstream media sites [60]. Kumar et al. [61] studied the profiles of 1.3 million blogs on livejournal.com, one of the most popular blogging sites. Livejournal gives its bloggers the opportunity to specify other bloggers who are their friends, basic demographic information, and to join groups based on their interests. The network is highly clustered, with a clustering coefficient of $C = 0.2$ (meaning that 20% of the time friends of the same blogger are friends themselves). This high clustering is consistent with the observation that 70% of the connections correspond to one of three factors: shared interest, same 5 year age group, and geographic location.

The geographic location information of livejournal users allowed for a very large scale study of the small world phenomenon [62]. It had been known for decades, ever since the 1960s, when Stanley Milgram conducted his famous small world experiment [63], that people are able to form short chains of acquaintances to reach a target individual using only information about their immediate network neighbors. Subsequently, Kleinberg [64] proved that

for certain spatial network topologies, individuals would be able to use a simple greedy algorithm, choosing an acquaintance closest to the target as the next person in the chain, to connect to targets in a short number of steps. This holds true for LiveJournal, where the probability of two LiveJournal users being connected is inversely proportional to the number of people who lived between them, and this spatial distribution of connections allows for successful formation of small world chains [62].

Blogs are continuously being updated and this temporal evolution is one of their most interesting aspects. Citations patterns between blogs change rapidly, even on the scale of a few months [76]. Kumar et al. [65] found that as blogging activity grew in popularity from 2002 to 2003, the addition of links became bursty – indicating that activity was centered around communities and events relevant to them. The degree distribution appeared to be power-law and asymptoting to a power-law exponent of 2.1. This inequality in link distribution was noted for the subset of political blogs [66,67] in the United States. The subgraph of political blogs also showed a strong tendency of blogs to link predominantly to blogs of similar political leanings [68,69].

The time resolved nature of blog entries enable researchers to track individual *memes*, topics that are contagiously spreading as information cascades from blog to blog. Several models [70,71] were developed to identify topics and their spread through the blog network. In a general cascade model [71] of the directed and weighted transmission graph of blogs, the weight of each edge corresponds to the probability that a blog both reads and copies an item from another blog, and this weight is learned through a machine learning algorithm from observed mentions by the two blogs of previous topics. By tacking 7,000 topics over a set of blogs, Kempe et al. [71] found that the amount of information flowing through blogs corresponded to the ranking of those blogs by blog search engines. Similarly, ordering the edges by the transmission probability on each edge closely corresponded to blogs that were directly linked through their blogroll links (links to other blogs that occur on the sidebar of the page). Adar et al. [72] used the presence of direct links between blogs, and overlap in text and links previously cited, to predict the path of information flow among a set of blogs mentioning the same URL. Leksovec et al. [73] developed an algorithm for near optimal selection of a set of blogs such that information cascades are detected quickly. Because of the skewed nature of participation and attention in the blogosphere, only a small fraction of blogs need to be monitored in order to capture a large portion of the activity.

Although blog content is highly distributed, and to a certain extent collaborative, with blogs filtering and generating news at rates often outpacing the mainstream media, the content is highly redundant, often inaccurate, and poorly interlinked. In contrast, Wikipedia is an example of collaboratively generated content that is accurate, non-redundant, and carefully structured.

Wikipedia

Wikipedia is a web-based encyclopedia that is collaboratively written by volunteers all around the world. Since its creation in 2001, Wikipedia has grown rapidly into one of the most prominent reference sources, its pages frequently appearing among the top search engine results. As of August 2007, there were 75,000 active contributors, with nearly 2 million articles in English alone. A study of the 30 largest language Wikipedias in January 2005 yielded degree distribution exponents ($\gamma_{\text{in}} = 2.15 \pm 0.13$ and $\gamma_{\text{out}} = 2.57 \pm 0.27$) reflective of the web overall.

Unlike the web, however, between 85 and 97 percent of the pages in the various languages belong to the strongly connected component [74] – indicating that Wikipedia content is more consistently bound than the WWW at large. It is also demonstrating densification, with the number of links growing superlinearly with the number of Wikipedia pages as $L \sim N^{\alpha}$, with $\alpha = 1.14 \pm 0.05$. Similar growth patterns, with more links than nodes being added over time, have been observed in the physical internet, patent and scientific citation networks[75] as well as blog citation networks [76].

The Wikipedia networks display a high degree of reciprocity, where one article linking to another frequently corresponds to another linking back. Using the unbiased mutual reciprocity measure specified in Eq. (3), the degree of reciprocity for the different language Wikipedias is $\rho = 0.32 \pm 0.05$, lower than for the crawl of pages in the nd.edu domain ($\rho = 0.52$) [19,20]. The Wikipedia also exhibits a high degree of clustering, with related articles linking together in triads. Specifically, the motif profile closely follows that of the web overall and falls in the same triad significance superfamily.

Beyond its structural characteristics, Wikipedia has been an ideal medium to validate the preferential attachment model. It was used to verify that the probability of acquiring new links is in fact proportional to the number of links already present on the page [77]. Such time resolved data had previously been difficult to obtain, but Wikipedia provides a complete history of edits for download. Two interesting observations were made. For pages with moderate numbers of inlinks and outlinks, the proportion is

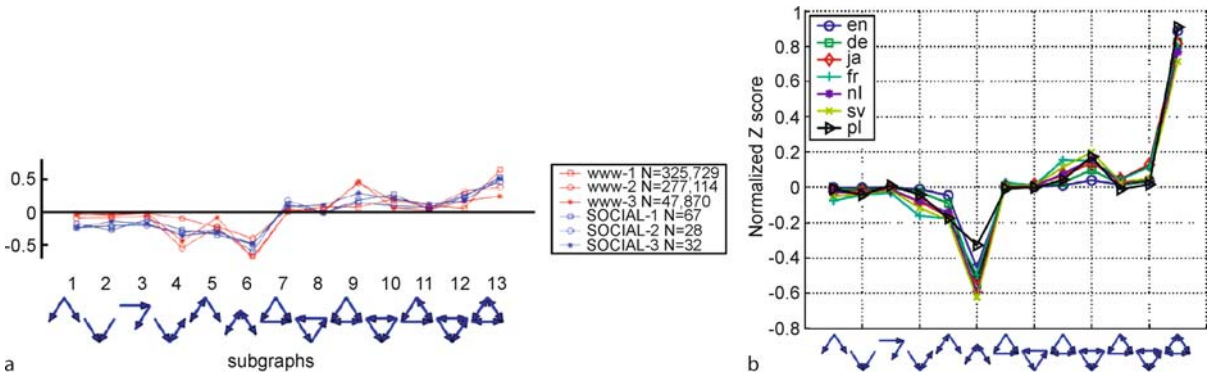

World Wide Web, Graph Structure, Figure 5

a Web motif profile b Wikipedia motif profile. The motif profiles of several web samples and social networks compared to profiles found in the different language Wikipedias. The *x* axis depicts the possible triads, and the *y* axis represents normalized *z*-scores comparing the frequency of each motif to randomized versions of the networks

slightly sublinear, but close to unity ($\nu = 0.9$ in Eq. (11)). Second, the probability reaches a peak, where pages that already contain many links are again not as likely to add more, and those that are already heavily referenced do not tend to attract more links. This could represent both mature pages and mature topics, where the links have already been established.

Future Directions

Web 2.0

The Wikipedia and the blogosphere are just two of the many exciting developments in Web 2.0. Web 2.0 refers to collaborative and interactive applications that are fast becoming integrated into the Web. Rather than simply being consumers of information that they retrieve on the web, web users now can easily contribute content through tools such as blogs and wikis. Unlike simple web pages, that are generated by a single author or through a script from a database, Web 2.0 allows users to interact with one another by modifying pages. For example, bloggers can interact with one another by leaving comments on each others posts. Wikipedians interact by collaboratively editing the same Wikipedia entires, and participating in discussions on talk pages dedicated to those entries. Many content providers for text, images, and video allow users to provide reviews in the form of ratings and comments on the content, or to respond by adding content of their own.

In addition to reflecting a change in the way that web content is generated, Web 2.0 sites allow the users to consume web content in a collaborative way. Through RSS feeds, users receive updates of web content rather than having to remember to visit individual sites. They can tag pages with keywords for easy subsequent retrieval, not just for themselves, but for others interested in the same topic. Through collaborative filtering, users can learn about products, news articles, and media that their friends enjoyed. The best recommenders need not be friends however. Collaborative tools such as RSS readers, bookmarking websites, and content providers can automatically identify users with like interests and make recommendations based on items that those users enjoyed. With Web 2.0 technologies, one can view the Web not only as a proliferation of information and content, but also as giant, distributed machinery that allows the most relevant, personalized content to bubble to the top. At the same time that the Web may have been producing a problem through over-production and over-participation, it has solved it through the emergence of a collection of exceedingly simple yet effective filtering tools.

At this stage, it is unclear whether social networking sites, which allow users to explicitly specify who their friends are, will become platforms that other content providers will be integrated into, or whether many different websites will be able to integrate an open, portable social network the users can bring with them. In either case, the Web graph may soon appear rather different, with webpages becoming mere containers of segmented, time-stamped, personalized content. One of the latest trends in combining content is that of mashups. On can use a mashup to overlay, for example, rental property listings or restaurant reviews on a neighborhood map. These mashups will be especially useful on location and user-aware devices.

As our lives become easier to track, from the sites we browse, to the people we know, to our physical movements tracked by cellphones, we may find our lives increasingly caught in the Web. Most people will readily trade

such information for improved access to personalized and geo-specific services and information. It does, however, also raise interesting privacy concerns, since the Web is a highly distributed storage and dissemination medium.

Mobile Web

Another forcing function for Web innovation is the fact that many individuals access the Web from small mobile devices. For many individuals in developing countries, this may be their primary way of accessing the Web. This has lead to the development of web services tailored to the mobile Web.

Access patterns from mobile devices have been shown to follow the same regularities as general web browsing [78]. This holds despite the fact that Web access from such devices presents challenges due to limited download speed and small display size [79]. It also presents opportunities, as users are able to contribute content wherever they may find themselves. Moblogging (mobile blogging) involves updating one's blog instantaneously with images and videos as they are captured. Currently several news sites encourage the contribution of such materials. The mobile Web also allows users to access Web content, such as maps, traffic and tourist information when they are on the go. Again this may change the structure of the Web graph, as its content is delivered in small, personalized chunks to a variety of devices. Even so, one would expect its main features to remain: billions of hyperlinked pieces of information, showing strong regularities that are a reflection of the real world that is now inextricably tied to the Web.

Bibliography

Primary Literature

1. Gray M (1997) Web growth summary. www.mit.edu/~mkgray/net/web-growth-summary.html. Accessed 9 March 2008
2. Walton M (2006) Web reaches new milestone: 100 million sites. www.cnn.com/2006/TECH/internet/11/01/100millionwebsites/index.html. Accessed 9 March 2008
3. Cho J, Garcia-Molina H, Haveliwala T, Lam W, Paepcke A, Raghavan S, Wesley G (2006) Stanford webbase components and applications. ACM Trans Inter Tech 6(2):153–186
4. Boldi P, Vigna S (2004) The webgraph framework I: compression techniques. In: Proceedings of the 13th International Conference on World Wide Web, New York, NY, pp 595–602
5. Boldi P, Codenotti B, Santini M, Vigna S (2004) UbiCrawler: a scalable fully distributed Web crawler. Softw Pract Experience 34(8):711–726
6. Adamic LA (1999) The Small World Web. Proc ECDL 99:443–452
7. Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener J (2000) Graph structure in the Web. Comput Netw 33(1–6):309–320
8. Joshi A, Kumar R, Reed B, Tomkins A (2007) Anchor-based proximity measures. In: Proceedings of the 16th International World Wide Web Conference, Banff, Canada
9. Fetterly D, Manasse M, Najork M (2004) Spam, damn spam, and statistics: using statistical analysis to locate spam web pages. In: Proceedings of the 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS 2004, Paris, France, pp 1–6
10. Zhang B, Li H, Liu Y, Ji L, Xi W, Fan W, Chen Z, Ma WY (2005) Improving web search results using affinity graph. In: SIGIR '05: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, New York, pp 504–511
11. Broder AZ, Lempel R, Maghoul F, Pedersen J (2006) Efficient PageRank approximation via graph aggregation. Inf Retr 9(2):123–138
12. Adamic LA, Huberman BA (2002) Zipfs law and the internet. Glottometrics 3:143–150
13. Mitzenmacher M (2003) A brief history of generative models for power law and lognormal distributions. Internet Math 1(2):226–251
14. Newman MEJ (2005) Power laws, Pareto distributions and Zipf's law. Contemp Phys 46(5):323–351
15. Barabási AL, Albert R (1999) Emergence of Scaling in Random Networks. Science 286(5439):509
16. Pennock DM, Flake GW, Lawrence S, Glover EJ, Giles CL (2002) Winners don't take all: Characterizing the competition for links on the web. Proc Natl Acad Sci 99(8):5207
17. Newman MEJ (2002) Assortative Mixing in Networks. Phys Rev Lett 89(20):208701
18. Dill S, Kumar R, McCurley KS, Rajogopalan S, Sivakumar D, Tomkins A (2002) Self-Similarity In the Web. ACM Trans Internet Technol 2(3):205–223
19. Albert R, Jeong H, Barabási AL (1999) Internet: Diameter of the world-wide web. Nature 401:130–131. doi:10.1038/43601
20. Garlaschelli D, Loffredo MI (2004) Patterns of link reciprocity in directed networks. Phys Rev Lett 93(26):268701
21. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network Motifs: Simple Building Blocks of Complex Networks. Science 298(5594):824–827
22. Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S, Ayzenshtat I, Sheffer M, Alon U (2004) Superfamilies of designed and evolved networks. Science 303(5663):1538–1542
23. Brewington BE, Cybenko G (2000) How dynamic is the Web? Computer Netw 33(1–6):257–276
24. Fetterly D, Manasse M, Najork M, Wiener JL (2004) A large-scale study of the evolution of Web pages. Softw Pract Experience 34(2):213–237
25. Cho J, Garcia-Molina H (2003) Estimating frequency of change. ACM Trans Internet Technol 3(3):256–290
26. Douglis F, Feldmann A, Krishnamurthy B, Mogul J (1997) Rate of change and other metrics: a live study of the world wide web. USENIX Symposium on Internet Technologies and Systems, vol 119
27. Dasgupta A, Ghosh A, Kumar R, Olston C, Pandey S, Tomkins A (2007) The discoverability of the web. In: Proceedings of the 16th International Conference on World Wide Web, Banff, Canada, pp 421–430
28. Huberman BA, Pirolli PLT, Pitkow JE, Lukose RM (1998) Strong Regularities in World Wide Web Surfing. Science 280 (5360):95

29. Lawrence S, Giles CL (1998) Searching the World Wide Web. Science 280(5360):98
30. Albert R, Barabási AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74(1):47–97
31. Simon HA (1955) On a Class of Skew Distribution Functions. Biometrika 42(3/4):425–440
32. Yule GU (1925) A Mathematical Theory of Evolution, Based on the Conclusions of Dr. JC Willis, FRS. Philos Trans Roy Soc Lond Ser B, Containing Papers of a Biological Character 213:21–87
33. Price DS (1976) A general theory of bibliometric and other cumulative advantage processes. J Am Soc Inf Sci 27(5–6):292–306
34. Adamic LA, Huberman BA (2000) Power-law distribution of the world wide web. Science 287(5461):2115a
35. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440–442
36. Albert R, Barabási A-L (2000) Topology of Evolving Networks: Local Events and Universality. Phys Rev Lett 85(24):5234–5237
37. Dorogovtsev SN, Mendes JFF (2000) Scaling behaviour of developing and decaying networks. Europhys Lett 52(1):33–39
38. Bianconi G, Barabasi AL (2001) Competition and multiscaling in evolving networks. Europhys Lett 54(4):436–442
39. Jeong H, Neda Z, Barabasi AL (2003) Measuring preferential attachment in evolving networks. Europhys Lett 61(4):567–572
40. Vázquez A (2003) Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. Phys Rev E 67(5):56104
41. Jackson MO, Rogers BW (2007) Meeting Strangers and Friends of Friends: How Random Are Social Networks? Am Econ Rev 97(3):890–915
42. Kumar R, Raghavan P, Rajagopalan S, Sivakumar D, Tomkins A, Upfal E (2000) Stochastic models for the web graph. In: Proceedings of the 41st Annual Symposium on Foundations of Computer Science, Redondo Beach, p 57
43. Leskovec J, Kleinberg J, Faloutsos C (2007) Graph evolution: Densification and shrinking diameters. ACM Trans Knowl Discov Data (TKDD) 1(1)
44. Menczer F (2002) Growing and navigating the small world Web by local content. Proc Natl Acad Sci 99(22):14014–14019
45. Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine. Comput Netw ISDN Syst 30(1–7):107–117
46. Litvak N, Volkovich Y, Donato D (2007) Determining factors behind the pagerank log-log plot. Lecture notes in computer science 4863:108
47. Haveliwala Taher H (2002) Topic-sensitive pagerank. In: WWW '02: Proceedings of the 11th International Conference on World Wide Web, New York, pp 517–526
48. Henzinger MR, Heydon A, Mitzenmacher M, Najork M (2000) On near-uniform URL sampling. Comput Netw 33(1–6):295–308
49. Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5):604–632
50. Pass G, Chowdhury A, Torgeson C (2006) A picture of search. Infoscale '06, Hong Kong. In: Proceedings of the 1st international conference on Scalable information systems. ACM, New York, pp 1
51. Fortunato S, Flammini A, Menczer F, Vespignani A (2006) Topical interests and the mitigation of search engine bias. Proc Natl Acad Sci 103(34):12684
52. Kleinberg JM, Kumar R, Raghavan P, Rajagopalan S, Tomkins AS (1999) The Web as a Graph: Measurements, Models, and Methods. Computing and Combinatorics: 5th Annual International Conference, Cocoon'99, Tokyo, Japan
53. Dourisboure Y, Geraci F, Pellegrini M (2007) Extraction and classification of dense communities in the web. In: Proceedings of the 16th International Conference on World Wide Web, Banff, Canada, pp 461–470
54. Kumar R, Raghavan P, Rajagopalan S, Tomkins A (1999) Extracting large-scale knowledge bases from the web. In: Proceedings of the 25th Very Large Data Bases Conference, Edinburgh, UK, pp 639–650
55. Flake GW, Lawrence S, Giles CL, Coetzee FM (2002) Self-organization and identification of Web communities. Computer 35(3):66–70
56. Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70(6):66111
57. Danon L, Díaz-Guilera A, Duch J, Arenas A (2005) Comparing community structure identification. J Stat Mech Theor Exp 9:P09008
58. Castillo C, Donato D, Gionis A, Murdock V, Silvestri F (2007) Know your neighbors: Web spam detection using the web topology. In: Proceedings of SIGIR, pp 423–430. ACM Press, Amsterdam
59. Leskovec J, Dumais S, Horvitz E (2007) Web projections: learning from contextual subgraphs of the web. In: Proceedings of the 16th International Conference on World Wide Web, Banff, Canada, pp 471–480
60. Anderson C (2006) The long tail. Hyperion, New York
61. Kumar R, Novak J, Raghavan P, Tomkins A (2004) Structure and evolution of blogspace. Commun ACM 47(12):35–39
62. Liben-Nowell D, Novak J, Kumar R, Raghavan P, Tomkins A (2005) Geographic routing in social networks. Proc Natl Acad Sci 102(33):11623–11628
63. Travers J, Milgram S (1969) An experimental study of the small world problem. Sociometry 32:425–443
64. Kleinberg J (2000) Navigation in a small world. Nature 406(6798):845
65. Kumar R, Novak J, Raghavan P, Tomkins A (2005) On the bursty evolution of blogspace. World Wide Web 8(2):159–178
66. Hindman M, Tsioutsiouliklis K, Johnson JA (2003) Googlearchy: How a Few Heavily-Linked Sites Dominate Politics on the Web. Annual Meeting of the Midwest Political Science Association
67. Drezner DW, Farrell H (2004) The power and politics of blogs. Download at http://www.danieldrezner.com/research/blogpaperfinal.pdf
68. Adamic LA, Glance N (2005) The political blogosphere and the 2004 US election: divided they blog. In: Proceedings of the 3rd International Workshop on Link Discovery, Aug 21–25, Chicago, IL, pp 36–43
69. Newman MEJ (2006) Modularity and community structure in networks. Proc Natl Acad Sci 103(23):8577–8582
70. Gruhl D, Guha R, Liben-Nowell D, Tomkins A (2004) Information diffusion through blogspace. In: WWW '04: Proceedings of the 13th International Conference on World Wide Web. ACM Press, New York, pp 491–501
71. Kempe D, Kleinberg J, Tardos É (2003) Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, pp 137–146
72. Adar E, Zhang L, Adamic LA, Lukose RM (2004) Implicit structure and the dynamics of blogspace. Workshop on the Weblogging Ecosystem, New York, NY

73. Leskovec J, Krause A, Guestrin C, Faloutsos C, VanBriesen J, Glance N (2007) Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, pp 420–429
74. Zlatic V, Bozicevic M, Stefancic H, Domazet M (2006) Wikipedias: Collaborative web-based encyclopedias as complex networks. Phys Rev E 74(1):016115
75. Leskovec J, Kleinberg J, Faloutsos C (2005) Graphs over time: densification laws, shrinking diameters and possible explanations. In: Conference on Knowledge Discovery in Data, Chicago, IL, USA, pp 177–187
76. Shi X, Tseng B, Adamic LA (2007) Looking at the Blogosphere Topology through Different Lenses. Proceedings of ICWSM'07. Boulder, CO, USA
77. Capocci A, Servedio VDP, Colaiori F, Buriol LS, Donato D, Leonardi S, Caldarelli G (2006) Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia. Phys Rev E 74(3):036116
78. Halvey M, Keane MT, Smyth B (2006) Mobile web surfing is the same as web surfing. Commun ACM 49(3):76–81
79. Buyukkokten O, Garcia-Molina H, Paepcke A, Winograd T (2000) Power browser: efficient web browsing for pdas. In: CHI '00: Proceedings of the SIGCHI conference on Human Factors in Computing Systems. ACM, New York, pp 430–437

Books and Reviews

Aiello W, Broder A, Janssen J, Milios E (eds) (2006) Algorithms and Models for the WebGraph: Proceedings of the 4th International Workshop, WAW 2006, Banff, Canada, 30 Nov–1 Dec

Bonato A (2005) A Survey of Models of the Web graph. In: López-Ortiz A, Hamel A (eds) Combinatorial and Algorithmic Aspects of Networking. Lecture Notes in Computer Science, vol 3405. Springer, Berlin, pp 159

Bonato A, Chung FRK (eds) (2007) Algorithms and Models for the WebGraph: Proceedings of the 5th International Workshop, WAW 2007, San Diego, CA, USA

Caldarelli G (2007) Technological Networks: Internet and WWW in Scale-Free Networks: Complex Webs in Nature and Technology. Oxford University Press, Oxford

Huberman BA (2001) The Laws of the Web: Patterns in the Ecology of Information. MIT, Cambridge

Z

Zero-Sum Two Person Games

T.E.S. Raghavan
Department of Mathematics, Statistics and Computer Science, University of Illinois, Chicago, USA

Article Outline

Introduction
Games with Perfect Information
Mixed Strategy and Minimax Theorem
Behavior Strategies in Games with Perfect Recall
Efficient Computation of Behavior Strategies
General Minimax Theorems
Applications of Infinite Games
Epilogue
Acknowledgment
Bibliography

Introduction

Conflicts are an inevitable part of human existence. This is a consequence of the competitive stances of greed and the scarcity of resources, which are rarely balanced without open conflict. Epic poems of the Greek, Roman, and Indian civilizations which document wars between nation-states or clans reinforce the historical legitimacy of this statement. It can be deduced that domination is the recurring theme in human conflicts. In a primitive sense this is historically observed in the domination of men over women across cultures while on a more refined level it can be observed in the imperialistic ambitions of nation-state actors. In modern times, a new source of conflict has emerged on an international scale in the form of economic competition between multinational corporations.

While conflicts will continue to be a perennial part of human existence, the real question at hand is how to formalize mathematically such conflicts in order to have a grip on potential solutions. We can use mock conflicts in the form of parlor games to understand and evaluate solutions for real conflicts. Conflicts are unresolvable when the participants have no say in the course of action. For example one can lose interest in a parlor game whose entire course of action is dictated by chance. Examples of such games are Chutes and Ladders, Trade, Trouble etc. Quite a few parlor games combine tactical decisions with chance moves. The game Le Her and the game of Parcheesi are typical examples. An outstanding example in this category is the game of backgammon, a remarkably deep game. In chess, the player who moves first is usually determined by a coin toss, but the rest of the game is determined entirely by the decisions of the two players. In such games, players make strategic decisions and attempt to gain an advantage over their opponents.

A game played by two rational players is called zero-sum if one player's gain is the other player's loss. Chess, Checkers, Gin Rummy, Two-finger Morra, and Tic-Tac-Toe are all examples of zero-sum two-person games. Business competition between two major airlines, two major publishers, or two major automobile manufacturers can be modeled as a zero-sum two-person games (even if the outcome is not precisely zero-sum). Zero-sum games can be used to construct Nash equilibria in many dynamic non-zero-sum games [64].

Games with Perfect Information

Emptying a Box

Example 1 A box contains 15 pebbles. Players I and II remove between one and four pebbles from the box in alternating turns. Player I goes first, and the game ends when all pebbles have been removed. The player who empties the box on his turn is the winner, and he receives \$1 from his opponent.

The players can decide in advance how many pebbles to remove in each of their turn. Suppose a player finds x pebbles in the box when it is his turn. He can decide to remove 1, 2, 3 or at most 4 pebbles. Thus a *strategy* for a player is any function f whose domain is $X = \{1, 2 \ldots, 15\}$ and range is

$R \subseteq \{1, 2, 3, 4\}$ such that $f(x) \leq \min(x, 4)$. Given strategies f, g for players I and II respectively, the game evolves by executing the strategies decided in advance. For example if, say

$$f(x) = \begin{cases} 2 & \text{if } x \text{ is even} \\ 1 & \text{if } x \text{ is odd} \end{cases},$$

$$g(x) = \begin{cases} 3 & \text{if } x \geq 3 \\ x & \text{otherwise} \end{cases}.$$

The alternate depletions lead to the following scenario

move by	*I*	*II*	*I*	*II*	*I*	*II*	*I*	*II*
removes	1	3	1	3	1	3	1	2
leaving	14	11	10	7	6	3	2	0 .

In this case the winner is Player II. Actually in his first move Player II made a bad move by removing 3 out of 14. Player I could have exploited this. But he did not! Though he made a good second move, he reverted back to his naive strategy and made a bad third move. The question is: Can Player II ensure victory for himself by intelligently choosing a suitable strategy? Indeed Player II can win the game with any strategy satisfying the conditions of g^* where

$$g^*(x) = \begin{cases} 1 & \text{if } x - 1 \quad \text{is a multiple of 5} \\ 2 & \text{if } x - 2 \quad \text{is a multiple of 5} \\ 3 & \text{if } x - 3 \quad \text{is a multiple of 5} \\ 4 & \text{if } x - 4 \quad \text{is a multiple of 5} \end{cases}.$$

Since the game starts with 15 pebbles, Player I must leave either 14 or 13 or 12, or 11 pebbles. Then Player II can in his turn remove 1 or 2 or 3 or 4 pebbles so that the number of pebbles Player I finds is a multiple of 5 at the beginning of his turn. Thus Player II can leave the box empty in the last round and win the game.

Many other combinatorial games could be studied for optimal strategic behavior. We give one more example of a combinatorial game, called the *game of Nim* [13].

Nim Game

Example 2 Three baskets contain 10, 11, and 16 oranges respectively. In alternating turns, Players I and II choose a non-empty basket and remove at least one orange from it. The player may remove as many oranges as he wishes from the chosen basket, up to the number the basket contains. The game ends when the last orange is removed from the last non-empty basket. The player who takes the last orange is the winner.

In this game as in the previous example at any stage the players are fully aware of what has happened so far and what moves have been made. The full history and the state of the game at any instance are known to both players. Such a game is called a game with *perfect information*. How to plan for future moves to one's advantage is not at all clear in this case. Bouton [13] proposed an ingenious solution to this problem which predates the development of formal game theory.

His solution hinges on the binary representation of any number and the inequality that $1+2+4+\ldots+2^n < 2^{n+1}$. The numbers 10, 11, 16 have the binary representation

Number		Binary representation
10	=	1010
11	=	1011
16	=	10000
Column totals (in base 10 digits)	=	12021 .

Bouton made the following key observations:

1. If at least one column total is an odd number, then the player who is about to make a move can choose one basket and by removing *a suitable number* of oranges leave all column totals even.
2. If at least one basket is nonempty and if all column totals are even, then the player who has to make a move will end up leaving an odd column total.

By looking for the first odd column total from the left, we notice that the basket with 16 oranges is the right choice for Player I. He can remove the left most 1 in the binary expansion of 16 and change all the other binary digits to the right by 0 or 1. The key observation is that the new number is strictly less than the original number. In Player I's move, at least one orange will be removed from a basket. Furthermore, the new column totals can all be made even. If an original column total is even we leave it as it is. If an original column total is odd, we make it even by making any 1 a 0 and any 0 a 1 in those cases which correspond to the basket with 16 oranges. For example the new binary expansion corresponds to removing all but 1 orange from basket 3. We have

10	=	1010
11	=	1011
1	=	0001
Column totals in base 10	=	2022 .

In the next move, no matter what a player does, he has to leave one of the 1's a 0 and the column total in that column will be odd and the move is for Player I. Thus Player I will be the first to empty the baskets and win the game.

For the game of Nim we found a constructive and explicit strategy for the winner regardless of any action by the opponent. Sometimes one may be able to assert who should be the winner without knowing any winning strategy for the player!

Definition 3 A zero-sum two person game has perfect information if, at each move, both players know the complete history so far.

There are many variations of nim games and other combinatorial games like Chess and Go that exploit the combinatorial structure of the game or the end games to develop winning strategies. The classic monographs on combinatorial game theory is by Berlekamp, Conway, and Guy [8] on Winning Ways for your Mathematical Plays, whose mathematical foundations were provided by Conway's earlier book On Numbers and Games. These are often characterized by sequential moves by two players and the outcome is either a win or lose kind. Since the entire history of past moves is common knowledge, the main thrust is in developing winning strategies for such games.

Definition 4 A zero-sum two person game is called a win-lose game if there are no chance moves and the final outcome is either Player I wins or loses (Player II wins) the game. (In other words, there is no way for the game to end in a tie.)

The following is a fundamental theorem of Zermelo [70].

Theorem 5 *Any zero-sum two person perfect information win-lose game Γ with finitely many moves and finitely many choices in each move has a winner with an optimal winning strategy.*

Proof Let Player I make the first move. Then depending on the choices available, the game evolves to a new set of subgames which on their own right are also win-lose games of perfect information. Among these subgames the one with the longest play will have fewer moves than the original game. By an induction on the length of the longest play, we can find a winner with a winning strategy, one for each subgame. Each player can develop good strategies for the original game as follows. Suppose the subgames are $\Gamma_1, \Gamma_2, \ldots, \Gamma_k$. Now among these subgames, let Γ_s be a game where Player I can ensure a victory for himself, no matter what Player II does in the subgame. In this case, Player I can determine at the very beginning, the right choice of action which leads to the subgame Γ_s. A good strategy for Player I is simply the choice s in the first move followed by his good strategy in the subgame Γ_s. Player II's strategy for the original game is simply a k-tuple of strategies, one for each subgame. Player II must be ready to use an optimal strategy for the subgame Γ_r in case the first move of Player I leads to playing Γ_r, which is favorable to Player II. Suppose no subgame Γ_s has a winning strategy for Player I. Then Player II will be the winner in each subgame. To achieve this, Player II must use his winning strategy in each subgame they are lead to. Such a k-tuple of winning strategies, one for each subgame, is a winning strategy for Player II for the original game Γ. □

The Game of Hex

An interesting win-lose game was made popular by John Nash in late forties among Princeton graduate students. While the original game is aesthetically pleasing with its hexagonal tiles forming a 10×10 rhombus, it is more convenient to use the following equivalent formulation for its mathematical simplicity. The version below is extendable to multi person games and is useful for developing important algorithms [28].

Let B_n be a square board consisting of lattice points $\{(i, j)\colon 1 \le i \le n, 1 \le j \le n\}$. The game involves occupation of unoccupied vertices by players I and II. The board is enlarged with a frame on all sides. The frame $\mathbf{F}$ consists of lattice points $\mathbf{F} = \{(i, j)\colon 0 \le i \le n+1, 0 \le j \le n+1$ where either $i = 0$ or $n+1$, or $j = 0$ or $n+1\}$. The frame on the west side $W = \{(i, j)\colon i = 0\} \cap \mathbf{F}$ and the frame on the east side $E = \{(i, j)\colon i = n+1\} \cap \mathbf{F}$ are reserved for Player I. Similarly the frame on the south side $S = \{(i, j)\colon 0 < i < n+1, j = 0\} \cap \mathbf{F}$ and frame on the north side $N = \{(i, j)\colon (0 < i < n+1, j = n+1)\} \cap \mathbf{F}$ are reserved for Player II. Two lattice points $P = (x_1, y_1)$, $Q = (x_2, y_2)$ are called *adjacent vertices* iff either $x_1 \le x_2, y_1 \le y_2$ or $x_1 \ge x_2, y_1 \ge y_2$ and $\max(|x_1 - x_2|, |y_1 - y_2|) = 1$. For example the lattice points $(4, 10)$ and $(5, 11)$ are adjacent while $(4, 10)$ and $(5, 9)$ are not adjacent. Six vertices are adjacent to any interior lattice point of the Hex board B_n while lattice points on the frame will have fewer than six adjacent vertices.

The game is played as follows: Players I and II, in alternate turns, choose a vertex from the available set of unoccupied vertices. The aim of Player I is to occupy a bridge of adjacent vertices that links a vertex on the west boundary with a vertex on the east boundary. Player II has a similar objective to connect the north and south boundary with a bridge.

Theorem 6 *The game of Hex can never end in a draw. For any $T \subseteq B_n$ occupied by Player I and the complement T^c occupied by Player II, either T contains a winning bridge for Player I or T^c contains a winning bridge for Player II. Further only one can have a winning bridge.*

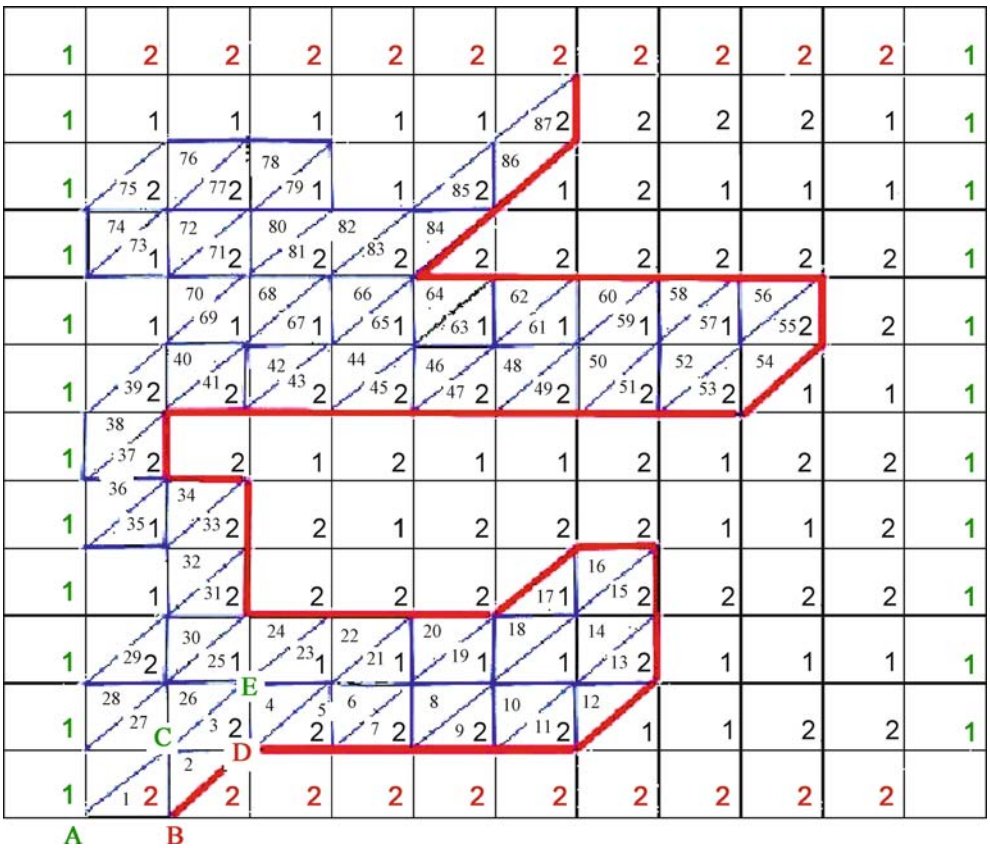

Zero-Sum Two Person Games, Figure 1
Hex path via mate triangles

Proof We label any vertex with 1 or 2 depending on who (Player I or Player II) occupies the vertex. Consider triangles Δ formed by vertices that are mutually adjacent to each other. Two such triangles are called *mates* if they share a common side. Either all the 3 vertices of the triangle are occupied by one player or two vertices by one player and the third by the other player. For example if $P = (x_1, y_1), Q = (x_2, y_2), R = (x_3, y_3)$ are adjacent to each other, and if P, Q, R are occupied by say, I, II, and I, they get the labels 1, 2 and 1 respectively. The triangle has exactly 2 sides (PQ and QR) with vertices labeled 1 and 2. The algorithm described below involves entering such a triangle via one side with vertex labels 1 and 2 and exiting via the other side with vertex labels 1 and 2. Suppose we start at the south west corner triangle Δ_0 (in the above figure) with vertex $A = (0, 0)$ occupied by player I (labeled 1), $B = (1, 0)$ occupied by player II (labeled 2), and suppose $C = (1, 1)$ is occupied by, player I (labeled 1). Since we want to stay inside the framed Hex board, the only way to exit Δ_0 via a side with vertices labeled 1 and 2 is to exit via BC. We move to the unique mate triangle Δ_1 which shares the common side BC which has vertex labels 1 and 2. The mate triangle Δ_1 has vertices $(1, 0)$, $(1, 1)$, and $(1, 0) - (0, 0) + (1, 1) = (2, 1)$. Suppose $D = (2, 1)$ is labeled 2, then we exit via the side CD to the mate triangle with vertices C, D, and $E = (1, 1) - (1, 0) + (2, 1) = (2, 2)$. Each time we find the player of the new vertex with his label, we drop out the other vertex of the same player from the current triangle and move into the new mate triangle. In each iteration there is exactly one new mate triangle to move into. Since in the initial step we had a unique mate triangle to move into from Δ_0, there is no way for the algorithm to reenter a mate triangle visited earlier. This process must terminate at a vertex on the North or East boundary. One side of these triangles will all have the same label forming a bridge which joins the appropriate boundaries and forms a winning path. The winning player's bridge will obstruct the bridge the losing player attempted to complete. The game of Hex and its winning strategy is a powerful tool in developing

algorithms for computing approximate fixed points. Hex is an example of a game where we do know that the first player can win, but we don't know how (for a sufficiently large board). □

Approximate Fixed Points

Let I^2 be the unit square $0 \le x, y \le 1$. Given any continuous function: $\mathbf{f} = (f_1, f_2)\colon I^2 \to I^2$, Brouwer's fixed point theorem asserts the existence of a point (x^*, y^*) such that $\mathbf{f}(x^*, y^*) = (x^*, y^*)$. Our Hex path building algorithm due to Gale [28] gives a constructive approach to locating an approximate fixed point.

Given $\epsilon > 0$, by uniform continuity we can find a $\delta > \frac{1}{n} > 0$ such that if (i, j) and (i', j') are adjacent vertices of a Hex board B_n, then

$$\begin{aligned} \left| f_1\left(\frac{i}{n}, \frac{j}{n}\right) - f_1\left(\frac{i'}{n}, \frac{j'}{n}\right) \right| &\le \epsilon\,, \\ \left| f_2\left(\frac{i}{n}, \frac{j}{n}\right) - f_2\left(\frac{i'}{n}, \frac{j'}{n}\right) \right| &\le \epsilon\,. \end{aligned} \tag{1}$$

Consider the the 4 sets:

$$H^+ = \left\{ (i, j) \in B_n \colon f_1\left(\frac{i}{n}, \frac{j}{n}\right) - \frac{i}{n} > \epsilon \right\}, \tag{2}$$

$$H^- = \left\{ (i, j) \in B_n \colon f_1\left(\frac{i}{n}, \frac{j}{n}\right) - \frac{i}{n} < -\epsilon \right\}, \tag{3}$$

$$V^+ = \left\{ (i, j) \in B_n \colon f_2\left(\frac{i}{n}, \frac{j}{n}\right) - \frac{j}{n} > \epsilon \right\}, \tag{4}$$

$$V^- = \left\{ (i, j) \in B_n \colon f_2\left(\frac{i}{n}, \frac{j}{n}\right) - \frac{j}{n} < -\epsilon \right\}. \tag{5}$$

Intuitively the points in H^+ under $\mathbf{f}$ are moved further to the right (with increased x coordinate) by more than ϵ. Points in V^- under $\mathbf{f}$ are moved further down (with decreased y coordinate) by more than ϵ. We claim that these sets cannot cover all the vertices of the Hex board. If it were so, then we will have a winner, say Player I with a winning path, linking the East and West boundary frames. Since points of the East boundary have the highest x coordinate, they cannot be moved further to the right. Thus vertices in H^+ are disjoint with the East boundary and similarly vertices in H^- are disjoint with the West boundary. The path must therefore contain vertices from both H^+ and H^-. However for any $(i, j) \in H^+, (i', j') \in H^-$ we have

$$\begin{aligned} f_1\left(\frac{i}{n}, \frac{j}{n}\right) - \frac{i}{n} &> \epsilon\,, \\ -f_1\left(\frac{i'}{n}, \frac{j'}{n}\right) + \frac{i'}{n} &> \epsilon\,. \end{aligned}$$

Summing the above two inequalities and using (1) we get

$$\frac{i'}{n} - \frac{i}{n} > 2\epsilon\,.$$

Thus the points (i, j) and (i', j') cannot be adjacent and this contradicts that they are part of a connected path. We have a contradiction.

Remark 7 The algorithm attempts to build a winning path and advances by entering mate triangles. Since the algorithm will not be able to cover the Hex board, partial bridge building should fail at some point, giving a vertex that is outside the union of sets H^+, H^-, V^+, V^-. Hence we reach an approximate fixed point while building the bridge.

An Application of the Algorithm

Consider the continuous map of the unit square into itself given by:

$$f_1(x, y) = \frac{x + \max(-2 + 2x + 6y - 6xy, 0)}{1 + \max(-2 + 2x + 6y - 6xy, 0) + \max(2x - 6xy, 0)}$$

$$f_2(x, y) = \frac{y + \max(2 - 6x - 2y + 6xy, 0)}{1 + \max(2 - 6x - 2y + 6xy, 0) + \max(2y - 6xy, 0)}.$$

With $\epsilon = .05$, we can start with a grid of $\delta = .1$ (hopefully adequate) and find an approximate fixed point. In fact for a spacing of .1 units we have the following iterations. The iterations according to Hex rule passed through the following points with $|f_1(x, y) - x|$ and $|f_2(x, y) - y|$ given by Table 1. Thus the approximate fixed point is $x^* = .4$, $y^* = .3$.

Extensive Games and Normal Form Reduction

Any game as it evolves can be represented by a rooted tree Γ where the root vertex corresponds to the initial move. Each vertex of the tree represents a particular move of a particular player. The alternatives available in any given move are identified with the edges emanating from the vertex that represents the move. If a vertex is assigned to chance, then the game associates a probability distribution with the the descending edges. The terminal vertices are called plays and they are labeled with the payoff to Player I. In zero-sum games Player II's payoff is simply the negative of the payoff to Player I. The vertices for a player

Zero-Sum Two Person Games, Table 1
Table giving the Hex building path

(x, y)	$\lvert f_1 - x\rvert$	$\lvert f_2 - y\rvert$	L
(.0, .0)	0	.6667	1
(.1, 0)	.0167	.5833	2
(.1, .1)	.01228	.4667	2
(0, .1)	.0	.53333	1
(.1, .2)	.007	.35	2
(0, .2)	0	.4	1
(.1, .3)	.002	.233	2
(0, .3)	0	.26667	1
(.1, .4)	.238	.116	1
(.2, .4)	.194	.088	1
(.2, .3)	.007	.177	2
(.3, .4)	.153	.033	1
(.3, .3)	.017	.067	2
(.4, .4)	.116	0	1
(.4, .3)	.0296	0	*

are further partitioned into *information sets*. Information sets must satisfy the following requirements:

- The number of edges descending from any two moves within an information set are same.
- No information set intersects the unique unicursal path from the root to any end vertex of the tree in more than one move.
- Any information set which contains a chance move is a singleton.

We will use the following example to illustrate the extensive form representation:

Example 8 Player I has 3 dice in his pocket. Die 1 is a fake die with all sides numbered one. Die 2 is a fake die with all sides numbered two. Die 3 is a genuine unbiased die. He chooses one of the 3 dice secretly, tosses the die once, and announces the outcome to Player II. Knowing the outcome but not knowing the chosen die, Player II tries to guess the die that was tossed. He pays \$1 to Player I if his guess is wrong. If he guesses correctly, he pays nothing to Player I.

The game is represented by the above tree with the root vertex assigned to Player I. The 3 alternatives at this move are to choose the die with all sides 1 or to choose the die with all sides 2 or to choose the unbiased die. The end vertices of these edges descending from the root vertex are moves for chance. The certain outcomes are 1 and 2 if the die is fake. The outcome is one of the numbers 1, ..., 6 if the die chosen is genuine. The other ends of these edges are moves for Player II. These moves are partitioned into information sets V_1 (corresponding to outcome 1), V_2 (corresponding to outcome 2), and singleton information sets V_3, V_4, V_5, V_6 corresponding to outcomes 3, 4, 5 and 6 respectively. Player II must guess the die based on the information given. If he is told that the game has reached a move in information set V_1, it simply means that the outcome of the toss is 1. He has two alternatives for each move of this information set. One corresponds to guessing the die is fake and the other corresponds to guessing it is genuine. The same applies to the information set V_2. If the outcome is in $V_3, \ldots, V_6$, the clear choice is to guess the die as genuine. Thus a *pure strategy* (master plan) for Player II is to choose a 2-tuple with coordinates taking the values F or G. Here there are 4 pure strategies for Player II. They are: (F_1, F_2), (F_1, G), (G, F_2), (G, G). For example, the first coordinate of the strategy indicates what to guess when the outcome is 1 and the second coordinate indicates what to guess for the outcome 2. For all other outcomes II guesses the die is genuine (unbiased). The payoff to Player I when Player I uses a pure strategy i and Player II uses a pure strategy j is simply the expected income to Player I when the two players choose i and j *simultaneously*. This can as well be represented by a matrix $A = (a_{ij})$ whose rows and columns are pure strategies and the corresponding entries are the expected payoffs. The payoff matrix A given by

$$A = \begin{array}{c} \\ F_1 \\ F_2 \\ G \end{array} \begin{array}{c} \begin{array}{cccc} (F_1, F_2) & (F_1, G) & (G, F_2) & (G, G) \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ \frac{1}{3} & \frac{1}{6} & \frac{1}{6} & 0 \end{pmatrix} \end{array}$$

is called the normal form reduction of the original extensive game.

Saddle Point

The normal form of a zero sum two person game has a saddle point when there is a row r and column c such that the entry a_{rc} is the smallest in row r and the largest in column c. By choosing the pure strategy corresponding to row r Player I guarantees a payoff $a_{rc} = \min_j a_{rj}$. By choosing column c, Player II guarantees a loss no more than $\max_i a_{ic} = a_{rc}$. Thus row r and column c are good pure strategies for the two players. In a payoff matrix $A = (a_{ij})$ row r is said to *strictly dominate* row t if $a_{rj} > a_{tj}$ for all j. Player I, the maximizer, will avoid row t when it is dominated. If rows r and t are not identical and if $a_{rj} \geq a_{tj}$, then we say that row r *weakly dominates* row t.

Example 9 Player I chooses either 1 or 2. Knowing player I's choice Player II chooses either 3 or 4. If the total T

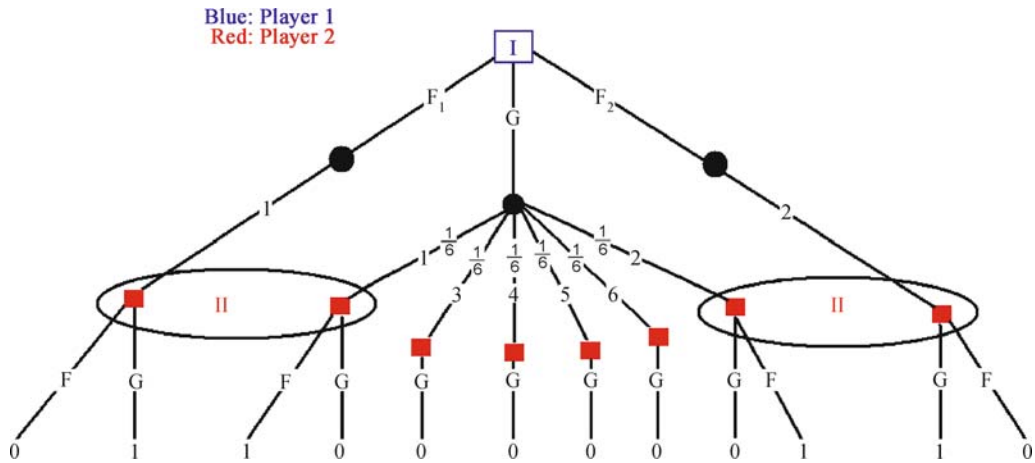

Zero-Sum Two Person Games, Figure 2
Game tree for a single throw with fake or genuine dice

is odd, Player I wins $\$T$ from Player II. Otherwise Player I pays $\$T$ to Player II.

The pure strategies for Player I are simply $\sigma_1 =$ choose 1, $\sigma_2 =$ choose 2. For Player II there are four pure strategies given by: τ_1: choose 3 no matter what I chooses. τ_2: choose 4 no matter what I chooses. τ_3: choose 3 if I chooses 1 and choose 4 if I chooses 2. τ_4: choose 4 if I chooses 3 1 and choose 3 if I chooses 2. This results in a normal form with payoff matrix A for Player I given by:

$$A = \begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{cccc} (3,3) & (4,4) & (3,4) & (4,3) \end{array} \\ \begin{bmatrix} -4 & 5 & -4 & 5 \\ 5 & -6 & -6 & 5 \end{bmatrix} \end{array}.$$

Here we can delete column 4 which dominates column 3. We don't have row domination yet. We can delete column 2 as it weakly dominates column 3. Still we have no row domination after these deletions. We can delete column 1 as it weakly dominates column 3. Now we have strict row domination of row 2 by row 1 and we are left with the row 1, column 3 entry $= -4$. This is a saddle point for this game. In fact we have the following:

Theorem 10 *The normal form of any zero sum two person game with perfect information admits a saddle point. A saddle point can be arrived at by a sequence of row or column deletions. A row that is weakly dominated by another row can be deleted. A column that weakly dominates another column can be deleted. In each iteration we can always find a weakly or strictly dominated row or a weakly or strictly dominating column to be deleted from the current submatrix.*

Mixed Strategy and Minimax Theorem

Zero sum two person games do not always have saddle points in pure strategies. For example, in the game of guessing the die (Example 8) the normal form has no saddle point. Therefore it makes sense for players to choose the pure strategies via a random mechanism. Any probability distribution on the set of all pure strategies for a player is called a *mixed strategy*. In Example 8 a mixed strategy for Player I is a 3-tuple $x = (x_1, x_2, x_3)$ and a mixed strategy for Player II is a 4-tuple $y = (y_1, y_2, y_3, y_4)$. Here x_i is the probability that player I chooses pure strategy i and y_j is the probability that player II chooses pure strategy j. Since the players play independently and make their choices simultaneously, the expected payoff to Player I from Player II is $K(x, y) = \sum_i \sum_j a_{ij} x_i y_j$ where a_{ij} are elements in the payoff matrix A.

We call $K(x, y)$ the mixed payoff where players choose mixed strategies x and y instead of pure strategies i and j. Suppose $x^* = (\frac{1}{8}, \frac{1}{8}, \frac{3}{4})$ and $y^* = (\frac{3}{4}, 0, 0, \frac{1}{4})$. Here x^* guarantees Player I an expected payoff of $\frac{1}{4}$ against any pure strategy j of Player II. By the affine linearity of $K(x^*, y)$ in y it follows that Player I has a guaranteed expectation $= \frac{1}{4}$ against any mixed strategy choice of II. A similar argument shows that Player II can choose the mixed strategy $(\frac{3}{4}, 0, 0, \frac{1}{4})$ which limits his maximum expected loss to $\frac{1}{4}$ against any mixed strategy choice of Player I. Thus

$$\max_x K(x, y^*) = \min_y K(x^*, y) = \frac{1}{4}.$$

By replacing the rows and columns with mixed strategy payoffs we have a saddle point in mixed strategies.

Historical Remarks

The existence of a saddle point in mixed strategies for Example 8 is no accident. All finite games have a saddle point in mixed strategies. This important theorem, called the *minimax theorem*, is the very starting point of game theory. While Borel (see under Ville [65]) considered the notions of pure and mixed strategies for zero sum two person games that have symmetric roles for the players, he was able to prove the theorem only for some special cases. It was von Neumann [66] who first proved the minimax theorem using some intricate fixed point arguments. While several proofs are available for the same theorem [29,34,44,48,49,65,69], the proofs by Ville and Weyl are notable from an algorithmic point of view. The proofs by Nash and Kakutani allow immediate extension to Nash equilibrium strategies in many person non zero sum games. For zero-sum two person games, optimal strategies and Nash equilibrium strategies coincide. The following is the seminal minimax theorem for matrix games.

Theorem 11 (von Neumann) *Let $A = (a_{ij})$ be any $m \times n$ real matrix. Then there exists a pair of probability vectors $x = (x_1, x_2, \ldots, x_m)$ and $y = (y_1, y_2, \ldots, y_n)$ such that for a unique constant v*

$$\sum_{i=1}^{m} a_{ij} x_i \geq v \qquad j = 1, 2, \ldots, n\,,$$
$$\sum_{j=1}^{n} a_{ij} y_j \leq v \qquad i = 1, 2, \ldots, m\,.$$

The probability vectors x, y are called optimal mixed strategies for the players and the constant v is called the value of the game.

H. Weyl [69] gave a complete algebraic proof and proved that the value and some pair of optimal strategies for the two players have all of their coordinates lie in the same ordered subfield as the smallest ordered field containing the payoff entries. Unfortunately his proof was non-constructive. It turns out that the minimax theorem can be proved via linear programming in a constructive way which leads to an efficient computational algorithm *a la* the simplex method [18]. The key idea is to convert the problem to dual linear programming problems.

Solving for Value and Optimal Strategies via Linear Programming

Without loss of generality we can assume that the payoff matrix $A = (a_{ij})_{m\times n} > 0$, that is $a_{ij} > 0$ for all (i, j). Thus we are looking for some v such that:

$$v = \min v_1 \tag{6}$$

such that

$$\sum_{j=1}^{n} a_{ij} y_j \leq v_1\,, \tag{7}$$

$$y_1, \ldots, y_n \geq 0\,, \tag{8}$$

$$\sum_{j=1}^{n} y_j = 1\,. \tag{9}$$

Since the payoff matrix is positive, any v_1 satisfying the constraints above will be positive, so the problem can be reformulated as

$$\max \frac{1}{v_1} = \max \sum_{j=1}^{n} \eta_j \tag{10}$$

such that

$$\sum_{j=1}^{n} a_{ij} \eta_j \leq 1 \quad \text{for all } j\,, \tag{11}$$

$$\eta_j \geq 0 \quad \text{for all } j\,. \tag{12}$$

With $A > 0$, the η_j's are bounded. The maximum of the linear function $\sum_j \eta_j$ is attained at some extreme point of the convex set of constraints (11) and (12). By introducing nonnegative slack variables $s_1, s_2, \ldots, s_m$ we can replace the inequalities (11) by equalities (13). The problem reduces to

$$\max \sum_{j=1}^{n} \eta_j \tag{13}$$

subject to

$$\sum_{j=1}^{n} a_{ij} \eta_j + s_i = 1\,, \quad i = 1, 2, \ldots, m\,, \tag{14}$$

$$y_j \geq 0\,, \quad j = 1, 2, \ldots, n\,, \tag{15}$$

$$s_i \geq 0\,, \quad i = 1, 2, \ldots, m\,. \tag{16}$$

Of the various algorithms to solve a linear programming problem, the simplex algorithm is among the most efficient. It was first investigated by Fourier (1830). But no other work was done for more than a century. The need for its industrial application motivated active research and lead to the pioneering contributions of Kantarowich [1939] (see a translation in Management Science [35]) and Dantzig [18]. It was Dantzig who brought out the earlier investigations of Fourier to the forefront of modern applied mathematics.

Simplex Algorithm Consider our linear programming problem above. Any solution $\eta = (y_1, \ldots, y_n), s = (s_1, \ldots, s_m)$ to the above system of equations is called a feasible solution. We could also rewrite the system as

$$\eta_1 C^1 + \eta_2 C^2 + \cdots + \eta_n C^n + s_1 e^1 + s_2 e^2 + s_m e^m = \mathbf{1}$$
$$\eta_1, \eta_2, \ldots, \eta_n, s_1, s_2, \ldots, s_m \geq 0\,.$$

Here $C^j, j = 1 \ldots, n$ are the columns of the matrix A and e^i are the columns of the $m \times m$ identity matrix. The vector $\mathbf{1}$ is the vector with all coordinates unity. With any extreme point $(\eta, s) = (\eta_1, \eta_2, \ldots, \eta_n, s_1, \ldots, s_m)$ of the convex polyhedron of feasible solutions one can associate with it a set of m linearly independent columns, which form a basis for the column span of the matrix (A, I). Here the coefficients η_j and s_i are equal to zero for coordinates *other* than for the specific m linearly independent columns. By slightly perturbing the entries we can assume that any extreme point of feasible solutions has *exactly m* positive coordinates. Two extreme feasible solutions are called *adjacent* if the associated bases differ in exactly one column.

The key idea behind the simplex algorithm is that an extreme point $P = (\eta^*, s^*)$ is an optimal solution if and only if there is no adjacent extreme point Q for which the objective function has a higher value. Thus when the algorithm is initiated at an extreme point which is not optimal, there must be an adjacent extreme point that strictly improves the objective function. In each iteration, a column from outside the basis replaces a column in the current basis corresponding to an adjacent extreme point. Since there are $m + n$ columns in all for the matrix (A, I), and in each iteration we have strict improvement by our nondegeneracy assumption on the extreme points, the procedure must terminate in a finite number of steps resulting in an optimal solution.

Example 12 Players I and II simultaneously show either 1 or 2 fingers. If T is the total number of fingers shown then Player I receives from Player II $\$T$ when T, is odd and loses $\$T$ to Player II when T is even.

The payoff matrix is given by

$$A = \begin{bmatrix} -2 & 3 \\ 3 & -4 \end{bmatrix}.$$

Add 5 to each entry to get a new payoff matrix with all entries strictly positive. The new game is strategically same as A.

$$\begin{bmatrix} 3 & 8 \\ 8 & 1 \end{bmatrix}.$$

The linear programming problem is given by

$$\max 1.y_1 + 1.y_2 + 0.s_1 + 0.s_2$$
such that
$$\begin{bmatrix} 3 & 8 & 1 & 0 \\ 8 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ s_1 \\ s_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

We can start with the trivial solution $(0, 0, 1, 1)^{\mathrm{T}}$. This corresponds to the basis e^1, e^2 with $s_1 = s_2 = 1$. The value of the objective function is 0. If we make $y_2 > 0$, then the value of the objective function can be increased. Thus we look for a solution to

$$s_1 = 0\,, \quad s_2 > 0\,, \quad y_2 > 0$$

satisfying the constraints

$$8y_2 + 0s_2 = 1$$
$$y_2 + s_2 = 1$$

or to

$$s_2 = 0\,, \quad s_1 > 0\,, \quad y_2 > 0$$

satisfying the constraints

$$8y_2 + s_1 = 1$$
$$y_2 + 0s_1 = 1$$

Notice that $y_2 = \frac{1}{8}, s_2 = \frac{7}{8}$ is a solution to the first system and that the second system has no nonnegative solution. The value of the objective function at this extreme solution is $\frac{1}{8}$. Now we look for an adjacent extreme point. We find that $y_1 = \frac{7}{61}, y_2 = \frac{5}{61}$ is such a solution. The procedure terminates because no adjacent solution with $y_1 > 0, s_1 > 0$ or $y_2 > 0, s_1 > 0$ or $y_1 > 0, s_2 > 0$ or $y_2 > 0, s_2 > 0$ if any has higher objective function value. The algorithm terminates with the optimal value of $\frac{1}{v_1} = \frac{12}{61}$. Thus the value of the modified game is $\frac{61}{12}$, and the value of the original game is $\frac{61}{12} - 5 = \frac{1}{12}$. A good strategy for Player II is obtained by normalizing the optimal solution of the linear program, it is $\eta_1 = \frac{7}{12}, \eta_2 = \frac{5}{12}$. Similarly, from the dual linear program we can see that the strategy $\xi_1 = \frac{7}{12}, \xi_2 = \frac{5}{12}$ is optimal for Player I.

Fictitious Play

Though optimal strategies are not easily found, even naive players can learn to steer their average payoff towards the value of the game from past plays by certain iterative procedures. This learning procedure is known as *fictitious*

play. The two players make their next choice under the assumption that the opponent will continue to choose pure strategies at the same frequencies as what he/she did in the past. If $x^{(n)}$, $y^{(n)}$ are the empirical mixed strategies used by the two players in the first n rounds, then in round $n+1$ Player I pretends that Player II will continue to use $y^{(n)}$ in the future and selects any row i^* such that

$$\sum_j a_{i^*j} y_j^{(n)} = \max_i \sum_j a_{ij} y_j^{(n)} .$$

The new empirical mixed strategy is given by

$$x^{(n+1)} = \frac{1}{n+1} I_{i^*} + \frac{n}{n+1} x^{(n)} .$$

(Here I_{i^*} is the degenerate choice of pure strategy i^*.) This intuitive learning procedure was proposed by Brown [14] and the following convergence theorem was proved by Robinson [56].

Theorem 13

$$\lim_n \min_j \sum_i a_{ij} x_i^{(n)} = \lim_n \max_i \sum_j a_{ij} y_j^{(n)} = v .$$

We will apply the fictitious play algorithm to the following example and get a bound on the value.

Example 14 Player I picks secretly a card of his choice from a deck of three cards numbered 1, 2, and 3. Player II proceeds to guess player I's choice. After each guess player I announces player II's guess as *"High"*, *"Low"* or *"Correct"* as the case may be. The game continues till player II guesses player I's choice correctly. Player II pays to Player I $\$N$ where N is the number of guesses he made.

The payoff matrix is given by

$$A = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccccc} (1,2) & (1,3) & (2) & (3,1) & (3,2) \end{array} \\ \begin{pmatrix} 1 & 1 & 2 & 2 & 3 \\ 2 & 3 & 1 & 3 & 2 \\ 3 & 2 & 2 & 1 & 1 \end{pmatrix} \end{array} .$$

Here the row labels are possible cards chosen by Player I, and the column labels are pure strategies for Player II. For example, the pure strategy (1, 3) for Player II, means that 1 is the first guess and if 1 is incorrect then 3 is the second guess. The elements of the matrix are payoffs to Player I. We can use cumulative total, instead average for the players to make their next choice of row or column based on the totals. We choose the row or column with the least index in case more than one row or one column meets the criterion. The total for the first 10 rounds using fictitious play is given in Table 2.

Zero-Sum Two Person Games, Table 2

Row choices	Total so far					Column choices	Total so far		
R_1	**1**	1	2	2	3	C_1	1	2	**3**
R_3	4	**3**	4	3	4	C_2	2	**5**	5
R_2	6	6	**5**	6	6	C_3	4	6	**7**
R_3	9	8	**7**	7	7	C_3	6	7	**9**
R_3	12	10	9	**8**	8	C_4	8	**10**	10
R_2	15	12	11	**9**	9	C_4	10	**13**	11
R_2	17	15	12	12	**11**	C_5	13	**15**	12
R_2	19	18	**13**	15	13	C_3	15	**16**	14
R_2	21	21	**14**	18	15	C_3	**17**	17	16
R_1	22	22	**16**	20	18	C_3	**19**	18	18

The bold entries give the approximate lower and upper bounds for the total payoff in 10 rounds giving $1.6 \le v \le 1.9$.

Remark 15 Fictitious play is known to have a very poor rate of convergence to the value. While it works for all zero sum two person games, it fails to extend to Nash equilibrium payoffs in bimatrix games even when the game has a unique Nash equilibrium. It extends only to some very special classes like 2×2 bimatrix games and to the so called potential games.

(See Miyasawa [1961], Shapley [1964], Monderer and Shapley [46], Krishna and Sjoestrom [41], and Berger [7]).

Search Games

Search games are often motivated by military applications. An object is hidden in space. While the space where the object is hidden is known, the exact location is unknown. The search strategy consists of either targeting a single point of the space and paying a penalty when the search fails or continue the search till the searcher gets closer to the hidden location. If the search consists of many attempts, then a pure strategy is simply a function of the search history so far. Example 14 is a typical search game. The following are examples of some simple search games that have unexpected turns with respect to the value and optimal strategies.

Example 16 A pet shop cobra of known length $t < 1$ escapes out of the shop and has settled in a nearby tree somewhere along a particular linear branch of unit length. Due to the camouflage, the exact location $[x, x+t]$ where it has settled on the branch is unknown. The shop keeper chooses a point y of his choice and aims a bullet at the point y. In spite of his 100% accuracy the cobra will escape permanently if his targeted point y is outside the settled location of the cobra.

We treat this as a game between the cobra (Player I) and the shop keeper (Player II). Let the probability of survival be the payoff to the cobra. Thus

$$K(x,y) = \begin{cases} 1 & \text{if } y < x\,, \text{ or } y > x + t \\ 0 & \text{otherwise}\,. \end{cases}$$

The pure strategy spaces are $0 \le x \le 1 - t$ for the snake and $0 \le y \le 1$ for the shop keeper. It can be shown that the game has no saddle point and has optimal mixed strategies. The value function $v(t)$ is a discontinuous function of t. In case $\frac{1}{t}$ is an integer n then, a good strategy for the snake is to hide along $[0,t]$, or $[t,2t]$ or $[(n-1)t,1]$ chosen with equal chance. In this case the optimal strategy for the shop keeper is to choose a random point in $[0,1]$. The value is $1 - \frac{1}{n}$. In case $\frac{1}{t}$ is a fraction, let $n = [\frac{1}{t}]$ then the optimal strategy for the snake is to hide along $[0,t]$, or $[t,2t], \ldots$ or $[(n-1)t, nt]$. An optimal strategy for the shop keeper is to shoot at one of the points $\frac{1}{n+1}, \frac{2}{n+1}, \ldots, \frac{n}{n+1}$ chosen at random.

Example 17 While mowing the lawn a lady suddenly realizes that she has lost her diamond engagement ring some where in her lawn. She has maximum speed s and will be able to locate the diamond ring from its glitter if she is sufficiently close to, say within a distance ϵ from the ring. What is an optimal search strategy that minimizes her search time.

If we treat Nature as a player against her, she is playing a zero sum two person game where Nature would find pleasure in her delayed success in finding the ring.

Search Games on Trees

The following is an elegant search game on a tree. For many other search games the readers can refer to the monographs by Gal [27] and Alpern and Gal [1]. Also see [55].

Example 18 A bird has to look for a suitable location to build its nest for hatching eggs and protecting them against predator snakes. Having identified a large tree with a single predator snake in the neighborhood, the bird has to further decide where to build its nest on the chosen tree. The chance for the survival of the eggs is directly proportional to the distance the snake travels to locate the nest.

While birds and snakes work out their strategies based on instinct and evolutionary behavior, we can surely approximate the problem by the following zero sum two person search game. Let $T = (X, \mathcal{E})$ be a finite tree with vertex set X and edge set $\mathcal{E}$. Let $O \in X$ be the root vertex. A hider hides an object at a vertex x of the tree. A searcher starts at the root and travels along the edges of the tree such that the path traced covers all the terminal vertices. The search ends as soon as the searcher crosses the hidden location and the payoff to the hider is the distance traveled so far.

By a simple domination argument we can as well assume that the optimal hiding locations are simply the terminal vertices.

Theorem 19 *The search game has value and optimal strategies. The value coincides with the sum of all edge lengths. Any optimal strategy for the hider will necessarily restrict to hide at one of the terminal vertices. Let the least distance traveled to exhaust all end vertices one by one correspond to a permutation σ of the end vertices in the order $w_1, w_2, \ldots, w_k$. Let σ^{-1} be its reverse permutation. Then an optimal strategy for the searcher is to choose one of these two permutations by the toss of a coin. The hider has a unique optimal mixed strategy that chooses each end vertex with positive probability.*

Suppose the tree is a path with root O and a single terminal vertex x. Since the search begins at O, the longest trip is possible only when hider hides at x and the theorem holds trivially. In case the tree has just two terminal vertices besides the root vertex, the possible hiding locations are say, O, x_1, x_2 with edge lengths a_1, a_2. The possible searches are via paths: $O \to x_1 \to O \to x_2$ abbreviated Ox_1Ox_2 or $O \to x_2 \to O \to x_1$, abbreviated Ox_2Ox_1. The payoff matrix can be written as

$$\begin{array}{cc} & \begin{array}{cc} Ox_1Ox_2 & Ox_2Ox_1 \end{array} \\ \begin{array}{c} x_1 \\ x_2 \end{array} & \begin{bmatrix} a_1 & 2a_2 + a_1 \\ 2a_1 + a_2 & a_2 \end{bmatrix} \end{array}.$$

The value of this game is $a_1 + a_2 =$ sum of the edge lengths. We can use an induction on the number of subtrees to establish the value as the sum of edge lengths. We will use an example to just provide the intuition behind the formal proof.

Given any permutation τ of the end vertices (leaves), of the above tree let P be the shortest path from the root vertex that travels along the leaves in that order and returns to the root. Let the reverse path be P^{-1}. Observe that it will cover all edges twice. Thus if the two paths P and P^{-1} are chosen with equal chance by the snake, the average distance traveled by the snake when it locates the bird's nest at an end vertex will be independent of the particular end vertex. For example along the closed path $O \to t \to d \to t \to a \to t \to x \to b \to x \to c \to x \to t \to O \to u \to s \to e \to s \to y \to f \to y \to g \to y \to s \to u \to h \to u \to j \to u \to O$ the distance traveled by the snake to reach leaf e is $(3 + 7 + 7 + \cdots + 8 + 3) = 74$. If the snake travels along the reverse path to reach e the distance traveled

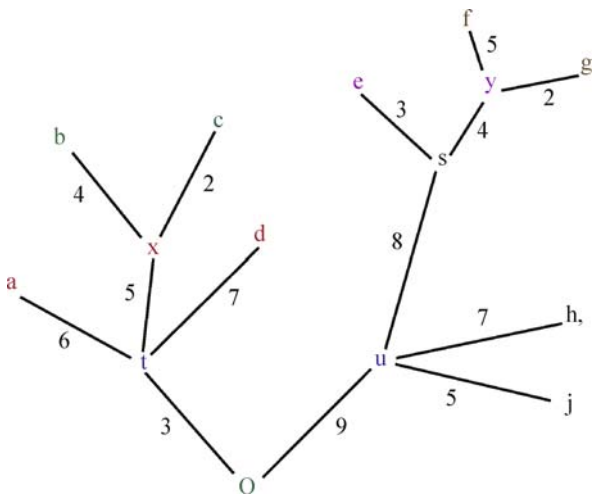

Zero-Sum Two Person Games, Figure 3
Bird trying to hide at a leaf and snake chasing to reach the appropriate leaf via optimal Chinese postman route starting at root *O* and ending at *O*

is $(9+5+5+\cdots+5+4+3)=66$. For example if it is to reach the vertex d then via path P it is $(3+7)$. Via P^{-1} it is to make travel to e and travel from e to d by the reverse path. This is $(66+3+\cdots+6+6+7)=130$. Thus in both cases the average distance traveled is 70. The average distance is the same for every other leaf when P and P^{-1} are used. The optimal Chinese postman route can allow all permutations subject to permuting any leaf of any subtree only among themselves. Thus the subtree rooted at t has leaves a, b, c, d and the subtree rooted at u has leaves e, f, g, h, j. For example while permuting a, b, c, d only among themselves we have the further restriction that between b, c we cannot allow insertion of a or d. For example a, b, c, d and a, d, c, b are acceptable permutations, but not a, b, d, c. It can never be the optimal permuting choice. The same way it applies to the tree rooted at u. For example h, j, e, g, f is part of the optimal Chinese postman route, but h, g, j, e, f is not. We can think of the snake and bird playing the game as follows: The bird chooses to hide in a leaf of the subgame G_t rooted at t or at a leaf of the subgame G_u rooted at u. These leaves exhaust all leaves of the original game. The snake can restrict to only the optimal route of each subgame. This can be thought of as a 2×2 game where the strategies for the two players (bird) and snake are:

Bird:
Strategy 1: Hide optimally in a leaf of G_t,
Strategy 2: Hide optimally in a leaf of G_u.

Snake:
Strategy 1: Search first the leaves of G_t along the optimal Chinese postman route of G_t and then search along the leaves of G_u.
Strategy 2: Search first the leaves of G_u along the optimal Chinese postman route and then search the leaves of G_t along the optimal postman route. The expected outcome can be written as the following 2×2 game. (Here $v(G_t)$, $v(G_u)$ are the values of the subgames rooted at t, u respectively.)

$$\begin{array}{cc} & \begin{array}{cc} G_tG_u & \qquad\qquad G_uG_t \end{array} \\ \begin{array}{c} G_t \\ G_u \end{array} & \begin{bmatrix} [3+v(G_t)] & 2[9+v(G_u)]+[3+v(G_t)] \\ 2[3+v(G_t]+[9+v(G_u)] & [9+v(G_u)] \end{bmatrix} \end{array}$$

Observe that the 2×2 game has no saddle point and hence has value $3+9+v(G_t)+v(G_u)$. By induction we can assume $v(G_t)=24, v(G_u)=34$. Thus the value of this game is 70. This is also the sum of the edge lengths of the game tree. An optimal strategy for the bird can be recursively determined as follows.

Umbrella Folding Algorithm Ladies, when storing umbrellas inside their handbag shrink the central stem of the umbrella and then the stems around all in one stroke. We can mimic a somewhat similar procedure also for our above game tree. We simultaneously shrink the edges $[xc]$ and $[xb]$ to x. In the next round $\{a, x, d\}$ edges $[a, t], [x, t], [d, t]$ can be simultaneously shrunk to t and so on till the entire tree is shrunk to the root vertex O. We do know that the optimal strategy for the bird when the tree is simply the subtree with root x and with leaves b, c is given by $p(b)=\frac{4}{(4+2)}$, $p(c)=\frac{2}{(4+2)}$. Now for the subtree with vertex t and leaves $\{a, b, c, d\}$, we can treat this as collapsing the previous subtree to x and treat stem length of the new subtree with vertices $\{t, a, x, d\}$ as though the three stems $[ta], [tx], [td]$ have lengths $6, 5+(4+2), 7$. We can check that for this subtree game the leaves a, x, d are chosen with probabilities $p(a)=\frac{6}{(6+9+7)}$, $p(x)=\frac{9}{(6+9+7)}$, $p(d)=\frac{7}{(6+9+7)}$. Thus the optimal mixed strategy for the bird for choosing leaf b for our original tree game is to pass through vertices t, x, b and is given by the product $p(t)p(x)p(b)$. We can inductively calculate these probabilities.

Completely Mixed Games and Perron's Theorem on Positive Matrices

A mixed strategy x for player I is called *completely mixed* if it is strictly positive ($x>\mathbf{0}$). A matrix game A is *completely mixed* if and only all optimal mixed strategies for Player I and Player II are completely mixed. The following elegant theorem was proved by Kaplanski [36].

Theorem 20 *A matrix game A with value v is completely mixed if and only if*

1. *The matrix is square.*
2. *The optimal strategies are unique for the two players.*
3. *If $v \neq 0$, then the matrix is nonsingular.*
4. *If $v = 0$, then the matrix has rank $n-1$ where n is the order of the matrix.*

The theory of completely mixed games is a useful tool in linear algebra and numerical analysis [4]. The following is a sample application of this theorem.

Theorem 21 (Perron 1909) *Let A be any $n \times n$ matrix with positive entries. Then A has a positive eigenvalue with a positive eigenvector which is also a simple root of the characteristic equation.*

Proof Let I be the identity matrix. For any $\lambda > 0$, the maximizing player prefers to play the game A rather than the game $A - \lambda I$. The payoff gets worse when the diagonal entries are reached. The value function $v(\lambda)$ of $A - \lambda I$ is a non-increasing continuous function. Since $v(0) > 0$ and $v(\lambda) < 0$ for large λ we have for some $\lambda_0 > 0$ the value of $A - \lambda_0 I$ is 0. Let y be optimal for player II, then $(A - \lambda_0 I)y \leq \mathbf{0}$ implies $\mathbf{0} < Ay \leq \lambda_0 y$. That is $\mathbf{0}$. Since the optimal y is completely mixed, for any optimal x of player I, we have $(A - \lambda_0 I)x = \mathbf{0}$. Thus $x > \mathbf{0}$ and the game is completely mixed. By (2) and (4) if $(A - \lambda_0 I)u = 0$ then u is a scalar multiple of y and so the eigenvector y is geometrically simple. If $B = A - \lambda_0 I$, then B is singular and of rank $n-1$. If (B_{ij}) is the cofactor matrix of the singular matrix B then $\sum_j b_{ij} B_{kj} = 0\,, \quad i = 1, \ldots, n$. Thus row k of the cofactor matrix is a scalar multiple of y. Similarly each column of B is a scalar multiple of x. Thus all cofactors are of the same sign and are different from 0. That is

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \det(A - \lambda I)\Big|_{\lambda_0} = \sum_i B_{ii} \neq 0\,.$$

Thus λ_0 is also algebraically simple. See [4] for the most general extensions of this theorem to the theorems of Perron and Frobenius and to the theory of M-matrices and power positive and polynomially matrices). □

Behavior Strategies in Games with Perfect Recall

Consider any extensive game Γ where the unique unicursal path from an end vertex w to the root x_0 intersects two moves x and y of say, Player I. We say $x \prec y$ if the the unique path from y to x_0 is via move x. Let $U \ni x$ and $V \ni y$ be the respective information sets. If the game has reached a move $y \in V$, Player I will know that it is his turn and the game has progressed to *some* move in V. The game is said to have *perfect recall* if each player can remember all his past moves and the choices made in those moves. For example if the game has progressed to a move of Player I in the information set V he will remember the specific alternative chosen in any earlier move. A move x is *possible* for Player I with his pure strategy π_1, if for some suitable pure strategy π_2 of Player II, the move x can be reached with positive probability using π_1, π_2. An information set U is *relevant* for a pure strategy π_1, for Player I, if some move $x \in U$ is possible with π_1. Let Π_1, Π_2 be pure strategy spaces for players I and II.

Let $\mu_1 = \{q_{\pi_1,\, \pi_1 \in \Pi_1}\}$ be any mixed strategy for Player I. The information set U for Player I is *relevant for the mixed strategy* μ_1 if for some $q_{\pi_1} > 0$, U is relevant for π_1. We say that the information set U for Player I is not relevant for the mixed strategy μ_1 if for all $q_{\pi_1} > 0$, U is not relevant for π_1. Let

$$S_v = \{\pi_1 \colon U \text{ is relevant for } \pi_1 \text{ and } \pi_1(U) = v\}\,,$$

$$S = \{\pi_1 \colon U \text{ is relevant for } \pi_1\}\,,$$

$$T = \{\pi_1 \colon U \text{ is not relevant for } \pi_1 \text{ and } \pi_1(U) = v\}\,.$$

The behavior strategy induced by a mixed strategy pair (μ_1, μ_2) at an information set U for Player I is simply the conditional probability of choosing alternative v in the information set U, given that the game has progressed to a move in U, namely

$$\beta_1(U, v) = \begin{cases} \dfrac{\sum_{\pi_1 \in S_v} q_{\pi_1}}{\sum_{\pi_1 \in S} q_{\pi_1}} & \text{if } U \text{ is relevant for } \mu_1\,, \\ \sum_{\pi_1 \in T} q_{\pi_1} & \text{if } U \text{ is not relevant for } \mu_1\,. \end{cases}$$

The following theorem of Kuhn [42] is a consequence of the assumption of perfect recall.

Theorem 22 *Let μ_1, μ_2 be mixed strategies for players I and II respectively in a zero sum two person finite game Γ of perfect recall. Let β_1, β_2 be the induced behavior strategies for the two players. Then the probability of reaching any end vertex w using μ_1, μ_2 coincides with the probability of reaching w using the induced behavior strategy β_1, β_2. Thus in zero-sum two person games with perfect recall, players can play optimally by restricting their strategy choices just to behavior strategies.*

The following analogy may help us understand the advantages of behavior strategies over mixed strategies. A book has 10 pages with 3 lines per page. Someone wants to glance through the book reading just 1 line from

each page. A master plan (pure strategy) for scanning the book consists of choosing one line number for each page. Since each page has 3 lines, the number of possible plans is 3^{10}. Thus the set of mixed strategies is a set of dimension $3^{10} - 1$. There is another randomized approach for scanning the book. When page i is about to be scanned choose line 1 with probability x_{i1}, line 2 with probability x_{i2} and line 3 with probability x_{i3}. Since for each i we have $x_{i1} + x_{i2} + x_{i3} = 1$ the dimension of such a strategy space is just 20. Behavior strategies are easier to work with. Further Kuhn's theorem guarantees that we can restrict to behavior strategies in games with perfect recall.

In general if there are k alternatives at each information set for a player and if there are n information sets for the player, the dimension of the mixed strategy space is $k^n - 1$. On the other hand the dimension of the behavior strategy space is simply $n(k-1)$. Thus while the dimension of mixed strategy space grows exponentially the dimension of behavior strategy space grows linearly. The following example will illustrate the advantages of using behavior strategies.

Example 23 Player I has 7 dice. All but one are fake. Fake die F_i has the same number i on all faces $i = 1, \ldots, 6$. Die G is the ordinary unbiased die. Player I selects one of them secretly and announces the outcome of a single toss of the die to player II. It is Player II's turn to guess which die was selected for the toss. He gets no reward for correct guess but pays \$1 to Player I for any wrong guess.

Player I has 7 pure strategies while Player II has 2^6 pure strategies. As an example the pure strategy (F_1, G, G, F_4, G, F_6) for Player II is one which guesses the die as fake when the outcome revealed is 1 or 4 or 6, and guesses the die as genuine when the outcome is 2 or 3 or 5. The normal form game is a payoff matrix of size 7×64. For example if G is chosen by Player I, and (F_1, G, G, F_4, G, F_6) is chosen by Player II, the expected payoff to Player I is $\frac{1}{6}[1+0+0+1+0+1] = \frac{1}{2}$. If F_2 is chosen by Player I, the expected payoff is 1 against the above pure strategy of Player II. Now Player II can use the following behavior strategy. If the outcome is i, then with probability q_i he can guess that the die is genuine and with probability $(1 - q_i)$ he can guess that it is from the fake die F_i. The expected behavioral payoff to Player I when he chooses the genuine die with probability p_0 and chooses the fake die F_i with probability p_i, $i = 1, \ldots, i = 6$ is given by

$$K(p,q) = p_0 \frac{1}{6} \sum_{i=1}^{6} (1 - q_i) + \sum_{i=1}^{6} p_i q_i \,.$$

Collecting the coefficients of q_i's, we get

$$K(p,q) = \sum_{i=1}^{6} q_i \left[p_i - \frac{1}{6} p_0 \right] + p_0 \,.$$

By choosing $p_i - \frac{1}{6}p_0 = 0$, we get $p_1 = p_2 = \ldots, p_6 = \frac{1}{6}p_0$. Thus $p = (\frac{1}{2}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12})$. For this mixed strategy for Player I, the payoff to Player I is independent of Player II's actions. Similarly, we can rewrite $K(p,q)$ as a function of p_i's for $i = 1, \ldots, 6$ where

$$K(p,q) = \sum_{k \neq 0} p_k \left[q_k - \frac{1}{6} \sum_{r=1}^{6} (1 - q_r) \right] + \frac{1}{6} \left(\sum_{k=1}^{6} (1 - q_k) \right) \,.$$

This expression can be made independent of p_i's by choosing $q_i = \frac{1}{2}, \quad i = 1, \ldots, 6$. Since the behavioral payoff for these behavioral strategies is $\frac{1}{2}$, the value of the game is $\frac{1}{2}$, which means that Player I cannot do any better than $\frac{1}{2}$ while Player II is able to limit his losses to $\frac{1}{2}$.

Efficient Computation of Behavior Strategies

Introduction

In our above example with one genuine and six fake dice, we used Kuhn's theorem to narrow our search among optimal behavior strategies. Our success depended on exploiting the inherent symmetries in the problem. We were also lucky in our search when we were looking for optimals among equalizers.

From an algorithmic point of view, this is not possible with any arbitrary extensive game with perfect recall. While normal form is appropriate for finding optimal mixed strategies, its complexity grows exponential with the size of the vertex set. The payoff matrix in normal form is in general not a sparse matrix (a sparse matrix is one which has very few nonzero entries) a key issue for data storage and computational accuracies. By sticking to the normal form of a game with perfect recall we cannot take full advantage of Kuhn's theorem in its drastically narrowed down search for optimals among behavior strategies. A more appropriate form for these games is the *sequence form* [63] and *realization probabilities* to be described below. The behavioral strategies that induce the *realization probabilities* grow only linearly in the size of the terminal vertex set. Another major advantage is that the sequence form induces a sparse matrix. It has at most as many non-zero entries as the number of terminal vertices or plays.

Sequence Form

When the game moves to an information set U_1 of say, player I, the perfect recall condition implies that wherever the true move lies in U_1, the player knows the actual alternative chosen in any of the past moves. Let σ_{u_1} denote the sequence of alternatives chosen by Player I in his *past moves*. If no past moves of player I occurs we take $\sigma_{u_1} = \emptyset$. Suppose in U_1 player I selects an action "c" with behavioral probability $\beta_1(c)$ and if the outcome is c the new sequence is $\sigma_{u_1} \cup c$. Thus any *sequence* s_1 for player I is the string of choices in his moves along the partial path from the initial vertex to any other vertex of the tree. Let S_0, S_1, S_2 be the set of all sequences for Nature (via chance moves), Player I and Player II respectively. Given behavior strategies $\beta_0, \beta_1, \beta_2$ Let

$$r_i(s_i) = \prod_{c \in s_i} \beta_i(c)\,, \quad i = 0, 1, 2\,.$$

The functions: $r_i : S_i :\to R : i = 0, 1, 2$ satisfy the following conditions

$$r_i(\emptyset) = 1 \tag{17}$$

$$r_i(\sigma_{u_i}) = \sum_{c \in A(U_i)} r_i(\sigma_{u_i}, c)\,, \quad i = 0, 1, 2 \tag{18}$$

$$r_i(s_i) \geq 0 \quad \text{for all } s_i\,, \quad i = 0, 1, 2\,. \tag{19}$$

Conversely given any such realization functions r_1, r_2 we can define behavior strategies, β_1 say for player I, by

$$\beta_1(U_1, c) = \frac{r_1(\sigma_{u_1} \cup c)}{r_1(\sigma_{u_1})} \qquad \text{for } c \in A(U_1)\,, \text{ and } r_1(\sigma_{u_1}) > 0\,.$$

When $r_1(\sigma_{u_1}) = 0$ we define $\beta_1(U_1, c)$ arbitrarily so that $\sum_{c \in A(U_1)} \beta_1(U_1, c) = 1$. If the terminal payoff to player I at terminal vertex ω is $h(\omega)$, by defining $h(a) = 0$ for all nodes a that are not terminal vertices, we can easily check that the behavioral payoff

$$H(\beta_1, \beta_2) = \sum_{s \in S} h(s) \prod_{i=0}^{2} r_i(s_i)\,.$$

When we work with realization functions $r_i,\ i = 1, 2$ we can associate with these functions the sequence form of payoff matrix whose rows correspond to sequence $s_1 \in S_1$ for Player I and columns correspond to sequence $s_2 \in S_2$ for Player II and with payoff matrix

$$K(s_1, s_2) = \sum_{s_0 \in S_0} h(s_0, s_1, s_2)\,.$$

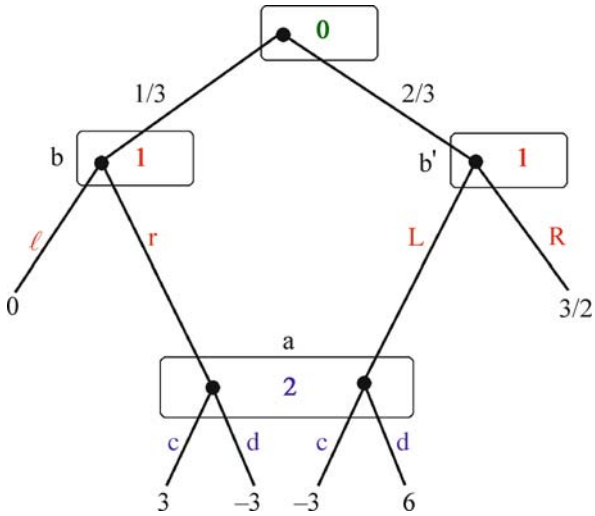

Zero-Sum Two Person Games, Figure 4

Unlike the mixed strategies we have more constraints on the sequences r_1, r_2 for each player given by the linear constraints above. It may be convenient to denote the sequence functions r_1, r_2 by vectors x, y respectively. The vector x has $|S_1|$ coordinates and vector y has $|S_2|$ coordinates. The constraints on x and y are linear given by $Ex = e, Fy = f$ where the first row is the unit vector $(1, 0, \ldots, 0)$ of appropriate size in both E and F. If $\mathcal{U}_1$ is the collection of information sets for player I then the number of rows in E is $1 + |\mathcal{U}_1|$. Similarly the number of rows in F is $1 + |\mathcal{U}_2|$. Except for the first row, each row has the starting entry as -1 and some 1's and 0's. Consider the following extensive game with perfect recall.

The set of sequences for player I is given by $S_1 = \{\emptyset, l, r, L, R\}$. The set of sequences for player II is given by $S_2 = \{\emptyset, c, d\}$. The sequence form payoff matrix is given by

$$K(s_1, s_2) = A = \begin{bmatrix} * & * & * \\ \mathbf{0} & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -2 & 4 \\ 1 & 0 & 0 \end{bmatrix}.$$

The constraint matrices E and F are given by

$$E = \begin{bmatrix} 1 & & & & \\ -1 & 1 & 1 & & \\ -1 & & & 1 & 1 \end{bmatrix}, \quad F = \begin{bmatrix} 1 & & \\ -1 & 1 & 1 \end{bmatrix}.$$

Since no end vertex corresponds $s_1 = \emptyset$, for Player I, the first row of A is identically 0 and so it is represented by $*$ entries.

We are looking for a pair of vectors x^*, y^* such that y^* is the best reply vector y which minimizes $(x^* \; Ay)$ among all vectors satisfying $Fy = f, y \geq 0$. Similarly the best reply vector is x^* where it maximizes (x, Ay^*) subject to $E^{\mathrm{T}}x = e, x \geq 0$. The duals to these two linear programming problems are

$$\begin{array}{ll} \max\,(f \quad q) & \min\,(e \quad p) \\ \text{such that} & \text{such that} \\ F^{\mathrm{T}}q \leq A^{\mathrm{T}}x^*\,. & Ep \geq Ay^* \\ q \text{ unrestricted.} & p \text{ unrestricted.} \end{array}$$

Since $E^{\mathrm{T}}x^* = e$, and $x^* \geq 0, Fy^* = f$, and $y^* \geq 0$ these two problems can as well be viewed as the dual linear programs

Primal: $\max(f \quad q)$
such that

$$\begin{bmatrix} F^{\mathrm{T}} & -A^{\mathrm{T}} \\ 0 & -E^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} q \\ x \end{bmatrix} \begin{bmatrix} \leq & 0 \\ = & e \end{bmatrix}, \qquad x \geq 0, \quad q \text{ unrestricted.}$$

Dual: $\min(e \quad p)$
such that

$$\begin{bmatrix} F & 0 \\ -A & E \end{bmatrix} \begin{bmatrix} y \\ p \end{bmatrix} \begin{bmatrix} = & f \\ \geq & 0 \end{bmatrix}, \qquad y \geq 0\,, \quad p \text{ unrestricted.}$$

We can essentially prove that:

Theorem 24 *The optimal behavior strategies of a zero sum two person game with perfect recall can be reduced to solving for optimal solutions of dual linear programs induced by its sequence form. The linear program has a size which in its sparse representation is linear in the size of the game tree.*

General Minimax Theorems

The minimax theorem of von Neumann can be generalized if we notice the possible limitations for extensions.

Example 25 Players I and II choose secretly positive integers i, j respectively. The payoff matrix is given by

$$a_{ij} = \begin{cases} 1 & \text{if } i > j \\ -1 & \text{if } i < j \end{cases}$$

the value of the game does not exist.

The boundedness of the payoff is essential and the following extension holds.

S-games

Given a closed bounded set $S \subseteq \mathbf{R}^m$, let Players I and II play the following game. Player II secretly selects a point $s = (s_1, \ldots, s_m) \in S$. Knowing the set S but not knowing the point chosen by Player II, Player I selects a coordinate $i \in \{1, 2, \ldots, m\}$. Player I receives from Player II an amount s_i.

Theorem 26 *Given that S is a compact subset of $\mathbf{R}^m$, every S-game has a value and the two players have optimal mixed strategies which use at most m pure strategies. If the set S is also convex, Player II has an optimal pure strategy.*

Proof Let T be the convex hull of S. Here T is also compact. Let

$$v = \min_{t \in T} \max_i t_i = \max_i t_i^*\,. \qquad \square$$

The compact convex set T and the open convex set $G = \{s\colon \max_i s_i < v\}$ are disjoint. By the weak separation theorem for convex sets there exists a $\xi \neq 0$ and constant c such that

$$\begin{array}{ll} \text{for all } s \in G\,, & (\xi\,, s) \leq c \quad \text{and} \\ \text{for all } t \in T\,, & (\xi\,, t) \geq c\,. \end{array}$$

Using the property that $\mathbf{v} = (v, v, \ldots, v) \in \bar{G}$ and $t^* \in T \cap \bar{G}$, we have $(\xi\,, t^*) = c$. For any $u \geq 0, t^* - u \in \bar{G}$. Thus $(\xi\,, t^* - u) \leq c$. That is $(\xi, u) \geq 0$. We can assume ξ is a probability vector, in which case

$$(\xi\,, \mathbf{v}) = v \leq c = (\xi\,, t^*) \leq \max_i t_i^* = v\,.$$

Now Player II has $t^* = \sum_j \mu_j x^j$ a mixed strategy which chooses x^j with probability μ_j. Since t^* is a boundary point of $T = \operatorname{con} S$, by the Caratheodary theorem for convex hulls, the convex combination above involves at most m points of S. Hence the theorem. See [50].

Geometric Consequences

Many geometric theorems can be derived by using the minimax theorem for S-games. Here we give as an example the following theorem of Berge [6] that follows from the theorem on S-games.

Theorem 27 *Let S_i, $i = 1 \ldots, m$ be compact convex sets in $\mathbf{R}^n$. Let them satisfy the following two conditions.*

1. *$S = \bigcup_{i=1}^m S_i$ is convex.*
2. *$\bigcap_{i \neq j} S_i \neq \emptyset$, $j = 1, 2 \ldots, m$.*

Then $\bigcap_{i=1}^m S_i \neq \emptyset$.

Proof Suppose Player I secretly chooses one of the sets S_i and Player II secretly chooses a point $x \in S$. Let the payoff to I be $d(S_i, x)$ where d is the distance of the point x from the set S_i. By our S-game arguments we have for some probability vector ξ, and mixed strategy $\mu = (\mu_1, \ldots, \mu_m)$

$$\sum_i \xi_i d(S_i, x) \geq v \quad \text{for all } x \in S \tag{20}$$

$$\sum_j \mu_j d(S_i, x^j) \leq v \quad \text{for all } i = 1 \ldots, m\,. \tag{21}$$

Since $d(S_i, x)$ are convex functions, the second inequality (1) implies

$$d\left(S_i, \sum_j \mu_j x^j\right) \leq v\,. \tag{22}$$

□

The game admits a pure optimal $x^\circ = \sum_j \mu_j x^j$ for Player II. We are also given

$$\bigcap_{i \neq j} S_i \neq \emptyset\,, \quad j = 1, 2 \ldots, m\,.$$

For any optimal mixed strategy $\xi = (\xi_1, \xi_2, \xi_m)$ of Player I, if any $\xi_i = 0$ and we choose an $x^* \in \bigcap_{i \neq 1} S_i$, then from (20) we have $0 \geq v$ and thus $v = 0$. When the value v is zero, the third inequality (22) shows that $x^\circ \in \bigcap_i S_i$. If $\xi > 0$, the second inequality (21) will become an equality for all i and we have $d(S_i,\ x^\circ) = v$. But for $x^\circ \in S$, we have $v = 0$ and $x^\circ \in \bigcap_{i=1}^m S_i$. (See Raghavan [53] for other applications.)

Ky Fan–Sion Minimax Theorems

General minimax theorems are concerned with the following problem: Given two arbitrary sets X, Y and a real function $K\colon X \times Y \to \mathbf{R}$, under what conditions on K, X, Y can one assert

$$\sup_{x \in X} \inf_{y \in Y} K(x, y) = \inf_{y \in Y} \sup_{x \in X} K(x, y)\,.$$

A standard technique for proving general minimax theorems is to reduce the problem to the minimax theorem for matrix games. Such a reduction is often possible with some form of compactness of the space X or Y and a suitable continuity and convexity or quasi-convexity of the kernel K.

Definition 28 A function $f\colon X \to R$ is upper-semi-continuous on X if and only if for any real c, $\{x\colon f(x) < c\}$ is open in X. A function $f\colon X \to R$ is lower semi-continuous in X if and only if for any real c, $\{x\colon f(x) > c\}$ is open in X.

Definition 29 Let X be a convex subset of a topological vector space. A function $f\colon X \to \mathbf{R}$ is quasi-convex if and only if for each real c, the set $\{x\colon f(x) < c\}$ is convex. A function g is quasi-concave if and only if $-g$ is quasi-convex. Clearly any convex function (concave function) is quasi-convex (conceive).

The following minimax theorem is a special case of more general minimax theorems due to Ky Fan [21] and Sion [61].

Theorem 30 *Let X, Y be compact convex subsets of linear topological spaces. Let $K\colon X \times Y \to \mathbf{R}$ be upper semi continuous (u.s.c) in x (for each fixed y) and lower semi continuous (l.s.c) in y (for each x). Let $K(x, y)$ be quasi-concave in x and quasi-convex in y. Then*

$$\max_{x \in X} \min_{y \in Y} K(x, y) = \min_{y \in Y} \max_{x \in X} K(x, y)\,.$$

Proof The compactness of spaces and the *u.s.c*, *l.s.c* conditions guarantee the existence of $\max_{x \in X} \min_{y \in Y} K(x, y)$ and $\min_{y \in Y} \max_{x \in X} K(x, y)$. We always have

$$\max_{x \in X} \min_{y \in Y} K(x, y) \leq \min_{y \in Y} \max_{x \in X} K(x, y)\,.$$

□

If possible let $\max_x \min_y K(x,y) < c < \min_y \max_x K(x,y)$. Let $A_x = \{y\colon K(x, y) < c\}$ and $B_y = \{x\colon K(x, y) > c\}$. Therefore we have finite subsets $X_1 \subseteq X$, $Y_1 \subseteq Y$ such that for each $y \in Y$ and hence for each $y \in \text{Con } Y_1$, there is an $x \in X_1$ with $K(x, y) > c$ and for each $x \in X$ and hence for each $x \in \text{Con } X_1$, there is a $y \in Y_1$, with $K(x, y) < c$. Without loss of generality let the finite sets X_1, Y_1 be with *minimum* cardinality m and n satisfying the above conditions. The minimum cardinality conditions have the following implications. The sets $S_i = \{y\colon K(x_i, y) \leq c\} \bigcap \text{Con } Y_1$ are non-empty and convex.

Further $\bigcap_{i \neq j} S_i \neq \emptyset$ for all $j = 1, \ldots, n$, but $\bigcap_{i=1}^n S_i = \emptyset$. Now by Berge's theorem (Subsect. "Geometric Consequences"), the union of the sets S_i cannot be convex. Therefore there exists $y_0 \in \text{Con } Y_1$, with $K(x, y_0) > c$ for all $x \in X_1$. Since $K(., y_0)$ is quasi-concave we have $K(x, y_0) > c$ for all $x \in \text{Con } X_1$. Similarly there exists an $x_0 \in \text{Con } X_1$ such that $K(x_0, y) < c$ for all $y \in Y_1$ and hence for all $y \in \text{Con } Y_1$ (by quasi-convexity of $K(x_0, y)$). Hence $c < K(x_0, y_0) < c$, and we have a contradiction.

When the sets X, Y are mixed strategies (probability measures) for suitable Borel spaces one can find value for some games with optimal strategies for one player but not for both. See [50], Alpern and Gal [1988].

Applications of Infinite Games

S-games and Discriminant Analysis

Motivated by Fisher's enquiries [25] into the problem of classifying a randomly observed human skull into one of several known populations, discriminant analysis has exploded into a major statistical tool with applications to diverse problems in business, social sciences and biological sciences [33,54].

Example 31 A population Π has a probability density which is either f_1 or f_2 where f_i is multivariate normal with mean vector μ_i, $i = 1, 2$ and variance covariance matrix, Σ, the same for both f_1 and f_2. Given an observation X from population Π the problem is to classify the observation into the proper population with density f_1 or f_2. The costs of misclassifications are $c(1/2) > 0$ and $c(2/1) > 0$ where $c(i/j)$ is the cost of misclassifying an observation from Π_j to Π_i. The aim of the statistician is to find a suitable decision procedure that minimizes the worst risk possible.

This can be treated as an S-game where the pure strategy for Player I (nature) is the secret choice of the population and a pure strategy for the statistician (Player II) is to partition the sample space into two disjoint sets (T_1, T_2) such that observations falling in T_1 are classified as from Π_1 and observations falling in T_2 are classified as from Π_2. The payoffs to Player I (Nature) when the observation is chosen from Π_1, Π_2 is given by the risks (expected costs):

$$r(1, (T_1, T_2)) = c(2/1) \int_{T_2} f_1(x)\mathrm{d}x,$$
$$r(2, (T_1, T_2)) = c(1/2) \int_{T_1} f_2(x)\mathrm{d}x\,.$$

The following theorem, based on an extension of Lyapunov's theorem for non-atomic vector measures [43], is due to Blackwell [10], Dvoretsky-Wald and Wolfowitz [20].

Theorem 32 *Let*

$$S = \left\{(s_1, s_2)\colon s_1 = c(2/1) \int_{T_2} f_1(x)dx\,,\ s_2 = c(1/2) \int_{T_1} f_2(x)dx\,,\ (T_1, T_2) \in \mathcal{T}\right\}$$

where $\mathcal{T}$ is the collection of all measurable partitions of the sample space. Then S is a closed bounded convex set.

We know from the theorem on S games (Theorem 22) that Player II (the statistician has a minimax strategy which is pure. If v is the value of the game and if (ξ_1^*, ξ_2^*) is an optimal strategy for Player I then we have:

$$\xi_1^*.c(2/1) \int_{T_2} f_1(x)\mathrm{d}x + \xi_2^*.c(1/2) \int_{T_1} f_2(x)\mathrm{d}x \geq v$$

for all measurable partitions $\mathcal{T}$. For any general partition $\mathcal{T}$, the above expected payoff to I simplifies to:

$$\xi_1^* c(2/1) + \int_{T_1} [\xi_2^*.c(1/2) f_2(x) - \xi_1^* c(2/1)] f_1(x)\mathrm{d}x\,.$$

It is minimized whenever the integrand is ≤ 0 on T_1. Thus the optimal pure strategy (T_1^*, T_2^*) satisfies:

$$T_1^* = \left\{x\colon \xi_2^*\, c(1/2) f_2(x) - \xi_1^* c(2/1) f_1(x) \leq 0\right\}$$

This is equivalent to

$$T_1^* = \left\{x\colon U(x) = (\mu_1 - \mu_2)^{\mathrm{T}} \sum^{-1} x - \frac{1}{2}(\mu_1 - \mu_2)^{\mathrm{T}} \sum^{-1} (\mu_1 + \mu_2) \geq k\right\}$$

and

$$T_2^* = \left\{U(x) = (\mu_1 - \mu_2)^{\mathrm{T}} \sum^{-1} x - \frac{1}{2}(\mu_1 - \mu_2)^{\mathrm{T}} \sum^{-1} (\mu_1 + \mu_2) < k\right\}$$

for some suitable k. Let $\alpha = (\mu_1 - \mu_2)^{\mathrm{T}} \sum^{-1} (\mu_1 - \mu_2)$.

The random variable U is univariate normal with mean $\frac{\alpha}{2}$, and variance α if $x \in \Pi_1$. The random variable U has mean $\frac{-\alpha}{2}$ and variance α if $x \in \Pi_2$. The minimax strategy for the statistician will be such that

$$c(2/1) \int_{-\infty}^{\frac{1}{\sqrt{\alpha}}(k - \frac{\alpha}{2})} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{y^2}{2}} \mathrm{d}y = c(1/2) \int_{\frac{1}{\sqrt{\alpha}}(k + \frac{\alpha}{2})}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{y^2}{2}} \mathrm{d}y\,.$$

The value of k can determined by trial and error.

General Minimax Theorem and Statistical Estimation

Example 33 A coin falls heads with unknown probability θ. Not knowing the true value of θ a statistician wants to estimate θ based ona single toss of the coin with squared error loss.

Of course, the outcome is either heads or tails. If heads he can estimate $\hat{\theta}$ as x and if the outcome is tails, he can estimate $\hat{\theta}$ as y. To keep the problem zero-sum, the statistician pays a penalty $(\theta - x)^2$ when he proposes x as the estimate and pays a penalty $(\theta - y)^2$ when he proposes y as the estimate. Thus the *expected loss or risk* to the statistician is given by

$$\theta(\theta - x)^2 + (1 - \theta)(\theta - y)^2 .$$

It was Abraham Wald [68] who pioneered this game theoretic approach. The problem of estimation is one of discovering the true state of nature based on partial knowledge about nature revealed by experiments. While nature reveals in part, the aim of nature is to conceal its secrets. The statistician has to make his decisions by using observed information about nature. We can think of this as an ordinary zero sum two person game where the pure strategy for nature is to choose any $\theta \in [0,\ 1]$ and a pure strategy for Player II (statistician) is any point in the unit square $I = \{(x, y)\colon 0 \le x \le 1, 0 \le y \le 1\}$ with the payoff given above. Expanding the payoff function we get

$$K(\theta, (x, y)) = \theta^2(-2x + 1 + 2y) + \theta(x^2 - 2y - y^2) + y^2.$$

The statistician may try to choose his strategy in such a way that no matter what θ is chosen by nature, it has no effect on his penalty given the choice he makes for x, y. We call such a strategy an *equalizer*. We have an equalizer strategy if we can make the above payoff independent of θ. In fact we have used this trick earlier while dealing with behavior strategies. For example by choosing x and y such that the coefficient of θ^2 and θ are zero we can make the payoff expression independent of θ. We find that $x = \frac{3}{4}, y = \frac{1}{4}$ is the solution. While this may guarantee the statistician a constant risk of $\frac{1}{16}$, one may wonder whether this is the best? We will show that it is best by finding a mixed strategy for mother nature that guarantees an expected payoff of $\frac{1}{16}$ no matter what (x, y) is proposed by the statistician. Let $F(\theta)$ be the cumulative distribution function that is optimal for mother nature. Integrating the above payoff with respect to F we get

$$K(F, (x, y) = (-2x + 1 + 2y) \int_0^1 \theta^2 \mathrm{d}F(\theta) + (x^2 - 2y - y^2) \int_0^1 \theta \mathrm{d}F(\theta) + y^2 .$$

Let

$$m_1 = \int_0^1 \theta \mathrm{d}F(\theta) \quad \text{and} \quad m_2 = \int_0^1 \theta^2 \mathrm{d}F(\theta) .$$

In terms of m_1, m_2, x, y the expected payoff can be rewritten as

$$L((m_1, m_2), (x, y)) = m_2(-2x + 1 + 2y) + m_1(x^2 - 2y - y^2) + y^2 .$$

For fixed values of m_2, m_1 the minimum value of $m_2(-2x + 1 + 2y) + m_1(x^2 - 2y - y^2) + y^2$ must satisfy the first order conditions

$$\frac{m_2}{m_1} = x^* , \text{ and } \quad \frac{m_2 - m_1}{m_1 - 1} = y * .$$

If $m_1 = 1/2$ and $m_2 = 3/8$ then $x* = 3/4$ and $y* = 1/4$ is optimal. In fact, a simple probability distribution can be found by choosing the point $1/2 - \alpha$ with probability $1/2$ and $1/2 + \alpha$ with probability $1/2$. Such a distribution will have mean $m_1 = 1/2$ and second moment $m_2 = (1/2\,(1/2 - \alpha)^2 + 1/2\,(1/2 - \alpha)^2) = 1/4 + \alpha^2$. When $m_2 = 3/8$, we get $\alpha^2 = 1/8$. Thus the optimal strategy for nature also called the *least favorable distribution*, is to toss an ordinary unbiased coin and if it is heads, select a coin which falls heads with probability $(\sqrt{2} - 1)/(2\sqrt{2})$ and if the ordinary coin toss is tails, select a coin which falls heads with probability $(\sqrt{2} + 1)/(2\sqrt{2})$. For further applications of zero-sum two person games to statistical decision theory see Ferguson [23].

In general it is difficult to characterize the nature of optimal mixed strategies even for a C^∞ payoff function. See [37]. Another rich source for infinite games do appear in many simplified parlor games. We give a simplified poker model due to Borel (See the reference under Ville [65]).

Borel's Poker Model

Two risk neutral players I and II initiate a game by adding an ante of \$1 each to a pot. Players I and II are dealt a *hand*, a random value u of U and a random value v of V respectively. Here U, V are independent random variables uniformly distributed on $[0, 1]$.

The game begins with Player I. After seeing his hand he either *folds* losing the pot to Player II, or *raises* by adding \$1 to the pot. When Player I raises, Player II after seeing his hand can either *fold*, losing the pot to Player I, or *call* by adding \$1 to the pot. If Player II calls, the game ends and the player with the better hand wins the entire pot.

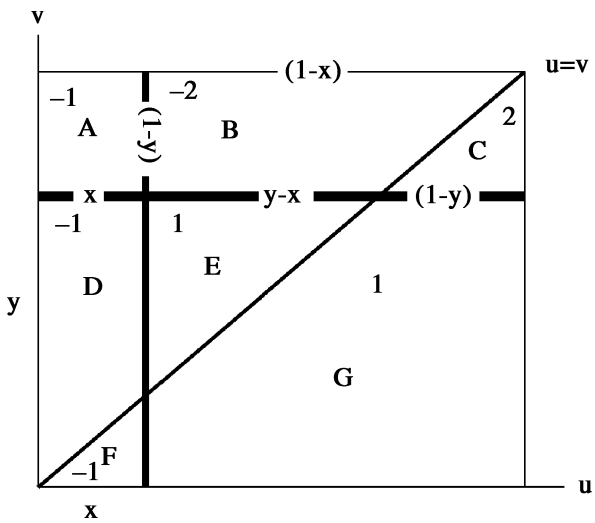

Zero-Sum Two Person Games, Figure 5
Poker partition when $x < y$

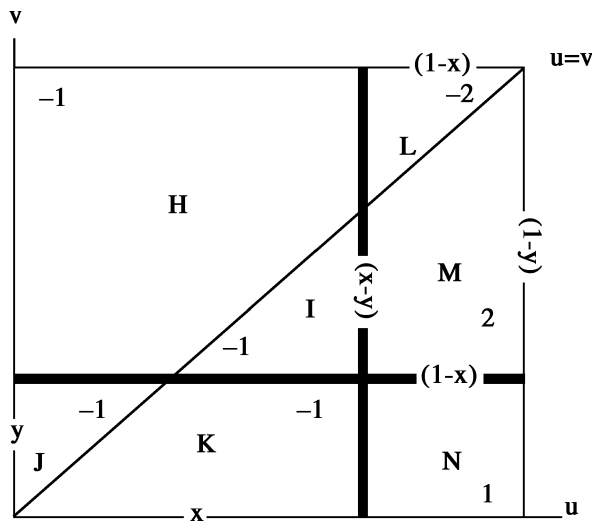

Zero-Sum Two Person Games, Figure 6
Poker partition when $x > y$

Suppose players I and II restrict themselves to using only strategies g, h respectively where

$$\text{Player I:} \quad g(u) = \begin{cases} \text{fold if } u \le x \\ \text{raise if } u > x\,, \end{cases}$$

$$\text{Player II:} \quad h(v) = \begin{cases} \text{fold if } v \le y \\ \text{raise if } v > y\,. \end{cases}$$

The computational details of the expected payoff $K(x, y)$ based on the above partitions of the u, v space is given below in Table 3.

The payoff $K(x, y)$ is simply the sum of each area times the local payoff given by

$$K(x, y) = 2y^2 - 3xy + x - y \quad \text{when } x < y\,.$$

The payoff $K(x, y)$ is simply the sum of each area times the local payoff given by

$$K(x, y) = \begin{cases} -2x^2 + xy + x - y & \text{for } x > y \\ 2y^2 - 3xy + x - y & \text{for } x < y\,. \end{cases}$$

Also $K(x, y)$ is continuous on the diagonal and concave in x and convex in y. By the general minimax theorem of Ky Fan or Sion there exist x^*, y^* pure optimal strategies for $K(x, y)$. We will find the value of the game by explicitly computing $\min_y K(x, y)$ for each x and then taking the maximum over x.

Let $0 < \bar{x} < 1$. (Intuitively we see that in our search for saddle point, $x = 0$ or $x = 1$ are quite unlikely.) $\min_{0 \le y \le 1} K(\bar{x}, y) = \min\{\inf_{y<\bar{x}} K(\bar{x}, y), \min_{y \ge \bar{x}} K(\bar{x}, y)\}$. Observe that $\inf_{y<\bar{x}} K(\bar{x}, y) = \min(-2\bar{x}^2 + \bar{x}y + \bar{x} - y) = y(\bar{x} - 1)$ + terms not involving y.

Clearly the infimum is attained at $y = \bar{x}$ giving a value $-\bar{x}^2$. Now for $y > \bar{x}$ we have $K(\bar{x}, y) = 2y^2 - 3\bar{x}y + \bar{x} - y$. This being a convex function in y, it has its minimum either at the boundary or at an interior point of the interval $[\bar{x}, 1]$. We have $K(\bar{x}, \bar{x}) = -\bar{x}^2, K(\bar{x}, 1) = 1 - 2\bar{x}$ or the minimum is attained at $y = (3\bar{x} + 1)/4$, the unique solution to $(\partial K(\bar{x}, y))/(\partial y) = 0$. At $y = (3\bar{x} + 1)/4$, the value of $K(\bar{x}, y) = (-9\bar{x}^2 + 2\bar{x} - 1)/8$. Clearly for $\bar{x} \ne 1$ we have $(-9\bar{x}^2 + 2\bar{x} - 1)/8 < -\bar{x}^2 < 1 - 2\bar{x}$. Thus $\min_y K(x, y) = (-9x^2 + 2x - 1)/8$. Now $\max \min_y K(x, y) = \max_x(-9x^2 + 2x - 1)/8 = -1/9$. It is attained at $x^* = 1/9$. Now $\min_y K(1/9, y)$ is attained at $y^* = 1/3$.

Thus the value of the game is $-1/9$. A good pure strategy for I is to raise when $x > 1/9$ and a good pure strategy for II is to call when $y > 1/3$.

For other poker models see von Neumann and Morgenstern [67], Blackwell and Bellman [5], Binmore [9] and Ferguson and Ferguson [22] who discuss other discrete and continuous poker models.

War Duels and Discontinuous Payoffs on the Unit Square

While general minimax theorems make some stringent continuity assumptions, rarely they are satisfied in modeling many war duels as games on the unit square. The payoffs are often discontinuous along the diagonal. The following is an example of this kind.

Zero-Sum Two Person Games, Table 3

Outcome	Action by players:	Region	Area	Payoff
$u \le x$	I drops out	$A \cup D \cup F$	x	-1
$u > x,\ v \le y$	II drops out	$E \cup G$	$y(1-x)$	1
$u < v,\ u > x,\ v > y$	both raise	B	$\frac{(1-y)(1-2x+y)}{2}$	-2
$u > v, u > x, v > y$	both raise	C	$\frac{1}{2}(1-y)^2$	2

Zero-Sum Two Person Games, Table 4

Outcome	Action by players:	Region	Area	Payoff
$u \le x$	I drops out	$H \cup I \cup J \cup K$	x	-1
$u > x,\ v \le y$	II drops out	N	$y(1-x)$	1
$u < v,\ u > x,\ v > y$	both raise	L	$\frac{(1-x)^2}{2}$	-2
$u > v, u > x, v > y$	both raise	M	$\frac{1}{2}(1-x)(1+x-2y)$	2

Example 34 Players I and II start at a distance $= 1$ from a balloon, and walk toward the balloon at the same speed. Each player carries a noisy gun which has been loaded with just one bullet. Player I's accuracy at x is $p(x)$ and Player II's accuracy at x is $q(x)$ where x is the distance traveled from the starting point. Because the guns can be heard, each player knows whether or not the other player has fired. The player who fires and hits the balloon first wins the game.

Some natural assumptions are

- The functions p, q are continuous and strictly increasing
- $p(0) = q(0) = 0$ and $p(1) = q(1) = 1$.

If a player fires and misses then the opponent can wait until his accuracy $= 1$ and then fire. Player I decides to shoot after traveling distance x provided the opponent has not yet used his bullet. Player II decides to shoot after traveling distance y provided the opponent has not yet used his bullet. The following payoff reflects the outcome of this strategy choice.

$$K(x,y) = \begin{cases} (1)p(x) + (-1)(1-p(x) = 2p(x) - 1 & \text{when } x < y \\ p(x) - q(x) & \text{when } x = y\,, \\ (-1)q(y) + (1-q(y)(1) \ = 1 - 2q(y) & \text{when } x > y\,. \end{cases}$$

We claim that the players have optimal pure strategies and there is a saddle point. Consider $\min_y K(x,y) = \min\{2p(x)-1, p(x)-q(x), \inf_{y<x}(1-2q(y))\}$. We can replace $\inf_{y<x}(1-2q(y))$ by $1-2q(x)$. Thus we get

$$\min_{0\le y\le 1} K(x,y) = \min\{2p(x)-1, p(x)-q(x), (1-2q(x))\}\,.$$

Since the middle function $p(x)-q(x)$ is the average of the two functions $2p(x)-1$ and $1-2q(x)$, the minimum of the three functions is simply the $\min\{2p(x)-1, (1-2q(x))\}$. While the first function is increasing in x the second one is decreasing in x. Thus the $\max_x \min_y K(x,y) =$ the solution to the equation $2p(x) - 1 = 1 - 2q(x)$. There is a unique solution to the equation $p(x) + q(x) = 1$ as both functions are strictly increasing. Let x^* be that solution point. We get $p(x^*) + q(x^*) = 1$. The value v satisfies

$$\begin{aligned} v &= p(x^*) - q(x^*) \\ &= \min_y \max_x K(x,y) \\ &= \max_x \min_y K(x,y)\,. \end{aligned}$$

Remark 35 In the above example the winner's payoff is 1 whether player I wins or player II wins. This is crucial for the existence of optimal pure strategies for the two players. Suppose the payoff has the following structure:

α	if I wins
β	if II wins
γ	if neither wins
0	if both shoot accurately and ends in a draw.

Depending on these values while the value exists only one player may have optimal pure strategy while the other may have only an *epsilon* optimal pure strategy, close to but not the same as the solution to the equation $p(x^*) + q(x^*) = 1$.

Example 36 (silent duel) Players I and II have the same accuracy $p(x) = q(x) = x$. However, in this duel, the players are both deaf so they do not know whether or not the

opponent has fired. The payoff function is given by

$$K(x,y) = \begin{cases} (1)x + (-1)(y)(1-x) = x - y + xy & \text{when } x < y\,, \\ 0 & \text{when } x = y, \\ (-1)y + (1)(1-y)x = x - y - xy & \text{when } x > y\,. \end{cases}$$

This game has no saddle point. In this case, the value of the game if it exists must be zero. One can directly verify that the density

$$f(t) = \begin{cases} 0 & 0 \le t < 1/3 \\ \frac{1}{4}t^{-3} & 1/3 \le t \le 1\,. \end{cases}$$

is optimal for both players with value zero.

Remark 37 In the above game suppose exactly one player, say Player II is deaf. Treating winners symmetrically, we can represent this game with the payoff

$$K(\xi,\eta) = \begin{cases} (1)\xi - (1-\xi)\eta & \xi < \eta \\ 0 & \xi = \eta \\ (-1)\eta + (1-\eta)(1) & \xi > \eta\,. \end{cases}$$

Let $a = \sqrt{6} - 2$. Then the game has value $v = 1 - 2a = 5 - 2\sqrt{6}$. An optimal strategy for Player I is given by the density

$$f(\xi) = \begin{cases} 0 & \text{for } 0 \le \xi < a \\ \sqrt{2}a(\xi^2 + 2\xi - 1)^{-3/2} & \text{for } a \le \xi \le 1\,. \end{cases}$$

For Player II it is given by

$$g(\eta) = \begin{cases} 0 & \text{for } 0 \le \xi < a \\ 2\sqrt{2}.\frac{a}{2+a}(\eta^2 + 2\eta - 1)^{-3/2} & \sqrt{6} - 2 \le \eta < 1 \end{cases}$$

with an additional mass of $\frac{a}{2+a}$ at $\eta = 1$. The deaf player has to maintain a sizeable suspicion until the very end!

The study of war duels is intertwined with the study of positive solutions to integral equations. The theorems of Krein and Rutman [40] on positive operators and their positive eigenfunctions are central to this analysis. For further details see [19,37,51].

Dynamic versions of zero sum two person games where the players move among several games according to some Markovian transition law leads to the theory of stochastic games [60]. The study of value and the complex structure of optimal strategies of zero-sum two person stochastic games is an active area of research with applications to many dynamic models of competition [24].

While stochastic games study movement in discrete time, the parallel development of dynamic games in continuous time was initiated by Isaacs in several Rand reports culminating in his monograph on Differential games [31]. Many military problems of pursuit and evasion, attrition and attack lead to games where the trajectory of a moving object is being steered continuously by the actions of two players. Introducing several natural examples Isaacs explicitly tried to solve many of these games via some heuristic principles. The minimax version of Bellman's optimality principle lead to the so called Isaacs Bellman equations. This highly non-linear partial differential equation on the value function plays a key role in this study.

Epilogue

The rich theory of zero sum two person games that we have discussed so far hinges on the fundamental notions of value and optimal strategies. When either zero sum or two person assumption is dropped, the games cease to have such well defined notions with independent standing. In trying to extend the notion of optimal strategies and the minimax theorem for bimatrix games Nash [48] introduced the concept of an equilibrium point. Even more than this extension it is this concept which is simply the most seminal solution concept for non-cooperative game theory. This Nash equilibrium solution is extendable to any non-cooperative N person game in both extensive and normal form. Many of the local properties of optimal strategies and the proof techniques of zero sum two person games do play a significant role in understanding Nash equilibria and their structure [15,32,39,52,53].

When a non-zero sum game is played repeatedly, the players can peg their future actions on past history. This leads to a rich theory of equilibria for repeated games [62]. Some of them, like the repeated game model of Prisoner's dilemma impose tacitly, actual cooperation at equilibrium among otherwise non-cooperative players. This was first recognized by Schelling [58], and was later formalized by the so called folk theorem for repeated games (Aumann and Shapley [1986], [3]). Indeed Nash equilibrium per se is a weak solution concept is also one of the main messages of Folk theorem for repeated games. With a plethora of equilibria, it fails to have an appeal without further refinements. It was Selten who initiated the need for refining equilibria and came up with the notion of subgame perfect equilibria [59]. It turns out that subgame perfect equilibria are the natural solutions for many problems in sequential bargaining [57]. Often one searches for specific types of equilibria like symmetric equilibria, or Bayesian equilibria for games with incomplete information [30]. Auctions

as games exhibit this diversity of equilibria and Harsanyi's Bayesian equilibria turn out to be the most appropriate solution concept for this class of games [47].

Zero sum two person games and their solutions will continue to inspire researchers in all aspects non-cooperative game theory.

Acknowledgment

The author wishes to acknowledge the unknown referee's detailed comments in the revision of the first draft. More importantly he drew the author's attention to the topic of search games and other combinatorial games. The author would like to thank Ms. Patricia Collins for her assistance in her detailed editing of the first draft of this manuscript. The author owes special thanks to Mr. Ramanujan Raghavan and Dr. A.V. Lakshmi Narayanan for their help in incorporating the the graphics drawings into the latex file.

Bibliography

1. Alpern S, Gal S (2003) The theory of search games and rendezvous. Springer
2. Aumann RJ (1981) Survey of repeated games, Essays in Game Theory and Mathematical Economics, in Honor of Oscar Morgenstern. Bibliographsches Institut, Mannheim, pp 11–42
3. Axelrod R, Hamilton WD (1981) The evolution of cooperation. Science 211:1390–1396
4. Bapat RB, Raghavan TES (1997) Nonnegative Matrices and Applications. In: Encyclopedia in Mathematics. Cambridge University Press, Cambridge
5. Bellman R, Blackwell D (1949) Some two person games involving bluffing. Proc Natl Acad Sci USA 35:600–605
6. Berge C (1963) Topological Spaces. Oliver Boyd, Edinburgh
7. Berger U (2007) Brown's original fictitious play. J Econ Theory 135:572–578
8. Berlekamp ER, Conway JH, Guy RK (1982) Winning Ways for your Mathematical Plays, vols 1, 2. Academic Press, NY
9. Binmore K (1992) Fun and Game Theory A text on Game Theory. DC Heath, Lexington
10. Blackwell D (1951) On a theorem of Lyapunov. Ann Math Stat 22:112–114
11. Blackwell D (1961) Minimax and irreducible matrices. Math J Anal Appl 3:37–39
12. Blackwell D, Girshick GA (1954) Theory of Games and Statistical Decisions. Wiley, New York
13. Bouton CL (1902) Nim-a game with a complete mathematical theory. Ann Math 3(2):35–39
14. Brown GW (1951) Iterative solution of games by fictitious play. In: Koopmans TC (ed) Activity Analysis of Production and Allocation. Wiley, New York, pp 374–376
15. Bubelis V (1979) On equilibria in finite games. Int J Game Theory 8:65–79
16. Chin H, Parthasarathy T, Raghavan TES (1973) Structure of equilibria in *N*-person noncooperative games. Int Game J Theory 3:1–19
17. Conway JH (1982) On numbers and Games, Monograph 16. London Mathematical Society, London
18. Dantzig GB (1951) A proof of the equivalence of the programming problem and the game problem. In: Koopman's Actvity analysis of production and allocationation, Cowles Conumesion Monograph 13. Wiley, New York, pp 333–335
19. Dresher M (1962) Games of strategy. Prentice Hall, Englewood Cliffs
20. Dvoretzky A, Wald A, Wolfowitz J (1951) Elimination of randomization in certain statistical decision problems and zero-sum two-person games. Ann Math Stat 22:1–21
21. Fan K (1953) Minimax theorems. Proc Natl Acad Sci Wash 39:42–47
22. Ferguson C, Ferguson TS (2003) On the Borel and von Neumann poker models. Game Theory Appl 9:17–32
23. Ferguson TS (1967) Mathematical Stat, a Decision Theoretic Approach. Academic Press, New York
24. Filar JA, Vrieze OJ (1996) Competitive Markov Decision Processes. Springer, Berlin
25. Fisher RA (1936) The Use of Multiple Measurements in Taxonomic Problems. Ann Eugen 7:179–188
26. Fourier JB (1890) Second Extrait. In: Darboux GO (ed) Gauthiers Villars, Paris, pp 325–328; English Translation: Kohler DA (1973)
27. Gal S (1980) Search games. Academic Press, New York
28. Gale D (1979) The game of Hex and the Brouwer fixed-point theorem. Am Math Mon 86:818–827
29. Gale D, Kuhn HW, Tucker AW (1951) Linear Programming and the theory of games. In: Activity Analysis of Production and Allocation. Wiley, New York, pp 317–329
30. Harsanyi JC (1967) Games with incomplete information played by Bayesian players, Parts I, II, and III. Sci Manag 14:159–182; 32–334; 486–502
31. Isaacs R (1965) Differential Games: Mathematical A Theory with Applications to Warfare and Pursuit. Control and Optimization. Wiley, New York; Dover Paperback Edition, 1999
32. Jansen MJM (1981) Regularity and stability of equilibrium points of bimatrix games. Math Oper Res 6:530–550
33. Johnson RA, Wichern DW (2007) Applied Multivariate Statistical Analysis, 6th edn. Prentice Hall, New York
34. Kakutani S (1941) A generalization of Brouwer's fixed point theorem. Duke Math J 8:457–459
35. Kantorowich LV (1960) Mathematical Methods of Organizing and Planning Production. Manag Sci 7:366–422
36. Kaplansky I (1945) A contribution to von Neumann's theory of games. Ann Math 46:474–479
37. Karlin S (1959) Mathematical Methods and Theory in Games, Programming and Econs, vols 1, 2. Addison Wesley, New York
38. Kohler DA (1973) Translation of a report by Fourier on his work on linear inequalities. Opsearch 10:38–42
39. Kreps VL (1974) Bimatrix games with unique equilibrium points. Int Game J Theory 3:115–118
40. Krein MG, Rutmann MA (1950) Linear Operators Leaving invariant a cone in a Banach space. Amer Math Soc Transl 26:1–128
41. Krishna V, Sjostrom T (1998) On the convergence of fictitious play. Math Oper Res 23:479–511
42. Kuhn HW (1953) Extensive games and the problem of information. Contributions to the theory of games. Ann Math Stud 28:193–216
43. Lindenstrauss J (1966) A short proof of Liapounoff's convexity theorem. Math J Mech 15:971–972
44. Loomis IH (1946) On a theorem of von Neumann. Proc Nat Acad Sci Wash 32:213–215

45. Miyazawa K (1961) On the convergence of the learning process in a 2 × 2 non-zero-sum two-person game, Econometric Research Program, Research Memorandum No. 33. Princeton University, Princeton
46. Monderer D, Shapley LS (1996) Potential LS games. Games Econ Behav 14:124–143
47. Myerson R (1991) Game theory. Analysis of Conflict. Harvard University Press, Cambridge
48. Nash JF (1950) Equilibrium points in n-person games. Proc Natl Acad Sci Wash 88:48–49
49. Owen G (1985) Game Theory, 2nd edn. Academic Press, New York
50. Parthasarathy T, Raghavan TES (1971) Some Topics in Two Person Games. Elsevier, New York
51. Radzik T (1988) Games of timing related to distribution of resources. Optim J Theory Appl 58:443–471, 473–500
52. Raghavan TES (1970) Completely Mixed Strategies in Bimatrix Games. Lond J Math Soc 2:709–712
53. Raghavan TES (1973) Some geometric consequences of a game theoretic result. Math J Anal Appl 43:26–30
54. Rao CR (1952) Advanced Statistical Methods in Biometric Research. Wiley, New York
55. Reijnierse JH, Potters JAM (1993) Search Games with Immobile Hider. Int J Game Theory 21:385–94
56. Robinson J (1951) An iterative method of solving a game. Ann Math 54:296–301
57. Rubinstein A (1982) Perfect equilibrium in a bargaining model. Econometrica 50:97–109
58. Schelling TC (1960) The Strategy of Conflict. Harvard University Press. Cambridge, Mass
59. Selten R (1975) Reexamination of the perfectness concept for equilibrium points in extensive games. Int Game J Theory 4: 25–55
60. Shapley LS (1953) Stochastic games. Proc Natl Acad of Sci USA 39:1095–1100
61. Sion M (1958) On general minimax theorem. Pac J 8:171–176
62. Sorin S (1992) Repeated games with complete information, In: Aumann RJ, Hart S (eds) Handbook of Game Theory, vol 1, chapter 4. North Holland, Amsterdam, pp 71–103
63. von Stengel B (1996) Efficient computation of behavior strategies. Games Econ Behav 14:220–246
64. Thuijsman F, Raghavan TES (1997) Stochastic games with switching control or ARAT structure. Int Game J Theory 26: 403–408
65. Ville J (1938) Note sur la theorie generale des jeux ou intervient l'habilite des joueurs. In: Borel E, Ville J (eds) Applications aux jeux de hasard, tome IV, fascicule II of the Traite du calcul des probabilites et de ses applications, by Emile Borel
66. von Neumann J (1928) Zur Theorie der Gesellschaftspiele. Math Ann 100:295–320
67. von Neumann J, Morgenstern O (1947) Theory of Games and Economic Behavior, 2nd edn. Princeton University Press, Princeton
68. Wald A (1950) Statistical Decision Functions. Wiley, New York
69. Weyl H (1950) Elementary proof of a minimax theorem due to von Neumann. Ann Math Stud 24:19–25
70. Zermelo E (1913) Uber eine Anwendung der Mengenlehre auf die Theorie des Schachspiels. Proc Fifth Congress Mathematicians. Cambridge University Press, Cambridge, pp 501–504

List of Glossary Terms

A

Ab initio approach 689
Abduction 201
Abelian group 8028
Abrupt climate change 2
ABS 1565
Absentee owners 2284
Absolute error 5161
Absolute temperature 2859
Absolute time 9831
Absolutely continuous curves 5440
Absorbing state 3360, 8380
Abstract game 1518
Abstract game of network formation with respect to irreflexive dominance 6024
Abstract game of network formation with respect to path dominance 6024
Ac 1307
AC Electroosmosis or induced charge electroosmosis (ICEO) 5573
Acceleration mechanism 21
Accessible bonds or sites 6538
Accessible hull 6521
Accretion disk 381
Accretionary wedge (prism) 10047
Accuracy 271, 1826, 5206
Accuracy (rate) 1982
ACF 4114
Achievable mate 9656
Action angle variables 2261
Action potential 6099
Action profile 8649
Action set 8649
Action type 7651
Action-angle variables 5064
Actions 6709
Activation function 1813
Activator 1352
Active fault 2497
Active gel 1739
Active membranes 5524
Activity, sensitivity 600
Activity-based approach 9536
Actors 5999
Actuator 9749
Actuators 1475
Acute hospitals 4562
Adaptation 113
Adaptive behaviors 8364
Adaptive planning framework 8439
Adaptive system 4869
Adaptive tension 8364
Additive cellular automata 64
Additively separable preferences 9655
Adiabatic approximation 4194
Adiabatic decoupling 6689
Adiabatic perturbation theory 6689
Adiabatic switching 5859
Adjacency matrix 4820, 5266, 8815
Adjacency matrix A 6048
Adjacent 8232, 8346
Adjoint conjugate operator 4317
ADMET 2196
Administrative leadership 8364
Adomian decomposition method 5161
Adomian decomposition method, Homotopy analysis method, Homotopy perturbation method and Variational perturbation method 5144
Adomian method 81
Adomian polynomials 5161

Adopters 1956
Adoption 8306
Adsorption 1307
Adverse selection (hidden information) 6977
Affiliation network 8291
Affine CA 2981
Affinity 4776
Age of a cluster or fuzzy rule 3244
Agency trade 1565
Agent 131, 148, 184, 5286, 7533, 8364, 8405
Agent architecture 184
Agent based models 8306
Agent (or software agent) 8375
Agent-based computational economics (ACE) 201
Agent-based computational models 8265
Agent-based model 131
Agent-based modeling 5286
Agent-based modeling (ABM) 113
Agent-based models 1041
Agent-based models (ABM) 201
Agent-based simulation 92, 160
Agents 201, 848
Agglomerative (hierarchical) clustering algorithm 2348
Aggregate models 5377
Aggregate ("sufficient unit") equation modeling 4609
Aggregation 1982, 8364
Aggregation mechanisms 8364
Aggregation operators 224
Aging 4210, 7520, 7583
Agmon metric 6747
Agree-and-pursue algorithm 7713
Aharonov Bohm effect 8597
Aharonov–Bohm (AB) effect 7290
AI 7931
Air traffic flow 234
AKNS method 4960
Alexander–Orbach conjecture 310
Alfvén speed 8010
Alfvén wave 8521
Algebra 8292
Algebraic models 8265
Algorithm 7361
Algorithmic complexity 5206
Algorithmic complexity of object x 243
Algorithmic self-assembly of DNA tiles 5631
Algorithmic trading 9565
Allele 9106
Allometric laws 3779
Allowable set of partners 9654
Almost conjugacy 8889
Almost equicontinuous CA 2232, 9246
Almost every, essentially 3083
Almost everywhere 4983
Almost everywhere (abbreviated a. e.) 3040
Almost everywhere (a.e.) 2883
Alphabet of a cellular automaton 63
α-Helix 9750
α-Level set and support 4047
Alternating independent-2-paths 8319
Alternating k-stars 8319
Alternating k-triangles 8319
Alternative splicing 4161
AM1 7931
Amoeboid motility 1738
Amorphous computer 257
Amygdala 1109
Analog 9127
Analog circuit 9706
Analytical method for solving an equation 5091
Analytical theories of turbulence 3642
Ancilla qubits 7325
Andesite 9763
Anisotropic elements 1719
Anisotropy 6565
Annealed law 7520
Annotations 131
Anomalous diffusion 310, 3878, 5218
Anomalous resistivity 8521
Ant colony optimization 2113, 8869
Ant colony optimization (ACO) 113
Anthropogenic climate change 3
Anthropogenic emissions 1071
Anthropomorphic 4643
Antibody 2113, 4776
Antigen 2113, 4776
Antigen presenting cells (APC) 4776
AO 7931
Aperiodic tile set 9159
Apparent exponent 6990
Apparent stress 2581
Applicability domain 7071

Approximate controllability 4804
Approximation 7787
Approximation space 7787
Aqueous fluid 2676
AR 4643
Arbitrage 3435
Arbitrage pricing theory (APT) 3475
Arbo viruses 1095
Arbo viruses transmitted by Aedes mosquitoes 1095
Arc 8232
Archimedean lattices 6579
ARCH(q) 3456
AR(k) 3456
ARMA 4114, 5206
Arrival time 2449
Arrow 8346
Arrow's impossibility theorem 9932
Arterial 9470
Artificial chemistry 113
Artificial financial markets 3374
Artificial immune systems 2113
Artificial intelligence (AI) 4869
Artificial life 8050
Artificial life (ALife) 113
Artificial neural network 1813, 3475
Artificial neural network (ANN) 113
Artificial neuron 1813, 2139
Artisan owners 2284
Ascendancy 2697
Aseismic 1689
Ask price 5392
Aspects 131
Assignment 9470
Assortative mixing and disassortative mixing 2531
Assortativity coefficient 6049
Astronomical unit (AU) 8521
Asymmetric information 6977
Asymptotic density 4497
Asymptotic negligibility 5360
Asymptotic series 6748
Asymptotic shape 4497
Asynchronous 848
Asynchronous circuit 5859
AT property of an automorphism 8555
Atmosphere 1
Atmospheric boundary layer 8140
Atomic force microscope (AFM) 689
Atomic force microscopy (AFM) 7026
Atomic units, a. u. 8575
Atomistic simulation 5746
ATP 3598, 9749
Attendance 5588
Attenuation factor $\boldsymbol{Q}^{-1}$ 7914
Attenuation relation (ground-motion prediction equation) 4435
Attraction basin 2232
Attractor 381, 2232, 2903, 3828, 4723, 8040, 8183, 8420, 8927, 9224, 9246
Attractors 600
Attribute 1826, 1983, 7787, 8232
Attribute (*also* feature *or* variable) 5317
Attribute node 1826
Attributed-based data model (ABDM) 4305
Attributes 131
Atypical graph 6013
Aubry set 4541
Auroral electrojet index (AE) 8521
Auroral zone 8521
Auto-catalysis 4920
Autocatalytic set 113, 326
Automata 4098
Automated guided vehicle (AGV) 8438
Automated planning 5188
Automaton 9411
Automorphism 2934, 8889
Autonomous 113
Autonomous information system 1533
Autonomous robot 5783
Autonomous system (AS) 4930
Autonomy 184, 3682, 4643, 6066, 7533
Autoparametric resonance 2323
Autopoiesis 8040
Autoregressive conditional heteroskedasticity (ARCH) 3475
Autoregressive equation 9900
Autoregressive model 442
Avalanche 8028
Avalanche photodiode (APD) 7249
Avalanches 644
Average bonds per object 1444
Average class accuracy 1982
Average distance 8910

Average partial effect 2770
Averaging 8588
Averaging algorithms 7713
Avida 113
AWE 1475
Axisymmetric (concentric) KdV equation 9967
Axon 3334

B

B cell 4776
B3LYP 7931
B92 7265
Babbage engine 5859
Babbling equilibrium 8125
Backbone 1395, 6538
Back-propagation 4869
Bag 4340
Bagging 3475
Balance equations 4378
Balance theories 8265
Balanced multiwavelet 5841
Ballistic transport 7401
Banach fixed point theorem 6611
Band structure 5746, 7401
Bandwidth 9470
Baroclinic fluid 6206
Barotropic fluid 6206
Barrier options 5404
Basal ganglia 1109
Basalt 9763
"Baseline" monitoring data 9862
Basic equation of fluid dynamics 3642
Basic notation 5926
Basic reproduction rate 4659
Basin of attraction 2903, 3828
Basins of attraction 64
Basis point 3456
Bayes' theorem; prior, likelihood and posterior distributions 456
Bayesian analysis 7853
Bayesian classifier 2159
Bayesian equilibrium 427
Bayesian game 426, 1587
Bayesian learning 5177
Bayesian parametric and non-parametric modeling 456
Bayesian–Nash equilibrium 5510
BB84 7265
Beach profile 9618
Beam splitter 7249
Behavior 344, 5999, 6066
Behavior mode 8948
Behavior strategy 5177, 8125
Behavioral finance 4972
Behavioral relations 2284
Behavioral strategy 7616
Behavioral type 8125
Behavior-based control 3682
BEKK 4114
Belief learning 5177
Bell inequality 7250
Bell states 7250
Bell's inequality 1012
Belousov–Zhabotinsky (BZ) reaction 7548
Benchmarks 4643
Benjamin–Feir instability of water waves 9967
BEP 1307
Bernoulli measure 2981
Bernoulli shift 2918, 4983
Bernoulli's equation 9967
Berry phase 4194
Besicovitch pseudometrics 2232
Bessel process 8708
β-Sheet 9750
Bethe lattice 608
BGK models 866
BGP (border gateway protocol) 4930
Bibliome 5535
Bid price 5392
Bid-ask spread 5392, 5404
Bi-directional QKD, or "Plug & play" QKD 7265
Bifurcation 381, 936, 1041, 2324, 4515, 5299, 6195, 6329, 6697
Billiard ball model 7696
Binarization 1983
Binary 8690
Binary forecast 2438
Binary neighborhood system (binary granular model; binary GrC model) 4339
Binary relation 8346
Bingham liquid 9763
Biochemical oscillation 525
Biochemical reaction network 511

Biochemically, genetically and genomically (BiGG) structured reconstruction 5535
Biocomposite 1283
Bioinformatics 2697
Biological function 719
Biological swimmers 549
Biologically-inspired computational algorithm 113
Biologically-inspired optimization 1775
Biomass function 5535
Biomimetics 1283
Biomineralization 1283
Biomolecular network 570
Biomolecule 570
Bionanomachines 7026
Bipartite graph 8346
Bipartite network 6048, 8232
BIRCH 5051
Bit 1011, 7201, 7361
Bit-flip 7266
Bit-flip error 7325
BKL 7931
Black box model 7088, 7119
Block 6913
Block mapping 148
Blocking word 2232, 9246
Blockmodel 6914, 8279
Blockmodel image 6914
Blockmodeling 6914, 8265, 8346
Blow up 2039
Blue-tongue disease in Europe 1096
Bluetooth 2072
BO 7931
Bode plot (phase and gain plot) 8948
Body waves 2497
Boltzmann equation 7807
Boltzmann–Gibbs distribution, the 2801
Boltzmann–Gibbs entropy 2859
Bond percolation 1443, 6538, 6579
Bonding mechanism 5618
Book-to-market ratio 3391
Boolean algebra 7361
Boolean functions 1789
Boolean network 511
Bootstrap 3456
Bootstrap percolation 608
Bore 9618
Bore solitons 8506
Bose–Einstein condensation (BEC) 9679
Bose–Einstein particles 7201
Boson 9679
Boson peak 2827
Bottleneck 3143, 9302
Bottom-up fabrication 5859
Boundary region 7787
Bounded rationality 3374
Bounded-rationality 92
Boussinesq equation 9967
Boussinesq equations 9618
Box 3924
Box diameter 3924
Box dimension 3827
Box splines 5800, 6855
BQP complexity class 7155
Bragg glass 2019
Braitenberg vehicle 3682
Branch 4820
Branch switching 6330
Branched polymers 6545
Branching 4820
Branching process 2556
Branes, D-branes 1642
Breaking waves 8465
Breakthrough time 6565
Breit–Wigner (BW) resonances 7290
Brillouin zone 689
Broker 1565, 9565
Broker's broker 1565
Broker-dealer 1565
Brokers 5404
Brouwer degree 6611
Brouwer fixed point theorem 6611
Brownian dynamics 8730
Brownian motion 3879, 5859, 6228, 7520, 8708
B-splines 6855
BSS 271
Bubble collapse 6956
Bubbly flow 9763
Bubbly liquid 9872
Bulk material 5889
Business dynamics 665
Business policy and strategy 665
Bytecode 131

C

C++ 131
C# 131
CA 866, 7931
CACTUS 5051
Cadherins 913
Calibration 4751
Call markets 5404
Camera-in-hand configuration 42, 4751
Camera-to-hand configuration 42, 4751
Camshift algorithm 5702
Canalization 525
Canard 7603
Canonical 1-form 2261
Canonical transformation 2261
Cantor set, Cantor dust, Cantor family, Cantor stratification 2261, 4515
Capillary dominated flow 6565
Capital 2284
Capital asset pricing model (CAPM) 3391, 3475
Capital intensive 2284
Capitalist sector 2284
Capitalist system 2284
CAPM 4114
Carbó (similarity) index 7422
Carbon dioxide fertilization and quality of food 1096
Carbon nanotube 5859
Carbon nanotube field effect transistors 7480
Car-following model 9430
Carrier sense multiple access (CSMA) 2054
Cartesian theater 1109
Cascade model 2538
Cascadia 1689
Case 1791
Case-based reasoning 4283
Casimir function 2261
Castaing distribution 8521
Catalysis and auto-catalysis 3284
Catalyst's activity 1307
Catalyst's selectivity 1307
Catastrophe 703
Cauchy problem 2039
Cauchy–Schwarz inequality 4318
Causal loop diagram 7052
Causal loops 9043
Causality 4609
Cause map 1845
CC 7931
CCC 4114
CCDF 4930
CE 7931
Čech–Stone compactification of $\mathbb{N}$, $\beta\mathbb{N}$ 3040
Celerity 4839
Cell 866, 922, 1738
Cell fate 1181
Cell lineage 1181
Cell type 1181
Cells 848
Cellular automata 63, 779, 823, 901, 2980, 6804, 8050, 8421
Cellular automata (CA) 113, 9159
Cellular automata rule 63
Cellular automaton 741, 755, 768, 800, 866, 3538, 4497, 4739, 5859, 6772, 7548, 7679, 8028, 8066, 8375, 8869
Cellular automaton (CA) 892, 3233
Cellular automaton with block rules 7679
Cellular programming 3233
Cellular-computing architecture 8869
Center 2261
Center-saddle bifurcation 2261
Central configurations 5959
Central executive agent (CEA) 1109
Central limit theorem, the 3318
Centrality 4820, 8279
Centroid 2449
Centroid moment tensor solution 9593
CGCM 9014
CHAMELEON 5051
Change determination process 5999
Change opportunity model 5999
Changes in the pollen season and new pollen 1096
Channel 7343
Chaos 4178, 4515, 7873, 8028, 8521, 8948
Chaotic 476
Chaotic hypothesis 989
Chaotic system 368
Chaotic systems 989
Characteristic function form game 1518
Characteristic or coalitional function 1509
Cheap-talk game 8125
Chemical Fokker–Planck equation (CFPE) 8730

Chemical gels 6545
Chemical Langevin equation (CLE) 8730
Chemical master equation (CME) 8730
(Chemical) Organization 326
Chemisorption 1307
Chemotaxis 913
Chern class 2261
Chinese room argument 1109
ChIP-chip 3965
Chiral, achiral and racemic 3284
Choice set of $f \in F$ from $A \subseteq W(Ch_f(A))$ 9655
Choked flow 9872
Chordal SLE 8709
Choreographical solution 5959
Chromatin 3894, 3965
Chronic illness 4577
Chronological algebra 1028
Chronological calculus 1028
Church–Turing (CT) computation 271
Church–Turing thesis, the 9127
Chymotrypsin 9750
CI 7931
City 7953
City size distribution 1041
C^k Conjugacy 3096
C^k topology 4723
Class 131
Classical computer 7119
Classical computing 7201
Classical limit 7135
Classical statistical mechanics 5667
Classically allowed and forbidden regions – turning points 6748
Classical-map hyper-netted-chain (CHNC) 8576
Classification 1826, 1983
Classification accuracy 1983
Classification of objects 4413
Classification task 1790
Classifier 1826, 1983
Clausius entropy 2859
Clearing 5405
Client 4474
Climate change 2215
Climate change related direct health effects 1095
Climate models 2, 7591
Climate simulation 2
Climate time scales 1
Climate variability pattern 2
Climate variables and forcing 2
CLIQUE 5051
Clique 8232
Clonal selection theory 4776
Close binary stars 381
Closing CA 9246
Closure 326
Closure property 800
Closure relation 4317
Club 5360
Cluster 608, 2348, 5051, 6380, 6538, 6783, 7822
Cluster analysis 1226
Cluster State 7250
Clustering 2348
Clustering coefficient 2531, 8910
Clustering coefficient C 6049
CMT centroid moment tensor 2657
(c_n)-Conservative system 3083
Cnoidal waves 9967
CNOT gate 7325
CO_2 9014
Coadjoint orbit 5483
Coarse-grained dissipation 8521
Coarse-graining 8521
Coarsening 3612
Co-clustering 5051
Cocycles and group extensions 8555
Coda attenuation factor $\mathbf{Q}_C^{-1}$ 7914
Coda waves 2497, 7914
Codimension 703, 6330
Coevolution 113, 3233, 8364
Coexistence 6626
Cognition 4893
Cognitive dynamics 9411
Cognitive map 1845
Coherent light 6388
Coherent structure 8521
Coherent superposition 7219
Cointegration 2758
Cointension 3985
Collective intelligence 6476
Collective or social choice problem 9932
Collective variables 1538
Collision 866
Collision and singularities 5960
Collision-based computer 7548

Column subshift 9246
Comb model 310
Combinatorial optimization 6407
Combinators 5286
Combi-QSAR – combinatorial QSAR 7071
Commercial 2284
Commission 1565
Commodity-based economy 8364
Common intermediate language 131
Common prior and consistent beliefs 426
Common values 9933
Communication channel 7361
Community 1141, 3110
Community empowerment 3110
Community safety group 3110
Compact visitation 310
Compacton 8457
Compacton-like solution 8457
Compartments 2711
Compatible observables 4317
Competition and cooperation 8040
Complementary metal-oxide-semiconductor (CMOS) 5859
Complete 8232
Complete graph 8380
Complete information 7651
Complete path planner 5783
Completely revealing strategy 7631
Complex adaptive system (CAS) 113
Complex and evolving structural systems 2421
Complex developmental system 1873
Complex disease 9090
Complex dynamics 7583
Complex (MHC) 4776
Complex networks 234
Complex phenotype 9106
Complex project 1845
Complex system 113, 3754, 7009
Complex systems 1268
Complex trait 3965
Complexity 149, 201, 3267, 3300, 4178, 8040, 9211
Complexity (descriptor) index 9211
Complexity in space 3574
Complexity in time 3574
Complexity measures 4820, 7713
Complexity mechanism 8364
Complexity organizational knowledge 8364
Complexity theory 8364
Component 6495, 8232
Composite 1283
Compositional strand asymmetry 3894
Compositionality 5286
Computable 5467
Computation 922
Computational algebra systems 131
Computational complexity 7089, 7119
Computational fluid dynamics 8153
Computational mathematics systems 131
Computational particle 257
Computational universality 901, 5859
Computer 7361
Computer architecture 5859
Computer generated imagery (CGI) 1335
Computer language 131
Computer programming language 131
Computing 922, 9127
Computing agent 7533
Concentration indices 4609
Condensation 8246
Condition D1 8125
Conditional measure 4983
Conditionally periodic 2262
Condorcet jury theorem 9933
Condorcet paradox 9933
Condorcet winner 9943
Conductance exponent (t) 3737
Conductance (G) 3737
Conductance quantization 7663
Conduit 9763, 9785
Confidence 5206
Confidence ellipse 4250
Configuration 5618, 7154, 8066
Configuration recognition 5618
Configuration space 8380
Configuration space and the shift 2980
Confining potential 8576
Conflicting constraints 8364
Conformal invariance 8709
Conformal transformation 8709
Confusion matrix 1983
Congested traffic 9302, 9355
Conjugacy 3096
Conjugate co-ordinates, canonical co-ordinates 2262

Connectance(C) 3661
Connected 8346
Connected graph 4820
Connectionist models 1109
Connective stability 2053
Connectivity 6565, 6833
Connectivity correlation 2711
Connectivity distribution or degree distribution 2531
Connectivity matrix 599, 5051
Connectivity-k (k-connected, k-cohesive, structural co- hesion, cohesive.blocks) 4608
Conservation law 866, 4713
Conservative, dissipative 8168
Conservative system 3083
Conservativeness 3056
Consistent estimator 6991
Constitutive equations 4378
Constraint 2827
Constraint-based analysis 511
Constraint-based reconstruction and analysis (COBRA) 5535
Construction 8067
Construction universality 5859
Constructive system 8040
Consumer behavior 5377
Consumer-resource interactions 3661
Continuation 6330, 6587
Continuation game 4098
Continuity equation 866, 9967
Continuous and discrete dynamical systems 936
Continuous phase transitions 7873
Continuous problem 7119
Continuous simulation 4562
Continuous space machine (CSM) 6388
Continuous stochastic process 6228
Continuous time random walk 5218
Continuous two-sided matching model 9654
Continuum 1475
Continuum forecast 2438
Continuum hypothesis 9967
Continuum limit 8709
Continuum media approximationj 8588
Continuum percolation 1444, 6565
Control 1475, 6066
Control function 1486
Control parameter 3597, 5718, 8926
Control parameters 1538
Control parameter(s) 3612
Control system 3524
Control systems and reachable sets 5440
Controllability 1027, 3524
Controllability to trajectories 4804
Controller 9470
Controlling feedback 2284
Convex 7520
Convex space 5536
Convolution 2657
Cooper pairs 9679
Cooperative control 7713
Cooperative game 1509
Coordination 9471
Coordination dynamics 1537, 8199
Copying-fidelity 3206
Core 1509, 1660, 5359
Core group 4659
Core set 7422
Corner frequency 2474, 2581
Corona 9268
Corporate 2284
Corporate (or commercial) customers 5405
Corrections to scaling 6991
Correlated equilibrium 427
Correlated equilibrium, a 8698
Correlated random effects 2770
Correlation 1616
Correlation dimension 3828
Correlation energy 8576
Correlation function 1268, 1616, 3360, 3779, 8521
Correlation hole 8576
Correlation length 1395, 6538, 8709
Correlation time 6991
Correlations 1268
Cortex 1109
Coseismic 1689
Cosmic evolution 3267
Cosmic microwave background 1634
Cosmic strings 1642
Cosmology 3267
Cost function 1660
Cost of an algorithm 7119
Cost of voting 9932
Cost sharing problem 1660
Cost sharing rule 1660

Cotangent bundle 5483
Couette flow 4997
Coulomb blockade 5859, 7309
Coulomb plasticity 10047
Counterpropagation neural network 2139
Coupled map lattice 2712
Coupler link rules 8815
Coupling constant 8576
Coupling of two measure spaces 4983
Coupons 9565
COV 4080
Covalent bond 6013
Covariance 3859
Covariates 5999
Covered interest arbitrage 5405
Cowles commission approach 5552
CP 7931
Crack wave 9872, 9900
Crapper's nonlinear capillary waves 9967
Crawling 5266
Creep 2019
Criterion 7753
Critical 476
Critical behavior 1443, 6772
Critical behavior of dynamical properties 1444
Critical curves 8709
Critical dimension 3360
Critical earthquake concept 7853
Critical equations 4541
Critical exponent 1395, 6783, 7822
Critical exponents 6538, 7873
Critical fractional volume 1444
Critical gels 6545
Critical part of the spectrum 936
Critical phenomena 866, 7873
Critical phenomenon 4903
Critical point 3779, 5021
Critical point, threshold, onset 3612
Critical properties and scaling 644
Critical slowing down 5667
Criticality 8521
Criticality (in physics) 7009
Cross national comparisons 2792
Cross section 9246
Crossbar array 5859
Crossing number 9268
Crossover 3754, 4134, 8521
Crowd 3142, 6476
Crowd disaster 3143
Crowd turbulence 6476
Cryosphere 1, 1705
Cryptography 7361
Crystal lattice 8465
Crystallization 9763
C-SIGN 7250
CSS code 7325
Culling process 608
Cultural evolution 3205
Culture 3205
Cumulated payoff 5588
Cumulative probability 2801
CURE 5051
Current mirror 9706
Current sheet 8521
Curse of dimensionality 719, 1775
Curvature 3040
Cusp 8521
Cusps 1642
Cut 8246
C-value 489
CWT 6352
Cybernetics 8974
Cycle 4820, 8346
Cycle time 9471
Cyclic mapping 149
Cyclic states 64
Cyclic triad 8320
Cyclicity 4820
Cylindrical Wiener process 6228
Cytoskeleton 1738

D

Darboux chart, Darboux co-ordinates, Darboux basis 2262
Dark count 7266
Data cleaning 1983
Data (information) model 4305
Data (information) table 4305
Data mining 1775, 1983, 2348, 4145, 4356, 4392, 7729
Data mining or physics of traffic – two different ap-proaches to traffic prediction 9502
Data object 1983

Data preprocessing/preparation 1983
Data set 1983
Data space 511, 2348, 2449
Data transmission system 2395
Data vector 2348
Database 7729
DBSCAN 5051
DCA 7931
DCC 4114
Deadlocked path planning 5783
Dead-reckoning 2072
Dealer 1565
Dealership market 5405
Debye layer 5573
Debye length 5573
Debye–Hückel approximation 5573
Decentralized control 113, 5618
Decidability 801
Decision making 4972
Decision node 1826
Decision problem 9160
Decision rule 1790, 4283, 4412, 7753
Decision system 4412
Decision systems 1790
Decision tree 2160, 4739
Decision under uncertainty 7753
"Decision window" 9862
Declarative language 131
Declustering 7873
Decoherence 7219, 7361
Deconvolution 2657
Decoupler link rules 8815
Deduction 201
Deep earthquake 2657
Deep water 8112, 8465
Deepness 5326
Defect 6195
Defect-tolerant 5859
Defender 4947
Deficit angle 1642
Defining system 6330
Defuzzification 4869
Degassing n. (degas v.) 9763
Degeneracy (or near-degeneracy) 8320
Degree 4487, 4930, 6495, 8232, 8816
Degree distribution 489
Degree distribution $P(k)$ 6048
Degree exponent γ 6048
Degree of a node 3924
Degree of freedom 1475, 2262
Degree rank 4930
Degrees of freedom 2827, 6426
Delay 9043, 9471
Delayed feedback 627
Delay-insensitive circuit 5859
δ-function 4317
Delta-hedge 5405
Demand game 1660
Demand revelation game 1660
Denaturation 9750
Dendrite 3334
Dendritic cell 4776
Dendrogram 1141
Density 8232
Density classification 3233
Density function (DF) tags 7422
Density-based partitioning 5051
Deoxyribozyme 5632
Deoxyribozyme-based automaton 5632
Deoxyribozyme-based logic gate 5632
Dependence assumption 8320
Dephasing 4195
Depth 9565
Dereferencing 8084
Description logic 4412
Design pattern 131
Design principle 719
Detailed balance principle 5667
Detector 6388, 9471
Deterministic 476
Deterministic automaton 4739
Deterministic chaos 2712, 2903, 6195
Deterministic continuous model 554
Deterministic evolutionary dynamic 3176
Deterministic system 368
Deterrence 4855
Development 1873
Developmental domains 1873
Developmental states, stages and ages 1873
Developmental Systems Theory 1916
Developmental time scales 1873
Deviations 5326
Device 5859
DEVS 344

DFT 1307, 7931
Diagram of congested patterns 9356
Dialectical relation 8183
Diameter D 6049
Dictator game 4609
Dielectric breakdown model 3720
Dielectric material 5859
Dielectrophoresis 5573
Diffeomorphism 703
Differentiable rigidity 3095
Differential adhesion 525
Differential gene expression 525
Differentiation 1181, 9763
Diffuse interface 9872
Diffuse interface theory 8153
Diffusion 866, 1956, 6783
Diffusion of innovation 5377
Diffusion processes 8265
Diffusion-limited aggregation 3720
Digital 9128
Digital circuit 9706
Digital images 10031
Digital organism 113
Digraph 6239, 8232, 8346
Dihedral angle rotation 6013
Dilation matrix 589
Dimension 3827
Dimension group 8889
Dimensionality reduction 1775, 1983
Dimensionless density 6505
Diode 9706
Diophantine approximation 3029, 3040
Diophantine condition, Diophantine frequency vector 2262, 4515
Diploid 9106
Dirac delta function 3334
Direct costs 1082
Direct method 8730
Direct scattering problem 4960
Direct search 2449
Directed graph 6048, 8346
Directed line 8346
(Directed) walk 6239
Directed/Undirected graph 1244
Directional dynamics 9246
Disaster 3110
Disaster lead time 3110
Disaster threat 3110
Discontinuous Galerkin method 7892
Discount factor 1082
Discount rate 1082, 3456
Discrete entity simulation 4562
Discrete least squares fitting 5800
Discrete problem 7119
Discrete ridgelet trasnform (DRT) 1718
Discrete symmetry 9644
Discrete two-sided matching model 9654
Discrete variational mechanics 2003
Discrete wavelet transform (DWT) 10007
Discrete multiwavelet transform (DMWT) 5841
Discretization 1983, 7797
Disjoint measure-preserving systems 5037
Disordered solids 8588
Disordered system 6407
Dispersal limitation 2697
Dispersion 2474
Dispersion, amplitude 9618
Dispersion, frequency 9618
Dispersion relation 9967
Dispersive estimate 2039
Disruption and delay 1845
Dissipative structures 8364
Distance 3924, 4820, 6833, 8232, 8910
Distance functions (metrics) 1790
Distortion estimate 8168
Distributed algorithm 7713
Distributed parameter system 1486
Distribution function 8690
Disulfide bond 9749
Diurnal cycle 8140
Dividend-price ratio and dividend yield 3448
Divinity 8125
DLA 866
DLVO 7931
DNA 417, 2089, 2104
DNA binary counter 5632
DNA circuit 5632
DNA computing 5632
DNA logic gate 5632
DNA Sierpinski triangle 5632
Domain 768, 10059
Domains of life 1738
Domain-specific language 131
Dome 9763

Domiciliary care 4562
Dominance 7753
Dominance-based rough set approach (DRSA) 4283
Dominant strategies 5510
Dominant structure 8948
Domination 1518
DOS 1307
Double-crossover (DX) motif 2104
Doubling map 3084
Downscaling 7591
Downsian model of political competition 9933
Downward causation 113
DQOS 7422
Drainage 4947
Drift Alfvén vortex 8521
DS flow 381
Dual graph 6579
Dual space 4317
Ductile shear zone 657
DWT 6352
Dyad independence 8320
Dyad-wise shared partners 8320
Dyke 9764
Dynamic cellular computing system 8869
Dynamic complexity 4562
Dynamic core 1424
Dynamic decomposition weights (DDW) 8948
Dynamic heterogeneity 4210
Dynamic method invocation 131
Dynamic modeling 7052
Dynamic network analysis 113
Dynamic planning 9536
Dynamic programming 6272
Dynamic programming principle 6219
Dynamic rupture model 4435
Dynamic stress change 2600
Dynamic stress drop 2538, 2581
Dynamic systems (developmental) 1873
Dynamic traffic assignment 9430
Dynamical complexity 8521
Dynamical phase 4194
Dynamical renormalization group 8521
Dynamical scaling and exponents 7873
Dynamical symmetry 9644
Dynamical system 368, 719, 866, 1932, 2933, 3016, 5450, 6175, 6329, 6939, 7603, 8420, 8630, 8927, 8974
Dynamical system (DS) 381
Dynamical systems theory 989, 4893
Dynamics 1948, 8040

E

E91 7265
EAC 271
Early adopters 5377
Early and late majority 5377
Earthquake 657, 4178
Earthquake catalog 2348
Earthquake early warning 9593
Earthquake early warning system (EEWS) 2396, 2497
Earthquake early-warning 2449
Earthquake forecasting 4179
Earthquake forecast/prediction 7853
Earthquake location 2396, 2449
Earthquake magnitude 2396
Earthquake mechanism 2556
Earthquake precursor 2497
Earthquake prediction 2497, 2621, 4179
Earthquake preparedness 4179
Earthquake quantities 5021
Earthquake size 2474
Earthquake source 2474
Earth-system models of intermediate complexity 2
Eccentricity 4820
Echo 113
Ecological community 2697
Ecological drift 2697
Economathematicians 2792
Econometrics 9034
Economic dualism 2284
Economic sector 2284
Economics 201, 1082
Economy, an 5360
Econophysicists 2792
Econophysics 2792
Ecosystem 2697
Ecosystem shifts 2712
Edge 8232, 8346

Edge betweennes 6049
Edge labeled graph 8346
Edge-wise shared partner distribution 8320
Edge-wise shared partners 8320
Education 1873
Effective dimension 8816
Effective mass 8576
Effective medium approximation 1395
Effective medium theory (EMT) 3737
Effective risk communication 3110
Effective temperature 4210
Effects 5999
Effectual bottleneck 9282, 9356
Efficiency 3327, 9933
Efficient markets hypothesis (EMH) 3374
Egalitarianism 9933
Ego-network 8232
Eigen energy and eigen wave function 4194
Eigenfield 3334
Eigenvalue 4930, 8948
Eigenvalue and eigenvector 4317
Eigenvalue and spectrum 936
Eigenvalue elasticity analysis (EEA) 8948
Eigenvector 8948
Eigenvector elasticity analysis (EVA) 8948
El Niño-southern oscillation (ENSO) 1071
Elastic 1689
Elastic interaction 8479
Elastic manifold 1126
Elastic networks 6545
Elastomer 294
Electric double layer (EDL) 5573
Electrical deep brain stimulation (DBS) 627
Electro-kinetic effect 2621
Electromechanical effects 5912
Electron gas parameter r_s 8576
Electronic communications networks (ECN) 9565
Electronic structure 5912
Electronic structure problem 6689
Electroosmosis 5573
Electrophoresis 5573
Electrophoretic mobility 5573
Electrowetting 5573
Electrowetting on dielectric (EWOD) 5573
Elementary set 7787
Elementary Swarm 8869
Elliptic 2262
Ellipticity/parabolicity 6218
Elsässer variables 8010
Embedding 8889
Embodied 4643
Embodied agent 2844
Emergence 113, 201, 4631, 6476, 8040, 8364, 8420
Emergency 3110
Emergent patterns 1041
Emergent phenomenon 768
Emergent properties 9411
Emobodied energy (emergy) 2697
Enabling behaviors 8364
Enabling rules 8364
Enantiomers and enantiomeric excess 3284
Encapsulation 132
End-effector 9749
Endocytosis 9750
Endogenous 8967
Endogenous and exogenous games 5588
Endogenous mechanisms 2792
Endogenous variable 5552
Endomorphic agents 344
Endosome 9750
Energy 3267
Energy rate density 3268
Energy release rate 2581
Energy shell 2262
Energy-momentum map 7135
Ensemble 1983, 5317
Entangled state 7250
Entanglement 7201, 7219, 7361, 8364
Enthalpic elasticity 7026
Entrepreneurship 2284
Entropic elasticity 7026
Entropic profile 489
Entropy 113, 4713, 5860, 6939, 7361, 9127
Entropy maximizing 1041
Entropy, measure-theoretic (or: metric entropy) 953
Entropy, topological 953
Enumeration algorithm 5618
Envelope broadening 7914
Environment 185, 4194, 7520
Environment-dependent tight-binding potential 9137
Envy-freeness 3327
Enzyme 2089

EO 1307
Epicenter 2449, 2497
Epidemic threshold 4659
Epidemiology 4562
Epigenesis 1916
Epigenetic inheritance 1916
Epigenetics 3965
Epimerization and racemization 3284
Epipolar constraint 4751
Epistasis 489, 3965, 9106
ε-Approximation 7119
Equicontinuity 243, 978
Equicontinuity point 243, 978
Equicontinuous CA 2232, 9160, 9247
Equicontinuous configuration 2232, 9246
Equidistributed 3040
Equilibrium 1041, 1268, 2215, 4098, 5667, 6329, 7631, 8364, 9225
Equilibrium, an 8698
Equilibrium concept 4788
Equilibrium outcome 8125
Equilibrium path 4098
Equilibrium state 6940
Equilibrium states 866
Equitability 3327
Equity premium puzzle 3456
Equity weights 1082
Equivalence class 7787
Equivalence of dynamical systems 936
Equivalence Relation 4340
Equivalence relation 7787
Equivariance 5483
Equivariant system 5299
Erdös–Rényi network 489
Erdős–Rényi random graph $\mathcal{G}(n, p)$ 7493
Ergodic 3040, 5450
Ergodic average 2934
Ergodic decomposition 3084
Ergodic law 7520
Ergodic measure preserving transformation 4983
Ergodic system 3084
Ergodic systems 989
Ergodic theorem 2934, 3084, 4983
Ergodic theory 3029, 3040, 6939
Ergodicity 866, 953, 3056, 5667, 7164
Ergodicity, weak mixing, mild mixing, mixing and rigidity of $\mathcal{T}$ 8555
Error 2450
Error correcting code (ECC) 5860
Error correction 7201
Error correction model 2758
Error syndrome 7325
Error-correcting code 7361
Escape rate formula 989
Estimate of y_i would be obtained as, the 4080
Estimator 6991
Euclidean distance (similarity) index 7422
Euclidean reconstruction 42
Eukaryote 3894
Eukaryotic cell 1738
Euler equations 9967
Evacuation 3110, 3142
Evaluation 5349
Evaluation of stage payoffs 8698
Evaluative responses 2242
Evanescent mode 7663
Evaporative cooling 9679
Event history analysis 8306
Evolution 201, 3268, 8040
Evolution (artificial) 113
Evolution equations 2039, 6228, 8926
Evolution in materio 3221
Evolution of fuzzy systems 4696
Evolution of wave functions 4194
Evolution rule 5524
Evolution strategies 113
Evolutionarily stable equilibria (ESS) 6639
Evolutionarily stable strategy (ESS) 3176
Evolutionary algorithm 113, 3221
Evolutionary algorithms/evolutionary computation 4133
Evolutionary computing 113, 4869
Evolutionary game 113
Evolutionary game theory 5677
Evolutionary optimization 6476
Evolving system 3243
Ex ante equity premium 3456
Ex post equity premium 3456
Exact controllability 4804
Exact equivalence 8333
Exact, inexact notion 1790
Exact solution 5144
Example (*also* instance *or* case) 5317
Exchange energy 8576

Exchangeability 456
Excitability 1352
Excitability level 1352
Excitable medium 7549
Excitable system 7549
Excluded volume 1444, 6505
Exclusion principle 866
Exobiology 3268
Exogenous 5552
Exogenous factors 2792
Exosystem 9078
Expandable limit order 5405
Expanded congested patterns 9356
Expansive CA 2232, 9247
Expansivity 243, 978
Expected value 3317
Experiment 2792
Expert systems 4869
Exp-function method 8453
Exploration 8709
Explosive eruption 9764
Exponent relation 8521
Exponential growth 1041, 4920
Exponentially stable Hamiltonian 5987
Extended memory 2758
Extended similarity scaling (ESS) 8521
Extended system 4902
Extensible markup language (XML) 4305
Extension principle 3985, 4060
Extensive form 4855
Extensive system 2859
External factors 1
External noise 3360
External stability 1519
Extinction thresholds 2712
Extracellular matrix (ECM) 913
Extreme event 8364
Extreme events 3300, 4179
Extreme pathways (ExPa) analysis 5536
Extreme value theory 8521
Extrinsic calibration parameters 42
Extrinsic effect 8104
Extrinsic noise 8730
Extrinsic uncertainty 4317
Extrusive flow or eruption 9764

F

Facilitator 4474
Factor map 2883, 8889
Factor model 3475
Factual statement 9862
Failure to predict 2438
Fainted laser source 7265
False alarm 2438
Fano resonances 7290
Far from equilibrium 8364
Far-field 2497
Farsighted stable set 1519
Fast acceleration of unemployment (FAU) 3301
Fast convergent (Newton) method 5064
Fast dynamics 1042
Fast slant stack (FSS) 1718
Faster-is-slower effect 6476
Fault 657, 2497
Fault model 7853
Fault parameters 9631
Fault slip 2497
Fault strength 2538
Fault-tolerant 5860
Fault-tolerant computation 7361
Fault-tolerant quantum computing 7325
FBm 4947
Fcc 1307
Feasible assignment 9655
Feature 1983, 5349, 7787
Feature extraction 1983, 2348
Feature point 42, 4751
Feature selection 1775, 5349
Feature space 2348
Feature transformation 1775
Feature vector 2348
Features 2348
Fecundity 3206
Feedback 1042, 4562, 7052, 8990, 9043
Feedback control 3944
Feedback loop 4577, 8967
Feedback loops 2284
Feedback networks 1813
Feedback trading 5405
Feedbacks 3
Feedforward control 3944
Feedforward network 1813

Fermi hole 8576
Fermi surface 9679
Fermi–Dirac particles 7201
Fermion 9679
FHP model 866
Fiber tracts 913
Fibonacci sequence 792
Field 132, 257, 3334
Field computation 3334
Field potential 6099
Field scale 6565
Field space 3334
Field theory 3360
Field transformation 3334
Field-effect transistor (FET) 5860
Field-programmable gate array (FPGA) 5860
Filament 1738
Filter 6855, 10007
Filter bank 10007
Filter process 6855
Filter taps 10007
Filtered probability of a regime 2745
Final set 4497
Financial customers 5405
Financial forecasting 3475
Financial market 2284
Fine-grained parallelism 5860
Finite automaton 4739, 5860
Finite configurations 755
Finite difference method 5161
Finite dynamical system 160
Finite element and macro-element 5800
Finite equivalence 8889
Finite size scaling 6565, 6991, 7873
Finite state machine 113, 3682, 8050
Finite tiling problem, the 9160
Finite time attractor 9247
Finite-size scaling 5667
FINRA 1565
FIPS 1565
Firing number 1352
First generation discrete curvelet transform (DCTG1) 1718
First and second laser thresholds 6116
First law of thermodynamics 9127
(First order) density function (DF) 7422
1st Postulate of quantum mechanics 4317
First-order cybernetics 1109
First-reaction method 8730
Fitness 489, 3206, 4134
Fitness landscape 6639
Fixed effects estimation 2770
Fixed point, stable 8927
Flatness 3524
Flatness (kurtosis) 8521
Flip-flop circuit 9707
FLm 4947
Floating-gate transistor 9707
Floppy mode 2827
Floquet theory 2324
Flow 3143, 4577, 8990, 9043, 9471
Flow map 3256
Flow payoff 7651
Fluctuating forces 8927
Fluctuations 5326
Fluency 2242
Flux 1395
Flux function 4713
Flux-balance analysis (FBA) 5536
Fluxon 8465
Focal mechanism 2497
Focus-focus type 2262
Fokker Planck equation 8927
Fokker–Planck equation 3574, 8690, 8766
Fokker–Planck equation, the 2801
Fokker–Planck–Landau equation 7807
Foliation 4723
Folksonomy 8084
Font conventions 2982
Fonts 2884
Food web 2697, 3661
Food webs 2711
F-optimal stable matching (respectively, payoff) 9656
Forced and/or self-organized criticality (FSOC) 8522
Force-free field 9268
Forecast 9862
Forecast point 9592
Foreign exchange dealers 5405
Foreign exchange market 3802
Formal concept analysis 8346
Formal language 800
Formal model 7052

Formal power series 6663
Fortuin–Kasteleyn representation (FK) 8709
Forward market 5405
Forward model 1689
Fourier and Laplace transforms 5218
FQQSPR equation 7422
Fractal 381, 476, 866, 2253, 3720, 3737, 3802, 3822, 3847, 3893, 4609, 7873, 8522
Fractal dimension 2253, 3822, 8709
Fractal geometry 3016
Fractal set 3779
Fractal structure 1042
Fractal system 3754
Fractality 3858
Fractals 6783, 7565, 7822
Fractional differintegration 5218
Fractional diffusion equation 310
Fractons 3879
Fracture 6565
Fracture network 6505
Freak waves 8465
Fredkin gate 7696
Free action 5483
Free cash flows 3456
Free energy 3598
Free energy functional 6195
Free flow 9302, 9355
Freezing-by-heating effect 6476
Frequency bands of electromagnetic waves 2622
Frequency domain 8948
Frequency-halving bifurcation 2262
Friedmann–Robertson–Walker model 1634
Friend request 4367
Front 866
Frustration 4210
FS 1307
F → S transition 9303
(Full) density function 7422
Full information maximum likelihood with optimal filtering 8990
Full shift 8889
Fully-developed turbulence 6195
Function pointers 132
Function spaces notation 5926
Functional 3334
Functional coefficient model 3475
Functional data model (FDM) 4305
Functional granulation 8333
Functional language 132
Functionalization 7026
Functions: bounded, continuous, uniformly Lipschitz continuous 3256
Fundamental diagram 3143
Fundamental diagram approach to traffic flow theory and modeling 9303
Fundamental diagram of traffic flow 9303
Fundamental modes 2474
Fundamental valuation 3456
Further normalization 6176
Fusion of knowledge 4413
Fusogenic 9749
Fuzzification 4869
Fuzzy control 4869
Fuzzy data mining 4146
Fuzzy estimators 8679
Fuzzy event and its probability 4060
Fuzzy if-then rule 3985
Fuzzy logic 4060, 4869, 6066
Fuzzy logic (FL) 3985
Fuzzy logic gambit 3985
Fuzzy number 4019
Fuzzy random variable 8679
Fuzzy reasoning 6086
Fuzzy relation 4060
Fuzzy set 3985, 4019, 4060, 4869, 8679
Fuzzy set and fuzzy logic 4324
Fuzzy set and fuzzy vector 4047
Fuzzy set theory 4019, 4145, 4355
Fuzzy sets 4283, 7729
Fuzzy statistical tests 8679
Fuzzy statistics 8679
Fuzzy system parameters 3244
Fuzzy system structure 3243
Fuzzy systems 4696
F-validity 3985
FX 5405

G

Galois lattice 8346
Game 5359
Game horizon 2215
Game of Life 8869
Game of life 892, 4240

Game of life, life 113
Game theory 1509, 1660, 4098, 5677
Gamma distribution, the 2801
Gap exponent 6991
GARCH 4114
GARCH(p, q) 3456
Garden-of-Eden 64
Gas slug 8153, 9873
Gate 5860
Gatekeeper 4474
Gaussian distribution 8522
Gaussian thermostats 989
GCM 9014
Gel 294
Gel point 6545
Gelfand triplet 989
Gene 4161
Gene conversion 4161
Gene frequency or allele frequency 5656
Gene network architecture and topology 1180
Gene network dynamics 1181
Gene regulatory network (GRN) 1180
Gene-protein-reaction association (GPR) 5536
General congested patterns 9356
General metabolic equation 2697
General practitioners (GPs) 4562
Generalized constraint 3985
Generalized constraint language 3985
Generalized dimension 8522
Generalized gradients and subgradients 6271
Generalized Langevin equation 8927
Generalized rigidity 1268
Generalized soliton 8457
Generalized tracking problem 9078
Generate-and-test 3221
Generating function 2262, 3361
Generation 892, 4240
Generative social science 113
Generic graph 6013
Generic rigidity 2827
Generic space 2232
Generics 132
Genetic algorithm (GA) 113, 3233
Genetic algorithms 2114, 4146, 4696, 4869
Genetic interaction 9106
Genetic interaction (epistasis) 570
Genetic locus 6639
Genetic network 9106
Genetic perturbation 9106
Genetic programming 3233
Genetic programming (GP) 113
Genetic transformation 9106
Genetic tree 3233
Genetic-fuzzy data mining 4146
Genome 511, 4161, 5656
Genome/chromosome 4134
Genome-scale 5536
Genotype 113, 3221
Genotyping 3965
Geodesic 4983, 8232, 8346
Geodesic (flow) 3040
Geodesics 3859
Geodesy 1689
Geometric integrator 2003
Geometric phase 4194
Geometrical simulation 6013
Geometrically weighted degree counts 8320
Geostrophic adjustment 8465
Geostrophic equilibrium 8465
Gevrey spaces 6649
GGA 1307, 7931
GHG 9014
Giant component 6495
Gibbard–Satterthwaite theorem 9933
Gibbs sampler 2745
Gibbs state 6940
Gillespie algorithm 8730
Gini coefficient, the 2801
Glass transition 4210
Glider 892, 4240, 7549
Global asymptotic stability 4671
Global bus 5618
Global climate models or General circulation models 2
Global expansion of tropical diseases 1095
Global Lyapunov exponent 3504
Global positioning system 9411
Global rigidity 3096
Global search 2450
Global seismographic network (GSN) 2347
Global solution 5926
Global workspace 1109
Glycoprotein 9749
GNSS 4250

Goal-directed agents 7533
Goal-oriented agents 7533
Goto statement 132
Goto–Nambu action 1642
GPAC 271
GPS 1689, 8851
GRACLUS 5051
Gradient 257
Graduation 3985
Grain boundary 6195
Granuland 3985
Granular computing 4283, 4305, 4355
Granular computing (GrC) 4324
Granular data 4402
Granular fluid 4378
Granular link 4402
Granular matter 4378, 4997
Granular neural network 4403
Granular neuron 4402
Granular structure 4324, 4339
Granular variable 3985
Granular weights 4402
Granularity 4324, 5736
Granulation 3985, 4324, 4339, 8333
Granulation of knowledge 4413
Granule 3985, 4305, 4324
Graph 1141, 2728, 4820, 5266, 5656, 6013, 6495, 8232, 8346, 8380, 8816, 9211, 10058
Graph center 4820
Graph clustering 5051
Graph cohesion 8279
Graph diameter 7493
Graph distance 4820
Graph grammar 8816
Graph invariant 9211
Graph, lattice, tree 4902
Graph metric function 8816
Graph (network) 5692
Graph path 4820
Graph structure 1244
Graph theory 8265
Graph topology 1244
Graph visualization 1244
Graph-rewriting automata 8816
Graph-theoretical distance 9211
Gravity 848
GRDDL 8084
Greenhouse gases 1071
Grid 149, 848, 2347
Grid generation 7892
Grid methods 5051
Ground deformation 6956
Ground motion intensity measures 4435
Ground motion uncertainty 4435
Ground state 6407
Group 8927
Group action 2262
Group model building (GMB) 7052
Group velocity 9967
Groups 778
Growth stocks 3391
GSR 4644
GT 9014
$G(t)$ 6810
Guaranteed possibility 6927

H

Haar measure (on a compact group) 3040
Hadamard gate 7155
Haken–Kelso–Bunz (HKB) model 5718
Halting 5524
Halting problem, the 9160
Hamilton operator 8927
Hamiltonian 1268, 4194, 5860, 6587
Hamiltonian dynamics 5064
Hamiltonian Hopf bifurcation 2262
Hamiltonian PDE 6709
Hamiltonian system 1932, 2262, 5064
Hamiltonian system with symmetry, a 6587
Hamilton–Jacobi equations 4540
Handicap principle 8125
Haploid 9106
Haplotype 9106
Haptotaxis 913
Harmonic measure 3720
Hartree–Fock 8576
Hausdorff $\boldsymbol{a}$-measure 3084
Hausdorff dimension 3029, 3827
Hazard 3110
Hazard zone 3110
HCI 4644
Headless 132
Headway 9471

Health 1095
Health and Social Care 4562
Health treatment 4562
Heavy bottleneck 9356
Heavy-tailed degree distributions 4487
Heisenberg picture 7155
Heisenberg picture in quantum mechanics 8927
Heisenberg uncertainty principle 5860
Helicity 8522, 9750
Helicity integral 9268
Heliosphere 21, 8522
Hemagglutinin 9749
Hermitean operator 4317
Heterochrony 1916
Heterogeneity 201, 8364
Heterogeneous networks 6024
Heterogeneous team 2072
Heterogeneously catalyzed reactions 7566
Heterojunction 8576
Heteroskedasticity 4114
HF 7931
Hierarchical clustering 5051
Hierarchical modeling 456
Hierarchical organization 2531
Hierarchy 4324
HIGFET 8576
Higher dimensional shift space 8889
Higher-order programming 132
High-temperature approximation 8588
Highway capacity of free flow at bottleneck 9304
Hilbert basis 2263
Hilbert space 4194, 4317, 6723
Hill's equation 6626
Hippocampus 6099
Hirota bilinear method 5138
Histones 3894
Holonomic constraint 9750
HOM interference 7250
Homeomorphism, diffeomorphism 4723
Homo oeconomicus 8183
Homo sociologicus 8183
Homogeneity assumption 8320
Homogeneous networks 6024
Homogeneous spaces 3029
Homogeneous team 2072
Homography 43, 4751
Homolog 9106
Homomorphism 8292
Homomorphism lattice 8292
Homophily 8306
Homotopy analysis method 5161
Homotopy perturbation method 5161, 8453
Hookean spring 7026
Hopf argument 8168
Horizon, cosmological 1642
Horizontal transmission 3206
Horocycle 4983
Hoshen–Kopelman algorithm 6521
Hot carriers 7401
Household income 2284
HPP model 866
HRI 4644
Hubble time 381
Hull 6521
Hull-generating walk 6521
Hulls 8709
Human activity system 4562
Human bioclimate 1095
Humanoid robots 1109
Hybrid 9106
Hybrid algorithms 8730
Hybrid closed-loop system 4671
Hybrid controller 4671
Hybrid intelligent systems 4696
Hybrid logic 8334
Hybrid two-sided matching model 9655
Hydraulic jump 8465
Hydrodynamic fields 4378
Hydrodynamic limit 7807
Hydrodynamical equations 7807
Hydrodynamics 4378
Hyperbolic 8168
Hyperbolic cellular automaton 741
Hyperbolic dynamical system 989
Hyperbolic geometry 792
Hyperbolicity 953
Hypercubic lattice 8380
Hypercycle 113
Hyperedge 8346
Hypergraph 8346
Hyper-netted-chain (HNC) 8576
Hyperreal 6285
Hyperstatic 6013
Hypertext transfer protocol (HTTP) 10059

Hypocenter 2450, 2497, 2658
Hypo-elliptic 2263
Hypostatic 6013
Hypothesis 6252
Hysteresis 8522

I

Ideal chain 6833
Idempotency 3334
i.i.d. 4114, 7520
Image compression 6855, 10031
Image denoise 6856
Image edge detection 6856
Image inpainting 10031
Image Jacobian 43
Image processing 7549
Image restoration 10031
Image segmentation 10031
Imitator 1956
Immunocomputing 4739
Imperative language 132
Imperfect monitoring 7651
Implementable social choice rule in an equilibrium concept (e.g. Nash equilibrium) 4788
Implicit solvation 9750
Implied volatility 8783
Importance sampling 2450, 5667, 6991
Imprecise data 8679
Impulse 6955
Impulse response 3334
Imputation 1518
Ince–Strutt diagram 2323
Incidence 8306
Incident 8232
Incoherent light 6388
Incomplete information 7651
Incomplete information game 8126
Incompressible word 243
Indegree 8232, 8346, 10058
Independent events 2556
Independent loop set (ILS) 8949
Indirect costs 1082
Indiscernibility 1790
Indiscernibility relation 7787
Individual-based model 114
Individual-based spatial models 8730
Induced automorphism 8555
Induced traffic breakdown (induced F → S transition) at Bottleneck 9304
Induction 201, 1826
Inelastic interaction 8480
Inertial frame 6426
Inference, prediction and decision-making; samplesand populations 457
Infinite dimensional control system 4804
Infinite measure preserving system 3084
Infinitesimal 6285
Infinity 6285
Inflation 1634
Inflation, cosmological 1642
Influence 5999
Information 1948, 7361, 7787
Information compression, inflation and adaptation 7953
Information content 4820
Information economics 6977
Information exchange 8199
Information flow 5286
Information fusion 224
Information granule 7787
Information integration 224
Information measures of statistical memory in complexsystems 1616
Information system 4412, 7729, 7788
Information systems 1790
Information theoretical descriptor (information index) 4820
Information theory 4820, 7729
Information-based complexity 7119
Infrasound 4839
Infrasound array 4839
Inheritance 132
Inhibitor 1352
Initial conditions 5161
Initial rupture process 2538
Initial set 4497
Initial value problem 3256
Initiating mechanisms 8364
Injective (reversible) CA 9160
Injectivity 64, 243, 978
Inner product 4317
Innovation 1956, 5377
Innovator 1956

Innovators 5377
Input-output map 6371
InSAR 1689
Inspector leadership 4855
Instability 8040, 8927
Instability pockets 6626
Instability tongues 6626
Installed base 1956
Instance 1826
Institutionalist economic models 2285
Integrability 4960, 7164
Integrable 1376
Integrable DS 381
Integrable Hamiltonian systems 5064
Integrable system 1932, 2263, 4515
Integral of motion 382
Integral (of motion) 2263
Intelligence 8869
Intelligent agent 1335
Intelligent agents 7533
Intelligent control 4696, 4869
Intelligent machine 4869
Intelligent query answering 1533
Intelligent transportation system 9471
Intensity, earthquake 2497
Intentionality 4893
Interacting particle system 2712
Interaction 184, 201, 719, 5286
Interaction network 511
Interactive visualization 2347
Intercross 9106
Interdependency 8364
Interface 6459
Interferometric synthetic aperture radar (InSAR) 4250
Intergovernmental coordination group 9592
Intermediate care 4562
Intermittency 8522
Internal model 9078
Internal noise 3361
Internal representation 7533
Internal solitons 8506
Internal stability 1518
Internal versus external influence 8306
Internal wave 8112
Internal waves 6206
International GNSS service (IGS) 4250
Inter-onset time 9831
Interseismic deformation 1689
Interstellar medium (ISM) 381
Inter-story drift 2371
Interval type-2 fuzzy sets 4009
Intrinsic calibration parameters 43
Intrinsic dynamics 8420
Intrinsic effect 8104
Intrinsic noise 8730
Intrinsic uncertainty 4317
Intrinsic universality 901
Intuitive criterion 8126
Invader 4947
Invariance principle 4671
Invariant 866
Invariant group of a tiling 792
Invariant manifold 936
Invariant measure 2981, 4724, 4983, 6939
Invariant set 9225
Invariant tori 5064
Invariant torus 1932
Invasion percolation 6565
Invention 1956
Inverse problem 9631
Inverse problem, inversion 2450
Inverse scattering problem 4960
Inverse scattering transform 4960, 5138
Inverse theory 1689
Invertible system 3084
Ion mobility 5573
Ion trap 7218
Ionosphere 2622, 4250, 8522
IP 4947
IPCC 9014
IPCC emission scenario 7591
Iron law of wages 2285
Irreducible matrix 6239
Irreflexive 8346
Irrelevant parameter 8522
IS 1307
ISAC 1109
Ising model 867, 6380, 6783, 6804, 8709
Ising-type model 149
Island 8246
Isocline 5536
Iso-energetic Poincaré mapping 2263
Isogenic 9106

Isomorphic configurations 5618
Isostatic 6013
Isostatic network 2827
Isotropic 2263
Isotropic dilation matrix 589
Isotropy 867
Iteration 2934
Iterative array 800
IUWT 6352

J

JAUS 8438
Java 132
Jellium 8576
Jensen–Shannon divergence 489
Johnson's equation 9967
Joining 5038
Joining of two measure preserving transformations 4983
Joint spectrum 7136
Jointly periodic configuration 9247
Julia set 3828
Jump process 8731

K

Kadomtsev–Petviashvili (KP) equation 9968
Kalman filter 5702
KAM 5064
KAM theorem 382
KAM theory 4516, 6673
Katabatic wind 8465
k-core 1254
KdV–Burgers equation 9968
Kerner's three-phase traffic theory 9501
Key distribution problem 7265
Keynesian 2285
Kibble mechanism 1642
k-independent-2-paths 8320
Kinematic GPS 4250
Kinematic rupture model 4435
Kinematics 1475
Kinetic scale 9727
k-in-star 8320
KKS (Kostant-Kirillov-Souriau) form 5483
kMC 7931
k-means 5051
k-Means clustering 2348
k-medoid methods 5051
K-nearest neighbors 1790
k-nearest neighbors 2160
kNN – k nearest neighbors 7071
Knowledge 1789, 4411, 7533
Knowledge base 1533, 5536
Knowledge discovery and data mining (KDDM) 1983
Knowledge discovery (KD) 1983
Knowledge economy 8364
Knowledge elicitation 4474
Knowledge representation 1789, 4412
Knowledge-based system 4869
Kolmogorov complexity 243
Kolmogorov group property 8555
Kolmogorov typicality 953
Kolmogorov–Arnold–Moser (KAM) invariant surface 6195
Kolmogorov–Sinai entropy per unit time 989
Kondo effect 7343, 9644
Koopman representation of a dynamical system $\mathcal{T}$ 8555
Korteweg–de Vries equation 5138, 5144, 5161
Korteweg–de Vries (KdV) equation 9968
k-out-star 8320
Kramers–Moyal coefficients 3574
Krein collision 2263
KS 7931
k-star 8320
k-triangle 8320
Kurtosis 4114

L

Lévy flight 5218
Lévy flights 6195
Lévy stable laws 5219
Lévy walks 5219
L_2 spaces 6856
Labor 2285
Labor intensive 2285
Labor productivity 2285
Lack of information on one side 7630
Lag 4114
Laggards 5377
Lagrange multiplier 5161

Lagrangean submanifold 2263
Lamb–Dicke regime 7219
Laminar flow 6195
Land surface 1
Land use transport model 1042
Landauer–Büttiker formalism 8597
Lane formation 3143
Langevin equation 3361, 3575, 8766, 8927
Langton's ant 114
Langton's loop 114
Laplace equation 9968
Laplacian matrix 9211
Large deviations principle (LDP) 7520
Laser cooling 9679
Lattice 608, 867, 1395, 3096, 8292
Lattice animals 6545
Lattice Boltzmann model 867
Lattice gas 867
Lattice property 9656
Lattice spacing 867
Lattice; unimodular lattice 3029
Laurent polynomial 10007
Law 7520
Law of large numbers (LLN) 7520
Law of requisite variety 8974
Lax method 4960
Lazy learning 2160
LB 867
LbL 7931
LCAO 7931
LCR algorithm 7713
LDA 7931
Leader election 7713
Learning (*also* training) set 5317
Learning and teaching 1873
Learning automaton 4739
Learning classifier system 4739
Learning classifier system (LCS) 114
Learning for problem-solving and planning 5188
Learning rule 1813
Leath algorithm 6521
Lebesgue space 5450
LEM2 algorithm 7797
Leptokurtosis 3404
Leray–Schauder degree 6611
Leray–Schauder–Schaefer fixed point theorem 6611
LERS machine learning/data mining system 7798
Level 5326
Leveling 1689
Levels of interoperability 344
Levels of system specification 344
Leverage 9044
LGA 867
Liapunov exponents 382
Liapunov–Schmidt reduction 6587
Lie algebras of vector fields 6761
Lie group 2003
Lie group action 5483
Lie groups 5440
Lie–Poisson structure 2263
Life 3285
Life cycle effects 1042
Lifting 5841
Lifting step 10007
Ligase 5632
Lightning 8465
Likelihood function 2450
Limit cycle 6329
Limit cycle, stable 8927
Limit order 5392, 5405
Limit set 9160
Limit sets 4724
Lindenmeyer system (L-system) 114
Lindstedt series 6663
Line 8232
Line graph 6579
Line of optimality 5536
Linear 5239, 5326
Linear CA 978, 2981
Linear cellular automata 64
Linear dispersion relation 9968
Linear dispersive waves 9968
Linear dynamical system 8949
Linear logic 5286
Linear matrix inequality 5702
Linear programming problem 5536
Linear Schrödinger equation 9968
Linear stability theory 6195
Linear system 368
Linear systems theory 8949
Linear TDMA broadcast 2054
Linear wave equation in Cartesian coordinates 9968

Linear wave equation in cylindrical polar coordinates 9968
Linearity in mean 3475
Linguistic variable 3985
Linguistic variable, fuzzy conditional statement, com-positional rule of inference 4060
Link analysis 5266
Link or arc 9429
Linkage disequilibrium 3965
Linkage metric 5051
Linking number 9269
Links 2728
Link-spamming 5266
Lipschitz norm 7520
Liquid crystal 295, 3221
Lithography 5860
Lithosphere 4179
Ljusternik–Schnirelmann category 6612
Local bifurcation 2263
Local equilibrium 867
Local field 8588
Local frequency 8588
Local intermittency measure, Lim 8522
Local Lyapunov exponent 3504
Local maps of a cellular automaton 64
Local neighborhood 1042
Local network 8291
Local properties 936
Local rigidity 3095
Local solution 5926
Local tsunami 2658
Local vertex invariant (LOVI) 2197
Locality 8709
Localization exponent 2253
Locally stationary process 8669
Location in a network 6913
Locked zone 1689
Lock-in 4631
Loewner equation 8709
Loewner evolution 8709
Logic gate 7361
Logic programming language 132
Logical gate 7549
Logically universal processor 7549
Logistic growth 1042, 4921
Lognormal distribution 1042
Long chain branching 6833
Long memory 4114
Long period seismic events 6956
Long position 5405
Longevity 3206
Long-lived player 7651
Long-period (LP) event 9873
Long-range power law correlations 3779
Long-term climate statistics 3
Long-term correlations 3755
Long-term prediction with time series 9501
Lookup table 867
Lookup table (LUT) 3233
Loop 6495, 8232
Loop eigenvalue elasticity analysis (LEEA) 8949
Loop erased random walk (LERW) 8709
Lorenz curve, the 2801
Lorenz equations 6116
Loss (cost) function 3475
Low density limit 7807
Low dimensional semiconductor nanostructures 5912
Low-dimensional chaos 8522
Lower and upper solutions 6612
Lower approximation 7787
Lower cutoff 6991
LP 4839
LSI, large-scale integration circuit 9707
Lu/Klimas model 8522
Lyapunov and control-Lyapunov functions 8617
Lyapunov exponent 2712
Lyapunov exponents 953, 989, 9247
Lyapunov function 43, 6272, 9225
Lyapunov stability theory 4671
Lymph node 4776
Lymphocyte 4776
LYP 7931

M

m^*, c^* 4080
Machine learning 114, 2160, 4739
Macro language 132
Macromolecule 6013
Macroscopic models 3143
Macroscopic scale 9727
Macrostate 4378
Mafic 9764

Magma 9764, 9785, 9873
Magma chamber 6956, 9764
Magma intrusion 9862
Magma, melt, liquid 6955
Magnetic reconfiguration (and/or reconnection) 8522
Magnetic terms 5483
Magnetic trap 9679
Magnetohydrodynamics 8010, 8522
Magnetohydrodynamics (MHD) 9269
Magnetosphere 21, 8522
Magnetotail 8522
Magnitude 2474, 2658
Magnitude, earthquake 2497
Magnitude saturation 9578
Major histocompatability 4776
Majority gate 5860
Majority relation, strict majority relation 9943
Majority voting 9933
Malaria and global warming 1096
Manakov system 1376
Manifolds 5440
Manipulability of a social choice function 9933
Manipulable mechanism 9656
Map germ 703
Mapping 6067
Marangoni effect 5573
Marginal costs 1082
Marginal factor cost 2285
Marginal link travel time 9430
Marginal of a probability measure on a product space 5038
Marginal revenue product 2285
Marginal stability 7583
Margolus neighborhood 867
Market 5359
Market anomalies 92, 3457
Market beta 3391
Market capitalization 3391
Market economy 2285
Market efficiency 3435
Market games 5360
Market maker 1565
Market microstructure 9565
Market model 3449
Market order 5392, 5405
Market structure 5392
Marketing efforts 5377
Markov chain 2745, 2918, 5860, 5999, 8766
Markov chain Monte Carlo 6991
Markov dependence assumption 8320
Markov intertwining 5038
Markov model 3457
Markov operator 8556
Markov partition 8889
Markov process 644, 5239, 8380, 8766
Markov property 8709
Markov shift 4983
Markov shift (topological, countable state) 953
Markov-switching autoregressive model 442
Markov-switching model 2745, 3475
Markov-switching models 5326
Mark-up and mark-down 1565
Marogram 2658
Marsupial mechanism 2072
Marsupial pair 2072
Marsupial robot 2072
Marsupial team 2072
Martingale 3436, 7520
Martingale process $\hat{Z}$ 5239
Martingale property 3475
Marxist economic theory 2285
Marxist system 2285
Mask 6856
Masking epistasis 9106
Master equation 3361, 5667, 8766
Matching graphs 6579
Matching mechanism 9656
Matching μ 9655
Mathematica 132
Mathematical framework 160
Mathematical model 6789, 9727
Mathematical modeling 5912
Mathieu equation 6626
MATLAB 132
Maximal spectral type and the multiplicity function of $\mathcal{U}$ 8555
Maximal attractor 9247
Maximal inequality 2934
Maximally uninformative (or ignorance) prior 2903
Maximin preferences 9655
Maximum entropy formalism 2903
Maximum entropy measure 953

Maximum likelihood 9034
Maxwell count 6014
MC 7931
MCMC 1226
MD 1307, 7931
Mean ergodic theorem 2934
Mean field 4902
Mean field theory 5021, 9411
Mean shift algorithm 5702
Mean-Field 7873
Mean-field approximation 3361, 6772
Measure 5450
Measure of maximal entropy 8889
Measure preserving system 3084
Measure preserving transformation 4984
Measure rigidity 3096
Measure space 2883
Measure theoretic entropy 4984
Measurements of traffic flow variables 9500
Measure-preserving map 2883
Measure-preserving transformation 2918
Measures 2981, 8710
Measure-theoretic entropy 2918, 3040
Mechanical connection 5483
Mechanical stability 7026
Mechanism 4788, 5467
Mechanism, a 5510
Mechanisms for power laws 7873
Median voter theorem 9933
Mediotype 3206
Medium 1352
Mega-thrust earthquake 4250
Melt 2676, 9764
Membership function 4869
Membrane fusion 9749
Membrane structure 5524
Meme 114, 3205
Memeplex (or meme complex) 3205
Memetics 3205
Memory 823
Memory effects in stochastic processes through correlations 1616
Memory functions 1616
Memotype 3206
Mendelian genetics 9106
Mental model 7052, 7831, 9057
MEP 1307
Mere-exposure effect 2242
Mereology 4413
Meso theory 8364
Message passing interface (MPI) 149
Message passing (MP) 149
Meta rules 8183
Metabolic network null space 5536
Metabolic scaling 2697
Metabolite 5656
Metabolome 511
Metabolomics 1226
Metacommunity 2697
Metadata 344, 8084
Metapopulation 2712
Metastability 1424, 1538, 4497, 7583
Meta-stability of man-made structures 2421
Meter 2310
Method 132
Method of moments 3457, 5999
Methodological individualism 132
Metric approach 4541
Metric number theory 3029
MHD turbulence 8010
Mh-precisiand 3985
Microarray 1226
Microarray hybridization 9106
Microdynamics 867
Microfeature 3334
Microlevel simulation 8375
Microlite 9764
Microscopic criterion for traffic phases 9303
Microscopic models 3143
Microscopic structure of wide moving jam 9356
Microscopic swimmers 549
Microsimulation 1042, 9536
Microworld 9044
Mind–body problem 114, 4893
Minimal energy method 5800
Minimal self-joinings 5038
Minimum robust representationalism (MRR) 1109
Minimum spanning tree 6495
Minority-carrier transport 9707
Misfit function 2450
Missing attribute values 7798
Mitigation 3110
Mitochondria 476

Mixed state 7250
Mixed strategy 4855
Mixed-star 8320
Mixing 3040
Mixing environment 7520
Mixing systems 990
Mixture modeling 457
Mm-precisiand 3985
MNDO 7931
MO 7931
Mobile agents 132
Modal analysis 2324
Modal logic 8333
Mode 7663, 8040
2-Mode matrix 8279
Model 2285, 4392, 6772, 8405
p_1 Model 8320
p_2 Model 8320
Model (*also* classifier) 5317
Model categories 2
Model of computation 7119
Model predictive control 3944
Model purpose 9000
Model simplification approach 8949
Model space 2450
Model test 2556
Model testing 2792
Model validation 1042, 7071
Model validity 9000
Model-based automation 344
Modeler 4474
Modeling 6565, 8153
Modeling and simulation ontology 345
Modeling process 9000
Model/model system 9000
Modelocked laser 6117
Modified Korteweg–de Vries 5144
Modular design of nucleic acid catalysts 5632
Modular robot 8050
Modularity 5692
Modularity Q 6049
Modulation matrix 5841
Module 132, 489
Molecular automaton 5632
Molecular barcode 9106
Molecular chaperons 7026
Molecular computing 922
Molecular descriptor 2197
Molecular descriptors 7422
Molecular dynamics 6014, 6689, 7026, 9137
Molecular dynamics method 5668
Molecular electronics 5860
Molecular finite state automaton 5632
Molecular graph 2197, 4820, 9211
Molecular logic gate 5632
Molecular Mealy automaton 5632
Molecular motor 1738
Molecular species 326
Molecule 326
Moment analysis 6991
Moment magnitude (M_w) 4250
Moment tensor 2497, 9873, 9900
Moment-tensor 8154
Momentum cocycle 5483
Momentum mapping 2263, 5483
Monetary valuation 1082
Monitoring structure 7616
Monodromy 2263, 7136
Monodromy matrix 2324
Monomial 10007
Monotone cellular automaton 4497
Monte Carlo 6991
Monte Carlo Markov chain maximum likelihood estimation (MCMCMLE) 8320
Monte Carlo (MC) method 689
Monte Carlo step 5668
Monte-Carlo 4903
Monte-Carlo method 8690
Moore neighborhood 867
Moore's law 5860, 7361
Moral dynamics 5677
Moral hazard (hidden action) 6977
Morphogenesis 3779
Morphological computation 2844
Morphology 5618
MOSFET 8576
MOSFETS 5860
Motif 489
Motility 1738
Movement 848
Moving blanks within wide moving jam 9356
Moving jams 9302
MP 7931
M-precision 3985

MRA 6856
MSRB 1565
μ-Compatible metric 3084
Multi-agent based modeling 8365
Multiagent systems (MAS) 1109
"Multi-color" network 570
Multidimensional scaling 2348, 8346
Multidimensional scaling (MDS) 8279
Multifractal 3720, 3737
Multifractal system 3754
Multifractals 8522
Multifunctionality 1283
Multi-granular computing 5736
Multilayer feedforward artificial neural network 2139
Multiparticle models 867
Multiple attribute (or multiple criteria) decision sup-port 7753
Multiple equilibria 3
Multiple instruction multiple data (MIMD) 149
Multiple network 8291
Multiple robotic system 2053
Multiple scales 5912
Multipliers 1042
Multipotency 1181
Multiresolution analysis 8669
Multiresolution, multiresolution approximation (MRA) 5841
Multi-resolutional clustering analysis 2348
Multiresolutions 1166
Multiscale problem 6663
Multiscaling function, multiwavelet function, multi-wavelet 5841
Multiset 326, 5524
Multivariate splines 5800
Mutation 4134
Mutualistic networks 2711
Mylonite 657
Myopic optimum 7651

N

Nanodevice 9749
Nano-electromechanical systems (NEMS) 689
NanoHUB 5746
Nanoparticles 5889
Nanostructures 5746, 7289
Nanotechnology 2089, 5912
NASD 1565
Nash equilibrium 1587, 4855, 5677, 6639, 7616, 7651, 8126
Nash reversion 7651
National health service (NHS) 4562
Natural asynchrony 8869
Natural patterns 3612
Naturalization 4893
Navier–Stokes equation 867
Navier–Stokes equations 548, 9618, 9968
Navier–Stokes hydrodynamics 4378
n-Body problem 5960
N-body solver 2348
NDDO 7931
NDO 7931
Nearest neighbor, a 1790
Near-field 2497
Nearly integrable system 4516
Nearly-integrable Hamiltonian systems 5064
NEB 1307
Necessity measure 6927
Negative feedback 4631
Neighbor 892, 4240, 8232
Neighborhood 64, 848, 867, 3233, 7154
Neighborhood system 4324, 4355
Neighborhood system (local granular model abb. local GrC model) 4339
Nekhoroshev theorem (1977) 5987
NEMO 3-D 5746
Neoclassical economic theory 2285
Neo-classical economics 1082
Neologism-proof equilibrium 8126
Nesting 9679
Net present value 1082
Net-centric environment 345
NetLogo 132
Network 476, 719, 1244, 1254, 2728, 4820, 5999, 6099, 8232, 8246
Network analysis 2728
Network configuration 8320
Network decomposition 8246
Network diameter 489
Network effects 1956
Network exposure 8306
Network flows 6407
Network glass 6014

Network model 8213
Network motif 570, 9106
Network motifs 2712, 5692
Network node 719
Network or graph 2530
Network population 8213
Network reconstruction 5536
Network scaling 1042
Network state 719
Network threshold 8306
Networks 645
Neural complexity 1424
Neural correlate of consciousness 1424
Neural network 5860
Neural networks 1110, 4869, 6066
Neuro-fuzzy systems 6086
Neuron 5860, 6099
Newman–Ziff algorithm 6521
Newtonian liquid 9764
Newtonian mechanics 4317
Newtonian viscous body 1705
Next-nearest neighbor 8816
Nilpotent CA 9247
Nilpotent cellular automata 9160
Nilpotent linear transformation 6649
NIST 4D/RCS 8438
NMOS FET 9707
No Dictatorship 9933
Node 8346
Node betweenness (betweenness centrality or load) 6049
Node degree distribution 1244
Node degree distribution n_k 7493
Node indegree (outdegree) 5266
Node-betweenness 8910
Node-degree 8910
Nodes 2728
Noise 476, 1983, 6229
Noise amplification 3504
Nomograph 271
Non revealing strategy 7631
Nonadditive entropy 2859
Non-archimedean 6285
Non-coding RNA (ncRNA) 3964
Noncooperative game 4855
Non-degenerate integrable Hamiltonian system, function 2263
Non-equilibrium 8040
Non-equilibrium green's function 7480
Non-equilibrium phase transitions 7874
Nonequilibrium random network 4486
Non-equilibrium systems 1269
Nonextensive system 2859
Non-hierarchical clustering algorithm 2348
Non-homogeneous cellular automaton 3233
Nonindependence 4608
Nonlinear 5239, 5326
Non-linear 2285
Nonlinear cointegration 2758
Nonlinear dispersion relation 9968
Nonlinear dispersive waves 9968
Nonlinear impulse response function 3504
Nonlinear polymer 6833
Nonlinear process 9873
Nonlinear Schrödinger equation 1376
Non-linear Schrödinger (NLS) equation 9968
Nonlinear system 4902
Nonlinear systems 8949
Nonlinear time series in macroeconomics 5326
Nonlinear waves 6206
Nonnegative matrix 6239
Nonparametric regression 3475
Nonparametric test 6252
Non-player character (NPC) 1335
Non-procedural language 132
Nonsingular dynamical system 3055
Non-smooth dynamical system 6697
Non-stationarities 3755
Nonsynonymous nucleotide substitution/amino acid replacement substitution 5656
Normal form 6175, 6330, 6709, 8927
Normal form game 3176
Normal form on invariant manifolds 6761
Normal form procedure 6310
Normal form truncation 4516
Normal frequency 2263
Normal linearization 2263
Normal modes 2263, 7219
Normal state 4378
Normal variables 990
Normalized discounted value 7651
Normalizing transformations and convergence 6761
North atlantic Oscillation (NAO) 1071

Notation 2884, 2982, 3084, 3256
Notations 2039
NP-complete 4609, 5860
NPT 7931
NTIP 4947
Nuclear magnetic resonance 6014
Nuclear regulation 9091
Nucleation 4497
Nucleation and probabilistic features of traffic breakdown 9282
Nucleation model for traffic breakdown 9282
Nucleation process 2538
Nucleation zone 2538
Nucleocapsid 9750
Nucleotides and nucleic acids 3284
Nuisance parameters 5326
Nullcline 3334
Numerical methods 7892
Nursing/residential home care 4562
nv 4080
NVE 7931
NVT 7931
NYSE 1565

O

Object 132, 7787
Object closest to a set 1790
Objective function 2450, 5536, 5999, 9034
Objective-C 132
Object-oriented language 132
Observability 3524, 6371
Observation scheme 8213
Observer 132, 6371, 9078
Occupation numbers 867
Ocean waves 9968
Oceans 1
OctArm 1475
ODE 271
Odometry 3682
Off-diagonal self-joinings 5038
Offset 9471
Oligonucleotide 5632
OMC 1307
$O(n)$ models 8710
One mode matrix 8346
One-mode network 8232
One-time pad 7265
On-line random graph 7493
On-line update and batch update 5588
Onset 9831
Onset of congestion 9501
Onset time 9831
On-the-run 9565
Ontogenesys 2844
Ontology 345, 1533, 8084
OP 4947
Open CA 9247
Open reading frame (ORF) 5536
Open system 867
Open/closed loop 3524
OpenGL 2347
Operator 2934, 3334
Operon 417
Opinion dynamics 6380
Opportunity cost 2285
Optical lattice 9679
Optimal algorithm 7119
Optimal control 3944
Optimal path 6495
Optimization 4019, 9034
Optimization algorithms 8869
Optimized direct method 8731
Option 3436
Orbit 3828, 6426, 9225
Orbit of x 2934
Order flow 5392, 5405
Order of a reaction 326
Order parameter 627, 1269, 3361, 5719
Order parameters 8927
Order parameter(s) 3612
Order-driven markets 5405
Ordered and chaotic motion 6427
Ordering 3233
Ordinal properties and monotonicity 4284
Ordinary differential equation 2285, 3256, 8630
Oregonator 7549
Organizational cybernetics 8974
Organizational level 8365
Origin time 2450
Orthogonal ensembles 7505
Orthogonalization 4739
Orthography 2311
Orthologous genes 5656

Orthonormal (ON) 3334
Oscillator 892, 4240, 7603
OTC securities 1565
Outcome 9656
Outdegree 8232, 8346, 10058
Outlier 1983
Outliers 4562
Out-of-equilibrium 201
Output elasticity 2285
Output function 3524
Overconfidence 5405
Over-the-counter market 5405
OWL 8084

P

P, NP, PSPACE, NC 6388
P system 5524
PAC 5206
PageRank 5266
Pair-correlation function 8576
Pair-distribution function 8576
Pairwise majority voting rule 9933
Palais–Smale condition for a C^1 function $\varphi\colon X \to \mathbb{R}$ 6612
Paleoclimate 3
Paleoseismology 1689
Panarchy 4631
Panel data 2770
Panic 6476
Parabolic 2263
Parabolic scaling law 1719
Paradox of voting 9933
Parallel 867, 5860
Parallel algorithms 7892
Parallel computation thesis 6388
Parallel computer architecture 149
Parallel computing 922
Parallel Turing machine 741
Parallelism 5525
Paralogous genes 5656
Parameter estimate, a 9831
Parameter space 4194
Parameter space and phase space 6992
Parameters 8320
Parametric excitation 2323
Parametric resonance 2323, 6626
Paranemic-crossover DNA 2105
Pareto distribution, the 2801
Pareto optimality or pareto efficiency 9933
Pareto-optimal matching 9656
Partial algebra 8292
Partial covering (global granular model; global GrC model) 4340
Partial dependence assumption 8320
Partial differential equation 1486
Partially molten rock 2676
Particle 768, 3233
Particle filter 5702
Particle swarm optimization 114, 2114, 8869
Particle-hole symmetry 7343
Particle-wave duality 4317
Partition 1141, 1254, 2884, 4340
Partition function of order $\boldsymbol{q}$ 8522
Partition of a non-empty set X 7788
Partitioned cellular automaton 7679
Partitioning 868
Partitioning networks 6913
P-a.s. 7521
Patch 132
Path 8232, 8346, 9211
Path dependence 4631
Path effects 4435
Path length 8346
Path planning 6067
Pathogen 4776
Pathway 719
Pathway participation metric 8949
Pattern 1983, 8926
Pattern recognition 2348
Pattern recognition of rare events 3301
Pattern searching 8246
Pauli operator 7155
Payoff 8649, 9034
Payoff vector 5359
Payoffs 2215
PBC 7931
PDB 9750
PDE 271
PDF 2438, 8522
Peano curve 8710
Pebble game 6014
Pedestrian 3142
Penalized least squares fitting 5800

Penetration and product life cycle 5377
Pentagrid 792
Peptides and amino acids 3284
Per capita boundedness 5359
Per capita instantaneous growth rate 2903
Perceptron 2139, 5860
Percolation 3737, 3847, 4902, 6783, 6804, 7822, 8710
Percolation and percolation threshold 6505
Percolation theory 1395, 6496
Percolation threshold 1395, 6496, 6565, 6579
Percolation threshold p_c 3737
Percolation transition 6538
Percolative system 1395
Perfect market 2285
Perfect monitoring 7651
Performance index/disutility index 9471
Period-doubling bifurcation 5299
Periodic graph 6579
Periodic orbit 6426
Periodic orbit, a 6587
Periodic orbit theory or trace formula 7164
Periodic tiling 9159
Periodic tiling problem, the 9160
Permeability 6565
Permutation 8346
Permutative CA 2981
Permutive CA 9247
Persistence 3755
Persistent property 4516
Personalization of ranking-scores 5266
Perturbation 6649
Perturbation expansion 3361
Perturbation problem 4516
Perturbation theory 6175, 6310, 6673, 6709
Perturbations 8365
PES 1307
PF 7931
Pharmacophore 2197
Phase 9471
Phase gate 7155
Phase portrait 368
Phase shift 8480
Phase space 2264, 6426
Phase transition 868, 5719, 6772, 8040, 8365
Phase transition and criticality 600
Phase transitions 7874, 8199
Phase velocity 9968
Phase-field method 9873
Phase-flip 7266
Phase-flip error 7325
Phenocryst 9764
Phenomenological field 3335
Phenomenological theories of the fine structure of tur-bulence 3642
Phenotype 114, 9106
Phenotypic phase plan (PhPP) analysis 5536
Pheromone 8869
Philosophy 9057
Philosophy of science 6789
Phone 2310
Phoneme 2310
Phonology 2310
Phonon bus 7219
Phosphodiesterase 5632
Photon 7250
Photosphere 9269
Photosynthesis and carbon fixation 3285
Phylogenesis 2844
Phylogenetic depth 489
Phylogenetics 3965
Physical gels 6545
Physical measure 953
Physical realizability 3335
"Physically controlled percolation" 1444
Picking-off risk 5405
Piezo-electric effect 2621
Piezoelectricity 5746
Pinch effect in synchronized flow 9356
Pinched torus 2264
Pinning 1126, 2019
Pitch 5861
Pitchfork bifurcation 5299
Pivotal 5326
Plane convex fractures 6505
Planned path 5783
Planning 7953
Planning knowledge 5188
Plasma 8465
Plasma analogy 8576
Plasma resonance 8522
Plasmid 417
Plasticity 1916
Plate tectonics 1689, 2497

Platform for agent-based simulation 185
Platoon 9471
Player 8649
Players 2215
Pleiotropy 9106
PM3 7931
P-model 8522
PMOS FET 9707
Poincaré map 6426
Poincaré operator 6612
Poincaré section 382, 6195
Poincaré section/map 6587
Poincaré's disc 792
Poincaré–Dulac normal form 6761
Point process 2556
"Pointedness" 1444
Pointwise and uniform convergence 3256
Pointwise ergodic theorem 2934
Poisson photon distribution 6117
Poisson reduction 5483
Poisson space, Poisson bracket, Poisson structure 2264
Poisson symmetry 2264
Poisson systems 6587
Political economy 2285
Political ignorance 9933
Polygenetic vent 9831
Polylogarithmic TDMA broadcast 2054
Polymer 6833
Polymorphism 133
Polynomial representation 4739
Polyphase decomposition, polyphase matrix 5841
Polyprotein 7026
Pontryagin Maximum Principle 6271
Pontryagin Maximum Principle of optimal control 1027
Pooling equilibrium 8126
Population coding 3335
Population density profile 1042
Population game 3176
Population network 8213
Population-based spatial models 8731
Pore scale 6565
Porous media 6565
Portfolio theory 3436
Pose 43, 4751
Position 6913
Positional analysis 6913
Positional equivalence 8333
Positive contraction 2934
Positive definite sequence 3084
Positive feedback 4631
Possibilistic logic 6927
Possibilistic number 4019
Possibility distribution 6927
Possibility measure 6927
Possibility theory 4019
Posterior distribution 442
Posterior pdf 2450
Postmodernism 8365
Post–Moore's law computing 3335
Postseismic deformation 1689
Potential 1395, 3244
Potential function 704
Potentiometer 271
Potts model 8710
Power 5326
Power law 4609
Power law distribution 3802, 7009, 7493
Power laws 1042
Power-law network 570
PP 7931
Pragmatic frame 345
Pragmatics 345
Pre- and postprocessing 5842
Precisiand 3985
Precisiend 3985
Precision 271, 1983
Precursory signal 2556
Predecessor state 64
Predicate logic 345
Predictability 3448, 5588
Prediction 9862
Prediction of spatiotemporal congested traffic pattern 9501
Prediction rule 5177
Preference model 7753
Preference, preference aggregation 9943
Preferential attachment 4487
Preferential attachment rule 2531
Pregame 5359
Premonitory patterns 3301
Premonitory seismicity patterns 4179
Presentationalism 9057

Preservation of structure 6310
Preslip model 2538
Pressure 6940
Pre-Topology 4339
Prevalence 954, 8306
Prevention 3110
Price discovery 9565
Price taking equilibrium 5360
Price-contingent orders 5405
Primary care trusts (PCTs) 4562
Primary dealers 9565
Primary system 271
Principal connection 5483
Principal fiber bundle 2264
Principal trade 1565
Principal–agent 6977
Principle diagonal 8346
Prior distribution 442
Prior pdf 2450
Probabilistic clustering 5051
Probabilistic graphical model 1226
Probabilistic seismic hazard analysis 2497
Probabilistic system 4870
Probability 7853
Probability density 2801
Probability density function – PDF 2396
Probability density function – pdf 2450
Probability distribution 3317, 4902
Probability distribution function 8927
Probability forecast 2556
Probability – frequentist and Bayesian 457
Probability space 2883, 4984
Probability vector 2883
Problem solving environment (PSE) 2347
Procedural language 133
Process in a measure-preserving dynamical systems 5038
Processing dynamics 2242
Processor 5861
Processor complexity 741
Product 7831
Product transformation 2918
Production factor 2285
Production level 8365
Production systems 1110
Productivity 2285
Profile 4367, 9943
Programming language 133
Progression 9471
Project 1845
Projection 3335, 7591
Projection normalization 43
Projection operator 4317
Promise 7119
Propagation 868
Proper action 5483
Proper degeneracy 2264
Proteasomes 7027
Protein 6014, 7026
Protein domain 5656
Protein engineering 7026
Protein fold 7026
Protein folding 7026
Protein nanomechanics 7027
Protein unfolding 7027
Protein-protein interaction 570
Proteolytic cleavage 9749
Proteome 511
Protonation 9750
Provably convergent cooperative controls 2053
Pseudo-chaotic systems 990
Pseudolattice 1395
Pseudo-likelihood estimation 8320
Pseudo-spin 8576
Public policy 7052
Pulsating stars 382
Purchasing power parity 2285
Pure state 7250
Pure strategy (or simply strategy) 1587
P-validity 3985
PW 7931
Pycnocline 6206
Pyroclastic flow or surge 9764
Pyrosequencing 3965

Q

QCA 7931
q-logarithmic and q-exponential functions 2859
QMA complexity class 7155
QO 7422
QOS 7423
QQSPR 7423
QS 7423

QS matrix 7423
QSAC 7423
QSAR & QSTR 7423
QSAR modeling workflow 7071
QSAR – quantitative structure activity relationships 7071
QSPR 7423
Quadratic return variation 8783
Qualitative system dynamics 4562
Quantitative structure-activity relationship (QSAR) 2197
Quantitative structure-activity relationships 2114, 2140, 2160
Quantitative structure-property relationship (QSPR) 2197
Quantitative structure-toxicity relationship (QSTR) 2197
Quantitative trait locus (QTL) 9106
Quantum algorithm 7361
Quantum bifurcation 7136
Quantum bit 7361
Quantum channel 7325
Quantum circuit 7174
Quantum circuit model 7088
Quantum circuits 7325
Quantum classical correspondence 8927
Quantum communication channel 7362
Quantum communications 7201
Quantum complexity class 7174
Quantum computation 7250
Quantum computer 7362, 9128
Quantum computing 7201
Quantum computing speedup 7119
Quantum correction 7401
Quantum cryptography 7362
Quantum dense coding 7201
Quantum dot 9644
Quantum dot cellular automaton 5861
Quantum dot device 7342
Quantum dot (QD) 7309
Quantum dots 5746
Quantum dots (QDs) 7289
Quantum entanglement 1011
Quantum error-correcting code 7362
Quantum error-correcting code (QECC) 7325
Quantum gate 1011, 7325
Quantum impurity system 7343
Quantum information 7362
Quantum interactive proof system 7174
Quantum interference 4194
Quantum key distribution 7201, 7265
Quantum Loschmidt echo or fidelity 7164
Quantum Mechanics 3859
Quantum mechanics 7362
Quantum network 7219
Quantum phase transition 7136, 7343
Quantum proof 7174
Quantum register 7219
Quantum setting 7119
Quantum similarity matrices (QS matrices) 7423
Quantum similarity measures (QSM) and indices (QSI) 7423
Quantum state 4317
Quantum statistical mechanics 5668
Quantum sub-band 7401
Quantum teleportation 7201, 7250
Quantum transport in semi-conductor devices 7480
Quantum tunneling 9644
Quantum Turing machine 7155
Quantum weirdness 7362
Quantum wire 1012
Quantum-classical correspondence 7136
Quantum-Monte Carlo (QMC) 8576
(Quasi) Ergodic hypothesis 3040
Quasi integrable Hamiltonian 5987
Quasi-attractor 9247
Quasi-integrable system 1932
Quasilinear models 8949
Quasi-periodic 2264, 6626
Quasiperiodic behavior 8949
Quasi-periodic motion 1932
Quasi-periodic motions 5064
Qubit 7155, 7201, 7219, 7250, 7266, 7325, 7361
Qubit complexity 7119
Qubit (or quantum bit) 1011
Qubus 7201
Quenched constraints 294
Quenched law 7521
Query 7119
Query complexity 7089, 7119
Queue 6521
Quota of an agent 9654
Quote-driven markets 5406

Quotient set 5736, 7788
Quotient space 5736
Quotient space model 5736
Quotient structure 4339

R

Rabi-oscillations 7219
Radial basis function network 2140
Radial basis function (RBF) 3335
Radial distribution function 3779
Radiated or seismic energy 2581
Radiative transfer theory 7914
Ramified torus bundle 2264
Ramp-up preparations 3110
Random dynamics approximation 8816
Random field 3475
Random graph 6495, 7493
Random graph model $\mathcal{G}(\mathbf{w})$ 7493
Random homogeneous mixing 4660
Random manifold 2019
Random matrices 7505
Random matrix theory 7164
Random media 7914
Random network 4486
Random networks 608
Random number generator (RNG) 5668
Random recursive graph (tree) 4487
Random resistor network 3737
Random sets 8679
Random system 368
Random variable 489, 3317
Random void and the inverted random void systems, the 1444
Random walk 868, 1395, 8380, 8710
Random-graph network 8910
Randomized network 10059
Randomized setting 7119
Random-phase approximation (RPA) 8576
Rank size rule 1042
Ranking 5349
Rank-ordered multifractal analysis, ROMA 8522
Rappture 5746
Rashba- and Dresselhaus spin-orbit coupling 8597
Ratchet 5861
Ratchets 8597
Rate 8306
Rate function 5999
Rational agents 7533
Rational choice 8183
Rational expectations 4972
Rational expectations (RE) 3374
Rational function, rational map 4984
Ray path 2450
Rayleigh number 6195
Rayleigh–Bénard convection 6195
RDF 8085
RDF(S) 8085
Reachable 8232
Reaction channel 8731
Reaction kinetics 7565
Reaction network 326
Reaction rate equation (RRE) 8731
Reaction-diffusion 1042
Reaction-diffusion equation 8731
Reaction-diffusion master equation (RDME) 8731
Reaction-diffusion mechanism 525
Reaction-diffusion system 114, 6195
Reaction-diffusion systems 868, 4921
Real option 7831
Realization 6371
Realized volatility 4114, 8783
Reasoning 1789, 4413
Recall 1983
Receiver 8126
Receiver or station 2450
Receptive field 3335
Receptor 2197
Recession 3301
Reciprocal-hemizygote 9106
Recombination 4161
Reconfiguration algorithm 5618
Reconnection 9269
Record 133, 7583
Recorder 4474
Rectified diffusion and rectified heat transfer 6956
Recurrence 3040, 9225
Recurrent random walk 1396
Recurring and non-recurring congestion 9501
Recursion 6521
Recursive and recursively enumerable (re) 9160
Recursive network 4487
Recursively generated object 114

Reduced form models 5552
Reduced phase space 2264
Reduction of network 8246
Reduction principle 937
Reduction process 1775
Reductionism 4562
Redundant bond 2827
Redundant edge 6014
Reference architecture (RA) 8438
Reference frame 4250, 6426
Reference Mode 9044
Reference mode 4474
Reference mode of behavior 4562
Refinable function vector, refinement equation 5842
Reflection 133
Refraction and inverse refraction diagrams (travel timemap) 9632
Refractory period 1353
Regime shifts 3
Regime-Switching Model 2745
Regional tsunami 2658
Regression 1983
Regular domain 3233
Regular equivalence 8279, 8333
Regular lattice 608
Regular solution 5926
Regularity 978
Regulatory network 9091
Relation 8346
Relational data model (RDM) 4305
Relational granulation 8333
Relational structure (relational granular model; rela-tional GrC model) 4340
Relative equilibrium 2264
Relative periodic orbit 6587
Relativity 3859
Relaxation 8691
Relaxation oscillation 6117, 7603
Relaxation time spectrum 6546
Relevant and irrelevant nodes 600
Relevant parameter 8522
Rematching proof equilibrium 9656
Remove the degeneracy 2265
Renormalization 1396, 7822
Renormalization group 1932, 6663
Renormalization group analysis 1269
Renormalization group (RG) 5021
Renormalization group theory 7874
Repast 133
Repast simphony 133
Repast simphony score 133
Repeated game 7616, 7651
Repeated game with incomplete information 7630
Repeated games 4098, 5177
Repeater 5861
Repellor 9225
Replication 3894
Replicator 3205
Replicator dynamic 3176
Replicator dynamics 5678
Repose 9831
Representations 6723
Representative agent 92
Reputation bound 7651
Requisite complexity 8365
Requisite variety 8365
R-Equivalence 703
Resilience 4631, 8040
Resistor networks 6546
Resonance 21, 2265, 4516
Resonance vs. Non-Resonance 6709
Resonant clock 5861
Resonant eigenvalues 6761
Resonant frequencies of vibration 2371
Resonant interaction 8480
Resonant or critical phenomenon 9968
Resonant tunneling device 5861
Response function 8522
Responsive preference 9655
Restricted totalistic rules 8816
Restriction 8710
Retain 1791
Retrieve 1791
Retrovirus 9749
Return 3404
Reuse 1791
Revelation game 9656
Revelation mechanism 9656
Reverse auction 2072
Reverse field pinch 8010
Reversibility 64, 755
Reversible cellular automaton 7679, 7696
Reversible computation 5861

Reversible logic 7362
Reversible logic element 7696
Reversible system 5299
Reversible Turing machine 7696
Revise 1791
Revision protocol 3176
Revisionist economic models 2285
Reynolds number 549, 5573, 6195
$r(f)$-Separable preference 9655
Rheology 9785
Rich configuration 243
Riemann's mapping theorem 8710
Riemannian manifolds 5440
Rigid region decomposition 6014
Rigidity percolation 6546
Rigidity percolation threshold 2827
Rings 778
Rise time 4435
Risk treatment options 3110
Riskless principal trade 1565
RNA 417, 2090, 3598
RNA interference (RNAi) 3965, 9106
Road network 9727
Road pricing 9430
Road problem 8889
Robot 1475, 3683, 4644, 6066
Robust identification 5702
Role algebra 8292
Rossby solitons 8506
Rotary element 7696
Rotations on $\mathbb{T}$ 3084
Rough inclusion 1790, 4413
Rough set 4284, 4324, 7753, 7788, 8333
Rough set theory 4355
Rough sets 7729
Rough sets: knowledge as classification 4412
Route or path 9429
RSA 7931
RTTRS (or TRS) 1566
Rule 892, 4240
Rule induction 7798
Rule table 64
Rules 848
Rules and RIF 8085
Run-length limited shift 8889
Runtime type identification 133
Run-up height 9578
Runup, or runup height 9618
Rupture 4250
Rupture duration 4436
Rupture velocity 4436

S

SA 7931
SAD 3457
Saddle 2265, 5377
Saddle point 4855
Saddle-node bifurcation 5299
SAM 7931
Sample complexity 5206
Sampling 1983
Sampling design 8213
Sampling frame 8213
SAR 4644
Satellite laser ranging (SLR) 4250
Satisfiability (SAT) 2090
Saturation 2474
Saturation flow 9471
Scalability 1141
Scalable 7201
Scale free 476
Scale free network 6495, 6639
Scale invariance 8710
Scale invariant process 7583
Scale invariants 8522
Scale of observation/representation 9727
Scale-free model 6049
Scale-free network 570, 3924, 4608, 5692, 6048, 8910
Scale-free networks 234, 1042, 2711, 4487
Scaling 271, 476, 1126, 1642, 2019, 9785
Scaling hypothesis 868
Scaling law 868, 3754
Scaling limit 8710
Scaling limits 7807
Scaling methods 8265
Scanning tunneling microscope (STM) 689, 5861
Scattering 7401, 7663
Scattering coefficient $\boldsymbol{g}$ 7914
Scattering data 4960
Scenario 7831
Scenario-driven planning (abbreviated SdP) 7831
SCF 7931

Schauder fixed point theorem 6612
Schelling model 6380
Schema 345
Schrödinger equation 6724
Schrödinger picture 7155
Schrödinger picture of quantum mechanics 8927
Schrödinger representation 6724
Schrödinger's cat 7362
Schramm's theorem 8710
Scientific visualization 2347
SCS property 8555
SDCA model hierarchy 8816
Seafloor geodesy 8851
Search 5349
SEC 1566
Second generationdiscrete curvelet transformx (DCTG2) 1719
Second law of thermodynamics 7362, 9127
2nd Postulate of quantum mechanics 4317
Secondary market 9565
Second-order cybernetics 1110
Section 1244
Sector 2285
Secular part of dipole–dipole interactions 8588
Segregation model 1042
Seismic body waves 2658
Seismic cycle 1689
Seismic data-logger 2396
Seismic energy 2474
Seismic hazard 2396, 2497, 7853
Seismic hazard analysis 2497
Seismic hazard map 2497
Seismic magnitude 9578
Seismic moment 2498, 2538, 2581, 2658, 5021, 9578
Seismic moment M_o 2474
Seismic phase 2450
Seismic potency 5021
Seismic radiation 2581
Seismic risk 2498
Seismic sensors 2396
Seismic spectrum 2581
Seismic surface waves 2658
Seismic tomographic image 2676
Seismic wave 2498
Seismic waves 2658
Seismicity 2450
Seismogenic 1689
Seismogram 2450, 2498
Seismograph 2498
Seismometer 2498
Selection 4134, 5999
Self- and cross-phase modulation 1376
Self healing 1283
Self organization 7953
Self-affine system 3754
Self-assembly method 5889
Self-avoiding random walk (SAW) 8710
Self-employment 2285
Self-exciting threshold models 5326
Self-joining 5038
Self-maintaining 326
Self-organization 114, 1537, 6476, 8198, 8420, 8926
Self-organized criticality 868, 3779, 7009, 7853, 8028, 8040
Self-organized criticality (SOC) 644, 7874, 8522
Self-organized dynamics 9411
Self-organizing map 2140
Self-reconfigurable robot 8051
Self-replication 114, 8051, 8066
Self-reproduction 8067
Self-signaling set 8126
Semantics 345, 1533, 4392
Semi-algorithm 9160
Semiclassical limit 6748
Semiclassical theory 8104
Semiconcave and semiconvex functions 4541
Semicycle 8232
Semi-decidability 755, 9160
Semigroup 8292
Semi-local bifurcation 2265
Semipath 8232
Semi-supervised learning 5349
Sender 8126
Sensation seeking 3457
Sensitive CA 2232, 9161
Sensitivity analysis 5536
Sensitivity on initial conditions 954
Sensitivity (recall) 1983
Sensitivity to initial conditions 243, 978
Sensor 345
Sensors 1475
Separating equilibrium 8126

Separatrices 4516
Sequential equilibrium 1587, 7616
Serial dependence 6252
Serial offering 1566
Service oriented architecture (SOA) 8439
Service-oriented architecture 345
Settlement 5406
Sexually transmitted infection 4660
SFF-LSE 4080
SFF-SVM 4080
Shadow price 5536
Shallow and deep water waves 8506
Shallow earthquake 2658
Shallow water 6206, 8112, 8465
Shallow water (long) waves 9631
Shallow water wave equations 8112, 9618
Shallow water waves 8112
Shape memory 295
Shape space 5483
Shapley value 1509, 5359
Sharpness 5326
Shear deformation 4997
Shell models 8010
Shift equivalence 8889
Shift of finite type 8889
Shift space 8889
Shilnikov scenario 382
Shock 21, 4713
Shor's algorithm 7250
Short memory 2758
Short position 5406
Shortest independent loop set (SILS) 8949
Shortest path 6495
Shortest path (geodesic path) 6049
Shortest-path problem 7549
Short-lived player 7651
Short-term correlations 3755
Short-term prediction with time series 9501
Shot noise 6117
Shrinkage 8669
Sieves 3475
σ-Algebra 4047
Signal 800
Signal subshift 9247
Signaling game 8126
Signaling pathway 9091
Signed graph 8232
Silicic 9764
Silicon nanowire field effect transistors 7480
Similarity 1791, 4412
Simple action type 7651
Simple graph 8232
Simple graph or network 6048
Simple mechanical system 2265
Simple random graphs, Bernoulli graphs, Erdös–Rényi graphs 8320
Simple sampling 5668
2d-Simplex 7521
Simplex of probabilities over a finite set, the 7631
Simplicity 5038
Simulated method of moments 3457
Simulation 1042, 5467, 6565, 8405, 9044
Simulation point 9592
Sinai–Ruelle–Bowen measure 6940
Sinai–Ruelle–Bowen measures 954
Sinai–Ruelle–Bowen (SRB) measure 990, 8168
Sincere strategy 9656
Single electron transistor 7309
Single electron tunneling device 5861
Single instruction multiple data (SIMD) 149
Single molecule experiments (SME) 3597
Single nucleotide polymorphism (SNP) 5536
Single photon detector 7265
Single photon source 7265
"Single-color" network 570
Single-crossing condition 8126
Single-electron circuit 9707
Single-molecule force spectroscopy (SMFS) 7027
Single-transient models 8949
Singular perturbations 7603
Singular reduction 2265, 5483
Singular value decomposition 8346
Singularity 703
Singularity spectrum $f(\alpha)$ 8522
Singularity theory 4516
Singwi–Tosi–Land–Sjölander (STLS) 8576
Sinusoidal (or exponential) wave 9968
SIR 4644
SIRN (synergetic inter-representation network) 7953
Site 868
Site effects 4436
Site percolation 1443, 6538, 6579
Situated agent 2844

Situation 7831
Size 5326
Size quantization 7401
$S(k)$ 6810
Skeleton 7549
Skewness 4114
Skin effect 2622
Slaving principle 3612, 8927
Sliding block code 8889
"Slime trail motion" 913
Slip 4250
Slip distribution 4436
Slip velocity 2538
Slippage 5406
Slip-velocity function (Slip-rate function) 4436
Slow dynamics 1042
Slow earthquake 8851
Small divisors/denominators 5064
Small group effectiveness 5359
Small-world models 8265
Small-world network 570, 3924, 6049, 6639, 8910
Small-world networks 1424
Small-worlds graph 1244
Smooth flow 4723
Smooth transition models 3476
Smooth transition threshold model 2745
Smoothed probability of a regime 2745
SMP-1 environmental scanning 7831
SMP-2 mission 7831
SMP-3 objectives 7831
SMP-4 policy 7831
SMP-5 strategy 7831
SMP-6 strategy design or formulation 7831
SMP-7 strategic action or implementation 7831
SMP-8 evaluation and control 7831
SMP-9 learning through feedback 7831
Sobolev inequality 6612
Sobolev space 6709
Sobolev spaces 2039
Social agent-based modeling 114
Social choice function 9933
Social choice function, a 5510
Social choice function, social choice correspondence 9943
Social choice rule 4788
Social circuit dependence 8320
Social contract 5678
Social force 6476
Social mechanism 8365
Social network 6913, 8213, 8291, 8333
Social network analysis (SNA) 114
Social network (or, in graph theory, a graph) 8279
Social networking community 4367
Social physics 1042
Social process 8405
Social services 4562
Social-cognitive systems 8183
Socialization 8183
Socio-cultural evolution 8183
Sociometry 8346
Sociophysics 6381
Sociotype 3206
Sofic shift 8889
Soft computing 4696, 4870
Software engineering 4356
Soil organic carbon 1071
Soil–structure interaction 2421
Soil-structure interaction (SSI) 2371
Sol 6546
Solar constant and albedo 3285
Solar corona 22, 8010
Solar flare 22
Solar wind 8010, 8522
Solenoid 2265
Solidification 4497
Solitary wave 1376, 8112, 8465, 8479
Solitary waves 6206
Solitary waves (or soliton) 9968
Soliton 1376, 4960, 5138, 5144, 6117, 8112, 8453, 8457, 8465, 8479, 8506
Soliton (colloquially as well as in certain domains) 8479
Soliton perturbation theory 8453
Soliton solution 8479
Solitons 80, 5091, 5161, 6206
Solution concept 1509
Solution space 5536
Somatic evolution 913
Source 2450, 6388
Source effects 4436
Source entropy 489
Source mechanism 2474
Source parameters of an earthquake 2498
Source (sink) SCC 5266

Soybean cyst nematode 1071
Space complexity 741
Space domain 1486
Space plasma 8522
Space-homogeneous 7154
Spaces between 8365
Spacetime 3858
SPARQL 8085
Sparsity 8669
Spatial chemical Langevin equation 8731
Spatial coordinate (vector x) 8927
Spatial interaction 1042
Spatial light modulator (SLM) 6388
Spatially extended systems 868
Spatiotemporal congested traffic pattern 9500
Special flow 8555
Species abundance 2903
Species richness 2903
Species strength 2712
Specificity 1983
Spectral clustering 5052
Spectral decomposition of a unitary representation 8554
Spectral density of states 3879
Spectral dimension 310, 3879
Spectral disjointness 8555
Spectral elements 7892
Spectral (or fracton) dimension 2253
Spectrum 8927
Spence-Mirrlees condition 8126
Spike, autowave 1353
Spiking neural P system 5525
Spillway 8465
Spin 868
Spin blockade 7309
Spin dynamics 8588
Spinodal decomposition 7874
Spin-orbit interaction 8104
Spin-orbit problem 6673
Spintronic device 5861
Spintronics 689, 8597
Splicing 4161
Split selection 1827
Spontaneous symmetry breaking 7136
Spontaneous traffic breakdown (spontaneous F → S transition) at bottleneck 9304
Spontaneous wide moving jam emergence in synchronized flow 9304
Sporulation 9106
Spot market 5406
Spreading set 2232, 9247
Spring embedder 1254, 8346
Squeezing 7201
SRO 1566
Stability 5065, 6426, 8199, 8616
Stability of a system 8927
Stability theory 8630
Stabilizability 3524
Stabilization 8616
Stabilizer code 7325
Stable manifold 990
Stable matching mechanism 9656
Stable North America 4250
Stable outcome 9656
Stack 6521
Stackelberg action 7651
Stackelberg type 7651
Stage game 7651
Stakeholder 7052
Standard signaling game 8126
Star-network 8232
Start of the homicide surge (SHS) 3301
State change complexity 741
State equation 1486
State of a system 4902
State of nature 426
State of the world 426, 4788
State space 1181
State splitting 8889
State transition diagram 64
State variables 1485
State vector 8926
States and observables 6723
Static stress change 2600
Static stress drop 2581, 4436
Static traffic assignment 9430
Station 2450
Stationary distribution and transient 6992
Stationary pattern 6459
Stationary process 2934, 7583, 8669
Stationary system 368
Statistical inferrence network 511
Statistical mechanics 1269, 4378

Statistical physics 7874
Statistical stability 954
Steady state 9078
Steady states of traffic flow 9303
Steady Stokes equations 549
Steady-State flow 5925
Steady-state spin current 8104
Steady-state spin density 8104
Steepness 5326
Stellar systems 8691
Stem cell 1181
Stigmergy 114, 8869
STM 1307, 7931
Stochastic 2556
Stochastic differential and difference equations 8750
Stochastic discrete model 554
Stochastic dynamical system 6229
Stochastic dynamical systems 8750
Stochastic dynamics (SD) method 689
Stochastic game, a 8698
Stochastic Loewner evolution (SLE) 6521, 8710
Stochastic process 2556, 4984, 5451, 8380, 8766
Stochastic process in scale 3575
Stochastic simulation algorithm (SSA) 8731
Stochastic stability 954
Stochastic variable 8766
Stochastic volatility 8783
Stochastically stable state 3176
Stock 4577, 8990, 9043
Stock return 3448
Stoichiometric matrix ($\boldsymbol{\nu}$) 8731
Stokes approximation 5573
Stokes expansion 9968
Stokes lines 6748
Stokes wave 9968
Stone–Wales transformation 9138
Stop-loss orders 5406
Strain 5746
Strainmeter 4250
Strange kinetics 5219
Strategic abstention 9933
Strategic complexity 4098
Strategic form game 1518
Strategic game 1660
Strategic level 8365
Strategic management process (abbreviated SMP) 7831
Strategic (or normal) form game 1587
Strategic voting 9933
Strategies 1948, 2215
Strategy 4098
Strategy, a 8698
Strategy of a player 7630
Strategy profile 7630
Stratification 2265, 6195
Stress drop (static stress drop) 2538
Stretched-exponential distribution 7009
Strike slip fault 5021
Strombolian activity 8153
Strong component 8232
Strong motion seismograph 4250
Strong shift equivalence 8890
Strong transitivity 978
Strong turbulence 6195
Strongly connected 8232
Strongly connected component 10059
Strongly connected component (SCC) 5266
Strongly connected digraph 6239
Strongly substitutable preferences 9655
Structural change 5326
Structural descriptor 2114, 2140, 2160
Structural equivalence 8279, 8333
Structural field 3335
Structural health monitoring 2371
Structural model 5552
Structural property 8232
Structural stability 954
Structurally dynamic cellular automata 8816
Structurally stable 4516
Structure 345
Structure contribution approach 8949
Structure function of order q 8522
Structure–activity relationships 2140
Structure-activity relationships 2114, 2161
Structured language 133
Subbands 8576
Subdiffusion 6195
Subduction 1689
Subduction zone earthquake cycle 10047
Subeconomy 2285
Subexcitability 1353
Subexcitable medium 7549

Subgame 7651
Subgame perfect equilibrium 7651
Subgraph 4820, 5692, 9211
Submarine landslide 8851
Subshift attractor 9247
Subshift of finite type 954
Subshift, shift space 5451
Subshifts 2981
Substitutable preferences 9655
Sufficient statistic 4608
Sugarscape 114
Sum resonance 6626
Sumatra-Andaman earthquake 2348
S-Unit theorems 3040
Supercooled liquid 4997
Superdiffusion 6195
Superintegrable system 2265
Superposition 7362
Superposition of states 5861
Superstrings, cosmic 1642
Supervised learning 5349
Supervisor of hybrid controllers 4671
Support vector machines 2160
Suprafault 657
Surface acoustic wave (SAW) 1012
Surface gravity waves 9968
Surface reaction center 1307
Surface site 1307
Surface wave 8113
Surface waves 2498
Surface-capillary gravity waves 9968
Surjective cellular automata 9160
Surjectivity 64, 243, 755, 978
Survey sampling 8213
Swap operation 7155
Swaps 5406
Swarm 114, 133
Swarm behavior 8365
Swarm Intelligence 8869
Swarm intelligence 114, 3683
Swarm optimization 8869
Swarm Robotics 8869
Swift–Hohenberg equation 3612
Swimming 548
Swimming microrobots 549
Symbol 5842
Symbolic processing 114
Symmetric 8346
Symmetric and asymmetric phase 5588
Symmetry 6175, 8927
Symmetry breaking 7136
Symmetry reduction 6175, 6310
Symplectic 2003
Symplectic ensembles 7505
Symplectic form 2265
Symplectic manifold 2265
Symplectic reduction 5483
Symplectic structure 5065
Symport/antiport 5525
Synapse 3335, 5861, 6099
Synchronization 627, 3233, 8910, 9471
Synchronized flow patterns 9356
Synchronized flow traffic phase 9303
Synchronous 848
Syndetic set 3084
Synergetics 8926
Synergies 1537
Synonymous nucleotide substitution/amino acid re-placement substitution 5656
Synoptic scale 8465
Syntax 345, 4392
Synthetic O-D estimation 9430
Synthetic sickness or lethality (SSL) 9106
System 4562, 7831, 8405, 8974
System behavior 4562
System dynamics 2285, 4562, 7052, 8365, 8967, 8974
System dynamics (abbreviated SD) 7831
System entity structure 345
System optimum traffic assignment 9430
System specification 345
System structure 4562
System theory 8974
Systematic risk 3436
Systems approach 8974
Systems biology 1226
Systems thinking 4562, 9044
Syzygy 2265

T

T Cell 4776
Tagged set 7423
Tail 3404

Tail index 3404
Take-profit orders 5406
Tangent and cotangent spaces 5440
Target system 4194, 6790
Tau-leaping method 8731
Taylor hypothesis 8522
Taylor–Couette flow 6195
TB 7931
Technical Trading 5406
Tectonics 2498
Tele-Operation 4644
Tele-operation 3683
Teleportation 7362
Teleseism 2498
Teleseismic 4251
Teletsunami 2658
Telluric current 2621
Templates 133
Temporal regularization 4671
Ternary heptagrid 792
Tessellation 792
Test function 6330
Testing object 1827
Testing set 1827, 5317
Thalamocortical system 1424
Theories of value 2285
Theory of neuronal group selection (TNGS) 1424
Thermal stability 8691
Thermal stress and mortality 1095
Thermocapillary effect 5573
Thermocline 8465
Thermodynamic variables 5668
Thermodynamics 3268, 9127
Thermohaline circulation 3
3rd Postulate of quantum mechanics 4317
Three traffic phases 9356
Three-body problem 6673
Three-dimensional (or 3D) flow 5925
Three-phase traffic theory 9303
Threshold autoregressive model 442
Threshold model 2745, 3476
Threshold result 7201
Threshold theorem 7325
Thresholding 8669
Thunder 8465
Thunderstorm 8465
Tidal bore 8465
Tie indicator 5999
Tierra 114
Tight binding 5746
Tight wavelet frame 5800
Tight-binding method 9137
Tight-binding molecular dynamics 9137
Tiles 9159
Tiling 792, 9159
Tiling $\{p, q\}$ 792
Tiling microarray 3964
Tiling problem, the 9160
Tiling problem with a seed tile, the 9160
Time complexity 741
Time complexity of algorithm 8246
Time division multiple access (TDMA) 2054
Time evolution of the scattering data 4961
Time reversibility 5326
Time series 3754, 5326, 6252
Time series of traffic variables 9501
Time step 868
Time-dependent traffic assignment 9430
Time-homogeneous 7154
Time-space diagram 3538
Time-varying transition probability 2745
TIP 4947
Tissue P system 5525
Titin 9750
TM 271, 1307
T-map 9655
Toeplitz space 2232
Tomography 7250
Top level domain 10059
Top-down administration 8365
Top-down fabrication 5861
Topological complexity 9211
Topological conjugacy 8890
Topological entropy 3040, 8890, 9161
Topological genericity 954
Topological index (TI) 2197
Topological mixing 978
Topological neighborhood system 4339
Topological sort 8246
Topology 2728
Torus 1932
Total adjacency of a graph 4820
Total costs 1082
Totally balanced game 5359

TRACE (formerly NASD TRACE) 1566
Trace velocity 4839
Trading volume 5406
Traffic assignment 9429
Traffic breakdown 9282, 9302
Traffic congestion at heavy bottleneck 9356
Traffic flow management 234
Traffic flow model 9303
Traffic loading 9430
Traffic routing 9429
Traffic stream motion model 9430
Training object 1827
Training set 1827
Trajectory 1486, 4902, 8927
Trajectory or path 3597
Tranformity 2697
Transcript 4161
Transcription 417, 3894
Transcription factor 1180
Transcriptome 511, 1181
Transfer function 2475, 3335
Transfer matrix 5702
Transfer principle 6285
Transformability 4631
Transient random walk 1396
Transient states 64
Transition function 3234
Transition probability 2745, 5668
Transitive triad 8320
Transitivity 243, 978, 9225
Translation 417
Transmission probability 7663
Transmission resonance 7663
Transport coefficients 990
Transversal gates 7325
Travel demand 9536
Travel time 2450
Treatment effects 5552
Tree 1932, 4487, 6495, 6521, 6833, 8346, 9211
Tremor 4839
Triad 8232
Triangle 8320
Triangular arbitrage 5406
Triangulation 1689
Trilateration 1690
Trip distribution 1042
Triples and triplestores 8085
Trophic cascades 2712
Trophic species (*S*) 3661
Troposphere 4251, 8465
Truncation 9656
TS 1307
TS-FR 4080
Tsunami 2658, 8113, 8465, 8506
Tsunami amplitude 9592
Tsunami early warning system 9592
Tsunami earthquake 2658, 9578
Tsunami inundation 9618
Tsunami magnitude 9578
Tsunami warning system 2658
Tuning parameters 5021
Tunnel effect 6748
Tunneling 5861, 7201, 7343
Tunneling percolation 1444
Tunneling phase logic 5861
Turbulence 7874, 8140, 8522
Turbulence modeling and large eddy simulations 3642
Turbulent cascades 3642
Turing equivalent 1376
Turing machine 489, 800, 5861, 9127
Turing machine (TM) 9159
Turing patterns 3612
Turtle 133
Twist number 9269
Two mode matrix 8346
2004 Indian Ocean tsunami, the 9631
Two-dimensional electron gas (2DEG) 1012
Two-dimensional (or planar, or 2D) flow 5925
Two-mode network 8232
Two-mode network data 6914
Two-scale similarity transform (TST) 5842
Two-sided matching model 9654
Type 426, 7651, 8126
Type of an agent 4788
Type-1 fuzzy sets 4009
Type-2 fuzzy sets 4009
Types II, II_1, II_∞ and III 3056
Types of solitons 80
Typical singularity 6697

U

Uehling–Uhlenbeck equation 7807

U-m 7424
U-m set 7424
UML 345
U-molecule 7424
Unbiased estimator 6992
Uncertainty 2450, 4317, 6565
Undecidability 755
(Un)decidability 9160
Undirected graph 4820, 8232
Unearned income 2285
Unfolding 703
Unfolding (folding) pathway 7027
Unfolding morphism 703
Uni-directional QKD 7265
Uniform distribution 2934
Uniform ellipticity/uniform parabolicity 6218
Uniform measure 2982
Uniform random sampling (Monte Carlo sampling) 5536
Uniform spanning tree (UST) 8710
Uniformizing maps 8710
Unintended consequences 4562
Unit doublet 3335
Unit quaternion 43
Unitary 7250
Unitary actions on Fock spaces 8556
Unitary ensembles 7505
Unitary operations 7201
Units 8588
Universal and non universal behavior 1443
Universal cellular automaton 7679
Universal constructor 8051
Universal inputs 6371
Universal modeling schema 8183
Universal properties 6407
Universal Turing machine (UTM) 114
Universality 755, 868, 1126, 5021, 5525, 6772, 6804, 7874
Unobserved effects 2770
Unpredictable system 8869
Unscented Kalman filter 5702
Unscented particle filter 5702
Unstable manifold 990
(Un)stable manifold 2265
Unstructured language 133
Unsupervised learning 5349
Unsystematic risk 3436
3′ Untranslated region (UTR) 9106
Update rule 114
Update table 7154
Updating 868
Upper approximation 7788
Upper cutoff 6992
Upper density 3084
Urban hierarchy 1042
Urban morphology 1042
Urban system 1042
Urbanism 7953
URI 8085
URL 10059
User equilibrium traffic assignment 9430
Utilitarianism 9933
Utility 457, 4855
Utility function 1587, 9411
Utility of voting 9933
UWT 6352

V

Validation 7591
Validation process 9000
Value 7630
Value stocks 3391
Vapnik–Chervonenkis dimension 5206
Variable or metric 7832
Variance and moments 3317
Variational approach 5960
Variational distance 7521
Variational iteration method 5161, 8453
Variational principle 9968
Variational principles 6587
Variety 8974
Varying cross-sectional rank (VCR) 3476
VCG mechanisms 5510
Vector computer/vector processor 149
Vector field 3256
Vector-borne diseases 1095
Vegetative index 1071
Vehicle communication protocols 2054
Vehicular traffic 9727
Velocity potential 9968
Velocity-weakening and strengthening 10047
Verifiable information game 8126
Verification 4855

Verification functions 6271
Vertex degree 4821
Vertex distance 4821
Vertex eccentricity 4821
Vertex (vertices) 8232
Vertex-adjacency matrix 9211
Vertical transmission 3206
Very large scale integration (VLSI) 5861
Very long baseline interferometry (VLBI) 4251
Very-long-period events (VLP) 8153
Very-long-period (VLP) event 9873
Virtual character 1335
Virtual machine 133
Virtual world 1335
Viscoelastic 1690
Viscosity 868, 1690
Viscosity solutions 4541, 6218
Vision 8365
Visual servo control system 43
Volatile 6955, 9764, 9873
Volatiles 9785
Volatility 3436, 3476, 4114, 5588
Volcanic crisis 9862
Volcano hazards 9862
Volcano monitoring 9862
Volcano risk 9862
Volcano status 9862
Volcano unrest 9862
Voltage encoding 5861
Volume 9471
von Neumann architecture 8869
von Neumann neighborhood 868, 5861
von Neumann universal constructor 8051
von Neumann–Morgenstern stable set 1519
von Neumann–Morgenstern utility function 1587
Voronoi diagram 7549
Vortex motions 3642
VOS model 1642
Voting cycle or Condorcet's paradox 9933
Voting procedure, voting theory 9943
Voting rule 9933
v-precision 3985
Vulnerability 3110, 9862

W

W3C 8085
Wage employment 2285
Walk 4821, 9211
Walk dimension 310
Warning 9862
Water waves 9968
Watson-Crick pairing 489
Wave dispersion 8113
Wave fragment 7549
Wave function 7424
Wave interference 7290
Wave number 4317
Wave pattern 6459
Wave vector 4317
Waveform inversion 9873, 9900
Wavelet 6856
Wavelet and scaling filters 8669
Wavelet and scaling functions 8669
Wavelet system 589
Wavelet transform 3893, 8522
Wavelet transform modulus maxima method 3893
Wavelets 5800, 10031
Waves on a running stream 9968
Weak coupling limit 7807
Weak (defect-mediated) turbulence 6195
Weak ergodicity breaking 5219
Weak solutions 6218
Weakly connected 8232
Weakly connected component 10059
Web graph 5266
Web of data 8085
Web science 8085
Web server 10059
WEB-IS 2347
Webpage 10059
Web-page ranking 5266
Website 10059
Weight matrix 8306
Weighted distribution 4821
Weighted graph 6495
Weighted network 6048
Weighted operator 8555
Well-mixed hypothesis 8731
(Well-posed) Hybrid dynamical system 4671
Weyl pseudometrics 2232
When issued 9565
Whistler wave 8522
White noise 8750

Whitham averaged variational principle 9968
Whitham's conservation equations 9968
Whitham's equation 9968
Whitham's equation for slowly varying wavetrain 9969
Whitham's nonlinear nonlocal equations 9969
Wide moving jam emergence 9501
Wide moving jam traffic phase 9303
Wiener process 8749
Wigner surmise 7164
Wigner transform 7807
Wigner–Seitz radius 8576
Winding number 9269
Wirtinger inequality 6612
WKB 6748
Wolfram classes 755
Word of mouth (w-o-m) 5378
Workbook 4474
Worldsheet 1643
Worst-case setting 7119
Wright–Fisher process 489
Writhe 9269
WT 6352
WT1D 1718
WT2D 1718

X

XC 7931
XML 345, 8085

Y

Yeast two-hybrid 570

Z

Zeno (and discrete) solutions 4672
Zero-one parameter 9034
Zero-sum game 4855
Zero-sum games 7630
0th Postulate of quantum mechanics 4317
Zeta function 8890
Zeta-potential 5573
Z-FRB 4080
Ziff model 868
Zinbiel algebra 1028
Zipf's law 490
z-Transform 10007

Index

A

A-not-B-Error 1888, 1893, 1901
A priori unstable systems 5088
A* path planning algorithm 1342
AAAI mobile robot competition 3691
Ab initio MD simulation method 692
Abacus 5473
Abduction 214
Abelian group 8028
Abel's identity 3265
ABM *see* Agent-based modeling, *see* Agent-based modeling
Abnormal glycolytic activity, tumor locus 919
Abrupt climate changes 1097
Absence of many-particle effects 8488
Absolute continuity 4730
Absolute coordination 1541
Absolutely continuous 2938
Absolutely continuous curves 5440
Absolutely continuous function 3265
Absolutely continuous invariant measure with non-maximum entropy 966
Absorbing boundaries 8768
Absorbing state 3363, 3368, 4917, 7574, 8385, 8387, 8394, 8395
ABSS *see* Agent based social simulation
Abstract computing machine model 5467
Abstract game 6024
  endogenous network formation 6025
  feasible set 6025
  player preferences 6025
  rules of network formation 6025
Abstract input-output map 6375
Abstraction 1120
  in amorphous computing 266
AC *see* Artificial chemistry (AC)
Accelerating moment release (AMR) 2442
Acceleration by parallel electric fields 24
Acceleration mechanism 21
Acceleration rate 25
Accelerograph 2501
Access 6245
Accessibility 1052, 3532
Accessible fraction 6541
Accommodation 1885
Accreting sources 392
Accretionary lapilli 9802
Accretionary prism 9588, 10048, 10049, 10057
Accretionary wedge (prism) 10047
Accuracy 271, 275, 281, 291, 1826, 5206, 5207, 5211–5213, 7731
  lower approximation 7731
  QSAR models 7079
  sensitivity 1843
  specificity 1843
  upper approximation 7731
Accuracy rate 1997
ACE *see* Agent based computational economics
Acetaldehyde 1329, 1331
Acetylcholinesterase (AChE) 2133
  benzodiazepine receptor 2145
Acid dissociation 856
Acidosis
  carcinogenesis 915
  tumor invasion 919
Ackoff 9050
ACL *see* Agent communication language
ACO 2120, 8875–8877, 8879, 8880
Acoustic signal 9818
Acoustic speed, infrasound 4841

Acoustic surveillance for hazardous eruptions (ASHE) 4850
Acoustic velocity, infrasound 4841
Acoustic wave 6455
Actin 1740, 1745–1757, 1761, 1762, 1765, 1768–1774
  F-actin 1745, 1748, 1750, 1751, 1754, 1761, 1769, 1771
  G-actin 1745, 1750, 1755, 1757, 1761
Actin and myosin filaments 1749
Actin-based 1769
Actin-based cell motility 1772
Actin-based motility 1753–1755, 1768, 1771–1774
Actin-based movement 1756
Actin-based propulsion 1753, 1756, 1757, 1771
Actin comet 1756, 1769, 1771, 1772
Actin cortex 1746, 1747, 1762, 1765
Actin cross-linkers 1750
Actin cytoskeleton 1751, 1771, 1774
Actin filament 1745–1748, 1750–1755, 1757, 1765, 1768–1774
Actin gel 1750, 1753, 1755, 1756, 1772, 1773
Actin network 1749, 1753–1756, 1768, 1769
Actin polymerization 1746, 1750, 1752–1754, 1756, 1768, 1772, 1773
Actin treadmilling 1757
Acting 1758
Action 2939, 3872, 5192, 6599, 8555
  affine 5485
  disjoint 8558
  ergodic 8555
  indefinite 6603
  Mackey 8563
  mildly mixing 8555
  mixing 8555
  nil 8560
  product 8562
  rigid 8555
  spectrally disjoint 8555
  weakly isomorphic 8570
  weakly mixing 8555
Action-angle variable 2261, 4530, 5064, 5988, 6676, 7140, 8641
Action function 6600
  definition 5503
Action identification 8422
Action levels
  2–1-choreography symmetry 5973
  Lagrange solutions in the 3 body problem 5977
  line symmetry 5973
  mountain pass solution in the 3 body problem 5977
Action map 9242
Action potential 283, 6100
Action profile 1669
Action research 3126
Action schema 5191
Action system 8186
Action theory 8233
Action/implementation threshold 7842
Activated processes 1310
Activation energy 1315
Activation field 1893
Activation function 1813, 1817, 2143
Activator 6461, 6467
Activator-inhibitor model 9713
Activator-inhibitor systems 3349
Active bond 8386, 8394
Active galactic nuclei 29
Active gels 1755, 1764, 1765, 1771
Active gradient, in amorphous computing 264
Active membranes 5524
Active modelocking 6166
Active perception 1119
Active phases 3368
Active polar gels 1764, 1765, 1770, 1771
Active polymer solutions 1761
Activity 8237
  irregular 6105
  potential 6105
  sensitivity 600
  spontaneous 6099, 6107
Activity fronts 3363
Activity paradigm 358
Actors
  aggregations of 8190
  collective 8190
Acts 7773
Acyclic graphs 1254
Acyclic relation 9946
Adamatzky 5475
Adaptability 290, 8370

Adaptation 128, 1072, 1076, 1874, 1880, 1883, 1885, 1896, 1901, 7993, 9033
p-Adaptation 7905
Adaption
autonomous 1076
planned 1076
Adaptive 2848
Adaptive actor architecture 154
Adaptive agent 1883
Adaptive law 51
Adaptive planning framework 8439, 8440, 8449
behavior specialist 8450
decision broker 8451
situation assessment specialist 8449
Adaptive real-time classifiers 3251
Adaptive visual servo control 1225
Adaptivity in amorphous computing 265
Adding machine 2958, 9250
Addition energy spectrum, charging diagram 7317
Additive cellular automata 66, 67, 73, 74
asymptotic randomization 2993
Additive pulse modelocking 6168
Additively optimal universal, representation system 244
Additivity 2862
Address 6397
encoding function 6397, 6401
ADER-DG method 7904
Adhesion 1743, 1747, 1750–1752, 1757, 1768–1770
Adhesive 525
Adiabatic algorithm 7387
Adiabatic approximation 4194, 4198, 4200, 4201
Adiabatic connection formula 8582
Adiabatic decoupling 6689
Adiabatic equilibrium 4596
Adiabatic invariant 32, 5994
Adiabatic limit 5024
Adiabatic perturbation theory 6689
Adiabatic switching 5869
Adic transformations 2974
Adjacency 8270
Adjacency matrix 4829, 5266, 5354, 6001, 6241, 6836, 8284
for the line graph 6837
neural network 503
polynomial coefficients of 6838
Adjoint
action 6314
operator 6312
Adjoint bundle 5505
Adjoint calculations 7910
Adjusted winner 3330
ADM 5144, 5146, 5156, 5158
ADMET (absorption, distribution, metabolism, excretion, and toxicity) 2196, 2210
Adolescence 1889, 1905
Adolescent health data 8313
Adomian decomposition method 81, 83, 5144, 5163, 5167, 8505
Adomian polynomials 81, 83
Adomian polynomials 5164
Adopters 1958
Adoption rate 1967
Adsorption 1307, 1319, 1320
chemisorption 1307
electronic effects 1319
geometric effect 1319
Adsorption-limited 7576
ADTree alternating decision tree 2165
Advanced driver assistance systems 3959
Advanced traveler information systems 9430, 9433
Advantages 6843
Adversary argument 7123
Adverse drug reactions 2170
Adverse selection 5414, 6983
repeated 6985
Advertising 8134
Adware 4370, 4371
Adyan SI 780, 784
Aerogel 3887
Affect 2244
Affiliation 8282
network 8291, 8293, 8295, 8301, 8304
Affine action, mechanical systems 5485
Affine CA 2981
Affine maximizer 5518
Affine models
affine continuous-time jump–diffusion models 8791
affine dynamic term structure models 8793

Affine structure 4527
Affine systems 171, 3531
Affinity chromatography 5659
Affordances 1119
AFM 7030
Aftershock 2368, 2605, 5022, 5025, 5034, 7854, 9584, 9589
Age of a cluster or fuzzy rule 3244
Agency 4624, 4626
Agency for International Development 2798
Agenda 21 4636
Agenda control 8136
Agent 1055, 1057, 1110, 1896, 4626, 5286, 7534, 8375, 9412
  autonomous 7534
  decision-making 7536
  definition 187
  deliberative 188
  goal-directed 7534
  goal-oriented 7534
  in neoclassical economic theory 178
  intelligent agents 7534
  mobile 5604
  multiagent systems (MAS) 1109
  rational 7534
  reactive 188
Agent actions 187
Agent architecture 188
Agent base (AB) cities 7967
Agent-based computational economics (ACE) 177
Agent-based computational methods 8266
Agent-based computational model 1432, 8269, 8273, 8276
Agent-based computer simulations 2805
Agent-based model 1042, 1064, 1065, 6477
Agent-based modeling (ABM) 115, 5384, 5386, 8317, 8371, 8376, 9051
  and neoclassical economics 177
  cellular automata approaches 186
  emerging behaviors 187
  grid computing 157
  guidelines for development 152
  large scale simulations 148
  limitations 152
  load balancing 154
  parallel computing 153
  program design 152
  protocol 153
  role of evidence 181
  role of physics 188
  scaling 152
  simulation 184
Agent-based models (ABMs) 187, 848, 3375, 5999, 8047, 8315, 8317
  analytical 3375
  computational 3375
Agent-based simulation 92, 362
  platforms for 193
Agent-based social simulation (ABSS) 177
Agent communication language (ACL)-based models 191
Agent complexity, role in simulation 153
Agent interaction 187, 190
  direct 191
  environment mediated 193
  indirect 192
  models, taxonomy 191
Agent theory 7534
  cognitive systems 7542
  computational agents 7536
  deliberative agents 7541
  endogenous 7534
  exogenous 7534
  game-theoretic models 7536
  goal-directed agents 7539
  limited autonomy 7544
  limited knowledge 7545
  neural nets 7538
  strategic action 7545
Aggregate 9800
  time effects 2771
Aggregated 4871
Aggregation 884, 4604
  cluster-cluster 3708
  diffusion-limited 3707
  kinetic 3707
  with injection 3813
Aggregation algorithms 7723
Aggregation operator 224, 226
Aggregative games 3194
AGIL-schema 8186
Agile computing, definition 5615
Aging 2036

Agmon metric 6747, 6756
AGNA 8848
Agostic bonds 1325
Agree-and-pursue algorithm 7713, 7722
Agree to disagree 436
Agricultural production 1071
Ah receptor 2146
Aharonov–Bohm (AB) effect 7213, 8598
Aharonov–Bohm (AB) flux 7671
Aharonov–Bohm (AB) interference 9650
  Aharonov–Casher effect 9651
Aharonov–Bohm (AB) ring 7290
AI-engineering 4336, 4337
AIDS epidemic 4660, 4662
Air pollution, allergic reactions 1100
Air traffic
  current 235
  future 235
Air traffic dynamic 1180
Air traffic flow 234
Air traffic network
  automation 242
  efficiency 242
  non-linear growth 242
Aircraft market models 685
Airport
  hub 239
  problems 1661
  secondary 239
AIRS – artificial immune recognition system 2124
Akaike Bayesian information criterion 1698
Akaike criterion, neuro-fuzzy systems 6095
Akaike information criterion (AIC) 9918
Akaike's information criterion 9598
AKNS method 4960, 4964, 8481
AKNS pair 4964
Alarm 3126, 3301
Albedo feedback 9
AlChemy 330
Alcohol consumption 6010
Alexander–Orbach conjecture 319, 6535
Alexander, Malcolm 8269
Alfvén speed 24
Alfvén waves 34, 8527
Alfvénic flux tubes 8526
Algebra 703, 5454, 8292, 8294, 8295, 8297–8299, 8301, 8305
  of vector fields 6184
  σ-Algebra 5454
Algebraic action, invariant measure classifying 4990
Algebraic branching equation 6344
Algebraic connectivity 6241
Algebraic dynamical systems 3049
Algebraic eigenspace 6247
Algebraic graph-theory 7720
Algebraic models 8273
Algebraic number field 7104
Algebraic product 4067
Algebraic Riccati equation 8756
Algebraic set theory 8266
Algebraic structures 6289
Algorithm 272, 4180, 5743, 6407, 8735, 8739
  adiabatic 7109, 7113
  black-box 7090, 7091, 7093
  coevolutionary algorithm 4136
  continued fractions 7094
  CSM 6399, 6400
  evolutionary algorithm 4133
  exact 7093, 7098
  genetic 4134
  Gillespie 8735
  Green's function reaction dynamics 8742
  Grover's 7113
  hybrid 8739
  in artificial chemistry 327, 330
  Kitaev's 7094
  M8 4180
  MSc 4180
  of Girvan and Newman 1150
  optical 6393, 6403
  particle-based 8743
  prediction 4186
  RTP 4181
  Shor's 7089, 7090, 7093, 7113
  Simon's 7092, 7100
  SSE 4181
  stochastic simulation 8735
  tau-leaping 8737
  topological 7110, 7113
  wave-front tracking 9745
  with approximation guarantees 9959

Algorithmic complexity 967, 7165
Algorithmic model 1280
aLife 1336
Aligned frame 6466
Alignments 7064, 9851
  Hough transform 9851
  two-point azimuth method 9851
Alkane 2199
All-atom 7034
Allen–Cahn theory 873
Allergenic neophytes 1100
Allergies, caused by pollen 1100
Allocation 1512, 1526
  individually rational 1526
  Pareto efficient 1526
Allocation rule 6044
Allometric 477
Allometric law, biological fractals 3784
Allometry, biological fractals 3780
Allophone 2310
Alloy
  bulk 1322
  electronic effect 1321
  geometric effect 1322
  near-surface alloy 1322
  surface 1322
Almost conjugacy 8900
Almost equicontinuous 9237
Almost global existence 6710
Almost-invariant subspace 6693
Almost-periodic 6627, 6632
Almost sure invariance principle 966
Alonso-Sanz, Ramon 8833, 8835
α-Cut 3990, 4064
α-Helices 9752
α-Level set 4047, 4064, 8682
  random 4052
Alpha-relaxation 4214
Alpha stable distribution 3413
Alphabet 5634
  code 2315
  symbol system 2315
Alternating offers bargaining 4105
Alternative futures 7835
Alternative method 5303
Ambrosetti–Prodi problem 6620
Amenable 2945
Amenable group
  Følner sequence 4994
Ames genotoxicity 2145
Amino acids 9752
Ammonia
  elementary steps 1313
  synthesis 1313
Amoeboid cell motility 1772, 1773
Amoeboid motility 1738, 1742, 1743, 1745, 1747, 1748, 1753, 1756, 1757
Amoeboid movement 1770
Amorphous computing 257, 929
  gradients as primitives 264
  in artificial chemistry 341
  linguistic framework for 263
  means of abstraction 266
  means of combination 266
  model 258
  modeling of classical partial-differential equations 264
  restriction to spatial regions 266
  supporting infrastructure 267
Amorphous computing paradigm 260
Amorphous particle, marker propagation 260
Amorphous solid 4211
Amplification
  of memory effects 1625
  relative 7849
  significant 7849
  supply chain 7849
  temporary 7849
Amplification ratio 7849
Amplification ratio magnitude 7849
Amplitude 7605
Amplitude amplification 7096
Amplitude dispersion 9978
Amplitude equation 392, 6461, 6469, 6470, 6474
  Newell–Whitehead–Segel (NWS) 6463
Amplitude estimation 7097
Amplitude instability phase instability 6117
Amplitude noise 6117
Amplitude parameter 9984
Amplitude ratios 6528
amplRes 6398, 6399, 6401
Analog 5474

Analog computation 276, 3335, 3354, 3356, 3357, 9705
general-purpose 274
Analog computing 3335
Analog equation solver 5474
Analog field computers 3340
Analog mechanical computing device 5474
Analog memory 280
Analog thinking 283
Analog VLSI 275
Analogue experimentation 9786
Analysis 10015
Analysis method 5023
Analytic approaches to the minority game 5598
Analytic signal 636
Analytical engine 285, 5476
Analytical learning 5193
Analytical methods for structural dominance analysis 8988
Analyzing wavelets 3898, 3899
Anasov's theorem 4595
Anchor text scoring 5271
Anchoring 1557, 4978
multiplicative 1557
Ancilla 7255
AND 4879
AND gate, reversibility 7366
Anderson model 7345, 8590
invasive tumor 917
multi-impurity 7357
Anderson theorem 6309
Anderson–Weiss–Kubo (AWK) theory 8595
Anderson's Wiener process 6307
Andness 228
Anelasticity 2695
Anesthesia 1437
Angell plot 4212
Angiotensin converting enzyme (ACE) 2131, 2145
Anharmonic forces 2258
Anharmonic oscillator 6739
generalized 6739
quartic 6739, 6740
Animals 3280, 7824
Animated work environment 1476, 1480
Anisotropic diffusion 10033, 10037, 10038
Anisotropic filtering 10042
Anisotropic properties 5921
Anisotropy 5024, 6480, 6572
Annealed law 7521
Annihilating random walkers 8386
Annihilation 1557
Annihilation operator 8933
Annihilation process 3366, 3367
Annihilation reaction 3366, 7573
Anomalous diffusion 3888
Anomalous resistivity 8547, 8549
Anomalous statistics 3591
Anomalous transmission of EM waves 2622, 2623
anomalous reception of VHF waves 2629
Anosov action 3097
Anosov and Anosov-like systems 992, 996, 999, 1007, 1008
Anosov diffeomorphism 3097, 6951
SRB measure 6952
Anosov flow 3046, 3098
ANS 7993
ANS involvement 7993
Ant colony optimization 123, 2120, 8869, 8875, 8876
Ant in the labyrinth 1398, 1418, 3855
Ant trails 1280
Antecedent 4871
Anterior cingulate cortex 8005
Anti-kindling 639
Anti-persistence 5589
Anti-phase 5720
Anti-Stokes line 6752
Anti-symmetry, icosahedral group 5970
Antibody, autoimmunity 4780
Anticancer 2190
Anticipation 6492, 9732
Anticipatory model 858
Anticrowds 5596
Antidepressant 2186
Antikythera mechanism 272, 5474
Antiphase 1545
Antisoliton 8494
Anxiety 3470, 3472
Anytime agent 1347
Apennines 2399
Apoptosis induction 2186
Apparent duration time 7916
Apparent stress 2483, 2485, 2585

Apparent threshold 6573
Applicability domains, QSAR models 7078
Applicative computation 5296
Approachability 7637
Approval voting procedure 9954–9956
Approximate controllability 4809
Approximate reasoning 4871, 4880
Approximation 1724, 4358, 7790
  Diophantine 3029, 3032, 3033
  M-term 1724
  metric Diophantine 3029
  non-linear 1724
  order, multiwavelet 5847
  space 7787, 8338
  with dependent quantities 3033
Approximatively controllable 3528
APS operator 2221
  property 2221
Aptamer technology 567
Aquatic toxicity, consensus QSAR models 7082
Aqueous solubility 2144, 2187
Aqueous toxicity 2183
AR *see* Assistive robotics
Ar fluid flow in SWCNT 696
Arbitrage 5410
Arbitrage pricing theory (APT) 3475, 3477, 4126, 4972
Arbitrary exponential rate observer 6374
Arbitration 6081
Arborescence, spanning 6250
Arc 6241, 8246, 8322
ARCH 3475, 3478, 3492
  model 5328
Archetypal swing 7833
Archimedean t-norm 4416
Arching 3151
Architectural spandrels 5699
Architecture
  CMOL 5881
  of agents 188
  quasi-regularity 5867
  random 5867
  reconfigurable 5867
  regular 5866
  unstructured 5867
  von Neumann 5879
  von Neumann architecture 5883
Arcing 5321
Area sampling 8226
Arenal volcano, tremor signal 4849
Arequipa earthquake analysis 4843
ARIMA 3439
Aristotelian physics 6663
Arithmetic mean 227, 228
Arithmetic operation 5475
  artificial chemistry 331
Arithometer 5476
Arm retraction, entangled polymer dynamics 6823
ARMA (auto-regressive moving average) 5206, 5208, 5216
Arms control and disarmament 4856
Arms race 4137
Armstrong's axioms 7737
Arnold 704, 5074
  work in reduction theory 5493
Arnold cat map 996, 999, 1007
Arnold diffusion 387, 409, 2279, 4519, 5081, 5987, 8642
Arnold tongues 5309
Arnold's scheme 5075
Arnold's transformation 5075
Arrest stress 5025
Arrhenius law 2035
Arrhythmia 480
  fibrillation 483
Arrow 9944
  of time 3269, 3278
Arrow's impossibility theorem 5513, 9934, 9937
Artifact mediated models 192
Artificial chemistry (AC) 120, 326, 928
  building block 328
  definition 326
  engineering motivation 327
  information processing 341
  inspired by Turing machine 333
  mechanical 336
  scientific motivation 326
  simulation by ODE 332
  theory 337
Artificial emotions 1120
Artificial immune system (AIS) 1303, 2124, 4776
  genetic variation 4778

Artificial intelligence (AI) 1111, 3334, 3358, 4894, 8093, 8873, 8886
good old-fashioned AI 8096
strong AI 1111
weak AI 1111
Artificial life 115, 1336, 8067
Artificial metabolic network, evolution 505
Artificial neural networks (ANN) 122, 1110, 1813, 2141, 3475, 3479, 3487, 3491, 4403
approach 7995
neural networks 1110
Artificial neuron 1813, 2143
Artificial organism 333
Artificial society 8416
Artificial-to-intelligence ratio 4138
Artificial worlds 4911
Arzela–Ascoli's lemma 3259
As-if defense 178
Ascendancy, ecosystem 2708
Ash 9884
ASHE *see* Acoustic surveillance for hazardous eruptions (ASHE)
Asia 7846, 7848
Aspect ratio 2688, 6573
equivalent aspect ratio 2691
Asperity 5026, 9588, 9640, 10055
contact 5025
model, earthquake waveforms 2551
Assembler automaton *see also* Von Neumann machine
Assessment
linguistic 4048
subjective 4048
Asset
pricing model 3378
risky 5400
Assignment games 1525
Assimilation 1885
Assistive robotics (AR) 4648
Associated flow
ergodic theory 3065
Association 4401
Association analysis 1991
Association mining 4363
Association rules 4147, 4363, 7738, 7777, 7778
confidence 7739
support 7739
Assortative interaction 4667
Assortativity
assortative 6054
coefficient 6054
disassortative 6054
Assumptions, neoclassical economics 179
Asteroid belt 6438, 6443
Asteroids 387
Astrobiology 3268
Astronomy 6663
Astrophysical jets 29
Asymmetric dynamic commons model 2226
Asymmetric information 6977
Asymmetric phase 5590, 5598, 5600
Asymmetric rule 618
Asymmetric tree, fractal dimension 3783
Asymmetry 1557
ecological 2706
Asymmetry parameter 7668, 7672
Asymptotic 9237
Asymptotic analysis 7607
Asymptotic behavior 3867, 6999
of the transition semigroup 6232
Asymptotic equipartition property 6944
Asymptotic freedom 6665
Asymptotic phase 8639
Asymptotic randomization
additive cellular automata 2981
linear cellular automata 2991
Asymptotic series 6739
strong 6739
Asymptotic stability 4897, 6427
Asymptotical treatments 8593
Asynchronous cellular automata 916
Asynchrony 8869, 8880–8883
Asystematic mobility 9426
Atmospheric model intercomparison project (AMIP) 1707
Atmospheric noise 1701
Atmospheric science 8113
Atmospheric water vapor 1701
Atom pair descriptors, chirality-sensitive 7075
Atom, ultra-cold 9680
Atomic ensemble 9679
Atomic force microscopy techniques 4220
Atomic information descriptors 4824

Atomic rate equations, coupled cavity-atom rate equations 6128
Atomic shell approximation (ASA) 7429
Atomic susceptibility 6123
Atomic unit 8575
Attachment 1903, 4487, 4489
  model, biological fractals 3790
  preferential 4487, 4489
Attacks on mobile agents 5610
Attention 1119
  parameter 8939
Attenuation 2482, 8271
Attenuation relation (Ground-motion prediction equation, GMPE) 4435
  directivity effect 4450
  empirical 4462, 4468
  ground-motion scaling relations 4440, 4462
  velocity-pulses 4450
Attracting heteroclinic 6629
Attracting tori 6629
Attractiveness 8240
Attractor dynamics 1438
Attractor reconstruction 3021
Attractor states, GRNs 1182
Attractors 733, 1002, 1547, 1881, 1886, 1898, 1899, 2234, 2247, 3828, 4723, 5721, 6106, 6108, 6111, 6134, 6952, 7992, 8040, 8192, 8420, 8422–8426, 8430, 8645, 8927, 9224, 9232, 9255
  as cell types 1199
  cell types 1203
  cultural 8195
  gene expression 1204
  genes 1190
  Lorenz attractor 3837, 3838
  self-esteem 8423
  socio-structural 8195
Attribute 1826, 1988, 5737, 7754, 7787
  condition 7756
  decision 7756
  decision node 1826
  dependency 4360
  node 1826
  sampling 4861
  types 1988
Attribute-based data model (ABDM) 4305
  attribute-value pairs 4305
  directory keywords 4307
  non-directory keywords 4307
  records 4306
Attribute-value blocks 7799
Atypical graph 6013
Aubry–Mather set 6217
Aubry set 4542, 4549–4551
  critical equation 4554
  dynamical properties 4554
  metric characterization 4553
Auction mechanisms 4112
Auction theory 429
Auctioneer, Kyle model 5400
Auctions 5393, 6978
  Dutch 6986
  English 6985
  first-prize 6978
  sealed-bid 6986
  second-prize 6978
Audition 2321
Auditory error maps 2320
Aumann model 429, 432, 440
Aumann–Shapley rule 1682
Auroral electrons 33
Auslander–Yorke dichotomy theorem 9237
Australia 3110, 3126, 3133, 3134, 3138
Austrian school of economics 207
Authentication codes, partial result 5612
Authority 5272
AUTO-07p 6349
AUTO, numerical studies 6596
Autocatalysis 485, 4920
  biochemical switches 485
  bistable 485
  hysteresis 485
Autocatalytic polymer chemistry 332
Autocatalytic process 208, 1889
Autocatalytic set 120, 337
  evolution of 339
AUTOCLASS 5054
Autocorrelation 3804
Autocorrelation function 3712, 3716, 3717, 6253, 6998, 7919
  fractal time series 3794
Autogenic training 7994, 7995
Automata 418, 892, 4240, 8869
Automata gas 9413
Automated highway systems 3960

Automation, of modeling and simulation 353
Automaton 5651, 5653
Automorphism 173, 2935, 8289, 8555
approximatively transitive 8555
Chacon 8561
GAG 8566
Gaussian 8565
Gaussian–Kronecker 8566
induced 8555
infinitely divisible 8571
interval exchange 8568
loosely Bernoulli 8567
Poissonian 8566
rank one 8564
Automorphism group 8899
Autonomous systems 61, 4931, 4933
dynamical systems 6178
Autonomy 187, 1889, 1905, 4899
human–robot interaction 4651
of interacting agents 192
Autoparametric resonance 2323, 2332, 6636
Autoparametric systems 6636
Autopoiesis 8040
artificial chemistry 338
Autoregression theory 6095
Autoregressive (AR) model 2746, 2753, 2754, 9598, 9914
linear 3509
nonlinear 3509
Auxiliary model 217
Avalanche 644, 1131, 1136, 2036, 5022, 7014, 8028, 8029
Avalanche models 1707
Avalanche photodiode 7251
Average bonds per object 1444
Average class accuracy 1997
Average entropy of information 4821
Average memory 824
Average partial effect 2776, 2782
Average structural function 2782
Average voting 9954
Averaged guiding function 6618
Averageness 2246
Averaging 4595, 6627, 8642
Averaging algorithm 7713, 7719
Averaging principle 8642
Averaging theorem 6326
Averaging theory 5076
dynamical system 1032
mechanical systems 1029
Avida 115, 335
Avoided criticality 4230
Aw–Rascle macroscopic traffic flow model 9335
Axelrod model 6385, 8393, 8397
Axelrod, Robert 8382, 8393
Axial mode 6124
Axiom A 8644
Axiom A flow 3047
Axiom of Archimedes 6290
Axiom of quality 8294, 8295, 8305
Axiomatic characterizations 1513
Axisymmetric KdV equation 9995
Axon 1814, 3334, 3343, 4871

B

B-Form 5802
B-Indiscernibility relation 4412, 4417
B-Lower approximation 1792
B-Spline 6361
B-Upper approximation 1792
b-Values 7927
B2B 7846
B2C 7846
Babbage, Charles 285, 5476
Babbage's difference engine 5476
Babbling 2320
Back-arc 7924
Backbone 1395, 1404, 1410, 1412, 3705, 3709, 3853, 6535, 6538, 6574, 7827
bonds 2840
fraction 6541
probability 1404
Backgammon 4144
Backpropagation 280, 1819
Backscattering enhancement 7921
Backward pass 1819
Backward search 5350
Bacterial colony growth 3787
Bacterial computer 9706
BADD-OWA 230
Bagging 3475, 3491, 3492, 5320
Bak–Sneppen model 3795
Bak–Tang–Wiesenfeld (BTW) model 646, 6992

Baker–Campbell–Haussdorf formula 6180
Baker diagram 75
Baker's map 996, 998, 999, 1007
Bala–Goyal rules 6026
Balance 6105, 8270
  equation 3643, 4382
  in graphs 8267
  laws, fluids in the Eulerian description 5928
Balanced environment 7526
Balancedness 1511, 1512
Baldwin effect 1925
Baldwin, Frank S. 5476
Baldwin, Joseph M. 9944
Baldwin's procedure 9948
Ballistic aggregation 3721
Ballistic behavior 7524
Ballistic computer 5470
Ballistic deposition (BD) 3788
Ballistic transport 7664
Balzer 3539, 3540
Banach algebra 2995
Banach principle 2937
Banach space 7442
Band detector 564
Band structure 7403
  calculation 5916
Bandwagon effect 1965
Bandwidth conservation 5614
Bang-bang principle 5443
Bank 2812
Banks' solution 9953
Banks winner 9953, 9958
Bar-joint graph 2838
Barabási–Albert model 654, 4489
Barabási, Albértó-Laszlo 8269, 8277
Barbalat's lemma 51
Bargaining 8135
Barkhausen noise 5022
Barnes, John 8268
Baroclinic fluid 6206
Barotropic fluid 6207
Barred galaxy 413
Barriers 2034, 3125
  bureaucratic process 3126
  entry 7838
  exit 7838
  inertia 7838
  innovator 3126
Bartholdi, L. 784, 789
Basal ganglia 641
Base models 5319
Base-superstructure 8184
Basic element 8987
Basic idea 7790
Basic minority game 5589, 5596, 5597, 5601
Basic reproduction number 4662
Basic reproduction rate 4659
BASICS 7835
Basin and range province 2612
Basin of a measure 958
Basin of attraction 1548, 1881, 3828, 6039, 8424, 8425
  Boolean networks 1199
Basis 5461, 5462
Baskin–Palyulin–Zefirov neural device 2150
Bass 8307
Bass diffusion model 681
Bass model 1957, 8307
Bässler law 4212
Batagelj, Vladimir 8272
Batch learning 1816, 1819
Bates stochastic volatility jump-difussion model 8792
Bathymetry 9587
BAttelle scenario inputs to corporate strategies 7835
Bayes affinity fingerprints 2171
Bayes factor 446
Bayes' theorem 442, 456, 2413, 9192
  application of Bayes' theorem to inverse problems 9192
Bayesian 2747, 2750, 2752
Bayesian approach 4874
  to estimation 2748
Bayesian classifiers 2170
Bayesian demand game 1681
Bayesian econometrics 443
Bayesian equilibrium 428, 436, 437, 439, 441
  Nash–Bayes equilibrium 437
Bayesian estimation 2749, 2751, 2753
Bayesian F-test 448
Bayesian game 427, 436, 437, 4096, 8665
Bayesian mechanism design 5519

Bayesian methods 2170
Bayesian–Nash equilibrium 5510, 5519
Bayesian network 518
  modeling 728
Bayesian probability theory 457, 7856
Bayesian regularized neural network (BRNN) 2153
Bayesian solution, belief invariant 1594
Bayesian statistical inference 1230
  Bayes factors 1239
  Bayesian computation 1230
  Bayesian information criterion 1239
  likelihood 1230
  marginal likelihood 1230
  posterior distribution 1230
  prior predictive distribution 1230
  prior probability distribution 1230
BBKGY hierarchy 7810, 7815
BD *see* Ballistic deposition
BDI model 188
BDS 3443
  test 3509, 3520, 6260, 6261, 6266
Beam splitter 7252
Bear market, fuzzy logic 3997
Beckmann's transformation
  Equivalency conditions 9439
  Uniqueness condition 9440
Bed capacity 7845
Beehive rule 8834
Behavior 346, 2846, 6079
Behavior-based control 3682, 3689
Behavioral 2845, 5431
Behavioral contract theory 6987
Behavioral economics 8998
Behavioral instability 8004
Behavioral science 1347
  biological fractals 3780
Behavioral strategies 4863
Belgrade 2797
Belief assessment 7622
Belief-desire-intention (BDI) model 188, 1337, 8378
Belief subspace (BL-subspace) 432, 433
Beliefs 3378, 7542
  coherent infinite hierarchies of 431
  fundamentalist 3378
  infinite hierarchies of 427, 428
  trend-following 3378
Bell basis 7256
  Bell measurement 7256
Bell inequality 1015
  quantum weirdness 7377
Bell state 1015, 1025
Belousov–Shabotinsky (BZ) reaction 8941
Belousov–Zhabotinsky (BZ) reaction 886, 887, 1355, 3293, 3357, 9709
Ben-Zion and Rice model 5025
Bénard convection 6464
Bénard system 4896
Benchmarking 7840
Bending flexibility 698
Bending rigidity, polymer physics 6815
Beneficial mutations 5657
Benford's law 3042
Benjamin–Bona–Mahoney (BBM) equation 8512
Benjamin–Feir instability 10003
Benzodiazepine receptor 2127, 2184
Berezinskii–Kosterlitz–Thouless phase transition 4493
Berkowitz, Steven 8269
Berners Lee, Tim 10059
Bernoulli endomorphism 3002
Bernoulli equation 9972
Bernoulli flow 4991
Bernoulli graph 8214, 8321, 8323
Bernoulli measure 2981
Bernoulli property 961, 2929
  isomorphism 4988
Bernoulli sampling 8216
Bernoulli shift 4984
  isomorphic transformation 4986, 4988
Bernoulli system 5455
Bernstein–Bézier polynomials 5803
Bernstein–Moser lemma 5079
Bernstein–Vazirani algorithm 7181
Berry curvature 8106
Berry–Esseen inequality 965
Berry phase 4194, 4196–4198, 4200–4203, 4206, 4208, 4209
Berry's conjecture 1008
Bertoluzza et al's distance 8686
Besicovitch pseudodistance 249
Besicovitch pseudometric 2236

Bessel equation 8722
Bessel process 8708, 8713, 8722
Best-guess estimation 54
Best modeling practices, QSAR models 7084
Best response dynamics 5178
β-Active nuclei 8590
Beta-effect, planetary 8494
Beta function 6665
β-Mixing 5214
β-Sheets 9752
Beta statistic 2604
Bethe-Ansatz 7347
Bethe lattice 611, 620, 1403, 1406, 7823
  theory 2835
Betweenness 6053, 8271
  centrality 2741, 6049, 8287
  edge betweennes 6049
  node 6049
Beyer 3559
BGP 4930, 4932, 4933, 4943–4945
Bi-fractal 3775
Bi-fractional diffusion equations 5231
BI Index 4830
Bi-orthogonal wavelet transform 6354, 6355
Bi-orthogonal wavelets 10033
Bianchi identity 5505
Bias 5198, 6785
Biased net theory 8321
Biclique 8288
Biclusters 1786, 1787
Biconvexity 7645
Bid-ask spreads 3472
Bieberbach conjecture 8718
Bifractality 3903, 3904
Bifurcation 10, 1542, 1898, 2324, 3271, 4515, 6134, 6136, 6195, 6198, 6311, 6331, 6590, 8929
  bifurcation parameters qualitative change 1542
  diagram 5721, 6331
  equation 6609
  equations 5300
  Hamiltonian Hopf 7145
  Hopf 6311, 6325, 6461, 6469, 6474
  Hopf–Neĭmark–Sacker 6311, 6325
  imperfect 7143
  liquids 5935, 5940
  mapping 5304
  organization of 7144
  period doubling 2330
  period multiplying 2331
  pitchfork 2330
  point 5487, 5935
  quantum 7136
  quasi-periodic Hamiltonian Hopf 6325
  quasi-periodic Hopf 6325
  saddle-node 7143
  symmetry-breaking 6461, 6469
  Taylor–Couette flow 5937
  theory 2330
Big-bang 3274
Bilateral-unilateral rules 6032
Bilinear form 8481
Bilinear model 5327
Bilinear process 6253
Bilinear realization 6376
Billiard ball model (BBM) 5875, 7690, 7700
Billiards, quantum 7166
Bimartingale 7644
Binary alloy 2819
Binary arithmetic 9705
Binary Boolean set-operations, definition 6286
Binary counter 5638, 5639
Binary granular data model (BGDM) 4346
Binary GrC model, third 4346
Binary Hamming code 7328
Binary neighborhood system (BNS) 4346
Binary output encoding 1821
Binary relation 8291, 8292, 8298, 8300
Binary search method 6534
Binary selection variable 2780
Binary, gravothermal oscillation 8693, 8694
Binding 1553
  problem 1439
Binning 7005
Binocular rivalry 1436
Bio-engineering methods 922
Bioastronomy 3268
Biobrick 421
Biochemical information 489, 494, 498, 500
Biochemical network models 512, 522
Biochemical reaction networks 518, 519, 521
  metabolic 519
  signaling 519
  transcriptional regulatory 519
Biochemical systems theory 559

Biocomposites 1289
bone 1291
exoskeleton 1289
exoskeletons 1289
leaf 1292
nacre 1290
spicules 1291
sponge 1291
wood 1291
Biodiversity 2903, 9104
Biofuel 1078
Bioinformatics 418, 923, 929, 1227, 3798, 4407
biological fractals 3780
ecological systems 2708
Biological complexity 489, 1537
measure 491
Biological coordination 1543, 1558
Biological data integration 513
Biological evolution 1876, 1879, 3270, 3280
fractals 3795
Biological fractals 3786
physical models 3785
Biological function 719
Biological information
definition 498
shared entropy 498
Biological motifs 5699
Biological motion 8205
Biological networks 720
Boolean networks 726
genetic regulatory networks 721
metabolic networks 723
neural 4403
protein–protein interaction networks 722
signaling network 727
signaling networks 723
Biological neuron 1814
Biological systems, complexity in 5693
Biologically-inspired 923
algorithm 4778
computational algorithms 121
hardware 929
optimization 1786
Biology 3278
fractals 3779
synthetic 258
Biomarkers 9096
Biomimetics 923, 1298
Biomineralization 1289
Biomolecular network 570
bipartite 574
directed 573
quasi-bipartite 574
undirected 574
Bionanomachines 7027
Bionanotechnology, biomaterials, energy 417
Biophysical coupling 1548
Biosensing 417
Biosphere 3280
Biotechnology 423
BioWatch 930
Bipartite 8281
Bipartite graph 8356
Bipartite matching 2839
Bipartite network 6048, 8258, 8322, 8330
Bipolar depressive, internal state 8426
Birefringent optical fiber 1388
Birkhoff ergodic theorem 4986
Birkhoff normal form 2279, 6710
Birkhoff pointwise ergodic theorem 957, 9241
Birkhoff theorem
ergodic probability measures 3017
Birth-and-death process 9292
Birth–death process 8770
Bisimulation 8340
Bismut–Elworthy formula 6234
Bispinor 3875
Bistability 566
networks 1193
Bistable 5720
Bistable dynamics 1193
Bistable genetic circuits 1193
1-Bit CA 3554
Bivalent logic 3986
Bivariate M-wavelet system 590
Bjerrum length, polymer physics 6816
Black and Scholes 4117
Black and Scholes formula 4117
Black-box 7090, 7091, 7850
groups 7187
models 1999, 7119
Black monday 3404

Black–Scholes model 8792
quadratic return variation 8786
volatility smile and smirte 8790
Black–Scholes option pricing model 3445
Blackboard architecture 192
Blackwell-set 7638
BLAS 5757
Blastula 526, 532, 533, 537, 538, 542
Blobs 1601, 3705, 3709
Bloch-sphere representation 7227
Bloch-vector, quantum computing 7227
Bloch's method 6732
ϵ_2 6733
ϵ_3 6733
Bloch's solution 8592
Block co-polymer 6827, 6828, 6843
vesicle, chain configuration 6830
Block density, fracture networks 6516
Block diagonal matrix 7516
Block mapping
parallel computing 154
Block-slider 2576
Block-tridiagonal matrix 8602
Blockage 3151
Blocking word 9250
Blocking, parallel computing 155
Blockmodel 8293, 8300, 8301
Blockmodeling 8263, 8266, 8273, 8289
indirect approach 6915
regular equivalence 6915
structural equivalence 6914
Blocks 1724, 4182, 4415, 7829
complete 6916
1-covered 6916
negative 6916
null 6916
partitioning 1724
positive 6916
Blogs
livejournal 10067
mobile blogging 10070
Blood-brain barrier (BBB) 2155, 2167
Blowing snow models 1706
BlueGene 5758
Bluegill sunfish, daphnia magna, tetrahymena pyriformis 2145
Blum 279
Body force, fluids 5928
3-Body problem 5970
mountain pass solutions 5976
planar 5972
planar symmetry groups 5971
Body-wave magnitude 2479, 2486
cumulative 2488
earthquake 2664
Bogoliubov transformation 9691
Bogomolnyi bound 1646
Bohm 3353
Bohm diffusion 27
Bohr recurrence 3092
Boids 1336
Boiling point 2198
Boiling temperature 2144, 2146, 2169
Boltzmann 3054
Boltzmann constant 2860
Boltzmann distribution 4912, 6119, 6129, 7011
Boltzmann equation 4908, 7415, 7817
Boltzmann–Gibbs distribution 2801–2803, 2805
Boltzmann kinetic equation 2808
Boltzmann machine 5871, 5882
Boltzmann semispace 7449
Boltzmann, Ludwig Eduard 2860, 2863
Boltzmann's ergodic hypothesis 957
Boltzmann's statistical thermodynamics 4821
Bolza problem 1502
Bonabeau model 8400
Bond angles 2829
Bond-bending models 6558
Bond-bending networks 2841
Bond number 5581
Bond percolation 6538, 6541
Bonding affinity, differential 1331
Bonding distance 6509
Bonds 5407, 7822
average per object 6509
US treasury market 5407
Book-to-market ratio 3391
Boole 3352
Boole, George 5477
Boolean algebra of sets 6300
Boolean circuits 905
Boolean function 726
Boolean grid 6492
Boolean logic 417, 5477

Boolean logic gate 9130
Boolean networks 165, 516, 517, 727
behaviors 1199
complex GRN 1197
dynamics 1199
ensembles 1201
GRN 1198
model formalism 1197
modelling real network dynamics 1199
state transition maps 1199
state transition table 1199
Boolean operations 813
Boom and bust dynamics 681
Boosting 5321
Bootstrap 3456, 3461, 7829
Bootstrap distribution 8688
Bootstrap percolation (BP) 609, 4507
asymmetric 618
modified 613
probabilistic 622
Bootstrap sample 8688
Bootstrap test 5340, 6266
Bootstrapping 5320
Borda 9944
Borda score 9948
Borda's procedure 9948, 9952, 9955, 9960
Borel property 6324
Borel summability 6670
Borel summation method 6740
Borel theorem 6324
Borel transform 6740
Borel, E. 6324
Born approximation 6742, 6743, 7919
Born model 6556
Born–Oppenheimer approximation 6689
Born series 6742, 6743
Borsuk–Ulam type theorem 7643
Bose–Einstein and Fermi–Dirac statistics 7817, 7819
Bose–Einstein condensation 9679
Bose–Fermi mixture 9680
Bose gas 9692
Boson 9679
Boson peak 2259
Bosonic channel, quantum information transmission 7393
Bosonization 9681
Bosonization-refemionization 7347
Bott, Elizabeth 8268
Bottleneck 6487, 9356, 9369, 9421
strength 9373
Bottom-up
approach 206, 207
simulation 152
Bound state 4966–4968, 7666, 8482
Boundary conditions 66, 1398, 3169
Dirichlet 1398
Dirichlet type 1487
Neumann 1398
Neumann type 1487
periodic boundary conditions 68
Boundary of the Hill's region 6604, 6605
Boundary region 7787
Boundary singularity 5978
Boundary to a concept 1792
Boundary value problem 5931
Bounded confidence 8393, 8397
Bounded-depth quantum circuits 7197
Bounded distortion, Markov map 6951
Bounded domains, flow 5931
Bounded function 3256
Bounded-rationality 92, 209, 1595, 4097
Bounded sum 4879
Bounds on probability 4053
Bourbaki school 3840
Bourgas indices 4832, 4833
Boussinesq 8114, 8117
Boussinesq equation 6207, 6455, 8117, 8122, 8512, 9983, 9984, 9986, 9987
Boussinesq systems 8117, 8122
Bow shocks 22
Bow tie structure 5269, 10060
Bowen ball 3018
Bowen–Ruelle–Sinai theorem 960
Bowen's formula 6952
Box counting 3926
Box counting method 3709, 3883
biological fractals 3788
Box covering algorithms 3939
compact-box-burning 3940
greedy coloring algorithm 3939
maximum-excluded-mass-burning 3941
random box burning 3941
Box dimension d_B 3926

Box splines 594
Boyd, John 8268
BQP
classical upper bounds 7184
error reduction 7182
subroutine theorem 7183
BQP-hard 7111
Bragg glass 2019
Brain 282, 286, 1546, 1553, 3336, 3341, 3356–3358, 4896, 6099, 7992, 8003, 8004
analogy 1553
blood oxygen level dependent (BOLD) 1546
computers, compared to 5882
coordination dynamics 1553
development 1902
energy demands 1546
function 8941
informatics 4409
information processing 1546
pacemaker 627, 8947
phase pictograms 1554
plasticity 1902
Brain size scaling, biological fractals 3784
Brain's fiber tract, glioma growth 918
Braitenberg vehicle 3682, 3689
Branch-and-cut 6420
Branch point withdrawal, polymeric fluid 6826
Branch switching
codimension 2 equilibria 6345
flip points 6345
homoclinic branch switching 6348
Hopf points 6345
simple branch points 6343
Branched DNA 2105
Branched polymers 6545, 6546, 6548
Branching 2198
Branching models 2561
Branching patterns 4831
Branching process 645, 647
Branching structures, self-similar 3781
Branes 1642, 1657
Bravais lattice vector 692
Breadth first search 8604
Break-down thresholds 5081
Breakdown in vehiculos traffic 9282
Breaking, symmetry 3864
Breakthrough 6574
Breeder's equation 2908
Breiger, Ronald 8269
Breit Hamiltonian 7446
Breit–Wigner 7290
Breit–Wigner dip 7668
Breit–Wigner resonances 7674
Bribery 9960
Bridges 8271, 8316
Briegel 7257
Brightness 6117
Brine shrimp population 9018
Brine solution 9822
Brittle-ductile transition 657
Brittle failure 9900
Brittle tectonics 1218
Broad transcription 2317
Broadband body wave, earthquake 2666
Broadband P-wave moment magnitude 2666
Broadband record 2482
Broadband vertical velocity seismogram 2667
Brockett condition 3534
Broken symmetry 6773, 7020
Brokers 5395
interdealer (IDB) 9565, 9567
Brønsted–Evans–Polanyi relationship 1313
Brown noise, biological fractals 3794
Brownian dynamics 4218, 8742
Brownian motion 2090, 3438, 3883, 5219, 5733, 5870, 6228, 7529, 8708, 8742
biological fractals 3785
single chain dynamics 6820
Brownian particle 8776
Brownian tension, polymer physics 6813
Brownian trajectory, biological fractals 3785
Brune's model 2583
Brunsviga 5476
BS 8790
BSS 271, 281
BSS model 279, 287
Bubble cloud 6972
Bubble dynamics 6957
Bubble growth 6966
Bubble nucleation 9791
Bubble raft 5475
Bubble sum 8580
Bubbles 95, 2616, 9874
Bucharest 2398

Bucklin's procedure 9954, 9956
Buckyball 9145
Budget balancedness 5519, 5520
Building block 2105
  of an artificial chemistry 328
Buildings damaged by earthquakes
  Imperial County Services Building 2377
  7-story RC building in Van Nuys 2377
Built-in self-test (BIST), applications 3240
Bulk modulus 700
  bubble 6962
  effective 6962
  of a bubble 6962
Bulk viscosity 2682, 4387
Bullwhip effects 7849
Bundle picture, mechanics 5495
Bundle reduction
  cotangent 5498
  tangent and cotangent 5495
Bundles
  Lagrange–Poincaré 5496
  principal 5500
Burden's BCUT 2209
Bureaucracy 8369, 8373
Burgers equation 84, 85, 3372, 3649, 6230, 6455, 6467
  controllability 4815
Burgers vector 2642
Burgers vortex 3646
Burning algorithm 8038
Burridge–Knopoff 5024
Burroughs, William S. 5476
Burst 5022, 6107, 6108
Burt, Ronald 8269, 8274
Burton–Steif theorem for multidimensional subshifts 2985
Business cycle 2745, 2746, 2748, 2750–2752, 2754, 2755, 7835
  asymmetry 5331, 5334, 5342, 5344, 5345
Business dynamics 665
  growth-producing feedbacks 668, 684
  negative feedback (constraints to growth) 684
  negative feedback, constraints to growth 668
Business ecosystems 7832
Business strategy 665, 666, 8989
  resource-based view 669
  role of models 667
  scientific model applied to 667
Business to business 7846
Business to customer 7846
Business units 7849
Butterfly effect 3830, 4638
Büttiker equation 7302
BV 8789
BZR model 5026
Båth's Law 2560
  'go-game' model 2560
  Griffiths' crack theory 2560
Båth's law
  elastic rebound model 2560

C

c-Contribution 5357
C. elegans brain, neural network 503
C-Regularization 4092
C-System 385, 402
C-Value paradox 492
C^1 subsolutions
  Hamilton–Jacobi equations 4559
C4.5 1830
  decision tree 2165
C_{60} 694, 9145, 9146
  formation mechanism 9146
CA 4240, 8875, 8879–8884, *see also* Cellular automata
  applications or examples 3235
  with three states 833
CA classification
  Culik–Yu 760
  Kůrka's 756
  Li–Packard 756
  Wolfram 756
Cable-stayed bridges 2334
Caccioppoli's inequality 1492
Caenorhabditis elegans 9108
Cahn–Hilliard equation 3636, 6231, 6457, 6467
  in three room dimensions 3637
  numerical solution 3637
  phase transition from liquid to gas 3636
Cairns 3126
Cairo 2398

Calculus 3861
of variations 6272, 6277
Caldera collapse 9824
Calibration 216, 5185, 6480, 9035
function, regional 2494
matrix 47
parameter mismatch 60
California 2398
Calipher diameter, fracture networks 6514
Caltagirone 8836
CAM 532
Camassa–Holm equation 87, 8118, 8512
CAMD 2116
Camera placement 2083
Campania 2399
Campbell–Baker–Hausdorff formula 1033
Campi Flegrei Caldera, Italy 9868
Monte Nuovo eruption 9868
Canadian lynx 2719
Canalization 1920
Canalizing 604
Canards 7605, 7610
Cancellation 6666
Cancer cells
characteristic traits 913
phenotypes 914
Cancer diseases 914
Cancer prediction and prognosis 2146
Cannon, Walter 8267
Canny 5472, 5473
Canonical 1-form 2261
Canonical anticommutation relations *see* CAR
Canonical co-ordinates 2262
Canonical coordinates 5446
Canonical symplectic form 5445
Canonical transformation 2261, 6769
Cantor distribution 7011
Cantor dust 2261, 4515
Cantor family 2261, 4515
Cantor set 2261, 3704, 4515, 6325
biological fractals 3781
Cantor stratification 4515
Cantorus 409
CAP *see* Chiral atom pair
Cap-and-trade 9029
Capability, organizational 7850, 7851
Capacitive coupling 7354
Capacitors 9780
Capacity
bottleneck 3149
dimension 3879
entanglement assisted 7393
full 7846
utilization 1967
Capillary electrophoresis 5580
Capillary number 9788
Capillary waves 9977
Capital 2284
Capital asset pricing model (CAPM) 3391, 3443, 4126, 4128, 4972
Capital intensive 2284
Capitalist sector 2284
Capitalist system 2284
Capture probe 2110
Car-following 9417
Car traffic models 6776
CAR, representations 3078
Carathéodory condition 3266
Caratheodory, Constantin 2860
Carbon 1285, 3285, 3298, 3299
nanotube 1284
polyaniline 1287
polypyrrole 1287
single-walled carbon nanotubes (SWNTs) 1286
Carbon allotrope 690
Carbon cycle 9015
Carbon market 9030, 9032
Carbon nanotube field effect transistors 7480
Carbon nanotubes 690, 1306, 9141, 9146, 9652
beams 692, 698
bundles 696, 700
capped 694
DWCNT 694, 697
DWCNTs 691
dynamic 696
fluid flow 695
gas adsorption 696
mechanical properties 698
MWCNT 691, 692, 694
MWCNTs 698, 699
static 696
thermal properties 692, 700
Carbonic anhydrase 2120

Carcinogenesis 913
metabolism and acidosis 915
Carcinogenic activity 2182
Cardiac physiology
action potential 480
electrocardiographic 479
His–Purkinje conduction 480
QRS complex 479
spectral reserve 479
sympathetic and parasympathetic stimulation 480
Cardinal ratings 9954
Cardinality of a fuzzy set
fuzzy 4065
nonfuzzy 4065
Cardinality, fuzzy logic 3990
Cardiovascular system, biological fractals 3784
Cardy's formula 6529
Carleman estimates 1490, 1493
Carnot cycle 2818
Carrier-envelope offset 6169
Carrier in amorphous computing 265
Carrier sense multiple access (CSMA) 2054
Carrington, Peter 8269
Carrying capacity 7843
CART 1835
Cartan structure equations 5504
Cartan subalgebras 5448
Cartesian product 4879
couple 6301
definition 6287
nonsingular maps 3075
of two fuzzy sets 4069
Cascade algorithm 592
multiwavelet 5844
Cascade-correlation network 1821
Cascade model, earthquake 2541
Cascade operator 592
Cascading 624
Cascading failures 6058
Case-based learning 5195
Case-based reasoning 1792, 1803, 7762
Case memory model 1804
Case representation 1804
Casimir function 2261
Casimir operators 9648
Castaing distribution 8533
Cataclysmic variable 392
Catalysis 3284, 3286, 3290, 3291, 3299
Catalyst 1308, 1314, 1317, 1327, 1329, 7570
activity 1314
heterogeneous 1308
homogeneous 1308
selectivity 1309, 1317, 1329
steam-reforming 1327
Catalytic oxidation of carbon monoxide 7573, 7574
Catalytic polymer, in artificial chemistry 334
Catalyzed reactions 7576
Catastrophe theory 1899, 4634, 8942
Catch of synchronized flow 9374
Catch-recapture sampling 8226
Catchments 7598
Categorical GrC model, eighth GrC model 4347
Categorical unrelated, QSAR modeling 7073
Category 173, 5352
Lusternik–Schnirelman 6601
mathematical systems 6289
Category-exemplar model 1804
Category of Turing machines 4347
Cauchy distribution 3422
Cauchy estimate 5071
Cauchy–Euler approximate solution 3258
Cauchy (or Lorentz) distribution 5221
Cauchy–Peano theorem 3258, 6299
Cauchy problem 3257, 6590, 6660
Cauchy–Riemann equation 8717
Caudate nucleus 8005
Cauliflower
anatomy, biological fractals 3782
fractal dimension 3781
Causal connection 7453
Causal density 1434
Causal dependency 5199
Causal feature selection 5358
Causal-loop diagrams 671, 673
Causal loops 9043
Causal mapping 7836
Causal mechanisms: phase transitions 1639
Causal processes
oscillatory 4610, 4623
time-lagged regression 4623
time-series modeling 4627
Causality 1901, 4609, 9069, 9070

Cause map 1847, 1848, 1853, 1858–1860, 1862
Cavity-QED, quantum computing 7245
Cayley graph 783, 785
Cayley tree 1403, 3856
Cayley's formula 6668
CBC algorithm 596
CBC (coset by coset) algorithm 596
CCC model 4127
CCNOT gate, quantum computing 7379
CCR *see* Convective constraint-release (CCR)
CDF 10013, 10014, 10019, 10020
Ceccherini-Silberstein, T. 780, 784, 786
Čech–Stone compactification 3040
Celestial calculation 5474
Celestial mechanics 372, 399, 6673, 6681
Celestial simulation 5472
Cell adhesion molecules 532
Cell adhesion proteins 7040
Cell-cell adhesion, tumor cell migration 917
Cell cycle 727
Cell cycle control 729
Cell differentiation 525, 529
Cell-ECM adhesion, tumor cell migration 917
Cell fate 1192, 1204
Cell heterogeneity 1205
Cell matrix 931, 5877
Cell migration, ECM influence 915
Cell model 6965
Cell pattern formation 525
Cell problem 6225
Cell processes, tumor invasion 914
Cell-scale models 515
Cell-surface receptor 9751
Cell types 1192, 1193
  attractors 1199, 1203
Cells 923, 2698
  grid 6113
  place 6112
Cellular array 3557
Cellular automata (CA) 63, 65, 66, 71, 116, 164, 188, 249, 752, 755, 760, 870, 892, 895, 901, 914, 916, 928, 1056, 1064, 1065, 2562, 2980, 4240, 4913, 5469, 5864, 5867, 6805, 7155, 7156, 7855, 8047, 8067, 8421, 8428, 9159, 9415
  $\mathbb{B}$-bubble 2983
  $\mathbb{B}$-diamond 2983
  $\mathbb{B}$-posexpansive 2985
$\mathbb{H}$-mixing 2990
$\mathbb{P}$-periodic 2985
χ-eigenfunction 3004
additive 2981
affine 2981
agent-based modeling 158, 186
agent-based simulations 4917
algebraic 2988
algebraic $\mathbb{Z}^D$-action 2989
asymptotic randomization 2993
asymptotically randomizes 2993
asynchronous 750
attractor 2997
average Lyapunov exponents 3006
ballistic annihilation model 2992, 3000
basin 2997
basins of attraction 64
binary Ledrappier 2993
bipermutative 2981
bounded-to-one 2983
Burton–Steif measure 2985
canonical factor 2986
Cayley graph 911
center 2998
Cesàro average 2993
Cesàro limit 2993
character 2994, 3004
cities 7966
classical 7155, 7156
collision rule 2991
column shift 2986
communication in 751
comparison with Turing machines 746
complete 752
complete connections 2988
complete history 2989, 3006
complexity measures 748
computationally universal 7156
conditional measure 2985
cone entropy 3007
cone entropy of 3007
connected 2999
cyclic 2991
cyclic addition and ballistic annihilation model 3001
cylindrical 66

d-dimensional measurable entropy density 3008
d-dimensional topological entropy density 3008
d-frame 3009
d-linear form 3010
d-plane 3008
defect 2991
densely recurrent 2999
densely recurrent subshift 2999
density one 3002
diffusive 2995
dimension 2986
Dirchlet boundary conditions 66
directional entropy of 3006
discrete spectrum 3004
dislocation 2991
domain 2991
domain boundary 2991
dual group 3004
eigencharacter 3004
eigenfunction 3003
eigenvalue 3003
elementary 66, 2991, 3001, 7156
embedding 910
emergent defect dynamics 3011
endomorphic 2988
entropy 9161, 9170
entropy geometry of 3008
entropy theory of 3004
equicontinuity 9170
equicontinuous 9160, 9161
ergodic 3002
essentially equals 2989
evolving 3234, 3236
expansive 2988, 3008, 9171
expansive component 3009
expansive subdynamics of 3008
extremally permutative 3003
full support 2993
game of life 4914
Garden of Eden 70–72, 9167
Garden-of-Eden problem 2983
generator 2987
group shift 2988
grouping 910
growth degree 2986
h-permutativity 2981
Haar measure 2988
harmonically mixing 2995
hexagonal lattice 911
history 909
hyperbolic plane 911
injective 2984, 9160, 9161, 9167
intrinsic universality 906
invariant measures of 2983
irreducible 2990
join 2993
lattice gas 4915
lean μ-attractor 2997
lean μ-center 2998
Ledrappier 2993
left-permutative 2981
left radius 2980
limit measures and other asymptotics 2993
limit set 2997, 9160
linear 64, 2981, 2993
local rule 2980
locally free 2995
Lyapunov exponent 3005, 4915
M-mixing 3002
machines 929
maxentropy measure 2984
(maximum) Lyapunov exponents 3005
measurable dynamics of 3002
measurable entropy 3005
measurable entropy dimension 3008
measure attractor 2999
measure of maximal entropy of 2984
measure-preserving dynamical system 3002
measure rigidity of 2988
minimal 2997, 2998
mixing and ergodicity of 3002
mixing of all orders 3002
modeling 848
μ-attractor 2997
μ-center 2998
μ-limit set 2998
multimixing 3002
multiplicative 2996
mutually erasable pair 2983
nearest-neighbor XOR 2993
neighborhood 2980
nilpotent 9160, 9161, 9169

null boundary conditions 66
odd sofic shift 3001
omega-limit set 2997
one-sided 2980
one-sided generator 2987
P-periodic 2985
particle 2991
particle weight function 3002
periodic 9171
periodic measures of 2985
permutative 2985
permutive 2981
phases 2991
posexpansive 2986
positively expansive 2986
pre-injective 2984
probabilistic 4916
proper 2993
quantum 911, 7155, 7156
quantum-dot 7162
quasi-invertible 3008
radius 2981
realm 2997
rescaling 911
reversible 911, 7156, 7157, 9160, 9161
right-permutative 2981
right radius 2980
rule 64
scalar coefficients 2981, 2993
self-timed 8082
sensitive 9170
σ-minimal 2997
signal subshift 2992
simulation among 909
small universal 909
sofic 2984
space complexity 746
space-time diagram 902
spectral group 3004
spectral properties of 3003
stable phases 2991
state change complexity 750
(strongly) mixing 3002
subgroup shift 2988
support 2995
surjective 2983, 9160, 9161, 9167
surjunctive 2984
time 746
toggle automata 2981
topological entropy 3005
topological entropy dimension 3008
topologically ergodic 3002
topologically transitive 3002
totally ergodic 3002
totally σ-ergodic 2989
tree 752
triangular lattices 911
universality 909
v-directional μ-entropy 3006
v-directional topological entropy 3006
vacuum configurations 3002
velocity 2992
weak* converges 2993
weakly mixing 3002
weakly σ-mixing 2997
well 2998
Cellular automata (CA) applications 64
biological pattern formation 64
cryptography 65
deterministic pattern generation 65
discrete reaction-diffusion systems 64
disease processes 64
disease transmission 64
DNA sequences 64
error correcting codes 65
forest fires 64
gene interactions 64
growth and aggregation processes 64
image processing 64
language and pattern recognition 64
parallel computers 64
parallel multipliers 64
prime number sieves 64
pseudoassociative memory 65
pseudoexhaustive pattern generation 65
pseudorandom sequences 65
signature analysis 65
social interaction networks 64
sorters 64
spin exchange 64
VLSI logic circuit design 65
Cellular automaton (CA) 610, 755, 768, 800, 803, 2982, 3158, 3233, 3234, 3538, 4497, 4739, 5024,

5438, 5874, 6772, 7156, 7548, 7679, 7680, 8028, 8375, 8869, 9250
additive 757
asynchronous 5877
in artificial chemistry 335
injective 7680
invertible 7680
modeling, tumor invasion 913
monotone 4497
neighborhood 3233, 3234
nilpotent 758
omnivorous 4500
partitioned 759, 7679, 7682
quantum dot 5875, 5876
reversible 758, 7679, 7681, 7710
self-reproducing 7694
self-timed 5879
solidification 4497
STCA 5879
supercritical 4500
surjective 758
synchronously 5877
threshold growth (TG) 4499
totalistic 4499
4-tuple definition 916
two-dimensional 3240
universal 7679
universal reversible 7683, 7689
with block rules 7679, 7681
Cellular automaton (CA) on a group 779, 781
G-equivariant 781
game of life 782
injective 783
Laplacian (discrete) 782, 786
linear 779, 780, 786
local defining map 779–781
Markov operator 787
memory set 779–781
minimal memory set 781
on $\mathbb{Z}^2$ 780
pre-injective 779, 783, 786
surjective 782
Cellular computing 922, 923, 925, 8869, 8878, 9706
Cellular engineering 258
Cellular functions 720
Cellular maps 5701
Cellular models 6959
Cellular network 554, 2984, 5699
Cellular neural network 930
Cellular programming 3233
Cellular programming algorithm 3238
Cellular, hyperbolic 752
Cenozoic climate 13
Center 241, 2261, 4523
area 241
Center-focus problem 8640
Center manifold 392, 631, 6184, 6335, 6337
construction of 942
for infinite dimensional flows 947
in discrete dynamical 943
theory 8931
Center of gravity method 4884
Center-of-sets (COS), type-2 fuzzy logic 4016
Center-saddle bifurcation 2261, 4521
Centered moving average (CMA) method 3767
Central charge 8711
Central dogma 418
Central executive agent 1109
Central force 2829
model 6556
networks 2841
Central limit theorem 965, 2874, 2876, 3318, 6993, 8712
stochastic processes 3579
Central pattern generator (CPG) 1544, 1553
Central place theory 7957
Centrality 8237, 8271, 8275, 8309, 8311
betweenness 2731
closeness 2731
degree 2731
hierarchical criteria 4832
information descriptors 4832
patterns of complexity 4832
Centralization 8237, 8271, 8311
Centralizer 6656, 6766
Centroid method 4884
Centroid moment tensor 2488
solution 9593
Centroid type-2 fuzzy logic 4016
Cerebral cortex
neural 1557
phenomenological modeling 1557
CFL criterion 7896

CGE *see* Computable general equilibrium
CGL equation 6470, 6471, 6474
CGPS networks 4253
Chacón maps 3061
Chain 9230
Chain components 9234
Chain configuration change 6820
Chain dynamics, single 6819
Chain elasticity, polymer physics 6815
Chain letters 3216
Chain rule 1819
  differentialbility 6298
Chain, ideal non-interacting 6813
Chalcogenide glasses 2832
Chan–Vese model 10039
Chan–Vese region-based active contour 10039
Chan–Vese's active contour without edge 10039
Chances, detection 3577
Change
  environmental 7832
  in regime 9845
  in regime, statistically identifiable 9846
  of variables 6660
Change order 1849
Change-point models
  CUSUM 2573
  hidden Markov models 2574
  stress-shadowing 2573
Change-point problems 9845
Channel additivity conjecture 7393
Channel modes 7664
Channel superconductivity 8493
Channels
  ionic 6101
  multiple use 7391
Chants 2314
Chaos 476, 1045, 2439, 2705, 3055, 3506, 4515, 6132, 6134, 6428, 6637, 8028
  attractor 483
  controlled chaotic attractor 483
  correlation integral 484
  deterministic chaos 484
  low-dimensional chaotic 484
  noisy 3509
  probabilities 4907
  stochastic processes 3586
  strange attractors 483
Chaos and fractals, applications of 3829
Chaos diagram 2423
Chaos theory 8942
Chaotic attractors 6629
Chaotic behavior
  networks 1201
  of the solar system 6445
Chaotic cities 7964
Chaotic diffusion 389
Chaotic dynamics 991, 2331, 2422
  statistical theory 6940
Chaotic galactic orbit 414
Chaotic hypothesis 989, 996, 999
  entropy production 6953
Chaotic in the sense of Devaney 954
Chaotic in the sense of Li and Yorke 955
Chaotic motions 2423, 6442, 6447, 6663, 6666
  particles 5961
Chaotic orbits 6428
Chaotic phase 6775, 6776
Chaotic region 6427
Chaotic rotation 6446
Chaotic systems 989, 1007
Chaotic vibration 2422
Chaoto-chaotic phase transitions 7995, 7998
Chapman–Kolmogorov equation 5250, 9292
  stochastic processes 3580
  stochastic time series 3582
Character-centric games 1346
Characteristic
  height 6248
  level 6248
  Segré 6248
  Weyr's 6248
Characteristic curves 9986
Characteristic exponents 6629
Characteristic factor 2949
Characteristic frequency 6962
  earthquake 2664
Characteristic function 1510–1512, 1520, 4061, 4877, 6260
Characteristic function form 1520
Characteristic length scale 3881
Characteristic multiplier 8637
Characteristic path length 1431
Characteristic polynomial 6836
  of the line graph 6848

Characteristic time 6966
Characteristic time scale 6957
Characteristic vectors 9950
Characterization, dynamic game 2220
Charge density waves 1128, 6414, 7586, 9681
Charged polymer 6817
Charges, non-abelian 3876
Chart of the nuclides 4825
Chartist 96
Cheap talk 1591, 8127
Chebyshev inequality 6935
Checkerboard lattice 6531
Checkerboard model 125
Checkerboarding 1697
Chelungpu fault, Taiwan 4261
Chemical bond 1308
Chemical communication, acetaldehyde 483
Chemical descriptor research, QSAR modeling 7074
Chemical differential equation 329
Chemical dimension 3702
Chemical direction sensing 1365
Chemical distance 3702, 3853
  sensing 1367
Chemical evolution 338, 3280
Chemical exclusive OR gate 1362
Chemical frequency filter 1364
Chemical gels 6545, 6547
Chemical gradient diffusion 259
Chemical graph theory 4827
Chemical logic AND gate 1361
Chemical logic OR gate 1361
Chemical media 9705
Chemical memory 1368
Chemical neuron 1371
Chemical potential 8580
Chemical reactions 1307, 3365, 3368, 8940
  elementary step 1310
  mechanism 1308
  products 1308
  reactants 1308
Chemical signal diode 1358
Chemical structure, multiple descriptors 7072
Chemical system, logic of 330
Chemical three-valued logic 1373
Chemical toxicology, QSAR models 7081
Chemical vapor deposition method 691
Chemical wave 886
Chemical work, ecological systems 2698
Chemisorption
  electronic effects 1319
  energy optimum 1314, 1316
  Hammer–Nørskov chemisorption model 1317
Chemistry, matrix-multiplication 332
ChemNet 2151
Chemotaxis 720, 729, 3349
Chen–Fliess series 1035, 1037
Chenciner–Montgomery symmetry group 5973
Chenciner–Venturelli Hip–Hop 5967
Chern class 2261
Chess 4139
Chess games 2098
Childhood 1903
Chinese whispers 3217
Chip-firing 8035
Chiral atom pair (cAP) 7075
Chiral factor, QSAR modeling 7075
Chirality descriptors, QSAR modeling 7074
Chirality indices, QSAR modeling 7075
Chirality-sensitive atom pair descriptors, QSAR modeling 7075
CHNC 8583
Choice 7753, 7754, 7765
Chomsky 1878, 1883
Chomsky hierarchy 925
  language family 808
Choquet integral 231
Chordal SLE 8709, 8724
Chords 6514
Choreography
  n-body problem 5959–5961
  non-planar 5969
Choreography symmetry 5973
2–1-Choreography symmetry 5972
Choroplasts 2698
Chromatin 3894, 3896, 3917, 3918
Chromatin immunoprecipitation (ChIP) 5661
  detection by micro array chip 3970
Chromatin modification 1195, 1196
  epigenetic marks 1195
Chromatographic separation 859
Chronic illness 4577
  asthma 4577
  depression 4586

diabetes 4577
disability 4578
disease management 4579
heart failure 4578
HIV/AIDS 4577
prediabetes 4579
premature mortality 4578
prevention 4578
risk factors 4583
Chronological algebra 1028, 1029, 1033
free 1033
Chronological calculus 1028, 1029, 9105
Chronological exponential 1031
Chronological identity 1033
Chronological logarithm 1032
Chronological product 1033
Chronology 3865
Chronotopoi 9412
Chuang 7256
Church–Turing 271
Church–Turing computable 3354
Church–Turing thesis 287, 925, 7089, 8191
quantum strong 7089, 7090, 7106, 7110
strong 7089
Cilia 1739, 1741, 1742, 1748, 1759–1761, 1769, 1770, 1774
Cilium 1761
CIMMSIM 127
CImmSim 163
Cinder cone 9821
Cirac–Zoller gate, quantum computing 7233, 7234
Cirac–Zoller scheme, controlled Z-gate 7233
Circadian clock 729, 731
Circadian rhythms 485
Circle group 5309
Circle method, biological fractals 3788
Circuit 6031, 9946
asynchronous 5868
delay-insensitive 5869, 5878
logspace uniform 6403
quasi-delay-insensitive 5869
self-timed 5869
speed-independent 5869
two-qubit 7380
Circuit diagram, quantum Fourier transform 7241
Circuit model of quantum 7090
Circulant matrices 68, 69
Fourier matrix 69
generalized circulant matrices 74
minimal annihilating polynomial 72
properties of 69
Circular and cumulative causation 8996
Circular causality 1538, 2853, 4897
Circular crack 2591
Circular ensembles 7507
Circular fault 2590
Circular orbits 6436, 6447
Circularity 4898
Circum-Antarctic seaways 13
Cis-regulatory elements 9094
Cities and complexity 7954
City 1047, 7953
as representation of society 7958
game 7974
planning 7953
size distribution 1049, 1051
Civilization 3281
C^k Conjugacy 3096
$C^{k,l,p,r}$ Locally rigid 3102
Claim 1859, 1861, 1862
Clarke mechanism 5514, 5519
Class
basic 6246
distinguished 6248
final 6246
initial 6246
of non-stationarity 1618
operator 4415, 4420
structure 8184
Classical augmented infinities 6289
Classical categories 7954
Classical computational complexity 7120, 7399
Classical control engineering 8952
Classical control theory 8954
Classical electron oscillator 6121
Classical information transmission quantum channels 7392
Classical integrable systems 5081
Classical-map hyper-netted chain 8576, 8582, 8583
Classical master and Langevin equations 3365
Classical MD simulation 694

Classical MD simulation method 692
canonical ensembles 695
constant-temperature 695
equilibrium 700
non-equilibrium 700
Classical Monte Carlo simulation method 692
Classical planning 5189
problem 5192
Classical solution 6218
Classical subsolution 6220
Classification 756, 760, 764, 1826, 1904, 4362, 4401, 7753–7755, 8243
and regression trees (CART) 2167
degree-based 764
methods 1799
modeling 7073
system 7803
theorem 713
Classifier 1791, 1826
Clausius 2860
Clean value, reduction theorem 5489
Cleaning 3683, 3690
Clear Lake region 7845
Clearing and settlement 5407
Clebsch–Gordan transformation 7113
CLF *see* Contour length fluctuation (CLF)
Client–therapist interaction 7996
Client, mobile system 5606
Climate 7592
climate simulations 7594
key parameters 1096
records 3713
Climate change 1071, 7601
abrupt 1097
human-induced 7845
impact on human health 1095–1097, 1099
impacts on plants 1102
indirect health effects caused 1101
migration 1102
natural 1097
vector-borne tropical infectious diseases 1101
Climate feedbacks 9015
Climate modeling 3713, 9023
Climate simulation 2
Climate statistics 3
Climate time scales 1
Climate variability 6
Clinical psychology 7991, 8006, 8947
CLIQUE 5058
Clique percolation 1156
Cliques 8253, 8267, 8271, 8275, 8287
Clock 5477, 8200, 8207
resonant 5868, 5870
Clock gating 5868
Clogging 3144
Clonal selection 4780
CLONALG – clonal selection algorithm 2128
Cloncurry 3128
Cloning 420
Close binary systems 392
Close encounter 8696
Closed loop 7054
Closed-loop control 632
Closed loop equation 1504
Closed-loop thinking 4565
Closed-world assumption 5191
Closeness 8271
centrality 8287
Closing CA 9253
Closing conditions 3100
Closure 8235
of sets 338
problem 3649
properties 800, 813
syntactic and semantic 341
Clouds 3700
CLR *see* Common language runtime
Club networks 6031
Club of Rome 8409
Clubs 4097
Cluster 610, 1598, 3849, 6382, 6538, 6541, 6542, 6566, 6669, 6784, 7822, 8247
average number 1598
density of sites in the infinite cluster 1598
finite 3708, 3849, 3850
fractal dimension 1599
in amorphous computing 266
infinite 3708, 3849, 3850
mean cluster size 1598
spanning 610, 3850
Cluster analysis 1235, 1991, 9851
distance metric 9851
hierarchical clustering 1236

model-based hierarchical clustering 1236
statistic 9851
Cluster assignment variable 505
Cluster expansions 8595
Cluster state 631, 632, 635, 7257
Cluster variables 631
Clustered samples
sample stratification 5569
social interactions 5570
spatial dependence 5570
survey sampling 5570
Clustering 730, 1232, 1431, 1778, 2351, 3772, 4362, 4401, 4489, 4663, 5415, 6005, 6053, 8247, 8674
algorithms 2356
coefficient 575, 725, 730, 2531, 2534, 2535, 2537, 4490, 5659, 6053, 8235, 8254, 8284, 8330, 10061
effect 7504
functional network 507
network density 6053
of extreme events 3713
procedure 8261
weighted 6057
CMOS 9706
CMT 9587
Cnoidal wave 8113, 8116, 8118, 8123, 9970, 9988, 9989
CNOT-gate, quantum computing 7236
CNOT quantum gate, truth table 7234
CNOT-wavepacket gate, quantum computing 7236
Co-sets 5546, 5548
CO_2-concentration, elevated 1102
Coadjoint orbit bundles 5501
Coadjoint orbits 5447, 5487
as symplectic reduced spaces 5490
cotangent bundle reduction 5501
Coalescence 9148, 9149
Coalition 1509–1511, 1514, 1515, 1520
losing 1522
minimal blocking 1522
minimal winning 1522
of constraints 1559
winning 1522
Coarea formula 10038
Coarse coding 282, 3344, 3351
Coarse-grained 3929, 7034
Coarse-grained dissipation 8529, 8549
Coarse grained observation 993
Coarse grained techniques 5917
Coarse graining 1006, 2368, 3367, 8711
Coarsening 6467, 6472
Coastline 3700
Coating 1296
COBE/CMB bound 1636
Coboundary cocycle 3067
Cobweb model 8994
Cocycle identity 5485
Cocycle superrigidity 3104
Cocycles 3099, 8555
affine 8560
dynamical systems 3067
momentum 5485
nil 8560
orbit classification 3068
Rokhlin 8563
weak equivalence 3067
COD 5061
Coda attenuation factor 7915
Coda interferometry 7927
Coda waves 7914
Code
mobile 5606
obfuscation 5612
on demand 5607
signing 5612
Codebook 1782
Codimension 703, 704, 713, 714
codim(f) 713
Coding DNA, fractal features 3797
Coefficient
assortativity 6049
clustering 6049
Coercivity 6601
Coevolution 3233, 3239, 4136, 8242, 8367, 8370
applications 3240
host-parasite 3239
of networks and behavior 6008
Coexistence 6626, 6631, 6633
CogCity (cognitive city) 7975, 7979
Cognition 1112, 1113, 1225, 1899, 1901, 2243
connectionist approach 1114

hybrid approach 1116
symbolic approach 1113
Cognitive 2845
Cognitive agents *see* Deliberative agents, 1338
Cognitive architectures 1112
Cartesian theater 1109, 1120
connectionist 1112
global workspace 1109, 1120
hybrid 1112
intelligent machine architecture 1121
symbolic 1112
Cognitive development 1903
Cognitive dissonance 2244
Cognitive dynamics 7543, 9415
filtering beliefs 7543
filtering goals 7544
Cognitive immunology 4782
Cognitive map 1853, 1859, 4898
Cognitive processes 8406
Cognitive robotics 1110, 1117, 1225
anthropomimetric robotics 1118
Braitenberg vehicles 1118
cognitive robot definition 1110
COGRIC 1110
ethics 1124
morphological computation 1118
Cognitive science 1347, 4403
Cognitive state 9415
Cognitive system 1889
Cogwheel 5476
Cohen, L.J. 6928
Coherence 1437, 2247, 3216, 3218, 6117
quantum computing 7238
Coherency condition 431
Coherent (effective) potential 8596
Coherent (laser) light 6388, 6390, 6391, 6395, 6396
Coherent optical wavefront 6390, 6394, 6398
Coherent states 3352, 3353, 3367
Coherent structures 394, 8523, 8526
Alfvénic 8527
Coherent systems 6391
Coherent type I motif 563
Coherent wavefronts 6395
Cohesive resistance 4627
collective resistance 4614
resistance 4626
resistant cohesion 4615
resistive groups 4626
Cohesive subgroup 8234, 8287
Cohomological functional equation 5458
Coil-collapse transition
block co-polymer 6830
polymer physics 6816
Coiled coil 9752
Coincidence detector 1364
Cointegration 2758, 3444
Cointension, fuzzy logic 3996, 4004
Cointensive precisiation, fuzzy logic 3994
Colbrar, Charles Xavier Thomas de 5476
Collaboration 9098
Collaborative cross 9118
Collapse
bubble 6956
of a bubble 6969, 6970
Collapse of the wave-function 7173
Collective behavior 1282, 6492
Collective choice problem 9935
Collective dynamics 6476
Collective event 5022
Collective excitations 1272
Collective firing 631
Collective intelligence 1281, 6483
Collective level of description 5721
Collective motion 610
Collective objects 4921
Collective phenomenon 6486
Collective pinning 2032
Collective transport 1127
Collective variables 1538, 1542
College admissions problem 4097
Collision-based reaction-diffusion computing 9720
Collision term 880
Collisional
cooling 4381
system 8696
Collisions 879, 880, 9413
asymptotic estimates 5979
between two or more particles 5961, 5978
explicitly simulated 330
of molecules 327
Collisions instants, isolatedness 5979
Colloid aggregation 3721

Colloidal suspensions 4215
Colloids 2025
Colony 8883
Comb model 310, 313
Combination 5741
  in amorphous computing 266
Combinatorial explosion 8736
Combinatorial generator 1891
Combinatorial library 2117
Combinatorial optimization problem 9950
Combinatorial QSAR 7079
  methodology, flow chart 7079
Combinatorial space-time 8844
Combinatorics 3055
Comfort control 4885
Commensurate density 9680
Commercial 2284
Commercial customers 5409
Commodity cycle models 678
Common agency 6983
Common knowledge 427
Common language runtime (CLR) 5607
Common-pool resources 4864
Common prior 428, 429, 435
Common prior assumption 1592
Common structure semigroup 8301
Common value 9935, 9937, 9940, 9941
Communicate 6245
Communicating finite state machines 165
Communication 1961, 3109, 3111, 3136, 4104, 6492
  about intentions 8127
  action statement 3117
  barrier to change 3125
  between particles 267
  channel, quantum information processing 7369
  complexity 7716
  core goal 3117
  electronic mass media 3114
  equilibrium 1588
  indigenous community 3111
  information flow 3117
  institutional barrier 3125
  integrating theory 3119
  language 3117
  make the threat real 3119
  mechanism, amorphous computing 263
  media 3117, 3119
  media support 3115
  medium and message 3114
  meteorological knowledge 3117
  non-English speaking household 3122
  paradigm shift 3126
  phone tree disaster warning model 3122
  policy and laws 3116
  politics 3119
  precautionary principle 3118
  risk communication theory 3132
  safety group 3126
  safety triangle 3111, 3123, 3126, 3138
  simulation 3114
  siren 3119
  target audience 3117
  web 3124
  world view 3117
Community 1141, 3110, 3131, 3135, 3136
  awareness 3110
  cultural knowledge base 3136
  dispersal limited 2706
  effect 529–531
  empowerment 3111, 3138
  hierarchy 1145
  immigration rate 2707
  indigenous 3120
  passivity to partnerships 3116
  personal contact 3124
  resilience 3114
  safety group 3130
  self-help 3131
  seven steps to community safety 3124
Community definition 1143
  global 1144
  local 1144
Community structure 1142, 2741, 2903, 6054, 8915, 10066
  modularity 10066
  non-overlapping 6054
  overlapping 6056
  species abundance 2903
  species abundance distributions 2913
  species richness 2903
Commutative-group, definition 6288
Commutative-semigroup, definition 6288

Commutator 4590
Compact 5029
Compact visitation 312
Compactification 3869
Compaction length 2685
Compactivity 5008
Compacton 81, 84, 88, 90, 8458, 8506
Comparative evaluation of prediction algorithms (CoEPrA) 2191
Comparative genomics 4173
Comparative molecular field analysis (CoMFA) 2209
Comparator 278, 280
Comparison principle 6221, 6222, 6224
Compartmentalization 8382
Compartments 2714
Compatibility condition 4963
Competing phase 9680
Competition 8040, 8771, 9701
  imperfect 7837
  model 874
Competitive advantage 7832
  sustainable 7851
Competitive equilibrium 1512
Competitive learning 4873
Competitive market 4108
Competitive (Walrasian) equilibrium 4108
Complement 4877, 4878
  of a fuzzy set 4066
Complementary 1539
Complementary cumulative distribution 7011
Complementary nature 1559
Complementation 4148
Complete basis 5462
Complete graph 8392
Complete integrability 6767
Completely dissipative 5459
Completely integrable 5447, 8113, 8114, 8641
Completely integrable equations 8113, 8114
Completely revealing strategy 7631
Complex 7837
Complex adaptive system 1877
Complex amplitude 9916
Complex behavior 5024
Complex data 4342
Complex dataset, QSAR 7072
Complex dynamics 2971, 3020
Complex energy landscapes 7586
Complex frequencies 9916
Complex Ginzburg–Landau (CGL) equation 6469
Complex GRN 1180, 1181
  Boolean networks 1197
Complex instability 6435
Complex natural systems 554
Complex networks 1191, 1244
  subgraphs in 5694
Complex organizations
  management 4564
  planning 4565
Complex social behavior 7579
Complex social system 6492
Complex systems 160, 235, 477, 2857, 3268, 4900, 6049, 7009, 7019, 8711, 8928, 8946
  approach 477
  as networks 5692
  non-Hamilton 1624
Complex topology polymer 6818
Complex wavelet 8675
Complexity 1424, 1805, 1980, 3276, 3859, 4098, 4101, 4821, 5740, 7064, 7089, 8040, 8711, 8943, 9105
  algorithmic 5206, 5207, 5213
  biochemical 1304
  biological 489
  biological sequence 494
  bounds 4104
  classes 9955
  computational 5740, 7089
  computational chemistry 1302
  conditions for 180
  costs 4112
  definition of 8372
  ergodic theory 6944
  evolution 505
  functional 491
  genome sequence 492
  in biological systems 5693
  in scale 3576
  in time 3575
  measures 1429, 4822, 6398, 6401, 7713
  mechanisms 8368, 8371
  metric 3277
  modeling 5377–5379, 5382, 5383, 5387
  molecular 498, 1302

of computation 4101
of contracting 4112
of implementation 4101
polylogarithmic 6403
polynomial 6402
QSAR based 7071
query 7089, 7091
reduction 4899
research 8007
resonance diagram 7999, 8000, 8003
sample 5206, 5207, 5211–5213, 5217
science (CS) 177, 9412
stochastic processes 3574, 3575, 3583, 3586–3589, 3591–3593, 4513
structural 491
theory 288
time 6398
topological 1302
Compliant motion planning 5473
Complicated 7837
Component separation 1719
morphological component analysis 1733
Components 4492, 6044, 6497, 8271, 8275
giant connected 4492
strong 8250
weak 8250
Componentwise egalitarian 6045
Composite error 2772
Composite-fermion 8586
Composition 4880
of relations 8293, 8295
operator 4871
Compositional complexity
DNA sequence 495
Compositional rule of inference 4070, 4875, 4882
Compositional strand asymmetry 3894
Compositionality 5286
Compressed sensing 10044
Compressibility 6960
Compressible 1719
Compression 699, 10032
Compressive sampling 10044
Compton scattering 6664
Computability theory 925
Computable 5467
Computable general equilibrium (CGE) 177
Computation 285
continuous 272, 276
discrete 272, 276
mobile 5605
quantum information processing 7365
reversible 5869
Computation at the edge of chaos 3237
Computation in CAs 3235
Computation of local travel times 9184
Computation theory 3234
Computation-universality 7694, 7697, 7705, 7709
Computation via CA particles 3237
Computational agent
BDI agents 7542
classifier systems 7538
genetic algorithms 7537
reinforcement learning 7538
strong 7536
weak 7536
Computational biology 923, 1227, 9094
Computational chemical toxicology, QSAR models 7081
Computational complexity 925, 7119, 8603, 8605
Computational complexity class 5468
Computational complexity theory 6402
Computational equivalence 763
Computational intelligence 4776
Computational maps 3341, 3350, 3353
Computational mechanics 4599
Computational model 4783
Computational particle 259
Computational problem 5468
Computational processes, agents as 188
Computational task 3235
Computational web intelligence 4408
Computer-aided drug design (CADD) 2141, 2161, 2210, 2211
Computer-assisted 6682
Computer-assisted proof 5082, 5088
Computer game(s) 1335
Computer generated imagery (CGI) 1336
Computer graphics 1335
Computer science 417, 923, 4216, 4894
Computer security 4342
discretionary access control model 4343

Computer simulation 4217, 7034, 7992, 8375, 8421, 8427, 8428, 8432, 9073, 9074, *see also* Simulation
as method 9073
of cultural evolution 3214
Computer systems, mobile 5605
Computer vision, applications 3240
Computerized tomographie
backprojection 9178
iterative solutions 9178
Computerized tomography
analytical solutions 9175
filtered projection 9177
historical overview 9173
parallel projection 9174
Computing
amorphous 257
noise driven 5870
paintable 262
pervasive 269
reversible 5875
Computing cell, growth 262
Computing cube 926
Computing predecessor states 73
Computing scenarios, process of 7850
Computing with solitons 1376
Computing with words (CW) 3986, 4005, 4071, 4362
Comte, Auguste 8380
Concavification 7634
Concavity 2861
Concentration control coefficients 559
Concentration expansion 8593
Concentric KdV equation 9995
Concept approximation 4352, 4359
Concept approximation space 4352
Concept dependent granulation 4428
Concept detection 5742
Concept lattice 8295
Conceptual model 15
Conceptual structure, fuzzy logic 3988
Conceptualization, context for 7850
CONCOR 8359
Concurrence relation, non standard analysis 6304
Concurrent Poisson process 6995
Concurrent relationship 4666
Condensation 4492, 8250
Condition
for optimality 1488, 1503
necessary 1502
Palais–Smale 6607
weighted Palais–Smale 6607
Condition A 6765
Condition E 5844
Condition ω 6765
Condition $\omega^{\#}$ 6767
Condition $S(\theta)$ 3532
Conditional complexity 493
Conditional entropy 495
Conditional expectation 5463
Conditional independence 2783
Conditional Kolmogorov complexity 493
Conditional measures 5463
Conditional phase shifter 1017
Conditional possibility distribution 6933
Conditional probability 495
of a fuzzy event 4072
Conditional variance 3440
Conditionally periodic 2262, 4516
Condorcet 9944
Condorcet jury theorem 9934, 9937, 9940
Condorcet–Kemeny problem 9949, 9956
Condorcet winner 9943, 9944, 9949
Condorcet's paradox 9935, 9937
Condorcet's procedure 9956
Conductance 8600
Conductance fluctuations 7172
Conductance matrix 3740
spanning tree polynomial 3741
Conductivity 6535, 7826
Conductivity exponent 1407, 1410, 2831, 3889
Conductivity (t) and superconductivity (s) exponents 1411
Conduit 9763, 9766, 9785, 9881
Cone field 4728
Confidence 5206, 5207, 5211, 5212, 5216
Confidence matrix 8397
Configurable logic blocks 930
Configuration
almost equal 783
almost periodic 761
finite support 761
initial 782
periodic 761, 785

recursive 761
rich 250
sufficient 4498
support of a 786
Configuration model 6052
Configuration reachability 169
Configuration space 5784, 6589, 6597, 6600, 6602–6604
Configuration space and the shift
alphabet 2980
Cantor space 2980
configuration space 2980
configurations 2980
shift map 2980
Configurational averaging 8594
Configurational entropy 4213, 5034
Configurational information 4827
Confinement potential 7310
electrostatic confinement potential 7313
Confining potential 8576
Confirmation 9067
Conflict
sociopolitical instability 4622
spatial population structure 4615
Conflict monitoring system 8005
Confluence 761
Conformal field theory 3731, 7347, 8711, 8713, 8726
Conformal invariance 8709, 8713, 8717, 8718, 8722, 8728
Conformal maps 8713, 8717
Hastings–Levitov scheme 3723
Laurent coefficients 3730
Conformal property 8721
Conformal repeller 6952
Conformal transformation 8709, 8713, 8717, 8723
Conformist pressure 3209
Conformity 3216, 3218
Confusion
conceptual 7836
operational 7836
Confusion matrix 1998
Congested patterns 9408
Congested traffic 9283, 9302, 9355
Congestion 6059, 9417
congestion-awareness 6059
congestion factor 6060
jamming transition 6059
Congestion games 3180, 3187
Conical intersection 6692
Conjecture
Littlewood's 3030, 3033
Oppenheim 3033
Oppenheim conjecture 3030
Conjugacy 3096
Conjugacy class 6659
Conjugate co-ordinates 2262
Conjugate gradient 5755
Conjugate Hamiltonian 4544
Conjugate pairing rule 1003
Conjugation function 1934
Conjunction 4879
Connected graphs 5268
Connected hole density 619
Connectedness 1889, 1905
length 1599
Connecting orbits 6347
Connectionism 924
Connectionist AI 1109
Connectionist approach
connectionist cognitive machine architecture 1115
synthetic neural modeling 1116
Connectionist learning 1822
Connections 4830
synaptic 639, 641
Connective stability 2053
Connectivity 1449, 2737, 6566
geological fractures 6505
pathological 628, 629
Connectivity graph 724
Connectivity index 6838
Connectivity matrix 6837
Connectivity theorems 560
Conscious 1556
Consciousness 1424, 4893, 8184
conscious content 1425
conscious level 1425
discrimination 1425
hard problem 1439
higher-order consciousness 1425
primary consciousness 1425

Consensus 7725, 7839
formation 8387, 8389, 8397
prediction, QSAR models 7083
QSAR models 7082
time 8387
Consequences, profound 7849
Consequent 4871
Conservation 1904
Conservation equation 5577
Conservation laws 882, 4382, 4713
n-body problem 5979
Conservation of mass 5541
Conservation understanding 1898
Conservative 5459
Conservative and progressive forces 1896
Conserved moieties 558
Conserved quantities 4715, 6182
Consistency 5183
assumption 435, 440
relation 5744
Consistent algorithm 5212, 5213
Consistent belief 435
Consistent BL-subspace 437
Consistent state of the world 436
Consonant 2310
Constant
cosmological 3868
of motion 6182
Constant proportion portfolio insurance (CPPI) 94
Constant relative risk aversion (CRRA) 96
Constitutive equation 4382, 4383
fluids 5929
liquids 5929
Constitutive relationship 2681
Constrained Pareto efficiency 6041
Constrained variable 3998
Constraint-based analysis 520
Constraint-based modeling 521, 5541, 5545
Constraint-based modeling approaches 5537
Constraint-based planning 1347
Constraint counting 2835
Constraint release (CR), entangled polymer dynamics 6824
Constraints 8239
fuzzy logic 3997, 3998
on multiple levels 1558
Constructibility 809
Construction universality 5879
Constructionism 207
Constructive system 8040
Consumer sovereignty 1671
Contact 5012
Contact angle 5584
hysteresis 5583
Contact map 7035
Contact tracing 4667
Contagion 8307, 8309, 8312–8315, 8317
Containment behavior 2057
Contemporaneously exogenous 2772
CONTENT 6349
Context data 4871
Contextual independence 1867
Contingency matrix 1865
Contingency table 1865, 2445
attribute 1865
conditional probability 1867
determinantal divisor 1869
elementary divisor 1870
elementary transformation 1870
rank 1866, 1867
rough set 1865
statistical independence 1867
subdeterminant 1868
submatrix 1868
Continuation 6332, 6590, 6608
Moore–Penrose 6334
of an orbit 6594
pseudo-arclength 6333
strategies 4100
Continued fraction 3042
expansion 5461
Continuity 1683, 1898, 1899
Continuity equation 881, 3614, 5576, 8106, 8108, 9971
Continuous deterministic 555
Continuous feedback 8621
Continuous function 3256, 6295
Continuous problem 7119
Continuous space machine (CSM) 6396
C_2 6400
Continuous spectrum 7165
Continuous stochastic process 6228
Continuous t-norm 4416

Continuous time 277, 3524
processes 3445
random walk 5222
stochastic volatility models 8786
SV jump–diffusion model, quadratic return variation 8787
Continuous wavelet transform 6353, 8671
Continuum 5576, 6567, 7666
asymmetric 2641
hypothesis 9728, 9971
limit 3365, 3367, 8709, 8718
mechanics 698
media approximation 8588
percolation 6507, 6508
robot 1225, 1475–1477, 1479, 1483
surface 1477
thresholds 6508
Contour length fluctuation (CLF), entangled polymer dynamics 6824
Contour of a convex image 4747
Contraceptive 8309
Contraction 2936, 3861, 8247
principle 5300
Contracts 6977
incentive 6978
long-term 6982
menu of 6984
short-term 6982
Contrarians 8388
Contrast enhancement 1719
Control 1486, 1488, 4671, 8617
boundary 1505
deterministic 6219
hierarchical 6068
hybrid 4671, 4674, 4678
juggling 4675
linear convex 1505
linear quadratic 1504
mobile robot 4690
networked 4679
optimal 1488, 1502
piecewise constant 6374
positional 1488, 1502
reset 4680
sample-and-hold 4678
stochastic 6219, 6220
supervisors 4684
temperature 4672
Control coefficients 559
Control function 1488
Control input 3524
Control-Lyapunov function 8624
Control node 155
Control parameter 1538, 5718, 7994, 8000, 8001, 8004, 8430, 8431, 8927, 8928
internal parameter 8430, 8432
internal states 8423, 8430, 8431
Control parameter α 5590
Control qubit
CNOT-gate 7236
Control space 713, 1488
Control surface 4887
Control system 284, 290, 3524, 4806
trajectories 5440
Control systems 289
optimal 5441
reachable sets 5440
Control theorems 560
Control theory 1542, 9014
Controllability 1027, 3524, 4811, 6372, 8619, 9104
approximate 1488, 4809
Burgers equation 4815
exact 4811
heat equation 4809
linear partial differential equations 4812
linear systems 4807
nonlinear systems 4814
one-dimensional heat equation 4813
regional null 1506
return method 4816
semilinear wave equation 4814
stabilization and optimal control 6217
wave equation 4808
zero 4811
Controllability Gramian 3527
Controllability to zero 4810
Controllability/observability duality 4811, 4812
Controllable 3526, 3531
Controllable along a trajectory 3529
Controlled interventions 7062
Controlled Lagrangians 2015
Controlled-NOT, CNOT 7329
Controlled-sign 7254

Controlled Z-gate, Cirac–Zoller scheme 7233
Controller, interval type-2 fuzzy logic systems 4015
Controlling feedback 2284
Convective constraint-release (CCR) 6824
Convenience samples 8225
Convergence 6177, 6189, 6761
  trajectories 1203
Convergent evolution 5662
Conversion relations 2476
45-Convex 4747
Convex combination, fuzzy logic 3992
Convex conditions 7429
Convex linear combinations 7429
Convex space 5536
Convolution 3345
Conway JH 782
Conway's game of life 5439, 8030
Conway's life rule 8837
Cooling 4381
  vibrational 7226
Cooper pair 9701
Cooperation 6477, 8040, 8382
  altruism 4616
  enforcement norms 4618
  evolution of 5679
  free-riders 4624
  punishment 4616
  reciprocity, reciprocators 4617
  self-interest 4625
  selfishness 4616
Cooperative
  grabbing 3691, 3692
  transport 3683, 3684, 3691, 3692
Cooperative control 7713
Cooperative cost game 1663
  sub-additivity 1663
Cooperative effect 9424
Cooperative game 1509, 1511, 1663
Cooperative game theory 1509, 1510, 4095
Cooperative outcome 4102
Cooperative phenomena 3361
Coordinate
  fractal 3861
  in amorphous computing 265
Coordinate transformation 712
Coordinated path 5786
Coordination 1539, 6477
Coordination dynamics 1438, 1537, 5719
  applications of 1558
  dynamical rules 1544
  nonlinear interaction 1544
  patterns of coordination 1547
  switching 1544
  therapy 1558
  understanding 1544
Coordination games 3196
Coordination law 1550, 1551
Coordination number 1403
  distribution, rock network 4955
Coordination patterns 5719
Coordination problem 6488
Coordination task 7721
Coordination variable 1538, 1539, 1551, 1553
Coordinative level of description 5721
Coordinative structures 1537
Coore's growing point language (GPL) 260
Coornaert M 786
Copeland and Galai's model 5397
Copeland score 9952
Copeland's procedure 9952, 9958, 9960
Coprocessors, secure 5612
Coproduct 1036
Copula 6256, 7020
  copula density 6256
  empirical copula 6256
Copy-cat strategy 5289
Copying 5290
Copying-fidelity 3208
Core 1509, 1510, 1512, 1513, 1520, 1660, 1662, 4095, 7756, 7761, 7769, 8253
  bipartite 8258
  main 8253
  p-core 8254
  (p,q)-core 8258
  stable 1520
  strict 1526
  symmetry groups 5969
Core collapse
  encounter 8693
  three-body 8693
Core GRN
  as a graph 1184

cell phenotype 1184
mammals 1182
Core group 4659, 4662
Core knowledge 1901
Core memory 7835
Core network 334
Core number 8253
Core-peripheral 2736
Core set 7425, 7431, 7436, 7462–7467
Core wars 334
$CORE_{PCT}$ 7769
$CORE_S$ 7772, 7773
Coriolis parameter 9604
Coriolis term, bundle reduction 5499
Corner frequency 2474, 2485, 2488, 2583, 2584, 2591, 5022
Corner period 2482
Corona 624
Coronal mass ejections 9278
Corotation 412
Corporate 2284
Corporate finance 3386
Correction, log-periodic 3865
Corrections to scaling 7000
Corrections to the energy 6728, 6731
ϵ_1 6728, 6732
ϵ_2 6728, 6732
ϵ_3 6728, 6732
ϵ_n 6728
effective order 6731
Corrections to the wave function 6728
$\psi^{[2]}$ 6729
first order 6728
through order 6728
Correlated equilibrium 439, 1587, 4096, 8658
canonical 1588
coalition proof 1594
existence 1589
objective 1590
subjective 1590
Correlated heterogeneity, invasion percolation 4956
Correlated percolation 6544
Correlated random effects 2775
Correlated random slopes 2776
Correlated reaction sets (co-sets) 5544
Correlation 727, 1268, 1616, 3345, 3362, 3363, 3367, 4188, 4490, 5007, 7024, 8586
anti 3368
assortative 4490
disassortative 4490
function 1616
long-range 4188
Correlation and memory in discrete non-Markov stochastic processes 1617
chain of connected kinetic finite-difference equations of a non-Markov type 1621
finite-difference kinetic equation of a non-Markov type 1620
set of dynamic parameters of non-stationarity 1621
short memory function 1621
Correlation coefficient, QSAR modeling 7076
Correlation device 1588
Correlation dimension 9843
Correlation dimension D2 7997
Correlation distance 7919
Correlation energy 8576, 8578, 8582
Correlation exponent 3759
Correlation functions 1278, 2574, 3361, 3363–3365, 3371, 3757, 3759, 3883, 4213, 4381, 4387, 6088, 8715, 9684
connected 3364
Ripley's K-function 2574
Correlation hole 8576
Correlation integral 3509, 6260
Correlation length 611, 1395, 1403, 3708, 3850, 3889, 5028, 5673, 6538, 6541, 6544, 6573, 8709, 8712, 8715
biological fractals 3792
Correlation length exponent 8713
Correlation time 6995, 6998
Correlations between degrees 3930
Correlations of local fields 8592
Correspondence 3873
compact-valued 2221
Correspondence analysis 8359
Correspondence principle 7164
Cortex 3341, 3358, 6100
Cortical coordination 1553
Cortical map 277, 284
Cortical pattern generation 1553

Cortico-striato-thalamo-cortical feedback-loops 8004
COS *see* Center-of-sets
Coseismic seafloor deformation (CSD) 10055
Coseismic slip distribution 1694
Cosine transform (DCT) 10032
Cosmic background temperature 3273
Cosmic conspiracy 29
Cosmic evolution 3268, 3278, 3282
Cosmic expansion 3276
Cosmic microwave background 395, 1634
Cosmic strings 1639, 1642, 1643
  elastic 1651
  global 1645
  lensing 1652
  local 1645
  loop production 1648, 1651
  numerical simulations 1648
  semilocal 1655
  superconducting 1655
  wiggly 1652, 1655
Cosmological gravitational background 1635
Cosmological principle 395
Cosmology 3272
Cost 1663
  final 1502
  marginal 1666
  running 1502
Cost function 1660, 1815
Cost functional 1502
Cost game
  cooperative 1663
  stand-alone 1663
Cost of an algorithm 7119
Cost of voting 9939, 9940
Cost share 1665, 1672
Cost sharing 1675, 4097
  additive 1675
Cost sharing problem 1660
Cost sharing rule 1660, 1665
  additive 1665
  additivity 1675
  anonymity 1666
  average 1674
  budget balance 1671
  constant returns 1675
  cost shares 1665
  cross-monotonic 1672
  decreasing serial 1675
  game-theoretical rules 1668
  independence of merging and splitting 1677
  irrelevance of reallocation 1677
  no advantageous reallocation 1677
  positive 1675
  separable cost remaining benefit solution 1664
  serial 1674
  Shapley–Shubik 1674
  strong monotonicity 1666
Cotangent bundle 5445
  reduction 5488, 5491, 5495, 5498
Coulomb blockade 7309, 7350, 9645
Coulomb coupling 1021
Coulomb-exchange operator 7448
Coulomb failure 2615
Coulomb failure function 2439
Coulomb failure stress 2608
Coulomb gas method 8713
Coulomb hole 8578
Coulomb plasticity 10047–10049
Coulomb wedge 10048, 10049, 10052–10054, 10056
  dynamic 10051, 10053, 10054
Count data models 5554
  count regression 5564
  poisson regressions 5566
Count-up wave 259
Countable state Markov shifts 968
Countably additive 5454
Countably subadditive 5454
Counter-charge, polymer physics 6817
Counter-harmonic mean 229
Counter machine 7692, 7710
  reversible counter machine 7710
  simulating reversible counter machines 7690
Counterflow 3145, 6489
Counterpropagation neural network 2148
Counting 7097
Counting complexity 7184
Couple, definition 6301
Coupled dynamic piezoelectricity 5921
Coupled effects 5915
Coupled electromechanical phenomena 5919
Coupled Fano resonance 7300
Coupled identical oscillators 8912

Coupled map lattice 287, 2720
Coupled systems 5916
  self-consistently 5916
Coupler 8837
Coupling 7606, 8201, 8203
  nearest neighbor 7614
Coupling constant 8576, 8579, 8582
Coupling of light and matter 5923
Coupling pattern 8916
Coupling strengths 7997
Courtroom litigation 7060
Couverage behavior, two-dimensional space with constrains 2060
Covalent bond 6013, 6016
Covariance
  scale 3865, 3869
  strong 3873
Covariance function 3757, 3759, 3772
Coven, Paul, Meester and Steif 2984
Coverage behavior
  forming an ellipse with constraints 2061
  three-dimensional plume 2063
  two-dimensional plume 2062
  two-dimensional space 2058
Coverage factor 4152
Coverage method 3882
Covering algorithm 1799
Covering lemma 2939
Covering number 5206, 5213
Covering space 6322
Cover's theorem 1822
C^p-Norm 5085
CR *see* Constraint release (CR)
Crack 3700, 9877
Crack model 5029, 9921
Crack oscillation 9924
Crack stiffness 9923
Crack wall 9922
Crack wave 9919
Crackling noise 5022
CRAFTY 4142
Crapper's nonlinear capillary wave solution 9980
Crashes 94
  market 3865
Crater 9889
Crawler 5266
Cray XT3/4 5758
Creation and termination of ties 5999
Creation operator 8933
Creative learning exchange 7061
Creativity 8370
Credibility 4798
Credit 2822
Creep 2019
Cremer–McLean mechanism 5521
Cricket localization system 265
Crime participation 8229
Crisis 7010
  economic 3865
Crisp boundary
  fuzzy logic 3988
Criteria 7754, 7760
Criteria isolines
  neuro-fuzzy systems 6097
Criteria isolines method
  neuro-fuzzy systems 6095
Criterion 7753
  Mahler's compactness criterion 3032
Criterion function
  block inconsistency 6917
  inconsistency 6917
Critical behavior 1443, 6772, 6777
Critical bottleneck strength for the mega-jam formation 9394
Critical concentration 8714
Critical curve 4555, 4556, 4559, 8709, 8715
Critical density, percolation 3790
Critical dimension 1130, 3368, 5029, 8713
  ergodic theory 3072
  scaling 3369
  upper 3368
Critical domain wall 8728
Critical dynamics, relaxational 3370
Critical equation
  Aubry set 4554
  Hamilton–Jacobi equations 4549
Critical exponents 477, 478, 648, 1395, 1404, 1443, 1598, 2029, 2831, 3850, 6500, 6538, 6542, 6543, 6571, 8712, 8713, 8715
  mean-field values 1404
Critical fluctuation 1544, 5723, 5734, 7999, 8005, 8199, 8201, 8209, 8928
Critical force 2033
Critical fractional volume 1444

Critical instability 8000, 8003
Critical interface 8716
Critical local speed 9361
Critical networks 1200
Critical opalescence, biological fractals 3792
Critical parameter 4542
Critical part of the spectrum 936
Critical path method 1847, 1850
Critical periods 1902
Critical phenomena 2024, 3847, 4913, 7013, 8711
  critical exponents 4913
  critical slowing down 4913
  power laws 4913
  scale invariance 4913
  self-organized criticality 4913
  universality classes 4913
Critical point 1402, 5011, 5021, 6784, 7855, 8712, 9232
  biological fractals 3791
Critical point universality 6663
Critical properties 644
Critical rate 5720, 5726
Critical slip distance
  earthquake 2552
  earthquake nucleation process 2539
Critical slowing down 1544, 1548, 5723, 7004, 8928
Critical speed 9982
Critical states 1898, 1899
Critical subsolution, Hamilton–Jacobi equations 4550, 4551, 4556, 4557
Critical surface 6586
Critical temperature 6773, 8712
Critical value 5722, 9232
Criticality 476, 7870
  critical 476
  scale-free characteristics 7865
CROCUS 1705
Cross 4103
Cross-correlation 9856
Cross-impact analysis 7835
Cross-impact matrices 7834
Cross-inhibition 3286, 3287, 3290, 3291, 3298
Cross-link density, polymer physics 6818
Cross-linkers 1756, 1761, 1763
Cross-linking 6819
Cross-national observations 2795
Cross-phase modulation 1376
Cross section 6741, 6743, 9254
Cross-validated correlation coefficient, QSAR modeling 7076
Cross-validation (CV) 1698, 1816, 3513
Crossbar array 5865, 5867, 5880, 5881
  storage capacity 5881
Crossbars
  diode 5880
  FET 5881
  resistor 5880
Crossing number 9270
Crossover 121, 3758, 3764, 4135, 4873
  operation 4150
Crosstalk 724
Crothers, Charles 8269
Crowd 5596, 6476
Crowd-anticrowd 5589
  approaches 5593
Crowd disaster 3146, 3154, 6484
Crowd dynamic 1180
Crowd modeling 186
Crowd motion 878
Crowd turbulence 6489
Crushing 6489
Crustal deformation 1690
  block modeling 4268, 4270
  dislocation sources 4258
  inferring source characteristics 4261
  modelling with GPS data 4258
  Mogi deformation source 4258
  sensitivity of GPS data to 4260, 4270
  source geometry 4258–4261
Crustal deformation monitoring 1219
"Cry wolf" issue 2417
Cryosphere 1705
Cryosphere models 1704
Cryptography 5477, 7089, 7090, 7097
  computationally secure 7089
  public key 7095
  public-key 7089, 7090
  quantum 7112
  quantum computationally secure 7113
  quantum information processing 7367
Crystal growth 9771
Crystal symmetry 5751

Crystallization 3727, 5475, 9763, 9766, 9767
latent heat 9775
CS *see* Complexity science
CSCA – clonal selection classification system 2130
CSS code 7334
CTP *see* Theory of perceptions
Cubic 712
Cubic Boltzmann equation, Uehling–Uhlenbeck equation 7818
Cubic nonlinear Schrödinger equation 9996
Cultural diffusion 3214
Cultural evolution 3270, 3281
Culture 3205, 8243
corporate 7841
Culture evolution 1879
Cumulated profits 1972
Cumulative probability 2801
CURE 5055, 5060
Current distribution 3850
Current-flow betweenness 1150
Current-mode logic 9720
Currents, synaptic 6103
Curse of dimensionality 214, 4403, 5350, 7123
Curta 5476
Curvature 3040, 8956
as a two-form on the base 5505
of a connection 5504
Curvature effect 691, 694
Curve fitting 5207, 5209, 5210
Curved plates 698
Curvelet transform
coronization 1727
first generation 1724
second generation 1727
Curvelets 1719, 8675
wedge 1727
Curves 714, 3833, 8713, 8722
fractal 3863
Menger sponge 3836
shortening 6606
space filling curve 3833
von Koch curve 3833, 3835
Cusp 2024
Cuspons 80
Customer base 7846
Cut-and-choose 3328
Cut-off function 937
Cut-out 8249
Cutoff
lower 6999, 7000
upper 6999
Cutoff function 5028
Cuts 8251
line-cut 8252
vertex-cut 8252
Cutting and stacking 2974
Cutting bonds 2840
Cutting plane algorithm 6420
CW *see* Computing with words
Cybernetics 1043, 1111, 1118, 8974–8977
first-order 1109
second-order cybernetics 1110
Cycle monotonicity 5517
Cycle period 70–72, 76
Cycles 2712, 4487, 6241, 8271
bistable 1385
multistable 1385
Cyclic actions, symmetry groups 5963
Cyclic CA 776
Cyclic group C_{6k} 5968
Cyclic group of order $2n$ 5966
Cyclic mapping, parallel computing 154
Cyclic set 9231
Cyclic states 64
Cyclic subspaces 5464
Cyclic tag system 7687, 7709
Cyclical pattern 7848
Cyclicity 9771
Cyclomatic number 2201
fracture network 6516
Cyclone Ingrid 3128
Cyclone Larry 3116, 3130
Cyclone Tracy 3134
Cyclooxygenase-2 (COX2) 2121, 2145, 2170
Cylinder set, ergodic theory 6943
Cylindrical shells 691
Cylindrical source 9906
Cylindrical Wiener process 6229
Cyprus 7842, 7843
Cyprus' Tourism Organization 7843
CYSTAT 7843
Cytochrome P450 2167

Cytoskeletal proteins 1745, 1751, 1752, 1756
prokaryotic 1752
Cytoskeleton 1738, 1739, 1741, 1743–1745, 1747, 1748, 1751, 1753, 1754, 1761, 1762, 1764–1770, 1772–1774
active filament bundles 1771, 1772
active filament solutions 1767, 1771
active filaments 1767
active polar filaments 1766, 1771
active polymer solutions 1765, 1767
active-polymer solutions 1767
eukaryotic 1745, 1768
polar films 1774
prokaryotic 1741, 1745, 1751

D

d-Dimensional full A-shift 8905
d-Dimensional rules 74
d-Dimensional shift of finite type X 8905
d-Dimensional shift space 8905
Da Vinci, Leonardo 5475
Dacite 9798
Dadic and Pisk 8844
Daganzo's cell transmission traffic flow model 9330
DAH 532, 533
Daily stock price, fuzzy syszem models 4091
Daisyworld model 3296, 3297
d'Alembert solution 9986
Damage 2371, 4190
diversity 4190
reduction 4190
DAML (DARPA agent markup language) 8092
DAML+OIL 8092
Damping 6962, 6969
acoustic 6960
control gains 59
distributed 1493
indirect 1498
memory 1500
viscous 6960
Danger theory 4782
Dangling ends 611, 2836, 2840, 3705, 3709, 3853
Dangling page 5277
Dansgaard–Oeschger cycle 11
DAO-E Sierpinski 5641
Darboux basis 2262
Darboux chart 2262
Darboux co-ordinates 2262
Darcy's law 1399
Dark energy 1644, 3273
Dark figure 8229
Dark matter 1644, 3273
DARPA 1477
DART 9594, 9635
Darwin 1883
Darwin, Charles 3662
Darwinian evolution 3291, 3294, 3297
Darwinian principle 8943
Darwinism 3270, 7583
Data
imprecise 4049
incomplete data 7798
Data analysis 6000
Data assimilation 9642
Data attribute descriptions 4399
Data bus, quantum computing 7231
Data classification 2163
Data cleanings 1989
Data collapse 6999
Data collection techniques 9434
Data errors, noise 1989
Data fusion 4404
Data (information) model 4305
Data-logger 2396
Data mining 128, 1827, 4314, 4325, 4328, 4332, 4392, 4404
approach to traffic prediction 9502
dependency 1864
granularity 1864
independence 1865
process 4398
technique 4147
Data mining frameworks
orange 2000
Weka 2000
Data preprocessing process 4398
Data reduction 1775
Data-space 512, 513
biological 512, 513
Data storage 2093

Data table 8259
keys 8260
records 8259
Data transmission system 2395
Database 2404, 4314, 7729
events 2406
instrumental 2404
publicly available 7084
waveforms 2406
Dataset 1828
complexity 7072
Datastructure 6399
continuous space machine 6391
Daubechies 4 10014, 10016, 10019, 10029
wavelet 10013
David-star 3704
Davis, Martin 288
Daylight saving time changes 3469, 3472
DBCLASD 5058
DBSCAN 5057
DCG *see* Dynamic commons game
DD *see* Duplication and diversification
DDOS *see* Distributed denial of service
De Bruijn diagrams 73
de Casteljau's algorithm 5803
De Hoffmann–Teller frame 36
De Vries, Gustav 81, 84, 8114
Deadlocking, parallel computing 155
Deadly earthquakes 2500
Dealer markets
multiple 5394
single 5394
Dealer model 3817
Dealers, market microstructure 5397
Dealing 2000-2 5395
Death spiral 9026, 9028
Deblurring 10031
Debt 2804, 2806, 2822
Debye equation 2694
Debye–Hückel approximation 5578
Debye screening length, polymer physics 6816
Decane (oil)-SWCNT 696
Decay
exponential 1496
of correlation 965, 2930
Decaying harmonic oscillations 9914
Decentralization 1559
Decentralized control 5618
Decidability 801, 816
Decimal 5451
Decision 4412, 7792, 8005
strategic 7850
Decision forest (DF) 1840, 2168
Decision making 4056, 4075, 4562, 8005
utility function 7536
Decision mechanisms 9413
Decision problem 9160
Decision rules 1795, 4412, 7753, 7756, 7761, 7767, 7769, 7771, 7774, 7778, 7792
approximate 7761
Bayesian confirmation measures 7762
certain 7757, 7761
confidence 7757, 7762
minimal 7757, 7762, 7767
possible 7757, 7761
relative support 7762
robust 7762, 7779
strength 7757, 7768
weak 7768
with a minimal number of descriptors 1795
Decision system 1793, 7791
Decision table 1865, 7756
Decision theory 461, 9433
Decision time 8390
Decision tree 1826, 1993, 2165, 4739, 4745
Decision under uncertainty 7753, 7773
Decision window 9862, 9864
Declarative graduation, fuzzy logic 3991
Decoding 1783
Decoherence 1008
quantum computing 7237
vibrational state 7239
Decoherence-free subspace 7340, 7389
quantum computing 7238
Decoherence-free subsystems 7340
Decomplexification, network 505
Decomposition 8247, 8670
of the motion 6599
Deconvolution 1719
Decoupled approaches 5788
Decoupler 8837
Decoupling of self-diffusion (D_s) and viscosity (η) 4219
Decreasing preimages 9256

Decreasing returns to scale 1679
Dedekind eta function 6529
Deduction 214
  fuzzy logic 3986, 3987, 4002
Deductive data mining 4354
Deep brain stimulation 8005
Deep earthquake 2493
Defect 699, 4228, 6195, 6462, 6463
  penta-hepta 6464
  spatial phase 771
  temporal phase 770
Defect chaos 6473
Defect dynamics 394
Defect-mediated turbulence 6201
Defect of information 4825
Defect-tolerant 5871
Defective tile 6777
Defence analysis 9042
Defense, neoclassical economics 178
Deference 8240
Deferred-acceptance algorithm 4097
Deffuant model 6385, 8398
Deficit angle 1642, 1643, 1652
Defining systems 6331
  branch point continuation 6344
  codimension 1 bifurcations of limit cycles 6340
  formulation as a BVP 6347
Deformation 8855
  basic 2639
Deformation nucleus 2641
Deformation of the crack wall 9925
Deformation rigidity 3102
Defuzzification 4073, 6935
  center-of-area 4073
  center-of-gravity 4073
Defuzzified 4882, 4884
Degassing 9767
Degeneracy 1428, 5764, 6731
  accidental 6736
  degree of 6736
  normal 6608
  residual 6731
Degenerate 8325
Degenerate eigenvalue 6730
Degenerate perturbation theory 6679
Degenerated 703, 715
Deglaciation 11
Degree 237, 764, 4487, 6241
  computation 237
  intermediate 764
  node 237
  Turing 764
Degree centrality 8287
Degree correlations 3935, 6054
  average degree of neighbors 6054
Degree–degree correlation 4489
Degree distribution 725, 2875, 3668, 3925, 4487, 4491, 5659, 5694, 6004, 6048, 6051, 8263, 8324, 8327, 10059
  biochemical information 502
  cutoff 4491
  degree exponent 6049
  hubs 2730
  Pareto 10060
  Poissonian 6050
  power law 10060
  power-law 6050
  Zipf 10060
Degree matrix 5354
Degree of compatibility 8681
Degree of dependence
  decrease 1872
  divisor 1869
  subdeterminant 1870
Degree of fracturation 6507
Degree of freedom 1540, 2262, 4515, 5024, 5066, 5719, 6426
Degree of granularity, dependence 1871
Degree (or connectivity) 575
Delauney triangulation 7898
Delay 1850, 1851, 9043, 9457
Delayed discharge model, sectors 4568
Delayed feedback 627, 629
  linear multi-site 635
  linear multisite 634
  linear single-site 635
  nonlinear 636, 638
Delayed hospital discharge model 4569
  examples 4568
  impact 4570
  outcomes 4568
Delayed hospital discharges 4567
Delayed synchronization algorithm 3550
Deleterious mutations 5657

Deliberative agents 188
Delivery
  dental care 4577
  HMO planning 4577
  mental health 4577
Delphi 7834
Delta rule 1817
Demand-controlled
  stimulation 634
  stimulus administration 632
  stimulus length 633
Demand elasticity 1977
Demand for health care managing 4572
Demand game 1660, 1662, 1670, 1678
Demand monotonicity 1679
Demand revelation game 1660, 1662
Demand revelation mechanisms 1671
Demixing 855
  polymer physics 6813
DEMTER 2632
Denali fault earthquake 2602
DENCLUE 5058
Dendrite 1814, 3334
Dendritic crystal growth 3725
Dendrogram 1141, 1146
Dengue fever, effects of climate change 1101
Denial of service, mobile agents 5611
Denoising 1719, 8671, 10031
  hard-thresholding 1721
  minimax rate 1730
Dense groups 8253
Density 2947, 3147, 4497, 4500, 8236, 8268, 8270, 8275
  dimensionless 6516
  isotropic fracture networks 6516
  probability 3874
Density classification 3233
Density classification (or majority classification) task 3235
Density classification task 3238
Density contours 1265
Density dependent control 9020
Density fluctuations 7570, 7573
Density function 1674
Density functional theory (DFT) 692, 1309, 8577
Density matrix 7412, 8109
  consistency 7187
  equation 8933
  purification 7391
  quantum mechanics 7373
Density of cars 9729
Density of states 1318, 1320, *see* Semicircle Law, *see also* Semicircle Law
Density wave 3145
Deoxyribozyme 5652
Deoxyribozyme-based full adder 5647
Deoxyribozyme-based half adder 5646
Deoxyribozyme-based logic gate 5641, 5643, 5644, 5647
Deoxyribozyme logic gate 5649, 5654
Dependence
  degree 1868
  rank 1868
Dependence graph 8322
Dependence measures 6254
  bicovariance function 6265
  bispectrum 6264
  Hellinger distance 6260, 6268
  Kendall's tau 6255
  mutual information 6266
  Spearman's rank correlation 6255
  spectral density 6264
Dependencies among attributes 1791
Dependency
  attribute 1865
  constant 4966
  graph 163, 166
Dependent Schrödinger solver 1023
Dephasing 4195, 4198, 4201, 4202, 4208, 6122
Depinning
  anisotropic medium 1136
  charge density wave 1135
  driven transition 1130
  interface 1135
Depinning transition 5027
Deployment algorithms 7723
Depression 3457
Depression (mood) 3470, 3472, 3473
Depth compensation 50
Depth electrodes 629
Depth estimation 4772
Depth-independent interaction matrix 56
Depth phases 2486
Depth ratio 47

Derivable function 6298
Derivates, futures 5410
Derivation, Navier–Stokes equations 5928
Derivative of Gaussian 3346
Derivatives 5407, 5411
  covariant 3877
  options 5415
Derrida plot 1200
Desai–Zwanzig model 5241
Descartes 418
Descent 8347
Description logic 4412, 8344
Descriptions of knowledge states 4398, 4400, 4401
Descriptive adequacy 1882, 1907
Descriptive data mining 4398
Descriptive expressions 4398
Descriptive language 4398, 4399
Descriptive model 4394, 4398
Descriptor values, QSAR 7072
Descriptors 7792
  chirality-sensitive 7075
  molecular 1303
  QSAR modeling 7073
Design-based and model-based inference 8226
Design concepts and details (ODD) protocol, overview 157
Design principle 719, 720
Designability 735
Desired velocity 6488
Desires 7542
Desulphurization, fuzzy system models 4091
Desynchronization 629, 635, 638
Detailed balance 5670, 6995
Detectable deviation 7623
Detection of bifurcations 6338
  codimension 2 bifurcations 6342
  homoclinic bifurcations 6348
  test functions for codimension 1 bifurcations 6338
  test functions for codimension 1 cycle bifurcations 6339
Determinantal divisor 1865
Determining equations 6178
Deterministic 5024
Deterministic automaton 4739
Deterministic cellular automata 872
Deterministic chaos 2719, 2903, 6195, 6196, 8423, 8426
Deterministic fractal 3880, 6787, 7568
Deterministic modeling 556
Deterministic perturbations 970
Deterrence 4855
Detrending 3762, 3763, 3765
Deutsch 7254
Deutsch–Jozsa algorithm 7380
  quantum computing 7239
Developing countries 2419
Development 1873, 2854
  historical 1876
  human 3865
  individuals 1876
  multi-cellular organism 1192
Developmental dynamic system 1873
Developmental-genetic toolkit 545
Developmental plasticity 1916–1921, 1923–1925, 1928
  behavioral development 1921
  morphological change 1921
Developmental psychology 1874
Developmental robotics 2856, 4648
Developmental state space 1882
Developmental states 1873
Developmental system 1882, 1888, 1890, 1918, 1919
Developmental theories, Fischer's theory of iterative embeddings of components 1877
Developmental timing 1905
Developmental trajectories 1907
Developmental/learning 2845
Deviations principle, ergodic theory 6941
Device scaling 5747
Device simulation 5749
Devices
  carbon nanotube (CNT) 5865
  molecular 5866
  resonant tunneling 5865
  single electron tunneling (SET) 5865
  spin logic 5866
Devil's staircase 2258
Devisor 1865
Dewey 7851
DFT/MD 5894
DGP 9373

Dhar 8035
Diabetes 5547
Diagnosis of mental health, modeling 4573
Diagnosis, modeling 4571
Diagnostic automaton 5641, 5643, 5644
Diagonal mapping 3560
Diagonal mapping scheme 3561
Diagonalization 5758
Diagram
  Bonzani and Mussone 9731
  fundamental 9730
  Greenberg 9731
  Greenshield 9731
  LWR 9730
  velocity 9730
Diagram of different congested patterns 9372
Diagrammatic method 8581
Diagrammatic rule 1936
  node factor 1937
  propagator 1937
Dialectics, multiperspective 7850
Dialog 9048
Diameter 4488, 6241, 10061
Diamoeba 4503
Diamond 690, 691, 700, 9252
Diblock
  co-polymer 6829
  morphology map 6829
Dichotomies 1559
Dichotomous worldview 8193
Dictator 9945
Dictionary 2317
Dictyostelium discoideum 3349
Dielectric breakdown model 3720
Dielectric constant 5577
Dielectric permittivity 1399
Dielectric susceptibility 4214
Dielectrophoresis 5580
Diffeomorphism 703, 709, 6659, 6951
  construction of 6593
  local 6593
Difference engine 5476
Difference equation modeling 8993
Difference equations (DEs) 3696
Difference, definition 6297
Differentiability 3860, 6297
  definitionb 6297
Differentiable dynamics 3054, 6950
Differentiable rigidity 3095
Differential 6286
Differential adhesion 525, 532, 545
Differential analyzer 273, 285, 3357, 5475
Differential equalizer 5477
Differential equation system 5960
Differential equations 3831, 8639
  chemical 329
  existence and uniqueness theorem 8631
  symmetry of 6176
Differential game 4096, 6284, 6478
Differential ground motion 2429
Differential inclusions 6277, 9105
Differential information content 4823
Differential motion 2426
Differentiation 525–529, 531, 532, 1881, 2247
  multi-cellular organism 1192
Differentiation of society 8184
Diffraction 8484
Diffusion 86, 284, 310, 530, 536, 537, 695, 854, 1278, 3285, 3288, 3289, 3293, 3337, 3348, 3357, 3936, 5672, 6785, 7528, 7829, 8272, 8714, 8940, 9792, 9796
  anomalous 310
  chemical gradient 259
  gradients 536
  in amorphous computing 259
  normal 310
  paradoxical 315
  time series 3585
Diffusion coefficient 8722
Diffusion constant 884, 7571
Diffusion equation 312, 2801, 3883, 5219
  normal 312
  on fractal lattices 320
Diffusion exponent 7569
Diffusion function 3656
Diffusion-limited 3368
Diffusion-limited aggregation (DLA) 3720, 3785, 3786, 3884, 7579
  advection-diffusion 3731
  branch competition 3733
  crossover 3730
  fixed scale transformations 3734
  noise reduction 3733
  numerical methods 3725

theory of DLA 3732
Turkevich–Sher relation 3732
Diffusion-limited CCA (DLCA) 3884
Diffusion-limited cluster-cluster aggregation (DLCA) 3884
Diffusion-limited reaction 7570, 7579
Diffusion matrix, stochastic processes 3584
Diffusion of innovation 1956, 5377–5379, 5384, 8307, 8990
Diffusion process 883, 8272
Diffusion solution 7917
Diffusion tensor 8591
Diffusion wavelets 10042
Diffusive coefficient 10037
Diffusive equation 8547
Diffusive shock acceleration 22, 25
Diffusivity 1399
Digital computation 271, 276
Digital computer 274
development 7366
Digital images 10031
Digraph 6239, 6241, 7714, 8281, 8353
minimally connected 6244
simple 6241
strongly connected 6241
weighted 6241
Dihedral action, symmetry groups 5963
Dihedral group $G=D_{2n}$ 5965
4-Dihedral periodic minimizers 5971
6-Dihedral symmetric periodic minimizers 5971
Dihedral (torsion) angles 2689, 2829
rotation 6013, 6016
Dihydrofolate reductase (DHFR) 2130
Dijkstra's algorithm 6409, 6411
Dike 9886
Dilation equation 6862
Dilation invariance 8722
Dilation matrix 589
Dilation symmetry 2253
Dilations 3859, 3861
Dimension 315, 3827, 3833, 3842, 6787
box 3827, 3842
chemical 3701
correlation 3509, 3828, 3842
critical 3870
fractal 314, 3701, 3827, 3863, 3865
Hausdorff 3029, 3036, 3827, 3836
information 3842
Lyapunov 3827, 3842
Lyapunov (Kaplan–Yorke) 3508
maximum determinable 375
metric 3835
not diagnostic of chaos 375
quantum system 7370
spectral 317
topological 3827, 3834, 3835, 3843, 3863
variable 3869
Walk's 315
Dimension group 8899
Dimension-like characteristics, topological entropy 3018
Dimension spectrum 3025
Dimension theory
higher-dimensional dynamical systems 3022
low-dimensional dynamical systems 3022
Dimensional complexity 8004
Dimensionality 3932
Dimensionality of a data set
curse of dimensionality 1989
Dimensionality problem 1802
Dimensionality reduction 1775, 1784, 1990, 5349, 7793
Dimensionless density 6516
percolation properties 6514
Diode 9705
Diophantine approximation 3055
Diophantine condition 1939, 2262, 4515, 6322, 6324
Bryuno condition 1939
Melnikov condition 1944
standard 1939
Diophantine frequency vector 2262, 4515
Diophantine property 6669, 6670
Diophantine vector 5066, 5990, 6681
Diploid-based synthetic lethality analysis with microarrays (dSLAM) 9109
Dipole–dipole interactions 8588
Dipole moment 6121
Dipolynomial representation 68
Dirac bracket, qubit 7370
Dirac system 2045
Dirac tensors, invariant forms 2644
Dirac's delta function 3334, 3343, 3346, 7453
Direct cascade 8727

Direct dominance 6027, 6035
Direct integration 8119
Direct interaction approximation 3650
Direct interaction models 191
Direct product 8298, 8299
Direct scattering problem 4960, 4962, 4965
Directed graphs 5268, 6001, 6030, 6048
  networks 1182
Directed networks 6056, 6788
  acyclic 6056
  bow-tie 6056
Directed percolation 3369, 3752, 4917, 6774, 7576
Directed search 2458
Direction fields 3350
Directional bivariate wavelets 598
Directional entropy 9258
Directional sensitivity 1719
Directional solidification 6466
Directivity 2612
  effects 4453
Dirichlet boundary conditions 4806, 6713
Dirichlet condition 6223, 6224
Dirichlet's principle 1406
Disassortativity 578
Disaster 3110, 3143
  impact severity 3114
  lead time 3114
  management 3112
  mitigation 3113
Discharge rate 9765
Disclination density 2645
Disclinations 2025, 6462, 6469
Disconnect 5426
Discontinous system
  escaping vector field 6699
  regularization 6700
  sliding vector field 6699
Discontinuities 1053, 1898
  discrete 6505
  in rock 6505
Discontinuous feedback 8625
Discontinuous Galerkin method 7892, 7897
Discontinuous system 6698
Discount factor 7618
Discovery system 1984
Discrete baker transformation 75, 76
Discrete choice 5553
  conditional logit (CL) model 5556
  Luce model 5556
  probit 5556
  random utility model 5555
Discrete Euler–Lagrange 2004
Discrete Euler–Poincaré 2005
Discrete event simulation 162, 330
Discrete event system specification (DEVS) formalism 349
Discrete Fourier transform 9914
Discrete Hamilton's principle 2004
Discrete interval valued type 2 fuzzy sets (DIVT2FS) 4090
Discrete Lagrange–d'Alembert principle 2012
Discrete Lagrangian and Hamiltonian mechanics 2004
Discrete Langevin processes 3577
Discrete Laplacian 3739
Discrete mechanical systems 5498
Discrete membership values 4090
Discrete multiwavelet transform 5846
  algorithm 5848
Discrete quantum object set 7435
Discrete scale invariance (DSI) 7568
Discrete stochastic approach 555
Discrete systems 5306
Discrete time 280, 3524
  evolution 3577
  series 1616
  SV models 8784
Discrete variational principle 2003
Discrete wave-number method 2415
Discrete wavelet transform 8671
Discreteness 2318, 2319, 5024
Discretization 1834, 3303, 4186, 6801, 7797, 7799
  based on cluster analysis 7799
  equal frequency per interval 7799
  equal interval width 7799
  minimal class entropy 7799
  of attributes 1798
Disease spread 8308
Diseased lungs, biological fractals 3784
Diseases 1075
Disequilibrium adjustment 7849
Disjoining pressure 3617

Disjointness 5040
ergodic theory 3076
Disjunction 4878
Dislocation 2025, 6462, 6464
Dislocation density
dislocation-relation 2645
dislocation-stress 2645
Dislocation patch 5027
Dislocation theory 7855
Disorder 2019, 5023, 5776, 6407
atomistic 5777
correlated 6409
strong 3850
weak 3850
Disordered solids 8588
Disordered system 2260, 6407
Dispersal limited community 2706
Dispersion 84, 1376, 2041, 2474, 6217, 8483
anomalous 1376
linear 84
management 2341
nonlinear 84, 88
normal 1376
relation 9969, 9975–9977, 9990, 9993, 9997, 9998
Dispersive equation 2041
Dispersive estimate 2041
Displaced maxwellian 7415
Displacement
field 6121, 6122
fluid-fluid 4948
spectrum 2482, 2485
Disruption 1848, 1850–1852
and delay in complex projects 8989
Dissipation 1494
McGehee coordinates 5980
relation 1495, 1496
Dissipative 4589
Dissipative cities 7960
Dissipative models 288
Dissipative non-linear form 5230
Dissipative range 3653
Dissipative scale 3642
Dissipative structures 478
Dissipative surface diffeomorphisms 3019
Dissipative system 288
Dissolution of a flow interruption interval 9398
Distal 9237
Distance 575, 6241, 8236, 8271
algorithmic 255
Cantor 249
epicentral distance 2479
of particles in amorphous computing 265
teleseismic distance 2479
Distance-based information descriptors 4831
Distance between two fuzzy sets 4066
Distance cutoff value 7078
Distance matrix 6837
diagonal elements of 6838
reciprocal 6838
Distinguished eigenvalue 6248
Distributed algorithm 7713, 7715, 9105
Distributed architecture for mobile navigation 8439, 8442
Distributed artificial intelligence 361
Distributed controls 1225
Distributed denial of service (DDOS) 5615
Distributed programming 5605
Distributed quantum information, trapped ion 7244
Distributed robotics 1225, 2073
Distributed sensor grids 5614
Distributed simulation 352
Distributed systems, mobile agent applications 5615
Distribution 1049, 1051, 3404, 5022, 8713, 8716
conditional 3425
fat-tailed 3404
financial market returns 3404
marginal 1866
stationary 6995
tail behavior 3407
Distribution function 236, 9735
hub 236
power law 236
Distribution of distances, biological fractals 3781
Distribution of forces 5016
Distributive justice 1661
Disturbance 711, 9361
Ditto 5473
DIVA model 2320
Divergence 3870, 9696
Divergence measures 6254, 6256
Cramér–von Mises criterion 6257

integrated functions of density differences 6258
Kolmogorov–Smirnov statistic 6258
Kullback–Leibler divergence 6258
quadratic forms 6262
Diversification, multi-cellular organism 1192
Divide-and-conquer genetic-fuzzy problem for items with a single minimum support (DGFSMS) 4155
Divide-and-conquer genetic-fuzzy problem for items with multiple minimum supports (DGFMMS) 4157
Divide-and-conquer strategy 3544, 4403
Dividend 95
yield 3462, 3464, 3465, 3468, 3469
Dividing-and-multiplying pattern 9719
DiVincenzo criteria 1018
DiVincenzo's checklist 7209
Division of labor 3691–3695
Divisive algorithm 1150
Divisive clustering 5054
Divisor, elementary 1870
DIVT2FS *see* Discrete interval valued type 2 fuzzy sets
d_k 3929
DLA 884, *see* Diffusion limited aggregation, 3853
DMWT *see* Transform
DNA 418, 2091
addition 2097
and chromatin 3923
as genetic logic unit 258
computation 932
computing 5479
ecological information 2699
entropy 494, 495
fractal features 3795
information content 4829
nanostructure 5479
origami 2102, 2108
replication 3908
self-assembly 2106
sequence 494, 495, 3795, 3893, 3895
sequence entropy 495
strand-asymmetry 3903
tiles 5867
tiling assembly 5479
walk 3896, 3904
Docking 2122
Dodgson 9944
Dodgson score 9953, 9959
Dodgson winner 9959
Dodgson's procedure 9953, 9958, 9959
homogeneous 9959
Domain 770, 6784
chaotic 771
filter 771
growth 4231
interface 775
multiple 772
of attraction 9227
problem, sequential compactness 2220
regular 770
size 9693
wall 6462, 6468, 8713, 8716
Domany–Kinzel 6774
Dome 9763, 9876
Dome growth 9773
Dominance 1669, 7753, 7760, 7769, 7770, 7773
based rough set approach (DRSA) 7760, 7784
classical 7773
cone 7760, 7762
overwhelmed 1670
solvable games 3189, 3192
stochastic 7773
successive elimination of (strictly) dominated strategies 1669
weak domination 1669
Dominance relation 1662, 6024, 7767, 7768
graded 7767
irreflexive 6024
multi-graded 7768
transitive 6036
Dominant logic 7850
Dominated splitting 4733
Dominates, stochastically 7774
Domination 1518
direct 1528
indirect 1528
strong 1526
weak 1526
Domination path 6036
Domino tiling 5478
Doob's theorem 6232
Dopamine antagonists 2169

Doping, modulation 5776
Doppler cooling
quantum computing 7226
Doreian, Patrick 8269, 8274
Double banana configuration 2838
Double-entry accounting 2806
Double layer 25, 9689
Double resonance 8490
Double series representation 6208
Double well model
splitting 6754
symmetric double well 6757
Double *Z*-characteristic 9391
for phase transitions in vehicular traffic 9327
Double Z-speed-density characteristic 9286
Doubling map 2920
Down-sampling 10035
Downarowicz–Serafin entropy realization theorem 967
Downsian model of political competition 9938
Downward causation 120
Drainage networks 3823
Dreicer field 35
Drift 3531, 3534
time series 3585
Drift function 3656
Drift terms, Gaussian noise 3583
Drift vector, stochastic processes 3584
Dripping handrail 378
Driven diffusive lattice gas 3372
Driven diffusive systems 3372
Driven nonequilibrium system 8727
Driver reaction time 9334
DRSA 7784
Drug bioavailability 2186
Drug design 1303
Drug discovery 1302, 2161
and development 2141
Drug-like compounds 2166
Drug synthesis 2102
Drug target 2161, 5546, 5548
DSGE model 5344, 5345
DTA, traffic stream behavior 9445
Dual game 7647
Dual lattice 1408
Dual problem 1489
Dual rail 7252
Dual recursive formula 7647
Dual sequence 6247
Duality 8724, 8726
controllability/observability 4811
DUEM model 1057, 1064, 1065
Duffie, Pan and singleton stochastic volatility jump–diffusion model 8792
Duffing's equation 6621
Dummy 1685
Dummy player 1663, 1665
Duopoly market game
Bertrand 1530
Cournot 1530
Duplication and diversification (DD), genomes 5699
Duplication and rewiring
model of 5700
Duplication model 7503
Duplication step 7503
Duration dependence 2751
Duration magnitude 2477
Dürer Pentagon 3704
Dusty gas 9806
Dutta's solution 9953
Dyad 6001, 8200, 8202, 8205
Dyadic 6007
Dyadic independence model 8323
Dyke 9766, 9778
Dynagraph 8842, 8843
Dynamic algorithm 1154
Dynamic circular crack 2592
Dynamic commons game (DCG) 2226
Dynamic complexity 1960, 4577, 7999, 8001, 8003
Dynamic core hypothesis 1427
dynamic core 1424
Dynamic coupling patterns 8919
Dynamic decomposition weight 8963
Dynamic diseases 7992
Dynamic environment 1975
Dynamic factor model 450, 2751, 2752, 2755
Dynamic fields 1893
Dynamic fractals 3793
Dynamic friction 5025
Dynamic game 4095, 4099
aggregative growth model 2225
argument 2218
asymmetry 2226

Bellman optimality equation 2220
business-as-usual equilibrium 2228
characterization 2220
climate change 2215
climate change model 2215
emission factor 2230
equilibrium 2215
equilibrium payoff vector 2229
feasible payoff 2223
finite argument 2218
Folk theorem 2222, 2224
game horizon 2215
global common 2226
global warming 2225
individually rational payoff 2224
model 2215
most rapid approach (MRAP) 2229
near-irreversibility 2226
nonlinearity 2226
payoff 2215, 2223
player 2215
second-best problem 2229
state 2216
state space 2225
strategic setting 2226
strategy 2215
time cycle strategy 2224
Dynamic graph calculus 8842
Dynamic graphs 8842
Dynamic instability 1543, 2425, 2428
Dynamic lengthscale ξ 4222
Dynamic market games 4108
Dynamic mechanical experiments 6553
Dynamic mechanism design 5522
Dynamic memory model 1804
Dynamic model (developmental) 1882
Dynamic models 520, 521
Dynamic networks 5786
Dynamic programming 1503, 4349, 5202, 6272, 6278, 6280, 9105
Dynamic programming models 5562
nested fixed point (NFP) algorithm 5563
Dynamic programming principle 1952, 6219, 6224
Dynamic recruitment 262
Dynamic regime 5022
Dynamic renormalization group 8543
Dynamic rule 1880
Dynamic speed limits 3954
Dynamic susceptibility 3371, 4221
Dynamic system 274, 8974, 8975, 8980
definition 1880
Dynamic term structure models
affine models 8793
empirical evidence 8794
empirical evidence on the volatility spanning condition 8795
quadratic yield variation in affine models 8795
volatility spanning condition in affine models 8795
Dynamic thinking 4564
Dynamic traffic assignment 1178, 9430, 9437
Dynamic traffic assignment solution approach
analytical modeling approach 9443
simulation solution 9443
Dynamic traffic management 3945
Dynamic travel time 9448
optimum locations 9450
spatial travel time measurement 9448
Spot speed measurement 9448
Time series models 9448
Dynamic triggering 1219
Dynamic weakening 5026
Dynamical 2846
Dynamical astronomy 399
Dynamical complexity 8523, 8530
Dynamical downscaling 7592
Dynamical exponents 3851
Dynamical heterogeneities 5006
Dynamical minimalism 8422, 8426–8428
Dynamical modeling, Lu–Klimas magnetic field reversal model 8546
Dynamical models 6795, 6798
in biology 1189
Dynamical noise, autocorrelation function 3588
Dynamical phase 4194–4196, 4201
Dynamical properties 1444
Dynamical robustness, biological models 1190
Dynamical social psychology 8421, 8432
Dynamical spectrum 410
Dynamical strengthening 5025
Dynamical symmetries 9644
discrete symmetries 9645
SO(n) symmetries 9648

Dynamical systems 728, 733, 762, 2233, 3053, 3830, 3832, 3838, 6175, 8199, 8420, 8421, 8423, 8424, 8429, 8432, 8630, 8974
cocycles 3067
continuous 4672, 4675
discrete 4675
effective 762
excitable dynamics 733
fractal geometry 3016
hybrid 4671, 4672, 4675
linearizable dynamical system 6185
multifractal analysis 3024
simple pendulum 3837
theory 1437
Dynamical systems, higher-dimensional 3022
Dynamical weakening 5025
Dynamically complete 2778
Dynamics 525, 1042, 1053–1055, 1057, 4380, 5473, 6655, 9777
as building block for AC 329
chaotic 6105
complex 3020, 3389
fast 7611
intricate 7849
molecular 6994
nonlinear 3389
of earthquakes 5022
reduction of 5488
scale 3867
slow 7612
Dynamism, environmental 7839
Dynamo 9278
Dynamorphisms 6798
dyRange 6398, 6399, 6401
Dyson series 6745
Dyson's Brownian process 8728
Dzyaloshinskii–Moriya interaction 9651

E

$E = m c^2$ 3274
E. coli 419, 8744
Min system 8744
E-Science 8087
EAC 278
Early stopping 1816
Early warning 2371
Early warning systems 1217
Earth 3279
orbital parameters 6
Earthquake 1214, 2449, 2476, 2622, 3824, 3865, 4178, 5022, 6490, 7018
a possible model 2552
Aegean Sea EQ, 2001 2626
as a critical phenomenon 2627
Athens EQ, 1999 2627
characteristic frequency 2664
Chi-Chi EQ, 1999 2629
critical slip distance 2552
direct-search location 2457
early-warning 2449, 2467, 2471
emergency response 2451
epicenter 2450
Eratini–Egio EQ, 1995 2626
estimate of source processes 2545
estimate of source time functions 2545
forecasting 4179
Great Chilean EQ, 1960 2627
Guam (Marianas) EQ, 1993 2623, 2627
Hawaii region 2663
hazard assessment 2451
hypocenter 2450
infrasound 4839
Kobe EQ, 1995 2623, 2627–2629
Kozani-Grevena EQ, 1995 2626, 2627
large earthquake 2542
linearized location 2457
local magnitude methods 2665
location 2449
Loma–Prieta EQ, 1988 2623
Loma-Prieta EQ, 1989 2626
magnitudes 2664
main phase 2542
measurements of initial parts 2542
model with small nucleation zones 2541
moment magnitude 2664
nucleation zone 2539
objective function 2450
observations of initial rupture processes 2542
observatory message 2660
ordinary phase 2542
origin time 2450
P-wave magnitude scale 2665
P-wave velocity pulses 2539, 2540

P-waves 2662, 2666
preparedness 4179
probabilistic location 2455
real-time location 2471
rupture growth 2541
rupture velocity 2539
S-waves 2664
schematic source models 2541
seismic swarm in Izu Island region, 2000 2625
size 2451
slow 2492, 8851
small 2543
source location computing 4845
source parameters 2657, 2664
source time functions 2550
Spitak (Armenia) EQ, 1988 2627
Spitak (Armenia) EQ, 1989 2623
strength profile 2553
Strofades EQ, 1997 2626
submarine earthquake 2476
theoretical models 2539
theta 2664
Tokachi-Oki EQ, 2003 2627
tsunami warning 2667, 2669
tsunamis 2658, 2659
velocity pulse 2539
waveforms 2541
Earthquake catalogs 2515, 7859
Earthquake clustering 2349, 2351, 2352, 2354, 2355, 2357, 2358, 2360, 2362, 2365, 2366
Earthquake clusters 1216, 2363
waveform 2549
Earthquake complexity 2484
Earthquake damage 1218
Earthquake damage detection 2373
literature review 2373
Earthquake doublets 7927
Earthquake dynamics 2358, 2439, 2592
Earthquake early warning (EEW) systems 2396, 2516, 2541, 9593
regional 2396
site-specific 2396
Earthquake engineering 1217, 2422, 2424
Earthquake forecasting 1217
Earthquake geodesy 1691
Earthquake-induced disasters 2419
Earthquake infrasound
case study Chile 2005 4844
case study China 2001 4843
case study Peru 2001 4842
Earthquake light 2622, 2623, 2627
Earthquake location 1215, 2512
Earthquake magnitude 1215, 2506, 2514
Earthquake monitoring 1215
in the digital era 2507
instrumentation 2500
regional and local networks 2510
Earthquake networks 1216, 2530, 2532–2537
connectivity distribution 2531, 2533
hierarchical organization 2531, 2535
scale-free 2532–2535
small-world 2532–2534, 2536
small-worldness 2535
Earthquake nucleation 1218
Earthquake nucleation process 2538
critical slip distance 2539
initial crack 2540
models 2541
preslip model 2541
Earthquake physics 1218
Earthquake prediction 2538, 2621, 2622, 4179, 7854
Earthquake prediction and forecasting 1215
Earthquake quantity 5021
Earthquake rupture process 2478, 4436, 4446
angle of slip, or rake angle 4441
asperities 4446, 4447
branching fault 4449
constitutive law (i. e. a friction model) 4446
directivity 4450
directivity effect 4450
dynamic rupture model 4435, 4445, 4446
dynamic stress drop 4436
dynamics of earthquake rupture 4439
earthquake magnitude 4441
earthquake scaling laws 4446
earthquake source dynamics 4449
earthquake-source inversions 4445
earthquake source-scaling 4464
elasto-dynamic equations of motion 4446
fault dimensions 4445
fault-zone maturity 4448
finite-source rupture model 4445, 4446

geometrical fault complexity 4448
hypocenter 4445, 4450
hypocenter location 4447
kinematic rupture model 4435, 4445
left-lateral strike-slip 4441
local slip-functions 4445
moment magnitude 4443
moment magnitude, M_w 4465
multi-scale dynamic simulations 4468
normal faulting 4441
physics of earthquake rupture 4437
point of rupture nucleation 4450
radiation pattern of P- and S-waves 4445
representation theorem 4445
right-lateral strike-slip 4441
rise time 4436, 4445
rupture duration 4436
rupture finiteness 4440
rupture nucleation 4450
rupture velocity 4436, 4445
scale-invariant static stress 4443
seismic moment 4436, 4441
seismic potency 4443
self-similar constant stress-drop scaling 4436
self-similar earthquake scaling 4465
self-similar earthquake source scaling 4443
slip distribution 4436, 4445
slip heterogeneity 4446
slip vectors 4441, 4445
slip-velocity function (slip-rate function) 4436, 4445
source properties 4439
source-scaling relations 4446, 4465
static stress drop 4436, 4443
style-of-faulting 4445
thrust faulting 4441
velocity-pulses 4450
Earthquake scaling laws 1219
Earthquake size 2474–2477
Earthquake source 1219, 2474, 2479
seismic methods 2662
Earthquake spectra 2583
Earthquake statistics 2350
Earthquake swarms 2353
Earthquake thermodynamics 2648
Earthquake visualization 2360
Earthquake waveforms *see also* Earthquake
slip area 2550
Earth's interior 1215
Earthworm 2402
Easley and O'Hara's model 5398
Eastern Europe 2820
EBITDA 7843, 7845
EBS 5409
EBS (electronic broker system 5395
Eccentric anomaly 6676
Echo 115
Eckmann–Ruelle conjecture 3023
ECM 532, 536
ECNs *see* Electronic crossing networks
Ecobot 3690
Ecoinformatics 2708
Ecological asymmetry 2706
Ecological city 7957
Ecological complexity 2697
describing solutions 2700
general types of phenomena 2698
Ecological diversity 2708
Ecological drift 2706
Ecological economics 8998
Ecological energetics 2698
Ecological information content 2699
Ecological information exchange 2699
Ecological kinetics 2699, 2701
Ecological network 1891
Ecological processes 2742
Ecological psychology 1901
Ecological relationships 1890
Ecological symmetry 2706
Ecological systems 1889, 2697, 9104
chemical work 2698
energy flow 2698
entropy 2700
feedback loops 2702
genetic information exchange 2699
organisms 2698
stability 2707
Ecological webs 7517
Ecologically sustainable development 3111
Ecology 8941
biological fractals 3780
spactial correlations 3792
Economathematics 2793

Econometric modeling 9035
Econometrics 3389, 9034
Economic city 7957
Economic determinants of volatility 8808
Economic dualism 2284
Economic efficiency 1661
Economic impact of climate change 1091
Economic modeling approaches 177
Economic sector 2284
Economic temperatures 2819
Economic theory 9432
Economics
  agent-based modeling 177
  and economic modelling 8989
Economy 8941
Econophysics 183, 2793, 2801, 8413
Ecosystem
  ascendancy 2708
  ecological diversity 2708
Ecosystem shifts 2724
Ecosystems
  energy flows 2704
  entropy 2707
  information theory 2707
  interacting species 2728
  material flows 2703
  nonlinear dynamics 2705
  statistical mechanics 2706
Eddy current 5030
Edge artifacts 10032
Edge betweenness 1150
Edge clustering coefficient 1151
Edge orbits 4827
Edges 1725, 6240, 8246, 8267, 8270, 8322
  contours 1735
  curved 1725
  discontinuities 1730
  incident 6240
Edgeworth, F. 1510
Education 1873
Education policy 7061
Educational robotics 4649
Edwards–Anderson model 6418
Edwards–Anderson order parameter 4214
Edwards Hamiltonian
  polymer physics 6815
Edwards–Wilkinson equation 2031
EEG 1555, 8004
EEPAS model, precursory swarm 2568
Effect size 8001
Effective bonds 6840
Effective conductance 1405
Effective conductivity 1397, 1400–1402, 1407, 1409
  effective medium approximation 1401, 1402
  Hashin–Shtrikman bound 1401
  Hashin–Shtrikman bounds 1401
  in one dimension 1406
  (parallel) bound 1401
  parallel bounds 1407
  series 1407
  upper bound of chayes and chayes 1407
Effective dimensionality 8825, 8846
Effective Hamiltonian 6694
Effective information 1433
Effective mass 5770, 7403, 8576
  Hamiltonian 7483
Effective measure complexity
  Grassberger's 495
Effective medium approximation 1395, 1397, 1408, 1409
Effective medium theorie
  Curie–Weiss effective field 3742
Effective medium theory 2682, 3738
  effective homogeneous medium 3742
Effective permeability 6576
Effective prediction domain 7078
Effective properties 1396
Effective resistance 1405
Effective stress 2681, 10049, 10050
Effective temperature 2033, 5008
Effectiveness, of neuro-fuzzy systems 6095
Effectivity 1521
Effector genes 1184
Effects of snow cover on frozen soil 1714
Effectual bottleneck 9374
Efficiency 2650, 3327, 8238, 9934, 9936, 9937
Efficient 3449
Efficient equilibria 4107
Efficient market believer (EMB) 97
Efficient market hypothesis 3438, 3449, 3450, 3454
Egalitarian solution 1662
Egalitarianism 9934, 9936, 9937

Ego-centric 8311
Ego-centric data 8315
Ego-centric network data 8315
Ego-network 8240
Egorov theorem 6758, 6759
Egypt 7844
Ehrenfest urn 8769
Ehrenfest wind-tree model 995
Eigenfrequencies 5727
Eigengap 5278
Eigenstate, degenerate 5770
Eigenvalue 405, 6593, 6597, 6651, 6761, 8949
Eigenvalue elasticity analysis (ELA) 8949
Eigenvalue problem 9916
Eigenvector 405
Eight-point algorithm 4752, 4755
Eight-shaped orbit, *n*-body problem 5966
Eight shaped three-body solution, *n*-body problem 5964
Eikonal equation 6220, 6466–6468
Einstein–deSitter solution 3273
Einstein formula, biological fractals 3785
Einstein relation 990, 991, 3371, 3936
Einstein, Albert 6664, 6805
El Farol bar problem 4977
El Niño 4
El Niño-Southern oscillation (ENSO) 1075
ElarmS 2398
Elastic 7826
Elastic approximation 2022
Elastic backbone 3853
Elastic constants 698
Elastic expansion 9770
Elastic exponent 2831
Elastic half space 5028
Elastic manifold 6414
Elastic percolation networks 6563
Elastic rebound 4268
Elastic rebound theory 1693
Elastic strain 698
Elastic wave 2687
  attenuation 2693
  dispersion 2693
Elastic wave propagation 2523
Elasticity 559, 1128, 8959, 8964
Elastodynamics 6721
Elections 3304, 6807
  presidential 3304
  senatorial 3304
Electric utility models 685
Electrical conductivity 1399
Electrical dissipative structures 9718
Electrical networks, complex 6189
Electricity market 9015
Electro-chemical potential 7311
Electro-kinetic effect 2621, 2632
Electrodipole transport 8592
Electrodynamics 6663
Electroencephalograms 6099
Electroencephelography (EEG) 1436
Electromagnetic signals 1216
Electromechanical digital computing 5477
Electromotive force 32
Electron-electron interactions 7343
Electron gas 7404, 8576
Electron layer 8577
Electron runaway 35
Electron shelving detection, transition 7230
Electron shelving, quantum computing 7230
Electron tunneling 9715
Electronic brokers 5395
Electronic communications networks 5409, 9565, 9567
Electronic crossing networks (ECNs) 5395
Electronic diary 8003
Electronic ecology 3793
Electronic elections 8387
Electronic structure 1309, 5751, 5754
Electroosmosis 5578
Electrophoresis 5579
Electrostatic aggregation 9803
Electrostatic charge, polymer physics 6816
Electrostatic shocks 34
Electrowetting on dielectric 5584
Elek G 781, 789
Elementary granules 7788
Elementary Jacobi rotations 7439, 7457
Elementary modes 5536
Elementary sets 7800
Elevated CO_2-concentration
  changes in food composition 1102
  nitrogen content in plants 1103
Eliasson 6667

Elitist strategy 4874
Ellipsoidal geodesics 5444
Elliptic 2262, 4520
Elliptic curve 7095
Elliptic family 6437
Elliptic orbits 6437, 6447
Elliptic resonant families 6439
Elliptic restricted three-body problem 6440
Ellipticity 6218, 6220
  uniform 6219
Elman network 1822
ElMo 5542, 5544
EM 5743
Embedded interval type-2 fuzzy sets 4014
Embedded state (conceptual state) 4348
Embeddedness 8237
Embeddedness of the observer 1875, 1879
Embedding 374, 375, 390, 8893
Embedding dimensions 7998
Embedding space 3879
Embedding theorem 8901
Embodied systems 1111
Embodiment 1888, 2845, 4901
Embryogenesis 525
Embryonic development 731
  fruit fly Drosophila 727
Embryonic period 1900
Embryonics project 930
Emerge 2847
Emergence 114, 1056, 1269, 1558, 1876, 3271, 4325, 4330, 4331, 4335, 4633, 6477, 7994, 8040, 8366, 8367, 8369, 8407, 8420–8423, 8426–8428, 8432
  adaptive tension 8367
  downward causality 1440
  enabling rules 8367
  heterogeneity 8367, 8370
  temporal emergent behavior 1280
Emergence of a new flow interruption 9398
Emergency 3110
Emergent behavior 1336, 9052
Emergent defect dynamics in cellular automata 2991
Emergent logic 5293
Emergent pattern 4898
Emergent phenomena 1282
Emergentism 1538
Emerging behaviors, agent-based modeling 187
Emissions 9457
  CMEM 9458
  VT-Micro model 9458
Emissions from focal regions 2622
  co-seismic signal 2623, 2625, 2630, 2632
  co-seismic wave signal 2623
  higher frequency electromagnetic emission 2627
  pre-seismic appearance of EM signals 2623
  pre-seismic signal 2632
  ultra low frequency (ULF) anomaly 2626
  VHF signal 2623
  VLF emission 2627
Emjoule 2705
EMM 8802
Emotion 1903, 2242, 2243
Empedokles 6805
Empires 4613, 4615
Empirical congested pattern 9367
Empirical congested traffic 9406
Empirical criteria for traffic phases 9308
Empirical features of traffic breakdown 9283
Empirical ground-motion prediction equations 4440
Empirical mode decomposition 3766
Empirical object set 7461
Empirical relationships, QSAR 7072
Empirical traffic congestion 1179
Empirical trajectory 6479
Empiricism 9064–9066
Enantiomers 3284, 3286, 3288
Enclosed area distribution 6528
Encoding 1782
Encrypted functions 5612
Enculturation 1926
End-to-end distance 6840
Endocytosed 9752
Endogeneity 5563, 7021
Endogeneous complexity 4102
Endogenity, control function 5563
Endogenous complexity 4112
Endogenous matching model 4110
Endogenous threats, mobile agents 5610
Endogenous variable 713
Endomorphic agents 352

Endomorphism
ergodic 2935
Endosome 9752
Endowment function 6011
Endpoint 2046
Energy 1493, 1500, 3270, 3276, 3281, 6408, 6596, 6805
seismic energy 2474
total 1499
Energy approach 700
Energy balance 2587
Energy band 6691
Energy barrier 1310
Energy consumption 9457
Energy decay rate 6122
Energy density 10000
Energy dissipation rate 3644
Energy equality, Navier–Stokes equations 5949
Energy equipartition 5994, 8482
Energy flow 3268, 3271, 3282
ecological systems 2698
ecosystems 2704
Energy-flow diagnostics 3280
Energy flux 10000
Energy gap 9685
Energy landscape 4910
Energy loss 4379
Energy magnitude 2483, 2486
Energy minimization methods 1254
Energy moment ratio 2586, 2589
Energy optimization
biological fractals 3784
Energy rate density 3277, 3279, 3281
Energy release 2650
Energy release rate 2593
Energy shell 2262, 4531
Energy systems 9015
Engagement 7064
Engineering 291, 2857
Engineering demand parameters 2418
Engineering of systems 8377
Enhancement of fluctuations 1548, 5723, 5734
Enigma machine 5477
ENO 10033
ENO-wavelet transforms 10040
Ensemble approaches
bagging 1995
boosting 1996
Ensemble averaging 8589
Ensemble axiom 5627
Ensemble methods 1841, 5317
Ensemble of random 7919
Ensemble, stress based 5015
Enslaving principle 7992
Entangled melt, polymeric fluid 6826
Entangled polymer dynamics 6821
Entanglement 1012, 1014, 1880, 1906, 6411, 6851, 9134
polymer physics 6812
quantum mechanics 7375
Entanglement assisted capacity 7393
Enthalpic elasticity 6559
Enthalpic forces 7030
Entrainment 7606, 7610, 7614
mutual 7610
Entrepreneurship 2284
Entropic effects 2841
Entropic elasticity 6559, 7033
Entropic forces 7030
Entropic profile
homeobox protein 500
rodent homeodomain 500
Entropic springs 2840, 7040
Entropy 728, 1046, 1047, 1049, 1430, 1830, 2444, 2804, 2859, 2860, 2870, 2874, 2878, 2944, 3040, 3055, 3086, 3271, 3276, 3353, 3812, 4713, 4717, 4821, 4897, 7025, 7731, 8715, 9249, *see also* Mean dimension
biomolecular sequence 494
completely positive 2928
conditional 495
DNA sequence 495
E. coli tRNA sequence 496
ecological system 2700
ecosystems 2707
ergodic theory 3071, 6942, 6943
K_2 3509
Kolmogorov 3508
Kolmogorov–Sinai 4986
measure-theoretic 953, 3040
of a measure 3018
of a subshift 790

of cellular automata 3004
topological 953, 3040
Venn diagram 498
Entropy formula 4737
Entropy maximization 1047, 1049, 1050, 4627, 9452
Entropy model 7995
Entropy of symbolic extensions 968
Entropy principle 8936
Entropy production 1006, 2861, 6953
ergodic theory 6941
Entropy spectrum 3024
Entropy theory of cellular automata
cone entropy 3007
directional entropy 3006
entropy geometry and expansive subdynamics 3008
Lyapunov exponents 3005
Entropy vanishing mechanism 4225
Enumeration algorithm 5626
Envelope 2412, 7915
Envelope broadening 7916
Envelope noise 6147
Environment 206, 4790, 7521, 7837, 7842, 7845
complex-dynamic 7837
complex-static 7837
difference between simulated and simulation 189
discontinuous 7837
disturbed-reactive 7837
energy and climate changesee 8989
extensional; transtensional 2608
independent-dynamic 7837
independent-static 7836
placid 7836
placid-clustered 7837
role for simulations 189
role in agent-based modeling 186
surprising 7837
synthetic 185
transpressional; compressional 2609
turbulent field 7837
Environment mediated interaction 193
Environment of evolutionary adaptedness 1925
Environmental applications 7062
Environmental footprint 7843
Environmental remediation 417
Environmental scanning 7838
Environmental stochasticity 8777
Environmental systems 9015
Envy-ensuring 3332
Envy-freeness 1680, 3327
Envy-possible 3332
Enzyme reactions 858
Enzymes 723, 2090, 9104
Epicenter 2474
Epicycles 6663
Epidemic communication *see* Gossip
Epidemic models 6060, 6778
susceptible–infected–recovered 6060
susceptible–infected–susceptible 6060
Epidemic simulations 7577
Epidemic spreading, percolation 3790
Epidemic threshold 4660
Epidemics 6498
Epidemiological system 9035
Epidemiology 4577, 9042
biological fractals 3780
system dynamics 4563
system dynamics studies 4571
Epidermal growth factor receptor (EGFR) 2121
Epigenesis 930
epigenetic inheritance 1925
predetermined 1919
probabilistic 1919
Epigenetic states 527
Epigenetics 1193, 1196, 3971
Epigenetis 527, 542
Epigenetis robotics 2856, 4648
Epilepsy 627, 629, 639
photosensitive 1627
Epileptic seizures 639
Epimerization 3284, 3290, 3291
Epipolar constraint 4752
Episodic representations 2320
Epistasis 9106
Epistatic miniarray profiles (E-MAP) 9109
Epistemological complexity 1906
Epithelial 532
Epithelioid 532
Epoch 1815
Epoxidation
catalyst 1329
polyols 1329

ε-Free Rokhlin lemma 3058
ε-Insensitive SVR 4086
ε-Insensitive value 4092
$\varepsilon_p(n,k)$ 703
Equal differential-time 2407
Equal treatment 1685
Equation 3874, 6596, 8114, 8733, 8734, 8738, 8740, 8741
  Boussinesq 8481
  chemical Fokker–Planck 8738
  chemical Langevin 8738
  chemical master 8734
  continuity 3874
  differential 6103, 6104
  diffusion 8740
  Dirac 3875
  dispersive 2041
  dynamics 3867, 3872
  Einstein 3855
  Euler 3874
  Euler–Lagrange 3866, 6600
  evolution 8481
  Fokker–Planck 6105, 6110
  geodesic 3870
  Hamiltonian 6588, 6589, 6603
  integrable 8481
  Kadomtsev–Petviashvili (KP) 8481
  Kirchhoff 3855
  Klein–Gordon 3875
  Korteweg–de Vries (KdV) 8481, 8511
  motion 3869
  nonlinear 8481
  nonlinear Schrödinger (NLS) 8481
  Pauli 3875
  reaction-diffusion 8740
  reaction-diffusion master 8741
  reaction rate 8733
  Schrödinger 3872
  sine-Gordon (SG) 8481
  spatial chemical Langevin 8740
Equation-based models
  dynamical 4613
  second-order dynamics 4614
  sufficient unit 4622
Equation of motion 377, 8105
  polymeric fluid 6825
Equation of state 4386
Equation with infinitely many variables 6237
Equi-partition set, density 6296
Equicontinuity point 9236
Equicontinuous 9236
Equicontinuous function 3256
Equicontinuous point 9250
Equicontinuous system 2234
Equidistributed 3040, 3041, 3043
Equidistribution of periodic points 959
Equilibirum selection 3194
Equilibrium 202, 1042, 1044, 1045, 1886, 2268, 4791, 4998, 7165, 7622, 7631, 7843, 7849, 8698, 9225
  Aubry set 4551
  Bayesian 4791
  belief free 7627
  communication 7622, 7625
  condition D1 8130
  correlated 5180, 7622, 7624, 8698, 8704
  CURB 5181
  divinity 8130
  ε-equilibrium payoff 8701
  ergodic theory 6945
  extensive-form correlated 8704
  intuitive condition 8130
  joint payoff 3178
  λ-discounted ε-equilibrium 8701
  limsup ε-equilibrium 8701
  mechanical 4998
  minimal prep sets 5181
  Nash 4791, 5179, 8127
  normal-form correlated 8704
  perfect 7618
  perfect Bayesian 8133
  pooling 6985, 8128
  public 7625, 7626
  pure equilibria 3178
  quantile response 5181
  self-confirming 5181
  separating 6985, 8128
  sequential 7622
  sociological factors/bounded rationality 4799
  strategic stability 8130
  subgame (Bayesian) perfect 6979
  subgame perfect 4797
  T-stage ε-equilibrium 8701

thermal 4999
uniform ε-equilibrium 8701
Equilibrium assignment, all-or-nothing 9436
Equilibrium distribution
ergodic theory 6942
Equilibrium in beliefs 8656
Equilibrium joint plan 7642
Equilibrium machines have one state each 4107
Equilibrium refinements 8128
Equilibrium selection 3197
Equilibrium selection argument 4103
Equilibrium state 3055
ergodic theory 6939, 6945
Equilibrium statistical mechanics 4909
Equipartion theorem 8942
Equipartition property 6944
Equitability 3327, 3330
Equity 3449
Equity premium
constant 3457, 3462, 3469
ex ante 3456, 3462, 3464
ex post 3456, 3462, 3464, 3467
puzzle 3456, 3457, 3462
structural breaks 3465, 3469
time-varying 3457, 3469, 3472
trends 3465, 3469
Equivalence class 6288
attribute 1866
Equivalence principle 1513
Equivalence relation 805, 5738, 6288
definition 6287
Equivalent body force 9902
Equivalent chain 6840
Equivariance 5307
infinitesimal 5484
Equivariant bifurcation theory 6177
Equivariant branching lemma 5305
Equivariant momentum maps 5484
Equivariant orbits 5963
Equivocality 7839
environmental 7837
Equlibrium
undominated nash equilibrium 4797
Erdős and Rényi 3856
Erdős–Rényi model 4488
Erdös–Rényi graph 8321, 8323
Erdös–Rényi network 502
motif information 508
Erdös, Paul 8277
Ergodic 385, 2922, 3040, 3054, 4589, 4985
Ramsey theory 3043
theorem 3041
unique 3043
Ergodic decomposition 2924
Ergodic decomposition theorem 3054, 3058
Ergodic flows 3059
Ergodic hypothesis 2919
Ergodic invariant measure 3017
Ergodic measures 5457
Ergodic probability 3017
Ergodic properties 8590
Ergodic Ramsey theory 3091
Ergodic system 402, 989, 992
Ergodic theorem 3054
Ergodic theory 2982, 3017, 3053, 4983
associated flow 3065
critical dimension 3072
disjointness 3076
entropy 3071
fractal geometry 3015
full groups 3065
Hausdorff dimension 3017
infinitely divisible stationary processes 3076
L^∞-spectrum 3069
maharam extension 3065
mixing 3062
mixing recurrence 3061
multiple recurrence 3061, 3063
non-singular transformations 3055
nonsingular joinings 3073
normalizer of the full group 3066
orbit 3064
orbit classification of type III systems 3065
outer conjugacy problem 3066
poisson suspensions 3076
polynomial recurrence 3063
pressure and equilibrium states 6939
quasi-invariance 3069
symmetric stable processes 3076
topological group 3064
topological pressure 3018

Ergodicity 953, 3043, 6232, 7164, 7165
quantum 7167
unique 2923
Ergodicity error 1882
Erikson 1884, 1898, 1905
Error 2450, 4186
diagram 3303, 4186
numerical 6997
quantum information processing 7388
Error-correcting code
quantum information processing 7389
Error-correcting output codes 5323
Error correction 7090, 7256
quantum 7090, 7112
quantum computing 7242
Error correction model 2758, 2759, 2762
Error function 1815, 1818
Error surface 1818
Error syndrome 7327
Errors/trembles in output 4107
Eruption 9876
signature, volcanic activity 4851
Eruptive volume 9832
Escape-rate formula 989, 1000, 1001
Escape route 6490
Escort distribution 2870
Essential matrix 4752, 4755, 4757
Essential tremor 627, 630
Estimated Euclidean homography 54
Estimation 216
Estimation methods
characteristic functions 8799
efficient estimation of general continuous-time processes 8800
efficient method of moments 8802
estimation from option data 8805
estimation via discrete-time model specification or approximation 8796
filtering the latent state variable directly during estimation 8797
indirect inference 8801
Kalman filter 8797
Markov chain Monte Carlo 8804
maximum likelihood with characteristic functions 8800
methods accommodating the lack of a closed-form transition density 8798
particle filter 8805
realized volatility 8808
reprojection method 8789, 8803
simulated maximum likelihood 8801
simulated method of moments 8800
Estimator 6996
consistent 6997
unbiased 6997
Estrogen receptor (ER) 2130, 2149, 2168
ETAS model
ground process 2567
Hawkes' processes 2566
process without ancestors 2567
Ethical theory, human–robot interaction 4650
Ethnography 4616, 4627
Ethylene epoxide 1329, 1331
Etna 9787
Euclidean algorithm 10027, 10028
Euclidean coordinate 46
Euclidean geometry 272, 279
Euclidean lattices, high-dimensional 623
Euclidean-space 44
Euclidean space of vectors of state 1619
Euclid's Elements 3840
Euler equations 3615, 3644, 9971, 9991, 10000
Euler–Hill symmetry 5973
Euler hydrodynamic equation 4382
Euler–Lagrange equation in integrated form 5443
Euler–Lagrange equations 10037, 10038
reduced 5495
Euler–Poincaré equations 5492
Euler–Poincaré reduction 5492
Euler pole 4257
European Modernity 8196
Evacuation 3110, 3134, 3137, 3143, 6483
arching 6484
clogging 6484
decision matrix 3112
egress 6483
jamming 6484
planning process 3131
precautionary 3118, 3130
pushing 6484
route 3127
social issue 3138
stampede 6483
trigger 3114

Evacuation dynamics 3143
Evacuation management 1180
Evacuation process 3143, 3165
  simulation 3165
  simulation example 3166
Evacuation time, calculation 3164, 3165
Evaluation 5351, 8698
  discounted evaluation 8700
  finite-horizon evaluation 8700
  limsup evaluation 8700
Evaluative response 2243
Evanescent wave 7293
Evaporation/condensation 6961, 6963
Even shift 8892
Event-event correlations 2532
Event history analysis 8313, 8314
Eventual set, definition 6290
Evidence-based medicine 9092
Evidence-based treatments 8003
Evidence theory 4049
Evidence, role in neoclassical economics 181
Evo-Devo 1917
Evolution 1553, 1900, 1916–1922, 1924–1926, 1928, 2854, 3270, 3278, 3282, 3375, 3865, 6639, 8040
  artificial metabolic network 505
  chemical 338
  extrinsic 339
  in materio 3221, 3222, 3224, 3225, 3228
  information in networks 502
  informational complexity 505
  intrinsic 339
  number of nodes 505
  of complexity 505
  of cooperation 3213
  of the systems approach 8975, 8987
  rule 5524
  strategies 121
  theory, biological fractals 3780
Evolution equation 1487, 8926, 8929
  approximately controllable 1489
  exactly controllable 1489
  null controllable 1489
Evolutionarily stable strategy 6643
Evolutionarily stable strategy (ESS) 3176, 3179
Evolutionary 2845, 3243
Evolutionary algorithm 6491
Evolutionary approach 2407
Evolutionary biology 1887
Evolutionary calibration 6479
Evolutionary computing 121, 4873
Evolutionary dynamic
  best response dynamic 3183
  Brown–von Neumann–Nash 3183
  logit dynamic 3183, 3189
  Markov process 3193
  mean dynamic 3183
  perturbed best response dynamic 3189
  projection dynamic 3187
  replicator dynamic 3176, 3183, 3187, 3188, 3190, 3192
  Smith dynamic 3191, 3192
  zero-sum game 3190
Evolutionary game 4097
Evolutionary game theory 125, 5677, 6640
Evolutionary granular cognitive neural network 4405
Evolutionary minority game 5590
Evolutionary models
  cooperation 4618
  dual inheritance 4618
Evolutionary optimization 6490
Evolutionary paths, motifs as fingerprints 5698
Evolutionary pressure 720
Evolutionary programming 121
Evolutionary robotics 2855
Evolutionary stable strategies 8667
Evolutionary strategy 2407
Evolutionary system, open-ended 341
Evolvability 735
Evolved spandrels 5699
Evolving clustering 3245
Evolving fuzzy classifiers 3247
Evolving fuzzy controllers 3249
Evolving fuzzy systems 3244
Evolving system 3243
Exact 2929
Exact control 732
Exact controllability 4811
Exact diagonalization
  Slater determinants 7318
Exact diagonalization of the many-electron Schrödinger equation 7318
Exact dimensional 961

Exact solution 5144, 5145, 5147, 5149, 5156, 5157
Exactly controllable 3528
Exactly solvable set 4502
Example 6602
  of a Hamiltonian 6596
Exaptation 4165
Excess volatility 98
Exchange 8586
Exchange and correlation
  finite temperature 8586
  magnetic field 8586
Exchange-correlation 8577
  energy 8582
Exchange economy 1512
Exchange energy 8576, 8582
Exchange energy *J* 7320
Exchange interaction 7353
Exchange of stability 5305
Exchangeability 456
Excitability 1354
Excitable 836
Excitable chemical medium 9705
Excitable lattices 9709
Excitable media 886, 3337, 3349
Excitation, recurrent 6106
Excitatory, synapses 6102
Excludability 7637
Excluded volume 1444, 6505
  calculation 6508
  continuum percolation 6508
  isotropic networks 6514
  polymer physics 6815
Excluded volume effect 3884, 6851
Exclusion principle 879, 884
Executive functions 1904
Exhaustive algorithms 1799
Exit route 3126
Exo-planets 397
Exogeneity 7021
Exogenous 4104
Exogenous factors 2796
Exogenous threats, mobile agents 5610
Exogenous variable 713, 1855
Exotic phase 9701
ExPa 5544
Expanded congested patterns 9374
Expanded traffic congested pattern 9507
Expanding circular pattern 9718
Expansibility 2861
Expansions 5451
Expansive 9238
Expansive subdynamics of cellular automata 3008
Expansive system 9254
ExPas and ElMos 5544
Expectation value 7371
Expectations 4105
  bounded rationality 3374
  rational 3374
Expected losses 2418
Expected value 3317
Experience curve 1967
Experiment
  evacuation 3152
  two-slit 3871
Experimental design 2452, 2456, 2465
  fuzzy system models 4091
Experimental fundamentals
  a brief history 7274
  channels 7275
  detection systems 7276
  sources 7275
  truly quantum random number generators 7278
Experimental implementation of BB84 protocol
  phase coding 7278
  polarization coding 7278
Experimental methodology 2793
Experimental petrology 9795
"Experiments" in social phenomena 2795
Explanation 5196, 9069–9073
Explanation-based learning 5188, 5193, 5196
Explanatory gap 1118
Explicitly simulated collision 330, 331
Exploration 8709, 8715
Exploration process 8716
Explorer 2075
Exponent 1131, 1134, 6243, 6410, 7826
  apparent 6999
  complex 3865
  correlation length 1131
  Diophantine 3031, 3033
  dynamical 1131
  multiplicative 3035
  multiplicative Diophantine 3035

roughness 1131
velocity 1131, 1134
Exponential Boltzmann–Gibbs function 2805
Exponential distribution 2804, 2823, 6003
Exponential family 8230
Exponential function 2801, 2803, 2813
Exponential growth 1054, 4921
Exponential random graph model 8275, 8317, 8319, 8320
Exponential set, definition 6302
Exponential smoothing 4116
Exponential stability 5991, 8618
Exponentially small 6324
Exponents of random power law graphs 7499
EXPTIME 5468
Extended analog computer (EAC) 277, 286, 3357
Extended Kalman filter 6076
Extended memory 2758, 2763
Extended phase space 6177
Extended self-similarity 3658
Extended similarity scaling 8537
Extended systems 4904
Extensible markup language (XML) 4305
elements 4309
schema 4309
tags 4309
unique 4309
Extension functor 4341
Extension group 8555
Extension hardening, entangled polymer dynamics 6822
Extension machines 7042
Extension principle 4071
fuzzy logic 4001
Extension problem 8906
Extension theorem 3260
Extensive form 1510, 1512, 4099
Extensivity 2862
Extent of reaction 556
Exterior covariant derivative 5504
External dynamics 2849
External (or barotropic) modes 6211
External parameters 6318
External perimeter 3709, 3853
External predictive accuracy, QSAR models 7082
External predictive power, QSAR models 7081
Externally validated QSAR model 7077
Extinction thresholds 2723
Extracellular matrix (ECM) 525, 914
tumor cell migration 917
Extrapolation techniques 7836
Extrasolar planetary systems 6446
Extrasolar systems 6428
Extrema 714
Extremal 3003
Extreme event 3771
Extreme pathways 5536
Extreme value distribution 3319, 3408
Extreme value statistics 3773, 7868
Extrinsic evolution 339

F

Fabric weakening 658
Fabry–Perot etalons 6126
FACE *see* Free air carbon dioxide enrichment
Facet, fuzzy logic 3989
Facilitated models 4226
FACS (Free agents on a cellular space) 7967
Factor code 8893
Factor map 5462, 8893
Factor model 3475, 3486
Factor subshift 9257
Factorization 8299, 8301
Factors 544, 2972
Failure stress 5025
Failures 5196
Fair division 3327, 4097
paradoxes 3331
Fall velocity 9801
False alarm 2417, 9869
Falsification 9003
Fama–French model 3392
Fama–MacBeth 3393
Familiarity 2244
Familiarization 2249
Family 711
of algebras 8292, 8294
of circular orbits 6438
of circular periodic orbits 6437
of elliptic periodic orbits 6437, 6441, 6445
of partial algebras 8292, 8296
of periodic orbits 6446
of resonant elliptic periodic orbits 6438

Family income 2815
Family income distribution 2816
Family resemblance 7954
Fano effect 7663, 7671, 9652
Fano interference 7290
Fano quadrupole 7292, 7297
Fano resonance 1305, 7290, 7668, 7674, 7675, 7677
Fano zero and resonance pole 7304
Far from equilibrium 8368, 8371
Far-from-equilibrium 1044
Faraday patterns 3618
Faraday waves 2340
Farsighted consistency 6027, 6038
Farsightedly consistent networks 6043
Fast Fourier transforms (FFT) 10033
Fast gate, quantum computing 7236
Fast multipole method 8694
Fast Radon transform 1723
Fast system 6700
Fast wavelet transform 591, 2382, 10035
Faster is slower effect 6487
Fat tails 6808
Fathead minnow 2145
Fault 6504
  intermittent 5871
  transient 5871
Fault creep 4270
Fault detection and prognostics 3251
Fault detectors 3244
Fault dip 9586
Fault impedance 5026
Fault model 7853
Fault parameters 9594, 9633
Fault slip 2483
Fault tolerance 8067
Fault-tolerant 5871
Fault-tolerant firing squad synchronization problem 3552
Fault-tolerant quantum computing, (FTQC) 7336
FBA 5536, 5542, 5544, 5546
FCBF 5352, 5353
FCCF *see* Fuzzy conjunctive canonical forms
FCM *see* Fuzzy C-means (FCM) clustering algorithm
FCM algorithm, special fuzzy functions 4084
FDCF *see* Fuzzy disjunctive canonical forms
Feasible payoffs 7618
Featural processing 2243
Feature 7787
  extraction 1990, 5349
  interaction 5356
  map 1823
  quality evaluation 5351
  ranking 5351
  selection 1990, 2122, 5349
  selection bias 5352
Feature points 46
Fecundity 3208
Feed-forward 7846
Feed-forward loop (FFL) 5662, 5697
  motifs 5698, 5699
Feedback 8, 280, 1043, 1055, 1493, 1949, 6008, 6273, 6279, 6281, 7592, 9014, 9043, 9104
  discontinuous 6272, 6282
  distributed 1494
  globally distributed 1494
  locally distributed 1494
  memory type 1496
  negative 7843
  nonlinear 1496
  optimal 6283
  positive 7843
  sample-and-hold 6282
  stabilizing 6281
  synthesis 6279
Feedback arc set problem 9957
Feedback control 3944
Feedback control systems 7835
Feedback controllers 2015
Feedback equivalent 8619
Feedback group 8620
Feedback law 3524, 8617
Feedback linearization 6080
Feedback loops 1975, 2284, 4638, 7055, 8184, 8991
  ecological systems 2702
Feedback neural network 1822, 4873
Feedback oscillators 562
Feedback stabilization 9104
Feedback stabilizer 8618
Feedback trading 5414
Feedforward network 1813
Feedforward neural networks 1816

Feedforward systems 563
Feller 6231
Feret diameter, fracture networks 6514
Fermi 6664
Fermi acceleration 25
Fermi distribution 7669
Fermi energy 8579
Fermi hole 8576, 8578
Fermi liquid 9696
Fermi-liquid parameter 8583
Fermi–Pasta–Ulam paradox 400
Fermi–Pasta–Ulam recurrence phenomenon 9983
Fermi surface 9679
Fermi systems 6669
Fermi, Enrico 2864
Fermion 9679
Ferny Creek 3135
Ferroelectric 2021
Ferromagnetic state 8712
Feshbach resonance 9680
Feshbach theory 7664
Fetal development 1900
Fetal stage 1900
Feynman 3860, 7251
Feynman diagrams 3365
Feynman graphs 6664, 6669
Feynman–Hellmann theorem 6729
Feynman–Kac path integration 7127
Feynman path integral, polymer physics 6814
FF-SVM *see* Special fuzzy functions with SVM (FF-SVM) method
FFL *see* Feed-forward loop
FFT 1720, 8696
FHP model 881
Fiber bundle 6190
Fiberglass laminate 7847
Fibonacci numbers, biological fractals 3780
Fick's law 991, 1399
Fictitious play 5178
Fidelity 7164, 7170
Field 779, 3334, 3335, 3877, 5712, 5713
  definition 6288
  gauge 3876, 3877
  in amorphous computing 264
  phenomenological 3335
  structural 3335
Field analogy method 272, 3336
Field computation 273, 277, 284, 286
Field computer 273, 3349, 3356
Field programmable gate arrays 930
Field scale 6568
Field-theoretic methods 3360
Field theory 3361, 3364, 3365, 3367, 3369, 3370, 3372, 8593
Figure-ground contrast 2245
Filament equation 3648
Filling scheme 2937
Filopodia 1742, 1743, 1746, 1750–1754, 1763
Filter bank 6362, 10007
Filters 1776, 6856, 10007, 10034
  band–pass 6860
  conjugate quadrature filter bank 6861
  high-pass 6860
  low–pass 6860
  quadrature mirror filter bank 6861
  subband coding 6860
Filtration 3848
Final prediction error (FPE) 3513
Final set 4497, 4499
Finance 3388
Financial accounting 7847
Financial customers 5408
Financial drawdowns 7019
Financial economics 3436
Financial market data, stochastic processes 3590
Financial markets 2284, 6808
  seismic activity 3716
Financial records 3700
Financial service industry models 685
Fine grained observation 993
Fine structure 6734
Fine structure of turbulence 3650
FinFET 7408
Finger movement experiment 8004
Finitary isomorphism 4990
Finite alphabet, Markov chain 6950
Finite automaton 4098, 4739, 7710, 9250
  reversible finite automaton 7710
Finite configuration 7680
Finite-difference 7895, 8601
Finite-difference method 2430, 5169, 9923
Finite-dimensional method
  Navier–Stokes equations 5932, 5939

Finite dynamical system 160, 165
Finite element 5800, 5807, 10053–10055
Finite-element method 7896
Finite field 167
Finite lattice, ergodic theory 6942
Finite normal mixtures 3423
Finite portion transport 3814
Finite rank, ergodic theory 3061
Finite size correction 7007
Finite-size scaling 1410, 1411, 3744, 6571, 6998
  fracture networks 6510
Finite-source inversions 4446
Finite state automaton 5634
Finite state code 8903
Finite state machine (FSM) 116, 1342, 3682–3684
Finite state transducer 5641
Finite statistical description 958
Finite system, ergodic theory 6941
Finite-temperature conductance 7669, 7676
Finite time estimates 5206, 5214, 5217
Finite time propagators 3581
Finite type, subshift 6947
Finitely additive 5454
Finitely equivalent 8900
Finitely fixed transformation 4991
Fiorenzi, F. 790
Firing set, type-2 fuzzy logic 4016
Firing squad synchronization 5439
Firing squad synchronization algorithm 3539
Firing squad synchronization problem 3538, 7694
  on two-dimensional arrays 3557
Firms
  active 7837
  as agents in neoclassical economic theory 178
  passive 7837
First arrival density becomes 5226
First-difference 2773
First Hund's rule 7312
First international workshop on rotational seismology and engineering applications 2525
First-order 611
  approach 6982
  nonlinear partial differential equations 6216
  phase transition 1639, 1640, 9297
  transition 2840
First passage 3749
  leapovers 5228
  time 3936
  time density 5234
First-past-the-post 9947
First-return map 5460
Fishburn's solution 9952
Fisher equation 3288, 6455
Fisher exponent, polymer physics 6819
Fisher–Lee relation 8600
Fisher–Tippet–Gumbel distribution 3773
Fisher's droplet model 1597
Fisher's theorem 6641
Fitness 3236, 4142, 4491
Fitness function 2855, 4134
  proportionate 4135
Fitness landscape 6640
FitzHugh–Nagumo system 6467, 6468, 6474
Five invariant manifolds 939
Fixed effects 2773
Fixed effects GLS 2776
Fixed effects models 5570
  conditional likelihood 5566
  incidental parameters 5566
  within model 5566
Fixed-path cost sharing rules 1682
Fixed point 172, 5306, 6426, 6435, 6443, 7829, 8927, 8930, 9692
Fixed-point attractor 281, 8424–8426
Fixed point drifts 5728
FJC *see* Freely jointed chain (FJC)
FL *see* Fuzzy logic
Flagella 1741, 1742, 1747, 1748, 1759–1761, 1768, 1771
Flagellum 1741, 1742, 1774
Flat free energy landscape 4233
Flatness 3524, 3535, 8535
  Kurtosis 8535
Fleming–Viot model 6231
Flemming, Walther 923
Flexible polymer, physics 6811
Flexion-extension symmetry 5724
Flexural rigidity 699
Flip bifurcation 5308
Flip polarization 5768
Floating car data 9512
Floating ice 1707

Flocking and formations 5794
Floods 3825
Floor field
dynamic 3160
static 3160
Floor field cellular automaton 3159
Floppy modes 2830, 2832, 2835, 2837
Floquet analysis 2329
Floquet exponent 4532
Floquet form 4516
Floquet multiplier 2262, 2326, 4523, 8638
Floquet normal form 8637
Floquet theory 2324, 6629, 6632, 8932
Florida 3137
Flory 7822
Flory exponent 2023
Flory–Huggins free-energy, polymeric fluid 6827
Flory's mean field theory 6787, 8726
Flow 2262, 3146, 3147, 4516, 6590, 7842, 8639, 8991, 9043
bidirectional pedestrian 3148
bottleneck 3149, 3150
horocycle 8567
in bounded domains 5931
induced 6598
isomorphism theory 4991
nil 8567
non-linear 6824
past a circular cylinder 5938
past an obstacle 5956
pedestrian stream 3146
rank one 8564
smooth, with singular spectra 8565
special 8556
special von Neumann 8569
three-dimensional flow 5939
translation 8569
uni- and multidirectional 3148
Flow-density correlations 9383
Flow equation 6576, 9686
Flow map 3263
Flow resistance 8467
Fluctuating force 8927
Fluctuation analysis (FA) 3761, 3763, 3766
detrended 3763
Fourier-detrended 3766
modified detrended 3766
Fluctuation analysis (MFDFA), multifractal detrended 3769
Fluctuation-dissipation relation 4222, 5233
Fluctuation-dissipation theorem 2031, 3371
Fluctuation exponent 3761
Fluctuation induced nonlinear instability 8549
Fluctuation records 7589
Fluctuation theorems 1000, 1007, 3599
dissipated work 3605
Jarzynski equaltity 3605
second law 3599
steady state FT 3606
transient FT 3606
work fluctuations 3604
Fluctuations 732, 1543, 1898, 3361–3363, 3368, 3371, 3372, 5007, 5722, 6642, 7024, 7522, 8001, 8711
broad distributions 7024
critical 3368
importance 3574, 3575, 3583, 3586–3589, 3591–3593, 4513
polymer physics 6811
1/f Fluctuations 1278
Fluency 2243
Fluid dynamics 8940, 8947
Fluid-filled crack 9919
Fluid flow 6566
Fluid–fluid displacement process 4948
Fluid inertia 5576
Fluid mechanics, Navier–Stokes equations 5927
Fluid motion 3642, 3831
Fluid–solid coupled waves 9921
Fluid–structure interaction 2340
Fluids 6481, 9588
analogies with 6481
balance laws 5928
body force 5928
constitutive equation 5929
incompressible 5929
mechanics modeling 5928
Newtonian 5927
steady-state flow 5930
stretching tensor 5929
Fluorescence 8596
imaging 2100
Flux 1395, 4386, 4713, 4715, 9729

Flux-balance analysis (FBA) 5536, 5540, 5543, 5663, 9116
Flux control coefficients 559
Flux equilibrium 8942
Flux function 4713
Flux lines 6410
Flux tube 9278
Flux vector 1398
Fluxon 8465
fMRI 8004, 8005
Focal depth 2474
Focal length 6398
Fock space 3365, 3366, 8844
Focus-focus type 2262
Fokker–Planck equation 6, 1000, 2801, 2811, 2817, 2877, 3371, 3591, 3656, 3657, 4589, 5734, 8392, 8723, 8772, 8927, 8931, 8937, 8947
  diffusion 8773
  jump moments 8772
  stochastic cascade process 3593
  stochastic processes 3581
Fokker–Planck–Landau equation 7813
Fold catastrophe 714
Fold point 7612
Foliations 4600
Folk theorem for repeated games 4103
Folk theorems 7619
Folklore theorem 5460
Folksonomy 8084
Følner E 784
Følner sequence 784, 2945
  amenable group 4994
Font conventions 2982
Food security 1071
Food webs 1248, 2704, 2714, 3661, 3662
  dynamics 3662
  energy flows 2731
  models 3665
  predator 2731
  prey 2731
  robustness 3674
  structure 3662
Football 6386
Footprint of uncertainty (FOU) 4013
For-simplicity defense 179
Foraging and coverage 5795
Foraging robots 1226
Forbidden links 2716
Force 4380, 5012
  scale 3867
Force approach 700
Force chain 5008, 5011, 5016, 5017, 6490
Force-clamp 7033
Force correlations 5009
Force distributions 5006, 5009
Force moment tensor 5000
  conserved 5000
  stoke's theorem 5000
Force spectroscopy 7032
Forced Korteweg and de Vries equation 9981
Forced mobility 5607
Forced nonlinear Schrödinger equation 9982
Forced vibration 2422
Ford–Fulkerson algorithm 6415
Ford–Kac–Mazur model 8751
Fore-arc 7924
Forecast 2443, 2745, 2746, 5411, 9862
  binary 2443
  continuum 2443
  long-term forecasts 9863
  short-term forecasts 9863
Forecast verification 2438
Forecasting 2747, 2755
  traditional 7850
Foreign exchange market 3805
Foreign wide moving jam 9376
Forest fire 8033
  biological fractals 3789
Forest fire model 3791
Forest thinking 4564
Formal concept analysis 8344, 8356
Formal language 800
Formal model 7057
Formal power series 1934, 6311
Formal series 6663, 6671
Formant 2315
Formation of crowds, anticrowds and anti-persistence in the symmetric phase 5595
Former Soviet Union 2820
Formulation
  Lagrangian 6589
  Poisson 6589
Forrester, Jay W. 9044, 9058, 9061, 9067
  industrial dynamics 9045

Forster's result 8589
Fortuin–Kasteleyn 8726
Fortuin–Kasteleyn representation 8709
Forward markets 5408
Forward pass 1819
Forward problem 9631
Forward search 5350
Foster–Miller TALON 1478
FOTO and ASDA models for traffic recognition 9509
FOU *see* Footprint of uncertainty
Foundations of granular computing 4365
Four-dimensional NMR 4220
Fourier characteristic
  finite-rank 6309
  hyperfinite-rank 6309
Fourier coefficients 2995
Fourier decompositions 10033
Fourier filtering 3774
Fourier integral operator 6757
Fourier slice theorem 9176
Fourier spectrum 6395, 6398, 6399
Fourier transform 1720, 3739, 6391, 6394, 7257, 7411
  quantum computing 7382
Fourier transformation 6390
Fourier/wavelets 10032
Fourier's law 1399, 4386
Fox H-function 5224
FPGA 5881, 8076
FPU paradox 8469
FQQSPR equation 7464
Fractal 67, 1043, 1412, 2575, 3720, 3721, 3822, 3879, 3881, 3893, 3923, 7565, 7567, 7827, 8536, 8714
  biology 3779
  conduction on 1412
  definition 3780
  deterministic 3701, 3702
  distribution 3822
  fractal-like kinetics 479
  geometry 3720
  percolation threshold 482
  random 3701, 3706
  real biological 3788
  scaling 3729
  spatial correlation coefficient 481
  spectral dimension 479
  time-lag correlation coefficient 481
  time series 479, 3793
  2D percolation theory 482
  (Zipf–Mandelbrot) distribution 479
Fractal cities 7954
Fractal dimension 478, 999, 1007, 1649, 2253, 2575, 3722, 3756, 3781, 3822, 3851, 3880, 3926, 6419, 6500, 6541, 6569, 6784, 7568, 8709, 8713, 8723, 8724, 9844
  biology 3780
  box-counting dimension 1000, 1002
  correlation dimension 1000
  entropy scores 2575
  Hausdorff 1005
  Hausdorff dimension 1000
  information dimension 1000–1002
  invasion percolation 4956
  measurements 3788
  polymer physics 6819
Fractal dynamics 2254
Fractal features 3795
Fractal geometry 2255, 3700
  dynamical systems 3016
  ergodic theory 3015
Fractal growth processes 3720
Fractal media 7579
Fractal metabolic rates 3784
Fractal networks 6055
Fractal noise 3794
Fractal reaction kinetics 7566
Fractal scaling behavior 3576
Fractal spatial structure 5023
Fractal structure 2253
Fractal tiling 3815
Fractality 3860, 3929
  renormalization 6056
  reverse renormalization 6056
  self-similarity 6055
Fractional Brownian motion 3442, 3711, 3900
Fractional Brownian signals 3899
Fractional Fokker–Planck equation 3581, 5223, 5232
Fractional GARCH model 3442
Fracto-emission 9801
Fracton 3888
Fracton dimension 3888, 3889

Fracton regime 2256
Fracturation, degree 6507
Fracture 3731, 9885
  discrete families 6515
Fracture mechanics 9589
Fracture networks 6506, 6570
  density 6516
  geological 6505
  percolation 6507
  three-dimensional models 6507
Fracture process, synchronization 2636
Fractured porous media 6505
Fractured system
  invasion percolation 4957
Fractures in rock 6504
Fragile glass-formers 4212
Fragment descriptors
  QSAR modeling 7074, 7076
Fragmentation 3823, 6502, 9793
Fragmentation bomb 9797
Fragmented landscapes, space 2736
Frame
  elements 1725
  tight 1726
Frame mapping 3563
Frame of reference 6427
Frame of relevance 289
Frame problem 1119
Frame-relay protocol 2402
Framework of granular computing 4366
Framingham heart study 9092
Frank vector 2642
Fréchet derivative 3340
Fréchet type 3320
Fredholm alternative 5310
Fredholm operator 5312
Fredkin 8032
Fredkin construction 8834
Free action, mechanical systems 5486
Free air carbon dioxide enrichment (FACE) 1102
Free energy 3636, 3638
  polymer physics 6813
Free-energy density
  polymer physics 6818
Free energy functional 5260, 6195, 6201
Free energy landscape 4224
Free energy rate density 3276, 3282
Free flow 9356, 9357
Free induction decay 8590
Free path encoding 5472
Free will 9412
Free word algebra 8293, 8296
Freely jointed chain (FJC) 6813, 6814
Freeway bottleneck 9282, 9308
Freeway capacities in free flow at the bottleneck 9298
Freeway traffic control 3945
Freezing by heating 6487
Freezing-thawing technique 3550
Freq 6398, 6401
Frequency band of electromagnetic waves 2622
Frequency band pass filter 564
Frequency control 638
Frequency dispersion 9978
Frequency distribution of past outcomes 178
Frequency domain 7, 1501, 8952, 9911
Frequency-halving bifurcation 2262, 4532
Frequency metrology 6169
Frequency-moment distribution 5030
Frequency response method 8955
Frequency size event statistics 5022
Frequency size statistics 5022
Frequency spectrum 6667
Frequency-splitting bifurcations 638
Frequency tagging 1437
Frequency transformation after barrier crossing 1360
Frequency vector 1934, 5066
  Bryuno vector 1939
  Diophantine vector 1939
Frequent set, definition 6290
Frequentist probability 470
Freud 1884
Friction 2615, 5026, 5473, 7858, 9589
  coefficient, effective 10050, 10052
  law 5024
  parameter 3160
Friedel sum rule 7358
Friedman–Moulin rule 1683
Friedmann–Keller equations 3649
Friedmann–Lemaitre equation 3272
Friedmann–Robertson–Walker model 1634
Friend-of-a-friend (FOAF) 8098
Friendship network 6007

Frobenius–König 6245
Frobenius normal form 6246
Frobenius–Victory theorem 6248
Front 6465, 6466
  combustion 6467
  solidification 6466
  weakly curved 6466
Frozen ground temperatures 1714
Frustrated systems 6418
Frustration 6407, 9697
FSM 3691, *see* Fuzzy system modeling
Full A-shift 8892
Full adder 5646
Full completeness 5293
Full configuration-interaction (CI) approach 7318
Full cost coverage 1969
Full group normalizer, ergodic theory 3066
Full groups, ergodic theory 3065
Full information maximum likelihood with optimal filtering (FIMLOF) 8997
Full r-shift 8892
Full rank
  individual 7626
  pairwise 7626
Full support monitoring 7659
Fullerene 690, 691, 9141, 9148
Fullerene molecules, endofullerene 9652
Fuller's phenomenon 5444
Fully-developed turbulence 6195, 6196
Function 3256, 3265, 6600
  definition 6287
  extension 6291
  fractal 3862
  Hamiltonian 6588, 6596, 6597, 6603
  hyperregular 6603
  invariant 6597
  Lagrange 3866
  Lagrangian 6589, 6603
  networks 1182
  of a molecule in AC 328
  of non-stationarity 1618
  ordering/level 8251
  strict evaluation 6301
Function-analytic method, Navier–Stokes equations 5933, 5939
Function and terminal nodes 3238
Function-extension 6294
Function generators 278
Function germs 703
Function set 4141
Functional 5582
  multivalued 6603
  penalized 6607
  Tonelli 6604, 6605
Functional cluster 1427
Functional coefficient 3475, 3488–3490
Functional complexity 491, 500
Functional coordination 1558
Functional data model (FDM) 4305
  entity 4308
  entity-valued functions 4308
  object 4308
Functional dependency 1796, 4417, 7736
Functional derivative 6711
Functional equations 227
Functional genomics 9102
Functional granulation 8338
Functional information, biological system 501
Functional integral 1273, 3365
Functional magnetic resonance imaging 1436
Functional modularity, GRN 1191
Functional network, clustering 507
Functional neuroimaging 8004
Functional order 1539
Functional renormalization groups 2020
Functional synthesis of finite automata 4739
Functional transformation of states 187
Fund flow 3472
Fundamental diagram 3147, 3148, 3150, 3947
  stochastic processes 3588
Fundamental diagram for traffic flow 9305
Fundamental empirical features of phase transitions 9408
Fundamental matrix of solutions 6430, 8637
Fundamental modes 2474, 2484
Fundamental QQSPR equation 7426, 7450, 7451, 7453, 7459, 7460, 7462, 7470
Fundamental solutions 5248
Fundamental theorem of calculus 6299
Fundamental value 97
Fundamentalist 96
Fundamentals 5418
Fungi, biological fractals 3788
Funnel 6492

Furstenberg structure theorem 9243
Furstenberg's class 3076
Furstenberg's conjecture 3105
Furstenberg's multiple recurrence theorem 5451
Further normalization 6176
Fusion 6081
Fusion gates 9720
Fusion rules 3655
Fuzzification 2368, 4010, 4871
Fuzzified 4883
Fuzziness 4072
Fuzzy 843, 3243
Fuzzy algorithms 4070
Fuzzy analysis 4051
Fuzzy association rule 4149, 4154
Fuzzy boundary 4877
Fuzzy C-means (FCM) clustering algorithm 1778, 4081
Fuzzy chip 4885
Fuzzy composition 4011
Fuzzy conditional statements 4069
Fuzzy confidence interval 8684
Fuzzy conjunctive canonical forms (FCCF) 4082
Fuzzy constraint 4075
Fuzzy control 4882, 6936
Fuzzy data-mining algorithm 4146
Fuzzy decision 4076
Fuzzy descriptor 4871
Fuzzy disjunctive canonical forms (FDCF) 4082
Fuzzy environment 4075
Fuzzy estimate 8684
Fuzzy estimator 8684
Fuzzy event 4050, 4053, 4072
  independent 4072
Fuzzy expected value 8687
Fuzzy functions 4055, 4080
  special 4083
Fuzzy-fuzzy set 4012
Fuzzy goal 4075
Fuzzy granulation-degranulation 1782
Fuzzy granules 4361
Fuzzy graph 3994
Fuzzy graph constraint, fuzzy logic 3999
Fuzzy-graph extension principle 4002
Fuzzy if-then rules 4004
Fuzzy implications 6088
Fuzzy inference 4696
Fuzzy inference reasoning
  logical approach 6088
  mamdani approach 6088
Fuzzy inference systems 6088
Fuzzy integrals 231
Fuzzy interval 6936
Fuzzy logic 1343, 3985, 3989, 4073, 4313, 4697, 4871, 4875, 6067, 6087
  equivalence 4074
  implication 4074
  intersection 4074
  negation 4074
  principal contributions 4003
  structure 3988
  type-2 4009
  union 4074
Fuzzy mean 8687
Fuzzy measure 231
Fuzzy membership function 4414
Fuzzy model 4708
Fuzzy neighborhood system 4361
Fuzzy neural network 4404
Fuzzy number 4071, 8681
  addition 4072
  division 4072
  inverse 4072
  multiplication 4072
  opposite 4072
  subtraction 4072
  trapezoid 4072
  triangular 4072
Fuzzy perceptions 8683
Fuzzy predicate 4871
Fuzzy-predicate approximate reasoning 4871
Fuzzy probability 4052
Fuzzy probability distribution functions 4055
Fuzzy probability space 4053
Fuzzy probability theory 4047
Fuzzy processor 4885
Fuzzy proposition 4878
Fuzzy quantifier 4871
Fuzzy random events 4052
Fuzzy random quantities 4051
Fuzzy random set 4000
Fuzzy random variable 8683
Fuzzy randomness 4049
Fuzzy realization 4051

Fuzzy reasoning 6087
Fuzzy relation 4068, 4879, 5744
Fuzzy-rough E-R diagram 7734
Fuzzy rough intersection 7735
Fuzzy rough join 7736
Fuzzy rough projection 7736
Fuzzy rough relational database 7733
Fuzzy rough relational difference 7735
Fuzzy rough selection 7735
Fuzzy rough set 7730
Fuzzy rough union 7735
Fuzzy rule 4879
Fuzzy rule bases 4080
  type 1 4083
Fuzzy sample mean 8687
Fuzzy set system 4360
Fuzzy-set-theoretic facet 3989
Fuzzy set theory 4148, 4355, 7730
  basics 3989
  fuzzy membership 7730
Fuzzy sets 1778, 3986, 4009, 4047, 4061, 4062, 4403, 4875
  contained in 4064
  degree of containment 4064
  degrees of equality 4064
  distance between 4050
  empty 4063
  equal 4063
  equal to a degree 4063
  interval valued 4063
  intuitionistic 4063
  L-fuzzy set 4062
  normal 4064
  subnormal 4064
  subset of 4064
  theory 4062
  type 2 4062
  type-2 fuzzy logic 4011
  type n 4063
Fuzzy singleton 4062
Fuzzy subset 8681
Fuzzy-support of the linguistic terms in large 1-itemsets 4157
Fuzzy system 4696, 6087
Fuzzy system modeling (FSM) 4081
Fuzzy system models
  evolution 4080
  future 4088
Fuzzy system structure 3243
Fuzzy test for hypothesis 8685
Fuzzy validity 4006
Fuzzy vector 4047
Fuzzy vector quantization 1782

G

G-Equivariant minimizers, *n*-body problem 5982
g-Factor 6841
  of any branched polymer 6842
  of randomly branched chains 6844
  of Y-shaped chain 6843
G-Invariance 6597
G-Protein-coupled receptor 2148
G-Space, Hamiltonian 5484
g-Twisted action 6602
Gabor 3351
Gabor uncertainty principle 3352
Gabor wavelets 3351
Gain ratio 1831
Gait control table 5627
Gaits 1554
Galactic clouds 3279
Galactic dynamics 6189
Galactic orbit 411, 412
Galam model 8387
Galam, Serge 6386, 8381, 8387
Galaxies 3274
Galerkin method 5932
Galileos' principle of relativity 6426
Galois correspondence 6796, 6803
Galois lattice 8275, 8295, 8303
Galton–Watson process 647
Game 1520
  absorbing games 8702
  abstract 1520
  balanced 1664
  common interests 3187
  concave 1664
  convex 1521
  demand revelation 1671
  dynamic 2215
  evolutionary games 4138

game of incomplete information 7653
games in normal form 8699
in extensive form 4862
in normal form 4857
matrix games 8699
multi-player games 8703
non-transferable utility games 1666
NTU- 1520
of life 66, 116, 775, 782, 869, 892, 894, 895, 927, 4502, 8869, 8880
of life rules 893
of perfect information 5647
on the unit square 4858
prisoner's dilemma 3178, 3179, 3198, 3199
prisoner's dilemmas 3200
repeated games 8699
rock–Paper–Scissors games 3200
simple 1522
standard fixed tree 1667
stochastic game 8698
stopping games 8699
strategic 1668
sub-modular 1664
TU- 1520
with complete information 433
with incomplete information 427, 428, 440
zero-sum 3188, 8701
Game horizon 2222, 2224
Game semantics 5290
Game theory 1509, 1510, 1660, 2215, 4095, 4098, 4610, 5677, 6492, 6978, 8127, 8381
dictator game 4609, 4616
dynamical 4624
evolutionary 4616
neoclassical economics 181
tit-for-tat 4617
Gamma-band activity 8003
Gamma-band synchrony 1439
Gamma distribution 2801, 2807, 2810
Gamma oscillations 6112
Gamma ray 22, 35
Gamma systems 928
Gap 407
Gap exponent 7002, 7003
Gap propagation 6489
Gap scaling 7002
Gapology, quantum information processing 7387
GapP function 7184
Garbage-less circuit 7699, 7703
Garbage-less computation 7697, 7706
GARCH models 3427
Garden of Eden (GOE) 2983, 2984, 3012, 5439, 7680
configuration 779, 783
pattern 783
theorem 780
Garman's model 5396
Gas 9880
Gas dynamics 4715
Gas like regions, critical point 3792
Gas volume fraction 9789
Gaseous model 8693
Gastrulae 532
Gastrulation 529, 532, 533, 536–538
Gate 7697
COPY 1380
FANOUT 1380
Fredkin gate 7689, 7698
NAND 1382
NOT 1381
ONE 1381
quantum computing 7231
reversible logic gate 7698
reversible logic gates 7689
switch gate 7689
Toffoli gate 7698
Gâteaux derivatives 3340
Gated transport spectroscopy 7350
Gauge theory of mechanics 5495
Gauge transformation 9275
Gauss 2867
Gauss linking number 9270
Gauss map 5461
Gaussian ACF 7921, 7925
Gaussian chain 6841
single chain dynamics 6819
with any type of branching 6848
Gaussian curve 6529
Gaussian distribution 5024, 8722
Gaussian ensembles 7507
Gaussian filter 10037
Gaussian kernel 10037, 10038
Gaussian minimum output entropy conjecture 7393

Gaussian noise 3583
Gaussian probability density function 5219
Gaussian random variable 6233
Gaussian random walk, polymer physics 6812
Gaussian systems 2963
Gaussian thermostats 989, 1000, 1002, 1007
Gaussian thermostatted systems 1001
GCL *see* Generalized constraint language
GDP 5326–5330, 5332, 5334, 5336, 5337, 5341, 5343, 5344
  GNP 5333, 5334, 5339, 5342
GDToolkit 1256
GEDI map, gene expression network 1183
Gel 1609, 1739
  active 1739
Gel electrophoresis 2093
Gel fraction 6549, 6550
Gel point 6545, 6547, 6554
Gelation 3848, 7822
  polymer physics 6818
Gelfand triplet 989, 1004
Gemetic tree 3238
Gender-role concepts 1905
Gene 418, 525–528, 538, 539, 541–543, 4161
  active 527
  alternative splice 4170
  alternative splicing 4161, 4169, 4171
  biological behavior 1183
  constitutively active 527
  developmentally regulated 527
  engrailed 540
  exons and introns 4169, 4170
  expression 525, 527, 541
  functionally isolated "islands" 1191
  gap 540, 542
  her 535
  housekeeping 526
  maternal 540
  miniclusters 1183
  pair-rule 539, 540, 542
  past-transcriptional regulation 4172
  post-transcriptional regulation 4171
  regulated 527
  regulating activity 1186
  RNA genes 4172
  segment polarity 540
  splicing 4169
  stencil genes 4172
  transcription 4169, 4173
Gene × enviroment interactions 1916
Gene × environment × development interactions 1922–1924
Gene × environment interaction 1918, 1920
Gene × environment interactions 1917, 1919
Gene × environment interactions 1923
Gene circuit
  architecture 1189
  dynamics 1189, 1190
  small 1188
Gene conversion 4161
Gene deserts and forests 4173
Gene detection 3906
Gene duplication
  and deletions 5659
  balancing selection 4167
  correlation complexity organism 4168
  DDC model 4168
  neofunctionalization 4167, 4168
  paleopolyploidy 4166
  polyploidy 4163, 4167
  processed genes 4166
  retroposition 4166
  segmental duplication 4166
  segmental duplications 4167
  subfunctionalization 4167, 4168
Gene-environment relationships 1883
Gene expression 721
  high-dimensional attractor 1204
  network state 1183
Gene expression microarray 9115
Gene expression pattern 1182
Gene expression profile 1182, 1188, 1204
Gene knockout 9097
Gene ontology 9097
Gene order 4173
Gene organization 3913, 3915
Gene regulation, cell fate determination 1180
Gene regulatory networks *see* GRN
Gene-specific information 1187
Gene therapy 423
Genealogy 8258
General circulation models 2
General electric 7835

General equilibrium 203
necessary condition for 180
General metabolic equation 2709
General Motors traffic flow model 9334
General operations among a family of sets, definition 6287
General pattern 9356, 9364
General practitioner services (GPs) 4566
General-purpose analog computer 277, 285, 3357
General-purpose computer 285, 3354
General purpose frameworks 194
General regression neural network (GRNN) 2154
General scheme for generalized blockmodeling 6916
General systems theory 207
General traffic congested pattern 9323, 9507
Generalization 1816
fuzzy logic 4003
Generalization operator 4397
Generalization relation 4397
Generalize 2851
Generalized autoregressive conditional heteroskedasticity 3440
Generalized blockmodeling 6915
criterion function 6917
criterion function requirements 6917
deductive use of blockmodeling 6918
direct approach 6916
generalized equivalence 6915
inductive use of blockmodeling 6918
measures of fit 6917
optimization 6917
pre-specification 6918
Generalized central limit theorem 3812
Generalized constraint language (GCL) 4000
fuzzy logic 3997
Generalized decision class 7792
Generalized decision function 7792
Generalized delta rule 1819
Generalized dimension 8538, 8539
Generalized eigenvectors 6247
Generalized equivalence, departures from permitted blocks 6917
Generalized estimating equations 2776
Generalized extreme value distribution 3320
Generalized firing squad synchronization algorithms 3565
Generalized firing squad synchronization problem 3547
Generalized gradients 6271, 6274
Generalized hip–Hops, *n*-body problem 5967
Generalized hyperbolic distribution 3417
Generalized impulse response function 5342
Generalized Langevin equation 3577, 8929
Generalized notions of solutions 6216
Generalized orbits 5976
Generalized OWA 230
Generalized Pauli matrix 7372
Generalized polynomial 3087
Generalized regularized long wave equation 6455
Generalized Riemann hypothesis 7096, 7104
Generalized rigidity 1275
Generalized solitons and compacton-like solutions 8460
Generalized solution, *n*-body 5978
Generalized Sundman–Sperling estimates, *n*-body problem 5980
Generalized Wiener index for a regular dendrimer graph 6846
Generalized Zakharov–Kuznetsov equation 6457
Generalizing equivalence, permitted block type 6916
Generated 703
Generating function 2262, 3364, 5066, 6499
Generating symbol 7429
Generative grammar 1878
Generative models of networks 10062
Barabasi–Albert model 10063
copying model 10063
cosine similarity 10064
forest fire model 10064
preferential attachment 10068
Generative social science 126
Generator matrix 7327
Generators 3877
GENERIC 5260
Generic 2268, 2923, 4515
Generic chaos 955
Generic dynamics 971
Generic graph 6013
Generic networks 2831
Generic point *see* Regular point
Generic rigidity 2835, 2838
Generic scale invariance 3372

Generic space 2238
Generic structures 8996
Genericity 6324
Genetic algorithm (GA) 121, 1431, 2116, 3233, 3235, 3236, 3384, 4150, 4873, 5549, 5754, 8832, 9042
Genetic assimilation 1925
Genetic circuits 284
  bistable 1193
Genetic determinism 1195
Genetic engineering 418
Genetic-fuzzy data-mining technique 4146
Genetic-fuzzy mining problem 4151
  way of processing items 4151
Genetic information exchange
  ecological system 2699
Genetic interaction 9107
Genetic network 9106
Genetic neural networks 4404
Genetic neural support vector machines 4407
Genetic operators 2855
Genetic perturbation 9106
Genetic programming 121, 3233, 3237, 3385, 4745
  strongly typed 4140
Genetic regulation 419, 2090
Genetic regulatory network 284
  combinatorial regulation 722
  regulatory loops 722
  thermodynamic models 722
Genetic tree 3233
Genetics 1227, 3243, 3971, 9106
  complex genetics 9107
  differential expression 1233
  functional genomics 1233
  gene regulatory networks 1227
  genome wide association studies 1241
  interactions 3973
  mass spectrometry 1239
  microarrays 1233
  nuclear magnetic resonance (NMR) spectroscopy 1239
  quantitative 3974
  regulation 1227
  single nucleotide polymorphisms 1228
  single nucleotide polymorphisms (SNPs) 1241
  systems genetics 9107
  transcribed 1227
  transcriptome 1233
  translation 1227
Genome annotation 5538, 5539, 5547
Genome dynamics 4165
Genome engineering 4173
Genome sequencing 4173
Genome size
  gene duplication 4166
  haploid 493
  microbial genome 4164
  microbial genomes 4163
  paleopolyploidy 4166
  polyploidy 4167
  selfish DNA hypothesis 4164
Genome structure
  chromosomal territories 4163
  diploidy 4163
  eukaryotic 4162
  isochore 4164
  organization 4174
  paleopolyploidy 4166
  polyploidy 4163
  prokaryotic 4162
Genomes 4161, 5536, 5538
  duplication and diversification 5699
Genomics 2708
  comparative 3966
  computational 3966
  functional 3965
Genotoxicity 2186
Genotype 118, 2855
Genotyping 3974
Geo-complexity 1216
Geodesic distances 6005
Geodesic flow 3098, 8645
Geodesic motion 993
Geodesics 3040, 3045, 3046
  fractal 3871
Geodesy 8856
  submarine 8857
Geodetic observations of ground deformation 1691
Geologic analogy method 8998
Geological formations
  transport processes 6505
Geological fractures 6505
Geometric 4195, 4197

Geometric complexity 3016
Geometric condition, multiplier 1494
Geometric coordinate-free approach to stochastic modeling 8756
Geometric frustration 4229
Geometric gate, quantum computing 7234
Geometric integration 2003
Geometric irregularities 5919
Geometric mean 228
Geometric mechanics 5483
Geometric OWA 230
Geometric phase 4194–4199, 4201, 4202, 4204, 4208, 4209
Geometric singular perturbation theory 6697
Geometric structures, reduction of 5497
Geometrical and material heterogeneity 5024
Geometrical condition
  piecewise multiplier 1495
Geometrical phase transition 8714
Geometrical simulation 6013, 6020–6022
Geometrodynamics 8843
Geometry
  curved 3867
  dynamical systems 3016
  noncommutative 3869
  space-time 3869
Geometry information in amorphous computing 265, 267
Geometry of group action 6187
Geophysical fluid dynamics 8118
Geostationary meteorological satellites 9613
Geothermal field 2605
Gerken 3539, 3541, 3542
German Indonesian tsunami early warning system 9596
Germinal stage 1900
Gestalt perception 8003
Gestalt psychology 4895, 7991, 8003, 8267
Gestalt theory 8939
Gestaltkreis 7991
Gevrey 6324, 6325
Gevrey-2 5989
Gevrey index 6656
Gevrey spaces 6649
Geyser 6958, 6969
Ghettos 6783
GHZ *see* Greenberger–Horne–Zeilinger (GHZ) state
GHZ experiment 7377
Giant component 6497
Giant SCC (GSCC) 5270
Gibbard–Satterthwaite theorem 9934, 9938
Gibbs distribution 4383
  discrete 6946
  ergodic theory 6946
Gibbs entropy 1002, 1006
Gibbs–Markov 965
Gibbs measure 3055
Gibbs' oscillations 10032, 10040
Gibbs' phenomena 10032, 10033
Gibbs property, ergodic theory 6942
Gibbs sampler 2749, 2751, 2753
Gibbs-sampling 450
Gibbs state
  ergodic theory 6940, 6948, 6949
  property 6949
  summable potential 6947
Gibbs–Thomson effect 6467
Gibbs, Josiah Willard 2860, 2863, 2864
Gibbsian systems 3087
Gibrat's law 180
Gillespie algorithm 330
Gini coefficient 2801, 2815, 2816
Gini index 1831
Ginzburg–Landau functional 8712
Ginzburg–Landau (RGL) equation 3624, 3626, 6168, 6463, 8931, 8934
Gipps-equation for safe speed in vehicular traffic 9336
Girth 6241
GL rule 895, 4241, 4245
Glaciers 1711
  glacier models 1711
Glass 1262, 2036, 9885
Glass transition 9790
Glassy dynamics 7584
Glassy effects 2023
Glauber–Sudarshan representation 8933
Glider 775, 4240, 4241, 5439, 8031
Glider gun 4243
Gliese 876 6447
Glimm–Effros theorem 3059
Glioma cell migration, cellular mechanisms 917

Glioma growth, brain's fiber tract 918
Glioma invasion, effects of fiber tracts 918
Global asymptotic stability, controllable 8618
Global attractor, Navier–Stokes equations 5953
Global catalog 9832
Global change 2712
Global climate change 2227
  issue 2225
  result 2227
Global climate models 7592
Global connectivity 6511
Global dynamics 174
Global function 7680
Global GrC model, second GrC model 4345
Global mean temperature, change rates 1097
Global minimum 1818
Global Pareto optimum (GPO) 2227
Global positioning system (GPS) 1219, 1691, 3688, 4251
  and earthquake response 4276
  and fault slip rates 4268
  and gravity data 4266
  and InSAR data 4264
  and seismic data 4264, 4275
  and transient deformation 4271
  and volcanic deformation 4262, 4266, 4275
  applications of GPS data 4255
  error sources 4252
  high sampling rate 4275
  inferring deformation source from 4261, 4263, 4266
  modeling of data 4258
  postseismic deformation measurement using 4273
  real-time data 4276
  relative precision 4255
  strength and weaknesses for deformation measurement 4267
  strengths and weaknesses for deformation measurement 4264, 4275
  strengths of 4255
  use with other datatypes 4264
Global rigidity 3096
Global-search 2449
Global seismographic network 2508
Global simplicity 968
Global warming 3115
  photochemical smog 1098
Global warning system 2367
Global welfare 2227
Global workspace theory 1438
  neuronal global workspace 1438
Globular cluster 8696
Globular form, polymer physics 6817
Globus pallidum interior 629
Gluckman, Max 8268
Glycine/NMDA receptor 2146
Glycogen phosphorylase B (GPB) 2129
Glycogen synthase kinase-3 2145
Glycoproteins 9751
GO-GARCH 4127
Go-like systems 7035
Goal 7539
  achievement 7539
  active vs. inactive 7543
  chosen 7543
  end-goal 7539
  executed vs. waiting 7543
  individual vs. social 7543
  innate 7539
  learned 7539
  maintenance 7539
  organizational 7839
  prosocial vs. aggressive 7543
  realizable 7539
  shared vs. collective 7543
  side-goals 7539
  sub-goal 7539
  unrealizable 7539
Goh condition 5449
Gold nanoparticles 1321
  model 1321
  reactivity 1321
  STM image 1321
Golden mean shift 8892
Golden number 6632
Goldreich–Julian" charge density 35
Goldstone mode 6466
Golubitsky 704
Goodness of fit 8329
Goodwin model 563
Google bombing 5279
Google PageRank 6239

Gordon S. Brown 7835
Gossip in amorphous computing 263, 264
Goto 3539, 3542
Goto–Nambu action 1642, 1647
Gottesmann 7256
Gottschalk W 780, 781, 785
Goudie 3119, 3130
Gouy–Chapman model 5578
Governance 9427
Government securities dealers 5395
Governor's Office of Regulatory Assistance (GORA) 7055
GPAC 274
GPCR 2148
GPL *see* Growing Point Language (GPL)
GPO *see* Global Pareto optimum
GPS 3690, 8851, 8856, 9424, 9635
GPs *see* General practitioner services, 9407
GPS data collection 4253
  campaign GPS measurements 4253
  campaign measurements 4253
  continuously recording GPS 4253
  semi-permanent instrument deployment 4277
Grade of membership 4062
Gradient 3340, 3352, 3354, 6711
  as primitives in amorphous computing 264
  in amorphous computing 267
Gradient ascent 3340
Gradient descent 281, 1817, 1820
Gradient expansion 4384
Gradient networks 6056
Gradient percolation 6530
Gradual development 3243
Graduation, fuzzy logic 3986
Grain boundary 6195, 6200
Grain of truth 5182
Gramian matrix 7461
Grammar 2318
Grand-canonical minority game 5590, 5598, 5602, 5603
Grand-canonical version 5589, 5590
Grand coalition 1511
Grand junction system 9954
Granger causality 1434
Granovetter 8272
Granovetter, Mark 8268
Granular agent 4426
Granular computing (GrC) 1778, 4305, 4313, 4324–4326, 4328, 4340, 4355–4357, 4403, 8334
  applications of 4366
  data table 4306
  granules 4305
Granular data 4402
Granular data model (GDM) 4314, 4346
Granular exact sequence 4350
Granular flows 5003, 9414
  inertial 5003
  quasistatic 5003
Granular fluid 4380
Granular kernel machines 4411
Granular link 4402
Granular mathematics (GrM) 4313
Granular matter 4378
Granular media 4215, 6490
Granular model 2688, 4393
Granular neural network 4403
Granular neuron 4402
Granular reflection 4427
Granular structure (GrS) 4324, 4329, 4337, 4341, 4348
Granular support vector machines 4411
Granular synthesizer 4426
Granular systems 4411
Granular weights 4403
Granularity 4396, 4397, 5736, 5749
  contingency table 1865
  degree 1871
Granulation 4357
  functions 3993
  fuzzy logic 3986, 3992, 3994
Granulation of knowledge 1798
Granule 1798, 4397, 4423
Granules of knowledge 7760
Graph 1141, 2201, 4487, 4488, 4494, 6030, 6240, 6411, 6415, 6417, 6419, 6421, 8232, 8246, 8281, 8322, 8351, 8604
  acyclic 8250
  bipartite graph 6241
  Cayley 778
  chemical 2201
  classical random 1245
  deterministic 4494
  directed/undirected 1244
  dual 6583

fracture networks 6510
labeled 785
matching 6583
maximum degree 8252
motifs in 5692, 5693, 5695, 5697–5700, 6369
non-planar 6419
planar 6419
recursive 4487, 4488
self-dual 6583
simple directed 8246
simple undirected 8246
small-worlds 1244
underlying 6241
Graph automorphisms 4822
Graph centrality 4832
Graph distance 4831
Graph drawing 1255
Graph grammar 8840
Graph layout algorithm 8283
Graph machine neural networks 2149
Graph partitioning 1143, 1147, 8604
Graph rewriting 336, 337
Graph rewriting automata 8840
Graph self-joinings (GSJ) 3074
Graph structure 10058
Graph theory 6050, 6836, 8266, 8270, 8604
structure/traversal isomorphism 4608
Graph visualization 1244
Graphael 1256
Graphcode 157
Graphene 8577
Graphene sheet 691, 692, 694, 700
Graphite 690–692, 700
Graphitic patch 9147
GraphViz 1256
Grashof number 5955
Grassberger model 6778
Grassberger's effective measure complexity 495
Grassmannian 7512
Gravitational model 1047
Gravitational wave 1653
Gravity
loop quantum 3869
quantum 3869
GRDDL (gleaning resource descriptions from dialects of languages) 8084
Great multiplier 5989
Greater London 1044, 1059
Greek fire 3116
Green fluorescence protein 562
Green function 9594
Green–Kubo formalism 995, 1001, 4386
Green matrix 6614
Greenberg–Hasting model 886
Greenberger–Horne–Zeilinger (GHZ) state
fluorescence measurement 7244
quantum computing 7242
Greenhouse effect 3294, 3295
Greenhouse gases 3294, 3295, 7592
Green's function 5246, 6741, 9640, 9902
Green's law 9606, 9637
Green's theorem 7443
Grey Walter 1118
Grid 1720, 6398, 6401
Cartesian 1720
exotic 1735
polar 1720
pseudo-polar 1721
rectopolar 1720
Grid computing, agent-based modeling 157
Grid environment 2365
Grid generation 7892, 7900
Grid search in space 9913
Grid/cloud computing 4314
Grid/cloud computing are granules 4344
Grim strategy 4101
Grim trigger strategy profile 7656
GRIP 1256
GRN
attractor states 1182
Boolean network 1198
complex 1180, 1181
dynamics 1188
evolution 1187
functional modularity 1191
in living cells 1206
medusa structure 1184
modularity 1184
state 1188
studies 1185
transcriptome 1181
Gröbner basis 6327
Gromov, M. 780, 781, 783, 785, 786
Gronwall's lemma 3264

Gross Pitaevskii 6721
Grossberg 278
Grossman–Aharony hull 6527
Ground motion 1217
Ground-motion complexity 4436, 4456
directivity effect 4450
rotational motions 4467
seismic coda 4437
Ground motion intensity measures 2417, 4435, 4462, 4464
ground-motion scaling relations 4440
peak ground 4435
peak ground acceleration 4435
peak ground acceleration close 4466
PGA-value 4466
precariously balanced rocks 4466
response spectra 4435
rotational motions 4467
S_A-values 4466
scalar intensity measures 4435
strong-ground motions 4436
strong-motion seismology 4437
Ground motion uncertainty 4435
directivity effect 4450
epistemic uncertainty 4465
ground-shaking variability 4468
inter-event variability 4465
intra-event 4465
M 6.0 Parkfield earthquake 4437
M 7.6 Taiwan (Chi-Chi) earthquake 4437
maximum shaking levels 4466
microzonation studies 4460
physical limits to maximum ground motions 4466
quantifying uncertainty 4465
random (aleatory) variability 4435
scientific (epistemic) uncertainty 4435
velocity-pulses 4450
Ground state 6407
Group 778, 4589, 6735, 8291, 8293, 8295, 8296
Abelian 778
action 6597
amenable 778, 783
black-box 7090, 7094, 7095
Cayley graph of a finitely generated 778, 783, 785
character 6597
class 7096
classes 4590
coset and quotient group 4590
dihedral 7101
direct product 4590
finite state automata 4591
finitely generated 778
Følner sequence in a 784
free 778, 784
free Burnside 784
Galilean 3863
Heisenberg 8560
homology 6602
homomorphism, isomorphism, and automorphism 4590
homotopy 6601, 6607
invariance 6736
L-surjunctive 788, 789
Lie 6597, 6598
Lie group and Lie algebra 4590
Lorentz 3868
mean on a 783
non-Abelian 7101
non-amenable 784
normal subgroup 4590
of sub-exponential growth 780
quotient manifold 4591
quotient set 4591
representation of 6735
residual finite 785
residually finite 778, 785
ring 779, 788
semidirect products 4590, 7103
semigroup 4590
semisymplectic 6597
sofic 780, 785, 786
solvable 7096
subgroup 4590
surjunctive 780, 785, 790
symplectic 6597
temporal character 6597
torsion-free 788
transformation group 4591
Group action 2262, 6795
Group cohesion 1555
Group dynamics 8267
Group interaction 7997

Group model building 4476, 4478, 7057, 8988
Group non-membership problem 7187
Group oracle 7187
Group selection 3213
Group shift 8907
Group test wavelet (GTW) 10036
Group theory 8935
Group velocity 8483, 9976
Group velocity dispersion 6139
Group workshops 1853, 1858
Grover's algorithm 7258
Grover's quantum search algorithm 7241, 7386
Growing domain size 4235
Growing point language (GPL) 260, 929
  sequencing decision 266
Growth
  accelerated 4493
  exponential 7843
Growth and competition principle 8943
Growth dynamics 682
  negative feedback (constraints) 684
  positive feedback 684
Growth process 884
Growth rate 1053, 1054, 1056
Growth stocks 3391
GSJ *see* Graph selfjoining
GTS 9613
Guaranteed possibility 6929
Guckenheimer 704
Guiding function 6618
Gum rosin 9815
Gumbel type 3320
Gutenberg and Richter 5022
Gutenberg–Richter (GR) law 661, 1219, 2360, 2413, 2559, 7018
  tapered Pareto distribution 2559
Gyroresonance 25
Gyroscopic systems 6634, 6636
Gyroscopic term
  bundle reduction 5499

H

h index 225
$\hbar$-Pseudodifferential calculus 6757
  elliptic symbol 6758
H^1 6600
H_2 696–699
Haar–Fisz transform 8674
Haar measure 3040
Haar transform 10009–10011, 10014, 10016
Haar wavelet 6365, 10034
Habitat 2737
Habitat fragmentation 2723
Hadamard 6664
  product 7428
HaikuBot 4143
Haken–Kelso–Bunz (HKB) equation
  coexisting tendencies 1550
  complementary 1549
  tendency 1550
Haken–Kelso–Bunz (HKB) model 5718, 5720
Hales–Jewett coloring theorem 3092
Half-adder 5646
Half relaxed limits 6224
Hall set 1037
Halpern 8832, 8836
Halpern and Caltagirone 8837
Halting 5524
Halting problem 287, 5468
HAM 5144, 5146, 5149, 5150, 5154, 5156
Hamachi's example, nonsingular Bernoulli transformations 3061
Hamaker constants 3633
Hamilton equations 5066, 5452
Hamilton filter 451
Hamilton–Jacobi–Bellman equation 6219, 6221
Hamilton–Jacobi–Bellman–Isaacs equation 6221
Hamilton–Jacobi equations 1504, 4540, 4542, 5076, 6237, 6272, 6278
  C^1 subsolutions 4559
  critical equation 4549
  critical subsolution 4550, 4551
  critical subsolutions 4557
  intrinsic metric 4551, 4552
  semiconcave functions 4559
  solutions 4545
  strongly semiconcave 4547
  strongly semiconvex 4547
  subsolution 4549
  subsolutions 4543, 4548
  time-dependent 4556
  viscosity solution 4558

viscosity solutions 4546
viscosity test functions 4551
Hamilton operator 8933
Hamilton–Poincaré equations 5496
Hamiltonian 1503, 2263, 4542, 5064, 5065, 5445, 5754, 6278, 6319, 6323–6325, 6408, 6633, 7139, 7405
conjugate 4544
dynamics 5064
effective 7138
flow 5065
left-invariant 5447
lift 5445
two-qubit interaction 7235
vector field 5445
Hamiltonian chaos 6634
Hamiltonian dynamical systems 6176
Hamiltonian function 6631
Hamiltonian *G*-space 5484
Hamiltonian Hopf bifurcation 2262, 4523
Hamiltonian path problem 2093, 9705
Hamiltonian system 383, 2262, 2330, 2963, 4515, 5064, 5452, 6190, 6433, 6633, 6637
billiard systems 2963
KAM-systems and stably non-ergodic behavior 2964
singular 5961
Hamiltonian vector field 6176, 6713
Hammersley–Clifford theorem 8323
Hammett equation 2200
Hamming distance 7327
Hand 8200
Hand calculation method 3157, 3164, 3165
complex 3165
dynamic flow 3164
fixed flow 3164
simple 3165
Hand clapping 8200, 8203
Handedness 1557
Handicap 8127
Handwriting recognition 4890
Hansch 7437
Hansch-like relationship 7454, 7470
Hansch QSAR models 7427
Hansch's lipophilicity 2200
Haploid genome size 493
HAR 8789
Harary index 6838
Hard ball systems 1006
Hard margin linear SVM classification 2171
Hard margin nonlinear SVM classification 2178
Hard spheres 4215
Hard thresholding 10033, 10036, 10037
Hardness proof 5468
Harmonic 3737
Harmonic analyzer 5474
Harmonic function 3739, 3750
Harmonic instability 2327
Harmonic mean 228
Harmonic measure 3720
growth probability 3725
Harmonic oscillator 5443, 8586, 8632
damped harmonic oscillator 8634
Harmonic synthesizer 5474
Harmonic tremor signal, volcanic sound 4849, 4850
Harrod growth model 8997
Harsanyi model 428, 429
Hartman–Grobman theorem 4726
Hartree–Fock 1309, 8576
Hartree–Fock approximation 8583
Harvesting 3683, 3690
Hasbrouck measure 5401
Hash functions 7097
Haskell model 2596
Hasse-diagram 338
Hausdorff dimension 3880
conformal repeller 6952
ergodic theory 3017
Hautus 3527
Hawaiian fire-fountaining 9785
Hawk–Dove game 3178, 3179
Hawthorne studies 8267
Hazard 3110, 3112, 9831
earthquake 3130
fire zone 3126
flood 3117
future flood 3128
map 2438, 3138
tsunami 3119
type 3112
Hazards warning 1223
^{3}He-rich events 30
Head 6241

Healing length 9694
Health
consequences of climate change 1096
social care management, system dynamics studies 4571
social care, overview 4566
World Health Organization (WHO) definition 1096
Health care 4577, 7061
costs 4578
delivery 4582
market models 685
providers 4582
system dynamics applications 4561
United Kingdom and Europe 4561, 8989
United States 8989
Health insurance 4584
Health management, system dynamics workforce planning models 4574
Health protection 4585
Health treatment, system dynamics 4563
Heat bath 8929
Heat dissipation 5867
Heat equation 1487
boundary control 4806
controllability 4806, 4809
one-dimensional 4813
Heavy bottlenecks 9392
Heavy elements 3279
Hebbian assemblies 1439
Hebbian learning 4873, 5882
Hector mine earthquake 2601
Hedlund G 782
Hegselmann–Krause model 8397
Height defuzzifier 6091
Heisenberg equation 8932
Heisenberg picture 8927
Heisenberg spin glass 4232
Heisenberg time 7168
Heisenberg's uncertainty principle 2091, 4313, 4343, 7446
Hele–Shaw problem 3892
Helfand moment 991, 1000
Heliosphere 21
Helmholtz–Smoluchowski (HS) slip 5578
Hemagglutinin 9751
Hemispherectomy 1903
HEMT 7402
Henderson, Lawrence 8267
Hendrickson's bond testing 2837
Hénon and Heiles system 386
Hénon-like maps 964
Hénon map 407, 3514
Hepp 6666
Hepp's theorem 6664
Heraclitus 3268
Herbert 3689
Herding 3153, 3154
behavior 6483
Hereditary information 3291, 3298
Hereman, W. 91
Herfindahl–Hirschman index 8235
hERG (human ether-á-go-go) 2166
hERG (human ether-á-go-go-related gene) 2146
Herman, Gazis et al. car-following model 9334
Hermann, J.A. 5474
Hessian matrix 703
Heston stochastic volatility model 8792
Heterochrony 1916–1918
Heteroclinic connection 6465
Heteroclinic orbit 2277, 4522
Heterodimensional cycle 4732
Heterogeneity 208, 3375, 4956, 4998, 5004, 5017, 5411, 5428, 6492, 8369
measurement 4956
spatial 5017
temporal 5017
Heterogeneous 610
Heterogeneous CA 837
Heterogeneous chemical reactions 7570
Heterogeneous connectivity 1558
Heterogeneous directed network 6030
Heterogeneous fault zone 5022
Heterogeneous good 3327
Heterogeneous linking network 6029
Heterogeneous market 4110
Heterogeneous teams 2076
Heterogeneously catalyzed reactions 7566, 7573
Heterojunction 8576
Heteroskedasticity 2773
Heterostructure 1018, 5775, 8577
Heuristic 2244
Hexagon 8934
Hexahedral grids 7900

Hexatic phase 2026
Hidden layers 1815
Hidden Markov model 9839, 9846
Hidden population 8225, 8230
Hidden subgroup problem 7095
  Abelian 7095, 7096
  dihedral 7101, 7102, 7106
  Heisengroup group 7102
  nilpotent group 7103
  non-Abelian 7100, 7102, 7113
  normal subgroup 7101
Hierarchical clustering 1148, 5054
  with relational constraints 8263
Hierarchical modeling 457, 6055
Hierarchical ordering, species 490
Hierarchical task networks (HTN) 5192, 5201
  HTN domain description 5193
  HTN planning problem 5193
  methods 5193
  non-primitive 5193
  operators 5193
  primitive tasks 5193
Hierarchy 4324, 4326, 4328, 4329, 5739, 6386, 8247
  complete 8247
  of beliefs 430
  of system specifications 346
HIGFET 8576, 8584
High density regions, crictical point 3792
High-pass filters 592, 595
High performance computing 7908
High-temperature approximation 8588
High-throughput data 5538, 9094, 9097
High-throughput docking 2170
High-throughput screening (HTS) 2161
Higher level instruction 4350
Higher order concept approximation 4353
Higher rank $\mathbb{Z}$ Anosov actions 3098
Highest posterior density interval 449
Highway bottleneck 9282
Hijacked links 5279
Hilbert
  basis 2263
  semispace 7447
  space 3338, 3340, 3353, 4194, 4196, 4197, 5464, 7442
  transform 636, 2940
  uniqueness method 1492
Hilbert–Banach spaces 7442
Hilbert–Sobolev spaces 7444
Hill
  clusters 1608
  equation 2325, 6626, 6629, 6631, 8637
  estimator 3410
Hillis, Danny 5477
Hillman 8844
Hip–Hops, *n*-body problem 5967
Hippocampus 6099, 6112, 6113
Hirota method 8119, 8481
  bilinear 81
Hisaoka 3564
Histone code 1196
Historical earthquakes
  characteristic earthquake 2562
  damage rheology 2563
Historical records, incomplete 9833
History 5026, 7617, 7622
  public 7625
History-independent subgame perfect equilibrium 4107
HITS 5272
Hitting time set 9238
HLA_GRID_Repast 157
HMETIS 5055
Hodgkin–Huxley 6102
Hofbauer's example 6950
Hölder condition 8720
Hölder exponent 3768, 3897
Hölder mean 228
Hölder norms 5087
Holism 207
Holland–Leinhardt model 8228
Holographic 6394
Holography 6395
Holonomy invariance of measures of maximum entropy 959
Holophrastic generator 1891
Homans' principle 8191
Homans, George 8268
Homeobox protein, entropic profile 500
Homeodynamics 482
Homeostasis 565, 6113
Homicide, surge 3301

Homo oeconomicus 8188
Homo sapiens 8400
Homo sociologicus 8188
Homochirality 3286, 3289, 3291, 3294, 3298
Homoclinic bifurcations 3019
Homoclinic orbit 406, 2265, 4522
Homoclinic tangency 4733
Homoclinic tangle 384, 406
Homogeneity 8323, 8326
Homogeneous 3234
Homogeneous cooling state 4381
Homogeneous coordinate 56
Homogeneous directed network 6029
Homogeneous linking networks 6028
Homogeneous market 4110
Homogeneous problems 1682
Homogenization 310, 1399, 6225
Homography 46, 49
Homological algebra 4341, 4349
Homological equation 5989, 6312, 6315, 6714, 6762
Homological operator 6179
Homology 1882, 6602
Homomorphism 8292, 8298–8301
  lattice 8292
  of measure algebras 5461
Homophily 8234, 8316, 8330
Homotopy 6601
  analysis method 5144, 5164, 5169
  perturbation method 5166
Homunculus 4894
Honda's fuzzy logic transmission 3991
Hong–Ou–Mandel (HOM) interference 7255
Hooke, Robert 923
Hooke's law 699, 10050
Hopf algebra 1037
Hopf-algebra 1037
Hopf bifurcation 4515, 5311, 6336, 7605
Hopf decomposition 3057
Hopf fibration 5492
Hopfield 3352
Hopfield network 164, 279, 1434, 4873
Hopping rate 2259
Horizon
  free 1502
  infinite 1502
Horizontal gene transfer 5664
Horizontal hostility 5610
Horizontal lift 5503
Horizontal projection 5503
Horizontal space, definition 5502
Horizontal transmission 3213
Horizontal vector fields
  definition 5504
Horocycle flow 3098, 4985
Horseshoe 3829, 3839
Horst and graben 9588
Hoshen–Kopelman algorithm 6524, 7824
Hospital bed utilization 4569
Hostility
  horizontal 5610
  vertical 5610
House Barter games 1526
Household
  budget data 5553
  income 2285
Households as agents in neoclassical economic theory 178
HPM 5144, 5146, 5147, 5151, 5154, 5156
HPP model 879
HRI
  anthropomorphism 4647
  autonomy 4651
  benchmarks and ethical issues 4649
  cost/benefit analysis 4652
  critical benchmark 4651
  design 4646
  embodiment 4646
  ethical theory 4650
  human factors 4646
  imitation 4651
  multi-modal perception 4645
  privacy 4651
  research challenges 4645
  safety 4650
  scalability 4650
  success relative to human caregiver 4652
  systems 4646
  task-oriented benchmarks 4652
  testbed 4645
HRI *see also* Human–robot interaction
HRI influences, popular culture 4644
HRI-specific sensors 4646
HTML (hypertext markup language) 8086

HTTP (hypertext transfer protocol) 8088
Hub 725, 5272, 8271, 8277
Hubbard model 7347, 9679
  Hubbard gap 7357
Hubbard–Stratonovich transformation 7512
Hubbert's lifecycle method 8998
Hull 8709, 8713, 8722, 8724, 8726
Hull and White stochastic volatility model, average integrated variance 8790
Hull-generating walks 6527
Human 149
Human activity systems, dynamic complexity 4563
Human brain core 1627
Human choice behavior 9432
Human development
  complexity 1874, 1875
  dynamic systems 1880
Human ethology
  cohesion vs. resistance 4612
  dynamics 4611
  ethological issues 4613
  human ethological resistance 4619
Human genome project 417, 9093
Human health
  climate change 1095
  ozone hole 1099
Human intestinal absorption (HIA) 2126, 2145, 2167
Human life 3280
Human perception 9434
Human–robot interaction (HRI) 1225, 4643, 4644
Human serum albumin 2121
Human serum protein 2187
Human serum protein binding 2145
Human sexual network 4660
Human sociopolitical dynamics 4610, 4623
  synchronization 4619
HUMIES 4138
Hund's first rule 4824
  breakdown of Hund's first rule 7317
Hurewicz pointwise ergodic theorem 3058
Hurricane Katrina 3132
Hurst exponent 3711, 3757, 3760, 3768, 3899, 9844
  generalized 3768
Hurst scaling exponent 8712
Hybrid additive rules 65
Hybrid approach, CAP2 1116
Hybrid arc 4676
Hybrid composite 1298
Hybrid computer 275, 276
Hybrid control 6284
Hybrid control systems 9105
Hybrid neural networks 4404
Hybrid oscillator 5725
Hybrid seed corn 8307
Hybrid time domain 4675
Hybrid, well-posed 4675
Hybridization 691
Hydraulic diffusivity 2685
Hydrodynamic field 4382
Hydrodynamic limit 7808
Hydrodynamic Stokes radius 6840
Hydrodynamics 549, 4382
Hydrogen plasma 8583
Hydrological cycle 7597
Hydrological modeling 9015
Hydrophilic 9752
Hydrophobic 9752
Hydrophobic effect 854
Hydrophobicity 2169, 2200
Hydrothermal system 6958, 6969
Hydrovolcanism 9793
Hyman, J. 88
Hyper-netted-chain 8576, 8583
Hyper-space, special fuzzy fucntions 4084
Hyper-universe 6302
Hyperbolic 2268, 8635
  infinitesimal 8635
Hyperbolic attractor 4731
Hyperbolic distribution 3420
Hyperbolic dynamical system 989
Hyperbolic ensembles 7507
Hyperbolic equilibrium 4516
Hyperbolic linear system 4725
Hyperbolic measure 3023
  pointwise dimension 3023
Hyperbolic set 4728
Hyperbolic system 999, 4716
Hyperbolic toral automorphisms 996, 8645
Hyperbolicity 953, 8644
Hypercycle 120, 3292, 8045
Hyperfine qubit 7229

Hyperfine state 9685
Hyperfine state qubit 7224
Hyperfine structure 6734
Hyperfinite coin-tossing, definition 6308
Hyperfinite equi-partition 6296
Hyperfinite random walk, definition 6308
Hyperfinite set, continous function 6296
Hyperfinite time-axis, definition 6307
Hypergraph 585, 586, 8355
Hyperion, chaotic rotation of 373
Hyperreals 6292, 6295
  definition 6292
Hyperredundant manipulators 1476
Hyperscaling relation
  breakdown 1599
  polymer physics 6819
Hyperstatic 6013, 6016
Hypo-elliptic 2263, 2269
Hypocenter 2396, 2474
Hypostatic 6013, 6016
Hypothalamic-pituitary-adrenal (HPA) system 1922
Hypothesis 6252
  on C^{∞}-norm 946
Hypothesis space 5349
Hypothesis testing 6000
Hysteresis 1548, 4894, 5023, 5720, 6636
Hysteresis effect 9284, 9373

I

ICAPM 3396
Ice sheets 1712
  ice sheet dynamics 1712
  land ice models 1713
Iceberg orders, market microstructure 5393
ICEO 5581
ICER 9242
Icosahedral crystal 4229
60-Icosahedral periodic minimizers 5970
ID3 1827
Ideal 703
  maximal 1030
Ideal chain 6818
Ideal non-interacting chain 6813
Idealism 9063–9065
Ideas and behaviors 8307
Idempotent network 8296
Identifiability 5553
Identification 444, 447, 449, 9105
  of the cellular automaton 4739
  partial identification 5560
  point identification 5560
Identity 1905, 8037
Idiosyncratic errors 2772
If-then rules 4011
Ig-like β-sandwich 7040
Ill-conditioned problems 9188
Ill-posed linear problems, solution of 9187
  Bayesian approach 9191
Ill-posed problems 10001
Illusion of life 1337
Image 1719, 3136, 6391, 6394
  binary symbol 6399
  column 6400
  complex-valued 6396, 6398
  geometry 1719
  matrix 6399, 6400
  processing 1719
  row and column 6399
  stack 6399
  two-dimensional 6396
Image-based feedback 1225
Image-based state estimation 1225
Image compression 10031
Image graph 8273
Image inpainting 10031, 10039
Image interpolation 10039
Image Jacobian 48, 52
Image processing 6856
  edge detection 6904
  image decomposition 6904
  image denoising 6904
  image reconstruction 6904
  LSI circuits 9713
Image restoration 10031
Image segmentation 10031, 10039
Image-space 44
Image-space trajectory 51
Imaging 7892
Imitation 6486, 8308
  human–robot interaction 4651
Imitators 1960
Immergence 8407

Immigrant agents 155
Immigration rate 2707
IMMSIM 127
Immune network 4777, 4781
Immune system, homeostasis 4779, 4784
Immunization 6498
Immunocomputing 4739, 4746
Immunology 4777
IMMUNOS 2132
Immunosuppressive peptide 2208
Impact 3112
  disaster 3126
Impact fracture 3803
Impasse 5196
Impedance of the vacuum 33
Imperfect information 4072
Imperfect rationality 5431
Imperial County Services (ICS) building 2385
Imperial Valley earthquake 2385, 2387
Impermeable 6209
Impersonating profiles 4376, 4377
Implementable social choice rule 4788
  multiple implementation 4799
  truthfully implementable 4794
Implementation 4097
Implication 4877, 4879
Implicit function theorem (IFT) 5066, 5083, 5300, 6588, 6593, 6594
  continuation 6594, 6595
Implied volatility
  short definition 8783
Importance sampling 2450, 2459, 5668, 5676, 6995
Impossibility theorem 205, 9944
Imprecise information 4060
Imprecise probabilities, fuzzy logic 4005
Imprecision
  fuzzy logic 3986
  type-2 fuzzy logic 4011
Impulse 6955, 6957
Impulse response analysis 2387, 2390
Impulse response function (IRF)
  linear 3515
  nonlinear 3516
Impulse responses 2383
Impulsive system 4693
Impurities 5750, 5770, 7343
Imputation 1520
Imputation set 1664
IMS *see* International monitoring system (IMS)
In 3294
In-phase 1545, 5720
In-phase synchronization 631
In-plane 700
In-plane stiffness 699
In-vehicle traffic prediction 9529
Ince–Strutt diagram 2323, 2327
Incentive 6978
  compatibility 5514, 5519
  compatibility constraint 6982
Incidence matrix 6241, 6837
Incipient 7827
Incipient infinite cluster 1412
Inclusion model
  crack model 2688
  oblate spheroid model 2687
  tube model 2687
Incoherent type I motif 564
Income 2823
Income diffusion 2817
Income distribution 2813, 2816, 4925
Income prediction, fuzzy system models 4091
Income temperature 2813, 2817, 2818
Incompatibility
  geometric 4184
  kinematic 4185
  principle, fuzzy logic 4006
Incomplete data 7798
Incomplete information 1588
  on one side 433
  on two sides 434
Incompressibility theorem for injective cellular automata 2984
Incompressible fluid motions 3643
Incompressible fluids 5929
Incompressible viscous fluids, Navier–Stokes equations 5926
Incompressible words, c-incompressible 247
Inconsistent belief 434
Inconsistent data 7800
Increasing returns 208
Increasing returns to scale 1680
Incremental approach 4313
Incubation distributions 9042

Indegree 6241
Independence 6998, 8722, 9944
  contextual 1867
Independent increment
  finite-rank 6308
  hyperfinite-rank 6308
Independent loop set (ILS) 8949
Independent sets 173
Independent two-path 8327
Indeterminacy 4049
Index 6246, 7011
  of a network 8296
  of cyclicity 6243
  of memory 1627
Indexical information 2315
Indian Ocean earthquake 8513
Indian rod trick 2338
Indian rope trick 2337
Indicator 3303
  economic 3307
Indifferent fixed point 6951
Indirect dominance 6027, 6035
Indirect evolutionary approach 5686
Indirect inference 217
Indirect interaction models 192
Indiscernibility 4412
Indiscernibility relation 7755, 7787, 7788, 8338
Individual rational level 7628
Individual rationality 1664
Individual timestep 8696
Individual timestep scheme 8696
Individually rational payoff 7618
Induced F ↑ S transition 9364, 9374
Induced fuzzy parameter 8684
Induced OWA 230
Induced traffic breakdown 9284, 9320
Induced transformation 2973, 4985
Induction 214, 1826
Inductive inference 1828
Inductive learning 5188, 5193, 5194
  generalization 5199
  negative examples 5198
  positive learning examples 5198
  specialization 5199
Industrial dynamics 9059
Inertia 6003
Inertia tensor, locked 5503
Inertial Alfvén waves 25
Inertial frame 6426, 6427
Inertial glut 9906
Inertial manifold 4593, 8646
Inertial range 3653
Inexact concept 1791
Infancy 1901
Infective individual 6779
Inference
  non-monotonic 6930
  non-parametric statistical inference 464
  parametric statistical inference 464
Infimum 5462
  definition 6290
Infinite cluster 2840
Infinite-dimensional control system 4805
Infinite-dimensional controllability 4804
  examples 4805
Infinite-dimensional Hilbert space
  quantum mechanics 7370
Infinite-dimensional torus 6716
Infinite ergodic theory 2971
Infinite iterated function systems 3020
Infinitely divisible stationary processes, ergodic
    theory 3076
Infinitesimal 4343, 4344, 6292
  non-Archimedian 6290
Infinitesimal decrease 6273, 6280
Infinitesimal difference 6286
Infinitesimal equivariance 5484
Infinitesimal generator 947
Infinitesimal granules 4313
Infinitesimal hyperreals 6293
Infinitesimal nonequivariance two-cocycle 5486
Infinitesimal rigidity 3102
Infinitesimal time evolution operator, quantum
    simulation 7385
Infinitesimally free action 5486
Infinitesimally hyperbolic 8635
Inflation 1634, 5329
Influence 5999, 6009, 8317
Influence diagram 1854, 1856, 1860
Influence measure 8959
Information 1552, 3275, 8947
  agency problems 5411
  asymmetric information (information
    asymmetry) 5410

biological 498
context-dependent 1556
coupling 1552
functionally meaningful 1540
informationally meaningful 1552
intention to change 1556
loss 3875
non-fundamental information 5418
on aminoacid composition 4827, 4828
on proton–neutron composition 4824
on vertex degree distribution 4829
on weighted distribution 4828
which-way 3871
Information adaptation 7953
Information-based complexity 7119
Information-based models, market microstructure 5397
Information-bottleneck, network module 505
Information compression 7953
Information content
of a chemical element 4822
of a nuclide 4822
of a structure 4822
of atomic nuclei 4824
of molecules 4826
of networks 4834
Information costs, market microstructure 5397
Information curve
modular network 506
NIB algorithm 506
Information descriptors of compositional complexity 4828
Information economics 6977
Information equations for periods and groups of chemical elements 4823
Information exchange 1555, 8199, 8200, 8203
Information extraction (IE) 8094
Information flow 5286
Information for atomic (chemical) composition 4827, 4828
Information fusion 224, 226, 4404, 5741
Information gain 1831
Information granularity 4360
Information granulation and integration 4341
Information granule 1777, 4403, 7740, 7787
Information hiding 4329, 4330, 4335, 4351
Information inflation 7953
Information integration 224, 226, 1433, 4351, 4366
minimum information bipartition 1433
Information integration theory of consciousness 1434
Information measure of statistical memory 1616
Information memory measure 1624
Information models 4305
Information overload 3138
Information processing 4893
in artificial chemistry 341
Information set 4417, 4420, 4862
Information spreading 6061
Daley and Kendall 6061
rumor 6061
Information system 1533, 1793, 4359, 4396, 4420, 7788, 7790
Information table 4314, 7755
Information technology 2349
Information theory 734, 2713
biomolecular sequence 494
ecosystems 2707
module 505
optimizing information flow 735
Informational complexity
evolution 505
Informational forcing 1556
complementary 1556
Informational rent 6984
Informed traders 5397, 5398
Infovis 1256
Infrasonic wave
propagation 4842
propagation modeling 4844
travel time curve 4842
Infrasound 1219, 4839
acoustic speed 4841
acoustic velocity 4841
analysis method 4841
basic principle 4840
earthquake 4839, 4841
measuring 4840
tsunami 4839, 4845
volcano 4839, 4848
Infrasound array 4840
INGV 2405
Inheritance, definition 6289

Inhibition
  feedforward 6112
  global 6106
  recurrent 6109
Inhibitor 6461, 6467, 6468
Inhomogeneous earthquake fault 5024
Inhomogeneous fault zone 5035
Initial-boundary value problem 5930, 5943
Initial configuration 3235
Initial crack, earthquake nucleation process 2540
Initial seeding 8315
Initial set 4497
  fills space 4499
Initial submanifold 5491
Initial value problem 3257, 9225
Injection mechanism 27
Injective 64, 65, 70, 71
  cellular automata 2983
Injectivity 7680
Ink jet printer 5574
Inner Lindblad resonance 412
Innovation 1956, 8308
Innovators 1958
Inpainting 1719, 10006, 10032
Input 6102, 8617
  rhythmic 6105
Input layer 1815
Input-output transformation function 214
Inquiry system 7850
InSAR 4264
Insects 284
Insight 7851
Inspection games 4096
Inspector leadership 4855
Instability 4181, 6117, 6434, 8001, 8003, 8040, 8927
  absolute 6470, 6473
  Benjamin–Feir (self-focusing) 6469
  Eckhaus 6463, 6464, 6469
  geometric 4184
  meandering 6471
  Mullins–Sekerka 6466
  of a fixed point 8930
  of a limit cycle 8932
  of tori 8932
  oscillatory 6468
  physical 4183
  traveling 6468
  zigzag 6463, 6468, 6469
Instability pockets 6626
Instability tongues 6626, 6630
Installed base 1965
Instant-runoff voting 9948
Institution
  market institutions 4624
  network economic 4624
Institutional barrier 3125, 3136
Institutional economics 8996, 8998
Institutional learning 7851
Institutionalist economic models 2285
Instrumental variables 5560
  causality 5560
  LATE 5565
  local instrumental variables 5565
  natural experiments 5560
  weak instruments 5564
Instrumentalism 9063
Insulator–conductor 3850
Insurance 6977, 8127
Integer factorization 5478
Integrability 6770, 7164, 8481
Integrable 2263
Integrable affine structure 4527
Integrable system 401, 995, 1376, 2263, 4516
  integrable 6176
Integral calculus 6298
Integral inequality 1495
  nonlinear 1497
Integral (of motion) 2263, 4515, 5447, 6429
Integral operators 3338
Integral scale 3642
Integrate-and-fire 6101
Integrated genetic-fuzzy problem
  for items with a single minimum support (IGFSMS) 4152
  for items with multiple minimum supports (IGFMMS) 4154
Integrated molecular circuit 5654
Integration 1430, 1881
  limit problem 2220
Integration error 8993
Intelligence 1901, 8869–8871, 8878
Intelligence systems 7851
Intelligent agent 1335

Intelligent control 1109
ISAC 1109
Intelligent data analysis 1776
Intelligent transportation systems 9433
Intelligent work environment 1479
Intensification of downstream congestion 9374
Intensity 6389, 6394, 6399
Intensity cross product 9855
Intensity of a process 9834
non-parametric 9834
Intensity spectral density 7921
Intensive 2865
Intentional
action 1876
behavior 1876
Intentional dynamics
goal-directedness 1556
spontaneous movements 1556
Intentional information
destabilizing 1556
stabilizing 1556
switching 1556
Intentionality 4893, 8948
Intentions 7542
Inter-cell communication 820
Inter-domain routing 4930, 4943, 4945
Inter-onset time 9831
Inter-person distance 3147
Inter-story drift 2371, 2378, 2388
outstanding issues 2379
INTERACT 5357
Interaction 208, 724, 727, 5286
attractive 8483
elastic 8480
human robot 4643
inelastic 8480
nonlinear 8480
pathological 629
repulsive 8483
resonant 8480
synaptic 639
valley–orbit 5771
Interaction-based simulation 161
Interaction center 8491
Interaction gain 2056, 2065, 5357
Interaction networks 514, 516
protein 514
Interactive dynamic visualization 1262
Interactive multiobjective optimization 7778, 7782
Interactome 723
Interception models 1706
Interconnect 5864
Interconnect topology 927
Interdealer market 5408
Interdependent valuations 5520
Interdisciplinary modeling 9032
Interdisciplinary research 8928
Interdisciplinary subject 3277
Interestingness 1986
Interface 2020, 6415, 8713
motion 873
Interface boundary conditions 5918
Interface growth 3372
Interference 7169, 7669, 7671
Interferogram 1692
Interferometric synthetic aperture radar 1692
Intergovernmental coordination group 9592
Interim stage 435
Interiorization 1885
Interlocking directorate 8281
Intermediate filament 1745, 1747, 1750–1752, 1754, 1768
Intermediate-filament 1747
Intermediate modularity 2732
Intermediate normalization 6728, 6735
ϵ_n 6728
Intermediate phase 2833, 2837
Intermediate-value, continous function 6296
Intermittency 3647, 3654, 4921, 6444, 7587, 8011, 8018
temporal intermittency 8019
Intermittent 7584
Intermittent flow 6487
Intermittent turbulence 8526, 8535
Internal correlation 7622, 7628
Internal dynamics 2849
Internal model 5719, 9079, 9085, 9086
asymptotic internal model property 9086, 9088
design of 9088
principle 9079, 9085
Internal (or baroclinic) modes 6211
Internal regulation 8041
Internal revenue service 2813

Internal solitary wave 8112, 8121, 8122
Internal state decoherence 7237
Internal waves 6206
Internality 226
International monitoring system (IMS) 4840
International phonetic alphabet 2310
International working group on rotational seismology 2525
Internet 3138
Internet topology 4930–4934, 4936, 4939, 4940, 4942, 4943, 4945
Interpersonal communication 8307
Interpersonal contact 8308
Interpersonal coordination 8200, 8201, 8422, 8423, 8429
  synchronization 8429, 8430, 8432
Interpersonal influence 8308
Interplanetary shocks 22
Interpolation 594
  fuzzy logic 3987
Interpolatory masks 595
Interpretation 7734
  redundant 7734
Intersection 3146, 4148, 4877
  of two fuzzy sets 4066
  per fracture 6511
  type-2 fuzzy sets 4012
Interseismic deformation 1694, 4267
Interstellar medium 393, 3279
Interval arithmetics 8682
Interval computing 4403
Interval exchange maps 2959
Interval exchange transformation 2920
Interval maps 962, 2962, 6951
  β-shifts 2962
  Continued fraction map 2962
  f-expansions 2962
  Farey map 2962
  piecewise C^2 expanding maps 2962
  Smooth expanding interval maps 2961
Interval type-2 fuzzy logic systems (IT2FLS) 4014
Interval type-2 fuzzy sets 4013, 4014
Interval-valued FSs 4013
Interval-valued type 2 fuzzy sets (IVT2FS) 4082
Intra-plate earthquakes 9584
Intra-species evolution 2701
Intractable conflict 8425
Intractable mechanical simulation 5473
Intrinsic absorption 7918
Intrinsic dynamic moduli 6850
Intrinsic dynamics 1556, 8420–8422, 8425, 8429, 8430, 8432
  internal states 8429
Intrinsic evolution 339
Intrinsic metric Hamilton–Jacobi equations 4551, 4552
Intrinsic viscosity
  in the freely draining condition 6848
  in the non-draining condition 6846
Intuitionistic sets 7739
Intuitive logics 7835
Invariance
  ergodic theory 6944
  scale 3864
Invariance principle 4676
Invariance relation 6188
  elementary 6188
Invariant curve 6436
Invariant manifold 937, 6189, 6768
Invariant measure 2982, 4724, 6232, 6943, 9240
  entropy 6944
  ergodic theory 3017, 6943
Invariant measures for cellular automata 2983
  domains, defects, and particles 2991
  Furstenberg conjecture 2990
  maxentropy measures 2984
  measure rigidity 2988
  periodic measures 2985
  posexpansive and permutative CA 2985
  uniform measure 2983
Invariant scaling 8543
Invariant set 9225
Invariant tests 6255
Invariants 9231
  method of 5497
  of the motion 4599
Invasion
  tumors 914
  Turkish 7843
Invasion percolation 4947, 4948, 6569
  algorithm 4949
  correlated heterogeneity 4956
  fractured system 4957
  gravity 4951

lattice topology 4955
method 4949
modification 4950
pore scale geometric correlation 4955
pore scale structure 4951
pore topology 4952
stratification 4957
trapping threshold 4955
viscous force 4951
wetting 4951
Invasion percolation simulation 4958
Invasion perlocation cluster 4957
Invasive tumor
Anderson's model 917
effects of directed cell motion 916
morphology 914, 916
spatial structure 916
Inventory 5410, 7835, 7847
Inventory models, market microstructure 5395
Inverse cascade 8727
Inverse cubic law 3413
Inverse function theorem 5084
Inverse limit 431, 2975
Inverse participation ratio (IPR) 8394
Inverse problem 2450, 6213, 9631
Inverse refraction diagram 9594, 9632, 9637
Inverse scattering problem 4960, 4962, 4963
Inverse scattering transform (IST) 4960, 8114, 8118, 8119, 8466, 8481, 8505
Inverse semigroup 8297
Inverse theory 1695
Inversion 2450, 8861
Inverted pendulum 2424
Inverted random void model 1450
Inverted-V electrons 34
Involution 6628
Inward matrix product 7428, 7444, 7454, 7455
Io–Jupiter interaction 33
Ion–ion entanglement, quantum computing 7244
Ion mobility 5577
Ion-photon entanglement, experimental setup 7245
Ion–photon entanglement, quantum computing 7244
Ion string, vibrational cooling 7226
Ion trap
diagram 7223
large-scale 7223
technology 7221
Ion-trap cavity-QED 7245
Ion, quantum computing 7220
Ionic qubit, state detection 7229
Ionization, quantum computing 7225
Ionosphere
GPS-TEC 2629, 2632
total electron content (TEC) 2622
IPCC 7845
fourth assessment report 1711
Iron law of wages 2285
IRR 4933, 4934, 4936, 4944
Irrational behavior 9416
Irrational decision making 4976
Irreducible representation 8935, 8936
Irreducible shift space 8894
Irrelevant feature 5349
Irreversibility 3859
Irreversible phase transitions (IPTs) 7573
IS-LM model 8992
IS-LM/AD-AS model 8994
Isaacs 1953
Ishii's lemma 6221, 6222
Ising 6783, 7825
Ising chain 6950
Ising domain wall 8713
Ising Hamiltonian 8712, 8715
Ising model 728, 873, 1602, 4911, 5671, 5676, 6381, 6806, 7569, 8709, 8713, 8714, 8726
critical exponents 1602
Ising clusters 1602
Ising droplets 1603
random field 6416
Ising spin glass 8727
ISIS communication system 265
Islands 8254
Hawaiian 8852
simple 8256
ISNet 2399
Iso-energetic KAM tori 5081
Iso-energetic Poincaré-mapping 4523
Iso-energetic poincaré mapping 2263
Isobaric 6209
Isochors, biological fractals 3795
Isochronous 8639
Isodefective nuclides 4825

Isoelectric focussing 5580
Isolated chain component 9235
Isolated invariant set 9233
Isolated state (internal state) 4348
Isolating neighborhood 9233
Isomorphic 5461
Isomorphic configurations 5618, 5625
Isomorphic transformation, Bernoulli shift 4988
Isomorphism 3055
  Bernoulli property 4988
  Bernoulli shifts 4986
  finitary 4990
  Ornstein's theorem 4987
Isomorphism invariant 4984, 4985
Isomorphism theory 4983
  action of amenable groups 4993
  transformation factor 4993
Isomorphy principle 8934
Isophote direction 10040
Isosceles symmetry 5973
Isostacity 5015
Isostatic 6013, 6016
Isostatic network 2834
Isostatic structures 2835
Isostaticity 5013, 5016
Isotope, quantum computing 7225
Isotopic number 4824
Isotropic 2263, 7527
Isotropic compression 5012
Isotropic diffusion 10037, 10039
Isotropic dilation matrix 589
Isotropic networks 6514
Isotropic scattering model 7923
Isotropic scattering process 7916
Isotropic undecimated wavelet transform 6360
Isotropic wavelet transform 6362
Isotropy subgroup 5487
Israel 7844
IST method 4962
Istanbul 2398
IT2 FS *see* Type-2 fuzzy systems
IT2FLS *see* Interval type-2 fuzzy logic systems
Italian symmetry 5970
Iterated exponential 618
Iterated function systems 3019
Iterated integral functionals 1035
Iteration of functions 3833
Iterative array 800, 803
Iterative diagonalization 7349
Iterative estimation methods 8230
Iterativeness 1880
Ithiel de Sola pool 8277
IThink 7843
ITPFI transformations 3068
IV
  continous-time SV jump–diffussion model 8787
  IV and cumureative squard jump component decomposition 8787
Iwo Jima, Japan 2605

J

J-Curve 4928
J-Measure 1831
Jackson–Moser–Zehnder, lemma 5079
Jackson–van den Nouweland rules 6026
Jackson–Wolinsky rules 6026
Jacobi elliptic 8116, 8119
Jacobi elliptic function 8116, 8119
Jacobi rotation 7441
Jacobi, C.G. 5444
Jacobian estimation 54
Jacobian ideal 704, 712
Jacobi's elimination of the node 5492
Jacquard, Joseph-Marie 5476
Jahn–Teller effect 6690
Jam interaction effect 9376
Jammed 5007, 5008, 5014
  collectively 5014
  locally 5014
  strictly 5014
Jamming 2834, 3144, 4997, 5011, 5017
  anisotropic 5017
  isotropic 5017
Japan 2398, 7846
Japanese firms 7845
Japanese Meteorological Agency 2398, 9593
JAUS 8438
JAUS reference architekture (RA) 8439
Java 423
Java programming and security issues 5610
Java-sandbox security model 5611
Java virtual machine (JVM) 5607

Jay W. Forrester 4577
Jaynes 8936
Jellium 8576, 8579
Jensen–Shannon information divergence 4829
Jensen's inequality 6946
Jerky response 5022
Jevons' logic 5477
Jevons, William Stanley 5477
JMA magnitude 9599
John Scott Russell experiments 8465
Johnson's equation 9996
Join 5462
Join of subshifts 9257
Joining 2976, 3055, 5040
  composition of joinings 5043
  ELF property 5049
  filtering problem 5049
  Markov intertwining 5043
  minimal self-joinings 5044
  pairwise independent joining 5048
  relatively independent joining 5042
  self-joinings 5044
  simple 5045
Joint architecture for unmanned systems (JAUS) 8439
Joint attention 1897
Joint homomorphism 8301
Joint plan 7642
Joint probability distribution 372, 3580
  of the matrix elements 7509
Jointly controlled lottery 1591, 7645
Jointly periodic point 9250
Jones polynomial 7110
Jordan canonical form 6651
Jordan chain 6247
Jost solution 4965, 4966
Jourjine 8844
JPEG 10032
JPEG2000 10032, 10040
Julia set 3706, 3828, 3841, 4992
Jump length distribution 5220
Jump process 8734
Justification 9065–9067
  in system dynamics 9067
JVM *see* Java virtual machine
JXP method 5281

K

K-Automorphism 3063
k-Cohesive groups
  connected 4608, 4611
  connectivity-k 4616
  k-components 4613, 4616, 4627
  k-connected groups 4618
  scalability 4611–4613, 4616–4618, 4625
k-Cohesive resistance 4611
k-Core percolation 2836
k-Core representation 1264
k-Cores 575
K-Descriptive language 4399
k-Means 5053
k-Means clustering approach 1149, 4153
k-Nearest neighbors 1992, 2168
k-plexe 8288
K-Property 2929
  ergodic theory 3062
k-Ring 8254
k-Satisfaction 4400
k-Satisfaction relation 4400
k-Skip bipower variation 8789
K-Space solution 7487
k-Star 8327, 8329
k-Subwebs 2716
K-System 401
k-Triangle 8327, 8329
K-Truth 4400
k-Two-path 8327
K_a/K_s 5657
Kadanoff block construction 8712
Kadomtsev–Petviashvili equation 6456, 8118, 8512, 9994
Kagan–Jackson models 2567
Kaiser–Stebbins effect 1653
Kakutani equivalence, isomorphism theory 4991
Kalman 3526
Kalman controllability decomposition 3527, 3534
Kalman filter 3688, 4753, 4769, 8797
Kalman's observer, state-linear systems 6377
KAM 6681, 6683
KAM *see also* Kolmogorov–Arnold–Moser (KAM)
KAM method 3103
KAM stability 5993

KAM theorem 404, 2342, 8641
KAM theory 1933, 4516, 6674, 6710, 7165
Kamikawa 3553
Kane's computer 5750
Kaplan–Yorke conjecture 3023
Kaplan–Yorke–Young formula 1002
Kaplansky I 781, 788–790
Kaplansky zero-divisor problem 788
Kaplansky's conjecture on stable finiteness of group rings 789, 790
Kardar–Parisi–Zhang equation, quenched 1138
Karman's equation 3657
Karush–Kuhn–Tucker (KKT) theorem 4088
Kasteleyn–Fortuin 7825
Katmai 2607, 2612, 2614
Kato bound 6739
Kato–Rellich theorem 6738
Kauffman bracket 7111
Kauzmann temperatures 4213
Kawahara equation 6456
KD systems
  first-generation 1985
  second-generation 1985
  third-generation 1985
KDD 4392
KDDM model 1985, 1986
KdV 5138–5142
KdV–Burgers equation 10002
KdV equation 5144, 5145, 5147–5156, 5162, 6213, 8466, 8511
  third-order 84
Keiretsu 7846
Kelso paradigm 1545
Kelso's finger experiment 8937
Kelvin 273, 285, 5474
Kemeny 9945, 9950
Kemeny index 9956
Kemeny score 9956
Kemeny winner 9950
Kemeny's procedure 9959
Kendall–Wei ranking 6250
Keplerian orbits 6428
Kepler's equation 1935, 6672, 6677
Kermack 6779
Kernel 2176, 3339
Kernel functions, support vector machines (SVM) 4087
Kerner–Klenov stochastic microscopic three-phase traffic flow model 9346, 9349
Kerner–Klenov–Wolf (KKW) cellular automata 9288, 9346
Kerner's three-phase traffic theory 9303, 9501, 9503, 9513
Kernighan–Lin algorithm 1147
Kerr 7250, 7254, 7258
Kerr effect 8475, 8484
Kerr-lensing 6137
Keynesian 2285
Keynesian cross model 8994
Keystone species 2713
  structural keystone 2735
Khephera 3691, 3692
Khinchin 2868
Kibble mechanism 1642, 1645, 1646
Kibble–Zurek mechanism 9680
Kilauea volcano, Hawaii (USA) 4262, 9868
  baseline monitoring 9869
Killer application 4139
Kin selection 3213
Kinase 2186
Kindling 629, 639
Kinematic models 1477
Kinetic energy 7404, 8582
  organisms 2701
Kinetic Monte Carlo 1315
Kinetic shaping 2015
Kinetic theory 1006
Kinetically constrained lattice gases 4227
Kinetically constrained models 4227
Kinetics 856
  mass-action 329
  steady state systems 2701
Kings 7017, 7018
Kink 80, 84, 85
Kinodynamic planning 5472
Kirchhoff coda migration 7923
Kirchhoff's Law 1405
Kirkwood gaps 6443
Kirkwook–Riseman chain 6839
KKS *see* Kostant, Kirillov, Souriau
KKW model 9290
KKW-model 9288
Kleiber's law, biological fractals 3784
Klein–Gordon equation 6455

KM algorithms, type-2 fuzzy logic 4017
K(n,n) equation 84, 88
Kneading invariants 968
Knight problem 2098
Knight relation 33
Knill 7257, 7258
Knot 7111
Knot theory 9274
Knowledge 4411, 7542, 7850
Knowledge base 4871, 5538
Knowledge discovery 4404, 7729
  data mining 1984
Knowledge economy 8366
Knowledge operation 4352
Knowledge representation 4352, 4412, 8447
Knowledge structures (KS) 4341, 4348
Knowledge system 4396
Koch curve 3702
Kochen, Manfred 8277
Kohn–Luttinger 5771
Kohonen map 1822
Kohonen network 4873
Kohonen self-organizing maps 1784
Kolmogoroff scaling exponent 8727
Kolmogorov 704, 5067, 5068, 5074, 6664
  normal form 5067
  set 5074
  transformation 5068
Kolmogorov -4/5 Law 3653
Kolmogorov–Arnold–Moser (KAM) invariant surface 6195
Kolmogorov–Arnold–Moser (KAM) theorem 383, 6203
Kolmogorov complexity 245, 492, 5439
  conditional 493
  given a representation system 244
  mutual 494
Kolmogorov endomorphism 3002
Kolmogorov equations 9225
Kolmogorov forward equation 2801
Kolmogorov group, property 8555
Kolmogorov operator 6231
Kolmogorov probability space 6301
Kolmogorov scales 3651
Kolmogorov–Sinai entropy 989, 992, 1001, 1006, 1007, 4737, 4986
Kolmogorov spectrum 29, 31
Kolmogorov typicality 953
Kolmogorov's algorithmic information 4829
Kolmogorov's method 6666
Kolmogorov's theory 3642, 3654
Kondo effect 7344, 7351, 7355, 9647
  Kondo energy 9647
  Kondo resonance 7352
  Kondo temperature 7352
  Kondo tunneling 9647
  spin-$N/2$ Kondo effect 7357
  SU(4) Kondo effect 7355
  two-stage Kondo effect 7354
Kondo model 7345
  two-channel 7356
  two-impurity 7353
König, René 8267
Koopman operator 2921
Korean family planning study 8311
Kornienko 8944
Korteweg and de Vries 8504
Korteweg–de Vries equation 5144, 5155, 6138, 6454, 6709, 6712, 8115, 8122, 9970, 9983, 9984, 9988, 9993
Korteweg, D.J. 81, 84, 8114, 8115
Kostant, Kirillov, Souriau coadjoint form 5487
Kosterlitz–Thouless transition 1276, 9680
Krakatau, 1883 9786
Kramers chain 6839
Kramers equation 8778
Kramers–Moyal coefficients 3584
Kramers–Moyal expansion 3593, 8773
Kramers problem 5229
Krause–Hegselmann model 6385
Krauß-stochastic microscopic traffic flow model 9336
Krein collision 2263
Krengel's entropy 3071
Krieger generator theorem 4986
Kripke semantics 8340
Kronecker factor 2927
Kruskal, M.D. 84
Krylov–Bogoliubov theorem 957, 6232
*k*th pOwer of a fuzzy set 4068
Kuhn, Thomas 2860
Kuhn's theorem 5179, 7617
Kuiper belt 387, 6445

Kullback–Leibler distance, stochastic processes 3585
Kullback–Leibler information 3518
Kuramoto model 5242, 8913
Kuramoto–Sivashinsky equation 6467
Kurokawa 8840
Kurtosis 3407, 5418
  excess 3407
Kutrib 3552
Kuznets hypothesis 2820
Kyle model 5400
Kyoto protocol 9029

L

L-Shaped mapping 3559
L^{∞}-spectrum, ergodic theory 3069
L-Systems 119
l_0 Minimization 10044
l_1 Optimization 10044
L2–norm 2407
Lab on a chip 5574
Labeled links SDCA model 8830
Labeled walk 8291
Labelled graph 9249
Labor 2285
Labor intensive 2285
Labor productivity 2285
Labour market model 8127
  Tesfatsions's 181
Lack of information
  on both sides 7640
  on one side 7630
Lacunarity 396, 478
Lacunary sequence 3087
Laflamme 7258
Lagrange 2870
Lagrange multiplier 5165, 10038
  support vector machines (SVM) 4087
Lagrange–Poincaré bundles 5496
Lagrange–Poincaré equations 5496
Lagrange symmetry 5973
Lagrangean submanifold 2261, 2263
Lagrange's principle 8632, 8640
Lagrangian 4555
Lagrangian equilibrium 5994
Lagrangian fluid dynamics 3658
Lagrangian map 3645, 3647
Lagrangian point 414, 6681
Lagrangian reduction 5495
Lagrangian torus 4519
Laing's molecular machine 334
Lake ice 1707
Laman's theorem 2830, 2838
Lamarckian evolution 3209
Lamb–Dicke limit, quantum computing 7236
Lamb–Dicke parameter, quantum computing 7232
Lamb–Oseen vortex 3645, 3646
Lamb shift 6664
Lambda-calculus 330
λ-Discounted game 7633
Lambda phage 422
Lambda set 8288
Lamellipodium 1742, 1743, 1746, 1749–1752, 1757, 1765, 1771, 1774
Laminar 6489
Laminar flow 6195, 6196
Lance-Williams updating formula 5054
Lanczos 5755
  block 5773
Land surface schemes in GCMs 1707
Land use 7983
  transport model 1062
Landau 6665
Landau–Ginzburg model 1645
Landau–Hopf bifurcation 8931
Landau level 8586
Landau theory 8931
Landau–Zener transitions 8609
Landauer–Büttiker formalism 7296, 8600, 8608
Landauer–Büttiker formula 7484
Landauer, Rolf 1013
Landauer's principle 7697
Landers earthquake 2601, 4264
Landesman–Lazer condition 6622
Landscape matrix 2737
Landslide 3824
  submarine 8851
Lane formation 6481
Langdon 3547
Langevin 6104
Langevin dynamics 3367, 3370

Langevin equation 2877, 3365, 3370–3372, 4589, 5224, 5232, 5246, 5733, 8712, 8714, 8722, 8751, 8775, 8927, 8931, 8937
dissipative 8752
generalized 3577
irreversible 8752
Markov diffusion process 8752
polymer physics 6813
Langevin force 2031
Langevin methods 730
Langevin processes
discrete 3577
short time propagator 3580
Langmuir–Hinshelwood mechanism 7573
of surface reaction 1311
Langton, Chris 8067
Langton's loop 118, 8070
Language 2310, 3111, 3353, 6386, 6808
fuzzy 7839
of dynamics 1559
Language development 1878, 1891, 1899, 1902
Language evolution 1879
Language family
bracketed context-free 812
context-free 808
Dyck languages 812
hierarchy 808
linear context-free 809
PSPACE complete 812
recursively enumerable 812
semilinear 812
Language learner, real-time speech processing 2318
Laplace 6663
Laplace equation 9972
∞-Laplace equation 6219, 6221, 6226
Laplace expansion 1868
Laplace operator 7443, 10037
Laplace operator on a group *see* Laplacian (discrete), *see* Laplacian (discrete)
Laplace pressure 3617
Laplace's limit 6672
Laplacian 6241, 10042
combinatorial 6241
normalized 6241
Laplacian (discrete) 782, 786
Laplacian growth 3723
Laplacian matrix 5354, 6250, 6837
Laplacian score 5354
Large deviations 7013, 7523
Large deviations problems 6233
Large earthquakes 2542
aftershock 2550, 2551
estimate of source processes 2549
initial rupture processes 2550
large amplitude waveforms 2543
main phase 2543
mainshock 2550, 2551
observed waveforms 2543
ordinary phase 2543
P-waves 2543
phases 2551
velocity waveforms 2543
waveforms 2549, 2550
Large eddy simulations 3650
Large-*N* expansion 7347
Large networks
architectural features 1201
dynamics 1201
Large-scale instability 6118
Large-scale ion trap, quantum computing 7223
Large scale simulations, agent-based modeling 148
Larkin 1129, 1134
Larkin model 2023
Laser 8928
Laser clocks 6169
Laser cooling, quantum computing 7225
Laser dynamics 6127
Late heavy bombardment 6451
Latency reduction 5614
Latent attractors 8425
Latent class models
finite mixture model 5567
information criterion 5567
nonparametric maximum likelihood 5567
Latent factor models, simulated likelihood 5564
Lateralization 1902
Lattice 1395, 3096, 6573, 6784, 8292, 8295, 8297–8299, 8301, 8303, 8304, 8356
Archimedean 6579
Galois 8356
hexagonal 6583
Kagomé 6585

square 6583
triangular 6583, 6585
unimodular lattice 3029, 3032
Lattice animals 6545, 6550
Lattice Boltzmann model 883, 916
Lattice coordination number 6507
Lattice dynamics 8469
Lattice gas 610
Lattice gas automata 879
Lattice-Gas reaction models 7576
Lattice mismatch 5914
Lattice molecular system 335
Lattice percolation 6507
Lattice property 9656
Laumann, Edward 8268
Laurent polynomial 10007
Lava domes 9765
Lava flow 9785
Lavoie's 8119
Lavoie's equations 8119
Law 1540, 1558
composition 3864
Darcy 9775
of requisite variety 8974, 8981
of web surfing 10062
scaling 3864
Lawn movers 3690
Lawton, W. 780, 785, 786
Lax method 4960, 4963
Lax pair 4963
Layout format 1266
Lazy learning 2168
LCB *see* Long chain branched" (LCB) melt
LCR algorithm 7713
Le Lann–Chang–Roberts (LCR) algorithm 7718
Lead compound 2197
Lead time 2407
Leader election 7713, 7717
Leader, amorphous computing 264
Leadership 9048
adaptive leadership 8369
administrative 8369
enabling 8369
strategic 8369
Learnability 5183
of language 1878
Learning 128, 280, 1557, 1827, 1886, 2143, 2854, 5177, 6083, 8370, 9105
argument 4103
asymmetrical avoidance behavior 6477
automaton 4739
Bayesian 5179
belief 5179
classifier system 121, 128, 4739
cycle, IPRD 7851
deterministic 5178
evolution 1879
game 1862, 4097
Kalai and Lehrer (KL) 5182
laboratory 9050
learnable best response property (LBR) 5182
machine learning 128
neuro-fuzzy systems 6094
non-supervised 8194
organization 7057
organizational 7851
paradigms 1815
process 6477
rate 1817, 1820, 1823
rational 3378
reinforcement 6084
reinforcement learning 128
rule 1813, 1815
sophisticated 5177
stochastic 5178
supervised 128, 6084
teaching 1873
unsupervised 128
Least-squares 3352, 3355
Least-squares estimator
damped 9189
generalized 9193
ordinary 9189
Least-squares method 5208
Leath 7824
Leath method 6523
Leave-one-out validation 1816
Lebesgue density points 2281
Lebesgue space 5461
LEDA 1256
Ledrappier–Young theorem 3023
Legendre 2863
Legendre polynomials 6677

Legendre transform 8540
Legendre transformation and reduction theory 5496
Legislative processes 7060
Lehmer mean 229
Lehmer, Derrick 5478
Lehmer's number sieve computer 5478
Leibniz 7837
Leibniz' product rule, differentiability 6298
Leibniz program, definition 6289
Leibniz, Gottfried Wilhelm von 5476
LEM2 algorithm 1799, 7798, 7799
Lemma of Neyman and Pearson 4860
Length
  bootstrap percolation 610
  Compton 3874
  fractal 3862
  Planck 3868
Length-clamp 7033
Length-scale 3868, 3929
  invariant 3868
Lennard–Jones polymer 6823
Lenses 1264, 6394, 6399
Leontief, Vassily 2793
Leptokurtosis 3404
LERS 7798, 7800
Lesche stability 2861
Level 6246
  of detail 1264
  of social reality, macro level 8427
  operational 3155
  strategic 3155
  tactical 3155
Level crossing 6727, 6737
Level repulsion 7166
Levels of interoperability 352
Levels of social reality
  levels of description 8427
  macro level 8422
  macro level of description 8427
  micro and macro levels 8427
  micro level 8427
Leverage 9044
Lévi-Strauss, Claude 8274
Levine, Joel 8268
Lévy flight 5223, 6203
Lévy flight foraging 3792
Lévy flights 3725, 5220, 6195
Lévy noise 3577
Lévy process
  stochastic processes 3578
Lévy processes 3584
  short time propagator 3580
Lévy–Smirnov distribution 5221
Lévy walk models 3792
Lévy Walks 5232
Lewin, Kurt 8267
Lewis, D. 6928
Lexical growth 1881
Lexical knowledge 1888
Li–Yorke pair 9237
Liapunov exponent 384
Liapunov function 2279
Liberalization 4928
Library screening 2170
Libration 6684
Librational Tori 6682
Lie 6313
Lie algebra 6317, 6761
  bundle 5505
  free 1036
  Nilpotent Lie algebra 6184
  semisimple Lie algebra 6184
  subalgebra 6317
Lie bracket 3530
Lie groups 2003, 5440, 6177
  derivative 6313, 6318
  group 6317
  left-invariant 5446
  Lie algebra 5446
  right-invariant vector fields 5446
  subgroup 6317
Lie–Poisson bracket 5492, 5494
Lie–Poisson bundle 5505
Lie–Poisson reduction 5492
Lie–Poisson structure 2263
Lie series, exponential 1037
Lie transforms 6180, 6633
Liénard differential equation 6615
Liesegang pattern 885
Life 1550
Life cycle 1059
Life Era 3274

Life forms 3274, 3280
variety, describing solutions 2700
Life span 1876
Life span overview 1900
Lifting
multiwavelet 5853
scheme 596, 6356
step 10007
Light bullets 6454
Light signal control 1178
Lighthill–Whitham–Richards traffic flow model 9329
Likelihood function 445, 448, 459, 2450
Likelihood test 2444
Lim sup 9228
Limit
quantum 5863
speed-of-light 5863
thermal 5863, 5867, 5869
Limit cycle 527, 528, 5724, 8927, 8930
oscillators 1557
stochastic processes 3586
Limit measures and other asymptotics for cellular automata
asymptotic randomization by linear cellular automata 2993
Hilmy–Hurley centers 2998
hybrid modes of self-organization 2996
Kůrka–Maass μ-limit sets 2998
Kůrka's measure attractors 2999
Milnor–Hurley attractors 2997
Limit order 5418, 5433
Limit order markets 5393
Limit point bifurcation 6336
Limit point of cycles 6336
Limit pricing 8135
Limit set 249, 4724, 6629, 9080, 9255
of a given point 9081
of a set 9081
orbit 249
Limit theorems, stochastic processes 3578
Limited dependent variable (LDV) models 5554
latent variable 5554
Limited hyperreals 6293
Limiting factors 8991
Limiting point
definition 6290
Limits to growth 4634, 8997
Linda coordination model 192
Lindemann criterion 2026
Lindenmayer systems 119
Lindstedt 5080
algorithm 6663
series 1933, 6667, 6668, 6670
Line diagram 8303
Line graph 4666, 6837
Line J for the downstream jam front 9387
Line-narrowing 6125
Line symmetry 5972
Linear 1817, 8954
Linear and quasilinear model 8960
Linear autoregressive (abbreviated, AR) model 442
Linear bounded automaton 170
Linear CA 2981
Linear cellular automata, asymptotic randomization 2993
Linear damping 6635
Linear dynamic system 8949, 9914
Linear free energy relationships (LFER) 2200
Linear functional, multiplicative 1030
Linear Hamiltonian system 6433
Linear interface model 1131, 1132, 1136
Linear inverse theory 9909
Linear ion trap
quantum computing 7222
Linear logic 5286
Linear matrix inequality 8755
Linear momentum 9907
Linear neural networks 1817
Linear nonequilibrium thermodynamics 5260
Linear orders, single-peaked 9958
Linear Paul trap, quantum computing 7221
Linear plane waves 9969
Linear polymers 6546
Linear prediction 10012
Linear program 6420
Linear programming 5536, 5543
Linear response 5233
Linear scale of events 1622
linear Schrödinger equation 9996
Linear stability 6200
Linear stability theory 6195
Linear superposition 8480

Linear system 3526, 8619
controllability 4807
Linear TDMA broadcast 2054
Linear thinking 7840
Linear transformation 171
Linear vector space 5713
Linear velocity 59
Linear vs. nonlinear system 8952
Linearization 6273, 6311, 6657
continuous 6313
formal 6312
holomorphic 6324
smooth 6313
Linearization principle 3528
Linearized equation 8634
controllability 4814
Linearized stability 8633
Lines 8246, 8267, 8270
Linewidth 6124, 6141
Linewidth broadening 6118
Linguistic expression 4883
Linguistic quantifier 4074
Linguistic structure 4342
Linguistic uncertainty
type-2 fuzzy logic 4011
Linguistic variable 4069
fuzzy logic 3986, 4004
Linguistic words 4404
Linguistically quantified statement 4074
Link addition 5700
Link-based ranking algorithm 10066
HITS 10065
PageRank 10065
Link deletion 5700
Link number 3810
Link popularity 5271
Link rules 8819, 8841
couplers 8819, 8828
decouplers 8819, 8828
outer-totalistic 8818
restricted totalistic 8819
substitution rules 8847
totalistic 8818
Link spam 5271
Link-spamming 5266
Link strength 2729
Link tracing 8214
Linkage 735
Linking 3135
Links 724, 727, 1600, 7111
Links-nodes-blobs model 1449
Lin's context based reasoning method 4314
Lin's method 5303
Linux cluster 5758, 5777
Liouville dynamics 4381
Liouville equation 997, 1003, 1621, 4381, 7810
finite-difference Liouville's equation of motion 1622
Liouville–Jacobi formula 6430
Liouville operator 1621, 4588
Liouville quasioperator 1619
Liouville theorem 4588, 5453, 6433
Lippmann–Schwinger equation 6741
Lipschitz continuous function 3256, 4543, 4545, 4552
Liquid 5929, 6955, 9874
constitutive equation 5929
non-Newtonian 5931
Oseen fundamental tensor 5939
planar motions 5930
pure shear flow 5930
steady bifurcation 5935
three-dimensional flow 5939
two-dimensional flow 5930, 5941
Liquid crystal model 5241
Liquid like regions
critical point 3792
Liquid-solid transition
polymer physics 6818
Liquidity 5408
customer effectively demands instantaneous liquidity 5423
overnight liquidity 5423
Liquidity traders 5398, 5412, *see also* Uninformed traders
Listeria 1744, 1745, 1753–1757, 1766, 1768, 1769, 1771–1773
Literacy 2321
coupled oscillators 2312
knowledge 2316
rhythm pattern 2311
words 2311
writing 2316
Lithosphere 4179, 7918

Lithosphere-atmosphere-ionosphere (LAI) coupling 2623, 2628, 2630–2632
Litigation 1846
Littlewood's conjecture 3106
Living things 1538, 1552
LLLE 7997
Lloyd Warner, W. 8267
LLS model 96
LMF *see* Lower membership function
Load balancing 7906
  agent-based modeling 154
Loading rate 7030
Local bifurcation 2263, 4523
Local control center 2400
Local-dissipation theory, Rouse 6812
Local drag, single chain dynamics 6820
Local earthquake 7915
Local earthquake magnitude methods 2665
Local entropy 967
Local equilibrium 994
Local ergodicity 961
Local field 8588
Local frequency 8588
Local function 7680
Local geometry in amorphous computing 265, 267
Local GrC model, first GrC model 4344
Local hamiltonian problem 7186
Local interaction 3197–3199
Local intermittency measure, LIM 8535
Local largest Lyapunov exponents 7997, 7998
Local limit theorem 965
Local linear estimator 3511
Local majority voting 3235
Local majority voting rule 3235
Local minimum 1818
Local modification, agent actions 188
Local network 8292, 8295
Local point of view, agent interaction 193
Local product structure 4729
Local rigidity 3096
Local rule 9250
Local site effects 4440
Local social neighborhood 8323
Local stability model 7042
Local structure, network 507
Local travel times, computation of
  bending methods 9185
  finite-difference approximation 9185
  Huygens's principle 9185
  pseudo-bending method 9185
  shooting method 9185
  shortest-path method 9186
Local uniqueness of invariant mainfolds 945
Local unstable manifold 4729
Local update function 165
Local vertex invariant 2201
Local well posedness 2040
Locality 8709, 8724
Locality property 8724
Localization 4921, 6071
Localization length 2256
Localized coefficients 6715, 6717
Localized electronic excitations 8592
Localized SP 9364
Locally compact groups, unitary representations 3079
Locally free actions 5486
Locally minimizing solution, *n*-body 5978
Locally stationary 8676
Location 2396, 5428
Lock-in 4640
Locke 1883
Locked inertia tensor 5503
Lockstep 8696
Loeb construction, non-standard stochastic analysis 6306
Loeb measure theorem 6306
Loewner equation 8709, 8719–8721
Loewner evolution 3729, 8709, 8720, 8722
Log fitness, network 505
Logarithmic correction 5027, 5028
Logarithmic type potentials, *n*-body problem 5981
Logarithmic waiting times 7586
Logic 7850
  of chemical systems 330
  possibilistic 6927
Logic circuit, quantum computation 7378
Logic gates 2090
Logic operation, artificial chemistry 331
Logical circuit 9705

Logical constructions
  symbols 6302
Logical control of computations 5477
Logical incrementalism 7840
Logical Kronecker symbol 7448
Logical mobility of programs 5608
Logical neuro-fuzzy system, architecture 6093
Logical quantum bit, error correction 7243
Logical systems 6088
Logical-type neuro-fuzzy systems 6089
Logically universal 7689, 7702
  logically universal 7698
Logistic curve 4921
Logistic function 1817, 8307
Logistic growth 1054, 4921
Logistic map 370, 406, 3506
  dynamic fractals 3793
Logistic regression 2569, 6004
Lognormal distribution 1050, 1051
Logons 3352
log(t) 5025
Loma Prieta 9201
Long chain branched" (LCB) melt 6822
Long period 6956
Long-range correlations 3758, 6574
  biological fractals 3791
Long-range elastic stress transfer 5025
Long-range interactions 4610, 4613
  long distance trade 4625
  maritime trade 4625
Long-range memory effect 1626
Long-range negative feedback 562
Long-run effects 7053
Long-term 1888, 1906
Long-term correlations 3764
Long-term evolution 7994
Long-term memory 3712, 3713, 3716
Long-term processes 1875
Long-term traffic prediction 9514
Long-time behavior, Navier–Stokes 5953
Long valley caldera 2605, 2610, 2613, 2617, 9868
Long-wave approximation 9604
Long wavelength parameter 9984
Longevity 2084, 3208
Longitudinal and transverse modes 9924
Longitudinal methods 8276
Longitudinal mode 6124
Longitudinal velocity increment, stochastic
    processes 3590
Lookup table (LUT) 871, 3233, 3234
  transition function 3233
Loop 4487, 4493, 6029, 8246
Loop contribution 8952
Loop eigenvalue Elasticity analysis (LEEA) 8948,
    8949
Loop erased random walk (LERW) 8709, 8713,
    8714, 8722, 8724
Loosely Bernoulli transformation 4991
LOQC 7257, 7258
Lord-of-the-Rings trilogy 1339
Lorentz gas 1001, 1002, 1005, 1006, 1009
Lorentzian lineshape 6123
Lorenz
  core 1667
  dominates 1666
  order 1666
Lorenz attractor 8645
Lorenz curve 2801, 2814
Lorenz equation 8645
  stochastic 3586
Lorenz–Haken equations 6133, 6172
Lorenz-like strange attractor 4734
Lorenz, Edward N. 6132
Lorenz's equations 6132
Lorrain, François 8268
Loschmidt echo 7170
Losonczi's means 229
Loss (cost) function 3475
Loss function 3479, 3491, 3493–3498, 3500
Lotka–Volterra model 3362, 3363, 3366, 3369
Louis Michel 6187
Love or Rayleigh wave 2615
Love wave 2611
Lovelace 5477
Low density limit 7808, 7809, 7819
Low density regions, critical point 3792
Low-dimensional dynamical systems 3022
Low-dimensional semiconductor nanostructures
    5912
Low dimensionality 8544
Low molecule numbers 555
Low-pass filter 592, 595
Lower and upper solutions 6619–6621

Lower approximation 4425, 7776, 7787, 7800
concept lower approximation 7802
singleton lower approximation 7802
subset lower approximation 7802
Lower dimensional tori 5068, 5082
Lower membership function (LMF) 4013
Löwner equation 7025
Löwner, Karl 8710
Lowry model 1062
ls-set 8288
Lubrication approximation 3633
Lubrication limit 5583
Lucas critique 205, 207, 218, 5330
Ludgate 5476
Luenberger observer 6374
Łukasiewicz *s*-norm 4067
Łukasiewicz *t*-norm 4067
Lumen 532
Lumping 4604
Lundgren hierarchy 3649
Lundgren spiral 3646
Lung
biological fractals 3783
diseased 3784
Lusternik–Schnirelmann category 6622
LUT 3234, 3238
Luttinger liquid 9680
LW *see also* APS operator
Lyapunov exponent 953, 989, 992, 998, 1001, 1003, 1006, 1007, 2862, 3023, 7169, 7608, 8912, 9251
global 3506
local 3506, 3514
Lyapunov function 51, 6272, 6273, 6280, 8639, 9225, 9231
converse theorems 4682
patchy control 4686
semiconcave 6281, 6283
strict 9227
Lyapunov functional 5242
Lyapunov-like index 3517
Lyapunov–Schmidt method 5300
Lyapunov spectrum 3025, 3507
Lyapunov's equation 8640
Lyapunov's theorem 8640
Lévi-Strauss, Claude 9411

M

M-Wavelet frame 590
M8 series 2569
MAC *see* Message authentication codes
Mach number 26
Mach reflection 8489
Mach stem 8489
Machì A 780, 784
Machine cognition 1111
Machine intelligence, human-competitive 4138
Machine learning (ML) 1303, 1827, 2161, 4403, 4739, 4777
Machine perspective 1542
Macro 1543
Macro element 5800
Macro-level 8185
Macro-micro-macro questions 6008
Macro-operator 5190, 5199
Macrocosm 7849, 7850
Macromolecule 6014, 6016, 6020
Macrophage 9098
Macroscale 1399
Macroscopic level 5733
Macroscopic modeling 9446
Macroscopic variable 2678, 7993
Macrostate 4378, 4380
polymer physics 6813
MACSYMA 6633
Madhavan, Richardson and Roomans model 5402
Maeda 3564
MAFIA 5061
Magic word 9253
Magma 2616, 6955, 9785, 9873
Magma chamber 6956, 6965, 9764, 9766, 9770, 9776
Magma discharge rate 9773
Magma extrusion 9777
Magma intrusion, subsurface intrusion 9870
Magmatic and hydrothermal fluids 9900
Magnet 5022
polymer physics 6815
Magnetic bubbles 2025
Magnetic field 22, 3825
fluctuation, quantum computing 7238
Magnetic helicity 9269
Magnetic length 8586

Magnetic moment 7343
Magnetic ordering 7355
Magnetic permeability 1399
Magnetic reconfiguration 8529
Magnetic reconnection 37
Magnetic resonance 8594
Magnetic spectroscopy 7350
Magnetic term 5491
  bundle reduction 5499
Magnetically diluted solid 8594
Magnetism 7823
Magnetization 3850
Magnetoencephalogram (MEG) 1627
Magnetoencephalography (MEG) 1436
Magnetohydrodynamic (MHD) equations 8523
Magnetohydrodynamics (MHD) 6721, 8010, 8526, 9272
  reduced magnetohydrodynamics 8013
Magnetosphere 21
Magnification curve 2475, 2480
Magnification, static 2477
Magnitude 2396, 2474–2476, 2481, 3765, 9594
  automatically determined 2491
  band-limited 2482
  broadband 2482
  earthquake 2664
  hawaii earthquake 2665
  local 2396, 2475, 2477, 2479
  local earthquake 2665
  rapid 2478
  Richter 2396
  series 3765
  teleseismic 2491
  traditional amplitude based 2665
Magnitude determination
  automatic 2494
  real-time 2494
Magnitude discrepancy 9580
Magnitude saturation 2484, 2486, 2487
Magnitude scale 2476, 2477, 2484
Magnitude units 2477
Maharam extension, ergodic theory 3065
Majercik 8829–8831
Major features: phase transition, volatility and predictability 5593
Major histocompatibility complex (MHC) 2121, 2146, 2191
Majorana 6805
Majoritarian voting 9947
Majority action cellular automaton 782
Majority-choice approval 9954
Majority logic 9722
Majority relation 9943
  strict 9943, 9949
Majority rule 872
Majority vote 9935, 9938
Majority voting 9935
Majorizing series 6764
Malaria, effects of climate change 1101
Malicious hosts 5611
Malicious mobile agents 5610
Malthus–Verhulst model 15
Malthus, Robert 3793
Malthusian effects of resource scarcity 4623, 4626
Malware propagation 4369
Mamdani approach, fuzzy inference reasoning 6088
Mamdani model 4708, 4882
Mamdani system 4882
Mamdani-type neuro-fuzzy systems 6090
Man-made structure 2421
Management exercises 7059
Management flight simulator 1980, 9050
Management laboratory 1980
Management science 9015
Managerial practice field 9050
Manakov (system) 1376
Mandelbrot–Given fractal 3705, 3880, 3891
Mandelbrot set 3706, 3841
Mañé 969
Manifold 1129, 1131, 1135, 5440
  contact 1135
  differentiable 6600
  driven 1131
  elastic 1126, 1129
  Hilbert 6600
  Poisson 6598
  Riemannian 6600
  Riemannian manifolds 5440
  symplectic 6587–6589, 6597, 6603
  tangent and cotangent spaces 5440
Manipulation 3331, 9960
Manna model 645
Manning condensation, polymer physics 6817

Mantle magnitude 2483
Mantle magnitude (M_m) method
  tsunami warning 2667
Mantle surface-wave data 2488
Manufacturing 1970
Many-body effect 8577, 8578
Many-body problem 7313
Many-chain fluid, polymer physics 6817
Many particle physics, stochastic processes 3587
Many-particle problems 7346
Many-spin clusters 8595
Many-valued implication 6936
Map 3034, 3036
  Gauss map 3030, 3032
  momentum 5484
  nondegenerate 3034, 3036
  of science 1258
  Poinaré 6592
  Poincaré 6593
Maple 8848
Mapping 4677
  locally bounded 4677
  outer semicontinuous 4677
  regularization 4679
  set-valued 4672
  structure-to-function 330
Mapping rules, DNA 3796
Mapping site 65, 67
Marakhovsky 3547
Marangoni effect 3620
Marchal's P_{12}–symmetry 5968
Marchenko integral equation 4968, 4969
Marchenko kernel 4968, 4969
Marchenko method 4968, 4969
Marcovian stochastic process 9291
Marginal 8952
Marginal cost pricing 1675
Marginal distribution 1866
Marginal factor cost 2285
Marginal likelihood 446, 449, 452
Marginal pricing rule 1675
Marginal rate of substitution 1678
Marginal revenue product 2285
Marginal stability 32, 7583, 7587
Margolus partitioning 7161
  generalized 7159
Margolus partitioning scheme 7157
Mark and destroy 2097
Marker propagation for amorphous particles 260
Market 3385, 6783, 9566–9575
  micro-structure 3385
  order-driven 3386
  secondary 9565, 9567, 9568, 9574
  US treasury 9566, 9568
Market anomalies 92, 3457, 3458
  daylight saving effect 3472
  daylight-saving effect 3458
  January effect 3458, 3469
  Monday effect 3469
  mondays effect 3458
  risk-free rate puzzle 3467
  SAD effect 3458, 3470, 3472
  turn of the month effect 3469
Market beta 3391
Market capitalization 3391
Market economy 2285, 2804
Market entry 1962
Market games 4097
Market inefficiency 3449
Market maker 3380, 5398
Market microstructure 5392
  Copeland and Galai's model 5397
  Easley and O'Hara's model 5398
  empirical 5401
  Garman's model 5396
  Hasbrouck measure 5401
  information-based models 5397
  inventory models 5395
  Kyle model 5400
  Madhavan, Richardson and Roomans model 5402
Market order 5433
Market power hypothesis 5432
Market saturation 681
Market share 1965, 1978
Marketing 5377–5384, 5386, 5388, 5390, 9042
Markov analysis 3658
Markov approximation 4908, 7915
Markov chain 167, 2745, 2747, 2749, 5670, 5999, 6000, 6840, 8276, 8768, 9849
  finite alphabet 6950
  maximum likelihood 9849

Markov chain Monte Carlo 450, 1231, 6995
Gibbs sampler 1231
metropolis–hastings 1231
Markov condition 6995
Markov decision problems 8699
Markov–Einstein coherence 3656
Markov graphs 8215
Markov hypothesis 8937
Markov map, uniformly expanding 6950
Markov models 3457, 3459
additive markov gordon model 3459
geometric markov gordon model 3459
Markov odometers 3060
Markov operator on a group 787
Markov partition 4730, 6941, 8891
Markov–perfect equilibrium 2228
Markov perfect equilibrium (MPE) 2218
Markov process 644, 2817, 2941, 5244, 5456, 8767, 9839
Chapman–Kolmogorov equation 8767
stochastic processes 3579
Markov property 8709, 8713, 8716, 8718, 8721–8724, 8728
stochastic cascade process 3592
Markov random graph model 8322, 8324, 8326, 8329
Markov shift 953, 4984, 5456
isomorphism theory 4992
Markov switching 3424, 3443
Markov switching autoregressive model 449
Markov switching model 2745–2748, 2750, 2752, 2754, 5326, 5333, 5334, 5336, 5337, 5339–5341, 5344, 5418
Bounceback model 5335, 5336, 5340, 5343
Markovian 5223
Markovian game 2216
Markovian properties
direct evolution 3582
joint probability distribution 3580
stochastic time series analysis 3582
Markovian splitting subspace 8757
minimal 8759
Markovian strategy 2217
pure stationary 2223
Markovitz portfolio theory 4973
Marsupial mechanisms 2079
Marsupial pair 2082
Marsupial teams 2077, 2087
Martin–Löf tests 247
Martingale 3475, 3476, 3478, 3480, 3481, 3495, 5253
Martingale difference 3438
Martingale of a posteriori 7634
Marxist city 7961
Marxist economic theory 2285
Marxist system 2285
MAS *see* Multi-agent systems
Mask 592
Mass-action kinetics 329, 556
Mass advection 9906
Mass conservation 5535, 5536
Mass density 5026
Mass diffusion 6961
Mass fraction 611
Mass media 8308
Massive software system 1339
Massively multi-agent systems (MMAS) 149
Massively multi-player online games 1341
Massively parallel computing 259
Massively parallel processors 9723
Massless fermion 8577, 8585
Master equation 3361, 3365, 3368, 3581, 5671, 6139, 6140, 6164, 8386, 8589, 8769, 8933, 9291, 9292
Master stability function 8912
Matching 2839, 6988, 7607, 9655
F-optimal stable matching 9656
pareto-optimal matching 9656
Matching law 1886
Matching mechanism
manipulable mechanism 9656
revelation mechanism 9656
stable matching mechanism 9656
Matching methods 5561
ATET 5561
Matching technology 4110
MATCONT 6349
Mate-In-N 4142
Material flows in ecosystems 2703
Material points, fluid 5928
Material surfaces, fluid 5928
Material unaccounted for 4860
Material volumes, fluid 5928
Materials, smart nanoscale self-assembled 1304

Maternal effects 1926
Mathematica 119, 8848
Mathematical biology 1227
Mathematical chemistry 6852
Mathematical discussion 6302
Mathematical extension, definition 6289
Mathematical finance 6231
Mathematical framework 160
Mathematical model 3694, 3696, 4783, 6792
  macroscopic 3694
  microscopic 3694
Mathematical modeling 3114, 3683, 3691, 3694
  Eigen plane 3115
Mathematical physics, and mechanical systems 5488
Mathematical structure 6797
Mathematical system 6301
  definition 6289
Mathematical table 5476
Mather 704
Mathieu equation 2325, 3618, 6626, 6628–6630, 6633
Matrix 2203, 9188
  adjacency 2201
  aperiodic 6243
  condition number 9188
  distance 2201
  fully indecomposable 6244
  irreducible 6239
  monodromy 6593
  Moore–Penrose generalized inverse 9188
  nearly reducible 6244
  nonnegative 6242
  null space 9188
  partly decomposable 6244
  positive 6242
  primitive 6242
  quantum mechanics 7371
  reducible 6239
  strongly connected digraph 6239
  symmetric 6241
  symplectic 6589, 6590
Matrix action 4985
Matrix diagonalization 8592
Matrix display 8248
Matrix element of Liouville's quasioperator 1623
Matrix element of the contracted description 1620
Matrix extension problem 595, 598
Matrix-matrix multiplication 6392
Matrix-multiplication chemistry 332
Matrix pencil 8620
Matrix reordering 8604
Matrix subshift 8906
Matrix-tree theorem 6250
Matrix-vector algebra 6393
Matrix-vector multiplier 6392, 6393
Matrix zoom 1258
MatrixExplorer 1258
Matter energy density 3274
Matter Era 3274
Matthew effect 6004
Maturation 2854
Mature localized fault 5035
Maurer–Cartan equations 5506
Max-cut problem 6420
Max-min composition of two fuzzy relations 4068
Maximal clique 2097
Maximal entropy measure 6949
  for cellular automata 2984
Maximal inequality 2936
Maximal KAM tori 5067
Maximin piece 3328
Maximin procedure 9951, 9956, 9960
Maximin strategy 3328
Maximizing decision 4076
Maximum 4067
Maximum a posteriori estimator 9193
Maximum entropy formalism 2903
Maximum entropy measure 953
Maximum entropy measures, existence of 961
Maximum entropy principle 728, 4897
Maximum flow problem 6415, 6417
Maximum likelihood 6006, 6007, 9034
Maximum likelihood estimates 2749, 2752
Maximum likelihood estimation 2748, 7016, 9036
Maximum likelihood estimator 9037
Maximum nonlinear response 2427
Maximum principle 1028, 5445, 6271, 6272, 6275, 6276, 9105
  abnormal 5445
  extremals 5445
  normal 5445
  Pontryagin 6276
Maximum spectral type 5464

Maximum transient lengths 72
Maxwell–Bloch equations 6133, 6167
Maxwell–Boltzmann distribution function 8936
Maxwell counting 2830, 6014, 6016, 6017
Maxwell distribution 4909
Maxwell equations 6121, 9272
MAYA-II automaton 5649, 5650, 5652, 5653
Mayer, Gottfried 8943
Mayo, Elton 8267
Mazoyer 3539, 3542
McCarthy, J. 3539
MCell 8743
McGehee coordinates, dissipation 5980
McKendrick 6779
MCMC 8784
MD *see* Molecular dynamics
Mealy machine 4106
Mean
  of maxima method 4884
  period 635
  phase 631
Mean anomaly 6676
Mean cluster size 1403, 6541
Mean coordination number, stochastic network 4955
Mean curvature flow 6219, 6221
Mean dimension 781, 786
Mean ergodic theorems 3058
Mean field 634, 636, 638, 1133, 4906, 5025, 6807, 7823
  approaches 3364
  approximations 3362, 6772
  methods 8046
  power laws 3369
  theories 3362
  theory 5021, 5027
Mean free path 7917
Mean intervertex distance 4488
Mean moment rate profile 5030
Mean path length 725
Mean reversion 98
Mean square displacement 3888, 5219, 8712
Mean square error 8672
Mean square forecast error (MSFE) 3505
Mean time delay 9296
Meaning postulate, fuzzy logic 4000
Measurability 4050
Measurable 3054
Measurable dynamical systems
  isomorphism 5041
Measurable dynamics of cellular automata 3002
  mixing and ergodicity 3002
  spectral properties 3003
Measurable functions, definition 6307
Measurable map 5463
Measure 1014, 2239, 2981, 3033, 3034, 3054, 3879, 5454, 6306, 8710, 8713, 8716, 8718, 8721
  absolutely friendly 3034
  Bernoulli measure 2981
  cylinder set 2981
  ergodic theory 6944
  extremal 3033, 3035
  Federer 3034
  friendly 3034
  full support 2982
  μ-preserving 2982
  Φ-invariant 2982
  Rajchman 8565
  spectral 8556
  spectral, of the process 8565
  strongly extremal measure 3035
  support 2982
Measure algebra 5461
Measure algebra endomorphism 5452
Measure-preserving dynamical system 2982, 5039
  commutant 5044
  empirical distribution 5047
  Gaussian dynamical systems 5049
  K-system 5040
  Krieger's finite generator theorem 5047
  mixing 5045
  multifold mixing 5048
  Ornstein's isomorphism theorem 5047
  pointwise convergence 5047
  Poisson suspension 5049
  rank-one transformations 5044
  spectral types 5041
  weak mixing 5040
Measure-preserving map 3054
Measure-preserving system 5452
Measure-preserving transformation 4984
Measure rigidity 3096
  of cellular automata 2988
Measure space 3054

Measure-theoretic entropy 4986
Measure-theoretically isomorphic 2921
Measured-mile approach 1847
Measurement scale 1682, 1988
Measurement (theory) 1879
Measures of effectiveness 9457
Measures of maximum entropy 958
Mechanical and analogue computers 9705
Mechanical artificial chemistry 336
Mechanical automaton 5477
Mechanical circuit 7032
Mechanical clamp 7036
Mechanical computing 5467
Mechanical computing device 5473
Mechanical connection 5493, 5495, 5503
  extremal characterization 5504
Mechanical failure
  bulk modulus 5002
  shear 5002
  shear modulus 5002
  unjamming 5002
Mechanical folding 7046
Mechanical forces 7027
Mechanical inference machine 5477
Mechanical proteins 7029
Mechanical stability 7033
Mechanical systems 5482
  and mathematical physics 5488
  discrete 5498
  nonholonomic 5498
Mechanical unfoldases 7042
Mechanics
  gauge theory 5495
  geometric 5483
  stochastic 3871
Mechanism 4788, 4790, 5467, 5510, 5514, 9070–9073
  direct 1672
  indirect 1672
  integer game 4797
  straightforward 1672
  strategy-proof 1673
Mechanism design 4096, 6978
Mechanism of aquatic toxicity (MOA) 2127, 2148
Mechanism of pre-seismic EM phenomena 2629
Mechanism of toxic action 2129
Mechanistic QSAR 7077
Mechanochemical enzymes 7042
Media 3128, 3133
Media campaign 8315
Median 227, 229
Median rule 227
Median voter theorem 9937
Mediated communication 1592
Medical innovation study 8311
Medical malpractice 7052
Mediotype 3209
Mediterranean Sea 7842
Medium resolution imaging spectrometer 1701
Mega-wide moving jam 9392
Melnikov conditions 5082
Melnikov theory 6637
Melt 6955, 9887
Melt segregation 2685
Melt viscosity
  entangled polymer dynamics 6822
  polymer physics 6813
Melting 2026
Melting temperature 2170
Membership function 4061, 4062, 4877, 8681
  fuzzy logic 3991
  secondary 4089
  type-2 fuzzy logic 4010
Membership grade 4062
  fuzzy logic 4010
Membership values
  discrete 4090
  special fuzzy function 4085
Membrane 691, 6101
Membrane computing 928
  in artificial chemistry 341
Membrane fusion 9751
Membrane permeability 859
Membrane structure 5524
Meme 126, 3205, 3206
  fitness 3212
  parasitic 3216
Meme replication 3211
Memeplexes 3210
Memes vs. genes 3208
Memetics 3205
Memory 1120, 6099, 6106, 6111, 6112, 7024
  episodic 1120
  long-term 1120

semantic 1120
short-term 1120
strong fluctuations 7024
working memory 1120
Memory effect, statistical 1624
Memory function 1616
Memory in other discrete time contexts 843
Memory organization packets 1804
Memotype 3209
Menarche 1905
Mendeleyev conjecture 7424
Mental disorders 627, 7992
Mental health diagnosis, modeling 4573
Mental model alignment 7063
Mental models 7057, 7841, 7850, 9048, 9060, 9062
Mere exposure effect 2244
Mereology 4419, 7789
Merge bottleneck 9374
Merging 7658, 7660, 9398
Mertens and Zamir system 7641
Merton jump–diffusion model, quadratic return variation 8787
Mesenchymal 532
Mesh partitioning 7906
Meso 1543
Meso-scale assembly 5479
MesoRD 8741
Mesoscale 1399
Mesoscopic description of a graph 1160
Mesoscopic level 5733
Mesoscopic modeling 9447
Mesoscopic physics 7663
Message authentication codes (MAC) 5612
Message-passing interface (MPI) 155
Meta-dynamical ODE frame work 332
Meta-model 216, 5320
Meta-modules 5627
Meta rules 8187
Meta-stability 2421
Metabolic control analysis 559
Metabolic network 504
Metabolic rates 3281
fractal 3784
Metabolic reaction networks 5656
Metabolic scaling 2708
Metabolic stability 2186
Metabolic systems biology 9104
Metabolism 5537, 5539, 9104
Metabolites 9104
Metabolomic 5537
Metacognition 2243
Metacommunity 2706
Metadata 355, 8084
Metaethnic frontier 4615
Metapopulation 2723
dispersing organisms 2737
Metastability 1438, 1538, 1549, 1551, 2023, 4497, 4505
symmetry breaking 1549
Metastable régime 1550
Metastable state qubit 7224
Metastable states 2022, 2035
Metastable synchronized flow 9361
Meter 2310
Methane 1078, 1322
selective dehydrogenation 1322
Method of moments 3457, 5999, 6006, 6007
Method of simulated moments 217, 3457, 3464
Methodological individualism 207
Methodology 9003, 9004, 9012
Metric
fractal 3870
Jacobi 6604
Riemannian 6605
Metric factor 6998
Metric multidimensional scaling (MDS) 8283
Metropolis 6784
Metropolis–Hastings algorithm 6003, 8325
Mexican hat function 3346
Mexico 2397
Meyer's wavelets 10033
Mezey's molecular hologram 2209
MI *see* Mutual information
Micelle 859
Michael Conrad 924
Micro 1543
Micro-electro-mechanical system (MEMS) 5478
Micro-level 8185
Micro-macro interaction 8406
Micro-simulator 163
Microanalytical simulation 8406, 8410
Microarray 721, 3968
Microbial camera 425
Microbial colony language (MCL) 260, 267

Microbial fuel cell (MFC) 3690
Microbiology, amorphous computing 258
Microcanonical ensemble 1270
Microcosm 7850
Microearthquakes 2543
  velocity pulses 2546, 2547
Microenvironment, breast cancer 919
Microfabrication, amorphous computer 258
Microfeature 3334
Microfoundations 5431
Microlevel simulation 8375
Microlite 9764, 9775
Microlocal analysis 1494
Micromotion, quantum computing 7222
Micron 5573
Microprocessors 7401
MicroRNAs (miRNAs) 1203
Microscale 1399
Microscopic modeling 9447
Microscopic simulation 92, 6478
Microscopic state, quantum information processing 7364
Microscopic structures of GPs 9397
Microscopic variable 2678
Microsoft robotics studio 2079
Microstate 4380
  polymer physics 6813
Microstructure 1396, 3390, 9565, 9570, 9573, 9575
  periodic 1396
  random 1396
  US treasury market 9573
Microtrap, quantum computing 7223
Microtubule 1744, 1746–1748, 1750–1755, 1759, 1761, 1762, 1766, 1768–1774
Microworld 4583, 9044, 9050
Mid-Pleistocene transition 4
Middleware 1349, 5762
Midrange 227
Migdall 7263
Mignosi F 780
Migration 6450
  climate change 1102
Milburn 7258
Mild solution 1488, 6234
Milestones 1902
Milgram experiment 8263
Milgram, Stanley 8269, 8277
Military 7061
Military intervention 7843
  Turkish 7845
Miller indices 1310
Miller's solution 9952
Millionaire calculator 5476
MIMD *see* Multiple instruction multiple data
MIMO *see* Multi-input multi-output
Min max payoff 7618, 7628
Mind 1550
  endomorphic modeling and simulation of 363
Mind–body problem 4893
Mind-body problem 121
Mind viruses 3206
Miniaturization 2091
  quantum information processing 7362
Minicluster of genes 1183
Minimal action principle 9413
Minimal covering set 9953, 9958
Minimal rational extension 8763
Minimal realizations 6375
Minimal self joining transformation 4989
Minimal system 9234
  observability 6373
Minimal time problem 6278, 6283
Minimax procedure 9952, 9956
Minimization algorithm 819
Minimizer
  n-body problem 5967
  planar equivariant 5968
  three-dimensional equivariant 5968
Minimum 4067
Minimum-cost-flow-problem 6411
Minimum cost spanning tree games 1664
Minimum cut problem 6415
Minimum-energy-path 1312
Minimum level, observability 6378
Minimum output entropy additivity conjecture 7393
Minimum-phase 9089
Minimum spanning tree algorithm 5054
Minimum spanning trees 6501
Minimum support value and membership functions 4154
Mining data streams 3252
Minkowski norm 7429
Minor 8956

Minority-carrier transport 9714
Minority game (MG) 4979, 5591, 6808, 7627
no strategy table and no information 5601
without information 5590, 5601
Minsky, M. 3539
MIP 7193
Mirror neuron 1898, 1904, 8205
Mirror system (HMS) 8205
MISO *see* Multi-input single output
Missed alarm 2417
Missing attribute values 7798, 7801
conditions 7798
do not care 7798
lost values 7798
Missing values 1841
Mitchell, Clyde 8268
Mitigation 1076, 3110, 3131, 9033
disaster mitigation 3113
Mitochondria, membrane potential 477
Mitochondrion 5547
Mittag–Leffler function 5222
Mixed behavior 2835
Mixed phase space 992
Mixed state 2869
Mixed strategies 4855, 8653
Mixed systems 1008
Mixed transition 2831
Mixing 385, 2531, 2533, 2535, 2925, 3049, 3647, 4589
assortative 2531, 2536
disassortative 2531, 2536
higer order 2929
mild-mixing 2928
strong-mixing 2926
weak-mixing 2925
Mixing recurrence, ergodic theory 3061
Mixing system 402, 990, 993, 994, 9248
weak mixing 996
Mixture modeling 457
Mixture-of-distributions hypothesis 8784
Mixture of experts models 5323
Miyagi-Oki earthquake, tsunami 4846
MLEM2 algorithm 7802
mm-precisiand, fuzzy logic 3997
MMAS *see* Massively multi-agent systems
Mobile agent systems
properties 5613
survey of 5613
Mobile agents 5604
agile computing 5615
application areas 5613
attacks on 5610
classification of 5605
components of 5609
definition 5607, 5608
denial of service 5611
distributed systems 5615
dynamic configuration 5609
endogenous threats 5610
exogenous threats 5610
global state of programming 5608
localities 5608
malicious 5610
orthogonal capabilities of 5605
requirements 5608
restricted connectivity 5608
security 5609
unmanned autonomous systems 5615
Mobile code 5606
Mobile computing 5605
theoretical foundations 5608
Mobile docking station 2085
Mobile search 5282
Mobile state 5606
Mobile systems, classification of 5607
Mobility 7402, 9412
Modal analysis 2324
Modal logic
alethic 8340
Graded 8341
hybrid logics 8342
propositional 8341
propositional multi-modal 8341
Modality, networks 1182
Mode 7663, 8040, 8280
mode competition 6124
of toxic action 2166
quantum computing 7231
1-Mode 8281
2-Mode 8280
Mode-coupling 3371
theory of the glass transition 4225

Mode-locking 6136
Mode skeleton 8936
Mode-space NEGF 7484
Mode-switching 5027, 5033, 5035
Model 289, 2285, 2441, 4392, 5743, 6104, 6412, 6772, 7088–7090
  Aw–Rascle 9733
  black box 7088
  black-box 7091
  Blue–Adler 3159
  bottom up 8186
  Colombo 9734
  continuous 7090
  continuous time 7107
  De Angelis 9732
  Delitala and Tosin 9739
  discrete 7090
  discrete velocity 9739
  Enskog-like 9738
  fictitious density 9732
  first order 9729
  fluid-dynamic 3156
  fracture networks 6506
  Fukui–Ishibashi 3159
  gas kinetic 3156
  Gipps–Marksjös 3159
  GPS measurements 2577
  Klar and coworkers 9738
  lattice Boltzmann (LB) model 870
  lattice-gas 3161
  learning from 8193
  Lighthill–Whitham–Richards 9730
  LWR 9730
  manna 651
  mathematical 9727
  measurement based 7114
  of biological networks 724
  optimal-velocity 3161
  Paveri Fontana 9737
  Payne–Whitham 9732
  precursor events 2577
  Prigogine 9736
  probabilistic 5743
  quantum circuit 7088
  quantum computation 7091
  rate 6100, 6106, 6109
  recursive scale-free trees 3932
  ring 6106
  second order 9730
  seismicity based 2441
  social-force 3157
  Song–Havlin–Makse 3930
  stress release model 2576
  subcritical 6106
  supercritical 6106
  Swiss cheese 3849
  top down 8186
  unobserved components model 5330
  validation 9067
Model A critical dynamics 3371
Model acceptability criteria 7079
Model analysis 157
Model-assisted estimators 8214
Model B 3371
Model calibration 9035
Model fusion 5319
Model hierarchy 5915
Model of computation 7119
Model of replication 3912, 3913
Model predictive control 3944
Model quality 9000, 9001, 9012
Model reduction vs. loop contribution 8952
Model selection 446, 448, 452, 5319
Model simplification approach 8949
Model specification 4056, 8784
Model validation 8988
  critical importance 7076
Model validity 8996
Modeling 2729, 2857, 3136, 8733, 8739
  agent-based 3389
  cellular automata 1304
  deterministic 8733
  forward and inverse approaches 1220
  fuzzy logic 3997
  non-spatial stochastic 8733
  spatial stochastic 8739
Modeling approach
  continuous 3155
  deterministic 3156
  discrete 3155
  force-based 3156
  high fidelity 3156
  low fidelity 3156
  macroscopic 3155

microscopic 3155
rule-based 3156
stochastic 3156
Modeling language, fuzzy logic 4005
Modeling physical systems, applications 3240
Modeling process 7838, 9000, 9003, 9010
antagonistic 7841
interactive 7841
Modeling space 337
Modeling volcanic eruptions 1223
Modeling workflow, QSAR 7071
Models 849, 1231, 2576, 6102, 6546–6548, 9069, 9412
evolution, fuzzy system 4080
Moderate resolution imaging spectroradiometer 1701
Modern linear systems theory 8954
Modified decomposition method 83
Modified Korteweg-de Vries (mKdV) equation 4962, 9991
Modifiers, fuzzy logic 3988
Modons 8494
Modular motors, RNA polymerase 3601
Modular network, information curve 506
Modular proteins 7032
Modular robot systems
ATRON 5623
CKBot 5622
M-TRAN 5623
Miche 5623
Programmable parts 5624
SUPERBOT 5623
Modular robotic docking station 2077
Modular self-reconfigurable 1226
Modular structures 2732
Modularity 730, 1144, 1145, 2716, 3934, 6049, 6054
clique 2730
groups 2730
matrix 6055
network 502
Modularity optimization 1151
extremal 1152
greedy technique 1151
reliability 1153
simulated annealing 1152
spectral 1152
Modulation matrix, multiwavelet 5848
Module 730, 1973, 4835, 6178, 9096
information theory 505
regulatory 730
Mokken, Rob 8269
Molar parallelism 3357
Molchan diagram 2446
Molecular automata 924
Molecular biology 1227
Molecular collision 327
Molecular complexity 498
Molecular computing 924, 2091, 3356
Molecular descriptor 2200, 2209
three-dimensional 2208
Molecular dissimilarity 2209
Molecular dynamics (MD) 2358, 6014, 6023, 6407, 6689, 7034, 7036, 8696
simulations 151, 2024
Molecular electronics 924, 5862
Molecular evolution, networks 1304
Molecular finite state machines 9705
Molecular frameworks 2838
Molecular full adder 5646, 5647
Molecular graph 4827
Molecular half-adder 5646
Molecular indicator
geometrical 2210
quantum-chemical 2210
topochemical 2210
topostructural 2210
Molecular logic gate 5654
Molecular machine 9750
Laing's 334
with fixed tape size 334
Molecular media 9705
Molecular motor proteins 1774
Molecular motors 1738, 1739, 1741, 1747, 1757–1759, 1761, 1766–1771, 1773, 1774, 3600, 8738
DNA polymerases 3600
rotary motor 3601
Molecular networks 5656
Molecular similarity 2209
Molecular structure, semi-realistic 336
Molecular system, lattice 335
Molecular topological descriptors, QSAR modeling 7075

Molecular transducing automaton 5647
Molecular weight 6547, 6549, 6550, 6553
  polymer physics 6811
Molecule
  as building block for AC 328
  cascades 5869, 5876, 5877
Mølmer–Sørensen gate
  quantum computing 7235, 7242
MolNet 2152
Moment 6145, 7000
  analysis 7001
  multiwavelet 5846
  numerics 7000
  seismic moment 2474
Moment condition models 5559
  generalized method of moments 5560
Moment history 2478
Moment magnitude 2478, 2484, 2485, 2594, 2595, 9598, 9633
  earthquake 2664
  fundamental earth's modes 2485
Moment map 6187
Moment measures 2574
Moment rate 2485
Moment-rate 2488
Moment rate shape 5029
Moment ratio 7004
Moment tensor 2488, 2582, 9903
Moments characteristic
  finite-rank 6309
  hyperfinite-rank 6309
Momentum 98
Momentum cocycles 5485
Momentum mapping 2263, 4531
Momentum maps 5484, 6597
  in symplectic reduction 5497
  nonequivariant 5485
Momentum one-cocycle 5485
Momentum term 1820
Monad hyperreals 6293
Monetary compensations 1671
Money 2804, 2805, 2823
  distribution 2803
  supply 5416
  temperature 2805
Monge–Ampère equation 6221
Monitoring
  almost perfect 7621
  almost public 7622, 7627
  canonical 7622
  perfect 7621
  public 7622
  real-time 8857
  semi-standard 7623
Monitoring structure 7621
Monodisperse fractures 6510
Monodisperse objects, continuum percolation 6509
Monodromy 2263, 4516, 7136
  classical 7146
  fractional 7147
  Hamiltonian 7140
  principal 2263
  quantum 7146
Monodromy matrix 2324, 2325, 6431, 6433
Monofractal 3717, 3893, 3898
Monogenic disease 2187
Monomer 6811
Monomer concentration 6817
Monomeric drag constant, single chain dynamics 6820
Monotone 8254
Monotone hazard condition 1673
Monotonic model 5026
Monotonicity 4794
Monotonous likelihood quotient property 6982
Monte Carlo 692, 6992, 7572, 7575, 7577
Monte-Carlo algorithm 421, 7123
Monte-Carlo computation 4909
Monte-Carlo dynamics 4218
Monte-Carlo estimation 8276
Monte-Carlo Markov chain maximum likelihood 8329
Monte-Carlo method 7413, 7917
Monte-Carlo simulation 2024, 3363, 3369, 3457, 3460, 3464, 4051, 6407, 8713, 8716
Monte-Carlo tests 6266
Montgomery example 5449
Mood 2242
Moore 3547, 5473
Moore EF 780, 783
Moore machine 4101
Moore neighborhood 872, 3234

Moore's law 276, 291, 924, 3358, 7399, 7401
first law 5861
second law 5866
Moral 3138
Moral dynamics 5677
Moral hazard 6981
repeated 6982
Morality 5678
Moran model 8771
Moran's *I* 8308
More informative action 7624
Moreno, Jacob 8267
Morgenstern, Oskar 1510, 8381
Mori–Zwanzig projections 4595
Morphism 703, 712
Morpho-dynamic fields 1878
Morphogen 540, 543
Morphogenesis 118, 423, 532, 539, 928, 3358, 8940
Morphogenetic 532
Morphological 2847
Morphology 1043, 1065
block co-polymer melt 6829
Morphotype transition 3788
Morrey 6664
Morse 704
Morse function 9227
Morse shift 8892
Morse theory 6622, 6623, 7138
Mortality
thermal stress 1098
weather extremes 1097
Mosaic picture 4225
Moser 704, 5079
MOSFET 7401, 8576, 8577
Motif frequencies, natural 5700
Motif information, network 507, 508
Motif intensity score 5698
Motif profile 10062
Motifs 726, 1247
and evolutionary paths 5698
biological 5699
FFL 5698
in graphs 5692, 5693, 5695, 5697–5700, 6369
of networks 5695
undirected 507
Motion 6427
coordination 5783, 5790, 7725
periodic 6587
planning 5470
Motivation 8001
Motor development 1902
Motor equivalence 5720
Motor programs 5719
Motor proteins 1741, 1745, 1747, 1750, 1752–1754, 1757–1760, 1764, 1766, 1769–1771, 1774
eukaryotic 1747
Moulin–Shenker rule 1683
Mount Pinatubo (Philippines) 9866
Mount St. Helens (USA) 9866
Mountain pass solution, 3-body problem 5976, 5977
Mountains 3700
Movement coordination 4894, 5718, 8004, 8947
Movement disorders 630, 641
Movement pattern 8938
Movement strategy 2738
Movements disorders 627
Moving blanks 9382
Moving jam 9357
emerging spontaneously 9363
Moving quantum dot 1022
Moving SP 9365
Moving synchronized flow pattern 9359
Moving traffic jam 9308
Moyal product 6758
MPE *see* Markov perfect equilibrium (MPE)
MPI 5755
MPI *see* Message-passing interface
MQA 7189
MRA *see* Multiresolution approximation
MRAP *see* Most rapid approach (MRAP)
MRJ 5014
MS fractional fluctuation 7919
MSFM problem 4149
MSP 1756, 1757, 1768, 1773, 1774
MSR-formalism 3650
Multi-agent simulation models 150, 185, 8406
Multi-agent systems (MAS) 161, 7534
architeture 188
Multi-binary GrC model, fourth GrC model 4346
Multi-bit communication cellular automata 3555

Multi-dimensional 8268
Multi-dimensional feature space 2354, 2355
Multi-dimensional scaling 2356, 8274, 8352
Multi-dimensional space 2361
Multi-frequency coordination 5734
Multi-input multi-output (MIMO), type-2 fuzzy logic 4016
Multi-input single output (MISO) 4082
Multi-modal perception 1898
Multi-objective optimization 9042
Multi-phase polymeric fluid 6827
Multi-photon ionization 7172
Multi-planner systems 5203
Multi-resolution 1164, 1166
Multi-resolution analysis (MRA) 2381, 6862, 10031, 10035
Multi-resolution approximation 6888
  multiwavelet 5845
Multi-robot path planning 5783
Multi-scale 2846, 2850
Multi-scale approach spin-density-function theory 7314
Multi-scaled phenomena 915
Multi-target observation 5795
Multi-wave mixing 8596
Multibody system 2332
Multicellular organism 929
  development 1192
  differentiation 1192
  DNA fractal features 3797
  stem cells 1192
Multicellular tumor spheroid 917
Multicore processor and amorphous computing 269
Multiculturality 8394
Multidrug resistance 2171
  reversal 2171
Multiferroic 2021
Multifractal analysis 3893, 3918, 9844
  dynamical systems 3024
Multifractal behavior 3576
Multifractal decomposition 3024
Multifractal formalism 3895, 3897
Multifractal model 3655
Multifractal records 3717
Multifractal spectra 3024, 3901
Multifractals 999, 3717, 3718, 3720, 3738, 3759, 3767, 3775, 3843, 3893, 3898, 8524, 8536, 8537
  binomial 3775
  distribution 3747
  D_q 3722
  $f(\alpha)$ 3722
  log binomial 3747
  multifractal measure 3722
  multifractal scaling 3745
Multifunctional composite 1284
Multifunctionality 1541
Multigraph 2201, 2203
Multilayer 2850
Multilayer feedforward artificial neural network 2143
Multilayer neural networks 1817, 1818
Multilayer perceptron (MLP) 2143
Multilevel 2846, 2850
Multilevel algorithms 1257
Multilevel simulation 8411
Multilevel theory 1559
Multinomial logistic regression 6004
Multiparticle entanglement, quantum computing 7242
Multiplayer preceptron 4872
Multiple attribute (or multiple criteria) decision support 7753
Multiple criteria sorting 7753, 7754
Multiple dealer markets 5394
Multiple equilibria 3
Multiple instruction multiple data (MIMD) 153
Multiple-minimum-support fuzzy mining (MSFM) 4147
Multiple network 8291, 8293, 8294, 8296, 8297, 8300, 8301, 8305, 8330
Multiple objective optimization 7753
Multiple-prover interactive proof 7192
Multiple realizability 285
Multiple recurrence 3044
  ergodic theory 3061, 3063
  polynomial 3044
Multiple robotic system 2053
Multiple rupture 2488
Multiple scale transformation of variables 6212
Multiple scales 5912
Multiple systems 7373
Multiple tunneling junction 9716

Multiple vents 9849
Multiplet 3876, 6734
Multiplexing 5872
Multiplicative ergodic theorem 2943, 4735
Multiplicative noise 6229, 7014, 8778
Multiplicity
  algebraic 6651
  of subgame perfect equilibria in the multiplayer game 4105
Multiplicity function 8555
Multiplicity sampling 8214
Multiplier 1062, 6593, 6597
Multiplier-accelerator model 8994
Multiplier method 1493, 1494
Multipotent cells 1192
Multirelational networks 8263
Multiscale 1719, 2351
  dyadic 1725
  geometry 1719
  pyramid 1725
  sub-band 1725
  transforms 1719
Multiscale analysis 1939, 6663, 6666, 6671
Multiscale complex systems 5914
Multiscale modelling of nanostructures 5923
Multiscale problems 6667
Multiscale simulation 5676
Multiscaling 7002
Multiscaling function 5845
Multiset 5524
Multistability 639, 641, 1548, 4894
  networks 1193
Multistable 1550
Multistage control process under fuzziness 4077
Multistage game with observable actions 4099
Multisymplectic reduction 5498
Multivariate 6882
Multivariate autoregressive modeling 1434
Multivariate GARCH, CCC model 4127
Multivariate network 8322
Multivariate regressions 2795
Multivariate splines
  bivariate spline 5801
  box splines 5800
  spherical 5801, 5802
  trivariate 5802
Multivariate stochastic volatility 4129
Multivariate wavelets
  biorthogonal boxsplines 6867
  multivariate prewavelets 6892
  orthonormal box spline wavelets 6908
  spherical tight wavelet frames 6901
  splines 6888
Multiview 1262
Multiwavelet 5841, 10006
  balanced 5850
  boundary 5850
  lifting 5853
  moment 5846
  polyphase factorization 5852
  pre and postprocessing 5849
  prefilter 5849
  symbol 5844
  TST 5853
Mumford–Shah segmentation model 10033, 10039
Munakata 5473
Murata 8840
Music 8200, 8202
  fractal noise 3794
Mutagenicity 2145, 2169
Mutation 121, 3971, 4135, 4780, 4873
Mutation operation 4150, 4874
Mutational analysis 7036
Mutual edges 5694
Mutual entropy, biological information 498
Mutual fund flows 3472
Mutual information (MI) 1430
  networks 1201
Mutual Kolmogorov complexity 494
Mutualistic networks 3677
Mutuality 4663
Mutually orthogonal subspace 1623
Mwp 9613
M_{wp} method, tsunami warning 2666
Myerson value 6045
Myhill, J. 780, 783, 3539
Mylonite zones 2651
Myopia 6003
Myopic profit maximization 1968

N

n-ary relation, definition 6301

n-Body problem 5960, 6588, 6681
absence of collision 5981
basic definitions 5962
choreographies 5959
conservation laws 5979
$D_{4q} \times C_2$ group 5966
$D_6 \times C_3$ group 5968
$D_q \times C_2$ group 5967
eight shaped three-body solution 5964
G-equivariant minimizers 5982
generalized solution 5978
generalized Sundman–Sperling estimates 5980
gravitational 383
locally minimizing solution 5978
Neumann boundary conditions 5982
rotating circle property 5964
Standard variation 5983
n-Body simulation 5472
n-Clan 8288
n-Clique 8288
n-Club 8288
n-Gram entropy, DNA sequence 495
N-Particle W-state, quantum computing 7243
N-Qubit Deutsch–Jozsa circuit 7382
N-Soliton solution 4969, 4970
∇-Extended density function 7446, 7447, 7449
∇-Extended wave function 7444
Nadel 8273
Nadel, Siegfried 8268
Nagel–Schreckenberg (NaSch) CA model 6776, 9289, 9335
Nakajima–Zwanzig 8589
Nano- 924
Nano-device system 2112
Nanocell 5882
Nanocluster 5889
Nanocomposites 1284
Nanocomputer 5862
chemical 5863
electronic 5863
mechanical 5863
programming methods 5884
quantum 5864
Nanodevices 7677
Nanodiamond 9154, 9155
Nanodot (minute dot) 9719
Nanoelectromechanical systems (NEMS) 5917
Nanoelectronic component
metallic nanoparticle 2111
metallic nanowire 2111
Nanofabrication 2105, 5866
bottom-up 5866, 5874
top-down 5866
NanoFabrics 5881
nanoHUB 5762
Nanoidentification problems 5918
Nanomachines 5923
Nanoparticle 5889
Nanopeapods 9149
Nanopiezotronics 5921
Nanoscale device 7307
Nanostructures 1305, 5747
Nanotechnology 2090, 2105, 8577
Nanotube 9149
Nanotube junctions 9149, 9150
Nanowires 5751, 5777
Nanson 9944
Nanson's procedure 9948
Napier, John 5476
Napier's bones 5476, 9705
Naples 2398
Narcissistic modeling, QSAR modeling 7076
Narrow moving traffic jam 9326, 9358
Narrow transcription 2317
Narrowing by motion 8595
NASA's Johnson Space Center 7845
NaSch CA model 9290
NASDAQ market 5395
Nash equilibrium (NE) 437, 1530, 1587, 1662, 1669, 2217, 3177, 3178, 3180, 3185, 3187, 4855, 4857, 6643, 7618, 7655, 8381, 8651
mixed 1588
strong 1669, 1681
with complexity 4100, 4101
Nash networks 6043
Nash-reversion equilibrium 7656
Nash stability 6027
Nash stationarity 3184
Nash, John F. 8381
Nasu's theorem for cellular automata 2986
National Airspace System (NAS) 237
National Bureau of Economic Research 2753
National Cancer Institute 2148, 2187
Natural climate changes 1097

Natural computing 283, 292, 4335, 4336, 9133
Natural experiments, exogenously variation 5560
Natural extension 2975
Natural language processing (NLP) 8094
Natural languages, fuzzy logic 3988
Natural mechanisms 4798
Natural selection 1923, 1924, 2908, 3209, 3278, 4873
Natural time 2622, 2623, 2625
Natural time analysis 2623
Naturalization 4896
Nature-inspired 923
Navier–Stokes
  derivation of some fundamental a priori estimates 5944
  long-time behavior 5953
  uniqueness and continuous dependence on the initial data 5946
Navier–Stokes approximation 4383
Navier–Stokes equations 549, 881, 882, 1493, 3642, 3644, 3843, 4383, 5576, 5925, 5926, 5929, 5930, 5943, 6217, 6229, 8727, 9769, 9971, 10040
  compressible Newtonian fluid 3615
  derivation 5928
  global attractor 5953
  incompressible fluids 3615
  less regular solutions 5949
  partial regularity results 5949
  regularity 5946
  times of irregularity 5946
  turbulence 5954
  weak solutions 5931
  weak solutions and related properties 5949
Navier–Stokes hydrodynamics 4388
Navigation, map-based 6070
NBER 5334, 5341, 5343
NBTree decision tree 2165
NC 6388, 6402, 6403
NE *see* Nash equilibrium (NE)
Near-miss 5353
Near real-time material accountancy 4861
1-Nearest neighbor 1792
Nearest neighbor method 67, 1791, 1800, 3847
  average connectivity 2531, 2536
Nearly-integrable Hamiltonian 4516, 5067
Nears-hit 5353
NEC 4107, 4111, 4112
Necessity 4425
Necessity measure 6927, 6929
Necessity relations 6930
Negation 4878
Negative feedback loop 678, *see* Balancing feedback loop
Negative feedback systems 560
Negative selection 4781
Negative-sum game 8381
Negative temperature states 3648
Negentropy 3275
  principle of information 4822
NEGF 5779
Negotiation models 4107
NEIC 2478
Neighborhood 164, 872, 1056, 4498, 7680, 8270
  clustering 577, 579
  configuration 65
  function 1823
  Margolus neighborhood 7681, 7690
  radius 1823
  system (NS) 4313, 4344, 4355, 4357
Neimark–Sacker 6336
  bifurcation 5309, 6629
Nekhoroshev
  estimates 6326
  stability 5993
  theorem 6714
  theory 405, 4519, 6681
NEMO 3-D 5747, 5752
Neo-liberalism 8381, 8382
Neo-Piagetian theories 1896
Neo-Piagetians 1898
Neoclassical economics 203, 2285
  agent-based modeling 177
  assumptions 179
  game theory 181
  simplicity 179
  tractability 179
Neologism proof 8132
Neoteny 1918
Nernst–Planck equation 5577
Nervous system 6070
Nervousness 6488
Nested summation symbol (NSS) 7444, 7455
Nestedness 2717
Nesting vector 9696

Nestling 7525
Net benefit 1671
Net flow score 7768
.NET infrastructure, Microsoft 5607
Net present value (NPV) 7846, 7847
NetDraw 8848
NetLogo, simulation tool 194
Network 209, 235, 237, 239, 477, 528, 541, 544, 645, 653, 719, 860, 1049, 1182, 1244, 1821, 2875, 3661, 3856, 3924, 4182, 4486, 4487, 4492, 4911, 5999, 6049, 6099, 6100, 6104, 6411, 6415, 6555, 6644, 7953, 8233, 8246, 8266, 8322, 9113, 9742
  Airport and Airspace Network (AAN) 237
  Airport Network (AN) 237
  analysis 2729
  arc 6024, 9742
  architectural 477
  as complex systems 5692
  autopoietic 478
  autoregulatory 529
  behavior 238
  biochemical 544, 545
  biological behavior 1183
  Booleans 1197
  ccdf 238
  cellular 5699
  community detection 4612
  complex 234, 1191
  connections of the C. elegans brain 506
  critical 1200, 1201
  decomplexification 505
  degree 235, 238
  directed 6024
  dynamical 477
  dynamics 236
  dynamics behavior 1182
  edge 235
  emergent 543, 544
  evolution of information 502
  fault 4182
  flows 6407
  fractal 3925
  gene 527, 542
  gene regulatory 1249
  genetic interaction network 9114
  graph 2729
  growth models 1249
  heterogeneous 6024
  hierarchic 544
  hierarchical 543, 725
  homogeneous 6024
  huge 8246
  in systems biology 1185
  label 6024
  labeled arcs 6024
  labeled links 6024
  large 236, 1201, 8246
  link 6024
  linking 6024
  local structure 507
  log fitness 505
  logarithmic 238
  metabolic 504
  micro-macro linkages 4610, 4613, 4614
  middle 8246
  motif information 507, 508
  multilevel models 4627
  mutual informations 1201
  neural 503, 6106
  nodes 235, 6024
  null model 1187
  protein–protein interaction 1186
  random 239, 503
  random graphs 4911
  randomly connected 6105
  real 236, 1202
  reciprocity, reciprocators 4617
  resilience 239
  scale-free 234, 3924, 4487, 4608, 4612, 4622, 4911, 9114
  seismic 2479
  self-similar 3925
  signaling network models 4611
  simple graph 6048
  small 8246
  small world 3925, 4911, 9114
  social 8266
  social networks 4911
  stability 236
  subgraph composition 507
  synchronization phenomena 1251
  targeted attack 239
  third-party intervention 4616

topological 477
topology 2729
transcription factor 528, 543
triad significance profile 5698
vertex 9742
weighted 4492
without descendants 6039
Network analysis
algebraic methods 8266
algebraic models 8273
visualization techniques 8266
Network analyzers 273, 3337
Network architecture 1182, 1186
Network autocorrelation 8307, 8308
Network automata 8846
Network barrier 1966
Network-based interventions 8317
Network biology 5700
Network closure 6005
Network complexity 502
levels of 5693
Network conditions
changing 5614
Network configuration 8323
Network design 2093
Network dynamics 8269
genes 1188
neuronal 6104
Network effects 1965
Network evolution 2742
Network exposure 8309, 8311–8315
Network externalities 1965
Network formation 4097
Network game 7625
Network glass 6014, 6020
Network (GRN) gene regulatory 1182
Network homomorphism 8300, 8301
Network inference 1186
Network information bottleneck (NIB) 505
Network matrix 8259
Network modularity 506
Network motifs 570, 572, 573, 578–580, 583, 584, 2718, 4835, 5660, 5694, 5695, 6053
dynamic behavior of 5698
Network payoff functions 6032
Network properties 730
Network reconstruction 5535, 5536, 5538, 5540
Network robustness 6497
Network sampling 8272
Network state
gene expression 1183
genes 1188
Network structure 8315
analyzing 1186
Network themes 572, 573, 581, 583, 584
Network theory and topology studies, applications 3240
Network topology 1182
Network utilization 2068
Network walk sampling 8221
Neumann boundary conditions 6713
n-body problem 5982
Neumann condition 6224
Neumann series 5072, 5085, 6743, 6744
Neumann, John von 257, 849
Neural complexity 1424
matching complexity 1433
Neural correlates of consciousness 1424
explanatory correlates 1440
Neural ensembles 1558
cellular 1558
neuroanatomical 1558
synaptic 1558
Neural network approach 2246
Neural network estimator 3514
Neural networks (NN) 276, 280, 282, 283, 288, 721, 1784, 1813, 1993, 2123, 2210, 3339, 3344, 3355, 3358, 4137, 4871, 4900, 5882, 6067, 6087, 7992, 7993
C. elegans brain 503
cellular 5875
for traffic prediction 9515
recurrent 280
Neural synchronization 8005
Neuraminidase 2146
Neuro-fuzzy 3243
Neuro-fuzzy control 1225
Neuro-fuzzy systems (NFS) 4411, 6086
architecture 6091
criteria isolines method 6095
effectiveness 6095
for pattern classification 6093
learning 6094
logical-type 6089

Mamdani-type 6090
simplified 6091
suboptimal 6095
Takagi–Sugeno 6092
with Lukasiewicz implication 6093
with weights 6092
Neuroeconomics 8199, 8209
Neuromodulators 1555
Neuron 1813, 1814, 2090, 6099–6102, 6104, 6105, 6111
biolobical fractals 3789
fractals 3794
Neuronal coalitions 1439
Neuronal network 4900
Neuronal synchronization 1439
Neuroscience 6100
Neutral directions 999
Neutral mutations 5657
Neutron flow mapping, polymeric fluid 6825
Neutrosophic computing 4403
New ideas 8195
New kind of science 8846
New principle 1552
Newell's optimal velocity traffic flow model 9334
Newhouse 969
Newhouse phenomenon 3019
Newmann–Moser problem 5444
News 5410, 5431
Newton 6663
equations 6588
law of gravitation 6427, 6429
method control law 2062
scheme 5068, 5071
tangent scheme 5084
Newton–Fourier method 3834
Newton method 6764
Newton–Raphson 7407
Newtonian fluids 5927
Newtonian viscous body 1705
Nexus sampling 8214
NFS *see* Neuro-fuzzy systems
NIB *see* Network information bottleneck (NIB)
NIB algorithm, information curve 506
Niche construction 1925
Nicholson–Bailey model 2712
NIED 2413
Nielsen 7257
Nilpotent perturbation 6650
Nilsystem 3091
Nintendo Wii-mote 2076
Ninth GrC model 4347
NIST 4D/RCS 8438, 8440
Nitrogen content, plant tissue 1103
Nitrogen fluxes 2702
NL-computation, fuzzy logic 4005
NLS equation 4962
No-cloning theorem 7328, 7395
quantum mechanics 7374
No dictatorship 9937
No envy test 1680
No free lunch 4136
No incentive to separate 8132
No-slip 695, 5576
No trade theorem 436
Noble numbers 6682
Nodal attribute 8330
Nodal density 241
Nodal distribution 240
250G 241
500G 241
750G 241
1000G 241
Node degree 5694
distribution 1244
Node indegree 5266
Node strength, distribution 6057
Node to node communication 155
Nodes 724, 1600, 4182, 4871, 8322
control 155
of a network 5694
parallel computing 153
physical mobility 5608
worker 156
Noise 281, 476, 730, 732, 1989, 3361, 3367, 6105, 6109, 8737
autocorrelation function 3588
colored noise 476
conserved 3372
correlation dimension 484
external 3361, 3369
extrinsic 8737
extrinsic fluctuations 732
fractal time series 3794
input noise 733

internal 3361
intrinsic 8737
intrinsic stochasticity 732
multiplicative nature 3594
noise in gene expression 732
output noise 733
quantum information processing 7388
single chain dynamics 6820
stochastic processes 3577
thermal 3369
white or brown noise 482
1/*f* noise 3805
Noise amplification 3510
Noise correlation 3372, 9680
Noise correlator 3369
Noise data, stochastic processes 3583
Noise-immune 1388
Noise-induced transition 6487
Noise measurements 6151
Noise probability distribution 3370
Noise sidebands 6118
Noise terms 83
Noise trader 98, 5412
Noise transfer function 6153
Noiseless subsystem 7389
Noisy circuits, stochstic processes 3586
Nominal 8342
Nomogram 272
Nomograph 271
Nonabelian routh reduction 5496
Nonanticipative strategy 1950
Non-archimedean infinitesimal 6290
Nonatomic 5461
Nonatomic game 1682
Noncanonical poisson brackets 5494
Noncoding DNA, fractal features 3797
Non-conservative force 4379, 4381
Non-cooperative game 4855
Non-cooperative game theory 1510
Non-cooperative implementation 1512, 1513
Non-cooperative or strategic game 1662
Non-cooperative rules 6034
Non-crossing condition 1131
Non-degeneracy condition 6681
Non-degenerate 2263, 5074
Non-determinism 820
Non-deterministic fractals 7568
Non-deterministic gate 7255
Non dictatorial 9938
Non-dictatorial 9937
Non-differentiable 3859, 3873
Non-emptiness problem 8906
Non-English-speaking household 3137
Non-equilibrium 209, 7025, 8040
asymptotic behavior 7024
Non-equilibrium Green's function (NEGF) 7413, 7480, 7483
Non-equilibrium phase transition 888, 1538, 8931, 8936
Non-equilibrium phenomena 5915
Non-equilibrium stationary state 1002
Non-equilibrium statistical mechanics 4379, 4380
Non-equilibrium steady state 6953
Non-equilibrium system 3270, 3282, 4897
Non-equilibrium thermodynamics 477, 3270, 8043
Non-equivariance one-cocycle 5485
Non-equivariant momentum maps 5485
Non-equivariant reduction 5490
Non-extensive systems 5253
Non-Fermi liquid behavior 7356
Non-generic networks 2831
Non-generic rigidity 2838
Non-holonomic integrator 6273, 8623
Non-holonomic mechanical systems 5498
Non-homogeneous boundary conditions, Navier–Stokes equations 5934
Non-homogeneous cellular automata 916, 3233, 3238
Non-integrable 384
Non-interacting chain, physics 6813
Non-invertible transformation, isomorphism theory 4992
Nonisotropic scattering 7920
Nonlinear 2285, 2846
Nonlinear brain dynamics 8003
Nonlinear cointegration 2758, 2760, 2763
Nonlinear correlations 3718
Nonlinear coupled effects 5922
Nonlinear coupling 1540
Nonlinear diffusion equation 5243
Nonlinear dispersion 9996
Nonlinear dispersion relation 9977
Nonlinear dispersive waves 9970

Nonlinear dynamical system 1554, 5719
Nonlinear dynamics 8001
  ecosystems 2705
Nonlinear effect 6956, 6964, 6969
Nonlinear equations 274
Nonlinear error-correction 3444
Nonlinear estimator 4773
Nonlinear field theory 8596
Nonlinear filtering 10042
Nonlinear flow, polymeric fluid 6824
Nonlinear group velocities 10001
Nonlinear mapping 5473
Nonlinear multi-modal model 8961
Nonlinear observer 4767
Nonlinear open systems 477
Nonlinear oscillation 6972
Nonlinear oscillator potential 5225
Nonlinear oscillators 1555, 5724
Nonlinear panel data models 2781
Nonlinear partial differential equations (PDEs) 6216, 8480, 8504
Nonlinear partial least-square 2123
Nonlinear periodic model 8960, 8961
Nonlinear phase transitions 7992
Nonlinear phenomena 1875
Nonlinear QQSPR equations 7453
Nonlinear response 2433
Nonlinear Schrödinger equation 2050, 6138, 6456, 6709, 6712, 6715, 8474, 9970, 10002
  NLS 1376
Nonlinear Schrödinger equation (NLS) 6461
Non-linear Schrödinger equations 7449
Nonlinear Schrödinger equations 6231
Nonlinear single-transient model 8960
Nonlinear site effects 4440
Nonlinear stiffening 2841
Nonlinear systems
  controllability 4814
  fluctuations 3575
  observability 6375
Nonlinear time series 3475, 3478, 3494, 3500
Nonlinear transmission lines 8473
Nonlinear wave equation 2048
Nonlinear wave propagation in a building 2430
Nonlinear waves 6206
Nonlinearity 2428, 3829, 3843, 3867, 5415, 7996
Nonlocal phenomena 5922
Non-manipulable 9938
Non-Markov kinetic equation 1621
Non-Markovity parameter 1624
Non-mechanical proteins 7029
Non-monotonic model 5026
Nonnegative basis theorem 6247
Nonnegative matrix 6239
Non-Newtonian 5931
Non-orthodox sampling methods 8214
Nonparabolicity 7410
Non-parallel model 156
Non-parametric 3475, 3479, 3485, 3488, 3494, 3500
Non-parametric estimator 3443
Non-parametric methods 7017
Non-parametric test 6252
Non-perturbative 7347
Non-planar choreography 5969
Non-planar minimizers 5975, 5977
Non-planar symmetric orbit 5969
Non-player characters 1336
Non-recursive algorithms 3546
Non-regular dynamic behavior of wide moving jams 9399
Non-regular spatiotemporal jam dynamics 9401
Non-representative samples 5569
Non-resonance 6612
Non-resonance condition 5311
Non-resonant motion 6667
Non-revealing game 7634
Non-revealing strategy 7631
Nonsingular automorphism 2938
Nonsingular Bernoulli transformations 3061
Nonsingular Chacón map 3061
Nonsingular coding, ergodic theory 3075
Nonsingular dynamical systems 3056
Nonsingular ergodic theory 3056
Nonsingular joinings, ergodic theory 3073
Nonsingular maps, Cartesian products 3075
Nonsingular odometers 3060
Nonsingular systems
  spectral theory 3069
  unitary operator 3070
Nonsingular transformation 5457
  ergodic theory 3055, 3056
  hopf decomposition 3057

ITPFI 3068
smooth 3068
Nonsmooth analysis 6271, 6272, 6274, 9105
Non-standard analysis 4313, 6285, 7605
idealization principle 6304
overflow 6305
set-terminology 6300
standardization principle 6304
transfer principle 6304
ultra-power construction 6303
underflow 6305
Non-standard stochastic analysis, Loeb construction 6306
Non-stationarity 3756, 3758, 3764, 7021, 7996
Non-stationary activity 9845
Non-stationary process
Euclidean space of the state vectors 1619
normalization and attenuation of correlation 1618
statistical theory of a discrete non-Markov process 1618
Zwanzig–Mori's kinetic equation 1618
Non-synonymous mutations 5657
Non-transferable utility game 1511
Non-trapping invasion percolation (NTIP) 4949
Non-uniform hyperbolicity 4733, 4735
Non-uniformly expanding maps 963
Non-uniformly hyperbolic systems 2968
Nonwandering 2267, 8644
set 8644
Noonan syndrome 9097
Normal 6271
Normal boiling point 2199, 2210
Normal diffusion 3888
Normal distribution 3422, 4383, 7011
Normal fault 9614
Normal form 404, 713, 1038, 6175, 6183, 6310, 6335, 6628, 6632, 6633, 6650, 6761, 7139, 8927
Birkhoff 6320, 6321, 6324
game 3176, 3177
near equilibrium points 6591
procedure 331
semi-local 6321
takens 6321
theory 6652
truncation 4516
Normal frequency 2263, 4530, 5082
Normal inverse Gaussian distribution 3420
Normal linearization 2263
Normal modes 2263, 4537
nonlinear 6591
quantum computing 7231
Normal state 4378
Normal stress effect, liquids 5931
Normal variables 994, 995
Normalization 6311, 6627, 6629
Normalize 6316
Normalized 6315
Normalized event memory function 1624
Normalized Laplacian matrix 5354
Normalized linear (Hamming) distance 4066
Normalized mutual information 5356
Normalized quadratic (Euclidean) distance 4066
Normally hyperbolic invariant manifolds 944
Norman 5473
Normed counting measure, defintion 6307
Normed space 5713
Norming constant 4966–4968
North Atlantic oscillation (NAO) 1075
North west Pacific tsunami advisory center 9596
Northridge earthquake 9180, 9194
Northumbria University 3131
Nosé–Hoover thermostat 695
NOT 4878
Notation 2982
Notrade equilibrium 5397
Novelties analysis 2391
Novelty 1896, 3216, 3218
Novelty analysis 2388
Novelty detection 2381
Nowotny and Requardt 8844
Nozzle 9882
NP 5468, 6388, 6393, 7091, 7096
complete 6402
hard 7103, 7110
NP-complete problem 1542, 2093, 6418
NP-hard 288
NTIP *see* Non-trapping invasion percolation (NTIP)
Nuclear binding energies 4826
Nuclear polarization 8590
Nucleate 615
Nucleation 4497, 9771
parameters 4506

Nucleation model for traffic breakdown 9283
Nucleation probability 616
Nucleation rate 9297
  for traffic breakdown 9296
Nucleation zone
  earthquake 2539
Nucleocapsid 9752
Nucleolus 1662
  per capita nucleolus 1668
Nucleosome 3894, 3896
Nucleosynthesis bound 1636
Nucleotide strand asymmetries 3918
Nucleus 9359, 9361
  for traffic breakdown 9287, 9321
  required for the nucleation of an S → J transition 9361
Nudged-elastic-band method 1313
Null controllable 3528
Null model, networks 1187
Null space 5536
Nullifying 6290, 6292
Number of large 1-itemsets 4152
Number of states 4098
Number theory 3055
Numerical approximation 1505
Numerical attributes 7798
Numerical integration 1038
Numerical methods 5144, 5146, 5163, 7892, 9015
Numerical simulation 8592
Numerical studies 5024, 6602
Numerically 5028
Nutrients 2702
Nyquist rate 10044
NYSE, order system 5393

O

Obesity 4577
Object 206, 7754
  detection 3688
  manipulation 3688, 3697
Object concept 1894
Object extraction 9712
Object level 7790
Object permanence 1877, 1901
Objective function 5536, 5541, 5999, 6002, 6009, 9034
Objectives 7839
Obligation 3138
Oblique condition 6224
Oblique extension principle (OEP) 597
Oboukhov model 3658
Observability 4811, 6371, 9105
  complete systems 6372
  deterministic systems 6371
  estimation error 6373
  geometric setting 6371
  indistinguishability relation 6373
  inequality 1490
  input-output map 6374
  linear systems 6374
  minimal systems 6373
  nonlinear systems 6372
  piecewise constant controls 6374
  rank condition 6374
  symmetric systems 6372
  weak indistinguishability relations 6373
Observable 3533
  pressure of 6945
  quantum mechanics 7371
Observable payoffs 7622
Observation 5195, 7594
  space 6373
Observational 2793
Observational medicine 9092
Observatory earthquake message 2660
Observed covariates 2770
Observer synthesis problem 6372
Observers 6373
  arbitrary exponential rate 6374
  injectable systems 6378
  linear systems 6379
  persistent-excitation 6378
  skew-adjoint state linear systems 6379
Obsessive-compulsive disorder (OCD) 8004
Obstacle, avoidance 3687, 3689
OCC model 1347
Occam's razor 5198
Occupation number 880, 8589
Oceanography 8113, 8121, 8123
Oct-tree 2409, 8695
Octahedral partition 5815
24-Octahedral periodic minimizers 5970
Octane number 2198, 2199

Octanol-water partition coefficient 2144
OctArm 1477, 1478, 1483
OctArm continuum manipulator 1475
ODD *see* Overview, design concepts and details
ODE frame work, meta-dynamical 332
Odhner, W.T. 5476
Odometers 9240
  Markov 3060
  nonsingular 3060
Odometry 3682, 3684, 3688, 3693
OEP 597
Off-ramp bottleneck 9374
Off-the-equilibrium payoffs 4112
Offline learning 1816, 1819
Ohm's law 1405, 9273
Oil market models 685
OIL (ontology interchange language/ ontology inference layer) 8092
Olfaction 6112, 6113
Oligopolistic innovation diffusion model 1964
Oligopoly 7837
OLSE *see* Ordinary LSE method
Olshanski, A.Y. 780, 784
Ω-Extended wave function 7445, 7447
Omega-limit 4557
Ω-Limit sets 4681
Omega-squared model 2584
Ω-Stability 4731
OMEN 5780
Omori 2614
Omori's law 2559, 5025
On-line 3243
$O(n)$ model 8710, 8713, 8726
On-ramp bottleneck 9372
Onchocerciasis, effects of climate change 1101
One-cocycles 5485
One-component plasma 8579
1D Brownian motion 8711, 8721, 8722
1D CA 3236
One-dimensional heat equation, controllability 4813
One-dimensional neighborhood 3234
1D random walk 8713
One-forms, principle connections 5502
One-fracton relaxation 2257
One-mode data 8272
One-out-of-n encoding 1821
One-parameter scaling formula 8531, 8541
One-particle distribution 7808
One-qubit gate 7253
One-round procedure 9947, 9955
One-sided 3545
One-step processes 8770, 9291, 9292
Online internet applications of traffic prediction 9525
Online learning 1816, 1819
Onsager coefficients 3371
Onsager–Machlup functional 8779
Onsager statistical theory 3648
Ontogeny 930, 1916–1919, 1922, 1926, 1928
Ontological approach, agent-based modeling 192
Ontology 8084
  for modeling and simulation 355
Oort cloud 388
OP *see* Ordinary percolation (OP)
Open cluster 8696
OPEN-DX 5761
Open-ended evolutionary system 341
Open-loop control 632
Open-loop steady-state gain 8955
Open physical system 8927
Open science grid 5763
Open system 3271, 8929, 9254
Open-system operation 7388
Operant conditioning 1886
Operating principles 730
Operation, disconnected 5614
Operational amplifier 273
Operational thinking 4565
Operator 3861
  closed linear 1487
  collisional 9736
  covariant derivative 3872
  dilation 3869
  Floquet 6592, 6596
  gain 9736
  loss 9736
  Markov 8556
  quantum error correction 7340
  universal set 5865
  weighted 8555
Operator norm 5085
Opinion dynamics 7536
Opinion formation 6492

Opinion leaders 8309
Oppenheim's conjecture 3105
Opportunity cost 2285
Optical clockwork 6117
Optical computing 3356, 9135, 9705
Optical lattice 7162, 7163, 9679
Optical neural networks 6392
Optical pattern recognition 6391
Optical tweezers 7030
OPTICS 5058
Optimal algorithm 7119
Optimal control 2004, 6272, 9105
Optimal controls, existence 1502
Optimal exponent 5992
Optimal path 6501
Optimal "stationary" policy 4103
Optimal velocity 9417
Optimality principle 1503
Optimization 3354, 4190, 4494, 4781, 4900, 6272, 6407, 6408, 6491, 7848, 8869, 8871, 8875, 8988, 9034
  combinatorial 6407
Optimization algorithms, QSAR modeling 7073
Optimization principles 734
Optimization problem 1900
Optimization theorem, support vector machines (SVM) 4087
Option pricing 4117
Option pricing models
  empirical evidence 8792
  volatility and jump risk premium 8793
  volatility smile 8790
  volatility smirk 8790
OR 4878
Oracle 7088, 7090, 8673
  risk 8674
Orbit 3053, 6426, 6427, 8631, 9225
  brake 6604, 6606
  closure relation 9229
  coadjoint 5487
  generalized 5976
  in closed non-simply connectedHill regions with boundary 6607
  in compact Hill regions with boundary 6606
  in compact Hill regions without boundary 6605
  internal periodic 6604
  non-planar symmetric 5969
  nondegenerate 6595
  normal 6595
  periodic 6595
  relation 9229
  relative periodic 6598
  sequence 9229
Orbit classification, type III systems 3065
Orbit reduced space 5491
Orbit reduction theorem, symplectic 5491
Orbit reduction, mechanical systems 5487, 5488, 5490
Orbit space 5306
Orbit theory, ergodic 3064
Orbital stability 6434, 6436, 7606
Order
  dominance 8354
  fractional 3863
  lexicographic 1667
  linear 9946
  median 9945
  of the reaction 329
  partial 9946
Order backlogs 679
Order-driven
  trading mechanism 5393
Order flow 5420
  commercial-customer 5423
  customer 5429
  financial-customer 5423
  interdealer 5420
Order parameter 627, 631, 635, 636, 648, 1275, 1538, 1539, 2723, 3361, 3626, 3850, 5719, 7993, 8004, 8713, 8715, 8927, 8928, 8930, 9683
Order parameter equation 8933
  complexity in time 3576
Order statistic, largest 3321
Order statistics 229
Order-to-order transition 7994, 7999
Ordered field
  definition 6288
  extension 6288
Ordered region 6427
Ordering 3233
  definition 6291, 6292
  task 3238
Ordinal classification 7753, 7754, 7760

Ordinality 1682
Ordinary differential equation (ODE) 1228, 2285, 3256, 8630, 8993
artificial chemistry 332
Ordinary least-squares estimator (OLSE) method 4084, 9193
Ordinary percolation (OP) 4948
Oregonator model 1355, 7549
Organization 6978
set theory 338
syntactically and semantically closed 339
Organization center 703, 714
Organization principles
DNA 3795
Organizational cybernetics 8974, 8976, 8980
Organizational learning 7849, 8988, 9044
Organizer
Nieuwkoop center 538
Spemann–Mangold 538, 539
vegetal pole 538
Orientation distribution 6570
continuum percolation 6509
Orientation, fracture networks 6506
Orientational order 2026
Origami-based self-assembly 261
Origami language 266
Origami shape language 929
Origin-Destination estimation 9450
dynamic synthetic 9450
static synthetic O-D 9450
synthetic O-D 9450
Origin of order 8045
Origin of replication 3894, 3908
Orness 228
Ornstein D 784, 786
Ornstein theory 3055
Ornstein–Uhlenbeck process 5225, 5234, 8775
Ornstein–Zernicke scattering function, polymer physics 6817
Ornstein's isomorphism theorem 4987
Ornstein's proof 4984
Orthogonal collocation 6334
Orthogonal mapping 3558
Orthogonal wavelets 10020
Orthogonality theorems 6736
Orthogonalization 4739
Orthography 2311
International phonetic alphabet 2315
Orthologous data 9094
Orthonormal basis 10031, 10034
Orthonormal M-wavelet basis 590
Oscillations 538, 540, 541, 545, 729, 3146, 6099, 6100, 6108, 6110, 6112, 8201, 8205
fast 6110
gamma 6109
nonlinear 7614
pedestrian 3146
relaxation 7603
self-sustained 7603
slow 6107, 6108
sparsely synchronized 6109
Oscillator 421, 535, 541, 892, 4240, 4241
harmonic 7137, 7603
Oscillator level 5724
Oscillatory 535, 536, 540, 543, 545
chaotic 527
stable 527
transient 527
unstable 527
Oscillatory dynamics 482
biological clocks 485
respiratory oscillations 483
temperature-compensated 483
ultradian clock 483
Oscillatory flow 6482
Oseen fundamental tensor, liquids 5939
Oseledets theorem 960
Oslo model 1136
Osmotic pressure, polymer physics 6817
Ostwald ripening 6467
Other memories 832
Other quantum key distribution protocols 7279
decoy state protocols 7280
differential-phase-shift-keying (DPSK) protocols 7282
entanglement-based protocols 7280
Gaussian-modulated coherent state (GMCS) protocol 7282
strong reference pulse protocols 7281
Other technique 1156
Our common future 4635
Out-of-equilibrium 7009
Out-of-equilibrium behavior 7586

Out-split graph 8897
Outcome 7773, 7999
Outdegree 6241
Outer Lindblad resonance 412
Outer measure 5455
Outer rise 9584, 9589
Outgassing 3286, 3295
Outlier 1989, 2409
Output elasticity 2285
Output function 3524
Output-injection, observers 6378
Output layer 1815, 1818, 1822
Outstanding problems 6672
Over-crowding 6487
Over-deceleration effect 9334
Over-dispersed 9838
Over-searching 5350
Overall connectivity index 4833
Overall Wiener index 4833
Overall Zagreb indices 4833
Overbooking 7844, 7845
Overconfidence 3473
Overconstrained cluster 2835
Overdamped Brownian motion 8776
Overfitting 1816
Overlap factor 4152
Overlap QSM 7448
Overlapping community 1146
Overlapping-generations model 2819
Overlapping of two Fano resonances 7297
Overruns 1846
Overshoot 5416
Overtime 1852
Overtraining 1816
OWA 229
OWL (Web ontology language)
  DL 8084
  full 8084
  lite 8085
Oxametallacycle, eclipsed 1330
Oxidative phosphorylation
  energy supply 480
  inner mitochondrial membrane ion
    channels 480
Ozone concentration, Germany 1099
Ozone hole, human health 1099

P

P 5468, 6388, 6393, 6402
P and *S* wave velocity models for Taiwan 9199
P-Boundary 7756, 7761, 7769
P-Coda 7926
P-Dimension 5216
P-Dominance 7767
P-Elementary sets 7755
P-Glycoprotein (Pgp) 2168
P-Glycoprotein (Pgp) 2127
p-Graph 8258
P-Lower and P-upper approximation 7769
P-Lower approximation 7755, 7760, 7767
p-Model 8539
p-n-p-n device 9714
P system 928, 5524
P-Upper approximation 7755, 7760, 7767
P-Wave 2662, 2666
  large earthquakes 2543
P wave 2476, 2482, 2607, 2611
P-Wave magnitude
  short-period 2481
P-Wave magnitude scale 2665
P-Wave velocity pulse 2539, 2540
p^* model 8319, 8320
p_1 model 8321
p_2 model 8321
PAC learning 5210, 5211
Pacific tsunami warning center 2478, 2484, 9596
Pacific tsunami warning center (PTWC) 2659
Packard 4500
Packing
  deformable 5000
  frictional 4999
  frictionless 4999
  isostatic 5000
  rigid 4999
Packing fraction 5015
Packing number 5206, 5213
Pad binding mechanism 5478
Padé approximants 6740, 6746
Padé-approximants 6634
Padé approximate 6207
PageRank 5273
Painlevé's theorem
  n-body 5978

Paintable computing 262
Pair cascade 35
Pair-correlation function 5016, 8576
Pair-distribution function 8576, 8578
Pair of biorthogonal M-wavelet bases 591
Pair of dual M-wavelet frames 591
Pairing 9683
Pairing symmetry 9698
Pairwise comparison 9948
  coefficient 9948
  method 9943
  table 7766
Pairwise independence, property 8558
Pairwise instance similarity 5354
Pairwise majority voting 9935, 9937
Pairwise stability 6027, 6042
Pajek 1256, 8848
Palais–Smale condition 6612, 6621–6623
Paleoclimate 3
Paleoseismological 2562
Paley–Littlewood partition of unity 2044
Palindromes 744
Palis conjecture 959
Palis–Yoccoz theorem 3019
Panarchy 4632
Panel data 5554, 6006, 8313
  fixed effects 5570
  incidental parameter 5566
  random effects 5554
Panic 3146, 3151, 3153, 3154, 6476, 9421
Pappus of Alexandria 232
Parabolic 1725, 2264, 4521
Parabolic equations
  degenerate 1506
  nonlinear 1493
Parabolic scaling law 1725
Parabolic systems 3020
Parabolic wave equation 7921
Parabolic wedge 1727
Parabolicity, uniform 6219
Paradigm 4187, 7832
  amorphous computing 260
  programming 259
  shift 7833
Paradox of voting 9934, 9939
Parallel algorithms 7892
Parallel cellular machines 3238
Parallel computation thesis 6402
Parallel computing 1702, 5439
  agent-based modeling 153, 157
  massively 259
Parallel dynamical system 165
Parallel electric field 22, 32
Parallel time 6402
Parallel Turing machines 748
  processor complexity 748
  relations between PTM complexity 749
  space complexity 748
  time complexity 748
Parallel, massively 6400
Parallelism 5525
Paramagnetic state 8712
Parameter 731
  optimization 6480
  space 1696, 6994
Parameter estimation 51, 6006
Parameter régimes 1554
Parameterized complexity 9959
Parametric amplification of vacuum fluctuation 1635, 1638
Parametric driving 5734
Parametric excitation 2323, 6626, 6627, 6629, 6633
Parametric oscillator 8940
Parametric resonance 2323, 6626
  combination 2334
Parametric stabilization 1557
Parametrically excited pendulum 6634
Parametrized core set 7462
Pareto distribution 2801
Pareto-dominates 3332
Pareto-efficient 9936
Pareto-optimal set 7778, 9042, 9936
Pareto property 9945
Pareto, Vilfredo 2793, 7011
Parity check matrix 7327
Parkfield segment 7855
Parkinson's disease 627, 630, 1626
PARPACK 5755
Parry measure 2984
Parry's entropy 3071
Parsimony principle, neuro-fuzzy systems 6095
Parsons, Talcott 8268
Part relation 4420

Partial algebra 8292, 8296, 8299
Partial algebraic 8293, 8305
Partial conditional dependence 8326
Partial containment 1796
Partial covering (PCov) 4345
Partial differential equations 5800, 6655, 7892
  biharmonic equations 5827
  fluid flow simulation 5830
  modeled by amorphous computing 264
  Navier–Stokes equations 5830
  Poisson equation 5823
Partial effects 2771
  at the average 2781
Partial FSSP solution 3553
Partial history within a stage 4106
Partial hyperbolicity 4733
Partial result authentication codes (PRAC) 5612
Partial role algebra 8296
Partial semigroup 8296
Partially hyperbolic difeomorphism 3099
Partially hyperbolic system 965, 2966
  group extensions 2968
  time-One maps of Anosov Flows 2968
Partially ordered algebra 8292, 8294
Partially ordered semigroup 8292, 8293, 8295, 8297
Partially revealing strategy 7632
Participative model building, system dynamics 4572
Particle 772, 3233, 3237
  collisions in the n-body problem 5961
  computational 259
  fluid 5928
  quantum 3863
Particle communication, in amorphous computing 267
Particle filter 8805
Particle identity in amorphous computing 267
Particle size distribution 9799
Particle swarm optimization (PSO) 123, 1788, 2122, 8869, 8875
Particle trajectories 4219
Partition 1141, 4103, 4359, 8247, 8290
Partition coefficient 2200
Partition function 1271, 3768, 5015, 8538, 8539, 8715, 8717
  Potts model 3741
Partition matrix 1779
Partitioning of elastic energy 2591
Parts 1540
Pascal 9412
Pascal, Blaise 5476
Pascaline 5476
Pass-transistor logic 9720
Passive abstractions, agents as 194
Passive elasticity 7040
Past a sphere at increasing Reynolds numbers, flow 5940
Patchwork evolution 5663
Path 2201, 2211, 3860, 6031, 6241, 8250
  augmenting 6415
  shortest 6409
  successive shortest path algorithm 6411, 6413
Path coordination 5789
Path dominance 6036
  core 6027, 6040
Path effects 4435, 4440
  basins and other deterministic deviations from a flat-layered model 4435
  broadband wave-propagation computations 4468
  far-field displacements of P- and S-waves 4453
  incoherently scattered wave energy 4454
  random heterogeneities in the three-dimensional velocity-density structure 4435
  ray theory 4453
  wave-propagation in heterogeneous media 4437
  waves in a flat-layered attenuating earth 4435
Path-generated cost sharing rules 1682
Path integral 8591, 8937
  polymer physics 6811
Path integral representation, stochastic processes 3580
Path integrals 3370, 8778
Path integration 7126
Path planning 1225, 6071
Pathogens 425
Pathways 719, 724, 4835
Patient flows, characteristics 4571
Pattern 782, 4179, 8048, 8926, 10032
  and interfaces in dissipative dynamics 8947
  chaotic 6464
  earth-specific 4190

glassy 6473
hexagonal 6462, 6464
mutually erasable 783
premonitory 4179
striped 6464
support 782
3D 6462
universal 4189
wave 6468
Pattern formation 3361, 8947
and pattern recognition in the city 7962
Pattern growth 5440
Pattern informatics (PI) 2441
Pattern modeling process 8996
Pattern pattern matching 9515
Pattern perception 8003
Pattern recognition 3314, 4185, 8938, 9855
Pattern searching 8257
Patterning 1901, 1904
Pattison, Phillippa 8274
Paul trap, spherical 7223
Pauli exclusion principle 4824
Pauli matrices 7333
Pauli potential 8583
Pavlov, I.P. 1886
Payne's macroscopic traffic flow model 9335
Payoff 9034
Markovian strategy 2223
Payoff dominant equilibrium 3196
Payoff function 1528, 9036, 9040
Payoff type 5184
Payoff vector
approachable 7625
decomposable 7626
PCL *see* Primary constraint language
PCTs *see* Primary care trusts
Peak ground acceleration 2400
Peakon 80, 84, 87
Peano curve 8710
Peapods 691
Pebble game 6014, 6017, 6019, 6020
Pebbles 2839
PEC 4101, 4107, 4111
PEC is always competitive in 4110
Pedersen conductivity 33
Pedestrian 3144, 6476
Pedestrian area module 3147
Pedestrian dynamics 9413
Pedestrian traffic 1180
Peer-to-peer (P2P) search engine 5281
Pendulum
damped pendulum 8634, 8635
pendulum with torque 8646
Pendulum equation 6621, 6622
Pendulum model 6792
Penetration pricing 1969
Penguin effect 1965
Penning trap, quantum computing 7224
Penrose tilings 3614
Pentagon-heptagon 699
Pentagonal tessellation 893–895
Per-site entropy
E. coli tRNA sequence 496
rodent homeodomain 500
Perceived implications 7839
Perception 1225, 3134, 6492
computational theory 4005
concept of 193
Perception–action cycle 4895
Perceptron 1817, 2143, 4427
learning rule 1817
Percolating networks 2255
Percolation 609, 1137, 1443, 3708, 3721, 3737, 5670, 5676, 6412, 6495, 6504, 6784, 6806, 7014, 7568, 8710, 8713–8715, 8726
backbone of the percolation cluster 1449
biological fractals 3789
bond 1443, 3849, 6579, 6580
cluster 1443
conducting backbone 3744
Coniglio's theorem 3747
continuum 1450, 3849
conventional 609
critical number of bonds 1452
directed 1137
exponent 3852
fracture networks 6507
hopping 3850
invasion 3848
inverted random void 1444
k-core 623
lattice 1448
nodes, links, and blobs picture 3743
oriented 610

percolation threshold 1447
polymer physics 6819
probability of 6511
quasi-one-dimensional "bubble" model 3748
rigidity 610
site 1443, 3849, 6579
site-bond 3849
threshold 3847, 3849
transition 6414
transition zone 6513
tunneling 1444
Voronoi 6586
Percolation cluster (IPC) 319, 7568, 8714
finite 319
probability of 6512
Percolation conduction problem 1405
general 1405
standard 1405
Percolation critical threshold 3791
Percolation models 2561
Percolation network 3880, 3882
Percolation parameter 6511
Percolation probability 6541
Percolation properties 6514
fracture networks 6510
geological fractures 6505
Percolation theory 1395, 1397, 2722, 6566
Percolation threshold 1395–1397, 1401, 1403, 4956, 6497, 6510, 6512, 6513, 6539, 6541, 6542, 6544, 6566, 6579, 6580, 8715
Percolative models 1396
Percolative system 1395
Perfect adaptation 731
Perfect equilibrium with complexity 4101
Perfect market 2285
Perfect matching 6419
Perfect recall 4863
Perfect reconstruction 10023, 10024
Performance 4187, 7839
algorithm 4187
Performance moderator functions 1347
Perihelion 6677
Period 6243, 7605
minimal 6590
refractory 7606
Period distribution 2533, 2536
Period doubling 377, 407
Period-doubling 6336
Period-doubling bifurcation 2271, 4523, 5308
Period-doubling route to chaos 2719
Period multiplication 6136, 6166
Period of a network 8296
Period solutions, perturbed 8643
Periodic attractors 8425, 8426
Periodic like trajectories 6710
Periodic medium 6416
Periodic orbit 405, 6426, 6428, 6429, 6674, 6684
Periodic orbit theory 7164
Periodic point 172, 5306, 9248
measure 6949
Periodic solution 5301, 6321
subharmonic 6322
Periodic structure 9250
Periodicity 9835
Periodicity conditions 6430
Periodicity of earthquakes 2558
earth tides 2558
Peripherality 8271
Permafrost
dynamics 1715
models 1714
Permeability 1399, 2616, 2679, 6567, 9767, 9775, 9792
fracture networks 6516
Permutation 5439, 8358
Permutation tests 6266, 6267
Permutative CA 2981, 9253
Perrier's loop 8072
Perron–Frobenius theorem 6242, 6243
Perron root 6242
Perron theorem 6242
Perron vector 6242
Perron's method 6222, 6223
Persistence 3757, 6311
Persistent 6311
Persistent current 7671, 7676
Persistent property 4519
Personal identity 1906
Personal mastery 9047
Personal network exposure 8309
Personalized medicine 9092
Perspective projection 56

Perturbation 731, 1541, 1881, 5722, 6310, 6311, 6649
attractors 1205
geometrical 7607
network parameters 731
problem 6311
regular 7607
singular 7604, 7607
state perturbations 731
Perturbation expansion 3364, 3365, 3367, 3369
Perturbation problem 4516
Perturbation theory 6163, 6175, 6673, 6724, 6730
free fall 6725
Rayleigh 6725
soliton 6163
vibrating string 6725
Perturbative expansions 3361, 3370
Perturbative methods 623, 7527
Pervasive computing, amorphous 269
Peschansky 3547
Pesin 2862
Pesin entropy formula 960
Pesin theory 960
Pest 1075
PFSM 3696
Phage λ 730, 8736
Phantom panic 6487
Pharmaceutical synthesis 417
Phase 3352, 4195, 4197, 5728
Phase condition 6331
Phase diagram 652, 5001, 5014, 5035, 9680
jammed 5001
point j 5001
Phase diffusivities 6463, 6469
Phase dynamics 6463, 6467
Phase equilibria 6465
Phase estimation algorithm 7384
Phase field 6465
Phase field models 3636
Phase locking 1544, 6636, 9680
Phase noise 6117
Phase plane 6131
Phase plots 8828
Phase portrait 6330, 7137
Phase protection mechanism 6447
Phase-response-curve 6108
Phase scattering 1550
Phase separation 6467, 6784, 9681
polymeric fluid 6828
Phase shift 1020, 7610, 8480
quantum computing 7238
Phase shift operator 5310
Phase shift oscillators 562
Phase space 166, 483, 695, 2264, 2860, 4516, 6136, 6177, 6426, 6994, 7997, 7999
Phase speed (celerity) 8483
Phase transition 1544, 1545, 1899, 3361, 3371, 3756, 3847, 3849, 4894, 4910, 4998, 5024, 5674, 5719, 6382, 6772, 7855, 8000, 8004, 8005, 8040, 8199, 8206, 8277, 8367, 8369, 8370, 8713, 8714, 8724, 8928, 9416, 9734
absorbing state 3366
active to absorbing 3368
continuous 611
equilibrium 4998
ergodic theory 6940
fluctuations 4998
jamming 4998
non-equilibrium 4998
nonequilibrium 3368, 3369
polymer physics 6819
Phase trapping 1550
Phase turbulence 6473
Phase velocity 9969, 9977
Phase wrapping 5733
phaseRes 6398, 6399, 6401
Phasic average 2679
Phasor amplitudes 6148
Phenomenological synergetics 8929, 8937
Phenomenology 4893
Phenotype 118, 9106
quantitative phenotype 9112
Pheromone 8869, 8871, 8877
Phi complex 8206, 8208
Philippine Institute of Volcanology and Seismology (PHIVOLCS) 9867
Phillips machine 8992
Philosophical background and underpinnings 8988
Philosophical structuralism 6796, 6797
Philosophy 4894
Phishing 4373
Phone 2310
Phoneme 2310

Phonology 2310
Phonon regime 2256
Phonons 700, 701
Phosphofructokinase 562
Phospholipidosis 2171
Photochemical smog, global warming 1098
Photodetection 6146
Photoelectric effect 7251
Photosynthesis 3280
  temperature dependence 1102
Phylogeny 930, 1916, 1926
Physical agents 1338
Physical biology, ecological systems 2700
Physical computation 3221
Physical evolution 3270, 3279
Physical gels 6546, 6547, 6563
Physical interaction force 6485
Physical measure 4731
Physical mobility of nodes 5608
Physical models, biological fractals 3785
Physical parameterizations 7594
Physical processes in volcanoes 1223
Physical systems 3836
Physically relevant measures and strange Attractors 2969
  Hénon Diffeomorphisms 2971
  Intermittent Maps 2971
  Unimodal Maps 2970
Physics 188
  of non-regular spatiotemporal dynamics 9403
  of the city 9411
Physiological synchronization 7995
Physiological time series 1624
Piaget, J. 1885, 1896, 1898, 1903
Piagetian 1896
Piaget's main stages 1885, 1888, 1899
Picard iteration 3261
Picard–Lindelöf theorem 3261
Picking 2411
Piezo-electric effect 2621, 2623
Piezo-electric polarization 2630
Piezoelectric 5765
Piezoelectricity 5752, 5763
Pimephales promelas 2148
PIN *see* Protein interaction networks
  easley and O'Hara model 5399
Pinch effect in synchronized flow 9289, 9358
Pinch region 9394
Pinch region of synchronized traffic flow 9323
Pinched torus 2264, 4531
Pinhole camera model 4753
Pinning 1126, 1129, 2019, 6469
  impurity 1129
  length 1129
Pinsker σ-algebra 2928
Pinwheel calculator 5476
Pipelining in amorphous computing 266
Pismen 8940
Pitch angle 24
Pitchfork 6633
Pitchfork bifurcation 5305, 5722
Plan abstraction 5190
Plan-space planning 5192
Planar 3-body problem, symmetry groups 5972
Planar equivariant minimizer 5968
Planar flow of a viscous liquid past a cylinder 5941
Planar graphs 1254
Planar homography 4752, 4756, 4767
  continuous 4767
  matrix 4753, 4757
Planar motions, liquids 5930
Planar symmetry groups 5972
  3-body problem 5971
Plane convex fractures, excluded volume calculation 6508
Planetary problem 6683
Planetary rover 3690
Planetary three-body problem 6428
Planetery exploration 3683
Planetesimal 8696
Planets 3274, 3279
Planforms 6461, 6462
Planimeter 5474
Planning 7850, 7953, 8377
  and decision making 8443
  interactive 7850
Planning knowledge
  domain-configurable 5192
  domain independent 5192
  domain-independent heuristics 5189
  domain-specific 5192
  domain-specific control knowledge 5189
Planning operators 5191

Planning systems, interactive 7850
Plant-pollinator networks 2736
Plants 3280
  impacts of climate change 1102
Plasma 8523
Plasma analogy 8576, 8586
Plasma instabilities 31
Plasma resonances 8526
Plastic flow 699
Plasticity 628, 639, 641, 6112
  synaptic 629, 641, 6111
Plate motions, measurement using GPS 4255
Plausibility ordering 6930
Player set 1520
Player/Stage 2078
Players 1660, 3327, 8698
Players' preferences 6026, 6031
  strong preferences 6026
  weak preferences 6026, 6031
Playground swing 2324, 2343
Pleiotropy 9106
Plug 9769, 9881
Plug-in probability, DNA sequence 495
Plural rationality 7834
Plurality rule 9947, 9954, 9955, 9960
Plurality rule with run-off 9948, 9955
Pluripotent cells 1192
Pluto 6445
PMFserv 1347
Pochhammer–Chree equation 6455
POEtic machines 930
Poinaré map 6435
Poincaré 704, 6663
  Domain 8635
Poincaré–Birkhoff fixed point theorem 6439
Poincaré domain 6653
Poincaré–Dulac normal form 6179
Poincaré–Dulac theorem 6720
Poincaré–Lindstedt 6627
Poincaré-Lindstedt 6627, 6629, 6630
Poincaré-Lyapounov-Nekhoroshev theorem 6186
Poincaré map 5306, 6336, 6426, 6435, 6447, 8638, 8644
Poincaré mapping 2265, 4518
Poincaré maps 6168
Poincaré Operator 6613
Poincaré operator 6612, 6614
Poincaré recurrence principle 3085
Poincaré Recurrence Theorem 3054
Poincaré renormalized form 6187
Poincaré section 384, 2265, 6195, 6203, 8639
Poincaré transformation 6178
Poincaré's theorem 6663
Point
  critical 6600, 6610
  minimum 6600
  multiperiodic 6610
Point j 5011, 5015, 5016
Point of view
  Eulerian 9729
  Lagrangian 9729
Point reduction, mechanical systems 5487, 5490
Point source 2485, 2488
Point source model 2485, 2488
Point system 9954
Point vortex motion 3647
Point vortex system 3647
Point/configuration, random 255
Pointedness 1444
Points 8267, 8270
  fixed 6104
Pointwise convergent sequence 3256
Pointwise D2 7997
Pointwise dimension
  ergodic theory 3017
  hyperbolic measure 3023
Pointwise ergodic theorems 3058
Poiseuille flow, biological fractals 3783
Poiseuille solution 9772
Poisson 5765
Poisson brackets 2264, 5447, 6587, 6589, 6713
  noncanonical 5494
Poisson distribution 2558, 7585
  Poisson process 2558
Poisson equation 5577, 7315, 8696
Poisson–Lie structure 5447
Poisson manifold 5447, 5493
Poisson probability distribution 6145
Poisson process 3445, 6002, 6009, 6144, 9835
  nonhomogeneous 9836
  simple (homogeneous) 9835
  stationary 9836
Poisson ratio 9776
Poisson space 2264

Poisson structure 2264
Poisson suspensions, ergodic theory 3076
Poisson symmetry 2264
Poisson's equation 7487
Poland 4927
Polarity in amorphous computing 265
Polarization 3218, 5766, 6121
Polarization field 6121
Polarization loop 8581
Polarization state 1378
Polarized 3218
Polaron 9680
Polaron problem 8591
Pole placement 8619
Pole shifting theorem 3534
Policy 7845
Policy aware web 8094
Policy design 7847
Policy making 8377
Policy optimization 9034, 9038
Policy parameter 7848
Political economy 2285
Political ignorance 9937, 9941
Political rhetoric 8136
Politics 3137
  fission 4618, 4619
  median duration 4620
Pollen allergies 1100
Pollution
  anthropogenic 7842
  carbon-dioxide (CO_2) 7845
Polydisperse fractures 6506, 6511
Polyelectrolyte 6816
Polygons, fracture networks 6506
Polylogarithmic TDMA broadcast 2054
Polymer blend 6827
Polymer chain, ideal non-interacting 6813
Polymer chemistry, autocatalytic 332
Polymer contour 6814
Polymer physics 6810
Polymer statistics 6813
Polymerase 419
Polymerase chain reaction 2093
Polymeric fluid 6824
  dynamics 6819
  multi-phase 6827
Polymerization 6811
Polymers 1304, 6409
  directed 6409
  nonlinearity in composite 1304
Polynomial 712
Polynomial recurrence ergodic theory 3063
Polynomial representation 4739, 4743
Polynomial time reduction 5468
Polyphase matrix, multiwavelet 5849
Polyproteins 7032
Polzik 7263
Pontryagin's maximum principle 1503, 2012
Popescu 7257
Population 3115, 4134
  concentration index 4620, 4622
  dynamical fractals 3793
  Malthusian effects of resource scarity 4622
  power-law growth 4622
  resource scarcity 4623
  resources 4626
  world 4622
Population averaged effect 2776
Population biology 8777
Population code 3350
Population coding 3350
Population density 1047, 1049
Population dynamics 3362, 3368, 4925, 6466
  biological fractals 3780
Population game 3176, 3180
Population growth 1057, 1058
Population inversion 6120
Pore coordination number 4952
Pore fluid pressure 10049, 10052
  Hubbert–Rubey fluid pressure 10050
Pore geometry 2687, 4952
  measurement 4952
Pore pressure 2616
Pore scale geometric correlation, invasion
    percolation 4955
Pore scale structure 4951
Pore scale topology, measurement 4952
Pore size distribution 4953
Pore topology, invasion percolation 4952
Pore tortuosity 4953
Poroelasticity 2686
Porosity 2687, 9780, 9792
Porosity wave 2685

Porous media 6566
fractured 6505
Porous media equation 6231
Port Douglas 3129
Pöschl–Teller potential 7666
Pose 48
Pose estimation 4765, 4770
Poset 9945
Position 8289
Position limits 5410
Positional equivalence
exact 8340
institutional 8344
observational 8343
regular 8339
structural 8339
Positive correlation 3185
Positive feedback 565
Positive feedback loop *see also* Reinforcing feedback loop, 683
Positive hole (p-hole) 2623, 2630–2632
Positive-real lemma 8756
Positively expansive, cellular automata 2985
Positively invariant set 6615, 6618
Positivism, management science 9066
Possibilistic constraint, fuzzy logic 3998
Possibilistic logic 6931
Possibility 4425
comparative 6928
conditional 6931
guaranteed 6927
Possibility correspondence 432
Possibility distribution 6927, 6929
subjective 6935
Possibility function 4877
Possibility measure 6927, 6929
Possibility relations 6930
Possibility theory
fuzzy logic 4005
qualitative 6930
quantitative 6932
Possible source geometries 9913
Possible worlds 8341
Post-earnings announcement drift 5416
Post Keynesian economics 8996, 8998
Post-Moore's law 291
Post–Moore's law computing 3335
Post-Moore's law computing 3336, 3358
Post-project 1862
lessons 1862
Post-transcriptional regulation, transcripts 4171
Posterior density 445, 448
Posterior distribution 443, 456
conditional posterior distributions 451
Posterior odds 446
PostgreSQL 2404
Postmodern city 7957
Postseismic deformation
afterslip 4273
and high-rate GPS 4276
poroelastic response 4273
viscoelastic response 4273
Postulate
Born 3874
von Neumann 3874
Potein tyrosine phosphatase 1B 2146
Potential 1395, 1548, 3244, 4962, 4967, 6101
extracellular 6100
gauge 3876
intracellular 6100
Potential customers 1960
Potential energy surface 1309, 1312
Potential fields 5792, 6071
Potential function 1666, 5720, 6003, 6044, 6429
Potential game 3186, 3197
Potential perturbation 7407
Potential shaping 2015
Potential surface 281
Potential systems, stochastic processes 3586
Potential well 8482
Potts 7825
Potts model 3740, 6420, 8710
Poverty alleviation 8989
Power law 22, 219, 477, 1049, 1050, 2715, 2812, 3863, 4664, 4934–4936, 4938, 4940, 6108, 6490, 7011, 8238, 8240, 8263, 8277, 8367
behavior 6773, 8544
distribution 1899, 4665, 5022, 5269, 7011, 7996
function 2801, 2813
inverse 477
network 570, 575, 577–579
scaling behavior 5027
size distribution 661
Power mean 228

Power series 6761
expansions 5074
Power set, definition 6302
Power spectral analysis 479
fast Fourier Transform 481
spectral exponent 479
Power spectrum 1279, 3760, 3774, 3804, 8712
Powersim 9015
PPP 5428
PRAC *see* Partial result authentication codes
Pragmatic frame 355
Prandtl number 3626
Pre-Robinson infinitesimals 6286
Preacceleration 27
Prebiotic evolution 8940
Precautionary principle 4639
Precession 6677, 6679
Precipitation 2612, 7596
monthly mean precipitation 7591
summer precipitation 7601
winter precipitation 7599
Precisiated natural language (PNL) 4005
Precisiation types, fuzzy logic 3995
Precisiation, fuzzy logic 3990, 3994, 4004
Precision 271, 273, 274, 276, 281, 283, 291
fuzzy logic 4004
limited 4048
Precursor film 3622, 3633
Precursors
complete model 2568
of volcanic eruptions 9855
partial model 2569
Predation 8355
Predator-prey interactions 8772
Predator-prey relations 6131
Predecessor states 64, 73
Predicate calculus 349
Predictability 3454, 4179
Predictable and reproducible pattern features 9378
Predictable and unpredictable phase 5598
Predictable features
of effectual road traffic bottlenecks 9519
of traffic time series 9519
Prediction 456, 2143, 3301, 9862, 10010, 10022
advance 3315
algorithm 3302
problem 3301
targets 3302
Prediction intervals 6935
Prediction rule 5178
Predictive distribution 456
Predictive modeling
classification 1991
regression 1991
Predictive network 9095
Predictive QSAR modeling
flowchart 7080
workflow 7079
Predictor-corrector 7407
Predictors 3244
Predominant period 2412
Preference aggregation 9943
Preference function 6009
Preference information 7760, 7782
Preference model 7753, 7782
Preference relation 1660, 7760, 7766
comprehensive 7766
graded 7766
strength of 7766
Preference supernetworks 6032
Preference voting 9948
Preferences 2243, 7769, 7770, 7774, 7778, 9655, 9943
additively separable 9655
bi-normal 1678
cardinal utility 1673
complete preorder 7770
for simple strategies 4110
interval scale 7770
maximin 9655
numerical non-quantitative scale 7770
ordinal scale 7770
over action profiles 1669
over consequences 1669
quantitative scale 7770
responsive 9655
$r(f)$-separable 9655
rich domain of preferences 1673
strictly convex 1677
strictly quasi-concave 1677
strongly substitutable 9655
substitutable 9655
utility function 1669

Preferential attachment 725, 6004, 6051, 8238, 8240
fitness model 6052
linear 6052
nonlinear 6052
Preferential attachment rule 8263
Preferential voting 9948
Preferential voting procedure 9955
Preferred-basis 6247
Preferred basis theorem 6247
Pregeometric theories of emergent space-time 8843
Prehension development 1899
Prenatal development 1900
Preorder, complete 9946
Preparation 3110, 3131
community 3111
hazard zone 3111
household 3111
household safety preparedness and action index 3114
individual 3111
precautionary action 3111
ramp-up 3110, 3113
Preparedness 3113
conceptual shift 3116
growth industry 3115
holistic 3122
household preparedness and safety action index 3115, 3122
institutional barrier to change 3115
knowledge base 3114
people's propensity to respond 3114
possible travel mode 3114
public participation 3116
well-being 3114
Preplay communication 1588
Preschool years 1901
PreSEIS 2398
Present-value constraint 3452
Present-value model 3454
Presentation duration 2245
Presentationalism 9063, 9064, 9066, 9067
in system dynamics 9067
Preserved fibrations 4600
Preserving property 5738
Preslip model
earthquake nucleation process 2541
Pressure 696, 697, 700, 701, 1851, 6484
ergodic theory 6939, 6942
observable 6945
variational principle 6946
Pressure recovery 6966
Pressure-stimulated current (PSC) 2629, 2632
Pressure-stimulated polarization effect 2623
Prestige 8240
Pretty good measurement 7102, 7106
Prevalence 954, 8316
Preventive 9096
Preventive medicine 9092
Price 2810, 4625, 4626
Price cascades 5416
Price changes, probability densities 3592
Price-contingent orders 5415
Price discovery 5432, 9570, 9573–9575
Price of anarchy 1681
Price wars 7838
Pricing strategies 1968
Primary care trusts (PCTs) 4566
Primary constraint language (PCL) 4000
Primary constraints, fuzzy logic 3999
Primary membership, type-2 fuzzy sets 4013
Primary resonance 6633
Priming 2245
Primitive
for amorphous computing 263
Primordial GWs 1635
Principal-agent 4096, 6977
Principal-agent problems 4856
Principal bundle, connections of 5500
Principal component analysis 1776, 2574
Principal connections 5493, 5500
definition 5502
Principal fiber bundle 2264
Principle
anisotropy 9733
conservation of the vehicles 9729
covariance 3875
Hamiltonian variational 6608
of detailed balance 2809
of systems 8996
of the least action 6600
variational 6600

Prior 448, 450
natural conjugate priors 448
Prior distributions 435, 443, 445, 456
natural conjugate prior 445
'non-informative' prior 445
Prioritized planning 5788
Prisoner's dilemma 181, 1528, 5679, 6645, 7617, 8381, 8382
iterated 8382
n-person 1529
spatial 8382
Privacy 8094
Private information 5427
Pro-activeness 187
Proactive persistent agent (PPA) architecture 1346
Probabilistic algorithms 9960
Probabilistic Boolean network 167, 518, 606
Probabilistic cellular automata 872
Probabilistic constraint, fuzzy logic 3998
Probabilistic features of traffic breakdown 9283, 9321
Probabilistic finite state machine (PFSM) 3696
Probabilistic graphical models 1229
causation 1229
conditional independence 1229
dependence 1229
directed 1229
graph 1229
undirected 1229
Probabilistic models 727, 1227
Probabilistic nature 9379, 9380
Probabilistic networks 585
Probabilistic neural networks (PNN) 2154
Probabilistic product 4067
Probabilistic reasoning 4874
Probabilistic rule 1865
Probabilistic SDCA 8836
Probabilistic seismic hazard analysis (PSHA) 2415, 4435, 4464
ground-motion prediction 4464
microzonation studies 4460
Probabilistic traffic breakdown model 9293
Probability 456, 7822, 9297
ergodic 3017
imprecise 4049, 6933
of a fuzzy event 4073
of passing 9736
of traffic breakdown 9284
of velocity class transitions 9740
presence 3874
statistical inference 1228
uncertainty 1228
Probability amplitudes 7253
Probability content, definition 6306
Probability current 3371, 8772
Probability density 2801
Probability density function 2396, 2399, 2407, 2450, 6992, 7015, 8931
Gaussian 9192
Probability distribution 3317, 3364, 3370, 4822, 5025, 7010, 8721
conditional 3582
fuzzy logic 3994
reduced density matrix 7374
stochastic processes 3579
Probability distribution functions 7601, 8530, 9192
non-Gaussian 8530
Probability forecasts 2571
entropy scores 2572
Kullback–Leibler distance 2572
probability gain 2572
R-score 2572
Probability measure 8716, 8718
Probability monads 6801
Probability space 6300
Probability theory, fuzzy logic 4004
Probit 2748, 2750
Problem
Abelian stabilizer 7095
associativity testing 7108
Banach 8557
Bernstein and Vazirani 7092
black-box 7090, 7091
class number 7096
claw-finding 7098
collision finding 7097
commutativity testing 7108
coset orbit 7112
database search 7097
decomposing Abelian group 7096
Deutsch–Jozsa 7092
Deutsch's 7095

discrete logarithm 7089, 7094, 7095, 7106
dynamic game 2220
Eigenvalue estimation 7094
element distinctness 7098, 7107
factoring 7089, 7091, 7093, 7096
fixed-energy 6604
Gauss sum 7106
generalized hidden shift 7105
generalized Simon's 7095
glued-trees 7108
graph automorphism 7100
graph isomorphism 7100, 7102
ground state 7100, 7114
hidden coset 7105
hidden linear function 7095
hidden linear structure 7104
hidden polynomial 7104
hidden radius 7104
hidden shift 7105
hidden translation 7105, 7113
hiding sublattice 7103
local search 7097
matrix product verification 7108
matrix sum 7103, 7106
maximum-finding 7098
median-estimating 7098
min–max tree 7109
NAND-tree 7108
of output regulation 9080
orbit coset 7105
order finding 7093, 7094
order-finding 7090, 7095
ordered search 7097
Pell's equation 7103
period-finding 7093, 7095
phase estimation 7094
principal ideal 7103
Rokhlin's multiple mixing 8558
Rokhlin's uniform multiplicity 8562
search 7090, 7096
self-shift-equivalence 7095
shift-equivalence 7105
shifted subset 7104
Simon's 7092
state distinguishability 7102
subset sum 7102, 7103
triangle-finding 7108
unit group 7104
Problem solving 5739, 7850
Problem structuring 4484
Process fragments 929
Process, T, T^{-1} 2929
Processes in time 3585
Processing dynamics 2243
Processing GPS 8856
Processor power, agent-based modeling 152
Product diffusion 681
Product launch 1968
Product life cycle concept 681, 1958
Product of a scalar and a fuzzy set 4068
Product of networks 8259
Product-space approach 7487
Production 7847
Production factor 2285
Production game 7625
Production market games 1523
Production mechanisms of relic GWs 1638
Production rules 3210
Production sharing 1660
Production systems 1110
Productivity 680, 1851, 1859, 2285
Products 2972
Profile 4367, 4368, 9943, 9945
similarity 8281
Profit 2818, 7847
Profitability 7842, 7843
Progenitor cells 1192
Program 280, 1542, 9441
CODESSA 2210
CONCORD 2209
DRAGON 2210
logical mobility 5608
mobile agents 5604
MOLCONN 2210
OASIS 2210
POLLY 2210
TOPOCLUJ 2210
TRIPLET 2210
Programmable multitude 929
Programming 3234
distributed 5605
genetic 4136

global state of 5608
localities in 5608
Programming abstraction, amorphous computing 262
Programming paradigm, amorphous computing 259
Projected function 594
Projected mask 594
Projection 44, 4880, 6730
Projection normalization 50
Projection operator
supplementary 1620
technique 1624
Projection-slice-theorem 1720
Projection technique 8589
Projective limit 431
Promise problems 7175
Pronation-supination movements 5730
Proof trees 5202
Proof verification 5612
Propagating structures 6468
Propagating wave 7295
Propagation 879, 1135
crack 1135
Propagation modeling, infrasonic wave 4844
Propagation of chaos 7808, 7809
Propagation operator 880
Propagation speed 6465, 6466
Propagator 8591
estimation 3583
finite time 3581
structure, polymer physics 6814
Proper action, mechanical systems 5486
Proper degeneracy 2264
Proper graph 6845
Properties 8246
Property (T) 3103
Proportional demand game 1678
Proportional-integro-differential (PID) feedback 638
Protein engineering 7036
Protein folding problem 7045
Protein interaction networks (PIN) 3789, 5656
Protein-ligand docking 2122
Protein nanomechanics 7029
Protein networks 5700
Protein–protein interaction networks 1186
Protein–Protein interactions 7516
Protein secondary structure, applications 3240
Protein translocases 7042
Proteins 418, 722, 6014, 6016, 6020, 6021, 6023
as logic signal 258
Proteome 5538
Proteomic 5537
Proto (programming language) 262
Protocol, ODD 157
Prototypicality 2246
Provably convergent cooperative controls 2053
Proximal 9237
Proximal subgradient 8627
Proximity data 8308
Prudent orders 9951
Prudent orders procedure 9956
Pruning
postpruning 1837
prepruning 1836
Pseudo-chaotic systems 990, 995, 1009
Pseudo-extensive 2865
Pseudo-intensive 2865
Pseudo-likelihood 8322, 8328
Pseudo-spectral method 7896
Pseudo-spin 8576
Pseudogas 9807
Pseudolattice 1395, 1403
Pseudopotentials 5751
Pseudorandom number generator (RNG) 5669
Pseudospectrum 7
PSMS *see* Pure stationary Markovian strategy
PSO 8875–8880, 8882
PSPACE 5468, 6388, 6393, 6402
complete 6402
Psycho physics 8939
Psychodrama 8267
Psychological balance 8233
Psychology 3134, 4893
subjective uncertainty 3133
Psychopathology 4900
Psychophysical relaxation 7993
Psychotherapy 7994, 7999, 8001, 8003, 8006
PTWC *see* Richard H. Hagemeyer pacific tsunami warning center
Puberty 1905
Pubeval 4144
Public communication 7627

Public goods 1674
Public health 8307
Public-key encryption 2093, 7254
Public policy 7062, 8989
Public policy problems 7058
Pucci operator 6219, 6221
Pull model of mobility 5606
Pulling geometry 7039
Pulsar magnetospheres 24
Pulsar timing bound 1637
Pulsars 34
Pulsating stars 390
Pulsation 9773
Pulse generator 564
Pumping arguments 806
Pumping lemma 819
Punishments 4102
Purchasing power parity 2285, 5426
Pure shear 5012
Pure shear flow, liquids 5930
Pure state 2869
Pure stationary Markovian strategy 2223
Purification, quantum channel 7391
Push model of mobility 5606
Pushdown stores 805
Pycnocline 6206, 8505
PyDSTOOL 6349
Pyroclastic flow 9803, 9822

Q

Q Factor 9916
q-Fourier 2872
q-Gaussians 2871
q-Independent 2873
q-Product 2872
q-State Potts model 1606
q-Triplet 2873
(q,α)-stable distribution 2874
QAM 7194
QCA
  block-partitioned 7160, 7161
  classically controlled 7161
  computationally universal 7155, 7156, 7158, 7161
  continuous-time 7160, 7161
  Grössing–Zeilinger 7157
  Hamiltonian 7160, 7161
  local-unitary 7159–7161
  nonunitary 7160
  partitioned Watrous 7158
  quantum-dot 7156
  reversible 7159, 7160, 7163
  Watrous 7158, 7161
QCCD *see* Quantum charge-coupled device (QCCD)
QCMA 7189
QIP 7190
QKD *see* Quantum key distribution (QKD)
QMA 7185
  error reduction 7187
QMA-hard 7114
QMAM 7194
QMC 8583
QMIP 7193
QQSPR fundamental equation 7427, 7438, 7439
QQSPR operators 7425, 7459
QRG 7194
QSAR 2114, 2160, 2200, 2210, *see* Quantitative structure activity relationships
  three-dimensional 2209
  two-dimensional 2209
QSAR modeling 1303, 7071
QSAR models
  applicability domains 7078
  consensus 7082
QSPR 2200, 2210, 7424
QSTR 2197
QSZK 7192
Quadratic error 7468
Quadratic error function 7440
Quadratic family 962
Quadratic mean 228
Quadratic return variation
  Black–Scholes model 8786
  definition 8786
  Merton jump–diffusion model 8787
  short definition 8783
Qualia 1425, 8003
Qualitative fuzzy sets 4347
Qualitative modeling 673
Qualitative research 8371, 8372
Qualitative theory 6709
Quality 1973

Quality function 1145
Quality of approximation 7756, 7761, 7769
Quanta 7364
Quantified Boolean Formula 752
Quantifying uncertainty 4465
Quantile regression, semiparametric estimators 5555
Quantitative modeling 5720
Quantitative parameter of the quality of treatment 1626
Quantitative structure activity relationship (QSAR) 2114, 2140, 2160, 7071
Quantitative thinking 4565
Quantitative trait locus (QTL) 9106
Quantum 843, 3859
Quantum advice 7194
Quantum algorithms 7121, 7239, 7399
  Deutsch–Jozsa 7204
  future 7388
  Shor's 7206
Quantum architectures, qubus 7211
Quantum bit *see also* Qubit
  error correction 7243
  quantum computing 7220
Quantum cellular automata 7155
Quantum channel 7332
  capacity 7391, 7392
  dephasing 7332
  depolarizing 7332
  quantum erasure 7332
Quantum chaos 1007, 7164
Quantum charge-coupled device (QCCD) 7223
Quantum circuit 7174, 7177, 7240
  distinguishability 7191
  natural representation 7177
  oracles 7180
  polynomial-time generated 7181
  universal gates 7178
Quantum circuit diagram 7380
Quantum classical correspondence 8927, 8933
Quantum coherence 7387
Quantum communication 7368, 7391
Quantum complexity class 7174
Quantum computer 5748, 7363, 9645, 9705
Quantum computing 1307, 5297, 5774, 7155, 7201, 7218, 7367, 7377, 7379, 9128, 9133
  adiabatic 7214
  cluster state 7212
  continuous variable (CV) 7211
  digital 7202
  error correction 7204, 7205
  one way quantum computer 7212
  topological 7213, 7390
Quantum computing speedup 7119
Quantum corrections 7413
Quantum cryptography 7368, 7395
Quantum data bus 7231
Quantum data compression 7113
Quantum discrete Fourier transform 1022
Quantum dot 1305, 5750, 5913, 7309, 7343, 7344, 9644
  complex 9645
  double 9645
  double quantum dot 7353
  in the quantum 7323
  in the ring 7295
  lateral or planar QD 7310
  linear chains of quantum dots 7355
  self-assembled structures 9645
  stacked 5769
  triple 9646
  vertical QDs 7310
Quantum effects 2091
Quantum electrodynamics 6664
Quantum entanglement 2878
Quantum error-correcting code 7389
  five-qubit code 7336
  nine-qubit code 7331
  QECC 7332
  seven-qubit code 7334
Quantum error correction 7242, 7243
Quantum Fourier transform 7094, 7100–7102, 7105, 7240, 7260, 7382
Quantum gate 1015, 7231
  CNOT gate 7203
  conditional phase shifter gate 1021
  controlled-NOT 1016
  Hadamard 7203
  Hadamard gate H 1015
  NOT gate 1015
  reversible 1015
  Toffoli gate 1016
  universal set of 1015
  universal set of gates 1017

Quantum hacking 7283
attacks 7283
counter-measures 7285
importance of quantum hacking 7285
Quantum holography 3358
Quantum impurity model 7344
Quantum impurity problems 7346
Quantum information 7172
carrier 7224
sending 7391
transmission 7368
Quantum information processing 7361
and control 5922
Quantum information science 7399
Quantum information theory, implication 7396
Quantum interactive proof systems 7174, 7189
parallel repetition 7191
public-coin 7193
Quantum interactive proofs with competing provers 7194
Quantum interference 4194, 4195, 4208, 4209, 7296
Quantum jump detection 7229
Quantum key distribution (QKD) 7267, 7286, 7395
battle-testing QKD systems 7287
calculation of the quantum key capacity 7287
equivalence between phase and polarization encoding 7270
example of a simple eavesdropper attack: intercept-resend attack 7270
ground to satellite QKD 7287
key distribution problem and one-time 7267
multi-party quantum key distribution and entanglement 7287
quantum no-cloning theorem and quantum key distribution (QKD) 7268
quantum repeaters 7286
security proofs with testable assumptions 7287
Quantum Loschmidt echo 7164
Quantum mechanical effects 5747
Quantum mechanical Langevin equation 8932
Quantum mechanics 1552, 3352, 4313, 4317, 5675, 7364, 7369
quantum postulates 4319
Quantum memories 7259
Quantum Monte Carlo 8576, 8582, 8586
Quantum non-demolition 7258
Quantum object set 7430
Quantum objects 7425, 7430
Quantum optics 8940, 8943
Quantum parallelism 7380
Quantum particle systems 6218
Quantum phase transition 7136, 7346, 7357
Quantum point contacts 8608
Quantum potential 3353
Quantum probe 7290
Quantum process tomography 7259
Quantum proof 7174, 7185
Quantum QSPR (QQSPR) fundamental equation 7434
Quantum refereed games 7194
Quantum register, initialization 7226
Quantum repeater 7259
Quantum search 7386
Quantum self-similarity measure 7446
Quantum setting with randomized queries 7130
Quantum similarity 1302, 7424
Quantum similarity Aufbau 7432
Quantum similarity matrix 7426, 7431, 7435, 7461
Quantum similarity measure 7446, 7452
Quantum simulation 7215, 7384
Quantum state distinguishability 7192
Quantum systems 5243
Quantum teleportation 7240, 7256
Quantum transport 7319, 7350
ballistic 7350
charge-ordering regime 7355
conductance quantum 7350
cotunneling 7350
coupled master (or rate) equations 7319
Fano resonance 7353
Meir–Wingreen formula 7351
Mott–Hubbard insulating behavior 7355
overlap matrix element 7319
scattering phase shift 7358
single-particle tunneling rate $\gamma_{L(R)}$ 7319
state occupation factors 7319
transition rates 7319
tunneling rate 7319
Quantum tunneling 9647
Quantum Turing machine 7196
Quantum uncertainties 4318

Quantum walk 7091, 7097, 7098, 7105, 7106, 7113, 7206, 7387
continuous 7107, 7108
Quantum weirdness, quantum information processing 7375
Quantum well 5751, 5774
Quantum wire 7404, 7664, 8608
Quasi-arithmetic means 228
Quasi-attractor 8000, 9233
Quasi-condensate 9693
Quasi-degenerate case 6734
Quasi-finite type 968
Quasi-homogeneous potentials, *n*-body problem 5982
Quasi-invariance 5457
ergodic theory 3069
Quasi-linear equations 38
Quasi-linear system 6721
Quasi-linear utilities 5514
Quasi-long-range order 9684
Quasi-monochromatic waves 7921
Quasi-OWA 230
Quasi-periodic 2264, 4516, 5066, 6626, 6627, 6632, 6637, 9239
Quasi-periodic motion 1934, 5064, 6663
Quasi-periodic oscillations 776
Quasi-periodic solution 5066
Quasi-periodic torus 6322
Quasi-solitons 8480
Quasi-stability 6037
Quasi-weighted mean 229
Quasibound state 7300, 7668
Quasicrystals 6462
Quasidynamic fault models 7859
Quasigroup 2996
Quasiprobability distribution 8933
Quasistatic fault models 7859
Quaternary volcanoes 7924
Quaternion estimate 55
Qubit 1014, 5774, 7202, 7253, 7369, *see also* Quantum bit
Bloch-sphere representation 7227
charge flying qubits 1012
complexity 7119
decoherence-free subspace 7238
hyperfine 7229
hyperfine state 7224
implementation 7225
mapping 7233
metastable state 7224
quantum computing 7220
state, vibrational structure 7232
storing 7223
universal set of gates 1013
Qubus 7211
Queensland 3125, 3137
QueensOD 9454
Quenched random "pinning" force 5026
Quenching 2333, 7521, 7606, 9689
Query answering 1533, 1535, 1536
Query complexity 7119
Query connection graphs 10067
Query language 1533, 1534
Quételet, Adolphe 8380
Quillen–Suslin theorem 595
Quorum sensing 423, 724
Quotient algebra 6289
Quotient by eventuality, definition 6292
Quotient-set, definition 6287
Quotient space 4360, 5736
Quotient space theory 4403
Quotient state (external state) 4348
Quotient structure (QS) 4341, 4348, 4350
QV 8786

R

R-Partition 5461
R&D-process 1974
Rabaul Caldera (Papua New Guinea) 9867
false alarm 9867
stage-2 alert 9867
Rabi-frequency 7229
Rabi-oscillation
ion 7229
quantum computing 7227
Racial and ethnic health disparities 4586
Radar 3832
Radcliffe-Brown, Alfred 8267
Radial basis function 3335, 3344, 3350, 3356
Radial basis function network (RBF network) 1822, 2146, 4872
Radial SLE 8724
Radiated energy 2585

Radiation 3273
Radiation condition 6209
Radiation energy density 3274
Radiation Era 3274
Radiation losses 8480
Radiation models 2587
Radiation pattern 2482, 2487, 2583
Radiative transfer theory 7916
Radio astronomy 22
Radiocarbon dating 9833
Radius 7826
  of gyration, higher moments of 6848
Radon transform 1720
  inverse 9175
Radon, domain 1720
Raghunathan conjecture 3105
Rainbow trout 2145
Raman-transition, Rabi-frequency 7229
Ramified torus bundle 2264
Ramp metering 3952
Randić index 6838
Randić molecular connectivity 2200
Random
  nested 6933
  numbers, independent 7585
Random-access magnetic storage 7835
Random attacks 8263
Random bond 2022
Random chessboard 1402
Random choice in amorphous computing 263
Random close packing 5013
Random cluster model 1607
  partition function 1612
Random coefficient models 5568
  quasi-random numbers 5569
  random coefficient logit 5568
  simulation-assisted estimation 5569
Random coefficients 2788
Random conductance 7529
Random curve 8721, 8724
Random effects 2774
Random effects models 5570
  numerical integration 5566
  random intercept model 5566
Random field 2022, 3475, 3484, 3485, 3648
Random field Ising model (RFIM) 5023
Random first order theory 4223
Random force 2023, 5729
Random forest 2165, 5322
Random forests method 4091
Random fractal 3880, 3882, 7572
Random fractal curve 8713
Random fuzzy set 8683
Random graph 2531, 2534, 2837, 6050, 6787, 7823, 8214, 8261, 8276
  Erdős and Rény 6050
  network 8911
Random homogeneous mixing 4660, 4661
Random inhomogeneity 7915
Random manifold 2019, 2023
Random map 8721
Random matrices 7025, 7505
  orthogonal ensembles 7505
  symplectic ensembles 7505
  unitary ensembles 7505
Random matrix theory 7164, 7505
Random midpoint displacement method 3774
Random multiplicative process 3813
Random network 503
  models 8266
Random noise 5722
Random percolation 1598
Random-phase approximation 8576, 8581
Random potential 2023, 6409, 6414, 6416
Random power law graphs 7500
Random proportionate growth 1053
Random resistor network 1398, 1405
Random search 2091
Random set, fuzzy 4000
Random spanning tree 8727
Random stress drop 5028
Random subspace method 5322
Random tree 2165
Random trend model 2777
Random utility model 5555
  conditional choice probabilities 5556
  conditional logit 5556
  multinomial logit 5556
  multinomial probit 5556
  nested logit 5556
Random variable 3317
Random void 1444
Random W-cascades 3900, 3903

Random walk 311, 884, 1396, 3706, 3749, 3760, 3805, 3884, 4907, 5027, 5670, 5733, 7025, 7521, 7568, 7569, 8710, 8713, 8768
betweenness 1150
biological fractals 3785
boundaries 3077
escape probability 3751
exit probability 3749
hypothesis 3438
in disordered media 8591
on random walk 314
Pólya's Theorem 3750
polymer physics 6813
recurrence 3077, 3751
transience and recurrence 3750
with non-stationary increments 3077
Random walker, definition 6308
Randomized network, motif information 508
Randomized Turing machine 5469
Randomly branched chains 6844
Randomness 4072
Randomness space 254
Randomness test 255
Range identification 4772, 4773
Range of interactions 5035
Range voting 9954
Rank 703
contingency table (2×2) 1866
contingency table ($m \times n$) 1867
definition 6308
matrix 1866
subdeterminant 1868
submatrix 1868
table 1866
Rank and degree dependence 1868
Rank-based tests 6254
Rank correlation 6255
Rank-one attractors 964
Rank-one transformations 3061
Rank-ordered multifractal analysis (ROMA) 8541
Rank-size distribution of cities 7957
Rank-size rule 180, 1051, 1053
Rank tests 6255
Ranked pairs procedure 9952, 9956
Rankine–Hugoniot equation 4717
Ranking 7753, 7754, 7765, 8240
Rao-Blackwellization 8221, 8230
Rapid burster 392
Rappture 5761
Rare event 6996, 7002
Ratchet 1754, 1755, 1758, 1760, 1768, 1771, 1772
Ratchet effect 6985
Rate constant 1312
Rate equations 728, 1314, 6128
elementary steps 1314
Rate function 5999, 6002, 6009, 7523
Rate-limiting step 1316
Rate models 6100
Rate of reaction 556
Rate parameter 8308
Rate, firing 6100
Rate/state variable friction law 657
Ratings summation 9954
Rational bubbles 3413
Rational choice 8188
Rational expectation 209, 4972
Rational map
isomorphism theory 4992
of the Riemann sphere 4985
Rationality 1520, 8381, 8662
coalitional 1521
group 1520
individual 1520
Rationalizable strategies 1590
Rationing method 1675
random priority 1676
uniform gains 1676
Rationing, problem 1675
Ratner's, property 8569
Rattlers 2834
Raussendorf 7257
Rawlsian maximin criterion 3332
Ray tracing 5472, 5473
Rayleigh–Bénard convection 6195, 6198, 6199, 6202
Rayleigh–Bénard experiment
stochastic processes 3575
Rayleigh–Bénard system 3647
Rayleigh number 3626, 6195, 6198–6200
Rayleigh oscillator 5725
Rayleigh–Plesset equation 6959
Rayleigh–Ritz theorem 59
Rayleigh–Schrödinger 6726
RS perturbation theory 6726

Rayleigh's monotonicity principle 1406
RBF kernel 5355
RCP 5014, *see* Rotating circle property
RD chip 9708, 9709
RDF (resource description framework) 8085
RDF/XML syntax 8092
RDF(S) (RDF schema) 8085
RDS 6469, 6471
Re-focusing
  quantum information processing 7389
Reachability 761, 8236
>-Reachable 6036
Reacting matching pairs 889
Reaction center 1322–1324
  agostic bond 1324
  Local chemical composition 1326
Reaction channel 8733
Reaction constant 885
Reaction-diffusion 525, 536–540, 543, 545, 1056, 7579
  Nieuwkoop center 538
Reaction-diffusion chemical computer 5475
Reaction-diffusion computing 278, 284, 3348, 3357
Reaction-diffusion equation 284, 948, 6230, 6465
Reaction-diffusion process 885
Reaction-diffusion system (RDS) 118, 647, 3365, 4921, 6195, 6204, 6461, 9708
Reaction front 885
Reaction kinetics 7565
  in fractals 7566
Reaction-limited 3368
Reaction-limited CCA (RLCA) 3884
Reaction network 328
  as example for artificial chemistry 327
Reaction order 329, 7571
Reaction propensities 8733
Reaction rate 885
Reaction rule as building block for AC 328
Reaction site 1326
  agostic bond 1324
  Local chemical composition 1326
Reaction zones 3368
Reactive agents 188
Reactive approaches 5792
Reactive oxygen species (ROS) 483
  oxidative stress 483
Reactivity 187
Real biological fractals 3788
Real estate 3126
Real Ginzburg–Landau equation 395
Real networks 1202
Real options 7839
Real-time 3243
Real-time monitoring 8003
Real-time seismology 2511
Real typologies 8996
Real-valued Turing machine 5469
Realism 9062, 9064
Realization
  input-output map 6375
  paracompact manifold 6376
  state space 8757
Realization theory 6371, 6375
Realized bipower variation, estimation of the squard tump quadratic variation component 8789
Realized volatility
  definition 8787
  discretization bias 8788
  estimation and specification of SV models 8808
  ex-past vs. ex-ante measures of volatility 8789
  Heterogeneaus auto regressive model 8789
  high-low price range estimator 8789
  long-memory 8789
  market microstructure noise 8788
  model-free ex-ante forecasts of volatility 8789
  short definition 8783
  unbaised and consistent estimater of the return quadratic variation 8788
Rearrangement, matrix 8268
Rebound release 2653
Recapitulation theory 1918
Receptive field 3335, 3344, 3345, 3347, 3351
Receptors 724
Recession 2746, 2748, 2752, 2754, 3301
  economic 3307
  end 3308
  start 3308
Recharge 2612, 2613
Reciprocal 7467
Reciprocal causality 1538
Reciprocity 6004, 6007, 8324, 10061

Recognition 10032
of faces 8938
Recognition memory 2315
Recombinant DNA 418
polymer physics 6814
Recombinant inbred (RI) 9118
Recommendation 1589
Reconfigurable device 930
Reconfiguration
off-line 5872
on-line 5872
Reconfiguration algorithm 5618
Reconnection 9273
Reconstruction 5535, 5538–5540, 5549
or knowledge base 5538
Record dynamics 7025, 7583, 7589
intermittent dynamics 7025
nested hierarchical 7587
subordinated 7589
Record statistics 7583
Recording damaging ground shaking 2511
Recovery 3114
Recruitment
discontinuous 1557
discrete 1557
Rectified diffusion 6956, 6958, 6964, 6969
Rectified heat transfer 6964, 6972
Recurrence 3041, 7522, 8488, 9225
Recurrence plot 8000, 8003
Recurrence theorem 2935
Recurrent 5460
Recurrent neural network 288, 1822
Recurrent random walk 1396
Recursion coefficients, multiwavelet 5844
Recursion relation, multiwavelet 5844
Recursive 3546
Recursive decomposition 8272
Recursive Fourier sampling 7092
Recursive function 281, 287
Recursive Greens function 8603, 8605, 8608
Recursive neural networks 2149
Recursive partitioning (RP) 2166
Recursive state 8036
Recursive structure 7626
Recursive zero-sum game 4862
Recursively computable 5468
Recursively generated objects 119
Red bonds 3705, 3709, 3853, 3854
Red digital camera 10006
Redistribution of energy levels 7149
RED_{PCT} 7769
RED_S 7772, 7773
Reduced density matrix, quantum mechanics 7374
Reduced distribution functions 994, 995, 998
Reduced Euler–Lagrange equations 5495
Reduced form 214, 5179
Reduced graph 6245
Reduced phase space 2264
Reduced space 5483
symplectic 5486
Reduced system 6311
Reducibility 5081
Reduct 1793, 1794, 4417, 7756, 7761, 7769, 7793, 8338
Reduction 6311, 6316, 6320, 6761, 7137, 8248, 9105
at zero 5500
Lagrangian 5495
Liapunov–Schmidt 6609
multisymplectic 5498
nonequivariant 5490
of dynamics 5488
of geometric structures 5497
semidirect product 5496
singular 5497
Reduction lemma 5488
Reduction principle in stability theory 945
Reduction process 9412
Reduction theorem 942
remarks on 5489
symplectic 5487
Reduction theory
definition 5483
historical overview 5492
symplectic 5483, 5484
Reductionism 202, 1538
Redundancy 1720, 5871
cascaded triple modular (CTMR) 5873
hardware 5872
information 5872
R-fold modular 5873
time 5872
triple modular (TMR) 5873

Redundant bond 2830, 2835, 2837
Redundant edge 6014, 6016
Redundant feature 5349
Redundant manipulators 1476
Reed–Muller expansion 9720
Reentry 1427
Reference frame 3860, 6426
Reference mode 9044
Reference plane 58
Reference set 4061
Referential 4061
Referential meaning 1902
Refinable function 592, 6864
  vector 5843
Refinement equation 592
  multiwavelet 5844
Reflection coefficient 4966
Reflection symmetry 6633
Reflective judgment 1893
Reflex 1902, 6069
Refraction diagram 9632, 9636
Refractory period 1354
Regeneration times 7524
Regenerative amplifier 6124
Regenerative gain 6125
Régime 1550
Regime shifts 3
Regime switching 3443, 7021
Regime-switching models 2745–2747, 2755
Regiment (programming language) 262
Region
  closed 6607
  compact 6605, 6606
  Hill 6604, 6605, 6607
  nonsimply connected 6607
  with boundary 6606
  without boundary 6605
Regional catalogues 2565
  declustering 2565
  residuals 2566
Regional climate models 7592
Registration 10032
Regression matrix 57
Regular 703
Regular domain 3233, 3237
Regular equivalence 8289
Regular language 5634
Regular lattice 3847
Regular perturbations 6738
Regular point
  reduction theorem 5489
  symmetry algebra 5487
Regular potential, Gibbs property 6947
Regular semigroups 8297
Regularity rules 2794
Regularized long-wave (RLW) equation 8116
Regularly varying 7012
Regulation 719, 5407, 9105
  of gene expression 719
Reif 5472, 5473
Reinforcement learning 1816, 5195
  policy 5202
  punishments 5202
  Q-learning 5202
  rewards 5202
Reinforcing effect 1977
Relation 9228
  Compton 3874
  definition 6301
  median 9950
  transitive 9229
Relation instance 4314
Relational data model (RDM) 4305
  foreign key 4307
  primary key 4306
  relation 4306
  tuples 4306
Relational databases (RDBs) 8089
Relational feature selection 5358
Relational granulation 8339
Relational GrC model, fifth GrC model 4346
Relative coordination 1541
Relative dispersional analysis 479
  long-term memory 484
Relative entropy 6941
  ergodic theory 6948
Relative equilibria 2264, 4535
  n-body problem 5961
Relative intensity (RI) 2441
Relative location SDCA model 8830
Relative majority 9947
Relative permeability, porous medium 4957
Relative phase 1553, 5720, 8200, 8201, 8204
Relative reducts 1795

Relatively finitely determined factor
isomorphism theory 4993
Relatively very weak Bernoulli factor
isomorphism theory 4993
Relativistic shocks 27
Relativity 3859
doubly-special 3869
scales 3863
special scale 3868
Relaxation 2694, 9732
Relaxation methods 3745
Relaxation modulus, single chain dynamics 6821
Relaxation oscillation 566, 6129, 6156
Relaxation phenomena 5672
Relaxation processes 5673
Relaxation time 6847, 6959, 9732, 9790
biological fractals 3794
Relaxation time approximations 5915
Relaxation time spectrum 6546, 6553, 6554
Relaxational dynamics 3371
Relay systems 6698
Relevant and irrelevant nodes 600
Relevant parameters 8544
Reliability 4186, 8067
algorithm 4186
Relication origins 3908
ReliefF 5353
Religious cults 3216
RELM test 2444
Remainder 6659
Remaining benefit 1664
Remote computation, difference to mobile computing 5605
Remote evaluation 5607
Remote ion–ion entanglement, quantum computing 7244
Remote method invocation (RMI) 5607
Remote visualization 2363
Remotely operated vehicles (ROVs) 3688
Remoteness Δ 9950
Remove the degeneracy 2265
Renegotiation 4798
Renewal models 2563
Renewal process 9837
gamma 9838
Gaussian 9838
likelihood 9837
log-logistic 9838
lognormal 9838
mixture of Weibull 9839
power-law 9839
tests 9837
Weibull 9838
Renormalization 1396, 1409, 2023, 3847, 3925, 6233, 6666, 7829
polymer physics 6811
Renormalization group (RG) 1134, 1270, 1932, 5021, 5024, 6663, 6669, 6671, 7020, 7344, 8388, 8713, 9680
cusp singularity 1134, 1135
flow 1134
functional 1134
Renormalization group (RG) methods 7347
coupling constants 7347
fixed point 7348
numerical renormalization group 7347
renormalization flow 7349
running parameter 7347
scaling 7347
Renormalization group (RG) technique 8711
Renormalization theory 6664, 6671
Rényi, Alfréd 8277
Repast, simulation tool 194
Repeated game 1594, 2216, 4096, 7651
complete information repeated game 7655
finitely 7618, 7620
imperfect monitoring 7654
incomplete information repeated game 7656
model 2216
no discounting 7618, 7619
perfect monitoring 7653
with discounting 7618, 7619
with imperfect monitoring 1594
with incomplete information 1594
Repeated prisoners' dilemma 4100
Repeller 989, 1001, 5721
Repellor 8423, 9225, 9233
Repetition 2244
Repetitive DNA 4164
recombination 4167
Replica symmetry breaking 4224
Replica theory 2020
Replication 3894, 5612
Replication domains 3913, 3915, 3917

Replication origin 3904, 3910, 3913–3916, 3918
Replication terminations 3913
Replication timing 3913, 3914, 3916
Replicator 3189, 3208
Replicator dynamics 1595, 5678, 5683
Replicator equation 2908, 6643
Replicon 3908, 3909
Replicon model 3914
Repose time 9831
Representation 4895, 6736
  CAR 3078
  equivalent 6736
  inequivalent 6736
  irreducible 6736
  Koopman 8555
  Maruyama 8567
  measure-preserving 8557
  of the center manifold 939
  reducible 6736
  unitary 8556
Representation theorem (RT) 4014, 4065
Representation theory 7100, 8935
Representationalism 1111
  minimum robust representationalism 1109
Representationww system 244
Representative agent 205, 6807
Representative chromosome 4154
Representativeness 7850
Repressilator 563
Reprojection method 8789
Repulsion between hubs 3934
Reputation 8135
Reputation bound 7657
Reputation effects 4096, 7651
Required number of large 1-itemsets (RNL) 4155
Required reserve ratio 2807
Requisite complexity 8370
Requisite variety 8370
Rescaled-range analysis 3760
Réseau-Lu 1256
Reservoir performance 6568
Residual degeneracy 6735
Resilience 1881, 3935, 8040
Resilient propagation (Rprop) 1820
Resistance 3854, 3936
  to change 7841
Resistor network 3738, 6546, 6559
  conductance exponent 3744
  Kirchhoff's law 3740
  maximum voltage 3747
  optimal defects 3748
  random 3739
Resolution 1697, 3860
  identity 4065
  matrix 9190
Resonance 17, 21, 2265, 4516, 5076, 6102, 6312, 6313, 6316, 6322, 6461, 6462, 6474, 6613, 6628, 6655, 7663, 8635
  sporadic resonance 6188
Resonance broadening 8529
Resonance condition 6591
Resonance line form function 8594
Resonance of a fluid-filled resonator 9900
Resonance overlap 384, 408
Resonance phenomenon 7290, 9994
Resonance relation 6179, 6188, 8641
Resonance tongue interaction 2338
Resonance tongues 2327, 6630
Resonance-type effect 9981
Resonant 6437
Resonant acceleration 37
Resonant (critical) phenomenon 9981
Resonant frequency 6959, 6962
Resonant monomial 6761
Resonant motion 6670
Resonant normal form 6715
Resonant perturbation theory 6678
Resonant scalar monomial 6188
Resonant scattering 23, 31
Resonant tunneling 7405, 8611
Resonant zone 5988
Resonator
  eigenmodes 6124
  quality factor, Q 6124
Resources 1978
  corporate 7840
  space-like 6398
  usage 6401
Respondent driven sampling 8214, 8219
Respondent (or classical) conditioning 1886
Response complexity 4106
Response functional 3370, 3372
Response to influences, agent actions 188

Rest point 8634, 8635
nondegenerate rest point 8635
Resting tremor 627
Restoration 10032
Restricted orbit equivalence, isomorphism theory 4992
Restricted three-body problem 6428, 6438
Restriction 8710, 8724
enzymes 2098
property 8724
Result
dynamic game 2218
stochastic game 2218
Resummation 1941, 6664, 6669, 6671
resummed series 1942, 6671
Resynchronization 637
Retail trading 5409
RETAIN 1804
Retard diffusion 8316
Reticular brainstem neurons 7994
Reticular rhythm 7994
Retrieval of cases 1804
RETRIEVE 1804
Retrograde evolution 5663
Retroviruses 9751
Return autocorrelation 103
Return interval 3714, 3716, 3718, 3771
Return map 776, 8632, 8638
Return method 1506, 3530
controllability 4816
Return predictability 3449
Return process and no-arbitrage condition, canoncial decomposition for special semi-martingales 8786
Return time map, isomorphism theory 4991
Return times
ergodic probability measure 3021
Returns 5420
on investment 1978
REUSE 1804
Reusing 1805
Reuters 5409
REV (representative elementary volume) 2678
Revelation game 9656
Revelation principle 4792, 6984
Revenue equivalence theorem 6986
Reverse auction 2085
Reverse fault 9614
Reversibility 64, 70, 820, 880, 1885, 1903, 7679, 7696, 8033
Reversible 6318, 6325, 6628
Reversible CA 837
Reversible computing 7089, 7090, 7366, 7680, 7696
Reversible logic gate 9130
Reversible logic, quantum computation 7378
Reversible SDCA 8833
Reversible systems 2330
Reversible Turing machine 5469
Reversing 4188
REVISE 1804
Revision protocol 3176, 3181, 3182
Revisionist economic models 2285
Rewiring 5659
Rework 1848
Rework cycle 1846, 1851
Rewriting systems 905
Reynolds equation 3650
Reynolds number 3643, 5576, 6195–6197
Reynolds stress 3650
Rheology 5017, 6959, 9771, 9777, 9785
Bingham 9777
Newtonian 9774, 9777
non-Newtonian 9774
of gels 6552
of supercooled liquids 4237
ρ-Hyperbolic 946
Rhyolite 9796
Rhythm 6099, 6107, 6109, 6110, 8200
Riccati equation 1505, 6617, 8620
Richness 5183
Ridge 1722
function 1722
Ridgelet transform 1719
fast slant stack 1723
local 1724
orthonormal 1722
rectopolar 1720
Ridgelets 1719
Riemann invariants 9986
Riemann–Liouville operator 5220
Riemann mapping theorem 8710, 8717–8719, 8721, 8724
Riemann problem 4717, 9743

Riemann solver 9743
Riemann sum, definition 6299
Riemann theorem 8720
Riesz M-wavelet basis 590
Riesz wavelets 597
RIF (rule interchange form) 8085
Rigged Hilbert space 989, 1004
Right closing factor code 8895
Right-equivalent 703, 709
Right invariant system, input-output mapping 6377
Right multiplication table 8295, 8300, 8301
Right resolving factor code 8895
Rights, property 4624
Rigid cluster 2836, 2839
Rigid region decomposition 6014, 6018, 6020
Rigidity 9585, 9776
  in gels 2833
  polymer physics 6815
  problem 2838
  transition 2830
Rigidity percolation 6546, 6557
Ring 703, 778, 7673
  Artinian 790
  commutative 779
  directly finite 779, 789
  group 779, 788
  oscillator 563
  stably finite 779, 789
  systems 6445
  zero divisor in a 779, 788, 790
4-Ring 8258
RISC 5883
Rising 4188
Risk 3110
  neoclassical economics 178
Risk analysis 1860
Risk assessment 1848, 1860
Risk aversion 3470
  time-varying 3471–3473
Risk aversity 3328
Risk communication 3110
Risk dominance 3195
Risk dominant equilibrium 3196, 3197
Risk estimation 3718
Risk management 7010
Risken 8932
Risky assets 5400
Rivalry, binocular 6108
River restoration 9021
Rivers 3700
RLP 5014
RLW equation 87
RMI *see* Remote method invocation
RNA 418, 2098, 4172
  editing 4171
  hnRNA 4169
  interference 3971
  miRNA 4172
  mRNA 4171, 4172
  mRNA (messenger RNA) 4169
  nonsense-mediated decay 4171
  piRNA 4173
  siRNA 4173
  snRNA 4169
  world 3286, 3291, 3298
Road pricing 9460
Road theorem 8901
Road traffic 5440
Roadmaps 5785
Roadwork planning 9526
Robbins–Monro algorithm 6007
Roberts theorem 5518
Robocode 4139, 4143
Robot control 9705
Robot design 2074, 4647
Robot evaluation 4650
Robot-room 1480
Robot sensors 4646
Robotic networks 7720, 7725, 9105
Robotic wall 1481
Robotics 1110, 1346, 1897, 2073, 8872, 8873
  and amorphous computing 269
  assistive 4648
  developmental 4648
  educational 4649
  epigenetic 4648
  humanoid robots 1109, 1121
  modular 5621
  socially assistive robotics 4648
Robots 1110, 1897, 4644
  autonomous 6067
  self-reconfigurable 5619
Robust 2846, 2848

Robust control 45
Robustness 565, 731, 3935, 4676, 4725
  resilience 6058
  vulnerability 6058
Robustness under incomplete information 4800
ROC diagram 2445
ROC graph 1998
Rock 6505, 9874
  discrete discontinuities 6505
Rock–Paper–Scissors games 3179
Rod networks 2833
Rodent homeodomain, entropic profile 500
Rogers, Everett 8272
Rogue advertising profiles 4374, 4375
Rogue trader risk 5410
Rokhlin's lemma 2975, 3058
Rokhlin's theorem 4986
Role 8289
Role algebra 8273, 8292, 8295, 8297, 8302
Role autonomy 8195
Rolle differentiation 6298
Romania 2399
Romanovsky theorem 6243
Root-mean-power (RMP) 228
Root-mean-square 228
Rosenau, P. 88
Rotary element 7702
Rotated l-shaped mapping 3562
Rotating circle property (RCP) 5964
Rotating spiral pattern 9718
Rotating Stäckel potential 403
Rotation
  of a compact group 2958
  quantum mechanics 7372
Rotation effects 2636
  seismic 2637
Rotation matrix 1017
Rotation motions, spin and twist 2641
Rotation number 1935, 3021
  best approximant 1939
Rotational dynamics 387
Rotational ground motions 2524
Rotational motion 6446
Rotational seismology 2523
Rothblum index theorem 6246
Rothemund 5479
Rotor 6634, 6636
Rotor dynamics 6634
Rough binary relation 7767
Rough entropy 7732
Rough functional dependency 7736
Rough inclusion 1797, 4413
Rough relation entropy 7733
Rough relational database 7730
  indiscernibility 7730
Rough schema entropy 7733
Rough set 4403, 7753, 7788, 7790, 8338
  contingency table 1865
Rough set system 4358
Rough set theory 1791, 1792, 4314, 4355, 4412, 7729, 7755
  classical 7755
Rough stress field 5029
Roughening exponent 2024
Roughening transition 3372
Roughly-redundant 7734
Roughness 3841, 7732, 7790
Roughness exponent 2022
Roulette wheel selection 4874
Round numbers 5415
Round-off error 8993
Rouse-chain motion 6824
Rouse dynamics 6821
Rouse local-dissipation theory 6812
Rouse matrix 6847
Rouse theory 6847
Rouse time, single chain dynamics 6820
Rousseau 1883
Route guidance 3955
Routes to chaos 3647
Routh–Hurwitz criterion 8636
Routh reduction, nonabelian 5496
Routh, work in reduction theory 5492
Routhian 5496
RSA public-key cryptosystem 7382
RSES package 1799
RT *see* Representation theorem
Rubber elasticity theory, entangled polymer dynamics 6821
Rubber elasticity, polymer physics 6813
Rubel 277, 286, 3357
Ruderman–Kittel–Katsuya–Yoshida (RKKY) interaction exchange 7353, 9651
Rudolph structure theorem 4988

Rudolph's counterexample machine 4989
Ruelle inequality 960
Ruelle–Pollicott resonances 1003, 1004
Ruelle's Perron–Frobenius theorem 6948
Ruelle's pressure formula 3019
Rugged energy landscape 1129
Rule
  Leibniz 3872
  probabilistic 1865
Rule 90 66, 72
Rule based systems 1343
Rule component 65
Rule discovery method 1865
Rule induction 1993, 7798
Rulebases
  fuzzy 4080
  type 1 fuzzy logic 4083
Rules supernetworks 6035
Run-length-limited shift spaces 8902
Runaway 5032
Runge–Kutta 8697
Running couplings 6665, 6671
Rupture 3622, 3639
Rupture area 2483
Rupture duration 2479, 2482, 2487, 2492
  automatic estimates 2491
Rupture evolution 7860
Rupture length 2483, 2485
Rupture model 2485
Rupture process 2482, 2489, 9581
Rupture slowness estimation, tsunami warning 2669
Rupture velocity 2485, 2486, 2488, 2593
  earthquake 2539
Russell, John Scott 81, 1376, 6138, 8113
Rüssmann 5079
  non-degeneracy condition 5081
RV 8787
RWRE 7521
Rényi, Alfréd 2867

S

S-Bots 3692
S-Coda waves 7915
S-Matrix 7672
s-Norm 4067
s-*t*-Norm composition of two fuzzy relations 4069
S-Unit theorem 3049
S wave 2476, 2478
S-wave 2664
Sabatier principle 1314
SAC 2405
Saccharomyces cerevisiae 483, 9108
  continuous culture 483
SAD seasonal affective disorder 3457
Saddle 2265, 4522
Saddle-node 6633
Saddle-node or limit point bifurcation 4536, 5304, 5728
Saddle point 1312, 4855, 6623
Saddle point criteria 4858
Saddle point theorem 6623
Saddle potential, Paul trap 7221
Saddlepack 2077
Safe languages 5611
Safe shelter 3134
Safety learning 5203
Saidani 8842
Sales force 7846, 7848
Salmon population 9015
SALSA 5275
Sample selection 5557
  censoring 5553
  self-selection 5557
  truncation 5557
  two-step Heckman procedure 5558
Sample-spanning cluster 6542, 6544
Sampling
  critical 1721
  design 8213, 8215, 8216
  sample size 1990
  theorem 1721
Samy worm 4369
San Andreas fault 2612
San Fernando basin 9194
  P wave velocity model 9194, 9195
  SCEC model 9194
San Fernando earthquake 9194
Sandbody 6570
Sandbox method 3709
Sandbox security model, Java 5611
Sandboxing 5611
Sandpile 8029

Sandpile cities 7968
Sandpile model 1135
SANS *see* Small angle neutron scattering (SANS)
Santa Fe approach 212
Santorini 9799
SAR 2114, *see* Socially assistive robotics
SAR research 4649
Sarcolemmal K_{ATP} channel 480
SAT-property, random walk 3077
Satellite altimeters 9635
Satellite positioning 4251
  GPS signal 4252, 4277
  integer ambiguity 4252
Satisfaction relation 4394
Satisfiability problem 2094
Saturable absorber 6137
Saturable absorption 6139
Saturation 2474, 2477, 2482, 2484
  principle, non-standard analysis 6305
  spectral saturation 2484
  time-window 2484
Saving propensity 2808
Saw-tooth pattern 7033
Scalability 1141
Scalar solitons 1376
Scale 3860, 6669
  kinetic 9728
  macroscopic 9728
  microscopic 9728
  minimal 3869
  of representation 9727
  Planck 3860, 3869
  stochastic processes 3589
Scale coupling 5917
Scale disorder 3575
Scale-free 2875
Scale-free distribution 8277
Scale-free law 4835
Scale-free model 6049, 6051
Scale-free networks 570, 575, 1050, 2715, 3856, 6048, 6050, 6496, 6787, 8263, 8912
  Barabási and Albert 6051
  heterogeneous 6050
  hubs 6050
Scale-free property 3924
Scale-free structure 3214
Scale index 6671
Scale invariance 725, 1682, 3700, 3802, 3879, 7012, 7568, 8710, 8712
Scale-invariant quantity 5022
Scale separation 4906
Scale transformation 312
Scalegram, wavelet power spectrum 378
Scaling 271, 277, 282, 320, 476, 1049, 1050, 1126, 1642–1644, 1646, 1648, 1657, 2019, 2821, 3865, 6410, 6608, 6777, 7826, 8266, 8268, 8712, 8713, 8718, 9692, 9785
  agent-based modeling 152
  in equilibrium 8712
  linear 1644, 1648
  scale free 478, 486
  self-similarity 478
Scaling behaviors 603
Scaling dimension 2024
Scaling exponent 3755, 3759, 3768, 6998, 9687
  Rényi 3768
Scaling factor 6669
Scaling function 1721, 6862, 6998, 8713
Scaling idea 8712
Scaling laws 2442, 2581, 3755, 5028, 6543, 6664, 8713
  near the gel point 6562
  of earthquakes 2581
  of seismic spectrum 2588
Scaling limit 6526, 8710, 8714
Scaling methods 8266
Scaling models 8274
Scaling relations 6773
  scaling laws 6777
Scaling result 5028
Scaling structure, polymer physics 6818
Scaling universality class 8711, 8716
Scalloping 1898
Scalogram 2672
Scar 1008, 7167
Scarabotti F 780, 784
Scarce resource 6492
Scattered data fitting
  discrete least squares fitting 5800, 5817, 5818
  L_1 spline method 5817
  L_1 spline methods 5822
  minimal energy method 5817
  penalized least squares fitting 5817, 5821

Scattering 4962, 6740, 7663
amplitude 6741, 6742
coefficient 4966
data 4960, 4962, 4966, 4968, 4969
length 8593, 9681
matrix 8600, 8608, 8609
potential 6741
slab 7924
solutions 6741
SCD *see* Social coordination dynamics (SCD)
Scenario 235, 240, 1859, 1860, 1970
base-run 7843
compound 240
computed 7832, 7848, 7850
decision 7836, 7839
diagrams 7044
environmental 7839
growth 7842
homogeneous 240
mixed environmental 7836
pollution 7845
risk 1860
tourism growth 7843
Scenario-driven planning 7849
Schaeffer 704
Schawlow–Townes limit 6143
Scheduling 2093
Schelling 6805
Schelling model 1061, 6381
Schema 4136
Schizophrenia 4900, 7992, 7994
Schlafkrankheit, effects of climate change 1101
Schmeling–Troubetzkoy theorem 3024
Schmitz–Schreiber method 3774
Schnirelman's theorem 1008
Schramm Loewner evolution 8710
Schramm's theorem 8710, 8713, 8721
Schrödinger cat state 7242
Schrödinger equation 1309, 2041, 2042, 4961, 4963, 6689, 7403
nonlinear 2050
Schrödinger operators 5081
Schrödinger picture 8927
Schur's lemma 6736
Schwartz, Anna 2793
Schwartz's solution 9953
Schwarzian derivative 3839
Science gateways 5763
Scientific progress, fuzzy logic 3988
Scientific theories, fuzzy logic 4003
Scientific thinking 4565
Score system 9954
Score vector 9947, 9960
Scoring procedures 9947, 9955
Scott, John 8269
Scout 2000 2075
Scrambled set 9237
Screening 8127
Screening cloud 9683
Screening length 8579
polymer physics 6812, 6817
Scroll wave 6471
SDCA and genetic algorithms 8832
SDCA with memory 8835
Sea ice
ice dynamics 1708
thermodynamic sea ice model 1709
Seafloor deformation 9633
Seamount 9587
Search and rescue operations 1476, 3683, 3698
Search control knowledge 5190
Search-control rule 5197
Search engine 10062
Search engine bias 10066
Search space 1835, 4135
Searching on complex networks 6061
Seasonal affective disorder (SAD) 3469, 3470
clinical incidence 3471
onset/recovery 3471
Seasonality 7843
tourism 7843
Second laser threshold 6134
Second law 4897
Second law of thermodynamics 3084, 3270
Second-order phase transition 3851, 5027, 6772, 6776, 8711
Secondary extinction, network fragmentation 2735
Secondary membership function
fuzzy system models 4088
Section
Poincaré 6599
Sector 2285
Secular equation 7437

Secular frequency, quantum computing 7221
Secular term 6212
Secure coprocessors 5612
Security proofs
  approaches to security proof 7271
  classical post-processing protocol 7273
  classification of eavesdropping attacks 7270
  composability 7274
  security proofs of practical QKD systems 7274
Sediments 9586, 9588
Seed adopters 8315
Segment 6841
Segmentability 1527
Segmentation 534, 536, 540–544, 10032
Segregation 1061, 1062, 2819, 3368, 6483, 7572
Segregation model 93, 210, 211
Seismic 2450
  phase 2450
  stations 2452
  travel-times 2454
  velocity model 2453
Seismic albedo 7918
Seismic coupling 9588
Seismic cycle 1219, 1690, 7863
Seismic efficiency 2585
Seismic energy 2479, 2483
Seismic gaps 1695
Seismic hazard 2396, 2438, 7853
  real-time 2396
Seismic inversion 9585
Seismic methods, earthquake source 2662
Seismic moment 2396, 2478, 2481, 2582, 2650, 5021, 9595, 9633, 10055
  scalar 2482, 2485
  small earthquakes 2546
Seismic monitoring of magmatic and hydrothermal activity 9925
Seismic network 2501
Seismic nucleation phase 2546
Seismic phases
  P-waves 2397
  S-waves 2397
  surface waves 2397
Seismic phenomenon 1626
Seismic potency 5021, 5022
Seismic sensors
  accelerometer 2397
  seismometer 2396
Seismic stations
  broadband 2400
  short period 2400
  strong motion 2400
Seismic tomography 2692, 9172
  arrival-time 9179
  arrival-time tomography 9172
  checkerboard model 9191
  checkerboard test 9199
  comparison with X-ray tomography 9179
  decoupling of the earthquake location and tomography problems 9183
  examples 9194
  Fermat's principle 9182
  local slowness tomography 9182
  local velocity tomography 9181
  solution iteratively 9190
  solution roughness 9190
  teleseismic tomography 9186
Seismic wave propagation 1215
Seismic waves 1215
Seismicity 1216
Seismicity as a critical phenomenon 2625
Seismicity parameters 2361
Seismicity patterns 7854
SeismNet manager 2403
Seismo-electromagnetics 2622
Seismogenic zone 10051–10056
Seismogram 9877
Seismograph 2475, 2477, 2481, 2500
  broadband seismograph 2477
  long-period 2481
  short-period 2481
Seismometer 2475
Selection 3278, 4103, 4135, 5999, 6009, 8317
Selection bias 5554
  self-selection 5557
Selection criteria 3215
Selection operation 4150
Self-affine fractal 3710, 3825, 3882
Self-assembly 2096, 5478, 5893
  directed 5866
  molecular 5866, 5874
  origami-based 261

Self-assembly after explosion 5628
Self-averaging 7002
Self-avoiding random walk 8710, 8713, 8726
Self-avoiding walk 3707
  biological fractals 3786
Self-calibrating intelligent sensors 3250
Self-compression 9370
Self-concept 1905, 8424, 8427
  certainty 8427
  self-esteem 8427
  self-structure 8427
Self-consistency 1683
  stochastic processes 3585
Self-consistent method 7482
Self-consistent solution 7315
Self-crossing random walk 8714
Self-development 3243
Self-efficacy 1905
Self-employment 2285
Self-energy 7666, 8602
Self-energy clusters 6669
Self-energy resummations 6671
Self-evaluation 8425
Self-exciting tar 3443
Self-healing materials 1293
  ceramic matrix composite 1293
  polymer composites 1293
Self-help 3111
Self-inspection 8082
Self-interacting chain, polymer physics 6815
Self-maintenance of sets 337
Self-modulation model 3816
Self-ordering 3271
Self-organization 284, 477, 1303, 1537, 1876, 1879, 1886, 3349, 3358, 3859, 4593, 4895, 5723, 6481, 7991, 7992, 7999, 8001, 8003, 8005, 8006, 8011, 8040, 8067, 8198, 8200, 8206, 8366, 8420, 8422, 8432, 8926, 8927, 8946
  and the city 7954
  complexity in time 3576
Self-organization processes 627
Self-organized 4921
Self-organized branching processes 649
Self-organized critical (SOC) system 3790, 7014, 7584, 8711
Self-organized criticality (SOC) 378, 392, 644, 663, 888, 1135, 1899, 2875, 3795, 7996, 8029, 8040, 8044, 8524, 8543, 8544
  forced 8523, 8543, 8544
Self-organizing 5241
Self-organizing communication system 7997
Self-organizing criticality (SOC) 2562
Self-organizing map (SOM) 1822, 2147
Self-organizing spatiotemporal congested pattern 1180
Self-organizing systems 8040
  models 914
Self-phase modulation 6139
Self-presentations 7996
Self-propulsion 549, 552
Self-reconfigurable robots 1226, 8842
Self-referential emotions 1905
Self-replicating robotic 1226
Self-replication 116, 1226, 8067
  computing machines 8067
Self-reproducing cellular automata 3538
Self-reproduction 5439, 7694, 8068
Self-repulsion, polymer physics 6817
Self-signaling 8132
Self-similar 3848, 3861, 3931, 7827
Self-similar approximants 6740, 6746
Self-similar branching structures 3781
Self-similar fractal patterns 67
  hierarchical iterated function systems 67
  matrix substitution systems 67
  Pascal triangle 67
  Sirpinski gasket 67
Self-similar solutions 2592
Self-similarity 2575, 3654, 3701, 3847, 3893
  biological fractals 3788
Self-similarity measure 7436
Self-sustained oscillations 8946
Selfish DNA, mitosis 4163
Selfish gene algorithm, applications 3240
Selfsimilar structures, complexity 3576
Selkov model 886
Selten game 437
Semantic approach 432
Semantic closure 341
Semantic coordination variables 1559
Semantic grid 8099
Semantic information 7953

Semantic model 4394, 4396, 4397
Semantic order parameters 1559
Semantic place map 2080
Semantic synergetics 8929, 8939
Semantic view 4341
  how-to-solve/compute-it 4341
  knowledge engineering 4341
  uncertainty theory 4341
Semantic web 8084
  commercial applications of 8096
Semantic web challenge 8095
Semantics 352, 1533, 1534, 1536, 4392
Semi-analytical methods
  homotopy analysis method (HAM) 8505
  homotopy perturbation method (HPM) 8505
  variational iteration method (VIM) 8505
Semi-classical regime 1008
Semi-conductor 8940
Semi-dilute solution, polymer physics 6817
Semi-endogenous sequential matching model 4111
Semi-jets 6220
Semi-local bifurcation 2265
Semi-perfect equilibrium 4101
Semi-realistic molecular structure 336
Semi-supervised feature selection 5355
Semialgebra 5453
Semiclassical limit 6748
Semiclassical methods 6749
Semiconcave functions, Hamilton–Jacobi equations 4559
Semiconcavity 4546
Semiconductor barrier 7300
Semicontinuous envelope 6223
Semiconvexity 4546
Semidecidability 816
Semidirect product reduction 5496
Semigroup 8273, 8292, 8293, 8295–8297, 8300, 8301
Semilinear wave equation, controllability 4814
Semiotics 3114
Semiparametric estimators 5555
  informative bounds 5558
Semiwalk 8250
Senge, Peter 9046
Sensation seeking 3457, 3470
Sensitive dependence 9237
Sensitive dependence on initial conditions 3506
Sensitive initial conditions 3830
Sensitive periods 1902
Sensitive system 2234, 9250
Sensitivity analysis 215, 4056, 5536, 5543
Sensitivity of initial conditions 3843
Sensitivity on initial conditions 954
Sensitivity to initial conditions 382
Sensor-based random trees 2081
Sensor fusion 6073
Sensor grids, distributed 5614
Sensor network and amorphous computing 269
Sensorimotor functioning 1901
Sensory–motor coordination 1112
Separable 1817
Separable cost 1664
Separation
  of time scales 3369, 6689
  of variables 5444
Separation, six degrees of 8263
Separatrices 4516
Sequence 3256
  formulation-solution-implementation 7850
  mathematical function 6292
  Morse 8561
  Rudin–Shapiro 8561
  spectral 8555
Sequence complexity 492
Sequence entropy, DNA 495
Sequence spaces
  coordinatewise operations 6291
  defintion 6291
Sequencing, triggering of 266
Sequential compactness, domain problem 2220
Sequential computation 3356
Sequential dynamical system 166
Sequential equilibrium 1587
Sequential exogeneity 2777
Sequential feature selection 5358
Sequential learning 1816, 1819
Sequential machine, reversible 7701
Sequential plan analysis 7995, 7997
Sequential selection 5350
Sequential time 279, 287
Serial choice function 1681
Serial correlation 2773
Serial demand game 1678

Serial dependence 6252, 9836
Serial extension 1683
Serial independence test 6253
  BDS test 6255
Serial principle 1683
Seriation 1904
Series expansion 1031
Serious games 1344, 8378
Server, mobile system 5606
Service industry dynamics 679
Service oriented architecture 8439, 8440
Service-oriented architecture 353
Service robotics 4648
SeSAm 189
Set
  autocatalytic 337
  complete 7762
  critical 6600
  fuzzy 3986, 3988, 4009, 4011–4013, 6927
  fuzzy-fuzzy 4012
  intersection 4012
  interval 4013
  Kronecker 8566
  minimal 7762
  of analyticity 6768
  of decision rules 7762
  of k-recurrence 3090
  of probability measures 4049
  of recurrence 3087
  union 4012
Set level 7790
Set partitioning in hierarchical trees (SPIHT) 10036
Set theory 8268
Settle 3547
Settling time 2247
Seven steps to community safety 3126, 3138
Seventh GrC model 4347
Sexual contact 4660
Sexual development 1905
Sexual intercourse 1905
Sexually transmitted infection 4660
SFF *see* Special fuzzy functions
SFF-LSE *see* Special fuzzy functions with LSE (SFF-LSE) method
SFFM *see* >Special fuzzy functions models
SGML (standard generalized markup language) 8091
Shackle, G.L.S. 6928
Shadow marketing 7833
Shadow price 5536, 5543
Shadowing lemma 4729
Shadowing property 4729, 9248
Shallow water 6206, 8505
Shallow water theory 6210, 9634
Shallow water wave and solitary waves 8123
Shallow water wave equation 3617, 8117, 8118, 9985
Shamir, Adi 5478
Shannon 277, 2867, 3357, 5477
Shannon–McMillan–Breiman theorem 4986, 6944
  Parry's generalization 3072
Shannon–Nyquist sampling theorem 10044
Shannon, Claude 285
Shannon's entropy of information 4821, 6934
Shannon's formula 1046, 1047
Shannon's information theory 1302, 8936
Shape
  asymptotic 4497
  asymptotic 4503
  fractures 6515
  linear 4503
  reverse 4505
  sublinear 4504
Shape factor, polydisperse fractures 6513
Shapley value 1509, 1510, 1514, 1515, 1662, 4095
Shapley, L. 1510
Shared entropy
  biological information 498
Shared vision 9047
Sharma–Tasso–Olver equation 6457
Sharpe ratio 3462, 3464, 3468, 3469
Shattered 5212
Shear 1723
  matrix 1729
Shear band 2649, 5002, 5007
  critical state 5002
  dilate 5002
Shear mechanical topology 7039
Shear viscosity 2682, 4387
Shear wave 2476

Shell models 8010, 8024
for MHD 8024
Shetland 3132
Shift-commuting transformations 66
Shift dynamical system 66
Shift equivalence 8898
Shift homeomorphism 9238
Shift map 4985
Shift of finite type 8894
Shift operator 65, 67, 5306
Shift space 6949, 8892
ergodic theory 6943
Shift transformation 5455
Shifts 2959, 3244
Chacon system 2961
coded systems 2961
context-free systems 2961
Prouhet–Thue–Morse 2960
Sofic systems 2961
Sturmian systems 2961
Toeplitz systems 2961
Shilnikov scenario 390
Shinahr 3567
Shinar 3559
Shinkansen 2398
Shinriki oscillator 3587, 3588
Ship dynamics 6189
Shipping market models 685
Shock 21, 4713, 4716, 6472
Shock absorbers 7040
Shock drift acceleration 36
Shockwave 6490
Shooting method 2014, 7411
Shor, Peter 1013, 7254
Shor's algorithm 7254, 7260
Shor's algorithm 7382
Short cuts 8277
Short memory 2758, 2763
process 2758
Short-range order 5903
Short-term 1888, 1893, 1896, 1906
processes 1875
traffic prediction 9514
Short time propagators 3580, 3582, 5243
estimation 3583
Shortcut 8258
Shortest path 6049, 10061
average 6053
diameter 6053
dimensions 3852
in a labyrinth 1356
problem 5471
Shot noise 6143
Shrinkage 8673, 10033
Shub 279
Shuffle product 1036
Sideband cooling, quantum computing 7226
Sideband excitation, quantum computing 7232
Siegel 6664
Siegel–Bryuno lemma 1941
Siegel domain 8635
Siegelmann 288
Siegel's method 6666
SIENA 6011
Sierpinski carpet 3702, 3703, 7567, 7575
Sierpinski gasket 3702, 3803, 3851, 3890, 7567, 7571, 7573, 7577
Sierpinski lattice 1412
Sierpinski sponge 3702, 3703
Sierpinski triangle 5639, 5640
Siersma's trick 713
Sieve 3475, 3479
Sigma-count, fuzzy logic 3991
σ-Finite ergodic invariant measures, classifying 3077
σ-Shells 7429
Sigmoid 276
Sigmoid function 1817
Sign ambiguity 55
Signal 800, 5869
token-based 5869
voltage-encoded 5869
Signal identification, infrasound 4840
Signal processing 2878
Signal strength 2247
Signal subshift 9257
Signal-to-noise ratio 2247, 2483
Signal transduction 724
Signaling games 4096, 8126
Signaling networks 484
amplitude-encoded 486
frequency-encoded 486
frequency filters 484

high-pass filter 485
low-pass filter 485
Signed infinities, definition 6289
Significance profile (SP) 5697
Signs, series 3765
Silica aerogel 2257
Silicon nanowire field effect transistors 7480
Sill 9881
Silly putty 9813
Šilnikov orbit 6637
SIMD *see* Single instruction multiple data
Similarity 4188, 4412
Similarity dimension 3879
Similarity matrix eigenvectors 7438
Similarity relations 1796
Simmel, Georg 8266
Simon 3547
Simple 6311, 6314
Simple choreographies, minimizing properties 5974
Simple mechanical system 2265
Simple minority game without Information 5597
Simple model 5024
Simple random graph model 8321
Simplicity 3216, 3218
neoclassical economics 179
Simplification
over 7839
unwarranted 7839
Simplified 6315
Simplified neuro-fuzzy systems 6091
Simplify 6311, 6314
Simpson's procedure 9956
SimSesam, simulation tool 195
Simulated annealing 278, 4913, 6574
Simulation 281, 849, 1862, 1975, 2078, 4474, 5467, 6567, 7089, 7098, 8276, 8325, 8405, 9044
agent-based modeling 184
collision 330
discrete-event 330
molecular dynamics 151
numerical 6105
of quantum systems 7132
role of environment 189
Simulation-based methods 5554
data augmentation 5557
Markov chain Monte Carlo 5554
maximum simulated likelihood 5557
smooth 5557
Simulation DTA tools
contemporary 9444
early 9444
Simulation methods 692
Simulation model 7842
multi-agent 150
Simulation modeling 4577, 7832
Simulation universality 5439
Simultaneous equations model (SEM) 5553
Simultaneous localization and mapping (SLAM) 3688
Sinai–Ruelle–Bowen (SRB) measure 954, 990, 1002, 6951, 6952
existence of 961
Sine-Gordon equation 84, 4962, 8475
Sine-Gordon model 9686
Single chain dynamics 6819
Single-crossing condition 8128
Spence–Mirrlees condition 8129
Single dealer markets 5394
Single digit counter 1370
Single-electron device 9715
Single-electron oscillator 9715
Single-electron penning trap 7224
Single-electron transistor 7309, 9715
Single-electron tunneling 9645
Single events 556
Single force 9906
Single instruction multiple data (SIMD) 153
Single-ion operation, quantum computing 7227
Single ion, quantum computing 7226
Single-layer neural networks 1817
Single-minimum-support fuzzy mining (SSFM) 4147
Single-molecule biology 7029
Single-molecule experiments 3607
free-energy differences 3607
optical tweezers 3607
Single-molecule force spectroscopy 7032
Single-molecule magnets 9652
Single-molecule manipulation techniques 7028
Single-molecule techniques 7029
Single nucleotide polymorphisms (SNPs) 3974, 5536
Single peak domains 5512

Single photon sources 7261
Single photon state 7262
Single polymer chain physics 6813
Single-qubit operation 7227, 7228
Single transferable vote 9948
Single transient behavior pattern 8961
Single vehicle characteristics 9383
Single-walled carbon nanohorn 691
Single-walled carbon nanotube 691
Single well potential 6753
Single, error correction 7243
Singlet state, quantum weirdness 7375
Singleton basins of attraction 6043
Singular hamiltonian systems 5961
Singular perturbation problem in *U* (SP-problem) 6704
Singular reduction 2265, 5497
  mechanical systems 5488
Singular trajectories 5449
Singular value decomposition 3766, 8352
  of a matrix 9183
Singular variable, fuzzy logic 3992
Singularity 61, 703, 1720, 3893, 3897
  analysis 8538
  and symmetry 5487
  focus-focus 7140
  lines 1720
  point 1720
  spectra 8538
  spectrum 3894, 8539
  theory 4533
Sink problem 5271, 5277
Sinusoidal forcing 7609
Sipper, Moshe 925
SIR *see* Susceptible-infective-removed (SIR)
SIR model 4660
Sir Philip Sidney game 8136
SIRN – synergetic inter-representation network 7953
Site 7822
Site-bond correlated percolation 1604
Site effects 4436, 4439, 4458
  asperity 4450
  bedrock motions 4458
  crack propagation 4450
  directivity 4450
  dynamic rupture modeling 4452
  effects of local site conditions 4437
  energy and the energy radiated by seismic waves 4450
  high-frequency seismic radiation 4451
  hypocenter location 4450
  kinematic source parameters 4452
  liquefaction effects 4462
  local directivity effects 4450
  local sedimentary cover 4436
  local slip-velocity function 4450
  new crack surface (fracture energy) 4450
  nonlinearity 4454, 4458, 4462
  nucleation phase 4451
  propagating crack tip 4450
  representation theorem 4453
  rupture velocity 4451
  self-similarly expanding crack model 4452
  site-amplification 4458
  site-correction factors 4461
  slip distribution 4450
  slip-velocity function and rise time 4452
  soil classification 4459
  source-site geometry 4463
  strong geologic contrasts 4436
  strong topographic effects 4456
  super-shear speeds 4452
  temporal rupture evolution 4450
  topography 4436
  water-table variations 4436
Site percolation 6539, 6541, 8726
  model 8714
Situated ABM approaches 189
Situatedness 2845
Situation
  assessment 8442
  competitive 3151
  formulation 7850
  nature 7841
  normal 3151
  structure 7841
Situational calculus 1347
Sixth GrC model
  categories of function 4347
  generalized function 4347
  random variable 4347
Size-consistency 6735
Size distribution 6573

Size-predictable 9842
Size quantization 7405
Skeleton 2681
Skew-adjoint, state linear systems 6379
Skew DNA walks 3904, 3908
Skew product 2972, 4602
  action 6799
  group extensions of dynamical systems 2973
  random Dynamical Systems 2973
  system 1945
Skew profile 3904, 3906, 3915
Skewness 5222
Skimming price 1969
Skin effect 2622
  skin depth 2622, 2626
Skin sensitization 2168
Skinner, B.F. 1886
Slab gel electrophoresis 5580
SLAM 6076
Slant stack 9176, 9177
Slater 9945
  index 9951
  order 9951
  problem 9951, 9957
  winner 9951
Slater's rule (transition state technique) 7316
Slaved 4593
Slaving principle 629, 631, 8927, 8928, 8930
SLE 8726
SLE evolution 8713
SLE parameter 8726
SLE property 8722
Slide rule 272, 279, 5476
Sliderblock 5024
Sliding block code 5456, 8893
Sliding state 5027
Slime mold 3349, 8940
Slip 2590, 9586
Slip-velocity 2488
SLM 6398, 6399
Slow earthquake 1221, 2493
Slow manifold 7612
Slow slip events 4271
  and nonvulcanic tremor 4272
  and seismic hazard 4272
  geodetic observations 4271
  temporal characteristics 4272
Slow system 6700
Slow wave 9919
Slowdown 7522
Slowly-varying envelope approximation 6128
Slowness 9179
Slug 9891
Slug flow 9818
Slugbot 3690
Slumps 9578
Smale 279
  work in reduction theory 5493
Smale horseshoe 4727
Small angle neutron scattering (SANS) 6814
Small denominator 5989, 6179, 6666, 6669, 6710, 6714, 6763
Small divisor 5064, 5071, 6322, 6674
Small divisor problem 1939
Small earthquake
  estimate of source processes 2545
  initial phase 2543
  main phase 2543
  waveforms 2544, 2547, 2548
Small gene circuits 1188, 1197
Small labeled-sample problem 5355
Small noise 6232
Small parameter ε 6210
Small-scale inhomogeneities 7928
Small-scale instability 6118
Small time locally controllable 3532
Small world 1431, 2730, 3856, 4663, 6053, 8263
Small-world cities 7968
Small-world models 6053, 8266, 8277
Small-world network 570, 577, 725, 4835, 6049, 6787, 8912
  ultra-small 6053
Small-world phenomenon 6053, 7504
Small-world properties 5270
Smallest available pore, invasion percolation 4948
Smallness assumption 5073
Smectic 2034
Smith dynamic 3184
Smith set 9952
Smith winners 9958
SMM 8800
Smoke 8313
Smoldyn 8742
Smoluchowski 8742

Smooth dynamical system 5452
Smooth flow 4723
Smooth nonsingular transformations 3068
Smooth transition 3443
Smooth transition autoregressive model 446
"Smooth transition" threshold models 2747
Smoothing 3310
Smoothness 1679
  exponent 592
  of the invariant manifold 942
SMS *see* Stationary Markovian strategy (SMS)
Snijders, Tom 8276
Snow cover 1705
Snow transport 883
Snowball and walk samples 8214
Snowball earth 3295
Snowball sampling 8214, 8217
Snowdrift game 6648
Snowflakes 4500
SNOWPACK 1706
SNP 5546, 5548
SNTHERM 1705
SO model 9441
SOAR architecture 1345
Sobolev space 1724, 7442, 7443, 10037
SOC *see* Self-organized critical system
Soccer 6387
Social 2856
Social ability 187
  agent-based modeling 191
Social actors 5999
Social and emotional development 1903
Social aspect 3138
Social choice correspondence 9943
Social choice function 1674, 5516, 5518, 9938, 9943
Social choice problem 9935
Social choice rule 4788
Social circuit dependence 8326
Social classes 2823
Social cognition 1889, 1904
Social comparison or competition 8309
Social contract 5678
  evolution of 5683
Social coordination dynamics (SCD) 8199, 8200, 8205
Social cost of carbon 1091
Social-cultural evolution 8194
Social darwinism 8381
Social development 1897
Social dilemmas 8188
Social dyads 1555
Social experiments, randomized trials 5554
Social field 6478, 8414
Social force 6478
Social influence 8311, 8313, 8330, 8422, 8427, 8429
  clustering 8428
  polarization 8428
Social insects 3683, 3684, 3691
Social interactions 3381
Social judgment 8422, 8423, 8425, 8427
Social learning 1926, 1927
  deferred imitation 1927
  emulation 1927
  imitation 1927
Social memory 8207, 8208
Social net 7833
Social network 3214, 4314, 4343, 5378, 5389, 5390, 8281, 8291–8293, 8295, 8339
Social network topology, agent-based modeling 180
Social networking community 4367, 4368
Social objectives 4790
Social phenomenon 3109, 3111
Social physics 1044, 8266, 8269
Social process 7840, 8406
Social robotics 4648
Social science 8376
Social selection 8330
Social self-help 3111
Social space 9414
Social structures 3213
Social system 7840, 8266
Social temperature 9416
Social theory 9068
Social-theory building 8376
Social volatility, agent-based modeling 180
Socialization 8185
Socially assistive robotics (SAR) 4648
Socially-situated learning 1898
Society 3281
Socio-cultural-cognitive algorithm 8195
Sociograms 8267

Sociomatrix 8191
Sociometric data 8311
Sociometric studies 8311
Sociometry 8267, 8349
Sociophysics 2802, 6381, 8376, 8380, 8413
Sociotype 3209
Soddy 6805
Sofic shift 8895
Sofic subshift 9249
Soft-assembled 1889
Soft assembly 1877
Soft computing 224, 4326, 4335, 4337, 4870
Soft constraints 6936
Soft magnet 5027
Soft margin linear SVM classification 2175
Soft margin nonlinear SVM classification 2179
Soft modes 5015
Soft thresholding 10033, 10037
Soft wavelet thresholding 10036
Software agent 8375
  mobile 5604
Software architectures 1225
Software engineering 4363
Software tool, comparison of commercial 3168
Soil carbon sequestration 1077
Soil-structure interaction 2371, 2421, 2425, 2434, 7899
  example 2379
  frequencies of vibration 2381
  Millikan library 2381
  shear wave velocity 2381
Solar corona 22
Solar energetic particles 29
Solar flare 22, 9269
Solar system 388
  Liapunov times 388
Solenoid 2265, 4531
Solenoid attractor 4727
Solid-on-solid 6412, 6416
Solitary structures 6474
Solitary wave solutions 80, 1376, 6206, 6453–6455, 8112, 8115–8118, 8123, 8476, 8479, 8504, 9970, 9982, 9989, 9997
Soliton 80, 84, 1376, 1394, 2341, 4961, 5138–5141, 5143–5145, 6137, 6206, 6453, 6454, 8113, 8465, 8479, 8504, 8511, 9705, 9970, 9984, 9988, 9989
  annihilation 8497
  antikink 8483
  breather 8483
  collision of two solitons 8504
  diffractive 8483
  dispersive 8483
  envelope 8484
  fission 8497
  fusion 8496
  hidden 8489
  historical discovery 8467
  interaction 8491
  internal 8483
  ion-acoustic 8483
  kink 8483
  line 8485
  Manakov 1378
  mathematical methods 8467
  nontopological 8482
  optical spatial 8483
  physical properties 8476
  plane 8485
  Rossby 8483
  Russell 8482
  semi-discrete 1394
  shallow-water 8482
  ship-induced 8499
  spatial 1394
  spiraling 8497
  surface 8483
  topological 8482
  vector 1377
  vector (composite) 8485
Soliton CA 773
Soliton lasers 6137, 6161
Soliton pertubation 8453
  exp-function method 8456
  homotopy perturbation method 8455
  parameter-expansion method 8456
  variational approach 8454
  variational iteration method 8455
Soliton pertubation theory 8505
Soliton solution 4961, 4962, 4969, 8479
  single- and multi-soliton solutions 8481
  two-soliton solution 8491
Solomon 6788
Solubility 853

Solution 1511, 1513, 1515
additive 1665
anonymity 1666
balanced contributions 1666
complete 4676
constrained egalitarian 1666
core 1667
discrete 4676
Hamilton–Jacobi equations 4545
long-time behavior 4556
maximal 4676
nontrivial 4676
per capita nucleolus 1668
periodic 6588, 6603, 6610
Shapley-value 1665
strong monotonicity 1666
symmetric 1665
Zeno 4676
Solution concept 1509–1511, 1513
Solution path 9225
Solution space 6790, 6793, 6794
Soma 1814, 4871
Somatosensory error maps 2320
Sompi method 9914
Songs 2314
Sophus Lie 6177
Sound spectrograms 2317
Source complexity 2484, 2488, 2492
Source depth 2493
Source durations 9582
Source effects 4439
Source entropy, DNA sequence 495
Source localization 4691
Source mechanism 2474, 2482, 2483, 2487
of VLP signals 9913
Source parameters 2399
Source spectrum 2485, 9581, 9585
seismic 2482
Source time function 2583, 9904
earthquake 2545, 2550
SP *see* Significance profile
Space 208, 6398, 6401
as programming object 262
configuration 6601
cyclic 8555
exponential 6402
homogeneous space 3029
$K(\pi,1)$ 6602
L^2 6600
modeling of 337
of paths 6601
of relative loop 6602
of relative loops 6601
Sobolev 6600
symmetric fock 8556
Veech moduli 8568
Space-bounded quantum computation 7196
Space complexity 7716
Space-fractional diffusion equation 5220
Space three-body problem 5974
Space-time
diagram 249, 803
fractality 3870
Spam 4371
Spam farm 5279
Span programs 7109
Spandrel 491, 5698
Spanning cluster 1448
SPARQL (SPARQL protocol and RDF query language) 8085
Sparre Andersen theorem 5227
Sparse representation 1720
Sparsely sampled, time series 3585
Sparsity 1719, 1989, 8675
Spatial autocorrelation 8273, 8311
Spatial bridge 4663
Spatial correlation 1698, 3792, 6570
Spatial databases 7737
Spatial discrete scale invariance 7579
Spatial DSI 7578
Spatial entropy 1047
Spatial graphs 2742
Spatial heterogeneity 7855
Spatial interaction 1043, 1045, 1047, 1055
Spatial light Modulators 6394
Spatial patterns 8946
Spatial periodicity, fracture networks 6510
Spatial random fields 4448
correlation lengths 4448
fractal dimension 4448
Hurst number 4448
von Karman auto-correlation function 4448
Spatial scaling structure, polymer physics 6817
Spatial sequencing, in amorphous computing 266

Spatial structure 2737
Spatially extended coevolution 3240
Spatially grounded models 192
Spatially separated GPs 9376
SpatialRes 6398, 6399, 6401
Spatio-temporal intensities
  Bayesian 9848
  kernel 9847
  nearest-neighbor 9847
  nearest-neighbor kernel 9847
Spatio-temporal pattern 6477, 8933
Spatiotemporal congested traffic pattern 9501
Spatiotemporal dynamics of traffic congestion 9409
Spatiotemporal stress field 7855
SPC (search path count) weights 8253
SPE *see* Subgame perfect equilibrium (SPE)
SPE outcome path 2225
  dynamics 2225
SPE payoff 2222
Spearman correlogram 9856
SPEC 5355
Special fuzzy functions
  with LSE (SFF-LSE) method 4084
  with SVM (FF-SVM) method 4088
Special fuzzy functions models (SFFM) 4092
Special fuzzy functions (SFF) 4083, 4090
Special symplectic reduction method 5497
Specific entropy 7523
Specific heat 692, 697, 700, 701
Specifically branched chains 6842
Spectral algorithm 1153
Spectral analysis 3760
Spectral bisection method 1147
Spectral dimension 2253, 3888, 3889, 7569, 7571, 7572
Spectral element scheme 7897
Spectral elements 7892
Spectral factorization problem 8755
Spectral feature selection 5354
Spectral form factor 7168
Spectral mapping problem 948
Spectral mapping property 8646
Spectral methods 1501
Spectral property 2921
Spectral radius 6242
Spectral representation 8719
Spectral solution 7411
Spectral theory, nonsingular systems 3069
Spectral transport 8596
Spectral weight 8719
Spectroscopy, volcanic eruption 4850, 4851
Spectrum 5022, 5464, 6653
  absolutely continuous 8555
  countable Lebesgue spectrum 2928
  discrete 8555
  Haar 8555
  of agents 1338
  rational 2928
  simple 8555
  singular 8555
Speech cues 2315
Speech perception 2321
Speech production 2320
Speech recognition 2315
Speech signal reconstruction 3252
Speech sound map 2320
Speed 7522
  and density disturbance 9282
  free 3152
  horizontal upward walking 3152
  upstairs 3152
Speed disturbance 9285
Speed subshift 9257
Speed-up theorems 808
Spherical Paul trap, quantum computing 7223
Spherical source 9905
Spikes 6100, 6102
Spiking neural P system 5525
Spin 5748
  quantum mechanics 7371
Spin blockade 7309
Spin-chain 7355
Spin chirality 9649
Spin conservation 8110
Spin current 8106, 8109, 8608, 8612
  pure spin current 8611
Spin density 8106
Spin-density functional theory
  Kohn–Sham (KS) equations 7314
  local spin density approximation 7315
Spin dipole 8108
Spin dynamics 8588
Spin filter 8598, 8608, 8610

Spin gap 9685
Spin glass 6418
Spin hall effect 8599, 8611
Spin interaction, polymer physics 6815
Spin models 610
Spin-orbit interaction 6681, 6682, 8105, 8598, 8607
  Rashba SO 8598, 8608, 8612
Spin pockets 8595
Spin polarization 8578
Spin precession 8110
Spin pumping 8599, 8611
Spin qubits 9645
Spin ratchets 8599, 8611
Spin-spin contact 7446
Spin splittings 5775
Spin-stiffness 8583
Spin-switch 8598
Spin switches 8607
Spin transport 8104
Spin transport current 8108
Spin wave stiffness 1275
Spinodal 3618, 5024
Spinodal dewetting 3633
Spiral defect chaos 6464
Spiral galaxy 413
Spiral wave 776, 6470, 6472
Splay faulting 9588
SPLERT *see* System for processing local earthquakes in real time
Spline spaces 5801
  approximation order 5808
  dimension of 5839
Spline wavelets 10033
Splines 1721
  B-splines 6856
  box splines 6856
Split brain patients 1426
Split selection 1827
Splitting lemma 7634
Splitting method
  head of the splitting 794
  leading tile 794
  region of the splitting 794
Splitting of a flow interruption 9398
Spontaneous emission 6118
Spontaneous emission noise 6141
Spontaneous symmetry breaking 6187
Spontaneous traffic breakdown 9320, 9380
Spontaneous wide moving jam emergence in synchronized flow 9326
Sprawling 9426
Spreading of an epidemic 3848
Spreading rate
  critical 6060
  endemic 6060
  epidemic threshold 6060
Spreading set 9256
Spreads 5408
Spring embedder 1254, 8352
Spyware 4370
Square radius of gyration 6840
Square synchronization algorithm 3567
Squeezing procedure 3329
Squirt flow 2694
SRB measure 1006, 1007, 5453
SRM *see* Structural risk minimization
sSelect 5356
SSFM problem 4149
Stability 1520, 1541, 1555, 2733, 3662, 4681, 4725, 5722, 6037, 6104, 6426, 6434, 8001, 8003, 8004, 8199, 8201, 8927
  external 1520
  global (pre-)asymptotic 4681
  internal 1520
  Liapunov 6590
  linear 6104, 6590
  multi 6105, 6111
  of periodic orbits 6590
  spectral 6590
  stability of rotating fluid bodies 8631
  stability of the solar system 8631
  vulnerability 2736
Stability analysis, biological models 1190
Stability criterion 9997
Stability house 2066
Stability index 6434
Stability theory 8631
Stabilizability 3524
Stabilization 1488, 1493, 4687, 8617
Stabilizer code 7335
Stable 713, 8631
  and unstable manifolds 998

asymptotically stable 8631
orbitally asymptotically stable 8631
Stable attractors 5720
Stable completion method 596
Stable distributions 3440, 3812
Stable games 3188, 3197
Stable laws 965
Stable librations 6447
Stable manifold 2265, 4516, 4727, 8644
Stable manifold theorem 4726, 4728
Stable matching problem 4097
Stable mode 8930
Stable outcome 9656
corewise-stability 9657
pairwise-stability 9657
setwise-stability 9657
strong-corewise-stability 9657
Stable set 1520, 4095, 4726, 6026, 6040, 9227
discriminatory 1522
farsighted 1528
objective 1522
sdom 1526
symmetric 1522
wdom 1526
Stable subset 9231
Stable trajectories, complex 6105
Stäckel potential 403
Stackelberg payoff, pure action 7657
Stacking 5322
Stag hunt 6645
Stag hunt game 3178, 3179
Stage 1898, 4099, 4105
Stage transitions 1899
Stairs 3152
Stampede 3153, 3154
Stand-alone core 1676
Standard constraints, fuzzy logic 3999
Standard deviation 6145
Standard map 384, 411, 1935
Standard refinable function 592
Standard symplectic form 5065
Standard variation, n-body problem 5983
Star 3274, 3279, 8324
Star clusters
gravothermal catastrophe 8691
self-similar solution 8693
thermal stability 8691
thermal timescale 8692
Star-extension 6294
super-structure 6303
STAR model
multiple regime STAR model 447
n dimensional two-regime STAR model 447
Stark effect 6737
Stark shift 5773
State 1311, 5191, 8698
adjoint 1503
biochemical 526, 531
border 803
clustered 631, 632
desynchronized 632
differentiated 526
dynamical 526, 527
equilibrium 533
final 1311
four-cluster 635
initial 1312
of nature 427–430
of the system, controllability 4805
of the world 428, 429, 431, 4788
quantum computing 7220
quiescent 803
stable 528, 529
steady 527, 538
steady/stationary 528
synchronized 632
two-cluster 635
unstable 528, 529, 538
State-change complexity 3542
State constraint 6224
State dependence 2771
State detection, ionic qubit 7229
State independent signaling 7646
State initialization, quantum computing 7225
State-linear realization 6376
State-linear skew-adjoint realization 6376
State-linear systems, observation space 6376
State observers 9088
high-gain observer 9088
State space 482, 713, 1887
high dimensional 1540
lower dimensional 1540
State-space grids 1882

State-space model 2751, 2753
State-space planning 5192
State-space realizations 8753
State transition 1312
  diagram 64, 67, 71, 75
  equation 4077
  map, Boolean networks 1199
  probabilities 3696
  table, Boolean networks 1199
State variable 1488
State vector 8926
  stochastic processes 3577
States
  alternative 528, 529
  biochemical 527, 531
  dynamical 527
  epigenetic 527
  functional transformation of 187
  localized 6104, 6111
  oscillation 545
  oscillatory 6104
  stable 527
  steady 527, 529
Static assignment algorithms
  equilibrium 9436
  non-equilibrium 9436
Static circular crack 2590
Static coupling patterns 8916
Static friction 5025
Static games 4095, 8649
Static model (developmental) 1882
Static O-D estimation 9452
Static routing 9434
Static strain energy 2590
Static traffic assignment 9436
Stationarity 8722
Stationary Markov process 2935
Stationary Markovian strategy (SMS) 2217, 2218
Stationary process 5039, 5453, 8669, 8768
  past and future subspaces 8757
  scattering pair 8758
Stationary stochastic process 2935, 8752
  distribution 2935
Stationary structures 6467, 6469
Statistical analysis of microarray data 1234
  model-based analysis of gene expression profiles 1237
Statistical control chart 9846
Statistical ensemble 5671
Statistical hydrodynamics 3648
Statistical inference 6000
Statistical inference modeling 522
Statistical inference models 512
Statistical inference networks 516
Statistical learning 5206, 5207, 5210, 5217
Statistical mechanics 400, 1269, 2803, 2859, 3364, 4380, 4999
  biological systems 2700
  ecosystems 2706
  edwards ensemble 4999
  entropy 4999
Statistical mechanics ensembles 695
Statistical methods 5999
Statistical models
  Bayesian mixture models 1240
  Bayesian time series models 1240
  clustering 1240
  Dirichlet process 1232
  for fault slip 2596
  mixture models 1231, 1232
  mixture of Dirichlet processes 1232
  model-based analysis of gene expression profiles 1237
  neural networks 1240
  partial least squares (PLS) 1240
  principal components analysis 1240
  random effect 1232
  regression models 1231
  state-space models 1231, 1232
  wavelets 1240
Statistical physics 721, 5024, 6407, 8711, 8714
Statistical stability 954
Statistical tests about fuzzy means 8688
Statistical thermodynamic 5668
Statistical zero-knowledge 7191
Statistically independent 370
Statistics 2802
  anomalous 3591
Statistics of earthquakes 5022
Staudinger's relation 6848
Steady bifurcation
  liquids 5935
  three-dimensional flow 5940
Steady flow, a liquid past a body 5938

Steady state 9080, 9082, 9841
behavior 9082
locus 9082
motion 9082
nonequilibrium 6953
Steady-state compliance 6849
Steady-state flow 5930
Steady-state solutions
Navier–Stokes equations 5930
Steady state systems
kinetics 2701
Steam 9880
Steam reforming
design of improved catalyst 1327
elementary reaction steps 1327
Steam-reforming 1326
Steen Rasmussen's 933
Steep function 5991
Steepest descent control law 2058
Steered 7036
Stefan problem 3724
Stella 9015
Stellar dynamics 389
Stellar evolution 3279
STEM 5892
Stem cells 1192
Step
collision 3161
function 1817
propagation 3161
100-Step rule 3336
Step size 1820
Stepover 5026
Stepped drum 5476
calculator 5476
Stepping stone patch 2741
Stern's behavioral explanation model 3120
Stick-pulling 3691, 3692, 3694
Stick-slip 5583
Sticker model 2101
Stickiness 409
Stiffness 691
Stiffness exponent 6419
Stigmergy 120, 8869, 8871, 8872, 8877
Stimulated emission 6118
Stimulation
anti-kindling 640, 641
cortical 627
current 634
deep brain 627
demand-controlled 630
desynchronizing 629, 639
double-pulse 630
high-frequency 627, 633
high-frequency deep brain 629, 641
kindling 640, 641
low-frequency 639
multisite coordinated delayed feedback 629
multisite coordinated reset 629–631, 634, 641
nonlinear delayed feedback 629
open-loop type 632
permanent deep brain 628
signal 630, 634–636
single-pulse 630
site 631, 635
Stimulation protocol 639, 640
Stimulation-registration setup 638, 639
Stimulation techniques 628, 629
Stimulus 6110, 6111
STING 5059
Stochastic 689, 692, 694, 3365, 3373, 7719
Stochastic acceleration 22, 23, 25
Stochastic actor-oriented model 6003
Stochastic analysis, field theoretic methods 7025
Stochastic averaging of the wave equation 7920
Stochastic branching model 5028
Stochastic cascade process 3591
scale 3592
Stochastic climate model 2
Stochastic complex systems 3361
Stochastic convolution 6233
Stochastic data analysis 3577
Stochastic differential and difference equations 8750
Stochastic differential equation 3445
Langevin-type 3361, 3369
Stochastic differential equations 8775
Stochastic differential games 6221
Stochastic dynamical system 6232, 8750
on Hilbert spaces 6217
Stochastic dynamics 729, 3367, 3369
Stochastic evolution equations 3578
Stochastic fluctuations 3368
Stochastic focusing 556, 8735

Stochastic forces 3583, 5733
Stochastic game 2216, 4096
folk theorem 2218
Stochastic game model
equilibrium 2217
history 2217
set-up 2216
state space 2217
state variable 2217
Stochastic growth 2811
Stochastic interacting particle systems 3365
Stochastic inverse 9193
Stochastic kinetics 3366
Stochastic Loewner evolution 8710, 8713, 8720
Stochastic Lorenz equation 3586
Stochastic matrix 5455, 8768
Stochastic methods 8266
Stochastic models 2556
GPS measurements 2557
instrumental catalogs 2557
statistical seismology 2557
Stochastic network, rock material 4955
Stochastic noise 3370
Stochastic optimization 4910
Stochastic partial differential equation 6229
Stochastic perturbations 970, 6232
Stochastic phase space 167
Stochastic processes 3361, 3365, 3368, 3574, 3575, 3583, 3586–3589, 3591–3593, 4513, 6009, 7569
statistical description 3579
Stochastic quantization 6231
Stochastic realization 8752, 8753
generating white noise w 8754
state process x 8754
stochastic realization of discrete-time random processes 8753
Stochastic resonance 278
Stochastic similarity matrices 7431, 7455, 7457
Stochastic similarity matrix 7456
Stochastic simulation 730
Stochastic simulation algorithm 729
Stochastic stability 954, 3194–3197, 4732
Stochastic time series analysis 3582
Stochastic transformation 7435
Stochastic volatility 4129
different notions of 8783
model 3446, 5328, 5329
short definition 8783
Stochastically stable state 3176
Stochasticity 3361
Stock 5407, 7842, 8991, 9043
London Stock Exchange 5408
Stock-flow diagrams 673
Stock management 676
Stock market 4925
crashes 7020
Stock return 3454
Stoichiometric matrix 328, 557, 558, 5541, 8733
Stoichiometry 556, 5539, 5548
Stokes approximation
viscous liquids 5941
Stokes equations 549
Stokes expansion 9979, 10000
Stokes flow 5576
Stokes line 6748, 6752
Stokes paradox, viscous liquids 5941
Stokes rule 6752
Stokes water waves 10003
Stokes waves 9977
Stokman, Frans 8269
Stone–Wales topological defect 699, 700
Stone–Wales transformation 9151
Stop and go 9371, 9422
Stop-and-go wave 6488
Stop-loss orders 5415
Stopping criterion 5351
Stopping rule 1816, 1819
Stopping time 8387, 8394–8396
Storm surge 3126
Story 3120, 3125
images tell a story 3125
story telling bonds neighbor 3125
Strain 694, 699, 1320, 5012, 5752, 5753, 5763
Strand asymmetries 3913
Strange attractor 383, 3087, 3829, 3837, 3838, 3842, 4593, 6134, 6629
Hénon scheme 3839
nonlinear dynamics 2705
Strategic 4098
Strategic abstention 9941
Strategic analysis of markets 4108
Strategic basin of attraction 6026
Strategic behavior 1980

Strategic decision processing system 7850
Strategic flexibility 7833, 7842
Strategic form 1510, 1528, 1587
Strategic game 1660
Strategic management 7836
Strategic manipulation 1676
Strategic situation formulation 7838
Strategic voting 9934, 9938, 9940
Strategically abstain 9941
Strategies and tactics net 7834
Strategy 1510, 1949, 2107, 4098, 5740, 7632, 8381, 8698
  behavioral 7617
  business 7835
  copycat 7840
  corporate 7835
  divisional 7835
  ε-optimal 8701
  functional 7835
  hierarchical 5740
  mixed 8700
  of a player 7630
  private 7627
  process 7835
  process or functional 7835
  public 7625
  pure 8700
  stationary 8700
  stationary Markovian 2217
  tradeoffs-free 7834
Strategy design 7839, 7850, 7851
  tradeoffs-free 7835
Strategy implementation 670, 672
  calibration 672, 674
Strategy-proof 1672
  group 1672
Strategy-proofness 1672, 5514
Stratification 2265, 4515, 6195, 6198
Stratified count, fuzzy logic 3991
Stratified fluid 6207
Stratified sampling 5569
  choice-based sampling 5569
  endogenously stratification 5570
  weighted GMM estimation 5570
Stratonovich definition 3578
Strength profile, earthquake 2553
Strength threshold 5023
Stress 2636, 5012, 9584
  and triggering, static 2601
  asymmetric 2636
Stress chains 5003
Stress change
  dynamic 2600
  static 2600
Stress concentration 2593
Stress corrosion 2615
Stress drop 2590, 10053
Stress glut 9902
Stress intensity factor 2593
Stress release 6490
Stress-release model 2564
  conditional intensity 2565
  coupled stress release model 2565
Stress sensitivity 7927
Stress transmission 4998
Stress vector, Navier–Stokes equations 5929
Stressed backbone 2838
Stretched exponential 3715, 3772
Stretched exponential distribution 7012
Stretched vortices 3646
Stretching tensor, fluids 5929
Strichartz estimates 2045
Strictly exogenous 2773
Strictly hyperbolic system 4716
Strike fault 9614
Strike slip fault 5021
String–string deformation 2636
Strings, cosmic 1642
Stripe formation 6482
STRIPS 5188
Strogatz, Steven 8269, 8277
Stromboli 9786
Strong and weak memory 1624
Strong congestion 9369
Strong coupling limit, cavity-QED 7245
Strong disorder 6501
Strong earthquake 1625
Strong electron–electron interactions 7350
Strong feller 6231
Strong glass-formers 4212
Strong ground motions 7894
Strong interactions 6663
Strong law of large numbers 957
Strong L_p inequality 2937

Strong mixing 2946
Strong shift equivalence 8897
Strong stability 6027, 6042
Strong ties 8254, 8263
Strong transversality property 8644
Strong turbulence 6195, 6201
Strong uncertainty 4633
Strongly connected components 5269, 6245
Strongly irreducible 8906
Strongly mixing 6232, 7523
Strongly nonlinear 5255
Strongly semiconcave, Hamilton–Jacobi equations 4547
Strongly semiconvex, Hamilton–Jacobi equations 4547
Structually dynamic cellular automata
  CA simulators 8831
  classes of behavior 8824
  hierarchy of models 8831
  labeled links model 8830
  models of computation 8829
  models of fundamental physics 8843
  random dynamics approximation 8838
  related graph dynamical systems 8840
  relative location model 8830
  statistical measures 8826
  symmetric links 8830
Structural balance 6005
Structural change 5326, 5328, 5330, 5331, 5344
Structural complexity 491
Structural damage 2371
Structural damage detection 2372
Structural descriptor 2114, 2140, 2160
Structural equivalence 8273, 8289, 8300, 8314
Structural form 214
Structural health monitoring 2371
Structural hierarchies 4593
Structural holes 8239, 8274, 8284
  non-redundant links 2735
Structural individualism 6003
Structural models 2375, 5559
  causality 5560
  conditional prediction 5559
  cowles commission approach 5559
  natural frequency of vibration 2375
  oscillator 2375
  reduced form 5559
  wave guide 2376
  wave travel 2376
Structural position 2735
Structural properties 654
Structural risk minimization (SRM) 4086
Structural robustness, biological models 1190
Structural stability 954, 4731, 6699, 8643
Structural system 2421
Structural unstable 711
Structurally dynamic cellular automata 840, 916, 8816
  basic model 8818
  numbers of possible rules 8822
Structurally dynamic disordered cellular Networks 8844
Structurally stable 3098, 4516, 6311, 8643
Structure 2371, 2433, 5737, 6310, 6317
  complex 7837
  complicated 7837
  excitation of 2433
  multi-symplectic 6609
  of a molecule in AC 328
  preservation 6317, 6322
  stock and flow 7847, 7849
  symplectic 6588, 6610
Structure-activity relationships (SAR) 2114, 2140, 2161
Structure causes performance 7841
Structure condition 6221
Structure contribution approach 8949
Structure (Coulomb "diamonds") 7311
Structure factor 2026
  biological fractals 3792
Structure functions 3652, 8537, 8542
  range-limited 8542
Structure indices 2642
  phase shift 2642
Structure-odor classification 2182
Structure-property models 2140
Structure-selectivity model 2149
Structure-to-function mapping 330
Structure/traversal isomorphism 4608
  robust information transmission 4610
  transmission quality 4613
  traversal 4611
Student *t* distribution 3421
Sturm–Liouville eigenvalue problem 7128

Style 1266
SU(2), quantum mechanics 7372
Sub- and supersolutions 6619
Sub-nets 2400
Sub-Riemannian geometry 6217
Subadditive
  ergodic theory 6944
Subband 7664, 8576, 8586
Subchain 6844
Subchain configuration 6818
Subdeterminant rank 1868
Subdiffusion 311, 2872, 5223, 5232, 6195, 6203
Subdigraph 6242
  induced 6242
Subdirect product 8298
Subdivision scheme 592
Subduction 9587, 9589
Subduction zone 2484, 2608
Subduction zone earthquake 10047
Subeconomy 2285
Subgame perfect equilibrium (SPE) 2218, 4100
Subgradient 6271
  ergodic theory 6945
  limiting 6274, 6275
  proximal 6274
  viscosity 6275
Subgraph 575, 8249, 8271
  and network motifs 5695
  complete bipartite 8258
  examples 5695
  in complex networks 5694
  induced 8249
  spanning 8249
Subgraph abundance, network 507
Subgraph-based information complexity descriptors 4833
Subgraph census 5693, 5694
Subgraph composition, network 507
Subgraph count 4833
Subgraph information content, network 507
Subharmonic instabilities 2327
Subharmonic solution 5306
Subinverse 8297
Subjective probability 9418
Subjectivity 1429
Submanifold, initial 5491
Submarine landslides 1221
Submatrix, rank 1868
Submeasures 2235
Submodular function 6421
Suboptimal neuro-fuzzy systems 6095
Suboptimum solution 1818
Subordination 5232
Subquadratic algorithms 8247
Subscript 1962
Subsequence ergodic theorem 2945
Subshift 790, 2234, 2981, 9249
  admissible transitions 2981
  entropy of a 790
  ergodic theory 6941
  finite type 790, 6947
  forbidden words of a 790
  irreducible 790
  language of a 790
  Markov subshift 2981
  of finite type (SFT) 2981
  one-sided Markov subshift 2981
  sofic 790
  strongly irreducible 2981
  topologically mixing 2981
Subshift of finite type 954, 9249
Subsidiaries, active 7835
Subsolutions
  Hamilton–Jacobi equations 4543, 4548, 4549
  regularity results 4547
Subspace algorithms 8763
Subspace identification 8760
Substance abuse
  cocaine prevalence 4577
  heroin addiction 4577
  tobacco 4577
Substrates 723
Subsumption 6070
Subsumption architecture 3689
Subthalamic nucleus (STN) 629
Subthreshold logic 9720
Sufficiency 8221
Sufficient statistics 4627, 6981, 8228
  sufficient unit 4609, 4610
Sugarscape 94
Sugeno integral 232, 6932
Sugeno model 4696, 4882
Sugeno–Yasukawa models 4092
Suitability of membership functions 4152

Suitability of the derived membership functions 4157
Sum resonance 6626, 6629, 6632, 6635
Sum rules 592
  multiwavelet 5847
Sum squared error 1818
Sumatra–Andaman earthquake 4265
Sumatra earthquake 4846
Summable potential
  Gibbs state 6947
Summation at the smallest term 5990
Summation rule 6663
Summation theorems 560
Summer mean temperatures, observed an modeled 1098
Sun 5473
Sundman–Sperling estimates, generalized 5980
Super-agents 150, 151
Super-Arrhenius behaviour 4212
Super-exchange 7352
Super-graphs 5786
Super-hydrophobic 5582
Super-individual approach 150
Super-localized wave functions 2254
Super-recurrence 8489
Super-structure, definition 6302
Super-Turing power 288
Supercell 5778, 8592
Supercomputer, quantum simulation 7385
Superconducting percolation networks 6555
Superconductivity exponent 1408
  superconductivity 1410
Superconductor 2021
  disordered 6413
Superconductor–normal conductor 3850
Supercooled phase 4211
Supercritical bifurcation, liquids 5935
Superdiffusion 311, 2871, 6195, 6203
Superexponential stability 5994
Superfield path integral 8591
Superfluid 9681
Superfluid stiffness 9689
Superintegrable system 2265
Superlattice 2648, 7405
Supermassive black holes 29
Supermodular games 3189, 3196, 3197
Supernetworks 6035
>-Supernetworks 6036
Supernova 22, 28
Superposition 7252, 7328
Superposition principle 1023
Superrigidity 3104
Supersolid 9685
Superstatistics 2878
Superstrings 1642, 1656, 3869
  cosmic 1642, 1656
Supersymmetry 3869
Supervised feature selection 5352
Supervised learning 1815, 1992, 2162
Supply chain dynamics 676
Support network 5681
Support vector machines (SVM) 2171
Support vector regression (SVR) 4086
Support vectors, fuzzy system model evolution 4088
Supremum, definition 6290
Surface 714, 1307, 1310, 1311, 6412, 8122
  models 1310
  of section 6426, 6435
  reaction center 1307
  site 1307
  three low Miller index 1311
Surface acoustic waves 1022
Surface defects 1320
Surface-electrode trap, quantum computing 7223
Surface evolver 5582
Surface Green's functions 7486
Surface patterning 5574
Surface reaction model 888
Surface tension 9788, 9818
Surface tension coefficient 5581
Surface wave 2483, 2601, 2607, 3616, 8113
Surface-wave magnitude 2479, 2480, 2484
Surjective 64, 65, 70, 71
  cellular automata 2983
Surjectivity 5439, 7680
Surplus demand 7846
Surprises, strategic 7834
Surrogates 7996, 7999
Survey sampling 8213, 8215, 8216
Survival probability 5226, 5233
Survivor distribution 7015
Susceptible individual 6779
Susceptible-infected-susceptible (SIS) model 654

Susceptible-infective-removed (SIR) 3791
model 4661
Suspension flows 2974
Sustainability 3126, 3138, 7842
Sustainability implementation research 3111, 3138
Sustainability management 9052
Sustainable development 3131, 4632
Sustainable solutions 7832
SV 8784
SVM *see* Support vector machines
SVM regression 2180
Svoboda 5475
SVR *see* Support vector regression
SWAP gate, quantum computing 7232
Swarm 212, 3683, 8869–8873
intelligence 3683, 3690
robotics 3690
Swarm behavior 8367
Swarm-bot 3692
Swarm intelligence 123, 8883
Swarm optimization 8869
Swarm robotics 8869
Swarming behaviors 7724
Swift–Hohenberg equation 3628, 8931
Swift–Hohenberg–Haken equation 3629
Swimming 548, 550
efficiency 551
optimal 549, 551
strategies 549
Swiss cheese model 1403
Switchable channel 1369
Switching 1548, 1555, 2090
Sybil attack 5279
Symbol 3353
multiwavelet 5844
Symbol binding 3353
Symbol grounding 8098
Symbolic action 1904
Symbolic AI 1110, 3353
Symbolic approach
ACT-R 1113
EPIC 1114
SOAR 1113
Symbolic dynamical system 8891, 9248
Symbolic dynamics 956, 4729, 5439, 6941, 6943
Symbolic grounding problem 1119
Symbolic models 1999
Symbolic processing 119
Symbolic space 9248
Symbolization 1901
Symlet 10020
Symmetric 3550
Symmetric difference distance δ 9949
Symmetric links SDCA model 8830
Symmetric periodic orbit 6430
Symmetric phase 5590, 5594, 5598, 5600
Symmetric stable processes, ergodic theory 3076
Symmetric theory 2837
Symmetrical uncertainty 5357
Symmetry 596, 1554, 2245, 5035, 6175, 6177, 6310, 6626, 6735, 6736, 6761, 8927, 9105
and perturbation theory 6736
and singularities 5487
approximate symmetry 6182
axial 6319
axially 6316
breaking 7136
ecological 2706
generalized symmetry 6190
Hamiltonian systems with 6597
mechanical systems 5483
quantum mechanics 7373
rotational 6315, 6319, 6320
toroidal 6311, 6316, 6320, 6322
torus 6315
Symmetry algebra 5487
Symmetry breaking 278, 658, 1549, 1551, 3288, 5727, 8928
spontaneous 7136
Symmetry-breaking field 6773
Symmetry groups 593, 1547, 1554, 5963, 6632
core 5969
invariant 1547
planar 3-body problem 5972
Symmetry reduction 6175, 6187, 6603
Symmetry state 8005
Symplectic 2003, 5066, 5988, 6318
matrix 5066
Symplectic algorithms 8697
Symplectic form 2265, 4528, 5084
Symplectic manifold 2265, 5065, 5445
Symplectic orbit reduction theorem 5491
Symplectic orthogonal 5488

Symplectic property 6433
Symplectic reduced space 5486
as coadjoint orbits 5490
Symplectic reduction 5488
special method 5497
theorem 5486, 5487
theory 5483, 5484
without momentum maps 5497
Symport/antiport 5525
Synapse 3335, 3343, 6099, 6102, 6103
Synaptic efficacy 3343
Synaptic plasticity 639
Synaptic weight 4871
Synchronizability 8911
Synchronization 627, 635, 637, 770, 1544, 1552, 1555, 3233, 3538, 4915, 6105, 6108, 7992, 7999, 8003, 8200, 8203, 8910
clapping 1555
manifold 8912
oscillatory 1552
pathological 627, 629
Synchronized flow 9356, 9357, 9370, 9417
Synchronized flow pattern 9356, 9364
Synchronized flow traffic phase 9303, 9308
Synchronized traffic congested pattern 9506
Synchronizing word 9249
Synchronous network 7714
Synchronous updates 3234
Synchrotron 22
Synchrotron losses 27
Syncopation-synchronization paradigm 5734
Syndemics 4586
Syndrome 8939
Synergetic cities 7962
Synergetic computer 8937, 8938
learning procedure 8938
Synergetic navigation system 8003, 8007
Synergetics 627, 1542, 4895, 5723, 7992–7994, 8003, 8005, 8043, 8381, 8413, 8926, 8927, 8942, 8946
lower dimensional dynamics 1542
Synergy 1537, 1540
Syngas 1326
Synonymous nucleotide substitution 5657
Syntactic approach 432
Syntactic closure 341
Syntactic generator 1891
Syntax 4392
Synthesis 10015
Synthetic aperture radar 1692
Synthetic biology 258, 417, 567, 5537
Synthetic catalogs 2359
Synthetic chemistry 417
Synthetic environments 185
Synthetic genetic array (SGA) 9109
Synthetic O-D generation 9451
dynamic formulations 9454
static formulations 9452
Synthetic seismograms 9585
Synthetic sickness or lethality (SSL) 9106
System $\mathbb{R}^{\mathbb{N}}$ 6291
System analysis 4364
System-as-cause thinking 4564
System commands, agents as 188
System design 4364
System dynamics (SD) 2285, 4475, 4565, 4577, 7842, 7849, 8371, 8406, 8974, 8975, 8983, 8991, 9000–9002, 9004, 9005, 9044, 9058
applications, health care 4561
as presentationalist 9067
case study 4567
components 4564
criticism 9061, 9068
definition 4563
foundations 9061
history 4563
justification in 9067
model, configuration 4567
model, health and social care 4567
modeling tools 4565
need 4564
origin of 9058
participative model building 4572
servomechanism 9068
simulation 9074
studies, epidemiology in Europe 4571
studies, health and social care management 4571
thinking skills 4564
workforce planning models, health management 4574
System entity structure 356
System frequency 2389

System identification 5206, 5208–5210, 5214, 5215, 5217, 8759
System implementation 4365
System optimum 9434, 9441
System size expansion 8780
System structure 353, 7841
System theory 8267, 8974, 8980
Systems 2234, 2263, 7798, 8405
  adjoint 1503
  affine 1034, 1035
  approximately controllable 1489
  approximating 1037
  autopoietic 7843
  discrete time 1038
  dynamical 1034
  exactly controllable 1489
  for processing local earthquakes in real time (SPLERT) 2665
  fully nonlinear 1038
  Hamiltonian 6588, 6594
  hierarchical 7833
  Lagrangian with symmetries 6601
  lifted 1038
  linear 1038
  natural mechanical 6600
  nonlinear 1030, 1038
  null controllable 1489
  of balance laws 4714
  of cities 7957
  of conservation laws 4714
  purposeful 7833
  universal 1035
  well-posed 4675
  with discontinuities 961
Systems affine 3530
Systems analysis
  biological network 9103
  ecological complexity 2701
Systems approaches 8974, 8975, 8978, 8982
Systems biology 420, 477, 1227, 3978, 5537, 9091, 9093, 9095, 9097, 9098, 9102, 9103
  data integration 9103
  definition 9101, 9103
  emergent 9102
  model 9103
  networks 9102
Systems biology workbench 567
Systems ecology 2705
Systems methodology 8979, 8982, 8983
Systems movement 8975, 8977, 8979
Systems neuroscience 7993
Systems sciences 8974, 8977
Systems theory 1044, 4634
Systems thinking 671, 9044
Systems thinking/system dynamics method 4565
  and modeling methodology 9045
Syzygy 2265
Szabó A 781, 789
Szemerédi's Theorem 9244
Szemerédi's theorem 3091, 5451
Sznajd model 6385, 8390, 8391
Sznajd-Weron, Katarzyna 8390
Sznajd, Józef 8390
Szwerinski 3547, 3564

T

Δt-Adaptation 7906
t-Conform 224, 4067
T-Equivalent 5458
T-Exponential 6745
T-Nonshrinkable 5458
t-Norm 224, 4067
T-Rank 5275
T-Stage game 7633
Table of games 9740
Tactics 7839
  competitive 7839
  cooperative 7839
  counterproductive 7840
  defensive 7839
  implementation 7839
  marketing 7846
  offensive 7839
  piecemeal 7840
2-Tag system 7688
Tail 6241
Tail entropy 967
Taiwan 2397
Takagi–Sugeno form 3243
Takagi–Sugeno fuzzy rulebase 4083
Takagi–Sugeno models 4092
Takagi–Sugeno neuro-fuzzy systems 6092
Take-profit orders 5415

Tangent bundle 5445
Tangent bundle reduction 5495
Tangent differentiability 6297
Tanh method 8120, 8121
TAR model
  n dimensional two-regime TAR model 444
  nested TAR model 444
Target 1502
Target genes 1184
Target knowledge system 4397
Target properties, QSAR modeling 7073
Targeted attacks 8263
Task allocation 2085, 7725
Task hierarchies 5190
Tatonnement 204
Taxonomy 7753–7755
Taylor-based Reynolds number 3652
Taylor–Couette experiment 3647
Taylor–Couette flow 6195–6197
  bifurcation 5937
Taylor–Couette instability 8941
Taylor hypothesis 8531
Taylor instability 8940
Taylor length 3652
Taylor vortices, liquids 5937
Taylor's conjecture 8527
TDMA coloring algorithm 2054
Technical analysis 3383
Technical capability 1975
Technical trading 5415
  support and resistance levels 5415
Technical trading rule (TTR) 3519
Technological knowledge 1974
Technological society 3281
Technology 7845
Telecom industry models 685
Teleoperation 3683, 3687, 3688
Teleportation
  quantum 7394
  quantum computing 7240
  quantum information processing 7368
Teleporting gates 7337
Teleseismic magnitude methods
  tsunami warning 2666
Telluric current 2621–2623
Temperament 1903
Temperature 695–697, 700, 701, 5001, 6382, 6805, 7591
  angoricity 5001
  dynamics 5000
  maximum temperature 7599
  temperature change 7591
Temperley–Lieb algebra 5295, 7111
Tempesti's loop 8072
Template 2110, 4414, 4415
Temporal disorder 3575
Temporal hierarchies 4593
Temporal networks 8263
Temporal physiological series 3798
Temporal regularization 4680
Temporal sequencing in amorphous computing 266
Temporary reputation effects 7659
Tensile crack 9904
Tension
  interfacial 532, 533, 545
  surface 533
Tensor product
  quantum mechanics 7373
Tent map 3505
Tenure 9099
TeraGrid 5761
Teramac 5872
Teraoka 3564
Term structures of interestrates, applications 8793
Terminal 4136
  complex 4140
  simple 4140
Terminal chain component 9235
Terminal set 4140
Terminus of replication 3894, 3908
Termite 3349
Ternary number 5740
Tesfatsions's labour market model 181
Tessellation 892, 894, 895, 4245
Test set 1816
Testing for linearity 446, 448, 452
Testing method 1158
  benchmark 1159
  Zachary's karate club 1159
Testing object 1827
Testing set 1827

Tetrahedral grids 7903
Tetrahedral partition 5807, 5815
12-Tetrahedral periodic minimizers 5970
Tetrahymena pyriformis 2147, 2166
  QSAR models 7081
Texas 7845
TF *see* Transcription factors
Thalamocortical system 1424
Thematic maps 583
Themes 580
Theorem
  Alexeyev's 8559
  Ambrose–Kakutani 8568
  averaging 8642
  Bernoulli 3615
  Bochner–Herglotz 8556
  Dirichlet's 3031, 3036
  Floquet's 8636
  flowbox 6311
  Foiaş–Stratila 8566
  Garden of Eden 780
  Hartman–Grobman 6313, 8635
  KAM 8641
  Khintchine–Groshev 3031, 3033, 3037
  Krein 6594
  Leonov's 8565
  Lusternik–Fet 6606
  Lyapunov 6591, 8637
  Marsden–Weinstein 6597
  Moore and Myhill 783
  mountain 6603
  Noether 3872, 6311, 6597
  Poincaré 6312
  Poincaré-Bendixson 8645
  Poincaré's 6313
  Poincaré's theorem 8636
  relative equilibria 6598
  Siegel's 8636
  singular reduction 6598
  Souriau 6597
  spectral 8557
  stable manifold 8644
  Sternberg's 6313
  structural stability 8644
  weak closure 8564
  Weyl 5067
Theory
  biological 1883
  Cantorized singularity 6325
  cognitive approaches to development 1885
  content-driven approaches 1887
  core knowledge theory 1884
  developmental agent models 1896
  developmental behavioral genetics 1884
  dynamic field theory 1878, 1893
  epigenetic robotics 1897
  ergodic theory 3029
  ethological theory 1884
  evolutionary developmental psychology 1884
  Fischer's 1903
  Fischer's model 1893
  gauge 3876
  historical 1883
  KAM 6323, 6326, 6592
  learning-theoretically inspired approaches 1886
  metric number theory 3029
  neo-Piagetian theorists 1886, 1903
  neural Darwinism 1427
  normal form 6326
  of coordination 1543
  of embodied action 1887
  of fuzzy sets 4062
  of games 8188
  of human development 1876
  of information processing 1886
  of mind 1877, 1904
  of neuronal group selection 1424, 1427, 1428
  of patterns and complexity 394
  of perceptions (CTP) 4005
  of value 2285
  parametrized KAM 6326
  perturbation 6310
  Piaget's 1877, 1878, 1880, 1881
  psycho-dynamically inspired theories 1884
  resource-dependent competition-support systems 1888
  singularity 6311, 6322, 6324
  social learning theory 1887
  socio-culturally inspired theories 1884
  systems and dynamically inspired theories 1887
  transactional development 1883

value 1428
Vygotsky's 1877, 1879, 1880, 1883
Werner's 1881
Yang–Mills 3876
Theory building 7850
Therapeutic automaton 5643–5645
Therapeutic effect 628, 630
Therapy Process Questionnaire [TPQ] 7999
Therapy process questionnaire (TPQ) 7998, 8000
Thermal blob, polymer physics 6816
Thermal conductivity 690, 700, 701, 1399, 2259, 4387
Thermal effects 7669
Thermal equilibrium 2813, 3370, 8752
Thermal gradient 3275
Thermal minority game 5590, 5595, 5597
Thermal physics 7825
Thermal properties 690
Thermal rounding 2032
Thermal stress
global changes 1097
mortality 1098
Thermocapillary effects 5584
Thermodynamic equilibrium 8381
Thermodynamic formalism 6940
ergodic theory 6941
Thermodynamic limit 8712
Thermodynamic phase 8712
Thermodynamic state 8752
Thermodynamics 2859, 3270, 3598, 4897, 5535, 5541, 8711, 9127, 9128
Brownian motion 3599
ecosystem 2705
first law of thermodynamics 9127
heat 3603
non-equilibrium states 3601
second law of thermodynamics 9127
small systems 3598
statistical mechanics 3598
work 3603
Thermoelectroelasticity 5922
Thermohaline circulation 3
Thermolysin 2132
Theta program, tsunami warning 2669
Thierrin–Vagner inverse 8297
Thin film equation 3633
Thinking, complex organizations 4564
Third integral 404
Thom 704
Thomas–Fermi approximation 7315, 8579
Thomas–Rashba interaction 9651
Thomson 5474
Thomson, James 5474
Thomson's principle 1406
Threat
internalize 3111
internalizing 3133
occasional 3133
probable maximum flood 3125
threat continuum 3110
Three-body collision 882
Three-body problem 3832, 5960, 6673
Three-body solution, eight shaped 5964
Three-cycle 6005, 6007
Three-dimensional equivariant minimizer 5968
Three-dimensional flow
in a bounded domain 5955
liquids 5939
steady bifurcation 5940
Three-dimensional game of life 4245
Three-dimensional microtrap, quantum computing 7223
Three-dimensional periodic orbits 6439
Three-phase traffic 9357
Three-phase traffic flow theory 9446
Three-phase traffic theory 9282, 9283, 9285, 9356
Three-scale model 1651
Three sisters volcanic center 9868
Three-terminal interferometer with an embedded QD 7301
Three traffic phases 9308
Three types theorem 3320
Threshold 164, 610, 1865, 2722, 6101, 6130, 7823, 8315
continuum percolation 6508
first 6130
Threshold autoregressive model 443, 3443
Threshold effects
critical mass 4617
transition thresholds 4618, 4620, 4621
Threshold function 1817
Threshold inversion 6130

Threshold model 2745–2747, 3476, 3480, 3481, 5326, 5333, 5336–5339, 5341
STAR model 5337, 5340
TAR model 5337, 5338, 5340, 5344
Threshold multiplication 332
Threshold theorem 7339
Threshold unit 1814, 1817
Threshold value, polydisperse fractures 6514
Threshold voter Automaton (TVA) 4507
Threshold voter automaton (TVA), *majority* 4508
Thresholding 8673
Thrombin 2134, 2145
Thrust earthquake 2484
Thümen's Isolated State 7957
TIB *see* Tsunami information bulletin
Tic-tac-toe automaton 5647–5651
Tick size 5408
Tidal bore 8465
Tide gauge 9596, 9635
Tideman's procedure 9956
Tides 2601
Tie indicator 5999
Tie variables 6001
Tierra 115, 334
Tight binding 5751, 5754, 7483
Tight-binding Hamiltonian 8601, 8608, 9138, 9140
Tight-binding potential
environment-dependent 9142, 9143
two-center 9139
Tight M-wavelet frame 591
Tightly-knit community (TKC) effect 5276
Tilings
acceptable for constrain 252
by local constraints 252
with Wang tiles 252
Time
complexity 6401
pre-movement 3153
re-scaling 6605
response 3153
Time- and slip-predictable models 2564
Time-averaged 695
Time-axis, hyperfinite 6307
Time complexity 3539, 7716, 7721, 8246
Time-correlation functions 700
Time delay 634, 636
Time-dependent Hamilton–Jacobi equations 4556
Time-dependent vectorfields 6628
Time-discrete dynamical systems 161
Time discrete scale invariance 7578, 7579
Time division multiple access (TDMA) 2054
Time DSI 7573
Time evolution 1019
Time-evolution operator 7389
quantum simulation 7385
Time-fractional diffusion equation 5220
Time frequency analysis 2387, 2392
Time-frequency analysis 2389, 2434
Time frequency distributions (TFD) 7994, 7995, 8000
Time graphs vs. eigenvalues or frequency response 8952
Time headways 9382
Time horizon 7841
Time-independent Schrödinger equation 7128
Time invariant 3526
Time-of-flight 9684
Time ordered product 6745
Time-predictable 9840
Time regularized long-wave (TRLW) equation 8116
Time reversal 880
Time-reversal symmetry 2809
Time reversibility 5326, 5332, 5342
Time scales 1873, 1875, 7603, 9779
Time series 390, 6252, 7996, 7997, 7999
biological fractals 3793
of traffic variables 9501
sparsely sampled 3585
Time-series 280
Time series analysis 3436
nonlinear 3436
stochastic 3582
Time-series data 4908
Time-series processes 3464
AR 3456, 3461
ARCH 3456, 3472
ARMA 3461
autocorrelation 3464
covariation 3464
GARCH 3456, 3461
Time signal measurements 3575

Time-space diagram 3538
TIME STEP 9036
Time-stepped simulation 162
Time-to-market 1975
Time-to-volume 1975
Time-varying 3527
Time-varying expected return 3449, 3454
Time-varying parameter model 5328, 5344
Time-varying transition probabilities 2750
Time-window component 2487
Timescales 1542
Timing jitter 6154
TIP *see* Trapping invasion percolation (TIP)
Tissue engineering 417
Tissue lesioning 629
Tissue P system 5525
Tisza, Laszlo 2864
Tit-For-Tat 8382
TiTAN 2076
Titin 7039
Titin elasticity 7040
Titin I27 domain 7035
TMG 5590
TNS action 3101
Toda lattice 279
Toddler 1901
Toeplitz space 2237
Toffoli gate, quantum computation 7379
Toggle circuits 563
Tokachi-Oki earthquake 4273
Tolerance class 4417
Tolerance relation 1796, 4413, 4416
Tom Thumb algorithm 8076
Tomita 8840
Tomography
  computerized 9173
  conventional 9173
  seismic 9172
  X-ray 9173
Tonks–Girardeau gas 9680
Tonotopic maps 3341
Tool use 1879
Top cycle 9952
Top-down causation 8046
Topographic maps 3350, 3351
Topography 3825
Topological automata 8817
Topological chirality indices, QSAR modeling 7075
Topological complexity 4835
Topological conjugacy 8893
Topological defect 1277, 2025, 9680
Topological divide 4349
Topological dynamical system 5452
Topological dynamics 3054
Topological effect
  entangled polymer dynamics 6821
  polymer physics 6812
Topological entropy 4737, 8896
  dimension-like characteristics ergodic theory 3018
Topological genericity 954
Topological group, ergodic theory 3064
Topological index (TI) 1303, 2197, 2203, 2204, 6838
  Balaban 2205, 2208
  Bonchev 2204
  centric 2204
  Diudea 2206
  Estrada generalized 2207
  Gordon–Scantlebury 2204
  Gutman 2206
  Gálvez 2208
  Hosoya 2204
  information theoretic 2205
  interpretation 2210
  Kier–Hall 2205
  mean square distance 2205
  orthogonal 2208
  path-count-based 2206
  Platt 2204
  Randić 2204
  reverse engineering 2211
  Rücker 2206
  Schultz 2204
  SMILES notation based 2208
  triplet 2206
  variable 2208
  Wiener 2203
  Zagreb 2204
Topological information 4827
Topological invariants 9269
Topological models 724
Topological order 2026

Topological pressure, ergodic theory 3018
Topological quantum computing 7390
Topological reconfigurations 8529
Topological rigidity 3104
Topological sort 8251
Topological space 4352
Topological trapping 5081
Topologically transitive 9234
Topology 724, 3831, 3930, 4313
  collapse 505
  generation 4940
  modeling 4938
  of uniform convergence 1030
  rules 8843
  social network 180
  strong and weak 5442
  usual 1030
  weak 1030
Toppling 8035
Toral automorphisms 997
Torque density 8106, 8107
Torque dipole 8107, 8108
Torres Quevedo, Leonardo 5475
Torsade de pointes (TdP) 2125, 2166
Torsional forces 2841
Tortuosity 2687
Torus 1932, 5084, 8930
  invariant 1934
  lower-dimensional 1943
  maximal 1943
  resonant 1943
Total adjacency 4829
Total clearance 2155
Total-ordering, definition 6288
Total relation 9234
Total scattering coefficient 7915
Total scattering cross-section 7915
Total stability 5081
Total variation (TV) 10038
Total variation (TV) restoration 10033
Total walk count 4833
Totalistic 8030
Totally ergodic 2923
Totally ordered field 6293
Tourism 7842
Tourism demand 7845
Tournament 6249, 9946
  majority 9943, 9945, 9949
Tournament equilibrium 9953, 9958
Tournament matrix 6249
Tournament solutions 9945, 9949, 9952
Tower transformations 3060
Tower, Rokhlin 8564
Toxicity of organic chemicals 2145
Toxicology, QSAR models 7081
Toy chemistry 336
Trace formula 7164, 7168
Trace intersections density, fracture networks 6514
Traces 6514
  of the attractor 9233
Tracking 4687, 9079
  adaptive tracking 9079
  dynamic inversion 9079
  tracking via internal models 9079
Tractability, neoclassical economics 179
Trade deficit 2818
Trade size 5408
Traders, market microstructure 5397
Trading mechanism, order-driven 5393
Trading network game, Tesfatsions's 181
Trading process, tree diagram 5399
Trading volume 98, 5407
  NYSE 5407
Traditional 8955
Traffic assignment in a traffic network 1177
Traffic breakdown 9296, 9297, 9302, 9316, 9355
Traffic congestion 1177, 3945, 9403
Traffic control 3945, 5791
  network-oriented traffic control systems 3956
Traffic control centers 9526
Traffic control parameters 9379
Traffic data, phase space 3589
Traffic flow 1177
  stochastic processes 3587
Traffic flow interruption 9380
Traffic flow interruption effect 9309
Traffic flow interruption intervals 9398
Traffic flow management 234
Traffic flow model 3950, 9303
Traffic flow theory 1179, 9283
Traffic growth 239
Traffic jam 1179, 3945

Traffic management 1178
Traffic model 876
Traffic modeling 9445
  macroscopic 9446
  mesoscopic 9446
  microscopic 9446
Traffic phase definitions 9357
Traffic routing 9430
  defined 9430
Traffic states 3589
Trailing spiral arm 414
Training 2143, 8377
Training object 1827
Training set 1827
Trajectory 1552, 3263, 4904, 6489
  Boolean networks 1199
  Brownian 3785
  convergence 1203
  coordinating elements 1552
  optimal 1502
Trajectory formation 1557
Transaction cost 6978
Transactional relationship 1901
Transcranial magnetic stimulation (TMS) 1437, 8004
Transcription 418, 544, 3894, 3968, 4161, 4169
  antisense 3969
  non-coding RNA 3967
Transcription factors (TF) 528, 529, 535, 539, 540, 542, 543, 721, 1182, 5661
  autoregulatory 528, 529, 535
Transcriptional regulation network 5656, 5661
Transcriptome 1181
Transcriptomics 5537, 9099
Transcritical bifurcation 5305
Transducing automaton 5645
Transfer entropy 1434
Transfer function 1813, 1817, 2382, 2475, 8955
Transfer map 3100
Transfer matrices 1410
Transfer operator
  multiwavelet 5844
  Ruelle's Perron–Frobenius theorem 6948
Transfer principle 2940
Transfer, heat 6962
Transferable utility game 1511
Transfinite fractals 3933
Transformation 6315, 6596, 6653
  bond-to-site 6585
  canonical 6320
  elementary 1870
  equivariant 6318
  finitely fixed 4991
  isomorphic to Bernoulli shifts 4988
  loosely Bernoulli 4991
  minimal self joining 4989
  non-invertible 4992
  normalizing 6310
  not isomorphic to Bernoulli shifts 4989
  pignistic 6934
  preliminary 3303
  scale 3861, 3864, 3877
  σ-preserving 6318
  star-triangle 6580
  $[T; T^{-1}]$ 4985
Transformation process, cutting and stacking 4990
Transforming 4188
Transfractals 3933
  transfractal dimension 3933
Transience 7522
Transient deformation 1701
Transient growth 15
Transient lengths 65, 70
Transient pressure disturbance 9900
Transient random walk 1396
Transient states 64
Transient time 2721
TRANSIMS 162
Transistor 420, 7401
Transition 620, 621, 1132, 1898, 3863, 3864, 7823
  between different traffic states 9389
  between existing traffic phases 9389
  between traffic phases 9389
  continuous 620
  depinning 1132
  electron shelving detection 7230
  first-order 620
  hybrid 621, 622
Transition diagram 5634
Transition function 169, 3234, 5634
Transition interval 2740
Transition metal 1317, 1318
  adsorbate interaction 1318

center 1318
d-band 1318
s-band 1317
Transition operator, multiwavelet 5844
Transition probability 2749, 7521, 9421
Transition rules 5874
Transition semigroup 6231
Transition state 1314, 1317, 1321, 1324, 1325, 1330
early 1314
late 1314
structural analysis 1324, 1325, 1330
Transition-state theory 1312
Transition zone
percolation 6513
Transitional metal
early 1307
late 1307
Transitive points 9234
Transitive system 9248
Transitivity 999, 6005, 6007, 8270, 8284, 8324, 8326, 9225
Translation 418
Translation vector 46
Transmission coefficient 4966, 4970
Transparency 5409
Transport 6961, 8600, 8606
heat 6961
mass 6961
spin-dependent transport 8598
Transport coefficients 990–992, 1000, 1001, 1008, 1398, 4386, 4388
continuum percolation 6507
Transport processes, geological 6505
Transposable elements 4164
defence against 4165
derived cellular features 4165
DNA transposons 4164
dynamic 4165
gene duplication 4166
LINE1 4164, 4166
regulatory network 4166
retrotransposons 4164
Transversal 6311, 8644
Transversal gates 7337
Transversality condition 5305, 6276
Trap frequency, quantum computing 7222
Trap models 4231
Trapped ion
distributed quantum information 7244
quantum computing 7218
Trapping invasion percolation (TIP) 4948
threshold 4950
universality class 4949
Trapping threshold, invasion percolation 4955
Traps 7523
Trash-collecting 3691
Travel decision behavior 1177
Travel time curve, infrasonic wave 4842
Traveling salesman problem 6408
Traveling wave solutions 5144, 5146
Traveling waves 2721
Traversability, component 2739
Treadmilling 1746, 1752, 1765, 1766
Treadmilling actin 1746
actin filament 1746
Treatment effect models 5560
average treatment effect 5561
potential outcome model 5559
Rubin causal model 5561
Treatment effectiveness 8001
Treatment evaluation 5555
counterfactual 5557
Treatment of mental health diagnosis
modeling 4573
Treatment of physical conditions
modeling 4571
Tree 1254, 1935, 4487, 4488, 6668
algorithm 8694
cluster 1939
diagram of the trading process 5399
finitely determined transformation 4992
fractal dimension 3783
height 71, 72
matrix 3740
number 6668
renormalized 1942
self-energy cluster 1939
tree of life 3865
value 6668
very week Bernoulli transformation 4992
Trefoil knots 7045
Treiber et al. intelligent driver traffic flow model 9335

Tremor 2610
Trench 9583, 9585
Trend 9836
  detections 3577
Trend-impact analysis 7835
Tri-gate MOSFET 7408
Triad significance profile (TSP) of networks 5698
Triadic 6007
Triads 8241, 8270
Triangle 8324
  density distribution 7516
Triangular connectivity 8254
Triangular lattice 8714, 8715, 8726, 9680
Triangular network 8254
Triangular tessellation 892, 893
Triangulation refinement 5801
  Alfeld refinement 5807
  Clough–Tocher 5801
  Powell–Sabin 5801
  quadrangulation 5807
  Wang refinement 5802
  Worsey–Farin refinement 5814
Trigger 7605
Triggering 6958, 6964
  dynamic 2600
Triggering mechanism 6957
Trip distribution 9451
  gravity model 9451
Tripartite scheme 1543
  hierarchy of timescales 1543
Triple density similarity measures 7453
Triple similarity measures 7470
Triplestores 8085
Tristability, networks 1193
Trivial symmetry 5972
Triviality conjecture 6665
TRNA sequence, entropy 496
Trochoid 9978
Trophic roles 2732
Tropic cascade 2718
Trotter formula 7099
  quantum simulation 7385
True anomaly 6676
True in GM 4401
Truncated power law 2716, 5035
Truncation 6310, 6315, 6655
Trust 8088
  evolution of 5686
Trustworthiness 4048, 4049
Truth set 4400
Truth table
  CNOT quantum gate 7234
Truthful scheduling 5518
Truthfulness 5512, 5514, 5516
Trypanosoma cruzi 2147
Tsallis entropy 5260
Tschirnhaus-transformation 712
TSK model 4890
TSP *see* Triad significance profile
TST *see* Two-scale similarity transform
Tsunameters 9635
Tsunami 1220, 2451, 2476, 2483, 2491, 2658, 2659, 8113, 8121, 8465, 8505, 8855, 9578, 10047, 10055
  infrasound 4839, 4846
  interaction with the coastline 4847
  Miyagi-Oki earthquake 4846
  propagation 2659
  Sumatra earthquake 4846
  tsunami wave heights 9587
Tsunami alarms 2492
Tsunami early warning 2491, 2493
Tsunami early warning system 9582, 9590, 9592
Tsunami earthquakes 2484, 2660, 2669, 9577, 9594, 9639
  model 9588
  slow character 9580
Tsunami forecasting and warning 1220
Tsunami generation 9633
Tsunami hazard 2494
Tsunami information bulletin (TIB) 2661
Tsunami infrasound 4845, 4848
Tsunami inundation 1220
Tsunami magnitude 9639
Tsunami observations 9635
Tsunami propagation 9634
Tsunami warning 2484
  a fixed regional bulletin 2661
  an expanding watch/warning bulletin 2661
  future directions 2671
  local earthquake magnitude methods 2665
  mantle magnitude (M_m) method 2667
  M_{wp} method 2666

rapid estimates 2657
retroactive performance 2662
rupture slowness estimation 2669
teleseismic magnitude methods 2666
theta program 2669
Tsunami warning center operations 2660
Tsunami warning centers 2660
Tsunami waves 2659
Tsunamigenesis 9585, 9586
Tsunamigenic earthquake 2367, 2660, 9578, 9639
Tube model, entangled polymer dynamics 6822
Tumor cell-extracellular matrix (ECM) interactions 914
Tumor cell migration 914
effect of heterogeneous environments 918
influence of the extracellular matrix 917
Tumor invasion
cell processes 914
cellular automata model 914
cellular scale 915
influence of metabolic changes 919
metabolism and acidosis 919
models 916
molecular scale 915
tissue scale 915
Tumor locus 919
Tumor morphology, invasive 914
Tumor spheroid, multicellular 917
Tuned damper 2333
Tuning parameter 5021
Tunneling between wells 6756
Tunneling energy 9692
Tunneling junction 9715
Tunneling spectroscopy 7311, 7319
linear transport regime 7311
non-linear transport regime 7311
source-drain bias 7311
Tuple space *see* Linda coordination model
Turbulence 23, 3643, 3831, 6489, 8947
cluster 6490
Eddy 6489
Navier–Stokes equations 5954
nonscaling behavior 3594
rupture 6490
stick-slip instability 6490
stochastic processes 3590
vortex cascade 6489
Turbulence modeling 3650
Turbulent flow 6489
Turbulent regime 9421
Turbulent velocity fluctuations 7019
Turing equivalent 1376
Turing machine 169, 271, 279, 280, 287, 288, 745, 800, 904, 2090, 3354, 5467, 6388, 6401, 7155, 7156, 7158, 7161, 7162, 7683, 7705, 8842, 9127, 9129, 9159, 9162
halting problem 9160
in artificial chemistry 333
logspace 6401
quantum 7155
reversible 7683, 7705
universal RTM 7709
Turing patterns 3349
Turing power 288
Turing structures 3624
Turing, Alan M. 923, 3349, 4134, 5467, 5478, 7251, 8940, 8942
Turing, intermediate 764
Turing's artificial chemistry 327
Turkey 2398
Turning points 6748
Tutte polynomial 7111
TV wavelet inpainting 10042, 10045
TWINKLE 5478
Twist 9271
Twist mappings 5079, 5087
Twist vector 2636
Two-ballot procedure 9948
Two-body collision 882
Two-body problem 5960, 6436
Two channel filter bank 10023
Two-class society 2819
Two-class structure 2813
Two-cocycle 5486
Two-cocycle identity 5486
Two-dimensional 3557
2D Brownian motion 8727
2D chirality, QSAR modeling 7074
Two dimensional electron gas 8608
Two-dimensional electron gas 1018
2D electron layer 8581
Two dimensional electron waveguide 7292
Two-dimensional flow, liquids 5930, 5941
Two-dimensional KdV 9994

2D layer 8584
2D quantum gravity 8724
2D spin glass 8713
Two-dimensional turbulence 8726
2D turbulence 8713
Two-dimensional wave equation 9986
Two-forms, curvature 5505
Two fracton relaxation 2258
Two-frequency mutual coherence function 7921
Two-gene circuit, expansion 1190
Two-level correlation function 7514, 7515
Two-mode data 8272
Two-mode network 8258
Two node simulation 156
Two-phase structure 5589
Two-qubit controlled Z-gate 7233
Two-qubit gate 7237
Two-qubit interaction 7231
Two-round procedure 9948, 9955
Two-scale similarity transform 5853
Two-sided 3545
Two-sided matching model 4097, 9654
  continuous 9654
  discret 9654
  hybrid 9655
Two-sphere 2009
Two-spin cluster 8595
Two-state quantum system, quantum computing 7220
Two-valley system 8579, 8585
Two-valuedness 3871
Twofold integral 232
Tygar 5472
Type 1 diabetes 9097
Type-1/type-2 fuzzy sets 4411
Type-2 fuzzy logic 4009
Type-2 fuzzy set 4011, 7739
Type-2 fuzzy systems 4011, 4089
  generalized 4013
Type of an agent 4788
Type-reduced set, interval type-2 fuzzy logic system 4016
Typogenetics used in artificial chemistry 333
Tyrosine kinase 2121

U

U-m 7460, 7463, 7466, 7468
U-Molecule 7460
U-Shaped growth 1890
U-Statistics 6260, 6261, 6266
UC models 2752
UE assignment 9437
  Beckmann formulation 9437
UE model 9439
  Beckmann's 9439
  mathematical program 9439
Uehling–Uhlenbeck equation 7818
Ugi multicomponent reaction 2117
Ulam, Stanislaw 779, 849, 7155
ULF electric signals 2569
  Cox regression model 2570
  RELM testing center 2570
  self-exciting (Hawkes type) process 2570
Ultimate-reals, construction of 6292
Ultimatum game 6648
Ultra-filter, infinite set 6292
Ultrasensitivity 485
Umeo 3542, 3552, 3553, 3564
Umeo, Maeda and Hongyo 3549
UMF *see* Upper membership function
Unanimity 9945
Unanimity game 4105
Unbalanced panel data 2780
Unbiased fuzzy-valued estimator 8688
Unbiased random walk 8714
Uncanny valley 1121
Uncertainty 2450, 3860, 4317, 4318, 6568, 7833, 8315, 9434
  aleatory 4049
  environmental 7838
  epistemic 4049
  extrinsic 4317
  future 7837
  fuzzy logic 4009–4013, 4015, 4017, 4513
  interdependence-induced 7839
  intrinsic 4317, 4318
  management of 7849
  mixed probabilistic/non-probabilistic 4048
  neoclassical economics 178
  strategic 7836
  type-2 fuzzy logic 4015

Uncertainty principle 3351
Uncertainty propagation 6936
Unconventional computing 9705
Uncorrelated process (white noise) 369
Uncoupled 5182
Uncoupled mode-space NEGF 7485
Uncovered elements 9958
Uncovered interest parity 5426
Uncovered set 9952
Undecidability 762
Undecidable 762, 764, 5468, 5473
Undecidable mechanical simulation problem 5473
Undecimated wavelet transform 6358
Under-dispersed 9838
Underground explosions 2476
Underground nuclear explosion 2485
Undetermined form 6290
Undirected graph 5268
Undirected motif 507
Undular bore 8116, 9989
Unearned income 2285
Unemployment
  acceleration 3301
  rate 5332, 5333, 5341–5343
Unequi-spaced FFT 1729
Unfoldases 7042
Unfolding 711, 5777
Unfolding dimension 703
Unfolding morphism 703
Unfolding of homoclinic tangencies 964
Unified modeling language (UML) 353
Uniform continuous function 3256
Uniform convergent sequence 3256
Uniform graph 8214
Uniform hyperbolicity 4724
Uniform measure 2982
Uniform Nash equilibrium 7633
Uniform random sampling 5536, 5542, 5544
Uniform sampling 5545
Uniform spanning tree 8710
Uniform value 7633
Uniformizing map 8710
Uniformly distributed 2938
Uniformly elliptic 7521
Uniformly expanding maps 959
Uniformly expanding Markov map 6950
Uniformly hyperbolic diffeomorphisms 959
Uniformly hyperbolic systems 2968, 4730
  Anosov Systems 2966
  Axiom A Systems 2967
  Geodesic Flow on Manifold of Negative Curvature 2965
  Horocycle Flow 2966
  Markov Partitions and Coding 2966
Uniformly rigid 9237
Unimodal maps 962
Uninformed traders 5397
Unintended consequences 7052
  secondary crisis 7053
Union 4148, 4877–4879
  type-2 fuzzy sets 4012
Union of two fuzzy sets 4067
Unipotent matrix 3105
Uniquely ergodic 2937, 3105, 5457
Uniqueness 7829
Unit quaternion 58
Unit shell 7429
Unitarily equivalent 5464
Unitary 7253
Unitary extension principle 6897
Unitary operator 5464
  nonsingular system 3070
Unitary representations, locally compact groups 3079
United States 2398
Universal 703, 712
  quality 906
Universal amplitude 7001
Universal approximation 3354–3356
Universal approximation theorem 1818
Universal approximator 2143
Universal behavior 1443
Universal belief space 430, 431
Universal CA 8886
Universal computation 1384, 3354, 5869
Universal computer 285
Universal constructor 5439
Universal formula for stabilization 8625
Universal history 3273
Universal model for computation 169
Universal modeling schema 8190
Universal property 6407, 8711
Universal quantum computation 7379

Universal representation theorem for expansive cellular automata 2987
Universal scaling function 5022, 5035
Universal set 4061
Universal threshold 8674
Universal Turing machine 121, 7709, 8069, 8191
Universal type space 431
Universality 758, 903, 905, 1126, 1402, 1404, 2821, 3314, 3743, 5021, 5030, 5525, 6576, 6772, 6804, 7167, 7823, 9944
  computational 901, 903
  consequences 905
  intrinsic 758, 901, 906
Universality class 609, 2023, 5026, 5035, 6412, 6772, 8712
Universality theorem 7179
Universe 1865, 3272, 4877
Universe of discourse 4061
Universe of objects 4417
Unmanned air vehicle (UAV) 3688, 4758
Unmanned autonomous systems 5615
Unmanned underwater vehicles – UUVs 3688
Unobserved components (UC) model 2751, 5330, 5335, 5337
Unobserved effect 2772
Unobserved heterogeneity 2770, 5565
  gamma distribution 5566
  independently 5566
  mixture model 5567
  omitted variable 5561
Unobserved shocks 2771
Unperturbed system 6310
Unpredictability 4909
Unpredictable 8869, 8870, 8885, 8886
Unstability in the quadratic family 970
Unstable and stable manifolds 990
Unstable manifold 2265, 4516, 4727
Unstable mode 8930
Unstable voids 611
Unsupervised feature selection 5353
Unsupervised learning 1816, 1822, 1992, 2162, 4873
Untrusted code, mobile agents 5610
Update 10010, 10022
  parallel 3158
  synchronous 3158
Update order 165
Upper approximation 4425, 7776, 7788, 7801
  singleton upper approximation 7802
  subset upper approximation 7802
Upper critical dimension 5028
Upper density 2947
Upper fractal limit 662
Upper membership function (UMF) 4013
Upper semicontinuous 8682
Upperbound 1683
Ur-element, definition 6302
Urban 3126
Urban dynamics 8997
Urban simulation models 7954
Urban traffic network 1178
Urbanism 7953
UrEDAS 2398
URI (uniform resource identifier) 8085
US business cycle 442
US Census Bureau 2815
U.S. Centers for Disease Control and Prevention (CDC) 4578
US geological Survey (USGS) 9866
US government bond market 5395
USA 7846, 7848
User equilibrium 9434
  traffic assignment 9437
  traffic assignment, incremental 9437
Usuality constraint, fuzzy logic 3999
USV 8795
UTA (urban traffic analysis) 9507
Utilitarian Borda-score criterion 3333
Utilitarianism 9936, 9939
Utility 4855
Utility function 1669
  expected utility 7536
Utility of voting 9939
Utility problem 5188

V

Vacant dislocations 2649
Vachaspati–Vilenkin algorithm 1648
Vacuum cleaners 3690
Vagueness 4060
  type-2 fuzzy logic 4011
Valence force field 5753
Valente, Tom 8273

Valid computations 817
Validation 174, 216, 217, 3303, 6802, 7017, 7803, 9000, 9001, 9003
Validation methods 9004, 9012
Validation set 1816
Validity 9000, 9001, 9003
Valley-degeneracy 5751
Valley-splitting 5775
Valuation 3457, 3458
  constant P/E model 3459
  Donaldson–Kamstra (DK) method 3457, 3460, 3461, 3464
  free-cash-flow method 3459, 3461
  fundamental 3456, 3458, 3459
  Gordon growth model 3459, 3461
  relative value method 3459
  residual income method 3459, 3461
Valuation functions 5517, 5521
Value 7631
  critical 6600
Value-at-risk 3429
Value chain
  customer-supplier 7842, 7849
  jet ski 7847
Value chain components 7849
Value-chain management structure, generic 7846
Value-chain participants 7845
Value function 1503, 6219, 9105
Value stocks 3391
Van-Allen belts 32
Van-Allen radiation belts 22
Van-der-Corput trick 3088
Van der Pol 7603
Van-der-Pol equation 3838
Van-der-Pol oscillator 5725
Van der Waals 696, 697, 699
Van-der-Waals equation 6664
Van-der-Waals forces 3633
Van der Waerden's theorem 9244
Van-Hove intermediate scattering function 1004
  incomplete 1005
VAN method 2622–2624
  conductive channel model 2630
  Ioannina station 2624
  seismic electric signal (SES) 2623
  seismic electric Signals (SES) 2622
  selectivity 2624
  selectivity map 2624
  SES 2632
  SES activity 2623, 2624, 2626
  VAN relation 2624
  VAN-type monitoring 2625
Van Nuys 7-story hotel 2389
Van Nuys building 2393
Vanishing 6290
Vanishing moment 591, 10034
Vanishing points 58
Vanishing viscosity method 6218, 6220, 6225
Vannevar Bush 273
van't Hoff–Arrhenius equation 1312
Vapor pressure 2144
VAR *see* Vector autoregressive model
VAR model 5329, 5330, 5344
Variability 1898, 1899, 1904
Variable 4491
  auxiliary constant 7843
  constraint 3998
  converter 7843
  hidden 4491
  locally meaningful 7850
  stochastic 3862
  type-2 fuzzy sets 4013
Variable focus liquid lens 5585
Variable sampling 4861
Variance 6145
  and moments 3317
Variance gamma distribution 3422
Variants of the minority game 5596
  variants 5589
Variation
  of parameters 1032, 1037
  seasonal 7843
Variation order 1849
Variational equations 6430, 6433
Variational formulation, Navier–Stokes equations 5931
Variational iteration method 5144, 5165, 5168
Variational methods 7347
Variational principle 958, 966, 3041
  ergodic theory 6945, 6946, 6949
Variational QQSPR 7456
Variations formula 1032
Varopoulos N 787
Varshavsky 3547

VC-dimension 5212, 5213, 5216
VCA 5778
VCG 5510, 5514, 5519
Vector autoregression (VAR) model 2750
Vector autoregressive (VAR) model 5402
Vector-borne tropical infectious diseases, increase by climate change 1101
Vector field 1030, 2261, 3257, 4517
  autonomous 6590
  C^0 equivalent 6698
  equivariant 6597
  germ-equivalent 6698
Vector moving average (VMA) representation 5402
Vector of cost shares 1664
  excess 1667
  stable 1664
Vector of final system's state 1622
Vector of initial state 1622
Vector payoff 7637
Vector potentials 9275
Vector semispace 7428
Vector space 703
Vectorization 5757
Vehicle
  candidate 9735
  field 9735
Vehicle cluster in free flow at the bottleneck 9285
Vehicle communication protocols 2054
Vehicle stops 9457
Vehicular traffic 1179, 9282, 9357
Vehicular traffic congestion 9304
Vehicular traffic management 1180
Velocity broadband record 2482
Velocity estimation 4767, 4770
Velocity increments 3651
Velocity pulse
  earthquake 2539
  microearthquakes 2546, 2547
  slow initial phas 2544
Velocity-strengthening 10047, 10052, 10055
Velocity Verlet algorithm 695
Velocity waveforms, large earthquakes 2543
Velocity-weakening 10047, 10051
Venice's Carnival 9419
Venn diagram 8268
  entropy 498
Vensim 9015
Vent 9889
Venttsel condition 6226
Verhulst equation 1889, 6616, 6617
Verhulst, François 7843
Verifiable information 8127
Verification 174, 4855
Verification functions 6271, 6277
Veristic constraint, fuzzy logic 3999
Versal 713
Versal unfolding 6632
Vertex 6240, 6241, 8246, 8267, 8270
  adjacent 6241
  and edge orbit's information 4827
  degree 2201, 8252
  in-degree 8252
  incident 6240
  initial 6241
  out-degree 8252
  reachable 8250
  strongly connected 8250
  terminal 6241
  weakly connected 8250
Vertex eccentricity 4832
Vertex orbits 4827
Vertex similarity 1144
Vertical hostility 5610
Vertical projection 5503
Vertical slice, type-2 fuzzy logic 4013
Vertical space, definition 5502
Vertical stability 6432
Vertical transmission 3213
Vertical vectors, definition 5504
Very long baseline interferometry 1691
Vibrational cooling, quantum computing 7226
Vibrational coupling, quantum computing 7231
Vibrational density of states 2255
Vibrational mode, quantum computing 7232
Vibrational quanta 7239
Vibrational state decoherence 7239
Vibrio fischeri 2145, 2147
Victoria 3135
Video games 1340
Video shot 5742
Video tracking 6479
VIM 5144, 5146, 5153, 5156, 5158
Violent relaxation 389

Viral protein linear (VPL) actuator 9751
Virial series 6664
Virtual characters 1335
Virtual enterprise network (VEN) 7846
Virtual fidelity 1338
Virtual human(s) 1336
Virtual machines (VM) 5607
Virtual screening 7079
  QSAR based 7071
Virtual screening of chemical libraries 2120
Virtual world(s) 1335
Viscoelastic 9790
Viscoelastic phase separation, polymeric fluid 6828
Viscoelasticity, polymer physics 6819
Viscosity 1611, 9766, 9771, 9776, 9787
  bacterial colonies 3788
  Newtonian 9787
  non-Newtonian 9787
  of gels 6555
Viscosity solutions 1505, 4547, 6217–6219, 6272, 6279
  Hamilton–Jacobi equations 4546, 4558
  semicontinuous 6225
Viscosity subsolution 6223
  L^p 6226
Viscosity test functions
  Hamilton–Jacobi equations 4551
Viscous expansion time 6960
Viscous fingering 3725, 6481
Vision 8200, 8201, 9415
Visual complexity 1256
Visual cortex 3352
Visual servo control 43
  2.5D 44
  homography-based 44
  image-based 44
  position-based 44
Visual stimulation paradigm 8004
Visualization 2357, 2358, 2360, 2366, 8274, 8283
Vlasov–Fokker–Planck equation 5243
VM *see* Virtual machines
VMA *see* Vector moving average representation
Vogel–Fulcher–Tamman law 4212
Voice-onset time 2315
Void instability 613
Volatile 6955, 9785, 9790
Volatility 2247, 3439, 3476, 3477, 3479, 3486, 3492–3494, 3498, 5412
Volatility and financial market innovation
  variance swaps 8807
  VIX index 8807
  volatility derivatives 8807
Volatility and jump risk premium, empirical evidence 8793
Volatility clustering 96, 3376, 3425, 3818, 4115
Volatility risk premium
  open questions 8808
Volcanic 9785
Volcanic activity, eruption signature 4851
Volcanic conduit 9769
Volcanic crisis 9862, 9865
Volcanic eruption 3824
  spectroscopy 4851
Volcanic eruption detection 4852
Volcanic explosivity index (VEI) 9785, 9832
Volcanic sound, modeling 4849
Volcano 1222, 1314
  infrasound 4839
  Kilauea 8855
  Mount St. Helens 9765, 9773
  Santiaguito 9773
  Shiveluch 9777
  Soufrière hills 9765
Volcano disaster assistance program (VDAP) 9867
Volcano hazards 9862
Volcano infrasound 4848, 4852
  basic principle 4848
Volcano monitoring 9862, 9869
  baseline monitoring 9862
Volcano plot 1314, 1316
Volcano risk 9862
Volcano-seismic signal 9818
Volcano seismicity 9900
Volcano seismology 1222, 9901
Volcano unrest 9862, 9864, 9866, 9869, 9870
Vollmar 3542, 3552
Voltage 6102
Voltage distribution 3745
Volterra series 1029, 1037
Volume preserving 6318
Volume, excluded 6505
Voluntary movements 1557

Voluntary participation 1671
von Kármán-type ACF 7920
von Neumann 418, 7251, 8869, 8878, 8885
von Neumann algebras, ergodic theory 3078
von Neumann architecture 2091
von Neumann machine *see also* Assembler automaton
  parallel 334
von Neumann mean ergodic theorem 3058
von Neumann–Morgenstern 1519
von Neumann neighborhood 872
von Neumann theorem 6232
von Neumann, John 779, 780, 783, 784, 1510, 2861, 7155, 8067, 8381
  universal Constructor 8067
Voronoi cells 7898
Voronoi diagram 3349, 9710
Voronoi tesselation 1802
Vortex 9689
Vortex core 9693
Vortex filament 3648
Vortex glass 6413
Vortex pair 1274, 9691
Vortex solutions 3646
Vortex structures 3645
Vortex unbinding 1275
Vortices 1272, 2025, 6472
Vorticity alignment 3646
Vorticity structure 8727
VOS model 1642, 1649
Voter model 6384, 7577, 8385
Voting 5320
0–99 voting 9954
Voting cycle 9935
Voting game 1521
  simple majority 1522
Voting procedure 9943
Voting rule 776, 9934, 9935, 9937–9939
Voting systems 4097
Voting theory 9943
Vowel 2310
Vulcanian 9807
Vulcanization 7822
Vulnerability 3110, 3113
  exit route 3114
Vygotsky 1884, 1896

W

W-Phase 9582
W-State, quantum computing 7244
Waddington's epigenetic landscape 1193
  formalism 1194
Wage employment 2285
Waiting time 4231
Waiting time PDF 5220
Waksman 3539, 3540
Wald maximin criterion 6932
Walk 2201, 2211, 4833, 6239, 6241, 6784, 8250
  biological fractals 3786
  closed 6241, 8250
  length 8250
  random 3860
Walrus 1256
Wandering 5458
Wandering effect 7920
Wang 5472, 5478
Wang tile 5635, 9159, 9162
  aperiodic 9159
  deterministic 9165
  periodic tiling 9159
  tiling 9159
  tiling problem 9160, 9163
War of attrition 3188, 8652
Ward identity 8581
Wardrop's first principle 9436
Warehouseman's problem 5470
Warning 3111, 9862
  active warning image 3134
  cyclone 3128
  spreading 3137
  timeliness 3134
  traditional weather warning 3136
Warning map 3136
Warranty 8127
Wasserman, Stanley 704, 8269, 8275
Water 852, 9874
Water availability 7599
Water resource 1073
Water-SWCNT system 696
Water tank experiments 8121
Water wave equations 9974, 9980
Water wave observations 8121, 8122
Water wave problem 6721

Watson's theorem 6740
Watts, Duncan 8269, 8277
Wave 2476, 8115, 8118, 8122, 9892, 9969
density 3154
"freak" or "giant" 8498
function 3872
linear 8480
longitudinal wave 2476
nonlinear 8480
ring 886
shock 3154
spiral 886
traveling 8480
Wave chaos 7172
Wave dispersion 8113, 8115, 8116, 8123
Wave equation 1487, 1492, 1493, 2042, 2043, 6121, 6231, 6709
boundary control 4806
controllability 4806, 4808
nonlinear 2048
strongly stable 1494
(uniformly) exponentially stable 1494
Wave function 3353, 3876, 7382
biquaternionic 3876
nondifferentiable 3874
Wave magnitude calibration function 2480
Wave maker 6208
Wave of synchronized flow dissolution 9391
Wave of upstream transitions 9389
Wave packet 2042, 3353, 8105
Wave propagation 7894
Wave propagation in complex media: path and site Effects 4453
incoherently scattered wave energy 4454
probabilistic seismic hazard analysis 4464, 4466
radiative transfer theory 4457
scattering in inhomogeneous media 4456
seismic coda 4456
wave propagation in random media 4455
Wave travel time analysis 2392
Wave travel times 2390
Waveform inversion 7895, 9909
Wavelength 2485
critical 2485
Wavelet analysis 3762
Wavelet based minimal energy methods 10041
Wavelet coefficients 10032
Wavelet functions 2382
Wavelet masks 592
Wavelet thresholding 10035
Wavelet transform 1719, 2384, 3893, 3894, 3897, 6352, 6357, 8019, 8534, 10032
decimated 1721
dyadic 1725
Haar wavelet 8534
Morlet wavelet 8534
orthogonal 1721
overcomplete 1721
pyramidal 1730
undecimated 1730
Wavelet transform modulus maxima (WTMM) method 3768, 3894, 3898
Wavelet transform skeleton 3894
Wavelets 7, 1165, 1719, 3893, 3923, 5800, 5801, 6856, 6882, 8669, 10031, 10032
band-limited 1721
bi-frames 6897
biorthogonal wavelets 6856
box spline tight wavelet frames 5837
compactly-supported 1721
Daubechies wavelets 6865
Haar wavelet 6862
Meyer 1729
mother wavelet 10006
multi-wavelets 6857
multiresolution analysis 10006
orthonormal box spline 5835, 5836
orthonormal wavelets 6863
pre-wavelets 6868
q-dilated orthonormal wavelets 6877
scaling function 10006
tight wavelet frames 6866
Way 8280
Weak approachability 7639
Weak congestion 9369
Weak-coupling limit 7809, 7817
for a single particle 7819
for classical systems 7809, 7817
for quantum systems 7814
Weak (defect-mediated) turbulence 6195
Weak disorder 6501
Weak KAM theory 4540, 4542
Weak L_1 inequality 2936

Weak links 2733
Weak localization 7921
Weak-mixing 957
Weak monotonicity 5516
Weak observability 6373
Weak solution 4716, 6218
  viscosity 4546
Weak spring 5026
Weak subsolution, definition 4543
Weak ties 8254, 8263, 8268
Weak topology, symbolic dynamics 6943
Weak turbulence 6201
Weak* topology 431
Weakly-minimal realization 6375
Weakly periodic point 9250
Weakly wandering 5458
Wealth 2810, 2811, 2823
Wealth distribution 2810, 2811
Wealth of nations 8994
Wealth temperature 2812
Weather 239, 241
  CCFP 241
  delay minute 241
  forecast 239
  impact 241
  scenario 239
  severe 241
Weather extremes
  mortality 1097
Web 2.0 8096, 10069
Web crawler 10062
Web graph 1246, 1248, 5266, 5269
Web link analysis 5267
Web science 8085
Web technology 4342
Weblogs 10067
  livejournal 10067
Wedge mechanics 1221
Weibull type 3320
Weidlich 6384
Weierstrass 6664
Weight 6501, 8246
Weight elasticity 8964
Weight matrix 8314
Weighted graphs 4828
Weighted holistic invariant molecular (WHIM) 2209
Weighted information
  on graph vertex distances 4831
  on vertex degree distribution 4829
Weighted Lehmer mean 229
Weighted mean 229
Weighted network 2718, 6048, 6057
Weighted probability 4828
Weiss B 780, 784–786
Weiss's microbial colony language 260
Weizsäcker's equation 4825
Weka 2124
Welfare loss 1673
Welfare policies 7059
Well log data 7918
Well-mixed hypothesis 8735
Well-posedness 6216
  local 2040
Wellman, Barry 8269, 8270
Wentzel–Bardeen instability 9687
West Coast/Alaska tsunami warning center 9596
Western electric research 8350
Wetting 2022
Wetting properties 3638
Weyl chamber flow 3099
Weyl pseudometric 2236
Weyl quantization 6758
Weyl's criterion 3088
What if scenarios 7058
Wheeled calculator 9705
Wheeler 8843
Whirlwind 7835
Whistlers 24
White noise 3577, 3804, 4595, 6233, 8776
  biological fractals 3794
White, Harrison 8268, 8273
Whitesides 5479
Whitham averaged Lagrangian 10000
Whitham equation 9998, 10000
Whitham's averaged variational principle 10000
Whitham's conservation equations 10000
Whitham's equation 9999, 10002
Whitney 704
Whitney smoothness 5080, 6325
Whitney umbrella 6632
Whole 1540
Whole-arm grasping 1476
Whole arm manipulation 1225

Wide moving jam 9356, 9357
Wide moving jam traffic phase 9303, 9308, 9337
Widening SP 9364
Wiedemann's psychophysical traffic flow model 9335
Wiener index 6838
  a physical meaning of 6848
  high-ordered 6849
  of a hydrocarbon molecule 6838
  of a linear graph 6845
  of graphs with linear branches 6842
  of stars 6843
Wiener–Khinchine theorem 6142, 6145
Wiener Lemma 8556
Wiener process 3438
  stochastic processes 3578
Wiener theorem, probability 6301
Wiggins SV model 8787
Wigner crystals 2027
Wigner distribution function 7412, 8933
Wigner equation 7814
Wigner function 7170
Wigner surmise 7164, 7506, 7514
Wigner transform 7814
Wigner's theorem 6729
  example of 6727
Wikipedia 10068
  degree distribution 10068
  motif profile 10068
Wilson 6665
Window 1724, 1801
  spatial 1724
Winfree 5478, 5479
Winner-takes-all 1782, 1821, 3209, 3218, 6106, 9947
Wireless local area network (WLAN) 3689
Wiring 5864
WKB approximation 6749
  connection formula 6751
  semiclassical solutions in the classically allowed region 6749
  semiclassical solutions in the classically forbidden region 6751
WLC *see* Worm-like chain (WLC)
Wold representation *see* Vector moving average representation
Wolfgang Köhler 8939
Wolfram 8846, 8847
Wolfram classes 768
Wolfram's rule 150 3553
Wolfram's rule 60 3553
Woodcrest 5758
Woodgate Beach 3113, 3116, 3130, 3134, 3135
Word-of-mouth 1959
Words 1036
Worker node 156
Workforce 7835
World Bank 2816, 2820
World dynamics 8997
World meteorological organization 9613
World Organization of Volcano Observatories 9870
World Wide Web consortium (W3C) 8085
World Wide Web (WWW) 8086, 10058
  deep 8089
  of documents 8086
  shallow 8089
Worldwide standardized seismograph network 2481, 2503
Worm 4369, 4370
Worm-like chain (WLC) model 6814, 7033
Worst-case studies 4051
WOVOdat 9870
WOWA operator 230
Wrapper 1776, 5351
Wrapping 1729, 5728
Wright–Fisher model 8771
Writhe 9271
Wronskian 3265
WSQ standard 10006
WT skeleton 3897, 3898, 3900, 3910
WTMM method 3904

X

X-Ray 22
X-Ray sources 392
XAFS 5890
Xanthan gum 9808
XML (extensible markup language) 355, 8085
  RDF/XML syntax 8092
XML schema 355
XOR problem 5357

xy-gate, quantum computing 7235
XY-model 1272

Y

Yanagihara 3553
Yang–Mills construction, mechanics 5495
Yang–Mills theories 6666
Yeast two-hybrid assay 5658
Yellowstone 2605, 2606
Yellowstone caldera 9868
Yield stress 9788
Yin–Yang computing 4403
Yomdin's theory 967
Yoshida 5472
Young–Dupré condition 5583
Young score 9954
Young winner 9954, 9959
Young's dimension formula 3022
Young's modulus 691, 698, 699
Young's procedure 9954, 9959
Yukawa potential 8579
Yunès 3549, 3552, 3553

Z

Z-Gate, quantum computing 7233, 7235
Z-shaped speed-flow characteristic for traffic breakdown 9295
Zabusky, N.J. 81
Zadeh fuzzy rule base 4083
Zadeh, L.A. 6929
Zadeh's extension principle 8682
Zakharov–Shabat system 4962
Zareckii lattice 8295
Zeeman 704
Zeeman effect 6737
Zeeman's tolerance stability conjecture 969
Zeipel's theorem
  n-body 5978
  structure in the collision set 5978
Zero-blocks 8274
Zero controllability 4810, 4811
Zero divisor 779, 788, 790
Zero dynamics 9085
Zero knowledge quantum interactive proofs 7191
Zero-noise limit 963
Zero-one parameter 9034, 9040, 9043
Zero padding 10009, 10014
Zero-photon signal, quantum computing 7230
Zero-shear-rate viscosity 6849
Zero-sum game 3190, 4855, 7630, 8381
Zero-sum two-person game 4096
Zero temperature 5023
Zero-tree scheme 10036
Zeta function 3046, 8897
Zhreshold effects, critical mass 4617
Ziff model 888
Zigzag design 6491
Zimm matrix 6837
Zinbiel algebra 1028, 1036
Zincblende 5764
Zipf's-law 6528
Zipper effect 3149
Zipper mechanical topology 7039
Zoë 3690
Zones 4182
  boundary 4182
  of proximal development 1877, 1880, 1881, 1885
Zuse 8844
Zygote 526